4

Linear Accelerators

Robert Jameson, Joseph Bisognano, and Pierre Lapostolle[†]

4.1 Introduction 125
4.2 General Principles and Brief History 126
4.2.1 The Original LINACs 126
4.2.2 Phase Stability 126
4.2.3 Focusing 127
4.2.4 Electron Accelerators 128
4.2.5 Proton Accelerators 129
4.3 Electron Linear Accelerators 131
4.3.1 Electron Source and Preacceleration (LINAC Injector) 131
4.3.2 Electron LINAC Characteristics 132
4.3.3 Electron Beam Intensity Limitations 133
4.3.4 Uses of Electron LINACs 136
4.4 Proton Linear Accelerators 138
4.4.1 Injection into a Proton LINAC 138
4.4.2 RFQ Accelerators 139
4.4.3 Drift-Tube-Type LINACs 140
4.4.4 LINACs for Higher Velocity Ions 141
4.4.5 Beam Dynamics in Nonrelativistic LINACs 142
4.4.6 Proton LINAC Applications 145
4.5 Heavy-Ion Linear Accelerators 146
4.5.1 Special Difficulties Relative to Ions 146
4.5.2 Accelerating Structures – Normal and Superconducting 147
4.5.3 Application 147
4.6 Other Types of Linear Accelerators and New Methods of Acceleration 148
4.6.1 Induction LINAC 148
4.6.2 Alternating-Phase-Focused LINACs 148

[†] Deceased

Encyclopedia of Nuclear Physics and its Applications, First Edition. Edited by Reinhard Stock.
© 2013 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2013 by Wiley-VCH Verlag GmbH & Co. KGaA.

4.6.3 New Methods of Acceleration 149
4.7 Engineering Aspects 149
4.7.1 Accelerator Structure Design 149
4.7.2 Magnetics 151
4.7.3 Radiofrequency Engineering 151
4.7.4 Availability and Maintainability 152
4.7.5 Computers 152
4.7.6 Wide-Bandwidth Electronics and Instrumentation 152
4.7.7 Automatic Controls 153
4.7.8 Civil Engineering 154
4.8 Concluding Remarks 154
Glossary 155
References 157
Further Readings 158

4.1 Introduction

Particle accelerators were originally invented as enabling tools for physics research, for example, nuclear and elementary particle physics. Many, if not most, of the important concepts were invented for almost all accelerators and, in particular, linear accelerators (LINACs), around 1930. However, technological capabilities at that time did not permit anything but a proof of principle of LINAC operation. By using the term *LINAC*, we mean devices where, typically, the particle passes only once, on a straight trajectory through a possibly long structure to gain energy, and contrast this to ''circular'' devices such as synchrotrons and cyclotrons, where the particle passes a great number of times through the same accelerating structure. Accelerators such as microtrons and recirculating LINACs, where particles travel only a ''few'' times through the accelerating structure, show the limitations of such distinctions. With one or only a few passes of acceleration, LINACs typically provide the highest throughput of particles at a given energy.

After the development of high-power short-wave and microwave radiofrequency (RF) sources for radar, it was possible to undertake the construction of useful LINACs. Successive improvements pushed by continuous requests for still more energetic and powerful beams of better and better quality led from the initial crude machines to the very elaborate LINACs of today. In the past 20 years, the use of superconducting materials such as niobium has been a most significant innovation.

Besides being used in nuclear and particle physics research, LINACs are now also used as a source of energetic particles in equipment for medical, biological, and industrial applications, including industrial irradiation, radiotherapy, medical tool sterilization, food conservation, and baggage inspection, and provide intense beams of neutrons or photons for use as probes for condensed matter, materials, and in biological sciences. In the future, LINACs will enable methods to deal with radioactive waste and other societally important applications.

In this chapter, Section 4.2 explains the general principles of linear acceleration. A historical approach has been taken. Both the evolving goals and the difficulties encountered that have led to the sophisticated and complex designs of today's electron and proton LINACs are explored. Electron LINACs are described with some details of their components, limitations, application areas, and advanced research for the future. Proton LINACs are

Encyclopedia of Nuclear Physics and its Applications, First Edition. Edited by Reinhard Stock.
© 2013 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2013 by Wiley-VCH Verlag GmbH & Co. KGaA.

described next, explaining in more detail the problems encountered at both low and high energies. Again, intensity and beam quality limitations are considered. Present and future applications are presented.

Heavy-ion LINACs, which have some special considerations and unique realizations, represent the final general category of LINACs.

The induction LINAC and alternating-phase-focused (APF) LINACs are covered briefly. The work being done on the search for new methods of acceleration aimed at reaching still higher energies in shorter distances for future generations of colliding beam machines for high energy physics and for tabletop devices for small laboratory research is described.

The final section, on engineering topics, covers the various aspects of LINAC construction: accelerating structures, magnetics, RF engineering, wideband electronics, instrumentation, automatic controls, and computers.

4.2
General Principles and Brief History

4.2.1
The Original LINACs

In the first part of the twentieth century, the development of nuclear physics and, later on, the discovery of induced radioactivity prompted the search for energetic projectiles to produce new nuclear reactions. The first particle accelerators were soon made from DC high-voltage generators, but source voltage and, therefore, particle energy were limited by electrical breakdown. To obtain more energetic projectiles, new methods had to be invented.

The proposal by Ising [1] that led to the development of LINACs called for an array of metallic tubes with an applied potential creating an electric field between tubes, but not inside them. Charged particles launched along the axis of the tubes were to be accelerated while traversing the gap between two adjacent tubes but drifted at constant velocity when inside a tube. Ising suggested that pulsed potentials be applied across the drift tubes at the time the charged particle bunches would enter the gap, and be switched off once they entered the tubes. Such a method, difficult to apply accurately enough then, is today the principle of the induction LINAC, mentioned in Section 4.6.1. Instead of using very accurately synchronized high-voltage pulses, Wideroe [2] proposed more simply to apply RF voltages on a succession of drift tubes, as shown in Figure 4.1. He tested a small device with two gaps and a single drift tube, and, with an RF peak voltage of 25 kV, accelerated Na^+ and K^+ ions produced by a heated filament to an energy of 50 keV.

Sloan and Lawrence [3] at Berkeley used 30 tubes of progressively increased length and a gap voltage of 42 kV at an RF frequency of 10 MHz to reach an energy of 1.25 MeV for Hg ions.

4.2.2
Phase Stability

To produce not only a few energetic particles but also a good useful beam, the following problems have to be resolved:

- how to obtain a large enough number of particles (*beam intensity*)?
- how to concentrate the particles – or focus them – into a thin enough beam (*beam quality*)?

Let us look further into this succession of accelerations between drift tubes at two

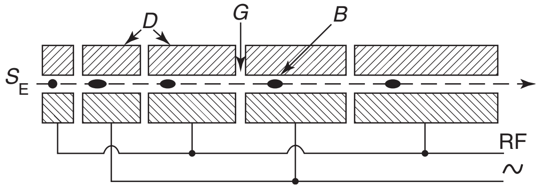

Figure 4.1 Principle of the Wideroe method of RF linear acceleration used in 1931 on the first machine of Sloan and Lawrence. *S* is a source of positive ions. *D* represents drift tubes alternatively connected to the two terminals of an RF voltage. The ion bunches *B* are accelerated in the gaps *G*. Because the velocity increases as the bunches are accelerated, the length of the drift tubes must increase along the machine.

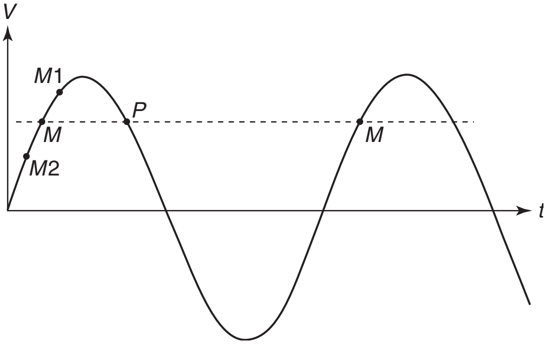

Figure 4.2 The voltage *V* in the gap *i* between the drift tubes has a sinusoidal variation with time *t*. To reach gap $i+1$ at the right time, the proper acceleration is necessary; the required amplitude occurs at two phases per period: *M* and *P*. While points *P* are unstable, points *M* are stable and particles form bunches around such phases.

consecutive gaps numbered i and $i+1$, where a sinusoidal RF voltage is applied (Figure 4.2). Suppose a particle arrives at gap i with the correct energy. The length of the drift tube between the gaps i and $i+1$ has been chosen for a given energy increase through gap i. Such a gain is produced according to Figure 4.2 at only two phases on the RF cycle: phases *M* and *P*. If particles crossing the gaps at these exact phases were the only ones that could be accelerated properly, the beam intensity would be extremely small. Let us examine, however, what happens around phase *M*, say at *M*1 and *M*2. A late particle at *M*1 is accelerated more and catches up with *M*. An early one at *M*2, on the contrary, is slower and also meets with *M*. It is said that there is phase stability around phase *M*. It can be seen that around *P* there would be instability. Thus, particles are caused to bunch around phases such as *M*, where the voltage is rising, and a more intense beam can be accelerated.

4.2.3
Focusing

Let us look now at focusing. Figure 4.3 sketches an electric field distribution in an accelerating gap; off-axis particles are focused at the gap input and defocused at the exit. For phase stability, however, at phase *M* (Figure 4.2), the field must be rising with time; defocusing is then larger than focusing. The overall effect of accelerating gaps is defocusing.

Additional focusing has to be provided, particularly for low-velocity particles [the velocity is small compared to the velocity of light: $v=\beta c$ (where $\beta \ll 1$)] that are nonrelativistic and are easily deflected by transverse electric fields. At the highest energies, focusing is introduced with

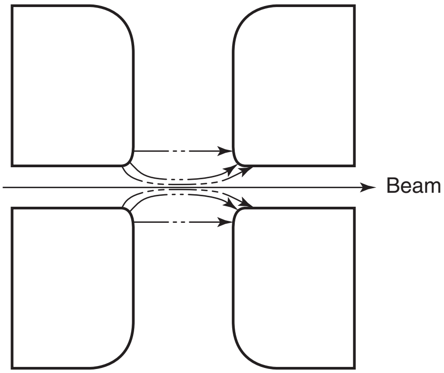

Figure 4.3 The electric field lines in the gaps terminate normally on the metallic surfaces of the drift tube holes. There is a focusing action at gap entrance and defocusing at gap exit if the gap is accelerating. If the voltage is increasing while particles cross a gap, the net effect is defocusing.

quadrupole magnets, typically in an alternate gradient configuration.

4.2.4
Electron Accelerators

In electron LINACs, the velocity rapidly approaches the velocity of light. The scheme of Wideroe is not easy to use; extremely high frequencies are needed, and the drift tubes become resonant antennas supporting very high currents with high losses (Section 4.4.3). Another method had to be invented. Using high-power microwave radar sources available in 1947, Ginzton *et al*. [4] designed an accelerating structure consisting of a stack of pillbox resonators with a central hole, called the *iris-loaded waveguide* (Figure 4.4). The holes provide a passage for the beam and also allow the electromagnetic energy to travel from resonator to resonator. The cavity array behaves as same as an RF filter of relatively narrow bandwidth; the phase relation between two successive cavities depends on the frequency. With a proper choice of the dimensions, at a given frequency, the field in successive cavities can be such that a traveling wave is generated with phase velocity equal to that of the electron beam: this synchronization ensures acceleration in every cell. The waveguide is interrupted at intervals to replenish the structure with energy and to install quadrupole magnets that provide the extra focusing needed for transverse stability.

Slater [5] at MIT, Walkinshaw in England, and others derived the theory of the iris-loaded waveguide and studied many other types of structures. Further developments at Stanford led to the longest electron LINAC to date; the 3 km long Stanford Linear Accelerator Center (SLAC) LINAC has accelerated electrons and positrons to 50 GeV.

In the 1980s, decades of work on superconducting accelerating structures came to fruition, with application to both circular [the CERN LEP (large electron–positron) collider] and the Continuous Electron Beam Accelerator Facility (CEBAF, now Jefferson Laboratory) recirculated LINAC [6]. For continually operating [continuous wave (c.w.)] or long-pulse LINACs of high Q, standing wave, multicell resonators rather than traveling wave structures are used. With the decision to use superconducting technology on the International Linear Collider (ILC) [7], free electron lasers (FELs) as optimal drivers [8], and energy recovery light sources [9], superconducting standing wave LINACs will likely become

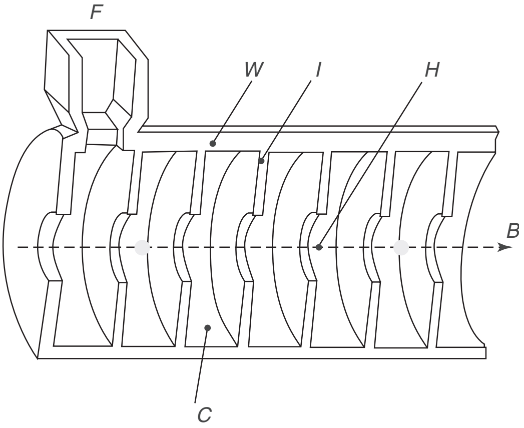

Figure 4.4 The cylindrical waveguide *W* is loaded by irises *I*; the holes *H* provide coupling between the cells *C* for electromagnetic energy transfer and beam passage along the axis *B*. The structure can also be considered as a stack of pillbox resonators *C* coupled through the axial holes *H*. Electromagnetic energy is fed at one end from a feeder *F*, through a ceramic window (not shown) to keep the interior of the structure under high vacuum.

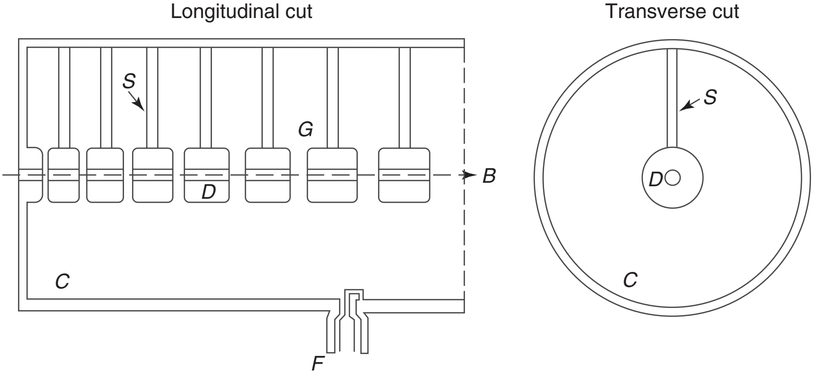

Figure 4.5 Alvarez structure. Drift tubes *D* supported by stems *S* are put along the axis of a cylindrical cavity *C*. The cavity is supplied with RF power from a coaxial line *F* (through a window not shown) by a coupling loop, and excited in a longitudinal mode such that the electric component of the electromagnetic field is directed along the axis and concentrated in the gaps *G*. The field varies sinusoidally with time but the drift tube lengths are chosen so that the beam is cumulatively accelerated along the axis *B*.

the dominant technology for large electron accelerators at the highest energies. Room-temperature standing wave structures (often of a side-coupled configuration) also have applications, especially as part of circular accelerators.

4.2.5
Proton Accelerators

For protons of low velocity, in 1947, Alvarez *et al.* [10] designed a 200 MHz structure at Berkeley, using surplus radar components. In a 12 m long, 1 m diameter resonant cavity, excited in a mode with uniform electric field on the axis, a succession of metallic drift tubes with axial holes provided cumulative acceleration of the beam (Figure 4.5). At first, additional focusing was provided with the help of flat foils or grids located at the exit of each gap (Figure 4.6). These devices suppressed the defocusing effect described in Figure 4.3, but stopped a nonnegligible fraction of

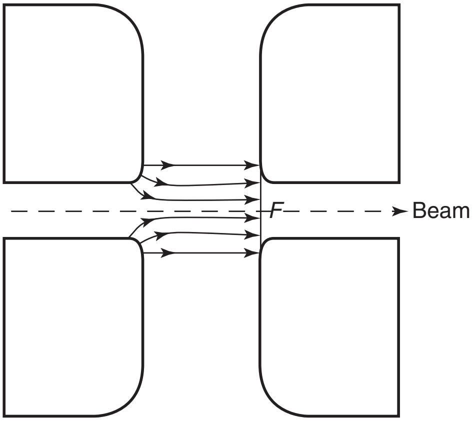

Figure 4.6 Foil (or grid) focusing. By closing the hole at the exit of the gap, the foil or grid suppresses the defocusing effect (Figure 4.3). The gaps are now focusing, but the beam is scattered and some particles are lost.

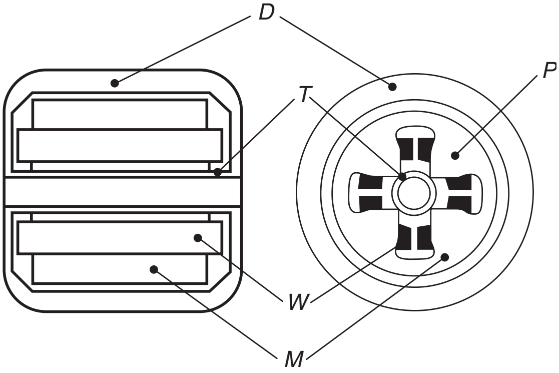

Figure 4.7 Magnetic quadrupole lens inside a drift tube. The windings *W* create a quadrupolar field between the four poles *P*. The magnetic flux returns through the magnet yoke *M*. The whole lens is encapsulated inside the drift tube *D* closed on the axis by a thin tube *T*.

the particles and deflected others in an irregular way, reducing the intensity and the beam quality.

It was then discovered by Courant *et al.* [11] that effective transverse focusing could be provided by a succession of focusing and defocusing quadrupole magnets, and Blewett [12] suggested installing such magnets inside the drift tubes. A four-pole magnet (Figure 4.7) constitutes a focusing lens in one transverse direction and a defocusing lens of equal strength in the orthogonal one. A succession of focusing and defocusing lenses, properly spaced, is globally focusing; with quadrupole magnets, one thus gets focusing in the two transverse directions [13]. This method is called *alternating-gradient* (AG) focusing.

The Alvarez structure, with various improvements, is used now for ion beam β values in the range 0.035–0.5; other types have been developed for lower and higher energies (Section 4.4.4).

As with electron accelerators, superconducting radiofrequency (SRF) LINACs have found important applications for protons and ions, with the Spallation Neutron Source (SNS) being the largest application to date [14]. A proposed rare isotope accelerator (RIA) using SRF is also under consideration to generate short-lived ion species.

4.3
Electron Linear Accelerators

4.3.1
Electron Source and Preacceleration (LINAC Injector)

Electrons are very light particles. According to Einstein's energy relation, one has $E = E_o + W = mc^2 = m_o c^2/\sqrt{1-\beta^2} = E_o/\sqrt{1-\beta^2}$, where the total energy E is the sum of the rest energy E_o (equal to $m_o c^2$) and kinetic energy W, where m is the total mass and m_o the rest mass; $\beta = v/c$ is the ratio of the velocity of the particle to the velocity of light. For electrons, $E_o = m_o c^2 = 511\,\text{keV}$. Table 4.1 shows how β changes with kinetic energy W.

As can be seen, with an energy gain in typical modern electron LINACs of $10\,\text{MeV}\,\text{m}^{-1}$ or more, the velocity is almost everywhere very close to the velocity of light, apart from a first section called *the injector*. A high-energy electron LINAC is then an injector followed by a number of identical sections of an accelerating structure, with the addition of some focusing. The simplest form of electron injector consists of a thermionic cathode emitting electrons into a DC voltage of 20–150 kV. A grid pulser modulates the beam to form "bunches" of adjustable length and amplitude in the range of microamperes to kiloamperes for certain applications. The conventional buncher section of an RF LINAC is an accelerating structure (Figure 4.8) with variable β bringing the energy to the order of 10 MeV and compressing the output bunches to 10° or less. Focusing is generally provided by magnetic solenoids.

Over the past few decades, electron injector technology has been revolutionized by the introduction of the laser-driven

Table 4.1 β as a function of W for electrons.

W (MeV)	0.020	0.05	0.10	0.20	0.50	1	2	5	10
β	0.272	0.413	0.548	0.695	0.863	0.941	0.979	0.996	0.999

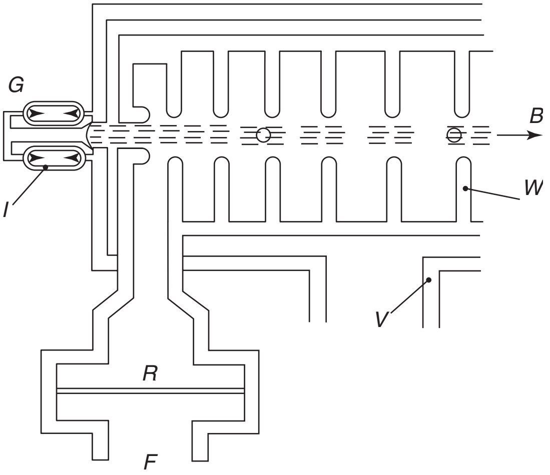

Figure 4.8 In this very schematic drawing of the bunching section of an electron LINAC, the gun *G*, at high voltage, is supported by an insulator *I*. The iris-loaded waveguide *W* is put inside a vacuum vessel *V* evacuated to high vacuum. A feeder *F* brings microwave power to the accelerating structure through the window *R*. The beam *B* is accelerated in the form of bunches. The details of the gun, the grid, and the chopper or prebuncher, as well as all the focusing elements, have been left out of this drawing.

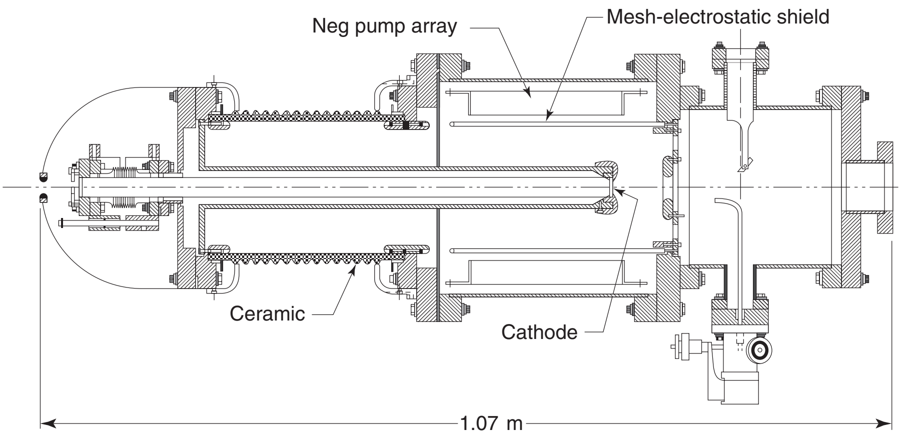

Figure 4.9 Schematic of Jefferson Laboratory polarized electron source. (Courtesy: Jefferson Laboratory.)

photocathode. This advancement has allowed the production of electron beams of significantly higher phase space density (smaller beam emittance) and shorter pulse lengths. For RF LINACs, the laser can be modulated to produce short electron bunches occupying only a narrow range about the proper RF phase. The photocathode is located at the entrance of an RF accelerating cell gap or a DC field (Figure 4.9). With high gradients, the electrons are accelerated very rapidly, helping to alleviate the effects of space–charge and preserve beam brightness. Bunch charges ranging from pico- to nanocoulombs can be obtained with normalized emittances below 10 μm at peak currents of tens of amperes. With the introduction of DC voltages approaching 1 MV or RF gradients of several tens of megavolts per meter, R&D is progressing toward submicrometer normalized emittances. The success of this effort will strongly impact the future of light sources such as FELs and energy recovery linear accelerators (ERLs).

Moreover, by using a photocathode and a laser, it is possible to produce longitudinally polarized electrons, which are of great interest in certain physics experiments. For example, circularly polarized near infrared (IR) light from a Ti:sapphire laser is projected onto a gallium arsenide (GaAs) cathode in the CEBAF polarized electron source [15].

4.3.2
Electron LINAC Characteristics

An RF electron LINAC consists of one or several sections of an accelerating structure, each generally several meters in length. Accelerating structures for short-pulse LINACs are of the iris-loaded waveguide, traveling wave type. They are usually operated in the microwave band around 10 cm wavelength (3 GHz). The shape of the cavities has been optimized to reduce (for a given accelerating gradient) both the peak electric field on the surfaces (to avoid the risk of electrical breakdown) and the RF power losses in the walls resulting from the finite conductivity of the material (usually copper). The SLAC cavities are pulsed for a few microseconds in a traveling wave mode, the first microsecond being used to fill the structure

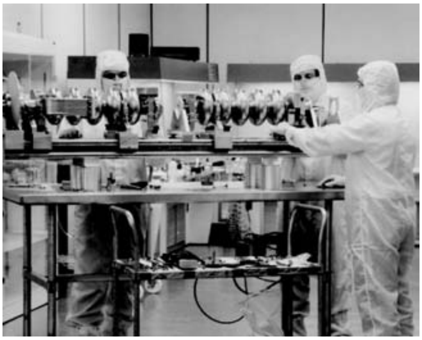

Figure 4.10 Assembly of CEBAF cavity in a clean room. (Courtesy: Jefferson Laboratory.)

from input to output (3 m). Only 1/10th or at most a few tenths of the power remains at the output, the rest being absorbed by losses in the walls or transferred to the beam. Electrons at relativistic velocity ride on the RF wave crest. As the RF cavity fields are raised by increasing the RF source power, the energy gain per unit length is raised accordingly.

For longer pulse and up to fully c.w. operation, RF LINACs are made with standing wave cavities. For optimal coupling to the electron beam, the side-coupled structure (Section 4.4.4) is typically chosen for normal conducting applications. The introduction of superconducting materials such as niobium has had a dramatic impact on standing wave LINAC technology. At RF, superconductors do still exhibit small losses, but resonators can offer *Q*s that approach 10^{10} at gradients of order $20\,\mathrm{MV\,m^{-1}}$. This is to be compared to a typical room-temperature copper structure where *Q*s of 10^4 are more typical. These low losses have enabled superconducting LINACs to be operated in a c.w. mode at these high gradients, which would be entirely impractical for room-temperature structures. True c.w. operation (and also the inherently high *Q*) provides at least an order of magnitude improvement in beam stability, which in turn has enabled applications that require smaller energy spread and beam temporal jitter. In fabricating superconducting cavities, much care must be taken to ensure the quality of the superconductor with minimization of imperfections in the material, meticulous assembly in clean-room environments, conditioning, and cool-down protocols (Figure 4.10).

Energy in big LINACs is mainly limited by cost considerations involving the length of the structure and the amount of RF power. In short-pulse machines such as SLAC, the energy can be increased economically by using a device called the *Stanford Linear Accelerator Center Energy Doubler* (SLED) [16, 17]. It makes use of high-*Q* resonant cavities that are filled with electromagnetic energy during the first part of the RF pulse. In the second part, the RF phase of the klystron is reversed and the cavities are automatically emptied, effectively resulting in pulse compression and a momentary reflected power increase by a factor of almost 3. With the help of 3 dB couplers, this power can be fed into the accelerating structures, and, over a very short pulse, can result in about 80% more acceleration.

4.3.3
Electron Beam Intensity Limitations

In applications such as high-energy LINACs for physics experiments or an

RF LINAC providing the driving beam for an FEL, beams of extremely good quality are required: very small transverse beam dimensions (measured in a tenth or hundredth of a millimeter or even in micrometers) and very small energy spread (0.1% or less). In these and other applications, beam intensity must also be as high as possible. Energy spread from the electron bunch phase width along the RF wave can be compensated using very short bunches and by ensuring that the bunches, on average, ride on the crest. However, to achieve the shortest bunch lengths, designs for FELs and colliders use bunch compression, which requires an energy chirp of the bunch that has passed through dispersive optics. The chirp is typically produced by off crest acceleration.

One limitation in short-pulse electron LINACs comes from beam loading. To be accelerated, the beam has to extract energy from the electromagnetic field, thus reducing the field strength in proportion to the beam intensity. As a result, under short-pulse operation, it is not easy to obtain a flat energy gain over the pulse. For long pulses or c.w., automatic control of the fields can be applied and then heavy beam loading (i.e., a large ratio of power supplied to the beam to that supplied to the accelerator structure losses) is used to raise the overall efficiency. In superconducting LINACs, for example, the structure losses are miniscule in comparison with the power transferred to the beam.

With the ever-decreasing bunch length and emittance and ever-increasing peak currents, wakefield phenomena and coherent synchrotron radiation (CSR) processes often limit the overall beam performance. As a generalization of beam loading, whenever a beam pulse passes through variations in its vacuum environment (e.g., the RF structure, bellows, and beam position monitors), the image charges change direction and radiate. The electromagnetic energy comes from the beam's kinetic and field energy. With a frequency content determined by the details of the environmental discontinuities, the electron bunch may find its momentum (both longitudinal and transverse) perturbed with a time variation. This is the basic notion of wakefield: the field generated by the beam, which is left behind and can act back on succeeding beam particles. A related but somewhat different notion is that of CSR. A single electron will radiate when it is accelerated, for example, on being steered along a curved path by a bending dipole magnet. If the electron is relativistic, this radiation has a very high frequency content (γ^3 of the typical revolution frequency), and the characteristic wavelength is typically much shorter than current modulations on the bunch or the bunch length itself. However, in the short bunch regimes typical of cutting-edge LINACs, the bunch length, in fact, is shorter than the characteristic wavelength of the radiation. The bunch can radiate as a single charge, with the radiation power enhanced by a factor corresponding to the number of electrons radiating in concert. Again, if the electron bunch is emitting energy, the beam energy must be diminished and will have a temporal variation. This is CSR.

The wakefields and CSR cause momentum changes that vary along the length of the electron bunch. This effectively increases the bunch energy spread and transverse emittance, which can destroy necessary resolution for spectroscopy and degrade performance of FELs. Moreover, the momentum variations can be exchanged for position variations in magnetic transport systems, yielding pulse lengthening, and spot size increases. Finally, beam instabilities can be generated, which can grossly degrade beam properties.

Another serious limitation comes from a class of phenomena called *beam breakup* (BBU), which results in beam deflection and eventual particle loss in the accelerator walls. The LINAC structure can be excited not only on the fundamental accelerating mode but, just the same way as an ordinary waveguide, in other modes, the most dangerous of which are transverse hybrid modes that deflect the beam. These modes can be excited by particles traveling off the cavity axis, or as a result of cavity asymmetries. Along a single accelerator section, the interaction of the beam with such a transverse mode can lead to uncontrollable beam deflection when the current is above a certain threshold. The deflected beam can couple its energy into the transverse mode, which has a longitudinal electric field component off axis. A feedback loop is generated, which can go unstable at high currents (effectively, high gain). This is called *regenerative BBU*. In a long accelerator, each cavity, although below the threshold of this unstable oscillation, can nevertheless behave as same as an amplifier. A long chain of such amplifiers, even if each gain is small, can eventually deflect the beam beyond the edge of the irises; this is called *cumulative BBU*. The wakefield effect discussed earlier leads to excitation of higher order deflecting modes, which can affect succeeding beam bunches. These wakefields can build as the chain of bunches get deflected and more strongly excite downstream cavity modes. Their integrated effect on particles passing through them can cause effective emittance growth that at least dilutes brightness, or may even cause beam loss. The transverse effects are combated by eliminating boundary discontinuities wherever possible, and by making geometrical changes in the accelerating cavities to prevent the beam from being synchronized with the offending modes. The effects in RF structures become more serious as the operating frequency is raised and the iris diameter becomes smaller, and this becomes a limitation in high brightness designs. In the design of c.w. LINACs using the latest generation of superconducting RF cavities, either cost optimization or energy recovery for high average currents favors configurations where the beam passes several times through the same accelerating structure. It was discovered during the commissioning of the Stanford Recyclotron that a new kind of instability, a hybrid between regenerative and cumulative BBU, can limit beam average currents. As in cumulative BBU, a series of marginally offset beam bunches can have their displacement amplified by a higher order mode (HOM) when passing through a long series of accelerating cavities. In addition, with recirculation, these offset bunches are allowed to pass through a given cavity for a second time, where the offset beam can excite the same HOM as happens in regenerative BBU. A closed feedback loop is formed, which can go unstable if the time delay is unfortunate and the beam current (which is equivalent to loop gain) is high enough. This ‘‘multipass BBU’’ is the principal limitation to average beam current in recirculated LINACs. Since superconducting cavity HOMs have relatively high Qs, prevention of BBU requires careful damping of the strongest modes with mode couplers. For applications in the milliampere range, this has proved to be sufficient to provide beam stability, for example, in the Jefferson Laboratory nuclear physics machine CEBAF. For ERLs to drive high-average-power FELs (e.g., the Jefferson Laboratory IR source) or for ERLs to generate incoherent synchrotron radiation (e.g., Cornell ERL) [9], average currents of order 100 mA are envisioned, and BBU and HOM damping must be monitored closely in cavity design and construction to ensure success.

4.3.4
Uses of Electron LINACs

Most of the pioneering electron LINACs in the world have been built for physics experiments. Table 4.2 lists four of the largest.

An important feature of large electron machines is the possibility of producing and accelerating positrons. Such particles are produced from the impact of high-energy (33 GeV at SLAC) electrons on a converter from which $e + e^-$ pairs emerge. Production efficiency is proportional to incident electron energy and is usually in the range of a small percent per incident gigaelectronvolt. Since the mid-1960s, energetic electrons and positrons from LINACs have been frequently used to fill collider rings (where counter-circulating beams collide, maximizing available energy) for physics experiments and to fill storage rings for the generation of synchrotron light.

In the early 1980s, the requirements for performing electron–nucleus coincidence experiments to look into the quark nature of the nuclear force and for very high resolution spectroscopy of nuclear states pointed out the need for high-duty factor electron accelerators. Although first conceived as a pulsed LINAC filling a pulse stretcher storage ring, it soon became clear that c.w. operation of a superconducting LINAC would offer the best, most upgradable solution. The CEBAF was designed as a multipass recirculated LINAC, with ultimately five beams of distinct energy being accelerated simultaneously through superconducting cavities originally designed for the Cornell electron positron storage ring (CESR) (Figure 4.11). Initially, the gradient in c.w. operation was specified at 5 MV m^{-1}. The electron beam (now from a photocathode polarized source) is sent up to 5 times through two antiparallel 400 MeV superconducting LINAC segments connected by beam lines to deliver a 4 GeV, 200 μA average current continuous beam as designed. The LINAC is constructed using 1.5 GHz, five-cell niobium cavities operating at a temperature of 2 K. Advances in SRF technology led almost immediately to operation at 6 GeV, and an upgrade to 12 GeV is planned for the near future, with average gradients nearing 20 MV m^{-1} for supplemental seven-cell cavity cryomodules. In addition, the technology was applied to smaller recirculated LINACs at Jefferson Laboratory, configured in an energy recovery mode, where the beam passes a second time through the accelerating structure 180°

Table 4.2 The four largest electron LINACs.

Laboratory	Energy (GeV)	Length (m)	Date of completion
Stanford (California)	50.0	3050	1966
Tsukuba (Japan)	2.5	320	1981
Orsay (France)	2.3	230	1959
Kharkov (USSR)	2.0	240	1964

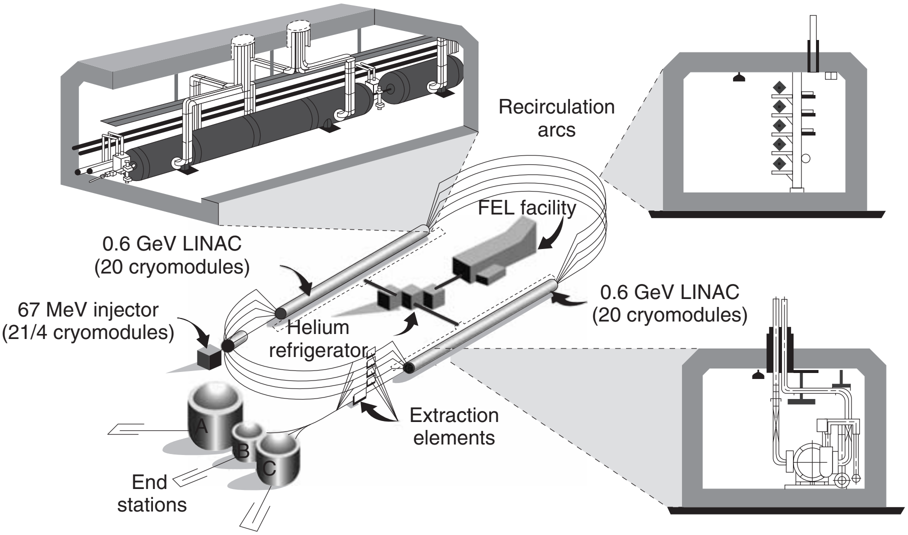
Recirculation arcs
FEL facility
0.6 GeV LINAC (20 cryomodules)
67 MeV injector (21/4 cryomodules)
Helium refrigerator
0.6 GeV LINAC (20 cryomodules)
Extraction elements
A
B
C
End stations

Figure 4.11 Configuration of the CEBAF recirculated LINAC at Jefferson Laboratory. (Courtesy: Jefferson Laboratory.)

out of phase with the field to deposit its kinetic energy for reuse. In this way, very high currents (approaching 10 mA) can be sustained for high-average-power FEL operation without making absurd demands on RF power. This notion of energy recovery is also of interest in the design of LINAC-based synchrotron light sources where fresh, low emittance electron bunches are sent through undulators, periodic arrays of magnets that coherently enhance the synchrotron radiation process.

After the initial findings of the large hadron collider (LHC) in the decade that followed, exploration of the high-energy frontier in detail will require a next generation of electron–positron colliders. Since 1980, high-energy physicists, seeking methods to collide electrons and positrons in the center-of-mass range of 100 GeV and higher, have reached the conclusion that above a certain energy level, linear colliders (two opposing LINACs whose beams collide at a target point) have an economic advantage over colliding rings. This advantage arises from the fact that a LINAC beam only loses a negligible amount of energy through synchrotron radiation because the particle trajectories are not bent in circles. Much R&D has gone into finding the best technology for linear colliders, and in the 1980s–1990s, work progressed on room-temperature short-pulsed, higher frequency copper LINACs, superconducting long-pulsed LINACs, and two-beam acceleration as potential approaches. In 2004, an international committee determined that long-pulsed superconducting RF technology was the most attractive candidate and an international collaboration has been created to bring this exciting idea to fruition.

The ILC is conceived as a positron–electron collider offering center-of-mass energies of 200–500 GeV, and upgradable to 1 TeV. Polarized electron beams are envisioned. In pulsed mode, the superconducting structures are expected to achieve gradients in excess of $35\,\mathrm{MV\,m^{-1}}$, with a theoretical limit of about $50\,\mathrm{MV\,m^{-1}}$.

Apart from physics experiments, many electron LINACs of lower energy are also now made for a host of applications. Medical or industrial irradiation, radiotherapy, industrial radiography, chemical reactions, or hardening processes of various materials induced by irradiation, medical sterilization, and conservation of agricultural products through ionization are some of the various applications that have been developed and are more and more extensively in use. Almost all hospital X-ray machines now in use contain small standing wave RF electron LINACs. For such applications, high power (several tens of kilowatts of beam power) and low energy (in most cases ≤10 MeV) are needed; beam quality is not critical, and the beam is scanned or widened to cover the needed irradiation area. The crucial parameter is average intensity, and costs are optimized in this respect. Sources based on laser Compton backscattering are also under development.

A special type of application of LINACs is flash or stroboscopic radiography; there, extremely intense but very short pulses (100 ns, for instance) are needed. Induction LINACs (Section 4.6.1) are used for this application, and RF LINACs can also be used. The Dual Axis Radiographic Hydro-Test (DARHT) Facility project at Los Alamos National Laboratory (LANL) represents the latest advance, with two induction LINACs producing 20 MeV electrons at 4 kA.

4.4
Proton Linear Accelerators

4.4.1
Injection into a Proton LINAC

Unlike electrons, protons need to have a very large kinetic energy to reach a velocity close to that of light. With a rest energy $E_o = 938\,\mathrm{MeV}$, the relation between β and W is shown in Table 4.3.

Even at the highest energy reached in proton LINACs today, the final velocity is <90% of the velocity of light. At injection, it is only a few hundredths of c, increasing along the machine at a typical accelerating gradient of around $2\,\mathrm{MeV\,m^{-1}}$ according to usual practice.

As explained in Section 4.2, protons, especially at low energy, are subject to a strong defocusing effect of the electric field, not at all compensated by the effect of the magnetic component. Furthermore, the magnetic field strength of static focusing lenses must increase as particle velocity decreases, and there is insufficient room in the drift tube to provide it. This led to the choice of at least 500 kV for the injection of protons in LINACs built before ~1980.

The source of protons, from a DC or RF discharge in hydrogen, was traditionally of the duoplasmatron type; that is, one in which the axial magnetic field is made to vanish at the extraction hole from an expansion cup. Such a source can produce a beam of very high quality with intensities of several hundred milliamperes of protons either pulsed or c.w. This source

Table 4.3 β as a function of W for protons.

W (MeV)	0.1	0.2	0.5	1	2	5	10	20	50	100	200	500	1000
β	0.015	0.021	0.033	0.046	0.065	0.103	0.145	0.203	0.314	0.428	0.566	0.758	0.875

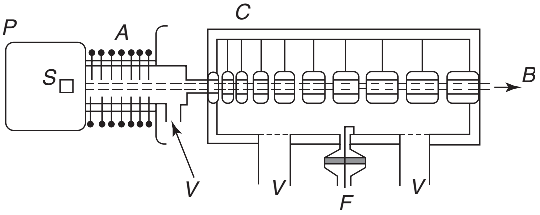

Figure 4.12 Schematic drawing of a proton LINAC with a high-voltage injector. The source *S* is placed on a high-voltage platform *P* and the beam is injected into the Alvarez structure (at ground potential) through a DC accelerating column *A*. A high vacuum is maintained all along the machine by vacuum pumps *V*. A coaxial line brings in *F* the RF power into the Alvarez cavity *C* by a coupling loop; vacuum tightness is provided by a window. Prebuncher and focusing elements are not shown in the drawing.

was traditionally put on a high-voltage platform at ~750 kV DC, and the beam injected from it into a drift-tube accelerator through an evacuated accelerating column (Figure 4.12).

Modern ion sources of the electron cyclotron resonance (ECR), the electron beam ion source (EBIS), and other types are used to produce intense beams of H^+, H^-, D^+, and other ion species; research continues – for example, using laser-driven sources – on producing higher currents of high-charge-state beams.

4.4.2
RFQ Accelerators

The first practical device using the idea that the RF field could be used to produce both an accelerating field and the necessary transverse focusing field, avoiding the need for magnetic elements, was invented by Teplyakov [18] and tested in the USSR by 1968. Radiofrequency quadrupole (RFQ) accelerators have replaced high-voltage DC injectors and moved injection energy into the drift-tube-type accelerating structure up to a few megaelectronvolts. The RFQ can provide strong AG focusing at low energy [19]. With it, the space–charge limit (Section 4.4.5) is sufficiently increased so that protons or other ions can be injected at energies from around 20–100 keV depending on the intensity and accelerated to around 2 MeV for injection into a drift-tube-type LINAC. Very high percentages (95–97%) of the input beam can be captured and accelerated with minimal growth in emittance.

Using this principle, the first device built outside the USSR was tested in 1980 [20, 21, 17]. Figure 4.13 shows the sketch of a four-vane RFQ; when excited on the quadrupolar RF mode, the field on the axis is mainly quadrupolar in configuration but also contains, owing to the modulation of the pole tips, a longitudinal component that is accelerating. The fields in the RFQ are continuous, rather than intermittent as in the drift tube approach. The period of the modulation is synchronized with the ion velocity. The four-vane structure is currently used at or above about 200 MHz (Figure 4.13). A lower frequency variant is the four-rod RFQ structure (Figure 4.14). There the tip profile is materialized by minivanes properly supported to achieve the chosen resonant frequency and the whole system is put into a cavity; such a structure can be operated at a lower frequency without becoming too large in size [22].

The efficiency of RFQ structures for focusing at low velocity comes from

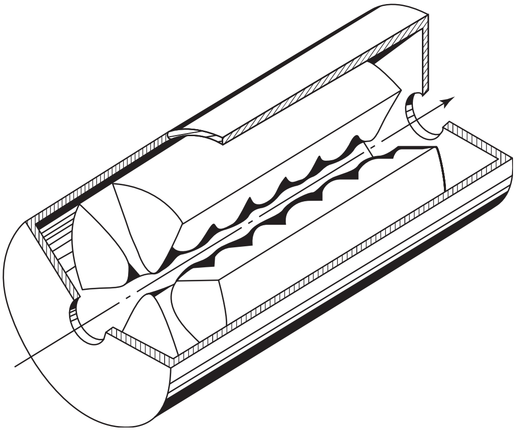

Figure 4.13 RFQ four-vane structure. The four vanes, in a quadrupolar arrangement, have their pole tips modulated; this produces a longitudinal RF field that allows the bunching and the acceleration of the beam.

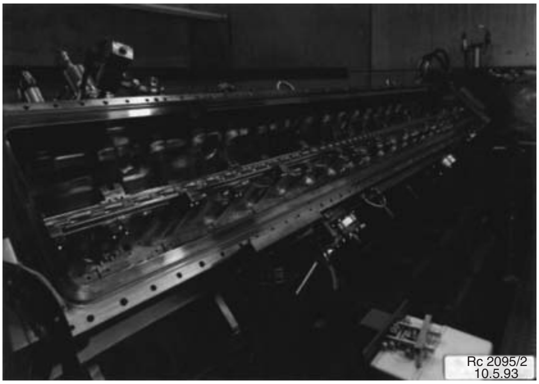

Figure 4.14 A four-rod-type RFQ built for GSI by Prof Dr A. Schempp, Institute for Applied Physics, Goethe University, Frankfurt.

the use of RF electric fields higher in strength than could be achieved with electrostatic fields and having a much larger focusing action than magnetic fields. RFQs are used to focus the beam emerging from the ion source to form bunches and accelerate them to an energy where adequate magnetic quadrupole focusing in the drift-tube-type structure can easily be provided (usually around 2 MeV). Because the RFQ has many cells in a reasonable length, the bunching and the acceleration characteristics can be slowly and carefully tailored. This nearly adiabatic process preserves the beam quality remarkably well and is a fundamental advantage of the RFQ with far-reaching consequences in modern high brightness proton and ion LINAC design.

4.4.3
Drift-Tube-Type LINACs

The RFQ loses efficiency as its cells become longer, necessitating a change in technique to accelerate to higher energies. The drift-tube principle is used.

The Alvarez type, with a focusing magnet located in every drift tube, provides the strongest and most uniform transverse focusing. The RF frequency of most of the large drift tube proton LINACs built in the 1970s is around 200 MHz, at which high-power triodes or tetrodes are available for long-pulse operation. New machines tend to use higher frequencies to achieve higher brightness. In a 200 MHz Alvarez structure, the length of a few tens of drift tubes varies from about 5 to 15 cm as the particles accelerate from 0.5 to 5 MeV.

Magnetic quadrupoles are enclosed in each drift tube, which, supported by stems, must be carefully aligned. Drift tubes are installed in long cylindrical cavities of typically a few meters; the cavity length is sized to match the power delivery capability of the amplifier units attached either singly or through multiple-feed ports.

The Alvarez structure is efficient for medium velocities: $0.05 < \beta < 0.3$. The large diameter drift tubes necessary to house the quadrupole focusing magnets require more RF power. The structure can be modified to save RF power. Groups of a few cells having small diameter drift tubes without magnets provide acceleration without focusing, interspersed with a group with larger drift tubes containing magnets. Thus this type of LINAC can be termed a *separated-function* LINAC; compared to the drift tube LINAC (DTL), where transverse focusing is provided in every cell, the separated-function LINAC uses a series of cells designed for high acceleration efficiency and low RF power consumption, with transverse focusing inserted periodically as needed. There are now many variations of this idea, including superconducting ones. The overall transverse focusing is weaker but often adequate, and the savings in RF power can be substantial.

4.4.4
LINACs for Higher Velocity Ions

Above ~100 MeV, for protons, the efficiency of these LINAC structures again becomes too low, and another structure is needed. The Los Alamos Meson Physics Facility (LAMPF) LINAC was the first high-intensity LINAC to push beyond ~100 MeV. The "separated-function" principle is used, and the side-coupled cavity structure [23] (Figure 4.15) was developed using the standing wave mode at an RF frequency that is an integer multiple of

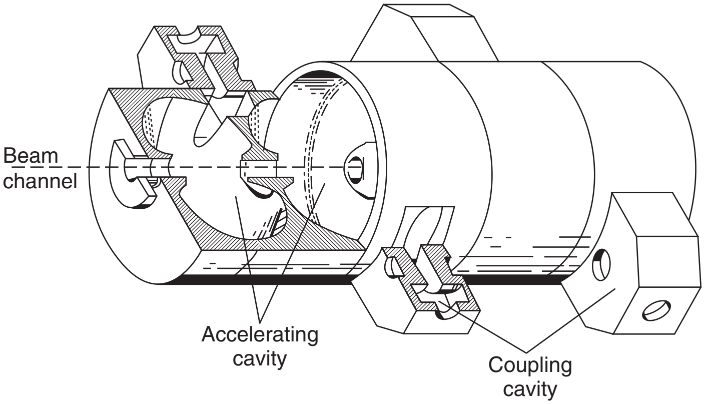

Figure 4.15 Side-coupled cavity structure. The on-axis accelerating cavities are designed with noses on the beam holes that prevent direct coupling and concentrate the accelerating field over a shorter length and time. The coupling between cavities is then made through lateral holes *H* via the side-coupling cavities.

2–4 times the frequency of the Alvarez section. In modern practice, this ''room-temperature'' structure type is used as a bridge between the structures of Section 4.4.3 to an energy at which a superconducting structure type becomes practical.

In Figure 4.15, one sees that the accelerating cells have noses on the beam holes in this structure. Such noses behave as same as drift tubes and concentrate the accelerating field over a shorter length, reducing the loss in acceleration resulting from the transit time through the gap. They also reduce the coupling between cells. The coupling necessary for the wave propagation is obtained through lateral cavities alternatively on one side and the other. There are two cells and two side-coupled cavities per wavelength. In a wavelength, there are two peaks and two nodes. The mode excited is such that the accelerating cells are at the peaks, that is, the field is high in them; the side cavities are on the nodes; so the RF losses are negligible in them. In this mode, the group velocity for energy propagation in the structure is high, and the cell fields are tightly synchronized – these features are very valuable in practice because the effect of errors is thereby reduced. The shaping of the cells optimizes the shunt impedance in room-temperature structures. In superconducting LINACs, the RF losses are low; so the cavity shape can be optimized on other criteria, and such structures can be used from below 100 MeV up to the velocity of light. The cell sequence is broken when it is necessary to place magnetic quadrupole lenses to provide focusing; special bridge coupler cells may couple RF power around the quadrupole.

Automatic control is applied to maintain the cavity field amplitude and phase constant by supplying extra drive when the beam is on. Because the required energy gain is fixed by the geometry, if beam loading were allowed to reduce the field, the trapping area in longitudinal phase space (called *acceptance*) would be smaller.

This side-coupled structure, with $\beta = 1$ for electrons, is presently used in most dental X-ray machines.

4.4.5
Beam Dynamics in Nonrelativistic LINACs

Almost all modern applications have to achieve an intensity as high as possible at the desired energy, along with a very good beam quality in terms of the beam confinement, aiming, or focusing. One figure of merit used is the beam brightness, defined as the *beam power* (or current, when the energy is fixed) divided by the phase space appropriate to the problem at hand. Phase space for the beam as a whole is six-dimensional, describing the physical size of the beam in x, y, and z, and the change in size with time or distance x', y', and z'; the area projected on one plane is called *emittance*.

Another figure of merit is to minimize the beam loss along the LINAC, and to concentrate any unavoidable beam loss to the lowest possible energies, as lost beam causes induced radioactivity in proportion to the lost particle energy.

The achievable beam intensity and quality are nonlinearly related, and raising the intensity tends to spoil the quality. Particle losses along the LINAC are to be strictly limited, as they would cause radioactivity buildup and complicate maintenance. The rule is that maintenance should be possible without remote manipulators. Achieving high brightness accelerated particle beams involves work on both the numerator and the denominator of the brightness equation – beam intensity divided by beam emittance.

The numerator can be raised by brute force, but the large power requirements and engineering problems that result can be formidable, and better system efficiency becomes a key issue. A given accelerator channel can accelerate only a certain amount of current, depending on the imposed constraints. If more current is desired, the constraints or basic parameters might be changed, or several accelerator channels or modules could be used. The latter approach would multiply the system power requirement directly; therefore, one would prefer to find a more efficient way to raise the current, if possible.

Decreasing the denominator of the brightness factor involves building low emittance (but also intense and, therefore, bright) particle sources and then preserving the source brightness through all the subsequent steps of beam transport and acceleration to the final energy and the beam target. Owing to basic beam dynamics, the minimum phase space volume anywhere in an accelerator is no smaller than that at the particle source.

The main problem at nonrelativistic velocities is space–charge effects. The beam particles must be charged to be accelerated, and all have the same sign; so they tend to repel each other. When the beam intensity is increased, the charge in the bunches becomes such that defocusing occurs. In a bunched beam of ellipsoidal shape, defocusing takes place in both transverse directions against the external focusing, and also in the longitudinal directions acting against the phase stability mechanism. Such effects mainly take place at low velocity when the particles are very sensitive to electric fields; the situation improves at higher velocity.

Simple but very effective design equations for nonrelativistic transported or accelerated beams can be obtained by writing general root-mean-square (RMS) envelope equations for the beam as it is constrained in the accelerator channel. The external channel-focusing forces are offset by the internal space–charge forces in the beam. For a beam that is matched to the machine and does not oscillate in shape, one finds then the following relations:

$$\varepsilon_\mathrm{t} = \frac{\sigma^\mathrm{t} a^2}{N\beta\lambda} \quad \text{and} \quad \varepsilon_\mathrm{l} = \frac{\sigma^\mathrm{l} b^2}{N\beta\lambda} \tag{4.1}$$

in terms of the transverse and longitudinal planes, where ε_t and ε_l are the transverse and longitudinal RMS emittance, a is the average transverse RMS beam radius, b is the physical RMS bunch length, and σ^t and σ^l are the phase advances of the oscillatory motion in phase space over the focusing period $N\beta\lambda$. Each of the σ terms can be expanded into two terms, one containing the machine parameters and the other containing the beam current and also the beam sizes a and b, explicitly. It is necessary to solve both parts of Eq. (4.1) simultaneously; the result is a beam that is ‘‘matched’’ to the RMS shape factors of the machine acceptance phase space. We know that RMS matching minimizes the growth of the effective emittances. It is important to note that these equations apply locally; thus, they should be applied not only as initial conditions at the entrance point of various sections of a machine but also with care to ensure that they remain true at every point.

In space-charge-dominated beam/accelerator systems, effective emittance growth could occur if the average energy in each of the coupled degrees of freedom is unequal [24, 25]. Requiring equality produces another set of equations:

$$\frac{\varepsilon_l \sigma^l}{\varepsilon_t \sigma^t} = 1, \text{ which also requires}$$
$$\frac{\varepsilon_l}{\varepsilon_t} = \frac{\sigma^t}{\sigma^l} = b/a \tag{4.2}$$

If Eqs. (4.1) and (4.2) is not satisfied, the effective emittance will shift between the degrees of freedom (with conservation in six dimensions). The physical mechanisms for the redistribution have been clarified as a charge density distribution occurring at an abrupt parameter change within about one-quarter of the plasma oscillation period of the system, and in a smooth channel, a slower kinetic exchange toward the equipartitioned state. Systems satisfying Eqs. (4.1) and (4.2) are both matched and ''equipartitioned.'' Again, the condition can be applied locally and may be applied at injection and all along the machine to ensure minimum emittance growth.

Another problem can be the formation of a diffuse halo of particles around the beam bunch. Halos are to be avoided, as these particles may strike the walls of the LINAC. A halo forms when single particles interact with resonant driving forces in the channel [26]. If Eq. (4.1) is not satisfied, the beam is ''mismatched'' and this is a prime source of a halo.

A primary design requirement is to maintain matching and to avoid resonances between the beam and the accelerator structure parameters that can cause emittance growth and particle loss. The equilibrium, equipartitioned beam has no free energy to drive a resonance; so, even if a resonance is traversed, the growth rate will be zero. Although the equilibrium condition is desirable, as with all LINAC parameters, it may be relaxed or abandoned in favor of a particular objective; in this case, if a resonance area is encountered in the LINAC, it should be crossed quickly to avoid excessive emittance growth.

From these equations, many of the tradeoffs and scalings in machine design can be explored to first order. For example, an important result is that if high brightness is the goal, the maximum current achievable for a fixed transverse emittance strongly favors higher frequency LINACs. As a practical example, the low-β section of a modern 100 mA average current proton LINAC would operate at a frequency around 400 MHz. Fortunately, there are also other factors that are advantageous at higher frequencies, such as smaller size, ability to withstand higher electric fields without breakdown, and reduction in the amount of energy that must be stored in a resonant structure. Engineering factors such as heat removal in high duty factor or c.w. applications, on the other hand, argue for lower frequency; so a balance must be struck. Optimization for low and minimum energy beam loss is much more difficult, still not expressible in compact form, but considerably aided using the equipartitioned equilibrium and resonance avoidance.

Thus, in the final design for an application, many factors constrain the brightness that can be achieved. However, in summary, a few key requirements and procedures have been postulated that have major effects on achieving lower beam loss and/or higher brightness. These include using higher frequencies for ion LINACs, maintaining external focusing fields as high as possible, maintaining local matching per Eq. (4.1), the use of beam equilibrium and resonance avoidance locally per the simultaneous solution of Eqs. (4.1) and (4.2), and the important techniques for initial acceleration from the source as embodied in the RFQ LINAC for ions. Further research is required to fully understand and exploit

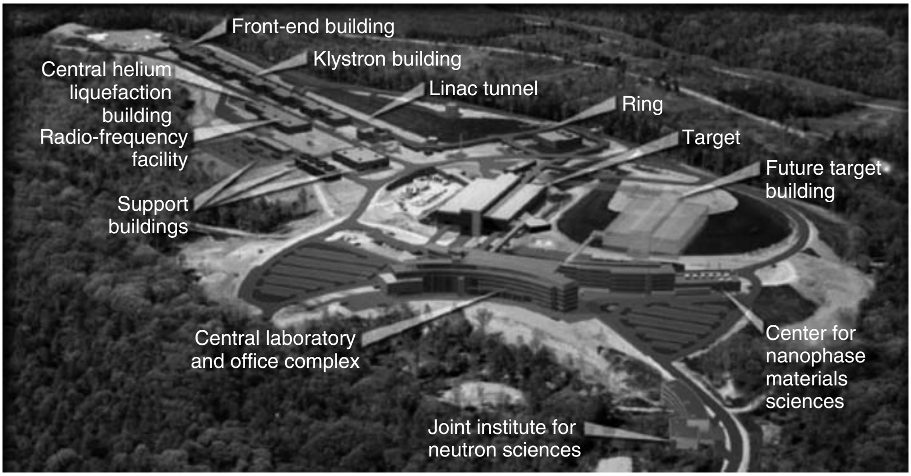
Front-end building
Klystron building
Central helium liquefaction building
Linac tunnel
Ring
Radio-frequency facility
Target
Future target building
Support buildings
Central laboratory and office complex
Center for nanophase materials sciences
Joint institute for neutron sciences

Figure 4.16 View of the ORNL SNS facility, indicating how the LINAC injects into the storage ring, from which the beam is extracted to the target.

these constraints, especially for controlling particles that tend to form a low-density ‘‘halo’’ around the beam and are likely to be scraped off, causing radioactivity problems for machine maintenance if their energy is above a few megaelectronvolts and to incorporate optimization algorithms.

4.4.6
Proton LINAC Applications

The LAMPF in New Mexico, producing an energy of 800 MeV with an average current of 1 mA of protons for medium-energy physics experiments, was completed in 1972; its energy is approximately 10 times and beam intensity 1000 times higher than any previous operational LINAC. Its length is close to a kilometer. It contains both 201.25 MHz Alvarez and 805 MHz side-coupled cavity structures. The SNS LINAC at Oak Ridge National Laboratory (ORNL), put into operation in 2005, has similar charge per bunch, and includes 402.5 MHz RFQ and DTL, and 805 MHz coupled cavity LINAC (CCL) and superconducting LINAC structures (SNSs, 2006) (Figure 4.16).

Ion LINACs are frequently used as injectors for storage rings or circular accelerators of higher energy. Frequently, H^- ions are accelerated for injection into a circular machine (e.g., SNS); a thin foil strips the two electrons, leaving the protons to circulate. An H^- LINAC just differs from a proton LINAC by the source and the sign of its extraction voltage.

Proton LINAC applications include drivers for intense neutron sources for physics and materials research such as the SNS, and the planned International Fusion Materials Irradiation Facility (IFMIF) for fusion-related materials development. Smaller proton and heavy-ion (mainly carbon) LINACs are being used for cancer therapy machines. Major new ideas were proposed at Los Alamos in 1990 [27] for closing of the fission power cycle by treatment of radioactive wastes, and for practical implementation of the thorium fuel cycle, using a high-intensity proton LINAC to drive a subcritical nuclear reactor. These proposals were extended to the international community [28], and an active program EUROTRANS [29] is in progress to construct an R&D facility. The initial accelerator would be able to accelerate up to ∼20 mA protons to 600–800 MeV; superconducting structures are ideal for a major part of the accelerator because a continuous beam is required. For such machines, the major design concerns are the minimization of beam losses at energies above a few megaelectronvolts, and strict prevention of sudden beam turn-off (discussed later).

More recently, advances in superconducting RF technology have been applied to the construction of proton LINACs. The main application is the SNS at the US ORNL [14]. The SNS is a second-generation pulsed neutron source, with a LINAC delivering 1 GeV protons at 1.4 MW average power to a liquid mercury target for neutron spallation with applications to the study of the structure of materials. The bulk of the LINAC is superconducting RF (0.812 GeV out of 1 GeV). In addition to greater efficiency and higher gradients for shorter lengths, the larger apertures allowable with SRF reduce beam loss, which is a critical concern because of activation. The superconducting LINAC utilizes two cavity types, for β of 0.61 and 0.81.

4.5
Heavy-Ion Linear Accelerators

4.5.1
Special Difficulties Relative to Ions

Under the term heavy ions, it is common practice to include all ionized atoms heavier than deuterons; accelerated ions usually range from lithium or carbon up to uranium.

Much of heavy-ion physics has been done in the past with the help of circular machines, usually cyclotrons, but there have also been several LINAC facilities, for instance, the heavy-ion linear accelerator (HILAC) at Berkeley – a big low-frequency Alvarez LINAC, later converted as an injector for the Lawrence Berkeley Laboratory Bevatron/Bevalac Facility (BEVALAC), a large heavy-ion synchrotron.

A center for heavy-ion physics was created at the Gesellschaft für Schwerionenforschung (GSI) laboratory in Darmstadt, Germany, and a heavy-ion LINAC complex called the universal linear accelerator (UNILAC) [30] was built. It is made of a Wideroe section, an Alvarez section, and an independent cavity section; more developments, such as special RFQs and H-type interdigital structure sections, have been added later, and the facility remains very active.

From this short description, it appears that a heavy-ion LINAC may be rather complex and differ appreciably from the proton machines.

A fundamental difference comes from the experimental requirements for a wide variety of ions with different masses and charge states. The charge-to-mass ratio (q/m) then differs from one ion to another and is always smaller than 1/2 (compared to the proton case) and often much smaller. In a given LINAC, where the geometry is fixed and only certain adjustments of the phase and amplitude of the accelerating fields are possible, only some combinations of q/m can be accelerated. For this reason, short, separately adjustable resonators are preferred to afford retuning for each separate ion beam.

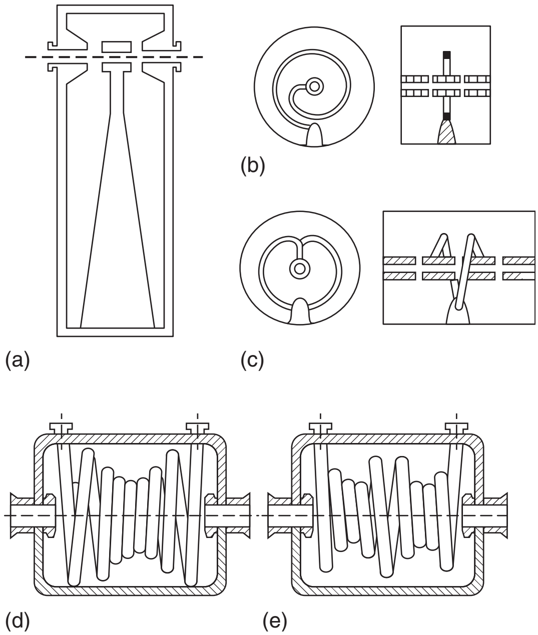

Figure 4.17 Superconducting cavities for heavy ions. (a) Quarter wave cavity: in this sort of coaxial resonator, the open end supports a drift tube and acceleration takes place across the gap. (b) Spiral ring resonator: to reduce the size, the resonant quarter wave line has been curved into a spiral. (c) Split ring resonator: to increase the energy gain, two spirals are put together; the gain should be doubled (in practice less), but the band width in velocity is reduced. (d) Half wave helix: the construction is easy, but there is a risk of mechanical vibrations and the field axis is not well defined. (e) Full wave helix: the rigidity is better, but the velocity bandwidth is smaller. Such cavities currently deliver >0.5 MV to the beam.

4.5.2
Accelerating Structures – Normal and Superconducting

Lower frequencies (20–100 MHz) offer more space for focusing; so structures such as the Wideroe structure (Section 4.2), or a variety of so-called H-type interdigital structures are used, some of which are represented in Figure 4.17. Independent short superconducting cavities are ideally suited, as they can be phased independently, affording complete flexibility in the velocity program. Fifty or one hundred cavities designed to have an optimum acceleration in several velocity steps (but with usable acceleration over a velocity range about a factor of 2 around the optimum) are arranged in succession. Their amplitude and phase are adjusted and controlled for the best possible acceleration of a chosen ion. The first machine of this type was made at Argonne near Chicago and is called *ATLAS* (Argonne tandem LINAC accelerator system). Its length has been progressively increased. Initially, as with the other machines of this type, it was used as a booster after an electrostatic tandem. Now, with proper low-β cavities ($\beta < 0.01$), beams from ECR sources at a relatively low voltage can be accepted directly [31].

4.5.3
Application

Heavy-ion accelerators have been mainly constructed for nuclear physics experiments; one of the largest facilities is the GSI in Darmstadt, Germany. More and more nonnuclear applications are appearing (solid-state physics, materials research, etc.), however, as also more and more industrial uses (implantation, surface treatments, ultrafiltration, etc.). An ambitious project that would reach the limits of the LINAC technology is a heavy-ion LINAC for inertial fusion. Fusion reaction could be induced in a small deuterium pellet if bombarded by a short and extremely intense bunch of energetic heavy ions (10 GeV, several kiloamperes with a 10 ns pulse length). It is clear that a special type of accelerator would be needed to reach that goal (Section 4.6).

4.6
Other Types of Linear Accelerators and New Methods of Acceleration

4.6.1
Induction LINAC

The induction LINAC is a candidate for short-pulse applications. Long induction LINACs are built up from a series of modules of the type shown in Figure 4.18. They behave as same as pulse transformers with the beam as one winding, and are accurately phased to produce acceleration (and possibly bunching) exactly at the time a bunch of particles crosses the gaps. The method is, in fact, the modern version of the original Ising proposal (Section 4.2). Induction LINACs are inherently pulsed devices of pulse length ∼50 ns. Longer pulses are limited by the energy that can be stored in the pulser cores. Repetition rates of up to a few kilohertz are being developed using special magnetic modulators. This device can accelerate extremely high intensities. The largest electron induction LINAC built is the Advanced Test Accelerator at Livermore, a 50 MeV machine that has accelerated up to 10 kA in 50 ns pulses [32]. The induction LINAC is also well suited for the heavy-ion fusion project [33] mentioned in the previous section (its length would be several kilometers).

4.6.2
Alternating-Phase-Focused LINACs

As noted in Section 4.4.2, the RF field can be used for both an accelerating and the transverse focusing field, avoiding the need for expensive magnetic elements. Except for the RFQ, it has been very difficult in general to design such LINACs, but recently a new practical method has been found [34].

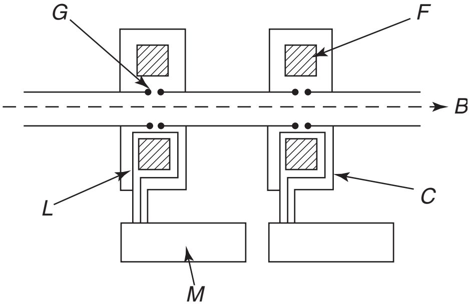

Figure 4.18 Induction LINAC modules. Along the beam path *B*, a succession of cylindrical cavities *C* are installed with an accelerating gap *G*. Inside these cavities, toroidal ferrite cores *F* are excited via a coupling loop *L* by the pulse-forming networks of the modulators *M*. The cavities behave as same as a transformer and the pulse voltage applied on the loop appears on the accelerating gap. Timing of the modulators must be accurately controlled so that the pulses are applied very correctly at the time the bunch crosses the gaps.

4.6.3
New Methods of Acceleration

Even in pulsed mode, microwave structures are limited to gradients of the order of 100 $\mathrm{MV\,m^{-1}}$, and as the need for higher energy beams and compact sources has grown, strong interest has developed in using the intense fields generated by lasers or plasma waves to accelerate beams. For example, plasma perturbations can yield field gradients of order 100 $\mathrm{GV\,m^{-1}}$. A device based on this technology could shrink the length of LINACs by three orders of magnitude! To generate these plasma waves, either a high-powered laser or a particle beam would be used to set up the plasma oscillation (laser wakefield or plasma wakefield concept). An alternative approach would use gratings, ministructures such as photonic bandgaps, or exotic materials (e.g., active or negative index of refraction) to couple a laser beam to a particle beam. Also, a particle beam passing through the alternating fields of an undulator can couple to a copropagating laser beam for acceleration in a device called an *inverse FEL*. Limitations of these ideas include structure damage, plasma instabilities, accelerating high currents, and beam quality maintenance. However, there have been significant experimental advances that hold the promise for future applications. See [35] for more details and references on these new methods of acceleration.

4.7
Engineering Aspects

The full spectrum of engineering disciplines is applied to accelerator technology; a close, multidisciplinary approach with the physical, mathematical, material, and computer sciences and technologies is essential. Deciding on trade-offs between the possibilities and constraints of the many aspects constitute a large part of the design, construction, and operation of accelerators. Broad engineering categories applied to most accelerator systems are described in the following sections.

4.7.1
Accelerator Structure Design

The structures in which beams are accelerated must efficiently provide precisely the correct fields and are mechanically complicated vacuum vessels that must be thermally stabilized, structurally rigid, extremely clean for sustaining high electromagnetic fields, precisely aligned, and maintainable. They must be structurally stiff to hold their shape under vacuum and component loading. RF accelerating structures are copper clad or copper plated, or of special material such as niobium if superconducting. Structures and beamlines must be constructed of materials that are hard-vacuum compatible because they must maintain a vacuum quality typically 10^{-6} to 10^{-9} Torr. Fabrication techniques used in accelerators are diverse. High-precision machining is done on numerically controlled milling machines and by other specialized fabrication techniques such as electron discharge machining (EDM). Specialized joining and forming techniques include brazing, electron beam and laser welding, explosive bonding of dissimilar metals, and explosive hydroforming. Very good mechanical tolerances and cooling systems are needed. Complex brazing is used in the latest vane-type RFQs (Figure 4.19) and side-coupled LINACs (Figure 4.20). The highest power RFQ built to date, the

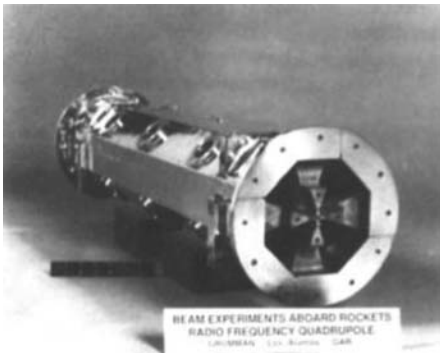

Figure 4.19 An RFQ manufactured in four segments.

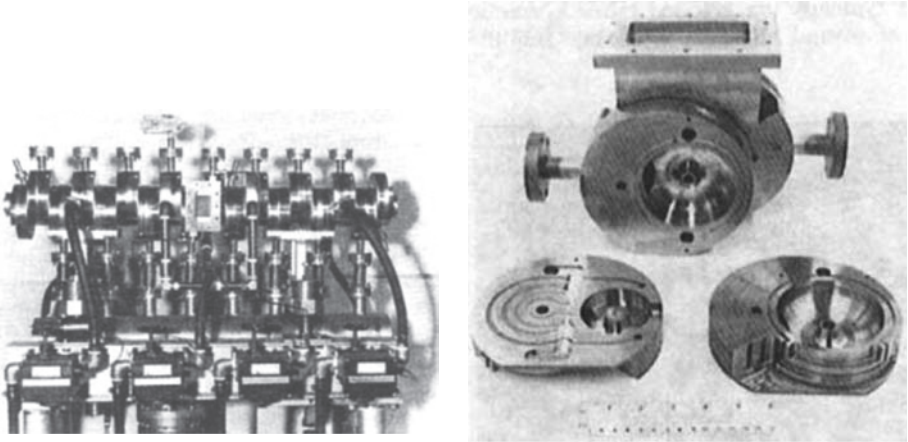

(a) (b)

Figure 4.20 (a) A side-coupled LINAC cell before brazing, and (b) a section of a 2400 MHz c.w. side-coupled LINAC.

Los Alamos Low-Energy Demonstration Accelerator (LEDA) c.w. 100 mA, 6.7 MeV proton LINAC (Figure 4.21), almost disappears in its support structure, which contains RF power, water-cooling, and vacuum connections.

Rapid advances in the performance of superconducting cavities have made RF superconductivity a key technology for accelerators. They are especially suitable for applications that require c.w. or long-pulse accelerating fields of many megavolts per meter. The RF surface resistance of an SRF cavity is many orders of magnitude less than copper, and even allowing for the inefficiencies of heat removal at cryogenic temperatures, a net gain in efficiency of 100 can be easily obtained. (Since superconducting Cooper pairs exhibit inertial effects in AC fields, they do not completely cancel the electric fields in a superconducting material. Nonsuperconducting electrons from ''ionized'' cooper pairs or impurities above absolute zero can respond to these residual fields and dissipate small amounts of energy.) Gradients of up to 20 MV m^{-1} and Qs approaching 10^{10} are practical today, and further improvement can be expected. SRF structures are

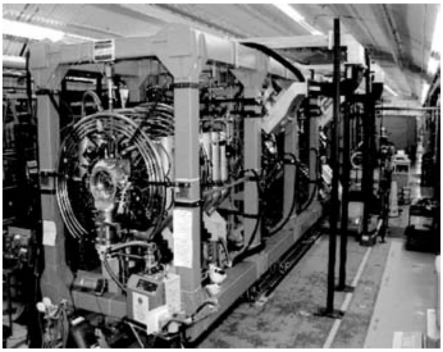

Figure 4.21 Low-Energy Demonstration Accelerator (LEDA) RFQ at Los Alamos.

shaped to control resonant electron effects (multipactoring). Control of contamination by macroparticles by assembly in clean rooms and careful surface preparation has eased the limits of field emission and thermal breakdown.

4.7.2
Magnetics

Complex magnet systems provide magnetic steering and focusing elements that control the position, size, and transport characteristics of particle beams. Modern transport systems use and control magnetic properties as high as fifth order in the optics. Many of the magnets used in accelerator systems are physically short because of packaging constraints or must deviate from ideal geometry because of the field coils. The fields must be measured, often to an accuracy of better than 1 part in 10^4, using magnetic mappers supported by sophisticated instrumentation and computers. Radiation-hardened magnets are also required.

Permanent magnets are quite applicable to high-frequency DTLs and considerably simplify the drift-tube design because they require no power or cooling. Small, accurate quadrupoles (with 1.2 cm bores operating at 18.33 kg cm^{-1} gradients) have been built and tested. They too can be built to a size appropriate for a beam-transport line. New design techniques can be used for adjustable-field multipole magnets and for magnets in which permanent magnets and normal-magnet steels are mixed. Recent experimental results on rare-earth cobalt materials show, however, that serious demagnetization occurs in some designs at 10^9 rad dose levels. Thus, although permanent-magnet technology grows rapidly in its own range of application, use in high radiation environments is precluded at present.

4.7.3
Radiofrequency Engineering

The complex electromagnetic requirements of the accelerator structures must be met, and high-power RF driver systems provided for them. Radiofrequency power systems typically use gridded tubes to frequencies of around 200 MHz, with solid-state amplifiers, klystrons, and inductive output tubes (IOTs) at higher frequencies. The power requirements can be very large; for example, a 40 MeV, 125 mA IFMIF RF LINAC requires 5 MW just for the beam. LINACs are divided into sections sized to the power sources; this ranges from single cells to tanks containing many cells. To maintain synchronism, all the sections must be tightly controlled to within typically 1% in amplitude and synchronized to ±1° in phase, by the automatic control of a reference oscillator. Bunch compression scenarios can demand more stringent RF control by up to a factor of 100. Development of efficient RF power sources is an urgent requirement for future applications.

4.7.4
Availability and Maintainability

LINACs for physics research machines have almost always been cutting-edge technology, always breaking new ground, and on a very tight budget. Usually a loss of beam for some period, from missing pulses to some hours, could be tolerated without any problem by the experiments. Thus, accelerator facilities to date have considered availability only as ''best-practice'' and not as a top-level specification. The users of research facilities such as SNS may come for short periods of a few days; so their productivity must be ensured; SNS has availability goals built into its development schedule over the next decade. For IFMIF, the question of beam trips resulted in the choice of building at least two parallel LINACs, in modules of 125 mA each, greatly increasing the probability that at least one would always be operating. Also, the material samples would not be damaged because the thermal cooling could be adjusted quickly. For transmutation, however, a beam trip would induce thermal stress on the subcritical core, and the number of unwanted ''beam trips'' should not exceed a few per year, a requirement that is several orders of magnitude above usual performance. For these and other future facilities, availability becomes a top-level specification and a major design challenge, as accelerators contain thousands of components, and often the failure of one will require the beam to be shut off.

4.7.5
Computers

Accelerators, in general, have long provided a major impetus for the development of advanced computer systems. Design teams and facilities use the full panorama of computers, from supercomputers to simulate particle motion and electromagnetic fields in the accelerator and magnet systems and on-line real-time computer control systems with distributed processing, to personal computer tools. Modern, fully integrated computer control using instrumentation and automatic control bring accelerator commissioning and operation under strict and very effective management, as witnessed by the recent commissioning of the SNS.

4.7.6
Wide-Bandwidth Electronics and Instrumentation

The pulsed nature of particle beams requires wide-bandwidth electronics, some at higher power, and various diagnostic instrumentation to monitor machine- and beam-operating parameters and control them. Most of these tasks can now be handled digitally.

An example of high-power, high-bandwidth electronics is the pulsed beam chopper shown in Figure 4.22. This device deflects part or all of a 750 keV H^- beam in the LAMPF injector transport line. A pulse rise time of <5 ns and pulse lengths from 15 ns to 750 μs are required at repetition rates up to 4.7 MHz. Chopping is accomplished by launching along this structure a pulse that propagates at the ion velocity of the beam, $1.2\,\mathrm{cm\,ns^{-1}}$. The very

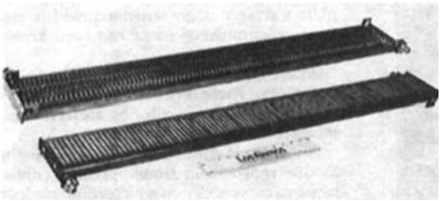

Figure 4.22 Deflector plate assemblies for a beam chopper.

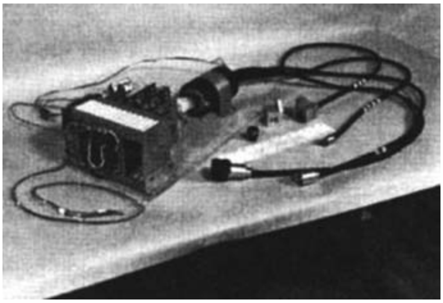

Figure 4.23 Four-armed fiber optics beam-sensing system.

fast high-power electronics uses planar triodes and vertical metal-oxide semiconductor fields-effect transistors (VMOS FETS) in a parallel cascade arrangement to deliver both positive and negative 600 V pulses to the opposing coaxial plate structures.

Diagnostic instruments must often measure beam parameters without intercepting the beam. This is essential with high-intensity beams, which would destroy intercepting materials. Beam parameters of interest include current, energy, density distribution, emittance, and centroid position. Beam density distribution is sometimes measured by optical sensors that view through optical ports the light generated by beam scattering from residual background gas. The optical system uses four-armed fiber-optic image scopes (Figure 4.23) to transmit four-beam profiles from a given beam tube/optical-port array to a microchannel light intensifier and then onto a pair of linear photodiode arrays, from which the light intensity is digitized. The use of a number of these ports distributed along the beam tube allows beam size and trajectory information to be reconstructed into beam emittance. The emittance reconstruction codes are executed in a microcomputer along with beam-transport dynamics codes and several display routines. Longitudinal beam phasing and shape information may be sensed using a variety of high bandwidth capacitive pickups, such as shown in Figure 4.24.

4.7.7
Automatic Controls

Automatic controls are used widely in accelerator technology. Examples include the following:

- control of parameters in complex subsystems such as the injector;
- cooling-system temperature control;
- RF amplitude and phase control;
- beam trajectory control.

Control of water temperature is fundamental to high-power accelerator tasks because operating temperature affords a vernier adjustment to the structure's RF.

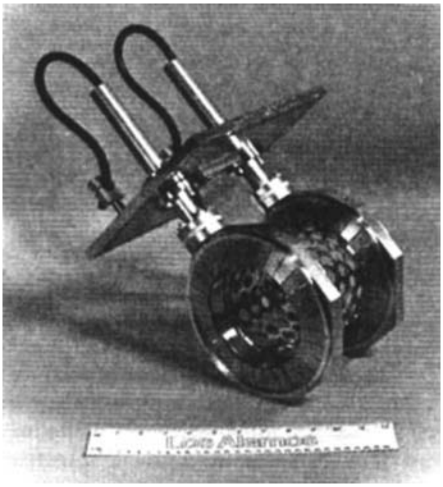

Figure 4.24 Capacitive pickup probes for time-of-flight measurement.

The proper RF field phase and amplitude are crucial to operation and must be controlled to within ±1° and ±1%, typically, under all possible transient and steady-state conditions. In some cases, multiple amplifiers drive a single structure; field control of these complex systems has been accomplished.

Many types of errors can affect the beam trajectory. These must be found and corrected or compensated for by correction elements located along the beamline.

A variety of controllers are used, covering a wide range from simple process control loops, to complex loops with feed forward, domain switching depending on system state, and adaptive and nonlinear systems; modern implementation is digital with computer processing. Work is currently being done using expert system and neural network techniques.

4.7.8
Civil Engineering

The accelerating structures and magnets defining the beam trajectory must be aligned relative to each other to exquisite tolerances (a few mils; down to micrometers for linear colliders) and also absolutely over the whole facility, again to mil-level tolerances. The alignment must be maintained over the lifetime of the facility. Civil engineering procedures must therefore take many factors into account, such as thermal cycling on stresses in the buildings, placement and rearrangement of heavy shielding, vibrations from utilities, and replacement procedures when accelerator parts are maintained. The surveying techniques incorporate the latest laser- and computer-supported techniques.

4.8
Concluding Remarks

Details on current developments can be found in the proceedings of the biannual US Particle Accelerator Conferences (PACs) published for many years as part of the IEEE Transactions on Nuclear Science, and recently as IEEE Conference Proceedings. Also biannually, LINAC Conferences have been held with proceedings published by a host laboratory. A third biannual conference series is "Application of Accelerators in Research and Industry." A

European Particle Accelerator Conference (EPAC) series started in 1988, and is now a triannual feature, rotating with the Asian Particle Accelerator Conference (APAC). Many of these proceedings are available at [36]. Russia also has a long-standing biannual particle accelerator conference series. These conference proceedings constitute the bulk of the accelerator technology literature, along with many laboratory internal reports and workshop proceedings (a number of these have been published as AIP Conference Proceedings, e.g., 156, 177, 184, 193, and 480). There is an active USPAC particle accelerator school and the CERN Accelerator School series. Since the previous (2007) edition, when conferences were almost the only source of accelerator literature, the American Physics Society has established its Division of Physics of Beams, and the Physical Review has a special topic component on Accelerators and Beams – PRSTAB.

Glossary

Acceptance: The maximum beam emittance that can be accepted without loss, through a transport-focusing channel or an accelerating system, determined by machine parameters, independent of the actual beam.

AG focusing (alternating-gradient focusing): A beam-focusing system made of a succession of focusing and defocusing lenses, normally of quadrupole type. Such a system provides net focusing.

Alvarez structure: An accelerating structure for protons, of the drift-tube type, invented by Luis Alvarez.

BBU (beam breakup): Intensity-dependent transverse phenomena in relativistic linear accelerators that limit the current in intense machines (Section 4.3.3).

Beam loading: Power absorption in RF accelerating cavities by the accelerated beam compensated by the application of RF power and control of the accelerator fields.

Bunch: The group of particles accelerated per RF period.

Disk-loaded structure: See Iris-loaded structure.

Drift tube: A tube connected to an RF voltage, through which the beam can drift. When a beam bunch is inside a drift tube, it drifts without being submitted to any RF field.

DTL (drift-tube LINAC): LINAC structure based on the use of drift tubes, for example, the Alvarez type.

Emittance: The positions of individual particles in an accelerated bunch can be expressed by three physical coordinates and their conjugate momenta. The areas occupied by the projections in the three-phase planes (see **Phase space**) are called the *beam emittances*; there are two transverse emittances (often equal in LINACs and expressed in terms of divergence angle rather than in terms of transverse momentum), and a longitudinal emittance (often expressed in terms of energy spread and time duration).

Equipartition: A balanced division of field energy between the three emittances of the beam that guarantees the best stability; the term is used similarly in the case of thermal equilibrium in a gas.

Focusing: The action of concentrating particles to form a beam, similar to focusing of light in optics.

GSI: Gesellschaft für Schwerionenforschung, Darmstadt, Germany.

Gap, accelerating gap: Interval between drift tubes where a strong RF field is concentrated, accelerating the particles of a bunch at the time they cross the gap.

IFMIF: International Fusion Materials Irradiation Facility, a complex of two 125 mA, 40 MeV, c.w. deuteron LINACs providing beams to a molten Li target, generating neutrons with a fusionlike spectrum for development of neutron-resistant materials for fusion reactors.

Iris-loaded structure (or disk-loaded): An accelerating structure for relativistic particles made of a circular waveguide loaded by irises to slow down the phase velocity of the RF wave, making it synchronous with the beam.

LAMPF: Los Alamos Meson Physics Facility, commissioned 1972, with 800 MeV, 1 mA average current proton LINAC, later known as LANSCE, Los Alamos Neutron Science Center.

LINAC: Linear accelerator.

Matching: The machine parameters define a preferred region in phase space (acceptance) for transport or acceleration. If the beam particle phase space distribution has a congruent orientation, called *matched*, beam passage will be smooth, without oscillations. A *mismatched* beam will oscillate in position or shape, which may cause loss of quality or actual particle loss.

Phase space: The six-dimensional space of physical coordinates and conjugate momenta.

Photocathode: A photocathode is a negatively charged electrode, often with a coating to improve emission, that generates electrons through photoemission when struck by a laser beam.

Quadrupole lens: Lenses, usually magnetic, with four poles around the beam: two of N polarity diametrically, and two S in between. Such a lens focuses in one transverse direction and defocuses in the other.

RFQ (radiofrequency quadrupole): An accelerating structure for low-velocity particles where the electrodes produce an RF field that is the sum of a transverse, quadrupolar, AG focusing field and an accelerating field.

SCL (side-coupled LINAC): A normal-conducting, standing wave LINAC structure used for accelerating electrons and protons when the RF pulse length is longer than a few microseconds, or for c.w. operation. The coupling between successive cells is arranged to take place through side cavities alternately on one side and then the other. The accelerating and side cavities resonate at the same frequency, resulting in a $\pi/2$ mode operation having maximum group velocity and field stability.

SNS: Spallation Neutron Source, Oak Ridge National Laboratory, USA, commissioned in 2006 [37].

Space–charge effects: Nonrelativistic particles in a bunch, all charged with the same sign, repel each other, tending to force the dimensions to grow, and requiring more focusing. This also leads to possible instabilities and couplings, and to the formation of a halo of particles around the bunches (Section 4.4.5).

Standing wave cavity: A hollow structure bounded by a conductive wall that supports a resonant electromagnetic mode when excited by an external RF source. The mode pattern is typically designed to accelerate a charged particle passing along the axis, although deflecting modes are also used in some accelerator applications.

Unit of energy: The energy of particles is expressed in electronvolts (eV), energy given to a unit charge (charge of the electron) through a potential difference of 1 V. Higher energies are expressed in kiloelectronvolts, megaelectronvolts, gigaelectronvolts, and teraelectronvolts.

Wakefield: The electromagnetic field left by a charged particle when it passes a discontinuity in its environment. For example, such discontinuities may be resonators or gaps, bellows, or steps in a conductive wall boundary. Resistive wall boundaries can also generate wakefields, which describe the loss of beam energy as image currents are driven in the wall.

References

1. Ising, G. (1924) Prinzip Einer Methode Zur Herstellung Von Kanalstrahlen Hoher Voltzahl. *Ark. Mat. Astron. Fyzik*, **18**, 1.
2. Wideroe, R. (1928) Über ein neues Prinzip zur Herstellung hoher Spannungen. *Arch. Electotech.*, **21**, 387.
3. Sloan, D.H. and Lawrence, E.O. (1931) *Phys. Rev.*, **38**, 2021–32.
4. Ginzton, E.L. *et al.* (1948) A linear electron accelerator. *Rev. Sci. Instrum.*, **19**, 89.
5. Slater, J.C. (1948) The design of linear accelerators. *Rev. Mod. Phys.*, **20**, 473.
6. Leemann, C., Douglas, D., and Krafft, G. (2001) The continuous electron beam accelerator facility. *Annu. Rev. Nucl. Part. Sci.*, **51**, 413.
7. Barish, B. (2006) The global design effort for an international linear collider. Proceedings of the 2006 European Particle Accelerator Conference, Edinburgh, Scotland.
8. Free Electron Laser (2007) The World Wide Web Virtual Library: Free Electron Laser Research and Applications, *sbfel3.ucsb.edu/, www/vl fel.html* (accessed 28 February 2013).
9. Merminga, L., Douglas, D., and Krafft, G. (2003) High current energy recovery linacs. *Annu. Rev. Nucl. Part. Sci.*, **53**, 387.
10. Alvarez, L.W. *et al.* (1955) Berkeley proton linear accelerator. *Rev. Sci. Instrum.*, **26**, 111–133210–219.
11. Courant, E.D., Livingston, M.S., and Snyder, H.S. (1952) The strong-focusing synchrotron – a new high energy accelerator. *Phys. Rev.*, **88**, 1190.
12. Blewett, J.P. (1952) Radial focussing in the linear accelerator. *Phys. Rev.*, **88**, 1197–1199.
13. Hereward, H.G. *et al.* (1956) Proceedings of the Conference on High Energy Accelerators, CERN, Genevap, p. 167.
14. Holtkamp, N. (2006) Status of the SNS linac: an overview. Proceedings of the 2006 European Particle Accelerator Conference, Edinburgh, Scotland.
15. Sinclair, C.K. *et al.* (2007) Development of a high average current polarized electron source with long cathode operational lifetime. *Phys. Rev. ST Accel. Beams*, **10** (2), 023501.
16. Farkas, Z.D. *et al.* (1975) Recent progress on SLED, the SLAC energy doubler. *IEEE Trans. Nucl. Sci.*, **NS-22** (3), 1299–1302.
17. Ahearne, J.T. (1980) RFQ is Alive and Well . . . Atom (Los Alamos Scientific Laboratory (LASL)) (Jul./Aug.), Vol. 17, No. (4).
18. Teplyakov, V.A. (2006) The first CW accelerator in USSR and a birth of accelerating field focusing. EPAC2006 Accelerator Prize, Edinburgh, Scotland, July 26–30, 2006.
19. Kapchinskij, I.M. (1970) *IEEE Trans. Nucl. Sci.*, **26**, 3462.
20. Stokes, R.H., Crandall, K.R., Stovall, J.E., and Swenson, D.A. (1979) RF quadruple beam dynamics. *IEEE Trans. Nucl. Sci.*, **NS-26**, 3462.
21. Potter, J.M., Williams, S.W., Humphry, F.J., and Rodenz, G.W. (1979) Radio frequency quadrupole accelerating structure research at Los Alamos. *IEEE Trans. Nucl. Sci.*, **NS-26**, 3745.
22. Schempp, A. (2002) Developer of more than 40 different RFQs. New development in high-duty cycle high-current RFQs. Proceedings of LINAC2002, Gyeongju, Korea, p. 539.
23. Knapp, E.A. *et al.* (1968) *Rev. Sci. Instrum.*, **39**, 979–91.
24. Jameson, R.A. (1981) Beam intensity limitations in linear accelerators. *IEEE Trans. Nucl. Sci.*, **28** (3), 2408.
25. Jameson, R.A. (2007) Equipartitioning in linear accelerators. Proceedings of the 1981 Linear Accelerator Conference, Report LA-9234-C, Los Alamos National Laboratory, February 1982, ORNL/TM-2007/001, p. 125.

26. Jameson, R.A. (1994) *AIP Proceedings of the 1994 Joint US-CERN-Japan International School on Frontiers of Accelerator Technology*, World Scientific, ISBN: 981-02-2537-7, pp. 530–560.

27. Bowman, C.D. *et al.* (1992) Nuclear energy generation and waste transmutation using an accelerator-driven intense thermal neutron source. *NIM Phys. Res.*, **AA320**, 336–367.

28. Jameson, R.A. Compiler, (1991) Specialist Meeting on Accelerator Driven Transmutation Technology for Radwaste and other Applications, June, 24–28, 1991, Saltsjöbaden, Stockholm, Sweden, spons. by LANL and Swedish National Board for Spent Nuclear Fuel (SKN), LA-12005-c, SKN Rpt. No. 54.

29. Knebel, J.U. (2006) EUROTRANS: European Research Programme for the Transmutation of High-Level NuclearWaste in an Accelerator Driven System. FI6 W-CT-2005-516520: Integrated Project on European Transmutation (EUROTRANS), Sixth Framework Programme.

30. Bohne, D. (1976) The UNILAC, development and present status. Proceedings of the Linear Accelerator Conference, Chalk River, Vol. AECL-5677, p. 2.

31. Bollinger, L.M. 1986 Linear Accelerator Conference, Stanford, California. CEBAF Report 89–001, p. 237.

32. Briggs, R.J. (1986) Linear induction accelerators. Proceedings of the 1986 Linear Accelerator Conference, SLAC-303, CONF-860629, UC-28, Stanford, California.

33. Fessenden, T. 1982 Linear Accelerator Conference, GSI 84-11, Darmstadt, Germany, p. 485.

34. Jameson, R.A. (2012) *Practical Design of Alternating-Phase-Focused (APF) Linacs* submitted for publication to NIM.

35. Blumenfeld, I. *et al.* (2007) Energy doubling of 42 GeV electrons in a metre-scale plasma wakefield accelerator. *Nature*, **445**, 741–44.

36. JACoW (2006) Joint Accelerator Conference Website, *www.JACoW.org* (accessed 28 February 2013)

37. Neutron Sciences (2006) The SNS Facility–Neutron Sciences, *neutrons.ornl.gov* and *www.sns.gov* (accessed 28 February 2013).

Further Readings

Accelerators and Beams (2012) *Tools of Discovery and Innovation*, Division of Physics of Beams of the American Physical Society, November 2012, *www.aps.org/units/dpb* (accessed 28 February 2013).

Bollinger, L.M. *et al.* (1984) Proceedings Linear Accelerator Conference, Darmstadt, Germany, Report GSI 84–11, p. 217.

Grunder, H. (1988) The continuous electron beam accelerator facility. Proceedings 1988 Linear Accelerator Conference, Williamsburg, Virginia, CEBAF Report 89–001, p. 3.

Hanna, S. (2012) *RF Linear Accelerators for Medical and Industrial Applications*, Artech House, Boston, MA March 31, 2012.

Lapostolle, P.M. *et al.* 1963 International Conference on High Energy Accelerators, Dubna, Atomizdat, Moscow, p. 513 (U.S. AEC Translation pp. 672–682).

Padamsee, H., Knobloch, J., and Hays, T. (1998) *RF Superconductivity for Accelerators*, John Wiley & Sons, Inc., New York.

Stokes, R.H. and Wangler, T.P. (1988) *Radiofrequency Quadrupole Accelerators and their Applications*, Annual Review of Nuclear and Particle Science, Vol. 38, Annual Reviews, Inc., Palo Alto, CA.

Vladimirskij, V. (1956) *Prib. Tekh. Eksp.*, **3**, 35.

Watson, J.M. (1985) The Los Alamos free-electron laser. *IEEE Trans. Nucl. Scin*, **NS-32** (5), 3363 (Proceedings 1985 Particle Accelerator Conference).

Wiedemann, H. (2007) *Particle Accelerator Physics*, Springer, London.

Wilson, P.B. (1989) Introduction to wakefields and wake potentials physics of particle accelerators. AIP Conference Proceedings 184, New York, p. 526.

ACCELERATORS, POTENTIAL-DROP

R. G. HERB AND G. M. KLODY, *National Electrostatics Corporation, Middleton, Wisconsin*

	Introduction	27
1.	**Description of Potential-Drop Accelerators**	28
1.1	Types of Potential-Drop Accelerators	28
1.1.1	Cockcroft–Walton Voltage Multiplier Accelerator	28
1.1.2	Symmetric Cascaded Voltage Multiplier Accelerator	29
1.1.3	Dynamitron Accelerator	29
1.1.4	ICT (Insulating Core Transformer) Accelerator	30
1.1.5	Van de Graaff Accelerator	31
1.1.5.1	Two-Stage (Tandem) Accelerators	31
1.1.6	Tandetron Accelerator	32
1.1.7	Pelletron Accelerator	32
2.	**Applications of Potential-Drop Accelerators**	35
2.1	Ion Implantation to Produce Semiconductor Devices	35
2.2	Ion Beam Processing for Reduced Corrosion and Wear	36
2.3	Materials Analysis with Ion Beams	36
2.3.1	PIXE (Particle Induced X-Ray Emission)	36
2.3.2	RBS (Rutherford Backscattering) Analysis	37
2.3.3	Channeling	38
2.3.4	ERD (Elastic Recoil Detection)	38
2.3.5	Resonance Scattering	38
2.3.6	NRA (Nuclear Reaction Analysis)	38
2.3.7	Activation Analysis	39
2.4	Microprobe: The Ion Beam Microscope	39
2.5	AMS (Accelerator Mass Spectrometry)	39
2.6	Electron Beam Processing	40
2.7	FEL (Free Electron Laser)	40
2.8	Electron Cooling	41
2.9	Electron Microscope	41
3.	**Summary**	41
	Acknowledgments	41
	Glossary	41
	Further Reading	41

INTRODUCTION

Potential-drop accelerators were the first to open nuclear physics to extensive experimentation. After more than half a century, they continue to play an important, and in many respects dominant, role in nuclear physics.

Basic research in nuclear and atomic physics was virtually the sole application of potential-drop accelerators until the 1980s, when their uses for applied physics began to expand rapidly in many fields. They became the new tools for the new techniques required for analysis and production in many emerging fields, especially in microelectronics and molecular biology. Advances in the designs and construction of potential-drop accelerators, which are described in Sec. 1, have given them the precision, reliability, simplicity, and versatility needed for broader application. Their unique capabilities, developed during decades of use in basic research, are now available to apply previously untapped knowledge and techniques.

The widespread and varied applied uses of potential-drop accelerators, described in Sec. 2, include ion implantation in the manufacture of semiconductor devices, age determination of samples thousands and even millions of years old, and nondestructive analysis to characterize computer chips, analyze biomedical specimens, monitor pollution, and locate oil. High power electron beams are used for enhancing insulation on wire, producing shrink tubing, and sterilizing dis-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

posable medical products, and for tunable, microwave lasers.

1. DESCRIPTION OF POTENTIAL-DROP ACCELERATORS

A potential-drop accelerator consists of a high-voltage terminal supported by an insulating column, a means of generating the high voltage, a source of charged particles, and an evacuated acceleration tube through which the particles are accelerated. The potential drop from the high-voltage terminal to ground potential accelerates the charged particles to an energy equal to the charge on the particle times the terminal voltage ($E = qV$).

Potential-drop accelerators that operate above 10^6 V (1 MV) are usually enclosed in a steel pressure vessel. The vessel contains electrically insulating gas at high pressure, which significantly reduces the size of these accelerators and is required for reliable operation above about 1 MV. SF_6 is the most widely used insulating gas for accelerators. Its dielectric strength is 2.7 times that of air at 1 atm (101 kPa) and rises approximately linearly with pressure to 7 atm, where SF_6 sustains fields of about 200 kV/cm.

For stable operation at high voltages, the insulators of the accelerator support structure and the acceleration tube must be subdivided by metal electrodes which are connected to a resistive voltage divider for voltage grading from the terminal to ground.

1.1 Types of Potential-Drop Accelerators

Potential-drop accelerators are often identified with the type of high voltage generator they use. Some systems generate high voltage by rectifying ac voltage. Others use electrostatic charging, in which a mechanical system carries the charge to the terminal.

Each type of accelerator described below has advantages for specific applications, and each is serving needs in the growing range of applied uses of accelerators. For analytical precision and for voltages above 5 MV, the electrostatic accelerator is far superior. For certain neutron-induced reactions, ion implantation to a few hundred keV, and electron beam applications, the Cockcroft–Walton or symmetric cascade rectifier accelerator may be the most convenient. Needs for very intense positive ion or electron beams of a few MeV energy are most easily met by the insulating core transformer (ICT) and Dynamitron accelerators.

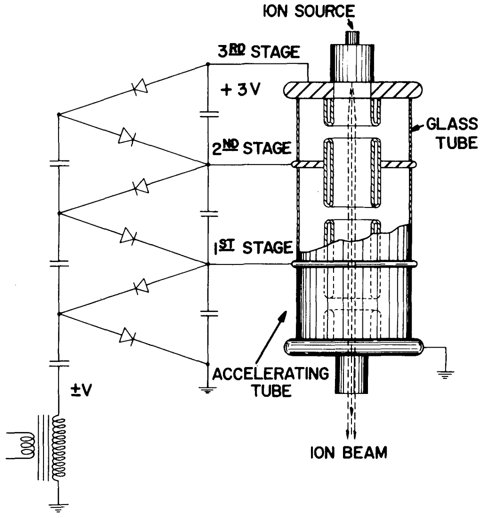

FIG. 1. Schematic of a Cockcroft–Walton accelerator with three voltage multiplier stages.

1.1.1 Cockcroft–Walton Voltage Multiplier Accelerator. This accelerator was the first to be used successfully for nuclear transmutation and gained wide recognition when results were published in 1932. It employs a circuit developed by H. Greinacher in 1920, which uses two stacks of series-connected capacitors as illustrated in Fig. 1. One capacitor stack is connected at one end to ground and at the other to the load, which in this case is an acceleration tube. The voltage across each capacitor in this stack is constant except for voltage ripple.

One terminal of the second capacitor stack is connected to a transformer giving peak voltages of ± V. Voltages at all points along this capacitor stack oscillate over a range of $2V$. Series-connected rectifiers link the two capacitor stacks. As voltages on the transformer-driven stack oscillate, charge is transferred stepwise through the rectifiers from ground to the terminal. Except for ripple from power drain and stray capacitance, the terminal voltage is constant and approximately equal to $2VN$, where N is the number of stages.

The terminal voltage accelerates charged particles from the ion (or electron) source

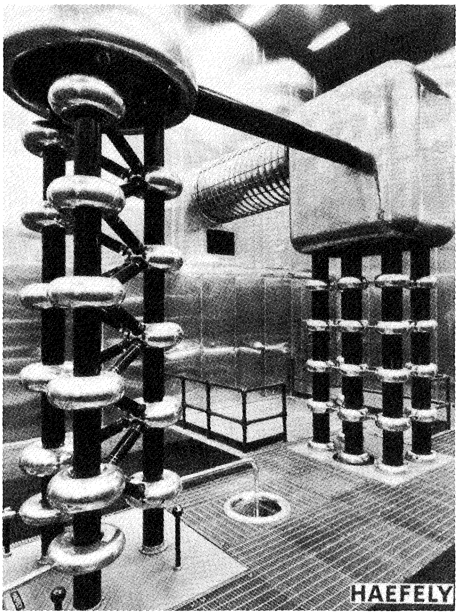

FIG. 2. 850-keV Haefely accelerator for injection of pulses of hydrogen ions into a linear accelerator at the Los Alamos National Laboratory (photo courtesy of Emile Haefely & Co., Ltd.).

through the acceleration tube, which is continually pumped. The acceleration tube sections may be connected to the power supply stages as in Fig. 1 or may be equipped with resistors for potential grading from high voltage to ground.

Cockcroft–Walton accelerators are in widespread use throughout the world and are especially advantageous for applications requiring beam currents of a few milliamperes in the voltage range of a few hundred kilovolts. Many are used to accelerate deuterons to bombard targets containing deuterium or tritium to yield neutrons, e.g., for neutron activation analysis. Neutron generators are available commercially from a number of manufacturers.

Cockcroft–Walton accelerators operating in open air near 1 MV are large and require a very large room to avoid voltage flashover. Figure 2 shows an 850-keV accelerator of this type. It generates pulsed hydrogen ion beams for injection into the high-intensity linear accelerator at the Los Alamos National Laboratory for production of intense meson beams. Air-insulated accelerators with Cockcroft–Walton charging have been manufactured by Emile Haefely & Co. Ltd. and N. V. Philips Gloeilampenfabrieken.

1.1.2 Symmetric Cascaded Voltage Multiplier Accelerator. G. Reinhold of Haefely has shown that, up to a few hundred kV, voltages developed by cascade voltage multipliers using the circuit of Fig. 1 do not depart greatly from $2VN$. Above about 500 kV, however, voltages achieved fall substantially below the value from this simple expression, because of stray capacitance and other effects. He developed a circuit called a symmetrical cascade rectifier, which eliminates most shortcomings of the circuit of Fig. 1. It employs two transformers and two capacitor stacks that oscillate in voltage. Both oscillating stacks feed one fixed-voltage capacitor stack. Using this symmetric system, Haefely built an open-air voltage generator without an acceleration tube that operated to 2.5 MV. Nearly all commercially available high voltage power supplies now use the symmetrical cascade rectifier circuit.

High Voltage Engineering–Europa and National Electrostatics Corp. build open-air accelerator systems with symmetric circuit charging to 500 kV for ion implantation and ion beam analysis. Air-insulated accelerators ranging from20 kV to over 500 kV are utilized as injectors for almost all MeV and GeV accelerators. Figure 9 shows the open-air 500-kV injector for the 25-MV tandem accelerator at the Oak Ridge National Laboratory.

Size and space requirements for open air systems increase rapidly with voltage and appear to set a practical upper limit of about 1.5 MV. Using high pressure gas for insulation greatly reduces the size of higher voltage accelerators. Haefely built accelerators insulated by gases such as SF_6 or a mixture of N_2 and CO_2 ranging in voltage from about 1 MV up to 4 MV. Nissin–High Voltage Co. builds SF_6-insulated Cockcroft–Walton accelerators from 500 kV to 3 MV for their industrial electron beam processing systems.

1.1.3 Dynamitron® Accelerator. Cleland invented a cascaded rectifier system called the Dynamitron, in which the rectifiers are driven in parallel (Fig. 3). Rectifiers connected in series between ground and the high voltage terminal are in two rows on opposite sides of the acceleration tube of the Dynamitron. The high voltage column is enclosed

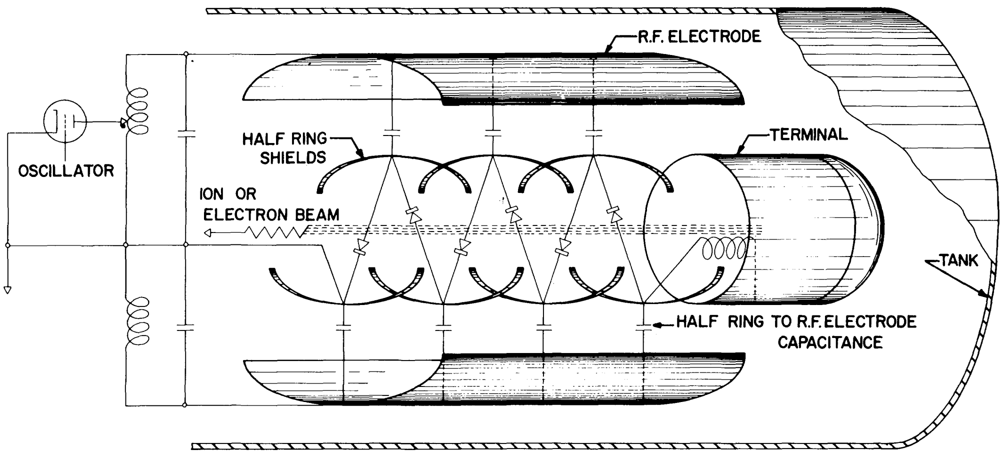

FIG. 3. Schematic of a Dynamitron accelerator in a pressure tank.

by half rings with smooth exterior surfaces to inhibit corona and spark discharge. Large, semicylindrical rf electrodes, mounted near the wall of the tank, surround the accelerator column. As indicated in Fig. 3, the half rings on the column provide capacitive coupling between the rectifiers and the large rf electrodes. The rf electrodes form the tuning capacitance of an *LC* resonant circuit driven by a separate high frequency (~100 kHz) power supply.

Since ac power is fed in parallel to each of the rectifiers, the large storage capacitors that are between successive stages in other cascaded rectifier systems are not required. This reduces stored energy in Dynamitrons to levels comparable with accelerators using electrostatic charging. Minimizing stored energy is important, especially in multimegavolt accelerators, because high stored energy released in a discharge can damage capacitors, rectifiers, column components, and acceleration tubes.

Dynamitrons are built by Radiation Dynamics, Inc. RDI has manufactured single-stage and two-stage Dynamitrons up to 4 MV for research applications requiring high current ion beams. Figure 4 shows a 4 MV positive ion Dynamitron accelerator. The tank for SF_6 insulating gas is open to show the accelerator column.

Most Dynamitrons are built to accelerate electrons for industrial polymerization and sterilization. These applications require electron energies up to a few MeV and high beam power. Systems built by Radiation Dynamics range from a 0.5 MV model with beam power of 50 kW and a model for 75 kW of 1.5 MeV electrons to a 5-MV Dynamitron with electron beam power of 200 kW.

1.1.4 ICT (Insulating Core Transformer) Accelerator. The High Voltage Engineering Corporation developed the ICT circuit for high-current electron beam applications. Rectifiers in the ICT power supply are on modular, flat, circular decks which are stacked one above the other. Nominal 460 V, three phase, 50 or 60 Hz line power drives the primary coil of the power supply transformer. Input power is magnetically coupled to secondary coils (three per deck) by a laminated three phase iron core which is electrically insulated between each deck. The secondary coils on each deck energize the rectifier circuit on the

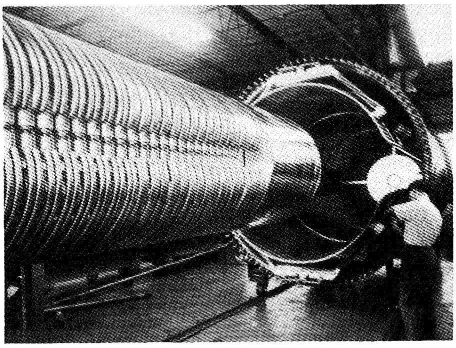

FIG. 4. A 4-MeV Dynamitron accelerator with pressure tank opened to show the high voltage column (photo courtesy of Radiation Dynamics, Inc.).

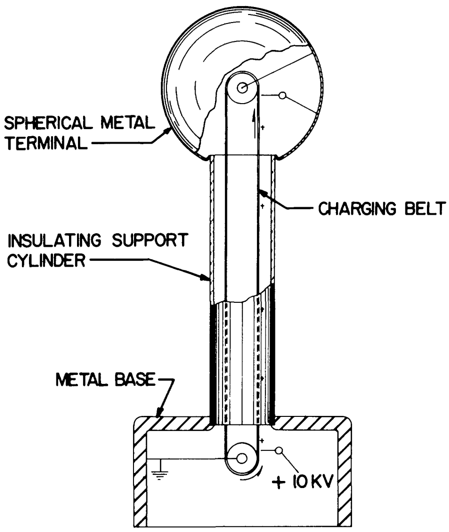

FIG. 5. Schematic of a Van de Graaff generator for operation in atmospheric air.

deck. This minimizes electrical stresses within the stack. HVEC builds ICT electron accelerators for energies from 0.3 to 3 MeV and beam power up to 100 kW, primarily for their industrial electron beam processing systems.

1.1.5 Van de Graaff Accelerator. Accelerators that use an insulating belt to carry charge to the high voltage terminal are named for Robert Van de Graaff, who first successfully demonstrated high voltage generation with belt charging in 1929.

In this device, illustrated in Fig. 5, electric charge is deposited on the insulating belt at the ground end. The motor-driven belt carries the charge to the smooth, well-rounded metal terminal, shown as a sphere in the figure.

To charge the belt, a corona discharge is maintained between a series of points or a fine wire on one side of the belt and the grounded lower pulley or a well-rounded, grounded plate on the other side of the belt. If the corona needles are at a positive potential, the belt intercepts positive ions moving from the needles toward the grounded pulley. The belt carries the positive charge into the sphere, where another array of needle points removes the charge, which passes to the outer surface of the sphere. The positive voltage on the terminal rises until discharged by a spark to ground or limited by corona or an electrical load, such as a resistor string along the insulating support or an ion current through an acceleration tube. Operating the needles at negative voltage produces a negative terminal voltage.

Charging electrodes must be well shielded from the electric field of the terminal, and charge must be carried well within the terminal before removal is attempted. Charging current is then independent of the voltage on the terminal.

The first use of a Van de Graaff voltage generator to accelerate ions was in 1932 at the Department of Terrestrial Magnetism of the Carnegie Institution, Washington, D.C. By 1934 this institution had built a second such accelerator, which was used extensively for nuclear physics research at energies to 1.2 MeV and is now on display at the Smithsonian Institute in Washington, D.C. Other belt-type accelerators built for 2–3 MV in open air were very large (to over 12 m high).

Compact belt-charged accelerators in steel tanks containing high pressure insulating gas were developed at the University of Wisconsin in the 1930s under the direction of Herb. Many basic design features of potential-drop accelerators, still applicable today, were determined in these early machines at Wisconsin. In 1940, Wisconsin completed a 4.5-MeV accelerator, which was used for precise nuclear physics measurements. It was the highest energy potential-drop accelerator for over a decade.

In 1947, High Voltage Engineering Corporation began commercially manufacturing belt-charged accelerators insulated by high pressure gas for electrons and positive ions to 4 MeV.

1.1.5.1 Two-Stage (Tandem) Accelerators. In 1958, High Voltage Engineering Corporation completed a potential-drop accelerator to produce high energy beams by accelerating ions twice. In this two-stage or "tandem" accelerator, illustrated in Fig. 6, injected negative ions are accelerated from ground potential to a positively charged high voltage terminal. In the terminal, a thin gas (or foil) target strips electrons from the ions. The resulting positively charged ions are accelerated a second time, away from the terminal, back to ground potential.

Tandems accelerate protons to an energy of $2eV$, where e is the electron charge and V is the terminal voltage. Heavier ions, which can be stripped to multiple positive charge states, are accelerated to even higher energies. For example, negative oxygen ions injected into a tandem operating at 10 MV are first accelerated to an energy of 10 MeV. If they are stripped of all electrons, the oxygen ions emerge from the terminal with a charge of $+8e$ and gain 80 MeV in the second acceleration for a total energy of 90 MeV.

The tandem principle was originally patented in 1935 by Bennett of Ohio State University, and Peter independently demonstrated keV protons from a tandem with negative-to-positive charge change at the Giessen University in 1936.

HVEC built many single-stage and tandem accelerators, which are in operation in laboratories around the world. The largest reach terminal voltages of over16 MV. Many are used for applied physics applications.

In 1988, Vivirad Corporation purchased the Accelerator Division of High Voltage Engineering Corporation to build accelerators for research and materials analysis. High Voltage Engineering–Europa also manufactures Van de Graaff accelerators to about 2 MV, primarily for ion implantation. Nissin–High Voltage Company manufactures single-stage and tandem accelerators with belt charging to 3 MV for basic research and materials modification and analysis.

1.1.6 Tandetron® Accelerator. In 1980, Purser designed the Tandetron line of compact tandem accelerators for materials analysis, accelerator mass spectrometry, MeV ion implantation (Fig. 7), and basic research. The high voltage generator is a Dynamitron-type parallel-driven cascade rectifier circuit with silicon rectifiers and a 50-kHz driver. It delivers several milliamperes of charging current at up to 3 MV with high stability. For accessibility, the high voltage stack is at right angles to the accelerator column. SF_6 at about 5 atm insulates the accelerator and power supply.

General Ionex Corporation manufactured Tandetrons, many with specialized accessories that they developed for materials analysis, until 1988, when Genus Corporation purchased General Ionex. In 1989 Genus sold the materials analysis line to High Voltage Engineering–Europa, retaining the line of ion implantation equipment.

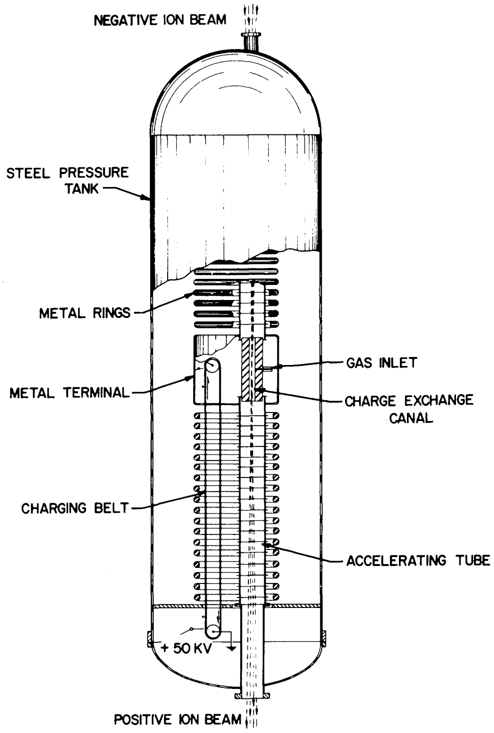

FIG. 6. Schematic of a two-stage (tandem) electrostatic accelerator in a pressure tank.

1.1.7 Pelletron® Accelerator. Pelletron accelerators use a charging chain, rather than a belt or rectifier stack, to generate high voltage on the accelerator terminal. The chain (Fig. 8) consists of steel cylinders (pellets) joined by links of solid insulating material such as nylon. The chain is intrinsically spark protected (undamaged in test sparks at 30 MV). Pellet charging current is adjustable and highly uniform, so that terminal voltages are easily maintained at very precise values with very little voltage ripple. This high charging stability is important and in some cases required for applied uses of accelerators. Many belt-charged accelerators have been converted to Pelletron charging.

The pellets are charged inductively. There is no contact with the charging electrodes (inductors) and no electrical discharge between the pellets and inductors. For positive terminal voltage, the charging inductor is at negative voltage (to -60 kV) to induce positive charge on the pellets at a motor-driven, grounded pulley.

The pellets carry the charge into the terminal, where the charge is removed at another pulley. The pellets are then negatively

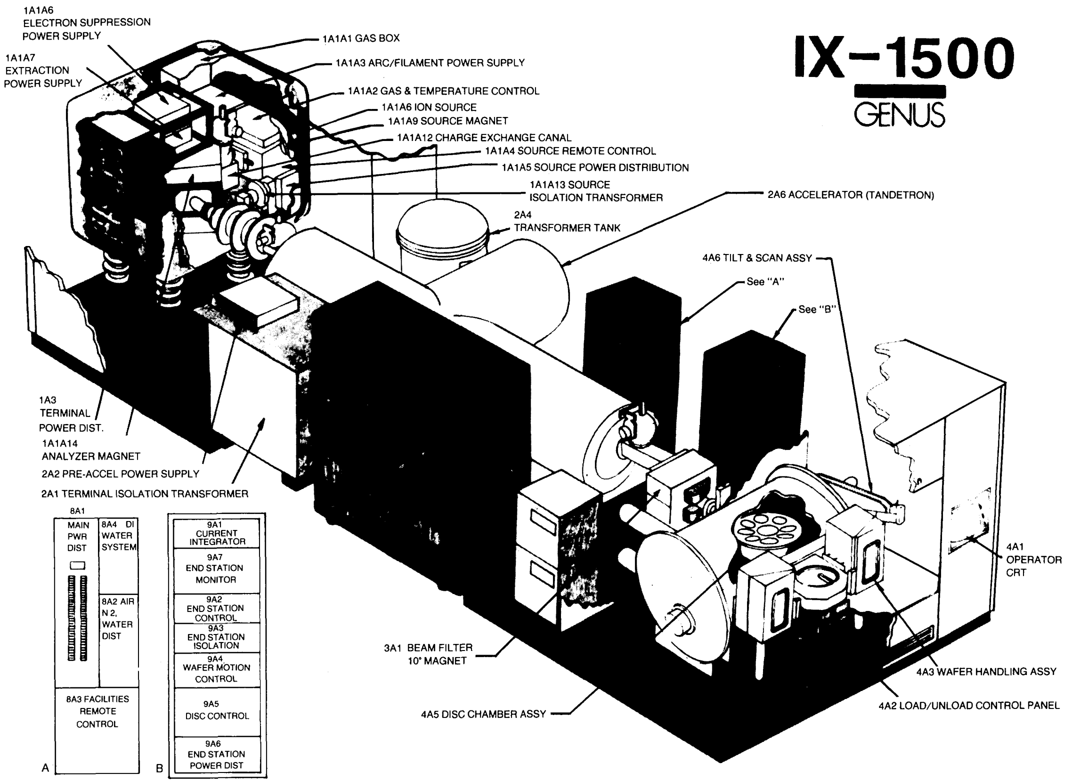
IX-1500
GENUS
1A1A6 ELECTRON SUPPRESSION POWER SUPPLY
1A1A7 EXTRACTION POWER SUPPLY
1A1A1 GAS BOX
1A1A3 ARC/FILAMENT POWER SUPPLY
1A1A2 GAS & TEMPERATURE CONTROL
1A1A6 ION SOURCE
1A1A9 SOURCE MAGNET
1A1A12 CHARGE EXCHANGE CANAL
1A1A4 SOURCE REMOTE CONTROL
1A1A5 SOURCE POWER DISTRIBUTION
1A1A13 SOURCE ISOLATION TRANSFORMER
2A6 ACCELERATOR (TANDETRON)
2A4 TRANSFORMER TANK
4A6 TILT & SCAN ASSY
See "A"
See "B"
1A3 TERMINAL POWER DIST.
1A1A14 ANALYZER MAGNET
2A2 PRE-ACCEL POWER SUPPLY
2A1 TERMINAL ISOLATION TRANSFORMER
8A1
MAIN PWR DIST
8A4 DI WATER SYSTEM
8A2 AIR N 2, WATER DIST
8A3 FACILITIES REMOTE CONTROL
A
9A1 CURRENT INTEGRATOR
9A7 END STATION MONITOR
9A2 END STATION CONTROL
9A3 END STATION ISOLATION
9A4 WAFER MOTION CONTROL
9A5 DISC CONTROL
9A6 END STATION POWER DIST
B
4A1 OPERATOR CRT
3A1 BEAM FILTER 10° MAGNET
4A3 WAFER HANDLING ASSY
4A5 DISC CHAMBER ASSY
4A2 LOAD/UNLOAD CONTROL PANEL

FIG. 7. Drawing of an MeV ion implanter with a Tandetron accelerator (photo by William Mackey Photography, Georgetown, MA, courtesy of K. Purser).

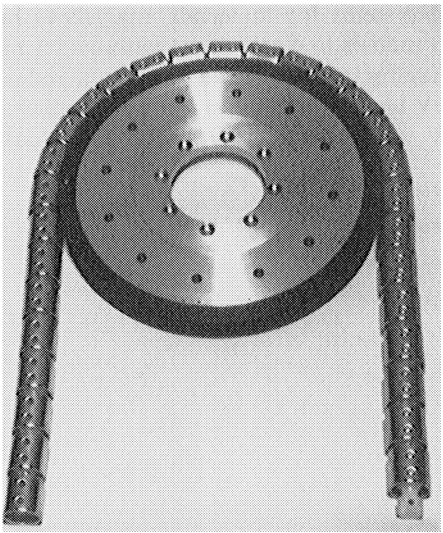

FIG. 8. Pelletron charging chain and pulley (National Electrostatics Corp. photo).

charged in the terminal, so that the up-going and down-going runs of the chain contribute equally to charging. Reversing inductor voltage polarities gives a negative terminal voltage.

Acceleration tubes in Pelletrons are unique. Their ceramic insulators and titanium electrodes are bonded together with a metal. Flanges are vacuum sealed with thin metal gaskets. These tubes are bakable for organic-free, ultra-high vacuum operation. This is important for reliability at high voltages and for contaminant-free vacuum applications. Earlier tube designs use organic glues to bond glass insulators to metal electrodes.

Pelletron accelerators were developed by the National Electrostatics Corporation in the 1960s. NEC manufactures single-stage and tandem Pelletron systems from 1 MV up to 25 MV for industries and laboratories around the world.

Most Pelletrons above 4 MV are vertically oriented for simplicity of support. The insu-

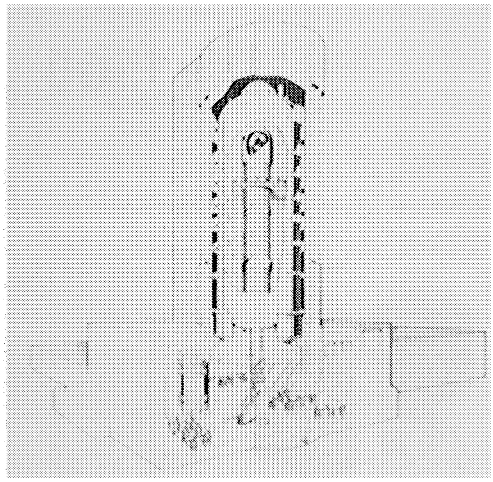

FIG. 9. Drawing of the 25 MV folded tandem at the Oak Ridge National Laboratory, showing 180° magnet in the high voltage terminal and raised service platform around the insulating support column. Multiple target stations are below accelerator. Cylinder (lower left) is a 500 keV ion injector (drawing courtesy of National Electrostatics Corp.).

lating support column is an assembly of standard column modules, each module holding 1 MV. The insulating column supports are posts of alumina ceramic and titanium.

The highest voltage potential-drop accelerator, the 25 MV tandem accelerator at the Oak Ridge National Laboratory (Fig. 9), utilizes this modular column design. It is also one of the few "folded" tandems. In the folded design, both acceleration tubes are mounted in a single insulating support column. Ions are accelerated up to the terminal, where a 180° magnet, shown in the cutaway in Fig. 9, reverses the beam direction for the second acceleration down to the target locations. The magnet completely filters ions from the stripper before the second acceleration, so there is no extra beam loading, and ion transmission is straightforward. Also shown in the figure is a service platform in a raised position in the tank for access to the high voltage column. The folded design, first successfully used by Naylor in New Zealand, conveniently locates both ion source and sample chamber near the system control panel. For very large tandems (20 MV and higher), it reduces the size and costs for the building, insulating gas, and pressure vessel.

Larger potential-drop accelerators (above ~8 MV) are used for basic research, not applied physics, because simpler, smaller cyclotrons and linacs are adequate for high energy beam applications like medical radiation therapy. In the 1990s, applied uses for large potential-drop accelerators may well be developed, e.g., for high-mass accelerator mass spectrometry (AMS) (Sec. 2.5) and shorter wavelength free electron lasers (FELs) (Sec. 2.7).

National Electrostatics also manufactures the S-Series line of small, simplified Pelletrons, mainly for applied physics uses. These accelerators have an inexpensive, cast acrylic support column. They range from 1-MV light-ion systems for materials analysis to heavy ion models to 4 MV, some with over 1 mA of charging current for high current industrial MeV ion implantation. Figure 10 shows a 3-MV tandem S-Series system with dual ion

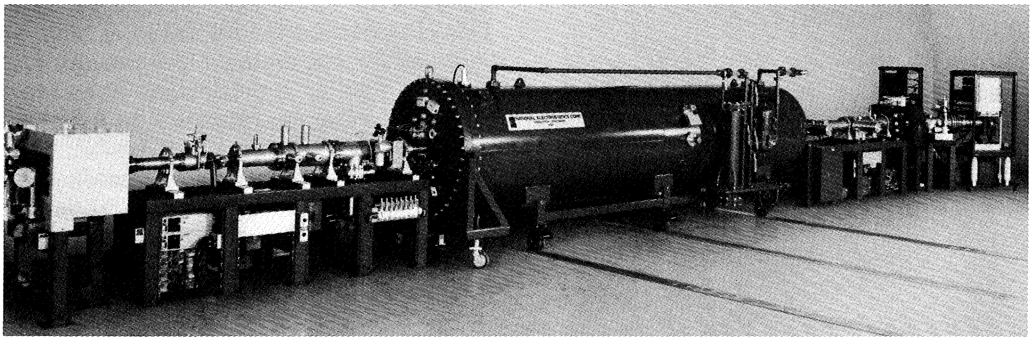

FIG. 10. S-Series Pelletron 3-MV tandem accelerator with two ion sources and high resolution beam handling (National Electrostatics Corp. photo).

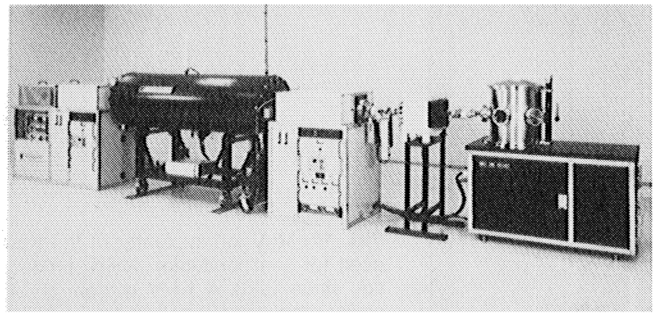

FIG. 11. Computer-controlled spectrometer with variable energy tandem accelerator (100-keV to 3.3-MeV helium ions) and multisample analysis chamber in use for materials analysis service (photo courtesy of Charles Evans & Assoc., Inc.).

sources and high resolution beam handling in use for materials modification and characterization.

Pelletrons with computer-automated, multisample end stations are used for production ion implantation, automated materials analysis, accelerator mass spectrometry, and biomedical analysis. Figure 11 shows a computer-automated Pelletron system used by a materials analysis service company.

2. APPLICATIONS OF POTENTIAL-DROP ACCELERATORS

From their beginnings in the early 1930s until well into the 1970s, the emphasis, development, and prolific construction of potential-drop accelerators centered around basic research in nuclear and atomic physics. In the 1980s a broad range of new applications of accelerators rapidly expanded their use into many fields, as illustrated in the examples below. By the end of the 1980s, nearly all new potential-drop accelerators were being built for applied physics uses. Many older accelerators, originally built for basic physics research, were also being converted for use for applied physics.

2.1 Ion Implantation to Produce Semiconductor Devices

Ion implantation is a process used to add specific "impurities" to bulk semiconductor materials to produce microscopic electronic circuits. In ion implantation, a beam of accelerated ions is scanned uniformly across the semiconductor material, e.g., a silicon wafer. The ions penetrate (are implanted) into the wafer. Photoresist is used to limit implantation to specific regions. Typical implant doses are from 10^{13} to 10^{16} ions/cm^2.

Since before 1970, manufacturers have made increasing use of ion implantation in the production of electronic devices. Implanters for energies from 20 to 200 keV are commonplace for creating *p*-type and *n*-type regions near the surface of the wafer. As manufacturers developed microchip circuits with more complex structures, higher energies were needed for deeper implants. Today many implanters are in use for implants at energies up to 400 keV. Most systems for implant energies to 400 keV use an air-insulated, positive voltage single-stage accelerator with a symmetrical cascade rectifier circuit power supply to generate the acceleration voltage. Eaton/Nova, Varian Extrion, Applied Materials, and other companies manufacture ion implanters operating at energies to a few hundred keV.

MeV implanters produce ion beams at energies from a few hundred keV to over 1 MeV. At these higher energies, ions penetrate deeper into the wafer. Deeper implants permit higher density circuits and can create conducting or insulating layers below surface circuitry to inhibit effects from stray charges, e.g., in computer memory chips. Multiple implants at different energies to over 1 MeV are used to produce deep complementary metal-oxide semiconductor (CMOS) wells, with economic and performance advantages over wells produced by slower, energy-consumptive thermal diffusion. MeV ion implantation in III–V materials (GaAs, etc.) produces very high speed switching circuits and light-emitting and light-sensitive solid state devices. MeV implants can program logic circuits late in their production cycle and enhance circuit resistance to radiofrequency and radiation interference.

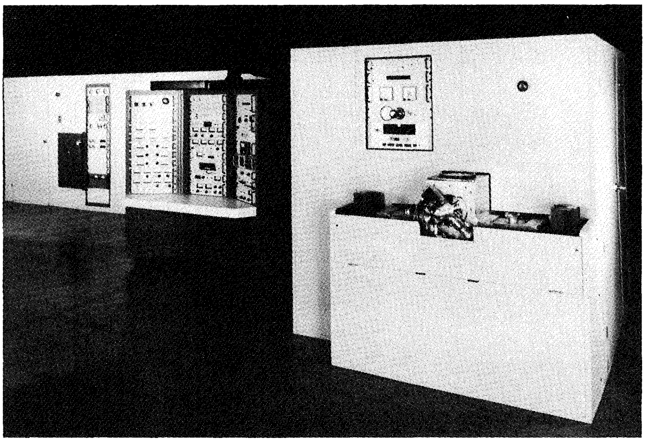

FIG. 12. MeV ion implanter of the type used for semiconductor device production since 1983. A 1-MV tandem accelerator produces ion beams for implantation at energies from 200 keV to over 3 MeV. Microcomputer-controlled cassette-to-cassette wafer handler is in right foreground (National Electrostatics Corp. photo).

Production use of ion implantation at energies of 1 to 3 MeV began in 1983 (see Fig. 12), and more than a dozen MeV implanters were installed in production lines during the next five years. National Electrostatics Corp., Genus Corporation, High Voltage Engineering–Europa, and Nissin–High Voltage Co. manufacture MeV implanters which use potential-drop accelerators for microelectronic device production applications.

2.2 Ion Beam Processing for Reduced Corrosion and Wear

Ion implantation is used to modify mechanical and chemical properties of surfaces to increase their resistance to wear or chemical corrosion. Implanting nitrogen into low carbon steel increases surface hardness for improved resistance to abrasive wear. Implanting titanium into steel reduces the coefficient of friction. Implants of such elements as Cr, Ta, Pd, P, and B into metals can reduce the chemical reactivity of the metals, e.g., for longer component lifetime in chemical processing systems.

These applications generally use high implant doses ($>10^{17}$ ions/cm^2) over large surface areas. This requires very high-current ion beams (10's of milliamps), but beam purity and energy dispersion are much less critical than in implants for electronic devices.

Harwell Laboratory in England has made notable contributions to the techniques and understanding of this applied use of accelerators.

2.3 Materials Analysis with Ion Beams

Materials analysis is a rapidly expanding application of potential-drop accelerators. These nondestructive techniques use MeV ion beams to identify elements and measure their concentrations in a wide range of materials.

2.3.1 PIXE (Particle Induced X-Ray Emission). PIXE identifies elements and their relative abundances from the energy spectrum of characteristic x rays emitted by atoms in the sample, which are excited by an incident MeV ion beam, usually protons. Compared to using electrons to excite the atoms, PIXE generates very little background, making it sensitive to trace concentrations ($<10^{-6}$ for elements of $Z \geqslant 10$).

PIXE was originally developed at the Lund Institute of Technology in Sweden in the 1970s. It is now used worldwide to measure trace elements in biomedical samples, ceramics, petroleum, polymers, and crystals, and on the surfaces of metals and semiconductor materials. PIXE analysis of air samples is used to monitor aerosol pollution. With a scanning proton microprobe (Sec. 2.4), PIXE provides high-resolution lateral mapping of trace element distributions in organic tissues, minerals, etc.

PIXE analysis uses standard equipment developed for atomic and nuclear physics research. Simple accelerator systems easily produce the few microamperes of beam at energies of a few MeV needed for most PIXE applications. Figure 13 shows a typical sample chamber with a multiple-sample holder for batch analysis. The metal block at an angle to

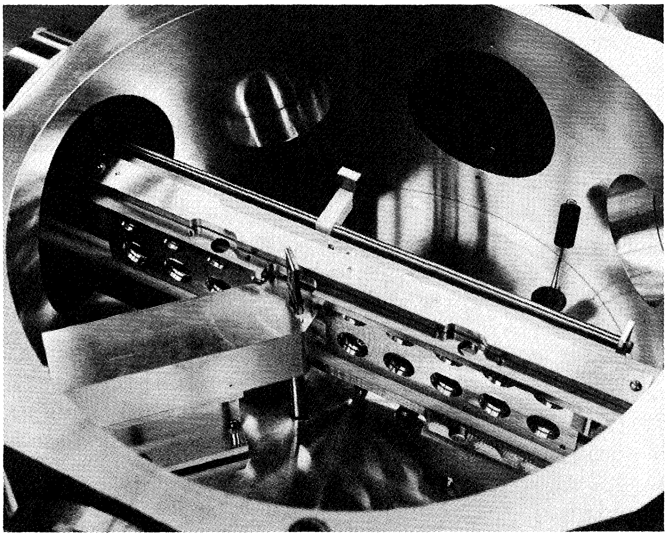

FIG. 13. PIXE target chamber with multiple-sample holder and Si(Li) detector housing with remotely selectable filters.

the sample changer houses a Si(Li) x-ray detector, positioned near the samples for detection efficiency. Selectable filters absorb background x rays to increase sensitivity.

For objects not compatible with analysis in a high vacuum chamber, PIXE is done outside the vacuum by directing MeV-energy protons through a thin foil "window" into the atmosphere. For example, the ALGAE accelerator facility at the Louvre Museum in Paris uses in-air PIXE to analyze art and archaeological objects.

2.3.2 RBS (Rutherford Backscattering) Analysis. RBS analysis measures depth and density profiles of elements at and near the surface of samples. Nuclei in the sample elastically scatter incident MeV helium ions, which are measured at a backward angle for increased sensitivity.

Rutherford scattering is simply elastic scattering due to the electrostatic repulsion of the positively charged incident and target nuclei. Kinematics and absolute cross sections can be calculated from first principles. For a given incident beam, the energy of the scattered ions increases with increasing target mass. Cross sections are proportional to the square of [the target atomic number (Z) divided by the incident energy]. From the calculated cross section and the measured incident beam current, RBS can measure elemental concentrations directly. Calibration standards are not required. RBS is, therefore, often used to measure concentrations of known target elements, e.g., in surface layers created by vapor deposition or ion implantation.

Figure 14 shows a RBS spectrum for bulk silicon with a thin insulating SiO_2 layer and a coating of tungsten silicide. The heaviest element (W in Fig. 14) gives the highest energy backscattered ions. Using known rates of energy loss in the materials, layer thicknesses are determined from peak widths. The area under each peak is converted to an elemental concentration profile, using the measured incident beam current and calculated cross sections.

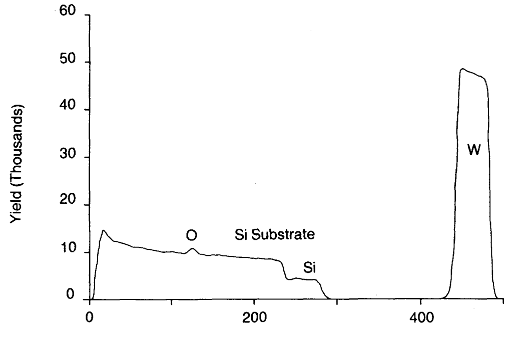

FIG. 14. RBS spectrum of 29% W, 71% Si layer (240 nm thick) over SiO_2 layer (50 nm) on bulk Si substrate, at 160° with 2.275 MeV incident helium ions. RBS measures silicide layer thickness, W/Si ratio, and impurities. (Spectrum from Charles Evans & Assoc., Inc.)

The fast, simple, quantitative, and nondestructive nature of RBS analysis has opened many applications, of which the following are but a few. RBS measures depth profiles of B, P, and As in silicon for electronics. Thin film analysis, a major use of RBS, characterizes metal silicide layers in very large scale integrated (VLSI) devices, measures thicknesses of laminated layers, measures relative concentrations, e.g., of the constituents in YBaCuO high temperature superconducting films, and analyzes nitride layers used as oxidation barriers and in surface hardening. RBS also measures the elemental composition at the interfaces between layers to clarify diffusion and reaction rates, catalytic effects, and adhesion and alloying processes.

RBS systems typically use small tandem accelerators, which produce microampere helium beams over a wide range of selectable energies. Lower energy beams increase depth resolution in thin films. Higher energies probe deeper and improve mass resolution for heavy elements, as in GaAs samples.

Charles Evans & Associates, Inc. and Universal Energy Services are two of the companies that use potential-drop accelerators to provide materials analysis services to industrial and research customers. Charles Evans & Assoc. utilizes the RBS spectrometer shown in Fig. 11 for automatic analysis of up to 100 samples per batch by RBS and related techniques, including elastic recoil detection (ERD), nuclear reaction analysis (NRA), and channeling.

2.3.3 Channeling. Channeling, an effect seen in RBS analysis of crystals, decreases backscatter yields by as much as 98% when the beam is aligned with a crystal axis. Ions incident parallel to the axis (within about 1°) are steered by correlated small-angle deflections along the axis, preventing the close approach to lattice atoms required for large angle scattering, except at imperfections and impurities in the crystal.

Channeling is used, especially in the semiconductor industry, to measure crystal perfection, amount and depth of lattice disorder, locations of substitutional and interstitial impurities, and lattice damage from ion implantation in silicon and subsequent crystal regrowth after annealing. Surface defects and amorphous surface layers increase the angular divergence of the beam, so that some of the beam cannot be channeled by the crystal below the surface. This effect is used to measure thickness and crystallinity in surface layers. Channeling improves RBS sensitivity to surface impurities by reducing backscatter from the substrate and permits deeper analysis with PIXE and NRA in crystals.

Most RBS systems can include channeling measurements by adding a goniometer for accurate sample orientation and collimators to limit the angular spread of the beam.

2.3.4 ERD (Elastic Recoil Detection). ERD is like RBS analysis, but ERD measures the recoiling target ions at forward angles. In most applications, the incident beam and detector are at grazing angles to the sample. Thin foil samples use transmission analysis with the beam at normal incidence and the detector at 0°, where recoil cross sections are largest. Absorbing filters in front of the detector suppress the (heavier) ions from the beam and other interfering recoils.

ERD is sensitive to light elements in heavy substrates and is used to profile hydrogen in metals and semiconductors. Nuclear reaction analysis gives better depth resolution, but ERD is faster and easy to analyze.

2.3.5 Resonance Scattering. Resonances increase the scattering cross section at particular energies and angles. Proton backscattering from carbon, resonant at 1.73 MeV, is used to measure photoresist thickness on wafers coated with a metal, such as gold. With helium RBS, the high background of backscatter from the Au obscures the carbon signal.

To measure elemental concentrations with resonance scattering, the detection system must be calibrated using standard targets of known concentrations.

2.3.6 NRA (Nuclear Reaction Analysis). As beam energy increases, the use of nuclear reactions becomes possible. The reaction ^{1}H(^{15}N, $\alpha\gamma$) is resonant at 6.4 MeV and provides good depth resolution for profiling hydrogen in metals, semiconductors, and glass. Unlike ions, gamma rays do not suffer energy loss in the material. Using channeled beams to induce gamma-emitting reactions, NRA locates impurities deep within a crystal. The sensitivity of reactions to specific isotopes is useful for analyses in several fields, including mineralogy and biomedicine.

2.3.7 Activation Analysis. Particle-induced nuclear reactions are used to generate trace quantities (10^{-10}) of radioactive isotopes at the surface of objects like gears or inside a pipe. Decreases in radioactivity in excess of natural decay measure the rate of surface wear or erosion.

Neutron activation analysis is used to locate underground oil deposits. Accelerators to generate the neutrons operate at about 100 kV and are small in diameter (to <45 mm) to fit into drill holes. Neutrons are also used to detect explosives in airplane cargo.

2.4 Microprobe: The Ion Beam Microscope

Typical MeV ion beams are about 1 cm in diameter. While ideal for many applications, this is too large for measuring lateral distributions of elements.

The microprobe produces a beam spot of from 10 to 20 μm down to less than 300 nm diameter for measurements with spatial resolution of a few microns or better. It uses energy-analyzed beams from an accelerator and includes a beam collimator, strong focusing lens, chamber with sample holder, detectors, and a means of scanning the beam across the sample. Data are collected at points in a line along the sample or over a grid of from 10×10 to 1000×1000 or more points. Microprobes mainly use MeV proton beams and PIXE analysis, primarily for biomedical studies, but also for archaeology, mechanical engineering, mineralogy, and solid state.

For high lateral resolution, microprobes require beams with very little angular divergence. The emittance (a measure of beam divergence in position-momentum phase space) of ion beams from accelerators is far too high for the microprobe. Reducing emittance requires a collimator to remove the more divergent ions, which also reduces beam current. Better accelerator voltage stability improves beam emittance for higher microprobe currents.

To focus the collimated beam, most systems use a quadrupole lens with micrometer adjustments to translate and rotate lens elements to minimize or eliminate aberrations. Magnetic or electrostatic scanning of the focused beam spot across the sample is generally used, because it is easier to control and less sensitive to beam current variations than mechanically moving the sample across a fixed beam. A computer is essential for data acquisition, storage of spectra from each grid point, and data reduction for each element detected. Color graphics displays show the variations in elemental concentrations within the microstructure (Plate 1).

In the 1980s over 30 laboratories built microprobe beamlines for their accelerators. The Universities of Oxford and Melbourne manufacture microprobe components and systems for sale to other laboratories.

2.5 AMS (Accelerator Mass Spectrometry)

AMS measures isotopic concentrations down to 10^{-16} with high precision using only milligram samples. Naturally occurring radioisotope concentrations are measured to date art and archaeological objects, ground water, lunar samples, and meteors and to study glacier and sediment formation and ocean mixing. AMS of tracer isotopes is used in hydrological and atmospheric studies.

The samples analyzed by AMS are not in a target chamber, but in the ion source, which creates ion beams from the sample. The ions are accelerated to MeV energies for high resolution isotopic analysis.

Carbon dating, the primary use of AMS, determines the age of samples from the concentration of ^{14}C, whose half-life is 5.73 ka (a=year). The beam from the ion source contains ^{12}C, ^{13}C, and ^{14}C ions in the same relative abundances as in the sample. The isotopes are accelerated to the same energy (~ 12 MeV), so that the electrostatic and magnetic analyzers following the accelerator give maximum mass resolution and eliminate all interference. Several of the over 30 AMS laboratories have automated their carbon-dating systems. Ages to10 ka require only 15 min for 1% accuracy, compared to 12 h for decay counting techniques, and ages to 50 ka can be measured.

Tandem accelerators are best suited to most AMS applications, because they can transport all of the isotopes without intermediate adjustments and accelerate them to high energy for high resolution. Molecular isobaric interferences, a common problem for other mass spectrometers, are eliminated in tandems, because the stripper dissociates molecules, and the fragments are easily resolved. The use of negative ions, required for tandems, eliminates potential interference from ^{14}N (much more abundant than ^{14}C) in carbon dating, because N^- ions cannot be created.

Other isotopes used in AMS measurements include ^{10}Be (1.6 Ma half-life) to measure ages to 15 Ma in polar ice core samples and ocean sediments, ^{26}Al (720 ka) for hydrology and cosmology, ^{36}Cl (310 ka) to map ocean currents and date ground water, ^{129}I (16 Ma) to study petroleum deposit formation, and ^{41}Ca (130 ka) for paleoanthropological studies.

AMS also measures traces of stable isotopes, e.g., $^{3}He/^{4}He$ ratios in geological samples. Impurity concentrations below 10^{-10} are measured by using a secondary ion mass spectrometer (SIMS) as the ion source for a tandem AMS system (Super-SIMS).

The first AMS measurements with tandem accelerators were at McMaster University (Canada) and the University of Rochester (New York). Dedicated systems, like those at the Universities of Toronto and Arizona, offer AMS service. ETH in Zurich, Switzerland, uses a 6 MV tandem for their AMS facility, which is unsurpassed in performance (over 2500 samples per year with 1% to 5% precision).

2.6 Electron Beam Processing

Industrial plants use electron beams from accelerators or gamma rays from ^{60}Co to induce physical and chemical changes to enhance the properties of materials. High throughput of thin materials favors electron beam (EB) processing, because of its high beam power and processing efficiency. Gamma rays are used for bulky items, e.g., sterilization of shipping cartons of medical disposables, because gamma rays penetrate deeper into materials than electrons.

Cross-linking of polymers, the primary industrial EB application, improves the resistance to heat and solvents of insulation on wires and cables and is vital to the production of heat shrinkable tubing and film. EB is used for polymerization to cure thin coatings and improve adhesives, for sterilization of single-item medical disposables, and for crystal modification for permanent color in glass and gems. Potential EB applications include the irradiation of food products for sterilization and preservation, treatment of drinking water, disinfection of municipal sewage, and treatment of SO_2 and NO_x in industrial smokestack exhausts.

High Voltage Engineering Corp. and Nissin–High Voltage manufacture potential-drop accelerators for EB processing systems at energies from 0.3 to 3 MeV with electron beam power up to 100 kW. They use electron beam scanning to process sheet material or multiple runs of wires or tubing.

Radiation Dynamics, Inc. has built many Dynamitron accelerators from 1.5 to 4.5 MV with electron beam power to 150 kW for EB facilities worldwide. They have also completed a Dynamitron which produces 200 kW of 5 MeV electrons. In addition to providing more power for electron processing, this system can generate x rays of intensities competitive with large gamma ray facilities (1–2 MCi of ^{60}Co) for processing bulky items.

2.7 FEL (Free Electron Laser)

The free electron laser (FEL) uses a high power MeV electron beam in an undulatory magnetic field (wiggler) to amplify electromagnetic radiation to produce high power, coherent "light" at wavelengths from millimeters to the visible region. Since the FEL is not restricted to atomic energy level transitions, laser frequencies are continuously tunable.

Electrostatic potential-drop accelerators produce electron beams with very low energy dispersion, which improves amplification efficiency in FELs. Electrostatic accelerators are also capable of dc electron currents. Linacs and electron storage rings, also used for FELs, produce electrons in short pulses.

The first facility for FEL applications was developed by Elias at the University of California, Santa Barbara (1984). It is used for studies in molecular biology, solid state, and atomic physics. The FEL utilizes a 6 MV electrostatic accelerator designed for recirculated electron beams. In this design, after the 6 MeV electron beam passes through the wigglers in the FEL system, the beam is directed back to the 6 MV terminal. By recollecting well over 90% of the beam in the terminal, beam currents in the ampere range are achieved with less than1 mA of charging current. Facilities are being designed with free electron lasers to provide shorter wavelengths, using higher energy electrostatic accelerators.

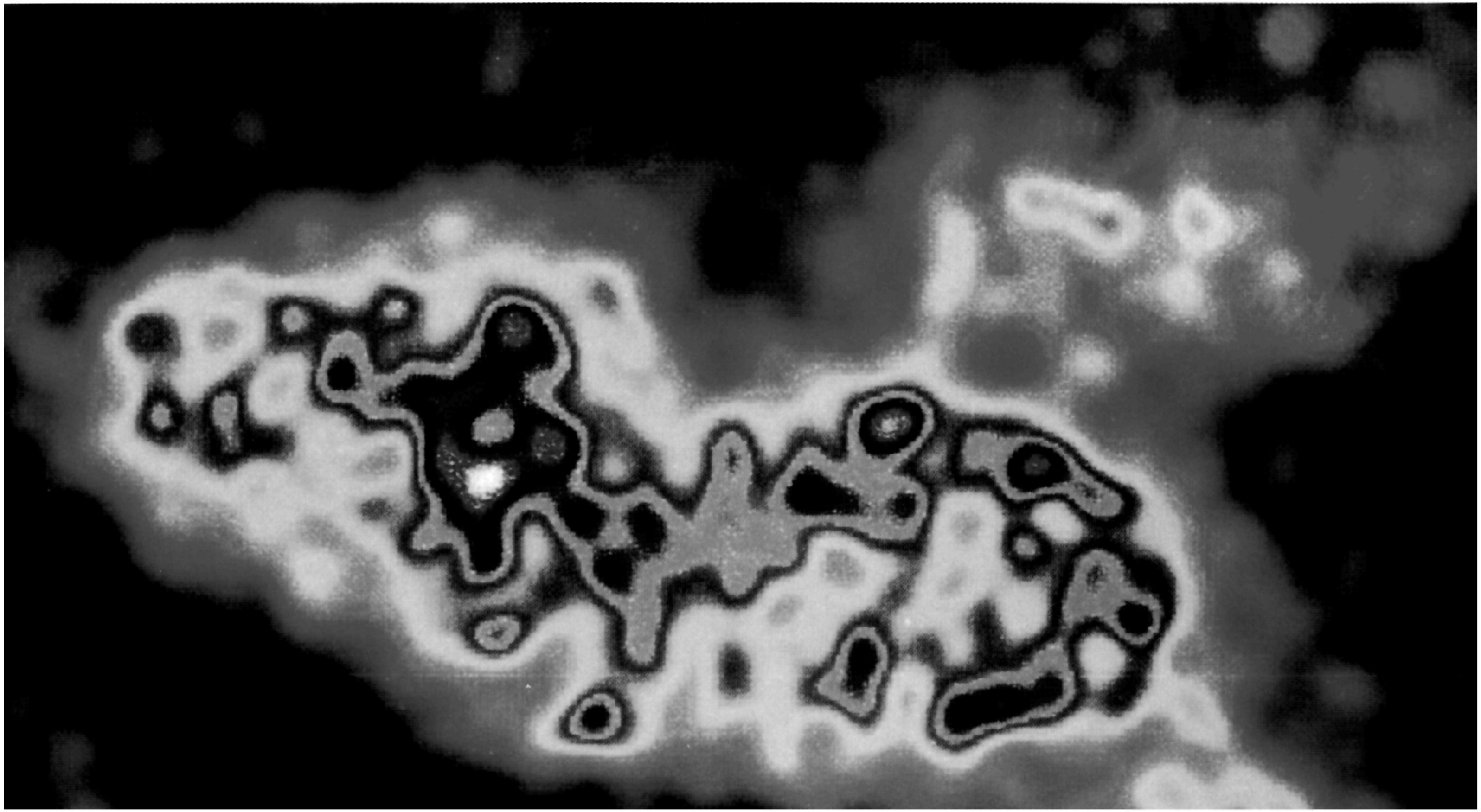

Vol. 1 – PLATE 1. PIXE map of copper in a freeze dried cryosection of mouse kidney tissue, using scanning proton microprobe. Color display (resistor color code) shows unusually high Cu concentrations in peripheral tubules, a factor related to Menkes disease, a fatal genetic disorder. (Photo courtesy of Macro Analytical Research Center, University of Melbourne.)

2.8 Electron Cooling

Electrons are used to increase the phase space density of beams from sources for high energy (GeV) accelerators. In the preacceleration storage ring, the beam repeatedly passes through a region of cool (low energy spread) electrons which have the same velocity as the beam. Collisions transfer momentum to the electrons, which are then extracted, until the energy spread of the beam has been reduced to that of the electrons.

To increase antiparticle beam intensities, higher antiparticle production energies are used. This increases to several MeV the electron energy required for cooling. Suitable electron beams were demonstrated in tests with an electrostatic accelerator at National Electrostatics Corp. in collaboration with the University of California, Los Angeles, and the Fermi National Accelerator Laboratory (supported by the U.S. Department of Energy). Using the circulated electron beam technique described for free electron lasers, the accelerator produced over 100 mA of 2 MeV electrons with 99.99% of the electrons recollected in the accelerator terminal.

2.9 Electron Microscope

Electron microscopes are like optical microscopes but use electrons at energies from tens of keV to several MeV to illuminate samples. The electron image is observed on a fluorescent screen, directed onto a photographic plate, or analyzed electrically. Scanning electron microscopes, which provide lateral imaging of samples, are in widespread and routine use in industry, biomedicine, and applied and basic research. A separate article in this *Encyclopedia* (ELECTRON MICROSCOPY) describes the significance and vast range of applications of this very important applied use of potential-drop accelerators.

3. SUMMARY

Potential-drop accelerators have emerged as reliable tools for an expanding range of applied uses in many industrial and applied science fields. The variety of potential-drop accelerators provides a spectrum of capabilities for high efficiency and high precision applications. The routine use of potential-drop accelerators continues to grow rapidly, often providing the only solution to meet existing needs.

ACKNOWLEDGMENTS

The authors are grateful to M. Cleland (Dynamitron) and C. Goldie (ICT) for information on electron beam processing and to all who provided photographs.

GLOSSARY

Beam: A directed stream of energetic electrons or atomic or molecular particles.

Charge State: The net charge on an ion divided by the absolute value of e (electron charge), i.e., the atomic number (Z) minus the number of electrons.

Charging: Pertaining to the means used to generate high voltage, as in charging system, charging current, etc.

Ion: A charged atomic particle, including charged molecular particles (radicals).

Stripping: The removal of electrons from energetic ions during their passage through a low density medium (foil, poor vacuum, etc.). Although stripping is generally undesirable in potential-drop accelerators, tandem accelerators utilize stripping in the terminal to advantage to achieve higher ion energies.

Terminal: A metal shell or frame with smooth, rounded exterior surfaces to prevent corona and flashover, supported by insulators, and operated at the high acceleration voltage. May also refer to the assembly of components housed in the terminal shell.

Voltage Grading: The controlled distribution of voltage along the acceleration tube and insulating support column, in most cases for a uniform voltage gradient, using a resistive voltage divider from the high voltage terminal to ground.

Voltage Flashover: A high voltage spark across the surface of an insulator.

Further Reading

Berejka, A. J., Läuppi, U. V., Thompson, C. C., Lyall, D., Mizusawa, K.-I., Cleland, M. R. (1988), *beta–gamma* **1**, 4–9.

Bromley, D. A. (Ed.) (1974), *Nucl. Instrum. Methods* **122**, 1–285.

Chu, W.-K., Mayer, J. W., Nicolet, M.-A. (1978), *Backscattering Spectrometry*, New York: Academic.

Cleland, M. R. (1986), "International Developments in Electron Accelerators for Industry and Medicine," presented at the Symposium on Electron Accelerators in Mexico, Salazar, Mexico, 26 September 1986.

Duggan, J. L., Morgan, I. L., (Eds.) (1989), "Proceedings of the 10th Conference on Applications of Accelerators in Research and Industry, Denton, TX," *Nucl. Instrum. Methods B* **40/41**, is one of many published proceedings from the numerous topical conferences and symposia held annually, which contain detailed and review articles on applied uses of accelerators.

Livingston, M. S., Blewett, J. P. (1962), *Particle Accelerators*, New York: McGraw-Hill.

Vis, R. D. (1985), *The Proton Microprobe: Applications in the Biomedical Field*, Boca Raton: CRC.

ACOUSTIC FILTERS AND RESONATORS

See FILTERS AND RESONATORS, ACOUSTIC

ACOUSTIC PROPERTIES OF LIQUIDS

STEPHEN V. LETCHER, *Department of Physics, University of Rhode Island*

Introduction 43

1. **Absorption and Dispersion of Sound in Liquids** 44

1.1 Propagation of Sound in Lossless Fluids 44

1.2 Viscosity and Thermal Conductivity 44

1.3 Relaxation Processes 45

1.4 Sound Propagation Near a Critical Point 47

2. **Experimental Techniques** 48

3. **Materials** 48

3.1 Simple Liquids 48

3.1.1 Liquefied Gases 48

3.1.2 Liquid Metals 48

3.2 Water and Associated Liquids 49

3.3 Seawater and Electrolytic Solutions 50

3.4 Organic Liquids 51

3.4.1 Rotational Isomerism 51

3.4.2 Monomer-Dimer Reaction 52

3.4.3 Vibrational Relaxation 52

3.5 Binary Mixtures 52

3.5.1 Nonassociated Liquids 52

3.5.2 Critical Phase Separation 52

3.5.3 Associated Liquids 53

3.6 Liquid Crystals 54

3.6.1 Intramolecular Relaxations 55

3.6.2 Acoustic Anisotropy 55

3.6.3 Relaxation Processes at Phase Transitions 56

3.6.4 Acousto-Optic Effect in Liquid Crystals 57

3.7 Amphiphilic Aggregates in Solution 57

3.7.1 Micelles 57

3.7.2 Microemulsions 58

3.7.3 Magnetic Liquids 58

3.8 Liquid Helium 58

3.8.1 Liquid ^{4}He 58

3.8.2 Liquid ^{3}He 60

Glossary 61

Works Cited 61

Further Reading 62

INTRODUCTION

The propagation of sound waves in fluid media has considerable practical and fundamental significance. The most widespread application of sound in liquids is underwater, or ocean, acoustics. Because seawater is electrically conducting, it is a poor medium for the propagation of electromagnetic waves and, as a result, sound waves are the best choice for long-range underwater detection and imaging. Another major application for sound-wave technology applied to a (mostly) liquid medium is medical diagnosis. Ultrasonic images complement those obtained with x rays, magnetic resonance, or positron emission and have the advantage of relatively low cost and low risk.

Knowledge of the acoustic properties of liquids is essential, not only because of the practical applications, but because sound provides a way to investigate some of the fundamental properties of condensed matter. In this article we will discuss the propagation of sound in a wide variety of liquids, in each case with anemphasis on what this reveals about the basic properties of the material. Sound waves are a gentle probe, varying the local temperature and pressure as they pass by. This energy can couple to internal degrees of freedom by, for example, altering the equilibrium distribution of the system among internal energy levels. A common theme, then, that occurs in all but the simplest liquids, is the relaxation of perturbed distributions towards equilibrium and the dissipation of acoustic energy that ensues. A secondary theme will be the type of relaxation that occurs at critical phase transitions—the dissipation of fluctuations of the parameter that specifies the degree of order of the ordered phase. We will be concerned throughout only with small-amplitude, linear sound waves;

digital Encyclopedia of Applied Physics

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

finite-amplitude effects are discussed in ACOUSTICS, NONLINEAR.

1. ABSORPTION AND DISPERSION OF SOUND IN LIQUIDS

1.1 Propagation of Sound in Lossless Fluids

The propagation of acoustic waves in fluids is governed by the principles of fluid mechanics and thermodynamics. When expressed as conservation equations, these are given by

$$\dot{\rho} = -\frac{\partial g_i}{\partial x_i} \quad \text{(mass conservation)}, \tag{1}$$

$$\dot{g}_i = -\frac{\partial \sigma_{ij}}{\partial x_j} \quad \text{(momentum conservation)}, \tag{2}$$

$$\dot{\epsilon} = -\frac{\partial q_i}{\partial x_i} \quad \text{(energy conservation)}, \tag{3}$$

where ρ is the fluid density, $g_i = \rho v_i$ is the mass current or momentum density, v_i is the particle velocity, σ_{ij} is the stress tensor, ϵ is the energy density, and q_i is the energy flux. Here we use the notation that repeated indices are summed over and i, $j = 1,2,3$ represent the x,y,z axes.

In a lossless fluid, we can ignore the energy conservation equation by assuming that the process is completely adiabatic. In an isotropic fluid without dissipation, the stress tensor is given by

$$\sigma_{ij} = P\delta_{ij} + \rho v_i v_j, \tag{4}$$

where P is the pressure of the fluid and δ_{ij} is the Kronecker delta. The pressure and density changes are in phase so that, to first order, the pressure is given by

$$P = P_0 + \left(\frac{\partial P}{\partial \rho}\right)_S (\rho - \rho_0), \tag{5}$$

where the derivative at constant entropy indicates an adiabatic process. If attention is limited to small-amplitude, longitudinal motion in the x direction, Eqs. (1), (2), (4), and (5) can be combined to form the one-dimensional acoustic wave equation, given by

$$\frac{\partial^2 v_1}{\partial x_1^2} - \frac{1}{c^2}\frac{\partial^2 v_1}{\partial t^2} = 0, \tag{6}$$

which has harmonic solutions for the acoustic variables of particle displacement and velocity and excess pressure and density. The particle velocity in the x direction, v_1, for example, is given by

$$v_1 = v_0 \exp[i(\omega t - kx)], \tag{7}$$

where the phase velocity is a constant, real number given by

$$c = \frac{\omega}{k} = \left(\frac{\partial P}{\partial \rho}\right)_S^{1/2}. \tag{8}$$

1.2 Viscosity and Thermal Conductivity

For a viscous liquid, a dissipative term must be added to the stress tensor in Eq. (4), in the form:

$$\sigma_{ij}^D = -2\eta A_{ij} - (\eta' - \tfrac{2}{3}\eta)\delta_{ij} A_{kk}, \tag{9}$$

where

$$A_{ij} = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right) \tag{10}$$

is the strain-rate tensor. The coefficient η is the shear viscosity, while η' is the volume viscosity coefficient associated with pure, shearless compressions. A finite volume viscosity means that, if the fluid is subject to a sudden pure compression, the density will not respond immediately, but will relax towards its new equilibrium value. There are a number of molecular and structural sources of this delay that will be discussed below. In fact, it is just the coupling of these internal processes to the sound wave through the volume viscosity that has made acoustic spectroscopy so useful in the study of the properties of liquids.

We will limit ourselves now to the effect of shear viscosity and will set $\eta' = 0$. Combination of the linearized, one-dimensional versions of Eqs. (1), (2), and (4) with Eq. (9) included will then yield the damped wave equation, given by

$$\rho_0 \frac{\partial^2 v_1}{\partial t^2} - \rho_0 \left(\frac{\partial P}{\partial \rho}\right)_S \frac{\partial^2 v_1}{\partial x^2} - \frac{4}{3}\eta \frac{\partial^3 v_1}{\partial x^2 \partial t} = 0. \tag{11}$$

We now assume a harmonic wave solution,

$$v_i = v_0 \exp[i(\omega t - k^* x)], \tag{12}$$

where the complex wave number is given by $k^* = k - i\alpha$, where α is the amplitude absorption coefficient. If $\alpha \ll k$, then one finds, to first order in α/k, that

$$c=\left(\frac{\partial P}{\partial \rho}\right)_S^{1/2} \tag{13}$$

and

$$\alpha=\frac{2\eta}{3\rho_0 c^3}\,\omega^2. \tag{14}$$

Conduction of heat from the compressed, warmer parts of a sound wave to the cooler, expanded parts dissipates acoustic energy and contributes a term to the absorption coefficient. The conservation of energy, Eq. (3), including thermal conductivity in the energy flux term, must be combined with the conservation of mass and momentum to yield the acoustic wave equation with a heat-conduction damping term. One finds that the acoustic absorption coefficient due to heat conductivity has the form

$$\alpha=\frac{(\gamma-1)}{2C_p\,\rho_0 c^3}\,\kappa\omega^2, \tag{15}$$

where κ is the thermal conductivity coefficient, C_p is the specific heat at constant pressure, and γ is the ratio of specific heats. Except for highly conducting fluids, such as liquid metals, the sound absorption due to heat conduction is generally negligible compared to that due to shear viscosity.

When simply added together, the combination of the shear viscosity and heat conductivity terms of sound absorption is called the classical sound absorption. The forms given in Eqs. (14) and (15) were developed in the 19th century by Stokes and Kirchhoff, respectively.

1.3 Relaxation Processes

All but the simplest liquids can be characterized by certain internal degrees of freedom that may consist, as we shall see below, of vibrational states, various structural configurations, dimerization, and a number of other possibilities. They have in common the ability to be described in thermal equilibrium by a distribution function, in which a certain fraction of the liquid is associated with one or another of the multiple states of the system. When the equilibrium conditions are perturbed, the distribution generally does not respond instantaneously, but relaxes more or less slowly towards its new equilibrium distribution. If the equilibrium conditions are changing sinusoidally, under the influence of a sound wave, the distribution lags in phase, continually trying to catch up with the changing equilibrium conditions. This phase lag results in a dissipation of energy and an attenuation of the sound wave.

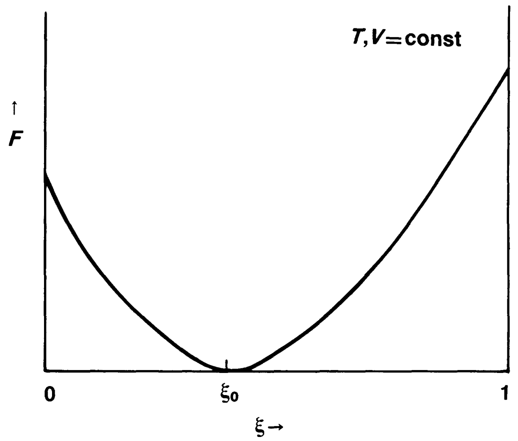

FIG. 1. Helmholtz free energy vs degree of reaction. The minimum at $\xi=\xi_0$ represents equilibrium.

A phenomenological description of this process can be obtained using the methods of irreversible thermodynamics (see THERMODYNAMICS, NONEQUILIBRIUM). A new thermodynamic variable must be introduced, called the order parameter or the degree of reaction, that describes the extent to which the internal degrees of freedom have approached their equilibrium distribution. The change of the Helmholtz free energy can then be expressed as

$$dF=-S\,dT-P\,dV-A\,d\xi, \tag{16}$$

where S, T, P, and V are the entropy, temperature, pressure, and volume, respectively, and where ξ is the degree of reaction and $A=-(\partial F/\partial \xi)_{T,V}$ is called the affinity. The free energy as a function of ξ for constant T and V will have the general form shown in Fig. 1, where ξ_0 represents the equilibrium distribution. When a sound wave perturbs the system, the new values of temperature and density result in a slightly changed curve with a new position of the minimum. It is assumed that the rate of change of ξ as it approaches its new equilibrium position is proportional to the negative slope of the curve, i.e., that

$$\frac{d\xi}{dt}=aA. \tag{17}$$

To find an expression for the relaxation time for this process, we can expand the affinity in terms of any three independent variables. A convenient choice for sound wave excitation is the combination of pressure, entropy, and degree of reaction. This expansion is given by

$$A(P,S,\xi)=\left(\frac{\partial A}{\partial \xi}\right)_{P,S}(\xi-\xi_0)+\left(\frac{\partial A}{\partial P}\right)_{S,\xi}\times(P-P_0)+\left(\frac{\partial A}{\partial S}\right)_{P,\xi}(S-S_0), \tag{18}$$

which can be written as

$$A(P,S,\xi)=\left(\frac{\partial A}{\partial \xi}\right)_{P,S}[\xi-\bar{\xi}\,(P,S)], \tag{19}$$

where

$$\bar{\xi}\,(P,S)=\xi_0+\left(\frac{\partial \xi}{\partial P}\right)_{A,S}(P-P_0)+\left(\frac{\partial \xi}{\partial S}\right)_{A,P}(S-S_0) \tag{20}$$

is the equilibrium value of ξ for given P,S. If we use this expression for A in Eq. (17), then we find that, for constant pressure and entropy, the relaxation time is given by

$$\tau_{P,S}=-\frac{1}{a}\left(\frac{\partial \xi}{\partial A}\right)_{P,S}. \tag{21}$$

This relaxation behavior can now be combined with thermodynamics (see, e.g., Beyer and Letcher, 1969) to provide an acoustic equation of state, given by

$$\tau_{P,S}\dot{\rho}+(\rho-\rho_0)=\frac{\tau_{P,S}}{c_\infty^2}\dot{P}+\frac{1}{c_0^2}(P-P_0). \tag{22}$$

Here

$$\frac{1}{c_\infty^2}=\left(\frac{\partial \rho}{\partial P}\right)_{\xi,S},\quad \frac{1}{c_0^2}=\left(\frac{\partial \rho}{\partial P}\right)_{A,S}. \tag{23}$$

This notation utilizes the fact that constant ξ, which refers to no reaction, occurs at very high frequencies and that constant (in this case, zero) A, which refers to equilibrium, occurs at very low frequencies.

When the acoustic equation of state is combined with the equations of mass and momentum conservation, one obtains a damped acoustic wave equation. Finally, if the absorption is much smaller than the wave number, one can find the velocity dispersion and the absorption to be given by

$$\frac{c^2}{c_0^2}=\frac{1+\omega^2\tau_{P,S}^2}{1+(c_0^2/c_\infty^2)\omega^2\tau_{P,S}^2} \tag{24}$$

and

$$\alpha\lambda=\frac{c^2}{c_0^2}\frac{\pi\epsilon\omega\tau_{P,S}}{1+\omega^2\tau_{P,S}^2}, \tag{25}$$

where $\epsilon=(c_\infty^2-c_0^2)/c_\infty^2$ is known as the relaxation strength. Generally, the dispersion is very small and one can let $c^2/c_0^2=1$ in Eq. (25). For a single relaxation process, the frequency dependence of the total absorption can be expressed as

$$\frac{\alpha}{f^2}=\frac{A}{1+f^2/f_r^2}+B, \tag{26}$$

where the first term is due to the relaxation process with relaxation frequency $f_r=(2\pi\,\tau_{P,S})^{-1}$ and the second term represents the background consisting of the classical absorption and any high-frequency relaxations. Three ways in which the absorption expression can be usefully plotted are shown in Fig. 2.

The properties of a particular liquid can be related to the acoustic absorption through the relaxation strength by noting that the sound velocity is given by $c=(\rho_0\kappa_s)^{-1/2}$. The adiabatic compressibility κ_s and other relaxing thermodynamic coefficients can be expressed as the sum of two terms—an instantaneous, constant ξ, part and a relaxing part. Then the relaxation strength can be written as

$$\epsilon=(\kappa_s^0-\kappa_s^\infty)/\kappa_s^0. \tag{27}$$

After considerable thermodynamic manipulation (see, e.g., Bauer, 1965 or Lamb, 1965), the relaxation strength can be put in the form

$$\epsilon=(\gamma-1)\frac{\delta C_P}{C_P^\infty}\left[1-\frac{\Delta V}{\Delta H}\frac{C_P}{V\beta}\right]^2, \tag{28}$$

where $C_p=C_p^\infty+\delta C_p$ is the specific heat at constant pressure, expressed in terms of its instantaneous and relaxing parts, β is the thermal expansion coefficient, and ΔV and ΔH are, respectively, the changes of the volume and enthalpy per unit change of the degree of reaction during the relaxation process at constant P and T.

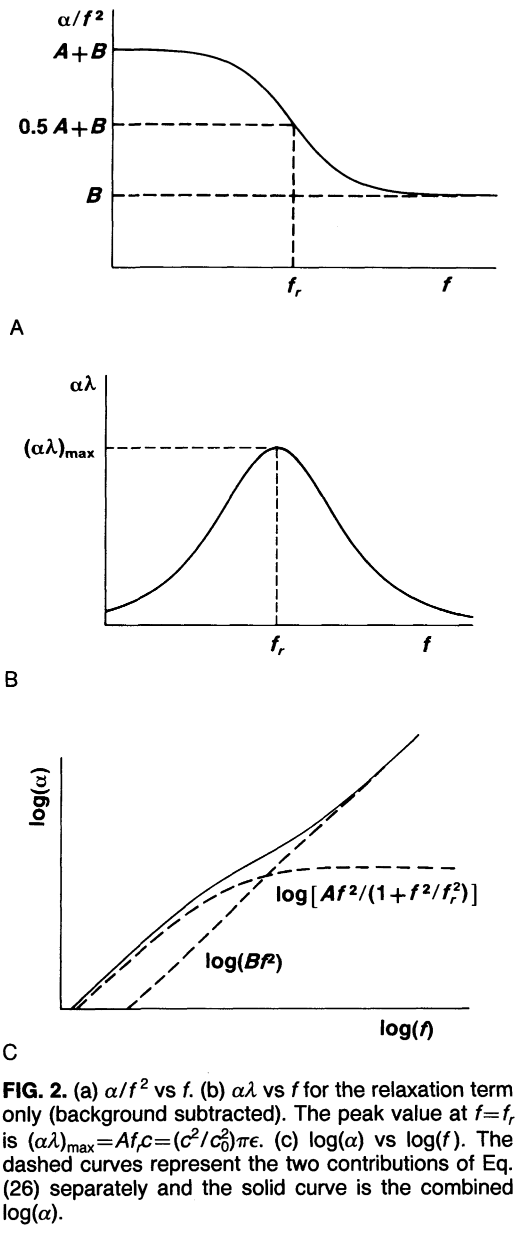

FIG. 2. (a) α/f^2 vs f. (b) $\alpha\lambda$ vs f for the relaxation term only (background subtracted). The peak value at $f=f_r$ is $(\alpha\lambda)_{\max}=Af_rc=(c^2/c_0^2)\pi\epsilon$. (c) log($\alpha$) vs log($f$). The dashed curves represent the two contributions of Eq. (26) separately and the solid curve is the combined log(α).

Equation (28) describes a relaxation process that is induced by either the pressure changes or the temperature changes of the sound wave. These are generally treated separately and are called structural and thermal relaxation processes, respectively. Structural processes occur most notably in water and alcohols and will be discussed further in Sec. 3.2. In temperature-induced relaxation process, ΔV in Eq. (28) may be negligible and the interpretation of the relaxation strength becomes a problem of finding an expression for $\delta C_p/C_p^\infty$ that describes a particular process.

For the relatively simple case of the two-state reaction illustrated in Fig. 3, Lamb (1965) has shown that the peak value of $\alpha\lambda$ [see Fig. 2(b)] is given by

$$(\alpha\lambda)_{\max}=\frac{\pi}{2}\frac{\gamma-1}{C_p}R\left(\frac{\Delta H}{RT}\right)^2 \times\exp\left[-\frac{\Delta H}{RT}\right]\exp\left[\frac{\Delta S}{RT}\right]. \tag{29}$$

Thus, sound absorption measurements can provide values for ΔH and ΔS, the differences of enthalpy and entropy between the two states of the reaction. Similarly, the relaxation frequency f_r provides values of ΔH_1 and ΔS_1, specifying the energy barrier between the two states.

1.4 Sound Propagation Near a Critical Point

General properties of critical phenomena are discussed in the article entitled CRITICAL PHENOMENA. A critical point marks the onset of a degree of long-range order, which is represented by an order parameter. The order parameter is finite in the ordered state and goes smoothly and continuously to zero at the critical temperature T_c. Local fluctuations of the order parameter have a characteristic size that diverges as $|T-T_c|$ goes to zero. Along with light scattering and thermodynamic measurements, propagation of ultrasound is a powerful way to study the nature of critical phase transitions. Near the critical point, the sound velocity generally has a frequency-dependent dip, but the most dramatic effect is

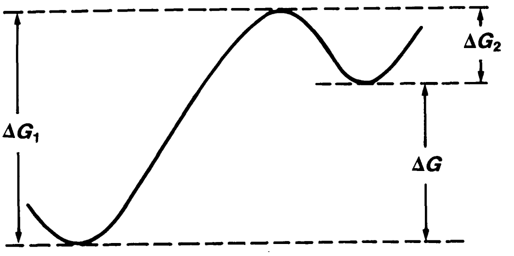

FIG. 3. Two-state energy diagram. ΔG is the difference in the Gibbs free energy between the two equilibrium states and ΔG_1 is the activation energy. The horizontal axis depends on the particular process considered.

a large increase in the absorption. The sound wave couples to the order-parameter fluctuations and the absorption can be calculated from the complex, frequency-dependent volume viscosity or specific heat that arises from the dissipation of the fluctuations. Ordinarily the relaxation time for the dissipation of fluctuations is very short, but the critical slowing down that occurs as T approaches T_c produces times that are accessible to ultrasound.

Early theoretical developments and experimental results (including those at the liquid-vapor critical point, which are not included in this article) have been discussed by Garland (1970). Recent theories have been based on mode-mode coupling and renormalization group theory. Theories for the case of binary mixture phase separation have been actively developed and have been summarized and compared by Tanaka and Wada (1985).

In addition to the coupling of the sound to fluctuations of the order parameter, which occurs on both sides of T_c, an additional source of absorption exists in the ordered phase near T_c. The value of the order parameter is strongly temperature dependent and the variation of temperature associated with the sound wave modulates the equilibrium condition. This is exactly the condition for a thermal relaxation process described in Sec. 1.3. In this context, it is referred to as the Landau–Khalatnikov mechanism.

2. EXPERIMENTAL TECHNIQUES

Techniques for measuring the velocity and absorption of sound waves in liquids will be mentioned here only briefly. For further details, see ACOUSTICAL INSTRUMENTATION, ULTRASONICS, and FILTERS AND RESONATORS, ACOUSTIC. A number of standard techniques exist, each with a range of frequencies over which it is reliable. To study the broad relaxation processes described below, it is convenient to have access to more than one technique.

At high frequencies (1–10 GHz), Brillouin scattering can be used. Laser light is scattering from thermally excited density fluctuations (phonons) and measurement of the frequency shift and the width of the resonance peaks provides the velocity and absorption of sound. At intermediate frequencies (1–1000 MHz), pulse methods are widely used. The absorption and velocity of the sound pulse are obtained from the decrease of amplitude and the time of flight over known distances. Phase-sensitive detection methods measure fractional changes of sound velocity of 10^{-4} to 10^{-6}. The pulse method is limited at high frequencies by the very large damping ($\alpha \sim f^2$) and at low frequencies by diffraction losses that overwhelm the very low absorption losses. At low frequencies (0.01–20 MHz), continuous-wave resonance techniques are used. The width of the resonance peak of a resonating system has a contribution from the absorption coefficient of the liquid sample. Parallel-plate resonators have been effective down to about 150 kHz. At lower frequencies, spherical or cylindrical resonant modes are used.

3. MATERIALS

3.1 Simple Liquids

From an acoustical point of view, simple liquids can be defined as those in which the speed of sound and α/f^2 exhibit no frequency dependence and for which α/f^2 has nearly the classical value.

3.1.1 Liquefied Gases. The acoustical properties of a few liquefied gases are given in Table 1. In each case, the absorption is slightly above the calculated classical value and is dominated by the viscous absorption. The small, positive $\alpha_{\text{meas}} - \alpha_{\text{classical}}$ suggests the existence of a relaxation process at high frequencies, perhaps a structural relaxation, but no frequency dependence of the absorption has been observed.

The superfluid properties of liquid helium disqualify it from being considered as a simple liquid; it will be considered separately in Sec. 3.8.

3.1.2 Liquid Metals. For a general description, see LIQUID METALS, STRUCTURE OF. The acoustic properties of a few liquid metals are included in Table 1. They share with liquefied gases the characteristic that α/f^2 is slightly (5%–30%) above classical and is frequency independent, but in this case the absorption is dominated by thermal conductivity. The excess absorption and the lack of observable frequency dependence again suggest

Table 1. Acoustic properties of some simple liquids.

Liquid	T (K)	c (m s^{-1})	dc/dT (m s^{-1} K^{-1})	$10^{15}\times\alpha/f^2$ (s^2 m^{-1}) Exp.	Classical
Ar	87	848	−8.7	14.5	11.5
N_2	70	925	−10.	14.5	10.4
O_2	70	1066	−7.9	11.6	7.6
Na	371	2531	−0.54	11.5	9.55
Sn	505	2474	−0.27	5.6	4.3
Pb	601	1820	−0.28	9.4	7.6

a form of structural relaxation with relaxation frequency above 100 MHz. Beyer and Ring (1972) analyzed the experimental results in a number of liquid metals and concluded that the uncertainty in the experimental results for sound absorption and for the thermophysical properties that appear in the classical absorption [Eqs. (14) and (15)] render the validation of any theory of the excess absorption extremely difficult.

The electrical conductivity of liquid metals leads to interesting magnetohydrodynamic effects when sound propagation is observed in the presence of an external magnetic field. When the sound is propagated perpendicular to the magnetic field, a linear, damped sound wave exhibits dispersion and a magnetically induced relaxation of the absorption, given by (Beyer and Letcher, 1969)

$$c=\left[c_\infty^2+\frac{\mu H_0^2/\rho}{1+f^2/f_m^2}\right]^{1/2}, \tag{30}$$

$$\alpha_m=\frac{\sigma B_0^2}{1+f^2/f_m^2}, \tag{31}$$

where σ is the electrical conductivity, μ is the magnetic susceptibility, H_0 is the applied magnetic field intensity, $f_m=\mu\sigma c_0^2/2\pi$, and $B_0=\mu H_0$. At low frequencies ($f\ll f_m$), the magnetic field lines are "frozen in" and cause a stiffening of the fluid and an increase in the sound velocity. At high frequencies ($f\gg f_m$), the fluid moves relative to the field lines and the induced currents and associated Joule heat losses produce an increase in sound absorption. At high frequencies, the sound velocity has its normal value. The relaxation frequency f_m is about 10 MHz for liquid alkali metals and 0.5 MHz for mercury.

3.2 Water and Associated Liquids

In water the absorption of sound is about three times greater than predicted by shear viscosity, but no frequency dependence of α/f^2 is observed at frequencies up into the gigahertz range. This suggests that a relaxation process is taking place with a very high relaxation frequency. This excess absorption persists even at 4 °C where the compression due to the sound wave causes no temperature change. This rules out a thermal relaxation as defined above; the excess absorption must be due to the pressure or volume changes associated with the sound wave. Furthermore, the ratio of the measured to classical absorption and, therefore, the ratio of volume viscosity to shear viscosity are nearly independent of temperature and pressure suggesting that the two processes originate in similar mechanisms. This is true even though the hydrogen bonding of water leads to the peculiar property that both viscosity coefficients have a minimum as a function of pressure at low temperatures (see Davis and Jarzynski, 1972).

Water and the alcohols are in the class called associated liquids, which have a strong tendency to form bonded aggregates even in the liquid state. The distribution of molecules in aggregates is pertubed by the pressure changes of a sound wave, resulting in a structural relaxation. Hall (1948) presented a two-state theory in which water is assumed to have two different arrangements of the molecules. In one state, the molecules are in the hydrogen-bonded, rather open, low density structure characteristic of the crystalline solid and in the other state the molecules are more closely packed. Hall assumed that the close-packed structure had a higher energy and that the states were separated by an energy barrier, as depicted in Fig. 3. The energy barrier is taken to be the same as that encountered in theories of shear viscosity, since in each case it is the breaking of hydrogen bonds that is the effective mechanism. The observed pressure dependence of the absorption and velocity of sound, on the other hand, requires that the icelike structure be assigned the higher free energy in the two-state model.

The two-state model has provided a very useful picture of structural relaxation processes, but the project of fitting this concept into a general model of the structure of water is still in progress. Application of such a model to the primary alcohols, which have similar acoustic properties as water, has been less successful.

3.3 Seawater and Electrolytic Solutions

The propagation of sound in the ocean, including the topics of scattering, refraction, the existence of sound channels, etc., is discussed in the article entitled ACOUSTICS, UNDERWATER. In this section, we will restrict our attention to the mechanisms for the absorption of sound in seawater.

When an electrolyte is dissolved in a polar solvent like water, the Coulomb attraction of the ions toward the solvent dipoles tends to disturb the short-range, crystallinelike order of the solvent and to replace it with a structure in which a relatively tightly bound sheath of solvent molecules surrounds each dissolved ion. In water, this results in a decrease in volume and in compressibility, which, for most electrolytes, results in an increase of sound velocity as the solute concentration increases. The sound absorption is changed rather little by the addition of electrolytes except in a few important cases.

The frequency dependence of the absorption of sound in seawater has traditionally been displayed in a log-log plot, as in Fig. 2(c), and is shown schematically in Fig. 4. At frequencies above a few hundred kilohertz, the absorption is proportional to the square of the frequency and has the value for pure water, due primarily to high-frequency structural relaxation and shear viscosity. The low-frequency absorption is increased substantially by a relaxation process involving magnesium sulfate ions with a relaxation frequency of 130 kHz. At lower frequencies another relaxation process occurs with a relaxation frequency of 1 kHz, which has recently been shown to be due to the minute concentration of boric acid present in seawater. Each of these processes has the general form described in Sec. 1.3, in which the equilibrium chemical distribution is perturbed by the volume changes induced by the sound wave, but they are both multistep processes that are more difficult to analyze.

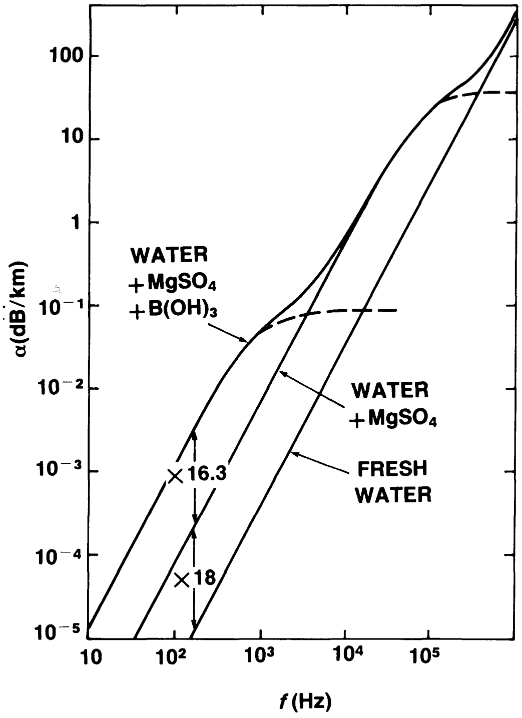

FIG. 4. Schematic representation of the absorption of sound in seawater. The top curve represents the total absorption, which is due to the fresh-water background (lowest curve) and the two independent relaxation processes (dashed curves). From Fisher and Simmons (1977).

The model for the magnesium sulfate relaxation and, by extension, that of other 2-2 electrolytes, is described in detail by Steuhr and Yeager (1965). It is based on the association of the Mg^{2+} and $(SO_4)^{2-}$ ions in a way that necessitates the displacement of water molecules from the hydration sheaths that surround each ion. Four different states in the process are identified: ions far removed from each other, ions approaching each other as close as intact hydration sheaths will allow, removal of one water molecule so that the remaining intervening molecule is shared by the two hydration sheaths, and, finally, direct ionic interaction following removal of

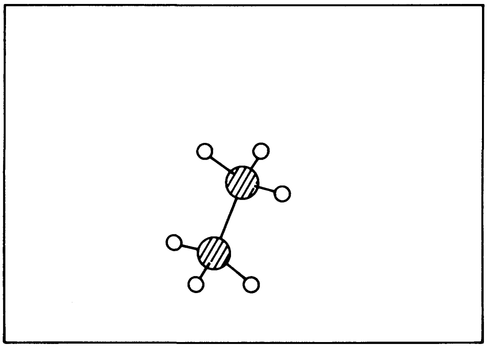

FIG. 5. Ethane molecule: carbon, ●; hydrogen, ○.

the last intervening water molecule. Each step of this process involves a volume change and takes some time, producing a relaxation effect. The final step in the $MgSO_4$ process is the slowest step that creates the relaxation at 130 kHz.

The source of the 1-kHz relaxation in seawater was much more difficult to establish. First of all, it was difficult to be certain that there was an effect. The absorption at 1 kHz is extremely small and the wavelength is 1.5 m, making laboratory acoustical experiments difficult. Only careful analysis of ocean data taken over long ranges produced definite evidence of the relaxation. A review of these ocean experiments has been provided by Waton and Candau (1983). The role of boric acid, which has a concentration in seawater of only 0.43 mmol L^{-1}, was identified by using temperature-jump experiments. Although the details of the process are still somewhat ambiguous, the general outline is now clear. Detailed discussion can be found in the reviews by Mellen *et al.* (1987) and Mallo *et al.* (1985). The dissociation of boric acid provides the correct relaxation time, but the sound absorption is negligible unless it takes place in the presence of magnesium carbonate or calcium carbonate. One is left, then, with a complex coupled exchange reaction involving this combination of components.

3.4 Organic Liquids

We are concerned here with liquids that are unassociated and that have a large relaxing sound absorption caused by the temperature variations of the sound wave. These are sometimes called Kneser liquids after Kneser (1938), who proposed that thermal relaxation is responsible for the excess absorption in liquids just as it is in gases (see ATMOSPHERIC ACOUSTICS). Initial observations of thermal relaxation in liquids were made in the 1930s; our current understanding owes much to the detailed measurements of Lamb and his colleagues and has been reviewed by Lamb (1965).

Several different molecular degrees of freedom can couple to the sound wave resulting in a relaxation process. Here we mention some examples of each.

3.4.1 Rotational Isomerism. A number of liquids are composed of molecules that can exist in two or more isomeric forms. If these forms have different energies, sound energy can be absorbed when the distribution among the isomeric forms is perturbed. A simple example is 1,1,2-trichloroethane. The ethane molecule is shown in Fig. 5. The two ends of the molecule can rotate relative to each other, but the rotation is not entirely free. The energy has local maxima and minima as a function of rotation angle, but in ethane, since all of the minima have the same energy, no acoustic energy is transferred. 1,1,2-trichloroethane, which is ethane with two hydrogens at one end of the molecule and one at the other replaced by chlorine, has the energy diagram shown in Fig. 6. In this case, sound energy is dissipated as the distribution is perturbed. The relaxing part of the absorption is shown in Fig. 7. From the temperature dependence of the maximum value of $\alpha\lambda$, one can find ΔH and from the relaxation time, the excitation energy ΔH_1. The results for 1,1,2-

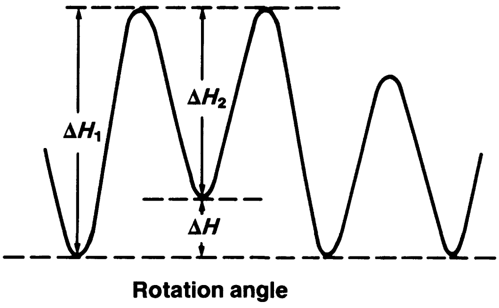

FIG. 6. Energy diagram for the rotational states of 1,1,2-trichloroethane. The higher-energy equilibrium corresponds to the three chlorines being on the same side of the molecular axis.

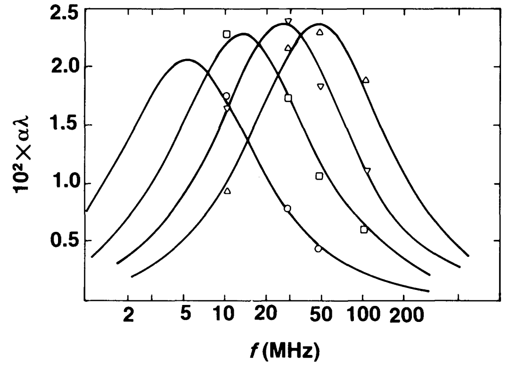

FIG. 7. Ultrasonic absorption due to rotational isomerism in 1,1,2-trichloroethane. Temperatures of the measurements are 0 (○), 20 (□), 40 (▽), and 60 °C (△). From Lamb (1965).

trichloroethane and a few other thermally relaxing liquids are shown in Table 2.

Various esters, aldehydes, and ketones also exhibit strong relaxation effects that can be attributed to hindered rotation about C–O or C–C bonds. Triethylamine and substituted cyclohexanes, which also have different isomeric forms, are other examples of this type of thermal relaxation.

3.4.2 Monomer-Dimer Reaction. Acetic acid shows a remarkably strong thermal relaxation. At 25 °C, the value of α/f^2 rises from 140×10^{-15} $s^2\,m^{-1}$ at frequencies well above the relaxation frequency of 725 kHz to $132\,000\times10^{-15}$ $s^2\,m^{-1}$ at low frequencies. Propionic acid and other carboxylic acids exhibit similar behavior.

Carboxylic acids tend to form dimers in which a molecule's O and OH group form double hydrogen bonds with a neighboring molecule. The equilibrium distribution of the monomer-dimer system is perturbed by the temperature variations of the sound wave. Analysis of the temperature dependence of the peak value of yields a value for the energy difference between the two states.

Table 2. Ultrasonic relaxation parameters for some thermally relaxing liquids.

Liquid	T (°C)	f_r (MHz)	$(\alpha\lambda)_{\max}$
1,1,2-trichloroethane	20	14	0.023
Corotonaldehyde	25	30	0.013
Acetic acid	25	0.725	0.0525
Carbon disulfide	25	78	0.26

3.4.3 Vibrational Relaxation. This process does not necessarily involve organic liquids, but is included here as another example of thermal relaxation. A large number of polyatomic liquids have been studied in which excess sound absorption and associated dispersion exist due to the transfer of energy between the translational and vibrational degrees of freedom of the molecules. For most liquids, the relaxation frequencies occur well above the typical range of measurements. Carbon disulfide is an exception with a relaxation frequency of about 80 MHz at room temperature. Results for CS_2 as well as for carbon dioxide, sulfur hexafluoride, cyclopropane, and a number of others are consistent with a picture of the excess sound absorption being due to the relaxation of the total vibrational specific heat with a single relaxation time. The relationship of the relaxation time to the collision rates in various theories of liquids is given by Herzfeld and Litovitz (1959).

3.5 Binary Mixtures

3.5.1 Nonassociated Liquids. When two thermally relaxing, but nonassociating, liquids are dissolved in one another, the absorption generally does not vary linearly with concentration. As is discussed by Sette (1961), this is due to the difference in the efficiency of collisions of like and unlike molecules in creating the transitions associated with relaxation. Two examples of absorption data and their fit by a theory based on this idea are shown in Fig. 8. A compilation of a large amount of early data is provided by Schaafs (1967).

3.5.2 Critical Phase Separation. Many binary mixtures of liquids display a critical phase separation. They are mutually soluble over only a certain range of concentration and temperature, but, for the critical concentration, a critical temperature exists, called the consolute point or T_c, at which the mixture separates in a second-order phase transition into two solutions with distinctly different concentrations. The order parameter, which is the difference in concentration of the two separated phases, changes continuously from zero at $T=T_c$. The critical behavior is similar to (in the same universality class

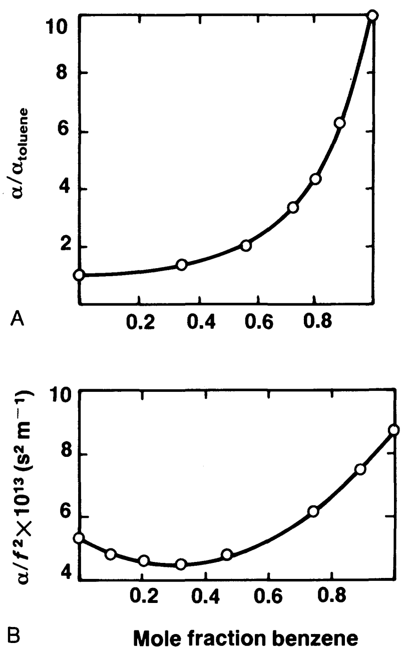

FIG. 8. Comparison of theoretical and experimental values of the absorption coefficient in binary liquid mixtures: (a) toluene-benzene; (b) carbon tetrachloride-benzene at 25 °C. From Sette (1961).

as) the liquid-vapor critical point. A few of the mixtures that exhibit this behavior are listed in Table 3.

As the temperature approaches T_c in a critical binary mixture, sound propagation is marked by velocity dispersion and by a frequency-dependent peak in the absorption. As in other critical phenomena, this is due to the adiabatic temperature changes of the sound wave coupling to the order-parameter (in this case, the concentration) fluctuations. As a case study of this type of behavior, we consider the application of the theory of Ferrell and Bhattacharjee (1985) [FB] to the data of Garland and Sanchez (1983) on the system of 3-methylpentane-nitroethane. The total absorption over a distance of one wavelength $\alpha\lambda$ is shown in Fig. 9 for the critical concentration as a function of temperature. The critical absorption at low MHz frequencies rises to values several times greater than the background absorption. The theory of FB uses dynamic scaling to calculate the dynamic specific heat associated with the dissipative relaxation of the concentration fluctuations. An approximate expression for the predicted temperature dependence of the absorption is given by

$$\left[\frac{\alpha_c(T)\lambda}{\alpha_c(T_c)\lambda}\right]^{1/2} \approx 1+\frac{0.414\Delta T}{\Delta T_{1/2}}, \tag{32}$$

where $\alpha_c(T)$ is the critical part (background subtracted) of the absorption, $\Delta T = T - T_c$, and $\Delta T_{1/2}$ is the value of ΔT for which $\alpha_c(T) = \alpha_c(T_c)/2$. The linear plot, shown in Fig. 10, is consistent with this prediction.

3.5.3 Associated Liquids.

It has been known for a long time, as indicated by Sette (1961), that water-alcohol mixtures have anomalously large attenuation that has frequency-dependent peaks as a function of concentration. Some examples are shown in Fig. 11. Although this behavior is similar to that displayed by critical mixtures, water-alcohol systems do not have a critical phase separation.

The exact source of the absorption in water-alcohol mixtures is still under discussion. Generally the absorption can be fitted by two separate relaxation processes and these have been attributed to reactions involving a liquid clathrate structure. The hydrophobic part of the alcohol molecule induces the water molecules to rearrange their hydrogen bonds to form a cell structure in which an alcohol molecule is trapped. Interaction of these structures or exchange of an alcohol molecule can produce a relaxation. On the other hand, the system water-propanol has been studied over a very large frequency range by Madigosky and Warfield (1987), who find good agree-

Table 3. Some binary mixtures whose critical phase separations have been studied acoustically.

Mixture	Critical concentration (mole fraction of a)	T_c (K)
a. Nitroethane b. Cyclohexane	0.453	296.3
a. Nitrobenzene b. *n*-heptane	0.470	293.7
a. *n*-amyl alcohol b. Nitromethane	0.385	301.1
a. Nitroethane b. 3-methylpentane	0.500	299.6

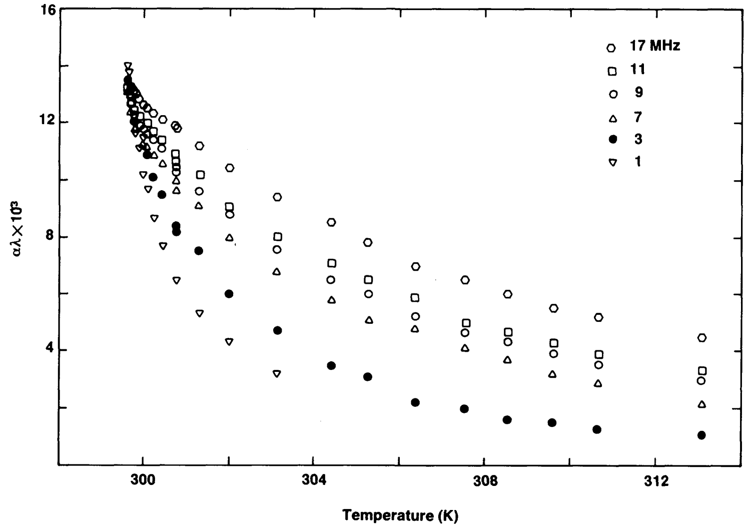

FIG. 9. Temperature and frequency dependence of the absorption per wavelength in a critical mixture of 3-methylpentane and nitroethane. From Garland and Sanchez (1983).

ment with the concentration-fluctuation theory of FB that is used for critical phenomena. They find that the pressure changes, rather than the temperature changes, of the sound wave couple to the concentration fluctuations and that a correlation length of 2.5 nm at room temperature suggests the formation of associated clusters.

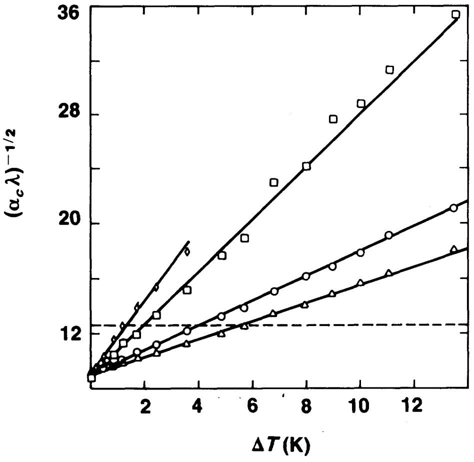

FIG. 10. Critical part of the absorption per wavelength vs ΔT for 3-methylpentane-nitroethane at the critical concentration for several frequencies: 1(△), 3 (○), 9(□), and 17 MHz (◇). The dashed line represents $\alpha_c(T)=\alpha_c(T_c)/2$. From Ferrell and Bhattacharjee (1985).

3.6 Liquid Crystals

The liquid crystalline phases are states of matter intermediate between the crytalline solid and the isotropic liquid that occur for a large number of organic materials. A liquid crystal has lost some or all of the solid crystal's long-range positional order, but it retains

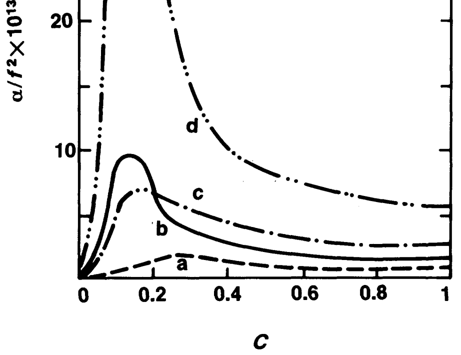

FIG. 11. Absorption as a function of concentration c for several water-alcohol mixtures: (a) ethanol; (b) n-propanol; (c) isopropanol; (d) tert-butanol. From Burton (1948).

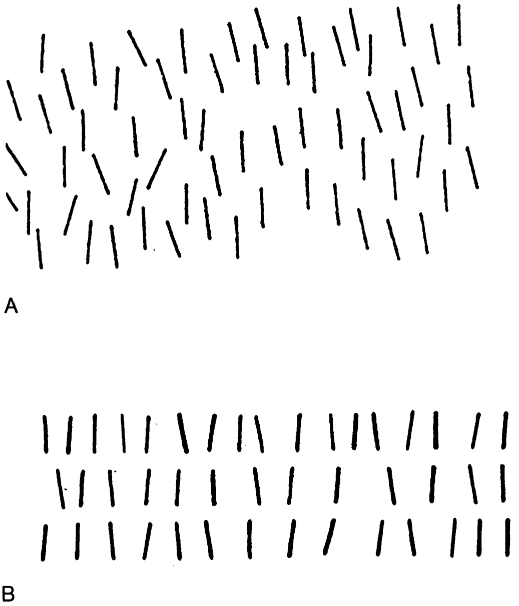

FIG. 12. Molecular arrangement in the (a) nematic phase and (b) smectic-A phase.

the long-range orientational order of its long, thin molecules. It is, thus, a liquid whose physical properties, including its acoustic properties, are anisotropic. Further acoustical interest derives from the behavior at the phase transitions between the liquid crystal states and the neighboring states and from intramolecular thermal relaxations.

For a description of the structure and properties of the variety of liquid crystalline phases that exist, see LIQUID CRYSTALS, STRUCTURE OF. Here we will mention only the acoustical properties of nematic and smectic-A phases. The molecular structures of these two phases are represented in Fig. 12. In the nematic phase, the molecules align parallel to one another, but otherwise have no long-range order. The smectic-A phase is characterized by one-dimensional positional ordering; the molecules lie in planes with the molecular axes aligned parallel to the plane normals.

3.6.1 Intramolecular Relaxations. Liquid crystal molecules generally contain rigid phenyl rings and long flexible end chains. The end groups can exist in various conformations that have different free energies, and thus are subject to rotational isomeric thermal relaxation processes, as discussed in Sec. 3.4.1. An example is the case of *p*-methoxy-*p*′-methylazoxybenzene (MBAB), which has been discussed by Natale (1978) and Candau and Letcher (1978). The frequency dependence of absorption can be fitted by a single relaxation with a relaxation frequency of about 15 MHz. The temperature dependence of the relaxation frequency gives values of the activation enthalpy and entropy. Comparison with pentane, which has the same hydrocarbon chain, shows a comparable activation enthalpy; on the other hand, the activation entropy of MBAB is much larger, which is consistent with the high degree of ordering of the end chains in nematic liquid crystals.

3.6.2 Acoustic Anisotropy. The hydrodynamics of ordered systems, including liquid crystals, has been presented by Martin *et al.* (1972) and reviewed by Miyano and Ketterson (1979). The dissipative part of the stress tensor, Eq. (9), must be expanded to reflect the symmetry. For a uniaxial system, such as nematic and smectic-A liquid crystals, five viscosity coefficients are required. From the revised hydrodynamic equations, one can derive the acoustic wave equation from which the sound absorption is found to have the form

$$\begin{aligned}\frac{\alpha}{\omega^2}\,2\rho_0c^3 &= (\nu_2+\nu_1)\sin^2\theta\\ &\quad+(2\nu_1+\nu_2-\nu_4+2\nu_5)\cos^2\theta\\ &\quad+(4\nu_3-2\nu_1-2\nu_2)\sin^2\theta\,\cos^2\theta, \qquad (33)\end{aligned}$$

where θ is the angle between the direction of propagation of the sound wave and the preferred direction of orientation of the molecules. The coefficients ν_1, ν_2, and ν_3 are shear viscosities and ν_5 and $\nu_4-\nu_2$ are volume viscosities. Thermal conductivity, which is also anisotropic, makes negligible contribution to the absorption.

In the hydrodynamic (low frequency) limit, the sound velocity in nematics is isotropic. In smectic phases, due to the ordering of the molecules in planes, the sound velocity is anisotropic and propagating shear modes can exist.

The anisotropy of absorption in the nematic phase is shown in Fig. 13, which shows the angular dependence of $\alpha(\theta)-\alpha(\pi/2)$. The solid curves are two-parameter fits of the angular dependence of the form of Eq. (33). In

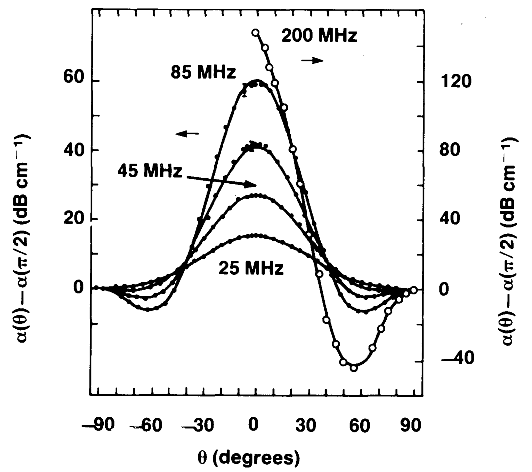

FIG. 13. Frequency dependence of the absorption anisotropy in the nematic phase of MBBA. Data in closed circles are from Wetsel *et al.* (1972); open circles, Bacri (1974). Figure from Miyano and Ketterson (1979).

all nematics studied the absorption anisotropy conforms to the hydrodynamic model, but with viscosity coefficients that can vary dramatically with temperature and frequency.

In the absence of intramolecular relaxations, one would expect that the viscosity coefficients are independent of frequency and then Eq. (33) states that the absorption coefficient in smectics and nematics is proportional to f^2. This is observed in all nematics after relaxation processes have been accounted for. Mazenko *et al.* (1983), however, have shown that, in smectics, anharmonic terms in the fluctuations of the compression and undulation modes of the layers lead to a $1/f$ divergence of some viscosities at low frequencies. This predicts a frequency dependence of sound absorption of the form $\alpha/f^2 = a(\theta) + b(\theta)/f$, where $a(\theta)$ is given by Eq. (33) and $b(\theta)$ contains elastic and viscous constants. Collin *et al.* (1986) have shown that this behavior is observable in the megahertz range only if the elastic constant for planar compression is sufficiently large compared to that for planar bending.

Measurements of sound velocity in the megahertz frequency range in both nematics and smectics show anisotropy, with the velocity for propagation parallel to the preferred molecular orientation greater than for normal propagation direction. This is illustrated in Fig. 14 for cyanobenzylidene octyloxyaniline (CBOOA), which can exist in both the nematic and smectic-A phases. In the low frequency limit, the anisotropy in the nematic phase vanishes and that in the smectic phase remains finite, in accordance with the hydrodynamic theory.

3.6.3 Relaxation Processes at Phase Transitions. Liquid crystals offer a wide array of liquid-liquid phase transitions for study by acoustic means. Here we will mention briefly only the nematic to isotropic liquid (NI) transition.

The NI transition is a weak first-order transition in the sense that the nematic order parameter changes discontinuously at T_{NI}, but the discontinuity is small. As a result of this weak first-order behavior, critical phenomena are still evident, but with a virtual critical temperature less than T_{NI}. The sound absorption has a frequency-dependent peak just below T_{NI} that, at low frequencies, dominates the absorption for five or ten degrees on either side of T_{NI}. The sources for this absorption are, for the most part, analogous to those that occur at the binary liquid phase separation (Sec. 3.5.2). In the isotropic and nematic phases, the sound couples to thermal fluctua-

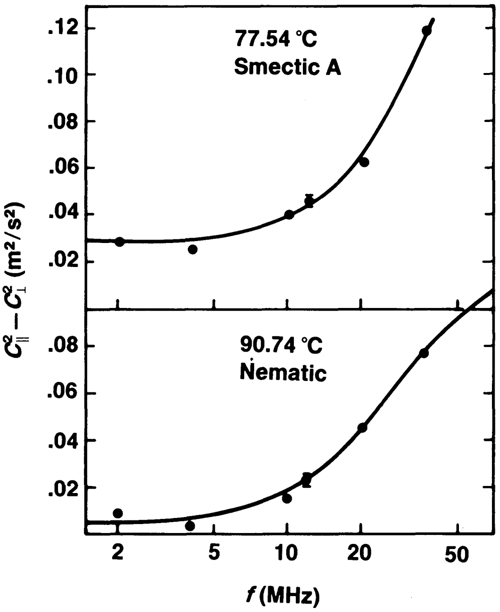

FIG. 14. Frequency dependence of the velocity anisotropy in the smectic-A (top) and nematic (bottom) phases of CBOOA. From Miyano and Ketterson (1975).

tions of the nematic order parameter and, in the nematic phase, the Landau–Khalatnikov mechanism (thermal relaxation of the acoustically perturbed nematic order parameter) is also present. In the nematic phase near the transition temperature, a third absorption mechanism has been identified that is due to fluctuations of the director that specifies the orientation direction.

3.6.4 Acousto-Optic Effect in Liquid Crystals. The combination of acoustic and optic anisotropy of liquid crystals suggests the possibility of a device that would allow the visualization and mapping of ultrasonic fields. This would be particularly useful in the fields of nondestructive evaluation and medical diagnosis. Consider a cell that consists of a thin layer of uniaxially birefringent nematic with the long molecular axes attached normal to the containing glass surfaces. The optic axis is then everywhere normal to the surfaces and, when the material is placed between crossed polaroids, incident light is blocked. If ultrasonic waves are incident on this cell, they exert a torque on the molecules and, at high enough intensity, induce streaming. Each of these effects will tilt the optic axis and allow light to pass through. One can thus map the sound field and, with a little more sophistication, form images of objects irradiated by the sound. Although only intensity, not phase, information is detected, this method has the advantage that a full two-dimensional picture of the sound field is observed in real time. This field was reviewed by Candau and Letcher (1978) and by Kapustina (1984).

3.7 Amphiphilic Aggregates in Solution

Amphiphilic, or surfactant, molecules have a polar head group that is attracted to water and a hydrocarbon tail that is hydrophobic. When dissolved in water, they tend to arrange themselves in structures that minimize the contact of the tails with the water. The large number of parameters that can be varied—including, for example, temperature, concentration of amphiphiles, whether the head group is ionic or nonionic, and concentration of added components, such as alcohol, salt or oil—leads to a rich variety of structures and properties. A full description is beyond the scope of this section; we will mention only a few of the acoustic properties.

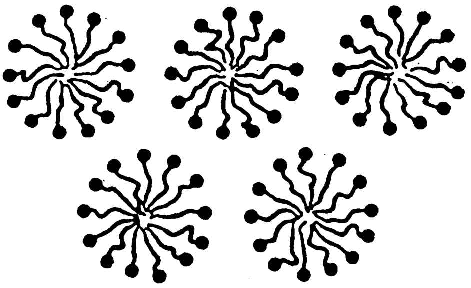

FIG. 15. Molecular structure of micelles.

3.7.1 Micelles. When the concentration of amphiphile molecules in water reaches a value called the critical micelle concentration (cmc), the molecules form themselves into spherical structures, called micelles, with the head groups on the outside, in contact with the water, and the tails on the inside. This is illustrated in Fig. 15. An equilibrium exists between the monomers left in solution and the micelles. When sound propagates through this system, one observes a large frequency-dependent maximum of the absorption as a function of concentration that begins at the cmc, rises to a maximum amplitude at 2–4 times the cmc, and then decreases toward zero at large concentrations. The relaxation frequency depends on the length of alkyl tail; for ultrasonic studies, one chooses an amphiphile with 12 or fewer carbon atoms in the end chain to have the relaxation frequency in the low megahertz region. This relaxation is attributed to the exchange of a surfactant molecule between the micelle structure and the surrounding water-rich environment.

When an alcohol with medium chain length is added to the solution, alcohol molecules are incorporated into the micelle structure and have their own exchange reaction with the environment, which leads to a second relaxation process with a somewhat higher relaxation frequency. These processes have been summarized by Lang *et al.* (1982).

With increasing concentration, eventually the micelles come into contact and interact with one another, leading to anisotropic structures. Common types are the lamellar phase, in which bilayers of surfactant molecules are separated by water layers, and the hexagonal phase, where long, cylindrical micelles are packed in a hexagonal array in a water medium. In this region, as the monomer-micelle exchange reaction disappears, a new relaxation process, with a lower relaxation frequency, is observed in some solutions. This has been attributed by Tiddy *et al.* (1982) to the exchange of water molecules between a "bound" state near the surfactant head groups and a "free" state.

3.7.2 Microemulsions. When oil is added to a water-surfactant-alcohol solution, the water and oil are generally separated by a structure of the surfactant with alcohol as a cosurfactant. For low oil concentrations, the oil may be held inside a spherical micelle that is floating in the water-rich environment. This is an oil-in-water microemulsion. At the other extreme of high oil concentration, one finds water inside an inverted micelle—a water-in-oil microemulsion.

The ultrasonic absorption of a large number of systems of both types has been reported by Lang *et al.* (1982). For surfactant and alcohol molecules with appropriate chain length, both structures are characterized by two relaxation processes in the low megahertz range. For the oil-in-water case, these have been attributed to the same exchange reactions found in micelles (see above). Similar exchanges take place in water-in-oil microemulsions, but the surfactant molecules can transfer to the interior water medium. Although the absorption can go through a maximum as a function of concentration in some microemulsions, no case has yet been found where this is attributable to concentration fluctuations associated with a critical process.

3.7.3 Magnetic Liquids. A magnetic liquid is composed of very fine magnetic particles, each coated with surfactant molecules, that are colloidally dispersed in a fluid. Since it is a fluid that can be controlled, to a certain extent, by a magnetic field, it has many practical uses, for example, as a sealant or lubricant. Magnetic fluids have a wide variety of acoustic properties, but since a large number of parameters are involved that can be varied, but not necessarily controlled, a detailed description that is generally applicable is not easily obtained. In the absence of an external magnetic field, increasing the concentration of particles causes a decrease in the sound velocity and an increase in the attenuation. The frequency dependence of α/f^2 indicates a relaxation process in the low megahertz region. In the presence of an external magnetic field, the absorption and velocity are anisotropic, but the shape of the anisotropy is not that of liquid crystals, given by Eq. (33). The situation is more complicated because the magnetic field causes the individual magnetic particles to agglomerate into large particles whose shape and size depend on the magnetic history.

3.8 Liquid Helium

Helium has two stable isotopes, ^{3}He and ^{4}He, each of which exhibits a superfluid state (see SUPERFLUIDITY: LIQUID HELIUM SYSTEMS). Liquid ^{3}He and ^{4}He behave quite differently and will be discussed separately.

3.8.1 Liquid ^{4}He. Between the normal boiling point of 4.21 K and the superfluid transition temperature of 2.17 K, ^{4}He is a simple, weakly bound van der Waals liquid that could be included in Sec. 3.1 except that its acoustical properties begin to be dominated by the approach to the superfluid transition as the temperature is lowered. Below 2.17 K (called the lambda point or T_λ), the superfluid state is a macroscopic quantum system, but many of its hydrodynamic properties can be described accurately with a two-fluid model. The fluid is assumed to consist of a mixture of two components, normal and superfluid, and the density of the fluid is simply the sum of the normal and superfluid densities: $\rho=\rho_n+\rho_s$. The superfluid component has zero viscosity and carries zero entropy, so that it can flow through narrow cavities (called superleaks) that immobilize the normal component. An increase in the ratio ρ_s/ρ_n corresponds to a decrease in temperature.

A number of propagating hydrodynamic modes exist in the superfluid state. Ordinary sound, in which the normal and superfluid components move together creating density and pressure variations, is called first sound. When the two components move out of phase in the bulk, so that the overall density remains unchanged, the resulting propagating temperature and entropy wave is called second sound. Third sound occurs in thin films of He when the normal component is immobilized by its viscosity and the superfluid component moves, creating a kind of surface wave with variations of the film thickness but not of total density. Fourth sound is a bulk

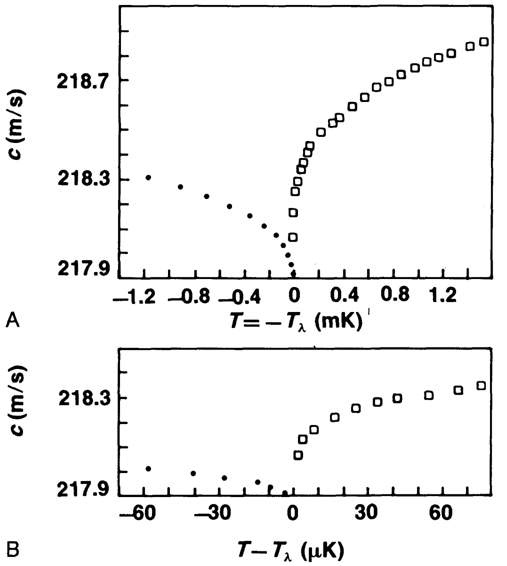

FIG. 16. Temperature dependence of the velocity of first sound at 22 kHz in liquid ^{4}He near the lambda transition on (a) the millikelvin and (b) the microkelvin scales. From Barmatz and Rudnick (1968).

pressure wave in which the normal component is locked by a superleak when, for example, the helium fills the narrow interstices of a packed powder. Finally, fifth sound is a thermal wave in which the normal component is immobilized by a superleak and the superfluid moves, but the pressure is not allowed to change. These modes are summarized by Rudnick (1983) and are discussed in the entry SUPERFLUIDITY: LIQUID HELIUM SYSTEMS. We will concentrate here on the properties of ordinary, or first, sound.

The speed of sound in liquid and solid helium is less than that in air at standard temperature and pressure, which is a manifestation of the loose packing and high compressibility of condensed helium that, in turn, is due to the weakness of the van der Waals force and the relative strength of the quantum mechanical zero-point energy. Near T_λ, the velocity has a sharp decrease associated with the critical nature of the phase transition. This is shown in Fig. 16.

The temperature dependence of the absorption of sound in ^{4}He is represented in Fig. 17. From about 2.5 K up to the boiling point, the absorption is due to the classical causes of viscosity and thermal conduction. A large peak occurs near the lambda point and then a maximum occurs at about 1 K, in the middle of the superfluid state. At low temperatures, the absorption goes to zero as T^4. The low attenuation at very low temperatures has been exploited in a scanning acoustic microscope (see MICROSCOPY, ACOUSTICAL).

The peak of absorption at 1 K is related to the elementary excitations that exist in superfluid ^{4}He. At very low temperatures, the only thermal excitation is the phonon, the quasiparticle associated with ordinary sound; above about 0.6 K, a new excitation, called the roton, can be found. The absorption peak represents a thermal relaxation process in which the sound wave disturbs the equilibrium distribution of phonons and rotons and the return to equilibrium is achieved with time-consuming collision processes.

The lambda transition in liquid ^{4}He is a critical process, as introduced in Sec. 1.4. The order parameter is a complex quantity that is related to the quantum mechanical wave function in the superfluid state. The sound velocity has a marked decrease near T_λ (Fig. 16), and the sound absorption has a dramatic increase (Fig. 17). The absorption for $T > T_\lambda$ is accurately described by the coupling of the sound to fluctuations of the order parameter, similar to critical binary mixture phase separations (Sec. 3.5). The peak in the sound absorption occurs slightly below T_λ, as can be seen in Fig. 18. This is attributed to the Landau–Khalatnikov mechanism in the superfluid state, which is the relaxation of the order parameter that has been perturbed by the temperature variations of the sound wave.

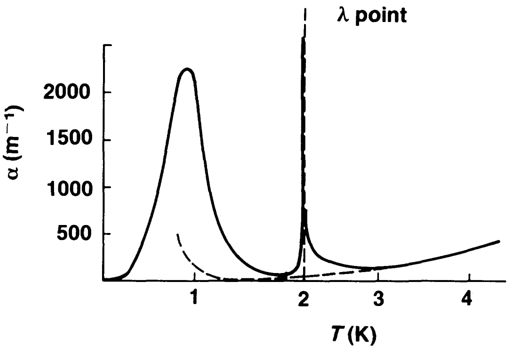

FIG. 17. Attenuation of sound in liquid ^{4}He at 12 MHz. Dashed curve is the classical viscous absorption. From Atkins (1959).

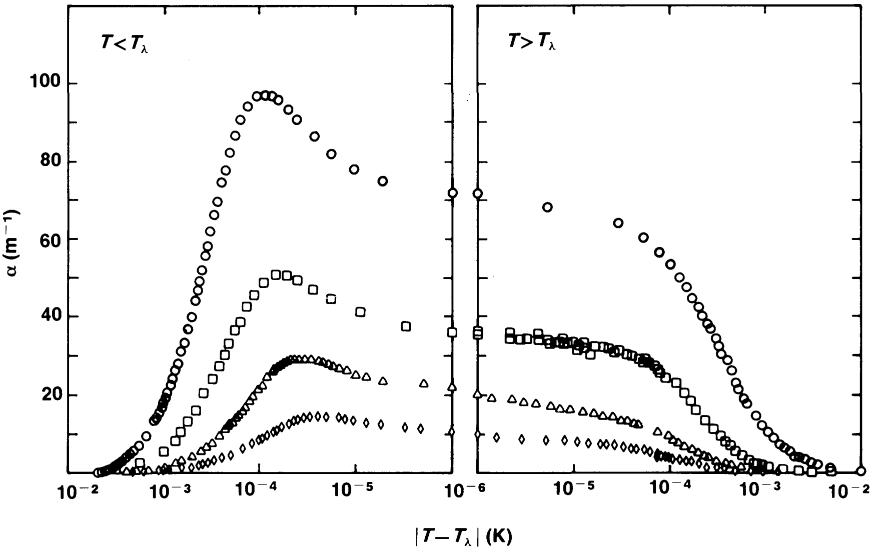

FIG. 18. Temperature and frequency dependence of the absorption of first sound in ^{4}He near the lambda transition. Starting from the lowest curve, the frequencies are 600 kHz, 1 MHz, 1.75 MHz, and 3.17 MHz. The data represent the critical absorption, after the background has been subtracted. From Williams and Rudnick (1970).

3.8.2 Liquid ^{3}He. Ultrasonic spectroscopy has been a valuable tool in un der standing the remarkable properties of liquid ^{3}He. At temperatures above 1 K, ^{3}He behaves like a simple liquid and its sound absorption is due to viscosity, which has a temperature dependence of T^{-2}, and is proportional to the square of the frequency. Helium-3 is, however, a Fermi liquid and at sufficiently low temperatures these properties begin to manifest themselves. The atoms fill discrete energy levels up to the Fermi energy and only those atoms with momentum near the corresponding Fermi surface participate in collisions. In this regime, the time between collisions τ varies rapidly with temperature according to T^{-2}. For high ultrasonic frequencies, a point is reached where $\omega\tau > 1$; the time between collisions is greater than the period of the sound wave. This is called zero sound. It is the Fermi liquid analog of sound propagation in rarified gases where the molecular mean free path is greater than the wavelength. The absorption of zero sound is proportional to T^2. The zero-sound regime occurs below about 20 mK and the first, or ordinary, sound, above 20 mK.

Below about 2 mK (depending on the pressure), liquid helium-3 enters one or another of three superfluid states: ^{3}He-A, ^{3}He-B, or, in the presence of a magnetic field, ^{3}He-A_1. The helium atoms form Cooper pairs, in exact analogy to the electron pairs associated with superconductivity. The pairs form a reduced energy state that creates an energy gap that broadens as the temperature is reduced. The superfluid states can be described by a two-fluid model that predicts second through fifth sound, but these have been difficult to observe in ^{3}He because of large damping.

As an example of the behavior of sound absorption in superfluid ^{3}He, we mention briefly results in ^{3}He-B; the reader is referred to Wilkes and Betts (1987) and references therein and to the article SUPERFLUIDITY: LIQUID HELIUM SYSTEMS for further discussion. The absorption of 60-MHz ultrasound in ^{3}He-B is shown in Fig. 19. The large peak just below the superfluid transition is due to acoustic energy being used to break up Cooper pairs. The other large peak and the smaller, sharper peak are called the squashing and real squashing modes and refer to the distortion by the sound wave of the symmetry of the order parameter that describes the state.

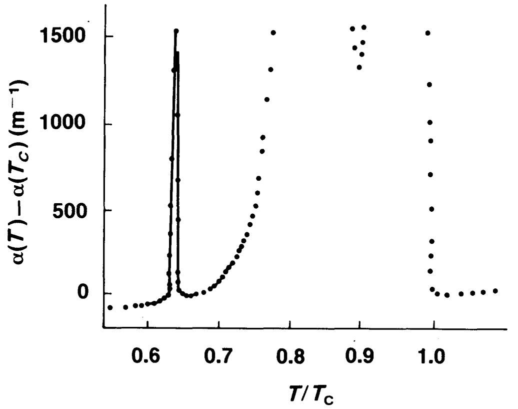

FIG. 19. Relative absorption of 60-MHz sound in the superfluid state of ^{3}He-B at 5.3 bar. The higher-temperature broad peak is due to the breaking of Cooper pairs. The other two peaks are due to the squashing and real squashing modes. From Lee (1983).

GLOSSARY

Associated Liquids: Materials such as water and the alcohols that tend to form molecular clusters or groups in the liquid state. They have a contribution to the sound absorption due to structural relaxation.

Classical Sound Absorption: The absorption of sound in a fluid that is due to shear viscosity and heat conduction.

Fermi Liquid: A liquid, such as ^{3}He, that obeys Fermi–Dirac statistics. See article on STATISTICAL MECHANICS, QUANTUM.

Irreversible Thermodynamics: Thermodynamics of systems that are not maintained in thermal equilibrium. The extent to which a system departs from equilibrium is treated as a new thermodynamic variable. See THERMODYNAMICS, NONEQUILIBRIUM.

Kneser Liquids: Liquids that undergo thermal relaxation processes of the same type that are found in gases.

Structural Relaxation: A relaxation process between different short-range structural configurations of a liquid that is induced by the pressure variations of a sound wave.

Thermal Relaxation: A relaxation process between different molecular energy states that is induced by the temperature variations of a sound wave.

List of Works Cited

Atkins, K. R. (1959), *Liquid Helium*, London: Cambridge University.

Bacri, J. C. (1974), *J. Phys. (Paris) Lett.* **35**, L141–L142.

Barmatz, M., Rudnick, I. (1968), *Phys. Rev.* **170**, 224–238.

Beyer, R. T., Letcher, S. V. (1969), *Physical Ultrasonics*, New York: Academic.

Beyer, R. T., Ring, E. M. (1972), in S. Z. Beer (Ed.), *Liquid Metals: Chemistry and Physics*, New York: Marcel Dekker.

Burton, C. J. (1948), *J. Acoust. Soc. Am.* **20**, 186–199.

Candau, S., Letcher, S. (1978), *Adv. Liq. Crystals* **3**, 167–235.

Collin, D., Gallani, J. L., Martinoty, P. (1986), *Phys. Rev. A* **34**, 2255–2264.

Davis, C. M., Jarzysnki, J. (1972), in: F. Franks (Ed.), *Water: A Comprehensive Treatise*, Vol. 1. New York: Plenum.

Ferrell, R., Bhattacharjee, J. (1985), *Phys. Rev. A* **31**, 1788–1809.

Fisher, F. H., Simmons, V. P. (1977), *J. Acoust. Soc. Am.* **62**, 558–564.

Garland, C. W. (1970), in: W. P. Mason, R. N. Thurston (Eds.), *Physical Acoustics VII*, New York: Academic.

Garland, C., Sanchez, G. (1983), *J. Chem. Phys.* **79**, 3090–3098.

Hall, L. (1948), *Phys. Rev.* **73**, 775–781.

Herzfeld, K. F., Litovitz, T. A. (1959), *Absorption and Dispersion of Ultrasonic Waves*, New York: Academic.

Kapustina, O. A. (1984), *Mol. Cryst. Liq. Cryst.* **112**, 1–164.

Kneser, H. O. (1938), *Ann. Physik* **32**, 277–285.

Lamb, J. (1965), in: W. P. Mason (Ed.), *Physical Acoustics IIA*, New York: Academic.

Lang, J., Djavanbakht, A., Zana, R. (1982), in: I. D. Robb (Ed.), *Microemulsions*, New York: Plenum.

Lee, D. M. (1983), *Int. J. Quantum Chem.* **23**, 1191–1199.

Madigosky, W., Warfield, R. (1987), *J. Chem. Phys.* **86**, 1491–1496.

Mallo, P., Waton, G., Candau, S. (1985), *Rev. Phys. Appl.* **20**, 403–408.

Martin, P. C., Parodi, O., Pershan, P. (1972), *Phys. Rev. A* **6**, 2401–2420.

Mazenko, G., Ramaswamy, S., Toner, J. (1983), *Phys. Rev. A* **28**, 1618–1636.

Mellen, R., Scheifele, P., Browning, D. (1987), *Global Model for Sound Absorption in Sea Water*, New London: Naval Underwater Systems Center.

Miyano, K., Ketterson, J. (1975), *Phys. Rev. A* **12**, 615–635.

Miyano, K., Ketterson, J. (1979), in: W. P. Mason, R. N. Thurston (Eds.), *Physical Acoustics XIV*, New York: Academic.

Natale, G. (1978), *J. Acoust. Soc. Am.* **63**, 1265–1278.

Rudnick, I. (1983), in: J. G. M. Armitage (Ed.), *75th Jubilee Conference on Helium-4*, St. Andrews: World Scientific.

Schaafs, W. (1967), Molecular Acoustics, in: K. Hellwege, A. Hellwege (Eds.), *Landolt-Bornstein*, Group II, Vol. 5, Berlin: Springer.

Sette, D. (1961), in: S. Flugge (Ed.), *Handbuch der Physik*, Vol. XI/1. Berlin: Springer.

Steuhr, J., Yeager, E. (1965), in: W. P. Mason (Ed.), *Physical Acoustics IIA*, New York: Academic.

Tanaka, H., Wada, Y. (1985), *Phys. Rev. A* **32**, 512–524.

Tiddy, G., Walsh, M., Wyn-Jones, E. (1982), *J. Chem. Soc. Faraday Trans. 1* **78**, 389–401.

Waton, G., Candau, S. (1981), *Rev. Cethedec.* **67**, 83–95.

Wetsel, G., Speer, R., Lowry, B., Woodard, M. (1972), *J. Appl. Phys* **43**, 1495–1497.

Williams, R. D., Rudnick, I. (1970), *Phys. Rev. Lett.* **25**, 276–280.

Further Reading

Mason, W. P., Thurston, R. N. (Eds.) (1964–1982), *Physical Acoustics*, New York: Academic. Volumes IA, IIA, IVB, VI, VII, XI, XII, XIV, XVI contain relevant material.

ACOUSTIC PROPERTIES OF SOLIDS

See SOLIDS, ACOUSTIC PROPERTIES

ACOUSTIC SENSORS

See SENSORS, ACOUSTIC

ACOUSTICAL HOLOGRAPHY

See HOLOGRAPHY, ACOUSTICAL

ACOUSTICAL INSTRUMENTATION

ILENE BUSCH-VISHNIAC, *Department of Mechanical Engineering, University of Texas, Austin*

ELMER HIXSON, *Department of Electrical and Computer Engineering, University of Texas, Austin*

	Introduction	63
1.	**Physical and Technical Principles**	64
1.1	Conversion of Energy	64
1.1.1	Capacitive Transducers	64
1.1.1.1	Quadratic Types	64
1.1.1.2	Piezoelectric	65
1.1.2	Inductive	66
1.1.2.1	Quadratic Types	66
1.1.2.2	Linear Types	67
1.1.3	Resistive	67
1.2	Intensity Measurement	68
1.3	Directivity	68
1.4	Radiation Efficiency	71
2.	**Function and Operation**	71
2.1	Condenser Microphones	71
2.2	Piezoelectric and Ceramic Transducers for Liquids	74
2.2.1	Hydrophones	74
2.2.2	Sound Sources	75
2.3	Loudspeakers	76
2.4	Geophones	77
2.5	Calibration	78
3.	**New Technologies**	78
3.1	Fiber Optic Transducers	78
3.2	Solid-State Sensors and Actuators	80
3.3	New Materials	81
3.3.1	Polymers	81
3.3.2	Magnetostrictive Materials	82
3.4	Novel Techniques	82
4.	**Applications**	84
4.1	Gases	84
4.2	Liquids	85
4.3	Solids	86
	Glossary	87
	Works Cited	88
	Further Reading	88

INTRODUCTION

Acoustical instrumentation is used to generate and monitor compressive waves in gases, liquids, and solids. In gases, acoustical instruments are used in communication systems such as telephones, public announcement systems, and recreational audio systems. They are also commonly employed in noise control of industrial or community sources and in research on the properties of gases. In liquids, acoustical instruments have been widely used for sonar systems. They are also receiving progressively more attention in biotechnology, where they aid in sonograms and lithotripsy, and in geophysics, where they are useful for off-shore prospecting. In solids, acoustical instruments are being used for nondestructive testing of structures, for geophysical prospecting in the earth, and for surface acoustic wave devices useful in a variety of signal processing schemes. With this wide range of applications and frequency regimes acoustical instrumentation has found a sizable market.

In this article we will describe the unique aspects of acoustical instruments, concentrating on the sensors and actuators which convert acoustical energy to another form (generally electrical) or vice versa. We discuss the physical and technical principles of energy conversion in acoustical transducers and the measurable field-related properties such as directivity, intensity, and radiation efficiency. We describe the practical manifestations of these physical principles in various transducers. Specific examples mentioned include condenser microphones, piezoelectric and ceramic hydrophones, electrodynamic loudspeakers, and linear magnetic geophones. We also discuss new technologies for acoustical instruments, including fiber optic devices, solid-state sensors and actuators, transducers using new materials, and novel transduction

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

techniques. Finally, we describe the practical use of acoustical instruments in various common applications.

1. PHYSICAL AND TECHNICAL PRINCIPLES

The main feature of all acoustical sensors and actuators is the conversion of energy from acoustical form to another form, usually electrical, or vice versa. Below we describe the various mechanisms used in achieving this energy conversion. All of the mechanisms discussed may be used for applications spanning a wide range of frequencies. We also discuss sound intensity, directivity, and radiation efficiency.

An acoustic wave is associated with a pressure fluctuation about the ambient pressure. A pressure gradient produces a particle velocity in the direction of the gradient. Thus, pressure and particle velocity are the basic quantities to be measured. The product of these two is the intensity, a vector which gives the direction and magnitude of power flow per unit area of a propagating acoustic wave. The square of the pressure is related to the potential energy density and the square of the velocity is related to the kinetic energy density in the wave. In a reverberant field resulting from the random summation of many waves, the sum of the potential and kinetic energy densities is the total energy density in the sound field.

1.1 Conversion of Energy

Acoustic pressures can be sensed or generated effectively and accurately over a wide frequency range. Measurement (or generation) of the pressure at two spatial positions allows the determination (or specification) of the pressure gradient and hence particle velocity. All other quantities describing the sound field can be derived from the pressure and velocity.

Pressure transducers use a diaphragm or another surface that acts to convert pressure to force (sensor) or force to pressure (actuator). It is this moving element that results in the conversion of energy from acoustical form to electrical form or vice versa. The conversion process may involve electric fields with capacitive characteristics, magnetic fields with inductive characteristics, or changes in conductance with resistive characteristics.

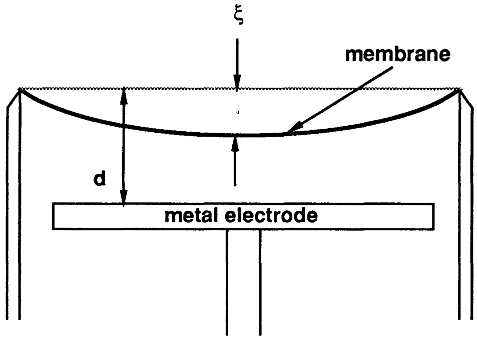

FIG. 1. Typical quadratic capacitive sensor using a membrane as the active transducing element.

1.1.1 Capacitive Transducers. There are two general classes of capacitive transducers: those which are quadratic and those which are linear. Both are described below.

1.1.1.1 Quadratic Types. In quadratic capacitive transducers energy conversion is accomplished by means of an electromechanical coupling which causes or results from a change in the energy stored in an electric field. In acoustical instruments this typically takes the following form. A very thin circular metal or plastic membrane is used to sense or generate an acoustic pressure. In the sensor the pressure acts as a uniformly distributed force to produce a parabolic displacement ξ of this membrane, as shown in Fig. 1. The average displacement in meters is then given by (Merhaut, 1981)

$$\xi=\frac{a^2}{8T}p, \tag{1}$$

where a is the diaphragm radius, T is the tension on the membrane (in units of force/length), and p is the acoustic pressure in Pa.

The instantaneous energy stored in the electric field is described by

$$W=\frac{Q_0^2}{2C}=\frac{Q_0^2(d+\xi)}{2\epsilon A}, \tag{2}$$

where Q_0 is the static charge which is assumed to be constant, C is the capacitance of the device, ϵ is the electric permittivity of the dielectric between the capacitor plates, A is the membrane area, and d is the distance between the two electrodes. The force on the membrane is given by

$$F=\frac{\partial W}{\partial \xi}=\frac{Q_0^2}{2\epsilon A}, \tag{3}$$

and the electric potential is given by

$$V=\frac{\partial W}{\partial Q_0}=\frac{Q_0(d+\xi)}{\epsilon A}. \tag{4}$$

Note that the force is quadratic in charge and thus depends quadratically on the electric field strength. In Eq. (4) we identify the time varying portion of the potential as v, and find through use of Eq. (1) that

$$v=\frac{a^2 V_0}{8Td}\, p, \tag{5}$$

where $V_0 = Q_0/C_0$, C_0 being the static capacitance. This linear relation holds for uniform electric fields and deflections ξ small compared to d so that Q_0 is essentially constant.

1.1.1.2 Piezoelectric. There is a class of dielectric anisotropic crystals that provide coupling between elastic and dielectric phenomena (Mason, 1964). This coupling is referred to as the piezoelectric effect. Natural and artificially grown piezoelectric crystals exhibit linear electromechanical coupling relationships. Certain rare-earth compounds when used to produce a ceramic yield polycrystalline ferroelectrics which have square-law electromechanical relationships. When a high potential is applied to the material at a temperature above the Curie point, a remnant polarization is retained. For small strains and potentials small compared to the remnant polarization, essentially linear piezoelectric coupling is obtained for these materials as well. See Kino (1987) for properties of several materials.

Since piezoelectric crystals are anisotropic elastic solids, the stress-strain relationship must be represented by matrix equations with a 6×6 array of elastic terms and a 3×6 array of electromechanical coupling terms. The electric field relations are a 3×3 array. Fortunately many of the off-diagonal terms of these arrays are zero and some useful configurations allow a one-dimensional representation.

For illustration, a thin plate used as a low-frequency pressure sensor is shown in Fig. 2. The pressure produces a force on the 3 surface. (The 3 direction is always defined as the principal direction in which physical properties are least like those in the other principal directions.) The top and bottom 3 surfaces are covered with a thin conducting material so that the electric field in the 3 direction can be sensed. The bottom surface is considered clamped. The equations of state for the piezoelectric plate simplify to the following:

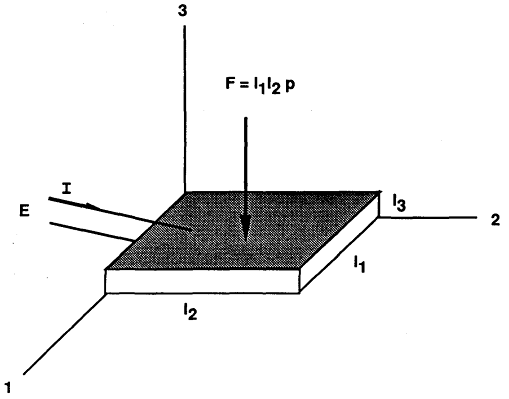

FIG. 2. Example of a piezoelectric pressure sensor. A force in the 3 direction causes the establishment of an electric field in that direction, which varies as the plate thickness changes in response to the pressure.

$$\begin{aligned} D_3&=\epsilon_{33}^{S}E_3+e_{33}S_3,\\ -\sigma_3&=e_{33}E_3-c_{33}^{E}S_3, \end{aligned} \tag{6}$$

where D is the electric flux density, σ is the stress, S is strain, ϵ_{33}^{S} is the clamped permittivity in the 3 direction, c_{33}^{E} is the short-circuit elastic modulus in the 3 direction, and e_{33} is the short-circuit charge density per unit strain or the clamped strain per unit potential gradient, all in the 3 direction. Note that Eq. (6) indicates that there is a linear coupling between mechanical and electrical fields, and that it is described by three matrices of constants: one for purely mechanical coupling, one for purely electrical coupling, and one for electromechanical coupling.

As an example, Fig. 3 shows a standard representation of the constants describing barium titanate ($BaTiO_3$) where dots indicate nonzero values for material constants and lines connect elements with equal values. Note that barium titanate has two useful piezoelectric couplings: the 3–3 or thickness mode, and the 1–5 and 2–4 or shear modes. In addition, through the presence of the e_{13} and e_{23}, Fig. 3 shows that the thickness piezoelectric mode is coupled to piezoelectric behavior (of a smaller magnitude) in the other principal directions.

When the mass of the plate $m_c = \rho l_1 l_2 l_3$ is included the two-port terminal equations become

$$I = Y_e V + \alpha U,$$

$$-F = \alpha V - Z_m U, \tag{7}$$

where I is the current, U is the velocity of one electrode relative to the other, F is the force in the 3 direction at the surface of the piezoelectric, Y_e is the electrical admittance $(j\omega \times \epsilon_{33}^{s} l_1 l_2 l_3)$, Z_m is the mechanical impedance $(j\omega m/3 + l_1 l_2 / j\omega_{33}^{e} l_3)$, and $\alpha = e_{33} l_1 l_2 / l_3$. The open-circuit potential then becomes

$$V = \frac{l_3}{e_{33}(\epsilon_{33}^{s} l_1^2 l_2^2 / \alpha^2 l_3^2 c_{33}^{E} + 1)} p \tag{8}$$

for low frequencies. Thus the output potential is linearly related to pressure.

1.1.2 Inductive. There are also two general classes of inductive acoustic transducers: those which are quadratic and those which are linear. Both are described below.

1.1.2.1 Quadratic Types. In quadratic inductive transducers the electromechanical coupling is produced by a change in the energy stored in a magnetic field. These devices are the duals of the capacitive instruments based on changes in the energy stored in an electric field.

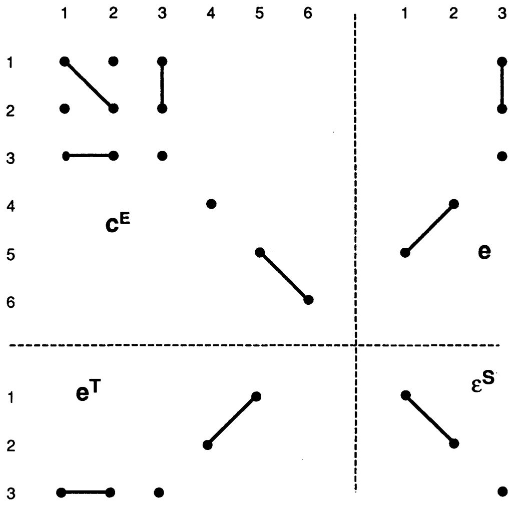

FIG. 3. Graphical representation of the physical parameters of barium titanate, a common piezoelectric material.

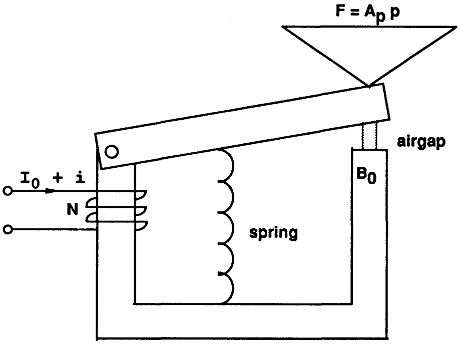

FIG. 4. Typical quadratic inductive sensor using a changing air gap thickness to generate an output signal.

Figure 4 shows a typical example of a quadratic inductive acoustic sensor. An acoustic pressure acts on a piston to change the length of the air gap in a magnetic circuit made up of material whose permeability μ is very much larger than that of air, μ_0. The air gap dimensional change is given by

$$\xi = p A_p C_m, \tag{9}$$

where A_p is the piston area and C_m is the mechanical compliance of the restoring spring.

The instantaneous energy stored in the magnetic field is

$$W = \frac{\lambda_0^2}{2L} = \frac{\lambda_0^2 (d + \xi)}{2N^2 \mu A}, \tag{10}$$

where λ_0 is the flux linkage of the device which is assumed to be constant, L is the inductance, d is the static air gap length, N is the number of turns of the solenoid, and A is the cross-sectional area of the gap. The force on the piston is given by

$$F = \frac{\partial W}{\partial \xi} = \frac{\lambda_0^2}{2N^2 \mu A} \tag{11}$$

and the current is given by

$$I = \frac{\partial W}{\partial \lambda_0} = \frac{\lambda_0 (d + \xi)}{N^2 \mu A}. \tag{12}$$

We note that the force is quadratic in λ_0 and thus depends quadratically on the magnetic field strength. In Eq. (12) we identify the time varying portion of the current as i, and find using Eq. (9) that

$$i=\frac{\lambda_0 A_p C_m}{N^2\mu A}p. \tag{13}$$

Thus a linear relationship results.

Another quadratic phenomenon is illustrated in Fig. 5. An acoustic pressure produces a force on the end of a bar of magnetostrictive material in which the permeability μ changes with density. With a polarizing current I_0 a magnetic flux is produced in the rod. The force on the end of the rod is found from Eq. (10) by recognizing that μ is now a function of ξ, and then transforming from displacement to density using conservation of mass. The result is

$$F=\frac{\lambda_0^2\rho}{2N^2\mu^2A}\frac{\partial\mu}{\partial\rho}. \tag{14}$$

The time varying portion of the current may be similarly found using Eq. (9):

$$i=-\rho\frac{\partial\mu}{\partial\rho}\frac{\lambda_0}{\mu^2}\frac{A_pC_m}{N^2A}p. \tag{15}$$

Again a linear pressure sensor is produced.

1.1.2.2 Linear Types. In linear types of inductive acoustic transducers the electromechanical coupling is produced by the interactions of magnetic fields. When a conductor of length l is placed in a magnetic field B the following coupling relations hold: There is a force on the conductor of $\mathbf{F}=(\mathbf{I}\times\mathbf{B})l$; if the conductor moves, a potential $\mathbf{V}=(\mathbf{U}\times\mathbf{B})l$ results.

When a constant flux density B_0 is used and the current, motion, and B_0 are placed orthogonal to maximize the cross products, the result is as follows:

clamped force, $F_c=B_0lI$, (16a)

open-circuit potential, $V_{oc}=B_0lU$. (16b)

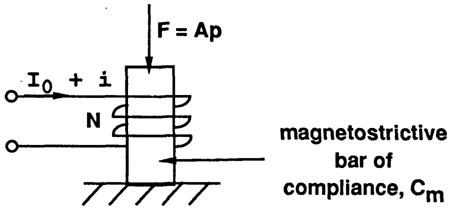

FIG. 5. Example of a magnetostrictive transducer. A force on the end of the magnetostrictive bar causes a change in the bar magnet's permeability and produces a time varying output current.

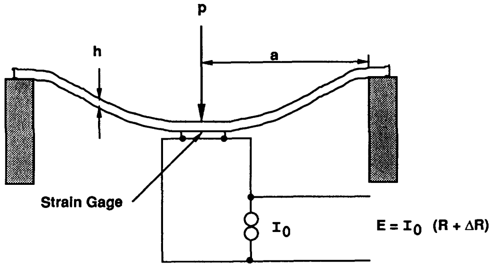

FIG. 6. Example of a resistive acoustic sensor. In this device the pressure strains a membrane which incorporates a strain gage. Hence the pressure causes a change in resistance which is typically monitored as a current change.

If a force is applied to this conductor through a piston exposed to an acoustic pressure, the mechanical parts being retained by a spring and the conductor terminated in an electrical resistance, the following two-port equations result:

$$-V=-IR=Z_eI-B_0lU,$$

$$F=pA_p=B_0lI+Z_mU. \tag{17}$$

Here Z_e is the electrical impedance of the conductor in the magnetic field and Z_m is the mechanical impedance of the moving mechanical structure. Note that these equations imply that there are linear relations between all four port variables: F, U, I, and V. The potential at the electrical terminals is

$$V=\frac{B_0lRA_p}{Z_e\,Z_m+(B_0l)^2+Z_mR}p. \tag{18}$$

Thus a linear pressure sensor results.

1.1.3 Resistive. Resistive acoustic sensors transform pressure to strain in a diaphragm or thin beam. The strain then acts on some material that changes resistance with strain causing a change in either a current or a potential. There are two mechanisms of resistance change which result from imposition of a strain on a material: an increase in resistance due to elongation of the material, and a piezoresistive effect in which the strain affects the crystal lattice of a material and alters its resistivity. The relative magnitude of the piezoresistive effect (compared to the dimensional effect), and its sign, are highly dependent upon the material.

In Fig. 6 acoustic pressure acts on a clamped-edge circular plate to deform it and produce a strain on the surface. That strain on the surface at the center is

$$S=\frac{a^2h/2}{16D_{\text{mod}}}\,p, \tag{19}$$

where D_{mod} is the bending modulus [$D_{\text{mod}}=Y_{\text{mod}}h^3/12(1-\nu^2)$ where Y_{mod} is Young's modulus, h the plate thickness, and ν Poisson's ratio].

The strain gage produces a change of resistance $\Delta R/R=KS$, where K is defined as the gage factor. Hence

$$\Delta R/R=\frac{Ka^2h/2}{16D}\,p. \tag{20}$$

Through a constant-current source a change in potential is linearly related to the pressure. For metals, the dimensional resistance change mechanism dominates and the gage factor is about 1.6. For silicon, the piezoresistive mechanism is dominant and the gage factor varies from 100 to 200.

1.2 Intensity Measurement

Over the last decade there has been increasing emphasis on measurement of intensity rather than sound pressure. There are two primary reasons for this shift in measurement approach: sound intensity measurement permits location of acoustic sources in a volume because intensity is a vector quantity, and intensity measurements can be made in the near field, eliminating the need for specialized testing chambers. As a result there are now intensity measuring devices commercially available. For further information on intensity measurements in acoustics see Schultz *et al.* (1975).

Most of the intensity measuring instruments currently on the market use a two-microphone technique to determine the intensity. The acoustic pressure is determined directly from microphones (or hydrophones). The particle velocity along the line on which the two microphones are located is found by differencing the two microphone signals. The linear inviscid force equation generally used for small-amplitude acoustic processes relates the particle acceleration to this pressure difference:

$$\frac{\partial U}{\partial t}=-\frac{1}{\rho}\frac{\partial p}{\partial t}. \tag{21}$$

The use of two-microphone intensity measurements is limited at the low-frequency end by the need to have the phase change in the signal exceed the sensor-pair phase mismatch. The typical phase mismatch in conventional intensity measuring microphones is ±0.005 rad. In order to have an error of less than 1 dB the signal phase must exceed this mismatch by at least a factor of 5. This defines a minimum separation between the two transducers which scales inversely with frequency. At 20 Hz the minimum separation is 6.8×10^{-2} m.

At the high-frequency end intensity measurements are limited by the need for the wavelength to exceed the separation between the microphones. The maximum separation distance at 10 kHz for a 1-dB error is about 6×10^{-3} m. If conventional microphones (1.25×10^{-2}-m diameter) are placed closer than about 5×10^{-3} m, then diffraction of the sound field by the microphones becomes a problem. Hence the upper limit for intensity measurements using conventional microphones is about 10 kHz. One can go to higher frequencies using smaller microphones or microphones with small-aperture probe tubes.

In addition to the conventional two-microphone intensity measuring instruments there are efforts under way to build surface-intensity and energy-density measuring instruments. Surface-intensity acoustical instruments determine the acoustic intensity in a liquid or gaseous medium which is attributable to a vibrating structure in the medium. They typically use a single microphone (or hydrophone) and a single accelerometer attached to a vibrating body, although attempts have been made to use noncontacting displacement measuring devices instead of accelerometers. Energy-density measuring devices are similar to conventional acoustic intensity measuring instruments but use four microphones to find the acoustic pressure and the particle velocity in three orthogonal directions simultaneously.

1.3 Directivity

Most real acoustical instruments exhibit spatially nonuniform response characteristics. For sound generators the nonuniformity is comprised of two parts: range dependence and angular dependence. For sound detectors

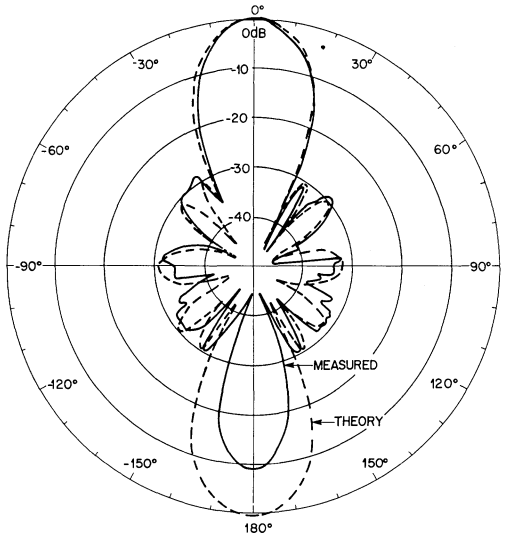

FIG. 7. Directivity of a typical linear microphone array: DF = 16 and DI = 12 dB.

it is the angular dependence alone which is significant. Since the range dependence of generators is determined by the medium, rather than the instrument, we concentrate here on angular variations.

The most commonly used measures of angular variation of acoustical transducers are the beam width, directivity factor, and directivity index. The beam width is typically experimentally determined and is the angular range embracing the peak over which the response is within a specified amount (usually a factor of 0.5) of the peak value. Here the term "response" refers to a quantity, such as sound intensity or mean squared pressure, which is proportional to energy or power. For example, for the transducer array whose angular response is shown in Fig. 7, the beam width is 16°.The directivity factor (DF) is the ratio of the peak response at radius r to the response averaged over a sphere of radius r. The directivity index is a logarithmic measure of the directivity factor: $\mathrm{DI}=10 \log_{10} (\mathrm{DF})$. For the response pattern shown in Fig. 7, DF = 16 and DI = 12 dB.

There are two general causes of directive response: the geometry of individual transducers and the result of summing multiple transducers in arrays. First consider the behavior of a single acoustical transducer. If the transducer is large compared to the wavelength of the sound it is not well approximated by a point receiver or radiator in space and it will generally exhibit directive response. The precise form of the directivity depends upon the exact geometry of the transducer. For example, many directional tranducers expose the energy-converting material on both the front and back faces. The result is an instrument which either responds to pressure differences or radiates out-of-phase sound pressures which add in the medium. Such transducers exhibit a far-field directivity factor which is generally described by

$$\mathrm{DF}(\theta)=\frac{1+\beta\cos\theta}{1+\beta}, \tag{22}$$

where β is defined by

$$\beta=\Delta l/CR\,. \tag{23}$$

Here Δl is the distance between the front and rear sound ports, C is the compliance of the system which is generally determined by the air in a cavity behind the transducing material, and R is the excess acoustic resistance of the rear sound port compared to the front port.

For conventional pressure elements, there is no rear port and so $\beta=0$ and DF = 1, i.e., they are nondirectional. For first-order gradient elements, there is negligible resistance in the rear port, so β is very large and $\mathrm{DF}(\theta)=\cos(\theta)$. Cardioid elements combine pressure and first-order gradient behavior by designing the rear sound port, to have $\beta=1$, in which case $\mathrm{DF}(\theta)=[1+\cos(\theta)]/2$. The three directivity patterns produced by these elements are shown in Fig. 8. Note that these patterns have broad main lobes with no side lobes, and are not frequency dependent.

Another example of this class of directive transducer is a small transducer surrounded by a parabolic reflector. This produces unidirectional sound radiation or reception if the transducer is located at the focus. For further information on directive acoustical transducers see Rossi (1988) and Beranek (1954,1986).

Now consider the effect of summing the output signals from transducers in arrays. For a sensor array with length large compared to a wavelength, there will be a wide variation in the phases and possibly the amplitudes of sound observed by each element due to path differences to the source. For such a sensor array one can write the response O (either a voltage or a current) as

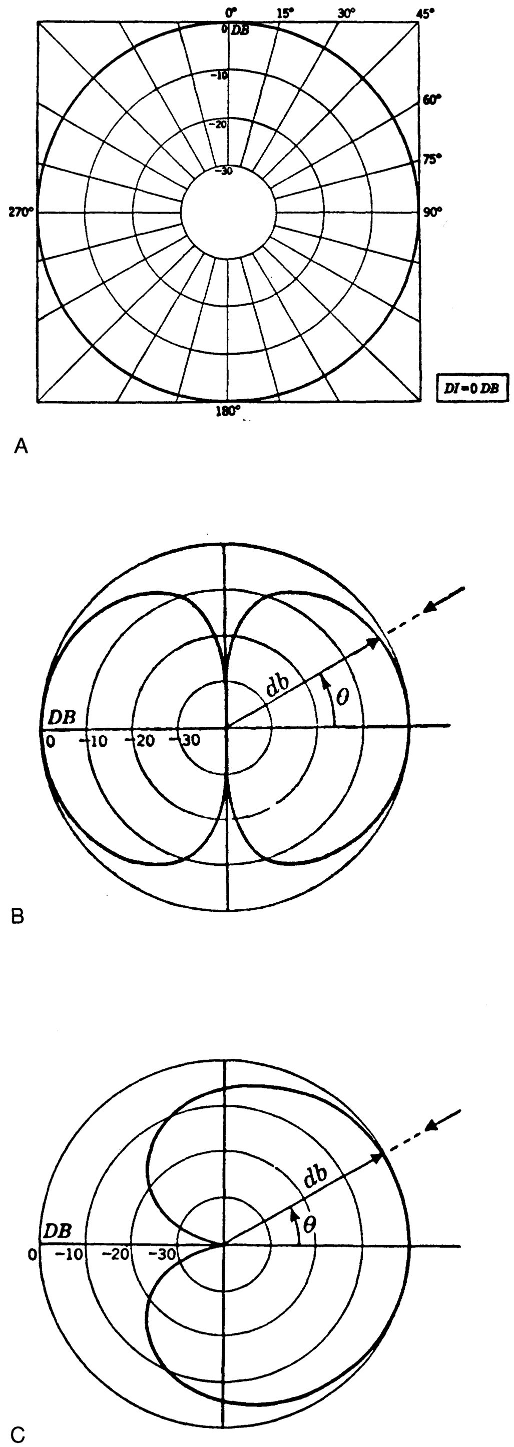

FIG. 8. Directivity patterns of typical microphones: (a) omnidirectional element; (b) first-order gradient element; and (c) cardioid element [from Beranek (1986)].

$$O = \sum \frac{a_i p}{r_i} \exp(jkr_i), \tag{24}$$

where a_i is the complex amplification gain associated with element i, p is the pressure amplitude of the source at a unit distance, r_i is the distance from the source to element i, and k is the wave number of the sound. Note that since k is proportional to frequency, the response is frequency dependent. If the sound source is in the far field of the transducer array, then the phase changes dominate over the amplitude variations and the response can be approximated by

$$O = \frac{p}{r} \sum a_i \exp(jkr_i). \tag{25}$$

For example, consider a line array of n omnidirectional sensors as shown in Fig. 9. Assuming the source to be in the far field, the response goes as

$$O = \frac{p}{r} \sum a_i \exp(jkx_i \sin\theta), \tag{26}$$

where x_i is the distance to element i from the array center, and θ is the angle between the normal to the array center and the source. If the amplifier gains and element spacing are uniform, Fig. 10 shows the resulting directivity pattern at a number of frequencies. Note that there is a main lobe whose beam width decreases with increasing frequency, and a number of side lobes. If the amplifier gains are not uniform, the directivity pattern is modified. For example, by introducing phase delays, one can steer the main beam away from the broadside ($\theta = 0$ rad) angle. Further, using various amplitude shadings one may narrow or widen the beam width, and raise or lower the side lobes. The commonly used array shadings are discussed in virtually all texts on antennas and filters.

For a radiating array a similar argument holds: even driving the elements in phase results in out-of-phase contributions at a point in the medium off the normal to the array. For example, consider a broadside source array which consists of a number of identical sources spaced uniformly in a line. The far-field array response is as shown in Fig. 10. By varying the amplifier gains for each element, the amplifier phase for each element, or the spacing between elements one may dramatically affect the directive response. This is a

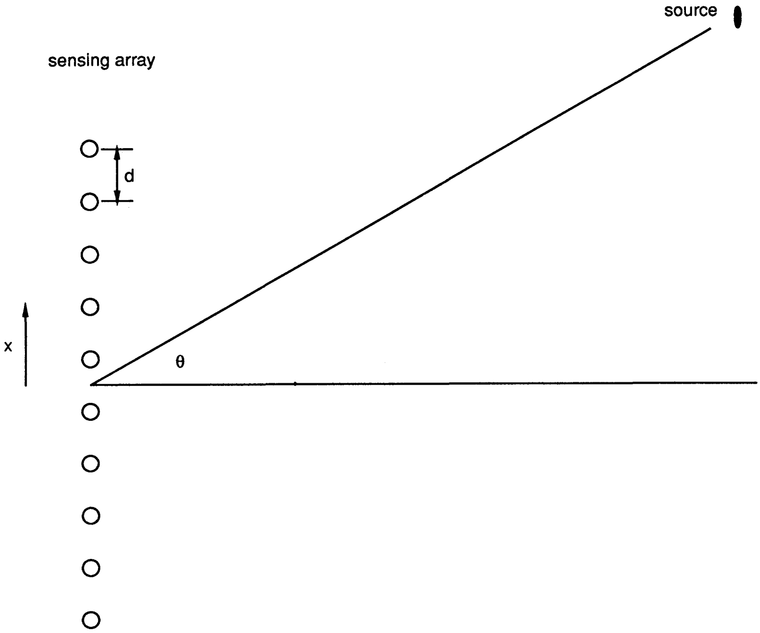

FIG. 9. Geometry of a typical linear array of transducers. In this array, spacing between elements is fixed at *d*.

routine practice in underwater sound sources and leads to beam forming, in which one chooses to emit sound in a specified direction.

Using these same principles one may build two- and three-dimensional arrays of various sorts. Typically, the directive response for arrays is frequency dependent, and includes side lobes as well as a main lobe. For more information on arrays, see Beranek (1954, 1986).

It should be noted that it is possible to build acoustical transducer arrays in which the elements are themselves directional. In this case, the product theorem guarantees that the directivity factor of the array is the product of the directivity of the elements and that of the array geometry.

1.4 Radiation Efficiency

It is important to characterize the ability of acoustical sources to couple vibratory motion of a transducing element with the surrounding medium in order to radiate sound. One means of describing the efficiency of this coupling is the radiation efficiency σ_{rad}, defined as the ratio of the acoustic power radiated by a source to that radiated at high frequencies by a baffled uniform circular piston source of the same size and mean squared velocity.

In general, the radiation efficiency of a vibrating object is less than 1 for low frequencies, and rises monotonically as frequency is increased to a value of 1 at frequencies for which the object is large compared to a wavelength. At low frequencies the energy imparted to the medium is stored in masslike or springlike behavior rather than being converted into acoustical waves. The above results are essentially independent of the shape of the vibrating object and the material of which the surrounding medium is comprised. This suggests that acoustical generators must be made large compared to the wavelengths of interest in the surrounding medium, or else their efficiency of radiation will be low.

2. FUNCTION AND OPERATION

Below we describe the practical operation of a number of acoustical transducers of various types.

2.1 Condenser Microphones

The condenser microphone shown in Fig. 11 uses a stretched membrane to convert acoustic pressure to displacement and as one plate of a capacitor. As was shown in Sec.

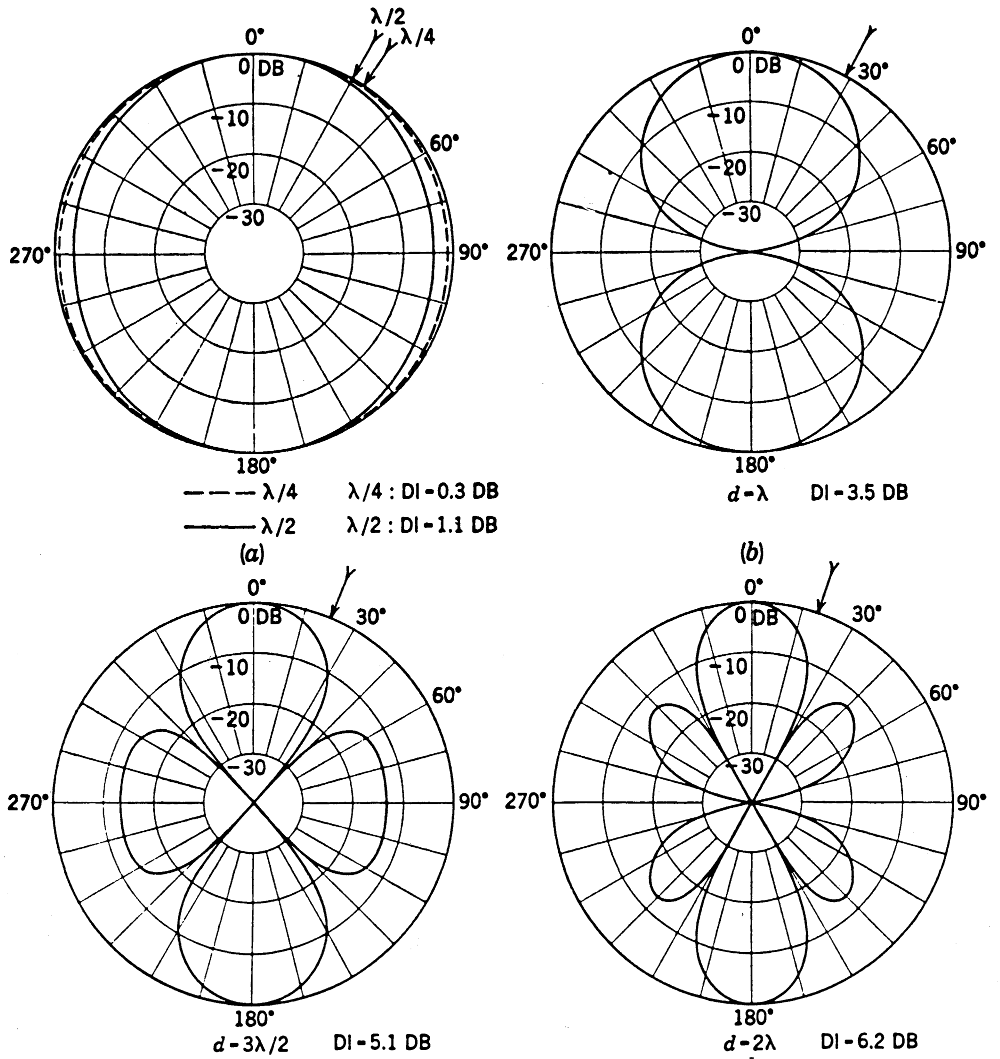

FIG. 10. Far-field directivity of the linear array shown in Fig. 9 as a function of frequency, from low frequency (a) to high frequency (d) [from Beranek (1986)].

1.1.1.1, the application of a static charge allows the pressure-induced change in capacitance to produce a potential proportional to pressure.

The static charge may be supplied by using a plastic membrane with embedded charge and a metallized outer surface. This arrangement is usually referred to as an electret microphone. Here tension and polarizing charge are reduced over those for the metal membrane but sensitivities can be large. Another alternative to the conventional condenser microphone shown in Fig. 11 is one in which the membrane tension is quite small. Under this condition the stiffness of the membrane is dominated by the stiffness of the surrounding air, which primarily depends on the air cavity depth and volume.

Because condenser microphones have a capacitive internal impedance, static or very low-frequency pressures are impractical to measure. This is because potential-sensing instruments have a finite input resistance. In general, the lower frequency limit is given by

$$f_1 = 1/2\pi R_{\mathrm{in}} C_0, \tag{27}$$

where R_{in} is the instrument input resistance. This low-frequency limit is averted by using

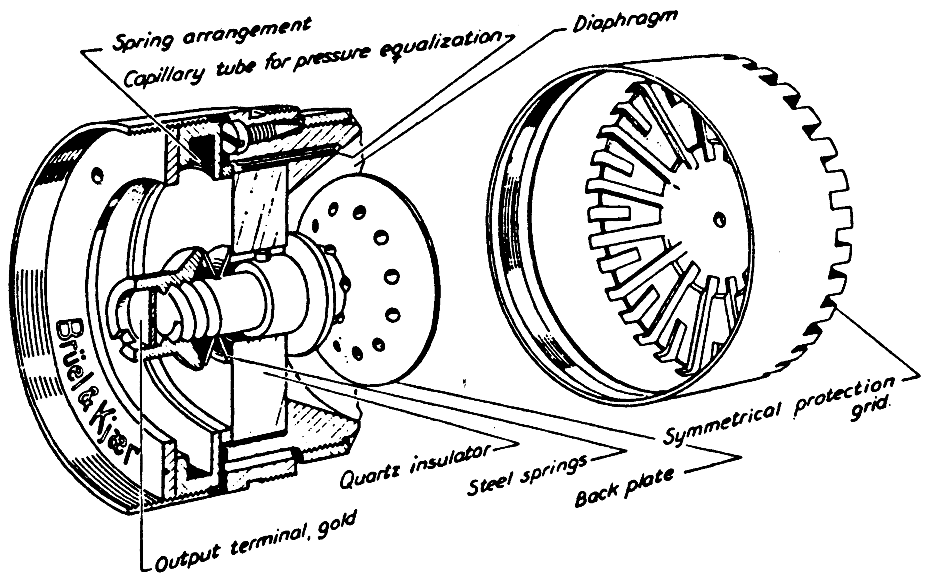

FIG. 11. Typical condenser microphone. Active element is a stretched diaphragm which is placed in proximity to a perforated backplate.

an ac polarization potential. When C_0 is made part of a parallel L-C circuit that determines the frequency of an oscillator, the change in C with acoustic pressure causes an oscillation frequency change which can be monitored using a frequency detector. The change with membrane displacement is

$$\Delta f = \frac{f_0}{2d}\xi, \tag{28}$$

where $f_0 = 1/(2\pi\sqrt{LC_0})$. Condenser microphones have been built to monitor sound at frequencies of fractions of 1 Hz using this method. However, it should be noted that in order to obtain a flat response to such low frequencies it is generally necessary to eliminate the rear pressure equalization port from the transducer because it is acoustically transparent at very low frequencies and makes the condenser microphone respond as a gradient microphone in this frequency regime. Note that even in the ac polarization case there is a static collapsing force on the membrane because the mean square polarizing potential is not zero.

The upper frequency limit of a condenser microphone is the lowest membrane resonant frequency and is determined by the tension and mass of the membrane. For a circular membrane the lowest resonance is given by

$$f_1 = \frac{1.20}{\pi a}\sqrt{T/\rho_s}, \tag{29}$$

where ρ_s is the surface density of the membrane (mass per unit area). The response at f_1 is controlled by acoustic resistance behind the membrane. Precision condenser microphones with flat response up to about 150 kHz are readily available.

When a microphone is used to measure a traveling wave (free field), diffraction affects the high-frequency response. When the diameter of the membrane is approximately equal to the acoustic wavelength, sound arriving at normal incidence experiences a pressure doubling at the membrane causing a 6-dB error. At grazing incidence, pressure averaging across the face can reduce sensitivity. By controlling the acoustic resistance the normal incidence response can be corrected; however, directivity still remains. For use in a diffuse sound field with waves arriving from random directions, a different compensation is used to account for directivity. Microphones can be compensated to yield a flat frequency response to free field, random incidence, and uniform pressure.

A condenser microphone made with a stainless-steel membrane and other metal parts is the most stable with temperature and other environmental changes. It holds its calibration for long periods of time and may be used as a secondary standard.

2.2 Piezoelectric and Ceramic Transducers for Liquids

2.2.1 Hydrophones. Devices for the measurement of acoustic pressure in liquids usually use solid sensors such as piezoelectric crystals or polarized ceramics. Often, as shown in Fig. 12, stacks of transducing material are used in order to enhance the overall transducer sensitivity. When the pressure is applied to one face, such as in Fig. 2, the potential is proportional to pressure as in Eq. (8). Because of the capacitive internal impedance, response down to dc is not practical without the use of an oscillator circuit as in condenser microphones. Since the ceramic is stiffness controlled, the frequency response is flat up to the first longitudinal resonance which is determined by the speed of sound in the material and its thickness. Piezoelectric and ceramic hydrophones for use in monitoring sound in the range of 20 Hz to 300 kHz are commonly available.

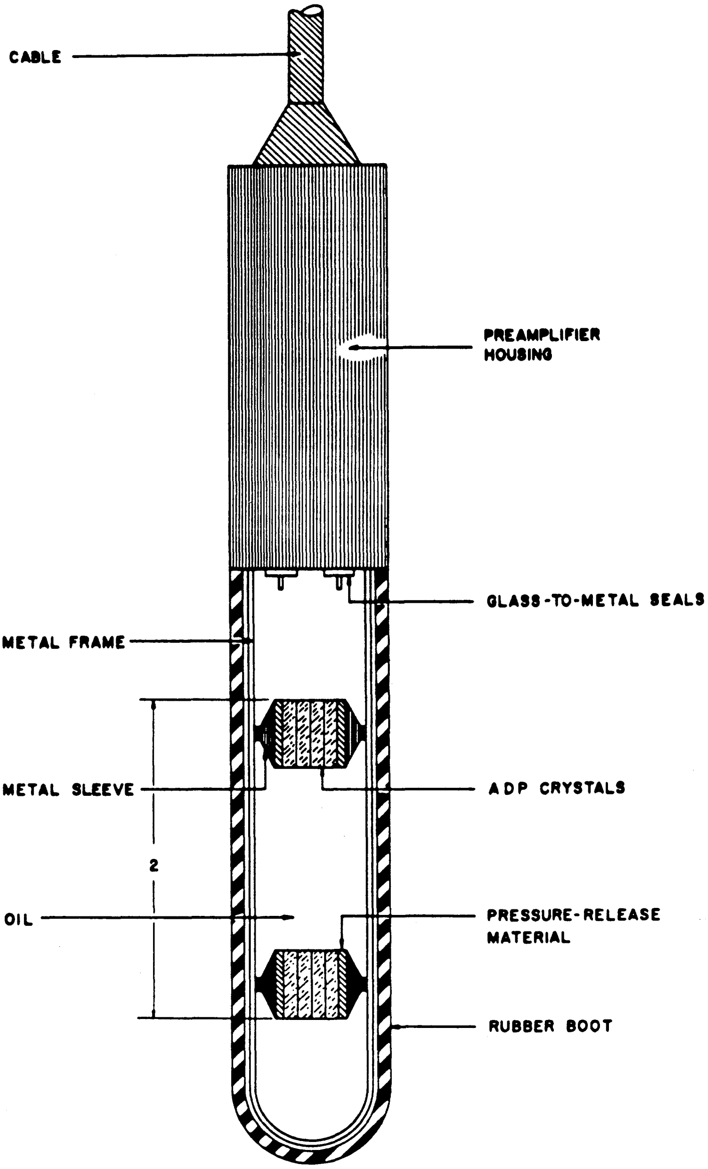
CABLE
PREAMPLIFIER HOUSING
GLASS-TO-METAL SEALS
METAL FRAME
METAL SLEEVE
ADP CRYSTALS
2
OIL
PRESSURE-RELEASE MATERIAL
RUBBER BOOT

FIG. 12. Typical piezoelectric hydrophone. Active transduction elements are the stacks of piezoelectric crystals [from Bobber (1988)].

In piezoelectric and ceramic hydrophones the maximum sensitivity would be obtained if one face were exposed to the pressure and the opposite face were clamped. However, because of the high mass and stiffness of the transducing materials, this cannot be done effectively. A compliant backing is usually used to give free conditions. Such a backing is shown in the commercial hydrophone in Fig. 12.

Ceramic elements can be exposed to acoustic pressure on all sides to give omnidirectional response if the device is small compared to the acoustic wavelength. The response then is proportional to the sum of the direct and transverse piezoelectric coupling terms $e_{33}+e_{13}$. The term e_{13} is negative but smaller than e_{33} so a net response results with a lower sensitivity than the single-sided case.

Ceramic materials also can be formed into cylinders and spheres to provide nearly omnidirectional pressure sensors (Wilson, 1985). Here the geometry offers a natural shielding of the back side of the sensing surface (interior electrode) from the acoustic pressure.

In addition to rigid ceramics and piezoelectrics, piezoelectric plastic film such as polyvinylidene fluoride (PVDF) can be used in pressure sensors. These polymers offer the advantage of flexibility but are usually inappropriate for sound sources since their piezoelectric constants are roughly an order of magnitude lower than those of conventional rigid piezoelectric ceramics.

2.2.2 Sound Sources. The crystals and polarized ceramics discussed above are reciprocal and can be used as sources of sound pressure. Because these solids are stiff compared to liquids, coupling to the liquid is usually poor. Below the first resonance, the devices are stiffness controlled which gives a sharply rising frequency response. For example, Fig. 13 shows the transmitting and receiving sensitivity of a ceramic underwater transducer. Thus, use of a piezoelectric and ceramic transducer is restricted to operation at resonant modes resulting in narrow-band operation.

Polarized ceramic can be driven with high potentials to produce large acoustic intensity at resonant modes. However, there are three factors which limit the usefulness of polarized ceramics for high-intensity projectors. First, the square-law nature of the source mechanism produces harmonics. This can be seen from Eq. (2) by recognizing that for high-amplitude driving the charge is no longer well approximated by the static polarization charge alone. Replacing Q_0 by Q_0+q in Eq. (4) yields a second-order term in the electric potential of $q\xi/\epsilon A$ which produces second-harmonic distortion and a change to the dc bias. Second, heating changes the material parameters of ceramics substantially. Third, poor material tensile strength can result in fracture of ceramics. For example, PZT-4, a common ceramic used in underwater transducer applications, can withstand a maximum stress of roughly 24 MPa. This corresponds to a maximum strain of about 10^{-3}.

For further information on hydrophones and underwater sound sources see Bobber (1988) and Wilson (1985).

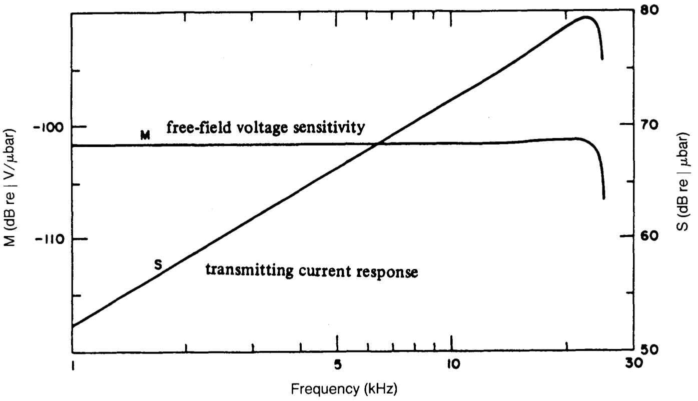

FIG. 13. Typical transmitting and receiving sensitivities of a piezoelectric underwater transducer as a function of frequency [from Bobber (1988)].

2.3 Loudspeakers

Sound pressure in gases can be generated by a vibrating piston. The piston size should be comparable to the acoustic wavelength for good coupling and the required motion is large because of the low acoustic impedance of gases.

In theory, the piston can be a thin membrane under tension and driven electrostatically. Examination of Eqs. (3) and (4) shows that the force then is quadratically related to the potential by

$$F=\frac{\epsilon A}{2(d+\xi)^2}V^2\sim\frac{\epsilon A}{2d^2}V^2, \qquad (30)$$

and a large polarizing charge is required for linearity. However, this produces a large dc electrostatic attraction force which tends to destabilize the system by causing the membrane to collapse onto the electrode. Hence large motions are difficult to obtain and electrostatic loudspeakers are rare.

Most loudspeakers use a light-weight piston which is driven to large motions electromagnetically. Such a loudspeaker is pictured in Fig. 14. Paperlike material in the form of a cone provides a light-weight stiff piston. The cone is usually supported by corrugated springs and driven at the apex. Current in a coil in a fixed magnetic field provides the driving force as in Eq. (16a). At frequencies below the resonances of the springs and cone-driver mass the motion is large causing significant distortion. Operation therefore is usually in the mass-controlled region and a relatively constant frequency response is obtained. The upper frequency limit usually is set by the first radial resonance in the cone material. Typical loudspeakers operate in the audio frequency range, from roughly 20 Hz to 20 kHz.

Operation at and below the suspension system resonance is possible if damping is added to limit the cone motion. If an aluminum coil form is used to hold the wire in the magnetic field, motion will produce eddy currents in the coil form. The mechanical resistance is then given by $R_m = B_0/R_c$ where R_c is the electrical resistance around the coil form. Additional electrical resistance in the electrical circuit can add mechanical damping.

In general the loudspeaker cone is mounted in an open frame. Thus the piston produces a pressure at the back which is π rad out of phase from that at the front. The effect is a negative source placed about a piston diameter behind a positive source. This dipole effect produces a frequency response which falls off at low frequencies:

$$p\sim(2\pi f)^2 I \qquad (31)$$

for frequencies where the piston is small compared to an acoustic wavelength. To eliminate this effect it is necessary to put the loudspeaker in a box or baffle so that the back side does not radiate. The box may be designed so that the acoustic impedance presented at the back of the cone can improve the low-frequency response of the loudspeaker system.

Small pistons driven electromagnetically often are placed at the throats of horns to improve the loudspeaker directivity. Special horn shapes provide frequency-independent directivity characterisitcs. The horns also serve to provide better matching between the piston and the air, resulting in improved efficiency. However, large piston excursions and high intensities in the throat lead to nonlinearities. The use of horns below 500 Hz is usually not practical.

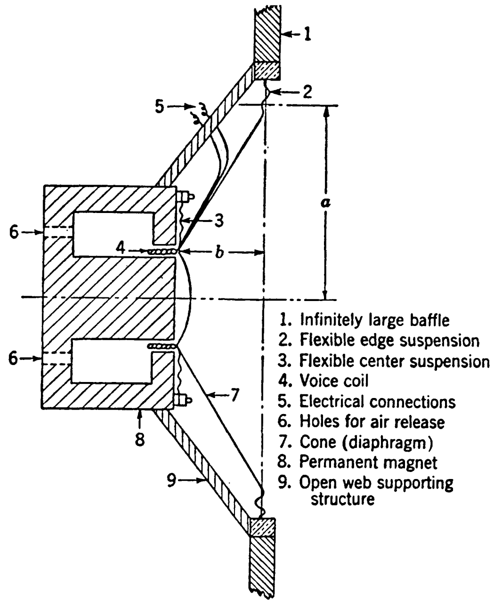
1
2
5
a
6
3
4
b
6
7
8
9
1. Infinitely large baffle
2. Flexible edge suspension
3. Flexible center suspension
4. Voice coil
5. Electrical connections
6. Holes for air release
7. Cone (diaphragm)
8. Permanent magnet
9. Open web supporting structure

FIG. 14. Typical dynamic loudspeaker. Device is quadratic inductive and uses the change in current through a voice coil on the speaker cone to cause a displacement of the cone [from Beranek (1986)].

2.4 Geophones

In geophysical exploration using seismic waves, it is desirable to measure motion of the ground resulting from a distant excitation. Because of frequency-dependent attenuation in the medium only waves below 1 kHz are useful. The surface motion can be determined by acceleration, velocity, or displacement sensors, but acceleration sensors exhibit poor low-frequency response and displacement sensors require a fixed displacement reference. Due to the frequency range and rigors of field operations, velocity sensors called geophones are used.

The geophone, an example of which is shown in Fig. 15, uses a compliantly supported seismic mass and the electromagnetic effect of Eq. (16b) to measure the relative velocity of the earth's surface with respect to the mass. As in the loudspeaker a coil of wire is placed in a constant magnetic field. This coil is attached to the structure that moves with the earth's surface. A permanent magnet and magnetic circuit conveniently provide the seismic mass. When the mass is stationary the output potential is directly proportional to the ground velocity.

Since the mass must be supported by springs essentially attached to the ground, a spring-mass resonator results. At and below the resonant frequency the mass motion is not zero. Thus the resonant frequency must be well below the frequency of interest. The resonant peak in the response can be controlled with eddy current and mechanical damping.

The upper frequency response limit is determined by the inductance of the wire in the sensing coil and the resistance in the coil and external measuring circuit:

$$f_h = R_t / 2\pi L, \tag{32}$$

where L is the inductance and R_t is the total circuit resistance. Typical geophones operate with relatively flat frequency response from a few Hz to 1 kHz.

Geophones can be made very rugged, and because of a low internal impedance, long cables can be used without on-board amplifiers.

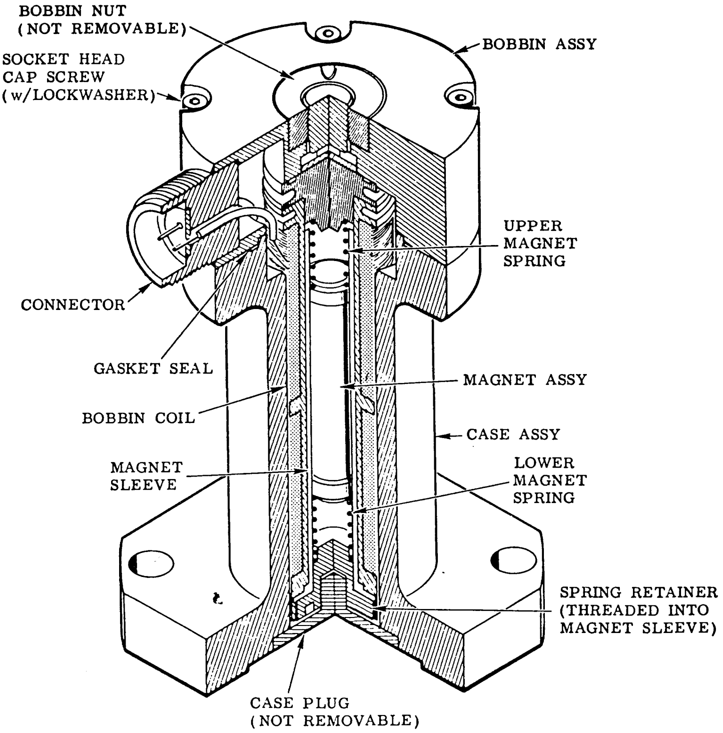
BOBBIN NUT (NOT REMOVABLE)
BOBBIN ASSY
SOCKET HEAD CAP SCREW (w/LOCKWASHER)
UPPER MAGNET SPRING
CONNECTOR
GASKET SEAL
MAGNET ASSY
BOBBIN COIL
CASE ASSY
MAGNET SLEEVE
LOWER MAGNET SPRING
SPRING RETAINER (THREADED INTO MAGNET SLEEVE)
CASE PLUG (NOT REMOVABLE)

FIG. 15. Electromagnetic geophone using the motion of an inertial mass relative to an earth-fixed coil to generate a current proportional to mass velocity [from Norton (1969)].

2.5 Calibration

There are generally three calibration methods used in acoustic instrumentation: comparison calibration, absolute direct calibration, and absolute reciprocal calibration. In comparison methods the transducer to be calibrated is compared to a similar transducer of known characteristics. This method is fast and eliminates the need for specialized equipment, but is the least precise of the techniques because errors in the calibration of the known source propagate.

In absolute direct calibration a specialized piece of equipment is used to provide a known calibration signal, which must be nearly independent of the load seen by the calibrator. A common acoustical example of this type of technique is the use of pistonphones for calibration of microphones. Pistonphones use reciprocating pistons to change the volume of a calibrated cavity in order to generate a well-known sound level. The advantage of absolute direct calibration methods are that they generally are fast and very precise. However, since they require the purchase of dedicated instrumentation, they tend to be expensive. Further, in order to obtain the desired precision, they generally offer absolute calibration at a very limited number of frequencies. For example, most pistonphones operate at a single frequency only. Complete calibration of a transducer then requires the use of the comparison method as well as absolute direct calibration.

The third calibration procedure which is commonly used for acoustical instrumentation, and which was historically developed for microphone calibration, is absolute reciprocal calibration. It is this method which separates the calibration of acoustical instrumentation from calibration methods common to other fields. Reciprocal calibration requires three transducers: one to be calibrated, one which is reciprocal (meaning it can operate either as a sensor or as an actuator), and one additional transducer which is complementary to the transducer to be calibrated (i.e., an actuator if a sensor is being calibrated and a sensor if an actuator is being calibrated). It is not necessary that any of the transducers have known calibration characteristics.

Although there are now many reciprocal calibration methods, all require a minimum of three measurements. For example, the following is a method for reciprocal calibration of a microphone or hydrophone. First, reciprocal transducer B is placed a distance l from source C. A current i_C is run through C and the open-circuit potential of B, v_B, is monitored. Next, transducer B is replaced by the transducer whose calibration we desire, transducer A, and the measurement is repeated yielding v_A. Finally, sound source C is replaced by B, a current i_B is run through it, and v'_A, the new open-circuit voltage from A is found. Defining φ, the sensitivity of A, as the ratio of the open-circuit voltage to the sound pressure, the measurement procedure yields

$$\varphi^2 = \frac{v_A v'_A}{v_B i_B} \frac{2l\lambda}{\rho_0 c_0}, \tag{33}$$

where λ is the wavelength of the sound, and ρ_0 is the mass density of the medium.

The advantages of reciprocal calibration compared to direct absolute and comparison calibration are that it is very accurate, can be applied at many frequencies, and requires only electrical and length calibration equipment. The major disadvantage is that it is more tedious than the other calibration methods. For more information on reciprocity calibration see Beranek (1954, 1986).

3. NEW TECHNOLOGIES

Over the past 15 years there have been rapid advances in acoustical sensors and actuators. These advances have matched advances in physics in general and can be classed into four areas: fiber optic transducers, solid-state sensors and actuators, new materials, and novel techniques. Below we describe each of these areas in some detail.

3.1 Fiber Optic Transducers

Fiber optic transducers offer many advantages over conventional nonoptical transducers: higher resolution, higher signal-to-noise ratio, immunity to electromagnetic interference, and compatibility with fiber optic communication systems. In addition they are easier to use than most other optical transducers. Hence they are rapidly finding their way into sensors spanning the frequency range from dc to over a MHz.

In acoustical sensors, there has been great interest in the use of fiber optic transducers underwater, where they are good candidates for undersea surveillance and seismic monitoring. They are also in demand for active vibration control of structures and for process control in chemical plants.

All fiber optic sensors have four main portions: a light source, a sensor, a detector, and a demodulator. The light source may be incoherent or coherent. The detector includes circuitry designed to process the optical signals.

Fiber optic sensors are divided into two classes: those in which the light carried in a fiber leaves the fiber, reflects off of a vibrating object, and then returns to the detector via a fiber (extrinsic), and those in which the light never leaves the fiber (intrinsic). Figure 16(a) shows a fiber optic lever, a simple example of an extrinsic fiber optic sensor. The main component of this instrument is a fiber optic probe consisting of multiple fibers, half of which transmit light and half of which carry the light reflected off of the target back to the detector. The detector generates an output voltage proportional to the intensity of the reflected light. No modulation is required. As shown in Fig. 16(b), the output voltage is null when the probe touches the target since the light reflects specularly into the sending fibers rather than the receiving fibers. As the distance between the target and the probe end increases, the output voltage rises rapidly until the entire surface of the probe is illuminated. Increases in separation beyond this distance result in a voltage drop due to square-law reduction in reflected light intensity on the probe surface.

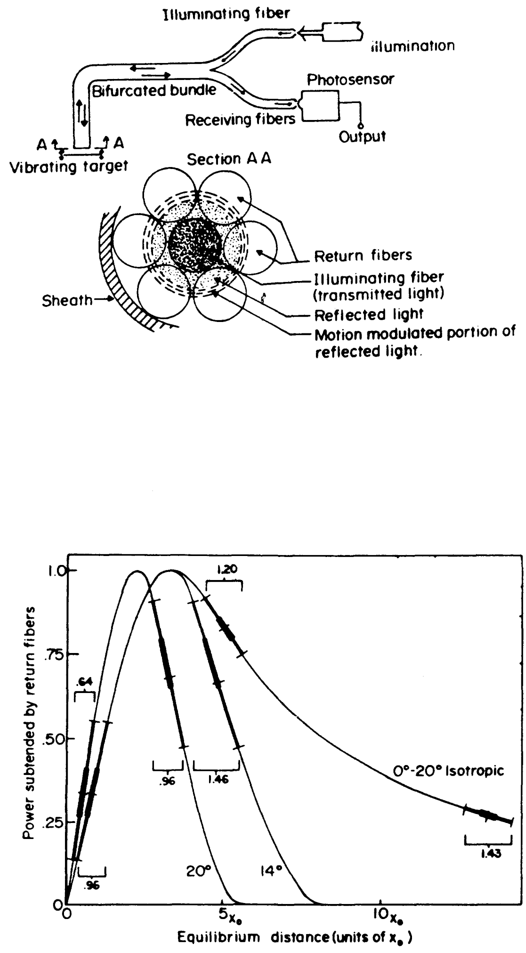

FIG. 16. Fiber optic lever: (a) system operation; (b) view of fiber bundle used in the probe; and (c) reflected power received from a planar target as a function of distance. The three curves are for three different fiber bundle configurations: concentric rings of transmitting and receiving fibers, semicircles of sending and receiving fibers, and random transmitting and receiving fiber orientation.

Intrinsic fiber optic sensors are designed so that the length of a fiber varies with changes in a parameter to be measured. Typically the length variation is determined by comparison to a reference length of fiber. An example of such a sensor is shown in Fig. 17, which illustrates a Mach-Zehnder interferometer. In this sensor a coherent light source is split so that part of the beam travels in a reference fiber and part in a sensing fiber. The reference fiber is wrapped around a piezoelectric crystal which is used as a phase modulator. By applying a potential across the piezoelectric, the reference fiber length is controlled, thus permitting the modulator to establish a bias point and to correct for any initial mismatch in the length of the two fibers. The sensing fiber is exposed to acoustic pressure which produces a strain in the fiber. This causes a change in the refractive index and thus a change in the effective length of the fiber. By recombining the reference and sensor signals and monitoring the phase shift from the original, the sensor fiber length change is very accurately determined.

The fiber optic devices discussed above are also representative of the common modulation schemes. Most fiber optic devices are based on amplitude modulation, as in the optic lever, or phase modulation, as in the

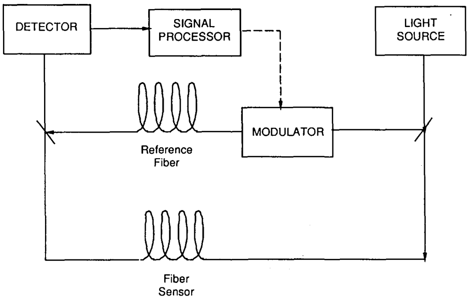

FIG. 17. Schematic of a Mach-Zehnder interferometer.

Mach-Zehnder interferometer. Phase-modulation fiber optic sensors are more sensitive but the signals are more difficult to modulate.

For more information on fiber optic sensors see Winsor (1980), Grant (1988), and Krohn (1988).

3.2 Solid-State Sensors and Actuators

For the past decade there has been substantial work in the area of solid-state sensors, and in the last few years there has been increasing interest in solid-state actuators. Solid-state transducers are generally devices made using the standard materials found in microelectronic devices such as silicon. The advantages of solid-state sensors and actuators are many: there is good fabrication process control, miniaturization is easily achieved, transducers may be integrated with the electronics on a single chip, and costs per item are low. The notable successes in solid-state sensors include devices for the measurement of acoustic pressure.

The construction of solid-state sensors differs substantially from the standard fabrication processes involved in microelectronics. Two items in particular have presented problems: size scales and feature shapes. The sizes of sensor and actuator transducing elements are generally much larger than standard microelectronics features and there have been problems obtaining consistently high yields of large-feature items. Further, most solid-state transducers require nonstandard feature shapes such as cavities. These nonplanar shapes can be difficult to obtain, although selective chemical etches, which follow preferred crystal orientations, have permitted rapid progress overcoming this stumbling block. Another difficulty with solid-state sensors and actuators has been the desire to integrate the processing electronics with the device. Although such an approach offers advantages from the perspective of signal-to-noise ratio, the yields for electronics alone are still substantially higher than those for the transducer alone. Hence, insisting upon integration usually results in loss of a large amount of viable electronics. In addition, many of the fabrication processes for sensors and actuators seem to be fairly incompatible with electronics processing. For further information about the fabrication of solid-state sensors see Petersen *et al.* (1988).

Acoustical solid-state transducers generally have been limited to pressure sensors (for air and water), although attempts to build sound generators have been reported. Solid-state pressure sensors have been designed using a piezoresistive effect, a piezoelectric effect, and the effect of stress on a *p–n* junction. A varying capacitance effect much as described in Eq. (5) is also used. Of these, the varying capacitance and piezoresistive have received the most attention.

In typical capacitive solid-state pressure sensors such as shown in Fig. 18, a moving membrane is coated with metal and a bias voltage is applied between the membrane and a fixed backplate. Devices such as this are miniaturized versions of conventional condenser microphones, except that the stiffness is usually determined by the size of the air cavity rather than the membrane tension. However, the solid-state capacitive transducers may be more sensitive and usually are useful up to much higher frequencies due to the increase in the first resonant frequency which accompanies the miniaturization. Their main disadvantage is a low capacitance of generally 1–3 pF, necessitating close location of processing electronics and causing poorer low-frequency response than their macroscopic counterparts. For further information on capacitive solid-state pressure sensors see Ko (1986).

Piezoresistive solid-state pressure sensors are more numerous than the capacitive type. In these devices, an example of which is shown in Fig. 19, resistors are typically diffused into a membrane and used in a half-bridge or full-bridge form. Membrane motions cause strain and thus result in differential resistances. These devices are easier to fabricate than the capacitive pressure sensors, and thus are less expensive, but are generally more limited in range due to hysteresis and nonlinearities. In addition, the piezoresistive solid-state sensors are susceptible to interference due to stressing of the package since they sense differential stresses.

3.3 New Materials

In recent years there have been great advances in materials which are finding their way into acoustical instruments with great rapidity. These advances are primarily in two areas: polymers and rare-earth magnetostrictive materials.

3.3.1 Polymers. Before the advent of polymer-based transducers, acoustical sensors and actuators used heavy transducer materials. With the discovery that some of these materials could be replaced by light-weight polymers without a drastic change in the transduction principles, new transducers were invented. Until recently, these new materials were not used to their full extent because they were viewed merely as replacements for heavier materials.

The first of these new transducers using polymers were electret acoustic sensors which make use of a permanently polarized polymer film (usually Teflon). Electret transducers operate like conventional condenser (capacitive quadratic) transducers, but the bias is supplied by the polarization rather than an external source. This results in a device which has very low power requirements, low distortion, and high sensitivity. Electret microphones are now the standard microphones found in telephones, hearing aids, and tape recorders. There are well over one hundred million electret microphones sold every year worldwide.

More recently piezoelectric transducers have been investigated in which ceramics and crystals are replaced by piezoelectric polymer films, particularly polyvinylidene fluoride (PVDF). PVDF acoustic transducers include microphones, earphones, hydrophones, and biomedical devices. In addition to conventional piezoelectric acoustical instruments, there have been a few transducer designs which take advantage of the flexibility of the film. For example, Bennett *et al.* (1977) designed a variable-focus hydrophone in which a source is located in a volume by varying the curvature of a PVDF membrane. PVDF sensors generally have higher sensitivity than their ceramic and crystal counterparts and are generally more accurately linear, but PVDF is expensive to fabricate and does not exhibit good temperature stability. New piezoelectric polymers and copolymers offer some of the same advantages of PVDF without the fabrication problems. For a review of PVDF transducers see Sessler (1981).

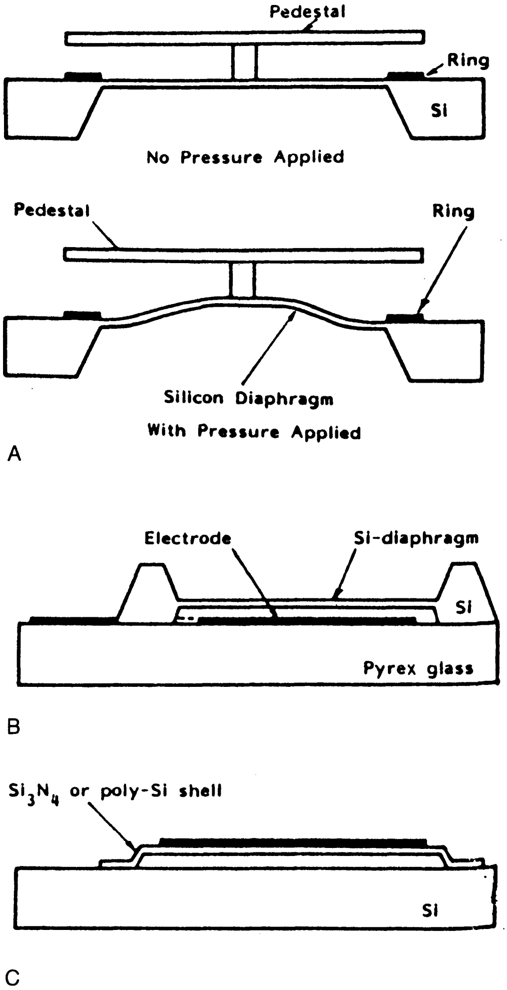

FIG. 18. Three examples of capacitive solid-state pressure sensors: (a) using a pedestal connected to a silicon diaphragm; (b) using a silicon diaphragm over a metallized piece of Pyrex glass; and (c) using a solid-state membrane over a silicon substrate [from Ko (1986)].

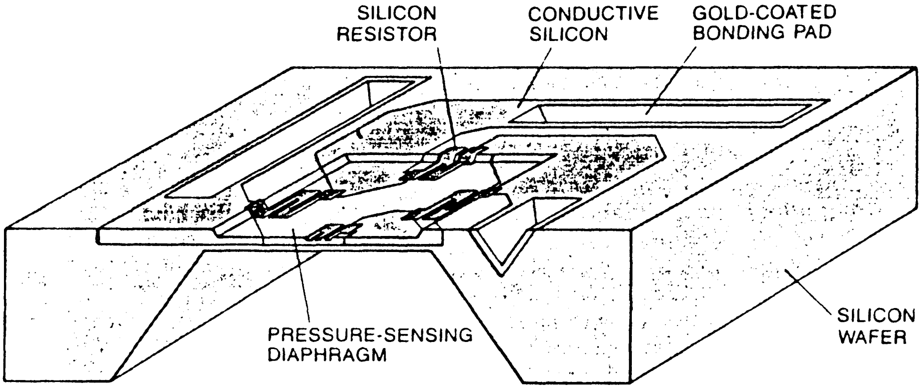

FIG. 19. Piezoresistive solid-state pressure sensor. Pressure strains a solid-state membrane into which strain gages have been diffused [from Angell *et al.* (1983)].

Recent work in polymer composites has also been leading to new acoustical instruments. In particular, work by Newnham (1986) has led to an ability to combine nonactive polymeric materials with transducing materials in a composite. The resulting composite has mechanical properties similar to the polymer, and transduction properties determined by the additive. An interesting example of this technology has been the invention of a material composite using barium titanate (a piezoelectric ceramic) as the additive. As shown in Sec. 1.1.1.2, one of the troubling properties of barium titanate is a relatively strong coupling between strains in the three principal directions; i.e., a forced strain in one direction will result in a strain in the orthogonal directions. By placing the material in a polymer matrix it has been possible to effectively decouple the principal strains while simultaneously raising the material effective piezoelectric constants due to increased flexibility. This material is currently being studied for underwater acoustic applications.

New polymers are being discovered and designed at a breakneck speed. Although many have not yet found a niche in acoustical instrumentation, several offer intriguing possibilities. For example, the magnetic materials used for disk storage and conducting polymers seem to be well suited to transduction.

3.3.2 Magnetostrictive Materials. In the early 1950s the bulk of the work done in acoustical transducers for underwater operation considered magnetostrictive materials (nickel compounds, for example). With the discovery of piezoelectric ceramics the magnetostrictive materials were largely abandoned, as they did not compare favorably in cost or sensitivity. Recent work in rare-earth materials has prompted new interest in magnetostrictive transducers.

The current interest in magnetostriction has focused on a material known as Terfenol-D ($Tb_{0.3}Dy_{0.7}Fe_{1.9}$). This material was developed by Clark (1980) and has just become commercially available. It can safely generate peak strain amplitudes of about 2×10^{-3}, which is roughly a factor of 3 times greater than that of conventional piezoelectric ceramics, and more than an order of magnitude better than nickel. The first attempts to use Terfenol-D in acoustical instruments are just being reported and modest successes are being observed.

3.4 Novel Techniques

In addition to advances in materials, recent years have seen advances in techniques used in acoustical instruments. Among these are laser-generated sound, magnetohydrodynamic transducers, acoustic heat engines and refrigerators, and acoustic microscopy.

Laser-generated sound is the process of using a laser to generate sound in a medium. Most of the work has involved fluid media, although there has been work on laser-generated sound in solids. Several transduction mechanisms have been identified for laser-generated sound in fluids, varying widely in efficiency and directivity. The two mechanisms that have received the greatest attention are the thermal mechanism and surface ablation.

In the thermal mechanism a laser shined into a fluid medium causes the medium to heat and expand. By modulating the laser, the expansion of the medium is modulated, resulting in sound wave propagation. The advantage of this approach to sound generation in a fluid is that no instruments need be located in the fluid medium where they are subject to corrosion. However, the process of sound generation is very inefficient. Attempts to increase the source sensitivity using a moving laser beam have met with moderate success.

In the surface ablation mechanism of laser sound generation the laser is focused onto the surface of a fluid where it induces a localized boiling. This causes a momentum transfer into the medium which radiates away as sound. Although much more efficient than the thermal mechanism of laser sound generation, this blast mechanism is less directive and has proved difficult to control and analyze. Surface ablation using a laser has also been used to create short duration acoustical pulses in solid materials.

Magnetohydrodynamic (MHD) acoustical instruments are transducers which rely on the coupling between magnetic, electric, and velocity fields in a conducting medium. A typical realization of an MHD acoustic transducer is shown in Fig. 20, and uses salt water in a duct. Permanent magnets establish a magnetic field in one direction, and an applied electric field in an orthogonal direction results in a mean velocity of

$$U = V_0 / d\mu H \tag{34}$$

for the medium in the third orthogonal direction which is generally taken to be along the axis of the duct. Here V_0 is the mean electric potential, H is the magnetic field established by the permanent magnets, and d is the diameter of the duct. In addition to the mean fluid flow (which is nonexistent in the absence of a dc bias voltage), there is an oscillating fluid motion which causes acoustic radiation into the medium surrounding the source. Alternatively the MHD transducer may be used as a receiver, in which case the fluid oscillation caused by incoming sound results in the establishment of an electric field. These MHD transducers offer the advantage of broadband operation, but they are less efficient than their piezoelectric counterparts.

Acoustic refrigerators and heat engines based on both MHD and thermally induced sound have been studied. In these devices thermal variations are used to generate sound waves which propagate in a duct. The thermoacoustic sound generating mechanism is analogous to the thermal mechanism for sound generation by a laser discussed above. The sound in the duct is coupled to electrical power through an MHD mechanism. The result is a compact instrument which has high-energy conversion efficiency. For example, Migliori and Swift (1988) have reported on a liquid-sodium acoustic engine in which an imposed temperature difference is used to generate sound; most of the sound energy is recovered in an MHD transduction process.

Acoustic microscopy (see MICROSCOPY, ACOUSTICAL) has been the subject of intense research for about 15 years. In acoustic microscopy a sound wave is reflected off of a submerged target and compared to the incident sound. The sound wave is generated using a piezoelectric crystal or a polarized ceramic, and the returning wave is typically sensed using the same device. The sound beam is focused using a lens such as shown in Fig. 21. This lens consists of a sapphire crystal with a spherical cavity cut into it. The lens moves in a raster pattern to scan the surface of the target. Acoustic microscopes operate at about 3 MHz in water and 8 GHz in mercury. They span over four orders of magnitude of wavelengths, much broader than optical microscopes. They offer about a 200-Å resolution, which is better than that for optical or x-ray microscopes, but inferior to ion or tunneling microscopes. Applications in which acoustic microscopy has been demonstrated to be the method of choice include examining fibers in composite materials, examining features in biological cells nondestructively, and detection of individual grains in polycrystalline samples of anisotropic metals. For further information on acoustic microscopy, see Quate (1985).

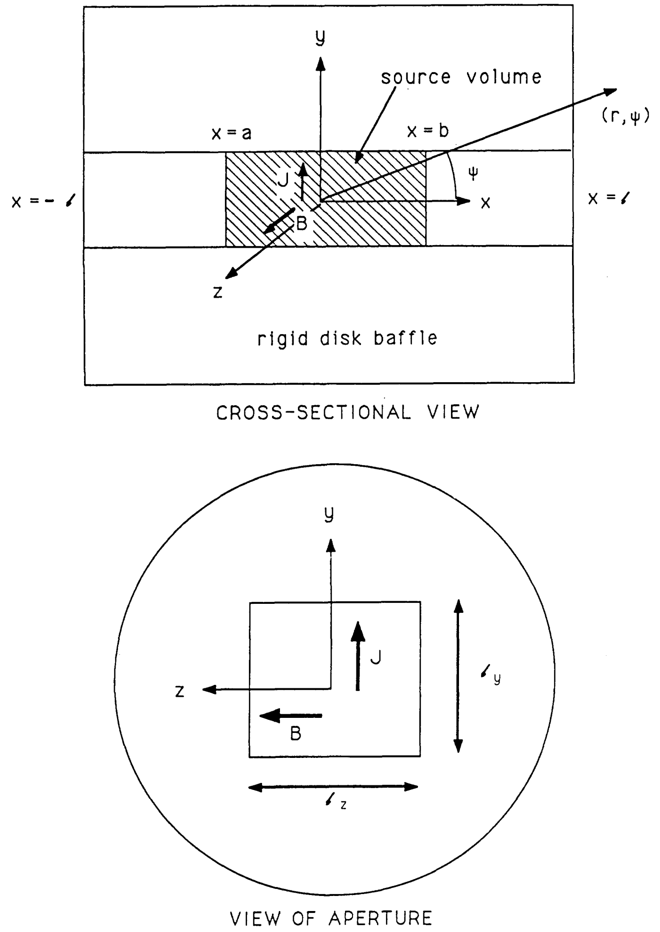

FIG. 20. Example of a magnetohydrodynamic transducer which couples a magnetic and electric field to velocity of a conducting fluid [from Quate (1985)].

4. APPLICATIONS

Acoustic instruments have numerous applications in communication, testing, signal processing, and research. Below we describe some of the most common of the applications.

4.1 Gases

The most common use of acoustical instruments is in telephony. Every telephone contains at least two acoustical transducers, a microphone and an earphone. Taken together these items account for most of the weight and cost of conventional handsets. For decades the standard telephone microphone was the carbon button microphone, a nonlinear resistive device. Now electret capacitive microphones are replacing the carbon microphone worldwide. Besides transducers, telephony research involves a number of other acoustical instruments including artificial ears which are microphones that are mounted at the bottom of standard-sized cavities, acoustical dummy heads which are used to simulate sound received binaurally, and microphone arrays which are used for teleconferencing.

Another very large segment of the acoustical instrument industry is dedicated to public announcement (PA) and home music entertainment (see ENGINEERING ACOUSTICS). PA systems typically concentrate on loudspeakers and speaker arrays suitable for speech communication. Transducer efficiency is more important than the quality of sound reproduction. Audio entertainment equipment includes a very wide range of acoustical instrument types and qualities. Standard components are loudspeakers, tuners (for radio reception), audio amplifiers, tape players (reel-to-reel and cassette), phonograph turn-

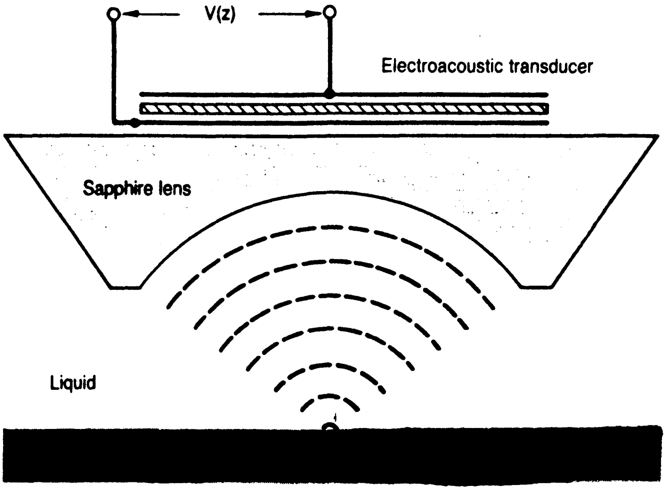

FIG. 21. Transducer portion of an acoustic microscope. A piezoelectric element generates a compressional wave which is focused by a lens and transmitted through a fluid to a target [from Quate (1985)].

tables, compact disk players, microphones, and earphones. The most recent advances have been the advent of compact disk technology and the development of very efficient polymer earphones which have permitted rapid growth of the portable audio equipment industry.

A third segment of the market has been the area of noise control. Due to government regulations worldwide, noise control work has concentrated on quieting the industrial noise sources which are sufficiently loud to cause hearing damage, and community noise sources most likely to interfere with speech communication and sleep. Typical acoustical instruments employed in studies to document noise sources are sound level meters and sound exposure meters. Sound level meters are acoustical instruments which combine a microphone, amplifier, logarithmic converter, filter, and a meter which displays a sound pressure level in dB, re 20 μPa. Most sound level meters provide an A-weighting filter network which is designed to simulate human hearing. In addition, some meters have octave band and one-third octave band filter sets available. In sophisticated sound level meters one may also vary the response time of the signal processing from very fast for impulsive sounds to slow for steady sounds.

Exposure meters are similar to sound level meters but are designed to determine the average A-weighted level to which a worker is exposed in order to check legal compliance. They thus use very slow averaging and do not have alternative filter sets.

4.2 Liquids

By far the largest application of acoustical instrumentation in liquids is in underwater sound systems (see ACOUSTICS, UNDERWATER). Historically the impetus for much of this work has been the heavy use of sonar systems for navigation, communication, and target location in the oceans. Working in oceans presents severe problems because the environment is highly corrosive to equipment, and opaque to visible light, infrared, radio, and radar signals. Acoustical signals are the only feasible means of transmitting information for distances greater than a few meters.

The early applications of underwater acoustical instruments concentrated on depth sounding and navigation. In the 1940s the emphasis shifted to target detection and sonar (sound navigation ranging) systems. There are two classes of sonar systems: passive sonar in which the received acoustic signal is that generated by a remote source, and active sonar in which a sound source is reflected off of a target and the echo detected. In both cases and in virtually all underwater communication systems used in oceans, a key problem is the detection of weak acoustical signals in the presence of noise. There are three standard means of overcoming this signal to noise problem. First, one may increase the power of an active sonar system. However, reverberation of the source signal off of the ocean bottom and other scatterers limits the effectiveness of power increases. Second, one may use a directive sound receiver so that sound arriving from undesired directions may be rejected. This is typically accomplished in sonar systems using both stationary hydrophone arrays and arrays which are towed by ships. Third, one may use a notch filter in the receiving circuit to reject a narrow band of frequencies which encompasses the greatest reverberation noise. This will also tend to reject the echo from a stationary source, but since moving sound sources result in Doppler shifted signals, moving targets still can be detected. For further information on sonar see Urick (1975).

In addition to underwater communication in the oceans, acoustical instruments are becoming progressively more important in biotechnology. Sonograms are routinely used in circumstances which do not permit x rays, such as in fetal monitoring. In sonogram equipment a directive, intense sound source is placed in contact with a body and the reflected sound signal is monitored, processed, and displayed graphically. Lithotripsy, the use of intense sound to break up kidney stones, has recently become more available and popular as a nonsurgical treatment. In lithotripsy a sound source is used to produce an intense, focused sound beam which selectively applies forces to hard objects such as kidney stones due to the large acoustical impedance mismatch with the soft tissue. Other biotechnical applications of acoustical instruments include ultrasound therapy for muscles, and acoustic microscopy in which sound waves are reflected from specimens rather than light.

Acoustical instruments for use in liquids are also widely employed in geophysical prospecting in oceans. In this application a source is used to generate a low-frequency sound wave in the water. This wave propagates through the water and partway into the ocean bottom. By monitoring the echo signals, a description of the ocean bottom at that location is obtained.

Acoustical instruments are also used for communication purposes in oil wells. Here, information is propagated in the form of compressional waves in the drilling mud which are monitored by hydrophones at the well surface. The information is sent at very low frequencies (under 3 Hz).

4.3 Solids

The applications of acoustical instruments in solid media parallel those in liquid and gaseous media. Among the major applications are nondestructive testing, length gaging, down-hole communication, geophysical prospecting, and surface acoustic wave devices.

Nondestructive testing of solids may be accomplished by generating an acoustic wave in the medium and looking for reflected acoustic signals. Such reflections indicate an impedance change located at a distance given by $t = \Delta t/2c_0$, where Δt is the time between the source signal and the echo. In this manner it is possible to check large structures for cracks or material voids such as occur in delaminations. Another popular application of this technology is the determination of lengths of objects which are only accessible at one end. In this case a transducer may send an acoustic wave into the object and monitor the time for the return of the signal reflected from the end. If the sound speed is well known for the material comprising the object, then its length may be accurately determined. Such a technique has been used, e.g., to determine lengths of muddy drilling pipes which are laid in stacks.

An alternative to down-hole oil well communication through acoustic wave propagation in drilling mud is the use of acoustic stress waves in drilling pipes or pumping rods. The generated waves carry information at much higher frequencies (audible range) and similarly do not require the locating of electrical instrumentation down the well. For further information on stress wave communication see Hixson (1966) and Drumheller (1989).

Geophysical prospecting in the earth's surface has been a significant application of acoustical instruments in solids. Typically a low-frequency wave is sent into the ground and the reflected signals monitored using geophones. There are generally two means of generating the acoustic wave: directed explosives and very powerful and massive shaker trucks.

Surface acoustic wave (SAW) devices have been used for a number of years for high-frequency filters and oscillators. In SAW devices a thin metal film is deposited onto a piezoelectric substrate typically of quartz or lithium niobate. The film consists of two electrically isolated contacts, each with fingers, deposited so that the fingers are interdigitally spaced. These fingers are used to excite an acoustic wave whose intensity decays exponentially with depth into the piezoelectric. Typical penetration depths and finger spacings are about an acoustic wavelength. By shaping the fingers and varying the spacing between them it is possible to control the time or frequency response of the device thus making them prime candidates for filters of various sorts. SAW devices are generally available in the 10 MHz to 2 GHz range, and are compatible with standard microelectronics processing techniques. For further information on SAW devices see Kino (1987).

GLOSSARY

Beam Width: Angular range of the main lobe of sensitivity of a transducer.

Bending Modulus: Ratio of the moment applied to a structure to its angular strain.

Cardioid Transducers: Sensors and actuators which expose the transducing element to sound on both the front and back face so that the net result is an even combination of a conventional pressure transducer and a gradient transducer.

Directivity Factor: Ratio of the peak response of a transducer, measured at a distance *r*, to its average response measured at the same distance.

Directivity Index: Decibel measure of the directivity factor: $\mathrm{DI}=10\ \log_{10}\mathrm{DF}$.

Electrical Admittance: Ratio of current to voltage (inverse of electrical impedance).

Electrical Impedance: Ratio of voltage to current (inverse of electrical admittance).

Gage Factor: Figure of merit for resistive strain gages. The gage factor is the ratio of resistance change to nominal resistance.

Geophone: A transducer designed to measure the motion of the ground in response to a remote excitation.

Gradient Transducers: Sensors and actuators in which the transducing element responds to the gradient of a state of the system, usually due to exposing the front and back sides of the transducing element to the environment.

Hydrophone: Underwater acoustic pressure sensor.

Magnetostriction: An effect in some materials in which the magnetic permeability is a function of the mass density.

Mechanical Admittance: Ratio of velocity to force (inverse of mechanical impedance).

Mechanical Impedance: Ratio of force to velocity (inverse of mechanical admittance).

Piezoelectricity: The coupling of the mechanical elastic field and electric fields in a material. An imposed strain of the material produces a changing electric field and an imposed electric displacement produces a mechanical stress in the material.

Piezoresistivity: The dependence of the electrical resistivity of some materials to the mechanical strain imposed on them. The strain affects the crystal lattice of the material and the material dimensions, thus changing the material resistance in two ways.

Radiation Efficiency: Ratio of the acoustic power radiated by a source to that radiated at high frequencies by a baffled uniform circular piston source of the same size and mean squared velocity.

Reciprocal Calibration: Calibration of a transducer which is accomplished using one or more other transducers, at least one of which is reciprocal (i.e., which can operate either as a sensor or as an actuator).

Reverberant Sound Field: A sound field which is uniformly (or nearly uniformly) distributed in a space, usually due to multiple reflections of source sound waves on boundaries.

Sound Level Meter: Device for measurement of sound pressure levels incorporating a microphone, amplification, filtering, and a display.

Surface Acoustic Wave (SAW) Device: High-frequency transducer in which the acoustic wave intensity decays exponentially with depth into the transducing material.

List of Works Cited

Angell, J. B., Terry, S. C., Barth, P. W. (1983), *Sci. Am.* April, 44–45.

Bennett, S.D., Chambers, J. (1977), *Electron. Lett.* **13**, 110–111.

Beranek, L. L. (1986), *Acoustics*, New York: AIP.

Bobber, R. J. (1988), *Underwater Electroacoustic Measurements*, Los Altos, CA: Peninsula.

Clark, A. E. (1980), in: E. P. Wohlfarth (Ed.), *Ferromagnetic Materials*, Vol. I. New York: North-Holland.

Cook, R. O., Hamm, C. W. (1979), *Appl. Opt.* **18**, 3230–3241.

Drumheller, D. S. (1989), *J. Acoust. Soc. Am.* **85**, 1048–1064.

Fischer, F. A. (1955), *Fundamentals of Electroacoustics*, New York: Interscience.

Grant, W. O. (1988), *Understanding Lightwave Transmission: Applications of Fiber Optics*, San Diego: Harcourt Brace Jovanovich.

Hixson, E. L. (1966), US Patent No. 3,252,225.

Kino, G. S. (1987), *Acoustic Waves*, Englewood Cliffs, NJ: Prentice-Hall.

Ko, W. H. (1986), *Sensors Actuators* **10**, 303–320.

Krohn, D. A. (1988), *Fiber Optic Sensors—Fundamentals and Applications*, Research Triangle Park, NC: Instrument Society of America.

Mason, W. P. (1964), *Physical Acoustics*, Vol. 1, Part A. New York: Academic.

Merhaut, J. (1981), *Theory of Electroacoustics*, New York: McGraw-Hill.

Migliori, A., Swift, G. W. (1988), *Appl. Phys. Lett.* **53**, 355–357.

Newnham. R. E. (1986), *Ferroelectrics* **68**, 1–32.

Norton, H. N. (1969), *Handbook of Transducers for Electronic Measuring Systems*, Englewood Cliffs, NJ: Prentice-Hall.

Petersen, K., Barth, P., Christel, L., Pashby, G. (1988), in: *Silicon Sensors and Microstructures*, Fremont, CA: Nova Sensor, Chap. 3.

Quate, C. F. (1985), *Physics Today* August, 34–42.

Rossi, M. (1988), Roe, P. R. W. (trans), *Acoustics and Electroacoustics*, Norwood, CT: Artech House.

Sessler, G. M. (1981), *J. Acoust. Soc. Am.* **70**, 1596–1608.

Schultz, T. J., Smith, P. W., Malme, C. I. (1975), *J. Acoust. Soc. Am.* **57**, 1263–1268.

Wilson, O. B. (1985), *An Introduction to the Theory and Design of Sonar Transducers*, Washington, DC: U. S. GPO.

Winsor, H. V. (1980), *SPIE* **239**, 252–258.

Further Reading

Allocca, J. A., Stuart, A. (1984), *Transducers: Theory and Applications*, Reston, VA: Reston.

Groves, Jr., I. D. (Ed.) (1981), *Acoustic Transducers*, Stroudsburg, PA: Hutchinson Ross.

Hanish, S. (1983), *A Treatise on Acoustic Radiation, Vol. II: Acoustic Transducers*, Washington, DC: U.S. GPO.

Korites, B. J. (1987), *Microsensors*, Rockland, MA: Kern.

ACOUSTICAL MICROSCOPY

See MICROSCOPY, ACOUSTICAL

ACOUSTICAL TOMOGRAPHY

BRUCE M. HOWE AND DANIEL ROUSEFF, *Applied Physics Laboratory, College of Ocean and Fishery Sciences, University of Washington*

Introduction 89
1. **Nontomographic Imaging** 91
2. **Straight-Ray Tomography** 92
2.1 Projection-Slice Theorem 93
2.2 Example of Straight-Ray Tomography 94
3. **Diffraction Tomography** 95
3.1 Scattering Problem 95
3.2 Fourier Diffraction Theorem 97
3.3 Inversion Algorithms 97
3.4 Examples of Diffraction Tomography 98
4. **Rays in a Waveguide: Ocean Tomography** 99
4.1 Forward Problem 99
4.2 Inverse Problem 102
4.3 Examples of Ocean Tomography 103
5. **Concluding Remarks** 105
Glossary 107
Works Cited 108
Further Reading 108

INTRODUCTION

Various forms of acoustic imaging are playing an increasing role in research and everyday life. It is becoming common practice for a pregnant woman to have a "sonogram" of her baby to be sure that all is well. Statistical quality control requires rigorous testing of manufactured items; much of the testing is done using acoustic nondestructive evaluation (NDE). In science and research, acoustic microscopes are being used that can often give a sharper image than optical microscopes. In clinical applications, research is providing simultaneous quantitative images of the sound velocity, attenuation, and fluid velocities. In the field of fluid dynamics, acoustic techniques are being used to map the temperature and velocity fields in laboratory to ocean scale flows.

The invention of x-ray tomography by Hounsfield in 1972 stimulated the current interest in all kinds of tomography. Since then the instrumentation and computer hardware and software have been improved so that accuracies of one part in 1000 are achievable from 10^6 data points. In the field of medicine, tomography has let doctors study the interior of a patient's body with precision and safety. While x rays are used to produce images of the x-ray attenuation coefficient, other forms or sources of energy including radio waves, magnetic resonance, positrons, radioisotopes, and sound are used to image different parameters.

Acoustic imaging can be loosely divided into tomographic and nontomographic systems. Nontomographic imaging generally consists of a single view of an object and presents the raw data with little or no processing. The resulting qualitative images highlight regions of strong scattering associated with large gradients in the refractive index or, equivalently, regions of high reflectivity, such as distinct boundaries. However, the "correct" interpretation of these qualitative images is largely dependent on the skill of a trained analyst. Nontomographic imaging is widely used in medicine and NDE.

In contrast, tomographic imaging is a distinctly two-step process. In its most pristine form, the object is first probed from many different angles. The measured data from the various views are then processed and combined to yield a quantitative reconstruction of some physical parameter such as the acoustic index of refraction. Tomographic reconstruction is therefore a much more experimentally and computationally intensive procedure, but the quantitative images that are generated

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

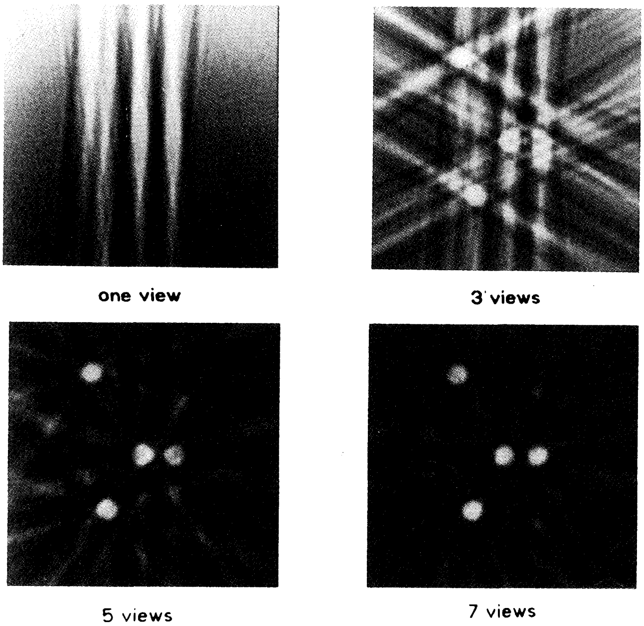

one view

3 views

5 views

7 views

FIG. 1. Reconstruction of four cylinders of radius five obtained by filtered backpropagation for a wavelength of unity. The quality of the reconstruction improves as additional views are incorporated (from Devaney, 1983; ©1983, IEEE).

can reveal subtleties that are not recovered by nontomographic methods.

In practice, the classification of a particular acoustic imaging system is not always clear. Many practical systems combine both tomographic and nontomographic features. For example, inherently nonquantitative medical images taken from different angles may be superimposed to create a superior image. In medical or geophysical applications of tomography, the number of views available may be far fewer than the number that are required for ideal inversion. The type of system used for a particular application depends on a variety of factors. In nondestructive evaluation, for example, the mere detection of a flaw may be sufficient, and quantitative information may not be needed.

The tomographic reconstruction process is illustrated in Fig. 1. In the simulation, a system of four cylinders is probed by an acoustic plane wave. An image based on a single view is inadequate for recovering the system. However, as the images generated from additional views taken at other angles are superimposed, the reconstruction is improved and the individual cylinders become apparent.

Acoustic tomography is related to several other topics covered in this volume. Acoustic microscopy (see MICROSCOPY, ACOUSTICAL) can be used in detecting surface cracks in NDE. Some forms of microscopy are in fact tomographic, but are beyond the scope of this article. The coherent field measurements required for diffraction tomography can be interpreted as acoustic holograms. Indeed, generalized acoustic holography (*q.v.*) can form the basis of a variety of both tomographic and nontomographic inverse scattering algorithms (Porter, 1989). Finally, because of the large amount of recorded data needed for tomographic inversion, efficient acoustic signal processing (see SIGNAL PROCESSING) techniques are required.

The contents of this article are divided primarily by imaging technique rather than by application. By taking this approach, we stress the similarity of the imaging problem for such seemingly disparate fields as medicine, geophysics, nondestructive evaluation, and oceanography. The first section briefly reviews nontomographic techniques as background for the tomographic methods discussed in the balance of the paper. The next section describes conventional straight-ray tomography. In this method, the probing energy is assumed to travel along straight ray paths through the object. Conventional tomography also illustrates the importance of multiple views for quantitative inversion. The third section discusses diffraction tomography. In diffraction tomography, the probing wave is assumed to obey a wave equation. Recent encouraging experimental results in dif-

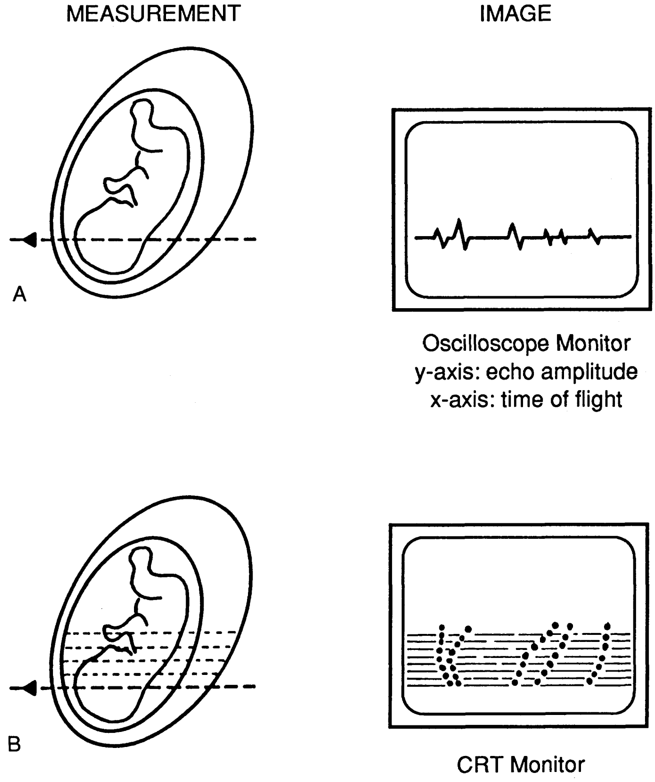

FIG. 2. Nontomographic imaging. (a) A-scan. Output displays echo amplitude vs distance into medium. (b) B-scan. Output is a cross-sectional image. The image brightness is related to the amplitude of the return pulse.

fraction tomography are discussed. Examples of current research from medicine, geophysics, and small-scale oceanography are included.

Straight-ray and diffraction tomography are similar because in many applications reconstructions are often "agnostic"; i.e., there is sufficient data that nothing need be assumed about the field being reconstructed. Section 4 deals with ocean acoustic tomography, in which the rays are refracted by the ocean sound channel. Because of a paucity of data, *a priori* information is necessary to constrain the reconstruction. Concluding remarks are in Sec. 5.

1. NONTOMOGRAPHIC IMAGING

Most methods of acoustical imaging currently used in clinical and NDE applications are nontomographic and are based on reflection or pulse-echo techniques. The different methods have historically been assigned letter designations: A-scan, B-scan, C-scan, etc. [Figs. 2(a) and 2(b)]. Many of the commercial systems incorporate some aspects of each method (Schueler *et al.*, 1984).

The A-scan is based on "time of flight": A pulse is transmitted and bounces off a scattering or reflecting boundary and the echo is detected, often by the transmitting element which also acts as a receiver [Fig. 2(a)]. The time between transmission and reception of scattered energy gives the range to the reflecting surface (assuming a constant speed of sound). The amplitude of the echo provides information on the strength of the scattering. The time corresponding to a large- amplitude peak is used to locate a major scattering boundary. It could be the interface between different organs in the body, a wall of a pipe, or the ocean bottom. Weaker amplitudes would be representative of, for instance, the relative severity of a flaw in a weld or casting, or scatter from schools of fish or zooplankton, or sediment layers beneath the ocean floor.

Time modulated (T/M) mode interrogation involves a succession of A-scan measurements taken at a fixed transducer location in even time increments. The output gives the

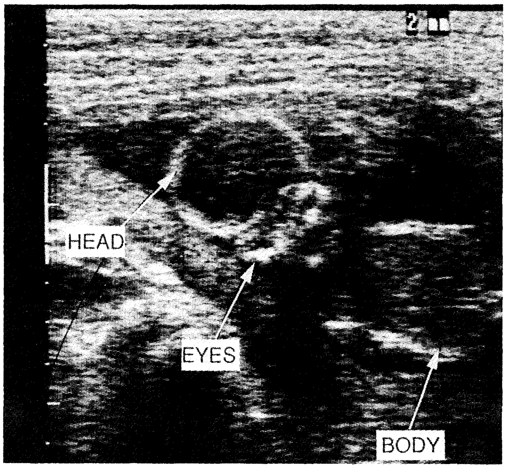

FIG. 3. B-scan image of the first author's son at the end of the first trimester.

range to a reflecting surface as a function of time and can be used to monitor cardiac motion.

By carefully moving the transducer along a line and taking a series of A-scan measurements, a cross-sectional "image" as a function of range and the scan dimension can be obtained; in medical systems, this is called a B-scan [Fig. 2(b)]. The transmitted acoustic wave is focused so that the imaged section is more nearly planar. The brightness of the display is related to the amplitude of the reflection. Scanning can be controlled either mechanically or electronically. Electronic scanning can be done very rapidly to give nearly real-time images. The output can be displayed on a television screen with a new image generated each 1/30 of a second. Figure 3 shows a typical sonogram of a fetus.

The corresponding oceanographic instrument is called Seabeam. A swath of the sea floor is mapped by an array of transducers on the bottom of the ship. As the ship moves along its track, topographic maps of the ocean bottom are produced with high accuracy. Seabeam is being used to chart the 200-mile Exclusive Economic Zone (EEZ) of the United States and other countries.

The objective in nondestructive evaluation is to locate faults in a structural component. While conceptually similar to medical A- and B-scan techniques, NDE has several important differences. Since the structural component is a solid, it can support shear and surface waves in addition to longitudinal waves. It is difficult to couple acoustic waves from air into a solid because most of the energy is reflected. Furthermore, the faults present in an object may vary widely in size, shape, and location; this diversity can complicate the interpretation of the output.

A similar procedure is also found in geophysics (see SEISMOLOGY). The objective in seismic inversion is to determine the structure of the earth beneath its surface. This information is crucial in oil exploration or in the disposal of nuclear waste. Low frequencies of less than 100 Hz are used to obtain information at depths of several kilometers. As with NDE, shear and surface waves are generated. Seismic imaging is simplified by the *a priori* assumption that structures vary primarily in depth by geologic layer.

The C-scan can operate in either the reflection or transmission mode. As a transmitter is moved over a plane, a receiver is also moved and the received intensity as a function of position is recorded. This is equivalent to a simple x-ray picture since both provide a "shadowgraph" of the object.

Doppler ultrasound uses the change in apparent frequency between the transmitted and received signals to measure relative motion. The technique is often used in medical applications to monitor blood flow. A typical transmission frequency is 3 MHz and the shift in frequency due to blood flow is usually in the audio range. Consequently, after demodulation the output signal is often sent to an audio speaker allowing the clinician to "hear" the blood flow. Continuous- wave Doppler systems require separate transducers for transmission and reception and give no information about the location of a flow. Pulsed Doppler systems do give range data, but require more elaborate timing and demodulation circuitry.

2. STRAIGHT-RAY TOMOGRAPHY

The nontomographic systems of the previous section operate primarily in the pulse-echo mode. Many tomographic systems operate in the transmission mode; the object is between the transmitter and receiver. In propagating through the object, the probing wave is perturbed in accordance with the properties of the medium. The simplest prop-

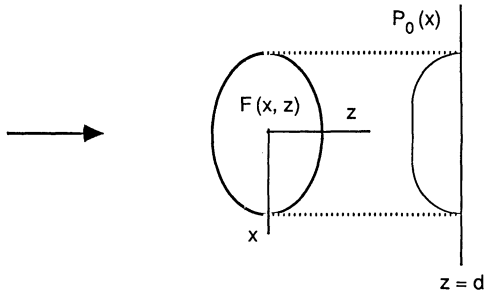

FIG. 4. Straight-ray tomography. Probing energy travels in straight-line paths through the unknown object F(*x*, *z*). The measured quantity $P_0(x)$ is the geometric projection or "shadowgraph." If the projection corresponds to travel time measurements along rays parallel to the *z* axis, then $F(x, z) = C_0/C(x, z) = n(x)$, the index of refraction.

agation model is to assume that the transmitted energy travels along straight ray paths through the object. This model permits the derivation of relatively simple inversion algorithms that parallel those common to x-ray tomography.

For simplicity, we begin with a two-dimensional imaging problem and later generalize to the full three-dimensional case. The typical experimental configuration is shown in Fig. 4. The object function $F(x, z)$ describes some physical parameter of an object such as the attenuation rate or the index of refraction that we wish to determine. For this view, the object is probed by energy traveling along a straight ray in the z direction. It is assumed that the measured quantity (such as intensity or travel time) is related to the geometric projection $P_0(x)$ of the object. The geometric projection at this angle is related to the object function by

$$P_0(x) = \int_{-\infty}^{\infty} F(x, z)dz. \tag{1}$$

The objective is to recover the unknown function $F(x, z)$ from projections measured at many different angles. If the data are measurements of intensity, then $F(x, z)$ is the attenuation field; if the data are measurements of travel time, $F(x, z)$ is the reciprocal of the sound speed. From a practical point of view, the projection can only be sampled at a finite number of points corresponding to a finite number of rays.

2.1 Projection-Slice Theorem

It is instructive to examine the relationship between the projection and the object function in the Fourier domain. Because of its relative simplicity, the analysis for the geometry shown in Fig. 4 is now presented. Defining the Fourier transform of the projection as $\tilde{P}_0(k_x)$, we find from Eq. (1) that

$$\begin{aligned} \tilde{P}_0(k_x) &\equiv \int_{-\infty}^{\infty} P_0(x)e^{-ik_x x}\,dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x, z)e^{-ik_x x}\,dx\,dz. \end{aligned} \tag{2a}$$

Defining the two-dimensional transform of the object as

$$\begin{aligned} \tilde{F}(k_x, k_z) \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} & F(x, z) \\ & \times e^{-ik_x x}\, e^{-ik_z z}\,dx\,dz, \end{aligned} \tag{2b}$$

it follows that

$$\tilde{P}_0(k_x) = \tilde{F}(k_x, k_z{=}0). \tag{2c}$$

Equation (2c) indicates that by taking the one-dimensional transform of the measured projection, we recover the two-dimensional transform of the object evaluated in Fourier space along the line $k_z = 0$. This result is easily generalized to an insonification angle ϕ as shown in Fig. 5. This fundamental relationship is referred to as the Projection-Slice theorem: The one-dimensional transform of the projection taken at angle ϕ maps onto a radial line at angle ϕ in the two-dimensional Fourier space of the object.

It is apparent from Fig. 5 that if we could probe the object at all possible directions over 180°, we would "fill" Fourier space and could then, in principle, recover the object function $F(x, z)$ by an inverse two-dimensional transform. The figure also illustrates a concept relevant to all types of tomography: A single view provides only limited information about an object and many views are necessary for an accurate quantitative reconstruction.

In practice, the reconstruction is usually performed in the spatial domain. Each projection is first filtered. Alternate filtering schemes are available, but the idea is to compensate for the relatively sparse sampling of $\tilde{F}(k_x, k_z)$ at high spatial frequencies when a finite number of projections are measured (Horn, 1978). These filtered projections are

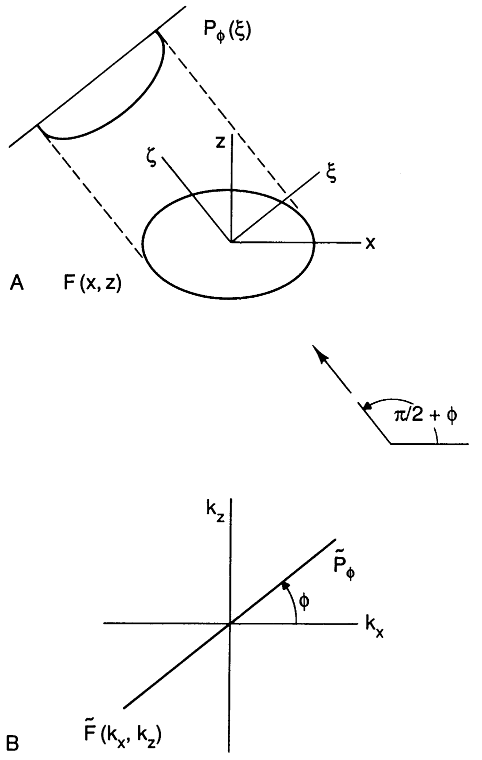

FIG. 5. The Projection-Slice theorem. The one-dimensional projection $P_\phi(\xi)$ maps onto a line in two-dimensional Fourier space of object *F*.

then backprojected along straight-ray paths. The resulting images from each view are then superimposed to yield the reconstruction. This inversion method is called *filtered backprojection* (Kak and Slaney, 1988). It avoids problems with interpolating in the Fourier domain, in favor of (usually linear) interpolation in the spatial domain. Most importantly, it is computationally efficient because calculations can start once the first projection is measured.

The generalization of these concepts to full three-dimensional inversion is straightforward. Because the energy is assumed to travel along straight ray paths, there is no out-of-plane scatter. Consequently, any cross section of the object can be reconstructed independent of all others by using the filtered backprojection algorithm. Reconstruction of the entire object is accomplished by reconstructing several parallel cross sections and stacking the results to give a three-dimensional effect. Thus the three-dimensional imaging problem is reduced to a series of independent two-dimensional problems. Each two-dimensional slice is reconstructed using one-dimensional projections.

2.2 Example of Straight-Ray Tomography

Schreiman *et al.* (1984) conducted a study of ultrasound-transmission computed tomography of the breast to evaluate its capability for detecting anomalies. The technique measures both the speed of sound and the attenuation within the breast by measuring the time of flight and reduction in amplitude of sound pulses. There was enough variability between patients in the imaged parameters that relative rather than absolute values were required; reference values were taken from "normal" adjacent tissue.

The scanning apparatus consisted of four focused transmitters and four receivers operating at 5 MHz; the corresponding acoustic wavelength was approximately 0.3 mm. Each transducer was 10 mm in diameter and separated from its neighbor vertically by 14 mm (Fig. 6). Using a translate/rotate scan, each transducer pair obtained a set of data used for reconstructing images in its plane of the breast. At each of 60 viewing angles, the instruments were translated horizontally to construct a projection made up of 200 parallel rays. A second set of scans was made with the transducers lowered 7 mm. This resulted in data for eight planes, 7 mm apart and approximately 3 mm thick. The entire procedure lasted about 5 min. Data were stored on magnetic media and processed off-line.

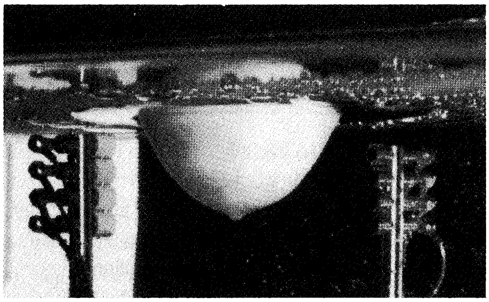

FIG. 6. The four transducer (transmitter-receiver) pairs used to scan a phantom breast. The patient lies face down on a table above a water tank and the breast is immersed in water during the scanning (from Schreiman *et al.*, 1984).

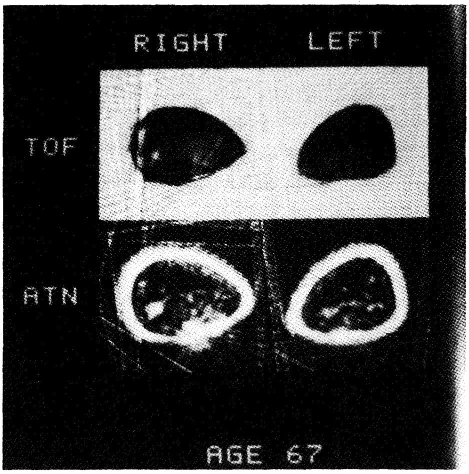

FIG. 7. The time of flight (TOF) images on top and the attenuation (ATN) images on the bottom show the small cancer. The sound speed change for the tumor over the background is 45 m/s (from Schreiman *et al.*, 1984).

The arrival time of the sound pulse at the receiver was measured using threshold detection of the signal after it had been passed through a full-wave rectifier and a bandpass filter. The amplitude of the received signal was estimated by integrating the first two microseconds of the signal after detection. Although this latter method averages the attenuation coefficient over frequency, its virtue is simplicity.

An example of the ultrasonic tomographic images is shown in Fig. 7. The reconstruction was done using the filtered backprojection method. It was found that malignant lesions could be identified by various ranges of sound speed and attenuation. The tumor shown has a low sound speed and a high attenuation. More work needs to be done to correlate tomographic results with tissue malignancy. It would also be interesting to see what effects ray bending and diffraction would have on these images. For additional examples of ultrasonic tomography as applied in medicine, the reader is referred to Robinson and Greenleaf (1985).

Tomographic techniques can also be applied to seismic data. Typically several boreholes are drilled between which either acoustic or electromagnetic signals are transmitted. Reconstruction methods often rely on *a priori* information and reduce to the solution of a system of equations. A similar procedure for the oceanographic case is discussed in detail in Sec. 4. For additional examples from seismology, see Daily (1986).

3. DIFFRACTION TOMOGRAPHY

The conventional tomography algorithms discussed in the previous section were largely developed for x-ray tomography (see X-RAY IMAGING: RADIOLOGY AND TOMOGRAPHY). The central assumption of the method, i.e., that the probing energy travels along straight ray paths, is certainly true for x rays. The successful application of such techniques to acoustics, however, has been limited. In acoustics, diffraction effects often cannot be neglected. Diffraction occurs whenever the acoustic wavelength is comparable to spatial scales associated with changes in the index of refraction of the medium. In diffraction tomography, the measured information is treated like a scattered field governed by the wave equation. We first derive an expression for the scattered field and then show how it can be used to reconstruct an object.

3.1 Scattering Problem

The two-dimensional scattering problem is shown in Fig. 8(a). A scatterer characterized by the index of refraction $n(\mathbf{x})$ is probed by a known incident acoustic plane wave. For a source-free region, we assume that the total acoustic pressure wave ψ satisfies the linearized, time-harmonic $[\exp(-i\omega t)]$ Helmholtz equation:

$$[\nabla^2+k^2n^2(\mathbf{x})]\psi(\mathbf{x})=0, \tag{3}$$

where $k = \omega/C_0$ is the wave number and C_0 is the speed of propagation in the background medium. In general, the index of refraction can be real or complex, the latter applying when the scatterer is lossy. Outside the object, the index of refraction is equal to that of the homogeneous background and $n(\mathbf{x}) = 1$.

The total field $\psi(\mathbf{x})$ outside the object can be written as the sum of the known incident field $\psi_i(\mathbf{x})$ and a scattered field $\psi_s(\mathbf{x})$. The incident field is that which would propagate in the absence of the scatterer. The scattered field contains the perturbations in the total

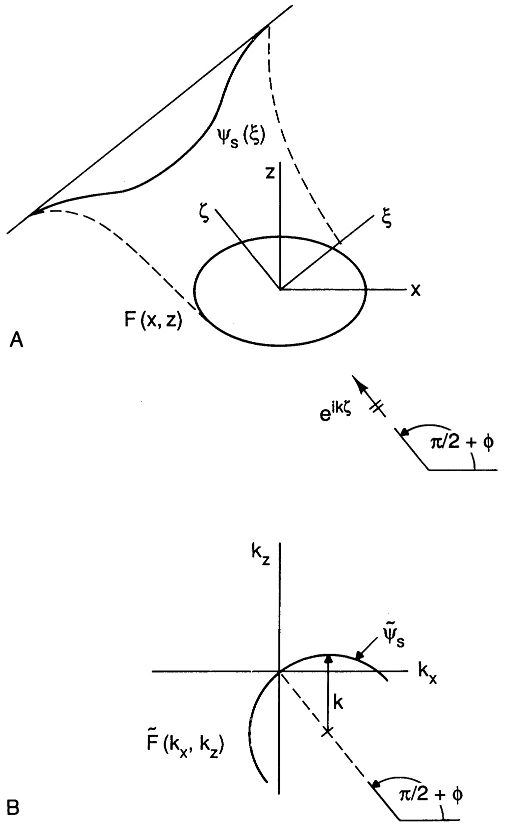

FIG. 8. Diffraction tomography. (a) An incident plane wave probes the unknown scatterer *F*(*x*, *z*). The measured field is governed by the wave equation. (b) The Fourier diffraction theorem. The one-dimensional transform of the scattered field $\psi_s(x)$ maps onto a semicircle in the two-dimensional Fourier space of the object *F*. The radius of the semicircle is the wave number *k* of the incident wave.

field due to the object and is given by (Ishimaru, 1978)

$$\psi_s(\mathbf{x}) = \int_V F(\mathbf{x}')\psi(\mathbf{x}')G_f(\mathbf{x}, \mathbf{x}')d\mathbf{x}'. \tag{4}$$

Here $F(\mathbf{x}) = k^2[n^2(\mathbf{x}) - 1]$ is the scattering function, V is the volume of the object, and G_f is the free-space Green's function. The objective is to recover the scattering function from measurements of the scattered field. This task is complicated by the presence of the unknown total field $\psi(\mathbf{x})$ in the integrand, so simplifying approximations are often made. Equation (4) can be linearized by approximating ψ in the integrand with the known incident field ψ_i. This is the so-called (first) Born approximation and is valid when the scattered field is significantly weaker than the incident field and multiple scattering can be ignored. The first Born approximation includes only single scattering. Higher order approximations correspond to the incident field being scattered twice, and so on.

As an alternative to the Born approximation, the total field $\psi(\mathbf{x})$ may be expressed as the *product* of the incident field $\psi_i(\mathbf{x})$ and a perturbation term:

$$\psi(\mathbf{x}) = \psi_i(\mathbf{x})e^{\phi_c(\mathbf{x})}, \tag{5}$$

where $\phi_c(\mathbf{x})$ is the complex phase. A series expression for $\phi_c(\mathbf{x})$ can be generated where the first term in the expansion is referred to as the (first) Rytov approximation:

$$\phi_c(\mathbf{x}) \approx \psi_s(\mathbf{x})/\psi_i(\mathbf{x}), \tag{6}$$

where $\psi_s(\mathbf{x})$ is given by Eq. (4). The Rytov solution describes how the wave fronts of the incident wave are distorted as they go through the object. If the complex phase is small, the exponential in Eq. (5) can also be expanded in a series and the Born and Rytov solutions for the total field are equivalent to second order.

It is generally believed that the Born approximation is most valid for large-angle scattering while the Rytov solution is most valid in forward, line-of-sight propagation (the usual tomographic configuration). Which formalism should be adopted for numerical implementation? There is an important practical problem associated with inversion algorithms based on the Rytov approximation. In the Born approximation, the scattered field is obtained from the measured total field by a simple subtraction of the known incident wave. In the Rytov approximation, the complex phase is obtained by taking a logarithm, which is not straightforward since the logarithm of a complex argument is multivalued and phase unwrapping is notoriously difficult. In most cases, the assumption is that the scattering is sufficiently weak that either the Born or the Rytov approximation adequately models the forward scattering process. Inversion algorithms are therefore usually based on the Born approximation with the understanding that the Rytov solution might be better if the phase can be correctly unwrapped.

3.2 Fourier Diffraction Theorem

In direct analogy with straight-ray tomography, diffraction tomography inversions are usually based on the *Fourier diffraction theorem,* illustrated in Fig. 8. An object characterized by scattering function $F(x, z)$ is insonified by a plane wave propagating in direction ϕ. The total field is measured along a linear receiving array orthogonal to the incident field direction, and the scattered component $\psi_S(\xi)$ is obtained by subtracting the known incident wave [Fig. 8(a)]. The theorem states that the one-dimensional Fourier transform of the scattered field maps onto a semicircle in the two-dimensional Fourier space of the object [Fig. 8(b)]. We observe that the radius of the semicircle is the wave number k of the probing wave. This result can be proved by substituting an appropriate integral representation of the Green's function into Eq. (4) and taking the transform of the scattered field (Devaney, 1982).

By rotating either the measurement system or the object, the scattered field can be recorded at other angles. Rotating the insonification vector corresponds in Fourier space to rotating the semicircle, much in the same way as a line in Fourier space is rotated in the nondiffracting case. If we were to take views at all directions over 360°, we would cover all spatial frequencies out to a cutoff wave number $\sqrt{2}k$. Higher spatial frequencies cannot be recovered from the forward scattered field, so it is only possible to recover a low-pass filtered version of the object. Physically the low-pass filtering constraint corresponds to the fact that we cannot see details of the object much smaller than the wavelength of the incident wave. Consequently, we can recover only a smoothed version of the original object.

As the frequency of the probing wave is increased, the wavelength becomes smaller and we can therefore recover more of the fine detail in the object. This can be observed from Fig. 8(b). Raising the frequency increases the wave number and hence the radius of the semicircle. In the high frequency limit, Fig. 8(b) reduces to Fig. 5(b) and diffraction can be neglected.

3.3 Inversion Algorithms

Having formulated the scattering problem, we now consider techniques for inverting the scattered field to determine the scattering function. Inversion is more complicated when diffraction effects are considered; the measured quantity ψ_s is related to the unknown scattering function by a volume integral containing a Green's function [Eq. (4)] rather than a simple line integral [Eq. (1)].

Given the Fourier diffraction theorem, there are two basic techniques for inverting the measured scattered fields and reconstructing the two-dimensional object. The first essentially reduces to interpolation in the frequency domain, while the second involves interpolation in the spatial domain.

The first and most obvious reconstruction technique makes use of the discrete Fourier transform. The scattered field, which is measured along a linear recording surface, is uniformly sampled. The continuous transform of the field is then efficiently approximated using the fast Fourier transform (FFT). The discrete transform maps onto sampled arcs in the two-dimensional discrete Fourier space of the object. We can then, in principle, perform an inverse transform into real space to yield the reconstruction. The inverse transform is best achieved by a two-dimensional FFT which requires that the data be distributed on a uniform grid. The reconstruction problem therefore becomes one of interpolating from sampled arcs onto a uniform grid. Bilinear interpolation appears to work satisfactorily in some cases (Pan and Kak, 1983).

The second reconstruction approach operates entirely in the spatial domain. The filtered backprojection method used in conventional straight-ray tomography is modified to include an additional depth-dependent filter (Devaney, 1982). The backpropagation filter attempts to model the acoustic wave propagation within a finite depth interval. An example of this reconstruction technique was given in Fig. 1.

The Fourier diffraction theorem suggests another possible approach to the inversion problem. As shown in Fig. 8(b), the radius of the mapped semicircle is the wave number k of the probing wave. Recalling that the wave number is proportional to the temporal frequency, this implies that if the object were probed by a plane wave with a wide temporal bandwidth, additional information would be recovered in spatial Fourier space. It can be shown that there is a duality between probing a weak scatterer from many directions using a single-frequency wave and probing at a single direction using a wideband pulse (Porter, 1989). It should be noted, however, that the validity of the Born approximation becomes particularly questionable at high frequencies.

The Fourier diffraction theorem is easily extended to three dimensions. The two-dimensional transform of the scattered field measured along a plane maps onto a spherical shell in the three-dimensional Fourier space of the object (Wolf, 1969). In three dimensions, the incident field will be scattered out of the plane containing the insonification vector. Consequently, unlike straight-ray tomography, three-dimensional reconstruction cannot be accomplished by a succession of independent two-dimensional problems. Optimal inversion requires that the object be insonified at all possible angles over a complete sphere. In practice, this substantially increases both the number of views needed and the computational effort required to invert the data.

3.4 Examples of Diffraction Tomography

Diffraction tomography has yet to find widespread clinical applications. The fact that the two-dimensional algorithms of the preceding section are not easily extended to full three-dimensional inversion has limited their usefulness. Another major difficulty is the weak-scattering assumption; the measured field must obey either the Born or Rytov approximation. In the medical application, bones are not weak scatterers.

The clinical application of diffraction tomography with the greatest potential is imaging the breast. As we saw in the previous section, cancerous lesions can often be characterized as regions with relatively small perturbations in the acoustic index of refraction (Schreiman *et al.*, 1984); it is therefore hoped that the forward scattering can be accurately modeled by weak scattering.

In cases where the scattering is strong, Eq. (4) cannot be linearized and diffraction tomography does not yield a quantitative reconstruction of an object's acoustic parameters. Rather, it gives the sources induced in the object which depend on the pressure field within. The qualitative images that are produced, however, may reveal features of the object that could not be resolved by nontomographic techniques.

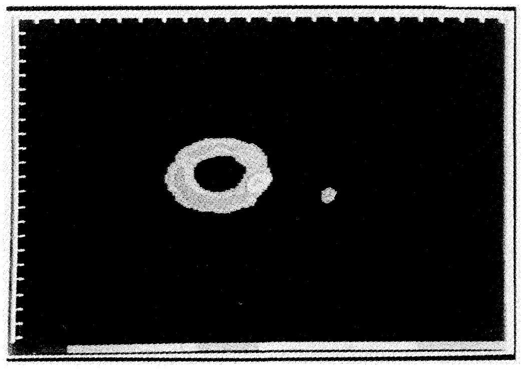

FIG. 9. A reconstruction of a rubber pipe and a nylon string. Each axis tick is 1 mm (from Duchêne *et al.*, 1988).

Duchêne *et al.* (1988) have reported on qualitative experimental reconstruction of cylindrical targets. They used an acoustic frequency and corresponding wavelength of 2 MHz and $\lambda_0 = 0.75$ mm, respectively. The targets were a rubber pipe and a nylon string. The inside and outside diameters of the pipe were 2 and 5 mm ($2.7\lambda_0$ and $6.8\lambda_0$), and the sound speed was 1560 m/s. The nylon string with a diameter of 0.6 mm ($0.8\lambda_0$) was highly refractive with a sound speed of 2600 m/s. A line receiver with 80 individual elements $0.7\lambda_0$ apart was used as the receiving array, with coupling between elements measured and taken into account in the processing. The magnitude of the complex reconstruction of a single pipe and string is shown in Fig. 9. This image was built up from 32 views. It appears to be qualitatively correct except for the bright spot in the pipe wall; this can be attributed to the nonoptimal element spacing. These authors employed the same technique to image the heart of a rabbit (size$\approx$24 λ_0), but the result is qualitative at best. Much remains to be done before high quality reconstructions can be realized.

Another potential application of diffraction tomography is imaging ocean microstructure, the perturbations in the sound speed field on scales less than a meter. Since the relative variance of these perturbations is on the order of 10^{-8}, the forward scattering problem is well modeled by either the Born or the Rytov approximation. Theoretical investigations have used diffraction tomography to recover realizations of such diffuse continua. Microstructure can also be modeled as an anisotropic random medium governed by a set of statistics. Tomographic techniques can then be used to estimate these statistics (Rouseff, 1989).

At larger oceanic scales, the effects of refraction dominate those of diffraction. This situation is discussed in the following section.

4. RAYS IN A WAVEGUIDE: OCEAN TOMOGRAPHY

Observing the ocean is an exceedingly challenging task because of the wide range of oceanic time and space scales. Spatial scales range from millimeters to megameters, while temporal variability ranges from seconds to years. Oceanographic measurements until recently have typically been point measurements in space and/or time. Aliasing because of undersampling has been the nemesis of oceanographers since the founding of the discipline over a hundred years ago. Satellite sensors will help a great deal in that they offer very rapid measurements over essentially the entire surface of the Earth, but they do not provide information on the interior of the ocean. However, the ocean is transparent to sound, and oceanic processes change the sound speed field by several percent. These facts led to the idea of ocean acoustic tomography, a method for using sound to probe the three-dimensional structure of the oceans.

This concept was formalized by Munk and Wunsch (1979). Their idea, reminiscent of medical x-ray tomography, was to transmit sound between instruments in a moored array over ranges of hundreds of kilometers. A mesoscale eddy or some other ocean feature would affect the travel times along some transmission paths but not others (Fig. 10). Because of practical limits on the number of instruments involved and thus the number of data, one must deal with many individual transmission paths rather than a projection made up of many paths. Geophysical inverse theory is used to sort out the magnitudes and locations of the perturbations. This measurement technique has many attractive features. Compared with conventional sampling methods, relatively few instruments are needed to sample a large volume because the number of data grows quadratically with the number of instruments, rather than linearly as with point measurements. Furthermore, because of the ocean sound speed channel (discussed in the next section), sound travels along many different paths sampling different depths, so information on the depth structure can be obtained from a single source-receiver pair. Because the measurement inherently averages the ocean along the path the sound takes, small-scale ocean processes are averaged out, thus improving the resolution of the larger scale features. Also, because sound travels faster with a current than against one, differential travel time is a measure of the water velocity along the path.

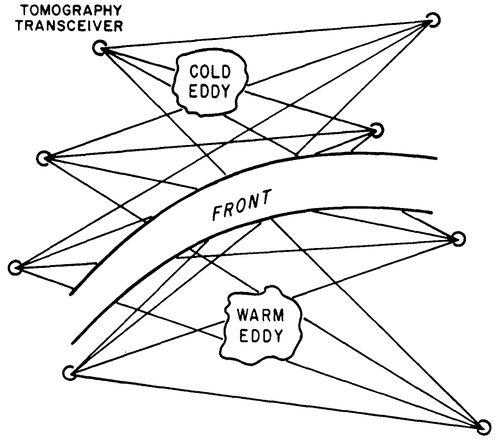

FIG. 10. Acoustic transceivers send and receive pulses along many paths. Inhomogeneities such as eddies and fronts alter the sound speed and thus change the travel times (from Spindel, 1986; ©American Society of Mechanical Engineers).

4.1 Forward Problem

Before one can use measured travel times to obtain maps of sound speed (roughly proportional to temperature) or currents, one must be able to predict the travel times for a given sound speed field. The feature that most distinguishes ocean tomography from other kinds of tomography is the ocean sound channel. The speed of sound (inversely proportional to the index of refraction) typically decreases from the surface as the water temperature decreases, but then increases due to the competing effects of pressure (see ACOUS-

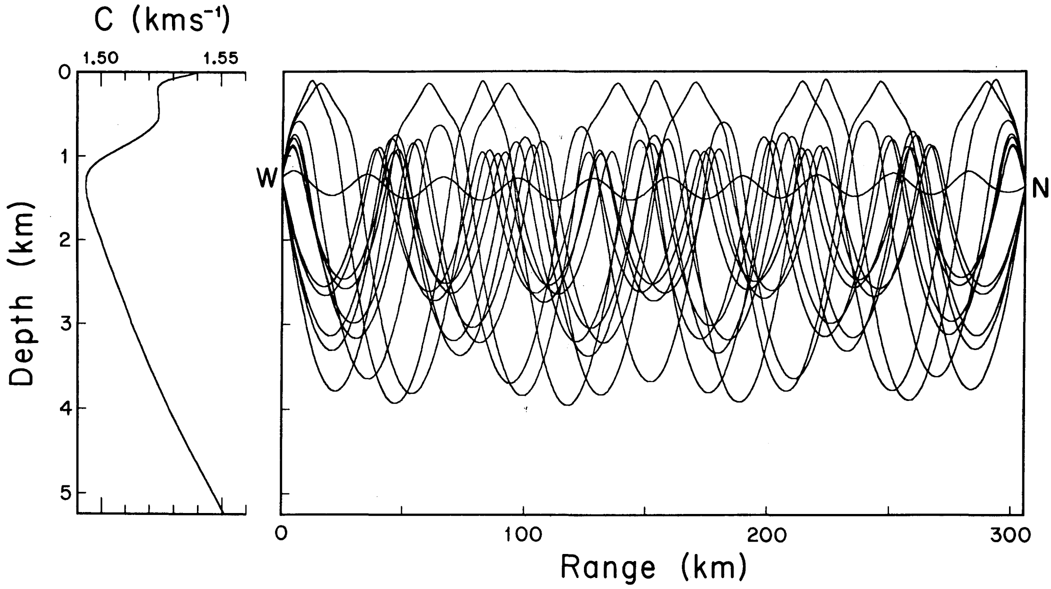

FIG. 11. (Left) Reference sound speed profile $C_0(z)$. This range-averaged profile was obtained by smoothing fairly deep historical data with profiles obtained every 20 km to 1800 m depth. (Right) Corresponding eigenrays. Note the ray groups with shallow and deep upper turning points as well as the axial ray (from Howe *et al.*, 1987).

TICS, UNDERWATER). The effect is to produce a waveguide; sound in the waveguide or channel is constantly being refracted back toward the axis of the channel and travels along many quasisinusoidal rays (Fig. 11). Typical ray "wavelengths" are 50 km. The sound leaves an omnidirectional sound source in all directions; only sound leaving at specific launch angles along eigenrays (Fermat paths) will intersect a receiver. The energy from a single transmitted pulse will be split up among the various rays and will arrive at the receiver at different times, not only because the different rays have different lengths but because they go through regions of the ocean with different sound speed. A typical arrival pattern is shown in Fig. 12.

The success of the method depends on several essential features of ocean acoustic propagation. Separate ray arrivals must be resolved, the arrival pulse must be identified with a known ray path, and the ray paths must be stable over the duration of the experiment. A 1979 experiment verified that arrival patterns remained stable over long periods of time; mesoscale and other oceanic processes merely distorted but did not confuse the pattern (Spiesberger *et al.*, 1980). The small changes in the travel times of the ray arrival peaks are the data for tomography. Typical changes in travel time due to sound speed perturbations and currents are 100 and 10 ms, respectively.

In the geometric optics approximation the travel time along a ray path Γ_i in the presence of a current $\mathbf{u}(\mathbf{x}, t)$ is

$$T_i(t) = \int_{\Gamma_i} \frac{ds}{C(\mathbf{x}, t) + \mathbf{u}(\mathbf{x}, t)\cdot\boldsymbol{\tau}}, \tag{7}$$

where $\boldsymbol{\tau}$ is a unit vector tangent to the ray and s is arc length. The sign of $\mathbf{u}\cdot\boldsymbol{\tau}$ depends on the direction of propagation, and the travel times in opposite directions differ because of the effects of currents. The ray paths Γ_i are obtained using a numerical eigenray code. Equation (7) is analogous to Eq. (1) but is for one ray rather than a multitude of parallel paths.

The travel-time path integral [Eq. (7)] is nonlinear in propagation velocity. Defining a reference sound speed field $C_0(\mathbf{x})$, one has

$$C(\mathbf{x}, t) = C_0(\mathbf{x}) + \delta C(\mathbf{x}, t), \tag{8}$$

where the perturbation sound speed $\delta C(\mathbf{x}, t) \ll C_0(\mathbf{x})$. Substituting in Eq. (7), forming perturbation travel times, and linearizing gives

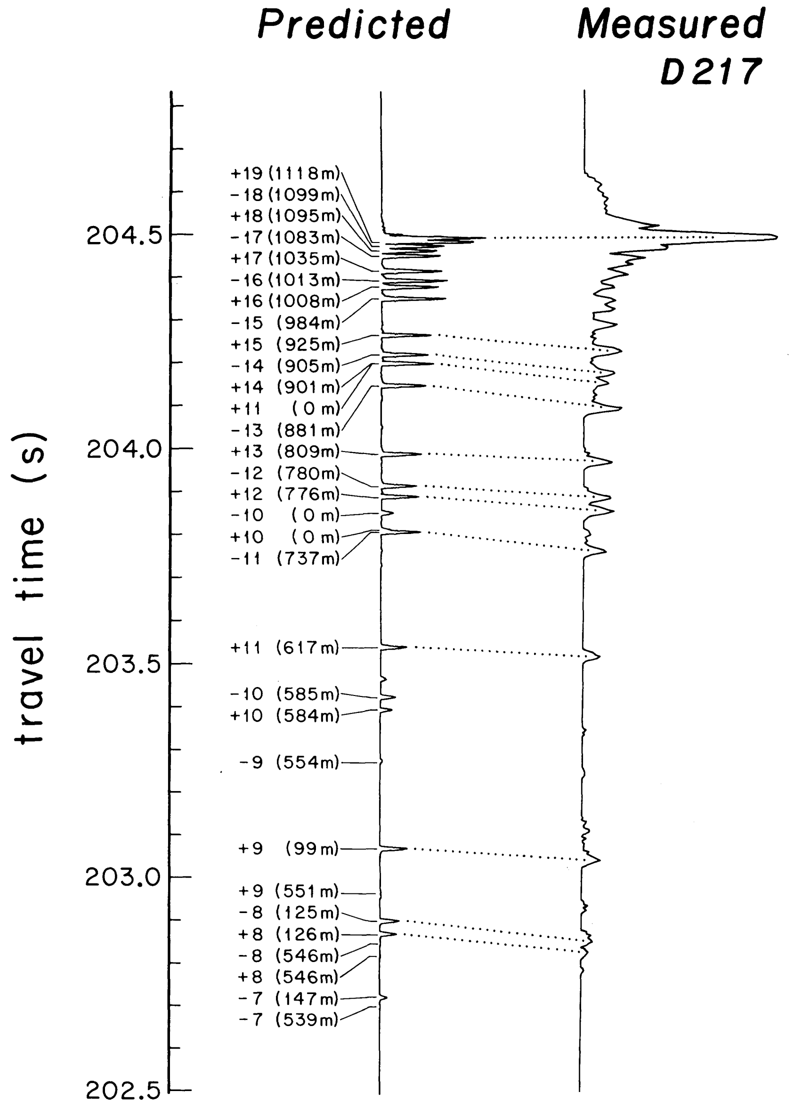

FIG. 12. Comparison of the predicted and measured arrival patterns. The patterns represent sound pressure amplitude at a hydrophone (arbitrary units) vs time. The measured arrival pattern is an average for day 217. The predicted arrival pattern was calculated from $C_0(z)$ using the WKBJ method of Brown (1982). Geometric eigenray arrivals are labeled $\pm n(\hat{z}^+)$ where positive (negative) rays depart upward (downward) from the source, have a total number of upper and lower turning points n, and have upper turning depths $\hat{z}^+$ (from Howe *et al.*, 1987).

$$\delta T_i^+ = T_i^+ - T_{0i} = -\int_{\Gamma_{0i}} \frac{[\delta C(\mathbf{x}, t) + \mathbf{u}(\mathbf{x}, t)\cdot\boldsymbol{\tau}]}{C_0^2(\mathbf{x})}\, ds, \tag{9}$$

$$\delta T_i^- = T_i^- - T_{0i} = -\int_{\Gamma_{0i}} \frac{[\delta C(\mathbf{x}, t) - \mathbf{u}(\mathbf{x}, t)\cdot\boldsymbol{\tau}]}{C_0^2(\mathbf{x})}\, ds. \tag{10}$$

The superscript plus (minus) refers to propagation in the $+(-)$ $\mathbf{x}$ direction; $\boldsymbol{\tau}$ is negative in the second case since propagation is in the $-\mathbf{x}$ direction. The ith reference ray path Γ_{0i} is determined in the reference ocean $C_0(\mathbf{x})$. The corresponding reference travel time is

$$T_{0i} = \int_{\Gamma_{0i}} \frac{ds}{C_0(\mathbf{x})}. \tag{11}$$

Forming sums and differences gives

$$s_i = \tfrac{1}{2}(\delta T_i^+ + \delta T_i^-) = -\int_{\Gamma_{oi}} \frac{\delta C(\mathbf{x}, t)}{C_0^2(\mathbf{x})}\, ds, \qquad (12)$$

$$d_i = \tfrac{1}{2}(\delta T_i^+ - \delta T_i^-) = -\int_{\Gamma_{0i}} \frac{\mathbf{u}(\mathbf{x}, t)\cdot\boldsymbol{\tau}}{C_0^2(\mathbf{x})}\, ds. \qquad (13)$$

The problem has separated: s_i is linearly related to the sound speed perturbation, while d_i is linearly related to the water velocity $\mathbf{u}\cdot\boldsymbol{\tau}$ along the ray path.

4.2 Inverse Problem

Equations (12) and (13) constitute the acoustic forward problem: Continuous δC and $\mathbf{u}$ fields are used to predict a finite data set s_i and d_i. The corresponding inverse problem uses the measured data to estimate the continuous fields. The problem is an underdetermined one because values of the δC and $\mathbf{u}$ fields must be specified at all $\mathbf{x}$ using only a finite number of data. To proceed, the fields δC and $\mathbf{u}$ are parametrized with a finite number of discrete parameters by a model that is oceanographically meaningful. This procedure has three advantages: The parametrization can be very efficient, the inverse solution can become more accurate because the model injects our oceanographic knowledge concerning permissible structure of the fields (reducing the degrees of freedom), and the procedure can be used to evaluate data-model consistency.

Current models are represented by a linear combination of basis functions describing the horizontal (x, y) and vertical (z) structures,

$$\delta C(x, y, z) = \sum_{i,j,k} a_{ijk}\, X_i(x) Y_j(y) Z_k(z). \qquad (14)$$

The coefficients a_{ijk} are the model parameters we are trying to determine. The horizontal dependence is often taken to be a truncated Fourier series or, even more simply, pixels. The functions describing the vertical dependence $Z_k(z)$ are chosen to efficiently parametrize expected variations in the ocean. Typically they are functions based on oceanographic theory or data which describe the dominant modes of variability (see OCEANOGRAPHY).

To obtain estimates of the model parameters and thus the sound speed anomaly from the measured travel times, the "stochastic inverse" of geophysics (see GEOPHYSICS) is used (Aki and Richards, 1980). The method assumes that an *a priori* estimate of the mean and covariance of the unknown field exists, and constructs an estimator that minimizes the square of the difference between the estimate of the unknown field and the true field. In the overdetermined case, this reduces to the weighted least squares fit where the sum of the squares of the data residuals (weighted by the inverse of the data error) is minimized and the *a priori* information is not necessary. In the mixed or underdetermined case, the problem is constrained by adding the *a priori* information about the model parameters. In the absence of any data, the best estimate of the field is simply the *a priori* estimate of the mean, and the associated error is just the *a priori* estimate of the covariance. The stochastic inverse provides a rational way of combining measurements and *a priori* information. Such a method is almost always necessary in geophysical, data poor, situations.

The generalized forward problem can be written

$$\mathbf{d} = \mathbf{G}\mathbf{m} + \mathbf{n}. \qquad (15)$$

The vector $\mathbf{d}$ contains the data, for instance the s_i in Eq. (12), and $\mathbf{m}$ contains the "true" model parameters [the a_{ijk} in Eq. (14)]. $\mathbf{G}$ is the matrix of partial derivatives $G_{ij} = \partial d_i/\partial m_j$ relating the model parameters to the data, obtained by substituting the model Eq. (14) into Eq. (12) and differentiating, while measurement noise and uncertainty in the forward problem physics and model are incorporated in the noise vector $\mathbf{n}$. The stochastic inverse operator is given by (Aki and Richards, 1980)

$$\mathbf{L} = \mathbf{W}\mathbf{G}^T(\mathbf{G}\mathbf{W}\mathbf{G}^T + \mathbf{E})^{-1}. \qquad (16)$$

The matrices $\mathbf{W}$ and $\mathbf{E}$ are the model and data error covariances, respectively. The estimated model parameters $\hat{\mathbf{m}}$ are given by

$$\hat{\mathbf{m}} = \mathbf{L}\mathbf{d}. \qquad (17)$$

The model parameters are a linear combination of the data.

In a mixed or underdetermined problem it is important to have a measure of how well each model parameter is resolved by the data. The resolution matrix $\mathbf{R}$ is defined by

$$\hat{\mathbf{m}} = \mathbf{R}\mathbf{m}. \qquad (18)$$

The $\mathbf{G}$ matrix along with the covariance matrices enter into the calculation of the resolution matrix (Menke, 1989). The resolution matrix shows how well the combined measuring system and the reconstruction method are able to determine each model parameter. Ideally **R** is the identity matrix so that the estimated model parameters $\hat{\mathbf{m}}$ equal the true model parameters **m**. Unfortunately, this is rarely the case.

If a Fourier spectral model is being used, the resolution matrix will tell how well each Fourier coefficient is resolved, i.e., how well Fourier space is filled, in a way analogous to the Projection Slice theorem. If a simple pixel model is used, one can directly obtain the impulse response, i.e., how well a unit perturbation in one pixel is reconstructed, and how the reconstruction depends on the values in all the other pixels.

When data are time dependent, one can use a Kalman filter to assimilate new data. The field at the new time is first estimated from the old field using either a rigorous physical model, such as a three-dimensional, primitive equation numerical model, or *ad hoc* methods such as letting the field decay to the *a priori* values with some characteristic time scale. The error covariance is estimated in a similar fashion. The inverse is then performed with these estimates of the mean and covariance of the field, producing a new estimate of the mean and error covariance.

Over long ranges it is no longer valid to assume the ray paths in the reference ocean are "close" to the ray paths in the actual ocean. This introduces a nonlinearity, the degree of nonlinearity being proportional to range and the average eddy energy along the path (Spiesberger, 1985; Munk and Wunsch, 1987). For a horizontal array of instruments with many crossing paths in the horizontal, the first inversion should be able to place the eddies in approximately the right location, at least within a fraction of a ray wavelength (implying some resolution at all wave numbers). Simple iteration should converge to the correct solution. For a single-slice experiment, there is some information in the travel times about range dependence, but it is not clear whether there will be enough information for an iteration to converge.

In the source localization problem (e.g., finding submarines), one takes the entire measured wave form over a large spatial array and solves for source position; this is called matched field processing and incorporates aspects of matched filtering, backpropagation, and holography (Baggeroer *et al.*, 1988). This is done by assuming various source positions and predicting the wave form at the array until the difference between the predicted and measured wave forms is small. This technique is in its infancy and will most likely have to include sound speed variability. In the typical ocean tomography case, the instrument positions are usually known to within a few tens of meters and this position residual can be included as a model parameter in the (linear) inverse. In the localization problem, the position is not known to even a convergence zone, making the problem very nonlinear with many local solutions. The links between the tomography and localization problems are being actively pursued.

4.3 Examples of Ocean Tomography

An ocean acoustic tomography system requires instruments capable of sending and receiving acoustic pulses whose duration is short enough to resolve separate ray arrivals, that are precisely timed to allow absolute measurement of travel time, and are strong enough to be received above the ambient noise. A short duration pulse implies a broad bandwidth signal; present sources have a bandwidth approaching 100 Hz which gives a pulse width of 10 ms. Precision time keeping (1 ms/yr) is accomplished by using rubidium frequency standards in the underwater instruments. Even the loudest acoustic sources available (193 dB re 1 μPa at 1 m) do not produce enough energy in a 10 ms pulse to be heard at any significant distance. The remedy has been to transmit a long, coded signal, spreading the signal energy out in time, and then in the signal processing to recombine the energy to make it appear as if it all had been sent in one loud pulse. The signals sent are pseudorandom, phase-modulated sequences; the processing is called replica cross correlation, which is closely related to matched filtering. (The Hadamard transform is used to do this efficiently.) A signal processing gain of 35 dB is typical.

One of the tomography transceivers presently being used is shown in Fig. 13. The instruments are deployed on single-point, subsurface oceanographic moorings. Although the moorings are designed to be quite stiff,

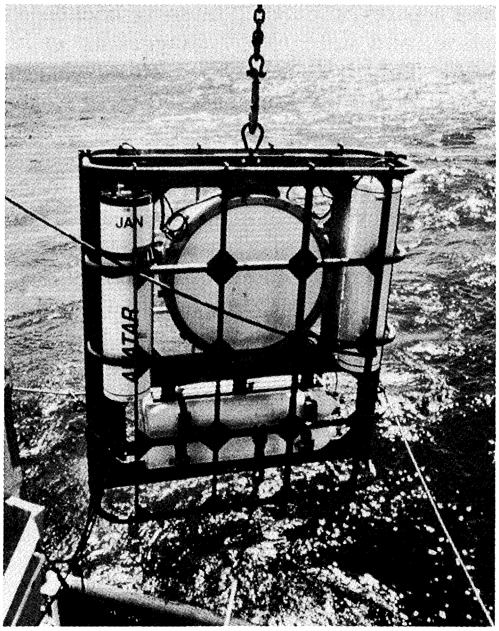

FIG. 13. Tomography transceiver used in the 1986/1987 Gyre-Scale Reciprocal Transmission Experiment. This is a second generation transceiver, using an HLF-5 acoustic source from Hydroacoustics Inc. and a newly designed set of electronics. The high-pressure nitrogen bottles in the bottom of the cage are used to pressure compensate the source at its operating depth of approximately 1000 m. The cage is 2.4 m high, 2.1 m wide, and about 1 m thick; it weighs roughly 1590 kg in air and 590 kg in water (from Munk and Worcester, 1988; © 1988, The Oceanography Society).

their motion in response to ocean currents is measured with a long-baseline acoustic navigation system to account for changes in range between the instruments.

The first ocean tomography experiment, billed a "demonstration," was conducted southwest of Bermuda in 1981. Four sources and five receivers were deployed on the perimeter of a 300-km square. Sources with a center frequency of 224 Hz and a bandwidth of 10 Hz were used, producing pulses with 100 ms widths. This was less than optimal, but these sources were the only ones available at the time. Nevertheless, there were sufficient data to construct maps of the sound speed field (Fig. 14) (Cornuelle *et al.*, 1985).

Travel time differences between co-located sources and receivers can be used as a measure of currents (Fig. 15). The signal-to-noise ratio of the travel time difference depends on how nearly reciprocal the paths in each direction are; if the paths are exactly reciprocal, then the effects of internal waves on the one-way travel times (3–10 ms rms) will cancel. In the 1983 reciprocal transmission experiment over 300 km, the reciprocal noise variance was smaller than the one-way noise variance by a factor of 10.

More recently, three tomography transceivers were deployed in the northeast Pacific in a 1 Mm triangle for four months. The primary purpose of the experiment was to measure large-scale gyre vorticity or swirl: By measuring the current velocity along each leg of the triangle using reciprocal transmissions, one can calculate the circulation around the triangle, $\int \mathbf{u} \cdot d\mathbf{l}$. This is in turn proportional to the average vorticity by Stokes's theorem. The vorticity is a desirable quantity because it is the basis for much of theoretical oceanography, yet it has never been measured directly on large scales. This experiment exploits the integrating (projection) nature of the method to filter out the mesoscale features, leaving the gyre-scale features of interest.

There are several ocean tomography projects in progress. One involves six moorings deployed in September 1988 in the Greenland Sea for one year. This area far to the north is interesting because it is thought that very cold, dense water forms during the winter in small-scale overturning events, then sinks, and finally spreads throughout the world ocean. This has never been directly observed; it is hoped that the tomography data will show indications of these events. A complicating feature of acoustic propagation in polar waters is that the sound speed monotonically increases with depth producing a surface-bounded waveguide; all rays are surface reflected. Acoustic energy propagating within several hundred meters of the surface is not well modeled by ray theory. The dual of rays, normal modes, is necessary to describe the propagation.

The moorings in the Greenland Sea are also being used for engineering tests of moving ship tomography (Cornuelle *et al.*, 1989). A simulated geometry is shown in Fig. 16. The idea in this application is to obtain a "snapshot" of the ocean by circumnavigating the moored array, dipping a hydrophone array every 4 h or 70 km to listen to the sources, and thereby building up a net of many rays. This is an obvious analogy to x-ray tomogra-

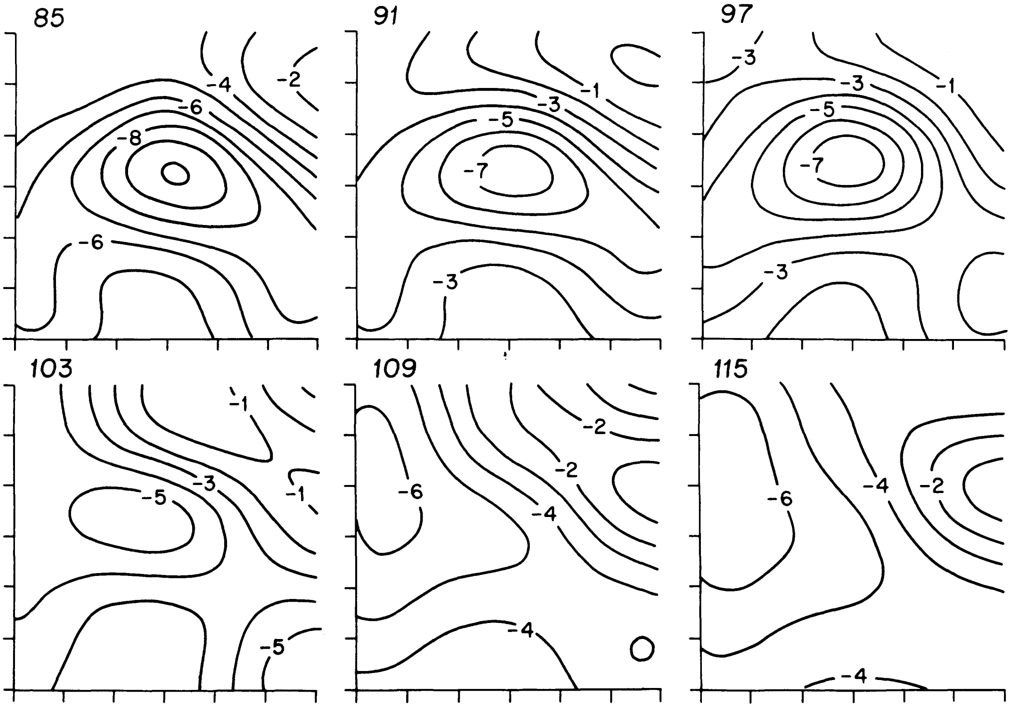

FIG. 14. The acoustic travel time data obtained in the 1981 Ocean Acoustic Tomography Experiment were used to draw maps of the sound speed field at 700 m depth in the 300 km by 300 km area of the experiment (each panel). Maps at 6-day intervals from year day 85 through year day 115 show a cold eddy with low sound speed remaining stationary for several days and then moving rapidly out of the experiment area beginning on year day 103. The sound speed contours relative to 1506 m/s are drawn at intervals of 1 m/s, which roughly corresponds to a temperature change of 0.2 °C. Maps at other depths can also be constructed from the acoustic data (from Munk and Worcester, 1988; ©1988, The Oceanography Society).

phy. For this technique to succeed, it is necessary to locate the receiving hydrophones to better than 10 m accuracy. Two acoustic tracking systems are being used in conjunction with the NAVSTAR Global Positioning System, a very long-baseline system using satellites.

Munk and Forbes (1989) have proposed placing a source off Heard Island south of Australia and listening to it as far away as Bermuda and California. This idea is motivated by the reanalysis of a 1960 experiment in which a 140-kg explosive charge was detonated off Perth, Australia, and heard in Bermuda some four hours later. Such long-range transmissions will average out much of the ocean variability, and it is possible that changes in travel time will be proportional to global-scale changes in ocean temperature. Munk and Forbes calculate that a 4 mdeg C/yr change corresponds to a 100 ms/yr change in travel time for a Heard Island–California path. A major unanswered question is whether the forward problem can be adequately solved so that a quantitative inversion can be made. Ultimately, a tomographic network is envisioned where sources and receivers will be distributed around the world, continually monitoring the global ocean.

5. CONCLUDING REMARKS

The ultimate physical limit to spatial resolution in any tomography experiment is the acoustic wavelength. The image quality and accuracy achieved in a particular experiment will depend on a multitude of factors including the number of transducers, the number of views, the size of the viewing aperture, and the signal-to-noise ratio. All of these are controlled to a large extent by practical considerations such as geometry, signal processing, and cost (Schueler *et al.*, 1984).

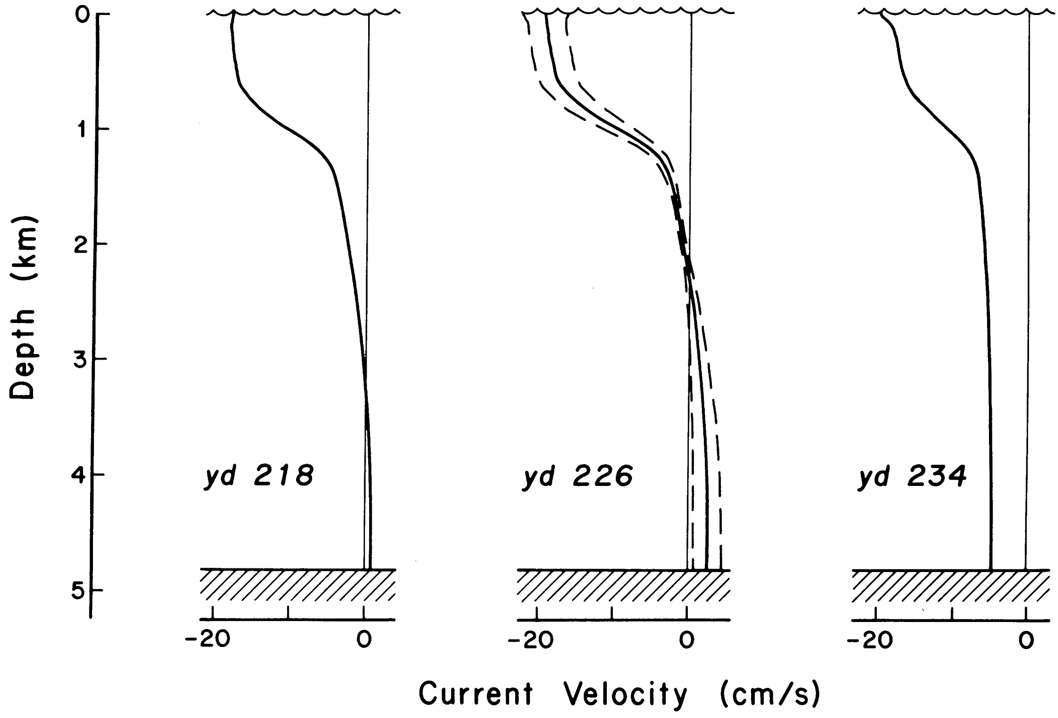

FIG. 15. The differences in travel times between sound pulses traveling in opposite directions in the 1983 Reciprocal Acoustic Transmission Experiment were used to compute the average current velocity in the 300-km-long vertical slice between the transceiver moorings as a function of depth. Three profiles obtained at 8-day intervals show the slow changes that occur in the current profiles in the ocean. All have stronger currents near the surface than at great depths, as is typical of ocean currents. The calculated errors in the profiles are approximately 1 cm/s, as shown by the dashed lines around the central profile. Note the change from a baroclinic mode structure on year day 218 to a strong barotropic (top-to-bottom) component on year day 234 (from Munk and Worcester, 1988; ©1988, The Oceanography Society).

However, the accuracy of any quantitative reconstruction will be directly related to the quality of the forward-problem solution. Any inaccuracies in the forward problem will appear in the inversion or reconstruction. This is the major problem facing diffraction tomography with strong scattering. The Born and Rytov approximations are often not valid, implying that higher order scattering needs to be taken into account as well as three-dimensional effects. This will make the problem nonlinear and thus even more difficult. With desktop minisupercomputers becoming more common, brute force numerical approaches may be appropriate.

Straight-ray tomography is relatively easy to perform, but again the accuracy of the forward problem must be checked. For instance, a ray trace through the reconstructed sound-speed field should be done to see if the differences in travel times along the curved rays are significantly different from those along the straight rays, or whether diffraction effects are important.

Ocean tomography is just evolving from the testing and demonstration stage and is at the point where useful oceanographic measurements are starting to be made. Current research is focused on improving horizontal resolution by making use of the special nature of the ocean waveguide and the ocean structure (Cornuelle and Howe, 1987), recasting the forward/inverse problems in terms of acoustic normal modes, creating synthetic apertures with moving instruments, and using long range transmissions aimed at basin and global monitoring.

Many difficult problems remain to be solved before acoustic tomography becomes a common tool of scientists and engineers. It is clear, though, that acoustic methods can yield information that is either unobtainable any

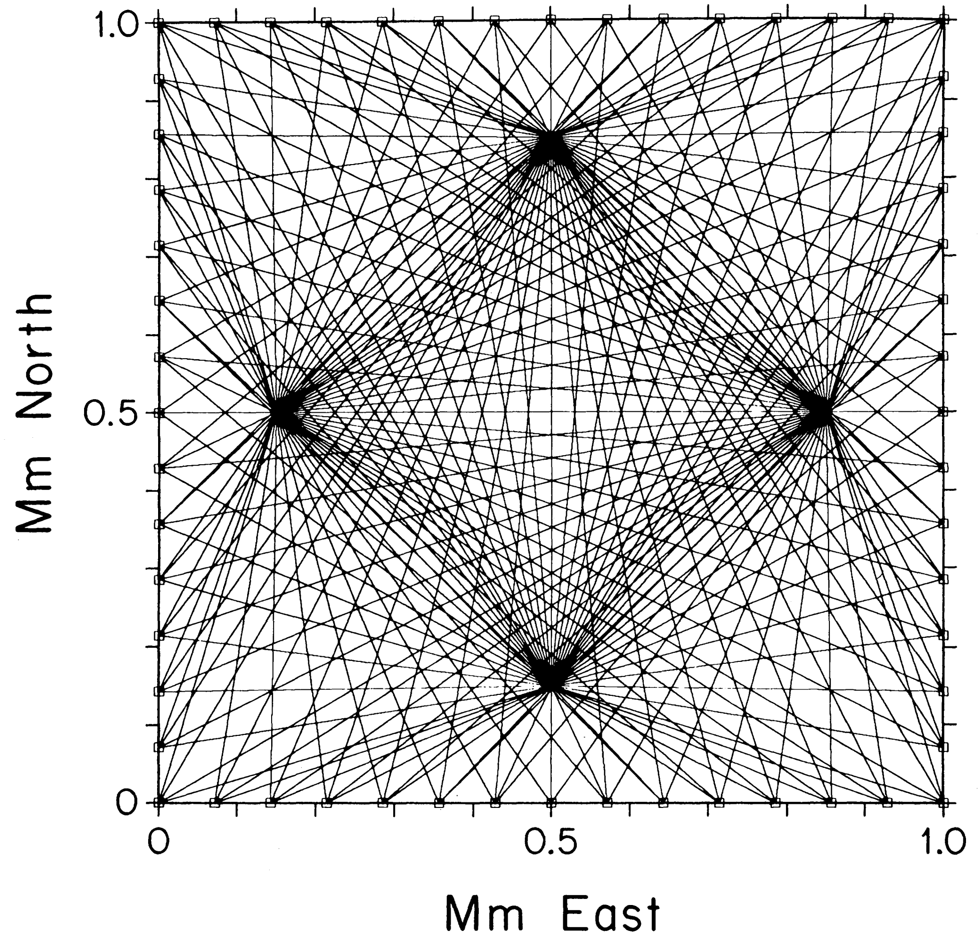

FIG. 16. Plan view of the ray geometry in a simulated moving-ship tomography experiment. Four sources are located on a 350-km-radius circle about the center of a 1000-km by 1000-km square. Receivers are located at approximately 70-km intervals around the square. In this simulation the ocean is assumed to be two dimensional, with no vertical structure; straight-line ray paths can therefore be used in the calculation. The inversion procedure (using realistic travel-time and positioning errors) leads to mapping which accounts for 95% of the mesoscale variance. Subsequent 3-D simulations with vertical multipaths lead to similar mapping errors. In this preliminary simulation all receptions are assumed to be simultaneous. The actual experiment will be conducted by lowering a receiver from a ship at roughly 70-km intervals as it transits around the square. This involves "nowcasting" in a 4-D (space + time) environment (from Munk and Worcester, 1988; ©1988, The Oceanography Society).

other way or complementary to results from other methods.

GLOSSARY

Born Approximation: Linearizing approximation in scattering theory. Approximates unknown total field in a scatterer by the known incident field.

Diffraction: Deviation of energy from straight ray paths due to effects other than reflection or refraction. Occurs when size of irregularities in a medium become comparable to the wavelength of the probing wave.

Forward Problem: Predicting data given the parameters which affect the data.

Fourier Diffraction Theorem: Fundamental theorem of diffraction tomography. Fourier-domain relationship between measured data and unknown scatterer.

Inverse Problem: Determining the parameters which affect measured data.

Projection-Slice Theorem: Fundamental theorem of conventional straight-ray tomography. Fourier-domain relationship between measured data and unknown object.

Ray: A curve whose tangent at any point lies in the direction of propagation of a wave.

Refraction: The bending of a wave as it passes through regions with varying wave velocities or indexes of refraction.

Tomography: Imaging procedure that uses multiple views to produce quantitative cross-sectional reconstructions of an object.

Waveguide: A physical situation which constrains or guides the propagation of waves by reflection off solid boundaries and/or by refraction within the medium.

List of Works Cited

Aki, K., Richards, P. G. (1980), *Quantitative Seismology: Theory and Methods*, San Francisco: Freeman.

Baggeroer, A. B., Kuperman, W. A., Schmidt, H. (1988), *J. Acoust. Soc. Am.* **83**, 571–587.

Brown, M. G. (1982), *J. Acoust. Soc. Am.* **71**, 1427–1432.

Cornuelle, B., Howe, B. M. (1987), *J. Geophys. Res.* **92**, 11 680–11 692.

Cornuelle, B., Munk, W., Worcester, P. (1989), *J. Geophys. Res.* **94**, 6232–6250.

Cornuelle, B., Wunsch, C., Behringer, D., Birdsall, T., Brown, M., Heinmiller, R., Knox, R., Metzger, K., Munk, W., Spiesberger, J., Spindel, R., Webb, D., Worcester, P. (1985), *J. Phys. Oceanogr.* **15**, 133–152.

Daily, W. D. (Ed.) (1986), *Proc. IEEE* **74**, 328–362.

Devaney, A. J. (1982), *Ultrason. Imaging* **4**, 336–350.

Devaney, A. J. (1983), *IEEE Trans. Biomed. Engr.* **BME-30**, 377–386.

Duchêne, B., Lesselier, D., Tabbara, W. (1988), *IEEE Trans. Ultrasonics, Ferroel., Freq. Control* **35**, 437–444.

Horn, B. K. P. (1978), *Proc. IEEE* **66**, 551–562.

Howe, B. M., Worcester, P. F., Spindel, R. C. (1987), *J. Geophys. Res.* **92**, 3785–3805.

Ishimaru, A. (1978), *Wave Propagation and Scattering in Random Media*, New York: Academic.

Kak, A. C., Slaney, M. (1988), *Principles of Computerized Tomographic Imaging*, New York: IEEE.

Menke, W. (1989), *Geophysical Data Analysis: Discrete Inverse Theory*, 2nd ed. Orlando, Florida: Academic.

Munk, W., Forbes, A. (1989), *J. Phys. Oceanogr.* **19**, 1765–1778.

Munk, W. H., Worcester, P. F. (1988), *Oceanography Magazine* **1**, 8–10.

Munk, W., Wunsch, C. (1979), *Deep-Sea Res.* **26A**, 123–161.

Munk, W., Wunsch, C. (1987), *Geophys. Astrophys. Fluid Dyn.* **39**, 1–24.

Pan, S. X., Kak, A. C. (1983), *IEEE Trans. Acoust. Speech Sig. Process.* **ASSP-31**, 1262–1275.

Porter, R. P. (1989), "Generalized Holography with Applications to Inverse Scattering and Inverse Source Problems," in: E. Wolf (Ed.), *Progress in Optics 27*, New York: Elsevier.

Robinson, B. S., Greenleaf, J. F. (1985), "Computerized Ultrasonic Tomography," in: R. A. Robb (Ed.), *Three-dimensional Biomedical Imaging*, Vol. 2. Boca Raton: CRC.

Rouseff, D. (1989), "Diffraction Tomography and the Image Projection Technique," Ph.D. Dissertation, University of Washington, Seattle.

Schreiman, J. S., Gisvold, J. J., Greenleaf, J. F., Bahn, R. C. (1984), *Radiology* **150**, 523–530.

Schueler, C. F., Hua, L., Wade, G. (1984), *IEEE Trans. Sonics Ultrasonics* **SU-31**, 195–217.

Spiesberger, J. L. (1985), *J. Acoust. Soc. Am.* **77**, 83–100.

Spiesberger, J. L., Spindel, R. C., Metzger, K. (1980), *J. Acoust. Soc. Am.* **67**, 2011–2017.

Spindel, R. C. (1986), in: T. McGuinness and H. H. Shih (Eds.), *Current Practices and New Technology in Ocean Engineering*, OED-Vol. 11. New York: ASME, pp. 7–13.

Wolf, E. (1969), *Opt. Commun.* **1**, 153–156.

Further Reading

Schueler, C. F., Hua, L., Wade, G. (1984), *IEEE Trans. Sonics and Ultrasonics* **SU-13**, 195–217.

Kak, A. C., Slaney, M. (1988), *Principles of Computerized Tomographic Imaging*, New York: IEEE.

Acoustic Holography, Vols. 1–7; *Acoustic Imaging*, Vols. 8–16. Annual series containing current research. Various editors, New York: Plenum.

Kaveh, M., Soumekh, M. (1986), "Computer Assisted Diffraction Tomography," in: H. Stark (Ed.), *Image Recovery: Theory and Application*, New York: Academic.

ACOUSTICS, APPLIED

See ENGINEERING ACOUSTICS

ACOUSTICS, ARCHITECTURAL

David P. Walsh and Thomas R. Norris, *Walsh ♦ Norris and Associates, Inc., San Francisco, CA*

	Introduction	109
1.	**Room Acoustics**	110
1.1	Sound in Enclosed Spaces	110
1.1.1	Room Modes	111
1.1.2	Steady-State Sound Field	112
1.1.3	Decay of the Sound Field	113
1.2	Sound Reflections	113
1.2.1	Geometrical Reflections	113
1.2.2	Diffraction and Diffusion	114
1.3	Sound Absorption	115
1.3.1	Sound Absorbing Properties of Materials	115
1.3.1.1	Porous Absorber	115
1.3.1.2	Resonant Panel Absorber	116
1.3.1.3	Resonant Air Cavity Absorber	117
1.3.2	Measurement Procedures	118
1.3.3	Absorption Coefficients	120
1.3.4	Noise Reduction by Sound Absorbing Materials	120
1.3.5	Air Absorption	120
1.4	Reverberation Time	121
1.4.1	Reverberation Equations	121
1.4.1.1	Sabine Equation	121
1.4.1.2	Eyring Equation	121
1.4.1.3	Fitzroy Equation	121
1.4.2	Design Criteria for Enclosed Spaces	122
1.4.3	Assisted Resonance Systems	122
1.5	Auditorium and Enclosure Design	123
1.5.1	Sound Reflection Design	123
1.5.1.1	Early versus Reverberant Sound	123
1.5.1.2	Lateral Reflections	124
1.5.2	Subjective Preference Scaling	124
1.5.3	Model Testing and Computer Simulation	124
2.	**Sound Isolation**	125
2.1	Airborne-Sound Isolation	125
2.1.1	Transmission Coefficient and Transmission Loss	125
2.1.2	Homogeneous Walls	126
2.1.2.1	Stiffness-Controlled Transmission Loss	126
2.1.2.2	Resonance-Controlled Transmission Loss	126
2.1.2.3	Mass Law	127
2.1.2.4	Critical Frequency	127
2.1.3	Multiple-Skin Walls	127
2.1.4	Flanking Transmission	128
2.1.5	Sound Transmission Class	129
2.2	Structure-Borne Sound Isolation	129
2.2.1	Impact Insulation Class	129
2.3	Measurement Procedures	130
2.3.1	Airborne-Sound Isolation	130
2.3.2	Impact-Sound Isolation	131
	Glossary	131
	Works Cited	134
	Further Reading	134

INTRODUCTION

The basic concern of architectural acoustics is to provide a satisfactory acoustical environment in any space of a building. Essentially, this requires an integration of the physical and technical aspects of acoustics and building technology with the psychological and subjective requirements of the users of the building. On the technical side, the physical behavior of sound both in a space and in the building structure itself must be understood. On the subjective side, it is necessary to understand how the users of a building react to an acoustical environment in any of a number of situations including privacy, intelligibility, annoyance, task interference, sleep disturbance, enjoyment of the performing arts, etc. Architectural acoustics must therefore relate human responses to the physical stimuli of sound, and provide such information to the architects, designers, and engineers responsible for a building.

The two basic areas of architectural acoustics discussed in this article are room acoustics and sound isolation. Room acoustics in-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

volves the characteristics of sound in the air, together with the effects on the sound by the shape, size, and materials of the boundary surfaces of the space. Sound isolation deals with the manner in which sound is transmitted between spaces in a building. In each of these areas, both the physical characteristics of, and the human response to, sound are discussed. This article concentrates on the characteristics of sound in buildings and its effect on building technology, and, as such, does not include such topics as the basics of sound which are covered in detail in other articles of this publication.

Architectural acoustics as a specific discipline had its beginnings at the turn of this century with Wallace C. Sabine at Harvard University, whose work primarily involved the development and understanding of reverberation. He was the first to develop relationships between the objective measurements of sound in spaces and criteria of human judgments. The field of architectural acoustics, however, did not show any significant advances from the late 1920s until the 1950s. During the 1930s extensive research was completed in acoustics, but on the physical aspects of sound, primarily supported by the communications industry. In the 1940s the emphasis in acoustics was on war-related activities such as sonar, radar, speech perception in high noise levels, etc. The post-war building boom of the 1950s finally gave architectural acoustics the opportunity to expand but, due to the lack of actual building experience, it was difficult for those in the field to apply effectively all of the acoustical research that had been completed over the prior 50 years. This was caused by the practical limitations on advances in architectural acoustics which are directly related to the building industry itself. Since a typical building project can take from five to ten years from design to occupancy, advances in the field tended to be slower than in other applications of acoustics. However, by the 1970s acoustical consultants had achieved extensive experience with a large number of buildings and building types. At present, significant research in architectural acoustics continues both by practitioners in the field as well as by university researchers, based not only on experience with actual buildings, but also on the advances in model testing and computer simulation in which theoretical and analytical studies can be applied without the expense and long time delay associated with the construction process.

Public awareness of the effects of noise has also played an important role in the development of architectural acoustics as evidenced by the numerous design standards currently mandated by such agencies and authorities as the Department of Housing and Urban Development, the Uniform Building Code, various Departments of Public Health, Department of Transportation, local building codes and ordinances, etc. In current practice, architectural acoustics is an important part of the building design process, and its influence extends to a wide variety of building types including multifamily dwellings, hotels, office buildings, work environments, educational facilities, hospitals, etc., rather than just to performing-arts projects, which traditionally have required architectural acoustics design assistance.

1. ROOM ACOUSTICS

Room acoustics is primarily involved with sound in three dimensions within an enclosed space. It is therefore concerned not only with sound traveling in the air, but also with the effects on sound by the enclosure, which include the influences of the size, shape, reflection, and absorption characteristics of the boundary surfaces. Based on the type of information desired in studying the acoustics of enclosures, two basic approaches for analyzing sound can be used. For situations in which transient events are important, as in rooms for speech and music, time domain analysis is the most informative. Where a steady-state sound field is important, as in noise isolation situations, frequency domain analysis is the most useful. The human ear may be considered as a frequency-sensitive pressure transducer with a sufficiently wide bandwidth so as to be very responsive to transients. The ear is also quite directionally sensitive, particularly in determining azimuth along the horizon.

1.1 Sound in Enclosed Spaces

A large space has dimensions that are typically many times the wavelength of any sound wave being considered. Therefore, a sound source will excite many normal modes of vibration in the sound field. The sound field in a large room can be described either in terms of the mean square sound pressure or by the intensity, or net power flow per unit area, of the sound waves that travel around the room. Since pressure is a scalar quantity it can be measured using a single pressure-sensitive microphone. Intensity, on the other hand, is a vector quantity and requires directional information, with measurements usually utilizing a computer plus either two pressure microphones or one pressure microphone and a velocity-sensitive microphone. (Ordinary directional microphones used in the entertainment field give the total, but not the net, sound power flow in a given direction.) Both the mean square sound pressure and the intensity are influenced by the sound source and by the properties of the enclosure. Near a sound source the sound field has a relatively high level, and in such regions there is an identifiable flow of energy away from the source. This is the direct field. Far from the source, the direct sound becomes weak in comparison to the reverberant sound, which consists of the energy reflected from the various surfaces of the room. This is the reverberant field.

For example, the first sound that a listener receives in a room is the sound that travels directly from the source to the listener. It will arrive at a time

$$t_0+r_d/c, \tag{1}$$

where t_0 is the time (in ms), r_d is the distance (in m), and c is the speed of sound (in m s^{-1}). The intensity of the sound will be equal to

$$I_0-x_d, \tag{2}$$

where I_0 is the intensity at the source position (in W m^{-2}) and x_d is the attenuation.

The attenuation x_d consists of both the attenuation due to spherical spreading of the sound and that due to the molecular absorption effect of the air. After the direct sound has arrived at the listener there will be a series of sounds that have traveled by other paths which include one or more reflections from the surfaces of the room. Each reflected sound will arrive at a time

$$\sum_{i=1}^{n} (t_0+r_i/c), \tag{3}$$

where r_i is the total distance traveled by the sound reflections.

The intensity of each reflected sound will be

$$\sum_{i=1}^{n} (I_0-x_i), \tag{4}$$

where x_i is the respective attenuation.

The attenuation of these reflected sound waves will not only include the two factors contained in Eq. (2), but will also include the effects of absorption encountered at the boundary surfaces.

1.1.1 Room Modes. When a sound is generated in an enclosure, a sound wave spreads out in all directions from the source until it encounters a boundary surface, where it is reflected off the walls, floor, and ceiling and travels around the room in every direction. If each path that the sound wave travels is traced around the room, there will be certain paths that will be repeated and room resonances or standing waves will occur. Each room has its normal modes of acoustic vibration or resonances, and its normal or resonant frequencies, with the total number of modes asymptotically approaching, at sufficiently high frequencies, a number which is directly proportional to the volume of the enclosure. There are three types of normal modes n in an enclosure of dimensions l_x, l_y, and l_z: axial, in which the sound waves are reflected only by two opposite surfaces (one dimensional; two of the indices n_x, n_y, and n_z are equal to 0); tangential, in which the sound waves are reflected from two pairs of surfaces (two dimensional; one of the indices n_x, n_y, and n_z is equal to 0); and oblique, in which the sound waves are reflected from all boundary surfaces (three dimensional; all of the indices n_x, n_y, and n_z have nonzero values). For simple rectangular spaces it is possible to calculate the frequencies of these room resonances. The characteristic frequencies or normal modes of a rectangular enclosure are given by

$$f_n=(c/2)[(n_x/l_x)^2+(n_y/l_y)^2+(n_z/l_z)^2]^{1/2}, \tag{5}$$

where f_n is the resonant frequency (in Hz); c is the speed of sound in air (in m s^{-1}); n_x, n_y, and n_z identify the mode of vibration and can be any integers, 0,1,2,...; and l_x, l_y, and l_z are the dimensions of the room (in m). At the lower frequencies the room resonances are widely spaced so that with broadband acoustic excitation, it is relatively easy to hear the tonal or spectrum peaks at the resonant frequencies. As the frequency increases the resonances become spaced very close together. In typical spaces with some sound absorption present at the boundary surfaces, the width of the resonant frequency peaks of contiguous normal modes becomes greater than the spacing between peaks, and eventually are so close together that the overlapping peaks coalesce into a continuum and achieve a relatively uniform transmission characteristic. It is interesting to note that for every mode of vibration in a rectangular enclosure the sound pressure is maximum at the corners.

Equation (5) applies only to rectangular spaces. This same type of simple mathematical analysis is not available for more complex room shapes, but although they cannot be predicted easily, resonant frequencies still exist. For example, a change in the shape of a room from rectangularity shifts the frequencies of the normal modes, making them extremely difficult to compute, although the total number of modes remains essentially the same, assuming the room volume remains constant.

1.1.2 Steady-State Sound Field. When a constant sound source begins to operate in a quiet architectural enclosure, it will initially emit energy in all directions and the sound pressure level will increase until it exponentially approaches a steady-state value in which the sound energy being removed from the space in each frequency band is equal to the sound energy produced by the source. The process is complicated by interference between the waves reflected from the room surfaces and the waves emitted directly from the source, thus producing spatial standing waves and also time-domain unevenness in the growth of the sound field which is sometimes heard as audible "beats." At steady state, the source excites normal modes. The degree of mode excitation within each frequency band depends on the acoustic power of the source, the source location with respect to any standing-wave patterns, source acoustic impedance, and the bandwidth of each mode. The direct field exists close to the sound source and the energy flow is away from the source. The steady-state energy density of the direct field is

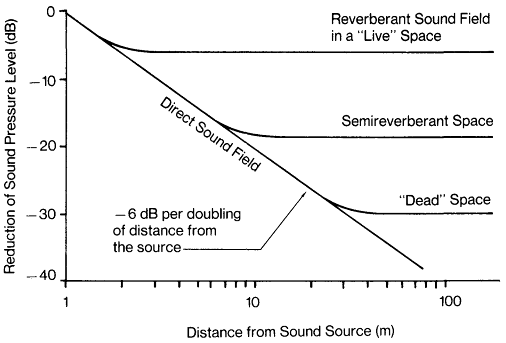

FIG. 1. Effect of sound absorbing boundary surfaces on the reverberant sound field in an enclosure. The direct field in a reverberant (or "live") space dominates for only a short distance out from the sound source, whereas in a highly absorptive (or "dead") space the direct field can predominate for long distances from the sound source.

$$D_0 = WQ_\theta / 4\pi r^2 c, \tag{6}$$

where D_0 is the energy density of the direct field (in W s m^{-3}), W is the acoustic power output of the source (in W), Q_θ is the directivity factor of the source in a particular direction θ (dimensionless), and r is the distance from the source (in m).

As the distance from the source increases the energy density of the direct sound field becomes equal to, and eventually less than, that of the reverberant field. In the reverberant field all of the sound waves have been reflected at least once from the boundary surfaces of the enclosure. The more absorptive the boundary surfaces are, the greater the reduction of the reflected sound energy that will occur, as indicated in Fig. 1. The steady-state energy density of the reverberant field is defined as

$$D_R = 4W/cR, \tag{7}$$

where D_R is the energy density of the reverberant field (in W s m^{-3}), R is the room constant, defined as

$$R = S\bar{\alpha}/(1-\bar{\alpha}), \tag{8}$$

where S is the total surface area of the room (in m^2) and $\bar{\alpha}$ is the average absorption coefficient, defined as

$$\bar{\alpha} = \frac{S_1\alpha_1 + S_2\alpha_2 + \cdots + S_n\alpha_n}{S_1 + S_2 + \cdots + S_n}. \tag{9}$$

At a certain distance from the source, the contributions to the sound pressure level from both the direct and reverberant fields will be equal. This distance is defined as the "room radius," r_r, and is equal to

$$r_r = (RQ_\theta/16\pi)^{1/2}. \tag{10}$$

Architecturally significant reduction, i.e., greater than 3 dB, of the sound level at a distance less than the room radius from an omnidirectional source cannot be achieved by adding sound absorbing material to the space since the direct sound predominates.

1.1.3 Decay of the Sound Field. When the sound source in an enclosure is stopped, the sound will persist for a time, as determined by the volume of the space and the amount of sound absorption present (see Sec. 1.4). The direct sound from the sound source will die away first, and the various components of the reverberant field will disappear one by one as the end of each field passes the observation point. The component waves will continue to impinge on the various boundary surfaces of the enclosure with the energy of the sound field being, for a continuum of modes, continuously and exponentially absorbed. During this decay, modes with the slowest decay rates dominate, and, for frequencies with widely spaced modes, this decay may result in very noticeable beats, "boominess," or "hollowness." The total energy in the enclosure at any instant equals the average energy density D_R times the room volume V. The number of times per second that a given sound wave will impinge upon a room boundary surface will depend on the average distance traveled between boundaries. This average distance is called the mean free path of the enclosure, which can be approximated by

$$4V/S, \tag{11}$$

where V is the room volume (in m^3) and S is the total surface area (in m^2). Thus the number of reflections occurring in 1 s will be equal to the speed of sound, c, divided by the mean free path.

Assuming a diffuse sound field in the enclosure, after a length of time the energy of the reverberant field will be reduced by numerous encounters with the boundary surfaces in a continuous process. Therefore, the reverberant sound field decays typically in small steps, beginning the instant that the last of the direct field is reflected from the boundary surfaces of the enclosure. The stepwise sound field decay is approximately exponential at a rate given by

$$1.086\,\frac{c}{V}\left(\sum_{i=1}^{n} S_i \ln(1-\alpha_i)\right)\quad \text{dB/s}, \tag{12}$$

where S_i is the surface area of each material in the enclosure (in m^2) and α_i is the absorption coefficient of each material.

A diffuse sound field is defined as one in which the sound pressure level is the same throughout, and it is equally probable that sound waves are traveling in every direction. This characteristic also ensures that the average energy density is the same at all points. In reality, a completely diffuse sound field cannot exist; however, the concept of a diffuse sound field is useful in spaces that are not highly absorptive, and in fact, near-diffuse sound fields can be achieved.

1.2 Sound Reflections

Sound reflections are the return of sound waves from the boundary surfaces of an enclosure and are essential for achieving good hearing conditions in a space, such as an auditorium. The signal strength, direction of arrival, time of arrival, and frequency content of the reflected sound are important factors, and in large spaces they provide the distinctive acoustical "signature," good or bad, associated with the space.

1.2.1 Geometrical Reflections. When a sound wave strikes a uniform boundary surface that is large compared with the wavelength of the sound, the reflection can be considered to be specular, and the ray can be traced similar to the reflection of light as used in geometrical optics. Analogous to light rays, sound rays are reflected from hard plane surfaces in accordance with the laws of reflection, i.e., the angle of the incident ray equals the angle of the reflected ray. Similarly, sound rays striking a curved surface will be focused or dispersed depending on whether the surface is concave or convex, as shown in Fig. 2. This approach is commonly used for investigating the effects of various shapes of proposed spaces on the distribution of sound within that space. For the most part, the use of concave surfaces as sound reflecting ele-

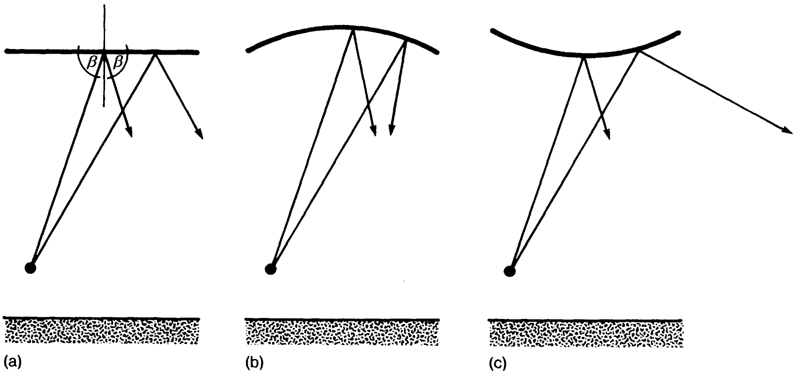

FIG. 2. Sound reflection characteristics from various types of boundary surfaces in an enclosure. The reflection patterns are similar to light reflectance in geometrical optics in that the angle of incidence of a sound wave is equal to the angle of reflection from a hard, plane surface (a), is focused from a concave surface (b), and is diffused from a convex surface (c).

ments should be avoided since this shape will cause focusing effects, time-smearing of transient sounds, discrete echoes, and nonuniform distribution of sound. It is often tempting for architects to use this shape for the rear wall of an auditorium since it follows the curve of the seating which is laid out in concentric arcs. Other examples of architectural spaces that typically exhibit focusing effects are those with barrel vaults or domed ceilings. Convex surfaces, on the other hand, disperse the reflected wave and are useful for achieving a diffuse sound field in the space. Caution must be used, however, in the application of the laws of reflection since this simplification of the behavior of sound is only valid for sound at the mid and high frequencies. In general, the minimum dimension of a surface should be on the order of at least two to four times the wavelength of the incident sound in order to achieve reflection. Surfaces smaller than this will not provide reflection, but rather diffraction of the incident sound wave.

1.2.2 Diffraction and Diffusion. When a sound wave strikes an object of finite size some of the energy is reflected, some is absorbed or transmitted, and, in general, some gets around the object, i.e., is diffracted. The extent of the diffraction and its scattering pattern depends on the relationship between the wavelength of the sound and the size of the object. In performance spaces the diffraction effects that accompany the reflection of sound are important and necessary. The architectural and decorative treatment of a room, such as columns, beams, balcony faces, statuary, niches, alcoves, etc., provides discontinuities in the boundaries of the space, and the pattern of the sound field distribution is altered by the scattering of the sound. This adds to the diffusion in the space, which helps to establish a more uniform distribution of the sound, ensures a perception of smooth growth and decay of the sound field, and improves the subjective degree of "liveness" of the room. Another means of increasing the diffusion of the sound in a space is through the use of randomly distributed areas of sound absorptive material on the boundary surfaces. Both diffusion and diffraction can be achieved by this means due to phase shifts in the diminished sound wave reflected from the absorbing material and from sound diffraction due to the edge effect (see Sec. 1.3.2).

An interesting development regarding diffraction and diffusion in broad directional patterns for use in performance spaces began in the 1970s with the use of number theory by Schroeder. Diffraction grating designs were developed for both two- and three-dimensional sound diffraction and diffusion, based on real or complex periodic number sequences fashioned from quadratic residues, primitive roots, primitive polynomials in finite number fields, the Legendre symbol, the index function, the Zech logarithm, etc. These diffusing elements consist of a periodic array of wells of equal width, but different depths, separated by thin dividers. The well-depth sequences provided by the diffraction grating achieve both a time distribution of the backscattered sound as well as a uniform angular dispersion over a broad, designable frequency range.

1.3 Sound Absorption

Sound absorption is the primary factor in reducing noise and controlling reverberation in a space. Sound absorption can occur both at the boundary surfaces of a space and by the air contained within the enclosure. At the boundary surfaces absorption is achieved by converting the acoustic energy of a sound wave into heat through thermal lag and viscous drag. When a sound wave encounters a boundary surface a portion of the energy is absorbed, the amount depending on the surface material, and the remaining energy is reflected back to the space. The amount of sound absorption of a material is defined in terms of its absorption coefficient α, which is a ratio of the sound energy absorbed to the sound energy incident upon that surface. A perfect absorber would have an absorption coefficient of 1, while a perfect reflector would have an absorption coefficient of 0. The unit of sound absorption, $S\alpha$, is called a sabin, which is dimensionless, and is equal to the area of a sound absorbing material times its sound absorption coefficient at each frequency of concern. The total sound absorption in a space, therefore, is equal to the sum of the products of the areas of all of the surfaces and materials, and their absorption coefficients, i.e.,

$$\sum S\alpha = S_1\alpha_1 + S_2\alpha_2 + \cdots + S_n\alpha_n, \qquad (13)$$

where $\Sigma S\alpha$ is the total sound absorption (in sabins); and $S_1\alpha_1, S_2\alpha_2, \ldots, S_n\alpha_n$ are the areas of each material in the space times its absorption coefficient.

The absorption coefficient of a material will vary with the angle of incidence of the sound wave. However, most published absorption coefficients for building materials are obtained from measurements in reverberation chambers which are designed to achieve a near-diffuse reverberant sound field, and therefore the reported absorption coefficients are for near-random incidence of sound. From these measurements, the absorption coefficients for materials are obtained for six standard octave band-center frequencies; 125, 250, 500, 1000, 2000, and 4000 Hz. In current practice, some manufacturers of sound absorbing materials do not report the absorption coefficients for their products, but rather publish a single-number rating called the noise reduction coefficient (NRC), which is the arithmetic average (to the nearest increment of 0.05) of the absorption coefficients at 250, 500, 1000, and 2000 Hz. The NRC can be used to select sound absorbing materials for noncritical spaces or for general noise control applications, but should not be used in the design of critical or acoustically sensitive spaces.

A common misconception continues to exist throughout the building industry regarding which materials to use for sound absorption and which for sound isolation. Sound absorption and sound isolation are entirely different functions and require materials with entirely different physical characteristics. Sound absorption typically requires a lightweight, porous material, which allows the air molecules to move in and out freely, but retarded mostly by viscous drag. Sound isolation, on the other hand, requires a heavy, impervious barrier to block air movement. Applying a sound absorbing material to the surface of a wall will reduce the reverberation in the space, but will not provide any meaningful increase in the amount of sound isolation provided by the wall itself.

1.3.1 Sound Absorbing Properties of Materials. All materials absorb sound to some degree. The three basic mechanisms through which sound absorption can be achieved, using standard building materials, are porous materials, resonant panels, and resonant air cavities. Each of these types provides absorption in a different manner and as a result has distinctive absorption characteristics. The generalized sound absorption characteristics of the three types of absorbing mechanisms are shown in Fig. 3.

1.3.1.1 Porous Absorber. Porous, fibrous materials are by far the most commonly used type of sound absorption in architectural acoustics design and include such standard finish materials as carpet, drapes, upholstered furniture, acoustical tile, etc. This type of material is characterized by its pores, which are exposed to the air of a room. Sound absorption is achieved when the air

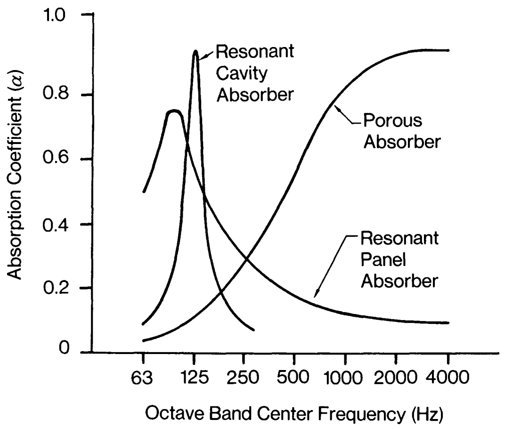

FIG. 3. Typical sound absorption characteristics of standard building materials. Porous materials absorb acoustic energy over a broad frequency range. Resonant absorbers, such as panel absorbers and cavity absorbers, absorb acoustic energy over narrow frequency ranges near their resonant frequencies.

pressure at the material increases and decreases due to successive sound waves which force the air molecules in the connected pores both to vibrate and also to undergo heating and cooling through compression. This rapid motion of the air molecules results in the loss of acoustic energy due to conversion to heat through viscous, linear friction. Another damping mechanism is irreversible heat conduction from the air into the material and back. The latter is significant at low frequencies in closed cavities, such as within walls. The amount of sound absorption achieved in each type of porous material is determined by such physical properties as thickness, density, porosity, fiber or pore orientation, and flow resistance. Of these properties, flow resistance and thickness are usually the most important. As a sound wave encounters a porous material the amplitude of the air particle vibration is progressively damped by friction against the sides of the pores, thus acting as an acoustic resistance. The specific flow resistance of a layer of porous material is defined as

$$R_1 = \Delta p / Tu, \tag{14}$$

where R_1 is the specific flow resistance (in rayls m^{-1}) (the unit "rayl" is equal to $\mathrm{N\,s\,m^{-3}}$), Δp is the air-pressure differential between the two sides of the material (in $\mathrm{N\,m^{-2}}$), T is the thickness of the material (in m), and u is the apparent particle velocity through the sample (in $\mathrm{m\,s^{-1}}$), calculated using the face area of the material and not the actual flow area. If R_1 is too small, insufficient friction will result and little absorption will occur. If the value of R_1 is too great, the particle velocity is reduced and too much reflection will occur at the face of the material due to too large an impedance mismatch. This is the reason why painting a porous material can significantly decrease its sound absorbing performance. It is interesting to note that in architectural acoustics applications, typical sound levels produce airflows small enough that essentially all sound absorption is through viscous and thermal losses, and is well described by linear equations. Turbulent dissipation, on the other hand, involves drag which is proportional to the square of the particle velocity and, if significant, can result in audible frequency distortion of the sound field. Experimentally, this distortion becomes audible with some materials at particle velocities associated with sound levels above approximately 120 dB, which are not typically encountered in architectural acoustics situations.

The sound absorbing performance of a porous material can be increased at the mid and lower frequencies by increasing the thickness of the material, as indicated in Fig. 4. Maximum sound absorption will occur in a porous absorber when the airflow, or particle velocity, is at a standing-wave maximum or antinode. This occurs when the distance from the wall is equal to one-quarter wavelength, $\lambda/4$. Thus, in order to absorb low-frequency energy the absorber must either be very thick or else be mounted $\lambda/4$ in front of a rigid wall.

1.3.1.2 Resonant Panel Absorber. Absorption can also be achieved by constructing a resonant panel system which consists of a thin, impervious, flexible material mounted at some distance out from a rigid wall. A sound wave incident on the flexible surface will cause it to resonate, and thereby absorb sound, at a characteristic frequency. At frequencies other than near the resonant frequency the absorption is poor. Sound absorption with this type of system is achieved by converting the sound energy incident on the panel into mechanical energy and heat. The resonant frequency of this system is determined by the panel stiffness, the stiffness of its supports, panel thickness, mass, and depth of air space. If the panel material is thin and

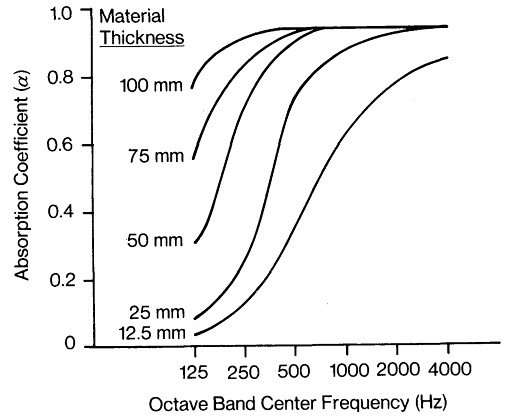

FIG. 4. Effect of thickness on the sound absorbing characteristics of porous materials. Absorption is most effective when the thickness of the porous material is equal to at least one-quarter wavelength of the sound. Thus the lowest frequency at which a material can effectively absorb acoustic energy decreases as the thickness of the material increases, or as the material is moved away from the wall.

flexible the resonant frequency can be approximated by

$$f_{\mathrm{res}}=60/\sqrt{Md}, \tag{15}$$

where M is the surface mass of the panel (in kg m^{-2}) and d is the depth of air space (in m). Increasing either the mass of the panel or the depth of the air space will lower the resonant frequency of the system. Resonant panel absorbers are most practical for absorption in the lower frequencies, i.e., 200 Hz and below, and generally will be thinner than a porous absorber of similar performance in the same frequency band. The frequency bandwidth over which the resonant panel absorber will provide effective absorption depends on the amount of damping within the system. A heavily damped system will act over a wider frequency range, but the amount of absorption provided at the resonant frequency will be less. Few resonant panel absorbers are commercially available, and typically must be designed and constructed as required for a specific installation. It can be noticed that spaces built with large areas of thin wood paneling on the walls are often unintentionally weak in low-frequency "bass" energy, resulting in a sound quality which is acoustically "bright" and lacking in "warmth." This is due to the excess absorption of the wood paneling acting as efficient resonant absorbers at the lower frequencies.

1.3.1.3 Resonant Air Cavity Absorber.

The third type of absorber is a resonant air cavity, volume resonator, or Helmholtz resonator, which consists of an enclosed volume of air connected to the room by a small opening or neck. Sound energy in the room near the resonant frequency will cause the air in the restricted neck to vibrate. Energy is absorbed primarily by friction of the air particles that pump back and forth along the neck. Architecturally, the neck of this type of absorber is usually either a round hole or a slot.

With this type of absorber, increasing the resistance to air movement by adding a porous material, either to the neck or to the air cavity, is a standard practice. The primary reason for this is that some sound energy is transferred into the air cavity which stores it instead of dissipating it. When the sound in the room is stopped the stored resonant energy is reradiated back into the room, thus prolonging the room reverberation. Adding a porous material to this type of resonator also has the effect of increasing the frequency range over which it will provide effective absorption, but as with the resonant panel absorber, this reduces the peak absorption at the damped resonant frequency.

The resonant frequency of an undamped Helmholtz resonator with a circular neck can be approximated by

$$f_{\mathrm{res}}=(c/2\pi)\sqrt{S/Vl}, \tag{16}$$

where c is the speed of sound (in m s^{-1}), S is the cross-sectional area of the neck (in m^2), l is the length of the neck (in m), and V is the volume of the enclosed air volume (in m^3).

As with the resonant panels, the architectural resonant air cavity provides useful absorption primarily at the lower frequencies. This type of absorber also has a somewhat limited application, but several products are commercially available in the form of structural masonry units that can be used for wall constructions. The two major types of masonry units currently available are ceramic blocks with holes leading into the enclosed air chamber or cell, and concrete blocks with slots leading into the cell. Both of these products can be provided with a porous absorbing material in the cells. These products can withstand moisture and other adverse weather conditions, and are therefore useful for some industrial, exterior locations such as enclosures for transformers and mechanical equipment.

1.3.2 Measurement Procedures. The absorption coefficients of acoustical materials are typically obtained by measuring the time rate of decay of the sound energy density in a certified reverberation room with and without a test specimen present. The procedures required for measuring the absorption coefficients are fairly strict and must follow those established by the American Society for Testing and Materials (ASTM), Designation C423, "Standard Test Method for Sound Absorption and Sound Absorption Coefficients by the Reverberation Room Method," or the International Organization for Standardization (ISO), Recommendation R354, "Measurements of Absorption Coefficients in a Reverberation Room." The ASTM standard requires that the minimum volume of the room shall not be less than $V=4\lambda^3$ based on the center frequency of the lowest one-third octave frequency band at which measurements are to be made. For the one-third octave band centered on 125 Hz, for which the wavelength λ is 2.75 m, this requires a minimum room volume of 82 m^3. The ISO recommendation is for a room volume larger than 180 m^3, preferably as close to 200 m^3 as possible. In order to achieve a near-diffuse reverberant sound field in the room, a number of angled sound reflective panels are required. These can be both fixed, which are hung and distributed at random angles, and moving, on a rotating shaft or other mechanism.

The size of the sample to be tested is standardized, as are the proportions or aspect ratio, so that test specimens can be compared on an equal basis. For the ASTM standard the size of the test specimen is 6.7 m^2, in a shape of 2.44 m $\times$2.74 m. For the ISO procedure the recommended specimen size is between 10 and 12 m^2, in a rectangular shape with the ratio of the sides between 0.7 and 1. The size and aspect ratio are specified because a test specimen with a narrow rectangular shape will generate excessive diffraction effects at the perimeter, or "edge effects," resulting in an unrealistically high absorption coefficient.

The reverberation time is measured in the test chamber with and without the presence of the test specimen. The absorption coefficient of the test specimen is then determined by dividing the excess absorption by the area of the test specimen. Note that since the sound absorption of a material depends not only on the physical properties of the material itself, but also on the manner in which the material is mounted or installed, the mounting used for the test specimen must also be known. According to ASTM Designation E795, "Standard Practice for Mounting Test Specimens During Sound Absorption Tests," there are eight standard mounting configurations that can be used for determining the absorption coefficients of a material. For example, most manufacturers of acoustical ceiling board products test their materials in a configuration approximating a typical suspended ceiling with an air space of 400 mm between the face of the material and backing surface (type E-400 mounting). This same ceiling board product will provide significantly different absorption coefficients, particularly much less at the lower frequencies, when tested in a mounting configuration where the material is adhered directly to a surface (type B mounting). It is important to use the absorption coefficients determined from the mounting configuration closest to that condition in which the material will actually be used.

Even with the strict requirements for the design of a reverberation room and standardized measurement procedures, the measured absorption coefficients will vary somewhat between testing laboratories due to such factors as microphone positions, loudspeaker positions, test specimen position, room geometry, and diffuser geometry. Another peculiarity that occurs in the measurement of absorption coefficients is that the value can exceed unity for a highly absorptive material. This is caused by diffraction effects at the edges of the material, i.e., the previously noted "edge effect," which increases the apparent absorption coefficient. The ASTM standard recommends that no adjustments be made to absorption coefficients greater than 1, but requires that the data be reported as measured.

Table 1. Sound absorption coefficients for common building materials and furnishings.

Material	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz	NRC
Walls							
Brick, unglazed	0.02	0.02	0.03	0.04	0.05	0.07	0.05
Concrete, rough	0.01	0.02	0.04	0.06	0.08	0.10	0.05
Concrete block	0.36	0.44	0.31	0.29	0.39	0.25	0.35
Concrete block, painted	0.10	0.05	0.06	0.07	0.09	0.08	0.05
Glass, heavy (large panes)	0.18	0.06	0.04	0.03	0.02	0.02	0.05
Glass, ordinary window	0.35	0.25	0.18	0.12	0.07	0.04	0.15
Gypsum board (25 mm)	0.29	0.10	0.05	0.04	0.07	0.09	0.05
Marble	0.01	0.01	0.01	0.01	0.02	0.02	0.00
Plaster, on brick	0.01	0.02	0.02	0.03	0.04	0.05	0.05
Plaster, on lath	0.14	0.10	0.06	0.05	0.04	0.03	0.05
Wood paneling (6.5 mm), with air space behind	0.42	0.21	0.10	0.08	0.06	0.06	0.10
Wood paneling (25 mm), with air space behind	0.19	0.14	0.09	0.06	0.06	0.05	0.10
Drapes							
Lightweight, 340 g m^{-2}, flat on wall	0.03	0.04	0.11	0.17	0.24	0.35	0.15
Mediumweight, 475 g m^{-2}, draped to half area	0.07	0.31	0.49	0.75	0.70	0.60	0.55
Heavyweight, 610 g m^{-2}, draped to half area	0.14	0.35	0.55	0.72	0.70	0.65	0.60
Floors							
Concrete or terrazzo	0.01	0.01	0.02	0.02	0.02	0.02	0.00
Linoleum, rubber, or asphalt tile on concrete	0.02	0.03	0.03	0.03	0.03	0.02	0.05
Wood parquet on concrete	0.04	0.04	0.07	0.06	0.06	0.07	0.05
Carpet							
Heavy, on concrete	0.02	0.06	0.14	0.37	0.60	0.65	0.30
Heavy, on foam rubber	0.08	0.24	0.57	0.69	0.71	0.73	0.55
Indoor-outdoor	0.01	0.05	0.10	0.20	0.45	0.65	0.20
Ceilings							
Acoustical ceiling board, 20 mm thick, in suspension system (Mtg. E-400)	0.76	0.93	0.83	0.99	0.99	0.94	0.95
Acoustical ceiling board, 20 mm thick, glued to solid surface (Mtg. B)	0.10	0.60	0.80	0.82	0.78	0.60	0.75
Sprayed cellulose fibers, 25 mm thick, on concrete (Mtg. A)	0.08	0.29	0.75	0.98	0.93	0.76	0.75

(continued)

Table 1. Sound absorption coefficients for common building materials and furnishings *(continued).*

Material	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz	NRC
Polyurethane foam, 25 mm thick, open cell, reticulated	0.07	0.11	0.20	0.32	0.60	0.85	0.30
Seats and audience[a]							
Fabric well-upholstered seats, with perforated seat pans, unoccupied	0.19	0.37	0.56	0.67	0.61	0.59	···
Audience, seated in upholstered seats	0.39	0.57	0.80	0.94	0.92	0.87	···
Chair, metal or wood seat, unoccupied	0.15	0.19	0.22	0.39	0.38	0.30	···

[a] These values are for the floor area occupied by the seating or audience (in m^2), and include an "edge effect" equal in area to a strip 0.5 m wide for each aisle bordering the seating area.

1.3.3 Absorption Coefficients. Table 1 lists typical absorption coefficients of common building materials and furnishings.

1.3.4 Noise Reduction by Sound Absorbing Materials. The installation of sound absorptive materials at the boundary surfaces of an enclosure can shorten the reverberation time, lower the level of the reverberant sound field, and help to localize the noise to the region of its source, i.e., its direct sound field. As indicated in Eq. (7), increasing the amount of absorption in a space decreases the energy density of the steady-state reverberant field. However, the level is never significantly reduced in the direct field closer to the noise source than the room radius [see Eq. (10)]. Noise reduction provided by the addition of sound absorbing material to a space can be determined by

$$D = 10 \log_{10}(A_2/A_1), \qquad (17)$$

where D is the amount of noise reduction (in dB), A_1 is the total amount of sound absorption before treatment (in sabins), and A_2 is the total amount of sound absorption after treatment (in sabins).

1.3.5 Air Absorption. In large enclosures sound energy can be absorbed by the air itself, particularly at the higher sound frequencies. For reverberation time calculations (see Sec. 1.4) the amount of sound absorption provided by the air must be included in the total absorption for the enclosure, the total absorption being the sum of all absorbing features, including those due to the losses at the boundary surfaces and the furnishings in the space. If an enclosure is small the number of reflections from the boundary surfaces is large, while the distance traveled by the sound waves between the boundaries is small. In this situation air absorption is typically not as important as the boundary absorption and can usually be ignored. However, if the enclosure is very large, for the same number of reflections from the boundary surfaces, the distance traveled by the sound waves through the air is much greater and the energy loss, or absorption, can be significant.

Air can absorb acoustic energy by two methods: classical absorption, which is caused by heat conduction and air viscosity; and molecular absorption, which is due to rotational and vibration relaxation of the air molecules. In architectural acoustics situations, the amount due to classical absorption is insignificant. The amount of absorption due to molecular absorption, however, can range from negligible amounts below 1000 Hz to large amounts in large spaces and at high frequencies. In practice, air absorption is typically considered for the octave band-center frequencies at 2000 Hz and above.

There are several methods used for predicting the amount of air absorption, which include such variables as room volume, air temperature, humidity, and frequency. Not all of these prediction methods are in agreement. The classic study, which empirically

Table 2. Molecular absorption of sound in air per cubic meter of room volume (at 20 °C).

Relative humidity	Octave band-center frequency 2 kHz	4 kHz	8 kHz
30%	0.012	0.038	0.136
50%	0.010	0.024	0.086
70%	0.009	0.021	0.060

determined the amount of absorption provided by air, was completed by Harris in 1966. The results of this study are summarized in Table 2 and show the amount of sound absorption provided by the air based on the volume of the enclosure and relative humidity.

1.4 Reverberation Time

Reverberation time is perhaps the most widely recognized term from the field of architectural acoustics. Sound which is generated in an enclosed space is repeatedly reflected by its boundary surfaces where it is absorbed, transmitted, or reflected. Once the sound source has ceased, a fraction of the acoustic energy is lost at each reflection. The time that it takes the sound energy to decrease by a factor of one million, or the sound pressure by a factor of 1000, i.e., 60 dB, is defined as the reverberation time, T_{60}. The reverberation time is therefore controlled by the volume of the space, i.e., the larger the enclosure, the farther the sound must travel between reflections, and the type and amount of sound absorption in the space, i.e., the higher the absorption coefficient, the greater the decibel loss per reflection.

1.4.1 Reverberation Equations. The three most commonly used reverberation equations, Sabine, Eyring, and Fitzroy, are discussed below.

1.4.1.1 Sabine Equation. Reverberation time was first defined by Sabine from his empirical work, beginning in 1895, on the acoustics of the then new lecture hall in the Fogg Art Museum at Harvard University. His work established the quantitative relationship between the room volume and the sound absorption provided by the room's finish materials, furnishings, and occupants. The Sabine equation for reverberation time is

$$T_{60}=0.161\,V/\Sigma S\alpha, \qquad (18)$$

where T_{60} is the reverberation time (in s), V is the room volume (in m^3), and $\Sigma S\alpha$ is the total sound absorption in the room (in sabins).

Note that the unit of sound absorption, $S\alpha$, is called the sabin in honor of Sabine's pioneering work.

1.4.1.2 Eyring Equation. For noncritical design situations, the Sabine equation predicts reverberation times with adequate accuracy. However, the equation is not completely accurate, particularly in spaces where the total absorption is very large, such as in broadcast studios or other acoustically "dead" spaces. In 1930, Eyring developed an equation that is a more accurate expression than the Sabine equation, based on multiple sound reflections at the enclosure's surfaces and uniform distribution of the sound absorption:

$$T_{60}=0.07V/[-S\log_{10}(1-\bar{\alpha})], \qquad (19)$$

where S is the total surface area of the room (in m^2) and $\bar{\alpha}$ is the average absorption coefficient.

1.4.1.3 Fitzroy Equation. Both the Sabine and Eyring equations assume uniform absorption at all surfaces as well as a diffuse sound field in the space. These conditions are rarely achieved in practice and the equations are not accurate in such nondiffuse spaces where pairs of surfaces are sound reflective and others are absorptive, as in a typically furnished space with carpeted floor, acoustical tile ceiling, and hard walls. In 1959, Fitzroy expanded the Eyring equation to take such uneven distributions of sound absorbing surfaces into account. This equation apportions the reverberation time between the three major axes of a rectangular space which exhibit decay rates dependent on the average absorption of the pair of surfaces in each axis:

$$\begin{aligned}T_{60}=&\frac{x}{S}\left(\frac{0.07V}{-S\log_{10}(1-\bar{\alpha}_x)}\right)\\&+\frac{y}{S}\left(\frac{0.07V}{-S\log_{10}(1-\bar{\alpha}_y)}\right)\\&+\frac{z}{S}\left(\frac{0.07V}{-S\log_{10}(1-\bar{\alpha}_z)}\right),\end{aligned} \qquad (20)$$

where x is the total side wall area (in m^2); y is the total ceiling and floor area (in m^2); z is the total front and rear wall area (in m^2); S is the total surface area of the room (i.e., $x+y+z$)

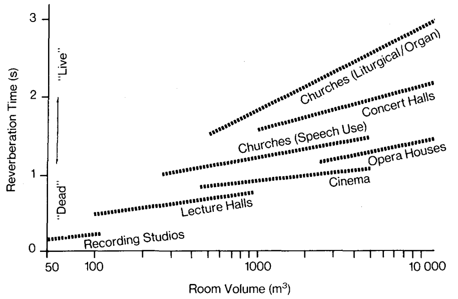

FIG. 5. Approximate range of optimum reverberation times at 500 Hz for fully occupied spaces as a function of the volume and use of the space. The shorter reverberation times in dead spaces provide a high degree of speech intelligibility. Longer reverberation times in live spaces provide "warmth" for music performance.

(in m^2); and $\overline{\alpha}_x$, $\overline{\alpha}_y$, and $\overline{\alpha}_z$ are the average absorption coefficients for the areas x, y, and z, respectively.

1.4.2 Design Criteria for Enclosed Spaces. There are no uniformly accepted values for the "ideal" reverberation time in any type of space. In general, however, a space that is to be used for speech should have a fairly short reverberation time to ensure good articulation and intelligibility. A space to be used for music performance should have a longer reverberation time to achieve adequate "blend" and "richness" of sound. The ranges of reverberation times generally considered acceptable for various types of spaces are shown in Fig. 5. However, since there are no absolute or "perfect" reverberation times for any space, these values should be considered as general guidelines. The values shown are for the reverberation times at mid-frequency, i.e., 500 Hz. For most performance spaces, they should be longer at the lower frequencies and somewhat shorter at the higher frequencies. For spaces which require a high degree of articulation, such as recording studios, music practice rooms, etc., the reverberation times should be fairly constant over the frequency range from 63 to 4000 Hz.

Although reverberation time is an important acoustical characteristic of enclosed spaces, it can be a notoriously unreliable descriptor of overall subjective quality. Comparably sized spaces can have identical reverberation times but, due to numerous other acoustical characteristics, one space can be well regarded while the other considered unsatisfactory. This is because the reverberation time essentially only describes the decay constant of the reverberant energy in the enclosure. Other factors such as reflection characteristics, ratio of early to reverberant energy, sound level, and directional distribution patterns are also important for achieving overall good acoustical characteristics of a space (see Sec. 1.5).

1.4.3 Assisted Resonance Systems. Assisted resonance is a term used to describe various methods of creating an artificial reverberation time in a space by electroacoustic means. This type of system has been used, with varying degrees of success, in performance spaces which require longer reverberation times to compensate for insufficient volume or shape dictated by nonacoustical criteria of the space. This approach has also been used as a specific design element for multiple-use spaces which must accommodate various functions requiring differing reverberation times.

One of the first large performance spaces to use assisted resonance, in this case as a remedial measure, was the Royal Festival Hall in London. Originally opened in 1951, an assisted resonance system was installed during renovation work in 1965 to compensate for the "dryness" associated with the acoustics of the hall. This system is based on regenerated reverberation in which multiple channels, each consisting of a microphone, amplifier, and loudspeaker, are located at pressure maxima for the particular modes whose reverberation needs to be "assisted." The signal for each channel is adjusted to be in phase around the loop, and the gain is adjusted to be approximately 6 dB below the feedback level. The microphones are located towards the rear of the hall, i.e., in the reverberant field, and the loudspeakers near the front. This provides the required time delay so that the reinforced sound occurs later than the direct sound. To achieve appropriate frequency filtering each microphone is installed in a tunable Helmholtz resonator located above the ceiling, with the neck extending into the hall and the bottom of the neck flush with the underside of the ceiling.

One advantage to this type of system is that, since each loudspeaker is operating at only one frequency, the frequency response uniformity of the loudspeaker is not critical.

The disadvantage is the large number of channels required. In Festival Hall there are 172 channels, covering the range of frequencies from approximately 60 to 700 Hz. As a general guideline for this type of system, Parkin, the designer of the system at the Royal Festival Hall, indicates that a system should consist of at least 100 channels. Using this number, a 100% increase in the reverberation time can be achieved at the lower frequency range and approximately a 60% increase at the mid to high frequencies, i.e., up to approximately 2000 Hz.

A second method of electroacoustic enhancement of reverberation is called ambiophony. This system uses microphones to pick up sound close to the source; the signal is mixed and passed through a time-delay element, such as a tape loop, reverberation plate, reverberation chamber, or digital delay device, and then transmitted back to the hall by numerous loudspeakers. The time delays and amplification of the signal from the various loudspeakers must be adjusted and evaluated subjectively until the desired increase in the reverberation time and "natural" sound quality are achieved. The major disadvantage to this type of system is the requirement for locating microphones close to the source, i.e., a very large number of microphones is required for an orchestral performance in a concert hall. This requirement has both practical and aesthetic drawbacks which are generally not acceptable to either typical performing groups or audiences.

A more recent development is the electronic reflected energy system (ERES), which takes a signal from a single microphone located in the sound field near the stage area. The signal is processed through amplifiers, equalizers, and delay systems, and then distributed to loudspeakers located at various positions in the auditorium. The signal is modified and the loudspeakers located so as to simulate sound reflections comparable to those of natural reflections in a performance space, including early reflections, lateral reflections, and reverberant energy. The processed sound provided by this system has a frequency range from approximately 100 to 6000 Hz. The advantage of this type of system is that the acoustical characteristics of a multipurpose space can be quickly modified to accommodate the requirements of various conflicting performance types.

1.5 Auditorium and Enclosure Design

The traditional acoustical characteristics that must be considered in the design of almost any type of indoor or outdoor performance space can be summarized by the following basic requirements:

1. quiet background, which ensures that the wanted signal is not masked by intruding noise from mechanical equipment systems, exterior sources, etc.;
2. adequate loudness and distribution, which ensures that an appropriate sound level of the source signal is achieved at the listening areas, typically through proper shaping of the enclosure surfaces providing reflected energy of suitable strength, direction, frequency content, and the appropriate balance of early to reverberant energy; and
3. reverberation time, to achieve appropriate decay characteristics.

1.5.1 Sound Reflection Design. There are many different approaches and design philosophies to auditorium design. Since the late 1950s, significant research has been completed both in the design of actual spaces and in simulated spaces. However, there is little consensus among designers and researchers on the relative importance of the various factors in auditorium design. Some of the major aspects regarding reflected sound characteristics are summarized below.

1.5.1.1 Early versus Reverberant Sound. The major area of discussion that continues regarding these two critical portions of the sound field is the appropriate balance between the amount and character of the early-arriving sound to the amount of reverberant sound energy. In general, the early sound, i.e., arriving at the listener within 50–80 ms after the direct sound, provides clarity and articulation. Too much early energy produces a harsh, "hi-fi" quality to the sound. The reverberant sound in a space provides blend and richness to the sound quality. Excessive reverberant energy, however, produces a muddy or indistinct sound quality. Establishment of an optimum pattern of early sound reflections should include such parameters as frequency content, energy content relative to the reverberant sound field, time and direction of arrival, etc. To achieve proper early sound reflections, appropriately designed and oriented reflecting surfaces near the performers are required. Note that if two sounds arrive within 10 ms of each other they cannot be distinguished as separate sounds by a normal listener. Further, if a reflected sound arrives within 40 ms of the direct sound, it is also not perceived separately, but rather produces a change of quality and increase in the perceived loudness of the original sound. This is called the Haas precedence effect.

1.5.1.2 Lateral Reflections. The other major factor that has been extensively discussed and researched for the past decade is the proper value of lateral sound reflections in performance spaces, including energy content, frequency content, time of arrival, and angle of arrival. It has been established that the subjective quality of auditory "spaciousness" is a direct function of lateral reflections. This subjective quality consists of such perceptual attributes as spatial impression, spatial responsiveness, apparent source width, envelopment, etc. For the formation of auditory spaciousness binaural dissimilarity of the sound at the listener's position is required, including both interaural arrival time and level differences. Subjectively, interaural coherence is highly positively correlated with the width of the hall, the side walls being the primary source of lateral reflections. This is also one of the major reasons why the early, "classical" European concert hall is well regarded as having good acoustical characteristics. This space, typically classified as a "shoebox" shape, consists of a rectangular space with a high ceiling, which provides adequate reverberant energy, and a narrow width, which provides early, strong lateral reflections from the close side walls.

1.5.2 Subjective Preference Scaling. As a means for establishing the relationship between the objective physical properties of the sound field and the subjective properties as heard by the listener, various subjective preference and evaluation scales have been developed. The main purpose of this type of study is to establish the desired acoustical characteristics of a performance space during the design phase of a project rather than having to correct any deficiencies after construction. This work was initially begun by Beranek in the early 1960s and included such subjective preference parameters as clarity, warmth, intimacy, loudness, balance, etc. More recently, extensive work has been carried out by Barron (1971), Ando (1985), and others to study the relationship of such subjective and objective parameters as preferred listening level, delay time of early sound reflections, reverberation time, and the magnitude of interaural cross correlation. These studies typically involve head-related stereophony, in which either sound fields are recorded in actual performance spaces, or special recordings of musical works are recorded in anechoic chambers. The microphones are located in the ears of a dummy human head to achieve highly accurate reproduction of the sound field, and the recordings therefore include such effects as diffraction of the signal caused by the head and outer ear configuration, time and level differences of the signal at the left and right ear, etc. These recorded data can be reproduced, simulated, or modified using digital computers to achieve direct comparisons of sound fields of various spaces, with subjective measures using paired comparison tests to establish multivariate scaling preferences.

1.5.3 Model Testing and Computer Simulation. Another design tool for studying the acoustical characteristics of a performance space is through the use of scale model testing and computer simulation. The very early acoustical model testing consisted of two-dimensional models such as ripple tanks, in which the paths of sound waves are represented by water wavelets, or optical models, in which the boundary surfaces are constructed of an optically reflecting material for light reflection to simulate sound rays. The use of scale models for actual acoustical measurements was first developed in the 1930s; however, current techniques can provide highly accurate objective evaluation of the acoustical characteristics of spaces. Typically, the preliminary work on a performance space model is completed in a model built at 1:50 scale. Since frequency scales in a linear manner, all elements of the acoustical model can be similarly scaled, including dimensions, wavelength, frequency, absorption coefficients, and time. This scaling factor requires careful selection of the materials used for the finishes of the space in order to achieve acceptably accurate sound absorption coefficients. The most difficult factor to achieve is the appropriate degree of air absorption. To compensate for the excess absorption of air in the scale models some researchers use models filled with nitrogen as the medium, others dry the air to approximately 2% relative humidity to minimize absorption.

After the initial design studies are completed, another model at 1:10 scale is commonly used to refine the objective measurements. The sound source used for the studies is typically an impulse source such as a spark discharge. The signal is measured using $\frac{1}{8}$-in.-diam or $\frac{1}{4}$-in.-diam condenser microphones, often located in a scaled, miniature dummy head (see Sec. 1.5.2), and the signal is processed using analog-to-digital converters and digital signal processing. Use of this type of system can provide accurate information on such acoustical design parameters as the ratio of early to late energy, lateral energy, echograms, reverberation time, loudness levels, etc. The use of scale models, combined with the signal processing capabilities of digital computers, is considered by many researchers to be far more accurate than simulation techniques that use a digital computer alone to calculate the sound field of a performance space. This is primarily due to the lack of proper algorithms for simulating scattering on diffusing surfaces.

2. SOUND ISOLATION

In architectural acoustics, intruding noise sources from both the exterior and interior of a building must be considered. Control of the transmission of intruding noise is of critical importance in the design of spaces, and the amount of noise reduction or transmission loss that must be achieved depends on such factors as the level of the intruding noise; transmission paths, i.e., airborne or structure-borne; frequency content of intruding noise; and the design criteria or level of ambient noise allowable in the space of concern. For noise-reduction design purposes, the type of intruding noise at a space must first be categorized by its primary transmission path, i.e., through the air or through the building structure. The basic sound isolation problem is to reduce the sound pressure level of the intruding noise by sound isolation techniques appropriate to the type of intruding signal.

2.1 Airborne-Sound Isolation

A basic example of airborne-sound isolation can be shown by two spaces located side by side with a wall separating them. In the source room, a noise source produces a reverberant sound field which, when it encounters the boundary surfaces, will be either absorbed, reflected, or transmitted. The amount of sound energy not absorbed or reflected back to the source room will propagate through the wall, the amount depending on the acoustic impedance of the wall compared to that of the air. At the boundary of the adjacent space, the relative impedances of the wall and the air will again determine the percentage of energy transmitted into the receiving room.

2.1.1 Transmission Coefficient and Transmission Loss. Airborne-sound isolation is essentially the reduction of energy transmitted from one space to another. One way to express this is by the flow of sound energy, or transmission coefficient τ, which is defined as the ratio of transmitted acoustic energy to the incident acoustic energy. For a given angle of an incident plane wave in an infinitely large panel, the transmission coefficient is given by

$$\begin{aligned}\tau(\Theta)&=I_t/I_i\\&=\left\{\left[1+\eta\left(\frac{\omega\rho_s}{2\rho c}\cos\Theta\right)\left(\frac{\omega^2B}{c^4\rho_s}\sin^4\Theta\right)\right]^2\right.\\&\quad+\left[\left(\frac{\omega\rho_s}{2\rho c}\cos\Theta\right)\right.\\&\quad\left.\left.\times\left(1-\frac{\omega^2B}{c^4\rho_s}\sin^4\Theta\right)\right]^2\right\}^{-1},\end{aligned}\tag{21}$$

where $\tau(\Theta)$ is the transmission coefficient, I_t is the sound intensity of the transmitted energy (in W m^{-2}), I_i is the sound intensity of the incident energy (in W m^{-2}), η is the composite plate loss factor (dimensionless), ω is angular frequency (in rad s^{-1}), $\rho_s=\rho_m t$ is the plate surface density (in kg m^{-2}), ρ_m is the density of the plate material (in kg m^{-3}), t is the thickness of the plate (in m), ρ is the density of the air (in kg m^{-3}), c is the speed of sound in air (in m s^{-1}), Θ is the incident angle (in degrees), and B is the plate bending stiffness per unit width (in N m).

A forcing sound field consisting of an incident plane wave is not typically encountered in practice, since the sound field in a space is more commonly a diffuse field. In this situation the total transmitted intensity is

$$I_t = \int_\Omega \tau(\Theta)\ I_i \cos\Theta\ d\Omega. \qquad (22)$$

The integration is over a hemisphere of solid angle Ω, where

$$d\Omega = \cos\Theta\ d\Theta\ d\Phi. \qquad (23)$$

Since the sound intensity of the incident energy is the same for all plane waves in a diffuse field, and the transmission coefficient is independent of the polar angle Φ, an average transmission coefficient is defined by

$$\overline{\tau} = \frac{\int_0^{\Theta_{\mathrm{lim}}} \tau(\Theta)\ \cos\Theta\ \sin\Theta\ d\Theta}{\int_0^{\Theta_{\mathrm{lim}}} \cos\Theta\ \sin\Theta\ d\Theta}, \qquad (24)$$

where Θ_{lim} is the limiting angle of incidence of the sound field which, for random incidence, is taken as $\pi/2$ or 90°.

To rate the sound isolating properties of a structure, the difference between the incident intensity level and the transmitted intensity level is used. This difference is called the transmission loss, TL, or sound reduction index, R (in dB), and is defined as

$$R = 10\ \log_{10}(1/\overline{\tau}). \qquad (25)$$

Commonly, it is found that a wall assembly will consist of several elements of different sizes and different sound isolating capabilities, such as a wall containing a door and windows. To determine the average transmission loss of this compound assembly the transmission coefficients of each element are averaged. The resultant value is the composite transmission loss of the overall assembly, and is defined as

$$R_{\mathrm{comp}} = 10\ \log_{10} \frac{S_1 + S_2 + \cdots + S_n}{S_1\tau_1 + S_2\tau_2 + \cdots + S_n\tau_n}, \qquad (26)$$

where R_{comp} is the composite transmission loss (in dB); S_1, S_2, ..., S_n are the areas of the various elements (in m^2); and τ_1, τ_2, ...,τ_n are the transmission coefficients of the various elements (dimensionless).

2.1.2 Homogeneous Walls. Sound transmission through a wall varies with frequency and is dependent upon the transverse vibration velocity response of the wall. A generalized transmission-loss curve for a single-layer, homogeneous partition, such as a window or concrete floor, is shown in Fig. 6 together with the frequency regions affected by the stiffness, resonances, mass, and coincidence effects.

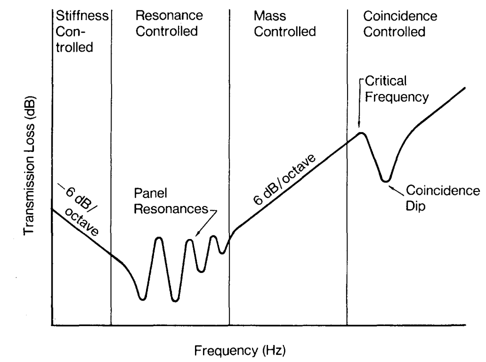

FIG. 6. Generalized sound transmission-loss curve of a single-layer, homogeneous partition mounted in an immobile frame. Different physical characteristics of a material control its effectiveness as a sound insolator in different parts of the frequency spectrum. The sound transmission of most common materials used to construct partitions in buildings is controlled by mass and the coincidence effect. The effects of stiffness and resonance are typically at frequencies below the range of human hearing.

2.1.2.1 Stiffness-Controlled Transmission Loss. At very low frequencies, below the lowest resonant frequency, the panel transmission loss R is considered to be stiffness controlled. The transmission loss increases with decreasing frequency at 6 dB per halving of frequency. However, for walls, floors, and ceilings, the lowest resonance is usually between about 5 and 15 Hz, and generally lower than 10 Hz for large single-pane windows. Thus, stiffness- controlled transmission loss is rarely a factor in architectural acoustics, perhaps with the exception of small panels such as a window in a door.

2.1.2.2 Resonance-Controlled Transmission Loss. At and near the fundamental resonant frequency, the transmission loss is at its lowest, limited only by damping. In practice, this frequency, usually from 5 to 15 Hz, is rarely an acoustic problem, with the possible exceptions of interference to neighbors by amplified music and structural response to ground vibration or sonic booms. The first few resonances above the fundamental resonant frequency can also reduce the transmission loss, but are usually not as strongly excited.

For a simply supported panel, i.e., with fixed edges, the resonant frequency due to bending waves is

$$f_{\text{res}}=\frac{\pi}{2}\left(\frac{B}{\rho_s}\right)^{1/2}\left[\left(\frac{m}{l_x}\right)^2+\left(\frac{n}{l_y}\right)^2\right], \tag{27}$$

where B is the bending stiffness of the panel per unit width (in N m), ρ_s is the mass per unit area of the panel material (in kg m^{-2}), m and n are integers (both finite), and l_x and l_y are the panel edge lengths (in m).

2.1.2.3 Mass Law. At a frequency above approximately twice the lowest resonant frequency, the transmission loss of a simple partition becomes dependent primarily on surface density and is said to be mass controlled. In this frequency region the vibrational response of the wall to sound is resisted primarily by the wall's inertia, and the panel acts as if it consists of many small masses connected only by a limp membrane. The sound energy transmission obeys the "mass law," which states that for each doubling of mass per unit area, or for each doubling of frequency, the transmission loss is increased by 6 dB. The mass law can be approximated by

$$R=20\log_{10}(Mf)-48, \tag{28}$$

where M is the surface mass of the partition (in kg m^{-2}) and f is the frequency of the incident sound (in Hz).

This equation gives the theoretical maximum sound transmission loss for random incidence and should only be used as an approximation.

2.1.2.4 Critical Frequency. For a given angle of incidence of an airborne sound wave there is a particular frequency at which the projected wavelength of incident sound matches the bending wavelength of the flexural wave of the wall. When this matching occurs, the incident energy is readily accepted and radiated, i.e., transmitted, with the only limit in energy transmission due to damping. The minimum frequency at which this matching occurs is for 90° incident sound, and is called the critical frequency, f_c, defined as

$$f_c=\frac{c^2}{2\pi}\left(\frac{\rho_s}{B}\right)^{1/2}, \tag{29}$$

where ρ_s is the mass per unit area of the material (in kg m^{-2}) and B is the bending stiffness per unit width (in N m).

This "coincidence effect" or "coincidence dip" particularly limits the performance of gypsum board constructions and glass windows between 2 and 4 kHz. To control the coincidence effect in the latter, a laminated glass utilizing an inner, plastic damping layer can be used for single panes of glass. In double-glazed assemblies, glass of different thicknesses can be used to achieve critical frequencies with less overlap.

The relationship described by the mass law between sound transmission loss, frequency, and mass can only be used at frequencies up to about one-half of the critical frequency, which is the lowest frequency at which the coincidence effect begins to lessen the transmission loss. Above about ten times the critical frequency for common building materials, the thickness of the panel becomes significant and vibrations other than transverse become significant.

2.1.3 Multiple-Skin Walls. To achieve any significant improvement in the sound transmission loss of a homogeneous or solid wall requires increasing its mass, since, in the most important frequency range of 50–2000 Hz, architectural materials such as glass, concrete, and masonry are "mass controlled." Adding mass to achieve an adequate degree of sound isolation, however, becomes both inefficient acoustically and uneconomical in current building technology. Some of the disadvantages of increasing the mass of a wall include increased wall thickness, with resultant loss of available floor space, and limited effectiveness, due to the law of diminishing returns. Using massive walls on the upper floors of a high-rise building will also significantly increase the cost of the building, since the design of the structural support system must also be made more massive to support this added weight, particularly if lateral seismic forces are anticipated. A high degree of transmission loss can be achieved, however, using lighter weight wall types, such as a stud wall, consisting of multiple layers of material. A stud wall is perhaps the most common type of wall assembly used in the building industry today. It consists of layers of material, typically gypsum board or plaster, installed on each side of a supporting member or stud, usually of metal or wood, typically spaced 40 or 60 cm apart. With this type of wall there are two major paths for transmission of sound energy: airborne radiation from the first panel to the second panel through the air of the stud cavity; and structure-borne transmission of vibrational energy from the first panel to the second panel through direct connection at the studs.

At low frequencies the trapped air of the stud space couples the two leaves of the wall together with a compliance (spring), and therefore, there is a resonant frequency. The frequency at which this mass-air-mass resonance occurs can be determined by

$$f_{\mathrm{res}}=60\left(\frac{w_1 w_2}{w_1+w_2}\,d\right)^{-1/2}, \qquad (30)$$

where f_{res} is the resonant frequency of the air cavity (in Hz), w_1 and w_2 are the surface masses of the leaves of the wall (in kg m^{-2}), and d is the width of air space (in m).

This resonance can be significant, and in practice, the mass-air-mass resonance should be designed to be below 100 Hz so that for leaves of low mass, a wide cavity is required. For a typical wood stud wall, there are also other resonances, usually ranging from 50 to 350 Hz. Note that in order to reduce the effect of the resonances of the wall, a porous sound absorbing material should be installed in the air cavity.

Transmission through the studs and around the fixed edges becomes an important factor when sound isolation greater than a sound transmission class (STC) (see Sec. 2.1.5) rating of 50 is required. Up to this level standard, single-stud walls can be used and the amount of isolation required can be achieved by varying the mass of the leaves and width of air space. Above approximately STC 50, it is usually necessary to vibration isolate the two sides of the wall from each other. One method is to stagger the studs, using a double row of studs, one supporting each side, installed on a common plate. The sound isolation of this type of assembly is limited, however, by the common, fixed top and bottom edges, which limit the vibration isolation. Higher sound isolation can be achieved by constructing a double-stud-wall assembly which consists of a double row of studs, each on its own separate edge plates, so that the two rows of studs are better vibration isolated from each other. Typical sound transmission loss curves for single-stud, staggered-stud, and double-stud walls are shown in Fig. 7.

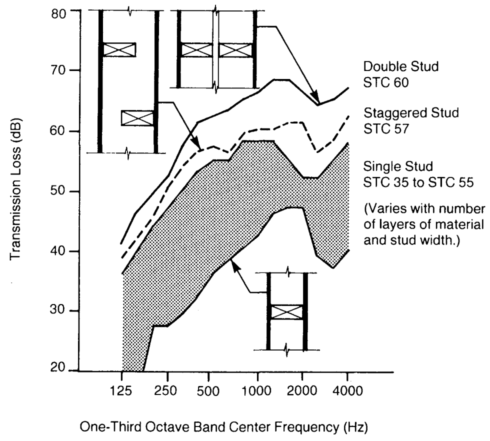

FIG. 7. Typical sound transmission-loss curves for single-stud walls (shaded region), staggered-stud wall (dashed line), and double-stud wall (solid line). Transmission loss increases with the depth of air space between the two leaves of the wall, the degree of isolation of the wall leaves, the number of layers of wall material, and the presence of sound absorbing material in the stud cavity.

Another method of isolation is to use metal studs instead of wood studs. A metal stud is somewhat like the letter "*C*" in cross section, and provides added resilience over that of a wood stud. In common practice, most high-rise buildings use metal studs.

2.1.4 Flanking Transmission. In sound isolation situations it is necessary to consider the total amount of sound energy transmitted to a space by all sound paths, not just through the common barrier. The total degree of sound isolation between spaces is typically limited, not by how much sound is transmitted through the intervening partition, but by how much sound is transmitted by every path, through *and around* the partition. The transmission of sound by indirect paths is called flanking transmission. For a given degree of sound isolation, the most economical construction can be achieved when all paths transmit approximately the same amount of sound energy. Typical sound flanking paths in construction systems include the following: unsealed joints between construction elements, ventilation grilles, interconnecting duct work, common corridors, continuous plenum space above suspended-ceiling systems, electrical outlet boxes located back to back on opposite sides of a wall, wall furnaces, and pipe penetrations.

A very small hole or crack can seriously reduce the effectiveness of a sound isolating construction. Since a hole has a transmission loss of near zero, it takes a very small percentage of hole-to-total wall area to negate the airborne sound isolation of the wall, considering that a residential wall may be required to reduce incident sound energy by 50 dB, or a factor of 100 000. The sound flanking paths must therefore be carefully sealed to achieve the full sound isolating capabilities of the construction assembly. Such leaks can be controlled by inserting solid, impervious barriers in the openings or, at the smaller cracks and holes, by filling with resilient materials and sealants to achieve an airtight closure.

2.1.5 Sound Transmission Class. For convenience in selecting or comparing the airborne sound isolation provided by various building assemblies, a single-number rating system called the sound transmission class (STC) is used. In general, the higher the STC rating the more efficient the construction is for isolating sound. The STC rating is determined by comparing the measured transmission-loss values of an assembly at the 16 one-third octave band-center frequencies from 125 to 4000 Hz with a standard reference contour, which is adjusted based on standard rules. The STC rating system is widely used throughout the U.S. building industry for airborne-sound isolation design and evaluation. However, selecting a building assembly solely on its STC rating does not always ensure a satisfactory degree of sound isolation. The STC rating system was originally developed to evaluate the isolation provided by construction for speech privacy situations. It is therefore not a reliable predictor for situations in which low-frequency isolation, i.e., below 125 Hz, is required, such as for mechanical equipment noise sources, or live or reproduced music.

2.2 Structure-Borne Sound Isolation

The other major transmission path of noise in buildings is as vibration directly through a rigid building structure. The resulting radiated noise is called "structure-borne sound." Structures commonly transmit vibration as shear, flexural (bending), and compressional waves. Structure-borne sound can be generated both by sources which act mechanically on the structure, such as the impact of footsteps, or cupboard doors, as well as by acoustic excitation from airborne sound. The energy may be felt as vibration if the level is high enough. Structure-borne sound may be particularly troublesome where lightweight, lightly damped radiating panels are rigidly attached to the vibrating structure, often far removed from the original source. Structure-borne energy is readily transmitted long distances with little attenuation through typical buildings. This is particularly true if the structural system consists of continuous timber, steel, or concrete framing members or large areas of rigid frame and panel assemblies with only small discontinuities to reflect or absorb energy. It is generally more effective to reduce the impacting or vibrating forces at their source, such as by providing resilient floor finishes, vibration isolation devices on mechanical and electrical equipment, resilient supports for plumbing and piping, etc. If it is necessary to achieve a high degree of impact noise reduction at acoustically sensitive spaces, more elaborate constructions are required, including specially designed, resiliently mounted double, or "floated," floors or resiliently suspended ceiling systems. Placement of vibration breaks in the structure is used most often in wood-frame floors of multifamily dwellings.

2.2.1 Impact Insulation Class. Since the sound generated by impact sources depends on the interaction between a specific source and a particular building structure, it is generally not possible to calculate the airborne noise levels that will be transmitted to an adjacent space. In practice, the degree of impact noise reduction required is specified for that particular type of occupancy based on excitation of the construction by a standardized impact force. In order to compare the effectiveness of various floor-ceiling assemblies for isolating an impact noise source, a single-number rating system is used, the impact insulation class (IIC), which is similar in concept to the sound transmission class rating for airborne sound isolation. The IIC rating is determined by comparing the normalized impact sound pressure level, L_n, in the space below the floor-ceiling assembly being tested against a standard reference spectrum contour. The measured impact-sound pressure-level data are generated using a standard impact-noise source called a "tapping machine," which produces a series of uniform impacts at a uniform rate on the test floor assembly (see Sec. 2.3.2). The IIC ratings should be used with caution, however, since it has been found that evaluations of various floor coverings on test floor constructions using the standard tapping machine do not always provide accurate assumptions for use of the same floor coverings on other floor constructions. In addition, because of its lightweight design, the tapping-machine impact source does not simulate the weight or characteristics of a human walker. Therefore, the "creaks" and "booms" of a flexible floor assembly caused by such footstep excitation is typically not reflected in the single-number IIC rating. Note, however, that at present the tapping-machine impact source is the only standardized device allowable for determining IIC ratings of floor-ceiling assemblies.

2.3 Measurement Procedures

2.3.1 Airborne-Sound Isolation. The transmission-loss values of a building assembly can be obtained by either laboratory or field testing, with the measurement procedures for each specified by various standards. Laboratory measurements are governed nationally, such as by the American Society for Testing and Materials (ASTM) Designation E90, "Standard Method for Laboratory Measurement of Airborne Sound Transmission Loss of Building Partitions," and also by the International Organization for Standardization (ISO) Recommendation R140, "Field and Laboratory Measurements of Airborne and Impact Sound Transmission." The sound reduction index R, or transmission loss (TL), is the ratio of incident to transmitted acoustic energy, and is measured by inserting the test specimen in an opening between two reverberant rooms whose minimum volume is 80 m^3 for the ASTM standard and 50 m^3 for the ISO standard (100 m^3 desirable). The volume and shape of the rooms are designed to achieve a diffuse sound field. Diffusing panels, both stationary and rotating, are also required (see Sec. 1.3.2). The minimum dimension of the test specimen is 2.4 m and the recommended overall size is 10 m^2. There is no specific minimum area requirement for the test specimen because various building elements are tested, e.g., walls, doors, windows, etc., and it is not possible to standardize the size.

In the source room, a near-diffuse sound field is generated and the space-average sound pressure levels L in both the source and receiving rooms are measured. The numerical difference between the measured sound pressure levels provides the noise reduction D at each test frequency:

$$D=L_1-L_2, \tag{31}$$

where L_1 is the average sound pressure level in the source room (in dB) and L_2 is the average sound pressure level in the receiving room (in dB).

The noise reduction represents the sound isolation achieved by the test specimen in that specific configuration, which includes the effects of specimen size and of receiving room absorption. So that the specimen can be compared with similar specimens from other test facilities, corrections are applied to the measured data to negate the effects of the size and absorption by converting the noise-reduction (D) data to transmission-loss (R) values by

$$R=D+10\log_{10}(S/A), \tag{32}$$

where S is the area of the test specimen (in m^2) and A is the sound absorption in the receiving room (in metric sabins). The single-number STC rating of the test specimen is determined from the transmission-loss values ranging from 125 to 4000 Hz (see Sec. 2.1.5).

Field measurements are also controlled by standards, such as ASTM Designation E336, "Standard Test Method for Measurement of Airborne Sound Insulation in Buildings," and ISO R140. However, because it is generally not possible to perform measurements in the field under the same ideal conditions encountered in laboratory test facilities, the sound isolation data can be measured and reported somewhat differently from the laboratory measurements. When measurements are performed between two spaces in the field, the measured noise-reduction values can be corrected for specimen size and absorption to determine the field transmission-loss (R_F) values at each frequency, and the field sound transmission class (FSTC) obtained for that specimen. In this situation the R_F and FSTC values apply to that specific construction assembly tested and are theoretically independent of the surroundings, quality of installation, and other acoustical characteristics of the space. However, the same measured noise-reduction data can also be reported to be measured without corrections and, in this case, the noise isolation class (NIC) rating is obtained. The NIC is determined from the noise-reduction data in exactly the same manner that the STC rating is determined from the adjusted transmission-loss values. The difference between the two rating systems is that the NIC rating provides information on the total amount of airborne sound isolation actually existing between the two rooms, taking into account *all* sound transmission paths, while the FSTC rating evaluates only the performance of a specific building assembly, independent of its surroundings.

2.3.2 Impact-Sound Isolation. As with airborne sound, the measurement procedures for impact-sound isolation are controlled by strict standards. For laboratory measurements these include ASTM Designation E492, "Standard Method of Laboratory Measurement of Impact Sound Transmission through Floor-Ceiling Assemblies Using the Tapping Machine." For field measurements, ASTM E1007 is used. The method for measuring impact noise and determining impact-sound isolation is significantly different from the methods used for airborne-sound isolation. For airborne sound the amount of sound isolation is determined by measuring the difference in sound pressure levels between the source room and receiving room. The greater the difference, the better the isolation. For impact isolation the sound pressure level in the source room is not important, but rather the average impact-sound pressure level L in the space below the floor-ceiling assembly being tested is important. It is measured at the 16 one-third octave band center frequencies from 100 to 3150 Hz by using a tapping machine as the source (see Sec. 2.2.1). The test specimen is inserted in an opening between two reverberant rooms similar to those used for airborne-sound isolation testing, but with the rooms located one above the other instead of side by side. The recommended minimum size of the test specimen is 9.3 m^2. Since the spectrum and level of the measured impact-sound pressure levels depend on the sound absorption in the receiving room, the average impact-sound pressure levels are normalized to a reference absorption, so that the results can be compared with those obtained in other spaces with different amounts of absorption by

$$L_n = L_r - 10 \log_{10}(A_0/A), \tag{33}$$

where L_n is the normalized impact-sound pressure level (in dB), L_r is the average impact-sound pressure level in the receiving room (in dB), A_0 is the reference absorption equal to 10 metric sabins, and A is the sound absorption in the receiving room (in metric sabins).

The single-number impact insulation class (IIC) rating is determined from the normalized L_n data by comparing them to a standard reference contour, similar to the method used to determine the sound transmission class (STC) rating for airborne-sound isolation. The higher the IIC rating, the more efficient the floor-ceiling assembly is for isolating impact sound.

GLOSSARY

Air Absorption: Absorption of sound energy caused by molecular relaxation of air molecules. Air absorption is significant only in large rooms and at sound frequencies of 2000 Hz and above.

Airborne Sound: Sound radiated initially into and transmitted through the air rather than through the structure of a building.

Ambient Noise: The all-encompassing noise associated with a given environment, being usually a composite of sounds from many sources near and far.

Anechoic Room: A room whose boundaries absorb effectively all incident sound, thereby providing essentially a free-field condition within a specified range of frequencies.

Architectural Acoustics: The area of acoustics that deals with the effects of sound and vibration in buildings and its effects on the building users. The main concerns in architectural acoustics are room acoustics, sound isolation, and mechanical equipment noise and vibration control.

Assisted Resonance: Artificial enhancement of sound reflections and reverberation time by electronic means.

Critical Frequency: The frequency at which the acoustic wavelength in air exactly matches a transverse bending wavelength of a sound isolating structure.

Dead Room: A space that is characterized by an unusually large amount of sound absorption.

Decibel (dB): A dimensionless, logarithmic unit denoting the ratio of a quantity to a reference quantity of the same kind, both proportional to energy or power, and given by 10 $\log_{10}(q_1/q_{\text{ref}})$. Examples of quantities include sound power, sound intensity, sound pressure squared, and particle or vibration velocity squared. For systems adequately described by linear equations, force-like quantities, such as sound pressure, acceleration, and force at levels usually found in architectural acoustics are usually used directly. For example, sound pressure (in dB) is

$$L_P = 10 \log_{10}(P^2/P_{\text{ref}}^2) = 20 \log_{10}(P/P_{\text{ref}}),$$

where P and P_{ref} are root mean square (rms) pressures. A commonly used approximation substitutes linear averages, zero-to-peak, and other wave-form descriptors for the rms. (Strictly speaking, a decibel is one-tenth of a bel, but the latter is now rarely used.)

Diffraction: The change in the direction of a sound wave, other than by reflection or refraction, which occurs when the wave encounters an obstacle or nonhomogeneity in a medium.

Diffuse Sound Field: A theoretical sound field in an enclosure in which the time average of the mean square sound-pressure level is the same throughout and the flow of energy in all directions is equally probable.

Direct Sound Field: The field of unreflected sound that exists close to a sound source.

Echo: In architectural acoustics, a wave of sound that has been reflected or otherwise returned with an adequately uniform and long delay and with sufficient magnitude to be detected as a sound that is distinct from the directly transmitted sound and its smoothly decaying reverberation.

Edge Effect: The excess or increased sound absorption of a test sample relative to what would be measured for an arbitrarily large sample of the same material. This effect is due to diffraction effects at the sample's edges or to discontinuities in the surface impedance.

Energy Density: The sound energy at a point in a sound field contained in a given infinitesimal part of the medium divided by the volume of that part of the medium.

Flanking Transmission: The transmission of sound from one space to another by indirect paths rather than directly through an intervening construction.

Helmholtz Resonator: A resonant acoustic device consisting of an air cavity connected to a larger space by a relatively small opening or neck.

Impact Insulation Class (IIC): A single-number rating used to evaluate the impact-sound isolating performance of a floor-ceiling construction system.

Live Room: A space that is characterized by an unusually small amount of sound absorption.

Loudness: The subjective response of human hearing to sound levels. The perception of loudness also depends on the frequency and wave form of the sound.

Masking: The ability of one sound to inhibit the perception of another sound. Also, the use of an unobtrusive background noise to cover other intruding sounds.

Mass-Air-Mass Resonance: The resonant system comprising the enclosed air between the two leaves of a partition and the mass of the two leaves. This resonance can result in a pronounced dip in the sound transmission loss, usually at low frequencies.

Mass Law: The sound transmission loss of a panel structure will increase by 6 dB for every doubling of mass per unit area. The mass law is reasonably accurate only in the absence of significant resonances and in the frequency region above twice the lowest resonant frequency and below one-half the critical frequency.

Mean Free Path: The average distance that a sound travels between successive reflections in an enclosure.

Natural Frequency: A frequency of free oscillation of an undamped system. The natural frequencies are the frequencies of the normal modes of vibration. See also resonance frequency.

Noise Isolation Class (NIC): A single-number rating used to evaluate the airborne-sound isolating performance of a completed construction system, determined from the measured noise-reduction values. See also sound transmission class.

Noise Reduction (NR): The reduction in level of a sound by any of several means. Also, the measured difference, in a specified frequency band, between the average sound pressure levels in two enclosed spaces due to a source in one of them and measured without adjustment for sound absorption or flanking noise paths. See also transmission loss.

Noise Reduction Coefficient (NRC): The arithmetic average of the sound absorbing coefficients of a material at 250, 500, 1000, and 2000 Hz, rounded to the nearest 0.05.

Octave: The interval between two sound frequencies having a ratio of two.

Octave Band: A frequency range that is one octave wide. Standard octave bands are designated by their center frequencies.

Pink Noise: A random or pseudorandom broadband noise whose spectrum level decreases with increasing frequency to yield constant time-averaged energy per constant percentage bandwidth, such as one-third or whole octaves. See also white noise.

Pitch: The subjective response of the human hearing mechanism to the perceived dominant frequency of a sound. The perception of pitch also depends on the sound pressure level and wave form of the sound.

Pure Tone: A simple sound whose instantaneous wave form is that of a sine wave, i.e., consisting of only one frequency.

Quadratic Residue Diffusor: A sound diffusing element consisting of a reflection phase grating that provides wide angles of sound dispersion within its designed frequency range.

Resonance Frequency: Frequency of oscillation at which the response of the system peaks. See also natural frequency.

Reverberant Sound Field: The sound field in an enclosure in which all of the sound waves have been reflected at least once from the boundary surfaces.

Reverberation: The persistence of sound in an enclosed space, as a result of multiple reflections or scattering, after the sound source has stopped.

Reverberation Time (T_{60}): The time in seconds required for the mean square sound pressure level of a sound originally in a steady state and within a specified frequency band to decay by 60 dB after the sound source has stopped.

Room Acoustics: The study of sound in an enclosure, including the sound in the air and the effects on the sound by the enclosure.

Room Modes: The normal modes of acoustic vibration in an enclosure that describe standing waves that may occur near the characteristic natural frequencies based on the size and shape of the enclosure.

Room Radius (r_r): The nominal distance from an ideally uniform sound source in an enclosure at which the sound pressure levels of both the direct and reverberant sound fields are equal.

Sabin ($S\alpha$): The dimensionless unit measure of sound absorption. A metric sabin is equal to the absorption provided by one square meter of a perfect absorber (i.e., one with an absorption coefficient of 1.0).

Sound Absorption: The change of sound energy into some other form, usually heat, in passing through a medium, including air, or striking a material or object.

Sound Absorption Coefficient (α): The fraction of incident sound energy absorbed by a material. Unless otherwise specified, a diffuse sound field is assumed.

Sound Intensity: The average sound power transmitted in a specified direction and location through a plane perpendicular to the specified direction and normalized to a unit area.

Sound Isolation: The use of building materials or construction systems to reduce the transmission of sound. Also, the degree of reduction achieved, as measured according to various standardized methods and expressed using corresponding standardized ratings.

Sound Power Level (L_w): The total sound energy per unit of time, such as produced by a sound source, usually expressed in decibels. For airborne sound, this is equal to 10 $\log_{10}$ of the ratio of the power of a sound in watts to the standard reference power of 10^{-12} W.

Sound Pressure Level (L_p): As used in architectural acoustics, a value expressed in decibels. For airborne sound, equal to 20 $\log_{10}$ of the time average of the root mean square of the instantaneous sound pressure divided by the reference pressure of 20 μPa.

Sound Reduction Index (R): See transmission loss.

Sound Reflection: The return of sound energy from a discontinuity in a medium, such as a boundary surface.

Sound Transmission Class (STC): A single-number rating used to evaluate the airborne-sound isolating performance of a specific construction element, determined from the transmission-loss values by standardized procedures. See also noise isolation class.

Standing Waves: See room modes.

Steady-State Sound Field: The sound field in an enclosure in which the time-averaged sound pressure level at any point is constant.

Structure-Borne Sound: That portion of sound in a room that has been transmitted from elsewhere through a solid medium, such as a building structure, rather than only through air.

Subjective Preference Scaling: Any of various systems used either to evaluate or relate objective physical properties of the sound field with subjective characteristics in music performance spaces.

Threshold of Audibility: The minimum sound pressure level at which a person can hear a sound of a given frequency. In children without auditory pathology, the threshold is about 20 μPa or 0 dB at 2000 Hz.

Threshold of Pain: The minimum sound pressure level at which a person begins to sense pain in the ear from sounds of high intensity. Pain or a tickling sensation is reported at 140 dB, while sounds at 120 dB are discomforting.

Transmission Coefficient (τ): The ratio of transmitted acoustic energy to the incident acoustic energy in airborne-sound isolation situations. This varies with frequency and angle of incidence.

Transmission Loss (R): The ratio, expressed in decibels, of the incident to transmitted airborne sound of a specified angle of incidence and a specified frequency content as it passes through a building construction. Strictly speaking, transmission loss must be determined in accordance with standardized procedures which take into account the size of the construction and the amount of sound absorption present. This is also called the sound reduction index.

Weighting Network: An electronic network that modifies the frequency response of a standard sound level meter, usually to enable it to approximate the subjective spectral response of human hearing to various sounds. Most frequently used is the "A" network, with the sound level designated as dBA.

White Noise: A broadband noise whose energy per unit bandwidth remains constant at all frequencies within the total bandwidth. See also pink noise.

Works Cited

Ando, Y. (1985), *Concert Hall Acoustics*, Berlin: Springer-Verlag.

Barron, M. (1971), *J. Sound Vib.* **15**, 475–494.

Harris, C. M. (1966), *J. Acoust. Soc. Am.* **40**, 148–159.

Further Reading

12th International Congress on Acoustics (1986), *Proceedings of the Vancouver Symposium on Acoustics and Theater Planning for the Performing Arts*, Ottawa: The Canadian Acoustical Society.

Beranek, L. L., Kopek, J. W. (1981), *J. Acoust. Soc. Am.* **69**, 1–16.

Beranek, L. L. (1962), *Music, Acoustics & Architects*, New York: Wiley.

Beranek, L. L. (Ed.) (1971), *Noise and Vibration Control*, New York: McGraw-Hill.

Beranek, L. L. (1975), *J. Acoust. Soc. Am.* **57**, 1258–1262.

Berendt, R. D., Corliss, E. L. R. (1976), *Quieting: A Practical Guide to Noise Control*, Natl. Bur. Stand. Handbook No. 119, Washington, DC: U.S. GPO.

Cavanaugh, W. J., Farrell, W. R., Hirtle, P. W., Watters, B. G. (1962), *J. Acoust. Soc. Am.* **34**, 475–492.

Cremer, L., Muller, H. A. (1982), *Principles and Applications of Room Acoustics*, London: Applied Science.

Cremer, L. (1989), *J. Acoust. Soc. Am.* **85**, 1213–1225.

Day, B. F., Ford, R. D., Lord, P. (Eds.) (1969), *Building Acoustics*, London: Elsevier.

Egan, M. D. (1988), *Architectural Acoustics*, New York: McGraw-Hill.

Forsyth, M. (1985), *Buildings for Music*, Cambridge: MIT.

Ginn, K. B. (1978), *Architectural Acoustics*, Bruel & Kjaer.

Green, D. W., Sherry, C. W. (1982), *J. Acoust. Soc. Am.* **71**, 90–96; **71**, 902–907.

Harris, C. M. (Ed.) (1979), *Handbook of Noise Control*, New York: McGraw-Hill.

Hedeen, R. A. (1980), *Compendium of Materials for Noise Control*, HEW Publication No. (NIOSH) 80-116, Washington, DC: U.S. GPO.

Jordan, V. L. (1980), *Acoustical Design of Con-*

cert Halls and Theatres, London: Applied Science Publishers.

Jordan, V. L. (1981), *Appl. Acoust.* **14**, 253–266.

Knudsen, V. O., Harris, C. M. (1950), *Acoustical Designing in Architecture,* New York: Wiley.

Kuttruff, H. (1979), *Room Acoustics,* London: Applied Science.

Lawrence, A. B. (1970), *Architectural Acoustics,* London: Elsevier.

Mackenzie, R. (Ed.) (1975), *Auditorium Acoustics,* New York: Wiley.

Ramsey, C. G., Sleeper, H. R. (Eds.) (1988), *Architectural Graphic Standards,* New York: Wiley.

Rettinger, M. (1973), *Acoustic Design and Noise Control,* New York: Chemical.

Sabine, H. J., Lacher, M. B. (1975), *Acoustical and Thermal Performance of Exterior Residential Walls, Doors and Windows,* Building Science Series No. 77, Washington, DC: U.S. GPO.

Sabine, W. C. (1922, reprinted 1964), *Collected Papers on Acoustics,* New York: Dover.

Schroeder, M. R., Gottlob, D., Siebrasse, K. F. (1974), *J. Acoust. Soc. Am.* **56**, 1195–1201.

Schroeder, M. R. (1975), *J. Acoust. Soc. Am.* **57**, 149–150.

Schultz, T. J. (1965), *IEEE Spectrum* **2**(6), 56–67.

Storer, R. A. (Ed.) (1990), *Annual Book of ASTM Standards, Vol. 04.06 Environmental Acoustics,* Philadelphia: American Society for Testing and Materials.

Yerges, L. F. (1978), *Sound, Noise, and Vibration Control,* New York: Van Nostrand Reinhold.

ACOUSTICS, ATMOSPHERIC

See ATMOSPHERIC ACOUSTICS

ACOUSTICS, LINEAR

ALLAN D. PIERCE, *Graduate Program in Acoustics, Pennsylvania State University*

Introduction 138

1. **Basic Equations of Continuum Mechanics** 139
1.1 Mass, Momentum, and Energy Equations 139
1.2 Newtonian Fluids 140
1.3 Equilibrium Thermodynamics 140
1.4 Bulk Viscosity and Thermal Conductivity 140
1.5 Navier–Stokes–Fourier Equations 141
1.6 Thermodynamic Coefficients 141
1.7 Ideal Compressible Fluids 142
1.8 Elastic Solids 142

2. **Equations of Linear Acoustics** 143
2.1 The Linearization Process 143
2.2 Linearized Equations for an Ideal Fluid 143
2.3 The Wave Equation 144
2.4 Wave Equations for Isotropic Elastic Solids 144
2.5 Linearized Equations for a Viscous Fluid 144
2.6 Acoustic, Entropy, and Vorticity Modes 145
2.7 Boundary Conditions at Interfaces 147

3. **Waves of Constant Frequency** 147
3.1 General 147
3.2 Spectral Density 148
3.3 Complex-Number Representation 149

4. **Plane Waves** 149
4.1 Plane Waves in Fluids 149
4.2 Plane Waves in Solids 150

5. **Attenuation of Sound** 150
5.1 General 150
5.2 Classical Absorption 151
5.3 Relaxation Processes 151
5.4 Attenuation of Sound in Air 153
5.5 Attenuation of Sound in Seawater 154

6. **Acoustic Intensity and Power** 154
6.1 General 154
6.2 Acoustic Energy Density and Intensity 155
6.3 Acoustic Power 155
6.4 Rate of Energy Dissipation 156
6.5 Energy Corollary for Elastic Waves in Solids 156

7. **Cylindrical and Spherical Waves** 156
7.1 Spherical Waves 156
7.2 Cylindrical Waves 157

8. **Simple Sources of Sound, Green's Functions** 157
8.1 Volume Sources 157
8.2 Small Piston in a Rigid Baffle . 158
8.3 Multiple and Distributed Sources 158
8.4 Thermoacoustic Sources 158
8.5 Green's Functions 159

9. **Dipoles, Quadrupoles, and Spherical Harmonics** 159
9.1 Multipole Series 159
9.2 Acoustically Compact Sources. 160
9.3 Spherical Harmonics 160

10. **Integral Equations in Acoustics** 161
10.1 The Helmholtz–Kirchhoff Integral 161
10.2 Integral Equations for Fields on the Surface 162

11. **Impedance** 163
11.1 Mechanical Impedance 163
11.2 Specific Acoustic Impedance 163
11.3 Characteristic Impedance 163
11.4 Radiation Impedance 163
11.5 Acoustic Impedance 164

12. **Reflection, Refraction, and Transmission** 164
12.1 Reflection at a Plane Surface 164
12.2 Reflection at an Interface 164
12.3 Theory of the Impedance Tube 165

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

12.4 Transmission through Walls and Slabs 165
12.4.1 Transmission through Limp Plates 166
12.4.2 Transmission through Porous Blankets 166
12.4.3 Transmission through Elastic Plates 166
13. Waveguides, Ducts, and Resonators 167
13.1 Guided Modes in a Duct 167
13.2 Cylindrical Ducts 167
13.3 Low-Frequency Model for Duct Propagation 168
13.4 Sound Attenuation in Ducts ... 168
13.5 Mufflers and Acoustic Filters 169
13.5.1 Nonreflecting Dissipative Muffler 169
13.5.2 Expansion Chamber Muffler 169
13.6 Helmholtz Resonators 170
14. Ray Acoustics 170
14.1 Wave-Front Propagation 170
14.2 Reflected, Transmitted, and Diffracted Rays 171
14.3 Wave Equation for Inhomogeneous Moving Media 171
14.4 The Eikonal Approximation 172
14.5 Amplitude Variation in Rectilinear Propagation 173
15. Parabolic-Equation Approximations 173
Glossary 174
Works Cited 179
Further Reading 181

INTRODUCTION

Acoustics, the science of waves of a mechanical nature, is typically concerned with small-amplitude phenomena that can be described to a good approximation in terms of linear differential equations. The analysis that accompanies most applications of acoustics consequently draws upon a linearized theory, briefly referred to as linear acoustics. The present article presents the underlying principles of linear acoustics, and discusses some general applications of the theory.

The reason why one chooses to view acoustics as a linear phenomenon, unless there are strong indications of the importance of nonlinear behavior, is the intrinsic conceptual simplicity brought about by the principle of superposition. Thus, for example, a complicated sound field caused by several sources can be regarded as a sum of fields, each of which is caused by an individual source. Moreover, if one has some assurance that the governing equations have time-independent coefficients, then the concept of frequency has great utility. Waves of each frequency propagate independently, so one can analyze the generation and propagation of each frequency separately, and then combine the result.

Nevertheless, linear acoustics should be always regarded as an approximation, and the understanding of nonlinear aspects of acoustics is of increasing importance in modern applications (see ACOUSTICS, NONLINEAR). Consequently, part of the task of the present article is to explain in what sense linear acoustics is an approximation. An extensive discussion of the extent to which the linearization approximation is valid is not attempted, but a discussion is given of the manner in which the linear equations result from nonlinear equations that are regarded as more nearly exact.

The historical origins (Hunt, 1978) of linear acoustics reach back to antiquity, to Pythagoras and Aristotle (Lindsay, 1972) and to other ancient civilizations. The mathematical theory began with Mersenne, Galileo, and Newton, and developed into its more familiar form during the time of Euler and Lagrange (Truesdell, 1955). Prominent contributors during the 19th century include Poisson, Laplace, Cauchy, Green, Stokes, Helmholtz, Kirchhoff, and Rayleigh. The last-named's book, *The Theory of Sound* (Rayleigh, 1878), is still widely read and quoted today. The subject is now a mature mainstay of applied physics, with a seemingly unbounded realm of application, and with a voluminous and ever growing scientific and technical literature. The maturity of linear acoustics, its long history, and its voluminous literature can at times be intimidating to a researcher, but the basic principles can be expressed relatively concisely, and the diversity of novel applications continues to afford ample opportunity for creative contributions.

1. BASIC EQUATIONS OF CONTINUUM MECHANICS

1.1 Mass, Momentum, and Energy Equations

The primary equations governing sound are those that account for conservation of mass and energy, and for changes in momentum. These may be written in the form of either partial differential equations or integral equations. The former version is the customary starting point for the derivation of approximate equations for linear acoustics. Extensive discussions of the equations of continuum mechanics can be found in texts by Thompson (1972), Fung (1965), Shapiro (1953), Batchelor (1967), Truesdell (1966), and Landau and Lifshitz (1959), and in the encyclopedia article by Truesdell and Toupin (1960).

The conservation of mass is described by the partial differential equation,

$$\frac{\partial \rho}{\partial t}+\nabla\cdot(\rho\mathbf{v})=0, \tag{1}$$

and the generalization of Newton's second law to a continuum is described by Cauchy's equation of motion, which is written in Cartesian coordinates as

$$\rho\frac{D\mathbf{v}}{Dt}=\sum_{ij}\mathbf{e}_i\frac{\partial\sigma_{ij}}{\partial x_j}+\mathbf{g}\rho. \tag{2}$$

Here the Eulerian description is used with each field variable regarded as a function of actual spatial position coordinates and time. The quantity ρ is the density of matter, mass per unit volume; $\mathbf{v}$ is the particle or fluid velocity; the σ_{ij} are the Cartesian components of the stress tensor. The quantities $\mathbf{e}_i$ are the unit vectors in a Cartesian coordinate system. The second term on the right-hand side of Eq. (2) is a body-force term associated with gravity, with $\mathbf{g}$ representing the vector acceleration due to gravity. In some instances, it may be appropriate to include body-force terms associated with external electromagnetic fields, but such are excluded from consideration in the present article. The time derivative operator on the left-hand side of Eq. (2) is Stokes's total time derivative operator (Stokes, 1845),

$$\frac{D}{Dt}=\frac{\partial}{\partial t}+\mathbf{v}\cdot\nabla, \tag{3}$$

with the two terms corresponding to (i) the time derivative as would be seen by an observer at rest and to (ii) the convective time derivative. This total time derivative applied to the particle velocity field yields the particle acceleration field, so the right-hand side of Eq. (2) can be regarded as the apparent force per unit volume on an element of the continuum.

The stress tensor's Cartesian component σ_{ij} can be regarded as the ith component of the surface force per unit area on a segment of the surface of a small element of the continuum, when the unit outward normal to the surface is in the jth direction. This surface force is caused by interactions with the neighboring particles just outside the surface or by the transfer of momentum due to diffusion of molecules across the surface. Considerations of the net torque that such surface forces would exert on a small element, and the requirement that the angular acceleration of the element be finite in the limit of very small element dimensions, lead to the conclusion

$$\sigma_{ij}=\sigma_{ji}, \tag{4}$$

so the stress tensor is symmetric.

The equations described above are supplemented by an energy conservation relation,

$$\rho\frac{D}{Dt}\left(\frac{1}{2}v^2+u\right)=\sum_{ij}\frac{\partial}{\partial x_j}(\sigma_{ij}v_i)+\rho\mathbf{g}\cdot\mathbf{v}-\nabla\cdot\mathbf{q}, \tag{5}$$

which involves the internal energy u per unit mass and the heat flux vector $\mathbf{q}$. The latter has the interpretation that its dot product with any unit vector represents the heat energy flowing per unit time and per unit area across any internal surface with the corresponding unit normal. An alternative version,

$$\rho\frac{Du}{Dt}=\sum_{ij}\sigma_{ij}\frac{\partial v_i}{\partial x_j}-\nabla\cdot\mathbf{q}, \tag{6}$$

of the energy equation results when the vector dot product of $\mathbf{v}$ with Eq. (2) is subtracted from Eq. (5). [In carrying through the derivation, one recognizes that $\mathbf{v}\cdot\{\rho D\mathbf{v}/Dt\}$ is $\frac{1}{2}\rho Dv^2/Dt$.]

1.2 Newtonian Fluids

The above equations apply to both fluids and solids. Additional equations needed to complete the set must take into account the properties of the materials that make up the continuum. Air and water, for example, are fluids, and their viscous behavior corresponds to that for Newtonian fluids, so that the stress tensor is given by

$$\sigma_{ij}=\sigma_n\delta_{ij}+\mu\phi_{ij}, \tag{7}$$

which involves the rate-of-shear tensor,

$$\phi_{ij}=\frac{\partial v_i}{\partial x_j}+\frac{\partial v_j}{\partial x_i}-\frac{2}{3}\,\nabla\cdot\mathbf{v}\delta_{ij}. \tag{8}$$

The quantity σ_n that appears here is the average normal stress component and μ is the shear viscosity, while δ_{ij} is the Kronecker delta, equal to unity if the two indices are equal and otherwise equal to zero.

1.3 Equilibrium Thermodynamics

Further specification of the equations governing sound requires some assumptions concerning the extent to which equilibrium or near-equilibrium thermodynamics applies to acoustic disturbances. In the simplest idealization, which neglects relaxation processes, one assumes that the thermodynamic variables describing the perturbations are related just as for quasiequilibrium processes, so one can use an equation of state of the form

$$s=s(u,\rho^{-1}), \tag{9}$$

which has the corresponding differential relation (Landau, Lifshitz, and Pitaevskii, 1980)

$$T\,ds=du+pd(\rho^{-1}) \tag{10}$$

in accordance with the second law of thermodynamics. Here s is the entropy per unit mass, T is the absolute temperature, and p is the absolute pressure. The reciprocal of the density, which appears in the above relations, is recognized as the specific volume.

For the case of an ideal gas with temperature-independent specific heats, which is a common idealization for air, the function in Eq. (9) is given by

$$s=(R_0/M)\,\ln(u^{1/(\gamma-1)}\rho^{-1})+s_0. \tag{11}$$

Here, s_0 is independent of u and ρ^{-1}, while M is the average molecular weight (average molecular mass in atomic mass units), and R_0 is the universal gas constant (equal to Boltzmann's constant divided by the mass of an atomic mass unit), equal to 8314 J/kg K. The quantity γ is the specific-heat ratio, equal to approximately 7/5 for diatomic gases, to 5/3 for monatomic gases, and to 9/7 for polyatomic gases whose molecules are not collinear. For air, γ is 1.4 and M is 29.0. [The expression given here for entropy neglects the contribution of internal vibrations of diatomic molecules, which cause the specific heats and γ to depend slightly on temperature. When one seeks to explain the absorption of sound in air, a nonequilibrium entropy (Pierce, 1978) is used which depends on the fractions of O_2 and N_2 molecules that are in their first excited vibrational states in addition to the quantities u and ρ^{-1}.]

For other substances, a knowledge of the first and second derivatives of s, evaluated at some representative thermodynamic state, is sufficient for most linear acoustics applications.

1.4 Bulk Viscosity and Thermal Conductivity

The average normal stress must equal the negative of the thermodynamic pressure that enters into Eq. (10), when the fluid is in equilibrium. Consideration of the requirement that the equations be the same regardless of the choice of directions of the coordinate axes plus the assumption of a Newtonian fluid lead to the relation (Stokes, 1845)

$$\sigma_n=-p+\mu_B\nabla\cdot\mathbf{v} \tag{12}$$

which involves the bulk viscosity μ_B. Also, the Fourier model for heat conduction (Fourier, 1822) requires that the heat flux vector $\mathbf{q}$ be proportional to the gradient of the temperature,

$$\mathbf{q}=-\kappa\nabla T, \tag{13}$$

where κ is the coefficient of thermal conduction.

Transport Properties of Air. Regarding the values of the viscosities and the thermal conductivity that enter into the above expressions, the values for air are approximated by

(Hilsenrath, 1960; Sutherland, 1893; Greenspan, 1959)

$$\mu_S=\mu_0\left(\frac{T}{T_0}\right)^{3/2}\frac{T_0+T_S}{T+T_S}, \tag{14}$$

$$\mu_B=0.6\mu_S, \tag{15}$$

$$\kappa=\kappa_0\left(\frac{T}{T_0}\right)^{3/2}\frac{T_0+T_A e^{-T_B/T_0}}{T+T_A e^{-T_B/T}}. \tag{16}$$

Here T_S is 110.4 K, T_A is 245.4 K, T_B is 27.6 K, T_0 is 300 K, μ_0 is 1.846×10^{-5} kg/(m s), and κ_0 is 2.624×10^{-2} W/(m K).

Transport Properties of Water. The corresponding values of the viscosities and the thermal conductivity of water are given by (Horne, 1969; Pinkerton, 1947)

$$\mu_S=1.002\times 10^{-3}e^{-0.0248\Delta T}, \tag{17}$$

$$\mu_B=3\mu_S, \tag{18}$$

$$\kappa=0.597+0.0017\Delta T-7.5\times 10^{-6}(\Delta T)^2. \tag{19}$$

The quantities that appear in these equations are understood to be in mks units, and ΔT is the temperature relative to 283.16 K (10 °C).

1.5 Navier–Stokes–Fourier Equations

The assumptions represented by Eqs. (9), (10), (12), and (13) cause the governing continuum mechanics equations to reduce to

$$\frac{\partial\rho}{\partial t}+\nabla\cdot(\rho\mathbf{v})=0, \tag{20}$$

$$\rho\frac{D\mathbf{v}}{Dt}=-\nabla p+\nabla(\mu_B\nabla\cdot\mathbf{v})+\sum_{ij}\mathbf{e}_i\frac{\partial}{\partial x_j}(\mu\phi_{ij})+\mathbf{g}\rho, \tag{21}$$

$$\rho T\frac{Ds}{Dt}=\frac{1}{2}\mu\sum_{ij}\phi_{ij}^2+\mu_B(\nabla\cdot\mathbf{v})^2+\nabla\cdot(\kappa\nabla T), \tag{22}$$

which are known as the Navier–Stokes–Fourier equations for compressible flow.

1.6 Thermodynamic Coefficients

An implication of the existence of an equation of state of the form of Eq. (9) is that any thermodynamic variable can be regarded as a function of any other two (independent) thermodynamic variables. The pressure p, for example, can be regarded as a function of ρ and T, or of s and ρ. In the expression of differential relations, such as that which gives dp as a linear combination of ds and $d\rho$, it is helpful to express the coefficients in terms of a relatively small number of commonly tabulated quantities. A standard set of such quantities includes the square of the sound speed.

$$c^2=\left(\frac{\partial p}{\partial\rho}\right)_s, \tag{23}$$

the specific heat at constant pressure,

$$c_p=T\left(\frac{\partial s}{\partial T}\right)_p, \tag{24}$$

the specific heat at constant volume,

$$c_v=T\left(\frac{\partial s}{\partial T}\right)_\rho, \tag{25}$$

and the coefficient of thermal expansion,

$$\beta=\rho\left(\frac{\partial(1/\rho)}{\partial T}\right)_p. \tag{26}$$

(The subscripts on the partial derivatives indicate the independent thermodynamic quantities that are kept fixed during the differentiation.)

The coefficients that are given here are related by the thermodynamic identity,

$$\gamma-1=T\beta^2c^2/c_p, \tag{27}$$

where γ is the specific-heat ratio,

$$\gamma=c_p/c_v. \tag{28}$$

In terms of the thermodynamic coefficients defined above, the differential relations of principal interest in acoustics are

$$d\rho=\frac{1}{c^2}\,dp-\left(\frac{\rho\beta T}{c_p}\right)ds, \tag{29}$$

$$dT=\left(\frac{T\beta}{\rho c_p}\right)dp+\left(\frac{T}{c_p}\right)ds. \tag{30}$$

Thermodynamic Coefficients for an Ideal Gas. For an ideal gas, which is the common idealization for air, it follows from Eqs. (9) and (10) (with the abbreviation R for R_0/M) that the thermodynamic coefficients are given by

$$c^2=\gamma RT=\gamma p/\rho, \tag{31}$$

$$c_p=\gamma R/(\gamma-1), \tag{32}$$

$$c_v=R/(\gamma-1), \tag{33}$$

$$\beta=1/T. \tag{34}$$

For air, R is 287 J/kg K and γ is 1.4, so c_p and c_v are 1005 and 718 J/kg K, respectively. A temperature of 293.16 K and a pressure of 10^5 Pa yield a sound speed of 343 m/s and a density ρ of 1.19 kg/m^3.

Thermodynamic Properties of Water. For pure water, the sound speed is approximately given in mks units by (Wilson, 1959)

$$c=1447+4.0\Delta T+1.6\times 10^{-6}p. \tag{35}$$

Here c is in meters per second, ΔT is temperature relative to 283.16 K (10 °C), and p is absolute pressure in pascals. The pressure and temperature dependence of the density is approximately given by (Horne, 1969)

$$\rho\approx 999.7+0.048\times 10^{-5}p-0.088\Delta T$$

$$-0.007(\Delta T)^2, \tag{36}$$

the coefficient of thermal expansion is given by

$$\beta\approx(8.8+0.022\times 10^{-5}p+1.4\Delta T)\times 10^{-5}, \tag{37}$$

and the coefficient of specific heat at constant pressure is approximately given by

$$c_p\approx 4192-0.40\times 10^{-5}p-1.6\Delta T. \tag{38}$$

The specific-heat ratio is very close to unity, the deviation described by

$$(\gamma-1)/\gamma\approx 0.0011(1+\Delta T/6+0.0024$$

$$\times 10^{-5}p)^2 \tag{39}$$

The values for sea water are somewhat different, because of the presence of dissolved salts. An approximate expression for the speed of sound in sea water is given by (Wilson, 1960)

$$c\approx 1490+3.6\Delta T+1.6\times 10^{-6}p+1.3\Delta S. \tag{40}$$

Here ΔS is the deviation of the salinity in parts per thousand from a nominal value of 35.

1.7 Ideal Compressible Fluids

If viscosity and thermal conductivity are neglected at the outset, then the Navier-Stokes equation, Eq. (21), reduces to

$$\rho\frac{D\mathbf{v}}{Dt}=-\nabla p+\mathbf{g}\rho, \tag{41}$$

which is known as the Euler equation. The energy equation, Eq. (22), reduces to the isentropic flow condition

$$\frac{Ds}{Dt}=0. \tag{42}$$

Moreover, the latter in conjunction with the differential equation of state, Eq. (29), yields

$$\frac{Dp}{Dt}=c^2\frac{D\rho}{Dt}. \tag{43}$$

The latter combines with the conservation-of-mass relation, Eq. (1), to yield

$$\frac{Dp}{Dt}+\rho c^2\nabla\cdot\mathbf{v}=0. \tag{44}$$

In many applications of acoustics, it is these equations that are used as a starting point.

1.8 Elastic Solids

For analyses of acoustic wave propagation in solids, the idealization of a linear elastic isotropic solid (Crandall, Dahl, and Lardner, 1978; Landau and Lifshitz, 1970; Love, 1944; Fung, 1965; Achenbach, 1975; Lindsay, 1960) is often used. The solid-particle displacements are regarded as sufficiently small that convective time derivatives can be neglected, so the Cauchy equation, given previously by Eq. (2), reduces, with the omission of the invariably negligible gravity term, to an equation whose ith Cartesian component is

$$\rho\frac{\partial^2\xi_i}{\partial t^2}=\sum_{j=1}^{3}\frac{\partial\sigma_{ij}}{\partial x_j}. \tag{45}$$

Here ρ is the ambient density; $\xi_i(\mathbf{x},t)$ is the ith Cartesian component of the displacement of the particle nominally at $\mathbf{x}$. This displacement is measured relative to the particle's ambient position.

The stress tensor components σ_{ij} are taken to be linearly dependent on the strain components,

$$\epsilon_{ij}=\frac{1}{2}\left(\frac{\partial\xi_i}{\partial x_j}+\frac{\partial\xi_j}{\partial x_i}\right). \tag{46}$$

The linear stress-strain relation for an isotropic elastic solid has the form

$$\sigma_{ij}=2\mu_{\mathrm{L}}\,\epsilon_{ij}+\lambda_{\mathrm{L}}\delta_{ij}\sum_{k=1}^{3}\epsilon_{kk}, \tag{47}$$

which involves two Lamé constants λ_{L} and μ_{L} (the latter being the same as the shear modulus G). These are related to the elastic modulus E and Poisson's ratio ν by

$$\lambda_L = \nu E/(1+\nu)(1-2\nu), \tag{48}$$

$$\mu_L = G = E/2(1+\nu). \tag{49}$$

In this idealized and simplified model of acoustic propagation in solids, the conservation of mass and energy is not explicitly invoked. The density that appears on the left-hand side of Eq. (45) is regarded as time independent, and as what is appropriate for the undisturbed state of the solid. Values of the elastic properties of solids can be found in various reference books. Steel, for example, has an elastic modulus E of the order of 20×10^{10} N/m^2, a Poisson's ratio ν of the order of 0.27, and a density ρ of the order of 7.8×10^3 kg/m^3.

2. EQUATIONS OF LINEAR ACOUSTICS

2.1 The Linearization Process

Sound results from a time-varying perturbation of the dynamic and thermodynamic variables that describe the medium. The quantities appropriate to the ambient medium are customarily represented (Pierce, 1989) by the subscript 0, and the perturbations are represented by a prime on the corresponding symbol. Thus one expresses the total pressure as

$$p = p_0 + p', \tag{50}$$

with corresponding expressions for fluctuations in specific entropy, fluid velocity, and density. The linear equations that govern acoustical disturbances are then determined by the first-order terms in the expansion of the governing nonlinear equations in the primed variables. The zeroth-order terms cancel out completely because the ambient variables should themselves correspond to a valid state of motion of the medium. Thus, for example, the linearized version of the conservation-of-mass relation in Eq. (1) is

$$\frac{\partial\rho'}{\partial t} + \nabla\cdot(\mathbf{v}_0\rho' + \mathbf{v}'\rho_0) = 0. \tag{51}$$

The possible forms of the linear equations differ in complexity according to what is assumed about the medium's ambient state and according to what dissipative terms are included in the original governing equations. In the establishment of a rationale for using simplified models, it is helpful to think in terms of the orders of magnitude and characteristic scales of the coefficients in more nearly comprehensive models. If the spatial region of interest has bounding dimensions that are smaller than any scale length over which the ambient variables vary by a respectable fraction, then it may be appropriate to idealize the coefficients in the governing equations as if the ambient medium were spatially uniform or homogeneous. Similarly, if the characteristic wave periods or propagation times are much less than characteristic time scales for the ambient medium, then it may be appropriate to idealize the coefficients as being time independent. Examination of orders of magnitude of terms suggests that the ambient velocity may be neglected if it is much less than the sound speed c. The first-order perturbation to the gravitational force term can be ordinarily neglected (Bergmann, 1946) if the previously stated conditions are met and if the quantity g/c is sufficiently less than any characteristic frequency of the disturbance. Analogous inferences can be made about thermal conductivity and viscosity. However, the latter may be important (Cremer, 1948) in the interaction of acoustic waves in fluids with adjacent solid bodies.

A discussion of the restrictions on using linear equations and of neglecting second-order and higher terms in the primed variables is outside the scope of the present article, but it should be noted that one regards p' as small (Pierce, 1989) if it is substantially less than $\rho_0 c^2$, and $|\mathbf{v}'|$ as small if it is much less than c, where c is the sound speed defined via Eq. (23). It is not necessary that p' be much less than p_0, and it is certainly not necessary that $|\mathbf{v}'|$ be less than $|\mathbf{v}_0|$.

2.2 Linearized Equations for an Ideal Fluid

The customary equations for linear acoustics neglect dissipation processes and consequently can be derived from the equations for flow of a compressible ideal fluid, given in Eqs. (41)–(44). If one neglects gravity at the outset, and assumes that the ambient fluid velocity is zero, then the ambient pressure is constant. In such a case, Eq. (44) leads after linearization to

$$\frac{\partial p}{\partial t} + \rho c^2 \nabla\cdot\mathbf{v} = 0, \tag{52}$$

and the Euler equation, Eq. (41), leads to

$$\rho\frac{\partial \mathbf{v}}{\partial t}=-\nabla p. \tag{53}$$

Here a common notational convention is used to delete primes and subscripts. The density ρ is understood to be the ambient density ρ_0, while p and $\mathbf{v}$ are understood to be the acoustically induced perturbations to the pressure and fluid velocity.

These two coupled equations for p and $\mathbf{v}$ remain applicable when the ambient density and sound speed vary with position.

2.3 The Wave Equation

A single partial differential equation for the acoustic part of the pressure results when one takes the time derivative of Eq. (52) and then uses Eq. (53) to reexpress the time derivative of the fluid velocity in terms of pressure. The resulting equation is

$$\nabla\cdot\left(\frac{1}{\rho}\nabla p\right)-\frac{1}{\rho c^2}\frac{\partial^2 p}{\partial t^2}=0. \tag{54}$$

If the ambient density is independent of position, then this reduces to

$$\nabla^2 p-\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}=0, \tag{55}$$

which is what is ordinarily termed the wave equation of linear acoustics. Often this is written as

$$\Box^2 p=0 \tag{56}$$

in terms of the d'Alembertian operator defined by

$$\Box^2=\nabla^2-c^{-2}\partial^2/\partial t^2. \tag{57}$$

2.4 Wave Equations for Isotropic Elastic Solids

When the ambient density and the Lamé constants are independent of position, the Cauchy equation and the stress-strain relations described by Eqs. (45)–(47) can be combined into the single vector equation

$$\frac{\partial^2 \boldsymbol{\xi}}{\partial t^2}=(c_1^2-c_2^2)\nabla(\nabla\cdot\boldsymbol{\xi})+c_2^2\nabla^2\boldsymbol{\xi}. \tag{58}$$

Here the quantities c_1 and c_2 are defined by

$$c_1^2=(\lambda_\mathrm{L}+2\mu_\mathrm{L})/\rho, \tag{59}$$

$$c_2^2=\mu_\mathrm{L}/\rho. \tag{60}$$

Because any vector field may be decomposed into a sum of two fields, one with zero curl and the other with zero divergence, one can set the vector $\boldsymbol{\xi}$ equal to

$$\boldsymbol{\xi}=\nabla\Phi+\nabla\times\boldsymbol{\Psi} \tag{61}$$

in terms of a scalar and a vector potential. This expression will satisfy Eq. (58) identically, provided the two potentials separately satisfy

$$\nabla^2\Phi-\frac{1}{c_1^2}\frac{\partial^2\Phi}{\partial t^2}=0, \tag{62}$$

$$\nabla^2\boldsymbol{\Psi}-\frac{1}{c_2^2}\frac{\partial^2\boldsymbol{\Psi}}{\partial t^2}=0. \tag{63}$$

Both of these equations are of the form of the simple wave equation of Eq. (55). The first corresponds to longitudinal wave propagation, and the second corresponds to shear wave propagation. The quantities c_1 and c_2 are referred to as the dilatational and shear wave speeds, respectively.

If the displacement field is irrotational, so that the curl of the displacement vector is zero (as is so for sound in fluids), then each component of the displacement field satisfies Eq. (62). If the displacement field is solenoidal, so that its divergence is zero, then each component of the displacement satisfies Eq. (63).

2.5 Linearized Equations for a Viscous Fluid

For the restricted but relatively widely applicable case when the Navier–Stokes–Fourier equations are presumed to hold and the characteristic scales of the ambient medium and the disturbance are such that the ambient flow is negligible, and the coefficients are idealizable as constants, the linear equations have the form appropriate for a disturbance in a homogeneous, time-independent, nonmoving medium, these equations being

$$\frac{\partial \rho'}{\partial t}+\rho_0\nabla\cdot\mathbf{v}'=0, \tag{64}$$

$$\rho_0\frac{\partial \mathbf{v}'}{\partial t}=-\nabla p'+\left(\frac{1}{3}\mu+\mu_B\right)\nabla(\nabla\cdot\mathbf{v}')+\mu\nabla^2\mathbf{v}', \tag{65}$$

$$\rho_0 T_0 \frac{\partial s'}{\partial t} = \kappa \nabla^2 T', \tag{66}$$

$$\rho' = \frac{1}{c^2} p' - \left(\frac{\rho \beta T}{c_p}\right)_0 s', \tag{67}$$

$$T' = \left(\frac{T\beta}{\rho c_p}\right)_0 p' + \left(\frac{T}{c_p}\right)_0 s'. \tag{68}$$

The primes on the perturbation field variables are needed here to distinguish them from the corresponding ambient quantities.

2.6 Acoustic, Entropy, and Vorticity Modes

In general, any solution of the linearized equations for a fluid with viscosity and thermal conductivity can be regarded (Kirchhoff, 1868; Cremer, 1948; Kovasznay, 1953; Pierce, 1989) as a sum of three basic types of solutions. The common terminology for these is (i) the acoustic mode, (ii) the entropy mode, and (iii) the vorticity mode. Thus, with appropriate subscripts denoting the various fundamental mode types, one writes the fluid velocity perturbation in the form

$$\mathbf{v}' = \mathbf{v}_{\mathrm{ac}} + \mathbf{v}_{\mathrm{ent}} + \mathbf{v}_{\mathrm{vor}} \tag{69}$$

with similar decompositions for the other field variables. The decomposition is intended to be such that, for waves of constant frequency, each field variable's contribution from any given mode satisfies a partial differential equation that is second order [rather than of some higher order as might be derived from Eqs. (64)–(68)] in the spatial derivatives. That such a decomposition is possible follows from a theorem (Wu, 1956; Courant, 1937) that any solution of a partial differential equation of, for example, the form

$$(\nabla^2 + \lambda_1)(\nabla^2 + \lambda_2)(\nabla^2 + \lambda_3)\psi = 0 \tag{70}$$

can be written as a sum,

$$\psi = \psi_1 + \psi_2 + \psi_3, \tag{71}$$

where the individual terms each satisfy second-order differential equations of the form

$$(\nabla^2 + \lambda_i)\psi_i = 0. \tag{72}$$

This decomposition is possible providing no two of the λ_i are equal. Thus one has replaced the seemingly formidable problem of solving a sixth-order partial differential equation by that of solving three second-order partial differential equations

Linear acoustics is primarily concerned with solutions of the equations that govern the acoustic mode, while conduction heat transfer is primarily concerned with solutions of the equations that govern the entropy mode. However, the simultaneous consideration of all three solutions is often necessary when one seeks to satisfy boundary conditions at solid surfaces. The decomposition becomes relatively simple when the characteristic angular frequency is sufficiently small that

$$\omega \ll \rho c^2 / (\tfrac{4}{3}\mu + \mu_B), \tag{73}$$

$$\omega \ll \rho c^2 c_p / \kappa. \tag{74}$$

These conditions are invariably well met in all applications of interest. For air the right-hand sides of the two inequalities correspond to frequencies of the order of 10^9 Hz, while for water they correspond to frequencies of the order of 10^{12} Hz. If the inequalities are not satisfied, it is also possible that the Navier–Stokes–Fourier model is not appropriate, since the latter presumes that the variations are sufficiently slow that the medium can be regarded as being in quasithermodynamic equilibrium.

The acoustic mode and the entropy mode are both characterized by the vorticity (curl of the fluid velocity) being identically zero. (This characterization is consistent with the theorem that any vector field can be decomposed into a sum of a field with zero divergence and a field with zero curl.) In the limit of zero viscosity and thermal conductivity [which is a limit consistent with Eqs. (73) and (74)], the acoustic mode is characterized by zero entropy perturbations, while the entropy mode is a static entropy perturbation, with zero pressure perturbation. These identifications allow the approximate equations governing the various modes to be developed in the limit for which the inequalities are valid. In carrying out the development, it is convenient to use an expansion parameter defined by

$$\epsilon = \omega\mu / \rho_0 c^2, \tag{75}$$

where ω is a characteristic angular frequency, or equivalently, the reciprocal of a characteristic time for the perturbation. In the ordering, ϵ is regarded as small, μ_B/μ and $\kappa/\mu c_p$

are regarded as of order unity, and all time derivative operators are regarded as of order ω.

Acoustic Mode. To zeroth order in ϵ, the acoustic mode is what results when the thermal conductivity and the viscosities are identically zero and the entropy perturbation is also zero. In such a case, the relations of Eqs. (52) and (53) are valid and one obtains the wave equation of Eq. (55). All of the field variables in the acoustic mode satisfy this wave equation to zeroth order in ϵ. This is so because the governing equations have constant coefficients.

When one wishes to take into account the corrections that result to first order in ϵ, the entropy perturbation is no longer regarded as identically zero, but its magnitude can be deduced by replacing the T' that appears in Eq. (66) by only the first term in the sum of Eq. (68), the substitution being represented by

$$T_{\mathrm{ac}}=(T_0\beta/\rho_0 c_p)p_{\mathrm{ac}}. \tag{76}$$

To the same order of approximation one may use the zeroth-order wave equation, Eq. (55), to express the right-hand side of Eq. (66) in terms of a second-order time derivative of the acoustic-mode pressure. The resulting equation can then be integrated once in time with the constant of integration set equal to zero, because the acoustic mode's entropy perturbation must vanish when the corresponding pressure perturbation is identically zero. Thus one obtains the acoustic-mode entropy as

$$s_{\mathrm{ac}}\approx\frac{\kappa\beta}{\rho_0^2 c^2 c_p}\frac{\partial p_{\mathrm{ac}}}{\partial t} \tag{77}$$

correct to first order in ϵ.

This expression for the entropy perturbation is substituted into Eq. (67) and the resulting expression for ρ_{ac} is substituted into the conservation-of-mass relation, represented by Eq. (64), so that one obtains

$$\frac{\partial p_{\mathrm{ac}}}{\partial t}-\frac{(\gamma-1)\kappa}{\rho_0 c^2 c_p}\frac{\partial^2 p_{\mathrm{ac}}}{\partial t^2}+\rho_0 c^2\nabla\cdot\mathbf{v}_{\mathrm{ac}}=0. \tag{78}$$

The vorticity in the acoustic mode is zero,

$$\nabla\times\mathbf{v}_{\mathrm{ac}}=0, \tag{79}$$

so the momentum balance equation, Eq. (65), simplifies to one where the right-hand side is a gradient. Then with an appropriate substitution for the divergence of the fluid velocity from Eq. (78), one finds that the momentum-balance equation to first order in ϵ takes the form

$$\rho_0\frac{\partial\mathbf{v}_{\mathrm{ac}}}{\partial t}=-\nabla\left[p_{\mathrm{ac}}+\frac{1}{\rho_0 c^2}\left(\frac{4}{3}\mu+\mu_B\right)\frac{\partial p_{\mathrm{ac}}}{\partial t}\right]. \tag{80}$$

The latter and Eq. (78), in a manner similar to that in which the wave equation of Eq. (54) is derived, yield the dissipative wave equation represented by

$$\nabla^2 p_{\mathrm{ac}}-\frac{1}{c^2}\frac{\partial^2 p_{\mathrm{ac}}}{\partial t^2}+\frac{2\delta_{\mathrm{cl}}}{c^4}\frac{\partial^3 p_{\mathrm{ac}}}{\partial t^3}=0. \tag{81}$$

The quantity δ_{cl}, defined by

$$\delta_{\mathrm{cl}}=(1/2\rho_0)\left[\tfrac{4}{3}\mu+\mu_B+(\gamma-1)(\kappa/c_p)\right], \tag{82}$$

is a constant that characterizes the "classical" dissipation in the acoustic mode.

Entropy Mode. For the entropy mode, the pressure perturbation vanishes to zeroth order in ϵ, while the entropy perturbation does not. Thus for this mode, Eqs. (67) and (68) approximate in zeroth order to

$$\rho_{\mathrm{ent}}=-(\rho\beta T/c_p)_0 s_{\mathrm{ent}}, \tag{83}$$

$$T_{\mathrm{ent}}=(T/c_p)_0 s_{\mathrm{ent}}. \tag{84}$$

The latter, when substituted into the energy equation, Eq. (66), yields the diffusion equation of conduction heat transfer,

$$\frac{\partial s_{\mathrm{ent}}}{\partial t}=\frac{\kappa}{\rho_0 c_p}\nabla^2 s_{\mathrm{ent}}. \tag{85}$$

All of the entropy-mode field variables, including T_{ent}, satisfy this diffusion equation to lowest order in ϵ.

The mass-conservation equation, with an analogous substitution from Eq. (83) and another substitution from Eq. (85), integrates to

$$\mathbf{v}_{\mathrm{ent}}=(\beta T_0\kappa/\rho_0 c_p^2)\nabla s_{\mathrm{ent}}, \tag{86}$$

so there is a weak fluid flow caused by entropy or temperature gradients in the entropy mode. The substitution of this expression for the fluid velocity into the linearized Navier–Stokes equation, Eq. (65), subsequently yields a lowest-order expression for the entropy-mode contribution to the pressure perturbation, this being

$$p_{\mathrm{ent}}=\frac{\beta T_0}{c_p}\left(\frac{4}{3}\mu+\mu_B-\frac{\kappa}{c_p}\right)\frac{\partial s_{\mathrm{ent}}}{\partial t}. \tag{87}$$

Vorticity Mode. For the vorticity mode, the divergence of the fluid velocity is identically zero,

$$\nabla\cdot\mathbf{v}_{\text{vor}}=0. \tag{88}$$

The mass-conservation equation, Eq. (64), and the thermodynamic relations, Eqs. (67) and (68), consequently require that all of the thermodynamic field quantities with subscript "vor" are zero. The Navier–Stokes equation, Eq. (65), accordingly requires that each component of the fluid velocity associated with this mode satisfy the vorticity diffusion equation

$$\rho_0\frac{\partial\mathbf{v}_{\text{vor}}}{\partial t}=\mu\nabla^2\mathbf{v}_{\text{vor}}. \tag{89}$$

2.7 Boundary Conditions at Interfaces

For the model of a fluid without viscosity or thermal conduction, such as is governed by Eqs. (52) through (56), the appropriate boundary conditions at an interface are that the normal component of the fluid velocity be continuous and that the pressure be continuous. At a rigid nonmoving surface, the normal component must consequently vanish, but no restrictions are placed on the tangential component of the velocity. The model also places no requirements on the value of the temperature at a solid surface.

When viscosity is taken into account, no matter how small it may be, the model demands additional boundary conditions. An additional condition invariably imposed is that the tangential components of the velocity also be continuous, the rationale (Lamb, 1932) being that a fluid should not slip any more freely with respect to an interface than it does with itself; this lack of slip is invariably observed (Shapiro, 1961) when the motion is examined sufficiently close to an interface. Similarly, when thermal conduction is taken into account, the rational interpretation of the model requires that the temperature be continuous at an interface. Newton's third law requires that the shear stresses exerted across an interface be continuous and the conservation of energy requires that the normal component of the heat flux be continuous.

The apparent paradox, that tiny but nonzero values of the viscosity and thermal conductivities should drastically alter the basic nature of the boundary conditions at an interface, is resolved by acoustic boundary-layer theory. Under normal circumstances, disturbances associated with sound are almost entirely in the acoustic mode except in oscillating acoustic boundary layers near interfaces. In these boundary layers, a portion of the disturbance is in the entropy and vorticity modes. These modes make minor contributions to the total perturbation pressure and normal velocity component, but, respectively, have temperature and tangential velocity components that are comparable to those of the acoustic mode. The boundary and continuity conditions can be satisfied by the sum of the three mode fields, but not by the acoustic-mode field alone. However, the entropy- and vorticity-mode fields typically die out rapidly with distance from an interface. The characteristic decay lengths when the disturbance has an angular frequency ω are (Pierce, 1989)

$$l_{\text{vor}}=(2\mu/\omega\rho)^{1/2}, \tag{90}$$

$$l_{\text{ent}}=(2\kappa/\omega\rho c_p)^{1/2}. \tag{91}$$

Given that the parameter ϵ in Eq. (75) is much less than unity, these lengths are much shorter than a characteristic wavelength. If they are, in addition, much shorter than the dimensions of the space to which the acoustic field is confined, then the acoustic boundary layer has a minor effect on the acoustic field outside the boundary layer. However, because dissipation of energy can occur within such boundary layers, the existence of a boundary layer can have a long-term accumulative effect. This can be manifested by the larger attenuation of sound waves propagating in pipes (Lambert, 1951) and in the finite *Q*'s (quality factors) of resonators.

3. WAVES OF CONSTANT FREQUENCY

3.1 General

One of the principal terms used in describing sound is that of frequency. Although certainly not all acoustic disturbances are purely sinusoidal oscillations of a quantity (such as pressure or displacement) that oscillates with constant frequency about a mean value, it is usually possible to characterize (at least to an order of magnitude) any such disturbance by one or a limited number of representative frequencies, these being the reciprocals of the characteristic time scales. Such a description is highly useful, because it gives one some insight into what detailed physical effects are relevant and what instrumentation is applicable. The simplest sort of acoustic signal would be one where a quantity such as the fluctuating part p of the pressure is oscillating with time t as

$$p=A\sin(\omega t+\phi). \tag{92}$$

Here the amplitude A and phase constant ϕ are independent of t (but possibly dependent on position). The quantity ω is called the angular frequency and has the units of radians divided by time (for example, rad/s or simply s^{-1} when the unit of time is the second). The number f of repetitions per unit time is what one normally refers to as the frequency (without a modifier), such that $\omega=2\pi f$. The value of f in hertz (abbreviated Hz) is the frequency in cycles (repetitions) per second.

The human ear responds (Fletcher, 1940) almost exclusively to frequencies between roughly 20 Hz and 20 kHz. Consequently, sounds composed of frequencies below 20 Hz are said to be infrasonic; those composed of frequencies above 20 kHz are said to be ultrasonic. The scope of acoustics, given its general definition, is not limited to only audible frequencies.

3.2 Spectral Density

The term frequency spectrum is often used in relation to sound. Often the use is somewhat loose, but there are circumstances for which the terminology can be made precise. If the fluctuating physical quantity p associated with the acoustic disturbance is a sum of sinusoidal disturbances, the nth having rms amplitude p_n and frequency f_n, no two frequencies being the same, then the set of mean squared amplitudes can be taken as a description of the spectrum of the signal. Also, if the sound is made up of many different frequencies, then one can use the sum of the p_n^2 that correspond to those frequencies within a given specified frequency band as a measure of the strength of the signal within that frequency band. This sum divided by the width of the frequency band often approaches a quasilimit as the bandwidth becomes small but with the number of included terms still being moderately large, with this quasilimit being a definite smooth function of the center frequency of the band. This quasilimit is called the spectral density p_f^2 of the signal. Although an idealized quantity, it can often be repetitively measured to relatively high accuracy; instrumentation for "measurement" of spectral densities or of integrals of spectral densities over specified frequency bands (such as octaves) is widely available commercially.

The utility of the spectral density concept rests on the principle of superposition and on a version of Parseval's theorem which states that, when the signal is a sum of discrete frequency components, then if averages are taken over a sufficiently long time interval,

$$(p^2)_{\text{av}}=\sum p_n^2. \tag{93}$$

Consequently, in the quasilimit corresponding to the spectral density description, one has the mean squared pressure expressed as an integral over spectral density,

$$(p^2)_{\text{av}}=\int_0^\infty p_f^2(f)\,df. \tag{94}$$

The spectral density at a given frequency times a narrow bandwidth centered at that frequency is interpreted as the contribution to the mean squared acoustic pressure from that frequency band.

If the signal is a linear response to a source (or excitation) characterized by some time variation $s(t)$ with spectral density $s_f^2(f)$, then the spectral density of the response is

$$p_f^2(f)=|H_{ps}(2\pi f)|^2 s_f^2(f), \tag{95}$$

where the proportionality factor $|H_{ps}(2\pi f)|^2$ is the square of the absolute magnitude of a transfer function which is independent of the excitation but which does depend on frequency. For any given frequency, this transfer function can be determined from the response to a known sinusoidal excitation with the same frequency. Thus the analysis for constant-frequency excitation is fundamental, even when the situation of interest may involve broadband excitation.

3.3 Complex-Number Representation

Insofar as the governing equations are linear with coefficients independent of time, disturbances that vary sinusoidally with time propagate without change of frequency. Such sinusoidally varying disturbances of constant frequency have the same repetition period (reciprocal of frequency) at every point, but the phase will in general vary from point to point.

When one considers a disturbance of fixed frequency or one frequency component of a multifrequency disturbance, it is convenient to use a complex-number representation, such that each field amplitude is written as (Rayleigh, 1878; Bouwkamp, 1946)

$$p=\mathrm{Re}(\hat{p}\,e^{-i\omega t}). \tag{96}$$

Here $\hat{p}$ is called the complex amplitude of the acoustic pressure and in general varies with position.

If one inserts expressions such as that of Eq. (96) into a homogeneous linear ordinary or partial differential equation with real time-independent coefficients, the result can always be written in the form

$$\mathrm{Re}(\Phi e^{-i\omega t})=0, \tag{97}$$

where the quantity Φ is an expression depending on the complex amplitudes and their spatial derivatives, but not depending on time. The requirement that Eq. (97) should hold for all values of time consequently can be satisfied if and only if

$$\Phi=0. \tag{98}$$

Moreover, the form of the expression Φ can be readily obtained from the original equation with a simple prescription: replace all field variables by their complex amplitudes and replace all time derivatives using the substitution

$$\frac{\partial}{\partial t}\rightarrow -i\omega. \tag{99}$$

Thus, for example, the linear acoustic equations given by Eqs. (52) and (53) reduce to

$$-i\omega\hat{p}+\rho c^2\nabla\cdot\hat{\mathbf{v}}=0, \tag{100}$$

$$-i\omega\rho\hat{\mathbf{v}}=-\nabla\hat{p}. \tag{101}$$

The wave equation in Eq. (55) reduces to

$$\nabla^2\hat{p}+k^2\hat{p}=0, \tag{102}$$

which is the Helmholtz equation (Helmholtz, 1860; Birkhoff, 1973) for the complex pressure amplitude.

4. PLANE WAVES

4.1 Plane Waves in Fluids

A solution of the wave equation that plays a central role in many acoustical concepts is that of a plane traveling wave, which is such that all acoustic field quantities vary with time and with one Cartesian coordinate, taken here as x, but are independent of y and z. The Laplacian reduces thus to a second partial derivative with respect to x and the d'Alembertian can be expressed as the product of two first-order operators, so that the wave equation takes the form

$$\left(\frac{\partial}{\partial x}-\frac{1}{c}\frac{\partial}{\partial t}\right)\left(\frac{\partial}{\partial x}+\frac{1}{c}\frac{\partial}{\partial t}\right)p=0. \tag{103}$$

The solution of this equation is a sum of two expressions, each of which is such that operation by one of the two factors in Eq. (103) yields zero. Such a solution is represented by

$$p(x,t)=f(x-ct)+g(x+ct), \tag{104}$$

where f and g are two arbitrary functions. The argument combination $x-ct$ of the first term is such that the function f with that argument represents a plane wave traveling forward in the $+x$ direction at a velocity c. The second term in Eq. (104) similarly represents a plane wave traveling backwards, in the $-x$ direction, also at a velocity c. For a traveling plane wave, not only the shape but also the amplitude is conserved during propagation. A typical situation for which wave propagation can be adequately described with traveling plane waves is that of low-frequency sound propagation in a duct. Diverging or converging (focused) waves can often be approximately regarded as planar within regions of restricted extent.

The fluid velocity that corresponds to the plane-wave solution above has y and z components that are identically zero, but the x component, in accordance with Eqs. (52) and (53), is

$$v_x=\frac{1}{\rho c}\,f(x-ct)-\frac{1}{\rho c}\,g(x+ct). \tag{105}$$

The general rule that emerges from this is that, for a plane wave propagating in the direction corresponding to unit vector $\mathbf{n}$, the acoustic part of the pressure is

$$p=f(\mathbf{n}\cdot\mathbf{x}-ct) \tag{106}$$

for some generic function f of the indicated argument, while the acoustically induced fluid velocity is

$$\mathbf{v}=(\mathbf{n}/\rho c)p. \tag{107}$$

Because the fluid velocity is in the same direction as that of the wave propagation, such waves are said to be longitudinal. (Electromagnetic plane waves in free space, on the other hand, are transverse.)

For a plane wave traveling along the $+x$ axis at a velocity c, the form of Eq. (106) implies that one can represent a constant-frequency acoustic pressure disturbance by the expression

$$p=|P|\cos[k(x-ct)+\phi_0]=\mathrm{Re}(Pe^{ik(x-ct)}) \tag{108}$$

with the angular frequency identified as

$$\omega=ck. \tag{109}$$

Alternatively, if the complex-amplitude representation of Eq. (96) is used, one writes the complex amplitude of the acoustic pressure as

$$\hat{p}=Pe^{ikx}, \tag{110}$$

where P is a complex number related to the constants that appear in Eq. (108) as

$$P=|P|e^{i\phi_0}. \tag{111}$$

Here $|P|$ is the amplitude of the disturbance, ϕ_0 is a phase constant, and k is a constant termed the wave number. The wavelength λ is the increment in propagation distance x required to change the argument of the cosine by 2π radians, so

$$k=2\pi/\lambda. \tag{112}$$

Also, the increment in t required to change the argument by 2π is the period T, which is the reciprocal of the frequency f, so one has

$$\lambda=c/f. \tag{113}$$

4.2 Plane Waves in Solids

Plane acoustic waves in isotropic elastic solids (Landau and Lifshitz, 1970) have properties similar to those of waves in fluids. Dilatational (or longitudinal) plane waves are such that the curl of the displacement field vanishes, so the displacement vector must be parallel to the direction of propagation. A comparison of Eq. (62) with the wave equation of Eq. (55) indicates that such a wave must propagate with the speed c_1 determined by Eq. (59). Thus, a wave propagating in the $+x$ direction has no y and z components of displacement, and has an x component described by

$$\xi_x=F(x-c_1t), \tag{114}$$

where F is an arbitrary function. The stress components can be deduced from Eqs. (46), (47), (59), and (60). These equations as well as symmetry considerations require, for a dilatational wave propagating in the x direction, that the off-diagonal elements of the stress tensor vanish. The diagonal elements are given by

$$\sigma_{xx}=\rho c_1^2F'(x-c_1t), \tag{115}$$

$$\sigma_{yy}=\sigma_{zz}=\rho(c_1^2-2c_2^2)F'(x-c_1t). \tag{116}$$

(Here the primes denote derivatives with respect to the total argument.)

The divergence of the displacement field in a shear wave is zero, so a plane shear wave must cause a displacement perpendicular to the direction of propagation. Shear waves are therefore transverse waves. Equation (63), when considered in a manner similar to that described above for the wave equation for waves in fluids, leads to the conclusion that plane shear waves must propagate with speed c_2. A plane shear wave polarized in the y direction and propagating in the x direction will have only a y component of displacement, given by

$$\xi_y=F(x-c_2t). \tag{117}$$

The only nonzero stress components are the shear stresses

$$\sigma_{yx}=\rho c_2^2F'(x-c_2t)=\sigma_{xy}. \tag{118}$$

5. ATTENUATION OF SOUND

5.1 General

Plane waves of constant frequency propagating through bulk materials have amplitudes that typically decrease exponentially with increasing propagation distance, such that the magnitude of the complex pressure amplitude varies as

$$|\hat{p}(x)| = |\hat{p}(0)| e^{-\alpha x}. \tag{119}$$

The quantity α is the plane-wave attenuation coefficient and has the units of nepers per meter (Np/m); it is an intrinsic frequency-dependent property of the material. This exponential decrease of amplitude is called attenuation or absorption of sound and is associated with the transfer of acoustic energy to the internal energy of the material.

5.2 Classical Absorption

The attenuation of sound due to the classical processes of viscous energy absorption and thermal conduction is derivable (Kirchhoff 1868) from the dissipative wave equation, Eq. (81), given previously for the acoustic mode. Dissipative processes enter into this equation through a parameter δ_{cl}, which is defined by Eq. (82). To determine the attenuation of waves governed by such a dissipative wave equation, one sets the perturbation pressure equal to

$$p_{\mathrm{ac}} = Pe^{-i\omega t}e^{ikx}, \tag{120}$$

where P is independent of position and k is a complex number. Such a substitution, with reasoning such as that which leads from Eq. (97) to Eq. (98), yields an algebraic equation that can be nontrivially (amplitude not identically zero) satisfied only if k satisfies the relation

$$k^2 = (1/c^2)\,\omega^2 + i(2\delta_{\mathrm{cl}}\omega^3/c^3). \tag{121}$$

The root that corresponds to waves propagating in the plus x direction is that which evolves to Eq. (109) in the limit of no absorption, and for which the real part of k is positive. Thus to first order in δ_{cl}, one has the complex wave number

$$k = \omega/c + i(\delta_{\mathrm{cl}}\omega^2/c^3). \tag{122}$$

The attenuation coefficient is the imaginary part of this, in accordance with Eq. (119), so

$$\alpha_{\mathrm{cl}} = \delta_{\mathrm{cl}}\omega^2/c^3. \tag{123}$$

This is termed the classical attenuation coefficient for acoustic waves in fluids and is designated by the subscript "cl." The distinguishing characteristic of this classical attenuation coefficient is its quadratic increase with increasing frequency.

5.3 Relaxation Processes

For many substances, including air and sea water, the variation of absorption coefficient with frequency is not quadratic (Beyer and Letcher, 1969; Markham, Beyer, and Lindsay, 1951; Lindsay, 1974), and the classical model is insufficient to predict the magnitude of the absorption coefficient. The substitution of a different value of the bulk viscosity is insufficient to remove the discrepancy, because such would still yield the quadratic frequency dependence. The successful theory to account for such discrepancies is in terms of relaxation processes. The physical nature of the relaxation processes varies from fluid to fluid, but general theory in terms of irreversible thermodynamics (Meixner, 1943, 1952a,b) yields appropriate equations.

The equation of state for the instantaneous (rather than equilibrium) entropy is written as

$$s = s(u, \rho^{-1}, \xi), \tag{124}$$

where ξ represents one or more internal variables. The differential relation of Eq. (10) is replaced by

$$T\,ds = du + p\,d(\rho^{-1}) + \sum_{\nu} A_\nu d\xi_\nu, \tag{125}$$

where the affinities A_ν are defined by this equation. These vanish when the fluid is in equilibrium with a given specified internal energy and density. The pressure p is the same as entered into the expression, Eq. (12), for the average normal stress, and the T is the same as entered into the Fourier law, Eq. (13), of heat conduction. The mass-conservation law, Eq. (1), and the Navier–Stokes equation, Eq. (21), remain unchanged, but the energy equation, expressed in Eq. (22) in terms of entropy, is replaced by the entropy balance equation

$$\rho\frac{Ds}{Dt} + \nabla\cdot\frac{\mathbf{q}}{T} = \sigma_s. \tag{126}$$

Here the quantity σ_s, which indicates the rate of irreversible entropy production per unit volume, is given by

$$T\sigma_s=\mu_B(\nabla\cdot\mathbf{v})^2+\frac{1}{2}\mu\sum_{ij}\phi_{ij}^2+\frac{\kappa}{T}(\nabla T)^2$$
$$+\rho\sum_\nu A_\nu\frac{D\xi_\nu}{Dt}. \tag{127}$$

One needs, in addition, relations that specify how the internal variables ξ_ν relax to their equilibrium values. The simplest assumption, and one which is substantiated for air and sea water, is that these relax independently according to the rule (Herzfeld and Rice, 1928)

$$\frac{D\xi_\nu}{Dt}=-\frac{1}{\tau_\nu}(\xi_\nu-\xi_{\nu,\mathrm{eq}}). \tag{128}$$

The relaxation times τ_ν that appear here are positive and independent of the internal variables.

When the linearization process is applied to the nonlinear equations for the model just described of a fluid with internal relaxation, one obtains the set of equations

$$\frac{\partial\rho'}{\partial t}+\rho_0\nabla\cdot\mathbf{v}'=0, \tag{129}$$

$$\rho_0\frac{\partial\mathbf{v}'}{\partial t}=-\nabla p'+(\tfrac{1}{3}\mu+\mu_B)\nabla(\nabla\cdot\mathbf{v}')+\mu\nabla^2\mathbf{v}', \tag{130}$$

$$\rho_0T_0\frac{\partial s'}{\partial t}=\kappa\nabla^2T', \tag{131}$$

$$\rho'=\frac{1}{c^2}p'-\left(\frac{\rho\beta T}{c_p}\right)_0s'+\sum_\nu(\xi'_\nu-\xi'_{\nu,\mathrm{eq}})a_\nu, \tag{132}$$

$$T'=\left(\frac{T\beta}{\rho c_p}\right)_0p'+\left(\frac{T}{c_p}\right)_0s'+\sum_\nu(\xi'_\nu-\xi'_{\nu,\mathrm{eq}})b_\nu, \tag{133}$$

$$\frac{\partial\xi'_\nu}{\partial t}=-\frac{1}{\tau_\nu}(\xi'_\nu-\xi'_{\nu,\mathrm{eq}}), \tag{134}$$

$$\xi'_{\nu,\mathrm{eq}}=m_\nu s'+n_\nu p'. \tag{135}$$

Here a_ν, b_ν, m_ν, and n_ν are constants whose values depend on the ambient equilibrium state.

For absorption of sound, the interest is in the acoustic mode, and so approximations corresponding to those discussed in the context of Eqs. (73)–(81) can be made here also. The quantities a_ν and b_ν are treated as small, so that one obtains, to first order in ϵ and in these quantities, a wave equation of the form

$$\nabla^2p_{\mathrm{ac}}-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\left(p_{\mathrm{ac}}-\frac{2\delta_{\mathrm{cl}}}{c^2}\frac{\partial p_{\mathrm{ac}}}{\partial t}\right.$$
$$\left.-2\sum_\nu\frac{(\Delta c)_\nu}{c}\tau_\nu\frac{\partial p_\nu}{\partial t}\right)=0. \tag{136}$$

The auxiliary internal variables that enter here satisfy the supplemental relaxation equations

$$\frac{\partial p_\nu}{\partial t}=-\frac{1}{\tau_\nu}(p_\nu-p_{\mathrm{ac}}). \tag{137}$$

Here the notation is such that the p_ν are given in terms of previously introduced quantities by

$$p_\nu=\xi'_\nu/n_\nu. \tag{138}$$

The sound-speed increments that appear in Eq. (136) represent the combinations

$$(\Delta c)_\nu=\tfrac{1}{2}a_\nu n_\nu c^2. \tag{139}$$

Equations (136) and (137) are independent of the explicit physical interpretation of the relaxation processes. Insofar as acoustic propagation is concerned, any relaxation process is characterized by two parameters, the relaxation time τ_ν and the sound-speed increment $(\Delta c)_\nu$. The various parameters that enter into the irreversible thermodynamics formulation of Eqs. (132)–(135) affect the propagation of sound only as they enter into the values of the relaxation time and the sound-speed increments. The replacement of internal variables by quantities p_ν with the units of pressure implies no assumption as to the precise nature of the relaxation process.

The attenuation of plane waves governed by Eqs. (136) and (137) is determined when one inserts substitutions of the form of Eq. (120) for p_{ac} and the p_ν. The relaxation equations yield the relations

$$\hat{p}_\nu=\frac{1}{1-i\omega\tau_\nu}\hat{p}_{\mathrm{ac}}. \tag{140}$$

When inserted into the complex amplitude version of Eq. (136), these yield the dispersion relation

$$k^2=\frac{\omega^2}{c^2}\left(1+i\frac{2\omega\delta_{\mathrm{cl}}}{c^2}+\sum_\nu\frac{2(\Delta c)_\nu}{c}\frac{i\omega\tau_\nu}{1-i\omega\tau_\nu}\right). \tag{141}$$

To first order in the small parameters, δ_{cl} and $(\Delta c)_\nu$, this yields the complex wave number

$$k=\frac{\omega}{c}\left(1+i\frac{\omega\delta_{\mathrm{cl}}}{c^2}+\sum_{\nu}\frac{(\Delta c)_{\nu}}{c}\frac{i\omega\tau_{\nu}}{1-i\omega\tau_{\nu}}\right), \tag{142}$$

which corresponds to waves propagating in the $+x$ direction. The attenuation coefficient, determined by the imaginary part of Eq. (142), can be written

$$\alpha=\alpha_{\mathrm{cl}}+\sum_{\nu}\alpha_{\nu}. \tag{143}$$

The first term is the classical attenuation determined by Eq. (123). The remaining terms correspond to the incremental attenuations produced by the separate relaxation processes, these being

$$\alpha_{\nu}=\frac{(\Delta c)_{\nu}}{c^2}\frac{\omega^2\tau_{\nu}}{1+(\omega\tau_{\nu})^2}. \tag{144}$$

Any such term increases quadratically with frequency at low frequencies, as would be the case for classical absorption (with an increased bulk viscosity), but it approaches a constant value at high frequencies.

The labeling of the quantities $(\Delta c)_{\nu}$ as sound-speed increments follows from an examination of the real part of the complex wave number, given by

$$k_R=\frac{\omega}{c}\left(1-\sum_{\nu}\frac{(\Delta c)_{\nu}}{c}\frac{(\omega\tau_{\nu})^2}{1+(\omega\tau_{\nu})^2}\right). \tag{145}$$

The ratio of ω to k_R is identified as the phase velocity v_{ph} of the wave. In the limit of low frequencies, the phase velocity predicted by Eq. (145) is the quantity c, which is the sound speed for a quasiequilibrium process. In the limit of high frequencies, however, to first order in the $(\Delta c)_{\nu}$, the phase velocity approaches the limit

$$v_{\mathrm{ph}}\rightarrow c+\sum_{\nu}(\Delta c)_{\nu}. \tag{146}$$

Consequently, each $(\Delta c)_{\nu}$ corresponds to the net increase in phase velocity of the sound wave that occurs when the frequency is increased from a value small compared to the relaxation frequency to one large compared to the relaxation frequency. Here the term relaxation frequency is used to denote the reciprocal of the product of 2π with the relaxation time.

5.4 Attenuation of Sound in Air

In air, the relaxation processes that affect sound attenuation are those associated with the internal vibrations of the diatomic molecules O_2 and N_2. The ratio of the number of molecules in the ground and first excited vibrational states is a function of temperature when the gas is in thermodynamic equilibrium, but during an acoustic disturbance the redistribution of this ratio to what is appropriate to the concurrent gas temperature is not itself instantaneous. The knowledge that the relaxation processes are vibrational relaxation processes and that only the ground and first excited states are appreciably involved allows the sound-speed increments to be determined from first principles. One need only determine the difference in sound speeds resulting from the two assumptions: (i) that the distribution of vibrational energy is frozen, and (ii) that the vibrational energy is always distributed as for a gas in total thermodynamic equilibrium. The resulting sound-speed increments are (Pierce, 1978)

$$\frac{(\Delta c)_{\nu}}{c}=\frac{(\gamma-1)^2}{2\gamma}\frac{n_{\nu}}{n}\left(\frac{T_{\nu}^*}{T}\right)^2 e^{-T_{\nu}^*/T}, \tag{147}$$

where n is the total number of molecules per unit volume and n_{ν} is the number of molecules of the type corresponding to the designation parameter ν. The quantity T is the absolute temperature, and T_{ν}^* is a characteristic temperature, equal to the energy jump ΔE between the two vibrational states divided by Boltzmann's constant. The value of T_1^* (corresponding to O_2 relaxation) is 2239 K, and the value of T_2^* (corresponding to N_2 relaxation) is 3352 K. For air the fraction n_1/n of molecules that are O_2 is 0.21, while the fraction n_2/n of molecules that are N_2 is 0.78. At a representative temperature of 20 °C, the calculated value of $(\Delta c)_1$ is 0.11 m/s, while that for $(\Delta c)_2$ is 0.023 m/s.

Because relaxation in a gas is caused by two-body collisions, the relaxation times at a given absolute temperature must vary inversely as the absolute pressure. The relatively small (and highly variable) number of water vapor molecules in the air has a significant effect on the relaxation times, because collisions of diatomic molecules with H_2O molecules are much more likely to cause a transition between one internal vibrational quantum state and another. Semiempirical expressions for the two relaxation times for air are given by (Evans, Bass, and Sutherland, 1972; Bass, Sutherland, and Zuckerwar, 1990)

$$\frac{p_{\mathrm{ref}}}{p}\frac{1}{2\pi\tau_1}=24+4.04\times 10^6 \times h\left(\frac{0.02+100h}{0.391+100h}\right), \tag{148}$$

$$\frac{p_{\mathrm{ref}}}{p}\frac{1}{2\pi\tau_2}=\left(\frac{T_{\mathrm{ref}}}{T}\right)^{1/2}(9+2.8\times 10^4 he^{-F}), \tag{149}$$

$$F=4.17[(T_{\mathrm{ref}}/T)^{1/3}-1]. \tag{150}$$

The subscript 1 corresponds to O_2 relaxation, and the subscript 2 corresponds to N_2 relaxation. The quantity h is the fraction of the air molecules that are H_2O molecules; the reference temperature is 293.16 K; and the reference pressure is 10^5 Pa. The value of h can be determined from the commonly reported relative humidity (RH) (expressed as a percentage) and the vapor pressure of water at the local temperature according to the defining relation

$$h=10^{-2}(\mathrm{RH})p_{\mathrm{VP}}(T)/p. \tag{151}$$

However, as is indicated in Eqs. (148) and (149), the physics of the relaxation process depends on the absolute humidity and has no direct involvement with the value of the vapor pressure of water. A table of the vapor pressure of water may be found in various references; some representative values in pascals are 872, 1228, 1705, 2338, 4243, and 7376 at temperatures of 5, 10, 15, 20, 30, and 40 °C.

5.5 Attenuation of Sound in Seawater

The relaxation processes contributing to the attenuation of sound are associated with dissolved boric acid $B(OH)_3$ (subscript 1) and magnesium sulfate $MgSO_4$ (subscript 2). From approximate formulas derived by Fisher and Simmons (1977) from a combination of experiment and theory one extracts the identifications

$$\delta_{\mathrm{cl}}/c^3=1.42\times 10^{-15}F(T_C)G(P_{\mathrm{atm}}) \tag{152}$$

with

$$F(T_C)=1-4.24\times 10^{-2}T_C+8.53\times 10^{-4}T_C^2 -6.23\times 10^{-6}T_C^3, \tag{153a}$$

$$G(P_{\mathrm{atm}})=1-3.84\times 10^{-4}P_{\mathrm{atm}} +7.57\times 10^{-8}P_{\mathrm{atm}}^2, \tag{153b}$$

$$(\Delta c)_1/c^2=1.64\times 10^{-9}(1+2.29\times 10^{-2}T_C -5.07\times 10^{-4}T_C^2)(S/35), \tag{154}$$

$$(\Delta c)_2/c^2=8.94\times 10^{-9}(1+0.0134T_C) \times(1-10.3\times 10^{-4}P_{\mathrm{atm}} +3.7\times 10^{-7}P_{\mathrm{atm}}^2)(S/35), \tag{155}$$

$$1/2\pi\tau_1=1320Te^{-1700/T}, \tag{156}$$

$$1/2\pi\tau_2=15.5\times 10^6 Te^{-3052/T}. \tag{157}$$

Here T_C is temperature in degrees C, T is absolute temperature, while P_{atm} is the absolute pressure in atmospheres; S is the salinity in parts per thousand (which is typically of the order of 35 for seawater). All of the quantities on the left-hand sides of these equations are in mks units.

6. ACOUSTIC INTENSITY AND POWER

6.1 General

A complete set of linear acoustic equations, regardless of the idealization incorporated into its formulation, usually yields a corollary (Kirchhoff, 1877; Pierce, 1989) of the form

$$\frac{\partial w}{\partial t}+\nabla\cdot\mathbf{I}=-\mathscr{D}, \tag{158}$$

where the terms are quadratic in the acoustic field amplitudes, the quantity w contains one term that is identifiable as a kinetic energy of fluid motion per unit volume, and the quantity $\mathscr{D}$ is either zero or positive. Such a relation is interpreted as a statement of energy conservation. The quantity w is the energy density, or energy per unit volume associated with the wave disturbance, while the vector quantity $\mathbf{I}$ is an intensity vector or energy flux vector. Its interpretation is such that its dot product with any unit vector represents the energy flowing per unit area and time across a surface whose normal is in that designated direction. The quantity $\mathscr{D}$ is interpreted as the energy that is dissipated per unit time and volume.

This interpretation of an equation such as Eq. (158) as a conservation law follows when one integrates both sides over an arbitrary fixed volume V within the fluid and reexpresses the volume integral of the divergence of $\mathbf{I}$ by a surface integral by means of Gauss's theorem. Doing this yields

$$\frac{\partial}{\partial t}\int_V w\,dV+\int_S \mathbf{I}\cdot\mathbf{n}\,dS$$
$$=-\int_V \mathscr{D}\,dV, \tag{159}$$

where $\mathbf{n}$ is the unit normal vector pointing out of the surface S enclosing V. This relation states that the net rate of increase of "acoustical energy" within the volume must equal the "acoustic power" flowing into the volume across its confining surface minus the energy that is being dissipated per unit time within the volume.

6.2 Acoustic Energy Density and Intensity

For the ideal case, when there is no ambient velocity and when viscosity and thermal conduction are neglected, the energy corollary results from Eqs. (52) and (53). The derivation begins with one's taking the dot product of the fluid velocity $\mathbf{v}$ with Eq. (53), and then using vector identities and Eq. (52) to reexpress the right-hand side as the sum of a divergence and a time derivative. The result yields the identification of the energy density as

$$w=\tfrac{1}{2}\rho v^2+\tfrac{1}{2}(1/\rho c^2)p^2, \tag{160}$$

and yields the identification of the acoustic intensity as

$$\mathbf{I}=p\mathbf{v}. \tag{161}$$

For this ideal case, there is no dissipative term on the right-hand side. The corollary of Eq. (158) remains valid even if the ambient density and sound speed should vary from point to point.

The first term in the expression for w is recognized as the acoustic kinetic energy per unit volume, and the second term is identified as the potential energy per unit volume due to compression of the fluid.

For a plane progressive wave, it follows from Eqs. (106) and (107) that the kinetic and potential energies are the same (Rayleigh, 1878) and that the energy density is given by

$$w=(1/\rho c^2)p^2. \tag{162}$$

The intensity becomes

$$\mathbf{I}=\mathbf{n}(p^2/\rho c). \tag{163}$$

For such a case, the intensity and the energy density are related by

$$\mathbf{I}=c\mathbf{n}w. \tag{164}$$

This yields the interpretation that the energy in a sound wave is moving in the direction of propagation with the sound speed. Consequently, the sound speed can be regarded as an energy propagation velocity. (This is in accord with the fact that sound waves in this idealization are nondispersive, so the group and phase velocities are the same.)

6.3 Acoustic Power

Many sound fields can be idealized as being steady, such that long-term time averages are insensitive to the duration and the center time of the averaging interval. Constant-frequency sounds and continuous noises fall into this category. For such sounds, the time derivative of the acoustic energy density will average out to zero over a sufficiently long time period, so the acoustic energy corollary of Eq. (158), in the absence of dissipation, yields the time average relation

$$\nabla\cdot\mathbf{I}_{\text{av}}=0. \tag{165}$$

This implies that the time-averaged vector intensity field is solenoidal in regions that do not contain acoustic sources. This same relation holds for any frequency component of the acoustic field or for the net contribution to the field from any given frequency band. In the following discussion, the intensity $\mathbf{I}_{\text{av}}$ is understood to refer to such a time average for some specified frequency band.

The relation of Eq. (165) yields the integral relation

$$\int_S \mathbf{I}_{\text{av}}\cdot\mathbf{n}\,dS=0, \tag{166}$$

which is interpreted as a statement that the net acoustic power flowing out of any region not containing sources must be zero on the time average and for any given frequency band.

For a closed surface which encloses one or more sources, such that the governing linear acoustic equations do not apply at every point within the volume, the reasoning above allows one to define the time-averaged net acoustic power of these sources as

$$\mathscr{P}_{\mathrm{av}}=\int_S \mathbf{I}_{\mathrm{av}}\cdot\mathbf{n}\, dS, \tag{167}$$

where the surface S encloses the sources. It follows from Eq. (166) that the acoustic power of a source computed in such a manner will be the same for any two choices for the surface S, provided that both surfaces enclose the same source and no other sources. The value of the integral is independent of the size and of the shape of S. This result is of fundamental importance for the measurement of source power. Instrumentation to measure the time-averaged intensity directly has become widely available in recent years and is often used in determining the principal noise sources in complicated environments.

6.4 Rate of Energy Dissipation

For the circumstances in which viscosity, thermal conduction, and internal relaxation are to be taken into account, the linear acoustic equations for the acoustic mode, which yield the dissipative wave equation of Eq. (136) and the relaxation equations described by Eq. (137), yield (Pierce, 1978) an energy corollary of the form of Eq. (158). To a satisfactory approximation the energy density and intensity remain as given by Eqs. (160) and (161). The energy dissipation per unit volume is no longer zero, but is instead

$$\mathscr{D}_{\mathrm{av}}=2\frac{\delta_{\mathrm{cl}}}{\rho c^4}\left(\frac{\partial p}{\partial t}\right)^2+2\sum_{\nu}\frac{(\Delta c)_{\nu}}{\rho c^3}\tau_{\nu}\left(\frac{\partial p_{\nu}}{\partial t}\right)^2 . \tag{168}$$

For constant-frequency plane waves propagating in the x direction, the time average of the time derivative of the energy density is zero, and the time averages of $\mathbf{I}$ and $\mathscr{D}$ will both be quadratic in the wave amplitude $|\hat{p}|$. The identification of the attenuation coefficient α must then be such that

$$\mathscr{D}_{\mathrm{av}}=2\alpha I_{\mathrm{av}}. \tag{169}$$

The magnitude of the acoustic intensity decreases with propagation distance x as

$$I_{\mathrm{av}}(x)=I_{\mathrm{av}}(0)e^{-2\alpha x}. \tag{170}$$

This identification of the attenuation coefficient is consistent with that given in Eq. (143).

6.5 Energy Corollary for Elastic Waves in Solids

An energy conservation corollary of the form of Eq. (158) also holds for sound in solids. The appropriate identifications for the energy density w and the components I_i of the intensity are

$$w=\frac{1}{2}\rho\sum_i\left(\frac{\partial\xi_i}{\partial t}\right)^2+\frac{1}{2}\sum_{i,j}\epsilon_{ij}\sigma_{ij}, \tag{171}$$

$$I_i=-\sum_j\sigma_{ij}\frac{\partial\xi_j}{\partial t}. \tag{172}$$

7. CYLINDRICAL AND SPHERICAL WAVES

7.1 Spherical Waves

A type of propagation of fundamental importance is that of a spherically symmetric wave spreading out radially from a source in an unbounded medium. The symmetry implies that the acoustic field variables are functions of only the radial coordinate r and time t. The Laplacian then reduces to

$$\nabla^2 p=\frac{\partial^2 p}{\partial r^2}+\frac{2}{r}\frac{\partial p}{\partial r}=\frac{1}{r}\frac{\partial^2(rp)}{\partial r^2}, \tag{173}$$

so the wave equation of Eq. (55) takes the form

$$\frac{\partial^2(rp)}{\partial r^2}-\frac{1}{c^2}\frac{\partial^2(rp)}{\partial t^2}=0. \tag{174}$$

The solution of this is given by

$$p(r,t)=r^{-1}f(r-ct)+r^{-1}g(r+ct). \tag{175}$$

Causality considerations (no sound before source is turned on) lead to the conclusion that the second term on the right-hand side of Eq. (175) is not an appropriate solution of the wave equation when the source is concentrated near the origin. The expression

$$p(r,t)=r^{-1}f(r-ct), \tag{176}$$

which describes the acoustic pressure in an outgoing spherically symmetric wave, has the property that listeners at different radii will receive (with a time shift corresponding to the propagation time) wave forms of the same shape, but of different amplitudes. The factor of $1/r$ is characteristic of spherical spreading and implies that the peak waveform amplitudes in a spherical wave decrease with radial distance as $1/r$.

Spherical waves of constant frequency have complex amplitudes governed by the Helmholtz equation, Eq. (102), with the Laplacian given as stated in Eq. (173). The complex pressure amplitude corresponding to Eq. (176) has the form of

$$p=Ar^{-1}e^{ikr}. \tag{177}$$

The fluid velocity associated with an outgoing spherical wave is purely radial and has the form

$$v_r=(1/\rho c)[-r^{-2}F(r-ct)+r^{-1}f(r-ct)]. \tag{178}$$

Here the function F is such that its derivative is the function f that appears in Eq. (176). Because the first term (a near-field term) decreases as the square rather than the first power of the reciprocal of the radial distance, the fluid velocity asymptotically approaches

$$v_r\rightarrow p/\rho c, \tag{179}$$

which is the same as the plane-wave relation of Eq. (107).

For outgoing spherical waves of constant frequency, the complex amplitude of the fluid velocity is

$$\hat{v}_r=\frac{1}{\rho c}\left(1-\frac{1}{ikr}\right)\hat{p}. \tag{180}$$

7.2 Cylindrical Waves

For cylindrically symmetric waves, there is no dependence on the azimuthal angle θ or on the axial coordinate z, so the Laplacian in cylindrical coordinates reduces to

$$\nabla^2=\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right), \tag{181}$$

where r is here the radial distance from the z axis. Consequently, the wave equation of Eq. (55) takes the form

$$\frac{\partial^2(r^{1/2}p)}{\partial r^2}-\frac{1}{c^2}\frac{\partial^2(r^{1/2}p)}{\partial t^2}+\frac{r^{1/2}p}{4r^2}=0. \tag{182}$$

and the Helmholtz equation of Eq. (102) takes the form

$$\frac{d^2\hat{p}}{dr^2}+\frac{1}{r}\frac{d\hat{p}}{dr}+k^2\hat{p}=0. \tag{183}$$

The solution of the latter that corresponds to an outgoing wave is a Hankel function of the first kind (Morse and Feshbach, 1953),

$$p=AH_0^{(1)}(kr). \tag{184}$$

For waves that are not of constant frequency, the wave-form shapes of outward propagating waves governed by the cylindrical wave equation, Eq. (182), tend to distort with increasing propagation distance, especially so at small r. However, at larger values of r it is often a good approximation to neglect the last term in that equation, so one obtains the approximate solution

$$p(r,t)\approx f(r-ct)/r^{1/2}, \tag{185}$$

which is similar to the expression of Eq. (176) for an outgoing spherical wave, only here the amplitude drops off with r as $r^{-1/2}$.

The latter approximate expression is consistent with the constant-frequency solution given by Eq. (184), since the Hankel function asymptotically approaches the limit

$$\lim_{kr\to\infty} H_0^{(1)}(kr)=\left(\frac{2}{\pi kr}\right)^{1/2}e^{-i\pi/4}e^{ikr}. \tag{186}$$

The fluid velocity induced by outgoing cylindrical waves is not as simply related to the corresponding acoustic pressure as that induced by a plane wave, although symmetry directs that the velocity must be in the appropriate radial direction when the propagation is cylindrically symmetric. For the constant-frequency case, an expression may be determined from the radial component of (101), such that the velocity is expressed in terms of the Hankel function of first order. However, a simple approximate result emerges in the limit of large radial distance r, this being the plane-wave relation of Eq. (179). (Here large r implies large compared to a characteristic wavelength, or compared to c divided by a characteristic angular frequency.)

8. SIMPLE SOURCES OF SOUND, GREEN'S FUNCTIONS

8.1 Volume Sources

Sources in acoustics are incorporated into the governing equations through either boundary conditions or source terms. Sources that are some distance from bounding surfaces and that are small compared to a wavelength can frequently be described by source terms. The simplest such source would be one which causes a net amount of mass of fluid to flow out of or into a surface that encases it. The mass passing out per unit time divided by the ambient density ρ is a quantity $Q_S(t)$ termed the source-strength function or the source volume velocity. If such a source is concentrated at a point $\mathbf{x}_0$, then the appropriate inhomogeneous wave equation which would replace Eq. (55) would be

$$\nabla^2 p - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = -\rho \dot{Q}_S(t)\delta(\mathbf{x}-\mathbf{x}_0), \qquad (187)$$

where $\delta(\mathbf{x})$ is the Dirac delta function, which has a volume integral of unity and which is concentrated at the point where its argument vanishes.

The solution of the above inhomogeneous wave equation for an isolated source in an unbounded region is

$$p = \frac{\rho}{4\pi r}\dot{Q}_S(t-[r/c]) = \frac{S(t-[r/c])}{r}, \qquad (188)$$

where

$$S(t) = \frac{\rho}{4\pi}\frac{d}{dt}Q_S(t) \qquad (189)$$

is called the monopole strength. The replacement of the argument by the retarded time $t - [r/c]$ accounts for the transit time lag r/c for the sound to propagate from the source to the listener.

The classic example of a monopole source is a pulsating small sphere. In such a case, $Q_S(t)$ can be taken as the time derivative of the instantaneous volume within the sphere. Thus the radiated acoustic pressure is proportional to the volume acceleration, the second derivative with respect to time of the sphere's volume, so that

$$p = \frac{\rho}{4\pi r}\left(\frac{d^2V}{dt^2}\right)_{t\to t-[r/c]}. \qquad (190)$$

This rule holds for bodies that are not necessarily spherically shaped, provided the largest dimensions are small compared to a characteristic wavelength and provided the acoustic pressure is measured at a sufficiently large distance from the source.

8.2 Small Piston in a Rigid Baffle

Another example of a volume source is that of a very small (relative to a wavelength) piston mounted in an infinite rigid baffle. If the piston has area A and outward normal velocity $v_n(t)$, then symmetry (or, equivalently, the inclusion of an image source) requires that the radiated sound be the same as from an isolated source with twice the volume velocity of the piston, so one has

$$p = (\rho/2\pi r)A\,\dot{v}_n(t-[r/c]). \qquad (191)$$

8.3 Multiple and Distributed Sources

For assemblages of sources, each concentrated at a point, the generalization of Eq. (187) is to replace the right-hand side by a sum of individual source terms; the solution to this inhomogeneous wave equation is given by

$$p = \sum_n \frac{S(t-[R_n/c])}{R_n}, \qquad (192)$$

where R_n is the distance of the listener from the nth source.

When the source is continuously distributed in space, the point-source term is replaced by a smoothly varying function, so the inhomogeneous wave equation is now of the form

$$\nabla^2 p - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = -\rho\,\dot{q}_S(\mathbf{x},t) = -4\pi s(\mathbf{x},t). \qquad (193)$$

Here the function q_S is termed the source-strength density (source strength per unit volume). For an unbounded medium, the solution of this latter equation is given by the definite integral

$$p(\mathbf{x},t) = \int \frac{1}{R} s(\mathbf{x}_0, t-[R/c])\,dV_0, \qquad (194)$$

where the volume integration ranges over source position $\mathbf{x}_0$, and where $R = |\mathbf{x} - \mathbf{x}_0|$ is the distance between listener and source positions.

8.4 Thermoacoustic Sources

The differential equation of Eq. (193) arises when one suddenly adds heat to a fluid (Chu, 1955; Westervelt and Larson, 1973), as with a laser or by combustion, so that the entropy s per unit mass changes according to the thermodynamic relation

$$\rho T\, Ds/Dt=h, \tag{195}$$

and so that, to first order, the equation-of-state relation, Eq. (68), yields

$$\frac{\partial p'}{\partial t}=c^2\frac{\partial p'}{\partial t}+\frac{c^2\beta}{c_p}h. \tag{196}$$

Here h is the heat added per unit time and unit volume, β is the volume expansion coefficient defined by Eq. (26), and c_p is the specific heat at constant pressure. Equation (196) changes the basic linear acoustic equations, Eqs. (52) and (53), to

$$\frac{\partial p}{\partial t}+\rho c^2\nabla\cdot\mathbf{v}=\frac{c^2\beta}{c_p}h, \tag{197}$$

$$\rho_0\frac{\partial\mathbf{v}}{\partial t}+\nabla p=0, \tag{198}$$

so the energy-conservation corollary of Eq. (158) becomes

$$\frac{\partial w}{\partial t}+\nabla\cdot\mathbf{I}=\frac{\beta}{\rho c_p}ph, \tag{199}$$

and the wave equation becomes

$$\nabla^2 p-\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}=-\frac{\beta}{c_p}\frac{\partial h}{\partial t}. \tag{200}$$

The appropriate identification for the monopole strength density function $s(\mathbf{x},t)$ in the integral expression of Eq. (194) is consequently given by

$$s(\mathbf{x},t)=\frac{1}{4\pi}\frac{\beta}{c_p}\frac{\partial h}{\partial t}. \tag{201}$$

8.5 Green's Functions

The idealization of a point source of constant frequency is of basic importance in formulating solutions to complicated acoustic radiation problems. The resulting complex pressure amplitude of the field resulting from such a single source can be expressed as a constant times a Green's function, where the Green's function is a solution of the inhomogeneous Helmholtz equation

$$(\nabla^2+k^2)G(\mathbf{x}\,|\,\mathbf{x}_0)=-4\pi\delta(\mathbf{x}-\mathbf{x}_0). \tag{202}$$

If the medium surrounding the source is unbounded, then the Green's function is the free-space Green's function, which can be identified from Eq. (188) to have the form

$$G(\mathbf{x}\,|\,\mathbf{x}_0)=R^{-1}\,e^{ikR}, \tag{203}$$

where R is the distance between the source and observation (listener) point. When external boundary conditions are imposed on the Green's function that satisfies Eq. (202), then the Green's function will have a form different from that of the free-space Green's function. However, in all such cases, the Green's function will approach $1/R$ plus a bounded function when the listener's position approaches the source point.

9. DIPOLES, QUADRUPOLES, AND SPHERICAL HARMONICS

9.1 Multipole Series

Radiation fields from sources of limited spatial extent in unbounded environments can be described in terms of either multipoles or spherical harmonics. That such descriptions are feasible can be demonstrated with the aid of the constant-frequency version of Eq. (194):

$$\hat{p}(\mathbf{x})=\int\hat{s}(\mathbf{x}_0)\frac{e^{ikR}}{R}\,dV_0, \tag{204}$$

which is the solution of the inhomogeneous Helmholtz equation with a continuous distribution of monopole sources taken into account.

The multipole series results when one expands $R^{-1}e^{ikR}$ in a power series in the coordinates of the source position $\mathbf{x}_0$ and then integrates term by term. Up through second order one obtains

$$\hat{p}=\hat{S}\,\frac{e^{ikr}}{r}-\sum_{\nu=1}^{3}\hat{D}_\nu\frac{\partial}{\partial x_\nu}\frac{e^{ikr}}{r}+\sum_{\mu,\nu=1}^{3}\hat{Q}_{\mu\nu}\frac{\partial^2}{\partial x_\mu\partial x_\nu}\frac{e^{ikr}}{r}, \tag{205}$$

where the coefficients are expressed by

$$\hat{S}=\int\hat{s}(\mathbf{x})dV, \tag{206}$$

$$\hat{D}_\nu=-\int x_\nu\hat{s}(\mathbf{x})dV, \tag{207}$$

$$\hat{Q}_{\mu\nu}=\frac{1}{2!}\int x_\mu x_\nu \hat{s}(\mathbf{x})dV. \tag{208}$$

The three terms in Eq. (205) are said to be the monopole, dipole, and quadrupole terms, respectively. The coefficients $\hat{S}$, $\hat{D}_\nu$, and $\hat{Q}_{\mu\nu}$ are similarly labeled. The D_ν are the components of a dipole moment vector, while the $Q_{\mu\nu}$ are the components of a quadrupole moment tensor. The general validity of such a description extends beyond the manner of derivation (Lamb, 1925; Lighthill, 1978) and is not restricted to sound generated by a continuous source distribution embedded in the fluid. It applies in particular to the sound radiated by a vibrating body of arbitrary shape.

9.2 Acoustically Compact Sources

If the source is acoustically compact, so that its largest dimension is much less than a wavelength, the multipole series converges rapidly, so one typically only need retain the first nonzero term. Sources exist whose net monopole strength is zero, and sources also exist whose dipole moment vector components are all zero as well. Consequently, compact sources are frequently classed as monopole, dipole, and quadrupole sources. The prototype of a monopole source is a body of oscillating volume. One for a dipole source is a rigid solid undergoing translational oscillations; another would be a vibrating plate or shell whose thickness changes negligibly. In the former case, the detailed theory shows that in the limit of sufficiently low frequency, the dipole moment vector is given by

$$\hat{D}_\nu=\hat{F}_\nu+m_d\hat{a}_{C,\nu}, \tag{209}$$

where $\hat{F}_\nu$ is associated with the force the moving body exerts on the surrounding fluid and where $\hat{a}_{C,\nu}$ is associated with the acceleration of the geometric center of the body. The quantity m_d is the mass of fluid displaced by the body.

The simplest example of a dipole source is that of a rigid sphere transversely oscillating along the z axis about the origin. If the radius of the sphere is a and if $ka \ll 1$, then the force and acceleration have only a z component, and the force amplitude is given by

$$\hat{F}_z=\tfrac{1}{2}m_d\hat{a}_{C,z}. \tag{210}$$

The dipole moment when the center velocity has amplitude $\hat{v}_C$ is consequently of the form

$$\hat{D}_z=-\tfrac{3}{2}i\omega(\tfrac{4}{3}\rho\pi a^3)\,\hat{v}_C\,. \tag{211}$$

Taking into account that the derivative of r with respect to z is cos θ, one finds the acoustic field from Eq. (205) to be given by

$$\hat{p}=-\hat{D}_z\cos\theta\frac{d}{dr}\frac{e^{ikr}}{r}\,. \tag{212}$$

When $kr \gg 1$, this approaches the limiting form

$$\hat{p}\to -ikD_z\cos(\theta)e^{ikr}/r. \tag{213}$$

The far-field intensity, in accord with Eq. (161), has a time average of $|\hat{p}|^2/2\rho c$ and is directed in the radial direction in this asymptotic limit. The drop of intensity as $1/r^2$ with increasing radial distance is the same as for spherical spreading, but the intensity varies with direction as $\cos^2\theta$.

Vibrating bodies that radiate as quadrupole sources (and which therefore have no dipole radiation) usually do so because of symmetry. Vibrating bells (Stokes, 1868) and tuning forks are typically quadrupole radiators.

A dipole source can be represented by two similar monopole sources, 180° out of phase with each other, and very close together. Since they are radiating out of phase, there is no total mass flow input into the medium. Such a dipole source will have a net acoustic power output substantially lower than that of either of the component monopoles radiating individually. Similarly, a quadrupole can be formed by two identical but oppositely directed dipoles brought very close together. If the two dipoles have a common axis, then a longitudinal quadrupole results; when they are side by side, a lateral quadrupole results. In either case, the quadrupole radiation is much weaker than that from either dipole radiating separately.

9.3 Spherical Harmonics

The closely related general description of source radiation in terms of spherical harmonics results from Eq. (204) when one inserts the expansion (Morse, 1936; Morse and Feshbach, 1953; Morse and Ingard, 1968)

$$\frac{e^{ikR}}{R}=\sum_{l=0}^{\infty}(2l+1)j_l(kr_0)h_l^{(1)}(kr)P_l(\cos\Theta), \tag{214}$$

where j_l is the spherical Bessel function and $h_l^{(1)}$ is the spherical Hankel function of order l and the first kind. (The expansion here assumes $r>r_0$; otherwise, one interchanges r and r_0 in the above equation). The quantity $P_l(\cos\Theta)$ is the Legendre function of order l; the angle Θ is that between the directions of $\mathbf{x}$ and $\mathbf{x}_0$. Alternatively, one uses the expansion

$$P_l(\cos\Theta)=\sum_{m=-l}^{l}\frac{(l-|m|)!}{(l+|m|)!}\times Y_l^m(\theta,\phi)Y_l^{-m}(\theta_0,\phi_0), \tag{215}$$

where the spherical harmonics are defined by

$$Y_l^m(\theta,\phi)=e^{im\phi}P_l^{|m|}(\cos\theta). \tag{216}$$

Here the functions $P_l^{|m|}(\cos\theta)$ are the associated Legendre functions. [The value of $P_0(\cos\Theta)$ is identically 1.]

If such an expansion is inserted into Eq. (204), and if r is understood to be sufficiently large that there are no sources beyond that radius, one has the relation

$$\hat{p}(r,\theta,\phi)=a_{00}h_0^{(1)}(kr)+\sum_{l=1}^{\infty}\sum_{m=-l}^{l}a_{lm}\times h_l^{(1)}(kr)Y_l^m(\theta,\phi) \tag{217}$$

with coefficients given by

$$a_{lm}=ik(2l+1)\frac{(l-|m|)!}{(l+|m|)!}\times\int\hat{s}(r_0,\theta_0,\phi_0)j_l(kr_0)\times Y_l^{-m}(\theta_0,\phi_0)dV_0. \tag{218}$$

The indicated volume integrations are to be carried out in spherical coordinates. The general result of Eq. (218) holds for any source of limited extent; any such wave field in an unbounded medium must have such an expansion in terms of spherical Hankel functions and spherical harmonics.

The spherical Hankel functions have the asymptotic (large r) form

$$h_l^{(1)}(kr)\rightarrow(-i)^{(l+1)}e^{ikr}/kr, \tag{219}$$

so the acoustic radiation field must asymptotically approach

$$\hat{p}\rightarrow\hat{F}(\theta,\phi)e^{ikr}/r, \tag{220}$$

where the function $\hat{F}(\theta,\phi)$ is a function of θ and ϕ that has an expansion in terms of spherical harmonics. In this asymptotic limit, the acoustic intensity is in the radial direction and given by

$$I_{r,\mathrm{av}}=\tfrac{1}{2}|\hat{F}|^2/\rho c r^2. \tag{221}$$

For fixed θ and ϕ, the time-averaged intensity must asymptotically decrease as $1/r^2$. The coefficient of $1/r^2$ in the above equation describes the far-field radiation pattern of the source, having the units of watts per steradian.

Although the two types of expansions, multipoles and spherical harmonics, are related, the relationship is not trivial. The quadrupole term in Eq. (205), for example, cannot be equated to the sum of the $l=2$ terms in Eq. (217). It is possible to have spherically symmetric quadrupole radiation, so an $l=0$ term would have to be included.

10. INTEGRAL EQUATIONS IN ACOUSTICS

10.1 The Helmholtz–Kirchoff Integral

The analysis of radiation of sound from a vibrating body of limited extent in an unbounded region is often aided by an integral corollary of the Helmholtz equation of Eq. (102) which dates back to 19th century works of Helmholtz (1860) and Kirchoff (1883). One considers a closed surface S where the outward normal component of the particle velocity has complex amplitude $\hat{v}_n(\mathbf{x}_S)$ and complex pressure amplitude $\hat{p}_S(\mathbf{x}_S)$ at a point $\mathbf{x}_S$ on the surface. For notational convenience, one introduces a quantity defined by

$$\hat{f}_S=-i\omega\rho\hat{v}_n, \tag{222}$$

where $\hat{v}_n(\mathbf{x}_S)$ is the normal component $\mathbf{n}(\mathbf{x}_S)\cdot\hat{\mathbf{v}}(\mathbf{x}_S)$ of the complex fluid velocity vector amplitude $\hat{\mathbf{v}}(\mathbf{x})$ at the surface. One can regard $\hat{f}_S$ as a convenient grouping of symbols, either as a constant times the normal velocity or as a constant times the normal acceleration, or as the normal component of the apparent body force, per unit volume, exerted on the fluid at the surface. Because of the latter identification, the use of the symbol f is appropriate. The subscript S is used to denote values appropriate to the surface.

Then, given that there are no sources outside the surface, a mathematical derivation, involving the differential equation of Eq. (202) for the free-space Green's function of

Eq. (203), involving the Helmholtz equation, and involving Gauss's theorem, yields

$$\hat{p}(\mathbf{x})=\mathscr{M}\{\mathbf{x},\hat{p}_S,\hat{f}_S\} \tag{223}$$

for the complex pressure amplitude $\hat{p}$ at a point $\mathbf{x}$ outside the surface. Here the right-hand side is given by the expression

$$\mathscr{M}=\frac{1}{4\pi}\int[\hat{f}_S(\mathbf{x}'_S)G(\mathbf{x}|\mathbf{x}'_S) + \hat{p}_S(\mathbf{x}'_S)\mathbf{n}(\mathbf{x}'_S)\cdot\{\nabla' G(\mathbf{x}|\mathbf{x}')\}_{\mathbf{x}'=\mathbf{x}'_S}]dS'. \tag{224}$$

In the integrand of this expression, the point $\mathbf{x}'_S$ (after evaluation of any requisite normal derivatives) is understood to range over the surface S, with the point $\mathbf{x}$ held fixed during the integration. The unit outward normal vector $\mathbf{n}(\mathbf{x}'_S)$ points out of the enclosed volume V at the surface point $\mathbf{x}'_S$.

In Eq. (224), the integral $\mathscr{M}$ is a function of the point $\mathbf{x}$, but a functional (function of a function) of the function arguments $\hat{p}_S$ and $\hat{f}_S$.

10.2 Integral Equations for Fields on the Surface

The functions $\hat{p}_S$ and $\hat{f}_S$ cannot be independently prescribed on the surface S. Specifying either one is a sufficient inner boundary condition on the Helmholtz equation. The corollary of Eq. (223) applies only if both functions correspond to a physically realizable radiation field outside the surface S. If this is so and if the point $\mathbf{x}$ is formally set to a point inside the enclosed volume, then analogous mathematics involving the properties of the Green's function leads to the deduction (Copley, 1968; Schenck, 1968)

$$\mathscr{M}\{\mathbf{x}_{\text{inside}},\hat{p}_S,\hat{f}_S\}=0, \tag{225}$$

which is a general relation between the surface value functions $\hat{p}_S$ and $\hat{f}_S$, holding for any choice of the point $\mathbf{x}$ inside the enclosed volume.

Equation (223) allows one to derive two additional relations (distinguished by subscripts I and II) between the surface values of $\hat{p}_S$ and $\hat{f}_S$. One results when the off-surface point $\mathbf{x}$ is allowed to approach an arbitrary but fixed surface point $\mathbf{x}_S$. For one of the terms in the integral defining the quantity $\mathscr{M}$, the limit as $\mathbf{x}$ approaches $\mathbf{x}_S$ of the integral is not the same as the integral over the limit of the integrand as $\mathbf{x}$ approaches $\mathbf{x}_S$. With this subtlety taken into account, one obtains

$$\hat{p}_S(\mathbf{x}_S)-\mathscr{L}_{\text{I}}\{\mathbf{x}_S,\hat{p}_S\}=\mathscr{H}_{\text{I}}\{\mathbf{x}_S,\hat{f}_S\}, \tag{226}$$

where the two linear operators $\mathscr{L}_{\text{I}}$ and $\mathscr{H}_{\text{I}}$ are

$$\mathscr{L}_{\text{I}}\{\mathbf{x}_S,\hat{p}_S\}=\frac{1}{2\pi}\int\hat{p}(\mathbf{x}'_S)\mathbf{n}(\mathbf{x}'_S)\cdot\{\nabla' G(\mathbf{x}_S|\mathbf{x}')\}_{\mathbf{x}'=\mathbf{x}'_S}dS', \tag{227}$$

$$\mathscr{H}_{\text{I}}\{\mathbf{x}_S,\hat{f}_S\}=\frac{1}{2\pi}\int\hat{f}(\mathbf{x}'_S)G(\mathbf{x}'_S|\mathbf{x}_S)dS'. \tag{228}$$

These linear operators operate on the surface values of $\hat{p}_S$ and $\hat{f}_S$, respectively, with the resultant in each case being a function of the position of the surface point $\mathbf{x}_S$.

The second type of surface relationship (Maue, 1949) is obtained by taking the gradient of both sides of Eq. (223), subsequently setting $\mathbf{x}$ equal to $\mathbf{x}_S+\epsilon\mathbf{n}(\mathbf{x}_S)$, where $\mathbf{x}_S$ is an arbitrary point on the surface, taking the dot product with $\mathbf{n}(\mathbf{x}_S)$, then taking the limit as ϵ goes to zero. The order of the processes, doing the integration and taking the limit, cannot be blindly interchanged, and some mathematical manipulations making use of the properties of the Green's function are necessary before one can obtain a relation in which all integrations are performed after all necessary limits are taken. The result is

$$-\mathscr{L}_{\text{II}}\{\mathbf{x}_S,\hat{p}_S\}=\hat{f}_S(\mathbf{x}_S)+\mathscr{H}_{\text{II}}\{\mathbf{x}_S,\hat{f}_S\}, \tag{229}$$

where the relevant operators are given by

$$\begin{aligned}\mathscr{L}_{\text{II}}&\{\mathbf{x},\hat{p}_S\}\\ &=[\mathbf{n}(\mathbf{x}_S)\times\nabla]\cdot\frac{1}{2\pi}\int[\mathbf{n}(\mathbf{x}'_S)\nabla'\hat{p}_S(\mathbf{x}'_S)]\\ &\quad\times G(\mathbf{x}_S|\mathbf{x}'_S)dS'+\frac{k^2}{2\pi}\int\mathbf{n}(\mathbf{x}_S)\\ &\quad\cdot\mathbf{n}(\mathbf{x}'_S)\,\hat{p}_S(\mathbf{x}'_S)G(\mathbf{x}_S|\mathbf{x}'_S)dS',\end{aligned} \tag{230}$$

$$\mathscr{H}_{\text{II}}\{\mathbf{x}_S,\hat{f}_S\}=\frac{1}{2\pi}\int\hat{f}_S(\mathbf{x}'_S)\mathbf{n}(\mathbf{x}_S)\cdot\{\nabla G(\mathbf{x}|\mathbf{x}'_S)\}_{\mathbf{x}=\mathbf{x}_S}dS'. \tag{231}$$

In regard to Eq. (230), one should note that the operator $\mathbf{n}(\mathbf{x}_S)\times\nabla$ involves only derivatives tangential to the surface, so that the integral on which it acts needs only be evaluated at surface points $\mathbf{x}_S$.

A variety of numerical techniques have been used and discussed in recent literature to solve either Eq. (225), (226), or (229), or some combination (Burton and Miller, 1971) of these for the surface pressure $\hat{p}_S$, given the surface force function $\hat{f}_S$. Once this is done, the radiation field at any external point **x** is found by numerical integration of the corollary integral relation of Eq. (223).

11. IMPEDANCE

11.1 Mechanical Impedance

The ratio of the complex amplitude of a sinusoidally varying force to the complex amplitude of the resulting velocity at a point on a vibrating object is called the mechanical impedance at that point. It is a complex number and usually a function of frequency. Other definitions (Beranek, 1972) of impedance are also in widespread use in acoustics.

11.2 Specific Acoustic Impedance

The specific acoustic impedance or unit area acoustic impedance $Z_S(\omega)$ for a surface is defined as

$$Z_S(\omega)=\hat{p}/\hat{v}_{\text{in}}, \tag{232}$$

where $\hat{v}_{\text{in}}$ is the component of the fluid velocity directed into the surface under consideration. Typically, the specific acoustic impedance, often referred to briefly as impedance without any adjective, is used to describe the acoustic properties of materials. In many cases, surfaces of materials abutting fluids can be characterized as locally reacting, so that Z_S is independent of the detailed nature of the acoustic pressure field. In particular, the locally reacting hypothesis implies that the velocity of the material at the surface is unaffected by pressures other than in the immediate vicinity of the point of interest. At a nominally motionless and passively responding surface, and when the hypothesis is valid, the appropriate boundary condition on the complex amplitude $\hat{p}$ that satisfies the Helmholtz equation is given by

$$i\omega\rho\hat{p}=-Z_S\nabla\hat{p}\cdot\mathbf{n}, \tag{233}$$

where **n** is the unit normal vector pointing out of the material into the fluid. A surface which is perfectly rigid has $|Z_S|=\infty$. The other extreme, where $|Z_S|=0$, corresponds to the ideal case of a pressure release surface. This is, for example, what is normally assumed for the upper surface of the ocean in underwater sound. Since a passive surface absorbs energy from the sound field, the time-averaged intensity component into the surface should be positive or zero. This observation leads to the requirement that the real part (specific acoustic resistance) of the impedance should always be non-negative. The imaginary part (specific acoustic reactance) may be either positive or negative.

11.3 Characteristic Impedance

For extended substances, a related definition is that of characteristic impedance Z_{char}, defined as the ratio of $\hat{p}$ to the complex amplitude $\hat{v}$ of the fluid velocity in the direction of propagation when a plane wave is propagating through the substance. As indicated by Eq. (107), this characteristic impedance is ρc, regardless of frequency and position in the field. The mks units (kg m^{-3} m/s) of specific acoustic impedance are referred to as mks rayls (after Rayleigh). The characteristic impedance of air under standard conditions is approximately 400 mks rayls, and that of water is approximately 1.5×10^6 mks rayls.

11.4 Radiation Impedance

The radiation impedance Z_{rad} is defined as $\hat{p}/\hat{v}_n$, where $\hat{v}_n$ corresponds to the outward normal component of velocity at a vibrating surface. For outward propagating spherically symmetric waves, where $\hat{p}$ is $Ar^{-1}e^{ikr}$, the surface can be taken as any sphere concentric with the origin, so the radiation impedance at radius r is given by

$$Z_{\text{rad}}=\hat{p}/\hat{v}_r=\rho c(1+i/kr)^{-1}. \tag{234}$$

The time-averaged intensity in the radial direction is therefore

$$\begin{aligned} I_{\text{av}}&=\tfrac{1}{2}\,\text{Re}(pv_r^*)\\ &=\tfrac{1}{2}\,|\hat{p}|^2\,\text{Re}(Z_{\text{rad}}^{-1})\\ &=(1/2\rho c)\,|\hat{p}|^2. \end{aligned} \tag{235}$$

The latter is the same as results for a plane wave, even though, for a spherical wave, the acoustic pressure is not in phase with the radial velocity, except in the asymptotic limit of large radial distances.

11.5 Acoustic Impedance

The term acoustic impedance Z_A is reserved (Beranek, 1972) for the ratio of $\hat{p}$ to the volume-velocity complex amplitude. Here volume velocity is the net volume of fluid flowing past a specified surface element per unit time in a specified directional sense. One may speak, for example, of the acoustic impedance of an orifice in a wall, of the acoustic impedance at the mouth of a Helmholtz resonator, and of the acoustic impedance at the end of a pipe.

12. REFLECTION, REFRACTION, AND TRANSMISSION

12.1 Reflection at a Plane Surface

When a plane wave reflects at a surface with finite specific acoustic impedance Z_S, a reflected wave is formed such that the angle of incidence θ_{I} equals the angle of reflection (law of mirrors). Here both angles are reckoned from the line normal to the surface and correspond to the directions of the two waves. If one takes the y axis as pointing out of the surface and the surface as coinciding with the $y=0$ plane, then an incident plane wave propagating obliquely in the $+x$ direction will have a complex pressure amplitude

$$\hat{p}_{\mathrm{in}}=\hat{f}\, e^{ik_x x}e^{-ik_y y}, \tag{236}$$

where $\hat{f}$ is a constant. (For transient reflection, the quantity $\hat{f}$ can be taken as the Fourier transform of the incident pressure pulse at the origin.) The two indicated wave-number components are $k_x=k\,\sin\theta_{\mathrm{I}}$ and $k_y=\cos\theta_{\mathrm{I}}$. The reflected wave has a complex pressure amplitude given by

$$\hat{p}_{\mathrm{refl}}=\mathscr{R}(\theta_I,\omega)\,\hat{f}\, e^{ik_x x}e^{ik_y y}, \tag{237}$$

where the quantity $\mathscr{R}(\theta_{\mathrm{I}},\omega)$ is the pressure-amplitude reflection coefficient.

Analysis that makes use of the boundary condition of Eq. (233) leads to the identification

$$\mathscr{R}(\theta_I,\omega)=\frac{\xi(\omega)\cos\theta_{\mathrm{I}}-1}{\xi(\omega)\cos\theta_{\mathrm{I}}+1} \tag{238}$$

for the reflection coefficient, with the abbreviation

$$\xi(\omega)=Z_S/\rho c, \tag{239}$$

which represents the ratio of the specific acoustic impedance of the surface to the characteristic impedance of the medium.

12.2 Reflection at an Interface

The above relations also apply, with an appropriate identification of the quantity Z_S, to sound reflection at an interface between two fluids with different sound speeds and densities (Green, 1838). Translational symmetry requires that the disturbance in the second fluid have the same apparent phase velocity ω/k_x (trace velocity) along the x axis as does the disturbance in the first fluid. This requirement is known as the trace velocity matching principle (Rayleigh, 1878; Pierce, 1989) and leads to the observation that k_x is the same in both fluids. One distinguishes two possibilities: the trace velocity is higher or lower than the sound speed c_2.

For the first possibility, one has the inequality

$$c_2<c_1/\sin\theta_{\mathrm{I}}, \tag{240}$$

and a propagating plane wave (transmitted wave) is excited in the second fluid, with complex pressure amplitude

$$\hat{p}_{\mathrm{trans}}=\mathscr{T}(\omega,\theta_{\mathrm{I}})\,\hat{f}\, e^{ik_x x}e^{ik_2 y\cos\theta_{\mathrm{II}}}, \tag{241}$$

where $k_2=\omega/c_2$ is the wave number in the second fluid and θ_{II} (angle of refraction) is the angle at which the transmitted wave is propagating. The trace velocity matching principle leads to Snell's law,

$$\sin\theta_{\mathrm{I}}/c_1=\sin\theta_{\mathrm{II}}/c_2. \tag{242}$$

The change in propagation direction from θ_{I} to θ_{II} is the phenomenon of refraction.

The requirement that the pressure be continuous across the interface yields the relation

$$1+\mathscr{R}=\mathscr{T}, \tag{243}$$

while the continuity of the normal component of the fluid velocity yields

$$\frac{\cos\theta_{\mathrm{I}}}{\rho_1 c_1}(1-\mathscr{R})=\frac{\cos\theta_{\mathrm{II}}}{\rho_2 c_2}\,\mathscr{T}. \tag{244}$$

From these one derives the reflection coefficient

$$\mathscr{R}=(Z_{\mathrm{II}}-Z_{\mathrm{I}})/(Z_{\mathrm{II}}+Z_{\mathrm{I}}), \tag{245}$$

which involves the two impedances defined by

$$Z_{\mathrm{I}}=\rho_1 c_1/\cos\theta_{\mathrm{I}}, \tag{246}$$

$$Z_{\mathrm{II}}=\rho_2 c_2/\cos\theta_{\mathrm{II}}. \tag{247}$$

The other possibility, which is the opposite of that in Eq. (240), can only occur when $c_2>c_1$ and, moreover, only if θ_{I} is greater than the critical angle

$$\theta_{\mathrm{cr}}=\arcsin(c_1/c_2). \tag{248}$$

In this circumstance, an inhomogeneous plane wave, propagating in the x direction, but dying out exponentially in the $+y$ direction, is excited in the second medium. Instead of Eq. (241), one has the transmitted pressure given by

$$\hat{p}_{\mathrm{trans}}=\mathscr{T}(\omega,\theta_{\mathrm{I}})\,\hat{f}\,e^{ik_x x}e^{-\beta k_2 y} \tag{249}$$

with

$$\beta=[(c_2/c_1)^2\sin^2\theta_{\mathrm{I}}-1]^{1/2}. \tag{250}$$

The previously stated equations governing the reflection and transmission coefficients are still applicable, provided one replaces $\cos\theta_{\mathrm{II}}$ by $i\beta$. This causes the magnitude of the reflection coefficient $\mathscr{R}$ to become unity, so the time-averaged incident energy is totally reflected. Acoustic energy is present in the second fluid, but its time average over a wave period stays constant once the steady state is reached.

12.3 Theory of the Impedance Tube

Impedance tubes are commonly used in the measurement of specific acoustic impedances; the underlying theory (Paris, 1927; Hall, 1939) is based for the most part on Eqs. (237)–(239). The incident and the reflected waves propagate along the axis of a cylindrical tube with the sample surface at one end. A loudspeaker at the other end creates a sinusoidal pressure disturbance which propagates down the tube. Reflections from the end covered with the test material create an incomplete standing wave pattern inside the tube. The wavelength of the sound emitted by the source can be adjusted, but it should be kept substantially larger than the pipe diameter, so that the plane-wave assumption holds. With k_x identified as being 0, the complex amplitude that corresponds to the sum of the incident and reflected waves has an absolute magnitude given by

$$|\hat{p}|=|\hat{f}|\,|1+\mathscr{R}e^{2iky}|, \tag{251}$$

where y is now the distance in front of the sample. The second factor varies with y and repeats at intervals of a half-wavelength, and varies from a minimum value of $1-|\mathscr{R}|$ to a maximum value of $1+|\mathscr{R}|$. Consequently, the ratio of the peak acoustic pressure amplitude $|\hat{p}|_{\max}$ (which occurs at one y position) to the minimum acoustic pressure amplitude $|\hat{p}|_{\min}$ (which occurs at a position a quarter-wavelength away) determines the magnitude of the reflection coefficient via the relation

$$\frac{|\hat{p}|_{\min}}{|\hat{p}|_{\max}}=\frac{1-|\mathscr{R}|}{1+|\mathscr{R}|}. \tag{252}$$

The phase δ of the reflection coefficient can be determined with use of the observation that the peak amplitudes occur at y values where $\delta+2ky$ is an integer multiple of 2π, while the minimum amplitudes occur where it is π plus an integer multiple of 2π. Once the magnitude and phase of the reflection coefficient are determined, the specific acoustic impedance can be found from Eqs. (130) and (131).

12.4 Transmission through Walls and Slabs

The analysis of transmission of sound through a wall or a partition (Cremer, 1942) is often based on the idealization that the wall is of unlimited extent. If the incoming plane wave has an angle of incidence θ_{I} and if the fluid on the opposite side of the wall has the same sound speed, then the trace velocity matching principle requires that the transmitted wave be propagating in the same direction. A common assumption when the fluid is air is that the compression in the wall is negligible, so the wall is treated as a slab which has a uniform velocity v_{sl} throughout its thickness. The slab moves under the influence of the incident, reflected, and transmitted sound fields according to the relation (corresponding to Newton's second law) given by

$$m_{\mathrm{sl}}\frac{\partial v_{\mathrm{sl}}}{\partial t}=p_{\mathrm{front}}-p_{\mathrm{back}}+\text{bending term}, \tag{253}$$

where m_{sl} is the mass per unit surface area of the slab. The front side is taken as the side facing the incident wave; the transmitted wave propagates away from the back side.

The "bending term" (discussed further below) accounts for any tendency of the slab to resist bending. If the slab is regarded as nonporous then the normal component of the fluid velocity at both the front and the back is regarded the same as the slab velocity itself. If it is taken as porous (Beranek, 1947, 1971) then these continuity equations are replaced by the relations

$$v_{\text{front}} - v_{\text{sl}} = v_{\text{back}} - v_{\text{sl}} = (1/R_f)(p_{\text{front}} - p_{\text{back}}), \tag{254}$$

where R_f is the specific flow resistance. The latter can be measured in steady flow for specific materials. For a homogeneous material, it is given by the product of the slab thickness h and the flow resistivity, the latter being a commonly tabulated property of porous materials.

In general, when one considers the reflection at and transmission through a slab, one can define a slab specific impedance Z_{sl} such that, with regard to complex amplitudes,

$$\hat{p}_{\text{front}} - \hat{p}_{\text{back}} = Z_{\text{sl}}\,\hat{v}_{\text{front}} = Z_{\text{sl}}\,\hat{v}_{\text{back}}, \tag{255}$$

where Z_{sl} depends on the angular frequency ω and the trace velocity $v_{\text{tr}} = c/\sin\theta_{\text{I}}$ of the incident wave over the surface of the slab. The value of the slab specific impedance can be derived using considerations such as those corresponding to Eqs. (254) and (255). In terms of the slab specific impedance, the transmission coefficient $\mathscr{T}$ is

$$\mathscr{T} = [1 + \tfrac{1}{2}(Z_{\text{sl}}/\rho c)\cos\theta_{\text{I}}]^{-1}. \tag{256}$$

The fraction τ of incident power that is transmitted is $|\mathscr{T}|^2$.

12.4.1 Transmission through Limp Plates. If the slab can be idealized as a limp plate (no resistance to bending) and not porous, the slab specific impedance Z_{sl} is $-i\omega m_{\text{sl}}$ and one obtains

$$\tau = \left[1 + \left(\frac{\omega m_{\text{sl}}}{2\rho c}\right)^2 \cos^2\theta_{\text{I}}\right]^{-1} \approx \left(\frac{2\rho c}{\omega m_{\text{sl}} \cos\theta_{\text{I}}}\right)^2 \tag{257}$$

for the fraction of incident power that is transmitted. The latter version, which typically holds at moderate to high audible frequencies, predicts that τ decreases by a factor of 4 when the slab mass m_{sl} per unit area is increased by a factor of 2. In the noise-control literature, this behavior is sometimes referred to as the mass law.

12.4.2 Transmission through Porous Blankets. For a porous blanket that has a specific flow resistance R_f, the specific slab impedance becomes

$$Z_{\text{sl}} = (1/R_f - 1/i\omega m_{\text{sl}})^{-1}, \tag{258}$$

and the resulting fraction of incident power that is transmitted can be found with substitution into Eq. (256), with τ equated to $|\mathscr{T}|^2$.

12.4.3 Transmission through Elastic Plates. If the slab is idealized as a Bernoulli–Euler plate with elastic modulus E, Poisson's ratio ν, and thickness h, the bending term in Eq. (253) has a complex amplitude given by

$$\text{bending term} = -B_{\text{pl}} k_x^4 \hat{v}_{\text{sl}}/(-i\omega), \tag{259}$$

where

$$B_{\text{pl}} = Eh^3/12(1-\nu^2) \tag{260}$$

is the plate bending modulus. The slab specific impedance is consequently given by

$$Z_{\text{sl}} = -i\omega m_{\text{sl}}[1 - (f/f_c)^2 \sin^4\theta_{\text{I}}], \tag{261}$$

where

$$f_c = (c^2/2\pi)(m_{\text{sl}}/B_{\text{pl}})^{1/2} \tag{262}$$

gives the coincidence frequency, the frequency at which the phase velocity of freely propagating bending waves in the plate equals the speed of sound in the fluid.

Although the simple result of Eq. (261) predicts that the fraction of incident power transmitted is unity at a frequency of $f_c/\sin^2\theta_{\text{I}}$, the presence of damping processes in the plate causes the fraction of incident power transmitted to be always less than unity. A simple way of taking this into account makes use of a loss factor η (assumed much less than unity) for the plate, which corresponds to the fraction of stored elastic energy that is dissipated through damping processes, during 1 rad (cycle period divided by 2π). The loss factor can be formally introduced into the mathematical model by the replacement of the real elastic modulus by the complex number $(1-i\eta)E$. When this is done one finds

$$Z_{\text{sl}} = \omega\eta m_{\text{pl}}(f/f_c)^2 \sin^4\theta_{\text{I}} - i\omega m_{\text{sl}}[1 - (f/f_c)^2 \sin^4\theta_{\text{I}}] \tag{263}$$

for the slab impedance that is to be inserted into Eq. (256). The extra term ordinarily has very little effect on the fraction of incident power that is transmitted except when the (normally dominant) imaginary term is close to zero. When the frequency f is $f_c/\sin^2\theta_I$, one finds the value of τ to be

$$\tau=[1+\tfrac{1}{2}(\omega\eta m_{\mathrm{pl}}/\rho c)\cos\theta_I]^{-2} \tag{264}$$

rather than identically unity.

13. WAVEGUIDES, DUCTS, AND RESONATORS

13.1 Guided Modes in a Duct

Pipes or ducts act as guides of acoustic waves, and the net flow of energy, other than that associated with wall dissipation, is along the direction of the duct. The general theory of guided waves applies and leads to a representation in terms of guided modes.

If the duct axis is the x axis and the duct cross section is independent of x, the guided mode series (Rayleigh, 1878; Hartig and Swanson, 1938) has the form

$$\hat{p}=\sum_n X_n(x)\Psi_n(y,z), \tag{265}$$

where the $\Psi_n(y,z)$ are eigenfunctions of the equation

$$\left(\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\Psi_n+\alpha_n^2\Psi_n=0 \tag{266}$$

with the α_n^2 being the corresponding eigenvalues. The appropriate boundary condition, if the duct walls are idealized as being perfectly rigid, is that the normal component of the gradient of Ψ_n vanishes at the walls. Typically, the Ψ_n are required to conform to some normalization condition, such as

$$\int \Psi_n^2 dA=A, \tag{267}$$

where A is the duct cross-sectional area.

The general theory leads to the conclusion that one can always find a complete set of Ψ_n which, with the rigid-wall boundary condition imposed, are such that the cross-sectional eigenfunctions are orthogonal, so that

$$\int \Psi_n\Psi_m dA=0 \tag{268}$$

if n and m correspond to different guided modes. The eigenvalues α_n^2, moreover, are all real and non-negative. However, for cross sections that have some type of symmetry, it may be that more than one linearly independent eigenfunction Ψ_n (modes characterized by different values of the index n) correspond to the same numerical value of α_n^2. In such cases the eigenvalue is said to be degenerate.

The variation of guided-mode amplitudes with source excitation is ordinarily incorporated into the axial wave functions $X_n(x)$, which satisfy the one-dimensional Helmholtz equation

$$\frac{d^2X_n}{dx^2}+(k^2-\alpha_n^2)X_n=0. \tag{269}$$

Here $k=\omega/c$ is the free-space wave number. The form of the solution depends on whether α_n^2 is greater or less than k^2. If $\alpha_n^2 < k^2$, the mode is said to be a propagating mode and the solution for X_n is given by

$$X_n=A_ne^{ik_nx}+B_ne^{-ik_nx}, \tag{270}$$

where k_n, defined by

$$k_n=(k^2-\alpha_n^2)^{1/2}, \tag{271}$$

are the modal wave numbers. If, on the other hand, the value of α_n^2 is greater than k^2, the mode is evanescent (not propagating), and one has

$$X_n=A_ne^{-\beta_nx}+B_ne^{\beta_nx}, \tag{272}$$

where β_n is given by

$$\beta_n=(\alpha_n^2-k^2)^{1/2}. \tag{273}$$

Unless the termination of the duct is relatively close to the source, waves that grow exponentially with distance from the source are not meaningful, so only the term that corresponds to exponentially dying waves is ordinarily kept in the description of sound fields in ducts.

13.2 Cylindrical Ducts

For the frequently encountered case of a duct with circular cross section and radius a, the index n is replaced by an index set (q,m,s), and the eigenfunctions Ψ_n are described by

$$\Psi_n = K_{qm} J_m\left(\frac{\eta_{qm} r}{a}\right)\left\{\begin{matrix}\cos\phi \\ \sin\phi\end{matrix}\right\}. \tag{274}$$

Here either the cosine ($s=1$) or the sine ($s=-1$) corresponds to an eigenfunction. The quantities K_{qm} are normalization constants, and the J_m are Bessel functions of order m. The corresponding eigenvalues are given by

$$\alpha_n = \eta_{qm}/a, \tag{275}$$

where the η_{qm} are the zeros of $\eta J_m'(\eta)$, arranged in ascending order with the index q ranging upwards from 1. The smaller roots (Abramowitz and Stegun, 1965) for the axisymmetric modes are $\eta_{1,0}=0.00$, $\eta_{2,0}=3.831\ 71$, and $\eta_{3,0}=7.015\ 59$, while those corresponding to $m=1$ are $\eta_{1,1}=1.841\ 18$, $\eta_{2,1}=5.331\ 44$, and $\eta_{3,1}=8.536\ 32$.

13.3 Low-Frequency Model for Duct Propagation

In many situations of interest (Stewart, 1922; Mason, 1927), the frequency of the acoustic disturbance is so low that only one guided mode can propagate, and all of the other modes are evanescent. Given that the walls can be idealized as rigid, there is always one mode which can propagate, this being the plane-wave mode for which the eigenvalue α_0 is identically zero. The other modes will all be evanescent if the value of k is less than the corresponding α_n for each such mode. This would be so if the frequency is less than the lowest cutoff frequency for any of the nonplanar modes. For the circular duct case discussed above, for example, this would require that the inequality

$$f < \frac{1.841}{2\pi}\frac{c}{a} \tag{276}$$

be satisfied.

When the single guided mode assumption is valid, and even if the duct cross-sectional area should vary with distance along the duct, the acoustic field equations can be replaced to a good approximation by the acoustic transmission-line equations,

$$\frac{\partial p}{\partial t} + \frac{\rho c^2}{A}\frac{\partial U}{\partial x} = 0, \tag{277}$$

$$\rho\frac{\partial U}{\partial t} = -A\frac{\partial p}{\partial x}. \tag{278}$$

Here U, equal to Av_x, is the volume velocity, the volume of fluid passing through the duct per unit time.

One of the implications of the low-frequency model described by Eqs. (277) and (278) is that the volume velocity and the pressure are both continuous even when the duct has a sudden change in cross-section area. The pressure continuity assumption is not necessarily a good approximation, but becomes better with decreasing frequency. An improved model sets the difference of the complex amplitudes of the upstream and downstream pressures just ahead of and after the junction equal to

$$\hat{p}_{\text{ahead}} - \hat{p}_{\text{after}} = Z_J \hat{U}_{\text{junction}}, \tag{279}$$

where the junctions's acoustic impedance Z_J is taken in the simplest approximation as $-i\omega M_{A,J}$, and where $M_{A,J}$ is a real number independent of frequency that is called the acoustic inertance of the junction. Approximate expressions for this acoustic inertance can be found in the literature cited in the text by Pierce (1989); a simple rule is that it is ordinarily less than $8\rho/3A_{\min}$, where $A_{\min}$ is the smaller of the two cross-sectional areas. One may note, moreover, that Z_J goes to zero when the frequency goes to zero.

When an incident wave is incident at a junction, reflected and transmitted waves are created in the two ducts. The pressure-amplitude reflection and transmission coefficients are given by

$$\mathscr{R} = \frac{Z_J + \rho c/A_2 - \rho c/A_1}{Z_J + \rho c/A_2 + \rho c/A_1}, \tag{280}$$

$$\mathscr{T} = \frac{2\rho c/A_2}{Z_J + \rho c/A_2 + \rho c/A_1}. \tag{281}$$

13.4 Sound Attenuation in Ducts

The presence of the duct walls affects the attenuation of sound and causes the attenuation coefficient ordinarily to be much higher than for plane waves in open space. A simple theory for predicting this attenuation uses the concepts of acoustic, entropy, and vorticity modes described in Sec. 2.6. The entropy- and vorticity-mode fields exist in an acoustic boundary layer near the duct walls and are such that they combine with the acoustic-mode field to satisfy the appropriate boundary conditions at the walls. With viscosity and thermal conduction thus taken into account in the acoustic boundary layer, and with the assumption that the duct walls are much better heat conductors than the fluid itself, the approximate attenuation coefficient for the plane-wave guided mode is given by (Mason, 1928)

$$\alpha=\alpha_{\text{walls}}$$
$$=\left(\frac{\omega}{8\rho c^2}\right)^{1/2}\left[\mu^{1/2}+(\gamma-1)\left(\frac{\kappa}{c_p}\right)^{1/2}\right]\frac{L_P}{A}, \tag{282}$$

where L_P is the length of the perimeter of the duct cross section.

A related effect is that sound travels slightly slower in ducts than in an open environment. The phase velocity (Henry, 1931) is given by

$$v_{\text{ph}}=c-(c^2\alpha_{\text{walls}}/\omega), \tag{283}$$

so the group velocity is midway between the phase velocity and the free-space sound velocity c (still less than c).

13.5 Mufflers and Acoustic Filters

The analysis of mufflers (Davies, 1964; Embleton, 1971; Bender and Bremmer, 1975) is often based on the idealization that their acoustic transmission characteristics are independent of sound amplitudes, so that they act as linear devices. The muffler is regarded as an insertion into a duct, which reflects waves back toward the source and which alters the transmission of sound beyond the muffler. The properties of the muffler vary with frequency, so the theory analyzes the muffler's effects on individual frequency components. The frequency is assumed to be sufficiently low that only the plane-wave mode propagates in the inlet and outlet ducts. Because of the assumed linear behavior of the muffler, and the single-mode assumption, the muffler conforms to the model of a linear two-port, the ports being the inlet and outlet. The model leads to the prediction that one may characterize the muffler at any given frequency by a matrix, such that

$$\begin{pmatrix}\hat{p}_{\text{in}}\\ \hat{p}_{\text{out}}\end{pmatrix}=\begin{pmatrix}K_{11} & K_{12}\\ K_{21} & K_{22}\end{pmatrix}\begin{pmatrix}\hat{U}_{\text{in}}\\ \hat{U}_{\text{out}}\end{pmatrix}, \tag{284}$$

where the coefficients K_{ij} represent the acoustical properties of the muffler. Reciprocity requires that the determinant of the matrix be unity. Also, for a symmetric muffler, K_{12} and K_{21} must be identical.

It is ordinarily a good approximation that the waves reflected at the entrance of the muffler back to the source are significantly attenuated, so that they have negligible amplitude when they return to the muffler. Similarly, the assumption is ordinarily valid that the waves transmitted beyond the muffler do not return to the muffler (anechoic termination). With these assumptions, the insertion of the muffler causes the mean squared pressure downstream of the muffler to drop by a factor

$$\tau=\left(\frac{1}{4}\left|K_{11}+K_{22}+\frac{\rho c}{A}K_{21}+\frac{A}{\rho c}K_{12}\right|^2\right)^{-1}, \tag{285}$$

where A is the cross section of the duct ahead of and behind the muffler.

Acoustic mufflers can be divided into two broad categories: reactive mufflers and dissipative mufflers. In a reactive muffler, the basic property of the muffler is that it reflects a substantial fraction of the incident acoustic energy back toward the source. The dissipation of energy within the muffler itself plays a minor role; the reflection is caused primarily by the geometrical characteristics of the muffler. In a dissipative muffler, however, a low transmission of sound is achieved by internal dissipation of acoustic energy within the muffler. Absorbing material along the walls is ordinarily used to achieve this dissipation.

13.5.1 Nonreflecting Dissipative Muffler. When a segment of pipe of cross section A and length L is covered with a duct-lining material that attenuates the amplitude of traveling plane waves by a factor of $e^{-\alpha L}$, the muffler's K matrix is given by

$$[K]=\begin{pmatrix}\cos(kL+i\alpha L) & -\frac{i\rho c}{A}\sin(kL+i\alpha L)\\ -\frac{iA}{\rho c}\sin(kL+i\alpha L) & \cos(kL+i\alpha L)\end{pmatrix}, \tag{286}$$

so the fractional drop in mean squared pressure reduces to

$$\tau=e^{-2\alpha L}. \tag{287}$$

13.5.2 Expansion Chamber Muffler. The simplest reactive muffler is the expansion chamber which consists of a duct of length L

and cross section A_M connected at both ends to a pipe of smaller cross section A_p. The K matrix for such a muffler is found directly from Eq. (286) by setting α equal to zero (no dissipation in the chamber), but replacing A by A_M. The corresponding result for the fractional drop in mean squared pressure is

$$\tau=[\cos^2 kL+\tfrac{1}{4}(m+m^{-1})^2\sin^2 kL]^{-1}, \qquad (288)$$

where m is the expansion ratio A_M/A_P. The relative reduction in mean square pressure is thus periodic with frequency. A maximum reduction occurs when the length L is an odd multiple of quarter wavelengths. The greatest reduction in mean squared pressure is when τ has its minimum value, which is that given by

$$\tau=4/(m+m^{-1})^2. \qquad (289)$$

13.6 Helmholtz Resonators

An idealized acoustic system (Helmholtz, 1860; Rayleigh, 1870) which is the prototype of commonly encountered resonance phenomena consists of a cavity of volume V with a neck of length l and cross section S. In the limit where the acoustic frequency is sufficiently low that the wavelength is much larger than any dimension of the resonator, the compressible fluid in the resonator acts as a spring with spring constant given by

$$k_{\text{sp}}=\rho c^2 S^2/V \qquad (290)$$

and the fluid in the neck behaves as a lumped mass of magnitude

$$m=\rho S l'. \qquad (291)$$

Here l' is l plus end corrections for the two ends of the neck. If l is somewhat larger than the neck radius a, and if both ends are terminated by flanges, then the two end corrections are each $0.82a$. The resonance frequency ω_r, in radians per second, of the resonator is given by

$$\omega_r=2\pi f_r=(k_{\text{sp}}/m)^{1/2}=(M_A C_A)^{-1/2}, \qquad (292)$$

where

$$M_A=\rho l'/S \qquad (293)$$

gives the acoustic inertance of the neck and

$$C_A=V/\rho c^2 \qquad (294)$$

gives the acoustic compliance of the cavity. The ratio of the complex pressure amplitude just outside the mouth of the neck to the complex volume velocity amplitude of flow into the neck is the acoustic impedance Z_{HR} of the Helmholtz resonator and given, with the neglect of damping, by

$$Z_{\text{HR}}=-i\omega M_A+1/(-i\omega C_A) \qquad (295)$$

and this vanishes at the resonance frequency.

If a Helmholtz resonator is inserted as a side branch into the wall of a duct, it acts as a reactive muffler that has a high insertion loss near the resonance frequency of the resonator. The analysis assumes that the acoustic pressures in the duct just before and just after the resonator are the same as the pressure at the mouth of the resonator. Also, the acoustical analog of Kirchhoff's circuit law for currents applies, so that the volume velocity flowing in the duct ahead of the resonator equals the sum of the volume velocities flowing into the resonator and through the duct just after the resonator. These relations, with Eq. (295), allow one to work out expressions for the amplitude reflection and transmission coefficients at the resonator, the latter being given by

$$\mathscr{T}=\frac{2Z_{\text{HR}}}{2Z_{\text{HR}}+\rho c/A_D}. \qquad (296)$$

The fraction of incident power that is transmitted is consequently given by

$$\tau=\left(1+\frac{1}{4\beta^2(f/f_r-f_r/f)^2}\right)^{-1}, \qquad (297)$$

where β is determined by

$$\beta^2=(M_A/C_A)(A_D/\rho c)^3. \qquad (298)$$

The fraction transmitted is formally zero at the resonance frequency, but if β is large compared with unity, the bandwidth over which the drop in transmitted power is small is narrow compared with f_r.

14. RAY ACOUSTICS

14.1 Wave-Front Propagation

The theory of plane-wave propagation predicts that wave fronts move with speed c when viewed in a coordinate system in which the ambient medium appears at rest. If the ambient medium is moving with velocity $\mathbf{v}$, the wave velocity $c\mathbf{n}$ seen by an observer moving with the fluid becomes $\mathbf{v}+c\mathbf{n}$ in a coordinate system at rest. Here $\mathbf{n}$ is the unit vector normal to the wave front; it coincides with the direction of propagation if the coordinate system is moving with the local ambient fluid velocity $\mathbf{v}$. A ray can be defined (Milne, 1921) as the time trajectory of a point that is always on the same wave front, and for which the velocity is

$$\mathbf{v}_{\text{ray}}=\mathbf{v}+\mathbf{n}c. \tag{299}$$

To determine ray paths without explicit knowledge of wave fronts, it is appropriate to consider a function $\tau(\mathbf{x})$, which gives the time at which the wave front of interest passes the point $\mathbf{x}$. Its gradient $\mathbf{s}$ is termed the wave slowness vector and is related to the local wave-front normal by the equations

$$\mathbf{s}=\mathbf{n}/(c+\mathbf{v}\cdot\mathbf{n}), \tag{300}$$

$$\mathbf{n}=c\mathbf{s}/\Omega, \tag{301}$$

where the quantity Ω is defined by

$$\Omega=1-\mathbf{v}\cdot\mathbf{s}. \tag{302}$$

Given the concepts of ray position and the slowness vector, a purely kinematical derivation leads to the ray-tracing equations

$$\frac{d\mathbf{x}}{dt}=\frac{c^2\mathbf{s}}{\Omega}+\mathbf{v}, \tag{303}$$

$$\frac{d\mathbf{s}}{dt}=-\frac{\Omega}{c}\nabla c-(\mathbf{s}\cdot\nabla)\mathbf{v}-\mathbf{s}\times(\nabla\times\mathbf{v}). \tag{304}$$

Simpler versions result when there is no ambient flow, or when the ambient variables vary with only one coordinate. If the ambient variables are independent of position, then the rays paths are straight lines.

The above equations are often used for analysis of propagation through inhomogeneous media (moving or nonmoving) when the ambient variables do not vary significantly over distances of the order of a wavelength, even though they may do so over the total propagation distance. The rays which connect the source and listener locations are termed the eigenrays for that situation. If there is more than one eigenray, one has multipath reception. If there is no eigenray, then the listener is in a shadow zone.

14.2 Reflected, Transmitted, and Diffracted Rays

The theory is readily extended to take into account solid boundaries and interfaces. When an incident ray strikes a solid boundary, a reflected ray is generated whose direction is determined by the trace velocity matching principle, such that the angle of incident equals the angle of reflection. At an interface a transmitted ray is also generated, whose direction is predicted by Snell's law.

An extension of the theory, known as the geometrical theory of diffraction and due primarily to J. B. Keller (1958), allows for the possibility that diffracting edges can be a source of diffracted rays which have a multiplicity of directions. The allowable diffracted rays must have the same trace velocity along the diffracting edge as does the incident ray.

The ray paths connecting source and listener satisfy Fermat's principle (Ugincius, 1972), that the travel time along an actual eigenray is stationary relative to other geometrically admissible paths. Keller's geometrical theory of diffraction (Keller, 1958) extends this principle to include paths which have discontinuous directions and which may have portions that lie on solid surfaces.

14.3 Wave Equation for Inhomogeneous Moving Media

The determination of the amplitude variation along ray paths requires a more explicit use of the equations of linear acoustics. If there is no ambient fluid velocity then an appropriate wave equation is that of Eq. (54). When there is also an ambient flow present, an appropriate generalization (Blokhintzev, 1946; Pierce, 1990) for the acoustic mode is

$$\frac{1}{\rho}\nabla\cdot(\rho\nabla\Phi)-D_t\left(\frac{1}{c^2}D_t\Phi\right)=0. \tag{305}$$

Here D_t, defined by

$$D_t=\frac{\partial}{\partial t}+\mathbf{v}\cdot\nabla, \tag{306}$$

is the time derivative following the ambient flow. The dependent variable Φ is related to the acoustic pressure perturbation by

$$p=-\rho D_t\Phi. \tag{307}$$

The above wave equation is derived from the linear-acoustics equations for circumstances when the ambient variables vary slowly with position and time, and neglects terms of second order in $1/\omega T$ and $c/\omega L$, where ω is a characteristic frequency of the disturbance, T is a characteristic time scale for temporal variations of the ambient quantities, and L is a characteristic length scale for spatial variations of these approximations. The nature of the entailed approximations is consistent with the notion that geometrical acoustics is a high-frequency approximation and yield results that are the same in the high-frequency limit as would be derived from the complete set of linear acoustic equations.

14.4 The Eikonal Approximation

The equations of geometrical acoustics follow (Sommerfeld and Runge, 1911) from Eq. (305), if one sets the potential function Φ equal to the expression

$$\Phi(\mathbf{x},t)=\Psi(\mathbf{x})F(t-\tau). \tag{308}$$

Here the function F is assumed to be a rapidly varying function of its argument. The quantities Ψ and τ vary only with position. (It is assumed here that the ambient variables do not vary with time.)

When such a substitution is made, one obtains terms that involve the second, first, and zeroth derivatives of the function F. The geometrical-acoustics or eikonal approximation results when one requires that the coefficients of F'' and F' both vanish identically, and thus assumes that the terms involving undifferentiated F are of lesser importance in the high-frequency limit. The setting equal to zero of the coefficient of the second derivative yields

$$(1-\mathbf{v}\cdot\nabla\tau)^2=c^2(\nabla\tau)^2, \tag{309}$$

which is termed the eikonal equation (Heller, 1953; Keller, 1954). The setting equal to zero of the coefficient of the first derivative yields

$$\nabla\cdot\left(\frac{\rho\Psi^2}{c^2}[c^2\nabla\tau+(1-\mathbf{v}\cdot\nabla\tau)\mathbf{v}]\right)=0, \tag{310}$$

which is termed the transport equation. The convenient property of these equations is that they are independent of the nature of the wave-form function F.

The eikonal equation, which is a nonlinear first-order partial differential equation for the eikonal $\tau(\mathbf{x})$, can be solved by Cauchy's method of characteristics. What results are the ray-tracing equations of Eqs. (303) and (304). However, the time variable t is replaced by the eikonal τ. The manner in which these equations solve the eikonal equation is such that if τ and its gradient $\mathbf{s}$ are known at any given point, then the ray-tracing equations determine these same quantities at all other points on the ray that passes through this initial point. The initial value of $\mathbf{s}$ must conform to the eikonal equation itself, as is stated in Eq. (309). If this is so, then the ray-tracing equations ensure that it conforms to the same equation, with the appropriate ambient variables, all along the ray.

Solution of the transport equation, given by Eq. (310), is facilitated by the concept of a ray tube. One conceives of a family of adjacent rays which comprise a narrow tube of cross-sectional area A, this area varying with distance l along the path of the central ray of the tube. Then the transport equation reduces to

$$\frac{d}{dl}\left(A\frac{\rho\Psi^2}{c^2}\Omega v_{\mathrm{ray}}\right)=0, \tag{311}$$

so the quantity within parentheses is constant along the ray tube.

Alternatively, the relation of Eq. (307) implies that the acoustic pressure is locally given to an equivalent approximation by

$$p=P(\mathbf{x})f(t-\tau), \tag{312}$$

where f is the derivative of F, and P is related to Ψ by

$$P=-\rho\Omega\Psi. \tag{313}$$

Then Eq. (311) requires that

$$B=AP^2v_{\mathrm{ray}}/\rho c^2\Omega \tag{314}$$

be constant along a ray tube, so that

$$\frac{dB}{dl}=0. \tag{315}$$

This quantity B is often referred to as the Blokhintzev invariant.

If there is no ambient flow then the constancy of B along a ray tube can be regarded as a statement that the average acoustic power flowing along the ray tube is constant, for a constant-frequency disturbance, or in general that the power appears constant along a trajectory moving with the ray velocity. The interpretation when there is an ambient flow is that the wave action is conserved (Hayes, 1968).

Geometrical acoustics, insofar as amplitude prediction is concerned, breaks down at caustics, which are surfaces along which the ray-tube areas vanish. Modified versions (Buchal and Keller, 1960; Ludwig, 1966), involving Airy functions and higher transcendental functions, of the geometrical acoustics theory have been derived which overcome this limitation.

14.5 Amplitude Variation in Rectilinear Propagation

A simple but important limiting case of geometrical acoustics is for propagation in homogeneous nonmoving media. The rays are then straight lines and perpendicular to wave fronts. A given wave front has two principal radii of curvature R_{I} and R_{II}. With the convention that these are positive when the wave front is concave in the corresponding direction of principal curvature, and negative when it is convex, the curvature radii increase by a distance l when the wave propagates through that distance. The ray-tube area is proportional to the product of the two curvature radii, so the acoustic pressure varies with distance l along the ray as

$$p=\left(\frac{R_{\mathrm{I}}R_{\mathrm{II}}}{(R_{\mathrm{I}}+l)(R_{\mathrm{II}}+l)}\right)^{1/2} f(t-[l/c]). \qquad (316)$$

Here R_{I} and R_{II} are the two principal radii of curvature when l is zero. The relations of Eqs. (176) and (188), for spherical and cylindrical spreading, are both limiting cases of this expression.

15. PARABOLIC-EQUATION APPROXIMATIONS

It is often useful in propagation analyses to replace the Helmholtz equation or its counterpart for inhomogeneous media by a partial differential equation that is first order, rather than second order, in the coordinate which corresponds to the primary direction of propagation. Approximations that achieve this are known generally as parabolic-equation approximations and stem from early work by Fock (1965) and Tappert (1977). Their chief advantage is that the resulting equation is no longer elliptic, but parabolic, so it is no longer necessary to check whether a radiation condition is satisfied at large propagation distances. This is especially convenient for numerical computation, for the resulting algorithms march out systematically in propagation distance, without having to "look far ahead" for the determination of the field at a subsequent point.

A simple example of a parabolic-equation approximation is that of two-dimensional propagation in a medium where the sound speed and density vary primarily with the coordinate y, but weakly with x. A wave of constant angular frequency ω is propagating primarily in the x direction but is affected by the y variations of the ambient medium, and by the presence of interfaces that are nearly parallel to the x direction. The propagation is consequently governed by the two-dimensional reduced wave equation,

$$\frac{\partial}{\partial x}\left(\frac{1}{\rho}\frac{\partial \hat{p}}{\partial x}\right)+\frac{\partial}{\partial y}\left(\frac{1}{\rho}\frac{\partial \hat{p}}{\partial y}\right)+\frac{\omega^2}{\rho c^2}\hat{p}=0, \qquad (317)$$

which can be obtained from Eq. (54) by use of the complex amplitude convention.

In the parabolic approximation, the exact solution $\hat{p}$ of Eq. (317), with a radiation condition imposed, is represented by

$$\hat{p}\,(x,y)=e^{i\chi(x)}F(x,y), \qquad (318)$$

where

$$\chi(x)=\int_0^x k_0\,dx. \qquad (319)$$

Here $k_0(x)$ is a judiciously chosen reference wave number (possibly dependent on the primary coordinate x). The crux of the method is that the function $F(x,y)$ is taken to satisfy an approximate partial differential equation (parabolic equation) that differs from what would result were Eq. (318) inserted in Eq. (317) with no discard of terms.

The underlying assumption is that the x dependence of the complex pressure amplitude is accounted for in a major part by the exponential factor, so the characteristic scale for variation of the amplitude factor F with the coordinate x is much greater than the reciprocal of k_0. Consequently, second derivatives of F with respect to x can be neglected in the derived differential equation. The y derivatives of the ambient variables are also neglected. The resulting approximate equation is given by

$$\frac{\partial}{\partial x}\left(\frac{k_0}{\rho}F\right)+\frac{k_0}{\rho}\frac{\partial F}{\partial x}=i\frac{\partial}{\partial y}\left(\frac{1}{\rho}\frac{\partial F}{\partial y}\right)+i\frac{k_0^2}{\rho}(n^2-1)F, \tag{320}$$

where n, the apparent index of refraction, abbreviates $k_0^{-1}\omega/c$. If the reference wave number k_0 is selected to be independent of x, then Eq. (320) reduces to

$$2k_0\frac{\partial F}{\partial x}=i\rho\frac{\partial}{\partial y}\left(\frac{1}{\rho}\frac{\partial F}{\partial y}\right)+ik_0^2(n^2-1)F. \tag{321}$$

Although the computational results must depend on the selection of the reference wave number k_0, results for various specific cases tend to indicate that the sensitivity to such a selection is not great.

GLOSSARY

Absorption: Process whereby some or all of the energy of a sound wave is transferred to the medium through which it is passing or to an object or surface on which it impinges.

Acoustic Boundary-Layer Theory: Theory which accounts for the typically rather abrupt changes in the nature of an acoustic field that take place near solid surfaces, so that the field outside the boundary layer more nearly conforms to idealizations appropriate to an unbounded medium.

Acoustic Compliance: Acoustic lumped parameter element analogous to capacitance in electric circuit theory. The volume velocity difference across a compliance element equals acoustic compliance times the time rate of change of pressure within the element.

Acoustic Fluid Velocity: Portion of the fluid velocity that is associated with the acoustic field.

Acoustic Impedance: Complex number pertaining to acoustic fields of constant frequency, equal to the ratio of a complex pressure amplitude to the complex amplitude of the volume velocity.

Acoustic Inertance: Acoustic lumped parameter element analogous to inductance in electric circuit theory. The pressure difference across an inertance element equals acoustic inertance times the time rate of change of volume velocity through the element.

Acoustic Intensity: Vector that describes the acoustic energy flowing per unit area and per unit time. The projection of this vector on any given direction gives the energy flow rate per unit area of surface normal to that direction.

Acoustic Mode: Acoustic part of a disturbance governed by the linearized equations of fluid mechanics. The three portions (acoustic mode, vorticity mode, and entropy mode) collectively superimpose to yield the overall perturbation disturbance.

Acoustic Pressure: The portion of the overall pressure that is associated with the acoustic field.

Acoustically Compact: An object or region of space that has dimensions that are small compared with a representative acoustic wavelength.

Acoustics: Science of sound, including its production, transmission, and effects, with sound defined as organized disturbances of a primarily mechanical nature that can propagate as waves or be regarded as a superposition of waves.

Affinity: Generic term for thermodynamic variables that appear as coefficients when the equation of state is written in differential form.

Ambient: Generic adjective referring to the medium and to its properties in the absence of an acoustic disturbance.

Ambient State: State of the medium or thermodynamic state in the absence of an acoustic disturbance.

Anechoic: Adjective used to imply that the subject does not reflect sound (without echos).

Angle of Incidence: Angle between the normal direction into a surface or interface and the direction of propagation of an incident acoustic wave.

Angle of Refraction: Angle between the normal direction into an interface and the direction of propagation of a transmitted acoustical wave. The phenomenon whereby the angle of refraction is not the same as the angle of incidence is referred to as refraction.

Angular Frequency: Frequency of a disturbance expressed in radians per second.

Associated Legendre Functions: Functions that are solutions of the associated Legendre equation (a second-order ordinary differential equation, which emerges when one solves the wave equation by separation of variables in spherical coordinates).

Asymptotic Limit: Simplified value or expression for a given function which holds to an increasingly higher accuracy when the independent variable becomes increasingly large.

Attenuation: Exponential decay in amplitude of a sound wave with increasing propagation distance as it passes through a medium.

Attenuation Coefficient: Exponential rate of decay with propagation distance of the amplitude of a sound wave, typically measured in nepers per meter.

Audible Frequencies: Those frequencies of acoustic disturbances that are ordinarily audible to the human ear, the range typically being taken as between 20 and 20 000 Hz.

Axial Wave Function: Factor describing the variation with distance along the waveguide axis when the wave function for a guided wave is expressed by the method of separation of variables.

Baffle: A relatively rigid extended surface surrounding an acoustic source.

Bending Modulus: Parameter describing the resistance (stiffness) of a beam or plate to bending. For a plate temporarily bent into a cylindrical shape, the bending modulus is the ratio of the curvature (reciprocal of the radius of curvature) to the bending torque per unit length.

Bernoulli–Euler Plate: Plate whose static and dynamic behavior adheres to the generalization of that model appropriate to a Bernoulli–Euler beam, including the assumption that points on any line normal to the central surface remain on a straight line, and that such lines are always normal to the surface throughout the motion.

Blokhintzev Invariant: Quantity proportional to the square of the magnitude of the wave amplitude which has the same value at every point along a geometrical acoustics ray path in inhomogeneous moving media.

Boundary Layer: Layer adjacent to a surface in which effects of viscosity or thermal conductivity are important.

Bulk Viscosity: The second viscosity coefficient, defined such that the nonequilibrium part of the average normal stress is bulk viscosity times the divergence of the fluid velocity.

Cauchy's Equation of Motion: The continuum generalization of Newton's second law which involves the distributed forces associated with stress gradient.

Caustic: Lines or surfaces along which adjacent rays intersect each other and along which the ray-tube area goes to zero.

Characteristic Frequency: Equivalent to representative frequency; whatever value of frequency is sufficient for making order-of-magnitude estimations and in assessing whether certain approximations are appropriate and whether certain terms in the governing equations are negligible. Equivalent to the reciprocal of characteristic time.

Characteristic Time: A value for a time interval that is representative of the source excitation or of the boundary conditions inherent to the determination of a wave disturbance; used in making order-of-magnitude estimations and in assessing whether certain approximations are appropriate and whether certain terms in the governing equations are negligible.

Classical Dissipation: The dissipation that is associated with viscosity and thermal conduction, in contrast to the dissipation associated with relaxation effects.

Compact: Within the context of acoustics, compact refers to an object or region that is small compared to a representative wavelength.

Compliance: Often used as an abbreviation for acoustic compliance.

Conduction Heat Transfer: The type of heat transfer that is governed by Fourier's law of heat conduction, as contrasted to heat transfer by convection or radiation.

Conservation of Mass Equation: Partial differential equation of fluid mechanics that accounts for the increase of density in a region in terms of the net flow of matter into that region.

d'Alembertian Operator: Operator equal to Laplacian minus product of square of reciprocal of sound speed with second time derivative.

Degenerate: Adjective used to characterize eigenvalues which are associated with more than one eigenfunction.

Diffusion Equation: Linear partial differential equation that accounts for diffusion of a concentrate, equating the time derivative of the concentrate to a constant times the Laplacian of the concentrate.

Dilatational: Having to do with the enlargement of the volume occupied by a definite amount of mass; opposite of compressional.

Dipole: Type of source formed by two closely spaced monopoles with equal and opposite source strengths.

Dispersion: Phenomenon whereby different frequencies propagate with different phase velocities.

Dispersion Relation: Any mathematical relation that involves angular frequency and wave number, such that one could, in principle, determine all of the possible values of angular frequency that correspond to a given wave number.

Dissipative Wave Equation: Any wave equation that accounts for dissipation of energy during wave propagation.

Eigenfunction: Nonzero solution of a homogeneous boundary-value problem with one adjustable parameter, when for almost all such values of that parameter the only solutions are identically zero; the values when a nonzero solution is possible are called the eigenvalues.

Eigenray: Ray connecting two specified points which satisfies the ordinary differential equations of geometrical acoustics.

Eikonal Equation: Nonlinear first-order partial differential equation arising in the theory of geometrical acoustics. Its solution is typically given parametrically in terms of ray paths.

Euler Equation: Counterpart of Newton's second law for an ideal fluid without viscosity, in which acceleration is caused by pressure gradients.

Eulerian Description: Scheme for a mathematical description of material motion in a continuum in which field quantities are expressed as functions of actual position coordinates rather than of the coordinates of initial positions of material particles.

Evanescent: Adjective referring to wave phenomenon that does not propagate but which dies out exponentially with distance, without necessarily being accompanied by any dissipation of energy.

Fermat's Principle: Actual path followed by an eigenray must be such that the travel time along that path is stationary (does not change to first order) to all small virtual changes along that path, originally stated (with less universality) as a principle of least time and sometimes stated as the principle that the travel time must be an extremal (maximum or minimum).

Flow Resistance: A counterpart of Ohm's law for flow through a porous slab is that the pressure difference across the slab is a constant (the flow resistance) times the apparent velocity of flow through the slab. The latter is the mass flowing through the slab per unit area and per unit time, divided by the ambient density of the fluid external to the slab.

Flow Resistivity: Flow resistance per unit thickness of the slab.

Fourier's Law of Heat Conduction: The heat flux vector is taken to be opposite in direction and proportional to the gradient of the temperature.

Geometrical Acoustics: Approximate theory of acoustics based on the premise that disturbances which locally resemble plane traveling waves propagate locally according to the relations appropriate to such waves; the theory leads to the concept of rays or ray paths which are, in general, curved and along which the dominant features in the disturbance travel.

Geometrical Theory of Diffraction: Generalization of geometrical optics or geometrical acoustics based on rigorous asymptotic analysis of wave theory, which includes as a basic tenet that waves diffracted by edges and curved objects propagate away from such edges or objects along rays, which are completely analogous to the rays along which direct and reflected waves propagate.

Group Velocity: Velocity at which energy is transported by a wave of nearly constant frequency.

Helmholtz Corollary: Alternatively known as the Helmholtz–Kirchhoff corollary to the wave equation, this applies to a constant-frequency disturbance and expresses the acoustic pressure (or an analogous field variable) at a point off the surface as an integral over the surface involving the field and its normal derivative, as well as a Green's function. The integral is often referred to as the Helmholtz integral.

Helmholtz Equation: Form of the wave equation which holds for disturbances of constant frequency, with the dependent variable being the spatially dependent complex amplitude of a field quantity such as acoustic pressure.

Impedance Tube: A device consisting of a long cylindrical tube with a sound source at one end and a sample at the other end, which is used to measure the surface impedance of the sample.

Inertance: A lumped parameter quantity which plays a role analogous to mass in the equations describing a system's dynamics.

Intensity: The energy flux vector of a wave disturbance, not to be confused with the amplitude or the energy density of the disturbance. The dot product of a given unit vector with the intensity vector yields the energy flow per unit area and time across a surface normal to that unit vector.

Kirchhoff Corollary: A corollary of the linear acoustic equations which expresses the acoustic pressure field off of a surface as an integral over the surface involving the field, its time derivative, and its normal derivative, as well as a Green's function, usually referred to as the Helmholtz–Kirchhoff corollary.

Lamé Constants: Coefficients in the linear relations between the stress and strain tensor components, which appear in theory of elasticity of isotropic materials.

Law of Mirrors: The rule that for reflection of waves at a smooth surface, the angle of incidence equals the angle of reflection.

Mass Law: Approximate rule that the fraction of incident energy transmitted through a wall or panel is inversely proportional to the square of the mass per unit area in the wall.

Medium: Abstract term referring to whatever a wave is propagating through, characterized by the nature and coordinate dependences of the coefficients in those equations that govern the propagation.

Momentum Balance Equation: Equation related to Newton's second law for a continuum which accounts for changes in momentum in terms of forces.

Monopole: Source of sound whose radiation characteristics are the same as those of an omnidirectional point source.

Monopole Strength: Descriptor of a constant-frequency source's ability to produce monopole radiation, equal to the coefficient of $r^{-1}e^{ikr}$ in the expression for the complex amplitude of the radiated pressure.

Muffler: Device for abating the noise accompanying the flow of gas through a pipe.

Multipath: Phenomena whereby more than one ray path can reach a given reception point.

Multipole: Adjective referring to a type of mathematical expansion of radiated sound fields from localized sources, the various terms being monopole, dipole, quadrupole, etc., terms.

Navier–Stokes Equation: Generalization of Newton's second law for a viscous fluid.

Navier–Stokes–Fourier Equations: Set of fluid dynamic equations that govern a compressible fluid with viscosity and thermal conductivity.

Newtonian Fluid: A fluid within which the stress tensor components are linearly dependent on the rate of strain tensor components.

Parseval's Theorem: The version for a Fourier series expresses the time average of the square of the function in terms of a sum over the squares of its Fourier coefficients; that for transients expresses the time integral of the square of the function in terms of the frequency integral of the square of the Fourier transform.

Phase Velocity: The velocity with which position would have to change such that the phase remains constant; if $\phi(x,t)$ is the phase, then the phase velocity is found by setting $x=x[t]$ and then setting the total derivative with respect to time of the phase equal to zero.

Plane Wave: Wave which depends on distance in only one direction, so that the field variables are constant along planes perpendicular to that direction.

Poisson's Ratio: Constant in the theory of elasticity which corresponds to the negative of the ratio of transverse strain to axial strain when a thin elastic rod undergoes axial stretching, with a simultaneous shrinking in cross-sectional area.

Pressure Release Surface: Surface on which the acoustic pressure vanishes.

Quadrupole: Source of sound whose radiation properties are similar to those of four closely spaced point sources with source strengths such that (i) their sum is identically zero, and (ii) their net dipole moment is identically zero.

Radiation Condition: A condition which a constant-frequency solution of a wave equation must satisfy at large distances from the source in order that the radiated field be the aftermath of the source having been turned on at some time in the past.

Radiation Impedance: Specific radiation impedance is the ratio of the complex amplitude of the acoustic pressure to the complex amplitude of the fluid velocity in the outward normal direction on the surface of a vibrating body.

Rate-of-Shear Tensor: Rate-of-strain tensor minus a multiple of the unit tensor, with the coefficient chosen such that the trace is identically zero; the tensor vanishes for rates of strain occurring during uniform compression or dilatation.

Rate-of-Strain Tensor: In the linear approximation, this is the same as the time derivative of the strain tensor.

Ray: Path in space, or a trajectory in space-time, along which characteristic features of an acoustic field propagate.

Ray Tube: Hypothetical surface enclosing a bundle of ray paths, such that if any given ray paths touches the surface at one point, then it lies entirely in the surface.

Reactive Muffler: Muffler whose primary mechanism for attenuating sound involves the frequency-dependent reflection of sound back toward the source.

Reciprocity: Principle in wave theory which asserts, with various conditions, that the amplitude detected by a given receiver due to a disturbance generated by a given source will be the same if the positions of the source and receiver are interchanged.

Reference Wave Number: Parameter that enters into the formulation of the parabolic equation approximation.

Reflection Coefficient: A complex number that describes the ratio of the complex amplitude of a reflected wave to the complex amplitude of the corresponding incident wave.

Relaxation Processes: Processes whereby a thermodynamic system evolves toward thermodynamic equilibrium.

Salinity: Measure of the total salt content of sea water, expressed as parts per thousand, or grams of dissolved salts per kilogram of water.

Spherical Spreading: The characteristic decay of wave amplitudes as $1/r$ when propagating out from a localized source in an unbounded medium.

Stokes's Time Derivative Operator: The time derivative that would be measured by someone moving with the fluid.

Strain Tensor: Tensor which describes the state of deformation within a continuum material. The diagonal elements correspond to stretching strains (elongation per unit length) and the off-diagonal elements correspond to shear strains.

Stress Tensor: Tensor which describes the distributed internal forces within a continuum material. The tensor product of the stress tensor with a given unit vector yields the force vector per unit area exerted by adjacent material on an internal surface with unit normal in that direction.

Thermal Conductivity: Material constant that appears in Fourier's law of heat conduction: heat flux vector is the negative product of the thermal conductivity with the temperature gradient.

Thermodynamic Coefficient: Thermodynamic quantity described by an expression involving one or more partial derivatives of one thermodynamic state variable with respect to another, with the evaluation being at the ambient thermodynamic state. In linear acoustics, the thermodynamic coefficients are independent of field amplitudes, but may vary with position.

Thermodynamic Field Quantity: Quantities such as pressure, density, specific entropy, and temperature which are associated with the thermodynamic state of the material, but which are perturbed during an acoustic disturbance, with the perturbations varying with position and time.

Trace Velocity: The apparent speed at which an acoustic disturbance moves along a specified line that is not necessarily aligned with the direction of propagation. If the disturbance is associated with a plane wave, then the trace velocity is the sound speed divided by the cosine of the angle between the direction of propagation and the direction of the line.

Trace Velocity Matching Principle: If a primary portion of a disturbance has a single fixed trace velocity along a surface or interface, then all of the other portions of the overall wave disturbance that can be regarded as caused by the primary disturbance must have the same trace velocity.

Unit Area Acoustic Impedance: Same as specific acoustic impedance, equal to the ratio of the complex amplitude of the acoustic pressure to the complex amplitude of the acoustic particle velocity component pointing into the surface.

Volume Acceleration: Time derivative of the volume velocity.

Volume Expansion Coefficient: Thermal expansion coefficient, equal to the change in volume per unit volume per unit change in temperature at constant pressure.

Volume Velocity: Apparent volume of fluid which crosses a specified surface per unit time, equal to the rate of mass flow across the surface divided by the ambient density of the fluid.

Vorticity: Curl of the fluid velocity.

Vorticity Mode: That part of a disturbance governed by the linearized equations of fluid mechanics which is characterized by a nonzero vorticity. The three portions (acoustic mode, vorticity mode, and entropy mode) collectively superimpose to yield the overall perturbation disturbance.

Wave: Disturbance moving through a medium, typically yielding little or no net transport of mass, but carrying energy.

Wave Action: For an oscillating one-degree-of-freedom system undergoing adiabatic changes that affect its oscillation frequency, induced by gradual changes of the environment or of the applied external forces, it is often true that the quantity which most nearly remains constant is the energy divided by the oscillation frequency, rather than the energy itself. Because this is equivalent to an integral of energy over a time period, which in turn is analogous to the integral of a Lagrangian over time in the principle of stationary action, this conserved quantity is called the system action. For a propagating wave, the wave action density is the energy density divided by the local frequency that would be perceived by someone moving with the ambient flow.

Wave Equation: In the narrow sense this is the ordinary wave equation that holds for the most idealized case and whose one-dimensional version appears in the elementary theory of the vibrating string; the term is also often used to refer to any equation that governs wave phenomena.

Wave Front: Moving surface associated with a propagating wave; the surface is one along which the wave's phase is everywhere the same or one along which a distinct waveform feature is being simultaneously received.

Waveguide: Device or circumstance that causes acoustic propagation to be confined to a restricted range of propagation directions.

Wave Number: Rate of change of phase with propagation distance, equal to 2π divided by the wavelength, and having units of radians per meter.

Wave Slowness: Reciprocal of phase velocity, equal to wave number divided by angular frequency.

Wave Slowness Vector: Wave-number vector divided by angular frequency.

List of Works Cited

Abramowitz, M., Stegun, I. A. (1965), *Handbook of Mathematical Functions*, New York: Dover.

Achenbach, J. D. (1975), *Wave Propagation in Elastic Solids*, Amsterdam: North-Holland.

Bass, H. E., Sutherland, L. C., Zuckerwar, A. J. (1990), *J. Acoust. Soc. Am.* **88**, 2019–2021.

Batchelor, G. K. (1967), *An Introduction to Fluid Dynamics*, Cambridge: Cambridge University.

Bender, E. K., Bremmer, A. J. (1975), *J. Acoust. Soc. Am.* **58**, 22–30.

Beranek, L. L. (1947), *J. Acoust. Soc. Am.* **19**, 556–568.

Beranek, L. L. (1971), *Noise and Vibration Control*, New York: McGraw-Hill.

Beranek, L. L. (1972), in: D. E. Gray (Ed.), *American Institute of Physics Handbook*, New York: McGraw-Hill, Sec. 3a.

Bergmann, P. G. (1946), *J. Acoust. Soc. Am.* **17**, 329–333.

Beyer, R. T., Letcher, S. V. (1969), *Physical Ultrasonics*, New York: Academic.

Birkhoff, G. (1973), *A Source Book in Classical Analysis*, Cambridge: Harvard University.

Blokhintzev, D. I. (1946), *J. Acoust. Soc. Am.* **18**, 322–328.

Bouwkamp, C. J. (1946), *Philips Res. Rep.* **1**, 251–277.

Buchal, R. B., Keller, J. B. (1960), *Commun. Pure Appl. Math.* **13**, 85–144.

Burton, A. J., Miller, G. F. (1971), *Proc. R. Soc. London, Ser. A* **323**, 201–210.

Chu, B. T. (1955), NACA Technical Note 3411, Washington: National Advisory Committee for Aeronautics.

Copley, L. G. (1968), *J. Acoust. Soc. Am.* **44**, 28–32.

Courant, R. (1937), *Differential and Integral Calculus*, New York: Wiley-Interscience, Vols. 1 and 2.

Crandall, S. H., Dahl, N. C., Lardner, T. J. (1978), *An Introduction to the Mechanics of Solids*, New York: McGraw-Hill.

Cremer, L. (1942), *Akust. Zh.* **7**, 81–104.

Cremer, L. (1948), *Arch. Elekt. Uebertrag.* **2**, 136–139.

Davies, P. O. A. L. (1964), *J. Sound Vib.* **1**, 185–201.

Embleton, T. F. W. (1971), in: L. L. Beranek (Ed.), *Noise and Vibration Control*, New York: McGraw-Hill, pp. 362–405.

Evans, L. B., Bass, H. E., Sutherland, L. C. (1972), *J. Acoust. Soc. Am.* **51**, 1565–1575.

Fisher, F. H., Simmons, V. P. (1977), *J. Acoust. Soc. Am.* **62**, 558–564.

Fletcher, H. (1940), *Rev. Mod. Phys.* **12**, 47.

Fock, V. A. (1965), *Electromagnetic Diffraction and Propagation Problems*, London: Pergamon.

Fourier, J. (1822, 1878, 1955), *Analytical Theory of Heat*, New York: Dover.

Fung, Y. C. (1965), *Foundations of Solid Mechanics*, Englewood Cliffs: Prentice-Hall.

Green, G. (1838), *Trans. Cambridge Philos. Soc.* **6**, 403–412, reprinted in: Lindsay (1972).

Greenspan, M. (1959), *J. Acoust. Soc. Am.* **31**, 155–160.

Hall, W. M. (1939), *J. Acoust. Soc. Am.* **11**, 140–146.

Hayes, W. D. (1968), *Phys. Fluids* **11**, 1654–1656.

Hartig, H. E., Swanson, C. E. (1938), *Phys. Rev.* **54**, 618–626.

Heller, G. S. (1953), *J. Acoust. Soc. Am.* **25**, 950–951.

Helmholtz, H. (1860), *J. Reine Angew. Math.* **57**, 1–72, partial translation in: Birkhoff (1973).

Henry, P. S. H. (1931), *Proc. Phys. Soc. London* **43**, 340–361.

Herzfeld, K. F., Rice, F. O. (1928), *Phys. Rev.* **31**, 691–695, reprinted in: Lindsay (1974).

Hilsenrath, J. (1960), *Tables of Thermodynamic and Transport Properties of Air*, Oxford: Pergamon.

Horne, R. A. (1969), *Marine Chemistry*, New York: Wiley-Interscience.

Hunt, F. V. (1978), *Origins of Acoustics*, New Haven: Yale University.

Keller, J. B. (1954), *J. Appl. Phys.* **25**, 938–947.

Keller, J. B. (1958), in: L. M. Graves (Ed.), *Calculus of Variations and its Applications*, New York: McGraw-Hill, pp. 27–52.

Kirchhoff, G. R. (1868), *Ann. Phys. Chem.* **134**, 177–193; translation in: Lindsay (1974).

Kirchhoff, G. R. (1877), *Mechanik*, Leipzig: Teubner.

Kirchhoff, G. R. (1883), *Ann. Phys. Chem.* **18**, 663–695; partial translation in: Birkhoff (1972).

Kovasznay, L. S. G. (1953), *J. Aeronaut. Sci.* **20**, 657–674, 682.

Lamb, H. (1879, 1932, 1945), *Hydrodynamics*, New York: Dover.

Lamb, H. (1925, 1960), *Dynamical Theory of Sound*, New York: Dover.

Lambert, R. F. (1951), *J. Acoust. Soc. Am.* **23**, 480–481.

Landau, L. D., Lifshitz, E. M. (1959), *Fluid Mechanics*, New York: Pergamon.

Landau, L. D., Lifshitz, E. M. (1970), *Theory of Elasticity*, New York: Pergamon.

Landau, L. D., Lifshitz, E. M., Pitaevskii, L. P. (1980), *Statistical Physics*, Pt. 1, New York: Pergamon.

Lighthill, J. (1978), *Waves in Fluids*, Cambridge: Cambridge University.

Lindsay, R. B. (1960), *Mechanical Radiation*, New York: McGraw-Hill.

Lindsay, R. B. (1972), *Acoustics: Historical and Philosophical Development*, Stroudsburg, PA: Dowden, Hutchinson and Ross.

Lindsay, R. B. (1974), *Physical Acoustics*, Stroudsburg: Dowden, Hutchinson, and Ross.

Love, A. E. H. (1892, 1927, 1944), *The Mathematical Theory of Elasticity*, New York: Dover.

Ludwig, D. (1966), *Commun. Pure Appl. Math.* **19**, 215–250.

Markham, J. J., Beyer, R. T., Lindsay, R. B. (1951), *Rev. Mod. Phys.* **23**, 353–411.

Mason, W. P. (1927), *Bell Syst. Tech. J.* **6**, 258–294.

Mason, W. P. (1928), *Phys. Rev.* **31**, 283–295.

Maue, A.-W. (1949), *Z. Phys.* **126**, 601–618.

Meixner, J. (1943), *Ann. Phys.* **43**, 470–487.

Meixner, J. (1952), *Acustica* **2**, 101–109.

Meixner, J. (1952), *Z. Phys.* **131**, 456–469.

Milne, E. A. (1921), *Philos. Mag.* **42**, 96–114.

Morse, P. M., (1936, 1948, 1976), *Vibration and Sound*, New York: Acoustical Society of America.

Morse, P. M., Feshbach, H. (1953), *Methods of Theoretical Physics*, Vols. 1 and 2. New York: McGraw-Hill.

Morse, P. M., Ingard, K. U. (1963), *Theoretical Acoustics*, New York: McGraw-Hill.

Paris, E. T. (1927), *Proc. Phys. Soc. London*, **39**, 269–295.

Pierce, A. D. (1978), *J. Sound Vib.* **58**, 189–200.

Pierce, A. D. (1981, 1989), *Acoustics: An Introduction to Its Physical Principles and Applications*, New York: Acoustical Society of America.

Pierce, A. D. (1990), *J. Acoust. Soc. Am.* **87**, 2292–2299.

Pinkerton, J. M. M. (1947), *Nature (London)* **160**, 128–129.

Lord Rayleigh (1870), *Philos. Trans. R. Soc. London* **161**, 77–118.

Lord Rayleigh (1878, 1896, 1945), *The Theory of Sound*, New York: Dover, Vols. 1 and 2.

Schenck, H. A. (1968), *J. Acoust. Soc. Am.* **44**, 41–58.

Shapiro, A. H. (1953), *The Dynamics and Thermodynamics of Compressible Fluid Flow*, New York: Ronald.

Shapiro, A. H. (1961), *Shape and Flow: The Fluid Dynamics of Drag*, Garden City: Doubleday, pp. 59–63.

Sommerfeld, A., Runge, J. (1911), *Ann. Phys.* **35**, 277–298.

Stewart, G. W. (1922), *Phys. Rev.* **20**, 528–551.

Stokes, G. G. (1845), *Trans. Cambridge Philos. Soc.* **8**, 75–102; partially reprinted in: Lindsay (1972).

Stokes, G. G. (1868), *Philos. Trans. R. Soc. London* **158**, 447–463; partially reprinted in: Lindsay (1974).

Sutherland, W. (1893), *Philos. Mag.* **36**, 507–531.

Tappert, F. D. (1977), in: J. B. Keller and J. S. Papadakis (Eds.), *Wave Propagation and Underwater Acoustics*, Berlin: Springer-Verlag, Chap. 5.

Thompson, P. A. (1972), *Compressible-Fluid Dynamics*, New York: McGraw-Hill.

Truesdell, C. (1955), *Leonhardi Euleri Opera Omnia*, Ser. 2, 13, Lausanne: Orell Füssli, pp. XIX–LXXII.

Truesdell, C., Toupin, R. (1960), in: S. Flügge (Ed.), *Handbook of Physics*, Berlin: Springer-Verlag, Vol. III/1.

Truesdell, C. (1966), *The Mechanical Foundations of Elasticity and Fluids Mechanics*, New York: Gordon and Breach.

Ugincius, P. (1972), *J. Acoust. Soc. Am.* **51**, 1759–1763.

Westervelt, P. J., Larson, R. S. (1973), *J. Acoust. Soc. Am.* **54**, 121–122.

Wilson, W. D. (1959), *J. Acoust. Soc. Am.* **31**, 1067–1072.

Wilson, W. D. (1960), *J. Acoust. Soc. Am.* **32**, 641–644, 1357.

Wu, T. Y. (1956), *J. Math. Phys.* (Cambridge, MA) **35**, 13–27.

Further Reading

Achenbach, J. D. (1973), *Wave Propagation in Elastic Solids*, Amsterdam: North-Holland.

Auld, B. A. (1973), *Acoustic Fields and Waves in Solids*, New York: Wiley-Interscience, Vols. 1 and 2.

Baker, B. B., Copson, E. T. (1950), *The Mathematical Theory of Huygens' Principle*, London: Oxford University.

Baxter, S. M., Morfey, C. L. (1986), *Angular Distribution Analysis in Acoustics*, New York: Springer-Verlag.

Beranek, L. L. (1954), *Acoustics*, New York: Acoustical Society of America.

Beranek, L. L. (1989), *Acoustic Measurements*, New York: Acoustical Society of America.

Beranek, L. L. (1960), *Noise Reduction*, New York: McGraw-Hill.

Beyer, R. T., Letcher, S. V. (1969), *Physical Ultrasonics*, New York: Academic.

Beyer, R. T. (1974), *Nonlinear Acoustics*, Washington: Naval Sea Systems Command.

Blackstock, D. T. (1972), in: D. E. Gray (Ed.), *American Institute of Physics Handbook*, New York: McGraw-Hill, Sec. 3n.

Blake, W. K. (1986), *Mechanics of Flow-Induced Sound and Vibration*, New York: Academic.

Bleistein, N. (1984), *Mathematical Methods for Wave Phenomena*, Orlando: Academic.

Bowman, J. J., Senior, T. B. A., Uslenghi, P. L. E. (1969), *Electromagnetic and Acoustic Scattering by Simple Shapes*, Amsterdam: North-Holland.

Boyles, C. A. (1980), *Acoustic Waveguides*, New York: Wiley.

Bragg, W. (1920, 1968), *The World of Sound*, New York: Dover.

Brekhovskikh, L. M. (1960), *Waves in Layered Media*, New York: Academic.

Brekhovskikh, L. M., Lysanov, Yu. (1982), *Fundamentals of Ocean Acoustics*, Berlin: Springer-Verlag.

Clay, C. S., Medwin, H. (1977), *Acoustical Oceanography: Principles and Applications*, New York: Wiley.

Coulson, C. A. (1941, 1955), *Waves*, Edinburgh: Oliver and Boyd.

Courant, R., Hilbert, D. (1937, 1953, 1962), *Methods of Mathematical Physics*, New York: Interscience, Vols. 1 and 2.

Cremer, L., Heckl, M. (1973), *Structure-borne Sound*, New York: Springer-Verlag.

Crocker, M. J., Price, A. J. (1975), *Noise and Noise Control*, Cleveland: Chemical Rubber.

Dowling, A. P., ffowcs Williams, J. E. (1983), *Sound and Sources of Sound*, Chichester: Ellis Horwood.

Fahy, F. (1985), *Sound and Structural Vibration: Radiation, Transmission and Response*, London: Academic.

Felsen, L. B., Marcuvitz, N. (1973), *Radiation and Scattering of Waves*, Englewood Cliffs: Prentice-Hall.

Friedlander, F. G. (1958), *Sound Pulses*, Cambridge: Cambridge University.

Garabedian, P. R. (1986), *Partial Differential Equations*, New York: Chelsea.

Goldstein, M. E., (1976), *Aeroacoustics*, New York: McGraw-Hill.

Gossard, E. E., and Hooke, W. H., *Waves in the Atmosphere*, Amsterdam: Elsevier.

Graff, K. F. (1975), *Wave Motion in Elastic Solids*, Columbus: Ohio University.

Hansen, R. C. (Ed.) (1981), *Geometric Theory of Diffraction*, New York: IEEE.

Harris, C. M. (1979), *Handbook of Noise Control*, New York: McGraw-Hill.

Helmholtz, H. (1885, 1954), *On The Sensation of Tone*, New York: Dover.

Hunt, F. V. (1954), *Electroacoustics*, New York: Acoustical Society of America.

Hunt, F. V. (1972), in: D. E. Gray (Ed.), *American Institute of Physics Handbook*, New York: McGraw-Hill, Sec. 3c.

Ingard, K. U. (1958), in: E. U. Condon and H. Odishaw (Eds.), *Handbook of Physics*, New York: McGraw-Hill, Chap. 8.

Junger, M. C., Feit, D. (1986), *Sound, Structures, and Their Interaction*, Cambridge, MA, MIT.

Kinsler, L. E., Frey, A. R. (1962), *Fundamentals of Acoustics*, New York: Wiley.

Knudsen, V. O., Harris, C. M. (1950, 1978), *Acoustical Designing in Architecture*, New York: Acoustical Society of America.

Kuttruff, H. (1973), *Room Acoustics*, London: Applied Science.

Lamb, H. (1879, 1932, 1945), *Hydrodynamics*, New York: Dover.

Lamb, H. (1925, 1960), *The Dynamical Theory of Sound*, New York: Dover.

Lee, D., McDaniel, S. T. (1988), *Ocean Acoustic Propagation by Finite Difference Methods*, Oxford: Pergamon.

Levine, H. (1978), *Unidirectional Wave Motion*, Amsterdam: North-Holland.

Lighthill, M. J. (1978), *Waves in Fluids*, Cambridge: Cambridge University.

Lindsay, R. B. (1960), *Mechanical Radiation*, New York: McGraw-Hill.

Lindsay, R. B. (1972), *Acoustics: Historical and Philosophical Development*, Stroudsburg: Dowden, Hutchinson and Ross.

Lindsay, R. B. (1974), *Physical Acoustics*, Stroudsburg: Dowden, Hutchinson and Ross.

Mason, W. P., Thurston, R. N. (Eds.) (1964–1988), *Physical Acoustics, Principles and Methods*, New York: Academic, Vols. 1–22.

Meyer, E., Neumann, E.-G. (1972), *Physical and Applied Acoustics*, New York: Academic.

Miklowitz, J. (1978), *The Theory of Elastic Waves and Waveguides*, Amsterdam: North-Holland.

Morse, P. M. (1936, 1976), *Vibration and Sound*, New York: Acoustical Society of America.

Morse, P. M., Feshbach, H. (1953), *Methods of Theoretical Physics*, New York: McGraw-Hill.

Morse, P. M., Ingard, K. U. (1961), "Linear Acoustic Theory," in *Handbuch der Physik: Akustik I*, S. Flügge (Ed.), Vol. 11/1. Berlin: Springer-Verlag.

Morse, P. M., Ingard, K. U. (1968), *Theoretical Acoustics*, New York: McGraw-Hill.

Officer, C. B. (1958), *Introduction to the Theory of Sound Transmission*, New York: McGraw-Hill.

Pierce, A. D. (1981, 1989), *Acoustics: An Introduction to Its Physical Principles and Applications*, New York: Acoustical Society of America.

Lord Rayleigh (1878, 1893, 1945), *The Theory of Sound*, New York: Dover, Vols. 1 and 2.

Lord Rayleigh (1964), *Scientific Papers*, New York: Dover, Vols. 1–6.

Redwood, M. (1960), *Mechanical Waveguides*, New York: Pergamon.

Rossing, T. D. (1982), *The Science of Sound*, Reading: Addison-Wesley.

Rudenko, O. V., and Soluyan, S. I. (1977), *Theoretical Foundations of Nonlinear Acoustics*, New York: Plenum.

Skudrzyk, E. J. (1971), *The Foundations of Acoustics: Basic Mathematics and Basic Acoustics*, New York: Springer-Verlag.

Sommerfeld, A. (1949), *Partial Differential Equations in Physics*, New York: Academic.

Stephens, R. W. B., and Bate, A. E. (1966), *Acoustics and Vibrational Physics*, London: Arnold.

Stump, F. B. (1980), *Analytical Acoustics*, Ann Arbor: Ann Arbor Science.

Temkin, S. (1981), *Elements of Acoustics*, New York: Wiley.

Tolstoy, I. (1973), *Wave Propagation*, New York, McGraw-Hill.

Towne, D. H. (1967), *Wave Phenomena*, Reading, MA: Addison-Wesley.

Urick, R. J. (1983), *Principles of Underwater Sound*, New York: McGraw-Hill.

Whitham, G. B. (1974), *Linear and Nonlinear Waves*, New York: Wiley-Interscience.

Wood, A. (1960), *Acoustics*, New York: Dover.

ACOUSTICS, NONLINEAR

ROBERT T. BEYER, *Department of Physics, Brown University*

Introduction 183
1. **The Equation of Motion** 184
2. **Nonlinear Propagation in Fluids** 185
2.1 Solution of Basic Equation 185
2.2 Burgers' Equation 186
2.3 Spherical and Cylindrical Coordinates 188
3. **Shock Waves and Aeroacoustics** 188
3.1 Rankine–Hugoniot Equation 188
3.2 Shock-Wave Thickness 189
3.3 N Waves 190
3.4 The Lighthill Equations 191
3.5 Monopole, Dipole, and Quadrupole Sources 193
4. **Radiation Pressure, Streaming, and Cavitation** 194
4.1 Radiation Pressure 194
4.2 Acoustic Streaming 195
4.3 Acoustically Induced Cavitation 195
5. **Nonlinear Interaction of Sound Waves** 197
5.1 Two Intersecting Sound Beams in a Fluid 197
5.1.1 Nondispersive and Dispersive Media 199
5.2 Two Collinear Sound Beams 199
5.3 Parametric Arrays 199
5.3.1 Far-Field Transmission 199
5.3.2 Far-Field Reception 200
6. **Nonlinear Effects in Solids** 201
6.1 The Scope of Nonlinear Behavior 201
6.2 Third-Order Elastic Constants 201
6.3 Interaction of Sound with Sound 204
6.4 Nonlinear Surface Waves 204
7. **Recent Applications** 204
7.1 General View 204
7.2 Radiation Pressure 204
7.3 Parametric Array 205
7.4 Lithotripsy 206
7.5 *B/A* in Biological Tissues 207
7.6 Surface Acoustic Waves (SAW) 207
Glossary 209
Works Cited 211
Further Reading 212

INTRODUCTION

Nonlinear acoustics is the analysis of mechanical waves and vibrations in a medium where the governing differential equations cannot be assumed to be linear. It is a rather backward definition of a field, defining it only in the negative, but this is a fact of historical experience.

Since the approximation of linearity extends to the entire field of acoustics, it follows that every branch has its nonlinear supplement. In addition, however, there are phenomena that are inherently nonlinear and form branches of acoustics in their own right (for example, shock waves and radiation pressure).

Ordinary linear acoustics presumes that the variations of local pressure, density, and temperature produced by a sound wave are negligible in comparison with the equilibrium values of these same quantities. In that sense, linear acoustics is said to be an infinitesimal theory. In the case of the displacement velocity of a particle in a continuous medium, subject to the passage of a sound wave, we cannot make the same comparison, since the basic treatment presupposes a con-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

tinuous medium that is at rest in equilibrium. However, we can compare the particle velocity with the velocity of propagation of sound in the medium.

This ratio is called the acoustic Mach number M. If the amplitude of the displacement velocity is u_0 and the sound velocity is c_0, then $M=u_0/c_0$. Also, since the excess pressure amplitude p_a is related to the displacement velocity amplitude by the relation $p_a=\rho_0 c_0 u_0$, where ρ_0 is the mean density of the medium, then $M=p_a/\rho_0 c_0^2$. This Mach number must be negligible in value if linear acoustics is to prevail. As an example, we consider the sound level of a rock concert, which, in decibel notation, is 120 dB *re*: 20 μm Pa. That is, $120=20\log[p_a/(20\times 10^{-6})]$; hence $p_a=20$ Pa. Hence $M=20/(1.4)(343)^2=1.2\times 10^{-4}$ (density of air $\rho_0=1.4$ kg/m^3, c_0 for air$=343$ m/s).

However, as a sound wave passes through such a medium, the nonlinear effects are cumulative. Thus it is meaningful to consider that the linear approximation will fail when the product of the number N of wavelengths traversed by the beam and the acoustic Mach number M becomes an appreciable fraction of unity. In the example considered, the wavelength of the sound will, at a frequency of 10 kHz, be equal to $(343\ \mathrm{m\ s^{-1}})/(10\times 10^3\ \mathrm{s^{-1}})=0.0343$ m. For $MN=0.1$, we obtain a value of $N=0.1/M=0.1/1.2\times 10^{-4}=830$. This corresponds to a distance of $830\times 0.0343=28.5$ m. If the frequency were 100 kHz, the wavelength would decrease by a factor of 10, and the distance required would shrink to 2.85 m. Such a distance would be even smaller in the case of liquids, so that the nonlinearity is almost always present in sound propagation to some degree.

Historical references to what we would now call nonlinear acoustics abound. In the Aeolian harp, an instrument that goes back to Biblical times, a flow of wind past a string or wire can set that string into oscillation. Today, we know these oscillations to be due to the growth and shedding of vortices of air downwind, with the creation of an alternating force on the string. A similar shedding of vortices in the flow of air past a narrow wedge is the source of so-called edge tones and is responsible for the operation of some musical instruments. In the nineteenth century, it was observed that there is a one-way flow of air near a sound source. Such acoustic streaming is another nonlinear sound effect. Cavitation, or the creation of holes (or gas bubbles) in a liquid due to the passage of intense sound waves, is another example of the nonlinear behavior of acoustic media. And finally, the creation of turbulent flow by jet engines, for example, is responsible for the emission of the strong sound waves that are created in the wake of jet aircraft.

1. THE EQUATION OF MOTION

Ordinary sound and vibration are treated by the use of linear differential equations (see ACOUSTICS, LINEAR). For example, for a one-dimensional vibrating string, and with neglect of dissipation, the equation is

$$c^2\frac{\partial^2 y}{\partial x^2}=\frac{\partial^2 y}{\partial t^2}, \tag{1}$$

where y is the transverse displacement of the string at the position x, and c is the speed of propagation of the wave along the string.

With such an equation, the general solution is of the form

$$y=A\cos(\omega t-kx)+B\sin(\omega t-kx), \tag{2}$$

where A and B are arbitrary constants. Such an equation and its solutions have the property of superposition; i.e., the linear combination of any two solutions of Eq. (1) is also a solution of that equation. Thus, if we analyze any initial disturbance into its Fourier components, the amplitude of every frequency component initially present in the wave will remain constant in time. The presence of dissipation can be taken into account by the addition of a term $R(\partial y/\partial t)$ on the left-hand side of Eq. (1). If, in addition, a forcing term $F(x,t)$ is introduced, the resultant equation takes the form

$$c^2\frac{\partial^2 y}{\partial x^2}+R\frac{\partial y}{\partial t}-\frac{\partial^2 y}{\partial t^2}=F(x,t). \tag{3}$$

If the forcing term is decomposed by Fourier analysis into its harmonic components, each harmonic term can be treated separately, and the complete solution will be the sum of all the solutions of the individual forcing terms.

We look first at the one-dimensional equation governing the propagation of plane waves in a nondissipative ideal gas:

$$\frac{\partial^2\xi}{\partial t^2}=\frac{c_0^2}{(1+\partial\xi/\partial x)^{\gamma+1}}\frac{\partial^2\xi}{\partial x^2}. \tag{4}$$

Here ξ is the displacement of a point on the wave that was initially at coordinate x (the Lagrangian coordinate), γ is the ratio of specific heats, and c_0 is the smallest-amplitude (small ξ) speed of sound.

The normal procedure of linear acoustics is to neglect the contribution of $\partial\xi/\partial x$. This makes the differential equation a linear one. In fact, it becomes the one-dimensional wave equation [Eq. (1)] with propagation velocity c_0.

If a solution of this linear equation is

$$\xi=-\xi_0\cos(\omega t-kx), \tag{5}$$

so that $\xi=\omega\xi_0\sin(\omega t-kx)$ is the displacement velocity, then

$$\left(\frac{\partial\xi}{\partial x}\right)_{\max}=k\xi_0=\frac{\xi_0\omega}{c_0}=\frac{u_0}{u}=M,$$

which is the acoustic Mach number of the displacement velocity. As noted in the Introduction, this quantity is extremely small in air for even the loudest of sounds, and for that reason it has traditionally been neglected.

The principle of superposition does not hold for the nonlinear equation of motion, and entirely different methods are necessary for its solution.

2. NONLINEAR PROPAGATION IN FLUIDS

2.1 Solution of Basic Equation

As far back as 1808, Poisson derived a form for the displacement velocity of the so-called acoustical particle which is

$$u=f(x-(c_0+u)t), \tag{6}$$

in the case of a plane wave traveling along the x axis, where u is the particle velocity and c_0 is the small-amplitude speed of sound. That is, points in the wave with a particle velocity u travel with the enhanced speed c_0+u. Physically, this means that the points with greater particle velocity move forward in the wave relative to those points with lesser particle velocities, thus building up an increasingly steep wave front. This behavior is similar to that of the particles in a water wave, where the crests overtake the troughs and finally break into surf.

There are two historic methods for dealing with such motions. In one, the Eulerian coordinates, one deals with the motions occurring at a fixed point in the medium. In such a case, for a nondissipative medium, the equation of motion becomes

$$\dot{\mathbf{u}}+(\mathbf{u}\cdot\nabla)\mathbf{u}=-\nabla p/\rho, \tag{7}$$

where $\mathbf{u}$ is the particle velocity at the point x. In one dimension this equations becomes

$$\ddot{\xi}+\dot{\xi}\,\frac{\partial\dot{\xi}}{\partial x}=-\frac{1}{\rho}\frac{\partial p}{\partial x}, \tag{8}$$

where $\dot{\xi}$ is the particle velocity in the x direction at the point x.

In the second system (the Lagrangian coordinates), one follows the behavior of a point of the fluid that was initially at rest at the point x. The one-dimensional equation governing the propagation of a plane wave in a nondissipative ideal gas in Lagrangian coordinates becomes

$$\frac{\partial^2\xi}{\partial t^2}=\frac{c_0^2}{(1+\partial\xi/\partial x)^{\gamma+1}}\frac{\partial^2\xi}{\partial x^2}, \tag{9}$$

where ξ is the displacement of the acoustical particle that was at rest at the point x and γ is the ratio of specific heats. If the displacement of a point of the fluid from its rest position is negligible (which is the basic assumption of linear acoustics), then the two sets of coordinates become identical.

It was shown by Earnshaw in 1858 that the implicit solution of Eq. (8) is of the form

$$\dot{\xi}=f(\omega(t\pm x/v)), \tag{10}$$

where

$$v=c_0\left(1+\frac{\gamma-1}{2}\frac{u}{c_0}\right)^{2/(\gamma-1)+1}$$

and $f(\omega t\pm\omega x/c_0)$ is the solution of the corresponding linear equation. More specifically, in the case of an initially sinusoidal wave,

$$\begin{aligned}\dot{\xi}&=u\\&=u_0\sin\left[\omega t-\frac{\omega x}{c_0}\left(1+\frac{\gamma-1}{2}\frac{u}{c_0}\right)^{-2/(\gamma-1)-1}\right],\end{aligned} \tag{11}$$

where u_0 is the initial amplitude of the particle velocity.

In 1935, Fubini-Ghiron solved this equation explicitly for the case of low acoustic Mach numbers (i.e., low values of u_0/c_0). He first recast in Eq. (10) in the form

$$u(x,t)=u_0 \sin\left[\omega t-\frac{\omega x}{c_0}\left(1-\beta \frac{u}{c_0}\right)\right], \tag{12}$$

or

$$\frac{u}{u_0}=\sin\left(\omega t-kx+\frac{x}{l}\frac{u}{u_0}\right). \tag{13}$$

This equation holds for small values of the Mach number. Here $1/l=\beta Mk$, $\beta=(\gamma+1)/2$, and $k=\omega/c_0$.

The quantity l is known as the *discontinuity distance,* being the distance from the source at which the steepness of the wave front in Eq. (13) becomes negatively infinite at the point where u is equal to zero.

The quantity β is known as the *nonlinearity coefficient.* To evaluate it for a liquid, we begin by expanding the pressure in the medium in a Taylor series in the density:

$$p=p_0+\left(\frac{\partial p}{\partial \rho}\right)_{S,\rho=\rho_0}(\rho-\rho_0) + \frac{1}{2}\left(\frac{\partial^2 p}{\partial \rho^2}\right)_{S,\rho=\rho_0}(\rho-\rho_0)^2+\cdots, \tag{14}$$

or

$$p=p_0+As+(B/2)s^2, \tag{15}$$

where s is the condensation, $s=(\rho-\rho_0)/\rho_0$,

$$A=\rho_0\left(\frac{\partial p}{\partial \rho}\right)_{S,\rho=\rho_0}=\rho_0 c_0^2, \tag{16}$$

$$B=\rho_0^2\left(\frac{\partial^2 p}{\partial \rho^2}\right)_{S,\rho=\rho_0},$$

and where we have cut off the series after the quadratic term. It can easily be shown that Eqs. (12) and (13) above remain valid if γ is replaced by $1+B/A$. Hence

$$\gamma=1+B/A,$$

$$\beta=\tfrac{1}{2}(\gamma+1)=1+B/2A. \tag{17}$$

The ratio B/A is known as the parameter of nonlinearity. Its determination has been the subject of considerable study. By thermodynamic analysis (Beyer, 1974), B/A can be expressed as

$$\frac{B}{A}=2\rho_0 c_0\left(\frac{\partial c}{\partial p}\right)+\frac{2\beta_T c_0}{c_p}\left(\frac{\partial c}{\partial T}\right). \tag{18}$$

Here β_T is the volume coefficient of thermal expansion, $\beta_T=(1/V)(\partial V/\partial T)$, and c_p is the specific heat at constant pressure. The ratio can also be determined from analysis of the harmonic content of a finite-amplitude wave as it progresses through the medium. A number of values of B/A are shown in Table 1.

An empirical formula, known as *Ballou's rule,* connects the values of B/A with the sound velocity c_0:

$$B/A=(1.2\times 10^4\ \mathrm{m/s})/c_0-0.5. \tag{19}$$

This formula is moderately successful for liquid metals and a number of other liquids, but it is not uniformly successful, and fails to predict the temperature dependence of the ratio. Some theoretical justification of this rule has been provided by Hartman (1979), who connected the behavior of B/A with the intermolecular potentials of the molecules.

The solution of Eq. (12) is given by (Fubini-Ghiron, 1935)

$$\frac{u}{u_0}=2\sum_{n=1}^{\infty}\frac{J_n(nx/l)}{nx/l}\sin n(\omega t-kx), \tag{20}$$

where $J_n(x)$ is a Bessel function of the first kind of order n. It is clear from Eq. (20) that the fundamental component falls off with distance, while each harmonic increases from zero to a maximum and then falls off. Examples of this behavior are shown in Fig. 1. The solution (20) is valid only as long as $x \ll l$ and as long as damping can be neglected. It is useful, over short ranges, in indicating the harmonic content of an intense sound wave, but is not very effective in realistic situations.

2.2 Burgers' Equation

A method that does take absorption into account is one proposed by Burgers (1948). In this case, Eq. (9) is replaced by an equation in a coordinate system that travels with the crest of the wave. That is, we replace t by $\tau=t-x/c$. Such a change in coordinates enables one to eliminate a number of small terms, so that a new differential equation is obtained—Burgers' equation:

Table 1. Values of B/A. Except where indicated, all values are at atmospheric pressure.

Substance	T (°C)	B/A	Substance	T (°C)	B/A
Distilled water	0	4.2	Benzyl alcohol	30	10.2
	20	5.0	Diethylamine	30	10.3
	40	5.4	Ethylene glycol	30	9.7
	60	5.7	Ethyl formate	30	9.8
	80	6.1	Heptane	30	10.0
	100	6.1	Hexane	30	9.9
Pressure			Methyl acetate	30	9.7
1 atm	30	5.2	Cyclohexane	30	10.1
200 kg/cm^2	30	6.2	Nitrobenzene	30	9.9
4000	30	6.2	Mercury	30	7.8
8000	30	5.9	Sodium	110	2.7
Seawater			Potassium	100	2.9
(3.5%)	20	5.25	Tin	240	4.4
Methanol	20	9.6	Idium	160	4.6
Ethanol	0	10.4	Bismuth	318	7.1
	20	10.5	Monatomic gas	20	0.67
	40	10.6	Diatomic gas	20	0.40
n-propanol	20	10.7			
n-butanol	20	10.7	Methyl iodide	30	8.2
Acetone	20	9.2	Sulfur	121	9.5
Benzene	20	9.0	Glycerol (4% H_2O)	30	9.0
Chlorobenzene	30	9.3	1.2-dichlorohexafluoro-		
Liquid nitrogen	b.p.	6.6	cyclopentene (DHCP)	30	11.8

$$c_0^3 u_z - \beta c_0 u u_\tau = \frac{1}{2}\frac{b\eta}{\rho_0} u_{\tau\tau} = \frac{c_0^3 \alpha}{\omega^2} u_{\tau\tau}. \tag{21}$$

Here a single subscript denotes differentiation by that variable, α=absorption coefficient, $b=\frac{4}{3}+\eta'/\eta$ is the viscosity number, $z=x$, $\tau=t-x/c_0$.

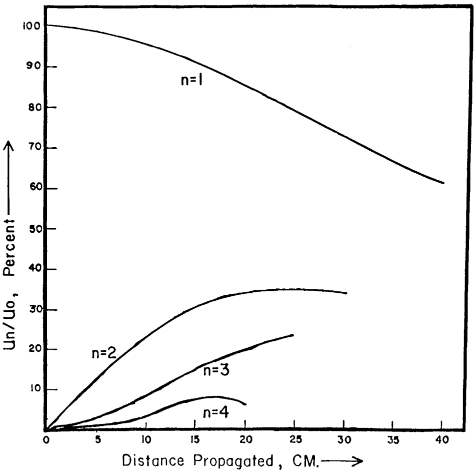

FIG. 1. Fundamental and several harmonics as a function of distance traveled in water. Frequency=2.5 MHz; pressure amplitude, 3.0 atm (Ryan, 1963).

Equation (21) can be put in dimensionless form by setting

$$W=u/u_0,\quad \sigma=z/l,\quad y=\omega\tau,$$

and

$$\Gamma=\beta(u_0/c_0\alpha)(2\pi/\lambda)=1/\alpha l:$$

$$W_\sigma - W W_y = (1/\Gamma) W_{yy}. \tag{22}$$

The quantity Γ is sometimes called the Goldberg number. It is widely used as a scaling parameter. Physically, its reciprocal measures the effective absorption coefficient over the discontinuity distance.

A further substitution, $W = (2/\Gamma\zeta)\zeta_y = (2/\Gamma)\partial \ln\xi/\partial y$, in Eq. (18) reduces that equation to the (linear) diffusion equation

$$\xi_\tau = (1/\Gamma)\xi_{yy}. \tag{23}$$

This equation can be solved exactly (see Beyer, 1974, and Rudenko and Soluyan, 1977).

Table 2. Notation in generalized Burger's equation.

Type of wave	a	W	f	E
Plane	0	u/u_0	σ	Γ^{-1}
Cylindrical	$\frac{1}{2}$	$\left(\frac{\sigma}{\sigma_0}\right)^{1/2}\frac{u}{u_0}$	$2[(\sigma\sigma_0)^{1/2}-\sigma_0]$	$\left(1+\frac{f}{2\sigma_0}\right)^{\Gamma-1}$
Spherical	1	$\left(\frac{\sigma}{\sigma_0}\right)\frac{u}{u_0}$	$\sigma_0 \ln\frac{\sigma}{\sigma_0}$	$\exp\left(\frac{f}{\sigma_0}\right)^{\Gamma-1}$

2.3 Spherical and Cylindrical Coordinates

A generalized form of Burgers' equation, good for the three standard coordinate forms, can be written as

$$W_f - WW_z = EW_z, \tag{24}$$

where

$$W = (\sigma/\sigma_0)^a u/u_0, \quad z = \omega(1 - x/c_0), \quad \sigma = x/l,$$

$\sigma_0 = x_0/l$, x_0 is the source radius (for spheres and cylinders), x is the field position, and the definitions of a, f, and E are given in Table 2. These lead to rather complicated calculations and also suffer from limits on their generality.

In all the analysis thus far, it has been assumed that one is dealing with unbounded waves. A technique for handling bounded beams in a nondissipative medium was developed by Zabolotskaya and Khokhlov (1969). This procedure led to the equation

$$\frac{\alpha}{2}\frac{\partial^2\rho}{\partial\tau^2} - \frac{\partial^2\rho}{\partial\tau\,\partial x} + \frac{c_0}{2}\left(\frac{\partial^2\rho}{\partial y^2} + \frac{\partial^2\rho}{\partial x^2}\right) = 0. \tag{25}$$

Here $\alpha = (\gamma+1)/2\rho_0 c_0$, $\tau = t - x/c$, and ρ is the density of the medium.

The dissipative case was considered by Kuznetsov (1970), leading to the equation

$$\frac{\alpha}{z}\frac{\partial^2\rho}{\partial\tau^2} - \frac{\partial^2\rho}{\partial\tau\,\partial x} + \frac{c_0}{2}\left(\frac{\partial^2\rho}{\partial y^2} + \frac{\partial^2\rho}{\partial z^2}\right) + K\frac{\partial^3\rho}{\partial\tau^3} = 0, \tag{26}$$

where $K = b\eta/2\rho_0 c_0^3$ and b is the combination of viscosity and thermal conductivity terms

$$b = \frac{\eta'}{\eta} + \frac{4}{3} + \kappa\left(\frac{1}{c_v} - \frac{1}{c_p}\right). \tag{27}$$

Here η' is the bulk viscosity, η is the shear viscosity, κ is the thermal conductivity, and c_v and c_p are the specific heats at constant volume and pressure, respectively. Equation (26) is referred to in the literature as the KZK equation. Numerous examples of solutions of this equation and Eq. (25) can be found in Bakhvalov *et al.* (1987) and Novikov *et al.* (1987). The analysis presented above has been confirmed experimentally in many instances. Studies of the harmonic content of an originally sinusoidal wave have produced results closely following those of the Fubini-Ghiron solution of Eq. (17). Examination of the distorted wave form has also been carried out. Analysis of the phenomenon of optical diffraction produced by a sound beam (the Debye–Sears effect) at high sound intensities indicates that the distortion of the wave form due to the production of higher harmonics follows closely the values calculated theoretically.

3. SHOCK WAVES AND AEROACOUSTICS

3.1 Rankine–Hugoniot Equation

The steepening of the wave front due to nonlinearities, discussed in Sec. 1, leads, in the absence of dissipation, to the formation of a front of infinite steepness, which is characteristic of a shock front. The development of this steepness is shown in Fig. 2. One could therefore consider finite-amplitude waves as a form of weak shocks.

The field of shock waves is an enormous one and will not be discussed in detail here (see SHOCK WAVES). The fundamental equation governing the fluid motion, relating the properties of the fluid ahead of the shock (denoted by the subscript a) and those behind

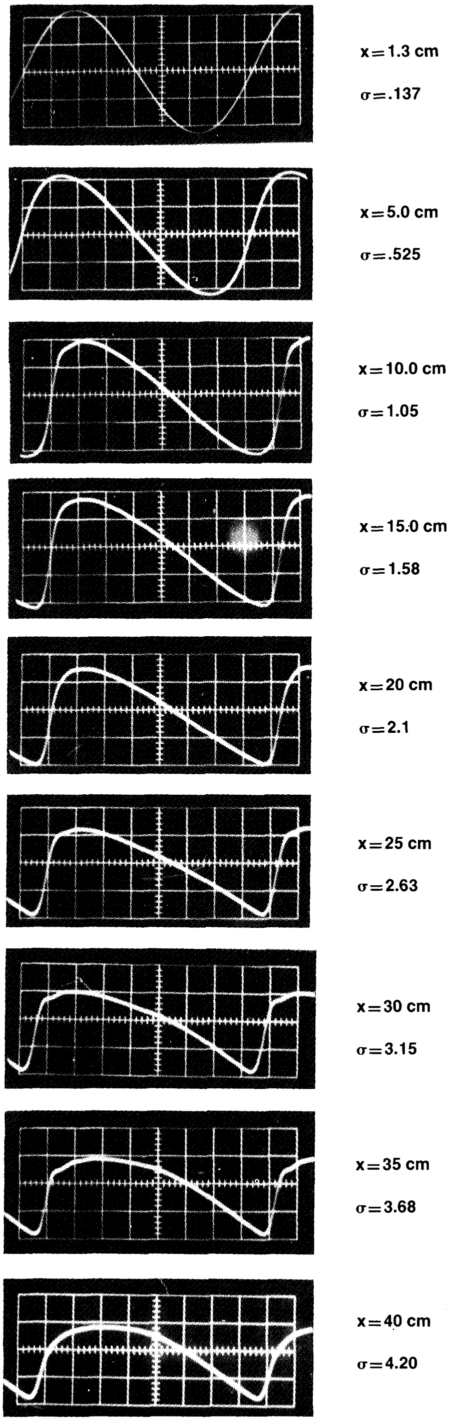

FIG. 2. Progressive distortion of a wave form in an originally sinusoidal signal. Frequency $\nu=2.58$ MHz; initial acoustic pressure, 6.0 atm. The medium was water at 22°C, x=distance from source, and σ=reduced length=x/l. Since the pictures are oscilloscope traces of the signal, the leading edge of the wave is at the left (Ryan, 1963).

the shock (subscript b) is the Rankine–Hugoniot equation (see Beyer, 1974):

$$\frac{\rho_b}{\rho_a}=\frac{p_b(\gamma+1)+p_a(\gamma-1)}{p_a(\gamma+1)+p_b(\gamma-1)}, \tag{28}$$

or using the coefficient of nonlinearity, $\beta=(\gamma+1)/2$,

$$\frac{\rho_b}{\rho_a}=\frac{p_b\beta+p_a(\beta-\frac{1}{2})}{p_a\beta+p_b(\beta-\frac{1}{2})}. \tag{29}$$

In the weak-shock approximation ($u_0<0.1c_0$), the speed of propagation of the shock front v, derived from Eq. (29), takes the form (Blackstock, 1966)

$$v=c_0+(\beta/2)(u_a+u_b). \tag{30}$$

If a shock begins to form at a point $\bar{x}$ and a time $\bar{t}$, then it will arrive at a point x at a time t given by

$$t=\bar{t}+\int_{\bar{x}}^{x}\frac{dy}{\bar{v}},$$

where y is a dummy variable. If we substitute the speed of propagation of the shock, Eq. (30), then we have, approximately,

$$\frac{1}{\bar{v}}=\frac{1}{c_0}-\frac{1}{2}\frac{\beta}{c_0^2}(u_a+u_b), \tag{31}$$

from which we can write down the derivative

$$\frac{d\tau_s}{dz}=-\frac{\beta}{2c_0^2}(u_a+u_b). \tag{32}$$

A plot of the successive distortions of the wave form is shown in Fig. 3.

3.2 Shock-Wave Thickness

The properties of the medium ahead of and behind the shock front must change continuously from one set into the other. The distance over which this change takes place is measured by the shock-wave thickness. The variation of the pressure through the front is usually approximated by the expression

$$p=\frac{1}{2}(p_a+p_b)+\frac{1}{2}(p_b-p_a)\tanh\frac{2x}{\delta}, \tag{33}$$

and the wave thickness δ is defined by the relation

$$\delta=\frac{p_2-p_1}{(dp/dx)_{\max}}, \tag{34}$$

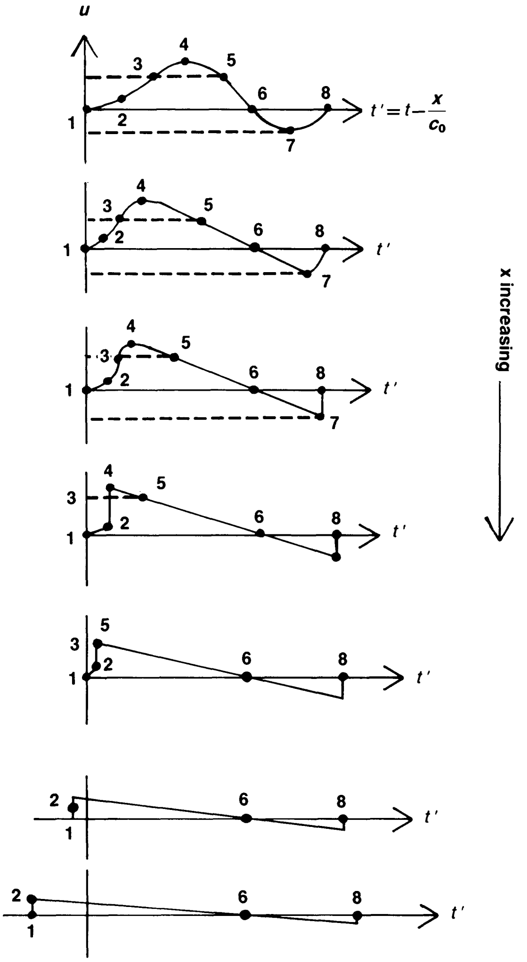

FIG. 3. Steepening of an originally sinusoidal pulse as a function of the reduced time $t' = t - xc_0$. The numbers refer to corresponding points on the wave. As in the case of Fig. 2, the leading edge of the wave (in space) is at the left (Blackstock, 1966).

which can be calculated from thermodynamic properties. It is in fact given by

$$\delta = \frac{[16b\eta + \kappa(1/c_v - 1/c_p)]}{\beta}\,\frac{c_0}{\Delta p}. \tag{35}$$

These results have been confirmed experimentally (see Eisenmenger, 1984).

3.3 N Waves

The buildup of a shock in the neighborhood of a speeding projectile in air (or the formation of the analogous wave in the case of a fast-moving boat on the surface of water) is a direct consequence of the medium's inability to move the wave out of the way of the oncoming source fast enough, so that there is a substantial compression of the medium just ahead of the projectile. The form of the wave front in this case can be obtained from a consideration of Fig. 4.

In the figure, a source of sound is moving to the right with velocity $v = 1.6c_0$. The points A, B, C, D, and E are the positions of the source as observed after each of a series of equal time intervals Δt.

The spherical waves set up by the source when it was located at each of these points are shown at corresponding later times. Thus, when the source reaches B, the sound from point A will lie along the curve B', of radius $c_0\Delta t$. When the source reaches C, the sound from A lies along the curve C' [radius $c_0(2\Delta t)$], while the sound from B lies along C'' (radius $c_0\Delta t$). The other curves are similarly drawn when the source is at D,E.

The trace of the resulting wave front is then given by the tangent curve shown. The angle made by the front with the direction of motion of the source is then

$$\sin\phi = c/u_0 = 1/M, \tag{36}$$

where M is the Mach number. The front of the wave therefore moves out over a conical surface. In the air, this is what is known as

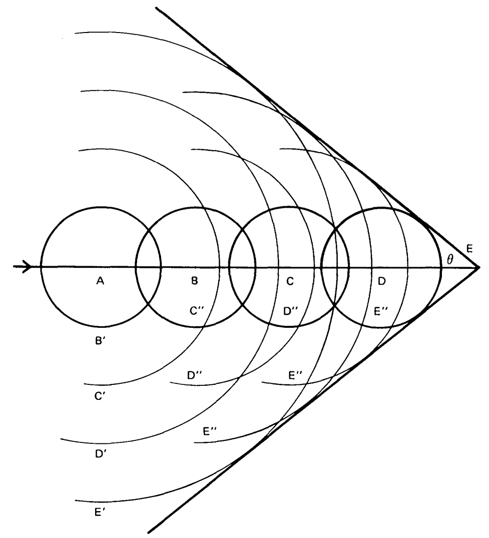

FIG. 4. Generation of a sonic boom (or bow wave). The source of the sound is moving at constant speed along the path *ABCD*, and the crests of the sound waves emitted from the four points *A*, *B*, *C*, and *D* are shown.

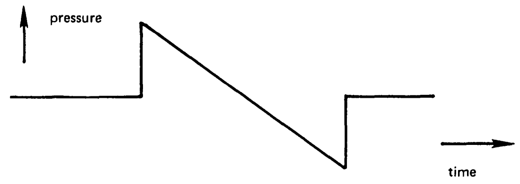

FIG. 5. N wave.

the sonic boom. The corresponding wave in water is called the bow wave.

In addition to this disturbance in front of the projectile, where the pressure undergoes a sudden rise, there is also a similar disturbance at the rear of the projectile, where the pressure returns to normal. The entire pressure cycle is shown in Fig. 5 and has been given the name of the N wave (after Dumond *et al.*, 1946).

The N wave differs from the ordinary shock wave in that the pressure is not constant behind the shock front. Therefore, some of the formulas applying to that case require modification. In the most elementary treatment, one starts with the Earnshaw relation [Eq. (11)] in terms of $\gamma=(B/A)+1$, in the form

$$u-u_0=\frac{2c_0}{\gamma-1}\left(\frac{p}{p_0}\right)^{(\gamma-1)/2\gamma} . \tag{37}$$

After numerous transformations, and with the approximation $p=p_0(1+\phi)$, where $\phi\ll 1$, Dumond *et al.* (1946) were able to derive the relation

$$-\frac{\gamma+1}{\gamma}c\phi\frac{\partial p}{\partial x}=\frac{\partial p}{\partial t} , \tag{38}$$

where terms of order ϕ and higher are neglected. The speed with which a particular pressure propagates can be found from the relation

$$\left.\frac{\partial x}{\partial t}\right|_{p=\text{const}}=-\frac{\partial p/\partial t}{\partial p/\partial x}=\frac{\gamma+1}{2\gamma}c_0\phi. \tag{39}$$

This is the speed with which the pressure p moves relative to the center of the N wave (where $p=0$ and $u=c_0$). Hence two points with pressures corresponding to the values φ and $-\varphi$ will separate at a rate that is twice the value given by Eq. (39) or $[(\gamma+1)/\gamma]c_0\phi$.

Figure 6 shows the behavior of the N wave in spatial coordinates. The discontinuity moves to the right with a speed $[(\gamma+1)/x\gamma]c_0\phi A$ relative to the forward motion of the point. In some small time interval δt, it therefore moves to the x position C''. Now the excess pressure $p_0\phi A$ moves to the right at twice this velocity, reaching the line AC'; i.e., the peak pressure in the shock has fallen to the value $p_0\,\phi''_A$. Since the reverse holds at the tail of the wave, the rate at which the entering wave increases its length is given by

$$\frac{dL}{dt}=\frac{\gamma+1}{2\gamma}c_0\phi . \tag{40}$$

Thus the N wave gets weaker in strength and longer in its duration as it propagates.

3.4 The Lighthill Equations

The starting point for the development of the Lighthill equations (Lighthill, 1954, 1978) is the study by Lord Rayleigh of the scattering of sound from small-scale inhomogeneities. In his development, Rayleigh (1945, Vol. II, p. 150) wrote down the expression for the D'Alembertian of the pressure

$$\Box^2 p\equiv\nabla^2 p-\frac{1}{c_0^2}\frac{\partial p}{\partial t^2} . \tag{41}$$

He then collected all other terms on the right-hand side of the equation. These included terms dependent on the relative inhomogeneities in the sound velocity $\Delta c/c_0$ and the density $\Delta\rho/\rho_0$ of the medium, as well as other nonlinear terms in the pressure and its derivatives. Since he was interested in the effectiveness of the inhomogeneities in scattering an incident sound beam, he neglected the higher-order terms not associated with the inhomogeneities, so that his equation took the form

$$\Box^2 p=-\frac{2\Delta c}{c_0}\frac{\partial^2 p_s}{\partial t^2}-\frac{\partial}{\partial y}\left(\frac{\Delta\rho}{\rho_0}\right)\frac{\partial p_s}{\partial y} , \tag{42}$$

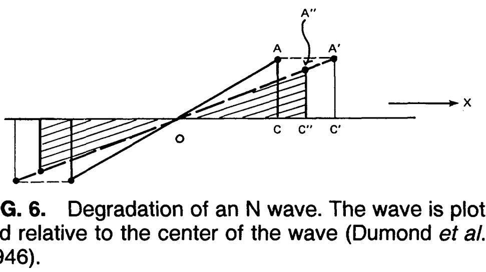

FIG. 6. Degradation of an N wave. The wave is plotted relative to the center of the wave (Dumond *et al.*, 1946).

where p_s is the pressure in the incident beam.

Lighthill's study began at this point. Since he was interested in the production of sound by turbulence, he started with the absence of an extended source (so that $p_s=0$); the terms kept by Rayleigh on the right-hand side of Eq. (42) disappear, while the terms that he neglected must now be considered.

In this analysis, Lighthill considered a fluctuating fluid flow that occupied a limited part of a large volume of fluid, the rest of which was at rest (Fig. 7). He then made a comparison between that system [Fig. 7(a)] and a similar one [Fig. 7(b)] whose density values were those appropriate to a uniform acoustic medium at rest. Outside the zone of low fluctuations, the two systems were therefore identical. Lighthill then stated that the difference between the systems was to be "considered as if it were the effect of a fluctuating external force field, known if the flow is known, acting in the said uniform acoustic medium at rest and hence radiating sound in it according to the ordinary laws of acoustics" (Lighthill, 1954, p. 564).

The momentum in a fixed region of space changes as a result of the combined effect of the stresses on its boundary and the momentum flow across that boundary. The boundary stresses are expressed by the Reynolds stresses $\rho v_i v_j$, while p_{ij} gives the real stress. Hence the contribution to the rate of momentum change in the shaded areas of Fig. 7(a) is of the form

$$T'_{ij}=p_{ij}+\rho v_i v_j. \tag{43}$$

A uniform acoustic medium at rest [Fig. 7(b)] would have stresses only in the form

$$T''_{ij}=\rho c_0^2 \delta_{ij}, \tag{44}$$

where δ_{ij} is the Kronecker delta; that is, only simple hydrostatic pressure would be present. The difference between these two terms then is the equivalent external stress acting on the uniform acoustic medium of Fig. 7(b):

$$T_{ij}=T'_{ij}-T''_{ij}=\rho v_i v_j+p_{ij}-\rho c_0^2 \delta_{ij}. \tag{45}$$

Under these conditions, the equation of motion then takes the form

$$\square^2 p=-\frac{\partial^2 T_{ij}}{\partial y_i \partial y_j}=c_0^2 \square^2 \rho=c_0^2 \nabla^2 \rho-\frac{\partial^2 \rho}{\partial t^2}. \tag{46}$$

This equation can also be written in terms of the density:

$$\nabla^2 \rho=-\frac{1}{c_0^2}\frac{\partial^2 \rho}{\partial t^2}=\frac{1}{c_0^2}\square^2 \rho=\frac{\partial^2 T_{ij}}{\partial x_i\, \partial x_j}. \tag{47}$$

The effect of viscosity on the processes involved here may be taken into account in the stress tensor p_{ij}. For a Stokesian gas (for which the bulk viscosity $\zeta=0$),

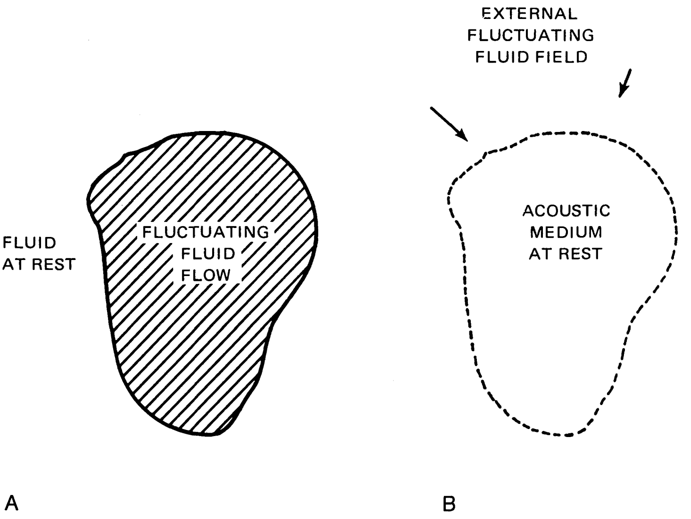

FIG. 7. (a) Lighthill concept of fluctuating fluid flow and (b) equivalent acoustic system.

$$p_{ij}=p_0\delta_{ij}+\eta\left[-\frac{\partial v_i}{\partial x_j}-\frac{\partial v_j}{\partial x_i}+\frac{2}{3}\left(\frac{\partial v_k}{\partial x_k}\right)\delta_{ij}\right], \qquad (48)$$

where η is the shear viscosity coefficient and p_0 is the equilibrium pressure.

Equation (46) together with Eq. (45) constitute the Lighthill equations. Equation (46) is a rather complicated one, but it can be reduced for the problems at hand. First, the stresses T_{ij} can be neglected outside the flow region. In such space, $v_i=0$, and one has only the motions of the medium due to the sound waves. Next, the viscous stresses in p_{ij} constitute small-scale absorption effects which can be neglected except for large-scale phenomena. And, finally, for low Mach numbers, T_{ij} can be approximated by $\rho_0 v_i v_j$.

The Lighthill equations have had great significance in the study of aerodynamic noise produced by turbulence (see Lighthill, 1978). Their application to nonlinear acoustics will be treated in Sec. 5.

3.5 Monopole, Dipole, and Quadrupole Sources

The following is a brief discussion of the mathematics of acoustic sources. The simplest possible sound is a small pulsating sphere whose radius varies sinusoidally with time ($u_s=u_0e^{i\omega t}$). By small is meant $ka\ll 1$, where a is the radius of the sphere and k is the wave number. In such a case the pressure at a distance r will be

$$p=i\omega\rho a^2\frac{u_0e^{i(\omega t-kr)}}{r}. \qquad (49)$$

We now define S as the *simple* or *point source strength.* It is equal to the amplitude of the volume rate of expansion the sphere, i.e., $S=4\pi a^2u_0$. Hence the pressure can be written as

$$p=\frac{i\omega\rho Se^{i(\omega t-kr)}}{4\pi r}. \qquad (50)$$

The variations of the density can be found from Eq. (50), since the excess pressure $p=c_0^2\Delta\rho$ or

$$\rho-\rho_0=\frac{i\omega\rho Se^{i(\omega t-kr)}}{4\pi c_0^2r}. \qquad (51)$$

Such a simple source is an example of a *monopole source.* Monopole sources usually involve the introduction of a new fluid into the source region (explosions or gas combustions), but they can also originate from the introduction of heat into a localized region (laser-induced sound) or from the presence of turbulence (pulsed jets).

If the source emits more generalized radiation than sinsuoidal, we may descreibe the density fluctuations, say, by the expression

$$\rho-\rho_0=\frac{1}{4\pi c_0^2}\frac{q'(t-r/c_0)}{r}. \qquad (52)$$

where q is the rate at which a point source adds mass to the medium, and $q'(z)=\partial q(x)/\partial z$.

In comparison with Eq. (51), $q[(t-r/c_0)]^{\cdot}$ has replaced $Se^{i(ut-kr)}$. For an extended distribution of sources of matter, we need to replace the point equation (46) by a volume integral, i.e., rewrite Eq. (46) as

$$\rho-\rho_0=\frac{1}{4\pi c_0^2}\int\frac{\partial}{\partial t}Q\left(\mathbf{R},t-\frac{|\mathbf{r}-\mathbf{R}|}{c_0}\right)\times\frac{dV_R}{|\mathbf{r}-\mathbf{R}|}, \qquad (53)$$

where $q\,dV_R$ is the time rate of mass production in the volume element dV_R marked out by the vector $\mathbf{R}$ (Fig. 8). The integration in Eq. (53) is over all space, but in practice is carried out only over the region in which there are sources. The quantity $q'(t)$ is called the *instantaneous strength* of the point source, and $\partial Q/\partial t$ is the (mass) *source strength per unit volume* or more briefly the (monopole) *source density.*

All the expressions for the density can be converted to pressure expressions by the linear relationship between pressure and density given in Sec. 1, i.e., $\Delta p=c_0^2\Delta\rho$.

While there are many monopole acoustic sources, in most physical situations the monopole strength vanishes, and attention

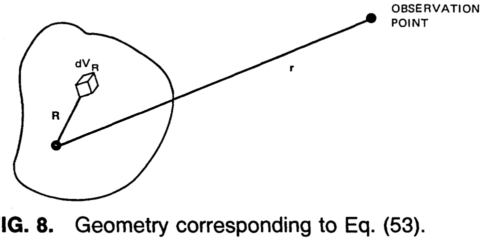

FIG. 8. Geometry corresponding to Eq. (53).

must be focused on the next higher level of sources: dipoles. If there are no matter sources, the sound is generated by a fluctuating force field F_i per unit volume in some part of the medium ($i=1,2,3$, corresponding to the vector components x,y,z). Such a force would appear as $-(\partial F_i/\partial x_i)$ on the right-hand side of Eq. (41). This force field would be equivalent to a source distribution with strength per unit volume equal to $-(\partial F_i/\partial x_i)$.

Paralleling the derivation of the monopole source, the density from such dipole sources can be obtained as

$$\rho-\rho_0=-\frac{1}{4\pi c_0^2}\frac{\partial}{\partial x_i}\int F_i\left(\mathbf{R},t-\frac{|\mathbf{r}-\mathbf{R}|}{c_0}\right)\times\frac{dV_R}{|\mathbf{r}-\mathbf{R}|}, \tag{54}$$

which, in the far field, takes the form

$$\rho-\rho_0\simeq\frac{1}{4\pi c_0^2}\int\frac{r_iR_i}{|\mathbf{r}-\mathbf{R}|^2}\frac{1}{c_0}\times\frac{\partial}{\partial t}F_i\left(\mathbf{R},\mathrm{t}-\frac{|\mathbf{r}-\mathbf{R}|}{c_0}\right)dV_R\,. \tag{55}$$

Thus the density changes at the point of interest depend on the time rate of change of the dipole strength F_i and are not due to simultaneous time changes in the local density, but depend on their time of arrival at the distant point.

While the fluctuating stresses T_{ij} in Eq. (41) produce a force per unit volume equal to their inward flux $-(\partial T_{ij}/\partial x_j)$, so that they generate like a dipole field of strength $-(\partial T_{ij}/\partial x_j)$ per unit volume, the sound radiated is not to be computed from the total dipole strength per unit volume, since this is, in fact, zero at any instant of time. Rather, the sound will come from the next higher order of terms; i.e., it will be the equivalent quadrupole field, a field that is the limiting case of four simple sources which obey the inverse square law of radiation. At large distances from the quadrupole producing the radiation, arguments similar to those above justify the approximation

$$\rho-\rho_0\cong\frac{1}{4\pi c_0^2}\int\frac{(R_i-r_i)(R_j-r_j)}{|\mathbf{R}-\mathbf{r}|^3}\times\frac{1}{c_0^2}\frac{\partial^2}{\partial t^2}T_{ij}\left(\mathbf{R},t-\frac{|\mathbf{r}-\mathbf{R}|}{c_0}\right)dV_R\,. \tag{56}$$

When the coordinate origin is within the flow region and the latter's dimensions are small compared with $\mathbf{r}$ (Fig. 8), we can disregard $\mathbf{R}$ and R_l in the difference terms of the integrand of Eq. (56), obtaining

$$\rho-\rho_0\cong\frac{1}{4\pi c_0^2}\frac{r_ir_j}{r^3}\times\int\frac{1}{c_0^2}\frac{\partial^2}{\partial t^2}T_{ij}\left(\mathbf{R},t-\frac{|\mathbf{r}-\mathbf{R}|t}{c_0}\right)dV_R. \tag{57}$$

Equation (57) has served as the starting point for the study of the interaction of two sound beams that will be described in Sec. 5.

4. RADIATION PRESSURE, STREAMING, AND CAVITATION

4.1 Radiation Pressure

In linear acoustics, the average pressure at a point due to a sound wave (which must be of infinitesimal amplitude in order that linearity prevail) must be equal to zero. The fact that there is a steady force associated with the passage of sound waves is evidence of finiteness of the waves and hence nonlinearity.

The situation is complicated by the fact that it becomes necessary to distinguish between the acoustic behavior of a fixed point in space (Eulerian coordinates) and the behavior of a point in the fluid that was originally at rest at some reference point (Lagrangian coordinates). Since the radiation pressure is the pressure measured at a surface which must be moving with the particles of the medium as they undergo oscillations, the Lagrangian representation is the more useful in calculating this pressure.

In addition, the literature discriminates between two definitions: the *Rayleigh radiation pressure*, defined as the difference between average pressure at a surface moving with the particle (i.e., the mean Lagrangian pressure), and the pressure that would have existed in the fluid of the same mean density at rest. For a gas, this quantity has the value $[(\gamma+1)/4]\langle E\rangle$, where $\langle E\rangle$ is the mean energy density in the sound wave. In the case of a liquid, the Rayleigh pressure is equal to $[(1+B/2A)/2]\langle E\rangle$.

The second definition is the *Langevin radiation pressure,* defined as the difference between the mean pressure (the Lagrangian pressure) at an absorbing or reflecting wall and in the same acoustic medium, at rest, behind the wall. This value is equal to $\langle E\rangle$.

The usual experimental setup leads to a radiation pressure of the Langevin type. While these pressures are relatively small, they are sufficient to produce instances of acoustic levitation of small particles and can also lead to the phenomenon of an acoustical fountain. Radiation pressure can be used as a measure of the intensity of a sound beam, since the mean intensity $\langle I\rangle$ is given by $\langle I\rangle = c\langle E\rangle$, where c is the speed of sound, and a sound-intensity meter has been developed on the basis of this principle. The measurement of the sound intensity can also be used for the determination of the sound absorption coefficient (Piercy and Lamb, 1954).

4.2 Acoustic Streaming

The flow of air in a strong sound beam was first observed by Faraday (1831), but the phenomenon came under more regular study only many years later. The eddies of air flow in a Kundt's tube, for example, were shown by Rayleigh (1883) to be due to this phenomenon. Analysis of streaming is carried out by expressing such variables as density, velocity, and temperature as series of zero-, first-, and second-order terms. Under such circumstances, the second-order vorticity $\mathbf{R}_2$ ($\mathbf{R}_2 = \nabla\times\mathbf{u}_2$, where $\mathbf{u}_2$ is the second-order displacement velocity) is given by (Medwin and Rudnick, 1953)

$$\Delta^2\mathbf{R}_2 = \frac{1}{\rho_0}\, b\, \Delta\rho_1 \times \Delta\, \frac{\partial\rho_1}{\partial\tau} - \frac{\rho_0}{\eta}\, \nabla\times(\mathbf{u}_1\times\mathbf{R}_1) - \frac{1}{\rho_0}\, \nabla\times(\rho_1\nabla\times\mathbf{R}_1). \qquad (58)$$

The first term on the right-hand side is due to Eckart (1948), the second to Rayleigh (1945, Vol. II, p. 333).

Solutions of this equation exist for special cases. In the case of a cylinder of liquid, the flow is similar to that shown schematically in Fig. 9. Figure 10 shows an experimental example of this streaming.

Since the Eckart streaming term depends on the quantity b [Eq. (27)], which is the basic constituent of the acoustic absorption coefficient, the amount of streaming depends directly on the acoustic absorption coefficient and can be used to measure this quantity.

4.3 Acoustically Induced Cavitation

The name cavitation applies to the phenomenon of the appearance of holes in liquids. This appearance is usually due to the stress of tensile forces of some kind. These latter may, in turn, be due to high-speed flow, the rapid motion of a solid (such as a propeller blade) through the liquid, or high-intensity sound. It is this last phenomenon that is of concern here.

Acoustical cavitation can be divided into two types. The first of these, called by various names such as gaseous cavitation or pseudocavitation, results from the growth of gaseous bubbles in the liquid during the pas-

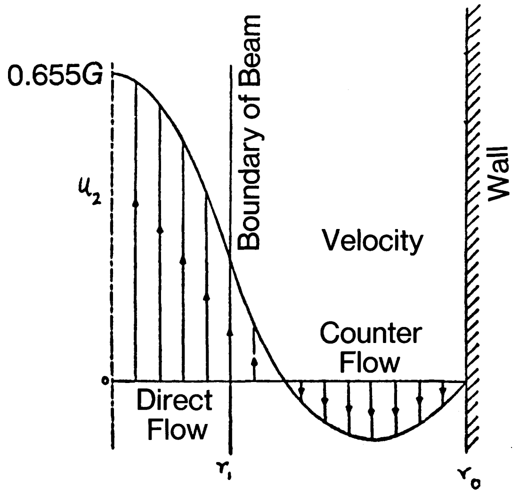

FIG. 9. Plot of theoretical value of streaming velocity as a function of radius r in a cylinder. (r_1 = radius of beam, r_0 = radius of cylinder. $r_1/r_0 = 1/3$, G is a parameter of the medium.) (Eckart, 1948.)

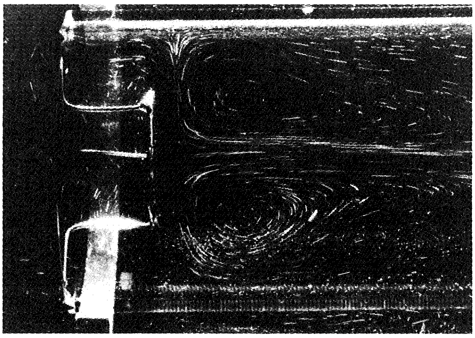

FIG. 10. Acoustic streaming from a sound source in water. The sound source is at the left center. The motion is made visible by a means of a suspension of finely divided aluminum particles (Liebermann, 1949).

sage of a sound wave, while the second, vapor cavitation, describes the appearance and growth of holes in a degassed liquid, where the only gaseous content is the vapor of the host liquid. Flynn (1965) has described these, respectively, as the stable bubble field and the transient bubble (or cavitation) field.

In the cavitation process, one sees a very small bubble that grows very rapidly and then collapses with a distinctive sound. What is not seen is how the tiny bubble came to be. Much of cavitation theory presumes the bubble was already there, clinging to some bit of impurity in the medium or created by the passage of some external radiation. Such analyses have led to an understanding of conditions under which a bubble will be stable in a liquid. A bubble of radius R will be stable if the pressure in the liquid p_0 is related to the pressure inside the bubble, p_i, by the equation

$$p_0=p_i-(2\sigma/R), \tag{59}$$

where σ is the surface tension of the liquid. In the case of vapor cavitation, $p_i=p_v$, the vapor pressure of the host liquid. Equation (59) indicates that the cavitation bubble will be stable only if the pressure in the liquid is lower than the vapor pressure by $2\sigma/r$, so that a very minute bubble would require enormous negative pressures to keep it stable. Such a bubble would dissolve itself rapidly if it were small and rise rapidly to the surface if it were large, being unstable in either case. Another possibility is the existence of a thin organic skin in the bubble, preventing solution of the entrapped gas in the surrounding liquid, and therefore giving greater stability to the minute bubble.

The dynamic behavior of bubbles has been studied extensively, and several equations have been developed to describe their behavior. In the Noltingk–Neppiras (1950, 1951) equation, the liquid is assumed to be incompressible, so that the analysis does not hold up in the stage of bubble collapse. The governing equation here is

$$\frac{dU}{dt}+\frac{3}{2}\frac{U^2}{R}=\frac{1}{\rho_{l_0}R}\left[\left(p_0+\frac{2\sigma}{R_0}\right)\left(\frac{R_0}{R}\right)^{3\Gamma}+p_v-\frac{2\sigma}{R}\right.$$
$$\left.-p_{l_0}+p_m\sin\omega t\right]. \tag{60}$$

In the Herring–Flynn (1941,1965) equation, compressibility of the liquid is taken roughly into account, but the results are not accurate if the rate of growth of the bubble $(U-dR/dt)$ is comparable with or large than the speed of sound. The Herring–Flynn equation is

$$\left(1-\frac{2U}{c_0}\right)\frac{dU}{dt}+\frac{3}{2}\left(1-\frac{4}{3}\frac{U}{c_0}\right)\frac{U^2}{R}$$
$$=\frac{1}{\rho_{l_0}R}\left[p_l(t)-p_\infty(t)+\frac{R}{c_0}\left(1-\frac{1}{c_0}U\right)\frac{dp_l}{dt}\right]. \tag{61}$$

Finally, there is the Kirkwood–Bethe equation (1942), developed for underwater explosions. This takes the form

$$\left(1-\frac{u}{c}\right)\frac{dU}{dt}+\frac{3}{2}\left(1-\frac{U}{3c}\right)\frac{U^2}{R}$$
$$=\left(1+\frac{U}{c}\right)H+\frac{U}{c}\left(1-\frac{U}{c}\right)\frac{dH}{dR}=0. \tag{62}$$

Here

$$H=\int_{p_\infty}^{p(R)}\frac{dp}{\rho}.$$

In the above equations, Γ is the effective specific-heat ratio, $p_\infty=p_{l_0}-p_m\sin\omega t$, R is the radius of the bubble, R_0 is the initial radius of the bubble, p_{l_0} and p_l are the pressure in the liquid at $R=R_0$ and $R=R_l$, p_v is the vapor pressure, $U=dR/dt$, σ is the surface tension, and p_m is the pressure amplitude of the sound wave of angular frequency.

The similarity of results from these various equations are indicated in the theoretical results shown in Fig. 11. The solid curve expresses in each case the behavior at 10 kH,

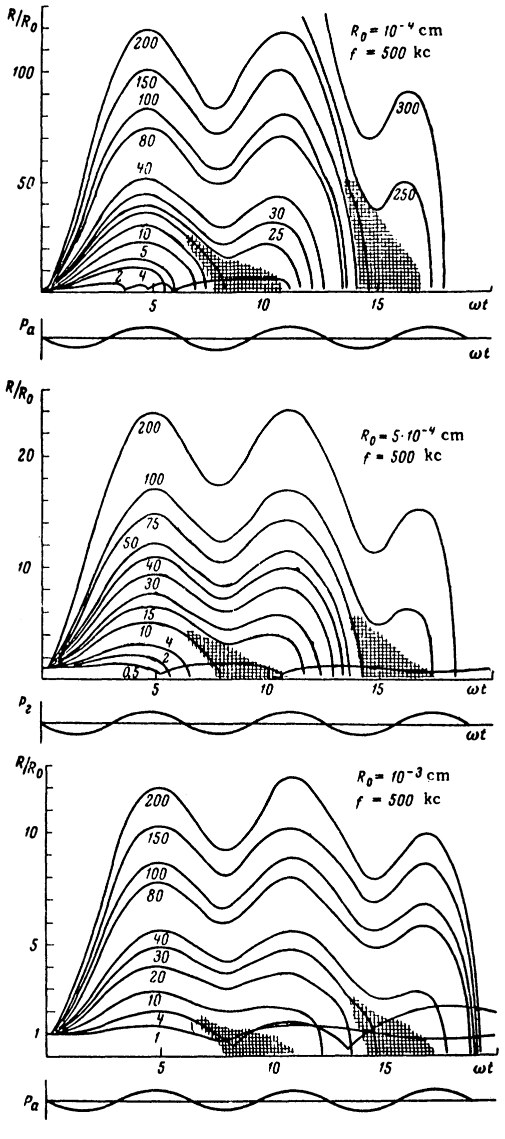

FIG. 11. Pulsation of cavitation bubble at 500 kHz for R_0 (initial bubble radius)$=10^{-4}$ cm. The numbers on the curves indicate the original sound-pressure amplitude in atmospheres. The sine wave at the bottom of the drawing indicates the sound signal (Akulichev, 1971).

while the broken curve indicates 500 kHz. The numbers on the individual curves indicate the value of the sound pressure p_m in atmospheres, while the sine wave at the bottom is the applied acoustic signal.

The threshold of vapor cavitation has been studied in a number of liquids. Galloway (1954) obtained the result that it takes a negative pressure of at least 200 atm to produce cavitation in degassed liquid, and this figure may still be low, for the existence of perfectly pure, degassed water is probably a fiction. To quote a worker in the field (Sirotyuk, 1971), "The production of absolutely pure water is impossible. One of the strongest of existing solvents, it dissolves the walls of the container, and on coming into contact with any gas, dissolves that gas."

Cavitation is also accompanied by the emission of visible radiation, which is called *sonoluminescence* (Kuttruff, 1962). It has been suggested that this light is due to the appearance of nonspherical bubbles in the cavitation which then develop positive and negative charges. These charges could then lead to an electrical microdischarge.

Such an explanation applies only to vapor cavitation, but the phenomenon has been observed in the case of gaseous cavitation also. Here the effect has been attributed to microshocks that develop in the process of bubble collapse. This would be the same phenomenon that produces luminescence in the case of cavitation bubbles produced by a continuous gas flow.

5. NONLINEAR INTERACTION OF SOUND WAVES

5.1 Two Intersecting Sound Beams in a Fluid

In a linear system, there can be no interaction of two intersecting sound beams. In line with the principle of superposition, the solution of two waves passing separately through a region will be equal to the solution of two waves passing simultaneously through the same region. No new terms will appear, and the interaction is therefore zero.

There is considerable evidence in support of the nonlinear interaction of two sound waves, but there still remains much disagreement on the basic governing relations. This stems from the fact that approximations or simplifications occur in every theory, and these may be sufficient to give erroneous results. Experimentally, the impossibility of obtaining two perfectly collimated beams, or of plane waves, raises the possibility that any interactions that are measured experimentally may be taking place on the face of the detecting instrument rather than in the "interaction region."

Ingard and Pridmore-Brown (1956) applied Lighthill's theory (Sec. 3) to the interaction of two perfectly collimated beams (Fig. 12) (frequencies ν_1, ν_2) and, neglecting viscous losses, obtained the following relation for the sum-frequency signal at a point outside the interaction region in the far field, defined by the radius r and angle θ:

$$p_+(r)=\frac{4\pi a^3}{\rho c_0^4}(\nu_1+\nu_2)^2\,\frac{p_1^{(I)}p_2^{(I)}}{r}$$

$$\times\left(\sin 2\theta+2\frac{B}{A}\right)\frac{\sin\alpha}{\alpha}\,\frac{\sin\beta}{\beta}. \tag{63}$$

B/A is the nonlinear parameter defined in Sec. 2, a is the radius of each source, and the angles α and β are defined by

$$\alpha=\frac{2\pi a}{c_0}[\nu_1-(\nu_1+\nu_2)\cos\theta],$$

$$\beta=-\frac{2\pi a}{c}[(\nu_1+\nu_2)\sin\theta-\nu_2].$$

These researchers performed experiments supporting these results, but their work has been criticized on the ground that the theory assumed perfect collimation of the beams, but the signals used in their experiments were relatively low, 10 and 100 kHz, so that there could be interaction of the two signals on the face of the microphone, wherever it was placed.

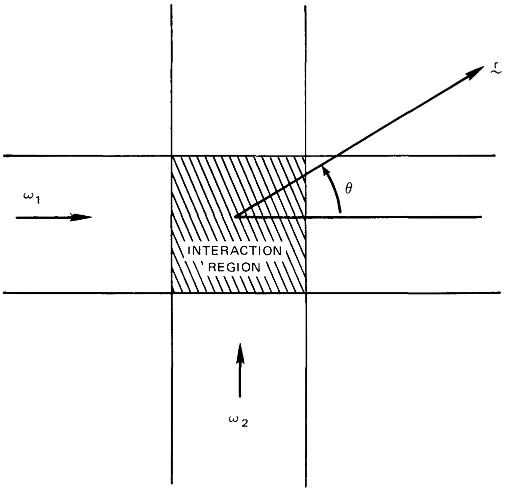

FIG. 12. Geometry for interaction of two crossed beams.

Westervelt (1957) used Lighthill's equation and deduced a relation for the lowest order in the scattering process:

$$\Box^2\rho_s=-\frac{1}{c_0^2}\frac{\partial^2}{\partial x_i\,\partial x_j}\left[\rho_0 u_i u_j+\frac{1}{2}\left(\frac{\partial^2 p}{\partial\rho^2}\right)_{\rho=\rho_0}\rho^2\delta_{ij}\right]. \tag{64}$$

Here $\Box^2$ is the D'Alembertian [Eq. (14)], and ρ_s is the density in scattered wave. Westervelt then recast Eq. (64) in the form for harmonic dependence of the initial waves:

$$\Box^2\rho_s=c_0^{-2}\Box^2E_{12}-\nabla^2(2T_{12}+qV_{12}), \tag{65}$$

where

$$T_{12}=\rho_0\mathbf{u}_1\cdot\mathbf{u}_2,\quad W_{12}=\left(\frac{c_0^2}{\rho_0\omega_1\omega_2}\right)\frac{\partial\rho_1}{\partial t}\,\frac{\partial\rho_2}{\partial t},$$

$$V_{12}=c_0^3\,\frac{\rho_1\rho_2}{\rho_0},\quad q=\frac{\rho_0}{c_0^2}\left(\frac{\partial^2 p}{\partial\rho^2}\right)_{\rho=\rho_0}, \tag{66}$$

$$E_{12}=T_{12}+V_{12},$$

$$\nabla^2V_{12}=\frac{1}{2}\,\frac{\omega_1^2+\omega_2^2}{\omega_1^2\omega_2^2}\,\Box^2W_{12}+\frac{1}{c_0^2}\,\frac{\partial^2T_{12}}{\partial t^2}.$$

In these relations ρ_1 and ρ_2 refer to the excess density in each of the primary beams, while T_{12}, V_{12}, and E_{12} are, respectively, the kinetic, potential, and total interaction energies.

For the case of two mutually perpendicular beams, considerable reduction takes place, and Eq. (65) becomes

$$\Box^2\rho_s=\Box^2\psi, \tag{67}$$

where

$$\psi=\frac{1}{c_0^2}\left[E_{12}-\frac{\rho_0}{2\omega_1\omega_2c_0^2}\left(\frac{\partial^2 p}{\partial\rho^2}\right)_{\rho=\rho_0}\nabla^2W_{12}\right]. \tag{68}$$

Westervelt concluded that $\rho_s = \psi$ is the complete solution of Eq. (61). Thus, if the two beams are perfectly collimated, W_{12} and E_{12} will vanish everywhere outside of the interaction region.

Westervelt's conclusions were challenged by Berktay (1969) who argued that the more general solution of Eq. (67) is

$$\rho_0=\psi+A, \tag{69}$$

where A is a homogeneous function satisfying the equation

$$\square^2 A=0. \tag{70}$$

On this basis, he obtained the following for the scattered pressure for the sum frequency $p_+(r)$ at distances that are large when compared with the interaction region:

$$p_+(r)=4\pi a^3 \frac{(\nu_1+\nu_2)^2}{\rho_0 c_0^4}\left(\frac{p_1(I)p_2(I)}{r}\right)\frac{B}{A} \times\frac{\sin\alpha}{\alpha}\,\frac{\sin\beta e^{-\alpha_+ r}}{\beta}, \tag{71}$$

which differs but slightly from the results of Ingard and Pridmore-Brown. A similar result was obtained by Zverev and Kalachev (1970).

5.1.1 Nondispersive and Dispersive Media. The disagreements regarding the interaction of two sound beams referred to above may be due to a failure to distinguish between two cases, nondispersive and dispersive media. In the first case, where there is no dependence of the sound velocity on frequency, the interaction of two noncollinear beams would appear to be impossible. But in the second case, there is substantial evidence that such interaction is possible at oblique angles between the beams. In particular, the use of bubbly media, where dispersion effects are very strong, makes possible a much larger interaction between two primary beams.

5.2 Two Collinear Sound Beams

In subsequent papers, Westervelt (1960, 1963) derived an expression for the interaction that would result if the two beams were collinear and perfectly collimated. He deduced an equation for the D'Alembertian of the scattered pressure for two such beams:

$$\square^2 p=-\rho_0\frac{\partial q}{\partial t}, \tag{72}$$

where

$$q=\frac{1}{\rho_0^2 c_0^4}\left(1+\frac{B}{2A}\right)\frac{\partial}{\partial t}p_i^2. \tag{73}$$

The quantity q is the simple source strength density resulting from the primary waves ρ. The solution to Eqs. (72) and (73) involves the appearance of two new components in the scattered signal, with frequencies equal to the sum and difference, respectively, of the two original frequencies.

For the case of two primary waves of the same pressure amplitude but different frequencies, Westervelt obtained the following expression for the scattered intensity in the far field (the geometry is shown in Fig. 13):

$$I_s=\frac{\omega_s^4\,P_0^4\,S_0^4(1+B/2A)^2}{2(8\pi)^2\rho_0^3\,c_0^9\,R_0^2}\,\frac{1}{\alpha^2+k_s^2\sin^4\theta/2}. \tag{74}$$

These latter results have been well confirmed experimentally (Bellin and Beyer, 1962).

One of the distinguishing characteristics of this interaction is the almost complete absence of sidelobes. The width of the main beam can be determined by Eq. (68).

5.3 Parametric Arrays

The success of the Westervelt collinear beam analysis, to which experimental arrangement he gave the name of "parametric array," has led to a large amount of research on the application of the array to the problems of transmission and reception of acoustic signals, especially in underwater sound. Of greatest interest has been work in the far field.

5.3.1 Far-Field Transmission. The following expression, due to Berktay (1965), describes the far field of a rectangular transducer radiating two beams of initial pressure amplitudes P_1 and P_2 (Fig. 14):

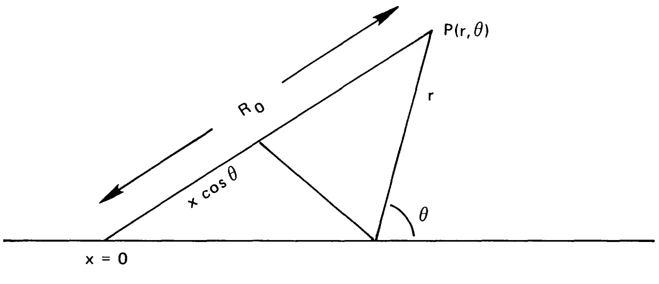

FIG. 13. Geometry for parametric array source.

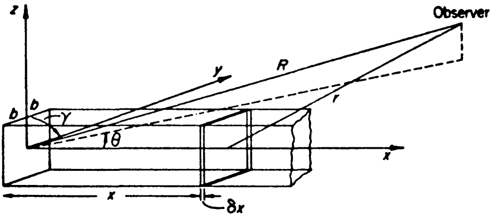

FIG. 14. Geometry for two collinear interacting sound beams traveling in the *x* direction (Berktay, 1965).

$$P_s=\frac{\beta P_1P_2S\omega_s^2\exp(-\alpha_s R)\psi(\gamma,0)}{4\pi\rho_0c_0^4R_0[A^2+4k_s^2\sin^2(\theta/2)]^{1/2}},\tag{75}$$

where

$$S=(2b)(2d),$$

$$A=\alpha_1+\alpha_2-\alpha_s\cos\theta,$$

$$\psi(\gamma,\theta)=\frac{\sin(dk_s\cos\gamma)}{dk_s\cos\gamma}\,\frac{\sin(bk_s\sin\gamma\sin\theta)}{bk_s\sin\theta\sin\gamma}.\tag{76}$$

A somewhat similar form can also be developed for the circular transducer. Experimental results indicate good agreement with the theory. Some typical experimental examples are shown in Fig. 15.

An important characteristic of the parametric array is the absence (or great reduction) of sidelobes. It is therefore useful as a search tool in situations where sidelobes on the beam can prevent accurate detection. This is especially the case in subbottom investigations in the shallow ocean (see Sec. 7).

5.3.2 Far-Field Reception. The parametric array can also be used as a receiving array in which an incoming wave interacts nonlinearly with a local sound source (the pump) and the scattered sum or difference frequency is detected. One such system is shown in Fig. 16 (Barnard *et al.*, 1972). The first-order field consisted of a spherical-wave source of angular frequency ω_1 and a plane-wave source of a lower frequency ω_2 ($\omega_1>\omega_2$). The spherical wave emanated from a baffled circular piston, and both sources were assumed to have harmonic time dependences. It is further assumed that (i) linear-field effects are disregarded; (ii) while absorption of the spherical wave is taken into account, the absorption in the plane wave is neglected; and (iii) the amplitudes of the primary waves are such that losses in the primary beams due to finite-amplitude effects can be neglected. The analysis began with the Westervelt equation for the scattered wave:

$$p_s(R_0,t)=-\frac{\rho_0}{4\pi}\int\frac{\partial q}{\partial t}\frac{\exp[i(k_s+i\alpha_s)]}{|\mathbf{R}_0-\mathbf{r}|}\times|\mathbf{R}_0-\mathbf{r}|\,dV,\tag{77}$$

where q is the source strength and the subscript s refers to quantities in the scattered wave. (The absorption factor was added to the Westervelt formula to take approximate account of the absorption of the secondary radiation.)

With $q=(\beta/\rho_0^2c_0^2)(\partial p_f^2/\partial t)$, where p_f is the first-order sound field p_f, the pressure p is given by

$$p_f=p_1\frac{r_0}{r}\left(\frac{2J_1(k_1a\sin\phi)}{k_1a\sin\phi}\right)e^{-\alpha_1 r}\cos(\omega_1t-k_1r)+p_2\cos(\omega_2t-k_2k),\tag{78}$$

and the scattered field becomes

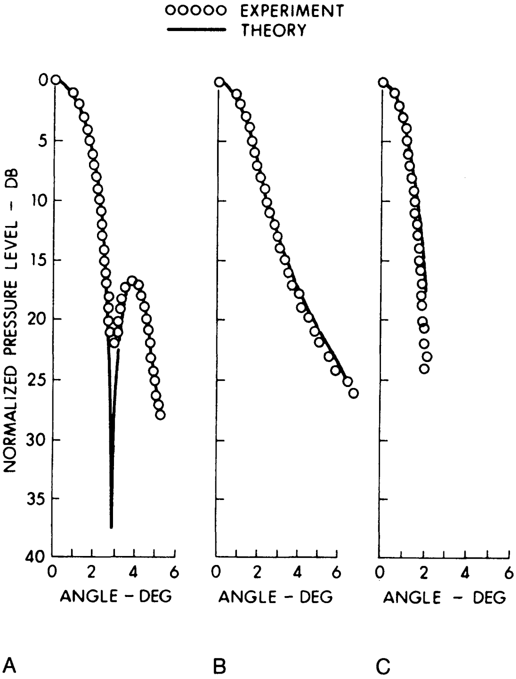

FIG. 15. Beam patterns for carrier and sum- and difference-frequency components in a transmitting array. (a) 482-kHz carrier; (b) 64-kHz difference-frequency radiation, and (c) 900-kHz sum-frequency radiation (Muir and Willette, 1972).

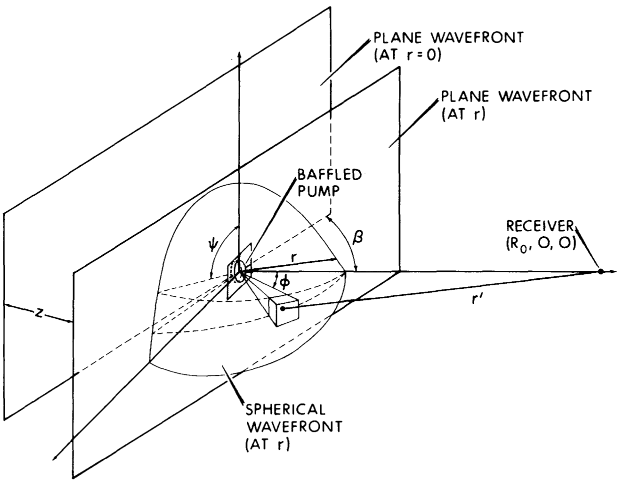

FIG. 16. Orientation of the plane- and spherical-wave primaries in the spherical coordinate system (r, ψ, φ) (Barnard *et al.*, 1972).

$$p_s \simeq \frac{\beta\omega_0^2 p_1 p_2 r_0}{2\rho_0 c_0^4 a k_1} \int_0^{R_0} \int_0^{\phi_{\text{eff}}} J(k_1 a \sin\theta) \times J_0(k_2 r \sin\theta \cos\gamma) \times \exp[i(k_1 \pm k_2 \sin\gamma \cos\theta + i\alpha_1) r] \times \frac{\exp[i(k_s + i\alpha_s) r]}{r'} \, r \, d\phi \, dr, \tag{79}$$

where $\omega = \omega_1 + \omega_2$ and the upper limit on the q integration, ϕ_{eff}, is the angle from the axis of the acoustic beam to the first zero of the pump far-field radiation pattern. Equation (79) has been solved numerically and the results tested experimentally for the sum frequency. Some results are shown in Fig. 17, while Fig. 18 shows a typical detection system.

6. NONLINEAR EFFECTS IN SOLIDS

6.1 The Scope of Nonlinear Behavior

In an ideally elastic medium, propagating waves are represented by functions that are solutions of the linear wave equation and therefore obey the principle of superposition. Even a very small-amplitude wave in a real solid, however, will only approximate this linear behavior. Fundamental nonlinearities are assumed to be negligible in order to derive the linear wave equation. It is a common occurrence, however, that elastic waves in solid do in fact interact with one another. Quantum mechanically, one can speak of phonon-phonon interactions. Such familiar phenomena as lattice thermal resistivity, thermal expansion, and the maintenance of an equilibrium distribution of thermal phonons are all possible only if the lattice waves behave in a nonlinear way. Thus much of nonlinear acoustical effects in solids is studied as a part of solid-state physics. Such subjects will not be treated here.

6.2 Third-Order Elastic Constants

The fundamental relations for strains in a solid are given by the strain tensor u_{ij}, whose components are

$$u_{ij} = \frac{1}{2}\left(\frac{\partial \xi_i}{\partial x_j} + \frac{\partial \xi_j}{\partial x_i} + \frac{\partial \xi_l}{\partial x_i}\frac{\partial \xi_l}{\partial x_j}\right). \tag{80}$$

If one deals with small deformations only, the linear Hooke's law of elasticity applies; this can be written in tensor form as (summation convention is used)

$$T_{ij} = c_{ijkl} u_{kl}. \tag{81}$$

When the direction of propagation is a pure-mode direction, one longitudinal and two transverse polarizations exist. The particle displacements are either parallel or perpendicular to the propagation direction. In other, nonpure-mode directions, the branch with the greatest phase velocity is usually identified as the longitudinal branch.

The elastic potential energy of a solid is given by

$$V = \tfrac{1}{2} c_{ijkl} \epsilon_{ij} \, \epsilon_{kl}. \tag{82}$$

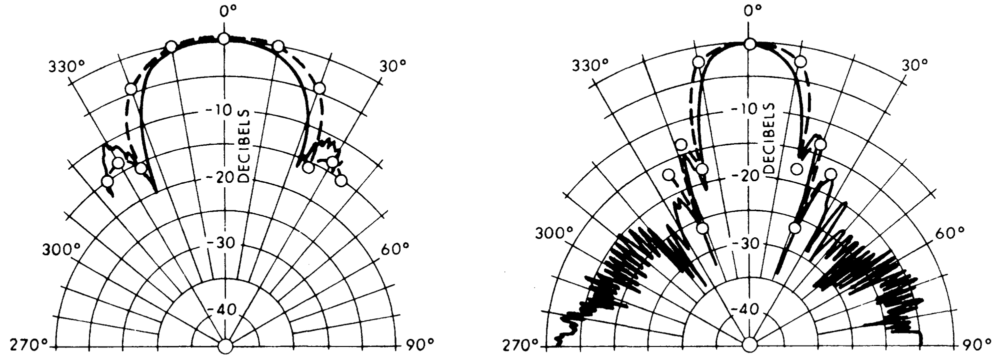

f_s = 91 kHz
EXPERIMENT: θ_{HP} = 27.5°
THEORY: θ_{HP} = 35.0°
SIGNAL FREQUENCY: 1 kHz

f_s = 92.8 kHz
EXPERIMENT: θ_{HP} = 17.5°
THEORY: θ_{HP} = 21.0°
SIGNAL FREQUENCY: 2.8 kHz

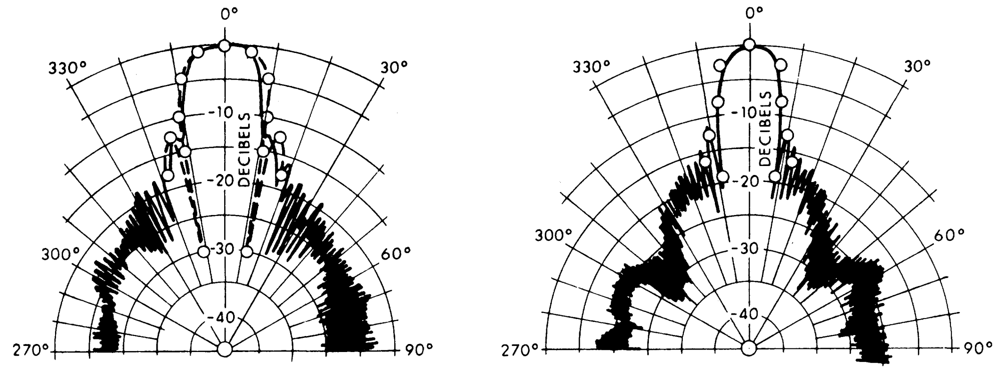

f_s = 95 kHz
EXPERIMENT: θ_{HP} = 13.5°
THEORY: θ_{HP} = 16.0°
SIGNAL FREQUENCY: 5 kHz

f_s = 100 kHz
EXPERIMENT: θ_{HP} = 10.5°
THEORY: θ_{HP} = 11.0°
SIGNAL FREQUENCY: 10 kHz

FIG. 17. Sum-frequency patterns for a pump-receiver separation of 48 ft. Theoretical values, broken lines with circles; experimental values, solid lines; F=90 kHz. P_1=101 dB *re*:1 μbar at 1 yd, and P_2=85 dB *re*:1 Mbar at input to parametric receiving array (Barnard *et al.*, 1972).

The quantities c_{ijkl} are commonly called the second-order elastic constants, for the reason that they appear in an expression that is second order in the strains. Nevertheless, they represent linear elastic phenomena. If we expand the internal energy $U(S,\xi)$ about the point of zero strain $\epsilon=0$ in the isentropic case, we obtain

$$\rho_0 U(S,\xi) - \rho_0 U(S,0) = \frac{1}{2!} c_{ijkl}\epsilon_{ij}\,\epsilon_{kl} + \frac{1}{3!} c_{ijklmn}\epsilon_{ij}\,\epsilon_{jk}\,\epsilon_{mn} + \cdots, \qquad (83)$$

where the third-order elastic constants are defined by

$$c^S_{ikl;\cdots} = \rho_0 \left(\frac{\partial^n U}{\partial\epsilon_{ij}\,\partial\epsilon_{kl}\cdots}\right)_S. \qquad (84)$$

[In the case in which Hooke's law is valid, only the first term on the right-hand side of Eq. (83) differs from zero, and the left-hand side becomes the elastic potential energy.] Because of inherent symmetries in crystal systems, it is possible to simplify the subscripts of the elastic constants. In particular, the six subscripts of the third-order constants can be

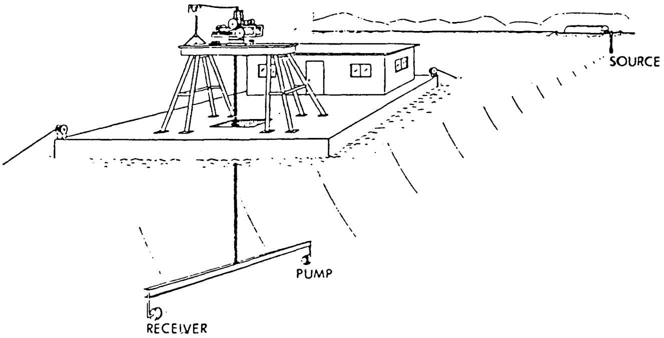

FIG. 18. Configuration for parametric reception. The source is located at some great distance, while the pump and receiver can be rotated for best reception (Barnard *et al.*, 1972).

reduced to three. In the case of a cubic crystal, there are three independent second-order constants and six independent third-order constants. If one is dealing with an isotropic solid, the representation is the same as for a cubic crystal, but with additional relationships that reduce the number of independent constants to two in the second-order constants and three in third order. Equation (83) then becomes

$$\begin{aligned}\rho_0 U(S,\xi)-\rho_0 U(S,0)\\ &=\tfrac{1}{2}c_{12}\epsilon_{ii}\,\epsilon_{jj}+c_{44}\epsilon_{ij}\,\epsilon_{ij}+\tfrac{1}{6}C_{123}\epsilon_{jj}\,\epsilon_{kk}\\ &\quad+C_{144}\epsilon_{jj}\,\epsilon_{ik}\,\epsilon_{ik}+\tfrac{4}{3}C_{456}\epsilon_{ik}\,\epsilon_{ij}\,\epsilon_{kj}\\ &\quad+\text{higher-order terms.} \qquad (85)\end{aligned}$$

Considerable attention has been paid to these third-order constants, and they have been evaluated for a large number of solids through the equation of motion

$$\frac{\partial^2 u}{\partial t^2}=\frac{1}{\rho_0}\frac{\partial^2 u}{\partial x^2}\left(K_2+(K_3+3K_2)\frac{\partial u}{\partial x}\right.$$
$$\left.+\ \text{higher-order terms}\right), \qquad (86)$$

where K_2 and K_3 are given in Table 3 for various crystalline directions.

Acoustic nonlinearities in solids are due mainly to two sources. First is the anharmonicity due to a departure of the medium from the simple linear behavior of Hooke's law, and the second is that of dislocation displacements. The first of these leads primarily to a small shift in the sound propagation velocity. The accurate measurement of this velocity therefore provides for the measurement of the third-order elastic constants.

The effect of dislocation displacements is associated with the point defects present in any crystal. A dislocation is a linear imperfection in a crystal. Such dislocations move under stress within a crystal, facilitating the propagation of lattice vibrations that are sound waves. The motion of the dislocation is hindered by the presence of point defects and by other dislocations. Internal stresses surrounding dislocations and point defects are both relieved by their interactions so that they tend to attract one another, causing the dislocations to be pinned at certain points.

In passing through a real crystal, a sound wave will be attenuated by interaction with the dislocations. The theory for such dislocation damping is based on the concept of the motion of dislocations between the two pin-

Table 3. Values of K_2 and K_3 for the [100], [110], and [111] directions.

Direction	K_2	K_3
[100]	C_{11}	C_{111}
[110]	$\dfrac{C_{11}+C_{12}+C_{14}}{2}$	$\dfrac{C_{111}+3C_{112}+12C_{166}}{4}$
[111]	$\dfrac{C_{11}+2C_{12}+4C_{44}}{3}$	$\dfrac{C_{111}+6C_{112}+12C_{144}+24C_{166}+2C_{123}+16C_{456}}{9}$

ning points as analogous to the motion of a vibrating string equal in length to the distance between the pinning points. The small-amplitude oscillations of such dislocation loops lead to an attenuation that has been well confirmed by experiment (see the work of Granato and Lücke, 1956).

As the amplitude of the oscillations becomes larger (Fig. 19), the dislocation loop will break away from many of the points, producing the bowing shown in the latter stages of the drawing. The results of the analysis of this phenomenon indicate that dislocation displacements can produce second harmonics that are comparable to those resulting from lattice inharmonicity, while the corresponding contribution to the third harmonic is enormously greater. Experimental investigation of the nonlinear behavior of dislocations is therefore most conveniently carried out by use of the third harmonic (Hikata and Elbaum, 1966).

6.3 Interaction of Sound with Sound

The phenomenon of the interaction of sound with sound also takes place in solids. A summary of possible interaction cases is shown in Table 4 (after Jones and Kobett, 1963). Many of these combinations have been confirmed experimentally.

6.4 Nonlinear Surface Waves

Surface waves on solids have been widely studied in acoustics. A typical transducer arrangement is shown in Fig. 20 (Kino and Matthew, 1971a; see also de Klerk, 1972). Detailed studies reveal that whenever the amplitudes of the surface waves become appreciable, nonlinear effects appear, paralleling those described for bulk waves. The nonlinear mixing of two waves of different frequency leads to combination waves (Luukala and Kino, 1971), while the correlation function of two signals can also be obtained (Quate and Thompson, 1970). Some applications are discussed in Sec. 7.

7. RECENT APPLICATIONS

7.1 General View

There are numerous applications of nonlinear acoustics to practical problems. Indeed, to some, the entire field of nonlinear acoustics is one of applied physics, and we have already mentioned a number of applications—for example, the use of the radiation pressure to measure the intensity of a sound. Below are given several more recent applications.

7.2 Radiation Pressure

The acoustic radiation pressure described in Sec. 4, though relatively small in comparison with the pressure amplitudes of intense sound waves, is sufficient to levitate small objects such as glass spheres and coins (Allen and Rudnick, 1947). An example of such levitation, using a high-intensity siren, is shown in Fig. 21 (Barmatz, 1982).

In an orbiting spacecraft, it is not possible to obtain a zero-gravity environment, but the acoustic radiation pressure can be used to establish and maintain such a situation. A three-dimensional, piezoelectrically controlled mechanism is shown in Fig. 22 (Wang *et al.*, 1974; Wang, 1986). The chamber is cubic. If the x-axis transducer excites the fundamental

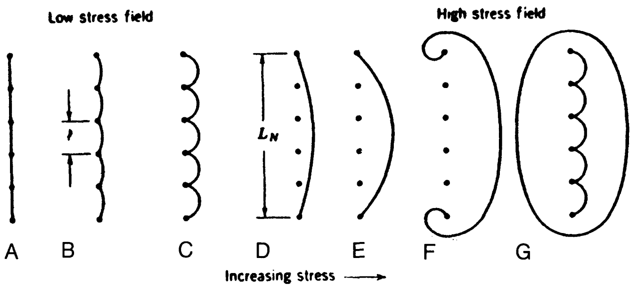

FIG. 19. Bowing out of a pinned dislocation line by an increasing applied stress. The loop length, determined by the impurity pinning, is l and the network length is L_N. At (d) the stress is large enough to cause the dislocation to break away from the impurity pinning sites (Granato and Lücke, 1956).

Table 4. Interaction cases which produce a scattered wave.[a]

Primary waves	Resonant wave type and frequency	Direction of scattered wave	$\cos\phi$ [b]	Frequency limits[c]
Two transverse	Longitudinal $(\omega_1+\omega_2)$	$\mathbf{k}_1+\mathbf{k}_2$	$c^2+\frac{(c^2-1)(a^2+1)}{2a}$	$\frac{1-c}{1+c}<a<\frac{1+c}{1-c}$
Two longitudinal	Transverse $(\omega_1-\omega_2)$	$\frac{\mathbf{k}_1-\mathbf{k}_2}{\omega_1-\omega_2}$	$\frac{c^2}{1}+\frac{(c^2-1)(a^2+1)}{2ac^2}$	$\frac{1-c}{1+c}<a<\frac{1+c}{1-c}$
One longitudinal and one transverse[d]	Longitudinal $(\omega_1+\omega_2)$	$\mathbf{k}_1+\mathbf{k}_2$	$c+\frac{a(c^2-1)}{2c}$	$0<a<\frac{2c}{(1-c)}$
One longitudinal and one transverse[d]	Longitudinal $(\omega_1-\omega_2)$	$\frac{\mathbf{k}_1-\mathbf{k}_2}{\omega_1-\omega_2}$	$c+\frac{a(1-c^2)}{2c}$	$0<a<\frac{2c}{(1+c)}$
One longitudinal and one transverse[d]	Transverse $(\omega_1-\omega_2)$	$\frac{\mathbf{k}_1-\mathbf{k}_2}{\omega_1-\omega_2}$	$\frac{1}{c}+\frac{(c^2-1)}{2ac}$	$\frac{1-c}{2}<a<\frac{1+c}{2}$

[a] From Jones and Kobett (1963).

[b] ϕ is the angle between $\mathbf{k}_1$ and $\mathbf{k}_2$ at resonance, a is the frequency ratio ω_2/ω_2, and c is the velocity ratio c_{tr}/c_{long}.

[c] When a is within the limits shown, it is possible to choose an angle ϕ that will give a scattered wave.

[d] The frequency of the longitudinal primary wave is ω_1.

acoustic resonance in the x direction, there would be a minimum radiation force in the x direction along the midplane. If the y-axis transducer excites the corresponding resonance in the y direction, the force minimum would be along the line of intersection of the two midplanes. By exciting the z-axis transducer, the particle could be constrained to the central point of the cube.

Such manipulations of the acoustic field make it possible to maintain a sample in the center of the cube without its being in contact with any support structure, so that containerless processing of materials becomes possible (Barmatz, 1982).

7.3 Parametric Array

The techniques described in Sec. 5 for parametric arrays have been applied in a number of ways in searching the bottom and

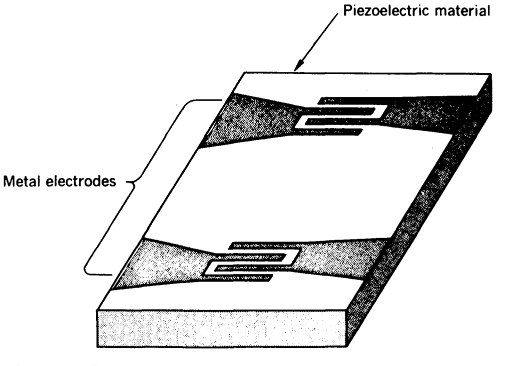

FIG. 20. Schematic representation of a typical piezoelectric surface-wave device with interdigital electrodes (Kino and Matthews, 1971).

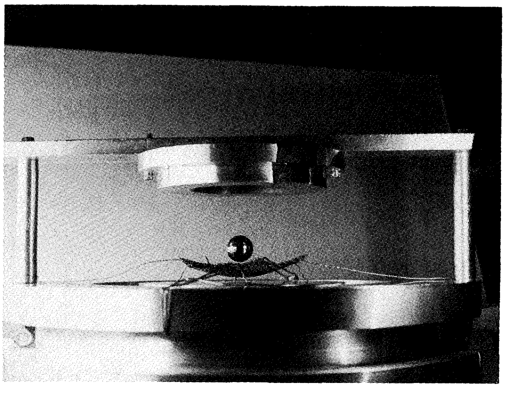

FIG. 21. Levitation of a steel ball in a ground-based siren. The diameter of the ball is 1.9 cm (Barmatz, 1982).

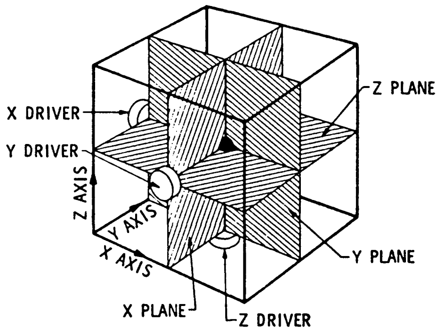

FIG. 22. Triple-axis acoustic levitator. The sample is positioned at the center by the simultaneous excitation of three orthogonal sound waves (Wang *et al.*, 1974).

subbottom layers of relatively shallow waters. Detailed descriptions of a number of systems appear in Novikov *et al.* (1987). Figure 23 shows the use of a difference-frequency beam (with no sidelobes) observing the bottom of a portion of the Gulf of Riga (Novikov *et al.*, 1987). Figure 24 shows an object buried in the mud sludge on the bottom of the Semevskii Lake, as well as a stone falling in that sludge, while Fig. 25 shows the detection of fish schools swimming in shallow waters. In all these cases, the parametric array beam has the advantage of a narrow, single lobe and a low frequency, so that its absorption is not a serious problem. In the several applications described above, the operating difference frequency ranged from 10 to 50 kHz, with half-angle beamwidths of 3°–4°.

7.4. Lithotripsy

An important recent application of shock waves (Sec. 3) has been in the field of lithotripsy, i.e., the use of focused shock waves to break up kidney and other stones embedded in body organs. The procedure for breaking up gallstones is shown in Fig. 26 (Sauerbrach *et al.*, 1986). Spark-produced shock pulses are focused by an ellipsoidal system into a patient in a water bath. The conditions within the gallbladder are monitored by acoustic imaging or, alternatively, by means of x rays. Similar systems have been used for kidney stones (Coleman *et al.*, 1987). In more refined systems, the shock pulses are generated electromechanically and focused by means of an

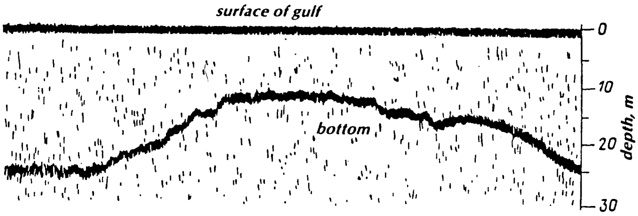

FIG. 23. Echogram of the bottom profile of a region of the Gulf of Riga, obtained by use of a parametric array system (Novikov *et al.*, 1987).

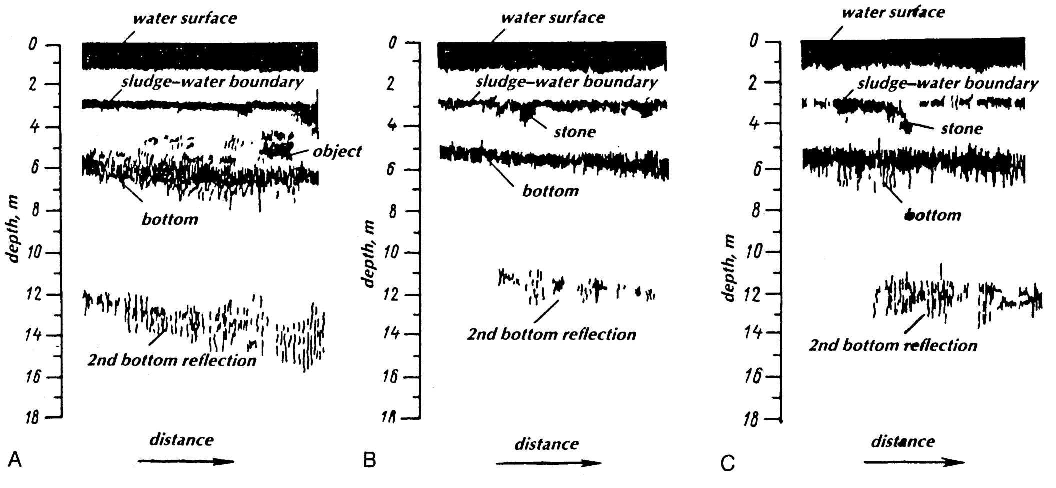

FIG. 24. Echograms of the bottom of Semevskii Lake, showing reflections from (a) a buried object, and (b), (c) a stone in different stages of sinking. The echograms were obtained with a parametric array system (Novikov *et al.*, 1987).

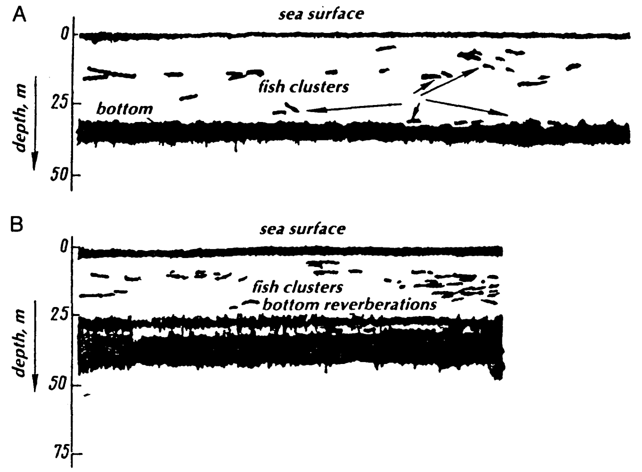

FIG. 25. Echograms of the bottom and of fish clusters, obtained with (a) a parametric array system and (b) a commercially available fish locator under conditions of bottom reverberation (Novikov *et al.*, 1987).

acoustic lens. Piezoelectric elements on a concave spherical shell have also been used and the need for a water bath eliminated (Crum *et al.*, 1989).

In these systems, the pressure signal that impinges on the stone consists of a positive pulse of about 1 μs in length, followed by a longer, negative trough (5–10 μs duration). The amplitude of a pulse may be as high as 100 MPa for the positive section, while the negative section reaches up to no more than 5 MPa. In a typical procedure, 1500 pulses will be directed on the stone.

7.5 *B/A* in Biological Tissues

Another biological application can be found in recent detailed studies of the nonlinear parameter *B/A* (Sec. 2) in biological tissues. Recently, Zhang and Dunn (1987) have measured *B/A in vitro* and *in vivo* for cat livers by measurement of the second-harmonic production, as suggested in Sec. 2. A typical setup is shown in Fig. 27. Their preliminary results suggest that there exists no significant difference in the two sets of measurements, which should open up the future for studies of tissue characterization by measurement of *B/A in vitro.*

Extensive measurements have been made on a variety of human tissues *in vitro*, using the thermodynamic method described in Sec. 2 (Errabolu *et al.*, 1988). A small but definite difference was noted by this same group (Seghal *et al.*, 1984) in the nonlinear parameter *B/A* in cancerous human liver tissues, so that evaluation of *B/A* in tissue samples may prove to be of diagnostic value, although the evidence so far is rather slight.

7.6 Surface Acoustic Waves (SAW)

The principal application of nonlinear SAWs (Sec. 6) is in its use as a convolver of two signals. The principle is illustrated in Fig.

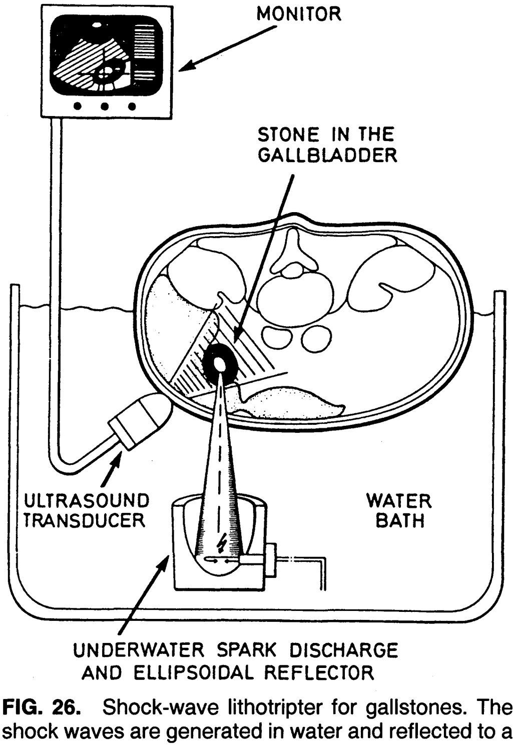

FIG. 26. Shock-wave lithotripter for gallstones. The shock waves are generated in water and reflected to a focal point where the stone has been positioned (Sauerbruch *et al.*, 1986).

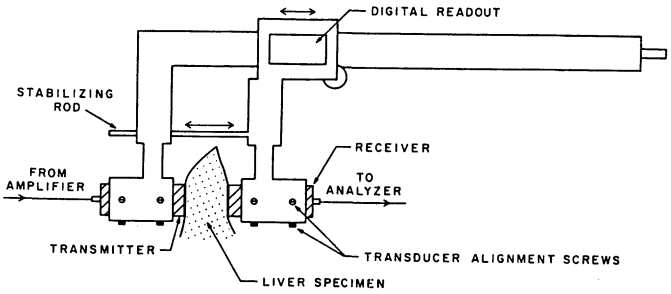
DIGITAL READOUT
STABILIZING ROD
RECEIVER
FROM AMPLIFIER
TO ANALYZER
TRANSMITTER
TRANSDUCER ALIGNMENT SCREWS
LIVER SPECIMEN

FIG. 27. Transducer-caliper assembly used in *in vivo B/A* measurements. The digital readout gives the separation distance between the transmitter and the receiver (Zhang and Dunn, 1987).

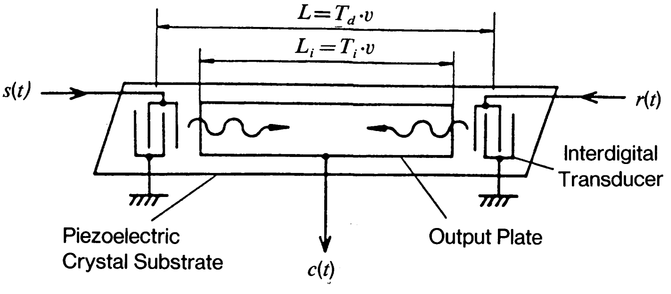
$L = T_d \cdot v$
$L_i = T_i \cdot v$
$s(t)$
$r(t)$
Interdigital Transducer
Piezoelectric Crystal Substrate
Output Plate
$c(t)$

FIG. 28. Principle of the surface-acoustic-wave (SAW) convolver. The two signals interact on the piezoelectric crystal substrate and are detected in the interval L_i (Grassl *et al.*, 1988).

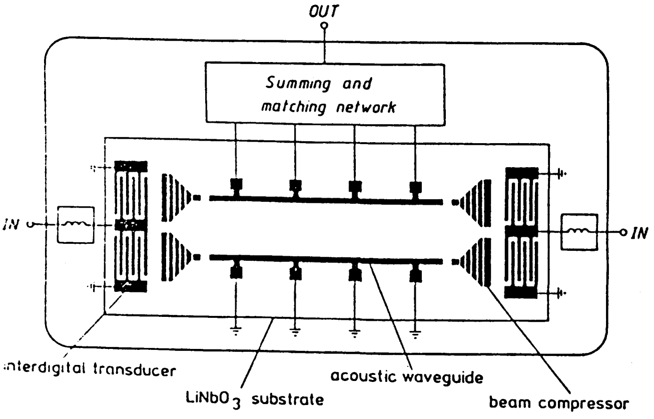
OUT
Summing and matching network
IN
IN
interdigital transducer
LiNbO$_3$ substrate
acoustic waveguide
beam compressor

FIG. 29. Schematic of efficient elastic convolver with self-convolution suppression (Grassl *et al.*, 1988).

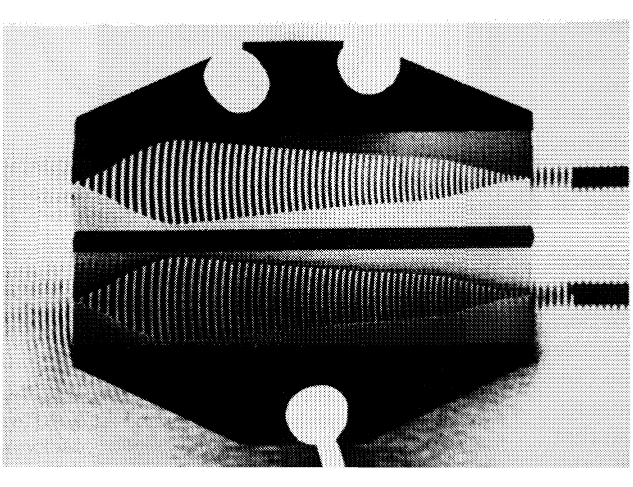

FIG. 30. Small-aperture focusing chirp transducers feeding three-wavelength-wide strip waveguides (Grassl *et al.*, 1988).

28 (Grassl *et al.*, 1987). Two SAWs are introduced onto a piezoelectric substrate by means of two interdigital transducers: the signal $s(t)$ and the reference $r(t)$. The two waves interact as described in Sec. 6, and the resultant signal $c(t)$ is detected by a third transducer. Analysis shows that the linear signals and the quadratic terms from each of the two original signals make a negligible contribution to the detecting transducer, but that the cross term remains. This term can be written in the form

$$c(t/2+T_d/2)=C_{\mathrm{NL}} + \int_{(t-T_i)/2}^{(t+T_i)/2} s(t-\tau)r(\tau)d\tau, \qquad (87)$$

where T_d, T_i, L, and L_i are defined in the figure, v is the velocity of the surface wave, and C_{NL} is a constant of integration. The integral is recognized as the convolution of the two signals.

The nonlinearity can be produced by interaction of the wave fields with free charge carriers in semiconductor substrates, in which case the system is called an acoustoelectric convolver. The nonlinearity can also be produced by taking advantage of the nonlinearity of the piezoelectric substrate. This is the so-called elastic convolver. The operation of an elastic convolver is shown in Fig. 29. The two parallel paths are used to suppress reflections at the transducers: The transducers are connected acoustically in phase at one end and with opposite phase at the other, so that the electrical responses in the connected transducers cancel. More recently, small-aperture, focusing chirp transducers have been employed. These have enough degrees of freedom so that the transfer function can be shaped to compensate for various distortions in the system. A sample of such transducers is shown in Fig. 30.

GLOSSARY

Acoustic Cavitation: The creation of holes (or gas bubbles) in a liquid, produced by the passage of intensesound waves.

Acoustic Mach Number: The ratio of the particle displacement velocity in a sound wave to the local sound velocity.

Acoustic Particle: A hypothetic small volume in a fluid, considered as a homogeneous medium, and subject to the forces that give rise to sound waves.

Acoustic Streaming (also called Quartz Wind and Hydrodynamic Flow): The one-directional flow of a fluid produced by the passage of a sound wave in it.

Aeolian Harp: An ancient musical instrument involving loose strings that can be driven by the flow of air past them.

Ballou's Rule: Empirical formula [Eq. (23)] connecting values of B/A with the sound velocity.

Burgers' Equation: An approximate equation [Eq. (18)] describing wave motions in a coordinate system that travels with the speed of sound.

Collimated Beam: A beam in which none of the radiation escapes from a limited, axial and near-axial region.

Convolution Integral: The integral of the product of two functions varying in the manner prescribed by Eq. (87).

Dipole Source: Source in which the monopole sources vanish, but positive and negative fluctuations within the system lead to sound radiation. Force field is equivalent to a distribution of dipole sources.

Discontinuity Distance: The distance from the source at which the steepness of the wave front becomes negatively infinite at the point at which the displacement velocity is zero.

Displacement Velocity: The time rate of change of the position of an acoustic particle in a sound wave.

Eulerian Coordinates: The system of coordinates in which one follows the behavior of the medium at a fixed point in the medium.

Far Field: Points in the sound field whose distance from the source is greater than the square of the size of the source divided by the wavelength. It is equivalent to the region where Fraunhofer diffraction prevails.

Fubini Solution: Explicit solution of the Earnshaw equation for propagation of finite-amplitude waves in a nondissipative medium.

Goldberg Number: Ratio of the extinction distance (reciprocal of the absorption coefficient) to the discontinuity distance in a finite-amplitude wave.

Herring–Flynn Equation: Equation describing bubble oscillations in an acoustic field in which the compressibility of the liquid is taken roughly into account.

In Vitro: Measurements on biological materials made outside the living organism.

In Vivo: Measurements on biological materials under natural conditions in the living organism.

Khokhlov–Zabolotskaya (KZ) Equation: Modification of Burgers' equation, taking diffraction effects into account.

Khokhlov–Zabolotskaya–Kuznetsov (KZK) Equation [Eq. (22)]: Modification of Burgers' equation, taking both diffraction and dissipation into account.

Kundt's Tube: A historic piece of apparatus, consisting of a glass tube, plugged at its ends, with an axial centered metal rod. The rod is excited by stroking to produce sound, and dry dust particles placed in the tube will be moved to accumulate at the nodes of the resulting standing waves.

Lagrangian Coordinates: The system of coordinates in which one follows the behavior of the medium at a point in the medium that was initially at rest at the point x.

Lagrangian Radiation Pressure: The difference between the mean pressure at an absorbing or reflecting wall and the pressure in the same acoustic medium at rest, behind the wall.

Levitation: The raising of an object against the pull of gravity.

Lighthill Equations: Equations of motion [Eqs. (39)- (42)] in which the fluctuating fluid flow in a small part of a system is used as the forcing term, determining the wave field throughout the system.

Monopole Source: Source of acoustic radiation that consists of an oscillating spherical surface. System involves matter sources.

N Waves: Finite-amplitude waves in which the wave profile (particle velocity versus time) resembles the letter N.

Near Field: Points in the sound field where the distance from the source is less than the square of the size of the source divided by the wavelength. It is equivalent to the region in which Fresnel diffraction prevails.

Noltingk–Neppiras Equation: Equation describing bubble oscillations in an acoustic field, in which the liquid is assumed to be incompressible.

Nonlinearity Coefficient β: The coefficient of correction terms when nonlinearity is taken into account:

$$\beta=1+\rho_0\left(\frac{\partial^2 p}{\partial\rho^2}\right)\bigg/2\left(\frac{\partial p}{\partial\rho}\right),$$

where the derivatives are taken at constant entropy and infinitesimal amplitude.

Parameter of Nonlinearity, B/A: The ratio of the coefficients of the quadratic to the linear term in the expansion of pressure changes in a power series of the density changes.

Parametric Array, Endfire: The interaction of two collinear finite-amplitude sound beams to produce a beam at the difference position.

Point Defect: A local irregularity in a crystal, viz., a vacancy (missing atom), a foreign atom (impurity), or an atom in an irregular position.

Quadrupole Source: Source in which both monopole and dipole sources vanish, but the fluctuations of pressure within the system can be equivalent to a distribution of quadrupoles.

Radiation Pressure: The time-independent pressure at a point resulting from the passage of a sound wave.

Rankine–Hugoniot Equation [Eq. (24)]: Fundamental equation of fluid motion, connecting the properties of the fluid ahead of a shock with those behind the shock.

Rayleigh Radiation Pressure: The difference between the average pressure at a surface and the pressure that would have existed in the fluid of the same density at rest.

Sonoluminescence: The emission of visible radiation during acoustic cavitation.

Sound-Pressure Level (SPL): Logarithmic scale of measuring sound pressure relative to a specified level:

$$\text{SPL}=20\ \log(p/p_0),$$

where p is the amplitude of the observed pressure, and p_0 is an arbitrarily chosen reference level, usually taken to be equal to 20 μPa.

Viscosity Number b: The ratio of the total viscosity effect in a material to the shear viscosity, as defined in Eq. (21). It is often generalized to include the effects of thermal conduction as in Eq. (27).

Vortex Shedding: When air flows past a solid edge, as in an airfoil or the reed of a musical instrument, small eddies or vortices appear at the edge and detach themselves from it, being carried away by the fluid flow.

List of Works Cited

Akulichev, V. A. (1971), in: L. D. Rozenberg (Ed.), *High Intensity Ultrasonic Fields*, New York: Plenum, pp. 203–259.

Allen, C. R., Rudnick, I. (1947), *J. Acoust. Soc. Am.* **19**, 857–865.

Bakhvalov, N. S., Zhileikin, Ya. M., Zabolotskaya, E. A. (1987), *Nonlinear Theory of Sound Beams*, translated from the Russian, New York: AIP.

Barmatz, M. (1982), in: G. E. Rindone (Ed.), *Materials Processing in the Reduced Gravity Environment of Space*, New York: North-Holland, p. 25.

Barnard, G. A., Willette, J. G., Truchard, J. J., Shooter, J. A. (1972), *J. Acoust. Soc. Am.* **52**, 1437–1441.

Bellin, J. L. S., Beyer, R. T. (1962), *J. Acoust. Soc. Am.* **34**, 1051–1054.

Berktay, H. O. (1965), *J. Sound Vib.* **2**, 435.

Berktay, H. O., Al-Temimi, C. A. (1969), *J. Sound Vib.* **9**, 295.

Beyer, R. T. (1974), *Nonlinear Acoustics*, Washington, DC: Naval Sea Systems Command.

Blackstock, D. T. (1966) *J. Acoust. Soc. Am.* **39**, 1019–1026.

Burgers, J. (1948), *Advances in Applied Mechanics*, New York: Academic.

Coleman, A. J., Saunders, J. E., Palfrey, E. L. H. (1987), *J. Med. Eng. Technol.* **11**, 4.

Crum, L. A., Church, C. C., Blackstock, D. T. (1989), *J. Acoust. Soc. Am.* **85**, 2225.

de Klerk, J. (1972), *Phys. Today* 25(11), 32–39.

Dumond, J. W. M., Cohen, E. R., Panofsky, W. K. H., Deeds, E. (1946), *J. Acoust. Soc. Am.* **18**, 97–118.

Earnshaw, S. (1858), *Philos. Trans. R. Soc. London* **150**, 133–148. Reprinted (1972), in: R. B. Lindsay (Ed.), *Acoustics*, Stroudsburg: Dowden, Hutchinson and Ross, pp. 291–308.

Eckart, C. (1948), *Phys. Rev.* **73**, 68.

Eisenmenger, W. (1964), *Acustica* **14**, 187.

Errabolu, R. L., Sehgal, C. M., Bahn, R. C., Greenleaf, J. F. (1988), *Ultrasound Med. Biol. U. K.* **14**, 137.

Faraday, M. (1831), *Philos. Trans. R. Soc. London* **175**, 1.

Flynn, H. G. (1965), in: W. P. Mason (Ed.), *Physical Acoustics*, Vol. 1B. New York: Academic, pp. 265–336.

Fubini-Ghiron, E. (1935), *Alta Freq.* **4**, 532–581. Partial English translation in: R. T. Beyer (Ed.) (1984), *Nonlinear Acoustics in Fluids*, New York: Van Nostrand Reinhold, pp. 118–177.

Galloway, W. J. (1954), *J. Acoust. Soc. Am.* **26**, 849–857.

Granato, A. V., Lücke, K. (1956), *J. Appl. Phys.* **27**, 583–593.

Grassl, H. P., Reindl, L., Wörz, H.-W. (1988), in: D. F. Parker, G. A. Maugin (Eds.), *Recent Developments in Surface Acoustic Waves*, New York: Springer, p. 2.

Hartmann, B. (1979), *J. Acoust. Soc. Am.* **65**, 1392–1396.

Herring, C. (1941), OSRD Report No. 236.

Hikata, A., Elbaum, C. (1966), *Phys. Rev.* **144**, 469–477.

Ingard, K. U., Pridmore-Brown, D. C. (1956), *J. Acoust. Soc. Am.* **28**, 367–369.

Jones, G. L., Kobett, D. R. (1963), *J. Acoust. Soc. Am.* **35**, 5–10.

Kino, G. S., Matthews, H. (1971), *Spectrum* **8**(8), 22.

Kirkwood, J. G., Bethe, H. A. (1942), OSRD Report No. 236.

Kuttruff, J. H. (1962), *Acustica* **12**, 230.

Kuznetsov, V. P. (1970), *Akust. Zh.* **16**, 548 [*Sov. Phys. Acoust.* **16**, 467].

Liebermann, L. (1949), *Phys. Rev.* **75**, 1415–1422.

Lighthill, M. J. (1952), *Proc. R. Soc. London Ser. A* **211**, 64.

Lighthill, M. J. (1954), *Proc. R. Soc. London Ser. A* **212**, 1.

Lighthill, M. J. (1978), *Waves in Fluids*, Cambridge, England: Cambridge University Press.

Luukala, M., Kino, G. S. (1971), *Appl. Phys. Lett.* **18**, 393–394.

Medwin, H., Rudnick, I. (1953), *J. Acoust. Soc. Am.* **25**, 538–540.

Muir, T. G., Wilette, J. G. (1972), *J. Acoust. Soc. Am.* **52**, 1481–1486.

Noltingk, B. E., Neppiras, E. A. (1950), *Proc. Phys. Soc. London* **63B**, 675.

Noltingk, B. E., Neppiras, E. A. (1951), *Proc. Phys. Soc. London* **64B**, 1032.

Novikov, B. K., Rudenko, O. V., Timoshenko, V. I. (1987), *Nonlinear Underwater Sound*, translated from the Russian, New York: Acoustical Society of America.

Piercy, J. E., Lamb, J. (1954), *Proc. R. Soc. London Ser. A* **226**, 43.

Poisson, S. D. (1808), *J. Ec. Polytech.* **7**, 319. Partial English translation in: R. T. Beyer (Ed.) (1984), *Nonlinear Acoustics in Fluids*, New York: Van Nostrand Reinhold, pp. 23–28.

Quate, C. G., Thompson, R. B. (1970), *Appl. Phys. Lett.* **16**, 494–496.

Lord Rayleigh (1883), *Phil. Trans. R. Soc. London* **175**, 1.

Lord Rayleigh (1945), *Theory of Sound*, Reprint, New York: Dover.

Rudenko, O. V., Soluyan, S. I. (1977), *Theoretical Foundations of Nonlinear Acoustics*, translated from the Russian, New York: Plenum.

Ryan, R. P. (1963), Ph. D. thesis, Brown University, Providence, RI.

Sauerbruch, T., *et al.* (1986), *New England J. Med.* 314, 818.

Seghal, C. M., Bahn, R. C., Greenleaf, J. F. (1984), *J. Acoust. Soc. Am.* **76**, 1023–1029.

Seghal, C. M., Brown, G. M., Bahn, R. C. Greenleaf, J. F. (1986), *Ultrasound Med. Biol.* **12**, 865.

Sirotyuk, M. G. (1971), in: L. D. Rozenberg (Ed.), *High Intensity Fields*, New York: Plenum, p. 278.

Wang, T. G., Saffren, M. M., Elleman, D. D. (1974), AIAA Paper No. 74-155.

Wang, T. G. (1986), in: D. Sette (Ed.), *Proceedings of the International School of Physics "Enrico Fermi,"* Amsterdam: North-Holland.

Westervelt, P. J. (1957), *J. Acoust. Soc. Am.* **29**, 199–203, 934–935.

Westervelt, P. J. (1960), *J. Acoust. Soc. Am.* **32**, 934(A).

Westervelt, P. J. (1963), *J. Acoust. Soc. Am.* **35**, 535–537.

Zabolotskaya, E. A., Khokhlov, R. V. (1969), *Akust. Zh.* **15**, 40; (1969), *Sov. Phys. Acoust.* **15**, 35.

Zhang, J., Dunn, F. (1987), *J. Acoust. Soc. Am.* **81**, 1635.

Zverev, V. A., Kalachev, A. I. (1970), *Sov. Phys. Acoust.* **15**, 322; (1970), **16**, 204.

Further Reading

Bakhvalov, N. S., Zhileikin, Ya. M., Zabolotskaya, E. A. (1987), *Nonlinear Theory of Sound Beams*, translated from the Russian, New York: AIP.

Beyer, R. T. (1974), *Nonlinear Acoustics*, Washington, DC: Naval Sea Systems Command.

Beyer, R. T. (Ed.) (1984), *Nonlinear Acoustics*, Benchmark Papers in Acoustics Series, New York: Van Nostrand Reinhold. Reprints and translations of 33 significant papers in the field.

Lighthill, M. J. (1978), *Waves in Fluids*, Cambridge, England: Cambridge University Press.

Novikov, B. K., Rudenko, O. V., Timoshenko, V. I. (1987), *Nonlinear Underwater Sound*, translated from the Russian, New York: Acoustical Society of America.

Parker, D. F., Maugin, G. A. (Eds.) (1988), *Developments in Surface Acoustic Waves*, New York: Springer.

Rozenberg, L. D. (1971), *High Intensity Ultrasonic Fields*, translated from the Russian, New York: Plenum.

Rudenko, O. V., Soluyan, S. I. (1977), *Theoretical Foundations of Nonlinear Acoustics*, translated from the Russian, New York: Plenum.

ACOUSTICS, PHYSIOLOGICAL

MARIO A. RUGGERO, *Department of Otolaryngology, University of Minnesota, Minneapolis*

MALCOLM N. SEMPLE, *Department of Anatomy and Neurobiology, University of California, Irvine*

Introduction 214
1. **Overview of the Peripheral Auditory System** 214
2. **The Output of the Cochlea: Auditory Nerve Activity** 216
2.1 Spontaneous Activity 216
2.2 Responses to Single Tones 217
2.2.1 Frequency-Domain Description 217
2.2.2 Time-Domain Description 218
2.3 Responses to Broadband Stimuli 219
2.4 Nonlinear Interactions in Responses of Cochlear Afferents to Tone Pairs 221
2.4.1 Two-Tone Suppression 221
2.4.2 Combination-Tone Generation. 222
2.5 Effects of Hair Cell Loss upon Cochlear Afferent Responses to Sound 223
2.6 Effects of Stimulation of the Olivocochlear Efferent System upon Cochlear Afferent Responses to Sound 224
3. **The Input to the Cochlea: The External and Middle Ears** 225
3.1 External Ear 225
3.2 Middle Ear Anatomy 225
3.3 Impedance Matching by the Middle Ear 225
3.4 Transfer Function of the Middle Ear 227
3.5 Middle Ear Muscles 228
4. **The Cochlea** 228
4.1 Anatomy of the Cochlea 228
4.1.1 Cochlear Compartments 228
4.1.2 Basilar Membrane and Supporting Cells 230
4.1.3 Inner and Outer Hair Cells 230
4.1.4 Afferent and Efferent Innervation 230
4.2 Passive Aspects of Basilar Membrane Motion 231
4.2.1 The Traveling Wave 231
4.2.2 Frequency Tuning and Tonotopicity 232
4.2.3 Physical Basis of Passive Cochlear Mechanics 232
4.3 Active Aspects of Basilar Membrane Motion 233
4.3.1 Basilar-Membrane Nonlinearities and their Lability 233
4.3.2 Mechanical Origin of Cochlear Afferent Response Properties 234
4.4 Hair Cell Receptor Potentials 235
4.5 The Cochlea as a Sound Generator 237
4.5.1 Otoacoustic Emissions 237
4.5.2 Outer Hair Cell Motility 238
5. **Overview of the Central Auditory System** 239
6. **The Cochlear Nuclear Complex** 239
7. **The Superior Olivary Complex** 241
8. **The Auditory Midbrain** 242
8.1 Nuclei of the Lateral Lemniscus 242
8.2 The Inferior Colliculus 242
9. **The Auditory Thalamus and Cortex** 243
9.1 The Medial Geniculate Body 243
9.2 The Auditory Cortex 243
9.3 Some Behavioral Consequences of Cortical Lesions 244
10. **Response Properties in the Central Auditory System** 244
10.1 Discharge Patterns 244
10.2 Phase Locking 244
10.3 Frequency Tuning 245
10.4 Response Latency 246
10.5 Level Sensitivity 246

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

10.6 Binaural Sensitivity and Mechanisms for Sound Localization.................. 248
10.6.1 Binaural Cues for Sound Localization.................. 248
10.6.2 Central Processing of Binaural Information.................. 248
10.7 Complex Sounds 250
10.7.1 Background Noise............ 250
10.7.2 Amplitude and Frequency Modulation 251
10.7.3 Vocalizations................. 252
11. Centrifugal Influences........ 253
11.1 Olivocochlear Efferent System. 253
11.2 Other Descending Pathways... 253
Acknowledgments 253
Glossary..................... 254
Works Cited 257
Further Reading............. 259

INTRODUCTION

The sense of hearing of mammals is a marvelous achievement of biological evolution. Even when considered as merely a physical sensor, the ear's capabilities are tremendous: humans can detect sounds at levels at which air particle displacements are as small as 1 pm, approximately equivalent to 1/100 of the radius of a hydrogen atom. The highest attainment of the sense of hearing arguably consists of carrying out psychoacoustical tasks, such as pitch perception and intensity discrimination, that ultimately permit speech communication among humans. A description of these psychoacoustical tasks is provided in the companion article, ACOUSTICS, PSYCHOLOGICAL.

The present chapter sketches the physiology of the auditory system, the physical basis of the hearing sensation. The auditory system consists fundamentally of three components: (1) the auditory sensory organ, the cochlea or inner ear, which, aided by the external ear and the middle ear, transduces and encodes the acoustic signal into electrical pulse trains in the peripheral terminals of the cochlear nerve; (2) the auditory or cochlear nerve, which conveys the electrical pulse trains to the brain; and (3) the auditory regions of the brain. Whether from a numerical standpoint, or from an assessment of organizational complexity, it is clear that it is in the auditory centers of the brain where "hearing" emerges out of the extremely intricate interactions of many millions of neurons. Because of their complexity, our understanding of central auditory mechanisms lags our knowledge of peripheral auditory physiology. Therefore, the present account of physiological acoustics places greater emphasis on peripheral processes than will likely be the case in future editions of this encyclopedia.

We begin our account of physiological acoustics with an overview of the peripheral auditory system (Sec. 1), before proceeding to more detailed accounts of the auditory nerve (Sec. 2), the external and middle ears (Sec. 3), and the cochlea (Sec. 4). Rather than systematically following the acoustic signal and its transformations from the environment toward the brain, we shall start with a description of the activity of the auditory nerve. This is justified on the grounds that the responses to sound of cochlear neurons are quite well known in comparison with the mechanical, electrical, and chemical processes of the cochlea that give rise to these neural responses. A coherent description of the neural code used by the auditory nerve to transmit acoustic information (Sec. 2) provides a convenient foundation for the subsequent exploration of how the encoding results from acoustical and mechanical transformations in the external and middle ears (Sec. 3), and from mechanoelectrical transduction in the cochlea (Sec. 4). In the final sections of the chapter we discuss what is known about how the brain channels and transforms the acoustic information received from the auditory nerve.

1. OVERVIEW OF THE PERIPHERAL AUDITORY SYSTEM

Figure 1 presents a block diagram of the peripheral auditory system. The arrows indicate the direction of signal transmission. The traditional view of the auditory periphery, current until less than a decade ago, was that the path of signal transmission and transduction proceeded steadily from the periphery

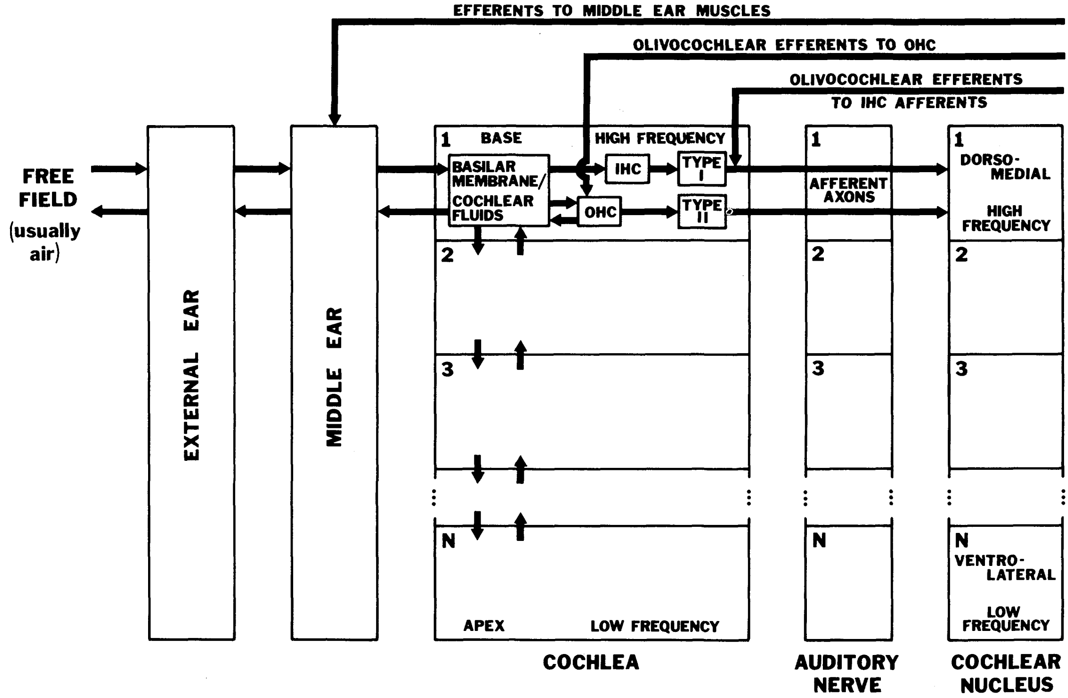
EFFERENTS TO MIDDLE EAR MUSCLES
OLIVOCOCHLEAR EFFERENTS TO OHC
OLIVOCOCHLEAR EFFERENTS TO IHC AFFERENTS
FREE FIELD (usually air)
EXTERNAL EAR
MIDDLE EAR
1 BASE
HIGH FREQUENCY
BASILAR MEMBRANE/ COCHLEAR FLUIDS
IHC
OHC
TYPE I
TYPE II
2
3
N
APEX
LOW FREQUENCY
COCHLEA
1
AFFERENT AXONS
2
3
N
AUDITORY NERVE
1
DORSO-MEDIAL
HIGH FREQUENCY
2
3
N
VENTRO-LATERAL
LOW FREQUENCY
COCHLEAR NUCLEUS

FIG. 1. A block diagram of the organization of the peripheral auditory system. Blocks are arranged from peripheral (i.e., closer to the environment, left) to central (i.e., closer to the brain, right). Arrows indicate the direction of signal transmission. Inner hair cells and outer hair cells are represented by the blocks labeled "IHC" and "OHC," respectively. The two types of cochlear afferent neurons are indicated by the blocks labeled "Type I" and "Type II." As the (initially, acoustic) signal proceeds inward from the periphery, it is successively transformed into mechanical vibrations of the middle ear, pressure fluctuations in the cochlear fluids, and displacements of the basilar membrane. Mechanical-to-electrical transduction takes place in the hair cells, whose deformation generates intracellular voltage analogs of the acoustic stimulus; these in turn cause chemical transmitter release at the hair cell basal surface, leading to the initiation of all-or-none impulses in the peripheral endings of the cochlear afferents. The neural impulses propagate along the axons of the auditory nerve toward the cochlear nucleus.

inward to the auditory central nervous system (with two exceptions, the olivocochlear efferents, which travel in the auditory nerve from the brain to the cochlea, and the efferents to the middle ear muscles). Thus, proceeding from left to right in Fig. 1: sound impinges upon the external ear and the eardrum, setting in motion tiny bones in the middle ear; these, in turn, create pressure fluctuations in the largely fluid-filled cochlea (inner ear), causing a slow displacement wave to propagate along the basilar membrane. The displacement of the basilar membrane stimulates cellular transducers, the inner and outer hair cells (IHCs and OHCs), to generate, much like tiny microphones, electrical receptor potentials that mimic the acoustic stimulus. Finally, these receptor potentials produce chemically mediated excitation in the peripheral terminals of cochlear afferent neurons, generating trains of electrical pulses (action potentials) which travel, via the auditory nerve, to the cochlear nucleus, the first station of the auditory central nervous system.

Note that at several stages of the auditory periphery (Fig. 1) signal transmission proceeds not only inward, in accordance with the traditional view, but also outward toward the environment. At the external ear the arrows signify that not only is sound conveyed from the environment into the middle ear, but also middle ear vibrations can broadcast sound into the environment. The discovery that the ear, in addition to serving as a passive receiver and transducer, can also act as an active generator of sound has contributed to a recent revolution in our understanding of cochlear physiology.

While the external ear and the middle ear are each represented in Fig. 1 as single, undivided boxes, the cochlea, auditory nerve, and cochlear nucleus are represented as being subdivided into segments numbered from 1 through N. These subdivisions represent the property of tonotopicity, i.e., the systematic mapping of frequency into space. Tonotopicity, ubiquitous throughout the auditory central nervous system, has its origin in the inner ear, where the basilar membrane, cochlear fluids, and sensory cells jointly perform a mechanical frequency analysis whereby waves of basilar membrane displacement resulting from low-frequency stimulation travel the length of the cochlea and peak toward its apex, while waves evoked by higher-frequency stimuli reach a peak toward the cochlear base and do not propagate farther. This mechanical frequency analysis causes the longitudinal array of hair cells to be stimulated by a spatial representation of the audible frequency spectrum. Thus, adjacent segments of the basilar membrane and, as a result, the hair cells on each segment and their associated cochlear afferents are tuned to slightly different frequencies. The inner ear may therefore be viewed as analogous to a bank of bandpass filters connected to long-distance signal transmission lines. These, the thousands of cochlear neurons and their axons, encode in their trains of action potentials all the information that, after processing by the many millions of neurons in the auditory central nervous system, make up the auditory experience.

Whenever possible, the ensuing account of peripheral auditory physiology will use the chinchilla as a representative mammalian species. Focusing on a single species makes it simpler to follow the acoustic-signal transformations from one stage of the peripheral auditory apparatus to another. The chinchilla has been chosen here as a counterpoint to reviews that have emphasized data for cats and guinea pigs, the other mammalian species whose ears and auditory nerves have been intensively studied.

2. THE OUTPUT OF THE COCHLEA: AUDITORY NERVE ACTIVITY

The axons of cochlear afferent neurons, whose cell bodies reside in the spiral ganglion of the cochlea, make up the bulk of the auditory (or cochlear) nerve [Figs. 1, 10, 13(a), and 13(b)]. There are two types of cochlear neurons, Type I and Type II, both of which carry signals from hair cells to the cochlear nucleus, where they terminate (Fig. 1; Spoendlin, 1988). Type I neurons, each of which innervates a single inner hair cell, constitute the vast majority of cochlear afferents (90% out of a total of about 30 000 neurons in chinchillas; Bohne *et al.*, 1982, and Ruggero *et al.*, 1982); each Type II neuron, on the other hand, innervates many outer hair cells. While nothing is known about what information Type II neurons transmit or what role they play in hearing, a great deal is known about the responses to sound of Type I neurons (reviewed by Kiang, 1984, and Pickles, 1988).

2.1 Spontaneous Activity

Like most neurons, Type I cochlear afferents transmit information by modulating the instantaneous frequency of occurrence of all-or-none electrical pulses (action potentials or "spikes") that travel nondecrementally along their cablelike axons at speeds of tens of m/s. These action potentials can be recorded by placing a microelectrode (usually a glass pipette with a tip smaller than 1 μm, filled with a conductive fluid) in the trunk of the auditory nerve near its exit from the cochlea, between the modiolus and the cochlear nucleus [Figs. 1, 10, and 13(a)]. Examples of oscillographic records of action potential trains may be seen in Fig. 6. In the absence of sound, cochlear afferents are nevertheless active, generating action potentials at average rates ranging from less than 1/s to 140/s. The distribution of rates of spontaneous activity across the population of cochlear Type I afferents is bimodal, with a substantial proportion of neurons being spontaneously active at rates less than 10/s and the remainder having higher, widely distributed rates, with a second mode around 60/s (Liberman, 1978; Dallos and Harris, 1978). The time of occurrence of each action potential is highly random: the distribution of interspike intervals closely resembles a Poisson distribution, except for the absence of intervals shorter than 0.7 ms (due to the period of inexcitability, the "refractory period," following each action potential).

2.2 Responses to Single Tones

2.2.1 Frequency-Domain Description. When effectively stimulated by sound, cochlear afferents increase their rate of discharge. At low intensities, such rate increases are evoked only by a narrow range of stimulus frequencies. The locus of intensity–frequency combinations for single tones that cause a just-measurable increase in firing above spontaneous rate is known as the frequency-threshold tuning curve, often abbreviated to "tuning curve" (Fig. 2). The tonal frequency at which responses are elicited at the lowest intensity is called the "characteristic frequency" (CF). The CF is a fundamental property of each cochlear afferent which provides accurate information about the cochlear location of the hair cell innervated by that afferent: there is a one-to-one CF-to-space mapping in the cochlea, such that high-CF cochlear afferents (i.e., those preferentially tuned to high frequencies) innervate inner hair cells at the basal end of the cochlea, while low-CF cochlear afferents innervate inner hair cells at the cochlear apex (see dual scale in abscissa of Fig. 8).

Figure 2 shows that the shapes of tuning curves change systematically as a function of CF. The tuning curves of high-CF cochlear afferents have very steep (e.g., 300 dB/octave) high-frequency slopes. The low-frequency sides of these tuning curves have two distinct segments. One segment has a relatively steep slope and extends from the tuning-curve tip to 40–70 dB above the CF threshold; the second, lowest frequency, segment has a relatively shallow slope. Tuning curves of cochlear afferents with lower CFs are broader and more symmetrical. A common measure of frequency tuning, adapted by auditory physiologists from the "quality factor" (Q_3) used by electrical engineers, is the Q_{10}, defined as the CF divided by the bandwidth at 10 dB above CF threshold. In the chinchilla auditory nerve, Q_{10}'s vary from 1 at CFs around 150 Hz to 7 at CFs close to 16 kHz (Dallos and Harris, 1978). Because of the strict CF-to-cochlear location correspondence, the brain could extract frequency information on the basis of the site of origin of excited neurons. This place code would be most precise at low stimulus levels and would become less specific for intense stimuli, which excite a widely distributed population of cochlear afferents.

There is a clear relationship between the rate of spontaneous activity of cochlear afferents and their thresholds at CF. High-spontaneous-rate afferents have the lowest thresholds, which, for a given CF, do not vary by more that 10 dB. The CF thresholds of low-spontaneous-rate afferents are substantially higher and more variable. In chinchillas, the lowest CF thresholds (10–20 dB SPL) are those of high-spontaneous-rate afferents with CF between 500 Hz and 8 kHz. When CF thresholds are plotted as a function of frequency, they describe a function which is roughly parallel to the behavioral audiogram for a particular species (Fig. 2). Afferents innervating inner hair cells at the extremes of

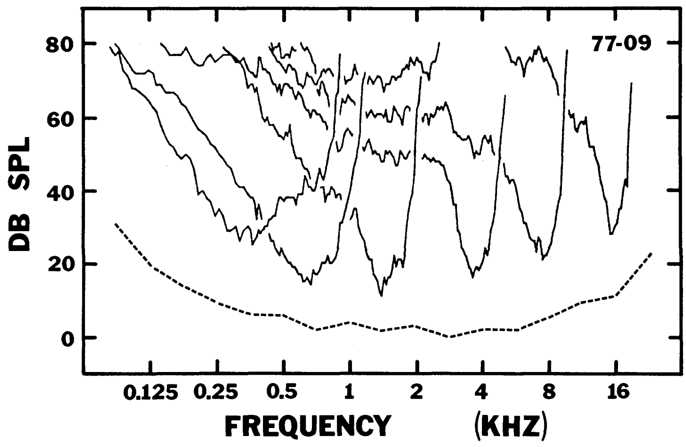

FIG. 2. Representative frequency-threshold tuning curves for cochlear afferents recorded from one chinchilla auditory nerve. Each tuning curve indicates the locus of frequency/intensity stimulus combinations that elicited a discharge rate just exceeding spontaneous activity. (Unpublished neural data of M. A. Ruggero.) Also shown (dashed line) is the behavioral audiogram for normal-hearing chinchillas, according to Miller (1970).

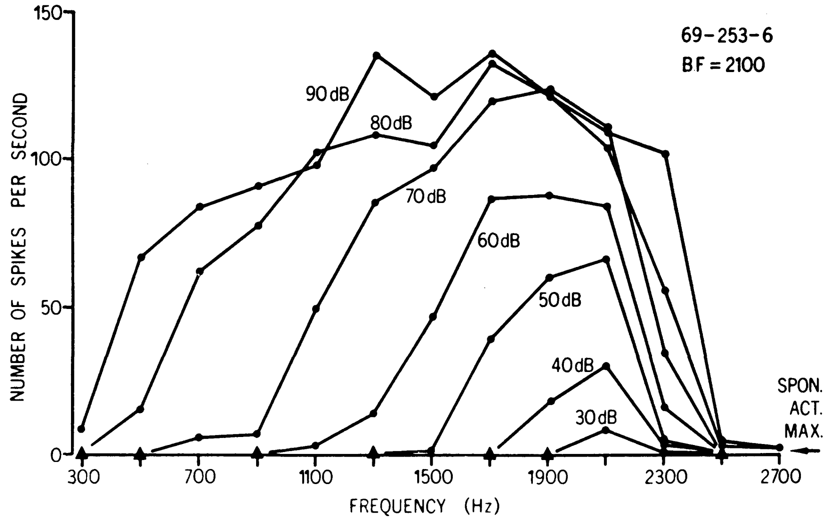

FIG. 3. Responses to tones, as a function of stimulus frequency and intensity, of a cochlear neuron in the squirrel monkey. The abscissa indicates the stimulus fundamental frequency. The ordinate indicates mean discharge rate. Responses to equal-intensity tones are connected, with intensity indicated as the parameter. Reproduced by permission from Rose *et al.* (1971), Fig. 2(b).

the chinchilla cochlea have CFs as low as 150 Hz and as high as 22 kHz; the absence of afferents with CF beyond these extremes probably determines the frequency limits of hearing for the chinchilla.

Because cochlear-nerve responses are nonlinear, to explore their sound-encoding capabilities adequately, it is necessary to use many different types of stimuli at various intensities. For stimulation with single tones, at levels above threshold, the average response rate increases roughly in proportion to the logarithm of sound pressure over a limited intensity range (the dynamic range), eventually reaching a rate plateau. A general description of the complex, intensity-dependent properties of a cochlear neuron's responses to single tones is obtained by plotting response rate as a function of frequency and intensity (Fig. 3). At low stimulus levels, the neuron responds to a narrow range of frequencies; at higher levels, responses extend to higher and lower frequencies. In accordance with the previously noted asymmetry in the slopes of threshold tuning curves of high-CF afferents (Fig. 2), increases in stimulus level incorporate wider response ranges at frequencies below CF than at frequencies above CF. In addition, at higher stimulus levels the highest response rates tend to shift toward lower frequencies. For tones at CF (2100 Hz), increments in response rate occur over a 40-dB range at a rate that is roughly proportional to the logarithm of pressure (i.e., proportional to decibels). At higher-than-CF frequencies the dynamic range is wider, and the maximum response rate lower, than at CF.

2.2.2 Time-Domain Description. Because of the stochastic nature of action potential generation noted above (Sec. 2.1), the spike trains evoked by identical stimuli are highly variable, and responses to sound must be described using statistical averages, such as the mean rate used for the construction of frequency-threshold tuning curves. In the time domain, the most commonly used statistical representation of spike trains is the so-called post-stimulus-time histogram (PSTH), constructed by averaging responses to repeated presentations of identical stimuli, binning spikes according to their latency relative to stimulus onset. Examples of PSTHs are shown in Fig. 4 for responses of a chinchilla cochlear neuron to a high-frequency (9450 Hz) tone burst and to a low-frequency (500 Hz) tone burst, both presented at a moderately intense level (70 dB SPL). The ordinate indicates the average rate of firing in responses to 200 repetitions of 25-ms tone bursts. Spontaneous activity is evident before each response onset and after the end of the

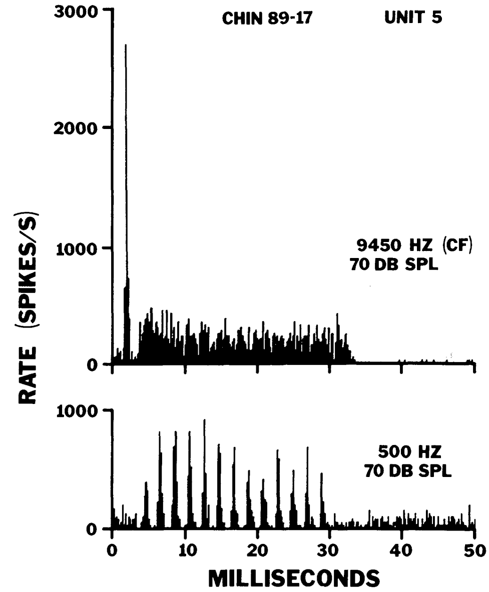

FIG. 4. Poststimulus time histograms for the responses of a single chinchilla cochlear neuron to 9450-Hz (top) and 500-Hz (bottom) tone pips. Stimuli had rise and fall times of 4 ms, lasted 25 ms, and were presented every 150 ms. Histograms are averages of responses to 200 stimulus repetitions. The neuron's CF was 9450 Hz and it had a CF threshold of 3 dB SPL and spontaneous activity of 59 spikes/s. The data for this figure were collected by B. Shivapuja.

500-Hz tone. In the PSTH for responses to the tone burst at 9450 Hz (Fig. 4, top), corresponding to the neurons's CF, activity attains an initially very high discharge peak, which is followed by a brief rate notch; after a few milliseconds, the response settles to a steady rate that is maintained until the end of the stimulus. The rate notch reflects the aforementioned refractory period following the action potential that occurs at the onset, in this case, of the majority of stimulus repetitions. The disparity between the initial discharge rate and the sustained rate is due to synaptic adaptation, which probably results from the inability of the inner hair cell to release chemical transmitter continuously at its synapses with the terminals of cochlear afferents. After the end of the stimulus, the spontaneous rate is temporarily depressed, eventually returning to prestimulus levels. Recovery of spontaneous activity is not completed for the 9450-Hz tone pip within the 50-ms window of the histogram, but is evident for the 500-Hz histogram.

The most prominent feature of the PSTH in response to the 500-Hz tone pip (Fig. 4, bottom) is the presence of multiple peaks, each corresponding to one period of the stimulus tone. Although cochlear afferents rarely fire at high enough rates to be able to discharge during every stimulus cycle, they always respond preferentially at a particular phase of any stimulus with fundamental frequency less than about 4 kHz. In responses to such low-frequency stimuli, intervals between action potentials may span one, two, or multiple stimulus cycles, but when spikes do occur, they tend to happen during a restricted portion of the stimulus period (i.e., at a specific phase). Thus, the temporal structure of the spike train reflects the periodicity of the stimulus wave form. When the phases of cochlear neurons' responses to low-frequency tones are plotted against their CFs [Fig. 8(b)], it is seen that phase lag increases as CF decreases. Because of the tonotopic organization of the cochlea (see Sec. 1), such an increasing lag implies that neural excitation occurs progressively later as a function of distance from the cochlear base. We shall show in Sec. 4.2 that this progressive delay in neural excitation arises from the finite time required for sound-evoked disturbances to travel from base to apex along the basilar membrane.

The ability of cochlear afferents to synchronize their responses to specific phases of low-frequency stimuli, known as "phase locking," provides a potential mechanism for encoding low-frequency (<4 kHz) information. In such a hypothetical temporal code, stimulus frequency would be conveyed to the cochlear nucleus in the temporal pattern of activity of all excited neurons. It is noteworthy that such a code would be effective irrespective of either the site of innervation of the cochlear afferent or the stimulus level, since any cochlear afferent that responds to low-frequency (<4 kHz) stimuli does so in a phase-locked manner.

2.3 Responses to Broadband Stimuli

Linear systems that are frequency tuned respond to an impulse or click with an oscillation whose period is the inverse of the center frequency and whose envelope decays at a rate that varies inversely with the sharpness of tuning. Although the response of cochlear neurons is highly nonlinear, their responses

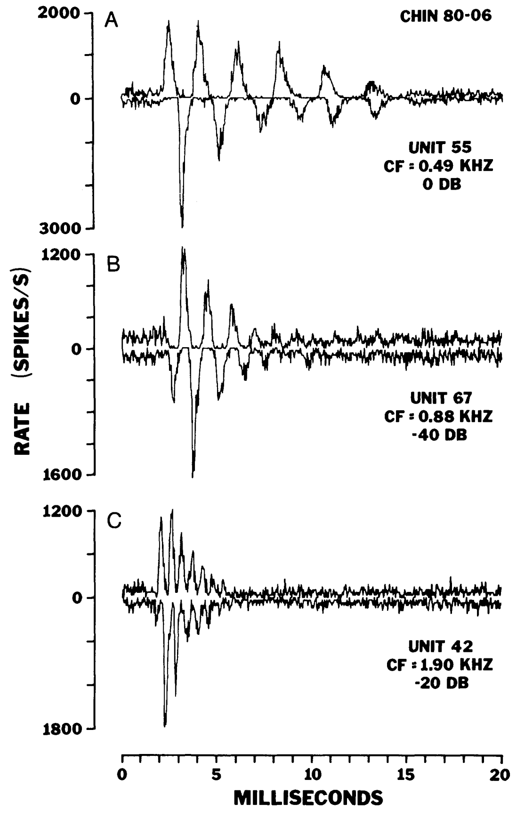

FIG. 5. Responses to clicks of three cochlear afferents from one chinchilla auditory nerve. Responses to stimuli of both polarities are shown jointly as compound poststimulus time histograms (PSTHs). In each of the three panels, the PSTH for responses to rarefaction clicks is plotted in the usual manner, and the PSTH for condensation clicks is plotted upside down. Characteristic frequency of the cochlear afferent and click intensity (expressed as attenuation) are indicated in each panel (unpublished data of M. A. Ruggero).

to clicks are to some extent predictable from the frequency tuning of their responses to tones (Fig. 2). All cochlear neurons can be excited by clicks (Kiang *et al.*, 1965). Figure 5 shows compound PSTHs of the responses to rarefaction and condensation clicks of three neurons (CFs of 490, 880, and 1900 Hz) in one chinchilla cochlear nerve. Each compound histogram is made up of two histograms: responses to rarefaction clicks are plotted as usual, while responses to condensation clicks are inverted. Upon presentation of the rarefaction click (at time zero), there is a brief delay during which the neuron continues to fire at its spontaneous rate. The response to the click consists of multiple peaks, spaced at intervals corresponding to the inverse of the neuron's CF. Upon reversing the polarity of the stimulus click, the times of occurrence of the responses remain periodic but shift to times at which there were few responses to rarefaction clicks: in other words, responses to clicks of opposite polarity interleave with each other. This is consistent with the phase locking noted above in responses to tones.

The shape of PSTHs for responses to clicks changes with the intensity of the stimuli. The PSTHs depicted in Fig. 5(a) are typical for intense clicks in that the earliest responses are elicited by rarefaction clicks and that the sizes of the peaks decay monotonically as a function of time. For lower-intensity clicks, the envelopes of the histograms are more spindle-shaped, with the largest peak being flanked by smaller ones [e.g., PSTHs for condensation in Fig. 5(b) and rarefaction in Fig. 5(c)]. The change in PSTH shape as a function of click intensity is, to a large extent, due to the adaptive properties of the synapse between the inner hair cell and the cochlear afferent terminal (Sec. 2.2), but it also probably reflects cochlear mechanical nonlinearities (Sec. 4.3).

Multiple peaks are characteristic of the PSTHs for responses to clicks of low-CF (<3 kHz) cochlear afferents; in higher-CF neurons, their inability to phase lock to CF tones is reflected in the absence of multiple peaks in their click-response PSTHs. In addition to the spacing between consecutive PSTH peaks, the latency of responses to intense clicks also changes systematically as a function of CF, diminishing from about 3 ms for very-low-CF neurons (e.g., 150 Hz) in the chinchilla to an asymptotic value of about 1 ms for neurons with CFs above 3 kHz. The 1-ms asymptotic value can be accounted for by two principal delays: the delay of chemical synaptic transmission, mostly due to the diffusion time of the chemical transmitter from the inner hair cell to the peripheral terminal of the cochlear neuron, and the axonal conduction time of action potentials from the terminal to the site of the recording microelectrode. Irreducible CF-dependent response delays longer than 1 ms correspond to the CF-dependent phase lags previously noted in responses to tones (Sec. 2.2.2); such delays will be shown in a later section (Sec. 4.2) to reflect corresponding delays in cochlear mechanics.

Responses of cochlear neurons to noise are understandable according to the same principles used to interpret responses to clicks. All neurons can be excited by acoustic white noise. The responses of low-CF neurons contain time fluctuations with a frequency spectrum similar to the frequency-tuning curves for their responses to tones. High-CF neurons, on the other hand, respond to white noise with an increased average discharge rate but with no periodicity evident in the train of discharges.

2.4 Nonlinear Interactions in Responses of Cochlear Afferents to Tone Pairs

2.4.1 Two-Tone Suppression. As seen in previous sections, cochlear afferent responses are not linear and, therefore, responses to stimuli with multiple frequency components (such as clicks) cannot be well predicted on the basis of responses to single tones. In order to gain some understanding of the underlying nonlinear processes that preclude linear synthesis, it is convenient to study the responses to relatively simple stimuli, such as tone pairs, in which nonlinear interactions are evident but whose complexities are still amenable to analysis. One of the better studied nonlinearities in cochlear nerve responses is "two-tone rate suppression" (Galambos and Davis, 1944; Sachs and Kiang, 1968). This is defined as the reduction in the response to one tone due to the presence of a second tone. Figure 6 shows the responses of a cat cochlear afferent to two tones presented individually [Figs. 6(a) and 6(b)] and together [Fig. 6(c)]. The neuron, which had little spontaneous activity, was moderately excited by both a continuous 360-Hz (CF) tone [Fig. 6(a)] and a 720-Hz tone pip [Fig. 6(b)]. When the two tones were presented simultaneously, however, the neuron almost ceased firing throughout the duration of the 720-Hz, suppressive, tone pip [Fig. 6(c)].

Whether or not one tone causes suppression of another depends strongly on the levels and frequencies of the two tones. Figure 7 shows, superimposed on a frequency-threshold tuning curve, shaded regions indicating tonal stimuli that suppressed responses to a near-CF tone (triangle). The regions of overlap of the excitatory tuning curve and the shaded suppressive areas in Fig. 7 indicate that tones that are excitatory by themselves can suppress responses to a CF tone. One of the most significant properties of two-tone suppression is its dependence on the frequency of the excitatory tone: only tones with frequencies close to the neuron's CF can be strongly suppressed. This frequency selectivity suggests that suppression arises in cochlear mechanics since, as we shall discuss later (Sec. 4.3), sharp frequency tuning is established early in the chain of transduction events, being evident in the vibrations of the basilar membrane. This suggestion is consistent with the latency of suppression onset, which is as short as the latency of response to excitatory tones (Arthur *et al.*, 1971), and with the demonstration that two-tone suppression persists after sectioning of the auditory nerve central to the site of recording from cochlear

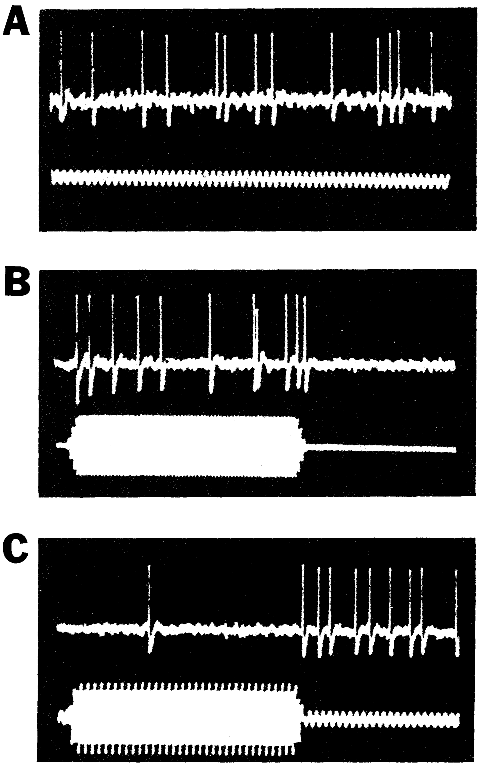

FIG. 6. Two-tone suppression in a cat cochlear afferent. The upper oscilloscope tracing in each panel shows spike trains elicited by stimulation with two tones presented singly (a and b) and together (c). The lower tracing in each panel indicates the duration of a continuous 360-Hz, 40-dB SPL tone (a and c) and/or of a 720-Hz, 60-dB SPL tone pip (b and c). Note diminished rate of firing when both tones are presented together (c). Reproduced by permission from Arthur *et al.* (1971), Fig. 2(a).

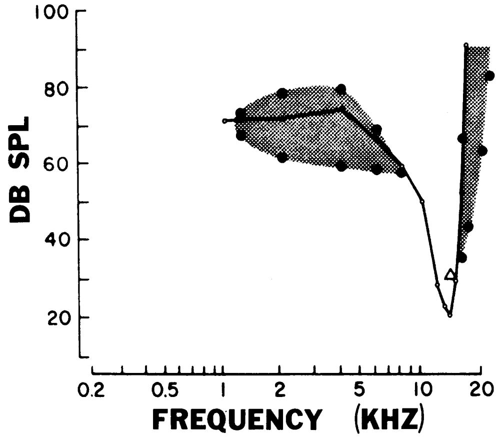

FIG. 7. Two-tone suppression areas and frequency-threshold tuning curve for one chinchilla cochlear afferent. The continuous line indicates the tuning curve in response to tone pips with frequency and intensity indicated in the abscissa and ordinate, respectively. The triangle indicates a probe tone, responses to which are suppressed when second tones are presented with frequency/intensity combinations demarcated by the shaded areas. Reproduced by permission from Harris (1979), Fig. 13.

afferents (Kiang *et al.*, 1965). These observations rule out involvement of the olivocochlear efferent system (see Fig. 1 and Secs. 2.6 and 11.1) in two-tone suppression, since sound-evoked activation of olivocochlear neurons has a long latency and sectioning of the auditory nerve prevents actions potentials traveling along olivocochlear axons from reaching the cochlea.

2.4.2 Combination-Tone Generation. Another well-studied example of nonlinear interactions between two tones is the generation of so-called "combination tones," also known as "intermodulation distortion products" or "Tartini pitches." Giuseppe Tartini, an 18th century Italian musician, is reputed to have first recognized that, upon presentation of two tones with frequencies f_1 and f_2 ($f_2 > f_1$), humans often perceive tones not physically present in the acoustic stimulus. Physiological correlates of these distortion products, the most prominent of which have frequencies equal to $2f_1 - f_2$ and $f_2 - f_1$, exist in the responses of cochlear afferents to pairs of tones with appropriate frequencies and intensities. Figure 8 presents evidence suggesting that the combination tone $f_2 - f_1$ is a product of nonlinear cochlear mechanics and that it propagates on the basilar membrane in a manner similar to that of a tone of the same frequency present at the input to the cochlea. The responses of cochlear afferents to a two-tone stimulus (3680 and 4000 Hz) were recorded from one auditory nerve in a chinchilla and they were compared to responses to a separately presented single tone with frequency of 320 Hz (i.e., equal to the combination frequency, $f_2 - f_1$). The spike trains were binned into time histograms, with the abscissa equal to the fundamental repetition period of the stimulus wave form. Subsequent Fourier transformation of the histograms revealed the magnitudes and phases of the spectral response components which were phase locked (Sec. 2.2.2) to 320 Hz, the frequency of the single tone (crosses and dashed lines) and of the combination tone $f_2 - f_1$ (filled circles and solid lines). Since there is a strict frequency-to-place mapping along the cochlea (Fig. 1), plotting the values of the spectral components against CF is equivalent to assessing the distribution of the spectral response components as a function of cochlear location. This equivalence is indicated in Fig. 8 by the two abscissa scales. For the two-tone stimulus, responses were strongly phase locked to the combination-tone frequency at two separate cochlear locations: at the basal regions with CFs equal to the primary frequencies (3680 and 4000 Hz) and at the apical location with CF equal to the combination frequency (320 Hz). In the absence in the stimulus of energy at the combination-tone frequencies, and given that the middle ear responds linearly, it is clear that neural responses phase locked to the combination-tone frequency indicate a cochlear origin of the combination tone. In the apical region, the spatial distributions of the 320-Hz components of the responses to the two-tone stimulus (filled circles and solid lines) and to the 320-Hz single tone stimulus (crosses and dashed lines) were very similar in both magnitude [Fig. 8(a)] and phase [Fig. 8(b)]. The similarity between the phase distributions is especially significant, since it suggests that the combination-tone component propagates from base to apex along the basilar membrane in a manner very similar to that of the single tone (see Secs. 2.2 and 4.2). Further, the presence of a strong $f_2 - f_1$ response at locations with CFs equal to the primary-tone frequencies, f_1 and f_2, suggests that the me-

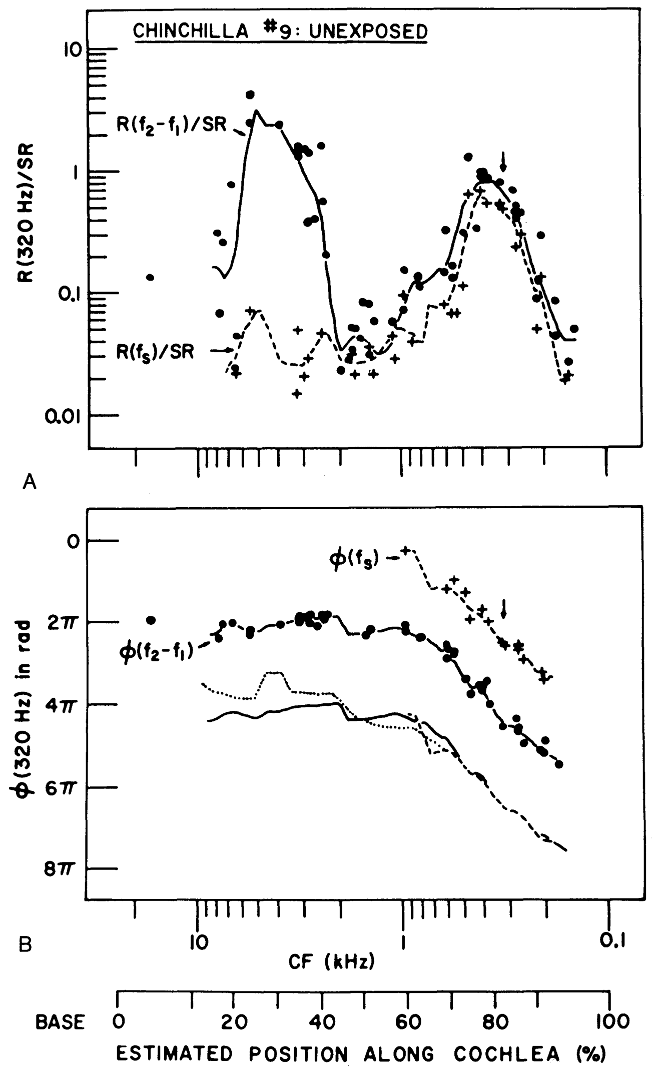

FIG. 8. Comparison of the spatial distribution of the 320-Hz Fourier component in responses of chinchilla cochlear afferents to 320-Hz single tones and to two-tone stimuli. The single-tone responses (crosses and dashed lines) were evoked by 320-Hz, 5-dB SPL stimuli; the two-tone data (filled circles and solid lines) were obtained with paired 3680-Hz and 4000-Hz tones, presented at 50-dB SPL. The magnitudes (a) and phases (b) of the 320-Hz response components are plotted as functions of characteristic frequency or, equivalently, cochlear location. The lines drawn through the data points are moving-window averages. In (b) these lines have been plotted twice, once after vertical displacements to facilitate comparison of their slopes. The dotted line indicates averages for responses to 50-dB, 320-Hz single tones. Reproduced by permission from Siegel *et al.* (1982), Fig. 4.

chanical distortion products originate at these primary sites and subsequently propagate, in an apical direction, toward locations with CFs matching the combination-tone frequencies. More convincing evidence for this hypothesis comes from studies of cochleae with hair cell loss.

2.5 Effects of Hair Cell Loss upon Cochlear Afferent Responses to Sound

Procedures such as exposure to very intense sound or chronic administration of certain antibiotics destroy some or many of the hair cells of the cochlea, while leaving most other cells intact. Recordings from the cochlear afferents of animals partially deafened by such procedures permit one to ascertain the role that the missing elements normally play in the encoding of sound (reviewed by Harrison, 1988). When inner hair cells are destroyed, the afferents formerly innervating them lose their spontaneous activity and do not respond even to very intense sounds at any frequency. When outer hair cells are destroyed, the CF thresholds for responses to tones of neurons innervating adjacent intact inner hair cells are selectively raised. In other words, the neurons lose some or most of their frequency tuning (Fig. 9). The implication is that the loss of outer hair cells somehow influences, in a frequency selective manner, the responses of Type I cochlear afferents, al-

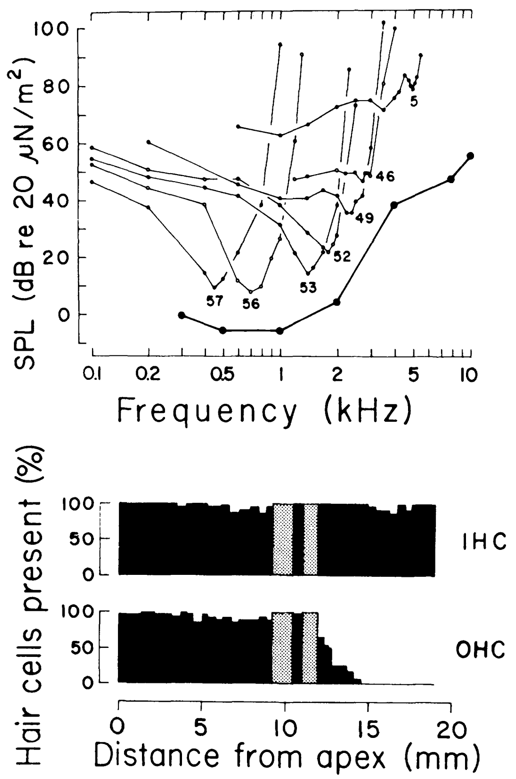

FIG. 9. Frequency-threshold neural tuning curves and behavioral audiogram (top panel) for a chinchilla whose basal cochlea is denuded of outer hair cells as a result of chronic administration of kanamycin, an ototoxic antibiotic (bottom panel). The thin lines in the top panel show three relatively normal tuning curves and four others with threshold elevations at CF. The thick line indicates the behavioral audiogram. The bottom panel graphically shows the location of surviving hair cells. (The gray bars indicate cochlear regions where, for technical reasons, hair cells could not be counted.) Note that threshold elevation of both the tuning curves and the behavioral audiogram occurs at high CFs, corresponding to the basal locations of outer hair cell loss. Reproduced by permission from Dallos and Harris (1978), Figs. 2 and 9.

though these innervate exclusively the (intact) inner hair cells.

Another effect of outer hair cell loss is to eliminate combination tones when the CF of the site of loss corresponds to *either* the primary frequencies or the frequency of the combination tone itself (Sec. 2.4). Such effects, demonstrated both in neurophysiological studies in animals (Siegel *et al.*, 1982) and in psychophysical measurements in hearing-impaired humans (Smoorenburg, 1972) indicate that the cochlear site with CF equal to the combination-tone frequency is the place that the brain attends to in detecting a combination tone, while the cochlear region with CFs equal to the primary tones is responsible for the generation of the mechanical distortion product: the combination tone propagates apicalward on the basilar membrane to reach the site with CF equal to its frequency, where it is transduced by the inner hair cells and its presence is relayed to the brain. The effect of hair cell loss upon combination-tone generation at the site of the primaries is very significant: it suggests that hair cells normally feed back upon, and thus influence, the mechanics of the basilar membrane (Goldstein, 1967; Smoorenburg, 1972). As we shall discuss in Sec. 4.5, it is now almost certain that a bidirectional (feedback) relationship indeed exists between outer hair cells and the basilar membrane.

2.6 Effects of Stimulation of the Olivocochlear Efferent System upon Cochlear Afferent Responses to Sound

Important clues to the processes of mechanical-to-neural transduction have emerged from experiments in which the responses of cochlear afferents to sound have been recorded during transient alterations of the functional state of the cochlea. One of these manipulations of the cochlea consists of electrical stimulation of the olivocochlear efferent system, which originates in the region of the superior olivary complex of the brainstem (see Secs. 7 and 11.1). A large proportion of these efferents make up the crossed olivocochlear bundle, which projects to the cochlea from the contralateral superior olivary complex and passes close to the dorsal surface of the brainstem, where it is easily accessible for electrical stimulation. The terminations of the crossed olivocochlear neurons are principally upon the outer hair cells and secondarily upon the Type I afferent neurons, close to their synapses with inner hair cells (Fig. 1 and Sec. 4.1). When the olivocochlear bundle is electrically stimulated, the responses to sound of the vast majority of cochlear afferents are inhibited in a frequency-selective manner: whereas the threshold of responses to tones near each afferent's CF may be raised by 20 dB or more, responses at other frequencies are hardly affected (Wiederhold, 1986). It should be noted that, while some terminals of the crossed olivocochlear bundle innervate Type I afferents directly (Fig. 1) and could conceivably inhibit them, this inhibition can hardly be CF specific, since neurons are essentially untuned signal-transmission devices. Qualitatively, therefore, the effect of olivocochlear efferent activation is similar to the effect of outer hair cell loss and poses the same puzzle: how can a CF-specific effect be mediated by the outer hair cells, given that Type I afferents innervate only inner hair cells? It will be seen in Secs. 4.3 and 4.5 that the answer to the puzzle appears to be that normal function of the outer hair cells is necessary for the maintenance of sharp tuning in the basilar membrane, which in turn determines the frequency tuning of inner hair cells and, ultimately, of cochlear afferents.

3. THE INPUT TO THE COCHLEA: THE EXTERNAL AND MIDDLE EARS

3.1 External Ear

The external ear of humans, chinchillas, and many other mammals consists of a roughly cylindrical canal, terminated internally by the tympanic membrane (eardrum), and flaring out externally into a cartilaginous, skin-covered pinna (Fig. 10). The geometry of the ear canal causes frequency-selective enhancement of sound-pressure waves at its tympanic-membrane end, relative to pressures at its entrance. The enhancement is greatest for sound frequencies with wavelengths about four times the length of the ear canal. In the chinchilla the canal length amounts to 3 cm and, correspondingly, as shown in Fig. 11, sound pressures at frequencies close to 2.8 kHz are increased by as much as 15 dB (von Bismarck, 1967). Figure 11 also shows that the frequency-selective pressure enhancement varies according to the location of the sound source relative to the ear. This directional selectivity results from such factors as shadowing of sound waves by the pinna and head (when the sound source is located on the opposite side of the head or widely off-axis from the ear) and destructive interference between directly arriving waves and waves scattered off the edge of the pinna. There is evidence that the external ear's directional selectivity permits the discrimination of sounds originating behind the head from those located in front. While human pinnas are essentially immobile, many mammals can orient their (often large) pinnas voluntarily, thus aiding in sound localization.

3.2 Middle Ear Anatomy

The middle ear of mammals (Fig. 10) consists of a mostly bone-enclosed, air-filled cavity containing the bony (ossicular) chain that transfers sound-induced vibrations from the peripheral "window" of the middle ear, the tympanic membrane, to the membranous oval window, which is the input to the fluid-filled cochlea; another membranous window, the round window, also communicates with the cochlea. The middle ear pressure is usually kept equalized with atmospheric pressure via the Eustachian tube, which opens into the pharynx. The ossicular chain is made up of three tiny bones: the malleus ("hammer"), the incus ("anvil"), and the stapes ("stirrup"). The malleus is firmly attached peripherally to the tympanic membrane and centrally to the incus; in some animals, such as the chinchilla, the malleus and incus are completely fused. Both the malleus and incus have single "long" projections which swing jointly around a shared axis. The long projection of the incus is joined to the stapes by means of a ligament. The stapes footplate abuts against the oval window, which it almost entirely fills. In addition to several ligaments that lightly suspend the ossicles from the middle ear's walls, two muscles, the stapedius and the tensor tympani, are attached to the stapes and to the malleus, respectively.

3.3 Impedance Matching by the Middle Ear

The function of the middle ear is to transfer efficiently the acoustic pressure waves from air into the fluid medium of the cochlea (see reviews by Zwislocki, 1975, and Møller, 1983). The cochlea consists of three fluid-filled, tubelike compartments, two of which (scala vestibuli and scala tympani) sandwich between them the third compartment, the scala media. The sensory organ proper, the organ of Corti, is located between scala media and scala tympani (Fig. 13). The middle ear ossicles direct the output of the external ear, namely, the vibrations of the tympanic mem-

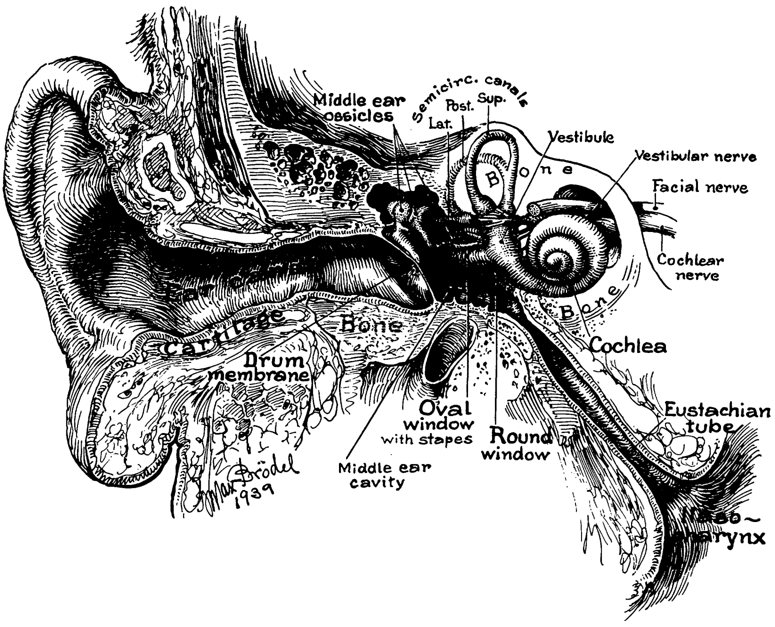
Middle ear ossicles
Semicirc. canals
Post.
Sup.
Lat.
Vestibule
Bone
Vestibular nerve
Facial nerve
Cochlear nerve
Cochlea
Eustachian tube
Naso-pharynx
Cartilage
Bone
Drum membrane
Oval window with stapes
Round window
Middle ear cavity
Brödel 1939

FIG. 10. A cut-out view of the right ear of a human. Reproduced from Brödel (1946).

brane, into one cochlear window, the oval window (adjacent to scala vestibuli), while shielding the second cochlear window, the round window (in contact with scala tympani), from the free-field sound waves that cause eardrum vibrations. Were the two cochlear windows exposed to the same sound field, pressure waves at the windows would interfere destructively and there would be little transfer of energy from the free field into the cochlea.

Because of the different acoustic impedance (i.e., pressure divided by volume velocity) of air and cochlear fluids, if sound waves were to impinge directly upon the oval window, most of their energy would be reflected and little would be available to drive the mechanical-to-electrical cochlear transduc-

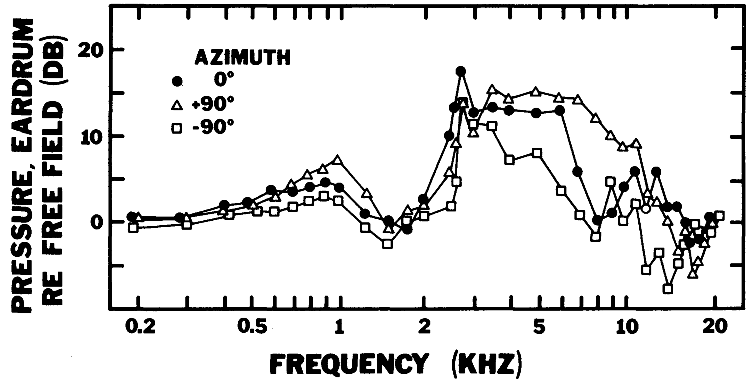

FIG. 11. The average acoustic pressure gain near the eardrum of the chinchilla, relative to pressure in the free field, plotted as a function of frequency. Symbols indicate measurements with the sound source facing the chinchilla's snout (azimuth zero, filled circles), closest to the experimental ear (azimuth of +90°, triangles) and closest to the opposite ear (azimuth of −90°, squares). Redrawn from von Bismarck (1967), Fig. 15.

tion processes. Thus, an important aspect of middle ear function is its role as an impedance-matching device that converts the relatively high-velocity, low-pressure air vibrations external to the tympanic membrane into higher-pressure, lower-velocity vibrations at the oval window. This impedance transformation is accomplished in two principal ways: (1) via the large area ratio between the middle ear input, the tympanic membrane, and its output, the footplate of the stapes; and (2) via the mechanical lever ratio of the malleus–incus complex. In the chinchilla, the area of the tympanic membrane is 0.56 cm^2 and the stapes footplate area is 0.02 cm^2 (Vrettakos *et al.*, 1988). Because the tympanic membrane vibrates little near the attachment to its bony frame, its effective area is reduced to about 0.37 cm^2. The pressure at the stapes footplate is therefore increased by a ratio of $0.37/0.02 = 18.5$ times. The long projections of the chinchilla malleus and incus are 0.45 and 0.16 cm in length, respectively, resulting in a lever ratio of $0.45/0.16 = 2.81$ (Vrettakos *et al.*, 1988). Because this lever ratio both reduces the magnitude of motion by a factor of 2.81 and multiplies force by the same amount, it increases the pressure-to-velocity ratio (i.e., the impedance) by $(2.81)^2 = 7.9$ times. Thus, the combined effects of the tympanic membrane/stapes footplate area ratio and of the malleus/incus lever ratio is to reduce the apparent cochlear input impedance by $18.5 \times 7.9 = 146$ times. The specific cochlear input impedance (i.e., pressure in scala vestibuli near the stapes divided by stapes velocity) at 1 kHz is about 16 000 dyn $\mathrm{s/cm}^3$ (Ruggero *et al.*, 1990). The transformer ratio of 146 reduces this impedance to $16\,000/146 = 110$ dyn $\mathrm{s/cm}^3$, which approaches (within 9 dB) the specific impedance of air, 42 dyn $\mathrm{s/cm}^3$.

3.4 Transfer Function of the Middle Ear

The calculation of the transformer ratio of the middle ear provides a first-order estimate of its signal transformations but it neglects relevant factors such as the compliance of the tympanic membrane and of the air trapped in the middle ear cavity, the elasticity and friction at the joints between the ossicles, and the inertia of the ossicles, ligaments, and muscles, all of which tend to impair transmission of vibrations to the cochlea. The effects of these factors are frequency dependent: in systems that include elastic and inertial components, elasticity tends to control low-frequency responses, while inertia is the dominant factor at high frequencies. The simplest way to assess the effectiveness of middle ear transmission is to measure its output, normalized to a constant input (i.e., constant sound pressure level at the eardrum), as a function of frequency. Since the middle ear ossicles vibrate linearly even at rather high stimulus intensities (e.g., 100 dB SPL), such a transfer function completely specifies its input-output characteristics and permits computation of the middle ear output for any arbitrary input. The output of the middle ear may be taken to be either the motion of the stapes footplate at the oval window or the pressure generated by this motion in scala vestibuli (pressure in scala vestibuli being given by the product of cochlear input impedance and stapes volume velocity, itself the product of the velocity of the stapes times the area of its footplate). The magnitudes of both these quantities in the chinchilla are plotted in Fig. 12 against frequency, expressed as stapes volume velocity per unit of stimulus pressure, and as pressure in scala vestibuli relative to stimulus pressure.

Both the velocity and pressure curves are mildly tuned, with similar bandpass characteristics in the frequency range 300–16 000 Hz; this similarity is consistent with the cochlear input impedance being resistive and constant in the midfrequency range (Zwislocki, 1975). The pressure curve has a maximum gain, at about 1 kHz, of 32 dB relative to pressure near the eardrum. This gain is very close to that predicted from the combined effects of the tympanic membrane/stapes footplate area ratio and of the force gain due to the malleus/incus lever ratio: $18.5 \times 2.8 = 51.8$ times, or 34 dB. However, at frequencies above and especially below 1 kHz, the pressure gain diminishes. The positive stapes velocity/frequency slope, which is almost exactly 6 dB/oct, is principally due to the air trapped within the middle ear, which acts like a cushion opposing the motion of the eardrum, shunting away some of the energy which otherwise would be spent in moving the ossicular chain. The milder negative slope of the pressure and velocity curves above 1 kHz is not yet well understood; in particular, it seems that it cannot solely reflect inertial

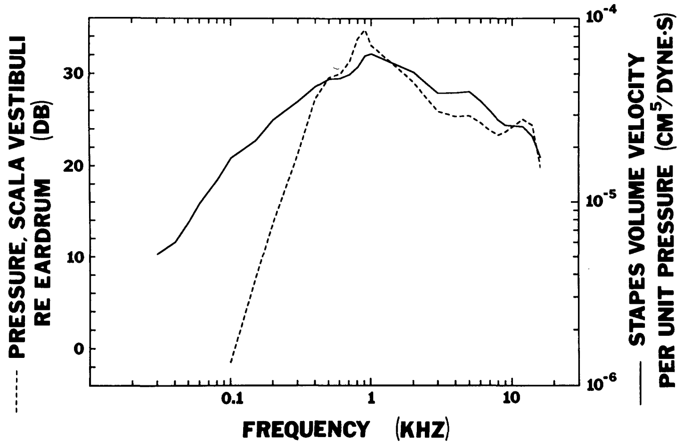

FIG. 12. Alternative transfer functions of the chinchilla middle ear. Shown are the volume velocity of the stapes per unit pressure in the external ear canal near the eardrum, expressed in c.g.s. units (continuous line, right ordinate), and the pressure in scala vestibuli near the stapes, expressed in decibels relative to pressure at the eardrum (dashed line, left ordinate). The curves are based on data from Ruggero *et al.* (1990) and Décory *et al.* (1990).

components of the middle ear, since their contributions to the transfer function should cause a −6 dB/oct slope.

3.5 Middle Ear Muscles

The effect of contraction of the middle ear muscles, the stapedius and the tensor tympani, is to stiffen the ossicular chain by pulling on the stapes and the malleus, respectively. Accordingly, middle ear muscle contraction interferes mainly with the transmission of low-frequency sounds (by as much as 15 dB) and has little effect on ossicular responses to sound frequencies above 1.5 kHz. Bilateral middle ear muscle contraction can be elicited by intense acoustic stimulation of either ear, by activity of the muscles involved in chewing and vocalization, and during the rapid-eye-movement stage of sleep. The function of the middle ear muscles is obscure. Their sound-activated contraction, the acoustic reflex, could play some role in protecting the ear against intense prolonged sounds (although not against impulsive sounds, because of the long latency of the reflex). However, such a role is questionable from an evolutionary perspective, due to the near absence of intense natural (i.e., non man-made) low-frequency sounds. Because contraction of the middle ear muscles produces selective attenuation of low-frequency transmission, their activation may be useful in reducing the masking of meaningful mid- and high-frequency sounds by low-frequency noise.

4. THE COCHLEA

4.1 Anatomy of the Cochlea

4.1.1 Cochlear Compartments. The cochlea derives its name from the Greek word for "land snail" whose coiled and cone-shaped shell it very much resembles in most mammals [Figs. 10 and 13(a)]. It consists of two fluid-filled tubes, with bony external walls and membranous inner walls (scala vestibuli and scala tympani), flanking a third tube (scala media) which is also fluid filled [Figs. 13(a) and 13(b)]. Scala vestibuli and scala tympani are separated for most of their length (18 mm in chinchillas) by scala media; because scala media stops short of the apical end of the cochlea, scala vestibuli and scala tympani communicate with each other via a channel at the apex, called the helicotrema. Scala media is separated from scala vestibuli by a thin cellular membrane (Reissner's membrane) and from scala tympani by the organ of Corti, which includes the sensory hair cells, and by the basilar membrane [Figs. 13(b) and 13(c)]. At the basal end of the cochlea scala vestibuli ends at the oval window, a flexible membrane against which the stapes footplate abuts (Fig. 10). Also at the basal end of the cochlea, another thin and flexible membrane, the round window, separates scala tympani from the middle ear (Fig. 10).

Perilymph, the fluid which occupies scala vestibuli and scala tympani, is similar in composition to cerebrospinal fluid (which surrounds and cushions the central nervous system) and other extracellular fluids in that it

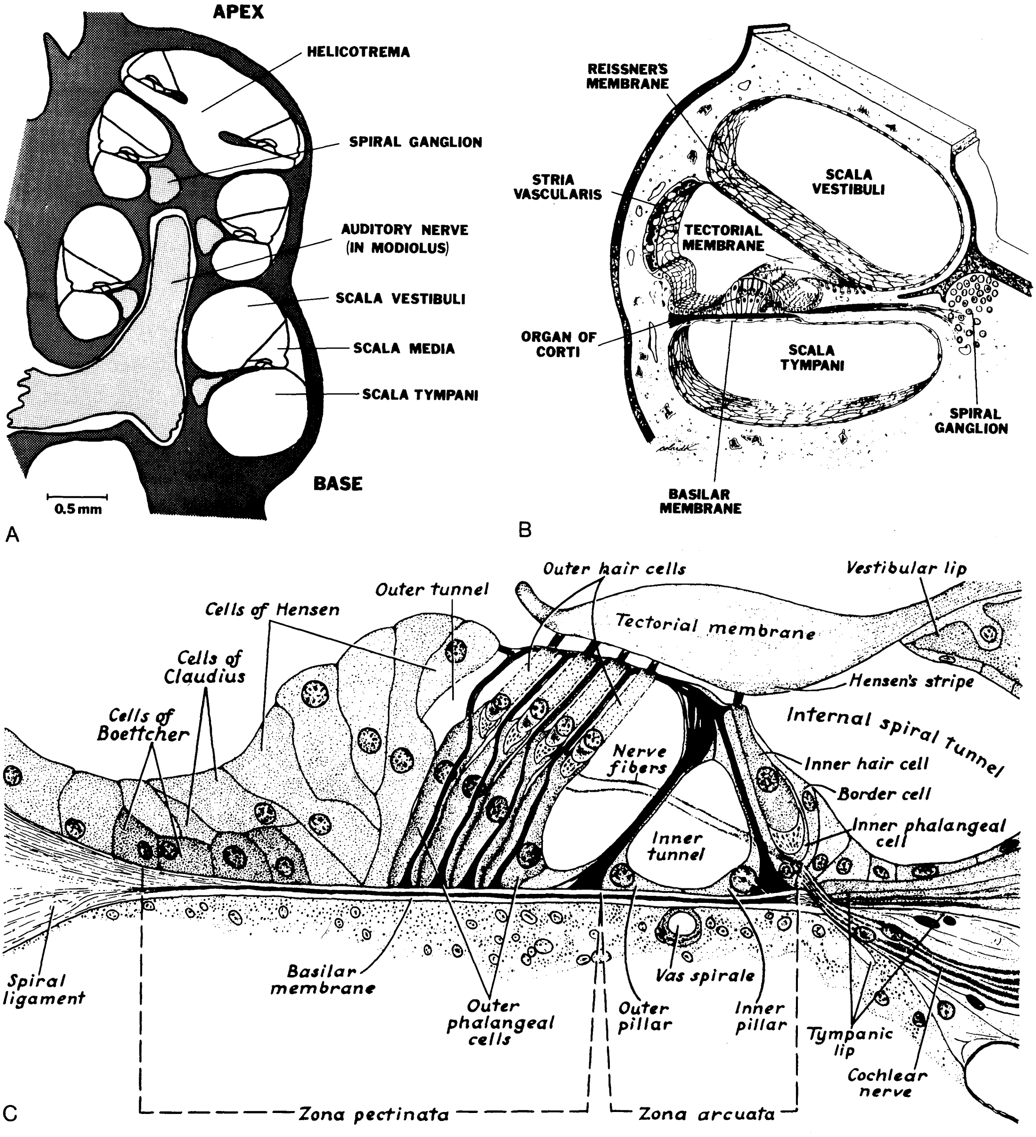
APEX
HELICOTREMA
SPIRAL GANGLION
AUDITORY NERVE (IN MODIOLUS)
SCALA VESTIBULI
SCALA MEDIA
SCALA TYMPANI
BASE
0.5 mm
A
REISSNER'S MEMBRANE
STRIA VASCULARIS
TECTORIAL MEMBRANE
SCALA VESTIBULI
ORGAN OF CORTI
SCALA TYMPANI
SPIRAL GANGLION
BASILAR MEMBRANE
B
Outer hair cells
Vestibular lip
Outer tunnel
Cells of Hensen
Tectorial membrane
Cells of Claudius
Hensen's stripe
Cells of Boettcher
Internal spiral tunnel
Nerve fibers
Inner hair cell
Border cell
Inner phalangeal cell
Inner tunnel
Spiral ligament
Basilar membrane
Outer phalangeal cells
Vas spirale
Outer pillar
Inner pillar
Tympanic lip
Cochlear nerve
C
Zona pectinata
Zona arcuata

FIG. 13. Anatomy of the cochlea. Magnification and detail increase from (a) through (c). (a) A cross section of the entire cochlea of a chinchilla. The darker shading indicates bone; the lighter shading represents the cochlear nerve, with its axons in the central pillar of the cochlea (modiolus) and its cell bodies (somata) in the spiral ganglion. Note that because of their spiral nature ($3\frac{1}{2}$ turns in 18 mm), the three scalae (vestibuli, tympani, and media) and the spiral ganglion are cut in several places as they coil around the central axis of the cochlear cone. (b) A slice of the three scalae at a single cochlear site. (c) Details of the organ of Corti. (b) and (c) are reproduced by permission from Bloom and Fawcett (1975), Figs. 36-11 and 36-14.

contains high sodium-ion and low potassium-ion concentrations. In contrast, endolymph—the fluid filling scala media—is unique among extracellular fluids: it contains high potassium-ion and low sodium-ion concentrations. Also in contrast with all other extracellular compartments in the body, scala media sustains a rather high electrical potential, +80 mV relative to the perilymph or distant extracellular sites. Both the unusual ionic composition and the high electrical potential of the endolymph are generated by energy-consuming metabolic pumps located in cells of the stria vascularis, a highly vascularized "ribbon" located in the lateral wall of scalamedia [Fig. 13(b)]. The high positive electrical potential of the endolymph drives positively charged potassium ions into the hair cells. As discussed in a later section (Sec. 4.4), the transduction process modulates this potassium current.

4.1.2 Basilar Membrane and Supporting Cells. The sensory hair cells are located in the organ of Corti, which rests on the basilar membrane and upon which, in turn, lies the tectorial membrane; the organ of Corti's main components also include the "supporting" cells and the peripheral terminals of afferent and efferent neurons [Fig. 13(c); reviewed by Lim, 1980]. The basilar membrane is held (but not under tension) at its central edge by the bony spiral lamina and at its peripheral edge by the spiral ligament. It consists largely of acellular materials, mostly protein fibrils arranged radially (relative to the modiolus) within an amorphous matrix. The thickness of the basilar membrane diminishes systematically from base to apex, while its width increases from the base to the apex. In the chinchilla the basilar membrane is 200 μm wide near the stapes and 420 μm wide at the helicotrema (Bohne and Carr, 1979; Lim, 1980).

Several types of supporting cells rest on the basilar membrane, some of them providing a pedestal for the hair cells [Fig. 13(c)]. The most prominent supporting cells are the inner and outer pillars, which contain fibrils that impart stiffness to them; together they divide the organ of Corti longitudinally into a narrow inner region, close to the spiral lamina, and a wider outer region. Because of its overall geometry, the outer region of the organ of Corti, which contains three rows of outer hair cells, is capable of relatively wide transverse movement; in contrast, the inner region, which includes a single row of inner hair cells, moves little as a whole, since it rests partly on the immobile edge of the bony spiral lamina. The outer phalangeal cells support the outer hair cells and, like the pillar cells, contain stiffening fibrils. A thin beam (phalanx) reaches up from each outer phalangeal cell to join the tops (apical surfaces) of inner and outer hair cells, inner pillars, and other supporting cells to form the regular cellular mosaic of the reticular lamina. The reticular lamina constitutes a (probably stiff) structural boundary of the scala media and also an ion barrier: the extracellular fluid bathing the basolateral walls of hair cells and supporting cells is perilymph or is perilymph-like in composition, while the apical surface of the hair cells and supporting cells are in contact with the endolymph of scala media.

4.1.3 Inner and Outer Hair Cells. Hair cells of the mammalian cochlea, like all other sensory hair cells (such as those in the vestibular organs of balance), are characterized by tufts of "hairs" or stereocilia, thin cylinders that protrude from their apical surface. Each stereocilium consists of a core, made of protein (actin) filaments, enveloped by plasma membrane, which is continuous with that of the rest of the cell. The actin molecules are organized into orderly arrays, almost crystalline in appearance, which make the stereocilium quite stiff. Inner and outer hair cells differ in the arrangement of their stereocilia, inner hair cells having fewer and thicker stereocilia. In both cases, stereocilia are arranged as a letter "v," with its concavity pointing toward the modiolus. In the inner hair cell, the "v" is quite shallow; in the outer hair cell, in which the stereociliar arrangement is reminiscent of the ranks of pipes in an organ, the "v" is deeper. The tallest of the outer hair cell stereocilia are well attached to the overlying tectorial membrane, whereas in inner hair cells such attachment is flimsy, if it exists at all. Inner and outer hair cells also differ in the shape of their bodies, in the organization of certain cytoplasmic organelles, and in their spatial relations to supporting cells. Inner hair cells are flask shaped, with a narrow apical region and a wider, round basal region. Their plasma membrane is completely surrounded by closely juxtaposed supporting cells. Outer hair cells, in contrast, have cylindrical bodies and only the basal (bottom) part of their basolateral plasma membrane is in direct contact with supporting (outer phalangeal) cells. Most of the outer hair cell basolateral plasma membrane is underlain by a membranous cistern; this structure is absent in inner hair cells.

4.1.4 Afferent and Efferent Innervation. The mode of innervation, both afferent and efferent, is fundamentally different in inner and outer hair cells (reviewed by Spoendlin, 1988). Inner hair cells, comprising only 20% of the hair-cell population, receive terminals from 90% of cochlear afferents, whereas only

10% of afferents innervate outer hair cells, which constitute the vast majority (80%) of hair cells. In the chinchilla, each of the 1800 inner hair cells is innervated by 10–20 of the 27 000 Type I cochlear afferents, with each afferent innervating a single inner hair cell (Bohne *et al.*, 1982; Ruggero *et al.*, 1982). In contrast, each of some 2700 Type II afferents innervates many of the numerous (7400) outer hair cells. While outer hair cells receive many synaptic endings of the olivocochlear efferents, no such terminals exist on inner hair cells. However, there are olivocochlear efferent endings on the terminals of Type I afferents, near their synapse with the inner hair cell (see Sec. 11.1). Thus one may justifiably think of an "inner hair cell system" and an "outer hair cell system." The inner hair cell system is the one whose output we have described in terms of the responses of the Type I afferents (Secs. 2.1–2.6); of the responses of Type II afferents, which presumably convey the signals transduced by the outer hair cells, we know almost nothing.

4.2 Passive Aspects of Basilar Membrane Motion

Until less than two decades ago, the first words of this section's title would have been redundant: all aspects of cochlear mechanics were viewed as "passive" and independent of the "active" metabolic processes of living cells and, hence, as essentially similar in live and dead animals. Whatever differences did exist between the mechanics of living and dead ears were deemed unimportant. That view of cochlear function had its roots, at least partially, in the historical accident that what was known empirically about basilar membrane motion was uncovered almost exclusively by measurements in ears of human cadavers and of dead animals. Such measurements were the solitary work of an extraordinary Hungarian-born physicist, George von Békésy. What follows in this section summarizes this work, for which Békésy was awarded the 1961 Nobel Prize for Physiology or Medicine.

4.2.1 The Traveling Wave. When the ossicular chain vibrates at very low (infrasonic) frequencies, the perilymph of the scala vestibuli and scala tympani is set in motion as a whole, with inward displacement of the stapes being accompanied by bulk flow of fluid from scala vestibuli, through the helicotrema, into scala tympani, and by an outward bulge of the round window into the middle ear cavity. At higher (audio) frequencies, the motion of the stapes generates pressure waves that propagate very rapidly (in a few microseconds) throughout the cochlea. Békésy, using relatively insensitive microscopy methods that required very intense stimulation (120–140 dB SPL), observed in cochleae of human cadavers that the basilar membrane sustained a much slower (transverse) displacement wave that always propagated (longitudinally) from the basal cochlea toward the apex, independent of the site of application of the vibration input [Fig. 14(a)]. Thus, although the usual physiological route for airborne sound is through vibration of the oval window by the stapes, Békésy found identical base-to-apex propagation patterns of

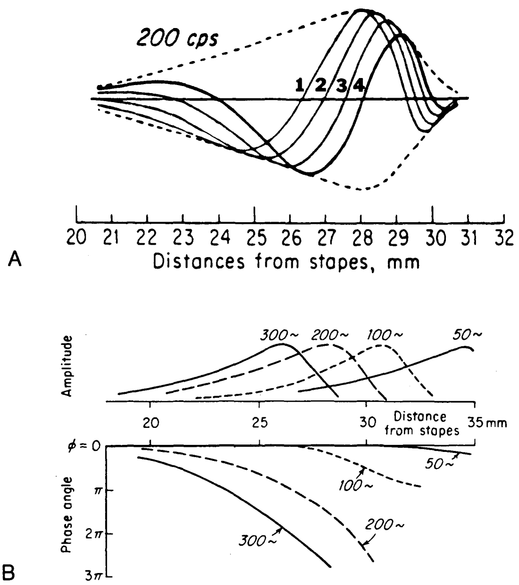

FIG. 14. The basilar-membrane traveling wave, as measured by Békésy in human cadavers. (a) "Snapshots" at four times (1 = earliest) of the basilar membrane displacement patterns elicited by a 200-Hz tone; the scale of the ordinate is greatly exaggerated relative to that of the abscissa. The dashed lines indicate the peak displacements of the basilar membrane toward scala vestibuli and scala tympani at each cochlear location. (b) Peak-to-peak displacements as a function of distance for responses to tones with four different frequencies (upper section), and the corresponding phase patterns (lower section). Reproduced by permission, with modifications, from Békésy (1960), Figs. 12-17 and 11-58.

the displacement wave when he applied sound to the cochlea via an artificial apical window or via bone conduction. The speed of travel of the displacement wave is not uniform as a function of distance: at the basal end propagation is fast and therefore, for sinusoidal stimulation, the wavelength is long; as the wave proceeds toward the apex, it slows down and the wavelength shortens. The finite speed of travel implies an accumulating delay as a function of distance from the stapes; for sinusoidal stimulation this translates into accumulating phase lag [Fig. 14(b); see also Fig. 8(b)]. The longer latency of responses to clicks of low-CF cochlear afferents relative to the latency of high-CF afferents (Sec. 2.3) originates in the time required by the traveling wave to reach the apical region of the basilar membrane. This time, about 2 ms in the chinchilla cochlea, is two to three orders of magnitude greater than the time required by sound to travel between the oval window and the helicotrema.

4.2.2 Frequency Tuning and Tonotopicity. Whatever the frequency of stimulation, the shape of the vibration envelope was found to be similar in that, in each case, the envelope's basal portion (corresponding to the region of long wavelengths) had a shallow spatial gradient, there was a maximum at a particular longitudinal location, and the maximum was followed on the apical side by a relatively steep slope [Fig. 14(a)]. For low-frequency stimuli, the displacement of the basilar membrane was largest near the apex. As stimulus frequencies were increased, the displacement envelopes were confined to progressively more basal locations [Fig. 14(b)]. Thus, a tonotopic map is laid out along the cochlea, in which each longitudinal location of the basilar membrane vibrates maximally at some particular frequency, its CF. Because Békésy found that basilar membrane responses grew linearly with stimulus level, it was considered justifiable to estimate, by extrapolation, basilar membrane displacements at levels much lower than those used in the experiments. At intensities corresponding to the lowest threshold of hearing in normal humans (about 0 dB SPL), such extrapolation suggested that the basilar membrane vibrates at CF with peak amplitudes of 1 pm, approximately equivalent to 1/100 of the radius of a hydrogen atom.

For several apical locations of the human basilar membrane, Békésy was able to measure responses to tones as a function of frequency. For a CF of 150 Hz, the tuning curve was broad, with a Q_{10} of only 0.8. [For comparison, recall that in chinchillas Q_{10}'s of cochlear afferents are as low as 1 for CFs around 150 Hz and as high as 7 for CFs of 16 kHz (Sec. 2.2).] When basilar membrane recordings were obtained in the basal (high-CF) cochlear region of live guinea pigs, they yielded responses that were as poorly frequency-tuned as those from Békésy's measurements in the apical (low-CF) region of human cadaver cochleae. The mismatch in the live guinea pig between the poorly tuned responses of the basal basilar membrane and the sharply tuned responses of afferents innervating the same cochlear region led to the belief that a "second filter," presumably located in the hair cells, must further filter the mechanical responses before encoding takes place in the cochlear nerve (Evans and Wilson, 1975).

4.2.3 Physical Basis of Passive Cochlear Mechanics. In addition to measurements of basilar membrane responses to sound, Békésy also measured the static compliance of the basilar membrane; he accomplished this by sealing the helicotrema, applying hydrostatic pressure to scala tympani, and measuring the deformation of Reissner's membrane at various locations. As might be expected from the longitudinal variations in basilar membrane width and thickness, Békésy found that at the cochlear base the (narrow and thick) basilar membrane is relatively stiff, becoming progressively more elastic toward its (wider and thinner) apical end. The variation of basilar membrane elasticity as a function of distance (amounting to a ratio of 1 to 100 between the extremes of the human cochlea) is thought to be the physical characteristic that dominates the passive aspects of the hydrodynamics of the cochlea. Accordingly, mathematical and hardware models of the cochlea that consist of little more than two fluid-filled chambers separated by a membrane with smoothly varying elasticity are able to mimic, at least qualitatively, the main features of cochlear motion observed by Békésy in cadaver ears, including the existence of displacement waves that travel from base to apex and a tonotopic or-

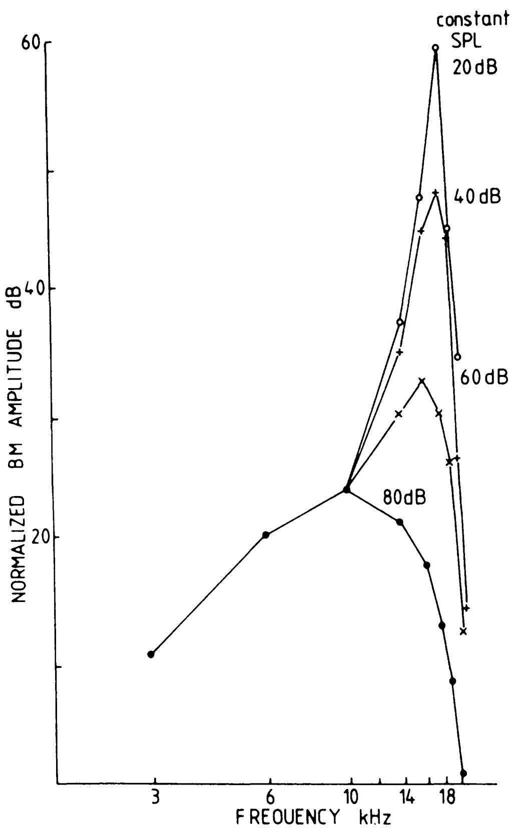

FIG. 15. Frequency response of the guinea pig basilar membrane at a site 3.4 mm from the cochlear base. Displacement gains (i.e., displacement magnitudes divided by stimulus levels) are expressed in decibels. Note that sensitivity near CF diminishes at higher stimulus levels (parameter). Because of this (frequency-selective and nonlinear) change in sensitivity, the frequency response is sharply tuned at low stimulus levels and poorly tuned at high levels. Reproduced with permission from Johnstone *et al.* (1986), Fig. 4.

ganization, with high-frequency responses localized to the basal end.

4.3 Active Aspects of Basilar Membrane Motion

4.3.1 Basilar-Membrane Nonlinearities and their Lability. The earliest indication that the metabolic state of the cochlea might be crucial in determining basilar membrane response came from experiments by Rhode (1971) in the cochleae of live (anesthetized) squirrel monkeys, using the then new Mössbauer technique. This technique, based on the detection of the Doppler shift of gamma radiation due to motion, permits the measurement of much lower velocities (e.g., 0.1 mm/s, equivalent to 1.6 nm at 10 kHz) than is possible using classical microscopy. In many monkeys, Rhode found that basilar membrane responses to near-CF tones grew with stimulus intensities at rates lower than linear, while response growth was linear at other frequencies. Because of the frequency-selective nature of the (compressive) nonlinearity, basilar membrane responses were more sharply tuned for low-level stimuli than for high-level ones (see Fig. 15). Most significantly, the CF responses became linear and less sensitive after the death of the animal, while off-CF responses remained largely unchanged (see Fig. 16).

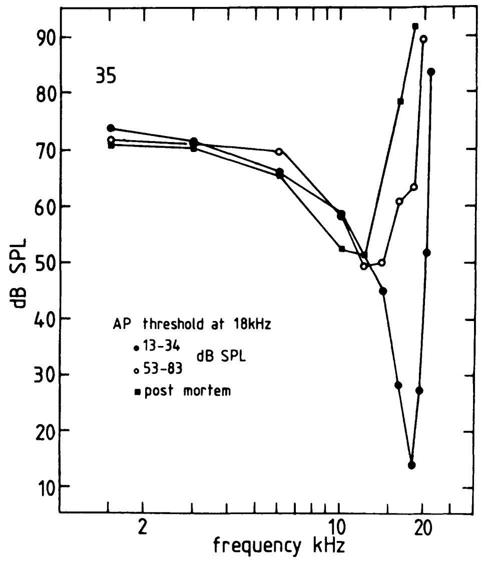

FIG. 16. Correlation between basilar membrane tuning and neural thresholds in a guinea pig cochlea. The basilar membrane recording site is near that of Fig. 15. The isovelocity curves indicate sound pressure levels, as a function of frequency, required to elicit a basilar membrane velocity of 0.04 mm/s. The state of the cochlea was assessed by recording the threshold levels for evoking compound action potentials ("AP": the synchronized response of many cochlear afferents) in response to tone pips with frequency identical to the CF of the basilar membrane recording site. Initially, the basilar membrane was sharply tuned (closed circles) and neural thresholds, accordingly, were low (13–34 dB SPL); as neural thresholds rose (53–83 dB SPL), basilar membrane responses deteriorated (open circles) and tuning was nearly lost. A further basilar membrane response loss at CF occurred upon the animal's death (squares). Reproduced by permission from Sellick *et al.* (1982), Fig. 15(b).

For about a decade, Rhode's findings were controversial because they could not be replicated (e.g., Evans and Wilson, 1975). In retrospect, it is evident that what prevented replication was irreversible damage to the cochlea caused unwittingly by experimenters during their surgical manipulations preparatory to mechanical recordings from the basilar membrane. Eventually, however, the nonlinear growth of near-CF basilar membrane responses was confirmed, first in the guinea pig (Sellick *et al.*, 1982) and then in the chinchilla (Robles *et al.*, 1986). Figure 15 shows response magnitude curves, as a function of frequency, for a healthy guinea pig basilar membrane site located 3 mm from the basal end. In order to facilitate comparison of tuning and sensitivity as a function of stimulus level, the curves have been normalized to stimulus magnitude (i.e., they are presented as gain curves, with the parameter being stimulus level). At moderate and intense stimulus levels, the tuning is broad. At low stimulus levels, however, responses are sharply tuned because of the selective sensitivity increase at near-CF frequencies; responses at off-CF frequencies are linear and remain constant.

Figure 16 demonstrates that, even before death of the animal, deterioration of the 3-mm guinea pig cochlear site, as indicated by threshold elevation of the compound response of the cochlear nerve ("AP threshold"), is accompanied by loss of sensitivity of the basilar membrane mechanical response at near-CF (18 kHz) frequencies. The sensitivity loss, which amounts to more than 40 dB, converts the normal, highly frequency-selective tuning curve into a shallow band-pass curve, reminiscent of those obtained by Békésy in human cadavers. The fact that the mechanical sensitivity loss occurs at near-CF frequencies also suggests a unifying explanation for the CF specificity of the effects of hair cell loss and olivocochlear efferent stimulation upon cochlear afferent responses (Secs. 2.5 and 2.6): it is hypothesized that both of these effects arise ultimately from alterations of the basilar membrane mechanical response.

4.3.2 Mechanical Origin of Cochlear Afferent Response Properties. Although it has not yet been directly shown that the effects upon cochlear afferents of outer hair cells loss or activation of the olivocochlear efferent system are due to underlying parallel effects upon the basilar membrane, it is clear that the normal frequency selectivity of cochlear afferent responses derives, largely or entirely, from mechanical (i.e., basilar membrane) frequency tuning. This is illustrated in Fig. 17, which shows the frequency tuning of the basilar membrane at a site 3.5 mm from the basal end of the chinchilla cochlea, together with an average neural tuning curve derived from responses of cochlear afferents with appropriate CF (about 8.4 kHz). Comparison between the neural and mechanical tuning curves suggests that, at neural threshold, the basilar membrane peak displacement is about 2 nm; at the behavioral hearing threshold, presumed to be 20 dB lower than neural threshold (Fig. 2), basilar membrane displacement would be 0.2 nm. Although very small, this estimate (equivalent to the diameter of a hydrogen atom or 1/50 of the thickness of a neuron's plasma membrane) is more than two orders of magnitude larger than the threshold displacement derived by extrapolation from Békésy's measurements in cadaver cochleae. Excellent matches between neural and mechanical tuning curves have been demonstrated in guinea pig (Sellick *et al.*, 1982) and cat (Khanna and Leonard, 1982), as well as chinchilla. It seems, therefore, that the previously hypothesized "second filter" (Sec. 4.2.2) is unnecessary, since the sharpness of tuning appears to be fully established in the "first" filter, i.e., the basilar membrane. In the same manner that the sharp frequency selectivity of cochlear afferent responses originates in basilar membrane mechanics, a mechanical counterpart of neural two-tone suppression (Sec. 2.4.1) also exists at the level of the basilar membrane: it is possible to reduce the basilar membrane response to a CF tone by introduction of a second tone with frequency either above or below CF, which may or may not cause basilar membrane vibration when presented alone (Robles *et al.*, 1989). As is the case for sharp frequency selectivity and nonlinear growth of responses, two-tone suppression in the basilar membrane is physiologically vulnerable, being reduced by cochlear deterioration and disappearing after the animal's death.

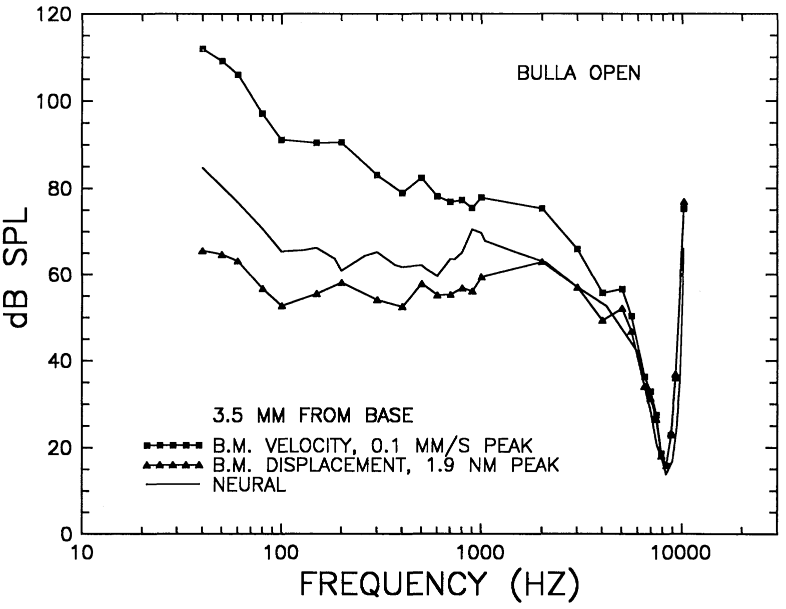

FIG. 17. Comparison of frequency tuning in the basilar membrane and in the auditory nerve of the chinchilla. The basilar membrane data, presented as both isovelocity (0.1 mm/s, square symbols) and isodisplacement (1.9 nm, triangular symbols) curves, are averages of recordings from five chinchillas at sites about 3.5 mm from the basal end of the cochlea. The neural curve (no symbols) is an average of recordings from many cochlear afferents with CFs close to that of the basilar membrane recording site. There is an excellent match between neural and mechanical tuning but it is unclear whether neural tuning is more closely related to basilar membrane displacement or to its velocity. Slightly modified from Ruggero *et al.* (1990), Fig. 22.

4.4 Hair Cell Receptor Potentials

In the absence of effective stimulation, inner and outer hair cells are electrically polarized across their plasma membrane, their cytoplasm being 45 and 70 mV, respectively, negative relative to the fluid that bathes their basolateral surfaces, or −125 and −150 mV relative to the endolymph that contacts their stereocilia. Upon displacement of their stereocilia bundles along their axis of symmetry (Sec. 4.1), hair cells respond by depolarizing or hyperpolarizing, depending on whether the displacement is toward the vertex of their "v" or away from it, respectively (Russell *et al.*, 1986). These electrical potential changes result from modulation of standing currents, carried by potassium ions, that flow passively across the apical surfaces of hair cells, from the endolymph (with its high positive potential) into their (negative-potential) cytoplasm (see reviews by Nuttall, 1986, and Roberts *et al.*, 1988). In the case of outer hair cells there is little doubt that basilar membrane displacement moves their bodies, hence causing shearing of the stereocilia tips against the tectorial membrane (Davis, 1965); thus, displacement of the basilar membrane toward the scala vestibuli leads to displacement of the stereocilia tips toward the "v" vertex, causing depolarization, while basilar membrane displacement toward scala tympani causes hyperpolarization. Such response polarity is in accordance with the firm attachment between the tallest outer hair cell stereocilia and the tectorial membrane; it is also in agreement with the radial position of outer hair cells in the region of the basilar membrane where transverse motion is greatest. In the case of the inner hair cell, the situation is unclear: their stereocilia are not firmly attached to the tectorial membrane and may not contact it at all; in addition, since inner hair cells are located at the edge of the bony spiral lamina, their bodies may move little upon basilar membrane displacement. Stimulation of inner hair cells, therefore, may be mediated by motion of the endolymph or the tectorial membrane.

Whatever the exact micromechanical mode of their stimulation, inner and outer

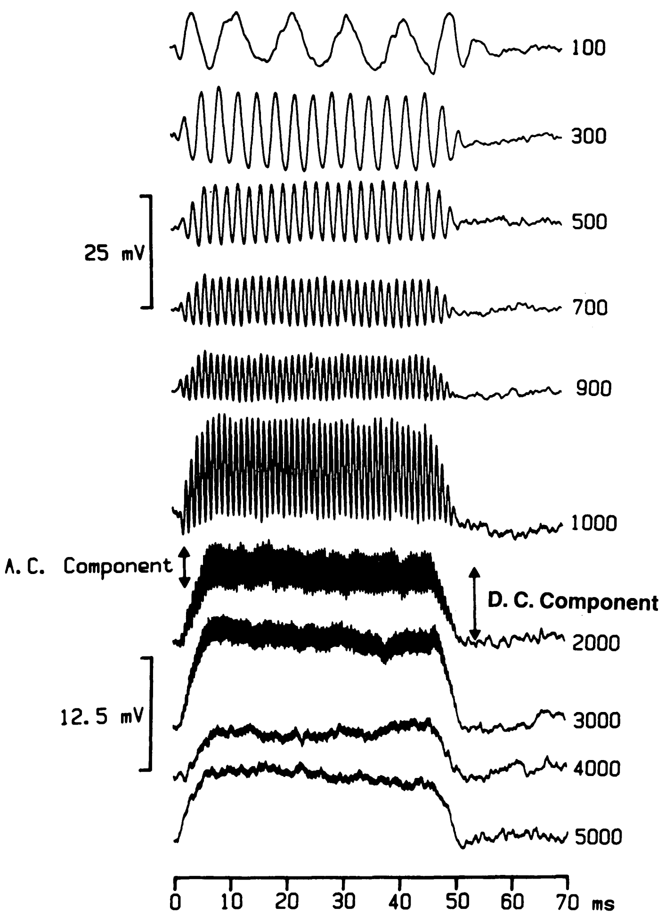

FIG. 18. Inner hair cell receptor potentials in response to tone pips. The hair cell was located at the base of a guinea pig cochlea and had a CF close to that of the inner hair cell in Fig. 19. The frequency (in Hz) of the tone-pip stimulus is indicated at the right of each trace. The lower ordinate scale applies to the bottom five traces. Note that the ac component is insignificant in responses to the higher-frequency tone pips. Reproduced by permission from Palmer and Russell (1986), Fig. 9.

hair cells respond to acoustic stimuli with intracellular voltage oscillations that, to some extent, mimic the stimulus waveshape (Fig. 18). The inner hair cell recordings of Fig. 18 were obtained from the same (basal) region of the guinea-pig cochlea from which the basilar membrane data of Figs. 15 and 16 were obtained. When the acoustic stimulus consists of a low-level, low-frequency tone pip (e.g., 300 Hz, top of Fig. 18), the intracellular voltage response consists of a relatively faithful replica of the stimulus. Such ac response is believed to account for the phase-locked responses of cochlear afferents; as in chemical synapses elsewhere in the nervous system, action potentials are triggered principally during the depolarizing half cycle of the receptor potential. As the stimulus frequency is raised, the ac voltage response diminishes progressively, probably as a result of capacitive shunting across the hair cell's plasma membrane. Nevertheless, high-CF inner hair cells are able to signal the presence of high-frequency acoustic stimuli by means of a (nonlinear) depolarizing dc shift. The frequency-dependent shift from an ac-dominated response to a dc-predominant response matches the progressive frequency-dependent deterioration of phase locking in high-CF cochlear afferents (Palmer and Russell, 1986).

By measuring the stimulus levels at which a given depolarization magnitude is attained as a function of frequency, it is possible to construct receptor-potential tuning curves comparable to those for the responses of cochlear afferents (Figs. 2, 7, and 9) or of the basilar membrane (Figs. 16 and 17). Such iso-response curves are presented in Fig. 19 for an inner hair cell from the basal end of the guinea pig cochlea. For dc responses of 2.7 mV, the inner hair cell is sharply tuned to a CF of about 23 kHz. For higher dc-response levels, tuning becomes less prominent be-

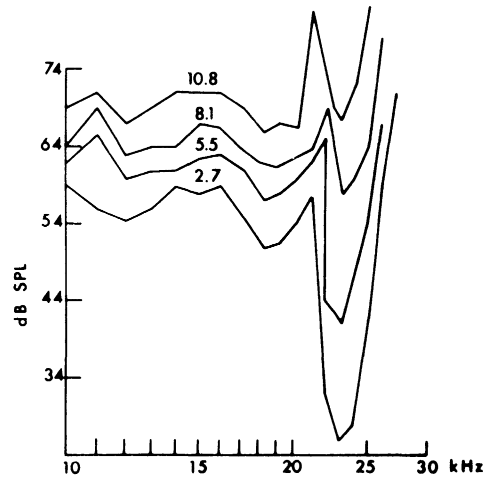

FIG. 19. Isopotential tuning curves for the dc responses to tone pips of a guinea pig inner hair cell. The parameter indicates the magnitude of the dc voltage response. Note the dissimilar rates of response growth at frequencies near CF and at lower frequencies; also note the sharper frequency tuning at lower response levels. Reproduced by permission from Russell and Sellick (1977), Addendum, Fig. 2(d).

cause responses at near-CF frequencies grow at slower-than-linear rates, while responses at other frequencies grow linearly. In most respects, including frequency tuning, the responses to sound of outer hair cells are similar to those of inner hair cells. Comparison of low-level dc tuning curves (such as those in Fig. 19) with threshold tuning curves of cochlear afferents with the same CF indicate that the frequency selectivity of cochlear afferents is a direct reflection of that of the inner hair cells they innervate. A further generalization is possible: with few exceptions, each inner hair cell's response to sound is an accurate voltage analog of the instantaneous rate of firing in cochlear afferents that innervate it. The most important exception is the absence of adaptation in hair cell potentials; adaptation of cochlear afferent responses, which is believed to arise at the chemical synapses between inner hair cells and cochlear afferents (Sec. 2.2.2), causes the "overshoot" at the onset of neural responses to tone pips and some of the nonlinear intensity-dependent aspects of neural responses to clicks (Sec. 2.3).

Among cochlear-nerve response properties that have clear counterparts in the voltage responses of inner hair cells (reviewed by Nuttall, 1986) are two-tone suppression (Sec. 2.4) and the inhibitory effects of activation of the olivocochlear efferents (Sec. 2.6). In the case of electrical stimulation of the olivocochlear efferents, which causes CF-specific inhibition of inner hair cell responses, intracellular recordings have shown that no transmembrane current change accompanies efferent stimulation; this is what is expected, given the absence of efferent synapses upon inner hair cells. These findings are significant because they strengthen the conclusion (based on cochlear afferent recordings; see Secs. 2.5 and 2.6) that the outer hair cells can somehow determine the responses of Type I cochlear afferents. While other modes of action of outer hair cells upon inner hair cells are conceivable, the most parsimonious mechanism is one in which outer hair cells are able to alter the mechanical input to the inner hair cell transduction process.

4.5 The Cochlea as a Sound Generator

4.5.1 Otoacoustic Emissions. The discovery by Kemp (1978) of otoacoustic emissions, sounds that are emitted by the ear, was one of the central developments that revolutionized our understanding of inner-ear physiology during the last decade. Kemp showed that when the normal human ear is stimulated by impulsive-type sounds (clicks or short tone pips) the ear responds, after a fairly long delay (about 10–15 ms), by broadcasting narrow-band, usually low-frequency echoes (Fig. 20). As stimulus levels are raised, these echoes grow at less-than-linear rates and stop growing altogether (saturate) when they reach levels of about 20 dB SPL. Subsequent to Kemp's initial discovery, he and others have demonstrated, in many species, identical or related phenomena, which are collectively known as otoacoustic emissions (Zurek, 1985). Among the most spectacular, particularly when intense, are the narrowband spontaneous otoacoustic emissions. As their name implies, spontaneous otoacoustic emissions do not require an external stimulus and emanate continuously from the ears of a substantial proportion, indeed perhaps the majority, of humans and primates, and also from the ears of some other animals. An extreme instance among the latter is a dog that produces intense (60 dB SPL), high-pitched "whistles" from both of its ears, audible from across a quiet room (Ruggero *et al.*, 1984).

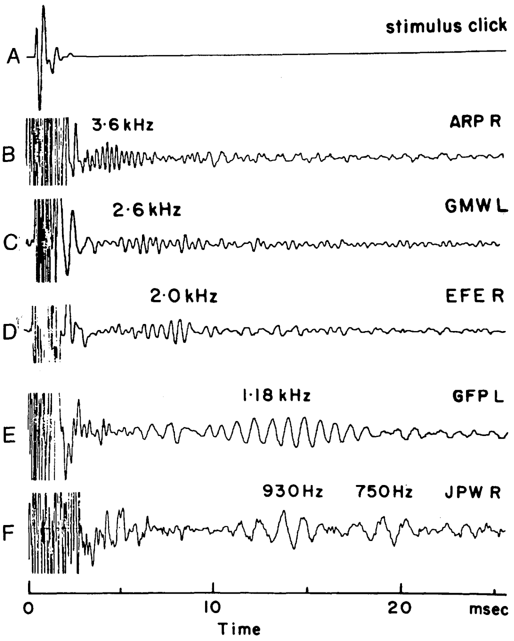

FIG. 20. Impulsively evoked otoacoustic emissions from human ears. The upper tracing indicates the time wave form of the acoustic-click stimulus. The lower five traces show averages of the acoustic wave forms recorded by microphones inserted in the external ear canals of the subjects (identified by name initials and "L" or "R" for left and right ear, respectively). Note that each ear produces different emissions and that lower-frequency emissions have longer latencies. Reproduced by permission from Wilson (1980), Fig. 2.

The more usual levels of spontaneous otoacoustic emissions, however, do not exceed 20 dB SPL.

While it is doubtful that spontaneous otoacoustic emissions play any role in normal hearing (and, in fact, there are reasons to believe that they may result from minor pathology of the organ of Corti), their existence is highly significant: it demonstrates that biological (biochemical) energy can be transduced in the ear into mechanical energy (i.e., sound). Indeed, upon stimulation by low-frequency tones (<2000 Hz), the cochlea responds by broadcasting tones at the same frequencies. The existence of such low-frequency emissions, known as "stimulus-frequency otoacoustic emissions," suggests that the cochlea may be capable of producing echoes to sounds throughout the audible frequency spectrum. The ability to generate sound implies the ability to generate motion. An important clue to the cellular nature of this motility is that subjects affected by sensorineural deafness (usually caused by hair cell loss) in a given frequency range lack otoacoustic emissions over the same frequency range, implying that hair cells are necessary for the generation of otoacoustic emissions.

Another clue to the origin of cochlear motility comes from the experimental manipulation of a fourth type of otoacoustic emissions, namely, distortion-product otoacoustic emissions. In Sec. 2.4.2 it was noted that cochlear afferents respond to certain pairs of tones (primaries) as if in fact three or four tones were present in the cochlea, namely, the two primaries plus one or two combination tones (Fig. 8). An acoustic correlate of such a phenomenon has been demonstrated in the external ear canals of humans and other animals: when two primaries (with frequencies f_1 and f_2) are presented, a sensitive microphone detects an extra (combination) tone, with frequency equal to $2f_1-f_2$. It is possible to alter the level of the combination tone by manipulation of the endolymphatic potential (Sec. 4.1.1) and, most significantly, by electrical stimulation of the crossed olivocochlear efferent system (Mountain, 1980). Since these efferents terminate mostly upon outer hair cells (see Sec. 4.1)—the remainder of their innervation being upon afferents, which can hardly be expected to act as effectors—it is concluded that the synaptic activity of the efferent terminals upon outer hair cells is capable of altering cochlear mechanics. A corollary is that outer hair cells must be motile (i.e., capable of self-induced or active motion), at least to the extent that they are capable of transducing a synaptic (chemical) efferent effect into pressure or displacement.

4.5.2 Outer Hair Cell Motility. Demonstration that outer hair cells are indeed motile, at least under certain circumstances, has been obtained in experiments in which the organ of Corti has been disaggregated into its cellular components and studied microscopically. When isolated outer hair cells are exposed to an electric current generated between two extracellular electrodes, they respond mechanically by shortening or lengthening, depending on the polarity of the current. Identical stimuli fail to cause motion responses in inner hair cells or in supporting cochlear cells (Brownell *et al.*, 1985). If the electrical stimulation consists of ac current, outer hair cell motion can follow the stimulus, albeit with amplitudes that are inversely related to its frequency, at least up to 8000 Hz (Ashmore and Brownell, 1986). There are many as yet unspecified details of the hypothetical feedback mechanism whereby the outer hair cells might act as frequency-tuned motion amplifiers. Nevertheless, the outer hair cell motility demonstrated *in vitro* goes a long way in accounting for the existence of otoacoustic emissions, the fragility of the sharply frequency-tuned response of the basilar membrane, and the hitherto poorly understood but undeniable role of outer hair cells in several other cochlear phenomena (see review by Dallos, 1988). In particular, if it is presumed that the motile response of the outer hair cell is responsible for boosting the low-level, near-CF responses of the basilar membrane *in vivo* (which has not yet been shown directly), then the apparent paradox posed by the frequency-specific effects upon cochlear afferents both of outer hair cells loss and of electrical stimulation of the olivocochlear efferents (see Secs. 2.5 and 2.6) is resolved thus: both effects are mediated by removal or alteration of the feedback normally provided by the outer hair cells upon basilar membrane motion; the altered basilar membrane motion causes concomitant changes in the receptor potential of the inner hair cells and in the responses of afferents that innervate them.

5. OVERVIEW OF THE CENTRAL AUDITORY SYSTEM

The remaining sections of this chapter, devoted to the central auditory system, are less extensive than the preceding account of the auditory periphery. This emphasis on peripheral, rather than central, auditory processes is admittedly disproportionate, because the connections of the auditory centers of the brain are vastly more complex and numerous than are those in the auditory periphery; however, the focus on the periphery is necessitated at present by our still limited understanding of central auditory mechanisms. The central auditory pathways are arguably the most complicated of all sensory systems. Even a simplified diagram of the major neural pathways (Fig. 21) reveals a maze of connections, loops, and branches: pathways conveying information from each ear converge at multiple levels of the brain and each terminus receives inputs from multiple sources. What cannot be appreciated from the diagram is the considerable heterogeneity of structural and functional properties within each central auditory nucleus. Presumably, the divergence and convergence of the multiple pathways permit the abstraction, within the auditory nuclei, of various features of the acoustic stimulus and, ultimately, their synthesis into the percepts embodied in the hearing sensation.

Our account of the central auditory system begins with an outline of the structural features of the principal auditory centers of the brainstem (in ascending order: the cochlear nuclear complex, Sec. 6; the superior olivary complex, Sec. 7; the midbrain, Sec. 8; and the auditory thalamus and cortex, Sec. 9). These sections provide a basis for Sec. 10, which deals with many of the functional aspects of central auditory processing, including responsiveness to frequency, level, sound location, and complex sounds. Although it is more usual to treat central auditory system physiology on a structure-by-structure basis, examining the properties of neurons at each level of the system in isolation from other levels, we shall adopt a perspective encompassing the entire central auditory system, comparing properties at different levels. The chapter concludes with a discussion of centrifugal auditory pathways (Sec. 11) that may regulate the afferent flow of acoustic information.

6. THE COCHLEAR NUCLEAR COMPLEX

After entering the brainstem, each auditory nerve fiber branches to provide innervation to three divisions of the ipsilateral cochlear nuclear complex (Fig. 22): one branch goes to the anteroventral cochlear nucleus (AVCN), while the second branch projects to the posteroventral cochlear nucleus (PVCN) before ascending to innervate the dorsal cochlear nucleus (DCN). The fiber bifurcation occurs ventrally for low-CF fibers and progressively more dorsally for higher-CF fibers. Each fiber then terminates in an orderly fashion, as illustrated in Fig. 22, to maintain its relative position within the frequency matrix

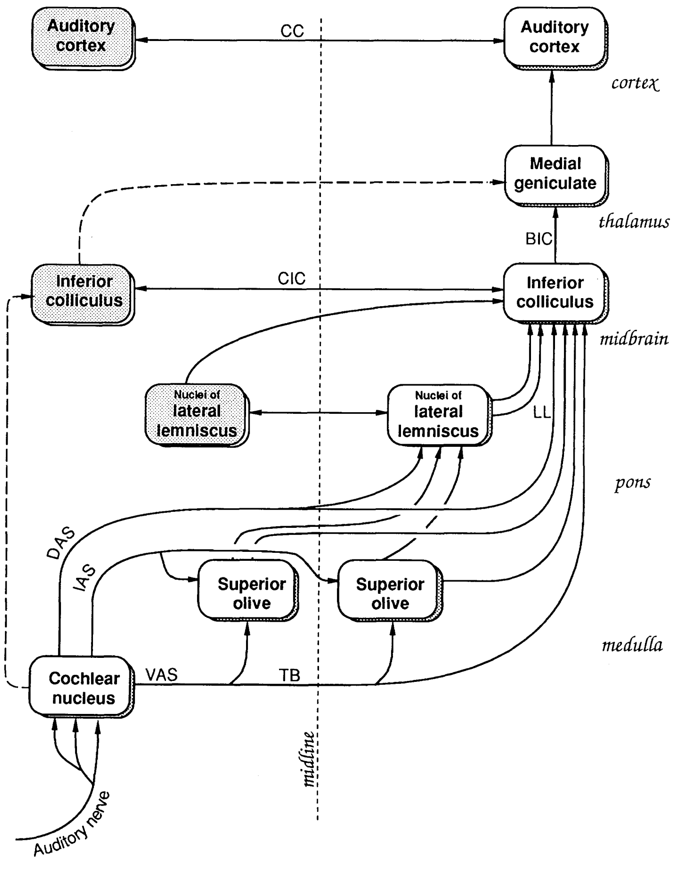
Auditory cortex
CC
Auditory cortex
cortex
Medial geniculate
thalamus
BIC
Inferior colliculus
CIC
Inferior colliculus
midbrain
Nuclei of lateral lemniscus
Nuclei of lateral lemniscus
LL
pons
DAS
IAS
Superior olive
Superior olive
medulla
Cochlear nucleus
VAS
TB
midline
Auditory nerve

FIG. 21. A simplified diagram of the central auditory system, illustrating some of its many branches and connections. Only primary ascending pathways are shown, although many secondary and descending projections are recognized. For clarity, and because pathways are generally routed to the contralateral side of the brain, projections are shown as arising from a single ear. At higher levels on the left side (gray background) only connections with the contralateral hemisphere are shown. Broken lines indicate minor pathways. DAS, IAS, VAS: dorsal, intermediate, and ventral acoustic stria, respectively; TB: trapezoid body; LL: lateral lemniscus; BIC: brachium of the inferior colliculus; CC: corpus callosum.

(and, hence, its relative location within a map of the cochlear partition). Because the input to each division of the cochlear nuclear complex derives from the entire cochlea, each division contains a complete spatial representation of stimulus frequency. This tonotopic order is maintained in the ascending outputs of the cochlear nuclear complex and throughout the ascending central auditory pathways.

The three divisions of the cochlear nucleus project to higher levels of the auditory system in three major fiber bundles (Fig. 21). Although there are sparse projections that ascend the auditory system on the same side, the three major tracts are directed medially and contribute to pathways that ascend primarily on the contralateral side. One tract, the ventral acoustic stria, carries fibers destined for binaural interaction at the core nuclei of the superior olivary complex. Another tract, the intermediate acoustic stria, provides input to "periolivary" regions surrounding the core binaural nuclei. The third tract, the dorsal acoustic stria, bypasses the superior

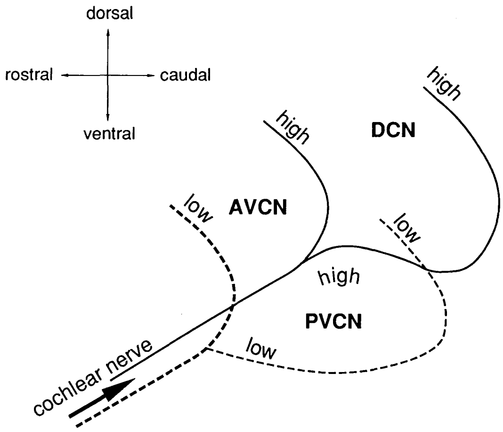
dorsal
rostral
caudal
ventral
high
high
DCN
low
AVCN
low
high
PVCN
low
cochlear nerve

FIG. 22. A schematic, sagittal view of the cochlear nucleus showing the pattern of innervation by auditory nerve fibers. Each fiber bifurcates as it enters, sending one branch to the anteroventral cochlear nucleus (AVCN). The second branch projects to the posteroventral cochlear nucleus (PVCN) and then continues to innervate the dorsal cochlear nucleus (DCN). The tonotopic organization of fibers (from low to high CF) in the auditory nerve is mapped onto each subdivision of the cochlear nucleus. Adapted from Osen (1970).

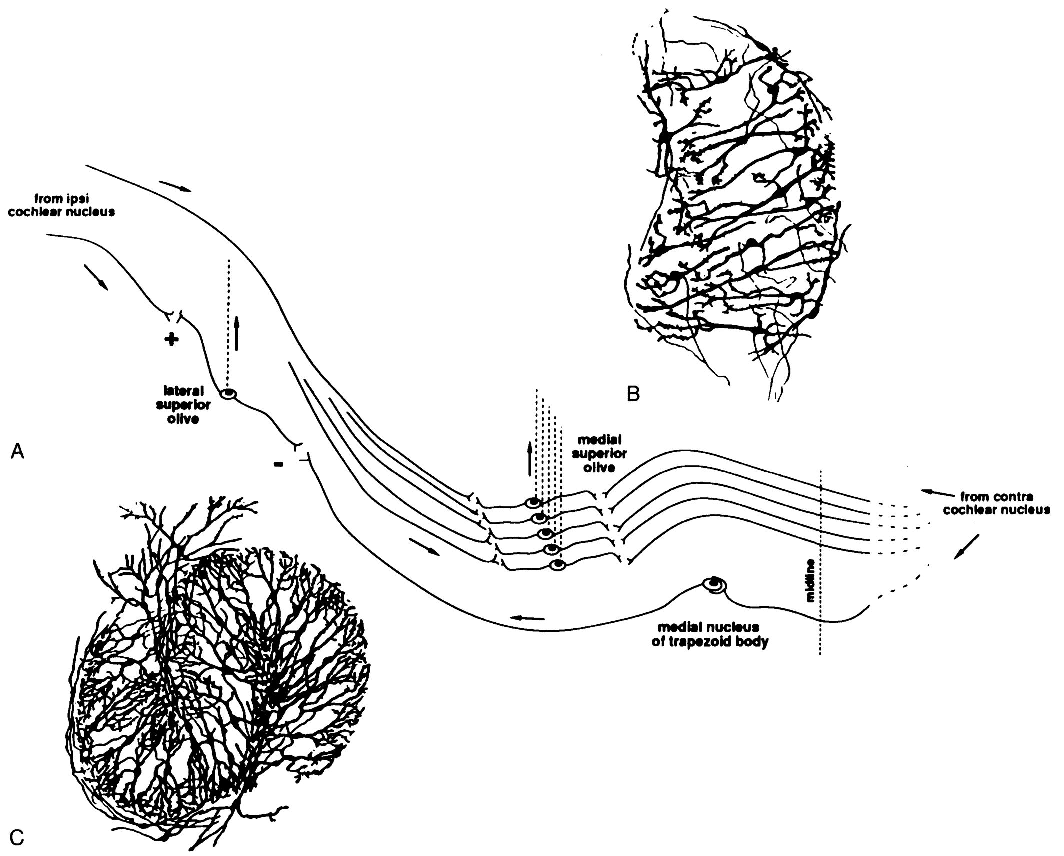
from ipsi cochlear nucleus
+
lateral superior olive
A
-
B
medial superior olive
from contra cochlear nucleus
midline
medial nucleus of trapezoid body
C

FIG. 23. Cellular organization and binaural convergence within the primary nuclei of the superior olivary complex (SOC). (a) A schematic illustration of the pathways converging from both cochlear nuclear complexes at SOC. Arrows indicate direction of information flow. Vertical dashed lines indicate the ascending outputs of the superior olive. (b) Lateral and (c) medial superior olivary neurons, as seen in a Golgi (silver-impregnation) preparation. Adapted from Ramón y Cajal (1909).

olivary complex entirely and has endings in the nuclei of the lateral lemniscus and in the inferior colliculus.

The projections arising from the cochlear nuclear complex originate from morphologically distinct neuron types. For example, the core binaural nuclei of the superior olive are innervated by neurons with bushy dendritic fields located in AVCN. Fibers of the dorsal acoustic stria arise primarily from fusiform neurons of the DCN whose dendrites are organized in a pyramidal form. A distinctive cell type of the PVCN, with a dendritic field described as octopus-shaped, sends its axon via the intermediate acoustic stria to the periolivary regions of the superior olivary complex. Thus, morphological specialization within the cochlear nuclear complex is associated with divergence in the pattern of its efferent pathways: the input provided by each auditory nerve fiber gives rise to multiple pathways that perform specialized functions in the processing of acoustic signals.

7. THE SUPERIOR OLIVARY COMPLEX

The superior olivary complex (Fig. 23) consists of three core nuclei—the lateral superior olive (LSO), the medial superior olive (MSO), and the medial nucleus of the trapezoid body—and several periolivary nuclei. The MSO and LSO provide the first opportunity in the auditory system for information received by the two ears to be compared and, hence, they may play a role in sound localization. In most species, the MSO is a thin sheet of elongate neurons whose dendritic fields appear bipolar when viewed in the frontal plane. As illustrated in Fig. 23(c), the major dendrites extend medially and laterally across the width of the nucleus, giving MSO the appearance of a highly organized stack of neurons. The LSO has a somewhat similar structure, but the neural sheet is folded [Fig. 23(b)]. The principal cells of LSO extend their dendrites in a bipolar fashion, perpendicular to the curvature of the nucleus. In both MSO and LSO, axons conveying information from the left ear project to the dendritic pole oriented towards the left, while axons derived from the right side influence dendrites oriented towards the right.

As illustrated schematically in Fig. 23(a), each neuron in MSO receives input from AVCN of both cochlear nuclear complexes (Warr, 1982). MSO neurons are sensitive to differences in the time of arrival of sounds at each ear, one of the major cues for sound localization. Although LSO is also a major site for binaural interaction, a direct projection from the cochlear nucleus comes only from the ipsilateral side [Fig. 23(a)]. Input derived from the contralateral ear comes indirectly via the third core nucleus of the superior olivary complex, the medial nucleus of the trapezoid body. It is generally assumed that the ipsilateral innervation has an excitatory influence whereas the contralateral input is converted to an inhibitory influence at the medial nucleus of the trapezoid body. The probability of an action potential being generated is increased by excitation and decreased by inhibition. Excitation and inhibition are indicated, respectively, as "+" and "−" in Fig. 23(a).

8. THE AUDITORY MIDBRAIN

8.1 Nuclei of the Lateral Lemniscus

The large tract of fibers known as the lateral lemniscus (Fig. 21) contains fibers from several sources in the cochlear nuclear and superior olivary complexes. Many of these fibers terminate at the inferior colliculus, but many also innervate the ventral and dorsal nuclei of the lateral lemniscus, two loose aggregations of neurons within the homonymous fiber tract. The ventral nucleus of the lateral lemniscus receives a strong projection from the contralateral cochlear nuclear complex, whereas the dorsal nucleus of the lateral lemniscus receives significant input from the ipsilateral medial superior olive and from both lateral superior olives (Glendenning and Masterton, 1983). The difference in the innervation patterns of the ventral and dorsal nuclei of the lateral lemniscus result in a major functional difference: the vast majority of neurons in the ventral nucleus respond only to stimulation of the contralateral ear, whereas those in the dorsal nucleus are usually binaural (Brugge *et al.*, 1970).

8.2 The Inferior Colliculus

The inferior colliculus is a prominence on the dorsal aspect of the midbrain. Its largest subdivision is the central nucleus (ICC), an ovoid core of densely packed cell bodies. Principal neurons are organized into a series of parallel sheets separated by afferent fibers, giving the central nucleus a laminated appearance. The dendritic fields of most neurons add to the lamination by ramifying only within the plane of their local neuronal sheet (Morest and Oliver, 1984). The ICC is the primary midbrain relay in the ascending auditory system and receives input from many sources in the lower brainstem (Fig. 21): contralaterally from the three divisions of the cochlear nuclear complex and the other inferior colliculus; ipsilaterally from the medial superior olive and ventral nucleus of the lateral lemniscus; and bilaterally from the lateral superior olive and dorsal nucleus of the lateral lemniscus (Aitkin, 1986).

Given its numerous sources of input, it is remarkable that ICC contains a single, very precise tonotopic organization. The foundation for this tonotopicity is the laminar organization: neurons within a given lamina have similar CFs; those located in dorsal laminae have low CFs and those located at successively more ventral sites have progressively higher CFs. Surrounding the ICC are fields of more diffusely distributed neurons that are not part of the primary auditory system and respond poorly to tones. Laterally, there is a region in which neurons responsive to sounds are intermingled with others sensitive to touch (Aitkin, 1986).

Most neurons in ICC project ipsilaterally via the brachium of the inferior colliculus to innervate the principal auditory relay of the thalamus, the ventral division of the medial geniculate body. Some fibers cross the midline in the commissure of the inferior colliculus and either ascend to the contralateral auditory thalamus or descend to innervate the contralateral inferior colliculus (Calford and Aitkin, 1983). In addition to the primary projection from ICC to the ventral division of the medial geniculate body, fringe regions of the inferior colliculus provide secondary projections to the dorsal and medial divisions of the medial geniculate.

9. THE AUDITORY THALAMUS AND CORTEX

9.1 The Medial Geniculate Body

A common parcellation scheme for the auditory thalamus divides the medial geniculate body into ventral, medial, and dorsal divisions. The ventral division contains cells of medium size organized in a laminated pattern similar to that found at the inferior colliculus. Intermingled with these principal cells, whose axons project to the auditory cortex, are interneurons with short axons that ramify only locally within the ventral division. These interneurons apparently modulate the activity of the principal cells by inhibitory circuits. Physiologically, the ventral division is characterized by sharp frequency tuning and tonotopicity. In contrast, the large, multipolar neurons of the medial division are often characterized by broad frequency tuning and complex response patterns that include habituation and variable discharge rate (Aitkin, 1973). Habituation and variability are also evident in responses of neurons in the superficial regions of the dorsal division, but sharper tuning and reliable, short latency responses can be found in deep dorsal regions (Calford and Aitkin, 1983).

9.2 The Auditory Cortex

The final destination in the primary central auditory pathway is the auditory cortex. Like the auditory midbrain and thalamus, the auditory cortex consists of a primary core surrounded by secondary fields. The primary auditory cortex (AI) is organized into functional cylinders or "columns" extending orthogonal to the cortical surface. Neurons in a given column usually respond to the same narrow band of frequencies. Columns representing similar frequencies are aligned, giving rise to a tonotopic organization of isofrequency slabs. In the cat, the species for which there is the most complete data concerning auditory cortex, there may be as many as seven cortical fields surrounding AI (Fig. 24); of these, the "anterior," "posterior," "ventral," and "ventro-posterior" fields are known to be organized tonotopically. In addition, there is a "secondary" auditory field (AII in Fig. 24) containing neurons that respond poorly to tones or have very broad frequency tuning. Little is known of the physiology of other auditory cortical fields.

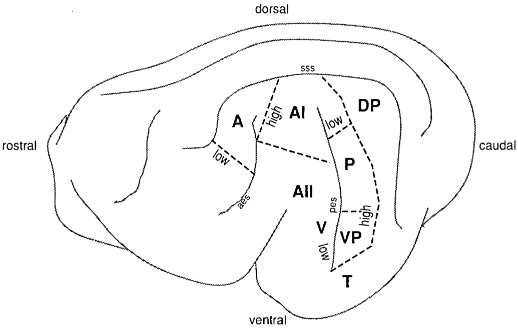

FIG. 24. A lateral view of the left cerebral hemisphere of the cat, showing the auditory cortical fields and their frequency organization. For the four tonotopic fields (A, AI, P, and VP), CF progressions from low to high are indicated, respectively, by "low" and "high." Significant portions of some fields are hidden within the three sulci (i.e., infoldings of the cortical surface): suprasylvian (sss), anterior ectosylvian (aes), and posterior ectosylvian (pes). Adapted from Imig and Reale (1980), Fig. 1.

Each cortical field receives input from multiple sources in the thalamus, but while the ventral division of the medial geniculate body provides the predominant source of input to the primary auditory cortex (AI), the nonprimary thalamic auditory nuclei figure more prominently in projections to the nonprimary auditory cortical fields. Apparently reflecting these differences in sources of input, auditory cortical neurons located in the fields surrounding AI often have lower discharge rates and greater response variability than are typical of AI. In addition to thalamocortical projections, there are numerous intracortical connections among the auditory cortical fields, and AI provides a significant projection within the corpus callosum to the contralateral AI (Imig and Reale, 1980).

9.3 Some Behavioral Consequences of Cortical Lesions

One way to explore the function of a brain structure is to examine the consequences of its removal. A striking result of extensive damage to the auditory cortex of both hemispheres is that a variety of auditory discriminations involving manipulations of pitch and loudness remain largely unimpaired. Deficits consequent to bilateral auditory cortical lesions include disturbances in detection of the duration of tone pips and the temporal patterning of tone sequences. It has been argued (Elliott and Trahiotis, 1972) that the common feature of such deficits is that they involve the differential response to more than one type of change in the auditory environment. One of the more profound deficits of unilateral auditory-cortical ablation is the inability to localize sounds in the contralateral sound field (Jenkins and Masterton, 1982).

10. RESPONSE PROPERTIES IN THE CENTRAL AUDITORY SYSTEM

10.1 Discharge Patterns

In contrast to the relative homogeneity of response patterns in the auditory nerve (Sec. 2), discharge patterns of central neurons are diverse. Patterns similar to those of auditory nerve fibers (Fig. 4) occur within lower brainstem nuclei, but are rare at levels above the midbrain. Even within the cochlear nucleus several other response types are common. An example of a "primarylike" pattern recorded in the superior olivary complex is shown in Fig. 25(a). A second pattern commonly observed in the cochlear nucleus and superior olive, but rarely seen at higher brain levels, is the "chopper" [Fig. 25(b)]. This pattern is characterized by regular, evenly spaced peaks that are reminiscent of phase locking. However, chopper peaks are not synchronized with the stimulus wave form, and the regularity is apparently generated by the resonant properties of the postsynaptic neural membrane (Wu and Oertel, 1984). "Onset" responses [Fig. 25(c)] are relatively uncommon at lower levels of the auditory system, but become increasingly more prevalent at higher levels. At auditory cortex, most responses in the anesthetized animal are of this form. The "pauser" pattern [Fig. 25(d)] is so named because of the pause in discharge that occurs just after the response onset. Pausers are common throughout the brainstem. Finally, there is the "buildup" pattern [Fig. 25(e)], which is unique in not containing a strong onset component. This pattern, seen at many levels of the auditory system, is most common in the dorsal cochlear nucleus.

Certain discharge patterns are well-correlated with cell morphology and innervation. For example, the primarylike patterns of spherical cells of AVCN are thought to arise from the specialized morphology of their input: a large synaptic terminal, the end bulb of Held, that may completely encapsulate the cell body (Brawer and Morest, 1975). The function of such a synapse appears to be to preserve temporal information. At the level of the cochlear nucleus, the onset pattern is rarely observed outside the "octopus cell" region of PVCN. These neurons are thought to integrate convergent inputs from fibers of different CF, thus enabling them to respond over an unusually broad range of frequency and intensity, with very precise onset timing (Smith, 1988). Chopper responses are the most common pattern observed in the ventral cochlear nucleus, apparently arising in multipolar neurons receiving a myriad of small bouton endings, which may derive from a small number of cochlear afferent fibers (Rhode and Smith, 1986).

10.2 Phase Locking

The temporal synchronization known as phase locking, evident in the discharges of auditory nerve fibers in response to low-frequency stimuli (Sec. 2.2.2), is preserved to some extent in the discharges of many central neurons within the brainstem. The upper frequency limit for phase locking is reduced progressively at each higher level of the system. At the auditory midbrain, synchronization is evident only at very low frequencies, and at the thalamus and cortex it is essentially absent. Phase locking is thought to play a critical role in binaural processing within the brain. In the cochlear nucleus, the end bulbs of Held (Sec. 10.1) minimize degradation of the temporal information at the synapse upon

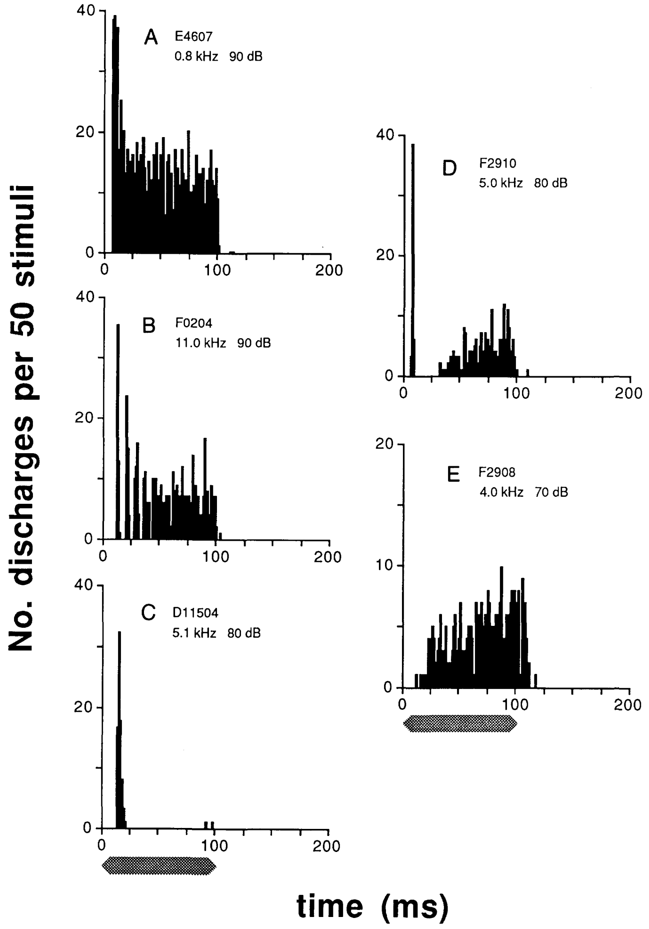

FIG. 25. Some discharge patterns of neurons in the superior olivary complex of the gerbil. Each post-stimulus-time histogram represents responses to 50 repetitions of a 100-ms CF-tone burst presented ipsilaterally at 2/s. The ordinate is the number of discharges per bin (see Sec. 2.2). Histograms illustrate five discharge patterns: (a) "primarylike," (b) "chopper," (c) "onset," (d) "pauser," and (e) "buildup." Unpublished data of M. N. Semple.

AVCN neurons. This permits the second-order fibers destined for binaural convergence in the superior olive to retain phase synchronization and hence to allow extraction of interaural phase difference information, a major cue for sound localization (Sec. 10.6.1). More generally, phase locking also offers a potential means for encoding low-frequency information, perhaps augmenting the frequency information inherent in the (tonotopic) central representation of cochlear place (Sec. 2.2.2).

10.3 Frequency Tuning

For a given CF, the frequency-tuning curves of most neurons of the primary (or principal) central auditory pathways are similar to those seen at the auditory periphery (Fig. 2), suggesting that frequency tuning in the central auditory system is essentially determined by the tuning at the basilar membrane (Fig. 17). Neurons whose responses reflect binaural influences usually have similarly tuned inputs from each ear (Semple and Kitzes, 1985). The sharp frequency tuning of the primary central pathways is contrasted by broad and sometimes quite irregular tuning in regions of the surrounding, diffuse auditory system (Secs. 8.2, 9.1, 9.2), such as the medial division of the auditory thalamus (Aitkin, 1973). Specification of a single CF is often difficult in these nonprimary areas.

When spontaneous activity is sufficiently high, it is possible to demonstrate that excitatory threshold tuning curves of many central neurons are flanked by inhibitory areas. These inhibitory sidebands, which can occur above and/or below the excitatory band, may serve to sharpen the excitatory tuning. Frequency selectivity at suprathreshold levels, assessed by determining discharge rate as a function of frequency for a variety of SPLs, is usually similar to that of auditory nerve fibers (Sec. 2.2.1 and Fig. 3). In some neurons, however, the frequency dependence of discharge rate varies considerably with SPL, so the "best" frequency can be more level dependent than is the case in cochlear afferents.

10.4 Response Latency

Response latency reflects mechanical delays within the cochlea (Sec. 2.3 and 4.2.1), conduction time along ascending axons, and the delay at each synapse. Thus, the minimum latency to the initial discharge at primary auditory cortex is several milliseconds longer than that occurring in the cochlear nucleus. For most central auditory neurons, the initial discharge latency is negatively correlated with stimulus level. For a given SPL, the latency is minimal at CF and becomes progressively longer as the frequency departs from CF. The effects of these determinants of latency are evident in Fig. 26 for a neuron recorded in the anteroventral cochlear nucleus.

The relative latency of responses in a population of stimulated neurons could well play a part in central processing, such as in the coding of level differences at the two ears, a cue for sound direction. The monotonic relationship between stimulus level and response latency suggests that when the information derived from each ear converges on a central cell, interaural level sensitivity could be achieved through a mechanism of temporal comparison.

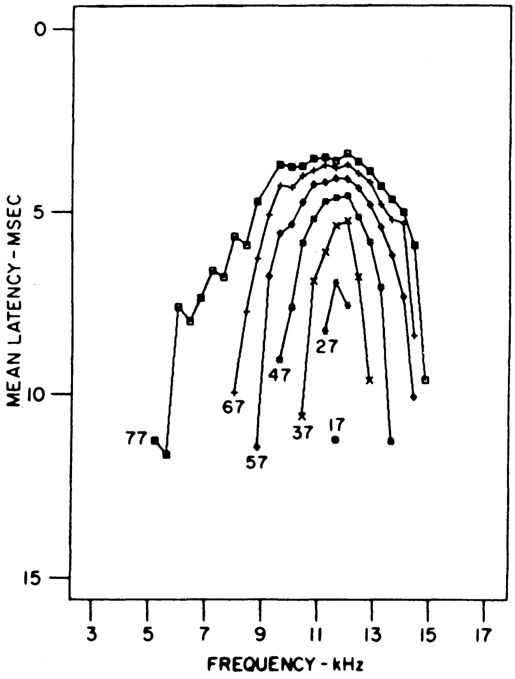

FIG. 26. Mean initial discharge latency as a function of frequency and SPL for a neuron from the anteroventral cochlear nucleus of the cat. As is typical of central auditory neurons, shorter latencies occur at higher SPLs and near CF. Data based on 50 presentations of a 50-ms tone burst at 4/s. Reproduced with permission from Kitzes *et al.* (1978), Fig. 1(b).

10.5 Level Sensitivity

The auditory system is extremely sensitive to sound level. A young and healthy human listener has a dynamic range of about 120 dB. The neural mechanism underlying this ability has long been an issue of contention. A single auditory nerve fiber representative of the majority (about 84%; Liberman, 1978) characterized by high (> 18 spikes/s) or medium (0.5–18 spikes/s) spontaneous rates has a dynamic range of only about 20–40 dB (Sec. 2.2), some 80–100 dB short of the dynamic range of behavioral discrimination. A broader dynamic range could be encoded by a population of fibers, each responding over different ranges of sound pressure level. This could be achieved if CF thresholds were broadly distributed, allowing recruitment of new fibers as a function of increasing stimulus strength. However, CF thresholds among high-spontaneous-rate fibers are uniformly low and occupy a very restricted range: for a given CF region, thresholds are contained within a range of about 10 dB (Liberman, 1978). The CF thresholds of medium-spontaneous-rate fibers are slightly higher, but they too have a quite restricted CF threshold range (about 15 dB for a given CF region), implying that even considered collectively the majority of auditory nerve fibers can encode only a limited dynamic range at relatively low SPLs. The encoding of higher SPLs may well depend on the smaller group of auditory nerve fibers characterized by low spontaneous rates. These fibers have relatively high CF thresholds with a 20–50 dB range for a given CF region and, in addition, have especially wide dynamic ranges. The operating range

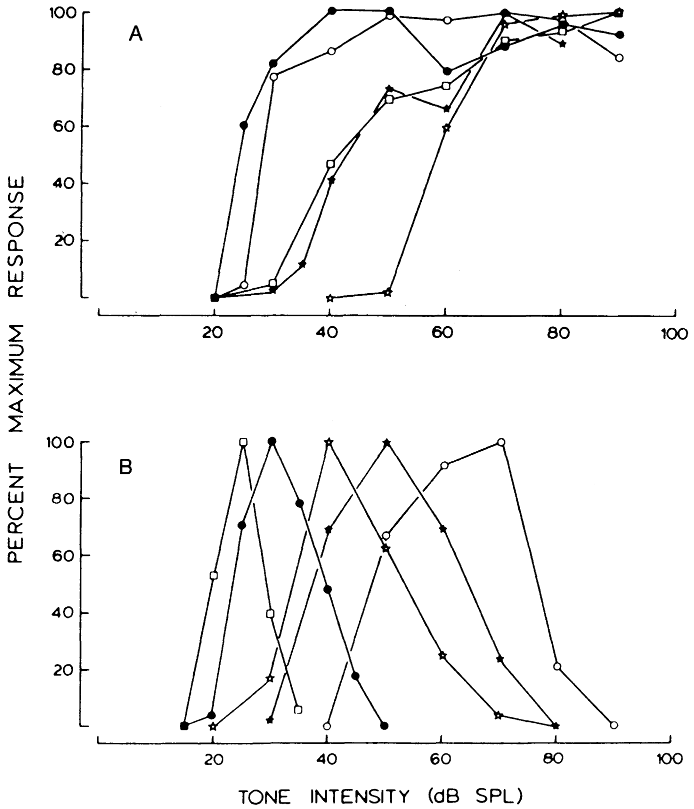

FIG. 27. Representative examples of (a) monotonic and (b) nonmonotonic rate-level functions, derived from primary auditory cortex in the cat. Discharge rates derived from 50 repetitions of CF tones are normalized. Reproduced with permission from Phillips *et al.* (1985b), Fig. 1.

may be extended further by progressive recruitment, at higher stimulus levels, of fibers tuned to lower and higher frequencies. Finally, it should be noted that the onset component of responses to sound can encode a much wider dynamic range than can the steady state component (Smith, 1988).

Perhaps the most striking transformation in level sensitivity within the central auditory system is the generation of nonmonotonic rate-level functions, in which discharge rate increases from threshold to some critical stimulus level, above which the rate declines. Although nonmonotonic rate-level relationships occur occasionally in the auditory nerve, the change in discharge rate is characterized by an abrupt decrease followed by a rapid recovery occurring within a very narrow range of SPLs (Kiang, 1984). By contrast, within the central auditory system the reduction in discharge rate can occur within 10 dB of threshold, is typically more gradual, and is not followed by a recovery of the rate at higher SPLs. In the lower brainstem, nonmonotonic responses are limited to DCN, but are very common at higher levels of the auditory system (e.g., approximately 50% of inferior colliculus neurons respond nonmonotonically; Semple and Kitzes, 1985). The mechanism for this phenomenon seems to be a convergence of excitation and inhibition tuned to the same frequency with the inhibitory threshold being higher. Examples of monotonic (a) and nonmonotonic (b) rate-level functions representative of primary auditory cortex in the cat (Phillips *et al.*, 1985b) are shown in Fig. 27. Note that for the monotonic functions [Fig. 27(a)], discharge rate rises from threshold to saturation within a variable range (10–40 dB). A similarly variable range exists for both the rising and falling slopes of nonmonotonic functions. Because each nonmonotonic function has a single, "best" SPL associated with peak responsiveness, and because best SPLs can vary considerably among a population of neurons, it has been suggested that intensity might be encoded centrally by a population of neurons tuned to a broad range of stimulus levels (Brugge and Merzenich, 1973).

Although the dynamic range of most central auditory neurons does not differ appreciably from that of most auditory nerve fibers, there are exceptions: some neurons of the cochlear nuclear complex, for example, respond over a wide range of levels, occasionally exceeding 80 dB (Rhode *et al.*, 1983). These neurons are characterized by transient discharge patterns and often by broad frequency tuning, suggesting that the broad rate-level functions result from convergence of cochlear fibers of different CFs.

10.6 Binaural Sensitivity and Mechanisms for Sound Localization

10.6.1 Binaural Cues for Sound Localization. Although sound localization is possible with only one ear, binaural cues significantly enhance localization resolution. Because of diffraction and reflection by the head, torso, and ears, sound waves rarely stimulate the left and right cochlea equivalently. Interaural differences in time and amplitude provide information about the location of a sound source. The interaural level difference (ILD) arises from sound scattering by the head and torso, coupled with the directionality of the pinnas (Sec. 3.1 and Fig. 11). Since these interactions are significant only when the physical dimensions of the body parts are large in relation to the incident wavelength, ILD is of biological relevance only at relatively high frequencies. Although under most conditions there is a considerably greater change in ILD with horizontal displacement than with vertical displacement, ILD does not specify a direction uniquely in either plane: a given ILD is produced by several angles of incidence.

The second major binaural cue is the interaural time difference (ITD), which depends on the difference in path length and the consequent difference in arrival time of the incident wave at each ear. For low frequencies, ITD is approximated by the equation

$$\mathrm{ITD}=\left(\frac{3a}{c}\right)\sin\theta_{\mathrm{inc}}, \qquad (1)$$

where a is the radius of a sphere approximating the head, c is the ambient speed of sound, and θ_{inc} is the angle of incidence (Kuhn, 1987). Interaural time differences are greatest for lateral angles of incidence at low frequencies. As a cue for source direction, ITD becomes ambiguous at high frequencies: when the period is less than the maximum possible interaural delay, a given phase disparity can occur for multiple angles of incidence.

Interaural time and amplitude disparities for low-frequency sounds are adequately approximated by modeling the head as a rigid sphere (Kuhn, 1987). At higher frequencies, however, the pinna becomes highly directional and reflections within the pinna and auditory meatus produce spectral maxima and minima (Fig. 11). For sounds incident in the midsagittal plane of the head, there are minimal interaural disparities. If the head, torso, and ears are perfectly symmetrical, disparity is zero. However, vertical displacement does change the spectral interference patterns at the ears and localization is possible in this plane if the sound is spectrally rich. In fact, artificial manipulation of spectral maxima and minima has been shown to produce the perception of shifts in the elevation of the sound source (Watkins, 1978). Thus, spectral cues are considered a third major source of directional information.

10.6.2 Central Processing of Binaural Information. Since each auditory nerve contains information that is exclusively monaural, the nervous system can only utilize binaural disparity cues by central comparison. As discussed earlier (Sec. 7 and Fig. 23), the first site of central comparison is the superior olivary complex. Neurons in the medial superior olive are sensitive to the relative phases of the stimuli at the two ears. The influence of ITD on the responsiveness of a typical MSO neuron is shown in Fig. 28(a). Discharge rate is a cyclic function of interaural delay, with a period equivalent to that of the stimulus wave form. Superimposed on the rate/delay function is a plot of vector strength, a measure (with a range from 0 to 1.0) of the magnitude of phase locking. Optimal delays produce maximum discharge rates and high vector strengths; unfavorable delays produce low discharge rates and low vector strengths. Sensitivity to ITD in the MSO arises from the convergence of phase-locked inputs from the anteroventral cochlear nuclei: ipsilateral inputs terminate on lateral dendrites of MSO neurons, while contralateral inputs terminate on medial dendrites. Optimal binaural stimulation occurs when the phase-locked discharges

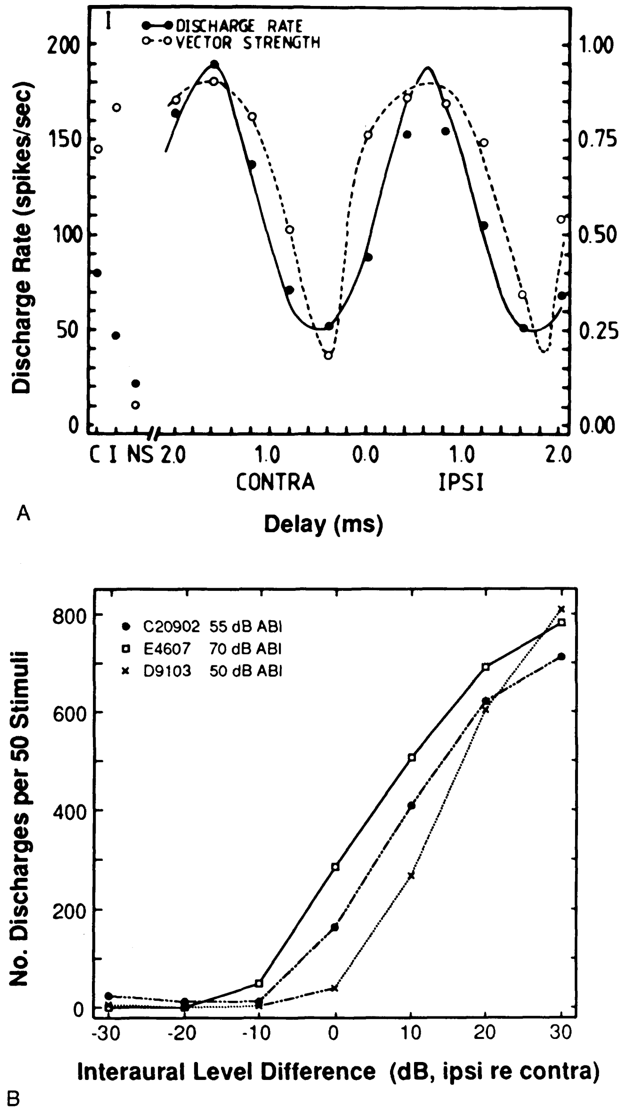

FIG. 28. Neural sensitivity to (a) the interaural time difference and (b) interaural level difference. (a) Discharge rate and phase synchrony (vector strength) are plotted as functions of interaural delay for a medial superior olivary neuron in the dog. Stimulus was a CF tone of 444.5 Hz at 70 dB SPL. Points at left of abscissa are the discharge rates for monaural contralateral (C), ipsilateral (I), and no stimulus (NS). Slightly modified, with permission, from Goldberg and Brown (1969), Fig. 8. (b) Number of discharges in response to 50 presentations of a 100-ms CF tone as a function of ILD for three neurons in the gerbil lateral superior olive. The average binaural level was maintained constant as ILD varied. The abscissa indicates the SPL at the ipsilateral ear minus that at the contralateral ear. The response is maximal when the excitatory (ipsilateral) input is relatively more intense and is reduced progressively as the ILD balance shifts to favor the inhibitory (contralateral) input. (Unpublished data of M.N. Semple.)

arrive simultaneously at the lateral and medial dendrites. Because the axons converging on a single MSO neuron are of unequal length, coincident arrival of the inputs does not necessarily occur at zero ITD: the optimal delay is different for different neurons. In addition, apparently because phase locking is limited to low frequencies (<4 kHz in most species), the CF distribution for ITD-sensitive neurons in MSO is biased towards low frequencies.

One useful system for classifying binaural responses is based on defining the monaural influences as either excitatory or inhibitory. Two major binaural classes are recognized: neurons receiving an excitatory input from one ear and an inhibitory input from the other ear are classified "EI"; bilaterally excitable units are classified "EE." At the superior olivary complex, the EI class is particularly sensitive to ILD and relatively insensitive to the mean binaural level, whereas the reverse is true for EE neurons (Goldberg and Brown, 1969). Sensitivity to the level difference potentially provides information about the angular displacement of the sound source relative to the ears of the observer, whereas sensitivity to mean binaural level provides intensity information that is less dependent on the incident angle. Thus, EI units may participate in the mechanisms for sound localization and EE units may be involved in the processing of intensity information. Within the superior olivary complex, the EE class is predominant for high-CF neurons within medial cell groups, including the medial superior olive (Goldberg and Brown, 1969), whereas EI neurons are dominant within the lateral superior olive (Tsuchitani, 1977). There is also evidence for differential distributions of EE and EI responses at higher levels of the auditory system (Semple and Aitkin, 1979). Figure 28(b) illustrates ILD sensitivity in three LSO neurons. In each instance, discharge rates are maximal when the SPL is greater at the ear providing excitation (i.e., ipsilateral), but the response diminishes progressively as the balance shifts in favor of the contralateral ear, whose influence is inhibitory. Consistent with the observation that ILD is most relevant for high-frequency signals, there is a marked skew toward high frequencies in the CF distribution of ILD-sensitive neurons within LSO and at other levels of the central auditory system.

Although the monaural responses of many neurons can readily be classified as excitatory or inhibitory, the responses of some neurons are more consistent with a monaural mixture of excitation and inhibition. Mixed monaural influences are more common at higher levels of the auditory system and are particularly apparent in responses of ITD-sensitive neurons, so the EE/EI classification is less useful for characterizing low-CF neurons. Mixed binaural interactions have also been noted for some high-CF neurons; thus, using a classification system closely related to the EE/EI scheme, Imig and his colleagues (1986) have identified three distinct types of topographically segregated binaural interaction in cat primary auditory cortex: regions of binaural summation, characterized by a greater response to stimulation of both ears than to stimulation of just one; regions of suppression, characterized by a reduction of responsiveness under binaural conditions; and regions of mixed interaction. An example of a suppression band in relation to the tonotopic organization of AI is shown in Fig. 29. Isofrequency lines from 4.9 to 22.7 kHz run vertically, with lower frequencies represented caudally (to the right in this figure). The elongated suppression band (solid outline) is oriented orthogonal to the isofrequency contours. Surrounding the suppression band is a belt in which interactions are summative.

The functional significance of cortical binaural bands remains unclear. They may well play a role in the central processing of sound source location and, accordingly, a given suppression band would presumably contain neurons sensitive to displacement in azimuth. However, these neurons would not respond exclusively to a specific spatial location and, hence, the band would not constitute a mapping of the acoustic field onto neural tissue. In fact, studies adopting the direct approach of presenting sounds from sources in the free field have failed to find neural maps of acoustic space in the central auditory system of the cat or other mammals. There is, however, convincing evidence for such a map in the barn owl (Konishi, 1986). The lower brainstem of this species contains two anatomically distinct pathways that respectively encode ITD and ILD. The convergence of these pathways at the auditory midbrain provides the basis for a representation of acoustic space. Individual neurons are selectively responsive to sounds deriving from a particular direction and are organized topographically; thus, sound emanating from a particular direction results in neural activation of a very limited region of the neural field.

10.7 Complex Sounds

10.7.1 Background Noise. The presentation of a tone against a background of broadband noise has different consequences for neural responses at peripheral and central levels of the auditory system. At both sites, background noise has the common effect of elevating tonal thresholds, but the threshold shift occurs equivalently across the frequency response area of auditory nerve fibers (Geisler and Sinex, 1980), whereas in cortical responses the shift is usually most pronounced at CF (Phillips, 1989). In most auditory nerve fibers, the threshold shift is accompanied by compression of the dynamic range for tonal stimuli: background noise elevates the baseline discharge rates of auditory nerve fibers, thus reducing the range of discharge rates available for encoding the intensity of the tone. In contrast, dynamic range compression rarely accompanies the noise-induced

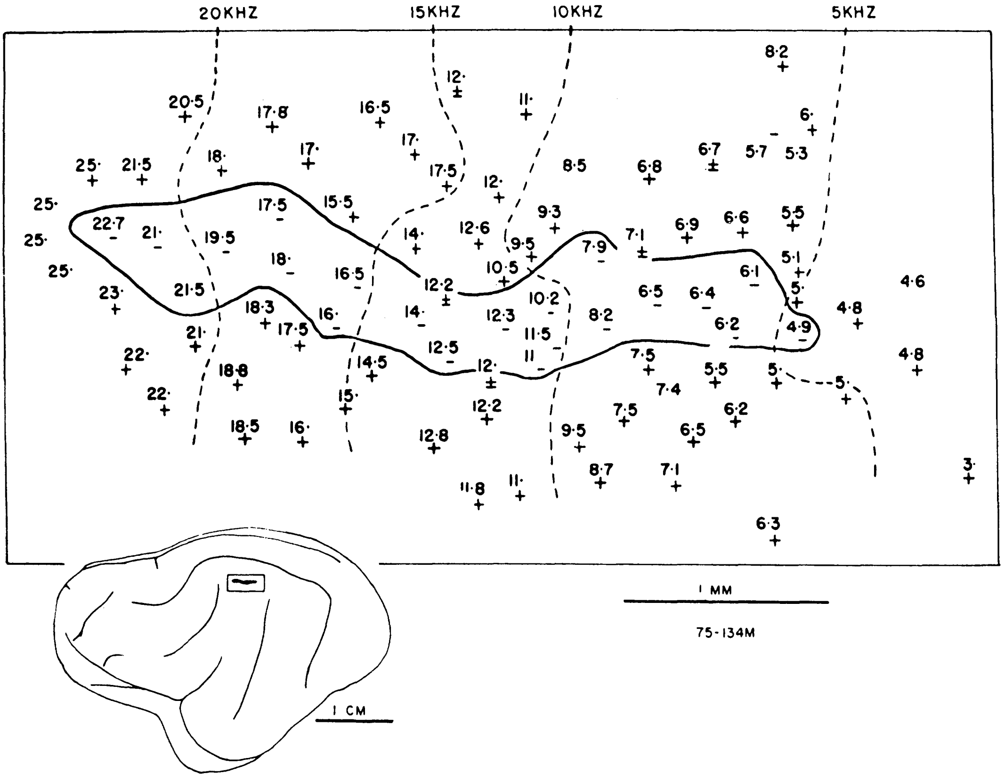
20KHZ
15KHZ
10KHZ
5KHZ
8·2 +
12· ±
20·5 +
17·8 +
16·5 +
11· +
6· +
25· +
21·5 +
18· +
17· +
17· +
17·5 +
6·7 ±
5·7
5·3
8·5
6·8 +
12· +
25·
17·5 -
15·5 +
9·3 +
6·9 +
6·6 +
5·5 +
22·7 -
21· -
19·5 -
14· +
12·6 +
9·5 +
7·9 -
7·1 ±
25·
18· -
16·5 -
10·5 +
5·1 +
25·
23· +
21·5
12·2 ±
10·2 -
6·1 -
5· +
4·6
18·3 +
16· -
14· -
12·3 -
8·2 -
6·5 -
6·4 -
4·8 +
21· +
17·5 +
11·5
6·2 -
4·9 -
22· +
18·8 +
14·5 +
12·5 -
12· ±
11 -
7·5 +
5·5 +
5· +
5· +
4·8 +
22· +
15· +
12·2 +
7·4
7·5 +
6·2 +
18·5 +
16· +
12·8 +
9·5 +
6·5 +
8·7 +
7·1 +
11·8 +
11· +
3· +
6·3 +
1 MM
75-134M
1 CM

FIG. 29. An example of a suppressive binaural band running orthogonal to the frequency organization of primary auditory cortex in the cat. A mapping electrode was inserted into the cortex many times, at different positions across the frequency organization of AI. At each recording site, the CF and the binaural interaction were determined. The low-to-high frequency progression in AI is from caudal to rostral (i.e., from right to left in this figure). Dashed lines indicate selected isofrequency contours; + and − indicate summative and suppressive interactions, respectively. Slightly modified, with permission, from Imig and Adrián (1977), Fig. 7.

threshold shift in auditory cortical neurons: cortical neurons do not respond tonically to continuous noise stimuli, so their full range of discharge rates remains available for encoding the tonal stimulus. In effect, the cortical neuron's discharge rate encodes the signal level not relative to an absolute sound pressure, but referenced to background (Phillips, 1989). In the minority of cortical neurons whose dynamic range for tonal stimuli is reduced by background noise, the compression is caused by an increase in the slope of the rate-level relationship. Hence, in contrast to auditory nerve fibers, these neurons have a greater range of discharge rates available for encoding the reduced dynamic range.

10.7.2 Amplitude and Frequency Modulation. There are clear differences in peripheral and central responses to amplitude modulation (AM) of the acoustic signal. Discharges in the auditory nerve are well synchronized to the wave-form periodicity of an AM signal over a broad frequency range (Smith and Brachman, 1980), but at higher levels of the central auditory system both the resolution and the upper limit of entrainment are significantly reduced (Creutzfeldt *et al.*, 1980). A major determinant of these differences is that each additional synapse in the ascending auditory pathway progressively adds temporal jitter, reducing the accuracy of entrainment to rapid temporal changes.

Neural responses to AM are customarily described using the so-called "temporal modulation transfer function," a measure that may be expressed in terms of rate or synchronization of the neural discharge as a function of modulation frequency. The modulation transfer functions of auditory nerve fibers are characteristically low-pass, regardless of

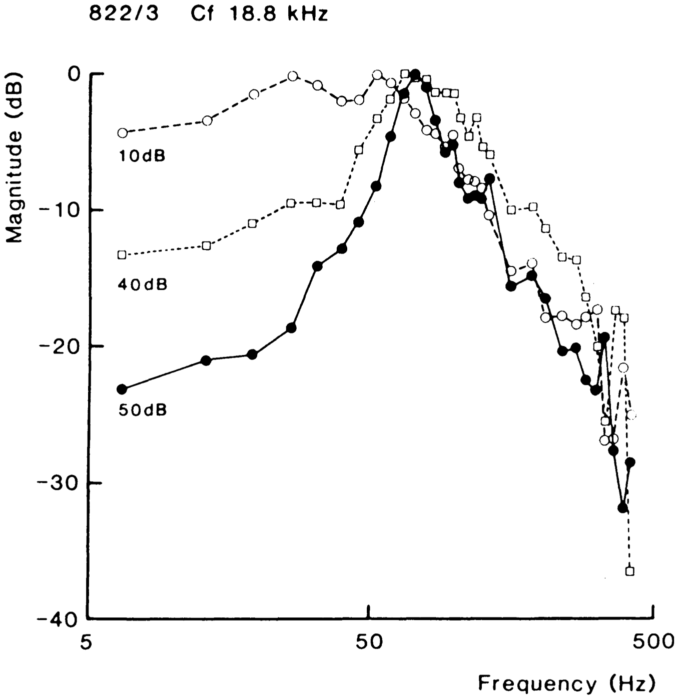

FIG. 30. Amplitude modulation transfer functions from a single neuron in the rat inferior colliculus, showing the effects of stimulus level. At three sound pressure levels, a CF tone carrier was amplitude modulated (depth 15% rms) with pseudorandom noise and presented for 5–10 min. Neural discharges were collected as 2048-bin histograms synchronized to the periodicity of the pseudorandom noise. Each histogram was then cross-correlated with the pseudorandom noise to yield a modulation transfer function. Reproduced with permission from Rees and Møller (1987), Fig. 2.

stimulus level, whereas those of at least some central neurons are quite level-dependent. Figure 30 provides an illustration of such level dependence in modulation transfer functions derived from a single neuron in the inferior colliculus of the rat (Rees and Møller, 1987); the response functions change from essentially low pass at 10 dB (i.e., comparable to the auditory nerve) to clearly bandpass at 50 dB. Modulation transfer functions of such neurons not only become more bandpass with increasing mean level, but their phase-locking peak is elevated to higher modulation frequencies. Hence, the concept of a "best modulation frequency," as proposed by some, is of limited utility.

Frequency modulation (FM), like AM, is an important component of speech and other complex sounds and, like AM, its representation is significantly transformed within the central auditory system. Within the auditory nerve, there is no sensitivity to the direction of a given FM sweep, nor is there any sensitivity to FM outside the steady-state (pure-tone) response area (Sinex and Geisler, 1981). Although the FM sensitivity of some central auditory neurons is similarly predictable on the basis of their responses to pure tones, for many neurons the influence of FM is more complex. Within the cochlear nuclear complex, the direction of modulation does not influence the responses of primarylike neurons, but many pauser, buildup, or onset neurons do respond asymmetrically to ascending and descending FM sweeps. The FM-induced responses of auditory cortical cells are frequently even more complex, with significant responsiveness occurring in frequency ranges outside the pure tone response area and many responses containing sharp sensitivity to the direction of modulation [Phillips *et al.*, 1985(a)].

10.7.3 Vocalizations. It may be argued that the most important acoustic stimulus is speech or, more generally, a species-specific vocalization. Conceivably, recognition of such vocalizations could be mediated by individual "feature-detector" neurons, each responding specifically to a particular vocalization feature. Of course, given a finite number of available neurons, the brain would be limited in the diversity of vocalizations that could be encoded by such feature detectors. An alternative, perhaps more likely, mechanism for the recognition of vocalizations may involve the integrated activities of many neurons, each responding somewhat broadly to components of the vocalization.

There appears to be a broad range of specificity in the responses of cortical neurons to vocalizations. Certain neurons are very sensitive to a small component of the vocalization or to the context in which the component occurs. For example, a neuron responding to a specific vocalization may not respond at all when the vocalization is played backwards (Glass and Wollberg, 1983), even though the spectral content of the stimulus is maintained. For other neurons, the temporal pattern of their discharge may be reversed or otherwise modulated.

11. CENTRIFUGAL INFLUENCES

The ascending components of the central auditory system, conveying acoustic information from the periphery to higher auditory processing centers, are complemented by a sizable system of descending projections whose role may be to regulate the incoming flow of information. Centrifugal control mechanisms may originate in auditory cortex and ultimately influence cochlear hair cells, via synapses at the inferior colliculus and periolivary regions. In addition, certain auditory reflexes (such as startling, pinna movements, and orienting responses) mediated via the brainstem may also be influenced by higher auditory centers (Aitkin, 1986).

11.1 Olivocochlear Efferent System

This projection arises bilaterally in the superior olivary complex and projects via the auditory nerve to terminate on cochlear outer hair cells and on the Type I afferents that innervate inner hair cells (Fig. 1 and Sec. 4.1.4). There are two separate (medial and lateral) subsystems of olivocochlear neurons (Warr *et al.*, 1986). The medial system arises near the medial superior olive from large neurons projecting via large axons that terminate on outer hair cells predominantly in the contralateral cochlea. Some of the medial olivocochlear axons branch en route to innervate the ventral cochlear nucleus, sometimes ramifying extensively within granule cell regions encapsulating the ventral cochlear nucleus. The lateral system arises in or near the lateral superior olive, from smaller cells whose fine-caliber axons terminate on Type I auditory nerve afferents, almost exclusively in the ipsilateral cochlea at the base of inner hair cells. Lateral olivocochlear fibers do not appear to provide collateral branches to the cochlear nuclear complex.

The functional significance of the two olivocochlear systems remains unclear. The efferent fibers terminate tonotopically and their responses to acoustic stimuli can be as sharply frequency-tuned as those of the cochlear afferents (Robertson and Gummer, 1985). The responsiveness of the efferent fibers is significantly depressed by anesthesia and, hence, the functions of the descending feedback systems in the awake and behaving animal might be difficult to infer from physiological studies that are usually conducted in anesthetized animals. The response-latency distribution of cochlear efferents is very broad, suggesting that efferent control of the cochlea may involve pathways arising at levels of the brain far higher than the superior olivary complex, perhaps at the cortex (Robertson and Gummer, 1985).

11.2 Other Descending Pathways

Each auditory cortical field sends descending projections to its source of input in the thalamus and to the auditory midbrain regions surrounding the central core nucleus of the inferior colliculus (Pickles, 1988). There are two major centrifugal projections arising in the inferior colliculus. One tract descends in the lateral part of the lateral lemniscus to innervate mostly nonauditory regions of the lower brainstem, whereas the second tract descends medially, terminating bilaterally in the dorsal cochlear nucleus and ipsilaterally in the periolivary area, apparently making contact with cells that give rise to the olivocochlear projection. The cochlear nuclear complex receives extensive descending input from several brainstem sources, including the contralateral cochlear nucleus, the periolivary nuclei bilaterally, the nuclei of the lateral lemniscus, and the inferior colliculus.

ACKNOWLEDGMENTS

We were supported by NIH Grants DC-00110 and DC-00419 (M.A.R.) and DC-00364 and DC-00450 (M.N.S.). Our thanks to B.

Shivapuja, W. D. Ward, and especially L. Kitzes and N. C. Rich, for their many helpful comments on previous versions of this manuscript. N. C. Rich, in addition, prepared many of the figures.

GLOSSARY

Acoustic Striae: Three fiber bundles projecting from the cochlear nuclear complex to higher levels of the central auditory system.

Action Potential: A brief all-or-none "depolarization" of the plasma membrane of a neuron that travels at speeds of tens of m/s along the neuron's axon; action potentials constitute the fundamental means of rapid signal transmission over long distances in the nervous systems of animals.

Afferent Neurons: Neurons that conduct "action potentials" from relatively peripheral locations, such as a sensory organ, toward more central locations.

Afferents: See "afferent neurons."

AI: The primary region of the "auditory cortex."

AP: See "compound action potential."

Apical: In cochlear anatomy, referring to the region of the cochlear "cone" near its apex.

Audiogram: The curve describing detection thresholds for tonal stimuli as a function of frequency.

Auditory Cortex: The auditory area of the cerebral cortex; it includes a primary region (AI) and several other regions.

Auditory Midbrain: See "inferior colliculus."

Auditory Nerve: See "cochlear nerve."

Auditory Reflex: An involuntary, prompt, and stereotyped muscular response to sound.

AVCN: Anteroventral cochlear nucleus, one of the three divisions of the "cochlear nuclear complex."

Axon: The typically long and thin process, emanating from a neuron's cell body, that conducts "action potentials" toward terminals that make synaptic contact with other neurons.

Basal: In cochlear anatomy, referring to the region of the cochlear "cone" near the oval and round windows.

Basilar Membrane: The relatively long and narrow membrane that lies between the "spiral lamina" and the "spiral ligament" of the cochlea and extends longitudinally from the cochlear base to its apex; upon it rests the organ of Corti, where the sensory "hair cells" reside.

Binaural Classification: Neural categorization based on the balance of excitation and inhibition received from the two ears; bilaterally excitable neurons are classified "EE" and those receiving an excitatory input from one ear and an inhibitory input from the other ear are classified "EI."

Caudal: Located "near the tail"; in brain anatomy, referring to locations towards the spinal cord; see "rostral."

CF: See "characteristic frequency."

Characteristic Frequency: The stimulus frequency to which an auditory structure ("basilar membrane," "hair cell," or "neuron") is most sensitive.

Chemical Synaptic Transmission: The process whereby neurons or sensory cells, such as "hair cells," excite or inhibit the generation of "action potentials" by other neurons by releasing a specific chemical substance that diffuses across the narrow gap between the presynaptic and postsynaptic membranes.

Cochlear Nerve: The bundle of axons of "afferent" and "efferent neurons" that connects the cochlea to the brain.

Cochlear Nuclear Complex: The first auditory center in the brain, receiving input from the cochlea.

Compound Action Potential: The synchronized response of many cochlear "afferent neurons"; it can be recorded by an electrode in the vicinity of the cochlea.

Condensation Click: An acoustic pulse that causes a brief increase in air density near the "tympanic membrane."

Contralateral: Located in the opposite side of the body.

Cytoplasm: The cell substance outside its nucleus, bounded externally by the "plasma membrane."

dB: See "decibel."

DCN: Dorsal cochlear nucleus, one of the three divisions of the "cochlear nuclear complex."

Decibel: A dimensionless unit expressing the ratio of two values of power or pressure. One decibel equals 10 times the logarithm to the base 10 of the power ratio or, equivalently, 20 times the logarithm to the base 10 of the pressure ratio. A commonly used pressure reference in physiological and psychological acoustics is 0.0002 dyn/cm^2, which corresponds approximately to the lowest threshold of hearing in normal humans. When this reference is used pressures are referred to as "decibels sound pressure level," abbreviated as "dB SPL."

Dendrite: A relatively short projection from a neuron's cell body that receives synaptic inputs from other neurons.

Depolarization: Decrease in the negative intracellular electrical potential (e.g., from −60 mV to −40 mV) of neurons or sensory cells.

Discharge: Synonymous with "action potential."

Dorsal: Located in the back of a structure.

Eardrum: See "tympanic membrane."

EE and EI: See "binaural classification."

Efferent Neurons: Neurons that conduct action potentials from relatively central locations to more peripheral sites.

Efferents: See "efferent neurons."

Endocochlear Potential: The +80 mV electrical potential of the "endolymph" in the "scala media" of the cochlea.

Endolymph: The potassium-ion-rich fluid that occupies the "scala media" of the cochlea.

Excitation: In neural systems, a process that increases the probability of action potential generation.

Fiber: Synonymous with "axon."

Frontal Plane: Any plane orthogonal to the sagittal plane and parallel with the body's long axis; see "sagittal plane."

Hair Cells: Mechanosensitive sensory cells that transduce motion of their "stereocilia" ("hairs") into intracellular electrical potential changes.

Helicotrema: The site where "scala vestibuli" and "scala tympani" meet at the apical end of the cochlea.

Hyperpolarization: Increase in the negative intracellular electrical potential of neurons or sensory cells.

ICC: Central nucleus of the "inferior colliculus."

IHC: See "inner hair cells."

Inferior Colliculus: The auditory structure on the dorsal aspect of the midbrain, whose major component is the central nucleus (the primary midbrain relay in the ascending auditory system).

ILD: See "interaural level difference."

Inhibition: In neural systems, a synaptic process that decreases the probability of action potential generation.

Inner Hair Cells: The mammalian cochlear "hair cells" located in a single row on the modiolar side of the pillar cells of the organ of Corti; they are innervated by the "Type I cochlear afferent neurons," which collectively make up the main signal outflow of the cochlea.

In Vitro: In an artificial environment.

In Vivo: In the living body.

Interaural Level Difference: The difference in sound pressure level at the two ears, due to diffraction and reflection by the head, pinna, and torso; it provides a cue for sound localization, particularly at high frequencies.

ITD: See "interaural phase difference."

Interaural Phase Difference: The difference in phase of the acoustic stimulus wave form at the two ears, generated by differences in arrival time; it provides a cue for sound localization, particularly at low frequencies.

Lateral Lemniscus: A large tract of fibers connecting auditory centers in the lower brainstem to the auditory midbrain; it contains the nuclei of the lateral lemniscus.

LSO: Lateral superior olivary nucleus, one of the divisions of the "superior olivary complex."

Medial Geniculate Body: Part of a brain region known as the thalamus, it relays information between the auditory midbrain (inferior colliculus) and the auditory cortex.

Modiolus: The hollow, bony central axis of the cochlea.

MSO: Medial superior olivary nucleus, one of the divisions of the "superior olivary complex."

Neuron: The fundamental cellular constituent of the nervous system, made up of a cell body from which projections ("dendrites" and "axons") extend; it is specialized to integrate synaptic inputs at the cell body and dendrites, and to transmit "action potentials" to "axon" terminals that synapse upon other neurons or effector cells, such as muscles fibers.

Octave: The frequency interval between a frequency and twice its value.

OHC: See "outer hair cells."

Olivocochlear Efferent Neurons: Neurons that originate in the superior olivary region of the brainstem and terminate in the cochlea upon "outer hair cells" and "Type I afferent neurons."

Ossicles: See "ossicular chain."

Ossicular Chain: A series of three tiny bones ("ossicles") in the mammalian middle ear that transmits vibrations of the "tympanic membrane" to the "oval window" of the cochlea.

Otoacoustic Emissions: Sounds emitted by the ears of animals.

Outer Hair Cells: The mammalian cochlear "hair cells" located peripheral to the pillar cells in the organ of Corti; they are innervated by "Type II cochlear afferent neurons."

Oval Window: A membrane at the basal end of the "scala vestibuli" of the cochlea, against which the "stapes" footplate abuts.

Parcellation: Division or distribution into separate parts or regions.

Perilymph: The fluid that fills the "scala vestibuli" and the "scala tympani" of the cochlea; it is similar to other extracellular fluids, having high sodium-ion concentration and low potassium-ion concentration.

Periolivary: Surrounding the "superior olivary complex."

Pharynx: The sac connecting the back of the mouth and nose with the esophagus.

Phase Locking: The property whereby cochlear "afferents" and certain other auditory neurons respond preferentially to specific phases of low-frequency (<4 kHz) periodic stimuli.

Plasma Membrane: The outermost structure of living cells, consisting of a bimolecular lipid layer plus various other molecules; in neurons, the latter include proteins that act as receptors for chemical transmitters and as ion channels.

Poststimulus Time Histogram: A histogram depicting the time of occurrence of "action potentials" relative to stimulus onset.

PSTH: See "poststimulus time histogram."

PVCN: Posteroventral cochlear nucleus, one of the three divisions of the "cochlear nuclear complex."

Q_{10}: A measure of the sharpness of tuning of a frequency filter, defined as its center frequency divided by the 10-dB bandwidth.

Rarefaction Click: An acoustic pulse that causes a brief decrease in air density near the "tympanic membrane."

Rate-Level Function: The function relating the evoked neural discharge rate to the stimulus sound pressure level.

Receptor Potential: The graded changes in electrical potential evoked in sensory receptor cells by effective stimuli.

Reissner's Membrane: The cellular membrane that separates the "scala vestibuli" of the cochlea from the "scala media."

Rostral: Located "near the face"; in brain anatomy, referring to locations far from the spinal cord; see "caudal."

Round Window: A membrane at the basal end of the "scala tympani" of the cochlea, in contact with the middle ear cavity.

Sagittal: Parallel to the median plane of the body; see "frontal."

Scala Media: The "endolymph"-filled, inner membranous tube of the cochlea; it lies between "scala vestibuli" and "scala tympani."

Scala Tympani: The "perilymph"-filled cochlear tube which is separated from "scala media" by the "basilar membrane" and the organ of Corti, the "spiral lamina" and the "spiral ligament."

Scala Vestibuli: The "perilymph"-filled cochlear tube which is separated from "scala media" principally by "Reissner's membrane."

SOC: See "superior olivary complex."

Sound Localization: In a strict sense, referring to the detection of both the angle of incidence and distance of a sound source; in a more general sense, referring to the detection of angle of incidence only.

Spike: See "action potential."

Spiral Ganglion: Spirally shaped group of cell bodies of the cochlear "afferent neurons" located within the "modiolus."

Spiral Lamina: a bony shelf protruding radially from the cochlear "modiolus," whose edge provides the central attachment of the "basilar membrane" and through which the peripheral terminals of cochlear neurons travel on their way to the "hair cells."

Spiral Ligament: The cochlear structure whose edge provides the peripheral attachment of the "basilar membrane."

SPL: Sound pressure level; see "decibel."

Spontaneous Activity: Neural discharges occurring in absence of stimulation.

Stapes: The innermost middle ear "ossicle"; its footplate abuts against the "oval window" of the cochlea.

Stereocilia: The "hairs" of the "hair cells"; they consist of thin cylinders with a tightly packed core of protein fibrils, enveloped by "plasma membrane."

Sulcus (pl. sulci): Infolding of the cortical surface.

Superior Olivary Complex: a group of auditory nuclei located near the midline of the lower brainstem.

Synapse: The specialized cellular structure where communication between neurons takes place; see "chemical synaptic transmission."

Tectorial Membrane: A noncellular gellike flap which overlies the organ of Corti and makes contact with the tallest "stereocilia" of the "outer hair cells."

Tonotopicity: The property whereby stimulus frequency is mapped systematically onto space in the cochlea, the "cochlear nerve," and the auditory regions of the brain.

Transduction: In sensory systems, the conversion of a signal from one form of energy into another.

Tympanic Membrane: The membrane at the central end of the external ear canal; it is set in motion by sound waves.

Type I Cochlear Neurons: The cochlear "afferent neurons" that transmit signals from single "inner hair cells" to the cochlear nucleus; the "axons" of these neurons constitute the vast majority of those in the "cochlear nerve."

Type II Cochlear Neurons: The cochlear "afferent neurons" that innervate "outer hair cells" and project to the cochlear nucleus; these neurons constitute 5%–10% of cochlear "afferents."

Ventral: Located near the belly surface of a structure.

Volume Velocity: The rate of flow of the medium perpendicularly through a specified area.

List of Works Cited

Aitkin, L. M. (1973), *J. Neurophysiol.* **36**, 275–283.

Aitkin, L. M. (1986), *The Auditory Midbrain: Structure and Function in the Central Auditory Pathway*, Clifton, NJ: Humana.

Arthur, R. M., Pfeiffer, R. R., Suga, N. (1971), *J. Physiol.* **212**, 593–609.

Ashmore, J. F., Brownell, W. E. (1986), *J. Physiol.* (London) **377**, 41P.

Békésy, G. von (1960), *Experiments in Hearing*, New York: McGraw–Hill.

Bismarck, G. von (1967), *The Sound Pressure Transformation Function From Free-field to the Eardrum of Chinchilla*, M. S. Thesis, Massachusetts Institute of Technology.

Bloom, W., Fawcett, D. W. (1975), *A Textbook of Histology*, Philadelphia: Saunders.

Bohne, B. A., Carr, C. D. (1979), *J. Acoust. Soc. Am.* **66**, 411–414.

Bohne, B. A., Kenworthy, A., Carr, C. D. (1982), *J. Acoust. Soc. Am.* **72**, 102–107.

Brawer, J. R., Morest, D. K. (1975), *J. Comp. Neurol.* **160**, 491–506.

Brödel, M. (1946), Three unpublished drawings of the anatomy of the human ear, Philadelphia: Saunders.

Brownell, W. E., Bader, C. R., Bertrand, D., de Ribaupierre, Y. (1985), *Science* **227**, 194–196.

Brugge, J. F., Anderson, D. J., Aitkin, L. M. (1970), *J. Neurophysiol.* **33**, 441–458.

Brugge, J. F., Merzenich, M. M. (1973), *J. Neurophysiol.* **36**, 1138–1158.

Calford, M. B., Aitkin, L. M. (1983), *J. Neurosci.* **3**, 2365–2380.

Creutzfeldt, O. D., Hellweg, F.-C., Schreiner, C. E. (1980), *Exp. Brain Res.* **39**, 87–104.

Dallos, P. (1988), in: G. M. Edelman, W. E. Gall, and W. M. Cowan (Ed.), *Auditory Function: Neurobiological Bases of Hearing*, New York: Wiley, pp. 153–188.

Dallos, P., Harris, D. (1978), *J. Neurophysiol.* **41**, 365–382.

Davis, H. (1965), *Cold Spring Harbor-Symp. Quant. Biol.* **30**, 181–190.

Décory, L., Dancer, A., Franke, R. (1991), *Intracochlear Pressure Measurements in the Cat, the Chinchilla and the Guinea Pig.* (to be published).

Elliott, D. N., Trahiotis, C. (1972), *Psychol. Bull.* **77**, 198–222.

Evans, E. F., Wilson, J. P. (1975), *Science* **190**, 1218–1221.

Galambos, R., Davis, H. (1944), *J. Neurophysiol.* **7**, 287–303.

Geisler, C. D., Sinex, D. G. (1980), *Hear. Res.* **3**, 317–334.

Glass, I., Wollberg, Z. (1983), *Hear. Res.* **9**, 27–33.

Glendenning, K. K., Masterton, R. B. (1983), *J. Neurosci.* **3**, 1521–1537.

Goldberg, J. M., Brown, P. B. (1969), *J. Neurophysiol.* **32**, 613–636.

Goldstein, J. L. (1967), *J. Acoust. Soc. Am.* **41**, 676–689.

Harris, D. M. (1979), *Hear. Res.* **1**, 133–154.

Harrison, R. V. (1988), *The Biology of Hearing and Deafness*, Springfield, IL: Charles C. Thomas.

Imig, T. J., Adrián, H. O. (1977), *Brain Res.* **138**, 241–257.

Imig, T. J., Reale, R. A. (1980), *J. Comp. Neurol.* **192**, 293–332.

Imig, T. J., Reale, R. A., Brugge, J. F., Morel, A., Adrián, H. O. (1986), in: F. Lepore, M. Ptito, and H. H. Jasper (Eds.), *Two Hemispheres—One Brain: Functions of the Corpus Callosum*, New York: Liss, pp. 103–115.

Jenkins, W. M., Masterton, W. B. (1982), *J. Neurophysiol.* **47**, 987–1016.

Johnstone, B. M., Patuzzi, R., and Yates, G. K. (1986), *Hear. Res.* **22**, 147–153.

Kemp, D. T. (1978), *J. Acoust. Soc. Am.* **64**, 1386–1391.

Khanna, S. M., Leonard, D. G. B. (1982), *Science* **215**, 305–306.

Kiang, N. Y. S. (1984), in: I. Darian-Smith (Ed.), *Handbook of Physiology. Section 1 (The Nervous System), Volume III (Sensory Processes, Part 2)*, Bethesda, MD: American Physiological Society, pp. 639–674.

Kiang, N. Y. S., Watanabe, T., Thomas, C., Clark, L. F. (1965), *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve*, Cambridge, MA: MIT.

Kitzes, L. M., Gibson, M. M., Rose, J. E., Hind, J. E. (1978), *J. Neurophysiol.* **41**, 1165–1182.

Konishi, M. (1986), *Trends in Neurosci.* **9**, 163–168.

Kuhn G. F. (1987), in: W. A. Yost and G. Gourevitch (Eds.), *Directional Hearing*, New York: Springer, pp. 3–25.

Liberman, M. C. (1978), *J. Acoust. Soc. Am.* **63**, 442–455.

Lim, D. J. (1980), *J. Acoust. Soc. Am.* **67**, 1686–1695.

Miller, J. D. (1970), *J. Acoust. Soc. Am.* **48**, 513–523.

Møller, A. R. (1983), *Auditory Physiology*, New York: Academic.

Morest, D. K., Oliver, D. L. (1984), *J. Comp. Neurol.* **222**, 209–236.

Mountain, D. C. (1980), *Science* **210**, 71–72.

Nuttall, A. L. (1986), in: R. A. Altschuler, R. P. Bobbin, and D. W. Hoffman (Eds.), *Neurobiology of Hearing: The Cochlea*, New York: Raven, pp. 47–75.

Osen K. K. (1970), *Arch. Ital. Biol.* **108**, 21–51.

Palmer, A. R., Russell, I. J. (1986), *Hear. Res.* **24**, 1–15.

Phillips, D. P. (1989), in: J. S. Lund (Ed.), *Sensory Processing in the Mammalian Brain*, Oxford: Oxford University, pp. 172–203.

Phillips, D. P., Mendleson, J. R., Cynader, M. S., Douglas, R. M. (1985a), *Exp. Brain Res.* **58**, 443–454.

Phillips, D. P., Orman, S. S., Musicant, A. D., Wilson, G. F. (1985b), *Hear. Res.* **18**, 73–86.

Pickles, J. O. (1988), *An Introduction to the Physiology of Hearing*, 2nd ed., London: Academic.

Ramón y Cajal, S. (1909), *Histologie du Système Nerveux de l'Homme et des Vertébrés*, Madrid: Instituto Ramón y Cajal.

Rees, A., Møller, A. R. (1987), *Hear. Res.* **27**, 129–143.

Rhode, W. S. (1971), *J. Acoust. Soc. Am.* **49**, 1218–1231.

Rhode, W. S., Smith, P. H. (1986), *J. Neurophysiol.* **56**, 261–286.

Rhode, W. S., Oertel, D., Smith, P. H. (1983), *J. Comp. Neurol.* **213**, 448–463.

Roberts, W. M., Howard, J., Hudspeth, A. J. (1988), *Ann. Rev. Cell Biol.* **4**, 63–92.

Robertson, D., Gummer, M. (1985), *Hear. Res.* **20**, 63–77.

Robles, L., Ruggero, M. A., Rich, N. C. (1986), *J. Acoust. Soc. Am.* **80**, 1364–1374.

Robles, L., Ruggero, M. A., Rich, N. C. (1989), in: J. P. Wilson and D. T. Kemp (Eds.), *Cochlear Mechanisms—Structure, Function and Models*, London: Plenum, pp. 369–375.

Rose, J. E., Hind, J. E., Anderson, D. J., Brugge, J. F. (1971), *J. Neurophysiol.* **34**, 685–699.

Ruggero, M. A., Kramek, B., Rich, N. C. (1984), *Hear. Res.* **13**, 293–296.

Ruggero, M. A., Rich, N. C., Robles, L., Shivapuja, B. (1990), *J. Acoust. Soc. Am.* **87**, 1612–1629 .

Ruggero, M. A., Santi, P. A., Rich, N. C. (1982), *Hear. Res.* **8**, 339–356.

Russell, I. J., Sellick, P. M. (1977), in: E. F. Evans and J. P. Wilson (Eds.), *Psychophysics and Physiology of Hearing*, London: Academic, pp. 71–84.

Russell, I. J., Richardson, G. P., Cody, A. R. (1986), Nature **321**, 517–519.

Sachs, M. B., Kiang, N. Y. S. (1968), *J. Acoust. Soc. Am.* **43**, 1120–1128.

Sellick, P. M., Patuzzi, R., Johnstone, B. M. (1982), *J. Acoust. Soc. Am.* **72**, 131–141.

Semple, M. N., Aitkin, L. M. (1979), *J. Neurophysiol.* **42**, 1626–1639.

Semple, M. N., Kitzes, L. M. (1985), *J. Neurophysiol.* **53**, 1467–1482.

Siegel, J. H., Kim, D. O., Molnar, C. E. (1982), *J. Neurophysiol.* **47**, 303–328.

Sinex, D. G., Geisler, C. D. (1981), *Hear. Res.* **4**, 127–148.

Smith R., Brachman, M. (1980), *Hear. Res.* **2**, 123–133.

Smith, R. L. (1988), in: G. M. Edelman, W. E. Gall, and W. M. Cowan (Eds.), *Auditory Function: Neurobiological Bases of Hearing*, New York: Wiley, pp. 243–274.

Smoorenburg, G. F. (1972), *J. Acoust. Soc. Am.* **52**, 615–632.

Spoendlin, H. (1988), in: A. F. Jahn and J. Santos-Sacchi (Eds.), *Physiology of the Ear*, New York: Raven, pp. 201–219.

Tsuchitani, C. (1977), *J. Neurophysiol.* **40**, 296–318.

Vrettakos, P. A., Dear, S. P., Saunders, J. C. (1988), *Am. J. Otolaryngol.* **9**, 58–67.

Warr, W. B. (1982), in: W. D. Neff (Ed.), *Contributions to Sensory Physiology*, New York: Academic, pp. 1–38.

Warr, W. B., Guinan, J. J., White, J. S. (1986), in: R. A. Altschuler, R. P. Bobbin, and D. W. Hoffman (Eds.), *Neurobiology of Hearing: The Cochlea*, New York: Raven, pp. 333–348.

Watkins, A. J. (1978), *J. Acoust. Soc. Am.* **63**, 1152–1165.

Wiederhold, M. L. (1986), in: R. A. Altschuler, R. P. Bobbin, and D. W. Hoffman (Eds.), *Neurobiology of Hearing: The Cochlea*, New York: Raven, pp. 349–379.

Wilson, J. P. (1980), *Hear. Res.* **2**, 233–252.

Wu, S. H., Oertel, D. (1984), *J. Neurosci.* **4**, 1577–1588.

Zurek, P. M. (1985), *J. Acoust. Soc. Am.* **78**, 340–344.

Zwislocki, J. J. (1975), in: D. B. Tower (Ed.), *The Nervous System, Vol. 3: Human Communication and its Disorders*, New York: Raven, pp. 45–55.

Further Reading

Altschuler, R. A., Bobbin, R. P., and Hoffman, D. W., (Eds.) (1986), *Neurobiology of Hearing: The Cochlea*, New York: Raven.

Edelman, G. M., Gall, W. E., Cowan, W. M., (Eds.) (1988), *Auditory Function: Neurobiological Bases of Hearing*, New York: Wiley.

Irvine, D. R. F. (1986), *The Auditory Brainstem*, Berlin: Springer.

Jahn, A. F., Santos-Sacchi, J., (Eds.) (1988), *Physiology of the Ear*, New York: Raven.

Kiang, N. Y. S. (1984), in: I. Darin-Smith (Ed.), *Handbook of Physiology. Section 1 (The Nervous System), Volume III (Sensory Processes, Part 2)*, Bethesda, MD: American Physiological Society, pp. 639–674.

Patuzzi, R., Robertson, D. (1988), *Physiol. Rev.* **68**, 1009–1082.

Pickles, J. O. (1988), *An Introduction to the Physiology of Hearing*, 2nd ed., London: Academic.

ERRATUM

Acoustics, Physiological (Vol. 1, p. 213). Mario A. Ruggero and Malcolm N. Semple.

Page 250. The first two sentences of section 10.7.1 Background Noise were incorrectly printed. They should read as follows:

The presentation of a tone against a background of broadband noise has somewhat different consequences for neural responses at peripheral and central levels of the auditory system. At both sites, background noise has the effect of elevating tonal thresholds, the threshold shifts being most pronounced at CF in responses of both auditory-nerve fibers (Geisler and Sinex, 1980) and cortical neurons (Phillips, 1989).

ACOUSTICS, PSYCHOLOGICAL

WILLIAM YOST, *Parmly Hearing Institute, Loyola University, Chicago*

	Introduction	261
1.	**The Auditory Code for Sound**	261
2.	**Psychophysics**	263
3.	**Absolute Sensitivity to Sound**	264
4.	**Differential Sensitivity to Sound**	267
5.	**Temporal Modulation Processing**	269
6.	**Interference Among Sounds (Masking)**	270
7.	**Locating Sound Sources**	274
7.1	Localization	274
7.2	Lateralization	276
7.3	Binaural Masking	279
8.	**Subjective Attributes of Sound**	281
8.1	Loudness	281
8.2	Pitch	283
8.3	Other Subjective Sound Attributes	285
9.	**Hearing Impairment**	286
9.1	Audiology and Otology	286
9.2	Tests for Hearing Disorders	286
9.3	Hearing Loss and Noise Exposure	287
9.4	Hearing Protection and Hearing Aids	288
	Glossary	289
	Works Cited	291
	Further Reading	292

INTRODUCTION

Psychological acoustics is the study of hearing: the examination of how the auditory systems of humans and other animals process sound. The study of psychological acoustics closely parallels that of physiological acoustics. The basic difference between the two fields is that physiological acoustics is concerned with how various parts of the nervous system function in processing sound, while psychological acoustics investigates how the entire organism processes sound. Psychological acoustics seeks to establish functional relationships between the physical characteristics of the sound input and the behavioral output of the responding organism. These relationships often depend on knowledge gained from physiological acoustics about the function of individual neural elements, while information gained from psychological acoustics often shapes the types of questions asked by physiological acoustics. Historically the behavioral study of sensory processing has been called "psychophysics," after the man, 1974a), a 19th century physicist and philosopher. "Psychoacoustics" is the description of the psychophysical study of hearing.

1. THE AUDITORY CODE FOR SOUND

Hearing in humans and other organisms can be divided simplistically into three stages: coding, processing, and integration. "Coding" refers to those mechanisms that transduce the physical characteristics of sound into a neural code. "Processing" refers to the organization of the neural code. "Integration" refers to the integration of the organized neural information with information from other neural mechanisms, such as those dealing with other sensory input, memory, or motor output. Thus, recognizing a familiar melody begins with neural coding of the frequency, amplitude, and phase/timing characteristics of the acoustic wave form; this neural code is

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

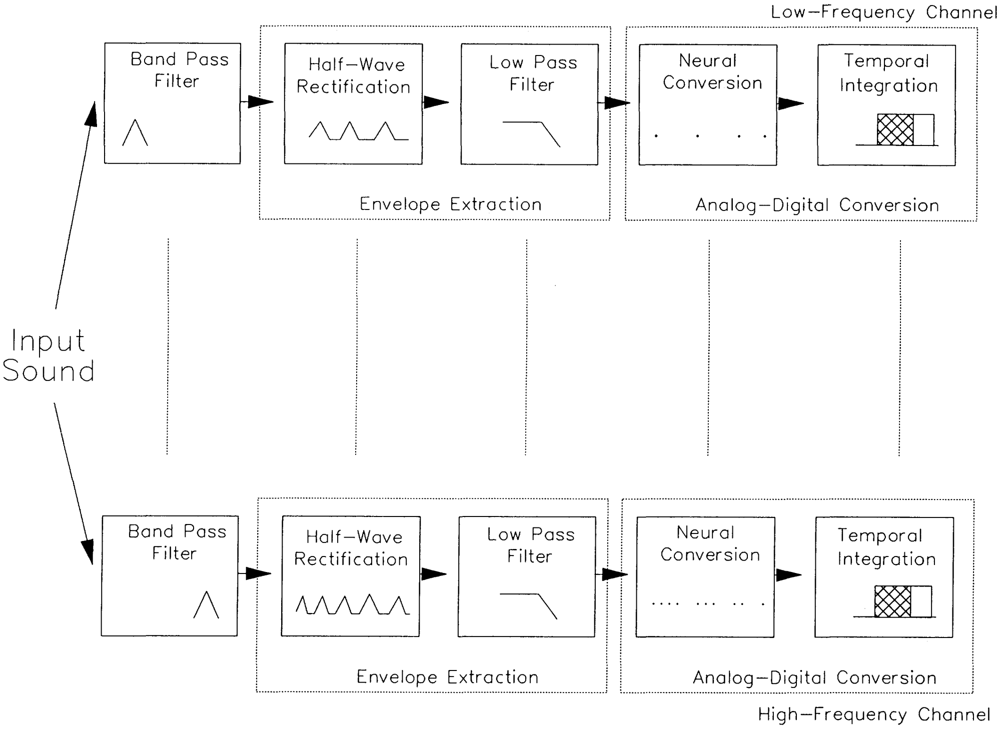
Low–Frequency Channel
Band Pass Filter
Half–Wave Rectification
Low Pass Filter
Neural Conversion
Temporal Integration
Envelope Extraction
Analog–Digital Conversion
Input Sound
Band Pass Filter
Half–Wave Rectification
Low Pass Filter
Neural Conversion
Temporal Integration
Envelope Extraction
Analog–Digital Conversion
High–Frequency Channel

FIG. 1. A schematic diagram depicting the various stages of sound processing assumed to occur within the peripheral auditory system. Sound is analyzed by a bank of bandpass filters with each filter having a different center frequency. The envelopes of the outputs of these bandpass filters are extracted through processes that mimic half-wave rectification and low-pass filtering. The analog nature of the vibratory signal is transduced to neural impulses where the rate at which the neural impulses occur is proportional to the frequency of the envelope rate. Certain processes accomplish various forms of temporal integration. The output of each bandpass filter is assumed to be a channel tuned to a particular frequency, carrying information about the intensity and timing/phase properties of the sound. In this figure the top row symbolizes a low-frequency channel and the bottom row a high-frequency channel. The actual processes are not necessarily organized in the manner depicted in the figure; however, these various aspects of sound processing are assumed to occur within the inner ear and bundle of nerves leaving the inner ear and traveling to the brain.

the melody; and then the organized code is integrated with memory and with other neural mechanisms allowing the listener to indicate that the melody is familiar.

Most of what has been studied in psychological acoustics, and, therefore, a great deal of what is known about hearing, concerns "coding." Less is known about "processing" and "integration." Figure 1 shows a simple, stimulus-processing model for establishing the neural code for sound. Sound waves vibrate the structures of the inner ear such that a spectral representation of sound is established, much as if the inner ear consisted of a bank of bandpass filters. The amplitude output of these filters is assumed to be squared, half-wave rectified, and low-pass filtered by processes that convert sound amplitude to energy and perform envelope extraction. The analog nature of these vibratory signals is converted to a digital code represented by all-or-none occurring neural impulses. Finally, some form of temporal integration is assumed to occur. The code for sound is a spectral code and many of the processes are nonlinear. The various processes described in Fig. 1 have been established by psychoacoustical and physiological experiments conducted over the past 100 years, and Fig. 1 contains components of the many "theories of hearing" (see Carterette and Friedman, 1974a, and Schubert, 1974) suggested over this period.

The neural output of the processes shown in Fig. 1 provides two types of codes for frequency: (1) the tuned channels containing the greatest amount of activity indicate which frequencies are present in the input (referred to as a "place code" for frequency), and (2) the periodicity of the neural activity in any one channel can code the periodicity and, hence, the frequency of the input (referred to as the "periodicity code" for frequency). Auditory neurons discharge in synchrony with the vibration pattern of the acoustic stimulus. Due to refractory limitations of neuronal function, such periodic information in a single nerve is available only up to approximately 5000 Hz. However, the nervous system may be able to extract timing information about an input stimulus for frequencies higher than 5000 Hz if it combines information from many similarly tuned neurons (referred to as the "volley principle," see Green, 1976).

Acoustic level is coded by increases in neural activity in any one tuned channel or from an increase in the number of tuned channels that are activated. Timing and phase information are coded by the time structure of the synchronous discharge firings of the neurons that make up the tuned channel.

The rest of this article will describe some of the basic characteristics of the auditory system as obtained from psychoacoustical data and will discuss some of the data that have been used to establish the model of coding shown in Fig. 1.

2. PSYCHOPHYSICS

A number of psychophysical paradigms have been established to measure the response of humans and animals to sound. When a human is asked, or an animal trained, to respond in a particular manner to sound, two classes of variables may affect the response: sensory variables and response bias variables (Watson, 1974). As an example, consider a situation in which a listener is asked to indicate if a tone has been presented. The tone may or may not be presented during a fixed trial marked by the onset and offset of a light and according to some random process. The listener indicates that the tone was present by saying "yes" and not present by saying "no." Thus, there are two possible stimulus situations (tone presented or tone not presented) and two responses ("yes" or "no"). Sensory variables are those dependent on processing the sound which might affect the response of the listener. Bias variables are all the other variables that may influence the listener to respond in a particular way (in this case to say "yes" or "no"). For instance, the properties of the assumed bandpass filters shown in Fig. 1 would be sensory variables. The instructions provided to the listener may be a bias variable. The instructions may be, "make sure the tone is present before responding 'yes,'" as opposed to, "respond 'yes' whenever the tone appears to occur." Under conditions in which the tone is not correctly detected 100% of the time, there would probably be more "yes" responses for the latter set of instructions than for the former. Clearly the number of "yes" responses depends on the instructions (bias) and on how the auditory system functions (sensory). Thus, unless the data are analyzed appropriately, measures of the listener's performance might reflect both types of variables (see Robinson and Watson, 1972).

Two general classes of psychophysical procedures are used to study sensory processing: *objective procedures*, usually used to measure thresholds; and *subjective techniques*, often used to obtain perceptual scales. Objective psychophysical procedures are designed to minimize or eliminate the influence of bias variables on the measured thresholds, while both bias and sensory variables can affect the listener's responses in subjective procedures. For instance, an objective psychophysical procedure would be used to measure the smallest difference in tonal frequency that a listener can discriminate. The smallest discriminable difference in frequency is called a threshold (in this case a difference threshold). A subjective procedure might be used to determine the perceived annoyance produced by aircraft noise. In this case, a scale of annoyance would be developed that might relate the magnitude of the perceived annoyance to the intensity of the airplane's acoustic output.

A threshold, as measured in objective psychophysical procedures, is obtained from a functional relationship between some measure of the listener's performance (e.g., percent correct in discriminating a low- from a high-frequency tone) and a physical measure of sound (e.g., the frequency difference between the low and high frequencies expressed

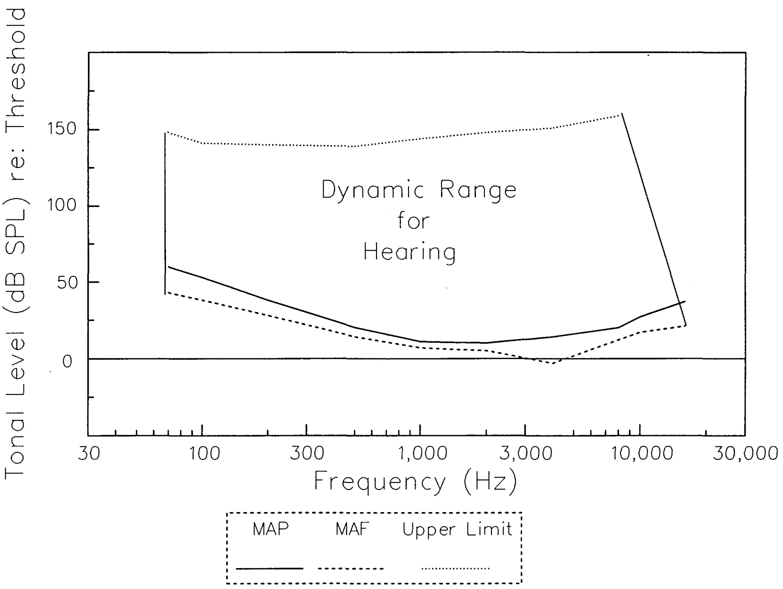

FIG. 2. The threshold levels (in dB SPL) for detecting the presence of a tone at different frequencies (MAP and MAF) are shown at the bottom of the figure, while the tonal levels leading to indications of damage to the auditory system (upper limit) are shown at the top of the figure. The two sets of functions define the dynamic range over which the auditory system operates. Thresholds in quiet have been measured when the tones have been presented from loudspeakers (MAF—Minimum Audible Field measures of thresholds) and over headphones (MAP—Minimum Audible Pressure measures of thresholds). Adapted from the data of Sivan and White (1933) as shown in Fig. 10.2 of Yost and Nielsen (1985).

in Hz). This function is called a "psychometric function." A threshold is the value of the physical measure (e.g., frequency in Hz) required for some arbitrary level of the listener's performance, chosen from the range of performance values between chance and perfect performance (e.g., 75% correct in discriminating a low- from a high-frequency tone). Thus, a threshold is a single stimulus value representing the approximate value for which the listener has difficulty performing the task. A threshold value is not an all-or-none level, meaning that if the stimulus exceeds the threshold the system always responds and if it is not exceeded the system never responds. This classical, all-or-none use of threshold is not supported by either psychophysical or physiological data (see Green and Swets, 1977). Rather, the performance of the listener represents a continuum, usually centered at the threshold, of the probability of responding.

Psychophysical scales can be obtained by a variety of subjective methods. In matching procedures, listeners are asked to adjust a physical variable (e.g., intensity) of one sound until this sound is judged to be subjectively equivalent (e.g., in loudness) to a standard sound. In magnitude estimation procedures, listeners are asked to estimate directly a subjective attribute of a sound by assigning numbers to the sound in proportion to the magnitude of the perceived perception (e.g., loudness). In both matching and scaling procedures, numeric scales are developed which relate the subjective perception (e.g., loudness) to a physical variable (e.g., sound level) (see Stevens, 1975).

3. ABSOLUTE SENSITIVITY TO SOUND

The absolute sensitivity of the auditory system to tones of different frequencies defines the frequency range over which the auditory system functions and determines the lower limit of sound pressure required for hearing. Thresholds for detecting the presence of sinusoidal tones are measured by two calibration methods using objective psychophysical paradigms. When sounds are presented over a loudspeaker, the thresholds are referred to as Minimal Audible Field (MAF) thresholds, while thresholds obtained from headphone-presented stimuli are referred to as Minimal Audible Pressure (MAP) thresholds. In the MAF procedure, the sound pressure level is measured in an anechoic (echo-free) room at the location where the listener would be (without the listener present). Both the direction of the sound relative to the listener and whether the listener uses one or both ears for listening determine the exact threshold values. In the MAP procedure, the sound pressure level is calibrated in an "artificial ear," which is a 6-cm^3 coupler affixed to the headphone. The artificial ear approximates the volume between the headphone and the eardrum, as well as simulating the combined acoustic impedance of the outer ear and eardrum for the average young listener. Thus, the MAP thresholds represent sound levels that would occur at the listener's eardrum, while the MAF thresholds represent those at the entrance to the outer ear.

The sound pressure levels are referenced to 20 micropascals (μPa), referred to as dB SPL (decibels of Sound Pressure Level); for instance,

$$20 \text{ dB SPL} = 20 \log(200\ \mu\text{Pa}/20\ \mu\text{Pa}),$$

where 200 μPa is the sound-pressure level being measured. 20 μPa is the international standard sound pressure level for the least intense sound the average young listener can detect. As a point of reference, the acoustic power represented by a sound pressure of 20 μPa (0 dB SPL), is approximately 20–30 dB more intense than the theoretic acoustic power existing at sea level due to thermal energy of the Brownian motion of air molecules (see Green, 1976). The term dB SL (SL for Sensation Level) is the level of sound relative to that measured for absolute threshold. Thus, the level of sound required for absolute threshold is determined and then used as the referent value for a SL decibel calculation.

Figure 2 depicts MAP and MAF thresholds as originally obtained by Sivian and White (1933), and Table 1 depicts standardized MAP thresholds. The standardized thresholds are used to determine hearing loss, and these procedures will be discussed in a later section of this chapter. The thresholds indicate that the human auditory system responds to frequencies over the range from 20 to 20 000 Hz and that the human auditory system is most sensitive to frequencies in the range of 1000–4000 Hz. The approximate 6-dB difference between the MAF and MAP measures of thresholds is due to the resonance, diffraction, and impedance properties of the outer ear and eardrum. Recall that MAF sound pressures are those at the entrance to the outer ear, while the MAP sound pressures are those at the eardrum.

Table 1. Standardized minimal audible pressure thresholds in dB SPL.

Frequency (Hz)	Range of thresholds[a]	Thresholds[b]
125	45.0–54.0	45.0
250	24.5–30.5	25.5
500	11.0–14.4	11.5
1000	6.5–8.0	7.0
2000	6.0–9.5	9.0
4000	3.5–13.5	9.5
6000	8.0–17.5	15.5
8000	9.5–20.5	13.0

[a] From American National Standards Institute (ANSI) 3.6-1989, American National Standard Specification for Audiometers, and ISO 389-1975, International Standards Organization (ISO) Standard Reference Zero for the Calibration of Pure-Tone Audiometers. A range of thresholds due to different headphones, cushions, and coupler combinations used in different countries.

[b] From ANSI 3.6-1989 for TDH-39 Headphones in a MX41/AR cushion and for a 9-A coupler.

The upper curve in Fig. 2 represents an estimate of the maximum sound pressure the human ear can tolerate. This estimate is obtained by increasing the level of a tone until the listener indicates that one of a number of sensations occurs: they "feel" the sound, there is a "tickling" sensation, or the sound is "uncomfortable." All of these sensations indicate that the sound pressure is close to the limit tolerated by the auditory system (a later section of this article will discuss some of the effects on hearing caused by intense sound presentations). The dynamic range of hearing can be as large as 120 dB, and the auditory system adapts to sound over this enormous range essentially instantaneously. A dynamic range of 120 dB exceeds that which can be achieved in almost all audio equipment.

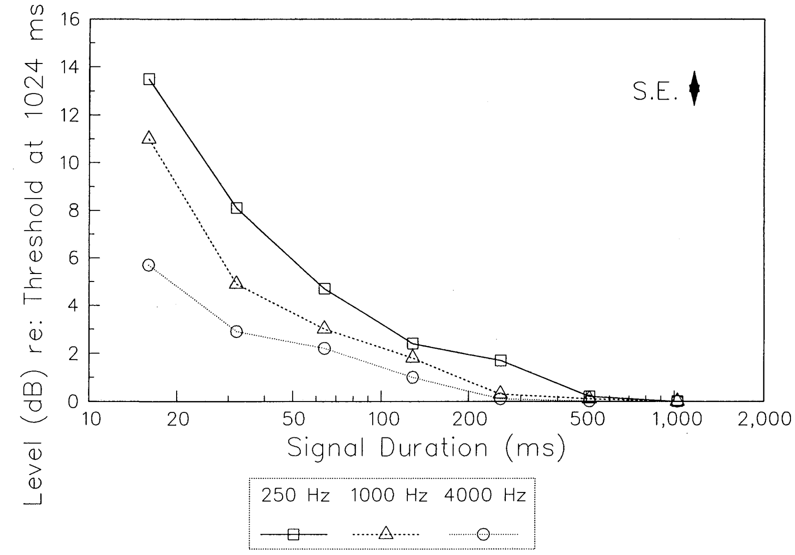

FIG. 3. The relative change in threshold level (in dB) required to detect the presence of a tone as a function of tonal duration. The change in thresholds is referenced to those obtained when the tones were 1024 ms long. Data for three frequencies (250, 1000, and 4000 Hz) are shown. As tonal duration decreases thresholds increase. The arrows labeled S.E. are an estimate of the standard error of measurement for these conditions and as such represent the expected variability about any data point. Adapted from the data of Watson and Gengel (1969) as shown in Fig. 10.3 of Yost and Nielsen (1985).

The thresholds displayed in Fig. 2 were obtained when the sounds were presented for half a second or longer. The thresholds forhearing are dependent on the duration of the sound, as shown in Fig. 3. When expressed as the total power of the signal, the thresholds decrease as duration increases to approximately 300 ms and then they remain constant as the duration is increased further. To a first approximation the thresholds change by approximately 10 dB for a tenfold change in duration down to durations of 20–50 ms; thus, between 50 and 300 ms the thresholds are actually total-equal-energy thresholds. At very short durations, the change in threshold is much steeper than 10 dB for a tenfold change in duration, and the slopes of the curves in Fig. 3 are frequency dependent.

The bandpass filtering, energy transformation, and temporal integration shown in Fig. 1 account for most of the effects of duration on absolute thresholds displayed in Fig. 3. As will be discussed in more detail in the section on masking, the auditory system is assumed to process information in channels (psychoacoustically referred to as "critical bands") narrowly tuned to each frequency, much as bandpass filters are tuned. Thus, detection of a tone is assumed to depend on the integrated energy in the channel tuned to the frequency of the tone. For detecting a pure tone, the integration time is assumed to be approximately 300 ms, accounting for the thresholds reaching an asymptote at a signal duration of 300 ms. For short durations, the energy of the signal is spread over a wider spectral range than that of the channel, which causes less energy to fall within the channel. If a certain amount of energy must be present within the channel for threshold performance, then the signal level must be increased at these short durations to obtain the necessary energy required for threshold. Thus, the increase in the slope of the functions in Fig. 3 at short durations is a result of energy spreading beyond the spectral region to which the processing channel is tuned. Because the spectral region over which the channels are tuned is different for each tuned channel, there would be a change in the thresholds as a function of frequency, resulting in the slopes of the functions changing with frequency as also shown in Fig. 3.

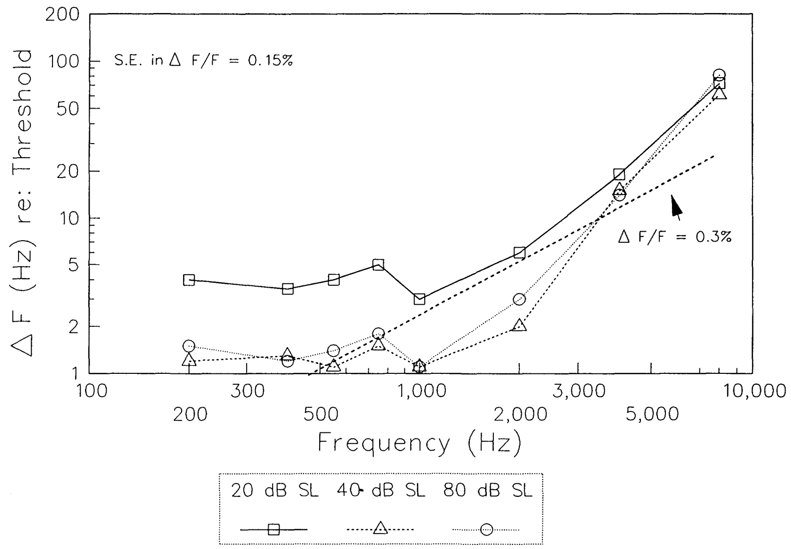

FIG. 4. The threshold for discriminating a change in frequency (Δ*F*) as a function of frequency. Data for three tonal levels (20, 40, and 80 dB SL) are shown. The dotted line labeled $\Delta F/F = 0.03\%$ indicates a Weber fraction showing a 0.03% change in frequency as a function of frequency. Below approximately 1000 Hz, a 1–2 Hz change in frequency can be discriminated, between 1000 and 5000 Hz approximately a 0.03% change in frequency is required for discrimination, and above 5000 Hz, frequency differences greater than 0.03% are required for discrimination. The Δ*F*/*F* labeled S.E. is an estimate of the standard error of measurement for these conditions and as such represents the expected variability about any data point. Adapted from the data of Weir, Jesteadt, and Green (1977) as shown in Fig. 10.6 of Yost and Nielsen (1985).

4. DIFFERENTIAL SENSITIVITY TO SOUND

The resolution of the auditory system can be estimated by measuring the smallest differences in frequency, sound level, and time that a listener can discriminate. The data plotted in Figs. 4–6 show the difference thresholds (sometimes labeled "just noticeable differences" or "jnds") for discriminating a change in frequency (*ΔF* as shown in Fig. 4), a change in sound level (*ΔI* as shown in Fig. 5), and a change in time (*ΔT* as in Fig. 6). As can be seen, the difference thresholds tend to increase as the variable under study increases. The ratio of the difference threshold to the variable (Δ*F*/*F*, Δ*I*/*I*, Δ*T*/*T*) is called the "Weber fraction." Over a significant range of frequency and intensity the Weber fractions for frequency (Δ*F*/*F*) and for sound level (Δ*I*/*I*) remain approximately constant. These data indicate that the resolution for discriminating a change in frequency is approximately 1–2 Hz for frequencies less than 1000 Hz and approximately 0.03% between 1000 and 5000 Hz. The threshold for discriminating a change in sound level for low and moderate level tones is approximately 1.0 dB [$10 \log_{10}(1+0.25)$]. Obviously the Weber fractions for frequency and sound level do not remain constant over the entire range of frequency and intensity. The departures of the data from a constant Weber fraction (sometimes referred to as the "near miss to Weber's law") are important for specifying completely the resolving capabilities of the auditory system (Laming, 1986).

The Weber fraction for time (Δ*T*/*T*) is constant only over a limited range of times, although the jnd (Δ*T*) does increase with increasing lengths of time. For all of these measures of discrimination, care must be used to ensure that the variable undergoing change is the only physical parameter actually varying. This is a difficult task when the duration of sound is changed. A sound's spectrum changes as a function of changing its duration. The auditory system is sensitive to the change in spectrum associated with a change in stimulus duration. Thus, the measure of time discrimination shown in Fig. 6 involves discriminating a change in the time separation between two pulsed noise stimuli.

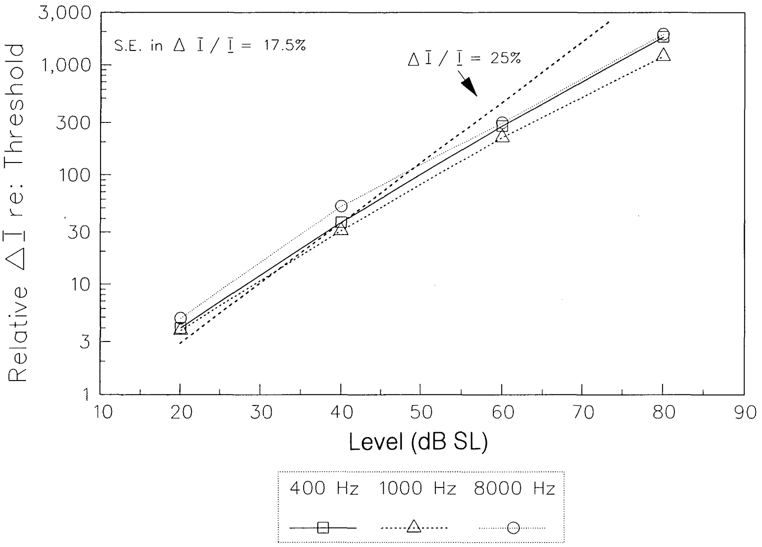

FIG. 5. The relative thresholds for discriminating a change in amplitude (ΔI) as functions of level (in dB SL). The values of ΔI were scaled such that 5 units of amplitude were required for threshold performance at 8000 Hz and 20 dB SL. Data for three tonal frequencies are shown (400, 1000, 8000 Hz). The dotted line labeled $\Delta I/I=25\%$ indicates a Weber fraction showing a 25% change in amplitude as a function of level. One can convert from $\Delta I/I$ for amplitude to the decibel difference between the two tonal levels being discriminated by $10 \log_{10}[1+(\Delta I/I)]$. A $\Delta I/I$ of 25% is approximately a 1-dB difference in tonal level. The $\Delta I/I$ labeled S.E. is an estimate of the standard error of measurement for these conditions and as such represents the expected variability about any data point. For tones presented at low stimulus levels listeners can discriminate approximately a 1-dB difference in level. At higher levels listeners can discriminate a smaller difference, representing a "near miss to Weber's law." Adapted from the data of Jesteadt, Weir, and Green (1977) as shown in Fig. 10.8 of Yost and Nielsen (1985).

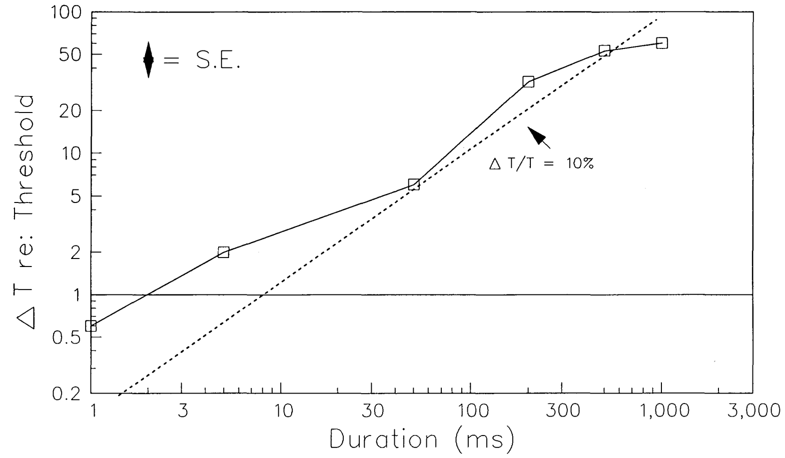

FIG. 6. The threshold for discriminating a change in the temporal separation between two sounds (ΔT) as a function of the temporal separation (in ms). The dotted line labeled $\Delta T/T=10\%$ indicates a Weber fraction showing a 10% change in time separation as a function of the separation. The data are described by a constant Weber fraction only over a small range of temporal separations. The arrows labeled S.E. are an estimate of the standard error of measurement for these conditions and as such represent the expected variability about any data point. Adapted from the data of Abel (1972) as shown in Fig. 10.9 of Yost and Nielsen (1985).

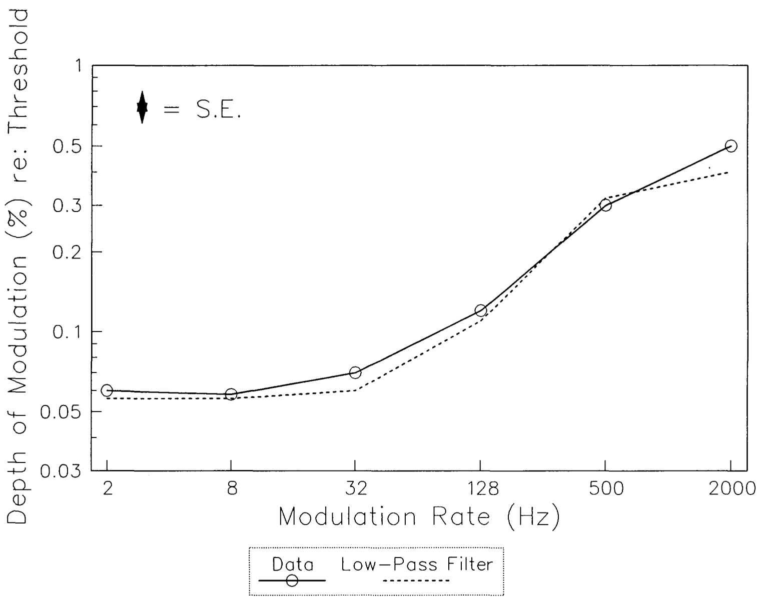

FIG. 7. The threshold depth (in %) of sinusoidal amplitude modulation (SAM) required to detect the presence of SAM for a noise stimulus, as a function of the rate of SAM (in Hz). The task was to discriminate between a nonmodulated noise $[n(t)]$ and the SAM noise $\{[1+m\cos(2\pi rt)]n(t)\}$. The depth of modulation (m expressed as a percentage) required for threshold performance was determined as a function of modulation rate (r expressed in Hz). The dotted curve represents the predictions based on a model that includes a low-pass filter with a time constant of approximately 3 ms. The arrows labeled S.E. are an estimate of the standard error of measurement for these conditions and as such represent the expected variability about any data point. Adapted from Fig. 1 of the chapter by Viemeister in Evans and Wilson (1977).

The spectral changes associated with varying the time between two acoustic markers are much less likely to influence discrimination performance than the spectral changes associated with varying the duration of a tonal signal.

5. TEMPORAL MODULATION PROCESSING

The data in Figs. 3 and 6 show some aspects of the auditory system's sensitivity to the duration of sound and the time separation between sounds. As explained above, obtaining estimates of the auditory system's sensitivity to changes in the temporal properties of sound is difficult due to the system's remarkable sensitivity to the spectral and sound level changes that accompany most temporal changes. Thus, a number of different procedures have been used to study auditory temporal processing (Green, 1971).

Procedures using amplitude modulated (AM) sounds offer a means of studying temporal processing, especially auditory processing of ongoing temporal changes of sound. Figure 7 shows the results from an experiment in which listeners were asked to discriminate between an unmodulated white, Gaussian noise $[n(t)]$ and a sinusoidally amplitude modulated (SAM) noise $\{[1+m\cos(2\pi rt)]\times[n(t)]\}$. The depth of modulation (m expressed as a percent) required to just discriminate the SAM noise from the unmodulated noise was determined for each frequency of modulation (r ranged from 1 to 2000 Hz). As the rate of modulation increases beyond approximately 50 Hz, greater and greater depth of modulation is required to process the temporal fluctuations due to amplitude modulation. Up to 50 Hz the temporal fluctuations are all approximately equally discriminable from the unmodulated noise. The low-pass filter characteristic of these data has been used to support the low-pass filter assumption shown in Fig. 1. The

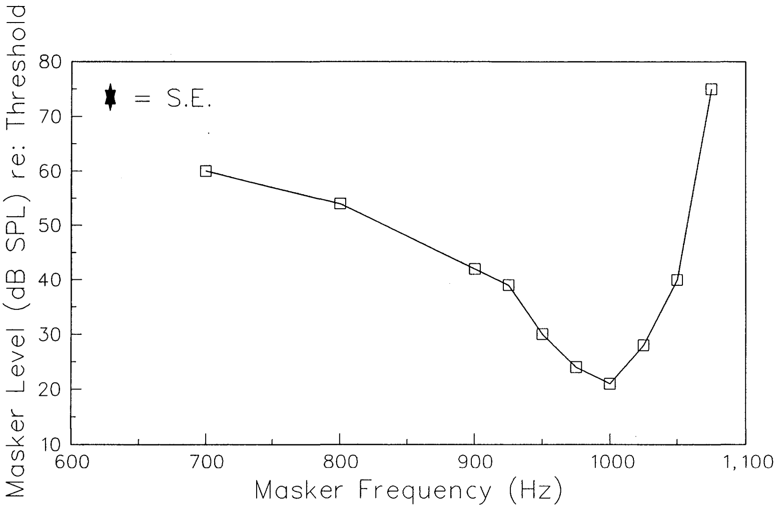

FIG. 8. The threshold level (in dB SPL) of a tonal masker required to mask a 1000-Hz tonal signal, as a function of masker frequency (in Hz). The lowest masker level is required for threshold performance when the masker and signal have the same frequency (1000 Hz), and greater masker level is required to mask the signal when the masker frequency differs from the signal frequency. These data are interpreted as indicating that the detection of the signal is mediated by neural channels tuned to the signal frequency. The arrows labeled S.E. are an estimate of the standard error of measurement for these conditions and as such represent the expected variability about any data point. Adapted from the data of Moore (1982a) as shown in Fig. 11.1 of Yost and Nielsen (1985).

curve fitted to the data was obtained from a model using a low-pass filter with a time constant of 3 ms. These and other data show that to a first approximation, the auditory system's ability to detect ongoing changes in sound can be modeled with a low-pass filter process as suggested in Fig. 1.

6. INTERFERENCE AMONG SOUNDS (MASKING)

The data discussed so far describe the auditory system's sensitivity to tones presented in isolation. Most of our daily experience with sound involves many sound sources. When there are many sound sources, it is possible that one sound will interfere with the ability to process other sounds. Masking is one procedure used in psychoacoustics to study interference. Masking is defined as the increase in the detection threshold of one sound (the signal sound) due to the presence of another sound (the masker sound). The amount of the threshold shift reflects the amount of masking (see Moore, 1986).

Figure 8 shows tonal masking data in which the signal was a brief duration (20 ms), 1000-Hz tone, presented at a low level (5 dB SL). The masker was a tone presented at different frequencies. The listener's task was to determine if the signal and masker were present or only the masker was present. Masker level was varied until a threshold for detecting the signal was obtained. The data are plotted as the masker level required to achieve threshold as a function of masker frequency. The form of the resulting data (Fig. 8) is explained by assuming that the auditory system monitors the neural channel tuned to the frequency of the signal. When the masker frequency is greatly different from the signal frequency, the neural channel tuned to the signal does not respond when the masker is present alone but does respond when the signal plus masker are present, requiring the masker level to be increased to obtain threshold. When the masker frequency is near that of the signal, the tuned channel responds to both masker and masker plus signal, requiring a decrease in the level of the masker to obtain a threshold. These data can be used to define the characteristics of the tuned channels which are usually modeled as bandpass filters (see Fig. 1).

Another procedure to study masking uses filtered white, Gaussian noise as a masker for tonal signals. In a use of this technique, a band-reject noise (a noise passed through filters such that the energy is reduced in a nar-

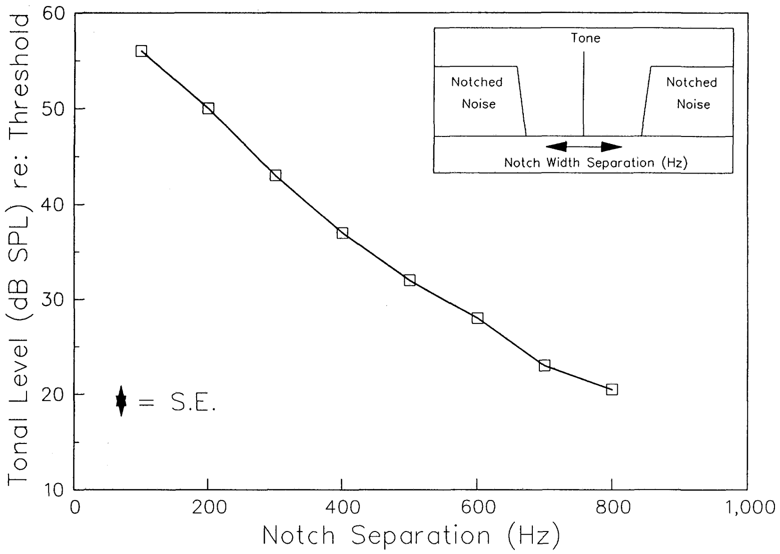

FIG. 9. The threshold level (in dB SPL) of a tonal signal as a function of the spectral notch separation (expressed in Hz) in a band-reject noise masker. The inset depicts the tonal signal with a frequency in the center of the spectral notch created by the band-reject noise. The threshold for detecting the signal is measured as the spectral width of the notch is increased. The threshold is lower at wide spectral separations between the two bands of noise that make up the band-reject noise. These data are consistent with the assumption that masking of a tonal signal is mediated by auditory channels tuned to the tonal-signal frequency. The arrows labeled S.E. are an estimate of the standard error of measurement for these conditions and as such represent the expected variability about any data point. Adapted from the data of Patterson (1976) as shown in Fig. 3.6 of the chapter by Patterson and Moore in Moore (1986).

row band of the noise) is used to mask a tonal signal, as shown in the inset of Fig. 9. The threshold for detecting the tonal signal is measured as a function of the spectral separation between the two bands of noise that comprise the band-reject noise. The results shown in Fig. 9 are explained in a manner similar to that used to explain the data of Fig. 8. When the frequency of the edge of the filtered noise is distant from the signal frequency (spectral notch is wide), the channel tuned to the signal frequency will respond when the signal plus masker are present, but not when the masker alone is present. Thus, the signal level can be reduced to obtain a threshold. When the frequencies of the edge of the filtered noise and that of the signal are close together (spectral notch is narrow), the tuned channel responds to both the masker presented alone and the signal plus masker. In this case, the signal level must be increased to obtain a threshold. The data from experiments like those depicted in Figs. 8 and 9 have been used to describe the characteristics of the tuned channels described as bandpass filters in Fig. 1.

Because only a band of masker frequencies near the signal frequency is critical for determining the amount of masking, the tuned channels have been called "critical bands" (see Moore, 1986). Figure 10 shows an estimate of the width of this critical band of frequencies as obtained for noise maskers and tonal signals of different frequencies. The critical bandwidth increases as signal frequency increases; however, the increase in bandwidth is not a fixed proportion of the center frequency as it is for physically realizable filters. To a first approximation, the bandwidth of the "critical band" filters changes as described in Fig. 10, but the shape of the filter does not change appreciably as a function of frequency.

Threshold detection in masking experiments is obtained when the signal-to-masker ratio remains approximately constant. The data of Fig. 11 show this constancy for conditions involving noise maskers and tonal signals. In addition, the estimated bandwidth of critical band filters is somewhat dependent on the overall level of the masker and signals used to measure the critical band. The

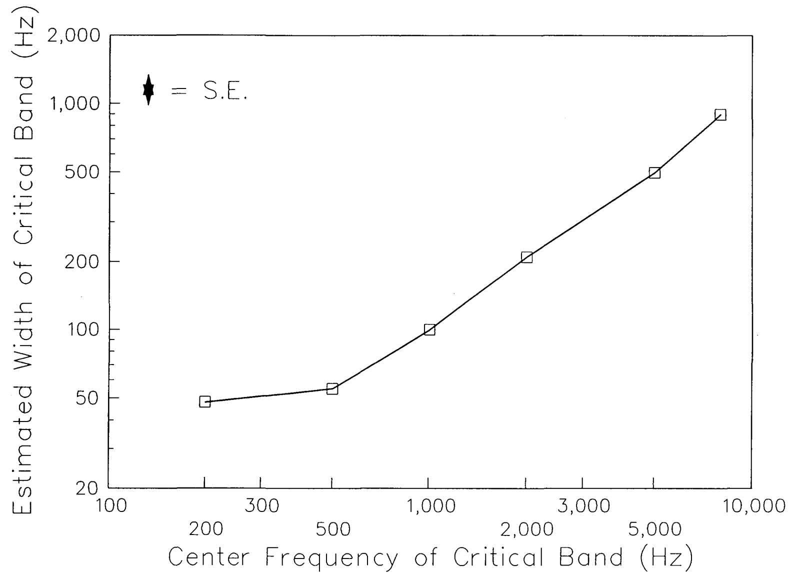

FIG. 10. The width (in Hz) of the band of frequencies critical for masking a pure tone, as a function of the center frequency of the band (in Hz), which would be the same frequency as the tonal signal. Only stimuli with frequencies within the critical band region are effective in masking tones at the center frequency. The arrows labeled S.E. are an estimate of the standard error of measurement for these conditions and as such represent the expected variability about any data point. Adapted from Moore and Glasberg (1983) as shown in Fig. 3.10 of the chapter by Patterson and Moore in Moore (1986).

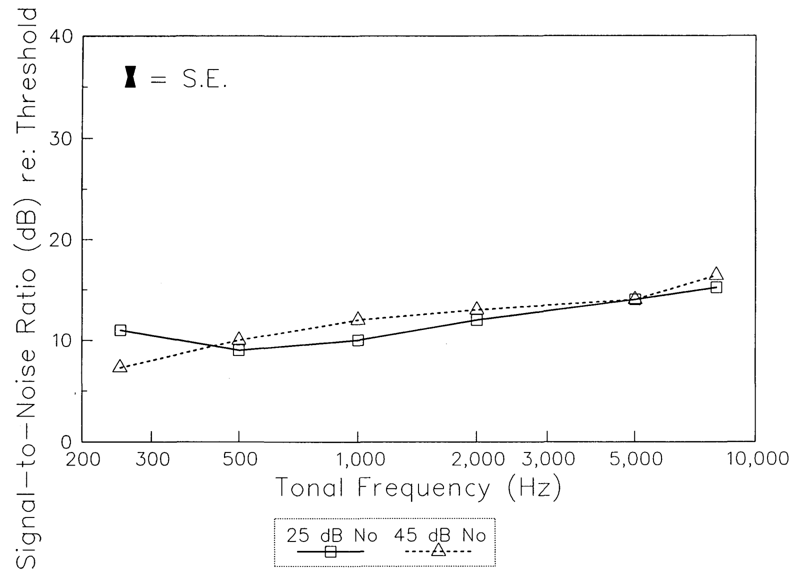

FIG. 11. The tonal signal-to-noise ratio (in dB) required for threshold detection of the signal, as a function of tonal-signal frequency (in Hz). Data for two noise spectrum levels are shown (25 dB No and 45 dB No). Spectrum level, designated as "No," is the noise power per unit band width assuming the noise has a relatively flat amplitude spectrum. The signal-to-noise ratio (energy of the signal at threshold divided by the spectrum level of the noise) required for threshold remains approximately constant in the range of 9–15 dB over a large range of frequencies and overall masker level. The arrows labeled S.E. are an estimate of the standard error of measurement for these conditions and as such represent the expected variability about any data point. Adapted from the data of Reed and Bilger (1973) as shown in Fig. 11.6 of Yost and Nielsen (1985).

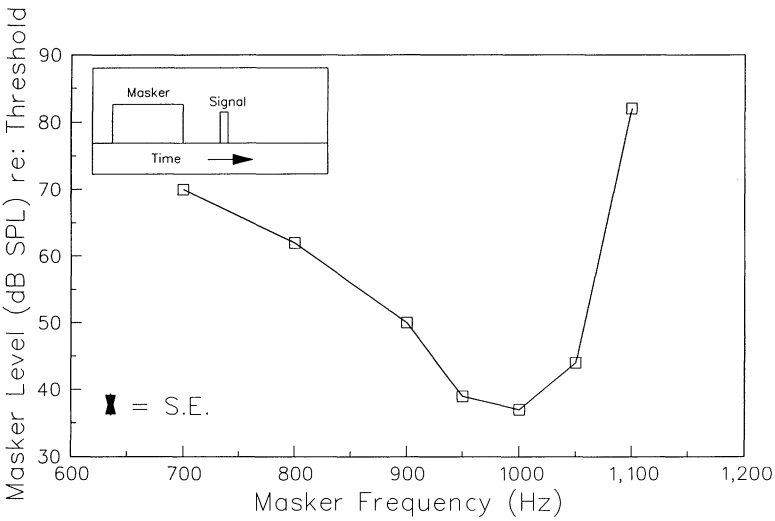

FIG. 12. The tonal masker level (in dB SPL) required for threshold detection of a 1000-Hz signal, as a function of the masker frequency (Hz). The tonal signal was presented after the offset of the masker as shown in the inset. In this forward-masking paradigm, the neural effect of the masker is assumed to remain after its offset, resulting in masking that is similar to that occurring when the signal and masker are presented simultaneously (see Fig. 8). The arrows labeled S.E. are an estimate of the standard error of measurement for these conditions and as such represent the expected variability about any data point. Adapted from the data of Moore, Glasberg, and Roberts (1984) as shown in Fig. 4.7 of the chapter by Moore and O'Loughlin in Moore (1986).

estimated critical bandwidth increases a small amount as a function of increasing stimulus level.

Figure 12 shows that results similar to those in Fig. 8 are obtained when the signal and masker do not occur at the same time. In the "forward masking" paradigm used to obtain these data, the short duration tonal signal occurred 1 ms after the masker was turned off. Forward masking is assumed to result from the neural effects of the masker persisting after the masker is terminated. The amount of forward masking decreases as the time separation between the masker and signal increases or as the signal increases in duration relative to that of the masker. Usually the data for the forward-masked signal imply narrower tuning (Fig. 12) than obtained when the masker and signal are presented simultaneously (Fig. 8). Explanations for the difference in the results assume that some nonlinear processes affect the signal and masker in the simultaneous conditions but not when the signal is presented after the masker in forward masking (see Moore and O'Loughlin, 1986). The shape of the data obtained in forward-masking experiments matches similar measures obtained from single auditory nerve fibers of the auditory nerve leaving the inner ear of the auditory periphery. This agreement suggests that the auditory system uses the neural information in the auditory nerve as the primary basis for detecting masked tonal signals.

In the masking studies described above, only the level of the signal or the masker was varied. In everyday life, sounds that are to be detected are often constantly changing in level and the task is to detect a change in the level of the signal *relative* to that of the masker. Figure 13 describes the stimulus conditions for a study that attempts to measure detection of a relative change in signal level for a masker whose level is also varied. In this example the masker consists of 11 equal amplitude tones spaced logarithmically in frequency, and the signal is a level increment to the middle, 1000-Hz component. The overall level of the masker was varied over a 40-dB range from condition to condition. The data in Fig. 14 show the results expressed as the level of the signal relative to that of the masking components as a function of the number of tonal components in the masker. The signal level required for threshold decreases as the number of components increases. With 11 components performance is nearly as good as that obtained with one masking tone. These

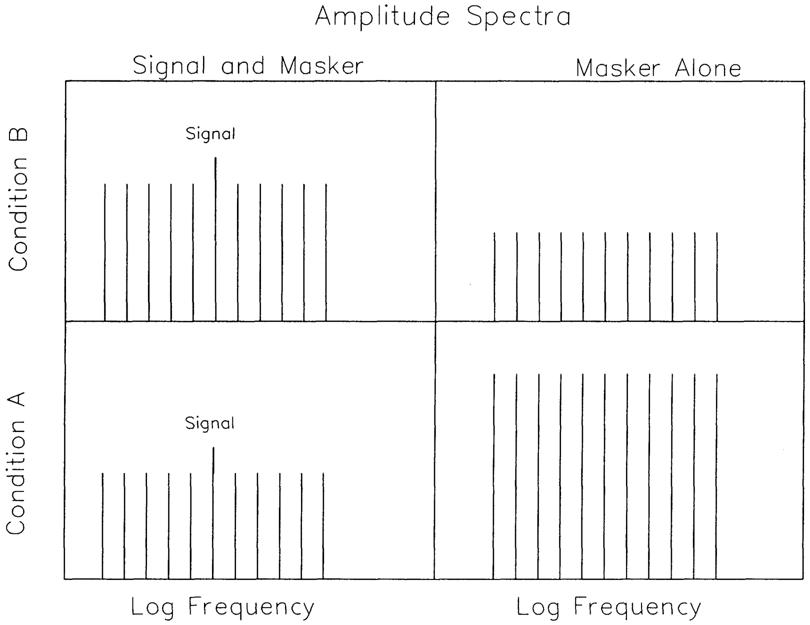

FIG. 13. A schematic diagram depicting the spectra of maskers and signals used in a masking study. The masker consists of 11 equal-amplitude tones with frequency spaced logarithmically. The signal is a level increment to the middle component of the masker. The listener is presented either the masker or the masker and signal and he/she must detect when the signal occurs. Conditions A and B represent two of the many possible comparisons (the overall level of the masker is randomly varied over a 40-dB range). In Condition A, the masker happens to be more intense during the signal presentation than when the signal was not present, while in Condition B the masker is less intense when the signal is present than when only the masker is present. As Fig. 14 will indicate, listeners are able to detect low-level signals in this procedure and, thus, detection is assumed to be based on comparing the outputs of many tuned channels to detect the relative increase in signal level in the spectral profile.

data are interpreted in terms of the auditory system's ability to compare the spectral profile when the signal is present to masker-alone conditions. When the signal is present, the spectral profile has a " bump" at the signal frequency. With many components in the masker, it is easier to define the spectral profile and hence to find the bump at the signal frequency. Such an explanation suggests that the auditory system compares the output of more than one tuned channel in order to define the profile. These data also indicate that the auditory system may not lose much resolving power when it is presented with a dynamically changing stimulus context in which to detect a signal (see Green, 1988).

7. LOCATING SOUND SOURCES

The experiments described above have helped shape our understanding of the auditory code for frequency, intensity, and temporal events. The auditory system is also capable of determining the spatial location of sound. A variety of mechanisms exist throughout the animal kingdom for locating sounds (see Yost and Gourevitch, 1987). In humans and most mammals, sound localization is achieved via binaural processing for determining the azimuthal location of sound and includes the spectral transform of sound as it travels from a source to the middle ear for vertical location. Little is known about our poor ability to judge distance when loudness is not a possible cue.

7.1 Localization

The cues for localizing sounds in the azimuth plane are the interaural difference in arrival time of the sound and the interaural level difference as depicted in Fig. 15. The interaural time difference is determined by the difference in sound travel time between the two ears, and the interaural level difference is

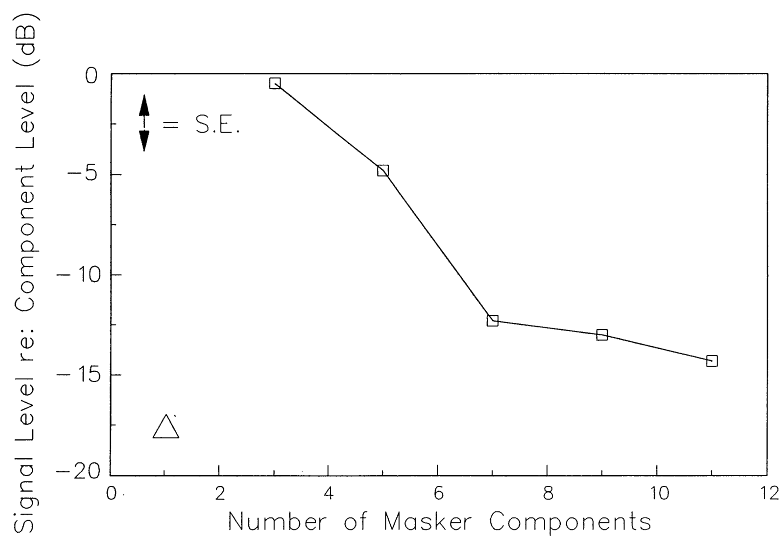

FIG. 14. The tonal-signal level (in dB relative to the level of the masker components) required for threshold detection, as a function of the number of equal-amplitude components in the masker (as shown in the spectral profiles of Fig. 13). The signal is always an increment in level to the middle component of the multitone masker when the components are spaced logarithmically in frequency. As described in Fig. 13, the overall level of the masker is randomly varied over a 40-dB range. As the number of components in the masker increases, the threshold for signal detection decreases. The triangle above one masker component represents the threshold for detecting an increment to a single tone. Thus, as the number of components in the masker increases, performance can approximate that obtained for a single tonal masker, even though the masker is varying randomly in level over a 40-dB range. The arrows labeled S.E. are an estimate of the standard error of measurement for these conditions and as such represent the expected variability about any data point. Adapted from the data of Green (1988, Fig. 4-2).

due to the sound shadow and diffraction produced by the head. In general, the further the sound source is from the center (0° in Fig. 15), the greater the interaural time difference and interaural level difference. Although there are cones of confusion where the interaural differences are the same for a number of positions (e.g., directly in front and directly behind a listener), asymmetries in the geometry of the head and torso and head movements eliminate most confusions by listeners. Figure 16 displays the complex patterns of interaural time differences [Fig. 16(a)] and interaural level differences [Fig. 16(b)] produced by the human head and torso for tones of different frequencies. These data indicate a complex relationship among the measures of interaural time and level, sound frequency, and sound location (see Kuhn, 1987).

Figure 17 displays the errors in localizing tones of different frequencies. The loss of acuity in the mid-frequency region near 2000 Hz is assumed to be due to the two cues (interaural time and interaural level) operating in different frequency regions (this theory has been called the duplex theory of sound localization; see Durlach and Colburn, 1978). Interaural level differences are most salient at high frequencies where the head's sound shadow is most pronounced. Interaural time differences for pure tones are ambiguous when the period of the tone is less than the maximum sound travel time difference between tween the ears. Thus, interaural time provides an unambiguous cue for frequencies below approximately 1600 Hz which corresponds to a period of 0.625 ms, approximately the maximum difference in travel time between the ears for a sound placed directly opposite one ear. The region near 2000 Hz (see Fig. 17) is the spectral region where neither cue works very well. Although the duplex theory of sound localization appears to account for the localization of sinusoidal stimuli, it does not predict localization performance for complex stimuli (see Yost and Hafter, 1987).

Figure 18 displays data from an experiment designed to determine the smallest separation between two sound sources that a listener can detect. These minimum audible angles (MAA) are plotted as a function of

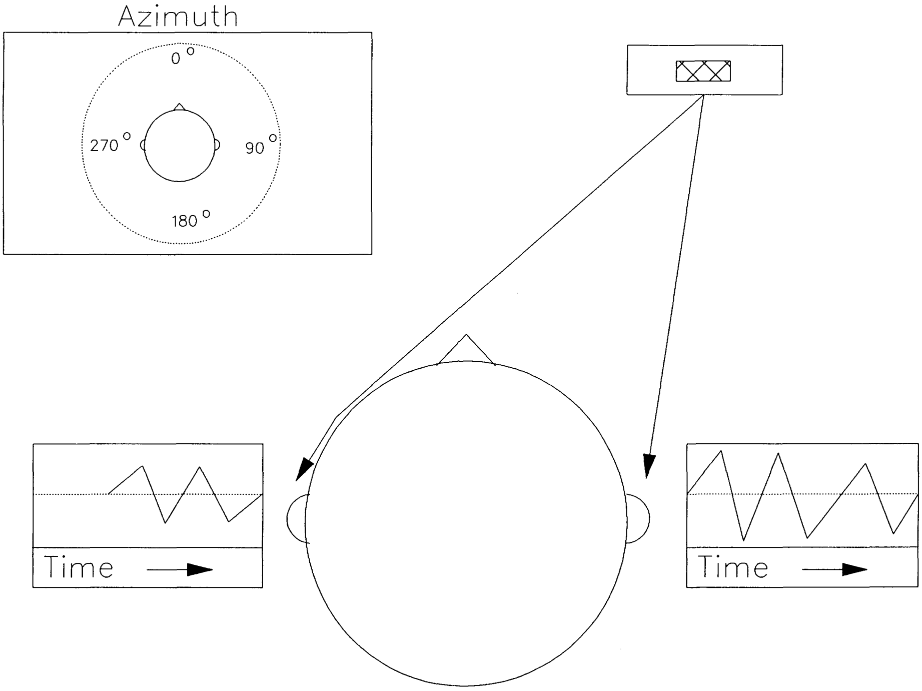

FIG. 15. A schematic diagram depicting the differences in interaural time and level at the ears when a sound source is presented to one side of a listener. The inset shows the normal reference system for describing azimuth position. Sound arrives first at the ear nearest the sound source, and because of the sound shadow produced by the head and body, the sound is also most intense at the ear closest to the source. Thus, as the source moves from in front to one side of a listener the interaural time and level differences increase.

tonal frequency and starting azimuth. Highest acuity is obtained for discriminations between sources directly in front of the listener and, as in Fig. 17, there is a loss of acuity in the mid-frequency region near 2000 Hz.

In most real-world listening environments sounds encounter many reflective surfaces before they reach a listener's ears. The sound waves reflected from these surfaces could be confused as additional sound sources. However, in most cases humans localize the sound as coming from the source and not from one or more of the reflective surfaces. It has been shown that the sound that reaches the ears first determines the location of the perceived source. In most situations, the first arriving sound is from the source. However, if the reflective environment prevents this, then a listener will hear a "phantom source" associated with the reflective surface which sends the sound to the ears first. The fact that the first arriving information at the ears determines sound localization is called the "precedence effect" (see Zurek, 1987).

7.2 Lateralization

In order to control the variables of interaural time and level, stimuli are often presented over headphones. Such headphone-delivered stimuli are usually perceived as being inside the head rather than externalized as real sound sources. Because such headphone-delivered stimuli are perceived to be located along an imaginary line between the ears, when the interaural time or level of a sound is varied, localization of headphone-delivered sounds is called "lateralization." In general, the lateralized image of the sound moves away from midline toward the ear receiving the sound first or receiving the higher level sound. Thus to a first approximation, the position of lateralized sounds concurs with their positions in real space as determined by their interaural time and level differences (see Yost and Hafter, 1987).

Figure 19 shows the difference thresholds for detecting a change in the interaural level [Fig. 19(a)] or interaural time differences of

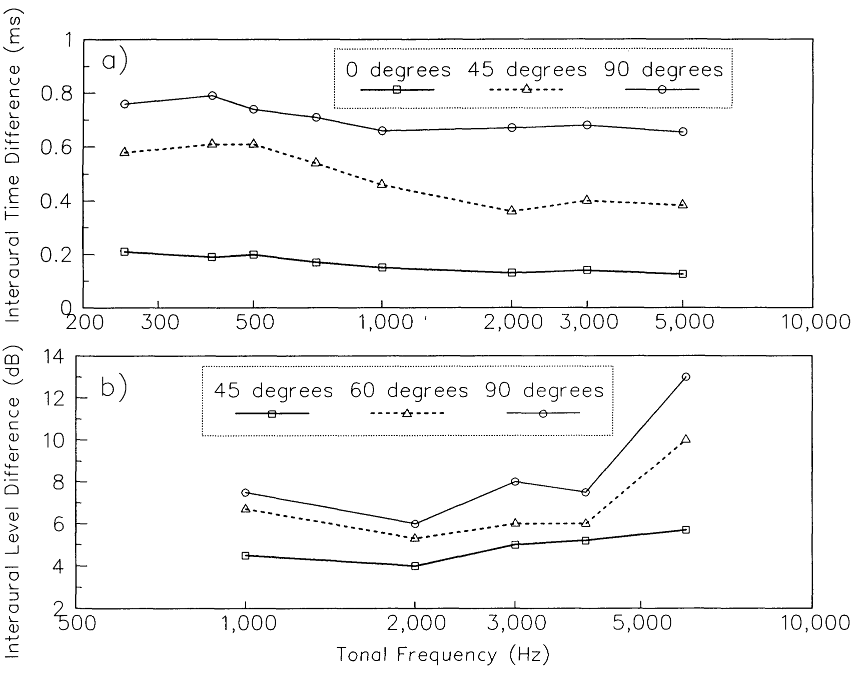

FIG. 16. The interaural time difference in ms (a) and interaural level difference in dB (b) measured between the ears of listeners, as functions of frequency (in Hz). The measurements were made at three azimuth positions [0°, 45°, 90° for (a); and 45°, 60°, 90° for (b); see Fig. 15]. The values of interaural time and level are complex functions of frequency and azimuth position. Adapted from the data of Kuhn (1987, Figs. 1.3, 1.7, and 1.8) in Yost and Gourevitch (1987).

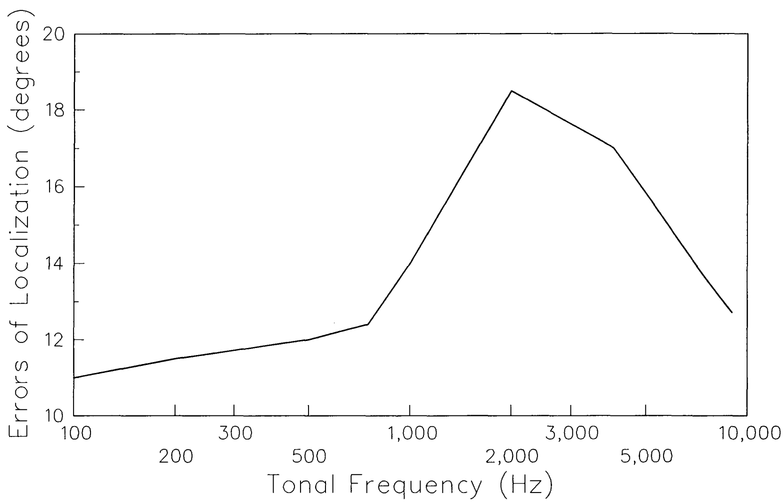

FIG. 17. Errors (expressed as degrees of azimuth) for locating tonal sound sources, as a function of tonal frequency (in Hz). Localization accuracy is worst in a frequency region between 1200 and 5000 Hz. These data are the basis for the duplex theory of localization. Adapted from the data of Stevens and Newman (1936) as shown in Fig. 2.1 of the chapter by Wightman, Kistler, and Perkins in Yost and Gourevitch (1987).

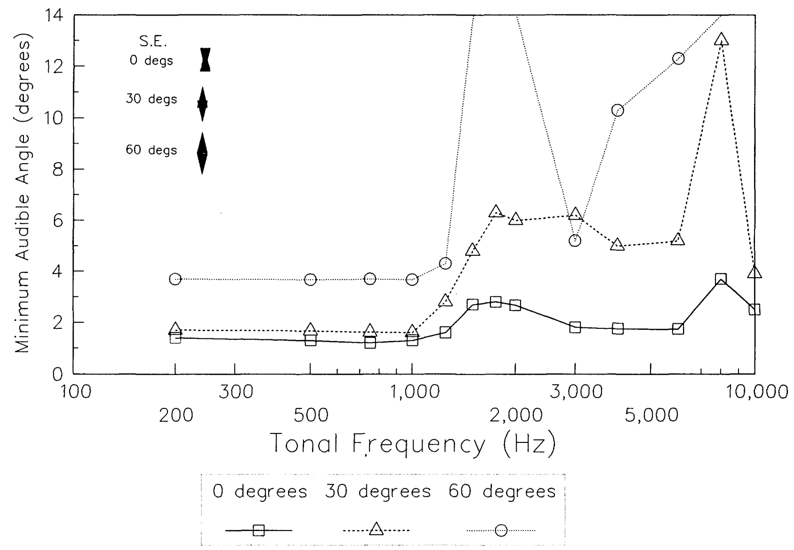

FIG. 18. The minimum audible angle (MAA shown in degrees of azimuth) required to discriminate between the location of two tonal-sound sources, as a function of frequency (in Hz). The MAAs were measured at three azimuths (0°, 30°, 60°). The MAA increases as the sources are moved from the front to one side of the listener, and performance is worst in the frequency region of 1200–5000 Hz. The arrows labeled S.E. are an estimate of the standard error of measurement for the three base azimuth conditions and as such represent the expected variability about any data point. Adapted from the data of Mills (1956) as shown in Fig. 2.1 of the chapter by Wightman, Kistler, and Perkins in Yost and Gourevitch (1987).

tones [Fig. 19(b)] as functions of frequency and the initial amount of interaural time or level. Best sensitivity to interaural level difference is approximately 0.5–1 dB across a large range of frequencies. However, sensitivity to changes in interaural time differences can be measured only for tonal frequencies less than 1600 Hz. Interaural time differences as small as 10 μs can be discriminated. When discriminations are made for sounds that have an initial interaural time or level difference (interaural differences greater than zero), the difference thresholds increase. Because increases in interaural time or level are perceived as images off midline, these data are interpreted as indicating that spatial acuity decreases as the location of a sound source is moved from midline or away from center (as was also shown in Fig. 18).

Complex high-frequency sounds with low-frequency temporal fluctuation(envelopes) can be discriminated on the basis of interaural differences of time even though high-frequency tones (above 1600 Hz) cannot be discriminated on the basis of interaural time. Figure 20 shows the results for lateralizing a 3900-Hz tone that was sinusoidally amplitude-modulated at 300 Hz. Listeners *cannot* discriminate a change in interaural time difference for the 3600-Hz tone but can make the discrimination when the 3900-Hz tone is amplitude modulated at 300 Hz. A 300-Hz sinusoidally amplitude- modulated 3900-Hz tone has a spectral component at 3600 Hz (the lower, side-band frequency associated with SAM). Thus, the lowest frequency in the SAM complex was 3600 Hz, and for this reason the 3600-Hz tone was used as the comparison frequency. The ability to localize high-frequency complex sounds on the basis of interaural time differences is assumed to be due to the binaural auditory system's use of the slow-frequency repetition in the envelope (300 Hz) of the amplitude modulated tone.

Thus, interaural level is potentially available as a localization cue at all frequencies, although in the real world very small interaural level differences (perhaps too small to be detected) actually exist at low frequencies. Interaural time differences for very narrow-band stimuli (e.g., pure tones) can be processed only for frequencies less than 1600 Hz. Interaural time differences can be discriminated for high-frequency sounds if the narrow-band stimuli contain a slow-frequency repetition. These conclusions per-

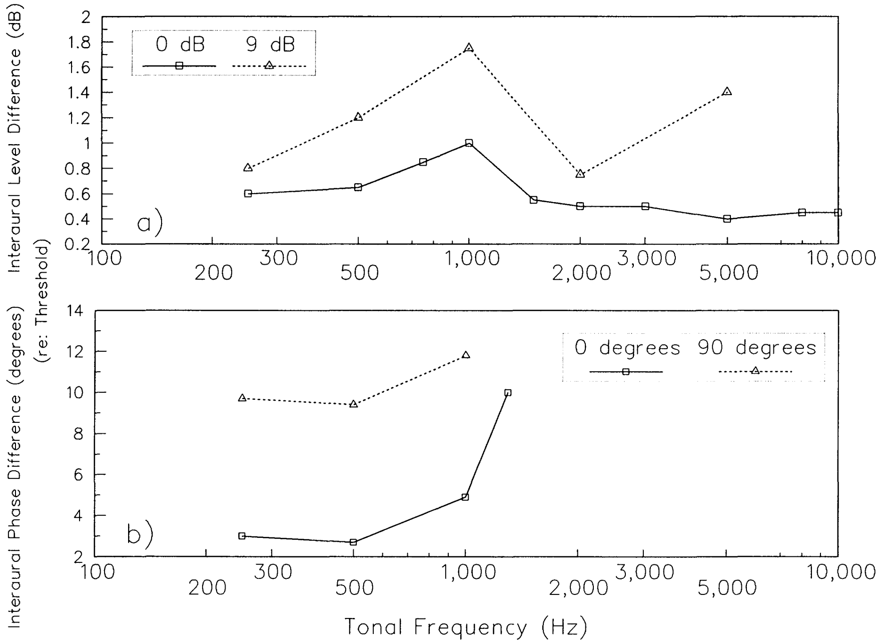

FIG. 19. The values of interaural level [in dB, (a)] and interaural phase [in degrees, (b)] required to discriminate a change in the interaural time (phase) or interaural level difference of a tone, as functions of tonal frequency (in Hz). The tones were presented over headphones with different base values of interaural level [0 dB or 9 dB in (a)] or interaural phase [0° or 90° in (b)]. The thresholds for discriminating a change in interaural level or phase increase as the value of the base interaural phase or interaural level differences increases. The smallest interaural level difference thresholds are between 0.5 and 1 dB except in a frequency region near 1000 Hz. Interaural phase-difference thresholds can be obtained only for frequencies less than 1600 Hz. The expected variability about any datum point in (a) is approximately 0.7 dB, while in (b) it is approximately 2°. Adapted from the data of Yost and Dye [1987, (a)] and Yost [1974, (b)] as shown in Figs. 3.9 and 3.10 of the chapter by Yost and Hafter (1987) in Yost and Gourevitch (1987).

tain primarily to localization of azimuth position (see Blauert, 1983).

Sound is transformed in a number of ways as it travels from its source to a listener's middle ear. These transformations can be determined by comparing the sound measured by small microphones placed in the outer ear to that originating at the source. "Head transfer functions" such as those shown in Fig. 21 can be established; they describe the magnitude (and phase) changes that a sound undergoes as it travels past the various parts of the head and torso.

When sounds are presented over headphones, they are not transformed by the head and body. When the magnitude and phase changes of the head transfer function are introduced back into the spectrum of a complex sound before being delivered to headphones, listeners usually report that the sounds are no longer internalized but are externalized at their real-world locations. Table 2 shows a comparison of localization judgments made when sounds were presented from loudspeakers in an anechoic room versus sounds presented over headphones after they were transformed by the inverse of the head transfer functions measured for the listener making the judgments. There is excellent correspondence for both azimuth and vertical judgments. These results suggest that for broadband stimuli, the transformations that sounds undergo while traveling to the ears are crucial for externalization and accurate localization of the sound source (see Wightman, Kistler, and Perkins, 1987).

7.3 Binaural Masking

The masking of one sound by another is reduced if the two sounds occupy different locations in space relative to conditions when the two sounds occupy the same location (see Blauert, 1983). Signals and maskers can also

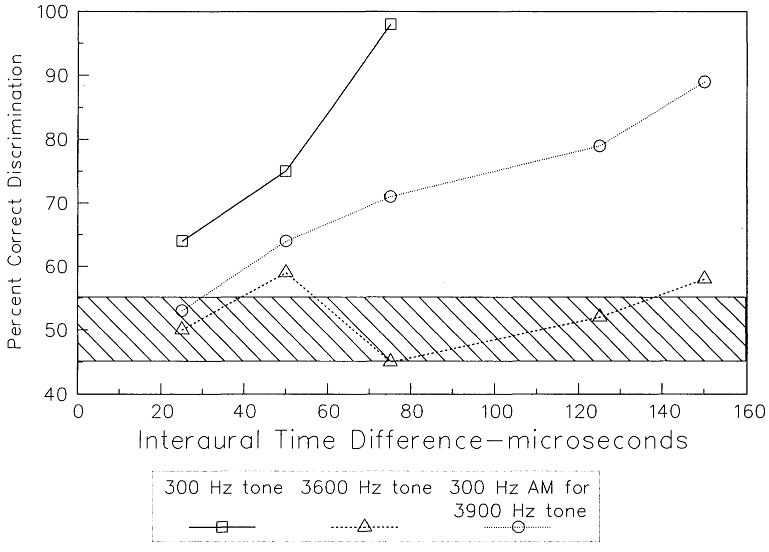

FIG. 20. The percentage correct discriminations of a change in interaural time differences as a function of the interaural time difference (in microseconds). Three stimuli were used: a 300-Hz tone, a 3600-Hz tone, and a 3900-Hz tone that was sinusoidally amplitude modulated (AM) at 300 Hz. As the interaural time difference increases, the listener's ability to discriminate a change in the interaural time difference for the 300-Hz tone and the 300-Hz AM 3900-Hz complex improves, but not for the 3600-Hz tone. The hatched area represents the region of chance performance indicating that the listener was unable to discriminate a change in the interaural time difference for the 3600-Hz tone. These data indicate that high-frequency [above approximately 1600 Hz, see Fig. 19(b)] sounds can be lateralized only when they contain a form of low-frequency temporal modulation. Adapted from the data of Henning (1977) as shown in Fig. 12.7 of Yost and Nielsen (1985).

be separated in lateral space when the sounds are presented over headphones. If the masking stimulus is presented with a different interaural difference than the signal, the signal is easier to detect (has a lower threshold) than when the masker and signal are presented with the same interaural difference. A nomenclature for defining the interaural differences of the masker (N, since the masker is usually a noise) and the signal (S) has been developed. Table 3 describes this nomenclature and shows the masked thresholds for a variety of binaural conditions. Whenever the signal and masker have different interaural values (e.g., the NoSπ condition) thresholds are lower than when the signal and masker have the same interaural values (e.g., the NoSo condition). This difference in masked threshold between a dichotic condition and a diotic condition is referred to as the binaural masking-level difference (BMLD). The BMLD has been used to describe the advantage in detection offered by spatially separating sound sources. This description is based on the observation that when the masker and signal with different interaural values are presented over headphones, they are perceived at different lateral positions within the head. Thus, binaural processing enables the auditory system to locate sounds and to detect one sound in the presence of another when the target sound is spatially separated from the interfering sound (see Durlach and Colburn, 1978).

Spatial localization is most often modeled by assuming that the nervous system performs a cross correlation between the neural information at each ear. A coincidence network has been suggested as a way in which cross correlation could be neurally implemented. The location of coincidence along the network indicates the location of sounds in space. It is usually assumed that such a coincidence network would occur at the output of each filter mechanism, as indicated in Fig. 1. Thus, the pattern of coincidences across the spectrally coded array of networks would form a type of neural spatial map for sound source location (see Colburn and Durlach, 1978).

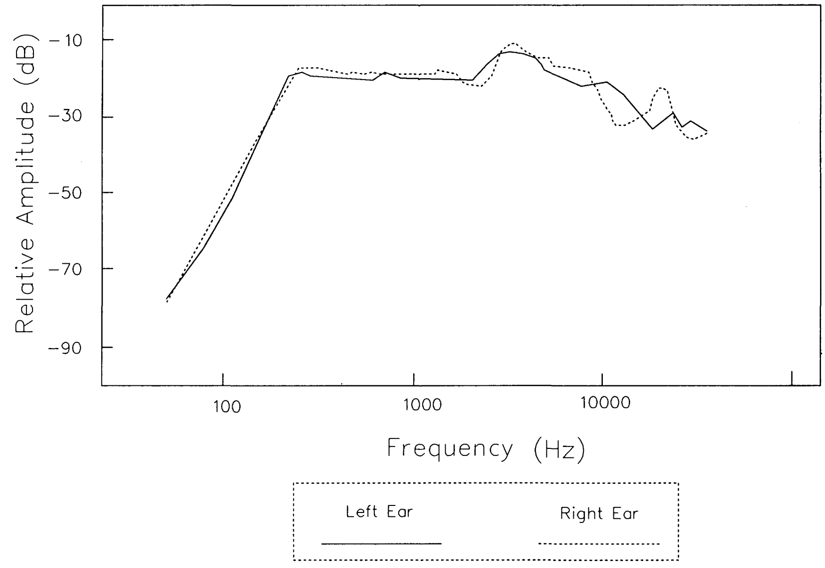

FIG. 21. The relative amplitude (in dB) measured in each outer ear of a listener as a function of frequency (in Hz). The sound source was located 2 m directly in front of the listener. These "head transfer functions" indicate the complex pattern (especially at high frequencies) that exist in the outer ears due to the filtering actions of the head and torso. The differences in the two functions are due to slight asymmetries in the head and torso of this listener. Adapted from the data of Wightman, Kistler, and Perkins (1987) as shown in Fig. 2.9 of Yost and Gourevitch (1987).

8. SUBJECTIVE ATTRIBUTES OF SOUND

The data discussed so far concern a listener's ability to detect stimuli in different stimulus contexts or to discriminate changes in stimulus events. The physical variables of sound—frequency, level, complexity, and time—have subjectively perceived counterparts: pitch, loudness, timbre, rhythm, etc. Subjective psychophysics is the study of the relationship between the physical variables of sound and their subjective attributes (see Stevens, 1975).

8.1 Loudness

Figure 22 shows equal-loudness contours for a range of tonal frequencies and sound levels. These data were obtained by asking listeners to adjust the sound level of a test tone until it is perceived equally loud to that of a comparison tone. The comparison tone is always 1000 Hz, but it is presented at different sound levels expressed in dB SPL. For each level of the comparison tone, the listener is presented test tones at a number of different frequencies and is asked to adjust the level of

Table 2. Correlation coefficients between absolute judgments of sound source location and actual sound source location for two listeners (S1 and S2). The listeners either judged the location of real sources (broad-band noise presented from one of 22 loudspeakers) in an anechoic chamber (free field) or over headphones when the spectra of the noise wave forms were transformed by the "head transfer functions" obtained from microphones in the outer ears of the listeners when they were in the free field (simulated free field). The correlations reflect judgments of both the azimuthal and vertical position of the 22 sound sources. The very high correlation coefficients indicate that listeners are very accurate in locating sound sources in both the free field and with headphones, if the headphone sounds are modified by the head transfer functions. Adapted from the data of Wightman, Kistler, and Perkins (1987).

Condition	Listener	Correlation[a]
Free field	S1	0.99
	S2	0.97
Simulated free field	S1	0.96
	S2	0.83

[a] Goodness-of-fit correlation coefficients ranging from −1.0 to 1.0, where 0 means there is no correlation.

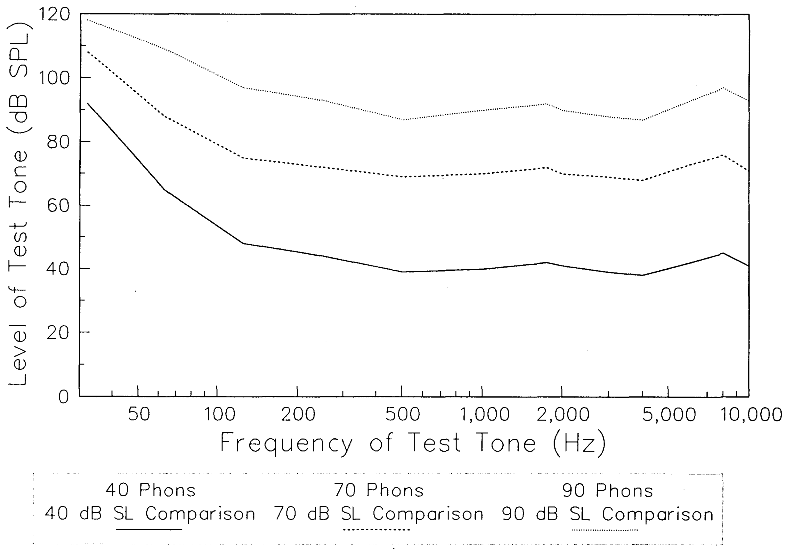

FIG. 22. The level of a test tone (in dB SPL) required to match the subjective loudness of a comparison, 1000-Hz tone, as a function of the frequency (in Hz) of the test tone. Data for three levels of the comparison tone are shown (40 dB SL, 70 dB SL, and 90 dB SL). All tones presented at the level and frequency along any one "equal loudness contour" are judged equally loud. These data are used to define a decibel scale of loudness, the "phon" scale. A tone has a level of *x* phons if it is judged equally loud as a 1000-Hz tone presented at *x* dB SPL. Hence, the three curves are the 40-, 70-, and 90-phon equal-loudness contours. Adapted from ISO Standard (1981) as shown in Fig. 13.1 of Yost and Nielsen (1985).

the test tone so that it appears "equally loud" to the 1000-Hz comparison tone presented at a particular level. The data shown along any one contour on Fig. 22 are, therefore, assumed to be equally loud, thus leading to the name of these curves as "equal loudness contours." All tones judged equally loud to the 1000-Hz comparison tone presented at 40 dB SPL are defined as having a loudness of 40 phons, those judged equally loud to the 1000-Hz comparison tone presented at 70 dB SPL are 70 phons loud, etc.

These data and the resulting phon scale of loudness have been standardized and have a number of practical uses. For instance, the 40-phon equal-loudness contour defines a frequency-weighting function used in sound level meters. That is, when making a sound level measurement that relates to human hearing, the sound needs to be filtered so that

Table 3. Masked thresholds in signal-to-noise ratios (dB) in a variety of binaural masking conditions. The difference in thresholds between a diotic or monotic condition and a dichotic condition is the binaural masking-level difference (BMLD).

Binaural condition	Description	Threshold (dB)	BMLD (dB)
NmSm (monotic)	Masker (N) at only one ear (m); signal (S) at only one ear (m)	13	Referent
NoSo (diotic)	Masker (N) the same at both ears (o); signal (S) the same at both ears (o)	13	0
NπSm (dichotic)	Masker (N) out of phase between ears (π); signal (S) at only one ear (m)	11	2
NoSm (dichotic)	Masker (N) the same at both ears (o); signal (S) at only one ear (m)	3	10
NπSo (dichotic)	Masker (N) out of phase between ears (π); signal (S) the same at both ears (o)	1	12
NoSπ (dichotic)	Masker (N) the same at both ears (o); signal (S) out of phase between ears (π)	−2	15

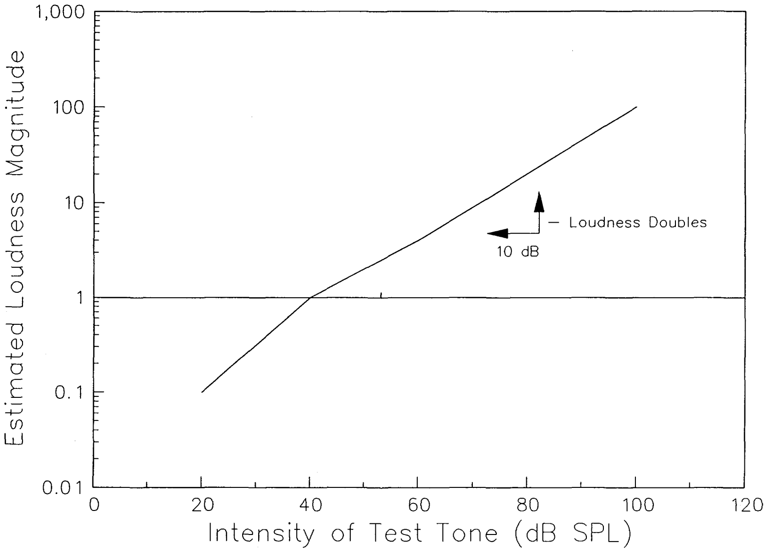

FIG. 23. The estimated magnitude of loudness for a 1000-Hz tone as a function of its level (in dB SPL). In a scaling procedure, the listener's task is to estimate the loudness of the tone by assigning a number proportional to the tone's subjective loudness, such that if the tone doubles in subjective loudness the estimated magnitude should also double. Most listeners indicate that a 10-dB increase in level is required for a doubling of loudness. A "sone" scale of loudness is derived from the scaling procedure such that a 40-dB SPL, 1000-Hz tone is 1 sone loud and any sound of *x* sones is judged to be *x* times as loud as 1 sone. Thus the ordinate is also loudness in sones. Adapted from Fig. 13.2a of Yost and Nielsen (1985).

only energy in the frequency region to which humans are sensitive is used in the sound level measurement. The 40-phon equal-loudness contour is used to define a weighting function that attenuates the levels of a sound at low and high frequencies. When the sound-pressure level has been measured based on the 40-phon weighting function, the sound-pressure level is expressed as dBA or A-weighted decibels. Most measurements of environmental sounds are based on dBA calculations.

The data of Fig. 23 show how the perceived loudness of sound changes when the sound level changes. These data were obtained by asking listeners to judge the loudness of a 1000-Hz tone. The listeners judged the loudness of the tone by using numbers to scale loudness such that if the listeners perceived the loudness to double, the value of the assigned number should double. This form of a "scaling procedure" indicates that the physical intensity of sound needs to be increased approximately 10 dB in order for the perceived loudness to double. Scaling has been used in a number of experiments to determine the relationship between the perceived subjective change in sound and the actual physical change in the sound, such as the "sone scale" explained in Fig. 23 caption.

8.2 Pitch

Pitch has also been measured with matching and scaling procedures. However, the scales or indicators of pitch most often used are those derived from music. Table 4 shows some of these scales.

Although the perceived pitch of a sound is highly correlated with the existence of spectral energy at frequencies equal to the reported pitch (i.e., the pitch of a 400-Hz tone is 400 Hz), many stimulus conditions can produce a pitch sensation when there is little or no energy in the frequency region of the sound that corresponds to the perceived pitch.

As stated previously, the transduction of sound into a neural code is nonlinear. Consequently, audible nonlinear distortion products are generated when sounds are presented, especially at high levels. Table 5 indicates the nonlinear distortion products that are audible when a two-tone complex is presented at a fairly high level. Under most circumstances the cubic difference tone is the

Table 4. The relationship among musical notes, cents, and fundamental frequency (in Hz) for three major ways of tuning for pitch. An octave is divided into 12 equal logarithmic intervals of 100 "cents" each, yielding 1200 cents per octave.

	Just intonation		Equal temperament		Pythagorean	
Musical note	Cents	Freq. (Hz)	Cents	Freq. (Hz)	Cents	Freq. (Hz)
C	0	264	0	264	0	264
D	204	297	200	296	204	297
E	386	329	400	332	408	334
F	498	352	500	353	498	352
G	702	396	700	395	702	396
A	884	440	900	444	906	445
B	1088	495	1100	498	1110	501
C (next octave)	1200	528	1200	528	1200	528

most audible of all of the distortion products. This indicates that the form of the nonlinearity is not a simple nonlinear transformation of the acoustic stimulus.

The existence of nonlinear distortion products has been known since the 18th century when musicians such as Tartini (1754) observed that a third tone (the difference tone) can be heard when two tones are played simultaneously. Tartini observed that the strength of the difference tone is most noticeable when the two generating tones had frequencies of simple ratios such as 2:3, 3:4, 4:5, and 5:6. Tartini and other 18th century composers wrote music based on the perception of the difference tones.

One procedure that indicates the existence of these nonlinear products is to present a slightly mistuned tone near the reported pitch of the suspected distortion tone. For instance, if a 1000-Hz tone is presented at a high level, then a listener may report hearing the aural harmonic pitch of 2000 Hz. If a 2003-Hz tone is added to the 1000-Hz original tone, then almost all listeners report that a faint sound with a pitch near 2000 Hz is beating in loudness three times per second. The 2000-Hz aural harmonic, produced by the auditory system's nonlinear response to the original 1000-Hz tone, and the 2003-Hz added tone are beating at a rate of 3 Hz. By adjusting the level of the 2003-Hz tone until the beating sensation is strongest, the level of the nonlinear aural harmonic at 2000 Hz can be estimated, since the strongest beating sensation occurs when two beating tones are equal in amplitude. This estimated level of the distortion tone can be confirmed and its phase estimated if a 2000-Hz tone is added to the 1000-Hz and 2003-Hz stimulus complex. Once the level of the 2003-Hz tone is adjusted until the listener perceives the strongest beat with the 2000-Hz aural harmonic, the listener then adjusts the level and phase of the 2000-Hz tone to cancel this beating sensation. That is, if the 2000-Hz tone is equal in level and out of phase with the 2000-Hz aural harmonic (produced by the original 1000-Hz tone), then the 2000-Hz aural harmonic and the 2000-Hz physical tone are cancelled. All the listener hears is the 1000-Hz tone and the 2003-Hz tone; there is no perception of a 2000-Hz tone

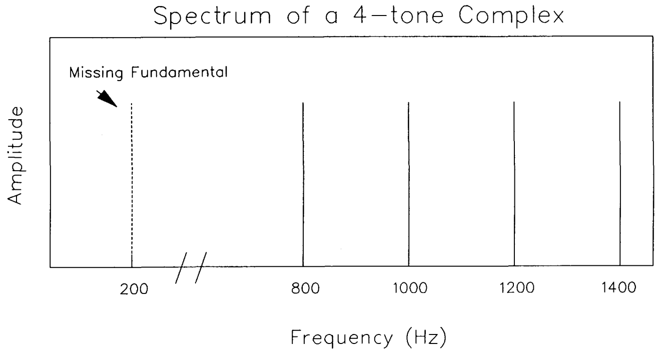

FIG. 24. A schematic drawing showing the amplitude spectrum of a complex sound consisting of the 4th, 5th, 6th, and 7th harmonics of 200 Hz. Despite the fact that 200 Hz is not in the stimulus, listeners assign a pitch of 200 Hz to this complex. The perceived pitch is at the "missing fundamental."

Table 5. Some of the audible pitches associated with the nonlinearity of the auditory system when two tones (f_1 and f_2) are presented. For the nonlinearity pitch column, $f_1 = 1000$ Hz and $f_2 = 1800$ Hz. In most situations, the cubic-difference tone is the easiest nonlinear pitch to detect.

Symbol	Nonlinearity pitch (Hz)	Label
$2f_1$	2000	Aural harmonic
$2f_2$	3600	Aural harmonic
$3f_1$	3000	Second aural harmonic
$3f_2$	5400	Second aural harmonic
f_2-f_1	800	Difference tone
$2f_1$-f_2	200	Cubic difference tone

(even though 2000 Hz is still in the stimulus) and there is no beating sensation. The fact that nonlinear distortion products, such as the 2000-Hz aural harmonic in the example above, can be manipulated as if the distortion products were actually in the stimulus, indicates that the nonlinearities occur early in the process of sound transduction. No simple description of the nonlinearity has been developed which can fully account for the estimates of the frequencies, levels, and phases of all the various distortion products (see Moore, 1982b).

The stimulus depicted in Fig. 24 indicates another stimulus condition for which listeners perceive a pitch when there are no frequency components in the stimulus corresponding to the reported pitch. The four-tone, complex stimulus consists of the 4th through 7th harmonics of 200 Hz. For this complex the listener reports a salient 200-Hz pitch; that is, the pitch of the sound is at the missing fundamental frequency of the complex. Listeners will match the pitch of this complex to that of a 200-Hz tone when the phases of the complex are randomized. Randomizing the phases of the four tones results in a time wave form with no apparent periodicity between the wave form peaks. Thus this "missing fundamental pitch" is not directly dependent on energy in the spectrum at 200 Hz or to any fundamental periodicity of 200 Hz between wave form peaks.

Figure 25 depicts another intriguing aspect of these complex pitch stimuli. For this condition listeners were asked to match the pitch of a tone to that of the four-tone complex when all the tones in the complex were shifted in frequency (e.g., 50 Hz as shown in Fig. 25). For instance, if the complex consists of tones at 750, 950, 1150, and 1350 Hz, the matched pitch is approximately 190 Hz. The frequency spacing is still 200 Hz, but the fundamental frequency is now 50 Hz, not 200 Hz. The pitch of this complex is neither 50 nor 200 Hz, but 190 Hz. A number of different models have been proposed to account for how the nervous system extracts the pitch of these complex stimuli. It is clear that mechanisms in addition to those that form the code for sound frequency are required to account for the pitch of many complex sounds such as the "missing fundamental pitch" described above (see Moore, 1982).

8.3 Other Subjective Sound Attributes

A number of subjective attributes have been measured with matching procedures similar to that used to obtain the equal loudness contours (Fig. 22). Measurements of sound volume, density, and annoyance are some of these attributes. Measures of volume and loudness have been used to calibrate sound level controls for radios and other audio and entertainment equipment. Annoyance measurements are used in dealing with

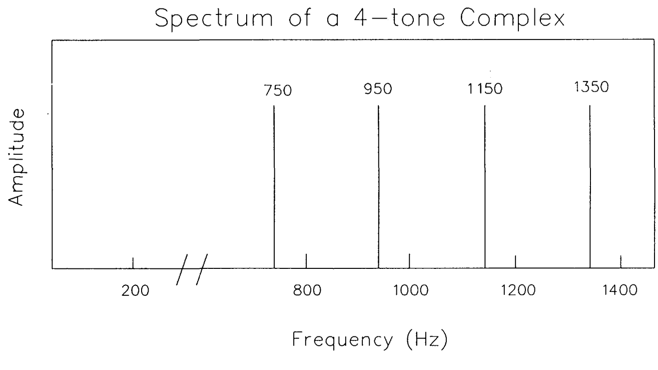

FIG. 25. This schematic amplitude spectrum is similar to that shown in Fig. 23, except the frequencies of the four tones have been shifted down 50 Hz relative to their frequencies in Fig. 23. The matched pitch is approximately 190 Hz for this complex, which is not equal to the missing fundamental (50 Hz).

indices of environmental noise pollution, especially in regard to airport noise (see Kryter, 1970).

The timbre of a sound is usually related to its spectral complexity, especially in the harmonic structure of periodic sounds. Thus, the difference in sound quality between a piano and a guitar playing the same pitch is referred to as a timbre difference. Each instrument will produce the same fundamental period of vibration if each produces the same pitch, but the amplitudes of the harmonics of this fundamental will differ between the two instruments resulting in the perceived timbre differences. No convenient metric has been developed to describe timbre.

Rhythm has been studied extensively in music and the musical notation of rhythm (i.e., specifying the relative durations of notes in terms of multiples of $\frac{1}{2}$, e.g., eighth notes, quarter notes) has proved to be among the most useful metrics for describing the subjective sense of rhythm (see Risset, 1978).

9. HEARING IMPAIRMENT

The data described so far relate to sound processing by the normal auditory system. For millions of people (approximately 20 million in the United States) auditory system disorders cause difficulties for functioning in everyday life. The primary means of determining the nature of a hearing disorder is by using psychoacoustic tasks similar to those described in this chapter.

9.1 Audiology and Otology

Two specialty areas deal with the diagnosis and treatment of hearing disorders: an otorhinolaryngologist (usually shortened to otolaryngologist or otologist and sometimes referred as ear, nose, throat or ENT specialist) is a surgical specialist of medicine, and an audiologist (or clinical audiologist) receives certification based on a master's degree and criteria set by the American Speech, Hearing, and Language Association, or in some special circumstances by state agencies. An audiologist often administers the hearing tests used for diagnosis and is usually responsible for rehabilitation following surgery or other corrective procedures such as the fitting of a hearing aid. An otolaryngologist is responsible for surgery or other medical treatments. For many patients, the two specialists form a team for treating hearing disorders.

9.2 Tests for Hearing Disorders

The neural site of a hearing disorder can be in the conductive pathways of the outer and middle ear, in the neural transduction mechanisms of the inner ear, or in the neural processes of the central nervous system. Many of the tests for detecting hearing disorders are designed to implicate one of these three sites as the location of the disorder.

The structures of the inner ear that signal the neural response to sound can be excited in two ways. The normal means is via the vibration of the middle ear bones by air-driven vibration of the eardrum, but the fluids and structures of the inner ear can be vibrated directly by the skull bone if the amplitude of bone vibration is great enough. Standard thresholds for both air-conducted (the thresholds shown in Fig. 2) and bone-conducted sound (by placing a vibrator on the mastoid or forehead) have been established (see Dirks, 1978). If a person has normal bone-conducted thresholds but abnormal air-conducted thresholds, it is assumed that the site of the impairment is in the conductive pathways of the outer and middle ears, most likely the middle ear bones. If both air-conducted and bone-conducted thresholds are abnormal, then the presumed site of the impairment is in the inner ear or the central nervous system.

This so-called "air-bone gap" has long been a standard hearing test. Another technique, frequently used to determine the integrity of the middle ear, is based on forming an acoustic impedance bridge, by reflecting sound off the eardrum and measuring the amount of reflected sound pressure as a function of changing the barometric pressure within the enclosed outer ear. The eardrum and bones of the middle ear are assumed to provide resistance and impedance to the reflected sound. Abnormal changes in the magnitude of the reflected sound with changes in the barometric pressure are interpreted as due to abnormal resistance or impedance changes, which can in turn implicate a middle ear structure as the cause of the abnormality. "Impedance audiometry" is a common test for deficiencies

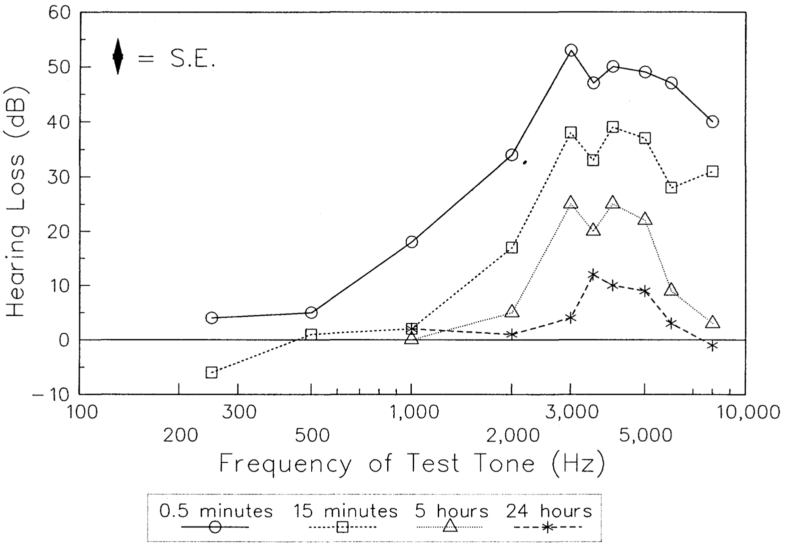

FIG. 26. The amount of hearing loss (in dB) for different tones as a function of the tonal frequency (in Hz) after the listener was presented a 115-dB SPL broadband noise for 20 min. Data are shown for four time intervals following the noise exposure (0.5 min, 15 min, 5 h, and 24 h). The amount of hearing loss (temporary threshold shift) was obtained by taking the difference between thresholds measured prior to the noise exposure and those measured at the various intervals after the noise exposure. The greatest temporary hearing loss is found in the region of 4000 Hz, and thresholds return to near normal 24 h after the exposure. The arrows labeled S.E. are an estimate of the standard error of measurement for these conditions and as such represent the expected variability about any data point. Adapted from the data of Postman and Egan (1949) as shown in Fig. 14.6 of Yost and Nielsen (1985).

in the conductive pathways for hearing (see Durrant and Lovrinic, 1977).

Tests for the neural sites of a hearing loss are more difficult to interpret. Psychophysical tests, often based on speech or binaural tasks (because binaural hearing and speech perception are assumed to depend on neural structures in the brain), are sometimes used to determine if a hearing disorder is due to abnormalities in the central nervous system. A variety of brain imaging techniques (e.g., measures of sound evoked electroencephalograms or PET) are also used if tumors or other such abnormalities are suspected as the cause of the hearing loss.

9.3 Hearing Loss and Noise Exposure

Hearing loss has many causes. The biological process of aging results in a progressive loss of hearing referred to as "presbycusis." Aging reduces sensitivity first to the highest frequencies, and then slowly, over time, the hearing loss progresses to lower and lower frequencies. Hearing abnormalities are due to genetic factors, diseases, infections, ototoxic drugs used in the treatment of some diseases and infections, damage to the nerves of the auditory system, or environmental factors, most notably exposure to intense sound.

Exposure to intense sound can produce a temporary hearing loss or a permanent hearing loss if the sound is very intense or the exposure is prolonged. Figure 26 shows data indicating the shift in thresholds of hearing following exposure to a 115-dB SPL noise presented for 20 min. The greatest loss occurs immediately after the noise is turned off and in the frequency region near 4000 Hz. Most temporary loss of hearing recovers to normal after a few hours following the exposure. However, repeated exposure to the intense sound may slow the recovery to the point that the loss is permanent. Table 6 shows the sound levels (expressed as dBA) and durations of exposure that will probably lead to a permanent loss of hearing if the exposure conditions are repeated. These "damage risk criteria" have been used in a number of laws and policies to limit the sound exposure that people can experience (see Miller, 1978).

9.4 Hearing Protection and Hearing Aids

In many job situations it is impossible to eliminate harmful exposure to intense sound levels. In these situations hearing protectors are used as a last resort for preventing a hearing loss. Most hearing protectors are passive, made of sound-attenuating material, and they fit into or over the outer ear. Some active noise-cancelling devices are also used to reduce noise, especially when speech communication is important (e.g., in the cockpit of a noisy airplane). These devices operate on a number of different principles, but generally attempt to cancel the noise background with an out-of-phase noise or to adaptively filter the unwanted noise, thus increasing the signal- to-noise ratio for the desired communication signal (see Lim, 1983).

Four general types of hearing aids or prosthetic devices are available to help a hearing-impaired listener: amplification aids, assisted listening devices, tactile aids, and cochlear prostheses. None of these aids restores hearing to a normal level, but each has its advantages and disadvantages (see Lutman and Haggard, 1983).

Amplification aids, made as small as possible for cosmetic purposes, act to amplify sound. They may amplify sound in particular frequency regions, limit the total amount of amplification to prevent discomfort by the listener, or provide some form of background-noise cancellation. Simply amplifying sound leads to two problems: (1) amplification causes all sounds to be increased in level, not just the signal the listener may want to hear, and (2) sounds can be increased only to a certain level before the listener experiences discomfort and the auditory system is jeopardized. Figure 27 shows the classic problem confronting an amplification aid. The abnormal increase in loudness for the hearing-impaired listener, sometimes called "recruitment," requires that the total output of the aid be limited. Many forms of compression are used to limit the sound level at high amplification. However, compression is a nonlinear operation that can distort the original sound. Many tricks of the trade are used to apply compression and reduce the distorting effects of the nonlinearity. Amplification aids are based on either analog or digital signal processes. Many digital amplification aids include a form of active noise cancellation, which often leads to a distorted signal requiring special circuitry to realize the noise cancellation while reducing distortion. The aid's small size and the close proximity of the microphone to the speaker also cause problems in providing a clear signal to the hearing-impaired listener. The battery life for aids is often very short, especially for digital aids that require a great deal of signal processing. Finally, properly fitting the aid to the ear of the individual listener is crucial both for comfort and to achieve the proper acoustic load for the hearing-aid speaker.

Table 6. Permissible noise exposure: sound levels (in dBA) and durations of exposures (in hours/day) that should not be exceeded. Repeated exposure to levels and durations greater than those of these damage-risk criteria could cause a measurable permanent hearing loss.

Duration (hours/day)	Sound level (dBA)
8	90
6	92
4	95
3	97
2	100
1.5	102
1	105
0.5	110
0.25 or less	115

Assisted listening devices enable sound to be fed directly to the listener. This can be accomplished over a cable, via FM broadcast, or by LED (light-emitting diode technology) transmission. These devices are found in classrooms, at theaters and churches, and in the home (e.g., connected to the television or as part of a telephone). Their main advantage is in magnifying the signal-to-noise ratio by carrying the signal directly to the listener with minimal amplification of the background noise. However, the talker as well as the listener must be fitted with a device.

Tactile aids provide vibration to the skin to translate various properties of sound to some part of the listener's skin surface (usually, the fingertip, forearm, back, or stomach). The devices usually divide sound into spectral bands through some form of bandpass filtering. The amount of energy in the various frequency bands may be coded according to different locations on the skin or by the rate or pattern of

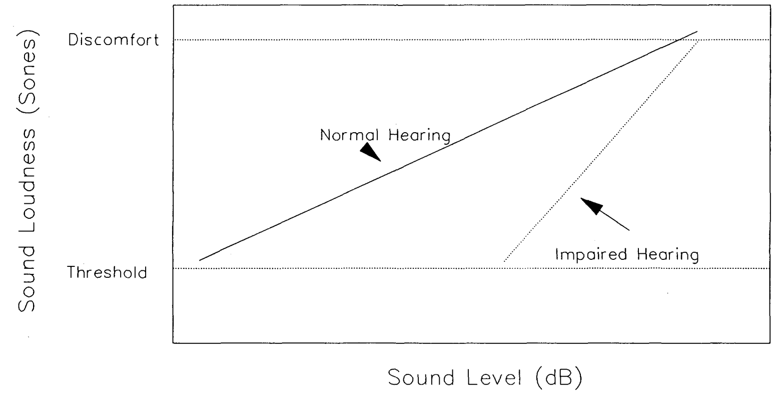

FIG. 27. A schematic loudness function represents the perceived change in loudness as a function of an increase in sound level (see also Fig. 23) for two types of listeners: a normal-hearing listener, and one with a hearing impairment. Both listeners judge very intense sounds to be approximately the same in loudness, especially near the sound level that is uncomfortable. However, because the hearing impaired listener has a hearing loss, the growth of his/her loudness function is steeper than for the normal hearing listener. Thus, the hearing impaired listener has a reduced dynamic range in which an amplification aid can operate in order to boost the level of sound without causing discomfort.

vibration. Tactile aids can assist the listener in lip reading and general sound awareness. For some users simple forms of speech can be transmitted through the skin, but usually only after considerable practice.

Cochlear prostheses are devices that deliver an electrical current directly to the nerves of the inner ear. Implanting the prosthesis is major microsurgery and is recommended only for patients with severe hearing loss. The aid consists of two main parts: a transducer which processes and translates sound into an electrical current, and an electrode(s) which delivers the current directly into the inner ear. The electrodes presumably stimulate neural tissue in the inner ear, triggering neural responses that mimic those that occur normally in neural sound transduction. For multiple-electrode devices, the transducer splits the sound into different frequency bands and delivers current to each electrode in proportion to the energy in each band. The electrodes are placed into different locations in the inner ear, such that those carrying information about the high frequencies are nearer the middle ear (the base of the inner ear) than those carrying low-frequency information (which are near the apex of the inner ear). In the normal inner ear, nerves near the base carry information to the brain about high frequencies while those near the apex carry lower-frequency information. Thus, the multiple-electrode devices attempt to mirror this form of neural coding. The success of cochlear prostheses is varied. Some users experience a range of sensations not dissimilar to those experienced by normal hearing listeners. However, the majority of the successful implants provide sensations useful for general sound awareness and aid in tasks such as lip reading (see Schindler and Merzenich, 1985).

GLOSSARY

All definitions are based on the American National Standards Institute's (ANSI) S3.20-1973(R 1986) Standard Terminology for Psychoacoustics. Terms not defined in the ANSI standard (and, therefore, not in this Glossary) but appearing in this chapter are defined within the chapter.

Aural Harmonic: An aural harmonic is a harmonic generated in the auditory mechanism.

Binaural Masking Level Difference (BMLD): Masking level difference is any decrease (improvement) in the masked threshold obtained when two ears are used instead of one. Masking level difference is expressed in decibels, as a function of the frequency of the masked tone.

Bone Conduction: Bone conduction is the process by which sound is conducted to the internal ear through the cranial bones.

Combination Tone: A combination tone is a (secondary) tone that can be perceived when two loud (primary) tones are presented simultaneously. The secondary tone may be a difference tone or a summation tone. Several combination tones may be produced from a single pair of primary tones. Combination tones can also be generated by aural harmonics. Combination tones are due to the nonlinear characteristics of the ear.

Coupler (6 cm^3): A coupler is a device for acoustic loading of earphones. It has a specified arrangement of acoustic elements, and is provided with a microphone for measurement of the sound pressure developed in a specified portion of the device.

Critical Band: The critical band for masking is that frequency band of sound, being a portion of a continuous-spectrum noise covering a wide band, that contains sound power equal to that of a pure tone centered in the critical band and just audible in the presence of the wide-band noise.

Dichotic: Dichotic refers to the condition in which the sound stimulus presented at one ear differs from the sound stimulus presented at the other ear.

Difference Tone: A difference tone is a combination tone with a frequency equal to the difference between the frequencies of two primary tones or of their harmonics.

Diotic: Diotic refers to the condition in which the sound stimuli presented at the two ears are identical.

Equal Loudness Contour: A loudness contour is a curve that shows the related values of sound pressure and frequency required to produce a given loudness sensation for the typical subject for a stated manner of listening to the sound.

Forward Masking: Forward masking describes the condition in which the masking sound appears before the masked sound.

Lateralization: Auditory lateralization is the determination by a subject that the apparent direction of a sound is either left or right of the frontal-medial plane within the head. Auditory lateralization is usually restricted to sounds presented by earphones or by bone-conduction vibrators.

Localization: Auditory localization is the determination by a subject of the apparent direction and/or distance of a sound source.

Loudness: Loudness is that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from soft to loud. Loudness depends primarily upon the sound pressure of the stimulus, but it also depends upon the frequency and wave form of the stimulus.

Masking: Masking is the process by which the threshold of audibility for one sound is raised by the presence of another (masking) sound. Masking is the amount by which the threshold of audibility of a sound is raised by the presence of another (masking) sound. The unit customarily used is the decibel.

Method of Magnitude Estimation: The method of magnitude estimation is a psychophysical method used primarily to scale sensations; in this procedure the subject assigns to a set of stimuli numbers that are proportional to some subjective dimension of the stimuli.

Minimal Audible Pressure (MAP): The minimum audible pressure is the sound pressure of a tone at the threshold of audibility that is presented by an earphone and measured or inferred at the tympanic membrane.

Minimal Audible Field (MAF): The minimum audible field is the sound pressure level of a tone at the threshold of audibility measured in a free sound field for a subject facing the sound source. The value of the sound pressure level is determined after the subject is removed from the field.

Monotic: Monotic refers to the condition in which a sound stimulus is presented at only one ear.

Phon: The (judged) loudness level of a sound, in phons, is numerically equal to the median sound pressure level, in decibels relative to 20 micropascals, of a 1000-Hz reference tone presented to subjects as a frontally incident free sound field, and judged by the subjects in a number of trials to be as loud as the sound under test.

Pitch: Pitch is that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high. Pitch depends primarily upon the frequency of the sound stimulus, but it also depends upon the sound pressure and wave form of the stimulus. The pitch of a sound may be described by the frequency of that pure tone that is judged by subjects to produce the same pitch.

Precedence Effect: The precedence effect is the phenomenon that occurs during auditory fusion when two (primary) sounds of the same order of magnitude are presented dichotically and produce localization of the secondary sound toward the ear receiving the first (primary) sound stimulus. The precedence effect occurs for stimulus time delays varying from fractions of a millisecond to the upper time limit for auditory fusion, beyond which separate and independent sounds are heard.

Psychoacoustics: Psychoacoustics is the science that deals with the psychological correlates of the physical parameters of acoustics. Psychoacoustics is a branch of psychophysics.

Psychometric Function: A psychometric function is a mathematical relationship in which the independent variable is a measure of a stimulus and the dependent variable is a measure of response.

Psychophysics: Psychophysics is the science that deals with the quantitative relationship between physical and psychological events. Psychophysics is a branch of psychology.

Recruitment: Recruitment is the otological condition in which weak sounds are not heard while strong sounds are heard as loudly as by a normal ear. The function that relates loudness to sound pressure level is steeper than normal and is often nonlinear. In recruitment, the dynamic range of hearing is narrowed.

Sensation Level (SL): The sensation level of a sound is the pressure level of the sound in decibels above its threshold of audibility for the individual subject or for a specified group of subjects.

Sensory-Neural Hearing Loss: Sensory-neural hearing loss is the impairment of hearing that occurs when there is damage to the cochlea or the cochlear nerve.

Sone: The sone is a unit of loudness. One sone is the loudness of a sound for which the loudness level is 40 phons. The loudness of a sound that is judged by a subject to be *n* times that of a 1-sone tone is *n* sones.

Sound Pressure Level (SPL): Sound pressure level, in decibels, of a sound is 20 times the logarithm to the base 10 of the ratio of the pressure of this sound to the reference pressure of 20 micropascals.

Temporary Threshold Shift (TTS): Temporary threshold shift is a temporary increase in the threshold of audibility for an ear at a specified frequency. The amount of temporary threshold shift is customarily expressed in decibels.

Threshold of Audibility: The threshold of audibility for a specified signal is the minimum effective sound pressure level of the signal that is capable of evoking an auditory sensation in a specified fraction of the trials. The characteristics of the signal, the manner in which it is presented to the subject, and the point at which the sound pressure level is measured must be specified. Unless otherwise indicated, the ambient noise reaching the ears is assumed to be negligible. The threshold of audibility is usually given as a sound pressure level in decibels, relative to 20 micropascals.

Timbre: Timbre is that attribute of auditory sensation in terms of which a subject can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar. Timbre depends primarily upon the spectrum of the stimulus, but it also depends upon the wave form, the sound pressure, the frequency location of the spectrum, and the temporal characteristics of the stimulus.

List of Works Cited

Abel, S. M. (1972), *J. Acoust. Soc. Am.* **51**, 1219–1224.

Blauert, J. (1983), *Spatial Hearing: The Psychophysics of Human Sound Localization*, Cambridge, MA: MIT Press.

Carterette, E. C., Friedman, M. P. (Eds.) (1974a), *Handbook of Perception*, Vol. 1, *Historical and Philosophical Roots of Perception*, New York: Academic.

Carterette, E. C., Friedman, M. P. (Eds.) (1974b), *Handbook of Perception*, Vol. 6, *Hearing*, New York: Academic.

Colburn, H. S., Durlach, N. I. (1978), in: E. C. Carterette, M. P. Friedman (Eds.), *Handbook of Perception*, Vol. 4. New York: Academic.

Dirks, D. D. (1978), in: E. C. Carterette, M. P. Friedman (Eds.), *Handbook of Perception*, Vol. 4. New York: Academic, Chap. 13.

Durlach, N. I., Colburn, H. S. (1978), in: E. C. Carterette, M. P. Friedman (Eds.), *Handbook of Perception*, Vol. 4. New York: Academic.

Durrant, J. D., Lovrinic, J. H. (1977), *Bases of*

Hearing Science, Baltimore: Williams and Wilkins.

Evans, E. F., Wilson, J. P. (Eds.) (1977), *Psychophysics and Physiology of Hearing*, London: Academic.

Green, D. M. (1971), *Psychol. Rev.* **78** (6), 542–551.

Green, D. M. (1976), *An Introduction to Hearing*, Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Green, D. M. (1988), *Profile Analysis: Auditory Intensity Discrimination*, New York: Oxford University Press.

Green, D. M., Swets, J. A. (1977), *Signal Detection Theory and Psychophysics*, Huntington, New York: Krieger.

Gulick, W. L. (1971), *Hearing: Physiology and Psychophysics*, New York: Oxford University Press.

Henning, G. B. (1977), *J. Acoust. Soc. Am.* **55**, 84–90.

Jesteadt, W., Wier, C. C., Green, D. M. (1977), *J. Acoust. Soc. Am.* **61**, 169–177.

Kryter, K. D. (1970), *The Effects of Noise on Man*, New York: Academic.

Kuhn, G. F. (1987), in: W. A. Yost, G. Gourevitch (Eds.), *Directional Hearing*, New York: Springer.

Laming, D. (1986), *Sensory Analysis*, London: Academic.

Lim, J. S., (Ed.) (1983), *Speech Enhancement*, Englewood Cliffs, New Jersey: Prentice-Hall.

Lutman, M. E., Haggard, M. P., (Eds.) (1983), *Hearing Science and Hearing Disorders*, London: Academic.

Miller, J. D. (1978), in: E. C. Carterette, M. P. Friedman (Eds.), *Handbook of Perception*, Vol. 4. New York: Academic.

Mills, A. W. (1956), *J. Acoust. Soc. Am.* **30**, 237–246.

Moore, B. C. J. (1982a), *J. Acoust. Soc. Am.* **63**, 524–532.

Moore, B. C. J. (1982b), *An Introduction to the Psychology of Hearing*, 2nd ed., London: Academic.

Moore, B. C. J. (1986), *Frequency Selectivity in Hearing*, London: Academic.

Moore, B. C. J., Glasberg, B. R., Roberts, B. (1984), *J. Acoust. Soc. Am.* **76**, 1057–1066.

Moore, B. C. J., O'Loughlin, B. J. (1986), in: B. C. J. Moore (Ed.), *Frequency Selectivity in Hearing*, London: Academic.

Patterson, R. D., Moore, B. C. J. (1986), in: B. C. J. Moore (Ed.), *Frequency Selectivity in Hearing*, London: Academic.

Postman, L. J., Egan, J. P. (1949), *Experimental Psychology: An Introduction*, New York: Harper & Row.

Reed, C. M., Bilger, R. C. (1973), *J. Acoust. Soc. Am.* **53**, 1039–1045.

Risset, J. (1978), in: E. C. Carterette, M. P. Friedman (Eds.), *Handbook of Perception*, Vol. 4. New York: Academic.

Robinson, D. E., Watson, C. S. (1972), in: J. V. Tobias (Eds.), *Foundations of Modern Auditory Theory*, Vol 2. New York: Academic.

Schindler, R. A., Merzenich, M. M. (Eds.) (1985), *Cochlear Implants*, New York: Raven.

Schubert, E. D. (1974), in: E. C. Carterette, M. P. Friedman (Eds.), *Handbook of Perception*, Vol. 4. New York: Academic.

Sivan, L. J., White, S. D. (1933), *J. Acoust. Soc. Am.* **4**, 288–321.

Stevens, S. S. (1975), in G. Stevens (Ed.), *Psychophysics*, New York: Wiley.

Tartini, G. (1754), *Trattato di musica secondo la vera scienza dell' armonia*, Padua: Stamperia del Seminario, pp. 10–19.

Tobias, J. V. (Ed.) (1972), *Foundation of Modern Auditory Theory*, Vol. 2. New York: Academic.

Viemeister, N. F. (1977), in: E. F. Evans, J. P. Wilson (Eds.), *Psychophysics and Physiology of Hearing*, London: Academic.

Watson, C. S. (1974), in: B. Wohlman (Ed.), *Handbook of Psychology*, Englewood Cliffs, New Jersey: Prentice-Hall.

Watson, C. S., Gengel, R. W. (1969), *J. Acoust. Soc. Am.* **46**, 989–997.

Weir, C. C., Jesteadt, W., Green, D. M. (1977), *J. Acoust. Soc. Am.* **61**, 178–183.

Wightman, F. L., Kistler, D. J., Perkins, M. E. (1987), in: W. A. Yost, G. Gourevitch (Eds.), *Directional Hearing*, New York: Springer.

Yost, W. A. (1974), *J. Acoust. Soc. Am.* **55**, 1299–1303.

Yost, W. A., Dye, R. H. (1987), *J. Acoust. Soc. Am.* **83**, 1846–1851.

Yost, W. A., Gourevitch, G. (Eds.) (1987), *Directional Hearing*, New York: Springer.

Yost, W. A., Hafter, E. R. (1987), in: W. A. Yost, G. Gourevitch (Eds.), *Directional Hearing*, New York: Springer.

Yost, W. A., Nielsen, D. W. (1985), *Fundamentals of Hearing: An Introduction*, 2nd ed., New York: Holt, Rinehart and Winston.

Yost, W. A., Watson, C. S. (Eds.) (1987), *Auditory Processing of Complex Sounds*, Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Zurek, P. M. (1987), in: W. A. Yost, G. Gourevitch (Eds.), *Directional Hearing*, New York: Springer.

Further Reading

Green, D. M. (1976), *An Introduction to Hearing*, Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Moore, B. C. J. (1982), *An Introduction to the Psychology of Hearing*, 2nd ed., London: Academic.

Yost, W. A., Nielsen, D. W. (1985), *Fundamentals of Hearing: An Introduction*, 2nd ed., New York: Holt, Rinehart and Winston.

ACOUSTICS, UNDERWATER

WILLIAM A. KUPERMAN, *Naval Research Laboratory*

Introduction 293
1. **The Ocean Environment** 295
2. **Sound Propagation Paths in the Ocean** 296
2.1 Volume Attenuation 297
2.2 Bottom Loss 298
2.3 Surface, Bottom, and Volume Scattering 301
2.4 Ambient Noise 302
3. **Sound Propagation Models** 305
3.1 The Wave Equation and Boundary Conditions 305
3.2 Ray Theory 306
3.3 Full Field Solution or "Fast Field Program" (FFP) 307
3.4 Normal Mode Model (NM) 308
3.5 Adiabatic Mode Theory 309
3.6 Parabolic Equation Model (PE) 309
3.7 Model Consistency 310
4. **Spatial Properties of Surface Generated Noise** 313
5. **Arrays and Array Gain for Signals in Noise** 317
6. **Signal Processing and Sonar Systems** 320
6.1 Passive Sonar Equation 320
6.2 Active Sonar Equation 322
7. **Underwater Acoustics and Remote Sensing** 322
Glossary 324
Works Cited 326
Further Reading 327

INTRODUCTION

Underwater acoustics is the science of sound in water (most commonly in the ocean) and encompasses not only the study of sound propagation, but also the masking of sound by interfering acoustic phenomena. Its origins date back centuries. For example, Leonardo da Vinci noted that sound travels to great distances in the sea and referred to interfering phenomena (in this case, self-noise) by suggesting the listener "...cause your ship to stop..." (Urick, 1983) before deploying a listening tube through the air-water interface. This listening tube was improved upon by Colladen and Sturm (Lasky, 1977) who, on Lake Geneva in 1827, employed a trumpetlike device as an underwater sound receiver together with a submerged bell ten miles away. By essentially a time of flight measurement, they determined the speed of sound to an accuracy of about two-tenths of one percent. In addition, "Colladen marvelled that so little energy at the source could be transmitted so great a distance through the water medium and could still be detected by the trumpet receiver." It is still believed that only sound will propagate signals to great distances in the sea.

The development of acoustics theory has followed a different historical path than optics and electromagnetic theory (Pierce, 1981). While there has been a long-running battle in optics between ray and wave theory, acoustics originally developed with wave and vibration concepts as far back as ancient Greek times. For example, musical sound is the result of air motion generated by a vibrating musical instrument. Thus sound is propagated in an analogous manner to water surface waves which propagate disturbances (signals), but not matter, long distances. The bending of sound around corners is then easily understood in the context of diffraction in the water surface wave analogy. In acoustics, the ray picture evolved as a mathematical approximation to wave theory. In underwater acoustics, in which the medium is inhomogeneous and bounded by complicated interfaces, ray theory has historically proved to be an indispensable tool for understanding and studying sound propagation. However, with the advent of computers and the recent developments of numerical physics, sound propagation in the ocean can now be studied and

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

quantitatively described in greater detail with the more exact wave theory. This latter trend is particularly important to the study of extracting signals from complex interfering fields in which quantitative understanding of the amplitude, phase, and higher order moments of all the relevant fields is necessary.

Research in the field of experimental underwater acoustics requires extensive equipment, with at least one ship and often an assortment of at-sea platforms equipped with sound projectors, receiving arrays, and sensors to measure the ocean environment. The requirements to measure the latter arise because the ocean is the medium which defines the acoustic index of refraction, and the ocean surface and bottom (and layers below) are the interfaces that define the boundary conditions of the partial differential equations descriptive of sound propagation in the ocean. Very often, the oceanographic and/or ocean geology measurements associated with an acoustic experiment rival or exceed the complexity of the acoustic measurements.

In this article the ocean environment will be introduced and described prior to the acoustics. Then we present underwater acoustics with minimum mathematics in a detailed, though descriptive, manner, associating various acoustic topics with the details of the previously discussed ocean environment. After completing the descriptive underwater acoustics part, more details are given on the theory of underwater acoustics and the models that have been developed to compute ocean acoustic phenomena. Since underwater acoustics cannot be divorced from the measurements of signals in complex media, some relevant array and signal processing ideas will also be reviewed together with a discussion of sonars and the sonar equation. We end by discussing the use of sonar technology for remote sensing of the ocean. We start by mentioning some relevant units and a brief discussion on underwater acoustic transducers.

Some Relevant Units

The decibel (dB) is the dominant unit in underwater acoustics and denotes a ratio of intensities (not pressures) expressed in terms of a logarithmic (base 10) scale. Two intensities, I_1 and I_2, have a ratio I_1/I_2 in decibels of 10 log I_1/I_2 dB. Absolute intensities can therefore be expressed by using a reference intensity. The presently accepted reference intensity is a micropascal (μPa): the intensity of a plane wave having an rms pressure equal to 10^{-5} dyn/cm^2. Therefore, taking 1 μPa as I_2, a sound wave having an intensity, of, say, one million times that of a plane wave of rms pressure 1 μPa has a level of $10 \log(10^6/1) \equiv 60$ dB re 1 μPa. Pressure (p) ratios are expressed in dB re 1 μPa by taking $20 \log p_1/p_2$, where it is understood that the reference originates from the intensity of a plane wave of pressure equal to 1 μPa.

The average intensity I of a plane wave with rms pressure p in a medium of density ρ and sound speed c is $I=p^2/\rho c$. In seawater, ρc is 1.5×10^5 g cm^{-2} s^{-1} so that a plane wave of rms pressure 1 dyn/cm^2 has an intensity of 0.67×10^{-12} W/cm^2. Substituting the value of a micropascal for the rms pressure in the plane wave intensity expression, we find that a plane wave pressure of 1 μPa corresponds to an intensity of 0.67×10^{-22} W/cm^2 (i.e., 0 dB re 1 μPa).

Acoustical Instrumentation

(See ACOUSTICAL INSTRUMENTATION in this volume.) A transducer converts some sort of energy to sound (source) or converts sound to energy (receiver), usually electric (Wilson, 1985). In underwater acoustics, piezoelectric and magnetostrictive transducers are commonly used. The piezoelectric effect, which occurs in certain crystals, such as quartz, involves electric polarization produced by mechanical strain and vice versa. There is also an electrostrictive effect which is common to all dielectrics but smaller than the piezoelectric effect: an electric field mechanically deforms the dielectric, independently of the direction of the electric field. The magnetostrictive effect is the change in dimensions of a ferromagnetic material when it is placed in a magnetic field and the change in magnetization when the material's dimensions change from an external force.

There are also electrodynamic transducers. For example, a receiver element operates by sound pressure oscillations moving a current-carrying conductor through a magnetic field causing a back emf. Electrostatic transducers operate on the principle that charged electrodes moving in a sound field change separation distance, thereby changing the capacitance of the system.

Explosive and airgun sources are high energy wideband types of sources. The technology is such that their signatures are fairly reproducible (Urick, 1983). Electric discharge and laser sources are also in use.

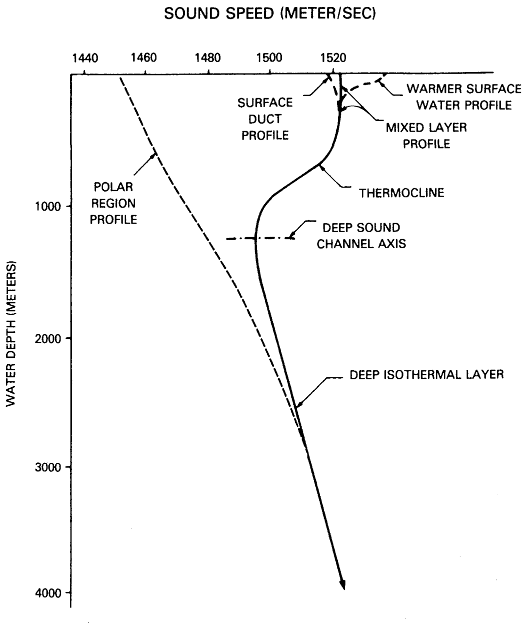

FIG. 1. Generic sound speed profiles.

1. THE OCEAN ENVIRONMENT

In underwater acoustics, the ocean is a waveguide and the speed of sound (c) plays the same role as the index of refraction does in optics. Sound speed is normally related to density and compressibility. In the ocean, density is related to static pressure (which increases with depth), salinity, and temperature (which varies significantly in the upper mixing layer). The sound speed in the ocean is an increasing function of temperature, salinity, and pressure. It is customary to express sound speed as an empirical function of three independent variables: temperature (T) in degrees centigrade, salinity (S) in parts per thousand, and depth (z) in meters. A simplified expression for this dependence is (Clay and Medwin, 1977)

$$c=1449.2+4.6T-0.055T^2+0.00029T^3 \\ +(1.34-0.10T)(S-35)+0.016z. \qquad (1)$$

For most cases, the above equation is sufficiently accurate; however, occasionally more accuracy may be required when distances are derived from travel time measurements.

Seasonal and diurnal changes affect the oceanographic parameters in the upper ocean. In addition, all of these parameters are functions of geography. Figure 1 illustrates a typical set of sound speed profiles indicating greatest variability near the surface as a function of season and time of day. In a warmer season (or warmer part of the day), the temperature increases near the surface and hence the sound speed decreases with depth. In nonpolar regions, the oceanographic properties of the water near the surface result from mixing activity originating from the air-sea interface. This near-surface mixed layer has a constant temperature (except in calm, warm surface conditions as described above). Hence, in this isothermal mixed layer, the sound speed profile can increase with depth due to the pressure gradient effect of the last term in Eq. (1). This is the "surface duct" region, and its existence depends on the near-surface oceanographic conditions. Note, the more agitated the upper layer, the deeper the mixed layer and the less likely there will be any departure from the mixed-layer part of the profile depicted in Fig. 1. Hence, an atmospheric storm passing over a region where a surface duct exists is likely to mix the surface waters and cause the duct to change.

Below the mixed layer is the thermocline in which the temperature decreases with depth and the sound speed decreases with depth. Below the thermocline, the temperature is constant (about 4 °C—a thermodynamic property of salt water at high pressure) and the sound speed increases because of increasing pressure. Therefore, a minimum sound speed which is often referred to as the axis of the deep sound channel exists between the deep isothermal region and the mixed layer. However, in polar regions, the water is coldest near the surface and hence the minimum sound speed is at the ocean/air (or ice) interface as indicated in Fig. 1. In continental shelf regions (shallow water) with water depths on the order of a few hundred meters, only the upper region of the sound speed profile in Fig. 1, which is dependent on season

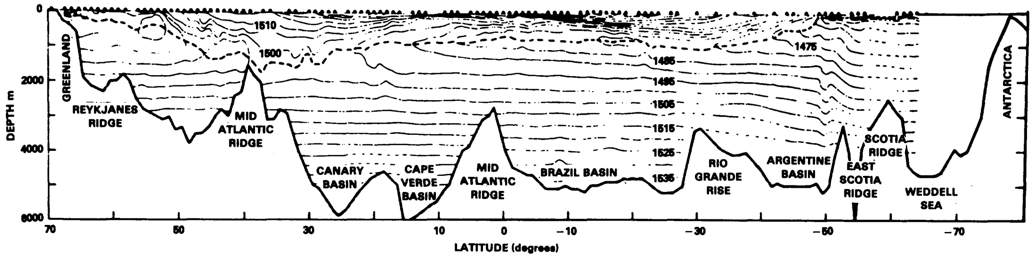

FIG. 2. Sound speed contours at 5 m/s intervals taken from North and South Atlantic along 30.50° W. Dashed line indicates axis of deep sound channel (from Northrup and Colburn, 1974).

and time of day, affects sound propagation in the water column.

Figure 2 is a contour display of the sound speed structure of the North and South Atlantic (Northrup and Colburn, 1974) with the deep sound channel indicated by the heavy dashed line. Note the geographic (and climatic) variability of the upper ocean sound speed structure and the stability of this structure in the deep isothermal layer. For example, as explained above, the axis of the deep sound channel becomes shallower toward both poles, eventually going to the surface.

2. SOUND PROPAGATION PATHS IN THE OCEAN

Figure 3 is a schematic of the basic types of propagation in the ocean resulting from the sound speed profiles (indicated by the dashed lines) discussed in the last section. These sound paths can be understood from a simplified statement of Snell's law: sound bends locally toward regions of low sound speed (or sound is "trapped" in regions of low sound speed). Simplest to explain are the paths about local sound speed minima: A, B, and C. Paths A and B correspond to surface duct propagation where the minimum sound speed is at the ocean surface (or at the bottom of the ice cover for the Arctic case). Path C, depicted by a ray leaving a deeper source at a shallow horizontal angle, propagates in the deep sound channel whose axis is at the shown sound speed minimum. As stated above, this local minimum tends to become more shallow towards polar latitudes converging to the Arctic surface minimum, path A. Hence for mild latitudes, sound in the deep channel can propagate long distances without interacting with lossy boundaries; propagation via this path has been observed over distances of thousands of kilometers. Also, from the above description of the geographical variation of the acoustic environment combined with Snell's law, we can expect that shallow sources coupling into the water column at polar latitudes will tend to propagate more horizontally around an axis which becomes deeper toward the mid latitudes. Path D, which is at slightly steeper angles than those associated with path C, is convergence zone propagation, a spatially periodic (~35–65 km) refocusing phenomenon producing zones of high intensity near the surface due to the upward refracting nature of the deep sound speed profile. Referring back

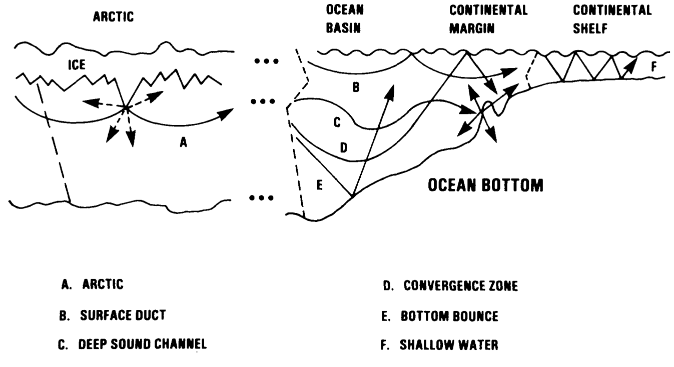

FIG. 3. Schematic representation of various types of sound propagation in the ocean.

to Fig. 1, there may be a depth in the deep isothermal layer at which the sound speed is the same as it is at the surface. This depth is called the "critical depth" and, in effect, is the lower limit of the deep sound channel. A "positive critical depth" specifies that the environment supports long distance propagation without bottom interaction, whereas "negative" implies that the bottom ocean boundary is the lower boundary of the deep sound channel. The bottom bounce path E, which interacts with the ocean bottom, is also a periodic phenomenon but with a shorter cycle distance and a shorter total propagation distance because of losses when sound is reflected from the ocean bottom. Finally, the right-hand side of Fig. 3 depicts propagation in a shallow water region such as a continental shelf. Here sound is channeled in a waveguide bounded above by the ocean surface and below by the ocean bottom. Because of the latter, negative critical depth environments exhibit much of the sound propagation physics descriptive of shallow water environments.

The modeling of sound propagation in the ocean is further complicated because the environment varies laterally (range dependence), and all environmental effects on sound propagation are dependent on acoustic frequency in a rather complicated way which often makes the ray type schematic of Fig. 3 misleading, particularly at low frequencies. Finally, a quantitative understanding of acoustic source mechanisms in the ocean is required for modeling sound propagation. These losses are, aside from geometric spreading: volume attenuation, bottom loss, and surface, volume (including fish), and bottom scattering loss.

2.1 Volume Attenuation

Volume attenuation increases with frequency. Returning to Fig. 3, the losses associated with path C, deep channel propagation, include volume attenuation and scattering because this path does not involve interaction with the boundaries. The volume scattering can be biological in origin or arise from the environmental condition that the upper boundary of the deep sound channel is typically the region of greatest internal wave activity in the ocean. Both of these effects are small for low frequencies, and hence deep sound propagation has been observed to distances of thousands of kilometers. This same internal wave region is also on the lower boundary of the surface duct, allowing scattering out of the surface duct, thereby constituting a loss mechanism for the surface duct. This mechanism also "leaks" sound into the Deep Sound Channel, a region which without scattering would be a "shadow zone" for a surface duct source. This type of volume scattering from internal waves is also thought to be a major source of fluctuation (Flatte, 1983).

Plane wave attenuation (α) is defined from a differential equation for a decay law as a function of distance:

$$\frac{dA}{dx}=-\alpha A \Rightarrow A=A_0 \exp(-\alpha x), \qquad (2)$$

where A_0 is the rms amplitude at $x=0$. The unit of α is nepers/distance. For example, a plane wave with wave number k that undergoes attenuation has the form $\exp(ikx-\alpha x)$. Attenuation is often expressed in decibels loss per unit distance, α', with units of dB/distance:

$$\text{Loss} \equiv 20 \log \frac{A_0}{A}=8.686\alpha x \Rightarrow \alpha'=8.686\alpha. \qquad (3)$$

The frequency dependence of attenuation can be roughly divided into four regimes of different physical origin as displayed in Fig. 4 (taken from Urick, 1979). The lowest frequency regime, Region I, is still not completely understood but it is conjectured that it is related to low-frequency propagation-duct cutoff, or, in other words, leakage out of the deep sound channel, as mentioned above. The main mechanisms associated with Regions II and III are boric acid and magnesium sulfate chemical relaxation. Region IV is dominated by the shear and bulk viscosity associated with freshwater. A summary of the approximate frequency dependence (f in kHz) of the attenuation (in units of dB/km) is (Urick, 1979; Fisher and Simmon, 1977; Thorpe, 1967)

$$\alpha'(\text{dB/km}) \approx 3.3\times 10^{-3}+\frac{0.11f^2}{1+f^2} + \frac{43f^2}{4100+f^2}+2.98\times 10^{-4}f^2, \qquad (4)$$

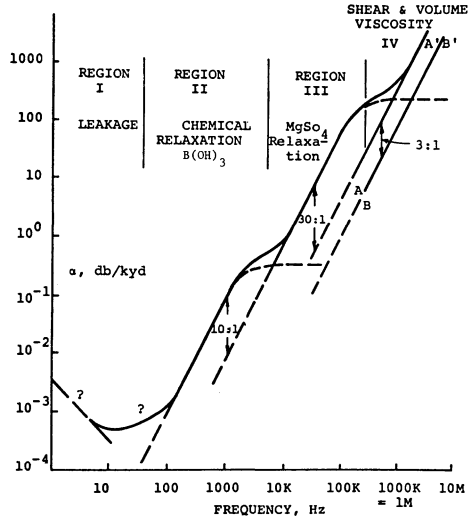

FIG. 4. Regions of different dominant processes of attenuation of sound in seawater (from Urick, 1979). The attenuation α is given in dB per kiloyard.

with the terms sequentially associated with Regions I–IV in Fig. 4.

2.2 Bottom Loss

When sound interacts with the ocean bottom, the structure of the ocean bottom becomes important. Ocean bottom sediments are often modeled as fluids, which support only one type of sound wave—the compressional wave. This is often a good approximation since the rigidity (and hence the shear speed) of the sediment is usually considerably less than that of a solid, such as rock. In the latter case, which applies to the "ocean basement," or the case where there is no sediment overlying the basement, the medium must be modeled as an elastic solid which means it supports both compressional and shear waves. In reality, the media are viscoelastic, as volume attenuation is important.

Reflectivity, the amplitude ratio of reflected and incident plane waves at an interface separating two media, is an important measure of the effect of the bottom on sound propagation. For an interface between two fluid semi-infinite half-spaces with density ρ_i and sound speed c_i, $i = 1, 2$, as shown in Fig. 5(a), the reflectivity is given by

$$\mathscr{R}(\theta) = \frac{\rho_2 k_{1z} - \rho_1 k_{2z}}{\rho_2 k_{1z} + \rho_1 k_{2z}}, \tag{5}$$

with

$$k_{iz} = (\omega / c_i) \sin\theta_i \equiv k_i \sin\theta_i; \quad i = 1, 2. \tag{6}$$

The incident and transmitted grazing angles are related by Snell's law,

$$k_\perp = k_1 \cos\theta_1 = k_2 \cos\theta_2, \tag{7}$$

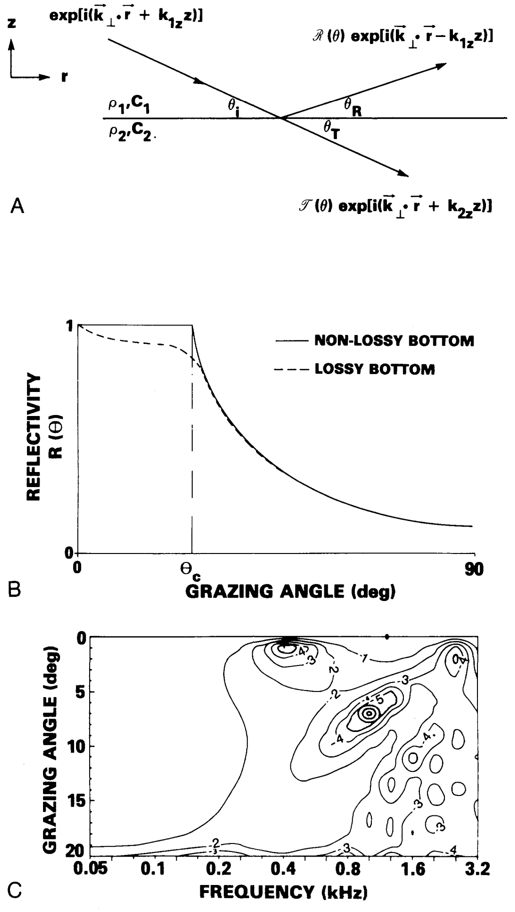

FIG. 5. The reflection and transmission process. Grazing angles are defined relative to the horizontal. (a) Plane wave impinges on interface separating two media with densities ρ_1, ρ_2 and sound speeds c_1, c_2. $\mathscr{R}(\theta)$ and $\mathscr{T}(\theta)$ are reflection and transmission coefficients. Snell's law is a statement that $k_\perp$ is the same for all three waves. (b) Rayleigh reflection curve indicating critical angle θ_c. The dashed curve shows that if the second medium is lossy, there is less than perfect reflection below the critical angle. (c) Contour of reflection loss for a layered bottom showing frequency and grazing angle dependence. Simpler Rayleigh reflection curve for each frequency is obtained from vertical slice.

where the incident grazing angle θ_1 is also equal to the angle of the reflected plane wave. $\mathscr{R}(\theta)$ is also referred to as the Rayleigh reflection coefficient and has unit magnitude (total internal reflection) when the numerator and denominator of Eq. (5) are complex conjugates. This occurs when k_{2z} is purely imaginary, and using Snell's law for determining θ_2 in terms of the incident grazing angle, we obtain the critical grazing angle, below which there is perfect reflection,

$$\cos\theta_c = c_1/c_2, \tag{8}$$

so that a critical angle can exist only when the speed in the second medium is higher than that of the first.

Using Eq. (6), Eq. (5) can be rewritten as

$$\mathscr{R}(\theta) = \frac{\rho_2 c_2/\sin\theta_2 - \rho_1 c_1/\sin\theta_1}{\rho_2 c_2/\sin\theta_2 + \rho_1 c_1/\sin\theta_1} = \frac{Z_2 - Z_1}{Z_2 + Z_1}, \tag{9}$$

where the right-hand side is in the form of impedances, $Z_i(\theta_i) = \rho_i c_i/\sin\theta_i$, which are the ratios of the pressure to the vertical particle velocity at the interface in the ith medium (Brekhovskikh, 1980). Written in this form, more complicated reflection coefficients become intuitively plausible. Consider the case in Fig. 5(a) where the second medium is elastic and thus supports shear as well as compressional waves with sound speeds c_{2s} and c_{2p}, respectively. The Rayleigh reflection coefficient is then given by

$$\mathscr{R}(\theta) = \frac{Z_{2\mathrm{tot}} - Z_1}{Z_{2\mathrm{tot}} + Z_1}, \tag{10}$$

with the total impedance of the second medium being

$$Z_{2\mathrm{tot}} \equiv Z_{2s} \sin^2 2\theta_{2s} + Z_{2p} \sin^2 2\theta_{2p}. \tag{11}$$

Snell's law for this case is

$$k_1 \cos\theta_1 = k_{2s} \cos\theta_{2s} = k_{2p} \cos\theta_{2p}. \tag{12}$$

It can be shown that the total impedance of the elastic medium is less than that of a liquid layer with the same compressional speed. Hence, the elastic interface appears to be "softer" than the equivalent liquid layer, which can be thought to arise from the fact that the shear wave provides an additional degree of freedom for the acoustic wave to penetrate into the second medium. A good approximation (neglecting interface waves) for a second medium, whose shear wave velocity is greater than the compressional wave velocity of the first medium, is to replace the elastic reflection coefficient with the two-fluid reflection coefficient using the shear wave of the second medium as the compressional wave of an "effective" second fluid medium.

In lossy media, attenuation can be included in the reflectivity formula by taking the sound speed as complex so that the wave numbers are, consequently, also complex, $k_i \to k_i + \alpha_i$. Then, a plane wave has the form exp $(ikx - \alpha x)$, just as in the volume attenuation case above. The acoustic properties of ocean bottom media have been estimated from measurements (see, for example, Hamilton and Bachman, 1982) and are given in Table 1 (as compiled in Jensen and Kuperman, 1983).

Figure 5(b) depicts a simple bottom loss curve derived from the Rayleigh reflection co-

Table 1. Geoacoustic properties of some ocean bottom types.

Bottom type	Porosity P(%)	Relative density $\frac{\rho_B}{\rho_W}$	Relative speed $\frac{C_P}{C_W}$	Compressional speed C_P (m/s)	Shear speed C_S (m/s)	Compressional attenuation β_ρ (dB/λ)	Shear attenuation β_ρ (dB/λ)
Clay-silt	60	1.6	1.01	1515	100	0.5	1.0
Sand-silt-clay	55	1.7	1.02	1530	150	0.8	1.5
Silt	50	1.8	1.05	1575	200	1.0	2.0
Sand-silt	40	1.9	1.1	1650	400	0.8	1.5
Coarse sand	35	2.0	1.2	1800	600	0.7	1.5
Chalk-limestone	···	2.2	1.5	2250	1000	0.4	1.0
Limestone	···	2.4	2.0	3000	1500	0.3	0.5
Basalt	···	2.6	3.5	5250	2500	0.2	0.5

efficient formula, where both the densities and sound speed of the second medium are larger than those in the first medium with unit reflectivity indicating perfect reflection. For loss in dB, 0 dB is perfect reflecting, 6 dB loss is an amplitude factor of one-half, 12 dB loss is one-fourth, etc. For a lossless bottom, severe loss occurs above the critical angle in the water column due to transmission into the bottom. For the lossy (more realistic) bottoms, only partial reflection occurs at all angles. With paths involving many bottom bounces (shallow water propagation), bottom losses as small as a few tenths of a dB per bounce accumulate and become significant because the propagation path may involve many tens of bounces.

Path E in Fig. 3, the bottom bounce path, often involves paths which correspond to angles near or above the critical angle; therefore, after a few bounces, the sound level is highly attenuated. On the other hand, for shallow angles, many bounces are possible. Hence, in shallow water, path F, most of the energy that propagates is close to the horizontal and this type of propagation is most analogous to waveguide propagation. In fact, as shown in Fig. 6, there exists a small cone from which energy propagates long distances (θ_c is typically 10°–20°). Energy outside the cone is referred to as the near field (or continuous spectrum), which eventually escapes the waveguide. The trapped field originating from within the cone is referred to as the normal mode field (or discrete spectrum) because there are a set of angles corresponding to discrete paths which constructively interfere and make up the normal (natural) modes of the shallow water environment. The modes within the cone are discrete because the nearly perfect reflectivity permits the existence of a set of discrete vertical standing waves analogous to those of a vibrating string or an organ pipe.

Very often, the ocean is more complicated than that described by the frequency independent reflection loss of Fig. 5(b). For example, the ocean may be a complex layered structure. For this bottom type, the reflectivity will be a complicated function of frequency and incident grazing angle with respect to the horizontal. Figure 5(c) displays some model results of loss contours of such a layered bottom. The more familiar curve of reflectivity versus grazing angle, such as depicted in Fig. 5(b), is obtained from a vertical cut through the loss contours (the loss contours are expressed in decibels). For example, at low fre-

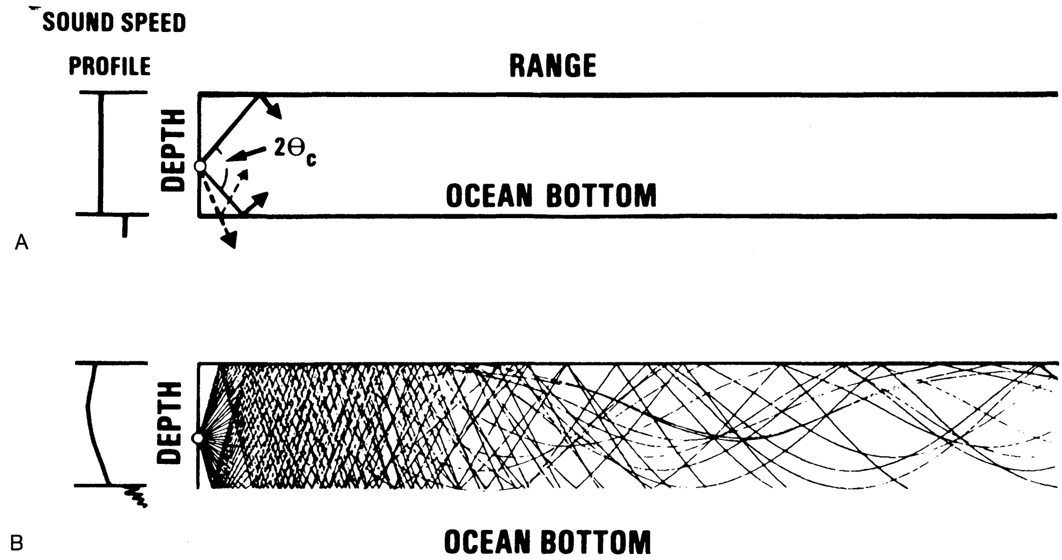

FIG. 6. Ocean waveguide propagation. (a) Long distance propagation occurs within a cone of $2\theta_c$. For angles greater than the critical angle θ_c, acoustic energy is partially transmitted into the ocean bottom at each bounce resulting in severe decay of the acoustic field in the water column with range. (b) The same as (a) but with a nonisovelocity water column showing refraction. Even below critical angle, higher angle paths have greater loss per bounce, as shown in Fig. 5(a).

quencies, Fig. 5(c) indicates a critical angle of about 18°, whereas at high frequencies we do not see a critical angle.

2.3 Surface, Bottom, and Volume Scattering

Scattering is a mechanism for loss, interference, and fluctuation phenomena. Ocean surface and bottom roughness attenuate the mean acoustic field. This attenuation increases with frequency. The field scattered away from the specular direction (reverberation) acts as interference for active sonar systems. Because the ocean surface moves, it generates acoustic fluctuations. Bottom roughness can also generate fluctuations when the sound source and/or receiver is moving. The effect of the roughness depends on the sound speed profile which determines the degree of interaction of sound with the rough surface.

The scattering theory literature is massive (see Ogilvy, 1987 for a review). Very often the effect of scattering from a rough surface is thought of in terms of a loss of a specularly reflected (coherent) component which results from scattering out of the coherent or specular field. The ocean bottom or surface can often be modeled as a randomly rough surface; if the roughness is small, relative to a wavelength, the scattering process can be modeled as a reduced reflection coefficient. A formula often used (Urick, 1979) for reflection from a rough interface is

$$\mathscr{R}'(\theta)=\mathscr{R}(\theta)e^{-2\Gamma^2}, \tag{13}$$

where $\mathscr{R}'(\theta)$ is the reduced reflection coefficient and Γ is the Rayleigh roughness parameter:

$$\Gamma=K\sigma\sin\theta. \tag{14}$$

In particular, the rough coherent sea surface reflection coefficient for the coherent field is

$$\mathscr{R}'(\theta)=-e^{-2\Gamma^2}. \tag{15}$$

For the ocean bottom, the appropriate parameters such as those from Table 1 are inserted into the expression for $\mathscr{R}$.

The coherent field is attenuated because of the generation of a scattered incoherent field, which is referred to as reverberation and can cause interference in the signal processing of the coherent waves. The excitation of this scattered field is characterized by a scattering cross section, which is defined as the ratio of the scattered power measured at a unit distance from a specified point to the intensity incident on a specified unit area (or unit volume in the case of volume scattering). Since power is the product of intensity and area, cross section is an effective area which multiplies the intensity of an incident field to give the scattered power. In underwater acoustics, the surface (area) and volume scattering strength $S_{s,v}$ (see Sec. 6.2) has become the more conventional parameter of reverberation and it is defined as the ratio (in dB units) of the intensity of the sound scattered by a unit surface area or volume, referenced to a unit distance (normally a yard), I_{scat} to the incident plane wave intensity I_{inc}:

$$S_{s,v}=10\log\frac{I_{\text{scat}}}{I_{\text{inc}}}. \tag{16}$$

Below we give empirical expressions for surface and bottom backscattering strengths $S_{S,B}$, which have been employed with some success.

At the ocean surface, the Chapman–Harris curves (Chapman and Harris, 1962) derived from measurements over the frequency range of 400 to 6400 Hz and wind speed up to 30 kn give a backscattering strength dependent on grazing angle θ in degrees, windspeed w in kn, frequency f in Hz, but independent of scattering angle:

$$S_S=3.3\,\beta\log\frac{\theta}{30}-42.4\log\beta+2.6;$$

$$\beta=158(wf^{1/3})^{-0.58}, \tag{17}$$

where the units are dB referenced to 1 yard.

At the ocean bottom, diffuse scattering described by Lambert's law together with an empirical scattering coefficient is used to estimate bottom scattering strengths from very rough ocean bottoms. Lambert's law states that the scattered power from a rough surface is proportional to the sine of the scattering angle ϕ. The incident power intercepted by the surface δA is proportional to the intensity $I_{\text{inc}}\times\sin\theta\delta A$. With Eq. (16) (per unit area) in the backscattered direction $\phi=\pi-\theta$, Lambert's law gives bottom scattering strength

$$S_B=-5+10\log\sin^2\theta, \tag{18}$$

where the first term is the proportionality constant arising from an assumption that all the incident energy is scattered and redistributed into the water column with no transmission into the bottom (Urick, 1983). Often, the first term of Eq. (18) is empirically adjusted according to a measured scattering strength.

In Sec. 2.1 we quantitatively summarized volume attenuation, part of which is caused by scattering. This scattering also produces a backscattered or reverberant acoustic field. However, most volume reverberation is thought to arise from biological organisms. A quantity often used to describe volume scattering is column strength. Returning to the discussion above Eq. (16), a surface (area) scattering strength can be related to a local volume scattering strength at depth h, $s_v(h)$:

$$S_s = 10 \log \int_0^H s_v(h)dh = S_v + 10 \log H, \qquad (19)$$

where S_v is an average volume scattering strength and H is a layer thickness in consistent units. When H is the entire water depth, $S_s \equiv S_c$ is called the column or integrated scattering strength.

In general, volume scattering decreases with increasing depth (about 5 dB per 300 m) with the exception of the deep scattering layer discussed below. For the lower frequencies (< 10 kHz), fish with air-filled swim bladders are the main scatterers. Above 20 kHz, zooplankton or smaller animals that feed upon the phyloplankton and the associated biological chain are the scatterers. The depth of the deep scattering layer varies throughout the day, being deeper in the day than in the night and changing most rapidly during sunset and sunrise. This layer produces a strong scattering increase of 5–15 dB within 100 m of the surface at night, and virtually no scattering in the daytime at the surface since it can migrate down to a depth of about 180 to 900 m at midlatitudes. Typical scattering results for the integral in Eq. (19) ($M_z \equiv s_v$) are shown in Fig. 7 (Chapman and Marshall, 1966) where the difference in day and night spectra of column scattering strength is thought to arise from the lower hydrostatic pressure nearer the ocean surface. This tends to lower the resonance frequency of the swim bladders, and the peak is lower in the night.

Finally, scattering off bubbles near the surface is sometimes referred to as either a volume or a surface scattering mechanism. These bubbles arise not only from sea surface action, but also from biological origins and ship wakes. Further, bubbles are not only a scattering mechanism, but bubble clouds may significantly alter local sound speed conditions. At the sea surface, the relative importance of roughness versus bubble effects is not yet resolved.

2.4 Ambient Noise

In underwater acoustics, ambient noise is an issue when it masks a signal of interest. Ocean acoustic signal processing (see Sec. 8) is a procedure for extracting a signal embed-

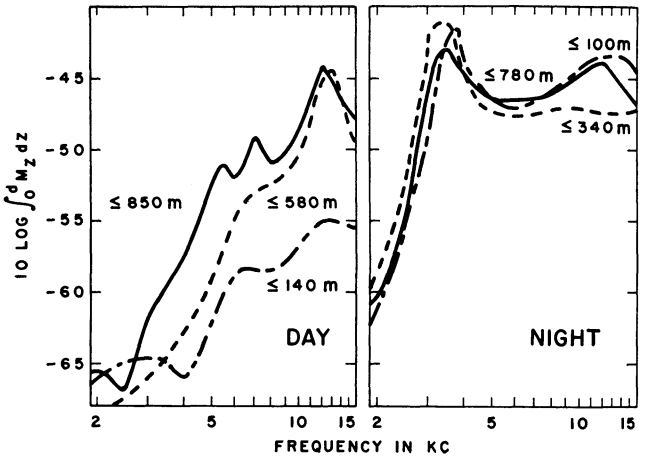

FIG. 7. Day and night scattering strength measurements using an explosive source, as a function of frequency (from Chapman and Marshall, 1966). The spectra measured at various times after the explosion are labeled with the depth of the nearest scatterer that could have contributed to the reverberation. The variable M_z corresponds to s_v in the text.

ded in noise. The noise is irrelevant if the signal is very strong. However, the more interesting case is the marginal situation of low signal to noise ratio. We would like to exploit the difference in the properties of the signal of interest and the noise to be rejected. Simple examples of this procedure are: omnidirectional noise can be reduced by a directional receiver with narrow look directions (beams); directional noise can be avoided by not looking in the direction of the noise. The more general case of achieving noise gain (i.e., enhancement of signal over noise) is to factor into the design of a receiving system knowledge of the general distribution of the ambient noise including its coherence properties. In this section we describe some of the properties of ocean ambient noise. In particular, we summarize the spectral character of noise. In Sec. 5 we discuss the relevant acoustic theory which explains the spatial distribution of noise.

There are essentially two types of noise: man-made and natural. Generally, shipping is the most important source of man-made noise, though noise from offshore rigs is becoming more prevalent. Usually, natural noise dominates at low frequencies (below 10 Hz) and at high frequencies (above a few hundred Hz). The field of ambient noise is reviewed rather extensively by Ross (1976), Urick (1984), and Kerman (1988), the last being a series of reviews and research papers on natural mechanisms of sea noise. Shipping noise fills in the region between ten and a few hundred Hz. A summary of the spectrum of noise is shown in Fig. 8 (Wenz, 1962). The higher frequency wind noise is usually parametrized according to sea state (also Beaufort number) and/or wind. Table 2 (Wenz, 1962) summarizes the relationship between sea state, wind, and wave height. The discussion of the spatial distribution of noise ap-

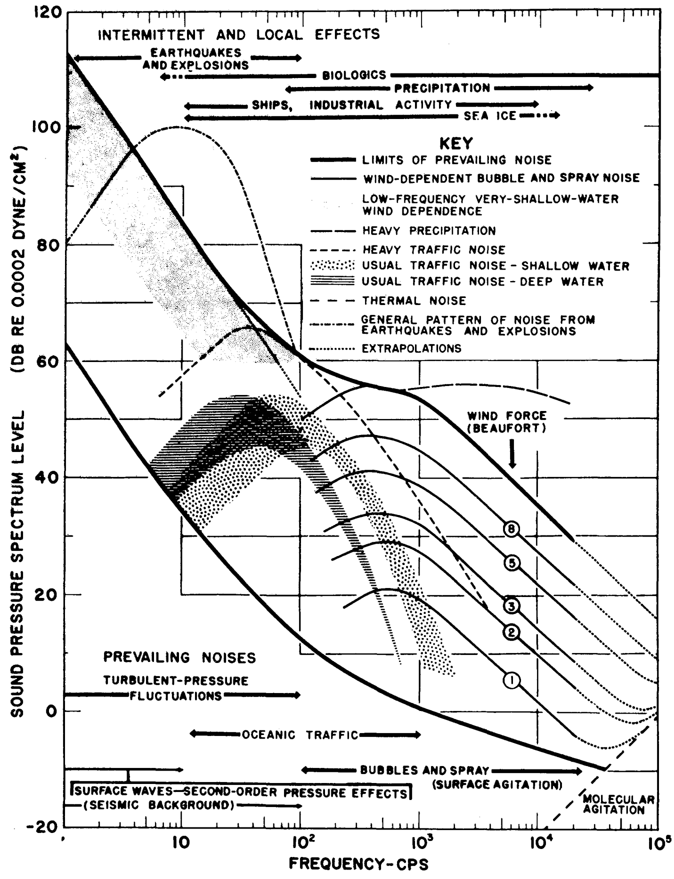

FIG. 8. Composite of ambient noise spectra (from Wenz, 1962).

Table 2. Descriptions of the ocean sea surface (from Wenz, 1962). Approximate relation between scales of wind speed, wave height, and sea state.

	Wind speed			12-h wind	Fully arisen sea			
Sea criteria	Beaufort scale	Range knots (m/s)	Mean knots (m/s)	Wave height[a,b] ft (m)	Wave height[a,b] ft (m)	Duration[b,c] h	Fetch[b,c] naut. miles (km)	Seastate scale
Mirrorlike	0	<1 (<0.5)						0
Ripples	1	1–3 (0.5–1.7)	2 (1.1)					$\frac{1}{2}$
Small wavelets	2	4–6 (1.8–3.3)	5 (2.5)	<1 (<0.30)	<1 (<0.30)			1
Large wavelets, scattered whitecaps	3	7–10 (3.4–5.4)	$8\frac{1}{2}$ (4.4)	1–2 (0.30–0.61)	1–2 (0.30–0.61)	<2.5	<10 (<19)	2
Small waves, frequent whitecaps	4	11–16 (5.5–8.4)	$13\frac{1}{2}$ (6.9)	2–5 (0.61–1.5)	2–6 (0.61–1.8)	2.5–6.5	10–40 (19–74)	3
Moderate waves, many whitecaps	5	17–21 (8.5–11.1)	19 (9.8)	5–8 (1.5–2.4)	6–10 (1.8–3.0)	6.5–11	40–100 (74–185)	4
Large waves, whitecaps everywhere, spray	6	22–27 (11.2–14.1)	$24\frac{1}{2}$ (12.6)	8–12 (2.4–3.7)	10–17 (3.0–5.2)	11–18	100–200 (185–370)	5
Heaped-up sea, blown spray, streaks	7	28–33 (14.2–17.2)	$30\frac{1}{2}$ (15.7)	12–17 (3.7–5.2)	17–26 (5.2–7.9)	18–29	200–400 (370–740)	6
Moderately high, long waves, spindrift	8	34–40 (17.3–20.8)	37 (19.0)	17–24 (5.2–7.3)	26–39 (7.9–11.9)	29–42	400–700 (740–1300)	7

---

[a] The average height of the highest one-third of the waves (significant wave height).

[b] Estimated from data given in U. S. Hydrographic Office (Washington, D.C.) publications HO 604 (1951) and HO 603 (1955).

[c] The minimum fetch and duration of the wind needed to generate a fully arisen sea.

pears in a following section after propagation models are described.

3. SOUND PROPAGATION MODELS

Sound propagation in the ocean is mathematically described by the wave equation whose parameters and boundary conditions are descriptive of the ocean environment. There are essentially four types of models (computer solutions to the wave equation) to describe sound propagation in the sea: ray theory, fast field program (FFP), normal mode (NM), and parabolic equation (PE). All of these models allow for the fact that the ocean environment varies with depth. A model that also takes into account horizontal variations in the environment (i.e., sloping bottom or spatially variable oceanography) is termed range dependent. For high frequencies (a few kHz or above), ray theory is the most practical. The other three model types are more applicable and usable at lower frequencies (below 1 kHz). The hierarchy of underwater acoustics models is shown in schematic form in Fig. 9 (modified from Kuperman and McDonald, 1986, which also includes a list of references to the specific models).

3.1 The Wave Equation and Boundary Conditions

The wave equation together with the appropriate boundary conditions describes acoustic fields in various media. It is typically written and solved in terms of pressure, displacement, or velocity potentials. This choice is not critical for fluid media but it is convenient, though not necessary, to formulate the elastic or shear-supporting medium problem in terms of displacement potentials. For a velocity potential φ, the wave equation in cylindrical coordinates with the range coordinates denoted by $\mathbf{r}=(x,y)$ and the depth coordinate denoted by z (and taken positive downward) for a source free region is

$$\nabla^2\varphi(\mathbf{r}, z, t)-\frac{1}{c^2}\frac{\partial^2\varphi(\mathbf{r}, z, t)}{\partial t^2}=0, \tag{20}$$

where c is the sound speed in the wave propagating medium and appears in the wave equation because its derivation involves the linearized equations of hydrodynamics and a linear adiabatic equation of state relating pressure changes to density changes, and c is related to the compressibility of the medium. [Nonlinear acoustics theory retains the convection term in the hydrodynamics equations as well as higher order terms in the equation of state (Kuperman and McDonald, 1986).] With respect to the velocity potential, the velocity v and pressure p are given by

$$\mathbf{v}=\nabla\varphi, \quad p=-\rho\frac{\partial\varphi}{\partial t}, \tag{21}$$

where ρ is the density of the medium. The wave equation is most often solved in the frequency domain; that is, a frequency dependence of $\exp(-i\omega t)$ is assumed to obtain the Helmholtz equation,

$$\nabla^2\varphi(\mathbf{r}, z)+K^2\varphi(\mathbf{r}, z)=0, \tag{22}$$

where $K\equiv\omega/c$.

In underwater acoustics both fluid and elastic (shear supporting) media are of interest. In elastic media the field can be expressed in terms of three scalar displacement potentials, $\phi(\mathbf{r}, z)$, $\psi(\mathbf{r}, z)$, and $\Lambda(\mathbf{r}, z)$, corresponding to compressional (P), vertically polarized (SV), and horizontally polarized shear waves (SH), respectively. For brevity, we write

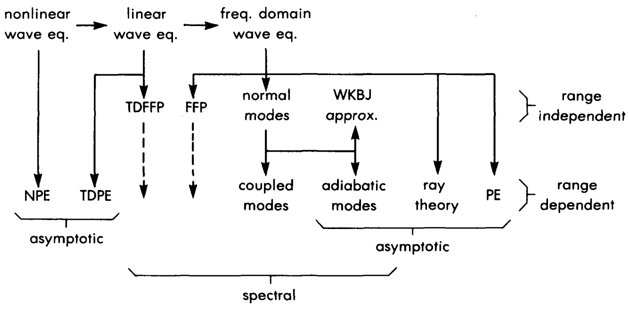

FIG. 9. Hierarchy of underwater acoustics models. (Note: TD refers to time domain.)

$$\Phi(\mathbf{r}, z)=\begin{cases}\phi(\mathbf{r}, z), & \text{P waves}\\ \psi(\mathbf{r}, z), & \text{SV waves}\\ \Lambda(\mathbf{r}, z), & \text{SH waves.}\end{cases} \tag{23}$$

In the limiting case of a fluid medium, where shear waves do not exist, $\Phi(\mathbf{r}, z)$ represents the compressional potential $\phi(\mathbf{r}, z)$. Most underwater acoustic applications involve only compressional sources which only excite the P and SV potentials, eliminating the SH potential $\Lambda(\mathbf{r}, z)$. The displacement potentials satisfy the Helmholtz equation with the appropriate compressional or shear sound speed, c_p or c_s, respectively,

$$c_p=\left[\frac{\lambda+2\mu}{\rho}\right]^{1/2}, \quad c_s=\left[\frac{\mu}{\rho}\right]^{1/2}, \tag{24}$$

where λ and μ are the Lamé constants. For the above defined fields, we can now write down the most common plane interface boundary conditions encountered in underwater acoustics. For the ocean surface, the pressure (normal stress) vanishes. This is often referred to as the pressure release condition equation involving the appropriate solution of the Helmholtz equation:

$$p=0, \quad \varphi=0, \quad \text{or} \quad \phi=0. \tag{25}$$

The interface between the water column (layer 1) and an ocean bottom (layer 2) is often characterized as a fluid-fluid interface. The continuity of pressure and vertical particle velocity, respectively, across a horizontal interface separating the layers yields the following boundary conditions for the Helmholtz equation in terms of pressure:

$$p_1=p_2, \quad \frac{1}{\rho_1}\frac{\partial p_1}{\partial z}=\frac{1}{\rho_2}\frac{\partial p_2}{\partial z}, \tag{26}$$

or velocity potential:

$$\rho_1\varphi_1=\rho_2\varphi_2, \quad \frac{\partial \varphi_1}{\partial z}=\frac{\partial \varphi_2}{\partial z}. \tag{27}$$

The boundary conditions applied to the plane wave fields in Fig. 5(a) yield the Rayleigh reflection coefficient given by Eq. (5).

For an interface separating two solid layers, the boundary conditions are continuity of vertical displacement $w(\mathbf{r}, z_i)$, tangential displacements $\mathbf{u}(\mathbf{r}, z_i) = (u_x, u_y)$, normal stress $n = \sigma_{zz}$, and tangential stresses $\mathbf{t} = (\sigma_{xx}, \sigma_{yz})$. (Note that the equations normal to the interface are scalar, whereas the tangential equations are vector equations.) In cylindrical coordinates with azimuthal symmetry, the radial and vertical components of displacements (in homogeneous media), u and w, respectively, are

$$u(r, z)=\frac{\partial \phi}{\partial r}+\frac{\partial^2 \psi}{\partial z^2}, \tag{28}$$

$$w(r, z)=\frac{\partial \phi}{\partial z}-\frac{1}{r}\frac{\partial}{\partial r} r \frac{\partial \psi}{\partial r}, \tag{29}$$

and the normal and tangential stresses are

$$\sigma_{zz}(r, z)=(\lambda+2\mu)\frac{\partial w}{\partial z}+\lambda\frac{\partial u}{\partial r}, \tag{30}$$

$$\sigma_{rz}(r, z)=\mu\left[\frac{\partial u}{\partial z}+\frac{\partial w}{\partial r}\right]. \tag{31}$$

Continuity of these quantities at the interface between two solids are the boundary conditions. For a fluid-solid interface, the rigidity μ vanishes in the fluid layers and the tangential stress in the solid layer vanishes at the boundary.

The Helmholtz equation for an acoustic field from a point source with angular frequency ω is

$$\nabla^2 G(\mathbf{r}, z)+K^2(\mathbf{r}, z)G(\mathbf{r}, z) = -\delta^2(\mathbf{r}-\mathbf{r}_s)\delta(z-z_s);$$

$$K^2(\mathbf{r}, z)=\frac{\omega^2}{c^2(\mathbf{r}, z)}, \tag{32}$$

where the subscript s denotes the source coordinates. The range dependent environment manifests itself as the coefficient $K^2(\mathbf{r}, z)$ of the partial differential equation for the appropriate sound speed profile. The range dependent bottom type and topography appears as boundary conditions on scalars and tangential and normal quantities as discussed above.

3.2 Ray Theory

Ray theory (Keller and Papadakis, 1977), a geometrical approximation, in its simplest form is just an application of Snell's law. It is a high frequency approximation to Eq. (22) and assumes a solution of the form

$$G(\mathbf{R})=A(\mathbf{R}) \exp[iS(\mathbf{R})], \tag{33}$$

where the exponential term allows for rapid variations as a function of range and $A(\mathbf{R})$ is a more slowly varying "envelope" which incorporates both geometrical spreading and loss mechanisms. Substituting Eq. (33) into Eq. (22) in a source free space, we obtain from real and imaginary parts

$$\frac{1}{A}\nabla^2 A-(\nabla S)^2+K^2=0;$$

$$\nabla A\cdot\nabla S+\frac{1}{2}A\nabla^2 S=0. \tag{34}$$

The geometrical approximation is that the amplitude varies slowly with range [i.e., $(1/A)\nabla^2 A \ll K^2$] so that the first part of Eq. (34) is approximated by the eikonal equation

$$(\nabla S)^2=K^2. \tag{35}$$

The ray trajectories are perpendicular to surfaces of constant phase (wave fronts) S and may be expressed mathematically as follows:

$$\frac{d}{dl}\left[K\frac{d\mathbf{R}}{dl}\right]=\nabla K, \tag{36}$$

where l is the arc length along the direction of the ray and $\mathbf{R}$ is the displacement vector. One can show that the direction of average flux (energy) follows that of the trajectories, and the amplitude of the field at any point can be obtained from the density of rays. Once S is obtained, the second part of Eq. (34) yields the amplitude. We mention here, also, that "corrected" ray theory assumes that A can be expanded in powers of inverse frequency—the leading term is the infinite-frequency result with the additional terms being frequency corrections.

The ray theory method is computationally rapid and extends to range dependent problems. Furthermore, the ray traces give a physical picture of the acoustic paths. It is helpful in describing how noise redistributes itself when propagating long distances over paths that include shallow and deep environments and/or midlatitude to polar regions. The disadvantage of ray theory is that it does not include diffraction and such effects that describe the low-frequency dependence ("degree of trapping") of ducted propagation. An example is given at the end of this section.

3.3 Full Field Solution or "Fast Field Program" (FFP)

The wave equation can be solved efficiently with spectral methods when the ocean environment does not vary with range. The term "Fast Field Program" is used because the spectral methods became practical with the advent of the fast Fourier transform (FFT). One of the possible derivations of the solution of Eq. (32) is to Fourier decompose the acoustic field into an infinite set of horizontal waves:

$$G(\mathbf{r},z)=\frac{1}{2\pi}\int_{-\infty}^{\infty} d^2k\, g(\mathbf{k},z,z_s) \times\exp[i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}_s)], \tag{37}$$

and from Eq. (22), the depth dependent Green's function $g(\mathbf{k},z,z_s)$ satisfies

$$\frac{d^2g}{dz^2}+[K^2(z)-k^2]g=-\frac{1}{2\pi}\,\delta(z-z_s). \tag{38}$$

Assuming azimuthal symmetry, we can integrate Eq. (37) over the angular variable obtaining Hankel functions; using their asymptotic form reduces Eq. (37) to (for simplicity, we take $\mathbf{r}_s=0$)

$$G(r,z)=\frac{\exp(-i\pi/4)}{(2\pi r)^{1/2}}\int_{-\infty}^{\infty} dk\,k^{1/2} \times g(k,z,z_s)\exp(ikr). \tag{39}$$

Note that the factor $r^{-1/2}$ arises from cylindrical spreading. We now discretize the above integral and transform to a form amenable to the FFT technique by setting $k_m = k_0 + m\Delta k$; $r_n = r_0 + n\Delta r$, where $n, m = 0,1,\ldots,N-1$ with the additional condition $\Delta r\Delta k = 2\pi/N$ and N is an integral power of 2. The discretization scheme limits the solution to outgoing waves and Eq. (39) becomes

$$G(\mathbf{r}_n,z)=\frac{\Delta k\,\exp[i(k_0r_n-\pi/4)]}{(2\pi r)^{1/2}}\sum_{m=0}^{N-1}X_m \times\exp(2\pi imn/N), \tag{40}$$

$$X_m=(k_m)^{1/2}g(k_m,z,z_s)\,\exp(imr_0\Delta k).$$

The above equation is now easily evaluated using the FFT algorithm with the bulk of the effort going into evaluating the Green's function by solving Eq. (38). Although the method is labeled "fast field" it is fairly slow because of the time required to calculate the Green's functions. However, it has advantages when one wishes to calculate the "near field" region or to include shear wave effects in elastic media (Schmidt and Jensen, 1985); it is also often used as a benchmark for other less exact techniques (Kuperman and Jensen, 1980). One such technique, not applicable to the near field but essentially exact for a large class of range independent far-field problems, is the computationally faster normal mode method.

3.4 Normal Mode Model (NM)

Rather than evaluate Eq. (38) for each g for the complete set of k's [typically solving Eq. (38) 1024 to 8196 times], one can utilize a normal mode expansion of the form

$$g(\mathbf{k},z)=\sum a_n(\mathbf{k})u_n(z), \tag{41}$$

where the quantities u_n are eigenfunctions of the following eigenvalue problem:

$$\frac{d^2u_n}{dz^2}+[K^2(z)-k_n^2]u_n(z)=0. \tag{42}$$

The eigenfunctions u_n are zero at $z=0$, satisfy the local boundary conditions descriptive of the ocean bottom properties, and satisfy a radiation condition for $z\rightarrow\infty$. They form an orthonormal set in a Hilbert space with weighting function $\rho(z)$, the local density. The range of discrete eigenvalues is given by the condition

$$\min[K(z)]<k_n<\max[K(z)]. \tag{43}$$

These discrete eigenvalues correspond to discrete angles within the critical angle cone in Fig. 6, such that specific waves constructively interfere. The eigenvalues k_n typically have a small imaginary part α_n, which serves as modal attenuation representative of all the losses in the ocean environment (see Kuperman and Ingenito, 1977, for the formulation of normal mode attenuation coefficients and other references on the subject).

We now substitute the normal mode expansion of g given by Eq. (41) into Eq. (37) (we take the source at the origin for simplicity) and multiply both sides by $\rho(z)u_n(z)$ and integrate over all z space to obtain, with the Hilbert space orthonormality conditions,

$$G(r,z)=\frac{\rho(z_s\,z)}{(2\pi)^2}\int_{-\infty}^{\infty}dk_x\int_{-\infty}^{\infty}dk_y \times\sum_n\frac{u_n(z_s)u_n(z)}{k^2-k_n^2}\exp(i\mathbf{k}\cdot\mathbf{r}). \tag{44}$$

We choose a contour about the poles which leads to a solution with outgoing waves from the point source. Each integral in Eq. (44) is proportional to the two-dimensional plane wave representation of the zero-order Hankel function of the first kind, and G can be expressed as

$$G(r,z)=\frac{i}{4}\rho(z_s)\sum_n u_n(z_s)u_n(z)H_0^1(k_nr). \tag{45}$$

The asymptotic form of the Hankel function can be used in the above equation to obtain the well known normal mode representation of a cylindrical (axis is depth) waveguide:

$$G(r,z)=\frac{i\rho(z_s)}{(8\pi r)^{1/2}}\exp(-i\pi/4) \times\sum_n\frac{u_n(z_s)u_n(z)}{k_n^{1/2}}\exp(ik_nr). \tag{46}$$

Equation (46) is a far field solution of the wave equation and neglects the continuous spectrum $\{k_n<\min[K(z)]\}$ of modes. For purposes of illustrating the various portions of the acoustic field, we note that k_n is a horizontal wave number so that a "ray angle" associated with a mode with respect to the horizontal can be taken to be $\theta=\cos^{-1}[k_n/K(z)]$. For a simple waveguide the maximum sound speed is the bottom sound speed corresponding to min $[K(z)]$. At this value of $K(z)$, we have from Snell's law $\theta=\theta_c$, the bottom critical angle. In effect, if we look at a ray picture of the modes, the continuous portion of the mode spectrum corresponds to rays with grazing angles greater than the bottom critical angle of Fig. 5(b) and therefore outside the cone of Fig. 6. This portion, as explained in Sec. 3.2, undergoes severe loss. Hence, we note that the continuous spectrum is the near (vertical) field and the discrete spectrum is the (more horizontal, profile dependent) far field (falling within the cone in Fig. 6). Note also that, in describing the acoustic paths in these two regimes, we use the result from Snell's law that, for a medium with a vertical variability in index of refraction like the horizontally stratified ocean, rays that start close to the horizontal are "more confined within their ducts" than the more vertical rays. Hence, we conclude that the continuous, more vertical portion of the sound field is less profile dependent than the discrete, more horizontal part.

The advantages of the NM procedure are that the solution is available for all source and receiver configurations once the eigenvalue problem is solved; it is easily extended to moderately range-dependent conditions using the adiabatic approximation; and it can be applied (with more effort) to extremely range-dependent environments using coupled mode theory.

3.5 Adiabatic Mode Theory

All of the range independent normal mode "machinery" developed for environmental ocean acoustic modeling applications can be adapted to mildly range dependent conditions using adiabatic mode theory. The underlying assumption (Brekhovskikh and Lysanov, 1982) is that individual propagating normal modes adapt (but do not scatter or "couple" into each other) to the local environment. The coefficients of the mode expansion, a in Eq. (41), now become mild functions of range, i.e., $a_n(k) \rightarrow a_n(k,\mathbf{r})$. This modifies Eq. (46) as follows:

$$G(r,z)=\frac{ip(z_s)}{(8\pi r)^{1/2}}\exp(-i\pi/4)$$

$$\times\sum_n \frac{u_n(z_s)v_n(z)}{\overline{k}_n^{\,1/2}}\exp(i\overline{k}_n r), \tag{47}$$

where the range-averaged wave number (eigenvalue) is

$$\overline{k}_n=\frac{1}{r}\int_0^{\mathrm{r}} k_n(r')dr', \tag{48}$$

and the $k_n(r')$ are obtained at each range segment from the eigenvalue problem Eq. (42) evaluated at the environment at that particular range along the path. The quantities u_n and v_n are the sets of modes at the source and the field positions, respectively.

Simply stated, the adiabatic mode theory leads to a description of sound propagation such that the acoustic field is a function of the modal structure at both the source and the receiver and some average propagation conditions between the two. Thus, for example, when sound emanates from a shallow region where only two discrete modes exist and propagates into a deeper region with the same bottom (same critical angle), the two modes from the shallow region adapt to the form of the first two modes in the deep region. However, the deep region can support many more modes; intuitively, we therefore expect the resulting two modes in the deep region will take up a smaller, more horizontal part of the cone of Fig. 6 than they take up in the shallow region. This means that sound rays going from shallow to deep tend to become more horizontal which is consistent with a ray picture of downslope propagation.

Fully coupled mode theory for range dependent environments has been developed (Evans, 1983). Numerical methods to evaluate coupled mode theories efficiently are under development.

3.6 Parabolic Equation Model (PE)

The PE method is presently the most practical and encompassing wave-theoretic range-dependent propagation model. In its simplest form, it is a far-field narrow-angle (~ ±20° with respect to the horizontal—adequate for many underwater propagation problems) approximation to the wave equation. Assuming azimuthal symmetry about a source, we express the solution of Eq. (32) in cylindrical coordinates in a source free region in the form

$$G(r,z)=\psi(r,z)\cdot J(r), \tag{49}$$

and we define $K^2(r,z)\equiv K_0^2 n^2$, n therefore being an "index of refraction" c_0/c, where c_0 is a reference sound speed. Substituting Eq. (49) into Eq. (32) in a source free region and taking K_0^2 as the separation constant, J and ψ satisfy the following two equations:

$$\frac{d^2J}{dr^2}+\frac{1}{r}\frac{dJ}{dr}+K_0^2\,J=0; \tag{50}$$

$$\frac{\partial^2\psi}{\partial r^2}+\frac{\partial^2\psi}{\partial z^2}+\left(\frac{1}{r}+\frac{2}{J}\frac{\partial J}{\partial r}\right)\frac{\partial\psi}{\partial r}+K_0^2n^2\psi-K_0^2\psi=0. \tag{51}$$

Equation (50) is a Bessel equation and we take the outgoing solution, a Hankel function $H_0^1(K_0r)$, in its asymptotic form and substitute it into Eq. (51), together with the "paraxial" (narrow angle) approximation:

$$\left|\frac{\partial^2\psi}{\partial r^2}\right| \ll \left|2iK_0\frac{\partial\psi}{\partial r}\right|, \tag{52}$$

to obtain the parabolic equation (in r),

$$\frac{\partial^2\psi}{\partial z^2}+2iK_0\frac{\partial\psi}{\partial r}+K_0^2(n^2-1)\psi=0, \tag{53}$$

where we note that n is a function of range and depth. We use a marching solution to solve the parabolic equation. There has been an assortment of numerical solutions but the one that still remains the most efficient is the so-called "split step algorithm" (Tappert, 1977), which we briefly describe below.

For this derivation, we take n to be a constant; the error this introduces can be made arbitrarily small by the appropriate numerical gridding. The Fourier transform of ψ with respect to z can then be written as

$$\chi(r,s)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\psi(r,z)\exp(-isz)dz, \tag{54}$$

which together with Eq. (53) gives

$$-s^2\chi+2iK_0\frac{\partial\chi}{\partial r}+K_0^2(n^2-1)\chi=0. \tag{55}$$

The solution of Eq. (55) is simply

$$\chi(r,s)=\chi(r_0,s)\exp\left[-\frac{K_0^2(n^2-1)-s^2}{2iK_0}\times(r-r_0)\right], \tag{56}$$

where the initial condition at r_0 must be specified. The inverse transform then gives the field as a function of depth:

$$\psi(r,z)=\int_{-\infty}^{\infty}\chi(r_0,s)\exp\left[\frac{iK_0}{2}(n^2-1)\Delta r\right]\times\exp\left[-\frac{i\Delta r}{2K_0}s^2\right]\exp(isz)ds, \tag{57}$$

where $\Delta r = r - r_0$. Introducing the symbol $\mathscr{F}$ for the Fourier transform operation from the z domain [as performed in Eq. (54)] and $\mathscr{F}^{-1}$ as the inverse transform, Eq. (57) can be summarized by the range-stepping algorithm:

$$\psi(r+\Delta r,z)=\exp\left[\frac{iK_0}{2}(n^2-1)\Delta r\right]\times\mathscr{F}^{-1}\left\{\left[\exp\left(-\frac{i\Delta r}{2K_0}s^2\right)\right]\times\mathscr{F}[\psi(r,z)]\right\}, \tag{58}$$

which is often referred to as the "split step" marching solution to the PE. The Fourier transforms are performed using FFTs. Equation (58) is the solution for n constant, but the error introduced when n (profile or bathymetry) varies with range and depth can be made arbitrarily small by increasing the transform size and decreasing the range-step size.

Methods of solving the parabolic equation, including extensions of it to higher angle propagation, elastic media, and direct time domain solutions including nonlinear effects, have recently appeared (see Kuperman and McDonald, 1986; Lee and McDaniel, 1988; Collins, 1989, from which an extensive set of appropriate references can be obtained).

3.7 Model Consistency

All of the models described above attempt to solve in one way or another, with and without assorted approximations, the Helmholtz equation. They therefore should be consistent and there is much insight to be gained from understanding this consistency (Kuperman and Jensen, 1980).

Figure 10 shows convergence zone type propagation for a simplified profile. The ray trace in Fig. 10(b) shows the cyclic focusing discussed in Sec. 3. The same profile is used to calculate normal modes as shown in Fig. 10(c), which, when summed according to Eq. (46), exhibits the same cyclic pattern as the ray picture. Figure 10(d) shows both the normal mode (wave theory) and ray theory result. Ray theory exhibits sharply bounded shadow regions as expected whereas the normal mode theory, which includes diffraction, shows that the acoustic field does exist in the shadow regions, and the convergence zones have structure.

Normal mode models sum the discrete modes which roughly correspond to angles of propagation within the cone of Fig. 6. The FFP solution can include the full field, dis-

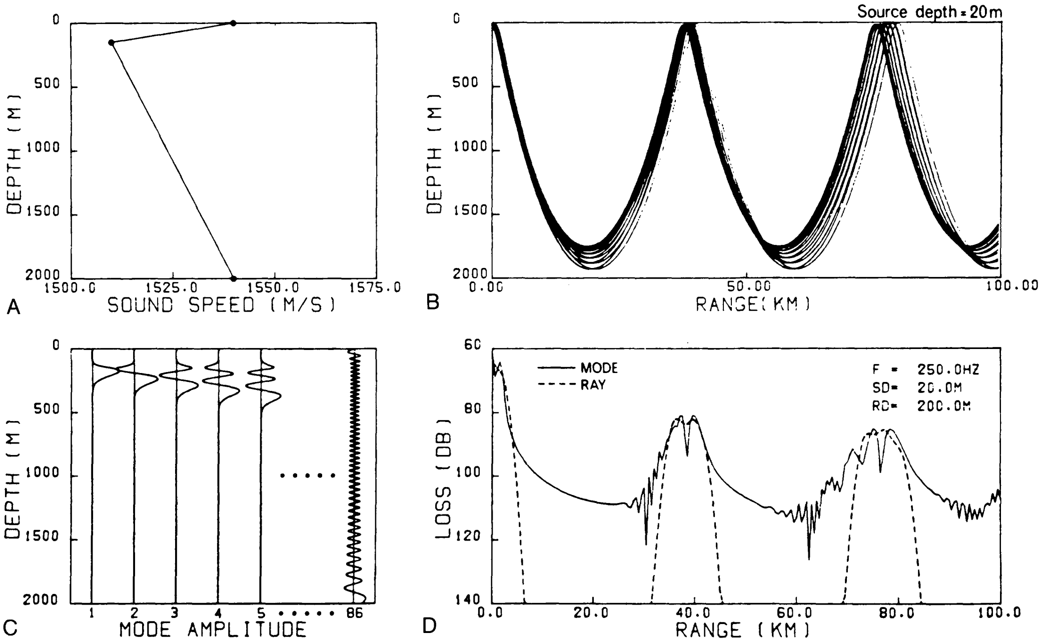

FIG. 10. Consistency between ray theory and normal mode theory. (a) Sound speed profile. (b) Ray trace plot. (c) The normal modes for this profile. Note that the lowest order modes are trapped in the duct around the sound speed minimum. (d) Normal mode and ray theory propagation calculations.

crete plus continuous, the latter corresponding to larger angles. The discussion below Eq. (46) defines these angles in terms of horizontal wave numbers, and the eigenvalues of the normal mode problem are a discrete set of horizontal wave numbers. Hence the integrand (Green's function) of the FFP method has peaks at the eigenvalues associated with the normal modes. These peaks are shown on the right of Fig. 11(a). The smoother portion of the spectrum is the continuous part corresponding to the larger angles. Therefore, the consistency we expect between normal mode and FFP and the physics of Fig. 6 is that the continuous portion of the FFP decays with range so that there should be complete agreement at long ranges between normal mode and FFP. The Lloyd's mirror effect, a near field effect, should also be exhibited in the FFP solution but not the normal mode solution. These aspects are apparent in Fig. 11(b). The PE solution appears in Fig. 11(b) and is in good agreement with the other solutions but with some phase error associated with the average wave number that must be chosen in

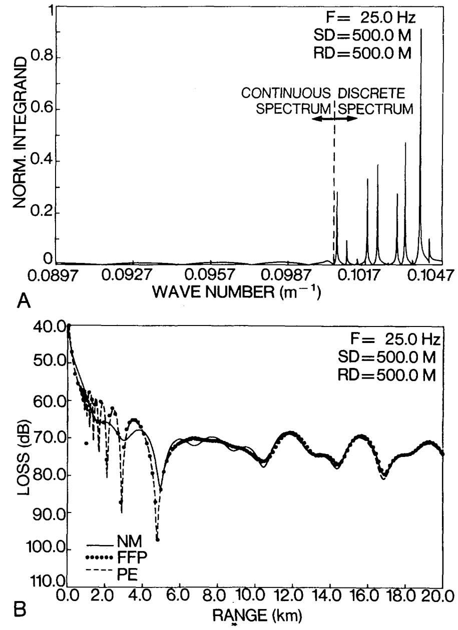

FIG. 11. Relation between full field (FFP), normal mode (NM), and parabolic equation (PE) computations. (a) A Green's function (integrand) of the FFP method. (b) Normal mode, parabolic equation and FFP propagation results showing some partial agreement of the latter two in the near field and agreement between all three in the far field.

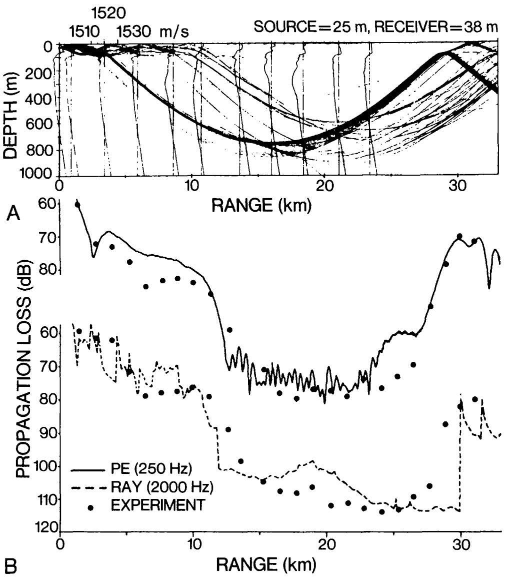

FIG. 12. Model and data comparison for a range-dependent deep water case (Kuperman and Jensen, 1980). (a) Sound speed profiles as a function of range and ray trace for the case of a surface duct disappearing. (b) PE comparison with data at 250 Hz and ray theory comparison with data at 2 kHz.

the split step method. The PE solution, which contains part of the continuous spectrum including the Lloyd's mirror beams, is more accurate than the normal mode solution at short range.

Range dependent results are shown in Fig. 12. A ray trace, a ray trace field result, a PE result, and data are plotted together for a range dependent sound speed profile environment. The models agree with the data in general with the exception that the ray results predict too sharp a leading edge of the convergence zone.

Upslope propagation is modeled with the PE (Jensen and Kuperman, 1980) in Fig. 13. As the field propagates upslope, sound is dumped into the bottom in what appears to be discrete beams. The flat region has three modes, and each is cut off successively as sound propagates into shallower water. The ray picture also has a consistent explanation of this phenomenon. The rays for each mode become steeper as they propagate upslope. When the ray angle exceeds the critical angle,

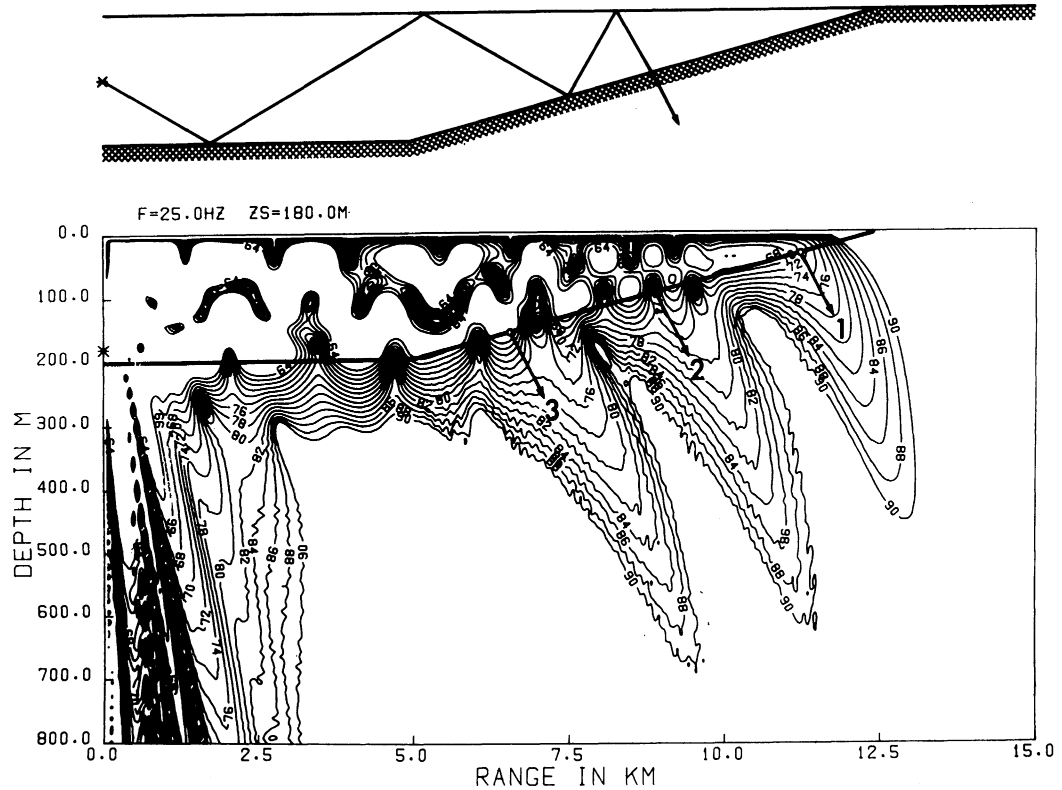

FIG. 13. Relation between upslope propagation (as calculated with the PE), showing individual mode cutoff and energy dumping in the bottom, and a corresponding ray picture.

the sound is significantly transmitted into the bottom. The locations where this takes place for each of the modes are identified by the three arrows.

4. SPATIAL PROPERTIES OF SURFACE GENERATED NOISE

The distribution of sound from surface noise sources can be calculated by coupling these noise sources into the water column using any of the above mentioned propagation models together with Green's theorem. We then would include both the physics of the noise sources and the environmental propagation effects. Figure 14 shows the geometry for such a model. The distribution of the field is characterized by the two-point "cross-spectral density," which reduces to the local sound intensity when the two field points are at the same position. Otherwise, this quantity is representative of the directional/spatial properties of the noise field (Kuperman and Ingenito, 1977) which would be measured by a distributed acoustic receiving array. The cross-spectral density is (⟨ ⟩ is the "expected value" symbol)

$$\langle \phi(\mathbf{r}_1, z_1)\phi^*(\mathbf{r}_2, z_2)\rangle$$

$$= \int\int d^2\mathbf{r}' d^2\mathbf{r}'' \langle S(\mathbf{r}')S^*(\mathbf{r}'')\rangle \times G(\mathbf{r}_1, \mathbf{r}', z_1, z')G^*(\mathbf{r}_2, \mathbf{r}'', z_1, z''), \qquad (59)$$

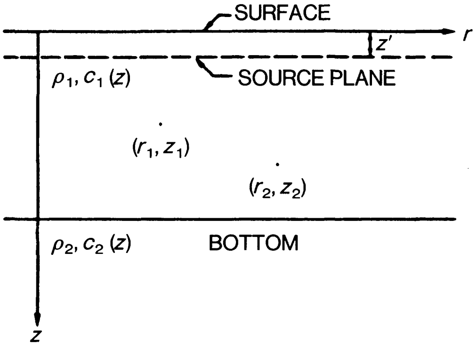

FIG. 14. Surface noise model geometry showing the noise source plane, at depth z′, below the surface, and the two field points.

where S is the surface noise source function and z'' is at the same depth as z'. The correlation function of the surface noise sources, which is the expectation quantity in the integrand of Eq. (50), is taken to be spatially dependent only on the separation ($\mathbf{s} \equiv \mathbf{r}' - \mathbf{r}''$) between sources. We denote this correlation function as $\langle S(\mathbf{r}')S^*(\mathbf{r}'')\rangle \equiv q^2 N(s)$, with q being surface source strength. We also express the noise field points through the difference of displacement vectors, $\mathbf{R} \equiv \mathbf{r}_1 - \mathbf{r}_2$, and denote the cross-spectral density given by Eq. (59) for angular frequency ω as $C_\omega(\mathbf{R}, z_1, z_2)$. Substitution in the full field representation of G given by Eqs. (37) and (38) into Eq. (59) and integration over the azimuth angle associated with k gives

$$C_\omega(\mathbf{R}, z_1, z_2)$$

$$= 2\pi q^2 \int ds\, N(s) \int k\, dk\, g(k, z_1, z') \times g^*(k, z_2, z') J_0(k|\mathbf{R} - \mathbf{s}|), \qquad (60)$$

where J_0 is the Bessel function of zero order. An alternative form to Eq. (60), which is particularly useful because it can directly call upon the algorithms associated with the FFP approach presented above, involves expressing the correlation function of the surface noise sources in terms of its spatial spectral representation $P(\mathbf{k})$. Then, if we substitute into Eq. (60) for N its Fourier transform,

$$N(\mathbf{s}) = \frac{1}{2\pi} \int d^3k\, P(\mathbf{k}) \exp(i\mathbf{k}\cdot\mathbf{s}), \qquad (61)$$

together with the other substitutions used to get Eq. (60), we obtain

$$C_\omega(\mathbf{R}, z_1, z_2)$$

$$= 2\pi q^2 \int d^3k\, P(\mathbf{k}) \exp(i\mathbf{k}\cdot\mathbf{R}) \times g(\mathbf{k}, z_1, z') g^*(\mathbf{k}, z_2, z'). \qquad (62)$$

Depending on the type of numerics being applied to a specific problem, Eqs. (60) and (62) are suitable for model computations. Note that the noise cross-spectral density evaluated at the same field points, $C_\omega(0, z, z)$, is the intensity of the noise at depth z for this range-independent formulation of the distribution of noise.

It has been shown (Liggett and Jacobsen, 1965) that the correlation function of surface sources which produces a far field radiation (of multipole order m) pattern $\cos^m\theta$ with respect to the vertical is given by

$$N(s)=\begin{cases}\dfrac{2\delta[K(z')s]}{[K^2(z')s]}, & m=1,\\[2ex] 2^m m!\,\dfrac{J_m[K(z')s]}{[K(z')s]^m}, & m>1.\end{cases} \tag{63}$$

In simple general terms, we know from the literature that many of the noise sources can be categorized as and are associated with (Ross, 1976):

1. monopole: mass addition, heat addition, volume change, compression—e.g., bubbles, rain droplet impact, etc.;
2. dipole: force, translation, acceleration (sloshing)—e.g., vibration of unbaffled rigid bodies;
3. quadrupole: moment, shear, distortion, rotation, turbulence.

These noise sources are then taken to be on the source plane in Fig. 14, where the surface is pressure released so that the radiation couples into the water column as a higher order pole radiator; for example, monopole sources near the surface couple as dipoles just by using the ocean sound propagation Green's function, which satisfies the pressure release condition. For this case we then use $m=1$ in Eq. (63), a dipole sheet at the surface corresponding to a monopole plane below the surface, which gives the $\cos^m\theta=\cos\theta$ radiation pattern with respect to the vertical. Then, the noise distribution of sound in the ocean from surface sources can be constructed from the above expressions using the propagation models described in Sec. 4 for G and the noise source distributions expressed in the multipole form of Eq. (63).

Of interest is a particular simplification when the noise can be modeled in a range independent environment and an expression for the distribution of noise is sought in terms of normal modes. Then, that part of the noise that corresponds to discrete propagation from mainly far away sources and corresponds to the inner cone of Fig. 6 and the inequality of Eq. (43) can be derived from Eqs. (60) and (46). A simple result can be obtained for this important case, with $m=1$ as given by Eq. (63). After a complex integration, we get the result for the discrete portion of the noise field (suppressing the subscript ω):

$$C_{\text{disc}}(\mathbf{R}, z_1, z_2)\approx\frac{\pi q^2\rho^2(z')}{2K^2(z')}\times\sum_n\frac{u_n^2(z')u_n(z_1)u_n(z_2)}{\alpha_n k_n}\times J_0(k_nR), \tag{64}$$

where α_n is the attenuation coefficient of the nth mode and, as described in Sec. 4, is the imaginary part of the eigenvalues resulting from the existence of loss in the system. It is clear from the physics of this problem that contributions from noise sources long distances away in a waveguide with only cylindrical spreading are unbounded if there is no loss in the system; hence, the attenuation coefficients in the denominator of Eq. (64) are essential.

As discussed in Sec. 3, the ocean is a waveguide. We can therefore draw some general conclusions about how sound sources at the upper boundary of a waveguide will couple into the waveguide. As mentioned earlier, the ray and mode pictures provide the most straightforward physical insight into the actual coupling process. First, for reference, let us consider the noise field in an unbounded semi-infinite isovelocity medium with constant distribution of monopole sources radiating into the medium from the upper boundary. Summing all the sources over an infinite area and noting that in this situation we have spherical spreading, it is easy to show that the resulting sound field will be finite and its directional properties will be isotropic in the upward direction (the only direction from which sound is coming) and the intensity of the sound will be independent of position. All these characteristics come from three simple considerations:

1. the sources are spherically symmetric;
2. the contributing source area goes up as the square of the radius;
3. the field of the sources decays, by spherical spreading.

However, cylindrical spreading occurs in the ocean and large areas from very long distances will contribute significantly to local noise [even for the real, lossy ocean, as discussed below Eq. (64)]. In a waveguide it makes sense that noise coming directly from above looks like the semi-infinite unbounded case discussed in the last paragraph whereas noise from long distances has the ocean channel properties discussed in Secs. 3 and 4.

With this background of sound propagation in the ocean and some intuition of how sources on a surface plane couple either into a semi-infinite space (spherical spreading) or into a waveguide (cylindrical spreading), we can understand much of how ambient noise is distributed in the ocean. Figure 15 depicts the possible propagation regimes by which surface generated noise can arrive at a receiver near the ocean bottom (where seismic paths might also be important). The continuous part is the vertical structured noise that is weakly affected by the sound speed profile. The discrete (or channeled) part corresponds to noise arriving horizontally from, typically, distant sources. We can write the continuous portion of the noise field as the right-hand side of Eq. (62), where the limit on the integration is $|k| < \min[K(z)]$ [see discussion below Eq. (46)]; we define this quantity to be C_{cont}. The total cross-spectral density of the noise field, which is also referred to as the spatial correlation or spatial coherence of the noise field, is given by

$$C_\omega(\mathbf{R}, z_1, z_2) = C_{\text{cont}}(\mathbf{R}, z_1, z_2) + C_{\text{disc}}(\mathbf{R}, z_1, z_2). \tag{65}$$

Equations (60) and (64) (which is in the discrete part of C) indicate that the noise field varies with depth; the cross-spectral density of the noise is a function of both the mean depth of the field points and the absolute depth location of the field points (i.e., noise is not homogeneous in depth). The cross-spectral density of the noise is homogeneous in range for a range-independent environment since there is no reference range. These conclusions are particularly true for horizontally stratified continental shelf areas in which waveguide theory is most applicable.

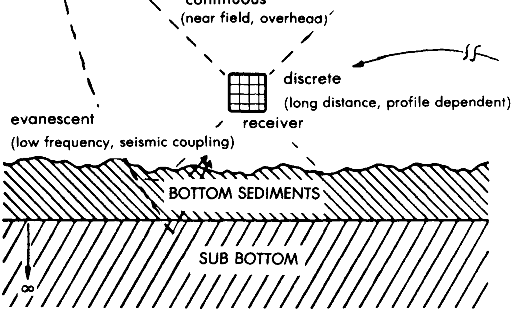

FIG. 15. Surface generated noise propagation paths to an acoustic receiver.

Figure 16(a) shows the above mentioned vertical inhomogeneity in a waveguide type environment by plotting the cross-spectral density of the noise related to receivers located at two different depths. The noise at receivers R_1 and R_2 is correlated with the noise at all other field points of the water column. We see that the two correlation functions appear different indicating the spatial inhomogeneity.

Also contrary to simplified noise models more applicable to high frequency, deep water environments (Cron and Sherman, 1962), the cross-spectral density of noise in a stratified waveguide should have a different structure than noise in a homogeneous semi-infinite medium. Figure 16(b) shows the modeled vertical spatial correlation (cross-spectral density for a receiver at the same horizontal location but different vertical positions) between a receiver located at 10 m depth and another receiver separated by distance D in units of wavelength λ. The particular environment for this case involves a 50 m water depth and an upward refracting sound speed profile. Here we see that the vertical correlation of the noise versus separation in units of wavelength is a function of frequency, in contrast to the simpler model for a semi-infinite medium where the spatial correlation scales with wavelength. This difference arises from the additional scales that are introduced when considering the stratified environment. Note also that the noise is more correlated over the water column than the simple case (the solid curve) because a few lower-order modes dominate the propagation due to upward refraction, and therefore there is very little interaction with the lossy bottom for the horizontal long distance contribution as opposed to the more vertical lossy propagation. That is, the noise tends to come from the horizontal rather than from all directions thereby increasing the spatial correlation.

Figure 16(c) shows the results for the horizontal correlation for the same kind of environment as for the case in Fig. 16(b) except

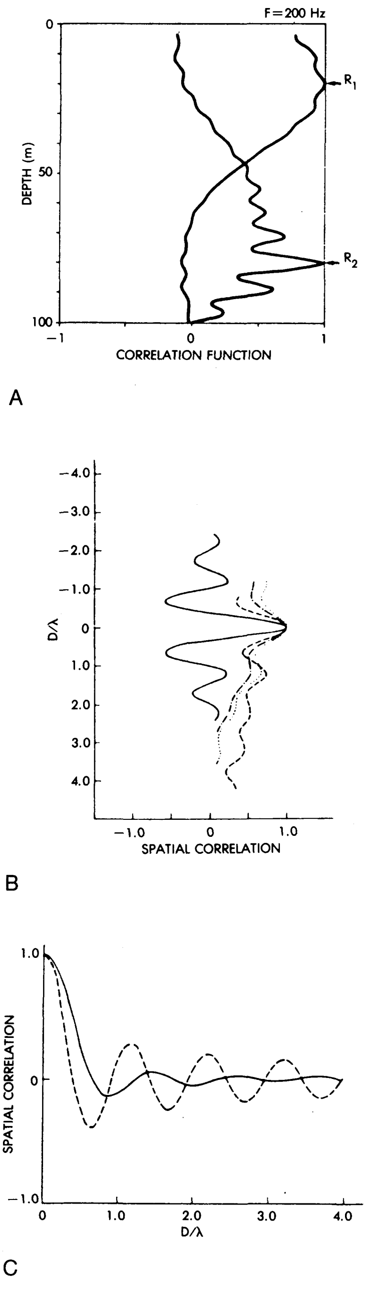

FIG. 16. Spatial distribution of noise (Kuperman and Ingenito, 1980). (a) Normalized spatial correlation in a waveguide environment for two different receiver depths. (b) Normalized vertical spatial correlation functions for an upward refracting environment with one receiver fixed at 10 m depth in a 50 m deep water column. Three frequencies are 200 (---), 400 (···), and 800 Hz (-·-·-). The solid line is the semi-infinite homogeneous result (Cron and Sherman, 1962). (c) Horizontal correlation. The solid line is the continuous contribution and the dashed line is the discrete contribution.

that the sound speed is constant in the water column. Here, we plot the discrete and continuous contributions separately. The continuous (vertical) part is more coherent for small receiver separations but quickly becomes less coherent (and looks like the simpler semi-infinite case) than the discrete part. Again, the physics of the propagation environment indicates the relative importance of the discrete and continuous parts. For low loss, the discrete modes dominate as they propagate very large distances from a very large area [see discussion below Eq. (64)]. In high loss cases, the continuous modes tend to dominate since they are important near the surface; the long range contributions of the discrete are severely attenuated. Therefore, a more vertical structure is expected for noise in a lossy environment.

In Secs. 2 and 3 we discussed the impact of the deep water sound speed structure on propagation. In particular, we mentioned in Sec. 3 the existence of a critical depth in the deep isothermal layer. For such a case we expect from the above discussion that, when there is a positive critical depth, sound from surface sources can travel long distances without interacting with the ocean bottom but that a receiver below this critical depth should sense less surface noise because any propagation from distant shallow sources to these depths involves interaction with the lossy boundaries, surface, and/or bottom. Indeed, Fig. 17 illustrates this effect using data from the Northeast Pacific (Morris, 1978). The noise in this figure comes from a combination of natural sources and shipping, but the propagation effects for these surface sources are the same. We should note here, also, that higher frequency noise (especially the multi-kHz region and above) diminishes with depth simply because volume attenuation increases with frequency.

The vertical directionality of deep water noise should also vary with depth as discussed above. Figure 18 shows some depth dependence of the directionality of low frequency noise in the Pacific (Anderson, 1979). The shallower depth is at the axis of the deep sound channel while the other is at the critical depth. We see, as expected, that the pattern is narrower at the critical depth where the sound paths tend to be horizontal since the rays are turning around at the lower boundary of the deep sound channel.

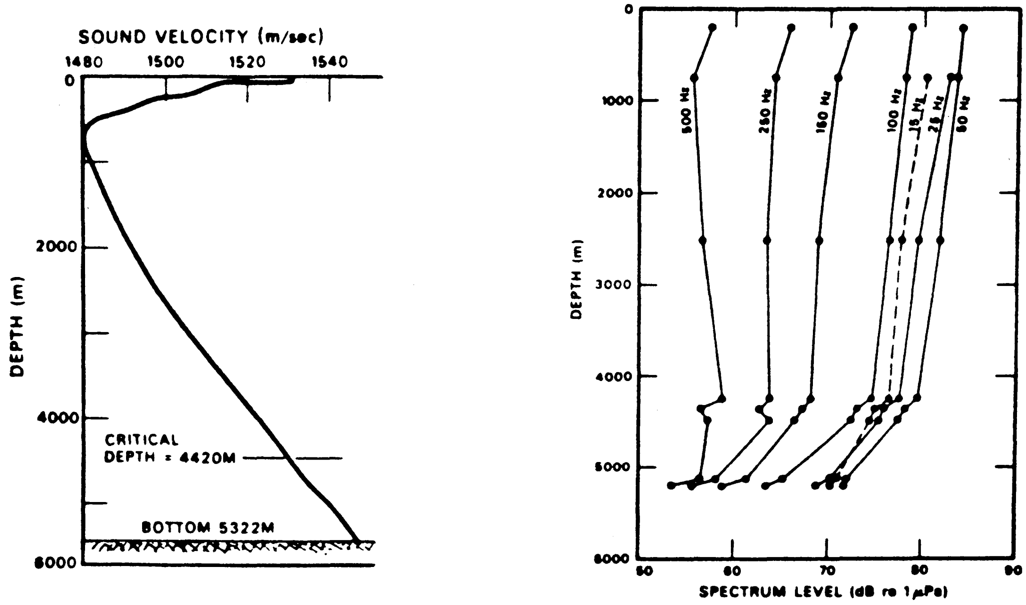

FIG. 17. Sound speed profile and noise level as a function of depth in the Pacific (from Morris, 1978).

In deep water, sound from surface sources would arrive, on the average, off the horizontal (~ ±12° to ±15°) if the ocean did not have a range varying environment. But, as explained in Sec. 4.4, noise from continental shelf areas tends to have paths which become more horizontal as the sound propagates to deeper water; an analogous effect is that polar latitudes have shallow sound channels so that sound propagating from polar latitudes will go deep and horizontal in the midlatitude regions. Hence, the accurate modeling of this horizontal component of the noise (Wagstaff, 1981) requires the use of range-dependent models. Adiabatic mode theory includes this effect.

Finally, we mention seismic noise. Very low frequency sound will couple directly into the ocean bottom. The surface generated "microseismic" noise (Akal *et al.*, 1988; Kibblewhite and Ewans, 1985) also obeys the basic physics of waveguide propagation we have discussed. Briefly, when the ocean surface sources are close (in terms of wavelength) to the ocean bottom, their evanescent field will couple into seismic waves. This coupling occurs below the acoustic cutoff frequency of the waveguide (Schmidt and Kuperman, 1988). Seismic noise shows a spectral peak around 0.25 Hz which is thought to originate from nonlinear surface wave-wave interactions corresponding to the surface wave peak of about 0.12 Hz.

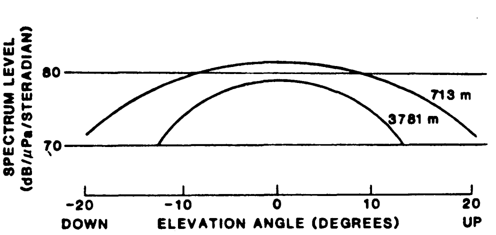

FIG. 18. Vertical directionality of noise at the axis of the deep sound channel and at the critical depth (from Anderson, 1979).

5. ARRAYS AND ARRAY GAIN FOR SIGNALS IN NOISE

An array of hydrophones constitutes an acoustic antenna which can be used to enhance the signal to noise ratio at a receiver. This enhancement is often referred to as array gain and is realized in three interrelated ways. First, the level of the signal at the receiver can be increased by summing the outputs of the hydrophones coherently. This coherent sum can be related to a "look direction" and a "beam width." Recall that the coherent sum refers to adding (complex) amplitudes, and the total energy is the square of the sum of amplitudes. Second, the incoherent part of the noise, the white noise, adds incoherently so that the energies, not the amplitudes, add. Finally, the coherent part of the noise has directional properties so that an appropriately phased coherent sum of hydrophone outputs can be constructed. This limits the directional noise input by looking away from the noise and further limits the noise input in the direction of the noise by the finite beamwidth associated with the look direction. We can illustrate these ideas quantitatively as follows: Consider the simple case of an m-element array with signal $s_i(t)$ and noise $n_i(t)$ at the ith element of the array. The average signal to noise ratio at the output of the array will be (the angular brackets denote time average)

$$\frac{S^2}{N^2}=\frac{\left\langle\left[\sum_{i=1}^{m} S_i(t)\right]^2\right\rangle}{\left\langle\left[\sum_{i=1}^{m} n_i(t)\right]^2\right\rangle}. \qquad (66)$$

Summing the outputs of the hydrophones using various phase and amplitude shading is called beamforming. The simplest example of array gain is the situation in which the signal is a plane wave traveling in one direction and the noise field is isotropic (i.e., it has components in all directions). A single omnidirectional hydrophone receives all the signal and noise. An array of hydrophones can be used to sum the signal coherently by phasing the array to look in the direction of the plane wave. Hence the signal output of the array is the coherent sum of the signal at the hydrophones. On the other hand, only the noise in the look direction (aside from white noise) is output from the array beamformer with the total energy of the noise related to the beamwidth. Hence an extremely narrow beam picks up all the signal but only a portion of the noise, and there is an enhancement of ratio of signal to noise relative to an omnidirectional receiver. Below, we briefly derive these results. The ratio of the average signal power to the average noise power at the output of a hydrophone array can be written as

$$\frac{S^2}{N^2}\equiv\frac{\left\langle\left[\sum_{i=1}^{m} s_i(t)\right]^2\right\rangle}{\left\langle\left[\sum_{i=1}^{m} n_i(t)\right]^2\right\rangle}=\frac{\sum_{i,j=1}^{m} s_{ij}}{\sum_{i,j=1}^{m} n_{ij}}, \qquad (67)$$

where i and j subscripts imply a double summation, and the array signal and noise crosscorrelation coefficients between the ith and jth hydrophones are given by $s_{ij} \equiv \langle s_i(t)\hat{s}_j^*(t)\rangle$ and $n_{ij}\equiv\langle n_i(t)\,\hat{n}_j^*(t)\rangle$, respectively.

In the simple and common case where the signal and noise have the same amplitude at each hydrophone, we define $s_{ij}\equiv s^2\hat{s}_{ij}$ and $n_{ij}\equiv n^2\hat{n}_{ij}$, where the carets denote normalized crosscorrelation coefficients. For this case, Eq. (67) simplifies to

$$\frac{S^2}{N^2}=\frac{s^2}{n^2}\,\frac{\sum_{i,j=1}^{m}\hat{s}_{ij}}{\sum_{i,j=1}^{m}\hat{n}_{ij}}, \qquad (68)$$

where, consequently, the second fraction on the right side of Eq. (68) is a measure of the array performance with respect to increasing the signal to noise ratio over a single omnidirectional hydrophone. Array gain (AG) is defined in decibels as

$$\mathrm{AG}=10\log\frac{S^2/N^2}{s^2/n^2}=10\log\frac{\sum_{i,j=1}^{m}\hat{s}_{ij}}{\sum_{i,j=1}^{m}\hat{n}_{ij}}, \qquad (69)$$

so that the array gain is a function of the signal and noise structure projected onto the geometry of the array. The directivity index (DI) is a special case of the array gain formula. DI equals AG when the signal is a single plane wave and the noise is isotropic (i.e., the noise is uniform distribution of plane waves in all directions and is obtained by simple integration of a constant over 4π solid angle). For this case, $\hat{n}_{ij}$ is stationary in space (only related to hydrophone separation distance) and is given by $\hat{n}_{ij} = (kd)^{-1}\sin(kd)$ for $i\neq j$, where d is the hydrophone separation and $k=2\pi/\lambda$ is the acoustic wavenumber. Note for isotropic noise, and an array hydrophone spacing of $d=\lambda/2$, the noise appears uncorrelated (i.e., cross terms vanish). For the purpose of illustration, we consider a two-element array. For this case the (coherent) sum of the numerator is 4 and the denominator is $2+2(kd)^{-1}\sin(kd)$, so that the DI of a two element array is 10 log $\{2/[1+(kd)^{-1}\sin(kd)]\}$, or is 10 log 2 if the hydrophone separation is $\lambda/2$. An important generalization for DI is that a linear array of m hydrophones with $\lambda/2$ spacing in isotropic noise has DI$=$10 log n.

Working in the frequency domain, we can shade the array by amplitude and phase factors at each hydrophone. The simplest example is phase shading to look in a specific direction (i.e., to match a plane wave coming from the look direction). This is referred to as steering the array, and this type of linear beamformer is often referred to as the Bartlett beamformer. We take θ_s (which in general is a two-dimensional angle) as the beamsteered angle or the look direction. Referring to Fig. 19, a plane wave $\exp(i\mathbf{k}\cdot\mathbf{R})$ across the array is summed in phase if the hydrophone at position $\mathbf{d}_i$ is multiplied by the factor $w_i=\exp(-i\mathbf{k}\cdot\mathbf{d}_i)$. The array gain does not change since both signal and noise are altered by this same factor. The output of this beamforming process $B(\theta_s)$ is a quadratic form obtained by summing the shaded signal plus noise from each hydrophone:

$$B(\theta_s)=\sum_{i,j=1}^{m} w_i^*(\theta_s)(s_{ij}+n_{ij})w_j(\theta_s). \tag{70}$$

If the field of the source at the receiving array is a plane wave in a particular look direction, the beamformer has its peak in that direction; however, there are also sidelobes because of the finite sampling of the acoustic field by the array.

We can write the above expression in matrix notation where the boldface lower case letters denote vectors and boldface upper case letters denote matrices. Define a steering column vector $\mathbf{w}$ whose ith element is w_i and which has the dimension of the number of array elements, and a cross-spectral density matrix $\mathbf{K}$ of the signal and noise with elements $s_{ij}+n_{ij}$. Then Eq. (70) can be simply written as

$$B(\theta_s)=\mathbf{w}^\dagger(\theta_s)\mathbf{K}(\boldsymbol{\theta}_{\text{true}})\mathbf{w}(\theta_s), \tag{71}$$

where "$\dagger$" denotes the complex transpose operation, and the data across the array as represented in the matrix $\mathbf{K}$ contain the information that the source is in the direction θ_{true}. This form of the beamforming equation has a succinct generalization to matched field processing. Often, $\mathbf{w}(\theta_s)$ is referred to as a replica, and the above beamforming process is then matching the received data across the array $\mathbf{K}$ with a replica. In the above case, the replica is a plane wave in the look direction. Now take $\mathbf{w}$ as solution to the acoustic wave equation as discussed in Sec. 4. It then represents a replica field at each element of the array for a specific ocean environment and for an arbitrary source location $\mathbf{x}$. Rather than find just the direction of the source from the array, we can search for its actual location by matching the received data, $\mathbf{K}(\mathbf{x}_{\text{true}})$ from the true source location, against solutions of the wave equation. The output of this matched field processor, denoted $S(\mathbf{x})$, at each point in space $\mathbf{x}$ is in analogy to Eq. (71) (Baggeroer *et al.*, 1988):

$$S(\mathbf{x})=\mathbf{w}^\dagger(\mathbf{x})\mathbf{K}(\mathbf{x}_{\text{true}})\mathbf{w}(\mathbf{x}), \tag{72}$$

where the peak of the output of the beamformer $S(\mathbf{x})$ is also referred to as the ambiguity function (or surface) of the matched field processor because it also contains ambiguous peaks analogous to the sidelobes of a conventional plane wave beamformer. Sidelobe suppression can often be accomplished by using a nonlinear beamformer such as the Maximum Likelihood Method (MLM) beamformer:

$$S_{\text{MLM}}(\mathbf{x})=[w^\dagger(\mathbf{x})\mathbf{K}^{-1}(\mathbf{x}_{\text{true}})w(\mathbf{x})]^{-1}, \tag{73}$$

where the above beamformer was first derived for the plane wave beamformer, as discussed above, so that $\mathbf{x}$ refers to angles or spatial coordinates (or any other unknown parameter of the problem to be estimated, e.g., bottom type!). This beamformer gives the same peak level output at the true source location as the Bartlett beamformer.

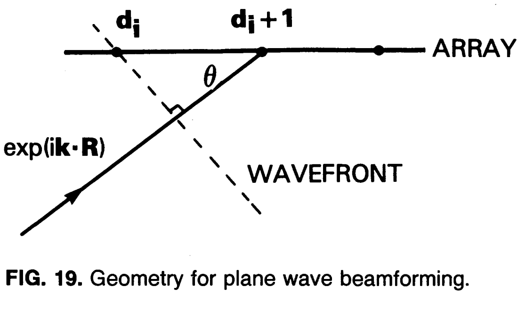

FIG. 19. Geometry for plane wave beamforming.

6. SIGNAL PROCESSING AND SONAR SYSTEMS

The main application of the field of underwater acoustics is to sonar system technology. There are two fundamental types of sonar systems: passive and active. Passive sonars listen for signals radiated by objects for the purpose of detection, classification, and localization of these acoustic sources. These signals are embedded in ocean noise. The performance of a passive sonar system requires sufficient signal above a threshold, such that the requisite signal information is extracted from the background noise. To give a simple example, a received signal of known source level can be used to estimate the distance between receiver and source if propagation conditions are known.

For the case of an active system, the sonar transmits a signal which is reflected or scattered from an object or ocean feature, and hence the modified signal carries information concerning the object or feature to a receiver (often colocated with the transmitter). This information must be extracted from the background interference, which for the active sonar is not only noise but also, and often more important, the reverberation generated by the original transmitted signal. The simplest type of active sonar system information is contained in a travel time measurement of an emitted and returning pulse from which the distance to an object or a feature (i.e., the ocean bottom for the case of a fathometer measuring water depth) can be estimated.

In either case, passive or active, the performance of a sonar system can to a large extent be described by a sonar equation. In this section, we shall introduce the sonar equation for passive and active systems. Because it is customary to use decibels in underwater acoustics, the forms of sonar equations are such that individual phenomena contributing to the performance of a sonar appear as additive terms and normally can be treated separately. Therefore all terms in the sonar equations presented below are expressed in decibels. The standard reference for this material is Urick (1983). More accurate representation of sonar system performance requires "simulation" of the relevant acoustic phenomena and subsequent insertion of these fields into a (computer) model of the system.

6.1 Passive Sonar Equation

A radiating object of source level "SL" (all units are expressed in dB) will be received at a hydrophone of a sonar system at a lower signal level "S" because of the transmission loss "TL" it suffers; hence,

$$\mathrm{S}=\mathrm{SL}-\mathrm{TL}, \tag{74}$$

where TL is a positive quantity; it is the negative of propagation loss which is defined to be 20 log (amplitude of the field from a unit source). Any of the propagation models in Sec. 4 or simpler geometric spreading laws can be used to estimate the field. The noise at a single hydrophone "N" is subtracted from Eq. (74) to obtain the signal-to-noise ratio at a single hydrophone:

$$\mathrm{SNR}=\mathrm{SL}-\mathrm{TL}-\mathrm{N}. \tag{75}$$

A sonar system normally consists of an array of hydrophones which are beamformed to obtain signal to noise enhancement. The directivity index "DI" [see discussion below Eq. (69)] is added to Eq. (75) to obtain the signal-to-noise ratio at the output of the beamformer:

$$\mathrm{SNR}_{\mathrm{BF}}=\mathrm{SL}-\mathrm{TL}-\mathrm{N}+\mathrm{DI}. \tag{76}$$

There exists a threshold SNR above which a sonar system will, with 50% probability (by convention), accomplish its detection purpose. Ideally, detection should occur when the signal is not marked by noise, i.e., $\mathrm{SNR}>0$, but normally because of the nature of the sonar, the signal, the interference, the operator training and alertment, and an assortment of other factors, a certain additional signal level above zero $\mathrm{SNR}_{\mathrm{BF}}$ is referred to as the recognition differential "RD" or the detection threshold "DT." Hence, if we include a term related to this detection threshold, we can write a sonar equation for the signal excess "SE," the difference between the actual received signal at the output of the beamformer and the minimum signal for detection (i.e., extra signal above what is needed for detection):

$$\mathrm{SE}=\mathrm{SL}-\mathrm{TL}-\mathrm{N}+\mathrm{DI}-\mathrm{DT}, \tag{77}$$

where even though all the quantities are in decibels, only SL and N must be defined with respect to the same reference pressure and frequency bandwidth (e.g., 1 Hz or a fraction of an octave such as the standard 1/3 octave filter convention). In any event, we will see below that an SE > 0 can be used to increase our effective probability of detection *p(D)*, which is defined as the probability that an existing signal will be detected.

Another useful quantity the sonar engineer uses is the figure of merit "FOM" which is the transmission loss that gives zero signal excess for a $p(D)=50\%$, i.e., the maximum allowable transmission loss for a detection. Hence, from Eq. (77),

$$\mathrm{FOM}=\mathrm{SL}-\mathrm{N}+\mathrm{DI}-\mathrm{DT}, \tag{78}$$

and we see that the FOM is a transmission loss number which actually encompasses target source level and sonar system parameters. Therefore, a specification of an FOM can be translated into a detection range if the propagation conditions (loss versus range) are known.

We note that the above sonar equations are for simplistic situations and must be modified accordingly to be more realistic. For example, in reality the signal and noise levels have vertical and azimuthal dependence so that they are a function of beam steered angle θ_s, and the signal is also likely to have a dispersed arrival structure because of multipath type propagation. Furthermore, the response of the sonar array is a function of θ_s. All of these complications can be approximated either by a more sophisticated sonar equation type analysis or by more precise combined model simulations of propagation, noise, and sonar signal processing packages. The latter requires considerably more detailed knowledge of the ocean acoustic environment and the sonar system under study.

Continuing with the simple sonar equation, we note that DT includes all the signal and noise processes associated with the sonar, including the operator, except for the array gain. Hence, we are in the realm of probabilistic phenomena. The detection threshold DT, in dB, is the ratio of the signal power S in the receiver bandwidth to the noise power N_0 in a 1-Hz band required for detection according to some detection decision criterion.

To understand the effect of detection threshold setting, we consider two possible cases. If the detection criterion is set too stringently and the DT is therefore high, the probability of detection is low. However, the probability of false alarm (the probability that noise crosses the detection threshold, DT) is also low. In this case only targets with high source level will be detected. Alternatively, for a low DT, low level targets are detected but there is a high probability of false alarm.

Also helpful in quantitatively characterizing the detection process are Receiving Operating Characteristic (ROC) curves, which are plots of probability of detection versus probability of false alarm using a detection index *d* as a parameter (Urick, 1983). The detection index is the standard deviation of the mean amplitude of the signal plus noise M_{S+N} about the mean noise amplitude M_N, normalized to the variance of the signal plus noise or noise alone. Both of the latter are assumed to be Gaussian with equal variance σ^2:

$$d=\frac{(M_{S+N}-M_N)^2}{\sigma^2}. \tag{79}$$

To understand further the utility of DT, we consider two extreme cases: Case 1 is for a completely known signal where the optimum receiver is a crosscorrelator (Urick, 1983); it is found that

$$\mathrm{DT}\equiv 10\log\frac{S}{N_0}=10\log\frac{d}{2t}, \tag{80}$$

where N_0 is the noise power spectrum in a 1-Hz bandwidth and t is the duration of the signal with power S. Case 2 is for an unknown weak signal in Gaussian noise and a long sample size (large time-bandwidth product) where the optimum receiver is a square-law detector (output proportional to the square of the input); it is found that

$$\mathrm{DT}\equiv 10\log\frac{S}{N_0}=5\log\frac{d\Delta f}{t}, \tag{81}$$

where Δf is the bandwidth of the signal with power S. However, the integration time T of the sonar system is an important additional factor: if it is too large, the signal peak is oversmoothed downward, hence increasing DT; if it is too small, the increased bandwidth also increases DT. Therefore, a filter mismatch to the signal always increases DT. In general, the time-bandwidth product is a very important parameter characterizing sonar performance. The detection threshold for the mismatched filter case can be estimated from a modified form of Eq. (81):

$$\mathrm{DT}=5\log\frac{d\Delta f}{t}+\left|5\log\frac{T}{t}\right|. \qquad (82)$$

Another important factor is that one normally tries to detect a signal over a number of independent (incoherently added) samples in order to increase the chances of detection. The incoherent aspect adds a 5 log n term to the received power and hence lowers the DT by that amount.

In reality, the performance of a sonar system has a more complicated behavior than the simplistic discussion above. For example, fluctuations of a signal alter the performance of the sonar and should be reflected in the ROC curve. For this particular example, one must consider modifications of the ROC curves that compensate for increased detection threshold crossings (for a fixed probability of false alarm). Hence, signal fluctuations in this case will result in a net increase in the probability of detection.

6.2 Active Sonar Equation

Active sonars have two fundamental configurations. A monostatic sonar transmits a pulse which interacts with a target, and a resulting echo return to the receiver array which is colocated with the transmitter. A bistatic sonar has the receiver in a different location than the transmitter. We will be mainly concerned with monostatic sonar but we will point out some of the pertinent differences between the two.

With respect to sonar equations, the main differences between the passive and active cases are that source level is modified by a target strength "TS"; reverberation is the dominant interference; and the transmission loss results from two paths, transmitter to target and target to receiver.

In the monostatic case, the transmission loss is 2TL, where TL is the one way transmission loss, and in the bistatic case, the transmission loss is $\mathrm{TL}_1+\mathrm{TL}_2$, where the transmission loss is a sum (in dB) over paths from the transmitter to target and from the target to the receiver. Reverberation is the additional interference term in the active problem that results primarily from scattering of the transmitted field from the boundaries and volume (including animal and vegetable matter) of the ocean. Hence, reverberation level at the receiver is a function of both the acoustic environment and the location of the receiver relative to the transmitter. The total interference term in the sonar equation is a sum of reverberation and noise terms, $\mathrm{RL}+\mathrm{N}$. The concept of detection threshold remains the same as for the passive case so that we can rewrite Eq. (77) for the active case as

$$\mathrm{SE}=\mathrm{SL}-\mathrm{TL}_1+\mathrm{TS}-\mathrm{TL}_2-(\mathrm{RL}+\mathrm{N})+\mathrm{DI}-\mathrm{DT}, \qquad (83)$$

where, in the monostatic case, the two transmission loss terms can be replaced by 2TL. The first two terms are for the outgoing path; the next two terms are for the returning echo; the quantities in the parentheses are the interference, and the last two terms are the system/operator parameters.

In most practical cases, reverberation is the dominant interference factor for active sonar performance. Hence we drop the noise term, noting that it may have to be retained in highly specialized cases. The directivity index used in the passive sonar equation approximated the interference (i.e., noise field) as isotropic; this is no longer a valid approximation for reverberation. It is therefore often convenient to define a quantity as equivalent plane-wave reverberation level "$\mathrm{RL}_{\mathrm{PW}}$," which replaces the terms $\mathrm{N}-\mathrm{DI}$ in passive sonar equations:

$$\mathrm{SE}=\mathrm{SL}-\mathrm{TL}_1+\mathrm{TS}-\mathrm{TL}_2-\mathrm{RL}_{\mathrm{PW}}-\mathrm{DT}. \qquad (84)$$

The corresponding FOM for an active system is defined similarly to the passive system case, Eq. (78), but is for the maximum allowable two-way transmission loss with $\mathrm{TS}=0$ dB. We can estimate the influence of surface and bottom reverberation on sonar performance by combining Eq. (17) or (18) with a propagation model from Sec. 4.

7. UNDERWATER ACOUSTICS AND REMOTE SENSING

Probably the most common usage for underwater acoustics is its application to determining water depth. A depth (echo) sounder or fathometer uses a travel time measurement on an emitted and bottom-reflected short acoustic pulse. The display is often a paper chart with time (which is associated with position) as the running axis versus two way reflection time interval which is translated

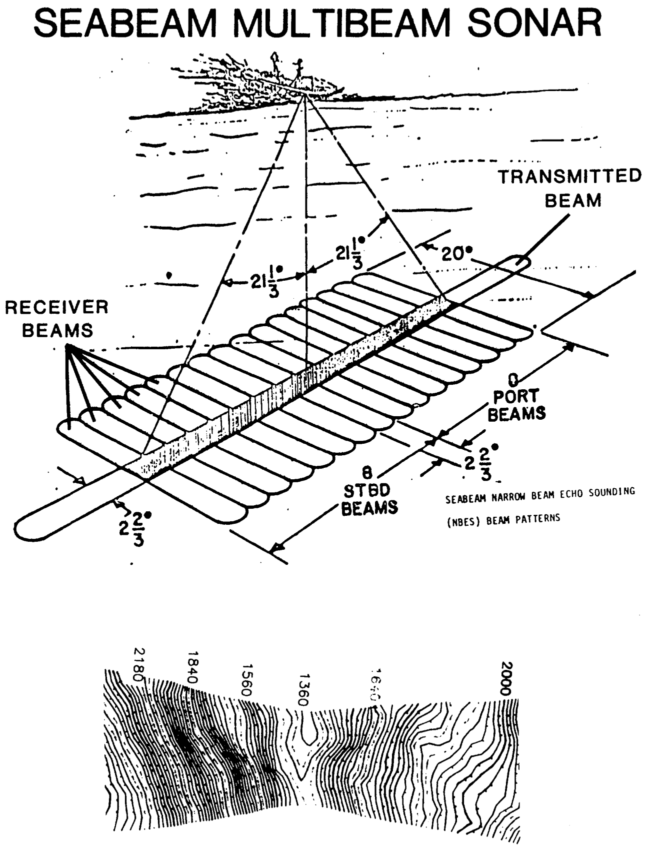

FIG. 20. Schematic of a multibeam bathymetry system with bathymetry contours as output showing a hill (shallow depth) in the center. [Courtesy P. R. Vogt.]

into depth (half-travel distance). Normally, the simplest depth sounders have a wide beamwidth so that they measure an average or minimum depth over a large area. Higher resolution depth sounders have beamwidths on the order of a few degrees and can be used for underwater positioning.

Underwater acoustics plays a major role in bathymetry, "the science and technology of ocean depth measurement," from which geologists/morphologists (who study undersea "geomorphology") infer basic understanding of undersea and earth geologic processes (Vogt and Tucholke, 1986; deMoustier, 1988). The simplest devices used for imaging the ocean bottom are the echo sounders described above. Bathymetric images are compiled from digitized data from the echo sounders combined with navigation/position data. More sophisticated sonar devices (see Fig. 20) have been developed which are essentially multi-narrow-beam ("multibeam") downward looking sonars. Some of these systems have dozens of beams. (The "SASS" system has 61 beams of width 1.5 degrees with a total swath width of up to 1.2 times the water depth.)

Recently introduced sidescan sonars (see Fig. 21), such as the SeaMARC (Sea Mapping and Remote Characterization) use phase measuring interferometric methods and give indication of bottom texture. The SeaMARC I (a 30 kHz system) is towed about 300 m above the ocean bottom at about 3 km/hr and has a swath width of about 2.5 km. Such a system can see objects less than a meter in height with a horizontal resolution of the order of 20 m. Longer range systems such as GLORIA (a 6.5 kHz system) have swath widths 15 to 30 km wide on each side of the ship and can see objects with heights of a few meters and horizontal resolution of the order of hundreds of meters. Higher resolution systems operate in the hundreds of kilohertz region with associated resolutions. In all cases, the key component in bathymetric analysis is the sophisticated data processing necessary to synthesize bathymetric/morphologic images.

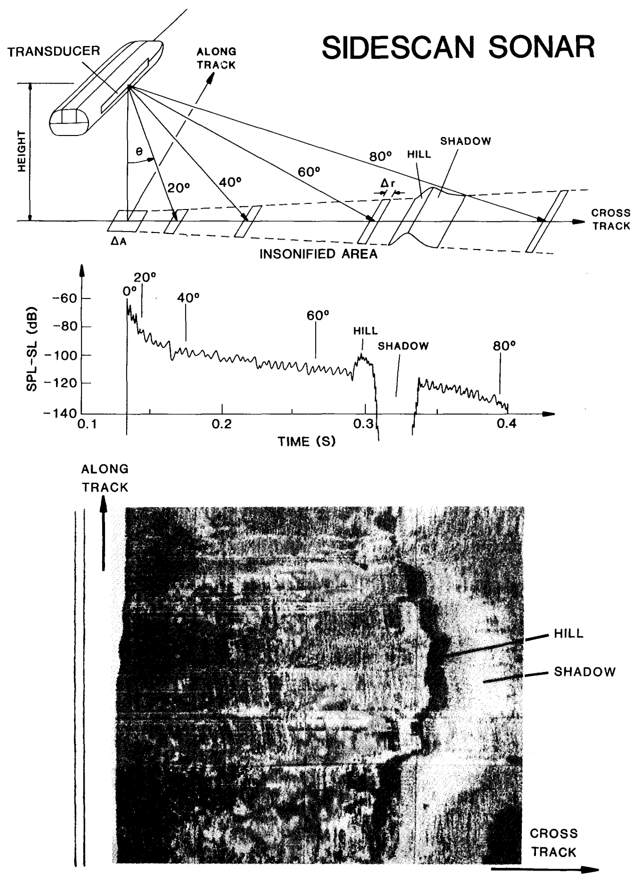
SIDESCAN SONAR
TRANSDUCER
ALONG TRACK
HEIGHT
θ
20°
40°
60°
80°
HILL
SHADOW
Δr
ΔA
INSONIFIED AREA
CROSS TRACK
SPL-SL (dB)
-60
-80
-100
-120
-140
0°
20°
40°
60°
HILL
SHADOW
80°
0.1
0.2
0.3
0.4
TIME (S)
ALONG TRACK
HILL
SHADOW
CROSS TRACK

FIG. 21. Schematic of a sidescan sonar system showing a received pressure level vs time crosstrack plot which is stacked along track to give textured picture of the ocean bottom. [Courtesy P. R. Vogt.]

With respect to the water column structure of the ocean, an extremely ambitious ocean acoustic remote sensing technology has recently emerged: "Ocean Acoustic Tomography" (Cornuelle *et al.*, 1985) (see ACOUSTICAL TOMOGRAPHY). Tomography is a remote sensing scheme which utilizes a series of acoustic time of flight measurements between sources and receivers on the periphery of a large ocean region (dimensions of a few hundred km or greater) of which knowledge of the sound speed structure is desired. This latter knowledge is primarily convertible to temperature (see Sec. 2). Oceanographers employ this technique to measure ocean mesoscale phenomena: temperature changes of the order of tenths of a degree over distance scales of the order of tens to hundreds of kilometers and times scales of days to weeks. The method utilizes the waveguide behavior of sound propagation in the ocean (Sec. 3), where the information contained in the multipath time-of-flight measurements is inverted using Radon's theorem to give the index of refraction of the internal region that is being insonified.

GLOSSARY

Active Sonar: A sonar which emits sound and receives its echo.

Adiabatic Mode Theory: Solution to the range-dependent wave equation under the assumption that the normal modes adapt to the gradually changing environment.

Array Gain (AG): The signal to noise ratio of an array divided by the average signal to noise of a hydrophone element expressed in decibels.

Beamforming: Phasing an array to form a set of "look directions."

Bistatic Reverberation: Reverberation at a point not colocated with the source.

Continuous Spectrum: The part of the solution spectrum of the waveguide problem which corresponds to the energy which is eventually radiated out of the waveguide.

Convergence Zone Propagation: Spatially periodic ($\approx$35–65 km) refocusing of sound from a shallow source producing zones of high intensity near the surface due to the upward refracting nature of the sound speed profile and the absence of bottom interaction.

Coupled Mode Theory: Solution to the range-dependent wave equation including mode conversion due to the changing environment.

Critical Depth: The depth (below the deep sound channel axis) at which the sound speed is the same as that at the surface.

Critical Grazing Angle: The angle with respect to the horizontal below which perfect reflection at an interface occurs.

Decibels: Ten times the logarithm in base 10 of a ratio of intensities.

Deep Scattering Layer: A layer in the water column populated by organisms which scatter sound and which typically undergoes diurnal variations in depth.

Deep Sound Channel: A sound channel occurring in deep water whose axis is at the minimum of the sound speed profile and in which propagation does not involve interaction with the ocean surface or bottom.

Detection Threshold (DT): The additional signal level above zero signal to noise ratio at the output of a sonar array necessary for a detection.

Directivity Index: The array gain under the assumption that the noise is isotropic.

Discrete Spectrum: The part of the solution of the waveguide problem which corresponds to long range normal mode propagation.

Fast Field Program (FFP): Direct numerical full field solution of the range-independent acoustic wave equation including both the continuous and discrete spectra.

Figure of Merit (FOM): The transmission loss that gives zero signal excess for a specified probability of detection.

Geometric Spreading Loss: For a point source, the amplitude decay with range (r) is spherical (r^{-2}) in free space and is cylindrical (r^{-1}) in a waveguide.

Lambert's Law: The scattering from a rough interface is diffuse.

Matched Field Processing: Beamforming by matching the data on an array with the solutions of the wave equation specific to the environment.

Micropascal: (μPa) 10^{-5} dyne per square centimeter.

Mixed Layer: The layer of water below the surface mixed by wave motion; it is essentially isothermal.

Monostatic Reverberation: Reverbation at the source.

Normal Modes: The eigenfunction solutions of the depth part of the range-independent wave equation.

Parabolic Equation (PE): Approximate version of the one-way wave equation for mildly changing environment including some mode coupling effects.

Passive Sonar: A sonar which only receives sound.

Probability of Detection [$p(D)$]: The probability that an existing signal will be detected.

Probability of False Alarm [p(FA)]: The probability that noise crosses the detection threshold and hence an incorrect detection is made.

Propagation Loss: In decibels, 20 log (amplitude of the field from a unit source).

Range-Dependent: An ocean environment which varies in range.

Range-Independent: A stratified ocean environment such that separation of variables, range, depth, and azimuth, can be employed to solve the wave equation.

Ray Theory: The high frequency limit of solving the acoustic wave equation which includes refraction but not diffraction effects.

Recognition Differential (RD): see Detection Threshold.

Reverberation: The scattered acoustic field from an active sonar source which acts as interference in the sonar system.

ROC (Receiving Operating Characteristic) Curves: Parametrized plots of probability of detection versus probability of false alarm.

Signal Excess (SE): The difference between the received signal at the output of the sonar array and the minimum signal for dectection.

Sound Channel: A local minimum in the sound speed profile in a stratified ocean causing ducted propagation according to Snell's law.

Sound Speed Profile: The speed of sound as a function of depth.

Split Step Algorithm: Extremely fast method to solve the parabolic equation based on fast Fourier transforms.

Surface Duct: A sound channel whose upper boundary is the ocean surface formed when there is a local speed profile minimum near the ocean surface.

Thermocline: A vertical negative temperature gradient typically below the mixed layer.

Transmission Loss: The negative of propagation loss and therefore a positive quantity.

List of Works Cited

Adair, R. G., Orcutt, J. A. (1986), *J. Acoust. Soc. Am.* **80**, 633.

Akal, T., Schmidt, H., Kuperman, W. A. (1988), in: B. R. Kerman (Ed.), *Sea Surface Sound: Natural Mechanisms of Surface Generated Noise in the Ocean*, Dordrecht: Kluwer Academic, p. 273.

Anderson, V. C. (1979), *J. Acoust. Soc. Am.* **66**, 1446.

Baggeroer, A.B., Kuperman, W. A., Schmidt, H. (1988), *J. Acoust. Soc. Am.* **83**, 571.

Brekhovskikh, L. M. (1980), *Waves in Layered Media*, 3rd ed. [Translated by R. T. Beyer], New York: Academic.

Brekhovskikh, L. M., Lysanov, Yu. (1982), *Fundamentals of Ocean Acoustics*, Berlin: Springer-Verlag.

Chapman, R. P., Harris, H. H. (1962), *J. Acoust. Soc. Am.* **34**, 1592.

Chapman, R. P., Marshall, J. R. (1966), *J. Acoust. Soc. Am.* **40**, 405.

Clay, C. S., Medwin, H. (1977), *Acoustical Oceanography*, New York: Wiley-Interscience.

Collins, M. D. (1989), *J. Acoust. Soc. Am.* **86**, 1459.

Cornuelle, B. D., Wunch, C., Behringer, D., Birdsall, T., Brown, M., Heinmiller, R., Knox, R., Metzger, K., Munk, L. W., Spiesberger, J., Spindel, R., Webb, D., Worcester, P. (1985), *J. Phys. Oceanogr.* **15**, 133.

Cron, B. F., Sherman, C. H. (1962), *J. Acoust. Soc. Am.* **34**, 1732–1736.

deMoustier, C. (1988), *International Hydrographic Review*, Monaco, **LXV** (No. 2), p. 25.

Evans, R. B. (1983), *J. Acoust. Soc. Am.* **74**, 188.

Fisher, F. H., Simmons, V. P. (1977), *J. Acoust. Soc. Am.* **62**, 558.

Flatte, S. M. (1983), *Proc. IEEE* **71**(11), November.

Hamilton, E. L. Bachman, R. T. (1982), *J. Acoust. Soc. Am.* **72**, 1891.

Jensen, F. B., Kuperman, W. A. (1980), *J. Acoust. Soc. Am.* **67**, 1564.

Jensen, F. B., Kuperman, W. A. (1983), *J. Acoust. Soc. Am.* **73**, 813.

Keller, J. B., Papadakis, J. S. (Eds.) (1977), *Wave Propagation and Underwater Acoustics*, Berlin: Springer-Verlag.

Kerman, B. R. (Ed.) (1988), *Sea Surface Sound: Natural Mechanisms of Surface Generated Noise in the Ocean*, Dordrecht: Kluwer Academic.

Kibblewhite, A. C., Ewans, K. C. (1985), *J. Acoust. Soc. Am.* **78**, 981.

Kuperman, W. A., Ingenito, F. (1977), *J. Acoust. Soc. Am.* **61**, 1178.

Kuperman, W. A., Ingenito, F. (1980), *J. Acoust. Soc. Am.* **67**, 1988.

Kuperman, W. A., Jensen, F. B. (1980), "Deterministic propagation modeling I and II," in: L. Bjorno (Ed.), *Underwater Acoustics and Signal Processing*, Dordrecht: Reidel.

Kuperman, W. A., McDonald, B. E. (1986), "Linear and Nonlinear Ocean Acoustic Propagation Models," in: T. Akal and J. M. Berkson (Eds.), *Ocean Seismo-Acoustics*, New York: Plenum.

Lasky, M. (1977), *J. Acoust. Soc. Am.* **61**, 283.

Lee, D., McDaniel, S. T. (1988), *Ocean Acoustic Propagation by Finite Difference Methods*, New York: Pergamon.

Liggett, W. S., Jacobsen, M. J. (1965), *J. Acoust. Soc. Am.* **38**, 302.

McDonald, B. E., Kuperman, W. A. (1987), *J. Acoust. Soc. Am.* **81**, 1406.

Morris, G. B. (1978), *J. Acoust. Soc. Am.* **64**, 581.

Northrup, J., Colborn, J. G. (1974), *J. Geophys. Res.* 5633.

Ogilvy, J. A. (1987), *Rep. Prog. Phys.* **50**, 1553.

Pierce, A. D. (1981), *Acoustics: An Introduction to its Physical Principles and Applications*, New York: McGraw–Hill.

Ross, D. (1976), *Mechanics of Underwater Noise*, New York: Pergamon.

Schmidt, H., Jensen, F. B. (1985), *Comp. and Maths. with Appls.* **11**, 699.

Schmidt, H., Kuperman, W. A., (1988), *J. Acoust. Soc. Am.* **64**, 2153.

Tappert, F. D. (1977), "The Parabolic Approximation Method," in: J. B. Keller, J. S. Papadakis (Eds.), *Wave Propagation and Underwater Acoustics*, Berlin: Springer-Verlag.

Thorpe, W. H. (1967), *J. Acoust. Soc. Am.* **42**, 270.

Urick, R. J. (1979), *Sound Propagation in the Sea*, Washington, DC: U.S. GPO.

Urick, R. J. (Ed.) (1983), *Principles of Underwater Sound*, New York: McGraw–Hill.

Urick, R. J. (1984), *Ambient Sea Noise in the Ocean*, Washington, DC: U.S. GPO.

Vogt, P. R., Tucholke, B. E. (1986), "Imaging the ocean floor: History and state of the art," in: P. R. Vogt, B. E. Tucholke (Eds.), *The Geology of North America, Volume M, The Western North Atlantic Region*, The Geological Society of America.

Wagstaff, R. A. (1981), *J. Acoust. Soc. Am.* **69**, 1009.

Wenz, G. M., (1962), *J. Acoust. Soc. Am.* **34**, 1936.

Wilson, O. B. (1985), *An Introduction to the Theory and Design of Sonar Transducers*, Washington DC: U.S. GPO.

Further Reading

Brekhovskikh, L. M. (1980), *Waves in Layered Media*, 3rd ed. (Translated by R. T. Beyer), New York: Academic.

Brekhovskikh, L. M., Lysanov, Yu. (1982), *Fundamentals of Ocean Acoustics*, Berlin: Springer-Verlag.

Clay, C. S., Medwin, H. (1977), *Acoustical Oceanography*, New York: Wiley-Interscience.

Keller, J. B., Papadakis, J. S. (Eds.) (1977), *Wave Propagation and Underwater Acoustics*, Berlin: Springer-Verlag.

B. Kerman, R. (Ed.) (1988), *Sea Surface Sound: Natural Mechanisms of Surface Generated Noise in the Ocean*, Dordrecht: Kluwer.

Lasky, M. (1977), *J. Acoust. Soc. Am.* **61**, 283.

Pierce, A. D. (1981), *Acoustics: An Introduction to its Physical Principles and Applications*, New York: McGraw-Hill.

Ross, D. (1976), *Mechanics of Underwater Noise*, New York: Pergamon.

Urick, R. J. (1979), *Sound Propagation in the Sea*, Washington, DC: U.S. GPO.

Urick, R. J. (1983), *Principles of Underwater Sound*, 3rd Ed., New York: McGraw-Hill.

Urick, R. J. (1984), *Ambient Sea Noise in the Ocean*, Washington, DC: U.S. GPO.

Wilson, O. B. (1985), *An Introduction to the Theory and Design of Sonar Transducers*, Washington, DC: U.S. GPO.

ACOUSTO-OPTICAL MODULATORS

See MODULATORS

ADHESION

See SURFACE PROCESSES

ADSORPTION

See SURFACE PROCESSES

AERODYNAMICS

John J. Bertin* and George B. Hept, *The U.S. Air Force Academy*

Introduction 329
1. **A Review of Fluid Mechanics** 330
1.1 Equations Governing Fluid Motion 331
1.1.1 Continuity Equation 331
1.1.2 Conservation of Linear Momentum 331
1.1.3 Conservation of Energy 332
1.2 General Remarks on the Equations of Motion 332
2. **Characterizing the Flow Field** 333
2.1 Inviscid Flows 335
2.1.1 Euler's Equation 335
2.1.2 Bernoulli's Equation 335
2.1.3 Bernoulli's Equation for Gaseous Flows and the Pressure Coefficient 336
2.1.4 Pressure Gradients 338
2.1.5 Irrotational Flows 338
2.1.6 Isentropic, Compressible Flows 339
2.2 Viscous Flow 339
2.2.1 Boundary-Layer Characteristics 341
2.2.2 Boundary-Layer Transition 342
2.2.3 The Effect of the Boundary Layer on Flow Separation 344
2.3 Shock Waves 345
3. **Characterization of Aerodynamic Forces** 346
3.1 Aerodynamic Forces and Moments 347
3.2 Parameters that Govern Aerodynamic Forces 348
3.3 Lift Coefficient 348
3.4 Moment Coefficients 349
3.5 Drag Coefficient 350
3.5.1 Zero-Lift Drag 352
3.5.2 Drag Due to Lift 352
3.5.3 Compressibility 352
3.6 Lift/Drag Ratio 352
4. **Aerodynamic Models by Speed Range** 352
4.1 Subsonic Aerodynamics 352
4.2 Transonic Aerodynamics 355
4.3 Supersonic Aerodynamics 359
4.3.1 Wave Drag at Supersonic Speeds 360
4.3.2 Design Considerations for Tactical Military Aircraft 361
4.4 Hypersonic Flow 362
4.4.1 General Characterization of Hypersonic Flows 362
4.4.2 Defining Hypersonic Flow 364
4.4.3 Chemistry of Hypersonic Flows 365
5. **Rotorcraft Aerodynamics** 366
5.1 Hovering Flight 367
5.2 Forward Flight 368
5.3 Descending Flight 369
5.4 Summary 370
6. **Computational Fluid Dynamics** 371
6.1 General Remarks 371
7. **Nonaerospace Applications of Aerodynamics** 372
7.1 General Remarks 372
7.2 Automotive Aerodynamics 372
7.3 Wind Loads on Buildings 373
7.4 Aerodynamics of Sports Balls 374
Acknowledgments 376
Glossary 376
Works Cited 376
Further Reading 377

INTRODUCTION

The aerodynamicist applies the principles of fluid mechanics to determine the forces and moments acting on an object moving through the air, e.g., a golf ball or an airplane, or due to the air moving past an object, e.g., the wind loads acting on a building. The proper determination of the aerodynamic forces and moments makes use of experimental measure-

---

*Currently with Sandia National Laboratories.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

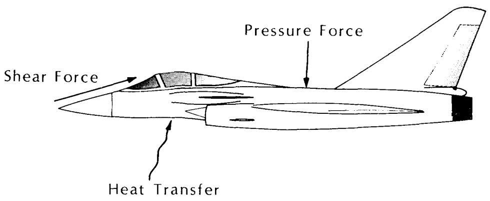

FIG. 1. Aerothermodynamic environment for a high-speed airplane.

ments and of computer-generated flow-field solutions. In most ground-based experimental investigations, e.g., wind-tunnel tests, the air moves around a model held stationary in the wind tunnel. In flight, however, the vehicle moves through the air. Thus, in one case, the analysis of the flow takes place in a vehicle-fixed coordinate system (the air moves around the stationary object), whereas in the other case, the analysis takes place in a ground-fixed coordinate system (the object moves through a stationary air mass). Although there are some significant differences between these two approaches regarding the treatment of the time-dependent terms in the analysis, the aerodynamic forces and moments acting on the object should be the same for either approach. Thus, in the discussion that follows, we shall usually refer to the flow around an object.

Aerodynamic forces acting on an object consist of the skin friction (or shear) forces and the pressure forces acting on its surface, as shown in Fig. 1. To determine the resultant pressure forces, one must know the local static pressure as a function of space and time, i.e., $p(x,y,z,t)$ where x, y, and z are also functions of time. For flight below 75 km (246 kft), the air may be considered a continuum. Because of viscosity, air particles that are immediately adjacent to the surface in this continuum move at the same speed as the surface, i.e., a "no-slip condition" exists at the surface of the object. As a result, there are large velocity gradients in the viscous boundary layer that exists between the no-slip boundary condition at the surface and the local "inviscid" flow away from the surface. These velocity gradients produce shear stresses which are transmitted to the surface as shown in the sketch of Fig. 1. Thus, to determine the shear forces acting on the vehicle, one must be able to define the velocity field, i.e., $\mathbf{V}(x,y,z,t)$, and the viscosity. For high-speed flight, on the order of three times the speed of sound or greater, point-to-point temperature variations in the flow field can cause significant variations in the fluid properties and can create sufficiently large temperature variations in the boundary layer, so that heat transfer to the vehicle becomes of concern to the analyst.

In the discussion that follows, the basic principles of fluid mechanics are discussed and then applied to aerodynamics. For an expanded discussion and for applications to a variety of projects, the reader is referred to Bertin and Smith (1989).

1. A REVIEW OF FLUID MECHANICS

We can define the aerothermodynamic environment of a vehicle, i.e., the aerodynamic forces and moments acting on the vehicle, and the heat transferred to the vehicle, if we can determine

$$p(x,y,z,t), \tag{1a}$$

$$\mathbf{V}(x,y,z,t), \tag{1b}$$

$$T(x,y,z,t). \tag{1c}$$

To obtain solutions for these variables (either experimentally or theoretically), one must develop flow models. Since we have three unknowns, we need three independent equations. The fundamental physical laws used to solve for the fluid motion in a general problem are as follows:

1. conservation of mass (or continuity equation),
2. conservation of linear momentum (or Newton's second law of motion),
3. conservation of energy (or the first law of thermodynamics).

Note that if we consider the velocity vector **V** as one of the unknown dependent variables, then we would use the vector form of the momentum equation (since the forces acting on a fluid particle and the resultant accelerations can be written as vector quantities). Therefore, the first and third physical relations are scalar equations, while the conservation of linear momentum is represented by a vector relation. However, the velocity field can be represented by its three components (u, v, and w). In this case, there would be three component momentum equations. Indeed, this will be the approach taken in the present development, since Eqs. (5a)– (5c) represent the three components of the conservation-of-momentum equation in Cartesian coordinates. In this approach, we have five scalar equations with five primary unknowns: p, u, v, w, and T. These laws can be expressed in differential form or in integral form.

1.1 Equations Governing Fluid Motion

1.1.1 Continuity Equation. If the principle of conservation of mass is applied to a small volume of space through which the fluid can move freely, the net outflow of mass through the surface surrounding the volume must be equal to the decrease of mass within the volume. In vector form, the general differential form of the continuity equation is

$$\frac{\partial \rho}{\partial t}+\nabla\cdot(\rho\mathbf{V})=0. \tag{2}$$

1.1.2 Conservation of Linear Momentum. The equation for the conservation of linear momentum is obtained by applying Newton's second law: The net force acting on a fluid particle is equal to the time rate of change of the linear momentum of the fluid particle. As the fluid element moves in space, its shape and volume may change, but its mass is conserved. Thus, using a coordinate system that is neither accelerating nor rotating, called an *inertial coordinate system,* we may write

$$\sum \mathbf{F}=m\frac{d\mathbf{V}}{dt}\,. \tag{3}$$

The velocity $\mathbf{V}$ of a fluid particle is, in general, an explicit function of time t as well as of its position x, y, z. Furthermore, the position coordinates x, y, z of the fluid particle are themselves a function of time. Since the time differentiation of Eq. (3) follows a given particle in its motion, the derivative is frequently termed the *particle, total,* or *substantial derivative* of $\mathbf{V}$:

$$\frac{d\mathbf{V}}{dt}=\frac{\partial\mathbf{V}}{\partial t}+(\mathbf{V}\cdot\nabla)\mathbf{V}, \tag{4}$$

where $d\mathbf{V}/dt$ is the total acceleration, $\partial\mathbf{V}/\partial t$ is the unsteady (time-dependent) acceleration, and $(\mathbf{V}\cdot\nabla)\mathbf{V}$ is the convective acceleration.

The principal forces with which we are concerned are those that act directly on the mass of the fluid element (the body forces) and those that act on its surface (the pressure forces and shear forces). The surface forces (per unit area) acting on the fluid particle are described by a stress tensor. In general, the various stresses change from point to point, producing net forces on the fluid particle which cause it to accelerate or to decelerate.

We can relate the stresses to the fluid motion. Stress components for a Newtonian fluid are a linear function of the components of the rate of strain. Thus, for a one-dimensional flow, $\tau=\mu(\partial u/\partial y)$. The viscosity μ is called the first viscosity coefficient.

With the appropriate expressions for the stresses substituted into the expressions for the forces in Eq. (3), we obtain

$$\rho\frac{\partial u}{\partial t}+\rho(\mathbf{V}\cdot\nabla)u=\rho f_x-\frac{\partial p}{\partial x}+\nabla\cdot(\mu\nabla u)+\nabla\cdot\left(\mu\frac{\partial\mathbf{V}}{\partial x}\right)-\frac{2}{3}\frac{\partial}{\partial x}(\mu\nabla\cdot\mathbf{V}) \tag{5a}$$

$$\rho\frac{\partial v}{\partial t}+\rho(\mathbf{V}\cdot\nabla)v=\rho f_y-\frac{\partial p}{\partial y}+\nabla\cdot(\mu\nabla v)+\nabla\cdot\left(\mu\frac{\partial V}{\partial y}\right)-\frac{2}{3}\frac{\partial}{\partial y}(\mu\nabla\cdot\mathbf{V})\,, \tag{5b}$$

$$\rho\frac{\partial w}{\partial t}+\rho(\mathbf{V}\cdot\nabla)w=\rho f_z-\frac{\partial p}{\partial z}+\nabla\cdot(\mu\nabla w)+\nabla\cdot\left(\mu\frac{\partial\mathbf{V}}{\partial z}\right)-\frac{2}{3}\frac{\partial}{\partial z}(\mu\nabla\cdot\mathbf{V}). \tag{5c}$$

These general, differential equations for the conservation of linear momentum are known as the *Navier–Stokes equations.* Navier in 1827 and Stokes in 1845 contributed to the inclusion of the frictional resistance terms in the equations of motion. Note that the viscosity μ is considered to be dependent on the spatial coordinates. This is done since, for a compressible flow, the changes in velocity and pressure, together with the kinetic-energy dissipation due to friction, bring about considerable temperature variations. The temperature dependence of viscosity in the general case should, therefore, be incorporated into the governing equations.

1.1.3 Conservation of Energy. The equation for the conservation of energy is obtained by applying the first law of thermodynamics to a small volume in space through which the fluid can move freely. The rate at which heat is added to the system of fluid particles, less the rate at which work is done by the system, is equal to the time rate of change of energy within the control volume plus the net efflux of energy across the system boundary. The energy of the system can take a variety of forms. They are usually grouped as follows:

1. *Kinetic energy* (e_k): energy associated with the directed motion of the mass.
2. *Potential energy* (e_p): energy associated with the position of the mass in an external field.
3. *Internal energy* (e): molecular and atomic energy associated with the internal fields of the mass.

The total energy of the system (e_t) may be written as

$$e_t = e_k + e_p + e. \tag{6}$$

Thus, the energy equation becomes

$$\frac{\delta q}{\delta t} - \frac{\delta w}{\delta t} = \frac{d(e_k)}{dt} + \frac{d(e_p)}{dt} + \frac{d(e)}{dt}. \tag{7}$$

The rate at which heat is transferred to the system and work is done by the system are path dependent, as indicated by $\delta/\delta t$. Introducing Fourier's law of heat conduction, the relations between stress and rate of strain, and the forms of energy, one obtains the general differential form of the energy equation as

$$\rho \frac{dh}{dt} - \frac{dp}{dt} = \nabla \cdot (k \nabla T) + \phi, \tag{8a}$$

where

$$\begin{aligned}\phi = & -\frac{2}{3}\mu(\nabla \cdot \mathbf{V})^2 \\ & + 2\mu\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2 + \left(\frac{\partial w}{\partial z}\right)^2\right] \\ & + \mu\left[\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y}\right)^2 \right. \\ & \left. + \left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z}\right)^2\right].\end{aligned} \tag{8b}$$

Equation (8b) defines the dissipation function ϕ. The specific enthalpy h is defined as

$$h = e + p/\rho, \tag{8c}$$

where e is the specific internal energy and p/ρ is related to the flow work.

1.2 General Remarks on the Equations of Motion

The continuity equation [Eq. (2)], the three components of the momentum equation [Eqs. (5a)–(5c)], and the energy equation [Eq. (8a)] represent a system of scalar equations in terms of five unknowns: p, T, u, v, and w. These five scalar equations of motion contain additional variables: density (ρ), enthalpy (h), viscosity (μ), and thermal conductivity (k). These four variables are, in general, functions of the pressure, the temperature, and the composition of the gas. Thus, our system of five equations of motion with five unknowns can be supplemented with additional relations from the kinetic theory of gases.

The evaluation of these four parameters presents varying degrees of difficulty, which depend on the application. For an application such as the Aeroassisted Space Transfer Vehicle (ASTV), which operates at velocities of 10 km s^{-1} (32.8 kft s^{-1}) at altitudes exceeding 90 km (295 kft), one must be able to model the noncontinuum and the nonequilibrium character of the flow. However, for low-velocity flows, i.e., less than 100 m s^{-1} (194 knots), temperature variations in the flow field are relatively small. Thus, for low-speed flows, the transport properties (μ and k) and the density (ρ) are usually assumed constant. Constant-density flows are termed incompressible flows.

For velocities in excess of 100 m s^{-1}, the density is often calculated using the equation of state for a thermally perfect gas:

$$\rho = p/RT. \tag{9}$$

In Eq. (9), the gas constant (R) is the universal gas constant ($\hat{R}$) divided by the molar mass ($\mathscr{M}$).

Clearly, in order to determine the aerodynamic forces and moments acting on an object, one must be able to define a complex flow field. For the purpose of the present discussion, we will say that there are four ways to generate design information about the flow field:

1. Experimental programs employing scale models in a wind tunnel.

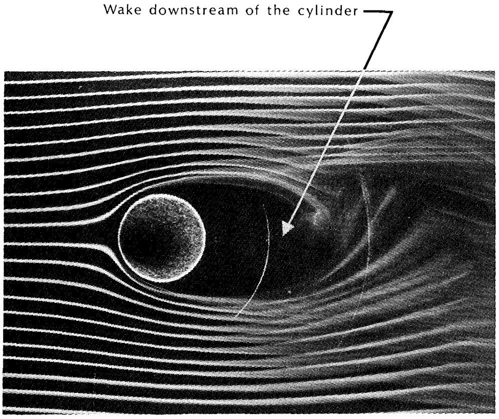

FIG. 2. Flow-visualization photograph of low-speed flow around a cylinder (flow from left to right).

2. Analytical-numerical formulations in which the flow field is divided into two regions, one in which the flow is essentially inviscid and a second in which the effects of viscosity are important.
3. Analytical-numerical formulations in which the entire flow field is determined simultaneously.
4. Flight-test programs employing a full-scale (or prototype) model.

The full-scale flight test is the only one of the four techniques that does not involve compromises. Because wind-tunnel tests usually employ small-scale models, there are often compromises even in such basic simulation parameters as the Reynolds number. Furthermore, the support system which holds the model in position in the wind tunnel may influence the data of interest. Although numerical codes incorporating a two-layer (inviscid-viscous) model of the flow field provide powerful design tools, the analyst must be sure that the flow model used in the code is appropriate. For instance, the two-layer model cannot be used to compute the flow field downstream of a blunt object, such as the flow in the wake of a cylinder, as shown in Fig. 2.

Because of the increased capability of computer hardware and of the improved software developed by numerous researchers, the third method is a rapidly developing tool. As discussed by Li (1989), the development and application of a computational fluid dynamics (CFD) code involves several steps:

1. Selecting the physical processes to be modeled.
2. Selecting the mathematical and topological models.
3. Defining the body geometry and constructing the spatial grid.
4. Developing a numerical solution algorithm.
5. Incorporating steps (1)–(4) in a computer code.
6. Calibrating and validating the code against benchmark experimental data or "classic" analytical solutions.
7. Using the CFD code to predict the desired aerodynamic coefficients, flow-field properties and/or heating-rate environment.

The development, calibration, and verification of a really useful CFD code may require several years.

In reality, the design process will integrate programs involving all four techniques. The data from experimental programs will be used to improve the flow models used in the computer codes. The computer codes can be used to generate flow-field solutions at conditions outside the test envelopes.

2. CHARACTERIZING THE FLOW FIELD

To predict accurately the aerodynamic forces and moments that act on a vehicle in flight, it is necessary to be able to describe the pattern of flow around the configuration. The resultant flow pattern depends on the geometry of the vehicle, its orientation with respect to the undisturbed free stream, and the altitude and speed at which the vehicle is traveling. Simplifying assumptions can be made for the equations of motion depending upon the flight regime of the aircraft. A discussion of some of these simplifying assumptions and their regions of applicability is presented in the following sections.

Consider the forces acting on the airfoil shown in the sketch of Fig. 3. We will assume that an airfoil is of infinite span, i.e., its length into the paper is infinite, and every cross section is identical to that appearing in the plane of the paper. Thus, the flow over an airfoil is two dimensional, i.e., the same in every plane parallel to the plane of the paper. Since a wing is a series of airfoil sections of finite span, the flow over a wing will be three di-

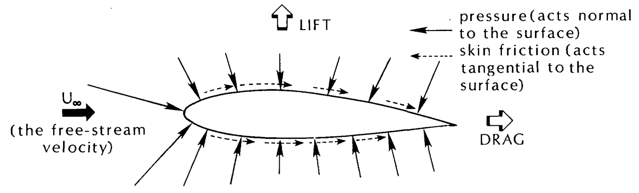
LIFT
pressure(acts normal to the surface)
skin friction (acts tangential to the surface)
U∞
(the free-stream velocity)
DRAG

FIG. 3. A sketch of the forces acting on a two-dimensional airfoil.

mensional, introducing additional complexity to its analytical description.

For simplicity, the present discussion will describe the phenomena and simplifying assumptions using two-dimensional flow fields. As discussed earlier and as depicted in Fig. 3, the forces distributed over the vehicle consist of pressure forces acting normal to the surface and skin-friction forces that are tangential to the surface. The varying lengths of the arrows representing the local pressure and the local skin friction reflect the fact that the air particles accelerate and decelerate as they move around the airfoil. Thus, even outside of the viscous boundary layer, the local velocity of the air particles at the edge of the boundary layer (u_e in Fig. 4) will be a function of the streamwise location. The aerodynamicist resolves the integrated effect of these forces into a lift (that acts normal to the free-stream velocity, upward being considered positive) and a drag force (which acts in the direction of the free-stream flow). Based on the sketch presented in Fig. 3, the pressure forces constitute the dominant contributor to the lift acting on the airfoil. Force coefficients for a two-dimensional flow will be designated with a lower-case subscript. Thus, C_l is the lift coefficient for an airfoil. Force coefficients for a three-dimensional flow will be designated with an upper-case subscript. Thus, C_L is the lift coefficient for a wing.

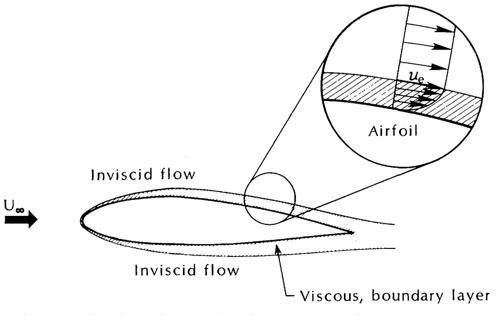

FIG. 4. A sketch of the flow around an airfoil for a two-region flow model.

As illustrated in the sketch of Fig. 4, there is a thin boundary layer adjacent to the surface of the airfoil, where the streamwise velocity component (u) varies from zero at the surface (the no-slip condition) to the (local inviscid) value outside the boundary layer (u_e). The large velocity gradients ($\partial u/\partial y$) in this region produce shear stresses which cannot be neglected. Let us examine the x-momentum equation [Eq. (5a)], and the y-momentum equation [Eq. (5b)], for our two-dimensional flow. Let us assume further that the density variations are negligible (i.e., the flow is incompressible). This implies, as noted earlier, that the velocity is less than 100 m s^{-1}. Thus, the continuity equation [Eq. (2)] becomes

$$\nabla \cdot \mathbf{V} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0. \tag{10}$$

Furthermore, since the density is assumed constant, the viscosity should be constant too. As a result, the x- and y-momentum equations for a two-dimensional, incompressible flow may be written

$$\rho \frac{\partial u}{\partial t} + \rho u \frac{\partial u}{\partial x} + \rho v \frac{\partial u}{\partial y} = -\frac{\partial p}{\partial x} + \mu \frac{\partial^2 u}{\partial x^2} + \mu \frac{\partial^2 u}{\partial y^2}$$

and

$$\rho \frac{\partial v}{\partial t} + \rho u \frac{\partial v}{\partial x} + \rho v \frac{\partial v}{\partial y} = -\frac{\partial p}{\partial y} + \mu \frac{\partial^2 v}{\partial x^2} + \mu \frac{\partial^2 v}{\partial y^2},$$

where the x coordinate is measured parallel to the airfoil surface and the y coordinate is measured perpendicular to it. Solving for the pressure gradients gives us

$$-\frac{\partial p}{\partial x} = \left(\rho \frac{\partial}{\partial t} + \rho u \frac{\partial}{\partial x} + \rho v \frac{\partial}{\partial y} - \mu \frac{\partial^2}{\partial x^2} - \mu \frac{\partial^2}{\partial y^2} \right) u,$$

$$-\frac{\partial p}{\partial y} = \left(\rho \frac{\partial}{\partial t} + \rho u \frac{\partial}{\partial x} + \rho v \frac{\partial}{\partial y} - \mu \frac{\partial^2}{\partial x^2} - \mu \frac{\partial^2}{\partial y^2} \right) v.$$

Near a solid surface, the normal component of velocity is usually much less than the streamwise component of velocity (i.e., v

$<u$). Since there is a similar correlation between the gradients of these velocity components,

$$\frac{\partial p}{\partial y} < \frac{\partial p}{\partial x}\,.$$

The pressure variation across the boundary layer is usually negligible. Therefore, the pressure distribution around the airfoil is essentially that of the inviscid flow (accounting for the displacement effect of the boundary layer).

The assumption that the pressure variation across the boundary layer is negligible breaks down for turbulent boundary layers at very high Mach numbers. Bushnell, Cary, and Harris (1977) cite data for which the wall pressure is significantly greater than the edge value for turbulent boundary layers where the edge Mach number is approximately 20. The characteristics distinguishing laminar and turbulent boundary layers will be discussed in a later section. Thus, we can obtain valuable information about the forces and moments, if we can describe the flow outside of the boundary layer where the velocity gradients associated with shear stresses are negligible, i.e., the *inviscid flow.*

2.1 Inviscid Flows

In the preceding section, we have seen that, for thin boundary layers having streamlines that are relatively straight, the pressure gradient across the boundary layer is negligible. Furthermore, we have seen that the lift generated by a streamlined shape depends primarily on the pressure distribution. In fact, the pitching moments about some reference point (which relate to the stability of the vehicle) also depend primarily on the pressure distribution for a streamlined shape. Thus, if we can determine the pressure distribution for the flow at the edge of the boundary layer, we can generate reasonable estimates for the lift and the pitching moment acting on a vehicle. Furthermore, even though the experimental pressure distributions both from flight tests and from wind-tunnel tests are determined from the pressures sensed by orifices located in the wall, they provide information about the flow outside of the boundary layer. This information can be used to verify computer-generated flow fields or to improve flow models used in these codes.

2.1.1 Euler's Equation.

As indicated in the sketch of Fig. 4, the velocity gradients that give rise to the viscous forces are negligible outside of the boundary layer. Thus, although the viscosity is the same within the boundary layer as outside of it (for a constant property flow), the shear forces (the products of the coefficient of viscosity and the velocity gradients) are small compared to the inertia and the pressure forces acting on the air particles. Thus, we can neglect all the terms involving the coefficient of viscosity (μ) in Eqs. (5a)–(5c). As a result, the components of the conservation of linear momentum equation for an inviscid flow become

$$\rho\,\frac{du}{dt} = \rho f_x - \frac{\partial p}{\partial x}\,, \tag{11a}$$

$$\rho\,\frac{dv}{dt} = \rho f_y - \frac{\partial p}{\partial y}\,, \tag{11b}$$

$$\rho\,\frac{dw}{dt} = \rho f_z - \frac{\partial p}{\partial z}\,. \tag{11c}$$

We have used the term *inviscid flow* to emphasize the fact that it is the character of both the velocity field and the fluid viscosity that allows us to neglect the viscous forces. In vector form, the equation is

$$\frac{d\mathbf{V}}{dt} = \frac{\partial \mathbf{V}}{\partial t} + (\mathbf{V}\cdot\nabla)\mathbf{V} = \mathbf{f} - \frac{1}{\rho}\,\nabla p. \tag{12}$$

No assumption has been made about the density at this point, so these equations apply to both a compressible and an incompressible flow. These equations, derived in 1755 by Leonhard Euler, are called the *Euler equations.*

2.1.2 Bernoulli's Equation.

If we further assume that the flow is steady, then

$$\frac{\partial \mathbf{V}}{\partial t} = 0.$$

The body force most often encountered is that due to gravity. Since gravity can be represented by a force potential,

$$\mathbf{f} = -\nabla F = -\frac{\partial F}{\partial z}\hat{k} = -g\hat{k}, \tag{13}$$

where we have used the convention that the z axis is positive when pointing upward and normal to the surface of the earth, and that

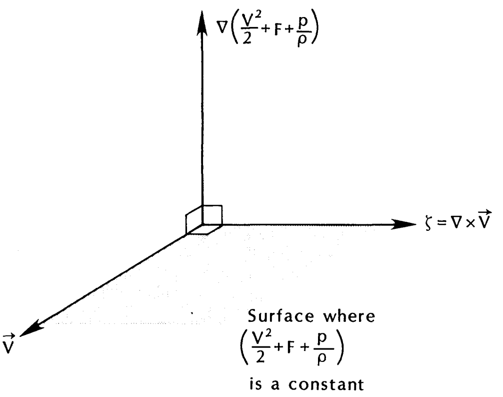
$\nabla\left(\frac{V^2}{2}+F+\frac{p}{\rho}\right)$
$\zeta=\nabla\times\vec{V}$
$\vec{V}$
Surface where
$\left(\frac{V^2}{2}+F+\frac{p}{\rho}\right)$
is a constant

FIG. 5. Geometry for Lamb's equation.

the force per unit mass due to gravity is directed downward and is of magnitude g.

Substituting these relations into Eq. (12) with the additional restriction that the flow is incompressible, we obtain the relation

$$\nabla\left(\frac{V^2}{2}+F+\frac{p}{\rho}\right)=\mathbf{V}\times\zeta, \tag{14}$$

where ζ, the curl of the velocity ($\nabla\times\mathbf{V}$), is the vorticity. This equation, which is called *Crocco's equation* or *Lamb's equation*, provides important information about the geometric relationships between the velocity vector $\mathbf{V}$, the vorticity vector ζ, and the quantity $(V^2/2)+F+(p/\rho)$. From the properties of cross products, the vector $\mathbf{V}\times\zeta$ must be perpendicular to the plane containing $\mathbf{V}$ and ζ. Thus, as shown in Fig. 5, the gradient of $(V^2/2)+F+(p/\rho)$ is perpendicular to the plane of $\mathbf{V}$ and ζ or, equivalently, $(V^2/2)+F+(p/\rho)$ is constant in the plane of $\mathbf{V}$ and ζ.

Furthermore, since the streamlines are parallel to the velocity $\mathbf{V}$, the plane of $\mathbf{V}$ and ζ contains the paths of the fluid flow. Since the streamlines are tangent to $\mathbf{V}$ and the vortex lines are tangent to ζ, the grid of streamlines and vortex lines must lie on the surface where $(V^2/2)+F+(p/\rho)$ equals a constant. Thus, for a flow that is inviscid, incompressible, steady, and irrotational (or, if the flow is rotational, we consider displacements along a vortex line or along a streamline), and for which the body force is gravity, which is conservative,

$$V^2/2+gz+p/\rho=\text{const.} \tag{15}$$

Equation (15) is known as *Bernoulli's equation.*

Because the density has been assumed constant, it is not necessary to include the energy equation in the procedure to solve for the velocity and pressure fields. Note that, in deriving Eq. (15), we have assumed that dissipative mechanisms do not significantly affect the flow. As a corollary, Bernoulli's equation is valid only for flows where there is no mechanism for dissipation, such as viscosity. In thermodynamics, the flow process would be called *reversible.*

Bernoulli's equation can also be derived using the integral form of the energy equation. For this derivation, the flow is assumed to be steady, inviscid, incompressible, and one dimensional with no heat transfer, no shaft work, and no perceptible changes in the internal energy. These assumptions are consistent with the five stated previously, when deriving Bernoulli's equation from the momentum equation. Again, we obtain

$$V^2/2+gz+p/\rho=\text{const},$$

where $V^2/2$ is the kinetic energy, gz is the potential energy, and p/ρ is the flow work, and where we have defined the individual terms by their physical significance based on the energy equation.

2.1.3 Bernoulli's Equation for Gaseous Flows and the Pressure Coefficient. The terms in Eq. (15) have units of energy per unit mass. The aerodynamicist usually multiplies through by ρ, so that the terms have units of force per unit area, i.e., pressure. For most gas flows, the changes in potential energy between two points in the flow field are negligible. (One exception would be in describing the atmospheric flow field where we must account for large changes in altitude.) Neglecting the change in potential energy, Eq. (15) may be written as

$$p+\tfrac{1}{2}\rho U^2=\text{const.} \tag{16}$$

This equation establishes a direct relation between pressure and velocity. Thus, if either parameter is known, the other can be uniquely determined, provided that the flow does not violate the assumptions required to apply Bernoulli's equation. This equation can be used to relate the flow at various points in the inviscid flow around a vehicle. Consider an incompressible flow around an airfoil, as

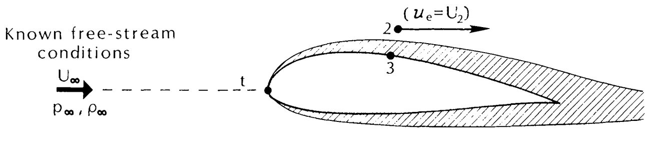

FIG. 6. Application of Bernoulli's equation to airflow around an airfoil.

shown in the sketch of Fig. 6. Recall that, if the airfoil is flying through a mass of air at rest, we superimpose the negative of the velocity of the airfoil so that the airfoil is at rest (in our vehicle-fixed reference frame), and the air is flowing around it. Typical free-stream properties (such as density and static pressure) are known as a function of altitude, e.g., *U. S. Standard Atmosphere* (1976). As the flow approaches the airfoil, it slows down, and there is one point near the leading edge where the local velocity is zero. Because the local velocity is zero, the pressure is equal to its maximum value in the flow field, which we will designate as the total pressure, and represent by the subscript *t*. Referring to Fig. 6,

$$p_\infty+\tfrac{1}{2}\rho_\infty U_\infty^2=p_t. \tag{17}$$

Aeronautical engineers refer to p_∞ as the free-stream static pressure, $\frac{1}{2}\rho_\infty U_\infty^2$ as the dynamic pressure (which is also represented by the symbol q_∞), and p_t as the total, or stagnation, pressure. Thus, for this steady, inviscid, incompressible flow, the total pressure is the sum of the static pressure plus the dynamic pressure. Thus, p_t is the constant of Bernoulli's equation as it appears in Eq. (17). The total pressure is not the sum of $p_\infty + \frac{1}{2}\rho_\infty U_\infty^2$ if density variations are significant, i.e., if the flow is compressible.

The fluid is said to be a *static fluid medium* if the fluid particles, when viewed as a continuum, are either all at rest or all moving with the same velocity. Thus, the term *static fluid properties* may be applied to situations in which elements of the fluid are moving, provided there is no relative motion between fluid particles. Since we have superimposed the velocity of the airfoil to create a steady flow in which all the free-stream air particles are moving at the same speed (U_∞), the free-stream static pressure (p_∞) is the ambient atmospheric pressure for the altitude at which the airfoil is operating. Furthermore, as given by Eq. (17), the pressure at the stagnation point, where the air particles have zero velocity with respect to the wing, is p_t, the total pressure. If we were to analyze this problem in the ground-fixed coordinate system, i.e., with the wing moving through a mass of air which is at rest, the ambient atmospheric pressure is not the total pressure, even though the air particles in the free stream are at rest in this coordinate system. In this reference frame, the time-dependent terms must be included in the flow model.

Consider an air particle that moves from point *t* to point 2 (in Fig. 6), remaining outside the boundary layer. Since the particle remains outside of the boundary layer, it is not affected by viscosity. With the five conditions for Bernoulli's equation satisfied,

$$p_\infty+\tfrac{1}{2}\rho_\infty U_\infty^2=p_t=p_2+\tfrac{1}{2}\rho_\infty U_2^2.$$

Note that, since the local inviscid velocity (U_2) may be greater than or less than the free-stream velocity (U_∞), the local static pressure (p_2) may be less than or greater than the free-stream static pressure (p_∞).

What about the pressure and the velocity at point 3, which is adjacent to the surface of the airfoil but at the same streamwise location as point 2? Since there is the "no-slip" boundary condition at the wall, $U_3=0$. From our discussion of thin boundary layers, $\partial p/\partial y \approx 0$ across the boundary layer. As a result, the static pressure at point 2 equals the static pressure at point 3, i.e., $p_2 = p_3$. This combination of pressure and velocity does not violate Bernoulli's equation, because point 3 lies buried within a viscous region. Therefore, the conditions for Bernoulli's equation are not met. Thus, we cannot apply Eq. (16) to determine the pressure or the velocity at point 3.

The engineer often uses experimental data or theoretical solutions for one flow condition to gain insight into the flow field that exists at another flow condition. Wind-tunnel data, where scale models are exposed to flow conditions that simulate the design flight environment, are used to gain insight to describe the full-scale flow field at other flow conditions. Therefore, it is most desirable to present (experimental or theoretical) correlations in terms of dimensionless coefficients that depend only on the configuration geometry and the angle of attack. This also eliminates problems associated with the choice of units. One such dimensionless coefficient is the pressure coefficient:

$$C_p = \frac{p - p_\infty}{\frac{1}{2}\rho_\infty U_\infty^2} = \frac{p - p_\infty}{q_\infty} . \tag{18}$$

This particular choice of a parameter to nondimensionalize the local static pressure can be understood by referring to Bernoulli's equation without the potential-energy term. Rearranging Eq. (16), we find

$$C_p = \frac{p - p_\infty}{\frac{1}{2}\rho_\infty U_\infty^2} = 1 - \frac{U^2}{U_\infty^2} . \tag{19}$$

Thus, at the stagnation point, where the local velocity is zero, $C_p = 1.0$ for an incompressible flow. Note that the stagnation point value is independent of the free-stream flow conditions, the configuration geometry, and the choice of units.

2.1.4 Pressure Gradients. Both Bernoulli's equation and Euler's equation indicate that, for an inviscid flow with negligible body forces,

$$dp = -\rho U\, dU \tag{20}$$

along a streamline. Thus, if dp/dx is positive, i.e., the pressure increases in the streamwise direction, it is called an adverse pressure gradient, since it acts against the fluid motion and the flow decelerates. Similarly, if dp/dx is negative, it is called a favorable pressure gradient and the flow accelerates. The sign of the pressure gradient is of great importance when the pressure field established by the inviscid flow interacts with the boundary layer. This will be discussed later.

2.1.5 Irrotational Flows. If $\nabla \times \mathbf{V} = 0$, the flow is said to be *irrotational.* According to Lord Kelvin's theorem, a flow remains irrotational in the absence of viscous forces or discontinuities provided that the fluid is homogeneous and the body forces can be described by a potential function. A homogeneous (or barotropic) fluid is one in which the density depends only on the pressure. Such is the case for incompressible flow (since ρ is a constant) and for isentropic, compressible flow (for which p is proportional to ρ^γ).

In most flow problems (certainly those of interest to this article), the undisturbed, free-stream flow is a uniform parallel flow in which there are no shear stresses. Kelvin's theorem implies that, although the fluid particles in the subsequent flow patterns may follow curved paths, the flow remains irrotational except in those regions where the dissipative viscous forces are an important factor.

For an irrotational flow, the velocity may be expressed in terms of a potential function:

$$\mathbf{V} = \nabla \phi . \tag{21}$$

For relatively low-speed flows (i.e., incompressible flows), the continuity equation is

$$\nabla \cdot \mathbf{V} = 0 . \tag{22}$$

Combining Eqs. (21) and (22), one finds that for an incompressible, irrotational flow,

$$\nabla^2 \phi = 0 . \tag{23}$$

Thus, the governing equation, which is known as *Laplace's equation,* is a linear, second-order partial differential equation.

Within certain constraints on geometric slope continuity, a bounded, simply connected velocity field is uniquely determined by the distribution on the flow boundaries either of the normal component of the total velocity, $\nabla\phi \cdot \hat{n}$, or of the total potential ϕ. These boundary-value problems are respectively designated Neumann or Dirichlet problems. For applications in this article, the Neumann formulation is most common since most practical cases involve prescribed normal velocity boundary conditions. Specifically, the flow tangency requirement associated with inviscid flow past a solid body is expressed mathematically as

$$\nabla\phi \cdot \hat{n} = 0 .$$

The computational solutions to the flow field around the vehicle may be obtained using surface-singularity methods or by field methods. For surface-singularity (or panel) methods, the configuration is modeled by a large number of panels. To each panel, there is attached one or more types of singularities, such as sources, vortices, and doublets. (Since Laplace's equation is linear, individual solutions can be added to each other provided the boundary conditions are satisfied.) The strengths of the singularities are determined by applying the appropriate boundary conditions. Once the singularity strengths have been determined, the velocity and pressure fields can be determined. For the field methods, solutions to the Euler equations or to Laplace's equation are generated for the flow around the configuration.

2.1.6 Isentropic, Compressible Flows. The use of the first law of thermodynamics to obtain the energy equation does not place any constraints on what types of processes are physically possible and what types are not. We know from experience that not all processes permitted by the first law actually occur in nature. It is the second law of thermodynamics which provides insight into which processes may occur. It is not possible for a process to occur if the entropy of the system and its surroundings decreases.

However, for an adiabatic, frictionless process, there is no change in entropy, i.e., the flow is isentropic. This is not an overly restrictive assumption, since heat transfer in aerodynamic problems usually occurs only in the boundary layer and has a negligible effect on the flow in the inviscid region. Thus, for an inviscid flow, one can derive the following from Gibbs equation:

$$p/\rho^{\gamma}=\text{const.} \tag{24}$$

The energy equation for a one-dimensional, adiabatic flow requires that the total enthalpy (H) is constant:

$$H=h+U^2/2, \tag{25}$$

where h (the static enthalpy) is equal to $e+p/\rho$. For a (calorically) perfect gas, the enthalpy is proportional to the temperature:

$$H=c_pT_t, \tag{26a}$$

$$h=c_pT, \tag{26b}$$

i.e., the specific heat at constant pressure is constant. The speed of sound is

$$a=\left[\left(\frac{\partial p}{\partial \rho}\right)_s\right]^{1/2}=\sqrt{\gamma RT}. \tag{27}$$

Combining Eqs. (24)–(27), we obtain

$$p_{t1}=p\left(1+\frac{\gamma-1}{2}M^2\right)^{\gamma/(\gamma-1)}. \tag{28}$$

In Eq. (28), p_{t1} is that stagnation pressure that is achieved when a flow whose Mach number is M and whose static pressure is p is brought to rest isentropically. We have used the subscript $t1$ (rather than merely t, as was done with Bernoulli's equation), since the total pressure changes when shock waves are present. As will be discussed later, shock waves are possible when the flow is supersonic. The ratio of specific heats (γ) is 1.4 for perfect air.

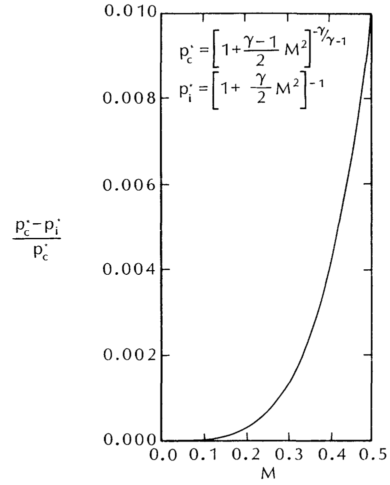

FIG. 7. The effect of compressibility on the theoretical value for the pressure ratio.

If we use Eq. (27) to express the parameters in Bernoulli's equation in terms of the Mach number, we obtain

$$p_i^*=\frac{p}{p_t}=\left(1+\frac{\gamma}{2}M^2\right)^{-1},$$

where the superscript $*$ represents the dimensionless pressure for incompressible flow. By rearranging Eq. (28), we obtain

$$p_c^*=\frac{p}{p_{t1}}=\left(1+\frac{\gamma-1}{2}M^2\right)^{-\gamma/(\gamma-1)}$$

as a dimensionless pressure ratio for compressible flow. Using these two parameters, we can estimate the effect of compressibility on the theoretical pressure ratio from the correlation presented in Fig. 7. This parameter increases rapidly with Mach number, reaching a value of 0.01 when the Mach number is 0.5. Thus, beginning with Mach numbers of approximately 0.5, one should take into account the effects of compressibility when solving for the flow field.

2.2 Viscous Flow

The general differential equations for the conservation of linear momentum, i.e., the Navier–Stokes equations, Eqs. (5a)–(5c), de-

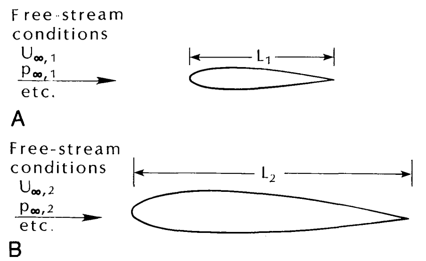

FIG. 8. Nomenclature for a flow around a prototype configuration and that around a scale model: (a) model flow; (b) prototype flow.

fine the motion in the viscous boundary layer adjacent to the surface. The Navier–Stokes equations for a general viscous flow are nonlinear partial differential equations. Even today, there are no general solutions for the complete Navier–Stokes equations. However, by evaluating the relative magnitude of the terms and eliminating the less significant terms for a particular application, one can introduce simplifying assumptions that allow us to obtain solutions of the viscous flow more easily. Advances in computer software and hardware allow us to generate flow-field solutions using reasonably rigorous models of the flow. Nevertheless, a turbulent boundary layer flowing over a complex airplane still provides a formidable challenge to the analyst.

Let us now consider the question of fluid dynamic similarity. Consider an incompressible flow around two objects that are geometrically similar (but of different size), as shown in Fig. 8. It is as if you took a photograph of the prototype and reduced it until it exactly matched the model configuration. For a constant-property flow with negligible body forces, the x component of the momentum equation [Eq. (5a)] becomes

$$\rho\frac{\partial u}{\partial t}+\rho u\frac{\partial u}{\partial x}+\rho v\frac{\partial u}{\partial y}+\rho w\frac{\partial u}{\partial z}$$
$$=-\frac{\partial p}{\partial x}+\mu\frac{\partial^2 u}{\partial x^2}+\mu\frac{\partial^2 u}{\partial y^2}+\mu\frac{\partial^2 u}{\partial z^2}. \tag{29}$$

Let us divide each of the thermodynamic properties by the value of that property at a point far from the vehicle (i.e., the free-stream value of the property) for each of the two flows. Thus, for the first flow

$$p_1^*=\frac{p}{p_{\infty,1}},\quad \rho_1^*=\frac{\rho}{\rho_{\infty,1}},\quad \mu_1^*=\frac{\mu}{\mu_{\infty,1}};$$

and for the second flow

$$p_2^*=\frac{p}{p_{\infty,2}},\quad \rho_2^*=\frac{\rho}{\rho_{\infty,2}},\quad \mu_2^*=\frac{\mu}{\mu_{\infty,2}}.$$

Note that the free-stream values for all three nondimensionalized ($*$) thermodynamic properties are unity for both cases.

Similarly, let us divide the velocity components by the free-stream velocity. Thus, for the first flow

$$u_1^*=\frac{u}{U_{\infty,1}},\quad v_1^*=\frac{v}{U_{\infty,1}},\quad w_1^*=\frac{w}{U_{\infty,1}};$$

and for the second flow

$$u_2^*=\frac{u}{U_{\infty,2}},\quad v_2^*=\frac{v}{U_{\infty,2}},\quad w_2^*=\frac{w}{U_{\infty,2}}.$$

With the velocity components thus nondimensionalized, the free-stream boundary conditions are the same for both flows. That is, at points far from the vehicle,

$$u_1^*=u_2^*=1 \quad\text{and}\quad v_1^*=v_2^*=w_1^*=w_2^*=0.$$

A characteristic dimension L is used to nondimensionalize the independent variables:

$$x_1^*=\frac{x}{L_1},\quad y_1^*=\frac{y}{L_1},\quad z_1^*=\frac{z}{L_1},\quad t_1^*=\frac{tU_{\infty,1}}{L_1}$$

and

$$x_2^*=\frac{x}{L_2},\quad y_2^*=\frac{y}{L_2},\quad z_2^*=\frac{z}{L_2},\quad t_2^*=\frac{tU_{\infty,2}}{L_2}.$$

In terms of these dimensionless parameters, Eq. (29) becomes

$$\rho_1^*\frac{\partial u_1^*}{\partial t_1^*}+\rho_1^* u_1^*\frac{\partial u_1^*}{\partial x_1^*}+\rho_1^* v_1^*\frac{\partial u_1^*}{\partial y_1^*}+\rho_1^* w_1^*\frac{\partial u_1^*}{\partial z_1^*}$$
$$=-\left[\frac{p_{\infty,1}}{\rho_{\infty,1}U_{\infty,1}^2}\right]\frac{\partial p_1^*}{\partial x_1^*}$$
$$+\left[\frac{\mu_{\infty,1}}{\rho_{\infty,1}U_{\infty,1}L_1}\right]\left(\mu_1^*\frac{\partial^2 u_1^*}{\partial x_1^{*2}}+\mu_1^*\frac{\partial^2 u_1^*}{\partial y_1^{*2}}\right.$$
$$\left.+\mu_1^*\frac{\partial^2 u_1^*}{\partial z_1^{*2}}\right) \tag{30a}$$

for the first flow. For the second flow

$$\rho_2^* \frac{\partial u_2^*}{\partial t_2^*}+\rho_2^* u_2^* \frac{\partial u_2^*}{\partial x_2^*}+\rho_2^* v_2^* \frac{\partial u_2^*}{\partial y_2^*}+\rho_2^* w_2^* \frac{\partial u_2^*}{\partial z_2^*}$$

$$=-\left[\frac{p_{\infty,2}}{\rho_{\infty,2}U_{\infty,2}^2}\right]\frac{\partial p_2^*}{\partial x_2^*}+\left[\frac{\mu_{\infty,2}}{\rho_{\infty,2}U_{\infty,2}L_2}\right]$$

$$\times\left(\mu_2^* \frac{\partial^2 u_2^*}{\partial x_2^{*2}}+\mu_2^* \frac{\partial^2 u_2^*}{\partial y_2^{*2}}+\mu_2^* \frac{\partial^2 u_2^*}{\partial z_2^{*2}}\right). \quad (30b)$$

Both the dependent and independent variables have been nondimensionalized, as indicated by the quantities with asterisks. The dimensionless *boundary-condition values* for the dependent variables are the same for the two flows around geometrically similar configurations. As a consequence, the solutions of the two problems in terms of the dimensionless variables will be identical, provided that the differential equations are identical. The differential equations will be identical if the dimensionless parameters in the brackets have the same values for both problems. In this case, the flows are said to be dynamically similar, as well as geometrically similar.

Let us now consider the first similarity parameter in Eq. (30):

$$p_\infty/\rho_\infty U_\infty^2 \,. \quad (31)$$

This similarity parameter can be interpreted in relation to gas flows. Recall that for a perfect gas the equation of state is

$$p_\infty=\rho_\infty RT_\infty,$$

and the free-stream speed of sound is given by

$$a_\infty=(\gamma RT_\infty)^{1/2}.$$

Substitution of these relations into Eq. (31) yields

$$\frac{p_\infty}{\rho_\infty U_\infty^2}=\frac{RT_\infty}{U_\infty^2}=\frac{a_\infty^2}{\gamma U_\infty^2}=\frac{1}{\gamma M_\infty^2}\,, \quad (32a)$$

since the *Mach number* is defined as

$$M_\infty=U_\infty/a_\infty \,. \quad (32b)$$

The inverse of the second similarity parameter is written as

$$\rho_\infty U_\infty L/\mu_\infty \,, \quad (33)$$

which is the *Reynolds number*, a measure of the ratio of inertia forces to viscous forces. Named after Osborne Reynolds, a British engineer, the Reynolds number is important for all flows in which we account for the effects of viscosity. Since the coefficient of viscosity is never zero, the Reynolds number is one of the most important similarity parameters. However, as noted earlier, there are many regions of the flow field where the product of the viscosity times the shear velocity gradient is negligible. Since the viscous terms of Eqs. (5a)–(5c) (i.e., the Navier–Stokes equations) are negligible for these inviscid flows, the Reynolds number would not be important for inviscid flows.

The Reynolds number may vary over several orders of magnitude during the entry flight of a hypersonic vehicle. During the high-altitude portion of the flight, i.e., the initial stages of reentry, the density and, therefore, the Reynolds number are very low. As the spacecraft descends towards the earth's surface, the density and the Reynolds number increase dramatically. Typical Reynolds numbers for some representative vehicles are presented in Fig. 9. The Reynolds number is an important parameter when trying to determine if the boundary layer is laminar or turbulent. Although there are many phenomena that influence the process by which the boundary layer transitions from a laminar state to a turbulent state, the lower the Reynolds number, the more likely the flow is to be laminar.

2.2.1 Boundary-Layer Characteristics.

When the boundary layer is laminar, the transverse exchange of momentum (i.e., the momentum transfer in a direction perpendicular to the principal flow direction) takes place on a microscopic scale. As a result of the molecular movement, slower-moving fluid particles from the lower layer (or lamina) of fluid move upward, slowing the particles in the upper layer. Conversely, when the faster-moving fluid particles from the upper layer migrate downward, they tend to accelerate the fluid particles in that layer. This molecular interchange of momentum for a laminar flow is depicted in Fig. 10(a).

For a turbulent boundary layer, there is a macroscopic transport of fluid particles, as shown in Fig. 10(b). Thus, in addition to the laminar shear stress described in the preceding paragraph, there is an effective *turbulent shear stress* that is due to the transverse trans-

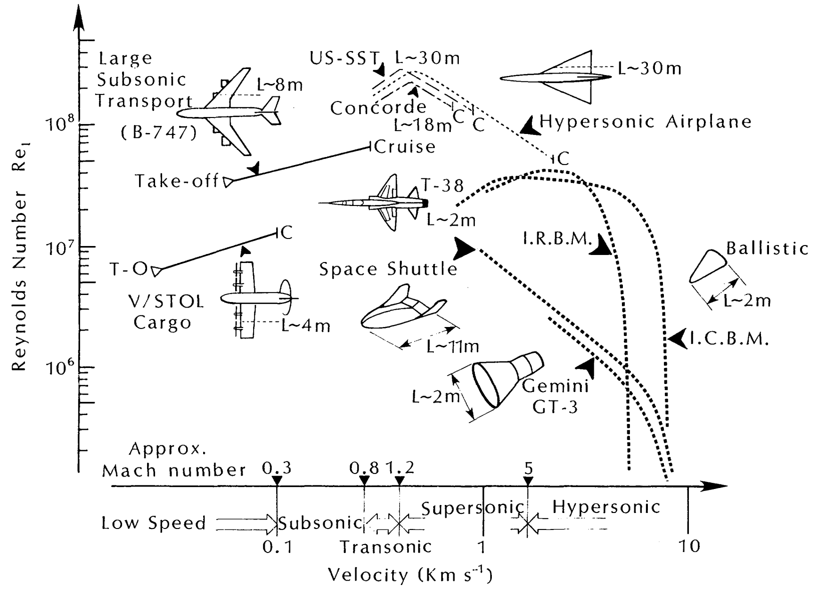

FIG. 9. Typical Reynolds numbers for some representative vehicles.

port of momentum which is very large. Because slower moving fluid particles near the wall are transported well upward, the turbulent boundary layer is relatively thick. Because faster-moving fluid particles (which are normally located near the edge of the boundary layer) are transported toward the wall, they produce relatively high velocities for the fluid particles near the surface. Thus, the shear stress at the wall for a turbulent boundary layer is larger than that for a laminar boundary layer, and yet the turbulent boundary layer is thicker. Because the macroscopic transport of fluid introduces large localized variations in the flow at any instant, the values of the fluid properties and the velocity components are (in general) the sum of the "average" value and a fluctuating component.

2.2.2 Boundary-Layer Transition. Near the forward stagnation point on a wing or near the leading edge of a flat plate, the boundary layer is laminar. As the flow proceeds downstream, the boundary layer thickens and the viscous forces continue to dissipate the energy of the airstream. Disturbances to the flow in the growing viscous layer may be caused by surface roughness, a temperature variation in the surface, pressure pulses, and so on. If the Reynolds number is low, the disturbances will be damped by viscosity and the boundary layer will remain

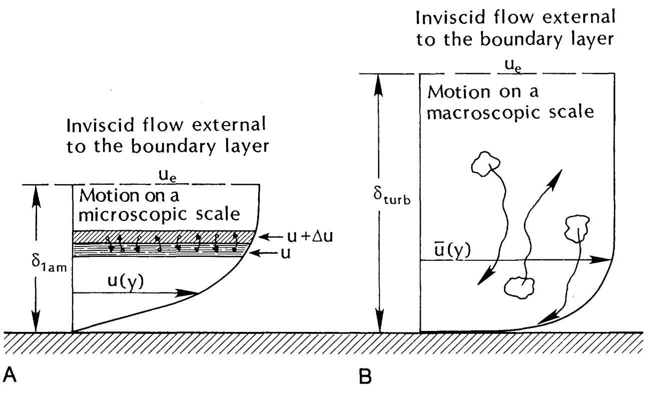
Inviscid flow external to the boundary layer
u_e
Motion on a microscopic scale
$u+\Delta u$
u
δ_{lam}
u(y)
A
Inviscid flow external to the boundary layer
u_e
Motion on a macroscopic scale
δ_{turb}
$\bar{u}(y)$
B

FIG. 10. Momentum-transport models: (a) laminar boundary layer; (b) turbulent boundary layer.

laminar. At higher Reynolds numbers, the disturbances grow. In such cases, the boundary layer may become unstable and, eventually, turbulent (i.e., transition will occur). The details of the transition process are quite complex and depend on many parameters.

The engineer who must develop a transition criterion for design purposes usually uses the Reynolds number. For instance, if the surface of a flat plate is smooth and if the external airstream has no turbulence, transition "occurs" at a Reynolds number (Re_x) of approximately 500 000. However, experience has shown that the Reynolds number at which the disturbances will grow and the length over which the transition process takes place depend on the magnitude of the disturbances and on the flow field. Let us consider briefly the effect of surface roughness, of the surface temperature, of a pressure gradient in the inviscid flow, and of the local Mach number on transition.

1. *Surface roughness.* Since transition is the amplification of disturbances to the flow, the presence of surface roughness significantly promotes transition (i.e., causes transition to occur at relatively low Reynolds numbers).
2. *Surface temperature.* The boundary-layer thickness decreases as the surface temperature is decreased. Cooling the surface usually delays transition.
3. *Pressure gradient.* A favorable pressure gradient (i.e., the static pressure decreases in the streamwise direction or, equivalently, the inviscid flow is accelerating) delays transition. Conversely, an adverse pressure gradient promotes transition.
4. *Mach number.* The transition Reynolds number is usually higher when the flow is compressible (i.e., as the Mach number is increased).

Stability theory, e.g., Mack (1984), can be an important tool that provides insights into the importance of individual parameters without introducing spurious effects that might be due to flow disturbances peculiar to a test facility (e.g., "noise" in a wind tunnel). For a more detailed discussion of transition, the reader is referred to a textbook on boundary-layer theory, e.g., Schlichting (1979).

If the skin friction is the dominant component of the drag, transition should be delayed as long as possible to obtain a low-drag section. To delay transition on a low-speed airfoil section, the point of maximum thickness should be moved aft so that the boundary layer is subjected to a favorable pressure gradient over a longer run. Consider the NACA 0009 section and the NACA 66-009 section (from the National Advisory Committee for Aeronautics), as shown in Fig. 11. Both are symmetric, having a maximum thickness of 0.09*c*. The parameter *c* is the *chord length* (as measured along a line from the leading edge to the trailing edge) of the airfoil. The maximum thickness for the NACA 66-009 section is at 0.45*c*, while that for the NACA 0009 section is at 0.3*c*. As a result, the minimum pressure coefficient occurs at $x=0.6c$ for the NACA 66-009 and a favorable pressure gradient acts to stabilize the boundary layer to this point. For the NACA 0009, the minimum pressure occurs near $x=0.1c$. The lower local velocities near the leading edge and the extended region of favorable pressure gradient cause transition to be farther aft on the NACA 66-009. Since the drag for a streamlined airfoil at low angles of attack is primarily due to skin friction, one would expect that the drag is lower for the laminar flow for the NACA 66-009. This is verified by the data from Abbott and von Doenhoff (1949), which are reproduced in Fig. 12. The subsequent reduction in

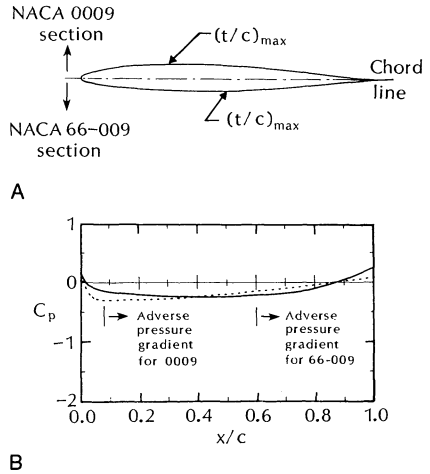

FIG. 11. Comparison of geometries and resultant pressure distributions for a "standard" airfoil section (NACA 0009) and for a laminar airfoil section (NACA 66-009): (a) comparison of cross section for an NACA 0009 airfoil with that for an NACA 66-009 airfoil; (b) static pressure distribution.

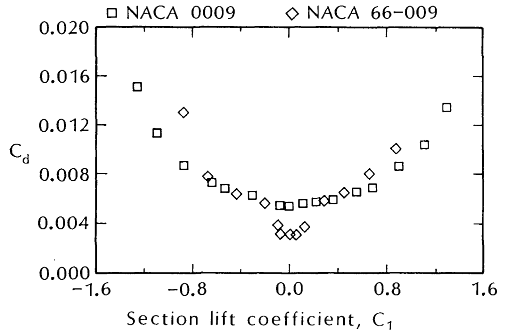

FIG. 12. Section drag coefficients for NACA 0009 airfoil and the NACA 66-009 airfoil, $Re_c = 6 \times 10^6$ [data from Abbott and von Doenhoff (1949)].

the friction drag for the NACA 66-009 section creates a *drag bucket*, i.e., that region of relatively low drag that occurs when the section lift coefficient is small. Note that the section drag coefficient varies only slightly with C_l for small excursions in angle of attack near zero lift, since the skin-friction coefficient varies little with angle of attack. At the very high Reynolds numbers that occur at some flight conditions, it is difficult to maintain a long run of laminar boundary layer, especially if surface roughnesses develop during the flight operations. However, a *laminar-flow section*, such as the NACA 66-009, offers additional benefits. Comparing the cross sections presented in Fig. 11, the cross section of the NACA 66-009 airfoil provides more flexibility for carrying fuel and for accommodating the load-carrying structure.

For larger angles of attack, the section drag coefficient depends both on Reynolds number and on angle of attack. As the angle of attack and the section lift coefficient increase, the minimum pressure coefficient decreases. The adverse pressure gradient that results as the flow decelerates toward the trailing edge increases. When the air particles in the boundary layer, already slowed by viscous action, encounter the relatively strong adverse pressure gradient, the boundary layer thickens and separates. The term *stall* is used to describe the flow situation when the boundary layer separates from the airfoil surface. Because the thickening boundary layer and its separation from the surface cause the pressure distribution to be significantly different from the inviscid model at the higher angles of attack, form drag dominates. Note that, at the higher angles of attack (where form drag is important), the drag coefficient for the NACA 66-009 is greater than that for the NACA 0009.

Thus, the aerodynamic forces and moments are a function of vehicle geometry and angle of attack as well as Mach number and Reynolds number. By carefully tailoring the configuration, one can obtain significant improvements in the aerodynamic coefficients for flight conditions of most importance.

2.2.3 The Effect of the Boundary Layer on Flow Separation. A turbulent flow is one in which irregular fluctuations (mixing or eddying motions) are superimposed on the mean flow. Thus, the velocity at any point in a turbulent boundary layer is a function of time, even for a flow where the free-stream conditions and wall-boundary conditions are steady. The fluctuations occur in the direction of the mean flow and at right angles to it, and they affect macroscopic lumps of fluid. Thus, whereas momentum transport occurs on a microscopic (or molecular) scale in a laminar boundary layer, it occurs on a macroscopic scale in a turbulent boundary layer. It should be noted that, although the velocity fluctuations may be only several percent of the local streamwise values, they have a decisive effect on the overall motion. The size of these macroscopic lumps determines the scale of turbulence.

The effects caused by the fluctuations are as if the viscosity were increased by a factor of 10 or more. As a result, the shear forces at the wall and the skin-friction component of the drag are much larger when the boundary layer is turbulent. However, because of the macroscopic momentum transfer, air particles near the wall have a relatively high velocity (see Fig. 10). Thus, a turbulent boundary layer can negotiate an adverse pressure gradient for a longer distance. As a result, boundary-layer separation may be delayed or even avoided altogether if the boundary layer is turbulent. Delaying (or avoiding) the onset of separation reduces the pressure component of the drag (i.e., the form drag). For a blunt body or for a slender body at angle of attack, the reduction in form drag usually dominates the increase in skin-friction drag.

Figure 13 illustrates how a turbulent boundary layer can flow without separation into regions of much steeper adverse pressure

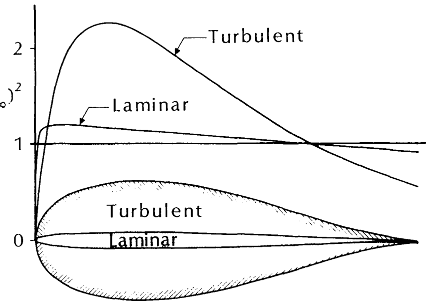

FIG. 13. Thickest symmetrical Joukowski airfoils capable of supporting fully attached laminar and turbulent flows. The angle of attack is 0°, and the Mach number is low. For turbulent flow, transition is assumed to occur at the velocity peak. The turbulent case is calculated for $Re_c = 10^7$. Results for laminar flow are independent of Reynolds number. Maximum thickness for laminar flow is about 0.046*c*, for turbulent flow, 0.31*c*. If displacement-thickness effects on pressure distribution were included, the turbulent airfoil would increase to about 0.33*c*. The change in the laminar case would be negligible. [From Cebeci and Smith (1974).]

gradients than can a laminar boundary layer. Incompressible boundary-layer solutions were generated for symmetrical Joukowski airfoils at zero angle of attack. The edge velocity and, therefore, the corresponding inviscid pressure distributions, are presented in Fig. 13. At the conditions indicated, boundary-layer separation will occur for any Joukowski airfoil that is thicker than 0.046*c*, if the flow is entirely laminar. However, if the boundary layer is turbulent, separation will not occur until a thickness of about 0.31*c* has been exceeded. The boundary layer effectively thickens the airfoil, especially near the trailing edge, since the boundary-layer thickness increases with distance. This thickening alleviates the adverse pressure gradients, which in turn permits somewhat thicker sections before separation occurs. To ensure that boundary-layer transition occurs and, thus, to delay or avoid separation altogether, one might use vortex generators or other forms of surface roughness, such as shown in Fig. 14.

2.3 Shock Waves

The formation of a *shock wave* occurs when a supersonic flow decelerates in response to a sharp increase in pressure or when a supersonic flow encounters a sudden, compressive change in direction. The shock wave is on the order of several mean free paths thick, and for flow conditions where the gas behaves as a continuum this distance is very small ($\sim 6 \times 10^{-6}$ cm at atmospheric pressure and temperature). Thus, there is an almost instantaneous change in the values of the flow parameters across a shock in a continuum flow. Because of the large streamwise variations in velocity, pressure, and temperature, viscous and heat-conduction effects are important within the shock wave. The difference between a shock wave and a Mach wave should be kept in mind. A *Mach wave* represents a surface across which some derivatives of the flow variables (such as the thermodynamic properties of the fluid and the flow velocity) may be discontinuous while the variables themselves are continuous. A *shock wave* represents a surface across which the thermodynamic properties and the flow velocity are essentially discontinuous. Thus, the characteristic curves, or Mach lines, are patching lines for continuous flows, whereas shock waves are patching lines for discontinuous flows.

Consider the curved shock wave illustrated in Fig. 15. The flow upstream of the shock wave, which is stationary in the body-fixed-coordinate system, is supersonic and designated by the subscript 1. The downstream values are designated by subscript 2. At the plane of symmetry, the shock wave is normal (or perpendicular) to the free-stream flow, and the flow downstream of the shock wave is subsonic. Away from the plane of symmetry, the shock wave is oblique and the downstream flow may be subsonic or supersonic, depending on the free-stream Mach number and the turning angle (δ). In order to have a

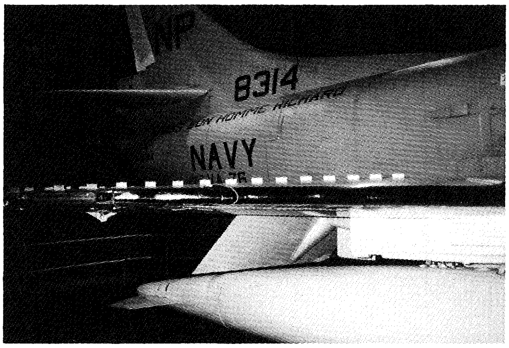

FIG. 14. Row of vortex generators on the wing of the A4 (author's collection).

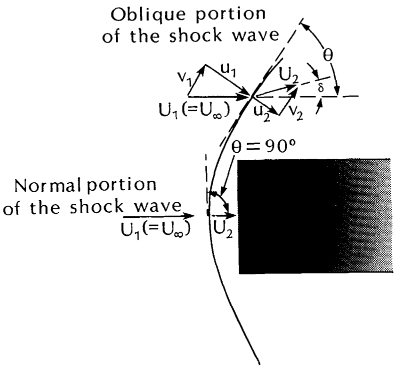

FIG. 15. Curved shock wave illustrating nomenclature for normal shock wave and for oblique shock wave.

shock wave, the upstream Mach number must be greater than 1, i.e., shock waves will not occur in subsonic flow. Referring to Fig. 16, the flow just downstream of a normal shock wave is subsonic, i.e., M_2 is less than 1. As noted earlier, shock waves occur in supersonic flow to accomplish compressive turns of the flow or to decelerate the flow (increase the pressure) to match boundary conditions. The entropy increases across the shock wave and the total (or stagnation) pressure decreases, i.e., p_{t2} is less than p_{t1}. The property variations across a normal shock wave are presented in Fig. 16.

Note that the density increases as the air particles cross the shock wave. The density gradients produced by a shock wave and the boundary layer in a compressible flow will bend a parallel beam of light passing through the flow field. The deflection of light by the changes in density will produce variations in the intensity of light impinging on a photographic plate. Thus, we can photographically record regions of the flow field with large density gradients. Shock waves are clearly evident in the schlieren photograph of a biconic configuration in a Mach 4.25 stream in the Trisonic Wind Tunnel at the U.S. Air Force Academy, presented in Fig. 17. Under certain circumstances, these density gradients allow us to "see" the shock waves for an airplane or for a missile in flight, as for the supersonic flight of a rocket-powered vehicle presented in Fig. 18.

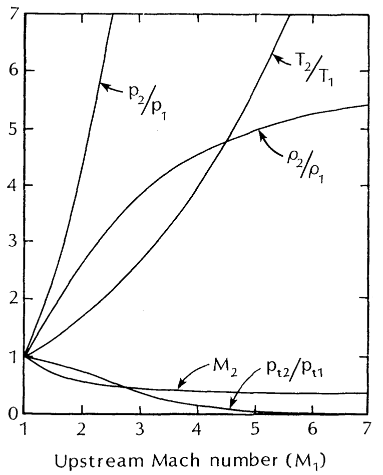

FIG. 16. Property variations across a normal shock wave.

3. CHARACTERIZATION OF AERODYNAMIC FORCES

The motion of the air around the vehicle produces pressure and velocity variations which produce the aerodynamic forces and moments. Although viscosity is a fluid property and, therefore, acts throughout the flow field, the viscous forces depend on the velocity gradients as well as on the viscosity. For many aerodynamics problems, the viscous forces are significant only within a narrow region near the wall where the velocity gradients are large (the boundary layer), or within the viscous wake should boundary-layer separation occur. When the viscous effects are confined to the boundary layer, it is possible

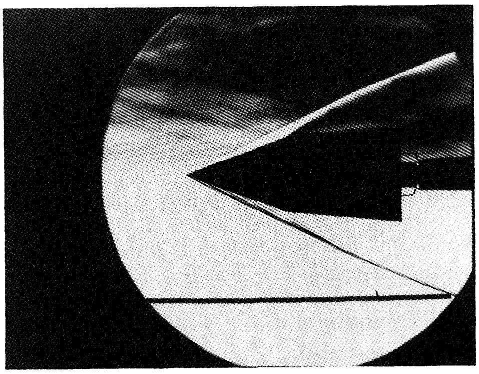

FIG. 17. Schlieren photograph of flow past a sharp biconic in a Mach 4.25 stream (photo courtesy of the U.S. Air Force Academy).

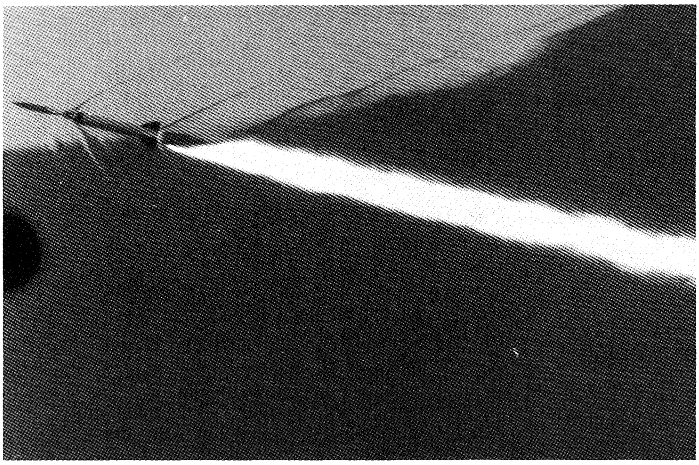

FIG. 18. Rocket-powered vehicle flying at Mach 3 with shock waves clearly visible (courtesy of R. C. Maydew of the Sandia National Laboratories).

to use the equations of motion for an inviscid flow to determine the pressure distribution around the vehicle and, hence, the velocity of the air particles at the edge of the boundary layer. With the inviscid flow field known, the velocity distribution across the viscous boundary layer, and hence the tangential shear forces at the wall, can be calculated. The normal and tangential forces, which act on the surface due to the fluid motion around an airfoil, are shown in Fig. 3.

3.1 Aerodyamic Forces and Moments

It is conventional to resolve the sum of the normal, or pressure, forces and the tangential, or viscous shear, forces into three components along axes parallel and perpendicular to the free-stream direction. The three components are designated as lift, drag, and side force. The relation of the lift and the drag forces with respect to the free-stream velocity is shown in Fig. 3.

1. *Lift.* Lift is the force component acting upward, perpendicular to the undisturbed free-stream velocity. The aerodynamic lift is produced primarily by the pressure forces acting on the vehicle surface.
2. *Drag.* Drag is the net aerodynamic force acting in the same direction as the undisturbed free-stream velocity. The aerodynamic drag is produced by the pressure forces and by skin friction forces that act on the surface.
3. *Side force.* Side force is the component of force in a direction perpendicular to both the lift and the drag. The side force is positive when acting toward the starboard wing (i.e., the pilot's right).

The resulting aerodynamic force usually will not act through the origin of the airplane's axis system (e.g., the center of gravity). The moment due to the resultant force acting at a distance from the origin may be divided into three components, referred to the airplane's reference axes. The three moment components are the pitching moment, the rolling moment, and the yawing moment, as shown in Fig. 19.

1. *Pitching moment.* The moment about the lateral axis (the y axis of the airplane-fixed-coordinate system) is the pitching moment. The pitching moment is the result of the lift and the drag forces acting on the vehicle. A positive pitching moment is in the nose-up direction.
2. *Rolling moment.* The moment about the longitudinal axis of the airplane (the x axis) is the rolling moment. A rolling mo-

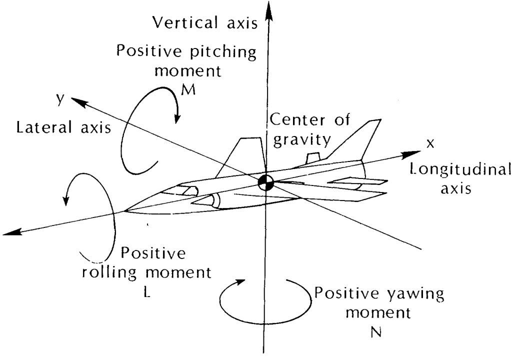

FIG. 19. Reference axes of the airplane and the corresponding aerodynamic moments.

ment is often created by a differential lift, generated by some type of ailerons or spoilers. A positive rolling moment causes the right, or starboard, wingtip to move downward.

3. *Yawing moment.* The moment about the vertical axis of the airplane (the z axis) is the yawing moment. A positive yawing moment tends to rotate the nose to the pilot's right.

3.2 Parameters that Govern Aerodynamic Forces

The magnitude of the forces and the moments that act on a vehicle depend on the combined effects of many different variables. Personal observations of the aerodynamic forces acting on an arm extended from a car window or on a ball in flight demonstrate the effect of velocity and of configuration. Pilot manuals advise that a longer length of runway is required if the ambient temperature is relatively high or if the airport elevation is high (i.e., the ambient density is relatively low). The parameters that govern the magnitude of aerodynamic forces and moments include the following:

1. configuration geometry,
2. angle of attack,
3. vehicle size or model scale,
4. free-stream velocity,
5. density of the undisturbed air,
6. Reynolds number (as it relates to viscous effects), and
7. Mach number (as it relates to compressibility effects).

The calculation of the aerodynamic forces and moments acting on a vehicle often requires that the engineer be able to relate data obtained at other flow conditions to the conditions of interest. Thus, the engineer often uses data from the wind tunnel, where scale models are exposed to flow conditions that simulate the design environment, or data from flight tests at other flow conditions. So that one can correlate the data for various stream conditions and configuration scales, the measurements are usually presented in dimensionless form. Ideally, once in dimensionless form, the results would be independent of all but the first two parameters listed above, configuration geometry and angle of attack. In practice, flow phenomena, such as boundary-layer separation, shock-wave–boundary-layer interactions, and compressibility effects, limit the range of flow conditions over which the dimensionless force and moment coefficients remain constant.

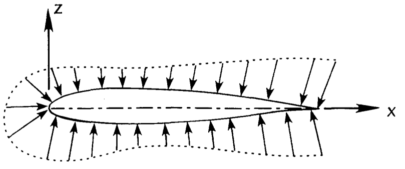

FIG. 20. Pressure distribution for a lifting airfoil section.

3.3 Lift Coefficient

To illustrate the physical significance of a dimensionless force coefficient, consider the normal force coefficient. The normal (or z) component of the resultant force is chosen since it is relatively simple to calculate and it has the same relation to the pressure and the shear forces as does the lift. For a relatively thin airfoil section at a relatively low angle of attack, it is clear from Fig. 3 that the lift (and similarly the normal force) results primarily from the action of the pressure forces. The shear forces act primarily in the chordwise direction (i.e., contribute primarily to the drag). Therefore, to calculate the force in the z direction, we need consider only the pressure contribution, which is illustrated in Fig. 20. The pressure force acting on a differential area of the vehicle surface is $dF = p\, ds\, dy$, as shown in Fig. 21. The elemental surface area is the product of ds, the wetted length of the element in the plane of the cross section, times dy, the element's length in the direction perpendicular to the plane of the cross section (or spanwise direction). Since the pressure force acts normal to the surface, the force component in the z direction is the product of the pressure times the projected planform area:

$$dF_z = p\, dx\, dy. \qquad (34)$$

Integrating over the entire wing surface (including the upper and the lower surfaces), the net force in the z direction is given by

$$F_z = \oint p\, dx\, dy. \qquad (35)$$

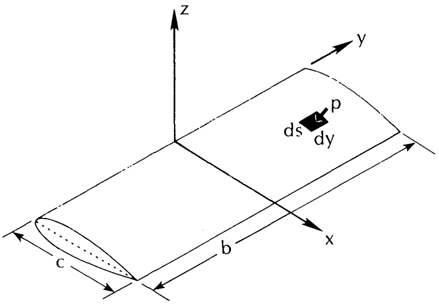

FIG. 21. Pressure acting on an elemental surface area.

Note that the resultant force in any direction due to a constant pressure acting over a closed surface is zero:

$$\oiint p_\infty \, dx \, dy = 0. \tag{36}$$

Combining Eqs. (35) and (36), the resultant force component is

$$F_z = \oiint (p - p_\infty) dx \, dy. \tag{37}$$

To nondimensionalize the factors on the right-hand side of Eq. (37), divide the pressure by q_∞ (the free-stream dynamic pressure), x by c (the chord length), and y by b (the span). Thus, we obtain the dimensionless parameter,

$$\frac{F_z}{q_\infty cb} = \oiint \frac{p - p_\infty}{q_\infty} \, d\left(\frac{x}{c}\right) d\left(\frac{y}{b}\right) .$$

Since the product cb represents the planform area S (i.e., the projection on a horizontal plane) of the rectangular wing of Fig. 21,

$$\frac{F_z}{q_\infty S} = \oiint C_p d\left(\frac{x}{c}\right) d\left(\frac{y}{b}\right) . \tag{38}$$

When the boundary layer is thin, the pressure distribution around the airfoil is essentially that of an inviscid flow. Thus, the pressure distribution is independent of Reynolds number and does not depend on whether the boundary layer is laminar or turbulent. When the boundary layer is thin, the pressure coefficient at a particular location on the surface given by the dimensionless coordinates $(x/c, y/b)$ is independent of vehicle scale and of the flow conditions. Over the range of flow conditions for which the pressure coefficient is a unique function of the dimensionless coordinates $(x/c, y/b)$, the value of the integral in Eq. (38) depends only on the configuration geometry and on the angle of attack. Thus, the resulting dimensionless force parameter, or force coefficient (in this case, the normal force coefficient), is independent of model scale and of flow conditions.

A similar analysis can be used to calculate the lift coefficient, which is defined as

$$C_L = L/q_\infty S. \tag{39}$$

3.4 Moment Coefficients

The moment created by the aerodynamic forces acting on a wing (or airfoil) is determined about a particular reference axis. The reference axis may be the leading edge, the quarter-chord location, the aerodynamic center, and so on. The aerodynamic center (or ac) is that point at which changes in lift effectively act. The procedure used to nondimensionalize the moments created by the aerodynamic forces is similar to that used to nondimensionalize the lift. To demonstrate this nondimensionalization, the pitching moment about the leading edge due to the pressures acting on the surface will be calculated (refer again to Fig. 21). The contribution of the chordwise component of the pressure force to the moment is small and is neglected. Thus, the pitching moment about the leading edge due to the pressure force acting on the surface element whose area is ds times dy, and which is located at a distance x from the leading edge, may be approximated by

$$dM_0 = p \, dx \, dy \, x, \tag{40}$$

where $dx \, dy$ is the projected area. Integrating over the entire wing surface, the net pitching moment is given by

$$M_0 = \oiint px \, dx \, dy. \tag{41}$$

When a uniform pressure acts on any closed surface, the resultant pitching moment due to this constant pressure is zero. Thus,

$$\oiint p_\infty x \, dx \, dy = 0. \tag{42}$$

By combining Eqs. (41) and (42), the resulting pitching moment about the leading edge is

$$M_0 = \oiint (p - p_\infty) \, x \, dx \, dy. \tag{43}$$

To nondimensionalize the factors on the right-hand side of Eq. (43), divide by $q_\infty c^2 b$, which has the units of force times length:

$$\frac{M_0}{q_\infty c^2 b}=\oiint \frac{p-p_\infty}{q_\infty}\frac{x}{c}\,d\left(\frac{x}{c}\right)d\left(\frac{y}{b}\right).$$

Since the product of *cb* represents the planform area of the wing, which is represented by the symbol *S*,

$$\frac{M_0}{q_\infty Sc}=\oiint C_p\frac{x}{c}\,d\left(\frac{x}{c}\right)d\left(\frac{y}{b}\right). \tag{44}$$

Thus, the dimensionless moment coefficient is

$$C_{M_0}=M_0/q_\infty Sc\,. \tag{45}$$

Since the derivation of Eq. (45) was for the rectangular wing of Fig. 21, the chord *c* is used. For a general wing, the mean aerodynamic chord is used together with *S* to nondimensionalize the pitching moment.

The section moment coefficient is used to represent the dimensionless moment per unit span (m_0) for an airfoil (i.e., for a two-dimensional flow):

$$C_{m_0}=m_0/q_\infty cc, \tag{46}$$

since the surface area per unit span is the chord length *c*.

The characteristic length (or moment arm) for the rolling moment and for the yawing moment is the wing span *b* (instead of the chord). Therefore, the rolling moment coefficient is

$$C_{\mathscr{L}}=\mathscr{L}/q_\infty Sb\,, \tag{47}$$

and the yawing moment coefficient is

$$C_N=N/q_\infty Sb. \tag{48}$$

3.5 Drag Coefficient

The drag force on a wing is due in part to skin friction and in part to the integrated effect of pressure. If $\mathbf{t}$ denotes the tangential shear stress at a point on the body surface, *p* the static pressure, and $\hat{n}$ the external normal to the element of surface *dS*, the drag can be formally expressed as

$$D=\oiint \mathbf{t}\cdot\hat{e}_\infty dS-\oiint p\hat{n}\cdot\hat{e}_\infty dS, \tag{49}$$

where $\hat{e}_\infty$ is a unit vector parallel to the free stream and the integration takes place over the entire wetted surface. The first integral represents the friction component and the second integral represents the pressure drag.

If an incompressible flow over a two-dimensional airfoil were truly inviscid, the pressure distribution would be such that the pressure coefficient would be 1, both at the leading edge and at the trailing edge, i.e., there would be a stagnation point at the trailing edge as well as at the leading edge. Furthermore, the pressure forces acting on the projected areas normal to the flow are balanced, so that the pressure drag would be zero. Of course, the drag due to skin friction is zero for a truly inviscid flow. Thus, for an incompressible, inviscid flow over a two-dimensional airfoil, the total drag is zero. Since this result is contrary to all experience, it is known as *d'Alembert's paradox.*

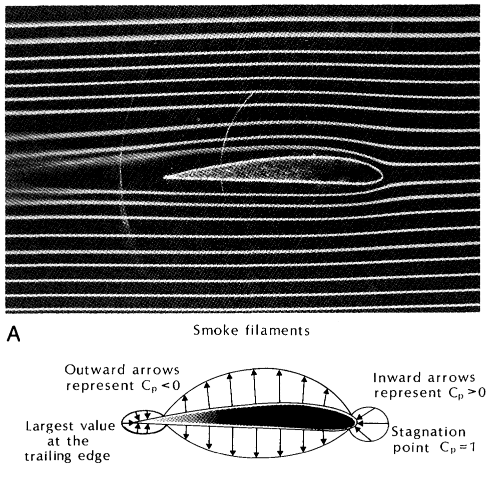
A
Smoke filaments
Outward arrows represent $C_p<0$
Inward arrows represent $C_p>0$
Largest value at the trailing edge
Stagnation point $C_p=1$
B
Pressure distribution (in terms of pressure coefficients)

FIG. 22. Incompressible flow past a two-dimensional airfoil at a small angle of attack.

In reality, the viscous forces in a boundary layer not only produce the drag force due to friction, but modify the pressure distribution, producing a pressure, or form, drag. Thus, for incompressible, viscous flow past a two-dimensional airfoil, both terms of Eq. (49) are due to the viscous effects. Equation (49) may be written in a slightly different fashion:

$$D=D_f+D_p. \tag{50}$$

At low angles of attack, the pressure distribution is similar to that shown in Figs. 11 and 22. The adverse pressure gradient that exists over much of the airfoil (see Fig. 11) thickens the boundary layer, altering the pressure distribution slightly from that for an inviscid flow. However, separation does not occur and D_p is relatively small. As a result, since the skin-friction component (D_f) is significant,

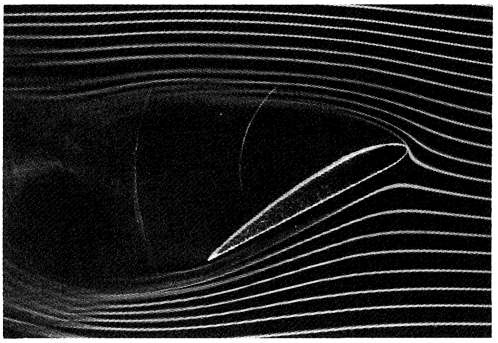

A Smoke filaments

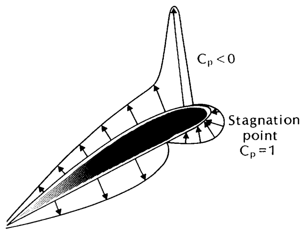

B Pressure distribution (in terms of pressure coefficients)

FIG. 23. Incompressible flow past a two-dimensional airfoil at a large angle of attack.

the drag is less when the boundary layer is laminar. This is illustrated in the data presented in Fig. 12. Note that the section drag coefficients for C_l near zero (i.e., at low angles of attack) are less for the NACA 66-009 laminar airfoil sections.

As the angle of attack increases, the flow accelerates rapidly over the top of the airfoil, creating relatively low pressures in this region (and generating lift). However, in a subsonic application, the flow decelerates toward the trailing edge. At these relatively large angles of attack, the air molecules already slowed by viscosity in the boundary layer cannot overcome the relatively large adverse pressure gradients, so they separate, i.e., break away from the surface, as shown in Fig. 23(a). Recall that this is known as stall. The pressure in the separated region is essentially constant at the (low) value corresponding to the point where separation occurred. Now there is a large difference between the pressure acting near the leading edge and that acting near the trailing edge. As a result, there is a large drag component due to pressure (D_p). This is indicated by the rapid increase in the drag coefficient with C_l (or, equivalently, the angle of attack) evident in the data of Fig. 12. Because a large adverse pressure gradient causes separation more quickly for a laminar boundary layer than for a turbulent one, the drag coefficient is greater for the laminar-flow airfoil section (NACA 66-009).

The drag coefficients presented thus far are for two-dimensional airfoils (i.e., configurations of infinite span). For a wing of finite span that is generating lift, the pressure differential between the lower surface and the upper surface causes a spanwise flow. The resultant three-dimensional flow field produces a drag component C_{D_v} which, as will be discussed later, is proportional to

$$C_L^2/\pi(\text{AR}). \qquad (51)$$

This is termed the *vortex drag*, the *induced drag*, or the *drag due to lift*. Let us introduce another term, designated ΔC_{D_M}, which will include compressibility effects and may, possibly, include the changes in the airflow accompanying thrust variations [according to Thomas (1985)]. Thus, the total drag for an airplane may be written as the sum of (1) the profile drag, which is independent of the lift coefficient, and therefore exists when the configuration generates zero lift (C_{D_0} or ΔC_{D_p}), (2) the drag associated with lift (kC_L^2), and (3) the compressibility-related effects that are correlated in terms of the Mach number and are known as *wave drag*:

$$C_D = C_{D_0} + kC_L^2 + \Delta C_{D_M}. \qquad (52)$$

Although the relative importance of the different drag sources depends on the aircraft type and the mission to be flown, approximate breakdowns (by category) for large, subsonic transports are presented in Fig. 24.

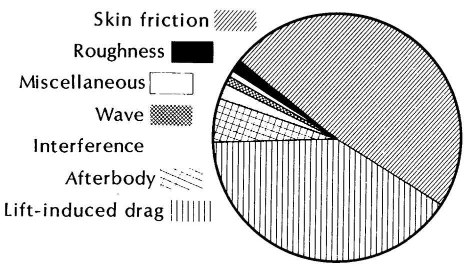

FIG. 24. Contributions of different drag sources for a typical transport aircraft. [From Thomas (1985).]

According to Thomas (1985), the greatest contribution arises from turbulent skin-friction drag. The next most significant contribution arises from lift-induced drag, which, when added to skin-friction drag, accounts for about 85% of the total drag of a typical transport aircraft. Thomas cited the pressure drag due to open separation in the afterbody and other regions, interference effects between aerodynamic components, wave drag due to the compressibility effect at near-sonic flight conditions, and miscellaneous effects, such as roughness effects and leakage, as constituting the remaining 15%.

3.5.1 Zero-Lift Drag. Skin friction and form-drag components can be calculated for the wing, tail, fuselage, nacelles, etc. When evaluating the zero-lift drag, one must consider interactions such as how the growth of the boundary layer on the fuselage reduces the boundary-layer velocities on the wing-root surface. Because of the interaction, the wing-root boundary layer is more easily separated in the presence of an adverse pressure gradient. Since the upper wing surface has the more critical pressure gradients, a low wing position on a circular fuselage would be sensitive to this interaction. Adequate filleting and control of the local pressure gradients can minimize the interaction effects.

A representative value of C_{D_0} would be 0.020, of which the wings may account for 50%, the fuselage and the nacelles 40%, and the tail 10%. Since the wing constitutes such a large fraction of the drag, reducing the wing area would reduce D_0 if all other factors are unchanged. Although other factors must be considered, the reasoning above implies that an optimum airplane configuration would have a minimum wing surface area and, therefore, highest practical wing loading (N/m^2).

3.5.2 Drag Due to Lift. The drag due to lift may be represented as

$$kC_L^2 = C_L^2/\pi(\text{AR})e, \tag{53}$$

where e is the airplane *efficiency factor* and AR the *aspect ratio*. Typical values of the airplane efficiency factor range from 0.6 to 0.9. At high lift coefficients (near $C_{L,\max}$), e should be changed to account for increased form drag. The deviation of the actual airplane drag from the quadratic correlation where e is a constant is significant for airplanes with low aspect ratios and sweepback.

3.5.3 Compressibility. Another factor to consider is the effect of compressibility. When the flight Mach number is equal to or greater than 0.5 (approximately), the drag increases very rapidly with speed. The correlation of Eq. (52) includes wave drag (i.e., the compressibility effects) as the third term. Low aspect ratios and sweepback are factors that delay and reduce the compressibility drag rise.

3.6 Lift/Drag Ratio

The configuration and the application of an airplane is closely related to the lift/drag ratio. Many important items of airplane performance are obtained in flight at $(L/D)_{\max}$. Performance conditions that occur at $(L/D)_{\max}$ include the following: (1) maximum range of propeller-driven airplanes, (2) maximum climb angle for jet-powered airplanes, and (3) maximum power-off glide ratio (for jet-powered or propeller-driven airplanes).

Representative values of the maximum lift/drag ratios for subsonic flight speeds are given below:

Type of airplane	$(L/D)_{\max}$
High-performance sailplane	25–40
Commercial transport	12–20
Supersonic fighter	4–9

4. AERODYNAMIC MODELS BY SPEED RANGE

The discussion of aerodynamics is divided into speed ranges: subsonic aerodynamics, transonic aerodynamics, supersonic aerodynamics, and hypersonic aerodynamics. Assumptions unique to these speed ranges will be discussed.

4.1 Subsonic Aerodynamics

At relatively low angles of attack, i.e., −10° to +10°, the section lift coefficient is independent of the Reynolds number and is a linear function of the angle of attack. This occurs because the boundary layer is thin and attached over the entire length of the profile. This is illustrated in the data for a NACA 2412 airfoil section, which are presented in Fig. 25.

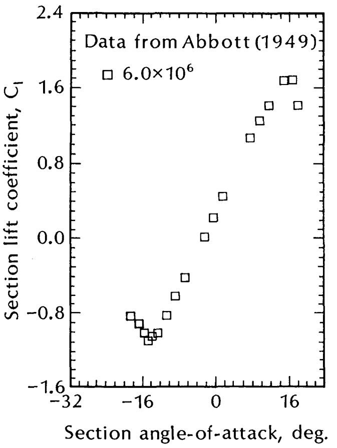

FIG. 25. Aerodynamic coefficients for a NACA 2412 airfoil section.

Note that, because this section is cambered, i.e., not symmetric about the chord line (or x axis), the angle of attack at which zero lift is generated is negative (specifically, −2°). As the angle of attack is increased, flow remains attached over the entire length of the profile [see Fig. 26(a)] and the lift is a linear function of the angle of attack.

In this linear range, the change in lift for a unit change in the angle of attack is known as the *lift-curve slope* and is an important design parameter. For inviscid flow past a two-dimensional airfoil, the lift-curve slope is 2π per radian. Viscous-induced changes to the pressure field (at low angles of attack, thickened boundary layers; at high angles of attack, boundary-layer separation) reduce the lift-curve slope. As will be discussed shortly, reducing the wing span, i.e., going from a two-dimensional (infinite-span) airfoil to a finite-span wing, further reduces the lift-curve slope.

As the angle of attack increases, the larger pressure gradients produce boundary-layer separation, starting at the trailing edge. (Note: Stall patterns are a function of the nose radius, section thickness, etc.) As a result, the lift is no longer a linear function of the angle of attack, as shown in Figs. 25 and 26(b). When the separation point reaches the midchord region, the maximum lift coefficient is achieved.

As the angle of attack is increased further, the separation point continues to move forward, as shown in Fig. 26(c). Lift actually decreases and the pressure drag increases rapidly with the angle of attack. One states that the airfoil is "stalled" when these conditions occur.

The lift is produced by pressure differences between the lower surface and the upper surface. For a wing of finite span, the high-pressure air beneath the wing spills out around the wing tips toward the low-pressure regions above the wing. As a consequence of the tendency of the pressures acting on the top surface near the tip of the wing to equalize with those on the bottom surface, the lift force per unit span decreases toward the tips. As another result of the spanwise pressure variation, the air on the upper surface flows inboard toward the root. Similarly, on the lower surface, air will tend to flow outward toward the tips. The resultant flow around a wing of finite span is three dimensional, having both chordwise and spanwise velocity components. Where the flows from the upper surface and the lower surface join at the trailing edge, the difference in spanwise velocity components will cause the air to roll up into a number of streamwise vortices, distributed along the span. These small vortices roll up into two large vortices just inboard of the wing tips. These features are illustrated in the sketches of Fig. 27. A photograph of smoke filaments illustrating the formation of the tip vortex as taken in the Aerodynamics Laboratory at the U.S. Air Force Academy is presented in Fig. 28.

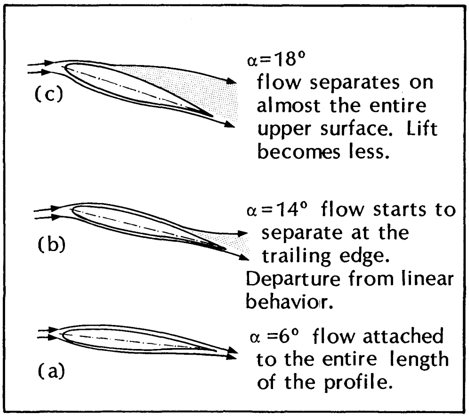
(c)
α=18°
flow separates on almost the entire upper surface. Lift becomes less.
(b)
α=14° flow starts to separate at the trailing edge. Departure from linear behavior.
(a)
α =6° flow attached to the entire length of the profile.

FIG. 26. Sketches of flow patterns as a function of the angle of attack.

The tip vortices trail back from the wing tips and have a tendency to sink and roll toward each other downstream of the wing. The

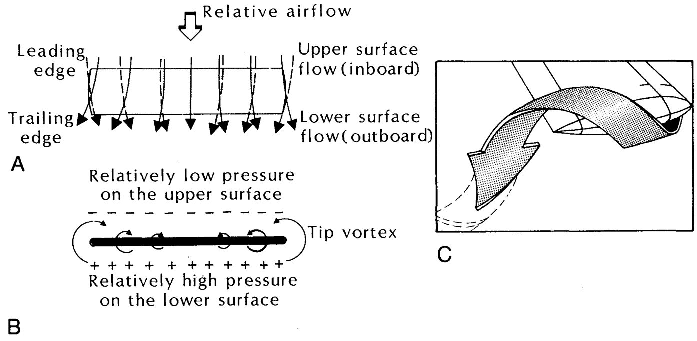
Relative airflow
Leading edge
Upper surface flow (inboard)
Trailing edge
Lower surface flow (outboard)
A
Relatively low pressure on the upper surface
Tip vortex
Relatively high pressure on the lower surface
B
C

FIG. 27. Generation of the trailing vortices due to the spanwise load distribution: (a) view from bottom; (b) view from trailing edge; (c) formation of the tip vortex.

air within the vortex system that is moving downward is called *downwash.* Eventually, the tip vortices weaken, their energy being dissipated by viscosity.

The problems of tip vortices are especially severe during the take-off and landings of the heavier jet airplanes. During these times, the speed of the airplane is low and the airplane is operating at high lift coefficients to maintain flight. The Federal Aviation Agency has shown that for a 0.27-MN (600 000 lb) plane, the tip vortices may extend back strongly for 8 km (5 miles) from the airplane and the downwash may approach 160 m min^{-1} (500 ft min^{-1}). Tests also show that a small light aircraft flying into such a vortex could be rolled over at rates exceeding 90° per second. These hazards are illustrated in the sketches of Fig. 29.

The vortex lines (or filaments) have an important role in the study of flow around wings. The lift produced by the pressure differences can be represented by a circulating (or vortex-like) flow. A computational technique for representing the lifting character of a wing may incorporate a lattice of vortices and is known as the vortex lattice method (VLM). The vortex theorems of Helmholtz provide a foundation for such techniques.

The tip vortices contribute to the downwash field at and downstream of the wing. To create downwash due to the finite-wing contribution requires the expenditure of energy per unit time (or power). The power required to induce this component of downwash is associated with an additional drag force, known as *induced drag.* This was discussed in Sec. 3.5.

The effect of finite wing span may also be viewed through changes in the flow field. The downwash has the effect of "tilting" the undisturbed air, so that the effective angle of attack is less than the geometric angle of attack for the airfoil section. As a result of the induced downwash velocity, the lift generated by a finite-span wing (C_L) having a particular airfoil section which is at the geometric angle of attack (α) is less than that for an infinite-span airfoil (C_l) having the same airfoil section and operating at the same angle of attack. This is illustrated in Fig. 30. Because the effective angle of attack is changed (in effect the flow is rotated), there is a component of the resultant force parallel to the actual free-stream flow which is, therefore, a drag force. This is the induced (or vortex) drag associated with the finite-span wing.

Recall that, according to Eq. (51), the induced drag varies inversely with the aspect ratio or, equivalently, as the $(\text{span})^{-2}$. Thus, consider two airplanes having rectangular wings as shown in the sketch of Fig. 31 [as taken from Talay (1975)]. Both wings are at the same lift coefficient and they have the same wing area. The only difference is that the wing span of the second wing is twice that of the first wing. Ideally, both wings should

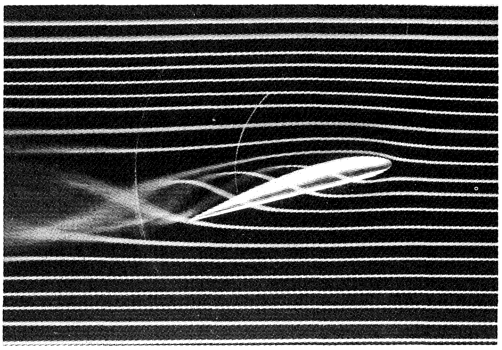

FIG. 28. Smoke filament patterns showing the formation of the tip vortex (taken at U.S. Air Force Academy).

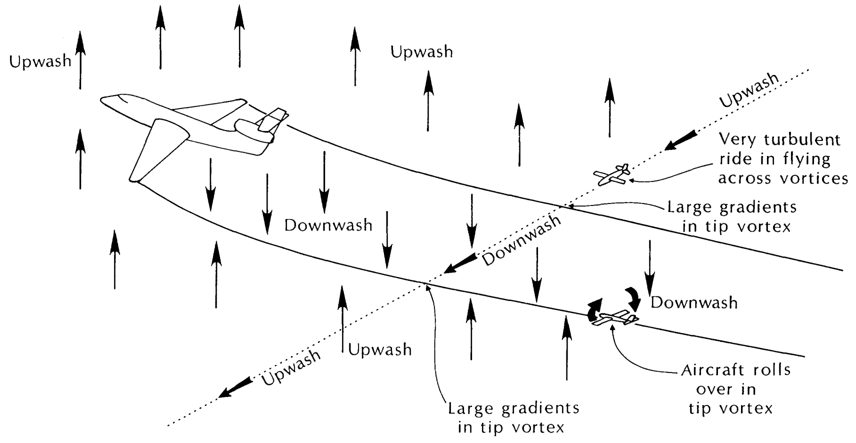
Upwash
Upwash
Upwash
Very turbulent ride in flying across vortices
Large gradients in tip vortex
Downwash
Downwash
Downwash
Downwash
Upwash
Upwash
Large gradients in tip vortex
Aircraft rolls over in tip vortex

FIG. 29. Upwash and downwash fields around an airplane.

produce the same lift. But the longer span wing (higher aspect ratio) has one-fourth the induced drag and, therefore, a greater efficiency. The thought arises that with less induced drag with longer wing spans, why not make a wing with an extremely long wing span (high aspect ratio). This would give nearly idealized flow, the tip vortex effects being very small. In fact, as the photographs of Fig. 32 illustrate, high-performance sailplanes that must rely on high efficiencies have large wing spans. As a result, sailplanes have aspect ratios in excess of 15, whereas single-engine light aircraft have wings with an aspect ratio of 7. Supersonic fighters typically have aspect ratios of 4, or less.

4.2 Transonic Aerodynamics

As the flight speed (or free-stream velocity) increases, a point is reached where the flow may no longer be considered incompressible. The pressure changes that occur as the air flows around the airfoil sections are sufficiently large that the density also varies. Even so, the flow is everywhere subsonic. Thus, the air accelerates from the (leading-edge) stagnation point, reaches a maximum velocity or, equivalently, the minimum pressure, then decelerates as it flows toward the trailing edge.

Eventually, the flight speed is sufficient for the Mach number to reach unity at some point. The free-stream Mach number for this condition is called the *critical Mach number.* Again, the air particles slow as they move toward the trailing edge, moving at subsonic speeds.

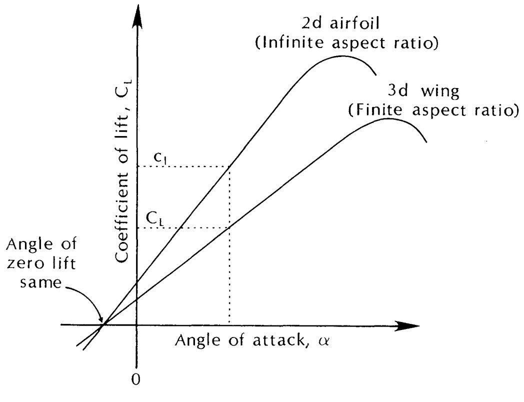

FIG. 30. Effect of aspect ratio on coefficient of lift.

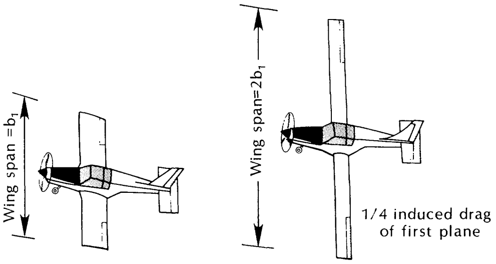
Wing span $=b_1$
Wing span$=2b_1$
1/4 induced drag of first plane

FIG. 31. Wing-span effect on induced drag for airplanes having the same wing area, same lift coefficient, and same dynamic pressure. [From Talay (1975).]

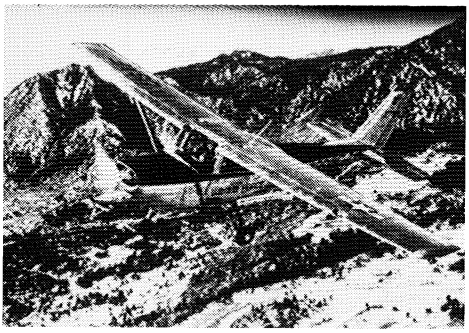

A Ordinary aspect ratio (T-41)

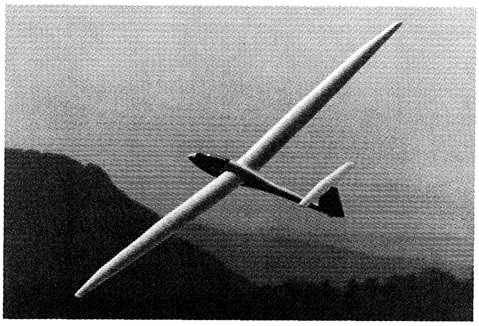

B High aspect ratio (High performance sailplane)

FIG. 32. Variations in aspect ratio.

However, once there are regions of locally supersonic flow, there exists the possibility that shock waves could occur. Three features which contribute to the location of a shock wave and, therefore, to the pressure distribution are the flow in the supersonic region ahead of the shock wave, the pressure rise across the shock wave itself (which involves considerations of the interaction with the boundary layer), and the subsonic flow downstream of the shock wave (which involves considerations of the boundary-layer development between the shock wave and the trailing edge, and of the flow in the near wake).

When the free-stream Mach number exceeds the critical Mach number, the flow is mixed subsonic/supersonic. Lift coefficient measurements [as taken from Farren (1956)] are presented as a function of the free-stream Mach number in Fig. 33. Line drawings made from schlieren photographs are presented in Fig. 34 to illustrate the following changes in the flow.

(a) When the free-stream Mach number is 0.75, the flow past the upper surface decelerates from local flow velocities which are supersonic without a shock wave. The lift coefficient is approximately 60% greater than the low-speed values at the same angles of attack.

(b) At $M_\infty = 0.81$, the lift coefficient reaches its maximum value, which is approximately twice the low-speed value. As indicated in Fig. 34(b), flow is supersonic over the first 70% of the surface, terminating in a shock wave. The flow on the lower surface is subsonic everywhere. Because the viscous flow separates at the foot of the shock wave, the wake is appreciably wider than for (a).

(c) At $M_\infty = 0.89$, flow is supersonic over nearly the entire lower surface and deceleration to subsonic speed occurs through a shock wave at the trailing edge. As a result, the lower surface pressures are lower at $M_\infty = 0.89$ than at $M_\infty = 0.81$. Flow on the upper surface is not greatly different than that for Fig. 34(b). As a result, the lift is drastically reduced. Separation at the foot of the upper surface shock wave is more conspicuous and the turbulent wake is wide. The shock wave at the trailing edge of the lower surface effectively isolates the upper surface from the lower surface. As a result, the pressure on the upper surface near the trailing edge is greater than that on the lower surface.

(d) When the free-stream Mach number is 0.98, the shock waves both for the upper

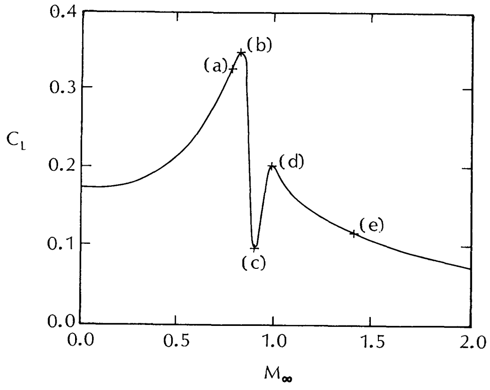

FIG. 33. Lift coefficient as a function of Mach number to illustrate the effect of compressibility. [From Farren (1956).] Refer to Fig. 34 for the flow fields corresponding to the lettered points on this graph.

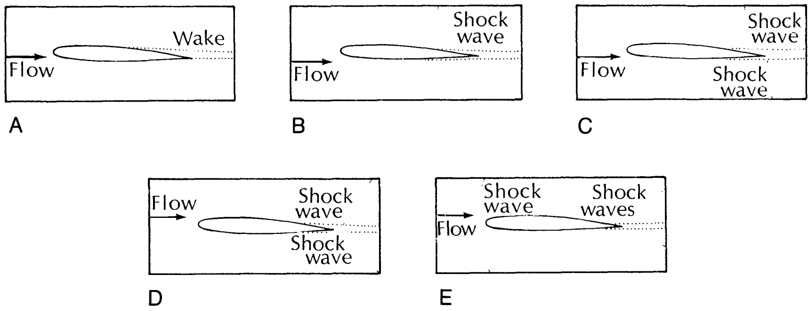

FIG. 34. Flow field around an airfoil in transonic streams based on schlieren photographs: (a) Mach number $M_\infty = 0.75$; (b) Mach number $M_\infty = 0.81$; (c) Mach number $M_\infty = 0.89$; (d) Mach number $M_\infty = 0.98$; (e) Mach number $M_\infty = 1.4$. [From Farren (1956).]

surface and for the lower surface have reached the trailing edge. The local Mach number is supersonic for most of the airfoil (both upper and lower surfaces).

(e) When the free-stream flow is supersonic, a bow shock wave (i.e., the detached shock wave in front of the leading edge) is generated. The flow around the airfoil is supersonic everywhere except very near the rounded nose. The shock waves at the trailing edge remain, but they have become weaker.

The data presented in Figs. 33 and 34 illustrate the effect of Mach number for a given airfoil section at a particular angle of attack. Parameters such as thickness ratio, camber, and nose radius also influence the magnitude of the compressibility effects. Transonic flows are very sensitive to the contour of the body surface, since changes in the surface slope affect the location of the shock wave and, therefore, the inviscid flow field as well as the downstream boundary layer.

Furthermore, the shock-wave–boundary-layer interaction and the possible development of separation downstream of the shock wave are sensitive to the character of the boundary layer, to its thickness, and to its velocity profile at the interaction location. Since a turbulent boundary layer can negotiate higher adverse pressure gradients than can a laminar one, the shock-wave–boundary-layer interaction is smaller for a turbulent boundary layer.

As the wing goes through the transonic speed regime, there is a rapid rise in the drag coefficient, as shown in Fig. 35. The Mach number corresponding to the drag increase is the *drag divergence Mach number* (M_{DD}). As noted in Sec. 3.5, the compressibility-related effects that are correlated in terms of the Mach number are known as the *wave drag*. Furthermore, there is a significant shift in the center of pressure which creates stability and control problems. Since these problems are associated with the formation of shock waves, the designer seeks configuration changes which eliminate the shock waves, move the shock waves to points where the Mach number is near unity (so that the shock wave is weaker), or reduce the pressure gradients which influence the shock strength. Two long-recognized solutions are thinner airfoil sections and sweeping the wing.

In the late 1930s, two aerodynamicists who had been taught by Prandtl, Adolf Busemann and Albert Betz, discovered that drag at transonic and supersonic speeds could be reduced by sweeping the wings. Betz, in 1939, was the first person to draw attention to the significant reduction in transonic drag which comes when the wing is swept back enough to avoid the formation of shock waves which occur when the flow over the wing is locally supersonic. The basic principle is that the component of the main flow parallel to the wing leading edge is not perturbed by the wing, so that the critical conditions are reached only when the component of the flow normal to the leading edge has been accelerated to the local sonic velocity at some point on the wing. This model is in fact a simplification, since the flow over a typical swept wing of relatively low aspect ratio is a complex, three-dimensional flow.

The effects of sweeping the wing on C_{D_0} and $C_{L_{\max}}$ are illustrated in Fig 36. As illustrated in these sketches, sweeping the wing

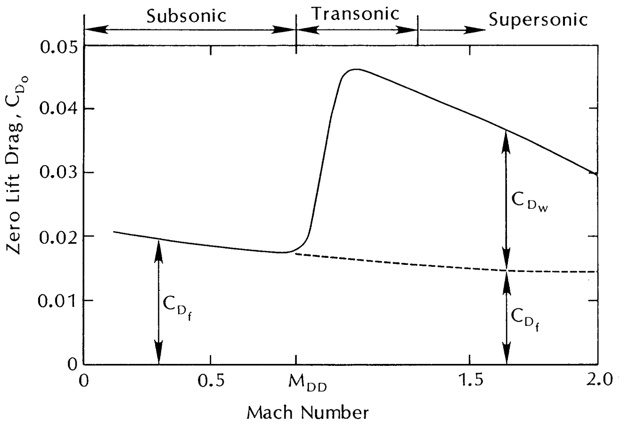

FIG. 35. Typical development of zero lift drag above the drag divergence Mach number.

not only delays the subsonic drag rise to a higher speed but reduces the peak drag coefficients. Benefits also occur at supersonic speeds for highly swept wings.

One can use a variable-geometry (or swing-wing) design to obtain a suitable combination of low-speed and high-speed characteristics. For the unswept, high-aspect-ratio configuration, one obtains efficient subsonic cruise and good maneuverability at the lower speeds needed for approaches and landings. For the highly swept, low-aspect-ratio configuration, the variable geometry wing provides low wave drag and eliminates the need for wing fold in the case of naval aircraft. Negative factors in a swing-wing design are complexity, a loss of internal fuel capacity, and the considerable weight of the hinge-pivot structure.

Where the fuselage and wing are joined together, the air flows around the two surfaces mutually interfere. This interference is especially important at high speeds because of the occurrence of shock waves, which give rise to wave drag. The interference effects total as much as a quarter of the sum of the individual drag contributions of the configuration's components.

Thus, the interaction between the central fuselage and the swept wing has a significant effect on the transonic flow over the wing. In the early 1950s, Küchemann (1957) recognized that a properly shaped waist-like indentation on the fuselage could be designed to produce a velocity distribution on the wing near the root similar to that on the corresponding infinite-span wing. Küchemann noted that "the critical Mach number can be raised and the drag reduced in the transonic and the supersonic type of flow, the wing behaving in some respects like a thinner wing." Whitcomb (1954) found that zero-lift drag rise is due primarily to shock waves. Furthermore, the shock-wave formations about relatively complex swept-wing–body combinations at zero lift near the speed of sound are similar to those which occur for a body of revolution with the same axial development of cross-sectional area normal to the airstream. Whitcomb concluded that "near the speed of sound, the zero-lift drag rise of a low-aspect-ratio thin-wing–body combination is primarily dependent on the axial development of the cross-sectional areas normal to the air stream." Therefore, the drag-rise increments near the speed of sound are less for fuselage-wing configurations which have a more gradual change in the cross-sectional

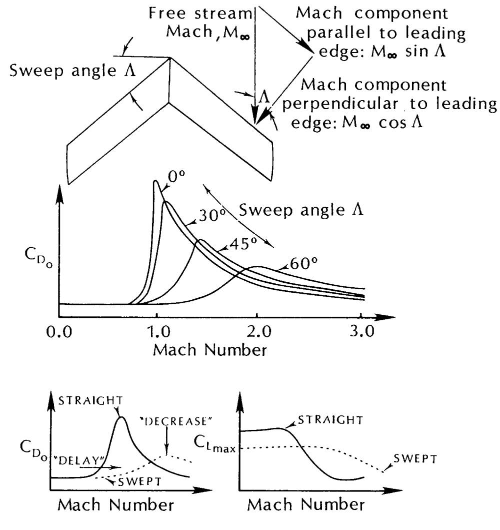
Free stream Mach, M_∞
Mach component parallel to leading edge: $M_\infty \sin\Lambda$
Sweep angle Λ
Λ
Mach component perpendicular to leading edge: $M_\infty \cos\Lambda$
0°
30°
Sweep angle Λ
45°
60°
C_{D_o}
0.0
1.0
2.0
3.0
Mach Number
STRAIGHT
"DECREASE"
$C_{L_{max}}$
C_{D_o}
"DELAY"
SWEPT
Mach Number
STRAIGHT
SWEPT
Mach Number

FIG. 36. The effects of sweeping a wing back.

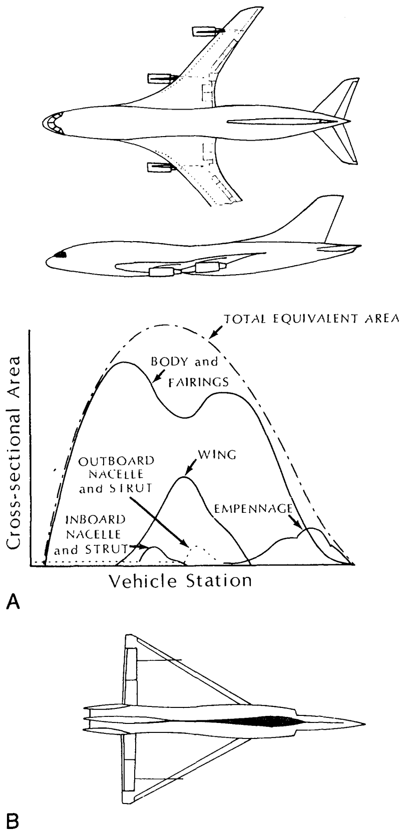
TOTAL EQUIVALENT AREA
BODY and FAIRINGS
Cross-sectional Area
OUTBOARD NACELLE and STRUT
WING
INBOARD NACELLE and STRUT
EMPENNAGE
Vehicle Station
A
B

FIG. 37. Application of the area rule to the axial distribution of the cross-sectional area: (a) area rule applied to the design of a near-sonic transport; (b) sketch of the Convair F 102 A.

area (including the fuselage and the wing) with axial position (as well as a reduction in the relative magnitude of the maximum area). Whitcomb noted that it would be expected that indention of the body of a wing-body combination, so that the combination has nearly the same axial distribution of cross-sectional area as the original body alone, would result in a large reduction in the transonic drag rise. Design applications of this "theory," which is often known as *Whitcomb's area rule* for transonic configurations, are illustrated in Fig. 37. Early flight tests revealed that the prototypes of the YF-102 had serious deficiencies in performance. A major redesign effort included the first application of Whitcomb's area rule to reduce the transonic drag. The modified YF-102A achieved supersonic flight satisfactorily, providing the U.S. Air Force with its first operational delta-wing design.

4.3 Supersonic Aerodynamics

It should be clear from the previous discussion of shock waves and transonic flows that supersonic flows are very different from subsonic flows. Consider the flat plate traveling at twice the speed of sound at zero angle of attack, i.e., $U=2a$, as shown in Fig. 38(a). A piece of tape, which is placed on the plate, generates an infinitesimal disturbance, which propagates radially at the speed of sound, i.e., at a.

After the passage of time Δt, the disturbance front has moved $r=a\Delta t$, and the leading edge of the tape has moved $s=2a\Delta t$. The disturbances generated by the tape as it moves from its position in Fig. 38(a) to its position in Fig. 38(b) are contained within the cone bounded by the Mach angle μ. The boundary of the cone is known as the Mach wave. Thus, the perturbed flow is limited to points within the "Mach cone." This region is known as the *zone of action.* Points ahead of the Mach wave have no knowledge of the perturbations produced by the tape. This region is known as the *zone of silence.*

The geometry is such that

$$\sin\mu=\frac{a\Delta t}{2a\Delta t}=\frac{1}{2};$$

thus, $\mu=30°$. For a general case

$$\mu=\sin^{-1}\frac{a}{U}=\sin^{-1}\frac{1}{M}. \tag{54}$$

As the Mach number goes to 1.0, the Mach angle approaches 90°. Since the disturbance is infinitesimal, flow properties do not change

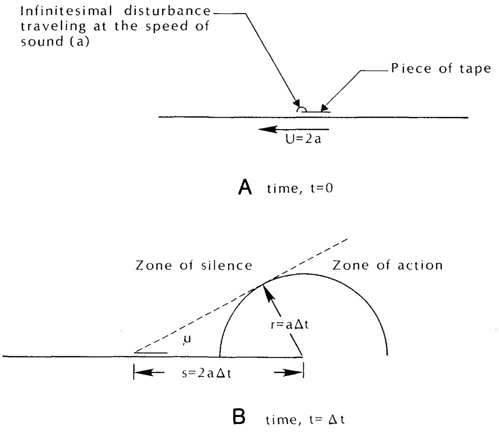
Infinitesimal disturbance traveling at the speed of sound (a)
Piece of tape
U=2a
A time, t=0
Zone of silence
Zone of action
r=aΔt
μ
s=2aΔt
B time, t=Δt

FIG. 38. Sketch illustrating the Mach angle in supersonic flow.

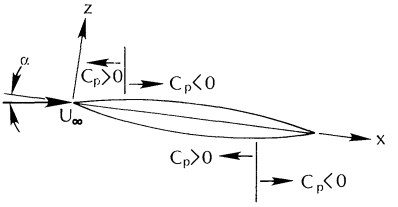

FIG. 39. General features for linearized supersonic flow past a thin airfoil.

across a Mach wave. If the disturbance increases in strength, i.e., is finite, it creates a shock wave. As shown in Fig. 15, shock waves make an angle θ with the free stream which is greater than μ. Of course, flow properties change across the shock wave, as discussed in Sec. 2.3.

4.3.1 Wave Drag at Supersonic Speeds. Experience has shown that the leading edge and the trailing edge of supersonic airfoils should be sharp and the section relatively thin. If the leading edge is not sharp (or only slightly rounded), the leading-edge shock wave will be detached and relatively strong, causing relatively large wave drag. Thus, configurations designed to operate at supersonic speeds are relatively long for a given cross section so that the flow deflections are small.

For small deflections of a supersonic flow, the pressure coefficient is proportional to the angle through which the flow is turned relative to the free-stream direction:

$$C_p=2\theta/(M_\infty^2-1)^{1/2}. \tag{55}$$

As shown in Fig. 39, a positive pressure coefficient is associated with a compressive change in the flow direction. Furthermore, for regions where C_p is positive, the local velocity is below the free-stream value. If the flow is turned toward the upstream Mach waves, it is a compressive turning of the flow. If the flow turns away from the upstream Mach waves, it is an expansive turning of the flow, i.e., an acceleration. Thus, for the double-convex-arc airfoil section shown in Fig. 39, the pressure is greatest at the leading edge, being greater on the lower surface when the airfoil is at a positive angle of attack. Note that the local flow accelerates continuously from the leading edge to the trailing edge both for the lower surface and for the upper surface. The pressure coefficient is zero, i.e., the local static pressure is equal to the free-stream value, at those points where the surface is parallel to the free stream. Downstream, the pressure coefficient is negative, which corresponds to an expansive change in direction.

The pressure distribution described in the preceding paragraph for a thin airfoil in a supersonic flow produces a drag force on a two-dimensional airfoil in an inviscid stream. Recall that, for incompressible, inviscid flow past a two-dimensional airfoil, the drag is zero. Typical pressure distributions for an airfoil in an incompressible stream and in a supersonic stream are presented in Fig. 40. Examining the sketch for the supersonic case, it can be seen that both the positive pressure coefficients on the forward portion of the airfoil and the negative pressure coefficients on the aft portion of the airfoil contribute to the drag. This pressure drag is known as *wave drag.*

The wave drag can be minimized by selecting the optimum area, or volume distribution, along the length of the body. Because of the complexity of the flow, there is no single equivalent body, as was the case for the *sonic area rule.*

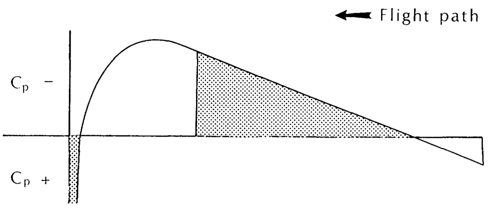

A Subsonic flow

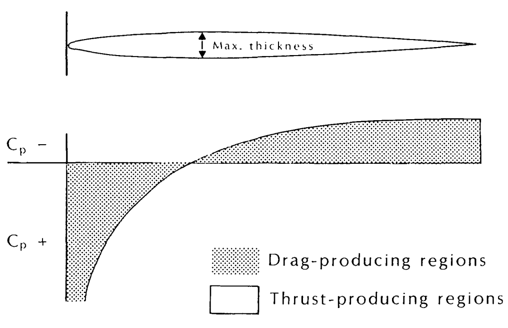

B Supersonic flow

FIG. 40. Typical pressure distributions for flow past a two-dimensional airfoil. [From Stinton (1966).]

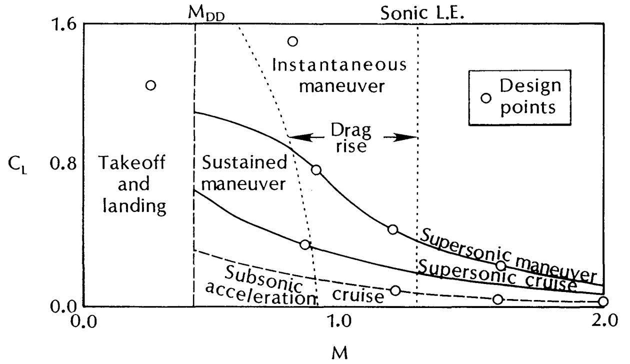

FIG. 41. Typical performance map for fighters. [From Bradley (1981).]

4.3.2 Design Considerations for Tactical Military Aircraft. As noted by Bradley (1981), the designer of a tactical military aircraft is faced with a multitude of design points throughout the subsonic-supersonic flow regimes plus many off-design constraints. The design goals for a tactical weapons system may include efficient cruise at both subsonic and supersonic Mach number, superior maneuverability at both subsonic and supersonic Mach number, and rapid acceleration.

Bradley states:

> The multiple design point requirement turns out to be the major driver for the designer of fighter aircraft. The aerodynamic requirements for each of the design points often present conflicting requirements. For example, the need for rapid acceleration to supersonic flight and efficient supersonic cruise calls for thin wing sections with relatively high sweep and with camber that is designed to trim out the moments resulting from aft ac movement at supersonic flight. However, these requirements are contrary to those requirements for efficient transonic maneuver, where the designer would prefer to have thicker wing sections designed with camber for high C_L operation and a high-aspect-ratio planform to provide a good transonic drag polar. Designers are thus faced with a situation of compromise. These conflicting requirements suggest the obvious solution of variable geometry, that is, variable sweep wing and/or variable camber. Although this is a satisfactory aerodynamic solution, in many cases the resultant weight increases to a configuration can be prohibitive. [In this paragraph "ac" refers to the aerodynamic center, which is defined in Sec. 3.4.]

A typical performance spectrum corresponding to mission requirements is illustrated in Fig. 41, which presents a map of lift coefficient versus Mach number. The low Mach end of the spectrum throughout the C_L range is typical of takeoff and landing for the configuration. The subsonic cruise and supersonic cruise portions are noted in the moderate lift range. Acceleration to high supersonic speeds occurs at the low lift coefficients. Sustained maneuver takes place in the C_L range of less than one for most fighter configurations. Above this lift coefficient, the aircraft is in the instantaneous maneuver regime. Drag rise occurs, depending on the wing geometry, in the range of 0.8 to 1.2 Mach.

The particular flow conditions that correspond to the flow map are shown in Fig. 42. At the cruise and acceleration points, the aircraft designer is dealing primarily with attached flow, and his design objective is to maintain attached flow for maximum efficiency. At the higher C_L's, corresponding to instantaneous maneuver, separated flow becomes the dominant feature. Current designs take advantage of the separated flow by forming vortex flows in this range. Intermediate C_L's corresponding to sustained maneuver are usually a mixture of separated and attached flow. Consequently, if the aircraft is designed with camber to minimize separation in the maneuver regime, the configuration will have camber drag and may have

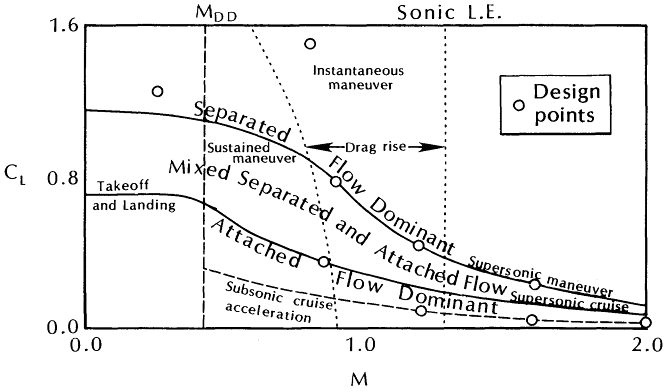

FIG. 42. General flow regimes encountered for fighters. [From Bradley (1981).]

lower surface separation, which increases drag at the low C_L's needed for acceleration.

Thus, the aircraft design is a compromise to achieve an optimal flow efficiency considering the numerous design points associated with the mission objectives. It is easily seen that the transonic flow problems that must be addressed in fighter design are driven to a very large extent by the constraints imposed at the other design points—supersonic and subsonic.

Bradley (1981) notes that "Supersonic design enjoys relatively precise computation and optimization thanks to wide applicability of linearized theory." Further, he noted:

> The design of efficient transonic configurations may proceed from two conceptual schools of logic. One acknowledges that the optimum low-drag flow must accelerate rapidly over the airfoil to supercritical flow and decelerate in a nearly isentropic manner, avoiding strong shocks and/or steep gradients that can lead to significant regions of separation. This approach sets attached flow or near fully attached flow as an intuitive design goal, and is typically used for aircraft that permit strong emphasis to be placed on transonic cruise or sustained maneuver design points.
>
> The second school of thought recognizes the inevitability of significant flow separations at design conditions and adopts a philosophy of controlling certain regions of separation through vortex flows to complement other regions of attached supercritical flow. This approach is appropriate for configurations constrained by multiple design points that emphasize added supersonic requirements. Current tactical fighters that rely on high wing loadings for transonic performance are good examples. The F-16 and F-18 employ a combination of controlled vortex flow and variable camber to achieve maneuverability.

Even though a designer has worked carefully to design a wing for attached flow, the flow fields associated with sustained and instantaneous maneuvers involve predominantly separated flows. The manner in which the flow separates, and how the separation develops over the configuration, will strongly affect the vehicle's drag and its controllability at the higher C_L's. Aircraft such as the F-16 and the F-18 employ strakes or leading-edge extension devices to provide a controlled separated flow. Controlled vortex flow can then be integrated with variable-camber devices on the wing surface to provide satisfactory high-lift, stability and control, and buffet characteristics. The resulting flow field is a complex one, combining attached flows over portions of the wing with the vortex flow from the strake.

4.4 Hypersonic Flow

4.4.1 General Characterization of Hypersonic Flows. The problems associated with determining the aerothermodynamic environment of a vehicle flying through the atmosphere offer diverse challenges to the designer. The characteristics of the hypersonic environment depend on the mission requirements and vehicle size constraints. Represen-

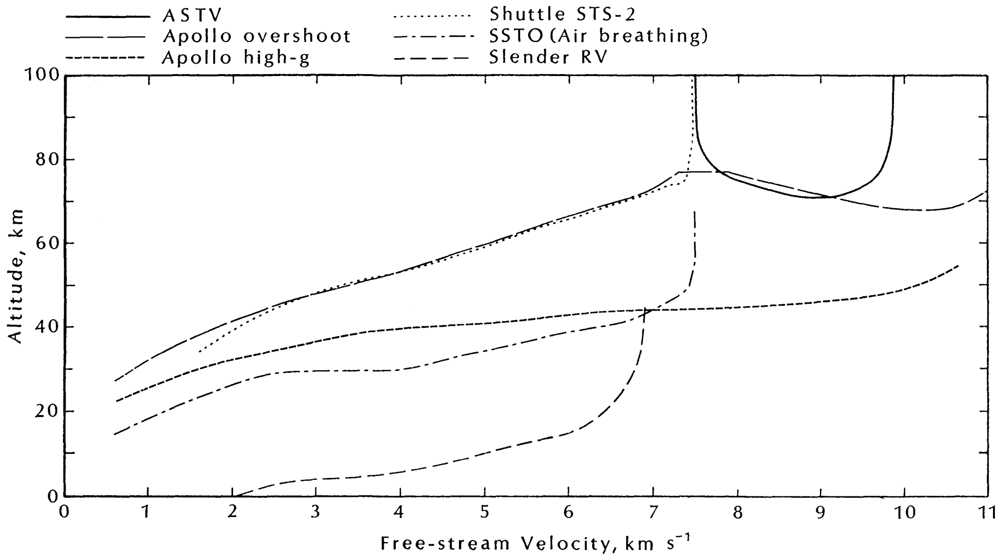

FIG. 43. Velocity-altitude parameters for several reentry vehicles. [From Bertin (1989).]

tative reentry trajectories for several different programs are presented in Fig. 43. Presented in Fig. 43 are the approximate velocity-altitude parameters for a trajectory for the Aeroassisted Space Transfer Vehicle (ASTV), for two trajectories for the Apollo Command Module (an overshoot trajectory and a 20-g limit trajectory), for the best-estimated trajectory for the STS-2 (Shuttle Orbiter) reentry, for the trajectory of a single-stage-to-orbit vehicle with an air breathing engine (SSTO), and for the trajectory for a slender (relatively low-drag) reentry vehicle. For some applications, the vehicle will be very blunt or fly at very high angles of attack, so that the drag coefficient is large and the hypersonic deceleration occurs at very high altitudes. Reentry vehicles from the U.S. manned spacecraft program, e.g., the Apollo and the Shuttle, are characteristic of these applications. The designer of such vehicles may have to account for nonequilibrium thermochemistry, viscous-inviscid interactions, and (possibly) noncontinuum flow models. Applications that require the time of flight to be minimal produce slender reentry vehicles that penetrate deep into the atmosphere before experiencing significant deceleration. The trajectory for the Slender RV, which is presented in Fig. 43, illustrates the velocity-altitude environment for a low-drag configuration. As a result, these relatively high ballistic coefficient β (i.e., W/C_DA) configurations experience severe heating rates and high dynamic pressures, but only for a short period of time.

For a free-stream velocity of 7 km s^{-1} (23 kft s^{-1}), the kinetic energy per unit mass, i.e., $U_\infty^2/2$, is 2.45×10^7 J kg^{-1} (1.05×10^4 Btu lb$_m^{-1}$). As the air particles pass through the shock wave ahead of the vehicle, they slow down, converting kinetic energy into internal energy. For a perfect gas, one for which the composition is frozen and one for which γ is constant, the increase in internal energy is in the form of random motion of the air particles (temperature). Thus, using a perfect-gas model for the air crossing a normal shock wave, the temperature would increase by approximately 2.4×10^4 K (4.4×10^4 °R). In reality, the change in kinetic energy of the air particles would go into increasing the random thermal motion of the air particles (temperature), exciting the vibrational (internal) energy of the diatomic particles, and dissociating the molecules (producing atomic species). If the density of the air is high enough, there are sufficient collisions between the air particles so that the air downstream of the shock wave quickly achieves the equilibrium state. For an equilibrium flow, the temperature of the air downstream of a normal shock wave is of the order 7×10^3 K (13×10^3 °R) for this case.

For most cases, the thermodynamic state will be between these limits of frozen flow and of equilibrium flow. Where the actual flow falls between these limits is a function of the density (or, equivalently, the altitude). At higher altitudes, there are fewer air particles and, therefore, fewer collisions between them. Thus, the air particles must travel further to experience the collisions required to excite the vibrational modes and to accomplish the dissociation process.

4.4.2 Defining Hypersonic Flow. Flow fields around blunt bodies begin to exhibit many of the characteristics of hypersonic flows when the Mach number is 4, or greater. By definition,

$$M_\infty = U_\infty / a_\infty \gg 1 \tag{56}$$

is the basic assumption for all hypersonic flow theories. Thus, the internal thermodynamic energy of the free-stream fluid particles is small compared with the kinetic energy of the free stream for hypersonic flows. In flight applications, this results because the velocity of the fluid particles is relatively large. The limiting case, where M_∞ approaches infinity because the free-stream velocity approaches infinity while the free-stream thermodynamic state remains fixed, produces extremely high temperatures in the shock layer. The high temperatures associated with hypersonic flight cannot be accommodated in ground-test facilities. Therefore, in wind-tunnel applications, hypersonic Mach numbers are achieved through relatively low speeds of sound, i.e., for the wind tunnel, M_∞ approaches infinity because the speed of sound (which is proportional to the square root of the free-stream temperature) goes to "zero" while the free-stream velocity is held fixed. As a result, the fluid temperatures remain below the levels that would damage the wind tunnel or the model.

Another assumption common to hypersonic flow is that

$$\epsilon = \rho_\infty / \rho_2 \ll 1, \tag{57}$$

which is known as the small-density-ratio assumption. Thus, this assumption relates primarily to the properties of the gas downstream of the shock wave. Recall that for a perfect gas [Ames Research Staff (1953)]:

$$\epsilon = \frac{\gamma - 1}{\gamma + 1} \tag{58}$$

for a normal shock wave as $M_\infty \to \infty$. Thus, ϵ is 1/6 for perfect air. Note that typical hypersonic wind tunnels operate at conditions where the test gas can be approximated by the perfect-gas relations. However, during the reentry flights of the Apollo Command Module, the density ratio approached 1/20. Thus, the density-ratio simulation in those wind tunnels where the air flow behaves as a perfect gas does not match the flight values. This will have a significant effect on the shock stand-off distance, as well as other parameters.

As the free-stream Mach number approaches infinity, the shock layer becomes very thin. As a result, one can assume that the speed and direction of the gas particles in the free stream remain unchanged until they strike the solid surface exposed to the flow. For this flow model (termed *Newtonian flow*, since it is similar in character to one described by Newton in the 17th century), the normal component of momentum of the impinging fluid particle is wiped out, while the tangential component of momentum is conserved.

The Newtonian flow model and the various theories for thin shock layers related to the Newtonian approximation are based on the small-density-ratio assumption. The small-density-ratio requirement for Newtonian theory also places implicit restrictions on the body shape in order that the shock layer be thin.

For slender configurations, such as sharp cones and wedges, the strong shock assumption is

$$M_\infty \sin\theta_b \gg 1, \tag{59}$$

where θ_b is the body deflection angle. Equation (59) is mixed in nature, since it relates both to the flow and to the configuration. The concept termed the *Mach number independence principle* depends on this assumption.

The Mach number independence principle was derived for inviscid flow by Oswatitsch. Since pressure forces are much larger than the viscous forces for blunt bodies or for slender bodies at relatively large angles of attack when the Reynolds number exceeds 10^5, one would expect the Mach number independence principle to hold at these conditions. This is demonstrated by the data presented in

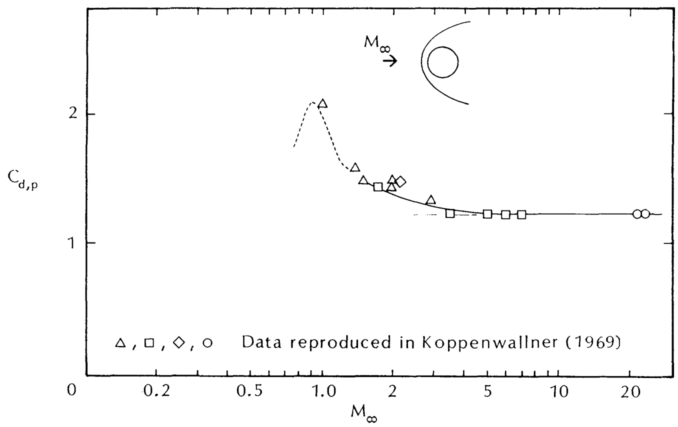

FIG. 44. Pressure drag of a right circular cylinder as a function of Mach number. [From Koppenwallner (1969).]

Fig. 44, which are taken from Koppenwallner (1969). Note that the pressure drag for the (blunt) right circular cylinder reaches its limiting value by $M_\infty = 4$.

4.4.3 Chemistry of Hypersonic Flows. It is well known that the air particles in the shock layer of a hypervelocity reentry vehicle undergo vibrational excitation, dissociation, and (possibly) ionization. These chemical phenomena absorb energy, limiting the temperature increase as the kinetic energy of the hypervelocity air particles crossing the bow shock wave is converted to thermal energy. At low altitudes, where the free-stream density is sufficiently high, these chemical phenomena tend to reach equilibrium, as shown in Fig. 45. At higher altitudes, where the free-stream density is relatively low, there are not sufficient collisions for the gas to reach an equilibrium state. For these nonequilibrium flows, the concentrations of individual species and the energy contained in the different internal modes must be calculated by integrating the conservation equations governing the phenomena.

For hypersonic flight, the gas molecules that pass through the bow shock wave are excited to higher vibrational and chemical energy modes. This lowers the specific-heat ratio of the gas below the free-stream value, if it is assumed that equilibrium exists and that dissociation is not driven to completion. A

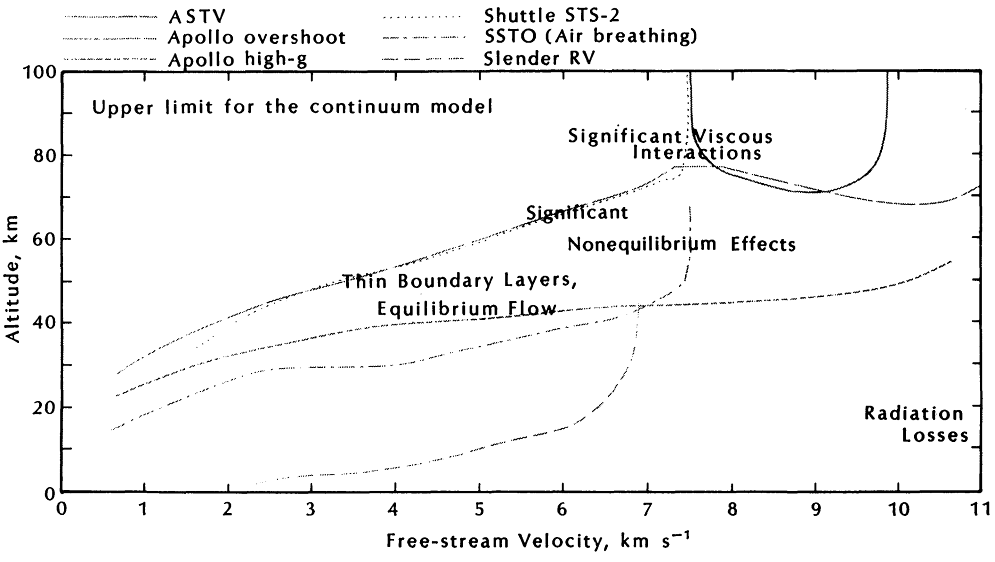

FIG. 45. Modeling flow phenomena for hypersonic flight.

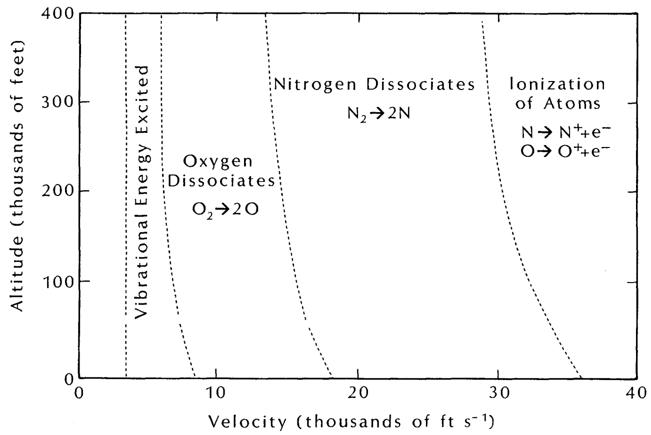

FIG. 46. Composition of equilibrium air in the stagnation region [Hansen and Heims (1958)].

large amount of the energy that would have gone into increasing the static temperature behind the bow shock wave for a perfect gas is used instead to excite the vibrational energy levels or to dissociate the gas molecules. The dominant chemical reactions that occur in the stagnation region for an equilibrium air flow [as presented in Hansen and Heims (1958)] are presented in Fig. 46. At about 914 m s^{-1} (3000 ft s^{-1}), the vibrational energy of the air molecules begins to become important. Oxygen dissociation begins when the free-stream velocity is between 1829 m s^{-1} (6000 ft s^{-1}) and 2438 m s^{-1} (8000 ft s^{-1}). Nitrogen dissociation occurs at velocities in excess of 4572 m s^{-1} (15 000 ft s^{-1}). Finally, ionization of atoms becomes important at velocities in excess of 9144 m s^{-1} (30 000 ft s^{-1}). Hansen and Heims (1958) note that

> the dissociation and ionization reactions are pressure dependent because each particle yields two product particles, and such reactions are inhibited by high pressure. Therefore, higher temperature and, consequently, higher velocity are required to produce the reactions at sea level than at high altitudes where much lower pressures occur.

5. ROTORCRAFT AERODYNAMICS

A complete dynamic analysis of the flow field for rotorcraft rotor blades is more complex than that for fixed wings. This is because the blades provide not only lift, but also propulsion and control. This complexity is accentuated by the fact that the long, thin rotor blades are very flexible and are subjected to a severely fluctuating aerodynamic environment. As a result, the structural dynamics and aerodynamics of rotor blades are tightly coupled and difficult to analyze.

A partial control loop for the dynamic analysis of rotor blades is presented in Fig. 47. The pilot controls the angle of attack of the blades via the "cyclic" (controls blade path angle) and the "collective" (controls blade pitch). Then the aerodynamic forces on the rotor blades create a flow field which can bend and twist the rotor blades, changing their angle of attack. So, the rotor blades' structural dynamics obviously feed back into the control of the blades. The bending and twisting of the rotor blades also affect the loads on the rotor "hub" which is connected to the power plant. This results in a torque which affects the power needed from the drive train and creates moments and shear

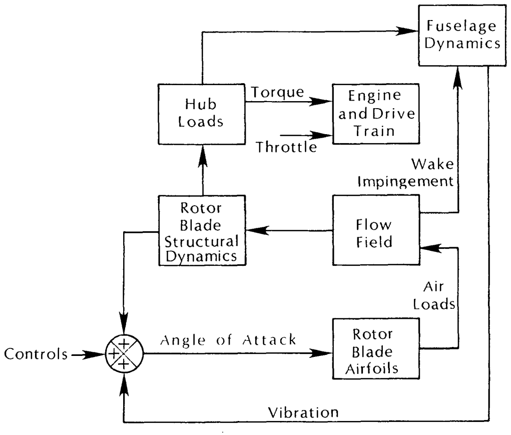

FIG. 47. Rotor-blade dynamic analysis.

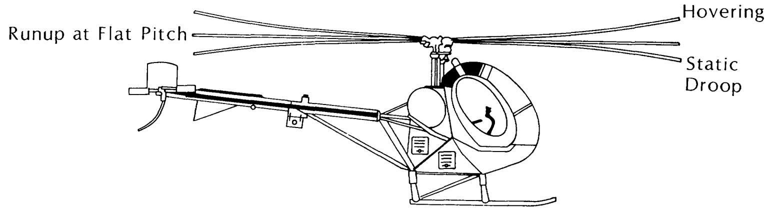

FIG. 48. Helicopter coning angles.

stresses transmitted to the fuselage which can also feed back into the control loop via vibration. The fuselage dynamics can also be affected directly by wake impingement from the rotor blade's flow field, resulting in a negative feedback loop.

This is only a basic view of the complexity of rotor-blade aero- and structural dynamics. This section will not attempt to address this very difficult problem fully, but will point out the most significant differences between the analysis of rotor blades and fixed wings.

All of the analysis of flow fields in Sec. 4.1 on subsonic flight also applies to the rotor blades of rotorcraft. The blades are, in fact, merely wings that rotate as well as translate through the air. This section will concentrate on some of the unique problems encountered in analyzing the aerodynamics of rotor blades by looking at three important phases of flight: hovering, forward motion, and descending flight.

5.1 Hovering Flight

An analysis of hovering flight in calm air is relatively straightforward since there is no translational motion and the blades only have to provide lift. Such an analysis does, however, point out the fundamental difference between a rotor blade and a fixed wing: each section of a rotor blade experiences a significantly different airspeed than the other sections along the span of the rotor. The "equivalent" free-stream velocity for each section of the rotor blade is related to the rotational speed of the rotor by the simple equation

$$v_r = r\omega. \tag{60}$$

Using RPM (revolutions per minute), the airspeed at any point along the rotor can be calculated by

$$v_r = \frac{\mathrm{RPM}\cdot r}{9.55} \quad (\mathrm{m\ s^{-1}}). \tag{61}$$

Since lift depends on the square of velocity (all other factors held constant), the tips of the rotor blades produce a disproportionate amount of lift, which is aerodynamically inefficient. The blade tips not only experience the largest lift force, but are also furthest away from the center hub. This results in a large torque applied to the tips which causes large "coning" angles about the hub, as shown in Fig. 48, which can cause structural problems.

To minimize the aerodynamic and structural problems associated with the differential airspeeds along the rotor blades, the shape of the blades is altered along the span. Typically, the blade pitch angle (the angle of the airfoil section's chordline relative to the plane of rotation) is 8°–10° smaller at the blade tips than near the rotor hub. This minimizes the lift differential between the rotor blade's center and tip, as shown in Fig. 49.

The total lift produced by the rotor blades during hover is more difficult to calculate than the lift produced by a fixed wing, developed in Sec. 3.3, and is given by

$$L = \tfrac{1}{2}\,\rho v^2 S C_L. \tag{62}$$

For the reasons discussed earlier, each section of the rotor blade will have a different velocity, a different blade path angle (affect-

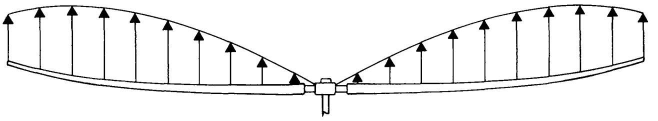

FIG. 49. Typical spanwise lift distribution for helicopter rotor blades.

ing its C_{l_i}), and a different cross-sectional area (for structural reasons). Therefore, the lift must be calculated by summing the contributions of each blade element:

$$\sum_i \Delta L = \sum_i \tfrac{1}{2} \rho v_{r_i}^2 S_i C_{l_i} = L_r, \qquad (63)$$

where L_r is the total lift due to rotation. Integration of Eq. (63) yields

$$L_r = \frac{1}{2} \rho \frac{v_{\text{tip}}^2}{3} S C_{L_{\text{ave}}}, \qquad (64)$$

where S is the total rotor blade area, and $C_{L_{\text{ave}}}$ is the average lift coefficient.

It is convenient, and customary, to replace several of the terms in Eq. (64) with terms more useful to the aerodynamicist. The solidity ratio represents the portion of the circular disk actually occupied by the blades, typically 4%–7%:

$$S^* = S/\pi R^2 \quad \text{(nondimensional)}. \qquad (65)$$

The tip velocity can be simplified using Eq. (61):

$$v_{\text{tip}} = \frac{\text{RPM} \cdot R}{9.55} \quad (\text{m s}^{-1}). \qquad (66)$$

The density ρ is usually replaced by a density ratio:

$$\rho^* = \rho/\rho_0 \quad \text{(nondimensional)}, \qquad (67)$$

where ρ_0 is the density at sea level.

Making all the substitutions into Eq. (64) and inserting the appropriate units, the total lift (in newtons) produced by the rotor blades during hover is

$$L_r = 1740 S^* \rho^* C_{L_{\text{ave}}} (\text{RPM})^2 R^4. \qquad (68)$$

This equation shows the linear dependence of lift on atmospheric density and the quadratic dependence on the RPM. The strong dependence on RPM is particularly important to the pilot since it is the only parameter in Eq. (68) that he has control over (although the pilot also has some control over $C_{L_{\text{ave}}}$ via the collective and cyclic).

The R^4 term is obviously the most important term in Eq. (68), but its usefulness is limited by two factors. The most obvious limitation is the structural problems created by the large torque at the blade tips. Another serious limitation is the speed of sound. Although it may seem strange to apply Mach-number-related effects to a stationary (hovering) helicopter, the blade tips can easily reach transonic speeds at high RPM. This causes all the problems encountered by transonic aircraft, as described in Sec. 4.2, most notably a huge increase in drag, which would put impractical loads on the drive train and power plant.

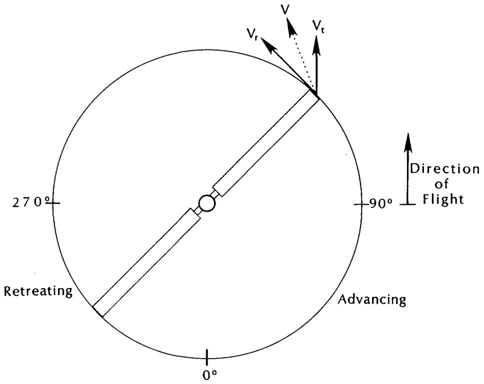

FIG. 50. Vector relation between rotational and translational velocities.

5.2 Forward Flight

The aerodynamic analysis of rotor blades in forward flight (or hovering on a windy day) is more complicated than that for hovering since, for each blade section, the magnitude and the direction of the relative wind are constantly changing. This results in an unsteady flow field. Each section experiences a translational velocity, v_t, in addition to its rotational velocity, v_r. The net velocity is the vector sum of v_t and v_r, as shown in Fig. 50. When the "advancing" blade is at the 90° position in Fig. 50, it has the greatest net velocity. Conversely, the "retreating" blade experiences the least net velocity at the 270° position. At intermediate angles, the magnitude of the net velocity is between these two values and its direction is not perpendicular to the blade, resulting in a condition very similar to a "swept" fixed wing, discussed in Sec. 4.2.

One unique problem caused by the vector addition of v_r and v_t is that there will be a part of the retreating blade where the flow will be reversed. This will occur since v_r decreases towards the hub, eventually becoming less than v_t. This causes the airflow to go from the trailing edge to the leading edge of the rotor blade. Obviously, this part of the rotor blade will be a very inefficient lift producer.

A more serious problem is the overall lift asymmetry between the advancing and retreating blades caused by the higher average airspeed on the advancing blade. In early rotorcraft, with rigid blades rigidly attached to the hub, the left rolling tendency caused by this lift asymmetry was often the chief limit on forward airspeed. The obvious solution to this problem is to reduce the angle of attack on the advancing blade and to increase it on the retreating blade, thus compensating for the velocity asymmetry. This is called *cyclic feathering*, and some helicopters do it by mechanically reducing the pitch on each advancing blade and increasing its pitch as it becomes the retreating blade.

However, most rotorcraft are able to avoid cyclic feathering by taking advantage of the natural structural dynamics of flexible rotor blades and hinges, in a process called *flapping*. As the rotor blade reaches the advancing side, the additional lift "upflaps" the blade (this happens because the moment applied to the blade is not instantaneously transmitted to the hub), giving a downward component to the relative wind, thereby decreasing the angle of attack. A similar "downflapping" occurs on the retreating blade which gives an upward component to the relative wind, naturally increasing the retreating blade's angle of attack. The natural frequency of this flapping motion is exactly equal to once per blade revolution regardless of the blade's lift characteristics or stiffness. Flexible blades attached by "flapping hinges" to the hub can be nearly as effective as cyclic feathering, and they are far simpler to implement and maintain.

A fundamental limit on the ability of flapping to correct the lift asymmetry between the advancing and the retreating blades is the *retreating blade stall*. The increase of the angle of attack on the retreating side eventually causes stalling of the blade, resulting in a large loss of lift and a severe lift asymmetry, and a need for additional power. Either the dramatic loss of lift or a lack of power could limit the forward speed of a rotorcraft. The analysis of a retreating blade stall is complex. First of all, this is a "dynamic" stall, i.e., the blade is only at a high angle of attack momentarily, so it does not necessarily stall at the static stall angle of attack. A second, more fundamental, complexity involves the spanwise "twisting" of the blades. An untwisted blade would have its maximum angle of attack at the blade tip and would stall there first. However, since the angle of attack of the blade is deliberately increased toward the center of the rotor blade, it is not obvious which part of the blade will stall first. Most helicopters have their maximum angle of attack in a region bounded by 270°–300° (see Fig. 50) at radii from 70% to 100% of the blade length.

5.3 Descending Flight

In level flight, the thrust of a rotorcraft is partially perpendicular to the direction of flight, to offset weight, and partially in the direction of flight, to offset drag. In a vertical descent, however, the thrust of the rotor is directly opposite the direction of flight, creating a very different flow field. The flow field for descending flight is usually categorized as one of the following four states: normal thrusting, vortex ring, autorotative, or windmill brake. An idealized flow field for each of the states is shown in Fig. 51. In all of the states, there will be a vortex due to pressure gradients (similar to the tip vortices discussed in Sec. 4.1) at the tip of the rotor blade and where the flow field reverses.

In the normal thrusting state, the velocity of the air induced through the rotor exceeds the rate of descent. This condition exists for hovering and low rates of descent up to approximately 70% of the ideal hovering velocity, given by

$$v_{i_{\text{hover}}} = \left(\frac{\text{Weight}}{2\rho A}\right)^{1/2} = \left(\frac{\text{Disk loading}}{2\rho}\right)^{1/2}. \tag{69}$$

All the flow is downward (relative to the rotor) in this state but of unequal magnitude because of the blade twist and the differences in rotational velocity along the span of the blade. The thrust is steady, and a constant amount of power is required from the engine to maintain RPM.

The vortex ring state occurs between 70% and 150% of $v_{i_{\text{hover}}}$. Thrust may vary by as much as plus or minus a third in this condition. There will also be increased vibration and a tendency to produce even higher descent rates. This state is somewhat analogous to a pilot in a fixed-wing aircraft flying through his own wake. This is because the

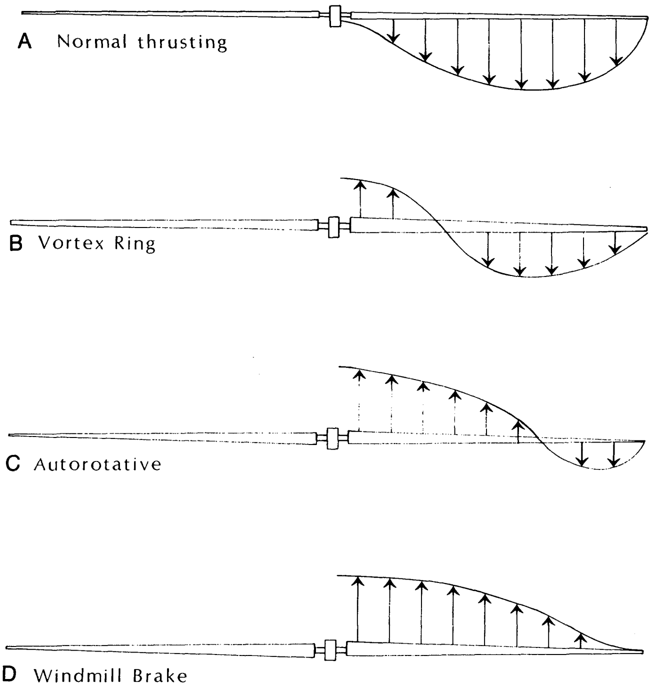

FIG. 51. Flow fields in vertical descents.

high rate of descent has overcome the relatively low downward-induced velocity near the rotor hub and caused a flow reversal. The vortex formed at the flow reversal is very turbulent, disturbing the flow field over a large area of the disk, resulting in a loss of thrust and thrust variations. Rotorcraft pilots are warned to avoid the vortex ring state.

At descent rates beyond the vortex ring state up to around 180% of $v_{i_{\text{hover}}}$, the vortex at the flow reversal will become steadier and much less turbulent. In this autorotative state, there will be some descent rate where no power will be required to maintain RPM. This state of autorotation is extremely important in case of engine failure because, although the pilot cannot maintain level flight, he can generate enough thrust to maintain a controlled equilibrium descent to the ground. In effect, the rate of loss of potential energy in autorotation is equal to the power required for the steady-state descent. The autorotative state can also be thought of as the boundary between conditions where power must be used to prevent RPM decrease, and conditions where power must be extracted from the rotor to prevent RPM increase.

The state where power must be extracted from the rotor to prevent RPM increase is called the *windmill brake state* and is, in fact, the state in which windmills operate in order to extract energy from the wind. This is not a normal operating state for any rotorcraft. However, pilots will purposely enter this state as a transient condition just prior to the landing from an autorotative approach. The increased rotational kinetic energy in the rotor can be used to "flare" the rotorcraft, reducing the rate of descent to nearly zero for landing.

5.4 Summary

All of the rotorcraft aerodynamics presented are idealized and simplified cases that are merely meant to show the most obvious differences in dealing with rotating wings versus a fixed wing. The flow field around rotor blades is, in general, very complex and is often handled these days through CFD (computational fluid dynamics), the topic of the next section.

6. COMPUTATIONAL FLUID DYNAMICS

6.1 General Remarks

Computational fluid dynamics (CFD) is certainly not pure theoretical analysis. CFD changes the fundamental nature of the analysis from calculus to arithmetic, and from the continuum domain to the discrete domain, so that it can be solved using a computer. The existing mathematical theory for numerical solutions of nonlinear partial differential equations is continuously being improved. In computational fluid dynamics, it is necessary to rely heavily on the mathematical analysis of simpler, linearized formulations and on heuristic reasoning, physical intuition, wind-tunnel experience, and trial-and-error procedures.

Miranda (1984) divides the flow fields into zones: (1) the inviscid zone, where the effects of viscosity are negligible; and (2) the viscous zone, where the effects of viscosity have a significant role, such as in the boundary layer and in the near wake. Kutler (1983) states that "The use of zonal methods is advantageous because (1) the regions over which complex equation sets are required are minimized, (2) the simpler equation sets have faster solution algorithms which result in improved computational efficiency, and (3) the computer storage requirements for the overall procedure are not as great."

Chapman (1979) cites four major stages of successively refined approximations to the full Navier–Stokes equations [Eqs. (5a)– (5c)] in order of their evolution and complexity: (I) linearized inviscid, (II) nonlinear inviscid, (III) Reynolds averaged Navier–Stokes, and (IV) full Navier–Stokes. Numerical methods using the stage-I approximation are termed *panel methods*, since complex aircraft geometries are modeled by a large number of contiguous surface panels. The stage-II approximations neglect only the viscous terms and have been used to predict transonic and hypersonic flow fields. For the stage-III approximations, the basic equations are averaged over a time interval, long compared to turbulent eddy fluctuations, yet small compared to macroscopic flow changes. Since terms representing the time-averaged transport of turbulent momentum and energy must be modeled, the principal inaccuracy of stage-III approximations is that of modeling turbulence. In its full complexity, the stage-IV approximation involves the direct simulation of turbulent eddies from the complete time-dependent Navier–Stokes equations. Chapman states that "Progressive advances in computer power and in numerical methods have made possible the development of a variety of computational codes ranging from stage-I codes for complex aircraft configurations to stage-IV codes for very simple geometry." Although there have been significant advances both in computer hardware and in computer software since that statement first appeared, it remains qualitatively correct.

Phenomenal progress in CFD has been made over the past decade and continues to be made. Thus, for design cases in which the flow is primarily attached with thin boundary layers such as transports in the cruise configuration, CFD can be used directly in the design cycle. However, as noted by Bradley (1987), "The CFD capability for more complex designs, such as tactical aircraft, has not progressed as far because the flow fields are dominated by flow separations and vortex interactions."

Boppe (1988) stated:

> It can be appreciated that subtle discrepancies between computed and experimental pressures will have different effects on drag and pitching moments obtained via pressure integration depending on local position and geometry. A small pressure anomaly near the middle of an aerodynamic configuration (where surface shaping is nearly aligned with the onset flow) will sum to produce a negligible contribution to total drag and if the location is near the moment center... a negligible effect on overall pitching moment. If the pressure anomaly, even though small, is positioned aft on the configuration, say on a nozzle boattail, a significant drag effect will register due to integration on an aft-facing surface and significant moment effects can register since the moment arm is large.

However, good agreement between the computed pressures and the measured values does not guarantee satisfactory predictions for the drag, since exact calculations of the boundary layer and, therefore, the skin friction are difficult. As noted by Jameson (1987), "Since a universally satisfactory turbulence model has yet to be found, current turbulence models have to be tailored to the particular flow."

7. NONAEROSPACE APPLICATIONS OF AERODYNAMICS

7.1 General Remarks

As noted in the Introduction, the aerodynamicist applies the principles of fluid mechanics to determine the forces and moments acting on an object moving through the air, e.g., a golf ball or an airplane, or due to the air moving past an object, e.g., the wind loads acting on a building. Some nonaerospace applications of aerodynamics are discussed in this section.

7.2 Automotive Aerodynamics

Aerodynamics has a significant effect on the road holding, driving performance, contamination of glass panels, and fuel economy of an automobile. In recent years, there have been extensive world-wide efforts to reduce fuel consumption by reducing the aerodynamic drag of motor vehicles. As shown in Fig. 52, there has been a considerable reduction in aerodynamic drag.

In order to achieve a low drag coefficient for a series production car, it is necessary to include the aerodynamic requirements into the styling and the construction concept throughout the development of the vehicle. For instance, side windows are an area in which aerodynamic optimization and production-cost control have conflicting requirements. Flush glass, desirable from the aerodynamic point of view, requires expensive manufacturing processes.

Design of the front end must accommodate the competing requirements of sufficient internal air flow to meet engine cooling and other requirements while minimizing drag of the external aerodynamics. Emmelmann, Berneburg, and Schulze (1990) note that "The optimum positioning of the front cooling air inlets on a vehicle plays an important part in achieving a low C_D value along with good engine-cooling characteristics. From an aerodynamic standpoint the optimum front would be completely closed, as air flow into the engine compartment causes drag.... Drag inducing airflow into the engine compartment is minimized by the use of flexible mouldings between the leading edge of the hood, and the headlamps and grill, and also between the headlamp assemblies and the bumpers." The size and location of the opening are other parameters in the front-end design.

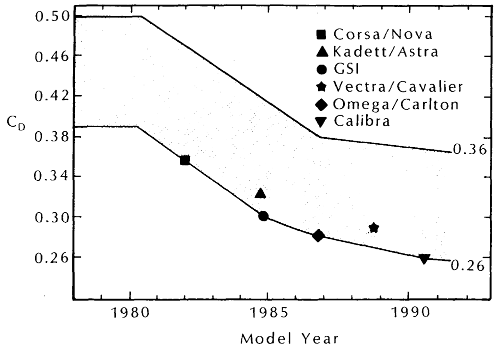

FIG. 52. The development of the drag coefficient of Opel/Vauxhall vehicles. [From Emmelmann, Berneburg, and Schulze (1990).]

Morel (1978) explored the dependence of the near-wake flow on the geometry of the afterbody characteristics of hatchback automobiles. The drag coefficient data showed a very strong variation of C_D with the slant angle. The data presented in Fig. 53, which are taken from Costelli (1984), illustrate the critically important parameter of back-window slant angle.

Nedley (1979) notes that subtle changes in the surface in critical areas influence drag significantly. Furthermore, the effect of a particular change is specific to the vehicle and may not apply to other vehicles. Emmelmann, Berneburg, and Schulze (1990) note that trend statements made on the basis of numerical code analysis of simplified body shapes cannot be expected to be valid for real vehicles.

As a result, extensive wind-tunnel tests are conducted throughout the development of an aerodynamically efficient vehicle. Scale models, typically 1:4 or 1:5, are tested early in the styling phase. Wind-tunnel tests using prototypes and pilot vehicles allow the designer to check that the obtained optimization features are effective. A photograph of a test conducted in the General Motors Aerodynamics Laboratory is presented in Fig. 54.

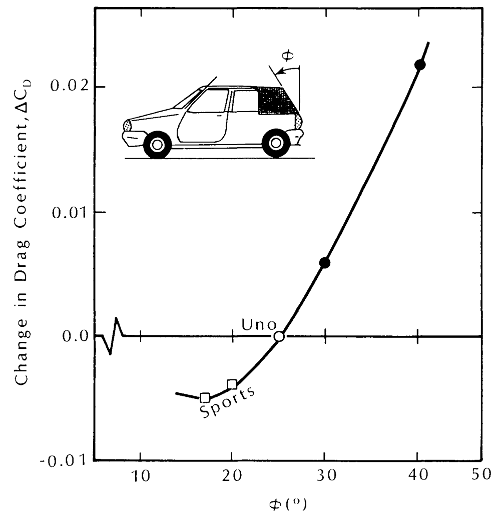

FIG. 53. The influence of back-window slant angle on the change in drag coefficient. [From Costelli (1984).]

7.3 Wind Loads on Buildings

To determine the wind loads acting on a building, one is confronted with a complex, unsteady, three-dimensional flow field containing turbulence, vortices, and large regions of separated flow. Even the approach flow away from the building (i.e., the free-stream flow of the previous discussions) is a complex, three-dimensional flow because it is in the atmospheric boundary layer adjacent to the earth's surface. The height at which viscous effects become negligible is the gradient height, which is a function of surface roughness. Typical values of the gradient height are 275 m for flat, open country, 400 m for suburban terrain, and 500 m for an urban area. Above the gradient height, the frictionless wind balance is established, and the air moves at the gradient wind velocity along isobars. Within the atmospheric boundary layer, the wind crosses the isobars. Thus, in the "free-stream" flow, the velocity varies with distance from the surface and forms a spiral.

Because the velocity approaching a tall building increases with distance above the ground, the pressures acting on the building differ from one floor to another. This is illustrated in the data of Dalgliesh (1970) that are reproduced in Fig. 55. On the windward surface, the positive pressure coefficients are much higher on the 33rd floor than on the 10th floor. Because subsonic flows cannot turn sharp corners, the flow separates from the side walls (those parallel to the free-stream flow) creating large negative pressure coefficients. Furthermore, the flow is unsteady in these separated regions with pressure fluctuations approaching $0.5q_\infty$. Windows located in these regions of large negative pressure coefficients may be sucked out of their frame, falling to the street below. Thus, the designer must be concerned not only with the integrated wind loads on a building but with local flow phenomena.

The proximity of adjacent buildings further complicates the flow field. As shown in the sketch of Fig. 56, a vortex forms off the lower building in the foreground. Air descending close to the windward wall flows through openings located at the ground level of the high-rise building. Regions of accelerated flow are produced around the vertical

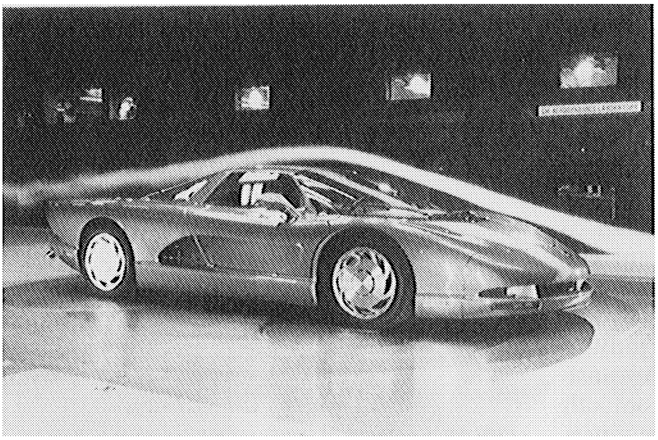

FIG. 54. Photograph of a test in the General Motors Aerodynamics Laboratory (photograph provided by K. Staggemeier).

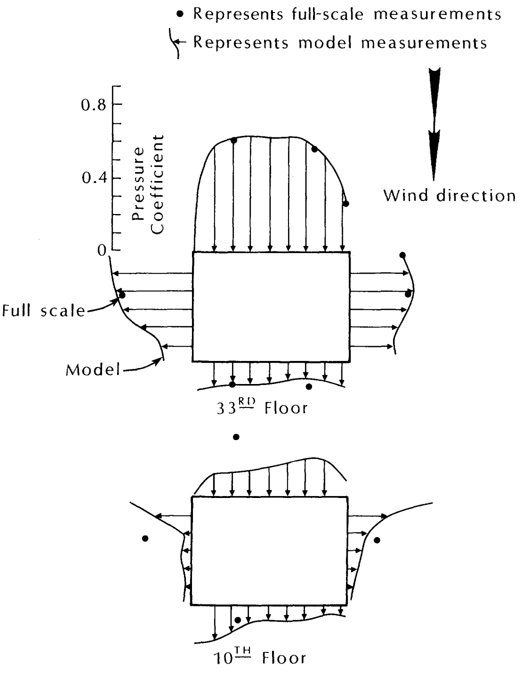

FIG. 55. Mean pressure coefficients acting on a 34-story office building. [From Dalgliesh (1970).]

and horizontal corners of the high-rise building. The flow separates from the roof and side walls, causing large negative pressure coefficients in these regions. Many design problems are presented in these regions containing vortices, separated flows, and through flow.

7.4 Aerodynamics of Sports Balls

Aerodynamics plays a prominent role in many sports where a ball is either struck or thrown through the air. Lateral deflection in flight (variously known as swing, swerve, or curve) is well recognized in cricket, baseball, golf, and tennis. In most of these sports, the swing is obtained by spinning the ball about an axis perpendicular to the line of flight, which gives rise to what is commonly known as Magnus effect. According to Mehta (1985), the Magnus effect has been attributed to asymmetric boundary-layer separation. The effect of spin is to delay separation on the retreating side and to enhance it on the advancing side. Clearly, this would occur only at postcritical Reynolds numbers, when transition has occurred on both sides. A smooth sphere rotating slowly can experience a negative Magnus force at precritical Reynolds numbers, when transition occurs first on the advancing side.

Mehta (1985) describes the aerodynamics of a cricket ball, a baseball, and a golf ball. Details of these balls, together with typical flow conditions, are presented in Fig. 57.

A cricket ball has six rows of prominent stitching, with typically 60–80 stitches in each row (primary seam). The stitches lie along the equator holding the two leather hemispheres together. The better cricket balls are, in fact, made out of four pieces of leather, so that each hemisphere has a line of internal stitching forming a "secondary seam."

Detailed pressure distributions along the equator of a cricket ball are reproduced in Fig. 58. The measurements on the seam side of the ball, which is at an angle of attack of 20°, are those for which θ is positive. At low Reynolds numbers, i.e., the lowest velocity, the pressure distribution is symmetric, so there would be no side force. When $V=25$ m s^{-1}, the pressure distribution exhibits an asymmetry, which would cause the ball to swing to the right. The maximum pressure difference between the two sides occurs when $V=29$ m s^{-1}, when presumably the boundary layer on the seam side is fully turbulent while that on the nonseam side is still laminar. The actual (critical) flight speeds at which the asymmetry appears or disappears were found to be a function of the seam angle, surface roughness, and spin rate of the ball.

In the aerodynamics of golf balls, an important objective is to obtain maximum distance of flight. Thus, the lift force (which is generated by imparting backspin to the ball), the drag force, and the gravitational force are critical parameters. The effect of dimples is to lower the critical Reynolds number. As shown in the data of Achenbach (1972),

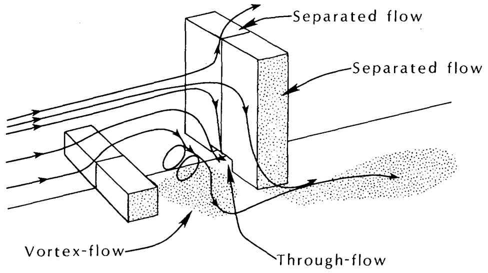

FIG. 56. Features of flow around adjacent buildings.

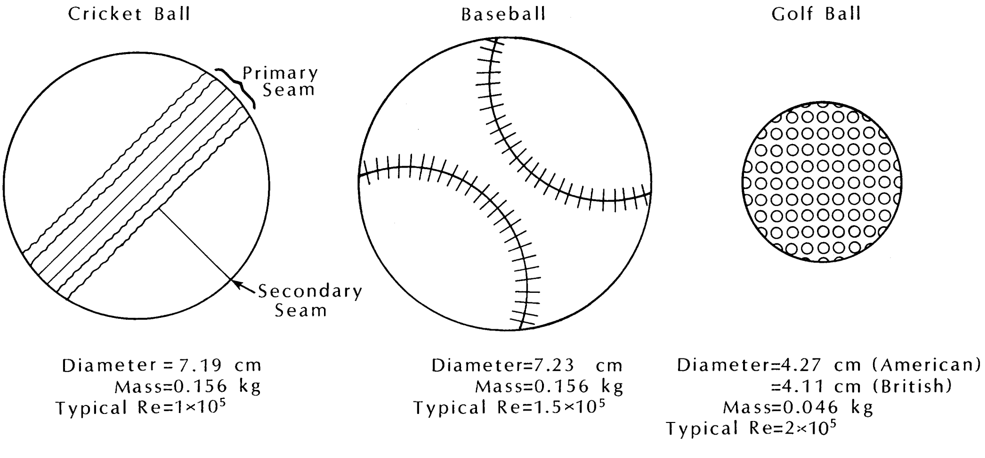
Cricket Ball
Primary Seam
Secondary Seam
Diameter = 7.19 cm
Mass=0.156 kg
Typical Re=1×10^5
Baseball
Diameter=7.23 cm
Mass=0.156 kg
Typical Re=1.5×10^5
Golf Ball
Diameter=4.27 cm (American)
=4.11 cm (British)
Mass=0.046 kg
Typical Re=2×10^5

FIG. 57. Typical ball dimensions and flow conditions.

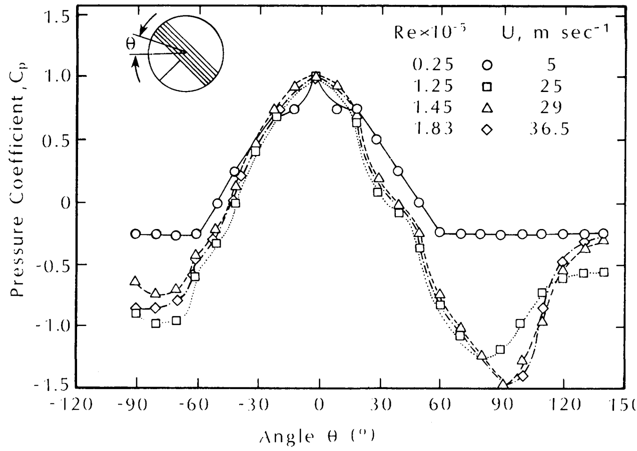

FIG. 58. Pressure distributions on a cricket ball held at an incidence angle of 20°. [From Mehta (1985).]

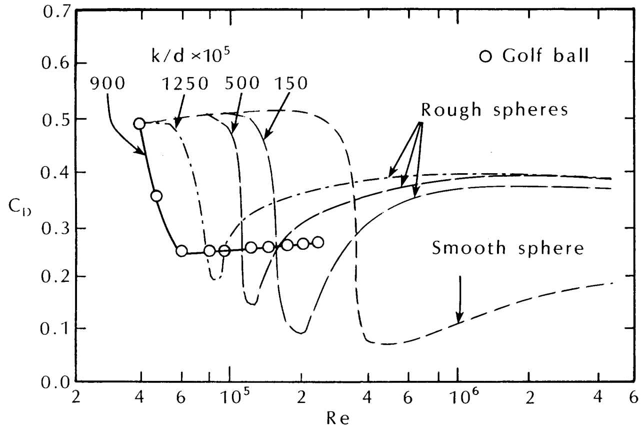

FIG. 59. Comparison of the drag for golf balls and for spheres with various heights of sand-grain roughness. The sand-grain roughness height is *k*, and *d* is the diameter of the ball. [From Achenbach (1972).]

which are reproduced in Fig. 59, dimples are more effective than sand-grain roughness at reducing the critical Reynolds numbers. The rapid rise in C_D for the sphere with sand-grain roughness is due to the forward movement of the transition point and the artificial thickening of the boundary layer by the roughness elements. The value of C_D for the golf ball does not rise in this way, which indicates that dimples are effective at tripping the boundary layer without causing the thickening associated with positive roughness. In golf, the Reynolds number changes continuously during flight, which means that C_D and C_L will also vary significantly.

ACKNOWLEDGMENTS

The authors would like to thank Jeanie Paterson for cheerfully typing the many drafts of this manuscript and Timothy Valdez for creatively illustrating the manuscript. The authors would also like to express their appreciation to DFSIV of the U.S. Air Force Academy for providing many of the photographs.

GLOSSARY

Aircraft Moments and Axes: The roll, pitch, and yaw moments and the vertical, lateral, and longitudinal axes are all illustrated in Fig. 19.

Airfoil Terminology: Standard terminology for the cross section of an airfoil is shown below:

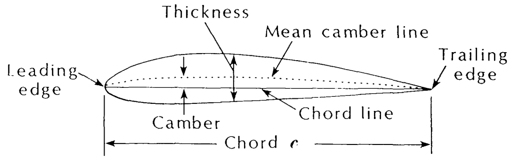

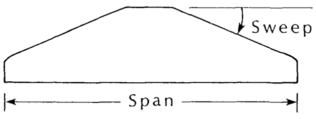

Drag: Drag is discussed in Sec. 3.5. A summary of the types of drag is as follows: (i) *Friction drag or skin-friction drag* are synonomous terms. *Form drag* and *pressure drag* are also used interchangeably. Skin-friction drag and form drag are different components of "zero lift drag," caused by shear forces. (ii) *Vortex drag* is used interchangeably with *induced drag* and is the type of drag caused by lift. (iii) *Wave drag* is another type of drag. It is associated with compressibility effects at Mach numbers above 0.5.

List of Works Cited

Abbott, I. H., von Doenhoff, A. E. (1949), *Theory of Wing Sections*, New York: Dover.

Achenbach, E. (1972), *J. Fluid Mech.* **54**, 565–575.

Ames Research Staff (1953), National Advisory Committee for Aeronautics Report No. 1135 (unpublished).

Bertin, J. J. (1989), in: Bertin, J. J., Glowinski, R., Periaux, J. (Ed.), *Hypersonics, Volume I: Defining the Hypersonic Environment*, Boston: Birkhäuser.

Bertin, J. J., Smith, M. L. (1989), *Aerodynamics for Engineers*, Englewood Cliffs, NJ: Prentice-Hall.

Boppe, C. W. (1988), Proceedings of the Meeting of the Advisory Group for Aerospace Research and Development on Validation of Computational Fluid Dynamics (unpublished).

Bradley, R. G. (1981), Procedings of the NASA/Ames Research Center Transonic Perspective Symposium (unpublished).

Bradley, R. G. (1987), Proceedings of the American Institute of Aeronautics and Astronautics 8th Computational Fluid Dynamics Conference (unpublished).

Bushnell, D. M., Cary, A. M., Jr., Harris, J. E. (1977), NASA Report No. SP-422 (unpublished).

Cebeci, T., Smith, A. M. O. (1974), *Analysis of Turbulent Boundary Layers*, Orlando: Academic.

Chapman, D. R. (1979), Proceedings of the American Institute of Aeronautics and Astronautics 17th Aerospace Sciences Meeting, New Orleans, LA (unpublished).

Costelli, A. F. (1984), Society of Automotive Engineers Paper No. 840297 (unpublished).

Dalgliesh, W. A. (1970), in *Wind Loads on Buildings and Structures*, National Bureau of Standards, Building Science Series 30, Washington, DC: U.S. GPO.

Emmelmann, H. J., Berneburg, H., Schulze, J. (1990), Society of Automotive Engineers Paper No. 900317 (unpublished).

Farren, W. S. (1956), *J. R. Aeronaut. Soc.* **60**, 431–449.

Hansen, C. F., Heims, S. P. (1958), National Advisory Committee for Aeronautics Technical Note No. 4359 (unpublished).

Jameson, A. (1987), Proceedings of the American Institute of Aeronautics and Astronautics 8th Computational Fluid Dynamics Conference (unpublished).

Koppenwallner, G. (1969), *Z. Flugwiss.* **21**, 81–88.

Küchemann, D. (1957), *J. R. Aeronaut. Soc.* **61**, 37–42.

Kutler, P. (1983), Proceedings of the American Institute of Aeronautics and Astronautics 2lst Aerospace Sciences Meeting, Reno, NV (unpublished).

Li, C. P. (1989), in: Bertin, J. J., Glowinski, R., Periaux, J. (Eds.), *Hypersonics, Volume II: Computation of Hypersonic Flow Fields*, Boston: Birkhäuser.

Mack, L. M. (1984), in: Michel, R. (Ed.), Special Course on Stability and Transition of Laminar Flow, AGARD Report No. 709 (unpublished).

Mehta, R. D. (1985), *Annu. Rev. Fluid Mech.* **17**, 151–189.

Miranda, L. R. (1984), *J. Aircr.* **21**, 355–369.

Morel, T. (1978), Society of Automotive Engineers, Paper No. 780267 (unpublished).

Nedley, L. (1979), Society of Automotive Engineers, Paper No. 790724 (unpublished).

Schlichting, H. (1979), *Boundary-Layer Theory*, New York: McGraw-Hill.

Stinton, D. (1966), *The Anatomy of the Airplane*, New York: American Elsevier.

Talay, T. A. (1975), National Aeronautics and Space Administration Report No. SP-367 (unpublished).

Thomas, A. S. W. (1985), Lockheed Horizons (unpublished).

U. S. Standard Atmosphere, 1976, Washington, DC: U.S. GPO.

Whitcomb, R. T. (1954), National Advisory Committee for Aeronautics Report No. 1273 (unpublished).

Further Reading

Anderson, J. D., Jr. (1984), *Fundamentals of Aerodynamics*, New York: McGraw-Hill.

Huenecke, K. (1987), *Modern Combat Aircraft Design*, Annapolis, MD: Naval Institute.

Kuechemann, D. (1978), *The Aerodynamic Design of Aircraft*, Oxford: Pergamon.

Kuethe, A. M., Chow, C. Y. (1986), *Foundations of Aerodynamics*, New York: Wiley.

Schlichting, H., Truckenbrodt, E. (1979), *Aerodynamics of the Airplane*, New York: McGraw-Hill.

Shevell, R. S. (1989), *Fundamentals of Flight*, Englewood Cliffs, NJ: Prentice-Hall.

Whitford, R. (1987), *Design for Air Combat*, London: Jane's.

AERONAUTICS

JOHN D. ANDERSON, JR., *Department of Aerospace Engineering, University of Maryland, College Park*

	Introductory Concepts	379
	Definition of Aeronautics	379
	The Source of Aerodynamic Force	380
	The Standard Atmosphere	380
	Historical Note	381
1.	**Aerodynamics of the Airplane**	381
1.1	Basic Airfoil Characteristics	381
1.2	Lift, Drag, and Moment Coefficients	384
1.3	Finite Wing Properties: Induced Drag	394
1.4	Swept Wings	396
1.5	Flaps—A Mechanism for High Lift	398
2.	**Airplane Performance**	399
2.1	Equations of Motion	400
2.2	Maximum Velocity	401
2.3	Rate of Climb	401
2.4	Gliding Flight	401
2.5	Maximum Altitude	402
2.6	Range and Endurance	402
2.7	Turning Flight	403
3.	**Airplane Stability and Control**	404
3.1	Definition of Stability and Control	404
3.2	Moments on the Airplane	408
3.3	Criteria for Longitudinal Static Stability	408
3.3.1	The Neutral Point and Static Margin	411
4.	**Hypersonic Flight**	411
	Glossary	414
	Works Cited	414
	Further Reading	414

INTRODUCTORY CONCEPTS

Definition of Aeronautics

Aeronautics is the name given to the general discipline of flight through the atmosphere. Aeronautics is synonymous with all aspects of the aircraft—the understanding of the principles of its flight, its design, and performance. The basic principles of aeronautics were of interest as early as the 15th century when Leonardo da Vinci (1452–1519) filled hundreds of pages of his notes with ideas and concepts for heavier-than-air flying machines. These involved designs with mechanically driven flapping wings—flapping for both lift and forward thrust. The first time that human beings left the ground in sustained flight was 21 November 1783, when a hot-air balloon designed by the Montgolfier brothers carried two people for thirty minutes across Paris. However, the first pivotal work on aeronautics, both as a science and with practical applications, was carried out by Sir George Cayley (1773–1857) in England. Cayley was the first person to make public the concept of the modern configuration airplane, namely, a vehicle with a fixed wing, fuselage, and tail. The fixed wing was for the production of lift; the mechanism of propulsion was considered by Cayley as a totally separate device. The tail was for stability. Although da Vinci made note late in his life that a flying machine with fixed wings, rather than flapping wings for both lift and propulsion, might be the appropriate approach for flying machine design, his notes were misplaced and were not available to the general scholarly audience until the 19th century. Hence it is George Cayley who deserves the credit for independently conceiving and building the first modern-concept flying machine in history—a yard-long, hand-launched glider in 1804. Moreover, he published two important scientific papers in 1809–1810 which were the first to point out the basic principles of flight. Cayley's ideas were extended by numerous investigators during the 19th century, and culminated in the work of Wilbur and Orville Wright, who achieved the first successful heavier-than-air,

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

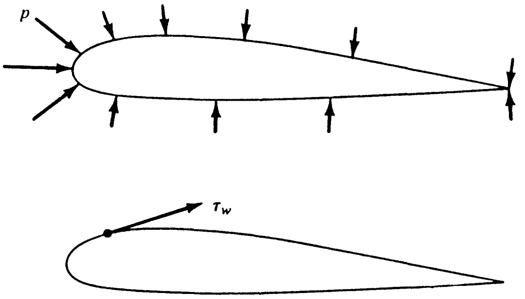

FIG. 1. Pressure and shear stress distributions acting on the surface of an airfoil. The flow is from left to right. The pressure and shear stress act at every point on the body. [From Anderson (1989).]

sustained, controlled, manned flight on September 17, 1903, at Kill Devil Hills, near Kitty Hawk, North Carolina. This article will focus on such heavier-than-air machines, and therefore will deal primarily with aerodynamic phenomena.

The Source of Aerodynamic Force

A flight vehicle moving through the atmosphere experiences a net, resultant aerodynamic force exerted by the air on the vehicle. This force stems from two natural sources: the pressure distribution exerted by the airflow over the surface, and the shear stress due to frictional action between the airflow and the solid surface. Referring to Fig. 1, we see that pressure exerted by the gas on the solid surface of an object always acts *normal* to the surface, as shown by the directions of the arrows. The lengths of the arrows denote the magnitude of the pressure at each local point on the surface. Note that the surface pressure varies with location. The net *unbalance* of the varying pressure distribution over the surface creates a force, an aerodynamic force. The second source, shear stress acting on the surface, is due to the frictional effect of the flow "rubbing" against the surface as it moves around the body. The shear stress τ_w is defined as the force per unit area acting *tangentially* on the surface due to friction, as shown in Fig. 1. It is also a point property; it also varies along the surface; and the net unbalance of the surface shear stress distribution also creates an aerodynamic force on the body. *No matter how complex the flow field, and no matter how complex the shape of the body, the only way nature has of communicating an aerodynamic force to a solid object or surface is through the pressure and shear stress distributions which exist on the surface.* These are the basic fundamental sources of all aerodynamic forces.

Finally, we can state that a primary function of theoretical and experimental aerodynamics is to predict and measure the aerodynamic forces on a body. In many cases, this implies prediction and measurement of p and τ_w along a given surface. Furthermore, a prediction of p and τ_w on the surface frequently requires knowledge of the complete flow field around the body. This is the main function of the science of *aerodynamics*. (See AERODYNAMICS.)

The Standard Atmosphere

In order to standardize flight data and wind tunnel data in aeronautics, it is essential to have standardized values of atmospheric properties, namely, the variation of air pressure, density, and temperature with altitude. In reality, the properties of the real atmosphere depend on geographic location around the earth, time of day, season, and even solar sunspot activity. Indeed, the real atmosphere is always a dynamically changing system. However, the standard atmosphere is based on experimental data averaged over long periods of time, along with some judicious applications of the laws of physics, as explained below.

The keystone of the standard atmosphere is a *defined* variation of T with altitude, based on experimental evidence. This variation is shown in Fig. 2. Note that it consists of a series of straight lines, some vertical (called the constant-temperature, or *isothermal*, regions) and some inclined (called the *gradient* regions). Given $T=T(h)$ as defined by Fig. 2, where h is altitude, then $p=p(h)$ and $\rho=\rho(h)$ follow from the laws of physics. The first of these laws is the hydrostatic equation

$$dp=-\rho g\, dh,$$

where g is the acceleration of gravity, and the second is the equation of state for a perfect gas,

$$p=\rho RT.$$

From these equations, we obtain for the pressure and density variations in the isothermal regions

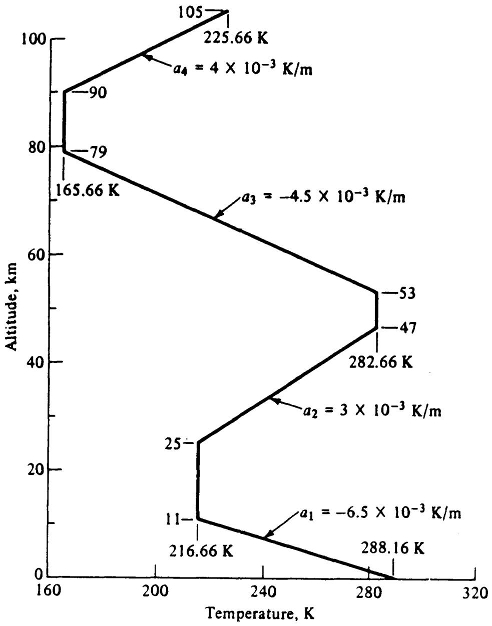

FIG. 2. Temperature distribution in the standard atmosphere. [From Anderson (1989).]

$$\frac{p}{p_1} = e^{-(g/RT)(h-h_1)}$$

and

$$\frac{\rho}{\rho_1} = e^{-(g/RT)(h-h_1)}.$$

In the gradient regions, we have

$$\frac{p}{p_1} = \left(\frac{T}{T_1}\right)^{-g/aR}$$

and

$$\frac{\rho}{\rho_1} = \left(\frac{T}{T_1}\right)^{-[(g/aR)+1]}.$$

Recall that in the gradient regions the variation of T is linear with altitude and is given by the specified relation

$$T = T_1 + a(h - h_1),$$

where a is the *lapse rate*, $a = dT/dh$. The standard atmosphere is pieced together from the above equations, starting from the standard sea level values of pressure, density, and temperature, which are

$$p_s = 1.0135 \times 10^5 \text{ N/m}^2 = 2116.2 \text{ lb/ft}^2,$$

$$\rho_s = 1.2250 \text{ kg/m}^3 = 0.002\,377 \text{ slug/ft}^3,$$

$$T_s = 288.16 \text{ K} = 518.69\ {}^\circ\text{R}.$$

Tables of the standard atmosphere are available in most technical libraries.

Historical Note

The driving philosophy of many advancements in aeronautics since 1903 has been to fly *faster and higher.* This is dramatically evident from Fig. 3, which gives the flight speeds for typical aircraft as a function of chronological time. Note the continued push for increased speed over the years, and the particular increase in recent years made possible by the jet engine. Singled out in Fig. 3 are the winners of the Schneider Cup races between 1913 and 1931 (with a moratorium during World War I). The Schneider Cup races were started in 1913 by Jacques Schneider of France as a stimulus to the development of high-speed float planes. They prompted some early but advanced development of high-speed aircraft. The winners are shown by the dashed line in Fig. 3, for comparison with standard aircraft of the day. Indeed, the winner of the last Schneider race in 1931 was the Supermarine S.6B, a forerunner of the famous Spitfire of World War II. Of course, today the maximum speed of flight has been pushed to the extreme value of 36 000 ft/s, escape velocity from the earth, by the Apollo lunar spacecraft.

As a companion to speed, the maximum altitudes of typical manned aircraft are shown in Fig. 4 as a function of chronological time. The same push to higher values is evident; so far, the record is the moon in 1969.

1. AERODYNAMICS OF THE AIRPLANE

1.1 Basic Airfoil Characteristics

The aerodynamic properties of a complete airplane are a synthesis of the properties of its parts—the wing, fuselage, tail, engine nacelle, etc. To describe the aerodynamics of the airplane, we start with the primary lifting surface, namely, the wing. The cross-sectional shape of the wing is called an *airfoil.* Such an airfoil is sketched in Fig. 5, which defines some basic terminology. The major design feature of an airfoil is the *mean camber line,* which is the locus of points halfway between

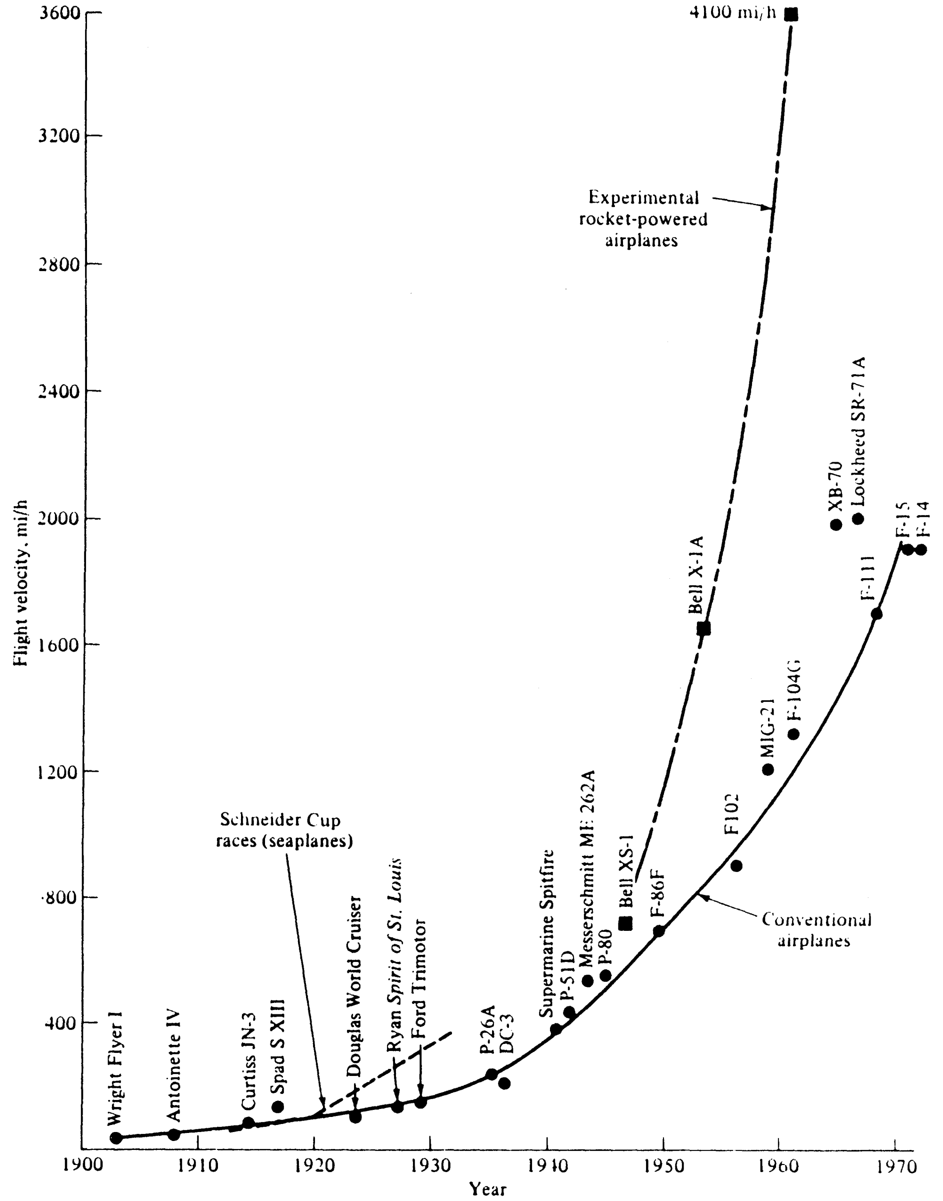

FIG. 3. Typical flight velocities over the years. [From Anderson (1989).]

the upper and lower surfaces as measured perpendicular to the mean camber line itself. The most forward and rearward points of the mean camber line are the *leading* and *trailing edges*, respectively. The straight line connecting the leading and trailing edges is the *chord line* of the airfoil, and the precise distance from the leading to the trailing edge measured along the chord line is simply designated the *chord* of the airfoil, given by the symbol *c*. The *camber* is the maximum distance between the mean camber line and the chord line, measured perpendicular to the chord line. The camber, the shape of the mean camber line, and, to a lesser extent, the thickness distribution of the airfoil essentially control the lift and moment characteristics of the airfoil.

More definitions are illustrated in Fig. 6, which shows an airfoil inclined to a stream of

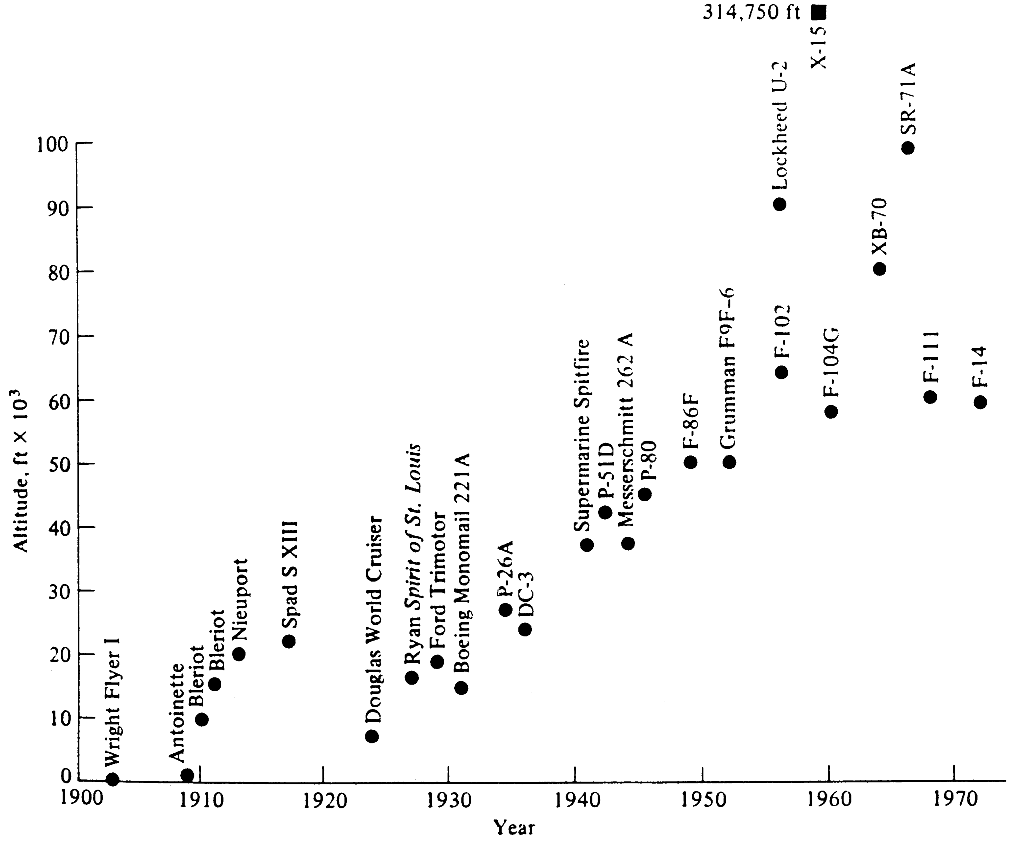

FIG. 4. Typical flight altitudes over the years. [From Anderson (1989).]

air. The freestream velocity V_∞ is the velocity of the air far upstream of the airfoil. The *direction* of V_∞ is defined as the *relative wind.* The angle between the relative wind and the chord line is the *angle of attack* α of the airfoil. As described earlier, there is an aerodynamic force created by the pressure and shear stress distributions over the wing surface. This resultant force is shown by the vector *R* in Fig. 6. In turn, the aerodynamic force *R* can be resolved into two forces, parallel and perpendicular to the relative wind. The *drag D* is always defined as the component of the aerodynamic force *parallel to the relative wind.* The *lift L* is always defined as the component of the aerodynamic force *perpendicular to the relative wind.*

In addition to lift and drag, the surface pressure and shear stress distributions also create a *moment M* which tends to *rotate* the

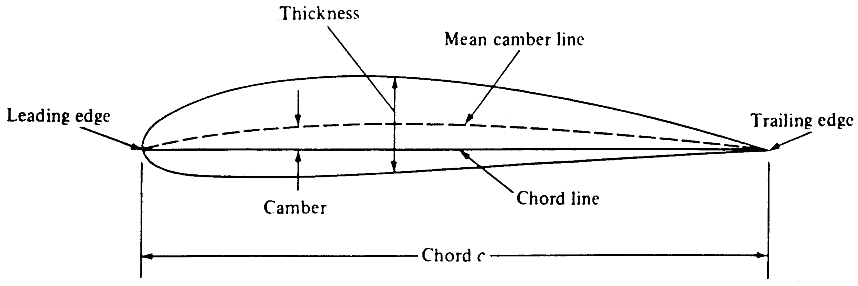

FIG. 5. Airfoil nomenclature. The shape shown here is a NACA 4415 airfoil. [From Anderson (1989).]

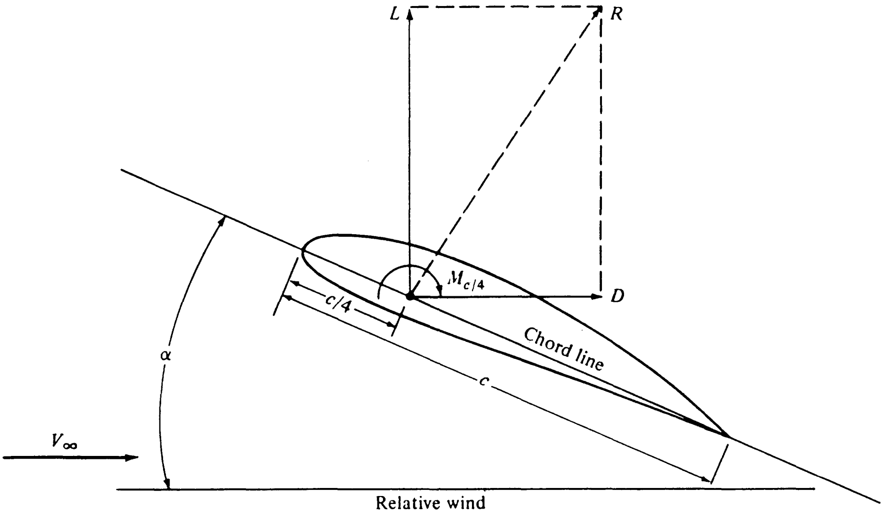

FIG. 6. Sketch showing the definitions of lift, drag, moments, angle of attack, and relative wind. [From Anderson (1989).]

wing. To see more clearly how this moment is created, consider the surface pressure distribution over an airfoil (we will ignore the shear stress for this discussion). Consider just the pressure on the top surface of the airfoil. This pressure gives rise to a net force F_1 in the general downward direction. Moreover, F_1 acts through a given point on the chord line, point 1, which can be found by integrating the pressure times distance over the surface (analogous to finding the centroid or center of pressure from integral calculus). Now consider just the pressure on the bottom surface of the airfoil. This pressure gives rise to a net force F_2 in the general upward direction, acting through point 2, where point 2 is generally a different location than point 1 on the airfoil. The total aerodynamic force on the airfoil is the *summation* of F_1 and F_2, and lift is obtained when $F_2 > F_1$. However, F_1 and F_2 will create a moment which will tend to rotate the airfoil. Moreover, the value of this aerodynamically induced moment depends on the point about which we chose to take moments. For example, if we take moments about the leading edge, the aerodynamic moment is designated M_{LE}. It is more common in the case of subsonic airfoils to take moments about a point on the chord at a distance $c/4$ from the leading edge, the *quarter-chord point*, as illustrated in Fig. 6. This moment about the quarter chord is designated $M_{c/4}$. In general, $M_{\mathrm{LE}} \neq M_{c/4}$. Intuition will tell you that lift, drag, and moments on a wing will change as the angle of attack α changes. In fact, the variations of these aerodynamic quantities with α represent some of the most important information that an airplane designer needs to know. We will address this matter in the following sections. However, it should be pointed out that, although M_{LE} and $M_{c/4}$ are both functions of α, there exists a certain point on the airfoil about which moments essentially *do not* vary with α. This point is defined as the *aerodynamic center*, and the moment about the aerodynamic center is designated M_{ac}. By definition,

$$M_{\mathrm{ac}} = \text{const}$$

independent of angle of attack. The location of the aerodynamic center for real aerodynamic shapes can be found from experiment. For low-speed subsonic airfoils, the aerodynamic center is generally very close to the quarter-chord point.

1.2 Lift, Drag, and Moment Coefficients

For an airfoil, it is convenient to define the lift, drag, and moment as those acting on a unit span of a wing, i.e., *L*, *D*, and *M* discussed earlier are per unit span. Referring again to Fig. 6, we visualize a section of a wing which extends a unit length perpendicular to the page. In turn, we can define a reference area,

S, as the planform area of this unit span, i.e., the projected area as seen looking down on the wing from above. Hence, $S=c\times 1=c$. Also, the freestream dynamic pressure, q_∞, is defined as $q_\infty=\frac{1}{2}\rho_\infty V_\infty^2$. With these quantities, nondimensional aerodynamic force and moment coefficients are defined as follows:

lift coefficient,

$$c_l=L/q_\infty S=L/q_\infty c; \tag{1}$$

drag coefficient,

$$c_d=D/q_\infty S=D/q_\infty c; \tag{2}$$

moment coefficient,

$$c_m=M/q_\infty Sc=M/q_\infty c^2. \tag{3}$$

In addition to their being dimensionless numbers, the use of c_l, c_d, and c_m is very fundamental to the science of aerodynamics. Dimensional analysis shows that, for a given shaped body, c_l, c_d, and c_m depend on angle of attack, α, and the dimensionless parameters, Mach number M and Reynolds number Re. The freestream Mach number is defined as

$$M_\infty=V_\infty/a_\infty,$$

where a_∞ is the freestream speed of sound, and the freestream Reynolds number is defined as

$$\mathrm{Re}=\rho_\infty V_\infty c/\mu_\infty$$

where μ_∞ is the coefficient of viscosity of the free stream. The results of dimensional analysis show that

$$c_l=f_1(\alpha,M_\infty,\mathrm{Re}), \tag{4}$$

$$c_d=f_2(\alpha,M_\infty,\mathrm{Re}), \tag{5}$$

$$c_m=f_3(\alpha,M_\infty,\mathrm{Re}). \tag{6}$$

This is a powerful result. It is the basis of the concept of *dynamic similarity* between two different flows. That is, for two flows with different freestream velocities, densities, and temperatures, over two geometrically similar bodies of different sizes at the same angle of attack, the lift, drag, and moment coefficients obtained with the two different flows will be the *same* as long as M_∞ and Re are the *same for the two different flows.* This dynamic similarity principle is used to relate the values of aerodynamic coefficients measured in a wind tunnel to those in free flight through the atmosphere. Specifically, if in the wind tunnel, the free-flight values of M_∞ and Re are reproduced, then c_l, c_d, and c_m measured in the wind tunnel will be the same as obtained in free flight. In this context, M_∞ and Re are called *similarity parameters.* More detailed considerations that apply to compressible viscous flows involve additional similarity parameters: ratio of specific heats, $\gamma=c_p/c_v$; Prandtl number, $\mathrm{Pr}=\mu\, c_p/k$; and the wall-to-freestream temperature ratio, T_w/T_∞, where c_p and c_v are the specific heats at constant pressure and constant volume, respectively, and k is the thermal conductivity. However, for considerations of lift, drag, and moment coefficients, M_∞ and Re are the dominant similarity parameters.

Reflecting for an instant, we find there may appear to be a conflict in our aerodynamic philosophy. On the one hand, we emphasized that lift, drag, and moments on an aerodynamic shape stem from the detailed pressure and shear stress distributions on the surface and that measurements and/or calculations of these distributions, especially for complex configurations, are not trivial undertakings. On the other hand, Eqs. (1)–(3) indicate that lift, drag, and moments can be quickly obtained from simple formulas. The bridge between these two outlooks is, of course, the lift, drag, and moment coefficients. All the physical complexity of the flow field around an aerodynamic body is implicitly buried in c_l, c_d, and c_m. Before the simple Eqs. (1)–(3) can be used to calculate lift, drag, and moments for an airfoil, wing, or body, the appropriate aerodynamic coefficient must be known. From this point of view, the simplicity of Eqs. (1)–(3) is a bit deceptive. These equations simply shift the forces of aerodynamic rigor from the forces and moments themselves to the appropriate coefficients instead. So we are now led to the following questions: How do we obtain values of c_l, c_d, and c_m for given configurations and how do they vary with α, M_∞, and Re? The answers are introduced in the following paragraphs.

A goal of theoretical aerodynamics is to predict values of c_l, c_d, and c_m from the basic equations and concepts of physical science, some of which were discussed in previous chapters. However, simplifying assumptions are usually necessary to make the mathematics tractable. Therefore, when theoretical results are obtained, they are generally not "exact." The use of high-speed digital computers to solve the governing flow equations is now

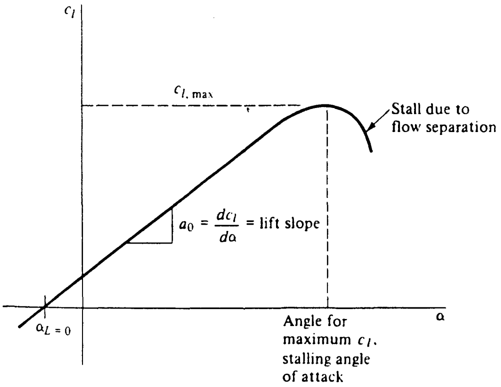
c_l
$c_{l,\max}$
Stall due to flow separation
$a_0 = \frac{dc_l}{d\alpha}$ = lift slope
$\alpha_{L=0}$
Angle for maximum c_l, stalling angle of attack
α

FIG. 7. Sketch of a typical lift curve. [From Anderson (1989).]

bringing us much closer to the accurate calculation of aerodynamic characteristics; however, there are still limitations imposed by the numerical methods themselves, and the storage and speed capacity of current computers is still not sufficient to solve many complex aerodynamic flows. As a result, the practical aerodynamicist has to rely upon direct *experimental* measurements of c_l, c_d, and c_m for specific bodies of interest.

A large bulk of experimental airfoil data was compiled over the years by the National Advisory Committee for Aeronautics (NACA), which was absorbed in the creation of the National Aeronautics and Space Administration (NASA) in 1958. Lift, drag, and moment coefficients were systematically measured for many airfoil shapes in low-speed subsonic wind tunnels. These measurements were carried out on straight, constant-chord wings which completely spanned the tunnel test section from one side wall to the other. In this fashion, the flow essentially "saw" a wing with no wingtips, and the experimental airfoil data were thus obtained for "infinite wings." (The distinction between infinite and finite wings will be made in subsequent sections.)

Let us examine the variation of c_l with α more closely. This variation is sketched in Fig. 7. The experimental data indicate that c_l varies *linearly* with α over a large range of angle of attack. Thin-airfoil theory, which is the subject of advanced books on aerodynamics, also predicts the same type of linear variation. The slope of the linear portion of the lift curve is designated as $a_0 \equiv dc_l/d\alpha \equiv$ lift slope. Note that in Fig. 7, when $\alpha = 0$, there is still a positive value of c_l, that is, there is still some lift even when the airfoil is at zero angle of attack to the flow. This is due to the positive camber of the airfoil. All airfoils with such camber have to be pitched to some negative angle of attack before zero lift is obtained. The value of α when lift is zero is defined as *zero lift angle of attack* $\alpha_{L=0}$ and is illustrated in Fig. 7. Refer again to Fig. 7, at the other extreme: for large values of α, the linearity of the lift curve breaks down. As α is increased beyond a certain value, c_l peaks at some maximum value, $c_{l,\max}$, and then drops precipitously as α is further increased. In this situation, where the lift is rapidly decreasing at high α, the airfoil is *stalled.*

The phenomenon of airfoil stall is of critical importance in airplane design. It is caused by flow separation on the upper surface of the airfoil. This is illustrated in Fig. 8, which again shows the variation of c_l vs α for an airfoil. At point 1 on the linear portion of the lift curve, the flow field over the airfoil is attached to the surface, as pictured in Fig. 8. However, the effect of friction is to slow the airflow near the surface; in the presence of an adverse pressure gradient, there will be a tendency for this viscous boundary layer adjacent to the surface to separate from the surface. As the angle of attack is increased, the adverse pressure gradient on the top surface of the airfoil will become stronger, and at some value of α—the stalling angle of attack—the flow becomes separated from the top surface. When separation occurs, the lift decreases drastically and the drag increases suddenly. This is the picture associated with point 2 in Fig. 8.

The lift curves sketched in Figs. 7 and 8 illustrate the type of variation observed experimentally in the data measured by the NACA. A sample of these data is shown in Figs. 9 and 10 for a particular airfoil denoted as the NACA 2412 airfoil. The code for these curves is given in Fig. 10. Note in Fig. 9 that the lift curves are all virtually linear up to the stall. Also note that c_l vs α is given for three different values of the Reynolds number from 3.1×10^6 to 8.9×10^6. The lift curves for all three values of Re fall on top of each other in the linear region, i.e., Re has little influence on c_l when the flow is attached. However, flow separation is a viscous effect, and Re is a governing parameter for viscous flow. Therefore, it is not surprising that the experimental

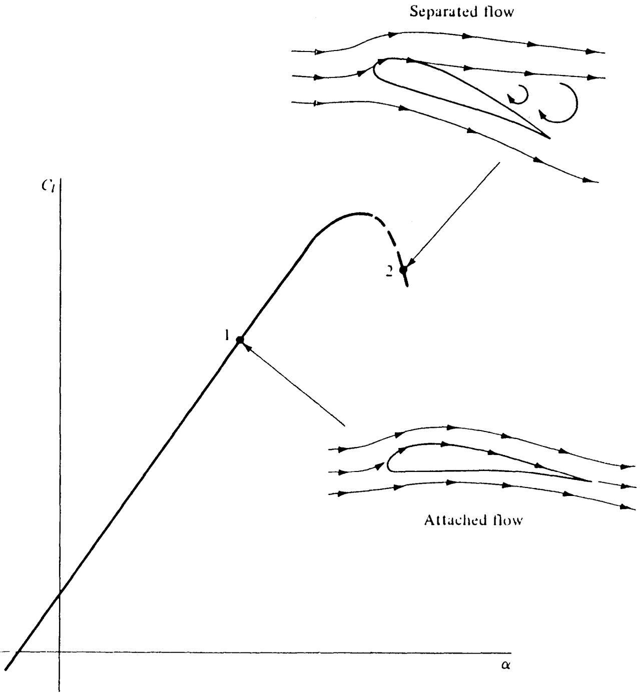

FIG. 8. Flow mechanism associated with stalling. [From Anderson (1989).]

data for $c_{l_{\max}}$ in the stalling region are affected by Re, as can be seen by the slightly different variations of c_l at high α for different values of Re. In fact, these lift curves at different Re answer part of the question posed in Eq. (4): the data represent $c_l = f(\text{Re})$. Again, Re exerts little or no effect on c_l except in the stalling region.

The variation of $c_{m,c/4}$ vs α is also given in Fig. 9. It has only a slight variation with α and is almost completely unaffected by Re. Also note that the values of $c_{m,c/4}$ are slightly negative. By convention, a positive moment is in a clockwise direction; it pitches the airfoil towards larger angles of attack, as shown in Fig. 6. Therefore, for the NACA 2412 airfoil, with $c_{m,c/4}$ negative, the moments are counterclockwise, and the airfoil tends to pitch downward. This is characteristic of all airfoils with positive camber.

The variation of c_d and $c_{m,ac}$ is given versus c_l in Fig. 10. Because c_l varies linearly with α as well, the shapes will be the same. Note that the drag curves have a "bucket" type of shape, with minimum drag occurring at small values of c_l (i.e., minimum drag occurs at small angles of attack). As α goes to large negative or positive values, c_d increases. Also note that c_d is strongly affected by Re, there being a distinct drag curve for each Re. This is to be expected because the drag for a slender aerodynamic shape is mainly skin friction drag, and Re strongly governs skin friction. With regard to $c_{m,\text{ac}}$ the definition of the aerodynamic center is clearly evident; $c_{m,\text{ac}}$ is constant with respect to α. Also, it is insensitive to Re and has a small negative value.

Refer to Eqs. (4)–(6); the airfoil data in Figs. 9 and 10 experimentally provide the variation of c_l, c_d, and c_m with α and Re. The effect of M_∞ on the airfoil coefficients will be discussed later. However, emphasis is made that the data in Figs. 9 and 10 were measured in low-speed subsonic wind tunnels. Hence, the flow was essentially incompressible. Thus, c_l, $c_{m,c/4}$, c_d, and $c_{m,ac}$ given in Figs. 9 and 10 are low speed, incompressible flow values. The final question associated with Eqs. (4)–(6), namely, how do the aerodynamic coefficients vary with Mach number, M_∞, requires some extended discussion, as given in the following paragraphs.

To understand the variation of lift and drag coefficients with Mach number, we first introduce the concept of the *critical* Mach number, denoted by M_{cr}. Consider the flow over an airfoil. As the gas expands around the top surface near the leading edge, the velocity and hence the Mach number will increase rapidly. Indeed, there are regions on the air-

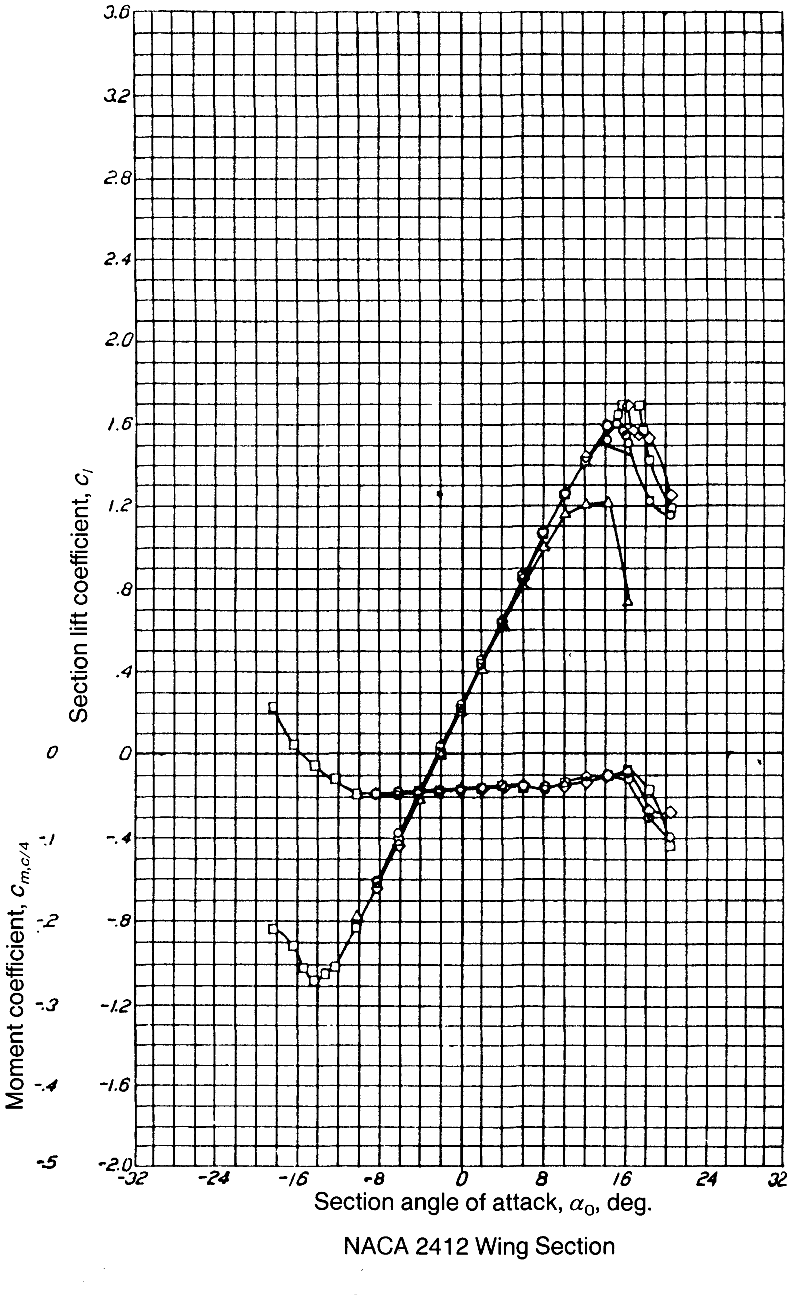

FIG. 9. Lift and moment coefficient curves for an NACA 2412 Wing Section. The moment coefficient is about the quarter-chord. [From Abbott and von Doenhoff (1949).]

foil surface where the local Mach number is greater than M_∞. Imagine that we put a given airfoil in a wind tunnel where $M_\infty = 0.3$ and that we observe the peak local Mach number on the top surface of the airfoil to be 0.435. This is sketched in Fig. 11(a). Imagine that we now increase M_∞ to 0.5; the peak local Mach number will correspondingly increase to 0.772, as shown in Fig. 11(b). If we further increase M_∞ to a value of 0.61, we observe that the peak local Mach number is 1.0, locally sonic flow on the surface of the airfoil. This is sketched in Fig. 11(c). Note that the flow over an airfoil can locally be sonic (or higher), even though the freestream Mach number is subsonic. By definition, that freestream Mach

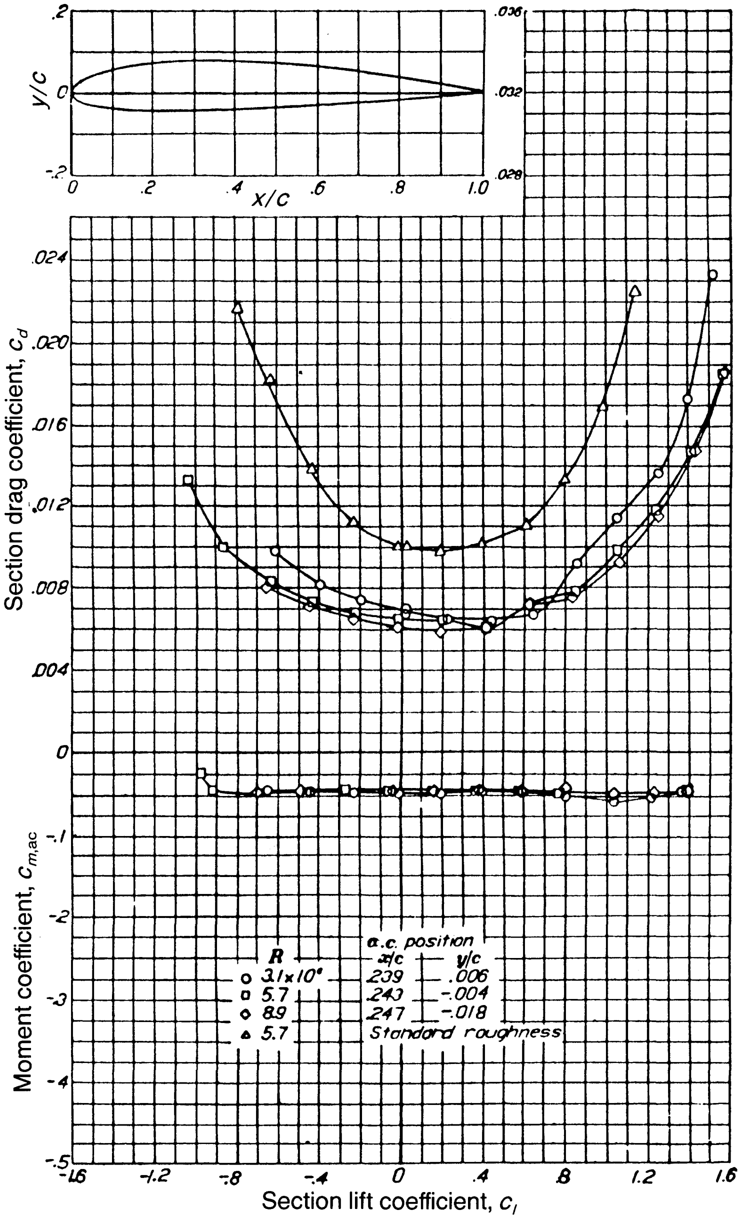

NACA 2412 Wing Section (*Continued*)

FIG. 10. Drag and moment coefficients for an NACA 2412 airfoil. The moment coefficient is about the aerodynamic center. [From Anderson (1989).]

number at which sonic flow is first obtained somewhere on the airfoil surface is called the *critical Mach number* of the airfoil. In the above example, the critical Mach number M_{cr} for the airfoil is 0.61. M_{cr} is an important quantity, because at some freestream Mach number above M_{cr} the airfoil will experience a dramatic increase in drag. Figure 12 sketches the variation of c_d with M_∞. At low Mach numbers, less than M_{cr}, c_d is virtually constant and is equal to its low-speed value given in Fig. 10. The flow field about the airfoil for this condition (say, point *a* in Fig. 12) is noted in Fig. 13(a), where $M < 1$ everywhere in the flow. If M_∞ is increased slightly above M_{cr}, a "bubble" of supersonic flow will occur

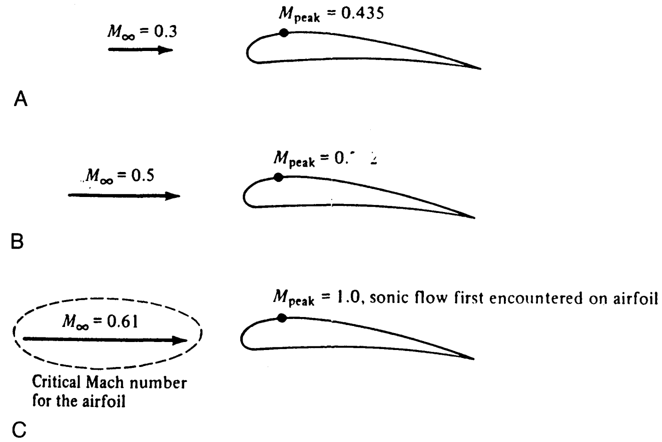

FIG. 11. Illustration of critical Mach number. [From Anderson (1989).]

surrounding the minimum pressure point as shown in Fig. 13(b). Correspondingly, c_d will still remain reasonably low, as indicated by point *b* in Fig. 12. However, if M_∞ is still further increased, a very sudden and dramatic rise in the drag coefficient will be observed, as noted by point *c* in Fig. 12. Here, shock waves suddenly appear in the flow, as sketched in Fig. 13(c). The shock waves themselves are dissipative phenomena, which result in an increase in drag on the airfoil. But in addition, the sharp pressure increase across the shock waves creates a strong adverse pressure gradient, hence causing the flow to separate from the surface. Such flow separation can create substantial increases in drag. Thus, the sharp increase in c_d shown in Fig. 12 is a combined effect of shock waves and flow separation. The *freestream* Mach number at which c_d begins to increase rapidly is defined as the *drag-divergence* Mach number and is noted in Fig. 12. Note that

$$M_{\mathrm{cr}} < M_{\text{drag-divergence}} < 1.0.$$

The shock pattern sketched in Fig. 13(c) is characteristic of a flight regime called transonic. When $0.8 \leqslant M_\infty \leqslant 1.2$, the flow is generally designated as transonic flow and it is characterized by some very complex effects only hinted at in Fig. 13(c). To reinforce these comments, Fig. 14 shows the variation of both c_l and c_d as a function of Mach number with angle of attack as a parameter. The airfoil is a standard NACA 2315 airfoil. Figure 14, which shows actual wind-tunnel data, illustrates the massive transonic-flow effects on both lift and drag coefficients. The analysis of transonic flows has been one of the major challenges in modern aerodynamics. Only in recent years, since about 1970, have com-

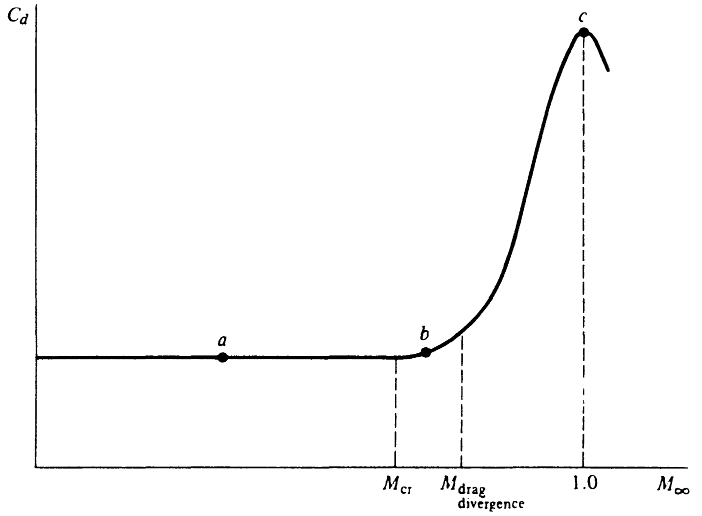

FIG. 12. Variation of drag coefficient with Mach number. [From Anderson (1989).]

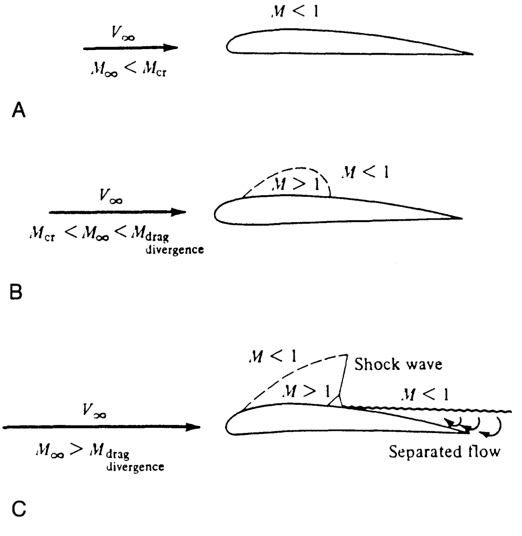

FIG. 13. Physical mechanism of drag divergence. (a) Flow field associated with point *a* in Fig. 12. (b) Flow field associated with point *b* in Fig. 12. (c) Flow field associated with point *c* in Fig. 12. [From Anderson (1989).]

puter solutions for transonic flows over airfoils come into practical use; these numerical solutions are still in a state of development and improvement.

The airfoils on modern subsonic jet aircraft, such as the McDonnell-Douglas DC-10 wide-body transport, are relatively thin profiles designed to increase the drag-divergence Mach number. In fact, NASA (mainly under the direction of Richard Whitcomb at the Langley Research Center) has developed a new *supercritical* airfoil, designed to place the drag-divergence Mach number extremely close to 1.0. Supercritical airfoils are of the general shape shown in Fig. 15. Here, the maximum thickness is designed close to the trailing edge. The flow over the airfoil is largely supersonic, and the airfoil shape is designed to discourage the formation of shock waves.

These considerations of M_{cr} and drag-divergence Mach number are important to subsonic transport design. The cruising

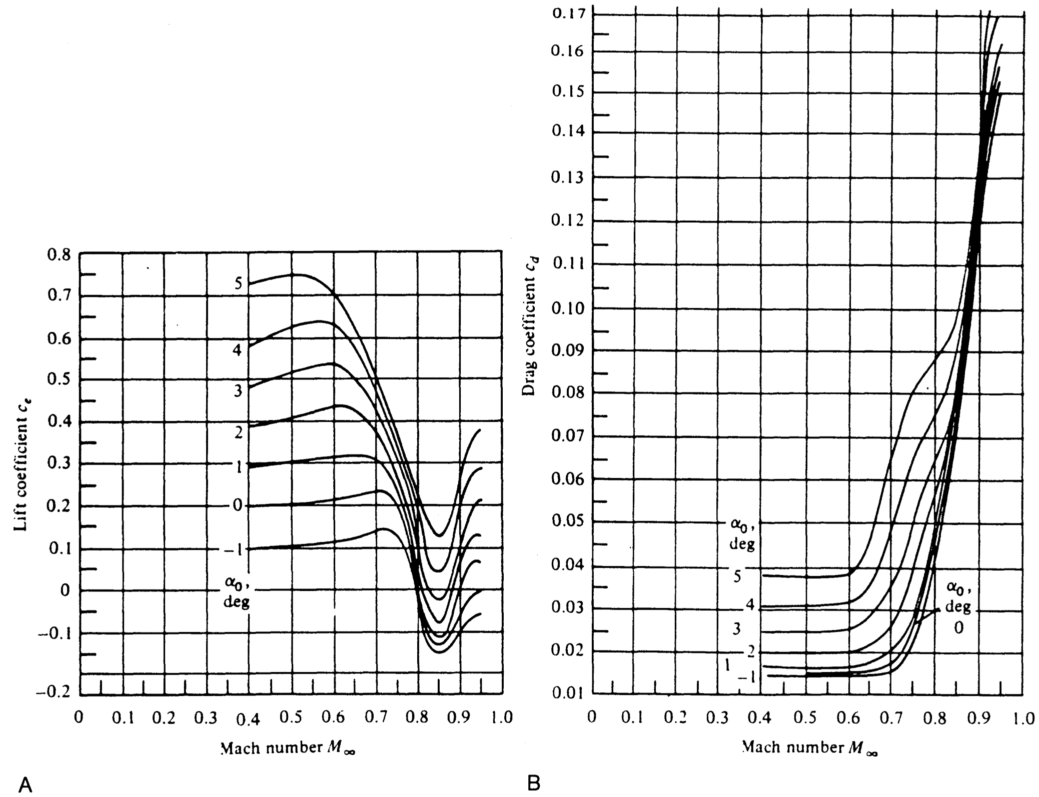

FIG. 14. Variation of (a) lift coefficient and (b) drag coefficient vs Mach number with angle of attack as a parameter of an NACA 2315 airfoil. (Wind-tunnel measurements at the NACA Langley Memorial Laboratory.)

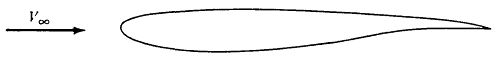

FIG. 15. Shape of a typical supercritical airfoil. [From Anderson (1989).]

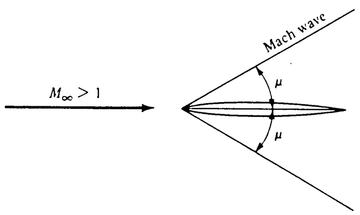

FIG. 16. Mach waves on a needlelike body. [From Anderson (1989).]

speeds of such airplanes can be increased by incorporating airfoils with high values of $M_{\text{drag-divergence}}$. Alternatively, in the spirit of energy conservation, supercritical airfoils should allow less drag, hence less fuel consumption, at the same cruising speed. In either case, the advantages are obvious, and this helps to explain why there has been a modern resurgence of interest in high-speed airfoil design.

As a final note in this section, the next time you have an opportunity to fly in a jet airliner and the sun is directly overhead (around noon and early afternoon), look out along the span of the wing. Due to the refraction of light waves through shock waves, you can sometimes see with the naked eye the transonic shock waves dancing about on the wing surface.

The preceding discussion has highlighted the variation of c_d with M_∞, where M_∞ ranges from low values near zero up to the transonic range around Mach one. In regard to the variation of lift coefficient with M_∞ for this same Mach number range, Fig. 14(a) gives typical results. Indeed, an approximate formula for the Mach number variation of c_l is given by the *Prandtl–Glauert rule,*

$$c_l = c_{l_0} / \sqrt{1 - M_\infty^2},$$

where c_{l_0} is the low speed, incompressible flow value of the airfoil lift coefficient. The moment coefficient also obeys the Prandtl–Glauert rule,

$$c_m = c_{m_0} / \sqrt{1 - M_\infty^2}.$$

The Prandtl–Glauert rule is based on subsonic thin airfoil theory; it is an approximate result that holds reasonably well for freestream Mach numbers less than 0.7. The Prandtl–Glauert rule is an example of a *compressibility correction,* intended to modify low speed, incompressible flow data to take into account the effects of compressibility at higher subsonic speeds. For more accuracy near Mach one, other, more sophisticated, compressibility corrections are available.

When the freestream Mach number is supersonic, shock waves will occur around the aerodynamic body; these shock waves create an additional source of drag called *wave drag.* Let us examine the source of this wave drag more closely.

A very thin object (such as a thin needle) moving at supersonic speeds ($M_\infty > 1$) creates a very weak disturbance in the flow, limited to a *Mach wave,* which makes the Mach angle μ with respect to the freestream direction. A Mach wave is illustrated in Fig. 16, where the Mach angle μ is defined as

$$\mu = \arcsin(1/M_\infty).$$

On the other hand, an object with some reasonable thickness, such as the wedge shown in Fig. 17, moving at supersonic speeds will create a strong disturbance called a *shock wave.* The shock wave will be inclined at an oblique angle β, where $\beta > \mu$, as shown in Fig. 17. As the flow moves across the oblique shock wave, the pressure, temperature, and density increase, and the velocity and Mach number decrease.

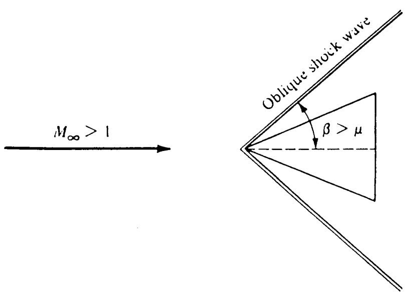

FIG. 17. Oblique shock waves on a wedge-type body. [From Anderson (1989).]

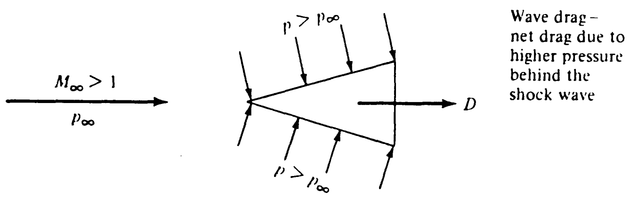
$p > p_\infty$
$M_\infty > 1$
p_∞
D
$p > p_\infty$
Wave drag – net drag due to higher pressure behind the shock wave

FIG. 18. Pressure distribution on a wedge at supersonic speeds; origin of wave drag. [From Anderson (1989).]

Consider now the pressure on the surface of the wedge, as sketched in Fig. 18. Since p increases across the oblique shock wave, then at the wedge surface, $p > p_\infty$. Since the pressure acts normal to the surface and the surface itself is inclined to the relative wind, there will be a net drag produced on the wedge, as seen by simple inspection of Fig. 18. This drag is called *wave drag*, because it is inherently due to the pressure increase across the shock wave.

In order to minimize the strength of the shock wave, all supersonic airfoil profiles are thin, with relatively sharp leading edges. (The leading edge of the Lockheed F-104 supersonic fighter is almost razor thin.) Let us approximate a thin supersonic airfoil by the flat plate illustrated in Fig. 19. The flat plate is inclined at a small angle of attack α to the supersonic freestream. On the top surface of the plate, the flow field is turned away from the freestream through an *expansion wave* at the leading edge; an expansion wave is a fan-shaped region through which the pressure decreases. At the trailing edge on the top side, the flow is turned back into the freestream direction through an oblique shock wave. On the bottom surface of the plate, the flow is turned into the freestream, causing an oblique shock wave with an increase in pressure. At the trailing edge, the flow is turned back to the freestream direction through an expansion wave. The expansion and shock waves at the leading edge result in a surface pressure distribution where the pressure on the top surface is less than p_∞, whereas the pressure on the bottom surface is greater than p_∞. The net effect is an aerodynamic force normal to the plate. The components of this force perpendicular and parallel to the relative wind are the lift and supersonic wave drag, respectively. Approximate relations for lift and drag coefficients are

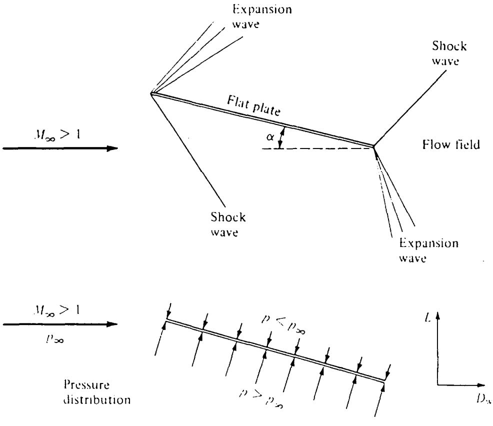

FIG. 19. Flow field and pressure distribution for a flat plate at angle of attack in supersonic flow. There is a net lift and drag due to the pressure distribution set up by the shock and expansion waves. [From Anderson (1989).]

$$c_l = 4\alpha/(M_\infty^2 - 1)^{1/2} \tag{7}$$

and

$$c_{d,w} = 4\alpha^2/(M_\infty^2 - 1)^{1/2}. \tag{8}$$

A subscript w has been added to the drag coefficient to emphasize that it is the wave drag coefficient. These are approximate expressions, useful for thin airfoils at small to moderate angles of attack in supersonic flow. Note that as M_∞ increases, both c_l and c_d decrease. This is not to say that the lift and drag forces themselves decrease with M_∞. Quite the contrary: for any flight regime, as the flight velocity increases, L and D usually increase because the dynamic pressure $q_\infty = \frac{1}{2}\rho V_\infty^2$ increases. In the supersonic regime, L and D increase with velocity, even though c_l and $c_{d,w}$ decrease with M_∞.

In summary, this section has discussed the definition and meaning of lift, drag, and moment coefficients, as well as their variations with angle of attack, Reynolds number, and Mach number. Of particular importance is the variation of drag coefficient with Mach number. To amplify on the aspects of drag, we can write the total drag of an airfoil as the sum of three contributions:

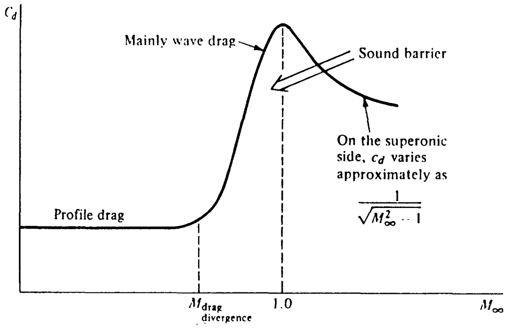

FIG. 20. Variation of drag coefficient with Mach number for subsonic and supersonic speeds. [From Anderson (1989).]

$$D_f + D_p + D_w,$$

where D=total drag on the airfoil, D_f=skin friction drag, D_p=pressure drag due to flow separation, and D_w=wave drag (present only at transonic and supersonic speeds; zero for subsonic speeds below the drag-divergence Mach number).

In terms of the drag coefficients, we can write

$$c_d = c_{d,f} + c_{d,p} + c_{d,w},$$

where c_d, $c_{d,f}$, $c_{d,p}$, and $c_{d,w}$ are the total drag, skin friction drag, pressure drag, and wave drag coefficients, respectively. The sum $c_{d,f} + c_{d,p}$ is called the *profile drag coefficient*; this is the quantity that is given by the data in Fig. 10. The profile drag coefficient is relatively constant with M_∞ at subsonic speeds.

The variation of c_d with M_∞ from incompressible to supersonic speeds is sketched in Fig. 20. It is important to note the qualitative variation of this curve. For M_∞ ranging from zero to drag divergence, c_d is relatively constant; it consists entirely of profile drag. For M_∞ from drag divergence to slightly above 1, the value of c_d skyrockets; indeed, the peak value of c_d around $M_\infty = 1$ can be an order of magnitude larger than the profile drag itself. This large increase in c_d is due to wave drag associated with the presence of shock waves. For supersonic Mach numbers, c_d decreases approximately as $(M_\infty^2 - 1)^{-1/2}$.

The large increase in the drag coefficient near Mach 1 gave rise to the term "sound barrier" in the 1940s. There was a camp of professionals who at that time felt that the sound barrier could not be pierced, that we could not fly faster than the speed of sound. Certainly, a glance at Eq. (8) for wave drag in supersonic flow would hint that the drag coefficient might become infinitely large as M_∞ approaches 1. However, such reasoning is an example of a common pitfall in science and engineering, namely, the application of equations outside their ranges of validity. Equation (8) is not valid in the transonic range near $M_\infty = 1$. Moreover, remember that nature abhors infinities. In real life, c_d does not become infinitely large. To get past the sound barrier, all that is needed (in principle) is an engine with enough thrust to overcome the high (but finite) drag.

1.3 Finite Wing Properties: Induced Drag

The airfoil properties discussed in the previous section apply to a wing which stretches to infinity along the span direction, i.e., there are no wingtips. Hence, airfoil data is frequently called *infinite wing* data. On the other hand, all real airplanes have wings that are finite in extent—so-called *finite wings*. The most important geometric parameter for a finite wing is the *aspect ratio*, AR, defined as

$$\text{Aspect ratio} = \text{AR} \equiv b^2/S,$$

where b is the *wing span*, i.e., the distance between the two wingtips, and S is the planform area (projected top-view area) of the wing.

The flow field about a finite wing is three-dimensional and is therefore inherently different from the two-dimensional flow about an infinite wing. As a result, the lift, drag, and moment coefficients for a finite wing with a given airfoil shape at a given α are different from the lift, drag, and moment coefficients for an infinite wing with the same airfoil shape at the same α. For this reason, the aerodynamic coefficients for a finite wing are designated by capital letters, C_L, C_D, C_M; this is in contrast to those for an infinite wing, which we have been designating as c_l, c_d, and c_m. Note that the data in Figs. 9 and 10 are for infinite (two-dimensional) wings, i.e., the data for c_l, c_d, and c_m.

The fundamental difference between flows over finite wings as opposed to infinite wings can be seen as follows. Consider the front view of a finite wing as sketched in Fig. 21(a). If the wing has lift, then obviously the average pressure over the bottom surface is greater than that over the top surface. Consequently, there is some tendency for the air to

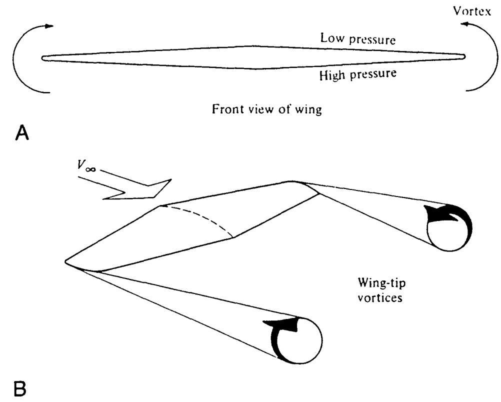

FIG. 21. Origin of wingtip vortices on a finite wing. [From Anderson (1989).]

"leak," or flow, around the wingtips from the high- to the low-pressure sides, as shown in Fig. 21(a). This flow establishes a circulatory motion which trails downstream of the wing. The trailing circular motion is called a *vortex.* There is a major trailing vortex from each wingtip, as sketched in Fig. 21(b).

These wingtip vortices downstream of the wing induce a small downward component of air velocity in the neighborhood of the wing itself. This can be seen intuitively from Fig. 21(b); the two wingtip vortices tend to drag the surrounding air around with them, and this secondary movement induces a small velocity component in the downward direction at the wing. This downward component is called *downwash,* given by the symbol w.

An effect of downwash can be seen in Fig. 22. As usual, V_∞ designates the relative wind. However, in the immediate vicinity of the wing, V_∞ and w add vectorally to produce a "local" relative wind which is canted downward from the original direction of V_∞. This has a couple of consequences:

1. The angle of attack of the airfoil sections of the wing is effectively reduced in comparison to the angle of attack of the wing referenced to V_∞.
2. There is an increase in the drag. The increase is called *induced drag,* which has at least three physical interpretations. First, the wingtip vortices simply alter the flow field about the wing in such a fashion as to change the surface pressure distributions in the direction of increased drag. An alternative explanation is that, because the local relative wind is canted downward (see Fig. 22), the lift vector itself is "tilted back"; hence it contributes a certain component of force parallel to V_∞, that is, a drag force. A third physical explanation of the source of induced drag is that the wingtip vortices contain a certain amount of rotational kinetic energy. This energy has to come from somewhere; indeed, it is supplied by the aircraft propulsion system, where extra power has to be added to overcome the extra increment in drag due to induced drag. All three of these outlooks of the physical mechanism of induced drag are synonymous.

The induced drag coefficient is defined as

$$C_{D,i}=D_i/q_\infty S,$$

where D_i is the induced drag. In turn, $C_{D,i}$ is related to the lift coefficient as

$$C_{D,i}=C_L^2/\pi e\mathrm{AR}. \tag{9}$$

The induced drag is the cost that is paid for the generation of lift; when the lift is zero, then D_i is zero. In Eq. (9), AR is the aspect ra-

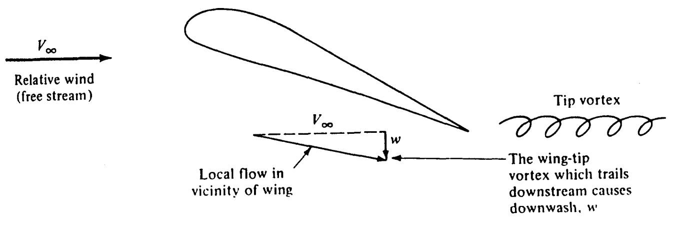
V_∞
Relative wind (free stream)
V_∞
w
Local flow in vicinity of wing
Tip vortex
The wing-tip vortex which trails downstream causes downwash, w

FIG. 22. The origin of downwash. [From Anderson (1989).]

tio, and e is the span efficiency factor determined by the way that lift is distributed over the span, which depends on the planform shape of the wing and is usually less than unity; at best it is equal to unity. A value of e for a typical wing is about 0.9. Note that AR plays a strong role in the induced drag; a higher AR leads to lower $C_{D,i}$.

The lift slope for a finite wing is lower than that for the airfoil section of the wing. If a_0 represents the airfoil lift slope,

$$a_0 = dc_l/d\alpha,$$

and a represents the lift slope of the finite wing,

$$a = dC_L/d\alpha,$$

then

$$a = \frac{a_0}{1 + 57.3a_0/\pi e_1 \mathrm{AR}},$$

where a is given in terms of the change in lift coefficient *per degree*, and e_1 is an efficiency factor slight less than unity; the value of e_1 depends on the planform shape of the wing, and is usually about the same in value as e given in Eq. (9).

The total drag coefficient for a finite wing, C_D, is the sum of the skin friction drag as well as pressure drag due to flow separation, both represented by the *profile drag*, c_d, and the induced drag, C_{D_i}. Hence,

$$C_D = c_d + C_L^2/\pi e \mathrm{AR}. \tag{10}$$

If the wing is in a freestream near Mach one, then c_d also contains the wave drag.

1.4 Swept Wings

Many high-speed aircraft, both subsonic and supersonic, have swept wings. The reason for the swept wing is the same in both cases, namely, to reduce drag. However, the aerodynamic means for reducing drag is slightly different for the subsonic as compared to the supersonic cases.

We will first consider subsonic flight. Consider the planview of a straight wing, as sketched in Fig. 23(a). Assume this wing has an airfoil cross section with a critical Mach number $M_{cr} = 0.7$. (Remember from our previous discussions that for M_∞ slightly above M_{cr}, there is a large increase in drag. Hence, it is desirable to increase M_{cr} as much as possible in high-speed subsonic airplane design.) Now assume that we sweep the wing back through an angle of, say, 30°, as shown in Fig. 23(b). The airfoil, which still has a value of $M_{cr} = 0.7$, now "sees" essentially only the component of the flow normal to the leading edge of the wing; i.e., the aerodynamic properties of the local section of the swept wing are governed mainly by the flow normal to the leading edge. Hence, if M_∞ is the freestream Mach number, the airfoil in Fig. 23(b) is effectively seeing a smaller Mach number, M_∞ cos 30°. As a result, the actual freestream Mach number can be increased *above* 0.7 before critical phenomena on the airfoil are encountered. In fact, we could expect that the critical Mach number for the *swept wing itself* would be as high as 0.7/cos 30° = 0.808. Therefore, we see the main function of a swept wing. *By sweeping the wings of subsonic aircraft, drag divergence is delayed to higher Mach numbers.*

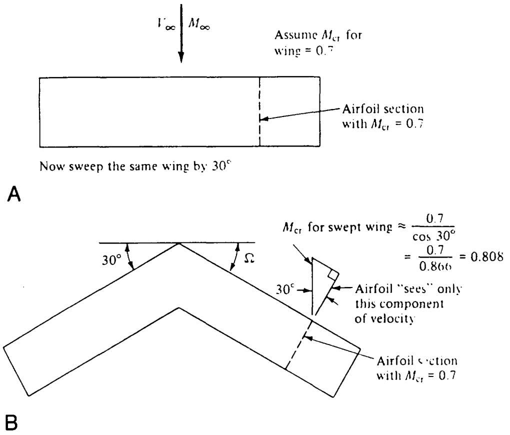
V_∞
M_∞
Assume M_{cr} for wing = 0.7
Airfoil section with M_{cr} = 0.7
Now sweep the same wing by 30°
A
30°
Ω
M_{cr} for swept wing ≈ $\frac{0.7}{\cos 30°}$ = $\frac{0.7}{0.866}$ = 0.808
30°
Airfoil "sees" only this component of velocity
Airfoil section with M_{cr} = 0.7
B

FIG. 23. Effect of a swept wing on critical Mach number. [From Anderson (1989).]

In real life, the flow over the swept wing sketched in Fig. 23(b) is a fairly complex three-dimensional flow, and to say that the airfoil sees only the component normal to the leading edge is a sweeping simplification. However, it leads to a good rule of thumb. If Ω is the sweep angle, as shown in Fig. 23(b), the actual critical Mach number for the swept wing is bracketed by

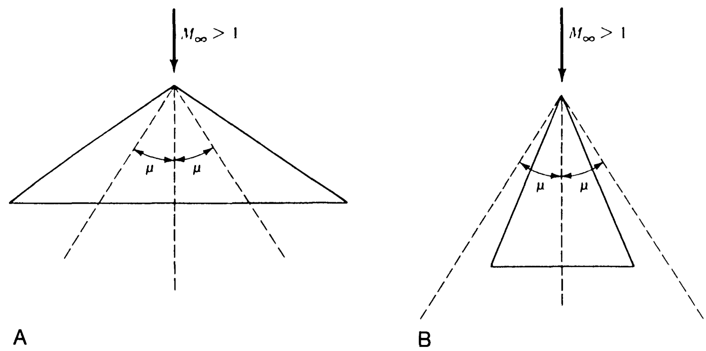

FIG. 24. Swept wings for supersonic flow. (a) Wing swept outside the Mach cone. (b) Wing swept inside the Mach cone. [From Anderson (1989).]

$$\begin{aligned}M_{\text{cr}} &\text{ for airfoil}\\ &< \text{actual } M_{\text{cr}} \text{ for swept wing}\\ &< (M_{\text{cr}} \text{ for airfoil})/\cos\Omega\,.\end{aligned}$$

For supersonic flight, swept wings are also advantageous, but not quite from the same point of view as described above for subsonic flow. Consider the two swept wings sketched in Fig. 24. For a given $M_\infty > 1$, there is a Mach cone with vertex angle μ, equal to the Mach angle. If the leading edge of a swept wing is *outside* the Mach cone, as shown in Fig. 24(a), the component of the Mach number normal to the leading edge is supersonic. As a result, a fairly strong oblique shock wave will be created by the wing itself, with an attendant large wave drag. On the other hand, if the leading edge of the swept wing is inside the Mach cone, as shown in Fig. 24(b), the component of the Mach number normal to the leading edge is subsonic. As a result, the wave drag produced by the wing is less. Therefore, the advantage of sweeping the wings for supersonic flow is in general to obtain a decrease in wave drag, and if the wing is swept inside the Mach cone, a considerable decrease can be obtained.

The quantitative effects of maximum thickness and wing sweep on wave-drag coefficient are shown in Fig. 25(a) and 25(b), respectively. For all cases, the wing aspect ratio is 3.5 and the taper ratio (tip to root chord) is 0.2. Clearly, thin wings with large angles of sweepback have the smallest wave drag. The effect of wing sweep on the lift-to-drag ratio is shown in Fig. 26. Note that for a subsonic airplane, sweeping the wing reduces C_L more than C_D and hence gives a lower L/D ratio. In contrast, for a supersonic airplane at Mach 2.2, sweeping the wing reduces both C_L and C_D correspondingly and has virtually no effect on L/D.

As a final note, we might observe that not all supersonic aircraft have swept wings. In supersonic flight, wave drag can also be reduced by using stubby, very *low*-aspect-ratio, straight wings with thin, sharp airfoils (such as on the Lockheed F-104). In supersonic

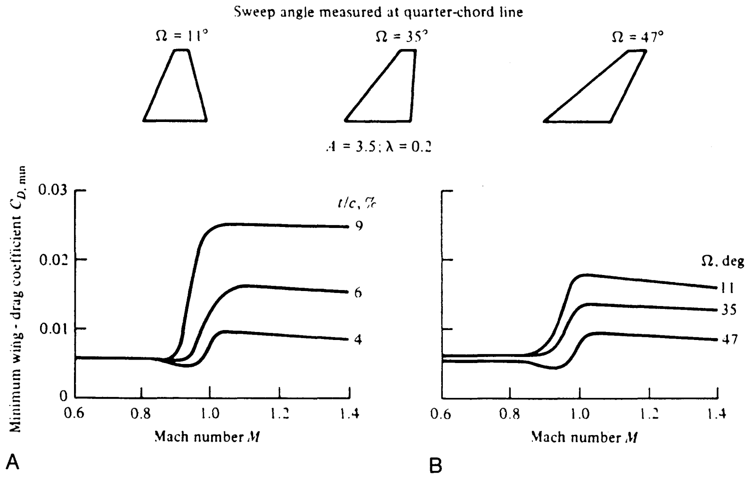

FIG. 25. Sketch of the variation of minimum wing-drag coefficient vs Mach number with (a) wing thickness as a parameter (Ω=47°) and (b) wing sweepback angle as a parameter (t/c=4%). [From Loftin (1985).]

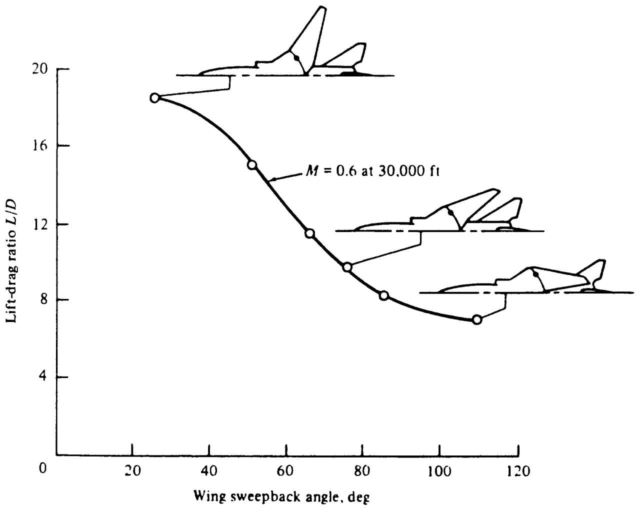

FIG. 26. Variation of lift-to-drag ratio with wing sweep. Wind-tunnel measurements at the NASA Langley Research Center. [From Loftin (1985).]

flight, wave drag is generally much larger, hence much more important, than induced drag. Thus, low-aspect-ratio wings for supersonic aircraft reduce wave drag much more than they increase induced drag. This results in a net decrease in total drag. Thus, low-aspect-ratio wings are efficient for supersonic flight, whereas high-aspect-ratio wings are efficient for subsonic flight.

1.5 Flaps—A Mechanism for High Lift

An airplane normally encounters its lowest flight velocities at takeoff or landing, two periods that are most critical for aircraft safety. The slowest speed at which an airplane can fly in straight and level flight is defined as the *stalling speed* V_{stall}. Hence, the calculation of V_{stall}, as well as aerodynamic methods of making V_{stall} as small as possible, are of vital importance.

The stalling velocity is readily obtained in terms of the maximum lift coefficient, as follows. From the definition of C_L,

$$L=q_\infty SC_L=\tfrac{1}{2}\rho_\infty V_\infty^2 SC_L.$$

Thus

$$V_\infty=\sqrt{2L/\rho_\infty SC_L}.$$

In steady, level flight, the lift is just sufficient to support the weight W of the aircraft, that is, $L=W$. Thus

$$V_\infty=\sqrt{2W/\rho_\infty SC_L}.$$

Examining this equation, for an airplane of given weight and size at a given altitude, we find the only recourse to minimize V_∞ is to maximize C_L. Hence, stalling speed corresponds to the angle of attack that produces $C_{L,\max}$:

$$V_{\text{stall}}=\sqrt{2W/\rho_\infty SC_{L,\max}}\,.$$

In order to decrease V_{stall}, $C_{L,\max}$ must be increased. However, for a wing with a given shape, $C_{L,\max}$ is fixed by nature, i.e., the lift properties of an airfoil, including maximum lift, depend on the physics of the flow over the airfoil. In order to assist nature, the lifting properties of a given airfoil can be greatly enhanced by the use of "artificial" high-lift devices. The most common of these devices is the simple flap at the trailing edge of the wing, as sketched in Fig. 27. When the flap is rotated downward, the camber of the airfoil is effectively increased, with a consequent increase in $C_{L,\max}$ for the wing. At the same time, the zero-lift angle of attack is shifted to a more negative value. These trends are illustrated in Fig. 28, which compares the variation of C_L with α for a wing with and without flaps.

The increase in $C_{L,\max}$ due to flaps can be dramatic. If the flap is designed not only to

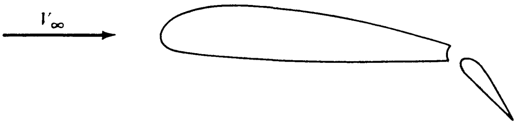

FIG. 27. Illustration of a single flap. [From Anderson (1989).]

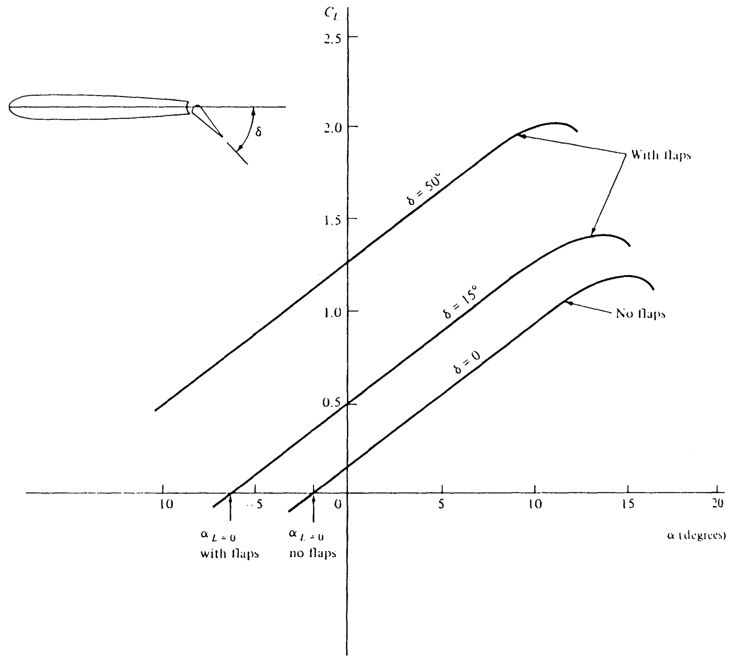

FIG. 28. Illustration of the effect of flaps on the lift curve. The numbers shown are typical of a modern medium-range jet transport. [From Anderson (1989).]

rotate downward, but also to translate rearward so as to increase the effective wing area, $C_{L,\max}$ can be increased by approximately a factor of 2. If additional high-lift devices are used, such as slats at the leading edge, slots in the surface, or mechanical means of boundary layer control, then $C_{L,\max}$ can sometimes be increased by a factor of 3 or more, as shown in Fig. 29.

2. AIRPLANE PERFORMANCE

Airplane performance is the study of such questions as: How fast can the airplane fly? How high can it fly? How long can it stay in the air starting with a full fuel load (endurance)? How far can it fly starting with a full fuel load (range)? To answer all these questions, the basic aerodynamics of the airplane are reflected through one important relation, similar to that developed in Sec. 1.3 for a finite wing, namely, the *drag polar* given by

$$C_D = C_{D,0} + C_L^2/\pi e \mathrm{AR},$$

where C_D and C_L are the total drag and lift coefficients, respectively, for the entire airplane, $C_{D,0}$ is the *zero-lift* drag coefficient (sometimes called *parasite drag at zero lift*) due to skin friction and pressure drag due to flow separation (sometimes called "form drag") acting over the entire airplane when the airplane is at zero lift, and $C_L^2/\pi e \mathrm{AR}$ is the *drag due to lift*, which is composed of the induced drag

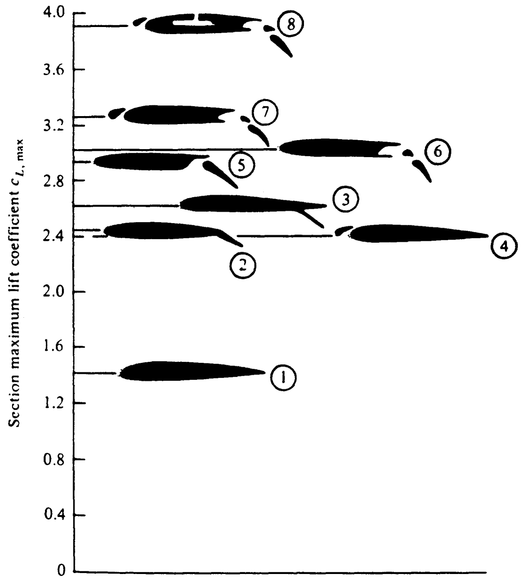

FIG. 29. Typical values of airfoil maximum lift coefficient for various types of high-lift devices: (1) airfoil only, (2) plain flap, (3) split flap, (4) leading edge slat, (5) single-slotted flap, (6) double slotted flap, (7) double-slotted flap in combination with a leading-edge slat, (8) addition of boundary-layer suction at the top of the airfoil. [From Loftin (1985).]

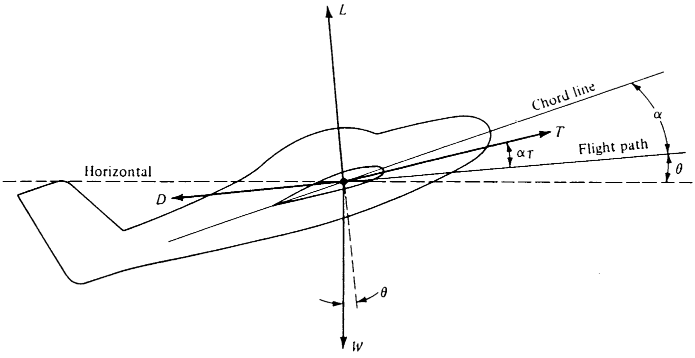

FIG. 30. Force diagram for an airplane in flight. [From Anderson (1989).]

plus the increased parasite drag when the airplane is at the higher angles of attack necessary to generate the required lift. In the above drag polar, AR is the conventional aspect ratio defined in Sec. 1.3, and e now represents the Oswald efficiency factor, different from the span efficiency factor defined in Sec. 1.3. A value of e for a typical airplane is about 0.6.

2.1 Equations of Motion

Consider an airplane in flight, as sketched in Fig. 30. The flight path (direction of motion of the airplane) is included at an angle θ with respect to the horizontal. The flight path direction and the relative wind are along the same line. The mean chord line is at a geometric angle of attack α with respect to the flight path direction. There are four physical forces acting on the airplane:

1. Lift L, which is perpendicular to the flight path direction.
2. Drag D, which is parallel to the flight path direction.
3. Weight W, which acts vertically toward the center of the earth (and hence is inclined at the angle θ with respect to the lift direction).
4. Thrust T, which in general is inclined at the angle α_T with respect to the flight path direction.

The flight path shown in Fig. 30 is drawn as a straight line. This is the picture we see by focusing locally on the airplane itself. However, if we stand back and take a wider view of the space in which the airplane is traveling, the flight path is generally curved. This is obviously true if the airplane is maneuvering; even if the airplane is flying "straight and level" with respect to the ground, it is still executing a curved flight path with a radius of curvature equal to the altitude h plus the radius of the Earth.

Summing forces parallel and perpendicular to the flight path in Fig. 30, and applying Newton's second law, we obtain

$$T \cos\alpha_T - D - W \sin\theta = m(dV/dt),$$

$$L + T \sin\alpha_T - W \cos\theta = m(V^2/r_c).$$

These are the *equations of motion* for an airplane in translational flight.

The above equations describe the general two-dimensional translational motion of an airplane in accelerated flight. However, in this discussion we are interested in a specialized application of these equations, namely, the case where the acceleration is zero. The performance of an airplane for such unaccelerated flight conditions is called *static performance.* This may, at first thought, seem unduly restrictive; however, static performance analyses lead to reasonable calculations of maximum velocity, maximum rate of climb, maximum range, etc.— parameters of vital interest in airplane design and operation.

With this in mind, consider level, unaccelerated flight. Referring to Fig. 30, level flight means that the flight path is along the horizontal; that is, $\theta = 0$. Unaccelerated flight means that the right-hand sides of the above equations are zero. Therefore, these equations reduce to

$$T \cos\alpha_T = D,$$

$$L + T \sin\alpha_T = W.$$

Furthermore, for most conventional airplanes, α_T is small enough that $\cos\alpha_T \approx 0$. Thus,

$$T = D,$$

$$L = W.$$

These are the equations of motion for level, unaccelerated flight. They can also be obtained directly from Fig. 30, by inspection. In level, unaccelerated flight, the aerodynamic drag is balanced by the thrust of the engine, and the aerodynamic lift is balanced by the weight of the airplane.

2.2 Maximum Velocity

The maximum velocity of an airplane at a given altitude occurs when the maximum thrust of the engine (called maximum thrust available) equals the thrust required to overcome the drag of the airplane at the maximum velocity. The thrust required (which is simply associated with the drag of the airframe) is

$$T_R = D = q_\infty S C_D.$$

The calculation of T_R for a given airplane at a given altitude is made as follows:

1. Choose a value of V_∞.
2. Calculate C_L from
   $$C_L = W/\tfrac{1}{2}\rho_\infty V_\infty^2 S,$$
   where W, ρ_∞ and S are known.
3. Calculate C_D from the drag polar
   $$C_D = C_{D,0} + C_L^2/\pi e \mathrm{AR}$$
   where $C_{D,0}$, e, and AR are known.
4. Calculate T_R from
   $$T_R = q_\infty S C_D.$$

The thrust required will vary with velocity. It can easily be shown that minimum T_R occurs when the airplane is flying at a velocity such that the lift-to-drag ratio, L/D, is a maximum. A typical variation of T_R versus velocity for a light, general aviation airplane is shown in Fig. 31. The velocity at which T_R equals the maximum thrust available (T_A) from the engine is the maximum velocity of the airplane.

A related concept is that of power required and power available, defined respectively as

$$P_R = T_R V_\infty,$$

$$P_A = T_A V_\infty.$$

The variation of P_R versus velocity is qualitatively similar to the T_R curve shown in Fig. 31.

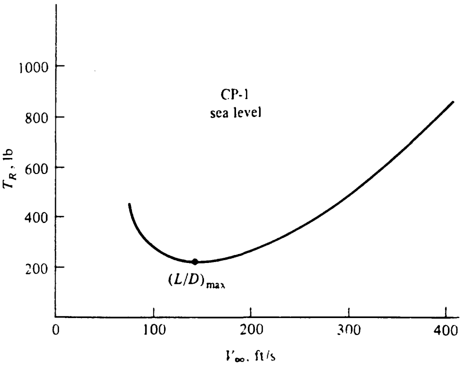

FIG. 31. Thrust-required curve. [From Anderson (1989).]

2.3 Rate of Climb

Returning to Fig. 30, when the airplane is in a steady, unaccelerated climb, the equations of motion become

$$T = D + W\cos\theta,$$

$$L = W\cos\theta.$$

Multiplying the first of these two equations by V_∞, we have

$$(TV_\infty - DV_\infty)/W = V_\infty \sin\theta.$$

However, $TV_\infty - DV_\infty = P_A - P_R$, which is the difference between maximum power available and power required. This difference is called *excess power*. Also, geometrically, $V_\infty \sin\theta$ is the *vertical* velocity of the airplane, called the *rate of climb*, *R/C*. Hence, from the above equation,

$$R/C = (\text{excess power})/W.$$

2.4 Gliding Flight

An airplane in powerless, steady, gliding flight is shown in Fig. 32. The equations of motion reduce for this case to

$$D = W\sin\theta,$$

$$L = W\cos\theta.$$

Hence

$$\tan\theta = \frac{1}{L/D}.$$

The equilibrium glide angle, θ, given by the above equation, is inversely proportional to the lift-to-drag ratio.

2.5 Maximum Altitude

As the altitude of an airplane increases, at the same velocity the drag decreases and the engine thrust generally decreases (both because ρ_∞ decreases with altitude). At some altitude, the maximum excess power will become at best zero, or negative, depending on the velocity. This altitude is the maximum altitude of the airplane. It is defined as the *absolute ceiling*. A more practical index is the *service ceiling*, defined as that altitude where the maximum rate of climb is 100 feet per minute.

2.6 Range and Endurance

Range is technically defined as the total distance (measured with respect to the ground) traversed by the airplane on a tank of fuel. A related quantity is endurance, which is defined as the total time that an airplane stays in the air on a tank of fuel. In different applications, it may be desirable to maximize one or the other of these characteristics. The parameters which maximize range are different from those which maximize endurance; they are also different for propeller- and jet-powered aircraft. Let us first consider the propeller-driven airplane, as discussed below.

One of the critical factors influencing range and endurance is the *specific fuel consumption*, a characteristic of the engine. For a reciprocating engine, specific fuel consumption (commonly abbreviated as SFC) is defined as the *weight of fuel consumed per unit power per unit time*. As mentioned earlier, reciprocating engines are rated in terms of horsepower, and the common units (although inconsistent) of specific fuel consumption are

$$\text{SFC}=\frac{\text{lb of fuel}}{(\text{bhp})(\text{h})},$$

where bhp signifies shaft brake horsepower.

First, consider endurance. On a qualitative basis, in order to stay in the air for the longest period of time, common sense says that we must use the *minimum* number of pounds of fuel per hour. On a dimensional basis this quantity is proportional to the horsepower required by the airplane and to the SFC:

$$(\text{lb of fuel})/\text{h} \propto (\text{SFC})(\text{hp}_R).$$

Therefore, minimum pounds of fuel per hour is obtained with minimum hp_R. Since minimum pounds of fuel per hour gives maximum endurance, we quickly conclude that *Maximum endurance for a propeller-driven airplane occurs when the airplane is flying at minimum power required.* It can be shown that this condition also corresponds to flying such that $C_L^{3/2}/C_D$ is maximum.

Now, consider range. In order to cover the longest distance (say, in miles), common sense says that we must use the minimum number of pounds of fuel per mile. On a dimensional basis, we can state the proprotionality

$$\frac{\text{lb of fuel}}{\text{mi}} \propto \frac{(\text{SFC})(\text{hp}_R)}{V_\infty}.$$

As a result, minimum pounds of fuel per mile is obtained with a minimum hp_R/V_∞. This minimum value of hp_R/V_∞ precisely corresponds to maximum L/D. Thus, *maximum range for a propeller-driven airplane occurs when the airplane is flying at a velocity such that C_L/C_D is maximum.*

Approximate expressions for the range and endurance of a propeller driven airplane are as follows:

range,

$$R=\frac{\eta}{c}\frac{C_L}{C_D}\ln\frac{W_0}{W_1};$$

endurance,

$$E=\frac{\eta}{c}\frac{C_L^{3/2}}{C_D}(2\rho_\infty S)^{1/2}(W_1^{1/2}-W_0^{1/2}),$$

where η is the propeller efficiency, c is the specific fuel consumption, W_1 is the weight of the airplane with empty fuel tanks, and W_0 is the full gross weight of the airplane. The units in the above equations must be consistent units, such as the SI system or the British engineering system.

Let us now consider a jet-propelled airplane. For a jet airplane, the specific fuel consumption is defined as the *weight of fuel consumed per unit thrust per unit time*. Note that *thrust* is used here, in contradistinction to power as in the previous case for a reciprocating-engine-propeller combination. The fuel consumption of a jet engine physically depends on the thrust produced by the engine, whereas the fuel consumption of a re-

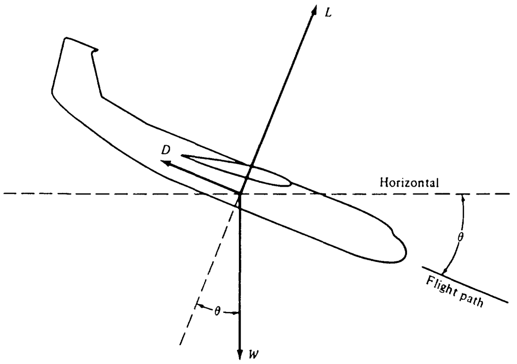

FIG. 32. Airplane in power-off gliding flight. [From Anderson (1989).]

ciprocating engine physically depends on the brake power produced. It is this simple difference which leads to different range and endurance formulas for a jet airplane. In the literature, thrust-specific fuel consumption (TSFC) for jet engines is commonly given as

$$\text{TSFC}=\frac{\text{lb of fuel}}{(\text{lb of thrust})(\text{h})}.$$

(Note the inconsistent unit of time.)

The maximum endurance of a jet airplane occurs for minimum pounds of fuel per hour, the same as for propeller-driven aircraft. However, for a jet,

$$(\text{lb of fuel})/\text{h}=(\text{TSFC})T_A,$$

where T_A is the thrust available produced by the jet engine. Recall that in steady, level, unaccelerated flight, the pilot has adjusted the throttle such that thrust available T_A just equals the thrust required T_R: $T_A=T_R$. Therefore, minimum pounds of fuel per hour corresponds to minimum thrust required. Hence, we conclude that *maximum endurance for a jet engine occurs when the airplane is flying at minimum thrust required.* We have already seen that minimum thrust required corresponds to flying at maximum L/D.

Now, consider range. As before, maximum range occurs for a minimum pounds of fuel per mile. For a jet, on a dimensional basis,

$$(\text{lb of fuel})/\text{mi}=(\text{TSFC})T_A/V_\infty.$$

Recalling that for steady, level flight, $T_A=T_R$, we note that minimum pounds of fuel per mile corresponds to a minimum T_R/V_∞. In turn, T_R/V_∞ is the slope of a line through the origin and intersecting the thrust-required curve; its minimum value occurs when the line becomes tangent to the thrust-required curve. The aerodynamic condition holding at this tangent point is obtained as follows. Recalling that for steady, level flight, $T_R=D$, then

$$\frac{T_R}{V_\infty}=\frac{D}{V_\infty}=\frac{\frac{1}{2}\rho_\infty V_\infty^2 S C_D}{V_\infty}=\frac{1}{2}\rho_\infty V_\infty S C_D.$$

Since $V_\infty=\sqrt{2W/\rho_\infty S C_L}$, we have

$$\frac{T_R}{V_\infty}=\frac{1}{2}\rho_\infty S\sqrt{\frac{2W}{\rho_\infty S C_L}}\,C_D\propto\frac{1}{C_L^{1/2}/C_D}.$$

Hence, minimum T_R/V_∞ corresponds to maximum $C_L^{1/2}/C_D$. In turn, we conclude that *maximum range for a jet airplane occurs when the airplane is flying at a velocity such that $C_L^{1/2}/C_D$ is maximum.*

Approximate relations for range and endurance for a jet airplane are as follows:

range,

$$R=2\sqrt{\frac{2}{\rho_\infty S}}\,\frac{1}{c_t}\,\frac{C_L^{1/2}}{C_D}\,(W_0^{1/2}-W_1^{1/2});$$

endurance,

$$E=\frac{1}{c_t}\,\frac{C_L}{C_D}\,\ln\frac{W_0}{W_1},$$

where c_t is the thrust-specific fuel consumption.

2.7 Turning Flight

A level turn is illustrated in Fig. 33. Here, the wings of the airplane are banked through the angle ϕ; hence the lift vector is inclined at the angle ϕ to the vertical. The bank angle ϕ and the lift L are such that the component of the lift in the vertical direction exactly equals the weight:

$$L\cos\phi=W$$

and therefore the airplane maintains a constant altitude, moving in the same horizontal plane. However, the resultant of L and W leads to a resultant force F_r, which acts in the horizontal plane. This resultant force is perpendicular to the flight path, causing the airplane to turn in a circular path with a radius of curvature equal to R. We wish to study this turn radius R, as well as the turn rate $d\theta/dt$.

From the force diagram in Fig. 33, the magnitude of the resultant force is

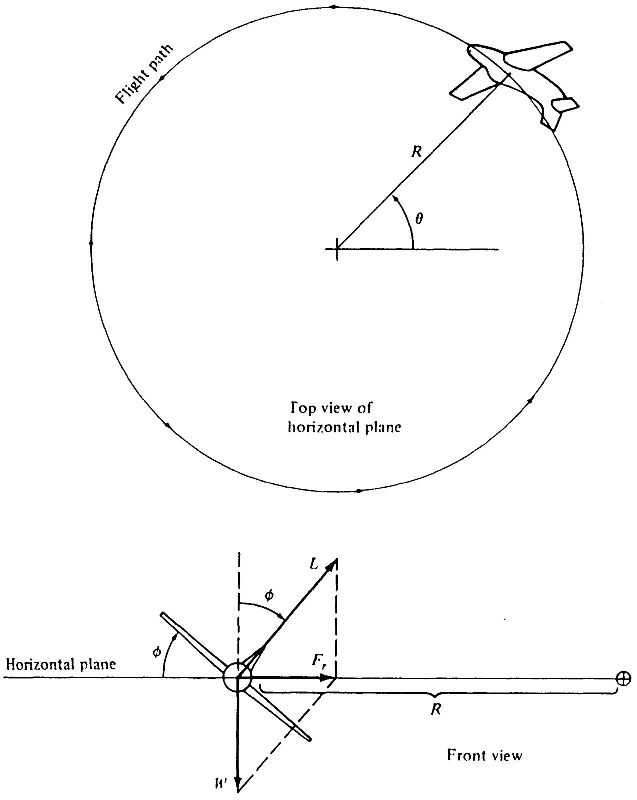

FIG. 33. An airplane in a level turn. [From Anderson (1989).]

$$F_r = \sqrt{L^2 - W^2}.$$

We introduce the *load factor n*, defined as

$$n \equiv L/W.$$

The load factor is usually quoted in terms of "*g*'s"; for example, an airplane with lift equal to five times the weight is said to be experiencing a load factor of 5*g*'s. Hence,

$$F_r = W\sqrt{n^2 - 1}.$$

The airplane is moving in a circular path at the velocity V_∞. Therefore, the radial acceleration is given by V_∞^2/R. From Newton's second law

$$F_r = m\frac{V_\infty^2}{R} = \frac{W}{g}\frac{V_\infty^2}{R}.$$

Combining, we have

$$R = V_\infty^2/g\sqrt{n^2 - 1}.$$

The angular velocity, denoted by $\omega = d\theta/dt$, is called the turn rate and is given by V_∞/R. Thus,

$$\omega = g\sqrt{n^2 - 1}/V_\infty.$$

For the maneuvering performance of an airplane, both military and civil, it is frequently advantageous to have the smallest possible R and the largest possible ω. The above equations show that to obtain both a small turn radius and a large turn rate, we want

1. the highest possible load factor (i.e., the highest possible L/W);
2. the lowest possible velocity.

3. AIRPLANE STABILITY AND CONTROL

3.1 Definition of Stability and Control

Airplane performance, as discussed in Sec. 2, is governed by forces (along and perpendicular to the flight path), with the translational motion of the airplane as a response to these forces. In contrast, airplane stability and control, discussed in the present section, are governed by moments about the center of gravity, with the rotational motion of the airplane as a response to these moments. Therefore, moments and rotational motion are the main focus of this section.

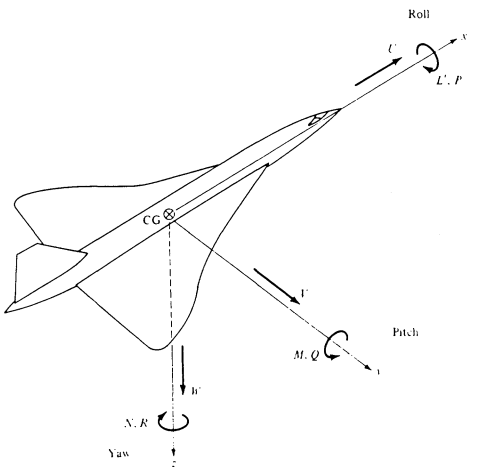

FIG. 34. Definition of the airplane's axes along with the translational and rotational motion along and about these axes. [From Anderson (1989).]

Consider an airplane in flight, as sketched in Fig. 34. The center of gravity (the point through which the weight of the complete airplane effectively acts) is denoted as cg. The *xyz* orthogonal axis system is fixed relative to the airplane; the *x* axis is along the fuselage, the *y* axis is along the wingspan perpendicular to the *x* axis, and the *z* axis is directed downward, perpendicular to the *xy* plane. The origin is at the center of gravity. The translational motion of the airplane is given by the velocity components, U, V, and W along the *x*, *y*, and *z* directions, respectively. (Note that the resultant freestream velocity V_∞ is the vector sum of U, V, and W.) The rotational motion is given by the angular velocity components P, Q, and R about the *x*, *y*, and *z* axes, respectively. These rotational velocities are due to the moments L', M, and N about the *x*, *y*, and *z* axes, respectively. (The prime is put over the symbol L so that the reader avoids confusing it with lift.) Rotational motion about the *x* axis is called *roll*; L' and P are the *rolling* moment and velocity, respectively. Rotational motion about the *y* axis is called *pitch*; M and Q are the *pitching* moment and velocity, respectively. Rotational motion about the *z* axis is called *yaw*; N and R are the *yawing* moment and velocity, respectively.

There are three basic controls on an airplane—the ailerons, elevator, and rudder—which are designed to change and control the moments about the *x*, *y*, and *z* axes. These control surfaces are shown in Fig. 35; they are flaplike surfaces that can be deflected back and forth at the command of the pilot. The ailerons are mounted at the trailing edge of the wing, near the wingtips. The elevators are located on the horizontal stabilizer. In some modern aircraft, the complete horizontal stabilizer is rotated instead of just the elevator (so-called flying tails). The rudder is located on the vertical stabilizer, at the trailing edge. Just as in the case of wing flaps discussed in Sec. 1.5, a downward deflection of the control surface will increase the lift of the wing or tail. In turn, the moments will be changed, as sketched in Fig. 36. Consider Fig. 36(a). One aileron is deflected up, the other down, creating a differential lifting force on the wings, thus contributing to the rolling moment L'. In Fig. 36(b), the elevator is deflected upward, creating a negative lift at the tail, thus contributing to the pitching moment M. In Fig. 36(c), the rudder is deflected to the right, creating a leftward aerodynamic force on the tail, thus contributing to the yawing moment N.

Rolling (about the *x* axis) is also called *lateral motion*. Referring to Fig. 36(a), we see that ailerons control roll; hence they are known as *lateral controls*. Pitching (about the *y* axis) is also called *longitudinal motion*. In Fig. 36(b), we see that elevators control pitch; hence they are known as *longitudinal con-*

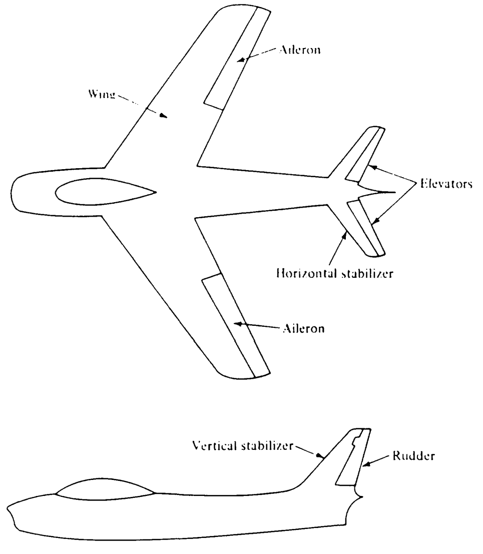

FIG. 35. Some airplane nomenclature. [From Anderson (1989).]

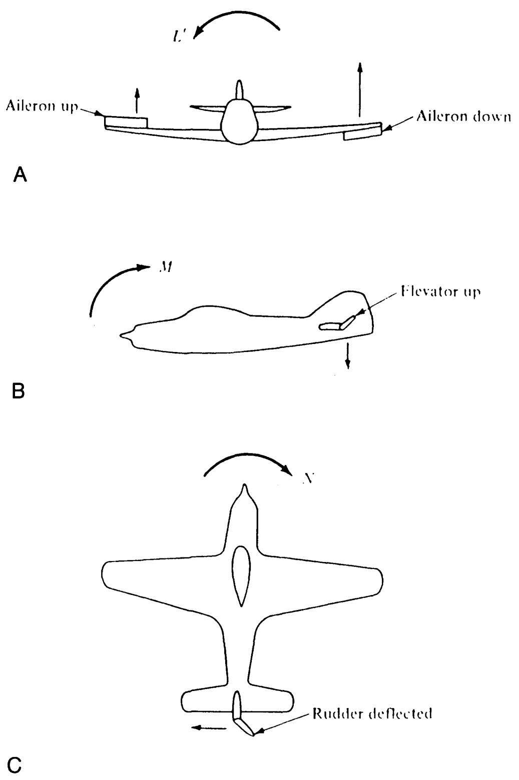

FIG. 36. Effect of control deflections on roll, pitch, and yaw. (a) Effect of aileron deflection; lateral control. (b) Effect of elevator deflection; longitudinal control. (c) Effect of rudder deflection; directional control. [From Anderson (1989).]

trols. Yawing (about the z axis) is also called *directional motion.* Figure 36(c) shows that the rudder controls yaw; hence it is known as the *directional control.*

There are two types of stability; static and dynamic. They can be visualized as follows. First, consider static stability. Consider a marble on a curved surface, such as a bowl. Imagine that the bowl is upright and the marble is resting inside, as shown in Fig. 37(a). The marble is stationary; it is in a state of *equilibrium,* which means that the moments acting on the marble are zero. If the marble is now disturbed [moved to one side, as shown by the dotted circle in Fig. 37(a)] and then released, it will roll back toward the bottom of the bowl, i.e., toward its original equilibrium position. Such a system is *statically* stable. In general, we can state that *if the forces and moments on the body caused by a disturbance tend initially to return the body toward its equilibrium position, the body is statically stable. The body has positive static stability.*

Now, imagine the bowl is upside-down, with the marble at the crest, as shown in Fig. 37(b). If the marble is placed precisely at the crest, the moments will be zero and the marble will be in equilibrium. However, if the marble is now disturbed [as shown by the dotted circle in Fig. 37(b)], it will tend to roll down the side, away from its equilibrium position. Such a system is *statically unstable.* In general, we can state that *if the forces and moments are such that the body continues to move away from its equilibrium position after being disturbed, the body is statically unstable. The body has negative static stability.*

Finally, imagine the marble on a flat horizontal surface, as shown in Fig. 37(c). Its moments are zero; it is in equilibrium. If the marble is now disturbed to another location, the moments will still be zero, and it will still be in equilibrium. Such a system is *neutrally stable.* This situation is rare in flight vehicles, and we will not be concerned with it here.

Emphasis is made that static stability (or the lack of it) deals with the *initial* tendency of a vehicle to return to equilibrium (or to diverge from equilibrium) after being disturbed. It says nothing about whether it ever reaches its equilibrium position, nor how it gets there. Such matters are the realm of dynamic stability, as follows.

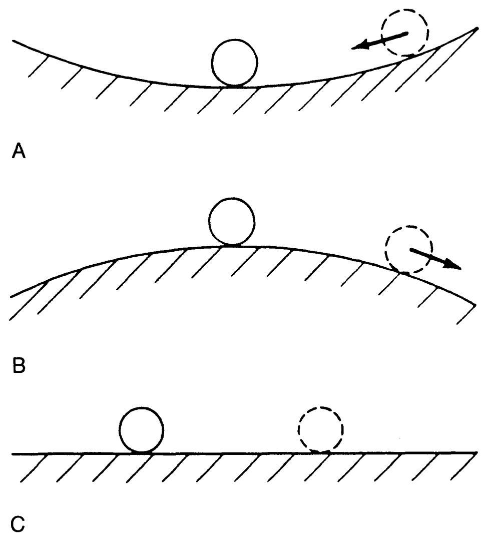

FIG. 37. Illustration of static stability. (a) Statically stable system. (b) Statically unstable system. (c) Statically neutral system. [From Anderson (1989).]

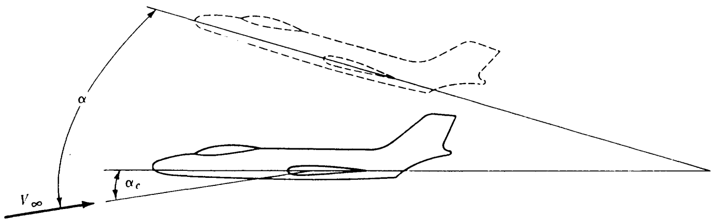

FIG. 38. Disturbance from the equilibrium angle of attack.

Dynamic stability deals with the *time history* of the vehicle's motion after it initially responds to its static stability. For example, consider an airplane flying at an angle of attack α_e such that its moments about the center of gravity are zero. The airplane is therefore in equilibrium at α_e; in this situation, it is *trimmed*, and α_e is called the trim angle of attack. Now assume that the airplane is disturbed (say, by encountering a wind gust) to a new angle of attack α as shown in Fig. 38. The airplane has been pitched through a *displacement* $\alpha-\alpha_e$. Now, let us observe the subsequent pitching motion after the airplane has been disturbed by the gust. We can describe this motion by plotting the instantaneous displacement versus time, as shown in Fig. 39. Here $\alpha-\alpha_e$ is given as a function of time t. At $t=0$, the displacement is equal to that produced by the gust. If the airplane is statically stable, it will *initially* tend to move back towards its equilibrium position, that is, $\alpha-\alpha_e$ will initially decrease. Over a lapse of time, the vehicle may monotonically "home-in" to its equilibrium position, as shown in Fig. 39(a). Such motion is called aperiodic. Alternately, it may first overshoot the equilibrium position and approach α_e after a series of oscillations with decreasing amplitude, as shown in Fig. 39(b). Such motion is described as damped oscillations. In both situations, the airplane eventually returns to its equilibrium position after some interval of time. These two situations are examples of *dynamic stability* in an airplane. Thus, we can state that *a body is dynamically stable if, out of its own accord, it eventually returns to and remains at its equilibrium position over a period of time.*

On the other hand, after initially responding to its static stability, the airplane may oscillate with increasing amplitude, as shown in Fig. 40. Here, the equilibrium position is never maintained for any period of time; the airplane in this case is *dynamically unstable* (even though it is statically stable). Also, it is theoretically possible for the airplane to pitch back and forth with constant-amplitude oscillations. This is an example of a *dynamically neutral* body; such a case is of little practical interest here.

It is important to observe from the above examples that a dynamically stable airplane must always be statically stable. On the other hand, static stability is *not* sufficient to ensure dynamic stability. Nevertheless, static stability is usually the first stability characteristic to be designed into an airplane. (There are some exceptions, to be discussed later.)

The conventional control surfaces (elevators, ailerons, and rudder) on an airplane were sketched in Figs. 35 and 36. Their function is usually to change the airplane from one equilibrium position to another and to produce nonequilibrium accelerated motions such as maneuvers. The study of the *deflection* of the ailerons, elevators, and rudder necessary to make the airplane do what we want and of the amount of *force* that must be exerted by the pilot (or the hydraulic boost system) to deflect these controls is part of a discipline called "airplane control."

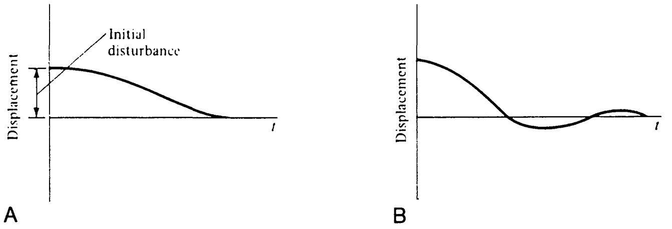

FIG. 39. Examples of dynamic stability: (a) Aperiodic; (b) Damped oscillations. [From Anderson (1989).]

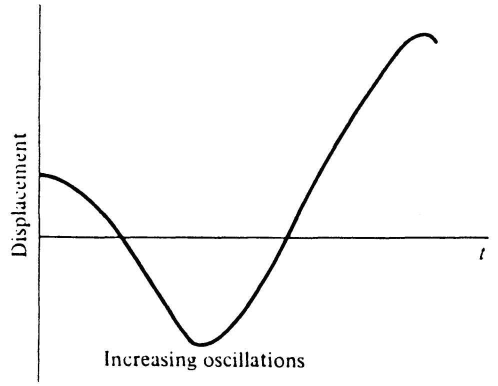

FIG. 40. An example of dynamic instability. [From Anderson (1989).]

3.2 Moments on the Airplane

A study of stability and control is focused on moments: moments on the airplane and moments on the control surfaces. Recall that the pressure and shear stress distributions over a wing produce a pitching moment. This moment can be taken about any arbitrary point (the leading edge, the trailing edge, the quarter chord, etc.).

Now consider the complete airplane, as sketched in Fig. 41. Here, we are most concerned with the pitching moment about the center of gravity of the airplane, M_{cg}. Clearly, by examination of Fig. 41, M_{cg} is created by L, D, and M_{ac} of the wing; lift of the tail; thrust; and aerodynamic forces and moments on other parts of the airplane, such as the fuselage and engine nacelles. (Note that weight does not contribute, since it acts through the center of gravity.)

The moment coefficient about the center of gravity is defined as

$$C_{M,\text{cg}}=M_{\text{cg}}/q_{\infty}Sc.$$

An airplane is in equilibrium (in pitch) when the moment about the center of gravity is zero; i.e., when $M_{\text{cg}} = C_{M,\text{cg}}=0$, the airplane is said to be *trimmed*.

3.3 Criteria for Longitudinal Static Stability

Static stability and control about all three axes shown in Fig. 34 is usually a necessity in the design of conventional airplanes. However, a complete description of all three types—lateral, longitudinal, and directional static stability and control—is beyond the scope of this article. Rather, the intent here is to provide only the flavor of stability and control concepts, and to this end only the airplane's longitudinal motion (pitching motion about the y axis) will be considered. This pitching motion is illustrated in Fig. 36(b). It takes place in the plane of symmetry of the airplane. Longitudinal stability is also the most important static stability mode; in airplane design, wind-tunnel testing, and flight research, it usually earns more attention than lateral or directional stability.

Consider a rigid airplane with fixed controls, e.g., the elevator in some fixed position. Assume the airplane has been tested in a wind tunnel or free flight and that its variation of M_{cg} with angle of attack has been measured. This variation is illustrated in Fig. 42, where $C_{M,\text{cg}}$ is sketched versus α. For many conventional airplanes, the curve is nearly linear, as shown in Fig. 42. The value of $C_{M,\text{cg}}$ at zero lift is denoted by $C_{M,0}$. The value of α where $M_{\text{cg}}=0$ is denoted by α_e; this is the equilibrium, or trim, angle of attack.

Consider the airplane in steady, equilibrium flight at its trim angle of attack α_e as shown in Fig. 43(a). Suddenly, the airplane is disturbed by hitting a wind gust, and the angle of attack is momentarily changed. There are two possibilities: an increase or a decrease in α. If the airplane is pitched upward, as shown in Fig. 43(b), then $\alpha > \alpha_e$. From Fig. 42, if $\alpha > \alpha_e$, the moment about the center of

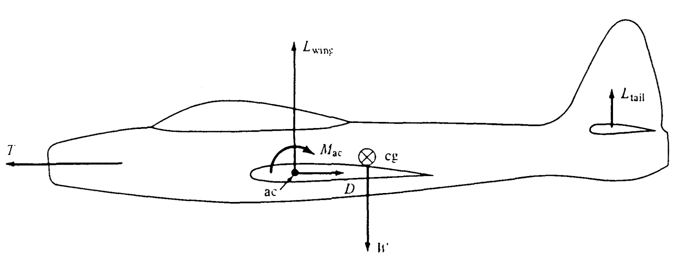

FIG. 41. Contributions to the moment about the center of gravity of the airplane. [From Anderson (1989).]

gravity is negative. A negative moment (by convention) is counterclockwise, tending to pitch the nose downward. Hence, in Fig. 43(b), the airplane will initially tend to move back toward its equilibrium position after being disturbed. On the other hand, if the plane is pitched downward by the gust, as shown in Fig. 43(c), then $\alpha < \alpha_e$. From Fig. 42, the resulting moment about the center of gravity will be positive (clockwise) and will tend to pitch the nose upward. Thus, again we have the situation where the airplane will initially tend to move back toward its equilibrium position after being disturbed. This is precisely the definition of static stability. Therefore, we conclude that an airplane which has a $C_{M,\text{cg}}$ vs α variation like that shown in Fig. 42 is *statically stable.* Note that from Fig. 42 that $C_{M,0}$ is positive and that the slope of the curve, $\partial C_{M,\text{cg}}/\partial \alpha_a$, is negative. Here, the partial derivative is used for the slope of the moment coefficient curve. This is because $C_{M,\text{cg}}$ depends on a number of other variables in addition to α, and therefore it is mathematically proper to use $\partial C_{M,\text{cg}}/\partial \alpha_a$ rather than $dC_{M,\text{cg}}/d\alpha_a$ to represent the slope of the line in Fig. 42.

Consider now a different airplane, with a measured $C_{M,\text{cg}}$ variation as shown in Fig. 44. Imagine the airplane is flying at its trim angle of attack α_e as shown in Fig. 45(a). If it is disturbed by a gust, pitching the nose upward, as shown in Fig. 45(b), then $\alpha > \alpha_e$. From Fig. 44 this results in a positive (clockwise) moment, which tends to pitch the nose even further away from its equilibrium position. Similarly, if the gust pitches the nose downward [Fig. 45(c)], a negative (counterclockwise) moment results, which also tends to pitch the nose further away from its equilibrium position. Therefore, because the airplane tends to diverge from equilibrium when disturbed, it is *statically unstable.* Note from Fig. 42 that $C_{M,0}$ is negative and $\partial C_{M,\text{cg}}/\partial \alpha_a$ is positive for this airplane.

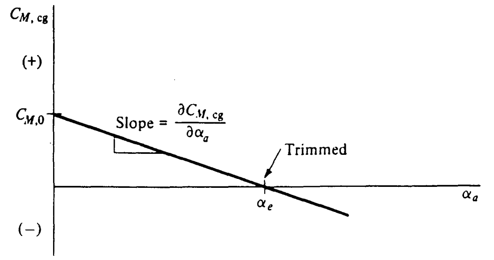

FIG. 42. Moment coefficient curve with a negative slope. [From Anderson (1989).]

For both airplanes, Figs. 42 and 44 show a positive value of α_e. An airplane moves through a range of angle of attack as it flies through its velocity range from V_{stall} (where α

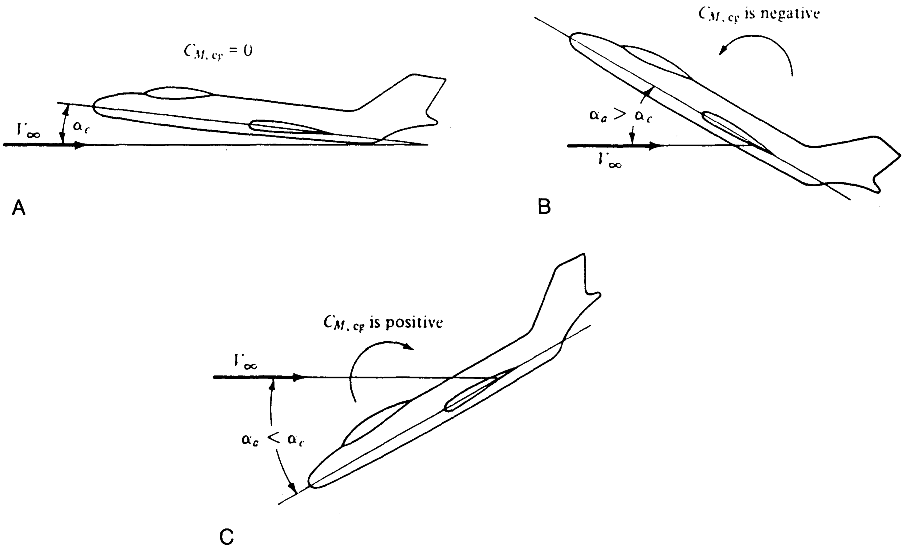

FIG. 43. Illustration of static stability: (a) Equilibrium position (trimmed). (b) Pitched upward by disturbance. (c) Pitched downward by disturbance. In both (b) and (c), the airplane has the initial tendency to return to its equilibrium position. [From Anderson (1989).]

is the largest) to $V_{\max}$ (where α is the smallest). The value of α_e must fall within this flight range of angle of attack, or else the airplane cannot be trimmed for steady flight. (Remember that we are assuming a fixed elevator position; we are discussing "stick-fixed" stability.) When α_e does fall within this range, the airplane is *longitudinally balanced.*

From the above considerations, we conclude the following. The necessary criteria for longitudinal balance and static stability are as follows:

1. $C_{M,0}$ must be positive.
2. $\partial C_{M,\mathrm{cg}}/\partial \alpha_a$ must be negative.

That is, the $C_{M,\mathrm{cg}}$ curve must look like Fig. 42. Of course, implicit in the above criteria is that α_e must also fall within the flight range of angle of attack for the airplane.

We are now in a position to explain why a conventional airplane has a horizontal tail (the horizontal stabilizer is shown in Fig. 35). First consider, an ordinary wing (by itself) with a conventional airfoil, say an NACA 2412 section. Note from the airfoil data in Figs. 9 and 10 that the moment coefficient about aerodynamic center is negative. This is characteristic of all airfoils with positive camber. Now, assume that the wing is at zero lift. In this case, the only moment on the wing is a pure couple; hence, at zero lift the moment about one point is equal to the moment about any other point. In particular,

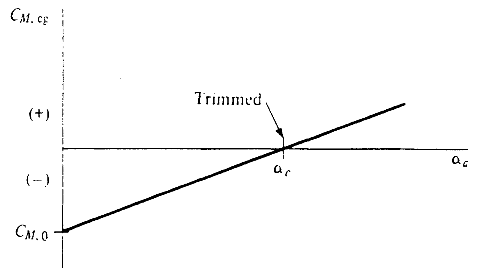

FIG. 44. Moment coefficient curve with a positive slope. [From Anderson (1989).]

$$C_{M,\mathrm{ac}} = C_{M,\mathrm{cg}} \text{ for zero lift (wing only).}$$

On the other hand, examination of Fig. 42 shows that $C_{M,0}$ is by definition the moment coefficient about the center of gravity at zero lift (when $\alpha = 0$). Hence,

$$C_{M,0} = C_{M,\mathrm{ac}} \text{ (wing only).}$$

This equation demonstrates that, for a wing with positive camber ($C_{M,\mathrm{ac}}$ negative) $C_{M,0}$ is also negative. Hence, such a wing by itself is *unbalanced.* To rectify this situation, a horizontal tail must be added to the airplane, as shown in Figs. 46(a) and 46(b). If the tail is mounted behind the wing, as shown in Fig.

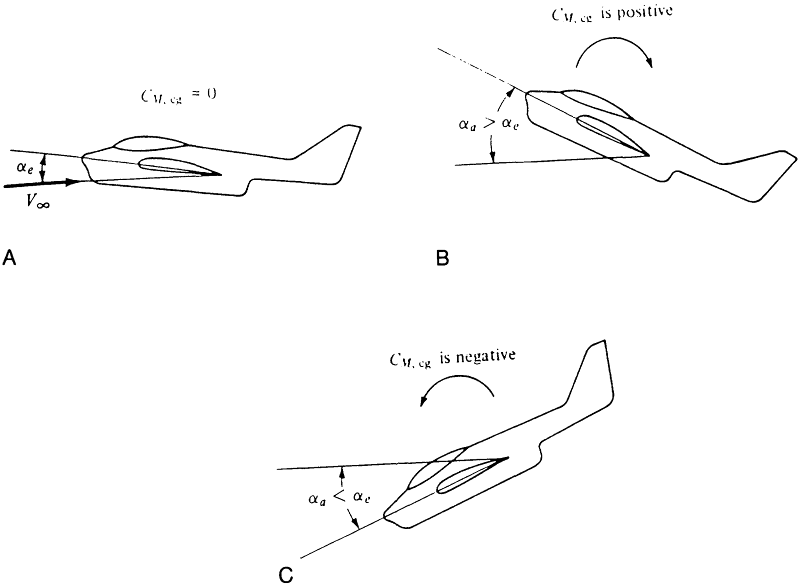

FIG. 45. Illustration of static instability. (a) Equilibrium position (trimmed). (b) Pitched upward by disturbance. (c) Pitched downward by disturbance. In both (b) and (c), the airplane has the initial tendency to diverge farther away from its equilibrium position. [From Anderson (1989).]

46(a), and if it is inclined downward to produce a negative tail lift as shown, then a clockwise moment about the center of gravity will be created. If this clockwise moment is strong enough, it will overcome the negative $C_{M,\text{ac}}$ and $C_{M,0}$ for the wing-tail combination will become positive. The airplane will then be balanced.

The arrangement shown in Fig. 46(a) is characteristic of most conventional airplanes. However, the tail can also be placed ahead of the wing, as shown in Fig. 46(b); this is called a *canard configuration*. For a canard, the tail is inclined upward to produce a positive lift, hence creating a clockwise moment about the center of gravity. If this moment is strong enough, $C_{M,0}$ for the wing-tail combination will become positive and again the airplane will be balanced.

3.3.1 The Neutral Point and Static Margin. An airplane will be statically stable in longitudinal motion if the slope of the moment coefficient curve is negative, as explained above. The location of the center of gravity plays a strong role in the value of this slope. The slope can be made more or less negative by moving the center of gravity forward or backward, respectively. There is some rearward location of the center of gravity at which the moment coefficient slope goes to zero, and hence the airplane becomes neutrally stable. This location is called the *neutral point*; it is a fixed point on the airplane. Hence, another criterion for stability is that the *center of gravity must be forward of the neutral point*. Basically, the neutral point is the *aerodynamic center* of the complete airplane; the moment coefficient about this point is independent of angle of attack.

The static margin is defined as the *distance* between the center of gravity and the neutral point, divided by the mean chord of the wing. When the center of gravity is forward of the neutral point, the static margin is positive. For static stability, the static margin must be positive.

It is interesting to note that the Wright brothers intentionally designed their airplanes to be statically unstable, i.e., with a negative static margin. This was to make the airplane immediately responsive to the controls. In contrast, a highly statically stable airplane naturally resists perturbations to its motion, and is therefore sluggish at the controls. On the other hand, the Wright airplanes were, as a result of the static instability, hard to fly. Virtually all conventional airplane designs since the Wright brothers were statically stable aircraft. However, with the advent of computer driven controls, modern military fighter aircraft such as the General Dynamics F-16 are once again being designed with a certain amount of negative static margin in order to enhance maneuverability.

4. HYPERSONIC FLIGHT

Flight in the 21st century will include airplanes which fly at many times the speed of sound—hypersonic flight. As a "rule-of-thumb" definition, hypersonic flight is that above a Mach number of five. However, hypersonic flow is best defined as that high-Mach number regime where certain physical flow phenomena become progressively more important, that are not important at lower supersonic or subsonic speeds. Among these are:

1. Thin shock layers. The shock waves lie very close to the body, and may even interact with the viscous boundary layer growing along the surface.
2. Very thick boundary layers, which can strongly interact with the outer inviscid flow. This is called *viscous interaction* problem.
3. High temperatures. The amount of kinetic energy is extremely large in a hypersonic flow. In regions on a body where the flow is slowed to low speeds, such as in the nose and leading-edge regions, and within a viscous boundary layer, the original kinetic energy is converted to internal energy in the gas, hence causing very high temperatures to occur in these regions. In such cases, the oxygen and nitrogen molecules in air will dissociate, and even ionization can occur. That is, the air can become a partially ionized plasma, and the flowfield over the body is a chemically reacting flow.

These effects distinguish hypersonic flight as a special regime in aeronautics.

Figure 47 shows a three-view diagram of a conceptual hypersonic cruise aircraft, e.g., a hypersonic transport. This NASA concept, in existence since the early 1970's, is a typical

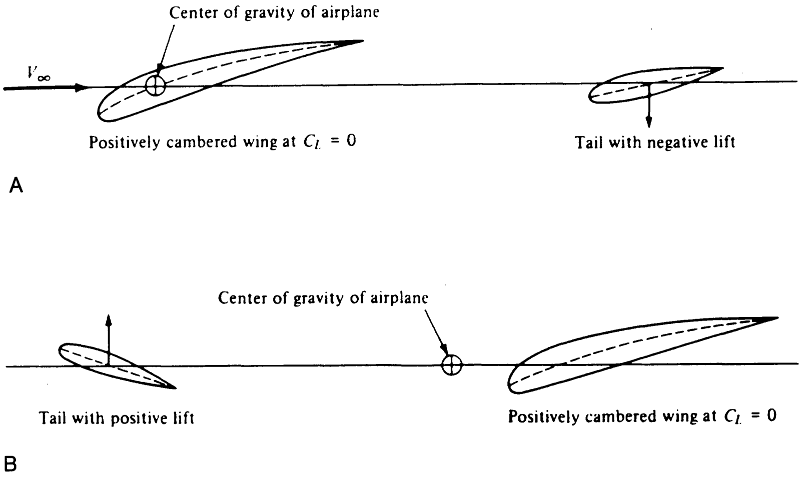
Center of gravity of airplane
V_∞
Positively cambered wing at $C_L = 0$
Tail with negative lift
A
Center of gravity of airplane
Tail with positive lift
Positively cambered wing at $C_L = 0$
B

FIG. 46. (a) Conventional wing-tail combination. The tail is set at such an angle as to produce negative lift, thus providing a positive $C_{M,0}$. (b) Canard wing-tail combination. The tail is set at such an angle as to produce positive lift, thus providing a positive $C_{M,0}$. [From Anderson (1989).]

example of a hypersonic airplane configuration. The solid lines show the wing-body combination, which was tested in a hypersonic wind tunnel; the dashed lines show the propulsion module and vertical tail surface, which are part of the airplane design but were not included in the wind-tunnel model.

As a final note in this section, we mention an important characteristic of any hypersonic-airplane design, namely, the necessity to integrate the propulsion system fully with the airframe. For subsonic airplane design, some attention is always paid to the aerodynamic interaction between the engine

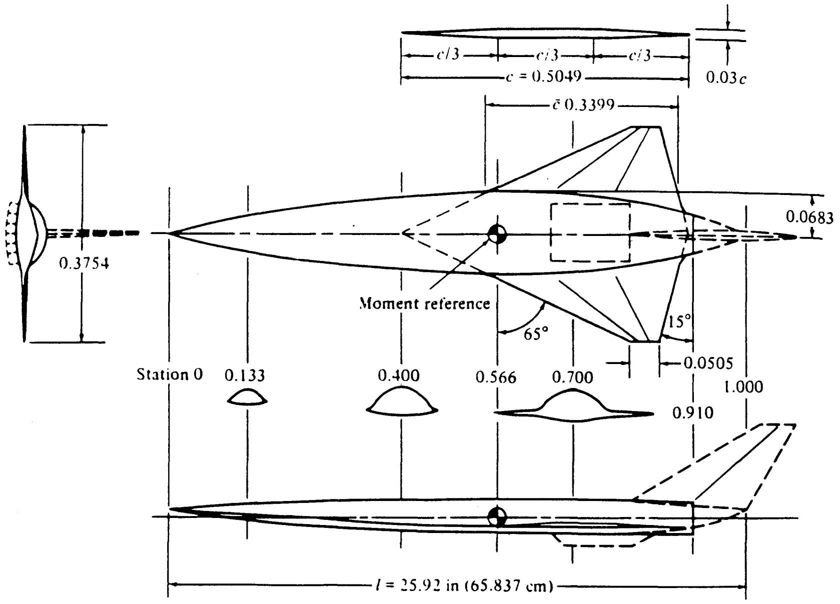
c/3
c/3
c/3
c = 0.5049
0.03c
c̄ 0.3399
0.0683
0.3754
Moment reference
65°
15°
0.0505
Station 0
0.133
0.400
0.566
0.700
1.000
0.910
l = 25.92 in (65.837 cm)

FIG. 47. A generic hypersonic-transport configuration. The dimensions pertain to a wind-tunnel model. Solid lines=wing-body model used for the wind-tunnel tests: dashed lines=tail and propulsion modules for the complete configuration. [From Penlund *et al.* (1983).]

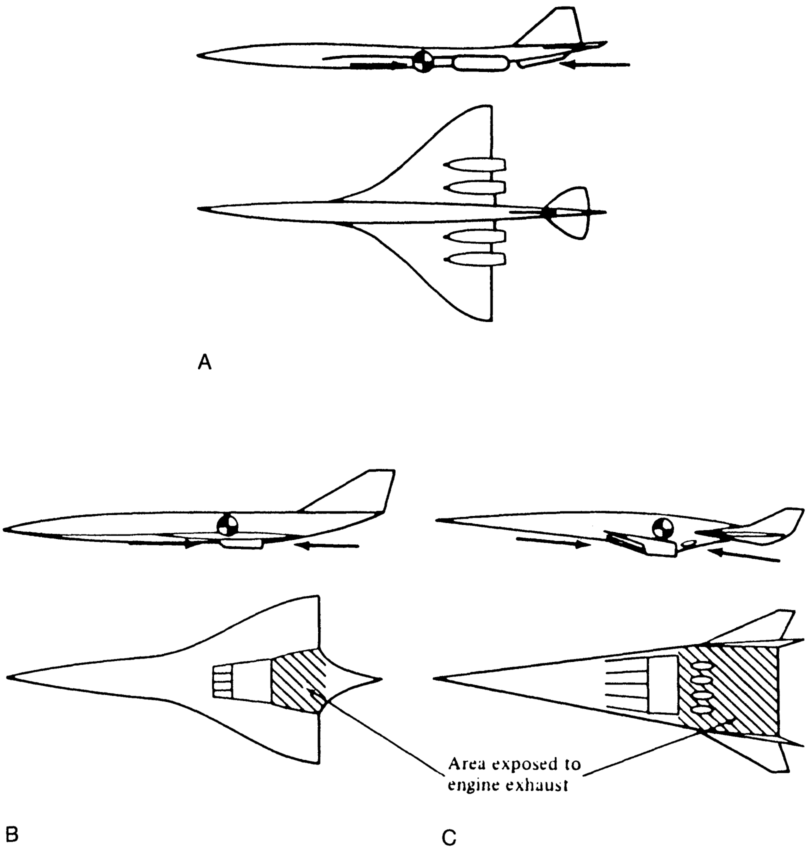

FIG. 48. Comparison of high-speed airplane design from Mach 3 to 12. (a) Supersonic transport ($M=3$), ram drag=54 500 lb, gross thrust=123 000 lb; (b) hypersonic transport ($M=6$), ram drag=220 000 lb, gross thrust=330 000 lb; (c) hypersonic cruise vehicle ($M=12$), ram drag=1950 000 lb, gross thrust=2100 000 lb. (Courtesy of NASA.)

nacelles and the rest of the airframe. However, this is not a driving aspect of airplane design, and in most subsonic airplanes the location of the engines is distinctly obvious. In contrast, at hypersonic speeds, extreme care must be taken to ensure that shock waves from one portion of the airplane, including the propulsion system, do not adversely impinge upon, and interact with, other portions of the airplane. Moreover, the flow that goes through the supersonic combustion ramjet engines has first passed through one or more shock-wave systems from the forward portion of the vehicle, and it is necessary to tailor the aerodynamic properties of this air so that they encourage the most efficient engine performance. Therefore, for a hypersonic airplane, the propulsion system and the airframe must be highly integrated. As an example, three typical generic configurations are compared in Fig. 48: a Mach 3 supersonic transport, a Mach 6 hypersonic transport, and a Mach 12 hypersonic cruise vehicle. Note that the supersonic airplane [Fig. 48(a)] still has fairly distinct propulsion nacelles and that none of the engine exhaust is designed to touch the airframe. In contrast, the Mach 6 hypersonic transport [Fig. 48(b)] has a more fully integrated propulsion system, where the rear part of the airframe acts as part of the engine nozzle expansion. Also, the wing and fuselage are more fully integrated, the wings being less distinct than for the Mach 3 airplane; i.e., the Mach 6 airplane is more of a blended wing-body configuration than the supersonic airplane. At Mach 12 [Fig. 48(c)] these features are even more pronounced; the undersurface area of the airframe exposed to the engine exhaust is much greater, and the engine is much more an integral part of the airframe. Since the wings have become much smaller, the Mach 12 vehicle is more like a *lifting body* than a wing-body combination. Clearly, the design of hypersonic vehicles is a marked departure from conventional airplane design, and this will pose many interesting challenges for the aerospace engineers of the future.

GLOSSARY

For the most part, all terms are defined within the text as appropriate. Here are a few of the important terms.

Aerodynamic Center: That point on an aerodynamic body about which the moments are independent of angle of attack.

Airfoil: A cross section of an airplane wing in the flow direction.

Angle of Attack: Angle between the chord line and the relative wind.

Aspect Ratio: Square of the wing span divided by the wing planform area.

Chord: Distance between the leading and trailing edges of an airfoil.

Critical Mach Number: That freestream Mach number at which sonic flow is first encountered at a point on the aircraft.

Drag: Component of the aerodynamic force parallel to the flight direction.

Lift: Component of the aerodynamic force perpendicular to the flight direction.

Planform Area: Projected area you see by looking down on the wing.

Relative Wind: The velocity and direction of the relative motion between the airplane and the atmosphere.

Span: Distance from one wing tip to the other.

Static Performance: Unaccelerated motion of the airplane.

Subsonic Flight: Mach number at every point the flow over the airplane is less than one.

Supersonic Flight: Mach number at almost every point in the flow over the airplane is greater than one.

Transonic Flight: The flowfield around the aircraft has mixed regions of both subsonic and supersonic speeds.

Trimmed: The condition when the moments about the center of gravity of the airplane are zero.

Wave Drag: Drag at transonic and supersonic speeds due to the presence of shock waves.

Works Cited

Abbott, J. H., von Doenhoff, A. E. (1949), *Theory of Wing Sections*, New York: McGraw-Hill.

Anderson, J. D., Jr. (1989), *Introduction to Flight*, 3rd ed., New York: McGraw-Hill.

Loftin, L. (1985), *Quest for Performance*, NASA Special Publication No. 468, Washington, DC: U.S. GPO.

Penlund, J. A., *et al.* (1983), *Wall Temperature Effects on the Aerodynamics of a Hydrogen-Fueled Transport Concept in Mach 8 Blowdown and Shock Tunnels*, NASA Technical Publication No. 2159, Washington, DC: U.S. GPO.

Further Reading

Anderson, J. D., Jr. (1989), *Introduction to Flight*, 3rd ed., New York: McGraw-Hill.

Anderson, J. D., Jr. (1991), *Fundamentals of Aerodynamics*, 2nd ed., New York: McGraw-Hill.

Bertin, J. J., Smith, M. (1989), *Aerodynamics for Engineers*, 2nd ed., Englewood Cliffs, NJ: Prentice Hall.

McCormick, B. W. (1979), *Aerodynamics, Aeronautics and Flight Mechanics*, New York: Wiley.

Shevell, R. S. (1983), *Fundamentals of Flight*, Englewood Cliffs, New Jersey: Prentice Hall.

AEROSOLS

BENJAMIN Y. H. LIU AND DAVID Y. H. PUI, *Mechanical Engineering Department, University of Minnesota, Minneapolis*

Introduction 415
1. **Aerosol Property and Behavior** 416
1.1 Mechanical Properties 416
1.2 Electrical Properties 416
1.3 Optical Properties and Light Scattering 418
2. **Instrumentation for Aerosol Studies** 418
2.1 Aerosol Generators 418
2.2 Aerosol Measuring Instruments 420
2.3 Aerosol Samplers 422
3. **Scientific and Technical Applications** 424
3.1 Lung Deposition and Respiratory Protection 424
3.2 Atmospheric Aerosols and Air Pollution 427
3.3 Industrial Gas Cleaning 428
3.4 Cleanrooms and Microelectronics 429
3.5 Material Synthesis and Processing 434
3.6 Radioactive Aerosols and Nuclear Safety 435
3.7 Atmospheric Science and Weather Modification 437
Glossary 438
Works Cited 439
Further Reading 440

INTRODUCTION

Small solid or liquid particles suspended in a gas have found numerous scientific and technical applications. Such gas-borne dispersions are usually referred to as aerosols and treated in a branch of study known as aerosol science.

Aerosol is one of three types of dispersed systems found in nature and in various technological applications: aerosol (solid or liquid particles dispersed in a gas), hydrosol (solid or liquid particles suspended in a liquid), and foam (small gas bubbles dispersed in a solid or liquid). Although the most common aerosol systems are those with air as the suspending medium, other gaseous media are also important. The term aerosol is now used to denote any gas-borne particulate suspensions, including those in nitrogen, argon, helium, etc., and various complex gas mixtures, such as products of combustion from rockets, internal combustion engines, and coal or oil fired power plants, among others.

The importance of aerosols stems primarily from two facts: some particles are harmful and must be controlled, while others can be put to good practical use. Aerosols due to air pollution, for example, lead to visibility degradation and may be injurious to human health. Airborne lead, asbestos, and other harmful substances can lead to serious health consequences when inhaled and must be controlled to prevent overexposure. Atmospheric dust particles ingested by an engine, a gas turbine, or other power generating equipment can cause wear and the premature failure of the equipment. Particles depositing on semiconductor wafers during manufacturing of large-scale integrated circuit chips can cause defects, leading to device failure and loss of product yield. Beneficial applications of aerosols include production of useful materials such as carbon black, ceramic powders, glass fibers, etc., and use of aerosols for medical therapy or diagnosis. Understanding aerosol behavior and properties is important for such applications.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. AEROSOL PROPERTY AND BEHAVIOR

The size range of interest in aerosol studies is approximately 0.002–100 μm. The lower limit of 2 nm (0.002 μm) is an operational limit, based on the size of the smallest particle that can be detected by an aerosol measuring instrument, viz., the condensation nucleus counter. Smaller particles do exist, but are usually in the form of molecular clusters, and normally not considered in aerosol studies. Particles larger than 100 μm will fall out quickly and are of less interest in aerosol science.

1.1 Mechanical Properties

The mechanical behavior of an aerosol particle is governed by the Stokes law (see FLUID DYNAMICS, EQUILIBRIUM), modified by the inclusion of a "slip correction," C, to account for the effect of gas slip due to the finite mean free path λ_g of the gas molecules:

$$F_d = 6\pi\mu aV/C\,. \tag{1}$$

In Eq. (1), F_d is the drag force exerted by the gas on the particle, V is the velocity of the particle relative to the gas, μ is the gas viscosity, and a is the particle (assumed spherical) radius. The slip correction is a function of the Knudsen number, $\mathrm{Kn} = \lambda_g/a$,

$$C = 1 + 1.246\,\mathrm{Kn} + 0.42\,\mathrm{Kn}\,e^{-0.87/\mathrm{Kn}}. \tag{2}$$

When subjected to a constant force F, an aerosol particle will achieve a terminal velocity V relative to the gas. The ratio V/F is known as the dynamic mobility:

$$B = V/F = C/6\pi\mu a\,. \tag{3}$$

The electrical mobility Z_p of a charged particle is the ratio of the terminal particle velocity to the applied electric field E,

$$Z_p = V/E = n_p eC/6\pi\mu a\,, \tag{4}$$

where n_p is the number of elementary units of charge on the particle, and e is the elementary unit of charge. The time period in which the particle will respond to the applied force to achieve the terminal velocity is of the order of the particle relaxation time,

$$\tau = 2a^2\rho_p C/9\mu\,, \tag{5}$$

where ρ_p is the particle density. Small particles also undergo Brownian motion due to molecular impact by the surrounding gas molecules. This leads to diffusion (see VISCOSITY AND DIFFUSION), described by a particle diffusion coefficient, which is related to the dynamic particle mobility B by Einstein's equation,

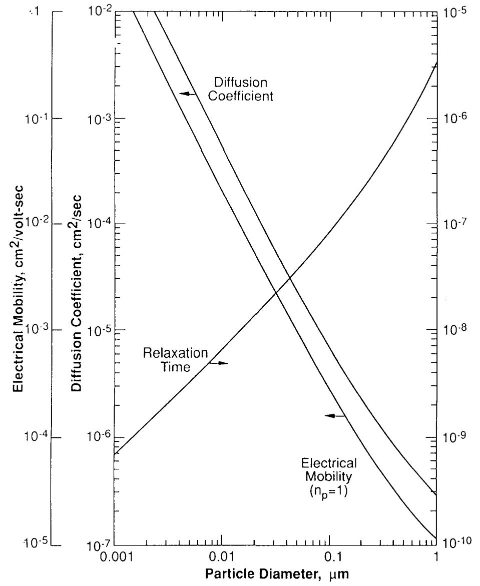

FIG. 1. Relaxation time, diffusion coefficient, and electrical mobility of submicrometer aerosols.

$$D = kTB = kTC/6\pi\mu a\,, \tag{6}$$

where k is Boltzmann's constant, and T is the absolute temperature. For large particles, gravitational settling is important. The settling speed of a particle in the Stokes law regime is given by

$$V_s = \tau g = 2a^2\rho_p Cg/9\mu\,, \tag{7}$$

where g is the acceleration of gravity. Figure 1 shows the relaxation time, electrical mobility, and diffusion coefficient of submicrometer particles and Fig. 2 shows the settling speed of particles in the >0.1 μm diameter range.

1.2 Electrical Properties

Aerosol particles found in the ambient air or generated in the laboratory are often charged. Important charging mechanisms include ion diffusion, ion drift in an electric field, and contact electrification. A high-voltage corona discharge will produce unipolar ions which can be used to charge the par-

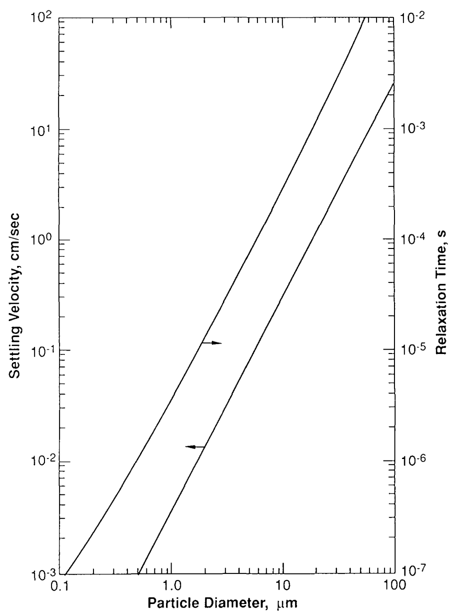

FIG. 2. Relaxation time and settling velocity of aerosols.

ticles by diffusion or ion drift, while surface contact is necessary for charging by contact electrification. Thus, particles reentrained from a solid surface or an aerosol formed by dispersing a powder into air are often charged due to contact electrification.

In the presence of ionizing radiation, the gaseous medium surrounding the particles may be ionized, leading to the appearance of positive and negative ions of approximately equal concentration in the gas. These ions will collide with the particles, causing the aerosol cloud to develop a bipolar charge. The steady-state bipolar charge distribution on the particles is given by the Boltzmann law and shown in Table 1.

Other mechanisms of charging include thermionic emission (*q.v.*) at high temperatures, photoelectric emission in the presence of light, gamma rays, or x rays, and induction charging when liquid sprays are formed in the presence of an electric field. The last mechanism is particularly important in industrial spray painting where charged paint droplets are produced by induction to increase the efficiency of paint utilization.

Figure 3 shows the aerosol charge levels for some of the charging mechanisms described above. Included in the figure is the Rayleigh limit for liquid droplets, which is the limiting charge at which the repulsive force due to the surface charge on the drop is just sufficient to overcome the surface tension to cause drop disintegration and, therefore, is the maximum charge a liquid drop can carry. The charge on atomized drops shown is based on the theory of Smoluchowski (Whitby and

Table 1. Boltzmann charge distribution on aerosols. (D_p is particle diameter.)

	Percent of particles carrying n_p elementary charge units								
$D_p(\mu m)$	$n_p = -4$	−3	−2	−1	0	+1	+2	+3	+4
0.01				0.34	99.32	0.34			
0.02				5.24	89.53	5.24			
0.03			0.04	11.57	76.79	11.57	0.04		
0.04			0.23	16.22	67.10	16.22	0.23		
0.05			0.64	19.30	60.11	19.30	0.64		
0.06		0.01	1.25	21.30	54.89	21.30	1.25	0.01	
0.07		0.03	1.98	22.58	50.82	22.58	1.98	0.03	
0.08		0.08	2.78	23.37	47.53	23.37	2.78	0.08	
0.09		0.15	3.59	23.85	44.82	23.85	3.59	0.15	
0.10	0.01	0.26	4.39	24.09	42.51	24.09	4.39	0.26	0.01
0.20	0.32	2.34	9.66	22.63	30.06	22.63	9.66	2.34	0.32
0.30	1.19	4.47	11.51	20.31	24.55	20.31	11.51	4.47	1.19
0.40	2.19	5.92	12.05	18.44	21.26	18.44	12.05	5.92	2.19
0.50	3.09	6.84	12.07	16.97	19.01	16.97	12.07	6.84	3.09
0.60	3.82	7.41	11.89	15.79	17.36	15.79	11.89	7.41	3.82
0.70	4.39	7.74	11.62	14.82	16.07	14.82	11.62	7.74	4.39
0.80	4.83	7.94	11.32	14.00	15.03	14.00	11.32	7.94	4.83
0.90	5.16	8.03	11.01	13.31	14.17	13.31	11.01	8.03	5.16
1.00	5.42	8.07	10.71	12.70	13.45	12.70	10.71	8.07	5.42

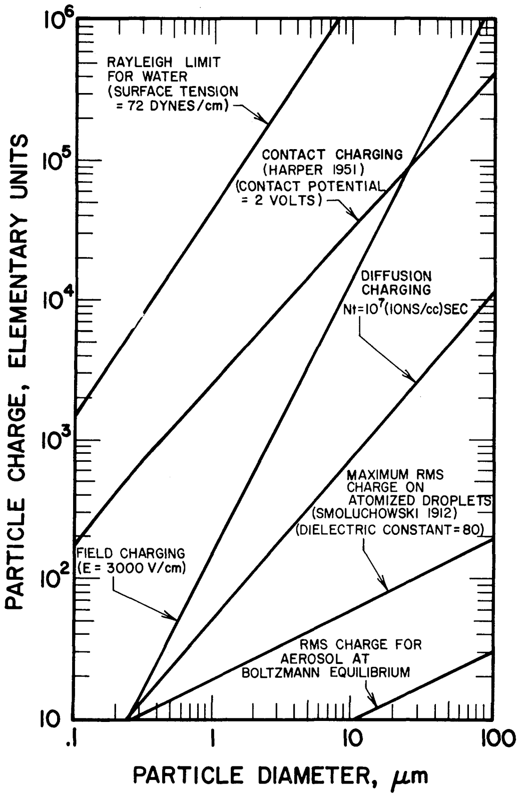

FIG. 3. Comparative charge level on aerosol particles.

Liu, 1966) for droplet charge resulting from the random fluctuation of ions in the liquid at the time of droplet formation during atomization.

1.3 Optical Properties and Light Scattering

The optical properties of aerosol particles depend on the sizes of the particles and the wavelength of light (see ELECTROMAGNETIC RADIATION). Light scattering by aerosol particles is responsible for many spectacular effects seen in the atmosphere. Colored sunsets, coronae, halos, and rainbows are all due to light scattering by atmospheric particles or water droplets. Calculations of light scattering by small spherical particles can be made by the Mie theory (Kerker, 1969). For small particles with radius a much smaller than the wavelength of light λ, the theory of Rayleigh scattering applies:

$$I_1 = \frac{8 I_0 \pi^4 a^6}{R^2 \lambda^4} \left(\frac{m^2 - 1}{m^2 + 2} \right)^2, \tag{8}$$

$$I_2 = \frac{8 I_0 \pi^4 a^6}{R^2 \lambda^4} \left(\frac{m^2 - 1}{m^2 + 2} \right)^2 \cos^2\theta, \tag{9}$$

for scattered intensities I_1 and I_2 polarized perpendicular and parallel to the scattering plane and for a receiver at the distance R from the scatterer and in the direction θ from the incident beam, where m is the refractive index of the particle. The light scattering by small particles is strongly dependent on light wavelength and particle size. The total scattering cross section of a particle increases as the sixth power of particle size and inversely as the fourth power of wavelength.

2. INSTRUMENTATION FOR AEROSOL STUDIES

Devices needed for aerosol studies include those for the generation, measuring, sampling, and analysis of airborne particles. Many such devices were developed in the 1970s and 1980s, in response to the need in air pollution studies. Figure 4 summarizes the principal devices used in aerosol studies. A comprehensive review of the topic has been given in two recent papers by Liu and Pui (1987) and Pui and Liu (1988). In this article, only selected instruments in each category will be briefly discussed to illustrate their operating principles and performance characteristics.

2.1 Aerosol Generators

The ability to generate an aerosol in a controlled and reproducible manner is an important requirement for laboratory aerosol studies. As shown in Fig. 4, a variety of aerosol generators have been developed to produce monodisperse (particles of the same size) or polydisperse (nonuniform sized) aerosols for experimental purposes. Two such devices are particularly important in basic aerosol science because they are used as standards of calibration for aerosol measuring instruments and for different experimental purposes.

The vibrating orifice monodisperse aerosol generator, described by Berglund and Liu (1973), and shown schematically in Fig. 5, is based on the controlled disintegration of a liquid jet through a small vibrating orifice to produce uniform droplets. The jet is formed

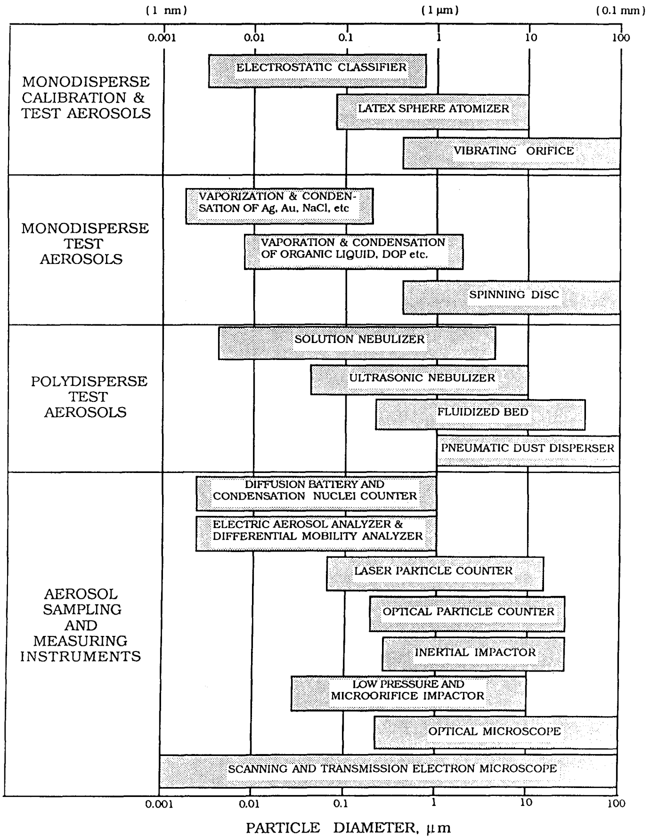

FIG. 4. Size range of aerosol generating, measuring, and sampling instruments.

by means of a syringe pump which forces the liquid through the small orifice which is vibrated at a suitable frequency by a piezoelectric ceramic. The jet is broken up into uniform droplets by the vibration. To prevent plugging of the orifice, which is usually quite small (5–20 μm), the upstream passage of the orifice must be flushed with pure solvent by removing the plug before the start of the jet.

The vibrating orifice generator produces uniform droplets of a known volume V_d, which is related to the vibrating frequency f and the liquid flow rate Q by the equation $V_d = Q/f$. By atomizing a solution of the aero-

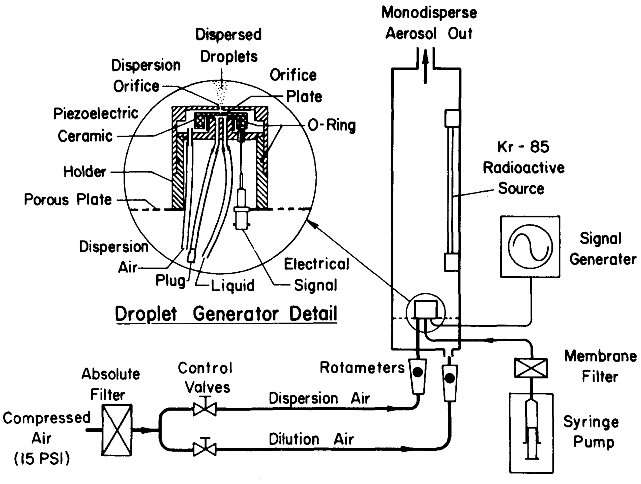
Monodisperse Aerosol Out
Dispersed Droplets
Dispersion Orifice
Orifice Plate
Piezoelectric Ceramic
O-Ring
Holder
Porous Plate
Dispersion Air
Plug
Liquid
Electrical Signal
Droplet Generator Detail
Kr - 85 Radioactive Source
Signal Generater
Membrane Filter
Syringe Pump
Absolute Filter
Control Valves
Rotameters
Dispersion Air
Dilution Air
Compressed Air (15 PSI)

FIG. 5. The Berglund–Liu vibrating orifice monodisperse aerosol generator.

sol material in a volatile solvent, and allowing the solvent to evaporate from the droplets, a monodisperse aerosol of diameter $D_p = (6QC/\pi f)^{1/3}$, is obtained, where C is the volumetric concentration of the dissolved material in the solution. The particle diameter can thus be determined accurately by measuring the quantities Q, f, and C. Some solid and liquid particles produced by the device are shown in Fig. 6. The device is capable of generating monodisperse aerosols from less than 1 μm to about 30 μm in diameter. With simple modifications, particles as large as several hundred micrometers have been generated by the principle. The device is particularly useful for generating aerosols of a known size for calibrating optical particle counters, inertial impactors, and other particle measuring instruments, and for fundamental studies such as those concerning turbulent deposition, electrical charging, and neutralization, etc.

To produce submicrometer aerosols, the electrostatic classification method of Liu and Pui (1974) shown in Fig. 7 is widely used. The system includes an atomizer to produce a polydisperse aerosol which is then classified electrically according to mobility to produce a monodisperse aerosol. The particle size can be calculated from the known electrical mobility of the particles, and the particle concentration can be measured by collecting the charged particles by filtration and measuring the resulting current flow with an electrometer. Making use of the fact that the particles are singly charged, the aerosol concentration can be determined by the equation $N = I/eq$, where I is the electrometer current, e is the elementary unit of charge, and q is the volumetric flow rate of the aerosol. The device has been used for generating monodisperse aerosols in the 0.008 to 1 μm diameter range to calibrate condensation nucleus counters, diffusion batteries, the electrical aerosol analyzer, etc., and for various experimental aerosol studies.

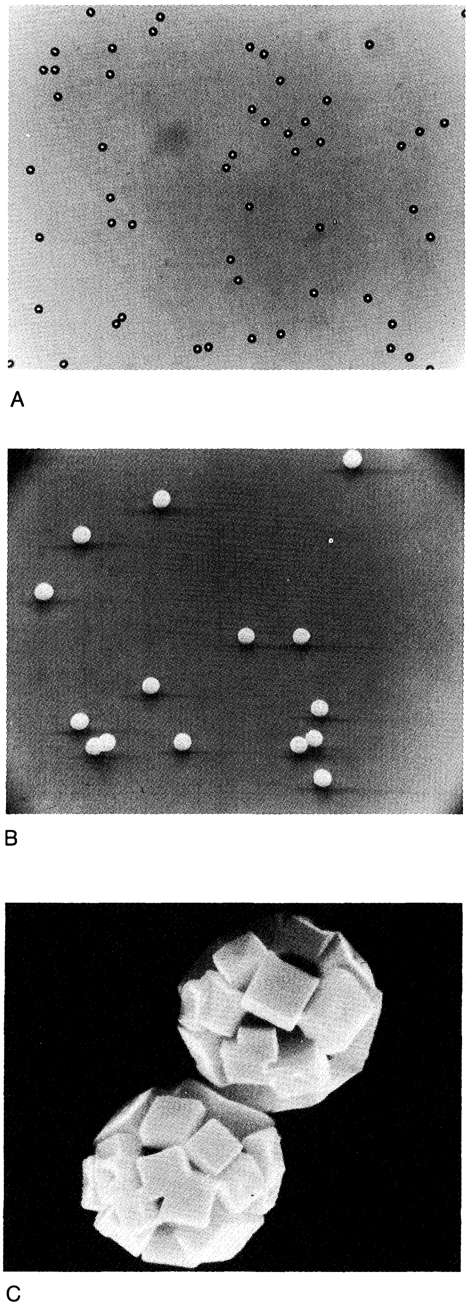

FIG. 6. Particles generated by the Berglund–Liu generator; particles are approximately 4, 5, and 10 μm in diameter (top to bottom).

2.2 Aerosol Measuring Instruments

Automatic instruments for concentration and size distribution measurement of aerosols are available to cover a wide range of particle sizes from 0.002 to approximately 50 μm. These instruments have fast time response and can follow rapid changes in both particle size and concentration. Good counting statistics can usually be obtained with automatic instruments because repeated measurements on a large number of particles will yield statistically accurate results. On the other hand, automatic instruments usually

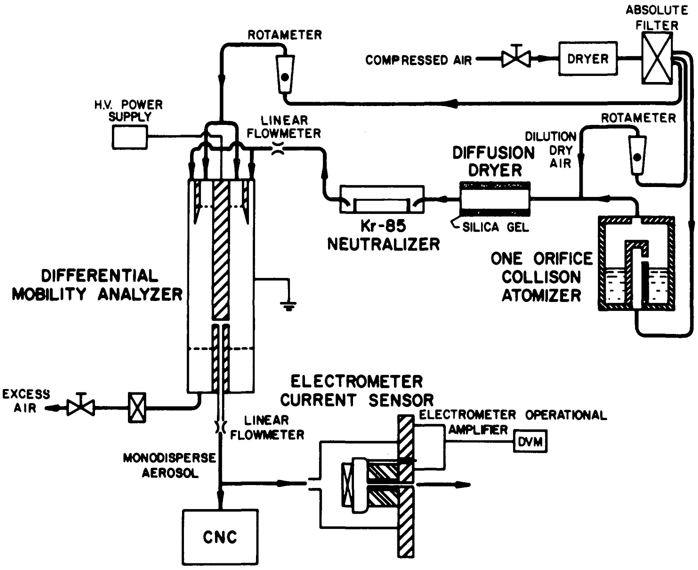
ABSOLUTE FILTER
ROTAMETER
COMPRESSED AIR
DRYER
H.V. POWER SUPPLY
LINEAR FLOWMETER
ROTAMETER
DILUTION DRY AIR
DIFFUSION DRYER
Kr-85 NEUTRALIZER
SILICA GEL
ONE ORIFICE COLLISON ATOMIZER
DIFFERENTIAL MOBILITY ANALYZER
ELECTROMETER CURRENT SENSOR
ELECTROMETER OPERATIONAL AMPLIFIER
DVM
EXCESS AIR
LINEAR FLOWMETER
MONODISPERSE AEROSOL
CNC

FIG. 7. Electrostatic classifier aerosol generator of Liu and Pui (1974), shown being used for CNC (condensation nucleus counter) calibration by comparison with an electrometer current sensor with a digital voltmeter (DVM).

rely on indirect sensing techniques and more calibration effort is usually required.

The optical particle counter, shown schematically in Fig. 8, is widely used for particle measurement in both the indoor—as in a cleanroom—and outdoor environments. Single, individual particles are carried by an airstream into the instrument and passed through an illuminated viewing volume where light is scattered by the particle as it traverses the volume. The scattered light is then sensed by a photomultiplier or a solid-state photodetector. The output of the sensor is in the form of a voltage pulse corresponding to the passage of the particle through the viewing volume. By counting the number of pulses and measuring the corresponding pulse amplitude, both the number and size of particles can be determined. Traditional counters using an incandescent light source can measure particles to approximately 0.3 μm in diameter. Counters using gas or solid-state laser sources can measure smaller particles—to approximately 0.05 μm in diameter. The upper size limit of the instrument is usually about 10 μm due to sampling limitations—particles much larger than 10 μm usually do not follow the streamlines faithfully into the sensing volume of the instrument for them to be counted. Calibration studies by Liu *et al.* (1974, 1985, 1986b), Szymanski and Liu (1986), and others show that

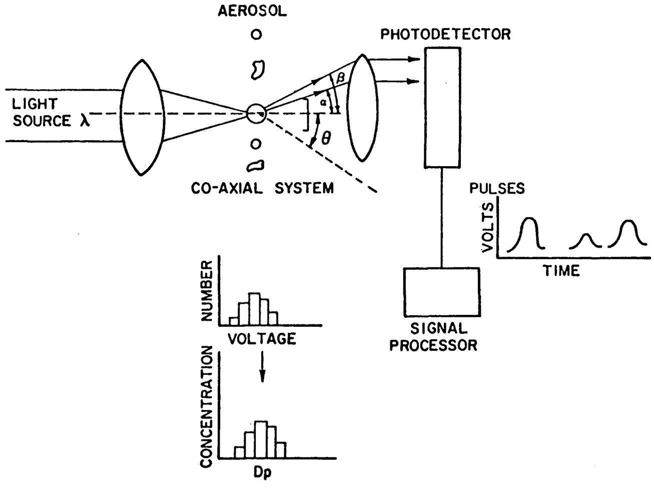
AEROSOL
PHOTODETECTOR
LIGHT SOURCE λ
β
α
θ
CO-AXIAL SYSTEM
PULSES
VOLTS
TIME
NUMBER
VOLTAGE
SIGNAL PROCESSOR
CONCENTRATION
Dp

FIG. 8. Schematic diagram of optical particle counters; α, β, and θ indicate major angles affecting instrument reponse to particles of a specific size with a specific wavelength, λ, light source.

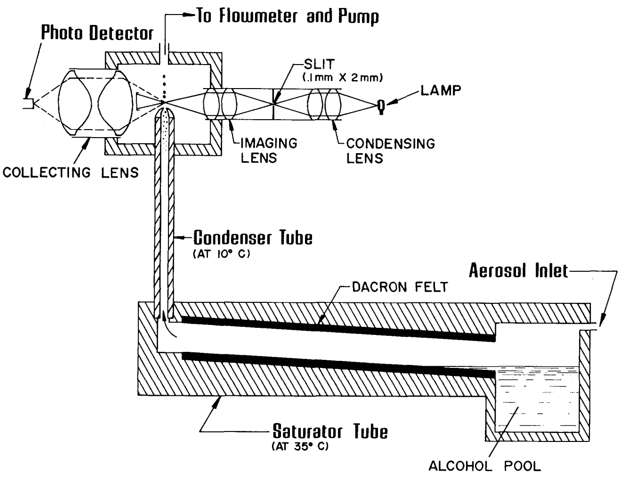

FIG. 9. Schematic diagram of the continuous flow condensation nucleus counter.

the influence of refractive index and particle size on counter response can be predicted accurately by Mie scattering calculations.

The principal instrument for counting particles smaller than 0.1 μm is the condensation nucleus counter (CNC). A schematic diagram of a commercially available CNC, described by Agarwal and Sem (1980), is shown in Fig. 9. The aerosol to be measured is first saturated with alcohol vapor by passing through a heated butyl alcohol saturator maintained at 35 °C, and then cooled in a thermoelectrically cooled condenser to 10 °C. This causes the vapor to become supersaturated and condense on the particles. Particles that are too small to be detectable by direct light scattering can then be detected by measuring the grown drops, which are usually about 12 μm in diameter. Either the particles can be counted individually or, in the case of high aerosol concentrations, the light scattering from the entire droplet cloud in the viewing volume can be measured in a "photometric mode" to determine the droplet concentration. Particle concentration in the range of <0.01 to more than 10^6 particles/cm^3 can be measured. The lower detection limit of the standard commercially available instruments is in the range of 0.005–0.01 μm. Recent advances in instrument design (Stolzenburg and McMurry, 1984) have led to instruments capable of measuring small, ultrafine particles to as small as 0.003 μm in diameter.

Since the CNC measures only the particle concentration, a different approach is needed to measure particle size. This is often done by combining the CNC with a size-selective filter in the form of a diffusion battery. By passing aerosols through metal screens of different sizes, particles with sizes below a selectable limit can be removed. Using a series of such diffusion screens and the CNC, the size distribution of the aerosol can be measured.

Another approach to sizing submicrometer aerosols is the electrical aerosol analyzer (EAA) described by Liu and Pui (1975) and shown schematically in Fig. 10. The instrument operates by first charging the aerosol particles, using unipolar ions produced by a corona discharge, and then measuring the electrical mobility of the charged particles. From the measured electrical mobility distribution and knowing the relationship between particle size and mobility by calibration, the size distribution of the particles can be determined. The instrument has been used for air pollution studies and for measuring the characteristics of pollution control devices, such as the electrostatic precipitator, bag-house filter, etc.

2.3 Aerosol Samplers

Sampling instruments for aerosols are particularly useful when particles are to be collected for subsequent analysis. The collection can be done with a simple filter to give an integrated sample for time-averaged mass or chemical measurement, or with an impactor or cyclone to give size-fractionated samples for mass concentration and chemical composition studies as a function of particle size. Analysis of a collected sample can be done by weighing with a microbalance, wet-chemical

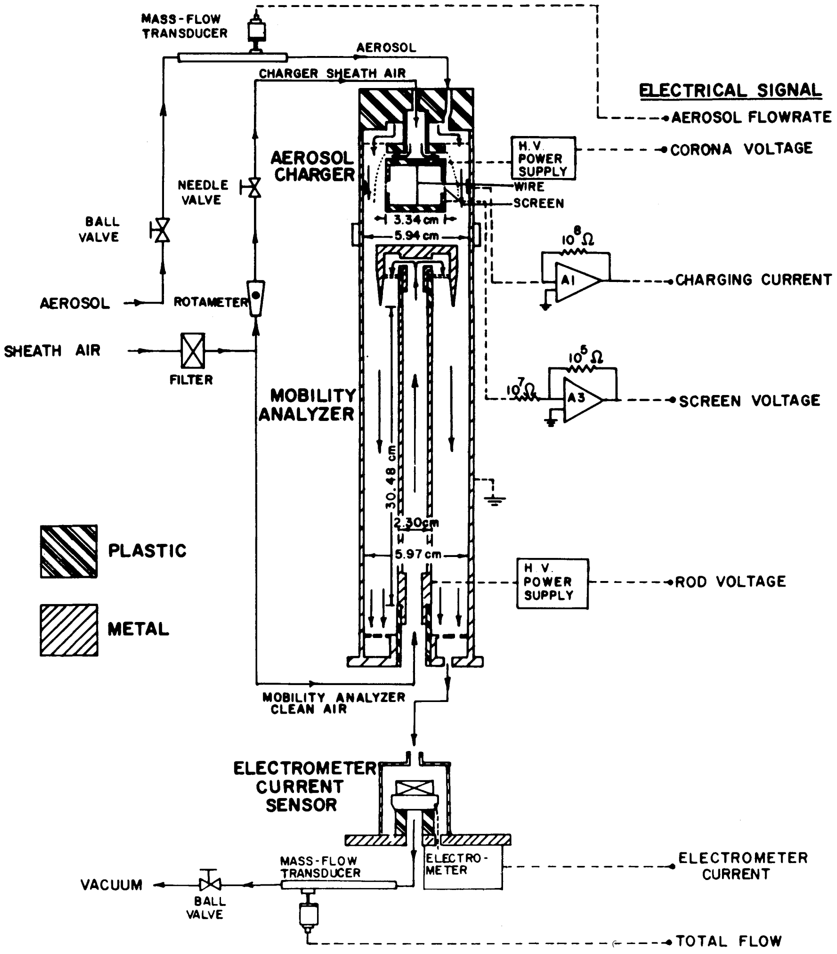
MASS-FLOW TRANSDUCER
AEROSOL
CHARGER SHEATH AIR
ELECTRICAL SIGNAL
AEROSOL FLOWRATE
H.V. POWER SUPPLY
CORONA VOLTAGE
AEROSOL CHARGER
WIRE
SCREEN
NEEDLE VALVE
BALL VALVE
3.34 cm
5.94 cm
$10^8\,\Omega$
A1
CHARGING CURRENT
AEROSOL
ROTAMETER
SHEATH AIR
FILTER
$10^6\,\Omega$
$10^7\,\Omega$
A3
SCREEN VOLTAGE
MOBILITY ANALYZER
30.48 cm
PLASTIC
METAL
2.30 cm
5.97 cm
H.V. POWER SUPPLY
ROD VOLTAGE
MOBILITY ANALYZER CLEAN AIR
ELECTROMETER CURRENT SENSOR
MASS-FLOW TRANSDUCER
ELECTROMETER
ELECTROMETER CURRENT
VACUUM
BALL VALVE
TOTAL FLOW

FIG. 10. Electrical aerosol analyzer (Liu and Pui, 1975).

analysis following sample extraction, or various nondestructive chemical analysis techniques, such as energy disperse x-ray (EDX), x-ray fluorescence (XRF), and proton induced x-ray emission (PIXE) analysis, among others.

The cascade impactor, shown schematically in Fig. 11, is particularly useful and widely used for size fractionating of airborne particles. A single-stage impactor, consisting of a nozzle directing a high-velocity jet against a solid surface, will cause particles above the stage "cut size" to impact on the surface while allowing smaller particles to pass through. This occurs because large particles traveling towards the impaction plate will have sufficient momentum to impact on the plate as the airflow is deflected sideways by the plate. In the meantime, small particles with insufficient momentum will follow the air flow and be carried by the airstream around the plate, thus avoiding impaction. The critical particle size at which this occurs is known as the stage cut size.

The stage cut size of an impactor can be varied by varying the nozzle diameter and the airflow velocity. By arranging several of these impactor stages in series with progressively smaller nozzles and higher jet velocities, progressively smaller particles can be removed by impaction. By this means, a polydisperse aerosol composed of particles covering a wide range of sizes can be size-fractionated into samples each covering only a narrow size range. The size-fractionated samples can then be analyzed by weighing to determine their mass, or by x-ray fluorescence, atomic

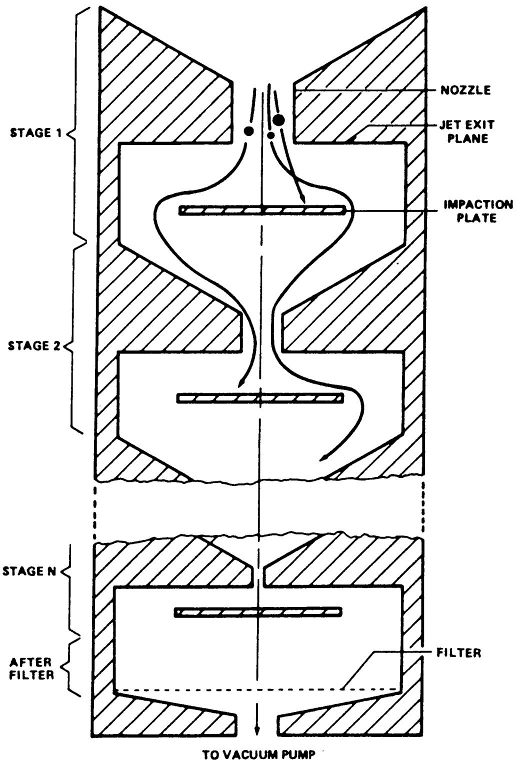

FIG. 11. Schematic diagram of cascade impactor.

absorption, etc., to determine their elemental chemical composition.

The lower size limit of the conventional impactor is about 0.3 μm. Smaller particles can be impacted by means of small nozzle diameters, or high jet velocities, or both. A recent development by Marple *et al.* (1981), using micro-orifice nozzles, allows the cut size to be extended to 0.05 μm. A typical micro-orifice impactor will have six to eight impactor stages and a final filter stage. The filter collects all particles below the smallest cutoff size of the impactor stage. A picture of the Micro-Orifice Uniform Deposit impactor described by Marple *et al.* is shown in Fig. 12.

Impactor samplers require particles to adhere to the surface upon impact. Since this may not occur for dry solid particles, the surface usually must be coated with an adhesive layer to prevent particle bounce. Another approach involves using a virtual impactor (Loo *et al.*, 1979), shown schematically in Fig. 13. In the virtual impactor, the particle-bearing airstream exiting the nozzle is directed at the void space of a receiving tube. Only a small fraction (about 10%) of the airflow is allowed to flow through the receiving tube, while the remainder is deflected through a side passage as shown. Particles larger than the cut size of the virtual impactor stage will not follow the major airflow into the side passage, but will impact into the receiving tube as a result of their momentum. These particles are then carried by the small airflow to a filter located downstream. Particle bounce therefore can be avoided. A single virtual impactor will separate particles into two size fractions, each fraction being collected on a separate filter. By arranging several of these virtual impactor stages in series, size-fractionated samples can be obtained.

3. SCIENTIFIC AND TECHNICAL APPLICATIONS

3.1 Lung Deposition and Respiratory Protection

Small airborne particles can be inhaled into, and be deposited in, the lung to cause damage or produce harmful health effects. The ability of an aerosol to enter the human respiratory system is governed by the aerodynamic particle size, defined as the size of a unit-density sphere having the same settling

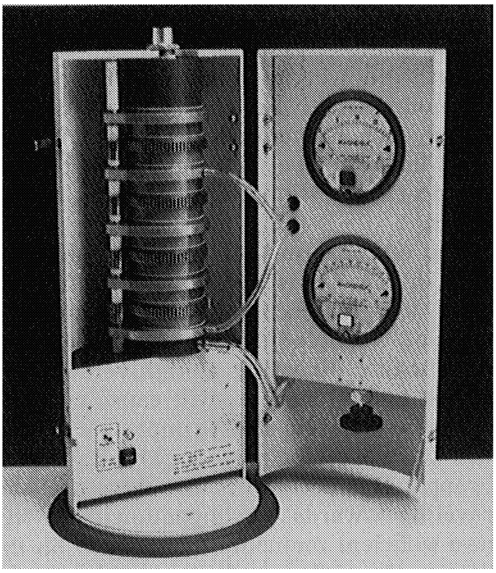

FIG. 12. The Micro-Orifice Uniform Deposit Impactor.

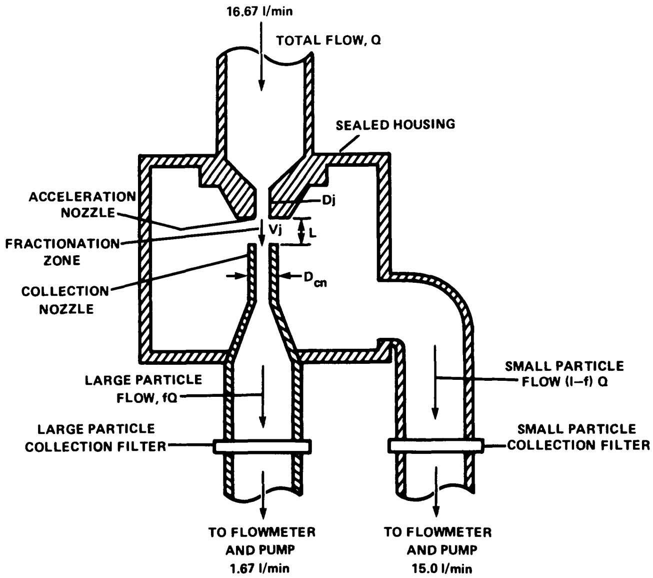
16.67 l/min
TOTAL FLOW, Q
SEALED HOUSING
ACCELERATION NOZZLE
Dj
FRACTIONATION ZONE
Vj
L
COLLECTION NOZZLE
D_cn
LARGE PARTICLE FLOW, fQ
LARGE PARTICLE COLLECTION FILTER
SMALL PARTICLE FLOW (1–f) Q
SMALL PARTICLE COLLECTION FILTER
TO FLOWMETER AND PUMP 1.67 l/min
TO FLOWMETER AND PUMP 15.0 l/min

FIG. 13. The virtual impactor particle classifier.

velocity as the particle in question. The entire human respiratory system can be divided into three regions: the head airways region, the tracheobronchial region, and the alveolar region (Hinds, 1982). Particles smaller than 100 μm will enter the nose during breathing, but most will deposit in the head region. Only those smaller than 10 μm can enter the trachea and those smaller than about 3.5 μm can reach the deep, gas exchange region of the lung. The ACGIH (American Conference of Governmental Industrial Hygienists) Air Sampling Procedures Committee (Phalen *et al.*, 1986) has established the size-dependent respiratory penetration curves shown in Fig. 14 to describe particle penetration into the human respiratory system, and has recommended measuring three mass fractions for assessing airborne inhalation hazard: the inspirable particulate mass (IPM) for materials which may be hazardous anywhere in the respiratory tract, thoracic particulate mass (TPM) for materials which may be hazardous anywhere within the lung airways and the gas exchange region, and respiratory particulate mass (RPM) for materials which may be hazardous in the gas exchange region of the lung. The U.S. Environmental Protection Agency has recently promulgated a PM-10 standard (particulate matter, 10 μm cut) for the thoracic particle concentration in the revised particulate air quality standard for ambient air. Devices for ambient PM-10 sampling have been described by Liu and Pui (1986) and McFarland *et al.* (1984).

In the workplace environment where airborne particulate concentration can reach unhealthy levels, respirators are often used to protect workers against particulate exposure. One of the earliest aerosol studies was in connection with developing respirators to protect miners and firemen. Early pioneers, such as Robert Brown in 1827 and John Tyndall in 1870, developed respirators using aerosol technologies. However, the most rapid advances made in respirator design were during World War I to develop respirators to protect personnel against chemical warfare agents (Pritchard, 1976).

Modern respirators come in a variety of shapes and configurations to protect the wearer against inhalation of harmful substances, including particles, gas, and vapors. Figure 15 shows the general classification of respirators according to whether they are atmosphere supplying or air purifying. The former is further divided into two major subcategories, self-contained breathing apparatus and air supplied respirator; and the latter into powered and nonpowered. There are further subclassifications according to their specific design, intended applications, or ability to remove gas, vapor, and/or particles.

An increasingly popular form of respirator for the workplace environment and for general use is the disposable mask made of fine

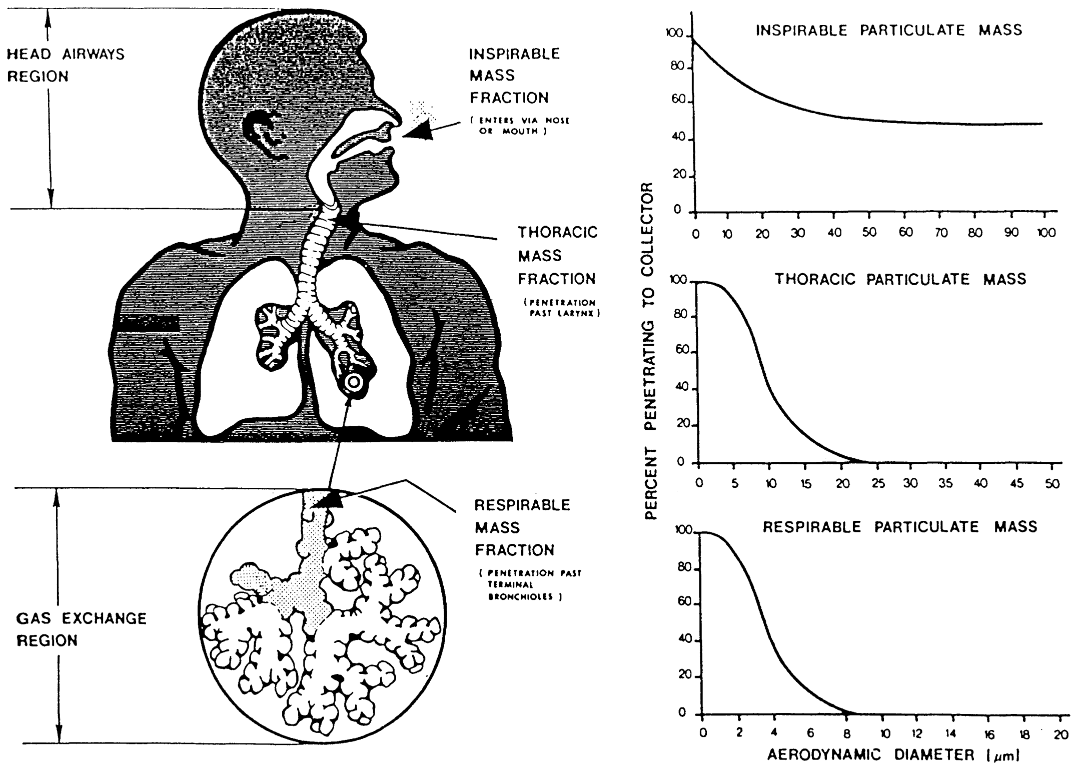
HEAD AIRWAYS REGION
INSPIRABLE MASS FRACTION
(ENTERS VIA NOSE OR MOUTH)
THORACIC MASS FRACTION
(PENETRATION PAST LARYNX)
RESPIRABLE MASS FRACTION
(PENETRATION PAST TERMINAL BRONCHIOLES)
GAS EXCHANGE REGION
PERCENT PENETRATING TO COLLECTOR
INSPIRABLE PARTICULATE MASS
100
80
60
40
20
0
0 10 20 30 40 50 60 70 80 90 100
THORACIC PARTICULATE MASS
0 5 10 15 20 25 30 35 40 45 50
RESPIRABLE PARTICULATE MASS
0 2 4 6 8 10 12 14 16 18 20
AERODYNAMIC DIAMETER [μm]

FIG. 14. Aerosol penetration into the human respiratory system and definitions of inspirable, thoracic, and respirable aerosols recommended by the ACGIH Air Sampling Procedures Committee (Phalen *et al.*, 1986).

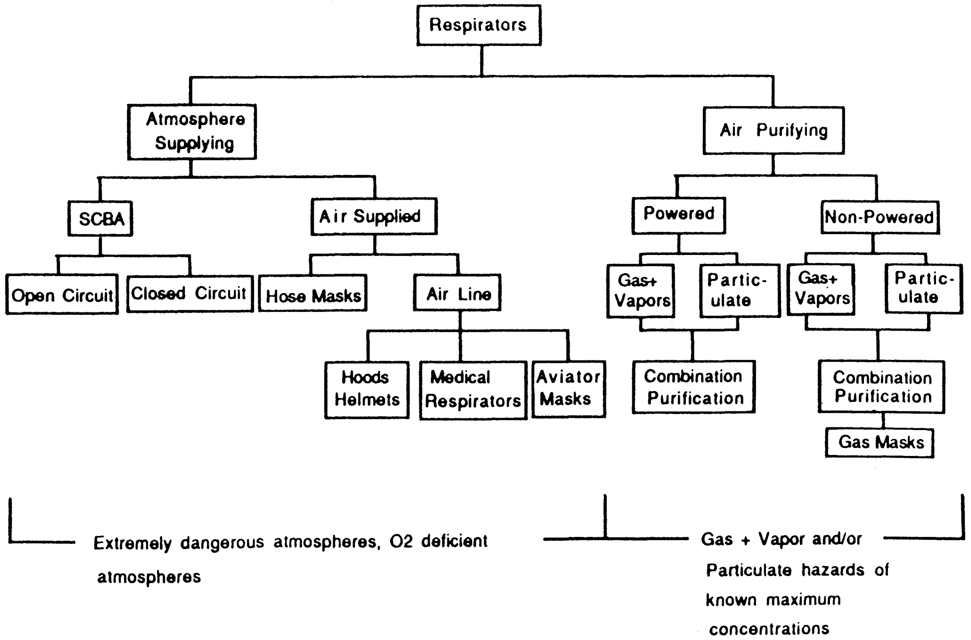
Respirators
Atmosphere Supplying
Air Purifying
SCBA
Air Supplied
Powered
Non-Powered
Open Circuit
Closed Circuit
Hose Masks
Air Line
Gas+ Vapors
Partic- ulate
Gas+ Vapors
Partic- ulate
Hoods Helmets
Medical Respirators
Aviator Masks
Combination Purification
Combination Purification
Gas Masks
Extremely dangerous atmospheres, O2 deficient atmospheres
Gas + Vapor and/or Particulate hazards of known maximum concentrations

FIG. 15. Respirator classification.

organic or glass fiber mats. A recent development is the dielectric fibers carrying a permanent dipole electric charge. The charge density is much higher than in the classical resin-impregnated wool fibers and gives stable performance at high temperature and in humid environments. On account of their high efficiency and low breathing resistance, the filters have found increasing use as the respirator filter media. A fundamental study of disposable respirators has been reported by Fardi and Liu (1989). A photograph of a disposable respirator face mask and a cartridge-type full face mask is shown in Fig. 16.

3.2 Atmospheric Aerosols and Air Pollution

Aerosols in the ambient atmosphere are complex mixtures of particles from a variety of sources. The principal sources include those that are natural, such as windblown dust, sea salt sprays, etc., and those that are man-made, such as direct emission from industrial plants and vehicular emissions. Gaseous emissions from both natural and man-made sources can also be converted into particles in the atmosphere through photochemical reactions. H_2S, NH_3, and other organic and inorganic species produced by decaying vegetation and hydrocarbon emissions from industrial plants and motor vehicles can react in the atmosphere under sunlight to form condensable species and aerosol particles (see AIR POLLUTION). Table 2 shows the typical atmospheric concentrations of particles in urban and rural areas (Seinfeld, 1986).

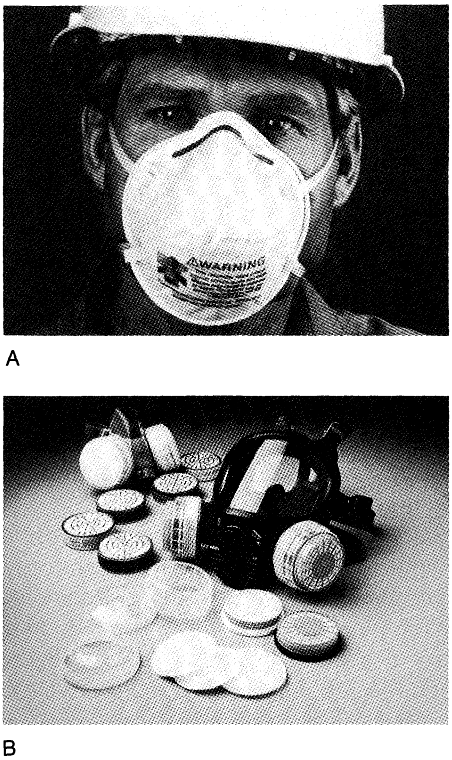

FIG. 16. Disposable and cartridge-type respirators (courtesy 3M Co.).

The size distribution of atmospheric aerosols is usually bimodal in shape with particles in the coarse (>2.5 μm) mode being primarily of soil origin, such as oxides of aluminum, iron, and silicon, and those in the fine particle mode consisting of products of com-

Table 2. Typical concentration of atmospheric particulate matter. [TSP is total suspended particulates (Seinfeld, 1986).]

Site	Date	Measurement	Concentration (μg m^{-3})
New York City	1977 (50 days)	TSP	70.5
		SO_4^{2-}	11.58
		NH_4^+	2.38
		NO_3^-	4.96
South Coast Air Basin, CA (Rubidoux)	1972–3 (24 hr. avg.)	SO_4^{2-}	10.7
		NO_3^-	31.5
Great Smoky Mountains (Elkmont, TN)	20–26 Sept. 1978	TSP(fine)	24
Washington, DC	Summer 1976	TSP	65

bustion or photochemical reaction. This bimodal size distribution is shown in Fig. 17 and illustrated in Fig. 18 for the distribution of pollutant species in the aerosol (Whitby, 1978).

3.3 Industrial Gas Cleaning

Particles suspended in a gas often need to be removed for purposes of product recovery, equipment protection, or emission control to meet environmental regulations. Important equipment used for industrial gas cleaning includes fibrous and fabric filters and electrostatic precipitators.

Fibrous and fabric filters remove particles by capturing the particles by adhesion when the particles are brought into physical contact with the fiber by the airflow. The overall efficiency of a fibrous filter is shown schematically in Fig. 19. For small particles, the diffusion mechanism is important, whereas for large particles, the interception and impaction mechanisms dominate. The minimum in the curve of filter efficiency versus particle size corresponds to the transition from the diffusion dominated regime to the interception and inertial impaction dominated filtration regimes. This minimum usually occurs in the 0.1 to 0.3 μm diameter range. Figure 20 (Liu *et al.*, 1985) shows the penetration (1 – efficiency) of the aerosol through a HEPA (high efficiency particulate air) filter and an ULPA (ultralow penetration air) filter. These filters are used extensively for cleanroom air filtration and they have efficiencies in excess of 99.99% in the most penetrating particle size range.

A popular form of industrial gas filter is the pulse-cleaned cartridge dust collector shown schematically in Fig. 21. The filter is in the form of a pleated paper element made into a cylindrical cartridge. A number of these cartridges are placed in the collector to provide the required flow rate. Particles are collected on the outside of the cylindrical cartridges as the airflow passes through them and these particles can be dislodged from the outside surface by a pulse of high-pressure air and drop into a dust bin below. The cartridge filters have efficiencies in excess of 99.999% for dry dust particles. Table 3 shows the performance of nine industrial dust collectors (Bergin *et al.*, 1989). The high efficiencies of these filters have made it possible to recirculate the treated air into the workplace environment for energy conservation and other purposes.

For high volume gas filtration requiring low pressure drop, the electrostatic precipitator invented by F. G. Cottrell (White, 1963) is widely used. A typical single stage electrostatic precipitator is shown in Fig. 22. It consists of a discharge electrode to produce unipolar negative ions, which charge the particle as they flow through the corona discharge region. The charged particles are then precipi-

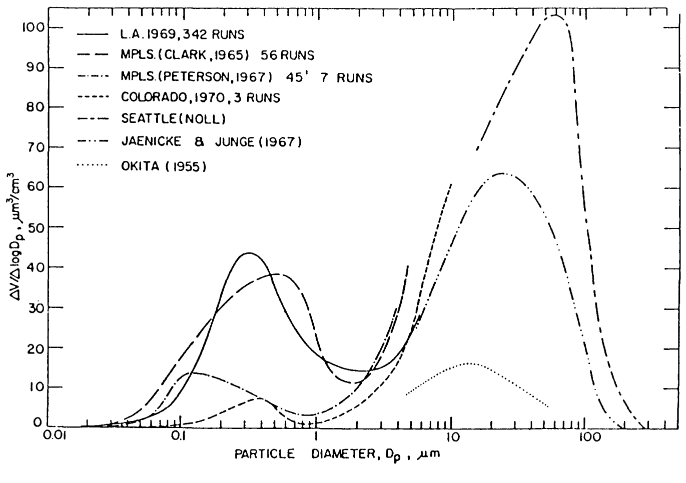

FIG. 17. The bimodal size distribution of atmospheric aerosols (Whitby, 1978).

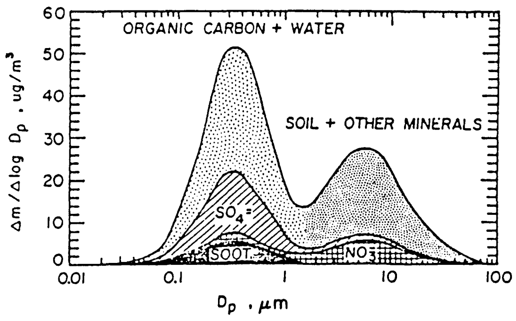

FIG. 18. Distribution of pollutant chemical species in the aerosol modes.

tated onto the outer electrode by the electric field. Electrostatic precipitators are used extensively for fly ash removal from coal-fired power plants and for various industrial applications. Theory of particle charging due to ion drift in the corona discharge region has been developed by Liu and Kapadia (1978).

Other important industrial gas cleaning devices include cyclones and Venturi scrubbers (Strauss, 1966). Cyclones operate by imparting a rotational motion to the gas to cause the particles to be separated by centrifugal force. The rotational gas motion can be created by a set of turning vanes or by introducing the gas into the cylindrical cyclone body in a tangential direction. Venturi scrubbers operate by accelerating the gas flow in a Venturi tube to a high velocity in the Venturi throat, to atomize water and cause particle impaction on the atomized drops. The water drops are then removed along with the collected particles by a cyclone or a mist eliminator.

3.4 Cleanrooms and Microelectronics

Clean manufacturing technologies for microelectronics constitutes a fast growing area for the application of aerosol science. The manufacturing of integrated circuit chips and data storage devices, such as those based on magnetic thin films, is a multibillion dollar industry that requires particulate contamination control, including the control of airborne contaminants in the form of aerosols (see SEMICONDUCTOR DEVICE PROCESSING). The yield and reliability of the manufactured products depend critically on contamination control and the avoidance of particulate contamination due to people, machinery, and process generated contaminants.

Due to device miniaturization, the critical contaminant size causing device failure also becomes smaller. Figure 23 (Larrabee, 1985) shows a sequence of photos comparing the semiconductor wafer, the integrated circuit chips on the wafer, and typical device feature size, to an insect—a ladybug. The state-of-the-art 1-megabit DRAM (dynamic random access memory) chip capable of storing one million bits of data has typical device feature sizes of approximately 0.9 μm (Cooper, 1986). The next-generation 4-megabit chips will have feature sizes of the order of 0.7 μm. Contaminant particles down to 1/10 of the characteristic device feature size are usually consid-

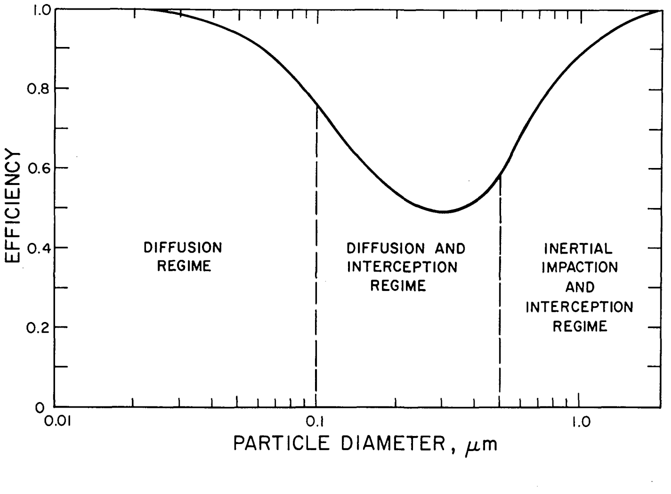

FIG. 19. Schematic diagram of filter efficiency and capture mechanisms.

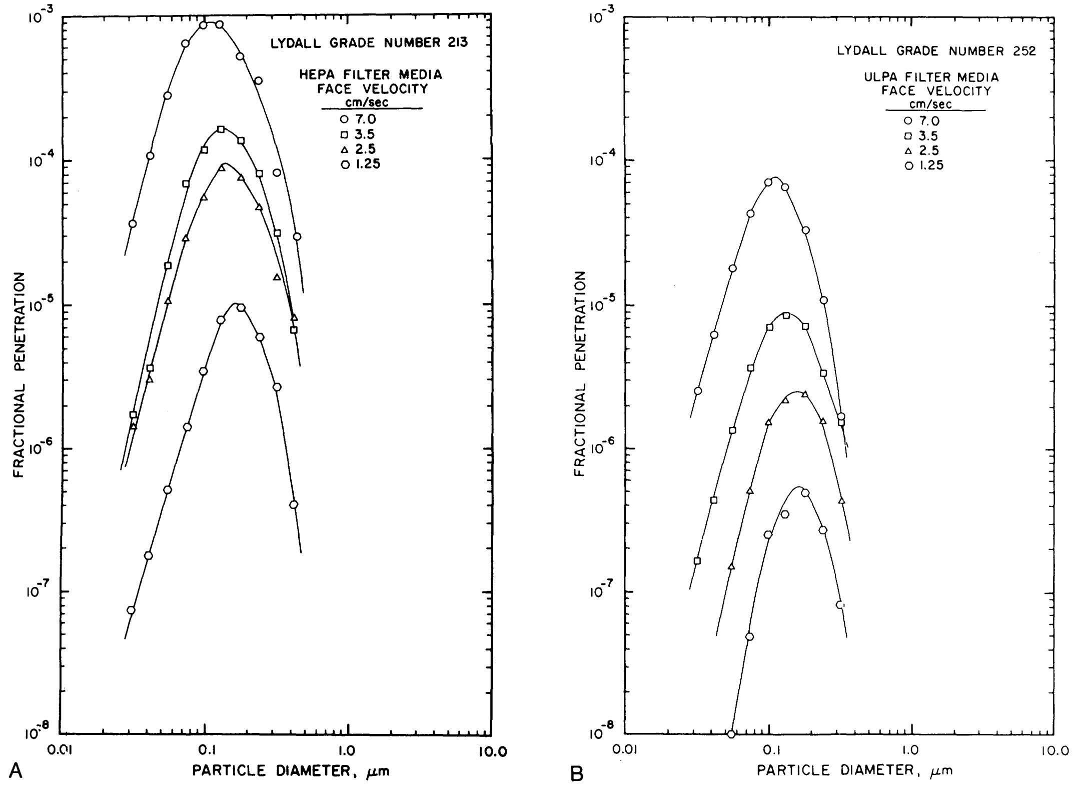

FIG. 20. Aerosol penetration through HEAP (high efficiency particulate air) and ULPA (ultralow penetration air) filters (Liu *et al.*, 1985).

ered harmful and must be avoided. The present concern is for particles down to approximately 0.1 μm. Future generations of integrated circuit chips will require the control of particles of all sizes down to the lower size limit of aerosol science, viz., 2 nm.

For magnetic storage disk drives, the read-write head is supported on a thin air film above a rotating disk. The spacing between the head and the rotating disk surface is currently about 0.2–0.3 μm. Figure 24 shows the flying height of the head as compared to the

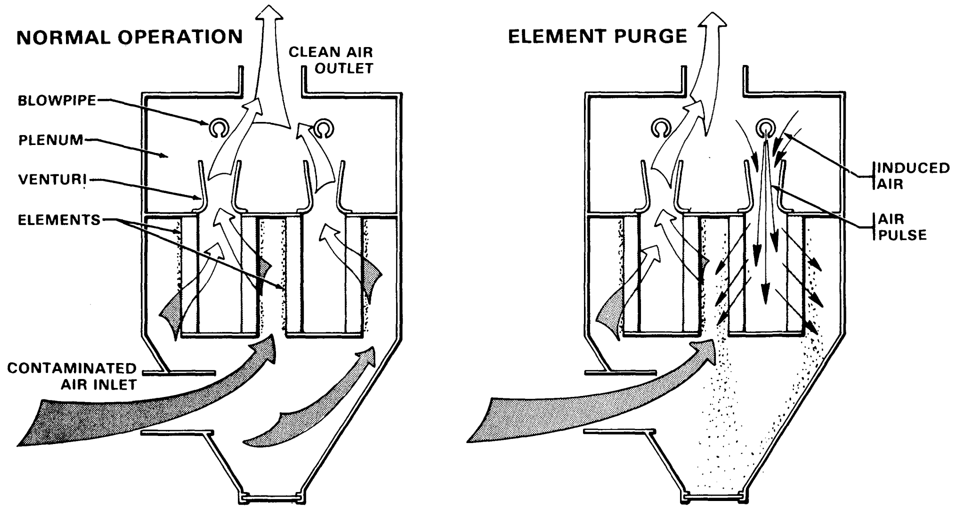
NORMAL OPERATION
CLEAN AIR OUTLET
BLOWPIPE
PLENUM
VENTURI
ELEMENTS
CONTAMINATED AIR INLET
ELEMENT PURGE
INDUCED AIR
AIR PULSE

FIG. 21. Schematic diagram of cartridge-type dust collector (Liu *et al.*, 1986a).

Table 3. Performance of industrial dust collectors.

(1) Designation	(2) Collector type	(3) Flow rate (cfm)	(4) Face velocity (ft/min)	(5) SS[a] time (hr)	(6) Pressure drop (in. H_2O)	(7) SS[a] fractional penetration $D_p = 0.5\ \mu m$	(8) 2 μm	(9) 8 μm	(10) Respirable[b] mass (mg/m^3)	(11) Figure of merit (1/in. H_2
A1	Cartridge HP[c]	3000	2.22	230	2.2	2(−5)[d]	3(−6)	4(−6)	0.005	2.2
A2	Cartridge HP	3000	1.00	420	1.8	3(−4)	2(−6)	1(−6)	0.002	3.0
A3	Cartridge HP	3000	···	380	1.4	2.5(−5)	4(−6)	5(−6)	0.004	3.5
B1	Baghouse HP	3000	8.85	170	8.0	1.2(−5)	1.2(−5)	1.5(−5)	0.02	0.6
B2	Baghouse HP	3000	10.6	20	7.0	1(−5)	1(−5)	2(−5)	0.03	0.7
B3	Baghouse HP	3000	6.94	310	6.0	1(−4)	1(−4)	4(−5)	0.07	0.7
B4	Baghouse LP[e]	3000	8.51	360	6.5	1.5(−3)	1(−3)	6(−4)	1.0	0.4
C1	ESP 2-stage	2500	···	<1	<0.5	2(−2)	4(−3)	2(−3)	0.3	···
C2	ESP 1-stage	2500	···	<1	<0.5	···	2.5(−1)	7(−2)	0.8	···

---

[a] SS=Steady-state operation.

[b] Upstream concentration=1 grain/ft^3 for A1–3 and B1–4; 0.04 grain/ft^3 for C1–2.

[c] HP=High-pressure pulsed unit.

[d] Fractional penetration $P=2(-5)=2\times10^{-5}$; efficiency$=1-P=$99.998%.

[e] LP=Low-pressure pulsed unit.

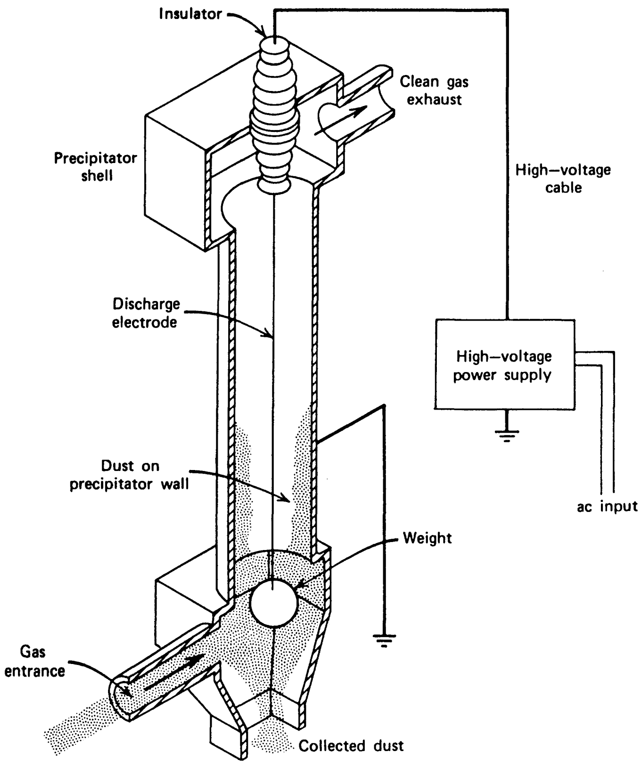
Insulator
Clean gas exhaust
Precipitator shell
High–voltage cable
Discharge electrode
High–voltage power supply
Dust on precipitator wall
ac input
Weight
Gas entrance
Collected dust

FIG. 22. Schematic single-stage electrostatic precipitator.

sizes of some typical particles. Particles larger than the "flying height" of the head can cause damage or the destruction of the device in a "head crash."

To control airborne contaminants, microelectronic manufacturing is usually done in a cleanroom. Figure 25 is a schematic diagram of a typical vertical laminar-flow (VLF) cleanroom (Larrabee, 1985). Air is introduced through HEPA (high efficiency particulate air) or ULPA (ultralow penetration air) filters covering the entire ceiling into the clean-

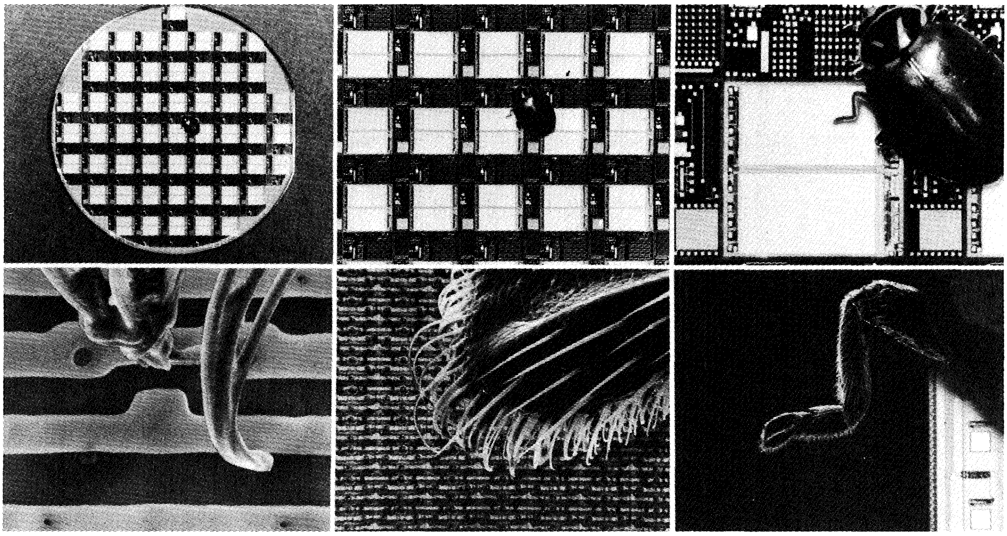

FIG. 23. Comparative size of semiconductor wafer, integrated circuit chip, and device features; increasing magnification clockwise from top left (Larrabee, 1985).

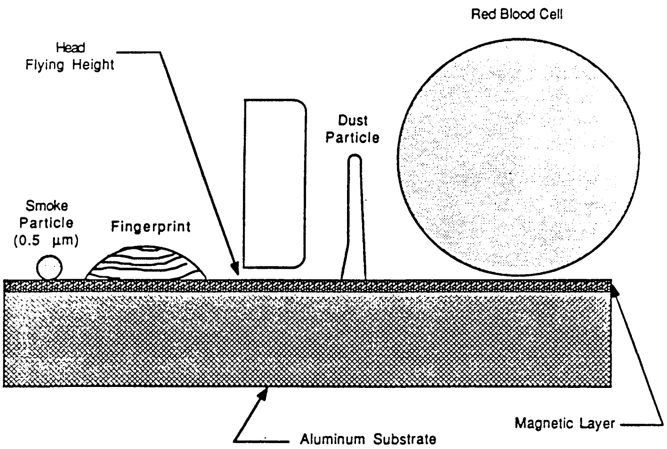
Red Blood Cell
Head Flying Height
Dust Particle
Smoke Particle (0.5 µm)
Fingerprint
Aluminum Substrate
Magnetic Layer

FIG. 24. Flying height of magnetic disk drives and size of contaminant and other particles.

room. Although the airflow in a VLF cleanroom is vertical, it is seldom laminar. As Fig. 25 shows, various factors can cause the airflow in a VLF cleanroom to become turbulent. Other types of cleanrooms include the horizontal flow cleanroom in which air flows horizontally from one end of the room to the other, and the mixed flow cleanroom in which the airflow is vertical, but different velocities are maintained across the ceiling filter to reduce airflow requirement and conserve energy.

Cleanrooms are usually categorized into classes in accordance with Federal Standard 209D (FED-STD-209D, 1988). A Class 100 cleanroom is one in which there are fewer than 100 particles per cubic foot >0.5 μm, Class 10, fewer than 10 particles per cubic feet >0.5 μm, etc. Since normal ambient air contains approximately 1 million particles

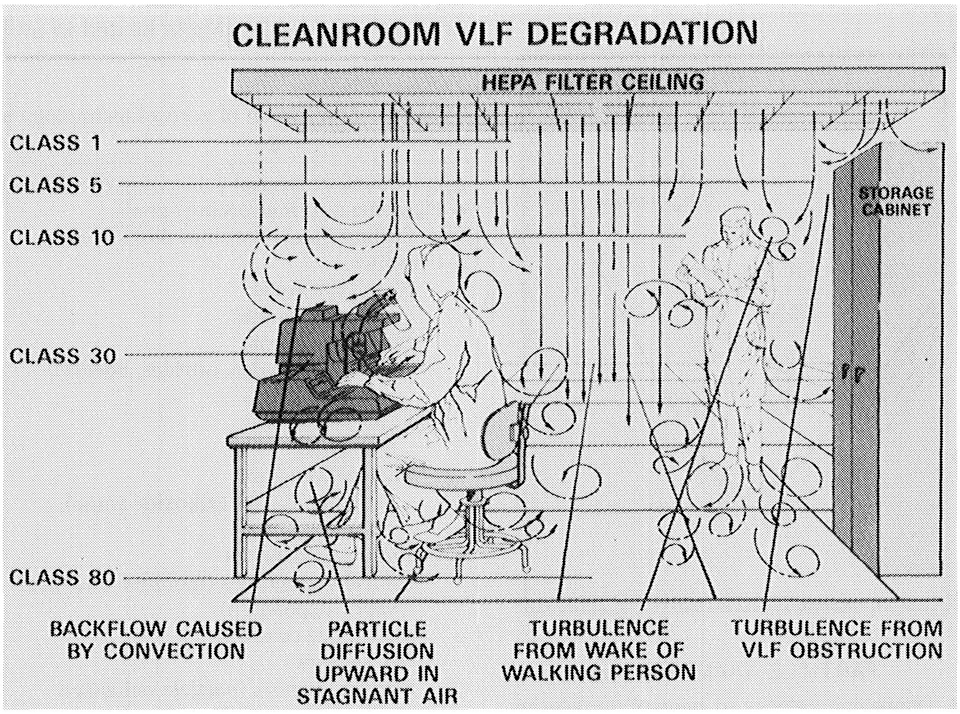
CLEANROOM VLF DEGRADATION
HEPA FILTER CEILING
CLASS 1
CLASS 5
CLASS 10
CLASS 30
CLASS 80
STORAGE CABINET
BACKFLOW CAUSED BY CONVECTION
PARTICLE DIFFUSION UPWARD IN STAGNANT AIR
TURBULENCE FROM WAKE OF WALKING PERSON
TURBULENCE FROM VLF OBSTRUCTION

FIG. 25. Schematic diagram of a vertical laminar flow (VLF) cleanroom (Larrabee, 1985).

per cubic foot >0.5 μm, a Class 100 cleanroom is 10 000 times cleaner than the normal ambient; a Class 10 cleanroom, 100 000 times cleaner. Current cleanrooms are being constructed in the Class 1 to Class 10 level for the state-of-the-art semiconductor manufacturing.

Deposition of aerosol particles on semiconductor wafers has been studied theoretically (Liu and Ahn, 1987) and by experiments (Pui *et al.*, 1990). Figure 26 shows the particle deposition velocity, the ratio of particle flux to airborne particle concentration, as a function of particle diameter. Particles larger than 0.2 μm are deposited by gravitational sedimentation and those below by Brownian diffusion.

3.5 Material Synthesis and Processing

Aerosol technology has played an important role in modern material processing. Aerosols are used in material synthesis to produce powders and bulk materials (Bowen, 1980), and for coating optical and microwave waveguides (Schober, 1987). Due to their simplicity and freedom from contamination, aerosol processes are used in the production of a variety of modern materials, such as advanced ceramics, carbon black, paint pigments, and high-temperature superconductors, and for coating waveguides, refractories, gas sensors, and semiconductors.

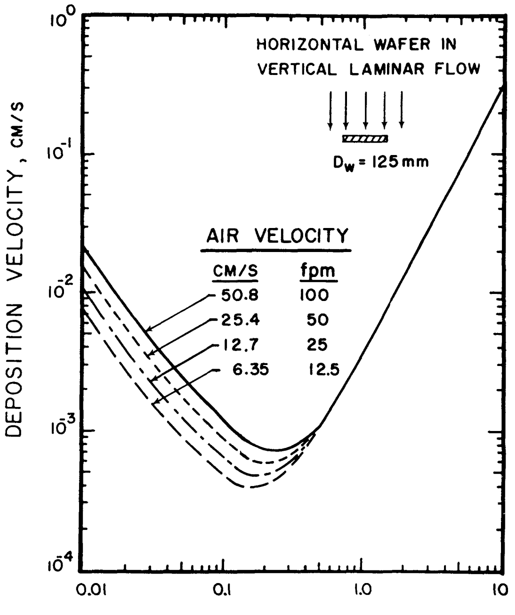

FIG. 26. Deposition velocity of aerosol particle on semiconductor wafers.

Table 4 summarizes the applications of aerosol processing of materials (Flagan, 1989). The largest commercial aerosol processes are flame synthesis reactors that are used in the production of carbon black, TiO_2 powder, and fume silica. Ceramics are produced by a number of aerosol processes including flame synthesis to produce oxides and carbides, plasma synthesis to produce oxides, carbides, nitrides, and borides, laser synthesis to produce SiC and Si_3N_4, and thermal reactors to produce Si, Si_3N_4, oxides, etc. Combustion processes are also used in a number of film growth technologies including the modified chemical vapor deposition process (see CHEMICAL VAPOR DEPOSITION) that is used to fabricate optical waveguide preforms (Nagel *et al.*, 1982). Lasers (Flint *et al.*, 1982) and plasmas (Pfender and Young, 1985) are also used to drive the chemistry in aerosol processes for powder synthesis or film growth.

Figure 27 shows a schematic diagram of a vapor-phase reactor developed for producing ultrafine MgO powder (Sheppard, 1987). The 99.98% ultrafine MgO can be used for such applications as sintering aids, electronics, catalysts, and rubber or plastic additions. To produce the powder, Mg is heated in an inert

Table 4. Application of aerosol technology in material processing.

Applications of aerosol processing of materials:
- Pigments and reinforcing agents
  - Carbon black, titanium dioxide, silica
- Ceramics
  - Flame synthesis
    - Oxides, carbides
  - Plasma synthesis
    - Oxides, carbides, nitrides, borides, etc.
  - Laser synthesis
    - SiC, Si_3N_4
  - Thermal reactors
    - Si, Si_3N_4, oxides (alkoxide route), etc.
- Coatings and deposits
  - Optical waveguides (modified chemical vapor deposition)
  - Refractory coatings
  - Gas sensors
  - Semiconductors (particle enhanced chemical vapor deposition)

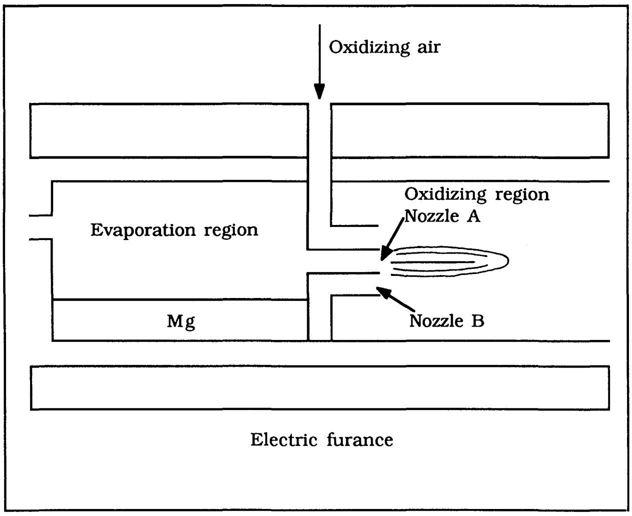

FIG. 27. Aerosol flame reactor for producing ultrafine MgO powders (Sheppard, 1987).

gas atmosphere above its boiling point by an electric furnace. The Mg vapor is then injected by argon into the oxidation region through the cylindrical nozzle, A. The oxidizing air is carried into the oxidation region through an annular nozzle B outside of nozzle A. A diffusion flame is formed as a result of reaction between the Mg vapor and the oxidizing air, which oxidizes the vapor to form MgO powder.

A second example involves the fabrication of a gas sensor using ultrafine tin dioxide particles. Figure 28 shows the schematic diagram of an ultrafine tin dioxide particle generator developed by Adachi *et al.* (1988). High-purity tin wire in a ceramic boat is heated in a nitrogen gas stream. The tin vapor is subsequently oxidized by oxygen gas in a mixing section to produce a supersaturated tin dioxide vapor. At a sufficient level of supersaturation, ultrafine tin oxide clusters are formed by homogeneous nucleation and by heterogeneous condensation of vapor molecules on the clusters. These tin dioxide particles are then impacted onto a glass substrate between two silver-plated electrodes to produce a deposit 2×3 mm in size and 0.01 to 0.2 mm in thickness. The gas sensor was found to be a sensitive leak detector for hydrogen in air.

3.6 Radioactive Aerosols and Nuclear Safety

Radioactive aerosols have received considerable attention by aerosol scientists because of safety concerns relating to nuclear power and fuel reprocessing plants. Radioactive aerosols also occur naturally as decay products of radon emanation from the soil attach to airborne particles. Current nuclear research has focused on three areas, namely, characterization of sources of nuclear aerosols, aerosol transport and deposition in confined systems, and air cleaning and nuclear aerosols. A summary of nuclear aerosol research can be found in the two international aerosol conference proceedings (Liu, Pui, and Fissan, 1984; GAeF and AAAR, 1986). Extensive research has been performed in the above three areas for fast breeder (LMFBR), light-water reactor (LWR), and reprocessing plant (RP).

In a nuclear power plant accident, the most severe form of which is the loss of all coolant to the reactor cores, the core would melt and vaporize, leading to the formation of highly radioactive aerosols. These aerosols would first be released into the containment vessel, and in the case of containment failure, escape to the ambient atmosphere. The source terms describing radioactive aerosol release can be classified as primary or secondary. The primary source terms are those due to the melting and vaporization of major parts of the core inventory, and the secondary, those in the form of delayed emissions such as hot sodium in LMFBR carrying fission and activation products which are released into the containment space (Schikarski, 1984). Brockmann and Tarbell (1984) have demonstrated that the pressurized ejec-

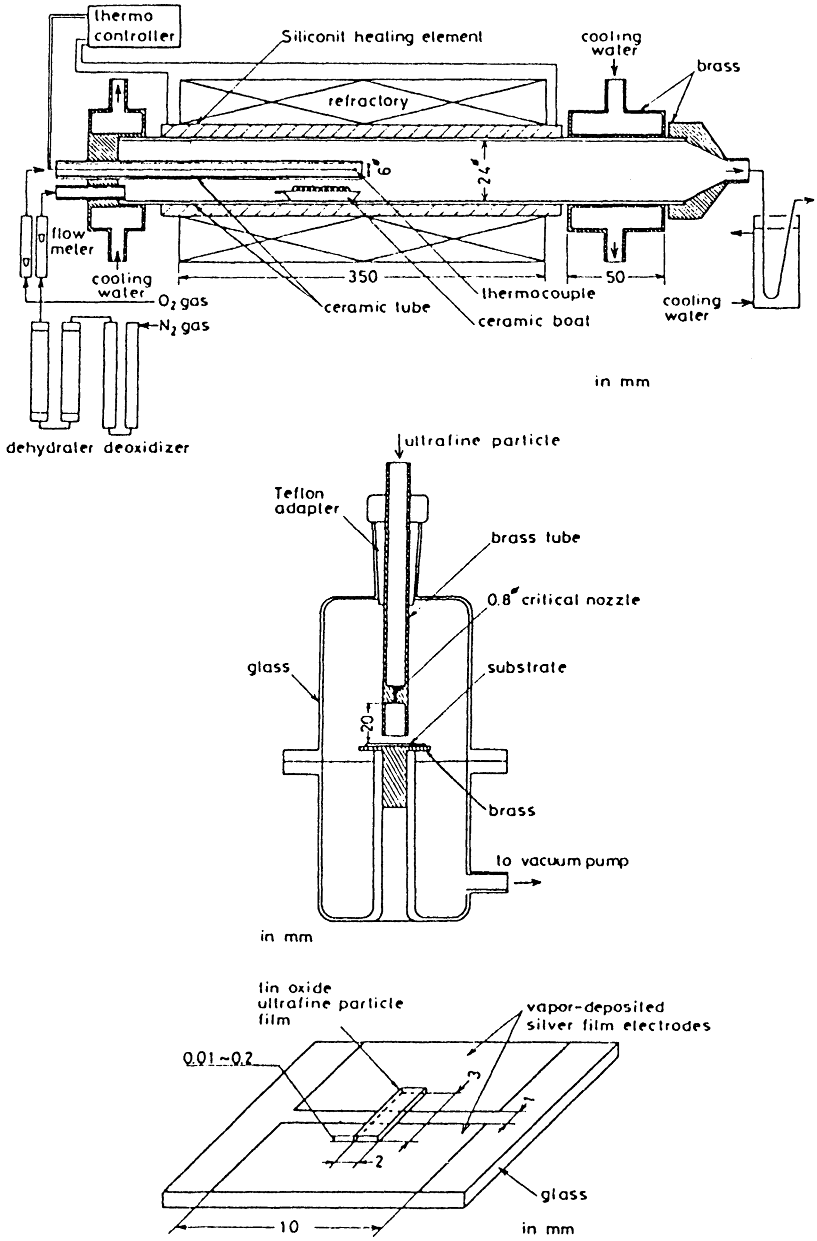
thermo controller
Siliconit heating element
cooling water
brass
refractory
24
8
flow meter
cooling water
O₂ gas
N₂ gas
350
50
thermocouple
ceramic tube
ceramic boat
cooling water
in mm
dehydrater deoxidizer
ultrafine particle
Teflon adapter
brass tube
0.8 critical nozzle
glass
substrate
20
brass
to vacuum pump
in mm
tin oxide ultrafine particle film
vapor-deposited silver film electrodes
0.01~0.2
3
1
2
glass
10
in mm

FIG. 28. Method of producing ultrafine tin oxide particle film as hydrogen gas sensor (Adachi *et al.*, 1988), showing the particle generator (top), impactor for collecting particles on the substrate (middle), and ultrafine tin oxide particle film on supporting substrate (bottom).

tion of molten core material from the vessel can lead to the formation of fine aerosols.

Computer codes for predicting aerosol transport and deposition in confined systems have been developed for nuclear safety studies beginning about 1965. Schikarski (1983) summarized the experimental programs and corresponding code development. Aerosol processes and properties, such as coagulation, sedimentation, diffusion, thermophoresis, diffusiophoresis, turbulent deposition, dynamic shape factor, and coagulation shape factor, all need to be considered in order to understand the behavior of nuclear aerosol in confined systems.

Air cleaning requirements in a nuclear facility are stringent. In addition to high removal efficiency, cleaning devices are required to handle high temperature and chemically aggressive nuclear aerosols. It is known that sodium aerosols together with high humidity can degrade ordinary HEPA filters rapidly. Alternative systems, such as granular bed filters, Venturi scrubbers, and electrostatic precipitators, are also investigated for the removal of nuclear aerosols (Schikarski, 1984).

Recent concerns about radon gas and its short-lived progeny in indoor environments and mine shafts have advanced significantly the measurement technology in radioactive aerosols and thus nuclear aerosols. Knutson *et al.* (1986) performed an intercomparison program for radon progeny measurement.

3.7 Atmospheric Science and Weather Modification

Clouds are formed when rising currents of air containing water vapor are cooled as they expand (see ATMOSPHERIC PHYSICS). Condensation of water vapor in the air requires aerosol particles to serve as nuclei of condensation. In the atmosphere, a particle of 0.1-μm diameter will serve as a nucleus of condensation for a cloud droplet at a supersaturation level of 1%. NaCl and other hygroscopic particles are particularly effective in serving as cloud nuclei due to their hygroscopic nature and their ability to absorb water vapor from the ambient atmosphere and grow before the relative humidity reaches 100%. The presence of aerosol particles in the ambient atmosphere therefore contributes to the cloud formation process. An important source of cloud nuclei over the ocean surface or in the coastal area is the dry sea salt particles resulting from the bursting of bubbles from breaking of waves on the ocean surface.

When clouds rise to a sufficiently high altitude, the temperature drops below 0 °C. However, the cloud droplets may remain in a supercooled liquid state due to the absence of freezing nuclei in the droplet. Very pure water droplets, only a few micrometers in diameter, may be supercooled to −40 °C without freezing and cloud droplets commonly exist in the supercooled state to −20 °C and, on occasion, down to −35 °C (Mason, 1975). The most important source of terrestrial ice nuclei appears to be soil particles, including clay minerals such as kaolinite, montmorillonite, and other closely related forms.

The possibility of seeding supercooled clouds by introducing artificial ice nuclei to form snow and initiate precipitation was foreseen in the 1930s, but cloud seeding experiments did not start in earnest until 1946, when Vincent J. Shaefer of the General Electric Company dropped dry ice into a supercooled cloud, precipitating snow that fell for 2000 ft before evaporating in the dry air. The discovery by B. Vonnegut in the same year that silver iodide crystals served as efficient ice nuclei further stimulated work on artificial rainmaking experiments in the U.S. and elsewhere. Cloud seeding is now done commercially, although only to a limited extent.

Atmospheric aerosols are often referred to as Aitken nuclei in honor of John Aitken, who in 1880 developed the first aerosol counter based on the adiabatic expansion of saturated air to condense water vapor on the invisible particles to form visible droplets, which are then counted. Atmospheric aerosols cover a wide range of sizes, from a few nanometers to as large as 100 μm. Atmospheric aerosols reflect sunlight back to space, and those with absorption characteristics will absorb solar radiation, causing a heating effect on the atmosphere. The climatic impact of air pollution is a subject of considerable contemporary interest, but the effect is not precisely quantified. When combined with the heating effect due to the "greenhouse gases," principally CO_2, which is also emitted by fossil fuel combustion, the net effect of air pollution on global climate remains to be determined. In the event of large scale nuclear war, very large amounts of aerosol material may be injected into the atmosphere to cause a significant cooling of the atmosphere and the onset of "nuclear winter." Both the nature and the extent of the nuclear winter effect are of some controversy and not universally accepted.

GLOSSARY

Aerosol: Solid or liquid particles suspended in a gas.

Aerodynamic Equivalent Diameter: Diameter of a unit-density sphere having the same settling speed as the particle in question; the particle itself may have a density different from one and be irregular in shape.

Aitken Nuclei: Aerosol particles measured by the Aitken type of condensation nucleus counter; generally refers to particles that can serve as nuclei of condensation in high vapor supersaturation, typically in the (2–5)% range.

Cascade Impactor: An aerosol size fractionation device whereby a polydisperse (nonuniform) aerosol is separated into discrete size-fractionated samples each covering a narrower size interval. Cascade impactor operates by the principle of inertial impaction onto a solid surface (inertial impactor) or into a void space (virtual impactor).

Cloud Condensation Nuclei: Aerosol particles that can serve as nuclei of condensation in cloud formation processes involving low water vapor supersaturation (usually $<1\%$).

Condensation Nuclei: Small particles that can serve as nuclei of condensation in a supersaturated vapor atmosphere.

Condensation Nucleus Counter: An aerosol particle counter based on the condensation of a vapor on the particle and subsequent growth of the particle to a large enough size for optical detection by light scattering or attenuation.

Corona Discharge: High-voltage electrical discharge maintained from the tip of a needle or a fine wire in a gas.

Diffusion Battery: An aerosol size distribution measuring device based on the diffusive collection of particles by metal screens, capillary tubes, or rectangular flow channels. The decrease in particle concentration is taken as a measure of particle diffusion coefficient, which is then converted to a particle size by Einstein's equation.

Dispersions: Solid or liquid particles dispersed in a gaseous or a liquid medium.

Drag Force: Force exerted by the fluid medium on a suspended particle during motion of particle through the medium.

Electrical Aerosol Analyzer: An aerosol concentration and size distribution measuring device based on the charging of particles by unipolar ions and subsequent measurement of the mobility distribution of the charged aerosol by an electrical condenser and an electrometer detector.

Electrostatic Precipitator: An aerosol particle removal device based on the charging of the particles by unipolar ions generated by a corona discharge and precipitating the charged particles in an electric field.

Energy Disperse X-Ray (EDX) Analysis: A particle analysis method based on irradiating a single particle or a small sample of particles by electrons and measuring the energy distribution of the emitted x rays by a solid-state detector.

Foam: Small gas bubbles dispersed in a solid or liquid matrix.

HEPA Filter: A high efficiency particulate air filter with efficiencies $>99.97\%$ for 0.3-μm-diam particles.

Hydrosol: Solid or liquid particles suspended in a liquid.

Ice Nuclei: Small particles that can serve as nuclei for ice formation in a subcooled cloud.

Knudsen Number: Ratio of mean free path of molecules to particle radius.

Inspirable Particles: Particles capable of entering the human nose or mouth during breathing; usually <100 μm in aerodynamic equivalent diameter.

Mie Scattering: Light scattering by particles with radius comparable to the wavelength of light as originally studied by Gustav Mie in 1908.

Mobility: Dynamic mobility is the ratio of the terminal velocity of the particle to the applied force; electrical mobility is the ratio of the terminal velocity of a charged particle in an electric field to the electric field intensity.

Molecular Clusters: Clusters of molecules consisting of two to several thousand molecules which are often unstable and may grow by collision with other molecules or disintegrate as a result of thermal energy or collision. Molecular clusters are usually smaller than an aerosol particle and larger than a single molecule.

Optical Particle Counter: A particle counting and sizing instrument based on the scattering of light by particles.

PM-10, or Particulate Matter, 10-μm Cut: Aerosol particles with aerodynamic equivalent diameter of 10 μm or smaller measured by a method approved by the U.S. Environmental Protection Agency for air pollution control regulations.

Particulate Matter or Particulates: Small, microscopic pieces of material from a few nanometers to several hundred micrometers in size. Particulate matter can exist in the form of an aerosol (solid or liquid particles suspended in a gas), a hydrosol (solid or liquid particles suspended in a liquid), or a powder (solid particles).

Proton Induced X-Ray Emission (PIXE) Analysis: A particle analysis method whereby a small sample of particles is irradiated by energetic protons from an accelerator to cause x-ray emission from the sample for elemental analysis.

Rayleigh Limit: Limiting charge on a liquid drop determined by the balance between the electrical repulsion between the charges on the drop and surface tension. Rayleigh limit is the maximum charge a liquid drop can carry without disintegration.

Rayleigh Scattering: Light scattering by particles much smaller than the wavelength of light.

Respirable Particles: Aerosol particles capable of reaching the gas-exchange, alveolar (pulmonary) region of the human lung; usually <3.5 μm in aerodynamic equivalent diameter.

Settling Speed: Terminal velocity of a particle in a gas or liquid under the action of gravity.

Terminal Velocity: The steady velocity of a particle in a gas or liquid under the action of a constant force such as gravity, electrical, etc.

Thoracic Particles: Particles capable of penetrating into the trachea in the human thorax; usually <10 μm in aerodynamic equivalent diameter.

ULPA Filter: An ultralow penetration air filter with 99.999% or higher efficiency when tested with particles in the 0.1 to 0.3 μm diameter range.

Ultrafine Particles: In aerosol science, usually refers to particles in the 2–20 nm range.

VLF Cleanroom: A vertical laminar flow cleanroom with entire ceiling covered with HEPA or ULPA filters used for semiconductor or microelectronic manufacturing.

X-Ray Fluorescence (XRF) Analysis: A particle analysis method whereby a small sample of collected particle is irradiated by x rays and the fluorescent x-ray emission of the sample is analyzed to determine its elemental composition.

List of Works Cited

Adachi, M., Okuyama, K., Kousaka, Y., Tanaka, H. (1988), *J. Aerosol Sci.* **19**, 253–263.

Agarwal, J. K., Sem, G. J. (1980), *J. Aerosol Sci.* **11**, 343–357.

Bergin, M. H., Pui, D. Y. H., Kuehn, T. H., Fay, W. T. (1989), *ASHRAE Trans.* **95**, Pt. 2, 3264 (RP-531).

Berglund, R. N., Liu, B. Y. H. (1973), *Environ. Sci. Technol.* **7**, 147–153.

Bowen, H. K. (1980), *Mater. Sci. Engr.* **44**, 1–56.

Brockmann, J. E., Targbell, W. W. (1984), "Aerosol Generation by Pressurized Melt Ejection SAND84-0246A," in: B. Y. H. Liu, D. Y. H. Pui, H. J. Fissan (Eds.), *Aerosols*, New York: Elsevier.

Cooper, D. W. (1986), *Aerosol Sci. Technol.* **5**, 287–299.

Fardi, B., Liu, B. Y. H. (1989), "Performance of Disposable Respirators," Publication No. 709, Particle Technology Laboratory, Mechanical Engineering Department, University of Minnesota, Minneapolis, MN.

Flagan, R. C. (1989), "Lecture Notes for Aerosol Technology," presented at the Mechanical Engineering Department, University of Minnesota, 22–26 May.

Flint, J. H., Marra, R. A., Haggerty, J. S. (1986), *Aerosol Sci. Technol.* **5**, 249–260.

GAeF and AAAR (1986), *Aerosols: Formation and Reactivity*, Oxford: Pergamon.

Hinds, W. C. (1982), *Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles*, New York: Wiley.

Kerker, M. (1969), *The Scattering of Light and Other Electromagnetic Radiation*, New York: Academic.

Knutson, E. O., Solomon, S. B., Miles, J. C. H., Holub, R. F. (1986), "Aerosol Generation/Measurement in the OECD-CEC Intercomparison Program for Radon Progeny Measurement," in:

GAeF, AAAR, *Aerosols: Formation and Reactivity*, Oxford: Pergamon.

Larrabee, G. B. (1985), *Chem. Eng.*, 10 June, 1985, 51–59.

Liu, B. Y. H., Ahn, K. H. (1987), *Aerosol Sci. Technol.* **6**, 215–224.

Liu, B. Y. H., Kapadia, A. (1978), *J. Aerosol Sci.* **9**, 227–242.

Liu, B. Y. H., Pui, D. Y. H. (1974), *J. Colloid Interface Sci.* **47**, 155–171.

Liu, B. Y. H., Pui, D. Y. H. (1975), *J. Aerosol Sci.* **6**, 249–264.

Liu, B. Y. H., Pui, D. Y. H. (1986), "On Aerosol Sampling and Sampling Inlets," in: S. D. Lee *et al.* (Eds.), *Aerosols: Research, Risk Assessment and Control Strategies*, Chelsea, MI: Lewis, pp. 175–184.

Liu, B. Y. H., Pui, D. Y. H. (1987), "Instrumentation for Aerosol Measurement," in: D. Tolliver (Ed.), *Handbook of Contamination Control in Microelectronics*, Park Ridge, NJ: Noyes.

Liu, B. Y. H., Berglund, R. N., Agarwal, J. K. (1974), *Atmos. Environ.* **8**, 717–732.

Liu, B. Y. H., Pui, D. Y. H., Rubow, K. L. (1983), "Characteristics of Air Sampling Filter Media," in: V. A. Marple, B. Y. H. Liu (Eds.), *Aerosols in the Mining and Industrial Work Environments*, Vol. 3, Michigan: Ann Arbor Science, pp. 989–1038.

Liu, B. Y. H., Pui, D. Y. H., Fissan, H. J. (1984), *Aerosols: Science, Technology, and Industrial Applications of Airborne Particles*, New York: Elsevier.

Liu, B. Y. H., Rubow, K. L., Pui, D. Y. H. (1985), "On the Performance of HEPA and ULPA Filters," in *Proceedings: 31st Annual Technical Meeting of the Institute of Environmental Sciences.*

Liu, B. Y. H., Pui, D. Y. H., Schaller, R. E., McDonald, B. N., Johnson, T. W. (1986a), *Part. Characterization* **3**, 68–73.

Liu, B. Y. H., Szymanski, W., Pui, D. Y. H. (1986b), *ASHRAE Trans.* 92, Pt. 2971 (RP-354).

Loo, B. W., Adachi, R. S., Cork, C. P., Goulding, F. S., Jaklevic, J. M., Landis, D. A., Searles, W. L. (1979), Report LBL-8725, Lawrence Berkeley Laboratory, CA.

Marple, V. A., Liu, B. Y. H., Kuhlmey, G. A. (1981), *J. Aerosol Sci.* **12**, 333–337.

McFarland, A. R., Ortiz, C. A., Bertch, R. W., Jr., (1984), *J. Air Pollution Control Assoc.* **34**, 544.

Nagel, J. B., MacChesney, J. B., Walker, K. L. (1982), *IEEE J. Quantum Electron.* QE-**18**, 459–476.

Pfender, E., Young, R. M. (1985), *Plasma Chem. Plasma Proc.* **5**, 1–37.

Phalen, R. F., Hinds, W. C., John, W., Lioy, P. J., Lippmann, M., McCawley, M. A., Raabe, O. G., Soderholm, S. C., Stuart, B. O. (1986), *Appl. Ind. Hyg.* **1**, 3–14.

Pritchard, J. A. (1976), "A Guide to Industrial Respiratory Protection," Publ. No. 76-189, U.S. Dept. of Health, Education, and Welfare, Washington, DC.

Pui, D. Y. H., Liu, B. Y. H. (1988), *Phys. Scr.* **37**, 252–269.

Pui, D. Y. H., Ye, Y., Liu, B. Y. H. (1990), *Aerosol Sci. Technol.* **12**, 795–804.

Schikarski, W. O. (1983), KfK-Report 3636/EIR-Report 502.

Schikarski, W. O. (1984), "Recent Developments in Nuclear Aerosol Research," in: B. Y. H. Liu, D. Y. H. Pui, H. J. Fissan (Eds.), *Aerosols*, New York: Elsevier.

Schober, M. (1987), *Ceram. Bull.* **66**, 1698–1699.

Seinfeld, J. H. (1986), *Atmospheric Chemistry and Physics of Air Pollution*, New York: Wiley.

Sheppard, L. M. (1987), *Adv. Mater. Proc. Metal Prog.* **4/87**, 53–58.

Stolzenburg, M. R., McMurry, P. H. (1986), "Counting Efficiency of an Ultrafine Aerosol Condensation Nucleus Counter: Theory and Experiment," in: GAeF, AAAR, *Aerosols: Formation and Reactivity*, New York: Pergamon, p. 786.

Strass, W. (1966), *Industrial Gas Cleaning*, New York: Pergamon.

Szymanski, W. W., Liu, B. Y. H. (1986), *Part. Characterization* **3**, 1–7.

Whitby, K. T. (1978), "The Physical Characteristics of Sulfur Aerosols," in: R. B. Husar *et al.* (Eds.), *Sulfur in the Atmosphere*, New York: Pergamon.

Whitby, K. T., Liu, B. Y. H. (1966), "Electrical Properties of Aerosols," in: C. N. Davies (Ed.), *Aerosol Science*, New York: Academic.

White, H. J. (1963), *Industrial Electrostatic Precipitation*, Reading, MA: Addison–Wesley.

Further Reading

Davies, C. N., (Ed.) (1966), *Aerosol Science*, New York: Academic.

Friedlander, S. K. (1977), *Smoke, Dust and Haze*, New York: Wiley.

Fuchs, N. A. (1964), *The Mechanics of Aerosols*, New York: MacMillan.

Green, H. L., Lane, W. R. (1964), *Particulate Clouds, Dust, Smokes and Mists*, 2nd ed., Princeton, NJ: Van Nostrand.

Hinds, W. C. (1982), *Aerosol Technology*, New York: Wiley.

Junge, C. (1963), *Air Chemistry and Radioactivity*, New York: Academic.

Kerker, M. (1969), *The Scattering of Light and Other Electromagnetic Radiation*, New York: Academic.

Liu, B. Y. H. (Ed.) (1976), *Fine Particles*, New York: Academic.

Marple, V. A., Liu, B. Y. H. (Eds.) (1983), *Aerosols in the Mining and Industrial Work Environments,* Ann Arbor, MI: Ann Arbor Science.

Seinfeld, J. H. (1986), *Atmospheric Chemistry and Physics of Air Pollution,* New York: Wiley.

Shaw, D. T. (Ed.) (1978), *Fundamentals of Aerosol Science,* New York: Wiley.

Shaw, D. T. (Ed.) (1978), *Recent Developments in Aerosol Technology,* New York: Wiley.

Van de Hulst, H. C. (1980), *Multiple Light Scattering,* New York: Academic.

White, H. J. (1963), *Industrial Electrostatic Precipitation,* Reading, MA: Addison–Wesley.

AEROSPACE MEDICAL PHYSICS

PERCIVAL D. MCCORMACK, *Life Science Division, NASA Headquarters, Washington, DC*
CHARLES E. SWENBERG, *Radiation Biochemistry Department, Armed Forces Radiobiology Research Institute, Bethesda, Maryland*

	Introduction	444
1.	**Biodynamics**	445
1.1	Physics and Physiology of High-Magnitude Acceleration	445
1.2	Physiological Effects of Sustained Acceleration	445
1.2.1	General Effects	445
1.2.2	Negative-Acceleration Effects	446
1.2.3	Cardiovascular Effects	446
1.2.4	Hydrostatic Effects	446
1.3	Positive Acceleration and Its Effects on Retinal Circulation, Cerebral Circulation, and the Pulmonary System	446
1.3.1	Retinal Circulation	446
1.3.2	Cerebral Circulation	446
1.3.3	Pulmonary System	447
1.3.3.1	Regional Lung Ventilation	447
1.3.3.2	Regional Pulmonary Blood Flow	447
1.4	Long-Duration Negative Acceleration	447
1.4.1	General Effects	447
1.4.2	Cardiovascular Effects	448
1.4.3	Carotid Sinus and Cardiac Output	448
1.4.4	Cerebral Circulation	448
1.4.5	Pulmonary Effects	448
1.5	Effects of Acceleration on Arterial Wall Mechanical Properties	448
1.5.1	Arterial Wall Viscoelastic Properties	449
1.5.2	Arterial Viscoelasticity—Complex Young's Modulus	450
1.5.3	Hysteresis	450
1.6	Calculation of Arterial Wall Performance in Aircraft Maneuvers	452
2.	**Physiological and Biological Responses to Weightlessness**	452
2.1	Physiological Responses to Weightlessness	452
2.2	Cardiopulmonary System	453
2.2.1	Fluid Shifts	453
2.2.2	Decreased Orthostatic Tolerance	454
2.2.3	Lower Body Negative Pressure	455
2.2.4	Venous Compliance	455
2.3	Bone and Calcium Metabolism	456
2.3.1	Calcium Balance Studies	456
2.3.2	Bed Rest Simulation	457
2.4	Biological Responses to Weightlessness	457
2.4.1	Tropic Systems of Plants	457
2.4.2	Gravitational Biodynamics	457
2.5	Statistical Distribution of Particles under Gravitational Force	458
2.6	Effects of Gravitation on Membranes	459
2.6.1	Effects of Hydrostatic Stress on a Membrane	459
2.6.2	Effects of Hydrostatic Stress on Active and Passive Transport through Semipermeable Membranes	459
3.	**Artificial Gravity: A Countermeasure to Weightlessness**	460
3.1	Characteristics of the Rotating Environment	460
3.2	Artificial Gravity Level	461
3.3	Coriolis Forces	461
3.4	Human Performance in Artificial Gravity	462
3.4.1	Self-Locomotion	462
3.4.2	Material Handling	463
3.5	Cross-Coupled Angular Accelerations	463
4.	**Spatial Disorientation and Motion Sickness**	466
4.1	Orientation Mechanisms	466
4.2	The Inner Ear	466
4.3	The Vestibular System	467
4.3.1	Semicircular Canals	467

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

4.3.2 Mathematical Modeling of the Cupula-Endolymph System.... 468
4.4 Nonlinear Characteristics of the Endolymph Cupula-Nystagmus System 469
4.5 The Otolith Organs 470
4.5.1 Dynamics of the Otolith Organs....................... 471
4.5.2 False Perception of Angular Motion: Somatogyral and Oculogyral Illusions 472
4.6 Space Motion Sickness........ 472
5. The Radiation Environment and Associated Hazards...... 474
5.1 Earth's Radiation Environment................. 474
5.1.1 Inner Zone................... 475
5.1.2 Outer Zone................... 476
5.1.3 Cosmic Rays 476
5.1.4 Galactic Cosmic Rays......... 477
5.1.5 Solar Cosmic Rays 478
5.2 Shielding 478
5.2.1 Geomagnetic Shielding 478
5.2.2 Shielding by Materials........ 480
5.3 Biological Effects of Radiation..................... 481
5.4 Expected Radiation Exposures in Space Missions 482
5.4.1 LEO Missions 482
5.4.2 Geosynchronous Orbits 482
5.4.3 Lunar Mission 482
5.4.4 Mars Mission 482
Glossary..................... 483
Works Cited 484
Further Reading 487

INTRODUCTION

The human body has evolved in the Earth's gravitational field and is well adapted to this environmental stress. Extremely fast, highly maneuverable modern aircraft expose air crews to acceleration forces of exorbitant magnitude, whereas spacecraft within Earth's orbit subject crews to an acceleration force of virtually zero magnitude. This article summarizes the physics and physiology of high-magnitude, sustained acceleration on retinal and cerebral circulation and discusses the effects of aircraft maneuvers on arterial wall performance. Human physiological adaptation to the microgravity space environment is summarized and basic biological responses to weightlessness are discussed.

A countermeasure to microgravity, the rotating environment, is examined. This artificial-gravity environment should provide an effective countermeasure for the bone demineralization, muscle loss, motor coordination disturbance, and cardiovascular deconditioning experienced after long periods in the zero gravity of space. Such physiological deterioration if not compensated for could have a disastrous impact on an Earth-Mars-Earth mission with the propulsion systems currently available. A rotating artificial-gravity environment has inherent and unique characteristics that may have debilitating effects on humans and their performance. Here we review current knowledge about artificial gravity. Its unique characteristics are examined and its effects on human performance and vehicle habitability are discussed.

Many aircraft and spacecraft crew members experience illusory sensations of craft altitude and motion, or fail to detect changes in craft orientation at some time during their flights. These events are normal and are due to physiological limitations of the sensory mechanism. Human sensory organs are functionally adapted to terrestrial conditions in a stable 1*g* environment. In the aerial and space environment humans are exposed to motion stimuli that differ from those experienced on the ground, so that misperception of spatial orientation is not surprising.

The discussion of motion sickness is restricted to air and space motion sickness (SMS). The cardinal symptom of motion sickness is nausea and the cardinal signs are vomiting, pallor, and sweating. SMS affects, to some degree, about half of those exposed to 0*g* environment. Its onset is rapid—within 30 min of entry—and in nearly all cases adaptation and recession of signs and symptoms occur within 48 to 72 h. The visual and vestibular systems play significant roles in the genesis of both spatial disorientation and motion sickness.

In addition to microgravity conditions, the Earth's natural radiation environment poses a hazard to manned space activities directly through biological effects and indirectly through effects on materials and electronics.

Long-term manned space flight and missions involving the colonization of the Moon or Mars would expose humans to types and levels of ionizing radiation neither encountered on Earth nor experienced in near-space orbits. Characteristics of the radiation environment unique to manned space flight, both in near space and beyond the Earth's geomagnetic field, and biological effects of the radiation are summarized briefly.

1. BIODYNAMICS

1.1 Physics and Physiology of High-Magnitude Acceleration (Leverett and Whinnery 1985)

Acceleration is defined as the rate of change of velocity, and is a vector quantity. It can result from a change in *speed* along a straight line (linear acceleration), from a change in the *direction* of motion (radial acceleration), or from changes in both. If s is the distance and t is the time, then

$$\text{acceleration } (f) = d^2s/dt^2. \qquad (1)$$

Acceleration can be positive (increasing velocity) or negative (decreasing velocity). In the Système International d'Unités (SI), the units are m/s^2. In aviation, acceleration is generally expressed as a multiple of that due to the Earth's gravitational force. The Earth's gravitational acceleration, g, has the value 9.81 m/s^2. The ratio of an applied acceleration to the gravitational constant is known as the g value. For example, $3g$ represents an applied acceleration of 29.43 m/s^2. It is also standard practice to express the acceleration vector g in terms of three orthogonal components, g_x, g_y, and g_z, with the axes oriented to the human torso (Fig. 1). The symbols $\dot{r}_x$, $\dot{r}_y$, and $\dot{r}_z$ denote *angular* acceleration around these orthogonal axes.

The applied acceleration and the resultant inertial force always act in diametrically opposite directions, and this holds for the $0g$ and non-$0g$ environments. Thus, a headward acceleration tends to force blood towards the lower extremities. The resultant force is designated as $+g_z$. Figure 1 illustrates the standard terminology for indicating the *direction* of acceleration and the inertial forces that act on aircraft crews.

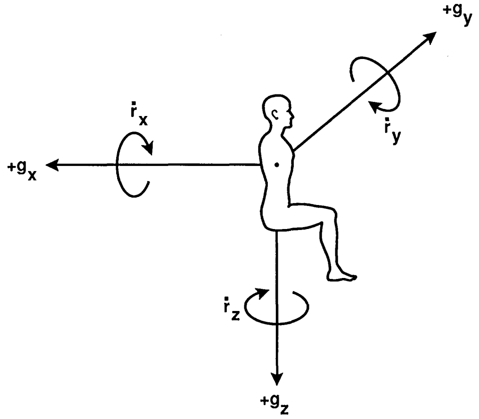

FIG. 1. Physiological acceleration system.

1.2 Physiological Effects of Sustained Acceleration

Crew members in high-performance aircraft are regularly exposed to sustained $+g_z$ acceleration. Accelerations of $(5–7)g$ for 10–40 s are common. Onset rates may be as high as $10g/s$. Even with the use of anti-g suits, human tolerance is the limiting factor in aircraft performance.

1.2.1 General Effects. With $+ g_z$ acceleration (Burton, Leverett, and Michaelson, 1974), the sequence of general effects as the g value rises is as follows:

$+g_z=2.0$: The subject is aware of an increase in weight of the trunk and limbs and increased pressure on the buttocks. The face and soft tissues of the body begin to sag.

$+g_z=2.5$: Difficulty in raising oneself from a sitting position is experienced.

$+g_z=3.0–4.0$: It is now impossible to raise oneself from a sitting position and difficult to raise the arms and legs. Progressive dimming of vision occurs after 3–4 s of exposure, progressing to tunnel vision.

The acceleration value at which a given loss of peripheral vision occurs varies among subjects. In one experimental series, the vision field was reduced to a cone with an angle of 45° at a mean value of $4.1g$ (standard deviation $+0.7g$). Blackout occurred at a mean value of $4.7g$ $(+0.8g)$.

g_z = 4.5–6.0: Diminution of vision continues and progresses to blackout after about 5 s. Loss of hearing and then loss of consciousness (LoC) occur if exposure continues. Mild to severe convulsions occur in about 50% of test subjects during or following LoC; subjects frequently experience bizarre dreams. LoC is accompanied by a loss of muscle tone, so the head and trunk slump. Recovery of consciousness is slow, with complete incapacitation for 15 s followed by a similar period of loss of orientation in time and space.

1.2.2 Negative-Acceleration Effects. Aircraft maneuvers that produce $-g_z$ are outside loops and spins and simple inverted flight. Tolerance of $-g_z$ is much less than $+g_z$. At $-g_z = 2.0–3.0$, severe facial congestion and a throbbing headache develop. Progressive blurring, graying, or reddening of vision ("redout") occurs after 5 s. The congestion resolves slowly and may leave petechial hemorrhages and edematous eyelids.

The limit of tolerance for $-g_z$ is 5.0 for 5 s—rarely reached by most subjects. Because of severe performance decrements and aircraft limitations (they are not usually stressed to withstand high $-g_z$), maneuvers involving $-g_z$ greater than $(1.0–1.5)g$ are seldom performed.

1.2.3 Cardiovascular Effects. The cardiovascular effects of acceleration are generally the most significant. Positive acceleration immediately changes the distribution of pressure in the arterial and venous systems, shifting blood toward the more dependent parts (heart and lungs) and away from the head. Reflex compensatory changes then occur.

1.2.4 Hydrostatic Effects. Positive acceleration increases the weight of the columns of blood above and below the heart, decreasing vascular pressure above the heart and increasing it below. The changes in intravascular pressure alter the lumen diameter of blood vessels. The lumen diameter is a function of transmural pressure, vessel distensibility, and volume of blood available to fill the vessel.

Alteration of vessel diameter, in turn, determines regional blood flow and blood content. As the increase in pressure is transmitted to the small arteries and arterioles (vascular "beds") below the heart, the peripheral resistance falls and local blood flow rises. Any major decrease in venous pressure above the heart can cause vessels to collapse and stop blood flow through them. The magnitude of the induced pressure change, ΔP, is proportional to the vertical distance (h) from the observation point to the heart and the acceleration level (ng). Thus,

Table 1. Pressure in the arterial tree of a seated subject.

Location	Arterial pressure (mm Hg) $+1g_z$	$+4.5g_z$
Heart	100	100
Cerebral artery (eye level, +30 cm)	78	1
Femoral artery (−60 cm)	144	300

$$\Delta P = \rho n g h, \tag{2}$$

where ρ is the density of blood.

Table 1 indicates the gauge pressures in the arterial tree of a seated subject at $+1.0g_z$ and $+4.5g_z$. The pressure in the right atrium of the heart (venous reference point) is approximately atmospheric (0 mm Hg) both at $+1.0g_z$ and $+4.5g_z$. At $+1.0g_z$ the pressure in the cerebral veins at eye level would be about −20 mm Hg. At $+4.5g_z$ one would expect it to be about −100 mm Hg. But, because of a compensatory change in diameter of the jugular veins in the neck, the pressure in the cerebral veins falls to about −50 mm Hg. Positive acceleration increases the pressure in veins below the heart; in the femoral vein it rises from about 40 mm Hg at $+1.0g_z$ to about 150-mm Hg at $+4.5g_z$.

1.3 Positive Acceleration and Its Effects on Retinal Circulation, Cerebral Circulation, and the Pulmonary System

1.3.1 Retinal Circulation. Visual disturbances are caused by ischemia of the retina. Internal pressure in the eye is 20 mm Hg. If the systolic arterial pressure in the eye falls below 20 mm Hg, blackout occurs. Once the arterial pressure falls below 20 mm Hg, loss of vision occurs in 4–6 s. The delay is caused by a small reserve of oxygen dissolved in the extravascular fluid of the retina. When the pressure in the central retinal artery is reduced by $+g_z$, the blood supply to the periphery of the retina fails first and produces loss of peripheral vision ("greyout").

1.3.2 Cerebral Circulation. Positive acceleration of $(3.4–4.0)g$ reduces the arterial pressure at the brain below that required to

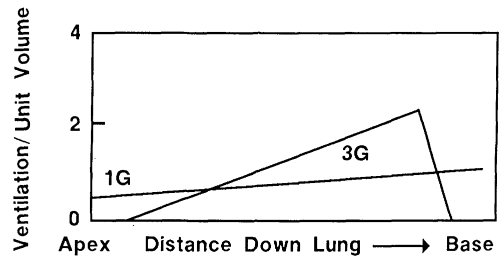

FIG. 2. Vertical distribution of alveolar ventilation.

maintain adequate cerebral blood flow. At 4.5*g* the cerebral arterial pressure is reduced to almost zero. The incidence of LoC with positive acceleration of (3.0–5.0)*g* is much lower than expected. Several mechanisms maintain an adequate flow of blood through the brain (even with arterial pressures of 0–20 mm Hg at head level).

First, the cerebrospinal fluid pressure falls so that the pressure differences across the walls of the intracranial vessels remain nearly normal. Second, active vasodilation of the cerebral arterioles reduces flow resistance. Third, the venous blood column in the neck sets off a siphon effect that maintains cerebral flow. The pressure difference between arterial and venous systems is maintained at 50–60 mm Hg at $+g_z$ of (4.0–5.0)*g* by a pressure in the jugular bulb of −50 mm Hg.

1.3.3 Pulmonary System (Glaister, 1970). Positive acceleration of up to +5.0*g* produces minimal respiratory difficulties. Total lung and vital capacities are unaffected at values up to +3.0*g* while exposure to +5.0*g* reduces them by about 15%. Positive acceleration causes the abdominal contents and diaphragm to descend, thereby increasing the functional residual volume by about 500 mL at $+3.0g_z$. The diaphragm descent is greatly reduced, and may be reversed by inflating the standard anti-*g* suit.

1.3.3.1 Regional Lung Ventilation. Positive acceleration increases regional differences in the distribution of ventilation in the lungs of an erect individual at $+1.0g_z$ (see Fig. 2). At $+5g_z$ the fluid pressure at the lung base is 30 cm water greater than at the apex. Alveoli at the apices are more distended and those at the bases are nearer to closure with respect to conditions at $+1g_z$. At minimum alveolar volume, the associated airways close. The lung volume at which airway closure can first be detected is termed the closing volume of the lung, and it increases linearly with acceleration (Fig. 3). Inflation of the abdominal bladder of an anti-*g* suit raises the diaphragm and reduces the functional residual capacity. Its use markedly increases the number of nonventilated alveoli in the lower part of the lung.

1.3.3.2 Regional Pulmonary Blood Flow. The distribution of blood flow in the lung is susceptible to $+g_z$, as pulmonary blood pressures are relatively low. The mean pulmonary artery pressure (15 mm Hg) at the junction of the middle and lower thirds of the lung is unaffected by $+g_z$. The vascular pressures above and below this level are determined by hydrostatic forces. Even at $+1g_z$ the mean arterial pressure (MAP) falls to zero 20 cm above this level (that is, near the apex). At $+4g_z$ the MAP is zero only 5 cm above this level. The proportion of the lung not perfused increases as $+g_z$ increases (Fig. 4).

1.4 Long-Duration Negative Acceleration

Tolerance to $-g_z$ is much lower than that for $+g_z$ and symptoms at even $-2.0g_z$ are unpleasant and alarming. Serious decrement in performance also occurs.

1.4.1 General Effects. A feeling of heaviness and interference with limb movements occurs as with $+g_z$. At $-1.0g_z$, a sense of fullness and pressure in the head occurs, which develops into a throbbing headache at

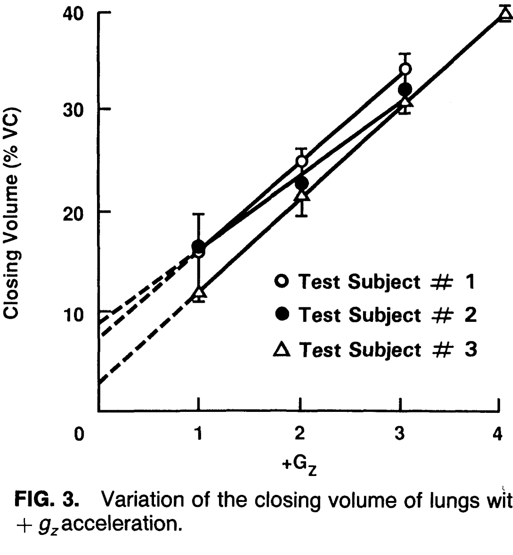

FIG. 3. Variation of the closing volume of lungs with $+g_z$ acceleration.

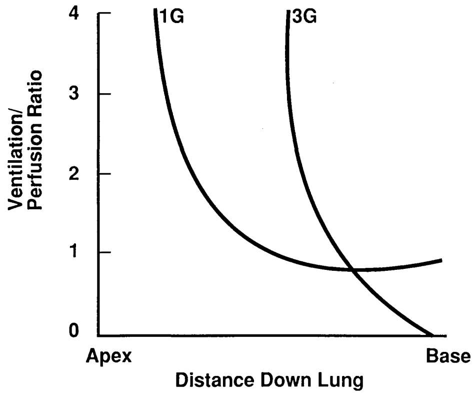

FIG. 4. Vertical distribution of ventilation/perfusion ratio for 1g and 3g accelerations.

$-2.0g_z$. At $-2.5g_z$, edema of the eyelids and petechial hemorrhages in the facial and neck skin occur. Nasal mucosal congestion may cause difficulty in breathing and lead to epistaxis. At $-(2.5\text{–}3.0)g_z$ the eyes feel as if they are about to pop out of their sockets, and excessive lachrymation causes blurring of vision. Exposure to more than $-4.0g_z$ for more than 6 s causes mental confusion and unconsciousness.

1.4.2 Cardiovascular Effects. The immediate hydrostatic effect of $-\ g_z$ is to increase vascular pressures in the regions above the heart and to decrease them below this level. The MAP at eye level increases by 20–25 mm Hg per g, so that it becomes 170 mm Hg at $-3g_z$. The *venous* pressures at eye level take several seconds to rise to a plateau, as there is a time delay involved in transcapillary flow. The venous pressure at eye level at $-3g_z$ rises to about 100 mm Hg.

1.4.3 Carotid Sinus and Cardiac Output. The increase in MAP in the neck stimulates the baroreceptors of the carotid sinus. The response is bradycardia and generalized arteriolar vasodilation. Cardiac arrhythmias usually occur on exposure to $-g_z$ over 1g. The arrhythmia reduces the cardiac output, and this effect and the dilation oppose the direct pressure rise due to the acceleration.

1.4.4 Cerebral Circulation. The increase in MAP produced by $-g_z$ is balanced internally by a similar increase in the pressure of the cerebrospinal fluid, so that there is no risk of rupture of vessels in the skull. The rise in MAP, venous engorgement, and the carotid sinus effect combine to reduce progressively the arteriovenous pressure differences across the cerebral vascular bed. The cerebral blood flow progressively slows and mental confusion and LoC result.

1.4.5 Pulmonary Effects. Negative acceleration produces a headward displacement of the diaphragm, and the vital capacity and residual capacity are reduced. Negative acceleration also reduces pulmonary ventilation. In $-g_z$ the apical region of the lung is better ventilated and perfused than the basal region. Indeed, most of the lung remains perfused under $-g_z$. Acceleration atelectasis induced by $-g_z$ occurs at the apices of the lung.

1.5 Effects of Acceleration on Arterial Wall Mechanical Properties

The viscoelastic behavior of arterial walls primarily reflects the properties of the fibrous components—smooth muscle, elastin, and collagen. Table 2 gives the amounts, quantitated by chemical, physiological, and histologic methods. The arterial wall has three layers—intima, media, and adventitia. The intima layer is formed by a single layer of endothelial cells, which holds the blood cells within the lumen and controls the passage of water and other substances from the plasma into the tissue.

A layer of elastic tissue—the internal elastic membrane—separates the intima from the media. The media is formed of smooth mus-

Table 2. Viscoelastic behavior of arterial walls.

	Arterial wall composition (% mass)		
Location	Elastin	Collagen	Smooth muscle
Arch of aorta	41	18	35
Main pulmonary artery	21	9	65

Table 3. Elastic moduli for elastin and collagen in carotid arteries.

Species	Elastin (dyne cm^{-2})	Collagen (dyne cm^{-2})
Dog	3.6×10^6	800×10^6
Rabbit	6.2×10^6	2400×10^6
Rat	5.6×10^6	1100×10^6
Human	4.5×10^6	1000×10^6

cle cells. The fibrous proteins—elastin and collagen—and the carbohydrates of the media are all secreted by smooth muscle cells.

The outer layer of the artery wall, the adventitia, is composed of fibroblast cells and their associated intercellular collagen and glycosaminoglycans. Table 3 summarizes the values of elastic moduli for elastin and collagen in carotid arteries for humans and three other species. The results of Roach and Burton (1957) indicate that the passive mechanics (no smooth muscle activity) of these vessels is closely related to the elastin content at low strain and to the collagen content at high strain (cf. BIORHEOLOGY).

In a published analysis of arterial wall mechanical properties by Westerhof and Noordergraff (1970), the incremental elastic modulus of arteries was assumed to be governed by an equation of the form

$$E=W_eE_e+f_cW_cE_c, \tag{3}$$

where E_e is the elastic modulus of elastin, E_c is the elastic modulus of collagen, W_e is the fractional mass content of elastin, W_c is the fractional mass content of collagen, and f_c is the fraction of collagen fibers supporting the wall load at a particular value of wall strain. The analysis assumes that at low strain, $f_c \to 0$ (i.e., no collagen supports the wall load), whereas at high strain, $f_c \to 1$ (i.e., all collagen fibers are recruited). Using Eq. (3), experimentally measured values for E_e and E_c, and the fractional tissue content, one can compute f_c for a range of arterial pressure (Fig. 5). Defining strain as

$$\epsilon=\text{strain}=(l_1-l_{10})/l_{10}, \tag{4}$$

where l_{10} is the unstretched (circumferential) length of wall element and l_1 is the stretched length, it is found that at low strains ($\epsilon<1.3$) the aortic wall behaves like smooth muscle, at intermediate strains ($1.3<\epsilon<1.7$) the aortic wall behaves like elastin, and at high strains ($\epsilon>1.8$) the aortic wall behaves like collagen.

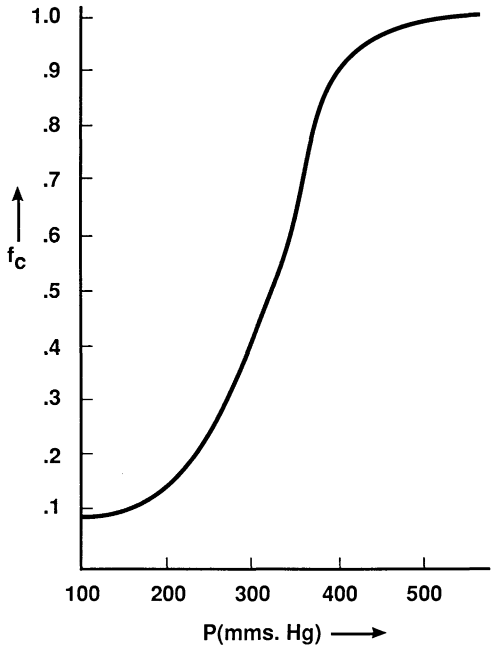

FIG. 5. Collagen recruitment fraction (f_c) versus arterial pressure *P*. [Redrawn from McCormack (1984).]

Gow and Taylor (1968) showed that for canines the aortic elastic modulus varied with MAP as

$$\text{modulus} \propto (\text{pressure})^{2.8}. \tag{5}$$

For the human femoral artery McCormack (1981) found a similar relation, namely,

$$\text{modulus} \propto (\text{pressure})^{2.6}. \tag{6}$$

For a $+g_z$ of 1, the MAP in the femoral is about 100 mm Hg, while at a $+g_z$ of 5, the pressure would be about 500 mm Hg. Using Eq. (3) for humans, the predicted modulus at $-5g_z$ would be 197×10^6 dyne cm^{-2}. This would imply a very stiff artery with the recruitment of many collagen fibers.

1.5.1 Arterial Wall Viscoelastic Properties. The arterial wall displays viscoelastic properties that differ from elastic material in that it exhibits the following.

Creep: The time-dependent deformation exhibited by an arterial strip subject to a constant stress or load. Figure 6 shows a general creep reference curve. Linear viscoelastic theory treats the total deformation, δ_1, as the sum of three components,

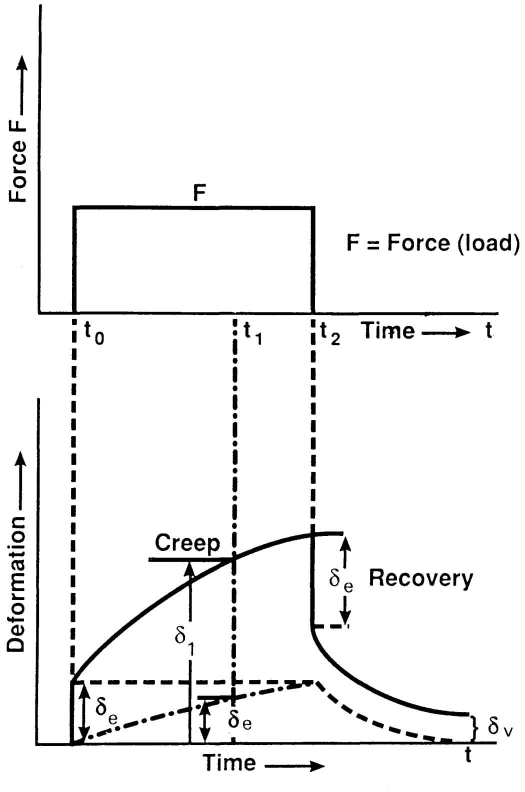

FIG. 6. General creep curve decomposed into different components; δ_e denotes elastic part of deformation, δ_{de} is the delayed part of deformation, δ_v represents the viscous component of deformation, and δ_1 the total deformation.

$$\delta_1=\delta_e+\delta_{de}+\delta_v, \qquad (7)$$

where δ_e is the *elastic* contribution, instantaneous and recoverable, δ_{de} is the *delayed* contribution, time dependent and recoverable, and δ_v is the *viscous* contribution, permanent and nonrecoverable.

Stress Relaxation: The time-dependent stress or load in an arterial strip subject to a constant *deformation.*

Hysteresis: This involves a nonelastic property of a material characterized by a particular stress-strain path during stress or load application, but taking a different path when the stress is removed (Fig. 7). The displacement remaining when the load is removed is called the "permanent set." Hysteresis can be related to creep or stress relaxation, depending on whether force or length is chosen as the independent variable.

1.5.2 Arterial Viscoelasticity—Complex Young's Modulus. The general definition of Young's modulus in the frequency domain is

$$E_c(s)=E_0 \prod_{l=1}^{M} b_l \prod_{n=1}^{N} (s+a_n) \times\left[\prod_{n=1}^{N} a_n \prod_{l=1}^{M} (s+b_l)\right]^{-1}, \qquad (8)$$

where $s=jw$; w is angular frequency; $E_c(s)$ is the complex Young's modulus; E_0 is the zero-frequency modulus ($s=0$); and N, M, a_n, and b_n are constants, still to be determined. The form of this expression is such that $E_c(s)$ reduces to E_0 at low frequencies.

For stress to be bounded, the condition $N \leqslant M$ must hold. For length to be bounded in creep experiments, the condition $M \leqslant N$ must hold, and, thus, $N=M$. Moreover, stress relaxation and creep measurements are nonoscillatory. Thus, a_n and b_n are greater than or equal to zero and are real.

1.5.3 Hysteresis. In high-performance aircraft maneuvers, the viscoelastic characteristic of prime interest is the hysteretic behavior of the arterial wall. In a typical maneuver, the stress builds to a maximum, re-

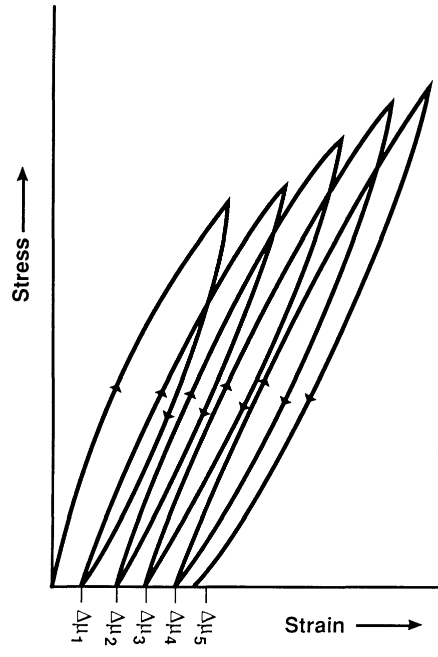

FIG. 7. Stress-strain hysteresis; $\Delta\mu_n$ denotes the permanent sets.

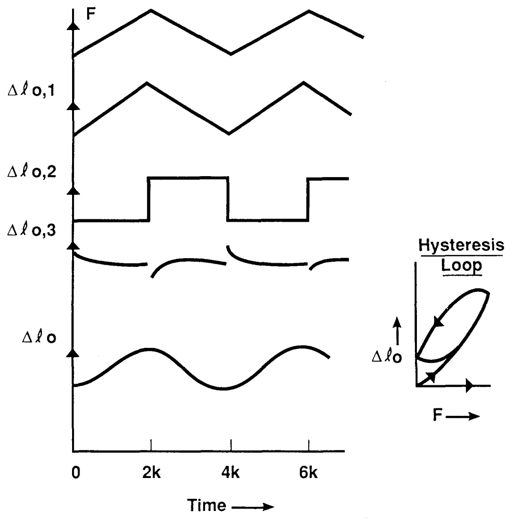

FIG. 8. Construction of hysteresis loop; $\Delta\ell_{0\ell}$, $\ell = 1,2,3$, are extension components versus time, $\Delta\ell_0$ is total extension versus time, and F is applied force versus time.

mains there for a finite time, and then falls to its initial value. The strain lags but follows the same pattern. Force, or length, can be used as the independent variable. One can use various shapes for the time variation of this variable—square wave, trapezoidal, triangular, etc. Consider an applied force that increases linearly with time up to $t=2k$ and then decreases linearly. At $t=4k$ the process then repeats. The applied force is thus

$$F(t)=F_0\left(tu(t)+2\sum_{m=1}^{\infty}(-1)^m(t-2mk)\times u(t-2mk)\right), \tag{9}$$

where $u(t)$ is a unit step function. Figure 8 illustrates the qualitative shape of $F(t)$ (upper curve), the resulting strain, Δl_0 (lower curve), and three components ($\Delta l_{0,j}$) thereof (middle three curves) versus time. The abscissas of each quantity at $t=0$ have been arbitrarily displaced from the origin for clarity and their magnitudes are arbitrary.

The Laplace transform of $F(t)$ is

$$F(s)=F_0 \tanh ks/s^2. \tag{10}$$

Using the basic definition for the complex Young's modulus $E_c(s)$ gives

$$E_c(\omega)=l_0F(\omega)/A\Delta l_0(\omega), \tag{11}$$

where $F(\omega)$ is the applied force, A is the initial cross-sectional area of the strip of material, and $\Delta l_0(\omega)$ is the change in length.

If a force of magnitude F_0 is suddenly applied to a strip of material (arterial wall), then it can be shown (McCormack, 1984) that in terms of the parametric constants used in Eq. (8), the change in length at $t=0$ is given by

$$\Delta l_0(0)=l_0F_0\prod_{n=1}^{N}a_n\left(AE_s\prod_{n=1}^{N}b_n\right)^{-1}, \tag{12}$$

whereas as for $t\to\infty$

$$\Delta l_0(t)=l_0F_0/AE_s. \tag{13}$$

Thus, Δl_0 jumps to the value given by Eq. (12) at $t=0$ and then creeps to the value given in Eq. (13). Inserting Eqs. (13) and (10) into Eq. (11) gives

$$\Delta l_0(s)=\frac{F_0l_0}{AE_s}\tanh ks\left(\frac{1}{s}-\frac{1}{s}\sum_{n=1}^{N}P_n+\sum_{n=1}^{N}\frac{P_n}{s+a_n}\right), \tag{14}$$

where P_n is constant.

Transforming this expression to the time domain yields

$$\begin{aligned}\Delta l_0(t)=\frac{F_0l_0}{AE_s}\Bigg[&\left(tu(t)+2\sum_{m=1}^{\infty}(-1)^m(t-2mk)u(t-2mk)\right)\\&-\left(\sum_{n=1}^{N}P_n\right)\left(u(t)+2\sum_{m=1}^{\infty}(-1)^mu(t-2mk)\right)\\&+\sum_{n-1}^{N}P_n\left(e^{-a_nt}u(t)+2\sum_{m=1}^{\infty}(-1)^me^{-a_n(t-2mk)}u(t-2mk)\right)\Bigg].\end{aligned} \tag{15}$$

Thus we can write the extension as a sum of three components,

$$\Delta l_0(t) = \Delta l_{0,1}(t) + \Delta l_{0,2}(t) + \Delta l_{0,3}(t). \qquad (16)$$

Figure 8 shows the graphical construction of $\Delta l_0(t)$ from its parts, $\Delta l_{0,1}$, $\Delta l_{0,2}$, and $\Delta l_{0,3}$. Δl_0 does not return to zero when the force returns to zero. The remaining extension in the absence of a net force is called the *permanent set*. It is not permanent in the true sense of the word, as it will decay with a time constant that is the same as the creep behavior of the material—a_1, for example, in a simple one–time-constant material. For this simplified case ($N=1$), $a_1 \neq b_1$ and $b_1 > a_1$, the above equations give at short times

$$\Delta l_0(t) = F_0 l_0 t / AE_s \qquad (17)$$

and at long times

$$\Delta l_0(t) = F_0 l_0 (t+D)/AE_s, \qquad (18)$$

where

$$D = [(b_1/a_1) - 1], \; a_1 \neq b_1. \qquad (19)$$

The permanent set after a hysteresis loop is, therefore, given by

$$\text{permanent set} = \frac{F_0 l_0 D}{AE_s} = \frac{F_0 l_0}{AE_s}\left(\frac{b_1}{a_1} - 1\right). \qquad (20)$$

Values of a_1 for the aortic wall have been calculated from the data of Hardnung (1953), Bergel (1961), Learoyd and Taylor (1966), and Tucker *et al.* (1969). The average of their values is 5.8 s^{-1}. For other arteries the data of Bergel (1961) and Learoyd and Taylor (1966) lead to a value for a_1 of 4.0 s^{-1}.

1.6 Calculation of Arterial Wall Performance in Aircraft Maneuvers

As stated previously, in aircraft maneuvers the viscoelastic characteristic of prime interest is the hysteretic behavior of the arterial wall. In typical maneuvers the stress builds to a maximum, remains there for a period, and then falls to its initial value. The strain lags, but follows the same temporal pattern.

It is usual procedure to enter the loop so that $+g_z$ is experienced. Below a fractional extension, or strain, ϵ of less than 1.3, it is the smooth muscle component in the arterial wall that is predominantly active. This is the situation at normal arterial pressures that produce strains in the region of $\epsilon \geqslant 0.3$. The smooth muscle is, of course, under autonomic nerve control.

As pointed out earlier, if the strain ϵ is over 1.7, collagen is predominant; the strain does *not* return to its initial value, and a permanent set is left. As noted above, the set is only quasipermanent because it decays slowly with a time constant of the order of 100 s.

McCormack (1984) predicted the number of sequential maneuvers (with less than 1 min separating the sequences) that would cause the complete failure of the collagen and rupture of the arterial wall. These are catastrophic events leading immediately to LoC (loss of consciousness). Step changes of 100% in collagen recruitment, at 6*g* for the external carotid and 5*g* for both the abdominal aorta and pulmonary arteries, will be considered. At these baseline *g* values, the elastic modulus is 400×10^6 dyne cm^{-2} and the number of sequential maneuvers to extend *all* the collagen by 20% can be computed. At 10*g* only 10 sequential maneuvers are required to cause catastrophic rupture of the pulmonary artery. The conclusion inferred from the above calculations is that in combat maneuvers in contemporary high-speed aircraft, significant damage to the collagen component in the arterial wall is quite likely.

2. PHYSIOLOGICAL AND BIOLOGICAL RESPONSES TO WEIGHTLESSNESS

2.1 Physiological Responses to Weightlessness

Figure 9 shows the time course of physiological changes observed in human adaptation to the microgravity environment for near-Earth orbit. It is assumed that all physiological systems are in a state of homeostatis in the 1*g* environment before entering orbit. While most systems appear to establish a new equilibrium at 0*g* steady state, calcium loss, loss of lean body mass, and the effects of exposure to ionizing radiation increase continually (Johnston, Dietlein, and Berry, 1975; Yorub'yer, Gazenko, and Shulzenko, 1986).

In readapting to the 1*g* environment, however, various physiological systems behave differently (Fig. 10). The variable value indicated as the clinical horizon is the value above which a clinically observable change occurs in the subject. The bone calcium re-

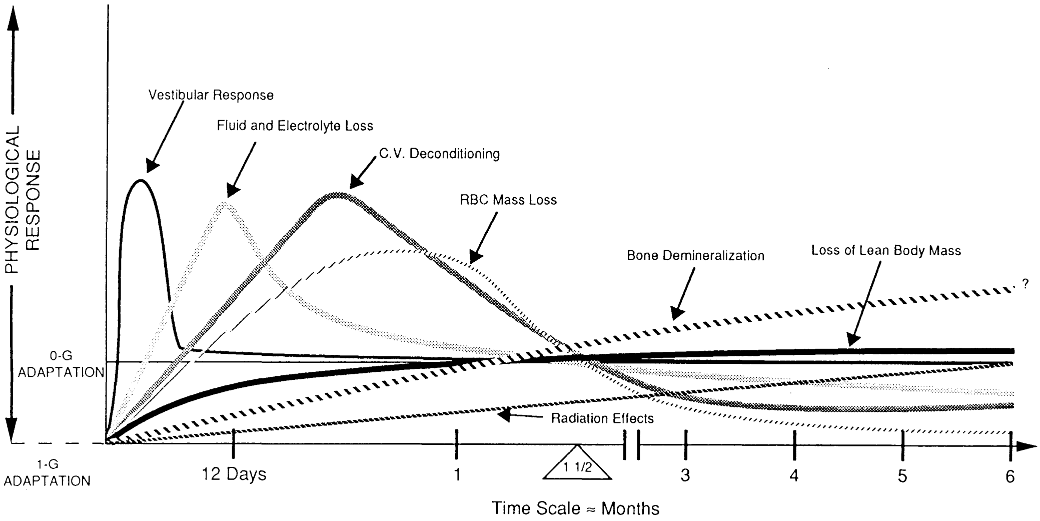

FIG. 9. Time course of adaptation to microgravity. RBC, red blood cells; C.V., cardiovascular.

covery curve is not shown because sufficient postflight data is not available. Evidence indicates that most of the bone loss is recovered (as indicated by bone density measurement), but the question of whether or not bone quality changes permanently (as measured by bone tensile strength, for example) has not been answered (Rambaut and Johnston, 1979).

Table 4 lists observations of selected systems from Skylab experiments. Three physiological systems are examined in some detail here, as they have been investigated extensively and are of significance operationally. A fourth system is considered in the context of significant potential change in 0*g*—the gastrointestinal system.

2.2 Cardiopulmonary System

Humans are a 1*g* species and the cardiopulmonary system has evolved an array of mechanisms to counter the stress of gravity. In adapting to the 0*g* of space flight, the cardiopulmonary system changes significantly. Probably the most overt symptom is the cephalad fluid shift.

2.2.1 Fluid Shifts. When the body experiences weightlessness, about 1.5–2.0 L of fluid move cephalad from the lower extremities. Visual evidence of this shift is evident in periorbital puffiness, facial edema, and thickening of the eyelids observed in astronauts and cosmonauts (Thornton, Hoffler, and

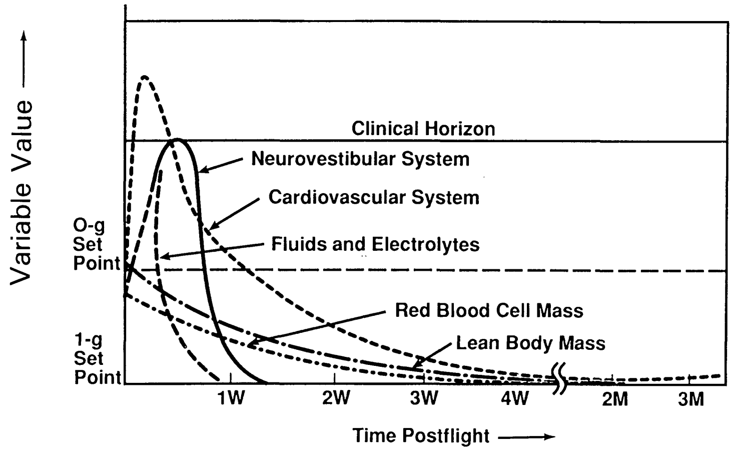

FIG. 10. Postflight time course (w=weeks; m=months) of physiological shifts during readaptation to 1*g* environment. Clinical horizon represents the level above which the variable produces a clinically observable change.

Table 4. List of effects on selected physiological systems from Skylab experiments.

Fluid-Electrolyte System
Rapid loss of water early in flight.
Much larger increase in leg fluid volume compared to bed rest.
Absence of measurable diuresis following headward fluid shifts.
Unusual combination of hormonal changes compared to bed rest and water immersion.
Presence of a hyponatremic plasma maintained throughout flight.
Increased water and sodium excretion throughout mission without continuous body losses.
Difference in ADH (anti-diuretic hormone) response between missions.
Decreased evaporative water loss in presence of hypobaric atmosphere.
Disproportionate losses of potassium and cell water.
Cardiovascular System
Acute cardiovascular disturbances expected from headward fluid shifts were not observed.
Decreased orthostatic tolerance as measured by lower-body negative pressure tests.
Maintained aerobic capacity during flight in spite of decreased blood volume and decreased orthostatic tolerance.
Decreased aerobic capacity immediately following flight.
Differences in aerobic capacity among the three crews.
Erythropoietic System
Inflight loss of red cell mass.
Delay in postflight recovery of red cell mass on shortest mission.
Significant differences in rates of red cell loss among the three crews.
Musculoskeletal System
Progressive loss of calcium from body.
Urinary calcium stabilizes at elevated level while focal calcium increases progressively with flight duration.
Changes in PTH (parathyroid hormone) during space flight not similar to changes in bed rest.
Rates of nitrogen losses appear to decrease as a function of flight duration and increasing levels of exercise.
Body Composition
Rates of water, muscle, and fat losses appear to be disproportionate to ratio in body tissues.
Body mass losses decrease with flight duration.
Water losses appear independent of flight duration.
Large differences in body composition changes between the three Skylab missions.
Energy requirements in 0g and 1g appear similar.

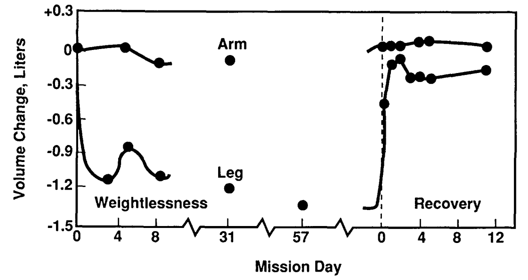

FIG. 11. Changes in left limb volumes of Skylab 4 Commander. [Redrawn from Thornton *et al.* (1977).]

Rummel, 1977). Efforts to quantify the shift have focused on measuring changes in calf girth and leg volume. In the Skylab and Apollo–Soyuz flights, circumferential measurements of the lower extremities were performed preflight, in flight, and postflight.

Figure 11, limb volume data (arm and leg) for the Skylab 4 commander, shows that the lower but not upper limb volumes decrease in flight and return to preflight values after returning to the 1*g* environment (Thornton, Hoffler, and Rummel, 1977). Observations in later flights showed that the major shift in fluid volume occurs rapidly upon entry into the 0*g* environment. The *rate* of this fluid shift appears to follow an exponential course, reaching a maximum in approximately 24 h and a new steady state in 3–5 days. Calf volume measurements performed in the Soviet Salyut-6 mission show similar behavior, with steady state achieved on the 12th day of the flight. For 140-, 175-, and 185-day missions, variations in leg volume follow an oscillatory time course of loss and recovery. The renal and endocrine systems are known to be involved in the adaptation process (Leach, 1983; 1984).

2.2.2 Decreased Orthostatic Tolerance. During the immediate postflight period, decreased orthostatic tolerance has been observed consistently. Symptoms range from a higher heart rate and decreased arterial pulse pressure to fainting (syncope). Tests commonly used to quantify this phenomenon are the tilt test, stand test, and lower body negative pressure (LBNP). The latter involves the application of a reduced pressure around the lower extremities. This will tend to reduce the cephalad fluid shift, thereby minimizing the loss of orthostatic tolerance on returning to a

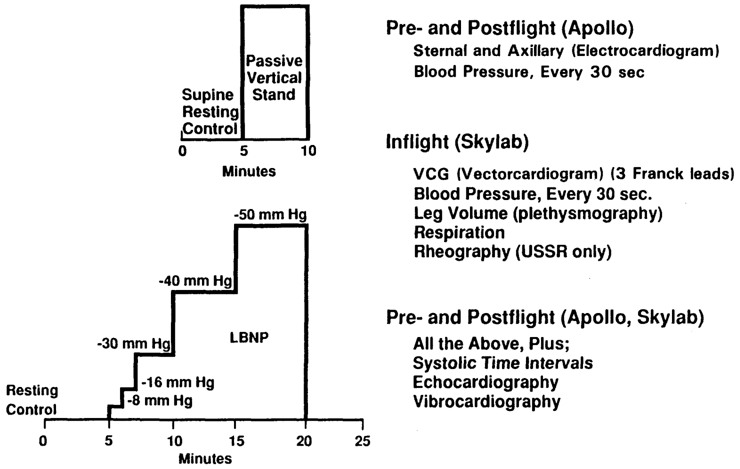
Supine
Resting
Control
Passive
Vertical
Stand
0
5
10
Minutes
-50 mm Hg
-40 mm Hg
-30 mm Hg
LBNP
-16 mm Hg
-8 mm Hg
Resting
Control
0
5
10
15
20
25
Minutes
Pre- and Postflight (Apollo)
Sternal and Axillary (Electrocardiogram)
Blood Pressure, Every 30 sec
Inflight (Skylab)
VCG (Vectorcardiogram) (3 Franck leads)
Blood Pressure, Every 30 sec.
Leg Volume (plethysmography)
Respiration
Rheography (USSR only)
Pre- and Postflight (Apollo, Skylab)
All the Above, Plus;
Systolic Time Intervals
Echocardiography
Vibrocardiography

FIG. 12. Protocols used in assessing orthostatic tolerance. [Redrawn from Hoffler and Johnston (1975).]

1*g* environment. Figure 12 describes the protocols used for the stand and LBNP tests and the cardiovascular parameter measurements.

2.2.3 Lower Body Negative Pressure. This involves the application of negative pressure to the lower part of the body resulting in a footward displacement of body fluid. The response can be measured by such parameters as change in leg volume (ΔLV) and changes in heart rate. Figure 13 shows the change in the leg volume and heart rate during pre- and postflight conditions of the Apollo Commander using the LBNP protocol (Hoffler and Johnston, 1975). These and other observations, for both Soviet and U.S. missions, showed that during flight the LBNP presents greater stress to the cardiovascular system than the *same* levels of LBNP preflight. This is presumably due to change attending the large cephalad movement of body fluids in flight. The change in immediate postflight orthostatically stressed heart rates from preflight mean reference value, as a function of change in total blood volume, has also been studied (Nicogossian, Leach-Huntoon, and Pool, 1989). Results indicate that the depleted blood volume is at least one factor in the exaggerated cardiovascular response to LBNP observed postflight. These observations led to the countermeasure of fluid loading of shuttle crews before returning to Earth.

2.2.4 Venous Compliance. Measurements of venous compliance were performed for the crew of Skylab 4 preflight, in flight, and postflight (Fig. 14) (Thornton and Hoffler, 1977). Vascular compliance increased in flight (under microgravity conditions) but did not attain maximum value until 10 days inflight. After approximately six weeks compliance fell and rapidly returned to preflight val-

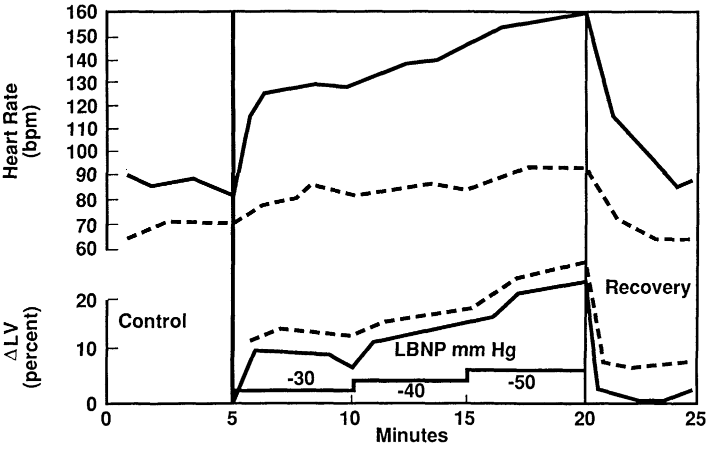

FIG. 13. Heart rate (bpm) and leg volume change, ΔLV, during lower body negative pressure (LBNP) test protocol (25 min). Preflight mean curves are dashed; the first postflight test data 3 h after splash down are solid lines; Apollo 8 Commander. [Redrawn from Hoffler and Johnston (1975).]

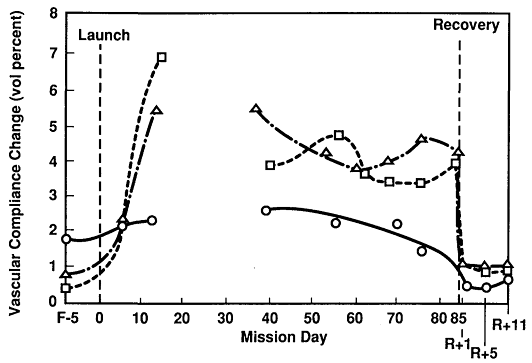

FIG. 14. Skylab 4 crew vascular compliance. [Redrawn from Thornton *et al.* (1977).] (○—○ Commander; □---□ scientist pilot; △--△ pilot.)

ues on return to ground. This change in mechanical properties of the venous walls in 0*g* has not yet been explained.

2.3 Bone and Calcium Metabolism

Reduction of gravitational forces acting on the weight-bearing structures of the body leads to a loss of calcium (Rambaut and Johnston, 1979) and other minerals essential for bone strength and integrity (Hattner and McMillan, 1968; Mack, 1971). Bone mineral losses are not observed from the distal compact radius, however. Soviet space flight measurements demonstrated that mineral loss occurs from the tubercle and plantar areas of the os calcis—mainly compact bone. Bone mineral loss increases approximately proportionally to mission length, varying from −0.9% to −19.8% over intervals from 75 days to 184 days (Stufakov, 1984). Calcaneal mineral recovery occurs gradually and at about the same rate as the mineral (Vogel and Whittle, 1976). The measured recovery was only partial in at least one Skylab 4 astronaut (50% replacement, 90 days postflight).

2.3.1 Calcium Balance Studies. Skylab studies showed that human exposure to microgravity is accompanied by increased excretion of calcium and phosphorus. Figure 15 illustrates the changes in urinary and fecal calcium content during Skylab 4. The calcium content of urine increased rapidly after entrance into a microgravity environment and plateaued after 30 days. Over the first 10 days in flight, preflight positive calcium balances decreased and eventually became negative. The rate of loss increased to almost 300 mg/day by the eightieth day of flight. For three Skylab 4 crewmen, the average loss (over 84 days) was 25 g of calcium from the overall body pool (about 1250 g). Rambaut and Johnston (1979) predicted, based on trends in calcium loss over the first 30 days in flight, that calcium loss for a 1-year mission might be sufficient to undermine bone strength.

The 4% loss observed in the os calcis after the 84-day mission represents a loss of about 100 mg of calcium. Thus, other weight-bearing skeletal sites account for most of the mineral loss. Unfortunately, bone mineral loss from nonweight-bearing bone structures has not yet been documented. Soviet experience with intensive exercise programs demonstrated that this countermeasure can reduce the bone mineral loss considerably (Gazenko, Grigor'yev, and Natochin, 1981). In recent 1-y flights, overall bone loss was kept to within 5%.

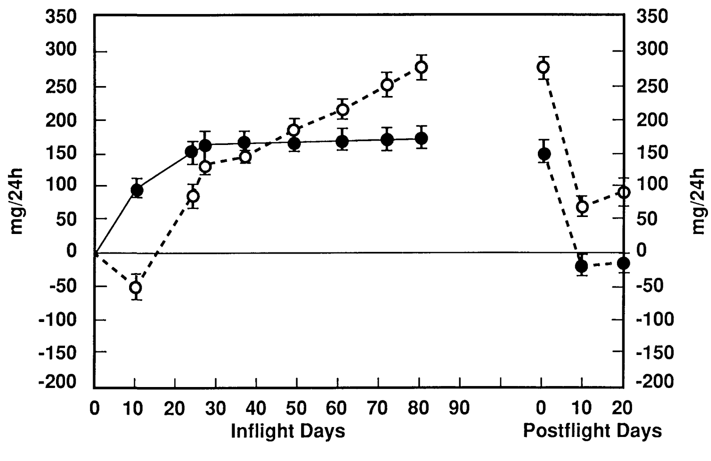

FIG. 15. Changes in urine (●—●) and fecal calcium (○---○) as a function of Skylab flight duration (mean ± SE). [Redrawn from Rambaut and Johnston (1979).]

Recovery of lost calcium began soon after returning to the $1g$ environment. Urinary calcium content decreased below the preflight baseline in 10 days postflight, but fecal calcium content remained elevated even after 20 days. Negative calcium balance also persisted for up to 20 days. Studies on the recovery of the os calcis mineral content postflight indicate that after several weeks (and sometimes up to months), the content returns to normal. This has been confirmed by bed rest studies.

2.3.2 Bed Rest Simulation. In bed rest the force of gravity on the longitudinal skeleton is reduced from $1g$ to $\frac{1}{6}\,g$. There are many uncontrollable variables involved in in-flight mineral balance studies. Bed rest studies, however, offer reliable and reproducible data. Such studies determined that calcium loss continues for at least 36 weeks with no evidence that a new steady state is being approached. Total body calcium decreases by 6 g per month after the first 30 days of bed rest, and by 9 months at least 50 g of calcium are lost.

The amount of mineral lost in bed rest subjects varies widely. This variability may be partially explained by two parameters: the *initial* os calcis mineral content, and the urinary hydroxyproline excretion rate. The most pronounced bone changes in $0g$ occur in weight-bearing bones. Mechanical stress apparently has a profound effect on bone metabolism and structure, as many studies involving bone strain measurement have shown (Hinsenkamp, 1981).

2.4 Biological Responses to Weightlessness

All life on Earth evolved in $1g$ environment. Hence, every cell and cell organelle has weight, however small. The basic question in gravitational biology is how nature has used weight to accomplish the objectives of life development.

All organisms must perceive the environment. The best-understood environment receptors are the chemoreceptors of the enteric bacteria, receptors that recognize a favorable chemical and ultimately enable the bacterium to swim toward the nutrient chemical. Many types of receptors have been investigated, including light, heat, moisture, osmolarity, and gravity. Unfortunately, gravity perception is one of the least-understood mechanisms.

2.4.1 Tropic Systems of Plants. Plants respond to both light and gravity. Light responses have been reviewed thoroughly elsewhere (Wareing, 1983). Compared to light receptors, there is considerably less information about the gravity-sensing component of plants (Bandurski and Schulze, 1983). It has, however, been shown that a plant can respond to gravity within minutes (Bandurski, Schulze, and Kaufman, 1984). In fact, the time period of the response in some cases is so rapid that it precludes the settling of a heavy organelle to the bottom of the plant cell. There is considerable literature on statoliths, which are formed by a heavy starchlike organelle. However, the time course of the response observed is so rapid as to make it unlikely that the statoliths are the major gravity sensing component in plants.

2.4.2 Gravitational Biodynamics. Many fundamental questions remain unanswered, including the following.

1. Is the perception of gravity a fundamental property of a cell ?
2. How does the absence of gravity or a reduction in gravity affect the molecular structure of the cell and hence the genome?
3. What happens to the mechanism of cell recognition under $0g$?
4. How is three-dimensional reaggregation of dissociated cells affected by $0g$?
5. Does $0g$ cause changes in the distribution of cell organelles during mitosis, meiosis, or budding?

The well-known law of universal gravitation in its simplest form states that every mass is attracted to every other mass. Hence, in response to the gravitational force, internal mechanical stress is produced in every biological system. In $0g$ this stress is zero. The question is, then, would a cell that has evolved in the presence of stress retain the same characteristics in a stress-free environment?

We note that, of the four types of forces in nature, only the electromagnetic and gravitational forces appear to play major roles in biological systems. Nuclear forces have no direct biological effects. Weak forces are responsible for beta decay and have no direct

Table 5. Comparison of thermal velocities for spherical particles executing Brownian motion in the absence of gravity at $T=300$ K with terminal velocities in a plasma using Stokes's law.

Radius (μm)	Thermal velocity ($\mathrm{m\ s^{-1}}$)	Terminal velocity ($\mathrm{m\ s^{-1}}$)
10^{-2}	1.4	10^{-10}
10^{-1}	4.5×10^{-2}	10^{-8}
1	1.4×10^{-3}	10^{-6}
10	4.5×10^{-5}	10^{-4}
100	1.4×10^{-6}	10^{-2}

biological effects. Electromagnetic forces are involved in molecular interactions, maintenance, and breakage of chemical bonds. Gravitational forces are the most pervasive, but the weakest of the fundamental forces.

The diffusion of molecules and very small particles is governed by random collisions with the local molecular constituents. It is a simple calculation to show the particle size at which the directional force of gravity overcomes this randomizing force. Molecular thermal motion is governed by the equation

$$\tfrac{1}{2}mv^2=\tfrac{3}{2}k_\mathrm{B}T, \tag{21}$$

where m is the particle mass (kg); T is the temperature (K); v is the particle velocity (m/s), and k_B is the Boltzmann constant, $1.380\,65\times10^{-16}$ ergs per kelvin.

Consider a sphere of plasma of radius r with density$=1.5\times10^3\ \mathrm{kg\ m^{-3}}$ at $T=300$ K (see Table 5). Under these conditions Eq. (21) implies that a ribosome has a thermal velocity of 1 m/s, a cell nucleus has a thermal velocity of 10 μm/s, and a free protist has a thermal velocity of only 1 μm/s. Using Stokes's law, the balance between a force of 1g and the viscous friction of a fluid allows one to calculate the terminal particle velocity (v)

$$\frac{4\pi r^3}{3}(\rho-\rho_a)g=6\pi\eta v r, \tag{22}$$

where η is the fluid viscosity, $\rho-\rho_a=0.5\times10^3\ \mathrm{kg\ m^{-3}}$, and ρ_a is the ambient mass density. For water at 20 °C, $\eta=10^{-3}\ \mathrm{kg\ m^{-1}\ s^{-1}}$; for blood at 37 °C, $\eta=3\times10^{-3}$. Comparisons given in Table 5 clearly demonstrate that gravitationally driven motion is insufficient to overcome the diffusional motion of a particle unless the particle diameter is over 10 μm, if the particle density is 50% greater than that of the surrounding medium.

2.5 Statistical Distribution of Particles under Gravitational Force

As noted above, random molecular motion competes with ordering produced by the force of gravity. If the average density in a cell is taken to be unity ($\rho_a=1$), and the density of a particulate unit is ρ, then for a volume v, the difference between the *downward* force of gravity and the buoyant *upward* force is $v(\rho-\rho_a)g$. At a temperature T the resulting statistical distribution is

$$n=n_0\exp[-v(\rho-\rho_a)gh/k_\mathrm{B}T], \tag{23}$$

where n_0 denotes the number of particulates per unit volume at a certain baseline level, and n is the number per unit volume at a height h above this baseline. For a bacterial cell with $h=1\ \mu$m at room temperature and taking the particulate to be the whole unit of DNA ($\rho=1.6\ \mathrm{g\ cm^{-3}}$ and $v=2\times10^{-5}\ \mathrm{cm^3}$), Eq. (22) gives

$$n/n_0=0.997.$$

Thus, the maximum intracellular difference in the statistical distribution produced by gravity for the largest and densest internal component of a cell is less than 0.3%. A similar calculation for ribosomes (volume$=4.2\times10^{-18}\ \mathrm{cm^{-3}}$, $\rho=1.5\ \mathrm{g\ cm^{-3}}$) gives an even smaller distribution gradient. One is forced to conclude that changes in the statistical distribution due to gravity for a single bacterial cell are insignificant. For a mammalian cell, however, 10 μm in diameter, the situation changes. For organelles, such as mitochondria, with a volume of approximately $10^{-12}\ \mathrm{cm^3}$ and a density of $1.1\ \mathrm{g\ cm^{-3}}$, Eq. (23) gives $n/n_0=0.09$ at room temperature. This is a significant difference and, hence, in this situation, the removal of the gravitational force could produce considerable intracellular changes.

The stratifying effect of gravity, however, is opposed by a second effect of gravity: convection, or streaming. This effect tends to mix material in a cell, distributing components equally. Nevertheless, it has been suggested that certain dynamic aspects of the cell may be exquisitely sensitive to the presence or absence of gravity.

The kinetics of transposable genes, the pairing and exchange of chromosome fragments, and the repair or misrepair of DNA could be influenced by gravity. The capacity for DNA to repair is intrinsically connected with life; in fact, life could not exist if DNA did not have a significant capacity to repair damage, because replication errors occur continually, as does radiation-induced damage. A basic property of a living system is self-assembly (Lima de Faria, 1986), as has been demonstrated experimentally (Fritz-Niggli, 1988). The process could be altered significantly in a weightless environment. In the D1 Spacelab Shuttle flight, a biostack experiment indicated synergy between gravity and radiation-induced cell damage (Bücker, Facius, and Horneck, 1986). A tentative explanation for this observation involves the effect of gravity on the DNA repair mechanism.

2.6 Effects of Gravitation on Membranes

All compartments of the body, both extracellular and intracellular, are delineated by biological membranes. The transport of all material (for example, enzymes, nutrients, and waste products) into and out of different compartments is controlled by membranes. Known physical properties of membranes include the basic unit, which is a pairing of lipid molecule; the electrical capacitance, the order of which is 1 μF/cm^2; the surface tension (values as low as 0.1 erg/cm^2 are known); and two intermolecular forces, *attraction* due to London dispersion force and *repulsion* due to electrostatic charges.

The London potential energy (in ergs) between "two cylinders" of lipids (Verwey and Overbeck, 1948) can be written as

$$U_a = 5\times 10^{-13} a/R, \tag{24}$$

where a is the radius of a cylinder (cm) and R is the surface separation (cm). The *repulsive* potential U_r can be written in the form (Verwey and Overbeck, 1948)

$$U_r = B \exp(-\kappa R), \tag{25}$$

where κ is the Debye–Hückel constant, and B is a constant (units of energy).

Calculations using these potentials, unfortunately, provide no direct information about the membrane behavior or the effects of gravity on membranes (Verwey and Overbeck, 1948).

2.6.1 Effects of Hydrostatic Stress on a Membrane. Hydrostatic stress caused by gravity does *distort* the cell's membrane integrity. For a cell 10 μm in diameter, pressure on the lower membrane due to gravity is about $10^{-3}g$, approximately 1 dyne/cm^2. The resulting strain in the membrane increases separation between molecules and may enhance membrane permeability. The increase in membrane permeability can be estimated as follows.

Consider a pressure of 1 dyne/cm^2 on a membrane 100 Å thick. This pressure causes the membrane to stretch and develop an elastic force. Weibull (1955) measured the changes in volume of a protoplast produced by changes in the concentration of sucrose. For a change in osmotic pressure (Δp) of 2×10^6 dyne/cm^2, the volume changes from 2.7×10^{-12} cm^3 to 2.0×10^{-12} cm^3. For a cell of radius 8.6×10^{-5} cm, this pressure causes the area to change by 1.54×10^{-8} cm^2. For 1 dyne/cm^2 pressure, the volume decreases by a factor of only 2.2×10^{-6}, corresponding to a volume change of roughly 3.5×10^{-14} cm^3 or to an area change of approximately 8.1×10^{-16} cm^2. This change is considered too small to have any significant effect. In comparison, we note that for a plant 1-m high the hydrostatic pressure on its lower membrane is about 10^5 dyne/cm^2. This pressure is significant and produces a relative change of approximately 0.1 in the cell's membrane area. In this case the average separation for a neighboring pair of molecules is increased by about 30%. Hence, the effect of gravity can be appreciable in plants of this size or larger.

2.6.2 Effects of Hydrostatic Stress on Active and Passive Transport through Semipermeable Membranes. The absorption of water and electrolyte occurs mostly in the jejunum and ileum across a semipermeable nonporous membrane with active transport. The absorption rates are a function of the osmolality of the intestinal contents. Vaughan (1960) showed that for a dog's intestine the absorption rates in the jejunum and ileum are linear functions of the osmolality (Fig. 16). Both sodium and water move against their concentration gradients, implying an active transport process. Figure 17 illustrates the most accepted mechanism for transport of sodium and water. Compartment 1 is the intestinal medium, 2 is the blood plasma, and 3 de-

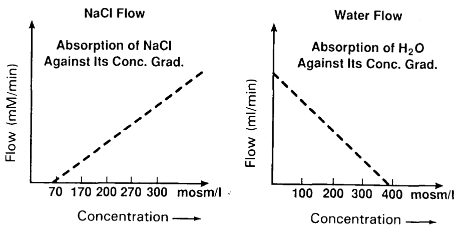

FIG. 16. Absorption flow of water (H_2O) and sodium (NaCl) versus intestinal osmolality. [Redrawn from Vaughan (1960).]

notes the epithelial layer of the intestinal wall (the membrane); A denotes a semipermeable nonporous membrane with active transport and B represents a passive (porous) and active membrane. It is assumed that no hydrostatic flow occurs across A, and no osmotic flow occurs across B.

The solution in compartment 3 would leak passively into compartment 2 under hydrostatic pressure. However, this process would not occur under 0*g* conditions. The balance equations for water and solute flows are as follows:

$$P_{AO}(C_3-C_1)=P_{BH}H+K_1, \quad (26)$$

$$T-P_{AS}(C_3-C_1)=P_{BH}HC_3+K_1C_3. \quad (27)$$

The corresponding 0*g* equations are as follows:

$$P_{AO}(C_3-C_1)=K_1, \quad (28)$$

$$T-P_{AS}(C_3-C_1)=K_1C_3, \quad (29)$$

where K_1 is the active transport of water (mosm s^{-1} cm^{-1}) (mosm is the abbreviation for milliosmole; see "Osmole" in the Glossary), H is the hydrostatic pressure (dyne cm^{-2}), T is the net active transport of solute (mosm^2 s^{-1} cm^{-1} L^{-1}), P_{AO} (L s^{-1} cm^{-1}) is the osmotic permeability of A to water, P_{AS} (mosm s^{-1} cm^{-1}) is the osmotic permeability of A to solute, P_{BH} (mosm cm s^{-1} dyne^{-1}) is the hydrostatic permeability of B to the solution, and C_i is the solute concentration in mosm L^{-1}. As Eqs. (26) and (27) differ from those in 0*g*, there could be significant differences in membrane transport (McCormack, 1989) in 0*g* (Gazanko, Grigor'yev, and Natochin, 1981) that have implications for organ processes ranging from the absorption of nutrients and medications to the function of the immune system. Unfortunately, no comparative experimental data exist on membrane transport characteristics in 0*g* and 1*g* environments. Furthermore, it is not known whether absence of gravity alters the osmotic permeability parameters.

3. ARTIFICIAL GRAVITY: A COUNTERMEASURE TO WEIGHTLESSNESS

3.1 Characteristics of the Rotating Environment

Several features of the rotating environment make artificial gravity different from the Earth's gravitational environment. These unique characteristics of artificial gravity fall into two categories—static and dynamic. The static characteristics include the magnitude of the artificial gravity, the gradient of the force in the radial direction, and the variation of the hydrostatic pressure in fluid systems caused by the gravitational gradient. The dynamic aspects include properties associated with motion: Coriolis acceleration, which occurs when a subject moves linearly with re-

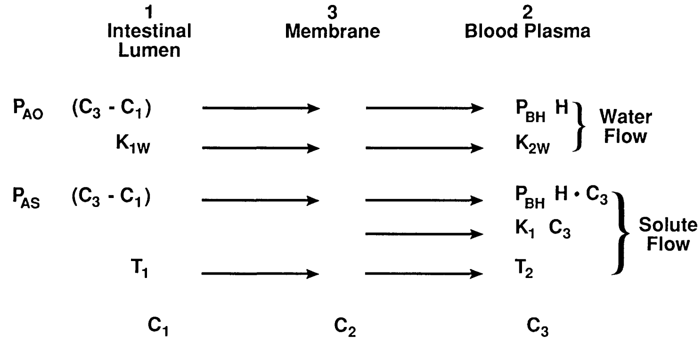

FIG. 17. Movement of sodium and water across a cellular membrane water. T denotes active transport flow of solute; P_{AO} and P_{AS} the osmotic permeability of A to water and solute, respectively; and P_{BH} is the hydrostatic permeability of B to solute.

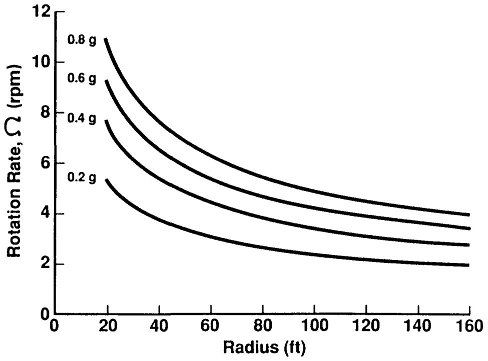

FIG. 18. Rotation rate [Ω (rpm)] versus radius (ft) for various *g* levels. [Redrawn from McCormack and Nicogossian (1987).]

spect to the rotating environment, and cross-coupled angular acceleration, which occurs when the subject rotates relative to the rotating environment. A major effect of rotation is that the apparent weight of an object is now a function of its position relative to the vehicle central axis of rotation. The object weight change is

$$\Delta W/W=(r_1-r_2)/r_2, \tag{30}$$

where subindices 1 and 2 imply two radial positions.

3.2 Artificial Gravity Level

Figure 18 shows the variation in rotation rate Ω (rpm) of the space habitat versus radius for a range of g levels. The curves tend to level off for radii greater than 70 ft. Adding 20 ft to a 90-ft-radius vehicle only reduces the rotation rate by about 0.3 rpm. At a vehicle rotation of 4 rpm and at g level of 1, a radius of 197 ft is required.

In the context of interplanetary travel, increasing vehicle size is expensive in terms of launch weight, inertia, momentum, and energy. It also introduces considerable difficulties in control, stabilization, and fuel requirements, to note but a few of the problems expected for large spacecraft. The variation in gravity gradient with vehicle radius also flattens as the radius increases. At $0.4g$ adding 20 ft to a 20-ft vehicle radius reduces the gravity gradient by $0.09g$/ft. Adding 20 ft to a 98-ft vehicle radius reduces the gradient by less than $0.007g$/ft.

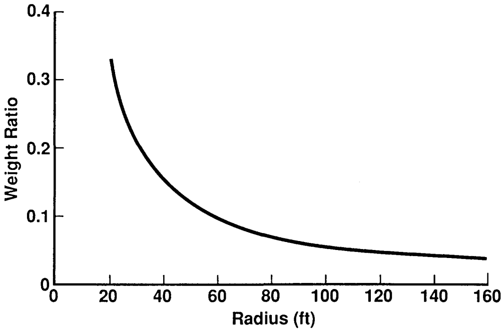

FIG. 19. Artificial weight change versus radius (6-ft change in radial position of object). [Redrawn from McCormack and Nicogossian (1987).]

Figure 19 illustrates the variation of the ratio of the weight of an object 6.6 ft off the floor to the weight when on the floor with vehicle radius. It is again evident that the advantage of increasing radius becomes insignificant for radii greater than 80 ft.

3.3 Coriolis Forces

This title is restricted to the force that is operative when *linear* motion is attempted in a rotating environment. When an angular motion is carried out in a rotating environment, cross-coupled angular acceleration occurs. This case is called the Coriolis effect.

When an object (animate or inanimate) moves tangentially or radially in a rotating environment, it experiences a force normal to the relative motion. Figure 20 shows the ratio of the tangential Coriolis acceleration to artificial gravity versus the radius of vehicle rotation for a range of artificial gravities. The

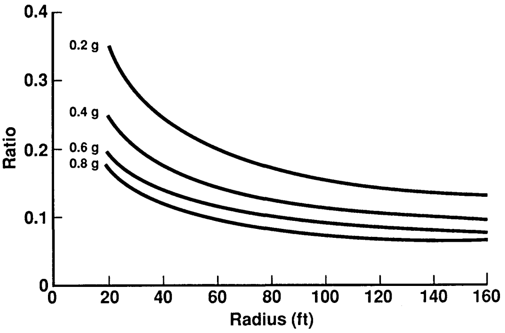

FIG. 20. Ratio of tangential Coriolis acceleration to artificial gravity versus radial position. [Redrawn from McCormack and Nicogossian (1987).]

Table 6. Potential criteria for artificial gravity.

Gravity level
  $g_{\max} \geqslant r\omega_1^2 + 2r\omega_1\omega_v + r\omega_v^2 \geqslant g_{\min}$
Coriolis forces
  Tangential movement
  $(1 + C_1)r\omega_v^2 \geqslant r\omega_1^2 + 2r\omega_1\omega_v + r\omega_v^2 \geqslant (1 - C_1)r\omega_v^2$
  Radial movement
  $2\dot{r}/r\omega_v \geqslant C_2$
  Dropped objects
  $r(\sqrt{r^2 - r_1^2}/r_1 - \tan^{-1}\sqrt{r^2 - r_1^2}/r_1) \leqslant C_3$
Gravity gradient
  Change in object ωeight $(r - r_1)/r \leqslant C_4$
  Along body $(r - r_h)/r \leqslant C_5$
  Hydrostatic pressure $(r_{\text{heart}} + r_h)/(r + r_h) \geqslant C_6$
Cross-coupled angular accelerations
  $\omega_v \leqslant C_7/\omega_h$

curves were computed for a subject having a radial velocity of 2 ft/s. Once again, a decrease in the effect as vehicle radius increases is obvious. At 0.4*g*, for example, adding 20 ft to a 20-ft radius vehicle reduces the ratio by just over 25%. Doubling a 98-ft radius vehicle, however, reduces the ratio by less than 10%. With tangential motion, changes in the radial force (weight) also occur. Similar graphs can be drawn for the ratio of radial Coriolis acceleration to artificial gravity versus vehicle radius. Again, a rapid reduction in change in the ratio with increasing radius occurs.

3.4 Human Performance in Artificial Gravity

The peculiar characteristics of the rotating environment would alter human performance and, in all likelihood, would degrade it to some extent. Over a period of exposure, adaptation may occur and performance could improve. Factors of concern are the radius and rate of rotation that allow effective performance and habitable conditions, and the cost of attaining this condition in terms of factors such as weight, moments of inertia, angular momentum, rotational energy, and stability.

The various aspects of human performance that may be influenced by artificial gravity are gross motor performance, including self-locomotion and material handling, and fine motor performance involved in dexterity and eye-hand coordination in the presence of head motion. Table 6 lists potential criteria for artificial gravity.

3.4.1 Self-Locomotion. Habitability of the space vehicle, with respect to orientation and arrangement of equipment in the vehicle, depends directly on self-locomotion capability. Figure 21 illustrates the influence of artificial gravity forces on a moving subject in a rotating spacecraft. In radial movement, both away from or toward the center of rotation, the tangential components of the Coriolis force operate. For tangential movement, both with and against the direction of rotation, changes in the radial forces (weight) occur.

Axial motion produces no Coriolis effect unless the limbs move laterally or the body sways. Figure 22 shows the four boundaries expected to limit tangential motion on the floor of a rotating vehicle. The 1*g* limit is shown for a walking speed of 3.3 ft/s (a reasonable upper walking speed). Tests on the Langley rotating Space Station simulator

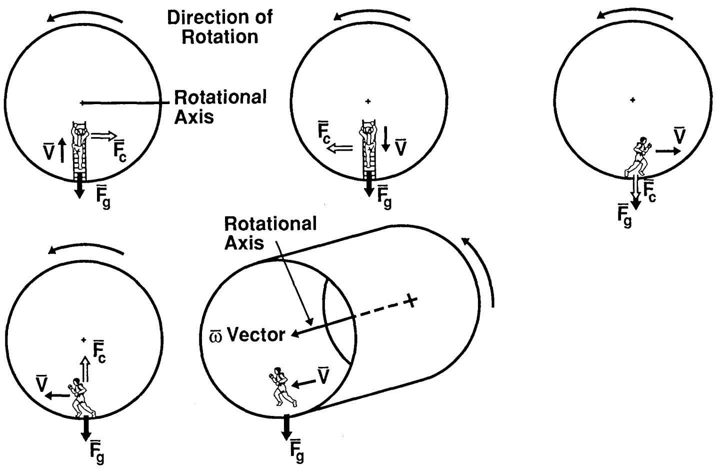

FIG. 21. Influences of artificial gravity on locomotion. $\overline{\omega}$ denotes rotation rate, $\overline{F}_g$ is the gravitational force, $\overline{F}_c$ the Coriolis force, and $\overline{V}$ the man's velocity.

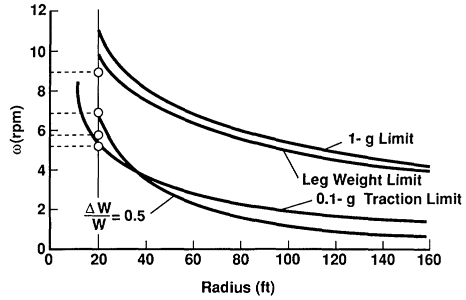

FIG. 22. Rotation rate and radius boundaries for tangential locomotions at a rate of 3.3 ft/s. [Redrawn from McCormack and Nicogossian (1987).]

showed that leg heaviness at this weight could become objectionable (Stone, Piland, and Letko, 1970; Hewes and Spady, 1964). The lower limit of 0.1*g* is set by the necessity for sufficient traction to allow efficient motion. Along the 20-ft radius line, the four points at which simulated studies were made on the Langley simulator are shown. The upper point for walking is 0.5*g*; at this point leg heaviness was noted and found to be undesirable. The lowest point is for $\frac{1}{6}$ *g* at which traction was poor.

At the fourth boundary,where $\Delta W/W=0.5$, the change in weight due to the relative motion in walking at 3.3 ft/s does not exceed half the artificial gravity weight. The best level for the four gravity levels tested (on a limited number of subjects) was 0.3*g*. It is important to note that the curves in Fig. 22 and the data of Stone, Piland, and Letko (1970) were for a curved floor of constant radius.

The boundaries for radial motion of a crew member can be established similarly (Stone, Piland, and Letko, 1970). One boundary is for a Coriolis force at a radial motion at 2 ft/s, 0.3 times the artificial weight. The upper boundary is for 0.3 times the Earth weight. In this case the Coriolis force acts tangentially, and arm loading, rather than leg loading, is involved. The choice of the 0.3 factor is arbitrary at this stage.

3.4.2 Material Handling. Movement in the rotating environment differs from movement in the Earth's gravitational environment. Objects will (a) change weight with radial position according to the equation $\Delta W/W=(r_1-r_2)/r_2$, (b) experience Coriolis acceleration when moved, and (c) experience cross-coupled angular acceleration when rotated.

Table 7. Lower limit of vehicle radius.

ω=rotation rate (rpm)	Radius (ft)
2	150
4	75
6	50

A reasonable boundary for radial motion is that the tangential Coriolis force is never greater than 0.25 times the artificial weight. Another boundary is for tangential motion such that ΔW will be less than 0.25 times the artificial weight. A third boundary is that the weight of an object when raised 6.5 ft (2 m) does not decrease to less than half its original artificial weight. For tangential motion this criterion may be written as

$$\frac{\text{Coriolis force}}{\text{Artificial weight}}=\frac{2\omega_r\omega_v}{\omega_v^2}=\frac{2v_r}{\omega_v r}\leqslant 0.25 \tag{31}$$

where v_r is the relative velocity of movement. For a relative velocity of 4 ft/s this criterion becomes

$r\omega_r > 32.$

If the rotation rate is fixed, the lower limit of the vehicle radius can be determined. Table 7 gives a few calculated lower vehicle radius limits.

3.5 Cross-Coupled Angular Accelerations

When a human subject makes a head movement out of the axis of rotation of the vehicle, cross-coupling of the angular accelerations occurs, stimulating the semicircular canals in the vestibular system. This stimulations results in perceptual misinterpretation of the path of head movement and, if strong enough, evokes symptoms of motion sickness. Adaptation to the cross-coupled angular acceleration generated by movement in rotating rooms has been studied by Stone and Letko (1964), Graybiel and Wood (1969), and Lackner and Graybiel (1986).

Each semicircular canal in the vestibular system is orthogonal to the other two and is essentially a torus filled with endolymph (a fluid having a specific gravity near unity). In a sense, each set of semicircular canals can be viewed as a gyroscopic component when the head is rotated inside a rotating vehicle. The canal lying *in* the plane of vehicle rotation is similar to a gyroscopic rotor spinning about its axis of symmetry. If the canal rotates about an axis normal to the vehicle rotation axis, then an *output* torque is generated about an axis perpendicular to both the vehicle rotation axis and the canal axis of rotation. This output torque (T_o) is given by

$$\mathbf{T}_o = I_v(\omega_v \times \omega_H), \tag{32}$$

where I_v is the moment of inertia of the rotating body about the rotation axis, ω_v is the angular velocity of the rotating body, and ω_H is the angular velocity of the semicircular canal involved (Graybiel and Wood, 1969). Hence, the torque magnitude is

$$T_o = I_v \omega_v \omega_H \tag{33}$$

and the resultant angular acceleration is

$$\dot{\omega}_{\text{res}} = \omega_v \omega_H. \tag{34}$$

Based on experiments performed at Langley in the 15-ft simulator, the torque value on the endolymph that resulted in the onset of malaise in all subjects was 4.40×10^{-10} ft lb (Stone and Letko, 1964). Reducing either ω_v or ω_H by 66% produced a tolerable state with a torque of about 3.0×10^{-10} ft lb. Figure 23 illustrates the rate of head movement versus the vehicle rotation rate for both torque values. Table 8 shows the corresponding values of rotation rates and angular accelerations. Experience with Slow Rotating Rooms (SRRs) has shown that a conservative criteria for tolerance of cross-coupled angular acceleration is 2 rad/s^2 (Graybiel and Wood,1969). Table 8 shows that this value for $\dot{\omega}_{\text{res}}$ limits head movement rates to between 100°/s and 200°/s (fairly rapid movement) for vehicle rotation rates between 11 rpm and 6 rpm.

Substituting a value of 2 rad/s^2 for the tolerable acceleration limit and a torque of 3.03×10^{-10} ft lb, which corresponds to the absence of malaise in Eq. (33), gives a moment of inertia for endolymph about its axis of spin

$$I = 1.51 \times 10^{-10} \text{ ft lb s}^2.$$

Tests in the SRR at Pensacola showed that the semicircular canal tolerance to the above gyro torque can be increased by adapting to incremental increases to an upper rotation rate of about 10 rpm (Graybiel and Wood, 1969). Readaptation to a nonrotating environment, however, is equally demanding. Figure 24 illustrates a set of curves for vehicle radius versus apparent gravity experienced by the crew (in units of g) for a range of vehicle rotation rates from 1 rpm to 10 rpm.

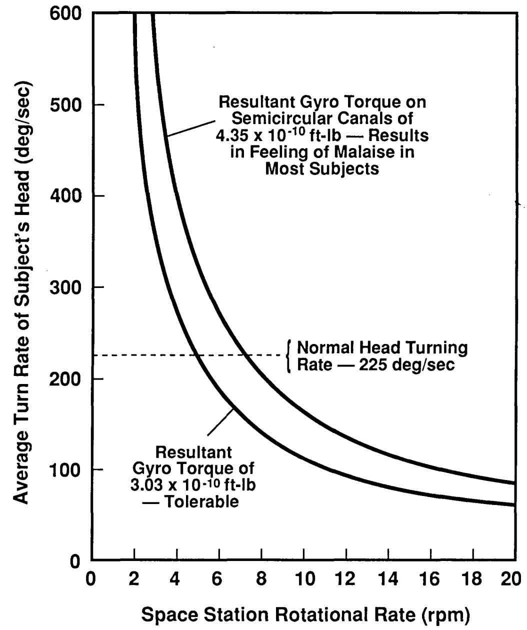

FIG. 23. Gyro torque limits on cross-coupled rotation of semicircular canals in rotating vehicle. [Redrawn from McCormack and Nicogossian (1987).]

The Coriolis acceleration and the centrifugal acceleration obey the relationship

$$A_{\text{Co}} = 2v(A_{\text{ce}}/gr)^{1/2}, \tag{35}$$

where v is the velocity of the body relative to the rotating vehicle (ft/s) and r is the radius (ft). Hence,

$$A_{\text{ce}} = 34 \times 10^{-5} N^2 r, \tag{36}$$

where N is rotation rate in rpm. For a Coriolis force ratio corresponding to 50% of the crewman's apparent weight (force ratio, F_c), an apparent gravity of 0.4 g implies a Coriolis force of 0.2g. Using Eqs. (35) and (36), this leads to an N value of about 10. The broken curves in Fig. 24 correspond to a range of F_c values from 10% to 50%.

Table 8. Tolerable and intolerable head rotation rates.

Rate of head rotation		Vehicle rotation rate (rpm)		Angular acceleration (rad/s^2)	
Rad/s	rpm	Tolerable	Intolerable	Tolerable	Intolerable
1.75	16.7	11.4	15.8	2.08	2.89
3.50	33.4	5.8	8.0	2.13	2.94
5.25	50.1	4.0	5.5	2.20	3.05
7.00	66.8	3.0	4.25	2.34	3.13
8.72	83.5	2.8	3.6	2.52	3.31

If 6 rpm is selected as the maximum tolerable angular rate, then the region bounded by the hatched lines in Fig. 24 corresponds to an artificial gravity range of 0.28*g* to 1*g* and a force ratio between 10% and 20%. This region defines the zone of tolerable and effective human performance for a rotating space vehicle.

Using the various constraints listed above, McCormack and Nicogossian (1987) optimized the basic physical parameters and arrived at the following permissible parameters for a rotating space vehicle: a vehicle radius of 80 ft, a rotation rate of 4.8 rpm, and an artificial *g* value of 0.62*g*. We do not know whether a reduced gravity level of 0.6*g* is sufficient to eliminate the various physiological degradations observed in long-term presence in 0*g*.

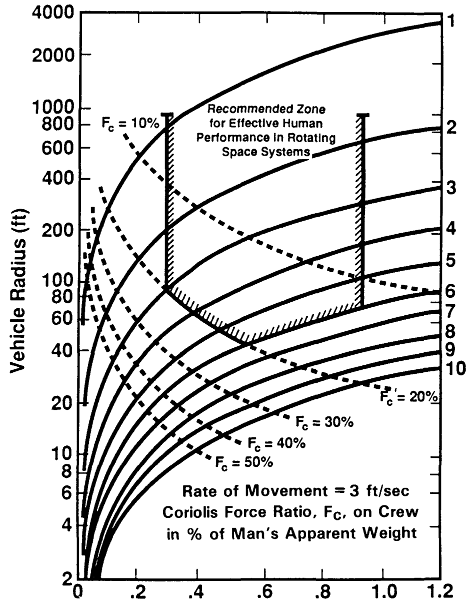

FIG. 24. Design parameters for effective human performance in rotating vehicle. F_C denotes Coriolis force. [Redrawn from McCormack and Nicogossian (1987).]

The boundaries on performance outlined above may be characterized by the criteria listed in Table 6. The various constants $C_1 - C_7$, defined by the inequalities in the table, have been given numerical values and it is useful to reiterate them.

C_1, which is involved in the criterion for tangential motion, is estimated to be 0.5 for man and 0.25 for materials.

C_2, which is involved in the criterion for radial motion, is estimated to be 0.3 for man and 0.25 for materials.

Note that these criteria are expressed as ratios of the force involved to the local artificial gravity level.

C_3, the distance a dropped object falls from its expected position, is assumed to be 1 ft when it is dropped from 3.3 ft.

C_4, change in object weight, is assumed to be 0.5 for objects raised 6.6 ft.

C_5 is related to change in gravity along the human body. This is not specified as it does not affect performance significantly. For the same reason C_6 is not specified.

The constant C_7 is estimated to be 2 rad/s^2. This rate, when associated with a maximum average head rate of rotation (ω_H) of 3 rad/s, gives a vehicle rate of rotation of 0.67 rad/s or about 6 rpm. Adaptation to this rate has been clearly demonstrated in ground-based experiments. Adaptation is still not well understood; however, the guiding criterion generally employed is to set a value as low as practically possible.

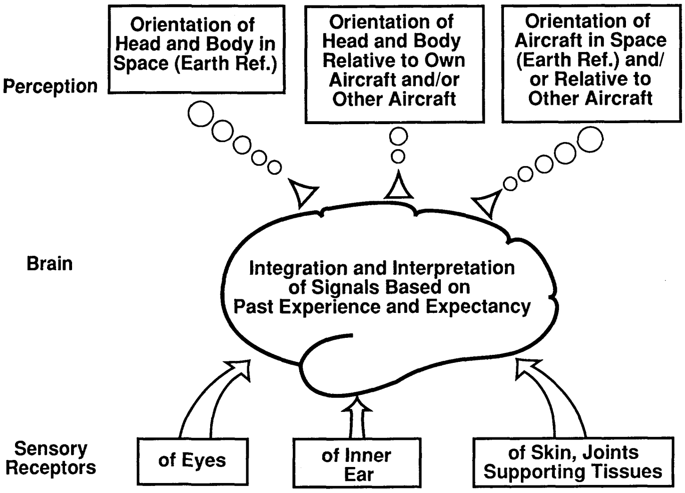
Perception
Orientation of Head and Body in Space (Earth Ref.)
Orientation of Head and Body Relative to Own Aircraft and/or Other Aircraft
Orientation of Aircraft in Space (Earth Ref.) and/ or Relative to Other Aircraft
Brain
Integration and Interpretation of Signals Based on Past Experience and Expectancy
Sensory Receptors
of Eyes
of Inner Ear
of Skin, Joints Supporting Tissues

FIG. 25. Spatial orientation, perception, sensation, and interpretation.

4. SPATIAL DISORIENTATION AND MOTION SICKNESS

4.1 Orientation Mechanisms

Humans perceive their orientation in three-dimensional space through the learned interpretation of the input to the cortex of many sensory receptors (Fig. 25). Sets of these receptors form specialized sense organs like the eye or the vestibular apparatus in the inner ear. Others are distributed through the skin, the joint capsules, and supporting tissues. When on the ground or in a nonmaneuvering spacecraft, whether the person is moving or motionless, accurate perception of the spatial orientation of his or her body to the immediate surroundings is readily achieved using visual inputs alone. These visual signals, plus those from nonvisual receptors, allow the person to sense his or her position, attitude, and motion relative to the stable frame of reference—the Earth's surface or the spacecraft—and to the local vertical.

In the case of aircraft in flight, perception is complicated by disorientation to the immediate surroundings in the aircraft (which is changing its orientation to the Earth's surface and to the Earth vertical). Parts of the aircraft are always visible, whereas the Earth's surface may not be visible at times (night, cloud, etc). When in orbit, the shuttle is the immediate reference for the crew. But during reentry the shuttle becomes an aircraft again and its orientation with respect to the Earth's surface becomes important.

Accurate perception of aircraft and deorbiting spacecraft orientation to Earth's surface, which is essential for control of the craft, depends primarily on correct interpretation of visual cues from outside the aircraft or spacecraft or from cockpit instruments. These cues are, of course, supported secondarily by signals from the nonvisual sense organs.

In this text the function of the inner ear and its vitally important vestibular apparatus, a basic ingredient in both spatial disorientation and SMS, will be dealt with exclusively.

4.2 The Inner Ear

Even when there is no visual input, humans can still perceive orientation with respect to gravity and to their state of motion. The most important sense organs that provide this nonvisual capability of orientation and capacity to maintain balance are those of the inner ear.

Figure 26 illustrates the two anatomic and functional parts of the inner ear. The first is the organ of hearing, the cochlea, which plays only a minor role in orientation and SMS. The second organ is the vestibular system, which is specially adapted to sense linear and angular head accelerations. The information it supplies to the cerebellum and cortex allows humans to orient themselves and maintain equilibrium on the Earth's surface or in space. If the motions to which the person is subjected are atypical and fall outside the normal dynamic range of the sensory system,

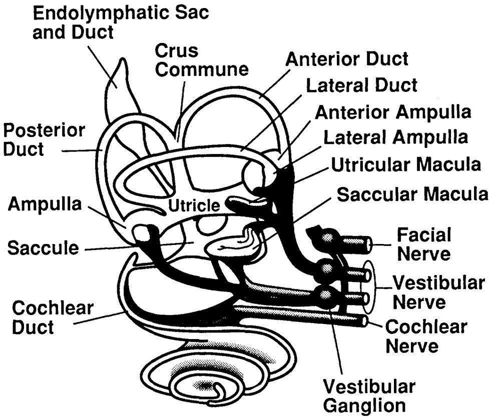
Endolymphatic Sac and Duct
Crus Commune
Anterior Duct
Lateral Duct
Anterior Ampulla
Lateral Ampulla
Utricular Macula
Saccular Macula
Posterior Duct
Ampulla
Utricle
Saccule
Facial Nerve
Vestibular Nerve
Cochlear Duct
Cochlear Nerve
Vestibular Ganglion

FIG. 26. Schematic illustration of human inner ear (right side).

the vestibular system may supply false information to the brain and lead to disorientation. If it supplies unbalanced information (for example, no linear acceleration information) or information that conflicts with visual cues, motion sickness may result. This is the situation in the space environment.

4.3 The Vestibular System

The labyrinth of the human inner ear consists of three main entities: the cochlea (organ of hearing), the otolith organs (sensors of linear acceleration), and the semicircular canals (sensors of angular acceleration). Collectively, these form the vestibular system. Angular accelerations as low as 0.05 deg/s^2 (0.9 mrad/s^2) and linear accelerations as low as 0.01g (0.32 ft/s^2) may be detected. Figure 26 illustrates the labyrinth of the left ear viewed from the medial aspect. These sensory organs are situated in hollowed-out channels in the petrous part of the temporal bone. The three ducts, or semicircular canals, open into the sac of the utricle. Below the utricle, and connected to it, is the saccule. The four otolith organs are in outpouches of the inner walls of the paired utricle and saccule. This system, or membranous labyrinth, is filled with a watery fluid, the endolymph, and is attached securely to the petrous bone. Rigidly coupled to the skull, its experiences the same angular and linear accelerations as the head.

The space between the membranous ducts and the bony canals is filled with a fluid called perilymph. Table 9 shows that the specific gravities of endolymph and perilymph are quite similar.

Table 9. Specific gravity (at 23 °C) and viscosity (at 20 °C) of canal fluids.

Fluid	Specific gravity	Viscosity (centipoise[a])
Endolymph	1.01	1.91
Perilymph	1.01	1.10

[a] Poise = dyne s cm^{-2}.

4.3.1 Semicircular Canals.

The membranous canal has an average cross-sectional radius of about 0.16 mm in humans, while the bony canal radius is approximately 1.2 mm. The cross sections are more elliptical than circular. The horizontal and posterior canals in humans have radii of curvature between 1.6 and 1.9 mm. The anterior canal has a somewhat larger radius of curvature, about 2.2 mm (Curthoys, Blanks, and Markhan, 1977). At one point in each canal, near its junction with the utricle, the lumen swells to form the ampulla, which contains the sensory epithelium crista ampullaris. A crista is a ridge of epithelium protruding into the lumen of the ampulla carrying many multiciliated sensory cells interspersed with structural cells. The cilia on all the cells project into a common gelatinous structure, or cupula. The cupula reaches to the far side of the ampullae (Fig. 27). When the head's axis of rotation is in the plane of a canal, the endolymph fluid lags relative to the walls of the canal and, thus, flows relative to the walls, counter to the head rotation direction.

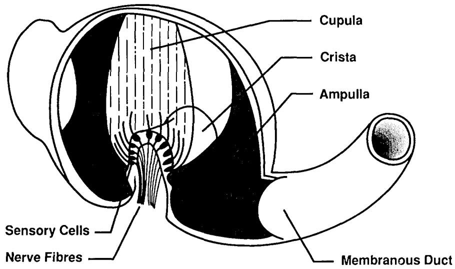
Cupula
Crista
Ampulla
Sensory Cells
Nerve Fibres
Membranous Duct

FIG. 27. Illustration of ampulla of human semicircular canal.

4.3.2 Mathematical Modeling of the Cupula-Endolymph System. Steinhauesen (1933) was the first to report that the cupula normally functions like a watertight, spring-loaded flaplike structure, whose deflection closely follows the flow of the endolymph fluid in the canal. He hypothesized that the cupula-endolymph canal system could be equated to a heavily damped torsion pendulum, in which the inertial and damping terms correspond to the mass and viscosity of the endolymph fluid, respectively, with a restoring force corresponding to the cupula elasticity. Support for this hypothesis comes from the research of Groen, Lowenstein, Vendirk (1952), Egmond, Groen, and Jongkees (1949), and Jones (1965). These and other workers elucidated the various physiological responses to rotational stimuli.

If I is the moment of inertia of the endolymph-cupula system and α denotes the angular acceleration with which a person rotates his or her head in the plane of a particular semicircular canal, then the torque acting on the cupula is $I\alpha$. This torque displaces the endolymph fluid and cupula through the angle θ relative to the skull. To a first approximation, variations in θ with time are governed by the damped torsion pendulum equation

$$\frac{d^2\theta}{dt^2}+\left(\frac{\Pi}{I}\right)\frac{d\theta}{dt}+\left(\frac{\Delta}{I}\right)\theta=\alpha(t), \tag{37}$$

where Π is the system rotational resistance at unit angular velocity (viscous damping) and Δ is the system rotational stiffness due to the cupula. McCormack (1988) calculated a value of 2.3×10^4 cm^2 g for the moment of inertia of the cupula and endolymph. This value compares quite well with the value 2.4×10^4 cm^2 g computed by Oman and Young (1972). The coefficient of viscous resistance may be calculated from the Poiseuille's law,

$$\Pi=16\pi^2\eta R^3, \tag{38}$$

where η is the viscosity of the endolymph and R is the radius of the canal torus. Steer *et al.* (1967) estimated $\Pi=0.043$ dyne cm^{-1} s^{-1}. Unfortunately, it has not been possible to calculate the elasticity of the cupula because its small displacement, at most 10 μm, is difficult to measure. It is known, however, to be much smaller than the viscous resistance, and, thus, to approximate one can write

$$\alpha I=\Pi\frac{d\theta}{dt}. \tag{39}$$

Taking I and Π as constant gives an angular velocity for the cupula proportional to the head acceleration. It is evident that the angular displacement of the cupula is proportional to the angular velocity of the head. The vestibular canals are, in effect, an integrating accelerometer, or angular speedometers. Making the head angular acceleration, $\alpha(t)$, sinusoidal and observing the phase lag between the $\alpha(t)$ and the eye movement (nystagmus) for any given response (for example, oculogyral illusion) creates an informative experimental technique. Further excitation stimuli can be provided by a variety of rotational devices, such as turntables, torsion swings, and rotating chambers.

The three coefficients I, Π, and Δ of Eq. (37) conveniently summarize the information concerning the mechanical properties of the semicircular canal. Furthermore, the system performance in response to a dynamic stimulus can also be interpreted in terms of an undamped characteristic angular frequency, ω_n, and a damping ratio, ρ. The damping ratio parameter ρ denotes the ratio of the actual damping Π to the critical damping Π_c,

$$\rho=\Pi/\Pi_c=\Pi/2\sqrt{I\Delta}, \tag{40}$$

where ω_n is defined as

$$\omega_n=(\Delta/I)^{1/2}. \tag{41}$$

In terms of these parameters Eq. (36) becomes

$$\frac{d^2\theta}{dt^2}+2\rho\omega_n\frac{d\theta}{dt}+\omega_n^2\theta=\alpha(t). \tag{42}$$

The overall system transfer function, relating the eye velocity to angular acceleration input, can be written (in the Laplace domain) as

$$V(s)/\alpha(s)=K_0G_0(s). \tag{43}$$

It can be shown that in terms of the parameters ρ and w_n

$$V(\omega)=\frac{-u\alpha_m}{\omega_n^2}\times\left(\frac{\angle G_0(j\omega)}{[(1-\omega^2/\omega_n^2)^2+(2\rho\omega/\omega_n)^2]^{1/2}}\right), \tag{44}$$

Table 10. Parameter values for the canal-nystagmus transducer system.[a]

Subject	ρ	ω_n	$2\rho/\omega_n$	k_2	$\angle G_0(j\omega)$
1	6.7	1.9	7.1	0.28	arctan $[7.1\ \omega/(1-0.28\ \omega^2)]$
2	5.0	1.0	10.0	1.00	arctan $[10.0\ \omega/(1-\omega^2)]$

[a] Egmond *et al.*, 1949.

where α_m is the maximum angular acceleration, $\angle G_0$ is the phase angle of $G_0(s)$,

$$\angle G_0(j\omega) = -\arctan\left(\frac{2\rho\omega_n}{1-(\omega^2/\omega_n^2)}\right) \tag{45}$$

and $K_0 = k_1 k_2$, where

$$k_1 = \omega_n^{-2} \text{ and } k_2 = -u. \tag{46}$$

4.4 Nonlinear Characteristics of the Endolymph Cupula-Nystagmus System

Table 10 lists typical parameter values measured in human test subjects (Egmond *et al.*, 1949). Figure 28 shows nystagmus phase shift versus frequency for a typical test subject (Niven and Hixson, 1961). Both the experimental points and the theoretical points inferred using the theory outlined above are given. The agreement is quite good. Figure 29 graphs the phase lag, $\phi_0(j\omega)$, versus peak angular acceleration for the same test subject at a stimulus frequency of 0.02 s^{-1}. The decrease in $2\rho/\omega_n$ with increasing peak implies that the ratio of system damping to system stiffness decreases as the cupula peak displacement (proportional to the peak angular acceleration) increases. This observation constitutes the first nonlinear characteristic identified for the system.

Additional nonlinear characteristics involve the use of the nystagmus transition technique (Hixson and Niven, 1962). Figure 30 illustrates two such transitions. When a subject is accelerated to the left, his or her eyes initially move slowly to the right, return quickly to the left, then start a new slow rightward movement. Conversely, acceleration to the right results in slow eye movement to the left with a fast return to the right. The point at which the eye has zero velocity denotes the transition point. This transition point is easily measured and is not initially dependent on the amplitude of the nystagmic eye motions. Similarly, the phase lag for each nystagmus transition is easily measured relative to the corresponding stimulus acceleration transition [$\phi(t)$ in Fig. 30]. These phase lags and frequency characteristics can be easily established. By experimentally measuring the frequency producing a 90° phase lag, the undamped frequency ω_n can be determined (Niven, Hixson, and Correia, 1965). Figure 29 shows phase lag/peak acceleration results ob-

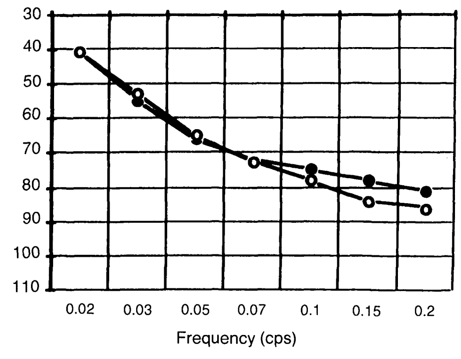

FIG. 28. Nystagmus phase shift, ϕ_0, versus frequency. Symbols: ● experimental; ○, theoretical. [Niven and Hixson (1961).]

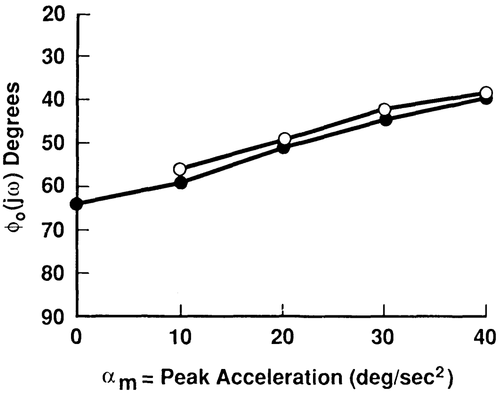

FIG. 29. Experimental nystagmus phase shift versus peak angular acceleration. Solid circles from Niven and Hixson (1961), open circles from nystagmic transition data. Stimulus frequency of 0.02 Hz. [Redrawn from Niven and Hixson (1961).]

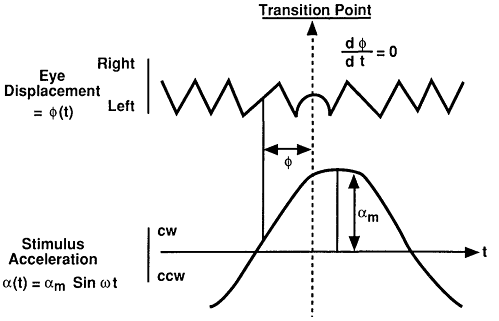

FIG. 30. Nystagmus transition in response to sinusoidal angular acceleration. [Redrawn from Niven and Hixson (1961).]

tained by this technique at a stimulus frequency of 0.02 Hz. These results agree with the results obtained by the previous techniques. Figure 31 shows the dependence of $\angle\phi_0(j\omega)$ on peak acceleration for stimulus frequencies of 0.02, 0.04, 0.06, and 0.2 Hz. For frequencies of 0.08 and 0.2 Hz the phase shift is virtually independent of stimulus magnitude. Thus, in this frequency range the cupula-endolymph system is operating in a linear mode. At stimulus frequencies of 0.04 and 0.02 Hz the phase shift decreases significantly with stimulus magnitude; the slope increases as the frequency decreases. This is a nonlinear characteristic and is likely due to a saturation or overloading effect. It is likely that the mechanical properties of the cupula are nonlinear (Oman, 1981; Oman and Young, 1972).

If the cupula is attached around its periphery, as suggested by Dohlman, then low-frequency stimuli should produce maximum displacement for small volume displacement in the base region. The dynamic range of motion of the midpoint of the cupula has been estimated to range from about 10 mm at the behavioral threshold of 0.1 deg/s^2 to approximately 3 mm for the upper limit of self-induced sinusoidal head motions (30 deg/s^2) (Oman and Young, 1972). As noted already, nonlinearity in the system response starts at a stimulus frequency of about 0.04 Hz. This is consistent with the observed time lag constant of the system response, Π/Δ, of approximately 20 s. At stimulus frequencies below 0.04 Hz, and for quite moderate peak acceleration stimuli (less than 50 deg/s^2), the cupula deflection can easily reach saturation, resulting in the observed nonlinearity.

The cupula is attached to hair cells (Fig. 27). During a rapid turn of the head, for semicircular canals that lie in the plane of the angular movement, the rings of endolymph tend to remain in their original position due to their inertia. This results in a shearing movement near the attachment of the hair cells to the cupula. The hair cells are thus stimulated and send nerve impulses to the vestibular nuclei located in the brain stem. The dynamics of the hydromechanical system are such that the cupula deflection and the associated signal from the sensory cells closely match the angular velocity of the head. The semicircular canal acts as a rapid rate-of-turn transducer.

Difficulties arise, however, when the speed of rotation remains steady for a time as occurs in a prolonged roll in flight. As soon as the rate of roll becomes constant, the cupulae of the stimulated canals begin to return to their neutral position, as there is no longer any angular acceleration. Typically, in a roll at 100 deg/s (≈2 rad/s) the sensation of roll vanishes in 10–15 s. There is also an erroneous signal of rolling in the opposite direction on recovery from a constant rate of roll.

4.5 The Otolith Organs

The four otolith organs appear as thickened portions on the inner walls of the paired utricle and macule. The latter are two endolymph-filled sacs, which lie below the canal and contain platelike congregations of sensory cells. These cells have many hairlike projections that, when deflected, alter the sensory cell activity. The cells in both maculae are grouped in a saucerlike area and are covered with a gelatinous layer. The outer sur-

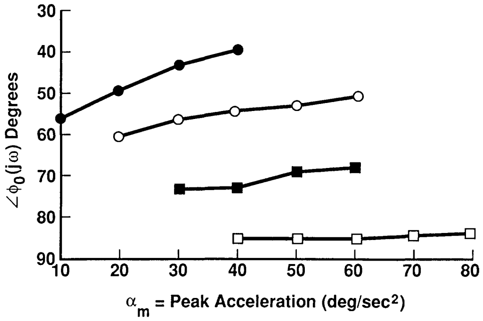

FIG. 31. Nystagmus transition data for four stimulation frequencies. See text for definition of terms. [Redrawn from Niven *et al.* (1965).]

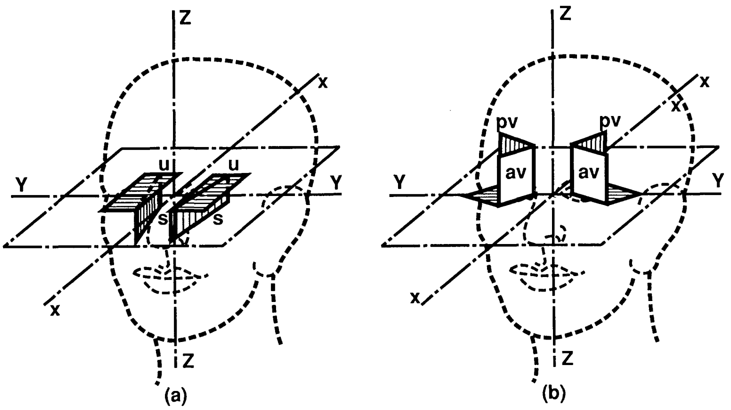

FIG. 32. Approximate planes of (a) otolithic maculae and (b) semicircular canals.

face is covered with small calcium carbonate crystals that look like a white stony plaque, whence the name otolith. The specific gravity of the otoliths is about 2.74. Displacement of the heavy otoliths produces shear forces on the hair cells resulting in output signals that are sent to the brain. Such displacements are induced by changes in the extent or direction of the linear acceleration of the head that are caused by changes in the magnitude or direction of the linear velocity of the head or by a tilt of the head with respect to gravity. Only a nonaccelerated person in the weightless condition of space is free of linear accelerative forces.

4.5.1 Dynamics of the Otolith Organs. Trinker (1962) recorded changes in potential in the macula of the guinea pig and demonstrated that only shearing forces produced by *tangential* displacement of the otolith membrane relative to the macular surface stimulate the afferent nerves. Pressure or tension *perpendicular* to the macular epithelium produces no effect. Shearing forces produce a response proportional to the sine of the angle of tilt of the macular surface from the horizontal.

The excess mass of the otolith membrane, β, is defined as its mass, m, minus the mass of an equivalent volume of endolymph. If the density of endolymph is taken as one ($\rho_a = 1$), then

$$\beta = m[1-(\rho_a/\rho)], \tag{47}$$

where ρ is the density of the otolith membrane.

The shearing force acting on the otolith organ equals the product of β and the linear acceleration acting in the plane of the macula (a). This is opposed by the elastic, or position dependent, resistance k; the viscous, or velocity-dependent, resistance μ; and the mass, or acceleration-dependent, resistance m. Hence, if x denotes the linear displacement of the otoliths,

$$\beta a = kx + \mu\frac{dx}{dr} + m\frac{d^2x}{dt^2}, \tag{48}$$

where m is the effective mass of the otolith membrane. DeVries (1950) measured these constants for fish and found that the otolith membrane of the saccule was displaced by about 33 μm by a 1g tangential force. If similar values are assumed for humans, and given a human threshold for saccular stimulation of 0.005g (Melvill Jones and Young, 1978), then the threshold displacement of the saccule for humans is about 0.15 μm. This displacement is of molecular dimensions and suggests that the critical event is a deformation of organic molecules.

Figure 32 shows the approximate planes of the otolithic maculae and the semicircular canals. There are two otolith organs on each side of the head; those of the utricle lie in an approximately horizontal plane when the head is in a normal upright position and those of the saccule lie in the vertical plane. By observing the displacement of the otolith membrane during oscillatory stimulation, de Vries (1950) deduced that the structure is critically damped and is fully displaced within 5 ms after the application of a constant tangential force. Furthermore, the membrane apparatus rapidly returns to its normal position upon removal of the stimulus.

Table 11. Physiological thresholds for the cupula-canal system.

Threshold	Angular acceleration (deg/s^2)
Oculogyric illusion[a]	0.1
Nystagmus illusion[b]	0.8
Self-induced head motion	30.0
Motion sickness (Coriolis)[c]	115.0

[a] Meiry (1965).
[b] Buys and Rijhaut (1939).
[c] Graybiel and Wood (1969).

There is, however, a basic ambiguity in signals from the otolith organs. Specifically, the effect of a horizontal linear acceleration, *a*, acting in the plane of the utricles is indistinguishable from the effect of tilting the head through an angle whose sine is *a*. Thus, if an aircraft gains speed, the otolith organs are stimulated to the same extent as if there was an upwards pitch of the aircraft. When flying in a cloud, the aircraft pilot may try to correct for this apparent upward pitch and fly the plane into the ground. However, input from the semicircular canals may help to correctly interpret the otolith inputs.

4.5.2 False Perception of Angular Motion: Somatogyral and Oculogyral Illusions. During a coordinated turn, a sustained roll, or a spin, the semicircular canal receptors only give correct information during the first few seconds of the maneuver. As the aircraft comes out of the spin, there is angular acceleration in the opposite direction. The semicircular canals are stimulated again and give a sensation of turning in the opposite direction. This effect is called the somatogyral illusion. The inappropriate signals from the canal affect visual perception and add to this disorientation. These phenomena were called oculogyral illusions by Graybiel (1952); they take the form of apparent motion and errors in the localization of targets. Table 11 lists some physiological thresholds for the cupula-canal system. The threshold for the oculogyric illusion is the lowest.

The threshold for Coriolis (cross-coupled acceleration) induced motion sickness most likely results in saturation of the cupular movement. As shall be seen in the next section, the sensory conflict theory offers the most likely explanation for this form of motion sickness.

4.6 Space Motion Sickness

Approximately 50% of those who enter the microgravity environment of low-Earth orbit experience the symptoms of SMS. Symptoms vary from stomach awareness to vomiting. In all cases studied these symptoms vanish almost completely within 48 h. The occurrence of SMS does not appear to correlate with previous sickness, such as parabolic flight sickness, air sickness in general, or sea sickness. There are some indications that adaptation occurs because the intensity of symptoms appears to lessen with successive flights. Common anti-motion-sickness drugs, especially scopolamine, lessen the intensity of symptoms, and many susceptible crewmembers use these drugs liberally.

Although considerable research has been performed in the last 20 years, no fully successful theory has been developed and we cannot predict who will be affected by SMS and who will not. For these reasons many researchers have attempted to model the relevant neurophysiological systems.

A system model of SMS is needed to categorize and order the observed facts into a rational structure, encode the measured data and dynamic processes efficiently, and clarify several key experiments needed to predict and/or prevent motion sickness on earth and in space.

One theory of the etiology of SMS is the sensory conflict theory, which suggests that the sensory orientation cues from the eyes, the vestibular system, and the proprioceptors are interpreted and modified by a "central comparator" in the brain with respect to a "neural store" of expected correlations in normal experience. A significant unexpected correlation would then result in SMS. The model must include the vestibular system; the visual system; a neurochemical system upon which drugs may act; a "learning center," including habituation and training effects; and a nauseogenic system that induces its symptoms. Figure 33 illustrates the comprehensive model called the "cue conflict" model.

The vestibular and visual systems in Fig. 33 have been replaced with linearized mathematical models. However, the limitation of this model should be apparent in that the cu-

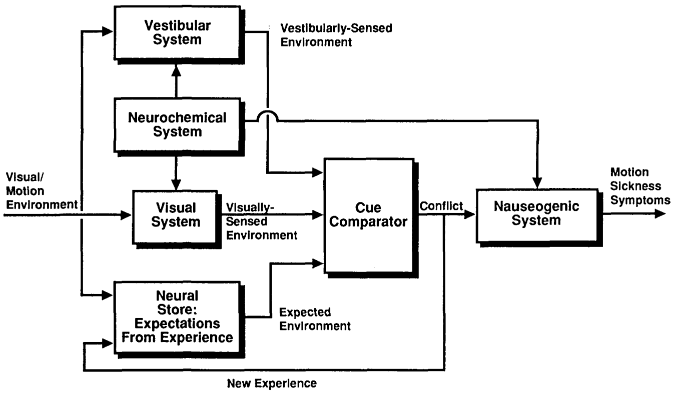
Vestibular System
Vestibularly-Sensed Environment
Neurochemical System
Visual/ Motion Environment
Visual System
Visually-Sensed Environment
Cue Comparator
Conflict
Nauseogenic System
Motion Sickness Symptoms
Neural Store: Expectations From Experience
Expected Environment
New Experience

FIG. 33. Comprehensive model based on "cue conflict" hypothesis.

pula-canal system has nonlinear characteristics. Moreover, the vestibulo-ocular reflex (wherein eye movements may be evoked by purely vestibular stimuli) and the eye stabilization system have not yet been incorporated. At finer levels of detail the following key facts must be included.

1. Vestibular signals dominate in the short-term, high-frequency determination of acceleration, while visual signals process long-term, low-frequency velocity information (Clark and Graybiel, 1963).
2. When vestibular cues confirm visual cues, the vestibular cues are weighted slightly; when a conflict occurs, vestibular sensations dominate (Young and Dichgaus, 1973).
3. Cross-coupled or Coriolis accelerations provoke motion sickness due to a canal-otolith conflict (Miller and Graybiel, 1970).
4. Rapid head movements in the weightless environment cause SMS, possibly due to the absence of otolith indication of a 1g environment (Miller and Graybiel, 1970).

These facts suggest a comparator structure composed of elements that compare vestibularly perceived velocity with visually perceived velocity, and compare canal-deduced orientation with otolith-deduced orientation. Conflict can arise from two signals or when one signal is absent. The three levels of comparison possible are direct canal-otolith comparison, actual and expected canal signal comparison, and actual and expected otolith signal comparison.

In recent years the sensory conflict theory has been replaced by the neural mismatch hypothesis (Oman, 1982). An essential feature of the neural mismatch hypothesis is that the presence of the sustained mismatch signal rearranges the internal model and evokes the sequence of neural and hormonal responses that constitute the motion sickness syndrome. Figure 34 shows the components and processes that the hypothesis embraces. The output from the comparator is the mismatch signal that updates the internal model and activates the neural centers and pathways that produce the symptoms of motion sickness.

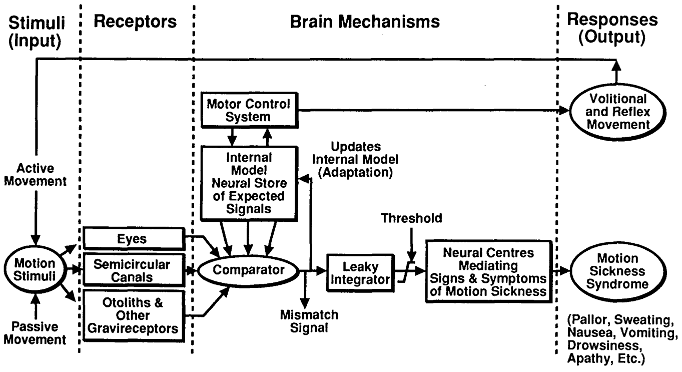
Stimuli (Input)
Receptors
Brain Mechanisms
Responses (Output)
Motor Control System
Volitional and Reflex Movement
Active Movement
Internal Model Neural Store of Expected Signals
Updates Internal Model (Adaptation)
Threshold
Eyes
Motion Stimuli
Semicircular Canals
Comparator
Leaky Integrator
Neural Centres Mediating Signs & Symptoms of Motion Sickness
Motion Sickness Syndrome
Otoliths & Other Gravireceptors
Passive Movement
Mismatch Signal
(Pallor, Sweating, Nausea, Vomiting, Drowsiness, Apathy, Etc.)

FIG. 34. Model of motion control, motion detection, and motion sickness; "neural mismatch" hypothesis.

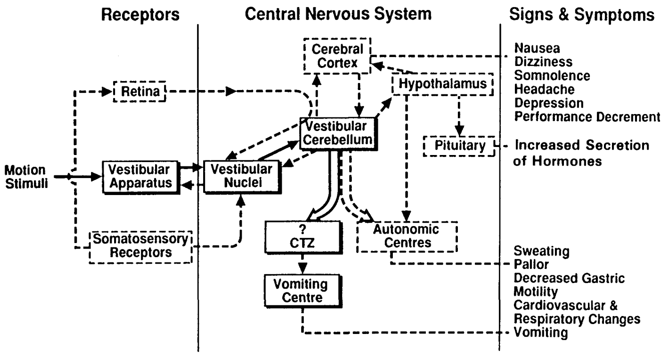
Receptors
Central Nervous System
Signs & Symptoms
Motion Stimuli
Retina
Vestibular Apparatus
Somatosensory Receptors
Cerebral Cortex
Hypothalamus
Vestibular Cerebellum
Pituitary
Vestibular Nuclei
? CTZ
Autonomic Centres
Vomiting Centre
Nausea
Dizziness
Somnolence
Headache
Depression
Performance Decrement
Increased Secretion of Hormones
Sweating
Pallor
Decreased Gastric Motility
Cardiovascular & Respiratory Changes
Vomiting

FIG. 35. Neural structures, signs, and symptoms involved in motion sickness. CTZ denotes the central trigger zone.

Unfortunately, the neural mismatch hypothesis fails to explain why motion sickness takes the form that it does, or why it occurs at all. However, the hypothesis provides a unifying concept and explains why motion sickness occurs in some motion environments and not others. It also explains the basic features of adaptation in the context of sensory rearrangement.

Animal research has yielded considerable information on the relevant neural mechanisms involved in SMS. Figure 35 summarizes the principal elements and pathways that have been identified as implicated in the genesis of motion sickness. There is strong evidence that the vestibular apparatus, and hence the vestibular nuclei and their projections to the cerebellum, are essential for motion sickness to occur. Convergence of vestibular, visual, and kinesthetic and cerebellar afferents occurs at the level of the vestibular nuclei, and so the comparison of actual and expected motion cues could occur in these brain stem nuclei. It is most likely that the cerebellum is the site of the comparator and the internal model. How the cerebellum initiates the sequence of sensory and autonomic responses that constitute the motion sickness syndrome, however, remains to be resolved.

5. THE RADIATION ENVIRONMENT AND ASSOCIATED HAZARDS

5.1 Earth's Radiation Environment

In addition to the Earth's two fundamental fields, magnetic (Sugiura and Heppner, 1968) and gravitational, the Earth is surrounded by a particle radiation field. From data acquired by the 1958 Explorer IV and Pioneer III missions, van Allen and Frank (1959a,b) inferred the first complete map of the trapped radiation (Fig. 36). This model led to the concept of inner and outer radiation zones or what are commonly called the "belts of radiation." The inner zone extends up to an altitude $R < 2R_e$, where R_e is the Earth's radius, and is comprised mostly of high-energy protons with energies up to tens of MeV. The outer zone, extending over distances R between $2R_e$ and $10R_e$, consists mainly of electrons with energies varying from a few tens of keV to several MeV. The energies and particle fluxes comprising these zones depend on the altitude distance R. Particles in the outer zone, especially the electrons, show large temporal variations, changing their fluxes by more than an order of magnitude in less than a day. The inner-zone properties, on the other hand, are quite stable in time. The outer zone edge is very sharp. For the special case of electrons having $E > 40$ keV the flux decreases to cosmic ray background levels, a factor of

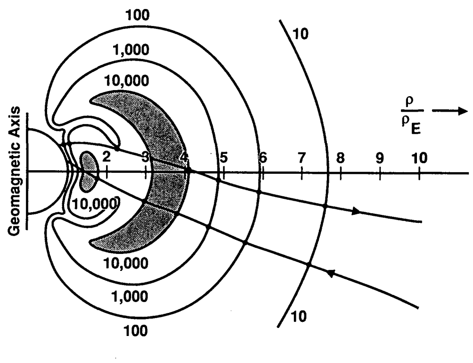
100
1,000
10,000
10
Geomagnetic Axis
$\frac{\rho}{\rho_E}$
2
3
4
5
6
7
8
9
10
10,000
10,000
1,000
100
10

FIG. 36. Van Allen's initial map of the Earth's radiation belt illustrating the inner and outer zones of high count rate. Contours are labeled by the Geiger counter rate of 1-cm^2 area behind 1g/cm^2 of lead. [Redrawn from Frank *et al.* (1964).]

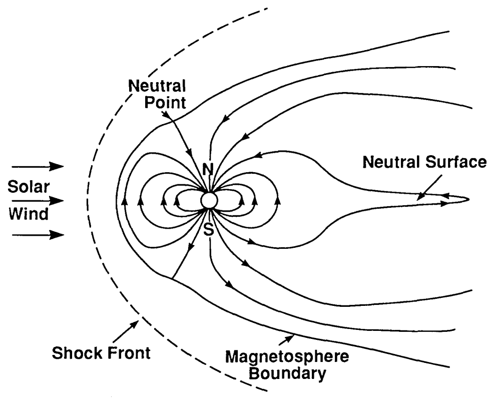

FIG. 37. Schematic of the deformation of the outer magnetosphere as derived from experiment and theory.

over 1000, in a distance as small as 100 km. This sharp spatial variation is a consequence of the solar wind. The solar wind consists of particles streaming continuously outward from the sun at speeds averaging 300–500 km/s; its composition is primarily protons and an approximately equal number of low-energy electrons, although a small percentage of higher element nuclei are also present. Its density at the Earth is about 10 ion/cm^3. The solar wind exerts a pressure, P, on the geomagnetic field, pushing the Earth's field ahead of it until it is balanced by the compression of the Earth's magnetic field. The boundary location, called the magnetopause (Hess and Mead, 1968), between the solar wind and geomagnetic field is determined by the hydromagnetic pressure balance equation

$$P^2+(B^2/8\pi)=0, \qquad (49)$$

where B (in gauss) denotes the magnetic flux density. Figure 37 illustrates the deformation of the outer magnetosphere caused by the solar wind. A general solution of the pressure balance equation for the magnetosphere boundary requires a solution of a nonlinear second-order partial differential equation (Beard, 1964). In the special case for a point on the equatorial plane, Eq. (49) gives the geocentric distance to the boundary of approximately $10R_e$ if typical values of 10 particles/cm^3 at 400 km/s are assumed for the solar wind.

5.1.1 Inner Zone.

Figure 38 shows the energy spectrum for the inner-zone protons as determined by Freden and White (1962). The solid line is the theoretical curve that was derived on the basis of cosmic-ray albedo neutron injection and ionization and nuclear collision losses. Several sources have been suggested for the origin of the trapped radiation belt particles, including primary cosmic rays, solar winds, and neutrons produced by cosmic ray protons colliding with primarily oxygen and nitrogen nuclei in the Earth's upper atmosphere. In the last case several neutrons are generated in high-energy stars (approximately seven per primary). A few of the neutrons (approximately 2.5 per star) escape from the top of the atmosphere and eventually decay according to the reaction

$$n\rightarrow p+e^-+\overline{\nu}, \qquad (50)$$

where $\overline{\nu}$ denotes an anti-neutrino. Calculations based on this process give a proton spectrum source near the Earth at the equator

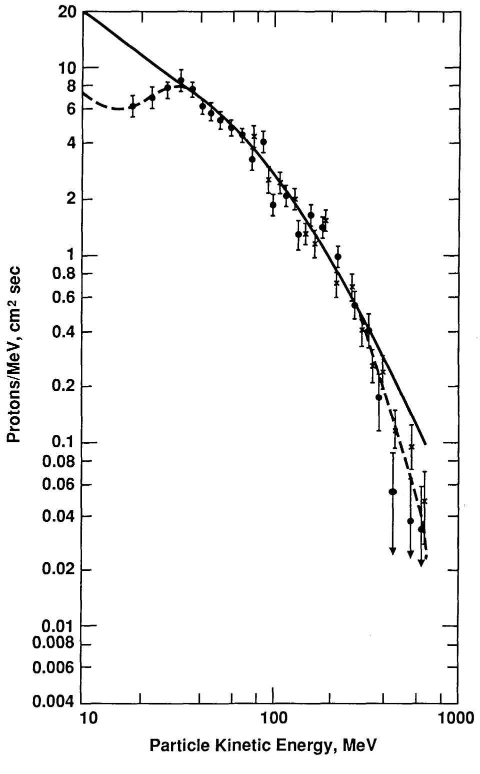

FIG. 38. Energy distribution of high-energy protons in the middle of the inner zone of the radiation belt. The solid circles are data from Freden and White (1962). The crosses are the data from the April 1959 flight normalized to data in region of 100 MeV. The dashed curve is a fit to the data. The solid curve is the theoretical spectrum expected from the cosmic-ray neutron albedo decay theory. [Redrawn from Hess, 1968.]

$$S_p(E)=0.8E^{-2.0}/\gamma v T_n, \tag{51}$$

where v is the neutron velocity; T_n is the neutron mean life; and γ is the time dilation factor, $\gamma = [1 - (v^2/c^2)]^{1/2}$. A quantitative description of the flux-energy spectrum can be obtained by solving the continuity equation

$$\frac{dN(E)}{dt}=S_p(E)-L(E)+\frac{\partial J(E)}{\partial E}, \tag{52}$$

where $N(E)$ is the equilibrium proton density, $S_p(E)$ denotes the proton source, $L(E)$ represents the loss term due to inelastic nuclear collisions, and $J(E)$ is the current

$$J(E)=N(E)\frac{dE}{dt}. \tag{53}$$

At equilibrium and assuming no losses $[L(E)=0]$, Eq. (52) gives a proton flux

$$J(E)=110E^{-0.72} \text{ protons/cm}^2 \text{ s MeV} \tag{54}$$

in the energy range $10<E<80$ MeV, assuming an air density of 2.8×10^{-18} g/cm^3 (corresponding to an atomic density of 10^5 atoms/cm^3); an injection coefficient, $\eta=0.30$, defined as the fraction of protons produced by neutron decay that are actually trapped; and using the source term given by Eq. (51). The low-energy region of the solid curve in Fig. 38 agrees with that given by Eq. (54). The general solution to Eq. (52) when loss terms are explicitly included was given by Freden and White (1960) and is denoted by the full solid line in Fig. 38. The agreement between theory and experiments offers strong support that secondary neutrons are an important source of the inner radiation zone. In addition to galactic cosmic ray (GCR) protons, protons in the solar winds with energies up to a few hundred MeV reach Earth in the polar regions. Neutrons are made when these protons interact with the polar atmosphere. The decay of these solar-proton–produced neutrons constitutes an additional source of energetic protons for the inner radiation zone. Protons produced from solar-cap neutrons initially were suggested to be responsible for the low-energy component at $L>1.6$ reported by Naugle and Kniffer (1961), although recent measurements showed that the large fluxes for $E<50$ MeV are not completely accounted for by either of the above processes. Nevertheless, neutron-decay electrons are a major source of the trapped inner-zone electrons. Here L is McIlwain's magnetic shell parameter (Scraf, 1966; McIlwain, 1961) and is a distance (expressed in Earth radii) defined in terms of integral invariants; in a dipole field L is the distance to the Equator for a particular magnetic field line. L is nearly constant along a field line and, therefore, replaces the commonly used equatorial radius with the real values of the Earth's magnetic field.

5.1.2 Outer Zone. The outer zone consists mainly of electrons and low-energy protons. Outer zone protons having energies 0.1 MeV$<E<$5 MeV were initially investigated on the Explorer XII, XIV, and XV missions. At these altitudes, the energy spectra largely obey the relationship

$$N(E)=k \exp(-E/E_0), \tag{55}$$

where k is a constant and the characteristic energy E_0 is proportional to L^{-3}. That $E_0 \propto L^{-3}$ has been interpreted in terms of the radial drift of the outer zone protons. Time variations in the proton fluxes are seen for $L>6$, but for lower L values the fluxes are quite stable.

In contrast to protons, electrons in the outer zone show large time variations. For $E>1.5$ MeV the electron flux can increase or decrease by a factor of 10 or more in hours. Typical intensities of electrons in the central part of the outer radiation zone ($L=4.0$ near the geomagnetic equator) are $J(E>40$ keV$)=3\times10^7$ (cm^2 s)$^{-1}$; $J(E>230$ keV$)=3\times10^6$ (cm^2 s)$^{-1}$; and $J(E>1.6$ MeV$)=3\times10^5$ (cm^2 s)$^{-1}$ (Frank, Van Allen, and Hills, 1964). For $E>0.5$ MeV the energy spectra are exponential with 0.2 MeV$<E_0<$0.6 MeV; electrons of a particular energy E peak at lower L as E increases.

5.1.3 Cosmic Rays. Cosmic rays are energetic nuclear particles found in interplanetary space; they arrive at the Earth's upper atmosphere from all directions. Cosmic rays can be considered as either solar or galactic, depending on their site of origin. There may be, in addition, a small high-energy (greater than 10^{15} eV) component that originates in extragalactic space. The charged particles are principally protons; heavier nuclei make up about 10% of the total and electron and positions contribute about 3% of the total. A small contribution from antiprotons has also been

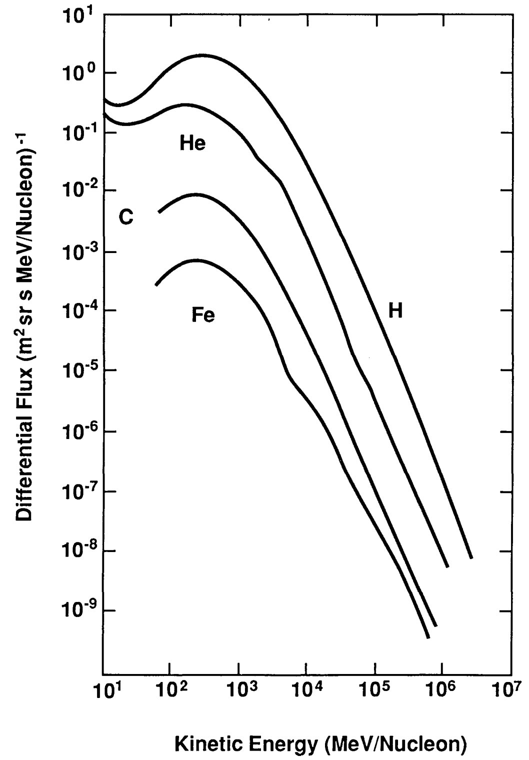

FIG. 39. Differential energy spectra of cosmic rays for the elements hydrogen (H), helium (He), carbon (C), and iron (Fe) measured near earth. [Redrawn from Gow and Taylor (1968).]

detected. The γ radiation accounts for approximately 10^{-6} of the total, and neutrinos predominate by many orders of magnitude, as they originate largely from our Sun. Although cosmic rays were detected by Hess in 1912, their origin and the processes by which they achieve their high energies are still subjects of debate. Energies as high as 3×10^{19} eV have been recorded.

5.1.4 Galactic Cosmic Rays. Galactic cosmic rays (GCRs) are high-energy particles that do not originate from our Sun. Their flux is approximately 2 particles $\text{cm}^{-2}\ \text{s}^{-1}$. Contrary to previously held views, recent measurements demonstrated that GCRs do not impinge on the earth isotropically but show slight anisotropies that increase with energy for $E>3\times10^{11}$ eV. Figure 39 shows the spectrum of the more pronounced components. Simpson (1983) reviews the elemental and isotopic composition of the GCRs extensively.

For total energy E greater than a few BeV, the flux of protons can be represented by the formula

$$N(>E)=CE^{-\gamma} \tag{56}$$

where the exponent γ varies from 1.4 for E between 5×10^{9} eV and 5×10^{11} eV to 1.5 for E between 10^{13} eV and 10^{15} eV, and is approximately 2 for $E>10^{15}$ eV. An important feature of the proton spectrum is its variability within the 11-year solar cycle. When the solar wind fields are minimal, modulation in the GCR spectrum is minimal. Solar modulation is more effective for low-energy nuclei than high-energy particles. Figure 40 shows this correlation between solar activity and GCRs in plots of the differential rigidity proton spectrum during solar minimum and solar maximum activity. The rigidity R of a particle of momentum p and charge Ze is defined as

$$R=pc/Ze, \tag{57}$$

where c is the speed of light. A larger rigidity corresponds to a particle with higher energy. Another important characteristic of GCRs is their constancy in time. An analysis of deep-sea core samples indicates a variation in the GCR flux of less than 20% over the last 30 000 years.

Several proposals have been given for the origin of GCRs (Morrison, 1961). For energies of 10–100 BeV, GCR particles are presumed to be produced in the atmosphere of

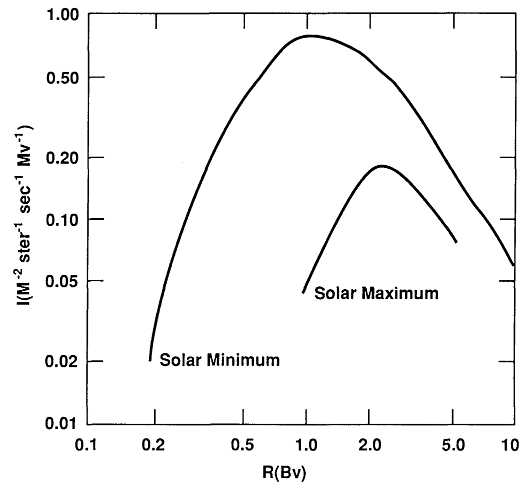

FIG. 40. Differential rigidity spectrum (*I*) for protons during solar minimum and solar maximum. [Redrawn from Heinrich *et al.* (1989).]

Sun-like (or possibly special) stars. For energies in the range of 10^2–10^9 BeV, the particles originating in Sun-like stars acquired these high energies by accelerating through the action of the weak random interstellar magnetic fields. For very energetic particles, several mechanisms, such as remnants of supernova, pulsars, and black holes have been suggested.

5.1.5 Solar Cosmic Rays. Solar cosmic rays (SCRs) are high-energy particles that originate at the Sun and are present near the Earth's upper atmosphere 10%–20% of the time. SCRs are distinct from the continuous stream of particles that comprise the solar wind. SCRs occur only in conjunction with solar flares (Hirman, 1988), which are disturbed regions on the Sun's surface. Solar flares are classified as either ordinary events (OR) or anomalously large events (AL). AL events are quite rare; only three occurred in the 19th solar cycle, one in the 20th solar cycle, and none in the 21st cycle. Flares release enormous amounts of energy in a short period of time (hours to days). For example, in the flare of August 1972, the 10-MeV proton fluxes at geosynchronous orbits exceeded 10^5 p $\mathrm{cm}^{-2}\ \mathrm{s}^{-1}\ \mathrm{sr}^{-1}$ and released over 10^{32} ergs of energy; this is more energy than mankind has used in all of history. As with GCRs, protons are the dominant species in SCRs, although higher element nuclei are also present. Not all solar flare activity on the Sun results in particle emission, and even these events may not emit particles directed toward the Earth. Solar flares that satisfy both criteria are termed solar particle events (SPEs). SPEs occur at a rate of about 100 per 11-yr solar cycle, predominately during solar maximum, although they can appear at any time in the solar cycle. For energies >10 MeV, they have proton fluxes of 10 p $\mathrm{cm}^{-2}\ \mathrm{s}^{-1}\ \mathrm{sr}^{-1}$ or greater as measured at geosynchronous orbits. Because a single AL solar particle event can be the dominant part of radiation dose received during a space mission, considerable research has been concerned with solar proton event forecasting (Hirman, 1988). The National Oceanic and Atmospheric Administration is responsible for space weather forecasting.

In summary, the space radiation environment is a function of location and time. Radiation in low-Earth orbits (LEO) is primarily trapped protons, whereas in geosynchronous earth orbits (GEO) trapped electrons (and bremsstrahlung radiation produced by electrons interacting with the spacecraft) dominate. For journeys beyond the Earth's radiation belts, the effects GCRs and large solar particle flares are the main sources of ionizing radiation. Table 12 lists the sources of radiation, their locations, and the secondary radiation produced by interaction of the primary component with a spacecraft.

5.2 Shielding

5.2.1 Geomagnetic Shielding. Incident particles originating from outside the geomagnetic field are deflected by the Earth's magnetic field. The concept of the rigidity, R, of a particle in a magnetic field can be used to study the influence of the Earth's magnetic field on the energy spectra of cosmic rays. The rigidity was defined by Eq. (57) as

$$R=B_{\perp}C=Pc/Ze, \tag{58}$$

where $B_{\perp}$ is the field component perpendicular to the particle's instantaneous velocity. For inhomogeneous fields, each point on the particle curve trajectory defines a radius of curvature, r. If one approximates the Earth's magnetic field as a dipole field, then it can be shown that particles must have a rigidity greater than R_s to reach the Earth's surface (Störmer, 1930), where

$$R_s=\frac{M\cos^4\lambda}{r^2[1+(1-\sin\theta\cos\phi\cos^3\lambda)^{1/2}]^2}. \tag{59}$$

Here M denotes the Earth's dipole moment and has a value of 8.1×10^{25} G cm^3; r is the radial distance from the center of the dipole; and λ, θ, and ϕ are the geomagnetic latitude, zenith, and azimuth angles, respectively. Hence, each location on the Earth defines a cutoff energy,

$$E=[(m_pc^2)^2+R_s^2\Lambda^2]^{1/2}-m_pc^2 \tag{60}$$

in the cosmic ray particle energy below which particles cannot reach the Earth's surface. Here $\Lambda=ez/A$ is the charge-to-mass number ratio of the particle and m_pc^2 is the rest energy of the proton. R_s is called the Störmer cutoff. In the special case of vertical incidence ($\theta=0$), $R_s = M\cos^4\lambda/4r^2$. At the geomagnetic equator ($\lambda=0$), R_s is maximal; $R_s = 0$ at the poles. More precise values of R_s can

Table 12. Sources of radiation, location, and types of secondary radiation.

Type	Description	Location	Secondary radiation
Trapped electrons	Low LET[a] Large temporal variations Not very penetrating Low dose rate	(3–12) R_e[b] GEO[c]	Bremsstrahlung Penetrating Low LET
Trapped protons	Ranges from low to high LET Penetrating Low dose rate	LEO[d]	Neutrons High LET
Solar particle events	Mostly protons Lesser amounts of heavier ions Occurs sporadically Occasional events of extremely high intensity	Outside Earth's magnetic field— polar orbit GEO Moon interplanetary space	Neutrons Nuclear fragments High LET Penetrating
Galactic cosmic rays	Protons, He, heavier ions (especially Fe) Very penetrating, high LET Low dose rate, isotropic	Outside Earth's magnetic field— polar orbit GEO Moon interplanetary space	Neutrons Nuclear fragments High LET Penetrating

---

[a] LET: linear energy transfer—a measure of the amount of energy deposited as radiation interacts with matter; for a given radiation dose, biological effects are strongly dependent on the LET of the radiation.

[b] R_e: Earth radius, approximately 6378 km.

[c] GEO: geosynchronous Earth orbit

[d] LEO: low-Earth orbit.

be calculated by including the effects of higher order magnetic multipoles. Including higher-order multipoles on R has been considered by Shea, Smart, and McCracken (1965) and Shea and Smart (1967), to which the reader is referred.

The radiation exposure received by an astronaut is altered by the effects of geomagnetic shielding and spacecraft attenuation. The effects of geomagnetic shielding on the incident flux $I_0(E)$, where E denotes energy, can be related to the energy-dependent flux at the desired location $I(E)$ by the equation (Heinrich and Spill, 1979)

$$I(E) = I_0(E)\,W(R), \tag{61}$$

where $W(R)$ denotes the probability that particles with rigidity R can penetrate the field. In general, $W(R)$ varies with location, and is, therefore, a function of the longitudinal position of the intersection of the orbit with the equator, the angle between the intersection point of the orbit with the equatorial plane and the perigee of the spacecraft, the space-

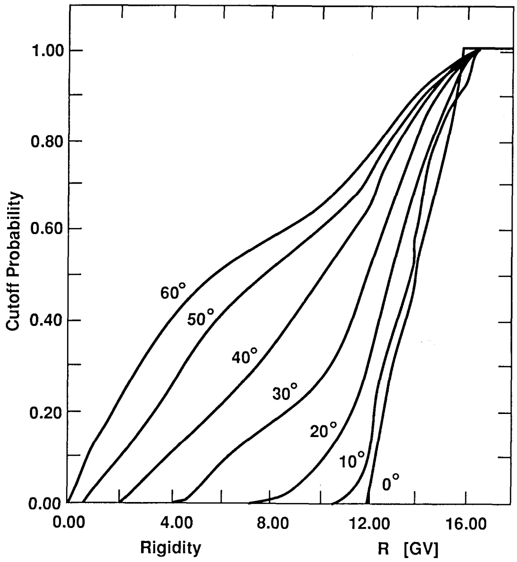

FIG. 41. Cutoff rigidity functions for a circular orbit at an altitude of 223 km for several inclinations (i). [Redrawn from Heinrich and Spill (1979).]

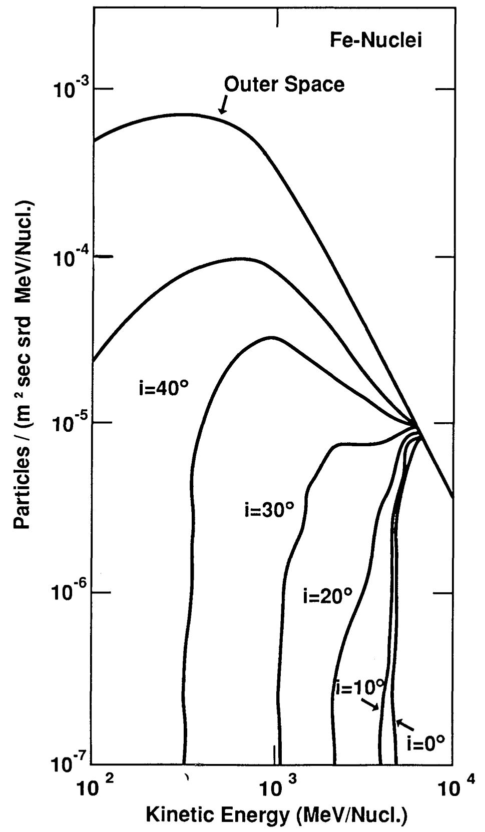

FIG. 42. Iron (Fe) nuclei energy spectra for circular orbits at 223-km altitude for different inclinations (*i*). [Redrawn from Heinrich and Spill (1979).]

craft altitude, its orbit inclination, and the solar cycle. Figure 41 illustrates the variation expected in $W(R)$ for a circular orbit at a fixed altitude (223 km) for several inclinations i (Heinrich and Spill, 1979). Figure 42 illustrates the effects of geomagnetic shielding on the energy spectra for iron nuclei. As expected, the cutoff rigidity by having a steeper slope for lower inclinations introduces a sharper and higher-energy cutoff in the energy spectrum. The uppermost curve in Fig. 42 gives the spectrum for iron nuclei outside the magnetosphere.

5.2.2 Shielding by Materials. Apart from nuclear collisions of an incident particle with material nuclei, a charged ion loses most of its kinetic energy through electromagnetic interaction, transferring more energy to the light electrons than to heavy nuclei comprising the target. The energy lost by this latter mechanism results in either excitation or ionization of target atoms. A quantity helpful in discussing the interaction of ionizing radiation with matter is the linear energy transfer (LET) of a particle, which is defined as the energy loss per unit of path length. LET is a function of energy and charge of the incident ion or nuclei, and it depends only weakly on the material composition (Ablen, 1980). The path length of a particle down to its stopping point is called the particle range.

In nuclear collisions of high-energy heavy ions with target nuclei, fast nuclear fragments are produced. These fragments have approximately the same velocity and direction as the incident particle. The fragmentation process alters the composition of a heavy-ion beam passing through the material and reduces the number of original beam nuclei. This reduction is almost independent of energy, because the interaction mean free paths λ_j (j labels the nuclei type) exhibit no strong energy dependence. In a more quantitative treatment of projectile fragments, one must solve the coupled equations

$$\frac{dN_j(x)}{dx}=\frac{-N_j(x)}{\lambda_j}+\sum_{i\neq j} P_{ij}\left[\frac{N_i(x)}{\lambda_i}\right], \qquad (62)$$

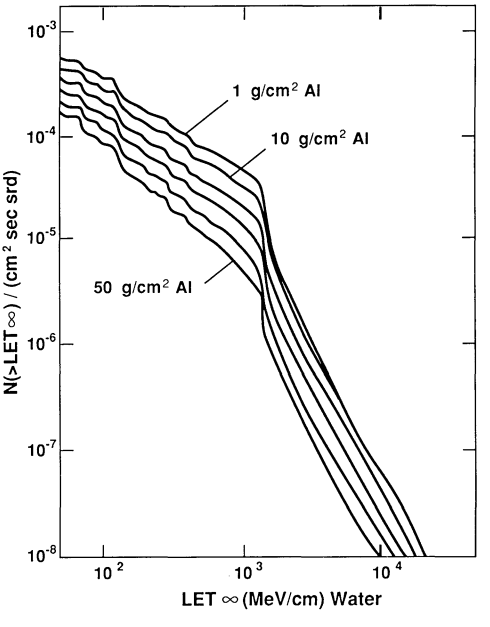

FIG. 43. Calculated integral linear energy transfer (LET) spectra illustrating effects of shield thickness of (1, 10, 20, 30, 40, and 50 g/cm²) aluminum (Al). Altitude 241 km, inclination 57°. [Redrawn from Heinrich *et al.*, 1989.]

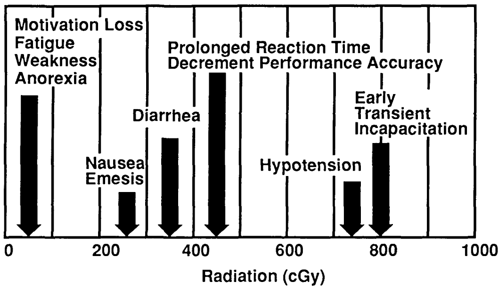

FIG. 44. Performance degradation events occurring at various doses of gamma irradiation in centigrays (cGy). (Redrawn from *Exploring the Living Universe: A Strategy for Space Life Science,* NASA, 1988.)

where $N_j(x)$ is the number of nuclei of type j at depth x, and P_{ij} denotes the probability that a nucleus of type i in a collision with a target nucleus produces a projectile fragment of type j,

$$P_{ij}=\sigma_{ij}/\sigma, \tag{63}$$

where σ is the total interaction cross section and σ_{ij} denotes the partial fragmentation cross sections. Our ignorance of the complete process lies in our limited knowledge of σ_{ij}, as these must be derived from experimental data. In the case of fragmentation in collisions with hydrogen, the empirical cross sections given by Silberberg and Tsao (1973) are useful. Other cross sections can be found by invoking the factorization law suggested by Lindstrom *et al.* (1975). Figure 43 illustrates the effects on the differential LET spectra for the SL-1 (SL-1 denotes the European Spacelab mission of November '83 which had a circular orbit at an altitude of 241 km and an inclination of 57°) behind varying thicknesses of aluminum (Al) shielding (Heinrich *et al.*, 1989); a difference in thickness of 10 g/cm^2 introduces a factor of 1.5 in the LET spectra.

5.3 Biological Effects of Radiation

Radiation is generally measured by its physical dose unit, the *rad*, which denotes the energy deposition by radiation per unit mass of absorber material (1 rad = 100 erg/gm) without reference to its effects. When the effects of radiation on living organisms are evaluated, the absorption of equal amounts of energy per unit mass under different irradiation conditions, for example, heavy ions or neutrons versus gamma irradiation, does not guarantee the same biological effect. To quantify the probable biological effect of a given radiation exposure, the concept of dose equivalent was introduced. The dose equivalent H is the product of the absorbed dose D and the quality factor QF that characterizes the specific radiation:

$$H=D\times \mathrm{QF}. \tag{64}$$

The quantity QF depends on the type of radiation, its energy, and the organ being irradiated. QF can vary from unity for x rays to greater than 20 for alpha particles of typical energies and multicharged particles (ICRU, 1986). If D is expressed in rad, then H is defined in units of rem. If D is expressed in SI units or grays (Gy) then the corresponding unit of dose equivalent is called the sievert (Sv). (See also BIOLOGICAL EFFECTS OF ELECTROMAGNETIC RADIATION.)

The effects of ionizing radiation are generally grouped into two categories: acute effects and long-term effects. Acute effects include radiation sickness and death; long-term effects are carcinogenesis, teratogenesis, and cataract formation. Figure 44 summarizes the types of acute physiological responses and the doses that typically cause their onset, and Figure 45 describes the temporal pattern of radiation effects following ^{60}Co exposure (Conklin and Walker, 1987). For a review of the effects of ionizing radiation on neural function the reader should consult Kimeldorf and Hunt (1965); for a view of the effects of ionizing radiation on behavior one should consult Maier and Landauer (1989) and references quoted.

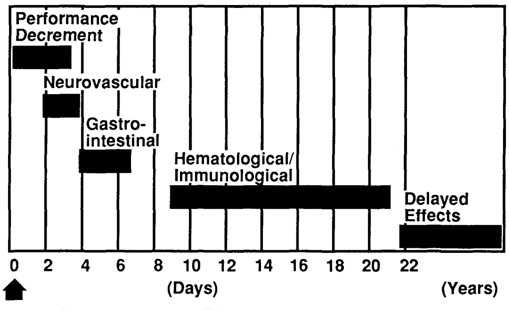

FIG. 45. Temporal relation of postexposure medical effects after gamma irradiation. (Redrawn from *Exploring the Living Universe: A Strategy for Space Life Science,* NASA, 1988.)

5.4 Expected Radiation Exposures in Space Missions

The types of radiation hazards experienced by space crews depend on the mission scenario. Each scenario determines the type and magnitude of the biological effects expected.

5.4.1 LEO Missions. The principal mission for LEO is the Space Station Freedom at orbital parameters of 450 km, 28° inclination. In this case the main source of radiation is trapped protons in the South Atlantic Anomaly (SAA) with GCR contributing a smaller fraction. SAA refers to the region where the van Allen belts are at lower altitudes. This depression of the magnetic field over the coast of Brazil occurs because the dipole term of the geomagnetic field is shifted about 11° from the Earth's axis of rotation and displaced approximately 500 km towards the Western Pacific. In LEO, SAA accounts for nearly all the trapped proton radiation astronauts receive with a small contribution of galactic cosmic radiation. Curtis (1974) estimated for the proposed LEO missions a dose rate of 97 mrem/day to blood-forming organs shielded by 1 g/cm^2 Al. For comparison, a total dose equivalent of 0.1 rem is about equal to that received on Earth from natural sources (other than radon) in 1 yr. For the Apollo missions, during which no major SPEs occurred, astronauts received a skin dose of less than one rad (Atwell, 1979).

5.4.2 Geosynchronous Orbits. During missions in GEO a crew would be exposed to several radiation sources: electrons in the outer radiation belt, bremsstrahlung from electron-shielding interactions, GCRs, and SPEs. The GCR contribution to the total dose is significant but smaller than that of fast electrons. Shielding does not profoundly affect the dose rate, which is estimated to be about 100 mrem/day with no shielding and 50 mrem/day behind 4 g/cm^2 of Al. The dose from SPEs varies with magnitude and duration of the solar flare. For SPEs in solar cycle 19, doses were estimated to be 100–200 rad/day for 2 g/cm^2 shielding and 15–30 rad/day for 5 g cm^{-2} shielding (Silberberg *et al.*, 1985). These are doses sufficient to produce acute effects.

5.4.3 Lunar Mission. In contrast to orbits about the Earth, the Moon offers little or no protection from GCRs or SPEs. At solar minimum the annual dose-equivalent rate from GCRs is estimated to be about 30 rem/year, and doses from SPEs are substantially larger. GCRs are penetrating and their interaction with target material generates a substantial number of neutrons that have a high biological effectiveness. QFs for neutrons between 10 and 20 have been proposed. Figure 46 illustrates this complex dose-depth relationship for the lunar environment. If individuals are to spend years on the Moon, substantial shielding, presumably provided by the surface itself, would be necessary if the lifetime exposure limits given in Table 13 are not to be exceeded. These career limits are based on a lifetime excess risk of cancer of 0.03.

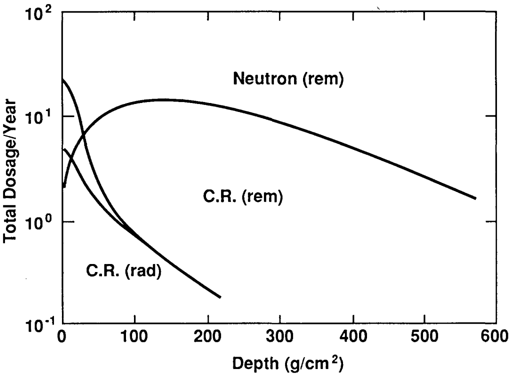

FIG. 46. A comparison of the annual dose equivalent due to secondary neutrons and cosmic-ray nuclei, as a function of shielding depth, and the absorbed dose rate due to cosmic-ray nuclei. (Redrawn from *Exploring the Living Universe: A Strategy for Space Life Science*, NASA, 1988.)

5.4.4 Mars Mission. The radiation space environment for a mission to Mars is essentially the same as for a lunar colony but without the advantage of building sleeping quarters below the surface. A baseline dose for a Mars trip has been calculated to be 43 rem/

Table 13. Career limits (in rem) for lifetime excess risk of fatal cancer at 3×10^{-2}.

	Age of first exposure			
	25	35	45	55
Male	150	250	325	400
Female	100	175	250	300

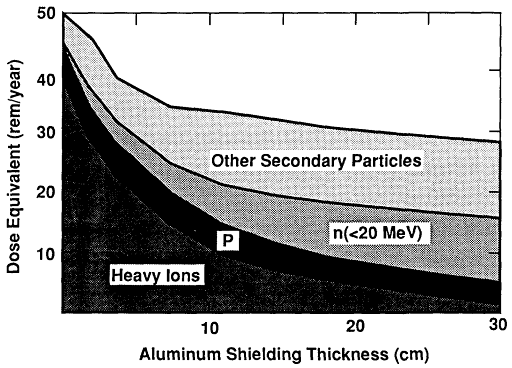

FIG. 47. Dose equivalent at 5-cm tissue depth versus aluminum shielding thickness at solar minimum showing radiation component individuality.

year in free space and 36 rem/year when shielded by 4 g/cm^2 of Al (Letaw, Silberberg, and Tsao, 1986). The capacity of Al for shielding astronauts from galactic cosmic radiation in free space is given by the dose versus depth curves shown in Fig. 47 under solar minimum conditions. It is evident that the heavy ions, the primary source of radiation, are attenuated *e*-fold by approximately 10 cm of shielding, whereas the secondary components, including neutrons, protons, and recoil heavy nuclei, tend to accumulate. For 30 cm of shielding the annual dose equivalent is reduced from 500 mSv to 20 mSv. An identical calculation for solar maximum conditions shows that shielding is even more ineffective (Letaw, Silberberg, and Tsao, 1988).

GLOSSARY

Afferent: Carrying toward a center; centripetal.

Alveoli: Plural of alveolus; small saclike structures that are the terminal elements in the airways (bronchioles) in the lungs.

Anthropomorphic: Having human form.

Atelectasis: Collapse of lung tissue, preventing the respiratory exchange of carbon dioxide and oxygen.

Autonomic Nerve: Part of the nervous system that regulates the activity of cardiac muscle, smooth muscle, and glands.

Baroreceptor: Pressure-sensitive nerve ending found in the walls of the atria of the heart, the vena cava, the aortic arch, and the carotid sinus.

Bradycardia: An abnormally low heart rate (below 60 beats per minute).

Cephalad: Toward the head.

Cerebellum: The inferior part of the brain located in the posterior cranial fossa behind the brain stem; it consists of two lateral lobes and a middle lobe.

Cerebrospinal: Pertaining to the brain and the spinal cord.

DNA: Deoxyribonucleic acid, the genetic material contained in all living cells.

Dorsalis Pedis Artery: Continuation of the anterior tibial artery starting at the ankle joint.

Dose-Equivalent (DE): The product of the dose in rads and a quality factor (QF), that varies with the LET of radiation.

Edema: The abnormal accumulation of fluid in the tissue spaces due to increased transudation of the fluid from the capillaries; dropsy.

Endocrine System: Network of ductless glands and other structures that secrete hormones into the blood.

Enteric: Pertaining to the small intestines.

Epistaxis: Nose bleed.

Genome: The complete set of genes in the chromosomes of each cell of a particular organism.

Gray (Gy): Système International unit of absorbed dose; one gray denotes the absorption of one joule of energy in a mass of one kilogram.

Homeostasis: A tendency to stability in steady states in the organism by coordinated physiologic processes.

Ischemia: Decreased blood supply to a body organ or part, due to inflow of arterial blood or to vasoconstriction; localized tissue anemia.

Jugular bulb: Inferior bulb or superior bulb of the internal jugular vein.

Kinesthetic: The sense of perception of movement, weight, resistance, and position.

Lachrymation: Secretion of tears by the lacrimal glands in the eyelids.

LET: Linear energy transfer; the average loss in energy per unit length of path of the incident radiation.

Lumen: A cavity or channel within any organ or structure of the body.

Meiosis: Division of a sex cell into two, then four, gametes.

Mitosis: Cell division in somatic cell resulting in two genetically identical daughter cells.

Nystagmus: Involuntary, rhythmic movement of the eyeball—horizontal, vertical, rotary, or mixed. It may be congenital, acquired, physiologic, or pathologic, due to neurogenic, myopathic, labyrinthine, or oscular causes.

Oculogyration: Movement of the eyeball about the anteroposterior axis.

Orthostatic Intolerance: Inability to maintain blood pressure in the brain when an individual assumes the standing position.

Os Calcis: Heel bone.

Osmole: The amount of substance that dissociates in solution to form one mole of osmotically active particles; e.g., 1 mole of glucose, which is unionizable, forms 1 osmole of solute, but 1 mole of sodium chloride forms 2 moles of solute. Symbol: osm. Also written osmol. 1000 mosm = 1 osm. Osmolarity refers to osmoles per liter.

Perfusion: Introduction of blood into tissue by their injection into blood vessels.

Petechia: Tiny purple or red spots that appear on the skin as a result of minute hemorrhages within the dermal or submucosal layers.

Proprioceptor: Any sensory nerve ending, in, for example, muscles, tendons, and vestibular apparatus, that responds to stimuli originating in the body regarding movement and spatial position.

Protist: The protoplasm of a cell without its containing membrane.

Quality Factor (QF): An LET-dependent factor by which absorbed dose is multiplied to obtain a quantity that corresponds more closely to the degree of biological effects produced by x or low-energy gamma rays.

Rad: Radiation absorbed dose; a unit of absorbed dose that represents the absorption of 0.01 J of energy per kilogram of absorbing material.

Relative Biological Effectiveness (RBE): Biological potency of one type of radiation as compared to another producing the same biological end-point. It is numerically equal to the inverse of the ratio of absorbed doses of the two radiations required to produce equal biological effects.

Residual Volume: The volume of gas in the lungs at the end of a maximum expiration.

Ribosome: A cytoplasmic organelle composed of ribonucleic acid and proteins.

Scotomata: Dark areas of lost or depressed vision.

Sievert (Sv): Système International unit of radiation dose equivalent; equal to dose in grays times a quality factor.

Smooth Muscle: One of two kinds of muscle, distinguished by being under autonomic (involuntary) control.

SMS: Space motion sickness.

Systole: The period of heart contraction when blood is forced into the aorta and pulmonary artery.

Transmural: Across the wall of an organ.

Urticaria: An itchy skin eruption characterized by transient wheals of varying shapes and sizes.

Venous Compliance: Change in diameter of vessel per unit blood pressure.

Venous Pressure: Pressure exerted by circulating blood on the walls of veins (60–100 mm of water).

Ventilation: Process by which gases are moved in and out of the lungs.

Vital Capacity: Amount of air that can be expelled after a maximum inspiration; normal values range from 4000–5000 mL.

Works Cited

Ablen, S. P. (1980), *Rev. Mod. Phys.* **52**, 121–173.

Adams, J. D., Theis, C. F., Steven, K. W. (1977), Denitrogenation and Renitrogenation Profiles, *Proceedjngs of the Aerospace Medjcjne Meetjng*, Las Vegas, NV.

Anliker, M., Van Ruskirk, W. (1970), *5th Symposium on the Role of Vestibular Organs in Space Exploration*, Washington, DC: U.S. GPO (NASA SP-314), pp. 135–141.

Atwell, W. (1979), Dosimetry in the Manned Space Program, *Internatjonal Solar-Terrestrjal Predjctjons Proceedjngs and Workshop Program*, Boulder, CO.

Bandurski, R., Schulze, A. (1983), *Adv. Space Res.* **3**, 229–235.

Bandurski, R., Schulze, A., Dayanandau, P., Kaufman, P. B. (1984), *Plant Physjology.* **74**, 284–288.

Beard, D. B. (1964) *Rev. Geophys.* **2**, 335–365.

Bergel, D. N. (1961), *J. Physjol.* **156**, 445–467.

Bücker, H., Facius, R., Horneck, G. (1986), *Adv. Space Res.* **6**, 12–115.

Burton, R. K., Leverett, S. D., Michaelson, E. D. (1974), *Aerospace Med.* **45**, 1115–1136.

Butler, B., Hills, B. (1979), *J. Appl. Physjol.* **47**, 537–543.

Buys, E., Rijhaut, P. (1939), *Arch. Int. Physiol.* **49**, 101–105.

Clark, B., Graybiel, A. (1963), *Am. J. Psychol.* **76**, 1–27.

Conklin, J. J., Walker, R. I. (1987), in: J. J. Conklin and R. I. Walker (Eds.), *Military Radiobiology*, Orlando: Academic, Chap. 12, pp. 231–240.

Curthoys, I., Blanks, R., Markhan, C. (1977), *Acta Otolaryngol.* **3**, 258–265.

Curtis, S. B. (1974), in: C. A. Tobias, P. Todd (Eds.), *Space Radiation Biology and Related Topics*, New York: Academic.

de Vries, H. (1950), *Acta Otolaryngol.* **388**, 262–273.

Egmond, A., Groen, J., Jongkees, L. (1949), *J. Physiol.* **110**, 1–17.

Frank, L. A., Van Allen, J. A., Mills, H. K. (1964), *J. Geophys. Res.* **69**, 2121–2191.

Freden, S. C., White, R. S. (1960), *J. Geophys. Res.* **65**, 1377–1383.

Freden, S. C., White, R. S. (1962), *J. Geophys. Res.* **67**, 25–29.

Fritz-Niggli, H. (1988), in: P. D. McCormack, C. E. Swenberg, and H. Bücker (Eds.), *Terrestrial Space Radiation and its Biological Effects*, New York: Plenum, pp. 61–70.

Gazenko, O. G., Grigor'yev, A. J., Natochin, Y. U. (1981), *Major Medical Results of the Salyut 6/Soyuz 185-Day Mission*, Washington, DC: U.S. GPO (NASA NDB-2747).

Glaister, D. H. (1970), *Agardograph* **133**.

Gow, B. S., Taylor, M. S. (1968), *Circ. Res.* **23**, 111–122.

Graybiel, A. (1952), *Arch. Ophthalmol.* **48**, 605–615.

Graybiel, A., Wood, C. D. (1969), *Aerospace Med.* **40**, 638–643.

Groen, J., Lowenstein, O., Vendirk, A. (1952), *J. Physiol.* **117**, 329–346.

Hardnung, V. (1953), *Helv. Physiol. Acta.* **10**, 482–498.

Hattner, R. S., McMillan, D. E. (1968), *Aerospace Med.* **39**, 849–885.

Heinrich, W., Spill, A. (1979), *J. Geophys. Res.* **84**, 4401–4411.

Heinrich, W., Wiegel, B., Ohrndorf, T., Bücker, H., Reitz, G., Schott, J. V. (1989), *Radiat. Res.* **118**, 63–82.

Hess, W. H. (1968), *The Radiation Belt and Magnetosphere*, Waltham: Blaisdell.

Hess, W. H., Mead, G. D. (1968), in: W. H. Hess, and G. D. Mead (Eds.), *Introduction to Space Science*, New York: Gordon and Breach.

Hewes, D., Spady, A. (1964), *Evaluation of a Gravity-Simulation Technique for Studies of Man's Self-Locomotion in Lunar Environment*, Washington, DC: U.S. GPO (NASA TND-2176).

Hinsenkamp, M. (1981), *Aviation, Space and Environmental Medicine* **52**, 95–103.

Hirman, J. W. (1988), in: P. D. McCormack, C. E. Swenberg, and H. Bücker (Eds.), *Terrestrial Space Radiation and its Biological Effects*, New York: Plenum, pp. 61–70.

Hixson, W., Niven, J. (1962), *Frequency Response of the Human Semicircular Canals II*, Pensacola: Naval School of Aviation Medicine Report No. 73.

Hoffler, J. W., Johnston, R. L. (1975), in: *Biomedical Results of Apollo*, Washington, DC: U.S. GPO (NASA SP-367).

International Commission on Radiation Unit and Measurements (1986), *Quality Factor in Radiation Protection*, Bethesda, MD (ICRU Report 40).

Johnston, R. S., Dietlein, L. F., Berry, C. A. (Eds.), (1975), *Biomedical Results From Skylab*, Washington, DC: U.S. GPO (NASA SP-368).

Jones, G. (1965), *J. Physiol.* **154**, 1–32.

Kimeldorf, D. J., Hunt, E. L. (1965), *Ionizing Radiation: Neural Function and Behavior*, New York: Academic.

Lackner, J., Graybiel, A. (1986), *Aviat. Space Environ. Med.* **57**, 229–235.

Leach, C. S. (1983), *Aviat. Space Environ. Med.* **54** (Suppl. 1), S50–S54.

Leach, C. S. (1984), *Aviat. Space Environ. Med.* **58** (Suppl.), A74–A79.

Learoyd, B., Taylor, M. (1966), *Circ. Res.* **18** (Suppl. 1), 278–292.

Letaw, J. R., Silberberg, R., Tsao, C. H. (1986), in *Manned Mars Missions*, Washington, DC: U.S. GPO (NASA-M002).

Letaw, J. R., Silberberg, R. Tsao, C. H. (1988), in: P. D. McCormack, C. E. Swenberg, and H. Bücker (Eds.), *Terrestrial Space Radiation and its Biological Effects*, New York: Plenum, pp. 663–673.

Leverett, S. D., Whinnery, J. E. (1985), in: R. L. de Hart (Ed.), *Fundamentals of Aerospace Medicine*, Philadelphia: Lea and Febiger, Chap. 9.

Lima de Faria, A. (1986), *Molecular Evolution and the Organization of the Chromosome*, New York: Elsevier.

Lindstrom, P. J., Greiner, D. E., Heckman, H. H., Cork, B., Bieser, F. S. (1975), *Isotope Production Cross-Sections for the Fragmentation of ^{16}O and ^{12}C at Relativistic Energies*, Livermore, CA: Lawrence Berkeley Laboratory (LBL-3650).

Mack, P. B. (1971), *Aerospace Med.* 42 (8), 828–833.

Maier, D. M., Landauer, M. R. (1989), *Aviat. Space Environ. Med.* **60**, 774–778.

McCormack, P. D. (1981), *Doppler Monitoring of Peripheral Vascular Status in Human Subjects Under Variable Intra-Arterial Pressure*, Pensacola: Naval Aerospace Medical Research Laboratory.

McCormack, P. D. (1984), *Aviat. Space Environ. Med.*, **55**(7), 620–631.

McCormack, P. D. (1988), *39th Congress of the International Astronautical Federation,* Bangalore, India (IAF/IAA-88-496).

McCormack, P. D. (1989), *40th Congress of the International Astronautical Federation,* Malaga, Spain (IAF/IAA 89-569).

McCormack, P. D., Nicogossian, A. E. (1987), *38th Congress of the International Astronautical Federation,* Brighton, England (IAF/IAA-7-533).

McIlwain, C. E. (1961), *J. Geophys. Res.* **66**, 3681–3691.

Meiry, L. (1965), *The Vestibular System and Human Dynamic Space of Orientation,* Washington, DC: U.S. GPO (NASA CR-628).

Melvill Jones, G., Young, L. (1978), *Acta Laryngol.* **85**, 45–53.

Miller, E., II, Graybiel, A. (1970), *Aerospace Med.* **41**, 1180–1184.

Morrison, P. (1961), in *Encyclopedia of Physics: Cosmic-Rays I,* Vol 46/I. Berlin: Springer, p. 1.

Naugle, J. E., Kniffer, D. A. (1961), *Phys. Rev. Lett.* **7**, 3–4.

Nicogossian, A., Leach-Huntoon, C., Pool, S. (Eds.), (1989), *Space Physiology and Medicine,* Philadelphia: Lea and Febiger, p. 189.

Niven, J., Hixson, W. (1961), *Frequency Response of the Human Semicircular Canals I,* Pensacola: Joint NASA/Naval School of Aviation Medicine Report.

Niven, J., Hixson, W., Correia, M. (1965), *An Experimental Approach to the Dynamics of the Vestibular Mechanism,* Washington, DC: U.S. GPO (NASA SP-77), pp. 43–45.

Oman, C. (1981), in: J. Gualtierotti (Ed.), *The Vestibular System,* New York: Springer, Chap. 5.

Oman, C. (1982), *Acta Otolaryngol.* **392**, 4–42.

Oman, C., Young, L. (1972), *Acta Otolaryngol.* **74**, 324–331.

Rambaut, P. C., Johnston, R. S. (1979), *Acta Astronautica* **6**, 1113–1122.

Roach, M., Burton, A. (1957), *Can. J. Biochem. Physiol.* **55**, 681–690.

Scraf, F. L. (1966), *Raumfahrtschung* **10**, 114–118.

Shea, M. A., Smart, D. F. (1967), *J. Geophys. Res.* **72**, 2021–2027.

Shea, M. A., Smart, D. F., McCracken, K. G. (1965), *J. Geophys. Res.* **70**, 4117–4130.

Silberberg, R., Tsao, C. H., (1980), *Astrophys. J.* (Suppl.) **25**, 315–368.

Silberberg, R., Tsao, C. H., Adams, J. H., Letaw, J. R. (1985), in: W. W. Mendell (Ed.), *Lunar Base and Space Activities of the 21st Century,* Houston: Lunar and Planetary Institute.

Simpson, J. A. (1983), *Ann. Rev. Nucl. Part. Sci.* **33**, 323–325.

Steer, R., Li, Y., Young, L., Meiry, J. (1967), *Physical Properties of the Labyrinthine Fluids,* Washington, DC: U.S. GPO (NASA SP-152).

Steinhausen, H. (1933), *Pflugers Arch. Physiol.* **232**, 500–512.

Stone, R. W., Letko, W. (1964), *1st AIAA Annual Meeting,* Washington, DC (AIAA Paper No. 64-128).

Stone, R. W. Jr., Piland, W., Letko, W. (1970), *4th Symposium on the Role of Vestibular Organs in Space Exploration,* Washington, DC: U.S. GPO (NASA SP-187).

Störmer, C. (1930), *Z. Astrophys.* **1**, 237–251.

Stufakov, G. P. (1984), *Kosm. Biol. Aviakosm. Med.* **18** (2), 33–37.

Sugiura, M., Heppner, J. P. (1968), in: W. N. Hess, and G. D. Mead (Eds.), *Introduction to Space Science,* New York: Gordon and Breach, pp. 5–92.

Thornton, W., Hoffler, W. (1977), in *Biomedical Results of Skylab,* Washington, DC: U.S. GPO (NASA SP-377).

Thornton, W., Hoffler, W., Rummel, J. (1977), *Anthropometric Changes and Fluid Shifts,* Washington, DC: U.S. GPO (NASA SP-377).

Trinker, D. (1962), in *Symposium of the Society of Experimental Biology,* Cambridge Univ. Press (No. 16, Biological Receptor Mechanisms).

Tucker, W. K., Ganicki, J., Plowman, F. Ratal, D. (1969), *J. Appl. Physiol.* **26**, 656–658.

Van Allen, J. A., Frank, L. A. (1959a), *Nature* **183**, 430–434.

Van Allen, J. A., Frank, L. A. (1959b), *Nature* **184**, 219–224.

Vaughan, B. E. (1960), *Am. J. Physiol.* **108**(b), 1255–1264.

Verwey, E., Overbeck, J. (1948), *Theory of Stability of Lyophobic Colloids,* New York: Elsevier, pp. 160–192.

Vogel, J. M., Whittle, M. W. (1976), *Aviat. Space Environ. Med.* **47**, 396–400.

Waligora, J. M., Horrigan, J. Jr., Conklin, J., Hadley, A. T., III. (1984), *Verification of an Altitude Decompression Sickness Protocol for Shuttle Operations Using a 10.2 psi Pressure State,* Washington, DC: U.S. GPO (NASA TM-5259).

Wareing, P. F. (Ed.), (1983), *Plant Growth Substances,* New York: Academic.

Weibull, C. (1955), *Exp. Cell Res.* **9**, 294–304.

Westerhof, N., Noordergraff, A. (1970), *J. Biomech.* **3**, 357–379.

Yorob'yev, Y., Gazenko, O. G., Shulzhenko, E. (1986), *Space Biology and Aerospace Medicine,* Vol. 2. Moscow: Medical Press, Chap. 5.

Young, L. (1974), in: V. Mountcastle (Ed.), *The Medical Physiology,* St. Louis: Mosby.

Young, L., Dichgaus, J. (1973), *Acta Otolaryngol.* **76**, 24–31.

Further Reading

McCormack, P. M., Swenberg, C. E., Bücker, H. (Eds.) (1988), *Terrestrial Space Radiation and Its Biological Effects,* New York: Plenum.

NCRP Report No. 98 (1989), *Guidance on Radiation Received in Space Activities,* Bethesda, MD: NCRP.

Nicogossian, A. E. (Ed.) (1989), *Space Physiology and Medicine,* Philadelphia: Lea and Febiger.

Van der Walt, D. J., Wolfendale, A. W. (1988), *Space Sci. Rev.* **47**, 1–45.

AIR POLLUTION: COMPONENTS, CAUSES, AND CURES

ARMISTEAD G. RUSSELL, *Department of Mechanical Engineering, Carnegie Mellon University*

Introduction 489
1. **History** 490
2. **Components and Scales of Pollutant Systems** 491
2.1 Pollutants, Sources, Effects, and Measurement 491
2.1.1 Formaldehyde, Benzene, and Other Organics 491
2.1.2 Nitrogen Oxides 492
2.1.3 Ozone and Other Photochemical Oxidants 493
2.1.4 Acids and Sulfur Dioxide 493
2.1.5 Carbon Monoxide 494
2.1.6 Chlorofluorocarbons 494
2.1.7 Carbon Dioxide 494
2.1.8 Aerosols 494
2.1.9 Radioactive Contaminants 494
2.2 Spatial Scales of Air Pollutant Systems 494
2.2.1 Indoor Air Pollutants 494
2.2.1.1 Indoor Anthropogenic Pollutants 495
2.2.1.2 A Natural Indoor Air Pollutant: Radon 495
2.2.2 Local Scale Air Pollution 495
2.2.2.1 Release of Noxious and Toxic Chemicals 496
2.2.3 Urban and Regional Scale Air Pollution 496
2.2.3.1 "London Smog" 496
2.2.3.2 Photochemical Smog: An Urban and Regional Phenomenon 497
2.2.3.3 Acid Deposition 499
2.2.4 Global Air Pollution 499
2.2.4.1 Global Warming 499
2.2.4.2 Ozone Depletion 501
3. **Air Pollution Meteorology and Pollutant Transport** 502
4. **Atmospheric Chemistry** 504
4.1 Tropospheric Chemistry 504
4.1.1 Urban Photochemistry 504
4.1.2 Regional Gas Phase Photochemistry and Acid Formation 505
4.1.3 Regional Aqueous Phase Chemistry 506
4.2 Stratospheric Chemistry 507
5. **Air Pollution Control** 508
5.1 Air Quality Regulations 508
5.2 Determining Source–Air-Quality Relationships 509
5.3 Control Devices 511
5.4 Developing Strategies to Improve Air Quality 512
6. **Summary** 512
Glossary 513
Works Cited 513
Further Reading 514

INTRODUCTION

There are literally thousands of pollutants present in the atmosphere. These pollutants are trace contaminants that, either by themselves or upon interaction with other pollutants, adversely impact our environment. Not all contaminants have universally adverse impacts, even though they affect the environment. Consider the environmental havoc precipitated by the buildup of free oxygen as the earth evolved. Many existing organisms had to seek refuge below the surface of the land or water. However, at the same time it allowed other organisms to flourish. To the previous inhabitants oxygen was a pollutant, to those that followed it is a necessity.

Most pollutants, however, have an overall negative impact on the environment as judged by health, environmental, and mate-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

rial degradation. In an economic or social sense, indoor radon, urban smog, regional acid deposition, destruction of stratospheric ozone, and rapid global warming are undesirable. For this reason the chemical compounds that lead to these phenomena are termed pollutants. A pollutant can be defined as a substance that is present at some location at an undesirably high concentration. This definition allows for the possibility that a compound may not be a pollutant at one concentration, although its buildup may have adverse consequences. CO_2 is necessary for plant life, but too much leads to global warming. Likewise, a compound may be beneficial in one location and not another; e.g., ozone is beneficial in the stratosphere but is a primary component of urban air pollution in the troposphere. Time is also important in determining how a contaminant will impact the environment. If the buildup of a pollutant is slow enough, the ecological system can adjust, as was the case with the buildup of oxygen over millions of years. However, man's activities have led to increases in pollutants at a rate that does not allow complete adjustment. This is one of the concerns of the greenhouse effect—not only that there is a warming trend, but that it is warming more quickly than the environment can adapt to the new conditions.

The approach taken here is to review the general characteristics of the more common air pollution systems and then to discuss their chemistry, physics, and control in more detail. More extensive information on certain aspects, such as meteorology, the climate, and atmospheric physics are contained in articles devoted to those fields (see ATMOSPHERIC PHYSICS; METEOROLOGY). Following this introduction is a brief history of air pollution. Next, the compounds commonly considered air pollutants are discussed, including their effects and where they are found. Included are brief discussions of the more pressing air pollutant problems we now face. Most air pollution systems are due to the interaction of pollutant transport and chemistry, as is discussed in the sections on meteorology and atmospheric chemistry. The last sections cover pollution control, including control devices, air quality regulations, and the development of effective control strategies.

1. HISTORY

Air pollution is not a new phenomenon, though it is growing in scope and complexity. Use of fire by early man was accompanied by smoke filling their dwellings. As people started to congregate into cities, the problem intensified. Accounts from early Rome testify to the soot and stench of the city (e.g., Stern *et al.*, 1984; Brimblecombe, 1984).

In London, the smoke and fumes from wood and coal burning have plagued the city since medieval times. In spite of a series of royal edicts and laws, the air continued to worsen. In 1661 John Evelyn wrote for Charles II and Parliament the classic "Fumifumigium, or the Inconvenience of the Aer, and Smoake Dissipated," describing the sources, effects, and remedies for the air pollution inundating the city. While the work was far-sighted, it had relatively little impact, and London continued to experience "Great Stinking Fogs," the result of coal smoke and fog.

The industrial revolution was spurred by the ability to use steam to drive machinery, and coal was used to generate the steam. The prosperity afforded by the increased manufacture of products accompanying these plants made smoke a tolerated inconvenience, a view still held today in developed and developing nations. In London, however, the oppressiveness of the smog caused sufficient pressure for Parliament to pass Great Britain's Public Health Act in 1848. Again, neither that legislation, nor later versions, were entirely effective. In 1952, the Great Smog encompassed London for four days, leading to an estimated 4000 excess deaths. London was not alone. The Meuse Valley (Belgium) and Donora, Pennsylvania, were hit by similar conditions in 1930 and 1948, respectively, leading to increased rates of illness and death.

While London, and indeed most large cities, were struggling to clean up coal smoke, Los Angeles was "cooking up" a new form of air pollution (Finlayson-Pitts and Pitts, 1987). Unlike London, Los Angeles smog was not a result of smoke mixing with fog. In this case, the severity increased with sunlight. Photochemical smog, as it has become known, was a result of automobile, power plant, refinery, and other source emissions interacting in the sunlight to form ozone, formaldehyde, and acids. The pollutants formed were more obnoxious than those originally emitted. It is now recognized that most major cities experience photochemical smog. Through aggressive control measures instituted, the magnitude of the problem has been reduced in Los Angeles, despite continued population growth.

While we are still struggling for solutions to urban air pollution, expansion of man's activities, accompanied by our ability to identify more subtle, global changes, has raised other problems. In the 1970s and 80s, increasing attention has been focused on regional and global changes in air quality. On the regional scale, power plant and automotive emissions are reacting to form atmospheric sulfuric and nitric acid. These compounds deposit on the ground, acidifying lakes and streams and stressing tree growth. Acidification has resulted in dramatic ecological changes in lakes in Canada, the northeastern United States, and Scandinavia. In addition to acids, it was found that the urban ozone and its precursors were being carried into rural areas.

Probably the most alarming discoveries involve global changes in the Earth's atmosphere arising from man-made emissions. Increased carbon dioxide emissions and other man-made emissions may be leading to global warming, or the "greenhouse effect." Emissions of chlorofluorocarbons (e.g., Freon from automobile air conditioners) are expected to continue to decrease stratospheric ozone. This layer of ozone protects the Earth's surface from harmful UV radiation. What is most ominous about the global problems is that they are very tough to attack, and even with very aggressive efforts they are expected to worsen. In contrast, controlling urban smog and acid deposition has begun with some success. As man's activities have diversified, the variety of impacts has multiplied as have the strategies necessary for remedial action.

2. COMPONENTS AND SCALES OF POLLUTANT SYSTEMS

Air pollutant systems are comprised of a variety of compounds, and differ markedly depending on spatial scale and location. Historically, air pollution was first an indoor and local phenomenon. Next, the scope grew to urban areas, then regional. Now the global consequences of man's actions are recognized. This provides a convenient framework for classifying pollutant systems. First, however, it is instructive to identify the compounds that are called pollutants, their sources and effects.

2.1 Pollutants, Sources, Effects, and Measurement

Airborne matter is present in all three phases: gas, liquid, and solid. In the condensed phases, the pollutants are found as small particles (called aerosol or particulate matter) less than about 10 μm in diameter. Particles larger than this rapidly settle out. Pollutants may be either primary or secondary. Primary pollutants are emitted directly. Secondary pollutants are formed in the atmosphere by chemical reactions of primary pollutants. A number of both primary and secondary pollutants are listed in Table 1, along with representative concentrations and sources (Brimblecombe, 1986; Seinfeld, 1987; Finlayson-Pitts and Pitts, 1987; Stern *et al.*, 1984; Schwartz, 1989). Many of the pollutants are both anthropogenic (man-made) and natural in origin, and it is evident that a single source is seldom the sole cause of an air pollution problem. Also included in the table is the lifetime of each pollutant, which is the approximate period before the compound is chemically destroyed in the troposphere. Also listed is the spatial scale where the compound acts. Below, various pollutants are discussed in more detail, along with their effects and techniques used to measure the atmospheric concentration. Their role in particular air pollutant systems and atmospheric chemistry are delineated in later sections.

2.1.1 Formaldehyde, Benzene, and Other Organics. Organic gases are emitted by a variety of processes and, except for individual species detailed below, are of concern as a precursor to photochemical smog. Hundreds of organics have been identified in the atmosphere. Their role in forming ozone can be viewed as providing the energy and chemical pathways necessary to oxidize NO to NO_2. Hydrocarbons, which comprise a majority of the organics emitted, can be measured by gas

Table 1. Pollutant lifetimes, concentrations, and sources.

Pollutant	Atmospheric chemical lifetime	Scale of action[a]	Concentrations[b] rural/urban (ppb)	Major sources (as a pollutant)
Methane	7 yr	R, **G**	1600/2000	Swamps, rice paddies, animal digestive systems
Butane	4 d	**U**, R	1/10	Automobiles, solvents
Toluene	2 d	**U**, R	0.1/5	Automobiles, solvents
Propene	4 h	**U**, R	0.1/5	Automobiles, solvents
Isoprene	3 h	U, **R**	10/0.1	Trees
Formaldehyde (HCHO)	4 h	I, **U**, R	0.4/20	Atmospheric reaction of other organics
Sulfur dioxide (SO_2)	13 d	U, **R**	5/100	Power plants, volcanoes
Sulfuric acid (H_2SO_4)	10 d[c]	R	$5/10\ \mu g\ m^{-3}$	Gas and liquid phase oxidation of SO_2
Nitric oxide (NO)	8 h[d]	I, U	0.01/100	Automobiles, power plants
Nitrogen dioxide (NO_2)	8 h	**I**, **U**, R	0.3/100	Atmospheric oxidation of NO
Nitric acid (HNO_3)	90 d[c]	R	0.1/15	Atmospheric oxidation of NO
Carbon monoxide (CO)	50 d	**I**, **U**, R	120/5000	Automobiles, incomplete combustion, methane oxidation
Carbon dioxide (CO_2)	7 yr	G	350 ppm	Power production, heating
Nitrous oxide (N_2O)	20 yr	G	0.3/0.3	Biogenic, fertilizer use
Carbon tetrachloride	70 yr	G	0.1	Solvent
Trichlorofluoroethylene (Freon 11)	100 yr	G	0.1	Air conditioners, foaming agent
Hydroxyl radical (OH)	0.01 s	**U**, **R**, **G**	10^{-3}	Atmospheric reactions of NO_x and organics
Hydrogen peroxide	2 d	U, **R**	0.1/0.1	Atmospheric reactions of NO_x and organics
Nitrogen trioxide (NO_3)	0.1 s	U, R, G	0.01/0.1	Atmospheric reactions of NO_2 and O_3
Ozone (O_3)	2 min	**U**, R	40/300	Atmospheric reactions of NO_x and organics
Peroxyacetyl nitrate	2 h	U, R	5/30	Atmospheric reactions of NO_x and organics
Ammonia (NH_3)	50 d	U, R	1/10	Biogenic, e.g., animal waste decomposition

[a] Letters refer to the scale of air pollutant system where species play a significant role: I, indoor; U, urban; R, regional; and G, global. The bold letter indicates scale where the compound is of particular interest.

[b] Approximate concentrations in rural areas and polluted urban areas affected by these pollutants. Many pollutants have strong diurnal and spatial variation.

[c] Long-lived species. Primary removal is not via chemical reaction.

[d] NO and NO_2 are in rapid steady state with ozone. Lifetime of the two combined in an urban atmosphere is about 8 h.

chromatography and flame ionization (Wellburn, 1988).

Formaldehyde is a major indoor pollutant because it is emitted by many building products. Outdoors, formaldehyde is formed by the oxidation of other organics. It photolyzes rapidly, leading to ozone formation. Formaldehyde is an eye irritant at low concentrations [~ 100 parts per 10^9 (ppb)]. Exposure to these levels would occur primarily indoors and very seldom outdoors in a polluted urban area. Formaldehyde also is a suspected carcinogen, though tests are not definitive. Benzene, which is a component of gasoline, is of concern because it is a carcinogen.

2.1.2 Nitrogen Oxides. Oxidized (or odd) nitrogen is a primary ingredient in photochemical smog and in the stratospheric ozone balance. Anthropogenic odd nitrogen is primarily emitted as nitric oxide (NO), which is oxidized in the atmosphere to NO_2, nitric acid (HNO_3), aerosol nitrate, nitrogen trioxide (NO_3), and dinitrogen peroxide (N_2O_5). Of these, NO and NO_2 (referred together as NO_x) are the most plentiful in urban atmospheres, followed by nitric acid and aerosol nitrate. Nitrous oxide, N_2O, is emitted biogenically by denitrification of other nitrogen compounds. It is relatively stable in the troposphere, and is a source of stratospheric odd nitrogen.

NO and NO_2 are measured by chemiluminescence with O_3. Nitric acid and aerosol nitrate are measured by collection on an absorbing medium followed by ion chromatography. NO_3, which is found at very low concentrations (part per trillion range) is measured spectroscopically.

NO is of concern as a pollutant because it is a precursor to NO_2, nitric acid, and ozone formation. NO_2 is an oxidant and can impair lung function, particularly in concert with other pollutants such as ozone. Nitric acid is a primary component of acid deposition, and is an irritant. Health impacts of aerosol nitrate, or individual components of most aerosols, are not well characterized. However, small particles can be respired easily into the lung, and be deposited there. Aerosol nitrate inhibits visibility and is one of the most abundant components causing the haze associated with Los Angeles smog.

2.1.3 Ozone and Other Photochemical Oxidants. Ozone is the highly reactive, triatomic form of oxygen, O_3. It is a strong oxidant, formed in the atmosphere by the combination of atomic oxygen with an oxygen molecule. In the troposphere, it is a primary constituent of photochemical smog, and the severity of photochemical air pollution episodes is often measured by the maximum ozone level reached (National Research Council, 1977). Ozone is a secondary compound, formed by reactions of organic gases and nitrogen oxides. Exposure to ozone impairs lung function and can cause pulmonary stress in humans. Recent studies have indicated that effects on humans are noted above 0.08 ppm. This can be compared to urban levels reaching maximum levels of 0.4 ppm. Ozone also attacks plants, inhibiting growth and increasing susceptibility to attack by pests. Ozone can also attack materials, such as rubber, causing premature aging and oxidation. As an indication of the attack of ozone on materials, a previous measurement technique relied on how rapidly a rubber band cracked from exposure. Currently, tropospheric ozone is measured by use of chemiluminescent reaction with ethylene.

In the stratosphere ozone is desirable. The depletion of stratospheric ozone by manmade pollutants is of concern because ozone shields the earth's surface from harmful UV radiation. Stratospheric ozone is measured by both ground-based and satellite-based spectroscopy systems. In this case, the absorption or emission of light is used to quantify the amount of ozone in a column of air.

Other photochemical oxidants include peroxyacetyl nitrate (PAN) and hydrogen peroxide (H_2O_2). These compounds are formed along with ozone by reactions of nitrogen oxides and organic gases. PAN is a phytotoxin and eye irritant. H_2O_2 is a strong aqueous phase oxidizing agent, converting SO_2 to sulfuric acid.

2.1.4 Acids and Sulfur Dioxide. Sulfuric and nitric acids are recognized as the two most abundant species contributing to acid deposition. Recent attention is also being given to organic acids. In addition to naturally occurring sulfuric acid, excess is formed from the oxidation of anthropogenic sulfur dioxide (SO_2) emissions. Natural sources of sulfur include volcanoes and biological emissions of carbonyl sulfide, carbon disulfide, hydrogen sulfide, and dimethyl sulfide. Nitric acid is a product of NO_x oxidation.

Atmospheric acids are blamed for acidification of lakes and streams in the northeastern United States, Canada, and Scandinavia. These areas are more susceptible to acidification than other areas because they are downwind of major SO_2 sources and they do not have the buffering capacity of other regions. Acidification of lakes and streams has led to dramatic decreases in game fish populations and, in some cases, eradication of the native ecosystem. Acid deposition has also been cited as a reason for forest decline and materials damage. Sulfur dioxide is a lung irritant. Health impacts appear to be exacerbated by the simultaneous exposure to ozone or particulates. Acid is probably a precursor to the compounds responsible for impaired health in London-type smog episodes. Atmospheric sulfuric acid is in a condensed phase, and can be measured by filtration followed by ion chromatography. SO_2 is measured by flame photometric and pulsed fluorescence.

2.1.5 Carbon Monoxide. Carbon monoxide (CO) is emitted by most sources burning fossil fuels. In urban areas, motor vehicles are the largest contributor. The most serious exposure of humans to CO, however, is cigarette smoking. Indoors, CO can result from automobile testing and repair, and from incomplete combustion of fuels in heaters, boilers, or cooking applications. Outdoors, CO builds up during stagnation episodes in large cities, and is particularly severe during the winter months around rush hour.

Carbon monoxide is absorbed into the blood stream, diminishing its ability to absorb oxygen, i.e., carbon monoxide poisoning. This results in fatigue, headache, and, at high levels (>600 ppm), convulsions and possibly death. CO is measured routinely by electrochemical techniques and infrared detection.

2.1.6 Chlorofluorocarbons. Chlorofluorocarbons (CFCs) are man-made compounds consisting of chlorine and fluorine atoms attached to hydrocarbon chains. Of these, Freon 11 and 12 (CCl_3F and CCl_2F_2) are the most well-known and most used. These compounds are very stable in the troposphere, and can migrate into the stratosphere. In the stratosphere, they photodissociate, releasing Cl atoms that lead to the enhanced destruction of ozone. They also absorb light in the infrared range and can lead to global warming.

2.1.7 Carbon Dioxide. A few years ago, carbon dioxide (CO_2) would not have been considered a serious pollutant. However, increased fossil fuel consumption for producing energy is leading to higher CO_2 concentrations, which in turn act to blanket the earth from losing heat, i.e., the "greenhouse effect." The threat of increasing world temperatures, coupled with desertification of major agricultural regions and flooding of coastal cities is of concern. Aside from global warming, ambient CO_2 buildup has no major health impact, and can enhance crop yields.

2.1.8 Aerosols. Atmospheric aerosol, also called particulate matter, is composed of many compounds, including sulfate, nitrate, organics, elemental carbon, metals, sea salt, and soil dust. The abundance of each component changes markedly with geographic location and time. Some aerosols are formed in the atmosphere by the condensation of gas phase species (e.g., sulfate, nitrate, and some organics), while others are emitted directly (e.g., soil dust and sea salt spray).

Health effects of aerosols are not well established, and will undoubtedly be closely linked to their composition. Small aerosol particles, about 0.1 to 3 μm, are easily respirable into the lung, intensifying their health impact. Some of the constituents of aerosols have been shown to be mutagenic and cancerogenic in animals. Atmospheric aerosol composition is usually measured by filtration followed by chromatography or spectroscopy.

2.1.9 Radioactive Contaminants. Radioactive air pollutants can be either anthropogenic or natural. Anthropogenic sources include accidental release from nuclear power plants and detonation of nuclear weapons. The recent Chernobyl, USSR, reactor accident is an example of such a release. Fallout from that accident has been measured across the northern hemisphere.

Radon, a natural, relatively short-lived radioactive gas, migrates through the soil and is emitted into the atmosphere or buildings. Products of radon decomposition, when inhaled, can lead to cancer. Radon is measured by alpha-track detectors, scintillation cells, and radiation detectors.

2.2 Spatial Scales of Air Pollutant Systems

Adverse impacts from the compounds discussed above can be identified on spatial scales from a few meters (e.g., indoors) to those encompassing the globe (1000s of km). In between are local, urban, and regional scale systems. At each scale, particular problems are of interest.

2.2.1 Indoor Air Pollutants. Because of their enclosed nature, indoor environments such as houses, office buildings, and garages can experience extremely elevated levels of many compounds. Examples of man-made indoor pollutants are formaldehyde, NO_x, and carbon monoxide. The most notorious natural indoor air pollutant is radon—a colorless, odorless, radioactive gas. It is ironic that the buildup of both anthropogenic and natural contaminants has been exacerbated by measures taken to reduce energy usage, which was motivated in part by environmental concern.

2.2.1.1 Indoor Anthropogenic Pollutants. Indoor sources of anthropogenic pollutants include the combustion of fossil fuel and the use of certain building materials. Cooking and heating lead to emissions of NO_x and CO. Aging building materials can give off formaldehyde.

Particle board and plywood, both common building materials, are formed using formaldehyde-based resins as adhesives. Also, urea-formaldehyde foam was once used as insulation. As these materials age, they emit formaldehyde. To a lesser extent, combustion of cigarettes and other consumer products may also emit formaldehyde.

Concentrations of formaldehyde (HCHO) in buildings can be greatly elevated above outdoor levels, even in heavily polluted urban areas. Peak levels of up to 4.8 ppm have been measured in mobile homes that used formaldehyde-based resins and insulation (National Research Council, 1981). Average levels are much lower: 0.4 ppm in houses that use formaldehyde foam insulation and 0.01 to 0.1 ppm in houses that do not (National Research Council, 1981). Peak levels in urban areas seldom exceed 0.05 ppm, and average levels are on the order of 0.01 ppm.

HCHO has been shown to cause cancer in rats, but not in similarly exposed mice, and carcinogenicity in humans has not been established. However, at the levels reached indoors, formaldehyde can cause eye irritation, headaches, nausea, fatigue, and thirst. This has led to what is termed the "sick house" syndrome. Reducing HCHO levels can be achieved by replacing the building materials with products that do not emit as much HCHO, by sealing existing walls, or by increasing ventilation either with outside air or with use of a filter that can remove HCHO. Also, new building materials have been developed to alleviate the sick house syndrome.

A second class of indoor anthropogenic pollutants is derived from combustion sources. Combustion in unvented gas appliances, wood burning, and smoking are the major sources of indoor CO and NO_x (Walsh *et al.*, 1984). Under extreme circumstances these sources can lead to indoor levels approaching short term standards. Wood burning can form polycyclic aromatic hydrocarbons (PAHs), such as benzo-(*a*)-pyrene, which are suspected carcinogens.

2.2.1.2 A Natural Indoor Air Pollutant: Radon. Radon is a colorless, odorless, gaseous decay product of radium. After formation, radon migrates through the soil and is emitted into the atmosphere and building basements. In the free atmosphere, it readily disperses and decays, and does not constitute a serious problem. However, in buildings without adequate ventilation, radon can build up to dangerously high levels.

Radon itself is not the direct actor. It further decays into radioactive polonium, which condenses out of the gas phase and attaches to small particles. These particles are easily respirable. Polonium further decays forming radioactive isotopes of lead, then bismuth, polonium again, then finally lead. The half lives of the first four radon progeny, also called "radon daughters," are less than one-half hour. Each decay of polonium is accompanied by the emission of an α particle. In the lung, this promotes cancer.

Radon levels in houses are highly variable, ranging from near zero to 2800 picocuries liter^{-1} (pCi/L) (Lafaviore, 1987). Concentrations above 4 pCi/L are considered hazardous. The reasons for the variability include location, building construction, and age. The large variability in building concentrations and the long times between exposure and the development of cancer lead to difficulty in identifying problem areas. Cracks in subterranean basement walls and floors can greatly increase radon migration into buildings. Control measures include sealing basement walls, increasing ventilation, and ventilation of contaminated air from below the building foundation.

2.2.2 Local Scale Air Pollution. Somewhere between the size of a typical house ($\sim$10 m) and cities (10s of km) lies a class of pollutant systems that are rather localized, impacting street canyons, sections of cities, and small towns. These systems are characterized by relatively concentrated releases of pollutants that have not had a chance to disperse over an area large enough to make their impact negligible. In most cases, the relatively short time involved in dispersion minimizes the role of chemical transformation. Examples of such incidences include inadvertent releases of noxious chemicals, the buildup of exhaust in street canyons, or the buildup of wood smoke in resort areas during the winter. Dynamics of CO and NO_x buildup in street canyons, and of particulates and wood smoke in mountain towns, are very similar to urban systems, though smaller and more concentrated. Concentrated releases of noxious chemicals pose a different type of threat, and are discussed here.

2.2.2.1 Release of Noxious and Toxic Chemicals. Locally high concentrations of noxious and, sometimes, toxic chemicals have resulted from inadvertent releases from storage or production facilities or during transport. Examples include ruptures of storage tanks and accidents involving railroad or truck transport. Contaminants emitted range widely, and include strong acids, bases, fertilizers, and oxidants—virtually any chemical used in processing. The effects are often intensified because the gas emitted is denser than air and tends to stay near the ground. An extreme example of the result of such releases is the Bhopal, India, disaster where methylisocyanate was released into the atmosphere, killing over 2000, and injuring thousands more. Methylisocyanate is an intermediate in the production of a pesticide. In Sarreto, Italy, an inadvertent release of organic gases was found to contain dioxin, an extremely potent toxin, requiring the evacuation of the area. Releases of toxic gases, such as acid, ammonia, and chlorine, also occur during transport.

2.2.3 Urban and Regional Scale Air Pollution. Urban (10s–100s of km) and regional (100s–1000s of km) scale air pollution systems are treated together here because the two are closely linked. The two systems can be characterized as having characteristic residence times on the order of one day to one week. Over this time, atmospheric transport, diffusion, chemical transformation, and deposition are all integral in determining the fate and impact of pollutant emissions. In many cases the pollutants with the greatest impacts are formed in the atmosphere from sources 10s to 100s of km away.

Three relatively distinct types of problems can be identified within this spatial range. "London smog" refers to the type of events that London has experienced for over 500 years, and is a result of particulate and SO_2 emissions. Photochemical air pollution is the type of smog associated with Los Angeles, CA, but also affects urban areas throughout the world. Recently, it has been recognized that the transport of these photochemical pollutants and their precursors impact regional scales. Acid deposition, often referred to as "acid rain," is a result of the deposition of sulfuric and nitric acid, which are products of atmospheric oxidation of emissions from fossil fuel burning.

2.2.3.1 "London Smog." London historically experiences extended periods of meteorological stagnation, inversion, and fog. The inversion, as discussed in Sec. 3, on pollutant transport, traps emissions near the ground. Over a period of a few days pollutants can build up to very high levels. In London, the buildup of the smoke with the fog (the origin of the term smog) and SO_2 led to periods of very dense hazes, a strong odor, and ultimately an increase in deaths. While the 1952 London episode that lasted four days is the most infamous, these events have been noted in London since the 13th century.

The toxic agent(s) directly responsible for the increased mortality rates in London and other areas has not been identified, though one can get a general sense of possible causes. SO_2 emissions are absorbed by the fog and, over the course of a few hours to days, can react. All the products of these reactions, and the mechanism involved, are not known. The presence of fog leads to products that would not be found (at least at such high levels) due to gas phase reactions of SO_2. Bisulfate, which is known to cause allergic reactions in some individuals, can be formed, though this is only one of a myriad of compounds produced by the aqueous phase reactions. Smoke may serve multiple roles in adding to the severity of the smog episodes. First, coal smoke contains metals which promote aqueous reactions of SO_2 products. Smoke and acid particles can also act as a condensation site for water vapor, enhancing fog formation. The fine particulate matter that makes up smoke is readily respirable and can irritate the lung, and the simultaneous presence of particulate matter increases irritation from SO_2 exposure. SO_2 exposure, even at levels below those found during severe smog events, can lead to lung impairment. CO would also be elevated during stagnations. It is possible that the increase in deaths is attributable to more than one agent acting in unison with others. Chronic exposure to London-type smog can also impact health. Prolonged exposure to smoke leading to blackening of the lungs has been cited as possibly accounting for 1 out of 20 deaths in early Britain (Brimblecombe, 1987). John Evelyn noted in 1661 that, of the children born in London, half died before the age of 2, while those living in the country could expect to live to 40. He blamed this on the "foul air."

Reducing the impact of London-type smog has been achieved by a variety of techniques. One of the first strategies was fuel-switching, which is still in practice today to fight not only smoke, but photochemical smog and SO_2. In the 13th century sea coal was recognized as producing excess smoke, and was banned. A second historical measure was to develop cleaner combustors with longer residence times to achieve more complete combustion. A third strategy was to build higher smokestacks in an attempt to disperse the smoke before the plume reached the ground. An obvious shortcoming with the latter maneuver was that it is, in essence, transporting the problem to someone else's backyard. It was found that during severe meteorological stagnation, the smoke did not make it very far. A final alternative was to relocate major sources. Early regulations, however, did not deal with the smoke from home heating.

Modern techniques to deal with London-type smog include refinements to those mentioned above, including building taller stacks and using natural gas or oil instead of coal. More efficient combustors were developed to reduce fuel consumption and emissions. Other measures used today include particle collection using cyclones, filters, and electrostatic precipitators. SO_2 is removed by scrubbing and coal cleaning. The theory behind these devices is discussed in Sec. 5.3.

2.2.3.2 Photochemical Smog: An Urban and Regional Phenomenon.

Urban Smog. In the mid-20th century, scientists noted injury to crops being grown downwind of Los Angeles, CA, and realized the injury was due to the air coming from the city. Haagen-Smit and co-workers went on to establish that this type of smog was different from the type found in London, and that the injury was due to gaseous oxidizing agents. Ozone and other oxidants were being formed by the interaction of reactive organic gases (ROG) and nitrogen oxides (NO_x) in the presence of sunlight. The products of these reactions were attacking the plants. Overall, the reaction can be summarized as

$$\mathrm{ROG} + \mathrm{NO}_x + \text{sunlight} + \text{time} \rightarrow \begin{cases} \mathrm{O_3},\ \mathrm{PAN},\ \mathrm{HCHO}, \\ \mathrm{HNO_3},\ \text{particulate matter} \\ \text{and other products.} \end{cases}$$

In addition to plant injury, the pollutants irritate the eyes and lungs and create the well-known haze. The haze is due to the particulate matter emitted directly by sources like diesel vehicles and to particles formed by the photochemical reactions. Unlike the London case, the haze is not a result of combining smoke and fog, though smog is still the most common term for it.

Photochemical smog is a soup of many gaseous and particulate compounds. The more prevalent pollutants include CO, ozone, SO_2, NO_2, NO, formaldehyde, other organics, peroxyacetyl nitrate (PAN), nitric acid, nitrous acid, soot, and particulate nitrates, sulfates, and ammonium (Watson *et al.*, 1985; Finlayson-Pitts and Pitts, 1987). Some of these pollutants are emitted directly, such as NO, CO, and SO_2, but many of the oxidants, acids, and oxygenated organics are formed primarily by gas phase reaction, which is explained in the section on atmospheric chemistry.

While photochemical air pollution is often associated with Los Angeles, virtually all major cities experience its effects. Reactive organic gases are emitted by a variety of sources including factories, solvent usage, and automobiles as well as by natural sources such as trees. Nitrogen oxides are emitted by virtually all combustion-powered devices, including automobiles, electricity generating facilities, and manufacturing plants. With sunlight, reactions proceed to form smog. Photochemical smog first being identified in Los Angeles is not surprising. The area is known for its sunlight and the large number of cars. In addition, Los Angeles experiences multiday periods of meteorological stagnation with a strong temperature inversion (as discussed in more detail in Sec. 3, stagnation is a period of minimal pollutant dilution). Smog precursors, in essence, react over multiple days to form high levels of ozone. As an

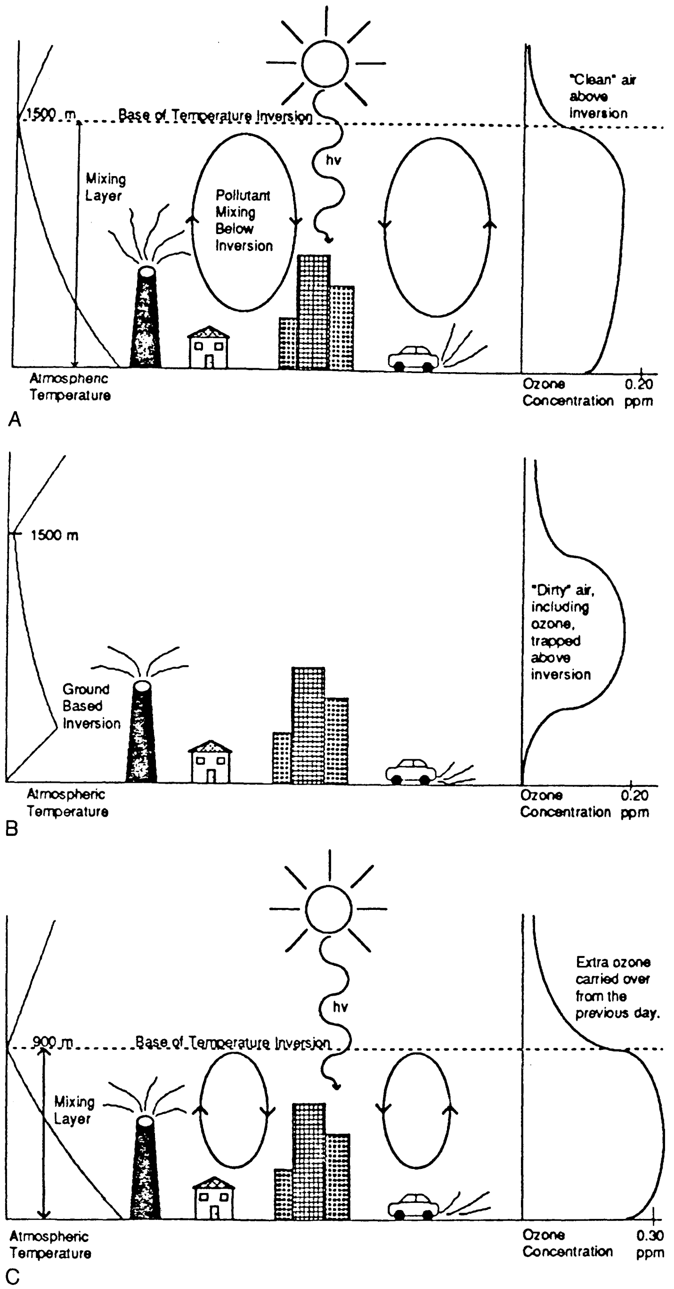
1500 m
Base of Temperature Inversion
Mixing Layer
Pollutant Mixing Below Inversion
hv
"Clean" air above Inversion
Atmospheric Temperature
Ozone Concentration
0.20
ppm
A
1500 m
Ground Based Inversion
"Dirty" air, including ozone, trapped above inversion
Atmospheric Temperature
Ozone Concentration
0.20
ppm
B
900 m
Base of Temperature Inversion
hv
Extra ozone carried over from the previous day.
Mixing Layer
Atmospheric Temperature
Ozone Concentration
0.30
ppm
C

FIG. 1. Evolution of a photochemical smog episode. Photochemical smog intensifies over a number of days. On the first day (a), the inversion is relatively high, confining pollutant emissions to a relatively large mixing volume and allowing dispersion. Sunlight driven reactions form moderate levels of ozone. (Slightly less than 0.2 ppm is shown, but levels vary dramatically depending on location.) During the day the pollutants are relatively uniformly distributed throughout the mixing layer. At night (b), ground cooling forms a ground-based inversion and inhibits pollutant mixing. Pollutants emitted near the ground are trapped, and the ozone aloft is separated from fresh emissions and deposition, allowing the ozone to remain overnight. The next day (c), sunlight again drives chemical reactions, producing more ozone, and the ozone aloft mixes down, intensifying the problem. Also, the inversion depth can be lower, decreasing pollutant dispersion. These events lead to higher ozone levels. This process can repeat over multiple days.

indication of the time delay required for the reactions to occur, the highest ozone levels are seldom found on Sunday or Monday because of reduced emissions over the weekend. The evolution of photochemical episodes is shown schematically in Fig. 1, and the chemistry is discussed in Sec. 4.

Regional Ozone. Though at somewhat lower levels than in cities, elevated ozone concentrations are now found far downwind of cities. This is due to the transport of ozone and reactive nitrogen compounds such as NO_2 and PAN. Organic gas emissions, such as isoprene and monoterpenes from trees, react

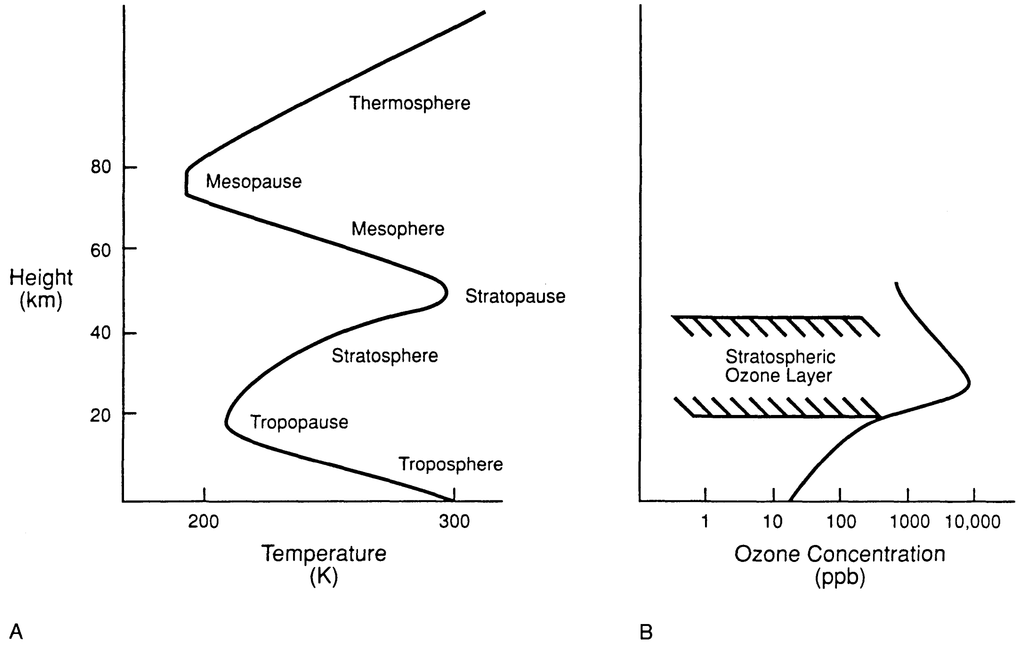

FIG. 2. Vertical (a) temperature and (b) ozone distributions in the atmosphere. Temperatures in the troposphere generally decrease with height, allowing for relatively rapid mixing. In the stratosphere, the temperature increases with height due to solar heating. This leads to a very stable atmosphere and slow transport. Ozone concentration peaks about 10 ppmv at 30–35 km. (Adapted from Anthes *et al.*, 1978; and Whitten and Prasad, 1985.)

in the presence of the nitrogen compounds to form ozone. Even though the pollutant levels are lower than in the cities, the persistence of ozone leads to forest decline as seen in the eastern United States, Canada, and Europe.

2.2.3.3 Acid Deposition. Acid deposition is the name given to the phenomena of the transfer of acidic components in the atmosphere to the earth. While commonly called "acid rain," that term is only half right. Much of the acid is deposited dry as sulfuric acid particles and nitric acid gas. While many areas are subject to acid deposition, eastern North America, China, Scandinavia, and Germany suffer the greatest impact. This is because those areas are downwind of large SO_2 and NO_x sources (e.g., power plants and shelters) and are least resilient to an acidic influx. These soils and water bodies have a smaller buffering capacity than other regions.

The increase in soil and water acidity upsets native ecosystems by the hydrogen ion stressing the biota and by leaching and mobilization of metals. This has resulted in dramatic declines in game fish populations and, in some cases, lakes dying. About 10 000 lakes in eastern Canada and 1000 in the northeastern United States have been acidified to the point that the endogenous fish population has been destroyed. Mobilization of the metals in the soils attacks plant life. Attempts have been made to reverse the trends by adding lime to the affected lakes. Long-term solutions rely on lowering emissions of SO_2 and NO_x in upwind areas.

2.2.4 Global Air Pollution. The 1970s and 80s witnessed a growing awareness of global atmospheric changes in response to man-made pollutants. It was recognized near the turn of the century that increased CO_2 could lead to a warmer earth. Our ability to identify and quantify the magnitude of those changes is an increasing concern. The second global threat is depletion of the stratospheric ozone layer (Fig. 2). The two problems are linked. Both are caused, at least in part, by chlorofluorocarbon emissions. Also, global tropospheric warming may increase stratospheric ozone depletion.

2.2.4.1 Global Warming. Atmospheric temperature is determined by a balance between incoming solar radiation and energy radiated by the earth's surface (Schneider,

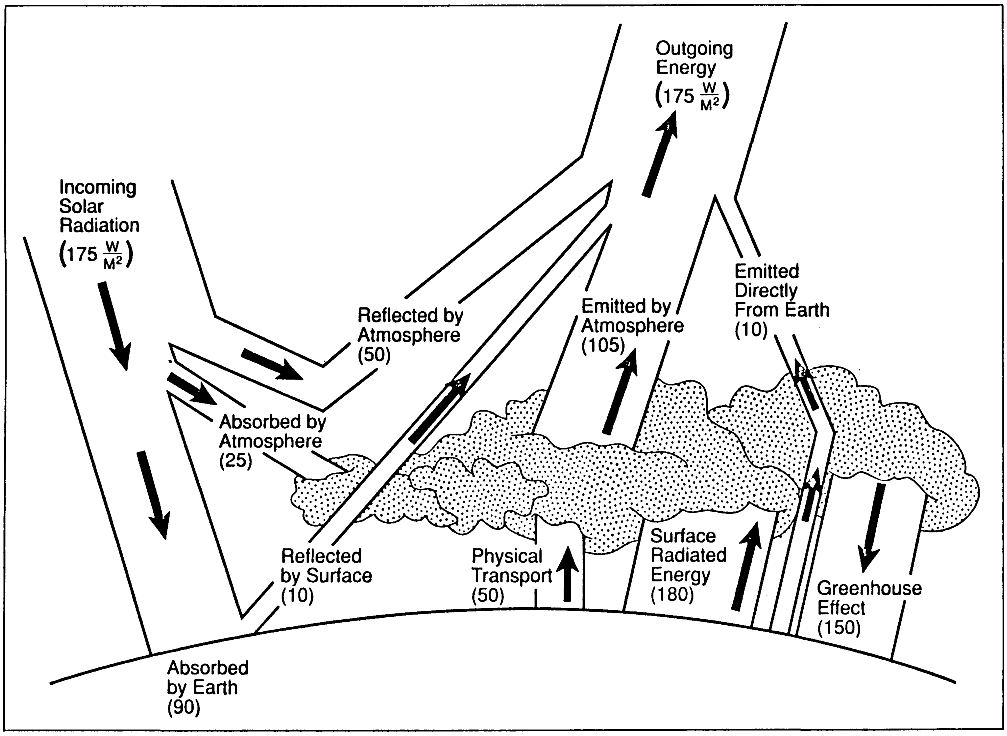
Outgoing
Energy
(175 $\frac{W}{M^2}$)
Incoming
Solar
Radiation
(175 $\frac{W}{M^2}$)
Reflected by
Atmosphere
(50)
Emitted by
Atmosphere
(105)
Emitted
Directly
From Earth
(10)
Absorbed by
Atmosphere
(25)
Reflected
by Surface
(10)
Physical
Transport
(50)
Surface
Radiated
Energy
(180)
Greenhouse
Effect
(150)
Absorbed
by Earth
(90)

FIG. 3. Radiative energy budget of the earth and atmosphere. Incoming solar radiative energy is approximately 175 W/m^2, and is balanced by energy reflected or emitted back into space. Approximately two-thirds of the incoming solar radiation is absorbed by the earth or atmosphere. Greenhouse warming occurs when the atmosphere absorbs radiant energy from the earth before the light can escape into space, and acts as a blanket. Trace cases, such as CO_2, increase the absorption by the atmosphere.

1989). The sun emits nearly blackbody radiation at a temperature of 6000 K, with most of the energy concentrated in the visible and near visible region. Only about 25% of the sunlight penetrates directly to the earth's surface and is absorbed. About 35% is reflected, 15% is absorbed in the atmosphere, and 25% is scattered by gas molecules and particles before being absorbed by the surface. Of the 50% absorbed, most is reradiated at an average temperature of 288 K, corresponding to light with wavelengths in the infrared range between 1 and 30 μm. Much of this energy (90%) is absorbed by water vapor, CO_2, and other trace gases in the atmosphere. This absorbed energy warms the atmosphere, and is reradiated both to space and back to the earth. This is shown, schematically, in Fig. 3. Energy being radiated from the earth's surface with wavelengths between 8 and 11 μm escapes most easily because H_2O and CO_2 do not absorb readily in this region. This region is referred to as the atmospheric window. The absorption of the earth's radiation by the atmosphere acts as a blanket and leads to an increased temperature of about 35 °C.

Increasing tropospheric concentrations of water vapor, CO_2, or gases that absorb infrared light (especially between 8 and 11 μm) would tend to increase the radiation blanket and tropospheric temperatures. Carbon dioxide was the first gas cited that might cause global warming. Industrialization has led to increased emissions of CO_2 because the necessary energy has required fossil fuel burning. A second contributor to CO_2 buildup may be deforestation. Trees metabolize CO_2 to oxygen and fixed carbon. CO_2 is also removed by oceanic absorption. Analysis of polar ice data and recent measurements indicate that atmospheric CO_2 has increased from 290 to 350 ppm since 1900.

Other gases, such as methane, N_2O, and CFCs, add to the greenhouse effect, though

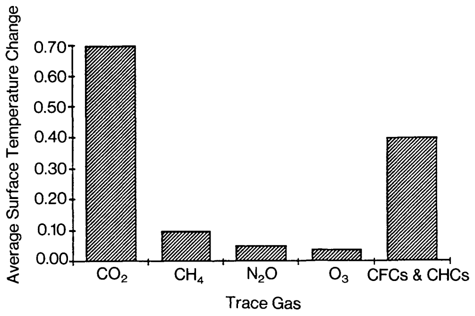

FIG. 4. Impact of trace gases on global warming. The increase in carbon dioxide is expected to be the largest contributor to global warming. Other trace gases combined equal CO_2, though the impact of the CFCs has a large uncertainty and may be much greater than shown here (modified from Ramanathan *et al.*, 1985).

they are found at much lower concentrations than H_2O and CO_2. They absorb light more efficiently at wavelengths other than the regions where CO_2 and H_2O operate, such as between 8 and 11 μm. This intensifies their effect, even though they are found at very low concentrations. Figure 4 shows the relative importance of various gases to the greenhouse effect as compared to CO_2 (Schneider, 1989).

It is at this time impossible to determine the outcome of increasing levels of CO_2 and other absorbing trace gases. The energy balance will certainly be perturbed, and the most advanced models predict an average warming of about 4 K corresponding to a doubling in CO_2. The change in temperature is expected to be greater near the poles, and some areas may see little change. The difficulty in predicting the actual outcome results from not fully understanding a variety of processes, such as how effectively the ocean will take up CO_2 and absorb heat, and how global warming will affect weather such as the formation of clouds and alteration of air currents. Increased cloud cover would reflect solar radiation which would lower the energy flux to the earth. Along with the increases in temperature, models predict changes in rainfall distribution and soil moisture.

Methane concentrations are rising faster than is currently explicable. Sources of the excess methane include termites and cultivation of rice paddies. A third possible cause cited for the rise is that its lifetime may be longer than has been estimated or there has been an increase in its atmospheric lifetime. Methane is removed from the troposphere by reacting with OH. The estimated lifetime, based on OH radical levels, is 7 yr, although it is less than 1 yr in the lower troposphere where OH radical levels are greater. If OH were being depleted by another compound, this could cause methane to build up.

Chlorofluorocarbons are used in spray cans (as a propellant), in refrigeration systems, for cleaning electronics instruments, and in production of polyurethane foams. In the early 1970s, the primary source was spray cans (Gribbin, 1988). Concern over ozone depletion reduced their use in spray cans, and now the major sources include plastic manufacture and automotive air conditioning systems (Fig. 5). While the emissions of CFCs from the most industrialized nations may decrease, emissions from developing nations are expected to offset these improvements.

N_2O is emitted by microbial activity, excess fertilizer use, and combustion. The importance of combustion as a major source of N_2O is being debated because emissions may be much less than assumed.

2.2.4.2 Ozone Depletion. Depletion of the stratospheric ozone layer over the Antarctic was first measured in the mid 1970s. By the early 1980s the loss was so large and so rapid that satellite measurements of the ozone were so low they were taken to be in error. In 1987, October ozone levels over the Antarctic fell to less than half of those found in 1974. This rapid loss of ozone has been found to be a result of catalytic cycles involving halogens that have slowly diffused into the lower stratosphere.

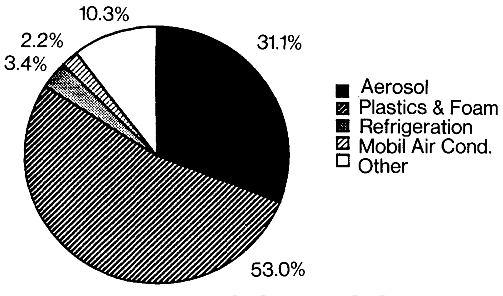

FIG. 5. Estimated use of CFC-11 and CFC-12 by product in 1984. Approximately 3.4×10^9 kg were produced by countries reporting to the Chemical Manufacturers Association.

Most rapid loss of stratospheric ozone is found over the Antarctic during its spring. This is a result of meteorological isolation of the stratosphere, cold temperatures, and the presence of chlorine and bromine atoms. Chlorine and bromine atoms promote the chemical destruction of ozone, as discussed in the section on stratospheric chemistry. The very cold temperatures allow the formation of polar stratospheric ice clouds (PSCs), which in turn provide a surface to enhance the production of reactive chlorine atoms. Meteorological isolation inhibits the transport of new ozone or warmer air to the ozone "hole" over the Antarctic. Similar dynamics can lead to ozone loss over the Arctic, but stratospheric temperatures are higher, decreasing the formation of PSCs.

Ozone loss over the mid latitudes is less pronounced than over the Antarctic because of the lack of PSCs, though a decreasing trend is noted. From 1979 to 1987, there was about a 2%–10% decrease in mid-latitude stratospheric ozone. There are uncertainties in the calculations of the response of the stratosphere to further chlorine input, but it is generally agreed that ozone levels will continue to decrease as more chlorine makes its way into the stratosphere. Increased absorption of light in the troposphere would decrease light transmission to the stratosphere. Thus, tropospheric warming could be accompanied by stratospheric cooling. A colder stratosphere would lead to increased PSC formation and faster ozone destruction. This links the greenhouse effect with ozone depletion.

The destructive chlorine comes from the slow diffusion of chlorofluorocarbons from the troposphere. These compounds are inert until they reach the stratosphere and the strong ultraviolet radiation breaks the molecules apart, releasing chlorine. There are few replacements for the many uses of CFCs, and their emissions will likely keep increasing (Fig. 5).

3. AIR POLLUTION METEOROLOGY AND POLLUTANT TRANSPORT

Air pollutants are transported in the atmosphere by advection, due to the bulk wind, and dispersion, which arises from atmospheric turbulence. Also associated with atmospheric transport is deposition, involving a complex series of interactions between the local turbulence structure of the atmosphere, physical surface characteristics, and the chemical capability of the surface to absorb pollutants. Along with chemical transformation, transport determines the ultimate fate and impact of pollutant emissions.

Pollutants, except for large particles, are effectively carried by local winds. In addition, pollutants are diffused vertically and, to a more limited extent, horizontally by turbulent dispersion. The region where vertical diffusion is rapid is called the mixed layer (or mixing layer) (Anthes *et al.*, 1978). The mixed layer is usually capped by a temperature inversion, described below. Atmospheric turbulence is caused by solar heating of the ground and mechanical generation by surface friction. Heating the ground leads to warm air generation at the surface. The warm air rises and is replaced by cooler air. Thermal mixing on a sunny day is generally greater than mechanically generated turbulence.

Air heated at the ground will continue to rise until it comes into equilibrium with the local atmospheric temperature. As it rises, it cools due to adiabatic expansion. This is shown in Fig. 6 and is called the adiabatic lapse rate. An atmosphere with a temperature structure that drops faster than adiabatically with altitude is unstable because a warm air parcel will continue to rise. Also, a cold air pocket will tend to fall. If the temperature rises with altitude faster than the adiabatic lapse rate, vertical exchange of air parcels is inhibited, and turbulent mixing is suppressed. This is referred to as a stable atmosphere. If the atmosphere cools vertically at the adiabatic lapse rate, it is called a neutral atmosphere.

During the morning and early afternoon of a sunny day, ground-level warming leads to an unstable atmosphere. The thermally generated turbulence rapidly disperses pollutants. If the vertical extent of the turbulent mixing is large, pollutants do not build up to very high levels. A feature commonly associated with urban air pollution is the development of a temperature inversion. In this case, the atmosphere at some height above the ground is stable, reducing vertical mixing and trapping pollutants near the ground. An inversion may develop due to heating of the atmosphere on previous days, or cloud cover

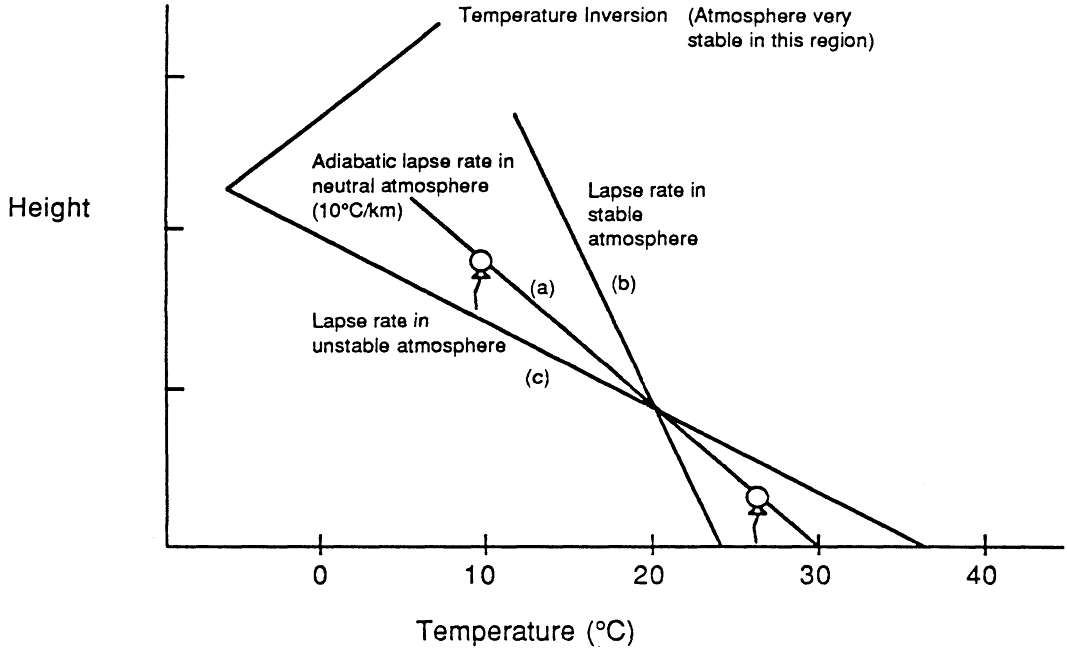
Temperature Inversion (Atmosphere very stable in this region)
Adiabatic lapse rate in neutral atmosphere (10°C/km)
Lapse rate in stable atmosphere
(a)
(b)
Lapse rate in unstable atmosphere
(c)
Height
0
10
20
30
40
Temperature (°C)

FIG. 6. An illustration of atmospheric stability. In a neutral atmosphere (a), the temperature decreases with height at the same rate that an air parcel cools due to adiabatic expansion. For example, a perfect balloon released from earth would expand as it rose because the pressure decreases with height. The expansion process leads to the gas cooling. Neutral atmospheres neither promote nor inhibit mixing. In a stable atmosphere (b), the temperature decrease with height is less than the adiabatic lapse rate. As an air parcel (or balloon) rises, it becomes colder and denser than the surrounding air and sinks back down. Conversely, if the temperature decreases faster than adiabatically (c), a rising air parcel remains warmer than the surrounding atmosphere and will continue to rise. This is an unstable atmosphere. As shown, the unstable atmosphere is capped by a region whose temperature starts to increase, or is "inverted." This region is called an inversion and is very stable, capping the dispersion of pollutants below.

that absorbs sunlight and insulates the surface from heating. Inversions may start at the ground or up to a few kilometers above. When accompanied by low wind speed, pollutant dilution is minimal, and over a period of days leads to smog episodes such as found in Los Angeles and London. Periods of light winds and strong inversion are called stagnation episodes. Low dilution can be limited to a relatively small area such as mountains, valleys, or urban basins, or may encompass large regions.

Transport of pollutants between the troposphere and stratosphere is slow because of the stability of the air above the troposphere (Fig. 2). Transport within the stratosphere is also reduced for this reason. Pollutants emitted at the surface take several years to reach the stratosphere. Except for the most stable compounds, chemical reaction and deposition (by rain-out or dry deposition) remove the pollutants before they pass through the troposphere. Because the transport to the stratosphere and deposition is so slow, virtually all the CFCs emitted are still present in the troposphere. Ozone depletion is a result of the small fraction that has passed to the lower levels of the stratosphere.

Deposition is the ultimate fate of the material emitted into the air. It is a two-part process. First, the compound is transported to a surface. Then the compound interacts chemically. Pollutants that are highly reactive, such as nitric acid, deposit rapidly. Less reactive pollutants deposit more slowly, allowing those compounds to be transported much further. In the case of ozone, this increases the extent of the exposure from urban to rural regions. Extremely unreactive compounds such as CO_2, methane, and CFCs can have very long lifetimes and build up across the globe. Deposition can be enhanced by rain and cloud droplets scavenging pollutants and then falling to earth.

4. ATMOSPHERIC CHEMISTRY

Smog, acid deposition, and stratospheric ozone depletion are a result of atmospheric chemical reactions. Primary pollutants react to alter the chemical composition of the air. These reactions may occur homogeneously in the gas or liquid phase, or heterogeneously, involving transport of gaseous pollutants to or from droplets and ice particles. This section shows how pollutants chemically evolve in the troposphere and the chemistry leading to ozone depletion.

4.1 Tropospheric Chemistry

4.1.1 Urban Photochemistry. There are two systems involved in the production of ozone and other photochemical oxidants. First is the photolytic cycle forming ozone, and the second is the oxidation of organics, leading to enhanced formation of NO_2. The two are tightly linked by the production and cycling of highly reactive radical species.

The single reaction responsible for virtually all ozone formation in the lower troposphere is the photolysis of NO_2 (see Finlayson-Pitts and Pitts, 1987):

$$NO_2 + h\nu \xrightarrow{k_1} NO + O^\bullet, \qquad \text{(R1)}$$

where $h\nu$ represents energy derived by interaction with photons from the sun. $O^\bullet$, a radical oxygen atom, rapidly reacts with molecular oxygen to form ozone,

$$O^\bullet + O_2 \xrightarrow{k_2} O_3. \qquad \text{(R2)}$$

The two reactions can be combined, giving

$$NO_2 + O_2 + h\nu \rightarrow NO + O_3.$$

The ozone can then react with NO,

$$NO + O_3 \xrightarrow{k_3} NO_2 + O_2. \qquad \text{(R3)}$$

These three reactions are the photolytic cycle, and summing leads to no net production of ozone. The amount of ozone predicted to be present due to these reactions alone is given by

$$[O_3] = \frac{k_1[NO_2]}{k_3[NO]}.$$

Approximately 95% of the NO_x is emitted as NO, and 5% as NO_2. Using this, and values of 0.007 s^{-1} and 0.5 s^{-1} for the reaction rates k_1 and k_3, respectively, leads to a predicted ozone concentration of 7×10^{-4} ppm—three orders of magnitude less than is observed.

Reactions of organic gases provide the necessary components to produce the ozone. For example, formaldehyde (and other aldehydes) readily photolyze in sunlight:

$$HCHO + h\nu \rightarrow CO + 2HO_2, \qquad \text{(R4)}$$

$$CH_3CHO + h\nu \rightarrow CH_3^\bullet + HCO^\bullet, \qquad \text{(R5)}$$

(oxygen molecules are in relative abundance, and their consumption in reactions often is not shown for brevity). The $CH_3^\bullet$ and $HCO^\bullet$ further decompose in the presence of oxygen and NO:

$$CH_3^\bullet + 2O_2 + NO \rightarrow NO_2 + HCHO + HO_2^\bullet, \qquad \text{(R6)}$$

$$HCO^\bullet + O_2 \rightarrow HO_2^\bullet + CO, \qquad \text{(R7)}$$

where reaction (R6) is the condensation of multiple reactions. The HO_2 produced will oxidize NO to NO_2:

$$NO + HO_2^\bullet \rightarrow NO_2 + OH^\bullet. \qquad \text{(R8)}$$

The NO_2 produced is then available to form ozone. Summing up reactions (R4) and (R8), and reactions (R5)–(R8) gives (neglecting oxygen molecules)

$$HCHO + h\nu + 2NO \rightarrow CO + 2NO_2 + 2OH^\bullet,$$

$$CH_3CHO + h\nu + 5NO \rightarrow 2CO + 5NO_2 + 4OH^\bullet,$$

and repeating reaction (R1)+(R2),

$$NO_2 + h\nu \rightarrow O_3 + NO.$$

Thus, the photolysis of one molecule of acetaldehyde, by this scheme, could produce five extra ozone molecules. There are, however, insufficient quantities of aldehyde emissions to form appreciable amounts of ozone, and most urban organic gases, such as alkanes and aromatics, do not photolyze. Additional chemistry is needed to overcome the shortfall.

In the late 1960s it was discovered that similar to combustion systems, hydroxyl radical attack accelerates the oxidation of organic gases such as alkanes and, simultaneously, the production of ozone. The photochemistry of propane ($CH_3CH_2CH_3$) provides a useful example:

$$CH_3CH_2CH_3+OH^\bullet \rightarrow CH_3CHCH_3+H_2O, \quad \text{(R9a)}$$

$$CH_3CHCH_3+O_2 \rightarrow CH_3CH(O_2^\bullet)CH_3 \text{ (peroxyalkyl radical)}, \quad \text{(R9b)}$$

$$CH_3CH(O_2^\bullet)CH_3+NO \rightarrow NO_2+CH_3CH(O^\bullet)CH_3 \text{ (alkyl radical)}, \quad \text{(R9c)}$$

$$CH_3CH(O^\bullet)CH_3+O_2 \rightarrow CH_3C(O)CH_3 \text{ (acetone)}+HO_2^\bullet; \quad \text{(R9d)}$$

the net reaction can be written as

$$CH_3CH_2CH_3+OH^\bullet+NO+2O_2 \rightarrow CH_3C(O)CH_3+NO_2+H_2O+HO_2^\bullet. \quad \text{(R9e)}$$

The OH is regenerated by reaction (R8). A net of two NO to NO_2 oxidations take place. The acetone produced further participates in photochemical reactions, increasing the oxidation rate of NO to NO_2 and, ultimately, ozone formation. The cycle can be summarized, rather simply, as

$$ROG+NO_x+h\nu+\text{time} \rightarrow O_3.$$

The $OH^\bullet$ is not shown because it is cycled to its original form. It is, however, quite necessary for the production of ozone.

The hydroxyl radical needed to initiate the chemistry above is formed in the atmosphere by reaction (R8), the photolysis of ozone, and the dissociation of nitrous acid:

$$O_3+h\nu \rightarrow O_2+O^\bullet(^1D) \text{ ("O singlet D" radical)}, \quad \text{(R10a)}$$

$$O^\bullet(^1D)+H_2O \rightarrow 2OH^\bullet, \quad \text{(R10b)}$$

$$HONO+h\nu \rightarrow OH^\bullet+NO. \quad \text{(R11)}$$

These reactions, along with reaction (R1), show the necessary role that sunlight plays, providing the critical energy to initiate the reaction sequence.

Nitric acid, peroxyacetyl nitrate, hydrogen peroxide (H_2O_2), and formaldehyde are also major products of photochemical reactions. Nitric acid is formed by daytime and nighttime mechanisms. During daylight, hydroxyl radicals oxidize NO_2:

$$NO_2+OH^\bullet \rightarrow HNO_3. \quad \text{(R12)}$$

This is an important sink for both NO_x and hydroxyl radical. Very high levels of NO_x tend to slow the photochemical system by absorbing OH radical. At night, NO_2 and ozone combine to form NO_3, which further reacts with another NO_2 molecule:

$$NO_2+O_3 \rightarrow NO_3+O_2, \quad \text{(R13)}$$

$$NO_3+NO_2 \rightarrow N_2O_5. \quad \text{(R14)}$$

During the daylight hours NO_3 photolyzes and reacts rapidly with NO. The N_2O_5 then hydrolyzes (either by homogeneous, gas phase reaction or on the surface of a particle):

$$N_2O_5+H_2O \rightarrow 2HNO_3. \quad \text{(R15)}$$

PAN and other organic nitrates are produced by peroxyalkyl and alkyl radical reactions with NO and NO_2; for example,

$$RCO_3+NO_2 \rightarrow PAN, \quad \text{(R16)}$$

$$RO_2+NO \rightarrow RNO_3, \quad \text{(R17)}$$

where R is a general alkyl group (e.g., CH_3CH_2–). Hydrogen peroxide is formed by the self-reaction of HO_2 radical:

$$HO_2^\bullet+HO_2^\bullet \rightarrow H_2O_2, \quad \text{(R18)}$$

acting as a radical sink from the system. While photochemical reactions do produce high levels of harmful pollutants, it should be recognized that they also cleanse the atmosphere from the buildup of NO_x, CO, and hydrocarbons.

4.1.2 Regional Gas Phase Photochemistry and Acid Formation. The most notable differences between urban and regional photochemistry are the importance of slower reactions in regional chemistry, the increased concentrations of natural pollutants (e.g., isoprene and other organics), and the relatively lower levels of pollutants, especially NO_x (see Table 1). NO_x and organics provide the fuel to drive the photochemistry and the hydroxyl radical still sustains the reactions. The hydroxyl radical is also responsible for the formation of atmospheric acids (Table 2). It was shown above that NO_x is required to cycle $HO_2^\bullet$ to $OH^\bullet$, and to photolyze to form ozone. In rural atmospheres, NO_x is found at levels two to three orders of magnitude less than in urban areas. This leads to rural environments being NO_x-limited, and to increases in oxi-

Table 2. Reaction pathways forming acids over regional scales.

	Reaction
Nitric acid	
Gas phase	
$NO+O_3 \rightarrow NO_2$	(R3)
$NO_2+OH \rightarrow HNO_3$	(R12)
Heterogeneous	
$2NO_2+O_3 \rightarrow N_2O_5$	(R13+R14)
$N_2O_5+H_2O \rightarrow 2HNO_3$	(R15)
Sulfuric acid	
Gas phase	
$SO_2+OH \rightarrow HOSO_2$	(R21a)
$HOSO_2+O_2 \rightarrow SO_3+HO_2$	(R21b)
$SO_3+H_2O \rightarrow H_2SO_4$	(R21c)
Net $SO_2+OH \rightarrow H_2SO_4+HO_2$	
Aqueous phase	
$SO_2+H_2O \leftrightharpoons HSO_3^-+H^+$	(R22a)
$HSO_3^-+H_2O_2 \rightarrow SO_4^{2-}+H^++H_2O$	(R22b)
Net $SO_2+H_2O_2 \rightarrow SO_4^{2-}+2H^+$	
Organic acids	
$HCHO+HO_2 \rightarrow OOCH_2OH$	
$OOCH_2OH+NO \rightarrow OCH_2OH+NO_2$	
$OCH_2OH+O_2 \rightarrow HO_2+HCOOH$ (formic acid)	

dized radicals such as $HO_2^\bullet$ and $RO_2^\bullet$ (i.e., general peroxyalkyl radicals formed by the oxidation organics such as propane shown above).

The buildup of HO_2 and RO_2 leads to the increased formation of hydrogen peroxide [reaction (R18)] and organic peroxides:

$$RO_2+HO_2^\bullet \rightarrow ROOH \text{ (general organic peroxide)}, \qquad \text{(R19)}$$

both of which are strong aqueous oxidants and act to convert SO_2 to sulfuric acid in cloud, rain, and fog droplets.

On a regional scale, the ozone present is due to formation from local emissions and the transport of ozone and its precursors from urban and industrial sources. Ozone can be transported hundreds of miles (Finlayson-Pitts and Pitts, 1987). NO_x from mobile sources, power plants, and other combustion processes is also transported, but is rapidly transformed into nitric acid by reactions (R12) and (R14). The nitric acid produced reacts readily with terrestrial surfaces and deposits out of the photochemical system.

The transport of NO_x into rural atmospheres can be facilitated by PAN. PAN is relatively stable in colder atmospheres, and is not depleted by hydroxyl radical attack. PAN acts as a reservoir of NO_x, releasing NO_2 as local concentrations drop:

$$CH_3CO_3NO_2 \overset{\text{(PAN)}}{\rightleftarrows} CH_3CO_3^\bullet+NO_2. \qquad \text{(R20)}$$

Many organics have significantly longer lifetimes than NO_x, and are transported with less loss (Table 1). Combined with local, natural emissions of isoprene and monoterpenes, and the ubiquitous presence of methane, rural atmospheres have a relative abundance of organic gases compared to NO_x (often characterized as having a high ROG/NO_x ratio). The oxidation of methane and isoterpene also is preceded by $OH^\bullet$ radical attack, forming radical intermediates that transform NO to NO_2. They may both be precursors to forming acid.

Sulfur dioxide reacts with the hydroxyl radical, in the gas phase, to form sulfuric acid [reaction (R21), Table 2]. This is a major pathway for the conversion of anthropogenic SO_2 emissions to sulfuric acid (Schwartz, 1989).

4.1.3 Regional Aqueous Phase Chemistry. Unlike urban photochemical episodes which thrive on sunlight, regional scale chemical systems can be enhanced by the presence of cloud droplets and fog providing alternative, faster paths for producing acids. In conjunction with gas phase photochemistry, these droplets act as tiny chemical reactors.

Sulfuric acid is produced by the oxidation of sulfur dioxide with H_2O_2, as follows. First, SO_2 gas is absorbed into the cloud or fog droplet, where it is at equilibrium:

$$SO_2(g) \rightarrow SO_2(aq) \overset{H_2O}{\rightleftarrows} HSO_3^-+H^+. \qquad \text{(R22)}$$

The HSO_3^- is oxidized to sulfuric acid by hydrogen peroxide, H_2O_2 [Table 2, reaction (R22)]. The necessary H_2O_2 is formed by gas phase photochemistry. Implicit in the reactions is the transport of the gas phase species to the droplet surface (Schwartz, 1989).

4.2 Stratospheric Chemistry

Stratospheric chemistry is significantly different from the photochemistry of the troposphere because of the flux of very high energy ultraviolet light ($\lambda < 400$ nm). Molecules in the stratosphere absorb this energy, allowing relatively little to pass to the troposphere. This does two things. First, it protects surface life from damaging sunlight, and second, it drives the photochemistry of the stratosphere.

When exposed to UV light below 190 nm, oxygen molecules photodissociate into oxygen atoms, forming ozone (Gribbin, 1988; McElroy and Salawitch, 1989)

$$O_2 + h\nu\ (\lambda < 190\ \text{nm}) \rightarrow O^\bullet + O^\bullet, \qquad \text{(R23)}$$

$$O^\bullet + O_2 + M \rightarrow O_3 + M, \qquad \text{(R24)}$$

where M is a third molecule that absorbs energy, allowing the newly formed ozone molecule to be energetically more stable. This also has the effect of heating the third-body molecules, and the stratosphere itself. The ozone, in turn, photolyzes by absorbing light between 230 and 350 nm:

$$O_3 + h\nu\ (230 < \lambda < 350\ \text{nm}) \rightarrow O_2 + O^\bullet. \qquad \text{(R10)}$$

This reaction absorbs most of the incoming solar radiation between 230 and 290 nm, and shields the earth's surface. The oxygen atom can then recombine with an oxygen molecule via reaction (R24), or with O_3:

$$O^\bullet + O_3 \rightarrow 2O_2. \qquad \text{(R25)}$$

Stratospheric NO is found in very trace quantities and can act catalytically to destroy ozone:

$$\begin{array}{ll} NO + O_3 & \rightarrow NO_2 + O_2, \qquad \text{(R3)} \\ NO_2 + O^\bullet & \rightarrow NO + O_2 \\ \hline \text{Net } O_3 + O^\bullet & \rightarrow 2O_2. \end{array} \qquad \text{(R26)}$$

These reactions lead to a dynamic equilibrium level of ozone that responds to the influx of radiation and NO_x level in the stratosphere. Solar activity has been shown to perturb ozone concentrations in the stratosphere.

Chlorofluorocarbons (CFCs) perturb the chemistry of the stratosphere and deplete the ozone layer by adding three more routes for the catalytic destruction of ozone. CFCs, which are inert in the troposphere, slowly diffuse into the stratosphere. In the presence of high levels of UV they dissociate, releasing chlorine:

$$CF_2Cl_2 + h\nu \rightarrow CF_2Cl + Cl^\bullet. \qquad \text{(R27)}$$

(Upon further reaction a second chlorine atom is released.) The chlorine then reacts with ozone:

$$\begin{array}{ll} 2(Cl^\bullet + O_3 & \rightarrow ClO^\bullet + O_2) \\ ClO^\bullet + ClO^\bullet & \rightarrow Cl_2O_2 \\ Cl_2O_2 + h\nu & \rightarrow Cl^\bullet + ClOO^\bullet \\ ClOO^\bullet & \rightarrow Cl^\bullet + O_2 \\ \hline \text{Net } 2O_3 & \rightarrow 3O_2. \end{array} \qquad \text{(R28)}$$

The second cycle has a similar outcome:

$$\begin{array}{ll} Cl^\bullet + O_3 & \rightarrow ClO^\bullet + O_2 \\ OH^\bullet + O_3 & \rightarrow HO_2^\bullet + O_2 \\ ClO^\bullet + HO_2^\bullet & \rightarrow HOCl + O_2 \\ HOCl + h\nu & \rightarrow OH^\bullet + Cl^\bullet \\ \hline \text{Net } 2O_3 & \rightarrow 3O_2. \end{array} \qquad \text{(R29)}$$

A third destructive cycle involves bromine and chlorine:

$$\begin{array}{ll} Cl^\bullet + O_3 & \rightarrow ClO^\bullet + O_2 \\ Br^\bullet + O_3 & \rightarrow BrO^\bullet + O_2 \\ BrO^\bullet + ClO^\bullet & \rightarrow Br^\bullet + Cl^\bullet + O_2 \\ \hline \text{Net } 2O_3 & \rightarrow 3O_2. \end{array} \qquad \text{(R30)}$$

Each of these cycles destroys two ozone molecules. The chlorine can be recycled thousands of times before being stabilized as HCl and chlorine nitrate. These reactions, as written, are homogeneous, gas phase reactions that can lead to ozone loss throughout the stratosphere. Heterogeneous reactions assist the more rapid depletion of ozone over the Antarctic by increasing the availability of free chlorine.

Reactions which proceed negligibly slowly in the gas phase can occur rapidly in the presence of an ice surface or in aqueous solution. During the austral winter, extensive ice clouds (polar stratospheric clouds, PSCs) form in the stratosphere over the Antarctic, providing a heterogeneous system. For example, two reactions releasing chlorine are

$$ClONO_2(\text{surface}) + HCl(\text{solid solution}) \rightarrow Cl_2(\text{gas}) + HNO_3(\text{surface}) \qquad \text{(R31)}$$

and

$$ClONO_2(\text{surface})H + H_2O(\text{ice}) \rightarrow HOCl(\text{gas}) + HNO_3(\text{surface}). \quad (R32)$$

Besides releasing chlorine, these reactions also trap nitric acid. Nitric acid scavenges OH, slowing reaction (R12), and the NO_x released by that reaction binds free chlorine as $ClONO_2$. The heterogeneous reactions then drive the depletion of O_3 from the Antarctic stratosphere more rapidly than is found over the rest of the globe.

At the time of this writing (1989), the relative importance of the reaction paths is being debated, and heterogeneous mechanisms are still being explored. There is general agreement that CFCs are primarily responsible for ozone depletion and that polar ice clouds accelerate depletion.

5. AIR POLLUTION CONTROL

Air pollution control is a three-step process. First, the acceptable air quality levels are stipulated via regulatory action. Then, source(s) of the pollutant emissions must be identified, along with how they affect air quality, i.e., developing source-to-air-quality relationships. Next, means for controlling those emissions are identified. Often, more than one type of control device might be suitable and more than one source may be responsible; so, finally, a strategy (or strategies) for implementing controls is constructed. It is desirable to compare the competing strategies and control alternatives in order to derive an optimal control strategy. Optimal may refer to least cost, greatest air quality improvement, highest probability for success, or more likely some combination of the three criteria.

5.1 Air Quality Regulations

Most developed countries have addressed problems of deteriorating air quality within their boundaries by setting air quality standards, emission standards, or both. An air quality standard mandates attaining a prescribed level of atmospheric cleanness, i.e., pollutant concentrations are to be at or below a specified level. An emission standard differs in that the mass (or volume) efflux from a source is limited to be below a specified value, or a certain type of control must be used. Often, the different approaches are not used independently, i.e., only those areas that do not meet air-quality-based standards are subjected to emission-control-based requirements.

Royal proclamations in England during the 13th century were some of the earliest regulatory actions to improve air quality (Brimblecombe, 1984). These edicts limited the use of imported "sea" coal which burned less cleanly than wood or charcoal. Later edicts and commentaries from the residents of London indicate that those actions were not entirely effective. The problem intensified through the years, and further measures were taken with varying results. Finally, the Great Fog of 1952 led to Britain's Clean Air Act of 1956 bringing about emission reductions and averting future events of similar magnitude. A notable aspect of the act was control of emissions from personal residences.

Air pollution control in the United States developed initially through local ordinances limiting smoke emissions (Bibbero and Young, 1974; Stern *et al.*, 1984). However, local control of air pollution in a mobile society did not prove entirely effective, and federal action was taken with the Clean Air Act of 1963, revisions to the act in 1970 and 1977, and the establishment of the Environmental Protection Agency (EPA). The amendments established national air quality standards and emission levels from certain sources, most notably automobiles. Canada similarly passed a Clean Air Act, setting ambient concentration standards for pollutants. Air quality standards for a variety of pollutants in a number of countries are listed in Table 3. Meeting the standards listed, however, has in many cases proved politically infeasible, and economic growth has been chosen over compliance.

In addition to air quality standards, the United States also set emission standards affecting all motor vehicles. Increasingly stringent standards call for decreasing emissions of reactive organic gases (ROG), CO, NO_x, and particulate matter. These regulations have forced the development of more efficient automotive emission controls and the use of catalysts on cars sold in the United States. In Europe, use of the same catalysts would require cooperation by the countries to provide un-

Table 3. Sample air quality standards and recommendations in several countries. Times given in parentheses are averaging periods.

	Pollutant standard (μg m^{-3})			
Country	CO	NO_x	O_3	SO_2
Austria	38 900 (1 h)			500 ($\frac{1}{2}$ h)
Canada	35 000 (1 h)	400 (NO_2) (1 h)	160 (oxidant) (1 h)	900 (1 h)
Japan	23 000 (8 h)	100 (24 h)	110 (1 h)	
USSR	3000 ($\frac{1}{2}$ h)	85 ($\frac{1}{2}$ h)		500 ($\frac{1}{2}$ h)
United States	40 000 (1 h)	100 (1 yr)	235 (1 h)	365 (24 h)

leaded fuel, or the devices would become ineffective. Instead, the European Council of Environmental Ministers has set vehicle emission limits that can be met without NO_x-reduction catalysts.

Pollutant problems that involve transboundary transport of contaminants have been difficult to deal with from a regulatory basis. In a few instances, countries have taken action against or in concert with polluting neighbors. In response to acid deposition, many European countries and Canada banded to form the "30% club," committing themselves to reduce SO_2 emissions 30% from 1980 levels by 1993 (O'Riordan, 1989). A similar agreement, the 1987 Montreal Protocol, calls for a reduction in CFC emissions, and was ratified by over 40 nations. This action was prompted in an attempt to slow down stratospheric ozone depletion. While this treaty does curb the increasing use of CFCs in signatory countries, it is not expected to return the stratospheric ozone to pre-1970 levels; and increased CFC use by developing nations is expected to overwhelm the reductions by those signing the agreement. Further international cooperation is needed to address both stratospheric ozone loss and acid deposition, though the mechanism for such action is uncertain.

5.2 Determining Source–Air-Quality Relationships

Three means are available for determining source-to-air-quality relationships. Field studies are the traditional means and have lately been supported or supplanted by mathematical and physical modeling. Currently, two or three of the techniques usually are used together. Field studies to establish source impacts can use tracers, either tagging a component of the pollutant emissions or emitting the tracer along with pollutants. Perturbing source emissions, chemical fingerprinting, and mass balance techniques have been employed effectively.

One of the most direct ways to establish the impact that a source has on air quality is to remove the source and see what results. However, this is seldom viable. A less drastic approach is to perturb the emissions and follow how the pollutant concentrations respond. Often this can be accomplished passively, as sources such as manufacturing facilities, shelters, and power plants vary emissions due to demand changes. However, changes in meteorological conditions can inhibit detailed source-impact reconciliation.

Many sources typically emit a well characterized assortment of pollutants that can be peculiar to a single type of activity. For example, particulate emissions from an oil-fired boiler contain about 7% vanadium, 6% iron, 5% sodium, 2% nickel, and trace amounts of many other metals. This composition acts as a fingerprint of oil-fired boilers, and if a particulate matter sample is rich in these elements in the same proportions, then the culprit (or one of the culprits) has been identified. When used with regression-type mathematical models, this is a powerful method for quantifying the contribution of sources to air pollutant concentrations. This method is detailed below.

Physical models simulate atmospheric processes with a scaled-down representation of the atmosphere in a laboratory setting. Wind tunnels and scale models of city structures are used to study pollutant transport through street canyons and around buildings. A second physical model is a smog chamber used to study atmospheric chemistry. Reactive gases, such as NO_x and organics, injected into glass chambers and large Teflon bags are exposed to natural sunlight or man-made light. The radiation then drives the chemistry, and the evolution of the pollutants is followed. Because of the use of confined volume, physical models are limited in application.

Currently, mathematical models are the most powerful tool available to establish source-to-air-quality relationships. They can embody our knowledge of the details of atmospheric transport and chemistry and integrate this knowledge into a structured framework. The primary limitations of mathematical models are the knowledge of atmospheric physics and chemistry, and the ability to describe the processes mathematically. Mathematical models are of one of two types: empirical/statistical or deterministic. Empirical models are based on establishing a statistical relationship between previous air quality and the corresponding emissions. Their predictive power is based on the assumption that future air quality will depend on the same set of relationships. This limits their utility to linear processes. One of the more powerful classes of empirical models is receptor-oriented models. By using the chemical fingerprint of source emissions, statistical regression is used to estimate the contribution of a variety of sources. Empirical models have been used very effectively to establish sources of particulate matter using the characteristic fingerprints, as well as nonreactive pollutants such as CO.

Deterministic models describe in a fundamental manner the individual processes that affect the evolution of pollutant concentrations. These models are based on solving the atmospheric diffusion/reaction equation (Seinfeld, 1987):

$$\frac{dc_i}{dt}+\mathbf{U}\cdot\nabla c_i=\nabla\cdot D_i\nabla c_i+R_i(c_1,c_2,c_3,\ldots,c_n)+S_i(\mathbf{x},t),\quad i=1,2,3,\ldots,n, \tag{1}$$

where c_i is the concentration of species i; $\mathbf{U}$ is the instantaneous wind velocity vector; D_i is the scalar molecular diffusivity corresponding to species i; R_i is the net production (depletion if negative) of species i by chemical reaction; S_i is the source rate of species i; and n is the number of species. In essence, this equation states that the time rate of change of pollutant (term 1) depends on convection (term 2), diffusion (3), reaction (4), and emissions (5). Turbulence, which occurs at time and spatial scales smaller than the local scale of the model, is often parametrized by an increased diffusivity rate K, replacing D. Except for the simplest source terms, reaction schemes, and turbulent diffusivity distributions (S, R, and K), there are no analytical solutions to Eq. (1), and the equation is solved computationally.

Within this framework, the ability to characterize accurately the evolution of pollutants is limited by the understanding of the physical processes, especially chemistry and pollutant deposition, and accurate specification of input data. In practice, these models require more data than is commonly available, and gaps must be approximated. For example, the models use wind fields specified at all points within the modeling domain, but the measurements seldom exist in such detail. A second, more fundamental limitation is that the models use volume- and time-averaged quantities, but measurements are point values.

Deterministic models have been used for characterizing and understanding pollutant evolution at all scales. As the scale and complexity of the system increases, computational limitations become a factor and the models are often simplified. This has been utilized in modeling global geochemistry as applied to ozone depletion where either the

Table 4. Commonly used control devices.

Source	Pollutant	Control Devices
Automobile	CO, ROG	Oxidation catalyst, alternative fuel
	ROG	Fuel blending
	NO_x	Exhaust gas recirculation (EGR), reduction catalyst
Diesel truck	Particulate matter	Alternative fuel, particulate trap
Stationary boilers and heaters	SO_2	Scrubber, fuel cleaning, alternate fuel
(e.g., power plants)	Particulate matter	Electrostatic precipitation (ESP), filter, cyclone separations
	NO_x	Ammonia injection with/without catalyst, staged combustion
Stationary internal combustion engines	NO_x	EGR, catalyst
Organic gas venting (e.g., coke production, petrochemical processing, waste disposal)	ROG	Afterburners, catalysts
Metallurgical operations	Particulate matter	Cyclone, ESP, filter
	CO, ROG	Flare/afterburner

ROG=reactive organic gases
NO_x=nitrogen oxides
CO=carbon monoxide
SO_2=sulfur dioxide

chemistry is simplified or a one-dimensional model (vertical column) is used. Exercising these models with varying inputs can establish the impact of source emissions on air quality, deposition of acids and oxidants, global warming, and stratospheric ozone. At present, greatest use has been made for modeling local and urban areas. The ability to model global chemistry and transport accurately is still being improved.

5.3 Control Devices

Emissions can be controlled by three means: removing the source (or at least decreasing its use), lowering the emissions via a device to alter emission characteristics, or moving the point of emission. Common control devices are listed in Table 4.

Many control devices are aimed at reducing pollutant mass flux. For combustion devices, controls include fuel switching, fuel cleaning and conditioning, combustion modification, and exhaust cleanup (Stern *et al.*, 1984; Flagan and Seinfeld, 1988). The most familiar control device is the automobile catalyst used to reduce NO_x, ROG, and CO. In addition, automobiles use exhaust gas recirculation to cool combustion and lower NO_x. Combustion has also been optimized to lower emissions. To control evaporation of gasoline, charcoal filled canisters are used to absorb the vapors. These devices have cut emissions by about 90% per vehicle. A less obvious control has been to reformulate gasoline to reduce the vapor pressure and lower emissions of the most reactive organics. Removal of lead from gasoline served two purposes. It reduced lead in the atmosphere, and it eliminated the poisoning of catalysts by the lead. Use of alternative, "clean" fuels such as methanol, natural gas, hydrogen, or electric powered vehicles is a means for further pollution reduction.

Organic gas emission control from stationary sources has relied on process modification, solvent reformulation, and catalytic oxidation. Process modification may involve application procedures (e.g., electrostatic spraying) or use of collection devices (e.g., fume hoods and afterburners). Solvent reformulation relies on the use of less reactive solvents, such as water. Charcoal filtering can also be used to absorb organic vapors.

Sulfur dioxide can be removed from exhaust gases, for example, in utility boilers, by wet or dry scrubbing. The exhaust stream is passed through a shower of alkaline material, such as a lime slurry. The particles in the shower absorb the pollutant. A drawback to this type of control is the production of solid and/or liquid waste that must be treated or disposed of.

Particulate matter control is much different from controlling gases. Inertial separation is very effective for removing larger particles (e.g., cyclone separators). Smaller particles can be removed by filtration and electrostatic collection. In the latter, a corona discharge is used to charge the particles in an electric field. Electrostatic forces cause the particles to migrate to a collection surface.

Controlling pollutants that cause tropospheric warming and ozone depletion is probably the most difficult task. Carbon dioxide emissions are tightly linked to power generation. Lowering CO_2 emissions requires developing and using suitable alternative sources (e.g., solar or nuclear), or energy conservation. CFC control relies on finding suitable replacements. While its use in some applications, such as spray can propellant, may decrease, its use in foam manufacturing and air conditioning will increase (Fig. 5).

Pollution control can be expensive. Control equipment on automobiles costs hundreds of dollars per vehicle. Emission control equipment on power plants can account for a third of the costs. Proposed control devices are estimated to cost tens of thousands of dollars per ton of pollutant removed. There is a definite need to identify the most cost effective control measure to improve air quality.

5.4 Developing Strategies to Improve Air Quality

Improving air quality can be very costly. In 1985, approximately $30 billion dollars was spent in the U.S. alone to control air pollutants. To achieve the air quality standards in the United States and elsewhere will require many times that amount. There is a critical need to be able to identify cost effective means for improving air quality. This task is complicated by the multitude of sources contributing to pollutant formation, and by the uncertainties in the relationships between source emissions and air quality.

As an example, urban ozone is formed by reactions between organic gases and nitrogen oxides. Both are required to form high levels of ozone. If one of the two precursors is in great excess, ozone levels may not respond to its decrease. The task is then to show if controlling one of the precursors or both is most effective for reducing ozone. A similar question arises in controlling acid deposition: Which sources of sulfur dioxide and nitrogen oxides are most advantageously controlled to reduce acid deposition in the most ecologically sensitive areas? In this case it is necessary to develop spatially accurate source–air-quality relationships. Both of these issues are being addressed using detailed, photochemical air quality simulation models. A large number of competing strategies may be tested using these models to determine an optimal path for improving air quality. As seen in Fig. 7, a vital component of a candidate strategy is that it be economically, technologically, and politically feasible. The control measures determined to be most effective can then be adopted for implementation.

6. SUMMARY

Air pollution problems have grown from very local disturbances to having global consequences. Indoor radon, urban smog, acid deposition, global climate change, and stratospheric ozone depletion pose the most severe threats to human health and the environment. Knowledge of the problems has led to identifying solutions and to significant improvements in air quality in some local and urban areas. The air quality in other areas and the problems of regional and global pollution have not seen significant progress. Control of global pollutants requires international cooperation.

Our ability to address issues in atmospheric pollution has developed rapidly in response to the need to improve air quality. Experiments and mathematical modeling have identified how pollutants are transported and interact in the atmosphere. This in turn has made it possible to develop strategies to reduce air pollution in economically and technically feasible ways. Control devices have also evolved to lower pollutant emissions more effectively. Many major pollutant sources now emit less than a tenth of previous levels. The persistence of urban smog in spite of these efforts testifies to the severity of

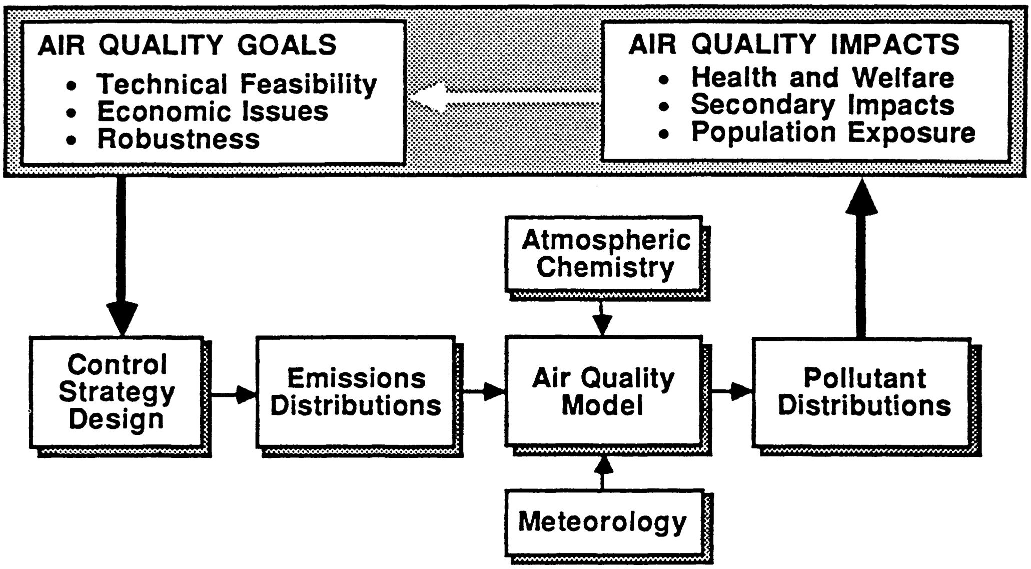
AIR QUALITY GOALS
• Technical Feasibility
• Economic Issues
• Robustness
AIR QUALITY IMPACTS
• Health and Welfare
• Secondary Impacts
• Population Exposure
Atmospheric Chemistry
Control Strategy Design
Emissions Distributions
Air Quality Model
Pollutant Distributions
Meteorology

FIG. 7. Schematic representation of the air quality planning process. Air quality goals are set to protect health and reduce environmental degradation. Control strategies are designed to meet these goals. Mathematical models are used to test the candidate strategies and to relate emissions to pollutant concentrations and environmental impact. The feedback process continues, as shown, until a set of technically, politically, and economically feasible controls are adopted.

the problem and the measures that will be required to achieve clean air. Regulatory action, our understanding of atmospheric dynamics, and technological advancement are driving continued improvement.

GLOSSARY

NO_x: Nitrogen oxides ($NO+NO_2$).

Ozone Layer: Global layer of high ozone concentrations in the stratosphere that effectively shields the earth from UV radiation.

Photolytic: Reaction driven by interaction with light, usually from the sun.

PSCs: Polar stratospheric clouds.

ROG: Reactive organic gas.

Stagnation: Meteorological condition having light winds and low mixing depth. Leads to buildup of pollutant emissions.

Stratosphere: Atmospheric layer above troposphere, extending from about 15 to 45 km.

Troposphere: Atmospheric layer closest to earth, about 15 km thick.

List of Works Cited

Anthes, R. A., Panofsky, H. A., Cahir, J. J., Rango, A. (1978), *The Atmosphere*, Columbus, OH: Merrill.

Bibbero, R. J., Young, I. G. (1974), *Systems Approach to Air Pollution Control*, New York: Wiley.

Brimblecombe, P. (1987), *The Big Smoke*, London: Methuen.

Evelyn, J. (1661), *Fumifumigium, or the Inconvenience of the Aer, and Smoake Dissipated*, London, England.

Finlayson-Pitts, B. J., Pitts, J. N. (1987), *Atmospheric Chemistry*, New York: Wiley.

Flagan, R. C., Seinfeld, J. H. (1988), *Fundamentals of Air Pollution Engineering*, Englewood Cliffs, NJ: Prentice–Hall.

Gribbon, J. (1988), *The Hole in the Sky*, New York: Bantam.

Lafaviore, M. (1987), *Radon, The Invisible Threat*, Emmaus, PA: Rodale.

McElroy, M. B., Salawitch, R. J. (1989), *Science* **243**, 763–770.

National Research Council (1981), *Formaldehyde and Other Aldehydes*, Washington, DC: National Academy.

National Research Council (1977), *Ozone and Other Photochemical Oxidants*, Washington, DC: National Academy.

O'Riordan, T. O. (1989), *Atmos. Environ.* **23**, 293–306.

Ramanathan, V. (1985), *J. Geophys. Res.* **90**, 5547–5561.

Schneider, S. H. (1989), *Science* **243**, 771–781.

Schwartz, S. E. (1989), *Science* **243**, 753–763.

Seinfeld, J. H. (1987), *Atmospheric Chemistry and Physics of Air Pollution*, New York: Wiley.

Stern, A. C., Boubel, R. W., Turner, D. B., Fox, D. L. (1984), *Fundamentals of Air Pollution*, Orlando, FL: Academic.

Walsh, P. J., Dundey, C. S., Copenhaver, E. D. (Eds.) (1984), *Indoor Air Quality*, Boca Raton, FL: CRC.

Watson, A. Y., Bates, R. R., Kennedy, D. (1988), *Air Pollution, The Automobile and Public Health*, Washington, DC: National Academy.

Wellburn, A. (1988), *Air Pollution and Acid Rain*, Essex, England: Longman.

Further Reading

Brimblecombe, P. (1987), *The Big Smoke*. London, Methuen. A historical account of the London smog.

Seinfeld, J. H. (1987), *Atmospheric Chemistry and Physics of Air Pollution*. New York: Wiley. A detailed text on the physics of air pollution.

Finlayson-Pitts, B. J., Pitts, J. N. (1987), *Atmospheric Chemistry*. New York: Wiley. Extremely complete coverage of tropospheric photochemistry.

Gribbon, J. (1988), *The Hole in the Sky*. New York: Bantam. A very readable account of the discovery and early research of antarctic stratospheric ozone depletion.

Schwartz, S. E. (1989), *Science* **243**, 753–763. An overview of the science, politics, and current status of acid deposition, particularly in North America.

ALGEBRAIC METHODS

A. J. COLEMAN, *Department of Math and Statistics, Queens University*

Introduction 515
1. **Groups** 516
2. **Fields** 516
2.1 The Characteristic of F 516
2.2 Algebraically Closed Fields 517
2.3 Rational Functions 517
3. **Linear Spaces** 518
3.1 Independence of Vectors 518
3.2 Change of Basis 518
3.3 Linear Maps and Their Associated Matrices 519
3.4 Determinants 520
3.5 Eigenvectors and Eigenvalues 520
3.6 Canonical Form of Matrices 521
3.7 Dual Space 522
3.8 Tensors 523
4. **Creating Algebraic Structures** 523
5. **Rings** 525
5.1 Examples of Rings 525
5.2 Polynomial Rings 527
5.2.1 Binomial and Multinomial Theorem 528
5.2.2 Fundamental Theorem of Algebra 528
6. **Algebras** 529
6.1 Examples of Algebras 530
7. **Modules** 534
7.1 Examples of Modules 534
7.2 Morphisms of Modules 535
Glossary 536
Works Cited 537
Further Reading 537

INTRODUCTION

The use of mathematics by physicists, and in particular of algebra, has increased in a remarkable degree during the last 50 years, both in the amount of space occupied in journal articles and in the type and abstractness of the methods employed.

Following N. Bourbaki, it is now conventional to characterize as algebraic structures those parts of mathematics that employ operations, such as addition, which act on a finite set of objects to produce a unique corresponding object. Such operations are contrasted with ideas like limit in calculus or closure in topology, which associate a number or other mathematical object to an infinite set or sequence. Thus, whereas the passage from (2,3) to $2+3 = 5$ is an algebraic operation, to go from the infinite sequence $n \rightarrow 1/n$ (where n is any positive integer) to the limit 0 is a topological operation. The present section is concerned chiefly with algebra.

In this brief article it is impossible to describe all the many algebraic structures which occur in the literature of applied physics. Therefore we have selected those which are absolutely essential for understanding the contemporary literature under the following rubrics: Groups; Fields; Linear Algebra; Rings; Algebras and Modules. As to style, we have attempted to steer a course between that which physicists would have liked 20 years ago and the austerity of contemporary pure mathematicians with which all physicists will be happy 20 years from now. This should leave all readers equally unhappy! Our definitions are seldom painstakingly detailed but rather highlight the essential ideas leaving the reader to use common sense to fill them out. We shall assume that the reader is familiar with elementary properties of vectors and matrices. Recall that a square matrix A is invertible or nonsingular if there is a matrix B such that $AB=BA=I$, where I is the identity matrix. In this case A and B are inverses of each other and we denote B by A^{-1}. Although, logically, rings should be discussed before fields, teaching experience suggests that the reverse order is pedagogically sounder.

NOTATION: We shall adopt the following widely used symbolism: $\mathbb{N}:=$the natural num-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

bers, $\{1,2,3,\ldots\}$; $\mathbb{Z}:=$the positive and negative integers and zero; $\mathbb{R}:=$the real numbers; $\mathbb{C}:=$the complex numbers; $i:=\sqrt{-1}$; $\mathbb{Q}:=$the rational numbers. We shall employ Einstein's summation convention in the restricted form that in any monomial an index which is repeated as a subscript and as a superscript will be interpreted as summed over its range unless the contrary is explicitly stated.

1. GROUPS

A group is a set, G, say, together with a binary operation which we temporarily denote by "$*$", which satisfies certain definite rules. A binary operation is one which combines two elements of G to obtain an element of G. Perhaps our first encounter with a group occurs when as babies we push our blocks around on the floor using the translation group in two dimensions! Later in grade school we learn the properties of the integers. Under addition the integers $\mathbb{Z}$ exemplify the axioms of a group:

(i) A group $(G,*)$ is a set, G, together with an operation, $*$, which to any two elements x and y of G associates an element $z=x*y$ of G. For example, in $(\mathbb{Z},+)$, $2+3=5$, $5+(-3)=2$. This property is described by saying that G is closed under the binary operation $*$.

However, for the structure $(G,*)$ to be dignified with the title "group," it must satisfy the additional properties:

(ii) The operation is associative, that is for any x,y,z in G, $(x*y)*z=x*(y*z)$.

(iii) There is a unique neutral or identity element, n, such that $x*n=n*x=x$ for all x in G.

(iv) For any element x in G there is a unique element y in G such that $x*y=n$. In this case, x and y are said to be inverses of each other.

Thus while $(\mathbb{N},+)$ satisfies (i) and (ii) it is not a group because (iii) and (iv) fail. However, $(\mathbb{Z},+)$ is a group when we take $n:=0$.

If G has a finite number of elements, the group is a finite group and the number of elements is called the order of the group. If $x*y=y*x$ for all $x,y \in G$, the group is Abelian or commutative.

The set of symmetries of any mathematical or physical structure constitutes a group under composition of symmetries. Such groups play a major role in physics for analyzing the properties of space-time, understanding crystal structure, and classifying the energy levels of atoms, molecules, and nuclei. Indeed, the role of groups is so important in physics that an article of the *Encyclopedia* is devoted to them. We therefore shall not explicitly pursue the detailed properties of groups further, even though they will occur as substructures in rings and fields.

2. FIELDS

Whereas a group consists of a set together with a single binary operation, a field consists of a set together with two binary operations linked together by a distributive law. The two operations are usually called addition and multiplication. The familiar fields are the real numbers, $\mathbb{R}$; the complex numbers, $\mathbb{C}$; and the rational numbers, $\mathbb{Q}$. We shall use the symbol $\mathbb{F}$ for an arbitrary field. Strictly speaking, we should employ a notation such as $(\mathbb{F},+,\times)$ to denote a field; however, the relevant operations are generally obvious from context in which case it is sufficient to use $\mathbb{F}$ alone.

$(\mathbb{F},+,\times)$ is a field if:

(i) $(\mathbb{F},+)$ is a commutative or Abelian group. That is, $x+y=y+x$ for any x and y in $\mathbb{F}$.

(ii) The elements of $\mathbb{F}$ other than zero form a group under multiplication.

(iii) Multiplication distributes over addition. That is, if a,b,c, belong to $\mathbb{F}$ then $a\times(b+c)=a\times b+a\times c$, and $(b+c)\times a=b\times a+c\times a$.

These properties are, of course, familiar for the reals, complexes, and rationals, but there are fields, such as the quaternions, for which multiplication is not commutative. There are also fields with only a finite number of elements.

A field always has at least two elements, 0 and 1.

2.1 The Characteristic of $\mathbb{F}$

Since a field is closed under addition, $\mathbb{F}$ contains $1+1$, which cannot equal 1 since this would imply that $1=0$, which we excluded. But $1+1$ might equal 0 in which case $(1+1)+1=1$ and one can easily check that $\mathbb{F}=\{0,1\}$ can serve as the set of a field of two elements. This is the smallest possible field and is both famous and useful since it plays a key role in the design of electric circuits, such as those which occur in computers.

More generally, if p is a prime number, we can obtain a field containing p elements in which the sums of j 1's are numbers which are distinct if $0 \leqslant j < p$ and equal to 0 if $j = p$. When this occurs in any field $\mathbb{F}$ we say that p is the characteristic of $\mathbb{F}$ and that $\mathbb{F}$ has finite characteristic. When there is no such p we say that $\mathbb{F}$ has characteristic zero. A field of characteristic zero has an infinite number of elements. If $\mathbb{F}$ has only a finite number of elements it will contain p^n elements, where p is a prime and n is a positive integer. If $n > 1$, $\mathbb{F}$ will contain a subfield of the above type with p elements. The fields with p^n elements are called Galois fields. They are important in coding and communication theory. A finite field is necessarily commutative.

2.2 Algebraically Closed Fields

We know that the square of a real number is positive, so there is no real number x such that $x^2 = -1$. In other words, in $\mathbb{R}$ there is no element x satisfying the equation $x^2 + 1 = 0$. If $\mathbb{F}$ has the property that for every equation of the form $a_j x^j = 0$, $0 \leqslant j \leqslant n$, where the a_j belong to $\mathbb{F}$, there is an element of $\mathbb{F}$ which satisfies the equation, we say that $\mathbb{F}$ is algebraically closed. Otherwise, it is not algebraically closed. Clearly $\mathbb{R}$ is not algebraically closed. If we assume that there is a "number" i such that $i^2 + 1 = 0$, then, as we know, the field containing $\mathbb{R}$ and i is the complex numbers, $\mathbb{C}$. It was proved by Gauss that $\mathbb{C}$ is algebraically closed.

Notice that if σ is a 1:1 map of $\mathbb{C}$ onto itself, such that $\sigma(x+iy) = x - iy$ for all $x, y \in \mathbb{R}$, then σ preserves all the properties of a field and is therefore an automorphism of $\mathbb{C}$. Recall that an isomorphism of two algebraic structures is a bijective (or one-to-one) correspondence between their sets, which preserves all the relations among their elements, and that an automorphism is an isomorphism of an algebraic structure onto itself. Note that $\sigma(x) = x$ if $x \in \mathbb{R}$ and that $\sigma(i) = -i$, which is the root other than i of the equation $x^2 + 1 = 0$. An automorphism of $\mathbb{C}$ must send 0 into 0 and thus must either leave i fixed (and so everything in $\mathbb{C}$ is fixed) or, like σ, send i to $-i$. The set consisting of σ and the identity map is a group of order two under composition of mappings. It is the Galois group of $\mathbb{C}$ over $\mathbb{R}$. Alternatively, it is also called the Galois group of the equation $x^2 + 1 = 0$ with respect to the reals. For more detail about fields, their algebraic extensions, and their Galois groups, the reader is referred to Jacobson (1964) or any of the multitude of algebra texts at the same level.

Are there fields containing $\mathbb{R}$ other than $\mathbb{C}$ which are at most finite dimensional over $\mathbb{R}$? The answer was given by Frobenius. There is one and only one, the quaternions, but in this field multiplication is not commutative. We shall see below that the quaternions can be realized as linear combinations with real coefficients of the Pauli matrices and the 2×2 identity matrix. The significance of the field of quaternions is dramatized by the observation that if it did not exist there would be no spin in physics, therefore no sigma and pi bonds in chemistry, and therefore no life on planet earth if, indeed, there were any stars or planets!

2.3 Rational Functions

If we adjoin a symbol x to any field $\mathbb{F}$ and form all possible sums, differences, products, and quotients involving x and the elements of $\mathbb{F}$, the result is a set which is closed under any finite sequence of these operations and forms a field, denoted by $\mathbb{F}(x)$, which we might describe as the field of rational functions in x over $\mathbb{F}$. There is a subset, denoted by $\mathbb{F}[x]$, of polynomials of the form $a_j x^j$ where j is summed from 0 to some $n \in \mathbb{N}$, where n is arbitary and the $a_j \in \mathbb{F}$. If a_n is not zero we say that the polynomial has degree n. As we shall remark below, the polynomials constitute a ring. As usual $x^0 := 1$, by definition, so when $n = 0$ the preceding polynomial reduces to a_0. Thus $\mathbb{F}$ is contained in $\mathbb{F}[x]$. The field $\mathbb{F}(x)$ consists of all possible quotients of elements of $\mathbb{F}[x]$.

Suppose that the rational function $R(x) = P(x)/Q(x)$, where P and Q are polynomials. Suppose further that $Q(x) = Q_1(x)Q_2(x)$, where Q_1 and Q_2 are polynomials with no common factor. Since we could have used long division to ensure that R is the sum of a polynomial and a rational function, the numerator of which has degree strictly less than the degree of Q, we may assume that deg(P)—the degree of P—is less than deg(Q). It is relatively easy to show that it is then possible to find polynomials P_1 and P_2 with $\deg(P_i) < \deg(Q_i)$ such that

$$P/Q = P_1/Q_1 + P_2/Q_2.$$

This is the fundamental theorem of the so-called method of partial fractions, by repeated application of which it follows that any rational function can be expressed as a sum of a polynomial and rational functions whose denominators have no nontrivial factors.

In particular, if $\mathbb{F}$ is algebraically closed (e.g., $\mathbb{F}=\mathbb{C}$), then Q is a product of factors such as $(x-a)^m$, where $a \in \mathbb{F}$. A summand in $R(x)$ of the form $g(x)/(x-a)^m$ with $\deg(g) < m$ can, by Taylor's theorem applied to $g(x)$, be expressed as the sum $c_j(x-a)^{-j}$, where $1 \leqslant j < m$ and $c_j = g^{(m-j)}(a)/(m-j)!$. Here $g^{(k)}$ is the kth order derivative of g.

The method of partial fractions is quite useful for finding integrals of rational functions. Books on calculus explain helpful tricks for obtaining the partial fraction decomposition of a given rational function.

3. LINEAR SPACES

The theory of linear space with its related concepts of linear transformation, eigenvector, matrix, determinant, and Jordan canonical form is certainly one of the most important and most useful part of mathematics. The abstract concept of linear space is frequently approached by a long discussion of the problem of solving systems of linear equations. We take a direct approach defining a linear space as consisting of a field $\mathbb{F}$ whose elements are called scalars, a set V, called vectors, and two operations called vector addition and scalar multiplication together with a set of rules governing the relation among these various entities. The vectors under addition form an additive Abelian group $(V,+)$. Under multiplication by scalars the set V is closed. Thus, $v \in V$ and $a \in \mathbb{F}$ imply that $av \in V$. Another important property is that multiplication by scalars distributes over addition of vectors. That is $a(v_1+v_2)=av_1+av_2$ for all $a \in \mathbb{F}$ and $v_i \in V$.

3.1 Independence of Vectors

This seems to be the most difficult idea in teaching elementary courses in linear algebra—possibly, the only difficult idea! Two nonzero vectors v_1 and v_2 are linearly dependent if there are scalars a^1 and a^2, not both zero, such that $a^1v_1+a^2v_2=0$, where, of course, by 0 we mean the zero vector. It is clear that neither a^1 nor a^2 is zero, and thus each vector is a scalar multiple of the other. More generally, if, given n vectors v_i, $1 \leqslant i \leqslant n$, there exist scalars a^i such that $a^iv_i=0$, where all $v_i \neq 0$ and not all $a^i=0$; then we say that the n vectors are linearly dependent. If no such relation holds, the vectors are linearly independent. For example, for $n \in \mathbb{N}$ there are no numbers a_n other than 0 such that $\Sigma_n a_n \times \cos(n\vartheta)=0$ for all ϑ. Thus the functions $\vartheta \rightarrow \cos(n\vartheta)$ are linearly independent.

If n vectors v_i are such that any vector v can be written as $v=a^iv_i$ for some choice of scalars a^i, we say that the set $\{v_i\}$ spans V. If the v_i are also linearly independent then the coefficients a^i are unique. We then say that $B=\{v_i\}$ is a basis of V, that the linear space V has dimension n, and that a^i are the components of v with respect to B. A basic theorem assures us that the dimension depends only on the space V and not on the choice of basis. If a linear space does not have a finite basis it is infinite dimensional.

3.2 Change of Basis

How do the components of a given vector change if the basis is changed? This was a key question which led to the theory of invariants in the mid-19th century and opened up the development of much of contemporary algebra. It also led to the emergence of the tensor calculus which was essential for Einstein's exposition of General Relativity Theory.

Suppose V is a linear space and that $B = \{v_i\}$ and $B' = \{v'_i\}$ are two different bases for V. Then there is a matrix, P^j_i called the transition matrix from B to B' such that $v'_i = P^j_i v_j$. Thus if the vector $x = x^j v_j = x'^{\,i} v'_i = x'^{\,i} P^j_i v_j$, it follows that $x^j = P^j_i x'^{\,i}$. If in the usual matrix notation we regard x^j as the jth component of a column vector $\mathbf{x}$, and P^j_i as the element in the jth row and ith column of the transition matrix, $\mathbf{P}$, from the old base B to the new base B', the preceding equation takes the form

$$\mathbf{x}=\mathbf{P}\mathbf{x}' \quad \text{or} \quad \mathbf{x}'=\mathbf{P}^{-1}\mathbf{x}.$$

There is an even more convenient notation. Define $P(B',B):=\mathbf{P}$; then the preceding equations imply that $\mathbf{P}^{-1}=P(B,B')$. Subsequently we shall need the formulas

$$\mathbf{x}=P(B',B)\mathbf{x}' \quad \text{and} \quad \mathbf{x}'=P(B,B')\mathbf{x}.$$

To understand tensor notation it will prove important to note that, whereas **P** sends the old to the new basis, it sends the new coordinates to the old ones. This observation underlies duality in homological algebra and the distinction between covariant and contravariant tensors, which we define below.

3.3 Linear Maps and Their Associated Matrices

Suppose that V and U are linear spaces of dimension n and m with bases $B_v=\{v_i\}$ and $B_u=\{u_j\}$, respectively. A transformation, function, or map from V to U sends each vector $x\in V$ to a vector, say, y of U, which we denote by $Ax:=y$. If A has the property that for any two vectors x and $x'\in V$ and arbitrary scalars a and $a'\in \mathbb{F}$, $A(ax+a'x')=aAx+a'Ax'$, we say that A is linear. The condition that a map be linear is very restrictive. Nonetheless, linear maps play a big role in the application of mathematics to physics (as well as statistics, economics, biology, etc.) for the same reason that the derivative is important in analysis. For example, if $f(x)$ is a real-valued function of the real variable x, such that $f(0)=0$, then $f'(0)x$ is the linear function of x which is the best possible linear approximation to $f(x)$ near 0.

A linear map or transformation, $A:V\rightarrow U$, can be completely described by an $m\times n$ matrix A^j_i, such that $Av_i=A^j_iu_j$, which we describe as the matrix associated to the linear transformation or map A with respect to the bases B_u and B_v. It has m rows and n columns. If we denote this matrix by $A(u,v)$ then if the bases in V and U are changed to B'_v and B'_u, respectively,

$$A(u',v')=P(B'_u,B_u)A(u,v)P(B_v,B'_v),$$

where we use the notation of Sec. 3.2. In terms of coordinates, if $\mathbf{y}=\mathbf{A}\mathbf{x}$, then $y^j=A^j_i\,x^i$, where, as follows from the context, $1\leqslant j\leqslant m$ and $1\leqslant i\leqslant n$.

In the particular case that $V=U$ of dimension n, with $B_u=B_v$, $A(u,u)$ is an $n\times n$ matrix which we denote by $A(u)$. We deduce that for a change of basis

$$A(u')=P(B'_u,B_u)A(u)P(B_u,B'_u).$$

We thus associate to any linear transformation a matrix which is unique, *once bases are chosen* for the domain and codomain of the transformation. But conversely, if the bases are given, then there is a unique linear transformation associated with a given matrix of the appropriate shape. Thus there is a bijection (i.e., a one-to-one correspondence) between $m\times n$ matrices with entries in $\mathbb{F}$ and linear maps from a linear space of dimension n into one of dimension m. We have found how the bijection changes when the bases are altered. It is this bijection which gives meaning to the familiar addition and multiplication of matrices.

A linear map between two spaces over the same field $\mathbb{F}$ has the property of preserving the linear structure and is said to be a homomorphism (i.e., a structure-preserving map), so it is common to denote by $\mathrm{Hom}(V_1,V_2)$ the set of all linear maps between linear spaces V_1 and V_2 where both have the same field. If $A\in\mathrm{Hom}(V_1, V_2)$ then V_1 is the domain of A and V_2 is the codomain of A. The kernel of A, frequently denoted by $\ker(A)$, is the set of all elements in the domain which are mapped onto 0 by A. The range of A consists of all elements of the codomain of the form Ax for some x in the domain. Of course these last four defintions are valid for any function, not merely linear maps. However, when A is linear it can be easily proved that both the kernel and the range are linear subspaces of their ambient spaces. This is probably the secret of the power and relative simplicity of the theory of linear spaces. When $V_1=V_2=V$, we denote $\mathrm{Hom}(V,V)$ by $\mathrm{Hom}(V)$.

If G is a map from V_1 to V_2 and F one from V_2 to V_3, we denote the composition of these two maps by FG and, having fixed bases in the spaces, we define the matrix corresponding to FG as the product of the matrices corresponding to F and G. This "explains" the usual rule for matrices that $(FG)^j_i=F^j_kG^k_i$, where i, k, and j range from 1 to the dimensions of V_1, V_2, and V_3, respectively.

The composition (or product) of two maps can be well-defined if the range of the first is in the domain of the second. The sum of two maps is only meaningful if the codomain is an additive group in order for the sum of Fx and Gx to be meaningful. In this case it is possible to let $F+G$ denote the map such that $(F+G)x=Fx+Gx$ for all x in the intersection of the domains of F and G. When the domain and codomain are fixed linear spaces over the same field $\mathbb{F}$ we can do even better and give $\mathrm{Hom}(V_1, V_2)$ the structure of a linear space over $\mathbb{F}$. This implies that the set of all $m\times n$ matrices with entries from $\mathbb{F}$ is a linear space of dimension mn over $\mathbb{F}$.

The dimension of the range of a linear operator is called the rank of the operator and also the rank of any matrix associated with the operator by a particular choice of bases. The dimension of the kernel of a linear transformation is called the nullity of the transformation and of its associated matrices. It follows from this definition that the various matrices obtained from one another by a change of basis all have the same rank and nullity. The rank of a product of operators or matrices is not greater than the minimum rank of its factors.

3.4 Determinants

If $\mathbb{F}$ is a commutative field, to any square matrix, it is possible to assign a number in $\mathbb{F}$ which is expressible as a polynomial in the elements of the matrix and which vanishes only if the matrix is not invertible. To two square matrices which are related as in Sec. 3.3 by a change of basis, we assign the same number, and therefore it is meaningful to also assign this number to the associated linear transformation belonging to $\mathrm{Hom}(V)$. The function, det, from $\mathrm{Hom}(V)$ into $\mathbb{F}$, has the following properties: (i) $\det(AB)=\det(A)\det(B)$; (ii) $\det(fI)=f^n$, where n is the dimension of V, I is the identity map, and f is any element of $\mathbb{F}$. The usual definition of the determinant follows from these properties (MacDuffee, 1943). In particular since, for a fixed basis, the equation $Ax=y$ is equivalent to the system of equations $A^j_i x^i=y^j$, Cramèr's rule implies

$$\det(A)x^i=\det(Y^i),$$

where Y^i is the matrix obtained from (A^j_i) by replacing its ith column by the column vector (y^k) where i, j, k run from 1 to n. Thus if $\det(A)\neq 0$ there is a unique x for every y so A is invertible; whereas if $\det(A)=0$, there is an x only for particular y satisfying the n conditions $\det(Y^k)=0$. Thus for a finite dimensional linear space V, $A\in\mathrm{Hom}(V)$ is invertible if and only if $\det(A)\neq 0$.

The theory of $\mathrm{Hom}(V_1,V_2)$ is really equivalent to the theory of systems of linear equations in several variables. This topic occurs in articles of this *Encyclopedia* devoted to NUMERICAL METHODS and to MATHEMATICAL MODELING and in at least one hundred elementary textbooks; so we shall not pursue it here.

3.5 Eigenvectors and Eigenvalues

If $A\in\mathrm{Hom}(V)$ then for any $x\in V$, $Ax\in V$. In general we shall not expect Ax to equal x or indeed, even, that Ax be parallel to x. However, in the latter case Ax would be a multiple of x, say, λx. The equation $Ax=\lambda x$ is equivalent to $(\lambda I-A)x=0$. By the preceding section, if the determinant of $\lambda I-A$ is different from zero, the only possible solution of this equation is $x=0$, which is of no great interest. When there is a nontrivial solution of this equation it will be somewhat unusual and is called an eigenvector of A and can occur only for special values of λ. Such a value of λ is the eigenvalue of A corresponding to the particular eigenvector x. The eigenvalue, λ, will satisfy the nth degree algebraic equation

$$f(z;A):=\det(zI-A)=0.$$

The nth degree polynomial $f(z;A)$ is called the characteristic function of A, and the preceding equation is the characteristic equation of A. Any eigenvalue of A satisfies its characteristic equation. For each zero of the characteristic equation there is at least one nontrivial eigenvector.

There is a one-to-one correspondence between the operators in $\mathrm{Hom}(V)$ and the set of $n\times n$ matrices over $\mathbb{F}$, and this set spans a linear space over $\mathbb{F}$ of dimension n^2. If we interpret A^0 as the identity operator, I, it follows that the operators A^k for $0\leqslant k\leqslant n^2$ are linearly dependent. That is, there are $c_j\in\mathbb{F}$ such that $c_j A^j=0$, where not all c_j are zero. Thus there exists at least one polynomial, $p(z)$, such that $p(\mathrm{A})=0$. From the algorithm for long division it easily follows that there is a unique monic polynomial (i.e., a polynomial with highest coefficient 1) of minimal degree with this property. We shall denote this so-called minimal polynomial of A by $m(z;A)$. A famous theorem of Hamilton asserts that A satisfies its characteristic equation. That is, $f(A;A)=0$.

Since $\deg(f)=n$, $\deg[m(z;A)]\leqslant n$. Since $m(z,;A$ divides any polynomial $p(z)$ such that $p(A)=0$, it follows that $m(z;A)$ divides $f(z;A)$.

The form of $m(z;A)$ provides information about A.

(i) $m(z{:}A)=z^p$ implies that $A^p=0$ but that $A^{p-1}\neq0$. Such an operator is called nilpotent, with nilpotency index p.

(ii) $m(z;A)=(z-1)^p$ implies that $A-I$ is nilpotent with index p. Thus in this case $A=I+N$, where N is nilpotent. An operator of this form is called unipotent.

(iii) Suppose the minimal polynomial of A has no multiple zeros, which is equivalent to saying that m and its derivative have no common factors. Then there is a basis of V consisting of eigenvectors of A. Equivalently, among the matrices associated with A there is one which is diagonal. In this case we say that A and its associated matrices are diagonalizable or semisimple.

(iv) If $m(z;A)=(z-\lambda)^p$, then, of course, $p\leqslant n$. A basis can be chosen so that the matrix corresponding to A has zero entries except along the diagonal where there are so-called Jordan blocks, which in case $n=4$, for example, would be

$$\begin{bmatrix}\lambda & 1 & 0 & 0\\ 0 & \lambda & 1 & 0\\ 0 & 0 & \lambda & 1\\ 0 & 0 & 0 & \lambda\end{bmatrix}.$$

That is $n_i\times n_i$ matrices with λ on the diagonal and 1's on the first superdiagonal, $\Sigma n_i=n$, $1\leqslant n_i\leqslant p$, and for at least one value of i, $n_i=p$.

In the preceding we have assumed that the entries of the matrix A could be arbitrary. However, if they are real and nonnegative the remarkable Perron–Frobenius theorem (Senata, 1973) about the eigenvalues and eigenvectors of A gives information which is useful in many contexts; we thus state it here. A matrix $M=(m_{ij})$ is connected or indecomposable if for any two indices i and j there is a sequence r_k, $1\leqslant k\leqslant s$, such that the continued product $m_{ir_1}m_{r_1r_2}m_{r_2r_3}\dots m_{r_sj}\neq0$. We write $M>0$ if all $m_{ij}>0$, and $M\geqslant0$ if all $m_{ij}\geqslant0$. Then, if $M\geqslant0$ is a real connected matrix, it has a largest simple positive eigenvalue, $r(M)=r$, and an associated column vector $x>0$, such that $Mx=rx$ where $r>0$; any other eigenvalue λ of M has absolute value less than or equal to r. Further, if $N\geqslant0$ is another real matrix of the same dimension, such that $M-N\geqslant0$, then $r(N)\leqslant r(M)$ with equality only if $N=M$. This theorem can be used to quickly give the basic classification of Kac–Moody algebras.

3.6 Canonical Form of Matrices

In Sec. 3.3 we noticed that distinct matrices were associated with the same linear operator, so there is a sense in which such matrices are "equivalent." Recall that by an equivalence relation a set is partitioned into distinct mutually exclusive subsets which exhaust the given set. One method of partitioning a set is into the orbits of a group which acts on the set. Thus if g belongs to a group G which is acting on a set S and we denote by gs the element of S into which g sends s, the orbit of s is the set $M_s=\{gs\,|\,g\in G\}$. It follows that $x\in M_s$ implies that $M_x=M_s$. Given an equivalence relation on a set of matrices, the problem considered in this section is that of choosing a canonical or "simplest" matrix in each equivalence class. There are different canonical forms depending on the types of matrices we consider and the different group actions contemplated.

Possibly the basic and most general situation is that considered by H. J. S. Smith in 1861. It is that of Sec. 3.3 where the equation $A(u',v')=PA(u,v)Q$ occurs in slightly different notation. There P and Q are arbitrary invertible $m\times m$ and $n\times n$ matrices, respectively. By choosing B'_u so that the last elements of the basis span the kernel of A and the first ones span a subspace which is complementary to the kernel, while the first elements of B_u span the range of A, one can arrive at Smith's canonical matrix which has 1's in the (i,i) positions for $1\leqslant i\leqslant r$ where r is the rank of A, and zero everywhere else. It would be difficult to demand anything "simpler." It follows that with this meaning of equivalence there are $p+1$ equivalence classes of $m\times n$ matrices where p is the minimum of $\{m,n\}$.

At first one is surprised that there are so few classes. However, on second thought, one notices that we have been acting on a space of matrices of dimension mn by the group $Gl(n,\mathbb{F})\times Gl(m,\mathbb{F})$ $(=G$, say), which has $n^2+m^2\geqslant2mn$ parameters; there is plenty of redundancy unless one of m and n is 1 and the other is 1 or 2.

If we consider an action on the set of $n\times n$ matrices by a smaller group we shall expect more equivalence classes. For $(P,Q) \in G$, subgroups of G can be defined by imposing restrictions on P and Q.

Recall the following definitions. If A is a square matrix the transpose of A, denoted by A^t, is obtained from A by interchanging rows and columns or by reflecting across the main diagonal. The operation of taking the transpose is an involution, that is $(A^t)^t=A$. If $A^t=A$, then we say A is symmetric. If $A^t=-A$, then A is antisymmetric or skew-symmetric. An important property of transposition is $(AB)^t=B^tA^t$. It is worth noting that once the basis of V has been fixed, the mapping defined by transposition of matrices can be transferred to the associated linear transformations, thus defining an involution on Hom(V).

If σ is an automorphism of $\mathbb{F}$, we can define an operation on the matrix A by replacing each of its elements by its conjugate under the automorphism, and denote the new matrix by A^σ. If the field is commutative $(AB)^\sigma=A^\sigma B^\sigma$. In particular, when $\mathbb{F}=\mathbb{C}$, complex conjugation is an automorphism of period two. We follow a common custom and denote the complex conjugate of A by $\overline{A}$, so $\overline{AB}=\overline{A}\,\overline{B}$.

The Hermitian conjugate of A is denoted by $A^*=\overline{A}^t$ and satisfies $(AB)=B^*A^*$. A matrix A is Hermitian if $A^*=A$ and anti-Hermitian if $A^*=-A$.

The approach of this section is based on that of Turnbull and Aitken (1932), a superb book which goes far beyond our brief summary. They distinguish five subgroups of G.

(i) The Collinearity Group is characterized by $PQ=I$. It arises in Sec. 3.3 when $v=u$ and $v'=u'$. Under the action of this group, a square matrix, A, can be reduced to Jordan canonical form, that is to a sum of diagonal blocks, each of which has the form $\lambda I+N$, where λ is an eigenvalue of A and N is a nilpotent matrix, all of whose entries are zero except for 1's along the first superdiagonal. A particular eigenvalue occurs on the diagonal of the canonical form as many times as its multiplicity in the characteristic equation. For any eigenvalue the dimension of the largest Jordan block is equal to the multiplicity of the eigenvalue in the minimal polynomial $m(z;A)$. Thus if the zeros of $m(z;A)$ are simple, A is diagonalizable.

(ii) The Congruent Subgroup is defined by the condition $P^t=Q$. Under this group, symmetry or antisymmetry of A is invariant. A symmetric matrix can be diagonalized. If $\mathbb{F}$ is closed under taking square-roots, we can choose as the canonical element of an equivalence class a diagonal matrix which has only 0's or 1's on the diagonal. If $\mathbb{F}=\mathbb{R}$, the diagonal could also contain -1. In the real case, Sylvester's Law of Inertia asserts that the number of 1's and the number of -1's are invariants. A nonsingular antisymmetric matrix has even rank r and there is a canonical form under the congruent group which contains zeros everywhere except for $r/2$ blocks of 2×2 antisymmetric matrices down the diagonal; each has 1 and -1 off the diagonal and 0 on the diagonal.

(iii) The Conjunctive Subgroup is defined by the condition $P=Q^*$. It changes Hermitian matrices into Hermitian matrices. For real matrices, the conjunctive and the congruent transformations are the same. For any $\mathbb{F}$, one may choose a diagonal matrix as canonical. If $\mathbb{F}=\mathbb{C}$, the diagonal can consist of 1's and 0's.

(iv) The Orthogonal Group is defined by $PQ=I$ and $P =Q^t$ and is thus a subgroup of the groups (i) and (ii). It will preserve symmetry or antisymmetry of a matrix. A symmetric matrix will be equivalent to a diagonal matrix whose diagonal elements are eigenvalues of the original matrix. An antisymmetric matrix will be equivalent to one with zeros everywhere except for 2×2 blocks on the diagonal, the determinants of these blocks being equal to the negatives of the squares of eigenvalues of the original matrix.

(v) The Unitary Subgroup is defined by $PQ=I$ and $P=Q^*$, and is thus a subgroup of (i) and (iii). It preserves the property of a matrix being Hermitian or anti-Hermitian. If $\mathbb{F}=\mathbb{R}$, groups (v) and (iv) are the same. Under this group, a Hermitian matrix is equivalent to a diagonal matrix whose nonzero elements are eigenvalues of the original matrix. An anti-Hermitian matrix is equivalent to one with 2-dimensional blocks on the diagonal whose determinants are the negatives of the squares of eigenvalues of the original matrix.

3.7 Dual Space

We have already noted that Hom(V,U), where V and U are linear spaces of dimension n and m, respectively, over a common field $\mathbb{F}$, can be given a structure of a linear space of dimension nm over $\mathbb{F}$. We can, of course consider $\mathbb{F}$ as a linear space of dimension 1 over $\mathbb{F}$. Thus, Hom(V,$\mathbb{F}$) is a linear space of dimension n over $\mathbb{F}$ and therefore isomorphic to $\mathbb{F}^n$ and hence also to V. It is called the dual space of V and usually denoted by V^*. This use of the asterisk can be distinguished from its use to indicate Hermitian conjugation by the context. The elements of V^* are linear functions on V with values in $\mathbb{F}$. We shall denote them by lower case Greek letters. Recall that the Kronecker symbol δ^i_j takes the value 1 if $i=j$ and 0 otherwise.

If $\alpha \in V^*$ and $x=x^jv_j$ is an arbitrary vector in V expressed in terms of the basis B_v, then $\alpha(x)=x^j\alpha(v_j)=a_jx^j$, where $a_j=\alpha(v_j)$. It is possible to define various bases for V^*. The basis which is said to be dual to B_v, and may be denoted by B^*_v, is defined as follows. Recall that a linear function on V is completely determined by the values it assumes for the elements of a basis of V.

Let α^i be a linear function such that $\alpha^i(v_j)=\delta^i_j$ for all j, $1\leqslant j\leqslant n$. Then $\alpha^i(x)=x^i$. Thus α^i is the ith coordinate function. It easily follows that α^i are linearly independent and that $\alpha=a_j\alpha^j$, where $a_j=\alpha(v_j)$. Thus any element of V^* is a linear combination of the n elements α^j, $1\leqslant j\leqslant n$, so that $B^*_v=\{\alpha^j\}$is a basis for V. Just as the x^i are coordinates of an arbitrary element of V with respect to B_v, so a_i are coordinates of an arbitrary element of V^*. Since $a_i=\alpha(v_i)$, when the basis of V is changed, a_i changes by the same transformation as, or cogrediently with, the basis. As we noted at the end of Sec. 3.2, the x^i transform contragrediently to the basis. This distinction reflects the fact that the definition of the linear function α: $x\rightarrow\alpha(x)$ is independent of the coordinate system used to describe it. A geometrical or physical entity which is described by a sequence of n numbers which transform like (a_i) is called a covariant vector. Similarly, an entity described by a sequence of n numbers which transform like (x^i) when the basis is changed is called a contravariant vector.

3.8 Tensors

Possibly it was algebraic geometers in the middle of the nineteenth century who first focused attention on the behavior of the coordinates of geometrical objects when the frame of reference is changed. But the first time this issue really impinged on physics was with the advent of Einstein's General Relativity Theory (GRT). The basic metric of GRT, $g_{ij}dx^idx^j$, is clearly independent of the coordinate system but since dx^i is a contravariant vector, g_{ij} will have to vary covariantly in both subscripts i and j. Then the n^2 symbols g_{ij} must be describing something (in fact, according to Einstein, the gravitational field!) which is a doubly covariant tensor.

The curvature of space-time, which allegedly explains black holes and how planets circle around the sun, is described by the Riemann–Christoffel tensor, $R^i{}_{jkl}$, which is contravariant in the index i and covariant in the other three.

The great advantage of the indicial notation, as it evolved in the writings of Eddington, Weyl, Synge, and other mathematical physicists between 1920 and 1940, is that it immediately indicates the behavior of the tensor when the underlying basis, or frame of reference, is changed. Thus if a_{ij} is a double covariant tensor and b^i is a contravariant vector (or first order tensor), then $a_{ij}b^k$ is a third order tensor covariant in two indices and contravariant in one. If we now contract on the indices j and k, we see immediately that $c_i=a_{ij}b^j$ is a covariant vector.

An algebraist would say that a_{ij} are the components of an element of $V^*\otimes V^*$, the tensor product of the dual space of V with itself. Similarly, $a^ib^j_k$ are the components of an element in the tensor product $V\otimes V\otimes V^*$. In general, the tensor product (see Sec. 4) of two linear spaces of dimension n and m is a linear space of dimension nm. In particular, $V^*\otimes U$ is isomorphic to Hom(V,U) and is spanned by a basis consisting of elements noted as $\alpha^i\otimes u_j$, where $1\leqslant i\leqslant n$ and $1\leqslant j\leqslant m$.

4. CREATING ALGEBRAIC STRUCTURES

What experimental apparatus is for the physicist, the Cartesian product and quotient structures are for the algebraist. These are the principal tools with which he makes new mathematical structures.

If A and B are two sets, the Cartesian product of A and B is denoted by $A\times B$ and defined as the set $\{(x,y)\,|\,x\in A,\ y\in B\}$. Thus it is a new set consisting of ordered pairs with the first element of the pair belonging to A and the second to B. If $A\neq B$, $A\times B\neq B\times A$, since by definition two ordered pairs (a,b) and (c,d) are equal only if $a=c$ and $b=d$.

Things become more interesting when A and B have some algebraic structure which can be used to impose structure on the Cartesian product. For example, suppose that $A=B=\mathbb{Z}$. We define the addition of pairs $\in \mathbb{Z}\times\mathbb{Z}$ by $(x,y)+(u,v):=(x+u,\ y+v)$. Notice that the plus signs on the right and left have quite different meanings. One acts on pairs of integers; the others on integers. If we think of $+3$ as a translation by 3 units along the number line, we can call $(\mathbb{Z},+)$ a translation group in one dimension. We could then think of $(\mathbb{Z}\times\mathbb{Z},+)$ as the translation group of a two-dimensional lattice. Another familiar example is the idea due to Gauss of imposing the structure of the complex numbers on $\mathbb{R}\times\mathbb{R}$.

The direct sum of two vector spaces provides us with another important example of this construction. Suppose X and V are two linear spaces over the same field $\mathbb{F}$ with bases $\{e_i\}$, $1\leqslant i\leqslant n$, and $\{f_j\}$, $1\leqslant j\leqslant m$ respectively. For $x,y\in X$, $u,v\in V$, and $\alpha\in\mathbb{F}$, define (i) $(x,u)+(y,v):=(x+y,u+v)$; (ii) $\alpha(x,u):=(\alpha x,\alpha u)$. By these definitions we have imposed on $X\times V$ the structure of a linear space for which the $n+m$ elements $\{(e_i,0),\ (0,f_j)\}$ form a basis. This new linear space is called the direct sum of the linear spaces X and V, and has dimension $n+m$, and is denoted by $X\oplus V$.

An apparently minor variation on the preceding definition leads us to an important but quite different object—the tensor product of X and V which is denoted by $X\otimes V$. This is a linear space which has a basis of elements belonging to the Cartesian product $X\times V$, but addition and scalar multiplication are different. (i) $(x_1+x_2,\ v_1+v_2)=(x_1,v_1)+(x_1,v_2)+(x_2,v_1)+(x_2,v_2)$; (ii) $\alpha(x,v)=(\alpha x,v)=(x,\alpha v)$. These conditions imply that $X\otimes V$ is a vector space over $\mathbb{F}$ of dimension mn, with $\{(e_i,f_j)\}$ as a basis.

Recall that an equivalence relation ρ on a set S partitions S into mutually exhaustive subsets which we call equivalence classes. A binary relation ρ on a set is an equivalence relation if it has the following three properties: (i) reflexive, $x\rho x$ for all $x\in S$, (ii) symmetric, $x\rho y$ implies $y\rho x$, (iii) transitive, $x\rho y$ and $y\rho z$ imply $x\rho z$. A subset of the partition of S contains exactly all the elements of S which are related by ρ to any one member of the subset. For example, the nails in a hardware store can be partitioned by length. Thus $x\rho y$ means length(x) $=$length(y).

Now consider the equivalence relation ρ on $\mathbb{Z}\times\mathbb{Z}$ such that $(a,b)\rho(u,v)$ if and only if $av=bu$. We have used only properties of the integers to partition $\mathbb{Z}\times\mathbb{Z}$ into equivalence classes. But the condition we used is identical with the equality of the rational numbers a/b and u/v. We have thus established a bijection, or one-to-one correspondence, between $\mathbb{Q}$ and the equivalence classes of $\mathbb{Z}\times\mathbb{Z}$ under the relation ρ.

For any set S with equivalence relation ρ, the new set whose elements are equivalence classes of S is denoted by S/ρ and called the quotient set of S by the relation ρ. Starting from $\mathbb{Z}$ we have just created the rationals $\mathbb{Q}$ as $(\mathbb{Z}\times\mathbb{Z})/\rho$. The notion of quotient structure frequently arises in physics when we have a group G acting on some set S.

Suppose that G acts transitively on S, that is if Q is a fixed point, and P is any point, there is at least one transformation in G which sends Q to P. The set of all transformations which leave Q fixed is a subgroup H of G—the so-called stabilizer of Q. For any two elements f and g of G we shall say that they are in the relation ρ, that is $f\rho g$, if $fg^{-1}\in H$. We easily prove that ρ is an equivalence relation and that the points of S are in one-to-one correspondence with the elements of G/ρ. Thus the physics of S can be transferred to G/ρ and the symmetries of the physical situation may become more transparent.

When the relation is defined by a subgroup H as above, G/ρ is usually denoted by G/H. Suppose we denote the equivalence class containing g by $\pi(g)$, if g is any element of G. That is π is a mapping from G to G/H, the so-called canonical map. We could ask whether it is possible to impose on G/H a structure of a group in such a way that for any $f,\ g\in G$, $\pi(fg)=\ \pi(f)\pi(g)$. The answer is yes—if and only if H is a normal subgroup of G. A normal subgroup is not only a group but has the additional property that for all $g\in G$, $gHg^{-1}=H$. Further $H=\{g\in G\,|\,\pi(g)=e\}$ where e is the neutral element of the new group G/H. When the subgroup H is not normal, G/H is not a group but is called a homogeneous space on which G acts transitively.

We shall meet below other examples of the use of quotienting as a method of creating new structures.

5. RINGS

A ring like a field consists of a set, R, together with two binary operations which are usually called addition and multiplication. $(R,+,\times)$ is a ring if

(i) $(R,+)$ is a commutative additive group with zero;

(ii) $(R,\times)$ is closed under multiplication and may or may not have a unit;

(iii) multiplication distributes over addition, i.e., $a(x+y)=ax+ay$ for all a, x, and y in R.

We do not require that nonzero elements of R have reciprocals in R, nor that multiplication be commutative or associative, but we do not exclude these properties. Thus a field is a ring but not all rings are fields.

5.1 Examples of Rings

We now list five rings and one "almost ring" which occur frequently in the physics literature.

(a) *The Integers* $\mathbb{Z}$. Perhaps it was this example which led to the emergence of the concept of ring. The integers form a group under addition and are therefore closed under addition and subtraction. They are also closed under multiplication, which distributes over addition. However, the solution, x, of the equation $mx=n$, where $m,n \in \mathbb{Z}$, is not,in general, an element of $\mathbb{Z}$. In contrast with some other rings there are no divisors of zero in the integers. That is you cannot find two integers, neither of which is zero, whose product is zero.

(b) *Square Matrices*. Suppose $A=(a^i_j)$, $B=(b^i_j)$, and $C=(c^i_j)$ are $n\times n$ matrices with entries in a field $\mathbb{F}$; then we define $A+B$ and AB or $A\times B$ to be $n\times n$ matrices whose entries in the ith row and jth column are, respectively, $a^i_j+b^i_j$ and $a^i_k b^k_j$. (Recall the summation convention in the Introduction.) Here $1\leqslant i,j,k \leqslant n$. Let $M_n(\mathbb{F})=M_n$ denote the set of all $n\times n$ matrices with entries in the commutative field $\mathbb{F}$. Then one can verify that $(M_n,+,\times)$ is an associative ring which is noncommutative if $n\geqslant 2$. The zero element of the ring is the matrix all of whose entries are 0, whereas the unit or identity for multiplication is the matrix (δ^i_j) which has 1 on the diagonal and 0 elsewhere.

Notice that if $n=2$,

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 3 & 7 \\ 0 & 0 \end{bmatrix}=\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix};$$

thus the ring of square matrices possesses zero divisors.

(c) *Quaternions* were invented by Sir William Rowan Hamilton in order to give a convenient description of rotations in 3-space.

In this section we shall use j,k,s,t as indices with their ranges restricted as follows: $1\leqslant j,\ k \leqslant 3$, and $0\leqslant s,\ t \leqslant 3$. The quaternions, H, form an associative ring with multiplicative identity, $I=e_0$, and contain three elements e_j satisfying the conditions $e_j e_k + e_k e_j = -2\delta_{jk}e_0$, so $e_j^2=-e_0$. Further, $e_j e_k=e_m$ where (j,k,m) is an even permutation of (1,2,3). As a ring, H will contain $e_0+e_0{:}=2e_0$, etc., so that H contains $\mathbb{Z}e_0$. More generally if R denotes any commutative ring, we could assume that H contains Re_0 and note this explicitly by denoting the quaternions as $H(R)$. Hamilton considered only the possibility that $R=\mathbb{R}$, the real numbers, since his concern was rotations in the 3-dimensional space of Newtonian physics—not some esoteric space of super string theory! Over R we can define H by

$$H=\{x^s e_s \mid x^s \in R\}.$$

Then it follows that H is closed under addition and multiplication. If we demand that the associative and distributive properties hold, we obtain a noncommutative associative ring. That it is consistent to demand the preceding properties follows from the fact that they are satisfied by 2×2 matrices with entries in R if we represent e_0 by the identity matrix and e_j by $-i\sigma_j$, where σ_j are the three Pauli matrices:

$$\sigma_1=\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},\quad \sigma_2=\begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix},\quad \sigma_3=\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Thus, if we set $E_0=I$ and $E_j=-i\sigma_j$, we find that

$$X=x^s E_S=\begin{bmatrix} x^0-ix^3 & -x^2-ix^1 \\ x^2-ix^1 & x^0+ix^3 \end{bmatrix},$$

and that if $x^0=0$, then $\det(X)=\delta_{jk}x^j x^k$, which equals the square of the Euclidean length of the vector with components x^j.

If T is any invertible 2×2 matrix and $Y=TXT^{-1}$, then trace of $Y=\mathrm{tr}(Y)=\mathrm{tr}(X)$ and the determinant $\det(Y)=\det(X)$. Since $\mathrm{tr}(X)=x^0$, it follows that $x^0=0$ implies $y^0=0$. Further, $\delta_{jk}x^j x^k=\delta_{jk}y^j y^k$, that is, Euclidean distance is preserved so the transformation from (x^1,x^2,x^3) to (y^1,y^2,y^3) is orthogonal. In particular if $T = \exp(\vartheta\sigma_3) = \cos\vartheta I + \sin\vartheta\sigma_3$, this transformation is a rotation though an angle 2ϑ about the x^3 axis.

If R is a finite commutative ring with m elements, $H(R)$ would be a noncommutative ring with m^4 elements.

When is $H(R)$ a field? Define X' by $X = x^0 I + X'$ and $\bar{X}$ by $\bar{X} = x^0 I - X'$. It then follows that $X\bar{X} = \delta_{st}x^s x^t I$. If $R = \mathbb{R}$, this vanishes only if $X = 0$. Thus $\bar{X}$ divided by $\delta_{st}x^s x^t$ is the reciprocal of X. It is not difficult to verify that $H(\mathbb{R})$ satisfies the requirements of an anticommutative or skew field. This is the field discovered by Hamilton to which we alluded in Sec. 2.2.

(d) *Boolean "Ring".* In studying what he called "The Laws of Thought," George Boole was led to introduce an algebraic structure on the subsets of any fixed set in which union, $\cup$, and intersection, $\cap$, are analogs of addition and multiplication, respectively. The original set acts as the identity for multiplication, and the empty set serves as the zero for addition. The reader can verify that most of the properties of a commutative ring are satisfied by Boole's structure, but a given subset does not have an additive inverse so that $\mathscr{P}(S)$, the set of subsets of S, is not an additive group under the binary operation $\cup$.

(e) *Lie Rings.* Let $(L,+,\circ)$ be a set L together with two binary operators such that $(L,+)$ is an additive commutative group such that the operation $\circ$ distributes over addition, so that

$$x\circ(y+z) = x\circ y + x\circ z.$$

However, the Lie product is neither commutative nor associative but satisfies the properties:

$$x\circ y + y\circ x = 0,$$

and

$$x\circ(y\circ z) + y\circ(z\circ x) + z\circ(x\circ y) = 0.$$

Because of the first of these conditions, we say that the Lie product is anticommutative. The second, which replaces the associativity property of the familiar rings, is referred to as the Jacobi identity. Lie groups are discussed in other articles of this work so we do not go into details here. We merely remark that the elements of a finite dimensional Lie group can be parametrized by continuous real variables and that Sophus Lie associated to such groups what he called an infinitesimal group which is a particular case of a Lie ring. Associativity of multiplication in the group implies the validity of the Jacobi identity in the corresponding Lie ring.

The Jacobi identity can be rewritten in the form

$$z\circ(x\circ y) = (z\circ x)\circ y + x\circ(z\circ y),$$

which is the same as

$$D(x\circ y) = (Dx)\circ y + x\circ(Dy),$$

if we set $Dw = z\circ w$, for fixed z and all $w \in L$. This last equation reminds us of the product rule in calculus, so we say that the linear map D: $w \to z\circ w$ is a derivation of L. The concept of Lie ring, which apparently (Witt, 1937) was first defined by Wilhelm Magnus in 1936, is a generalization of the concept of Lie algebra introduced under the name "infinitesimal group" by Lie and Killing independently before 1880.

(f) *Grassmann Ring.* As a final example of the concept of ring we briefly describe an algebraic structure invented by Hermann Grassmann about 1840 which is basic to the theory of fermions as well as the geometry of many dimensions.

Given a field, $\mathbb{F}$, and a finite vector space $(V,\mathbb{F},+)$ of dimension n, it is possible to define a new vector space, $V^\wedge$, of dimension 2^n over $\mathbb{F}$ and a binary operation, denoted by $\wedge$, called the wedge or Grassmann product, which distributes over addition. $(V^\wedge,\mathbb{F},+,\wedge)$ will be the Grassmann or exterior algebra. In order to define the product $\wedge$, which is the same as that for fermion creation operators in second quantization, we proceed by induction on the grade of homogeneous elements of the algebra. Recall that in the ring $\mathbb{F}[x,y]$ of all polynomials in x and y there are special subspaces such as $ax+by$, or $ax^2+bxy+cy^2$, or $ax^3+bx^2y+cxy^2+dy^3$, of homogeneous elements of dimension 2,3,4, respectively. Any polynomial can be expressed as a sum of homogeneous polynomials, and the summands are unique. It turns out, analogously, that if $\dim(V)=n$, $V^\wedge$ contains $n+1$ subspaces V^p, $0\leqslant p\leqslant n$, such that any element x of $V^\wedge$ can be expressed in precisely one way as $x = \sum_0^n x^p$, where $x^p \in V^p$. An element of V^p is said to be homogeneous of grade p. If x and y are homogeneous of grades p and q, respectively, then $x\wedge y=(-1)^{pq}y\wedge x$ is of grade $p+q$. In particular, $V^0=\mathbb{F}$ and $V^1=V$ by definition. It follows that if x and y are of grade 1 , that is belong to V, $x\wedge y=-y\wedge x$. So if $\mathbb{F}$ has characteristic other than 2 it follows that $x\in V$ implies that $x\wedge x=0$.

If $\{v_i\}$ is a basis of V, the $n(n-1)/2$ elements $v_i\wedge v_j$ for $i<j$ are linearly independent and span the subspace V^2 of $V^\wedge$. Similarly V^3 is spanned by $v_i\wedge v_j\wedge v_k:=(v_i\wedge v_j)\wedge v_k=v_i\wedge(v_j\wedge v_k)$ with $i<j<k$ between 1 and n. The $\dim(V^3)=n(n-1)(n-2)/6$. Proceeding in this way we define all the $n+1$ homogeneous subspaces. As is known, the sum of the coefficients of the nth power of a binomial is $(1+1)^n = 2^n$, so $V^\wedge$ has dimension 2^n.

The preceding terse abstract definition does not immediately suggest that the Grassmann ring is significant for fermion physics. However, this becomes plausible when one realizes that the above basis elements of grade p correspond to the Slater determinants for a system of p electrons which can be formed from a basis set of n linearly independent spin-orbitals.

5.2 Polynomial Rings

For everyday applications there is little doubt that the integers $\mathbb{Z}$ constitute the most important ring which is not also a field. Perhaps the next most important is the ring of polynomials involving one or more variables. Suppose R is any ring; then we consider all expressions of the form $P(x)=a_sx^s$, where $0\leqslant s\leqslant n$, x^s denotes the sth power of the variable x, and $a_s\in R$. If $a_n\neq 0$ we say that $P(x)$ is a polynomial of degree n in x. The set of all such polynomials of arbitrary finite degree will be denoted by $R[x]$. (Note the square bracket which distinguishes the ring from the field $R(x)$ of rational functions.) Assume that the powers of x commute with the elements of R and define addition and multiplication in the obvious way. Then $(R[x],+,\times)$ is a ring which is commutative if and only if R is commutative. For example, if $R=\mathbb{Z}$, $R[x]$ is the ring of all polynomials with integer coefficients. If R is the ring of 2×2 matrices with complex entries, $R[x]$ consists of all 2×2 matrices whose entries are polynomials in x with complex coefficients. In this case the variable is often called λ. The theory of this particular ring is discussed by Turnbull and Aitken (1932), for example, under the title λ-matrices.

An obvious extension of the preceding is to adjoin two or more variables to R. Thus $R[x,y]$ denotes the set of polynomials in x and y with coefficients in R. A term such as $3x^2y^5$, formed by multiplication without addition, is called a monomial. The sum of the powers of x and y is called the degree of the monomial. Thus the degree of the preceding monomial is $2+5 = 7$. Clearly there are 8 different monomials of degree 7 in two variables. Any sum of these with coefficients in R is a homogeneous polynomial in x and y of degree 7. When R is a field we see that the homogeneous polynomials of degree 7 form a linear space of dimension 8.

More generally, it is of considerable interest to determine how many distinct monomials of degree n can be obtained from r variables. It is not difficult to see that the possible such monomials occur as the coefficients of t^n in the expansion of the r-fold product $\Pi(1-x_it)^{-1}$, where $1\leqslant i\leqslant r$ and x_i are distinct variables. Setting all $x_i=1$, we see that the required number is the binomial coefficient $\binom{r+n-1}{n}$.

This is an opportune point at which to explain the concept of a graded ring which appeared in Sec. 5.1(f) and has recently entered quantum physics in connection with supersymmetry. It is clear that any element of $R[x,y]$ is a unique sum of homogeneous terms and that the product of two homogeneous terms of degree p and q, respectively, is homogeneous of degree $p+q$.

A graded ring (R,Σ) is a ring together with a set of grades, Σ, closed under addition, such that R contains special homogeneous elements to which a grade from Σ is assigned; any element of R can be expressed in a unique manner as a sum of homogeneous elements; the product of two homogeneous elements of grade α and β is homogeneous of grade $\alpha+\beta$. For polynomial rings we usually take Σ to be the non-negative integers. For the so-called Laurent polynomials $\mathbb{C}[t,t^{-1}]$, Σ is the set $\mathbb{Z}$ of all integers. For a Grassmann ring of r generators, Σ is the set of non-negative integers. However in that case there are no terms of grade greater than r, and for $0\leqslant n\leqslant r$ the subspace of homogeneous elements of grade n has dimension $\binom{r}{n}$.

5.2.1 Binomial and Multinomial Theorem. For $r \in \mathbb{Z}$, $r \geqslant 0$, $(x+y)^r = C^{\,r}_{n} x^{r-n} y^n$, where n is summed from 0 to r. The binomial coefficients $C^{\,r}_{n}$ satisfy the recurrence relation $C^{\,r+1}_{n} = C^{\,r}_{n} + C^{\,r}_{n-1}$, with which starting from $C^{\,0}_{0} = 1$ we can generate the famous Pascal triangle.

Define $n! := 1 \times 2 \times 3 \ldots \times n$, which we read as "$n$-factorial" or "factorial n." $C^{\,r}_{n}$ is often denoted by $\binom{r}{n}$ and is given by

$$\binom{r}{n} = \frac{r!}{n!(r-n)!}.$$

In this and most other formulas 0! is interpreted as 1.

The binomial coefficient is also the number of subsets of n elements in a set of r elements, or the so-called number of combinations of r things taken n at a time.

In the preceding, r and n are non-negative integers but, as Newton realized, the binomial coefficient can be defined as follows for any real r:

$$\binom{r}{n} = \frac{r(r-1)(r-2)\cdots(r-n+1)}{1 \times 2 \times 3 \ldots \times n}.$$

Here the numerator is a product of n factors which begin with r and are decremented successively by 1. For example with $r = -1$, $\binom{r}{n} = (-1)^n$. Hence,

$$(1-x)^{-1} = 1 + x + x^2 + x^3 \ldots,$$

which is valid when the infinite sum exists, that is if the absolute value of x is less than 1. This form was used in the preceding section to obtain the number of distinct monomials of degree n in r variables, viz. $\binom{-r}{n}$.

The binomial coefficient is a particular case of the multinomial coefficient which arises in powers of sums of more than two variables. Thus,

$$(x_1 + x_2 + \ldots + x_n)^r = C_{r_1, r_2, \ldots, r_n} x_1^{r_1} x_2^{r_2} \ldots x_n^{r_n},$$

where $0 \leqslant r_i \leqslant r$ and the summation is over all $r_i \in \mathbb{N}$ such that $\Sigma r_i = r$. It is not difficult to see that

$$C_{r_1 r_2, \ldots, r_n} = r!/\Pi_i r_i!,$$

with the product for $1 \leqslant i \leqslant n$.

Like the binomial coefficient, to which it specializes when $n = 2$, this number has a combinatorial interpretation which explains its appearance in certain arguments of statistical mechanics. Suppose a set S of r distinct elements is partitioned into the union of n subsets S_i such that S_i contains exactly r_i elements; then there are $C_{r_1, r_2, \ldots, r_n}$ distinct ways in which such a partitioning can be effected.

5.2.2 Fundamental Theorem of Algebra. Since the square of any real number is positive or zero, there is no $x \in R$ such that $x^2 = -1$. Suppose there is a commutative ring R which contains $\mathbb{R}$ and also an element i such that $i^2 = -1$; then R contains all elements of the form $x + iy$, where x and y are arbitrary real numbers. As a ring, together with i, R will contain $-i$. If we require that R has no divisors of zero, since $x^2 + 1 = (x+i)(x-i)$, there are two and only two possible solutions of $x^2 + 1 = 0$, namely, i and $-i$. Further since $(x+iy)(x-iy) = x^2 + y^2 \neq 0$, any element of R of the form $x + iy$ has a reciprocal of the same form. In particular $i^{-1} = -i$. Thus the ring $\mathbb{R}[i]$, generated by the reals together with i, is in fact a field. It is of course the complex numbers which we are denoting by $\mathbb{C}$. Thus any commutative ring R which contains the reals and one solution of the equation $x^2 + 1 = 0$ contains $\mathbb{C}$.

We found $\mathbb{C}$ by starting with $\mathbb{R}$ and demanding that a simple second degree equation have a solution. If one starts with $\mathbb{Z}$ and asks for solutions of equations such as $5x + 7 = 0$, one soon realizes that one needs more numbers than the integers. This leads us to the rationals $\mathbb{Q}$. If, like the ancient Greeks, we ask for solutions of equations such as $x^2 - 3 = 0$, we are forced beyond the rationals and are led to define the reals, $\mathbb{R}$.

The extraordinary property of the complex numbers, first proved by Carl Friedrich Gauss, is that any equation of finite degree, $a_j x^j = 0$, $0 \leqslant j \leqslant n$, whose coefficients a_j belong to $\mathbb{C}$ has a solution which also belongs to $\mathbb{C}$. This result is so important that it has been called the Fundamental Theorem of Algebra. A field $\mathbb{F}$ which contains the zeros of all finite-degree polynomials which can be formed with coefficients in $\mathbb{F}$ is called algebraically closed, as we noted in Sec. 2.2.

If $P(x)$ is a polynomial of degree n with coefficients in a field $\mathbb{F}$ such that $P(\alpha) = 0$ for some $\alpha \in \mathbb{F}$, then one easily proves that $P(x) = (x - \alpha)Q(x)$, where $Q(x)$ is a polynomial of degree $n - 1$. If, like $\mathbb{C}$, $\mathbb{F}$ is algebrically closed, nothing prevents us from continuing this process so that $P(x)=c\Pi_i(x-\alpha_i)$, $0\leqslant i\leqslant n$, if P has degree n, and c is the coefficient of x^n, where the product is over the zeros of $P(x)$. If $\mathbb{F}=\mathbb{R}$ this process will not carry through in general; however, it can be shown that a polynomial with real coefficients can always be expressed as a product of factors of first or second degree with real coefficients.

The Fundamental Theorem of Algebra assures us of the existence of a zero for any polynomial $P(x) \in \mathbb{C}[x]$, but it does not give us an effective procedure for finding such a zero. Indeed, Evariste Galois showed that it is only for polynomials of degree less than five that an explicit formula, analogous to that for solving a quadratic equation, exists. It is worth looking in detail at equations of degree less than 5.

(i) $\deg[P(x)]=1$. Take $P(x)=ax+b$, where $a,b \in \mathbb{F}$ and $a\neq 0$. Then there is a unique zero, $x=-b/a$, which belongs to $\mathbb{F}$.

(ii) $\deg[P(x)]=2$. Take $P(x)=ax^2+2bx+c$, $a\neq 0$, $a,b,c \in \mathbb{F}$. Clearly $P(\alpha)=0$ is equivalent to $aP(\alpha)=0$, but

$$aP(x)=a^2x^2+2abx+ac=(ax+b)^2+ac-b^2.$$

Thus if α is a zero of P, $(a\alpha+b)^2=b^2-ac=D$. To find α it will be necessary to find the square root of D. Since by various choices of a, b, and c, D will vary over the whole of our ground field $\mathbb{F}$, it follows that quadratic equations with coefficients in $\mathbb{F}$ will always have solutions in $\mathbb{F}$ if and only if $\mathbb{F}$ is closed under the operation of taking square roots.

In elementary school we learn to manipulate with square and cube roots and come to take them quite casually. But note that even the world's largest super-computer is not able to find the square root of 2 exactly since it is a nonrecurring decimal. To find it requires an infinite process. In other words, moving from talking about square roots to actually calculating them takes us out of algebra as defined in the Introduction of this essay!

For equations of degree n, if $P(x)=a_jx^j$, $0\leqslant j\leqslant n$, $a_n\neq 0$, we could divide $P(x)$ by a_n, so there is no loss of generality in assuming that $P(x)$ is monic; that is, it has leading coefficient equal to unity. It then follows that if we replace x by $y-b$, where $nb=a_{n-1}$, the resulting polynomial in y has zero as the coefficient of y^{n-1}.

(iii) $\deg[P(x)]=3$. By the preceding argument we may assume that $P(x)=x^3+px+q$. It is easy to see that if $(x-\alpha)^2$ or a higher power of $(x-\alpha)$ divides a polynomial then $x-\alpha$ also divides its derivative. By seeking for a common factor of $P(x)$ and its first derivative we find that two or more of the zeros of $P(x)$ will be equal if and only if $4p^3+27q^2=0$.

This conclusion is valid for p and q belonging to any field, $\mathbb{F}$. If $\mathbb{F}=\mathbb{R}$ we know from a simple graphical argument that $P(x)$ has at least one real zero. If $p,q \in \mathbb{R}$, it is possible to show that the roots will all be real or that there will be one real and two conjugate imaginary roots according as $4p^3+27q^2$ is, respectively, less than or greater than 0. When it is zero all the roots are real and at least two are equal.

It is interesting to note that if $pq\neq 0$, by setting $x=k\cos\vartheta$, the solution of $x^3+px+q=0$ can be obtained by comparison with the identity $4\cos^3\vartheta-3\cos\vartheta-\cos(3\vartheta)=0$. With k and ϑ such that $3k^2+4p=0$ and $pk\cos(3\vartheta)=3q$, the three roots are $k\cos\vartheta$, $k\cos(\vartheta+2\pi/3)$, and $k\cos(\vartheta+4\pi/3)$. When $pq=0$, the solution is trivial.

(iv) $\deg[P(x)]=4$. In this case the solution of the equation $P(x)=0$ can be made to depend on the solution of a so-called resolvent equation of degree 3. Then four zeros are obtained as expressions involving two square roots of rather complicated combinations of the coefficients of $P(x)$ and one zero of the resolvent cubic equation. These formulas are of little theoretical value so we do not display them. Nowadays anyone wanting explicit values for the zeros would obtain them by a computer-implementable algorithm.

There are elegant and deep arguments connecting the solution of equations of degree 5 and 6 with the theory of elliptic functions.

Of course, even for arbitrarily high degree there are particular equations of simple form which have convenient explicit solutions. For example, $x^n=a$ has n solutions: $r\exp(\vartheta+2\pi ki/n)$, $0\leqslant k<n$, with $r>0$ and $r^n\exp(n\vartheta)=a$.

6. ALGEBRAS

The word "algebra" is used in two distinct senses. On the one hand it refers to the subject which the reader began to study with more or less enthusiasm around the age of 13. Recall that when Einstein asked his uncle

"What is algebra?" the latter replied "In algebra you let x be the unknown and then you find x." Bourbaki would not accept this as a satisfactory definition but it started Einstein on the right path. Who knows? Without that initial encouragement we would never have had General Relativity which contains a good deal of algebra!

But *an* algebra also refers to a rather specific type of mathematical structure: a ring, which is also a linear space over some field. It is therefore a very rich structure. As a linear space, an algebra has a dimension—finite or infinite. In quantum mechanics the observables generate an associative algebra of operators on an appropriate Hilbert space, which is finite or infinite over $\mathbb{C}$ depending on the physical system.

As we noted in Sec. 5.1, rings differ among themselves according to whether the "product" is or is not associative or commutative. Because an algebra has linear-space structure and therefore a basis, it is possible to characterize different types of algebras by their structure constants which are defined as follows. Let $\{e_i\}$ be a basis for the algebra over $\mathbb{F}$ and denote the product merely by juxtaposition. Then e_ie_j is a linear combination of the basis such that $e_ie_j= c_{ij}^ke_k$, where the coefficients belong to $\mathbb{F}$. It then follows by linearity that if $x=x^ie_i$ and $y=y^je_j$ are any elements of the algebra, the product:

$$xy=x^iy^jc_{ij}^ke_k=z^ke_k,$$

where

$$z^k=c_{ij}^kx^iy^j.$$

The nature of the product is determined by the structure constants c_{ij}^k. For example, if for all i and j they are antisymmetric in i-j, then $xy+yx=0$, as in a Lie ring [Sec. 5.1(e)].

In understanding and classifying groups there is a class called simple groups with a key role in the sense that they are the building blocks with which other groups are constructed. For complex Lie groups the classification of all possible simple groups was essentially completed by Wilhelm Killing in 1888. The solution of the analogous problem for finite groups was achieved only in 1980 as a result of gigantic efforts over decades by hundreds of mathematicians—undoubtedly one of the truly outstanding intellectual achievements of all time. A simple group is one that contains no normal subgroups other than the full group and the one-element subgroup consisting of the identity. Another way of defining a simple group G is to say that any homomorphism of G into another group either is an isomorphism or sends every element of G onto a single element. Finally, this is the same as saying that if we denote the homomorphism by π then its kernel, $K=\{g\,|\,\pi(g)=e\}$, is either G or $\{e\}$.

The concept of homomorphism—that is, structure-preserving maps—applies to rings and algebras as well as groups. If R and R' are rings and π: $R\to R'$ is a map of R into R' then π is a homomorphism if $\pi(x+y)=\pi(x)+\pi(y)$ and $\pi(xy)=\pi(x)\pi(y)$ for all x and y in R. These are the conditions for a ring-homomorphism. If R is in fact an algebra, then π will also have to satisfy the condition $\pi(\alpha x+\beta y)=\alpha\pi(x)+\beta\pi(y)$, where α and β are in $\mathbb{F}$, if π is to qualify as an algebra-homomorphism. If we then define $K=\{x\in R\,|\,\pi(x)=0\}$, we easily see that K is a subring (or subalgebra) of R. But more! If $x\in K$ and $z\in R$ then both xz and zx belong to K. A subring of R which satisfies this last condition is called an ideal. If K is the kernel of a homomorphism π, and we define an equivalence relation on R by $x\rho y\Leftrightarrow x-y\in K$, then the image $\pi(R)$ is isomorphic to R/ρ. Conversely, if K is an ideal in R there is a homomorphism of R onto R/ρ with K as kernel. A ring (or algebra) is simple if it has no ideals other than $\{0\}$ and itself.

The famous Wedderburn theorem which emerged between 1895 and 1905 asserts that a simple finite-dimensional associative algebra is isomorphic to a complete matrix algebra for some n—that is, the algebra of all $n\times n$ matrices which we considered in Sec. 5.1(b).

6.1 Examples of Algebras

The algebra of $n\times n$ matrices over some field $\mathbb{F}$ to which we have just alluded is undoubtedly one of the most important and useful types of algebras. We mention three other algebras which are widely applied by physicists.

(a) *Frobenius or Group Algebra.* Suppose G is a finite group of order n. We can decree that the elements g_i of G are a basis for a linear space over some field $\mathbb{F}$. Then the set $A=\{x^ig_i\,|\,x^i\in\mathbb{F}\}$ is closed under addition and also under multiplication by elements of $\mathbb{F}$ and therefore is a linear space of dimension equal to the order of G. If $x=x^ig_i$ and $y=y^ig_i$ are arbitrary elements of A, we can define $xy:=x^iy^jg_ig_j=\Sigma_kz^kg_k$, $1\leqslant i, j, k\leqslant n$, where $z^k=c_{ij}^kx^iy^j$. Here, c_{ij}^k is 1 if $g_ig_j=g_k$ and 0 otherwise. With these definitions of addition and multiplication A is an algebra over $\mathbb{F}$. It is named after the Berlin mathematician G. Frobenius who made basic contributions to the theory of group characters, differential equations, and other branches of mathematics.

There is a better way to display the product xy. Denote x^i by $x(g_i)$ so that $x=\Sigma_gx(g)g$ where the sum is over all $g\in G$. It then follows that $xy=\Sigma_gz(g)g$, where $z(g)=\Sigma_hx(h)y(h^{-1}g)$ for fixed g with summation on $h\in G$. Viewed in this way there is a bijection between the elements of A and functions $x(g)$ on G. The sum $x+y$ is mapped onto the function $x(g)+y(g)$, and the product xy onto the convolution, $\Sigma_hx(h)y(h^{-1}g)$, of the functions $x(g)$ and $y(g)$.

Now suppose that G is a continuous group on which there is a concept of volume or measure which is invariant under translation by elements of the group—a so-called Haar measure. Then it is possible to extend the notion of Frobenius algebra from finite to continuous groups where the elements of the algebra are functions on G, addition is point-wise addition of functions, and the product of $x(g)$ and $y(g)$ is the function $\int x(h)y(h^{-1}g)dh$. This convolution of functions occurs in the theory of Fourier and Laplace transforms. A rigorous treatment involves deep problems of continuity, convergence, and measure theory which analysts love. But the algebraic structure is apparent and explains why the idea of convolution is so useful.

(b) *Clifford Algebras* are generalizations of quaternions. They are the E-numbers of Sir A. S. Eddington's Fundamental Theory. They play a key role in relativistic quantum mechanics as the Dirac matrices. They are very useful in the discussion of the orthogonal group in n-dimensions.

Suppose that E_i, $1\leqslant i\leqslant n$, are elements of an associative algebra over the field $\mathbb{F}$ and satisfy the conditions

$$E_iE_j+E_jE_i=2\delta_{ij}I,$$

where the Kronecker delta is 1 if $i=j$ and 0 otherwise and I is the identity of the algebra. Since $E_i^2=I$ and distinct E's anticommute, it follows that the algebra generated by the E's has a basis consisting of I, E_i, E_iE_j, $E_iE_jE_k$,... $1\leqslant i<j<k<...\leqslant n$, and therefore like the Grassmann algebra has dimension 2^n. However, if $X=x^iE_i$ we see that $X^2=\delta_{ij}x^ix^jI$, displaying the Euclidean metric! It follows that if T is any fixed invertible element of the Clifford algebra, the transformation $X\rightarrow TXT^{-1}$ gives rise to a linear transformation of the coordinates x^i under which the metric is invariant. It is therefore an orthogonal transformation.

In the case $n=4$, $2^n=16$ which is the number of entries in a 4×4 matrix. Indeed, there is then a faithful representation of the Clifford algebra by 4×4 matrices among which the Dirac matrices appear in a natural manner.

(c) *Lie Algebras.* A Lie algebra $(L, \mathbb{F}, +, \circ)$, abbreviated as LA, is a Lie ring as defined in Sec. 5.1, which is also a linear space over some field $\mathbb{F}$. As such it could have finite or infinite dimension. The smallest subalgebra of L which contains all the elements of the form $x\circ y$, for $x,y\in L$, is an ideal in L, which we denote by $L\circ L:=L'$. Defining an equivalence relation ρ by $x\rho y\Leftrightarrow x-y\in L'$ leads to a quotient algebra L/ρ which is Abelian; that is the product of any two elements of L/ρ is 0. A Lie algebra is said to be solvable if the descending sequence of subalgebras $L^{p+1}\subset L^p$ where $L^0=L$, $L^1=L'$ and $L^{p+1}=(L^p)'$, terminates in $\{0\}$ in a finite number of steps. Every finite dimensional Lie algebra has a maximal solvable ideal, R, called the radical of L, which is such that the quotient algebra L/R is semisimple—that is, a direct sum of simple LA's. Recall that a simple ring is one with only trivial subideals. Thus simple and solvable Lie algebras are the basic building blocks needed for the analysis of any LA.

The complete classification of simple finite dimensional LA's over $\mathbb{C}$ was obtained by W. Killing (1888) and expounded clearly by Elie Cartan in his thesis (1894). The classification of simple finite dimensional LA's over $\mathbb{R}$ was achieved by Cartan (1913).

The concept of a simple LA was given a far-reaching extension by the Canadian R. V. Moody (1967) and the Russian Victor Kac (1967) quite independently of one another. These infinite dimensional algebras, which are now called Kac–Moody (K-M) algebras, quickly proved to play a key role in fundamental particle theory as does the closely related Virasoro algebra which can be thought of as a set of derivations of a K-M algebra

(Goddard and Olive, 1988). The most accessible K-M algebras are the so-called affine or Euclidean algebras which, together with the finite-dimensional simple algebras, can be classified rather conveniently by means of the Coxeter–Dynkin diagrams, which are exhibited in Table 1. In this table the left-hand column names the finite-dimensional simple algebras and the right-hand column, the affine algebras. These diagrams encode so much useful information that it is worthwhile explaining them in some detail.

Up to isomorphism, a K-M algebra is characterized by a square matrix (a_{ij}), called a Cartan matrix even though it was first defined by Killing, which has the following properties: (i) $a_{ij} \in \mathbb{Z}$, $1 \leqslant i,j \leqslant n$. (ii) $a_{ii}=2$. (iii) For $i \neq j$, $a_{ij} \leqslant 0$ and if $a_{ij}=0$ then $a_{ji}=0$. Then L is a LA generated by $2n$ elements $\{e_i, f_i\}$ satisfying the following conditions: (a) $e_i \circ f_j = h_i \delta_{ij}$; (b) $h_i \circ e_j = a_{ij} e_j$; (c) $h_i \circ f_j = -a_{ij} f_j$; (d) $h_i \circ h_j = 0$; (e) the h_i, $1 \leqslant i \leqslant n$, are linearly independent; (f) $\hat{e}_i^{1-a_{ij}} e_j = \hat{f}_i^{1-a_{ij}} f_j = 0$.

In (f) we use the notation $\hat{x} = ad(x)$ to denote a linear map of L into L defined by $\hat{x}\, y = x \circ y$. The h_i span an Abelian subalgebra, H, called the Cartan subalgebra of L of dimension n, which is also the rank of L. For $h = t^i h_i$, define $\alpha_j(h) = a_{ij} t^i$, so α_j is a linear function on H. Then (b) and (c) imply that $h \circ e_j = \alpha_j(h) e_j$ and $h \circ f_j = -\alpha_j(h) f_j$.

The algebra L is spanned by monomials which are products of the e's and f's but (a)–(d) imply that nonzero products involve only the e's or only the f's or belong to H. The algebra is graded by ascribing a grade α_i to e_i, 0 to $h \in H$, and $-\alpha_i$ to f_i. Hence the possible grades for nonzero elements of L are 0 or $\alpha = k^i \alpha_i$, where the k^i are all non-negative integers, in which case α is said to be positive and noted $\alpha > 0$, or where the k^i are all nonpositive integers, in which case α is negative or $\alpha < 0$. Thus L is a direct sum of homogeneous terms which have positive, negative, or zero grade. That is, $L = L^- \oplus H \oplus L^+$ where L^- is spanned by products of f's and L^+ by products of e's.

The grades form a lattice $k^i \alpha_i$ with $k^i \in \mathbb{Z}$. This lattice is an additive Abelian group, generated by α_i for $1 \leqslant i \leqslant n$, whose elements can be pictured as the points of a crystallographic lattice. The grades for which there is a nonzero homogeneous element of L are called roots and span a sublattice of the lattice of grades. The set of roots other than 0 is generally denoted by Δ which has the obvious partition into the positive and negative roots $\Delta = \Delta^- \cup \Delta^+$, with $\Delta^- = -\Delta^+$. A nonzero element x_α of L with grade α is called an α-root vector.

As mentioned above, the K-M algebras and the simple finite LA's are classified by means of the Coxeter–Dynkin diagrams (CDD) of Table 1. These were first employed by Coxeter (1931) in his study of finite groups generated by reflections and applied by him to LA's in 1934. They were also introduced independently by Dynkin (1946). Bourbaki named them after Dynkin, having first learned about them from Dynkin's important work concerning LA's. It was only in 1949 that Claude Chevalley, a founding member of Bourbaki, learned of Coxeter's papers from the author of the present article.

In 1888 or earlier, Killing noticed that the operation S_i: $\alpha \to \alpha - \alpha(h_i)\alpha_i$ effects a permutation of Δ such that $S_i^2 = I$. The n operations S_i, $1 \leqslant i \leqslant n$, generate a group which is now usually called the Weyl group because in 1923 Hermann Weyl popularized a particular representation of this group. I prefer to call it the Killing–Weyl group. A particular element of this group, $R = S_1 S_2 S_3 \ldots S_n$, which is usually called the Coxeter transformation, was used by Killing to effect the classification of the finite simple LA's over $\mathbb{C}$. He exhibited the orders of all the "Coxeter" transformations some years before Coxeter was born. It is perhaps only fair to follow B. Kostant's usage (1959) and call this operator the Killing–Coxeter transformation. The order of the R associated with a simple LA on the LHS of Table 1 is the sum of the digits appearing on the diagram immediately to the right. For example, the order of $R(G_2)$ is $1+2+3=6$. The Killing–Coxeter transformation is discussed in Coleman (1989), where additional references may be found.

The CDD gives the Cartan matrix (a_{ij}) and also the relations among the generators of the Killing–Weyl group for the corresponding LA. The nodes of the diagram are numbered by the indices 1 to n for the finite algebras and 0 to n for the affine algebras. A branch between i and j indicates that a_{ij} and a_{ji} are different from zero and, therefore, both are negative integers. A simple branch indicates that the product $a_{ij} a_{ji}$ is 1 and, therefore, each factor is -1; a double branch indicates that the

Table 1. Coxeter–Dynkin diagram of the finite and affine Lie algebras.

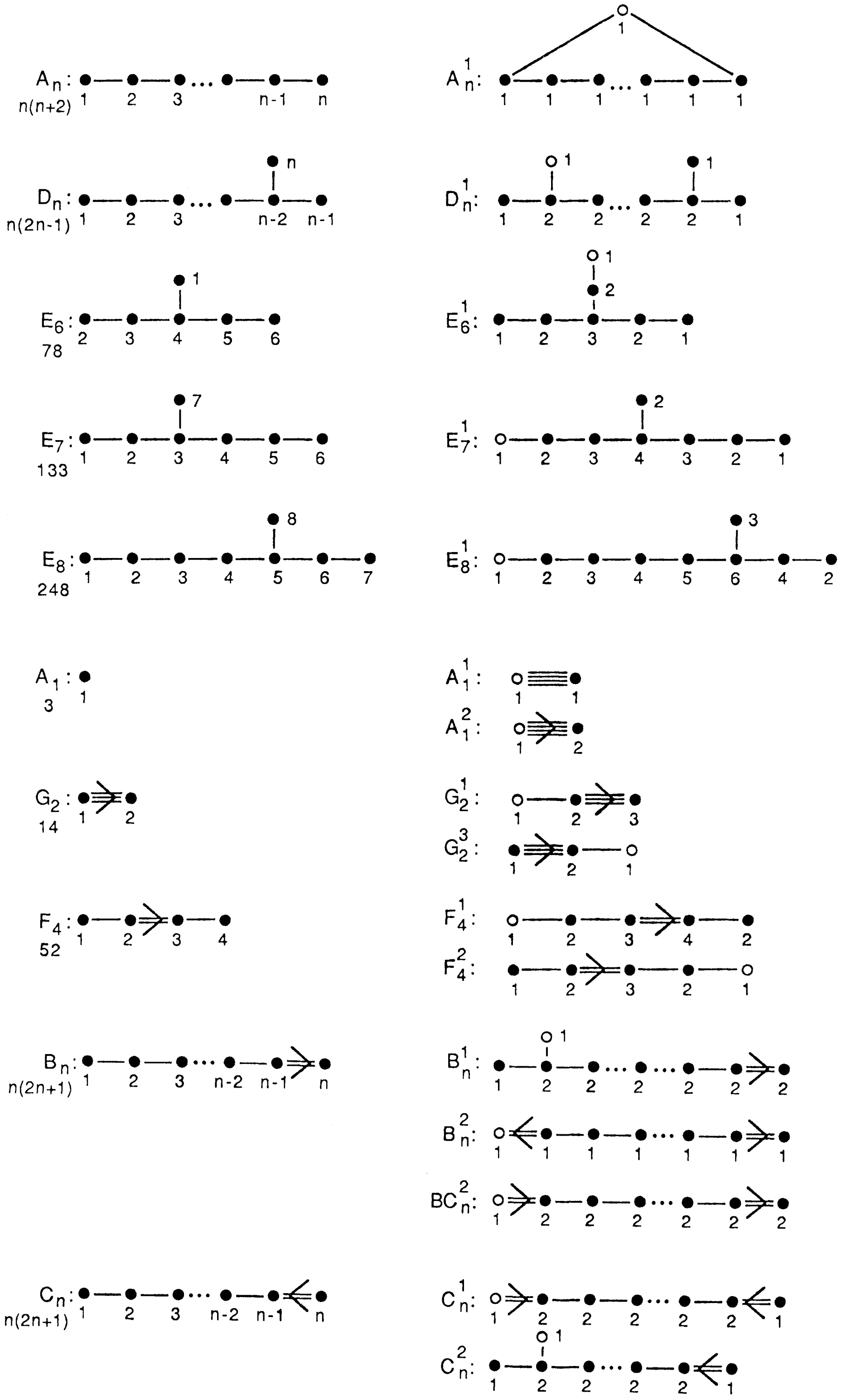
A_n : 1 2 3 … n-1 n
n(n+2)
A_n^1: 1 1 1 … 1 1 1; 1
D_n : 1 2 3 … n-2 n-1; n
n(2n-1)
D_n^1: 1 2 2 … 2 2 1; 1; 1
E_6 : 2 3 4 5 6; 1
78
E_6^1: 1 2 3 2 1; 2; 1
E_7 : 1 2 3 4 5 6; 7
133
E_7^1: 1 2 3 4 3 2 1; 2
E_8 : 1 2 3 4 5 6 7; 8
248
E_8^1: 1 2 3 4 5 6 4 2; 3
A_1 : 1
3
A_1^1 : 1 1
A_1^2 : 1 2
G_2 : 1 2
14
G_2^1 : 1 2 3
G_2^3 : 1 2 1
F_4 : 1 2 3 4
52
F_4^1 : 1 2 3 4 2
F_4^2 : 1 2 3 2 1
B_n : 1 2 3 … n-2 n-1 n
n(2n+1)
B_n^1 : 1 2 2 … 2 … 2 2 2; 1
B_n^2 : 1 1 1 1 … 1 1 1
BC_n^2 : 1 2 2 2 … 2 2 2
C_n : 1 2 3 … n-2 n-1 n
n(2n+1)
C_n^1 : 1 2 2 2 … 2 2 1
C_n^2 : 1 2 2 … 2 2 1; 1

product is 2, so the factors are -1 and -2; a triple branch indicates that the factors are -1 and -3. An arrow pointing from i to j indicates that a_{ij} is numerically larger than a_{ji}. Thus in the graph G_2, $a_{12}=-3$ and $a_{21}=-1$. The symbols A_n, B_n, C_n, D_n were introduced by Killing and Cartan to denote the four infinite classes of Lie algebras corresponding, respectively, to $SL(n+1)$, the general linear group on $n+1$ variables with determinant 1; the orthogonal group on $2n+1$ variables; the symplectic group on $2n$ variables; and the orthogonal group on $2n$ variables. E, F, G denote the exceptional LA's and the subscripts 2,4,6,7,8,n denote the rank of the algebra. For the finite algebras their dimension is noted in Table 1 under the name of the algebra. The left-hand column thus encodes the Cartan matrix for all the finite simple LA's.

But these diagrams also enable us to infer the relations among n or $n+1$ generators, S_i, of the Killing–Weyl group of the finite simple LA's or the affine algebras. $S_i^2 = I$, and for $i \neq j$, $(S_iS_j)p=\mathrm{I}$, where p equals 2, 3, 4, or 6 according as the i-node and the j-node are joined by a 0-, 1-, 2-, or 3-fold branch. For the simple LA's in the first column the determinants of the Cartan matrix and of all its principal subminors are strictly positive, whereas for the matrices in the second column the determinant of the matrix itself is zero but those of all principal subminors, positive. This corresponds to the fact that can be observed from the diagrams that if one node is removed from a diagram in the second column we obtain one or two diagrams of the first column. From this we infer immediately a class of finite semisimple LA's which are subalgebras of the affine algebras.

An open node is numbered 0 and the others retain their numbers from the diagram immediately to the left. The numbers attached to the nodes of the affine diagrams are the coefficients of the canonical null-root of the affine algebra. The affine algebras have infinite dimension, Killing–Weyl groups of infinite order, and an infinite number of roots. The dimension of a root space is finite and called the root multiplicity. Roots are distinguished according as they are in the orbit under the K-W group of a simple root α_i or not. The former are called real, the latter, imaginary. The real roots all have multiplicity 1, whereas imaginary roots can have arbitrarily high finite multiplicity.

By the height, ht(α), of the positive root $\alpha = k^i\alpha_i$ we mean Σk^i. A finite dimensional simple LA has a unique root of greatest height which can be read from Table 1. The numbers attached to nodes 1 to n in the diagram X_n^1 give the coefficients of α_i for the root of greatest height in the corresponding finite algebra X_n. Thus $\alpha = 2\alpha_1+3\alpha_2+4\alpha_3+2\alpha_4$ is the highest root of $\Delta(F_4)$.

We should note that some authors use the symbols $A_2^{(2)}$, $D_4^{(3)}$, $E_6^{(2)}$, $D_{n+1}^{(2)}$, $A_{2n}^{(2)}$, and A_{2n-1}^2 for our algebras A_1^2, G_2^3, F_4^2, B_n^2, BC_n^2, and C_n^2, respectively.

7. MODULES

So far we have not discussed exact sequences, commutative diagrams, or the game called diagram chasing which have played an increasingly important role in algebra in recent decades. The theory of modules provides a convenient context in which to introduce these ideas and is also significant in its own right.

Recall that a linear space is an additive group $(V,+)$, which is closed under multiplication by elements of a field $\mathbb{F}$ (usually commutative). We frequently denote the space $(V,\mathbb{F},+,\times)$ by V. A module is an additive group $(M,+)$ which is closed under multiplication by elements of a ring, R (often noncommutative). Frequently, when the context is clear, a module $(M,R,+,\times)$ is denoted by M. Essential conditions in the definition of an R-module M are that for $r \in R$, and $m_1,m_2 \in M$, (i) $rm_1 \in M$, (ii) $r(m_1+m_2) = rm_1+rm_2$.

7.1 Examples of Modules

(a) Suppose $R=\mathbb{Z}$ and $M=\{3n \mid n \in \mathbb{Z}\}=3\mathbb{Z}$. Clearly M is an additive group. Define the action of R on M as multiplication of integers. Obviously, if $n,m \in \mathbb{Z}$, $3n \in M$ and $m(3n) = 3mn \in M$.

(b) Let $M=\mathbb{Z}/3\mathbb{Z}:=\mathbb{Z}_3$. M has three elements which we could denote [0], [1], and [2]. For example $[1]=\{..-5,-2,1,4,7,...\}$. Take $R=\mathbb{Z}$. For $n \in R$ and $m \in \{0,1,2\}$, define $n[m]=[nm]$. For example, $7[2]=[14]=[2+3\times 4]=[2]$. We easily check that M is an R-module.

(c) For the physicist perhaps the most important example of a module occurs when R is the Frobenius algebra of a group G and M is a linear space over $\mathbb{R}$ or $\mathbb{C}$. In this case the action of R on M is called a linear representation of G. This example will be treated in other articles in this *Encyclopedia.* Of course it was via the representations of groups that Hermann Weyl and Eugene Wigner introduced the "group pest" into physics around 1930. The algebraization of physics then took a dramatic leap forward, and we were able to understand the periodic table, relativistic wave equations, and the classification of fundamental particles.

7.2 Morphisms of Modules

A mapping f: $M_1 \to M_2$ of one R-module into another could merely map one set into the other set. As such it could be (i) injective, such that $x \neq y \Leftrightarrow f(x) \neq f(y)$, (ii) surjective such that for each $y \in M_2$ there is an $x \in M_1$ for which $f(x)=y$, or (iii) bijective, that is both injective and surjective—or, one-to-one and onto. However, f might also have the property that for all $x,y \in M_1$, $f(x+y) = f(x)+f(y)$. We would then say that f is a homomorphism, or, now more frequently, simply, a morphism of additive groups. Even better, f might not only be a morphism of additive groups but also have the property $f(rm)=rf(m)$ for all $r \in R$ and $m \in M_1$. It would then have the distinction of being called an R-module morphism.

The kernel $K=\{x \in M_1 \mid f(x)=0\}$ of an R-module morphism f is an R-submodule of M_1. The image or range of f, $f(M_1) \subset M_2$, is an R-submodule of M_2 and is isomorphic to the quotient module M_1/K. When f is surjective, this situation is now frequently described by saying that the following diagram portrays an exact sequence of mappings:

$$0 \to K \to M_1 \xrightarrow{f} M_2 \to 0.$$

The preceding statement presupposes that the reader realizes that given our context (i) an arrow denotes a morphism of R-modules, (ii) the sequence is exact at K, M_1, and M_2. By exact is meant that the image of an incoming morphism is the kernel of the outgoing morphism. Hence, (i) the mapping from K to M_1 is injective since its kernel is 0, (ii) the mapping f is surjective since its image must be the whole of M_2 because this is the kernel of the final map. Since f is an R-module morphism the diagram implies that $M_2=f(M_1)=M_1/K$.

Though we chose to introduce the concept of an exact sequence for modules, it has wider application. Consider the diagram:

$$1 \to K \to G \to H \to 1,$$

where now the arrows represent homorphism of groups. Then if the sequence is exact, K is injected into G and can be identified with the normal subgroup of G which is the kernel of the map from G to H. We conclude that H is isomorphic to G/K. So this particular exact sequence encapsulates the familiar and basic First Isomorphism Theorem of group theory.

Returning to modules or additive groups for which the neutral element is denoted by 0, we see that for any exact sequence:

$$\dots \xrightarrow{\partial_{n-3}} M_{n-2} \xrightarrow{\partial_{n-2}} M_{n-1} \xrightarrow{\partial_{n-1}} M_n \xrightarrow{\partial_n}$$

$$M_{n+1} \xrightarrow{\partial_{n+1}} M_{n+2} \xrightarrow{\partial_{n+2}} \dots,$$

$\partial_{n+1}\partial_n M_n=0$, since $\partial_n M_n$ is the kernel of ∂_{n+1} and is mapped onto 0.

A topologist will inevitably recall that the boundary of a boundary is empty; a physicist will recall that the curl of a gradient is zero and that the divergence of a curl is zero. Indeed, these observations are the key to the algebraization of topology and of differential forms which led to homology and cohomology theory. The algebraic core of these topics has proliferated as Homological Algebra in which exact sequences and commutative diagrams are rampant.

Diagrams of maps can become quite complex and frequently admit two or more paths from object A to object B. If for all pairs A and B the compositions of maps along all different paths from A to B are identical, the diagram is said to be commutative. The game, art, or science of diagram chasing is the process of verifying that a given diagram is or is not commutative. Here, finally, is a simple but important example of a commutative diagram. It illustrates a method now widely used by algebraists to define universal objects.

Suppose G and $\hat{G}$ are given topological groups and $\pi: \hat{G} \to G$ a given homomorphism of $\hat{G}$ onto G. Suppose further that for any group H and any homomorphism τ such that $\tau(H)=G$ there exists a homomorphism σ which makes the accompanying diagram commutative.

$$\begin{tikzcd}
H & \hat{G} \arrow[l, "\sigma"'] \arrow[d, "\pi"] \\
 & G \arrow[from=1-1, "\tau"']
\end{tikzcd}\,.$$

Then $\hat{G}$ is the universal covering group of G, the kernel K of π is the Poincaré group of G, and the cardinality of K is the connectivity index of G. An example of this, which "explains" the occurrence of spin in physics, is that if G were SO(3) then $\hat{G}$ would be isomorphic to SU(2).

GLOSSARY

This article contains so many words which may be unfamiliar to the reader that if we gave their definition here the glossary would be almost as long as the article. Therefore, we list the most important concepts followed by the section number in which the concept can be found.

Abelian: see Sec. 1
Affine Lie Algebra: see Sec. 6.1(c)
Algebra: See Sec. 6
Algebraically Closed: see Sec. 2.2
Anti-Hermitian: see Sec. 3.6
Antisymmetric: see Sec. 3.6
Associativity: see Sec. 1
Automorphism: see Sec. 2.2
Basis: see Sec. 3.1
Bijective: see Sec. 7.2
Binary: see Sec. 1
Binomial Coefficient: see Sec. 5.2.1
Canonical Matrix: see Sec. 3.6
Cartan Matrix: see Sec. 6.1(c)
Cartan Subalgebra: see Sec. 6.1(c)
Chasing Diagrams: See sec. 7.2
Characteristic Function: see Sec. 3.5
Characteristic of a Field: see Sec. 2.1
Clifford Algebra: see Sec. 6.1(b)
Cogredient: see Sec. 3.7
Commutative: see Sec. 1
Commutative Diagram: see Sec. 7.2
Component: see Sec. 3.1
Connected Matrix: see Sec. 3.5
Contravariant: see Sec. 3.7
Contragredient: see Sec. 3.7
Contragredient Vector: see Sec. 3.7
Convolution: see Sec. 6.1(a)
Covariant: see Sec. 3.7
Covariant Vector: see Sec. 3.7
Coxeter–Dynkin Diagram: see Sec. 6.1(c)
Degree: see Sec. 2.3
Derivation: see Sec. 5.1(e)
Determinant: see Sec. 3.4
Diagonalizable: see Sec. 3.5
Dimension: see Sec. 3.1
Direct Sum: see Sec. 5
Distributivity: see Sec. 2
Domain: see Sec. 7
Dual Space: see Sec. 3.7
Eigenvalue: see Sec. 3.5
Eigenvector: see Sec. 3.5
Equivalence Relation: see Sec. 4
Exact Sequence: see Sec. 7.2
Field: see Sec. 2
Frobenius Algebra: see Sec. 6.1(a)
Fundamental Theorem of Algebra: see Sec. 5.2.2
Galois Field: see Sec. 2.1
Galois Group: see Sec. 2.2
Graded Ring: see Sec. 5.2
Grassmann Product: see Sec. 5.1
Grassmann Ring: see Sec. 5.1(f)
Group: see Sec. 1
Height of a Root: see Sec. 6.1(c)
Hermitian: see Sec. 3.6
Homogeneous Space: see Sec. 4
Homogeneous Subspace: see Sec. 5.1(f)
Homomorphism: see Secs. 3.3 and 6
Ideal: see Sec. 6
Identity Element: see Sec. 1
Imaginary Root: see Sec. 6.1(c)
Indecomposable: see Sec. 3.5
Injective: see Sec. 7.2
Invertible: see Introduction
Isomorphism: see Sec. 2.2
Jacobi Identity: see Sec. 5.1(e)
Jordan Block: see Sec. 3.5
Jordan Canonical Form: see Sec. 3.6(i)
Kernel: see Secs. 3.3 and 7.2
Killing–Coxeter Transformation: see Sec. 6.1(c)
Killing–Weyl Group: see Sec. 6.1(c)
Laurent Polynomials: see Sec. 5.2
Lie Algebra: see Sec. 6.1(c)
Lie Ring: see Sec. 5.1(e)

Linear Dependence and Independence: see Sec. 3.1
Linear Map: see Sec. 3.3
Linear Space: see Sec. 3
Matrix Ring: see Sec. 5.1(b)
Minimal Polynomial: see Sec. 3.5
Module: see Sec. 7
Monic: see Sec. 3.5
Monomial: see Sec. 5.2
Morphism: see Sec. 7.2
Multinomial Coefficient: see Sec. 5.2.1
Neutral Element: see Sec. 1
Nilpotent Matrix: see Sec. 3.5
Normal Subgroup: see Sec. 4
Nullity: see Sec. 3.3
Order of a Group: see Sec. 1
Partial Fractions: see Sec. 2.3
Pauli Matrices: see Sec. 5.1(c)
Polynomial of Degree n: see Sec. 5.2
Quaternion: see Secs. 2.2 and 5.1(c)
Quotient Set: see Sec. 4
Radical: see Sec. 6.1(c)
Range: see Sec. 3.3
Rank: see Sec. 3.3
Rational Functions $\mathbb{F}(x)$: see Sec. 2.3
Real Roots: see Sec. 6.1(c)
Reflexive Relation: see Sec. 4
Ring: see Sec. 5
Root: see Sec. 6.1(c)
Root Multiplicity: see Sec. 6.1(c)
Root Vector: see Sec. 6.1(c)
Scalar: see Sec. 3
Semisimple Lie Algebra: see Sec. 6.1(c)
Simple Group: see Sec. 6
Skew Field: see Sec. 5.1(c)
Solvable Lie Algebra: see Sec. 6.1(c)
Span: see Sec. 3.1
Stabilizer: see Sec. 4
Structure Constants: see Sec. 6
Sum of Maps: see Sec. 3.3
Surjective: see Sec. 7.2
Symmetric Matrix: see Sec. 3.6
Symmetric Relation: see Sec. 4
Tensor: see Secs. 3.5 and 3.8
Tensor Product: see Sec. 3.8 and 4
Transition Matrix: see Sec. 3.2
Transitive Group: see Sec. 4
Transitive Relation: see Sec. 4
Transpose: see Sec. 3.6
Unipotent: see Sec. 3.5
Universal Covering Group: see Sec. 7.2
Wedderburn Theorem: see Sec. 6

List of Works Cited

Cartan, E. (1984), *Sur la structure des groupes de transformation finis et continus*, Thèse, Paris: *Oeuvres Complètes* (1984) **I**, 137–257.

Coleman, A. J. (1989), *Invent. Math.* **95**, 447-477.

Coxeter, H. S. M. (1931), *J. London Math. Soc.* **6**, 132–136.

Dynkin, E. B. (1946), *Mat. Sbornik* **18**, 347–352.

Jacobson, J. (1964), *Lectures in Abstract Algebra* (3 Vols.), New York: van Nostrand.

Kac, V. (1967), *Func. Anal. Appl.* **1**, 328–329.

Killing, W. (1888–90), *Math. Annalen* **31**, 252–290; **33**, 1–48; **34**, 57–122; **36**, 161–189.

Kostant, B. (1959), *Am. J. Math.* **81**, 973–1032.

MacDuffee, C. C. (1943), *Vectors and Matrices*, Oberlin, OH: Mathematical Association of America.

Moody, R. V. (1967), *Bull. Am. Math. Soc.* **73**, 217–221.

Senata, E. (1973), *Non-negative matrices and Markov chains*, New York: Springer.

Turnbull, H. W., Aitken, A. C. (1932), *An Introduction to the Theory of Canonical Matrices*. London: Blackie and Sons.

Witt, E. (1937), *J. Reine Angew. Math.* **177**, 152–160.

Further Reading

Albert, A. A. (1956), *Fundamental Concepts of Higher Algebra*, Chicago: University of Chicago.

Bourbaki, N. (1942), *Algebra*, Paris: Hermann.

Goddard, P., Olive, D. (1988), *Kac-Moody and Virasoro Algebras*, Singapore : World Scientific.

Kac, V. (1985), *Infinite Dimensional Lie Algebras*, New York: Cambridge University.

Lidl, R., Niederreiter H. (1985), *Finite Fields*, Reading, MA: Addison-Wesley.

MacDuffee, C. C. (1943), *Vectors and Matrices*, Oberlin, OH: Mathematical Association of America.

Maclane, S., Birkhoff, G. (1967), *Algebra*, New York: MacMillan.

Maltsev, A. I. (1963), *Foundations of Linear Algebra*, San Francisco: Freeman.

Nash, C., Sen, S. (1983), *Topology and Geometry for Physicists*, London: Academic.

Shaw, R. (1983), *Linear Algebra and Group Representations*, London: Academic.

ALUMINUM

J. W. EVANS AND J. W. MORRIS, *Department of Materials Science and Mineral Engineering, University of California, Berkeley*

	Introduction	539
1.	**Properties and Characterization**	540
1.1	The Engineering Uses of Aluminum and its Alloys	540
1.1.1	Containers and Utensils	540
1.1.2	Construction	541
1.1.3	Aerospace and Transportation.	541
1.1.4	Electronics	541
1.2	Crystal Structure, Density, and Electron States	542
1.3	Electrical Conductivity	542
1.4	Mechanical Properties	543
1.4.1	Elasticity	543
1.4.2	Strength	543
1.4.3	Ductility	544
1.4.4	Fracture and Fatigue	545
1.5	Corrosion	546
1.6	Research Directions	546
2.	**Resources and Raw Materials**	547
3.	**Production of Aluminum**	548
3.1	The Bayer Process for the Production of Aluminum Oxide	548
3.2	The Hall–Héroult Cell for the Reduction of Oxide to Metal	548
3.3	Alternative Technologies for Producing Aluminum	551
4.	**Processing of Aluminum**	552
4.1	Alloying	552
4.2	Casting	552
4.3	Rolling and Extrusion	554
5.	**Environmental, Occupational, and Economic Issues**	555
5.1	Environmental Issues Related to Aluminum	555
5.2	Occupational and Legal Concerns	555
5.3	Economics of Aluminum	555
	Glossary	556
	List of Works Cited	557
	Further Reading	558

INTRODUCTION

The worldwide production of aluminum and its alloys exceeds that of any other metal except iron and its alloys. This economic significance and the widespread application of aluminum in transportation, defense, containers, utensils, construction, and electrical and electronic devices make aluminum one of the most important materials. Yet compared to steel, copper, zinc, and other common metals, aluminum is a relative newcomer. Worldwide production of aluminum at the beginning of this century was only six thousand tons per year, which is almost four orders of magnitude less than current production. The start of the commercial production of aluminum, a little over a century ago, was brought about by simultaneous discovery, by Hall in the USA and Héroult in France, of a suitable molten salt solvent for aluminum oxide, permitting the electrolytic generation of aluminum in what have become known as Hall–Héroult cells. This, coupled with a rapidly growing availability of electricity in the late nineteenth century, led to the establishment of the aluminum industry.

That industry has profited from the suitability of aluminum for a number of uses, ranging from cooking utensils through electrical wiring to aircraft alloys. This suitability arises from aluminum's low density, high electrical conductivity, and good corrosion resistance (see Table 1). The metal is usually regarded as having inferior mechanical properties to steel (at least in applications where density is a secondary consideration) but is superior to most other metals in this respect, particularly when alloyed with magnesium, silicon or lithium. The real cost of aluminum has declined by a factor of approximately one quarter in this century, but it is still significantly more expensive than other structural

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Table 1. Physical properties of aluminum. (Data from Hatch, 1984.)

Lattice constant	0.404 96 nm at 298 K
Density	2697–2699 kg m^{-3} (polycrystalline solid) at 298 K 2357 kg m^{-3} (liquid) at 973 K
Coefficient of expansion	23×10^{-6} K^{-1} at 293 K
Thermal conductivity	237 W m^{-1} K^{-1} at 298 K
Resistivity	2.655×10^{-8} Ω m at 293 K
Melting point	933.5 K
Boiling point	2767 K
Heat of fusion	397 kJ kg^{-1}
Heat capacity	0.90 kJ kg^{-1} K^{-1} at 298 K
Surface tension	0.868 N m^{-1} at 933.5 K
Viscosity	0.0012 kg m^{-1} s^{-1} at 933.5 K

materials (such as steel or reinforced concrete) in large part due to the amount of electrical energy used in its production. That energy has declined (on a unit mass basis) by a factor of more than one half over this century and further improvements can be anticipated. Nevertheless, expansion in the production of aluminum is occurring in countries with abundant supplies of inexpensive electrical energy.

The present article emphasizes the physical metallurgy and production of aluminum. Its scope includes the uses of this metal and its alloys; its physical and mechanical properties (crystal structure, density, electrical conductivity, elasticity, strength, ductility, fracture, fatigue, and corrosion); its occurrence in nature as bauxite, the production of aluminum oxide from bauxite by the Bayer process and the subsequent electrochemical reduction of the oxide in Hall–Héroult cells; alloying, casting, rolling and extrusion of the metal; and environmental, occupational, legal, and economic issues concerning aluminum.

1. PROPERTIES AND CHARACTERIZATION

Aluminum, in the broad sense of the metal and its alloys, is among the most widely used of engineering materials. The industry that is concerned with its manufacture and fabrication is among the largest of the basic metals industries. There are literally hundreds of aluminum alloys and product forms in daily commercial use. Their designations and properties are tabulated in a number of reference works [for example, Hatch (1985)]. This section is intended to provide a brief summary of the principal engineering uses of aluminum and its alloys, the properties that govern those uses, the microstructures that determine those properties, and the metallurgical characteristics of aluminum and its alloys that make it possible to achieve those microstructures.

1.1 The Engineering Uses of Aluminum and its Alloys

Most of the commercial applications of aluminum and its alloys exploit one or more of four important properties of aluminum: low density, ductility, corrosion resistance, and electrical and thermal conductivity. With some license the commercial applications can be gathered under the following four headings.

1.1.1 Containers and Utensils. This group of products includes such items as beverage cans, aluminum foil, cooking utensils, and small appliances. The first two, beverage cans and aluminum foil, constitute very large markets for aluminum; the current (1988) domestic consumption of beverage cans alone is approximately 80 billion units per year, of which 98% are presently made of aluminum. The property that permits aluminum alloys to dominate these markets is its ductility (formability), which allows it to be rolled into thin sheets or foil and deep-drawn into thin-walled containers. Density and corrosion resistance are also relevant; aluminum cans are lightweight, which minimizes transportation costs, and can be made relatively inert to most common foods and beverages so that the product does not acquire the taste of the container. Aluminum is attractive for cooking utensils because of its good thermal conductivity, low density, and corrosion resistance, and is attractive for both utensils and small appliances because of its formability. However, aluminum is chemically attacked by some cooking media, and may slightly contaminate the food. The most common problem is a slight metallic taste.

1.1.2 Construction. Aluminum alloys are widely used on the exteriors of buildings because of their resistance to corrosion. Their low density is also an advantage. For purely structural uses, however, their relatively high cost and low strength are competitive disadvantages with respect to structural steel and concrete.

1.1.3 Aerospace and Transportation. Aluminum alloys have been the dominant structural materials in both commercial and military aircraft for more than 50 years. They are advantageous with respect to other metals because of their low density, their formability, and their damage tolerance. In recent years aluminum alloys have been replaced in some important parts of aircraft structures by fiber-reinforced epoxy composite materials. These materials offer lower density, higher elastic moduli, and higher strength (at least in the fiber direction) but are relatively expensive and have low damage tolerance. The aluminum industry has countered the challenge of composites by developing new ingot alloys, such as the modern aluminum-lithium alloys, that offer lower density and higher moduli, and aluminum-matrix composite materials, such as Al-SiC whisker composites and aluminum-polymer fiber laminates, that provide a combination of the properties of aluminum alloys and fiber composites. Aluminum is less widely used in terrestrial vehicles, primarily because of its high cost relative to steel, but is used in an increasing variety of specific applications where its low density and corrosion resistance offer compensating advantages. There is a growing demand for structural materials for cryogenic aerospace systems, such as tankage for liquid hydrogen and oxygen fuel. Aluminum alloys are the dominant candidates for these structures because of their low density, their relatively low permeability and sensitivity to hydrogen, and their excellent toughness at cryogenic temperatures. There is also a growing demand for lightweight materials that can be used at moderate to elevated temperatures in automotive and aircraft engines and in the structures and external skins of hypersonic vehicles, which reach high temperature in flight. New aluminum alloys have been created for use at moderate temperature. There is active research on aluminum intermetallic compounds to achieve good structural properties at elevated temperatures.

1.1.4 Electronics. While the conductivity of pure aluminum is less than that of copper or gold, its good conductivity and relatively low cost have led to its wide use in electrical wiring. In addition, aluminum is the most common conductor for electrical lines on semiconductor chips. In this case it is deposited as a thin film either in pure form or lightly alloyed with copper or, sometimes, with copper and silicon. Aluminum is also widely used in electronic packaging, and in specific applications in microelectronics such as substrates for memory disks.

Pure aluminum is a superconductor with a maximum superconducting transition temperature (T_c) of 1.2 K, which is too low for practical applications. However, aluminum forms intermetallics with the A15 crystal structure, Nb_3Al and $Nb_3(Al,Ge)$ that have superconducting critical temperatures near 20 K and critical fields near 40 T, at 4.2 K, which are among the best properties obtainable with A15 superconductors. These compounds also offer exceptional critical current at 4.2 K in fields above 20 T. Their drawback is a manufacturing difficulty. The "bronze process" that is commonly used to manufacture multifilamentary wire of Nb_3Sn and V_3Ga is inapplicable to Nb_3Al and $Nb_3(Al,Ge)$. Rapid solidification and powder processes have been used with some success in the manufacture of Nb_3Al superconducting wire for use in the windings of high-field superconducting magnets.

While pure aluminum is not currently useful as a superconductor, its low residual resistivity at cryogenic temperatures makes it useful as a cryoconductor in conventional circuits at temperatures of 77 K or below. The combination of low density and high conductivity at low temperature make it particularly useful as a cryoconductor in electronic circuits in space vehicles.

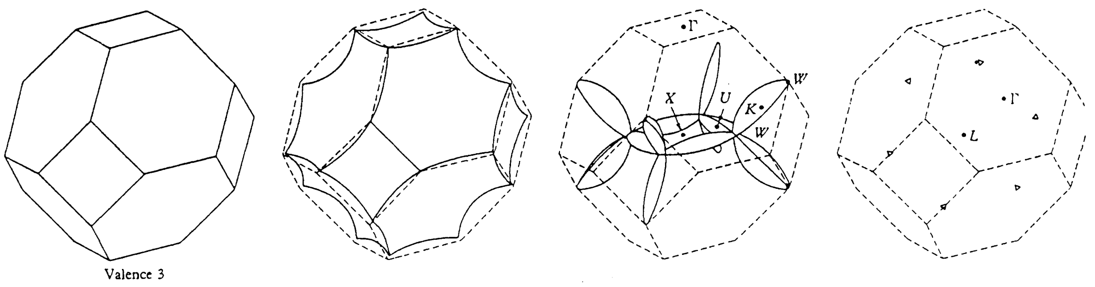

FIG. 1. Free-electron sphere for an fcc metal with valence 3, after Harrison (1960). The surfaces in the third and fourth zones are displaced by half a reciprocal lattice vector in the directions ΓX and ΓL, respectively, to better illustrate their forms.

As indicated by the applications listed above, the properties of aluminum and its alloys that have the greatest technological significance are its density, electrical conductivity, mechanical properties, and corrosion resistance. These are discussed in turn.

1.2 Crystal Structure, Density, and Electron States

Aluminum is the second element in group IIIB of the periodic table. Its crystal structure is face-centered cubic (fcc), both in the pure state and in all alloys in which it is the primary constituent. The lattice parameter of the fcc cell is 0.41 nm at 298 K, which corresponds to an atomic diameter of 0.29 nm. The density of pure aluminum is approximately 2700 kg m^{-3}, which is the lowest value of any cubic material that is useful as a structural metal. The density increases with the addition of most common alloying elements, but is decreased by alloying with the light elements Li, Be, and Mg, which makes these particularly desirable alloy additions in aerospace materials. The most effective of these alloy additions is Li; commercial Al-Li-Cu and Al-Li-Mg aerospace alloys have densities significantly below that of pure aluminum.

Aluminum behaves as a classic free-electron metal, and, as a consequence, is discussed in some detail as the prototype example of a trivalent metal in many treatments of the physical properties of metals (for example, Ashcroft and Mermin, 1976; Heine, 1969; Shoenberg, 1969; Alonso and March, 1989). In keeping with its nearly free-electron character the electron density of states is almost perfectly parabolic up to the Fermi energy. The electron states and many of the physical properties of the pure metal can be calculated with considerable accuracy using pseudopotential and other theoretical techniques (Harrison, 1966; Heine, 1969; Shoenberg, 1969; Hafner, 1989).

The Fermi surface of pure aluminum is almost, but not quite, a free-electron sphere, and is drawn and discussed by Ashcroft and Mermin (1976) and by Shoenberg (1969), among others. In an fcc crystal with valence 3 a free-electron sphere fills the first Brillouin zone, and partly fills the second, third, and fourth, as shown in Fig. 1. The portion of the Fermi surface in the third Brillouin zone forms an interconnected figure of thin, cigar-shaped elements, that is commonly called the "monster." The fourth Brillouin zone contains small, isolated elements of the Fermi surface. In fact, small deviations from ideal free-electron behavior in aluminum eliminate the segments of Fermi surface in the fourth Brillouin zone and break the "monster" in the third Brillouin zone into disconnected "rings of four," as shown in Fig. 2.

The complicated shape of the Fermi surface in the third Brillouin zone is reflected in a complex orientation dependence of the very low de Haas–van Alphen frequencies of aluminum (Shoenberg, 1969). The unfilled state of the second Brillouin zone dominates the high-field Hall effect in aluminum, which behaves as if it had one hole per atom rather than three electrons (Ashcroft and Mermin, 1976).

1.3 Electrical Conductivity

The electrical properties of aluminum are dominated by its crystal structure and metallic bonding, and resemble those of the fcc

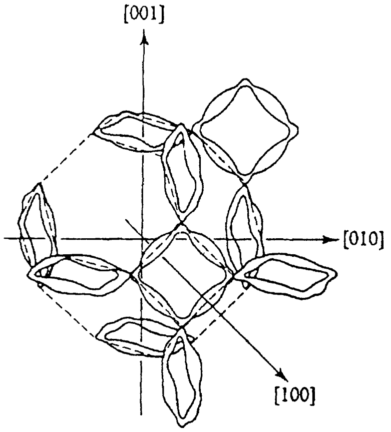

FIG. 2. The "ring of four" shape of the Fermi surface of Al in the third Brillouin zone (after Shoenberg, 1969).

metals Cu, Ag, and Au more than those of the nearby group IIIB elements B, Ga, and In. The electrical resistivity of ultrahigh-purity aluminum is 2.65 $\mu\Omega$ cm at 293 K, and varies approximately linearly with T with a slope $d\rho/dT \sim 0.011$ $\mu\Omega$ cm K^{-1} near room temperature. Its conductivity (reciprocal of the resistivity) is about 60% that of copper and 80% that of gold. As in all metals, the resistivity of aluminum increases with alloy content and is very sensitive to purity at low temperature. The resistivity is isotropic and remains nearly so in deformed samples; it is only slightly affected by grain size and dislocation density. The conductivity of aluminum is high enough to justify its use in electrical wiring, in which it is employed in nearly pure form to maximize its conductivity, and in thin film interconnects in microelectronic devices, in which it is often alloyed with copper or silicon.

The advantage of aluminum for thin film interconnects is the ease with which it can be deposited as a thin film and the relative ease with which it can be processed and etched into controlled complex geometries (Fraser, 1983). However, Al does form a eutectic with silicon. As a consequence it is often lightly alloyed with silicon to decrease reactivity at the interface or is deposited as a composite film that includes a thin diffusion barrier layer, such as W or TiN, that separates it from the silicon substrate. Aluminum interconnects are also liable to electromigration failures in service (d'Heurle and Ho, 1978). While the mechanism of electromigration failure is incompletely understood, it involves the rupture of the aluminum line through the growth of cracks or voids. The electric current contributes to void growth by increasing the effective self-diffusion constant. Electromigration failure is inhibited by the addition of copper, although the mechanism is not yet clear. Electromigration can also be inhibited by depositing a composite film in which thin layers of aluminum are separated by a diffusion barrier layer that prevents the line from rupturing by void growth. However, the increased thickness of the composite film limits miniaturization of the device.

1.4 Mechanical Properties

The important mechanical properties of aluminum alloys include their elasticity, strength, and ductility, and resistance to failure by fatigue and fracture. The mechanical properties of aluminum alloys govern both their formability into practical shapes and their mechanical performance in service and have, therefore, been the subject of extensive research and development.

1.4.1 Elasticity. Pure aluminum is nearly isotropic in its elastic properties. Its tensile (Young's) modulus at ambient temperature (E) is 69 GPa, which is about one-third the modulus of iron, and the value of Poisson's ratio is $\nu = 0.34$. The modulus is increased by most alloy additions to aluminum. However, the increase per unit of alloying element is relatively small except in the case of lithium; commercial Al-Li alloys have tensile moduli more than 10% above that of pure Al. The modulus can be increased further by introducing whiskers or fibers of high-modulus materials such as SiC.

1.4.2 Strength. Many of the structural uses of aluminum alloys are limited by a relatively low yield strength. It is difficult to increase the strength to values that resemble those of the steels and titanium alloys with which aluminum often competes, for two inherent reasons. First, the yield strength of a metal or alloy, with a given microstructure, scales very nearly with its elastic modulus. Since the elastic moduli of aluminum alloys are much lower than those of structural steels and Ti alloys, aluminum alloys are weaker even when they are tailored metallurgically to have similar microstructures. Second, the microstructures of high-strength steels and Ti alloys are generated, in part, through the creative use of the structural phase transformations that both materials undergo on cooling. There is no parallel transformation in the aluminum system.

The principal strengthening mechanisms that are available for use in aluminum alloys include solution hardening, work hardening, grain refinement, and precipitation hardening. In many alloys several mechanisms are used simultaneously.

The most important solution strengthening element in Al alloys is Mg. Mg is believed to strengthen the Al matrix in two ways: fixed Mg atoms act as obstacles to dislocation motion, and Mg also tends to segregate to dislocations, impeding their motion. All aluminum alloys can be strengthened by work hardening; however, the increased strength is often achieved at the price of a significant loss in ductility. Solution-strengthened Al-Mg alloys respond particularly to work hardening; the Mg addition inhibits the agglomeration of dislocations into cells or tangles during deformation so that it is possible to attain much higher dislocation densities. Al-Mn alloys, such as those used in the deep-drawing of beverage cans, also respond well to work hardening; their properties are due in part to a dense dispersion of Mn-based phases that form at high temperature and are retained.

Grain refinement is also used to strengthen Al alloys. Since there is no structural phase transition in the system the grain size is controlled by recrystallization treatments. However, since grain boundaries in Al alloys tend to be weak because of the accumulation of impurities or large precipitate particles, it is often preferable to adjust the alloy so that recrystallization and grain growth are impeded and a fine subgrain structure is created by recovery of the dislocations introduced at high temperature. The subgrains contribute to alloy strength very much as grain boundaries do. Alloy additions that form small, stable particles in the microstructure, such as Mn, Cr, Zr, and Sc, are particularly useful for inhibiting grain growth and controlling the grain and subgrain size after thermomechanical treatment.

Precipitation hardening is the most important of the strengthening mechanisms that are used in Al alloys. The alloys that can be hardened in this way are called "heat-treatable," and include most of the important structural and aerospace alloys. The common alloy additions that lead to controllable precipitates include Cu, Mg, Zn, and Li. While these alloying elements form equilibrium intermetallic phases with Al, the maximum strengthening is ordinarily not associated with the equilibrium phases, but with small, coherent metastable precipitates that appear at a relatively early stage in the precipitation process.

The most thoroughly studied of the precipitation sequences is that in Al-Cu. In the Al-Cu case the initial precipitate (and the one responsible for age hardening at room temperature) is a Guinier–Preston zone designated GP(1), which consists of disk-shaped copper-rich layers on {100} planes. On aging at higher temperature these particles are replaced by a different, multilayer zone, called GP(2) (whose structure is still not entirely clear), which is the predominant precipitate at peak hardness. Further aging creates the transient precipitate θ', which is a coherent modification of the incoherent, equilibrium precipitate phase $\theta(\mathrm{CuAl_2})$. A qualitatively similar behavior is observed in the other alloy systems. Al-Li is a special case that is hardened by the coherent, metastable precipitate $\delta'(\mathrm{Al_3Li})$, which is a well-defined phase in the strengthening precipitates at peak hardness. However, the high-strength aerospace alloys also contain significant concentrations of Cu, Mg, and Zr, and gain most of their strength from fine, coherent precipitates based on their combinations with Al and Li.

1.4.3 Ductility. To ensure ductility an alloy should have a high work hardening rate compared to its strength, because plastic instability occurs when the work hardening rate becomes comparable to the flow stress, and it should be free of large inclusions, precipitate particles, or zones of weakness that may initiate voids that cause early rupture. Solution-hardened alloys generally have good ductility, partly because of their relatively low strength. Precipitation-hardened alloys must be carefully engineered to achieve good ductility. The best properties are obtained when the alloy is aged below or to peak strength since over-aging leads to the formation of large, equilibrium precipitates, particularly in grain boundaries, that lead to premature rupture. The precipitates should be such that dislocations loop around them rather than shearing through them, since the planar slip that is associated with precipitate shearing leads to stress concentrations and relatively low work-hardening rates. It is also important to avoid precipitate-free zones near grain boundaries due to preferential precipitation on the boundary itself, since these constitute layers of weakness that lead to premature rupture. Given these considerations, precipitation-hardened alloys with good ductility are usually aged to near peak strength, adjusted in chemistry so that the dominant precipitate is mechanically hard and nonspherical in its morphology, and carefully heat treated or deformed prior to aging to eliminate precipitate-free zones.

One of the most important reasons to seek good ductility is in forming. Formability generally requires isotropy in plastic deformation; anisotropy causes wrinkling or "earing" (uneven elongation along the edges) in formed parts. The isotropy of plastic deformation is controlled by the crystallographic texture, or tendency for the grains of the metal to align their crystallographic axes. There are two important sources of texture in Al alloys: deformation, which invariably forces some crystallographic correlation between the grains of a polygranular microstructure, and recrystallization, in which new grains grow with a preferential crystallographic alignment to one another. The recrystallization and deformation textures are generally different, and can both be modified by altering the chemical composition. Moreover, the final texture can be adjusted by careful processing that mixes recrystallization and deformation steps so that their textures balance; this is a major objective of the rather elaborate processing that is used in the manufacture of Al sheet for beverage cans.

Several important aerospace aluminum alloys can be processed so that they become superplastic at elevated temperature and can be formed directly into complex shapes. The superplastic state is achieved by modifying the microstructure to obtain a very fine grain size that is preserved during deformation. In Al alloys this is most easily accomplished by adding alloying elements such as Cr, Ca, Zr, or Sc that create small, stable particles that inhibit grain boundary migration and, hence, stabilize a very fine grain size in recrystallized material.

1.4.4 Fracture and Fatigue. The fracture toughness of a material is controlled to first order by its fracture mode. Typical aluminum alloys fracture in one of three basic modes: ductile rupture by the formation and coalescence of voids in the microstructure, ductile rupture through plastic instability in shear, or intergranular fracture through separation on weakened grain boundaries. In the latter case the toughness is very low. Alloys that exhibit this fracture mode must be metallurgically modified to eliminate it in order to obtain high toughness. Intergranular fracture can be minimized by eliminating contaminants, such as the alkali metals, that embrittle the grain boundaries, by controlled processing to prevent the formation of precipitate-free zones or large precipitate particles along the grain boundaries, or by elongating the grains so that no continuous intergranular paths exist perpendicular to the applied load.

When the fracture mode is basically ductile the toughness of the alloy is controlled, first, by the distribution of inclusions that may nucleate voids. The predominant inclusions are large second-phase particles that are particularly likely to reside at grain boundaries. Since the most prominent and deleterious of these often include Fe and Si, the contents of these elements are carefully controlled in alloys that are designed for high fracture toughness. A second major source of inclusion particles that decrease toughness is the equilibrium precipitation that occurs on grain boundaries, especially in overaged alloys. To inhibit intergranular precipitation, alloys that are meant to have high toughness are often used in a slightly underaged condition. Precipitate-free zones near grain boundaries can also lead to early failure, and are minimized by mechanical deformation prior to precipitation to create a dense distribution of preferential nucleation sites near the boundary. When sources of premature fracture are eliminated the toughness is controlled by the inherent ductility of the alloy, which can be influenced by modifying the composition and the type and distribution of the hardening precipitates to control the strength and the work hardening rate.

The fatigue properties of aluminum alloys differ from those of steel in that Al alloys do not exhibit a fatigue limit in cyclic fatigue tests on smooth-bar specimens. Hence fatigue-resistant structures of Al alloys must be designed against crack growth to failure. The fatigue crack growth behavior of Al alloys is somewhat sensitive to composition; for example, Al-Cu-Mg alloys exhibit somewhat lower fatigue crack growth rates than Al-Zn-Mg alloys and are preferred for certain applications in aerospace systems for this reason. However, the most significant improvements in fatigue crack growth resistance occur in alloys, such as several of the Al-Li grades, whose microstructures are anisotropic and tailored so that propagating fatigue cracks blunt or branch at microstructural barriers such as perpendicular grain boundaries. The fatigue crack growth rate at very high stress intensity values is controlled by the fracture toughness.

Aluminum-based structural composites can have very good crack growth resistance when the second phase in the composite blunts or deflects growing cracks. New composites in which aluminum and polymer layers alternate have particularly good fatigue resistance.

1.5 Corrosion

Many of the important uses of Al alloys in structures and containers are motivated by its outstanding corrosion resistance in most common environments. It is important to recognize, however, that the corrosion resistance of aluminum is not due to any inherent inertness of the metal itself (it is among the most reactive of all metals) but to the fact that it forms a very stable and impermeable coherent oxide coating on exposure to the environment which shields it from further attack. In effect, aluminum paints itself. While this behavior makes aluminum alloys very suitable for use in many applications, it has the consequence that the corrosion resistance of aluminum may be poor or even very bad in situations in which the protective layer cannot form or cannot protect. Al alloys can be severely corroded by highly acidic or basic solutions, by chemicals such as Hg and its salts that can prevent the formation of the protective oxide film, and by strong galvanic couples with dissimilar metals.

Even relatively mild corrosion can produce cosmetic defects in appliances and structural parts, and can lead to failure by perforation or by stress corrosion cracking of parts that are subjected to high loads in corrosive environments. These problems can be addressed by the usual methods of painting, coating, cathodic protection (with Mg), or electrical insulation to break galvanic couples. However, there are also metallurgical solutions which are rather specific to aluminum alloys.

The most serious corrosion problem in Al alloys is the intergranular corrosion of high strength alloys that can lead to mechanical failure by stress corrosion cracking. Intergranular corrosion is most severe when a galvanic couple is established between the grain boundary and the grain interior. The problem is usually associated with the formation of precipitate phases in the boundary. Solution-strengthened Al-Mg alloys are relatively immune to intergranular corrosion since they have few intergranular precipitates. Al-Mn alloys are also relatively immune since the intergranular precipitates do not cause a significant galvanic couple. Precipitation-hardened Al-Mg alloys are treated to minimize intergranular attack by decreasing precipitate size and dispersing precipitates through the matrix. This approach is not as successful in Al-Cu alloy; copper in solution makes Al cathodic to Al-Cu precipitates in the grain boundary. Intergranular corrosion in Al-Cu alloys can be minimized by overaging the alloy to deplete copper from the grain interior, as is done in some alloys, but this procedure causes a loss of strength, ductility and toughness. High-strength Al-Cu alloys are hence often protected by a method known as "Alcladding," in which a thin layer of a relatively anodic Al alloy is plated onto the surface to provide cathodic protection.

For other applications Al alloys are protected by coating. A common method is called anodizing, and involves a surface treatment in an electrolytic cell to thicken the natural oxide coating. Since the visible color of the oxide coating can be adjusted by controlling its thickness, the anodizing process has cosmetic value as well.

1.6 Research Directions

Important directions for current research in the physical metallurgy of aluminum and its alloys derive primarily from the need for light structural alloys for use in aerospace vehicles and improved conductors for microelectronic devices. New, low-density structural materials that can perform structurally at the extremes of high and low temperature, and new conductors that can withstand exceptionally high current densities are particularly required. Much of the required research is applied and very specific to particular material requirements. However, there are important needs for supporting fundamental research. Many important research directions can be identified. Among them are three that specifically require development of the corresponding physics.

Intermetallic compounds based on aluminum are the critical hardening phases in precipitation-hardened aluminum alloys. They are also increasingly important as monolithic materials for use as low-density structure materials at high temperature. The structure and stability of the intermetallic phases are important for both applications, and new theoretical understanding is needed to guide development. In the case of hardening phases, these often have metastable structures or shapes that appear because of the elastic constraint of the matrix or preferential kinetics of formation, and hence are difficult to predict or interpret from equilibrium data. In the case of the monolithic intermetallics, many of the most attractive from the perspective of density and stability have noncubic structures that makes them relatively brittle. Theoretical developments are needed to guide the search for more attractive intermetallic phases.

The deformation of aluminum alloys governs strength, formability, and fracture under most conditions. The critical mechanism that governs deformation at moderate to large strains is work hardening. The hardening of a polygranular material during deformation is known to be influenced by both the evolution of the dislocation distribution within it and the development of texture in its polygranular structure. Work hardening is a complex, statistical, nonequilibrium process, and very little progress has been made in its quantitative theory. Probative research is needed.

The reliability of aluminum lines in microelectronic devices is strongly affected by the phenomenon of electromigration, which leads to failure by creating a break in the line. Since electromigration life tends to decrease as a line is made smaller or the current density is increased, the engineering problem of electromigration failure increases in severity with device miniaturization. A sound mechanistic theory of electromigration failure is needed both to guide design for reliability and to suggest material modifications that might improve reliability. The fundamental mechanisms of electromigration are poorly understood, and there is no established theory of electromigration failure. Research is badly needed in this area.

Table 2. World bauxite resources (only countries with largest resources, millions of metric tons).

Current economic reserves	
Jamaica	2 000
Brazil	2 250
Guyana	700
Surinam	575
Greece	600
Cameroon	680
Guinea	5 600
India	1 000
Indonesia	750
Australia	4 440
World total	21 000

2. RESOURCES AND RAW MATERIALS

The principal ore of aluminum is bauxite, which is used as a term for deposits containing gibbsite ($Al_2O_3 \cdot 3H_2O$), boehmite, or diaspore (both $Al_2O_3 \cdot H_2O$). Although this ore gets its name from Les Baux in Provence, large deposits are typically found in tropical regions. Aluminum (in the form of oxides and silicates) is present in the Earth's crust at a mean level of 8.3% and is the most abundant metal. However, leaching of other crustal constituents (principally silicon compounds) by humic acid, produced in tropical conditions by abundant rainfall and vegetation, leads to much higher levels of aluminum in bauxite ores (typically in excess of 30% Al_2O_3).

Table 2 provides data on the world's resources of bauxite, estimated by the U.S. Bureau of Mines for 1985 (U.S. Bureau of Mines, 1985). Major mining activities for bauxite take place in Australia, Guinea, Brazil, and Jamaica.

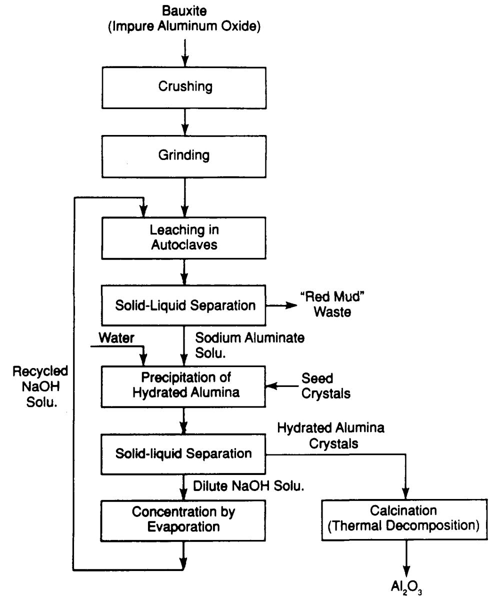

FIG. 3. Schematic diagram of the Bayer process for the production of aluminum oxide from bauxite.

3. PRODUCTION OF ALUMINUM

3.1 The Bayer Process for the Production of Aluminum Oxide

The starting point for the production of nearly all of the world's supply of new (as opposed to recycled) aluminum is the ore bauxite, the occurrence of which was described in the previous section. Bauxite is hydrated aluminum oxide containing oxides of silicon, iron, titanium, and other elements as impurities. The refining of this ore to aluminum oxide of greater than 99% purity is carried out in the Bayer process which relies on the selective dissolution of the aluminum oxide into a concentrated sodium hydroxide solution. Having been separated from insoluble impurities, this solution is then diluted with water to precipitate hydrated aluminum oxide.

Figure 3 is a schematic diagram of the Bayer process. The ore is first crushed and ground to a particle size (typically below 1 mm) to promote rapid dissolution of the aluminum oxide, following which it is contacted in autoclaves with a recycled concentrated solution of sodium hydroxide. At typical autoclave temperatures (200–240 °C) the reaction

$$Al_2O_3{\cdot}xH_2O+2NaOH$$
$$=2NaAlO_2+(x+1)H_2O \qquad (1)$$

proceeds to completion in approximately 100 min. The sodium aluminate solution is then separated from the undissolved impurities by decantation and filtering. The solid residue, known as "red mud," poses a disposal problem that is discussed in Sec. 5.1. The dissolution of aluminum oxide in sodium hydroxide solution can be reversed by cooling and diluting the solution and the precipitation of the solid hydrated aluminum oxide is nucleated by the addition of particles of the oxide. The precipitate is separated by filtration and dried in rotary kilns or fluidized beds while the remaining solution (largely sodium hydroxide) is concentrated by evaporation and recycled to the autoclaves. The product of this Bayer process is aluminum oxide powder (alumina) with a particle size usually in the range 10–200 μm. The principal impurity is sodium oxide (or aluminate) at the level of approximately half of 1%. The crystalline form is gamma alumina with some alpha alumina depending on the temperature reached in the kilns.

3.2 The Hall–Héroult Cell for the Reduction of Oxide to Metal

The most important step in the production of aluminum, in terms of economics, is the electrolytic reduction of the oxide produced by the Bayer process to metal. This reduction is carried out in cells invented simultaneously in the 1880s by Hall in the USA and Héroult in France. While Hall–Héroult cells have evolved over the intervening decades, particularly with regard to size and energy efficiency, the technology would still be recognizable by these inventors. Figure 4 (adapted from Jarrett, 1983) charts the progress of this technology and it is seen that much less energy is now required to produce unit mass of aluminum than early in the century. Nevertheless, in recent years Hall–Héroult cells have accounted for approximately 4% of the electricity generated in the U.S.A. with even higher percentages in some other countries. This electrical energy cost is reflected in the price of newly produced aluminum and has stimulated much research and development aimed at reducing this energy consumption.

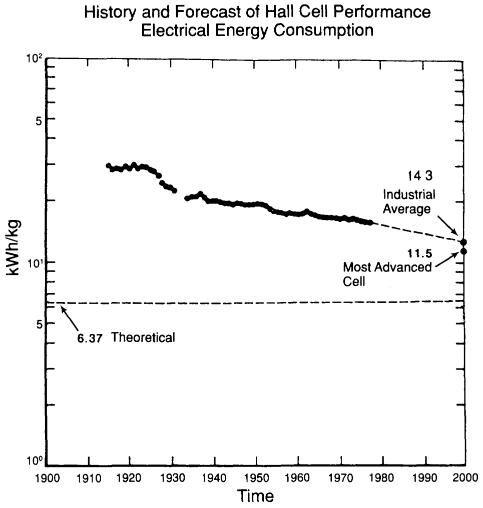

FIG. 4. History and prediction of the consumption of electrical energy in the production of aluminum (adapted from Jarrett, 1983).

The scope for improvement is revealed by the fact that the thermodynamic requirement for the reduction is only 40%–50% of the energy actually expended.

Figure 5 depicts a typical Hall–Héroult cell in cross section. The width of the cell might be three meters, the length (in the direction perpendicular to the paper) ten meters and the cell current 200 kA (dc). The voltage across the cell would be somewhat over 4 V and the cells, connected in series, are arranged in long double rows (usually with their long sides adjacent) known as "potlines." A typical potline would contain over a hundred cells with a representative aluminum plant having a few such potlines and generating 100 000–200 000 metric tons per year of metal.

Current flows from an adjacent cell via aluminum conductors and steel rods to carbon anodes dipping into the molten salt electrolyte. This electrolyte is a mixture of sodium and aluminum fluorides containing dissolved aluminum oxide and additives such as magnesium and lithium fluorides which, as will be described below, are intended to increase the current efficiency of the cell. The sodium and aluminum fluorides are in a ratio approximating the naturally occurring mineral cryolite (Na_3AlF_6) but additional aluminum fluoride is introduced to lower the freezing point to approximately 950 °C (and yet further with more additives) (see the phase diagram, Fig. 6). A typical operating temperature for a cell would be 960 °C and the electrolyte would contain a few percent of dissolved aluminum oxide.

As the current passes through the surface of the carbon anode into the electrolyte the oxidation of carbon to carbon dioxide occurs. The oxidation product expected on thermodynamic grounds (carbon monoxide) is produced in only small amounts because, at the current densities in question (0.5–1 A cm^{-2}), the conditions are far from electrochemical

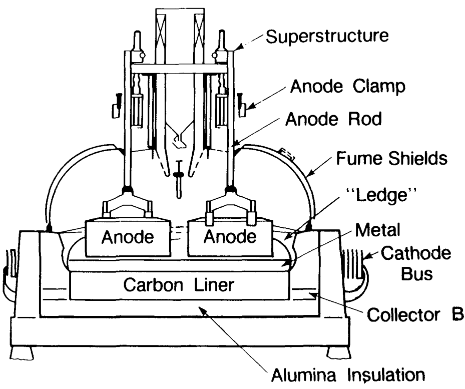

FIG. 5. Cross-sectional drawing of the Hall–Héroult cell for the production of aluminum.

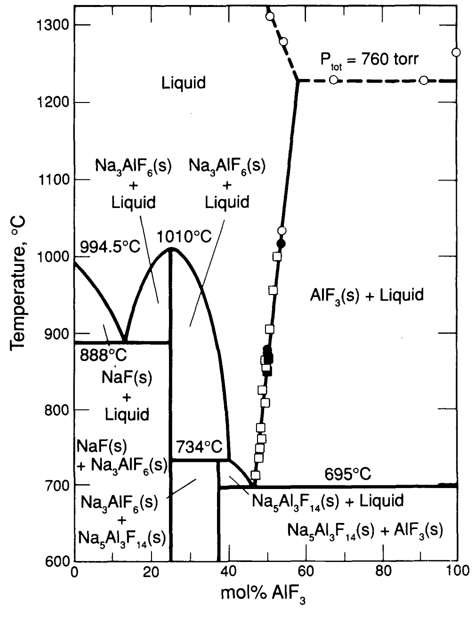

FIG. 6. Phase diagram for the system NaF-AlF_3 (from Grjotheim and Welch, 1980).

equilibrium. The conduction through the electrolyte is ionic and terminates with the current flowing into a pool of molten aluminum beneath the electrolyte. At the electrolyte-metal interface the cathodic reduction of aluminum ions (or, more correctly, fluorinated complexes of aluminum ions) occurs, producing the metal. The current then flows through a carbon lining into steel "collector bars" positioned in the lining and thence via bus bars to the next cell in line.

The overall reaction within the cell is

$$2Al_2O_3+3C=3CO_2+4Al. \qquad (2)$$

Consequently, the anodes are consumed and the aluminum pool grows in depth. Part of the pool is siphoned from the cell at intervals of one or two days. The cell of Fig. 5 has "prebaked" anodes manufactured in a separate operation by baking mixtures of carbon and carbonaceous binding agents. A cell would have 10–20 such anodes and it would be usual to replace the oldest one each day. In an alternative technology ("Soderberg" or "self-baking" anodes) one massive carbon anode is used and is formed continuously in the cell by the baking of carbonaceous materials added to the top of the anode. In both types of cells it is necessary to adjust the anode-cathode distance frequently to compensate for anode consumption, metal removal etc. and this is achieved by raising or lowering the superstructure from which the anodes are suspended.

Alumina is fed onto a frozen crust on top of the electrolyte and provides thermal insulation as well as trapping some of the fluoride vapors emanating from the electrolyte. Periodically the crust is broken, allowing alumina to fall into the electrolyte. Alumina is denser than aluminum and can sink beneath the metal if an excess is fed to the cell, with undesirable interference with the current path. Consequently, the oversupply of alumina to the cell is to be avoided. Neither on-line measurement nor routine laboratory analysis of samples for alumina have proved practical and until recently the alumina content of the electrolyte was controlled by periodically allowing it to approach zero. At low alumina contents, the wetting of the anode by the electrolyte diminishes to the point where carbon dioxide bubbles generated at the anode cover large parts of the anode, greatly increasing the cell resistance. The cell voltage then rises rapidly to several tens of volts, a phenomenon known as an "anode effect." The operator is then aware that the alumina concentration in the electrolyte is near zero and can extinguish the anode effect by (usually) feeding alumina. More recently computer-based data acquisition systems have been used to monitor cell voltages and currents and the alumina fed, so that anode effects can be avoided. Apart from this and intermittent measurement of cell temperature, the cells are not instrumented.

The inefficiency of Hall–Héroult cells has already been mentioned; it is convenient to discuss it in terms of two topics: current inefficiency and excess cell voltage. The current efficiency (the ratio of actual aluminum production to that expected from Faraday's electrochemical law) of a modern cell would typically lie in the range 90–95%. A small part of the current inefficiency can be accounted for by electronic conduction paths paralleling the ionic one through the electrolyte. However, the major cause of inefficiency is due to the "dissolution" of part of the aluminum product into the molten salt electrolyte and its subsequent reoxidation at, or near, the anode. The nature of this "dissolution" is still unclear and is likely to involve chemical reaction of the metal with the electrolyte to produce sodium and subvalent aluminum species (Grjotheim and Welch, 1980). However, laboratory measurements show solubilities in the range 0.05–0.15 wt. %, increasing with increasing temperature and with ratio of sodium fluoride to aluminum fluoride in the electrolyte. Improvements in current efficiency have been sought by using electrolyte compositions (and temperatures) that reduce aluminum solubility and by cell designs which minimize convective mass transport in the interelectrode region. The latter is driven partly by gas bubbles evolved from the anode but also by electromagnetic forces (Moreau and Evans, 1984). The electromagnetic forces arise from the interaction of the cell current with the magnetic field, of the order of 10^{-2} T, generated within the cell by the cell current and that in nearby bus bars and cells (Lympany and Evans, 1983). In designing modern cells great attention is paid to minimizing these magnetohydrodynamic effects (Potocnik and Evans, 1986). Current efficiencies today are superior to those of, say, the 1950s when they were usually in the range 82%– 86%.

The voltage of a Hall–Héroult cell cannot be reduced below a minimum determined by the enthalpy of the electrochemical reaction(s) and the overpotentials needed to "drive" the reactions at acceptable rates. Thermodynamically, the minimum electrical energy input to the cell is the greater of the enthalpy of reaction and the Gibbs free energy of reaction, multiplied by the production rate. Because the overall reaction entails production of gas (carbon dioxide plus monoxide in small amounts) the entropy of reaction is positive and the enthalpy of reaction therefore exceeds the Gibbs free energy of reaction and is the relevant quantity. To this minimum must be added a quantity of heat necessary to bring the reactants (aluminum oxide and carbon) from the temperature at which they enter the cell to the operating temperature (950–1000 °C), less any of that heat supplied by hot gases leaving the cell. Grjotheim and Welch (1980) report a result of 6.34 kWh kg^{-1} of aluminum product for a cell at 977 °C fed aluminum oxide and carbon at 25 °C with no heat transfer from the gas and a current efficiency of 100%. This minimum is increased somewhat if the current efficiency is less than 100% because the reoxidation of the aluminum proceeds with the production of carbon monoxide, rather than carbon such as would occur by reversal of the overall reaction. Adding overvoltages necessary to drive the anodic and cathodic reactions (particularly the former) at acceptable rates, a minimum cell voltage is reached of approximately 1.7 V, far below the actual cell voltages of 4–4.3 V. The difference represents ohmic potential drops in various parts of the cell, and the energy efficiency of the cell could be improved by their elimination. (In reality their complete elimination may be impossible because they lead to generation of heat that balances the heat losses from the boundaries of the cell. Some heat losses from the sides of the cell are mandated by the fact that materials presently in use for the sidewalls must be protected from exposure to molten electrolyte by the frozen "ledge" of electrolyte appearing in Fig. 5.) Means for reduction of some of these voltage drops are obvious (e.g., the use of bus bars of larger cross section) but one voltage drop is particularly troublesome, namely, the voltage drop through the electrolyte. The electrolyte is a poor conductor (2.4–3.2 Ω^{-1} cm^{-1}) and passage of the high current density through the 4–5-cm gap between anode and cathode results in potential drops of 1.4–1.7 V. Clearly significant energy savings would result from a reduction in the size of this gap. Unfortunately, the reduction of this gap is difficult in an industrial situation. Loss of current efficiency results if the anode is brought too close to the cathode; the loss is either by short-circuiting or by promotion of the mechanism discussed in the previous paragraph. Consumption of carbon from the face of the anodes is uneven and therefore the precise location of the faces is not known to the operators. Furthermore, the densities of the electrolyte (2100 kg m^{-3}) and metal (2300 kg m^{-3}) are close and electromagnetic forces perturb the metal-electrolyte interface, sometimes deforming it by a few centimeters. The normal 4–5-cm separation between the carbon and metal therefore represents a margin of safety, determined by experience, at which cell performance is optimized.

This description of the Hall–Héroult cell has necessarily been brief and there has been no opportunity to describe the research that is being pursued on this topic by aluminum companies and a limited number of universities. Nevertheless, the reader will recognize that the cell is governed by a complex mixture of electrical and magnetic phenomena, electrochemistry, turbulent multiphase flow, heat, and mass transport and is therefore a rich field for applied research.

3.3 Alternative Technologies for Producing Aluminum

Considerable effort is currently being expended on modifying Hall–Héroult cells in ways that depart significantly from present practice. Foremost among these are ones aimed at inert anodes and the use of materials for cell linings that are wetted by aluminum. From the discussion of the last section the cell voltage can be reduced by reducing the anode-cathode distance and this may be achieved if the geometry is better known to the operator (or a control computer). An anode that is not consumed by reaction is one step in this direction and cermet anodes, with a copper alloy metallic phase and a nickel ferrite/nickel oxide nonmetallic phase, have shown reasonable lifetime in laboratory tests

(Weyand *et al.*, 1986). With such anodes the anodic reaction is the generation of oxygen, rather than of carbon dioxide. A second step in the direction of a reduced anode-cathode distance is the use of titanium diboride, and composites of TiB_2, to line the bottom of the cell. These materials are wetted by aluminum (in contrast to the conventional carbon lining) and it becomes possible to operate with the aluminum present only as a film (draining to a sump) that is not perturbed by electromagnetic forces (Dorward and Payne, 1985).

The production of aluminum by electrolysis of the chloride, rather than the oxide, was developed to an industrial scale (Russel, 1981) but it is first necessary to produce the chloride from the oxide and unfavorable economics are suggested by the closure of the plant.

Manufacture of aluminum by chemical, rather than electrochemical, routes has been examined many times in the past. A favorite topic has been the "carbothermic" reduction of alumina, i.e., reduction by carbon at elevated temperatures. The standard free energy of reaction becomes negative at approximately 2100 °C but, at these temperatures, aluminum exerts a significant vapor pressure and a volatile suboxide (Al_2O) is stable. Consequently aluminum vapor and the suboxide migrate out of the reaction zone and back-react as they cool. The problem is aggravated if the necessary heat of reaction is supplied by fuel combustion (as in a blast furnace type of reactor) but is alleviated if the activity of aluminum in the reaction zone is reduced by a suitable alloying element. Aluminum-silicon alloys have been produced in the USSR by carbothermic reduction in electric arc furnaces. Use of tin to lower the aluminum activity, followed by separation of the tin, has recently been examined on a laboratory scale (Frank, Finn, and Elliott, 1989).

4. PROCESSING OF ALUMINUM

Following its production the molten aluminum is alloyed with other elements, cast, and processed (frequently by extrusion or rolling) into the many forms familiar to the end user.

4.1 Alloying

The aluminum from the Hall–Héroult cell typically is better than 99.5% aluminum; much higher purities can be produced by an electrolytic refining step. Such aluminums are soft and find their largest application in the production of aluminum foil where the primary consideration is the simplicity with which thin gauges can be rolled. These applications require aluminum free of nonmetallic inclusions and dissolved gases which lead to imperfections in the foil. The molten aluminum is commonly filtered and gases of low solubility (such as argon) blown through it to remove soluble gases before casting.

The bulk of aluminum is used to produce aluminum alloys. Common alloying elements are silicon, copper, magnesium, zinc, and lithium; their effects on the properties of aluminum have been described in Sec. 1. In most cases alloying consists of merely adding the requisite alloying elements to a furnace containing molten aluminum delivered from the Hall–Héroult cells. Sometimes aluminum is melted in these furnaces (either recycled scrap or new aluminum that has been cast for intermediate storage). The furnaces are usually gas-fired, although electric induction furnaces are used when absorption of impurities from furnace atmospheres is to be minimized. The exception to the above simplicity is alloying with lithium. This element is highly volatile at the temperatures (circa 700 °C) at which molten aluminum is processed and also highly reactive. Technologies for producing aluminum-lithium alloys are proprietary.

4.2 Casting

A distinction must be drawn between the casting of parts (lawn mower housings, cylinder heads, domestic utensils, etc.) and the casting of bulk aluminum for subsequent processing by extrusion, rolling and drawing (to window frames, beverage containers, aluminum foil, etc.). The former is conducted in foundries by die casting and sand mold casting and is largely self-evident, except at the level of detail; it will not be treated further here. The latter is carried out by every aluminum plant because little aluminum is supplied to the customer in liquid form. We may subdivide this latter casting into "continuous"

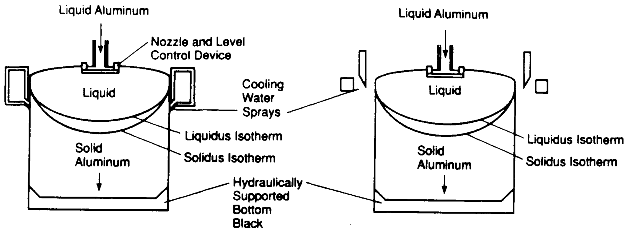
Liquid Aluminum
Nozzle and Level Control Device
Liquid
Cooling Water Sprays
Liquidus Isotherm
Solidus Isotherm
Solid Aluminum
Hydraulically Supported Bottom Black
Liquid Aluminum
Liquid
Solid Aluminum
Liquidus Isotherm
Solidus Isotherm

DC Casting

Electromagnetic Casting

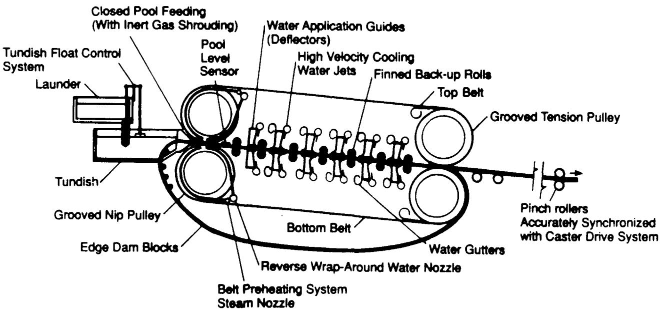
Closed Pool Feeding (With Inert Gas Shrouding)
Pool Level Sensor
Water Application Guides (Deflectors)
High Velocity Cooling Water Jets
Finned Back-up Rolls
Top Belt
Grooved Tension Pulley
Tundish Float Control System
Launder
Tundish
Grooved Nip Pulley
Edge Dam Blocks
Bottom Belt
Water Gutters
Pinch rollers Accurately Synchronized with Caster Drive System
Reverse Wrap-Around Water Nozzle
Belt Preheating System Steam Nozzle

Hazlett Strip Caster

FIG. 7. Three methods for the casting of aluminum into shapes suitable for subsequent rolling or extrusion.

and noncontinuous casting. The latter consists of merely pouring the metal into a mold, usually of cast iron, and allowing solidification to proceed. It is in common use only for producing metal for intermediate storage, i.e., for applications where the metal must be remelted.

The three principal continuous casting technologies for aluminum are "direct chill" (DC) casting, electromagnetic (EM) casting, and roll or belt casting. All three are illustrated in Fig. 7. They are continuous in the sense that they produce solidified metal from a continuous inflow of liquid metal for periods of the order of one to several hours. In DC casting, ingots of sections perhaps 1 m by 2 m and several meters high or slabs of smaller thickness are produced by direct contact between the liquid aluminum and a water-cooled copper mold. The solidified aluminum is supported by a "bottom block" that is moved downwards at a few meters per hour as solidification proceeds. The casting is started with the bottom block placed within the mold. Because solidification relies on contact with the mold, the solidified metal periodically breaks free from the mold surface resulting in surface imperfections and these imperfections must be removed by machining before further processing. It is to avoid this wasteful and expensive "scalping" that electromagnetic casting was devised, originally in the USSR. In EM casting the molten aluminum is confined by electromagnetic forces, rather than an actual mold, and the cooling water is sprayed directly onto the metal surface just below the point of solidification. The forces arise from the currents induced in the metal by current in an inductor surrounding it. The inductor current, of several kiloamperes at several kilohertz, generates a magnetic field and the net effect is a pinch force that supports the molten aluminum against gravity. The electromagnetic forces also cause a stirring of the melt and this is therefore a second example of magnetohydrodynamics in the production/processing of aluminum. Figure 8 provides a comparison between measurements of aluminum velocity made by Vives and Ricou (1985) and those computed by Sakane, Li, and Evans (1988). EM cast aluminum usually need not be scalped and over a million tons of aluminum is cast in this way per year in the West alone. EM cast ingots would be similar in size to DC cast ingots.

Figure 7 shows a Hazlett twin belt caster although related machines operating on similar principles exist. The cross sections cast in such machines (rectangular or near square) would typically be smaller (a few cm) than those of DC and EM casting. The molten

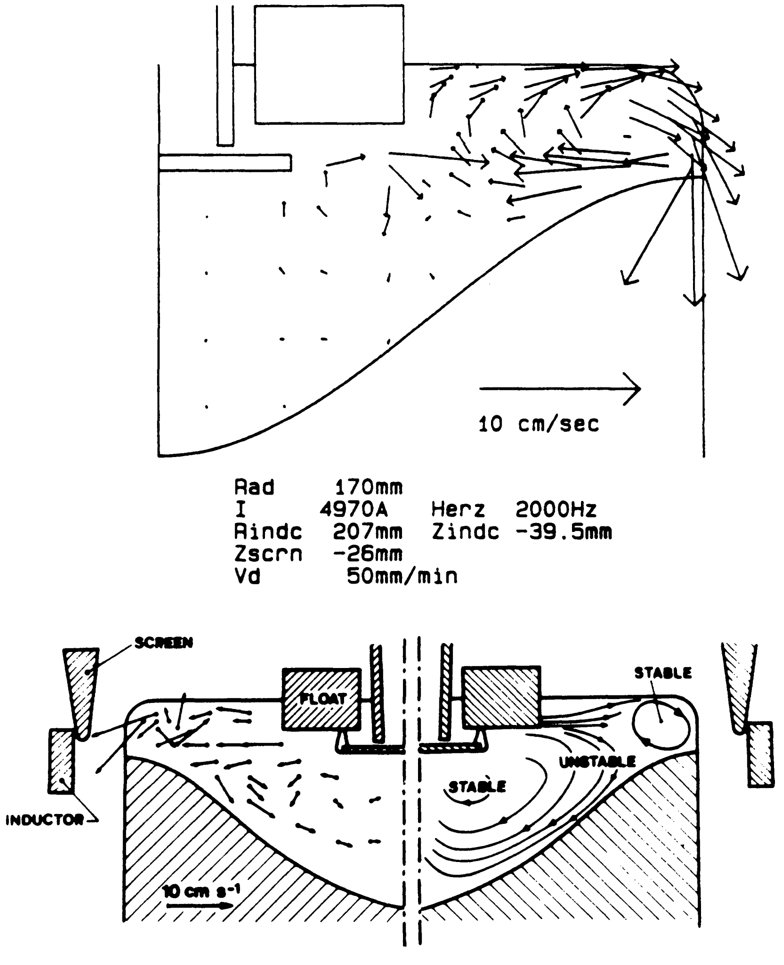

FIG. 8. Electromagnetically driven flows calculated by Sakane, Li, and Evans (1988) and measured by Vives and Ricou (1985) in electromagnetic casting of aluminum.

metal runs between belts driven by wheels at either end of the machine and is confined on the side by vertical dams moving along with the belts. The belts and dams are cooled by water sprays.

4.3 Rolling and Extrusion

The solid aluminum ingots generated by the continuous casting units of the last section are reduced in cross section by rolling or extrusion. The former consists of passing slabs repeatedly between parallel steel cylinders perhaps as much as one meter in diameter ("rolls") that are held at a preset distance apart in a steel frame. Slabs are fed to these machines hot. With each passage the aluminum becomes wider, longer, and thinner. Simple cross sections (rounds, channels, I beams, etc.) can be produced by machines that bend flat sheet subsequent to rolling (e.g., for aluminum guttering). Thin sheet produced in this way is made into beverage containers by "deep drawing" in which a piston-like device forces the sheet into a cylindrical die. Aluminum wire is produced by drawing rounds through dies with cylindrical holes of narrowing diameter.

Extrusion is used to produce cross sections of more complex shape (i.e., finned tube). A length of aluminum rod (heated to a temperature where it has become soft) is forced through a die of required cross section.

5. ENVIRONMENTAL, OCCUPATIONAL AND ECONOMIC ISSUES

5.1 Environmental Issues Related to Aluminum

The poor degradability of the aluminum can is well known; environmentalists and the aluminum industry have promoted the recycling of aluminum cans and other aluminum scrap vigorously. The effort has met with some success; aluminum recycled from users or within aluminum plants represented 25% of total US production in 1988, up from 13% in 1980. The recycling has the additional advantage of reducing energy use since the energy cost of recycled metal is only a fraction of that to produce new metal.

The other major environmental concerns with aluminum are fluoride emissions from Hall–Héroult cells, the red mud from Bayer plants, and the disposal of spent linings from cells. Fluoride emissions can be reduced by the use of "dry scrubbers." These are beds of alumina, en route to the cells, through which the cell exhaust gases are passed before discharge. In this way cell emissions can be reduced to below 0.95 kg per ton of aluminum produced.

Red mud consists of finely divided particles, largely of hydrated iron oxides with some silicon and titanium oxides. The aqueous phase admixed with these particles is highly alkaline and difficult to separate from the particles because of their small size. Consequently, contamination of ground water by the aqueous solution is a concern. Disposal at sea or into rivers has been practiced in the past but environmental concerns now limit such action. Impoundment in dams is now widely employed with abandonment being the eventual fate of the impoundment. In many cases the dams are designed so that, following drainage and evaporation of the liquid, they can be converted to agriculture or recreational use, following coverage by topsoil. Alternatives to disposal (such as use for neutralizing acid soil or manufacturing ceramics and refractories) have been suggested but none appear to have been adopted.

Carbon linings of Hall–Héroult cells have a lifetime of several years but must ultimately be disposed of. Several hundred thousand tons of spent lining are disposed of throughout the world each year. The disposal problem is aggravated by traces of cyanide formed by the reaction of the carbon lining with atmospheric nitrogen.

Table 3. Major uses of aluminum in the U.S. (From Mineral Commodity Summaries, 1987, U.S. Bureau of Mines.)

Packaging	29%
Building	22%
Transportation	21%
Electrical	10%
Consumer durables	9%
Other	11%

5.2 Occupational and Legal Concerns

Apart from exposure to fluorides and the usual hazards of working with hot metals and large pieces of equipment, there is a danger that appears peculiar to the aluminum industry. Contacting molten aluminum with water may lead to more serious accidents than contacting water with other molten metals; it has been suggested (e.g., Epstein and Miller, 1987) that, in addition to explosions due merely to the expansion of the steam, more violent explosions involving chemical reaction of the aluminum with water or steam frequently occur. A number of accidents of this nature take place each year, some involving fatalities. Melting and casting operations where ingots have been stored outside, or scrap is being processed, appear to be particularly susceptible to this hazard.

Allegations that there is a relationship between aluminum and Alzheimer's disease and that aluminum electrical wiring is less safe than copper have been made. Trade organizations such as the Aluminum Association in the U.S. have responded to these allegations.

5.3 Economics of Aluminum

Table 3 summarizes the major uses of aluminum. These have been described in detail in Sec. 1.1. Aluminum use has grown considerably since the start of its wide scale production a century ago, although naturally that growth has suffered occasional reverses. Future growth in worldwide aluminum production appears likely; as indicated by Fig. 9; anticipated future increase in the per capita GNP in developing countries leads to this conclusion. However, a large part of the cost of

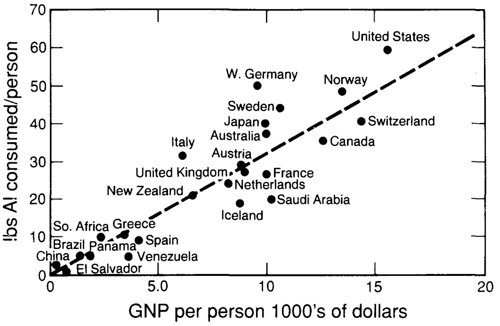

FIG. 9. Aluminum consumption versus gross national product in 1984.

production of aluminum is represented by the electrical energy cost. For example, in the USA this cost fraction is one quarter. Consequently growth in aluminum production is likely to occur in those countries where the cost of generating electricity is low, or the cost to aluminum smelters is limited by political action, self-generation of power, etc. With moderately high power costs the USA, Japan and many European countries do not fall into this category. Recent smelter construction has been in such countries as Venezuela, Brazil, and Australia. Despite this, economic events, particularly the low U.S. dollar, have resulted in U.S. smelters operating at close to capacity. At the end of 1987 aluminum producers in the U.S. were operating at 90% of capacity, compared to 72% at the beginning of that year. In 1988 production reached 96% of capacity.

GLOSSARY

Alumina: Aluminum oxide.

Anode Effect: Sudden increase in voltage across a Hall–Héroult cell brought about by near-zero alumina content in the electrolyte.

Anodes: Positive electrodes (presently of carbon) in Hall–Héroult cells.

Anodizing: In the processing of aluminum alloys, a process in which the alloy is treated in an electrolytic cell to thicken the natural oxide coating.

Bauxite: The ore (impure hydrated aluminum oxide) from which aluminum is produced.

Bayer Process: Process for producing pure aluminum oxide from bauxite.

Carbothermic Reduction: The reduction of aluminum oxide to metal by chemical (rather than electrochemical) reaction.

Collector Bars: Steel bars inset in the carbon lining of Hall–Héroult cells to conduct current out of the cell.

Cryolite: Sodium aluminum fluoride (no longer occurring naturally in significant quantities); the principal ingredient in the molten salt electrolyte of the Hall–Héroult cell.

Crystallographic Texture: Preferential alignment of the crystal axes of the grains of a polycrystalline material.

Current Efficiency: Ratio of aluminum production rate in Hall–Héroult cell to that expected from Faraday's law.

Damage Tolerance: The ability of a material to sustain small flaws, such as small internal cracks, without structural failure.

Deep Drawing: Production of thin-walled aluminum shapes (e.g., beer cans) by forcing sheet aluminum into a cup-shaped die using a piston.

Direct Chill Casting: A common method of continuous casting, whereby the aluminum solidifies as it passes through a copper mold with which it is in contact.

Dry Scrubbers: Devices for trapping fluoride emissions from Hall–Héroult cells.

Ductility: Ability to be deformed; in original usage, ability to be drawn into a thin wire.

Electromagnetic Casting: A common method of continuous casting whereby aluminum is solidified by water jets as it passes through an inductor which serves to support the molten metal.

Electromigration: A coupling between electrical current and chemical diffusion that leads to an enhanced rate of diffusion when a significant current is present.

Extrusion: Production of shapes of uniform cross section (e.g., I beams) by forcing heated solid aluminum through an aperture (die) of the requisite shape.

Fiber Composite: A composite material in which the reinforcing phase has the form of long, thin fibers.

Fiber Laminates: A composite material in which layers of organic fibers or cloth are alternated with layers of metal foil.

Guinier–Preston Zone: Microscopic precipitate within a solid, usually a small cluster of like atoms.

Hall–Héroult Cell: Electrolytic cell used for the reduction of aluminum oxide to metal.

Inert Anodes: Positive electrodes for the Hall–Héroult cell (still under development) that are not consumed by the anodic reaction.

Intermetallic Compound: A material of two or more metals with an ordered crystal structure.

Malleability: Deformability by plastic deformation; in original usage, ability to be hammered or beaten into a thin sheet.

Potlines: Rows of Hall–Héroult cells.

Prebaked Anodes: Anodes produced by baking carbonaceous material in a separate operation from the Hall–Héroult cell.

Precipitate-Free Zone: A layer of material parallel to a grain boundary in a precipitation-hardened material that is free of precipitates due to precipitation on the boundary itself.

Precipitation Hardening: Increasing the strength of a material by precipitating small particles of a second phase in its interior.

Recovery: A process in which dislocations are reconfigured or eliminated by diffusional climb at moderate temperature, softening the material.

Recrystallization: A structural transformation in which crystal defects are largely eliminated from a deformed polygranular material by the nucleation and growth of new grains.

Red Mud: Waste product of the Bayer process.

Roll (or Belt) Casting: A common method of continuous casting, whereby aluminum is solidified as it flows between, and comes in contact with, two moving belts or rolls.

Rolling: Reduction of the thickness of hot aluminum slabs by squeezing them between large-diameter steel rolls.

Soderberg (Self-Baking) Anodes: Anodes produced within the Hall–Héroult cell itself.

Solution Hardening: Increasing the strength of a pure material by adding solute species.

Subgrains: Subvolumes of the crystal grains of a polycrystalline material that are separated from one another by walls of dislocations which form low-angle tilt or twist boundaries.

Superplasticity: Very large plastic elongation prior to failure.

Thermomechanical Treatment: The combination of thermal and mechanical treatments to control the microstructure of a material.

Whisker Composite: Composite material in which the matrix phase is usually a ductile metal and the reinforcing phase is a high-modulus compound in the form of thin, whisker-shaped crystals.

Work Hardening: Increasing the strength of a material by deforming it, which increases the density of dislocations and other defects.

List of Works Cited

Alonso, J. A., March, N. H. (1989), *Electrons in Metals and Alloys*, London: Academic.

Ashcroft, N. W., Mermin, N. O. (1976), *Solid State Physics*, New York: Holt, Rinehart and Winston.

d'Heurle, F. M., Ho, P. S. (1978), in: J. M. Poate, K. N. Tu, and J. W. Mayer (Eds.), *Thin Films–Interdiffusion and Reactions*, New York: J. Wiley and Sons.

Dorward, R. C., Payne, J. R. (1985), *An Evaluation of a Titanium Diboride-Composite Material as Cathodes for Low Energy Alumina Reduction Cells*, report to U.S. Dept. of Energy, DOE/CS/40215-2, Pleasanton, CA: Kaiser Aluminum Corporation.

Epstein, S. G., Miller, R. E. (1987), in: R. D. Zabreznik (Ed.), *Light Metals 1987*, Warrendale, PA: The Metallurgical Society, p. 693.

Frank, R. A., Finn, C. W., Elliott, J. F. (1989), *Metall. Trans.*, **20B**, 161–173.

Fraser, D. B. (1983), in: S. M. Sze (Ed.), *VLSI Technology*, New York: McGraw-Hill.

Grjotheim, K., Welch B. J. (1980) *Aluminum Smelter Technology*, Dusseldorf: Aluminium-Verlag GmbH.

Hafner, J. (1989), "Quantum Theory of Structure: sp-Bonded Systems," in: F. R. de Boer and D. G. Pettifor (Eds.), *The Structures of Binary Compounds*, Amsterdam: Elsevier.

Harrison, W. A. (1966), *Pseudopotentials in the Theory of Metals*, New York: Benjamin.

Harrison, W. A. (1960), Phys. Rev. **118A**, 1190–1208.

Hatch, J. E. (Ed.) (1984), *Aluminum: Properties and Physical Metallurgy*, Metals Park, OH: Am. Soc. Metals.

Heine, V. (1969), "Electronic Structure of Metals," in: J. M. Ziman (Ed.), *The Physics of Metals*, Cambridge: Cambridge University Press.

Jarrett, N. (1983), in: T. J. O'Keefe and J. W. Evans (Eds.), *Electrochemistry Research Needs for Mineral and Primary Materials Processing*, Rolla, MO: University of Missouri.

Lympany, S. D., Evans, J. W. (1983), *Metall. Trans.* **14B**, 63–70.

Moreau, R., Evans, J. W. (1984), *J. Electrochem. Soc.* **131**, 2251–2259.

Potocnik, V., Evans, J. W. (1986), in: E. Ozberk, D. W. Macmillan, and R. I. L. Guthrie (Eds.), *The Production of Liquid Aluminum*, Series 25-7, No. 2, Montreal: The Metallurgical Society of the Canadian Inst. of Mining and Metallurgy.

Russell, A. S. (1981), *Metall. Trans.* **12B**, 203–215.

Sakane, J., Li, B. Q., Evans, J. W. (1988), *Metall. Trans.* **19B**, 397–408.

Shoenberg, D. (1969), "Electronic Structure: The Experimental Results", in: J. M. Ziman (Ed.), *The Physics of Metals*, Cambridge: Cambridge University Press.

U.S. Bureau of Mines (1985), *Mineral Facts and Problems*, Washington, DC: U.S. GPO.

Vives, C., Ricou, R. (1985) *Metall. Trans.* **16B**, 227–235.

Weyand, J. D., DeYoung, D. H., Ray, S. P., Tarcy G. P., Baker, F. W. (1986), *Inert Anodes for Aluminum Smelting*, report to U.S. Dept. of Energy, DOE/CS/40158-20., Alcoa Center, PA: Aluminum Company of America.

Further Reading

Grjotheim, K., Krohn, C., Malinovsky, M., Matiasovsky, K., Thonstad, J. (1977), *Aluminum Electrolysis: The Chemistry of the Hall–Héroult Process*, Dusseldorf: Aluminum-Verlag GmbH.

Grjotheim, K., Welch, B. J. (1980), *Aluminium Smelter Technology: A Pure and Applied Approach*, Dusseldorf: Aluminium-Verlag GmbH.

Hatch, J. E. (Ed.) (1984), *Aluminum: Properties and Physical Metallurgy*, Metals Park, Ohio: Am. Soc. Metals.

Mondolfo, L. F. (1976), *Aluminum Alloys: Structures and Properties*, London: Butterworth.

Vasudevan, A. K., Doherty, R. D. (1989), *Aluminum Alloys–Research and Applications*, Boston: Academic.

AMORPHOUS MATERIALS, PROPERTIES AND APPLICATIONS

See PROPERTIES AND APPLICATIONS OF AMORPHOUS MATERIALS

AMORPHOUS MATERIALS, STRUCTURE OF

S. R. Elliott, *Department of Physical Chemistry, University of Cambridge, UK*

	Introduction	559
1.	**Definitions**	559
1.1	The Amorphous State	559
1.2	Structural Order	560
1.3	Short-Range Order	560
1.4	Medium-Range Order	562
1.4.1	Short-Range MRO	562
1.4.2	Intermediate-Range MRO	563
1.4.3	Long-Range MRO	563
2.	**Experimental Techniques for Structure Determination**	563
2.1	Diffraction	563
2.2	X-Ray Absorption Spectroscopy	567
2.3	Nuclear Resonance Techniques	568
2.3.1	Nuclear Magnetic Resonance	569
3.	**Structural Modeling**	570
3.1	Serial-Addition Models	570
3.2	Monte Carlo Methods	571
3.3	Molecular Dynamics	571
4.	**Examples of the Structure of Amorphous Materials**	571
4.1	Amorphous Hydrogenated Silicon	571
4.2	Oxide Glasses	576
4.3	Chalcogenide Glasses	581
4.4	Amorphous Metals	585
5.	**Future Prospects**	587
5.1	X-Ray Absorption Spectroscopy	587
5.2	Nuclear Magnetic Resonance	587
	Glossary	588
	Works Cited	590
	Further Reading	592

INTRODUCTION

The determination of the atomic structure of amorphous materials is a nontrivial undertaking for a number of reasons. First, the fact that the structure can only be defined essentially in terms of a "unit cell" containing an infinitely large number of atoms means that a statistical description of the structure is unavoidable. Second, the absence of translational periodicity means that many simplifying features associated with the mathematical treatment of, say, diffraction are inappropriate. Finally, resulting from the above considerations, the structure of an amorphous solid can never be determined unambiguously, a situation reinforced by the fact that, in many cases, the structure of a noncrystalline material, at both microscopic and macroscopic levels, depends on details of the method of preparation.

In the light of the above problems, the use of a *single* experimental technique is insufficient to determine the structure of amorphous solids; rather a number of complementary structural probes must be employed in order to obtain as much experimental information as possible, before recourse is made to theoretical structure modeling studies with which the structural features present in the experimental data may be interpreted.

1. DEFINITIONS

1.1 The Amorphous State

We commence with a definition of amorphous materials (Elliott, 1984, 1990): *Amorphous materials are condensed phases which do not possess the long-range translational or orientational order (periodicity) characteristic of a crystal.*

The terms amorphous and noncrystalline are synonymous under this definition, and are used interchangeably in this article. Furthermore, no distinction is made between liquid and solid phases of matter, although here we will consider only the case of *solids* (having viscosities in excess of 10^{13} P at room temperature). Although the problems involved in the determination of the *average* structure are the same in both cases, additional complica-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

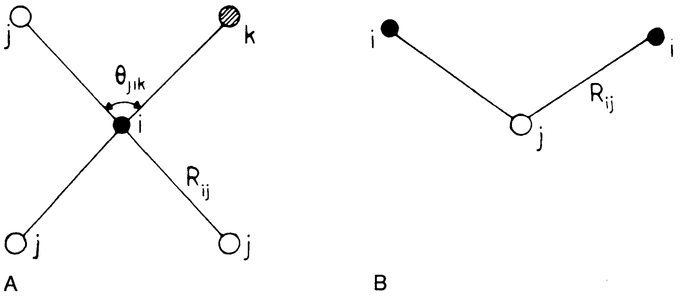

FIG. 1. Schematic illustration of the parameters describing short-range order in covalent amorphous materials.

tions exist in the case of the liquid state because of the dynamical disorder also present.

It should be stressed that here only *topological disorder* is considered, viz., aperiodic disorder in the atomic positions. Other types of disorder, which are associated with underlying *crystalline* lattices, are neglected, such as spin disorder (as in spin "glasses"), substitutional disorder (as in alloys), and either vibrational or orientational disorder, where dynamic vibrational or rotational disorder (which may be frozen in at sufficiently low temperatures in the latter case) may exist with respect to the equilibrium positions of a regular lattice.

Finally, we should emphasize that, with the definition given above, the terms amorphous and noncrystalline are the most general and all-embracing. In particular, we regard the term "glassy" to be a subcategory, with the following definition: *A glass is an amorphous solid which exhibits a glass transition.*

The glass transition is the phenomenon whereby a solid amorphous phase exhibits a more or less abrupt discontinuity in *derivative* thermodynamic properties (e.g., heat capacity or thermal expansivity) from crystal-like to liquidlike values with increasing temperature. Thus, glassy (or equivalently vitreous) materials are always amorphous, but conversely not all amorphous solids are glasses.

1.2 Structural Order

In order to proceed further with a description of the structure of amorphous solids and the techniques employed to determine it, it is convenient to consider the various types of structural order which can exist in such materials at various length scales. Such a categorization is convenient in two regards: the classification is hierarchical so that a particular type of order at one length scale can be dictated by order at a smaller scale (but not necessarily the converse); this division is also pragmatic in that different experimental structure probes are generally sensitive to structural correlations at various length scales.

1.3 Short-Range Order

We start with a discussion of covalent materials since the directed nature of the bonds characteristic of these systems considerably simplifies the picture. In such cases, the short-range order (SRO) is relatively easy to define in terms of well-defined local coordination polyhedra (see Fig. 1). Thus, the parameters which are sufficient to describe topological SRO are the number N_j of nearest neighbors j around an origin atom of type i, the nearest-neighbor bond length R_{ij}, the bond angle subtended at atom i, θ_{jik} (when the type of atom k is different from j), and the corresponding quantities when atom j is regarded as the origin, viz., N_i and θ_{ijk}. Note that in this description of SRO, nothing is said concerning the *connectivity* of the coordination polyhedra (e.g., corner-, edge-, or face-sharing); the form of connection of the polyhedra essentially determines the type and extent of the medium-range order (MRO), as will be seen in the next section.

An additional parameter is required if the degree of *chemical* SRO needs to be described as well, e.g., in the case when *different types* of atoms constitute the coordination polyhedron around a given origin atom. Thus, for nonstoichiometric compositions, for example, excess atoms must be accommodated by the introduction of "wrong" (i.e., homopolar) bonds (in the absence of coordination or valence changes), and the chemical order which might otherwise occur at the stoichiometric composition is thereby broken; the parameter of relevance here would then be the proportion of wrong bonds. Such considerations can also apply in the case of stoichiometric compositions, since chemical ordering may not exist even in that case.

A related type of chemical order is when the different types of atomic species in the coordination shell around a given origin atom are in fact the *same* element but each atom can have a different charge state, bonding connectivity, etc. An example of this is the case of nonbridging oxygens (NBOs) introduced in, say, silicate glasses by the incorporation of alkali modifier atoms, denoted Q_n species, where n (=0–4) is the number of *bridging* oxygens around a central Si atom. In fact, chemical ordering is often strong in such systems, and at the stoichiometric compositions only a *single* type of Q_n unit exists to a first approximation, whereas for compositions in between two stoichiometric compositions, the two species Q_{n+1}, Q_n, characteristic of the end-member compositions, coexist.

The question of chemical ordering in covalent systems is most simply discussed for the case of binary alloy systems, e.g., A_xB_{1-x}, where, if the elements A and B are in columns a and b of the Periodic Table, coordinations of $8-a$ and $8-b$, respectively, will occur, thereby satisfying their normal valence and the so-called "$8-N$" rule (strictly valid only for groups IV–VII). We neglect for the time being the effect of any coordination defects, such as dangling bonds. In general, A–A, A–B, and B–B bonds can coexist in an alloy of arbitrary composition, and two models can describe the distribution of such bond types (Lucovsky *et al.*, 1977).

The first is the *random covalent network* (RCN) model, in which the distribution of bond types is treated as being purely statistical, determined only by the local coordinations $N_a=8-a$ and $N_b=8-b$ and the concentration variable x. Thus, any effects leading to preferential ordering (e.g., arising from differences in bond energies) are neglected. The RCN model, therefore, admits A–A, A–B, and B–B bonds at *all* compositions except at $x=0$ and 1. The other model is the *chemically ordered network* (CON) model which assumes that heteropolar A–B bonds are favored; a completely chemically ordered phase thus occurs at the stoichiometric composition $x_c = N_a/(N_a + N_b)$ (e.g., As_2Se_3, $GeSe_2$). In contrast to the RCN model, only A–A and A–B bonds are allowed for A-rich compositions ($1 > x > x_c$), and conversely only B–B and A–B bonds for $1 < x < x_c$. Obviously the RCN and CON approaches are valid only for the case of covalent systems with well-defined directed bonds, i.e., not for metallic or ionic systems. Furthermore, they are appropriate only for glass-forming systems, i.e., excluding "modified" glasses in which network modifiers have either depolymerized the network, by introducing nonbridging anionic sites, or changed the coordination of the network-forming cation.

The above description of SRO in terms of rather well-defined coordination polyhedra is obviously inappropriate for those systems in which the bonding is nondirectional, e.g., in the case of metals or ionic systems in which the structure can be regarded simply in terms of a random packing of spheres. In such cases, a description of the local order is more difficult.

One particularly prevalent type of coordination polyhedron which can occur in randomly close-packed arrangements of spheres is the *icosahedron*. Thus, small neutral inert-gas clusters appear to exist comprised of "magic numbers" of atoms (13, 19, 23, 26, 29, etc.); all of these can be obtained by the serial addition of atoms to a starting 13-atom icosahedral cluster seed, giving so-called Mackay icosahedra. Moreover, it has long been asserted that supercooled liquid metals should contain icosahedral clusters (which cannot therefore subsequently act as nucleation centers for crystallization). It is consequently natural to consider the extent of local icosahedral order in solid materials rapidly quenched from the melt, viz., metallic glasses, and to do this some kind of order parameter is required.

One such concerns the so-called "Voronoi polyhedra" (or equivalently Wigner–Seitz cells) associated with such clusters. In three dimensions, the polyhedron is defined as the smallest convex polyhedron formed by the planes which bisect perpendicularly the "bonds" from a given origin atom to its nearest neighbors. The number of edges of a particular face of a Voronoi polyhedron is equal to the number of tetrahedra packed around the "bond" bisected by that face, i.e., pentagonal faces are a signature of icosahedral order.

Another monitor of local icosahedral order is associated with its *orientational* properties. Thus, Steinhardt *et al.* (1983) have proposed the following order parameter:

$$Q_{lm}(r)=Y_{lm}(\theta(r),\phi(r)), \tag{1}$$

where θ and ϕ are the polar angles of a "bond" between two nearest-neighbor atoms whose center is at r and $Y_{lm}(\theta,\phi)$ are the usual spherical harmonic functions. The particular direction of a bond is unimportant if only even-l spherical harmonics are considered, since these are inversion-invariant. Furthermore, rotationally invariant combinations are more discriminatory, such as

$$Q_l=\left[\frac{4\pi}{2l+1}\sum_{m=-l}^{l}|\overline{Q}_{lm}|^2\right]^{1/2}, \tag{2}$$

where $\overline{Q}_{lm}$ represents an average over a number of bonds in the structure. Clusters with icosahedral symmetry have nonzero spherical harmonics only for $l=6,10,12,\ldots$, etc., whereas various forms of cubic close packing, for example, have nonzero values of Q_l for $l=4$ as well. A dominant contribution to Q_l for $l=6$ can therefore signify local icosahedral order.

In the absence of reasonably well-defined coordination polyhedra, as discussed above, SRO can only be discussed in terms of averaged quantities; this is the case, for example, in glassy metal-metal alloys. An order parameter can be defined in terms of a function involving only the coordination numbers of the first shell (Sadoc and Wagner, 1983), viz.,

$$\alpha_1=1-\frac{N_{12}}{x_2(x_2N_1+x_1N_2)}, \tag{3}$$

where N_{ij} is the coordination number of atoms of type j surrounding atom i, and $N_i = \sum_{j=1}^{2} N_{ij}$. Negative values of α_1 indicate chemical ordering, i.e., the occurrence of unlike nearest neighbors. While Eq. (3) is strictly only valid for equally sized atoms, Cargill and Spaepen (1981) have also proposed an expression describing the extent of SRO in terms of nearest-neighbor coordination numbers, viz.,

$$\eta_{ij}=\frac{N_{ij}\overline{N}c}{x_jN_iN_j}-1, \tag{4}$$

where $\overline{N}$ is the average coordination number given by $\overline{N} = x_iN_i + x_jN_j$. For complete chemical disorder, $\eta_{ij} = 0$; otherwise, if chemical ordering exists (AB nearest neighbors favored), $\eta_{ij} > 0$, and in the other limit of chemical clustering (i.e., AA, BB pairs favored), $\eta_{ij} < 0$.

1.4 Medium-Range Order

Medium-range order (MRO) is less easy to define than SRO, and consequently several definitions exist in the literature. In this section we will adopt the following definition of MRO (Elliott, 1987): MRO can simply be regarded as constituting the next higher level of structural organization beyond that of SRO, existing on a length scale of, say, 5–50 Å. It is convenient in practice, particularly for covalent materials where MRO is more prevalent, to divide the types of MRO into three categories, corresponding to progressively increasing length scales.

1.4.1 Short-Range MRO. If SRO is defined in terms of well-defined coordination polyhedra, as above, short-range MRO (SRMRO) is therefore concerned with the type of connection of such polyhedra as well as their relative orientation. Thus, corner-, edge-, and face-sharing of polyhedra lead to very different ordering schemes at a local scale, as well as to pronounced differences at larger distances in terms of the network dimensionality (see Sec. 1.4.3).

MRO is most pronounced when the degrees of freedom associated with the relative orientation of neighboring polyhedra are restricted in some way, e.g., by the occurrence of edge-sharing connections. However, even in the case of corner-sharing units, where free rotation about the common bonds could occur in principle, certain orientations may be more probable than others. A structural parameter which is a measure of such orientational correlations for covalent systems (with well-defined bonds) is the "dihedral" (or torsion) angle, ϕ, characteristic of a four-body correlation function; it is the angle of rotation about a common bond required to bring into coincidence the projections, onto the plane perpendicular to this bond, of the other bonds in the connected units (Fig. 5). Departures from a uniform (i.e., random) distribution $P(\phi)$ are correspondingly a hallmark of MRO. Favored values of ϕ consequently are associated with features of MRO such as (small) rings. However, this need not necessarily always be the case: for example, it appears from structural modeling studies and considerations of features in the electronic density of states that the structure of *a*-Se is not representative of a freely rotating chain [with constant $P(\phi)$], but instead ϕ takes on more or less fixed values ($\sim 102°$, as in the crystal), with the sign of ϕ varying either alternately or randomly down a chain.

1.4.2 Intermediate-Range MRO. MRO at the next larger length scale, say $\geqslant 5$ Å, is obviously determined by the type and extent of SRMRO present. However, we make the distinction that intermediate-range MRO (IRMRO) is characterized by the presence of "superstructural" units (comprising several basic SRO units), e.g., rings or clusters of atoms of particular shapes or sizes, which occur in a considerably higher proportion than would be expected on a purely statistical (random) basis. In certain cases, such superstructural units can be regarded as forming the essential building blocks of the amorphous structure. Thus, IRMRO is characterized by correlation functions involving between, say, 6 and 12 atoms, and occurs if successive dihedral angles are themselves correlated.

1.4.3 Long-Range MRO. On a yet larger length scale, say greater than 10 Å, long-range MRO (LRMRO) can be regarded as being associated with the local dimensionality of a covalently bonded amorphous network, or the existence of domains in metallic systems.

A local dimensionality different from 3D can arise in a covalent network from two causes: either the type of connection between coordination polyhedra may impose such a reduced dimensionality, or it may result from network depolymerization resulting from the introduction of network modifiers.

It is often assumed that structural ordering at intermediate length scales is a *consequence* of a well-defined SRO. This picture is probably valid for the case of covalently bonded systems where the SRO is often extremely well-defined since it is determined by chemical bonding constraints. In the case of amorphous materials in which nondirectional bonding is predominant, e.g., metals, such constraints no longer apply, yet in some cases, e.g., in transition-metal metalloid (T-M) glasses, a well-defined topological and chemical SRO nevertheless still appears to be present. In such cases, it may be that the SRO itself is a *consequence* of the existence of a well-defined MRO, i.e., the converse of the case for covalent systems (Gaskell, 1987). An example of this in the crystalline state is the occurrence of tetrahedra and octahedra arising from ordered close-packed arrays of atoms. In a similar manner, the trigonal prismatic coordination of metalloid atoms in crystalline T-M materials can be regarded as arising from the introduction of "chemical twinning" planes containing the M atoms into close-packed crystalline arrangements of the metal atoms. Dubois *et al.* (1985) have proposed that similar considerations apply equally to the structure of *glassy* T-M alloys, in which the LRMRO is associated with domains of size 10–20 Å, within each of which more or less ordered arrays of chemical twinning planes exist, and along which are therefore located the trigonal prismatic units.

2. EXPERIMENTAL TECHNIQUES FOR STRUCTURE DETERMINATION

A variety of techniques can be used to investigate the structure of amorphous solids. Indeed, because of the limit on the amount of structural information obtainable from any one method, due to the aperiodicity of the structure, it is often necessary to employ a variety of complementary techniques in order to obtain the maximum amount of structural information. In this regard, the most useful structural techniques are those which are "atom-specific," viz., which probe the local structural environment around a given type of atom.

2.1 Diffraction

Diffraction, using x rays, neutrons, or electrons, is the technique most widely used to determine the structure of crystalline solids. However, for the case of amorphous materials, the absence of translational symmetry means that the diffraction theory must be developed from first principles without any of the simplifying features which can be employed in the case of crystalline materials. For the case of *monatomic* amorphous materials, diffraction provides the maximum amount of available structural information, whereas for the case of multicomponent systems the information is much less easily interpretable.

The scattering intensity from a sample is measured as a function of scattering angle 2θ, or equivalently the scattering vector (or momentum transfer)

$$k=\frac{4\pi}{\lambda}\sin\theta, \tag{5}$$

where λ is the wavelength of the radiation used. (k is conventionally used in x-ray applications and Q for neutron scattering.) For monatomic amorphous materials, the scattering intensity $I(k)$ is related to the structure factor $S(k)$ via

$$S(k)=I(k)/Nf^2, \tag{6}$$

where N is the number of scattering atoms and $f(k)$ is the k-dependent x-ray scattering factor (or equivalently b is the k-independent neutron scattering length). Thus, $S(k)$ is *independent* of the nature of the scattering radiation and is determined only by the atomic structure of the amorphous material. A reduced scattering factor can also be defined in terms of these two quantities, viz.,

$$F(k)=k[S(k)-1]. \tag{7}$$

$F(k)$ oscillates about an average value of zero, whereas $S(k)$ oscillates about unity and $I(k)$ about Nf^2.

It can be shown from first principles (see, e.g., Elliott, 1984, 1990) that, for isotropic amorphous materials, $F(k)$ is related to the real-space correlation function $G(r)$ (the reduced radial distribution function), via the expression

$$F(k)=\int_0^\infty G(r)\sin kr\,dr. \tag{8}$$

Correspondingly, there is an inverse relation through the Fourier transform:

$$G(r)=\frac{2}{\pi}\int_0^\infty F(k)\sin kr\,dk. \tag{9}$$

In practice, the scattering intensity cannot be measured over an infinite range of k values as required by Eq. (9); the maximum value is $k_{\max}=4\pi/\lambda$ from Eq. (5).

This cutoff in measured values of k leads to termination errors in the Fourier-transformed real-space correlation function: spurious small oscillations occur on either side of prominent peaks in the correlation function, and the resolution (i.e., width of peaks) is worsened.

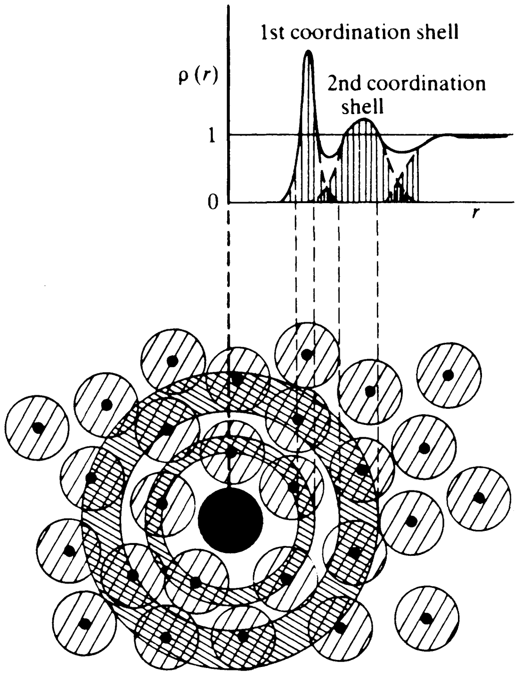

FIG. 2. Schematic illustration of the structural origin of features in the density function $\rho(r)$ for an amorphous solid.

The reduced radial distribution function (RDF) is related to the density correlation function $\rho(r)$ and the RDF itself $J(r)$ via the relations

$$\begin{aligned}G(r)&=4\pi r[\rho(r)-\rho_0]\\&=J(r)/r-4\pi r\rho_0,\end{aligned} \tag{10}$$

where ρ_0 is the average (macroscopic) atomic density. The RDF, defined as

$$J(r)=4\pi r^2\rho(r), \tag{11}$$

is the probability per unit length of finding an atom at a distance between r and $r+dr$ from a given origin atom. The structural origin of features in $\rho(r)$ [or $J(r)$] is illustrated in Fig. 2: it can be seen that the RDF is a one-dimensional representation of an (assumed isotropic) three-dimensional structure. Figure 3 shows various experimental correlation functions, in both real and reciprocal space,

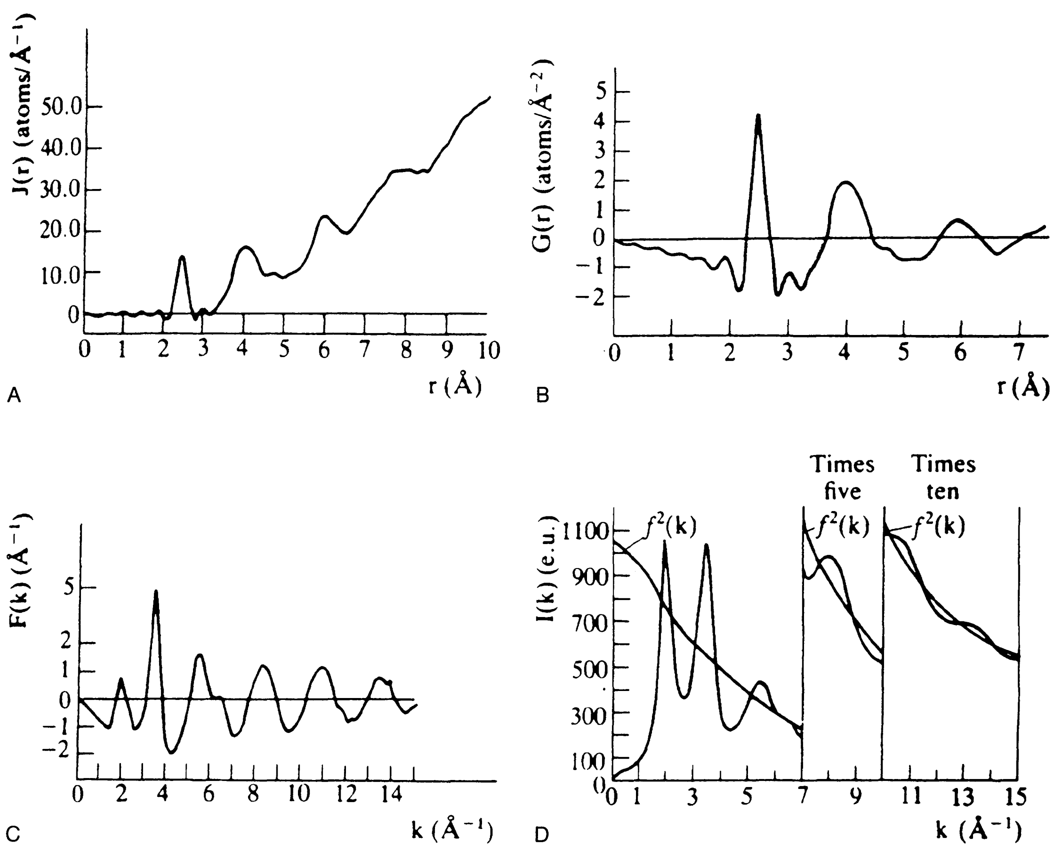

FIG 3. Structural correlation functions in real and reciprocal space obtained by x-ray diffraction for sputtered *a*-Ge films (Temkin *et al.*, 1973): (a) RDF; (b) reduced RDF; (c) reduced intensity; (d) scattering intensity.

for the case of the monatomic system *a*-Ge (Temkin *et al.*, 1973).

The more common situation involving multicomponent amorphous systems is considerably more complicated as far as diffraction is concerned. For a material containing n elements, a total of $n(n+1)/2$ distinct pair correlation functions are needed to describe the structure fully. The density function can now be defined, in terms of an origin atom i and surrounding atoms j, as

$$\rho(r)=\sum_{ij} W_{ij}\,\rho_{ij}(r)/x_j, \tag{12}$$

where the weighting factors (k-dependent for the case of x rays) are given by

$$W_{ij}=x_i\,x_j\,f_i\,f_j/\langle f\rangle^2, \tag{13}$$

and where $\langle f\rangle=\sum_{i=1}^{n} x_i\,f_i$. The reduced scattering intensity now becomes

$$F(k)=k\left[\frac{I/N-\langle f^2\rangle}{\langle f\rangle^2}\right]$$

$$=\int_0^\infty 4\pi r[\rho(r)-\rho_0]\sin kr\,dr \tag{14}$$

(where $\langle f^2\rangle=\sum_{i=1}^{n} x_i\,f_i^2$), which is superficially similar to Eqs. (8) and (10) for the case of monatomic systems.

The above treatment is developed in terms of atom-atom pair correlation functions (the so-called Faber–Ziman formulation). For the case of *binary* alloys, the expression for the scattering intensity can be reformulated in terms of number-number, number-concentration, and concentration-concentration correlation functions, $S_{NN}(k)$, $S_{NC}(k)$, and $S_{CC}(k)$, respectively (Bhatia and Thornton, 1970), where

$$S_{NN}(k)=x_1^2I_{11}(k)+x_2^2I_{22}(k)+2x_1x_2I_{12}(k), \tag{15a}$$

$$S_{\mathrm{NC}}(k)=x_1x_2\{[x_1I_{11}(k)+x_2I_{12}(k)] \\ -[x_1I_{21}(k)+x_2I_{22}(k)]\}, \qquad (15b)$$

$$S_{\mathrm{CC}}(k)=x_1x_2\{1+x_1x_2[I_{11}(k)+I_{22}(k) \\ -2I_{12}(k)]\}, \qquad (15c)$$

and where

$$I_{ij}(k)-1=\int_0^\infty 4\pi r^2\{[\rho_{ij}(r)/x_j]-\rho_0\} \\ \times\frac{\sin kr}{kr}\,dr. \qquad (16)$$

The overall scattering intensity in the Bhatia–Thornton formalism can then be written as

$$I/N=\langle f\rangle^2S_{\mathrm{NN}}(k)+2\langle f\rangle\Delta fS_{\mathrm{NC}}(k) \\ +\frac{[\langle f^2\rangle-\langle f\rangle^2]}{x_1x_2}\,S_{\mathrm{CC}}(k), \qquad (17)$$

where $\Delta f = f_1(k) - f_2(k)$. $S_{\mathrm{NN}}(k)$ represents the *topological* SRO, and it oscillates about unity and tends to that value as $k\rightarrow\infty$. $S_{\mathrm{CC}}(k)$ describes the *chemical* SRO, and for a random distribution of the two types of atoms in a binary alloy $S_{\mathrm{CC}}(k) = x_1x_2$; otherwise, it oscillates about this value and tends to it as $k\rightarrow\infty$. *Positive* values of $S_{\mathrm{CC}}(0) - x_1x_2$ indicate a preference for *like* nearest neighbors (or clustering), whereas *negative* values indicate a preference for unlike neighbors (or chemical SRO). Finally, $S_{\mathrm{NC}}(k)$ is zero for random mixtures; otherwise it oscillates about this value and vanishes as $k\rightarrow\infty$.

For the case of multicomponent amorphous systems, the maximum amount of structural information, viz., the partial pair correlation functions, is obtained, in principle, if as many scattering experiments are performed as there are distinct partial functions in each of which the weighting factor W_{ij} [Eq. (13)] is somehow varied. Since a change in composition (x_i or x_j) usually means a concomitant change in structure, this means that it is the scattering factors which must be varied, and at least five different methods for achieving this can be identified for the simplest case of binary systems where there are *three* distinct partial pair correlation functions.

The first is simply to use three different types of radiation, viz., neutrons, x rays, and electrons, although this method is often impractical in view of the disparate quantities of material required by each technique. The second method is that of isotopic substitution in neutron scattering, wherein the isotopic components of the elements involved are varied while keeping the overall composition unchanged (Neilson and Enderby, 1983). The largest changes in scattering factor arise when a particular element has a *negative* neutron scattering length, as is the case for the elements H, Li, Ti, and Ni, for which two special cases are possible: in the case of "null elements" (e.g., where $\bar{b}_1 = 0$), $I_{22}(k)$ or $G_{22}(r)$ is obtained directly from a neutron diffraction experiment; in the case of "zero alloys," i.e., where $\langle b\rangle = x_1b_1 + x_2b_2 = 0$, $S_{\mathrm{CC}}(k)$ is obtained directly, and if $b_1 = b_2$ (i.e., $\Delta f=0$), $S_{\mathrm{NN}}(k)$ is obtained directly [cf. Eq. (17)]. A third possibility of changing neutron scattering lengths is for the case of magnetic amorphous metals, where the scattering of *polarized* neutrons is different depending on whether the neutron spin polarization is parallel or antiparallel to the magnetization direction of the amorphous metal (Sadoc and Dixmier, 1976). A fourth technique, applicable to x-ray scattering, is isomorphous substitution, wherein one of the elements in a binary (usually metallic) alloy is partially replaced by a third element having the *same* size and chemical characteristics as that being replaced (so that the structure is unperturbed), but because of the difference in atomic number, the x-ray scattering factors can vary significantly. Finally, the method of anomalous scattering can be used to vary the scattering factors systematically in certain cases. The technique utilizes the fact that near an absorption edge for the scattering radiation involved, the scattering factor is modified by the dispersion correction which is complex, viz.,

$$f(k,\omega)=f^0(k)+f'(k,\omega)+if''(k,\omega). \qquad (18)$$

The first term in Eq. (18) is the normal scattering factor, which for the case of x rays is simply the Fourier transform of electronic charge density of the atom. The imaginary component of the dispersion correction is proportional to the absorption coefficient $\alpha(\omega)$, viz., $f''(\omega) \propto \omega\alpha(\omega)$, and f' and f'' are related via Kramers–Kronig transformations. Although the method is more generally useful for the case of x-ray scattering, where every element has a series of x-ray absorption edges at different energies and these are accessible using a broadband tunable x-ray synchroton radiation source, it can also be employed for the case of neutron scattering in the rare cases where a thermal neutron undergoes capture by a nucleus (e.g., in ^{149}Sm).

In conclusion, diffraction techniques provide, through the RDF, the maximum amount of structural information for amorphous materials for the case of monatomic systems, and a similar situation prevails in the case of multicomponent systems if the partial correlation functions can be extracted. Structural information can be interpreted most easily from the RDF, particularly for the first two peaks which give SRO information. Higher-lying peaks are increasingly difficult to interpret unambiguously in the absence of structural modeling studies (see Sec. 3).

2.2 X-Ray Absorption Spectroscopy

The first of the atom-specific structural probes, complementary to diffraction, to be discussed is x-ray absorption spectroscopy (XAS). In this, the x-ray absorption coefficient is measured as a function of x-ray photon energy in the vicinity of an absorption edge of one of the elements in a multicomponent amorphous material, and three regions in the resulting spectrum can be distinguished, consisting of, respectively, pre-edge structure, just beyond the absorption edge, "x-ray absorption near-edge structure" (XANES) or equivalently "near-edge x-ray absorption fine structure" (NEXAFS), and at higher photon energies "extended x-ray absorption fine structure" (EXAFS). Figure 4(a) shows experimental x-ray absorption spectra exhibiting features in all three regions of energy.

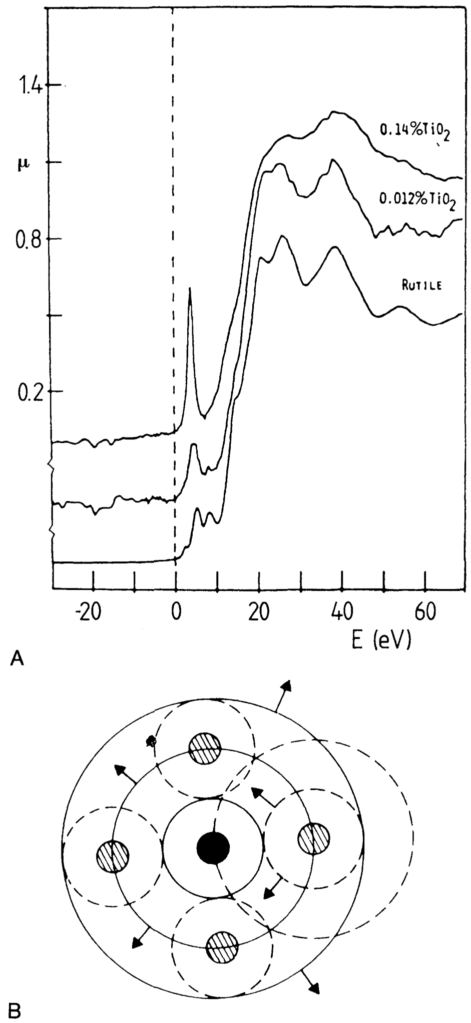

FIG. 4. (a) X-ray absorption spectra for Ti *K* edge in two TiO$_2$-containing silicate glasses, together with that for crystalline rutile for comparison (Greegor *et al.*, 1983). (b) Schematic illustration of the process leading to EXAFS. An atom (solid circle) absorbing an x-ray photon creates an outgoing photoelectron wave which backscatters off neighboring atoms (hatched circles). Interference between outgoing and backscattered electron waves results in the fine-structure oscillations in the x-ray absorption coefficient.

The physical origin of the postedge fine structure is illustrated in Fig. 4(b). The photoelectron created upon the absorption of an x-ray photon by a particular atom propagates outwards and is backscattered from the surrounding atoms; interference between outgoing and backscattered electron waves causes a modulation of the x-ray absorption coefficient with increasing photon energy and hence increasing photoelectron kinetic energy. The periodicity of the EXAFS oscillations is therefore clearly determined by the interatomic spacing. The origin of the XANES oscillations (up to ~50 eV from the edge) differs somewhat from that for EXAFS; multiple-scattering events characterize XANES, whereas only single-scattering processes are involved in EXAFS (in this regard EXAFS and diffraction are similar, yielding only information concerning *pair-correlation* functions. The origin of the pre-edge structure is rather different: core-level electrons are photoexcited into low-lying bound states of the atom (e.g., *d* levels in transition metals). XAS is clearly an atom-specific structural probe since each element has x-ray absorption edges at different energies.

XAS measurements may be performed in a number of ways. The x-ray absorption coefficient $\mu(E)$ may be measured directly by monitoring the transmission of the x-ray beam through the sample. Alternatively, $\mu(E)$ may be measured *indirectly* by monitoring the recombination processes of the photocreated core hole, either radiatively (via x-ray fluorescence) for heavy atoms, or nonradiatively (via Auger electron emission) for lighter atoms. The XANES or EXAFS amplitude is obtained from the measured x-ray absorption coefficient by subtracting the background absorption μ_0 and normalizing to it, viz.,

$$\chi(E)=\frac{\mu(E)-\mu_0(E)}{\mu_0(E)}. \tag{19}$$

The multiple scattering inherent in XANES precludes a simple theoretical treatment and an analytic expression for the XANES amplitude (Durham *et al.*, 1981). On the other hand, the single-scattering processes characteristic of EXAFS, together with the fact that the plane-wave approximation can be invoked for photon energies $\geqslant 100$ eV from the edge, means that an analytic expression for the EXAFS amplitude $\chi(k)$ can be obtained (Sayers *et al.*, 1971), namely,

$$\begin{aligned}\chi(k)=\sum_j &\frac{|f_j(\pi)|}{k}\exp(-2R_j/\lambda)\\ &\times\exp(-2\sigma_j^2 k^2)\\ &\times\sin[2kR_j+2\delta(k)+\eta(k)],\end{aligned} \tag{20}$$

where the photoelectron wave vector k is defined as

$$k=\frac{2\pi}{\lambda}=[2m_e(E-E_0)/\hbar^2]^{1/2}, \tag{21}$$

and where m_e is the electron mass and E_0 the photoelectron threshold (absorption edge). [Note that k for EXAFS is *half* that for diffraction; cf. Eq. (5).] In Eq. (20), the summation is taken over all shells of atoms j surrounding the absorbing atom, with N_j being the coordination number, R_j the interatomic distance, and σ_j its rms fluctuation characterizing each shell. The EXAFS amplitude is also determined by the backscattering amplitude $f_j(\pi)$ of atoms j, and is damped by two terms, one a Debye–Waller factor involving σ_j, and the other a term accounting for inelastic losses of the photoelectron where λ is the electronic mean free path. Finally, the EXAFS oscillations arise from the sine term in Eq. (20) reflecting the interference processes, the argument of which depends on the phase shift associated with the passage of the photoelectron to the backscattering atom and back, as well as the phase shifts suffered by the photoelectron traversing the emitter potential and scattering off the backscatterer potential.

The advantages of XAS as a structural tool for amorphous solids lie in its atom specificity, the sensitivity to dilute components (if fluorescence detection is used), the accuracy with which bond lengths can be measured, the sensitivity to symmetrical environments (particularly in the case of pre-edge structure), the ability to investigate the surface structure by means of glancing-angle techniques, and the possibility to undertake time-resolved studies by utilizing the pulsed nature of synchrotron radiation and using energy-dispersed detection. There are considerable disadvantages as well, however. Foremost amongst these is the fact that EXAFS is only sensitive to the very local structural order in an amorphous solid, principally the first coordination shell; higher-lying shells in noncrystalline materials have increasingly large amounts of static disorder and correspondingly their contribution to the EXAFS amplitude becomes increasingly damped [cf. Eq. (20)] and confined to values of k for which the single-scattering EXAFS formalism is invalid. Furthermore, the uncertainties in the value of coordination numbers resulting from fitting EXAFS data are rather large.

2.3 Nuclear Resonance Techniques

Two nuclear resonance techniques can be used to provide structural information on amorphous solids—these are nuclear magnetic resonance and Mössbauer spectroscopies, which by their very nature are atom specific. However, they both suffer the disadvantage that not all nuclei are amenable to investigation in this way. Moreover, since NMR is far more widely used and is applicable to a wider range of nuclei than is the case for Mössbauer spectroscopy, only NMR will be discussed further.

2.3.1 Nuclear Magnetic Resonance. NMR is the resonant excitation by rf photons of the magnetic-field-split Zeeman levels of the nucleus (see MAGNETIC RESONANCE, NUCLEAR). Early work used cw spectrometers, in which the rf frequency was fixed and the magnetic field was varied to establish the resonance condition (see, e.g., Bray, 1978). The field has been rejuvenated following the introduction of modern Fourier-transform NMR spectrometers, in which the magnetic field is kept fixed, and the frequency effectively swept by the use of pulses of rf power containing many Fourier components. This approach has the advantage that the signal-to-noise ratio can be greatly improved by the repeated application of pulses and the use of signal averaging; moreover, particular pulse sequences can be used to eliminate the effect of certain interactions involving the nuclear spins which serve to broaden the resonance lines.

Further significant progress in the use of NMR to study solids, particularly amorphous materials, has come through the use of the "magic-angle spinning" (MAS) technique combined with Fourier-transform spectrometers (Andrew, 1981). In nonmetallic systems, it is the chemical-shift interaction, in which the external magnetic field is partially shielded from the nuclear spins by the electronic charge distribution in the atoms, that mainly carries the structural information. However, other interactions involving the nuclear spins, e.g., dipolar and quadrupolar, can broaden the resonance lines of solids so much that their position and shape are no longer discernible; the same effect can also arise from the chemical-shift interaction itself in the case when the resonant nucleus is at a very nonsymmetrical structural site. All these broadening effects can be reduced to a greater or lesser extent by the use of the MAS technique, in which the sample is rotated at a high frequency (usually 1–10 kHz), comparable to the strength of the interaction which it is desired to remove, at an angle θ to the external magnetic field. In this case, it is found that the Hamiltonian for all interactions (but only to first order for the quadrupolar interaction) is multiplied by the factor (Andrew, 1981):

$$\beta=(3\cos^2\theta-1). \tag{22}$$

Thus, spinning the sample at the "magic angle" $\theta=54°44'$ ($=\cos^{-1}1/\surd 3$) means that the factor β vanishes, with the result that the line-broadening interactions are removed, and a previously broad line collapses to a single narrow line situated at the isotropic value of the chemical shift tensor, viz., $\sigma_i=\frac{1}{3}\operatorname{Tr}(\sigma)$, exactly as in liquid-state NMR where the improved resolution arises from the isotropic thermal rotational motion of the molecules in the liquid.

NMR is a powerful atom-specific structural probe, being particularly sensitive to the symmetry of sites and the coordination thereof, and often also sensitive to dilute components. However, it does have a number of disadvantages, principal among which is the fact that relatively few nuclei are NMR-active, and of these the natural abundance or sensitivity may be very low. Furthermore, *direct* structural information is mostly not obtainable from NMR spectra, although coordination-number information can sometimes be deduced by a "fingerprinting" process, namely, by comparison of the unknown (amorphous) NMR spectrum with that for a (crystalline) standard whose structure is known. In some cases, direct structural information can be obtained, e.g., in the case of glassy silicates. The ^{29}Si MASNMR spectra of various crystalline polymorphs of SiO_2 exhibit a systematic variation of chemical shift with oxygen bond angle θ, obeying the empirical relation (Smith and Blackwell, 1983)

$$\sigma=a+b\sec\theta. \tag{23}$$

Assuming that the same relation holds for the *continuous* distribution of oxygen bond angles in vitreous SiO_2 resulting from the static disorder which is present, Eq. (23) can be used to convert the observed ^{29}Si NMR line shape into a bond-angle distribution (Dupree and Pettifor, 1984; Gladden *et al.*, 1986).

An additional limitation to the applicability of the MASNMR technique to (amorphous) solids is in the case of *quadrupolar* nuclei (those for which the nuclear spin is $\geqslant 1$), where magic-angle spinning only removes the first-order broadening effects of the interaction; if the nuclear quadrupolar moment is very large, little resolution enhancement of lines is achievable using MAS, and the existence of a multiplicity of structural sites with different chemical shifts can therefore be obscured. In this regard, the very recent advance of double sample rotation (Samoson *et al.*, 1988) holds out considerable promise. In this, the sample is rotated about *two* axes simultaneously, one being at the magic angle, and in this way second-order quadrupolar broadening effects can be removed as well. Although significant increases in the resolution of NMR spectra of quadrupolar nuclei in crystalline materials have been achieved (Chmelka *et al.*, 1989), as yet the technique has not been applied to the study of amorphous materials.

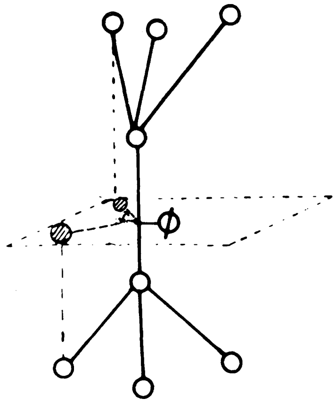

FIG. 5. Definition of dihedral (or torsion) angle ϕ for a covalently bonded material.

3. STRUCTURAL MODELING

It should be apparent from the above discussion that in the determination of the structure of amorphous solids, the use of experimental probes alone is insufficient, since only a rather limited amount of structural information can be obtained unambiguously in this way. An essential complementary activity, therefore, is to undertake modeling studies in which the structure of an amorphous material is simulated theoretically, using as input the limited amount of structural information (usually concerning SRO) available from experiment.

The cause of nonperiodicity in solids can be different depending on whether directional (covalent) or spherically symmetric (e.g., metallic) bonding is involved. For the case of *covalent* materials, it seems to be the case that amorphous structures arise as a result of a plethora of bonding configurations being available, which therefore allow crystallization to be bypassed. In other words, topological disorder results from the presence of a (quasi-)random distribution of dihedral (or torsion) angles, viz., the relative angles of rotation about a common bond to bring two neighboring structural units into coincidence (see Fig. 5). In addition, some static disorder in the bond-angle distribution $P(\theta)$ is also necessary to ensure complete connectivity (i.e., no dangling bonds) within the interior of such a disordered covalent structure. The ideal amorphous structure conforming to this picture is the "continuous random network" (CRN), first proposed by Zachariasen (1932).

In the case of amorphous metallic, molecular, or ionic systems, where central forces (metallic, van der Waals, or electrostatic, respectively) are predominant, other factors may influence glass formation, principal among which is that of "frustration." Small clusters of equally sized atoms tend to have noncrystallographic configurations as the lowest-energy structures, e.g., containing fivefold symmetry elements such as in the dodecagonal icosahedron, and aggregates of such clusters (for example, present in the liquid state) cannot form three-dimensional translationally periodic structures (i.e., crystals) on cooling. The ideal amorphous structure pertaining to this situation is thus the "dense random packing" (DRP) of spheres, first advanced by Bernal (1959, 1964).

A considerable number of modeling studies of the structure of amorphous materials have been performed for systems containing variously covalent, van der Waals, ionic, and metallic interactions. Furthermore, a variety of modeling techniques have been employed in such studies, the most important of which will be summarized below.

3.1 Serial-Addition Models

One of the conceptually most simple methods of constructing noncrystalline models is by the method of serial addition of atoms to an aperiodic seed. This approach has been followed in numerous studies involving the construction of hand-built models, following the pioneering work on covalent materials by Polk (1971) for the case of *a*-Ge and Bell and Dean (1972) for the case of *v*-SiO_2, and in the case of metallic (or other central-force systems) by Bernal (1959, 1964) and Finney

(1970). [For a review of such models, see Elliott (1984).] This adatom-by-adatom construction technique can also be achieved using computer algorithms, as in the work by Bennett (1972) on amorphous metallic systems, or that by Gladden and Elliott (1987, 1989) for covalent systems (for *v*-$SiSe_2$) and Gladden (1989) for *v*-SiO_2. The difficulty with the use of computers in constructing structural models in this way is to ensure that the resulting model has a sufficiently large density (and in the case of covalent systems a low concentration of dangling-bond defects).

3.2 Monte Carlo Methods

At the other extreme from the serial-addition method is the Monte Carlo technique, in which an existing structure (usually a crystalline polymorph of the material in question) is progressively disordered by a series of suitable randomization steps, involving, say, bond-switching for covalent structures. An example of this procedure is that used to generate the "sillium" model for *a*-Si starting from the diamondlike crystalline silicon lattice (Wooten and Weaire, 1984), and that employed in the simulation studies of the same system by Beeman and Bobbs (1975) in which the starting structure used was the even-ring tetrahedrally coordinated CRN model of Connell and Temkin (1974), and varying proportions of odd-membered rings were introduced by Monte Carlo means.

3.3 Molecular Dynamics

Potentially, the most physically realistic method of simulating the structure of amorphous solids is by using the molecular dynamics (MD) technique. In this, the time evolution of the classical motion of a collection of atoms (or molecules), interacting by means of a given interatomic potential, is calculated as a function of a change in temperature (or some other thermodynamic variable). A computer "liquid" is first generated by raising the temperature of a starting configuration of atoms, usually having a crystalline structure, confined within a box having periodic boundary conditions. A "glassy" structure can then be created by rapidly decreasing the temperature of the system. Attractive though this approach is in principle, in practice it suffers from a number of disadvantages insofar as the study of the properties and structure of amorphous solids (rather than of liquids) is concerned. One major limitation is the unavoidably very high rate of cooling ($>10^{10}$ K s^{-1}) characteristic of MD simulations; these values are essentially dictated by the minimum duration of the time step for a given configuration for which the simulation must be run ($\sim 10^{-15}$ s) and economic constraints on the total amount of computer time available. It is relatively easy to perform MD simulations using nondirectional (e.g., ionic or van der Waals) interactions, and a number of appropriate glass-forming systems have been simulated in this way (see, e.g., Angell *et al.*, 1977). However, such spherically symmetric (ionic) potentials are inappropriate for use in simulating the structure of network glasses in which covalency effects play a significant role. It is only very recently that three-body bond-bending terms have been included in the interatomic potential in an attempt to account for covalent interactions, as, for example, in *v*-SiO_2 (Feuston and Garofalini, 1988).

4. EXAMPLES OF THE STRUCTURE OF AMORPHOUS MATERIALS

In this section we will discuss the salient features of the structure of some important representative families of amorphous materials in the light of the preceding discussion of experimental and theoretical techniques for structure determination.

4.1 Amorphous Hydrogenated Silicon

By far the most important member of the family of amorphous tetrahedrally coordinated materials is hydrogenated silicon, *a*-Si:H, deposited in thin-film form by various chemical-vapor-deposition techniques from silane. The importance of *a*-Si:H lies in its various technological applications, arising from the fact that it can be doped electrically *p*- or *n*-type. Problems relating to the atomic structure of this material may be divided into three categories: those relating to the structure of the Si framework, those concerned with the H sublattice, and those relating to the structural environment of other alloying elements (e.g., C, Ge, or N) and dopants (e.g., P, B). An extensive review of recent structural investigations of *a*-Si:H has been given by Elliott (1989), and earlier work has been reviewed by Knights (1984).

Structural information concerning the Si framework of *a*-Si:H alloys can be obtained most straightforwardly from diffraction experiments. X-ray diffraction of *a*-Si:H is sensitive only to Si–Si correlations, i.e., the Si framework, since the light H atoms do not scatter x rays. The same information can be obtained from neutron scattering experiments utilizing isotopic substitution for the hydrogen. Since H and D isotopes have *oppositely* signed values of neutron scattering length, the appropriate null-element mixture of H and D results in scattering from Si–Si correlations only. Neutron diffraction for other isotopic mixtures will contain contributions from H–H and Si–H correlations as well; the use of three such isotopic substitution experiments therefore allows the individual partial distribution functions (PDFs) to be extracted, as shown in Fig. 6 (Bellisent *et al.*, 1985 a,b). The position of the first peak in the Si–Si PDF yields a value for the Si–Si bond length, namely, 2.34 Å. The area under the peak gives an estimate for the coordination number for Si atoms surrounding a given Si atom of $N_1 \approx 3.8$; the discrepancy between this value and the value of 4 expected for ideal tetrahedral coordination reflects, of course, the bonding of hydrogen to some Si atoms as well as, to a lesser extent, the influence of coordination defects (e.g., dangling bonds). EXAFS measurements, taken at the Si *K* edge (Menelle *et al.*, 1986; Filipponi *et al.*, 1986a) also confirm these findings for the first coordination shell of the Si framework structure.

A well-defined second peak lying at 3.85–3.9 Å also occurs in the Si–Si PDFs obtained from diffraction data (see Fig. 6); from the values of the first- and second-neighbor distances, a value for the Si bond angle of $\theta = 110°$ can be inferred. However, the width of the second peak is considerably greater than that of the first peak in the Si–Si PDF; the extra width is due to the *static* disorder in the structure associated with a finite width of the Si–Si–Si bond-angle distribution, in addition to the broadening factors due to experimental resolution (finite value of $k_{\max}$) and thermal vibrational disorder of the bonds. Typically, rms fluctuations of the Si bond angle of $\Delta\theta \approx 7°–10°$ are inferred from diffraction data.

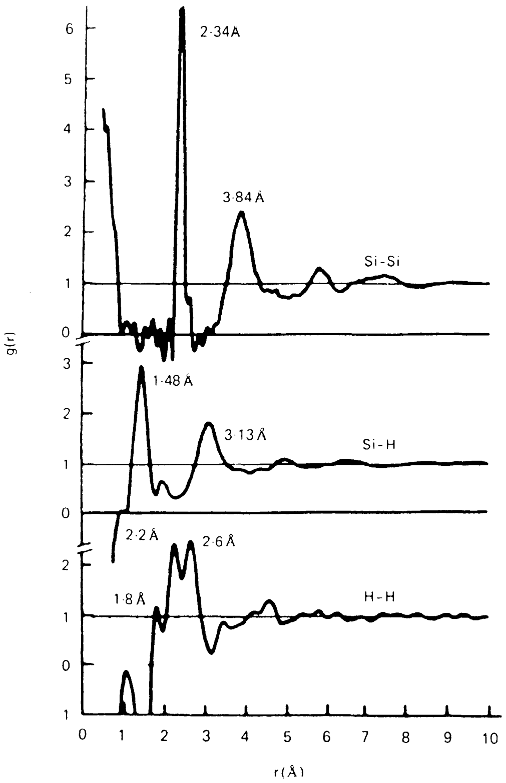

FIG. 6. Partial pair distribution functions for sputtered a-Si$_{84}$(H,D)$_{16}$ obtained from isotopic-substitution neutron-scattering experiments (Bellisent *et al.*, 1985 a,b).

Estimates for the magnitude of Si bond-angle fluctuations in *a*-Si:H can also be obtained from vibrational spectroscopy, notably Raman scattering. The quantities of interest are the width (FWHM) Γ and position ω_{TO} of the TO-like Raman peak at ~500 cm^{-1} [see Fig. 7(a)], which experimentally are observed to depend on the amount of (bond-angle) strain in the films (Maley *et al.*, 1983; Tsu, 1987). The relationship between Γ and $\Delta\theta$ has been investigated theoretically for a number of CRN models of *a*-Si containing varying amounts of bond-angle strain by Beeman *et al.* (1985) [see Fig. 7(b)] and the behavior of the calculated Raman peak widths can be accounted for by two empirical relations, namely,

$$\Gamma = a + b\Delta\theta, \qquad (24a)$$

proposed by Beeman *et al.* (1985), and the similar expression proposed by Tsu *et al.* (1984),

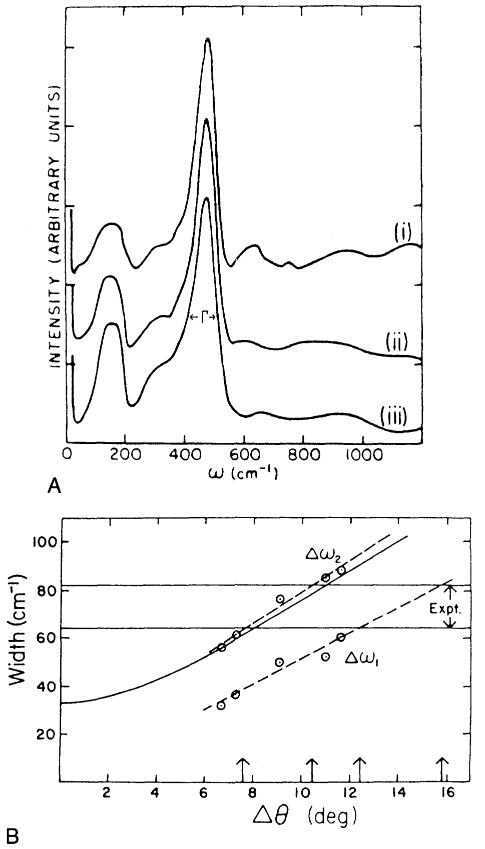

FIG. 7. (a) Experimental VH-polarized Raman spectra of a-$Si_{1-x}H_x$ (Maley *et al.*, 1983): (i) $x=0$; (ii) $x=0.4$ at. %; (iii) $x=32$ at. %. (b) Theoretical TO-like peak widths $\Delta\omega_1$ and $\Delta\omega_2$ (calculated for two different models of the Raman mechanisms), plotted against the rms bond-angle fluctuation $\Delta\theta$ for a number of models of a-Si (Beeman *et al.*, 1985). The curved solid line is given by Eq. (24b) and the dashed straight lines are fits according to Eq. (24a), where $\Gamma=\Delta\omega$ in both cases.

$$\Gamma^2=c+d(\Delta\theta)^2, \tag{24b}$$

where a, b, c, and d are appropriate constants. The intersections of the experimental TO-like peak widths with the theoretical predictions shown in Fig. 7(b) provide estimates for the rms bond-angle fluctuations in the range $\Delta\theta\sim8°$–$16°$, depending on the exact Raman mechanism assumed to be operative. Tsu (1987) has shown that there is also a correlation between the experimental values of the position of the TO peak and its width (and hence $\Delta\theta$), obeying the relation

$$\omega_{\mathrm{TO}}=\omega_0-k\Gamma_\theta^2, \tag{25}$$

where ω_0 is the unshifted (strain-free) frequency of the TO peak, k is a constant, and Γ_θ is the peak-width contribution associated *directly* with the bond-angle strain, namely, $\Gamma_\theta^2 = d(\Delta\theta)^2$ [cf. Eq. (24b)]. The form of Eq. (25) can be understood theoretically in terms of the vibrational characteristics of anharmonic potentials (Tsu, 1987).

Although information on the SRO of the Si-related structure of a-Si:H can be obtained fairly unambiguously from diffraction or EXAFS data (and to a limited extent, Raman spectra), the same is *not* true for the case of MRO. Although tetrahedrally coordinated amorphous materials exhibit considerably less MRO than other amorphous systems, e.g., chalcogenide glasses, nevertheless a small feature in the Si–Si PDF at ~4.8 Å has been regarded as being a signature of MRO. For instance, Dixmier *et al.* (1985) have suggested that the incorporation of two Si–H monohydride groups into the structure, such that the two Si atoms involved are members of the same ring, invariably leads to the creation of an ovoid-shaped void around the hydride units formed by rings characterized by intermediate values of dihedral angle, viz., $\phi=35°$–$45°$. However, detailed modeling studies to confirm this prediction are wanting.

Of interest also is the structural environment of the hydrogen in a-Si:H, the presence of which (up to ~10 at. %) is necessary to suppress the Si dangling-bond density in order to reduce concomitantly the electronic density of states in the band gap of the material so as to achieve the optimum electronic performance. Three distinct H-related configurations can exist in principle in a-Si:H, namely, monohydride (Si–H), dihydride (Si–H_2), and trihydride (Si–H_3) groups. In addition, there is evidence that material containing a large amount of H should be regarded as being a "polysilane," denoted $(SiH_2)_n$.

One technique which can distinguish between the various H configurations is infrared (IR) vibrational spectroscopy, and it is useful in this regard since the phonon modes associated with the Si framework are restricted to the range of frequencies $0<\omega<550$ cm^{-1} (or energies $0<E<70$ meV), the maximum phonon frequency being given by the expression

$$\omega_{\max}=(8\alpha/m)^{1/2}, \tag{26}$$

where a is the bond-stretching force constant and m is the atomic mass, whereas the vibrational modes involving H in a-Si:H occur at considerably higher frequencies ($600 < \omega < 2100$ cm^{-1}) due to the light mass of H. Therefore, H-related vibrational modes can be identified unambiguously from IR absorption spectra. For reviews of this topic, see Lucovsky and Pollard (1984) and Lucovsky (1985).

Figure 8(a) shows a schematic representation of the assignment of features in the IR absorption spectrum of a-Si:H in terms of the various H-related configurations, the local and resonance modes of which are indicated in Fig. 8(b). It can be seen from Fig. 8(a) that the different H-related configurations can be distinguished in this way. Although it was originally thought that trihydride groups existed in H-rich material, it is now generally accepted that in such cases polysilane groupings are predominant (Lucovsky, 1985). Thus, IR spectroscopy can readily be used as a diagnostic tool in order to check the quality of a-Si:H films: the best electronic-grade material, containing essentially only monohydride species, is easily distinguished by the lack of features in the IR spectrum at ~900 cm^{-1} due to scissor-bending modes of SiH$_2$ groups.

Although IR spectroscopy can give quantitative information concerning the concentration of H-related configurations present in a sample of a-Si:H, it can provide no other direct structural information. For this purpose, isotopic-substitution neutron scattering can beemployed, using isotopic mixtures of hydrogen and deuterium, as in the work of Bellisent *et al.* (1985a,b), the results of which are shown in Fig. 6. From the Si–H partial PDF shown in Fig. 6, the Si–H bond length is found to be ~1.48 Å, and the Si–Si–H bond angle (averaged over all the Si–H and Si–H$_2$ configurations in the samples) is ~107.9°, appreciably less than the tetrahedral angle of 109.5°. The first peak in the H–H partial PDF at 2.2–2.6 Å can be interpreted as being due to H–H correlations within dihydride Si–H$_2$ units; if this interpretation is correct, a value for the average H–Si–H bond angle of 109.5° can be inferred (Elliott, 1989). However, neutron experiments are complicated in the case of H-containing materials in view of the large corrections necessary to account for the substantial incoherent scattering of H.

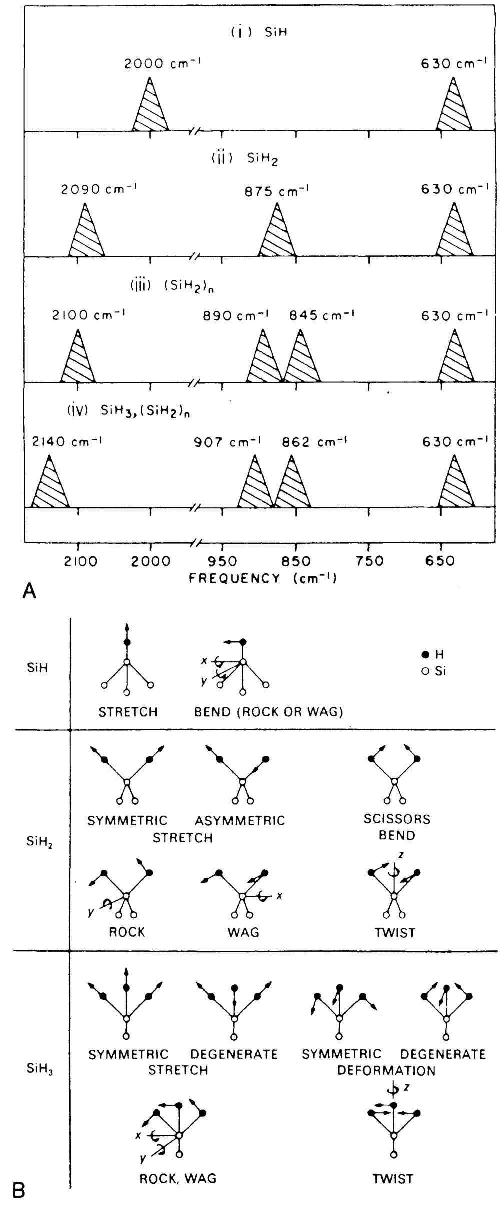
(i) SiH
2000 cm^{-1}
630 cm^{-1}
(ii) SiH$_2$
2090 cm^{-1}
875 cm^{-1}
630 cm^{-1}
(iii) (SiH$_2$)$_n$
2100 cm^{-1}
890 cm^{-1}
845 cm^{-1}
630 cm^{-1}
(iv) SiH$_3$,(SiH$_2$)$_n$
2140 cm^{-1}
907 cm^{-1}
862 cm^{-1}
630 cm^{-1}
2100
2000
950
850
750
650
FREQUENCY (cm^{-1})
A
SiH
STRETCH
x
y
BEND (ROCK OR WAG)
● H
○ Si
SiH$_2$
SYMMETRIC
ASYMMETRIC
STRETCH
SCISSORS BEND
y
ROCK
x
WAG
z
TWIST
SiH$_3$
SYMMETRIC
DEGENERATE
STRETCH
SYMMETRIC
DEGENERATE
DEFORMATION
x
y
ROCK, WAG
z
TWIST
B

FIG. 8. (a) Schematic representation of the assignment of features in the IR absorption spectrum of a-Si:H in terms of various H-related structural groupings (Lucovsky and Pollard, 1984). (b) Local and resonance modes for various H-related configurations in a-Si:H (Lucovsky *et al.*, 1979).

Proton NMR can also be used to provide some structural information, particularly concerning the H-related structure (see Reimer, 1981, for a review). This is in contrast to the case for ^{29}Si NMR, where the static disorder (in the Si bond-angle distribution) results in broad, featureless resonance lines which

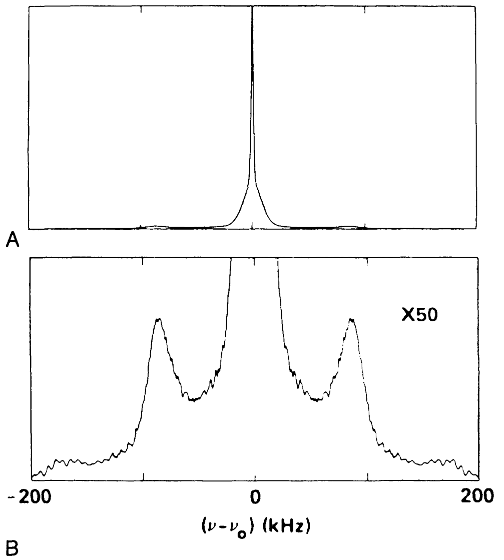

FIG. 9. (a) ^{1}H NMR spectrum of *a*-Si:H taken at 1.43 K showing the narrow and broad components of the central line, and the satellite lines due to the Pake doublet characteristic of solid H_2 (Boyce and Stutzmann, 1985). (b) An enlargement of (a), showing the Pake doublet more clearly.

are devoid of any discernible structural features, although the *width* of the ^{29}Si NMR lines appears to be governed by the fluctuations in the Si bond angle and consequently may be used to obtain an estimate for $\Delta\theta$ (Jeffrey *et al.*, 1981). The ^{1}H NMR spectrum is much more informative (see Fig. 9). At all but the very lowest temperatures, the NMR line is characterized by a combination of both a broad and a narrow component, the relative proportion of which depends both on the total concentration of H in the material and on the relative abundance of H-related configurations. Since the width of the NMR line in this case is determined by the homonuclear dipolar interaction between the nuclear spins of neighboring protons ($I=1/2$), which depends on the internuclear distance r as r^{-6}, clustered protons will experience a large dipolar interaction and hence lead to a broadening of the ^{1}H NMR line, whereas isolated (monohydride) groups will experience essentially no dipolar interactions and hence these H-related configurations are the origin of the narrow component of the line (Carlos and Taylor, 1982a,b).

At temperatures below $\sim$10 K, another component to the ^{1}H NMR line emerges [see Fig. 9(b)]: this has been identified as the Pake powder-pattern signature of solid *molecular* ortho-H_2 trapped in microvoids of size 10–20 Å in the amorphous silicon framework (Boyce and Stutzmann, 1985). This finding has been confirmed by a number of other techniques, including calorimetric measurements and IR spectroscopy (see Elliott, 1989, for a review).

Finally, we consider the structure of those alloying components (e.g., C, Ge, or N substituting for Si, or F substituting for H) and dopants (e.g., P or B) introduced deliberately into the material so as to alter, say, its optical or electrical properties. In such cases, where at least ternary materials are involved, atom-specific structural probes are essential in order to obtain structural information. In this regard, EXAFS and NMR have proved to be useful.

In the case of Si-substitutional alloy systems, EXAFS measurements have provided a great deal of information (see Mobilio and Filipponi, 1987, for a recent review). In Fig. 10(a), Fourier-transformed Si *K*-edge EXAFS data are shown for *a*-$Si_{1-x}N_x$:H thin-film alloys; the emergence of a Si–N peak as the N content increases is clearly apparent. From these results, the compositional dependence of the Si–Si and Si–substituent bond lengths and coordination numbers may be deduced. Figure 10(b) shows the compositional dependence of the average relative first-shell Si coordination numbers, $C = N_A/(N_{\mathrm{Si}} + N_A)$, so obtained for Si–Ge and Si–N alloys, from which it can be seen that Ge randomly substitutes for the Si whereas the N prefers a chemically ordered environment, both as in the corresponding crystals.

In the case of dopant impurities (e.g., P or B), knowledge of the SRO structural environment around the dopant can provide important information concerning the mechanism of doping in *a*-Si:H. For instance, As *K*-edge EXAFS studies of As-doped *a*-Si:H by Knights *et al.* (1977) have indicated that the average coordination of As is nearer to three than to four, implying that most As atoms occupy electrically inactive, normally bonded threefold coordinated sites rather than tetrahedrally coordinated, substitutional doping sites, in agreement with the experimental finding that the doping efficiency of *a*-Si:H is very low. NMR studies of, say, ^{31}P in P-doped *a*-Si:H also provide evidence for the finding that the majority of impurity atoms reside in

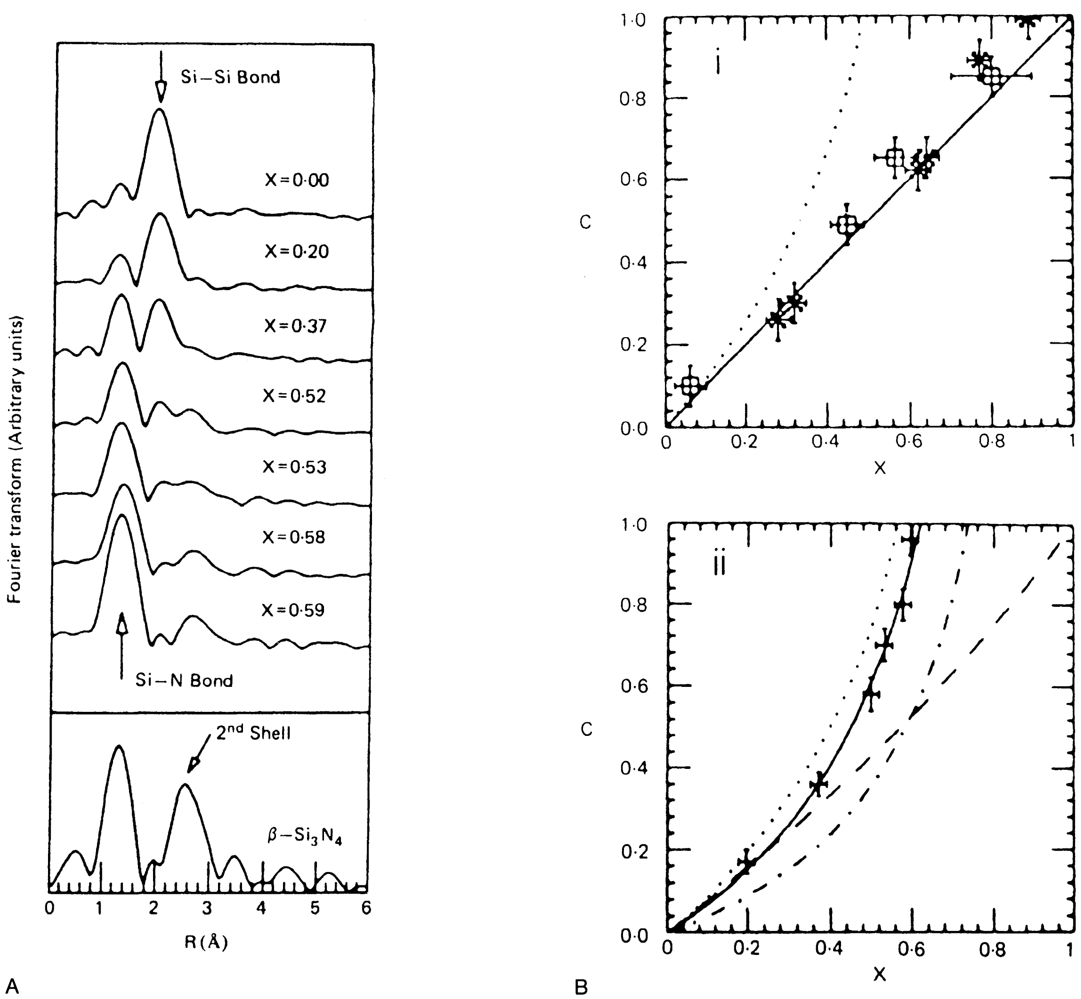

FIG. 10. (a) Fourier-transformed Si *K*-edge EXAFS data for *a*-$Si_{1-x}N_x$:H films and crystalline β-Si_3N_4 (Mobilio and Filippone, 1987). The Fourier transforms have been performed not taking into account phase-shift factors. (b) Average relative first-shell Si coordination numbers, $C = N_A/(N_{Si} + N_A)$, obtained from EXAFS data as a function of composition for amorphous silicon alloys (Mobilio and Filipponi, 1987): (i) *a*-$Si_{1-x}Ge_x$:H (solid line, random alloy model; dotted line, chemically ordered model); (ii) *a*-$Si_{1-x}N_x$:H [chemically ordered model incorporating H (–); chemically ordered model (···); random-alloy model (---) chemically ordered model with wrong bonds (-··-)].

sites satisfying their normal valence (Reimer and Duncan, 1983).

4.2 Oxide Glasses

Oxide materials, particularly silicates, are perhaps the canonical type of glass-forming systems. In view of this fact, and since these materials also have very many technological applications in a wide range of fields, it is important to understand their atomic structure and its relation to other physical properties.

At the outset it is useful to distinguish between two types of oxide materials: "network formers" are those materials which by themselves are glass formers (e.g., SiO_2, GeO_2, B_2O_3, P_2O_5, etc.), whereas "network modifiers" cannot by themselves form glasses, but do so when combined with a network former, and in so doing modify in some way the structure of the glass. There are other oxide materials which fall into neither of these categories. Such material, e.g., Al_2O_3, TiO_2, PbO_2, etc., are not, by themselves, glass formers, but they can be made into glasses if mixed with a network-forming material. Hence, they may be termed "conditional" glass formers, since the cations involved play a network-forming rather than -modifying role. It is important, therefore, to investigate the structure of both unmodified and modified oxide glasses.

For stoichiometric compositions of unmodified network-forming glassy systems, heteropolar bonds are dominant, and thus the principal factor determining the SRO is the number N_{co} of oxygen ligands surrounding each glass-forming cation or, in other words, the shape of the cation-centered coordination polyhedron characterizing the structure. For simple (binary) oxide glasses, the use of diffraction techniques is sufficient in order to ascertain N_{co} from the measured average coordination number (i.e., the area under the first peak in the *total* RDF), since it is safe to assume that the oxygen atoms will always be in "bridging sites," i.e., be twofold coordinated by cationic species (viz., $N_{oc}=2$). Thus, it is well established, for example, that Si in SiO_2 and Ge in GeO_2 are tetrahedrally coordinated and B in B_2O_3 is (planar) trigonally coordinated by bridging oxygen atoms, whereas P in P_2O_5 is tetrahedrally coordinated by three bridging oxygens and one double-bonded nonbridging oxygen atom. The tetrahedral coordination of Si by O, for example, is not only indicated by diffraction data (e.g., x-ray by Mozzi and Warren, 1969, and neutron by Wright and Sinclair, 1985), but also corroborated by ^{29}Si MASNMR experiments, where the chemical shift of the peak of the NMR line in vitreous SiO_2 (~ -111.5 ppm referenced to tetramethyl silane; see Gladden *et al.*, 1986) is practically identical to that characteristic of various crystalline polymorphs of silica in each of which Si is tetrahedrally coordinated (Smith and Blackwell, 1983).

Another SRO parameter of great interest is the oxygen bond-angle distribution $P(\theta_o)$. (The average cation-centered bond angle θ_c is constrained by symmetry to take only certain values in the case of very symmetrical coordination polyhedra, such as tetrahedra or planar triangles.) In the absence of isotopic-substitution neutron scattering data (or other diffraction results similarly yielding partial PDF information), extraction of $P(\theta_o)$ is very difficult to achieve using *total* diffraction data because of the problems of coordination-shell overlap in the region of the second peak in the RDF. However, Mozzi and Warren (1969) managed to obtain an estimate for $P(\theta_o)$ in the case of v-SiO_2 from their x-ray diffraction results after making a number of simplifying assumptions. This distribution, shown in Fig. 11, peaks at an angle $\theta_o \approx 145°$, and it can be seen that $P(\theta_o)$ is both broad and markedly asymmetric, being skewed to the high-angle side.

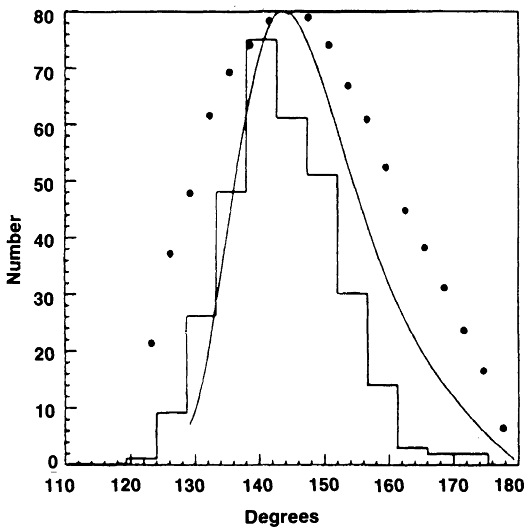

FIG. 11. Si–O–Si bond-angle distribution for v-SiO_2 obtained from x-ray diffraction data (·) (Mozzi and Warren, 1969), ^{29}Si MASNMR (–) (Gladden *et al.*, 1986), and the histogram distribution of a computer-generated CRN model (Gladden 1989).

For the case of v-SiO_2, MASNMR can also be used to obtain quantitative information concerning $P(\theta_o)$ by using the fact that the bond angle θ_o subtended at the oxygen ligands surrounding the resonant ^{29}Si nucleus subtly affects the ^{29}Si isotropic chemical shift σ. The empirical relation between σ and θ_o [Eq. (23)], deduced for the case of *crystalline* polymorphs of SiO_2 (Smith and Blackwell, 1983), can be assumed also to hold in the case of the inhomogeneously broadened oxygen bond-angle distribution, and hence NMR line, characteristic of v-SiO_2 (Dupree and Pettifor, 1984). Use of Eq. (23) therefore allows an estimate of $P(\theta_o)$ to be obtained from an analysis of the line shape of the ^{29}Si NMR signal, as shown in Fig. 11 (Gladden *et al.*, 1986). It can be seen that $P(\theta_o)$ obtained in this way is as similarly asymmetric as, but somewhat narrower than, that obtained from x-ray diffraction (Mozzi and Warren, 1969).

The fact that there is a wide distribution of oxygen bond angles in v-SiO_2 (see Fig. 11) implies that few constraints exist in the interconnection of neighboring SiO_4 tetrahedra other than those imposed by the steric constraints arising from the avoidance of a nonbonded oxygen atomic overlap (Galeener, 1985). This is consistent with *corner*-sharing connections between SiO_4 tetrahedra. This relative freedom of connection of structural units implies the existence of a minimal degree of MRO in the structure of v-SiO_2, and this is confirmed by the extensive structural modeling studies of Gladden (1989), in which a good fit to the experimental RDF obtained from neutron diffraction is achieved for models having a range of ring sizes, with a preference for 8- and 12-atom rings.

The situation in v-B_2O_3 is rather different, since here pronounced IRMRO in the form of superstructural units (see Sec. 1.4.2) seems to exist. The particular type of superstructural unit involved is the so-called "boroxol" ring, namely, a planar six-atom regular ring, which can be regarded as being constructed from three corner-sharing planar BO_3 triangles connected together in a triangular arrangement (see Fig. 12). The experimental evidence for the existence of boroxol rings in v-B_2O_3 comes from Raman scattering: the Raman spectrum is dominated by a very narrow, intense, strongly polarized line at $\sim 808\ cm^{-1}$ [Fig. 12(a)]. Isotopic substitution experiments by Windisch and Risen (1982) indicate that there is no change in the 808-cm^{-1} line on changing the B isotope, but that there *is* a pronounced shift in frequency when the O isotope is changed [Fig. 12(b)]. These results imply that only oxygen motion is involved in the vibrational mode responsible for the 808-cm^{-1} Raman peak, and the 1:3:3:1 intensity ratio for the resolved quadruplet line when a 50:50 mixture of ^{16}O and ^{18}O isotopes is used is consistent with a random isotopic substitution in a structural unit containing just three oxygen atoms. The strongly polarized nature and narrowness of the 808-cm^{-1} peak [see Fig. 12(a)] are also indicative of a highly symmetric, vibrational mode which is decoupled from the rest of the network.

The case of *modified* oxide glasses is considerably more complicated. Here, information on the structural influences of the modifying cation (often an alkali atom) on both the network-forming cation-centered coordination polyhedra and also on their interconnections is sought, together with that concerning the SRO structural environment of the network-modifier cations themselves. Thus, modifier-induced coordination changes of network-forming cations can occur, as can a depolymerization of the network by the transformation of bridging oxygen (BO) twofold-coordinated sites to "nonbridging" oxygen (NBO) singly coordinated (and negatively charged) anionic configurations. Atom-specific structural probes are obviously more useful in the structure determination of such ternary or more complicated systems than are conventional diffraction techniques.

In the case of alkali silicate glasses, addition of the alkali modifier ions causes a depolymerization of the silicate framework with the formation of NBO sites but with no change in the coordination of the Si. This observation is supported by the results of ^{29}Si MASNMR experiments (see, e.g., Dupree *et al.*, 1984), where the value of the isotropic chemical shift is approximately the same as that for pure v-SiO_2, indicating that the Si coordination remains unaltered. However, there *are* subtle changes to the ^{29}Si NMR line shape, as seen in Fig. 13(a), indicating changes in the oxygen coordination shell. There are *five* possible configurations involving the incorporation of NBO atoms in the tetrahedrally coordinated Si polyhedra, conventionally denoted Q_n species where $n=0$–4 is the number of BO atoms around a given Si atom. It might be thought that the incorporation of network-modifying cations in the glassy network, and concomitantly the distribution of NBO and Q_n species, would occur in a purely statistically random way; in such a case, for a given modifier content, it would be expected that *all* five Q_n species would occur, the relative proportion of each being determined by the composition. The MASNMR results [Fig. 13(a)] indicate otherwise; at stoichiometric compositions, essentially only *one* type of Q_n species exists, with the two species Q_n and Q_{n+1} (corresponding to the endpoint compositions) predominating at intermediate compositions [Fig. 13(b)]. Nevertheless, *static* ^{29}Si NMR experiments have indicated some discrepancies with this simple ordered picture (Stebbins, 1987). For the case of binary $(Na_2O)_x(SiO_2)_{1-x}$ glasses, it was found from static NMR spectra that Q_4 species existed for compositions close to stoichiometry ($x=0.33$) where the simple picture would predict the presence of only Q_3 units. Figure 14 clearly shows the emergence of a

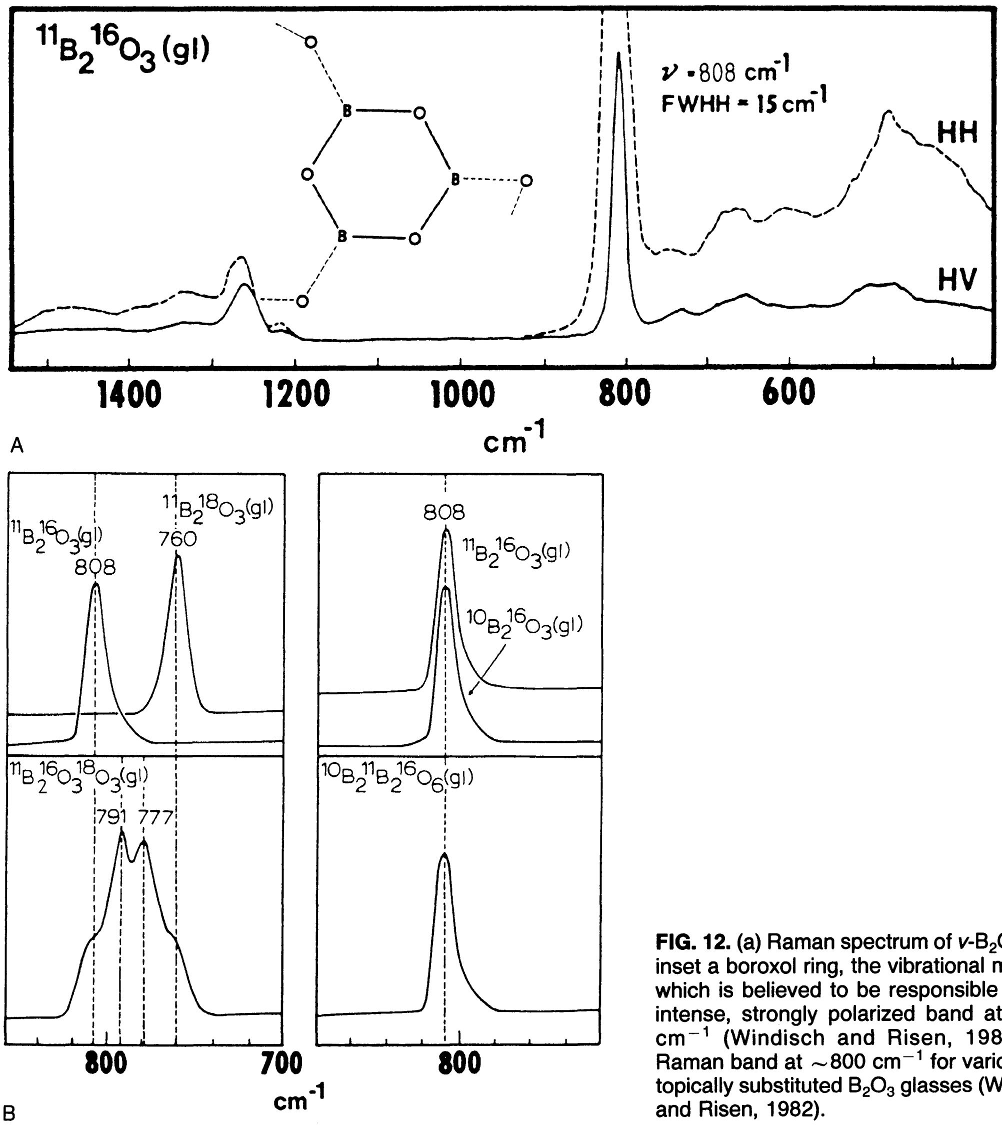

FIG. 12. (a) Raman spectrum of v-B_2O_3, with inset a boroxol ring, the vibrational mode of which is believed to be responsible for the intense, strongly polarized band at ~808 cm^{-1} (Windisch and Risen, 1982). (b) Raman band at ~800 cm^{-1} for various isotopically substituted B_2O_3 glasses (Windisch and Risen, 1982).

peak, identified as being due to Q_4 units, superimposed upon the axially symmetric powder pattern characteristic of Q_3 units (with C_{3v} symmetry, the axis of symmetry being along the bond connecting a Si atom and a neighboring NBO).

In certain exceptional cases, a change in the coordination of the Si can be effected with the addition of other network formers and modifiers, the coordination becoming octahedral as in the high-pressure polymorph of silica, stishovite, and in the crystal $Si_2P_2O_7$. For example, it has been found from ^{29}Si MASNMR studies (Dupree *et al.*, 1987) that $Na_2Si_2O_5$ glasses containing in excess of ~30 mol % of P_2O_5 contain an increasingly large proportion of octahedrally coordinated Si sites (see Fig. 15). The strongly polarizing P(V) atoms can stabilize electrostatically the change in coordination.

In other oxide glass-forming systems, where the radius ratio of the network-forming cation is more favorable, compositionally induced coordination changes occur more easily. Thus, for example, a transformation to octahedral coordination of Ge occurs in alkali germanate glasses with increasing alkali content (up to ~35 at. % alkali), as evidenced by Ge *K*-edge EXAFS studies (Sakka and Kamiya, 1982); at higher alkali contents,

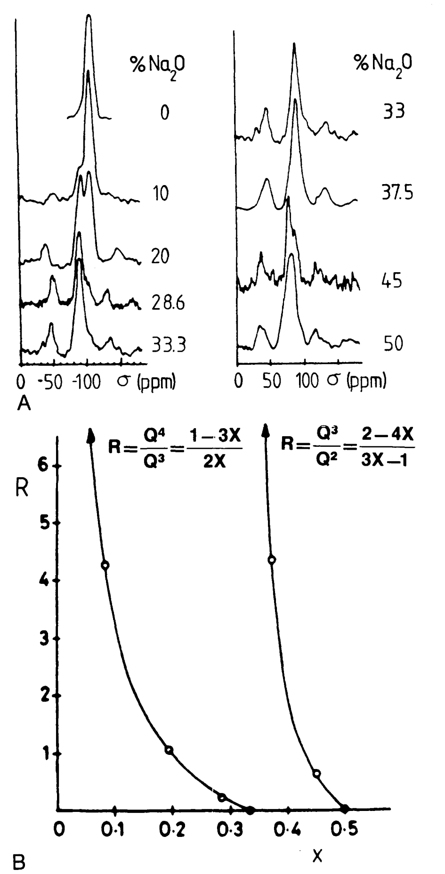

FIG. 13. (a) ^{29}Si MASNMR spectra for binary $(Na_2O)_x(SiO_2)_{1-x}$ glasses (Dupree *et al.*, 1984). (b) Calculated (curves) and experimental (circles) intensity ratios for Q_n species contributing to the ^{29}Si MASNMR line shape for different compositions (Dupree *et al.*, 1984).

NBO sites are created at the expense of octahedral Ge sites. The case of borate glasses is very similar. As discussed previously, B is planar threefold coordinated in pure v-B_2O_3, but the addition of alkali oxides causes a change from trigonal to tetrahedral coordination of B up to a limiting content of ~35 at. % alkali; thereafter the B coordination reverts to being trigonal and NBOs are created as charge-compensating centers for the alkali ions. ^{11}B NMR experiments, even static ones, are very sensitive to such coordination changes because of the concomitant changes in the quadrupolar interaction between the electric field gradient (EFG) at the nucleus caused by the electronic charge distribution in the atom and the nuclear quadrupolar moment, and the broadening effect this has on the NMR line. Trigonally coordinated B atoms will have large EFG values (and hence very broad lines), whereas the symmetry associated with tetrahedral coordination significantly reduces the value of the EFG, leading to characteristically narrow NMR lines. A review of this topic has been given by Bray (1978).

Thus far, we have concentrated mainly on aspects of the SRO of (modified) oxide glasses without discussing MRO in any detail. With the exception of v-B_2O_3, short-range and intermediate-range MRO (Secs. 1.4.1 and 1.4.2) is not particularly pronounced in oxide glasses, principally because of the inherent flexibility imparted to the structure by the corner-sharing connections of coordination polyhedra characteristic of most oxide glasses. However, the network depolymerization in, say, silicates caused by the modifier-induced creation of NBO centers can lead to the formation of long-range MRO (see Sec. 1.4.3). In particular, for, say, the composition $(Na_2O)_{0.33}(SiO_2)_{0.67}$, where Q_3 units predominate, the structure locally is forced to be *two dimensional* on topological grounds, as in the corresponding layerlike crystal. This type of reduced dimensionality, delineated by chan-

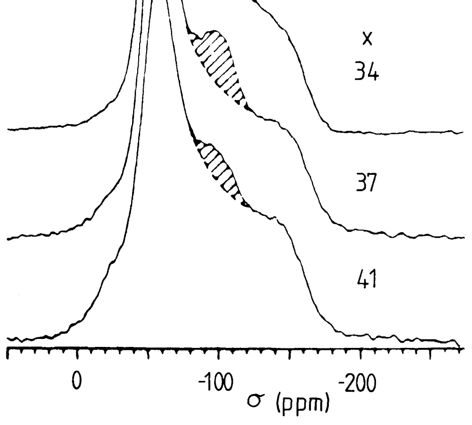

FIG. 14. Static ^{29}Si NMR spectra for $(Na_2O)_x(SiO_2)_{1-x}$ glasses with compositions as marked (Stebbins, 1987). Note the increasing hatched contribution due to Q_4 species.

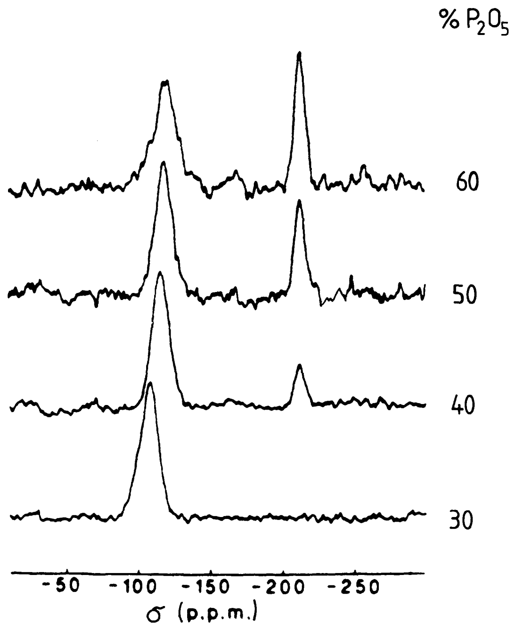

FIG. 15. ^{29}Si MASNMR spectra of $Na_2Si_2O_5$ glasses containing varying amounts of P_2O_5 (Dupree *et al.*, 1987). At low P_2O_5 contents, only tetrahedrally coordinated Si is present (with a chemical shift of ~ -110 ppm), whereas at higher concentrations, a peak at ~ -210 ppm appears, characteristic of octahedral coordination.

nels containing the modifier cations surrounded by NBOs, is the basis of the concept of the "modified" CRN proposed by Greaves (1985) (see Fig. 16).

4.3 Chalcogenide Glasses

Chalcogenide glasses are those inorganic compounds of chalcogen (group VI) elements (viz., S, Se, and Te) with elements such as P, As, Si, or Ge. Although oxygen is also a group VI element, and so technically a chalcogen, nevertheless it is conventional to discuss separately oxide and other chalcogenide materials, partly for historical reasons, and also because oxide materials are much more ionic than the more covalentlike chalcogenide materials. Furthermore, because the Si- and Ge-containing chalcogenide analogs of silica can have edge-sharing as well as corner-sharing connections between neighboring tetrahedra, the MRO of these glasses can be much richer than that found in the corresponding oxide glasses; the local two dimensionality imparted to the structure by the inclusion of pnictogen atoms (P, As, Sb), resulting from the appreciable "bond-free solid angle" associated with their threefold coordination, has much the same effect. Finally, since chalcogenide materials have predominantly covalent bonding, the degree of chemical ordering (heteropolar bonding) can be much less pronounced than is found in oxide systems, particularly when the elements involved have similar electronegativity values.

Diffraction experiments on chalcogenide glasses, particularly for stoichiometric compositions, confirm that in general the basic coordination polyhedra in these materials conform to the expected values of the network-forming cations, viz., tetrahedral for Si- and Ge-containing materials, and trigonal pyramidal for pnictogen-containing glasses. The larger sizes of the chalcogen atoms compared with oxygen, however, mean that increases in coordination of the network-forming cations caused by the addition of network-modifying atoms, which are found in oxide systems (Sec. 4.2), are not observed in chalcogenide glasses.

However, it is much more difficult to ascertain the extent of chemical ordering in chalcogenide glasses using diffraction methods alone. In the case of Ge chalcogenide glasses, Mössbauer spectroscopy has been used to investigate the breakdown of chemical order (see Boolchand, 1985, for a review), using ^{119}Sn and ^{129}I as probes for Ge and chalcogens, respectively. In the former, Sn is assumed to substitute isomorphously for Ge at

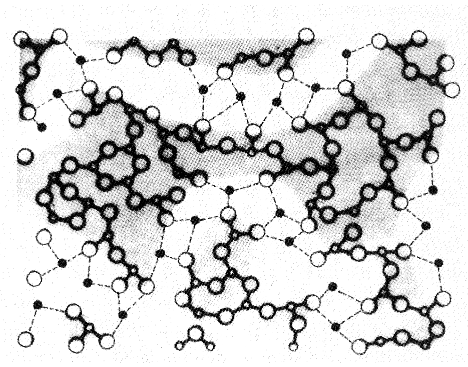

FIG. 16. Modified CRN model for alkali oxide glasses, proposed by Greaves (1985). The unmodified regions of the network are shown shaded, and the bonds between the modifier cations (solid circles) lying in the channels and nonbridging oxygen anions are shown as dashed lines.

the 1% level, and ^{119}Sn absorption Mössbauer spectra indicate the presence of two Ge sites even at the stoichiometric composition (e.g., $GeSe_2$) (Boolchand, 1985): the so-called A site is characterized by a single Mössbauer line and is ascribed to a symmetrical tetrahedral $Ge(Se_{1/2})_4$ unit; the other, B site is characterized by a weaker quadrupolar doublet Mössbauer line, and has been ascribed to an "ethanelike" structural configuration, viz., $Ge_2(Se_{1/2})_6$, i.e., containing *homopolar* Ge–Ge bonds. The Mössbauer experiments involving ^{129}I are somewhat complicated but yield similar complementary results. In these, radioactive $^{129}Te^m$ atoms, incorporated into, say, a $GeSe_2$ glass in an assumed isomorphous manner, decay by β emission into $^{129}I^*$ atoms which subsequently emit recoilless γ rays, the spectrum of which is detected by a monoenergetic ^{129}I absorbing detector (e.g., NaI). Such Mössbauer results provide evidence for two types of site originally occupied by the Te (and hence Se) atoms before their transmutation to I atoms even at the stoichiometric composition (Boolchand, 1985): the A sites correspond originally to chemically ordered $Te(Ge_{1/4})_2$ sites, and the B sites correspond to sites in which the chemical order is broken, viz., Ge–Se–Se configurations, with Te substituting for one or other of the Se atoms.

As usual, although the SRO in chalcogenide glasses is more or less well understood, the type and extent of MRO remains controversial. Diffraction evidence for MRO remains scanty and difficult to interpret, although one almost universal feature observed in the diffraction pattern has been ascribed to the existence of MRO: this is the so-called "first sharp diffraction peak" (FSDP) or "prepeak," observed at $k \sim 1$ Å^{-1} in the structure factor of chalcogenide glasses. The term prepeak is used because removal of this first peak in the structure factor causes no discernible difference in the real-space RDF obtained by Fourier transformation of $S(Q)$; the FSDP therefore seems to be a signature of structural ordering (i.e., MRO) beyond that which is registered in the RDF. In fact, oxide glasses such as v-SiO_2 and B_2O_3 also exhibit FSDPs in their structure factors but at somewhat larger values of scattering vector than is the case for chalcogenide glasses. However, when plotted on a common reduced scale, viz., with the scattering vector multiplied by the nearest-neighbor bond length r_1, scattering data for both chalcogenide and oxide glasses are very similar (see Fig. 17), indicating perhaps a common cause for the FSDP. It is noteworthy also that the tetrahedrally coordinated materials, a-Si and Ge, as well as the pure chalcogen glass, a-Se, do *not* exhibit FSDPs in their diffraction patterns.

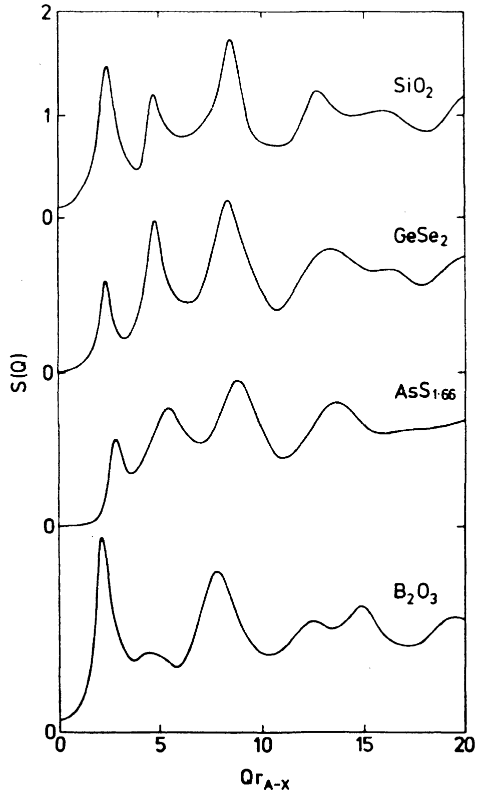

FIG. 17. Structure factors for chalcogenide and oxide glasses, plotted on a common reduced scale involving the nearest-neighbor bond length r_1, showing similar first sharp diffraction peaks (Wright *et al.*, 1985).

The structural origin of the FSDP remains obscure, however. One explanation treats the FSDP in reciprocal space as a single Bragg-like peak resulting from a quasiperiodic structural repeat in real space as, for example, in a layerlike structure. Busse (1984) has proposed this model for the case of As-chalcogenide glasses, and more recently Cervinka (1988) has extended this approach to the case of Ge-chalcogenide glasses where it is suggested that parallel arrangements of clusters of between two and five tetrahedral coordination polyhedra are responsible for the FSDP. However, the fact that the FSDP is observed even in the *liquid* state of chalcogenide materials (Uemura *et al.*, 1978) makes it unlikely that well-defined interlayer correlations over an appreciable correlation length of ~20 Å (necessary to account for the narrowness of the FSDP) are in fact responsible. Another explanation for the FSDP is based on intercluster correlations, as used in the case of molecular liquids. Here, the measured structure factor is regarded as arising from two terms:

$$S(Q)=f_m(Q)+D_m(Q), \qquad (27)$$

where $f_m(Q)$ is the "molecular" form factor describing the *intra*cluster scattering, and $D_m(Q)$ is the molecular-packing structure factor describing the *inter*cluster interference involving scattering from clusters separated by an average distance r_m. In this picture, therefore, the FSDP is simply the first (and largest) of a series of strongly damped peaks in $D_m(Q)$, the Fourier transform of a single broad peak at r_m in real space. [At high values of Q, the oscillations of $f_m(Q)$ dominate.] Moss and Price (1985) have proposed, in accord with this picture, that the FSDP in oxide and chalcogenide glasses arises from correlations involving the packing of coordination polyhedra.

The MRO in chalcogenide glasses is richer than that found in the corresponding oxide glasses due to the reduced local dimensionality associated with the trigonal pyramidal coordination of pnictogens and with the edge-sharing connections of tetrahedral coordination polyhedra believed to be present in Si and Ge chalcogenides. In the Si chalcogenides, SiS_2 and $SiSe_2$, edge sharing is predominant, leading to a locally *one-dimensional* structure as in the corresponding crystalline structures. A comprehensive theoretical modeling study of these systems, using computer-generated "ball-and-stick" models, has allowed those parameters to be identified which are essential for a model to match experimental (diffraction) results (Gladden and Elliott, 1989a,b). The results of this study are shown in Fig. 18, where the fit of the model to the experimental real-space correlation function derived from neutron diffraction is shown, together with the structural configurations believed to contribute particularly to certain regions of the RDF. It can be seen that the structure of v-SiS_2 or $SiSe_2$ appears to be well described in terms of amodel containing clusters of cross-linked parallel chain (edge-sharing) configurations.

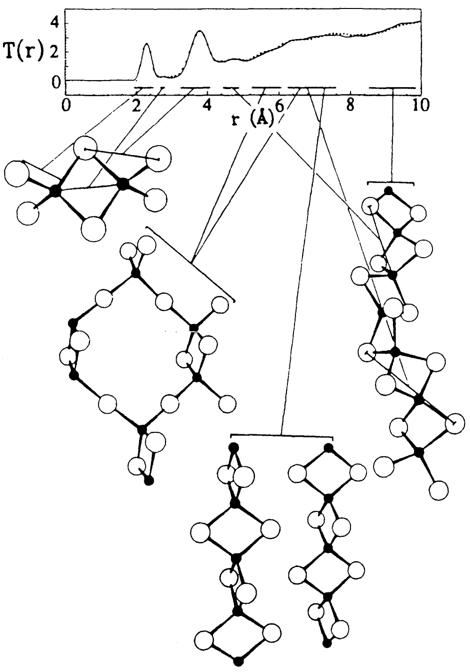

FIG. 18. Comparison of the experimental correlation function $T(r)[=J(r)/r]$ for v-$SiSe_2$ with that of a computer-generated model containing the structural configurations also shown (Gladden and Elliott, 1987).

Edge-sharing connections of tetrahedra are also believed to be present in Ge chalcogenide glasses, but to a lesser extent than in Si chalcogenide materials. The structure of one crystalline polymorph of GeS_2 or $GeSe_2$ is essentially layerlike, with the layers formed by edge-sharing tetrahedra cross-linking ribbons of corner-sharing units. It has been claimed (Phillips, 1981) that similar structural arrangements also exist in the corresponding glassy materials, termed "rafts," bordered by homopolar Se–Se bonds (with the conjugate ethanelike groupings involving Ge–Ge bonds occupying unspecified positions elsewhere in the structure).

Glassy Ge chalcogenides exhibit particular features in their Raman spectra which have been ascribed to the existence of MRO in the structure, namely, a sharp "companion" line accompanying the main A_1 stretching mode of the GeX_4 tetrahedra (X=S, Se); for the case of v-$GeSe_2$, for example, the A_1 line occurs at ~200 cm^{-1} and the companion line is at ~220 cm^{-1}. The companion line is distin-

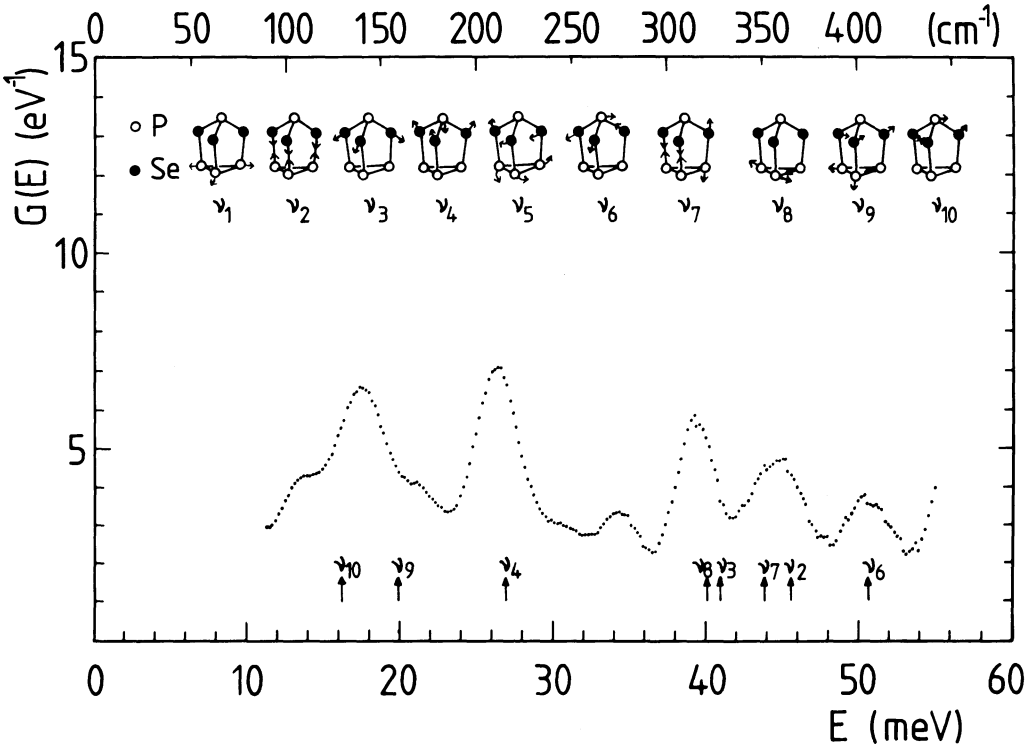

FIG. 19. Vibrational density of states for v-P_2Se obtained from inelastic neutron scattering (Verrall and Elliott, 1988). The assignment of the peaks in terms of various vibrational modes of P_4Se_3 molecules is indicated.

guished by its very strong compositional dependence as the composition departs from stoichiometry, indicating perhaps that the MRO configurations responsible for the companion line are particularly susceptible to breakup upon change of composition. In the Phillips raft model, the companion line was ascribed to modes involving the Se–Se bonds bordering the rafts. However, it is more likely that the companion line in Ge and Si chalcogenides arises from vibrations of the four-membered rings associated with edge-sharing connections of tetrahedra (Sugai, 1987). However, matrix-element enhancement of Raman lines, resulting from symmetric vibrational modes such as that associated with four-membered (edge-sharing) rings, can occur; unfortunately, the energy resolution of current inelastic neutron spectrometers is insufficient to be able to resolve the companion peak in v-$GeSe_2$ (Gladden *et al.*, 1988) in order to ascertain whether the anomalous composition dependence of the companion Raman line is in fact a result of matrix-element enhancement.

The final example of MRO in a chalcogenide glass to be discussed here concerns the phosphorus chalcogenide glass, P_2Se. This material is of interest since it appears to be a rare example of an inorganic molecular glass, i.e., a *zero-dimensional* solid. Inelastic neutron scattering results for v-P_2Se (Fig. 19) reveal a vibrational density of states (VDOS) which is highly structured and where the individual peaks are particularly narrow insofar as a glass is concerned (Verrall and Elliott, 1988). In fact, all but one of the peaks in the VDOS can be assigned to cagelike molecular vibrations of P_4Se_3 molecules (Fig. 19), in which the vibrational modes of the molecules are essentially decoupled from the overall network; the remaining vibrational mode at $\sim 250\ \mathrm{cm}^{-1}$ is ascribed to the vibrations of the excess phosphorus. Raman spectra support this picture of molecular modes (Elliott and Verrall, 1989), and allow some overlapping peaks in the neutron VDOS to be resolved. Conventional neutron diffraction experiments on v-P_2Se yield a structure factor which can be fitted perfectly at large values of

Q in terms of a P_4Se_3 molecule (Verrall and Elliott, 1989); features at smaller values of Q reflect intermolecular correlations [cf. Eq. (27)], and contain information on the IRMRO, viz., the packing of the P_4Se_3 molecular clusters.

4.4 Amorphous Metals

Amorphous metals are invariably alloys, and these glass-forming systems can be divided broadly into two categories: all-metal compositions include alloys of "simple" metals (e.g., $Ca_{67}Mg_{33}$) or of early and late transition metals(e.g., $Nb_{60}Ni_{40}$) or rare earths; in addition, alloys of transition metals with "metalloids" (e.g., C, B, Si, P) can also be glass formers (e.g., $Ni_{80}B_{20}$). Thus, since amorphous metallic systems are always at least binary alloys, atom-specific structural probes or diffraction methods yielding partial PDF data are necessary in order to obtain the maximum amount of interpretable structural information. In general, the structural information of interest in the case of amorphous metals concerns mainly the SRO around each constituent, although obviously this is not well defined as in the covalent materials discussed previously; in the case of the least ordered metal-metal alloy systems, only a statistical definition of SRO is possible, as given by Eqs. (3) and (4). A review of structural aspects of metallic glasses has been given by Sadoc and Wagner (1983), Gaskell (1983, 1987), and in the book edited by Luborsky (1983).

The transition metal-metalloid (T–M) glassy alloy systems have been investigated perhaps the most extensively due to their ease of fabrication and their technological applications, e.g., as transformer cores (Fe–B) and as record/playback heads in magnetic tape recording systems (Co-based alloys). An example of such a study is shown in Fig. 20, where the results of an isotopic-substitution neutron diffraction investigation of $Ni_{81}B_{19}$ are shown (Lamparter *et al.*, 1982). Use of the "null element" isotopic mixture $^{\phi}$Ni allows the B–B partial structure factor to be obtained directly [see Figs. 20(a) and 20(b)]. It can be seen from Fig. 20(c) that the nearest-neighbor coordination shell around the metal consists of metalloid atoms, and vice versa. In particular, it should be noted that there are *no* metalloid-metalloid near neighbors, a finding confirmed for another T–M glassy system, namely, $Co_{80}P_{20}$, using polarized neutron magnetic scattering (Sadoc and Dixmier, 1976).

The experimental observation of metalloid-metalloid avoidance in T–M glassy alloys rules out one possible model for their structure. In the case of the dense random packed (DRP) model, first proposed by Bernal (1959) for liquid structures and later extended to the case of amorphous metals, it was suggested by Polk (1970) that the structure of binary alloys such as T–M glasses could be represented by a DRP model consisting of two sizes of atom, the smaller (metalloid) atoms occupying the interstices between the larger (metal) atoms in a DRP arrangement. However, this model would predict that the metalloid atoms would be distributed in a statistical fashion, with at least some close metalloid-metalloid contacts, in contradiction with experimental results.

An alternative, stereochemical structural model for T–M glasses has been proposed by Gaskell (1979) in terms of local trigonal prismatic coordination. In *crystalline* T–M systems, e.g., cementite (Fe_3C), the dominant structural motif consists of trigonal prismatic coordination by metal atoms of the metalloid atoms, with the possibility of capping of three faces of the trigonal prism by further metal atoms (see Fig. 21). The occurrence of MT_6 trigonal prismatic units as the dominant structural motifs in glassy T–M alloys naturally explains, therefore, the absence of M–M nearest-neighbor bonds. This model gives reasonable agreement with experimental metal-metal partial PDFs, but rather poor agreement with metal-metalloid partial PDFs (Gaskell, 1981).

However, although in the case of covalent glasses, the directionality of the covalent bonding dictates the local SRO, i.e., the shape of the coordination polyhedra, in the absence of strong covalent interactions, as in the T–M alloy glasses, it is not clear precisely what determines such orientational ordering. Whereas it is generally accepted that, in *covalent* glasses, structural order on different length scales is *hierarchical*, i.e., the type of SRO present determines in turn the structural organization at the next level, viz., MRO, the converse of this has been suggested by Dubois *et al.* (1985) and Gaskell (1987) for the case of T–M glasses, namely, that the (stereochemical) SRO may be a *consequence* of the exist-

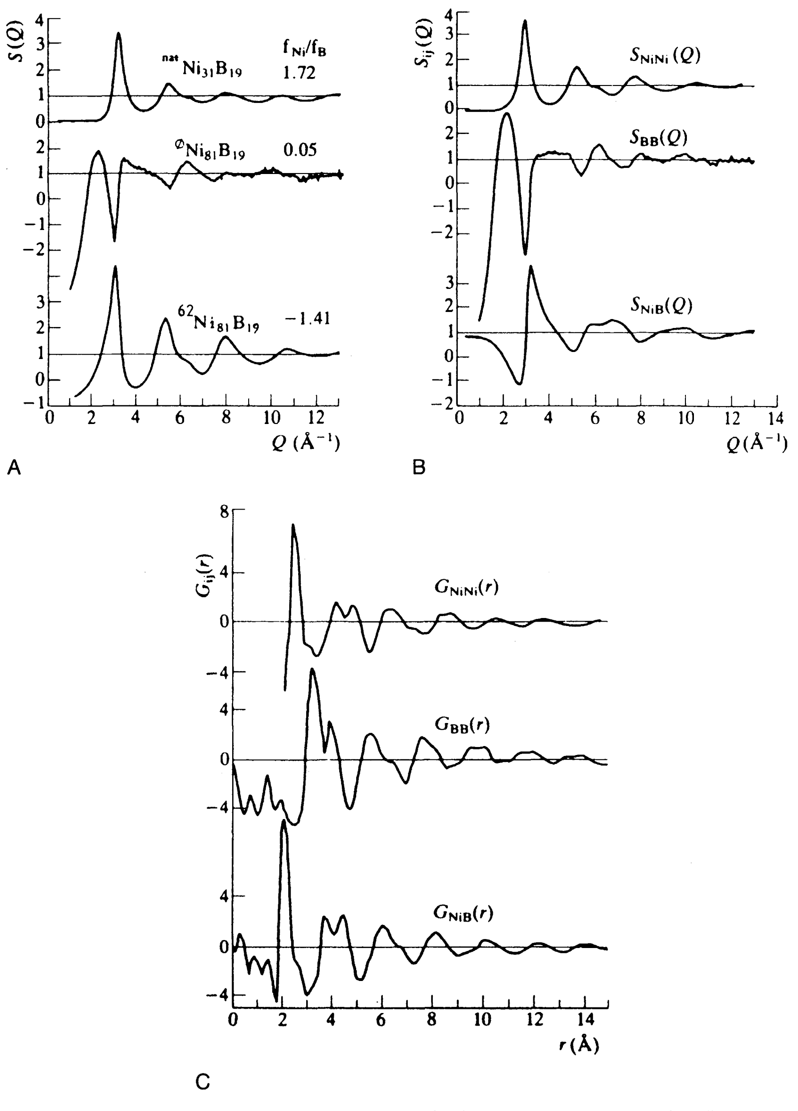

FIG. 20. Structural studies of glassy $Ni_{81}B_{19}$ using isotopic-substitution neutron diffraction (Lamparter *et al.*, 1982). (a) Structure factors for different isotopically substituted alloys, including the "null element" isotopic mixture $^{\phi}Ni$. (b) Partial structure factors obtained from the data shown in (a). (c) Partial reduced RDFs obtained by Fourier transformation of the $S_{ij}(Q)$ shown in (b).

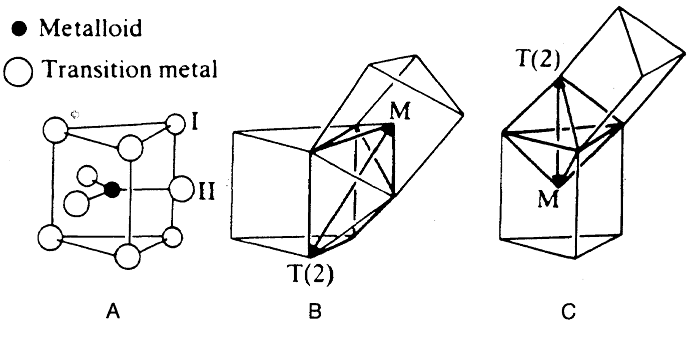

FIG. 21. (a) Trigonal prismatic packing of transition metals around a metalloid atom observed in many crystalline alloys. (b) Edge-sharing arrangement in the cementite (Fe_3) structure. (c) Edge-sharing arrangment in the Fe_3P structure with different metalloid–second nearest neighbor transition metal configuration.

ence of a well-defined MRO. An example of such a circumstance in the crystalline state is the occurrence of tetrahedra and octahedra arising from ordered close-packed arrays of atoms. In the case of crystalline T–M materials, the trigonal prismatic coordination of metalloid atoms can be regarded as arising from the introduction of "chemical twinning" planes (containing the M atoms) into a close-packed crystalline arrangement of the metal atoms (Hyde *et al.*, 1979); the composition of the T–M alloys is varied by changing the periodicity of the chemical twinning planes. Dubois *et al.* (1985) have suggested that a similar situation prevails in the case of glassy T–M alloys as well; here, it is proposed that the LRMRO is associated with domains of size ~10–20 Å, within each of which exist more or less ordered arrays of chemical twinning planes and along which are therefore located the trigonal prismatic units. Disorder is introduced into the model by the assumption that the orientation of twinning planes is different in each domain.

By contrast, the structural studies which have been carried out on glassy metal-metal alloys have revealed much less tendency for chemical ordering compared with that observed in T–M materials. This can be seen by reference to Fig. 22, in which are shown the first peaks in the total RDF for two amorphous alloys of late transition metals (T) with rare earths (RE), namely, *a*-$Fe_{64}Gd_{36}$ and *a*-$Co_{82}Gd_{18}$ (Cargill, 1975). The first peaks are clearly observed to be split into three components, corresponding, respectively, to T–T, T–RE, and RE–RE nearest neighbors in ascending order of distance. There is thus no evidence for strong chemical ordering as in glassy T–M alloys, and the structure of amorphous T–RE alloys can be described approximately by a random mixture of the constituent metal atoms in a DRP structure.

5. FUTURE PROSPECTS

In this final section, likely future trends in techniques for the determination of the structure of amorphous solids are outlined, laying emphasis on those methods discussed previously in this review. It is apparent that these new developments, both experimental and theoretical, will provide new and increasingly detailed structural information for amorphous materials in the near future.

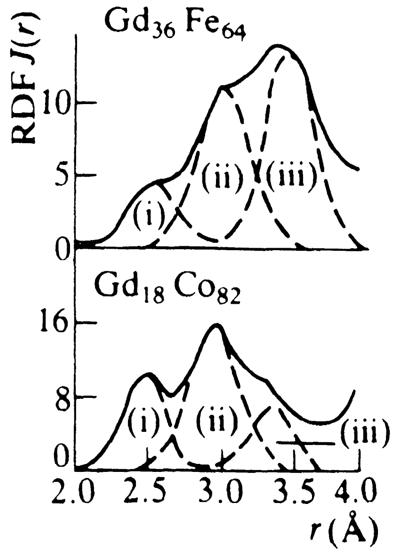

FIG. 22. First peaks of the total RDF for two late transition metal (T)–rare earth (RE) amorphous alloys (Cargill, 1975). In each case (i), (ii), and (iii) refer to T–T, T–RE, and RE–RE correlations, respectively.

5.1 X-Ray Absorption Spectroscopy

The use of the higher-brightness synchrotron sources shortly to be commissioned will enable more dilute systems to be investigated using fluorescence detection. Furthermore, such high-brightness x-ray sources will allow time-dependent XAS (and diffraction) measurements to be made using energy-dispersive detectors, so that the kinetics of structural changes may be followed. In addition, the newly developed technique of glancing-angle XAS, performed in total-reflection geometry, permits a depth-selective structural investigation of a surface region to be carried out using the evanescent x-ray wave as a probe. Preliminary experiments of this type have already been used to study leaching and corrosion processes at glass surfaces (Greaves, 1986), but higher x-ray photon fluxes should allow such experiments to be carried out in real time.

5.2 Nuclear Magnetic Resonance

It is perhaps in the field of NMR that the most significant developments are happening which promise to provide much new structural information for amorphous solids in the near future, despite the fact that the number of NMR-accessible nuclei is presently rather limited (but should be extended in the future). Three advances may be highlighted.

The first significant development to be discussed offers the prospect of a significant increase in resolution beyond that presently achievable using conventional MAS techniques. In the case of quadrupolar nuclei with values of nuclear spin $I \geqslant 1$ (which make up the majority of naturally occurring isotopes), MAS only removes the *first*-order term in the quadrupolar interaction (as long as the strength of the interaction is appreciably less than the spinning speed); the remaining second-order contribution to the quadrupolar interaction can still be strong enough so that the NMR lines are insufficiently narrowed by MAS operation to allow site-dependent chemical-shift differences to be discerned, although operation at high magnetic fields helps in this regard. The new technique of double sample rotation (Samoson *et al.*, 1988), in which a sample is rotated about *two* axes simultaneously, one being at the magic angle, allows even second-order quadrupolar interactions to be removed, leading to a significant improvement in resolution. This approach has been demonstrated for the case of ^{17}O in minerals (Chmelka *et al.*, 1989) but has not so far been applied in the case of glasses.

A second significant advance in NMR technique has been the development of the "multiple-quantum" interference method (Baum *et al.*, 1977). Normally, the Zeeman selection rule, $\Delta M = \pm 1$, applies to single-quantum transitions between nuclear spin levels. However, by means of the application of a particular rf pulse sequence, a system of N spins (each with $I = \frac{1}{2}$) can be forced to undergo essentially a cooperative transition in which the effective change in magnetic quantum number can be $\Delta M = N$ instead. The dipolar coupling between spins is employed to bring about the multiple-quantum transitions; the magnitude of this interaction depends on the internuclear distance as r^{-6}. Thus, for the case of completely isolated clusters each containing N spins, the maximum quantum transition possible is $\Delta M = N$, whereas for close clusters, or connected networks of spins, ΔM increases progressively as the "preparation" time in the rf pulse sequence is increased to allow time for more distantly separated spins to interact. This technique has been applied already to the case of protons in *a*-Si:H (Baum *et al.*, 1986; Gleason *et al.*, 1987), where it was found that device-quality material (with $\leqslant 10$ at. % H) appears to contain clusters of ~ 6 H atoms. In principle, this method could also be used to investigate clustering of other $I = \frac{1}{2}$ nuclei in glasses.

The final development to be discussed here offers the prospect of obtaining *direct* (bond-length) structural information from NMR measurements; this technique is "spin-echo double resonance" (SEDOR). In this, the dipolar interaction is again used to provide spatial information. A double resonance experiment is performed, in which the response at the resonant frequency of one type of nucleus is monitored while another type of nucleus is irradiated with rf radiation at its resonant frequency to cause spin flips. If the two nuclei are in dipolar spin contact, which essentially means in close physical proximity since the mediating dipolar interaction varies as r^{-6}, an influence of irradiation will be detected at the nucleus being monitored. From the behavior of the spin echo, an estimate can be obtained for the probable distribution of internuclear distances. This technique has been used so far to study the structural environment of dopants (e.g., B and P) in *a*-Si:H (Boyce *et al.*, 1987; Boyce and Ready, 1988), but the method could in principle also be used for other types of atoms in different amorphous systems.

GLOSSARY

Amorphous: A material lacking the long-range translational periodicity of a crystal (synonymous with noncrystalline).

Anomalous Scattering: Diffraction with scattering radiation having an energy close to that of an absorption edge for the radiation involved of an atom in a compound (alloy) material, with the result that the scattering factor contains complex and energy-dependent dispersion corrections.

Atom-Specific Structural Probe: An experimental technique for structure determination which is sensitive to the local structure around a selected element in a compound (alloy) material.

BO: Bridging oxygen.

Bridging: An anion which satisfies its normal twofold coordination (e.g., oxygen or a chalcogen) and connects two cation-centered coordination polyhedra.

Chalcogen: Element from group VI of the periodic table (e.g., S, Se, Te).

Chalcogenide: Compound material containing a chalcogen.

Chemical Order: The favoring of heteropolar bonds in a covalent compound material.

Chemically Ordered Network (CON): A continuous random network in which heteropolar bonds are favored.

Chemical-Shift Interaction: An interaction in nuclear magnetic resonance in which the resonant frequency of a nucleus is modified by the partial shielding of the external magnetic field due to the electronic charge distribution associated with the bonds to the atom containing the resonant nucleus.

Continuous Random Network (CRN): An idealized structure for a covalently bonded amorphous material, in which each atom adopts a coordination in accordance with its normal valence and there are no dangling bonds.

Coordination Number: The number of nearest-neighbor atoms surrounding an origin atom.

Coordination Polyhedron: The geometrical shape characterizing the fundamental structural unit of a covalently bonded solid, comprising a cation surrounded by nearest-neighbor anions.

Dangling Bond: The nonbonding orbital corresponding to a broken bond of an atom in a covalent material.

Dense Random Packing: An idealized model for the structure of a material held together by central forces (e.g., a metal), represented by a disordered aggregate of spheres.

Dihedral Angle: The torsion angle about a common bond required to bring into coincidence the projections onto the plane perpendicular to this bond of the other bonds comprising two connected coordination polyhedra.

Diffraction: The scattering of waves (e.g., x-rays, neutrons, or electrons) by pointlike objects (e.g., atoms).

Dipolar Interaction: An interaction between the dipole moments of two nuclei.

EFG: Electric field gradient.

EXAFS: Extended x-ray absorption fine structure.

Form Factor: The scattering intensity from a pointlike scattering entity.

FSDP: First sharp diffraction peak.

Glass: An amorphous solid exhibiting a glass transition.

Glass Transition: The point marking the transition from a liquid to a glass at which there is a discontinuity in derivative thermodynamic quantities (e.g., specific heat).

IR: Infrared.

IRMRO: Intermediate-range medium-range order.

Isomorphous Substitution: Partial replacement of an element in a compound (alloy) material by another element having a similar size and chemical characteristics, so that the overall structure is unperturbed.

Isotopic Substitution: Variation of the isotopic components of the elements in a material while keeping the overall atomic composition unchanged.

LRMRO: Long-range medium-range order.

Magic-Angle Spinning (MAS): A technique in nuclear magnetic resonance in which, by spinning a sample at a "magic" angle with respect to an external applied magnetic field, the interactions (dipolar, chemical-shift anisotropy, etc.) which otherwise would broaden the resonance signal are removed.

Medium-Range Order (MRO): The next higher level of structural organization at distances larger than that corresponding to short-range order.

Metalloid: Nonmetallic elements (e.g., C, P, Si, B, etc.) which form glassy alloys with metallic elements.

Molecular Dynamics (MD): A method for the simulation of the structure and atomic dynamics of liquids by following the time evolution of the classical motion of a collection of atoms interacting by means of a particular interaction potential.

NBO: Nonbridging oxygen.

Network Depolymerization: Change in the connectivity of the structure of a network former caused by the introduction of nonbridging (dangling-bond) sites accompanying the incorporation of network modifier cations.

Network Dimensionality: The number of dimensions traced out in a percolation process along covalent bonds in the structure of a covalently bonded material.

Network Former: A covalent material which, by itself, will form a glass.

Network Modifier: A cation (e.g., alkali) which causes coordination changes in the network-former structure, e.g., of the cation (charge in coordination polyhedron) or of the anion (network depolymerization with formation of nonbridging sites).

Neutron Scattering Length: The amplitude of scattering of neutrons from an atom.

NEXAFS: Near-edge x-ray absorption fine structure.

NMR: Nuclear magnetic resonance.

Nonbridging: An anion (e.g., oxygen or a chalcogen) which has a dangling bond or, equivalently, which is only singly coordinated.

Null Element: An isotopic mixture of an element for which the net neutron scattering length is zero.

Partial Correlation Function: The contribution to the total correlation function (e.g., radial distribution function) arising from pairs of specific types of atoms in a compound (alloy) material.

PDF: Pair distribution function.

Pnictogen: Element from group V of the periodic table (e.g., P, As, Sb).

Quadrupolar Interaction: An interaction between the quadrupole moment of a nucleus and the electric field gradient of the electronic charge distribution associated with the bonds to the atom.

Radial Distribution Function (RDF): The probability of finding an atom at a particular distance from an origin atom.

Random Covalent Network (RCN): A continuous random network structure of a compound material in which the distribution of bond types (heteropolar and homopolar) is determined statistically, i.e., by the composition.

RE: Rare earth.

SEDOR: Spin-echo double resonance.

Short-Range Order (SRO): Structural order of nearest-neighbor atoms around a given atom.

SRMRO: Short-range medium-range order.

Structure Factor: The interference function resulting from the diffraction (scattering) of waves from an array of scatterers (atoms) normalized to the scattering intensity (form factor) for each scatterer.

Superstructural Unit: A collection of coordination polyhedra comprising an effective structural unit in the structure of a covalently bonded amorphous solid.

TM: Transition metal.

TO: Transverse optic (phonon).

VDOS: Vibrational density of states.

VH: Vertical-horizontal (polarized Raman scattering).

Wrong Bond: Like-atom (homopolar) bond in a compound material.

X-Ray Scattering Factor: The amplitude of scattering of x rays from an atom.

XANES: X-ray absorption near-edge structure.

XAS: X-ray absorption spectroscopy.

List of Works Cited

Andrew, E. R. (1981), *Philos. Trans. R. Soc. London Ser. A* **299**, 505.

Angell, C. A., Cheeseman, P. A., Clarke, J. H. R., Woodcock, L. V. (1977), in: P. H. Gaskell (Ed.), *The Structure of Non-Crystalline Materials*, London: Taylor & Francis, p. 191.

Baum, J., Munowitz, M., Garroway, A. N., Pines, A. (1977), *Chem. Phys. Lett.* **83**, 2015.

Baum, J., Gleason, K. K., Pines, A., Garroway, A. N., Reimer, J. A. (1986), *Phys. Rev. Lett.* **56**, 1377–1380.

Beeman, D., Tsu, R., Thorpe, M. F. (1985), *Phys. Rev. B* **32**, 874–878.

Bell, R. J., Dean, P. (1972), *Philos. Mag.* **25**, 1381.

Bellisent, R., Chenevas-Paule, A., Chieux, P., Menelle, A. (1985a), *J. Non-Cryst. Solids* **77–78**, 213.

Bellisent, R., Menelle, A., Chenevas-Paule, A., Chieux, P. (1985b), *J. Phys. Paris* **46**, C8-93.

Bennett, C. H. (1972), *J. Appl. Phys.* **43**, 2727–2734.

Bernal, J. D. (1959), *Nature* **183**, 141.

Bernal, J. D. (1964), *Proc. R. Soc. London Ser. A* **280**, 299.

Bhatia, A. B., Thornton, D. E. (1970), *Phys. Rev. B* **2**, 3004–3012.

Boolchand, P. (1985), in: D. Adler, B. B. Schwartz, M. C. Steele (Eds.), *Physical Properties of Amorphous Materials*, New York: Plenum, p. 221.

Boyce, J. B., Ready, S. E. (1988), *Phys. Rev. B* **38**, 11008–11018.

Boyce, J. B., Stutzmann, M. (1985), *Phys. Rev. Lett.* **54**, 562–565.

Boyce, J. B., Ready, S. E., Tsai, C. C. (1987), *J. Non-Cryst. Solids* **97–98**, 345.

Bray, P. J. (1978), in: L. D. Pye, V. D. Frechette, N. J. Kreidl (Eds.), *Borate Glasses: Structure, Prop-*

erties, Applications, New York: Plenum.

Busse, L. E. (1984), *Phys. Rev. B* **29**, 3639.

Cargill, G. S. (1975), in: F. Seitz, D. Turnbull, H. Ehrenreich (Eds.), *Solid State Physics*, Vol. 30, New York: Academic, p. 227.

Cargill, G. S., Spaepen, F. (1981), *J. Non-Cryst. Solids* **43**, 91.

Carlos, W. E., Taylor, P. C. (1982a), *Phys. Rev. B* **25**, 1435–1438.

Carlos, W. E., Taylor, P. C. (1982b), *Phys. Rev. B* **26**, 3605–3616.

Cervinka, L. (1988), *J. Non-Cryst. Solids* **106**, 291.

Chmelka, B. F., Mueller, K. T., Pines, A., Stebbins, J., Wu, Y., Zwanziger, J. W. (1989), *Nature* **339**, 42.

Connell, G. A. N., Temkin, R. J. (1974), *Phys. Rev. B* **9**, 5323–5326.

Dixmier, J., Derouet, P., Essamet, M., Laridjani, M. (1985), *Philos. Mag. B* **52**, 943.

Dubois, J. M., Gaskell, P. H., LeCaer, G. (1985), *Proc. R. Soc. London Ser. A* **402**, 323.

Dupree, R., Pettifor, R. F. (1984), *Nature* **308**, 523.

Dupree, R., Holland, D., McMillan, P. W., Pettifer, R. F. (1984), *J. Non-Cryst. Solids* **68**, 399.

Dupree, R., Holland, D., Mortuza, M. G. (1987), *Nature* **328**, 416.

Durham, P. J., Pendry, J. B., Hodges, C. H. (1981), *Solid State Commun.* **38**, 159.

Elliott, S. R. (1984), *Physics of Amorphous Materials*, 1st ed. New York: Longmans, Green; (1990) 2nd ed.

Elliott, S. R. (1987), *J. Non-Cryst. Solids* **97–98**, 159.

Elliott, S. R. (1989), *Adv. Phys.* **38**, 1.

Feuston, B. P., Garofalini, S. H. (1988), *J. Chem. Phys.* **89**, 5818.

Filipponi, A., Della Sala, D., Evangelisti, F., Balerna, A., Mobilio, S. (1986a), *J. Phys. Paris* **47**, C8-375.

Finney, J. L. (1970), *Proc. R. Soc. London Ser. A* **319**, 479.

Galeener, F. L. (1985), *Philos. Mag.* **B51**, L1.

Gaskell, P. H. (1979), *J. Non-Cryst. Solids* **32**, 207.

Gaskell, P. H. (1981), *Nature* **289**, 474.

Gaskell, P. H. (1983), in: H. Beck, H. J. Güntherodt (Eds.), *Glassy Metals II*, New York: Springer (Topics in Applied Physics, Vol. 53), p. 5.

Gaskell, P. H. (1987), *Key Eng. Mater.* **13–15**, 71.

Gladden, L. F. (1989), in: O. V. Mazurin (Ed.), *Proceedings of the XV International Congress on Glass*, Vol. 1a. Leningrad: Nauka, p. 198.

Gladden, L. F., Elliott, S. R. (1987), *Phys. Rev. Lett.* **59**, 908–911.

Gladden, L. F., Elliott, S. R. (1989a), *J. Non-Cryst. Solids* **109**, 211.

Gladden, L. F., Elliott, S. R. (1989b), *J. Non-Cryst. Solids* **109**, 223.

Gladden, L. F., Carpenter, T. A., Elliott, S. R. (1986), *Philos. Mag.* **B53**, L81.

Gladden, L. F., Elliott, S. R., Sinclair, R. N., Wright, A. C. (1988), *J. Non-Cryst. Solids* **106**, 120.

Gleason, K. K., Petrich, M. A., Reimer, J. A. (1987), *Phys. Rev. B* **36**, 3259.

Greaves, G. N. (1985), *J. Non-Cryst. Solids* **71**, 203.

Greaves, G. N. (1986), *J. Phys. Paris* **47**, C8-819.

Greegor, R. B., Lytle, F. W., Sandstrom, D. R., Wong, J., Schultz, P. (1983), *J. Non-Cryst. Solids* **55**, 27.

Hyde, B. G., Anderson, S., Bakker, M., Plug, C. M., O'Keefe, M. (1979), *Prog. Solid State Chem.* **12**, 273.

Jeffrey, F. R., Dubois-Murphy, P., Gerstein, B. C. (1981), *Phys. Rev. B* **23**, 2099–2101.

Knights, J. C. (1984), in: J. D. Joannopoulos, G. Lucovsky (Eds.), *The Physics of Hydrogenated Amorphous Silicon I*, New York: Springer (Topics in Applied Physics), p. 5.

Knights, J. C., Hayes, T. M., Mikkelsen, J. C. (1977), *Phys. Rev. Lett.* **39**, 712.

Lamparter, P., Sperl, W., Nold, E., Rainer-Harbach, G., Steeb, S. (1982), in: T. Masumoto, J. Suzuki (Eds.), *Proceedings of the 4th International Conference on Rapidly Quenched Metals*, Japanese Institute of Metals, p. 343.

Luborsky, F. E. (Ed.) (1983), *Amorphous Metallic Alloys*, Washington, DC: Butterworth.

Lucovsky, G. (1985), *J. Non-Cryst. Solids* **76**, 173.

Lucovsky, G., Pollard, W. B. (1984), in: J. D. Joannopoulos, G. Lucovsky (Eds.), *The Physics of Hydrogenated Amorphous Silicon II*, New York: Springer-Verlag (Topics in Applied Physics, Vol. 56), p. 301.

Lucovsky, G., Galeener, F. L., Geils, R. H., Keezer, R. C. (1977), in: P. H. Gaskell (Ed.), *Structure of Non-Crystalline Materials*, London: Taylor & Francis, p. 127.

Lucovsky, G., Nemanich, R. J., Knights, J. C. (1979), *Phys. Rev. B* **19**, 2064–2073.

Maley, N., Lannin, J. S., Scott, B. A. (1983), *J. Non-Cryst. Solids* **59–60**, 201.

Menelle, A., Flank, A. M., Lagarde, P., Bellisent, R. (1986), *J. Phys. Paris* **47**, C8-379.

Mobilio, S., Filipponi, A. (1987), *J. Non-Cryst. Solids* **97–98**, 365.

Moss, S. C., Price, D. L. (1985), in: D. Adler, H. Fritzsche, S. R. Ovshinsky (Eds.), *Physics of Disordered Materials*, New York: Plenum, p. 77.

Mozzi, R. L., Warren, B. A. (1969), *Acta Crystallogr.* **2**, 164.

Neilson, G. W., Enderby, J. E. (1983), *Proc. R. Soc. London Ser. A* **390**, 353.

Phillips, J. C. (1981), *J. Non-Cryst. Solids* **43**, 37.

Polk, D. E. (1970), *Scr. Metall.* **4**, 117.

Polk, D. E. (1971), *J. Non-Cryst. Solids* **5**, 365.

Reimer, J. A. (1981), *J. Phys. Paris* **42**, C4-715.

Reimer, J. A., Duncan, T. M. (1983), *Phys. Rev. B* **27**, 4895–4901.

Sadoc, J. F., Dixmier, J. (1976), *Mater. Sci. Eng.* **23**, 187.

Sadoc, J. F., Wagner, C. N. J. (1983), in: H. Beck, H. J. Guntherodt (Eds.), *Glassy Metals II*, New York: Springer (Topics in Applied Physics, Vol. 53), p. 51.

Sakka, S., Kamiya, K. (1982), *J. Non-Cryst. Solids* **49**, 103.

Samoson, A., Lippmaa, E., Pines, A. (1988), *Mol. Phys.* **65**, 1013.

Smith, J. V., Blackwell, C. S. (1983), *Nature* **303**, 223.

Stebbins, J. F. (1987), *Nature* **330**, 465.

Steinhardt, P. J., Nelson, D. R., Ronchetti, M. (1983), *Phys. Rev. B* **28**, 784–805.

Sugai, S. (1987), *Phys. Rev. B* **35**, 1345.

Temkin, R. J., Paul, W., Connell, G. A. N. (1973), *Adv. Phys.* **22**, 581.

Tsu, R. (1987), in: G. A. Thomas, S. R. Ovshinsky, M. Kastner (Eds.), *Disordered Semiconductors*, New York: Plenum, p. 479.

Tsu, R., Hernandez, J. G., Pollak, F. H. (1984), *J. Non-Cryst. Solids* **66**, 109.

Uemura, O., Sagara, Y., Muno, D., Satow, T. (1978), *J. Non-Cryst. Solids* **30**, 155.

Verrall, D. J., Elliott, S. R. (1989), *J. Non-Cryst. Solids* **14**, 34.

Windisch, C. F., Risen, W. M. (1982), *J. Non-Cryst. Solids* **48**, 307.

Wooten, F., Weaire, D. L. (1984), *J. Non-Cryst. Solids* **64**, 325.

Wright, A. C., Sinclair, R. N. (1985), *J. Non-Cryst. Solids* **76**, 351.

Wright, A. C., Sinclair, R. N., Leadbetter, A. J. (1985), *J. Non-Cryst. Solids* **71**, 295.

Zachariasen, W. H. (1932), *J. Am. Chem. Soc.* **54**, 3841.

Further Reading

Cusack, N. E. (1988), *Physics of Structurally Disordered Matter*, London: Hilger.

Elliott, S. R. (1984, 1990), *Physics of Amorphous Materials*, New York: Longmans, Green.

Waseda, Y. (1980), *Structure of Non-Crystalline Materials: Liquids and Amorphous Solids*, New York: McGraw–Hill.

Wong, J., Angell, C. A. (1976), *Glass: Structure by Spectroscopy*, New York: Dekker.

Zallen, R. (1983), *The Physics of Amorphous Solids*, New York: Wiley.

AMORPHOUS SILICON

See SILICON, AMORPHOUS

AMPLIFIERS, ELECTRONIC

A. A. ABIDI, *Electrical Engineering Department, University of California, Los Angeles*

	Introduction	593
1.	**Specifying Amplifier Characteristics**	594
2.	**Types of Single Gain Stages**	595
2.1	Common-Emitter Amplifier	595
2.2	Common-Collector Stage	596
2.3	Common-Base Stage	596
2.4	FET Gain Stages	597
3.	**Differential Amplifiers**	597
4.	**Operational Amplifiers**	598
5.	**Feedback Amplifiers**	600
6.	**Power Amplifiers**	602
7.	**Narrow-Band Amplifiers**	603
8.	**Distributed Amplifiers**	604
9.	**Noise in Amplifiers**	604
	Glossary	605
	Further Reading	605

INTRODUCTION

An amplifier magnifies the strength of a signal or wave form. It does so by extracting energy from a power supply (usually dc) and converting it into signal energy. This function makes it the single most important component in electronic engineering, its applications ranging from communications systems to instrumentation. In analog circuits, linear amplifiers are used most often to reproduce a weak input signal at a higher power while faithfully preserving its waveform. Amplifiers are also essential building blocks of digital circuits (*q.v.*) but here their nonlinear, or saturation, characteristics are used to implement logic gates. This article considers only linear amplifiers.

To appreciate the role of amplification, consider the familiar example of radio transmission and reception. In the broadcast studio, a microphone converts a voice or music into a weak electrical signal by responding to pressure variations in the air, which, after modulating the amplitude or frequency of a carrier sine wave selected for its favorable propagation characteristics, is then magnified by many amplifier stages to kilowatts of electromagnetic energy, and finally transmitted into space by an antenna. The intensity of these electromagnetic waves, monitored as power per unit area, will diminish as the inverse square of the distance away from the transmitter, inducing only a few microvolts, or perhaps millivolts, on the antenna of a receiver situated tens of miles away. For easy audibility, the received signal wave form, after extraction from the carrier, must be faithfully amplified in voltage to drive a loudspeaker coil, causing enough air to be displaced by the cone attached to that coil that a listener situated many feet away can easily hear the sounds being produced in the studio of the radio station. This example illustrates the alternating chain of of transducer and amplifier that recurs time and again in practice—microphone, amplifier, transmitter antenna, receiver antenna, amplifier, and loudspeaker. At the receiver, the role of the amplifier is to enhance the power level of a weak signal produced by one transducer (the antenna) in order to drive another transducer (the loudspeaker) at a power level which may be orders of magnitude higher. The amplifier does this by extracting signal-less power from a supply (the battery or ac mains) and converting it into signal power which is a magnified replica of the weak input.

The first electronic amplifiers relied on the modulation of electronic beams flowing in vacuum between electrodes held at a large potential difference, by applying the input signal to an intermediate electrode capable of modulating the beam. This was the vacuum tube (or valve in British terminology) based amplification device (see ELECTRON TUBES). Vacuum tubes were used widely for amplification until they were preempted in 1948 by the invention of what has turned out to be a far cheaper, more reliable and compact am-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

plifying device, the semiconductor transistor. The rapid proliferation of this amplifying device heralded the electronics revolution.

1. SPECIFYING AMPLIFIER CHARACTERISTICS

There are four important characteristics that specify an amplifier: *gain, bandwidth, input impedance,* and *output impedance.* An amplifier may be designed to be driven by an input *voltage source,* and produces at its output a *voltage* which is a magnified replica of the input: this is called a voltage amplifier. To reduce a voltage loss across the series resistance of the source, the amplifier must have a high input impedance, while to drive an arbitrary load impedance with the amplified voltage without attenuation, the amplifier should have a low output impedance. The gain of a voltage amplifier is specified as a dimensionless quantity (volts/volt). Similarly, a current amplifier may be made, having a low input resistance to sample a current source driving its input, and a high output resistance to drive into the load the magnified input current without attenuation. The two remaining combinations of input and output stimulus type are a transimpedance amplifier, with current input and voltage output (whose gain is specified in volts/amp), and a transconductance amplifier, with voltage input and current output (Table 1). Choice of the amplifier type is determined by the application: while the voltage amplifier is most commonly encountered, a transimpedance amplifier must be used when the source is a current, as, e.g., the output current produced by a photodiode. It may be difficult in many instrumentation applications to determine whether a signal source with a certain resistive or capactive impedance should be considered a voltage or current source for the purpose of choosing the type of amplifier to follow it. One criterion is that the input impedance of the amplifier should not seriously attenuate the signal to be measured across the complete frequency range of interest. If, e.g., the source appears capacitive across the frequency range of interest, an amplifier with a high r_{in} will produce a very narrow bandwidth and cause the amplifier to perceive at its input the time integral of the source signal; in this case, an amplifier with low r_{in}, like a transresistance or current amplifier is more appropriate, for it will short out the capacitance and substantially extend the bandwidth at the input.

Table 1. Characteristics of amplifiers.

Type	Input	r_{in}	Output	r_{out}
Voltage	V	High	V	Low
Current	I	Low	I	High
Transresistance	I	Low	V	Low
Transconductance	V	High	I	High

Ideally, a linear amplifier should offer the same gain to an input signal, independent of the signal frequency. There are practical limits to this, however, for the amplifying devices and other components of the amplifier have capacitance in parallel with the signal path, which shunts signals to ground away from the output by an increasing degree as the signal frequency increases, until the shunting effect causes a substantial decrease in gain. The upper cutoff frequency of amplifiers is usually specified as that at which, for a fixed amplitude at the input, the power delivered by the amplified output signal to a resistive load decreases by a factor of 2. There may be a lower cutoff frequency as well, caused by other capacitors that have been inserted in series with the signal path for the purpose of decoupling the dc levels between different stages of the amplifier. Direct coupled amplifiers do not employ these series capacitors, and are capable of amplifying signals from dc to higher frequencies with uniform gain, limited only by the upper cutoff frequency. The difference between the upper and lower cutoff frequencies is the *bandwidth* of the amplifier in which signal frequencies experience a uniform gain.

Linear amplifier circuits are made out of amplifying devices, such as transistors or vacuum tubes, that are inherently nonlinear. Strictly speaking, amplifiers are always only approximately linear as long as the output signal swing is confined within certain limits. They become strongly nonlinear in saturation or cutoff, usually when the amplifier output reaches limits imposed by the positive and negative power supplies and no longer changes in response to the input signal. When this occurs, the amplified wave form is a clipped version of the input. To accommodate the largest input signal before distortion sets in at the output, the amplifier must be *biased* by choosing its dc operating point so that in the absence of the input signal, the output voltage lies in the middle of the positive and negative power supplies. Sometimes, when the bias points of two amplifier stages in cascade are unequal, the output dc of one stage must be blocked from the input of the next by a series capacitor, whose value is large enough to pass the signal frequencies of interest unattenuated, but which also introduces a lower cut-off frequency below which signals will be blocked.

Distortion, the deviation of the output wave form from the input wave form except for an amplitude scaling, is a measure of the nonlinearity in the amplifier. A distorting amplifier subject to a sinusoidal input at some frequency will produce at its output, in addition to amplified energy at this frequency, energy at integer multiple frequencies, or harmonics. The *total harmonic distortion*, the total power created in harmonics at the output relative to the power contained in the fundamental frequency, increases with the amplitude of the input.

2. TYPES OF SINGLE GAIN STAGES

The most common amplifier circuits using one transistor as the gain element are now described. Complex amplifier circuits may be decomposed into units of these stages connected together in some way. Bipolar transistor circuits (see TRANSISTORS, BIPOLAR) are illustrated here, although historically these circuits were built with vacuum tubes, and now may often use field-effect transistors (see TRANSISTORS, FIELD-EFFECT).

A brief introduction to bipolar transistor action is first given. This device is formed from a continuous semiconductor crystal containing two back-to-back junctions, in the form of either *npn* or *pnp* layers. In normal operation when the transistor is biased in the forward active region, the junction between the emitter and base (EB) is forward biased, injecting current carriers from the emitter into the base, which are swept into the collector across the favorable electric field at the reverse-biased collector-base (CB) junction. The magnitude of collector current (I_C) is controlled by the external current injected into the base (I_B) according to the current gain β. In addition, the electrical characteristics of the EB junction dictate that $I_C = I_S \exp(V_{BE}/V_T)$, where I_S, the reverse saturation current, is determined by the physical structure of a particular transistor and the physical constant $V_T = kT/q = 25$ mV at room temperature (300 K).

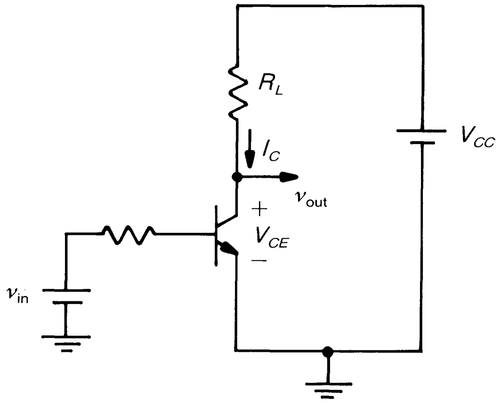

FIG. 1. The common-emitter bipolar transistor amplifier.

2.1 Common-Emitter Amplifier

This single-transistor amplifier (Fig. 1) affords the largest voltage gain. The input voltage is applied at the transistor base-emitter junction, which acts as a control on the current (I_C) that flows from collector to emitter. The circuit is called a common-emitter circuit, because the input and output voltages are referenced to the common emitter node. The voltage developed by the collector current flowing in the load resistor R_L defines the amplifier output. The amplifier is clearly nonlinear, being exponential due to the transistor characteristics in the forward active region, and its output voltage is limited by the positive and negative power supplies (V_{CC}, ground). More precisely, the transistor is cut off when I_C diminishes to zero, and the output current cannot go any lower; or it is saturated when V_{CE} tends to zero, when the output voltage cannot go any lower in response to the input. However, because most often a very small signal is applied at the input, the transistor current varies only over a small portion of the exponential nonlinearity, and so the transfer function around the quiescent operating, or bias, point may be approximated by a linear function, the derivative of the nonlinearity. Amplifier characteristics like the voltage gain (A), input resistance (r_{in}), and output resistance (r_{out}) may thus be expressed for small signals as follows:

$$A=R_L\frac{dI_C}{dV_{BE}}=\frac{I_C}{V_T}R_L, \tag{1}$$

$$r_{\text{in}}=\frac{dV_{BE}}{dI_B}=\frac{V_T}{I_B}, \tag{2}$$

$$r_{\text{out}}=R_L. \tag{3}$$

It is customary to denote bias currents and voltages by capital letters (e.g., I_C,V_{BE}) and small signals superimposed on these biases by small letters (e.g., i_c,v_{be}). The small-signal quantity dI_C/dV_{BE} is called the transconductance (g_m), while dV_{BE}/dI_B is called the diffusion resistance (r_π).

The common-emitter is the most frequently used gain stage, offering the largest gain of all single-transistor amplifiers with resistive loads. The transistor capacitances, whose magnitude is determined by the physical structure of the device, together with the source resistance R_S and load resistance R_L determine the bandwidth. When driven by a voltage source, r_{in} of the amplifier is made large by biasing it at a small base current (I_B). The collector current is then set by the value of transistor current gain β. To obtain a desired voltage gain, set by the I_CR_L product, an appropriate R_L must be chosen, which then sets the output resistance of the amplifier.

2.2 Common-Collector Stage

While the common-collector amplifier stage (Fig. 2) offers a voltage gain of almost 1, its use is principally as a buffer, because its input resistance is much higher and its output resistance much lower than the same transistor connected in the common-emitter configuration. These quantities are given by

$$A=\frac{g_mR_E}{1+g_mR_E}\sim 1, \tag{4}$$

$$r_{\text{in}}=r_\pi+\beta R_E, \tag{5}$$

$$r_{\text{out}}=(g_m+1/R_E)^{-1}. \tag{6}$$

A buffer stage may be connected, e.g., between two common-emitter gain stages in cascade, offering a high resistance to the output of the first stage, transferring the output voltage almost unaltered to the next gain stage with a low source resistance, so that the input resistance of the latter does not load down, and thus attenuate the gain of the first stage.

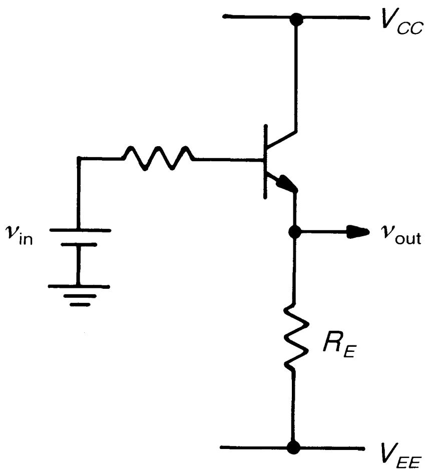

FIG. 2. The common-collector amplifier, or emitter follower.

Because the emitter of the transistor follows the applied input voltage with a gain of almost 1, this configuration is also known as an emitter follower.

2.3 Common-Base Stage

The common-base stage (Fig. 3) is used as a current buffer, in contrast to the common-collector, which acts as a voltage buffer. The common-base stage offers a low input resistance, and therefore is suitable to be driven by a current source. In normal operation, the emitter current of a transistor is very slightly larger than the collector current, so that the effective current transfer, or gain, is slightly less than unity. However, the output resistance at the collector is high, making the stage appear as an almost perfect current source.

One application of the common-base stage is as a buffer following a common-emitter stage, which, by offering a low resistance to the latter, suppresses the multiplication of the capacitance (C_{cb}) by the voltage gain, which is very often the dominant factor in reducing the bandwidth of a common emitter amplifier. This combination is called the cascode.

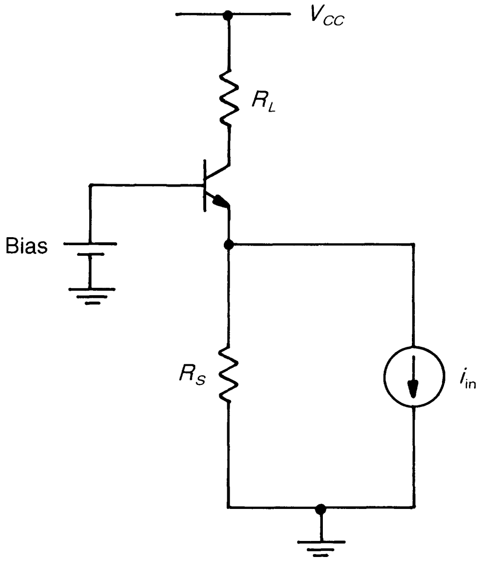

FIG. 3. The common-base amplifier.

2.4 FET Gain Stages

Field-effect transistors (FETs) work on different physical principles than the bipolar, but may be used as amplifiers in identical topologies to the bipolar transistor circuits described above. The field-effect transistor has a capacitive input which takes no dc current for operation, and is therefore used to make amplifiers that have a high input impedance at low frequencies. FET amplifiers have also become common recently because sometimes only these devices are available on complex integrated circuits.

3. DIFFERENTIAL AMPLIFIERS

In many instances, it is desirable to amplify only the *difference* in two signals, or the difference between a signal and a reference, while rejecting the average, or common-mode, component. The wires leading from a sensor like a thermocouple, e.g., that converts some physical quantity to a small voltage and transmits it to a remote amplifier may pick up greater interference and noise from the surroundings than the voltage generated by the thermocouple. A differential amplifier connected at the end of these wires will amplify only the thermocouple voltage, which appears as a potential difference between the wires, while rejecting the interference which would appear as an equal voltage on each wire relative to ground potential.

The classic differential amplifier (long-tailed pair in British terminology) is a pair of transistors biased by a high resistance or current source at their common emitter (Fig. 4). Two input signals (v_{i1},v_{i2}) are applied at the bases and the output is taken as the difference of the two collector voltages (v_{o1},v_{o2}). The amplifier responds in one way to balanced, antiphase inputs, and in another to equal, inphase signals at the inputs. For small balanced antiphase input signals, called a differential excitation, the emitter potential will tend to remain constant (by symmetry), so that $v_{be1} = v_{i1}$ and $v_{be2} = v_{i2}$, providing the full gain of a common-emitter stage at each collector. Equal and opposite signal currents, $g_m v_{i1} = -g_m v_{i2}$ will flow in the emitters. The difference of the collector voltages will double the signal amplitude. For equal, in-phase signals at the two inputs, called a common-mode excitation, the emitter will tend to follow the inputs, inducing a small current in the large resistance from the emitter to supply, which will split equally between the two transistors. The differential and common-mode gains are given by

$$A_{\mathrm{DM}} = g_m R_L, \tag{7}$$

$$A_{\mathrm{CM}} = R_L/2R_E, \tag{8}$$

where the fact that $g_m R_E = I_E R_E / 2V_T \gg 1$ implies $A_{\mathrm{DM}} \gg A_{\mathrm{CM}}$. The common-mode rejection

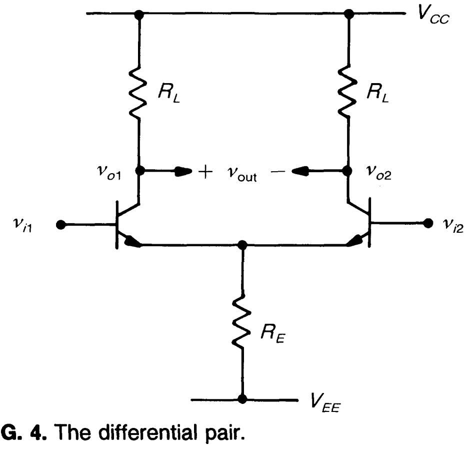

FIG. 4. The differential pair.

ratio is the factor by which the differential gain is greater than the common-mode gain.

In general, using linear superposition, two independent small input signals may be resolved into a balanced, antiphase component and a common-mode component. The response of the amplifier is the linear superposition of each component subjected to the appropriate gain.

4. OPERATIONAL AMPLIFIERS

Differential voltage amplifiers with a very high voltage gain and common-mode rejection ratio (ideally infinite), very high input resistance, and low output resistance are used as building blocks in feedback circuits that perform precise mathematical functions, like algebraic summation, integration, filtering, comparison, etc. These circuits are termed operational circuits, and the amplifiers used to make them operational amplifiers, or op amps. The operations are implemented by the electrical characteristics of passive elements used in feedback around the operational amplifier which serves either to decouple the interaction of two elements constituting the feedback loop, to invert the characteristics of one element, or simply to provide gain. In most cases, one or more passive elements connected between the output and inverting input introduce negative feedback around the op amp, implying that as long as the output of the op amp is not saturated and the gain of the op amp is assumed infinite, the differential voltage at its input will necessarily be zero. Any tendency of the differential input to deviate from zero would be overcome by the action of the negative feedback. Op amp circuits are analyzed and synthesized on the assumption that the differential voltage is zero in normal operation; this is called the virtual ground property. Thus, in the inverting amplifier [Fig. 5(a)] the voltage at the negative op amp input is held at 0 V by the action of feedback, forcing the current flowing from the applied input voltage to equal v_{in}/R_1. The output voltage, determined by this same current flowing through the feedback resistor R_0, is given by $v_{\text{out}}=-(R_0/R_1)v_{\text{in}}$, noting that no current flows into the very high input resistance of the op amp. The same circuit may be converted into a noninverting amplifier [Fig. 5(b)] by retaining the negative feedback connection, but applying the input v_{in} instead at the positive input. Again, negative feedback forces the differential voltage between the two op amp inputs to zero, implying that $v_{\text{out}}=(1+R_0/R_1)v_{\text{in}}$. This topology may be generalized to an algebraic weighted sum circuit [Fig. 5(c)] where a weighted resistive voltage divider sets up the voltage at the positive (noninverting) input of the op amp v_+, as given by

$$v_+\left(\frac{1}{R_0}+\frac{1}{R_{p1}}+\frac{1}{R_{p2}}+\cdots+\frac{1}{R_{pk}}\right)=\frac{v_{p1}}{R_{p1}}+\frac{v_{p2}}{R_{p2}}+\cdots+\frac{v_{pk}}{R_{pk}}. \tag{9}$$

Negative feedback through R_0 from the output to the negative (inverting) input of the amplifier then enforces the voltage $v_-=v_+$. Thus, the output voltage is given by

$$v_{\text{out}}=R_0\left(\frac{v_{p1}}{R_{p1}}+\frac{v_{p2}}{R_{p2}}+\cdots+\frac{v_{pk}}{R_{pk}}\,\frac{1+R_0\Sigma_1^{l+1}(1/R_{mi})}{1+R_0\Sigma_1^{k+1}(1/R_{pi})}-\frac{v_{m1}}{R_{m1}}-\frac{v_{m2}}{R_{m2}}-\cdots-\frac{v_{ml}}{R_{ml}}\right). \tag{10}$$

An arbitrary weighted sum may be obtained by selecting the resistor ratios $R_{m,n}/R_0$, and by making the sum of all the R_m equal to R_n after choosing appropriate values, as necessary, of the two free resistors R_{ml+1},R_{nk+1} connected to ground.

Time-domain functions like integration and differentiation may be performed with op amps which use the electrical properties of capacitors [Figs. 5(d) and 5(e)]. Finally, useful nonlinear functions may be implemented by using nonlinear devices in feedback. For example, an absolute value circuit may be constructed using two op amps and the rectification properties of diodes in the feedback loops [Fig. 5(f)]. For positive v_{in}, diode D_1 conducts but diode D_2 does not, so v_{in} appears unmodified at the output after passing through two unity-gain inverting amplifiers. For negative v_{in}, diode D_2 conducts but D_1 does not, and there is feedback from the output to both op amps. If the amplifiers are not saturated, the sum of feedback currents at the inverting input of the first op amp must equal

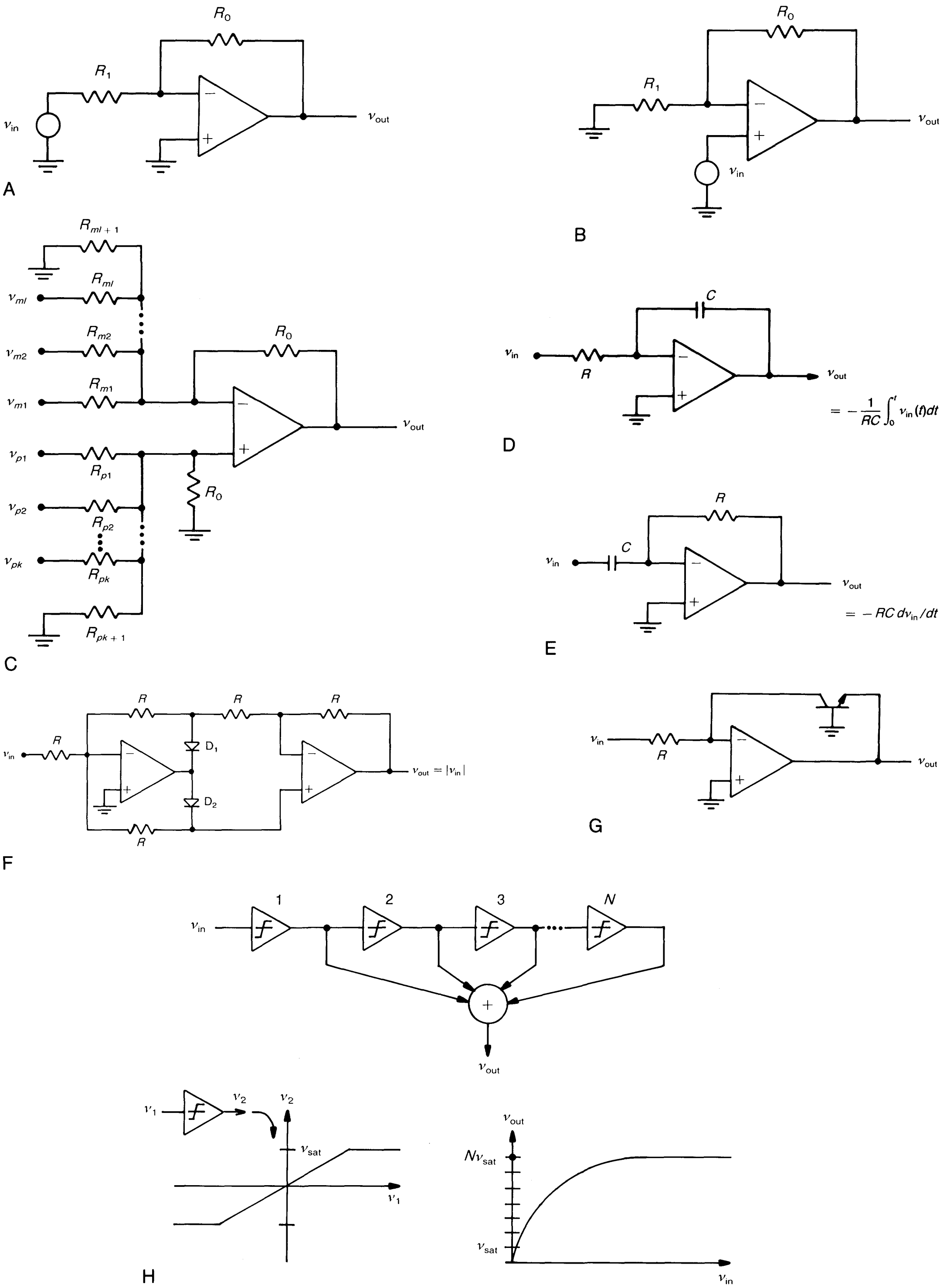
R_0
R_1
v_{in}
v_{out}
A
R_0
R_1
v_{out}
v_{in}
B
R_{ml+1}
R_{ml}
v_{ml}
R_{m2}
v_{m2}
R_{m1}
v_{m1}
R_0
v_{out}
v_{p1}
R_{p1}
R_0
v_{p2}
R_{p2}
v_{pk}
R_{pk}
R_{pk+1}
C
C
v_{in}
R
v_{out}
$= -\frac{1}{RC}\int_0^t v_{in}(t)dt$
D
R
C
v_{in}
v_{out}
$= -RC\, dv_{in}/dt$
E
R
R
R
R
v_{in}
D_1
D_2
$v_{out} = |v_{in}|$
R
F
v_{in}
R
v_{out}
G
1
2
3
N
v_{in}
v_{out}
v_1
v_2
v_2
v_{sat}
v_1
v_{out}
Nv_{sat}
v_{sat}
v_{in}
H

FIG. 5. (a) Inverting voltage amplifier using op amp; (b) noninverting amplifier using op amp; (c) weighted algebraic sum circuit; (d) a voltage integrator; (e) a voltage differentiator; (f) an absolute value circuit; (g) a logarithmic amplifier dependent on transistor characteristics; and (h) a logarithmic amplifier using the characteristics of a cascade of saturating amplifiers. One saturating amplifier is shown with its input-output characteristic. The logarithmic amplifier is obtained by summing the outputs of many such identical amplifiers in cascade.

v_{in}/R, which is only possible if $v_{out}=-v_{in}$. This produces the absolute value function.

The *I-V* characteristics of a *pn* junction may be used to make a logarithmic amplifier [Fig. 5(g)]. The output voltage $v_{out} = -V_T \log_e(v_{in}/RI_s)$ is indeed logarithmic but suffers from a very strong temperature dependence due to the terms I_s and V_T. A variety of ways have been used to reduce or eliminate this temperature dependence, principally relying on taking the difference in the output voltages of two logarithmic amplifiers, one driven by v_{in} and the other by a normalizing reference V_{ref}. This produces a ratiometric logarithmic output which varies proportionally with absolute temperature. This temperature dependence may be removed more simply by other methods. An altogether different method may also be used to produce a logarithmic amplifier, based on the saturation characteristics of an amplifier. If the outputs of a cascade of identical amplifiers are summed [Fig. 5(h)], an approximation to a logarithmic characteristic is obtained which is inherently independent of temperature.

Operational amplifiers are widely available as low-cost integrated circuits (*q.v.*), at times costing only as much as discrete transistors. For dc and low-frequency operation, their terminal characteristics are very close to the ideal, with input resistances of tens of megohms, output resistances of a few ohms, and dc gains exceeding 100 000. These characteristics deteriorate at high frequencies due to the frequency limits of the transistors constituting the op amp circuits, but more significantly due to a deliberate rolloff introduced in gain to ensure stable operation when feedback is applied. The gain may have decreased to unity from its large dc value at frequencies typically lying in the range of 1 to 10 MHz. This may introduce significant nonidealities in operational circuits at signal frequencies approaching the unity-gain frequency.

5. FEEDBACK AMPLIFIERS

Operational circuits are only one, somewhat ideal, instance of the use of negative feedback around amplifiers. Feedback evolved as a means of stabilizing amplifier characteristics, by subtracting a portion of the output signal from the applied input, and using the difference to drive the amplifier. The terminal characteristics are then strongly governed by the external, passive elements used in the feedback network, and become relatively independent of the lack of absolute accuracy in the gain of the active elements constituting the amplifier. Black, working at Bell Telephone Laboratories, first applied the notion of feedback in 1934 to stabilize fluctuations in the gain of vacuum tube amplifiers used in telephone repeaters by subtracting a fraction of the output voltage from the applied input, and feeding the *difference* to the amplifier input (Fig. 6).

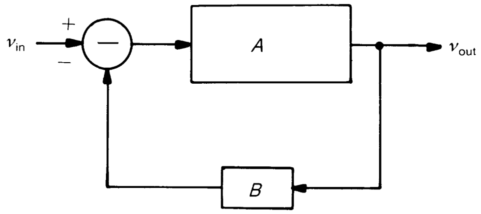

FIG. 6. Block diagram of a negative feedback loop. The quadranted circle at the input indicates the differencing operation between input and output variables (both assumed to be of the same type).

The transfer function of a voltage amplifier with gain *A*, around which negative feedback is applied by subtracting a fraction *B* of the output, is given by

$$\frac{v_{out}}{v_{in}}=\frac{A}{1+AB}. \tag{11}$$

As long as the *loop gain AB* is large, the closed-loop transfer function of the amplifier with feedback is almost equal to $1/B$, thus relatively insensitive to fluctuations in the open-loop amplifier gain *A*.

The use of feedback is illustrated here in a single-transistor amplifier, the emitter-degenerated voltage gain stage [Fig. 7(a)]. This resembles the standard common-emitter amplifier except that a resistor R_E has been inserted in the emitter circuit. Realizing that v_{in} applied at the base of the transistor modulates a current between the emitter and collector, and that this current flowing through R_C produces v_{out}, the implicit feedback loop becomes apparent; for an increase in v_{in} produces an increase in i_c through the transistor, which increases the voltage v_e, *subtracting* from the applied input to reduce the drive to the transistor $v_{be}=v_{in}-v_e$. A block diagram

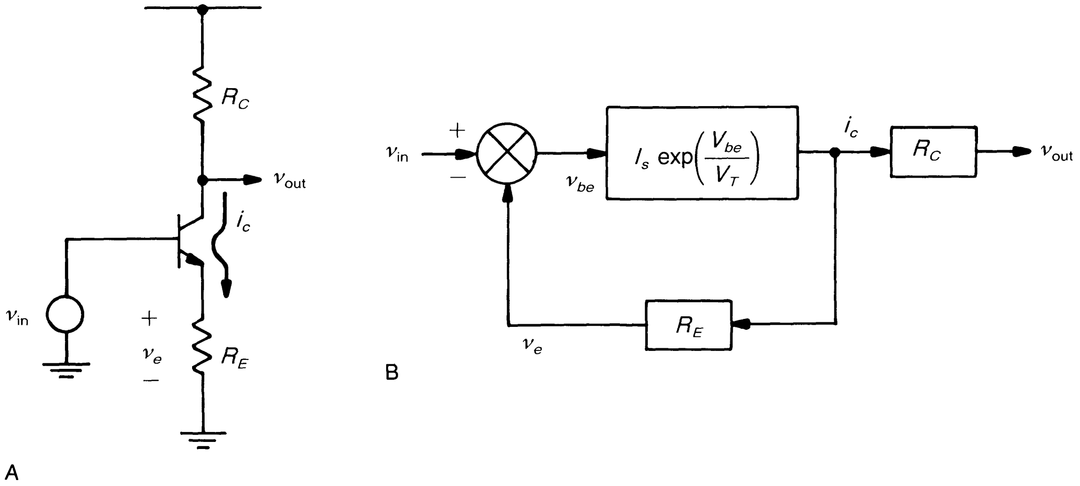

FIG. 7. (a) A single transistor feedback amplifier: the emitter degenerated stage; (b) feedback block diagram of emitter degenerated stage.

specifying the elements of the feedback loop is also shown [Fig. 7(b)]. If the loop gain is high, $i_c \approx v_{in}/R_E$, which implies that the gain

$$v_{out}/v_{in} \approx -R_C/R_E \tag{12}$$

is determined by the feedback element and becomes independent of the transistor characteristics. This is the virtue of feedback: the transistor in the circuit could be replaced by another of different characteristics (different I_s, that is); the temperature, on which transistor current depends exponentially, could be changed, or the bias current could be changed; but as long as the transistor remains in the active region, the gain of the amplifier does not change, being determined by the ratio of two resistors chosen by the user. This desensitization to the characteristics of the active device motivated the very first use of feedback in amplifiers, because vacuum tubes showed a notoriously large spread of characteristics in manufacture, and the characteristics of a given tube were liable to drift with aging.

After application of negative feedback, the gain available from the stage is reduced, because a smaller quantity than the applied input, $v_{in} - v_e$, appears as the effective drive to the amplifying element, whereas in the common-emitter amplifier, all of the input voltage drives the transistor. This reduction in gain is one disadvantage of feedback.

The signal in a feedback amplifier has to flow in a closed loop, forward through the amplifier and back through the feedback path, and after subtracting itself from the applied input, back into the amplifier. At some high signal frequency, the delay through this feedback loop may become comparable to the period at which the signal is varying: if the delay is equal to half the period of the sinusoid of frequency f_0, the signal will experience a phase lag of 180° after traveling through the feedback loop, and instead of subtracting from the applied input as had been intended at low frequencies, the fed back signal will *reinforce* it. If the magnitude of the loop gain (AB) at f_0 is 1 or more, a signal at this frequency can sustain or build itself without the application of an input. Thus, such a feedback amplifier is capable of *oscillation*, undesirable because a sustained sinusoidal output exists for no applied input. The amplifier is, in a manner of speaking, out of control.

Oscillation in a feedback amplifier may be prevented by introducing a stabilizing, or compensating, network in the feedback loop, whose function is to reduce the loop gain to less than 1 at the critical frequency f_0 where the loop phase shifts by 180°. Without this, the feedback amplifier is unusable. The exact conditions for the occurrence of oscillation in a feedback loop are described by the Nyquist criterion, which in its simplest form reduces to the conditions for oscillation described above. Operational amplifiers are compensated to remain stable under the largest dc loop gain a user may subject them to, which occurs when all of the output is subtracted from the input, corresponding to a closed-loop dc gain of 1 (because $B=1$, and the closed-loop transfer function is $1/B$). This means that at the frequency where the open-loop gain of the compensated op amp has rolled off to 1, the phase shift through the op amp is less than the critical 180° (typically, it is kept less than 135° for an added factor of safety). The resulting *unity gain frequency* together with the *dc gain* specify the two most important characteristics of the op amp. Compensation in this way inevitably compromises the bandwidth of an amplifier.

6. POWER AMPLIFIERS

While voltage amplifiers are encountered most often, power amplifiers are another class of circuits having many important applications. Their function is to deliver a signal to a load at an amplified *power level* which may be obtained with a voltage gain of only 1, but a much larger current swing than available directly from the driving source. Thus, the *product* of the signal voltage and current is amplified by this circuit. A power amplifier usually drives a large output signal, and may *distort* at the output due to the saturation or cutoff of the transistors in the circuit.

The simplest power amplifier is an emitter follower (Fig. 8). This has a voltage gain slightly less than 1, but converts current drawn from the power supplies into a signal driving the load. An *npn* emitter follower may provide a very large current for a positive output swing (because the exponential I_C-V_{BE} characteristic requires only a small increase in V_{BE} to produce a large change in output current), but the maximum current to produce a negative swing on the load is limited by the current source I_{EE}. If the output voltage swings more negative than $I_{EE}R_L$, the transistor will be deprived of all emitter current, severing the follower action between the input and output voltages. For a given load, a large undistorted swing requires a large power supply voltage and large I_{EE}, implying that in the absence of a signal, the transistor dissipates considerable power. Those power amplifiers which carry the maximum possible signal current in the quiescent condition are called *Class A* amplifiers. Their power conversion efficiency, defined as the maximum undistorted signal power delivered to the load for unit dc power extracted from the power supplies, is low (25% for the emitter follower) because a large fraction of the power is dissipated in the transistor itself. In that a large heat sink, necessarily bulky and expensive, capable of dispersing the generated heat to keep the transistor temperature at a reasonable value must be used, a Class *A* power amplifier is undesirable.

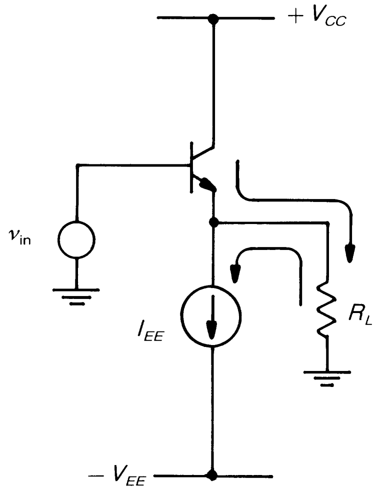

FIG. 8. The emitter follower as a class A power amplifier.

An improvement in efficiency is afforded by *Class B* power amplifiers. These circuits are designed to be normally off, only starting to conduct when current must be delivered to a load. Unlike the Class A amplifier described above, a standing current is not diverted to provide the load current; instead, two complementary transistors are biased back to back, so that one may act as an emitter follower when current must be sourced into the load from the positive power supply, while the other turns on to act as a current sink from the load to the negative power supply (Fig. 9). Only one of the transistors is on at a time when delivering current to the load; no standing current exists in the circuit in the absence of the load. The power conversion efficiency of this circuit is 75%.

The signal current must cross over from being provided by the *npn* transistor to the *pnp* transistor, and this action can give rise to some distortion. For to cause the *npn* to con-

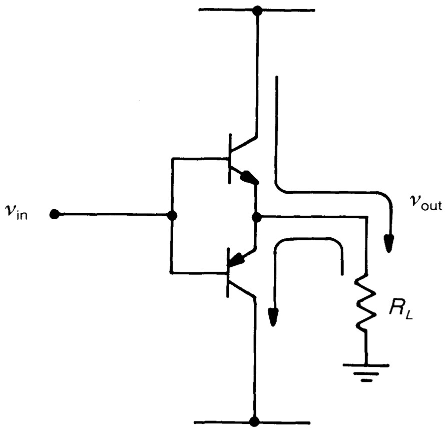

FIG. 9. A class B power amplifier using complementary transistors.

duct, v_{in} must become more positive than $V_{BE}(ON)$, while to cause the *npn* to conduct v_{in} must be more negative than $-V_{BE}(ON)$. This produces a dead band of $2V_{BE}(ON)$ around zero v_{in} over which the output voltage does not respond to the input. The crossover distortion may be overcome by inserting a fixed voltage source of 2 $V_{BE}(ON)$ between the two base terminals to bias both transistors at the edge of conduction. The signal is transmitted to the two bases separated by this offset voltage. The resulting power amplifier has some standing current flowing through the two transistors in series, which is usually quite small to bias them just at the edge of conduction, and so is called a *Class AB* power amplifier. These are most common in low-distortion audio power amplifier circuits.

7. NARROW-BAND AMPLIFIERS

Radio and TV communications signals, consisting of a carrier modulated by a signal with a bandwidth only a small fraction of the carrier frequency, are narrow band in the sense that useful energy is contained in only a small range around the carrier frequency. The carrier frequencies may extend from hundreds of kilohertz (AM) to tens of gigahertz (microwave). Amplifiers for such signals are also narrow band, i.e., their frequency response resembles the signal spectrum in that they offer gain only in the neighborhood of some well-defined center frequency, and the gain rolls off above and below this frequency. This characteristic may be obtained by using, e.g., a parallel *L-C* tuned circuit as the load to a common-emitter transistor amplifier stage (Fig. 10). Such an amplifier may also have a tuned input circuit, which provides an additional measure of *selectivity*, an important feature when the circuit is required to amplify only one of many closely spaced communications channels while filtering out, or rejecting, all the neighboring ones.

Using a small, low-loss inductor, it is possible to make a resonant circuit using the collector capacitance of the transistor itself in shunt with the inductor. This affords the possibility of obtaining gain across a bandwidth dictated by the Q factor of the tuned circuit at resonance, at frequencies well beyond the capability of a standard wide-band common-emitter amplifier with a resistive load, whose collector capacitance is detrimental in shunting high frequencies to ground. This principle of tuning out the detrimental capacitance underlies microwave amplifiers, which are almost always narrow band and offer modest amounts of gain centered at very high frequencies. The highest frequency at which gain may be obtained by this method is limited by transistor physics, for if the input signal varies at a rate faster than the time taken for a charge carrier to transit through the active region (e.g., the base for a bipolar), the effect of the input will tend to smear at the out-

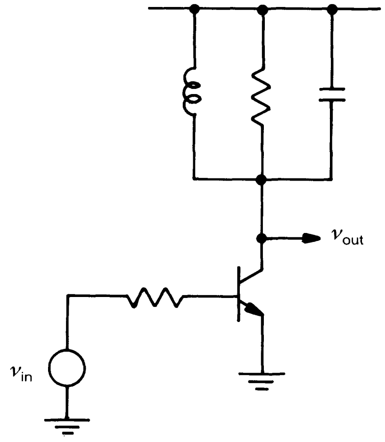

FIG. 10. A tuned amplifier using *L-C* resonant-circuit load.

put, reducing the gain. Thus, to make narrowband amplifiers at ever increasing frequencies, transistor structures with shorter carrier transit times through the active region are necessary, requiring ever more sophisticated technology for fabrication. Heterostructure transistors whose base regions are defined by growing layers a few molecules thick are the latest development in making very high-frequency active devices.

8. DISTRIBUTED AMPLIFIERS

Gains from dc to very high frequencies, often stretching into the microwave spectrum, may be obtained by embedding many stages of amplification into an artificial transmission line terminated in its characteristic impedance. A transmission line with a uniform inductance and capacitance per unit length possesses, ideally, an infinite bandwidth, limited in practice at many gigahertz by losses due to the nonzero resistance of the conductors and losses in the dielectric. With the inputs of many stages of identical amplifying devices connected in parallel by inductors, and their output terminals also connected in parallel by inductors, two transmission lines result consisting of discrete inductors in series and (amplifier) capacitors in parallel. If these are terminated in their characteristic impedance, a very wide-band signal may flow along the input line without attenuation, and the resulting output signals may be made to add up in phase at the output transmission line across a broad bandwidth.

This principle of distributed amplification has been used to make very wide-band microwave amplifiers and, more recently, monolithic microwave integrated circuit (MMIC) amplifiers.

9. NOISE IN AMPLIFIERS

Distortion due to the active devices sets the limit on the largest signal that is reproduced with fidelity by an amplifier. What sets the lower limit? Will an indefinitely small signal appear amplified at the output with fidelity? Random electrical noise in the amplifier components sets the lower limit, and so if a very small signal were applied at the input, the output would indeed contain an amplified replica of the input, which, however, would be overwhelmed in magnitude by the amplified noise. The most common instance of this phenomenon is in radio receivers, which may be receiving a very weak signal of only a few tens of microvolts from a distant transmitter, but which may be overwhelmed by noise at the output.

One source of noise, Johnson or thermal noise, is produced in any resistor by the random thermal agitation of electrons. On average, electrons stay uniformly distributed through the body of the resistor, but instantaneously more electrons may aggregate on one side than the other, producing an electric field between the resistor terminals whose magnitude and polarity varies randomly, but is zero on average. The laws of thermodynamics dictate that the maximum root mean square (rms) noise power available from the resistor is kTB, where k is Boltzmann's constant, T the absolute temperature, and B the bandwidth of observation. This means that the rms noise voltage at the terminals of an open-circuit resistor is

$$v_n(\mathrm{rms}) = (4kTRB)^{1/2}. \tag{13}$$

Another form of noise arises from the random diffusion of carriers across a potential barrier, as is the case in a diode or a junction in a bipolar transistor carrying a direct current. This shot noise appears as fluctuations in the direct current, with zero average but an rms value of

$$i_n(\mathrm{rms}) = (2qI_{\mathrm{dc}}B)^{1/2}. \tag{14}$$

A noise signal has a flat frequency spectrum extending from dc to extremely high frequencies. It may be thought of as constituted of sine waves at all possible frequencies of equal amplitude, added together in random phase. The wave form of noise in time appears like a "fuzz," whose strength is determined by the rms values given in the equations above.

The design of low-noise amplifiers involves choosing the operating points of the active devices and values of the resistors in the circuit so that the sum of their noise contributions amplified at the output does not overwhelm the smallest amplified signal. Radio and television receivers must be designed like this, and take advantage of the fact that they are narrow-band circuits, thus having a small noise bandwidth *B*. Much effort and ingenuity were expended in the early days of radio receivers to make highly sensitive low-noise amplifiers, culminating in the superheterodyne technique, where low-noise amplification is obtained by modulating the envelope of a strong, low-noise, single-frequency local oscillator signal in the receiver with the weak input signal close to the antenna, making the noise due to subsequent stages unimportant. This is now the standard technique employed in all communications receivers.

GLOSSARY

Amplifier: A device capable of magnifying signal power by using an input signal to modulate a reservoir of energy.

Analog Circuit: A circuit in which signal wave forms vary continuously in time and in magnitude.

Gain: Magnification.

Bandwidth: The range of signal frequencies across which uniformly maximum amplification is obtained.

Bipolar Transistor; Bipolar Junction Transistor (BJT): An amplifying device relying on the modulation of current at the output by the control of a potential barrier.

Bias: A nominal current or voltage to place the operating point of a device in a favorable region for amplification.

Common: A connection common to the input and output ports, usually ground.

Field-Effect Transistor (FET): An amplifying device relying on the modulation of a resistive region by an electric field to obtain amplification.

Feedback: Driving an amplifier derived from the difference of an applied input signal and the prevailing output signal (or some fraction, or some filtered version, or some nonlinearly modified version thereof).

Narrow Band: A small fractional bandwidth centered on some (usually high) frequency.

Distributed Circuit: Any circuit becomes distributed when the physical dimensions of its components and interconnections become comparable to the wavelength of signals propagating in it. Also, a circuit with a uniform impedance level throughout its extent.

Noise: Random fluctuations in current and voltage, usually very small, due to the microscopic agitations of charge carriers.

Further Reading

Several good textbooks describe amplifier circuits, a selection from which is listed below.

Gray, P. R., Meyer, R. G. (1984), *Analysis and Design of Analog Integrated Circuits*, New York: Wiley. A very widely referred to intermediate level text on amplifier circuits as they appear in monolithic integrated circuits.

Millman, J., Halkias, C. C. (1972), *Integrated Electronics: Analog and Digital Circuits and Systems*, New York: McGraw-Hill. A comprehensive introductory text on discrete transistor amplifier and oscillator circuits.

Sedra, A., Smith, K. (1987), *Microelectronic Circuits*, New York: Holt, Rinehart and Winston. Another good introductory text on the design and analysis of amplifier circuits, both discrete and integrated.

Terman, F. E. (1955), *Electronic and Radio Engineering*, New York: McGraw-Hill. The classic qualitative approach to vacuum-tube amplifiers and other radio circuits which set the tone of good texts for years to come.

IEEE J. Solid State Circuits, particularly the special Issue on Analog Circuits published in December every year. The main venue for the latest research results in integrated circuit design.

ANALYTIC METHODS

CHARLIE HARPER, *Department of Physics, California State University, Hayward*

Introduction 624

1. **Functions of a Complex Variable** 625
1.1 Introduction 625
1.2 Complex Variables and Their Representations 625
1.3 Analytic Functions of a Complex Variable 626
1.4 Contour Integrals 626
1.5 The Taylor and Laurent Expansions 627
1.6 The Cauchy Residue Theorem 629
1.7 The Cauchy Principal Value and Dispersion Relations 631
1.8 Conformal Transformations 632

2. **Ordinary Differential Equations** 633
2.1 Introduction 633
2.2 First-Order Linear Differential Equations 634
2.2.1 Separable Differential Equations 634
2.2.2 Exact Differential Equations 635
2.2.3 Solution of the General Linear Differential Equation 635
2.3 Second-Order Linear Differential Equations 636
2.3.1 Homogeneous Differential Equations with Constant Coefficients 636
2.3.2 Nonhomogeneous Differential Equations with Constant Coefficients 637
2.3.3 Homogeneous Differential Equations with Variable Coefficients 638
2.3.4 Nonhomogeneous Differential Equations with Variable Coefficients 639
2.4 Some Numerical Methods for Ordinary Differential Equations 640
2.4.1 The Improved Euler Method for First-Order Differential Equations 640
2.4.2 The Runge–Kutta Method for First-Order Differential Equations 641
2.4.3 Second-Order Differential Equations 641

3. **Partial Differential Equations** 641
3.1 Introduction 641
3.2 The Time-Independent Schrödinger Wave Equation 642
3.3 One-Dimensional Mechanical Wave Equation 643
3.4 One-Dimensional Heat Conduction Equation 644
3.5 The Two-Dimensional Laplace Equation 645
3.6 Fourier Transform Method 646
3.7 Green's Functions in Potential Theory 647
3.8 Numerical Methods for Partial Differential Equations 648
3.8.1 Fundamental Relations in Finite Differences 648
3.8.2 Two-Dimensional Laplace Equation: Elliptic Equation 648
3.8.3 One-Dimensional Heat Conduction Equation: Parabolic Equation 648
3.8.4 One-Dimensional Wave Equation: Hyperbolic Equation 648

4. **Integral Equations** 649
4.1 Introduction 649
4.2 Integral Equations with Degenerate Kernels 651
4.3 Integral Equations with Displacement Kernels 652
4.4 The Neumann Series Method 652
4.5 The Abel Problem 652

5. **Applied Functional Analysis** 653
5.1 Introduction 653
5.2 Stationary Values of Certain Definite Integrals 653
5.3 Hamilton's Variational Principle in Mechanics 656

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

5.3.1 Introduction 656
5.3.2 Generalized Coordinates 656
5.3.3 Lagrange's Equations 656
5.3.4 Format for Solving Problems by Use of Lagrange's Equations 657
5.4 Formulation of Hamiltonian Mechanics 657
5.4.1 Derivation of Hamilton's Canonical Equations 657
5.4.2 Format for Solving Problems by Use of Hamilton's Equations 658
5.4.3 Poisson's Brackets 658
5.5 Continuous Media and Fields . 658
5.6 Transitions to Quantum Mechanics 659
5.6.1 The Heisenberg Picture 659
5.6.2 The Schrödinger Picture 659
5.6.3 The Feynman Path-Integral Approach to Quantum Mechanics 659

Glossary 661

Further Reading 661

INTRODUCTION

The article on analytic methods is subdivided into the following five broad and interrelated subjects: functions of a complex variable, ordinary differential equations, partial differential equations, integral equations, and applied functional analysis. Throughout the article, emphasis is placed on methods of application involving physical problems and physical interpretations of solutions rather than on a rigorous mathematical presentation. Special cases of linear relations in one and two Cartesian dimensions are used to explain techniques. Extensions to more general cases and different coordinate systems are straightforward, in principle.

Section 1 is devoted to some aspects of complex variable theory needed in mathematical physics. The section begins with a discussion of complex variables and their representations, analytic and singular functions of a complex variable, important integral relations, and the Taylor and Laurent expansions. The Cauchy residue theorem is applied to obtain the Cauchy principal value of an integral and dispersion relations. A discussion of the uses of dispersion relations throughout physics is also given. The section is concluded with a brief discussion of physical applications of conformal transformations and Riemann surfaces.

Section 2, on ordinary differential equations, treats classes of physical problems that lead to first- and second-order ordinary linear differential equations. Procedures for obtaining solutions for first- and second-order ordinary linear differential equations are presented. Methods of applying initial and boundary conditions are discussed. Green's functions are introduced in connection with the variation of parameters method for solving second-order nonhomogeneous differential equations with variable coefficients. A brief introduction to numerical methods for solving first- and second-order ordinary differential equations is also presented.

In the section on partial differential equations (Sec. 3), some important partial differential equations involving the Laplacian operator are presented and explained. Separation of variables and Fourier transform methods for solving partial differential equations are illustrated. Green's functions for three-dimensional problems are discussed in this section. Extensions to cylindrical and spherical coordinates and to certain special functions in mathematical physics are discussed. The section is concluded with a brief presentation of numerical methods for solving partial differential equations.

An introduction to one-dimensional linear integral equations is given in Sec. 4. Discussions of classifications and methods of solution of integral equations are given. The essential difference between an integral- and a differential-equation formulation of a physical problem is discussed. The Abel problem is presented as an example of a physical problem that leads directly to an integral equation.

The focus of Sec. 5 is on applied functional analysis. The method of the calculus of variations is introduced in connection with finding the extremum of the definite integral of a functional, and techniques of variational calculus are applied to Hamilton's variational principle of mechanics. The Feynman path integral approach to quantum mechanics is presented as an example of functional integration.

1. FUNCTIONS OF A COMPLEX VARIABLE

1.1 Introduction

The imaginary number, $i = \sqrt{-1}$, was introduced into mathematics during the latter part of the sixteenth century. Imaginary numbers are needed since certain equations, for example, $x^2 + 1 = 0$, have no solutions that involve only real numbers. In physics, one writes the solution of the equation of motion for the linear harmonic oscillator, $\ddot{x} + \omega^2 x = 0$, in the form $x(t) = A \exp(i\omega t)$. Index of refraction is written in complex (containing real and imaginary parts) form in modern optics, and the wave function in quantum mechanics is often a complex quantity. How physical results are obtained from complex numbers or functions of a complex variable will be explained below. Complex variables are used throughout physics, and this section is devoted to discussions of some properties of complex variables that are useful in physical applications.

1.2 Complex Variables and Their Representations

A complex variable may be written in the general form

$$z=x+iy=re^{i\theta}. \qquad (1)$$

In Eq. (1), x and y are the respective real and imaginary parts of z and are written as $x=\text{Re}z$ and $y=\text{Im}z$; θ is the argument (phase) of z and is written as $\theta = \arg z = \theta_p + 2\pi n$ for $n=0,1,2,\ldots$; θ_p is the principal argument of z and varies from 0 to 2π; $e^{i\theta} = \cos\theta + i \sin\theta$ (Euler's formula); and $r=|z|$ is the absolute value (magnitude, modulus) of z where $r = (x^2 + y^2)^{1/2}$. The complex conjugate of z is denoted as z^* (for notational convenience, $\bar{z}$ is sometimes used to denote complex conjugate) and is obtained by changing the sign of the imaginary part (or imaginary terms) of z, $z^*=x-iy$. It is clear, and useful to note for physical purposes, that z^*z is a real quantity. Complex variables are subject to the same algebraic laws as real variables. The Argand diagram (z-plane diagram) is a convenient geometrical representation of a complex variable and is illustrated in Fig. 1.

On raising Eq. (1) to the nth power, one obtains

$$z^n=r^n(\cos\theta+i\sin\theta)^n=r^ne^{in\theta}$$
$$(n=0,\pm1,\pm2,\ldots \text{ and } z\neq0). \qquad (2)$$

Equation (2) is de Moivre's theorem and is often written as

$$(\cos\theta+i\sin\theta)^n=\cos(n\theta)+i\sin(n\theta). \qquad (3)$$

de Moivre's theorem may be used to obtain relations involving sines and cosines of multiple angles. On considering de Moivre's theorem for $n=2$, expanding the left-hand side, and equating corresponding real and imaginary parts, the following well-known relations are obtained: $\cos2\theta=\cos^2\theta-\sin^2\theta$ and $\sin2\theta=2\cos\theta\sin\theta$. For $n>2$, the binomial expansion may be used to expand the left-hand side of de Moivre's theorem, Eq. (3).

By use of de Moivre's theorem, the nth root of z may be written as

$$z^{1/n}=r^{1/n}\left[\cos\left(\frac{\theta+2\pi k}{n}\right)+i\sin\left(\frac{\theta+2\pi k}{n}\right)\right];$$
$$k=0,1,2,\ldots,n-1. \qquad (4)$$

The quantity $r^{1/n}$ represents the positive nth root of r. The square root of i, where $r=1$ and $\theta_p = \pi/2$, is found to illustrate the procedure for applying Eq. (4). Roots for $k=0$ and $k=1$, respectively, are

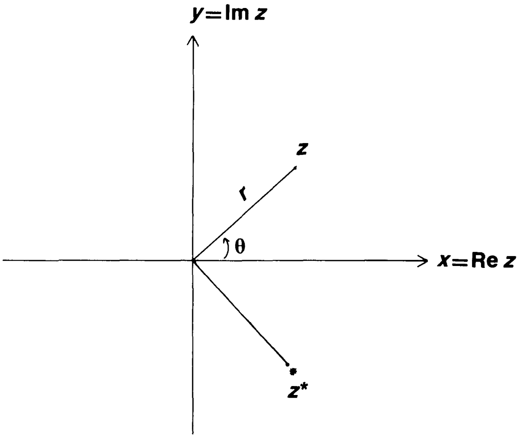

FIG. 1. Argand diagram.

$$z=\cos\left(\frac{\pi}{4}\right)+i\sin\left(\frac{\pi}{4}\right)=\frac{1+i}{\sqrt{2}} \tag{5}$$

and

$$z=\cos\left(\frac{3\pi}{4}\right)+i\sin\left(\frac{3\pi}{4}\right)=-\frac{1+i}{\sqrt{2}}. \tag{6}$$

The above two roots may be checked for correctness. The procedure used to calculate the square root of i can be applied to calculate the nth root of any quantity z ($z\neq 0$).

1.3 Analytic Functions of a Complex Variable

A function $f(z)$ of a complex variable is itself a complex quantity and may be written in terms of real and imaginary parts in the following manner:

$$f(z)=u(x,y)+iv(x,y). \tag{7}$$

The Argand diagram representations of z and $f(z)$ are respectively called z-plane and w-plane diagrams. The number $w=f(z)$ is the value of $f(z)$ at z. A single-valued function $f(z)$ is analytic (regular, holomorphic) at z_0 if it has a unique derivative at z_0 and at every point in the neighborhood of z_0. If a function fails to be analytic at some point z_0 but is analytic at points in the neighborhood of z_0, then z_0 is said to be a singular point (singularity) of $f(z)$. In this connection, note that the function $1/z$ is analytic everywhere except at $z=0$ (singular point). Liouville's theorem states that a function which is analytic for all z (including infinity) must equal a constant.

By analogy with the case of real variables, the derivative of a function of a complex variable is defined as

$$f'(z)=\lim_{\Delta z\to 0}\left(\frac{f(z+\Delta z)-f(z)}{\Delta z}\right)$$

$$=\lim_{\Delta z\to 0}\left(\frac{\Delta u+i\Delta v}{\Delta x+i\Delta y}\right). \tag{8}$$

The evaluation of Eq. (8) for the two paths (a) $\Delta x=0$ and $\Delta y\to 0$ and (b) $\Delta x\to 0$ and $\Delta y=0$ leads to

$$\frac{\partial u}{\partial x}+i\frac{\partial v}{\partial x}=-i\frac{\partial u}{\partial y}+\frac{\partial v}{\partial y}. \tag{9}$$

The Cauchy–Riemann conditions for analytic $f(z)$ result from equating corresponding real and imaginary parts of Eq. (9); the results are

$$\frac{\partial u}{\partial x}=\frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial u}{\partial y}=-\frac{\partial v}{\partial x}. \tag{10}$$

If u and v possess continuous partial derivatives to second order, Eq. (10) leads to

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0 \quad \text{and} \quad \frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}=0. \tag{11}$$

The equations in Eq. (11) are two-dimensional Laplace equations, and functions u and v (called harmonic or conjugate functions) are, therefore, solutions of the two-dimensional Laplace equations. The theory of analytic functions is extremely useful in solving problems in electrostatics, fluid mechanics, or whenever the two-dimensional Laplace equation occurs. The function $f(z)$ also satisfies the two-dimensional Laplace equation.

1.4 Contour Integrals

The integral of a function of a complex variable $f(z)$ is defined in a manner analogous to the case of real variable theory and may be written as

$$\int_C f(z)dz$$

$$\equiv \lim_{\substack{n\to\infty \\ \max|z_j-z_{j-1}|\to 0}} \left(\sum_{j=1}^{n} f(\xi_j)\,(z_j-z_{j-1})\right)$$

$$\equiv \int_{z_0}^{z'} f(z)dz. \tag{12}$$

The path (contour) of integration C is divided into n segments by points z_j, and ξ_j is a point between z_j and z_{j-1} (see Fig. 2). In complex variable theory, the integral in Eq. (12) is referred to as the contour integral of $f(z)$ along the path C from z_0 to z'. The integral around a closed path is denoted as $\oint f(z)dz$. The sign convention for contour integrals is as follows: When the path of integration is traversed such that the region of interest is on the left, the integral is considered positive. Regions in the complex plane are classified as either simply connected or multiply connected. Simply connected regions possess the following three equivalent properties: every closed path within the region contains only points that belong to the region, every closed path within the region can be shrunk to a point, and every scissors cut starting at an arbitrary point on the boundary and finishing at another point

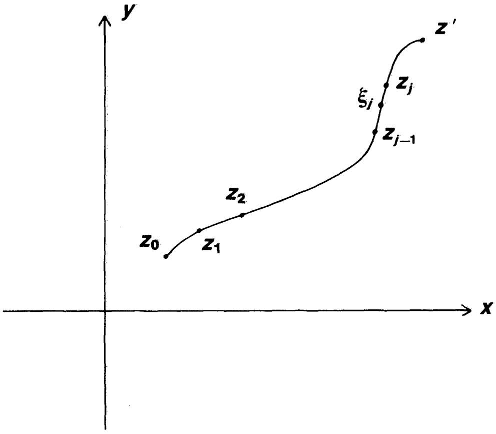

FIG. 2. Path for the contour integral in Eq. (12).

on the boundary separates the region into two unconnected pieces. Regions that are not simply connected are said to be multiply connected.

Two extremely important relations involving integrals of a function of a complex variable, the Cauchy integral theorem and the Cauchy integral formula, will now be discussed.

The CAUCHY INTEGRAL THEOREM: If $f(z)$ is analytic throughout a simply connected region Γ and C is a closed path within Γ, then

$$\oint_C f(z)dz=0. \tag{13}$$

Cauchy's integral theorem applies to special cases that are important in physics where the value of the integral of a function depends only on end points and is independent of the path taken between end points. The inverse of this theorem is also valid.

The CAUCHY INTEGRAL FORMULA is written as

$$\oint_C \frac{f(z)dz}{z-z_0}=2\pi\, if(z_0). \tag{14}$$

The function $f(z)$ in Eq. (14) is analytic within C, z_0 is within C, and the integrand is not analytic at $z = z_0$.

By use of the definition of $f'(z)$ and Cauchy's integral formula, the nth derivative of $f(z)$ evaluated at $z = z_0$ may be written as

$$f^{(n)}(z_0)=\frac{n!}{2\pi\, i}\oint_C \frac{f(z)dz}{(z-z_0)^{n+1}}. \tag{15}$$

Equation (15) will be used below in developing the Taylor expansion for $f(z)$.

1.5 The Taylor and Laurent Expansions

Two important series expansions, Taylor's series and Laurent's series, are valid for a function of a complex variable. If $f(z)$ is analytic in a region Γ and C is a circle within Γ with center at z_0 (see Fig. 3), then the Taylor expansion of $f(z)$ where $f^{(n)}(z_0) = n!a_n$ is

$$f(z)=\sum_{n=0}^{\infty} \frac{(z-z_0)^n}{n!} f^{(n)}(z_0) = \sum_{n=0}^{\infty} a_n(z-z_0)^n. \tag{16}$$

The Taylor expansion of $f(z)$ is obtained and applied in a manner similar to that in real variable theory. Classification of the zeros of $f(z)$ is made by use of Taylor's expansion of $f(z)$ as follows: (a) If $f(z)=0$ at $z = z_0$, the point z_0 is said to be a zero of $f(z)$. (b) If $a_0 = a_1 = \cdots = a_{m-1} = 0$ but $a_m \neq 0$, then $f(z)$ has a zero of order m at $z = z_0$. It is therefore clear that the conditions $f(z_0) = 0$ and $f'(z_0) \neq 0$ indicate the existence of a zero of order $m=1$ (called simple zero) at $z = z_0$.

The Laurent expansion of $f(z)$ has no real-variable counterpart and is key in the discussion of singularities and residues. If $f(z)$ is analytic in the interior and on the boundary of a circular ring between two circles C_1 and C_2 (see Fig. 4), it may be represented as a Laurent expansion which has the form

$$f(z)=\sum_{n=0}^{\infty} a_n(z-z_0)^n+\sum_{n=1}^{\infty} \frac{a_{-n}}{(z-z_0)^n}. \tag{17}$$

In Eq. (17), the coefficients a_n and a_{-n} have the forms

$$a_n=\frac{1}{2\pi i}\oint_C \frac{f(z)dz}{(z-z_0)^{n+1}},\quad n=0,1,2,\ldots, \tag{18}$$

and

$$a_{-n}=\frac{1}{2\pi i}\oint_C (z-z_0)^{n-1}f(z)dz,\quad n=1,2,\ldots. \tag{19}$$

The first series in the Laurent expansion, Eq. (17), is called the analytic part, and it converges everywhere within C_1. The second series in the Laurent expansion is the principal part which converges everywhere outside C_2. The quantity a_{-1} is the residue of $f(z)$ at $z = z_0$ and is given by

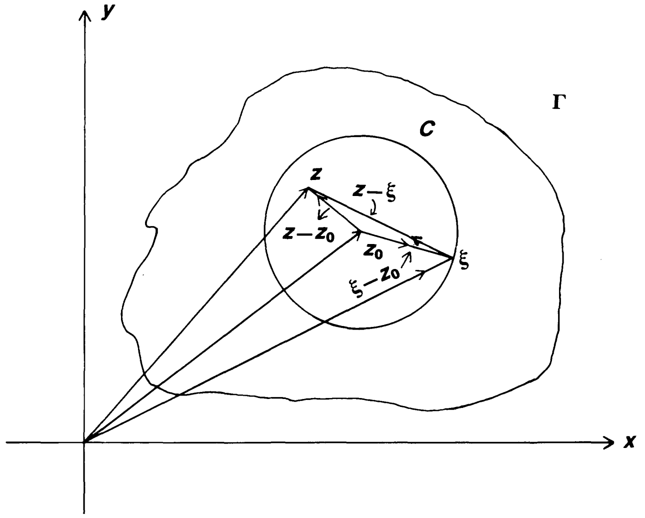

FIG. 3. Diagram for the Taylor expansion.

$$a_{-1}=\frac{1}{2\pi i}\oint_C f(z)dz. \tag{20}$$

In the above three equations, C is any circle between C_1 and C_2 that encloses z_0. Note that $2\pi i a_{-1}$ is the value of the integral in Eq. (20). For cases where the residue can be determined directly, an indirect method of evaluating definite integrals may be developed. First, the classification of isolated singularities and calculations of corresponding residues are considered.

A singularity at z_0 is said to be isolated if a circle of radius ϵ, containing no other singularities, can be drawn with z_0 as its center. Singularities are classified using the principal part of the Laurent expansion. If the first m terms in the principal part are different from zero but the remaining terms equal zero, then $f(z)$ has a singularity (a pole) of order m at z_0. When $m=1$, the pole is called a simple pole. If m is infinite, the singularity at z_0 is said to be an essential singularity. The residue of $f(z)$ at z_0 may be obtained by use of the following three methods.

1. The Laurent expansion directly [the coefficient of the $1/(z-z_0)$ term]. In the Laurent expansion

$$f(z)=\frac{1}{z^3}-\frac{1}{3!z}+\frac{z}{3!}-\frac{z^3}{7!}+\cdots, \tag{21}$$

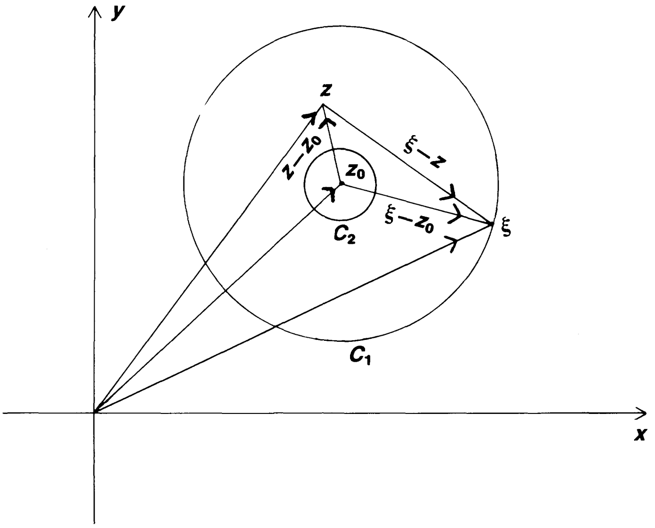

FIG. 4. Diagram for the Laurent expansion.

there is a third-order pole at $z=0$ with residue $a_{-1}=1/3!$.

2. The general formula

$$a_{-1}=\lim_{z\to z_0}\left(\frac{1}{(m-1)!}\,\frac{d^{m-1}\phi(z)}{dz^{m-1}}\right), \tag{22}$$

where

$\phi(z)=(z-z_0)^m f(z)$

for $\lim_{z\to z_0}[\phi(z)]$ analytic and nonzero.

To illustrate the procedure for applying the general formula, let us classify the singularities and calculate the residues of

$$f(z)=\frac{1}{(z^2+a^2)^2}$$

$$=\frac{1}{(z+ia)^2(z-ia)^2} \quad \text{for} \quad a>0. \tag{23}$$

There are singularities at $z=\pm ia$. Note that $m=2$, $\phi(ia)$ is nonzero and analytic, and the residue is $1/(4ia^3)$ when $z=ia$. In a similar manner, the residue for the singularity at $-ia$ is $-1/(4ia^3)$.

3. If $f(z)=g(z)/h(z)$ where $g(z_0)\neq 0$, $h(z_0)=0$, but $h'(z_0)\neq 0$, then

$$a_{-1}=\frac{g(z_0)}{h'(z_0)}\,. \tag{24}$$

For analytic $A(z)$ in $f(z)=A(z)/\sin z$, method **3** may be used to calculate the residue. There are singularities at $z=n\pi$ for $n=0, \pm 1, \pm 2,\ldots$; here the quantity $h(n\pi)$ equals zero, but $h'(n\pi)$ is different from zero. These poles are therefore simple poles and the residue is $a_{-1}=(-1)^n A(n\pi)$.

1.6 The Cauchy Residue Theorem

The Cauchy residue theorem and Cauchy principal value of an integral result from the applications of the Cauchy integral theorem and the Cauchy integral formula. The residue theorem and principal value of an integral are extremely important in physics.

CAUCHY'S RESIDUE THEOREM: If $f(z)$ is analytic within and on a closed region Γ (except at a finite number of isolated singular points z_j within Γ), then (see Fig. 5)

$$\oint_\Gamma f(z)dz$$

$$=2\pi i \sum_{j=1}^{n} a_{-1}z_j$$

$$=2\pi i \sum_{j=1}^{n} [\text{enclosed residue of } f(z)]. \tag{25}$$

For a simple pole on the path of integration (see Fig. 6), the residue theorem yields

$$\oint_C f(z)dz=\pi i a_{-1}. \tag{26}$$

The extension of the residue theorem to cases of simple poles on the path of integration is

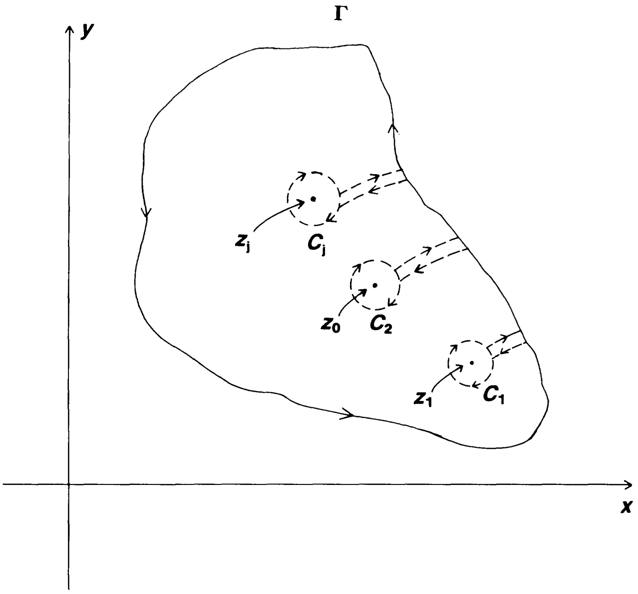

FIG. 5. Diagram for the Cauchy residue theorem.

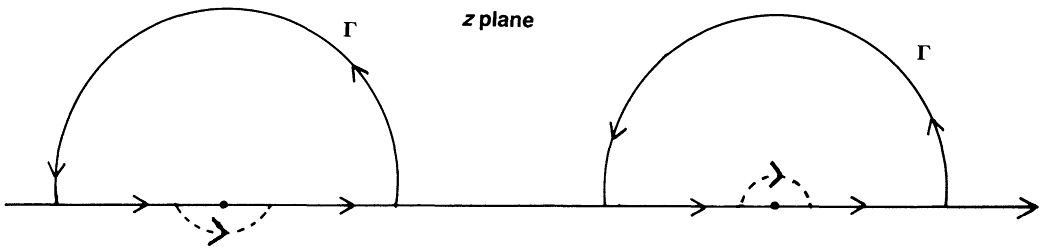

FIG. 6. Simple pole on the path.

important in physics, and the residue theorem is written as

$$\oint_C f(z)\,dz = 2\pi i \sum_{j=1}^{n} (\text{enclosed residue}) + \pi i \sum_{k=1}^{m} (\text{residue of simple poles on path}). \qquad (27)$$

The residue theorem may be used to evaluate certain definite integrals that occur when solving physical problems, and the procedure for evaluating four types of integrals will now be illustrated.

TYPE 1 INTEGRALS:

$$I_1 = \int_0^{2\pi} f(\sin\theta, \cos\theta)\,d\theta.$$

It is assumed that the integrand $f(\sin\theta, \cos\theta)$ contains no singularities other than poles. If $z = e^{i\theta}$ (unit circle), then

$$\sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i} = \frac{z^2 - 1}{2iz}$$

and

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2} = \frac{z^2 + 1}{2z}.$$

In terms of z, the integral I_1 becomes

$$I_1 = -i \oint_{\substack{\text{unit}\\ \text{circle}}} f\left(\frac{z^2-1}{2iz}, \frac{z^2+1}{2z}\right)\frac{dz}{z} = 2\pi \sum (\text{residue within the unit circle}). \qquad (28)$$

The analysis for Type 1 integrals may be used to evaluate the following integral:

$$I_1 = \int_0^{2\pi} \frac{d\theta}{5 + 4\cos\theta} = -i \oint_{\substack{\text{unit}\\ \text{circle}}} \frac{dz}{(2z+1)(z+2)} = \frac{2\pi}{3}. \qquad (29)$$

TYPE 2 INTEGRALS: If (a) $f(z)$ is analytic in the upper-half plane except for a finite number of enclosed singularities z_j and/or simple poles x_k on the real axis, and (b) $zf(z)$ approaches zero as $|z|$ approaches infinity, then

$$I_2 = \int_{-\infty}^{\infty} f(x)\,dx = 2\pi i \sum_{j=1}^{n} a_{-1,z_j} + \pi i \sum_{k=1}^{m} a_{-1,x_k}. \qquad (30)$$

By use of the analysis for Type 2 integrals, let us evaluate

$$I_2 = \int_{-\infty}^{\infty} \frac{dx}{1+x^2} = \oint_{\text{semicircle}} \frac{dz}{(z-i)(z+i)} = \pi. \qquad (31)$$

TYPES 3 AND 4 INTEGRALS: If (a) $f(z)$ is analytic in the upper-half plane except at a finite number of enclosed singular points and/or simple poles on the real axis and (b) $f(z)$ approaches zeros as $|z|$ approaches infinity, then integrals of the form

$$\int_{-\infty}^{\infty} f(x)\exp(imx)\,dx$$

yield I_3 and I_4 where

$$I_3=\int_{\infty}^{\infty} f(x)\cos mx\, dx=-2\pi \sum_{\text{enclosed}} \mathrm{Im}\{\text{residue}[\, f(z)\exp(imz)]\}$$
$$-\pi \sum_{\substack{\text{on the}\\ \text{path}}} \mathrm{Im}\{\text{residue}[\, f(z)\exp(imz)]\} \qquad (32)$$

and

$$I_4=\int_{\infty}^{\infty} f(x)\sin mx\, dx=2\pi \sum_{\text{enclosed}} \mathrm{Re}\{\text{residue}[\, f(z)\exp(imz)]\}$$
$$+\pi \sum_{\substack{\text{on the}\\ \text{path}}} \mathrm{Re}\{\text{residue}[\, f(z)\exp(imz)]\}. \qquad (33)$$

The application of Type 3 integrals is similar to that for Type 4 integrals. Type 4 may be used to evaluate

$$I_4=\int_{-\infty}^{\infty} \frac{\sin x\, dx}{x}$$
$$=\pi \sum \mathrm{Re}\{\text{residue}[\exp(iz)/z]\}=\pi. \qquad (34)$$

1.7 The Cauchy Principal Value and Dispersion Relations

On returning to the case of a simple pole on the real axis, note that it is useful to express the result in terms of the Cauchy principal value of an integral. The integral of a function $f(x)$ which has a simple pole at $x=x_0$ for x_0 within $[a,b]$ may be written as

$$\int_a^b f(x)dx=\lim_{\epsilon\to 0}\left[\int_a^{x_0-\epsilon} f(x)dx + \int_{x_0+\epsilon}^{b} f(x)dx\right]$$
$$\equiv \mathrm{P}\int_a^b f(x)dx. \qquad (35)$$

The symbol P in front of an integral indicates the Cauchy principal value of the integral and means carry out the limiting process in Eq. (35). Note that the Cauchy principal value may exist even if the regular value of the integral does not exist; for example, $\mathrm{P}\int_{-1}^{1} dx/x^3=0$.

Dispersion relations (also known as spectral representations, Kronig–Kramers relations, and Hilbert transforms) result from the analytic properties of the complex representation of physical quantities and the Cauchy residue theorem. Originally, Kronig and Kramers were concerned with the dispersion of light and the relation between the real (dispersive) and imaginary (absorptive) parts of the index of refraction at different frequencies. The basic idea of dispersion relations is applied in areas ranging from electronic design to quantum field theory. Here, general forms for dispersion relations will be presented. For a physical quantity $\chi(\omega)$ which approaches zero as ω approaches infinity and is analytic in the upper-half plane (see Fig. 7), consider the evaluation of the integral

$$\oint_{\Gamma} \frac{\chi(\omega)d\omega}{\omega-\omega_0}. \qquad (36)$$

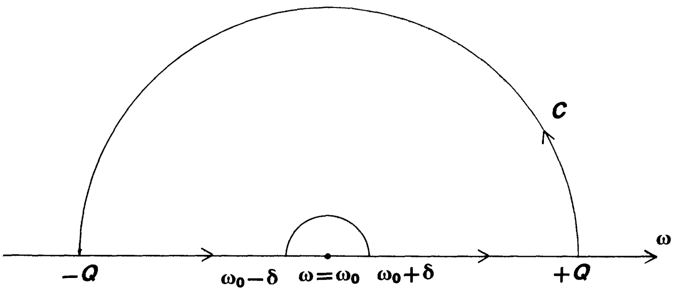

FIG. 7. Contour for Eq. (36).

By use of the Cauchy residue theorem for a simple pole at ω_0 on the contour and the physical property that $\chi(\omega)$ approaches zero as ω approaches infinity, Eq. (36) yields

$$-\chi(\omega_0)=\frac{i}{\pi}\,\mathrm{P}\int_{-\infty}^{\infty}\frac{\chi(\omega)d\omega}{\omega-\omega_0}. \tag{37}$$

On equating corresponding real and imaginary parts in Eq. (37), the dispersion relations are obtained:

$$\begin{aligned}\mathrm{Re}\chi(\omega_0)&=\frac{1}{\pi}\,\mathrm{P}\int_{-\infty}^{\infty}\frac{\mathrm{Im}\chi(\omega)d\omega}{\omega-\omega_0},\\ \mathrm{Im}\chi(\omega_0)&=-\frac{1}{\pi}\,\mathrm{P}\int_{-\infty}^{\infty}\frac{\mathrm{Re}\chi(\omega)d\omega}{\omega-\omega_0}.\end{aligned} \tag{38}$$

The equations in Eq. (38) express one part of an analytic function in terms of an integral involving the other part and are called dispersion relations. In electronics, one has $Z(\omega)=R(\omega)+iX(\omega)$ where Z is impedance, R is resistance, and X is reactance. Dispersion relations may be used to express resistance in terms of reactance. Dispersion relations for light (complex index of refraction $\eta = n_c + ia\alpha$) yield relations between dispersive power and absorption. In addition, a large number of definite integrals may be evaluated by use of the dispersion relations. Dispersion relations applied to $f(z)=\cos x + i \sin x$ lead to values of integrals with integrands of forms $(\sin x)/x$ and $(\cos x)/x$ for limits of integration from minus infinity to plus infinity.

1.8 Conformal Transformations

An analytic function $w=f(z)=u(x,y)+iv(x,y)$ for $z=x+iy$ is completely characterized by two pairs of variables (x,y) and (u,v). Riemann developed a mode of visualizing the relation $w=f(z)$ which uses two separate complex planes, z plane for (x,y) and w plane for the corresponding (u,v). By use of the two-plane picture, the equation $w=f(z)$ defines the transformation (relation, correspondence, or mapping) between the two planes. That is to say, $w=f(z)$ may be used for mapping a set of points (locus, figure) in the z plane (or w plane) into the corresponding figure in the w plane (or z plane). For physical problems, the basic idea involves transforming the geometry of a complicated problem in the z plane into a simpler geometry in the w plane, solving the problem with the simpler geometry, and inverting the transformation to obtain the desired solution in the z plane. The most important class of transformations used in solving physical problems are those that preserve the angle between two straight lines (conformal transformations). The angle-preserving property of conformal transformations will now be illustrated: Assume that two lines intersect at $z=a$ in the z plane and at $w=f(a)$ in the w plane, with elements of length along two lines given respectively by $dz_1=|dz_1|\exp(i\theta_1)$ and $dz_2=|dz_2|\exp(i\theta_2)$. The corresponding elements of length in the w plane are $dw_1=|dz_1|\,|f'(z)|\exp[i(\phi+\theta_1)]$ and $dw_2=|dz_2|\times|f'(z)|\exp[i(\phi+\theta_2)]$ since $dw=dz|f'(z)|\times\exp(i\phi)$. Finally, note that the direction of the corresponding lines in the w plane is rotated by ϕ, but the angle between the lines in the z plane $(\theta_2 - \theta_1)$ equals the angle between the lines in the w plane $[(\phi+\theta_2)-(\phi+\theta_1)]$.

Four often used elementary transformations are the following.

1. TRANSLATION: $w = z + z_0$ for z_0 constant. The transformation equations are

$$\begin{aligned}&w=(x+x_0)+i(y+y_0),\\ &u=x+x_0,\\ &v=y+y_0.\end{aligned} \tag{39}$$

2. MAGNIFICATION: $w=az$ for constant and real a. The transformation equations are

$$w=ax+iay,\quad u=ax,\quad \text{and}\quad v=ay. \tag{40}$$

3. ROTATION: $w = z_0 z$. Here one may write

$$w=\rho\exp(i\phi)=r_0 r\exp[i(\theta+\theta_0)]. \tag{41}$$

The angle in the w plane is $\phi=\theta+\theta_0$ where θ_0 is the angle of rotation and r_0 corresponds to the magnification.

4. INVERSION: $w=1/z$. In polar form, w may be written as

$$w=\rho\exp(i\phi)=\frac{1}{r}\exp(-i\theta). \tag{42}$$

The transformation equations for inversion are

$$\begin{aligned}&u=\frac{x}{x^2+y^2},\quad v=-\frac{y}{x^2+y^2},\\ &x=\frac{u}{u^2+v^2},\quad y=-\frac{v}{u^2+v^2}.\end{aligned} \tag{43}$$

The following transformation is also useful in physical applications: $w = z^2$ which yields $w = \rho \exp(i\phi) = r^2 \exp(i2\theta)$or $\rho = r^2$ and $\phi = 2\theta$ with transformation equations given by $u = x^2 - y^2$ and $v = 2xy$. Here one finds that a circle with radius r_0 is mapped into a corresponding circle with radius $R = r_0^2$, and θ_0 is mapped into $2\theta_0$. In potential theory, the two-dimensional Laplace equation is to be solved with appropriate boundary conditions. Note that the transformation $w = z^2$ maps the right angle in the z plane into a straight line in the w plane (see Fig. 8) where boundary conditions may be applied more conveniently.

In connection with the transformation $w = z^2$ (and other multivalued functions), note that the transformation is conformal except at $w=0$ which is called a branch point, and separately maps the upper- and lower-half planes of the z plane into the whole w plane (points z and $-z$ are mapped into the same points in the w plane). The inverse transformation $z = \sqrt{w}$ cannot be unique. The quantity z may be written as

$$z = \sqrt{\rho} \exp\left(i \frac{\phi}{2}\right) = \sqrt{\rho} \exp\left(i \frac{\phi_p}{2} + i\pi k\right). \tag{44}$$

Odd and even values of k in Eq. (44) yield opposite signs for z. In describing values of a unit circle about the origin in the z plane for $k=0$, it is found that (a) $z = 1$ for $\phi_p = 0$ and (b) $z = -1$ when $\phi_p = 2\pi$. When $k=1$, the values become (a) $z = -1$ when $\phi_p = 0$ and (b) $z = 1$ when $\phi_p = 2\pi$. One may avoid the double values by assuming a cut, which may not be crossed, from zero to infinity along the u axis in the w plane. Riemann introduced the scheme of two planes (sheets, surfaces) joined edge to edge at the cut as a way to combine cases for $k=0$ (all evens) and $k=1$ (all odds) and to eliminate the cut. For example, a lower sheet contains the set of values for $k = 0$ and an upper sheet contains the values for $k = 1$ (see Fig. 9).

The function $\sqrt{w}$ is analytic over the whole Riemann surface (two sheets) except at the branch point, $w=0$. In summary, it is found that the w plane is mapped into two sheets (Riemann surface). The concept of Riemann surfaces has broad application in physics.

2. ORDINARY DIFFERENTIAL EQUATIONS

2.1 Introduction

A differential equation is an equation which contains derivative(s), and it may be either an ordinary or a partial differential equation. Ordinary differential equations contain derivative(s) with respect to one independent variable, and partial differential equations contain partial derivatives with respect to two or more independent variables.

The order of a differential equation is the order of the highest derivative appearing in the equation. The degree of a differential equation is the power of the highest derivative after fractional powers of all derivatives have been removed. If the dependent variable and all of its derivatives are to first power without a product of the dependent variable and a derivative, the differential equation is said to be linear. If a differential equation is not linear, it is classified as nonlinear.

Applications of appropriate physical laws to a large number of physical problems lead to differential equations. In general, a physical process is described by use of a differential equation with appropriate boundary conditions in space and/or initial conditions in time and/or an integral equation. The boundary and/or initial conditions determine from

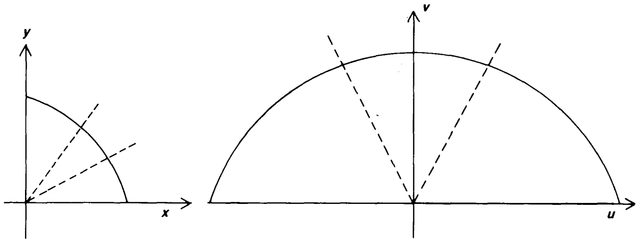

FIG. 8. Diagram for the transformation $w = z^2$.

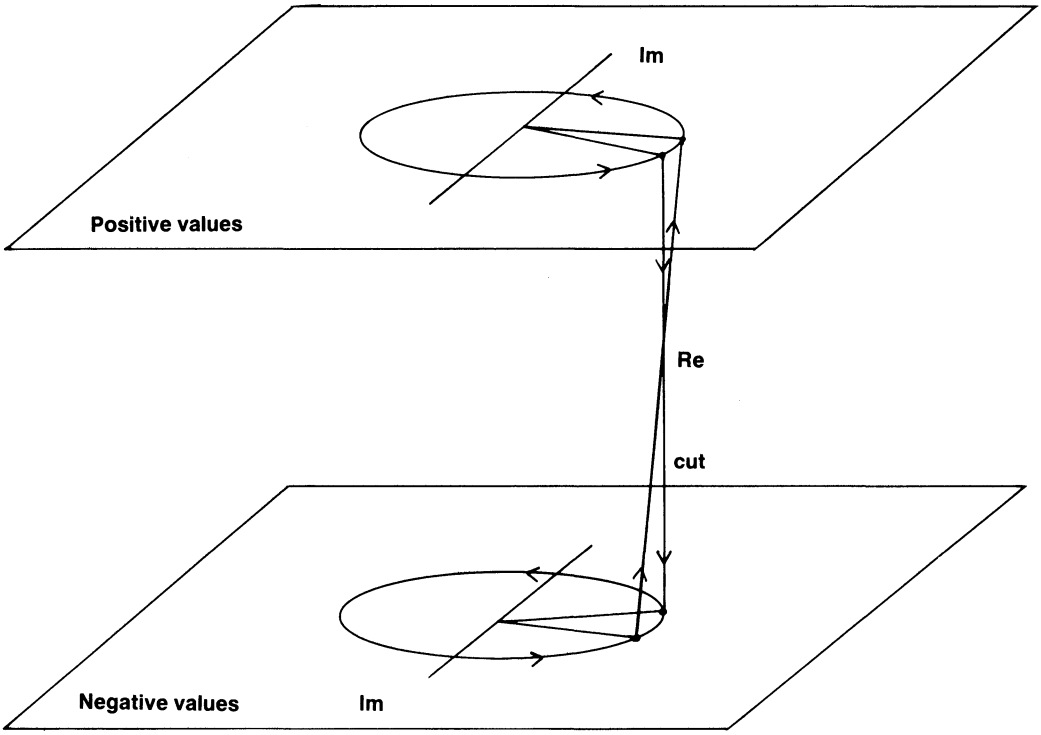

FIG. 9. Riemann surface.

the many possible solutions the one that describes the specific physical phenomenon involved.

The main purpose here concerns the development of solutions for differential equations which adequately describe physical processes under investigation. The mathematical subjects of existence and uniqueness theorems for solutions of differential equations will not be discussed.

An elementary introduction to the subject of ordinary differential equations, as it relates to the needs in solving physical problems, can be reduced to that of treating linear (or reducible to the linear form) first- and second-order differential equations. This presentation is devoted to the construction of solutions and physical applications of such ordinary differential equations.

First- and second-order linear ordinary differential equations have the following standard forms, respectively:

$$\frac{dy}{dx}+p(x)y=Q(x) \quad \text{or} \quad y'+p(x)y=Q(x) \tag{45}$$

and

$$\frac{d^2y}{dx^2}+p(x)\frac{dy}{dx}+q(x)y=f(x)$$

or

$$y''+p(x)y'+q(x)y=f(x). \tag{46}$$

In Eqs. (45) and (46), the notations $y' = dy/dx$ and $y'' = d^2y/dx^2$ have been used. When time t is the independent variable, one writes $\dot{y} = dy/dt$ and $\ddot{y} = d^2y/dt^2$. If the right-hand side of Eqs. (45) or (46) equals zero, the equation is classified as homogeneous; otherwise, the equation is classified as nonhomogeneous (inhomogeneous).

2.2 First-Order Linear Differential Equations

The formulation of many physics problems leads to first-order differential equations, and this section is devoted to solutions of such problems.

2.2.1 Separable Differential Equations. Differential equations that can be put in the form $g(y)dy=f(x)dx$ are called separable differential equations since the left-hand side is a function of y only and the right-hand side is a function of x only. For $dy=f(x)dx$, the general solution is

$$y=\int f(x)dx+C. \tag{47}$$

Since the general solution of a first-order differential equation results from one integration, it will contain one arbitrary constant. Similarly, the general solution of a second-order ordinary differential equation will contain two arbitrary constants. Values of arbitrary constants are determined by use of physical boundary or initial conditions.

EXAMPLE 2.1: In the radioactive decay of nuclei, the process is governed by the following differential equation: $dN/dt=-\lambda N$ with initial condition $N(t=0)=N_0$. The number of parent nuclei present at time t is represented by $N(t)$, and the decay constant λ is characteristic of the particular nuclei involved. The negative sign is used to indicate that the number of nuclei decreases with time. Let us find $N(t)$ subject to the indicated initial condition.

SOLUTION: The differential equation may be written in the form

$$dN/N=-\lambda\, dt$$

with general solution

$$\ln N=-\lambda t+C_1 \text{ or } N(t)=C_2 \exp(-\lambda t). \tag{48}$$

The value of the constant of integration is determined by use of the initial condition, $N(t=0)=N_0$; the initial condition leads to $N(0)=N_0=C_2$. The specific (particular) solution of the problem is the familiar relation

$$N(t)=N_0e^{-\lambda t}. \tag{49}$$

2.2.2 Exact Differential Equations. The general first-order differential equation, $dy/dx=f(x,y)$, may be written in the form

$$M(x,y)dx+N(x,y)dy=0. \tag{50}$$

The total (exact) differential of $F(x,y)=C$ (where F is continuous with continuous derivatives) is

$$dF=\left(\frac{\partial F}{\partial x}\right)_y dx+\left(\frac{\partial F}{\partial y}\right)_x dy=0.$$

Note that the general differential equation in Eq. (50) is exact if

$$M(x,y)=\left(\frac{\partial F}{\partial x}\right)_y \quad \text{and} \quad N(x,y)=\left(\frac{\partial F}{\partial y}\right)_x . \tag{51}$$

Since it is assumed that $F(x,y)$ has continuous first derivatives, note that

$$\left(\frac{\partial M}{\partial y}\right)_x=\left(\frac{\partial N}{\partial x}\right)_y . \tag{52}$$

The condition indicated in Eq. (52) is both necessary and sufficient for Eq. (50) to be an exact differential equation.

EXAMPLE 2.2: Determine whether the following differential equation is exact and find its solution if it is exact: $(4x^3+6xy+y^2) \times dx/dy=-(3x^2+2xy+2)$.

SOLUTION: The standard form of this differential equation is $(4x^3+6xy+y^2)dx+(3x^2+2xy+2)dy=0$; it is exact since the condition in Eq. (52) is satisfied. The solution of the original differential, therefore, has the form $F(x,y)=C$. The function $F(x,y)$ is obtained as follows:

$$\frac{\partial F}{\partial x}=4x^3+6xy+y^2$$

or

$$F(x,y)=x^4+3x^2y+y^2x+f(y) \tag{53}$$

and

$$\frac{\partial F}{\partial y}=3x^2+2xy+2$$

or

$$F(x,y)=3x^2y+y^2x+2y+g(x). \tag{54}$$

Functions $f(y)$ and $g(x)$ arise from integrating with respect to x and y, respectively. For consistency, it is required that $f(y)=2y$ and $g(x)=x^4$. The solution of the original differential equation is $x^4+3x^2y+xy^2+2y=C$.

2.2.3 Solution of the General Linear Differential Equation. A good feature of first-order linear differential equations is that the general equation in this category can be solved. It can be shown that the general solution of Eq. (45) may be obtained from the formula

$$y(x)=\exp\left(-\int p(x)dx\right)\times\int Q(x)\exp\left(\int p(x)dx\right)dx+C\exp\left(-\int p(x)dx\right). \tag{55}$$

If the first-order linear differential equation is separable, the method of Sec. 2.2.1 for separable equation should be followed, and the method of Sec. 2.2.2 for exact differential equations yields solutions for exact differential equations. The formula in Eq. (55) will now be applied to obtain the general solution of the differential equation generated by applying Kirchhoff's loop method to the circuit in Fig. 10.

EXAMPLE 2.3: The appropriate differential equation and initial condition for the circuit in Fig. 10 are

$$L\frac{dI}{dt}+RI=E, \text{ where } I(0)=0. \tag{56}$$

On applying the formula in Eq. (55) for $p(x)=R/L$ and $Q(x)=E/L$ and initial condition, the solution of Eq. (56) reduces to

$$I(t)=\frac{E}{R}[1-\exp(-Rt/L)]. \tag{57}$$

Differential equations of the form $y' + p(x)y = Q(x)y^n$ where $n>1$ (Bernoulli's equation) often occur in physical problems. Bernoulli's type of nonlinear first-order differential equation can be reduced to the linear form by use of the transformation $z = y^{1-n}$, $z' + (1 - n)p(x)z = (1 - n)Q(x)$. The differential equation with dependent variable z can be solved by use of the formula in Eq. (55).

EXAMPLE 2.4: The motion of a particle in a viscous fluid with Stokes damping, av, and Newtonian damping, bv^2, is characterized by an equation of motion of the form $\dot{v} + \alpha v = -\beta v^2$ subject to $v(0) = v_0$. This equation of motion is of the Bernoulli form where $n=2$, $Q(t)=-\beta$, and $p(t)=\alpha$, and the general solution is obtained from

$$z=v^{-1}=-\beta/\alpha+C\exp(\alpha t). \tag{58}$$

The particular solution is

$$v(t)=\frac{\alpha v_0}{(\alpha+v_0\beta)\exp(\alpha t)-\beta v_0}. \tag{59}$$

A graph of Eq. (59) characterizes the speed of the particle as a function of time. To obtain position as a function of time, replace v with $\dot{x}$ and solve the resulting first-order differential equation for $x(t)$.

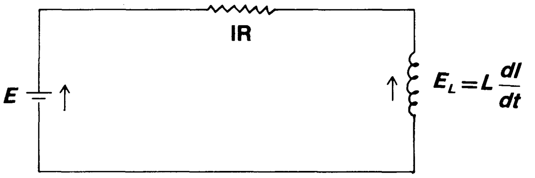

FIG. 10. Diagram for Example 2.3.

2.3 Second-Order Linear Differential Equations

The superposition of solutions principle, stated here in the form of two theorems, will be assumed valid for second-order linear homogeneous differential equations.

THEOREM 1: The set of all solutions of an nth-order linear homogeneous differential equation forms an n-dimensional vector space (see ALGEBRAIC METHODS).

For a second-order differential equation, Theorem 1 means that $y = y_1 + y_2$ is a solution of $y'' + p(x)y' + q(x)y = 0$ if y_1 and y_2 are two linearly independent solutions of the original differential equation.

THEOREM 2: A necessary and sufficient condition that solutions y_1 and y_2 of a second-order linear differential equation be linearly independent is that the Wronskian of these solutions be different from zero. The Wronskian of y_1 and y_2 is the determinant with elements $a_{11}=y_1$, $a_{12}=y_2$, $a_{21}=y_1'$, and $a_{22} = y_2'$.

2.3.1 Homogeneous Differential Equations with Constant Coefficients. The standard form for the general second-order homogeneous differential equation with constant coefficients is $y'' + p_0 y' + q_0 y = 0$ which may be written as $(D^2 + p_0 D + q_0)y = 0$ where $D= d/dx$. The procedure for solving differential equations in this category involves treating $D^2 + p_0 D + q_0 = 0$, the auxiliary or characteristic equation, algebraically and using techniques for solving first-order differential equations. The roots of the auxiliary equation (quadratic) may be real and unequal, real and equal, or a complex-conjugate pair. For real and unequal roots a and b of the auxiliary equation, the differential equation may be written in the symbolic form $(D-a)u=0$ where $u=(D-b)y$. The form of the general solution becomes $y(x) = c_1e^{ax} + c_2e^{bx}$ when the two indicated first-order differential equations are solved. If the roots of the auxiliary equation are a complex-conjugate pair $a^*=b$, the solution of the differential equation has the same form as the case for real and unequal roots with a^* replacing b. The solution of the differential equation for real and equal roots of the auxiliary equation is obtained from solving the two indicated first-order differential equations $(D-a)u=0$ where $u=(D-a)y$; the form of the general solution is $y(x) = (c_1x + c_2)e^{ax}$.

EXAMPLE 2.5: Consider the motion of a particle of mass m initially at rest and subject to a restoring force of $-kx$ and a damping force of $-a\dot{x}$. The equation of motion of this particle is $m\ddot{x}=-kx-a\dot{x}$. The equation of motion in standard form is $\ddot{x}+2\delta\dot{x}+\omega^2x=0$ where $\omega^2 = k/m$ and $2\delta=a/m$ (the factor 2 is used for convenience). Find the solution of the equation of motion for the following cases:

1. $\delta=0$ (no damping);

2. $\delta = \omega$ (critical damping);

3. $\delta<\omega$ (light damping); and

4. $\delta>\omega$ (heavy damping).

Case 1: The equation of motion for $\delta = 0$ reduces to $\ddot{x} + \omega^2x = 0$ with solution

$$\begin{aligned} x(t) &= c_1e^{i\omega t}+c_2e^{-i\omega t} \\ &= A\cos\omega t+B\sin\omega t \\ &= X_0\cos\omega t \text{ for } x(0)=X_0 \\ &\quad \text{and } \dot{x}(0)=0. \end{aligned} \tag{60}$$

The motion without damping is oscillatory and periodic (with constant amplitude X_0).

Case 2: For $\delta\neq 0$ and $\omega\neq 0$, the solutions of the corresponding auxiliary equation are $-\delta+\Delta$ and $-\delta-\Delta$ where $\Delta = \sqrt{\delta^2-\omega^2}$. The solution of the equation of motion for critical damping $\delta=\omega$ is

$$\begin{aligned} x(t) &= (c_1t+c_2)e^{-\delta t} \\ &= X_0(\delta t+1)e^{-\delta t} \text{ for } x(0)=X_0 \\ &\quad \text{and } \dot{x}(0)=0. \end{aligned} \tag{61}$$

Here the motion is not oscillatory and approaches equilibrium at a rapid rate.

Case 3: The solution for light damping $\delta<\omega$ using $\Delta' = \sqrt{\omega^2-\delta^2}$ is

$$\begin{aligned} x(t) &= (A\cos\Delta' t+B\sin\Delta' t)e^{-\delta t} \\ &= X_0\left(\cos\Delta' t+\frac{\delta}{\Delta'}\sin\Delta' t\right)e^{-\delta t} \\ &\quad \text{for } x(0)=X_0 \text{ and } \dot{x}(0)=0. \end{aligned} \tag{62}$$

In this case, the motion is oscillatory with decreasing amplitude (not periodic).

Case 4: The solution for heavy damping $\delta>\omega$ is

$$\begin{aligned} x(t) &= \frac{(\delta+\Delta)}{2\Delta}X_0\exp[(-\delta+\Delta)t] \\ &\quad +\frac{(\Delta-\delta)}{2\Delta}X_0\exp[-(\delta+\Delta)t]. \end{aligned} \tag{63}$$

The motion in this case is not oscillatory and approaches equilibrium at a rate less rapid than for critical damping.

2.3.2 Nonhomogeneous Differential Equations with Constant Coefficients.

The standard form for second-order non-homogeneous differential equations with constant coefficients is $y''+p_0y'+q_0y=f(x)$, and the two widely used methods for solving differential equations in this category are (a) $y = y_h + y_p$ where y_h is the solution of the corresponding homogeneous equation and y_p is any solution of the original nonhomogeneous differential equation, and (b) successive integration. The method of successive integration involves writing the differential equation in the factored form $(D-a)u=f(x)$ where $u=(D-b)y$, and solving the two indicated first-order differential equations.

Physical problems are often solved by use of the first method since y_p can often be obtained without difficulty. Systematic methods for finding y_p for three types of nonhomogeneous terms (polynomial, exponential, and sine and/or cosine) will now be given.

1. The nonhomogeneous term $f(x)$ is a polynomial of degree $n\geqslant 0$.

(A) If zero is not a root of the characteristic equation, then assume

$y_p=A_0+A_1x+\cdots+A_nx^n$.

(B) If zero is a single root of the characteristic equation, then assume

$$y_p=x(A_0+A_1x+\cdots+A_nx^n).$$

(C) If zero is a double root of the characteristic equation, then assume

$$y_p=x^2(A_0+A_1x+\cdots+A_nx^n).$$

2. The nonhomogeneous term $f(x)$ is of the form $C\exp(kx)$.

(A) If k is not a root of the characteristic equation, then assume

$$y_p=A\exp(kx).$$

(B) If k is a single root of the characteristic equation, then assume

$$y_p=Ax\exp(kx).$$

(C) If k is a double root of the characteristic equation, then assume

$$y_p=Ax^2\exp(kx).$$

3. The nonhomogeneous term $f(x)$ is of the form $\sin kx$, $\cos kx$, or $\sin kx+\cos kx$.

(A) If ik is not a root of the characteristic equation, then assume

$$y_p=A\cos kx+B\sin kx.$$

(B) If ik is a single root of the characteristic equation, then assume

$$y_p=Ax\cos kx+Bx\sin kx.$$

Values for constants in the assumed expression for y_p are obtained when that expression is substituted into the original nonhomogeneous differential equation.

EXAMPLE 2.6: The equation of motion for a mass attached to the end of a vertical spring fixed at the other end is $\ddot{y}+\omega^2y=-g$ where g is the acceleration due to gravity. The general solution of the homogeneous equation, $\ddot{y}+\omega^2y=0$, is $y_h=A\cos\omega t+B\sin\omega t$. By use of inspection, it is clear that $y_p=-g/\omega^2$ is a solution of the original nonhomogeneous equation. The solution of the equation of motion for $y(0)=Y_0$ and $\dot{y}(0)=0$ is

$$\begin{aligned}y(t)&=y_h+y_p\\&=A\cos\omega t+B\sin\omega t-\frac{g}{\omega^2}\\&=\left(Y_0+\frac{g}{\omega^2}\right)\cos\omega t-\frac{g}{\omega^2}.\end{aligned}\tag{64}$$

A graph of Eq. (64) characterizes the motion, position as a function of time, of this particle.

2.3.3 Homogeneous Differential Equations with Variable Coefficients. The general procedure used to solve differential equations of the form $y''+p(x)y'+q(x)y=0$ is the power-series method. The power-series method due to Frobenius and Fuchs yields the following two kinds of information concerning the nature of the solution for $x\neq0$: form of the solution as a result of the nature of $p(x)$ and $q(x)$, and form of the solution as indicated by the nature of the solution of the indicial equation. As normally needed in solving physical problems, the general form of the power series solution is

$$y(x)=\sum_{\lambda=0}^{\infty}a_\lambda x^{\lambda+k}\text{ for }a_0\neq0.\tag{65}$$

EXAMPLE 2.7: Consider the differential equation $xy''+2y'+xy=0$. By use of the power-series method, obtain the indicial equation and its two solutions, recursion formula, and general solution of the differential equation. On substituting Eq. (65) into the differential equation to be solved, one obtains

$$\sum_{\lambda=0}^{\infty}a_\lambda(\lambda+k+1)(\lambda+k)x^{\lambda+k-2}+\sum_{\lambda=0}^{\infty}a_\lambda x^{\lambda+k}=0.\tag{66}$$

The basic plan at this stage is to write the result using a single sum. On replacing λ with $\lambda'+2$ in the first sum, the power of x in the first sum becomes the same as that in the second sum. Equation (66) now becomes

$$a_0k(k+1)x^{k-2}+a_1(k+1)(k+2)x^{k-1}+\sum_{\lambda=0}^{\infty}\{a_{\lambda+2}(\lambda+k+3)(\lambda+k+2)+a_\lambda\}\times x^{\lambda+k}=0.\tag{67}$$

Since terms in Eq. (67) are linearly independent, it is required that

$$a_0k(k+1)=0\quad\text{(indicial equation)},\tag{68}$$

$$a_1(k+1)(k+2)=0,\tag{69}$$

and

$$a_{\lambda+2}(\lambda+k+3)(\lambda+k+2)+a_\lambda=0\quad\text{(recursion formula)}.\tag{70}$$

The indicial equation results from equating the coefficient of the lowest power of the variable to zero. In this case, the solutions of the indicial equation are $k=0$ and $k=-1$. When $k=0$, $a_1=0$ because of Eq. (69). The coefficient a_1 is arbitrary when $k=-1$, and two independent solutions of the original differential equation may be obtained by use of $k=-1$ since a_0 is arbitrary by hypothesis. The form of the solution becomes

$$y(x)=\sum_{\lambda=0}^{\infty} a_\lambda x^{\lambda-1}. \tag{71}$$

Coefficients in Eq. (71) are obtained from the recursion formula using $k=-1$. The general expressions for even and odd expansion coefficients, respectively, are

$$\begin{aligned} a_{2j}&=\frac{(-1)^j a_0}{(2j)!}, \\ a_{2j+1}&=\frac{(-1)^j a_1}{(2j+1)!}, \quad j=0,1,2,\dots . \end{aligned} \tag{72}$$

The general solution of the original differential equation is obtained by substituting coefficients in Eq. (72) into Eq. (71).

2.3.4 Nonhomogeneous Differential Equations with Variable Coefficients. Variation of parameters and Green's-function methods are normally used to solve nonhomogeneous linear differential equations with variable coefficients that occur in physics. The standard form for these differential equations is

$$y''+p(x)y'+q(x)y=f(x). \tag{73}$$

The method of variation of parameters due to Lagrange will now be used to solve Eq. (73) subject to the conditions given below. Assume the solution has the form

$$\begin{aligned} y(x)&=C_1 y_1+C_2 y_2 \\ &=v_1(x)y_1+v_2(x)y_2. \end{aligned} \tag{74}$$

In Eq. (74), y_1 and y_2 are two linearly independent solutions of the corresponding homogeneous differential equation, and constant parameters C_1 and C_2 are replaced with functions v_1 and v_2. Functions v_1 and v_2 are unknown parameters to be determined. If $v_1' y_1 + v_2' y_2 = 0$, and $f(x)$ is continuous in the region of interest, then the solution of the original differential equation, Eq. (73), is obtained by use of

$$y(x)=-y_1\int\frac{f(x)y_2dx}{W(y_1,y_2)}+y_2\int\frac{f(x)y_1dx}{W(y_1,y_2)}. \tag{75}$$

The quantity $W(y_1,y_2)$ is the Wronskian of y_1 and y_2.

On using Eq. (75) to solve $y'' - (2y'/x) + (2y/x^2) = (\ln x)/x$ for $x\neq0$, it is found that $y_1=x$ and $y_2=x^2$ are two linearly independent solutions of the corresponding homogeneous equation, the Wronskian equals x^2, and the solution becomes

$$\begin{aligned} y(x)=&-x[(\ln x)^2/2+\ln x+1] \\ &-C_1x+C_2x^2. \end{aligned}$$

Equation (75) will now be put in the form of a definite integral that is useful in solving initial or boundary value problems. Let x be a point in the closed interval $[a,b]$ such that the first term in Eq. (75) is replaced by a definite integral from x to b and the second term in Eq. (75) is replaced by a definite integral from a to x. In terms of the indicated two definite integrals, Eq. (75) becomes

$$\begin{aligned} y(x)&=\int_a^x \frac{y_1(t)y_2(x)f(t)dt}{W(t)} \\ &\quad+\int_x^b \frac{y_1(x)y_2(t)f(t)dt}{W(t)} \\ &=\int_a^b G(x,t)f(t)dt. \end{aligned} \tag{76}$$

The function $G(x,t)$ in Eq. (76) is called the Green's function for Eq. (73) subject to the appropriate boundary conditions. The Green's function is defined by

$$G(x,t)=\begin{cases} \dfrac{y_1(t)y_2(x)}{W(t)}\equiv G_1 & \text{for } a\leqslant t\leqslant x, \\ \dfrac{y_1(x)y_2(t)}{W(t)}\equiv G_2 & \text{for } x\leqslant t\leqslant b. \end{cases} \tag{77}$$

Note that the Green's function depends only on y_1, y_2, and the Wronskian. The quantity $W(t)$ means $W(y_1(t),y_2(t))$. The value of the Green's-function approach is related to the fact that initial or boundary conditions are incorporated in the formulation of the problem in a natural manner. At $t=a$, $G_1(x,t)$ satisfies the boundary condition imposed on $y(x)$, and $G_2(x,t)$ satisfies the boundary condition for $y(x)$ at $t=b$. On applying the Green's- function method to solve $y''=6x$ subject to

$y(0)=y(1)=0$, it is found that $y_1 = x$ and $y_2 = x - 1$ are two linearly independent solutions of the homogeneous equation, the Wronskian equals unity, the Green's functions become $G_1(x,t) = t(x-1)$ for $0 \leqslant t \leqslant x$ and $G_2(x,t) = x(t-1)$ for $x \leqslant t \leqslant 1$, and the solution of the differential equation is $y(x) = \int_0^1 G(x,t) 6t\, dt = x^3 - x$.

2.4 Some Numerical Methods for Ordinary Differential Equations

Numerical methods are treated in detail elsewhere in this *Encyclopedia* (see NUMERICAL METHODS), and a summary of essential features related to solutions of ordinary differential equations is given in this section. In general, the numerical solution of a differential equation consists of a table of values of the dependent variable for corresponding values of the independent variable.

2.4.1 The Improved Euler Method for First-Order Differential Equations. The basic idea of Euler's method for solving first-order ordinary differential equations is to convert the differential equation (continuous) to a difference equation (discrete). The general form for a first-order ordinary differential equation will now be written as

$$\frac{dy}{dx}=f(x,y). \tag{78}$$

By use of the definition of a derivative, one may write

$$\begin{aligned}\frac{dy}{dx} &= \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} \\ &= \lim_{\Delta x \to 0} \left(\frac{y(x+\Delta x)-y(x)}{\Delta x} \right) = f(x,y).\end{aligned} \tag{79}$$

The scheme of the finite difference method involves writing Eq. (79) as

$$y(x_{n+1})=y(x_n)+f(x_n,y_n)\Delta x. \tag{80}$$

Equation (80) is the Euler algorithm for solving first-order ordinary differential equations. The notations in Eq. (80) have the following meanings: $x_{n+1} = x_n + \Delta x$, and $y(x_{n+1}) = y_{n+1}$. To apply Euler's method, select the interval size Δx, evaluate $y(x)$ at x_0, and evaluate $f(x,y)$ at x_0, y_0; the result for $y(x_1)$ is

$$y(x_1)=y(x_0)+f(x_0,\ y_0)\Delta x. \tag{81}$$

A second iteration with inputs $y(x_1)$ from Eq. (81) and $f(x_1, y_1)$ yields $y(x_2)$; the result is

$$y(x_2)=y(x_1)+f(x_1,\ y_1)\Delta x. \tag{82}$$

The iteration is continued to yield a numerical solution of the required first-order ordinary differential equation in the region of interest. A systematic procedure for calculating the error involved during each iteration does not exist for Euler's method.

To improve the simple Euler method, the class of first-order differential equations is restricted to those whose solutions can be expanded in a Taylor series. Neglecting terms of order $O((\Delta x)^3)$, one obtains

$$\begin{aligned}y(x_{n+1}) &= y(x_n)+\Delta x f(x_n,y_n) \\ &\quad + \frac{(\Delta x)^2}{2}\left(\frac{\partial f(x_n,y_n)}{\partial x} \right. \\ &\quad \left. + f(x_n,y_n)\frac{\partial f(x_n,y_n)}{\partial y} \right).\end{aligned} \tag{83}$$

Equation (83) is referred to as the improved Euler method and will be used to obtain the solution of first-order ordinary differential equations.

EXAMPLE 2.8: The equation of motion for a certain particle is $\dot{v} + \alpha v = g$ where $\alpha=0.01$ s^{-1}, $g=9.8$ m s^{-2}, and $v(0)=0$. The analytical solution of this equation of motion is

$$v(t)=\frac{g}{\alpha}\,(1-e^{-\alpha t}). \tag{84}$$

Find the numerical solution of this equation of motion by use of the improved Euler method.

SOLUTION: The general form of the improved Euler method for the differential equation $\dot{v} = g - \alpha v$ is

$$\begin{aligned}v(t_{n+1}) &= v(t_n)+f(t_n,v_n)\Delta t+\frac{(\Delta x)^2}{2}\left(\frac{\partial f(t_n,v_n)}{\partial t} \right. \\ &\quad \left. + f(t_n,v_n)\frac{\partial f(x_n,v_n)}{\partial v} \right).\end{aligned} \tag{85}$$

For arbitrary Δt in Eq. (85), the quantities reduce to

$$\begin{aligned} &f(t_n,v_n)=g-\alpha v(t_n), \quad \frac{\partial f(t_n,v_n)}{\partial t}=0, \\ &\frac{\partial f(t_n,v_n)}{\partial v}=-\alpha. \end{aligned} \tag{86}$$

The essential programming statement for calculating the numerical solution of the original differential equation is

$$v(n+1)=v(n)+[g-\alpha v(n)][\Delta t-\alpha(\Delta t)^2/2]. \tag{87}$$

2.4.2 The Runge–Kutta Method for First-Order Differential Equations. There exist many methods for finding numerical solutions of first-order ordinary differential equations, and the fourth-order Runge–Kutta method is probably the most often used method. As with the Euler and the improved Euler methods, the essential problem is to generate a table of values for x and y for the differential equation $y'=f(x,y)$ when $y(x)$ at $x = x_0$ is given. The task is to develop a method for finding y_1 at $x_0+ \Delta x$, y_2 at $x_0+2\Delta x$, and successive values for y_n throughout the range of interest. For calculating successive values of $y(x_n)$ in the differential equation $y'=f(x,y)$, Runge–Kutta methods use a recurrence formula in the form

$$y_{i+1}=y_i+\Delta x \sum_{i=1}^{n} a_i k_i. \tag{88}$$

Of the many parameters a_i and k_i in Eq. (88), some are chosen arbitrarily and others are obtained by use of the Taylor series involving one and two variables. The order of the Runge–Kutta approximation is indicated by the value of n in Eq. (88). Evaluation of the parameters in Eq. (88) for $n>4$ in the Runge–Kutta approximation is straightforward but involves tedious algebraic manipulations. For $h=\Delta x$, the formula for the fourth-order Runge–Kutta method reduces to

$$y_{i+1}=y_i+\tfrac{1}{6}(k_1+2k_2+2k_3+k_4)+O(h^5). \tag{89}$$

The parameters in Eq. (89) are determined by use of

$$k_1=hf(x_i,y_i),\ k_2=hf\left(x_i+\frac{h}{2},y_i+\frac{k_1}{2}\right),$$

$$k_3=hf\left(x_i+\frac{h}{2},y_i+\frac{k_2}{2}\right),$$

$$k_4=hf(x_i+h,y_i+k_3).$$

EXAMPLE 2.9: Find the numerical solution of the differential equation in Example 2.8 by use of the fourth-order Runge–Kutta method.

SOLUTION: The general form of the Runge–Kutta method for the differential equation $\dot{v} = g - \alpha v$ is

$$v(n+1)=v(n)+\tfrac{1}{6}(k_1+2k_2+2k_3+k_4). \tag{90}$$

The k parameters reduce to $k_1 = h[g - \alpha v(n)]$, $k_2 = h\{g - \alpha[v(n + h/2)+k_1/2]\}$, $k_3 = h\{g - \alpha[v(n + h/2) + k_2/2]\}$, and $k_4 = h\{g - \alpha[v(n + h) + k_3]\}$.

2.4.3 Second-Order Differential Equations. Numerical solutions of second-order differential equations are obtained by first reducing them to a system of first-order differential equations and applying the methods for solving first-order differential equations. The general second-order differential equation may be written as

$$\frac{d^2y}{dx^2}=f(x,y,y'). \tag{91}$$

For $z=dy/dx$, Eq. (91) reduces to the following pair of first-order differential equations:

$$\frac{dz}{dx}=f(x,y,z) \quad\text{and}\quad \frac{dy}{dx}=z=g(x,y,z). \tag{92}$$

The procedure for solving Eq. (91) is to solve the first equation in Eq. (92) with condition $y'(0)$ and use that result as an input for the second equation in Eq. (92) to obtain the solution $y(x)$ with condition $y(0)$.

3. PARTIAL DIFFERENTIAL EQUATIONS

3.1 Introduction

Physical problems involving two or more independent variables are often described by use of partial differential equations. Partial differential equations contain partial derivatives with respect to two or more independent variables. The procedures for determining order, degree, whether linear or nonlinear, and whether homogeneous or nonhomogeneous for partial differential equations are the same as for ordinary differential equations. Some methods for solving partial differential equations are direct integration, characteristics, separation of variables, Fourier and Laplace transforms, and Green's functions. Appropriate boundary (space) and/or initial (time) conditions must be applied to the general solution of a partial differential equation to obtain a suitable solution for the problem under investigation. Three common types of boundary conditions are Dirichlet, specification of the solution at each point on the boundary; Neumann, specification of the normal derivative of the solution at each point on the boundary; and Cauchy, specification of both initial value(s) and the Dirichlet or Neumann condition.

The following equations are examples of important partial differential equations in physics involving the Laplacian operator,

$$\nabla^2=\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}:$$

$$\nabla^2 u=0;\ \text{Laplace's equation.} \tag{93}$$

The function $u(x,y,z)$ in Eq. (93) may represent electric potential in a charge-free region, gravitational potential in a region free of matter, or steady-state (time-independent) temperature in a region without a heat source.

$$\nabla^2 u=f(x,y,z);\ \text{Poisson's equation.} \tag{94}$$

The function $u(x,y,z)$ in Eq. (94) may represent electric potential, gravitational potential, or steady-state temperature in regions with respective sources denoted by $f(x,y,z)$.

$$\nabla^2 u=\frac{1}{\sigma}\frac{\partial u}{\partial t};\ \text{heat conduction (or diffusion) equation.} \tag{95}$$

In Eq. (95), the function $u(x,y,z,t)$ may represent a time-dependent temperature in a region without a heat source or concentration of a diffusing substance. The constant σ is called the diffusivity.

$$\nabla^2 u=\frac{1}{v^2}\frac{\partial^2 u}{\partial t^2};\ \text{mechanical wave equation.} \tag{96}$$

The function $u(x,y,z,t)$ in Eq. (96) may represent the motion of a vibrating string or membrane, and v is the speed of the wave motion.

$$\left\{-\frac{\hbar^2}{2m}\nabla^2+V(x,y,z)\right\}\Psi=i\hbar\frac{\partial\Psi}{\partial t};\ \text{Schrödinger's equation.} \tag{97}$$

Schrödinger's wave equation is the basic equation of motion of a microscopic particle of mass m, and $\Psi(x,y,z,t)$ is called the wave function. The potential energy of the particle is represented by $V(x,y,z)$, and other quantities in this equation have their usual meaning.

This section on partial differential equations is mainly devoted to the physical applications of linear second-order homogeneous partial differential equations in two independent variables; the general form for equations in this category is

$$A\frac{\partial^2 u}{\partial x^2}+2B\frac{\partial^2 u}{\partial x\partial y}+C\frac{\partial^2 u}{\partial y^2}+D\frac{\partial u}{\partial x}+E\frac{\partial u}{\partial y}+Fu=0. \tag{98}$$

In Eq. (98), the coefficients may be functions of x and y, and properties of the solution of the differential equation depend on the relative magnitudes of the coefficients. Based on the coefficients, partial differential equations are classified as elliptic, hyperbolic, or parabolic for $AC-B^2$ greater than zero, less than zero, or equal to zero, respectively. This classification is related to the general equation of a conic section ($Ax^2+2Bxy+Cy^2=1$) representing an ellipse, a hyperbola, or a parabola. According to these classifications, note that Laplace's equation is elliptic, the mechanical wave equation is hyperbolic, and the heat conduction (diffusion) and Schrödinger equations are parabolic. The geometrically related classifications are not of primary importance when solving the differential equation by use of analytical methods but do reflect the nature of the boundary conditions. Solutions of elliptic equations must satisfy conditions on a closed boundary. In this section, the focus will be on separation of variables and Fourier transforms as methods for solving the partial differentials involved in physical applications. The method of separation of variables is illustrated in the following four sections. The Fourier transform method is presented in Sec. 3.6, and Sec. 3.7 is devoted to the Green's-function method for three-dimensional problems.

3.2 The Time-Independent Schrödinger Wave Equation

The method of separation of variables will now be used to obtain the time-independent Schrödinger wave equation. Assuming that

$\Psi(x,y,z,t) = \psi(x,y,z)T(t)$ in Eq. (97) and dividing both sides of the resulting equation by ψT, the result obtained is

$$\frac{1}{\psi}\left(-\frac{\hbar^2}{2m}\right)\nabla^2\psi + V(x,y,z) = \frac{i\hbar}{T}\frac{dT}{dt} \equiv E. \tag{99}$$

Since the left-hand side of Eq. (99) is a function of space only and the right-hand side is a function of time only (time has been separated from the space variables), each side must equal a constant (separation constant) that is independent of space and time. The separation constant is a physical parameter when solving physical problems and has the dimensions of energy, denoted by E, in Eq. (99). Equation (99) leads to

$$T(t) = C\exp(-iEt/\hbar), \tag{100}$$

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V(x,y,z)\right)\psi = E\psi. \tag{101}$$

Equation (101) is the time-independent (steady-state) Schrödinger wave equation. Analyses of solutions of Eq. (101) for various potentials and use of fundamental postulates of quantum theory form the major part of the study of quantum mechanics.

3.3 One-Dimensional Mechanical Wave Equation

Here the one-dimensional mechanical wave equation characterizing the motion of a string fixed at the ends $u(0,t) = u(L,t) = 0$ with initial configuration such that $u(x,0) = 2hx/L$ for x in the closed interval $[0,L/2]$ and $u(x,0) = 2h(L-x)/L$ for x in the closed interval $[L/2,L]$ is solved. The string is initially at rest which means that the partial derivative of $u(x,t)$ with respect to t evaluated at $t=0$ equals zero, $u_t(x,0) = 0$ (see Fig. 11). The method of separation of variables is applied to the equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 u}{\partial t^2}. \tag{102}$$

Assume $u(x,t) = X(x)T(t)$ in Eq. (102) and divide the resulting equation by XT. The result is

$$\frac{1}{X}\frac{d^2X}{dx^2} = \frac{1}{v^2}\frac{1}{T}\frac{d^2T}{dt^2}. \tag{103}$$

Since the left-hand side of Eq. (103) is a function of x only and the right-hand side is a function of time only, the two sides must equal a constant (separation constant). The separation constant is denoted by $-\lambda^2$. The square is used for convenience as will be seen below. The negative sign is selected since an oscillatory solution is anticipated. Boundary conditions, however, will determine the required sign for the separation constant. Equation (103) leads to the following two ordinary differential equations

$$\frac{d^2X}{dx^2} + \lambda^2 X = 0 \text{ with solution}$$

$$X(x) = A\cos\lambda x + B\sin\lambda x \tag{104}$$

and

$$\frac{d^2T}{dt^2} + \lambda^2 v^2 T = 0 \text{ with solution}$$

$$T(t) = C\cos\lambda vt + D\sin\lambda vt. \tag{105}$$

The general solution of Eq. (102) is

$$u(x,t) = (A\cos\lambda x + B\sin\lambda x)(C\cos\lambda vt + D\sin\lambda vt). \tag{106}$$

Boundary and initial conditions will now be used to determine the values of the arbitrary

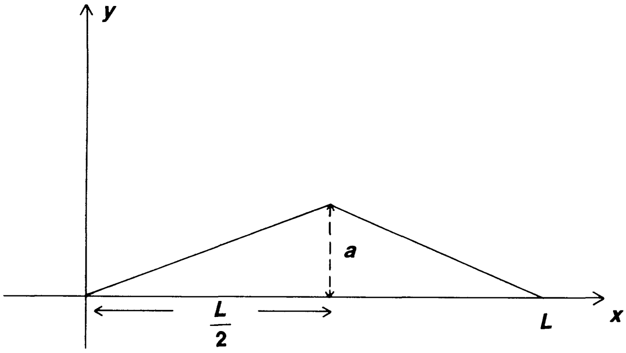

FIG. 11. Initial configuration of the string.

constants in Eq. (106). The first end-point condition $u(0,t)=0$ in Eq. (106) leads to $A=0$. The second end- point condition $u(L,t)=0$ requires that $\sin \lambda L=0$ for a nontrivial solution or $\lambda_n=n\pi/L$ where n ranges from unity to infinity. The solution now reduces to

$$u(x,t)=\sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi x}{L}\right)\left[C_n \cos\left(\frac{n\pi vt}{L}\right) + D_n \sin\left(\frac{n\pi vt}{L}\right)\right]. \qquad (107)$$

Condition $u_t(x,0)=0$ substituted into the partial derivative of $u(x,t)$ with respect to t requires that $D_n=0$ for all n, and the resulting solution becomes

$$(x,t)=\sum_{n=1}^{\infty} B_n' \sin\left(\frac{n\pi x}{L}\right)\cos\left(\frac{n\pi vt}{L}\right); \qquad B_n'=B_nC_n. \qquad (108)$$

The B_n' coefficients in Eq. (108) are evaluated by use of the Fourier sine series. A detailed discussion of the Fourier series method is given elsewhere in this *Encyclopedia* (see FOURIER AND OTHER MATHEMATICAL TRANSFORMS). Here a summary of Fourier series concepts needed in solving boundary valued problems is presented.

The Fourier representation of $f(x)$ in the closed interval $[-L,L]$ is

$$f(x)=\frac{a_0}{2}+\sum_{n=1}^{\infty}\left[a_n \cos\left(\frac{n\pi x}{L}\right)+b_n \sin\frac{n\pi x}{L}\right]. \qquad (109)$$

Coefficients in Eq. (109) are determined by use of (Euler's formulas)

$$a_n=\frac{1}{L}\int_{-L}^{L} f(x)\cos\left(\frac{n\pi x}{L}\right)dx, \quad n=0,1,2,\ldots, \qquad (110)$$

and

$$b_n=\frac{1}{L}\int_{-L}^{L} f(x)\sin\left(\frac{n\pi x}{L}\right)dx, \quad n=1,2,3,\ldots. \qquad (111)$$

Equation (109) is valid in $[-L,L]$ when $f(x)$ is single valued, is bounded, has at most a finite number of maxima and minima, and has at most a finite number of discontinuities. If $f(x)$ is an even function, $f(x)=f(-x)$, the Fourier cosine series results, and the Fourier sine series results when $f(x)$ is odd, $f(x)=-f(-x)$.

The final condition for Eq. (108), $u(x,0)=2hx/L$ for $[0,L/2]$ and $u(x,0)=2h(L-x)/L$ for $[L/2,L]$, leads to a Fourier sine series from which the B_n' may be obtained. The expression for the B_n' coefficients is

$$\begin{aligned}
B_n' &= \frac{2}{L}\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)dx \\
&= \frac{4h}{L^2}\left[\int_0^{L/2} x\sin\left(\frac{n\pi x}{L}\right)dx + L\int_{L/2}^{L} \sin\left(\frac{n\pi x}{L}\right)dx - \int_{L/2}^{L} x\sin\left(\frac{n\pi x}{L}\right)dx\right] \\
&= \frac{8h}{n^2\pi^2}\sin\left(\frac{n\pi}{2}\right) \quad \text{for } n \text{ odd}, \\
&= 0 \quad \text{for} \quad n \text{ even}.
\end{aligned} \qquad (112)$$

The particular solution of Eq. (102) reduces to

$$u(x,t)=\frac{8h}{\pi^2}\sum_{\text{odd}}^{\infty}\left[\frac{(-1)^{(n-1)/2}}{n^2} \times \sin\left(\frac{n\pi x}{L}\right)\cos\left(\frac{n\pi vt}{L}\right)\right]. \qquad (113)$$

The motion of the string is such that only odd-harmonics occur and is symmetrical about the midpoint.

3.4 One-Dimensional Heat Conduction Equation

The method of separation of variables will now be applied to solve the one-dimensional heat conduction equation for the temperature distribution $u(x,t)$ in a rod of length L such that $u(0,t)=u(L,t)=0$ and $u(x,0)=T_0 \times\exp(-ax^2)$. The one-dimensional heat conduction equation is

$$\frac{\partial^2 u}{\partial x^2}=\frac{1}{\sigma}\frac{\partial u}{\partial t}. \tag{114}$$

In Eq. (114), substitute $u(x,t)=X(x)T(t)$ and divide the resulting equation by XT; the resulting two ordinary differential equations for separation constant $-\lambda^2$ are

$$\frac{d^2X}{dx^2}+\lambda^2X=0 \quad \text{with solution}$$

$$X(t)=A\cos\lambda x+B\sin\lambda x \tag{115}$$

and

$$\frac{dT}{dt}+\lambda^2\sigma T=0 \text{ with solution}$$

$$T(t)=C\exp(-\lambda^2\sigma t). \tag{116}$$

The general solution of Eq. (114) is

$$u(x,t)=(A\cos\lambda x+B\sin\lambda x) \times[C\exp(-\lambda^2\sigma t)]. \tag{117}$$

Conditions $u(0,t)=u(L,t)=0$ lead to $A=0$ and $\lambda_n=n\pi/L$ for $n=1, 2,\ldots$, respectively. The final condition yields

$$u(x,0)=T_0\exp(-ax^2)=\sum_{n=1}^{\infty} B'_n\sin\left(\frac{n\pi x}{L}\right). \tag{118}$$

Equation (118) is just a Fourier sine series, and the B'_n coefficients are given by

$$B'_n=\frac{2}{L}\int_0^L T_0\exp(-ax^2)\sin\left(\frac{n\pi x}{L}\right)dx$$

$$=\frac{4T_0}{n\pi} \text{ for } n \text{ odd}, \tag{119}$$

$$=0 \text{ for } n \text{ even}.$$

The particular relation for the temperature distribution in the rod is therefore given by

$$u(x,t)=\frac{4T_0}{\pi}\sum_{\text{odd}}^{\infty}\frac{1}{n}\sin\left(\frac{n\pi x}{L}\right)\exp\left(-\frac{n^2\pi^4\sigma t}{L}\right). \tag{120}$$

3.5 The Two-Dimensional Laplace Equation

Laplace's equation is an example of an elliptic differential equation, and solutions of Laplace's equations are called harmonic functions. The electric potential $u(x,y)$ at points inside a rectangle (see Fig. 12) will now be determined from the solution of the two-dimensional Laplace equation with the indicated boundary conditions:

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0; \quad u(0,y)=u(a,y)=u(x,\infty)=0$$

$$\text{and} \quad u(x,0)=u_0. \tag{121}$$

Separation of variables with separation constant $-\lambda^2$ yields

$$\frac{d^2X}{dx^2}+\lambda^2X=0 \text{ with solution}$$

$$X(x)=A\cos\lambda x+B\sin\lambda x \tag{122}$$

and

$$\frac{d^2Y}{dy^2}-\lambda^2Y=0 \text{ with solution}$$

$$Y(y)=C\exp(\lambda y)+D\exp(-\lambda y). \tag{123}$$

The general solution of Eq. (121) is

$$u(x,y)=(A\cos\lambda x+B\sin\lambda x)[C\exp(\lambda y) +D\exp(-\lambda y)]. \tag{124}$$

Condition $u(x,\infty)=0$ requires that $C=0$, condition $u(0,y)=0$ leads to $A=0$, and condition $u(a,y)=0$ gives $\lambda_n=n\pi/a$ for $n=1, 2, 3,\ldots$. The general solution now reduces to

$$u(x,y)=\sum_{n=1}^{\infty} B'_n\sin\left(\frac{n\pi x}{a}\right)\exp\left(-\frac{n\pi y}{a}\right),$$

$$B'_n=B_nD_n. \tag{125}$$

The final condition is used to determine the values of B'_n as follows:

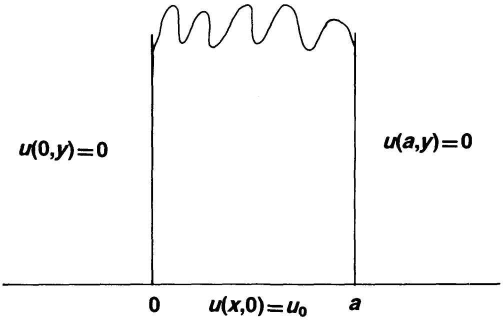

FIG. 12. Boundary configuration for Eq. (121).

$$u(x,0)=u_0=\sum_{n=1}^{\infty} B_n' \sin\left(\frac{n\pi x}{a}\right). \tag{126}$$

Equation (126) is just a Fourier sine series, and the B_n' are given by

$$\begin{aligned} B_n' &= \frac{2}{a}\int_0^a u_0 \sin\left(\frac{n\pi x}{a}\right)dx = -\frac{4u_0}{n\pi} \text{ for } n \text{ odd,} \\ &= 0 \text{ for } n \text{ even.} \end{aligned} \tag{127}$$

The particular solution, expression for the potential at points within the rectangle in Fig. 12, is therefore

$$u(x,y) = -\frac{4u_0}{\pi}\sum_{\text{odd}}^{\infty}\frac{1}{n}\sin\left(\frac{n\pi x}{a}\right)\exp\left(-\frac{n\pi y}{a}\right). \tag{128}$$

The extension to more than two independent variables is straightforward. While the presentation has been restricted to Cartesian coordinates, inclusion of other coordinate systems (for example, cylindrical and spherical) may be carried out in the usual manner. In general, time-independent equations involving the Laplacian operator may be put in the form of Helmholtz's differential equation, $\nabla^2 u + k^2 u = 0$, when the appropriate k is used. Hence, solutions of Helmholtz's equation in various coordinate systems apply to all problems involving the Laplacian operator. In spherical coordinates (r,θ,ϕ), use of separation of variables, the power-series method, and the appropriate k for Helmholtz's equation lead to the following special functions: spherical harmonics, associated Legendre polynomials and Legendre polynomials, associated Laguerre polynomials and Laguerre polynomials, and spherical Bessel functions. Bessel functions result when cylindrical coordinates (ρ,ϕ,z) are used in Helmholtz's differential equation.

3.6 Fourier Transform Method

Methods of integral transforms are treated in detail elsewhere in the *Encyclopedia* (see FOURIER AND OTHER MATHEMATICAL TRANSFORMS). This section is devoted to the technique for solving differential equations (ordinary and partial) by use of the Fourier transform method. The one-dimensional Fourier transform pairs in symmetrical notation are given by

$$\begin{aligned} F(k) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)e^{ikx}dx, \\ f(x) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} F(k)e^{-ikx}dk. \end{aligned} \tag{129}$$

In Eq. (129), $F(k)$ is referred to as the Fourier transform of $f(x)$, and $f(x)$ is the inverse transform of $F(k)$. The convention for quantum-mechanical problems involves a sign change in the exponents. Relations in Eq. (129) may be extended to multiple dimensions in a natural manner. The basic idea of the Fourier transform method for solving differential equations (ordinary or partial) is to transform the original equation (ordinary or partial) into a simpler equation (algebraic or ordinary differential) that can be easily solved. The required solution of the original differential equation is then obtained by finding the inverse transform of the solution of the simpler equation which is in transform space.

EXAMPLE 3.1: By use of the Fourier transform method, solve the ordinary differential equation, $\ddot{x} + 2\alpha\dot{x} + w_0^2 x = f(t)$ subject to conditions that $x(t)$ and $\dot{x}(t)$ go to zero as t goes to plus and minus infinity.

SOLUTION: On taking the Fourier transform of each term in the original differential, the result is

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \ddot{x}(t)e^{i\omega t}dt + \frac{2\alpha}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \dot{x}(t)e^{i\omega t}dt + \frac{\omega_0^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x(t)e^{i\omega t}dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)e^{i\omega t}dt. \tag{130}$$

By use of partial integration and the conditions that $x(t)$ and $\dot{x}(t)$ approach zero as t approaches plus and minus infinity, Eq. (130) reduces to the algebraic equation

$$-\omega^2 X(\omega) - 2\alpha i\omega X(\omega) + \omega_0^2\, X(\omega) = F(\omega). \tag{131}$$

On solving the algebraic equation in Eq. (131) for $X(\omega)$ and inverting the transform, the solution $x(t)$ is obtained:

$$x(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{F(\omega)e^{-i\omega t}d\omega}{\omega_0^2 - \omega^2 - 2i\alpha\omega}. \tag{132}$$

The integral in Eq. (132) can be evaluated by use of the methods of calculus of residues when $f(t)$ is known.

EXAMPLE 3.2: By use of the Fourier transform method, solve the one-dimensional heat conduction equation for the temperature distribution $T(x,t)$ such that $T(x,t)$ and $T_x(x,t)$ approach zero as x approaches plus and minus infinity and $T(x,0)=T_0\exp(-ax^2)$ for constant a.

SOLUTION: Here, one transforms out the space variable so that the resulting equation will be a first-order ordinary differential equation in t. On taking the Fourier transform of each term in the one-dimensional heat conduction equation, one obtains

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{\partial^2 T}{\partial x^2}e^{ikx}dx$$

$$=\frac{1}{\sigma}\frac{1}{\sqrt{2\pi}}\frac{\partial}{\partial t}\int_{-\infty}^{\infty}T(x,t)e^{ikx}dx. \qquad (133)$$

By use of partial integration and the conditions that $T(x,t)$ and $T_x(x,t)$ approach zero as x approaches plus and minus infinity, Eq. (133) reduces to

$$\frac{\partial T(k,t)}{\partial t}+\sigma k^2 T(k,t)=0. \qquad (134)$$

The solution of Eq. (134) is

$$T(k,t)=A(k)e^{-\sigma k^2 t}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}T(x,t)e^{ikx}dx. \qquad (135)$$

Substituting the condition $T(x,0)=T_0\times\exp(-ax^2)$ into Eq. (135) yields

$$A(k)=\frac{T_0}{\sqrt{2a}}\exp\left(-\frac{k^2}{4a}\right). \qquad (136)$$

The solution in transform space (k space) is therefore

$$T(k,t)=\frac{T_0}{\sqrt{2a}}\exp\left(-\frac{1+4\sigma at}{4a}\right)k^2. \qquad (137)$$

The solution in x space is obtained when the k-space solution in Eq. (137) is inverted; the result is

$$T(x,t)=\frac{T_0}{\sqrt{1+4\sigma at}}\exp\left(\frac{-ax^2}{1+4\sigma at}\right). \qquad (138)$$

The convolution theorem,

$$\int_{-\infty}^{\infty}f(x-\xi)g(\xi)d\xi$$

$$=\int_{-\infty}^{\infty}F(k)G(k)\exp(-ikx)dk,$$

may be used to find the inverse transform when the solution in transform space is the product of two functions $F(k)G(k)$.

3.7 Green's Functions in Potential Theory

Here the three-dimensional Fourier transform method will be used to solve Poisson's equation for the electric potential $\phi(\mathbf{r})$ due to a volume charge density $\rho(\mathbf{r})$, and the three-dimensional Green's function will be defined. Poisson's equation is written as

$$\nabla^2\phi(\mathbf{r})=-\rho(\mathbf{r})/\epsilon_0\,. \qquad (139)$$

The quantity ϵ_0 is the permittivity of free space. The Fourier transform of $\phi(\mathbf{r})$ has the form

$$\Phi(\mathbf{k})=\frac{1}{(2\pi)^{2/3}}\int_{-\infty}^{\infty}\phi(\mathbf{r})\exp(i\mathbf{k}\cdot\mathbf{r})d^3r. \qquad (140)$$

A shorthand notation for triple integral, $d^3r=dx\,dy\,dz$, is used in Eq. (140). On taking the Fourier transform of both sides of Eq. (139), the solution in transform space [subject to the conditions that $\phi(\mathbf{r})$ and $\partial\phi/\partial r$ approach zero as r approaches plus and minus infinity] becomes

$$\Phi(\mathbf{k})=\rho(\mathbf{k})/k^2\epsilon_0. \qquad (141)$$

The inverse transform of $\Phi(\mathbf{k})$ yields the solution $\phi(\mathbf{r})$, and the result is

$$\begin{aligned}\phi(\mathbf{r})&=\frac{1}{(2\pi)^{3/2}}\int_{-\infty}^{\infty}\frac{\rho(\mathbf{k})}{k^2\epsilon_0}\exp(-i\mathbf{k}\cdot\mathbf{r})d^3r\\&=\frac{1}{(2\pi)^3}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\rho(\mathbf{r}')}{k^2\epsilon_0}\\&\quad\times\exp[i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')]d^3k\,d^3r'\\&=\frac{1}{(2\pi)^2\epsilon_0}\int_{-\infty}^{\infty}\rho(\mathbf{r}')G(\mathbf{r},\mathbf{r}')d^3r'. \qquad (142)\end{aligned}$$

The function $G(\mathbf{r},\mathbf{r}')$, Green's function for the operator ∇^2, is given by

$$G(\mathbf{r},\mathbf{r}')=\int_{-\infty}^{\infty}\frac{\exp[-i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')]d^3k}{k^2}.$$

When spherical polar coordinates are chosen where $d^3k=-\,k^2d(\cos\theta)d\phi\,dk$ and $\mathbf{r}-\mathbf{r}'$ is assumed to be along the polar axis, the expression for the Green's function reduces to $G(\mathbf{r},\mathbf{r}')=2\pi^2/|\mathbf{r}-\mathbf{r}'|$.

Physically, the Green's function $G(\mathbf{r},\mathbf{r}')$ is the electric potential at point $\mathbf{r}$ due to a point charge located at $\mathbf{r}'$. For a volume charge density $\rho(\mathbf{r}')$, the potential at $\mathbf{r}$ is given by $\int\rho(\mathbf{r}')G(\mathbf{r},\mathbf{r}')d^3r'$. In differential equation form, this analysis may be written as $\nabla^2G(\mathbf{r},\mathbf{r}')=-4\pi\delta(\mathbf{r}-\mathbf{r}')$ subject to appropriate boundary conditions for $G(\mathbf{r},\mathbf{r}')$. The Dirac delta function $\delta(\mathbf{r}-\mathbf{r}')$ means

$$\delta(x-x')\delta(y-y')\delta(z-z')$$

with properties $\delta(\mathbf{r}-\mathbf{r}')=0$ for $\mathbf{r}-\mathbf{r}'\neq0$ and $\int_{-\infty}^{\infty}\delta(\mathbf{r}-\mathbf{r}')d^3r'=1$. For Dirichlet boundary conditions, $G(\mathbf{r},\mathbf{r}')=0$ on the boundary surface enclosing the charge distribution $\rho(\mathbf{r}')$. It can be shown that the Neumann problem requires appropriate nonzero values for the normal derivative of the Green's function on the boundary surface. Use of the Green's-function method simplifies the problem of applying boundary conditions.

3.8 Numerical Methods for Partial Differential Equations

Numerical methods in partial differential equations form a vast subject and are treated in detail elsewhere in this *Encyclopedia* (see NUMERICAL METHODS). Here the focus is on essential concepts involved in converting a partial differential equation to its corresponding difference equation by use of finite difference methods. One should consult the references for a detailed discussion of the various special techniques for finding numerical solutions, convergence of solutions, and stability of the various methods.

3.8.1 Fundamental Relations in Finite Differences.

First differences Δ_xu and Δ_yu for positive h and k are defined by

$$\Delta_xu=\frac{u(x+h,y)-u(x,y)}{h},$$

$$\Delta_yu=\frac{u(x,y+k)-u(x,y)}{k}.$$

The corresponding second differences are defined by

$$\Delta_{xx}u=\frac{u(x+h,y)-2u(x,y)+u(x-h,y)}{h^2}$$

and

$$\Delta_{yy}u=\frac{u(x,y+k)-2u(x,y)+u(x,y-k)}{k^2}.$$

For notational convenience, Δx and Δy are replaced with h and k, respectively, in the above finite difference equations, and k replaces Δt in Secs. 3.8.3 and 3.8.4.

3.8.2 Two-Dimensional Laplace Equation: Elliptic Equation.

The two-dimensional Laplace equation in terms of finite differences reduces to

$$u(x,y)=\tfrac{1}{4}[u(x+h,y)+u(x-h,y)+u(x,y+h)+u(x,y-h)].$$

The computational procedure involves replacing $u(x,y)$, for example, a potential, at a particular grid point (see Fig. 13) by the average value of its four closest neighbors. The function $u(x,y)$ or its derivative must be specified at all points surrounding a given region.

3.8.3 One-Dimensional Heat Conduction Equation: Parabolic Equation.

In terms of finite differences, the one-dimensional heat conduction equation reduces to

$$u(x,t+k)=\frac{\sigma k}{h^2}[u(x+h,t)-2u(x,t)+u(x-h,t)]+u(x,t). \qquad (143)$$

The numerical solution involves determining the initial values of $u(x,t)$ at various x locations (see Fig. 14 for the space-time grid) and applying Eq. (143) to obtain the $u(x,t)$ at other times.

3.8.4 One-Dimensional Wave Equation: Hyperbolic Equation.

The finite-difference representation of the one-dimensional wave equation reduces to

$$u(x,t+k)=\frac{k^2v^2}{h^2}[u(x+h,t)-2u(x,t)+u(x-h,t)]+2u(x,t)-u(x,t-k). \qquad (144)$$

The starting value for $u(x,t+k)$ is determined from the initial conditions (see Fig. 15), and remaining values are determined by use of Eq. (144).

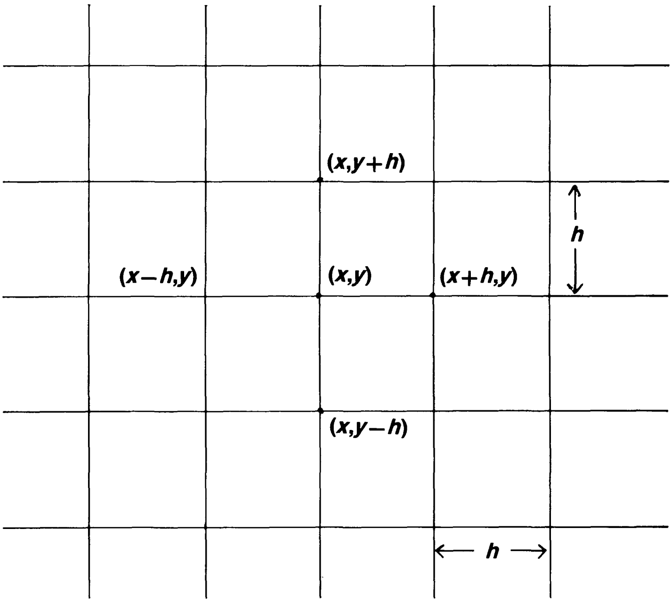

FIG. 13. Grid representation for Laplace's equation.

4. INTEGRAL EQUATIONS

4.1 Introduction

This section is devoted to a discussion of solutions and applications of one-dimensional linear integral equations of the first and second kinds. The formulations of many problems in physics lead to either differential or integral equations. Certain problems can only be represented by integral equations of the general form

$$u(x)=f(x)+\lambda\int_a^b k(x,s)u(s)ds. \tag{145}$$

Equation (145) is an integral equation since the unknown function $u(x)$ appears in the integrand. Functions $f(x)$ and $k(x,s)$ are to be given, and λ is a known parameter used here for convenience. The function $f(x)$ is called

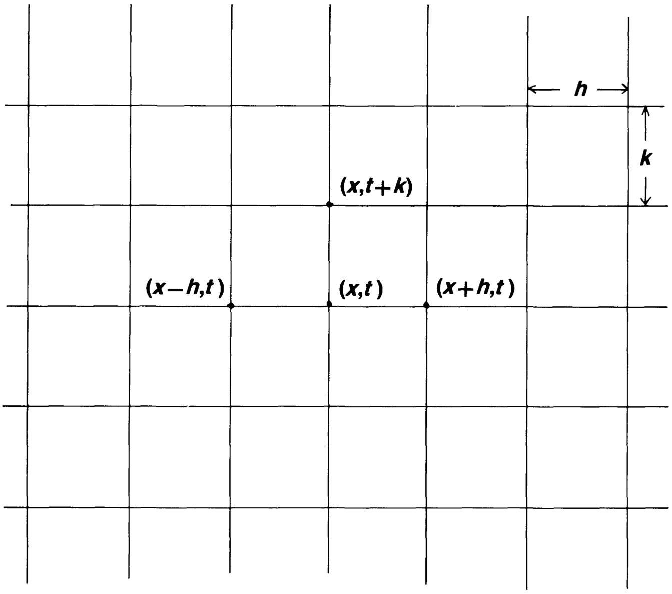

FIG. 14. Space-time grid for the heat equation.

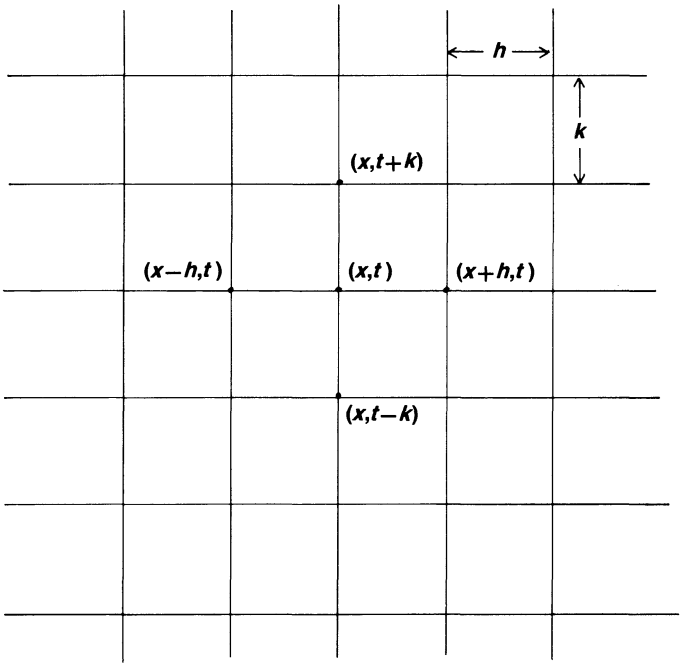

FIG. 15. Space-time grid for the wave equation.

the free term, and $k(x,s)$ is referred to as the kernel (nucleus). Quantities $f(x)$, $k(x,s)$, and λ may be either real or complex but are considered real in this section. Equation (145) is a linear integral equation since u is linear. An integral equation is singular if either (or both) of the limits of integration is infinite and/or if the kernel becomes infinite in the range of integration. When $f(x)$ equals zero, Eq. (145) is classified as a homogeneous integral equation. If the kernel is continuous in the closed region $[a,b]$, then Eq. (145) is classified as a Fredholm-type integral equation of the second kind. The equation (where the upper limit is a variable)

$$u(x)=f(x)+\lambda\int_a^x k(x,s)u(s)ds \qquad (146)$$

is known as a Volterra-type integral equation of the second kind. Fredholm integral equations of the first kind have the form

$$f(x)=\int_a^b k(x,s)u(s)ds. \qquad (147)$$

Volterra-type integral equations of the first kind have the form

$$f(x)=\int_a^x k(x,s)u(s)ds. \qquad (148)$$

In summary, classifications are Fredholm type if the limits of integration are fixed and Volterra type if one limit is variable, and first kind if the unknown function appears only in the integrand and second kind if the unknown function appears both in the integrand and outside the integrand.

Physical problems may be formulated as differential equations with appropriate boundary and/or initial conditions, integral equations, or either differential or integral equations. An essential difference in the formulation is that boundary conditions are imposed on general solutions of differential equations while boundary conditions are incorporated within the formulation of integral equations. While there exist procedures of converting differential equations to integral equations, use of integral equations seems more appropriate when formulations of problems lead directly to integral equations, or when solutions of the corresponding integral equations are easier to obtain than those for the corresponding differential equations. Laplace and Fourier transforms as well as dispersion relations are examples of singular integral equations of the first kind.

It is important to note that certain problems in classical mechanics, transport and diffusion phenomena, scattering theory, and other areas of physics can be formulated only by use of integral equations; the number of such problems is very small when compared to those leading to differential equations. In general, the theory of solution techniques needed in solving integral equations is not as familiar to physicists as techniques for solving differential equations. Integral equations are seldom treated in detail in introductory mathematical physics textbooks but are, however, discussed in advanced books in theoretical physics and mathematical physics. See Further Reading for some excellent books on integral equations. Many integral equations encountered in physics are normally solved by use of intuitive analytical methods, intuitive approximation methods and numerical techniques, or Laplace or Fourier transform methods.

Some systematic methods for solving nonsingular and linear integral equations are transform theory, Neumann series, separable kernel, Schmidt–Hilbert theory, Wiener–Hopf theory, and numerical. The Wiener–Hopf method is a different type of transform method which may be applied to certain integral equations with displacement kernels, $k(x,s)=k(x-s)$. Schmidt–Hilbert theory is an approach that applies to integral equations with Hermitian kernels, $k(x,s)=k^*(s,x)$. Fredholm theory involves representing the kernel as an infinite series of degenerate kernels (Sec. 4.2) and reducing the integral equation to a set of algebraic equations. Numerical solutions of Volterra equations involve reducing the original equations to linear algebraic equations, successive approximations and numerical evaluation of integrals. Numerical techniques for Fredholm equations involve solving a system of simultaneous equations.

4.2 Integral Equations with Degenerate Kernels

A subset of Fredholm equations of the first and second kinds with degenerate (separable) kernels can be solved by reducing them to a system of algebraic equations. In general, degenerate kernels may be written as

$$K(x,s)=\sum_{j=1}^{N} g_j(x)\phi_j(s). \tag{149}$$

In Eq. (149), it is assumed that $g_j(x)$ and $\phi_j(s)$ are linearly independent quantities, respectively. Substituting Eq. (149) into Eq. (145) yields

$$u(x)=f(x)+\lambda\sum_{j=1}^{N} g_j(x)C_j. \tag{150}$$

The coefficients C_j are given by

$$C_j=\int_a^b \phi_j(s)u(s)ds. \tag{151}$$

The solution of Eq. (145) has now been reduced to finding the C_j from the indicated algebraic equations and substituting the C_j into Eq. (150).

EXAMPLE 4.1: By use of the degenerate kernel method, find the solution of $u(x)=x+\lambda\int_0^1 xsu(s)ds$. The integral equation becomes

$$u(x)=x+\lambda x\int_0^1 su(s)ds=x+\lambda xC. \tag{152}$$

The coefficient C reduces to

$$C=\int_0^1 su(s)ds=\int_0^1 s(s+\lambda sC)ds=\frac{1}{3-\lambda}. \tag{153}$$

The second step in Eq. (153) results when the second step of Eq. (152) is substituted into the first step of Eq. (153). From Eqs. (152) and (153), the solution of the original equation is $u(x)=3x/(3-\lambda)$. It is seen that solutions exist for values of λ different from 3.

EXAMPLE 4.2: By use of the degenerate kernel method, find the solution of $u(x) = x + \frac{1}{2}\int_{-1}^{1}(s + x)ds$. The equation becomes

$$u(x)=x+\frac{C_1}{2}+\frac{xC_2}{2}. \tag{154}$$

The coefficients C_1 and C_2 are

$$C_1=\int_{-1}^{1} su(s)ds=\frac{2-C_2}{3} \quad\text{and}$$

$$C_2=\int_{-1}^{1} u(s)ds=C_1 \quad\text{or}\quad C_1=C_2=1. \tag{155}$$

On substituting the values for C_1 and C_2 into Eq. (154), the solution of the original equation becomes $u(x)=(3x+1)/2$.

4.3 Integral Equations with Displacement Kernels

If the kernel is of the form $k(x-s)$, it is referred to as a displacement kernel. Fredholm equations of the first and second kinds with displacement kernels and limits from minus infinity to plus infinity or from zero to plus infinity can normally be solved by use of Fourier and Laplace transform methods, respectively. Here the Fourier transform approach for solving integral equations with displacement kernels will be illustrated. Taking the Fourier transform of each term in Eq. (145) yields

$$\int_{-\infty}^{\infty} u(x)\exp(ikx)dx = \int_{-\infty}^{\infty} f(x)\exp(ikx)dx + \lambda\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} K(x-s)u(s)ds\right) \times\exp(ikx)dx. \tag{156}$$

In transform space, Eq. (156) is $u(k)=F(k)+\lambda K(k)u(k)$. The solution in x space is obtained when the inverse transform of $u(k)$ is taken, and the result becomes

$$u(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{F(k)\exp(-ikx)dk}{1-\lambda K(k)}. \tag{157}$$

4.4 The Neumann Series Method

Consider the set of Fredholm equations of the second kind such that

$$\int_a^b |f(x)|^2dx \quad \text{and} \quad \int_a^b |K(x,s)|^2ds$$

are bounded. Assume the solution may be written as a power series, Neumann series (also known as Liouville–Neumann series), with form

$$u(x)=\sum_{n=0}^{\infty}\lambda^n u_n(x). \tag{158}$$

Terms in the successive approximation are obtained by substituting the Neumann series into the Fredholm equation, Eq. (145), and equating coefficients of like powers of λ; the results are

$$u_0=f(x);$$

$$u_1(x)=\int_a^b K(x,s)u_0ds;\ \ldots;$$

$$u_n(x)=\int_a^b K(x,s)u_{n-1}(s)ds. \tag{159}$$

It can be shown that the Neumann series converges for all values of λ for Volterra equations and converges for small values of λ for Fredholm equations; techniques exist for improving the convergence in the Fredholm case. Numerical techniques may be used to evaluate terms in the Neumann series.

4.5 The Abel Problem

The section on integral equations is concluded with the earliest application of integral equations to a physical problem, Abel's problem. The Abel problem is as follows: Consider a bead sliding on a smooth wire under the influence of gravity and find the curve for which the time of descent is a given function of the initial position.

Let the starting position of the bead be (x_0, y_0) and position of the bead at time t be (x,y) such that y equals zero at the end of the fall. The speed of the bead at (x,y) for ds an element of arc length along the path is determined from the conservation of energy principle and is given by

$$\frac{ds}{dt}=\sqrt{2g(y_0-y)}.$$

If the shape of the curve is $u(y)$, then $ds=u(y)dy$ and the time of descent is given by

$$T=\int_0^{y_0}\frac{u(y)dy}{\sqrt{2g(y_0-y)}}.$$

The Abel problem is to find the curve $u(y)$ for which the time T of descent is a given function $f(y_0)$ of the initial vertical position, and the result is obtained from the integral equation (Abel's equation):

$$f(y_0)=\int_0^{y_0}\frac{u(y)dy}{\sqrt{2g(y_0-y)}}. \tag{160}$$

It can be shown that the curve in question is a portion of a cycloid.

5. APPLIED FUNCTIONAL ANALYSIS

5.1 Introduction

Concepts of functions (of one variable) and operators were introduced into mathematics in connection with the development of calculus during the latter part of the seventeenth century. In general, an operator applied to a given function yields a new function. The problem of finding an extremum (maximum or minimum) of a function is carried out in the usual manner by use of ordinary calculus, but the general problem of finding the stationary value (an extremum) of certain definite integrals that occur in mathematical physics is the subject matter of the branch of mathematics called the calculus of variations.

In relation to the calculus of variations, the process of connecting (mapping) each function $y(x)$ in $[a,b]$ with a number represented by the definite integral $\int_a^b F(y,y',x)dx$ (where $y'=dy/dx$) which depends on $y(x)$ was given the name functional during the end of the nineteenth century. The basic idea of functional analysis is that problems are often easier to solve if a function is considered to be a member of a whole space of functions, X. The space X is assumed to carry a metric, have a linear space structure, and be infinite dimensional. The concept of a metric involves topological and geometrical language while linear operators on X involve concepts of linear algebra, and relations among these concepts constitute linear functional analysis.

A function which depends on one or more functions rather than on discrete variables is referred to as a functional. The domain of a functional is a space of admissible functions. More precisely, functionals are continuous linear maps, from a normed space into itself or into some other normed space. The basic ingredient of the various definitions of a functional and of functional analysis is the existence of a linear space with a topology.

The main topics in Secs. 1–4 (functions of a complex variable and analytic functions, ordinary and partial differential equations, Fourier series and Fourier transform theory, and integral equations) are technically topics in functional analysis even though the topology and geometry of the linear spaces involved were not stressed. Mathematically, a valid argument can be made that concluding this article with a discussion of functional analysis is analogous to putting the cart before the horse. This argument, however, neglects the applications-of-techniques approach emphasized throughout the article.

In mathematical physics, functional analysis often involves discussions connected with the calculus of variations; theory of ordinary and partial differential equations; integral equations and transform theory; spectral theory involving eigenvalues, eigenfunctions, and Fourier series expansion theory involving orthogonal functions; functional calculus used in the path integral formulation of quantum mechanics, quantum field theory, and statistical mechanics; C^* algebra; and the theory of distributions. In mathematics, functional analysis often involves the general theory of linear normed spaces, the topological structure of linear spaces and continuous transformations, measure spaces and general theories of integration, spectral theories, C^* algebra, distribution theory, and number theory.

In this section, the original problem of functional analysis (the calculus of variations) and applications of functional integration to quantum mechanics, quantum field theory, and statistical mechanics will be discussed.

5.2 Stationary Values of Certain Definite Integrals

Consider the following definite integral of the functional $F(y,y',x)$ where F is a known function of y, y' (where $y'=dy/dx$), and x, but $y(x)$ is unknown:

$$J=\int_{x_1}^{x_2} F(y,y',x)dx. \tag{161}$$

A fundamental problem in the calculus of variations (a problem which occurs frequently in mathematical physics) is that of finding a function $y(x)$ such that the functional J is stationary (an extremum; a minimum in most cases of physical interest). The basic procedure here is to evaluate the integral for a slightly modified path $y(x,\alpha)=y(x,0)+\alpha\eta(x)$ where $\eta(x_1)=\eta(x_2)=0$ (all paths pass through the end points) and show that the change in the value of the integral due to the change in the path becomes zero. The function $\eta(x)$ is an arbitrary differentiable function, and α is a small scale factor (see Fig. 16). The function $y(x,\alpha)$ describes neighboring paths where $\delta y = y(x,\alpha) - y(x,0) = \alpha\eta(x)$ is the variation (hence the name calculus of variations) of $y(x,0)$ at some x. The delta symbol, δ, was introduced by Lagrange to denote a variation (a virtual change) and means a change made in an arbitrary manner. Both dy and δy denote infinitesimal changes in y, but dy means an infinitesimal change in $y(x)$ produced by dx while δy is an infinitesimal change which produces $y+\delta y$. It is straightforward to show that $d\delta y/dx = \delta\, dy/dx$ and $\delta\int_{x_1}^{x_2} F(y,y',x)dx = \int_{x_1}^{x_2} \delta F(y,y',x)dx$. On substituting $y(x,\alpha) = y(x,0) + \alpha\eta(x)$ into Eq. (161) and differentiating both sides of the resulting equation with respect to α, one obtains

$$\frac{dJ(\alpha)}{d\alpha} = \int_{x_1}^{x_2} \left(\frac{\partial F}{\partial y}\eta(x) + \frac{\partial F}{\partial y'}\eta'(x) \right) dx. \tag{162}$$

Integrating the second term in Eq. (162) by parts and using the fact that $\eta(x)$ is arbitrary yield

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'} \right) = 0. \tag{163}$$

Equation (163) is known as Euler's equation and its solution yields the $y(x)$ which makes J an extremum (minimum). An alternative and often used approach for obtaining Euler's equation makes use of expanding the right-hand side of δF in a Taylor's series with two variables; the result becomes

$$\delta F = F(y+\alpha\eta, y'+\alpha\eta', x) - F(y,y',x)$$
$$= \alpha\left(\frac{\partial F}{\partial y}\eta + \frac{\partial F}{\partial y'}\eta' \right). \tag{164}$$

Higher-order terms in the Taylor expansion may be used to determine the nature of the extremum (maximum or minimum), and neglected here since α is a small parameter. As a result of substituting Eq. (164) into the integrand for δJ, integrating the second term by parts as before, and setting $\delta J/\alpha = 0$, one obtains the Euler equation in Eq. (163).

The above processes of obtaining Euler's equation, Eq. (163), may be extended to functionals involving several dependent and/or independent variables; for example, the variational process applied to $F(y_1,\ldots,y_n, y_1', \ldots, y_n', x)$ yields the following set of Euler's equations:

$$\frac{\partial F}{\partial y_k} - \frac{d}{dx}\left(\frac{\partial F}{\partial y_k'} \right) = 0 \quad (k=1,2,\ldots,n). \tag{165}$$

EXAMPLE 4.3: By use of the variational calculus method (Euler's equation), determine the equation of the shortest path between two points (x_1,y_1) and (x_2,y_2) in Cartesian coordinates.

SOLUTION: The element of distance along the path between the two points is given by

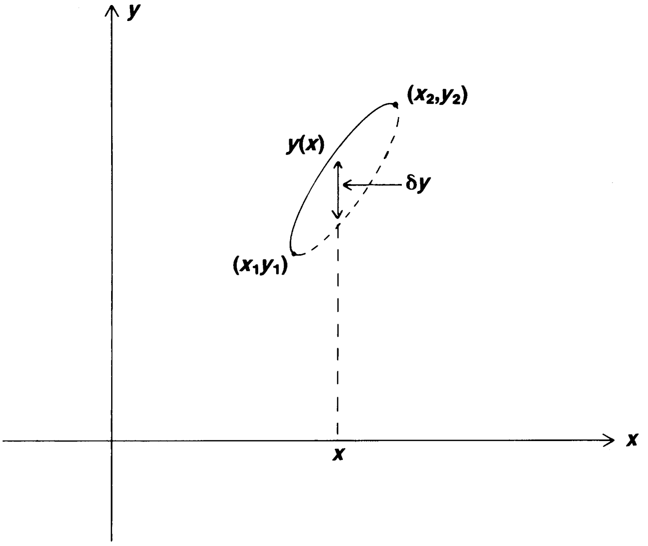

FIG. 16. A varied path between (X_1,Y_1) and (X_2,Y_2).

1. $ds=\sqrt{dx^2+dy^2}$.

The expression for the distance between the two points is therefore

2. $s=\int_{x_1}^{x_2}\sqrt{1+(y')^2}dx$, where $y'=\frac{dy}{dx}$.

For $F(y,y',x)=\sqrt{1+(y')^2}$, the differential equation for the equation of the shortest path between the two points, Euler's equation, reduces to

3. $\frac{dy}{dx}=A$ since $\frac{\partial F}{\partial y}=0$ and

$$\frac{\partial F}{\partial y'}=\frac{y'}{\sqrt{1+(y')^2}}.$$

The equation of the shortest path between the two points is therefore that of a straight line, $y(x)=Ax+B$.

EXAMPLE 4.4: (THE BRACHISTOCHRONE PROBLEM). The brachistochrone (shortest time) problem, first formulated and solved by Johann Bernoulli in 1696, is one of the first variational problems. The problem is as follows: Consider a bead of mass m which slides, under the influence of gravity, down a frictionless wire bent into the appropriate shape. The goal is to find the equation (shape of the wire) of the path along which the bead travels so that the time is a minimum.

SOLUTION: For convenience, it is assumed that the bead starts from rest at the origin of a coordinate system (see Fig. 17). Since this is a conservative system, the following relations are valid: $T_1+V_1=T_2+V_2$, $V_2=-mgy$, $T_1=V_1=0$, $T_2=\frac{1}{2}mv^2$, and $v=\sqrt{2gy}$. The expression for the time required for the bead to travel from the origin to point (x,y) is therefore given by

1. $t=\int\frac{\sqrt{dx^2+dy^2}}{\sqrt{2gy}}$

$$=\int_0^{y_2}\frac{\sqrt{1+(x')^2}dy}{\sqrt{2gy}},\quad x'=\frac{dx}{dy}.$$

The unknown function $y(x)$ must be determined such that the time is a minimum. On applying Euler's equation with

$$F=\left(\frac{1+(x')^2}{y}\right)^{1/2}$$

and independent variable y, one obtains

2. $x=\int\frac{Ay\,dy}{\sqrt{y-A^2y^2}}$ since

$$\frac{\partial F}{\partial x}=0\quad\text{and}\quad\frac{\partial F}{\partial x'}=\frac{x'}{\sqrt{y[1+(x')^2]}}=A.$$

On letting $A=1/\sqrt{2a}$ and making the change of variable $y=a(1-\cos\theta)$, the above integral reduces to $x=a(\theta-\sin\theta)$

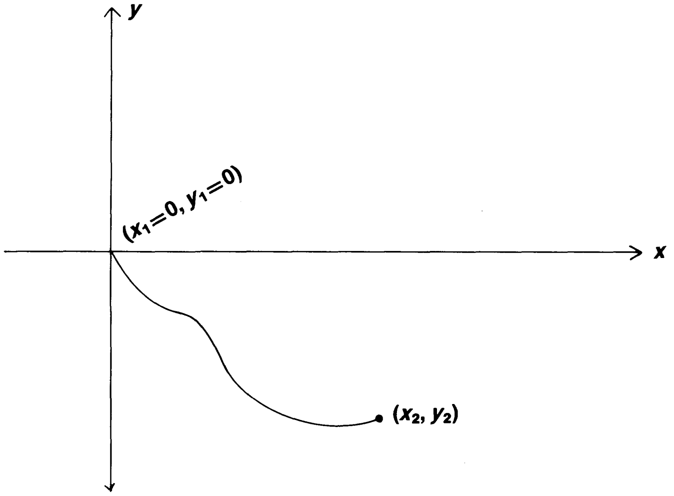

FIG. 17. Diagram for the brachistochrone example.

+const. The path that yields a minimum time of travel is in the form of parametric equations $x=a(\theta-\sin\theta)$ and $y=a(1-\cos\theta)$, equations for a cycloid that passes through the origin.

5.3 Hamilton's Variational Principle in Mechanics

5.3.1 Introduction. Mechanics is the study of the motions (including rest) of physical objects. The laws of classical mechanics are valid for macroscopic objects (size larger than 10^{-10} m), and the laws of quantum mechanics are valid in the microworld (object size smaller than 10^{-10} m). In this section, the focus is on the study of classical mechanics. Widely used equivalent formulations of classical mechanics are Newtonian mechanics (1687), Lagrangian mechanics (1788), Hamiltonian mechanics (1834), and Hamilton–Jacobi theory (1837).

Formulations of classical mechanics developed since Newtonian mechanics are generalizations and equivalent representations of Newtonian mechanics. These generalizations do not lead to new information but offer different ways of approaching problems. Certain problems can be solved by use of all four approaches with equal amounts of ease (or difficulty). Other problems are more amenable to solution by use of one approach than by use of the others. The specific nature of the problem to be solved usually dictates the approach that should be used.

Newton's second law is the basic equation of motion in the Newtonian picture of mechanics. In Lagrangian mechanics, Lagrange's equations are the required set of equations of motion for the system (particle or group of particles) under investigation. Hamilton's canonical equations are basic to Hamiltonian mechanics, and the Hamilton–Jacobi equation is the foundation of the Hamilton–Jacobi theory.

The approach in this section begins with Hamilton's variational principle for conservative systems (where the forces acting on the system may be derived from a potential function) from which Lagrange's equations will be developed by use of the variational calculus method. By use of a Legendre transformation, the Hamiltonian and subsequently Hamilton's canonical equations are obtained.

The variational technique used in mechanics was developed mainly by Euler and Lagrange and is a mathematical formulation of mechanics in which kinetic energy and potential energy play an essential role. In Newtonian mechanics, forces play the central role.

5.3.2 Generalized Coordinates. Linearly independent quantities $\{q_k\}=q_1,\ldots,q_k$ that completely define the position (configuration) of a system as a function of time are called generalized coordinates. Quantities $\{q_k\}$ are said to be linearly independent if $\sum_k \alpha_k q_k=0$ implies that $\alpha_k=0$ for all k. Generalized coordinates may be selected to match the conditions of the problem to be solved. The number of generalized coordinates that must be used to define uniquely the position of a system represents the number of degrees of freedom for the system. The corresponding quantities $\{\dot{q}_k\}$ are called generalized velocities.

The simultaneous specification of $\{q_k\}$ and $\{\dot{q}_k\}$ for a system determines the mechanical state of the system at that time, and subsequent motion is obtained from the solutions $q_k(t)$ of the appropriate equations of motion. The appropriate second-order differential equations expressing the relations among generalized coordinates q_k, generalized velocities $\dot{q}_k$, and generalized accelerations $\ddot{q}_k$ are called equations of motion for the system under investigation.

Although the set of generalized coordinates used to solve a problem is not unique, a proper set of generalized coordinates is that set which leads to an equation of motion whose solution has a straightforward physical interpretation. No general rule exists for obtaining a proper set of generalized coordinates.

5.3.3 Lagrange's Equations. Hamilton's variational principle asserts that the actual motion of a particle or system of particles (conservative system) from its initial configuration at time t_1 to its configuration at time t_2 is such that

$$\delta S=\delta\int_{t_1}^{t_2} L(q_k,\dot{q}_k)dt=0. \qquad (166)$$

In Eq. (166), $q_k=q_k(t)$, $L=T-V$ is defined as the Lagrangian for the system under investigation, $L\,dt$ is called the action, and

$$S=\int_{t_1}^{t_2} L\,dt$$

denotes the action integral. The quantities T and V are kinetic and potential energy, respectively.

Among the infinite number of trajectories $q(t)$ that connect the end points $q(t_1)$ and $q(t_2)$, the physical (actual) path yields a stationary value for the action integral. The action is therefore a functional of the functions $q_k(t)$ satisfying the boundary conditions that all trajectories pass through the end points. By use of the variational technique leading to Eq. (165), one finds that $q(t)$ is obtained from the following set of differential equations:

$$\frac{\partial L}{\partial q_k}-\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right)=0, \quad k=1,2,\ldots,n. \qquad (167)$$

The equations in Eq. (167) are called Lagrange's (or Euler–Lagrange) equations. Lagrange's equations, the equations of motion for the system under investigation, are a set of n second-order differential equations. The general solutions of these equations contain $2n$ arbitrary constants of integration. The values of these $2n$ arbitrary constants are determined when the initial state (initial values for the q_k and $\dot{q}_k$ at $t=0$) of the system is specified.

Quantities $\partial L/\partial \dot{q}_k$ and $\partial L/\partial q_k$ are defined to be canonical momenta (also called conjugate or generalized momenta) and generalized forces, respectively,

$$p_k=\frac{\partial L}{\partial \dot{q}_k} \quad \text{and} \quad F_k=\frac{\partial L}{\partial q_k}. \qquad (168)$$

By use of the definitions in Eq. (168), it is observed that Lagrange's equations may be considered a generalized version of Newton's second law where generalized force equals the rate of change of canonical momentum.

Basic conservation laws of mechanics result from invariance of the Lagrangian under time translation— conservation of energy, coordinate translation—conservation of canonical momentum, translation in space—conservation of total linear momentum, and rotation in space—conservation of angular momentum. In spite of the important role of the Lagrangian, it is not a unique function for a system since the equations of motion for the system, Lagrange's equations, are unchanged if $df(q_k,t)/dt$ is added to the Lagrangian.

5.3.4 Format for Solving Problems by Use of Lagrange's Equations. The following steps should be used when applying Lagrange's equations.

1. Draw a detailed diagram. Specify the degrees of freedom and the level where potential energy V equals zero.
2. Write down the appropriate expressions for T, V, and L.
3. Write down the specific set of Lagrange's equation(s).
4. Work out the terms in the set of equations in Step 3.
5. Solve the resulting equation(s) of motion subject to the given initial conditions.

5.4 Formulation of Hamiltonian Mechanics

It has been shown that Hamilton's variational principle combined with techniques of the calculus of variations transforms the process of finding the solution of a mechanical problem to that of obtaining solutions for Lagrange's equations. Hamilton developed a procedure for transforming Lagrange equations to a simpler (canonical) form by replacing them (a set of n second-order differential equations) with a set of $2n$ first-order differential equations now called Hamilton's canonical equations of motion.

5.4.1 Derivation of Hamilton's Canonical Equations. The Lagrangian is a function of q_k and $\dot{q}_k$; now the change of variable $\dot{q}_k \rightarrow p_k$ where $p_k=\partial L/\partial \dot{q}_k$ will be made. By use of a Legendre transformation [new function equals the old function minus (the derivative of the old function with respect to the old variable) times the old variable; the physical and geometrical content of the new and old functions is the same], one obtains

$$-H\equiv L-\sum_{k=1}^{n} p_k\dot{q}_k. \qquad (169)$$

The negative sign in Eq. (169) is by convention. The new function $H(q_k,p_k)$ contains the same geometrical and physical content as $L(q_k,\dot{q}_k)$ and is called the Hamiltonian of the system. Note that the action integral may now be written as

$$S=\int_{t_1}^{t_2}\left(\sum_{k=1}^{n} p_k\dot{q}_k-H\right)dt. \qquad (170)$$

Applying the variational techniques of Sec. 5.2 to Eq. (170) yields

$$\dot{q}_k=\frac{\partial H}{\partial p_k} \quad \text{and} \quad \dot{p}_k=-\frac{\partial H}{\partial q_k}. \tag{171}$$

The equations in Eq. (171) are referred to as Hamilton's canonical equations of motion (or simply Hamilton's equations). Hamilton's equations can be used to develop the specific set of equations of motion for the system under investigation in terms of the phase space variables q_k and p_k. Note that Lagrange's equations consist of n second-order differential equations whereas Hamilton's equations form a set of $2n$ first-order differential equations. For a conservative system, it can be shown that the Hamiltonian equals the total energy of the system ($H=T+V$).

5.4.2 Format for Solving Problems by Use of Hamilton's Equations. In solving problems by use of Hamiltonian mechanics, the following five-step procedure is highly recommended.

1. Write out the Lagrangian as in Lagrangian mechanics, $L=T-V$.
2. Solve the equation $p_k=\partial L/\partial \dot{q}_k$ for $\dot{q}_k$ and eliminate $\dot{q}_k$ from the Lagrangian.
3. Construct the Hamiltonian for the system, $H = \sum_{k=1}^{n} \dot{q}_k\, p_k - L$.
4. Obtain Hamilton's equations, $\dot{q}_k=-\partial H/\partial p_k$ and $\dot{p}_k=-\partial H/\partial q_k$.
5. Solve the $2n$ first-order differential equations (equations of motion) developed in step 4.

5.4.3 Poisson's Brackets. The total time derivative of a function $f(q_k, p_k)$ is

$$\begin{aligned}\frac{df}{dt}&=\sum_{k=1}^{n}\left(\frac{\partial f}{\partial q_k}\dot{q}_k+\frac{\partial f}{\partial p_k}\dot{p}_k\right)\\ &=\sum_{k=1}^{n}\left(\frac{\partial f}{\partial q_k}\frac{\partial H}{\partial p_k}-\frac{\partial f}{\partial p_k}\frac{\partial H}{\partial q_k}\right)=\{f,H\}.\end{aligned} \tag{172}$$

Hamilton's equations were used in obtaining Eq. (172). The last quantity in Eq. (172) is called a Poisson bracket. A Poisson bracket is defined by

$$\{f,g\}=\sum_{k=1}^{n}\left(\frac{\partial f}{\partial q_k}\frac{\partial g}{\partial p_k}-\frac{\partial f}{\partial p_k}\frac{\partial g}{\partial q_k}\right). \tag{173}$$

Hamilton's canonical equations in terms of Poisson brackets are given by

$$\begin{aligned}\dot{q}_k&=\frac{\partial H}{\partial p_k}=\{q_k,H\},\\ \dot{p}_k&=-\frac{\partial H}{\partial q_k}=\{p_k,H\}.\end{aligned} \tag{174}$$

Two variables ξ_i and ϕ_i are said to be canonically conjugate if

$$\{\xi_i,\xi_k\}=\{\phi_i,\phi_k\}=0 \quad \text{and} \quad \{\xi_i,\phi_k\}=\delta_{ik}. \tag{175}$$

The Kronecker delta function is defined by

$$\delta_{ik}=\begin{cases}1; & i=k,\\ 0; & i\neq k.\end{cases} \tag{176}$$

The quantities q_j and p_j are canonically conjugate variables since $\{q_j,p_k\}=\delta_{jk}$ and $\{q_j,q_k\}=\{p_j,p_k\}=0$; these three Poisson brackets are referred to as fundamental Poisson brackets.

5.5 Continuous Media and Fields

Thus far, only conservative systems composed of discrete particles have been considered. The Lagrangian of a system composed of N free particles may be written as

$$L=\sum_{i=1}^{N} L_i. \tag{177}$$

The extension of the above analysis to a system with an infinite number of degrees of freedom (a continuous medium) is achieved by replacing the subscript k with a continuous variable (say $\mathbf{x}$), q_k with a new function $q_k \rightarrow Q(\mathbf{x},t)$, the sum with an integral $\sum_i \rightarrow \int d^3x$, and canonical momenta with canonical momentum density given by $\pi(\mathbf{x})=\partial\mathscr{L}/\partial\dot{Q}$ where $\mathscr{L}$ is the Lagrangian density. The quantity $Q(\mathbf{x},t)$ is called a field. To denote several fields, the notation $Q_\alpha(\mathbf{x},t)$ may be used. The parameter α distinguishes among the different fields. From a mathematical point of view, a field is a set of functions of space-time, and these functions satisfy a set of partial differential equations. The corresponding Hamilton's variational principle is

$$0=\delta\int_{t_1}^{t_2}\sum_{i=1}^{N}L_i(q_k,\dot{q}_k)dt$$

$$=\delta\int_{t_1}^{t_2}\int_{\text{physical space}}\mathscr{L}\{Q_\alpha(\mathbf{x},t),$$

$$\dot{Q}_\alpha(\mathbf{x},t)\}d^4x;$$

$$d^4x=dx\,dy\,dz\,dt. \qquad (178)$$

Assuming that fields interact only with infinitesimal neighbors, the Lagrangian density should be a function of $Q_\alpha(\mathbf{x},t)$, $\dot{Q}_\alpha(\mathbf{x},t)$, and $\partial Q_\alpha(\mathbf{x},t)/\partial x_k$ or $Q_\alpha(x^\mu)$ and $\partial_\mu Q_\alpha$ in four-vector notation. By use of appropriate boundary conditions, the variation in Eq. (178) leads to the following set of equations of motion:

$$\frac{\partial\mathscr{L}}{\partial Q_\alpha}-\partial_\mu\left(\frac{\partial\mathscr{L}}{\partial(\partial_\mu Q_\alpha)}\right)=0,\quad \mu=0,1,2,3. \qquad (179)$$

The equations in Eq. (179) are the Lagrange's equations for classical fields.

5.6 Transitions to Quantum Mechanics

The laws of classical mechanics are not in general valid for the microworld, and new laws (quantum theory) that are appropriate for the microworld were developed during the period 1900–1927. In this section, the transition from classical mechanics to quantum mechanics in the Heisenberg picture, in the Schrödinger picture, and by use of the action functional (path integral) approach due to Dirac and Feynman will be made. For notational convenience, the discussion is restricted to the case of one nonrelativistic particle. The starting point in both the Heisenberg and Schrödinger pictures is Hamiltonian mechanics while the Feynman (Dirac–Feynman) approach begins with Lagrangian mechanics.

The postulates of quantum mechanics may be stated as follows.

1. Each state of a physical system corresponds to a normalized vector in Hilbert space called the state vector, Ψ or $|\Psi\rangle$.
2. Physical quantities are represented by linear Hermitian operators in Hilbert space.
3. If a system is in a state $|\Psi\rangle$, then the probability that a measurement (consistent with quantum theory) of the quantity corresponding to $\hat{A}$ will yield one of the eigenvalues a_k (where $\hat{A}|\Psi\rangle=a_k|\Psi\rangle$) is given by $|\langle a_k|\Psi\rangle|^2$. The system will change from state $|\Psi\rangle$ to $|a_k\rangle$ as a result of the measurement. The quantity $\langle a_k|\Psi\rangle$is the amplitude.

5.6.1 The Heisenberg Picture. In the Heisenberg approach, a system is quantized by letting q_k and p_k be Hermitian operators in a Hilbert space such that $q_k\to\hat{q}_k$ and $p_k\to-i\hbar\,\partial/\partial q_k$, and replacing Poisson brackets with commutators, $\{A,B\}\to[\hat{A},\hat{B}]/i\hbar$ where $[\hat{A},\hat{B}]=\hat{A}\hat{B}-\hat{B}\hat{A}$. If $[\hat{f},\hat{g}]=i\hbar$, the operators $\hat{f}$ and $\hat{g}$ are said to be canonically conjugate. The resulting Heisenberg equations of motion for a quantum mechanical system are

$$i\hbar\dot{p}_k=[p_k,H] \quad\text{and}\quad i\hbar\dot{q}_k=[q_k,H]. \qquad (180)$$

The equations in Eq. (180) are basic for Heisenberg (matrix) mechanics.

5.6.2 The Schrödinger Picture. From a classical mechanical point of view, the Hamiltonian of a particle subject to conservative forces equals the total energy of the particle, and one may write

$$H=E=\mathbf{p}^2/2m+V(x,y,z). \qquad (181)$$

The transition to quantum mechanics in the Schrödinger picture is achieved by use of the replacements $E\to i\hbar\,\partial/\partial t$ and $\mathbf{p}\to-i\hbar\nabla$; by use of these replacements, Eq. (182) is transformed into an operator equation. Operating on some function $\Psi(x,y,z,t)$ or $|\Psi\rangle$ in Hilbert space yields

$$i\hbar\frac{\partial\Psi}{\partial t}=-\frac{\hbar^2}{2m}\nabla^2\Psi+V\Psi \quad\text{or}\quad i\hbar\frac{\partial|\Psi\rangle}{\partial t}=\hat{H}|\Psi\rangle. \qquad (182)$$

Schrödinger's equation, Eq. (182), is the basic equation of motion of a particle in quantum mechanics in the Schrödinger picture.

5.6.3 The Feynman Path-Integral Approach to Quantum Mechanics. The special case of one particle with one degree of freedom is considered here to simplify the notation and make the explanations clear. Feynman's formulation of quantum mechanics was stimulated by some work of Dirac (1933) and is based on the following two postulates:

1. The amplitude $\langle q(t'')|q(t')\rangle$ for a particle to be found at $q(t'')$ at time t'' if its initial position is $q(t')$ at time t' equals a sum of complex contributions (amplitudes) for

each space-time path starting at $q(t')$ and ending at $q(t'')$.

2. All paths connecting $q(t')$ and $q(t'')$ contribute equally in magnitude, but the phase (weight) of their contribution is $\exp(iS/\hbar)$ where S is the classical action integral for the corresponding paths.

The measure on the functional space of paths $q(t)$ is denoted by $\mathscr{D}[q(t)]$, and appropriate normalization factors for the amplitude are contained in $\mathscr{D}[q(t)]$. Feynman's interpretation of the indicated functional integration is as follows: Divide the time interval $t''-t'$ into N equal parts, each with duration $\epsilon=t_{k+1}-t_k$; and in the limit $N\to\infty$ ($\epsilon\to 0$), it is assumed that the sequence of points $q(t_0),\ldots,q(t_n)$ approximates the path $q(t)$. The action functional associated with the classical path joining $q(t_k)=q_k$ and $q(t_{k+1}) = q_{k+1}$ is

$$S[q_{k+1},q_k]=\int_{t_k}^{t_{k+1}} L(q,\dot{q})dt.$$

Feynman's postulates thus assert that the amplitude $\langle q(t'')|q(t')\rangle$ is a sum of all amplitudes for all paths connecting $q(t'')$ and $q(t')$, and it may be written as (see Fig. 18)

$$\langle q(t'')|q(t')\rangle=\lim_{\substack{\epsilon\to 0\\ N\to\infty}}\int\cdots\int\prod_{k=0}^{N}\exp\left(\frac{i}{\hbar}S(q_{k+1},q_k)\right)\frac{dq_k}{A_k}=\int\cdots\int\exp\left(\frac{i}{\hbar}\int_{t'}^{t''}L(q,\dot{q})dt\right)\mathscr{D}[q(t)]. \tag{183}$$

The normalization factors A_k in Eq. (183) are independent of the path from q_k to q_{k+1} but depend on the mass of the particle and on the time interval ϵ. Equation (183) is a mathematical statement that the amplitude for a particle at $q(t')$ at time t' to move to $q(t'')$ at time t'' equals the sum of all possible paths between the two points times $\exp(iS/\hbar)$; the probability is the absolute square of the amplitude.

The path integral approach to quantum mechanics can be extended to include formulations of quantum field theory (a combination of quantum mechanics and special relativity), the partition function in statistical mechanics, and systems obeying Bose–Einstein and Fermi–Dirac statistics. The path integral method is the foundation for Feynman diagrams.

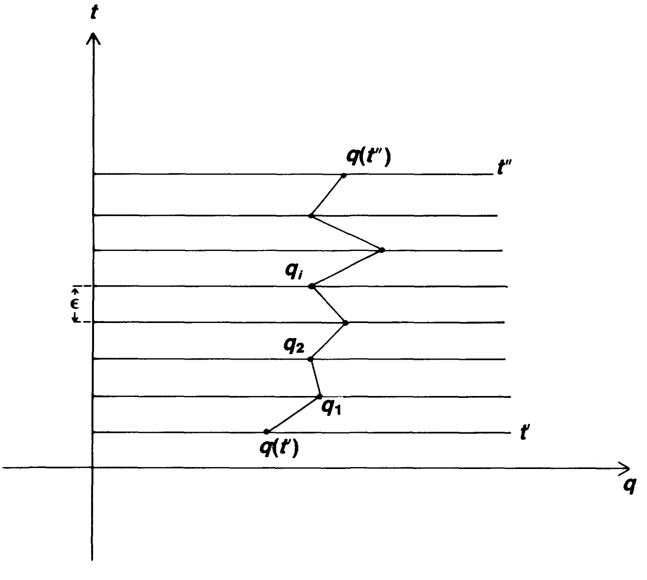

FIG. 18. A representative sequence of paths between $q(t')$ and $q(t'')$.

GLOSSARY

Complex Variable: An ordered pair of real variables ($z=x+iy$) with a real and an imaginary part.

Ordinary Differential Equation: An equation containing derivative(s) with respect to one independent variable.

Partial Differential Equation: An equation containing partial derivatives with respect to two or more independent variables.

Integral Equation: An equation where the unknown function appears in an integrand.

Functional: A function which depends on one or more functions.

Further Reading

Arfken, G. (1985), *Mathematical Methods for Physicists*, New York: Academic.

Bradbury, T. C. (1984), *Mathematical Methods with Applications to Problems in the Physical Sciences*, New York: Wiley.

Churchill, R. V., *et al.* (1974), *Complex Variables and Applications*, New York: McGraw-Hill.

Feynman, R. P., Hibbs, A. R. (1965), *Quantum Mechanics and Path Integrals*, New York: McGraw-Hill.

Glimm, J., Jaffe, A. (1987), *Quantum Physics: A Functional Integral Point of View*, New York: Springer.

Goldstein, H. (1950), *Classical Mechanics*, Reading, MA: Addison-Wesley.

Harper, C. (1976), *Introduction to Mathematical Physics*, Englewood Cliffs, NJ: Prentice-Hall.

Itzykson, C., Zuber, J. B. (1980), *Quantum Field Theory*, New York: McGraw-Hill.

Jones, L. M. (1979), *An Introduction to Mathematical Methods of Physics*, Reading, MA: Benjamin/Cummings.

Kreyszig, E. (1962), *Advanced Engineering Mathematics*, New York: Wiley.

Kreyszig, E. (1978), *Introductory Functional Analysis With Applications*, New York: Wiley.

Lanczos, C. (1970), *The Variational Principles of Mechanics*, New York: Dover.

Lovitt, W. V. (1950), *Linear Integral Equations*, New York: Dover.

Mikhlin, S. G. (1964), *Integral Equations*, New York: MacMillan.

Moiseiwitsch, B. L. (1977), *Integral Equations*, New York: Longman.

Morse, P. M., Feshbach, H. (1953), *Methods of Theoretical Physics*, Vols. 1 and 2. New York: McGraw-Hill.

Pipes, L. A., Harvill, L. R. (1970), *Applied Mathematics for Engineers and Physicists*, New York: McGraw-Hill.

Potter, M. C., Goldberg, J. (1978), *Mathematical Methods*, Englewood Cliffs, NJ: Prentice-Hall.

Press, W. H., *et al.* (1987), *Numerical Recipes*, New York: Cambridge University.

Pryce, J. D. (1973), *Basic Methods of Linear Functional Analysis*, London: Hutchinson.

Schulman, L. S. (1981), *Techniques and Applications of Path Integration*, New York: Wiley.

Schutz, B. F. (1980), *Geometrical Methods of Mathematical Physics*, New York: Cambridge University.

Simmons, G. F. (1972), *Differential Equations*, New York: McGraw-Hill.

Sommerfeld, A. (1949), *Partial Differential Equations in Physics*, New York: Academic.

ARTIFICIAL INTELLIGENCE

SCOTT H. CLEARWATER, *Computer Science Department and Physics and Astronomy Department, University of Pittsburgh, Pittsburgh, Pennsylvania, U.S.A.*

	Introduction	1
1.	**Problem Solving**	3
1.1	Overview	3
1.2	Knowledge Representation	3
1.3	Control	3
1.4	Search	4
2.	**Knowledge Representation**	4
2.1	Overview	4
2.2	Schemes	5
2.2.1	Logic	5
2.2.1.1	Formal Logic	5
2.2.1.2	Fuzzy Logic	5
2.2.1.3	Nonmonotonic Logic	5
2.2.2	Semantic Networks	5
2.2.3	Frames and Scripts	6
2.2.4	Production Systems	6
3.	**Control**	6
3.1	Strategies	6
3.2	Control Methods	7
4.	**Search**	8
4.1	Overview	8
4.2	Blind Search	8
4.2.1	Breadth-First Search	8
4.2.2	Depth-First Search	8
4.3	Heuristic Search	8
4.3.1	Hill-Climbing	9
4.3.2	Beam Search	9
4.3.3	Best-First	9
4.4	Game Search	9
5.	**Machine Learning**	9
5.1	Overview	9
5.2	Concept Learning	10
5.3	Learning by Example	11
5.3.1	Induction	11
5.3.2	Analogical and Case-Based Reasoning	11
5.3.3	Genetic Algorithms	11
5.4	Connectionist Approaches	12
5.5	Other Methods	12
5.6	Current Issues	13
6.	**Implementation Issues**	13
6.1	LISP	13
6.2	PROLOG	14
7.	**Expert Systems**	14
7.1	Overview	14
8.	**Problem Solving Applications of AI**	16
8.1	Scientific Applications	16
8.1.1	Physics	16
8.1.2	Biology	17
8.1.3	Chemistry	17
8.2	Technical Applications	17
8.2.1	Engineering and Industry	17
8.2.2	Medicine	17
8.3	Game-Playing	18
8.4	Defense	18
8.5	Space Exploration	18
8.6	Automatic Programming	18
8.7	Mathematics	19
8.8	Parallels with Human Perception and Language	19
8.8.1	Cognition	19
8.8.2	Vision	19
8.8.3	Language	19
8.8.4	Robotics	20
9.	**Relation to Other Fields**	20
10.	**Future Directions into the 21st Century**	21
11.	**Appendix: Additional Information**	21
11.1	Funding Sources	21
11.2	Centers of Research	21
11.3	Selected Societies, Journals, Magazines, and Conferences	21
	Glossary	22
	Works Cited	22
	Further Reading	23

INTRODUCTION

Traditionally, computers have performed data processing tasks but lacked the ability for reasoning, acquiring, and using knowledge; i.e., they were unable to solve problems. Artificial Intelligence (AI) goes beyond conventional data processing by focusing on the

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

development of programs that can solve problems. They do this, in part, by explicitly representing and using knowledge. Before discussing AI in more detail, we need to understand what is meant by the terms *computer, artificial,* and *intelligence.*

Computers belong to a class of objects known as physical symbol systems. A physical symbol system contains a set of symbols that it manipulates through creation and destruction (Simon, 1981). Copying symbols is considered to be a form of creation while changing an existing symbol into another symbol is considered to be destruction of the old symbol and creation of a new one. Thus, the symbol manipulation in a physical symbol system can be seen as analogous to the evolution of a particle system through allowable quantum states in time. Physical symbol systems consist of an architecture containing a storage area or memory, a processing unit, and input and output units for communicating with the outside world. Figure 1 shows a simple diagram of a physical symbol system.

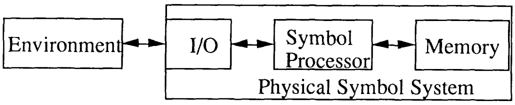

FIG. 1. The three components of a physical symbol system and its interactions with its environment.

Artificial refers to the human-made nature of computers and their programs. Although this term has a somewhat derogatory connotation with respect to food additives and the like, it is, in the context of this article, only descriptive. That is, "artificial" intelligence should be no less intelligent than "natural" intelligence.

Intelligent refers to the type of behavior that people exhibit when solving problems. Solving problems, then, is seen as the ability to achieve goals, where goals are defined as desired final states. Further, when people solve problems they are

1. often able to come up with an answer within a reasonable amount of time (even though it may not be the optimal one, just one that is good enough);
2. able to solve problems even when the data given to them are incomplete, inaccurate, or noisy; and
3. able to improve their performance at solving the task after repeated trials—in short, they learn.

Intelligence is also the ability to use knowledge to constrain a search through an enormous space of possibilities. In fact, one of the most important elements of AI is the study and use of heuristics, or techniques that constrain a search. Unlike traditional algorithmic search, heuristics involve a more exploratory type of search that is better suited towards problems that do not have an obvious algorithmic solution.

Hence, we define AI as the study of principles and methodologies for making intelligent computers. However, others have broadened the definition to include all forms of intelligent behavior, including human. The principles and methods of AI involve representation (of knowledge), control or strategy (of procedures), and search (through a problem space) directed towards solving problems (achieving goals), each of which will be discussed later.

Although computing engines, such as the abacus, have been around for hundreds of years, it was not until the development of computers in the 1950s that intelligence by machine became possible to study. Work in AI began after the development of the computer, so AI is still a young discipline, still in search of exactly what it is. However, just as intelligence is "practical," so too must be the results of AI.

AI at its lowest level is still just computer code, so what distinguishes it from conventional methods? Unlike conventional custom solutions to problems, AI is interested in problem solving in general. In fact, formalization of problem solving techniques that will have applicability across arbitrary domains is very difficult. On the other hand, AI differs from purely formal systems like logic and pure mathematics in that the emphasis is on problem solving which is verifiable through experimentation, which includes using imperfect data. AI methodology is better suited to problems where the structure is ill-defined than are conventional methods that depend on well-structured domains. Hence, the results of AI should be reusable across many domains. Thus, any field that desires automation in problem solving tasks is a candidate for an AI application.

In particular, AI should be a natural fit for science and engineering applications, since these two disciplines are most interested in solving problems in a systematic way. Scientists and engineers know all too well the difficulties with solving their problems—each one seems to have its own idiosyncrasies that demand custom techniques. AI's responses to these difficulties are its flexible problem solving techniques that are reusable in a variety of situations.

One goal of AI is to make science more automated. This may seem somewhat counterintuitive in that one usually thinks of science as something that is not routine, but requires gifted insight and clever experimentation to make progress. However, this is exactly one of the goals of AI, to bring such seemingly unfathomable qualities as creativity and discovery into the definable and programmable.

AI is important to people because it can free them from routine or mundane chores. Some of these monotonous chores include assembly line work, agriculture, and clerical tasks. In the context of science and engineering this means, for example, automatic monitoring and diagnosing of equipment or analysis of data.

The economic impact of computers is already pervasive and continues to have enormous impact on society. But this impact has been mostly in areas where human intelligence is weak—storage of large amounts of records (e.g., medical records) and complex mathematical computation (e.g., weather simulation). These types of applications make use of a computer's hardware capabilities of mass storage and the fast switching time of electronic components. So far the application of man-made systems to problems of reasoning has been small. The advent of intelligent computers will thus have even greater implications in all aspects of human society. The remainder of this article discusses how AI is realized and some of its applications.

1. PROBLEM SOLVING

1.1 Overview

Problem solving is defined as the performance of actions whose purpose is the attainment of some goal. A solution to a problem may be either the goal itself or the steps performed that led to the goal, depending on whether the goal is explicitly stated or not. Note that given an arbitrary problem, there is no guarantee of a solution, or even that the found solution be the optimal one.

What AI offers is a methodology for finding solutions to problems that are intractable because no formal mathematical technique exists, or are prohibitively expensive using brute force methods like exhaustive search. Methods that reduce search but do not guarantee solutions are called heuristic. The notion of heuristic methods is essential to the theme of AI. The tireless computer, through the use of heuristics, offers the possibility of exploring multiple, though not necessarily all, good enough solutions, whereas in human problem solving the solver often stops after finding the first good enough solution.

Problem solving involves several parts, some of which may overlap. The parts described in this article are referred to as representation, control, and search. An obvious implicit component in any problem solving system is the critic that evaluates a result and decides whether it is good enough, or modifies one or more of the components if it is not good enough.

1.2 Knowledge Representation

Knowledge may be implicit, as in procedures, or explicit, as in a database. Implicit knowledge is less easily understood, viewed, or manipulated than explicit knowledge. The behavior of the system depends on how the knowledge is represented. For example, a rule-based representation uses antecedent–consequent, or condition–action (IF–THEN), rules to test whether certain conditions have been met. The rules may chain together; i.e., the action of one rule is part of the condition of another rule so that they form a link of inferences. An advantage of this kind of representation is that the chain of inferences can be used as an explanation for why a particular final action was taken by the system.

1.3 Control

A problem-solving system needs a model in which to solve its problems. One popular paradigm is called generate-and-test. In the generate-and-test paradigm, a generator is used to create hypotheses that are then tested against the data and knowledge available to the system. Methods for constraining the generator, to search only the more plausible hypotheses, are provided through the search specification. A test program checks whether the generated hypothesis solves the problem. The system may return all the solutions or only some desired subset (e.g., the first solution).

Controlling a system also involves the detailed interactions of the data and procedures. This includes the order in which procedures are called and the data that are passed between them. Some systems operate under very predictable conditions, whereas others operate in a less predictable fashion. For example, the order in which data are received may be important because the actions of the system will vary according to the presentation of the data.

1.4 Search

In order to solve a problem, a program must search through a space of possibilities. These possibilities can be thought of as subproblems of the overall goal to be achieved. The difficulty in search is that the subproblems can interact so that what may seem like a good path to follow at one point may later turn into a dead end. Thus premature commitment to a path results in costly backtracking when the path is not the right one. In many real-world problems the potential search space is enormous and cannot be searched completely in any finite time by any physically realizable computer. Thus, there is not an explicit construction of the search space, but only an implicit one based on the parts of the space that are actually searched. Chess provides an example of a huge search space where humans have learned to solve the problem of playing the game in a finite amount of time. The way to avoid exhaustive search is to constrain it through knowledge of the problem domain. This is where AI's use of heuristics comes into play.

2. KNOWLEDGE REPRESENTATION

2.1 Overview

Knowledge in AI is not simply a set of statements, numbers, or facts stored in a database. It also refers to the data structures and the procedures used to manipulate and interpret the data as well. It includes the syntax and semantics of the objects in the knowledge base. In terms of problem solving, knowledge and search are related in an almost uncertainty principle–like way: the greater the amount of knowledge that can be brought to bear on a problem, the less search will be required, and the greater the amount of search, the smaller the amount of knowledge that is required to solve the problem (Berliner and Ebeling, 1986). However, unlike physics there is more than one value for Planck's constant, corresponding to the complexity of a given problem.

In AI, representation is the mapping between the symbol system and the environment; it is how the symbol system sees or describes the world. In physics and engineering the mapping between nature and nature's model, physics, is given by parameters such as mass, energy, charge, etc. Just as in physics problems, the correct representation can greatly facilitate problem solving. For example, in good representations, important things should be made explicit, including any constraints that may apply. A good representation contains all the things that are needed to solve the problem. Also, the grain size of the knowledge must be appropriate for the type of problem to be solved. Additionally, representations should present their knowledge in a clear and concise way. For instance, unnecessary information should be suppressed unless needed to solve the problem. The explicitness, clearness, and conciseness of the representation mean that understanding and modifying it will be much simpler, less time-consuming, and less error-prone. While all this may sound like standard practice in software engineering, the difference is that in AI what is being represented is the problem solving knowledge. However, unlike physics AI is not constrained to using numerical parameters, but may use symbolic data such as "big" and "heavy."

There are several types of knowledge that must be represented for a system to be capable of intelligent behavior. First, the system must have knowledge about objects—things that have properties (e.g., cars have wheels). Second, the system must have knowledge about events—happenings that occur to objects (e.g., the car got a flat tire). Also, an intelligent system must know about behavior—how objects interact with each other in events. Finally, an intelligent system should have metaknowledge—knowledge about knowledge (e.g., "knowing what you know").

As mentioned earlier, knowledge may be either explicit (declarative) or implicit (procedural). Most scientific programming hides knowledge in procedures. An example of the two representations is given by the problem of comparing musical notes. In a declarative representation the knowledge "F is higher than B" and "B is higher than A" are stored as facts in a knowledge base. In a procedural representation the sorting is done, say with numeric comparison of the sound frequencies, and the program itself has the explicit knowledge of how to do numerical comparisons. It should be pointed out that declarative representations still require procedures to manipulate them.

2.2 Schemes

Numerous schemes for representing knowledge have been developed throughout history. Some of these are useful for AI applications, as are ones developed as a direct result of AI research.

2.2.1 Logic. Logic offers a rigorous way of representing and proving things about situations (cf. LOGIC METHODS). However, there are also disadvantages to using "formal" logic, as will be discussed in sections on fuzzy and nonmonotonic logic.

2.2.1.1 Formal Logic. The formalism of logic has developed a language that provides a powerful means of knowledge presentation. Rules of inference (deductive rules) are used to derive facts that are true given other facts known to be true. This guarantee of truth is a powerful feature of formal logic. One useful application of formal logic is in its ability to prove theorems. Unfortunately, proof itself is subject to an exponential explosion in the number of nodes in the search space. Second, the search may not terminate unless there actually is a proof, which may not be known *a priori.*

2.2.1.2 Fuzzy Logic. Another one of the drawbacks of formal logic is the binary nature of its statements, either true or false. One way around the difficulties of the rigid constraints of formal logic is provided by fuzzy logic (Zadeh, 1979). In this logic, objects belong to a class with some range rather than being either in a set or not in a set; i.e., in a range 0 to 1. Fuzzy sets are useful in natural language applications, for instance, when describing a concept (e.g., "fast"). Fuzzy representation allows a gradual transition from fast to non-fast. Thus, if A is (FAST 0.9) and B is (FAST 0.6) then one can say A is "more fast" than B, whereas in formal logic both A and B may belong to the same class (either FAST or NON-FAST) or different classes.

2.2.1.3 Nonmonotonic Logic. When building a logic system, there is no guarantee that the axioms will be complete or noncontradictory. In nonmonotonic logic, one is able to get around the problem of conflicting conclusions drawn from the same set of axioms. Namely, if new assumptions have been made that may conflict with existing assumptions, then new inferences may be made. Thus, nonmonotonic logic does not assume that what was true at one time is necessarily true for all times. Nonmonotonic logic could be thought of as a time-dependent logic.

2.2.2 Semantic Networks. Semantic nets use a node and link graph structure to represent objects and their interrelations. Nodes represent concepts such as objects or events. The links between nodes represent interrelations between the objects. Figure 2 shows an example of a semantic net for a car. Unfortunately, there is some ambiguity in this representation. For instance, one may

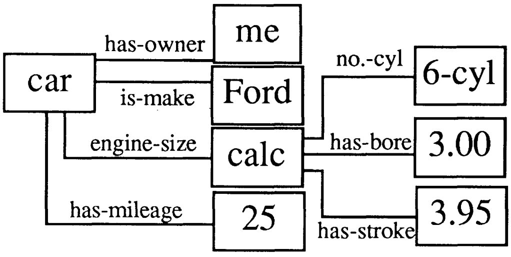

FIG. 2. A semantic network for a car.

conclude that all cars have these properties rather than just my car, or that all cars are owned by me. However, this problem can be eliminated by making a distinction between the car class and an instance of a particular car.

2.2.3 Frames and Scripts. In a frame representation declarative knowledge (facts) and procedural knowledge (programs) exist together in a predefined relationship to describe an object or event. Slots are used to hold values for attributes associated with a particular frame. An example of a frame for a particular car might be the following:

FRAME-NAME: CAR

OWNER: ME

MAKE: FORD

COST: 10,500$

MILES-PER-GALLON: 25

BORE: 3.00

STROKE: 3.95

NUMBER-OF-CYLINDERS: 6

DISPLACEMENT: PROCEDURE(CALC-DISP BORE STROKE NUMBER-OF-CYLINDERS)

Note that slot values need not be just numbers or symbols but, as in the *DISPLACEMENT* slot, can call out to procedures to obtain computed quantities as needed. Also, all objects belonging to the car class will have the same structure.

Scripts are frames that correspond to a particular situation or action (e.g., what happens when you go to a restaurant). The script includes slots for the low-level acts that can occur in a particular situation (e.g., waiting for a table or leaving a tip in the case of a restaurant script).

2.2.4 Production Systems. In a production system, knowledge takes the form of rules such as: IF condition exists, THEN perform action. An advantage of this representation is that the condition for its applicability is explicit. Additionally, rules are very modular and easy to modify and replace. Rule-based systems may be either antecedent-driven, that is, the system matches the condition parts of the rule and then takes the actions specified by the consequents, or consequent-driven, where the matching is done on the action parts of the rule, and the condition parts are performed.

However, when using a rule-based approach, several potential problems must be addressed. First, what should be done when multiple rules fire whose conclusions may conflict? In this case, the rules may be ranked by their previously determined accuracy in making correct classifications, or by their rate of misclassification, or by the ratio of correct classification rate to incorrect classification rate. Also, the rules could be ordered by the benefit of performing the recommended action, assuming it is correct, or by the inverse cost of performing the action if it is the wrong action. A second problem occurs when no rules match the current situation and the best that can be done is a partial match. One possibility, in this case, is to use the fraction of correct conjunctions matching to rank the rules.

3. CONTROL

3.1 Strategies

In order to perform a search, a strategy represented by some set of operators is needed to transform the problem from one state to another so that the goal can be satisfied. In logic, operators are the rules of inference, such as *modus ponens*. In game playing, the operators are the rules of the game.

A state-space search can be represented by its set of states, set of operators, and set of goal states. The state space itself is represented as a directed graph, or net, or tree, with the nodes representing the states and the links, or branches, between nodes representing the application of operators. A typical tree is shown is Fig. 3. A node with no parents is called the root node. Terminal nodes are the nodes with no children. Goal nodes are terminal nodes but not all terminal nodes are goal nodes. The difference between control paradigms is their different methods of traversing the search space to arrive at a solution.

More specifically, operators may employ forward, or data-driven, reasoning in which the object is to transform the current state forward toward the goal state. Another approach employs backward, or goal-directed,

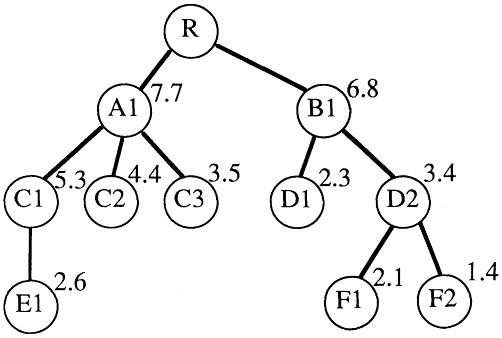

FIG. 3. The letters in the nodes identify the states. The numbers refer to the "distance" from the goal node. Distance is a metric used for measuring the cost in time or money, or whatever other parameter is relevant to solving the problem. When choosing a particular control strategy it is important to consider the cost of finding a path as well as the cost in actually traversing it. R is the root node and F2 is the only goal node. The goal of the search is to find a solution within some cost bound.

reasoning where the object is to work backward from the goal state by breaking the problem into smaller, more easily solvable subproblems. Finally, it is possible to combine forward and backward reasoning into a single strategy. One popular such method, means-ends analysis, finds the difference between the current goal state and the current state. The difference is used to fire a forward operator. If the most relevant forward operator does not decrease the difference then a backward operator is used to set up subgoals.

3.2 Control Methods

This section is intended mostly for multiprocessing systems where numerous procedures may be running concurrently, possibly on different machines. Control methods are how the procedures in a symbol system interact. All the procedures require certain input knowledge and resources in order to give an output. Control methods specify the architecture, resource allocations, and interaction among procedures. Three paradigms for control are discussed.

In the first control paradigm, each procedure has explicit knowledge of what other procedures to call so that it can complete its task. This is the method used in most scientific computing. Figure 4 shows an example of this kind of control.

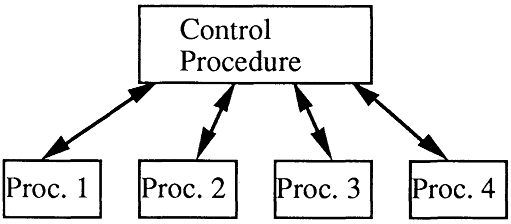

FIG. 4. A control paradigm where one procedure knows about all the other procedures.

In the second control paradigm, the entities in the symbol system belong to particular classes. In this case, it is possible to refer to a class of objects rather than to a particular procedure. The knowledge of which procedures are available and best suited for the particular problem is stored within a class object, as shown in Fig. 5.

In the third kind of control paradigm the procedures have no knowledge of other procedures, only that they can respond to certain requests from outside. Control occurs in response to statements posted on some centralized location, as depicted in Fig. 6. The centralized knowledge source, or blackboard, can be thought of as being analogous to a group of scientists brainstorming in front of a blackboard, each one writing his ideas down and drawing on the collective wisdom represented by the writings of the other contributors to solve the problem.

The problem of resource allocation may be handled in a number of ways. This problem arises when numerous procedures are all trying to solve some task using the same resources. One solution is simply to give each

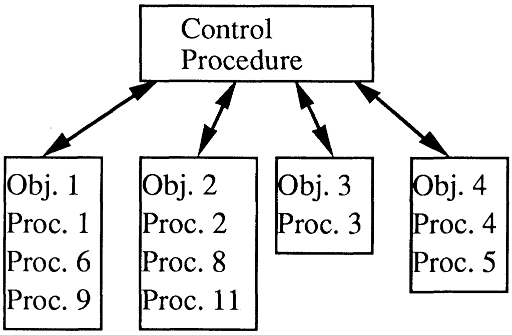

FIG. 5. Control between objects, showing links between procedure and objects.

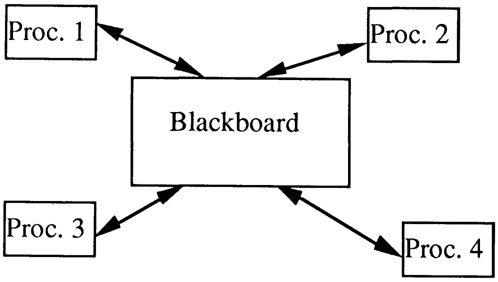

FIG. 6. Blackboard control, showing the links between the procedures and the centralized knowledge source (blackboard).

procedure an equal amount of the resource (processing time, memory, etc.). Another method allocates the most resource to the procedure that will have the highest "payoff" towards completing the task. Yet another scheme allocates on the basis of the inverse cost of performing the procedure, i.e., the more it costs to do it, the less resource it gets. More complicated resource "payoff" functions may be constructed that are combinations of the cost and benefit relevant to the problem solving activity. Further discussion of control methods can be found in Winston (1985).

4. SEARCH

4.1 Overview

Search is the engine that drives a system through a space of alternatives toward its goal. Thus, search is one of the most basic functions in any problem solving activity. It is the path through that choice of alternatives that the system follows on its way toward solving a problem. However, search must also address the problem of time and space limitations inherent in any physical system. Before discussing particular search algorithms we need to find a representation in which to conduct the search. The rest of this section presents several search techniques, each with different strengths and weaknesses. Further discussion of search methods can be found in Barr, Cohen, and Feigenbaum (1982) and Nilsson (1980).

4.2 Blind Search

Blind search is a class of search strategies wherein the search space is explored via the syntactic or structure of the space. In blind search there is no preference for one branch of the search over another. The search ends when the first path to the goal is found, which is not necessarily the best one. One of the problems with blind search is the combinatorial explosion of paths that occurs for all but the smallest of problems.

4.2.1 Breadth-First Search. In breadth-first search, the search path proceeds across nodes at a given level in the tree before going a level deeper. In Fig. 3 a breadth-first search would be: R, A1, B1, C1, C2, C3, D1, D2, E1, F1, F2. The disadvantage of this type of search occurs when multiple solutions exist roughly at the same level for many nodes.

4.2.2 Depth-First Search. In depth-first search the search proceeds by following the children of a particular node down the search tree before trying other paths at the same level. Thus, a depth-first search in Fig. 3 would proceed as follows: R, A1, C1, E1, C2, C3, B1, D1, D2, F1, F2. The disadvantage of this type of search is that, for very deep search trees, the system spends most of its time exploring fruitless paths that are far deeper than the solution to the problem. Depth-first tends to work best for problems where breadth-first is poorest and vice versa.

4.3 Heuristic Search

Recall the definition of the popular (in the AI world) and also controversial word "heuristic," to mean a strategy that enables a reduction in search through a space. The reduction in search means that the plausible part of the search space is reduced either by pruning unpromising nodes in the search tree or by focusing the search in a promising area. However, heuristic search does not guarantee a solution. Also, heuristic solutions are not necessarily optimal ones. Rules of thumb are heuristics as are other "tricks" used to solve problems. Heuristic search can be used to reduce the combinatorial explosion of possible search paths that exist by using purely blind search methods. However, in order to use heuristic methods there must be a reliable measure of cost or distance (e.g., from the current node to the goal node) so that the potential value of the heuristic can be determined. These distances are used to order the possible expansion nodes in the tree. The heuristic itself tells what the next node to expand should be and depends on the particular heuristic. Note that blind search methods can be modified so that the solutions can be found sooner.

4.3.1 Hill-Climbing. Hill-climbing search is basically parameter optimization and is the most familiar type of search used in optimization problems in physics. This is the type of search used by numerical optimization programs. Here, the hills in the search space are the relative extrema of the performance function. The profile of the surface depends on the performance function. The direction of the search path depends on the optimization method (e.g., gradient descent in numerical domains). Thus, the particular next node chosen would be the best of the children of the last node explored. Some of the problems with hill-climbing are that it is not exhaustive and that the complexity of the search surface increases with the dimensionality of the problem. This complexity leads to a greater prevalence of nonoptimal relative extrema such as foothills, plateaus, and ridges. In Fig. 3 a hill-climbing search would proceed as R, B1, D1. This example shows how hill-climbing can get stuck in local optima. To get around this problem requires costly backtracking.

4.3.2 Beam Search. Beam search is a modified version of breadth-first search where the path followed keeps active the "best" nodes at each level in the search, where "best" is based on the cost required to reach the goal. In Fig. 3 above, suppose that only the two best (lowest cost from the current node to the goal node) nodes are expanded to the next level. Then the beam search would be R, A1, B1, D1, D2, F1, F2. Note that in this case the C nodes were not expanded because they were "not good enough" according to the heuristic "expand best 2."

4.3.3 Best-First. Best-first search is similar to hill-climbing, except that in best-first the next node is chosen from the best of all the children of all the nodes explored so far, where "best" is as above. Best-first can often converge more quickly than hill-climbing at the expense of having to sort all the children at each expansion. In Fig. 3 the search would proceed as R, B1, D1, D2, F2. Note that unlike the hill-climbing example, the best-first search did not get caught in the local optimum, although, in general, for any heuristic search it is possible to get caught in a local minimum.

4.4 Game Search

Game search differs fundamentally from the above-mentioned techniques in that it involves the actions of two or more players whose actions affect the paths taken in the search.

Minimax search is a depth-first search where the depth is limited to a certain number of moves ahead. The path is then evaluated according to its value to the player. The idea is to maximize the evaluation function for the player and minimize it for the opponent. A maximizing evaluation function operates on alternating layers of the game move tree with a minimizing evaluation function. Both of the evaluation functions must be aware of the moves the other is likely to make. This search, like any heuristic technique, requires a measure of goodness to judge the moves.

5. MACHINE LEARNING

5.1 Overview

Learning is one of the most fundamental aspects of intelligent behavior. We define learning as the process by which a system is able to perform a given task more efficiently on subsequent trials. More efficiently means that either the accuracy (quality) of the solution is improved or the time to find a solution is decreased. The need for machines that learn is motivated by the desire to get away from the tedium involved in human learning and the difficulty in transferring human problem solving ability to a machine. While human learning is certainly an existence proof for intelligence, it should be kept in mind that it may be no more useful for the construction of learning machines than birds were for the construction of flying machines. Besides human learning, other examples of learning systems are species evolution, where the goal of learning is survivability; and the immune system of a living thing, where again the goal of learning is survivability.

Learning systems must create some representation, or model, for the world in which they will be learning. During learning, the model is refined to represent more closely the environment it is trying to learn. In some systems the model may take the form of differential equations in which learning takes place through the modification of coefficients. Adaptive control systems, neural networks, and pattern recognition systems fall into this category of learning system.

Learning can be divided into two categories. The first, knowledge acquisition, is the ability to construct new models and apply them in a variety of problem solving situations. The second, skill refinement, comes about from continued practice of some task with feedback on the performance.

Additionally, learning systems may be classified as to whether they perform their learning in a supervised or unsupervised mode. A supervised environment is akin to a stimulus–response system where the correct output for each input is available to the learner. In unsupervised learning there is no such teacher or oracle available. This latter situation leads to one of the most important issues in learning. That is, what component of the learning system receives credit or blame during training so that the system can avoid making the same mistake next time or can be more confident in getting it correct the next time? In other words, where does the feedback from the training example go? This credit-assignment problem remains one of the fundamental issues of learning.

Pattern recognition systems are learning programs used to do classification of some pattern or sequence of data. Many learning tasks can be viewed as pattern recognition or concept formation problems. A typical example of pattern recognition is recognizing written characters. There are also many tasks in experimental physics that require the classification of patterns, from oscilloscope tracings to energy depositions in large scale calorimeters.

More information about learning can be found in the *Machine Learning* books (Michalski, Carbonell, and Mitchell, 1983) and Vol. 3 of the *AI Handbook* (Barr, Cohen, and Feigenbaum, 1982).

5.2 Concept Learning

Learning systems in one important class, called concept formation systems, alter the symbolic structures in their representation of the environment, called the domain theory, in an attempt to form generalizations about classes of objects. A typical learning system contains the following components: environment, instance selector, learning element, domain theory, critic, and performance element.

Environment. This is where the training instances come from and where the ultimate feedback comes from.

Instance Selector. This element chooses the examples that the learning system is going to use. It may be as simple as accepting all examples given to it, or it may be intelligent itself and only choose "critical" examples that will be most efficient for learning.

Learning Element. This element is responsible for taking the recommendations of the performance element and suitably modifying the domain theory of the system. For example, based on the poor predictive power of a set of rules, the learning element will modify the rules in an attempt to improve performance to a level acceptable as specified in the domain theory.

Domain Theory. This element contains the current representation of the "world" or environment. It may be in the form of rules or parameters, for example. It also contains the semantics and syntax used in the domain representation as well as other constraints on the concept description language, such as the overall minimum performance level to be achieved during learning.

Critic. The role of the critic is to analyze the overall performance of the learning system and make recommendations about how and where changes should be made. This is where the credit-assignment problem is handled. The critic may be a local critic responsible only for one element of the learning system or a global critic responding to the overall performance of the system.

Performance Element. The purpose of this element is to test the current domain theory against examples given by the environment to see if the current theory is good enough. It also performs any actual manipulation to the environment requested by the theory in response to the environmental situation (e.g., robot movement).

5.3 Learning by Example

Learning by example is one of the most powerful methods of learning for people and machines. Presenting a system with preclassified examples (with respect to their class) allows the system to form generalizations based on the regularities it has seen in the individual examples of the training data. Of course, the utility of the generalizations depends on how well the training instances model the actual situation to be learned.

5.3.1 Induction. Inductive inference is one of the most effective tools in learning. Here, an inference is defined as a legal (with respect to a set of symbols and rules) translation from one statement to another. Induction is defined as an inferential process whose goal preserves truth and expands knowledge. Whereas deduction guarantees the truth of a conclusion given the truth of the premise, induction provides no such guarantee. Also, the inferences from induction are a function of the structure of the induction system and not purely a function of their premises as in deduction. In a learning system, there must be an amplification of the premise so that deduction alone cannot be used in a learning system; i.e., in deduction the premise and conclusion are equivalent so that no expansion of knowledge can occur. Below is an example of the algorithm for a rule-based induction program.

1. *Input Partial Domain Theory and Examples.*
2. *Recall previously saved rules.*
3. *Any untested rule? If no, then save "good enough" rules, done.*
4. *Evaluate next untested rule.*
5. *If rule too specific then go to 3.*
6. *If rule too general then add specializations and go to 3.*
7. *Add rule to "good enough" rules and go to 3.*

Further insight into induction can be found in Holland *et al.* (1986).

5.3.2 Analogical and Case-Based Reasoning. Analogy is one of the most powerful methods of human reasoning, but also one of the most difficult to codify. Analogy involves forming representations in one domain based on similarities with another domain. The difficulty lies in codifying what is meant by "similarities," particularly when non-numeric data are involved. Also, in making an analogy not all the features in the one domain will be mapped over to the other, and the problem is finding the correct ones. In the special case where the domains are the same the technique is called case-based reasoning.

5.3.3 Genetic Algorithms. Genetic algorithms are induction programs that learn, in analogy to biological evolution, through "mutation" of rules. Here, each gene in the population represents a string and each segment in the gene represents conditions and actions of rules. The basic idea is the "survival of the fittest" where, after a time, only the best rules will be active. The number of rules is fixed so that the poorly performing ones are eventually weeded out by better performing ones.

The program starts off by initializing a fixed population of genes, usually randomly in the absence of any prior knowledge. The current population is modified by operators to generate the next generation. There are several genetic operators that modify the population to produce offspring for the next generation. The modifiers are analogous to the "specialize" and "generalize" operators in the previous rules-based induction example. One of the most important operators is called crossover which takes portions of two genes and exchanges them between the two genes. Note that there is a random component not just in the mutation operator but in where the crossover operator crosses. Another operator causes random mutations in a gene. This operator functions at a much lower rate than the crossover operator. This is because, just as in biological systems, most mutations are not beneficial but they are necessary to insure that the entire search space can be covered. Operators are applied probabilistically to the strongest rules.

Once the new genes have been produced, a fitness evaluation function measures how well the output of the program agrees with the output produced by the training examples. The rules are ordered by fitness and only some number of the best are kept for the next generation. The survivors are then used as the parents for the offspring (i.e., the next generation of rules) so that the overall population is kept constant.

5.4 Connectionist Approaches

Connectionist approaches cover a class of learning programs commonly referred to as neural networks (*q.v.*), Boltzmann machines, simulated annealers, etc. The basic idea behind these programs is to employ the collective behavior of a large set of simple processing elements connected together in some way. It is through adjustments in the connectivity of the processors, reinforced positively or negatively with respect to training examples, that the system adapts and learns. The basic methods involve some sort of hill-climbing search through a performance space where the highest peak would theoretically be the one where the output of the program, given the training example inputs, was identical with the corresponding training outputs. Once the connection strengths have been learned, the network can be presented with cases it has not seen before and make predictions based on what it has seen before. Thus, the program behaves as a function interpolator. Like any induction-type learning system, the learner must be presented with a representative set of training examples. Changes in the network can be represented with differential equations (or, more accurately, as finite difference equations) which gives them a certain appeal for physicists.

The connectionist approach differs markedly from the traditional "computer science" type of approach in that the knowledge is represented implicitly in the connection strengths rather than explicitly in a knowledge base. Also, neural networks lack the modularity of design and flexibility of representation that is a hallmark of knowledge based systems. Additionally, neural networks only work with numeric inputs so that any symbolic inputs must be first converted to numerical form in an *ad hoc* fashion. This lack of explicitness also prevents neural networks from providing a reasoning traceback of how a particular answer was found. The analogic nature of neural networks permits a gradual shift in classification, more than is usually allowed in rule-based systems. However, rules can be given strengths or fitness and appear "fuzzy" just as the outputs of a neural network do.

One drawback to the neural network approach is the usually significant amount of computing time required. This is because prior knowledge of the domain (e.g., heuristics) cannot be explicitly codified (recall the relationship between knowledge and search mentioned in Sec. 3) so that learning even a trivial concept like exclusive-or (i.e., $00 \Rightarrow 0$, $01 \Rightarrow 1$, $10 \Rightarrow 1$ and $11 \Rightarrow 0$) can take hundreds of iterations through the example set. Another drawback is the possibility that the system performance will oscillate and never converge because the input parameters were not set correctly. The input parameters that must be chosen are the number of input units, output units, and "hidden" units (used to model higher order correlations), as well as their connectivity; i.e., not every unit needs to be influenced by all the other units. Also, the initial rate of change, or step, in the search space must be chosen—too small and learning will be agonizingly slow, too fast and the system may oscillate or behave chaotically. Choosing the correct parameters is based completely on intuition and experience. However, neural networks may be more suited to perceptual tasks such as image understanding than rule-based techniques. Such a system could be used to identify interesting events in a physics apparatus. A standard reference for neural networks is Rumelhart and McClelland (1986).

5.5 Other Methods

In explanation-based learning a concept can be learned from only one example (Mitchell *et al.*, 1986). This method has the advantage over induction which requires exponentially large numbers of examples before a conclusion can be drawn with some confidence. Developed in the 1980s, this method offers practical techniques for utilizing prior knowledge represented in a domain theory to reduce the training set size to a tractable value. Unfortunately, if the example is noisy then no concept, or the wrong concept, may be learned, although further research may alleviate this limitation.

In learning by discovery, the learner searches without the aid of a teacher. The learner must form its own clustering of concepts or classes based on the regularities it sees in the data (Langley *et al.*, 1983). In learning by instruction, a teacher or oracle explicitly tells the program what is to be learned. In this case, the teacher performs the role of an intelligent instance selector to give only the most relevant examples to the program. In rote learning there is explicit storage of the training examples. New examples are then compared against the training examples to find the closest match. The classification of the closest match then becomes the classification for the new example.

5.6 Current Issues

One area of current research in machine learning is aimed at improving the scaling of learning systems with respect to the number of examples and their complexity. Incremental learning is an approach to the problem of learning when not all of the training instances are processed at one time. Recent results from learning theory need to be broadened to increase the understanding of the relationship between the number of training examples, the tolerance for error, and the probability of successful learning.

6. IMPLEMENTATION ISSUES

Obviously one of the considerations in the development of intelligent systems is how to actually build one. Building includes the hardware and software to be used. While it may be true that "you can write anything in anything," meaning that an intelligent program could be written in any computer language, not all languages facilitate the construction of intelligent systems. The issues of implementation are those of facility, as in physics when choosing a coordinate system or gauge in which to solve a problem.

AI very often deals with symbol manipulation, rather than purely numerical manipulations as is usually the case in scientific programming. Thus, a language with extensive symbol manipulation capabilities would be desirable. Also a language supporting recursion, modular programming, and easy representation of formal logic would be useful (see COMPUTER PROGRAMMING LANGUAGES).

6.1 LISP

LISP (LISt Processing), first developed in the 1950s about the same time as FORTRAN, has found wide application in many areas of AI. Common LISP (all further references to LISP will refer to Common LISP) is the dialect that has become the standard that was long lacking. The basic data structure in LISP is a list which allows LISP to represent nearly any other data structure. In addition, LISP supports a large variety of data types such as integers, floating point numbers, complex numbers, ratios, characters, strings, symbols, arrays, vectors, bit-vectors, hash tables, etc. LISP also supports the full range of trigonometric, exponential, complex, and comparison operators found in "scientific languages" like FORTRAN. Functions are represented as lists where the first element in the list is the name of the function and the other elements in the list are the arguments and body of the function. Lists may contain lists of lists of lists, etc. The standard LISP reference is Steele (1984).

Originally, LISP was run interpreted and although this facilitates debugging, execution time is longer. However, there are LISP compilers that produce fast-running LISP code. Specialized hardware designed to run LISP code provides a high productivity environment with such features as a language-sensitive editor (checks syntax automatically, does indentation), incremental compilation (only compiles the lines of code that changed), and automatic linking, all of which act to reduce dramatically the edit–compile–link program development cycle.

Now consider a trivial example in LISP. In the example a database of facts will be stored and then inferences will be used to test hypotheses about the database using a rule. Suppose the property list in which the data will be stored is called JOHN. The function SETF is used to assign a value to the indicator TYPE to be the value PERSON, and so on.

```
(SETF (GET 'JOHN 'TYPE) 'PERSON)

(SETF (GET 'JOHN 'JOB) 'FUN)

(SETF (GET 'JOHN 'HEIGHT)' > 75)

(SETF (GET 'JOHN 'SMILE) 'ALOT)

(SETF (GET 'JOHN 'FROM) 'CALIFORNIA)
```

We can examine the resulting data structure with

```
(SYMBOL-PLIST 'JOHN)
```

which returns

```
(FROM CALIFORNIA SMILES ALOT HEIGHT >75
JOB FUN TYPE PERSON)
```

Now we can write a rule to see if a person is happy or not. Thus,

```
(DEFUN HAPPY-RULE (WHO?)
(COND ((AND (EQ 'PERSON(GET WHO? 'TYPE))
            (EQ 'FUN (GET WHO? 'JOB))
            (EQ 'ALOT (GET WHO? 'SMILES))
            (EQ   'CALIFORNIA   (GET   WHO?
            'FROM))
        )T)
       (T NIL)))
```

Now executing the function (possibly from inside another function),

```
(HAPPY-RULE 'JOHN)
```

returns T since all the conditions of the COND were true. In this example note how the data are separate from the rule procedure, and the mixture of numeric and symbolic data. Also note that the rules themselves are modular units that are easily modified.

6.2 PROLOG

Another popular "AI language" is PROLOG (PROgramming in LOGic), which is an implementation of first-order predicate calculus. Logic programming differs from conventional programming in that the programmer supplies facts in a database and lets the computer draw any conclusions, rather than explicitly writing a program that tells how to solve the problem. Put more simply, logic programming is more concerned with *what* a program computes rather than the details of *how* it computes. The standard PROLOG reference is Clocksin and Mellish (1981).

Given a question, PROLOG will do a depth-first search through its database of facts and rules in an attempt to resolve the question put to it. For example, we redo the previous LISP example in PROLOG.

Given the following database:

```
PERSON(JOHN).
JOB(JOHN, FUN).
HEIGHT(JOHN, >75).
SMILES (JOHN, ALOT).
FROM(JOHN, CALIFORNIA).
```

which translates to "John has a fun job, is over 75 inches tall, is from California, and smiles a lot." Now given the rule:

```
HAPPY_PERSON(X):-
    PERSON(X),
    JOB (X, J),
    SMILES(X, Y),
    FROM(X,California).
```

which translates to "X is a happy person if X is a person who has a job, smiles a lot, and is from California." Then inputting the question to PROLOG gives

```
?- HAPPY_PERSON(JOHN).
```

which yields the answer,

```
YES
```

This trivial example shows how PROLOG uses facts and rules to draw conclusions. Again note the modularity of the facts and rules.

7. EXPERT SYSTEMS

7.1 Overview

There is almost always a need for expert advice in any organization. Expertise, almost by definition, is a rare quantity (i.e., there is not enough of it or it is in the wrong place), and anyone able to make it more available could expect improvement in productivity. Expert systems are part of a proven technology that addresses the problem of making expert problem solving more available. Expert systems are programs that seek to achieve a high level of performance at some task through the use of codified knowledge of human experts. While there is no hard and fast criterion for defining what is and what is not an expert system there are numerous desiderata for what an expert system should contain. Several typical criteria for an expert system are listed below (Buchanan and Smith, 1988).

1. Able to reason with domain-specific knowledge that may include both symbolic and numeric data.
2. Able to use domain-specific techniques that may include both heuristic (plausible) and algorithmic (certain) methods.
3. Can perform comparably to experts in both accuracy and timeliness of response in its specific problem domain.
4. Can provide reasoning (backtraces) for how it arrived at a particular decision.
5. Has a flexible architecture that facilitates modification of domain knowledge and control structures.

Expert systems, by criterion 3, work well in their domains but their performance usually drops off sharply when the domain is changed or modified. Thus, although expert systems seek to embody general methods of reasoning and problem-solving, the task-specific nature of the knowledge base severely limits their applicability outside the domain for which they were originally designed. Recently, though, an attempt began to overcome this fragility so that knowledge bases would be reusable (Lenat, Prakash, and Shepherd, 1986). The standard references for expert systems are Hayes-Roth, Waterman, and Lenat (1983), Waterman (1986), and Buchanan and Shortliffe (1984).

An expert system contains several components: a knowledge base of facts, heuristics, rules, and procedures; an inference engine for applying the rules using the facts in the current situation; a knowledge acquisition system for incorporating knowledge into the knowledge base; and a unit for inputting and outputting data. A typical expert system architecture is shown in Fig. 7. Let us now discuss each component of an expert system in some detail.

First, there must be a way to get data from the environment and a way to manipulate the environment based on the response of the system. This input/output (I/O) module includes the human interface, since that is also often where the knowledge comes from. Also, a good interface is needed for debugging and other modifications that invariably occur over time. The man–machine interface also influences whether the system will be used by people, apart from how well the overall system solves problems.

Another feature (and shortcoming) of expert systems is knowledge acquisition. Traditionally, knowledge is acquired by a "knowledge engineer," a programmer with special training in building expert systems, who interviews the domain expert. Usually only one expert is chosen, which eliminates the problem of conflicting expert opinion but biases the expert system to one particular source of knowledge. Further, the knowledge engineer often has very little training in interview methodology, and the expert often has difficulty relating to the questions of a novice. Additionally, the expert's time is at a premium, so even obtaining the knowledge is quite laborious and time-consuming. However, this custom crafting has the virtue that it works.

The inference engine uses the knowledge (e.g., rules) in the knowledge base together with the data obtained from the environment to take actions and make conclusions. These actions may use other rules in the knowledge base, forming a chain of inferences before finally reaching some conclusion.

The knowledge base contains the facts, rules, etc., that the system uses to make its decisions. A key feature of expert systems is that the specialized problem knowledge is kept

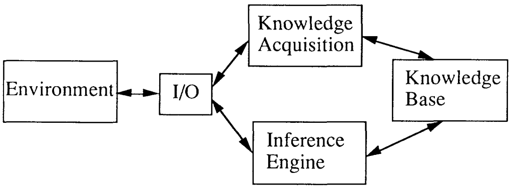

FIG. 7. Schematic diagram of an expert system showing the components and the links between them.

separate from the general reasoning methods in the inference engine.

A powerful feature of expert systems is their explanation capability which makes them much more believable to human users. When someone can explain his reasoning, we usually have much more confidence in his conclusion, because he seems to "know it" rather than our having to take it "on faith."

Not all problems are well suited to expert system methodology. For example, problems that are not well circumscribed are much more difficult to work with than ones that are clearly defined. Another problem encountered during expert system building occurs when the assumptions in the domain are not explicit. This can lead to poor performance when a situation occurs outside the assumptions. Debugging these assumptions can also be very difficult.

The organizations that successfully rode the first wave of AI are now clamoring for even more automated methods for problem solving. Clearly, overcoming the "knowledge acquisition bottleneck" is one area that is receiving great interest. This is the weakest link in the expert system chain. One of the most promising solutions comes from the area of machine learning; see Sec. 6.

Current research in expert systems includes development of more generally usable shells (an empty expert system where the knowledge base is filled in by the user) and less brittle systems that do not fail miserably when applied in a slightly different domain. Also, as expert systems become more pervasive there is a need to integrate their activities. Much of the work has to do with improving user interfaces (easier for nonexperts to use them), simplifying knowledge acquisition (learning), maintenance of knowledge bases (adding new knowledge without ruining the performance of the system and that is consistent with the old knowledge), and real-time monitoring (reasoning with time as a variable).

8. PROBLEM SOLVING APPLICATIONS OF AI

This section discusses some of the applications of AI. Although not meant to be exhaustive, it is intended to provide an idea of the areas where problems have been addressed by AI as well as the complexity of such problems. As indicated by the wide variety of applications, AI is indeed fulfilling its goal of wide applicability. In some domains, estimates of the search space (the number of possible answers) are given ranging typically from millions to $>10^{100}$. Again, this shows how AI can be used to reduce a search space from a size too large for blind search to one that can be heuristically examined to yield useful results.

8.1 Scientific Applications

Science provides a rich and challenging source of problems for AI. Indeed, science is an intelligent activity, so that AI has something to learn from the scientific method as well as being able to help scientists do science.

8.1.1 Physics. Although the physics community has been slow to incorporate AI techniques, there are nevertheless huge numbers of possible uses. For example, any large experiment in physics uses computers for data taking. As has been discussed above, expert systems provide a methodology for expert-level problem solving, such as automating the diagnosis of problems that occur with a detector or electronics. The Automated Beam Line Experiment (ABLE) (Clearwater and Lee, 1987) expert system has been used to locate and classify magnet and monitor errors in particle beam lines. Together with a knowledge-based induction system it has also been able to learn the diagnostic rules for classifying magnet and monitor errors in a beam line (Buchanan *et al.*, 1988) without the need for extensive knowledge engineering. Expert systems have also been used to diagnose problems in FASTBUS electronics (Booth and Carroll, 1987).

Other applications under way include learning rules ("cuts") for classifying events in high energy physics experiments. In this application the domain theory contains a list of parameters that are candidates for the cuts. As in any induction system, there is no *a priori* assumption that the domain theory contain all the relevant parameters or that it is free of irrelevant parameters. The training data for the learning program are preclassified examples (real or Monte Carlo) of the event type to be learned (signal events), as well as examples not belonging to the concept (background events). Because the data are noisy (e.g., misclassifications and finite resolution in the detector), the physicist sets the desired level of accuracy for the rules to be learned. The program then finds all the rules that satisfy the performance level given a domain theory. If certain parameters in the domain theory do not produce good enough rules, then those parameters may be left out of subsequent domain theories to reduce search. This procedure is similar to what physicists do anyway, except that now the process is formalized and the complicated hypothesis space is explored systematically rather than intuitively.

On a more down-to-earth level, the ISAAC program (Novack, 1977) uses a knowledge base of objects from elementary mechanics, such as masses and levers, to solve problems straight out of a college level physics text.

8.1.2 Biology. The recently begun Human Genome Project (National Research Council, 1988) and other related topics in molecular biology are excellent challenges for AI. The structures of biological molecules such as DNA are so complex that it may be necessary to use the techniques outlined in this article to learn the structure–functions relationship of sequences of amino acids in DNA. For example, properties of amino acids and the order in which they occur in proteins may be used to predict the location of three-dimensional structures such as helices in a protein, a precursor to predicting the overall properties of the protein. The utility of AI comes from the fact that at present it is very time-consuming to obtain these properties by conventional means.

8.1.3 Chemistry. The DENDRAL program (Lindsay *et al.*, 1980) was developed at Stanford University for interpreting mass spectrograph data from organic compounds. Meta-DENDRAL (Buchanan and Mitchell, 1978) was used to learn rules for determining the structure of organic compounds using the mass spectrograph as input and is one of the classic examples of an AI learning program. The input to both programs were properties of organic chemistry such as mass, valence, and stability, along with the mass spectrograph data. TQMSTUNE (Rauch-Hindin, 1986) was developed at Lawrence Livermore Laboratory to specify parameter settings for a triple quadrupole magnet spectrometer.

8.2 Technical Applications

AI has found great application in technical fields, particularly in the area of diagnosis. Diagnostic applications are relevant to physicists because diagnosis is a problem-solving activity that occurs over and over again when debugging experiments and in classifying processes in a detector. Several examples from different areas are given below. Additional and more detailed examples can be found in Feigenbaum, McCorduck, and Nii (1988).

8.2.1 Engineering and Industry. In the area of data interpretation, Schlumberger Inc. developed the Dipmeter Advisor System (Smith and Young, 1984) to interpret borehole data to aid oil prospecting. The program uses up to 650 geological categories in as many as 500 different depth intervals giving a search space of over 500^{650}.

Design systems, as exemplified by the following two examples, are relevant to scientists because today's experiments are so complicated that getting a good enough design is a huge task in itself that is still done largely by the "seat of the pants" approach. The XCON system (Rauch-Hindin, 1986) was developed to aid in the design of computer systems for VAX computers. The program had to select 50–150 items from a catalog of 20 000 parts giving a search space of over 10^{200}. Xerox Corporation developed the PRIDE (Mittal, Dym, and Morjaria, 1985) system which is used to design paper handling systems for copiers.

8.2.2 Medicine. Numerous systems have been developed to aid doctors in diagnosis of patient symptoms. The MYCIN (Buchanan and Shortliffe, 1984) system was developed to diagnose infectious blood diseases and has achieved expert diagnostician level performance during controlled tests. The INTERNIST (Schaffner, 1985) program was developed to aid in the diagnosis of internal disorders. Both MYCIN and INTERNIST reasoned in a space containing over a million possibilities.

8.3 Game-Playing

One of the earliest and still active areas of AI applications is in the realm of games. Very successful systems have been developed for many popular and complex games.

Recently, the Deep Thought (DT) system defeated an International Master at chess. DT consistently has a United States Chess Federation rating of 2500 in tournaments with humans. DT is currently (1990) the fastest chess machine and program and can analyze 700 000 moves per second. One of the classic results in AI and learning in particular comes from a checker playing program developed by Arthur Samuel in the late 1950s and 1960s (Samuel, 1963).

Waterman (1970) developed a poker playing program that learned how to play five-card draw poker. This program learned how to bet or fold depending on its hand, the amount of money in the pot, the last bet, and the bluffability and style of the opponent. The game of hearts has also been used as a test bed for learning by taking advice.

8.4 Defense

The U.S. Defense Department has long been a strong supporter of AI research. Recently this support has led to development of intelligent systems for use in military applications. Most of these applications are in the area of battle management, autonomous systems, intelligent analysis systems, etc. Some of the skills required for these applications include the ability to reason about scheduling, logistics, resource management, etc. Although many applications exist, only a small number are discussed here. Many military applications involve the same sort of pattern recognition skills needed in physics.

The Pilot's Associate (PA) is the U.S. Air Force's expert system for aiding fighter pilots. The purpose of PA is tactical planning during a combat situation. PA must take into account the threat environment, rules of engagement, currently available resources, etc., to allocate resources to provide overall tactics. All of these actions must occur in the real-time environment of combat.

The Force Requirements Expert System (FRESH) (Senator, Witte, and Kral, 1989) is currently in use by the Commander of the U.S. Pacific Fleet. FRESH is a knowledge-based system used to aid in the Navy's battle management decision making. Specifically, FRESH is used as an aid in resource and force employment allocation. The Navy reports higher readiness and higher accuracy of predictions with the use of FRESH.

The Autonomous Land Vehicle (ALV) is the U.S. Army's program to develop a vehicle that can perform tasks in the battlefield. Some of these tasks are smoke laying, deep reconnaissance, obstacle overwatch, mine laying, etc., all of which are performed without endangering soldiers and thus freeing them for other critical operations.

Part of the Strategic Defense Initiative (SDI) will require extensive use of AI for monitoring, targeting, and interpreting data from space-based devices as well as resource allocation.

8.5 Space Exploration

The National Aeronautics and Space Administration (NASA) is incorporating AI systems into its control rooms and future mission scenarios. As with the military systems, much of the work involves scheduling and resource allocation. One such space-based system is the Hubble Space Telescope whose limited viewing time must be optimally scheduled. Additionally, plans for a Mars Rover will involve many of the same problems of sensor interpretation and reasoning problems faced by the ALV.

8.6 Automatic Programming

Currently, computer programs must be laboriously developed by human programmers. This time-consuming and error-prone activity becomes even more critical as computer hardware leaps ahead in performance every few years. This "software gap" may be alleviated by the use of automated programming techniques.

Automatic programming tries to develop programs or systems that can automatically create computer-executable programs given some input of what to do, other than the code itself. Thus the goals of artificial intelligence may be thought of as some form of automatic programming. Automatic programming seeks to speed up the programming process by letting the programmer stylize and formulate the problem in a more abstract and natural way, unlike conventional programming languages that force the programmer to work within a rigid syntax and semantics in excruciating detail.

A familiar example of automatic programming comes from compilers, i.e., the source code of some programming language is converted into a form that is executable by a computer. More advanced automatic programming systems try to use natural-language input of a problem description that is then turned into executable code. Other methods use induction-like techniques of giving the automatic programming system examples of the desired input and output, as well as traces of the intermediate processing.

8.7 Mathematics

The AM program (Lenat, 1980) was designed to discover concepts in elementary finite-set theory. The program used empirical data and defined new concepts based on the regularities in the data. The data consisted of set theory concepts, operators, and heuristic rules for constructing plausible new concepts. While AM did not discover any new mathematics, it did find many well-known relationships such as DeMorgan's laws and led to the development of much shorter proofs for some laws of composite numbers.

The EURISKO program was an extension of the AM program and was designed to find useful heuristics and new representations in a variety of domains. EURISKO has been used in set theory, number theory, games, heuristics, device physics, etc. A couple of examples of the types of heuristics used by the program are these (Lenat, 1983):

If the results of performing f have not been useful, then lower the value of f.

If the results of performing f have sometimes been useful, then make new specializations of f.

8.8 Parallels with Human Perception and Language

8.8.1 Cognition. Much of the research in AI done by psychologists has been to model the human cognitive process. The idea is that information processing systems can thus be used to test cognitive models. This modeling includes models for knowledge representation and memory. Additional information on human problem solving can be found in Newell and Simon (1972).

8.8.2 Vision. Vision is included as an application of AI because it involves image interpretation, not just signal processing. Specifically, we define an image as a two-dimensional representation of an event in space. The image may be black and white, gray-scale, or color. Vision may be broken into three categories: signal processing—enhancing an image through a better signal-to-noise ratio with no regard to the content of the image; classification—where the classes are predetermined as in character recognition; and understanding—or scene analysis where the context of the entire image is important. One of the contributions of vision research has been the 2D sketch in which the surface knowledge is made explicit. Besides surface characteristics such as texture, reflectivity, etc., a vision system must be able to distinguish where one object stops and another begins, or where one side of an object changes to another side, i.e., edge detection.

AI vision systems might be useful in physics for interpreting spectra or event topologies in addition to being able to notice visible anomalies in a detector (e.g., a misalignment in an apparatus).

8.8.3 Language. Language is included as one of the applications of AI because it involves speech recognition and text understanding. Understanding language can be used to improve the performance of intelligent systems through better expression, recognition, and use of knowledge.

Understanding language means being able to interpret some language input and perform an action that is appropriate. For example, an appropriate response to "Where is your car?" might be "It's next to the big tree." Language understanding is difficult in part because of the complex interactions of the components of a sentence or group of sentences. For example, a mapping from a sentence may have many possible interpretations. Consider what must be known to understand a sentence. First, in understanding syntax, the order of the words must conform to the legal ordering specified by the language. Computer compilers are usually just syntax mappers; they do not know whether the program will work, only that there are no illegal statements. Second, the semantics of the sentence must have some meaning; that is, merely putting together a subject, verb, and object does not qualify a sentence as being semantically correct, such as "The car swims book." Finally, pragmatically the sentence must be interpreted; for example, "Did you drive to work today?" is usually interpreted to mean "Could I get a ride from you?" rather than as a yes/no question. Understanding not only includes individual words but sentences, paragraphs, etc.

An AI language system would provide a much more productive interface that could make programming and debugging much more efficient. As physics has become ever more dependent on extensive programming to run and analyze an experiment, a high-efficiency interface becomes even more critical.

8.8.4 Robotics. Robotics can be included as an AI application because it involves motion understanding. Robots must have both kinematic and dynamical knowledge which forms part of their spatial reasoning capability. Not only must a robot move through an environment, it must also have the ability to apply the correct amount of force, a tactile feedback. The robot must take into account the forces of motion such as centripetal and Coriolis forces as well as friction.

In science, an AI robot would be able to perform many mundane tasks around the experiment. In chemistry this could be mixing and performing actual experiments. In large physics experiments, the robot could go where people may not be able to because of high radiation and could possibly fix things without having to shut down a reactor or accelerator.

9. RELATION TO OTHER FIELDS

AI formalizes problem solving for a wide variety of domains. Intelligent systems such as evolutionary systems, where changes occur on a very long time scale, for example, species evolution (genetics), societal evolution, or national economies, are ill-structured and more difficult to parametrize and thus more suited to modeling using AI's methodology.

Operations research has made important strides in the last decade (Gill *et al.*, 1981), but these optimization programs work only with numbers and typically use some sort of hill-climbing search through the problem parameter space. They also make assumptions about the characteristics of the functions involved, such as invertibility, differentiability, continuity, etc., which in real life may be unknown or not applicable.

In statistics, Bayesian classification uses prior and posterior probabilities to represent less than certain data. The advantage of this technique is that it has a formal basis. The drawback is that it requires knowledge of distributions or other subjective estimates or requires a particular distribution to hold, which may not be the case in the real world. Recently, a nonparametric method of using trees, Classification and Regression Trees (CART) (Breiman, Olsen, and Stone, 1984), has been developed to do classification. These trees are similar in structure to the search trees discussed earlier and are constructed based on "best" partitioning the training data into two predefined classes. The nodes in the tree can be interpreted as antecedents of rules. The CART methodology removes the constraint of knowledge concerning the distributions in the data.

The areas of synergetics (Haken, 1983) and computational ecosystems (Huberman, 1988) have applications in almost any large complex dynamical system. These areas include the study of the behavior of such systems and their approach to stationary states, limit cycles, or chaos.

Cybernetics (*q.v.*) (Wiener, 1961) is the study of regulatory mechanisms. These systems are limited by their inability to take into account that the components of the system may have goals, nor do they require that the system be able to improve its performance (ability to achieve goals), only that it perform in an "optimal" way. In fact, closed feedback systems are subject to the second law of thermodynamics and therefore can only lose information and never gain. Although cybernetic systems may be adaptive, they are not so in a way that implies any understanding.

10. FUTURE DIRECTIONS INTO THE 21ST CENTURY

As has been discussed in this article, there are innumerable applications for AI. Almost any of them could be used as a grand challenge for AI—for example, the development of a machine and program that becomes the world champion chess player. Another example is a program that is able to help a handicapped person function more as a nonhandicapped person. Consider also machine systems that are able to self-organize and self-replicate, or to discover a new scientific result, or a truly natural language interface for communication between man and machine, or to write a poem.

By early in the next century, it can reasonably be expected that powerful computers will be readily available throughout the world. With the future advent of intelligent interfaces that can communicate through voice or handwritten character recognition, much of the knowledge of the world necessary for creating wealth can thus be made available to almost anyone who desires it (Reddy, 1988). Again this will have an enormous effect on global politics and the material well-being of all people.

Finally, we briefly summarize what AI has to offer to physics. On the one hand physics may seem like a straightforward approach to learning about nature because nature provides a ready test bed to test hypotheses. On the other hand, although the scientific method has been refined over the centuries it remains a hit or miss method whose progress relies on clever representations, heuristic search, and nonmonotonic reasoning, precisely the things that AI has to offer. Also, AI's emphasis on flexible programming methods facilitates construction of problem solving systems. In summary, the AI approach to problem solving offers a methodology for investigating situations where there is no known algorithmic solution by using heuristic knowledge to constrain the search for a solution.

11. APPENDIX: ADDITIONAL INFORMATION

11.1 Funding Sources

Traditionally, the Defense Advanced Research Projects Agency (DARPA) has been the largest funder of AI research over the years. Additionally, the Office of Naval Research (ONR), the National Aeronautics and Space Administration (NASA), the National Institutes of Health (NIH), and the National Science Foundation (NSF) are also large funders. Together these sources contribute over $100M/yr to research in artificial intelligence.

11.2 Centers of Research

Since the beginning of AI, three universities, Stanford, Carnegie–Mellon, and the Massachusetts Institute of Technology, have been at the forefront of AI research. These universities have faculty who were the founding fathers of the field as well as many of their disciples. Other schools with strong programs in AI include University of California at Irvine, University of California at San Diego, University of Illinois, University of Massachusetts at Amherst, University of Michigan, Rutgers University, University of Southern California, and University of Texas at Austin. Other centers include the industrial laboratories at AT&T Bell Laboratories, IBM Watson Laboratory, Schlumberger, Inc., and Xerox Palo Alto Research Center.

11.3 Selected Societies, Journals, Magazines, and Conferences

There are numerous societies involved in artificial intelligence and related areas. The most popular is the American Association for Artificial Intelligence (AAAI), which is open to anyone interested in AI.

Similarly, there are numerous publications available for those interested in the scientific and technical side of AI. *Artificial Intelligence, Machine Learning,* and *Neural Networks* are all scientific journals. The *AI Magazine* is a quarterly magazine which gives semitechnical accounts of recent work in AI, book reviews, announcements, etc. This a good source of information for AI products and upcoming AI events as well.

The annual meeting of the AAAI typically draws over 5000 people from universities, industry, and government from all over the world. The biennial International Joint Conference on Artificial Intelligence (IJCAI) is also a widely attended meeting and every fourth year cosponsors the AI conference with AAAI. The IEEE International Conference on Neural Networks and the International Conference on Genetic Algorithms are also offered. The AAAI and IEEE usually offer numerous other conferences, symposia, and workshops throughout the year catering to more specialized areas of interest.

GLOSSARY

AI: The abbreviation for Artifical Intelligence.

Computer: A machine that manipulates symbols.

Control Strategy: A strategy (set of rules) for directing the way a search will proceed.

Expert System: A program that is able to perform a task requiring human expertise.

Heuristic: A piece of knowledge that is used to reduce the amount of search in a problem, but is not guaranteed to work.

Knowledge: Facts about things, the way they interact with other things, etc.

Knowledge Base: Place where data structures and procedures used to manipulate and interpret knowledge are kept.

Learning System: A program that improves its ability to do some task without explicit programming.

Means-Ends Analysis: A control strategy that directs search in such a way as to minimize the difference between the current state in the search and the goal state.

Search: Exploration of a parameter space (also referred to as a hypothesis space or problem space) according to well-defined rules.

List of Works Cited

Barr, A., Cohen, P., Feigenbaum, E. (Eds.) (1982), *Handbook of AI*, Vol. 1. Los Altos, CA: Kaufmann.

Berliner, H. J., Ebeling, C. (1986), *Artif. Intell.* **28**, 1.

Booth, A. W., Carroll, J. T. (1987), *Comput. Phys. Commun.* **45**, 67–76.

Breiman, L., Friedman, L. H., Olsen, R. A., Stone, C. J. (1984), *Classification and Regression Trees*, Belmont: Wadsworth.

Buchanan, B. G., Mitchell, T. M. (1978), in: D. A. Waterman, F. Hayes-Roth (Eds.), *Pattern-Directed Inference Systems*, New York: Academic.

Buchanan, B. G., Shortliffe, E. H. (1984), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, Reading: Addison–Wesley.

Buchanan, B. G., Smith, R. G. (1988), *Annu. Rev. Comput. Sci.* **3**, 23–58.

Buchanan, B. G., Sullivan, J., Cheng T. P., Clearwater, S. H. (1988), in: *Proceedings of the Seventh National Conference on Artificial Intelligence, St. Paul, MN*, San Mateo, CA: Morgan Kaufmann, pp. 552–557.

Clearwater, S., Lee, M. (1987), in: E. R. Lindstrom, L. S. Taylor (Eds.), *Proceedings of the 1987 Particle Accelerator Conference, Washington, DC*, New York: IEEE, pp. 532–534.

Clocksin, W. F., Mellish, C. S. (1981), *Programming in Prolog*, Heidelberg: Springer.

Feigenbaum, E., McCorduck, P., Nii, H. P. (1988), *The Rise of the Expert Company*, New York: Times.

Gill, P. E., Murray, W., Wright, M. H. (1981), *Practical Optimization*, New York: Academic.

Haken, H. (1983), *Synergetics: An Introduction*, Berlin: Springer.

Hayes-Roth, F., Waterman, D. A., Lenat, D. (Eds.) (1983), *Building Expert Systems*, Reading: Addison-Wesley.

Holland, J. R., Holyoak, K. J., Nisbett, R. E., Thagard, P. R. (1986), *Induction*, Cambridge: MIT.

Huberman, B. A. (Ed.) (1988), *The Ecology of Computation*, Amsterdam: North-Holland.

Langley, P., Bradshaw, G. L., Simon, H. A. (1983), in: R. S. Michalski, J. G. Carbonell, T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*, Vol. 1. Palo Alto: Tioga, pp. 307–329.

Lenat, D. B. (1980), in: R. Davis, D. B. Lenat (Eds.), *Knowledge-based Systems in Artificial Intelligence*, New York: McGraw-Hill.

Lenat, D. B. (1983), in: R. S. Michalski, J. G. Carbonell, T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*, Vol. 1. Palo Alto: Tioga, pp. 276–282.

Lenat, D. B., Prakashand, M., Shepherd, M. (1986), *AI Magazine* **5** (4), 65–85.

Lindsay, R., Buchanan, B. G., Feigenbaum, E. A., Lederberg, J. (1980), *Applications of Artificial Intelligence for Organic Chemistry, The DENDRAL Project*, New York: McGraw-Hill.

Michalski, R. S., Carbonell, J. G., Mitchell, T. M. (Eds.) (1983), *Machine Learning: An Artificial Intelligence Approach*, Vols. 1 and 2. Palo Alto: Tioga.

Mitchell, T. M., Keller, R. M., Kedar-Cabelli, S. T. (1986), *Mach. Learning* **1** (1), 47–80.

Mittal, S., Dym, C., Marjaria (1985), in: C. L. Dym (Ed.), *Applications of Knowledge Based Systems to Engineering Analysis and Design*, New York: ASME.

National Research Council (1988), *Report of the Committee on Mapping and Sequencing the Human Genome, Board on Basic Biology, Commission on Life Sciences*, Washington, DC: National Academy.

Newell, A., Simon, H. (1972), *Human Problem Solving*, Englewood Cliffs: Prentice-Hall.

Nilsson, N. J. (1980), *Principles of Artificial Intelligence*, Los Altos, CA: Kaufmann.

Novack, G. (1977), in: *Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Cambridge, MA, August 1977*, San Mateo, CA: Morgan Kaufmann, pp. 286–291.

Rauch-Hindin, W. B. (1986), *Artificial Intelligence in Business, Science, and Industry*, Vols. 1 and 2. Englewood Cliffs: Prentice-Hall.

Reddy, R. (1988), *AI Magazine* **9**(4), 9–21.

Rumelhart, D. E., McClelland, J. (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vols. 1 and 2. Cambridge, MA: MIT.

Samuel, A. (1963), in: E. A. Feigenbaum, J. Feldman (Eds.), *Computers and Thought*, New York: McGraw-Hill, pp. 71–105.

Schaffner, K. F. (1985), *Logic of Discovery and Diagnosis in Medicine*, Berkeley, CA: University of California.

Senator, T. E., Witte, T. M., Kral, T. (1989), in *Proceedings of the Conference on Innovative Applications of Artificial Intelligence, Stanford University*, San Mateo, CA: Morgan Kaufmann, pp. 151–157.

Simon, H. A. (1981), *The Sciences of the Artificial*, Cambridge: MIT.

Smith, Young (1984), in: *Proceedings of the ACM Annual Conference, New York, October 1984*, New York: Association for Computing Machinery, pp. 15–23.

Steele, G. L. (1984), *Common LISP, The Language*, Bedford: Digital.

Waterman, D. A. (1970), *Artif. Intell.* **1**, 121–170.

Waterman, D. A. (1986), *A Guide to Expert Systems*, Reading: Addison-Wesley.

Wiener, N. (1961), *Cybernetics: or, Control of Communication in the Animal and the Machine*, Cambridge: MIT.

Winston, P. (1985), *Artificial Intelligence*, Reading: Addison-Wesley.

Zadeh, L. A. (1979), in: J. E. Hayes, D. Michie, L. I. Mikulich (Eds.), *Machine Intelligence 9*, Chichester: Ellis Horwood.

Further Reading

Anderson, J. R. (1983), *The Architecture of Cognition*, Cambridge: Harvard University.

Belzer, J., Holzman, A. G., Kent, A. (Eds.) (1981), *Encyclopedia of Computer Science and Technology*, New York: Marcel Dekker.

Harris, M. D. (1985), *Introduction to Natural Language Processing*, Reston: Reston Publishing.

Minsky, M. (1986), *The Society of Mind*, New York: Simon and Schuster.

Rich, E. (1983), *Artificial Intelligence*, New York: McGraw-Hill.

Shapiro, S. C. (1987), *Encyclopedia of AI*, New York: Wiley.

Waterman, D. A., Hayes-Roth, F. (1978), *Pattern-Directed Inference Systems*, Orlando: Academic.

ASTRONAUTICS

Donald Rapp, Charles E. Kohlhase, Jr., *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, U.S.A. (Editors)*

Charles E. Kohlhase, Jr., *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, U.S.A. (General Introduction)*

Brian K. Muirhead, Donald Rapp, *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, U.S.A. (Sec. 1)*

Kenneth L. Atkins, *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, U.S.A. (Sec. 2)*

Philip W. Garrison, *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, U.S.A. (Sec. 3)*

William G. Breckenridge, *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, U.S.A. (Sec. 4)*

Richard H. Stanton, *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, U.S.A. (Sec. 5)*

Lincoln J. Wood, *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, U.S.A. (Sec. 6)*

General Introduction 26
1. **Spacecraft Design Architecture and Principles** .. 27
1.1 Spacecraft Systems and Architecture 27
1.2 Spacecraft Subsystems 27
1.3 Design Requirements and Constraints 32
1.4 System Design Principles 33
1.5 System Configuration Drivers . 33
1.6 Future Missions 36
2. **Spacecraft Power Systems** ... 37
2.1 Functions of Spacecraft Power Systems 37
2.2 Energy Sources 37
2.3 Energy Conversion 38
2.3.1 Photovoltaic Conversion 38
2.3.2 Thermal-to-Electric Conversion 39
2.4 Energy Storage 41
2.5 Management and Distribution. 42
3. **Spacecraft Propulsion** 42
3.1 Introduction 42
3.2 Physics of Rocket Propulsion.. 44
3.3 Cold-Gas Propulsion Systems.. 44
3.4 Chemical Propulsion Systems . 44
3.5 Electric Propulsion Systems... 47
3.6 Advanced Propulsion Systems. 48
4. **Attitude Control** 48
4.1 Introduction 48
4.2 Attitude Sensing 49
4.3 Attitude Determination 51
4.4 Actuators 51
4.5 Controller 52
5. **Sensors for Spacecraft Control** 52
5.1 Introduction 52
5.2 Optical Sensors 52
5.2.1 Sun Sensors 53
5.2.2 Star Trackers 53
5.2.3 Horizon Sensors 54
5.2.4 Optical Target Trackers 54
5.3 Inertial Sensors 55
5.3.1 Accelerometers 55
5.3.2 Gyroscopes 55
6. **Navigation and Guidance** 57
6.1 Introduction 57
6.2 Modeling of Spacecraft Orbital Motion 57
6.3 Types of Navigational Data.... 58

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

6.4 Computation of Spacecraft Orbits........................ 60
6.5 Trajectory-Correction Maneuvers.................. 61
Glossary..................... 62
Works Cited................. 62
Further Reading............ 62

GENERAL INTRODUCTION

Astronautics deals with the design, engineering, science, and operation of vehicles intended for travel in space beyond Earth's atmosphere. As such, it provides the underlying technology that has led to many successful missions from Earth-orbiting satellites to lunar and planetary flybys, orbiters, landers, and rovers. Countries involved in exploring this exciting frontier include the United States, the Soviet Union, the European Space Agency consortium of member countries, Japan, Canada, and others.

The first three decades of space exploration came to a close with the epic Voyager Grand Tour mission, which dazzled the world with its flawless encounter with the remote Neptune kingdom in the late summer of 1989, but more ambitious missions were already under way or in their planning stages. These new missions will place great demands on the continuing evolution of performance improvements in the various technical disciplines comprising the astronautics field. The United States Space Exploration Initiative aspires to the landing of humans on Mars before another three decades have passed, but such an undertaking will depend upon available funding and may warrant that we first establish a Space Station, a Lunar Outpost, and a series of robotic precursor missions to the moon and Mars. In parallel with this challenging goal, a variety of other spacecraft will conduct more advanced explorations of the planets, their satellites, and the solar system's small bodies (asteroids and comets).

The 1990s will also usher in the era of the great Earth-orbiting observatories—the Hubble Space Telescope, the Gamma Ray Observatory, the Advanced X-Ray Astronomical Facility, and the Space Infra-Red Telescope. If successful, these sophisticated craft will enable astronomers to provide humankind with an unprecedented understanding of how the universe began, how it is unfolding, and how it might end. Where missions to the outer solar system will learn about worlds several billion kilometers away, the great observatories will probe the far reaches of the universe, billions of light years distant from Earth, gathering new knowledge about stars and galaxies, and about the physical and chemical nature of the stars and dust that pervade the cosmic sea.

All space missions require the same types of basic elements. In order to conduct the desired *science*, it is necessary to build a *spacecraft* capable of transporting the scientific payload to the destination, and provide the spacecraft with power, environmental protection, the proper instructions and orientation for performing measurements, and the necessary communications technology for transmitting the resulting data back to Earth. As the *launch vehicle* used to escape Earth's gravity well does not have an error-free guidance system, the process of *navigation* is necessary to deliver the spacecraft to a precise location at its destination, and to continue to control the flight path and accurately point the scientific sensors until the end of the mission. Until spacecraft have become entirely autonomous, a large complex of people, computers, communication lines, and tracking antennas at an *Earth base* must support the mission. Finally, for those nonrobotic missions involving the presence of a human crew, *life-support systems* are also required.

In the context of this article, the spacecraft and navigation elements will be described in some detail, with principal focus on the spacecraft design and the technical areas of power, propulsion, sensors, and attitude control. The Earth base, launch vehicle, science, and human crew elements will not be treated here.

1. SPACECRAFT DESIGN ARCHITECTURE AND PRINCIPLES

1.1 Spacecraft Systems and Architecture

A spacecraft is one part of a complex, highly interdependent set of systems that enable spaceflight, including the launch vehicle, launch facility, communications network, and control center. A spacecraft is defined by the mission-unique functions it performs, including

1. transporting the payload to a proper trajectory or orbit,
2. providing electrical power to the payload,
3. aligning and orienting the payload in space,
4. providing environmental protection and control to the payload and other supporting subsystems,
5. storing data, as necessary, from the instruments aboard, and
6. providing a communication link with Earth to receive instructions and to send back data.

Spacecraft can be categorized in terms of a number of different characteristics. The four most important characteristics are the following:

1. piloting (robotic or human),
2. mission type (e.g., orbiter, flyby, lander, rover),
3. attitude control (e.g., three-axis stabilized, spinning), and
4. propulsion (e.g., cold gas, chemical, electric).

Table 1 contains a selected listing of robotic and piloted spacecraft, mission descriptions, characteristics, masses, launch dates, and spacecraft-developing nations or agencies to illustrate the wide range and diversity of spacecraft designs that have been developed since the beginning of spaceflight in 1957.

The basic job of spacecraft system designers is to lay out the architecture of an electromechanical system that carries out the required mission objectives. This process requires that the spacecraft design include all the subsystems necessary to operate the spacecraft, taking into account all the interactions and constraints that are imposed by subsystems on one another. Such constraints include the fact that the communications antenna must be pointed toward Earth during periods of communication, the solar arrays must be pointed toward the sun, the radiators for heat rejection must be pointed toward space, and the payload must be aimed at its target and held steady. In general, the job of the spacecraft system designer involves integration of subsystems into a functional whole that minimizes launch mass and power consumption, both of which are almost always in high demand and short supply. The process of designing a spacecraft usually takes several years, during which time careful accounting of mass budgets and power budgets for every element of the spacecraft must be made, and great ingenuity used in optimizing the designs and interfaces.

1.2 Spacecraft Subsystems

To facilitate design and development, complex spacecraft are subdivided along specific functional lines into so-called subsystems. Some of the subsystems used in a typical robotic spacecraft developed by NASA's Jet Propulsion Laboratory (JPL) are as follows:

Attitude and articulation control. Maintains spacecraft and appendage stability and target pointing (see Sec. 5).

Adapter. The hardware interface between the spacecraft and the launch vehicle.

Cabling. The subsystem of wires, insulation, shielding, and connectors that interconnect the electronic subsystems. Cabling design is driven by such issues as electrical conductivity, grounding, electromagnetic interference shielding, and voltage drops.

Command and data handling. The command and data-handling subsystem provides the real-time or preprogrammed functions: decoding ground commands, distributing discrete and coded commands to subsystems, responding to interrupts generated on board; collecting, formatting, and managing science and engineering data (analog or digital); and the on-board clock. This subsystem consists mainly of electronic computers, memory, and input/output devices. Processor speed, quantity and access speed of memory, and fault-detection and -correction techniques are primary design considerations. For bulk data storage a magnetic tape recorder with various record and playback rates is the current state of the art.

Table 1. Selected spacecraft mission classes, masses, and launch dates (from Hart, 1987; Yenne, 1985).

Spacecraft	Mission	Characteristics	Mass (kg)	Date	Nation or agency
		Robotic Missions			
Mariner 4	First Mars flyby	Three-axis, cold gas	261	1964	U.S.
Luna 9	First Lunar landing	Three-axis, cold gas	1 538	1966	U.S.S.R.
Venera 4	First atmospheric probe	Three-axis, cold gas	1 106	1967	U.S.S.R.
Luna 17	First surface rover	Three-axis	5 700	1970	U.S.S.R.
Pioneer 10	First Jupiter flyby	Spin, chemical	258	1972	U.S.
Mariner 10	First gravity assist; first Mercury flyby	Three-axis, chemical	503	1973	U.S.
Viking	First Mars landing	Orbiter and lander, three-axis, chemical	2 991	1975	U.S.
Voyager 2	First Uranus and Neptune flyby	Three-axis, chemical	825	1977	U.S.
Ariane 1	Commercial launch vehicle	Three-axis, chemical	1 500 (ground to orbit)	1979–present	ESA[a]
Giotto	Halley's Comet flyby	Dual spin, chemical	575	1986	ESA
Galileo	First Jupiter orbiter and probe	Dual spin, chemical	2 700	1989	U.S.
Hubble Space Telescope	Visible observatory	Orbiter, three-axis, reaction wheels	11 400	1990	U.S.
		Piloted Missions			
Vostok	First piloted orbiter	Three-axis, chemical	4 713	1961	U.S.S.R.
Apollo L/M	First crewed lunar landing	Three-axis, chemical	43 900	1969	U.S.
Salyut	First space station	Three-axis, chemical	19 000	1971	U.S.S.R.
Space Shuttle	First reusable launch vehicle	Three-axis, chemical	100 000	1981–present	U.S.

---

[a] ESA = European Space Agency.

It is interesting to note that, due to contraints imposed by the environment (mainly radiation) and the need for reliability and limited available power, spaceflight computers lag far behind the terrestrial state of the art in speed and storage capacity (this is especially true for planetary missions). For example, each of the redundant processors launched on Galileo in 1989 was a 1975-vintage 8-bit machine rated at 0.015 million instructions per second (1 MIP=performance of a VAX 11/780) but only used 5 W of power and had 96 kbyte of RAM memory.

Contamination monitoring and control. A part of the design procedure that specifies appropriate materials, as well as venting and bakeout procedures to reduce the impact of outgassing of vacuum-condensible contaminating materials on critical surfaces and sensors. Contribution of propulsion modules to contamination is predicted by analysis of rocket-plume flow fields and calculating deposition rates. Plume shields are used to reduce this effect. Contamination sensors (usually quartz-crystal microbalances) are placed on board the spacecraft to record the buildup of contaminant layers on surfaces near sensitive areas in order to determine when and if it will be necessary to take action to mitigate contamination effects (including the use of heaters to drive off contaminant layers).

Electronic packaging. Electronic packaging is the technology by which the various digital and analog devices are packaged into high-surface-density, multilayer configurations. With present packaging techniques, the overall density is more limited by packaging of chips on boards than by packing of devices onto chips. For example, on the bus interface units for Mars Observer (to be launched in 1992), only 1.7% of the area of the circuit boards is covered with silicon devices; the remainder is used for chip housings, mountings, and connections. Such practices add greatly to the bulk and weight of spacecraft. In the case of highly optimized robotic spacecraft, packaged electronics represent on the average 30% of the dry spacecraft mass; however, the dry mass is often 20% to 50% of the total mass including stored propellants. Integrated-circuit technologies are changing rapidly toward smaller, higher-lead-count devices. Significant problems exist in mounting, attaching leads, inspecting, and flight qualifying these devices.

The combined spaceflight extremes of radiation, temperature, and vibration make the incorporation of state-of-the-art electronics a slow and difficult process. Soldered joints seem to lead to an endless series of problems for electronic packages on spacecraft. This is exacerbated by the practice of embedding electronics boards in plastic, which due to its coefficient of thermal expansion applies stresses to soldered joints. There are a number of new techniques being developed in industry that offer the promise of progressive increases in packaging density and higher speed. These new techniques offer great promise, but there are questions as to whether such boards are capable of being inspected, whether parts are replaceable, and ultimately how reliable they are.

Independent probe. Some planetary spacecraft carry an independent probe, penetrator, or lander that is released from the spacecraft and directed at a target (atmosphere or surface) for *in situ* measurements.

Power. The power subsystem deals with the processing, control, and distribution of power on board the spacecraft. Energy sources may include solar-cell arrays, radioisotope thermoelectric generators (RTGs), or batteries (see Sec. 3).

Propulsion system. The propulsion subsystem provides propellants and engines for mass-expulsion methods of trajectory changes and attitude control (see Sec. 4).

Mechanisms and pyrotechnics. Mechanical and pyrotechnic devices satisfy such open-loop functions as separation, latching and unlatching, deployment, scanning, etc. The nature of these devices generally makes them all candidates for single-point failures. For example, on Galileo, if a single pyrotechnic separation bolt or boom deployment pin puller had failed to actuate, the entire mission would have been lost at that instant. Design practices emphasize large strength or force margins, redundancy where practical (e.g., bearings within bearings), vacuum-compatible materials and lubricants, and special attention to thermal expansion and contraction effects. All mechanical devices are extensively tested to verify margins and lifetime, and intentional "fouling" tests are often conducted to gain extra confidence in their "bulletproof" performance.

Science payload. Scientific payloads perform both remote sensing and *in situ* measurements of planets and the interplanetary medium. Remote sensing requires precision pointing of optical instruments and antennas to do imaging, spectroscopy, and/or radiometry in the visible, UV, IR, and radio wavelengths. As a minimum, space missions usually include visible-wavelength cameras and often ultraviolet and infrared instruments. *In situ* instruments perform measurements on various magnetic fields, charged particles, and plasmas, including their distribution, charge, and energy. In the case of reentry probes (e.g., Venera 4) or surface missions (e.g., Viking, Lunakhod), direct measurements are made of atmospheric properties (e.g., pressure, temperature, humidity, particulates) and surface characteristics (e.g., elemental composition, chemistry, mineralogy). These instruments are usually developed separately from the development of the spacecraft itself, but they must be integrated into the spacecraft and the spacecraft designed to provide the instruments with the required articulation and pointing capabilities, appropriate power, proper thermal environment, and adequate data storage and processing.

Structure. The structure provides support and alignment of all flight subsystems and equipment during launch and flight. It also provides protection from natural and induced environments such as vibration, meteoroids, radiation, and electromagnetic interference. As can be seen from Table 2, the structure on a spacecraft tends to account for about 30% of the total dry mass. It should be pointed out, however, that for recent planetary and cometary orbiting spacecraft, the dry mass is typically less than that of the stored propellant aboard the spacecraft.

The primary requirement for the structure is to withstand the launch loads. For typical launch vehicles, there is a spectrum of force versus frequency with specific peaks in the low frequencies of 1–50 Hz. Generally, structural elements are designed to meet one of two criteria: strength or stiffness. A stiffness (frequency tuning) criterion is often used to minimize loads on individual resonant subsystems or instrument structures by shifting resonances away from regions of high input energy. For example, the Galileo RTG boom designer chose to shift primary modes to around 22 Hz. This reduced RTG vibration response to levels that the RTGs could withstand without internal damage.

Table 2. Typical distribution of mass on dry spacecraft (excluding liquid propellants) for solar system exploration.

Mechanical systems	Percentage of dry spacecraft mass
Structure	32
Electronics packages	32
Devices, sensors, and optics	14
Cabling	7
RTG, battery, and solar arrays	10
Propulsion (dry)	5

Because of the criticality of mass, materials with high specific stiffness (elastic modulus divided by the density) and high specific strength (yield strength divided by the density) are very important. With use of aluminum 6061-T6 as a reference standard, the following list shows the characteristics of some common and advanced aerospace structural materials.

Material	Specific stiffness	Specific strength
Al 6061-T6	1	1
Titanium	1	2.4
Beryllium	6	1.4
Graphite-epoxy composite	5	4.7
Metal-matrix composite	5.3	3

As the structural design of a spacecraft proceeds over its multiple-year developmental stage, a detailed finite-element model is maintained. For complex spacecraft, these models may have many thousands of mass degrees of freedom. To verify model accuracy, a full-scale *dynamic test model* of the spacecraft is built that has all the dynamic attributes of the real spacecraft but does not contain real electronic subsystems or instruments. This test model is extensively tested on the ground to verify the predictions of the computer model.

For new astronomical missions involving large telescopes, the issues regarding the structure may be very different. Such structures may weigh thousands of kilograms and deform significantly in a 1g field, necessitating special ground measurement and testing techniques. For deployable telescopes with large reflectors, the principal issue may be stiffness and thermal distortion when deployed in space, rather than stiffness and strength when stowed for launch. Techniques are being developed that use passive and/or active control techniques to maintain tolerances in the deployed state.

Telecommunications. The telecommunications subsystem provides communication to and from Earth for the purpose of vehicle tracking, range measurement, commanding, and telemetry. The telecommunications subsystem is comprised of antennas, transmitters, receivers, modulators/demodulators, and amplifiers. A key parameter that determines the capability of the communications subsystem is the number of bits per second (bps) that can be transmitted. The performance of a telecommunications subsystem is proportional to the transmitter power, the area of the transmitting and receiving antennas, and the square of the transmitting frequency. Over the past 25 years, there has been a trend toward higher frequencies for spacecraft telecommunications. The present state of the art is undergoing a transition from X band (8.4 GHz) to Ka band (32 GHz). Researchers are now developing much-higher-frequency optical communication systems with vastly increased capacity. The telecommunications subsystem involves data formatting and coding, data compression, modulation of the carrier signal, radio-frequency transmission, and similar steps in receiving. Spacecraft transmitters for planetary spacecraft have been limited to power levels of about 20 W. The current state of the art for long-distance data transmission rate is about 10^5 bps, from Jupiter range [5 astronomical units (AU)]. For new Earth-observing spacecraft planned for the Global Change Initiative of the late 1990s, NASA will require much higher data rates (up to 300 Mbit/s), and the telecommunications subsystems will be very challenging. (See TELECOMMUNICATIONS.)

Temperature control. The temperature-control subsystem maintains mechanical and electrical elements of the spacecraft within allowable operating and nonoperating temperatures. This includes meeting the special temperature requirements of components and subsystems such as batteries and cryogenic instruments. Temperature control of spacecraft must be effective throughout the prelaunch, launch, cruise, and operational phases of the mission. For hazardous conditions such as carrying propellants in the Shuttle, temperatures must be maintained within safe limits even in the event of a sudden mission termination.

The basic strategy for planetary spacecraft thermal design is the following:

1. minimize temperature sensitivity due to external and internal heat fluctuations,
2. emphasize passive (e.g., blankets, radiators, coatings) over active techniques (e.g., heat pipes, fluid loops, thermostatically controlled heaters), and
3. keep the design simple for ease of analytical modeling and ground testing.

The process of thermal control begins with the conceptual design of the spacecraft. A sophisticated thermal model including solar inputs, planetary infrared and visible heat loads, heat sources and sinks, conductances, capacitances, and radiation view factors is used for steady-state and transient analyses.

Temperature-control design tools include selection of absorption and emissivity properties of surfaces and coatings, thermostatically actuated low-emittance louvers that can open to expose a high-emittance heat-rejection surface, multilayer (typically 5–20 layers of aluminized Mylar separated by a Dacron net) insulation blankets, electrical heaters, radioisotope heater units (RHUs), heat pipes, radiators, shades, stored cryogens, and active refrigerators.

In recent years, greater emphasis is being placed on instruments making infrared wavelength measurements. Such instruments require cryogenically ($\sim$55–80 K) cooled sensors and optics. In low Earth orbit, radiators can only cool to about 150 K. For lower temperatures, cryostats have been developed that can store 80–200 L of liquid helium for as long as a year [e.g., IRAS (Infra-Red Astronomical Satellite) and COBE (Cosmic Background Explorer) missions], and much larger cryostats are now being planned. A number of mechanical coolers using split Stirling cycles are under development for longer life missions. By 1995 they are expected to be able to provide about 1 W of cooling in the 50–80-K temperature range for <30 W of electrical power.

Some missions have very severe requirements for temperature-control technology. For example, the Galileo mission involves three gravity assists, one by Venus and two by the Earth. In the Venus gravity assist, the spacecraft approached within $\sim$0.7 AU of the sun, and achieves its final orbit around Jupiter at ~5 AU. The solar input varies by a factor of ~50 during this extended journey. Despite the huge changes in solar heating, most of the sensitive electronics and instruments had to be kept near room temperature throughout the mission.

1.3 Design Requirements and Constraints

Space-vehicle design is set apart from the design of other electromechanical systems by simultaneous application of the following extreme conditions:

1. the demanding performance requirements and constraints of space missions,
2. the severe environments of space and the launch vehicle,
3. the general permanent inaccessibility for repair, and
4. the extreme lack of political tolerance for *any* failure.

Item 4 drives the entire spacecraft design attitude and creates a factor of 100 or more higher cost than would be required to produce equally complex terrestrial systems.

Spacecraft functional requirements and constraints evolve through an iterative process involving mission and science objectives, performance objectives, and resource tradeoffs. The subsystem-level functional requirements include specific statements about the following: performance, configuration, interfaces, lifetime, safety for personnel and hardware, alignments, compatibility, accuracies, etc. In order to meet the functional requirements, a number of key parameters must be controlled and managed as part of the functional design of a spacecraft: mass, power, volume, schedule, on-board data rates, spacecraft-to-ground data rates, computer memory, computer processor speed, propellant, design margins, surface area, inertia properties, operating and nonoperating temperature ranges, and cost.

Design environments for spacecraft are specified for particular mission phases such as prelaunch ground operations, launch, flight, and, for some missions, operations in and on the atmosphere or surface of another body. Natural environmental hazards include ionizing radiation, heavy ions, nuclear radiation, magnetic fields, charged particles, meteoroids, atmosphere, planetary ring particles, temperature, and pressure. For example, the Galileo spacecraft electronics were designed (with a factor of safety of 2) to survive an integrated mission dose of 75 krad from Jupiter's radiation belts. Most interplanetary electronics need only be designed to 7.5 krad. Induced environments are associated with manufactured or vehicle-related phenomena such as structural vibration, random vibration, acoustic noise, pyrotechnic shock, electrostatic discharge, space debris, and radiofrequency radiation. Launch vibration is typically the driving structural design consideration.

Once a spacecraft leaves the launch pad, there is no opportunity to repair it (a notable exception is the unique capability of the Space Shuttle to capture and repair some vehicles from low Earth orbit, as was done in 1984 with the Solar Maximum Mission). Conservative design practices, quality control, and preflight testing are the primary tools to avoid flight failures. Reliability is also enhanced by utilizing block, functional, and alternate-mode redundancy such that individual element failures, called single-point failures, cannot cause significant loss of data or of the mission. However, reliability can only be enhanced at the cost of other resources: time, money, mass, memory, power, and performance.

Advanced technology plays a critical role in spacecraft design and performance. Advances in materials, electronics, and sensors have enabled giant steps forward in subsystem and component technologies, which have, in turn, enabled more and more sophisticated missions. However, due to the demands of space reliability and qualification, the technology level in space missions generally lags the terrestrial state of the art by many years. Space-radiation environment effects severely restrict the selection of acceptable electronic components. For example, the processor being proposed for the Comet Rendezvous Asteroid Flyby (CRAF) mission (development started in 1989) has an effective speed of only 0.25 MIPS, a factor of 10 or more below the state of the art for desktop computers on Earth. However, it is radiation resistant and immune to single-event upsets.

1.4 System Design Principles

The design of spacecraft is an evolutionary process, paced by the inflexible demands of a fixed launch date for solar system exploration missions. Project development is an iterative cycle rather than sequential. The design proceeds not in an unbroken flow from functional requirement to design requirement, to design to finished product, but as a series of design cycles as both requirements and interfaces evolve.

Principles of spacecraft system design can be enumerated, but the reality imposed by parallel requirements and design evolution results in many exceptions to these principles. The following items are guiding principles in each cycle of the recursive design–requirements iteration loop:

1. Define the mission and subsystem requirements and constraints as early as possible, preferably before starting design.
2. Establish and manage margins of critical resources (mass, power, memory, etc.), and make the design robust to accommodate uncertainties and last-minute changes.
3. Identify major problems early and search for designs that avoid them instead of using brute-force solutions.
4. Maximize performance, quality, reliability, adaptability, and precision within the constraints.
5. Utilize proven technology.

Optimum performance is obtained by tightly integrating many functional elements. In a highly optimized design, there are no "clean" interfaces between elements, and changes made by one element tend to strongly affect others. For example, to achieve minimum mass of electronics and structure, JPL utilizes an electronics subchassis to stabilize the primary bus structure. In contrast, the separation of elements such as electronics units into discrete, stand-alone components can lead to lower cost, easier modification late in the design cycle, and shorter development schedules, but all at the cost of performance, especially mass. For example, mounting "black boxes" of electronics on a structural plate and cabling them together may be simpler from an interface point of view, but it is not mass efficient. The key to a good system design is optimization in selected areas where performance is crucial and only a few elements are integrated or interactive.

When subsystem development must occur in parallel, many subsystem design parameters must be simultaneously managed at the system level to assure meeting overall system requirements. This process generally requires establishing budgets for critical resources and making allocations to the various subsystems. The two most difficult resources to manage are mass and power. Very early in the design process, the total mass is established by the trajectory requirements and the estimated launch vehicle performance. Mass allocations are made with an adequate contingency and reserve to accommodate inevitable growth and possible shortfalls in launch-vehicle performance. Invariably, problems arise, changes occur, trade-offs are made, and the total allocated mass rises from the initial estimate, while the contingency and reserve diminishes.

1.5 System Configuration Drivers

The spacecraft configuration provides the primary constraints on the design of subsystems that integrate the spacecraft, such as the structure, thermal control, electronics packaging, cabling, and mechanisms. Because the configuration defines all the basic geometric relationships and many of the physical relationships between the various subsystems and the payload, it is driven by a number of competing functional requirements and interfaces. The basic configuration requirements of the Voyager and Galileo spacecraft provide a good illustration because even though they were essentially the same, different emphases between them resulted in vastly different configurations for the two spacecraft. The configuration drivers for both missions are summarized below:

- stable platform for imaging;
- maximum field of view for field and particle instruments;
- large high-gain antenna integration and pointing;
- radioisotope power source;
- propulsion tankage and location of attitude control thrusters;
- instrument shielding from thruster plumes,

stray light, and RTG heat and radiation;
- compatibility with launch-vehicle envelope and dynamic loads;
- inertia properties balancing;
- spacecraft fabricability, maintainability, and transportability.

Figures 1 and 2 show the configurations of the Voyager and Galileo spacecraft, respectively, in free flight.

The Voyager spacecraft was designed as a flyby mission of Jupiter and Saturn by two spacecraft, but the extended mission of one

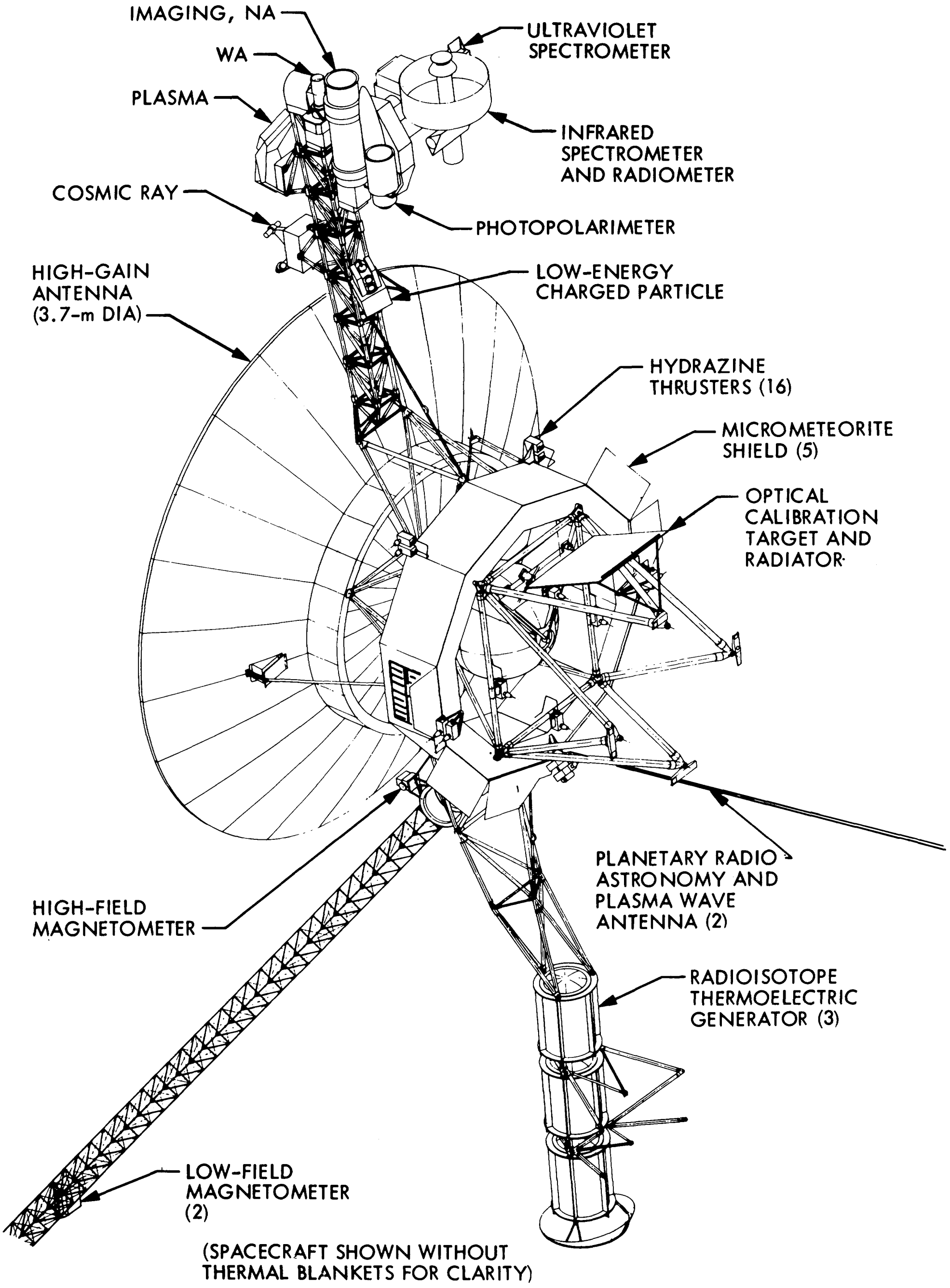
IMAGING, NA
WA
PLASMA
ULTRAVIOLET SPECTROMETER
INFRARED SPECTROMETER AND RADIOMETER
COSMIC RAY
PHOTOPOLARIMETER
HIGH-GAIN ANTENNA (3.7-m DIA)
LOW-ENERGY CHARGED PARTICLE
HYDRAZINE THRUSTERS (16)
MICROMETEORITE SHIELD (5)
OPTICAL CALIBRATION TARGET AND RADIATOR
PLANETARY RADIO ASTRONOMY AND PLASMA WAVE ANTENNA (2)
HIGH-FIELD MAGNETOMETER
RADIOISOTOPE THERMOELECTRIC GENERATOR (3)
LOW-FIELD MAGNETOMETER (2)
(SPACECRAFT SHOWN WITHOUT THERMAL BLANKETS FOR CLARITY)

FIG. 1. Voyager deployed configuration.

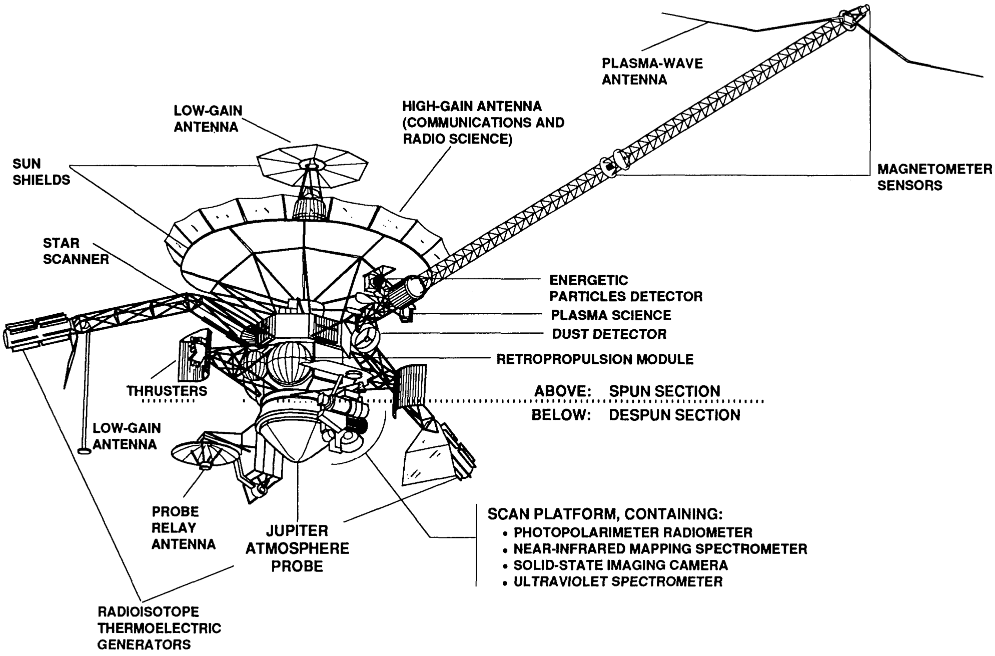
PLASMA-WAVE ANTENNA
LOW-GAIN ANTENNA
HIGH-GAIN ANTENNA (COMMUNICATIONS AND RADIO SCIENCE)
SUN SHIELDS
MAGNETOMETER SENSORS
STAR SCANNER
ENERGETIC PARTICLES DETECTOR
PLASMA SCIENCE
DUST DETECTOR
RETROPROPULSION MODULE
THRUSTERS
ABOVE: SPUN SECTION
BELOW: DESPUN SECTION
LOW-GAIN ANTENNA
PROBE RELAY ANTENNA
JUPITER ATMOSPHERE PROBE
SCAN PLATFORM, CONTAINING:
• PHOTOPOLARIMETER RADIOMETER
• NEAR-INFRARED MAPPING SPECTROMETER
• SOLID-STATE IMAGING CAMERA
• ULTRAVIOLET SPECTROMETER
RADIOISOTOPE THERMOELECTRIC GENERATORS

FIG. 2. Galileo deployed configuration.

spacecraft was to achieve a "Grand Tour" including Uranus and Neptune. The configuration was dominated by a 10-bay electronics bus, a science boom, an RTG boom, a payloadof 10 science instruments, and a 3.7-m-diam fixed high-gain antenna capable of transmitting 115 kbit/s to Earth from Jupiter (Anonymous, 1979).

Like all of JPL's previous spacecraft, Voyager was a three-axis stabilized design. Thrusters were located to provide fully redundant torques about three orthogonal axes. A small propulsion module (104 kg of hydrazine) was located in the middle of a 10-bay electronics bus for attitude-control and small trajectory-correction maneuvers.

Voyager's mission deep into the solar system required the use of three RTGs. Each RTG utilized a 2500-thermal-watt PuO_2 heat source and silicon germanium thermoelectric devices to produce 150 W electrical power (4 W/kg) at launch (Anonymous, 1979). Because of the health risks associated with plutonium, significant efforts were made to ensure that, in the event of a launch-vehicle or spacecraft failure, each of the PuO_2-carrying heat sources would return safely to Earth.

A few failures occurred on the Voyager spacecraft. Voyager 2 lost one of its two telecommunications receivers (possibly due to a blown fuse), and the other receiver required special control procedures to maintain 0.25 °C temperature stability. This stability requirement was needed to enable ground controllers to predict the uplink frequency at which Voyager 2 could be commanded. Voyager 2 also had its scan platform actuator temporarily stall after passing through the Saturn ring plane (the failure was not related to the ring crossing). Thorough analysis and ground simulation indicated that the azimuth actuator probably lost lubricant. Procedures were implemented that limited operations ($\frac{1}{12}$-deg/s slew rate and only a few total rotations of the actuator output), and no additional stalling occurred.

The Galileo spacecraft is in reality two spacecraft: an orbiter and a Jupiter atmospheric probe. The configuration is dominated by an 8-bay electronics bus, two RTG booms, a 9.5-m science and magnetometer boom, a 4.8-m-diam unfurlable high-gain antenna, a probe (with atmospheric science instruments), a 935-kg (wet) propulsion module, a scan platform, four bays of electronics opposite the scan platform, and nine science instruments (Anonymous, 1988).

The fundamental distinction between the designs of the Galileo and Voyager spacecraft was in the choice of Galileo's approach to satisfy field and particle science viewing requirements, by spinning part of the spacecraft. In this dual-spin concept, part of the Galileo spacecraft spins, allowing nearly a full 4π-sr field of view for its field and particle instruments, while the remainder of the spacecraft maintains an inertial pointing platform for remote sensing instruments on the despun section. In order to provide vehicle stability and to meet antenna and scan platform pointing requirements, the spinning section was designed to have a polar moment of inertia 1.16 times greater than the maximum transverse moment of inertia of the entire spacecraft. This resulted in using long booms for the RTGs, actively positioning the RTGs in flight to adjust products of inertia, adding 31 kg of ballast, and adding a tuned nutation damping mechanism to the magnetometer support (making it the largest nutation damper ever built). By contrast, Voyager, with its three-axis stabilized design, was simpler to design and build, but it did not provide as good a field of view for field and particle instruments as did Galileo. These two very different spacecraft provide insight into the conflict between the desire for better performance (more science) and the engineering design complexity that it produces.

1.6 Future Missions

Space telescopes are spacecraft in which the optical system, particularly the primary reflector, tends to dominate the design requirements. During the decade of the 1990s, NASA's Astrophysics Division is planning to launch four "Great Observatories." The first of these is the Hubble Space Telescope (HST), a reflecting telescope with a 2.4-m aperture and an extremely high-quality surface mirror. The limitations of the envelope of the Space Shuttle cargo bay (4.5×18 m^2), the light baffle forward of the secondary mirror, and the space required by instruments behind the primary mirror dictated a very compact Cassegranian optical system for the telescope itself. The primary mirror is mounted on a graphite-epoxy truss structure that acts as the main optical bench. The remainder of the spacecraft provides the normal spacecraft functions including large solar arrays that generate about 4 kW of power. The instruments of the HST are designed for periodic changeout and refurbishment.

Another of the Great Observatories is the Space Infra-Red Telescope Facility (SIRTF), scheduled for launch in the late 1990s. This 0.9-m-diam telescope will operate in the infrared region and will be cooled to an internal temperature of 1.7 K by immersing the optics and detectors in a toroidal cryostat filled with superfluid helium. The spacecraft will be deployed in high Earth orbit at an altitude of about 100 000 km to reduce the Earth thermal input to the cryostat and to improve the viewing efficiency of the telescope. The cryostat will hold about 4000 L (vs 200 L for IRAS, its predecessor) of superfluid helium and is to be designed for a five-year life. Two of the major design challenges are to develop a structure that will maintain its dimensional stability and to develop mechanisms that will allow active mirror pointing at these very cold temperatures.

The two other Great Observatories are the Gamma Ray Observatory and the Advanced X-Ray Astronomical Facility.

Toward the end of the century, it is hoped that construction will begin on a far-IR and submillimeter observatory of about 20 m diameter, called the Large Deployable Reflector (LDR). The reflector on this telescope will be shaded and will reach a passive temperature in the range 100–150 K in high Earth orbit. Unlike other space telescopes, which are small enough to be designed to be rigidly assembled on the ground and launched as a unit, the LDR is inherently much larger than any current launch vehicle cargo bay. One of the major challenges is to develop the capability to either deploy or construct the telescope reflector in space. In addition, the design must include active and passive figure control techniques to maintain dimensional stability within a few microns in the face of thermal gradients produced by the nonuniform solar inputs.

A Mars rover and sample return mission has been under study by NASA for a number of years. Early concepts include a moderately autonomous surface rover (rolling and walking designs are being studied), collecting about 5 kg of samples for *in situ* analysis, and return to Earth. This mission requires some unique techniques such as "aerobraking" in the Martian atmosphere to slow the spacecraft down and achieve orbit. This technique requires an "aeroshell" to protect and control the spacecraft's flight through the atmosphere but saves thousands of kilograms of propellant. Such a mission requires many elements, including a communications orbiter, the aeroshell system, the rendezvous vehicle, the lander, the ascent vehicle, and the Earth return vehicle.

2. SPACECRAFT POWER SYSTEMS

2.1 Functions of Spacecraft Power Systems

Spacecraft power systems (SPS) provide electrical power for the operation of a spacecraft. A SPS consists of an energy source (or sources), a system for conversion of energy to electrical power, energy storage (if necessary), and a system for management and distribution of electrical power on the spacecraft. The electrical power must be conditioned for particular end users, distributed to users (loads), made stable, and safely controlled to protect the entire space system from power failure. These capabilities must be arranged in an architecture allowing inclusion of the SPS within a limited volume, while supplying the required electrical power at the user load. In addition, the "noise" or electromagnetic signals carried on the power lines must be controlled to minimize impact on and disruption of sensitive scientific sensors. This entire "orchestra" of component players must then be directed. This is accomplished by a set of computer instructions, or programs, resident either in the spacecraft command computer or a special computer embedded in, and thus part of, the SPS.

2.2 Energy Sources

There are really only two major sources of energy on spacecraft, solar radiation and nuclear energy. Batteries can be used to augment these power sources for temporary periods. However, batteries do not have the capacity to supply the long-term power needs of spacecraft.

At a distance of 1 AU from the sun, the solar intensity is 1368 W/m^2. Solar radiation energy is readily available as long as a spacecraft stays within about 2 AU of the sun. The solar intensity, however, falls off as the square of the distance from the sun, $E \sim 1/R^2$, and at distances from the sun of more than a few AU, use of solar energy becomes impractical. In most cases, it is difficult to avoid having the spacecraft periodically "shadowed" or eclipsed by a planet or moon. This makes it necessary to include an energy storage device, such as a battery, or to include a nuclear energy source in those missions that must place the spacecraft in shadow or in regions of the solar system where the sunlight is very dim.

Nuclear sources take two primary forms, based either on fission or on radioactive decay of unstable isotopes. In either instance the energy is provided as heat. Energy is released by the fission reaction in the form of heat, electromagnetic radiation (e.g., γ rays), and other energetic particles. The γ rays are extremely energetic and can cause damaging ionization to materials in their surroundings. Because fission reactors generate large amounts of ionizing radiation, the shields to protect sensitive components or humans will tend to be massive.

Another alternative involves radioactive decay of unstable isotopes. In early applications on Earth, strontium-90 (^{90}Sr) and cesium-137 (^{137}Cs) were attractive because of their long half-lives and their ready availability as fission products from nuclear reactors. However, both ^{90}Sr and ^{137}Cs are strong γ-ray emitters and therefore require heavy shielding, which detracts from their value as energy sources on spacecraft. However, plutonium-238 (^{238}Pu) is an emitter of lower-energy α particles (stopped by a sheet of paper) and emits very little γ radiation. This, coupled with its relatively long half-life and high melting point ($\sim$2300 K), makes it the isotope of choice, even though it is expensive and has a much lower power density (4 W/g) than isotopes of cesium, cobalt, and polonium. Polonium-210 (^{210}Po) was used by the U.S.S.R. to power their moon-sampling Lunakhod. It has a very high power density (140 W/g) and a low radiation field, making it an ideal energy source for short missions. However, the half-life is only 138 days, making it unsuitable for long, deep-space missions.

Since the beginning of space exploration, nuclear SPS have included both radioisotope and fission reactor options. A series of navigational satellites were the first to employ nuclear power in space, beginning in 1961. They used an RTG in which some of the heat released by a radioactive material was converted to electricity by thermoelectric converters. These early RTGs provided 2.2 W of electrical power and weighed 2.3 kg each. Although their capability was modest, they performed well and raised confidence that RTGs could meet future mission lifetime requirements.

A spacecraft carrying an experimental fission reactor was successfully launched into orbit in 1965. After 43 days of producing 600 W of power, a spacecraft malfunction (having nothing to do with the reactor) caused the reactor to shut down automatically. A twin of this reactor accumulated 10 000 hours at power in a ground test, until it was shut down in 1966. The U.S. continued development of reactors at higher power levels (30 to 60 kW) for use in space, until all space reactor programs were terminated in 1973. In the mid-1980s a new U.S. space reactor program was begun, called the SP-100.

Beginning with their first publicly identified launch of nuclear power in 1965, the Soviet space program has actively developed small space reactors. They have launched about 35 reactors into low Earth orbit, primarily for military reconnaissance spacecraft. Several of these have accidentally reentered the Earth's atmosphere, the most notable being the reentry over Canada of the Radar Ocean Reconnaissance Satellite (RORSAT) known as Cosmos 954. This reentry in January 1978 focused world attention on the environmental consequences of the use of nuclear power in space. Although the Soviets have used nuclear power more extensively than the U.S., the U.S. continues to use RTGs for deep-space missions.

2.3 Energy Conversion

As shown previously, the sources of energy available for SPS are nuclear-generated heat and solar radiation. As with all energy sources these two forms must be converted to the form that is most useful on a spacecraft, namely electricity. (See ENERGY CONVERSION.)

2.3.1 Photovoltaic Conversion. For spacecraft that operate within about 2 AU of the sun, the most common SPS use arrays of photovoltaic cells to directly convert sunlight to electrical power (see PHOTOVOLTAIC DEVICES). In some ways, photovoltaic conversion of solar luminosity to electrical current is analogous to the photoelectric effect, which produces ejection of electrons from metal surfaces by photon impact. In semiconductor materials such as silicon, the impact of photons releases electrons that flow in opposite direction to "holes" (electron vacancies) within the material, creating an electric potential. Small wafers of such semiconductor materials, or "cells," are used in photovoltaic-space-power systems to convert incoming light energy into electrical power, *W*, with

$$W = I\eta,$$

where *I* is the incident solar-radiation intensity and η is the efficiency of the photovoltaic conversion cell. Typical conversion efficiencies in the range 10%–15% are now routine. This means that one can obtain roughly 150 W of electrical power for each square meter of solar cells at 1 AU from the sun.

To supply power to a spacecraft, arrays of large numbers of solar photovoltaic cells are integrated on lightweight structures called "substrates." The design of the substrates must be such that the entire mechanism can be stored in a small volume and withstand the launch loads imposed by the launch vehicle. After launch, the solar array is deployed and oriented toward the sun to begin producing electricity. The electrical power generated by hundreds or thousands of solar cells on an array must be collected and conducted to the electrical power bus, and usually to a storage-battery system. Each solar cell has an interconnect grid to connect its voltage to the next cell and, at the end of a series string, to a terminal or tie point. Most arrays include some isolation devices (diodes) for circuit fault isolation and to prevent nonilluminated cells from dissipating power.

The primary issues in solar-cell-array development are the conversion efficiency, the areal density, the degradation resulting from exposure to natural radiation in space, and methods for stowing and deploying arrays. These combine to provide the main figure of merit for array performance, specific power, typically given in W/kg. Typical performance of current solar arrays with rigid substrates is in the range 33 to 46 W/kg. A large experimental 12.5-kW flexible substrate "wing" was tested in space with a specific power of about

66 W/kg (see Fig. 3). The Advanced Photovoltaic Solar Array (APSA) now under development has advanced cells ($\eta \approx 13\%$) and composite, low-mass structures, with an estimated potential specific power of $\sim$ 130 W/kg.

2.3.2 Thermal-to-Electric Conversion. Primary energy sources that produce heat, such as nuclear fission reactors or radioisotopes, require a system to convert the heat into electricity. There are basically two techniques available, one being dynamic and the other being static.

Systems for dynamic conversion of heat to electricity in space build on our terrestrial experience in electric power generation using thermal power cycles. On Earth, boilers are heated by fossil-fueled or nuclear heat sources to produce steam, which is passed through a turbine to drive an electric generator in a *Rankine cycle*. The theoretical conversion efficiency of this cycle is determined by the temperature at which heat is absorbed by the working fluid in the boiler, and the temperature at which excess heat is rejected in the condenser. The actual efficiency is degraded below the theoretical efficiency by inevitable losses in the converting machinery. On Earth, excess heat unable to be converted is simply dumped into the atmosphere or ocean.

Proposed dynamic space-power-conversion techniques use variants of this technique in a variety of possible thermodynamic cycles, including the organic Rankine, Brayton, and Stirling cycles. Such cycles have the potential for conversion efficiencies of over 20%. A SPS must include radiative heat rejection (no air or water heat sinks) and redundancy to increase reliability. Such dynamic power-conversion systems have not been actually used in space for several reasons. One reason is that they tend to be complex and involve mechanical systems with moving parts, which are considered to be problematical for systems that require ex-

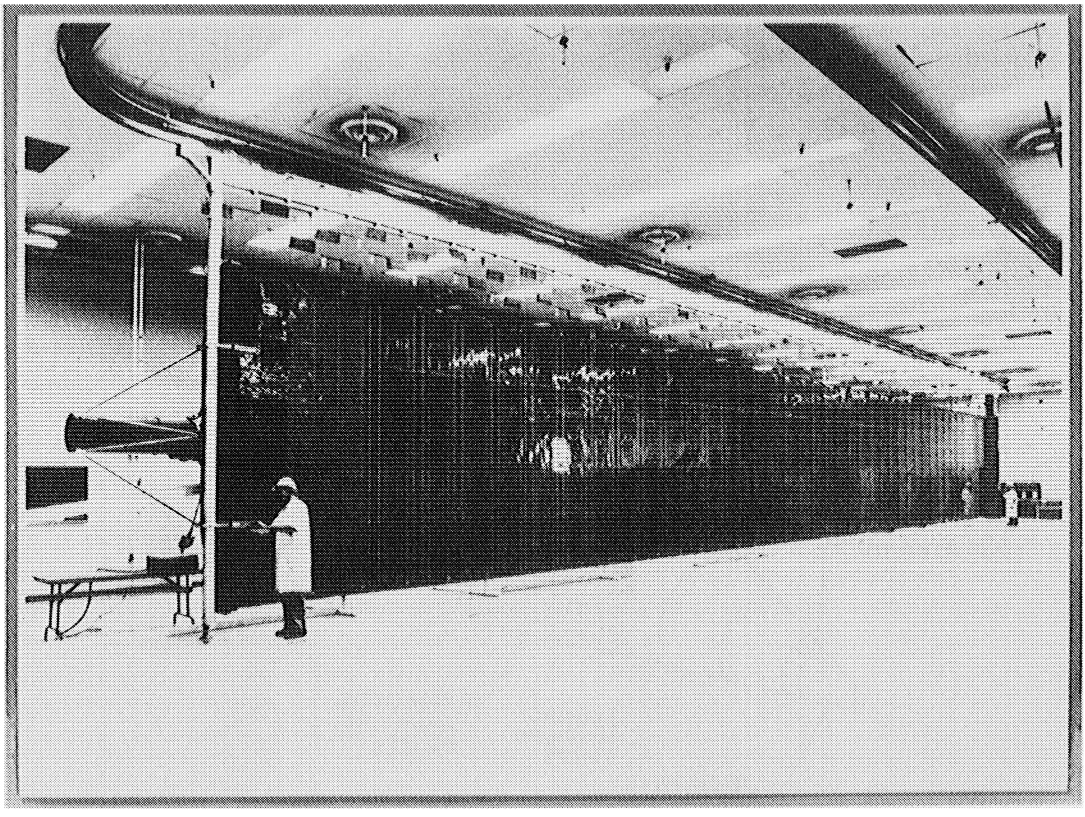

FIG. 3. NASA 12.5-kW SEP array wing.

tremely high reliability. A second reason is that the bulky and heavy nature of the components used in these systems (turbines, pumps, compressors, heat pipes, radiators, etc.) results in a poor specific power for systems that produce less than a few kW. For such low power levels, thermoelectric converters are the preferred choice because of their high reliability and low mass, despite their low conversion efficiency.

The second conversion technique does not employ moving machinery components, but instead, uses the phenomenon of temperature-induced current flow in materials (Seebeck effect) (see THERMOELECTRIC DEVICES). In this system, a pair of dissimilar semiconductor materials in contact, with each end at different temperatures, is joined in a closed circuit to produce a voltage. This type of system is called a thermoelectric converter. Unfortunately, their conversion efficiency is quite low (typically 5%–8%) and therefore the amount of power they produce is limited for a given thermal input. For thermoelectric SPS that must produce more than several kW of electrical power, such as SP-100, this leads to a requirement for large radiators that dominate the configuration of SPS because the large amount of excess heat not converted to electricity must be radiated into space (see Fig. 4 for the SP-100 reactor that uses this technology). It should be noted that the Seebeck effect is also used with metals, e.g., as thermocouples for temperature sensing.

For relatively low-power SPS (<1 kW of electrical power) the radiator penalty can be tolerated. For example, the RTG used to power today's crewless planetary spacecraft uses thermoelectric converters to provide about 285 W of electric power from heat released by radioactive isotopes. Galileo (the U.S. planetary exploration mission now in flight on its way to Jupiter) uses two such RTGs. RTGs utilizing thermoelectric converters are the standard power source for spacecraft that travel beyond Mars in the solar system. A cutaway drawing of an RTG is shown in Fig. 5. This RTG is a slightly modernized version of those that powered the successful Voyager spacecraft during its decade-plus tour of the solar system. RTGs were used on Ulysses and are planned for use on the CRAF Cassini spacecraft.

The focus of most U.S. space missions has been on applications that require typically less than about 1 kW. As a result, the energy sources have typically been RTGs and solar-

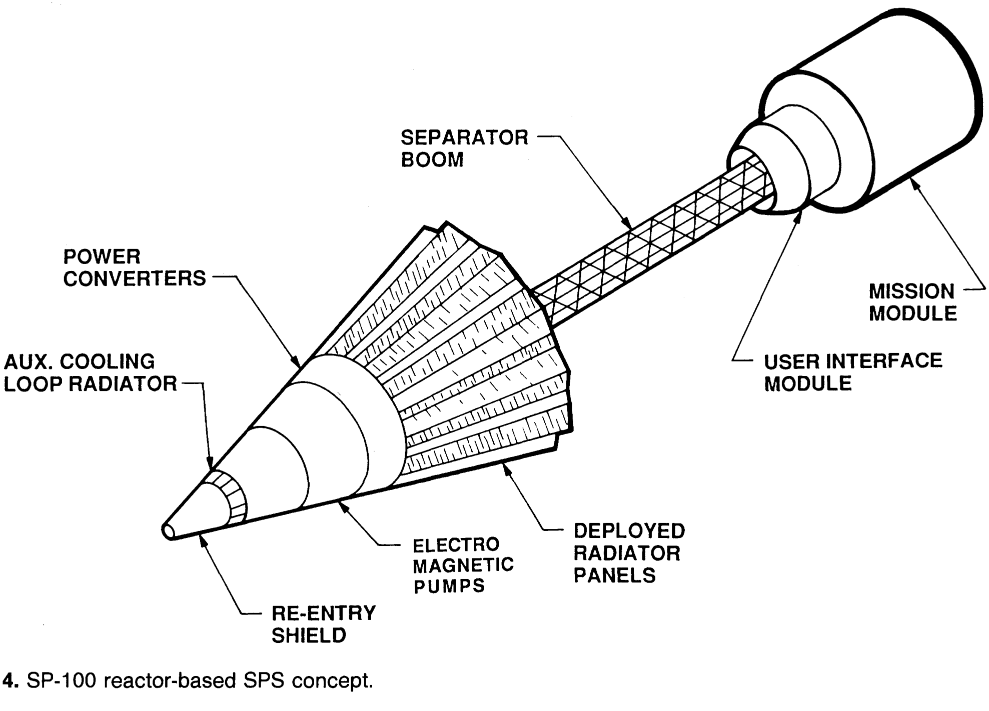

FIG. 4. SP-100 reactor-based SPS concept.

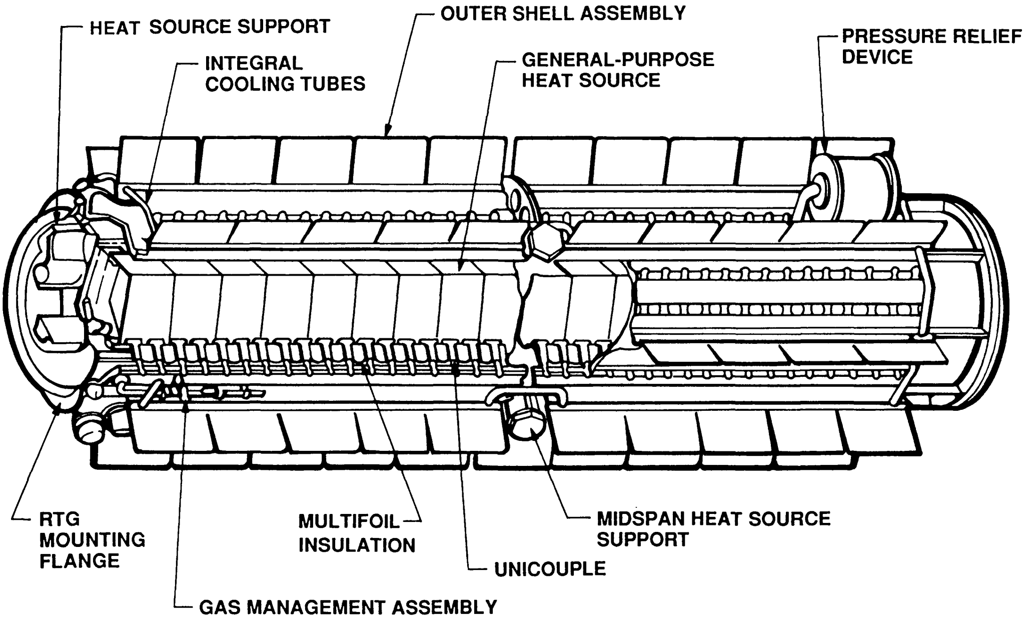

FIG. 5. ~250 We Radioisotope Thermoelectric Generator (RTG).

cell arrays. However, for more ambitious missions such as establishing human habitats on the moon and Mars, the power requirements will increase dramatically. In this forward-looking context, the SP-100 Program seeks to develop and demonstrate the technology for a nominal 100-kW reactor that can be easily scaled to higher power levels, using either static or dynamic conversion, depending on their state of readiness. A spacecraft powered by an SP-100 would have the form shown in Fig. 4, consisting of two major elements separated by a deployable boom structure. The forward main body, or power-generation unit, contains the reactor, shielding, the auxiliary cooling loop, static thermoelectric converters, radiators, and other equipment. The large radiator consists of 12 panels that fold forward and overlap when stowed for launch. When deployed, they form the "shuttlecock" configuration shown in Fig. 4.

Several new technologies for static conversion of heat to electricity that may be viable for space application are now under development in laboratories. One of the most attractive approaches is the alkali-metal thermoelectric converter (AMTEC) that uses liquid sodium as a working fluid and porous β-alumina to convert sodium atoms to ions, thus directly producing electricity from a pumped sodium circulation system. The pumping of sodium is done with an electromagnetic pump, which avoids dynamic machinery, a major source of vibration and unreliability in dynamic conversion systems.

2.4 Energy Storage

Rechargeable batteries contain cells that can be discharged and recharged (cycled) hundreds or thousands of times. They are used primarily to store energy when an Earth-orbiting spacecraft passes in the Earth's shadow. The most common battery for space applications uses nickel-cadmium (Ni-Cd) electrochemistry. Because of its favorable recharge characteristics and high cycle life, it has been used in the majority of Earth orbiting spacecraft. The cycle life is defined as the number of charge-discharge cycles that a battery will tolerate before failure. The cycle life, in turn, depends on the depth of discharge (DOD), which is the percentage of rated capacity removed from a cell during discharge. Ni-Cd batteries undergo a sharp drop in cycle life at high depths of discharge, so the DOD must be constrained in order to lengthen mission life. However, with this technique, the Solar Max spacecraft (orbiting the earth every 90 min) achieved eight years (~40 000 cycles) with three Ni-Cd batteries. Batteries are also characterized by their specific energy (energy stored per unit mass) and energy density (energy stored per unit volume). Table 3 displays these parameters for a selection of available batteries and a few batteries under development. (See also ENERGY STORAGE.)

2.5 Management and Distribution

The integration and interfacing of the energy source with a SPS to operate various spacecraft loads are accomplished with a power management and distribution system (PMAD). The PMAD

1. connects the energy source to the power grid or bus,
2. manages and regulates the bus current and voltage,
3. switches user loads on and off line,
4. monitors the status of the SPS, and
5. detects and manages faults in the system.

Almost all SPS employ direct current (dc) at around 28 to 30 V. A generalized PMAD architecture is given in Fig. 6. This example employs a photovoltaic solar array and a RTG as the power sources. The power control section of the SPS includes a microprocessor (computer) to direct the power. It does this by commanding various switches and components in accordance with its internal fault-detection and protection algorithms and instructions received from the ground.

Redundant batteries are included to provide power during short periods when the spacecraft requirements exceed the available combined RTG and solar-array power. When requirements are below the available power, a shunt-regulator unit regulates the bus by shunting the excess power to a radiator.

A "control" voltage (CV), in proportion to the variation of the 30-V bus, is developed by the shunt regulator and is used to control the switching of the solar array and the bidirectional converter (BDC). The latter handles charging and discharging of the batteries in response to the CV by providing appropriate voltage levels and charge rates. When the RTG and solar-array power is insufficient, the BDC provides supplemental battery power. When excess power is available, the BDC ensures that the batteries are charged.

The spacecraft engineering and science loads are connected to the power bus by the power distribution assembly of switches (relays). These switches must be designed to handle faults and protect the load.

3. SPACECRAFT PROPULSION

3.1 Introduction

The propulsion system provides thrust that is used

1. to accelerate or decelerate the spacecraft,
2. to counter the effects of atmospheric drag and other perturbations, or
3. to change or control the attitude of the spacecraft.

There are many different types of propulsion systems, and the selection of a propulsion system for a particular application depends on a number of considerations. Launch- or lander-vehicle propulsion systems must provide high thrust to escape or counter the pull

Table 3. Secondary cell characteristics.

Cathode–anode	Electrolyte	Open-circuit voltage	Nominal voltage	Specific energy (W h/kg)	Energy density (W h/cm^3)	Operating temperature (°C)	Cycle life (@ 80% DOD)
NiOOH–Cd	KOH	1.29	1.25	30	80	−10 to 35	2000
NiOOH–H_2	KOH	1.32	1.25	55	60	−10 to 35	2000
NiOOH–Zn	KOH	1.73	1.69	60	120	−20 to 60	200–300
AgO–Cd	KOH	1.41	1.2	55	110	−25 to 70	400
AgO–Zn	KOH	1.86	1.6	90	180	−20 to 60	100
AgO–H_2	KOH	1.39	1.2	80	90	0 to 50	400
PbO_2–Pb	H_2SO_4	2.1	2.0	30	70	−40 to 60	300–500
Li–TiS_2	organics	2.4	2.0	125	150	−20 to 55	new
Na–S	alumina	2.1	1.75	187	304	350	new

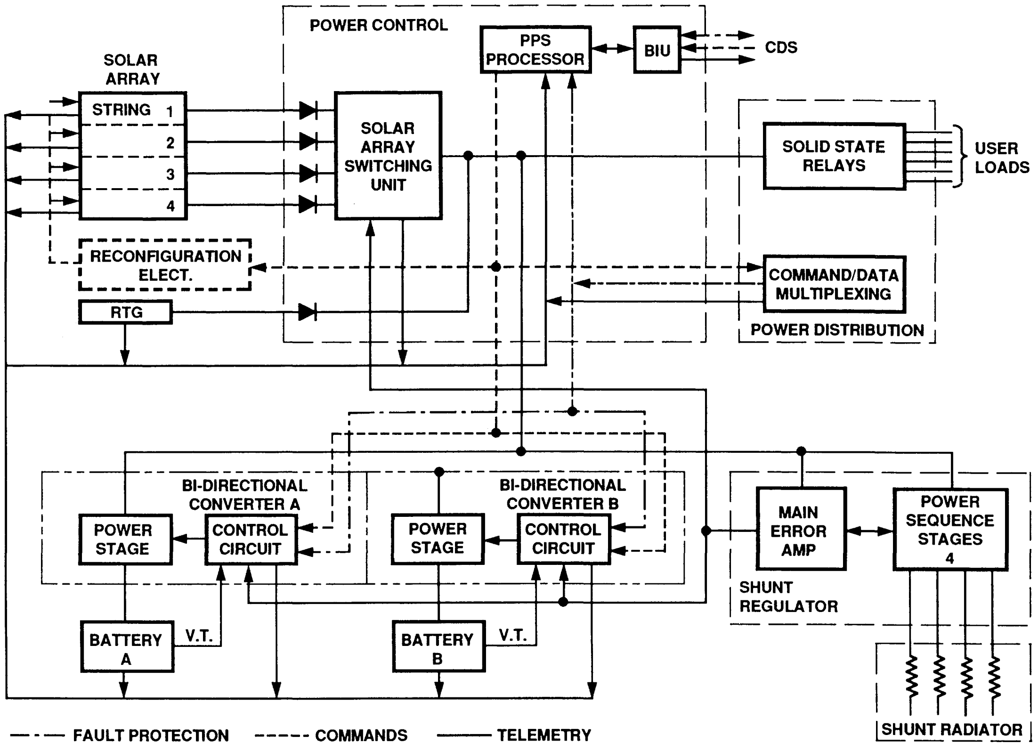
POWER CONTROL
SOLAR ARRAY
STRING 1
2
3
4
PPS PROCESSOR
BIU
CDS
SOLAR ARRAY SWITCHING UNIT
SOLID STATE RELAYS
USER LOADS
RECONFIGURATION ELECT.
RTG
COMMAND/DATA MULTIPLEXING
POWER DISTRIBUTION
BI-DIRECTIONAL CONVERTER A
BI-DIRECTIONAL CONVERTER B
POWER STAGE
CONTROL CIRCUIT
POWER STAGE
CONTROL CIRCUIT
MAIN ERROR AMP
POWER SEQUENCE STAGES 4
SHUNT REGULATOR
BATTERY A
V.T.
BATTERY B
V.T.
SHUNT RADIATOR
FAULT PROTECTION
COMMANDS
TELEMETRY

FIG. 6. Power processor subsystem functional block diagram. Note that BIU = bus interface unit, CDS = command and data system, and V.T. = voltage and temperature data for the batteries.

of gravity. Thrust is less important for other applications, and for these spacecraft the choice of propulsion system is driven by other considerations. High-energy missions such as unmanned outer-planet orbiter missions or fast manned missions to Mars must have propulsion systems that provide very high performance. For some missions, science objectives drive the choice of propulsion system. Sensitive optical surfaces, cryocooled sensors, and low-temperature thermal radiators are vulnerable to contamination produced by the propulsion system. Spacecraft using such devices require propulsion systems that do not exhaust condensable species that can accumulate on these sensitive surfaces. In some cases, a propulsion option is eliminated because a spacecraft instrument is designed to detect the very same chemical species that are present in the exhaust of the propulsion system.

All propulsion systems consist of

1. a source of energy,
2. a supply of propellant, and
3. a device for accelerating the propellant to high exhaust velocity.

In the chemical rocket, the propellant is both the source of energy and the expellant; in the electric propulsion system, nuclear rocket, or laser propulsion system the energy source is independent of the propellant. It is not even necessary that the source of energy be on board the spacecraft; it may be the sun, or a remote laser or microwave transmitter. Most rockets use a thermodynamic expansion process to accelerate the propellant to high velocities, but electric propulsion systems employ electrostatic or electromagnetic fields for this purpose. All propulsion systems must provide for the storage and transfer of the propellant to the rocket engine. For solid-propellant rockets, this is conceptually straightforward. Liquid-propulsion systems can be as simple as a system that stores a single propellant at room temperature and transfers this propellant to the engine by fluid pressure, or as complex as a system that applies sophisticated thermal control technology to store cryogenic propellants and turbine-driven pumps to transfer the propellant to the engine.

The preceding has been a general discussion of rocket propulsion. After a brief discussion of the basic physics of rocket propulsion, the characteristics of several representative propulsion systems will be addressed.

3.2 Physics of Rocket Propulsion

The propulsion system produces thrust (F) by expelling mass, i.e., propellant, at rate $\dot{m}$, at high velocity (v):

$$F=\dot{m}v.$$

The power (P) required to produce thrust is given by

$$P=\tfrac{1}{2}Fv.$$

This power can be supplied by energy stored on board the spacecraft in a chemical or nuclear fuel or by a source remote from the spacecraft such as the sun or a laser or microwave transmitter.

The power required by a launch-vehicle propulsion system is truly formidable. For example, the combined power of the main engines and solid rocket motors of the Space Shuttle at liftoff is 43 GW, which is equal to the output of 43 terrestrial electrical power plants. The combined thrust of these engines and motors is 29 MN. At the other end of the spectrum are small attitude control thrusters that produce a few millinewtons of thrust.

An important figure of merit for propulsion systems is specific impulse, which is a measure of the efficiency with which the system uses the expelled mass. Specific impulse (I) is defined as

$$I=F\frac{dt}{dm}=\frac{F}{\dot{m}}$$

and is equivalent to the exhaust velocity (v). By convention, specific impulse is often quoted in units of lbf s/lbm or "seconds," which is equivalent to the exhaust velocity in m/s divided by the gravitational constant (g). The specific impulse provided by different propulsion systems varies greatly. For chemical or nuclear systems, the maximum theoretical performance is limited by the energy release per unit mass of the fuel. The specific impulse that is really available, however, is often much lower than this theoretical value because of temperature constraints imposed by combustion chamber or reactor materials. Values of specific impulse for different propulsion systems are presented in Table 4.

The change in spacecraft velocity (V) that results from expelling propellant mass (m) is

$$V=v\ln[M/(M-m)],$$

where M is the mass of the spacecraft before the propulsive event. The mass M includes the payload as well as the "dry" mass of the propulsion system (i.e., the engine, tanks, and the power supply for nonchemical systems). Maximum performance requires that the mass of the nonpayload hardware be kept to a minimum. This equation illustrates the significance of specific impulse; the amount of propellant that must be expended to achieve a given velocity is less for a high-specific-impulse (i.e., high exhaust velocity) propulsion system.

The spacecraft velocity increases as the fraction of the vehicle mass devoted to propellant increases. As shown in Fig. 7, the maximum velocity achievable by a single-stage vehicle is 3 to 4 times the exhaust velocity of the propulsion system. Velocity requirements for typical spacecraft missions are given in Table 5.

3.3 Cold-Gas Propulsion Systems

The cold-gas system is the simplest propulsion system. This system requires no chemical reaction, but relies on the expulsion of a high-pressure gas such as nitrogen or helium to produce thrust. The performance of this system is low and the mass of the high-pressure tanks is high. Such systems are used for spacecraft with minimum propulsion requirements and for applications that require a very clean expellant.

3.4 Chemical Propulsion Systems

Almost all operational rocket propulsion systems are either solid or liquid chemical systems. These systems use a wide variety of propellants, some of which were identified in Table 4. Modern solid rocket propellants consist of an oxidizer, typically a perchlorate, and a metal fuel in an organic binder. Solid

Table 4. Performance of typical propulsion system.

Type	Vacuum thrust, N (power, kw)	Vacuum specific impulse (lbf s/lbm)	Application
		Chemical, Solid	
HC–NH_4ClO_4–Al	1.9×10^{5}[a]	293[a]	Inertial upper stage, first stage
HC–NH_4ClO_4–Al	1.0×10^{7}[a]	263[a]	Shuttle, solid rocket motor
		Chemical, Liquid	
N_2H_4	0.9	230	Voyager
NTO–MMH	400	308	Galileo
NTO–MMH	2.7×10^{4}	312	Shuttle, orbital maneuvering system
O_2–H_2	2.1×10^{6}	454	Shuttle, space shuttle main engine
		Nuclear, Thermal	
Solid core	10^5 to 10^6	800–900	Experimental
		Electric, Resistojet	
N_2H_4	0.2 (0.5)	300	Communication satellites
H_2	0.5 (3.0)	840	Experimental
		Electric, Arcjet	
N_2H_4	0.2 (1.8)	550	Telestar 4
NH_3	2.5 (30)	800	Experimental
		Electric, MPD	
Ar	1–100 (10–500)	2000–5000	Experimental
		Electric, Ion	
Xe	0.2 (5.0)	4000	Experimental

---

[a] Average values.

rockets are simple, reliable, and easily stored. The principal disadvantage of the solid rocket is poor controllability; once ignited these motors cannot be throttled or shut down easily. The Magellan spacecraft used the PAM D/Star 48D solid rocket motor shown in Fig. 8 to insert the spacecraft into orbit about Venus. This motor produces a thrust of 68.4 kN.

Liquid-propellant rockets use a wide range of propellants from single-component monopropellants such as hydrazine (N_2H_4) to two-component bipropellants such as nitrogen tetroxide–monomethylhydrazine (NTO-MMH) and cryogenic bipropellants such as oxygen–hydrogen. Liquid-propulsion systems can be relatively simple systems in which the propellant is fed to the engine by the propellant tank pressure or more complex systems in which propellant is fed to the combustion chamber by turbopumps that are driven by combustion gases or vaporized propellant.

Monopropellant propulsion systems are often used on planetary flyby spacecraft with modest propulsion requirements and for attitude control and precision trajectory adjustment requirements on spacecraft that use a solid rocket motor or a bipropellant system for large maneuvers. A monopropellant thruster uses a catalyst to decompose exothermically a propellant like hydrazine, and ex-

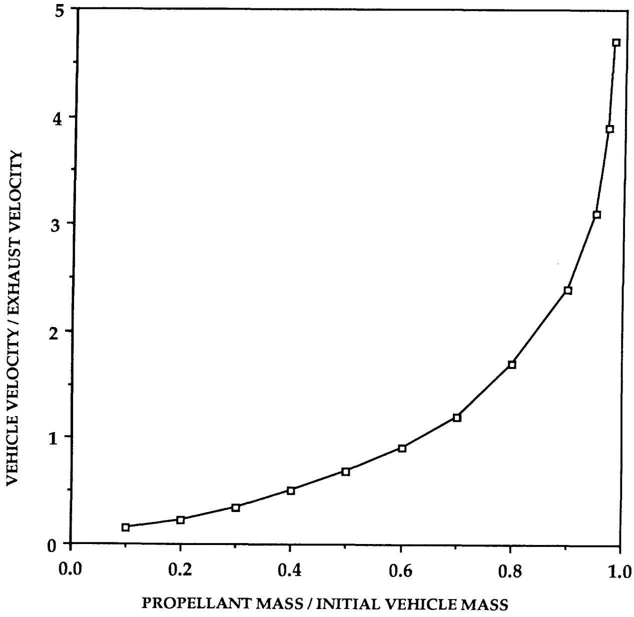

FIG. 7. Spacecraft velocity increase provided by a single-stage rocket.

hausts the high-temperature decomposition products through the thruster nozzle. A 0.9-N hydrazine thruster like that used on the Voyager spacecraft is shown in Fig. 9.

Propulsion requirements for planetary orbiter spacecraft are larger and more complex, and require a higher-performance propulsion system. For example, the Mars Observer propulsion system is composed of bipropellant and monopropellant subsystems. The bipropellant system is pressure regulated with four 490-N thrusters and four 22-N thrusters and contains 1360 kg of NTO-MMH propellant. The monopropellant system consists of eight 4.45-N and four 0.9-N thrusters and contains 84 kg of N_2H_4 propellant. Both of these systems consist of propellant storage tanks, con-

Table 5. Velocity requirements for typical spacecraft missions.

Missions	Velocity (km/s)
Planetary flyby—Voyager	0.2
Communications satellite	0.5
Planetary orbiter—Mars observer	2.5
LEO[a] to Earth escape	3.2
LEO to Lunar orbit (7 days)[b]	3.9
LEO to GEO[c]	4.2
LEO to Mars orbit (0.7 year)	5.7
Surface to LEO	7.6
LEO to solar escape	8.7
LEO to Neptune orbit (30 years)	13.4
LEO to Neptune orbit (5 years)	70.0
LEO to α-Centauri (50 years)	30 000

[a] LEO=low earth orbit.
[b] ( )=trip time.
[c] GEO=geosynchronous earth orbit.

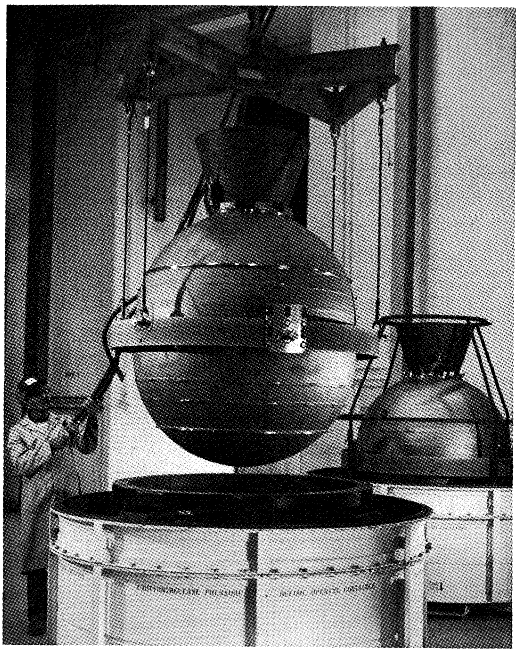

FIG. 8. 68.4-KN Star 48B solid rocket motor. (Photograph courtesy of Morton Thiokol Corporation.)

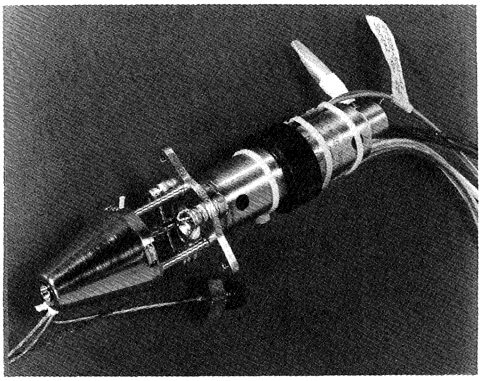

FIG. 9. 0.9-N hydrazine monopropellant thruster. (Photograph courtesy of Rocket Research Corporation.)

trol valves, pressure transducers, filters, and other components.

The Space Shuttle Main Engine (SSME), shown in Fig. 10, is an example of a state-of-the-art high-performance liquid rocket booster engine. This oxygen–hydrogen engine produces 2.1 MN of thrust under vacuum conditions and can be throttled between 65% and 109% of rated power. The turbopumps of this engine are driven by high-temperature hydrogen supplied from the propellant tank and heated by the combustion chamber. The SSME is designed for repeated reuse, up to 55 flights, before requiring a major overhaul. The engine is controlled by a pair of computers that monitor its operation and respond to commands from the Shuttle guidance and navigation computers.

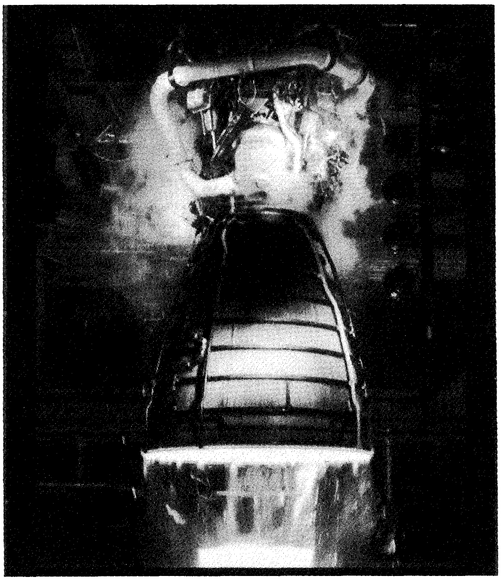

FIG. 10. 2.1-MN Space Shuttle main engine. (Photograph courtesy of Rockwell International-Rocketdyne Division.)

3.5 Electric Propulsion Systems

Electric propulsion systems rely on a source of energy other than that contained in the propellant to accelerate the propellant. Such systems can achieve very high performance, but require a solar or nuclear power supply to provide the required electrical energy. The mass of the power supply can be large, and its effect on the performance of these propulsion systems can be significant. Space-power technology is discussed in Sec. 3 of this article. Electric propulsion devices have been under development for many years and are finding increasing application in operational systems.

The hydrazine resistojet is a device that uses a resistive element to heat the hydrazine decomposition products and thereby to increase performance. The 0.2-N thruster, shown in Fig. 11, provides a specific impulse of approximately 300 lbf s/lbm, equivalent to bipropellant performance, at a power level of 500 W. This type of thruster is currently being flown on the INTELSAT, SATCOM, G-Star, and Spacenet communication satellites.

The arcjet uses an electrical arc rather than a resistive element to heat a propellant that is then expanded through a conventional nozzle. These devices have been successfully operated in the laboratory at power levels

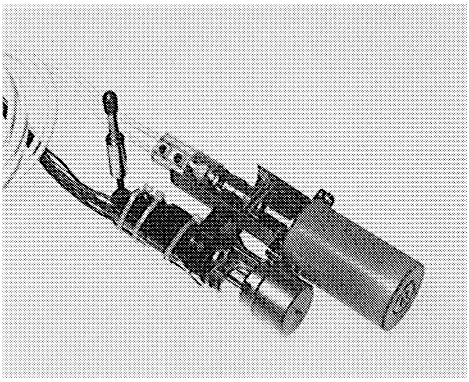

FIG. 11. 0.2-N hydrazine resistojet. (Photograph courtesy of Rocket Research Company.)

from 1 to 200 kW with a variety of propellants. The Telstar 4 communications satellite, to be launched in the late 1990s, will use two 1.8-kW arcjets for station keeping.

The magnetoplasma dynamic (MPD) thruster and the pulsed plasma thruster (PPT) accelerate a highly ionized propellant with an electromagnetic force. Micronewton-class PPTs are operational on the U.S. Navy Nova satellites, and MPD thrusters are operational in the laboratory at steady-state power levels of up to 500 kW. High-power MPD thrusters may be an attractive option for multimegawatt electric propulsion vehicles of the future.

The ion thruster employs an electrostatic field to accelerate propellant ions to high exhaust velocities. Because of its potential for providing very high exhaust velocity and high efficiency, it is well suited to meet the primary propulsion requirements of high-energy planetary missions. Mercury-ion thrusters were flown as experiments on U.S. spacecraft during the 1960s and 1970s. More recently, 5-cm ion thrusters were flown on ETS-III, a Japanese satellite. Ongoing U.S. research is directed toward improving thruster life and efficiency with inert-gas propellants such as xenon. A 25-cm-diam xenon thruster is shown in Fig. 12. This engine produces a thrust of 0.06 N at a specific impulse of 2900 lbf s/lbm and requires 1.3 kW of electrical power.

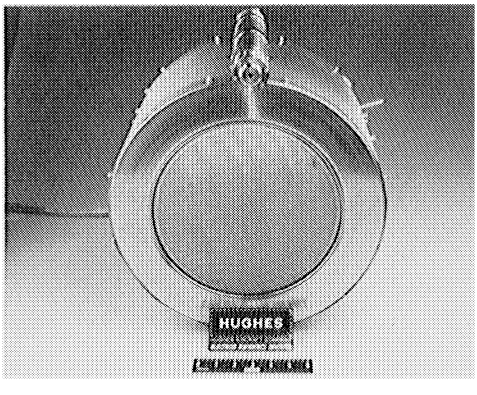

FIG. 12. 0.06-N xenon ion thruster. (Photograph courtesy of Hughes Corporation.)

3.6 Advanced Propulsion Systems

The energy released by nuclear fission is more than 100 times greater than that released by the most energetic chemical reaction. As previously stated, nuclear fission is a very mass-efficient source of energy for generating the electrical power required by electric propulsion systems. Nuclear energy may also be used to directly heat the propellant, which is then accelerated by a thermodynamic expansion process. In the solid-core nuclear rocket, heat is generated in nuclear fuel assemblies and transferred to a propellant such as hydrogen, which flows through the core. The maximum specific impulse, 800–900 lbf s/lbm, is limited by the maximum operating temperature of the fuel assembly materials. This performance is a factor of 2 greater than the best chemical systems. Solid-core nuclear rockets were developed and tested in the U.S. during the 1960s at thrust levels up to 0.9 MN, but this work stopped short of full-scale development when the U.S. plans for Mars exploration were put on hold. Even higher performance is theoretically possible with the gas-core reactor that bypasses the limitations of the solid-core rocket by using a fissioning plasma to heat the propellant. Fusion and antimatter-annihilation processes are even more energetic sources of power that may someday be harnessed for space propulsion.

4. ATTITUDE CONTROL

4.1 Introduction

Attitude control involves the control of the orientation of a spacecraft with respect to an inertial reference coordinate frame or with respect to a local relative coordinate frame, such as that defined by the direction to a nearby body. This includes the orientation of the spacecraft body structure, articulated appendages, and sensors and instruments attached to the various parts of the spacecraft. Maneuvers, turns, and articulations required to achieve desired orientations are part of attitude control.

The purposes of attitude control are varied. For example, orientation with respect to the sun may be for thermal control of the spacecraft, for power generation with solar-cell arrays, or because the sun is an easily detected attitude-reference body in our solar system. High-gain antennas need to be oriented to establish and optimize communication between the spacecraft and the Earth. Scientific instruments in the payload of the spacecraft require orientation toward the phenomena to be observed. Articulated masses on the spacecraft are used to control the mass properties (center of gravity and moments of inertia), particularly on spinning spacecraft. Propulsion system thrusters must be oriented inertially in order to perform trajectory-correction maneuvers.

Attitude-control systems are concerned with stabilization of the motions of the spacecraft, and with pointing and orienting the spacecraft or parts of the spacecraft. Because of the various forces that act on a spacecraft (gravity gradients, torques due to propulsion systems and mechanical mechanisms, solar pressure, micrometeorite impact, etc.), a spacecraft that is not stabilized will tend to acquire rotational motion and will tumble and move in a chaotic fashion if it is not explicitly stabilized by some technique. There are two general types of stabilization: three-axis and spin stabilization. Spin stabilization works on the same principle as a child's top. If a rigid body is spun to moderate rotational speed about its axis of maximum moment of inertia, the angular momentum vector and the spin axis will stay fixed in inertial space, stabilizing two rotational degrees of freedom. The third degree of freedom, about the spin axis, may have only rate control and not position control. Now, if any modest torques are applied to the spinning body, the net resultant angular momentum vector will be a vector sum of the large initial angular momentum plus the small perturbation, and will lie very close to the original vector. Thus the spacecraft becomes stabilized by virtue of its high initial angular momentum. The three-axis-stabilized spacecraft controls the position and rate in all three rotational degrees of freedom, and the usual state is holding a fixed orientation with respect to its reference coordinate frame. If perturbations tend to move it away from its preset orientation, forces are automatically applied by the spacecraft to compensate and keep it oriented as desired. Turns from one orientation to another are controlled but are considered a temporary or transient state. A hybrid of the two general types is the dual-spin spacecraft, which is predominantly spin stabilized but has a part that has a rotational degree of freedom about the spin axis. This part can be despun to allow position control about this third degree of freedom so that it has three-axis-like stabilization and allows inertially fixed pointing for instruments and sensors on the despun part. Any of these types can carry articulated appendages.

An attitude-control system has three basic components:

1. a complement of sensors to determine or measure the attitude in three degrees of freedom;
2. information processing to compare the sensor data to the commanded or desired attitude and to transform the error into the action necessary to correct the attitude; and
3. an actuation system to apply the required corrective torques to the spacecraft.

The control of articulated parts requires the equivalent three components but with probably fewer degrees of freedom.

4.2 Attitude Sensing

To measure the attitude of a spacecraft it is necessary to measure the direction, in a spacecraft fixed-coordinate frame, of an external body whose direction is known in an external reference coordinate frame. Each measured direction provides two degrees of freedom of attitude knowledge. Full attitude knowledge involves three degrees of freedom and thus requires observing two different directions, though some particular tasks, e.g., high-gain antenna pointing, require only two components of attitude knowledge.

Some of the external entities that can be used as references include light from stars, the sun or solar-system bodies, radio-frequency beacons, or directional fields like the Earth's magnetic field. Sun sensors are most commonly used because of the sun's highly visible presence in our solar system and its high luminosity. Even if totally disoriented, a spacecraft can rely on the sun as an unmistakable beacon. Its great brightness and relatively compact size (0.5° at Earth) provide an excellent attitude reference. A sensor staring at the sun, or scanning over its disk periodically from a spinning spacecraft, can provide two of the three axes of information needed to fully determine inertial orientation. A solar reference is also the first external reference sought in recovering from a serious spacecraft anomaly. After the solar reference is acquired, the spacecraft can position itself in a known orientation, where the solar power and thermal input are maintained at acceptable levels. A two-axis sun sensor usually consists of two single-axis channels providing measures of the sun direction offset from each of two orthogonal null planes. Both photovoltaic and photoconductive detectors have been used for sun sensors.

While sun sensors usually play a key role in establishing an initial two-axis attitude reference for both nominal operations and recovery from anomalies, star trackers often play an equally important role by either filling in the missing axis of information (as was done with the Mariner and Voyager Canopus tracker), or providing more accurate three-axis references once a safe orientation is established. The latter approach is used for the Galileo spacecraft in which the sun sensor keeps the spin axis pointed roughly at the sun (or the Earth for some phases of the mission), while the star scanner provides accurate three-axis data into the attitude-estimation process. Sensing the direction to distant objects like stars or bodies in our solar system requires the signal enhancement of a telescope and a sensitive detector. A single-axis star tracker, measuring the offset from a null plane, is sufficient to provide the third knowledge component to augment a two-axis sun sensor. A star tracker with a two-degree-of-freedom focal plane detector, e.g., a vidicon or charge-coupled device (CCD), can provide a complete star-direction measurement of one or more stars. Two or more stars in the field of view can thus provide full three-axis attitude knowledge. However, to be usable, the stars observed must be positively identified, since observation of an unknown star provides no information.

For a spinning spacecraft, a star scanner with two nonparallel (V) slit fields of view, approximately orthogonal to the spin axis, provides a series of timed events when the slit fields of view pass stars. The times of the pulse pair arising from the two slits crossing an identified star provides two pieces of information, and observation of two stars is sufficient to determine the spacecraft attitude. (A similar V-slit sun sensor provides only the angle between the spin axis and the sun direction, since there is only one sun.)

For an orbiting spacecraft maintaining an axis along the local vertical, a horizon scanner, measuring the angles from the spacecraft axis to the horizon in several (at least two) different planes, provides two axes of information on the offset of the spacecraft axis from the local vertical. By continuously or periodically measuring the vector to a planet's limb, this device provides horizon information that can be used to maintain a desired spacecraft orientation, or to point instruments, antennas, etc. For highest accuracy, corrections must be made for orbit altitude and inclination, planet oblateness, and sensed horizon altitude, including diurnal and seasonal variations. The sensed horizon altitude depends on the part of the electromagnetic spectrum used, usually the infrared.

While optical trackers are required for providing known references to celestial or "absolute" coordinates, inertial sensors, which measure vehicle motion without reference to any external body, are also required when the spacecraft attitude is changing. *Accelerometers* measure changes in linear motion that generally cannot be measured by optical means. *Gyroscopes*, measuring angular motion, are often required to supplement optical sensors to

1. provide high-bandwidth data when needed, particularly during critical mission phases such as propulsive maneuvers;
2. operate when normal optical references are unavailable, such as during eclipse;
3. instantly take over control should optical references be lost or some other anomaly occur.

Measurement of the deviation of the attitude from that determined by optical sensors is accomplished with gyroscopes, which measure the inertial attitude rate of the spacecraft and integrate this to compute the current attitude. Occasional optical measurements are used to correct for accumulated gyroscope and integration errors.

The position of movable joints, whether rotational or translational, must be measured if they are to be controlled. A number of sensors, such as potentiometers, resolvers, or digital encoders, are available. It may also be desirable to measure the joint velocity and the force (or torque) being applied. Actual selection depends on requirements such as accuracy, power consumption, mass, volume, and analog versus digital compatibility with the system design.

4.3 Attitude Determination

The attitude sensors described in the previous section provide measurements of quantities that are dependent on attitude, but not the attitude itself. The processing of these sensor measurements to obtain the attitude is called *attitude determination.*

The simple spacecraft, with no on-board attitude determination, is controlled to keep the sensor outputs at null or a specified biased value. In this case the attitude is known to the ground control but the spacecraft only knows about deviations from this unknown target attitude. The equivalent, using gyroscopes for turns, is to try to null each gyroscope axis error signal while the gyroscopes are biased by rate commands. This ignores the noncommutativity of the angular variations.

With on-board processing capability now available, on-board, autonomous attitude determination can be done. The sensor measurements, the knowledge of the directions to the objects seen, and the geometric relationships between attitude and measurements are combined to derive an estimate of the spacecraft attitude. This is only an estimate of the attitude since it is based on noisy data and can be improved with the processing of more measurements. The complexity of the estimator can range from simple, fixed-gain suboptimal to a full multistate Kalman filter.

If the measurements are not simultaneous but distributed in time, an attitude state model is required to propagate between measurements and must be updated with each new measurement.

Autonomous attitude determination has many advantages, from enhancing spacecraft fault tolerance to reducing the workload in ground control by being able to respond to high-level commands, working out the details on board.

4.4 Actuators

To effect a change in the spacecraft attitude it is necessary to apply a torque. This torque may be generated by mass expulsion, reaction against an external field, or momentum transfer to an internal component of the spacecraft.

Mass expulsion through small thrusters oriented such that the line of thrust does not pass through the spacecraft center of mass will provide a torque. Thruster locations and orientations are usually chosen so that the torques from various thrusters are aligned with the spacecraft principal axes. To provide torque without a translation force, they are usually mounted in opposing pairs (couples) to cancel the forces and combine the torques. To conserve the mass available, attitude maintenance using thrusters will have a very low duty cycle, with occasional small pulses to reverse the spacecraft angular rate as it is allowed to drift between specified limits (dead band). Higher duty cycles are used when a larger torque impulse is needed as for control during the firing of a larger thruster. For such a larger thruster, two axes of control torque can be provided by a two-degree-of-freedom gimballed nozzle that moves the line of thrust with respect to the spacecraft center of mass. A low duty cycle is not suitable for high-precision attitude control. For a spinning spacecraft, torque normal to the spin axis changes the direction of the angular momentum vector, and thus the spin axis, in inertial coordinates.

Solar-radiation pressure on the spacecraft causes a torque that can be controlled by changing the configuration, e.g., with movable vanes, to move the center of pressure. A configuration with the center of pressure behind the center of mass provides passive, stable control of two spacecraft axes. The third-axis torque can be controlled by changing the "windmill" component of the configuration but does not provide passive, stable control.

An orbiting spacecraft with its axis of least moment of inertia vertical is inherently stable due to torques from the gravity field gradient.

A spacecraft in a magnetic field can create torque by activating magnetic dipoles on board. However, only two axes of torque are possible at any one time with no torque about the field direction, so it is not suitable for continuous three-axis control, but can be used for "momentum dumping", described later.

A spacecraft can be made to turn about an axis by torquing against an internal reaction wheel that rotates about that axis. Momentum is transferred between the wheel and the spacecraft body. To stop, the momentum is transferred in the other direction. Control moment gyroscopes are also used for this momentum transfer. This type of actuation is well suited for high-precision, high-bandwidth control. If an external torque acts on a spacecraft that is to be held still, the momentum accumulates in the wheel and occasionally must be "dumped" by applying an external torque.

Actuators (torquers) for rotary joints are usually electric motors. They may be direct drive with the rotor and stator attached to opposite sides of the joint, or they may be geared to increase torque and positioning precision. Bearings must be designed to have low friction and still carry the appropriate loads. A common translational actuator is a ball screw that converts rotary motion of a screw to translation of a nut with ball-bearings. When high precision is required within limited ranges of movement, piezoelectric or magnetostrictive devices can be used.

4.5 Controller

We now have the attitude from the sensors and attitude determination to compute the error with respect to the desired attitude, as well as a means to torque the spacecraft, but we must still close the control loop. The control law is the computation of the torque command required to null the error, in a stable manner, given the attitude error. If attitude rate error is not available, it may need to be derived in order to provide stable control.

For low-duty-cycle systems using thrusters, the analysis and design of the control law is most easily done using piecewise-linear phase-plane techniques. For continuous control, the classical control theory of Bode and Nyquist or root locus is needed for design of a stable controller. Discrete digital control systems have sampling, quantization, and delays to consider as destabilizing factors and require other techniques such as *Z* transforms in lieu of the continuous Laplace transform.

Spacecraft evolution has made its own contribution to the difficulty of control law design. A small compact spacecraft is easily modeled as a rigid body. Flexible appendages, e.g., solar panels and science booms, complicate the model and the control law required to stabilize the spacecraft without unduly exciting appendage motion. As the spacecraft gets larger, it may no longer be considered a rigid body, but it has structural resonance modes of its own that must not be excited. On a large spacecraft the attitude sensors and the actuators may be separated by flexible structure (noncollocated). All these factors contribute to the increasingly difficult task of designing stable control laws for increasingly more complex spacecraft.

5. SENSORS FOR SPACECRAFT CONTROL

5.1 Introduction

An extensive range of sensors is needed for controlling all but the most simple spacecraft. This is true whether the vehicle is in orbit about the Earth or some other planet, or far from all other objects in interplanetary space. These sensors support a number of critical spacecraft functions:

1. control of attitude (and spin) with respect to inertial coordinates,
2. control of orbital motion (propulsive maneuvers), and
3. control of instrument pointing.

Although there may be much overlap in the sensors supporting these functions (e.g., a gyroscope can be important to all three) it is often helpful to keep these functional differences clearly in mind. Many of these same systems will play an important role as crewless exploration moves to surface vehicles for exploration of other planets and satellites. Other unique sensors, such as a three-dimensional mapper to aide in robot navigation, are still in early phases of development.

5.2 Optical Sensors

Optical tracking instruments, sensitive to visible or infrared wavelengths, are used for all but the most rudimentary spacecraft (e.g., simple spinners). Several types of instruments usually work together with inertial sensors to fully control the spacecraft.

5.2.1 Sun Sensors. Relatively simple sun-sensor configurations can provide full sky coverage for acquisition, or higher accuracy tracking over a more limited field. Figure 13 shows the layout and typical response curve for the fine sun sensor that was used on Mariner and Viking missions to Mars. More recent off-the-shelf designs incorporating digital signal processing electronics typically achieve 1.0 arcmin accuracy over a 64×64 deg^2 field, and 15 arcmin over 4π sr in scanning (spinning vehicle) applications.

One challenge for outer-planet missions has been the wide range of solar illumination. The intensity of sunlight at Saturn is roughly $\frac{1}{100}$ of that at Earth, dropping to $\frac{1}{900}$ at Neptune and Pluto. Most sensors designed for operation at 1 AU will not operate satisfactorily at these low intensity levels. As a result, a special high-dynamic-range detector was needed for Voyager. Substitution of linear CCDs for the monolithic detectors previously used is an approach to solving this problem. In addition, by precisely measuring the position of the solar limb this approach leads to significantly improved measurement accuracy.

5.2.2 Star Trackers. Early star-tracker designs (through Galileo) employed image dissectors or photomultiplier tubes for starlight detection. Both devices use electron multipliers for low-noise amplification, and in many cases achieve quantum-limited performance. For spinning spacecraft, a simple V-slit scanner using a photomultiplier to detect star pulses can provide excellent results. Each star crossing produces two pulses, their arrival time giving clock information (around the spin axis) and their separation providing cone data (angle from the spin axis). Complete (three-axis) knowledge of the spacecraft orientation is derived by tracking several known stars on each rotation.

Image dissectors develop two-axis data by electronically scanning a small sensitive aperture over the field of view. This enables star coordinates to be measured for an inertially fixed spacecraft. Image dissector tubes were used throughout the 1970s and 1980s in both "fixed-head" and "focal-plane" star trackers. A fixed-head tracker includes all elements (optics, stray-light baffle, detector, and electronics) needed to provide a stand-alone capability ready to mount at any appropriate orientation on the spacecraft. In contrast, a focal-plane tracker examines a portion of the image formed by an external optical system, tracking stars or other objects in that field. The Hubble Space Telescope's fine guidance sensor is an example of a focal-plane tracker, part of which uses image dissectors to develop "coarse" error signals. A dissector was also used for the NASA standard fixed-head tracker, which generates two-axis measurements of a star position by rapidly scanning the dissector's sensitive aperture over the star image. Multiple stars could also be tracked by time-sharing of the aperture over several images.

While the image dissector has many strengths for star tracking, it also suffers from a number of drawbacks that substantially limit its usefulness. First, by sensing only one point at any instant, the tracker is very susceptible to image motion. This is particularly true when attempting to track more than one star in the field. Another serious limitation to

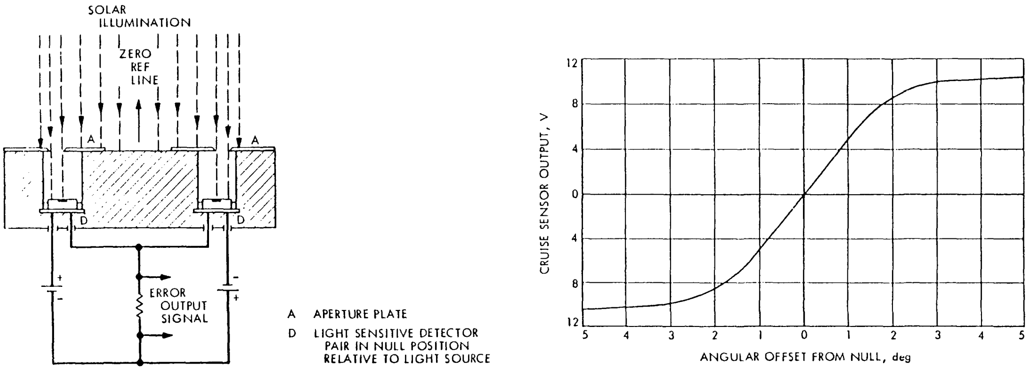

FIG. 13. Fine sun sensor used on Mariner and Viking missions to Mars.

accuracy is the positional distortion and nonlinearity associated with the electromagnetic deflection of the sensing aperture. Although these effects can be calibrated to some extent, significant errors remain ($>1/10^3$ of the field), which can be further exacerbated by variation in external magnetic and electrostatic fields.

These limitations have been removed to a remarkable degree by the advent of area-array CCDs [see CCDs (CHARGE-COUPLED DEVICES)] in the star-tracking arena. Instead of relying on variable electromagnetic deflection to derive positional information, the precise photolithography of the CCD provides a geometric reference frame accurate to substantially better than one part in 10^4 over the field. Much of this resolution is achieved by measuring the centroid of each star image to an accuracy much finer than the pitch of an individual pixel. Also, by integrating all image pixels simultaneously, a CCD can track much fainter stars and avoid the nonsimultaneity problem associated with image dissectors. Other advantages, such as insensitivity to magnetic fields, avoidance of high voltages, insensitivity to damage by overexposure, etc., have helped make CCD-based trackers the new *de facto* standard. Many of these advantages can be appreciated in comparing the specifications of a typical CCD tracker, flown on the Astro-1 shuttle-based observatory, to the earlier NASA standard (see Table 6). A schematic of the Astro-1 tracker is shown in Fig. 14.

5.2.3 Horizon Sensors. Horizon sensors, like sun and star trackers, can be divided into scanning and staring types. Staring sensors usually employ an array of detectors that locate the limb by radiance-balancing techniques. Many Earth satellites, at both low and geosynchronous altitudes, employ scanning horizon sensors, where the scanning motion is provided either by the spacecraft itself or by rotating mirrors, optical wedges, etc.

Infrared wavelengths are used to avoid significant day-night differences in the appearance of the limb, as well as to minimize the effects of clouds. For the Earth, bolometers operating in the 14–16-μm carbon dioxide band have proven most effective. At these wavelengths, the effective limb appears roughly 40 km above the surface, well above the weather in the troposphere. There are still season- and latitude-dependent errors, but these can be largely removed analytically. As one moves to other planets, detectors covering the mid- to long-wavelength infrared region are still required. The weather on Mars is substantially less dramatic than Earth's, but it can produce substantial albedo variations and horizon errors in optical wavelengths. Mars's atmosphere of mostly carbon dioxide dictates wavelengths similar to those used on the Earth.

5.2.4 Optical Target Trackers. The advent of the CCD has also provided the opportunity to track accurately targets that are much more complex than simple star fields. A capability to locate and track extended targets, limb points, or particular surface fea-

Table 6. Comparison of CCD star tracker with earlier NASA standard image dissector star tracker.

Parameter	CCD[a]	ID[b]
Field of view (deg^2)	4.0×4.5	4.2×4.2
Limiting star magnitude	8.2	5.7
Probability of detecting threshold star	>0.99	0.96
Accuracy after calibration (arcsec)	0.5	17.5
Noise-equivalent angle threshold star (arcsec)	0.5	11.3
Other factors	May require CCD cooling	• Damaged by sunlight • Sensitive to magnetic and electric fields • Requires high voltage

[a] Stanton *et al.* (1987) (scaled for field of view comparable to ID case).
[b] Deters (1979).

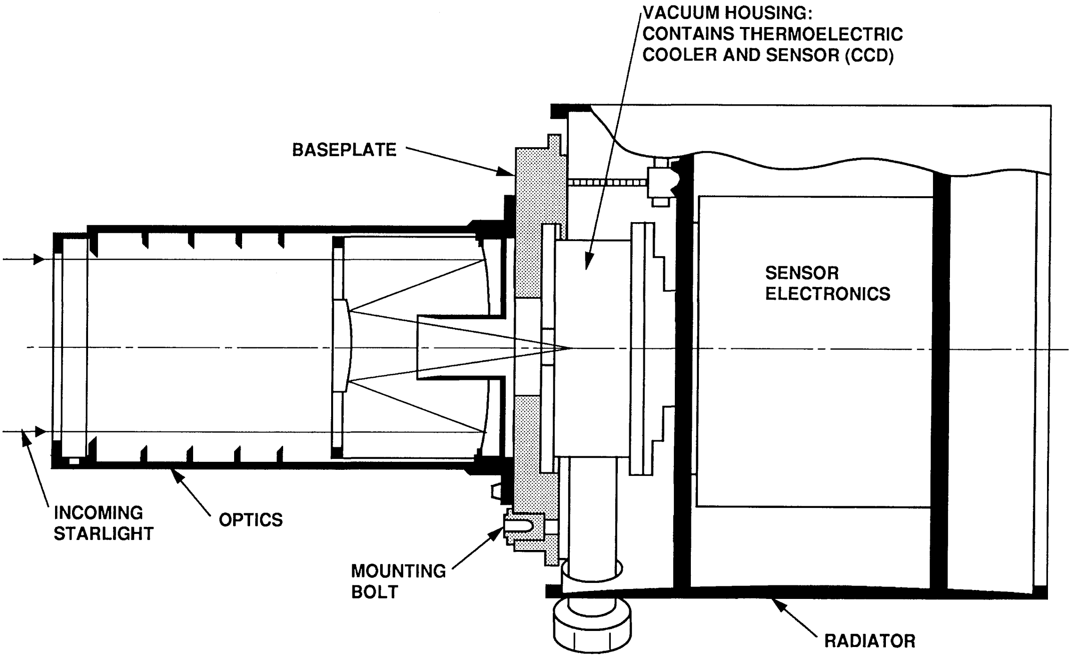

FIG. 14. Schematic of Astro-1 star tracker.

tures can all be derived from the information in a CCD image. By using two cameras, or a succession of frames with a single camera, three-dimensional maps can be generated for rover navigation, sample acquisition, or other functions required for robotic exploration of planet surfaces.

5.3 Inertial Sensors

As is the case with optical sensors, a new generation of inertial instruments is being developed that could be vital to the more challenging, extended life missions envisioned for the future.

5.3.1 Accelerometers. Typical spacecraft accelerometers provide single-axis information by measuring the force a small mass exerts on its surroundings, generally providing sensitivities in the micro-*g* range (10^{-5} m/s^2). Substantially more sensitive devices have been reported (10^{-8} m/s^2), for special "quiet" environments such as a "drag-free" spacecraft.

New-technology microaccelerometers have been recently demonstrated using a variety of micromachining techniques in silicon. A promising development uses the detection of a tunneling current to detect very minute motion of the proof mass (on the order of nanometers). Since these devices draw little power and can be made extremely small (see Fig. 15), their use in a wide variety of applications measuring vibration and rotations, as well as linear acceleration, can be envisioned.

5.3.2 Gyroscopes. *Spun Mass Gyroscopes.* To date, conventional floated and tuned rotor ball-bearing gyroscopes have been widely used in spacecraft applications due to their advanced state of development and adequate performance. Operating lifetimes of more than three years have been demonstrated, but wear-out mechanisms eventually lead to failure. Electrostatic suspension has been thoroughly investigated as one technique for limiting wear-out mechanisms and improving accuracy. However, high cost and difficulty in making them sufficiently rugged for the launch environment have kept this approach from being adopted.

A very successful conventional design, incorporating dual-axis gyroscopes in a redundant three-axis configuration, has been the NASA Standard Dry Rotor Inertial Reference Unit (DRIRU II). The characteristics of this instrument are listed in Table 7. It has been

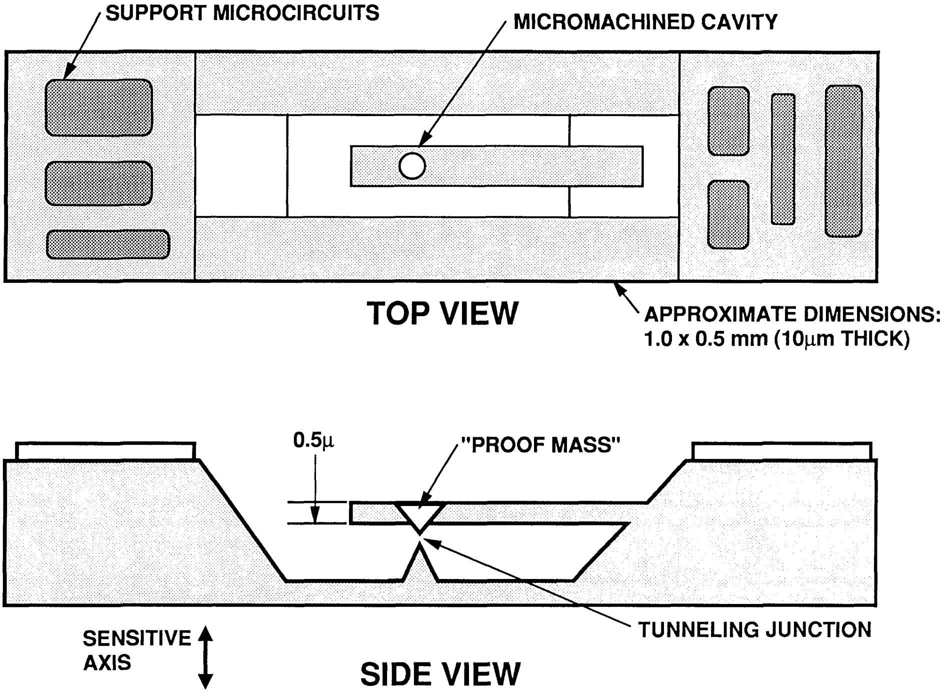

FIG. 15. Micromachined tunneling accelerometer.

incorporated into a number of NASA spacecraft including the Solar Maximum Mission, Hubble Space Telescope, and the Astro Ultraviolet Spacelab observatory.

Optical Gyroscopes. A substantial development effort has been undertaken in recent years to replace spun mass gyroscopes with optical devices. Although the main advantage for many spacecraft applications is expected to be the reduction in failure rate associated with the elimination of mechanical wear-out, other advantages, such as lower mass or power consumption, may also be substantial.

Both *laser gyroscopes* and *fiber-optic gyroscopes* are based on the principle discovered by G. H. Sagnac in 1913. The Sagnac effect states that, when two counterrotating beams of light are subjected to a rotation in their planes, a phase shift is generated between the two beams that is proportional to the rotation rate. By measuring this phase shift, it is possible to infer the rotation rate, usually over a range much larger than can be achieved with mechanical devices.

The fiber-optic approach (Fig. 16) is generally believed to offer the most promise for long-life applications. Laser gyroscopes require maintenance of a helium-neon gas laser operating within a precisely machined quartz block. In contrast, fiber-optic gyroscopes can

Table 7. Characteristics of NASA Standard Dry Rotor Inertial Reference Unit (DRIRU II).

Parameter	High-rate mode	Low-rate mode
Rate range (deg/s)	0–4.0	0–0.11
Bandwidth (Hz)	15	15
Short-term stability (arcsec/s) (six-hour period after calibration in a benign environment)	...	0.003
Long-term stability (arcsec/s) (30 days after calibration)	0.03	0.02
Nominal operating life (years)	5	5

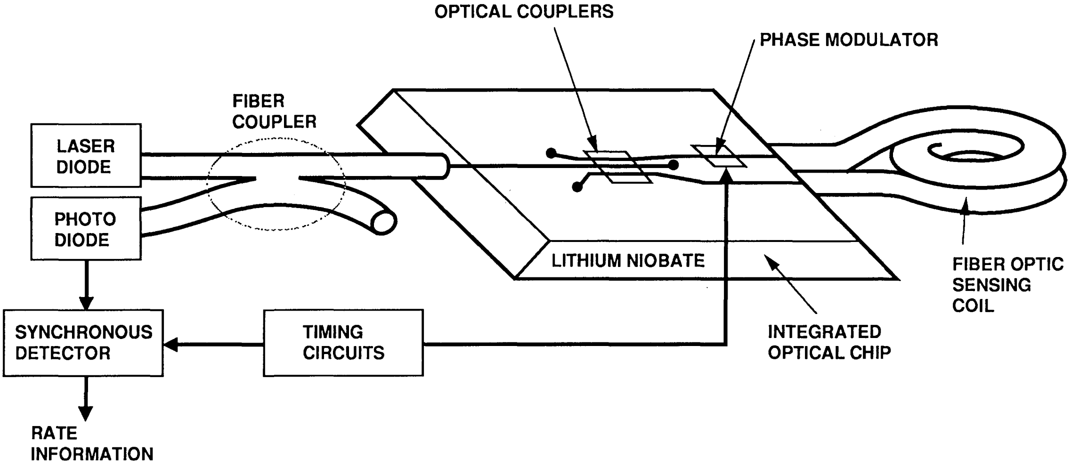

FIG. 16. Fiber-optic gyroscope.

use solid-state lasers with a length of fiber (meters to kilometers) commensurate with the accuracy needed. The use of redundant lasers for each gyroscope axis can be made to increase lifetime potential.

At present, many development challenges, such as the ability to manufacture reliable electro-optical switching and modulation chips, and the temperature sensitivity of the optical fiber, are being worked on, along with other problems. Progress in these and other developments is pointing the way to the realization of a new generation of completely solid-state control system combining CCD star and sun sensors with fiber-optic gyroscopes and micromachined accelerometers.

6. NAVIGATION AND GUIDANCE

6.1 Introduction

Navigation of an aerospace vehicle has traditionally been defined as the process of determining the current position and the predicted path of the vehicle and correcting that path so that it stays within acceptable limits of some desired trajectory. The latter portion of the navigation function is referred to as guidance. (Sometimes, though, navigation and guidance are defined as two distinct functions, with navigation comprising only flight-path determination.)

Navigation of a space vehicle is a complex process involving the collection of data containing information about the position and velocity of the vehicle (and other pertinent quantities as well), followed by the processing of these data to yield estimates of the vehicle's position and velocity as functions of time. Further computation is then needed to plan corrections for the trajectory dispersions away from the desired flight path that inevitably occur. The computational process requires accurate modeling of the motions of the vehicle and various planets and natural satellites and accurate modeling of the observational data. The following sections describe the dynamic modeling of a spacecraft's orbital motion, the various observational data that may be available for navigational purposes, and the processing of these data to estimate the spacecraft's position and velocity and to compute trajectory-correction maneuvers.

6.2 Modeling of Spacecraft Orbital Motion

The acceleration experienced by an orbiting spacecraft can be attributed to a number of effects. The dominant effect is usually gravitational, with one celestial body at a time typically exerting the dominant gravitational influence. The orbital motion of a spacecraft (or a natural body) subject to the influence of a single massive object with a spherically symmetric mass distribution can be described by simple geometric shapes such as ellipses and hyperbolas, as described several centuries ago by Kepler, Newton, and others. Most other forces, gravitational and nongravitational, are small in relative terms and perturb the orbit slightly from these simple geometric shapes. Accurate propagation of predicted spacecraft orbits generally requires numerical integration.

Additional gravitational effects that must be taken into account for the accurate description of spacecraft motion include the gravitational influence of nondominant bodies (such as the moon and the sun for an Earth-orbiting spacecraft), asymmetries in the gravitational field of the dominant body, and (in very high-precision applications) general relativistic effects. For example, the oblateness of the Earth, induced by its spinning motion, causes significant changes in the orbit of a low-altitude spacecraft over a period of time. This effect may be used to advantage to make the orbit plane of the spacecraft precess in synchronism with the Earth's orbital motion about the sun, so that the ground track of the spacecraft is always illuminated in the same way. The Earth and other terrestrial bodies in the solar system possess inhomogeneous mass distributions; hence they have complex gravitational fields, which are often described by spherical harmonic expansions. Coefficients in these expansions are often estimated as part of the navigation process. General relativity causes subtle differences in the motion of a spacecraft and in the propagation of radio signals relative to what would be expected on the basis of classical physics. Today's measurement systems are sufficiently sensitive to detect some of these small effects and consequently can be used to test predictions of general relativity.

From time to time, the rocket engines on board a spacecraft must be fired to modify the trajectory of the vehicle. In addition, the motion of a spacecraft is affected by mass expulsion that cannot be precisely modeled or predicted in advance. The attitude of a planetary spacecraft is typically maintained by firing small rocket engines or gas jets. Although these firings are intended to change the orientation of the vehicle, they also change its trajectory. In addition, gas leaks unrelated to commanded attitude-control maneuvers can cause deviations in the trajectory. All of these propulsive effects, whether intended or unintended, are often estimated as part of the navigation process.

Spacecraft in low orbits about atmosphere-bearing bodies are subject to aerodynamic drag perturbations. The resulting accelerations are relatively small except in launch and reentry applications and can be estimated as part of the navigation process.

Solar-radiation pressure exerts a small perturbing influence on orbiting spacecraft. The resulting acceleration can be modeled, to the extent that the absorptivity and reflectivity properties of the vehicle are known in advance, or can be estimated from navigational data.

6.3 Types of Navigational Data

The data that are collected and processed for navigating a spacecraft may be broadly classified as radiometric, optical, or inertial. The relative values of these classes of data and the specific forms taken on by each vary considerably from application to application. Missions to other planets in the solar system impose some of the most stringent navigational challenges yet encountered in space exploration. Consequently, representative navigational data types will be discussed first in the context of NASA's unmanned planetary program. Differences associated with Earth-orbiting missions will then be addressed.

Some of the measurements used to navigate planetary spacecraft are derived from the radio link between the spacecraft and the Earth. The Earth-based part of the telecommunications system is a network of large radio telescopes, or tracking stations, located geographically so that virtually any distant spacecraft is continuously in view. There are three principal tracking station sites, located in California, Spain, and Australia. Each site is equipped with several antennas that are 34 or 70 m in diameter. The tracking stations maintain a radio-communication link with the Space Flight Operations Facility at the Jet Propulsion Laboratory in Pasadena, California. The facility originates commands to the spacecraft and receives data from the spacecraft by way of the tracking stations.

The traditional radio-navigation measurements for planetary missions consist of two-way Doppler and range, acquired at S- and X-band frequencies by the Deep Space Network. The simplest form of Doppler measurement employs a downlink, or one-way mode, in which the frequency of the signal received by the tracking station is compared with the best estimate of the frequency of the signal that was sent by the spacecraft. Since the clocks on board the spacecraft and at the tracking station do not run at precisely the same rate (with the latter much more accurate than the former), the accuracy of one-way Doppler data is limited. Instead, a two-way coherent transmission mode is normally used, in which the tracking station continuously transmits and receives. The signal transmitted from the Earth is received by the spacecraft's radio system, which increases its frequency by some multiple (to avoid interference between the outgoing and incoming signals) and coherently transmits it back to the Earth (by means of a phase-locked loop). If the distance between the spacecraft and the tracking station is increasing (decreasing), the spacecraft radio system receives and subsequently retransmits a signal whose frequency is shifted to a slightly lower (higher) value by the Doppler effect. On the downlink trip from the spacecraft to the Earth, the frequency of the signal received by the tracking station is again modified by the Doppler effect. The difference between the frequency received by the tracking station (compensated for the deliberate downlink-uplink frequency multiplication in the on-board transponder) and the frequency transmitted at a time corresponding to one round-trip earlier is a measure of the rate of change in the distance between the spacecraft and the tracking station. (In practice, the frequency of the received signal is compared with the frequency of the signal currently being transmitted instead of the signal transmitted one round-trip time earlier. The error introduced by this approximation is negligible because the stability of the transmitting frequency, better than one part in 10^{13}, prevents any significant buildup of frequency error during a time interval of one round trip.) The Doppler shift can be measured in 1-min counts to an accuracy of about 0.1 mm/s.

The other conventional radio-navigation measurement, range, is obtained by determining the round-trip time of the radio signal. The current ranging accuracy is about 10 ns in round-trip time, which is equivalent to 1–2 m in one-way range.

In recent years the technique of very long-baseline interferometry (VLBI) has been developed for use as a navigation data source. VLBI measures the difference in signal reception times at two widely separated ground stations to an accuracy of several nanoseconds. When compared with the signal reception time difference from a background extragalactic radio source (or quasar), these measurements allow the spacecraft direction to be determined to an accuracy of 50 nrad with respect to the direction to the background quasars.

The discussion above has concentrated on the use of radio navigation in deep-space missions. Range and Doppler data, in various forms, are also used in navigating Earth-orbiting missions. The radio links can be one way or two way and can be from a spacecraft to a ground tracking station or from one vehicle to another. Interferometric data types are being considered for use in certain high-precision Earth-orbiting applications. Spacecraft in relatively low orbits can be tracked by ground-based radar.

In many planetary missions, optical measurements can be obtained from the imaging system on board the spacecraft. These instruments, although originally designed to meet the stringent high-resolution and dynamic-range specifications required for scientific imaging of the planets, have also proved to be excellent for optical navigation. Images of the target planets and their natural satellites against stellar backgrounds yield valuable navigational data on approach to these planets. Spacecraft angular position fixes to accuracies of 5 μrad relative to the target body can be obtained.

Optical observations are sometimes used to track Earth-orbiting spacecraft also. Telescopes on the Earth can be used to track satellite motions relative to background stars.

Inertial navigation sensors include gyroscopes and accelerometers. Of these, only accelerometers measure the translational motion of a vehicle. They detect all forces acting on a spacecraft that are not gravitational in nature. They are essential in launch and reentry applications, where aerodynamic forces are quite large. They are also useful for measuring (and sometimes controlling) vehicle thrusting. In most cases, they are not sufficiently accurate to measure the nongravitational forces experienced by a nonthrusting spacecraft far from a planetary atmosphere.

6.4 Computation of Spacecraft Orbits

For most spacecraft, navigational computations are performed on the Earth. In some applications, though, particularly where short reaction times or on-board autonomy may be required, the computations are performed on board. As in the previous section, the techniques used in NASA's planetary missions will be described. The computational procedures used in connection with Earth-orbiting spacecraft are generally similar.

Figure 17 illustrates the elements of measurement, communications, computation, and propulsion that comprise the overall navigation system that is employed for NASA's deep-space missions. The left side of the diagram shows a spacecraft acquiring optical images of a target body. Optical data are encoded at the spacecraft and telemetered to Earth. Radio data (Doppler, range, and VLBI) are extracted from the incoming radio signal at the tracking stations. Also shown on the left is the capability for spacecraft propulsion. The right side of the diagram depicts the ground processing system used to compute the spacecraft orbit and trajectory-correction parameters.

Several major functions must be performed as part of the orbit determination process. The first is the accurate numerical integration of the trajectory from assumed initial conditions, along with partial derivatives of the position and velocity components with respect to parameters on which their evolution depends. Next, simulated measurements are computed corresponding to each actual observation, based on the modeled trajectory. Partial derivatives of the measurements with respect to trajectory initial conditions, and parameters affecting the motion of the spacecraft and the observations, are computed. The estimates of the initial conditions and other parameters are then adjusted until the differences between the actual measurements and the expected measurements, based on the values assumed for the parameters, are minimized in some sense. The estimation algorithm may be a straightforward least-squares procedure, or it may be sequential in time, with stochastic accelerations of the spacecraft modeled as piecewise-constant functions. The estimation process can be performed iteratively until convergence is obtained. The final

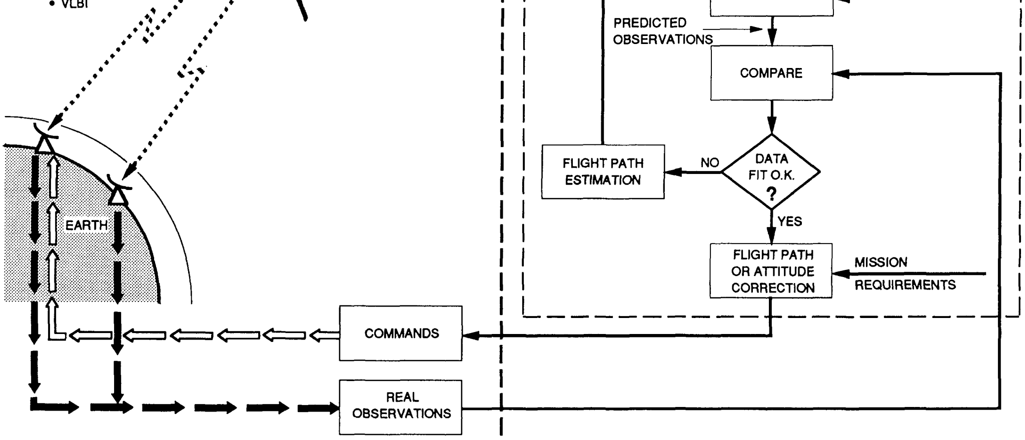
OPTICAL TRACKING BY SPACECRAFT IMAGING SYSTEM
• TARGET-STAR ANGLES
SPACECRAFT PROPULSION SYSTEM
RADIO TRACKING BY GROUND ANTENNAS
• DOPPLER
• RANGE
• VLBI
SPACECRAFT EARTH COMMUNICATIONS SYSTEM
EARTH
COMMANDS
REAL OBSERVATIONS
GROUND DATA PROCESSING SYSTEM
TRAJECTORY MODEL
PLANETARY EPHEMERIDES
CALIBRATIONS
EARTH ROTATION
STATION LOCATIONS
OBSERVATIONAL MODEL
PREDICTED OBSERVATIONS
COMPARE
FLIGHT PATH ESTIMATION
NO
DATA FIT O.K. ?
YES
FLIGHT PATH OR ATTITUDE CORRECTION
MISSION REQUIREMENTS

FIG. 17. Overall navigational system.

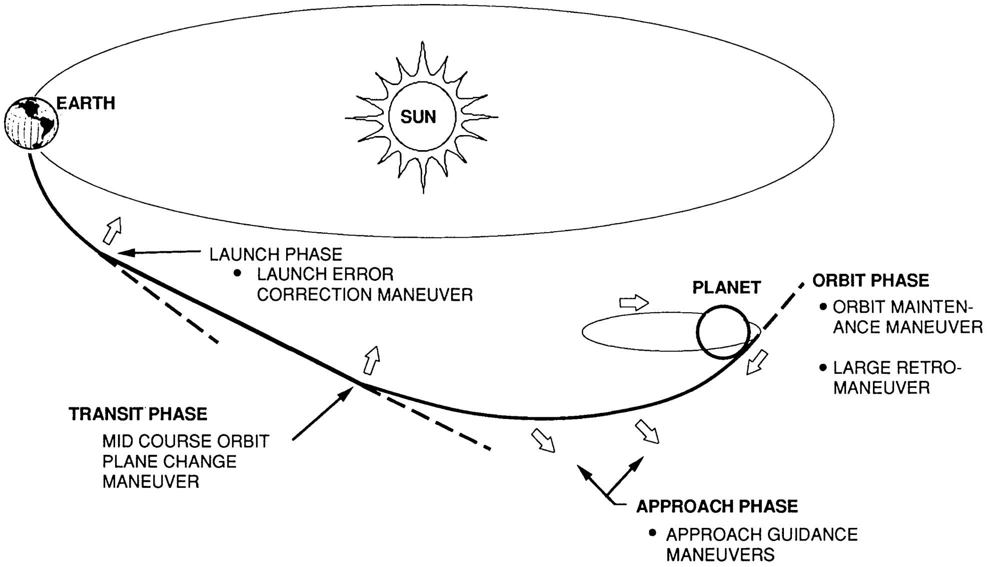
EARTH
SUN
LAUNCH PHASE
• LAUNCH ERROR CORRECTION MANEUVER
PLANET
ORBIT PHASE
• ORBIT MAINTENANCE MANEUVER
• LARGE RETRO-MANEUVER
TRANSIT PHASE
MID COURSE ORBIT PLANE CHANGE MANEUVER
APPROACH PHASE
• APPROACH GUIDANCE MANEUVERS

FIG. 18. Trajectory-correction maneuvers.

product of the estimation process is a numerically integrated trajectory that best fits the observations.

The high accuracy required for deep-space navigation is achieved with

1. precise observational measurements and precise knowledge of the timing of the measurements,
2. numerically accurate computational algorithms, and
3. a precise modeling knowledge of all physical phenomena that affect the values of the observational measurements.

These phenomena include those that affect the values of the measurements directly and those that affect the values of the measurements through their influence on the motion of the spacecraft. Examples of these phenomena that require highly accurate knowledge include the locations of the tracking stations on the Earth, the Earth's rotational motion, the motions of the Earth, moon, planets, and their natural satellites through the solar system, the gravity-field structure of the Earth and other solar system bodies, and the effects of the transmission media on the radio signal. All of these phenomena are modeled in the orbit determination software.

6.5 Trajectory-Correction Maneuvers

The best-estimate trajectory, derived from navigational data, serves as the basis for computation of velocity-correction parameters. A stylized interplanetary flight path is depicted in Fig. 18, in which several propulsive maneuvers that may be required to navigate a typical deep-space mission are indicated. Launch injection errors are usually corrected in the first month of travel. Often, during the transit from Earth to the distant target, an orbit-plane adjustment is required to redirect the spacecraft flight path to intersect the path of the moving planetary target. As the spacecraft approaches its distant target, one or two maneuvers are required to guide the spacecraft to the desired encounter geometry. If the spacecraft is to orbit the encountered planet (or other celestial body), a large retromaneuver is required near the point of closest approach. Subsequent orbit maintenance is usually required throughout the useful lifetime of a planetary orbiter.

In all space missions, the timing of trajectory correction maneuvers is governed by the desire to control the trajectory as accurately as possible, while minimizing the expenditure of propellant. Precise determination of the orbit of the spacecraft is essential prior to the computation of each maneuver. The error in this determined orbit usually dominates the flight-path error after the maneuver.

GLOSSARY

Astronomical Unit (AU): The average distance of the Earth from the sun (1.5×10^8 km).

BDC: Bidirectional converter. A power converter that operates in two directions to control a spacecraft power bus automatically by shunting excess power (e.g., from a solar array) to a battery, or supplying power from the battery to the bus, to maintain bus voltage.

Bipropellants: A pair of substances that are reacted in a rocket with the release energy in the reaction products.

Bus: Spacecraft central module containing subsystem electronics; typically excludes the scientific payload.

Charge-Coupled Device (CCD): Linear or area array of photosensitive elements that transfers the image charge from element to element during readout.

Doppler Shift: Change in the frequency with which electromagnetic (or sound) waves from a given source reach an observer when the source and observer are in motion with respect to each other.

Image Dissector (ID): Photomultiplier tube incorporating an electrostatically scanned internal aperture that permits an image to be scanned one point at a time.

Monopropellant: A single substance that decomposes in a rocket with the release of energy in the decomposition products.

Photomultiplier: Vacuum tube incorporating a photosensitive cathode and electron-multiplying stage that can detect extremely low light levels (some detect individual photons).

RTG: Radioisotope thermoelectric generator, employing thermal-to-electric converters to produce electric power from heat released by radioactive isotopes.

Specific Impulse: The impulse produced by a rocket per mass of propellant expelled (a measure of propellant efficiency).

Vacuum Specific Impulse: Specific impulse for a rocket operating in a vacuum.

List of Works Cited

Anonymous (1979), *Functional Requirements Book, Voyager Spacecraft*, JPL Document 618-205, Pasadena: Jet Propulsion Laboratory.

Anonymous (1988), *Functional Requirements Book, Project Galileo*, JPL Document 625-205, Pasadena: Jet Propulsion Laboratory.

Deters, R. A. (1979), *A User's Guide for the Standard Star Tracker*, AAS Paper 79-022, American Astronautical Society Meeting, San Diego.

Hart, D. (1987), *The Encyclopedia of Soviet Spacecraft*, New York: Bison.

Stanton, R. H., Alexander, J. W., Dennison, E. W., Glavich, T. A., Hovland, L. F. (1987), *Opt. Eng.* **26**, 930–938.

Yenne, B. (1985), *The Encyclopedia of U.S. Spacecraft*, New York: Bison.

Further Reading

Anonymous (1987), *Nuclear Power In Space*, U.S. Department of Energy Brochure, No. NE-0071, Washington, D.C.: U.S. G.P.O.

Anonymous (1987), *Atomic Power in Space: A History*, U.S. Department of Energy Report No. NE-32117-H1, Springfield, VA: National Technical Information Service, U.S. Department of Commerce.

Barrere, M., Jaumotte, A., De Veubeke, B. F., Vandenkerckhove, J. (1960), *Rocket Propulsion*, New York: Elsevier.

Battin, R. H. (1987), *An Introduction to the Mathematics and Methods of Astrodynamics*, New York: American Institute of Aeronautics and Astronautics.

Bennett, G. L. (1989), "A Look at the Soviet Space Nuclear Power Program," in: W. D. Jackson (Ed.), *Intersociety Energy Conversion Engineering Conference, 24th, Washington, D.C.*, New York: IEEE.

Brewer, G. R. (1970), *Ion Propulsion. Technology and Applications*, New York: Gordon and Breach.

Bussard, R. W., DeLauer, R. D. (1965), *Fundamentals of Nuclear Flight*, New York: McGraw-Hill.

Casani, J. R., Atkins K. L., Stapfer, G. (1987), "Applications of Radioisotope Thermal Generators in the United States Space Program," in: Irene Macke (Ed.), *Transactions of the American Nuclear Society*, Vol. 55. La Grange, IL: American Nuclear Society.

Caveny, L. H. (Ed.) (1984), *Orbit-Raising and Maneuvering Propulsion: Research Status and Needs*, Progress in Astronautics and Aeronautics, Vol. 89. New York: American Institute of Aeronautics and Astronautics.

Cochran, C. D., Gorman, D. M., Dumoulin, J. D. (Eds.) (1985), *Space Handbook*, Maxwell Air Force Base, Alabama: Air University Press.

Corliss, W. R. (1965), *Space Probes and Plane-*

tary Exploration, Princeton: Van Nostrand.

Cornelisse, J. W., Schoyer, H. F. R., Wakker, K. F. (1979), *Rocket Propulsion and Spaceflight Dynamics*, London: Pitman.

Deters, R. A. (1979), *A User's Guide for the Standard Star Tracker*, AAS Paper 79-022, American Astronautical Society Meeting, San Diego.

Dix, G. P., Voss, S. S. (1985), "The Pied Piper—An Historical Overview of the U.S. Space Power Reactor Program," in: M. S. El-Genk and M. D. Hoover (Eds.), *Proceedings of the First Symposium on Space Nuclear Power Systems, Albuquerque, NM, 1984*, Malibar, FL: Orbit Book Co.

Halpert, G., Surampudi, R., Rowlette, J. (1986), *The NASA Aerospace Battery Safety Handbook*, JPL Publication No. 86-14, Pasadena: Jet Propulsion Laboratory.

Hobbs, M. (1962), *Fundamentals of Rockets, Missiles and Spacecraft*, New York: Rider.

Jahn, R. G. (1968), *Physics of Electric Propulsion*, New York: McGraw-Hill.

Jordan, J. F., Wood, L. J. (1987), "Navigation, Space Mission," in *Encyclopedia of Physical Science and Technology*, Vol. 8. New York: Academic, pp. 742–767.

Krauthamer, S. (1988), *State of the Art of DC Power Distributions/Components for Space Applications*, NASA JPL Publication No. 88-26, Pasadena: Jet Propulsion Laboratory.

Melbourne, W. G. (1976), *Sci. Am.* **234**, 59–74.

Purser, P. E., Faget, M. A., Smith, N. F. (Eds.) (1964), *Manned Spacecraft: Engineering Design and Operation*, New York: Fairchild.

Stanton, R. H., Alexander, J. W. Dennison, E. W., Glavich, T. A., Hovland, L. F. (1987), *Opt. Eng.* **26**, 930–938.

Stuhlinger, E. (1964), *Ion Propulsion for Space Flight*, New York: McGraw-Hill.

Sutton, G. P. (1986), *Rocket Propulsion Elements*, 5th ed. New York: Wiley-Interscience.

Timnat, Y. M. (1987), *Advanced Chemical Rocket Propulsion*, New York: Academic.

Wilding-White T. M. (1976), *Jane's Pocket Book of Space Exploration*, New York: Collier.

Wilson, A. (1987), *Solar System Log*, London: Jane.

ASTRONOMY

David Layzer, *Department of Astronomy, Harvard University, Cambridge, Massachusetts, U.S.A.*

	Introduction	99
1.	**Astronomical Data**	100
1.1	Sources of Astronomical Data	100
1.2	From 2D to 3D	100
1.3	Stellar Spectra	103
1.4	Dimming and Reddening of Starlight	105
1.5	Measuring the Velocities of Astronomical Objects	106
2.	**Gravity and the Hierarchy of Self-Gravitating Systems**	106
2.1	Newtonian Gravity	106
2.2	Mass, Size, and Motion	107
2.3	Tidal Effects	107
2.4	Self-Gravitating Clusters	108
2.5	Mass, Radius, and Cohesive Energy	108
3.	**Astronomical Systems**	109
3.1	The Solar System	109
3.2	Origin of the Solar System	112
3.3	Stars and Star Clusters	113
3.4	The Interstellar Medium	114
3.5	Giant Molecular Clouds	114
	Works Cited	115
	Further Reading	115

INTRODUCTION

The goal of astronomy is to understand the nature, origin, and evolution of the physical universe. Astronomy is a physical science, relying on physical measurements for its data and on physical theories for their interpretation. It is also a historical science, because, as in biology, questions of origin and evolution occupy central positions.

In modern astronomical research, theory and observation interact closely. Data are given meaning and coherence by the theories used to interpret them. Observational evidence guides the construction and revision of theory. Without theory, observation is blind; without observation, theory degenerates into story-telling.

An internally consistent theoretical story is likely to be true to the extent that it meets two criteria: it must be consistent with strongly confirmed physical and astronomical theories, and it must be consistent with relevant observational data. As observations provide increasingly precise and increasingly varied data, they progressively narrow the range of viable hypotheses. At the same time, these hypotheses become increasingly secure. Our knowledge of the astronomical universe may be likened to a ball whose core is made up of strongly corroborated theories and whose ever-expanding surface, the interface between knowledge and ignorance, is made up of weakly corroborated hypotheses and empirical generalizations. The hypotheses and empirical generalizations at the surface come and go, but a few of them survive to become part of the growing core.

This article is divided into three parts. The first part describes the various sources of astronomical data and discusses the way in which astronomers analyze these data to derive information about the distribution in space and the physical properties of astronomical systems and the intervening medium. The second part discusses in general terms how gravitational attraction shapes the astronomical universe. The third part discusses specific astronomical systems, with particular attention to the solar system and excluding galaxies and systems of galaxies, which are discussed in GALAXIES AND COSMOLOGY. Stars and stellar evolution are discussed in greater detail in ASTROPHYSICS.

Astronomy has undergone an information explosion during the past 40 years. The present article, however, stresses the most basic methods and ideas, and their evolution, rather than results.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. ASTRONOMICAL DATA

1.1 Sources of Astronomical Data

Astronomical observation is largely a matter of collecting, measuring, and analyzing light incident on the Earth. Until the middle of the twentieth century, astronomers relied almost entirely on ground-based visual and photographic observations. Photography extends the wavelength range of detectable light beyond the range of the human visual system, but only slightly. Nowadays astronomers collect light over a range of wavelengths extending from less than a millionth of a micron (1 micron$=10^{-6}$ m) to more than a hundred meters; the accessible spectral range has been expanded from a single octave (a factor of 2 in wavelength or frequency) to well over 14 decades (powers of ten). Differences in the information obtained at different wavelengths are exemplified in Fig. 1. Gamma rays, x rays, ultraviolet light, visible light, infrared light, millimeter and microwave radiation, and radio radiation are collected and analyzed not only by ground-based instruments but also by instruments based on balloons, rockets, and satellites.

There are three other important carriers of astronomical data: meteorites, cosmic rays, and neutrinos.

Chemical analyses of meteorites yield estimates of the age of the solar system and provide clues to its origin.

Cosmic rays consist of highly energetic protons, heavier nuclei, and electrons. They contain information about the processes that accelerate charged particles in the Sun, in supernovae, in interstellar space, and (possibly) in external galaxies.

Neutrinos are emitted in the nuclear reactions that burn hydrogen to helium in the cores of the Sun and other stars. Because neutrinos interact so weakly with matter, nearly all of them escape into interstellar space; but for the same reason they are exceedingly difficult to detect. Experiments to detect neutrinos from the Sun have, however, been successful. The results of these experiments have not been in accord with theoretical expectations. Their interpretation is still controversial. Theories of stellar evolution predict that neutrinos are also copiously produced in supernova explosions. This prediction was confirmed by successful detections of neutrinos emitted by Supernova 1987a (see, e.g., Abell *et al.*, 1991).

Gravitational radiation is another potential source of astronomical data. Einstein's theory of gravitation (general relativity) predicts that certain objects, including binary stars and asymmetrically imploding and exploding objects, emit gravitation waves. These waves interact very weakly with matter. Since the early 1960s, progressively more sensitive gravitational-wave detectors have been developed, but until now (1991) no confirmed detection has been reported.

Finally, space probes have been able to explore and send back (or, in the case of the Moon, bring back) new and valuable information about the planets, their satellites, and the contents of interplanetary space.

1.2 From 2D to 3D

The earliest astronomers mapped the sky, noting the position and brightness of stars on the spinning celestial sphere and the motions (angular displacements) of the Sun, Moon, and planets among the stars. To this map of naked-eye stars and planets astronomers have added objects up to a million or more times as faint as the faintest naked-eye star—new planets and satellites, stars, multiple stars, star clusters, galaxies, galaxy clusters, and clouds of gas and dust. They have measured the brightness of discrete objects and made continuous brightness maps in broad and narrow spectral bands over 14 decades of the electromagnetic spectrum. They have measured angular displacements of thousands of relatively nearby stars. They have categorized, counted, and cataloged stars, star clusters, galaxies, and galaxy clusters in untold numbers. Most important of all, they have succeeded in translating these varied two-dimensional maps into three-dimensional models of the astronomical universe and the systems that compose it.

Aristarchus, in the third century B.C., and later Copernicus, in the 16th century A.D., used observations of planetary motions to construct a tolerably accurate scale model of the planetary system. Not until 1672 did astronomers succeed in estimating the scale of the model. In that year French astronomers measured the distance between Earth and Mars by triangulating Mars from Paris and

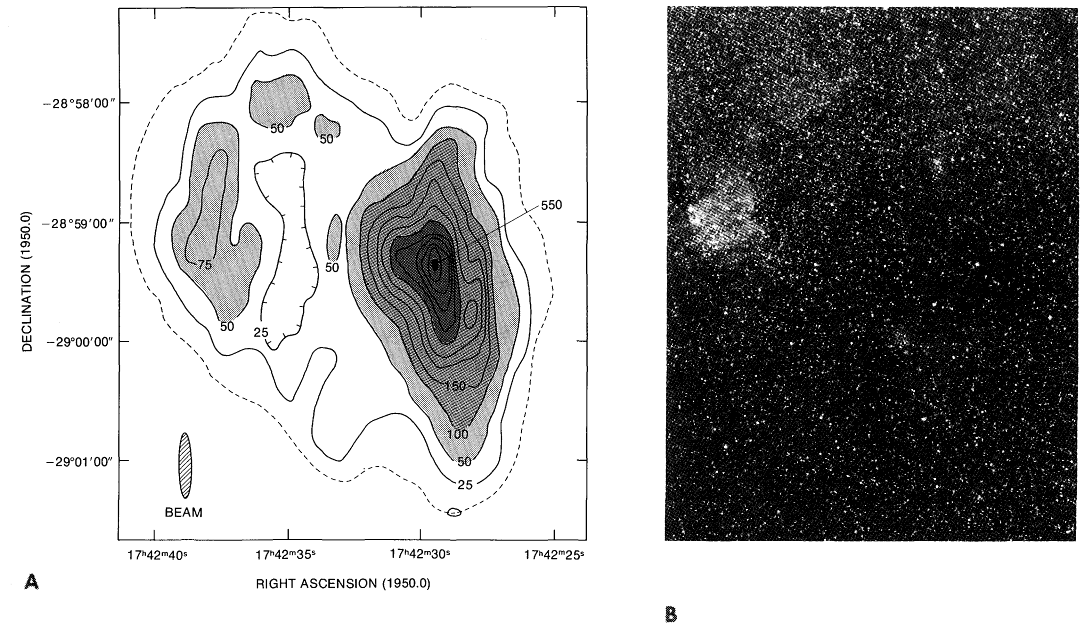

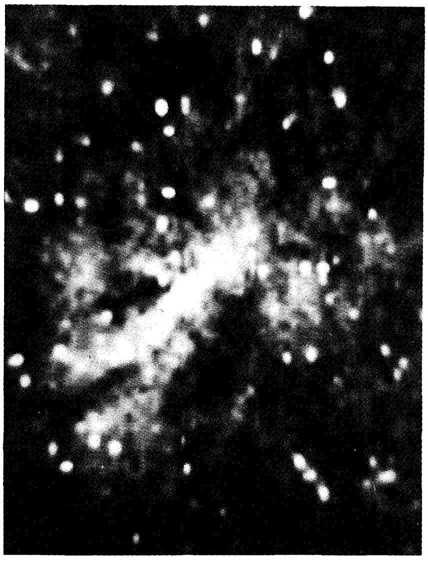

FIG. 1. Views of the center of the Milky Way galaxy at three different wavelengths: (A) 6 cm, (B) visible light, (C) 2.2 μm. The scale of (C) is somewhat different from that of (A) and (B). [Pasachoff and Kutner (1978), Fig. 25-4 and unnumbered.]

Cayenne (in French Guiana). At about the same time, Christiaan Huygens and James Gregory made independent estimates of the distance of the star Sirius, using the optical law that relates an object's apparent brightness to its distance (the inverse-square law) and making the qualitatively correct but quantitatively unreliable assumption that Sirius is intrinsically as bright as the Sun. (Many modern procedures for estimating distances rely on the same law, but use more reliable estimates of the distant object's intrinsic brightness.) The first successful triangulations of stars were accomplished in 1838 by Friedrich Bessel (61 Cygni), Thomas Henderson (Alpha Centauri) and Friedrich Struve

(Vega). These measurements created the first rung of the astronomical distance ladder.

Knowing the distances and hence the intrinsic brightnesses of nearby stars, astronomers sought indicators of intrinsic brightness that could be used to estimate distances (via the inverse-square law relating true and apparent brightness) of stars beyond the range of triangulation. They discovered that a star's color, which is a direct indicator of its *effective temperature,* the temperature of the layers from which we receive light, is also a reliable indicator of intrinsic brightness, provided it is combined with subtle spectral features that serve to distinguish bright stars with distended envelopes (giants) from faint stars (dwarfs) with the same effective temperature. (A plot of apparent brightness against color or effective temperature is known as a Hertzsprung-Russell diagram, after Ejnar Hertzsprung and Henry Norris Russell, who demonstrated its utility early in the twentieth century.)

The next, and in some ways most crucial, rung of the distance ladder was added in 1912–1913. Henrietta Swan Leavitt (1912), a research assistant at the Harvard College Observatory, measured the periods of light variation of certain variable stars, called Cepheids, in the Small Magellanic Cloud, an irregularly shaped object in the southern sky containing stars, star clusters, and bright gas clouds. She found that the measured periods were tightly correlated with apparent brightness, and hence, because all the stars in the Cloud could be assumed to be at nearly the same distance, with intrinsic brightness. The following year Hertzsprung (1914) succeeded in calibrating Leavitt's period-brightness relation by estimating the distances of 13 relatively nearby Cepheids in the Milky Way. The period-brightness relation for Cepheids played the key role in establishing the truth of a hypothesis put forward by Immanuel Kant in 1755: that the "spiral nebulae," fuzzy, often S-shaped patches of light, are distant stellar systems comparable in size to our own stellar system, whose central plane is defined by the Milky Way. An example is shown in Fig. 2. (Such stellar systems are now called galaxies.) With greatly improved calibration, the period-brightness relation (see Fig. 3) is an important rung in the modern distance ladder.

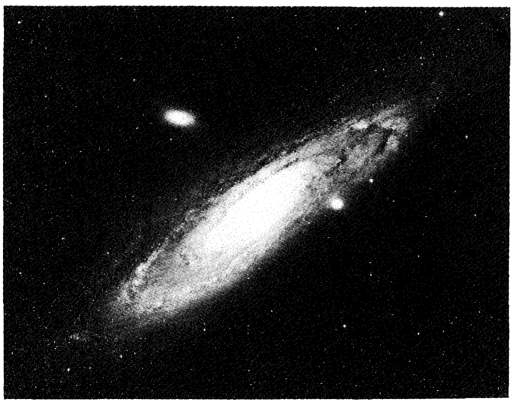

FIG. 2. M31, the Great Galaxy in the constellation Andromeda, a type Sb spiral, with its accompanying elliptical galaxies NGC 205 (upper left) and M32 (lower right). [Pasachoff and Kutner (1978), Fig. 26-3.]

Individual stars can be resolved—even with modern telescopes—only in relatively nearby stellar systems. In the 1920s Edwin Hubble began to use characteristics of whole galaxies as indicators of their intrinsic brightness (Hubble, 1936). The most reliable characteristic of this kind was discovered by R. Brent Tully and J. Richard Fisher (1977). They found that the width of a certain spectral line is a precise indicator of a spiral galaxy's intrinsic brightness. This line, at a wavelength of 21 cm in the radio region of the electromagnetic spectrum, is emitted by clouds of neutral atomic hydrogen concentrated toward the central planes of spiral galaxies.

Many other kinds of objects have been used as distance indicators, including novae and supernovae. Novae are now known to be close binary stars whose interaction gives rise to the characteristic brightenings of the system. Supernovae are exploding, massive single stars. Theoretical calculations show that stars whose masses lie between 10 and 70 solar masses explode when the supply of nuclear fuel in their core has been exhausted. These stars become so-called type-II supernovae; they are found in the disks of spiral galaxies. Type-I supernovae, which are found in the spheroidal halos of spiral galaxies and in elliptical galaxies, are less well understood. They are thought by some workers to result from explosions of stars with masses between 4 and 8 solar masses.

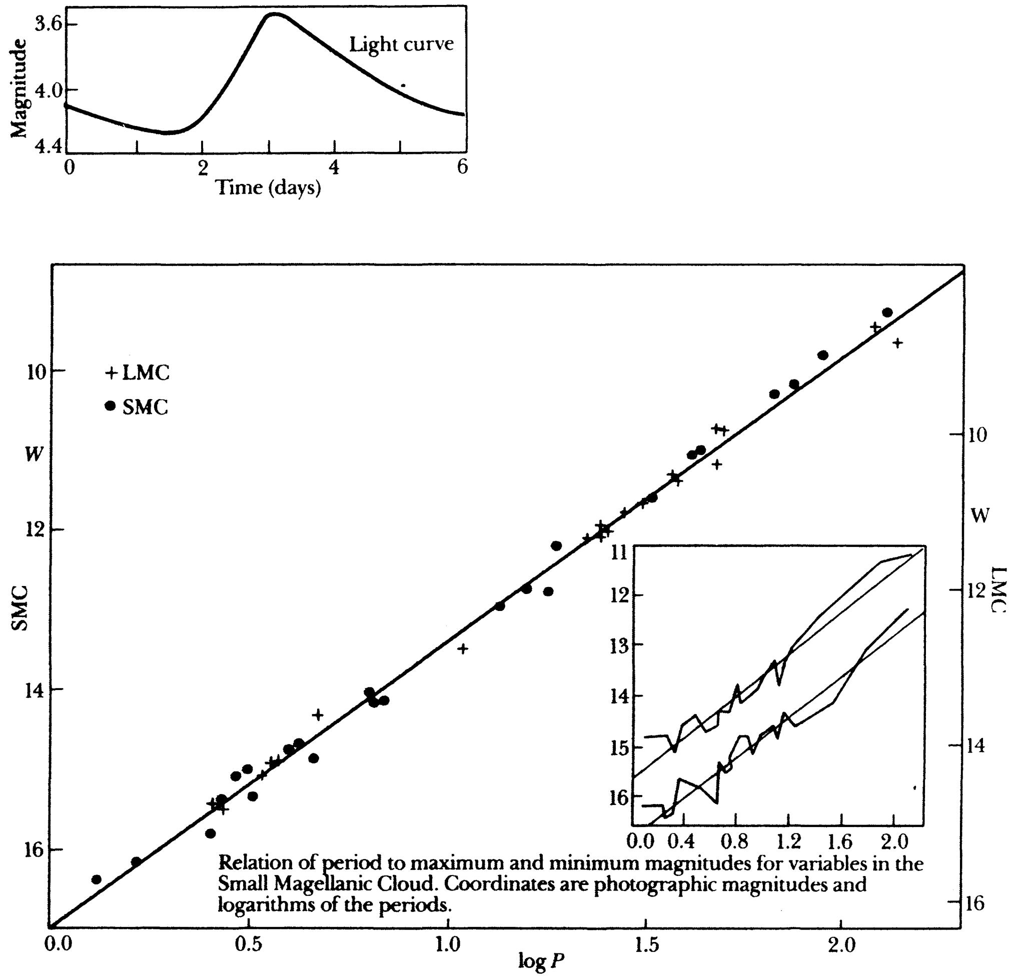

FIG. 3. Top, light curve of the star δ Cephei. Bottom, Leavitt's original plot of brightness vs period for Cepheids in the Small Magellanic Cloud (inset), and a modern plot for Cepheids in the Large and Small Magellanic Clouds (LMC and SMC, respectively). The ordinate is W, a linear combination of blue and apparent visual magnitudes that is insensitive to reddening by interstellar matter. [Layzer (1984).]

1.3 Stellar Spectra

A prism or diffraction grating spreads light into a spectrum, separating rays of different colors or frequencies. The spectra of astronomical objects are rich sources of information about their physical characteristics.

Stars, like incandescent solids, emit continuous spectra whose shape (which defines the distribution of radiant energy in frequency or wavelength) is completely determined by the temperature of the emitting surface. Physical theory predicts that the peak wavelength, which defines the color of the light, is inversely proportional to the (Kelvin or absolute) temperature. The temperature of sunlight is around 6000 K.

By contrast, a hot, transparent gas, like the gas in a fluorescent lamp, emits light only in very narrow, well separated bands of frequency or wavelength. (These narrow bands are called spectral lines. See ATOMIC SPECTROSCOPY.) For example, heated sodium vapor emits two closely spaced, narrow bands of yellow light. During the 19th century astronomers discovered that "spiral nebulae" have continuous spectra like stars (see Fig. 4), thereby confirming Kant's 100-year-old hypothesis that they are stellar systems. Other nebulae, which have bright-line spectra, are

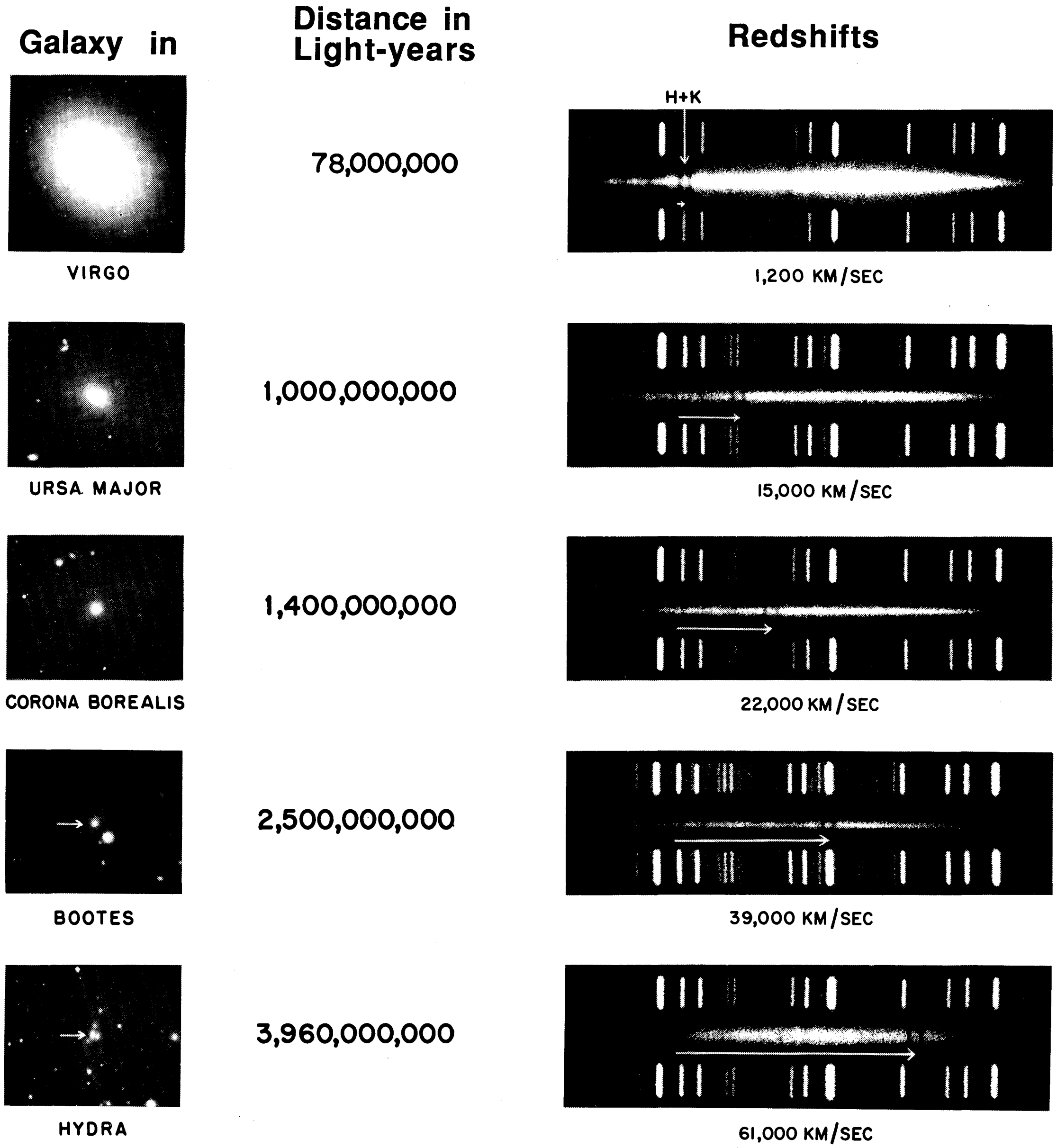
Galaxy in
Distance in Light-years
Redshifts
H+K
VIRGO
78,000,000
1,200 KM/SEC
URSA MAJOR
1,000,000,000
15,000 KM/SEC
CORONA BOREALIS
1,400,000,000
22,000 KM/SEC
BOOTES
2,500,000,000
39,000 KM/SEC
HYDRA
3,960,000,000
61,000 KM/SEC

FIG. 4. Spectra (right) for the galaxies shown on the left. Distances, in the middle column, are based on a value of 55 km/(s Mpc) for Hubble's constant. [Pasachoff and Kutner (1978), Fig. 26-20.]

true gas clouds, excited by embedded hot blue stars. Most of these gaseous nebulae belong to our own stellar system, but some are bright enough to be seen in neighboring galaxies. There are also dust clouds that, like the planets, shine by light reflected from nearby (or embedded) stars.

In 1823 Joseph Fraunhofer discovered that the spectra of the Sun and bright stars are crossed by very narrow dark lines. These lines represent light absorbed and scattered by the tenuous outer layers of a star. Every chemical element and every molecule produces its own characteristic pattern of spectral lines. In the spectrum of a hot transparent gas cloud these lines appear bright. If the gas lies between an incandescent source and the observer, it produces dark lines at exactly the same frequencies or wavelengths. The relative strengths of the lines, and hence the appearance of the spectrum, is determined by the temperature of the layers where the spectrum is formed. Conversely, from the appearance of the spectrum one can infer the temperature of the layers where the spectrum is formed. The temperature deduced in this way coincides with the temperature deduced from the continuous spectrum (the "color temperature")—provided the characteristics of the light have not been altered by its passage through an intervening medium. In fact, as discussed below, interstellar space is populated by clouds of dust and gas, which leave their imprint on starlight. Dust particles of sizes comparable with the wavelength of visible light (around a tenth of a micron) scatter blue light more effectively than red light. Hence light that has traversed a transparent dust cloud appears reddened. The relative strengths of the dark spectral lines, however, are not altered, and from them the true temperature of the emitting layers can be inferred. By relating the amount of reddening, as measured by the difference between the true and the apparent color of a distant object, to the amount of dimming, astronomers have been able to allow for the attenuation of light in its passage from source to observer.

Although the appearance of the line spectrum depends primarily on the temperature of the layers where the spectrum is formed, it also depends, though much more weakly, on the pressure or density. As mentioned above, by paying attention to subtle pressure-sensitive spectral features, astronomers have been able to distinguish bright stars with distended envelopes (giants) from faint stars (dwarfs) with the same surface temperature.

The appearance of a stellar spectrum depends not only on the temperature and pressure of the layers where the spectrum is formed but also on the relative abundances of the chemical elements and their isotopes in those layers. By subjecting careful and detailed spectral measurements to elaborate theoretical analyses, astrophysicists have been able to carry out quantitative chemical analyses of stellar atmospheres, interstellar gas clouds, and quasars. The results of such analyses have important implications for theories of cosmic evolution and for theories of the origin of the chemical elements.

1.4 Dimming and Reddening of Starlight

Long before the true nature of "spiral nebulae" had been definitely established, astronomers noticed a peculiarity in their distribution on the celestial sphere. Spiral nebulae are found everywhere except in the Milky Way itself, a narrow, irregular, bright band of stars and diffuse starlight centered on a great circle of the celestial sphere. In short, the spiral nebulae avoid the region of the sky where stars are most strongly concentrated. To those who accepted Kant's interpretation of the Milky Way as a great disk-like stellar system viewed from the perspective of an observer in a peripheral region, and his hypothesis that the spiral nebulae are distant stellar systems comparable to our own and lying well beyond its limits, the interpretation of this state of affairs was obvious: external stellar systems that lie in the direction of the Milky Way must be hidden from view by a layer of dust strongly concentrated toward our system's central plane, the plane of the Milky Way. Spiral nebulae seen edge on do indeed have conspicuous dark central lanes. Nevertheless, until the 1920s most astronomers preferred to believe that we have a clear, unimpeded view of the heavens—that the light we receive from stars and galaxies is dimmed by distance alone.

In the late 1920s, Edwin Hubble, armed with data from photographs made with the world's largest telescope, the 100-in.-diam reflector on Mt. Wilson in California, showed that the hypothesis of a thin, uniform obscuring layer centered on the central plane of our galaxy could explain not only the absence of external galaxies along lines of sight that make small angles with this plane but also the rate at which the density of galaxy images on the plane of the sky decreases with decreasing elevation of the line of sight relative to the plane. Around the same time, Robert J. Trumpler (1925) found clear evidence of dimming by a thin layer of galactic dust in the linear diameters of star clusters that lie close to the plane of the Milky Way. Trumpler deduced the linear diameters of these clusters from their angular diameters and estimated distances. He found that diameters estimated in this way increased systematically with estimated distance in all directions from the observer, an effect that could most plausibly be attributed to a systematic error in the distance estimates, arising from the dimming of starlight by intervening dust.

Finally, Joel Stebbins (1933), using the newly developed technique of photoelectric photometry, found that globular star clusters become systematically redder with decreasing elevation of their line of sight relative to the central plane of the Milky Way. The reddening of starlight indicated that a substantial fraction by mass of the solid particles responsible for dimming the light have dimensions comparable with the wavelengths of visible light. Because reddening and dimming are related in a known manner, measurements of reddening allow one to predict how much the light from a star or stellar system is dimmed in any given direction and thus to correct distance estimates for the effects of dimming. In this way Stebbins was able to deduce that the globular clusters form a spherical system strongly concentrated toward its center. The Sun lies on the central plane of this system, roughly 30 000 light years from the center.

1.5 Measuring the Velocities of Astronomical Objects

The frequency of monochromatic (single-frequency) light, or of any other periodic signal, is shifted by relative motion along the line joining source and observer. The fractional change in frequency or wavelength is equal to the ratio between the line-of-sight velocity and the speed of light. The shift is toward the red for recession (increasing separation), toward the blue for approach. This effect is a simple consequence of the finite speed of propagation of light. Although it is named after Christian Doppler, who demonstrated it for sound waves early in the 19th century, it was discovered in 1675 by Ole Roemer, who attributed systematic variations in the period of Io's eclipses by Jupiter to the Earth's orbital motion and the finite speed of light. Roemer's analysis established for the first time that light signals travel with finite speed, and yielded a remarkably accurate estimate of this speed. The Doppler effect has continued to play an important part in astronomy. Astronomers deduce the line-of-sight velocities of stars, galaxies, quasars, and other astronomical objects by measuring the displacements in the frequencies or wavelengths of their spectral lines from the standard laboratory frequencies or wavelengths (see Fig. 4). Theory predicts and observation confirms that the fractional shift is the same at every frequency.

As discussed below, estimated line-of-sight velocities, interpreted in the light of Newton's theory of gravitation, form the basis for estimates of the masses of stars, galaxies, and galaxy clusters. Systematic redshifts in the spectra of distant galaxies led to, and continue to support, our present picture of an expanding universe. (See GALAXIES AND COSMOLOGY.)

2. GRAVITY AND THE HIERARCHY OF SELF-GRAVITATING SYSTEMS

2.1 Newtonian Gravity

The domain of astronomy coincides with the domain of gravity. Astronomical systems, from planets and their satellites to clusters and superclusters of galaxies, are held together by the mutual gravitation of their parts. Moreover, the cosmic expansion—the expansion of the universe as a whole—is a gravitational phenomenon (see GALAXIES AND COSMOLOGY).

Unlike the electrostatic forces that hold together atoms, molecules, liquids, and solids, and the strong forces that hold together atomic nuclei and their constituent protons and neutrons, gravity has no preferred scale. The scale-free character of gravity is a consequence of two properties enunciated by Newton. First, every particle attracts and is attracted by every other particle in the universe. The strength of attraction is proportional to the product of the masses. Second, the force of attraction diminishes slowly with increasing separation (as the inverse square of the separation). As a consequence of these two properties, the cohesive energy per unit mass of a stable self-gravitating system—a system that is neither expanding nor collapsing—is proportional to its average mass density (mass divided by volume) and to the square of its diameter. So a system of low average density can have a large gravitational cohesive energy per unit mass if it is large enough.

In principle, the average mass density of a self-gravitating system can have any value at all. The actual range of observed and inferred average mass densities is indeed very large. The most diffuse (and largest) self-gravitating systems have average mass densities around 10^{-28} g cm^{-3}. The most compact astronomical objects have mass densities ranging from around 1 g cm^{-3} for ordinary stars to around 10^{14} g cm^{-3} for neutron stars. So the mass densities of stable self-gravitating systems span a range of 42 decades. By contrast, the mass densities of ordinary solids and liquids span a range of about two decades. The average densities of atomic nuclei and nucleons span an even narrower (but higher) range.

2.2 Mass, Size, and Motion

Newton's theory of gravitation relates the motions within a self-gravitating system to the distribution of mass in the system and its size. For example, with the help of Newton's theory we can deduce the Sun's mass (2×10^{30} kg) from the orbital period and the major diameter of a single planetary orbit. (Newton's theory predicts and observation confirms that the calculated mass is the same for each orbit.) Similarly, we can infer the mass of a planet from the orbital period and orbital diameter of one of its satellites. (For example, the Moon's period and orbital diameter give us the Earth's mass.) The same relation has allowed astronomers to infer the masses of stars in binaries (mutually gravitating pairs of stars).

Applied to more complex self-gravitating systems, Newton's theory yields a relation between the average squared velocity of the "particles" (molecules, stars, galaxies) that make up the system and the system's mass and diameter: The mean squared velocity is approximately equal to Newton's gravitational constant multiplied by the mass of the system and divided by its diameter. This relation is the basis of astronomical estimates of the masses of stellar systems. The diameter of a distant object can be inferred from its angular diameter which is directly measurable, and an estimate of its distance. The internal velocity dispersion can be inferred from the object's spectrum. The spectrum of a self-gravitating system is a composite of the spectra of its components. Each component contributes a line spectrum shifted to the red or violet according to that component's line-of-sight velocity. The composite spectrum has broadened lines, whose widths are simply related to the velocity dispersion of the components. Lines in the spectrum of a single star are broadened by the thermal motions of the atoms and molecules that give rise to the lines. Lines in the spectrum of a galaxy are further broadened by the line-of-sight motions of individual stars relative to the galaxy's center of mass. The 21-cm line of neutral hydrogen in a spiral galaxy is broadened by the rotational motion of hydrogen clouds in the galaxy's central plane. Hence its width—the criterion for intrinsic brightness proposed by Tully and Fisher (see above)—is an index of the galaxy's mass.

2.3 Tidal Effects

A major triumph of Newton's theory of gravitation was its explanation of ocean tides. The lunar tide-raising force at a given point on the Earth's surface is the difference between the gravitational pull of the Moon at that point and the gravitational pull at the Earth's center of mass. The Sun also exerts a tide-raising force, defined in exactly the same way. But because the tide-raising force turns out to be inversely proportional to the cube of the distance, the lunar contribution to the total tide-raising force dominates the solar contribution. The lunar tide-raising force is directed toward the Moon on the side of the Earth facing the Moon and away from the Moon on the opposite side. This accounts for the observation that high tides are separated by half a lunar day. The tides are highest when the lunar and solar contributions reinforce one another, lowest when they tend to cancel. Thus the highest tides occur when the Moon lies between the Sun and the Earth (at new Moon), while the lowest tides occur when the Earth lies between the Sun and the Moon (full Moon).

Tidal forces, defined more generally as the spatially varying part of an external gravitational force, are ubiquitous in astronomy. Their effect is always disruptive. The small objects that make up the rings of Saturn, Jupiter, and Uranus (see below) are prevented from coalescing into larger self-gravitating bodies by the tide-raising forces exerted by the respective planets. In 1850 Eduard Roche showed that a self-gravitating liquid satellite whose density is equal to the planet's mean density can resist tidal disruption only if its distance from the center of the planet exceeds

2.44 planetary radii. Satellites and ring particles that lie inside the "Roche limit," as this critical distance is called, are held together primarily by chemical cohesion, not gravity. Star clusters in the plane of the Milky Way are continually disrupted by the tide-raising force exerted by mass interior to their galactic orbits. Tide-raising forces limit the sizes of globular star clusters. Tide-raising forces also limit the sizes of galaxies in clusters and tend to disrupt subclusters within large galaxy clusters.

2.4 Self-Gravitating Clusters

The distribution of mass in the astronomical universe is strongly clustered on a wide range of levels. Self-gravitating systems tend to belong to larger self-gravitating systems and to have smaller self-gravitating systems within them. Because tidal forces tend to disrupt hierarchic structure, we may conclude that the clustering tendency must have been more conspicuous in the past than it is now and that the clustering tendency is a central feature of the process(es) by which self-gravitating systems were formed. (See GALAXIES AND COSMOLOGY.)

Stable astronomical systems are conventionally assigned to a number of broad categories: planets and their satellites, planetary systems, stars, small multiple stellar systems, star clusters (subdivided into open clusters and globular clusters), galaxies, groups and clusters of galaxies, and galaxy superclusters.

The extent to which the kind of structure exhibited by our own planetary system is reproduced around other stars is not yet known; and theories for the origin of stars and planetary systems are not yet secure enough to fill this gap in our knowledge. Many astronomers believe, however, that planetary systems are a common occurrence. There are continuing efforts to find evidence of planets around nearby stars. A massive planet with a tight orbit would perturb the orbited star's motion to an extent that is probably detectable with current techniques.

Most stars belong to double or triple systems. Simulations based on Newton's theory predict and observation confirms that triple systems tend to evolve into a close binary whose center of mass forms a binary system with the more distant third star.

Star clusters occupy the next hierarchical level. The largest and most compact clusters are nearly spherical systems with 100 000 to 10 million members. As mentioned above, the globular clusters in our own galaxy form a centrally concentrated spherical system. Open or galactic clusters are confined to our galaxy's central plane. They are much smaller, with at most a few hundred members.

Galaxies define a major level, in the sense that a large fraction of the mass in the universe is bound in galaxies. They range in mass from less than 10^{10} solar masses to more than 10^{12} solar masses. There also exist dwarf galaxies with masses of only 10^6 to 10^7 solar masses.

Galaxies, like stars, exhibit a strong clustering tendency. They are found in small groups, clusters, and superclusters.

2.5 Mass, Radius, and Cohesive Energy

Although a star and a stellar system, such as a globular star cluster, are both held together by gravity, the conditions that determine their respective sizes or mass densities are profoundly different. Most globular star clusters have masses in the range 10^5 to 10^7 solar masses and radii from 10 to 100 light years. But there is no known physical (as opposed to historical) reason why the range of cluster radii should not be 1000 or more times greater than it actually is. The relatively narrow range of cluster radii must be a consequence of initial conditions and evolutionary history.

By contrast, the radius of a star is largely independent of its initial conditions and evolutionary history. A star is in a condition of equilibrium in which its central temperature and density are just high enough for nuclear reactions in the deep interior to release energy at a rate that balances the rate at which energy is lost by radiation from the surface layers. When its supply of nuclear fuel is exhausted, a star (or what is left of it after some fraction of its mass has been more or less violently expelled) becomes a white dwarf, a neutron star, or a freely collapsing black hole. Knowing the initial mass, chemical composition, and age of a star, we can calculate its radius and predict its internal structure. Initial conditions are largely irrelevant, except insofar as they specify the mass and chemical composition.

From a theoretical standpoint, *cohesive energy* is a more significant characteristic than radius, because energy, like mass, is conserved in physical interactions. Cohesive energy is the energy per unit mass needed to disperse a system. For a self-gravitating system, it is proportional to the mass of the system divided by its radius. Mass and cohesive energy are the two most basic characteristics of a self-gravitating system. The present masses and cohesive energies of relatively tightly bound star clusters, galaxies, and galaxy clusters are probably close to their original values. It is reasonable to suppose that the diffuse clouds from which stars formed were comparable in size to the solar system. Under this assumption, the initial binding energies and masses of self-gravitating systems, from stars to galaxy clusters, obey a simple relation: the cohesive energy is proportional to the one-third power of the mass. Thus stars, globular clusters, galaxies, and rich galaxy clusters, with masses in the ratios $1:10^6:10^{12}:10^{15}$, have cohesive energies in the ratios $1:10^2:10^4:10^5$.

3. ASTRONOMICAL SYSTEMS

3.1 The Solar System

Since the early 1960s, observations with instruments mounted on spacecraft have greatly enriched our store of factual knowledge about the solar system. Thanks to these observations, we now have a wealth of information about geometric, physical, and chemical properties of planets and planetary atmospheres, satellites, asteroids, comets, and other small bodies. During the same period, the development of fast and powerful computers with large memories made it possible for theorists to study fundamental but previously unanswerable questions concerning the origin and evolution of the solar system. In particular, it became possible to investigate the long-term stability of orbits with a variety of initial conditions and to model the evolution of systems composed of large numbers of mutually gravitating components. Although these experimental and theoretical advances have not yet given us a secure theory for the origin of the solar system, they have yielded a greatly improved understanding of the conditions that such a theory must satisfy and of the difficulties it must overcome.

The objects that make up the solar system (apart from the Sun) may be classified according to the character of their orbits, according to their sizes, and according to their physical and chemical characteristics. That these three methods of classification do not yield the same groupings was known before the 1960s. For example, early telescopic observations showed that some satellites are bigger than some planets; and the fact that at least one satellite, Titan, a satellite of Saturn, has an atmosphere was discovered in 1944, by Gerard Kuiper. But information collected by instruments mounted on spacecraft has clarified the relations among the three sets of properties.

Orbits can be conveniently classified as *stable* or *unstable*. A stable orbit keeps its qualitative character for a period comparable to the age of the solar system, about 4.6 billion (=thousand million) years. Unstable orbits change significantly during periods shorter than 4.6 billion years. The point of this distinction is that unstable orbits do not reflect conditions in the early solar system, while stable orbits do. Stable orbits are dominated by either the Sun or one of the planets. As discussed below, instability arises from gravitational disturbances by a third body (or several "third bodies"). For example, gravitational disturbances by Jupiter create zones of unstable orbits in the asteroid belt.

All the planets except the tiny double planet (or planet-satellite system), Pluto-Charon, have stable orbits, as do most of the minor planets or asteroids whose orbits lie between Mars and Jupiter. These stable orbits are nearly circular, only slightly inclined to the central plane of the solar system, and prograde (counterclockwise as viewed from "above"). Pluto's orbital eccentricity of 0.25 and orbital inclination of 17° are considerably larger than those of the other planets. Its orbit may well have been very different in the remote past from what it is now.

Among satellite orbits, those that are circular, slightly inclined, and prograde (relative to the planet's spin axis and direction of spin) predominate. Of Jupiter's 16 known satellites, the eight nearest the planet move in prograde, essentially circular orbits in Jupiter's equatorial plane; the next four have prograde, nearly circular, orbits with moderate inclinations (27°–29°), while the outer four have retrograde orbits. The outermost of Saturn's 17 known satellites, Phoebe, has a retrograde orbit, as does Triton, by far the most massive of Neptune's 8 known satellites. By contrast, the orbits of all of Uranus's 15 known satellites (10 of which were discovered by Voyager 2 in 1985–1986) are prograde and circular, and lie in the planet's equatorial plane. The satellites with retrograde orbits are widely thought to have been captured relatively late in the solar system's evolution.

Although the great majority of asteroids have nearly circular, slightly inclined, prograde orbits whose radii lie between 2.2 and 3.3 astronomical units, some asteroids have eccentric, highly inclined orbits. (The astronomical unit is the radius of Earth's orbit.) Some asteroid orbits cross the Earth's orbit.

Meteoroids are small asteroids that pass close to Earth. Those that enter the Earth's atmosphere and become incandescent through friction with the atmosphere are called meteors; surviving fragments that reach the surface of the Earth are called meteorites.

The orbits of *comets* present a striking contrast to the orbits of the planets and satellites (apart from the exceptions noted above). These sometimes spectacular visitors to the night sky were first identified as solar orbiters by Tycho Brahe in 1577. Newton's theory of gravitation allowed his friend Edmund Halley to calculate the orbits of several comets and to predict the reappearance, on Christmas night of 1758, of a bright comet whose previous appearances had been recorded at 76-year intervals since 1531. Cometary orbits have large eccentricities (the orbits are nearly parabolic) and are randomly inclined to the central plane of the solar system. (Prograde and retrograde orbits are therefore equally numerous.)

Oort (1950) noted that the aphelion distances of cometary orbits tend to be close to 50 000 astronomical units. (An object's aphelion distance is its greatest distance from the Sun.) This is about one-quarter the average distance between stars in the Sun's neighborhood. It is also the distance at which an object's orbit is most sensitive to disturbance by a passing star. Comets with smaller aphelion distances are so tightly bound that they are not greatly disturbed by the tidal field of the Galaxy or by passing stars; comets with greater aphelion distances are disturbed so much that they are likely to be torn loose from the solar system. Oort surmised that comets with aphelion distances near 50 000 AU are most easily perturbed by passing stars into orbits that bring them close enough to the Sun to make them visible.

A comet becomes visible when it grows a tail or *coma*. This is produced by gases that boil off under heating by solar radiation. Spectroscopic analyses of these gases have shown that the volatile constituent of comets is largely (about 79%) water ice, with smaller quantities of many other ices, including ammonia, methane, carbon monoxide, and carbon dioxide. The nonvolatile constituents include hydrocarbons and silicates.

Where do comets come from? Their chemical composition suggests that they were formed in the outer solar system, beyond Jupiter's orbit. (The inner planets have a preponderance of nonvolatile elements, in the form of metal and rock; in the outer planets ices predominate.) Recent theoretical studies and numerical simulations strongly suggest that they originated as planetesimals between the orbits of Uranus and Neptune, and were then perturbed into their present highly elongated and highly inclined orbits by interactions with the outer planets (Duncan *et al.*, 1987, 1988). Numerical simulations have shown that once Jupiter and Saturn had formed, planetesimals on intervening orbits would have been quickly ejected from the solar system (Franklin *et al.*, 1989).

Some of the most interesting orbits in the solar system are those of the particles that make up *planetary rings*. Observations from spacecraft have shown that not only Saturn but also Jupiter, Uranus, and Neptune have ring systems. Saturn's ring system (Fig. 5) is by far the widest (20 000 km) and the most massive (10^{17}–10^{18} kg). Saturn's rings are made up of icy particles whose diameters range from a centimeter or so to a few meters. Uranus's rings are made of similar-sized but much darker particles. Uranus's ring system (Fig. 6) is only 100 km wide, and its mass is 1% the mass of Saturn's ring system. Jupiter's ring system is faint, transparent, and diffuse. It consists of micron-sized dust particles. Neptune has partial rings or arcs whose properties are still largely unknown.

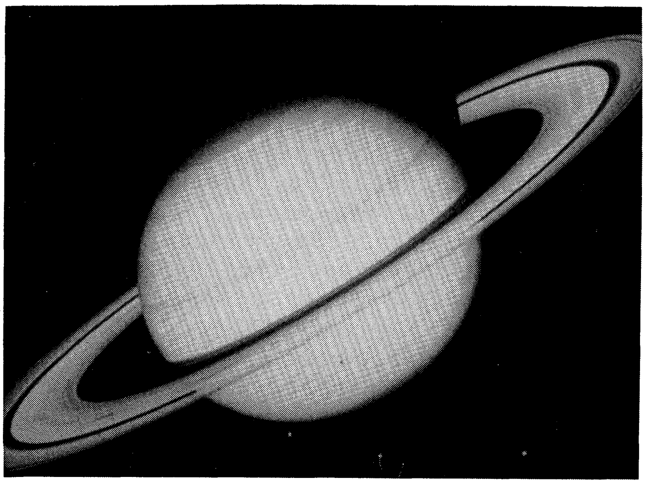

FIG. 5. Saturn and its rings as photographed by Voyager. (NASA/JPL photo.)

All or nearly all the material in each ring system lies within the planet's Roche limit. The rings may be composed wholly or in part of material that entered the zone of tidal disruption relatively recently or of primordial material that was prevented by the tidal field from coagulating into larger bodies.

The rings of Saturn and Uranus have an extremely intricate (and beautiful) fine structure (Fig. 7). Bright rings of varying widths alternate with narrow dark lanes. The dark lanes are caused in part by resonant interactions with the planet's satellites. A resonant interaction occurs between objects whose orbital periods are in the ratio of (usually small) whole numbers, such as 1:2 or 2:3. A ring particle whose orbital period is commensurable in this sense with that of a nearby satellite experiences a periodic disturbing force that causes its orbit to depart progressively more from its initially circular form. For example, there is a gap in Saturn's ring system where an orbiting particle would have exactly half the orbital period of the nearby moon Mimas.

Resonant interactions with Jupiter give rise to the so-called Kirkwood gaps in the asteroid belt. The gaps occur at distances from the Sun at which the ratio between an asteroid's period and Jupiter's period are small whole numbers.

Size and mass. Classification by these criteria cuts across classification by orbit. The planet Pluto is smaller than six satellites: Jupiter's four Galilean satellites, Titan (Saturn's giant satellite), and Triton (Neptune's giant satellite). Mercury is about the same size as, though denser and hence more massive than, the three largest satellites. Mars's two satellites and the innermost satellites of the outer planets, with diameters ranging from ten to a few tens of kilometers, are comparable in size with medium-sized asteroids. Chiron, an ob-

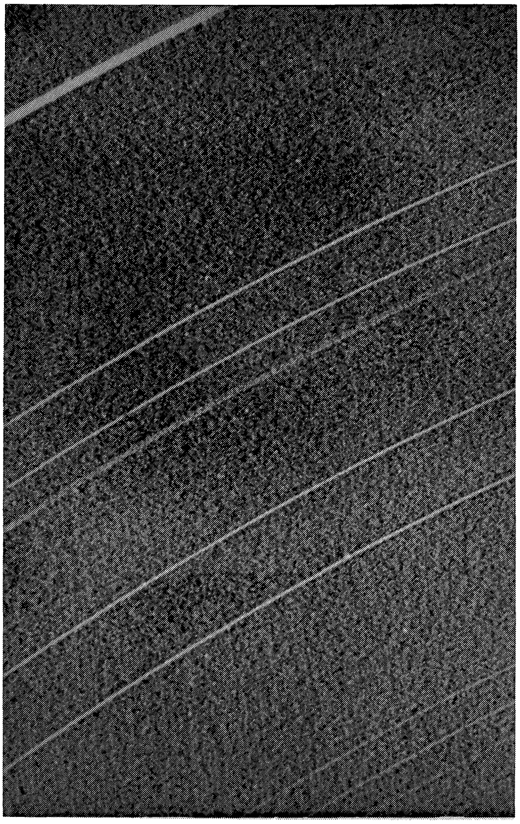

FIG. 6. Voyager image of the narrow, dark rings of Uranus. (NASA/JPL photo.)

ject with a diameter of 150 km (the size of a large asteroid) moves in an elliptic orbit between Saturn and Uranus, in one of the chaotic zones mentioned above. In 1988 it exhibited a faint but unmistakable coma, thereby qualifying for cometary status and the title of the largest known comet.

The number of small bodies in the solar system increases steeply with decreasing mass. Quantitative studies indicate that the number is proportional to the $-\frac{2}{3}$ power of the mass. Such a "spectrum" is consistent with the hypothesis that the small bodies formed through collisional fragmentation.

The ubiquity of cratering indirectly supports this hypothesis. The Moon, Mercury, Venus, Mars and its two satellites, and the large satellites of Jupiter, Saturn, and Uranus have all been found to have heavily cratered surfaces, produced by impacts with comets and asteroids. The data suggest that throughout the solar system the number density of small solid bodies was much higher in the past than it is now. An initial period of intense cratering activity lasted only 500 million years. The present distribution of small bodies subsequently evolved through accretion and collisional fragmentation.

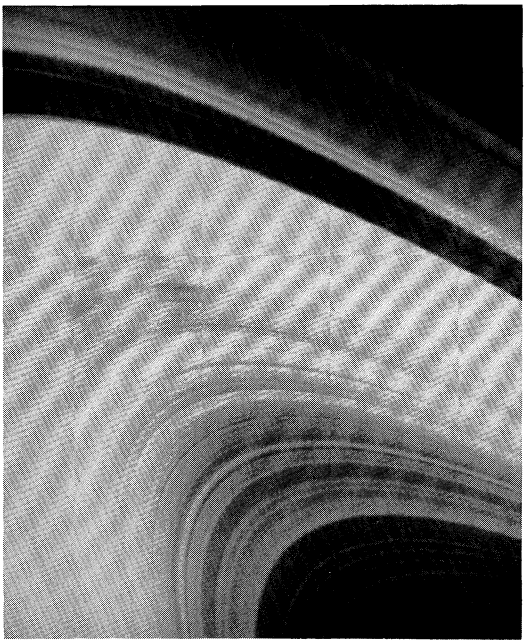

FIG. 7. Saturn's B ring in detail. The structure seen here has scales of tens to hundreds of km. Most of this structure has not been explained. (NASA/JPL photo.)

Meteorites. Most meteorites, unlike rocks on Earth, are chemically undifferentiated; that is, the various elements have not separated out according to their chemical properties, such as melting temperature. Moreover, these meteorites have the same chemical composition as the solar atmosphere, apart from a few volatile elements like the noble gases. Radiometric determinations of the ages of undifferentiated meteoritic material shows that it is about 4.5 thousand million years old. This is taken to be the age of the solar system. From studies of the structure and chemical composition of meteorites and experimental and theoretical information about the effects of heating and cooling in objects of different sizes, scientists have inferred that meteorites are fragments of larger bodies ranging in diameter from about a kilometer to a few hundred kilometers. Analysis shows that there must have been at least twenty or so different parent bodies, all the size of large asteroids, rather than a single large parent body of planetary size.

Extrapolations of observed meteor orbits indicate that the great majority of meteorites originate in the asteroid belt. Recently, however, some meteorites have been shown to have been formed on the surface of the Moon by relatively recent cratering events that also gave these objects enough kinetic energy to overcome the Moon's gravitational pull and reach the Earth. There is even evidence that some meteorites were produced by cratering events on Mars!

Meteorites encode an enormous quantity of information about the history of the inner solar system. So far, only a small part of the information has been decoded, but already it has served to constrain the range of hypotheses about the origin and evolution of the solar system.

3.2 Origin of the Solar System

Since the early 1960s, powerful computational techniques, sophisticated physical and chemical considerations, and data supplied by spacecraft and ground-based astronomical observations have all been brought to bear on this ancient problem. The most common theoretical approach, which goes back to Kant and Laplace, posits that the Sun contracted from an extended cloud of dust and gas, leaving behind a thin disk. Supporters of this picture cite recent high-resolution observations of young stars in molecular clouds (see below); but the interpretation of these observations is still controversial. However, there is firm evidence that dust particles with diameters of less than a micron are formed in gases ejected by supernovae and other massive stars.

The first step in the genesis of a planet is the aggregation of submicron-sized dust particles into self-gravitating "planetesimals" a few kilometers in diameter. A sufficiently thin layer of dust would indeed break up into just such planetesimals. Unfortunately, it has recently been found that interaction between the dust and the ambient gas effectively inhibits this process (Weidenschilling, 1980). The alternative process for forming planetesimals—simple coagulation (chemical binding)—also encounters severe difficulties. At the moment (1991), no way around this impasse is in sight.

Assuming that planetesimals have somehow been formed, theorists have used computers to simulate their subsequent evolution. With appropriate initial conditions, such simulations can be made to yield a realistic inner solar system (Mercury, Venus, Earth, Mars). But the models encounter as yet unsolved difficulties in predicting the outer planets and the asteroid belt. There is also an important difficulty with the inner planets: their predicted spin axes are randomly oriented, instead of being accurately perpendicular to the central plane.

3.3 Stars and Star Clusters

Stars are self-gravitating gaseous bodies whose centers are hot and dense enough to ignite thermonuclear reactions. The second condition requires the minimum mass of a star to be approximately one-tenth the mass of the Sun. The first condition implies that the maximum mass of an isolated star is around 70 times the mass of the Sun: in a more massive object, radiation pressure exceeds the pull of gravity.

The brightness and effective temperature of a homogeneous, nonrotating, nonmagnetic star are determined by its mass and chemical composition. Hence homogeneous stars with the same chemical composition should form a one-dimensional sequence. Now, it is reasonable to suppose that stars in the same cluster were formed at the same time with the same initial chemical composition. They are also at nearly the same distance (if the cluster subtends a small angle in the sky). So if we plot their apparent brightnesses against their effective temperatures, the plotted points should fall along a curve (rather than scattering over an area). Experience confirms this prediction, showing not only that the assumptions of equal age and initial composition are valid but also that rotation and magnetic fields, both of which are normally present and which vary from star to star, have relatively little effect on brightness and surface temperature.

The shape of the temperature-brightness (or Hertzsprung-Russell) diagram for a star cluster depends on both its age and initial chemical composition. With the help of calculations of stellar evolution (see ASTROPHYSICS) one can infer the age and initial chemical composition of a cluster from the shape of its temperature-brightness diagram. The ages of clusters in our own and other galaxies span a wide range. Globular clusters (see Fig. 8), with estimated ages between 12 000 and 15 000 million years, are the oldest. They form a vast spherical or spheroidal system. The youngest clusters (see Fig. 9) are less than a million years old. They lie in the central plane of spiral galaxies.

The age of a cluster is correlated with the initial abundances of elements heavier than helium. The oldest clusters have very low heavy-element abundances—as low as 0.01% or less. The youngest clusters have heavy-element abundances of 3% or more. The initial or primordial helium abundance seems to be much the same in all stars. It is currently estimated at between 22% and 24% by mass. (The difference between the upper and lower bounds is of crucial cosmological significance; see GALAXIES AND COSMOLOGY.)

Helium and heavier elements are formed in the interiors of massive short-lived stars. These stars go through an explosive evolutionary stage, ejecting their "processed" gases into the interstellar medium. New stars form from gas that has been more or less "enriched" in this way. Because gas and dust settles quickly into the plane of a spiral galaxy, it is understandable that stars in young clusters, which are also strongly concentrated toward this plane, have relatively high abundances of heavy elements.

3.4 The Interstellar Medium

The region between the stars is populated by clouds of dust and gas, in varying proportions. As discussed above, dust manifests itself in the dimming and reddening of light from distant stars and galaxies. Gas clouds in the line of sight to a more distant star impose their own absorption lines on its spectrum.

Hot blue stars are often embedded in interstellar clouds. They emit ultraviolet light copiously enough to ionize a large surrounding volume of hydrogen. In this ionized region, recombining hydrogen atoms emit light not only at the original ionizing frequency but also at a variety of other, lower frequencies, in the visible and infrared parts of the spectrum. As a result, such regions of ionized hydrogen are bright in visible and infrared light.

Spiral galaxies contain large quantities of atomic hydrogen in clouds strongly concentrated to their central plane. These clouds emit radiation at a wavelength of 21 cm. (The atomic transition responsible for this radiation connects the two hyperfine sublevels of the lowest level in hydrogen's energy spectrum. The two sublevels correspond to different relative orientations of the electron's and proton's magnetic moments. See ATOMIC SPECTROSCOPY.) Clouds of neutral hydrogen in our own and other galaxies are concentrated in spiral arms. As discussed above, estimates of the masses of spiral galaxies are based on measurements of the Doppler shifts of the 21-cm line emitted by hydrogen in their central planes.

3.5 Giant Molecular Clouds

Spiral galaxies contain massive (up to 10^7 solar mass), dense clouds in which hydrogen is bound predominantly in molecules such as OH, CO, and H_2. (Larger molecules, notably formaldehyde, have also been detected.)

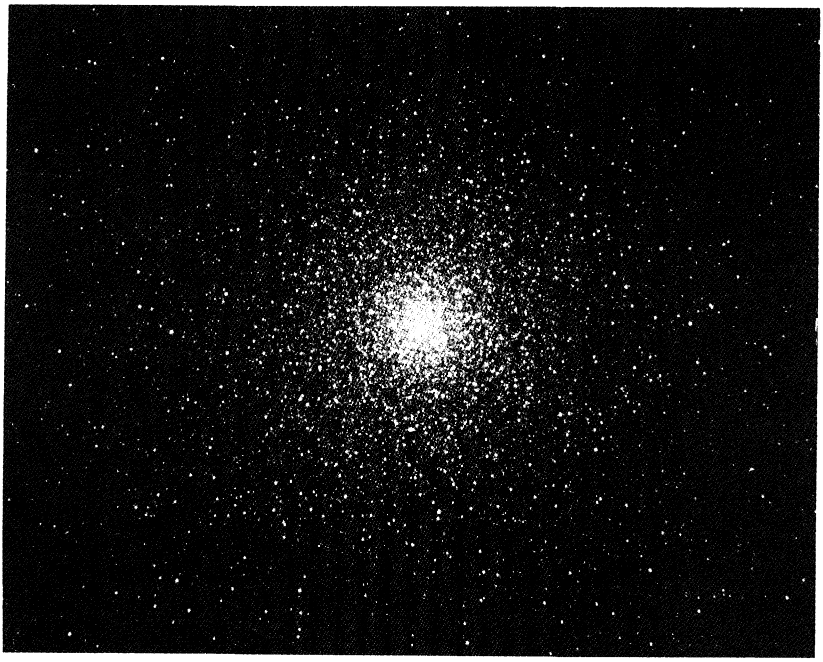

FIG. 8. The globular cluster M3 in the constellation Canes Venatici, the Hunting Dogs. [Pasachoff and Kutner (1978), Fig. 7-13.]

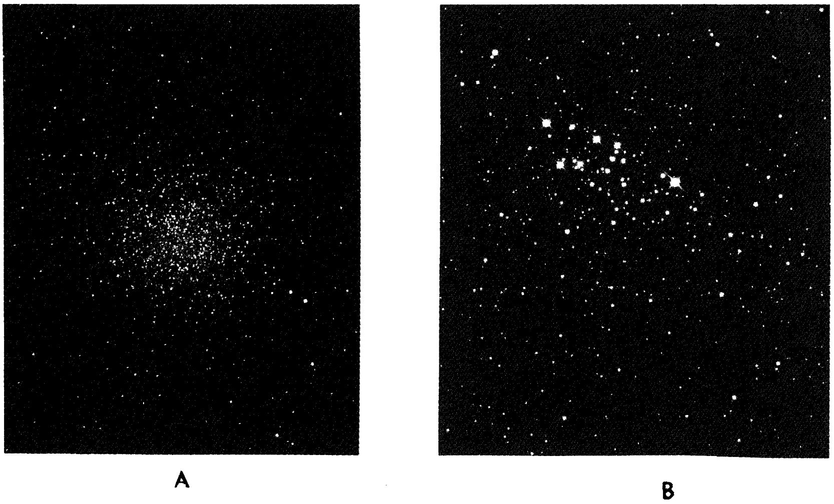

FIG. 9. (A) A rich galactic cluster, NGC 5897, in the constellation Libra. (B) A poor galactic cluster, NGC 7510. [Pasachoff and Kutner (1978).]

These molecules have emission lines in the radio part of the spectrum (see MOLECULAR SPECTROSCOPY). Molecular clouds or cloud complexes are particularly interesting because they contain newly formed stars and, presumably, stars in the process of formation. Among such stars are the so-called T Tauri stars, which have been intensively studied. They are always deeply embedded in dust and gas, emit strong winds, vary strongly and erratically in brightness, and give evidence of intense surface activity. The Sun is widely believe to have gone through a T Tauri stage in its evolution.

List of Works Cited

Abell, G., Morrison, D., Wolff, S. (1991), *Exploration of the Universe,* Philadelphia: Saunders, pp. 513–515.

Duncan, M., Quinn, T., Tremaine, S. (1987), *Astron. J.* **94**, 1330–1338.

Duncan, M., Quinn, T., Tremaine, S. (1988), *Astrophys. J. Lett.* **328**, L69–L73.

Franklin, F., Lecar, M., Soper, P. (1989), *Icarus* **79**, 223–227.

Hertzsprung, E. (1914), *Astron. Nachr.* **196**, 203–210.

Hubble, E. P. (1936), *The Realm of the Nebulae,* New Haven: Yale University.

Layzer, D. (1984), *Constructing the Universe,* New York: Scientific American Books.

Leavitt, H. S. (1912), Harvard College Observatory Circular, No. 173.

Oort, J. (1950), *Bull. Astron. Soc. Netherlands* **11**, 91–110.

Pasachoff, J. M., Kutner, M. L. (1978), *University Astronomy,* Philadelphia: Saunders.

Stebbins, J. (1933), *Proc. Natl. Acad. Sci. U.S.A.* **19**, 222–228.

Trumpler, R. J. (1925), *Proc. Astron. Soc. Pacific* **37**, 307–320.

Tully, R. B., Fisher, J. R. (1977), *Comments Astrophys.* **7**, 85–93.

Weidenschilling, S. J. (1980), *Icarus* **44**, 172–189.

Further Reading

There are many excellent modern texts, among them:

Abell, G. O., Morrison, D., Wolff, S. C. (1991), *Exploration of the Universe,* 6th ed. Philadelphia: Saunders.

Shu, F. (1982), *The Physical Universe,* Mill Valley, CA: University Science Books.

For review articles intended for general astronomical readers:

G. Burbidge, D. Layzer, A. Sandage (Eds.) (1963–present), *Annual Reviews of Astronomy and Astrophysics,* Palo Alto, CA: Annual Reviews.

ASTROPHYSICS

VIRGINIA TRIMBLE, *Physics Department, University of California, Irvine, California, U.S.A.*

	Introduction	117
1.	**Stellar Structure and Evolution**	118
1.1	Basic Equations and Their Solution	118
1.2	Main Sequence (Core Hydrogen-Burning) Stars	120
1.3	Evolution Off the Main Sequence	121
1.4	The End Points of Stellar Evolution	122
1.4.1	Planetary Nebulae and White Dwarfs	122
1.4.2	Supernovae, Neutron Stars, and Pulsars	123
1.4.3	Black Holes	124
1.5	Binary Stars	125
1.5.1	Main-Sequence and Post-Main-Sequence Binaries	125
1.5.2	Cataclysmic Variables, X-Ray Binaries, and Binary Pulsars	126
2.	**Interstellar Matter and the Formation of Stars**	128
2.1	Structure and Kinematics of the Interstellar Medium (ISM)	128
2.1.1	Phases of the Interstellar Medium	128
2.1.2	Large-Scale Distribution and Motions of the ISM	131
2.1.3	Energetics and Stability	132
2.2	Star Formation	132
2.2.1	Contraction and Condensation of and in Molecular Clouds	133
2.2.2	Protostellar Cores	135
2.2.3	Pre-Main-Sequence Stars and Planetary Formation	135
2.3	Star Clusters and Stellar Populations	135
2.3.1	Open Clusters	137
2.3.2	Globular Clusters	138
2.3.3	Stellar Populations	139
	Glossary	140
	Works Cited	142
	Further Reading	143

INTRODUCTION

Astrophysics is the branch of physics that attempts to understand the appearance and behavior of astronomical objects, especially those outside the solar system, including stars and galaxies, assemblages of these, and the material around them. Alternatively, one can describe astrophysics as the portions of astronomy that the physicists of any given generation regard as worthy of their attention.

The word first appears in the late nineteenth century to mean the application of principles of spectroscopy, atomic physics, and thermodynamics to measure the luminosities, temperatures, and chemical compositions of stars and nebulae, while astronomy meant positional measurements and their applications to the solar system, binary stars, and definitions of coordinate systems based on the earth's motions. All of these topics would today be regarded as parts of astronomy and rather old-fashioned ones at that.

Meanwhile, astrophysics has evolved to denote investigations of the internal structure, evolution, and formation of stars and galaxies. Nuclear astrophysics focuses on nuclear reactions as energy sources in stars and as synthesizers of the chemical elements. High energy astrophysics takes as its territory objects and processes where either the energy per particle or photon or the total energy is large (compared to thermal equilibrium or to average objects). Examples include supernovae, pulsars, and x-ray sources; cosmic rays; and quasars and radio galaxies.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Study of the structure and evolution of the universe as a whole is generally designated by the separate name cosmology, but is also addressed here. The properties of galaxies and these large-scale topics are covered in the article on GALAXIES AND COSMOLOGY.

1. STELLAR STRUCTURE AND EVOLUTION

The observed properties of stars include their masses, luminosities, radii, surface temperatures, surface chemical compositions, and (sometimes) an indication of age. A model star is a table giving values of mass, radius, luminosity, temperature, entropy, density, pressure, and composition from center to surface. It is arrived at by solving a set of four coupled nonlinear differential equations, constrained by three auxiliary tables or analytic relations for equation of state, opacity to radiation, and rate of generation of nuclear energy, and subject to suitable boundary conditions (mass=0 at the center; density=0 at the surface; etc.).

A sequence of model stars, reflecting changes with time in the fundamental quantities throughout the star, is an evolutionary track. The initial model is normally chemically homogeneous, in hydrostatic equilibrium (gravity balanced by pressure gradients), and representative of a star that has just begun fusion of hydrogen to helium at its core.

Clearly the primary goal of stellar structure and evolution studies is to obtain a one-to-one correspondence between real stars and model ones. This has, to a considerable extent, been achieved, though with many detailed effects due to rotation, magnetic fields, loss of material from the stellar surface, and presence of a close stellar companion still to be sorted out. Thus, we can be said to understand stars.

The most important determinant of a star's history is its initial mass, which largely sets its temperature, brightness, and lifetime, the sequence of nuclear reactions that will occur, and its end point as a white dwarf, neutron star, or black hole. Initial chemical composition, especially the percentage of helium and total heavier elements, has significant but rather subtle effects.

The classic treatments of stellar structure and evolution are the monographs by Chandrasekhar (1939) and Schwarzschild (1958). Recent texts include Collins (1989) and Rolfs and Rodney (1988). Late phases of evolution are addressed in Philip *et al.* (1987), Kwok and Pottasch (1987), Abbott and Conti (1987), and Torres-Peimbert (1989). Supernovae and their products have been reviewed by Woosley and Weaver (1986), Trimble (1988), Weiler and Sramek (1988), and Arnett *et al.* (1989). Special effects considered include pulsation and mass loss (Stalio and Willson, 1988), rotation and magnetic fields (Hartmann and Noyes, 1987), and binary companions (Pringle and Wade, 1985; Drechsel *et al.*, 1987; Sahade *et al.*, 1990).

1.1 Basic Equations and Their Solution

Let P, T, and ρ represent pressure, temperature, and density at a distance r from the center of a spherically symmetric star, and $M(r)$ and $L(r)$ represent the mass and luminosity interior to that point. Then the increase of mass with radius can be described as

$$\frac{dM(r)}{dr}=4\pi r^2\rho. \tag{1}$$

The condition that the star be neither expanding nor contracting rapidly (hydrostatic equilibrium) is that the inward force of gravity on a fluid element at r be balanced by the pressure gradient across it, that is,

$$\frac{dP}{dr}=-\rho\frac{GM(r)}{r^2}. \tag{2}$$

Nuclear energy generation is represented by the quantity ϵ in erg g^{-1} s^{-1} as a function of T, ρ, and chemical composition. Thus for a static star, the increase of luminosity outward is given by

$$\frac{dL(r)}{dr}=4\pi r^2\rho\epsilon. \tag{3a}$$

If some portion of the star is expanding or contracting rapidly compared to the time scale of the nuclear processes, then

$$\frac{dL(r)}{dr}=4\pi r^2\rho\left(\epsilon-\frac{dU}{dr}-\frac{P}{\rho^2}\frac{d\rho}{dt}\right), \tag{3b}$$

where the additional terms are the rate of change of internal energy per gram U, and the PdV work being done, where specific volume V has been replaced by its reciprocal ρ. At this point it becomes clear that integrating the equations together will require an equation of state, that is, analytic expressions or tables for $P(\rho,T)$ and $U(\rho,T)$ for each chemical composition of interest.

Finally, the gradient dT/dr follows for each possible mode of energy transport. Conduction is unimportant in stars (except for degenerate white dwarfs and neutron stars). Energy transport by radiation is opposed by the opacity of the gas, expressed by κ, defined so that $\kappa\rho dx$ is the fraction of the energy of a beam of light lost in traversing a distance dx. Then

$$\frac{dT}{dr}=\frac{3\kappa\rho L(r)}{16\pi r^2 acT^3}, \tag{4a}$$

where a is the radiation constant and c the speed of light.

If this temperature gradient is so steep that a test volume of fluid displaced from equilibrium position keeps going, then convection sets in and establishes a temperature gradient very close to the adiabatic one,

$$\frac{dT}{dr}=\left(1-\frac{1}{\gamma}\right)\left(\frac{T}{P}\right)\frac{dP}{dr}, \tag{4b}$$

where γ is the ratio of specific heats. This equation and the criterion for whether Eq. (4a) or Eq. (4b) applies become more complex in regions where there is a gradient of chemical composition.

A large fraction of the work in studies of stellar structure and evolution goes into getting the right numerical values for κ, ϵ, and the equation of state. Near the center of the sun, ϵ is 10–100 erg g^{-1} s^{-1}, κ is 0.01–0.1 cm^2/g, and T, ρ, and P are 1.5×10^7 K, 150 g/cm^3, and 6×10^{15} dyn cm^{-2}.

The main difficulties in solving these equations come from their nonlinearity, the large range of orders of magnitude involved, and the fact that not all the boundary conditions are known at either center or surface. In addition, there are numerical instabilities where M, L, and r go to zero together at the center. The usual strategies involve

1. a change of variables from physical ones to ratios like $q=M(r)/M$ and $p=P(GM^2/4\pi R^4)^{-1}$,
2. guessing the unknown boundary conditions, including P and T at the center and L and r at the surface,
3. standard approximations for getting started at the center and guessing suitable surface T and P, and
4. starting integrations from both directions and adjusting the guessed boundary conditions until the integrations meet up adequately at some intermediate point.

An initial stellar model must be obtained by this sort of (Schwarzschild) method. Numerically more satisfactory methods were developed about 1960 by Louis Henyey (see, e.g., Collins, 1989, for details). These methods require an initial approximate model (for instance, a chemically homogeneous Schwarzschild model or the one just obtained at the previous step in an evolutionary calculation). The independent variable is normally M rather than r, and the equations are expressed as finite difference equations across zones $i=1$–N, for instance

$$\frac{P_{i+1}-P_i}{M_{i+1}-M_i}=\frac{GM_{i+1/2}}{4\pi r^4_{i+1/2}}. \tag{2'}$$

The initial model is used to evaluate the quantities on the right-hand sides of all the equations, and the surface boundary conditions come from a separate model atmosphere calculation. There are then $4N-4$ equations in $4N$ unknowns, which can be solved by a Newton–Raphson iteration scheme.

To calculate an evolutionary track, the modeller takes a small forward time step, calculates the compositional changes resulting from nuclear reactions, and uses the previous model as the starting point for a Henyey calculation of the correct structure at the later time.

Most of the necessary cleverness goes into doing this in an acceptably short time. When Schwarzschild and others began calculating stellar evolution tracks in the late 1940s, they used desk calculators. A starting model took a day, and a track to the point where a star had exhausted its hydrogen fuel took a month. By the late 1960s, Henyey codes were running at many university and observatory computer centers, to which you carried your box of cards and returned later in the day for the printout resulting from about an hour of CPU time. Stellar evolution at this level of sophistication in the late 1980s fits on an office machine, and the answer comes out before you have had time to think what it is going to mean. Current research is thus focused on more complex cases, including nonspherical stars, ones that are rotating, pulsating, losing mass, or interacting with companions, and late evolutionary phases where many different nuclear reactions are simultaneously releasing and soaking up energy and changing the chemical composition (none of them having had time to come into equilibrium).

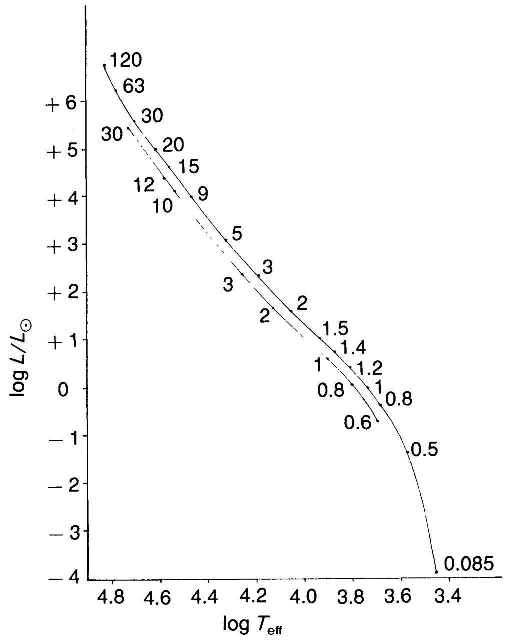

FIG. 1. Plot of log ($L/L_\odot$) ($L_\odot$=luminosity of the sun=3.86×10^{33} ergs/s) vs log T_{eff} (approximately surface temperature) of homogeneous, main sequence stars, so called because stars spend most of their lives in this core-hydrogen-burning phase, and thus most of the stars in a volume-limited sample will be in this phase. The parallel sequence to the left represents stars with 1%–10% of the metal abundance found in normal main sequence stars. Such subdwarfs dominate the old, halo population of our galaxy. Masses (in units of the mass of our sun) are indicated for representative stars along each sequence.

1.2 Main Sequence (Core Hydrogen-Burning) Stars

Models of chemically homogeneous stars of about 73% hydrogen, 25% helium, and 2% heavier elements (by mass and in standard proportions like those found in meteorites and spectroscopically at the solar surface) have the range of mass, luminosity, surface temperature, and radius and the correlations of those properties characteristic of real main sequence stars (Fig. 1, and see ASTRONOMY). The mass range is about $(0.08–100)M_\odot$ ($M_\odot$= the mass of the sun $=2\times10^{33}$ g). The lower limit is set by the need for the central temperature in hydrostatic equilibrium to be high enough for proton kinetic energies to get them over the Coulomb barrier so they can fuse, and the upper one by the stars becoming so luminous and unstable that material is blown off their surfaces by radiation pressure and pulsation.

Numerical models confirm analytic results (Schwarzschild, 1958) that, roughly, $L\propto M^3$ for bright stars and $M^{4.5}$ for fainter ones, $R\propto M^{1/2}$, and $T\propto M^{3/4}$. Since the fuel store is proportional to M, stellar lifetimes scale as M^{-2} or steeper (except close to $100M_\odot$, where $L\propto M$ and all stars have about the same 10^6-yr lifetime, compared to 10^{10} yr for $1M_\odot$).

A cloud of mostly hydrogen gas becomes a star when it has contracted and heated enough for hydrogen fusion to release as much energy as is being radiated from the surface, so that contraction stops. The nuclear reactions involved are those originally specified by Hans Bethe in 1939 (Rolfs and Rodney, 1988). Hydrogen atoms fuse either directly via the weak interaction ($H^1+H^1\rightarrow H^2+e^++\nu_e$, plus subsequent reactions) or via the strong interaction (but with intervening beta decays) with C, N, and O as catalysts (starting with $C^{12}+H^1\rightarrow N^{13}+\gamma\rightarrow C^{13}+e^++\nu_e$). These sets of reactions are known as the *proton-proton* (*pp*) and the *CNO chain cycle*, respectively. Some of the subsidiary reactions produce neutrinos of MeV energy, and the detection of them coming from the sun (albeit at only about 1/3 the expected level) constitutes direct confirmation that stars run on nuclear energy.

In either case, $4H\rightarrow He+30$ MeV (several percent of which is lost in neutrinos). Since a star can exhaust the hydrogen in about the inner 10% of its mass before drastic structure changes occur, the main sequence lifetime of a $1M_\odot$ star is about 10 Gyr. Diffusion and convective processes can lengthen stellar lifetimes by somewhat uncertain amounts (perhaps 10%) by bringing in fresh fuel.

Main sequence stars of less than about 1.5 $M_\odot$ run primarily on direct *pp* fusion. They have radiative cores, but convective envelopes, because hydrogen goes from neutral to ionized just below the surface, producing a zone of high opacity and steep temperature gradient [Eq. (4a)]. Stars above $1.5M_\odot$ have hydrogen ionized throughout, and thus radiative envelopes, but convective cores because CNO-catalyzed fusion produces most of the energy, and its reaction rate is a steep function of temperature, resulting in a steep central dT/dr. There are associated differences between low- and high-mass stars in how they generate magnetic fields, hot layers (chromospheres and coronae) above their surfaces, surface activity (like sunspots and flares), and winds (see SOLAR RADIATION). No star has convection in both envelope and core, and so the surface composition remains unaffected by interior nuclear reactions and reflects what the star was born with.

An initial composition less rich in heavy elements than the solar 2% produces stars that are somewhat bluer for a given luminosity (because you see deeper into them) and differ in the sizes of their convective envelopes and cores. They are called subdwarfs, make up the stellar population of the halo of our galaxy, and are left from early galactic evolutionary phases when very few heavy elements had yet been made in stars. The lifetime of a subdwarf that is just now leaving the main sequence sets the best lower limit we now have to the age of the universe. Unfortunately, uncertainties in the distances (hence luminosities) of these stars, in their abundances of helium and certain heavy elements, and in the effects of diffusion and convection still leave those lifetimes uncertain over a range of at least 11–15 Gyr.

1.3 Evolution Off the Main Sequence

Exhaustion of the central hydrogen fuel supply causes the core of a star to contract, since energy is still flowing out of it. The outer layers of the star expand, giving rise to red giants and supergiants, in which the main energy source is hydrogen fusion (always via CNO catalysis) in a thin layer around the helium core. Subsequent events and their time scales depend largely on the initial mass of the star and produce types of stars called horizontal branch, clump, and asymptotic giant branch stars, from their positions on plots of luminosity versus surface temperature (see Fig. 2, Renzini and Fusi Pecci, 1988, and ASTRONOMY).

Initial masses less than about $0.4M_\odot$ burn only hydrogen and leave behind degenerate cores (white dwarfs, see below) of nearly pure helium. The universe is not yet old enough for this to have happened to single stars, but such helium remnants are known in binary systems, where transfer of material from one star to the other has interrupted the evolution of an initially more massive star.

In stars above $0.4M_\odot$, the temperature of the contracting core, fed by hydrogen shell burning, rises to 10^8 K and helium fusion sets in (as a somewhat explosive helium flash in stars below $1.5M_\odot$ whose cores are already partially degenerate, peacefully in larger masses). The reactions $3\text{He}^4 \rightarrow \text{C}^{12} + \gamma$ and

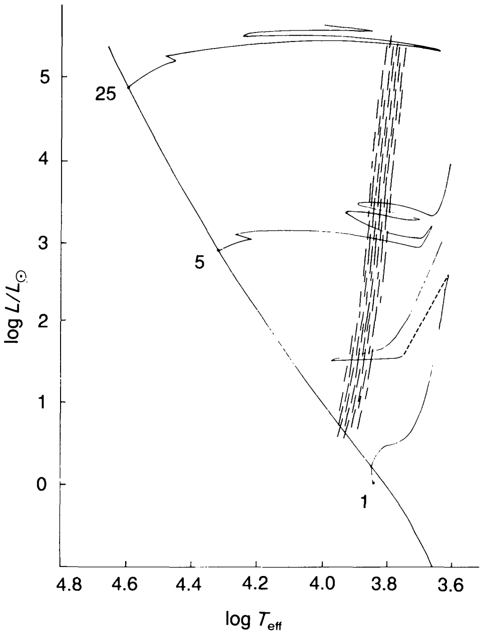

FIG. 2. Post-main-sequence evolution of stars of $25M_\odot$ (from Chiosi and Maeder, 1986), $5M_\odot$ (from Collins, 1989), and $1M_\odot$ (from Bowers and Deeming, 1984; posthorizontal branch from Chiosi *et al.* in Kwok and Pottasch, 1987). Hatched zone is the pulsational instability strip, with Cepheid variables at log $L = 3$–5, RR Lyrae variables near log $L = 1.5$, and δ Scuti stars on the main sequence. Solar composition (2% heavy elements) except for the $1M_\odot$ and its horizontal branch descendent of $0.62M_\odot$ which have 0.1% heavy elements. Evolution from red giant to horizontal branch phase (helium flash) is rapid and has not been modeled in full detail; thus it is shown as a dotted line.

$C^{12}+He^4 \rightarrow O^{16}+\gamma$ occur with comparable probabilities, so that helium burning produces roughly equal amounts of carbon and oxygen (as we observe). Interactions of zones burning hydrogen and helium can release free neutrons which are captured by iron already present in second-generation stars to produce about half of the still heavier elements (called *s*-process nuclides because the neutrons are captured slowly compared to beta-decay time scales).

Following helium core exhaustion, the star is fueled by both hydrogen and helium burning in thin shells, and becomes bright, distended, and unstable to convection in several layers. As a result, first, products of interior nuclear reactions, for the first time, reach the surface of the star, and, second, a large fraction of the initial mass is blown off in a wind. The ejecta are visible as a circumstellar shell, called a planetary nebula if the residual core is hot enough to ionize the material. Nuclear reactions stop at this point, leaving a carbon–oxygen core (CO white dwarf) if the star is stripped down to less than the mass needed to get hot enough to burn carbon (about $1.3M_\odot$ and 2×10^8 K). This happens for all or most stars initially less than at least $(5–6)M_\odot$ and possibly up to $(7–8)M_\odot$.

More massive stars fuse carbon. The dominant reactions are $C^{12}+C^{12} \rightarrow Ne^{20}+He^4$, $Na^{23}+p$, or $Mg^{23}+n$. At slightly higher temperatures, neon burns, then oxygen (with Mg, Si, and S among their products). Mass loss from the stellar surface continues, and a few stars are quenched during these reactions to leave degenerate cores of O, Ne, and Mg. More often, temperatures continue to rise toward 10^9 K, where silicon (made in Ne and O burning) in turn burns via $Si^{28}+\gamma \rightarrow 7He^4$, $Si^{28}+7He^4 \rightarrow Ni^{56}$. The Ni^{56} undergoes beta decay to Fe^{56}, producing a gradually growing iron core at 10^9 K and 10^9 g/cm^3. The core has no further nuclear energy sources available, and, when its mass has grown to the maximum value supportable by degenerate electron pressure, it will collapse suddenly and catastrophically (see Supernovae, below).

In addition to the nuclear reactions already mentioned, rapid capture of neutrons (*r*-process) produces additional elements heavier than iron, probably during supernova explosions. Cosmic rays hitting interstellar atoms contribute lithium, beryllium, and boron, and other rare processes make nuclides with relatively high p/n ratios. The sum of known processes, mostly in stars, accounts for the full range of stable nuclides, in roughly the proportions we see them. Some fine tuning occurs as supernova shock waves pass out through the layers of overlying heavy elements (see Rolfs and Rodney, 1988, for further details).

The last few burning stages in stars above $10M_\odot$ occupy less than 10^3 yr; thus core instability can overtake a star either when it is a red supergiant or on one of its swings back to higher surface temperature and blue color (as was the case for the progenitor of supernova 1987A).

Post-main-sequence evolution takes many stars through a range of L and T where the hydrogen and helium ionization zones just below the surface act like taps or faucets, releasing energy at high temperature and bottling it up at low. The resulting instability causes stars, including the Cepheid and RR Lyrae variables, to pulsate in radius, temperature, and luminosity on a time scale close to the free-fall time, $t=(G\rho)^{-1/2}$, and so to display relationships among their pulsation periods (hours to weeks), colors, and luminosities. Thus we can measure their distances and use them to trace out the structure of our galaxy and the distances to other galaxies. Detailed modeling of pulsation properties also provides an independent check of the correctness of interior models.

1.4 The End Points of Stellar Evolution

1.4.1 Planetary Nebulae and White Dwarfs. The existence of stars fainter than the main sequence was recognized early in the century. Their colors and small sizes (about the same as the earth, from measured temperatures and luminosities) prompted the name white dwarfs, though blue, yellow, and red ones are known. Well-observed white dwarfs trace out almost as tight a temperature-luminosity relation as the main sequence (it is a cooling curve), but about a factor of 10^4 fainter. Some of the hottest, brightest ones are still encircled by remnants of the envelopes of their parent stars. These are called planetary nebulae only because of their vague resemblance to Uranus and Neptune when seen through small telescopes.

They are enriched in He, N, C, and O, products of nuclear reactions in their parent stars.

A few white dwarfs in binary systems have measured masses of $(0.4\text{–}1.2)M_\odot$. Spectroscopic analysis of many more indicates an average near $0.6M_\odot$ (Philip *et al.*, 1987). Most rotate more slowly than conservation of angular momentum from the main sequence would suggest; braking by magnetic winds must continue throughout stellar life. A few percent have present magnetic fields of 1–200 MG (enough to remove the degeneracy of even hydrogen atomic states and produce remarkably complex Balmer line spectra); most of the rest are less than 1 MG.

Measured masses and radii yield average densities near 10^6 g/cm^3. At this density and the interior temperature of 10^7 K, electrons are degenerate. That is, there are more of them than there are available quantum states in phase space (two electrons per box of size $\Delta x \Delta p = \hbar$). The electrons thus necessarily occupy states of higher than thermal momentum (the energy having come from nuclear reactions and gravitational potential energy as the stellar core contracted during post-main-sequence evolution). Such electrons cannot readily change their momenta and have long mean free paths, leading to energy transport by conduction and isothermal interiors.

Degenerate electrons have a unique $P(\rho)$ and $U(\rho)$ equation of state, giving rise to a unique mass-radius relation for white dwarfs (Fig. 3; Chandrasekhar, 1939). In the absence of general relativistic effects, the radius would shrink to zero at a mass of about $1.44M_\odot$ (for He, CO, or ONeMg composition), called the Chandrasekhar limiting mass. In fact, relativistic and particle physics effects lead to rapid gravitational collapse for a white dwarf with mass just below that limit. And indeed we observe no larger ones.

Among the outstanding problems in white dwarf research (Philips, 1987) are the rate of cooling of the reddest (which should lead to an independent estimate of the age of the galaxy), the origin and thickness of the very thin layers of H and He found on most white dwarfs, and the existence of at least three categories of pulsating white dwarfs, each with its own instability strip in the temperature-luminosity plane.

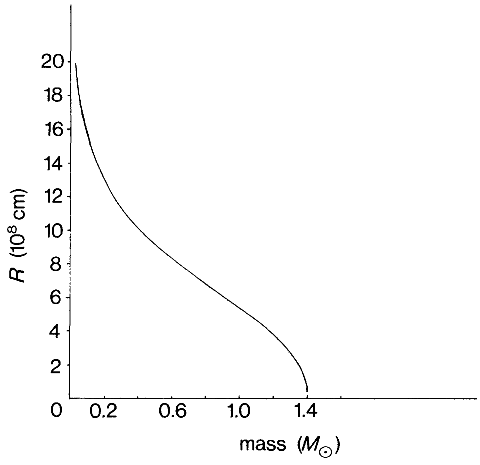

FIG. 3. Radius vs mass (Chandrasekhar, 1939) for zero-temperature degenerate dwarf with mean molecular weight per electron $\mu_e=2$ (e.g., pure He^4, C^{12}, O^{16}, Ne^{20}, etc.). In real stars, radii are increased 10% or so by finite temperatures, nondegenerate atmospheres, rotation, and magnetic fields. Effects of general relativity and particle interactions cause core collapse at about 90% of the mass at which radius vanishes in the idealized calculation. For other compositions, radius coordinates scale as μ_e^{-1} and mass as μ_e^{-2}.

1.4.2 Supernovae, Neutron Stars, and Pulsars. That neutrons as well as electrons could be degenerate was recognized very shortly after the discovery of the neutron in 1932; and the collapse of a normal star to a star made largely of neutrons was suggested as the energy source for stellar explosions by Baade and Zwicky in 1934. The 1968 discovery of a rapidly rotating magnetic neutron star (pulsar) in the center of the remnant of an explosion seen by the Chinese in 1054 (the Crab Nebula) triumphantly confirmed their main ideas (see Israel in Hawking and Israel, 1987, for more history).

About 500 pulsars are now known, including a handful in binary systems from which masses near $(1.4\text{–}1.5)M_\odot$ can be measured. The pulsed radio emission represents a searchlight effect when radiation beamed along the direction of powerful magnetic field lines sweeps across us. Most fields are $10^{12\pm1}$ G, and rotation periods range from 1.55 ms to 4 s. Fields apparently decay in $10^{6\text{–}7}$ yr, so that our galaxy must have 1000 or more times as many dead pulsars as live ones. A somewhat smaller number of neutron stars are found in binary systems (Sec. 1.5.2). The total number of known pulsars and binaries and their lifetimes are consistent with the measured supernova rate of a few per century in a largish, bright galaxy like ours.

In fact, there are at least two types of the powerful stellar explosions called supernovae. Those seen among young, massive stars result from iron core collapse (Sec. 1.3), liberate $(1–3)\times 10^{53}$ ergs (mostly in neutrinos), and leave both neutron stars and expanding gas clouds behind, the energy being the binding energy of the $1.5M_\odot$, 10 km neutron star. Those seen among older, less massive stars draw their energy from the nuclear deflagration of about $1M_\odot$ of C and O to Fe, and liberate only about 10^{51} ergs, mostly in kinetic energy of the disrupted stars, thus leaving only an expanding gaseous remnant. The progenitors seem to be white dwarfs in close binaries. Core collapse events are called Type II, nuclear events Type I or Ia.

In February 1987, a supernova in the nearby Large Magellanic Cloud galaxy rose to naked-eye visibility. Called 1987A (only because it happened to be the first supernova seen that year), it has confirmed our basic understanding of core collapse events (Trimble, 1988; Arnett *et al.*, 1990). The detected neutrino flux implied a total energy of $(2–3)\times 10^{53}$ ergs, and optical observations of both the progenitor star and the postexplosion object indicate that the collapse occurred inside a star of initial mass near $20M_\odot$ that had lost about half its envelope in winds.

Supernova explosions distribute through interstellar space the heavy elements synthesized by their progenitor stars and contribute much of the heating and turbulence that keeps interstellar gas in a state where star formation proceeds on a time scale of 10^{10} rather than 10^6 yr. Perhaps the thorniest problems connected with supernovae are

1. the precise progenitors of nuclear-deflagration events (white dwarf accreting from a companion? merging pair of white dwarfs?) and
2. the physical process that transfers 1% of the energy from core collapse to the envelope of a massive star to blow it off.

1.4.3 Black Holes. The maximum mass that can be supported by degenerate neutron pressure is much less certain than the Chandrasekhar limit to white dwarf masses. It depends on the poorly known, repulsive, short-range part of the nuclear potential and could be as small as measured neutron star masses [$(1.4–1.5)M_\odot$] or as large as $(2.5–3.0)M_\odot$. It must be near the low end of this range if any neutron star is to be compact enough to spin 2000 times a second. Notice that the masses measured by an observer at infinity are about 10% smaller than the baryonic masses that collapsed to form them, because of binding energy of the neutron star.

In any case, we expect at least some massive stars to be left with cores above that limit, and/or for later accretion to drive the remnant mass above it. Such cores, as was shown by Oppenheimer and Snyder (1939), will experience continued gravitational collapse until they reach radii less than $R=2GM/c^2$ and their escape velocities exceed the speed of light. The impossibility of receiving electromagnetic or other signals from within this radius, called the Schwarzschild horizon, is responsible for the name black hole.

The detailed properties of space-time in and around these collapsed objects present intriguing problems in general relativity (Thorne *et al.*, 1986) and are by no means uncontroversial. Quantum mechanical properties of black holes, including their potential for boiling away via Hawking radiation (Israel in Hawking and Israel, 1987), are not definitely connected with any observable phenomena.

From an astrophysical point of view (Blandford in Hawking and Israel, 1987) the important (and not controversial) properties are the existence of objects with sizes close to their Schwarzschild radii, so that accretion onto them must liberate an appreciable fraction of the rest mass of the accreted material, which will inevitably be heated thereby and radiate copiously. Isolated, nonaccreting black holes would seem to be detectable only as gravitational lenses, bending the light of background objects. No promising candidates are known.

The first, and still most persuasive, candidate for a stellar-mass black hole is the x-ray binary Cygnus X-1. The main points are that

1. it is an x-ray source with luminosity of about the maximum permitted before radiation pressure begins to blow a $(1–10)M_\odot$ object apart [Eddington luminosity $=(4\pi c m_p G/\sigma_t)\times M=6\times 10^{24}$ ergs/g,

where σ_t is the Thomson cross section], and so must contain a compact star,

2. the minimum mass of the compact component consistent with the binary orbit, light curves, and so forth is about $6M_\odot$, too large for a neutron star, and
3. the system is too compact for the x-ray component to consist, for instance, of a neutron star (providing the x rays) and a $5M_\odot$ main sequence star.

A handful of less well established cases include other systems in the Milky Way (SS 433 and A0620-00) and two (LMC X-3, LMC X-1) in the Large Magellanic Cloud.

The other context in which the existence of astrophysical black holes is generally regarded as established is as the central powerhouse in active galactic nuclei (see GALAXIES AND COSMOLOGY).

1.5 Binary Stars

Between 50% and 90% of stars are part of binary or multiple systems (distant faint companions being difficult to detect, their statistics are uncertain). In about half of these, in turn, the stars are close enough together to transfer material back and forth or otherwise influence each others' evolution (= close binaries). The result is a variety of phenomena not elsewhere possible.

Binaries are detected as such and studied by several methods, including

1. measuring positions of the images of the two stars over most of an orbit period (visual binaries, possible only for wide systems quite close to us),
2. following the changing position of an image centroid (astrometric binaries),
3. following the wavelength changes (Doppler shifts) of the absorption or emission lines of one or both stars (spectroscopic binaries), and
4. measuring the changing brightness through the orbit period as one star passes in front of the other (eclipsing binaries).

To the rest of astronomy, the chief importance of binary stars lies in the possibility of measuring stellar masses from the orbital periods P and semimajor axes a_1 and a_2 of the stars about their common center of mass using Kepler's third law,

$$G(M_1+M_2)P^2=4\pi^2(a_1+a_2)^3, \tag{5}$$

and the definition of center of mass,

$$M_1a_1=M_2a_2. \tag{6}$$

In general, this is possible only for visual binaries of known distance and for eclipsing spectroscopic binaries where two sets of lines are seen. The number of accurately known stellar masses is less than 200, and more are badly needed, especially for very bright, very faint, and highly evolved stars.

The evolutionary course of a binary system depends primarily upon the masses of the two stars and their separation, all of which can and do change from their initial values by large factors. The resulting richness of observational phenomena (variability, nonthermal emission, emission lines,...) has given rise to something like 50 different recognized types of binaries (often named for a prototype star like Algol, RS Canum Venaticorum, or SS Cygni) and some risk of botanization of the subject. Modeling of binary evolution consists of trying to keep track of the changes in mass and separation while the components attempt to pursue normal single-star evolutionary tracks and are frustrated in this by their partners. Nevertheless, it seems possible to fit most observed types into physically plausible scenarios and to find observed systems corresponding to most predicted evolutionary stages (Pringle and Wade, 1985; Sahade *et al.*, 1990).

About one binary in five or six actually consists of three or more stars, the highest known multiplicity being eight. Systems are hierarchical, with the ratio of one separation to the next being at least a factor of seven. The main effect seems sometimes to be to complicate orbit analysis beyond deconvolution.

1.5.1 Main-Sequence and Post-Main-Sequence Binaries. The classical problem in binary-star astronomy was the existence of large numbers of Algol variables, eclipsing spectroscopic pairs in which the more evolved (red giant) component was invariably less massive than the less evolved (main sequence) component, violating the well-established and -understood correlation of stellar lifetimes being proportional to about M^{-2}. The solution lies in the possibility of efficient transfer of material from one star to the other, which, in turn, arises from the geometry of the gravitational equipotential surfaces in a coordinate system that rotates with the system (Fig. 4, which is a two-dimensional cut along the axis through the centers of the two stars and should be visualized as rotated around an axis through the points L_1 and L_2—horizontal in the figure).

The closest-in equipotential enveloping both stars is called the Roche lobe (or inner Lagrangian surface); the point on it between the stars, L_1; and the point where an equipotential first connects with the outside world, L_2. Two time scales are important for the evolution of stars within this geometry—the thermal (or Kelvin–Helmholtz) time on which a star would radiate away its internal energy and on which it can readjust its structure when mass or energy is added or subtracted (10^7 yr for the sun), and the nuclear time scale, over which a significant fraction of the hydrogen or other fuel is used up (10^{10} yr for the sun on the main sequence).

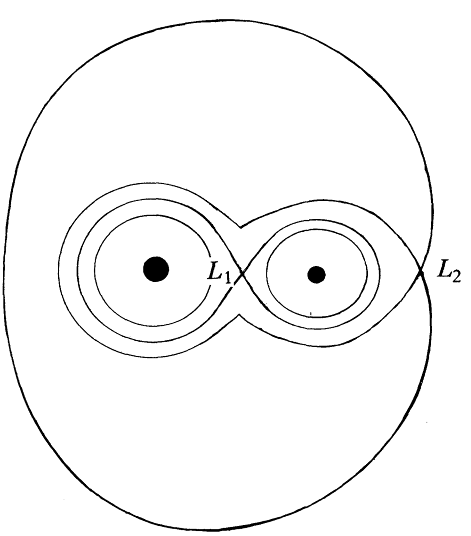

FIG. 4. Equipotential surfaces in the rotating coordinate system of a binary with initial masses in the ratio 2:3. The larger Roche lobe surrounds the more massive (and larger) star, the two lobes meeting at point L_1. A star which expands to fill its lobe can drizzle material onto its companion with high efficiency. There also exist contact systems in which both stars fill their lobes (orbit periods generally less than one day) and common envelope systems in which (usually as a result of rapid mass transfer) material extends out as far as L_2 and can be lost from the system, carrying energy and angular momentum with it. If the stars are assumed to be main sequence stars of (3 and 2)$M_\odot$ with the radii to scale as shown, then the orbit period would be about six days.

Consider two main-sequence stars separated by less than 100 or so stellar radii. The more massive, M_1, will use up its H first and begin to expand as it tries to become a red giant. Eventually it will fill its Roche lobe (subsequent details depending a good deal on whether it does this during core H burning, shell H burning, or shell He burning), and material is then free to stream exoergically down from L_1 onto M_2. The Roche lobe around M_1 shrinks, and so mass transfer occurs rapidly, on the thermal time scale, until the initial mass ratio is reversed and $M_2' > M_1'$. The Roche lobe then expands with further transfer, which therefore continues over the much longer nuclear time scale. Hence we see most systems with the mass ratio already reversed, resolving the traditional Algol paradox. Mass and angular momentum can also be lost completely from the system when, as often happens, transferred material forms a common envelope around both stars and reaches L_2.

In subsequent evolutionary phases, M_1' becomes a white dwarf, neutron star, or black hole (depending upon its mass); and M_2' completes its main sequence life, begins to expand, transferring material back onto the now-compact M_1'. M_2' eventually evolves also to a white dwarf, neutron star, or black hole, the system sometimes being disrupted in the process, but not always. Notice that at the time M_1' sheds a planetary nebula or becomes a core-collapse supernova, it contains less than half the total system mass; thus symmetric ejection from it cannot unbind the system. M_2', on the other hand, contains more than half the total mass at its demise.

1.5.2 Cataclysmic Variables, X-Ray Binaries, and Binary Pulsars. Transfer from the main-sequence or red-giant secondary back onto a white dwarf, neutron star, or black hole primary occurs for orbit periods of minutes to months, and the gas stream from L_1 always has far too much angular momentum to settle directly onto the compact recipient. Thus the gas is spun into an accretion disk (Fig. 5), and detectable light is possible from the disk, the gas stream, and a hot spot where they meet, as well as from the two stars. All these components are seen in some systems. Variability in luminosity and spectrum can be expected on time scales associated with the system orbit period, rotation of each of the

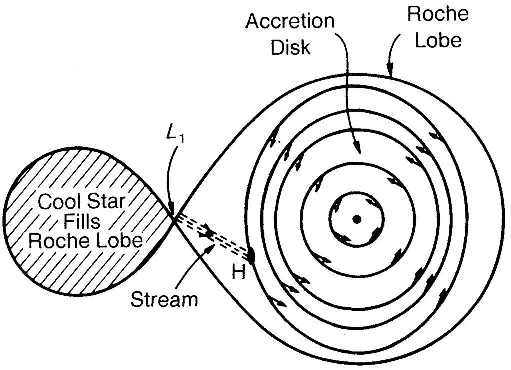

FIG. 5. Idealization of a highly evolved binary, in which an extended secondary is transferring material through L_1 back onto a compact (white dwarf, neutron star, black hole) primary. Diagram is drawn for a mass ratio compact:normal of about 1.5. Transferred gas has too much angular momentum to settle directly onto the compact object and so forms an accretion disk, with a hot spot where the gas stream hits the disk. Gradual outward transport of angular momentum in the disk (presumably by magnetic fields and turbulence) allows accretion onto the star, liberation of gravitational potential energy, and radiation of that energy.

stars, orbit periods in the disk, and assorted instabilities in both stars and disk. Again all of these are seen somewhere. The depth of the gravitational potential well determines that accreting white dwarfs will be sources of visible, UV, and soft x-ray radiation, while neutron stars and black holes will produce largely x rays and sometimes γ rays.

Systems with white dwarf recipients are the cataclysmic variables, including novae, dwarf novae, polars (where the white dwarf has a 10–40 MG magnetic field), and many others. Most of the light variations come from instabilities in the stars and disk, but the largest, nova, outbursts, in which 10^{44-45} ergs are emitted over a few weeks, represent explosive burning of a layer of degenerate hydrogen that has accumulated on a white dwarf surface over the past 10^{2-5} yr.

Most or all of the accumulated material is blown off in most cases. If the white dwarf can burn and keep the transferred material, this would provide a mechanism to drive an old white dwarf above the Chandrasekhar limit, producing a Type Ia supernova. Merger of two white dwarfs after the second star has completed its evolution is another possibility. At the moment, many binary white dwarfs are known, but none has sufficiently large total mass and short enough orbit period that an explosive merger could occur over the next 10^{10} yr.

Systems with neutron-star and black- hole recipients are the x-ray binaries. The black hole candidates are distinguished by

1. masses too large for stable neutron stars and
2. absence of periodic variations on the 1–1000-s time scale of neutron star rotation.

Again there are many subtypes, including young ones with companions more massive than the neutron star and old ones with less massive (down to $0.05M_\odot$) companions. Accreted hydrogen normally burns steadily, but the product helium can accumulate and explode in some systems, producing x-ray bursts lasting a few tens of seconds. Finally we have detected (initially with the Vela satellites, meant for monitoring nuclear explosions) several hundred short bursts of keV to MeV gamma rays. None of the sources has an unambiguous optical, x ray, or radio identification, and the population is remarkably isotropic around the sky (while pulsars and x-ray binaries are concentrated in the plane and toward the center of our galaxy). These gamma ray burst sources are successfully modeled as nuclear explosions on magnetic accreting neutron stars, but they could be something entirely different and perhaps extragalactic.

Neutron stars in close binaries are never seen as pulsars as long as the companion is a main-sequence or red giant star because stray gas from it keeps the local plasma frequency above the range of pulsar radiation. Once the second star becomes a white dwarf or neutron star, pulsar radiation can emerge. The number of known binary pulsars (12 at epoch 1989.7) is increasing rapidly. There are a few each with established white dwarf and neutron star companions, and orbit periods range from hours to months. In most cases, the pulsar displays a weak magnetic field near 10^9 G (as a result of more than 10^7 yr of decay) but a short rotation period of 1–100 ms (as a result of spin-up by gas transferred from the companion before it became compact). The handful of known weak-field, short-period single pulsars are thought to be former members of similar systems, liberated by the explosion of the companion or its gradual erosion. In principle, systems with two pulsars, one old and one young, should be possible, but none is currently known.

2. INTERSTELLAR MATTER AND THE FORMATION OF STARS

2.1 Structure and Kinematics of the Interstellar Medium (ISM)

Diffuse interstellar gas and dust make up roughly 10% of the (luminous, baryonic) mass of our galaxy. Of this $\sim 10^{10} M_\odot$, about 99% is gas (molecular, atomic, and ionized) and 1% is micron and submicron sized dust particles. The chemical composition is that of young stars, 73% hydrogen, 25% helium, and 2% heavier elements (in the same proportions as in the sun and meteorites), half of which is locked up in the grains of dust. The ISM shares the rotation of the galactic star disk and is the best tracer of it, since gas, being dissipative, cannot long sustain large chaotic noncircular motions.

Other spiral galaxies also have 5%–15% of their luminous material in gas and dust with properties much like that in the Milky Way. Irregular galaxies can be up to 50% ISM by mass, and the heavy-element abundance is normally lower. Elliptical galaxies are often free of neutral interstellar gas down to limits much less than 1% of their total masses, but many show emission lines from ionized gas (typically near the center), dust lanes (implying dense molecular gas), or x-ray emission implying up to 10% of their masses in ionized gas above 10^6 K.

Scheffer and Elsaesser (1987) provide a general introduction to the ISM; many of the detailed observations and processes are discussed in Hollenbach and Thronson (1987) and in Tenorio-Tagle *et al.* (1990).

2.1.1 Phases of the Interstellar Medium. Table 1 summarizes the components presently believed to coexist in the ISM. All, except probably the coronal, are very clumpy in both physical and velocity space. The components are many, partly because sources and sinks of both gas and energy are patchily distributed and partly because gas cooling rates (Fig. 6) are a complex function of temperature. With a cooling rate like that shown and constant heat input, pressure versus density takes the form shown in Fig. 7, so that there are two stable (and two unstable) phases in pressure equilibrium for the conditions that obtain in the ISM (for which the heating rate is very roughly $10^{-25} n_H$ erg/s, and the pressure in the equilibrium phases $nT = 3000$ K cm^{-3}).

In addition to the pressure equilibrium (cool neutral, warm neutral, warm ionized) phases, there are dense clouds of molecular gas, at higher pressure, but self-gravitating, so that, in the absence of additional turbulent and magnetic pressure support, they would collapse in their free-fall times, $t \approx (G\rho)^{-1/2} \approx 10^6$ yr. If this happened, the gas would all long since have been converted to stars (as in elliptical galaxies). These clouds are disrupted by the first generation of stars they produce and reform where less dense gas is swept together by magnetic pressure and expanding supernova remnants and stellar wind shells. A typical cloud has a mass of $10^5 M_\odot$ but is disrupted after forming a cluster of only 10^{3-4} stars, the process thus being not very efficient.

The chemistry of molecular clouds is exceedingly complex (Dalgarno in Tenorio-Tagle *et al.*, 1990) involving a combination of gas-phase reactions (usually with one of the reactants ionized) and molecule formation on dust grain surfaces followed by sputtering off. Some 50 distinct molecular and radical species have been found, primarily from emission lines in the microwave and infrared spectral regions. The largest numbers of distinct species occur in the densest clouds, at the galactic center and in the Orion Nebula. Many isomers (e.g., HCN and HNC) and isotopic variants ($C^{13}O$, CO^{18}, etc.) also exist.

H_2 is obviously the commonest molecule and observable only with difficulty (since it has no dipole moment). CO is the most abundant tracer, and the uncertainty in converting from number of CO molecules seen to number of H_2 molecules present is responsible for most of the disagreement about whether the molecular gas is at least half the total or only about one-quarter. Detected molecules include familiar substances like water, ammonia, methyl and ethyl alcohol, formaldehyde, and formic acid, and a number of the obvious radicals (OH, CH, CN), but also some combinations unknown in the laboratory like HC_9N.

Table 1. Phases of the interstellar medium.

Phase	Temp. (K)	Particle density (cm^{-3})	Scale height (pc)	Fraction of mass	Gas bulk speeds (km/s)	Fraction of volume	Observed phenomena	Heating processes	Cooling processes
Dust	5–40	10^{-12}	40	1%	2–4	Small	Reflection, extinction, polarization, reddening of starlight	Collisions, cosmic rays, photoabsorption	Thermal IR emission
Molecular clouds	10–30	10^{2-5}	60	1/3–1/2	2–4	Small	Molecular emission lines	Stellar radiation; collisions	Molecular lines, esp. CO
Neutral gas				1/2–2/3		35%			
Cool clouds[a]	80	30	100	1/4–1/3	7	5%	21 cm absorption & emission	Photoelectric heating by dust	Fine-structure lines of C, C^+, O, etc.
Warm intercloud[a]	8000	0.1–0.3	200	1/4–1/3	9	30%	H II emission lines	Photoelectric heating by dust	Atomic emission lines, esp. H
Ionized gas				10%		65%			
Warm ionized[a]	8000	0.1–0.3	⩾1000	10%	10–20	15%	Diffuse Hα emission	Stellar photoionization	Line emission of C, N, O, Ne
Coronal[a]	10^{5-6}	$10^{-3\pm1}$	⩾2000	Small	Up to 1000	50%	Soft x rays, coronal lines	Supernovae	Line emission and continua of H, C, O, He
H II regions	8000	⩾1–100	60	Small	10–30	Small	H recombination & ion emission lines	Individual stars	Emission lines

---

[a] Phases in approximate pressure equilibrium.

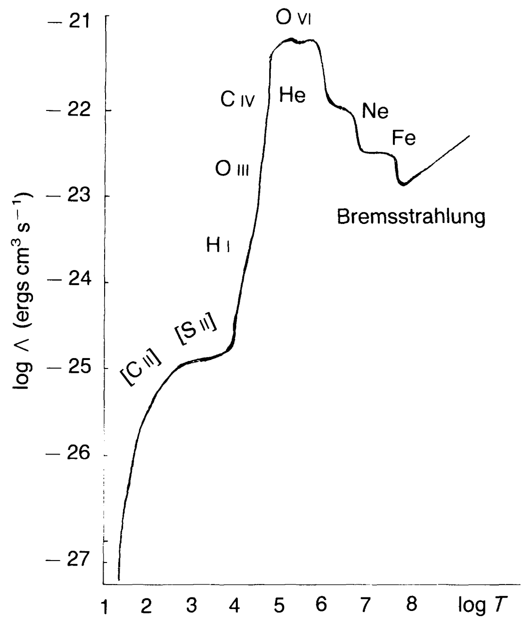

FIG. 6. Equilibrium cooling rate of interstellar gas in ergs cm^{+3} s^{-1} (The rate in ergs cm^{-3} s^{-1} is the plotted value times the square of the particle density) vs log T for average composition and ionization conditions. Ions shown are those that dominate the cooling at each temperature. [C II], etc., indicate forbidden transitions like the 158-μm line from a low-lying state of ionized C. Data from J. M. Shull in Hollenbach and Thronson (1987) and from B. Lazareff in Audouze *et al.* (1983).

The dust needed for H_2 formation consists (judging from its emission and absorption properties) of cores of silicates and carbon in several forms (graphitic, amorphous, and organic, perhaps polycyclic aromatic hydrocarbons) which form in the cool winds from evolved stars and in the ejecta from novae and supernovae. These are coated by mantles of ice (H_2O, CH_3OH, NH_3, CO, etc.) which they acquire in the dense molecular clouds.

Coronal gas, at the other end of the temperature range, emits the same sorts of lines of highly ionized species as does the solar corona (hence the name). It arises from interconnection of expanding supernova remnants. A topic of hot dispute is whether it occupies rather less than half, more than half, or essentially all of interstellar space. That is, do expanding supernova remnants sweep up all the diffuse gas they encounter into dense clouds, or do they manage to coexist with a comparable warm neutral and ionized phase? A good case can be made for the latter situation (Cox in Tenorio-Tagle *et al.*, 1990; Kulkarni and Heiles in Hollenbach and Thronson, 1987), which has been assumed in Table 1.

H II regions are residual molecular gas, left from star formation, ionized by the new, hot stars. Their sizes are set by the number of Lyman continuum photons emitted by the stars. Their pressures exceed that of the surrounding medium, so that they expand, opening up holes in the cool gas around them, as do winds from new stars. Where dust is copious, essentially all the radiation from these regions may emerge in the infrared.

The warm ionized gas contains most of the free electrons in the galactic disk. These account for the temporal dispersion of pulses received from pulsars and for the smearing of images of extragalactic radio sources. With n_e=0.3 cm^{-3} in that phase and $\langle n_e \rangle$=0.03 along the average pulsar line of sight, the filling factor must be about 10%. Given this distribution, the problems of heating and ionizing the interstellar medium largely decouple. The interstellar matter immediately surrounding the solar system belongs to this warm ionized phase (Cox and Reynolds, 1987).

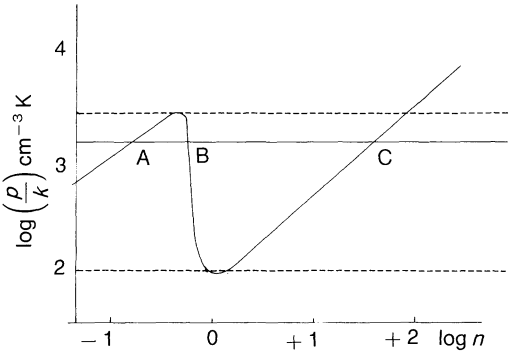

FIG. 7. Pressure (nT in cm^{-3} K) vs density (in cm^{-3}) for interstellar gas with constant heating rate and cooling vs temperature as shown in Fig. 6. For pressures between the dashed lines, several states have the same pressure. At any one P, like that indicated by the solid line, the phase B is unstable (because compression leads to lower P, not higher), while C represents the diffuse clouds and A the intercloud medium of neutral gas. There is a fourth, unstable, state at still lower density and T near 10^5 K, but gas at this temperature is in any case ionized, and the coronal component is not in equilibrium. Data from Scheffer and Elsaesser (1987) and from Shull in Hollenbach and Thronson (1987).

Pressure (or energy density) equilibrium of another sort exists among

1. motions of neutral gas clouds,
2. galactic cosmic rays, and
3. the galactic magnetic field.

All values are of order 1 eV/cm^3, the equipartition presumably resulting from physical interactions. The disk density of starlight energy is also about 1 eV/cm^3, which may be a coincidence. Cosmic rays are addressed in the article ASTRONOMY.

The magnetic field of the Milky Way (and of many other spiral galaxies) has a strength of a few μG in the galactic plane, with comparable energy densities in an ordered azimuthal pattern (aligned either with the spiral arms or with the circular gas motion) and in a chaotic component of scale length 10–100 pc (1 pc$=3\times10^{18}$ cm). The azimuthal component reverses direction at least several times across the galactic disk. Origin of the field has been attributed both to rotationally amplified primordial field and to a self-excited dynamo, with the dynamo models currently favored by most theorists (Zweibel in Hollenbach and Thronson, 1987). In addition to the azimuthal disk field, at least some spirals (including ours) have also field lines extending more or less vertically out of their planes to 1–2 kpc or more. This is thought not to be a separate component but to be a disk field dragged out by expanding supernova remnants and other gas flow structures. The field in the inner 50 pc is a dipole of mG strength (Morris, 1989, which addresses also many other important aspects of the galactic center).

2.1.2 Large-Scale Distribution and Motions of the ISM. The cooler gas phases, as indicated in Table 1, are largely confined to the galactic disk. That is, the distances over which their densities drop to 1/*e* of the in-plane values (scale heights) are 200 pc or less. In addition, the molecular clouds, especially the large ones, are concentrated

1. toward the galactic center,
2. in a ring at 4–6 kpc from the galactic center (on a scale where the sun is at 8.5 kpc), and
3. in the spiral arms of the galaxy (Scoville and Sanders in Hollenbach and Thronson, 1987).

Neutral hydrogen, on the other hand, is relatively sparse inward of 3 kpc but has roughly constant density from there out to 12 kpc and is measurable still further out. Both the molecular and the H I disks flare in their outer reaches. The outer part of the H I disk in addition is warped and rippled. These deviations from uniform flatness reflect the spherical shape of the outer galactic gravitational potential and interactions with the companion Large and Small Magellanic Cloud galaxies.

The extent of the galactic gaseous halo is an ancient topic of debate. The current consensus is (Tenorio-Tagle *et al.*, 1990) that coronal gas and the electrons persist out to at least several kpc. In addition, 15% or so of the H I clouds are outside 300 pc, some of them having large (often negative) velocities that are not obviously part of the general galactic rotation.

Gas is our best tracer of the galactic disk rotation because it does not readily sustain noncircular motions in the face of cloud collisions and dissipation. Out to the position of the sun, the most useful velocities come from Doppler shifts of the 21 cm line (hyperfine structure transition) of neutral hydrogen (H I), because there is no ambiguity about the position of the gas with the largest velocity along any given line of sight (see Mihalas and Binney, 1981, or Binney and Tremaine, 1987, for how the analysis is done). Outside the solar circle, this convenient property is lost. The best velocities are probably those from CO millimeter lines, because the distance to each CO complex can be found from young stars recently formed in it (Blitz in Athanassoula, 1983). The resulting plot of circular velocity versus radius (Fig. 8) is still flat or rising outside the solar position. This is one of the important pieces of evidence for dark matter in our galaxy, since the mass interior to point R with circular velocity V_c is roughly $M=V_c^2R/G$.

Gas flows both in and out of the galactic disk. Outflow is in the form of expanding bubbles (driven by supernova remnants and stellar winds) that pierce the disk layer and penetrate the halo. Most or all of that gas eventually returns in a sort of galactic fountain or champagne (Tenorio-Tagle *et al.*, 1990). In addition, some high velocity clouds of H I appear to be coming in on parabolic or hyperbolic orbits, and could be adding (1–2)$M_\odot$ of

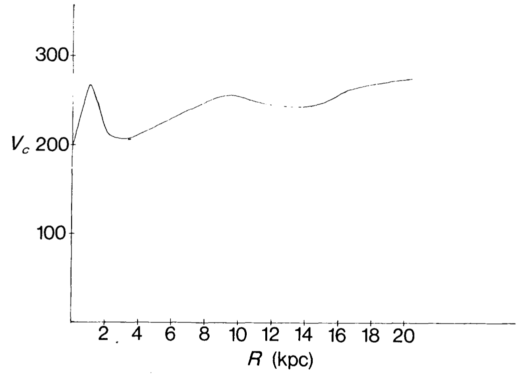

FIG. 8. Rotation curve of the Milky Way, showing circular velocity in the disk vs radius in kiloparsecs (1 kpc= 3×10^{21} cm), on the old scale where the sun is taken to be 10 kpc from the center and the circular velocity there 250 km/s. Ripples are expected from noncircular motions in spiral arms. Velocities are those of the 21 cm H I line out to 10 kpc (from Scheffer and Elsaesser, 1987, and Binney and Tremaine, 1981) and those of CO emission further out (from Blitz in Athanassoula, 1983). Still closer to the center, velocities have been measured for Ne II infrared emission lines (Morris, 1989). They imply a very compact (star cluster and/or black hole) central mass of about $10^6 M_\odot$.

unprocessed ($\frac{3}{4}$ H, $\frac{1}{4}$ He) material to the galaxy each year from the extreme halo or intergalactic space. Models of galactic chemical evolution (Sec. 3.1.4) indicate the need for such input.

2.1.3 Energetics and Stability. Supernovae contribute a few times 10^{49} ergs/yr of high-grade kinetic energy and heat to the ISM and maintain the coronal gas phase largely by boiling material away from cooler clouds. Winds from massive stars contribute comparable amounts of kinetic energy but at lower velocities (Lamers in Audouze *et al.*, 1983). Because stars form in clusters that often hold together for the lifetimes of the most massive stars in them, this energy input is very patchy and tends to blow large holes in the cooler gas phases (Tenorio-Tagle *et al.*, 1990).

Gas outflow speeds are in the range 10–100 km/s for stellar winds and 100–1000 km/s for aging supernova remnants. But sound speeds in the interstellar medium, $c_s = \gamma kt/\mu m_H$, are 1–10 km/s in the cool components (up to 100 km/s in the coronal gas). Thus interactions of outflow with surroundings take the form of shocks, whose structure and visible manifestations have received a great deal of attention (McKee and McCray in Dalgarno and Layzer, 1987; Shull and Draine in Hollenbach and Thronson, 1987).

Ionization in the diffuse gas is produced largely by ultraviolet photons from stars. Cosmic rays also contribute and are the main ionization source in molecular clouds, where they can penetrate but UV photons cannot. The several ISM components attempt to cool themselves, largely by line radiation, at the same rate they are heated. This line radiation is conspicuous in the spectra of other spiral galaxies, though the luminosity available from reprocessed supernova and stellar wind energy is only a few percent of the stellar continuum luminosity. Where there is vigorous star formation in dusty regions, up to half the luminosity of a galaxy may emerge as thermal infrared; it is probably about 10% for the Milky Way.

The multiphase interstellar medium is surprisingly stable (Cowie in Hollenbach and Thronson, 1987). Runaway heating would occur only for a star formation rate nearly 1000 times the present one. A smaller increase automatically damps itself down by stirring gas up too much to condense. Decreases in star formation rate seem to be similarly self-regulating (Dopita in Faber, 1986), though there exists one galactic-mass concentration of H I, in the galaxy group around M96 in Leo, where apparently no star formation has ever occurred.

2.2 Star Formation

About $10 M_\odot$/yr must have condensed from gas into stars on average over the age of our galaxy. The present rate is probably somewhat lower, (3–5)$M_\odot$/yr. If the interstellar gas were not replenished, the galaxy would be within the last 10%–20% of its life as a star-forming entity. About $1 M_\odot$/yr is returned by evolving stars and (1–2)$M_\odot$ of chemically pristine gas may be falling in from the outer halo or intergalactic space, but the Milky Way is clearly past its prime as a star maker.

Modern studies of star formation are largely the domain of infrared and radio astronomers, because the processes occur in dense, dusty clouds of molecular gas (opaque at optical wavelengths) and at low temperature. Evidence for recent or ongoing star formation in a given region includes

1. massive, short-lived stars,
2. less massive stars with colors and luminosities away from the main sequence relation and with surface activity and vigorous winds (T Tauri variables),
3. compact infrared sources with color appropriate to contracting clouds or reradiation of starlight by dust, and
4. molecular clouds that are unusually warm or that contain compact cores of stellar mass and density $\gtrsim 10^4$ cm^{-3} detectable in the lines of NH_3, H_2CO, CS, HC_3N, etc.

A peculiarity of star formation is its low efficiency. Large molecular clouds have masses in excess of $10^5 M_\odot$. Young star clusters in the Milky Way rarely reach $10^4 M_\odot$, and many clouds where we see young stars are already in the process of dissipating, with most of the gas uncondensed. This is clearly the norm, or all spirals would have exhausted their gas long ago. But a few "star burst" galaxies display current (and obviously temporary) star formation rates of $(100\text{–}1000) M_\odot/\text{yr}$ without having enormously larger quantities of gas than does the Milky Way. They are, therefore, converting gas to stars with very high efficiency, that is, on the gas free-fall time scale. Bound clusters of stars result only from relatively efficient ($\gtrsim 10\%\text{–}50\%$) star formation. Less than 10% of all stars seem to have formed in such bound clusters, and most of them have been torn apart since.

The ultimate goal of star formation studies is to be able to predict from the properties of a given gas cloud or assemblage of clouds in a galaxy the rate at which it will convert gas into stars and the spectrum of stellar masses, $N(M)$, it will produce (including the numbers, mass ratios, and separations of binary pairs). We cannot currently do any of these. Thus star formation rate and mass spectrum are treated as adjustable parameters in studies of galactic evolution, though each must really be determined by the gas properties.

The observed spectrum of stellar masses (Scalo, 1986) varies from one cluster to another, though the average over a whole galaxy may be fairly constant. Broadly, $N(M) \propto M^{-2}\text{–}M^{-2.5}$ over the range $(1\text{–}100) M_\odot$. The power law becomes less steep (Gaussian fits are at least as good) below $1 M_\odot$, with a broad maximum from $(0.6\text{–}0.15) M_\odot$. Very few stars below $0.1 M_\odot$ are known. Whether the deficiency is real or the result of observational selection acting on a population that really continues to increase toward smaller masses is a topic of as hot debate as can be expected for such cool stars (Liebert and Probst, 1987).

Star formation is probably not a single process. In particular, high- and low-mass stars form in different kinds of clouds and perhaps with different dominant physics. In addition, we see present star formation only in disk (spiral and irregular) galaxies. Very different conditions may have been obtained during the formation of spheroidal populations (elliptical galaxies, halos of spirals, globular clusters, gas flows in x-ray emitting galaxies). Next to nothing is known observationally about star formation in these contexts. The argument has, in fact, been made (Larson in Faber, 1987) that essentially all star formation really occurred in disk systems, the spheroids resulting from later mergers of them.

Shu *et al.* (1987) have reviewed the general subject of star formation with an emphasis on the importance of magnetic fields. Overviews of specific processes can be found in Faber (1987), Hollenbach and Thronson (1987), Lonsdale-Persson (1987), Gilmore and Carswell (1987), Peimbert and Jugaku (1987), Boss (1988), Tenorio-Tagle *et al.* (1990), and Capuzzo-Docetta *et al.* (1990).

2.2.1 Contraction and Condensation of and in Molecular Clouds. Molecular clouds are self-gravitating and gravitationally bound, with respect to both the tidal field of the galaxy and their own internal energies. Another way to say this is that, in average neutral diffuse clouds, the Jeans mass (the smallest perturbation that will grow),

$$M = \left(\frac{5kT}{Gm_H}\right)^{3/2} \left(\frac{3}{4\pi\rho}\right)^{1/2} = 1.82 \times 10^{23} T^{3/2} \rho^{-1/2}\ \text{g}, \tag{7}$$

is equal to a typical molecular cloud mass, and within the clouds it ranges from $100 M_\odot$ on down. Admittedly gravitation alone cannot create density perturbations, but in the present interstellar medium, there are plenty of cloud–cloud collisions, expanding wind shells, and so forth to do this. Yet star formation is inefficient, and clouds do not fragment into stars on either their dynamical or their thermal time scales.

Additional pressure support comes from three sources (Audouze *et al.* 1983; Scalo in Hollenbach and Thronson, 1987)—turbulence within the clouds, rotation, and magnetic fields. It has been widely believed that most clouds were in equilibrium between gravitation and these pressures, and so the onset of star formation had to be triggered by a collision. Collisions with shock waves at the leading edges of spiral arms would then give rise to star formation concentrated in well-defined arms (as seen in some spirals), while collisions with other clouds, expanding H II regions, and supernova remnants would give rise to star formation propagating away from a starting point. Differential rotation then twirls the zones of star formation into the semblance of arms (as seen in some other spirals). Such triggered star formation should start on the impacted side of the cloud and work its way inward. Such patterns are seen, but are by no means universal (Lonsdale-Persson, 1987; Peimbert and Jugaku, 1987). In particular, the molecular cores that precede low-mass stars seem to be studded at random through their parent clouds (Emerson in Peimbert and Jugaku, 1987).

The alternative is that star formation is an intrinsic (but slow!) property of molecular clouds, and that the real effect of density waves and collisions is to collect the gas into spiral arm structures, so that star formation is indeed seen there (Elmegreen in Peimbert and Jugaku, 1987). One implication is that large clouds are not necessarily long-lived. They may assemble out of smaller clumps within spiral arms and get torn apart again within a few crossing times ($\approx 10^7$ yr). Since the binding energy of a typical cloud is only 10^{50-51} ergs, one strategically placed supernova or the winds of a few massive stars can accomplish the dissection.

Star formation will be intrinsic if magnetic pressure dominates (Shu *et al.*, 1987; Mouschovias in Tenorio-Tagle *et al.*, 1990), that is, if

$$M/R^2 \leqslant CB/G^{1/2}, \tag{8}$$

where C is a combination of small numbers $= 0.13$. Notice that this is a condition on surface or column density through the cloud. For typical cloud fields of 10–100 μG, the limit is 0.005–0.05 g/cm^2 or a column density of 10^{21-22} atoms/cm^2. Clouds more compact than this will be able to take the field with them, and may yield efficient triggered or propagating star formation via hierarchical fragmentation (since the Jeans mass drops as density rises). Fragmentation stops when its products are optically thick and cannot radiate efficiently, at about $0.01M_{\odot}$.

Less compact clouds might seem able to collapse at least along the field lines. In practice, the presence of a field means that any form of turbulence generates magnetohydrodynamic waves and so pressure along, as well as perpendicular to, the field. Most of the gas is neutral and, in order to contract, must diffuse relative to the field-carrying ionized component (maintained by cosmic ray ionization even when no UV photons are present) in the face of very short time scales for ion-neutral collisions. The process is called ambipolar diffusion, and, for ionization fractions, densities, and thermal velocities characteristic of molecular clouds, has a time scale 5–10 times the dynamical one, that is, $\sim 10^7$ yr. Density fluctuations then grow quasistatically until they exceed the criterion of Eq. (8) and can collapse more rapidly. Thus fragmentation and star formation still occur, but more slowly.

Support for this two-mode picture of star formation comes from several directions. First, surveys of other galaxies indicate that the rate of formation of massive stars is proportional to gas density squared, as you would expect for a process involving collisions and such, while that of low-mass stars is linear in the gas density, as expected for an intrinsic process (Scoville and Sanders in Hollenbach and Thronson, 1987). Second, newborn, low-mass (T Tauri) stars and the cores that precede them are found both in relatively low-mass cloud complexes (like that in Taurus) and in giant ones (like the Orion Nebula region), while only a fraction of the largest clouds contain new massive stars and protostars. Since small, cold clouds that can form low-mass stars are found throughout the galactic disk, only massive star formation should be heavily concentrated in spiral arms, as indeed seems to be the case.

Two points may be noted. First, in neither form of star formation is the physics closely related to what the product star is going to do. Thus there is no reason to expect the mass spectrum to terminate at the $0.085M_{\odot}$ and $100M_{\odot}$ limits for stable, nuclear-burning stars. Larger entities blow themselves apart, but smaller ones should persist. In some sense, we expect substellar objects. Second, angular momentum must be disposed of at all stages of contraction and collapse. Observed rotation speeds of the clumps in molecular clouds are small (1–2 km/s) but imply far more angular momentum than could be accommodated by the product stars. Braking is universally assumed to be magnetic, but details (including the density at which ambipolar diffusion allows most of the flux to escape) remain obscure.

2.2.2 Protostellar Cores. Molecular maps of nearby cloud complexes have revealed hundreds of stellar-mass cores with densities $\gtrsim 10^4$ cm^{-3}. About half of these contain infrared sources detected by the IRAS satellite (Myers in Peimbert and Jugaku, 1987) and so already have luminous objects inside. Statistical arguments indicate that, for an eventual $1M_\odot$ star, the times to form a core, produce a protostar at the center, and clear away enough residual material for the star to be seen are each of order 2×10^5 yr.

A core produced by gravitational collapse (after ambipolar diffusion has removed the magnetic field) has approximately the structure of an isothermal sphere, $\rho \propto 1/r^2$; thus all time scales are shortest at the center, and evolution proceeds from the inside out. Accretion of surrounding material onto the core occurs at a rate that is crudely independent of time and core mass, though not of gas density, temperature, etc., and is about 2×10^{-6} $M_\odot$/yr for conditions inside the Taurus molecular cloud. Detailed models of emergent infrared flux imply a core-halo structure, where interpreting the halo as an accretion disk yields at least as good a fit as other symmetries (Myers in Peimbert and Jugaku, 1987; Shu *et al.*, 1987; Tenorio-Tagle *et al.*, 1990).

Apparently the mass of a star is set by whatever process terminates the accretion (Larson in Tenorio-Tagle *et al.*, 1990), since the molecular cloud provides an effectively infinite reservoir of accretable material. It is not clear how the termination of accretion is related to other differences, previously noted, between the sites of high- and low-mass star formation.

The transition phases in which accretion terminates and gravitational energy production gives way to nuclear energy production are not well observed. For massive objects, the accretion time scale is longer than the Kelvin–Helmholtz time, and we do not expect to observe any true protostars. Even among low-mass objects, there are no unambiguous cases of contraction- or accretion-dominated luminosity. The problem may be largely observational, arising from poor angular resolution at long wavelengths and obscuration at shorter ones.

2.2.3 Pre-Main-Sequence Stars and Planetary Formation. All or most stars retain traces of their protostellar disks when deuterium burning begins (Rodriguez, Kaifu, Nakano in Peimbert and Jugaku, 1987). While the deuterium-to-hydrogen ratio is only $(1\text{–}2)\times10^{-5}$, deuterium fusion releases enough energy to stop or slow down contraction and to drive convection, which enables dynamo-generated magnetic fields to turn rotational energy into surface activity and winds (Shu *et al.*, 1987). The wind emerges first at the poles and gradually fans outward until most of the residual circumstellar material has been cleared away. The intervening phase of bipolar outflow gives rise to a variety of mappable and modelable emission and reflection features, which are the subject of vigorous investigation, partly (but only partly) because the collimation of the flows by the residual disks seems to have some relevance to the collimation of jets from quasars and active galaxies (Koenigl in Belvedere, 1989; Pudritz and Fich, 1989).

From an anthropocentric point of view, the most important aspect of these disks is that many of them appear to have the masses, Keplerian structure, and sizes attributed to the protoplanetary disk from which our own solar system formed. Something like half of $1M_\odot$ pre-main-sequence stars have disks of this kind in one survey (Sargent in Tenorio-Tagle *et al.*, 1990).

2.3 Star Clusters and Stellar Populations

These subjects fit naturally together because the two classes of star cluster, globular and open (or galactic), represent the extremes of galactic stellar populations, consisting respectively of the oldest and the youngest stars. Figure 9 is a schematic representation of an edge-on view of our Milky Way galaxy.

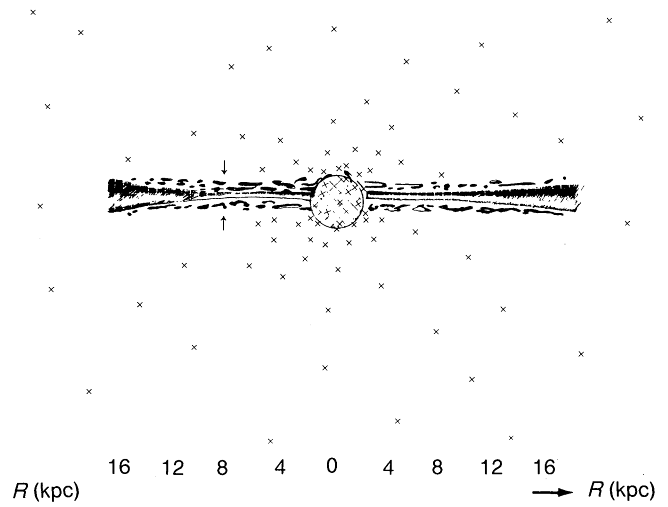

FIG. 9. Schematic representation of our galaxy seen edge-on. Central nuclear bulge contains old, metal-rich stars. The thin disk (flaring outside the position of the sun, shown by arrows at 8.5 kpc radius) includes the interstellar medium and young stars (formed in clusters, studded through the spiral arms in the disk). Globular clusters, shown as X's, trace the halo population (though making up only 1% of its stars) and extend out much further than shown, to at least 50 kpc, though the majority are closer to the center than is the sun. There is probably also a transitional thick disk population with a scale height of about 1 kpc as shown.

The youngest stars, along with the interstellar gas and dust, are confined to a thin disk. These are called Population I objects, are relatively rich in heavy elements ($\gtrsim$ 2 %), and will look blue from a distance because the brightest stars are hot main sequence ones. In a face-on view of the galaxy, star formation regions and the youngest open clusters would be dotted like beads along the spiral arms.

Globular clusters of $\approx 10^5$ stars each are the most conspicuous members of the spheroidally distributed Population II, though 99 % of the halo stars are not cluster members. The stars are all $\gtrsim 10^{10}$ yr old and have heavy element contents $\lesssim 0.5$ % (down to 10^{-6} or less but never zero). The brightest Pop II stars are red giants, so that from a distance, the halo would look redder than the disk. The extreme halo population is essentially nonrotating, and as many stars and clusters have retrograde as prograde orbits, many of them quite eccentric.

Older, but still metal-rich, Population I objects inhabit a somewhat thicker disk than the youngest ones, and there may be a transition population, between disk and halo, averaging about 0.3% heavy elements and having a scale height near 1 kpc. This thick disk includes some of the more metal-rich globular clusters and rotates more slowly than the thin disk whose circular velocity versus radius is shown in Fig. 8. Notice that astronomers use "metals" to mean all elements apart from hydrogen and helium (even though C, O, and Ne make up about half the total), because the heavy elements with strong lines at stellar temperatures really are mostly metals—Na, Ca, Fe, Ti, etc.

The two cluster types have in common that each represents a sample of stars of a single age and chemical composition. As a result

1. plots of surface temperature versus brightness can be used to measure their ages (by comparison with models for the time taken to exhaust hydrogen fuel, Sec. 1.3) and to test our understanding of stellar evolution, and
2. counts of stars versus brightness in them tell us that the mass spectrum of stars $N(M)$ is not a unique function.

In addition, a few clusters of each type contain 2–20 stars brighter and bluer than should still be there for their ages measured from the majority of stars. These so-called blue stragglers could be the products of merged binary pairs, stars that have somehow managed to mix large amounts of fresh fuel down into their cores, or straying non-cluster-members of lesser age.

Properties of star clusters and their use for age determinations and tests of stellar evolution are discussed in Hesser (1980) and Renzini and Fusi Pecci (1988). The globular clusters in our own and other galaxies are covered by Grindlay and Philip (1988) and the dynamical evolution of globular clusters by Saslaw (1985), Spitzer (1987), and Elson *et al.* (1987). A wealth of information on stellar populations is to be found in Norman *et al.* (1986), Faber (1987), Freeman (1987), and Frogel (1988).

2.3.1 Open Clusters. More than 1000 clusters in the galactic plane have been catalogued, though only about 100 have detailed data on stellar motions, brightnesses, colors, and so forth. Closely related are a couple of hundred known stellar associations, consisting of newly formed stars and probably not gravitationally bound. The inventory of these clusters is nowhere near complete. Star counts range from 100 to 10^4 per cluster and ages from $\lesssim 10^6$ yr (for associations like that in Orion) to (5–7)$\times 10^9$ yr (for NGC 188 and Mel 66). The scarcity of these oldest open clusters and the absence of older ones may mean either that the galactic disk is much younger than the halo or that the older clusters have all been torn apart by tidal encounters with molecular clouds or other massive objects.

Figure 10 is a plot of luminosity versus temperature for the stars of a nearby open cluster (the Pleiades at about 130 pc; notice that the Seven Sisters—or Six from most urban sites—are merely the brightest few of many stars). The cluster contains one white dwarf; stars of $6M_\odot$ have exhausted their hydrogen fuel and are in the process of becoming red giants (telling us that the cluster is about 6×10^7 yr old) and stars of less than about $0.5M_\odot$ have not yet had time to complete their pre-main-sequence contraction and start hydrogen burning. In fact, we can never see the complete main sequence of any one cluster. Stars that differ by more than 10^3 in luminosity have such different time scales that the bright ones are gone before the faint ones begin nuclear fusion.

At least some young clusters and associations appear to be deficient in low-mass stars even when this effect is allowed for. The deficiency is presumably another aspect of the two-mode star formation addressed in Sec. 2.2. In some cases, however (including the nearby 10^9-yr-old Hyades cluster), the low-mass stars have simply drifted further from the center than the bright ones and tend to get missed in counting. The phenomenon is called mass segregation and is established for a handful of clusters.

Low-mass stars will inhabit the cluster outskirts if kinetic energy is shared equally among all stars (smaller masses implying then larger velocities around the average of 1–2 km/s for most clusters). An occasional star will, on the other hand, acquire far more than its fair share of energy through close encounters, and leave the cluster at high speed. This process accounts for some or most of the high-velocity young stars seen in our galaxy. Such runaway stars have also been attributed to the disruption of binary systems by supernova explosions of the primary star. But the runaway population includes some binaries, and, in addition, the exploding star will be the

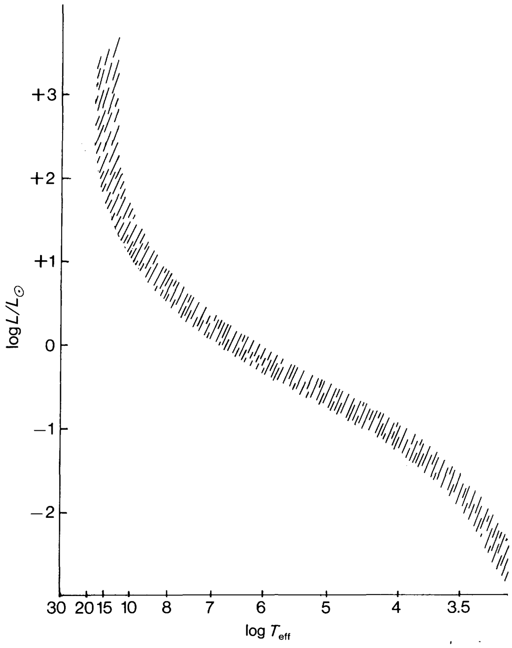

FIG. 10. Plot of brightness [log $(L/L_\odot)$] vs surface temperature (nonlinear scale) for stars in the Pleiades, the brightest 6 or 7 of which are naked-eye stars. Data have been published in a number of different places for about 100 stars, including one white dwarf, whose presence in the cluster indicates that stars of at least $6M_\odot$ (the mass that is just now leaving the main sequence) can shed their outer envelopes down to below the Chandrasekhar mass (Sec. 1.4) rather than becoming supernovae and neutron stars. Observations of stars are reported as brightness within a finite waveband (rather than as bolometric luminosity) and as a logarithmic measure of color temperature (rather than in K). Comparison of theory with data should properly be done in the observed coordinates (converting modeled quantities to observed ones rather than the other way around). This diagram is, therefore, somewhat schematic. Because of the coordinates properly used, such plots are called color–magnitude diagrams, or, in honor of the first people to attempt them, Hertzsprung–Russell or HR diagrams.

less massive member (Sec. 1.5.2) and so unlikely to disrupt the system.

2.3.2 Globular Clusters. Historically globular clusters were important because the distances to them measured by Harlow Shapley in the 1930s demonstrated that the sun is nowhere near the center of the galaxy (Fig. 9). The clusters are strongly concentrated around a point about 8.5 kpc away from us which later work at radio wavelengths established conclusively as the dynamical center of the galaxy.

About 150 globular clusters are known, and another 50 or so may lurk in obscured regions near the galactic center. Other galaxies have similar populations with anywhere from a handful of members (around a dwarf galaxy that in turn orbits the Milky Way) to 10^3 or more (around a large elliptical galaxy, M87, in the Virgo cluster). Some galaxies with very large gas contents may still be forming globular clusters at the present.

Recent studies of globular clusters have focused heavily on

1. determining their ages, as the best lower limit we have to the age of the universe, and
2. following their dynamical evolution.

Everyone would agree that the ages are in the range $(7\text{–}17)\times 10^9$ yr. Unfortunately, this is just the same range of uncertainty as is presented by other approaches, including backward extrapolation of the expansion of the universe (see GALAXIES AND COSMOLOGY). The difficulties in narrowing the range arise from imperfect knowledge of distances, chemical compositions, and physics of gas convection and diffusion inside stars (Renzini and Fusi Pecci, 1988). A related interesting question is the origin of the heavy elements in the globular clusters. For those with the smallest abundances, the entire cluster metal content could have come from only a couple of supernovae, which might well have occurred in the cluster itself and polluted its gas before the stars we now see had condensed. If this had happened, we would expect to see wider variation in heavy element content from one star to another in a single cluster. On the other hand, it is hard to see how to avoid such self-enrichment or to understand precisely where else the heavy elements might have come from.

The dynamical importance of globular clusters arises from their having evolutionary time scales comparable with the age of the universe. The observed distribution of stellar density on the sky through a cluster can be fitted heuristically as

$$\sigma(r)=c\{[1+(r/r_c)^2]^{-1/2} - [1+(r_t/r_c)^2]^{-1/2}\}^2, \qquad (9)$$

where σ is surface density versus radius, r_c is a core radius, c a scaling parameter, and r_t the limiting tidal radius set by the gravitational field of the rest of the galaxy. Typical values of r_c are 1–10 pc and $r_t/r_c \sim 10$. A cluster whose orbit carries it through the galactic plane or close to the galactic center may find r_t approaching r_c, so that it is disrupted. The present cluster population represents the survivors (anything from 90% down to 10% or less in various investigations) and so may not be typical of the entire original population.

No true equilibrium configuration is possible for a star cluster, since a star that acquires the escape velocity leaves, taking its kinetic energy with it, so that the remaining stars become monotonically more tightly bound. Effects of this sort from two-body encounters will become noticeable in a relaxation time given by

$$t_R=\frac{v^3R^3}{24\, G^2m^2N\ln(N/2)}, \qquad (10)$$

where v is the central velocity dispersion (10–30 km/s), R is a characteristic radius (2–10 pc), m is the mass of an average star ($0.8M_\odot$), and N the number of stars in the cluster ($10^{5\pm1}$). For the actual range of cluster parameters, t_R straddles the age of the universe.

On the time scale t_R, one expects the center of the cluster to collapse. Dynamical evolution of a stellar population can be followed using Eq. (11), variously called the Liouville, Vlasov, collisionless Boltzmann, or Fokker–Planck equation, depending on one's background:

$$0=\frac{\partial f}{\partial t}+u\frac{\partial f}{\partial x}+v\frac{\partial f}{\partial y}+w\frac{\partial f}{\partial z}-\frac{\partial v}{\partial x}\frac{\partial f}{\partial u}-\frac{\partial v}{\partial y}\frac{\partial f}{\partial v}-\frac{\partial v}{\partial z}\frac{\partial f}{\partial w}, \qquad (11)$$

where $f(x,y,z,u,v,w,t)$ is the time-dependent density of points in a phase space with three spatial dimensions, x,y,z, and three velocity dimensions, u,v,w. Many assumptions of symmetry and other simplifications are normally made before tackling solutions of the equation. The problem has, in recent years, also become approachable numerically through computer simulations that treat the motions of test particles in a predetermined potential—or, better, the simultaneous interaction of N point masses.

Both analytic and numerical studies confirm the onset of a "gravithermal catastrophe" in which the core collapses, bounces, and then oscillates. Surprisingly, the cluster survives, and [apart from a steeper central density gradient than implied by Eq. (9)] may not even look greatly different from a precollapse cluster.

The details are heavily dependent upon the exchange of energy between the cluster core as a whole and close binary systems that form and get torn apart during the approach to collapse. This interrelation between dynamical evolution and binaries couples to observations in two ways. First, globular clusters (and perhaps other halo populations) are notoriously lacking in confirmed close binaries on or near the main sequence. On the other hand, they are richly endowed with the highly evolved kinds discussed in Sec. 1.5.2 (Trimble in Hesser, 1980). Formation by late capture is an obvious interpretation. Recently, it has transpired that the clusters are rich not only in cataclysmic variables and x-ray binaries with low-mass companions, but also in binary pulsars and in the single (short period, low magnetic field) kind that are probably binary relics. These systems require either that the clusters keep for 10^{10} yr the neutron stars left by their massive stars until they can capture normal companions as the core evolves, or that white dwarfs do the capturing and get driven to collapse by accretion of material from their acquired companions. Each scenario presents difficulties.

2.3.3 Stellar Populations. The traditional dichotomy, laid down by Walter Baade more than 40 years ago, is Population I = young, thin disk, metal rich, associated with gas and dust, brightest stars blue, rotates with the disk (so the stars seem to have low velocities as seen from the solar system), and Population II = old, halo, metal poor, no gas or dust, brightest stars red, little rotation (so the stars seem to have high velocities as seen from the solar system). The observed increase in stellar velocity dispersions with age has also a "real" component, caused by the stars' encountering molecular clouds, star clusters, each other, etc., and so getting scattered into noncircular orbits.

In addition to extreme Pop I and II, the Milky Way clearly contains an old disk population (to which the sun belongs) consisting of stars with ~2% heavy elements, ages of [1–6 (or more)] × 10^9 yr, and velocities that can carry them a couple of hundred parsecs out of the galactic plane. The existence of a thick (1–2 kpc) disk, transitional between the old disk and the halo, and separable in age, kinematics, and metallicity, is hotly debated (Gilmore and Carswell, 1987).

The dichotomy really applies only to stars that come close to the sun, and breaks down in other contexts. Massive elliptical galaxies and the center of our own contain old, metal-rich stars. Irregular galaxies and the extreme outer disk of our own include very young, metal-poor stars. And, if we could look at sufficiently distant galaxies to catch the first generations of star formation, we would expect them all to be dominated by young, massive, metal-poor objects.

The average properties of stars thus vary

1. with time,
2. from one galaxy to another, and
3. from place to place in a single galaxy.

The dominant trends are that metallicity and degree of population evolution increase with time, with total mass of a galaxy, toward the centers of spheroidal populations, and toward the planes of disk ones. The central tasks of stellar population studies are to make these relationships quantitative, to identify and understand deviations from them, and to model the properties (as the sum of successive contributions of stellar generations) in a way that is consistent with what we know about behavior of individual stars (e.g., Mathews, 1988).

GLOSSARY

Algol Variables: Pairs of closely orbiting stars in which one star has become a red giant and transferred enough material to its companion that, although it started out as the more massive, it is now the less massive. Most known systems eclipse. Named for the prototype in Perseus.

Black Hole: The most extreme possible result of gravitational contraction or collapse. The escape velocity is at least the speed of light and gravitational redshift for light attempting to escape is infinite, so that no signals can be propagated to infinity.

Cataclysmic Variable: Close pair of orbiting stars in which one is a white dwarf. Material is accreted erratically from the more normal star and burns explosively from time to time, giving rise to a variety of kinds of rapid brightness changes, including novae and dwarf novae.

Cepheid Variable: A massive star which has completed core hydrogen burning and so evolved to a zone in the temperature–luminosity plane where its outer layers act like a periodic tap, holding and releasing luminous flux, so that the star pulsates in and out, changing its brightness by a factor of four or more over a period of 1–100 days.

Chandrasekhar Limit: The largest mass of cold, dense material in which degenerate electron pressure can balance gravity; about 1.4 times the mass of the sun for most chemical compositions.

Chromosphere: The thin layer at about 10^4 K just above the visible surface (photosphere) of the sun or other stars, named because of its red glow visible during solar eclipses.

CNO Cycle: A set of three coupled, catalyzed cycles of nuclear reactions, in which four hydrogen atoms fuse to form one helium atom with C, N, and O as catalysts. A typical C or N atom may catalyze formation of 100 or more helium atoms in the life of a star. Also referred to as CN cycle and CNO tricycle.

Corona: The thick layer at about 10^6 K above the chromosphere of the sun or other stars. The gas is too thin to be visible except during solar eclipses, but significant x rays are radiated.

Coulomb Barrier: The height, in energy units, of the electromagnetic repulsion between two nuclei when they are separated by the smallest distance allowed by classical mechanics. Quantum mechanical penetration through this barrier is required for nuclear reactions to occur in stars.

Dark Matter: Any gravitating mass which does not radiate roughly its fair share of one solar luminosity per solar mass.

Galactic Halo: The spheroidal component of the galaxy, consisting primarily of old stars. The phrase is occasionally also used to mean a spheroidal distribution of dark matter, probably somewhat more extended in radius than the stars.

Hawking Radiation: A somewhat hypothetical quantum mechanical phenomenon that enables very small black holes to boil away over the age of the universe, though having thermodynamic temperatures related to the sizes of the Schwarzschild horizons.

Hertzsprung–Russell Diagram: Plot of luminosity versus surface temperature for stars. Since these are the two most readily observed stellar properties, such diagrams are useful for comparing observations and theoretical models of stellar structure and evolution. Temperature increases to the left in such diagrams, just to show that astronomers do not think quite the way other people do.

H I Disk: The flat portion of the Milky Way or other spiral galaxies when viewed in the 21-cm radio emission line of neutral hydrogen.

H II Regions: Clouds of ionized gas. Since astronomical gas is everywhere mostly hydrogen, it is the ionization potential of H which dominates the conditions in such clouds.

Jeans Mass: The smallest mass that can contract under its own self-gravitation in a medium of uniform density and temperature.

Kelvin–Helmholtz Time: The thermal time scale on which a star would radiate away its gravitational potential energy or thermal energy content if no nuclear reactions released additional energy.

Magnetic Wind: Gas flowing off the surface of a star with appreciable surface magnetic field, so that the gas is constrained to corotate with the star out to some distance from its surface and so transports angular momentum outward.

Main Sequence: The phase in a star's life when its energy is derived from hydrogen fusion in its core. This phase occupies about 90% of a star's life, about 10^{10} yr for the sun.

Nebula: Any diffuse-appearing astronomical object, including star clusters, galaxies, and ionized gas clouds. Though modern practice tends somewhat to restrict the term to the last category, historical usages, like Andromeda Nebula for a nearby large galaxy, persist.

Neutron Star: A degenerate configuration in which pressure of neutrons balances the force of gravitation. Typical remnant of core collapse in massive stars.

Planetary Nebula: The outer layers of an intermediate-mass star, expelled at the end of the helium-burning phase, leaving the inert core as a white dwarf. The name derives from the appearance in a small telescope, which resembles that of Uranus or Neptune, and has no further significance.

Population I: The stars in the solar neighborhood that are relatively young, rotate with the galactic disk, and contain about 2% of elements heavier than hydrogen and helium.

Population II: The stars in the solar neighborhood that are about 10^{10} yr old, have orbits within the galactic spheroid, and contain less than about 0.2% of elements heavier than hydrogen and helium.

Proton-Proton Chain: The series of nuclear reactions converting four atoms of H to one atom of He, in which the first step is the weak interaction $p+p \rightarrow \mathrm{H}^2+e^+ + \nu_e$.

Pulsar: Neutron star that reveals itself through coherent radio emission beamed into a narrow cone. The pulses appear as the beam sweeps across us with the rotation period of the neutron star.

Quasar: The radio-emitting subset of quasistellar objects (QSOs). The QSOs in general are characterized by broad emission lines at high redshift, compact optical appearance, and total luminosities 100 or more times that of typical galaxies.

***r*-Process:** The rapid addition of neutrons to iron (usually) to produce the neutron-rich isotopes of elements with Z greater than 26. Captures alternate with beta decays; the most likely site is highly processed layers of stars undergoing core-collapse supernovae.

Radio Galaxy: A galaxy (typically, though not always, a giant elliptical) whose luminosity in radio waves exceeds about 10^{43} ergs/s.

Red Giant: A star whose primary source of energy is hydrogen fusion in a thin layer around an inert helium core. This structure results in a distended cool envelope (hence the name), and the phase lasts about 10% of the life of a typical star.

Roche Lobe: The innermost equipotential surface in a binary system that (in the rotating coordinate system) envelops both stars.

RR Lyrae Variable: A low-mass star which has completed core hydrogen burning and evolved to a zone in the temperature-luminosity plane where its outer layers act like a periodic tap, holding and releasing luminous flux, so that the star pulsates in and out, changing its brightness by a factor of two or more over a period of 0.3–1.5 days.

Runaway Stars: Stars whose velocities are sufficiently large to require some explanation other than motion in the general galactic potential. Some are thought to result from disrupted binary systems, others from close encounters with other stars in clusters.

***s*-Process:** The slow addition of neutrons to iron or other elements to produce the most tightly bound isotopes of elements with Z greater than 26. Captures alternate with beta decays; the most likely site is the nuclear-burning zones of highly evolved, intermediate-mass stars.

Schwarzschild Horizon: The spherical (imaginary) surface around a black hole from which the escape velocity $=c$ and gravitational redshift is infinite. Named for Karl Schwarzschild.

Schwarzschild Model: Solution of the coupled equations of stellar structure achieved by mathematical methods codified by Martin Schwarzschild (son of Karl).

Stellar Flares: Sudden releases of energy on the surfaces of stars leading to brightening at optical, radio, and x-ray wavelengths. The basic energy source is thought to be magnetic.

Supernova: The sudden brightening of a star by a factor of more than 10^4 to a peak luminosity of order 10^{44} ergs/s. Total energy release is 10^{51-53} ergs, and the primary energy source is collapse of a dense core of a massive star (Type II supernova) or the explosive burning of about $1M_\odot$ of carbon and oxygen to iron (Type I supernova).

T Tauri Variable: Late phase of the star formation process, in which nuclear reactions have already begun in the core, but the surface appearance of the star is still dominated by winds and other phenomena involving the gas from which it formed.

White Dwarf: A degenerate core of carbon and oxygen (usually; occasionally helium or heavier elements) left when an intermediate-mass star expels its outer layers at the end of nuclear fusion. Gravitation is balanced by the pressure of degenerate electrons.

X-Ray Binary: A pair of closely orbiting stars, one of which is a neutron star or a black hole and which accretes material from its more normal companion. The gravitational well of the compact star is deep enough that the accreted gas inevitably heats to x-ray emitting temperatures.

List of Works Cited

Abbott, D. C., Conti, P. S. (1987), *Annu. Rev. Astron. Astrophys.* **25**, 113.

Arnett, W. D., Bahcall, J. N., Kirshner, R. P., Woosley, S. E. (1990), "Supernova 1987A," *Annu. Rev. Astron. Astrophys.* **28**.

Athanassoula, E. (Ed.) (1983), *Internal Kinematics and Dynamics of Galaxies*, Dordrecht: Reidel (IAU Symposium No. 100).

Audouze, J., Lequeux, J., Levy, M., Vidal-Madjar, A. (Eds.) (1983), *Diffuse Matter in Galaxies*, Dordrecht: Reidel.

Baade, W., Zwicky, F. (1934), *Phys. Rev.* **45**, 138; *Proc. Natl. Acad. Sci. USA* **20**, 254, 259.

Beck, R., Kronberg, P., Wielebinski, R. (Eds.) (1990), *Galactic and Intergalactic Magnetic Fields*, Dordrecht: Kluwer (IAU Symposium No. 140).

Belvedere, G. (Ed.) (1989), *Accretion Disks and Magnetic Fields in Astrophysics*, Dordrecht: Kluwer.

Binney, J., Tremaine, S. (1987), *Galactic Dynamics*, Princeton: Princeton University.

Boss, A. P. (1988), *Comments Astrophys.* **12**, 169.

Bowers, R., Deeming, T. (1984), *Astrophysics I: Stars*, Boston: Jones & Bartlett.

Capuzzo-Docetta, R., Chiosi, C., Di Fazio, A. (Eds.) (1990), *Physical Processes in Fragmentation and Star Formation*, Dordrecht: Kluwer.

Chandrasekhar, S. (1939), *Stellar Structure*, Chicago: University of Chicago; New York: Dover (reprint).

Chiosi, C., Maeder, A. (1986), *Annu. Rev. Astron. Astrophys.* **24**, 329.

Collins II, G. W. (1989), *The Fundamentals of Stellar Astrophysics*, San Francisco: Freeman.

Cox, D. P., Reynolds, R. J. (1987), *Annu. Rev. Astron. Astrophys.* **25**, 303.

Dalgarno, A., Layzer, D. (Eds.) (1987), *Spectroscopy of Astrophysical Plasmas*, Cambridge, England: Cambridge University.

Drechsel, H., Rahe, J., Kondo, Y. (Eds.) (1987), *Cataclysmic Variables*, Dordrecht: Reidel (IAU Colloq. No. 93).

Elson, R., Hut, P., Inagaki, S. (1987), *Annu. Rev. Astron. Astrophys.* **25**, 565.

Faber, S. (Ed.) (1987), *Nearly Normal Galaxies*, Berlin: Springer-Verlag.

Freeman, K. C. (1987), *Annu. Rev. Astron. Astrophys.* **25**, 603.

Frogel, J. A. (1987), *Annu. Rev. Astron. Astrophys.* **26**, 51.

Gilmore, G., Carswell, R. (Eds.) (1987), *The Galaxy*, Dordrecht: Reidel.

Grindlay, J. E., Philip, A. G. C. (Eds.) (1988), *The Harlow Shapley Symposium on Globular Cluster Systems in Galaxies*, Dordrecht: Kluwer (IAU Symposium No. 126).

Hartmann, L. W., Noyes, R. W. (1987), *Annu. Rev. Astron. Astrophys.* **25**, 271.

Hawking, S. W., Israel, W. (Eds.) (1987), *300 Years of Gravitation*, Cambridge, England: Cambridge University.

Helfand, D. J., Huang, H.-H. (Eds.) (1987), *Origin and Evolution of Neutron Stars*, Dordrecht: Reidel (IAU Symposium No. 125).

Hesser, J. (Ed.) (1980), *Star Clusters*, Dordrecht: Reidel (IAU Symposium No. 85).

Hollenbach, D. J., Thronson, H. A. (Eds.) (1987), *Interstellar Processes*, Reidel: Dordrecht.

Kwok, S., Pottasch, S. R. (Eds.) (1987), *Late States of Stellar Evolution*, Dordrecht: Reidel.

Liebert, J., Probst, R. G. (1987), *Annu. Rev. Astron. Astrophys.* **25**, 473.

Lonsdale-Persson, C. J. (Ed.) (1987), *Star Formation in Galaxies*, NASA Conf. Publ. 2466.

Mathews, G. J. (Ed.) (1988), *The Origin and Distribution of the Elements*, Singapore: World Scientific.

Mihalas, D., Binney, J. (1981), *Galactic Astronomy*, 2nd ed., San Francisco: Freeman.

Misner, C. W., Thorne, K. S., Wheeler, J. A. (1973), *Gravitation*, San Francisco: Freeman.

Morris, M. (Ed.) (1989), *The Center of the Galaxy*, Dordrecht: Kluwer (IAU Symposium No. 136).

Norman, C. A., Renzini, A., Tosi, M. (Eds.) (1986), *Stellar Populations*, Cambridge, England: Cambridge University.

Oppenheimer, J. R., Snyder, H. (1939), *Phys. Rev.* **56**, 455–459.

Peimbert, M., Jugaku, J. (Eds.) (1987), *Star Forming Regions*, Dordrecht: Reidel (IAU Symposium No. 115).

Philip, A. G. D., Hayes, D. S., Liebert, J. W. (Eds.) (1987), *Second Conference on Faint Blue Stars*, Schenectady: Davis.

Pringle, J. E., Wade, R. A. (Eds.) (1985), *Interact-*

ing Binaries, Cambridge, England: Cambridge University.

Pudritz, R., Fich, M. (Eds.) (1989), *Galactic and Extragalactic Star Formation*, Dordrecht: Reidel/ Kluwer.

Renzini, A., Fusi Pecci, F. (1988), *Annu. Rev. Astron. Astrophys.* **26**, 199.

Rolfs, C. E., Rodney, W. S. (1988), *Cauldrons in the Cosmos*, Chicago: University of Chicago.

Sahade, J., Kondo, Y., McCluskey, G. (Eds.) (1990), *Close Binaries*, Dordrecht: Kluwer.

Saslaw, W. D. (1985), *Gravitational Physics of Stellar and Galactic Systems*, Cambridge, England: Cambridge University.

Scalo, J. (1986), *Fundam. Cosmic Phys.* **11**, 1.

Schwarzschild, M. (1958), *Structure and Evolution of the Stars*, Princeton: Princeton University; New York: Dover (reprint).

Scheffer, H., Elsaesser, H. (1987), *Physics of the Galaxy and Interstellar Medium*, Berlin: Springer-Verlag.

Shu, F. H., Adams, F. C., Lizano, S. (1987), *Annu. Rev. Astron. Astrophys.* **25**, 23.

Spitzer, L. Jr., (1987), *Dynamical Evolution in Globular Clusters*, Princeton: Princeton University.

Stalio, R., Willson, L. A. (Eds.) (1988), *Pulsation and Mass Loss in Stars*, Dordrecht: Kluwer.

Tenorio-Tagle, G., Moles, M., Melnick, J. (Eds.) (1990), *Structure and Dynamics of the Interstellar Medium*, Berlin: Springer-Verlag (IAU Colloq. No. 120).

Thorne, K. S., Price, R. H., Macdonald, D. A. (1986), *Black Holes: The Membrane Paradigm*, New Haven: Yale University.

Torres-Peimbert, S. (Ed.) (1989), *Planetary Nebulae*, Dordrecht: Kluwer (IAU Symposium No. 131).

Trimble, V. (1988), *Rev. Mod. Phys.* **60**, 859–872.

Weiler, K. W., Sramek, R. A. (1988), *Annu. Rev. Astron. Astrophys.* **26**, 295.

Woosley, S. E., Weaver, T. A. (1986), *Annu. Rev. Astron. Astrophys.* **24**, 205.

Further Reading

Athanassoula, E. (Ed.) (1983), *Internal Kinematics and Dynamics of Galaxies*, Dordrecht: Reidel (IAU Symposium No. 100).

Beck, R., Kronberg, P., Wielebinski, R. (Eds.)(1990), *Galactic and Intergalactic Magnetic Fields*, Dordrecht: Kluwer (IAU Symposium No. 140).

Collins, G. W., II (1989), *The Fundamentals of Stellar Astrophysics*, San Francisco: Freeman.

Faber, S. (Ed.) (1987), *Nearly Normal Galaxies*, New York: Springer.

Gilmore, G., Carswell, R. (Eds.) (1987), *The Galaxy*, Dordrecht: Reidel.

Hesser, J. (Ed.) (1980), *Star Clusters*, Dordrecht: Reidel (IAU Symposium No. 85).

Mathews, G. J. (Ed.) (1988), *The Origin and Distribution of the Elements*, Singapore: World Scientific.

Misner, C. W., Thorne, K. S., Wheeler, J. A. (1973), *Gravitation*, San Francisco: Freeman.

Morris, M. (Ed.) (1989), *The Center of the Galaxy*, Dordrecht: Kluwer (IAU Symposium No. 136).

Norman, C. A., Renzini, A., Tosi, M. (Eds.) (1986), *Stellar Populations*, Cambridge, England: Cambridge University.

Pringle, J. E., Wade, R. A. (Eds.) (1985), *Interacting Binaries*, Cambridge, England: Cambridge University.

Rolfs, C. E., Rodney, W. S. (1988), *Cauldrons in the Cosmos*, Chicago: University of Chicago.

Saslaw, W. D. (1985), *Gravitational Physics of Stellar and Galactic Systems*, Cambridge, England: Cambridge University.

Schwarzschild, M. (1958), *Structure and Evolution of the Stars*, Princeton: Princeton University; New York: Dover (reprint).

Shu, F. H., Adams, F. C., Lizano, S. (1987), *Annu. Rev. Astron. Astrophys.* **25**, 23.

Trimble, V. (1988), *Rev. Mod. Phys.* **60**, 859–872.

ATMOSPHERIC ACOUSTICS

Henry E. Bass, *Physical Acoustics Research Group, University of Mississippi, University, Mississippi, U.S.A.*

Introduction 145
1. **Sound Sources** 147
1.1 Aircraft Noise 147
1.2 Traffic Noise 149
1.3 Explosions 150
1.4 Supersonic Projectiles 151
1.5 Sounds in Nature 151
2. **Propagation** 154
2.1 Absorption 154
2.2 Effects of the Ground Surface 156
2.3 Refraction and Diffraction 160
2.4 Effects of Topography 162
2.5 Effects of Turbulence 164
2.6 Nonlinear Effects 169
2.7 Infrasound Propagation 169
2.8 Numerical Techniques 170
3. **Applications** 172
3.1 Acoustic Sensing of the Atmosphere 173
3.2 Perception of Sound 176
Glossary 178
Works Cited 179
Further Reading 179

INTRODUCTION

The science of acoustics encompasses the production, transmission, and reception of sound. When applied to the atmosphere each of these fundamental aspects of acoustics is affected to some extent by the properties of the atmosphere itself. Atmospheric properties vary in both space and time giving rise to a close interaction between atmospheric acoustics and the broader field of atmospheric physics. The interaction of sound with this complex atmosphere as well as the surface of the earth requires that a wide spectrum of physical phenomena be understood to describe completely a sound field in the atmosphere.

The need to communicate audibly in the atmosphere has traditionally motivated studies of atmospheric acoustics. The physical phenomena which affect atmospheric acoustics were generally identified in the 1800s although many of the details and mathematical techniques for dealing with these phenomena have evolved more recently. An excellent example of early experimental work in atmospheric acoustics was that of Tyndall using the large horn shown in Fig. 1.

The large size of the horn results from the relatively low density [1.29 kg/m^3 at standard temperature and pressure (STP): 273 K, pressure of 1 atm with 0% relative humidity] and speed of sound in air (331.6 m/sec at STP with 0% relative humidity). More recently, studies of atmospheric acoustics have been motivated by concerns for environmental noise. Most activities associated with modern society result in some noise but perhaps none is more pervasive than that associated with transportation. Noise in and around airports and along highways has led to attempts to increase transmission loss as well as reduce source generation. A detailed treatment of many of the topics addressed in this article can be found in Pierce (1981).

A variety of natural processes give rise to atmospheric sound. Some of the more common sources are thunder associated with lightning, noise generated by wind, and sounds produced by animals of all types.

Sound is often classified in terms of amplitude and frequency. Amplitudes of acoustic waves are generally given in dB ref 20 micropascal (μPa). A very quiet anechoic chamber might have a sound level of 0 dB. Outdoors on a quiet day, the sound level is typically 40 dB. Above 90 dB, audible sound begins to result

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

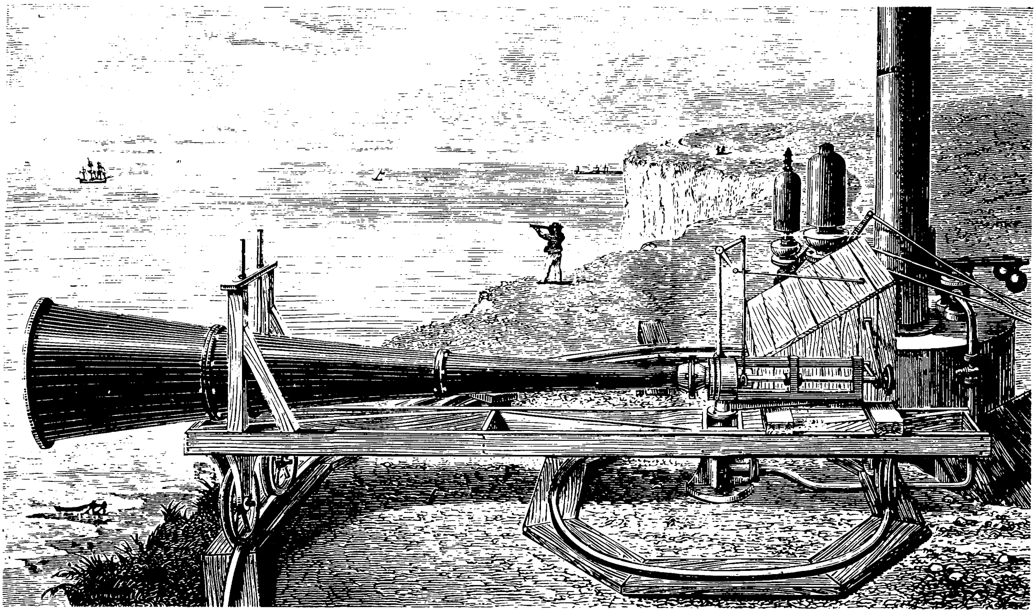

FIG. 1. Tyndall's apparatus for studying sound propagation in the atmosphere (from Tyndall, 1874).

in hearing loss in the case of very long exposures. A jet aircraft taking off might produce 120 dB. Above 130 dB, acoustic waves begin to exhibit nonlinear behavior. Above 140 dB, immediate and permanent hearing loss occurs.

Another way to categorize sound is according to frequency. Acoustic waves with a frequency above 20 Hz (units of $\sec^{-1}$) but below 20 kHz are referred to as audible sound. Lower frequencies are classified as infrasound; higher frequencies as ultrasound. At low frequencies acoustic waves couple into the buoyant restoring force of gravity in the atmosphere and two modes result (Brown, 1987). At very low frequencies ($\sim 10^{-2}$ Hz) the propagation is dominated by the buoyant force (internal gravity waves). At the opposite end of the frequency spectrum, ultrasonic waves are rapidly attenuated in the atmosphere with an attenuation coefficient approximately proportional to frequency squared. As a result, ultrasound does not travel over great distances although over short distances it can be important (echolocation used by bats and echoranging by cameras, for example).

Once sound is generated, there are a variety of physical processes which affect transmission. The most obvious of these is spreading of the wave. The amplitude of a spherically spreading wave decreases as $1/r$ (6 dB per doubling of distance). Even highly directional sources (such as a bat) begin to show this type of spreading some distance from the source. For this reason, 6 dB/doubling of distance is generally taken as a standard loss and all other factors which affect the received amplitude are referred to in terms of excess attenuation (attenuation in addition to 6 dB/doubling of distance).

Physical processes which contribute to excess attenuation are molecular absorption, interaction with the ground, refraction due to temperature and wind gradients, diffraction due to those gradients or topography, and scattering. In all cases, except for the ground interaction (and even there to some degree), the effects depend strongly on specific local atmospheric conditions. This strong dependence of propagation on meteorological condition can be used as a basis to sense the atmosphere remotely using echo sounders or radio sounders.

Application of atmospheric acoustics usually defines the receiver. The detection and location of aircraft might involve an array of microphones sensitive to phase changes between individual elements. Noise control applications typically consider a human ear as the receiver. These applications plus the atmospheric sounders already mentioned continue to provide motivation for studies of atmospheric acoustics.

1. SOUND SOURCES

A number of sources of atmospheric sound were mentioned in the Introduction. In this section, specific sources will be considered in more detail to illustrate the different phenomena which give rise to sound and the atmospheric properties which affect the sound generated. First, man-made sources will be considered, then natural sources.

1.1 Aircraft Noise

Jet aircraft represent a pervasive noise source. Aircraft noise is generated by both propulsive and nonpropulsive sources (Smith and Schien, 1989). Propulsion sources normally result in the dominant noise. Propulsion sources include turbomachinery noise generated internal to the engine and jet exhaust mixing and shock noise attributed to processes external to the engine. Nonpropulsive noise is generated in flight by the air flowing past the vehicle structure and is called airframe noise.

Turbomachinery noise sources include fan noise, compressor noise, combustion noise, and turbine noise.

Fan noise is caused by the unsteady component of the flow which interacts with rotor or stator blade rows. When a perturbation in flow pressure or velocity encounters a blade or strut surface, an acoustic pressure pulse is generated. Primary fan noise generation mechanisms are

1. inlet flow inhomogeneities interacting with the rotor blades, and
2. rotor wakes and pressure fields interacting with the stator blades.

Sound is generated primarily at the blade passing frequency and its harmonics, although there is a broadband component as well.

Key parameters affecting fan noise generation are the number of rotor and stator blades, rotor-stator axial spacing, blade loading, fan rpm, stage pressure ratio, and rotor tip speeds. Upstream stators (inlet guide vanes) increase acoustic levels. At high engine power settings, supersonic tip speeds give rise to tones at harmonics of the shaft rotation speed, which are known as multiple pure tones.

Compressor noise is similar in nature to fan noise, and has the same key parameters as the fan, applied to the compressor stages. Compressor noise is usually of higher frequency than fan noise and is attenuated as it passes through adjacent blade rows. Similar in nature to fan and compressor noise, turbine noise is generated downstream of the combustor and radiates from the core exhaust duct.

Instabilities in combustors give rise to sound waves. These instabilities are usually a result of flow turbulence, so that combustion noise is broadband in nature, normally peaking at 400–500 Hz. Key engine operating parameters affecting combustion noise generation are engine mass flow rate, compressor pressure and temperature ratio, and combustor temperature rise.

Jet exhaust noise is generated by two mechanisms in the jet plumes aft of the exhaust nozzles: turbulent mixing and shock cell perturbations.

Jet mixing noise is generated by the turbulent mixing process between the exhaust flow and the atmosphere. Noise generation is strongest in the turbulent shear layer between the laminar core and the ambient air. This source correlates most strongly with the velocity of the jet at the plane of the exhaust nozzle, varying as approximately the eighth power of the velocity. Secondary parameters are the jet flow density and exhaust nozzle area. Forward motion of the aircraft has a strong attenuation effect on jet mixing noise, because it reduces the magnitude of the shear layer velocity gradient.

Jet mixing noise is spectrally broadband in nature, usually peaking at frequencies below 1000 Hz, depending on the nozzle diameter and jet velocity. It is most intense in the downstream direction and peaks at a radiation angle of approximately $\theta = 140°$, measured from the aircraft nose (i.e., tail-to-nose vector is $\theta = 0°$).

The jet mixing noise generated by air vehicles and high-performance military aircraft is a significant source of noise. This source is difficult to suppress, since it is generated entirely outside the aircraft. Most successful reductions in jet noise have been due to reductions in jet velocity, which generally requires an increase in fan bypass ratio to maintain thrust. Modern commercial aircraft (referred to as Stage 3 aircraft) have large bypass ratio engines. Such engines are visually greater in diameter.

When an air vehicle's convergent exhaust nozzles are operated supersonically, a quasi-periodic shock cell structure will form in the jet. Shock cell noise is produced by the unsteady interaction between the flow turbulence and this structure. The intensity of shock noise is a function of the Mach number (ratio of speed to total speed of sound) of the fully expanded exhaust flow.

Shock-associated noise possesses characteristics distinct from those of turbulent mixing noise. In contrast to the mixing noise radiation pattern, broadband shock cell noise will be most prominent in the forward arc ($\theta < 90°$) directions for a static engine, due to the high jet mixing noise levels. However, for high flight speed, shock noise may dominate the signal in the forward arc and at all radiation angles.

Airframe or nonpropulsive noise is generated by the airflow over the wings, fuselage, vertical and horizontal stabilizers, slats, gears, or any element of an aircraft. The interaction of aircraft elements with large scale turbulence in the atmosphere results in large fluctuations in lift and drag. These forces and the subsequent radiation occur with any lifting element of the aircraft. The sound power radiated from this source of noise increases with the size of the lifting surface, the turbulence intensity, and the fourth power of the aircraft speed. This source of noise has been considered as insignificant for low flight speeds but could be a significant contributor to the aircraft signature when the aircraft speed and the turbulence intensity are high. For this case the radiated power would be high and the frequencies would be low resulting in little excess attenuation due to propagation through the atmosphere.

There are four components of noise associated with radiation from the turbulent boundary layer. These include the following:

1. a surface distribution of dipoles associated with the fluctuating pressures within the flow causing fluctuating forces on a rigid surface;
2. a volume distribution of quadrupoles generated by the fluctuating forces in the boundary layer;
3. reflected quadrupole noise due to the presence of the surface;
4. noise radiation resulting from the motion of the surface beneath the turbulent layer.

Current thinking, based on commercial aircraft experience at low Mach number, considers only the surface distribution of dipoles. For such a distribution, the sound pressure is taken to vary as AU^6 (Reed, 1977) where U is the flow velocity and A is the radiating area. At low Mach numbers this source is less than that predicted for the trailing edge.

Although the surface distribution may be a valid model throughout the subsonic Mach number range there is no evidence to show that dipole-type radiation will be dominant at high subsonic Mach numbers. That is, quadrupole-type sources may become the dominant source at high subsonic Mach numbers. This has not been satisfactorily answered at this time.

The acoustic radiation from structures excited by turbulent boundary layers has been investigated primarily in terms of the noise levels in aircraft cabins and noise radiation underwater. There has been little emphasis on this source of sound with respect to airframe noise. There is something of a conflict as to its importance at low Mach numbers.

The noise radiated directly from wakes and vortices is quadrupole in nature and, as such, is not a significant source. However, when wakes or vortices impinge on a structure or rotating blade, significant levels of noise can be generated. These sources are similar to whole body noise, but at higher frequencies because of the smaller scales.

Trailing edge noise is generated by turbulence being swept over the trailing edge with the creation of half-baffled dipoles. The source of the turbulence is the turbulent boundary layer. There is a large data base at low Mach number ($M < 0.4$) which appears to confirm that trailing edge noise is proportional to the aircraft velocity to the fifth or sixth power.

Rotary wing and propeller-driven aircraft generate sound quite different from that of jet aircraft even though many of the noise generation mechanisms are the same. Machinery noise, for example, is common to both. The biggest difference is the role that the helicopter blade or aircraft propeller plays. In many cases, for these types of aircraft, this blade noise is dominant and characterizes the audible sound.

Four mechanisms are important for noise generation provided the blade tip speed does not exceed the speed of sound. One is the discrete loading noise which dominates the lower harmonics of the blade passage frequency. Next is the blade-vortex interaction which is a discrete source of impulsive character. At higher frequencies, broadband noise due to blade interactions with turbulence in and around the rotor wake begin to dominate. Broadband self-noise due to blade interaction with boundary-layer and near wake turbulence controls the high-frequency part of the spectrum.

The relative importance of each of the above mechanisms depends upon aircraft operating conditions. In most cases, however, the aircraft is characterized by sound associated with the blade passage frequency. This frequency is variable for a propeller-driven aircraft depending upon flight operations. Rotary wing aircraft tend to operate at a lower, fixed, blade passage frequency. A military UH-1D helicopter main rotor rotates at 324 rpm and has twin blades, which gives a fundamental frequency of 10.8 Hz. The tail rotor rotates more rapidly producing a fundamental of 60.5 Hz at a level much lower than for the main blade. Since the human ear is not sensitive to 10.8 Hz, audible sound is due to harmonics of the fundamental blade passage frequency.

1.2 Traffic Noise

Another pervasive source of man-made noise is that due to traffic (Dept. of the Army, 1978). Traffic noise results primarily from the engine and exhaust system and interaction of tires with the road surface. The maximum noise emitted by an automobile increases approximately with the third power of vehicle speed. This is due primarily to tire noise created by the tire-roadway interaction. (Figure 2 illustrates automobile noise spectra at different speeds.)

The noise output of trucks is a more complicated phenomenon. First, trucks should be considered in three distinct classes according to their noise emission characteristics: light, medium, and heavy. *Light trucks* are two-axle, four-wheel vehicles such as panel and pickup trucks; their noise characteristics are similar to those of automobiles. *Medium trucks* are typically gasoline-powered two-axle, six-wheel vehicles, such as city trucks without a vertical exhaust muffler. The noise generation characteristics of these vehicles are also similar to those of automobiles. However, medium trucks are usually 10 dB noisier than automobiles for the same flow and speed.

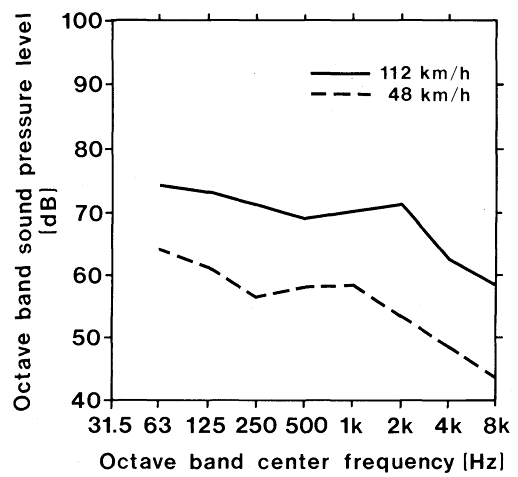

FIG. 2. Typical automobile spectra for two average speeds.

Heavy trucks are a more complex noise source. These diesel-powered, three- or more-axle vehicles have a multitude of noise mechanisms, i.e., tire noise, exhaust noise, intake noise, engine noise, and gear noise. Shown in Fig. 3 is a typical truck noise spectrum for the three major component sources: tires, engine, and exhaust. Tire-roadway interaction, the major noise source for automobiles and light and medium trucks, occurs at ground level. For heavy trucks an additional noise source, the exhaust stack opening, is nominally located eight feet above the ground. Heavy truck noise does not exhibit great variability. While tire noise varies with speed, the engine noise sources generally show little dependence upon road speed. Furthermore, drivers tend to maintain relatively constant engine speed for all road speeds.

As previously discussed, roadway noise exposure is a function of the traffic flow and pa-

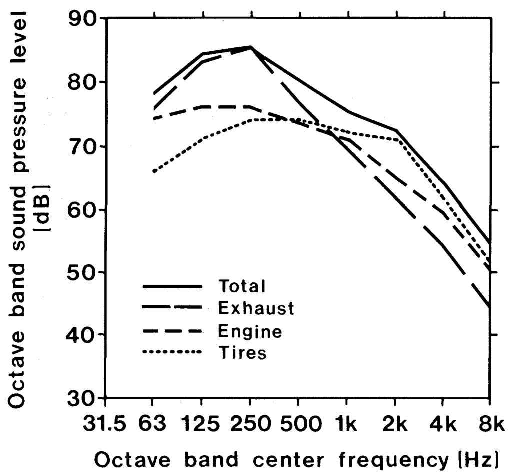

FIG. 3. Typical spectra for diesel truck and component sources.

rameters of the classes of vehicles using the roadway. Where vehicles are uniformly distributed along a single-lane roadway that is straight, infinitely long, flat, level terrain, the noise exposure is a function of the volume flow and average speed of each group of vehicles.

In practice, roadway factors will affect noise levels. Noise exposure is *increased* by uphill grades (for heavy trucks and transport vehicles). The noise exposure is decreased by buildings, land forms, or other barriers located between the roadway and the observer.

The drop-off of noise levels with distance from a roadway will typically range from 4 to 5 dB per doubling of distance. This drop-off rate is affected substantially by ground cover. Beyond two to three thousand feet, the drop-off can increase to about 6 dB per doubling of distance, due to the additional effect of atmospheric attenuation. However, the noise levels from roadway traffic will rarely be high enough to be of concern at these larger distances.

1.3 Explosions

Less common, though more dramatic, sources of sound are impulsive. Man-made explosions ranging in amplitude from nuclear blasts to the blast from a hunter's rifle can provide very large peak sound pressure levels. The variation of pressure with time from an explosive source can be approximated by the wave form shown in Fig. 4.

Near the explosion, the wave form is not so fully developed. In the far field, the most noticeable features of the impulse are the rise time, peak overpressure, and wave duration. The overpressure is given approximately as (Reed, 1977)

$$\Delta p = 0.0469 W^{b/3} R^{b}, \qquad (1)$$

where the value 0.0469 of the coefficient for yield W is given in kg of high explosives, the distance R in km, and Δp in kPa. Values for b are determined empirically to be 1.1–1.2. The wave duration determines the fundamental frequency $\nu_f = 1/\tau$ where τ is the duration in sec. The duration for a 0.57-kg charge of C-4 is typically 0.025 sec (at 100 m). Assuming spherical spreading, ν_f should be proportional to $W^{-1/3}$. Application of these scaling laws to atmospheric conditions other than

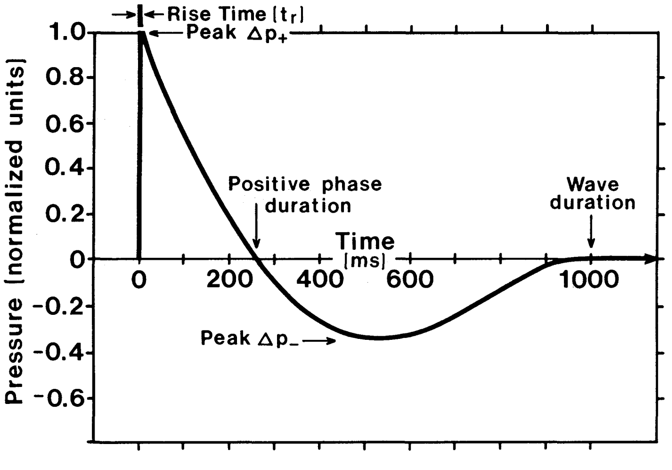

FIG. 4. Explosion-wave pressure/time signature [Reed (1977)].

STP can be accomplished, assuming the charge does *PdV* work against an ideal gas atmosphere.

1.4 Supersonic Projectiles

When an object is traveling at a speed v greater than the local speed of sound c, a wake forms behind the object. Since the object travels a distance vt while pressure waves travel a distance ct, a Mach cone develops ahead of which there is no sound and behind which the pressure rises very rapidly. The angle between the edge of the Mach cone and the projectile flight path is given as θ_M where $\cos\theta_M = c/v$. As the Mach cone passes over an observer on the ground, a sonic boom is heard.

The duration of the shock wave generated by supersonic projectiles depends upon the size of the projectile and the distance to the observation point. Near a projectile with simple geometry the duration $\tau \cong -l/v$ where l is the projectile length. As described later, propagation effects will cause the shock wave to increase in duration as it propagates.

1.5 Sounds in Nature

Natural processes provide a wide variety of atmospheric sounds (Brown, 1987). Referring to Fig. 5, we note that in the infrasound regime, earthquakes, large explosions (volcanoes, for example), and weather fronts give rise to very low-frequency atmospheric waves. Since these low-frequency waves are attenuated very little, they propagate great distances. It is not unusual to detect infrasound from these events after full circuits around the earth.

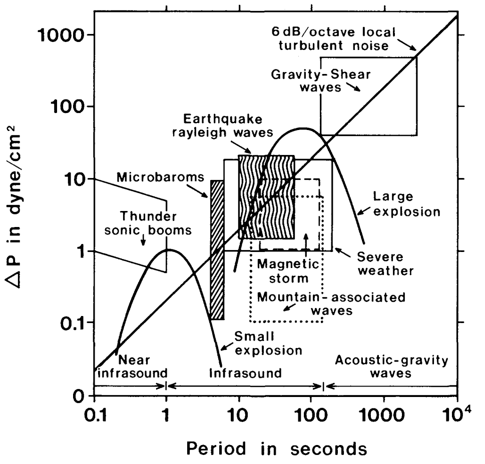

FIG. 5. Regions for infrasound sources. (Courtesy of A. J. Bedard.)

Earthquakes and other strong infrasound sources have the strange properties that the initial part of the received wave can arrive as much as a half hour before the main feature. The explanation for this effect is that high-speed Rayleigh waves traveling through the surface of the earth at a speed much greater than the speed of sound in air constantly leak energy into the atmosphere which can be detected long before the infrasound wave propagation through the atmosphere. It is frequently reported that birds and animals are startled by such arrivals, giving a warning of the upcoming infrasound.

Wind generates noise as it blows across leaves, wires, etc. The primary mechanism for the sound generation is turbulent vortex shedding. In addition to acoustic noise generated at one region and propagating to a receiver, a receiver such as a microphone or the ear responds to pressure fluctuations carried by wind. These fluctuations are not propagating but, since they result in variations in pressure as they flow past the receiver, the receiver cannot distinguish them from propagating waves. A receiver placed in the flow can also give rise to turbulent eddies providing additional noise.

Since wind noise results from the flow of turbulence, one would expect the amplitude to increase as U^2 where U is the mean flow velocity. In addition, wind noise decreases in amplitude as the frequency increases as $\nu^{-5/3}$ for moderate wind speeds. At very large wind speeds, the distribution of turbulence (discussed in more detail later) dictates fewer turbules available to flow by the wind screen resulting in a leveling off in amplitude with frequency. This effect is most pronounced at low frequencies.

An extreme case of wind noise is the sound of tornadoes (Arnold *et al.*, 1976). A model of a severe-traveling-storm cell chosen for discussion purposes is shown in Fig. 6.

Rotation of the air masses is chosen to be cyclonic in the model. The sources of sound S_i located at the coordinate (X_i, Y_i, Z_i) considered are presented in Fig. 6 and may be described as

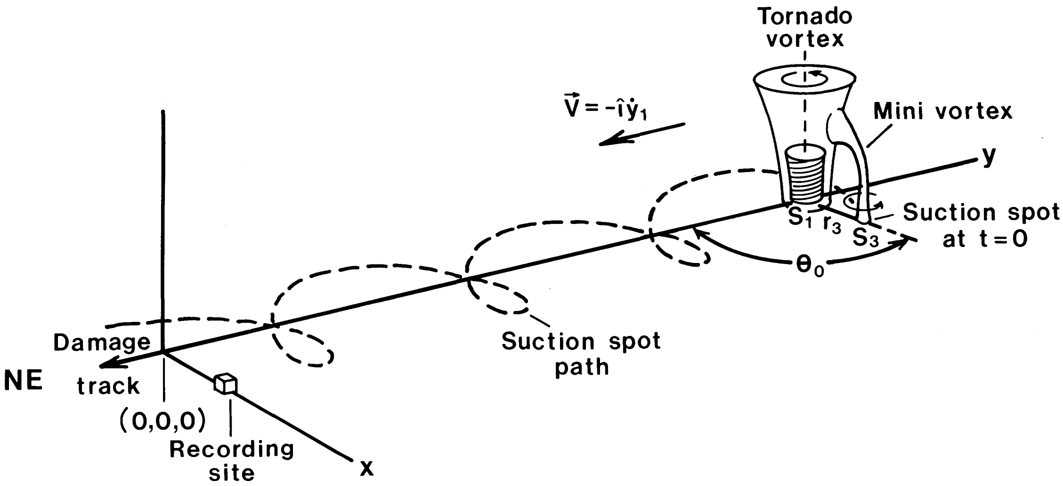

FIG. 6. Geometry assumed for the track of the Guin, Alabama, tornado and a possible minivortex. The terrain has been assumed flat although there were variations in elevation of approximately +15–20 m from the elevation at the recording site.

1. the wind-surface noise S_1 generated by interaction of the high-speed tornado winds with the ground,
2. atmospheric noise associated with the tornado vortex,
3. wind-surface noise S_3 generated by the suction spot of a minivortex,
4. atmospheric noise associated with a minivortex,
5. atmospheric noise associated with the parent storm, and
6. general nonlocalized atmospheric noises associated with the parent storm(s).

We define I_i as the acoustic power from the ith source and I_0 to be the total acoustic power (expressed in dB) when the noise (integrated over frequency) is a maximum. Maximum power probably occurs when the sound source S_1 (wind-surface noise generated by the tornado vortex) is located at or near the origin (0,0,0), and therefore the source S_1 is chosen to be located at the origin when the maximum (integrated over frequency) power is recorded at the recording site (located at position r_0,0,0).

For discussion purposes, only two principal sound sources S_1 and S_3 (which represent wind-surface interactions for the tornado vortex and suction spot, respectively) need to be considered. When the storm is most distant, the sources can be treated as spherical or point sources of sound.

One tornado studied in detail was the one which destroyed Guin, Alabama in 1974. There were no good photographic records for the Guin storm, so the distance of the suction spot from the center of the tornado vortex was assumed to be 0.6 km, and the small vortex was assumed to be moving in a circular path around the rotation axis of the tornado with a period of 60 sec. This means that the rotational (linear) speed of the suction spot was 240 km/h (150 miles/h). Combining this with the translational speed [assumed to be 96 km/h (60 miles/h)] indicates that the suction spot would have maximum wind speeds towards the recording site of 340 km/h and minimum speeds away from the receiver of 140 km/h. The high speeds involved here immediately suggest that Doppler effects should be included. For the present, however, assume that the tornado was a white-noise source. In this case, the distance from the suction spot to the receiver is given by

$$R_3=\{[6.4-26t-640\cos(\pi t/30)]^2 + [160-640\sin(\pi t/30)]^2\}^{1/2}. \qquad (2)$$

Now consider contributions at the receiver from both sources S_1 and S_3. It should be possible to combine noise from both sources and achieve a reasonable account for the qualitative and quantitative variations in the acoustic power level as the tornado approached and receded. For convenience we have chosen the level generated by each to be equal. Figure 7 presents a comparison of the relative acoustic power data and the relative power at the recording site if noise from the spherical (or point) sources S_1 and S_2 are combined. The agreement between the data and the calculated relative power is better than considering the sources separately.

Audibly, the tornado sounds like low-frequency noise (not unlike a freight train or jet aircraft landing) modulated by the rotation of the suction spot.

Thunder represents an excellent example of natural impulsive sound. As lightning discharges, the gas along the discharge channel rapidly heats and expands. This expansion propagates outward, away from the channel. To a good approximation, the frequency of maximum emission for a lightning discharge is given by (Bass, 1980)

$$\nu_n = c/[2.6(E_l/P_0\pi)^{1/2}], \tag{3}$$

where E_l is the energy per unit length dissipated in the discharge (typically 3 kJ/m) and P_0 is the ambient pressure.

Near the discharge channel, the pressure wave is a cylindrically spreading, highly nonlinear wave. For a truly cylindrical wave, all observers would hear a single clap associated with the point along the discharge path nearest to them. Real lightning, however, is very tortuous. This tortuosity gives rise to a "string of pearls" effect where thunder can be modeled as a long line of spherically spreading sources. As the thunder propagates, nonlinearity and absorption shift the spectrum to low frequencies. These combined effects give rise to the extended low-frequency rumbles associated with distant lightning and explain the sharp crack heard when lightning strikes nearby.

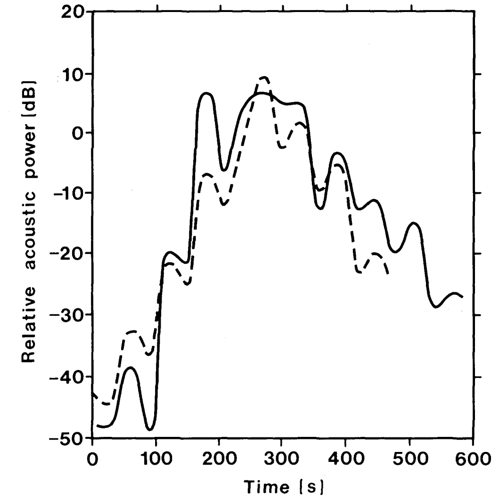

FIG. 7. Temporal variations in the relative acoustic power at 350 Hz for the Guin, Alabama, tornado compared to the variation expected from point sources S_1 and S_3. Solid line represents experimental data corrected for automatic volume control and the dashed line is predicted.

The animal world produces diverse sounds for communications (Busnel, 1977). Many animals use these sounds for warning and mating signals. The range of emission organs is quite varied and are found on all parts of the body. Invertebrates may use a toothed file to stimulate a vibrating element, vibration of nonspecialized wings (mosquitoes), semiresonant plates activated by muscle contractions, or reedlike organs activated by aspiration and respiration of air. In higher vertebrates specialized organs generate sound using forced or aspirated air. The frequency of the sound generated is controlled by varying the geometry or tension on the organs. At this point in time, sound generation mechanisms are well understood but there has been very little research which carefully characterizes total sound energy emitted. The recent discovery of communications between elephants via infrasound further emphasizes our general lack of knowledge.

Whether an animal equipped with sound emission organs will or will not activate them depends upon a number of factors such as temperature, humidity, light, and hormonal state. When sound is emitted, it follows a code involving rhythm, pulse, repetition rate, amplitude variation, and frequency. In lower animals, many of these quantities are controlled by the physical properties of the sound-producing mechanisms. In higher-order species, the code is more controlled by the central nervous system.

Messages broadcast by animals contain two specific elements of information. The first indicates presence and in a general sense, location. Hierarchical status is included for gregarious species. The second element of the message is motivation. It might also contain information giving location of an object, a territory, or a predator. Vocabularies in the animal kingdom vary widely in their complexity.

Perhaps the most interesting application of acoustics by animals is the sonar used by bats to capture flying prey. The sonar operates in two modes, search and home. In the search mode, the 35-kHz signal propagates reasonably well and has a beam width sufficiently wide to cover a large volume. Once an echo indicates a potential meal, the bat switches to a high frequency chirp (40–90 kHz). The higher frequencies provide better spatial resolution, and as the bat approaches the target, the added atmospheric absorption characteristic of high frequencies is less important. Spatial resolution automatically improves as the range to target decreases.

2. PROPAGATION

2.1 Absorption

Atmospheric absorption occurs by two basic mechanisms (Bass *et al.*, 1984):

1. classical losses associated with the transfer of acoustic energy (or the organized systematic motion associated with the kinetic energy of the molecules) into equivalent heat energy (uncoordinated random motion associated with thermal agitation) by a redistribution of translational energy among the molecules; and
2. relaxation losses associated with the redistribution (relaxation) of translational or internal energy of the molecules.

A complete calculation of sound absorption would necessarily include not only the contribution of each mechanism singly but also their interdependence; e.g., the effect of molecular relaxation processes on the classical loss mechanisms. Fortunately, it has been demonstrated by successive approximations to a solution to the Boltzmann equation and measurements at low pressures that, for frequencies below 10 MHz, absorption due to classical losses and molecular relaxation are additive. The effects of absorption will be represented by an absorption coefficient $\alpha(m^{-1})$ such that the acoustic pressure $P = P_0 e^{-ar}$. Often, the absorption coefficient is given with units of neper/m. The neper is not a unit in the normal sense but, rather, specifies that it is associated with the Naperian base *e*. The component of absorption due to viscous effects is

$$\alpha = (\omega^2/2\rho_0 c^3)(4\mu/3), \tag{4}$$

where ω is the angular frequency ($2\pi f$) in s^{-1}, c is the speed of sound in m/s, ρ_0 is the density in kg m^{-3}, and μ is the coefficient of viscosity in kg m^{-1} s^{-1}.

Energy loss from the sound wave can be separated into two distinct processes. First, since the sound wave causes a sinusoidal variation in temperature in space, heat will be conducted from the regions of high temperature to those of lower temperature. This heat conduction process converts the energy of the sound wave into random thermal motion of the gas molecules. Heat conduction increases the absorption to

$$\alpha = (\omega^2/2\rho_0 c^3)[4\mu/3 + (\gamma-1)\kappa/(\gamma c_v)], \tag{5}$$

in m^{-1}, where κ is the coefficient of thermal conductivity in J (kg mol)$^{-1}$ K^{-1} kg m^{-1} s^{-1}, and c_v is the specific heat at constant volume (J/K).

In addition to energy loss due to heat conduction, energy can also be lost due to relaxation of internal degrees of freedom. As the local temperature increases, the energy of internal modes also increases. If the temperature changes very slowly, the internal energy changes by an amount equal to $c_{\text{int}}\Delta T$, where c_{int} is the specific heat of the internal energy modes (e.g., due to rotational and vibrational processes), and ΔT is the local difference between the ambient temperature and the temperature in the presence of the sound wave. For low frequencies, any energy which goes into the internal modes while the local temperature is at a maximum will be diminished as soon as the local temperature begins to decrease. However, this transfer of energy from the translational energy of the gas to internal energy modes and back can only take place during collisions between the gas molecules. During any one collision, there is the probability of energy being transferred which depends on the particular internal mode of concern. So, as the frequency is increased, the process of transferring energy back and forth between internal energy modes and translation introduces a time lag between the time needed for the local temperature to decrease and the time needed for the internal energy modes to give energy back to the translational modes. This time delay results in a decrease in the energy of the sound—an attenuation. For a particular energy mode, at very low frequencies, no absorption will result. As the frequency is increased to a value referred to as the relaxation frequency f_r of the mode (or alternatively as the frequency of maximum absorption), the absorption increases. Then, as the variations in the local temperature become so rapid that the internal mode is always behind in transferring energy, the relaxation absorption becomes constant. For any particular relaxation process, the absorption can be expressed by a relation having the form

$$\alpha=\frac{\pi s_v}{c}\,\frac{f^2/f_r}{1+(f/f_r)^2} \tag{6}$$

in Np m^{-1}, where s_v is the relaxation strength in nepers, which depends on the specific heat of the relaxing mode.

For air, three relaxation processes are important; those associated with O_2 vibration, N_2 vibration, and rotation. Rotational relaxation occurs at very high frequencies so that f/f_r in the denominator of Eq. (11) can be ignored below 10 MHz at 1 atm of pressure (limit decreases in proportion to the local pressure). In this case, rotational relaxation has the same form as classical absorption and the two can be combined into a single term. Using atmospheric values for density, viscosity, and the speed of sound, the combined term is

$$\alpha_{\mathrm{cr}}=1.83\times 10^{-11}\,\frac{(T/T_0)^{1/2}f^2}{(P/P_0)}, \tag{7}$$

where $T_0=293.15$ K and $P_0=1$ atm.

Oxygen and nitrogen relaxation depend strongly on the mole fraction of water vapor. In terms of the relative humidity h_r, in percent, the mole fraction, h_f in percent, can be written as

$$h_f=h_r(P_{\mathrm{sat}}/P_0). \tag{8}$$

The ratio of saturated vapor pressure to atmospheric pressure (P_{sat}/P_0) adopted by the World Meterological Organization is

$$\begin{aligned}\log_{10}(P_{\mathrm{sat}}/P_0)=&10.795\,86[1-(T_{01}/T)]-5.028\,08\log_{10}(T/T_{01})\\&+1.504\,74\times 10^{-4}\{1-10^{-8.296\,92}[(T/T_{01})-1]\}\\&+0.428\,73\times 10^{-3}\{10^{4.769\,55}[1-(T_{01}/T)]-1\}-2.219\,598\,3,\end{aligned} \tag{9}$$

where T is the air temperature in kelvins, and T_{01} is the triple-point isotherm temperature with the exact, internationally agreed value of 273.16 K.

The values for the relaxation frequencies of oxygen $(f_{r,\mathrm{O}})$ and nitrogen $(f_{r,\mathrm{N}})$ can be extracted from the experimental data. The best available values for these quantities are

$$(f_{r,\mathrm{N}})=(P/P_0)(9+200h_f) \tag{10}$$

and

$$(f_{r,\mathrm{O}})=(P/P_0)\{24+4.41\times 10^4 h_f[(0.05+h_f)/(0.391+h_f)]\}. \tag{11}$$

With these relaxation frequencies and assuming that dry air is composed of 20.9% O_2 and 78.1% N_2, the total absorption becomes

$$\begin{aligned}\alpha=f^2\Bigg[&1.83\times 10^{-11}(P_0/P)(T/T_0)^{1/2}\\&+(T_0/T)^{5/2}\left(1.278\times 10^{-2}\,\frac{e^{-2239.1/T}}{f_{r,\mathrm{O}}+(f^2/f_{r,\mathrm{O}})}\right)\\&\times\left(1.069\times 10^{-1}\,\frac{e^{-3352/T}}{f_{r,\mathrm{N}}+(f^2/f_{r,\mathrm{N}})}\right)\Bigg]\end{aligned} \tag{12}$$

in Np m^{-1}, with $f_{r,\mathrm{O}}$ given by Eq. (11) and $f_{r,\mathrm{N}}$ by Eq. (10) and with the percent mole fraction h_f related to relative humidity h_r by Eqs. (8) and (9). In Eq. (12), air temperature T is in kelvins; $T_0=293.15$ K; atmospheric pressure P is in N m^{-2}; and $P_0=1.013\,250\times 10^5$ N m^{-2} (1 atm). The combined effects of these different loss terms are shown in Fig. 8.

Fogs and aerosols can introduce additional sound absorption mechanisms (Wei *et*

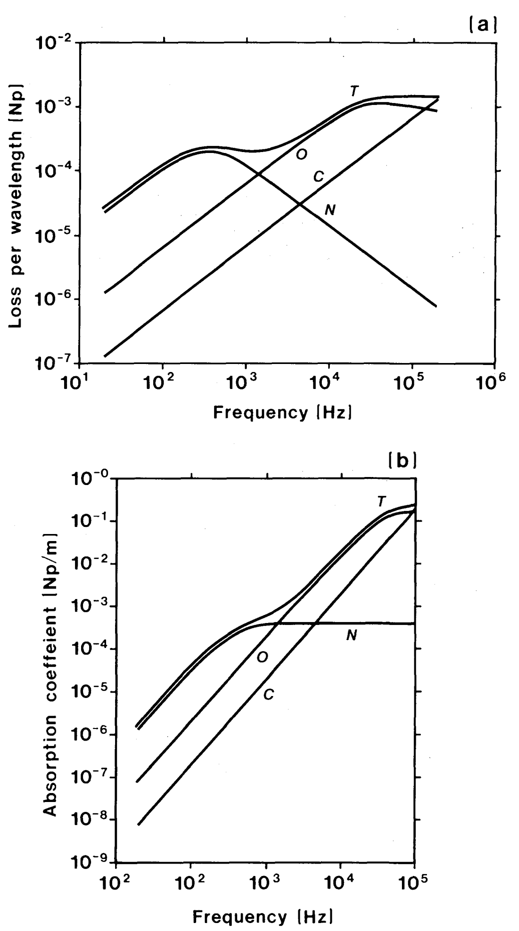

FIG. 8. Components and general behavior of total air absorption in air in terms of (a) loss per wavelength and (b) loss per unit distance. The contributions to the total absorption *T*, are *C*, classical and rotational relaxation; O, oxygen relaxation, and N, nitrogen relaxation.

al., 1987). These include evaporation and condensation around the droplets, partial entrainment of droplets by velocity fluctuation of the sound wave, heat transfer between droplets and surrounding air, the excitation of shape oscillations in the droplets, and gradients in the temperature and velocity along with heat fluxes and viscous effects. There have been little data taken in real fogs so our understanding of the mechanisms relies on theory and laboratory data.

Real fogs are composed of droplets with a wide range in size and number per unit volume. An average droplet radius is probably 8 μm and a typical ratio of liquid water content to air mass per unit volume might be 1.5×10^{-3}. A typical curve of fog absorption per unit length is given in Fig. 9 for different size droplets.

For comparison, molecular absorption at 100% relative humidity is also included. It can be seen, by comparison of these, that the two are comparable at low frequencies but that standard molecular absorption begins to dominate at high frequencies.

2.2 Effects of the Ground Surface

Research results of the past decade have greatly advanced understanding of sound propagation near the surface of the earth. Two major contributions have led to this rapid increase in understanding. The first was development of an accurate mathematical description of the sound field from a point source near a complex impedance boundary (Attenborough *et al.*, 1980). The second was the development of a microscopic description of the interaction of airborne sound with the earth (Sabatier *et al.*, 1986). Due to these advances, the physical understanding of sound propagation above plane homogeneous earth rivals that of atmospheric absorption. Referring to Fig. 10, the total acoustic field at the receiver can be written as

$$\phi_{\mathrm{TOT}}=\frac{e^{ik_1R_1}}{R_1}+\frac{e^{ik_1R_2}}{R_2}\{R(\theta_i)+B[1-R(\theta_i)]F(w)\}. \qquad (13)$$

In this expression for ϕ_{TOT}, k_1 is the propagation constant for air (m^{-1}), R_1 is the direct distance from the source to receiver, and R_2 is the path length from source to receiver that

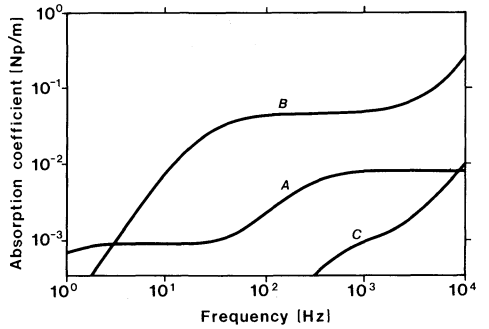

FIG. 9. Sound attenuation coefficient α_f, in units of Np/m, vs frequency. Curve A, $r=8$ μm, curve B, $r=1$ μm, curve C, molecular absorption [see Bass (1980)].

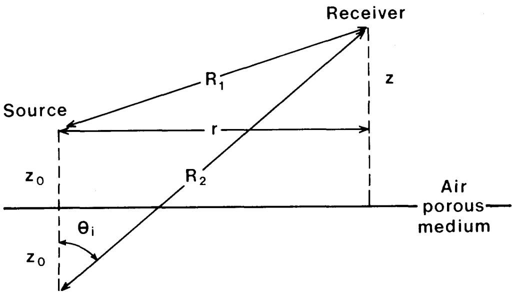

FIG. 10. Geometry of point-to-point propagation over a porous half-space.

includes one ground reflection. The quantity in curly braces is a reflection coefficient which has been modified for spherical waves reflecting from a complex-impedance flat boundary. Inside the curly braces is the usual plane-wave reflection coefficient $R(\theta_i)$, plus a term that depends on B and $F(w)$. The term B takes into account phase changes upon reflection from a complex impedance surface. For a locally reacting surface; (typical outdoors) $B\cong 1$. $F(w)$ arises from the need to match the curved wave front to a plane boundary. The quantity w is referred to as the numerical distance. It is useful to consider the numerical distance as representing the propagation distance scaled by the impedance. As w increases, wave fronts become more planar and $F(w)$ decreases.

The plane-wave reflection coefficient is given by

$$R(\theta_i)=\frac{\cos\theta_i-m(n^2-\sin^2\theta_i)^{1/2}}{\cos\theta_i+m(n^2-\sin^2\theta_i)^{1/2}}, \qquad (14)$$

where the angle θ_i is the angle of incidence of the incoming wave with the ground, measured from the vertical. The symbol n is the wave-number ratio

$$n=k_2/k_1, \qquad (15)$$

where k_2 is the propagation constant in the ground. The value for m may be deduced from the relation

$$\beta_c=mn=1/Z_c, \qquad (16)$$

where β_c and Z_c are known as the normalized characteristic admittance and the normalized characteristic impedance, respectively. The characteristic impedance can be written in terms of real and imaginary parts as $Z_c = R + iX$.

The equations for B and $F(w)$ are presented in their extended reaction form, both for added generality and a slight increase in accuracy. For the extended reaction case,

$$B=\left(\frac{[\cos\theta_i+\beta_c(1-\sin^2\theta_i/n^2)^{1/2}](1-n^{-2})^{1/2}}{[\cos\theta_i+\beta_c(1-n^{-2})^{1/2}/(1-m^2)^{1/2}](1-\sin^2\theta_i/n^2)^{1/2}}\right)\times\left(\frac{[(1-m^2)^{1/2}+\beta_c(1-n^{-2})^{1/2}\cos\theta_i+\sin\theta_i(1-\beta_c^2)^{1/2}]^{1/2}}{(1-m^2)^{3/2}(2\sin\theta_i)^{1/2}(1-\beta_c^2)^{1/4}}\right). \qquad (17)$$

For local reaction, which is often a good approximation, $B\cong 1$. The equation for the "sphericity" factor, $F(w)$, is

$$F(w)=1+i\pi^{1/2}we^{-w^2}\,\mathrm{erfc}(-iw), \qquad (18)$$

where erfc denotes the complementary error function, w is the so-called "numerical distance" defined by

$$w^2=ik_1R_2\{1+[\beta_c\cos\theta_i(1-n^{-2})^{1/2} -\sin\theta_i(1-\beta_c^2)^{1/2}]/(1-m^2)^{1/2}\}, \qquad (19)$$

and all square roots are taken so that their real parts are non-negative.

The first and second terms in Eq. (13) are readily identified as the direct and reflected waves. The last term ($[1 - R(\theta_i)]F(w)$ recalling that $B\cong 1$) arises mathematically from the need to match the boundary conditions, in particular the variation in wave-front curvature with distance (Piercy *et al.*, 1977). The function $F(w)$ in Eq. (18) is plotted for various values of the phase angle [$\phi=\tan^{-1}(X/R)$] in Fig. 11.

In order to understand the third term in Eq. (13) consider the case where the source and receiver are both on the ground ($\theta_i=90°$). For this case, $R(\theta_i)$ in Eq. (13) becomes -1 and, since $R_1 = R_2$, the first two

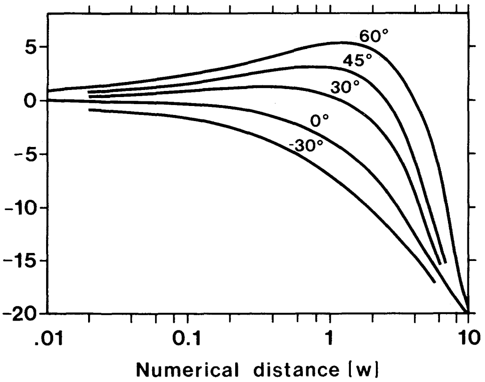

FIG. 11. Amplitude factor $F(w)$ of the ground and surface waves vs numerical distance w for various values of the phase angle ϕ of the surface impedance.

Table 1. The propagation distance d_w for numerical distance = 1, and height z_{sw} for the amplitude of the surface wave to decrease by $1/e$ from its value at the boundary, calculated from the impedance for mown grass.

f (Hz)	d_w (m)	z_{sw} (m)
50	4000	58
100	1500	23
200	270	7.3
500	24	0.9
1000	3.6	0.5
2000	1	0.15

terms in Eq. (13) completely cancel to form a shadow zone. This shadow zone is penetrated by the third term. For $\phi=0$ (purely resistive surface), the third term is referred to as a ground wave in analogy to electromagnetic propagation. The curve for $\phi=0$ in Fig. 11 indicates that for short distances, the ground wave suffers no attenuation in excess of spherical spreading but for longer distances falls off with an additional 6 dB per doubling of distance.

It will be shown later that for real ground surfaces, ϕ varies between 30° and 60° so F in Fig. 11 differs from that for $\phi=0$ by the pressure of a substantial increase for $w \geqslant 1$. This increase is due to the contribution of a surface wave in the air which propagates with an amplitude that decreases exponentially with height above the boundary. For $w<1$, the contribution of the surface wave is smaller than that of the ground wave because the surface-wave amplitude decreases at only 3 dB per doubling of distance. For $w>1$, the amplitude of the surface wave again becomes smaller because of attenuation by viscous losses in the pores of the boundary.

The scale of the distance and heights of the ground and surface waves is given in Table 1 for a grassy surface. Figure 12 gives relative contributions of the various waves.

Values of w, ϕ, and $R(\theta_i)$ all depend upon the complex normalized characteristic impedance of the sound. The four-parameter model of Attenborough (1985) can be used to calculate ground impedance. For ground surfaces commonly encountered, the complex normalized characteristic impedance of the ground can be approximated by

$$Z_c \cong \left(\frac{4q^2}{3\Omega} + \frac{iS_f^2\sigma}{\omega\rho_0}\right) \Big/ k_b, \tag{20}$$

where $q^2 = \Omega^{-n'}$ and n' is a constant which relates tortuosity (q) to porosity (Ω), S_f is the pore shape factor ratio, Ω is the porosity of the ground, σ is the flow resistivity of the ground (kg s^{-1} m^{-3} or g s^{-1} m^{-3}), ω is the angular frequency (s^{-1}), ρ_0 is the density of air (kg m^{-3}), and k_b is the normalized wave number (m^{-1}).

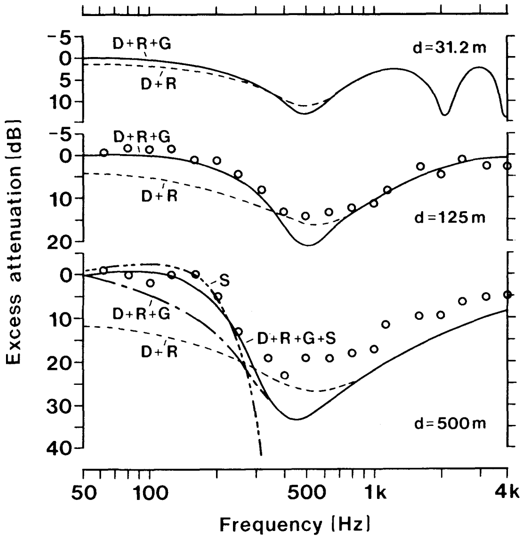

FIG. 12. Excess attenuation for propagation from a point source over mown grass, source height = 1.8 m, receiver height = 1.5 m. The calculated curves show the contributions from the various waves—direct D, reflected R, ground G, and surface S. The points are measurements of jet noise at comparable distances. The excess attenuation is relative to that for a point source placed on a perfectly hard surface.

The normalized wave number k_b is computed as

$$k_b \cong (\gamma\Omega)^{1/2}\left[\left(\frac{4}{3}-\frac{\gamma-1}{\gamma}N_{\text{pr}}\right)\frac{q^2}{\Omega}+\frac{iS_f^2\sigma}{\omega\rho_0}\right]^{1/2}, \tag{21}$$

where γ is the ratio of specific heats (1.4), and N_{Pr} is the Prandtl number (0.72).

The variables S_f, Ω, and n' have been varied individually until agreement between theoretical and measured values for impedance has been achieved for a variety of surfaces. As a result of such comparisons, n' is always set equal to 0.75 and S_f is always set equal to 0.875. Recommended values for porosity and flow resistivity are given in Table 2.

At intermediate to high frequencies, the semiempirical expression by Delany and Bazley (1977) and Chessel (1977),

$$Z_c=(R+iX)/[1-(k/k_\alpha)]^{1/2}, \tag{22}$$

where

$$\begin{aligned} R&=[1+9.08(f/\sigma_e)^{-0.75}],\\ x&=11.9(f/\sigma_e)-0.73,\\ k_\alpha&=(\omega/c_0)[1+10.8(f/\sigma_e)^{-0.70}\\ &\quad -i10.3(f/\sigma_e)^{-0.59}], \end{aligned} \tag{23}$$

has been found to give adequate agreement between predictions and experiment. In these equations, f is in Hz and σ_e is in cgs rayls. The effective flow resistance $\sigma_e \cong S_f^2\sigma/\Omega$. For large σ_e, Attenborough's four-parameter model reduces to

$$Z_c \cong 0.218(1+i)(\sigma_e/f)^{1/2}. \tag{24}$$

Some ground surfaces have a layered structure which results from the gradual deposition of material over a soil base. The effective impedance $Z(d)$ for a layer of depth d and impedance Z, above a semi-infinite layer of impedance Z_2 is given by

$$Z(d)=\left(\frac{Z_2-iZ_1\tan(k_bd)}{Z_1-iZ_2\tan(k_bd)}\right)Z_1, \tag{25}$$

where k_b is the bulk propagation constant in the upper layer. The quantities Z_1, Z_2, k_b are computed from Eqs. (21) and (22) or (23).

The effect of ground reflection can be seen in Fig. 13. The excess attenuation is given relative to what it would be in the free field (positive excess attenuation means lower level). The computed curve is for a source height of 15 m, receiver height of 0.55 m, range of 500 Hz, and specific impedance Z of $9.34-8.75i$.

The first dip is the so-called ground effect dip caused by interference between direct and reflected sound paths. At frequencies above the ground effect dip, constructive interference leads to an enhanced signal level followed by additional interference minima. At very large ranges, the direct and reflected paths will be nearly equal leading to a surface-induced shadow region.

Table 2. Parameters required for the four-parameter model; n' is always set equal to 0.750 and S_f (shape factor ratio) is always set equal to 0.875.

Description of surface	Ω	σ (cgs)
Dry snow, new fallen 0.1 m over about 0.40 m old snow	0.850	23
Sugar snow	0.825	48
In forest, pine or hemlock	0.825	60
Grass: rough pasture, airport, public buildings, etc.	0.675	330
Roadside dirt, ill-defined, small rocks up to 0.1 m mesh	0.575	960
Sandy silt, hard packed by vehicles	0.475	3 470
"Clean" limestone chips, thick layer (0.01–0.025 m mesh)	0.425	6 470
Old dirt roadway, fine stones (0.05 m mesh) interstices filled	0.400	7 500
Earth, exposed and rain-packed	0.350	17 100
Quarry dust, fine, very hard-packed	0.300	41 700
Asphalt, sealed by dust and light use	0.250	120 000

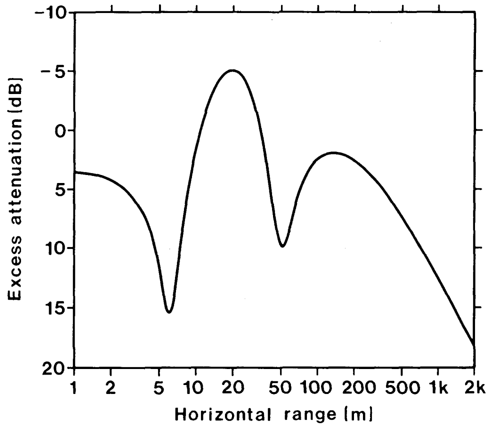

FIG. 13. Excess sound attenuation due to surface.

2.3 Refraction and Diffraction

The atmosphere is never quiet, but in some cases it is convenient to describe it in terms of temperature or, in the case of a steady wind, speed of sound as a function of altitude (Brown and Hall, 1978). Throughout the troposphere and stratosphere, the atmosphere has a relatively constant (with some variation with humidity) molecular weight of 28.97. Above infrasonic frequencies, the speed of sound c will be given by

$$c=343.3(T/293.15)^{1/2}\ \mathrm{m\,s^{-1}} \tag{26}$$

and the density will be given by the ideal gas law, $\rho = P_0/RT$ where P_0 is the static pressure and R is the universal gas constant. Figure 14, from Lindsey (1974), shows the speed of sound as a function of altitude for a model atmosphere. A model atmosphere, however, seldom if ever exists (Brown and Hall, 1978). The first few kilometers above the ground undergo significant diurnal variations. Inversions caused by radiative heating at night give rise to a sound speed which in creases with altitude near the surface. Following sunrise, rapid surface cooling causes surface air to rise pushing this inversion higher before destroying it. Wind causes the speed of sound to be dependent upon direction and can eliminate inversions by mixing air from different levels.

When the speed of sound as a function of altitude varies linearly, high-frequency sound follows a curved path represented by the arc of a circle shown in Fig. 15.

Wind causes the local speed of sound to become dependent upon direction. Sound propagating upwind will be refracted upward; downwind, waves are diffracted downward. When computing the local speed of sound, the important quantity for near horizontal propagation is

$$c=c+w\cos\theta, \tag{27}$$

where θ is the angle between the direction of propagation (propagation speed c) and the wind (wind speed w). Gradients in this quantity give rise to enhanced refraction and, at low frequencies, diffraction.

For horizontally stratified media, the ray paths are given by Snell's law,

$$\frac{\cos\theta(z)}{c_z}=s=\frac{\cos\theta_s}{c_s}, \tag{28}$$

where c_z is the sound speed at altitude z, c_s is the sound speed at the source, and s is a constant. The initial angle measured from horizontal is θ_s with upward-sloping rays having positive launch angles. When the sound speed gradient ($g_s = dc/dz$) is linear,

$$\cos\theta_f=s(c_s+g_sh), \tag{29}$$

where θ_f is the angle made with the adjoining interface and h is the separation between interfaces. Solving for h yields

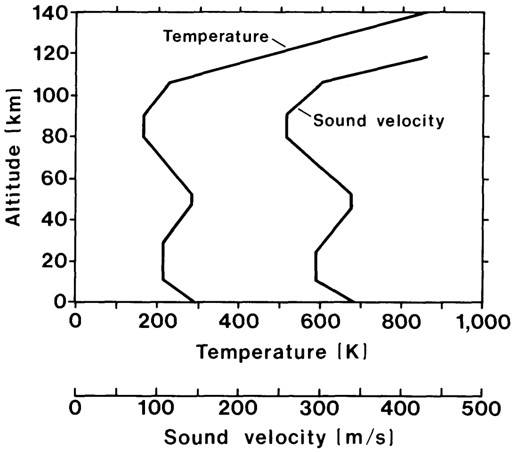

FIG. 14. Temperature and sound velocity in the model atmosphere as developed by the Air Research and Development Command in 1959. Details of the real atmosphere vary with location on the earth's surface and with the seasons. Wind speed also influences sound velocity (Lindsey, 1974).

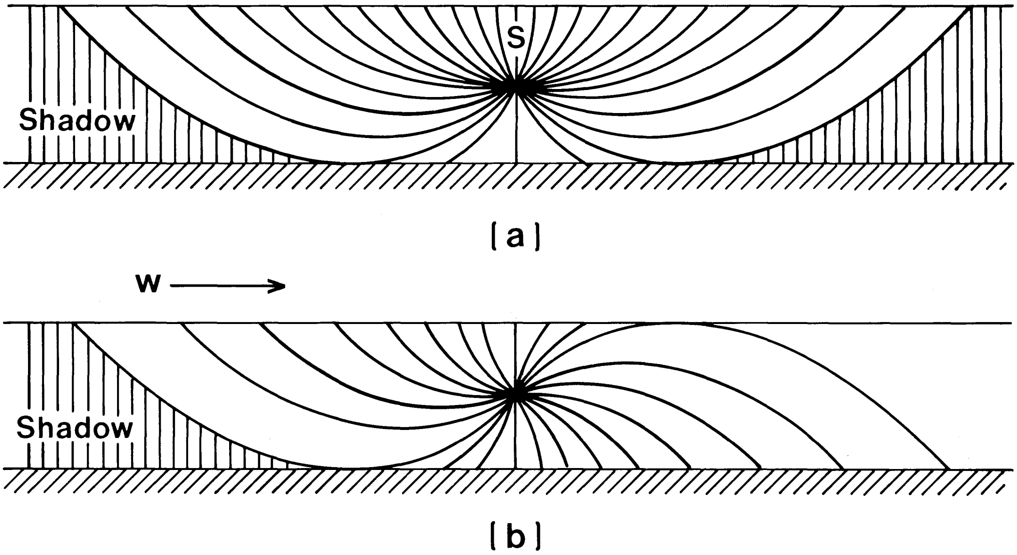

FIG. 15. Acoustic rays and shadow zones for (a) decreasing temperature and (b) increasing horizontal wind.

$$h=\frac{1}{sg_s}\cos\theta_f+\text{const.} \tag{30}$$

We note that the ray path is circular with radius $1/sg_s$. A single linear variation in c seldom applies in the first hundred meters or so. In addition, when significant (greater than 1%) variations in the speed of sound occur in a distance comparable to the acoustic wavelength, diffraction results. Ray tracing is not valid in this case and one must resort to numerical or approximate solutions to the wave equation.

The effects of diffraction are most pronounced near a shadow boundary. For an upward-refracting atmosphere, rays emanating from a source will bend upward creating an ever enlarging zone on the surface where rays cannot penetrate. In this shadow boundary, diffracted and scattered waves dominate.

A residue series has been obtained to treat the case of a stratified atmosphere where the sound speed varies linearly with height above a flat surface (Pierce, 1981; Berengier and Daigle, 1988). When $r > h_s + h$, where $z=h$ is the height of the receiver (see Fig. 16), the solution can be written

$$p(r,z)=\frac{\pi e^{i\pi/6}}{l}\sum_n H_0^{(1)}(k_nr)\,\frac{\text{Ai}[b_n-(h_s/l)e^{2i\pi/3}]\text{Ai}[b_n-(z/l)e^{2i\pi/3}]}{[\text{Ai}'(b_n)]^2-b_n[\text{Ai}(b_n)]^2}, \tag{31}$$

where

$$b_n=\tau\exp(2i\pi/3)=(k_n^2-k_0^2)l^2\exp(2i\pi/3) \tag{32}$$

are the zeros of the expression

$$\text{Ai}'(b_n)+q[\exp(i\pi/3)]\text{Ai}(b_n)=0. \tag{33}$$

The notation $H_0^{(1)}$ signifies the zeroth-order Hankel function of the first kind and Ai signifies an Airy function with primes denoting derivatives.

The abbreviations in Eqs. (31)–(33) are

$$q=ik_0\,\rho_0 lc/Z_c,$$
$$l=(R/2k_0^2)^{1/3},$$
$$\tau=(k^2-k_0^2)l^2,$$

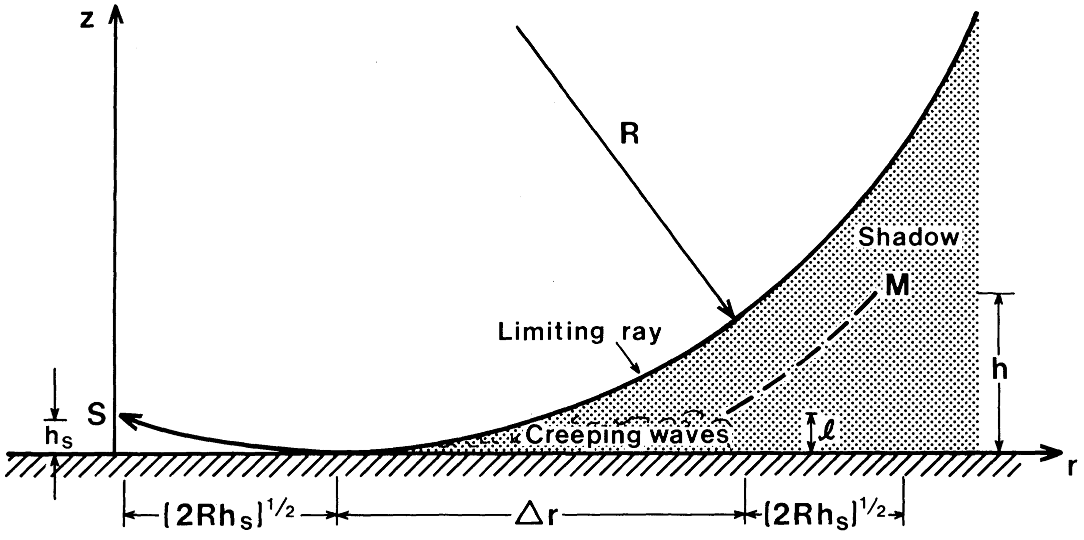

FIG. 16. Sketch showing the geometrical definitions for propagation in a stratified atmosphere above flat ground.

where $k_0 = \omega/c(0)$ and Z_c is the specific impedance of the surface.

The solutions of Eq. (33) for the b_n are simple only in the limiting cases of a rigid surface ($Z_c \approx \infty$) and that of a pressure release surface ($Z_c \approx 0$). In the intermediate cases of a surface of complex finite impedance Z_c, Eq. (33) must be solved numerically for its zeros for each value of n required in the series. The expression in Eq. (33) is a complex function of the complex variable b. Finding the zeros involves solving two nonlinear equations with two real unknowns for which a number of numerical methods exist.

The theory [Eq. (31)] is valid for propagation above a curved surface if z is interpreted as the height h transverse to the curved surface, r is interpreted as the arc length along the surface, and the radius of curvature R becomes, for the case studied here, the radius of the curved cylindrical surface.

When the receiver is deep within the shadow, but not close to the ground, the expression in Eq. (31) can be approximated by the first term in the series and, further, the Hankel and Airy functions can be approximated by their asymptotic behavior. For the purposes of this article it is sufficient to restrict the analysis to the magnitude of the pressure and these approximations yield the following expression:

$$|p(r,z)| = \left(\frac{1}{2k_0 r l^2}\right)^{1/2} \frac{1}{K(Z_c)} \left(\frac{1}{h_s}\right)^{1/4} \times \left(\frac{1}{z}\right)^{1/4} \exp[-\alpha(Z)\Delta r], \tag{34}$$

where k_0 and l are defined above. The function $K(Z)$ essentially describes the contribution of the denominator in Eq. (31) and has limiting values of $K(\infty) = 1.04$ in the case of a rigid surface and $K(0) = 1.74$ in the case of a pressure release surface. Similarly, the function $\alpha(Z)$ describes the attenuation of the creeping wave along the portion of the path Δr and takes on the following limiting values:

$$\alpha(\infty) = 0.7(k_0/R^2)^{1/3}, \tag{35}$$

in the case of a rigid surface, and

$$\alpha(0) = 1.61(k_0/R^2)^{1/3}, \tag{36}$$

in case of a pressure release surface.

As the frequency gets lower, the creeping wave approximation requires more and more terms in the expansion favoring a numerical solution to the wave equation.

2.4 Effects of Topography

Up to this point, the surface of the Earth has been assumed flat. The presence of hills, trees, buildings, and constructed sound barriers will modify, to some extent, the amplitude of the received sound field. The sound field behind a semi-infinite, thin, rigid plate with no ground surface can be predicted using the integral equations of Fresnel. In this simple treatment, the top edge of the barrier acts as a line source diffracting energy into the region behind the barrier. In practice, the problem is much more difficult.

If one considers a long barrier (Isei *et al.*, 1980), the sound paths shown in Fig. 17 will all contribute to some extent to the field at the receiver *R*. The total field Φ at receiver *R* depends on the fields from the real source *S* and the image source in the ground *T*, and the values of these fields at the image of the receiver position *X*, namely

$$\Phi=\Phi_{SR}+\Phi_{TR}+\Phi_{SX}+\Phi_{TX}, \qquad (37)$$

where Φ_{SR} is the direct diffracted field due to source *S* at receiver *R* (shortest path $SBR = d_{SB} + d_{BR}$), Φ_{TR} is the diffracted field due to image source *T* in the ground at receiver *R* (shortest path $TBR = d_{TB} + d_{BR}$), Φ_{SX} is the diffracted field due to image source *T* at image receiver *X*, both in the ground (shortest path $TBX = d_{TB} + d_{BX}$). More rigorously, the contribution from image source *U* in the barrier should not be overlooked. Then, instead of Eq. (37), we have

$$\Phi=\Phi_{SR}+\Phi_{TR}+\Phi_{SX}+\Phi_{TX}+\Phi_{UR}+\Phi_{UX}, \qquad (38)$$

where Φ_{UR} is the diffracted field due to image source *U* in the barrier at receiver *R* (shortest path $UBR=SBR$) and Φ_{UX} is the diffracted

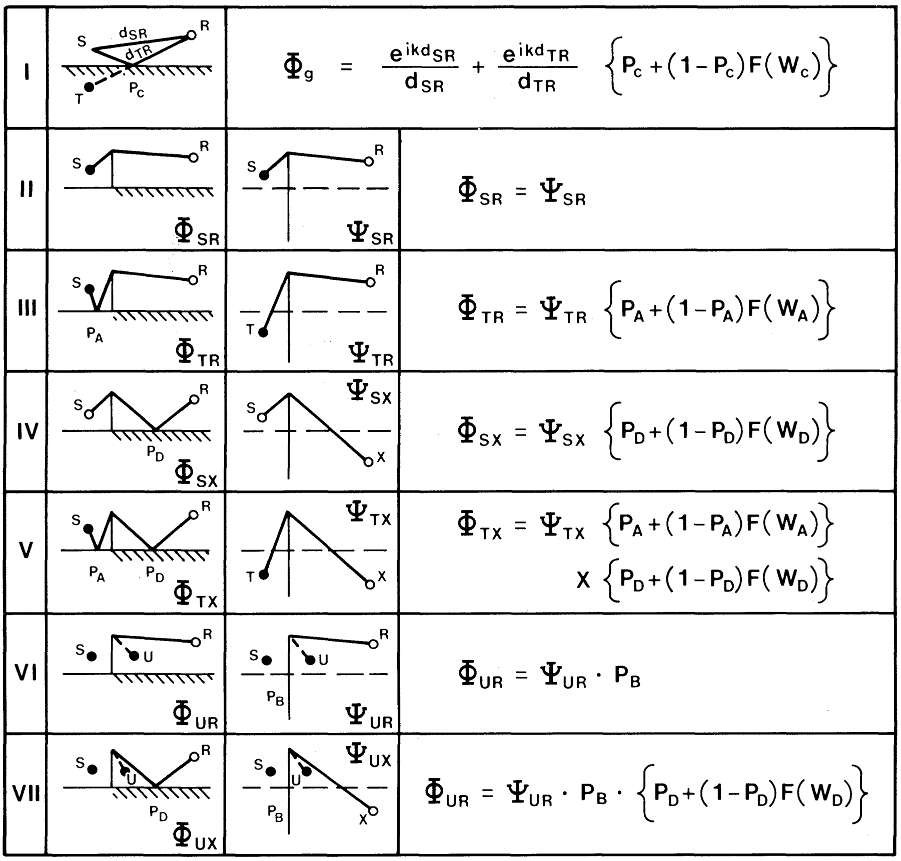
I
S
d_{SR}
d_{TR}
R
T
P_c
$\Phi_g = \frac{e^{ikd_{SR}}}{d_{SR}} + \frac{e^{ikd_{TR}}}{d_{TR}} \left\{P_c + (1-P_c)F(W_c)\right\}$
II
Φ_{SR}
Ψ_{SR}
$\Phi_{SR} = \Psi_{SR}$
III
P_A
Φ_{TR}
Ψ_{TR}
$\Phi_{TR} = \Psi_{TR} \left\{P_A + (1-P_A)F(W_A)\right\}$
IV
P_D
Φ_{SX}
Ψ_{SX}
X
$\Phi_{SX} = \Psi_{SX} \left\{P_D + (1-P_D)F(W_D)\right\}$
V
Φ_{TX}
Ψ_{TX}
$\Phi_{TX} = \Psi_{TX} \left\{P_A + (1-P_A)F(W_A)\right\}$
$\times \left\{P_D + (1-P_D)F(W_D)\right\}$
VI
U
P_B
Φ_{UR}
Ψ_{UR}
$\Phi_{UR} = \Psi_{UR} \cdot P_B$
VII
Φ_{UX}
Ψ_{UX}
$\Phi_{UR} = \Psi_{UR} \cdot P_B \cdot \left\{P_D + (1-P_D)F(W_D)\right\}$

FIG. 17. Ray path diagrams and equations used for calculating the six components (lines II–VII) of the diffracted sound field at *R*. Line I shows how to calculate, without the barrier, the reference sound field needed for describing barrier insertion loss. Column 1 shows the actual ray path, including reflections, that relates to the field component Φ. Column 2 shows the equivalent ray path used for calculations of the diffracted field corrected for reflection and possible presence of ground waves, using the equation of column 3, in order to obtain the desired field component Φ.

field due to image source *U* at the image receiver *X* (shortest path *UBX*=*SBX*). Furthermore, if the receiver *R* is directly illuminated by the source *S*, or its image *T*, then another two fields, due to geometrical spreading from each source, should be added to the field Φ.

The evaluation of each of these terms gives rise to predictions such as that shown in Fig. 18. As expected, at low frequencies, barriers provide little shielding.

Natural barriers such as hills can be represented by wedge shaped barriers or a wedge with a horizontal top (wide barrier). Again, the largest insertion loss is observed at high frequencies.

In practical applications, the insertion loss shown in Fig. 18 near 5 kHz is seldom achieved. Refracted and scattered sound often fill in interference dips providing less attenuation. Typical results are shown in Fig. 19 (from Daigle, 1982) with highway noise as a source and a receiver on the opposite side of a barrier 1 m above an asphalt surface. Note that at high frequencies where diffraction theory predicts the greatest insertion loss, the experimental measurements liewell above theoretical predictions. Scattering degrades the effect of the barrier.

For noise control, it is often desirable to make use of plants (trees, shrubs, etc.) (Price *et al.*, 1988). At low frequencies (below ~800 Hz), the primary effect of plant growth along the propagation path is a modified surface impedance due to deposited organic matter. The surface flow resistance is typically decreased in such regions to below 100 cgs rayls. Further, the organic layer requires the use of a two-layer model for the surface impedance. At higher frequencies, scattering by trees can become significant for dense growths.

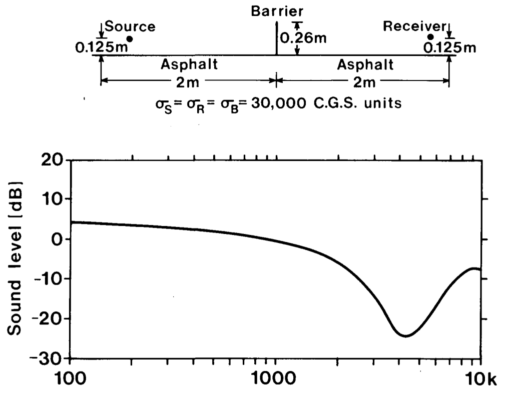

FIG. 18. Attenuation spectrum predicted by diffraction theory for the particular symmetrical configuration of source, receiver, barrier, and hard ground sketched at the top of the figure.

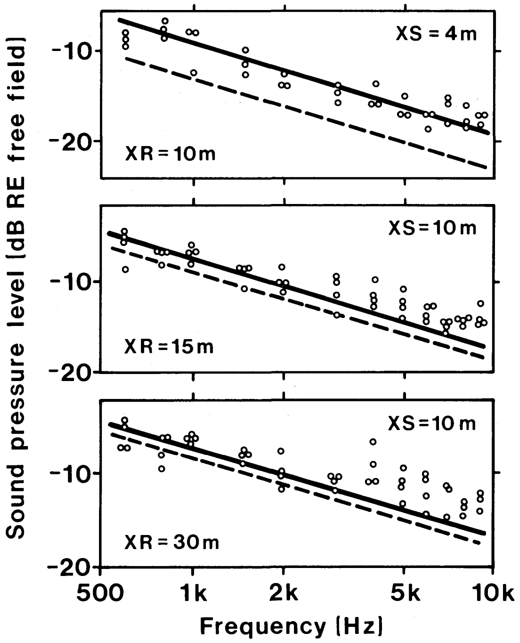

FIG. 19. The points are the measured sound pressure levels from a point source diffracted at the edge of a barrier 2.44 m high. XS and XR are, respectively, the horizontal distances from source to barrier and receiver to barrier (on hard ground—asphalt). The dashed curve is the prediction of diffraction theory.

Even at high frequencies, however, insertion loss due to trees is typically small. Figure 20 gives results for a very dense coniferous site where measurable sound attenuation is achieved. The trees can be modeled as scattering cylinders which are assumed to be identical, vertical, and infinitely long. Multiple scattering is allowed. The use of two arrays is intended to represent scattering both by tree trunks and by foliage.

2.5 Effects of Turbulence

The sound level at a receiver is seldom the value expected assuming the mechanisms discussed to this point. The real atmosphere is far from stationary with changes in acoustic index of refraction occurring on a time scale of milliseconds to seconds due to atmo-

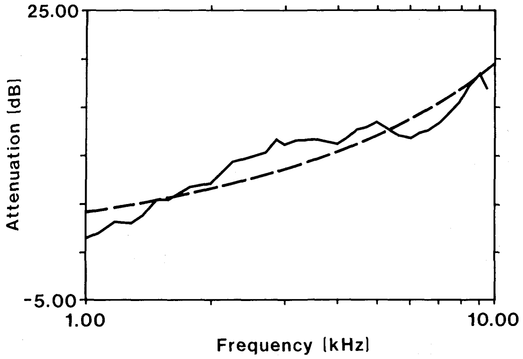

FIG. 20. Mean measured attenuation in dB per 24 m at site 3 (mixed coniferous) compared with a prediction (broken line) based upon the sum of two scattering arrays of cylinders: (a) rigid, radius 0.066 m, density 0.0724 m^{-2} and (b) nonrigid, radius 0.001 m, density 100 m^{-2}; σ_e (of surface) 860 $g^{-1}\,cm^{-3}$.

spheric turbulence. There are two sources for the turbulence. The first is that due to heating and cooling of the surface which gives rise to thermal plumes (thermally driven turbulence). The second is due to the presence of the ground surface which causes a gradient in wind leading to shear and turbulent eddy formation. In most cases, this wind-driven turbulence dominates. Acoustically, this turbulence serves to scatter sound into shadow zones and causes sound beams to spread.

First, consider a simple model of turbulence effects on the acoustic field. A simplified representation is shown in Fig. 21. As an acoustic ray travels toward the receiver, its direction and phase will be modified slightly when it encounters a turbule. The received signal will be some vector sum of these individual path segments plus any random noise at the same frequency. The next (either later or adjacent) acoustic ray will follow a slightly different path arriving with different amplitude and phase. The net result is that the received amplitude and phase will form a pattern in phase space rather than a point.

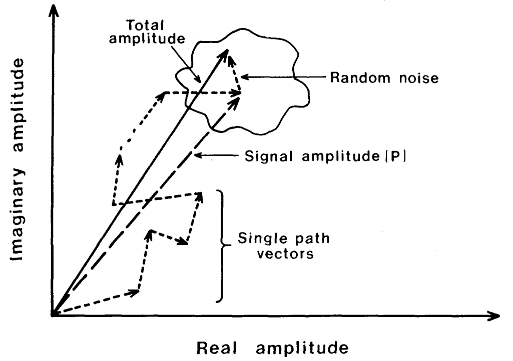

FIG. 21. Representation of acoustic ray scattered by turbulence.

If we now look along the amplitude vector, the probability (P) of observing the received sound with a given amplitude, p, can be represented by a Ricean distribution (Rayleigh at high frequencies, Gaussian at low frequencies) as shown in Fig. 22 and given by

$$P(p)=\frac{p}{\delta^2}\,(e^{-(p^2+p_0^2)/\delta^2})I_0(pp_0/\delta^2), \tag{39}$$

where p_0 is the average sound pressure, δ is a measure of the width of the distribution, and I_0 is the modified Bessel function.

It is sometimes convenient to represent phase and log-amplitude fluctuations in terms of structure functions D_s and D_x, respectively, in a plane perpendicular to the direction of propagation. For two microphones separated by distance ρ at a distance r from the source,

$$D_x(r,\rho)=\langle[x(\mathbf{r}+\rho)-x(\mathbf{r})]^2\rangle \tag{40}$$

and

$$D_s(r,\rho)=\langle[\phi(\mathbf{r}+\rho)-\phi(\mathbf{r})]^2\rangle, \tag{41}$$

where $x(\mathbf{r})$ is the log-amplitude measured at one microphone and $x(\mathbf{r}+\rho)$ is the log-amplitude measured at the second.

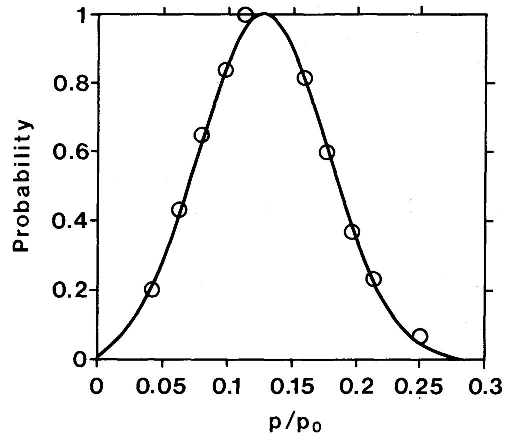

FIG. 22. Probability of observing an acoustic amplitude p, as a function of that amplitude divided by the free-field amplitude p_0. Open circles are experimental points. Solid line computed for Eq. (39) with $p_0=0.117$, $\delta=0.052$.

For $(\lambda r)^{1/2} \gg L$ where L is the scale of turbulence, simple single-scattering theory predicts that

$$D_x(r,\rho) = 2[\langle x^2\rangle - B_x(\rho)], \tag{42}$$

$$D_s(r,\rho) = 2[\langle s^2\rangle - B_s(\rho)], \tag{43}$$

where

$$\langle x^2\rangle = \langle s^2\rangle = \left(\frac{\sqrt{\pi}}{2}\right)\langle \mu^2\rangle k^2 rL, \tag{44}$$

μ is the rms fluctuation in the acoustic index of refraction, and k is the propagation constant. $B_x(\rho)$ and $B_s(\rho)$ are, respectively, the covariances of the log-amplitude and phase fluctuations,

$$\frac{B_x(\rho)}{\langle x^2\rangle} = \frac{B_s(\rho)}{\langle s^2\rangle} = \frac{\Phi(\rho/L)}{\rho/L}, \tag{45}$$

where

$$\Phi(\rho/L) = \int_0^{\rho/L} e^{-u^2} du. \tag{46}$$

In practice Eq. (44) agrees with experimental results for $\langle s^2\rangle$. However, measurements show that at large distances $\langle x^2\rangle \ll \langle s^2\rangle$ and, in addition, the log-amplitude fluctuations quickly saturate.

The two parameters that are used to characterize fluctuating wind speed and temperature are the strength and the scale (Johnson *et al.*, 1982). The strength of the turbulence is represented statistically by the standard deviation of the fluctuations (denoted by σ_T or σ_v for the temperature or wind speed, respectively). The scale of the fluctuations (denoted by L) is determined from the spectral density function or the correlation function.

To model the variation in scale and magnitude of convective and mechanical turbulence with altitude under unstable conditions, we must introduce two types of scaling: Monin–Obukhov scaling and mixed-layer scaling.

Monin–Obukhov's similarity theory allows the development of relationships between the turbulence variables and the surface roughness, wind speed, and convection to be expressed in terms of the single scale L_{MO} and the altitude z. The Monin–Obukhov length is a function of wind speed, heat flux, and surface roughness.

In practice, L_{MO} can be estimated from the Turner classes, which were developed for use in air pollution studies. The Turner class is determined from the wind velocity and an estimate of solar radiation (see Table 3). Since L_{MO} is also a function of surface roughness, this parameter must also be included in estimating L_{MO} from the Turner class. In Table 4 we present the range of Monin–Obukhov lengths for a given Turner class for a site which has a roughness length of about 0.05 m. For all conditions investigated, L_{MO} is negative.

The planetary boundary layer (PBL) represents the farthest extent of surface heating and cooling. This layer varies in thickness from a few hundred meters at night to a few kilometers on a sunny day. The region in the planetary boundary layer above $-L_{\mathrm{MO}}$ is known as the "mixed" layer. The scaling parameter in this layer is the mixing depth h, the vertical extent of the planetary boundary layer. This type of scaling is called mixed-layer scaling. In the mixed layer, certain turbulence statistics can be written as universal functions of z/h. Some statistics, in particular those associated with vertical velocity fluctuations, still obey Monin–Obukhov scaling in the mixed layer.

With these two scaling factors in mind, we can now examine the theoretical variation of

Table 3. Estimation of Turner classes (daytime only).

Surface wind speed (at 10 m) (m/s)	Incoming solar radiation		
	Strong	Moderate	Light
<2	1	1	2
2–3	1–2	2	3
3–5	2	2–3	3
5–6	3	3–4	4
>6	3	4	4

Class 4 should be assumed for overcast conditions.

Table 4. Turner class versus L_{MO} for a flat site in Illinois.

Turner class	$-L_{\mathrm{MO}}$ (m)
1	8–12
2	12–20
3	20–60
4	60

convective and mechanical turbulence with altitude for the case of an unstable to near neutral daytime atmosphere.

The dominant term in the vertical velocity fluctuation has scales which are a monotonically increasing function of height. As $-L_{\mathrm{MO}}$ increases, this function becomes approximately linear. The magnitude of the scale is independent of stability for $z>-0.7L_{\mathrm{MO}}$.

Horizontal velocity fluctuations obey mixed-layer scaling. The scale of the horizontal fluctuations is determined by the height of the mixed layer and is nearly constant, as a function of altitude, in the planetary boundary layer. These scales will be larger than the vertical velocity scales for unstable measurement conditions.

The temperature fluctuations are closely related to the vertical velocity fluctuations and are expected to obey Monin–Obukhov scaling. Since the temperature fluctuations are caused by vertical velocity fluctuations acting on the temperature gradient, the temperature fluctuation scale is expected to increase monotonically with altitude under unstable conditions.

The variance of the wind speed will also be determined by fluctuations in both the horizontal and vertical. Examining the theoretical variance in the surface layer, the variance in the vertical wind component is expected to increase slowly with height for unstable conditions as a function of z/L_{MO}:

$$\sigma_w \propto [1-3(z/L_{\mathrm{MO}})]^{1/3}. \tag{47}$$

The larger L_{MO}, the slower the variation. The variances of the horizontal components should be constant in the unstable surface layer. Above the surface layer, all wind variances should decrease as the altitude approaches the top of the planetary boundary layer.

The temperature variance in unstable air is a more rapidly decreasing function of height in the surface layer than the variance of the vertical wind velocity. Johnson *et al.* (1982) use the form

$$\sigma_T \propto [1-16(z/L_{\mathrm{MO}})]^{-1/2}. \tag{48}$$

[Note that Eq. (48) corrects an error in Eq. (2) of Johnson *et al.* (1982).] This implies, for example, that at $-L_{\mathrm{MO}}$ the variance should be one-fourth of that close to the ground. Above L_{MO}, the variance in temperature should decrease as $(z/L_{\mathrm{MO}})^{-1/3}$.

A summary of the average results of measurements made at Bondville Field Station in Illinois serves as indicators of typical results. This summary represents typical measured values of turbulence strength and scale at a location that has good fetch. A rougher terrain with poor fetch would have values of L and $\langle\mu^2\rangle$ that would be slightly larger near the ground. The variation of these variables with height for poor fetch would probably be similar to these results at heights away from the ground, but there are no data to confirm this.

(a) Variation of scale.

For overcast, light wind conditions (implying a shallow PBL),

$L\approx 1.0$ m @ 1 m to $L\approx 7.0$ m @ 33 m.

For sunny, windy conditions (implying a deep PBL),

$L\approx 1.0$ m @ 1 m to $L\approx 12.0$ m @ 33 m.

The variation of L with height is approximately linear.

(b) Variation of strength.

Sunny, light wind (<2 m/sec) conditions: $\langle\mu^2\rangle\approx 5\times 10^{-6}$.

Sunny, moderate wind (2–4 m/sec) conditions: $\langle\mu^2\rangle\approx(9\text{–}10)\times 10^{-6}$.

Sunny, strong wind (>4 m/sec): $\langle\mu^2\rangle\approx(15\text{–}25)\times 10^{-6}$.

Overcast, light wind conditions: $\langle\mu^2\rangle\approx 3\times 10^{-6}$.

Overcast, moderate wind conditions: $\langle\mu^2\rangle\approx(8\text{–}9)\times 10^{-6}$.

Overcast, strong wind conditions: $\langle\mu^2\rangle\approx(15\text{–}25)\times 10^{-6}$.

There is very little variation of $\langle\mu^2\rangle$ with height through 10 m. Above 10 m, $\langle\mu^2\rangle$ may be slightly increasing, as determined by the wind speed variance.

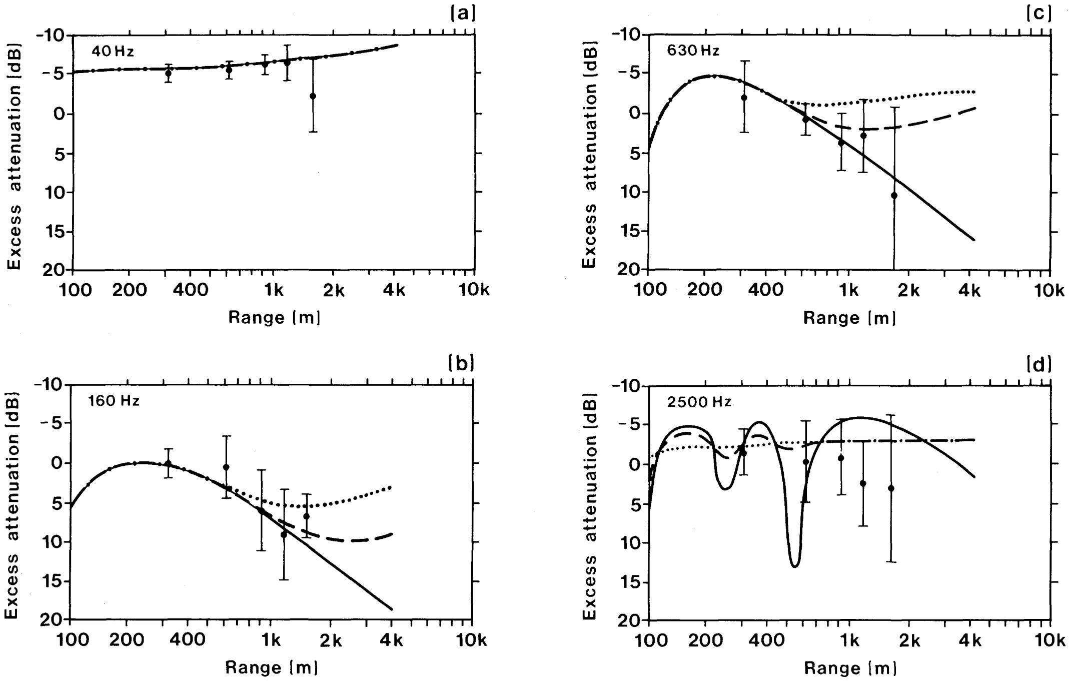

FIG. 23. Excess attenuation vs horizontal distance for 40, 160, 630, and 2500 Hz. The data are the points with error bars. The data at 1550 m are limited since wind noise reduced the sample size. The solid lines are the excess attentuation prediction for a nonturbulent atmosphere. The dashed lines are predictions for $L_0 = 1.0$ m, $\langle\mu^2\rangle = 3\times10^{-6}$, and the dotted lines are for $L_0 = 1.0$ m, $\langle\mu^2\rangle = 15\times10^{-6}$.

Figure 23 displays the variation in excess attenuation with distance for four frequencies measured at Bondville Field Station, Illinois. The source is a set of eight speakers on top of a tower at a height of 31 m. The receivers were 1.2 m high. The data displayed are the average of eleven 15-min measurements of excess attenuation selected as representing a nearly homogeneous atmosphere. These sets had a logarithmic sound speed gradient between -0.11 m/sec and 0.11 m/sec and the standard deviation of the gradient was less than 0.05 m/sec. The predicted quiet-atmosphere excess attenuation with distance for the average ground impedance condition is also displayed in Fig. 23 as the solid lines. These calculations display the expected asymptotic decay of excess attenuation at $1/r^2$ for large distances. The transition to $1/r^2$ depends on frequency, source height, and impedance.

The data for 160 and 630 Hz behave more coherently than was expected. The 40-Hz data and calculation display little variation from coherent behavior and will not be discussed further. The 2500-Hz data, on the other hand, display near incoherent behavior at all ranges.

Calculations based on the extremes of the variance of the index of refraction are displayed as the dashed lines in Fig. 23. These results are calculated for a scale length of 1 m, typical of the scale lengths measured at 1-m height. It is observed that the turbulence effects are less than would be observed for ground to ground propagation for the same scale length. Further, the calculation indicates a larger difference in the coherent and turbulent results than was measured.

For ground to ground propagation, in a quiet homogeneous atmosphere, there is cancellation between the direct wave and plane-wave reflected wave and only the lower-amplitude ground wave is measured. Turbulence reduces the coherence between the equal-amplitude waves, and therefore increases the amplitude of the resultant wave.

2.6 Nonlinear Effects

Large-amplitude waves and impulses will give rise to nonlinear propagation effects. If we consider an impulse as shown in Fig. 4, the point on the pressure wave form corresponding to maximum pressure will travel more rapidly than points of lower pressure according to the relation

$$\Delta v = c[(\gamma+1)/(\gamma-1)][(p/P_0)^{(\gamma-1)/2\gamma}-1], \tag{49}$$

where Δv gives the speed at which a point on the pressure wave form with pressure p travels in excess of the $p=P_0$ point; Δv may be positive or negative. The amount the positive phase of the wave form lengthens can be found from

$$\Delta L = \int_{t_0}^{t} \Delta v \, dt'. \tag{50}$$

The term Δv is a function of time, since from Eq. (49), Δv depends on p/P_0, and $\Delta p/P_0 = (p - P_0)/P_0$ decreases as r raised to a negative power depending upon the spreading law involved, which in turn depends on t. The relationship between r and t is not simple, since

$$r = vt + r_0 \tag{51}$$

and $v = c + \Delta v$, which is a function of (p/P_0) from Eq. (49).

As the wave propagates, the wave steepening predicted by Eq. (49) will tend to transfer energy from low frequency to high. Since atmospheric absorption increases approximately as the frequency squared, these high frequencies will be rapidly attenuated. The overall rate of decay of the wave form, then, will depend upon how rapidly low-frequency energy is transferred to high frequencies which in turn depends upon the amplitude of the acoustic pressure.

One indicator of this wave steepening versus high-frequency attenuation competition is the shock rise time discussed earlier. Referring to Fig. 24, we can see that the rise time decreases as the over-pressure increases. Since the rise time is a measure of the highest frequencies in the impulse, the small rise time (high frequency) associated with high amplitude shows that nonlinear steepening is replacing energy more rapidly for high amplitude waves.

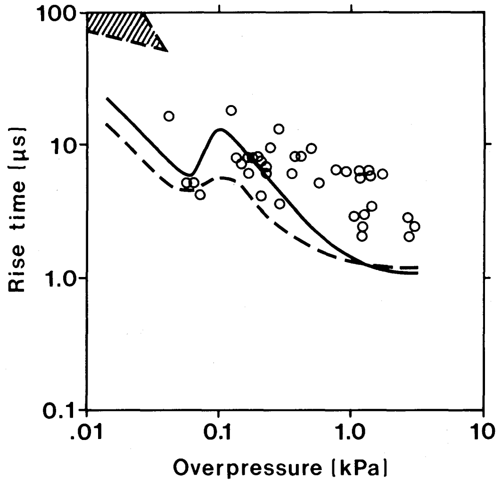

FIG. 24. Experimental and computed rise times. The hatched area and circles represent experimental data from Ref. 1. The solid line was computed for $T=305$ K, relative humidity 90%. The dashed line was computed for $T=293$ K, relative humidity 50%.

Harmonic waves are affected by nonlinearity in slightly different ways. Rather than pulse lengthening, harmonic waves exhibit subharmonic generation. As the subharmonics grow (they are attenuated less than the harmonics), they can produce harmonics and subharmonics until, provided the original amplitude was great enough, the harmonics and subharmonics will completely fill the frequency spectrum and appear as noise.

2.7 Infrasound Propagation

The primary mode of propagation for infrasound at long ranges is by multiple hops; each hop is typically 250 km. This process results from temperature inversions in the upper atmosphere (altitudes of 50–120 km) and reflections at the earth's surface. Referring to the section on absorption, we find that at low frequencies absorption is very small; hence long propagation paths are allowed. The surface impedance is very large at low frequencies so very little energy is lost to the surface on each hop. After a few hops, surface-generated infrasound is constrained to a channel and spreads cylindrically rather than spherically. Empirically it is found that (ANSI, 1983)

$$p \sim R^{-n} \text{ with } n \cong 1.1. \tag{52}$$

Infrasound propagation depends strongly on winds in the atmosphere which change seasonally. As a result, propagation is more effective toward the east in the winter when atmospheric winds tend to blow in that direction. The velocity of propagation for infrasound also varies seasonally with wind and temperature of the primary transmission channel. As shown in Fig. 14, the local sound velocity can rise dramatically at high altitudes but the fact that the waves bounce back and forth between the surface and the upper transmission channel increases the effective propagation distance. Therefore, infrasound velocities are typically a little less than the speed of sound near the earth's surface.

2.8 Numerical Techniques

The conditions for which ray tracing is applicable are

1. $\lambda n' \ll n^2$ where n' is the rate of change in the index of refraction n, and λ is the acoustic wavelength;
2. $\lambda(\ln A)' < 1$ where $(\ln A)'$ is the rate of change in the logarithm of the acoustic amplitude; and
3. $(\lambda/2\pi R)^{1/3} < \theta$ where R is the propagation range and θ is the grazing angle.

Note that each condition becomes increasingly difficult to meet as λ increases (low frequency). Two common sources of low-frequency sound are explosions and rotary wing aircraft.

Air pressure waves from very energetic explosions [the Mount St. Helen's eruption (Reed, 1987), for example] travel through the upper atmosphere and are refracted slowly so that n' is always very small. In these events, ray tracing routines accurately predict quiet zones relatively near the sources and very dramatic damage much further away due to refractive focusing. Low-flying helicopters do not produce sufficient energy to result in detectable sound levels after reflection by the upper atmosphere. The received sound depends, primarily, on index of refraction gradients near the ground surface. In this case, n' can be quite large. For such cases ray tracing can lead to poor predictions of long-range sound propagation and a numerical solution to the wave equation becomes necessary.

Full wave solutions take a variety of forms. The most accurate is a finite element solution. For propagation distances greater than a few meters, such algorithms take many hours of supercomputer time. Most computations involve modified ray tracing routines, the parabolic approximation, or the fast field program. At long ranges, corrected ray tracing routines are typically more time consuming than the other two approaches so ray tracing is generally limited to cases where extensive modifications are not required (sources nearly overhead or an almost neutral atmosphere). Of the remaining two approaches, the PE (parabolic equation) has the capability of including the largest number of effects but is least developed. At very long ranges, the FFP (fast field program) has an advantage in run time but can still consume several minutes on a minicomputer.

For purposes of this contribution, only the FFP will be described in detail (Raspet *et al.*, 1985). The problem considered is sketched in Fig. 25. In the space $0<z<z_{n-1}$, there exist $n-1$ liquid layers. A time-harmonic point source of constant pressure is located at the lower interface of layer S with cylindrical coordinates ($r=0$, $z=z_S$). The problem at hand is to determine the pressure at the observation point located at the lower interface of layer D with coordinate ($r=r$, $z=z_D$).

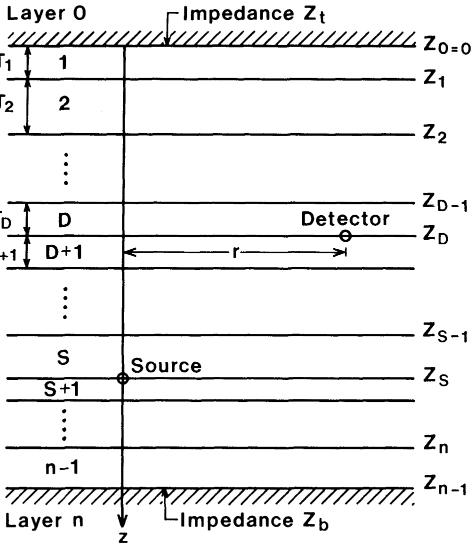

FIG. 25. Liquid layered medium with source and detector bounded by impedance surfaces.

In a typical layer m, the acoustic field is described by pressure $p_m(r, z)$ and a vertical particle velocity $\dot{w}_m(r, z)$. Both of them are expressed in terms of the familiar Fourier–Bessel transform

$$\begin{bmatrix} p_m(r, z) \\ \dot{w}_m(r, z) \end{bmatrix} = \int_0^\infty \begin{bmatrix} p_m(k, z) \\ \dot{w}_m(k, z) \end{bmatrix} J_0(kr)dk. \quad (53)$$

Here, the transform variable k can be interpreted as the propagation constant. Two surface-impedance boundary conditions are to be enforced at $z=0$ and at $z = z_{n-1}$.

At a surface $z=a$, the surface impedance in the transform domain is defined by

$$Z(k) = \text{sgn}\,\frac{p(k,z=a)}{\dot{w}(k,z=a)}, \quad (54)$$

where

$$\text{sgn} = \begin{cases} +1 & \text{if } \dot{w} \text{ moves into impedance surface,} \\ -1 & \text{if } \dot{w} \text{ moves away from the surface.} \end{cases}$$

For the configuration in Fig. 25, sgn$=-1$ if $a=0$ for top impedance Z_t, and sgn$=+1$ if $a=z_{n-1}$ for bottom impedance Z_b. It is assumed that the space above $z=0$ is a liquid half-space, i.e.,

$$Z_t = \omega\rho_0/(k_0 - k^2)^{1/2}, \quad (55)$$

where, for $m=0,1,2,\ldots,n$,

ρ_m = density of layer m in kg/m^3,

c_m = acoustic wave speed in layer m in m/s,

$k_m = \omega/c_m$ = wave number in layer m in m^{-1}.

Use the exp($+i\omega t$) time convention; then (54) assumes the following values:

$$Z_t(k) = \begin{cases} +|Z_t(k)|\,(\text{resistive}) & \text{if } k_0 > k, \\ +i|Z_t(k)|\,(\text{inductive}) & \text{if } k_0 < k. \end{cases}$$

The surface impedance Z_b at the bottom surface has already been described.

An integral solution for the pressure $p(r) = p_D(r, z=z_D)$ at the detector's location can be written as

$$\frac{p(r)}{p_0} = (-i)\int_0^\infty F(k)J_0(kr)dk, \quad (56)$$

where

$$F(k) = \frac{2k}{\omega\rho_s}\,[(D_D)_{21} - Z_t(D_D)_{22}]\left(\frac{(D_s)_{22}(D_{21}+Z_bD_{11}) - (D_s)_{21}(D_{22}+Z_bD_{12})}{(D_{21}+Z_bD_{11}) - Z_t(D_{22}+Z_bD_{12})}\right); \quad (57)$$

p_0 = source pressure at a unit distance away in a homogeneous medium.

In Eq. (57), we have used propagation matrices

$$D_s = d_s d_{s-1} \cdots d_2 d_1,$$

$$D = d_{n-1} d_{n-2} \cdots d_2 d_1,$$

where d_m is a 2×2 matrix with elements given by

$$(d_m)_{11} = (d_m)_{22} = \cos P[\xi_m(z_m - z_{m-1})],$$

$$(d_m)_{12} = (i\xi_m/\omega\rho_m)\sin P[\xi_m(z_m - z_{m-1})],$$

$$(d_m)_{21} = (i\omega\rho_m/\xi_m)\sin P[\xi_m(z_m - z_{m-1})],$$

$$\xi_m = \sqrt{k_m^2 - k^2},$$

D_{12}, for example, is the (1,2) element of matrix D.

Now consider the evaluation of the integral involved in Eq. (56). Of primary interest is the far field solution, i.e., the solution when the observation point is far away from the source. To this end, $J_0(kr)$ can be rewritten as

$$J_0(kr) = \tfrac{1}{2}\,[H_0^{(1)}(kr) + H_0^{(2)}(kr)]. \quad (58)$$

The component involving the incoming wave factor $H_0^{(1)}(kr)$ does not contribute significantly, and can be neglected. Thus, Eq. (56) is approximated by

$$\frac{p(r)}{p_0} \approx \frac{1}{2i}\int_0^\infty F(k)H_0^{(2)}(kr)dk. \quad (59)$$

Two further approximations are used in (59), namely,

1. for the far field, $H_0^{(2)}$ is replaced by its asymptotic expansion:

$$H_0^{(2)}(kr) \sim (2/kr)^{1/2} e^{-i(kr-\pi/4)}, \quad kr \to \infty;$$

2. the range of integration is truncated at $k = k_{\max}$. The final expression suitable for numerical evaluation is

$$\frac{p(r)}{p_0} \approx \frac{1-i}{2\sqrt{\pi r}} \int_0^{k_{\max}} \frac{F(k)}{\sqrt{k}} e^{-ikr} dk, \quad r \to \infty, \tag{60}$$

which is the normalized pressure at a distance r from the source.

Next consider the numerical evaluation of (60) by the fast Fourier transform. To this end, four parameters are required:

$k_{\max}$ = maximum horizontal wave number in integration range,

N = number of steps of integration,

k_0 = starting point of integration,

r_0 = minimum range.

The integral in Eq. (60) is approximated by a summation:

$$\frac{p(r_{n+1})}{p_0} = (\Delta k) \frac{1-i}{2(\pi r_{n+1})^{1/2}} \times \sum_{m=0}^{N-1} \left(\frac{F(k_{m+1})}{k_{m+1}^{1/2}} e^{-imr_0 \Delta k} \right) \times e^{-i2\pi mn/N}, \tag{61}$$

for $n = 0,1,2,\ldots,N-1$, where

$$\Delta k = (k_{\max} - k_0)/(N-1), \tag{62}$$

$$\Delta r = 2\pi/N(\Delta k), \tag{63}$$

$$k_m = k_0 + m(\Delta k), \quad m = 0,1,2,\ldots,N-1,$$

$$r_n = r_0 + n(\Delta r), \quad n = 1,2,3,\ldots,N.$$

Consider the case that c decreases from 335 m/s at the ground to 327 m/s at 8-m height. This is a sample model of shadow zone propagating with a noninfinite negative gradient up to 8 m. The source and the detector heights are 1 and 4 m, respectively. The detector is located in the shadow zone when $r > 71$ m. Regions near the ground are approximated by 16 layers of 0.5-m thickness. For source frequency $f = 40$ Hz, the pressure versus range curve calculated by FFP is shown in Fig. 26 (solid line). The quantity

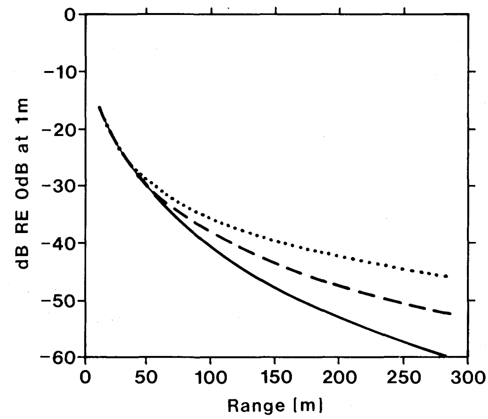

FIG. 26. Normalized pressure vs range for the inhomogeneous half-space sketched in Fig. 25 at a frequency of 40 Hz. In the diagram the dotted line is for the homogenous case, $s = 0$, the dashed line is for $s = 1.567 \times 10^{-3}$ m^{-1}, and the solid line is for $s = -2.985 \times 10^{-3}$ m^{-1}.

$$s = \Delta c / c \Delta h \tag{64}$$

is a measure of the speed of sound gradient. The curve for a homogeneous atmosphere ($c = 335$ m/s for all heights, $s = 0$) is shown by the dotted line. Much larger attenuation is predicted for the inhomogeneous case due to the upward refraction of the sound. In the region $r < 71$ m this reduction in level is principally caused by the reduction in grazing angle; beyond 71 m the detector is in the shadow zone and receives only diffracted sound energy. For ranges greater than 71 m, a simple ray trace treatment would predict infinite attenuation. In Fig. 26, a third curve is shown corresponding to an intermediate gradient $s = -1.567 \times 10^{-3}$ m^{-1}. As expected, this curve falls between the other two.

3. APPLICATIONS

Applications of atmospheric acoustics are quite varied. For purposes of this article only two will be considered. The first is the use of acoustics to determine the state of the atmosphere. Acoustic sensing of the atmosphere takes advantage of the fact that acoustic propagation is very sensitive to atmospheric conditions. The second, and most widespread, application is for prediction of noise levels. In addition to the nature of the source and propagation path discussed to this point, the receiver must be considered. When the receiver is a human listener, human perception of sound becomes an important factor. Based upon a knowledge of the source and transmission path, the sound level at the receiver can be accurately predicted. Human response to that sound, however, is much less predictable.

3.1 Acoustic Sensing of the Atmosphere (Brown, 1987)

The invention of the echosonde (also called the sodar), which McAllister patented in 1972, added a powerful tool for atmospheric research and for applications to forecasting especially in the atmospheric boundary layer. McAllister combined improved antenna design with the use of the facsimile recorder patented by Marti in 1919, which shows graphically the turbulent structure in the boundary layer. The use of the echosonde has become so widespread that results derived with it have become a ubiquitous part of the environmental impact studies required in many countries around the world. In its more complex form—the Doppler echosonde—the instrument can provide measurements of mean winds, sea breezes, and the flows in and out of valleys. An outgrowth of the instrument, called the echometer, shows promise of providing vertical profiles of mean temperatures as well. The monostatic echosonde relies on backscatter (scattering angle $\theta_s = \pi$) from a field of turbulence-driven temperature fluctuations. The bistatic echosonde ($\theta_s \neq \pi$) includes separated transmitter and receiver, with a signal that results from both temperature and velocity fluctuations. A usable return signal from a transmitted pulse in the monostatic geometry requires both a gradient in the temperature and a turbulent velocity field to shift air parcels to different positions, where the parcels will assume different actual temperatures T from their neighbors and produce a random field of refractive index fluctuations.

The equations of fluid mechanics, together with two assumptions valid in acoustic remote sensing, that the turbulence remains a nonzero-vorticity and incompressible field that produces no sound, while the acoustic wave remains a vorticity-free longitudinal compression field that produces no changes in the turbulence, lead to the governing acoustic remote-sensing equation. If the turbulence "eddy" sizes (here, an eddy represents a Fourier space component, not a physical swirl or blob) lie within the inertial range (tenths of meters to tens of meters), and Taylor's "frozen turbulence" hypothesis* holds, the "single scatter," or Born, approximation is valid. Then, the scattering cross section σ_s (in reciprocal meters) as a function of scattering angle θ_s becomes

$$\begin{aligned}\sigma_s(\theta_s) = {} & 1.52 k_0^{1/3} \cos^2(\theta_s) \\ & \times [0.13 C_n^2 + \cos^2(\theta_s/2) C_v^2/(4c^2)] \\ & \times [2 \sin(\theta_s/2)]^{-11/3}, \end{aligned} \qquad (65)$$

with k_0 the initial wave number, C_v^2 the structure parameter for velocity fluctuations ($\mathrm{m}^{4/3}\,\mathrm{s}^{-2}$), and C_n^2 the structure parameter for refractive index fluctuations ($\mathrm{m}^{-2/3}$) given by

$$\begin{aligned} C_n^2 = {} & C_T^2/4T_0^2 + 2(0.307) C_{eT}/4P_0T_0 \\ & + (0.307)^2 C_e^2/4P_0^2, \end{aligned} \qquad (66)$$

where subscript e refers to water vapor, and T to temperature, fluctuations (C_T^2 and C_{eT} are in $\mathrm{K}^2\,\mathrm{m}^{-2/3}$ and $\mathrm{K\,N\,m}^{-8/3}$, respectively). One should note that a single fluctuation would have negligible effect on the scattered wave. However, the nonzero correlation and cross-correlation distances for turbulence velocity, vapor pressure, and temperature fluctuations imply a quite different kind of interaction than that from a field of randomly spaced point scatterers. Therefore, wavelets scattered from nearby regions can interfere constructively and destructively to produce a partially coherent scattered wave resulting from the integration of many wavelets over a nonzero region of space. The same physics underlies the importance of the spatial Fourier transform decomposition of the random fluctuation field.

A typical monostatic echosonde employs a single antenna for both transmission and reception, with an electronic switch to alternate between the two. The antenna includes a high-power pressure transducer (similar to those used in airport public address systems)

---

*Taylor's "frozen turbulence hypothesis" maintains that temporal changes in local meteorological parameters caused by turbulence are related to spatial changes through the mean wind.

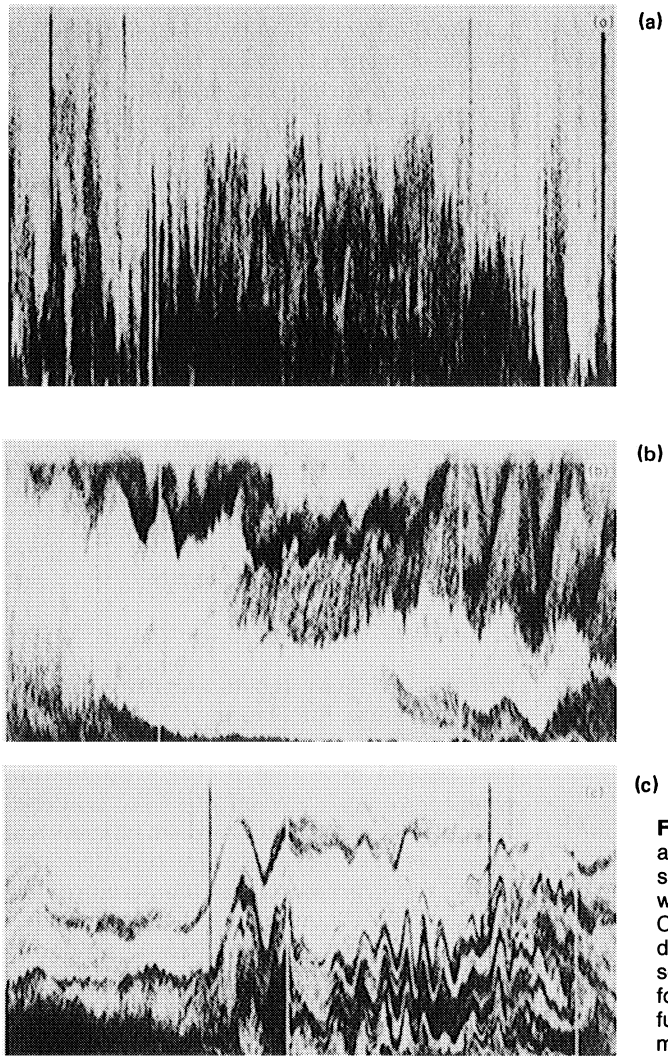

FIG. 27. Echosonde records of the boundary layer. Vertical scale: 0–500 m; horizontal scale: time, increasing to the right; vertical white streaks are hour markers. (Courtesy of C. W. King.) (a) Daytime thermal plumes; (b) decreasing ground layer (below) and descending inversion (above) in early evening, followed by triggering of internal waves; (c) fully developed internal wave activity after midnight.

together with an exponential feed horn (for impedance matching) mounted at the focus of a parabolic dish, with the entire assembly enclosed in a shield, or "cuff," made from plywood lined with an acoustic foam that includes a lead-sheet septum. The cuff provides substantial side-lobe suppression and, if the sides flare, an increase in the effective aperture of the antenna. Typical antennas achieve angular beamwidths of 8°–10°. The reason for using the same antenna for reception rather than a more sensitive microphone lies in this narrow beamwidth and the resulting much greater signal-to-noise ratio.

Figure 27 shows a sequence of boundary layer structures over a typical diurnal period. (Time in the figure moves from left to right.) The top of the figure corresponds to an altitude of 500 m; vertical thin white stripes represent hour markers. Figure 27(a) shows a typical unstable, daytime record of thermal plumes carried through a vertical echosonde beam by the mean wind. Such plumes result from the conductive random heating of air near the ground, itself heated by solar radiation. Randomness may result from variations in surface albedo and from the presence of clouds. Figure 27(b) shows the boundary layer in the late afternoon and early evening. The decrease and then disappearance of solar radiation and the consequent cooling of the ground brings about the vanishing of the thermal plumes, a concurrent creation of a ground-based shear layer that on a finer time scale would show a "herringbone" structure, and the slow descent of an elevated inversion previously off the top of the record. Near its end, the figure shows the beginning of internal gravity wave activity. In Fig. 27(c) a series of large-scale internal gravity waves dominate the record. Especially noteworthy, the particle displacement double amplitude (peak to peak) reaches values of the order of 300 m. In the descending part of the cycle such large displacements require a horizontal outflow near the ground that would appear similar to wind gusts, except for the regularity of their return with the period of the waves.

Although the bistatic configuration has many other uses, its greatest value lies in vertical profiles of Doppler measurements of mean winds. Figure 28 outlines the geometry of one leg of such a system. (With two horizontal legs at right angles to each other and the addition of a vertical Doppler measurement, one can map the entire three-dimensional wind field at all heights within the echosonde range.) A number of different arrangements have shown various advantages and drawbacks. The following describes a particular design that seems one of the more flexible and dependable and has the advantage that the vertical narrow receiver "pencil beam" rejects much of the noise coming in on low-elevation paths from road traffic, distant aircraft, and residential areas. The typical transmitter antenna uses a "fan beam" geometry, one that sprays an initial pulse over a range of angles. Time-gating of the received signal then gives a Doppler value at each desired height. Other configurations use a set of three monostatic echosondes pointing in three different directions, allowing greater simplicity in the hardware but producing greater errors when the atmosphere lacks horizontal homogeneity because of the separations between the three different scattering volumes. The wind velocity will be denoted by a vector **w** or a Mach vector defined as $\mathbf{m}=\mathbf{w}/c$. For the monostatic Doppler echosonde, when **m** is directed along the beam, and with no refraction, the frequency shift Δf satisfies

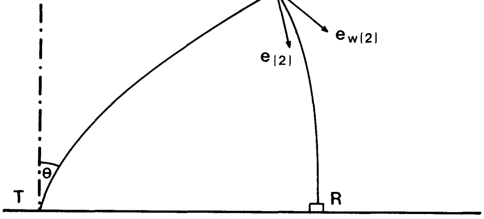

FIG. 28. Geometry for Doppler wind measurements. Transmitter *T* and receiver *R*. Scattering volume center *S* at height *z*. Other symbols defined in the text.

$$\Delta f/f=\pm 2m, \tag{67}$$

where the plus sign holds when the wind blows toward, and the minus sign away, from the antenna, and $m=|\mathbf{m}|$. For the bistatic system, shown in Fig. 28 (together with some simplifying assumptions, such as horizontal stratification and negligible vertical wind or cross wind) the horizontal component of the Mach vector m_x, with $m_x = \pm|m_x|$, and with an obvious choice of sign, satisfies

$$\Delta f/f=-m_x \sin\theta+(m_x \sin\theta)^2. \tag{68}$$

Frequently, an average over, say, 20 min will cancel the effect of any vertical wind; that is, one can consider it zero. However, a complete horizontal wind measuring system may often face conditions where cross winds on one leg or the other become important, or some other simplifying assumption no longer holds.

Researchers prefer various methods for determining the frequency shift Δf from the raw data, especially to cope with the ubiquitous presence of noise. In the frequency domain, many use the FFT (fast Fourier transform), at the same time trying to measure or estimate the spectral contribution of the noise. However, the FFT requires a fairly powerful computer. In the time domain, the CXCV (complex covariance) method provides good data with a much less expensive microprocessor but requires careful choice of sampling frequency and filter bandpass; a choice of sampling frequency such that the Nyquist frequency remains near the filter half-power point minimizes aliasing.

With the increasing amount of inexpensive computer memory becoming available, echosonde designers now have the choice of turning to completely coherent signal processing, for example; averaging the signal in both amplitude and phase, and, thus, reducing random noise to a negligible problem. The noise of rain hitting the receiving antenna makes echosonde observations impossible during a storm. In contrast, the temperature-measuring echometer, by virtue of the (here necessary) coherent processing, operates successfully through a rain squall, producing data in close agreement with those from a nearby *in situ* thermometer on a tower. No obvious reasons exist that would rule out similar coherent processing of echosonde signals to obtain a large increase in signal-to-noise ratios.

The boundary layer can have either an "open" structure for the unstable case, acting as a pathway for fluxes of momentum, heat, and water vapor to reach higher levels and perturb mesoscale processes, or a "closed" structure for the stable case, acting as an impenetrable container disconnected from large-scale events. In the first case, information on the structure may have an impact on future weather-prediction methods. In the second, boundary layer data should assist in predicting the strength and duration of air pollution events. Many of the parameters needed, however, derive from higher-order measurements that have a slightly shaky theoretical basis, and experimental values of unknown precision. Thus, they represent work still in progress, on the leading edge of the research in acoustic remote sensing.

Among other parameters, dependable measurements of mean temperature profiles $T(z)$ still elude researchers. One method, called RASS (radio acoustic sounding system), uses an acoustic pulse containing of the order of 50 phase fronts to reflect a radar signal coherently back to a receiver. With an acoustic wavelength equal to one-half the radar wavelength, strong returns arrive from altitudes throughout the boundary layer. The Doppler shift in the radar return derives—with a correction for mean vertical wind w_z—from the speed of the acoustic waves.

When an elevated inversion lies above an unstable atmosphere near the ground, the height z_i of the inversion becomes a useful parameter. The upper part of the boundary layer, the well-mixed layer, as its name implies, has very little fluctuation in temperature except in the thermal plumes and thus shows no echosonde returns elsewhere, even if (velocity) turbulence exists. Within the inversion the atmosphere remains stable, that is, without turbulence, even though the temperature T changes, and thus again the echosonde shows no returns. However, the rising air in the plumes has a tendency to overshoot and, before descending, entrain air with a different temperature from the bottom of the inversion in its turbulent motion. As a result echosonde facsimile records frequently show a dark elevated band corresponding to this entrainment layer above the region of thermal plumes. Comparisons with radar or FM-CW radar seem to support good agreement between their measurements of the base of an elevated inversion z_i and the echosonde measurement of the height of the bottom of the dark band produced by an entrainment layer.

3.2 Perception of Sound

Measurements and measures associated with atmospheric acoustics are typically related to the hearing response of the human ear. Such a treatment is especially applicable when dealing with acoustic noise and noise control. The loudness level is defined as the sound pressure level of a 1000-Hz tone that sounds as loud as the sound in question (see ACOUSTICS, PSYCHOLOGICAL). The unit of loudness is the phon. The famous Fletcher–Munson equal-loudness contours shown in Fig. 29 give an indication of the loudness level at different sound pressure levels. Note that a sound pressure level of 70 dB at 25 Hz has the same loudness as a 10-dB tone at 1000 Hz.

When measuring noise in the atmosphere, it is common to express the level in dBA, dBB, or dBC. A sound level meter with these ratings has a frequency response given in Fig. 30. The A-weighted curve approximates the hearing response shown in the previous figure for relatively low sound pressure levels ($\sim$40 dB). The B weighted curve is applicable at higher levels (80–90 dB). C-weighted sound levels show less frequency compensation and, therefore, are a better measure of sound pressure level.

Measures of noise are designed to give an indication of human response. Typically, the

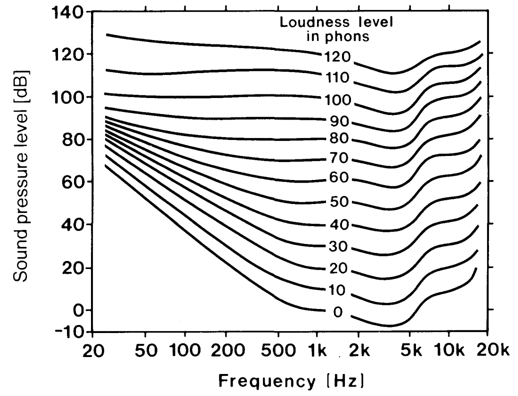

FIG. 29. Equal-loudness contours for pure tones by Fletcher and Munson. (American Standard for noise measurement, Z24.2-1942.)

goal is to reduce annoyance. Since different types of sound present different annoyances at the same C-weighted instantaneous level, a variety of noise measures have been advanced. Some of the more commonly used are perceived noise level (PNL), sound exposure level (SEL), and day-night sound level (L_{dn}).

In contrast to the A level, the perceived noise level weighting function is based upon subjective assessment of the noisiness of the various frequency components of the signal, rather than upon *loudness* considerations. There is more emphasis on the upper portion of the noise spectrum (2000–4000 Hz) in this weighting function than in the A-weighting function. The perceived noise level, in units of PNdB, has been used for many years in the U.S. as a measure of aircraft noise.

Subjective tests indicate that human response to noise is not only a function of the maximum level, but of the duration of the signal and its temporal variation. Time-related changes may range from a sound level constant over time, as produced by a continuously operating machine, to the typical haystack-shaped time history produced by an aircraft flyover, to the constantly varying noise level perceived near highways.

Historically, several methods have been used to introduce time characteristics. With recent advances in electronics and instrumentation technology, there are now instruments which can integrate, or sum, noise signals as a function of time. (Integrated noise levels are a measure of the physical energy in the noise signal.)

Significant evidence indicates that two signals with equal sound energy will produce the same subjective response. For example, a noise with a constant level of 85 dB occurring for 10 min would be judged equally as annoying as an 82-dB noise signal lasting for 20 min, i.e., one-half the energy lasting for twice the time period. This is known as the "equal energy" principle. In practice, the integration process is often replaced by a summation of levels occurring at one-half second intervals over the upper 10 dB of the noise signal.

Integration of the A-weighted noise level over the period of a single event (such as an aircraft flyover) gives the sound exposure level, in dB. The day-night average sound level (referred to as L_{dn}) is obtained by energy-averaging noise levels over a 24-hour period, with a 10-dB penalty to nighttime noise levels. The day-night average sound level can be applied to all sources of noise. For discrete event noise sources, such as aircraft operations, L_{dn} may be computed by decibel summation of noise levels occurring during daytime and nighttime periods (with the nighttime penalty). Areas subjected to L_{dn} in excess of 75 are useful only for selected commercial applications. Areas of a 70–75 L_{dn} may be used for a wider variety of commercial applications while areas impacted by 65–70 L_{dn} may be used for most applications including residential with proper sound treatment. Below 65 L_{dn} some treatment may be required for uses requiring a quiet background.

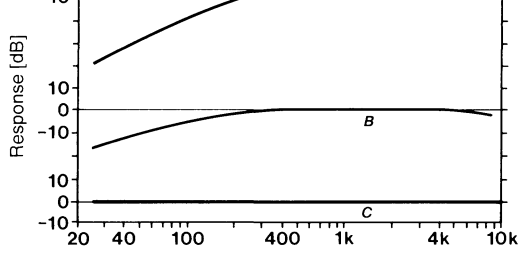

FIG. 30. Frequency-response characteristics taken from the American Standard for sound-level meters, Z24.3-1944.

Combining perception with propagation efforts described earlier provides practical guidelines for predicting community objections to different types of noise sources. In most cases involving community noise (Makarewicz, 1987), the source and receiver heights are much smaller than the source–receiver separation so $|z|\cos\theta_i \ll 1$. For this geometry, the A-weighted mean-square sound pressure at distance d can be given as

$$P_A^2 = \widetilde{P}_A^2[S(\sigma_e,\mu,\beta d)/d^2], \qquad (69)$$

where $\widetilde{P}_A^2$ is the A-weighted mean-square sound pressure for the direct wave, σ_e is the effective flow resistivity of the surface, β is the absorption coefficient α divided by the frequency f, and μ is a parameter which describes the source spectrum $[P(f) = P(0)\exp(-\mu f)]$. Far away from a source ($d \gg 500$ m), $P_A^2 \sim d^{-7}/1$ m which suggests that the A-weighted sound level will decrease at a rate of 21 dB per doubling of distance. In practice, turbulence does not allow complete cancellation of direct and reflected waves so such a rapid falloff is not expected. Nonetheless, atmospheric absorption and surface effects combine to give very rapid attenuation of sources close to the horizon.

GLOSSARY

(For complete definition of commonly used acoustical terms see American National Standards Institute, ANSI S1.1-1960 [R 1976])

Absorption Loss: That part of the transmission loss due to the dissipation or conversion of sound energy into other forms of energy.

Admittance: Inverse of impedance.

atm: One standard atmosphere of pressure $= 1.01 \times 10^6$ dyn/cm$^2 = 1.01 \times 10^5$ Pa $= 1.01 \times 10^5$ N/m^2.

Attenuation Coefficient: Logarithmic rate of change in sound pressure (dB/unit length or neper/unit length).

Creeping Wave: Dominant term in an approximate solution (applicable at long ranges) for sound diffracted into a shadow zone.

Diffracted Wave: One whose front has been changed in direction by an obstacle or other inhomogeneity in a medium.

Echo Sounder (Acoustic Sounder): A device which measures atmospheric conditions by sending and receiving acoustic pulses.

Excess Attenuation: Attenuation in excess of 6 dB/doubling of distance (+dB means added loss).

FFT: Fast Fourier transform.

Impedance: The complex ratio of the sound pressure to the particle velocity. The specific impedance, the ratio of impedance to impedance of air, is generally used.

Infrasound: Sound at frequencies less than 20 Hz.

Insertion Loss: The increase in transmission loss caused by the addition of a barrier.

Mach Number: Ratio of speed to local speed of sound.

neper: Unit of level when the logarithm is on the Naperian base.

Nyquist Frequency: Highest frequency allowed in a digitally sampled signal.

Octave Band: Band of frequencies extending from $f_c/\sqrt{2}$ to $\sqrt{2}\,f_c$, where f_c is the center frequency in Hz.

Pascal: mks unit of measure for pressure (1 N/m^2).

Radio Sounder: Similar to echo sounder but the acoustic wave is monitored with a radar system.

rayl: Unit of measure for impedance; 1 mks rayl $= 1$ kg/m^2 s $= \frac{1}{10}$ cgs rayl.

Rayleigh Waves: Surface waves associated with the free boundary of a solid such that a surface particle describes an ellipse.

Refraction: The process by which the direction of sound propagation is changed due to spatial variation in the speed of sound.

Scattering: Acoustic scattering is the irregular reflection, refraction, or diffraction of sound in many directions.

S/N: Signal-to-noise ratio typically measured in dB.

Sound Pressure Level: 20 times the logarithm to the base 10 of the ratio of the pressure to a reference pressure (conventionally 20 μPa).

Standard Temperature and Pressure: (STP) 273 K, 1 atm. Note that in atmospheric physics the standard atmosphere at sea level (288 K, 1 atm) is commonly used.

Transmission Loss: Decrease in sound pressure level between two points measured in dB re 20 μPa.

Turbule: A single element of turbulence defined in a statistical sense.

Ultrasound: Frequencies above 20 kHz.

List of Works Cited

American National Standards Institute (ANSI) S2.20-1983, *Estimating Airblast Characteristics for Single Point Explosions in Air with a Guide to Evaluation of Atmospheric Propagation and Effects*, New York: American National Standards Institute.

Arnold, R. T., Bass, H. E., Bolen, L. N. (1976), *J. Acoust. Soc. Am.* **60**, 584–593.

Attenborough, K. (1985), *J. Sound Vib.* **99**, 521–544.

Attenborough, K., Hayek, S. I., Lawther, J. M. (1980), *J. Acoust. Soc. Am.* **68**, 1493–1501.

Bass H. E. (1980), *J. Acoust. Soc. Am.* **67**, 1959–1966.

Bass, H. E., Sutherland, L. C., Piercy, J., Evans, L. (1984), in: *Physical Acoustics XVII*, New York: Academic, pp. 145–232.

Berengier, M., Daigle, G. A. (1988), *J. Acoust. Soc. Am.* **84**, 1055.

Brown, E. H. (1987), *Encyclopedia of Physical Science and Technology 3*, New York: Academic, pp. 147–163.

Brown, E. H., Hall, F. F., Jr. (1978), *Rev. Geophys. Space Phys.* **16**, 47–110.

Busnel, R. G. (1977), in: T. A. Sebeok (Ed.), *How Animals Communicate*, Bloomington: Indiana University.

Chessel, C. I. (1977), *J. Acoust. Soc. Am.* **62**, 825–834.

Daigle, G. A. (1982), *J. Acoust. Soc. Am.* **71**, 847–854.

Delany, M. E., Bazley, E. N. (1977), *J. Sound Vib.* **16**, 315–322.

Dept. of the Army (15 June 1978), *Dept. of the Army Technical Manual 5-803-2*.

Isei, T., Embleton, T. F. W., Piercy, J. E. (1980), *J. Acoust. Soc. Am.* **67**, 46–58.

Johnson, M. A., Raspet, R., Bobak, M. T. (1982), *J. Acoust. Soc. Am.* **81**, 638–646.

Lindsey, R. B. (1974), *Physical Acoustics*, Benchmark Papers in Acoustics, 69. New York: American Institute of Physics.

Makarewicz, R. (1987), *J. Acoust. Soc. Am.* **82**, 1706–1711.

Pierce, A. D. (1981), *Acoustics: An Introduction to its Physical Principles and Applications*, New York: McGraw-Hill.

Piercy, J. E., Embleton, T. F. W., Sutherland, L. C. (1977), *J. Acoust. Soc. Am.* **61**, 1403–1418.

Price, M. A., Attenborough, K., Heap, N. W. (1988), *J. Acoust. Soc. Am.* **84**, 1836–1844.

Raspet, R., Lee, S. W., Kuester, E., Chang, D. C., Richards, W. F., Gilbert, R., Bong, N. (1985), *J. Acoust. Soc. Am.* **77**, 345–352.

Raspet, R., Bass, H. E., Ezell, J. (1983), *J. Acoust. Soc. Am.* **74**, 267–274.

Reed, J. W. (1977), *J. Acoust. Soc. Am.* **61**, 39–47.

Reed, J. W. (1987), *J. Geophys. Res.* **92**, 11 979–11 992.

Sabatier, J. M., Bass, H. E., Bolen, L. N., Attenborough, K., Sastry, V. V. S. S. (1986), *J. Acoust. Soc. Am.* **79**, 1345–1352.

Smith, D., Schien, D. (1989), Description of jet aircraft noise provided by employees of the Northrop Corporation.

Tyndall J., (1874), *Philos. Trans. R. Soc. London, Ser. A* **164**, 183–244.

Wei, R. J., Tian, Y. R., Lu, Q. J. (1987), *J. Acoust. Soc. Am.* **81**, 1350–1354.

Further Reading

Bass, H. E., Sutherland, L. C., Piercy, J., Evans, L. (1984), in: *Physical Acoustics XVII*, New York: Academic, pp. 145–232.

Brown, E. H., Hall, F. F., Jr. (1978), *Rev. Geophys. Space Phys.* **16**, 47–110.

Busnel, R. G. (1977), in: T. A. Sebeok (Ed.), *How Animals Communicate*, Bloomington: Indiana University.

Pierce, A. D. (1981), *Acoustics: An Introduction to its Physical Principles and Applications*, New York: McGraw-Hill.

Piercy, J. E., Embleton, T. F. W., Sutherland, L. C. (1977), *J. Acoust. Soc. Am.* **61**, 1403–1418.

ATMOSPHERIC PHYSICS

Fredric W. Taylor, *Dept. of Atmospheric, Oceanic, and Planetary Physics, Oxford University, Oxford, England*

	Introduction	181
1.	**Fundamentals**	182
1.1	Definitions	182
1.2	Origin and Evolution of the Atmosphere	182
2.	**Atmospheric Structure**	183
2.1	Thermal Structure and Nomenclature	183
2.2	Clouds and Precipitation	185
3.	**Atmospheric Composition**	185
3.1	Principal Constituents	185
3.2	Trace Gases, Pollution, and Photochemistry	185
3.3	The Ozone Layer	186
3.4	The Ionosphere	186
4.	**Radiative Transfer in the Atmosphere**	186
4.1	Solar Radiation	186
4.2	Thermal Radiation	187
4.3	The Greenhouse Effect	187
5.	**Atmospheric Dynamics**	188
5.1	The General Circulation	188
5.2	Waves and Turbulence	188
5.2.1	Gravity Waves	189
5.2.2	Rossby Waves	189
5.2.3	Turbulence	189
6.	**Planetary Atmospheres**	189
6.1	The Terrestrial Planets	189
6.1.1	Composition	189
6.1.2	Clouds	190
6.1.3	Dynamics	191
6.2	The Outer Planets	191
6.2.1	Composition	191
6.2.2	Clouds	192
6.2.3	Dynamics	192
6.3	Satellites with Atmospheres	193
7.	**Atmospheric Observations**	193
7.1	Ground-Based and Airborne Observations	193
7.1.1	The Dobson Ozone Spectrophotometer	193
7.1.2	The Radiosonde	193
7.2	Measurements from Space	193
7.2.1	Remote Temperature Sounding	194
7.2.2	Vertical Resolution	194
7.2.3	Composition Measurements	194
8.	**Climate and Climate Change**	195
8.1	The Ozone Layer and Ozone Depletion	195
8.1.1	Production and Loss Mechanisms for Ozone	195
8.1.2	Anthropogenic Effects	195
8.2	The Greenhouse Effect	196
8.3	Prospects for Forecasting Climate	198
	Glossary	198
	Works Cited	199
	Further Reading	199

INTRODUCTION

Atmospheric physics deals with the application of physical laws to the study of the structure, composition, and motion of the atmosphere. It forms the basis for, but is not limited to, the study of weather and climate (see METEOROLOGY; WEATHER FORECASTING); for this reason, a major goal of the discipline is to develop an understanding of atmospheric behavior that is good enough to allow predictions of the evolution of the atmosphere over time scales from days to decades, on spatial scales from a few kilometers to the whole globe.

Attempts to understand the effects of the atmosphere on man's environment date back to the origin of science itself. At the present time, considerable attention is being focused by the public, politicians, and media on fears that the habitability of the Earth is being altered as an inadvertent side-effect of increasing industrialization. As a result, atmospheric physics is one of the most rapidly growing and widely discussed areas of applied physics at the present time.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

In the simplest physical terms, atmospheric science is basically the study of a relatively thin, radiatively heated and cooled layer of mixed gases on a large, rapidly rotating sphere with an uneven surface and with chemical activity and all three phases of water present. Because it deals with basic processes, atmospheric physics as a field is not restricted to the single example of the Earth. Eight bodies in the solar system are known to have substantial atmospheres and exhibit structure and behavior analogous to that on the Earth. Information about these derived from astronomical observations and space experiments can be used to make powerful tests of theories and models of atmospheric systems.

The three main cornerstones of the discipline are the subfields of atmospheric radiation (that is, electromagnetic radiation from the sun and terrestrial infrared radiation, and their quantum interactions with atmospheric molecules), geophysical thermodynamics, and gas-phase chemistry, especially photochemistry. The coupling between the various processes in these categories is extensive and important, and progress with a theoretical understanding of atmospheric processes, as well as attempts at prediction (weather and climate forecasting), depends to an increasing extent on vast computer models. The experimental side of the field is now dominated by the development of increasingly large and complex optical instruments, radiometers and spectrometers, deployed in space on satellites. Such experiments produce formidable quantities of raw data, and processing and analyzing the information from satellites also call for a very high level of computing power and sophistication.

Thus the field of atmospheric physics unites many areas of basic physics and engineering to address a topic of great complexity and relevance to society.

1. FUNDAMENTALS

1.1 Definitions

Atmospheres are the envelopes of gas that surround a planet or large satellite. In the case of the Earth, the atmosphere forms a very thin skin, of the order of 100 km in depth, and has a mass of 5.27×10^{18} kg, about one-millionth that of the solid body. This amount of gas is sufficient to exert a mean pressure p_0 of $1.013\,25\times10^5$ pascals (often called 1 atmosphere) on the surface. Other common units of pressure are those based on the *bar*, which is defined to be equal to 10^5 Pa; atmospheric pressures are most frequently quoted in millibars (mbar). The pressure falls off with height according to the formula

$$p=p_0\exp\left(-\int_0^z \frac{1}{H}dz\right), \tag{1}$$

which is easily derived from the perfect gas law and the equation of state. In Eq. (1), the quantity H, the height in which the pressure falls by a factor of e^{-1}, is called the *scale height* (strictly the *pressure* scale height, since other scale heights are sometimes defined; for example, that describing the change with height of the abundance of a species that is not uniformly mixed). H is equal to RT/Mg (M=mean molecular weight, R=gas constant, g=acceleration due to gravity) and so H varies with the temperature T and humidity of the air, but is typically about 8 km. In the approximation H=const, Eq. (1) reduces to $p=p_0e^{-z/H}$, and height is proportional to $\log_e$(pressure). The latter is often a convenient vertical coordinate in atmospheric calculations and sometimes known as the pseudoheight.

1.2 Origin and Evolution of the Atmosphere

The Earth's atmosphere is believed to be of secondary origin (Henderson-Sellers, 1983). This means that the gases surrounding the solid planet at the time it condensed were lost to space at an early stage and the present atmosphere subsequently formed as a result of outgassing from the crust. This process was augmented to an unknown degree by the infall of icy material as comets. The primitive atmosphere thus formed would have had a composition quite different from that which exists today, consisting primarily of carbon dioxide, methane, ammonia, and water vapor with no free nitrogen or oxygen. The large proportion of nitrogen now present arose as a consequence of the photodissociation of ammonia into nitrogen and hydrogen. The latter gas is so light that it can escape from the Earth even at the present time, whereas most heavier gases are gravitationally trapped now that the Earth is cooler and the Sun less active. Most of the methane was oxidized to form carbon dioxide and water vapor, although enough is produced (for example, by the decay of vegetable matter) to leave a substantial admixture in the atmosphere to the present day. Free oxygen exists as a consequence of the presence of life in the form of plants, which convert carbon dioxide into oxygen by photosynthesis. Thus, we owe the present state of the atmosphere to a series of complex and dynamic processes. There is evidence that the atmosphere is not entirely stable but continues to evolve (Sec. 8), perhaps towards a state less well suited to human comfort and survival.

2. ATMOSPHERIC STRUCTURE

2.1 Thermal Structure and Nomenclature

The various regions of the atmosphere are named for their temperature structure (Fig. 1). In simplified terms, these regions exist for the following reasons. Much of the electromagnetic radiation emitted by the Sun and reaching the Earth is at wavelengths in or near the visible part of the spectrum. These are not strongly absorbed by the atmosphere and so, unless there are clouds to scatter the photons and reflect a portion of the radiation back to space, most of the energy reaches the ground where it is absorbed. The lower atmosphere is then heated by the ground and will become unstable against convection if a layer nearer to the ground is warmer than that above it by a sufficient amount (calculated below) to become less dense in spite of the vertical pressure gradient. Convective instability exists in the lowest 10 km or so of the atmosphere, which is known as the troposphere ("turning region"). The upper boundary is the level where the overlying atmosphere is of such a low density that a substantial amount of radiative cooling to space can occur in the thermal infrared region of the spectrum. At this level, called the *tropopause*, radiation cools rising air so efficiently that the temperature tends to become constant with height and convection ceases. The tropopause varies in height over about 6 km with latitude, being highest (around 16 km) in the tropics, where solar heating is greatest, but it is generally quite a sharp feature in the temperature profile everywhere.

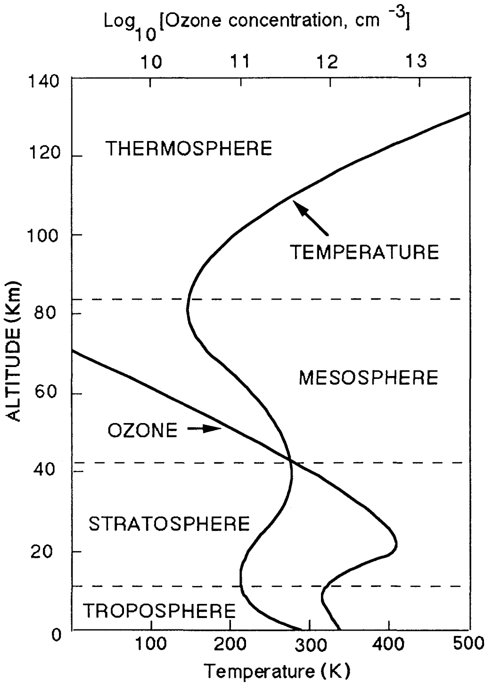

FIG. 1. Standard temperature and ozone profiles, showing the naming of the regions. After NASA (1986).

Assuming that hydrostatic equilibrium applies, then

$$dP = -g\rho\, dz. \tag{2}$$

Furthermore, if we assume that there is no net exchange of energy between a parcel of air and its surroundings, then the specific heat at constant volume c_v, temperature T, pressure P, and volume V are related by

$$c_v\, dT + P\, dV = 0. \tag{3}$$

Using the perfect gas law $PV = RT/M$ (R = gas constant per mole, M = molecular weight) and the relationship $c_p - c_v = R/M$ (c_p = specific heat at constant pressure), we find that the temperature gradient with height dT/dz is given by

$$\frac{dT}{dz} = -\frac{g}{c_p}. \tag{4}$$

This gradient, called the *lapse rate*, is constant and, since c_p is equal to about 1000 J kg^{-1} K^{-1}, equal to approximately 10 K km^{-1}

for dry air. For moist air, dT/dz is less and can be as small as 3 K km^{-1}. A useful average value for Earth is 8 K km^{-1}. Figure 2 shows a realistic temperature profile together with the steeper profile that would exist if there were no vertical convection.

The lapse rate above the tropopause, where convection stops, tends to zero (i.e., constant temperature with height) because there is no longer enough absorption above the layer to stop emitted photons reaching space. Then each layer is heated by radiation from the optically thick atmosphere below and cooled by radiating to space; to first order, height is no longer important. This region is called the stratosphere (it is stratified in the sense that density decreases monotonically with height, and therefore the layers do not try to move up or down through each other as in the troposphere).

If the stratosphere is modeled as an optically thin slab of (vanishingly small) emissivity ϵ, its temperature T_{strat} can be related to the effective radiative, or equivalent blackbody, temperature of the Earth, T_{Earth} (see Sec. 8.2). The stratosphere is heated from below by a flux proportional to $\epsilon T_{\text{Earth}}^4$ and cooled by identical fluxes in the upward and downward directions, each proportional to $\epsilon T_{\text{strat}}^4$. Thus we must have $\epsilon T_{\text{Earth}}^4 = 2\epsilon T_{\text{strat}}^4$ whence

$$T_{\text{strat}} = \frac{T_{\text{Earth}}}{2^{1/4}} = \frac{250}{2^{1/4}} \approx 210 \text{ K}. \tag{5}$$

Although these simple arguments suggest that the stratospheric temperature remains constant to a great height, in fact it is observed to increase above about 20 km (Fig. 2) as a consequence of the absorption of solar ultraviolet radiation by photochemically produced ozone in the stratosphere (Sec. 3.3). Ozone absorbs in the Hartley band, which forms a continuum from 0.2 to 0.3 μm. Below 70 km, virtually all of the energy absorbed is converted to heat.

The ozone concentration in the stratosphere peaks near 25 km (Fig. 1), but a calculation of the heating rate at height z where $\rho(z)$ is the absorber amount,

$$c_p \rho(z) \frac{dT(z)}{dt} = \int_{0.2}^{0.3} \frac{dF_\lambda(z)}{dz}\, d\lambda, \tag{6}$$

where

$$F_\lambda(z) = F_\lambda(\infty) \exp\left(- \int_z^\infty k_\lambda \rho(z) dz \right)$$

is the flux of radiant energy, and k_λ the absorption coefficient at wavelength λ, finds that this peaks at a height of about 50 km. The temperature is also maximum at this level, which is known as the *stratopause.* Above the stratopause, the temperature declines again, reaching an absolute minimum at the *mesopause* near 85 km. The pressure here is only a few microbars. With such low densities of gas above, very energetic solar photons in the extreme ultraviolet range, and particles too, penetrate into the region, causing ionization and dissociation and releasing kinetic energy. The heating thus produced causes the temperature to increase rapidly with height, leading to the name *thermosphere.*

Between the tropopause, where large-scale vertical convection ceases, and the mesopause, the atmosphere remains fairly well mixed by turbulence produced by a variety of instabilities in wave motions and the mean flow (Sec. 5.2). At a fairly small distance up

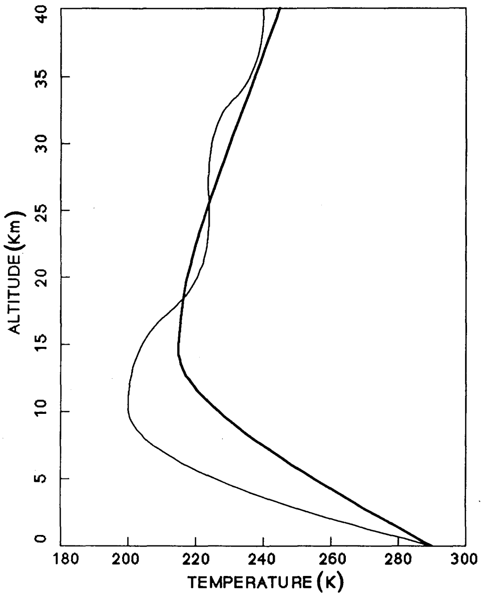

FIG. 2. Atmospheric temperature profile calculated assuming only radiative processes (light line), compared to the observed mean global profile (heavy line). After Kuhn (1985).

into the thermosphere, around the 100-km altitude level, diffusion takes over as the dominant process, and the atmosphere starts to separate into its lighter and heavier components. For many practical purposes, this level (the *homopause*) may be considered to be the effective top of the atmosphere.

2.2 Clouds and Precipitation

Substantial clouds of water droplets are usually confined to the lowest scale height of the atmosphere (0 to 8 km), although very thin "noctilucent" clouds have been observed as high as the mesopause (85 km). Clouds fall into many different classifications, but nearly all are formed by some version of the same basic process: convection upwards of moist air and subsequent condensation of water droplets or ice crystals around dust nuclei (Mason, 1974).

A typical cloud droplet is about 10 μm in diameter, while a typical raindrop is 100 times larger, i.e., contains as much water as a million cloud particles. The process by which cloud droplets grow is basically diffusion and condensation of water vapor until they reach about 30 μm diameter, then collision and coalescence, until they eventually rain out.

In addition to their obvious importance in meteorology and the terrestrial water cycle, clouds play an important role in controlling the global climate by scattering, absorbing, and emitting radiation. They interact strongly with both incoming solar and outgoing thermal (planetary) radiation, and so are an important component of the atmospheric greenhouse (Sec. 4.3). Because the radiative-transfer properties of clouds depend on variable and relatively inaccessible quantities such as the number density and size distribution of particles, their role is a particularly difficult one to quantify in climate models (Sec. 8).

3. ATMOSPHERIC COMPOSITION

3.1 Principal Constituents

Table 1. Composition of clean dry air near the surface.

Species	Mixing ratio (by volume)
Nitrogen, N_2	0.780 83
Oxygen, O_2	0.209 47
Argon, Ar	0.009 34
Carbon dioxide, CO_2	0.000 35
Neon, Ne	18.2×10^{-6}
Helium, He	5.2×10^{-6}
Krypton, Kr	1.1×10^{-6}
Xenon, Xe	0.1×10^{-6}
Ozone, O_3	0.4×10^{-6}
Hydrogen, H_2	0.5×10^{-6}
Methane, CH_4	2×10^{-6}
Nitrous oxide, N_2O	0.3×10^{-6}
Carbon monoxide, CO	0.1×10^{-6}

Table 1 shows the composition of dry air, the basic reasons for this particular mixture having been outlined above (Sec. 1.2). The amount of water that is present in addition to these noncondensible species varies, but can be as high as a few percent. Some of the other species, particularly ozone, are also very variable even in nonpolluted environments.

3.2 Trace Gases, Pollution, and Photochemistry

In addition to the principal gases nitrogen and oxygen and the common *minor* constituents argon, methane, carbon dioxide, and water vapor, the atmosphere contains a number of important *trace* constituents. Many of these are of anthropogenic origin, that is to say they are produced by industrial processes of various kinds at the surface. Well-known examples are the oxides of nitrogen (which are present naturally, but are being enhanced by production from artificial fertilizers and automotive and aircraft engines) and the chlorofluorocarbons used as refrigerants and in various other ways (see AIR POLLUTION: COMPONENTS, CAUSES, AND CURES). Gases such as nitrous oxide (N_2O) and trichloromonofluoromethane (CCl_3F) are source molecules for stratospheric concentrations of products produced by photochemistry in the stratosphere, for example, nitric oxide (NO) and chlorine oxide (ClO). Many other trace species have been observed or postulated to exist that have key roles in important atmospheric chemical processes, for example, those that maintain the ozone layer (Sec. 8.1). Some of them are shown in Fig. 3.

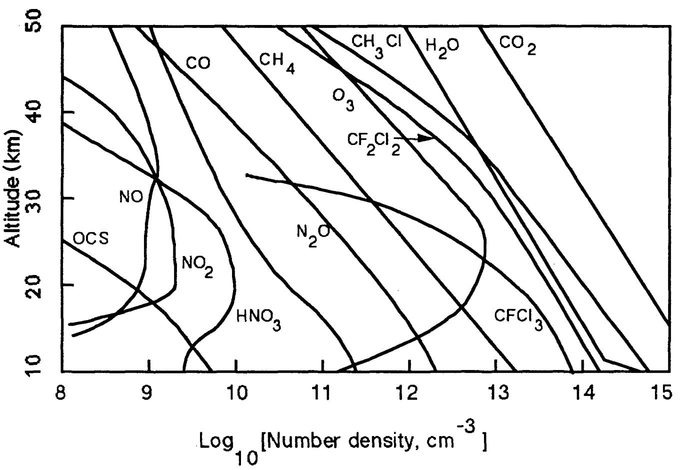

FIG. 3. Some atmospheric minor and trace constituent abundances. After NASA (1986).

3.3 The Ozone Layer

The importance of the ozone layer in determining the temperature structure has been mentioned above (Sec. 2), and its contribution to global habitability, by reducing the amount of energetic ultraviolet radiation that reaches the surface, is well known. Ozone is a very unstable molecule, and the amount that is present at any place and time depends on a balance between production and loss mechanisms. In their simplest form, as stated originally by Chapman, the equations may be written (Brasseur and Solomon, 1986)

$$\begin{aligned} &O_2+h\nu=O+O, \\ &O_2+O+M=O_3+M \end{aligned} \qquad (7)$$

for production, where M is any molecule, usually nitrogen, acting as a collision partner to conserve momentum, and

$$\begin{aligned} &O_3+h\nu=O+O_2, \\ &O+O_3=O_2+O_2, \\ &O+O+M=O_2+M \end{aligned} \qquad (8)$$

for destruction. To obtain reasonable agreement with measurements of the ozone abundance as a function of height, the Chapman scheme has to be augmented to include the effects of catalytic reactions involving trace constituents such as oxides of nitrogen and chlorine. Even then, the amount of ozone varies substantially, not only with latitude and season (as a consequence of the changing solar zenith angle and variations in the circulation of the atmosphere) but also in an apparently random fashion. Research driven by the current concern about ozone depletion (Sec. 8.1) is made complicated by the existence of these large natural fluctuations, many due to causes that are as yet imperfectly understood.

3.4 The Ionosphere

The atmosphere contains substantial numbers of charged particles, especially at those levels (in the mesosphere and above) where dissociation of molecules most frequently occurs (Rees, 1989) (see IONOSPHERE). The ionized fragments, mainly positive ions (of which the most abundant are NO^+ and O_2^+) and free electrons, collect into layers which, particularly in the days before communications satellites, were intensely studied for their effects on radio propagation. The distribution of electrons and the names given to the various regions of the ionosphere are shown in Fig. 4. Since most ions are singly charged and charge neutrality applies overall, the number of ions is approximately equal to the number of electrons at each level.

4. RADIATIVE TRANSFER IN THE ATMOSPHERE

4.1 Solar Radiation

As stated above, the most energetic solar photons at UV wavelengths and shorter are removed at various, mostly high, levels in the atmosphere, where they participate in the dissociation and ionization of oxygen and other species. The longer solar infrared wave-

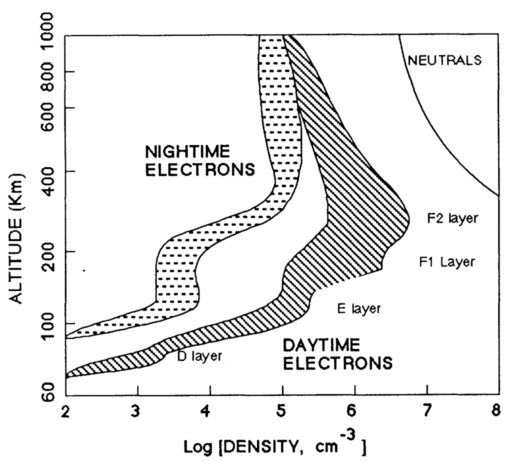

FIG. 4. Middle-latitude electron density profiles in the ionosphere, during the day (shaded area) and at night (stippled area). The areas represent the range of fluctuations due to solar variability.

lengths (1 to 5 μm) are absorbed in some regions of the near-infrared spectrum by molecular vibration-rotation bands, particularly those of water vapor and carbon dioxide. The remainder, radiation at or near visible wavelengths, mainly propagates to the ground unless obstructed by clouds (Fig. 5). On a global average, about 30% of the energy from the Sun is reflected back to space; 25% is absorbed in the atmosphere, and 45% at the ground.

4.2 Thermal Radiation

Because of its low temperature relative to that of the Sun, the Earth's surface re-emits the absorbed energy at a much longer wavelength than it was received (Fig. 6). Some of this thermal infrared radiation is transmitted directly to space by the atmosphere, but at most wavelengths substantially longer than visible the atmosphere is very opaque. Interestingly, the source of this opacity is not, for the most part, the main constituents nitrogen and oxygen, but the minor and trace constituents, especially water vapor, carbon dioxide, ozone, methane, and nitrous oxide. The reason for this is of course the fact that the two homopolar molecules N_2 and 0_2 have no permanent dipole moment and hence no strong vibration-rotation bands in the infrared region. The polar molecules, on the other hand, have a very rich infrared spectra and collectively absorb photons at all wavelengths except in a few small "window" regions, such as those near 12 and near 3.7 μm.

4.3 The Greenhouse Effect

The relative transparency of the atmosphere to sunlight and the great opacity to most infrared wavelengths has the effect of trapping solar energy near the surface. The temperature of the troposphere has to be high enough so that the cooling to space from its upper reaches conserves energy overall; on

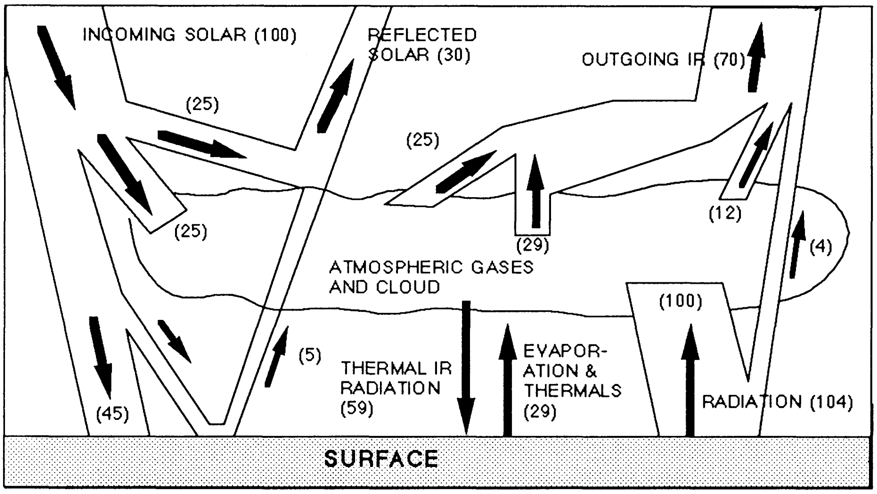

FIG. 5. Contributions of various processes to the energy balance of the atmosphere and surface, in units of percentage of the incoming solar flux.

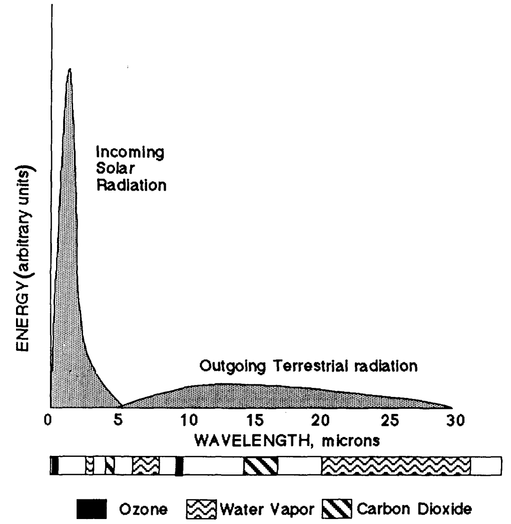

FIG. 6. Simplified emission spectra for the Earth (245 K) and the Sun (5750 K), showing how the energy emitted by these bodies depends on wavelength. Below, the approximate locations in the spectrum of the strongest absorption bands of carbon dioxide, ozone, and water vapor.

the Earth, this has the effect of raising the surface temperature by about 33 K above the equilibrium temperature that would prevail on an airless globe with the same albedo. Needless to say, without this phenomenon, somewhat irrationally known as the greenhouse effect, life as we know it would not be possible on the Earth. Changes in the concentrations of the gases contributing to the greenhouse effect, most notoriously carbon dioxide from industrial processes, are slowly driving the surface temperature up with possibly dire consequences. The physics involved are considered in more detail below (Sec. 8.2).

5. ATMOSPHERIC DYNAMICS

5.1 The General Circulation

The general circulation of the atmosphere is driven principally by pressure gradients that arise as a result of the uneven solar heating of the surface. Because the incoming energy per unit area falls off as the cosine of latitude, the heated air near the surface tends not only to rise, due to convection, but also to move polewards in such a way as to reduce the temperature gradient between equator and pole. The net result is, on the average, a general overturning of the troposphere, with rising motions at low and descending motions at high latitudes (Fig. 7). This *Hadley cell* does not extend all the way to the pole (although it may on Venus; see Sec. 6.1), probably because the rapid rotation of the Earth and the resulting Coriolis forces render such a simple general circulation unstable. A second cell (the *Ferrel cell*) occurs at middle to high latitudes, and a third near the poles, although the main mechanism for the horizontal transport of energy at middle and high latitudes is not a general overturning but instead large-scale wave motions. The relationship between large- and small-scale cells, waves, and flows is an extremely complicated research field (see, for example, Gill, 1982).

5.2 Waves and Turbulence

Atmospheric flows are subject to basic instabilities, generally classified as either *vertical, barotropic*, or *baroclinic* instabilities. An example of the first of these would be the convective instability, which, as discussed above, gives rise to regular overturning motions in the troposphere. Barotropic instabilities are

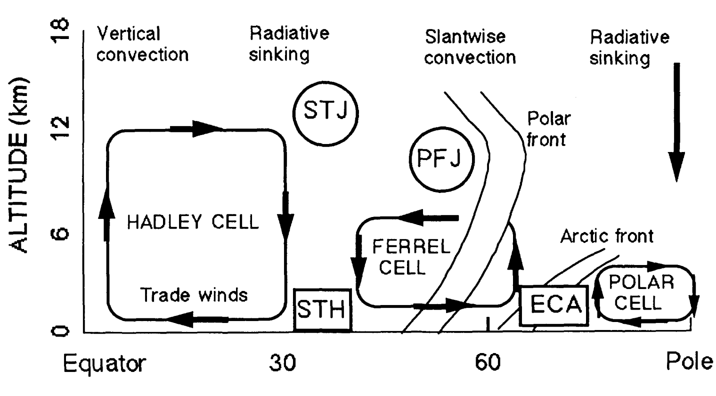

FIG. 7. Schematic features of the general circulation of the atmosphere. STJ = subtropical jet; PFJ = polar front jet; STH = subtropical high; ECA = extratropical cyclones and anticyclones.

those that grow by receiving energy from the mean flow. Baroclinically unstable disturbances extract energy from the horizontal temperature gradient. Both barotropic and baroclinic disturbances can give rise to oscillations in the basic atmospheric fields (temperature, pressure, density, and wind) known as waves. These range in scale from planetary waves, which span the globe in a few wavelengths and are a major factor in the general circulation, to small-scale motions, whose behavior is essentially chaotic, and which mix the atmosphere and act as a frictional term in the equations of motion (see TURBULENT FLOWS). Nonlinear effects in waves can, under certain conditions, roll waves into vortices, some of which become the familiar cyclones seen in weather charts.

Waves in the atmosphere are classified into at least twenty different types (Beer, 1974). Here we consider two prominent examples.

5.2.1 Gravity Waves. These waves arise when parcels of air of different density compared to their surroundings oscillate under the control of buoyancy forces. The motion is therefore at right angles to the direction of propagation. A particular example would be the lee waves that form behind a mountain, as a result of relatively dense air being forced to ascend to cross the obstruction, and then falling freely on the other side. The frequency of the oscillations depends on the wave number k, but for large k tends to the *Brunt–Vaisala* frequency ω, which is given by (Tritton, 1988)

$$\omega = \sqrt{(\gamma - 1) g^2 / c^2}, \tag{9}$$

where g is the acceleration due to gravity.

5.2.2 Rossby Waves. Many of the large-scale features that are seen, for example, in satellite maps of atmospheric temperature fields are *Rossby waves.* These arise as a result of the Coriolis force, which gives rise to a restoring force on an eastward flow that is perturbed, for any reason, away from purely zonal flow. The force on a westward flow is away from the original direction, so westward-propogating Rossby waves are not seen. This explains the absence of planetary-wave activity in the stratosphere in the summer, when the mean flow is predominantly to the west.

5.2.3 Turbulence. The measure of the onset of turbulence is the Reynolds number $\mathscr{R}$, defined as LV/ν, where L is a typical length scale of the motion, V its velocity, and ν the kinematic viscosity. The flow is turbulent if $\mathscr{R}$ exceeds a value of order 10^4. For air at standard temperature and pressure (STP), ν is of the order of $10^{-5}\ \mathrm{m^2\ s^{-1}}$, so the product LV must be as small as 10^{-1} for laminar flow. In fact, typical atmospheric values would be at least 1 m and $1\ \mathrm{m\ s^{-1}}$, respectively, so turbulent flow is common.

In fact, turbulent transfer of heat, minor constituents, and momentum is important at every level in the atmosphere. In the lowest 1 km (the *boundary layer*) it is the dominant process responsible for exchange of these quantities between the land and ocean surfaces and the atmosphere. In the stratosphere, it is found (Andrews *et al.*, 1987) that calculations of the zonal flow and mean meridional motions, which are important for the transport of minor constituents like ozone, resemble reality only if the heat and momentum fluxes due to eddies are incorporated. This is generally accomplished using mixing-length theory, or an empirically determined eddy diffusion coefficient.

6. PLANETARY ATMOSPHERES

The atmospheres of the solar system can be separated into two main classes: those of the terrestrial or Earth-like planets Venus, Earth, and Mars, and those of the outer gas giants Jupiter, Saturn, Uranus, and Neptune. Mercury, Moon, and Pluto, along with most of the satellites of the outer planets, could be considered another class of smaller, virtually airless planetary bodies. Saturn's large satellite Titan, however, possesses a thick Earth-like atmosphere, while Io of Jupiter and Triton of Neptune have very thin atmospheres maintained by active vulcanism.

6.1 The Terrestrial Planets

6.1.1 Composition. Venus, Earth, and Mars are rocky planets with thin *secondary* atmospheres; that is to say, any atmosphere which formed with the planet has been lost, and the gaseous envelopes that now exist were exhaled from the interior at a later date.

Table 2. Surface pressures and fractional abundances (%) in the atmospheres of the terrestrial planets Venus, Earth, and Mars.

	Venus	Earth	Mars
Surface pressure	90 bar	1 bar	0.007 bar
Nitrogen	3.5	77	2.7
Oxygen	~0	20	0.13
Water vapor	~0.01	~1	~0.03
Argon	0.007	0.93	1.6
Carbon dioxide	96	0.035	95
Principal trace gases present	SO_2, CO, Ne HCl, HF	Ne, He, Kr, Xe O_3, CH_4, SO_2, CO, N_2O, H_2	O_2, CO, Ne, Kr, Xe, O_3

Table 2 shows the principal constituents and their relative abundances at the present time.

The composition of Earth's atmosphere is, of course, much modified by the presence of life, and to a lesser extent by the existence of free water on the surface. Neither Mars nor Venus has either feature, and their atmospheric compositions are similar, consisting of carbon dioxide and nitrogen with traces of the rare gases and some photochemically produced species such as carbon monoxide and ozone. Carbon dioxide is a heavy gas and thus easily retained; it is a major product of planetary outgassing and, in the absence of liquid water to promote the production of carbonates, relatively stable. Venus, therefore, has huge amounts of the gas, forming a thick envelope that raises the surface temperature, through the greenhouse effect (Sec. 4.3), to around 730 K. This high surface temperature encourages further outgassing, and active vulcanism on present-day Venus is also a real possibility. Venus has a surface pressure nearly 100 times that on Earth, where most of the originally outgassed CO_2 is now in mineral form. Mars is something of an enigma in having such a low surface pressure, another factor of 100 lower than that on Earth. In part, this might be explained by the relatively small size of the planet, causing vulcanism to end early in its history and making escape of gases to space easier. Another factor is that Mars is cold enough so that CO_2 condenses on the winter pole, a phenomenon that leads to a huge annual swing in the surface pressure all over the planet, which means that some fraction of the atmosphere can be trapped at all times as ice. An additional and, in some models of Mars' evolution, an essential feature is that Mars too may have formed carbonate rocks during an early epoch when the surface was partially flooded. Considerable evidence can be seen in the geological patterns of the surface, which in places on Mars are strongly suggestive of rivers, lakes, and seas. How these were formed and how long they persisted is a matter of much active inquiry and debate.

It is also interesting to consider why it is that Venus and Mars now appear to be so dry compared to Earth. It is thought that on Mars, most of the water that once flowed on the surface still exists on the planet, but now as polar cap ice and lower-latitude subsurface frosts. Venus appears to have lost most of the water it ought to have outgassed along with its carbon dioxide, although a little is still present in the cloud droplets and as vapor. In part, the loss is due to Venus' closeness to the Sun. The high atmospheric temperature and intense ultraviolet fluxes allow greater upward transfer of water vapor, a higher rate of dissociation in the upper atmosphere, and a higher rate of loss to space of the hydrogen atoms produced as a result, compared to the Earth. Here again, however, detailed models suggest that this concept is too simple to account quantitatively for currently observed abundances, and planetary atmospheric water budgets are a research focus.

6.1.2 Clouds. On Venus, three main layers cover the entire planet, occupying the height range from about 20 to 60 km above the surface. The main constituent appears to be concentrated sulfuric acid solution, about 75% H_2SO_4 and 25% H_2O, at least in the uppermost layer where spectroscopy and polarimetry measurements have been made. The droplets there are fairly constant in size, at about 1.1 μm radius; lower down, larger particles up to about 30 μm in radius have been detected by instruments on descending probes. A rather complex, altitude-dependent cloud chemistry, involving photochemical processes, gaseous SO_2 and possibly elemental sulfur, is implied to produce and maintain clouds with this composition and structure.

Mars's atmosphere exhibits a variety of thin clouds at various locations, most spectacularly leeward of the giant volcanoes, which tower up to 25 km above the mean surface height. Clouds of both water and CO_2 ice occur. More important than either of these, in terms of its effect on conditions on the surface, is the load of dust that the atmosphere carries at all times. During storms, this obscures the surface from above, and extends over huge areas, sometimes the entire planet, to heights of 30 km. On Mars, the dust alone exerts a considerable "greenhouse" warming of the surface, as much as 20 K under average conditions.

6.1.3 Dynamics. Venus is a slowly rotating planet (1 Venus siderial day=243 Earth days) and as a consequence the Hadley-type circulation (Sec. 5.1) is more stable than it is on Earth. Cloud motions photographed from spacecraft show equator to pole motions, with velocities of a few m s^{-1} on average, quite clearly; what is less certain is the vertical extent of the cell, and whether or not there are several Hadley cells stacked on top of each other.

The atmosphere as a whole on Venus rotates much more rapidly than the solid planet. This effect, called *superrotation,* is greatest at heights near about 60 km, roughly where the observable cloud patterns lie. At that level, the pressure is about 100 mbar, and the zonal wind (parallel to the equator) is around 100 m s^{-1}, about 50 times faster than the rotation of the surface beneath. Superrotation also occurs in the Earth's upper atmosphere and on the outer planets; it is driven by the transfer of momentum upwards by wave motions. One result is that the poles of Venus are giant vortices, and they have a curious double structure that led to the coining of the term *polar dipole.*

The Martian general circulation is remarkable for its dependence on the surface topography, which is very pronounced, especially for a small planet. As on the Earth, flow over mountains and other structures forces wave motions that propagate and grow, an important part of the mean circulation. Other major factors are the very large day–night temperature contrasts, which can be as much as 100 K, because of the thin air, the variability in dust loading, and the seasonal pressure fluctuation due to the condensation of CO_2 at the winter pole. How these and other processes combine, and how the resulting Martian circulation responds, remain to be fully elucidated.

6.2 The Outer Planets

6.2.1 Composition. The gas giant planets Jupiter, Saturn, Uranus, and Neptune when formed were large enough and cold enough to retain the lighter elements, predominantly hydrogen and helium, which were present in the primordial cloud from which the solar system formed but which were lost to the inner planets. These planets are, in consequence, mostly atmosphere, increasing in density and temperature with depth to very high values, and with the heavy elements concentrated near the core. The composition of Jupiter's atmosphere, the best studied, is shown in Table 3.

As far as is known at present, the elemental abundances in Jupiter are the same as in the Sun, as would be expected if star and planets formed from the same mass and Jupiter retained all of its original material by virtue of its strong gravity field. Most of the elements are combined with hydrogen to form the most stable compound, usually the fully hydrogenated form; for example, most of the nitrogen is found as ammonia, NH_3, and most of the carbon as methane, CH_4. Many other species, in particular more complex hydrocarbons, are present in the thinner

Table 3. Percentage composition of Jupiter's atmosphere.

Hydrogen	86.1
Helium	13.8
Methane	0.09
Ammonia	0.02
Water vapor	0.008
Principal trace gases present	C_2H_2, C_2H_4, CO, HCN, GeH_4, PH_3, C_3H_4, C_3H_8

upper atmosphere where enough solar radiation penetrates to drive photochemical reactions.

6.2.2 Clouds. All of the gas giant planets are cloud-covered, as rising motions from the interior bring air rich in condensibles such as water and ammonia up to cooler levels where they condense. Jupiter is thought to have at least three layers of clouds in the region visible to the outside observer (Fig. 8) and many more below that. The uppermost layer consists of ammonia ice crystals, the one below that of hydrogen sulfide in combination with ammonia as ammonium hydrosulfide, NH_4SH, and the lowest of water. All are bound to have impurities, and the bright coloration of many of the Jovian clouds is testimony to that fact. The so-called chromophores are not yet definitely identified, but various forms of elemental sulfur may contribute to the yellows and browns, while the red spot may be colored by phosphorus produced photochemically from phosphine.

6.2.3 Dynamics. All of the gas giants have internal heat sources (except perhaps Uranus) due most likely to slow contraction of the bodies, either in an overall sense or as a result of fractionation and phase changes in the interiors, with the consequent release of gravitational potential energy. These sources are comparable to the energy arriving from the Sun, and give rise to convective overturning of the atmospheres on a planetary scale. This, and the rapid rotation rates of these large objects, dominate the general circulation, which takes the form of elongated convection cells stretched around the planet in the equatorial direction. The rising branches of the cell have denser, higher clouds, and this gives the planets their banded appearance, most prominent on Jupiter where the clouds form highest in the atmosphere.

After the bands, the most striking characteristics of outer-planet cloud markings are the various spots associated with giant eddies. These come in a variety of colors, sizes, and patterns of behavior and in general are fairly mysterious. One, the Jovian Great Red Spot, is known to have persisted for hundreds of years, at least. On Neptune, a small number of large spots dominates the appearance of the planet with the Great Dark Spot resembling its Jovian counterpart superficially, but

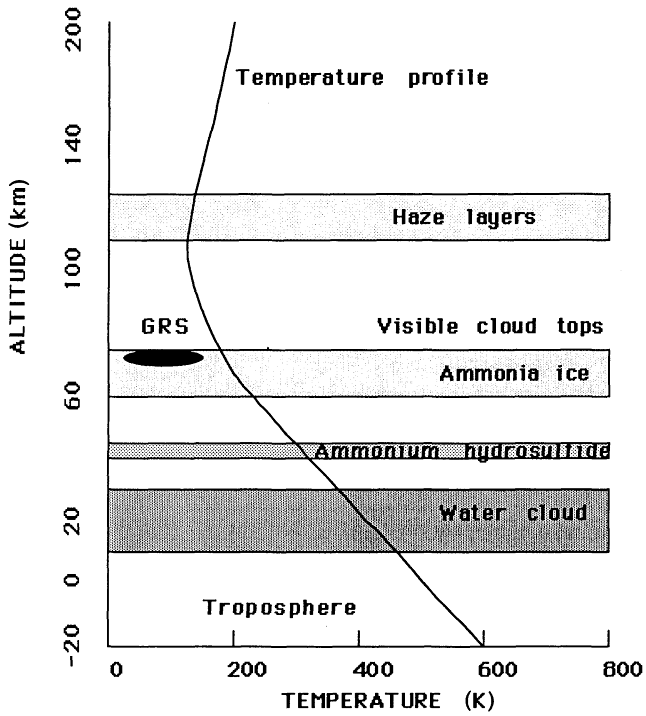

FIG. 8. Inferred structure of Jupiter's atmosphere, based on existing observational data and geochemical models. GRS denotes the Giant Red Spot.

not in detailed behavior. Both are cousins to terrestrial hurricanes but a theoretical understanding of the relationship remains elusive.

6.3 Satellites with Atmospheres

Saturn's Titan has a cloudy atmosphere of nitrogen with a thick layer of orange clouds and a surface pressure of over 2 bars. Why this satellite should be so endowed while, for example, the similarly sized Jovian satellites Ganymede and Callisto are essentially airless is a mystery. The nitrogen on Titan was probably once ammonia that was photochemically dissociated, allowing the hydrogen to escape. The clouds are probably drops of oily hydrocarbons, built up from the photochemical products of methane. It is possible that rain from these clouds has collected on the surface to form an ocean of liquid ethane and other condensible hydrocarbons.

Titan is unique in being the only satellite with an atmosphere that is thick compared to Earth's. Two others, Io (Jupiter) and Triton (Neptune), have thin, transient atmospheres fueled by volcanoes, probably expelling sulfur dioxide and hydrogen sulfide, and nitrogen, respectively. The volcanoes are driven by tidal heating of the interior of the body in both cases, rather than primordial and radioactive heat as with the Earth.

7. ATMOSPHERIC OBSERVATIONS

Although it was one of the earliest subjects of scientific study, our current understanding of the atmosphere of the Earth is mostly quite young, due to the difficulties inherent in obtaining sufficiently extensive and reliable data. To a considerable degree, the current picture, still far from complete, dates back only to the advent of Earth satellites (and computers capable of processing the masses of data that they produce) in the 1960s. Before that, spectrometric instruments on the ground and instruments on balloons and rockets produced isolated profiles of temperature, pressure, and composition, and aircraft explored the lower layers. These techniques are still valued for the high spatial resolution that they afford in studying basic processes, and for their relatively low cost.

7.1 Ground-Based and Airborne Observations

Of the many techniques that have been applied and are still being developed, we mention two classic examples of meteorological instrumentation that are still in daily use.

7.1.1 The Dobson Ozone Spectrophotometer. A worldwide network of about 150 Dobson spectrometers makes daily measurements of total ozone in the atmosphere. It was with this device that the Antarctic ozone "hole" was detected for the first time in the mid-1980s.

The instrument operates by observing the Sun from the ground in the near-ultraviolet part of the spectrum (0.3 to 0.33 μm). Two wavelengths at which ozone absorbs by different amounts are isolated using a diffraction grating and recorded on a photomultiplier detector. A calibrated optical wedge is used to attenuate the longer, less absorbed wavelength, and the wedge position that gives equal attenuation in both channels, together with a knowledge of the solar zenith angle and zero-path brightnesses, is interpreted in terms of total ozone.

7.1.2 The Radiosonde. Radiosondes are packages of inexpensive instruments that are released routinely on small balloons to measure temperature, pressure, humidity, and wind up to heights of about 20 km. Temperature is measured using a fine-wire resistance thermometer, pressure by the deformation of a sealed, flexible capsule, and humidity by the extension of a stressed membrane of a suitable material such as gold-beater's skin. The wind is measured by tracking the balloon during its ascent using radar. Again, a very extensive network in space and time is necessary to provide useful coverage.

7.2 Measurements from Space

Satellites and planetary probes employ infrared and microwave molecular spectroscopy techniques to obtain information about the physical properties of atmospheric gases. This is known as *remote sensing* or, where quantitative measurements of, for example, temperature profiles are concerned, sometimes *remote sounding*. The basic principle of temperature sounding is described in the next section. In applications such as weather forecasting, satellite observations are used to complement surface and balloon data by providing global, frequent coverage. The *in situ* measurements by contrast are sporadic, but have better resolution and often better precision.

7.2.1 Remote Temperature Sounding. The radiance I_λ leaving the top of the atmosphere at a particular wavelength λ is given by the *radiative transfer equation* as

$$I_\lambda = \int_0^\infty B_\lambda(T)\, \frac{d\tau_\lambda(z,\infty)}{dz}\, dz + B_\lambda(T_0)\tau_\lambda(0,\infty), \tag{10}$$

where B is the Planck blackbody function of temperature T at wavelength λ and $\tau_\lambda(z_1,z_2)$ is the transmission of the atmosphere between the two levels z_1 and z_2 at the same wavelength. The second term is the contribution from the surface at temperature T_0. The quantity $d\tau_\lambda(z,\infty)/dz$ is called the *weighting* or *contribution function*, and is usually written $K_\lambda(z)$. The weighting function has a maximum value at the height $z_{\max}$, where the optical depth of the atmosphere is unity, which of course depends on the choice of wavelength. A selection of wavelengths in the wing of a strong absorption band of a species such as carbon dioxide, whose mixing ratio is known and constant up to high levels, allows a family of weighting functions to be calculated with values of $z_{\max}$ ranging from the surface to 50 km or so (Fig. 9). A set of measurements of I_λ can then be turned into a temperature profile by the solution of the set of equations (10). The derived profile is a smoothed version of the actual atmospheric temperature profile, because of the averaging effect of the weighting functions. Some of the detailed structure can be extracted using many spectral channels corresponding to overlapping weighting functions; finding the optimum solution for the profile from a given set of measurements with their associated errors is therefore a complicated matter known as the *retrieval problem* (Houghton *et al.*, 1986).

7.2.2 Vertical Resolution. The vertical resolution of remote sounding measurements is limited by the fact that, in order to get a measurement of I_λ that is not dominated by instrument noise, narrow spectral intervals corresponding effectively to single wavelengths cannot be used in practice. Instead, spectral bands containing many molecular vibration-rotation lines must be employed. This has the effect of further broadening the weighting function relative to that for an idealized, monochromatic observation.

Techniques have been developed, for example using microwave radiometers, that do produce quasimonochromatic weighting functions, but these are still broader (the width at half maximum amplitude is around two atmospheric scale heights, or about 14 km) than desirable for modern atmospheric studies, forecasting, or other applications. This has led to the development of *limb scanning* techniques, in which the instrument views the atmosphere tangentially to the surface and images a thin slab of atmosphere onto the detector. This can improve the vertical resolution to about 2 km, at the expense of poorer horizontal resolution, but tropospheric clouds and hazes generally limit the usefulness of limb scanning to the stratosphere and above.

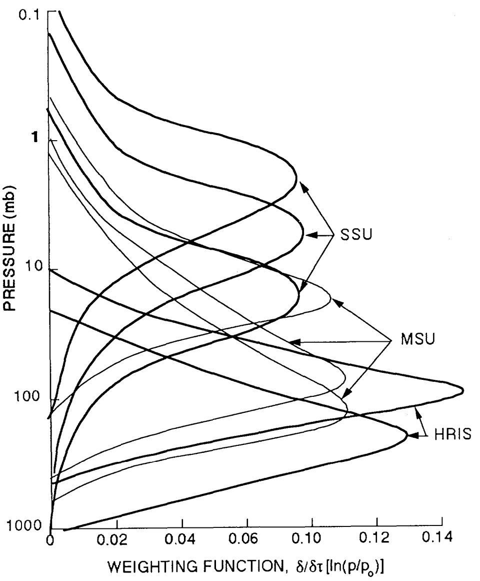

FIG. 9. Weighting function for selected spectral intervals measured by the three instruments that make up the TIROS operational vertical sounder. SSU: stratospheric sounding unit; MSU: microwave sounding unit; HRIS: high-resolution infrared sounder.

7.2.3 Composition Measurements. The transmission function in Eq. (10) is a known function of the vertical profile of the mixing ratio of the emitting species as well as temperature. Once $T(z)$ has been determined using a species of known abundance, measurements in a spectral band of a second species of unknown distribution can be used to obtain its profile. Water vapor or the trace gases such as oxides of nitrogen involved in ozone photochemistry are common targets of this technique. As with temperature, limb measurements improve the vertical resolution when the middle atmosphere is the objective and add the further benefit of a much longer absorbing path, by nearly a factor of 100, when trace constituents are to be measured.

8. CLIMATE AND CLIMATE CHANGE

A knowledge of atmospheric physics can be applied not only to the understanding of the natural state of the environment at the surface and above, but also to considerations of past and future changes in this state. At the present time, there is much concern over changes in the chemical composition of the atmosphere, particularly the stratospheric ozone layer, and changes in the mean temperature due to anthropogenic modifications to the natural "greenhouse" that maintains the status quo. There is also great interest in establishing the physical basis for long-range weather (or short-term climate) forecasting. We consider these topical problems in this final section.

8.1 The Ozone Layer and Ozone Depletion

8.1.1 Production and Loss Mechanisms for Ozone. It was noted above (Sec. 1.2) that the temperature actually *increases* with height above the tropopause and has a maximum value at about 50 km of around 270 K, some 50 K more than the calculated value for an atmosphere without ozone. In fact, it has been known since the early part of the present century that the stratosphere contains substantial concentrations of the triatomic version of oxygen, produced by those photons from the Sun that are at sufficiently short ultraviolet wavelengths (less than about 0.246 μm) to have enough energy to dissociate the oxygen in the atmosphere. The ozone itself also absorbs ultraviolet photons at wavelengths that match its internal energy states; as with CO_2 at longer wavelengths, this ultimately appears as kinetic energy and so warms the atmosphere. Sufficiently energetic UV photons dissociate ozone and produce oxygen again. The amount of ozone that is present at any time depends on the balance between these production and loss mechanisms; so it varies with latitude, height, time of day, and season. This leads to an irregular and variable distribution of ozone around the Earth, with more on the whole at high latitudes. The existence of large amounts of ozone in the stratosphere is important for the biosphere, because its opacity is important for removing ultraviolet radiation that would otherwise reach the ground. This very energetic radiation is harmful to animals and plants.

In addition to the molecular nitrogen that makes up most of the atmosphere, there is a very small amount of nitrogen in its reactive forms, such as the various nitrogen oxides or nitric acid. These come mainly from the surface, from soils containing nitrogen-fixing bacteria, among other sources. Once in the stratosphere—in tiny quantities—they take part in reactions like the following:

$$\begin{array}{l} NO+O_3=NO_2+O_2 \\ \underline{NO_2+O=NO+O_2} \\ O+O_3=2O_2 \end{array} . \tag{11}$$

The importance of this nitrogen oxide or NO_x cycle was not fully realized until the 1970s, because the amounts are so tiny that no one thought nitrogen oxides could possibly be important. The reason why they are important is clear from Eqs. (11)—they are involved in *catalytic* cycles in which the NO_x itself is not consumed.

8.1.2 Anthropogenic Effects. Human activities such as the supplementing of nitrogen-fixing bacteria with artificial fertilizers are increasing the amount of NO_x in the atmosphere. Recently, atmospheric cycles involving chlorine compounds, which mostly start life as Freon propellants in aerosol spray cans or as the working gas in refrigerators and air conditioners, have been identified, and the fact that these also affect the way ozone is created and destroyed has been recognized. More recently still, the importance of hydrogen chemistry through molecules such as H_2O_2 (hydrogen peroxide), which are present in tiny but still important amounts, has been realized (see AIR POLLUTION: COMPONENTS, CAUSES, AND CURES). Stratospheric chemistry is now an extremely complex subject, and realistic attempts to calculate the ozone distribution must include scores of reactions involving dozens of species (Fig. 3).

It is also necessary to realize that all of these reactions, including the catalyzed ones, are very temperature dependent. They also depend indirectly on atmospheric motions, because the ozone and the species with which it is reacting vary considerably in space and time, and the circulation of the atmosphere is very important for moving them around and mixing them up. The intensity of the radiation available is crucial, of course, for both temperature and photochemistry, and varies with altitude and time of day. The heating rate for a given flux depends on the amount of absorbing minor constituent, and so a circular relationship known as the "coupled radiation-chemistry–dynamics (RCD) problem" exists (Fig. 10). This is at the center of most of the problems that the theoreticians in the atmospheric sciences community are tackling at present, including the particularly dramatic, although localized in space and season, problem of the severe depletion of ozone over Antarctica—the so-called ozone "hole." The solution lies in building computer models in which the time-dependent equations of radiative transfer, fluid dynamics, and chemical reactions are all incorporated as accurately as possible and tested against detailed observations.

8.2 The Greenhouse Effect

The role of the greenhouse effect in maintaining the temperature of the Earth's surface may be appreciated by first considering the overall energy balance of the planet (Fig. 11). The effective temperature of the surface of the Sun is about 5750 K, and its distance from the Earth is about 150 million kilometers. Applying the *Stefan–Boltzmann* law, which states that the power E emitted as radiation by any body whose emissivity is ϵ is proportional to the fourth power of its temperature, i.e.,

$$E=\epsilon\sigma(T_{\mathrm{Earth}})^4, \tag{12}$$

where σ is a constant equal to

$5.670\times10^{-8}\ \mathrm{W\ m^{-2}\ K^{-4}}$.

In order for the planet to be in equilibrium overall, E must also be equal to the incoming power, i.e.,

$$E=S\pi R^2, \tag{13}$$

where S is the solar constant, which is defined as the power per cm^2 arriving at the Earth from the Sun. Knowing that the temperature of the Sun is approximately 5750 K, we have (also from the Stefan–Boltzmann law)

$$S=\sigma T^4(R_{\mathrm{Sun}}/D_{\mathrm{Earth\text{-}Sun}})^2 \tag{14}$$

from which we find that S is approximately 1.3 kW m^{-2}. Equating the expressions for E in Eqs. (13) and (14), we obtain the answer $T_{\mathrm{Earth}}=245$ K (approximately). This must be the temperature of a solid sphere the same-size as the Earth at the same distance from

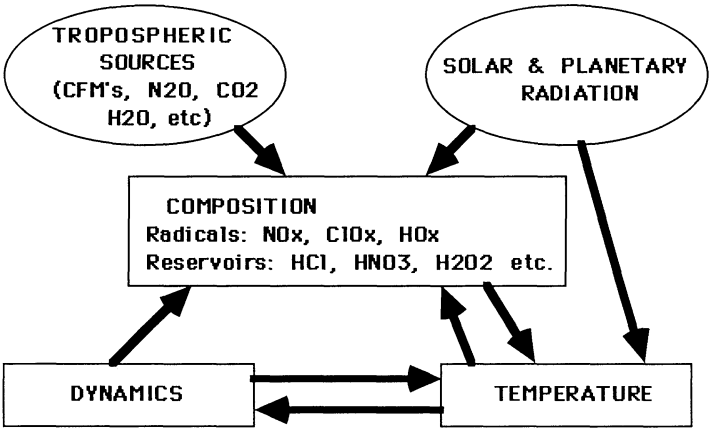
TROPOSPHERIC SOURCES (CFM's, N2O, CO2 H2O, etc)
SOLAR & PLANETARY RADIATION
COMPOSITION
Radicals: NOx, ClOx, HOx
Reservoirs: HCl, HNO3, H2O2 etc.
DYNAMICS
TEMPERATURE

FIG. 10. Coupled radiation-chemistry–dynamics problem in the stratosphere.

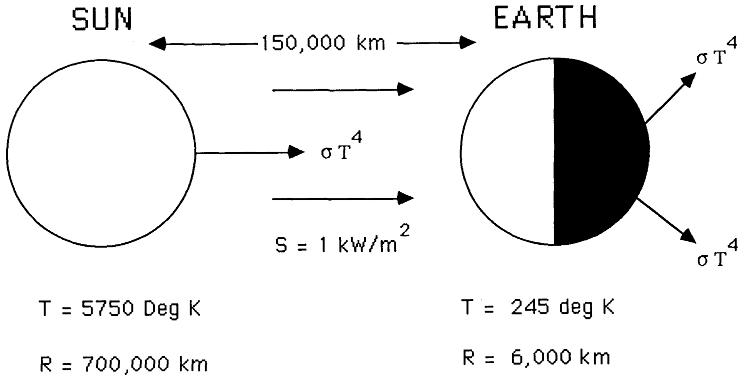

FIG. 11. The overall energy balance of the Earth.

the Sun, but as a value for Earth's surface temperature it is clearly wrong. The reason for the discrepancy is the atmosphere (Fig. 12). The energy from the Sun reaches the surface of the planet because solar radiation consists mostly of photons at visible wavelengths, and the atmosphere is nearly transparent at those wavelengths. The balancing, cooling radiation, however, is in the infrared part of the spectrum because the source is so much cooler, a few hundred degrees instead of a few thousand.

Many of the most abundant minor constituents, such as water vapor and carbon dioxide, absorb *infrared* photons, and do so prodigiously, at most infrared wavelengths (Fig. 6). It also happens that one of the strongest absorption bands of CO_2 covers the wavelength region where emission by typical Earth atmospheric temperatures occur. It is therefore the *atmosphere* that cools by radiating to space, and not the surface, and it is the atmosphere which has the characteristic temperature of 250 K. Of course, the atmosphere will radiate downwards as well as up, so the surface gets an extra σT^4 or 1 W/m^2 in addition to the 1 W/m^2 we get directly from the Sun. By the Stefan–Boltzmann law, this is enough to make the temperature go up to about 290 K, which is much closer to what is observed. The exact solution of the problem involves proceeding from this very simple model to one that includes all of the complex physics of the muitilayer, inhomogeneous atmosphere, with clouds and other factors incorporated (cf. Fig. 5).

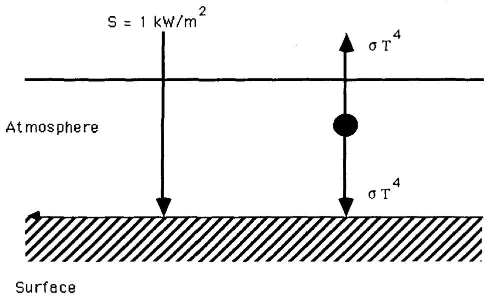

FIG. 12. Energy balance of a single-layer model of the atmosphere.

The above discussion shows how the "greenhouse" effect, caused by minor constituents in the atmosphere, especially water vapor and carbon dioxide, makes the Earth habitable in the first place. The amount of carbon dioxide in the atmosphere is governed by the balance between production, which is mainly by combustion of fossil fuels—coal, oil, and wood—and loss, which is mainly by photosynthesis—trees and other green plants taking in CO_2 and giving off oxygen. Examples of the ways in which human activity is altering this balance are the steady increase in industrial activity, which increases the production of CO_2, and the equally remorseless reduction in the extent of forests, which reduces the rate at which atmospheric CO_2 is recycled.

Figure 13 shows the increase in the abundance of CO_2 that has taken place over the last 100 years or so. A recent discovery is that methane is an important greenhouse gas, too, less so than CO_2 but still significant. The amount of methane in the atmosphere is also, for reasons less well known but probably related to increased agricultural development, going up steadily.

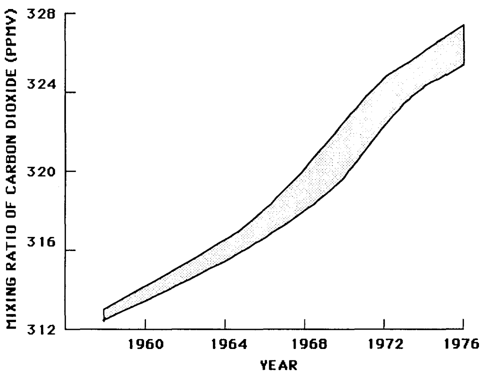

FIG. 13. Range of observed change in CO_2 abundance at various sites. (After Kellogg and Schware, 1981.)

Calculations of the atmospheric energy balance with the amount of CO_2 doubled, which at the present rate will happen in the next century, find that the tropopause will be a few hundred meters higher, and the surface about 3 or 4 degrees hotter. That is a very serious increase indeed and would lead to widespread harmful changes in the environment. For near-equatorial countries, increased desertification would follow; for high-latitude, low countries like parts of Western Europe, the worst problem may be a rise in sea level sufficient to cause extensive flooding of inhabited areas.

8.3 Prospects for Forecasting Climate

General extrapolations of the kind described above are useful for predicting possible modifications to the environment forced by externally imposed changes. A different, but related, problem is the prediction of the natural variability of the climate system, i.e., annual and seasonal fluctuations in the world as we find it now.

Weather forecasting using physical models and data from satellites is now routine, but climate forecasting (predictions of weather averaged over fairly broad space and time scales, but extending months to decades into the future) is at best still some way off. The latest research projects include models of atmosphere–ocean coupling, with which the existence of interactions and fluctuations on long time scales can be investigated. The latest research satellites carry very advanced radiometers of considerable complexity to measure, for example, the tiny amounts (measured in parts per billion in some cases) of reactive gases involved in ozone photochemistry, or the long-term changes in carbon dioxide and methane in the upper atmosphere. Attempts to reproduce the seasonal variation in these quantities are good tests of theoretical models. In due course, long-term trends may be related to temperature and cloud-cover changes to provide an early detection of changes in the climate.

GLOSSARY

Bar: A measure of pressure. 1 bar$=10^5$ pascals. Atmospheric pressures are usually cited in millibars.

Brunt–Vaisala Frequency: A natural frequency of atmospheric gravity waves.

Ferrel Cell: A large-scale overturning of the atmosphere at middle to high latitudes.

Greenhouse Effect: The warming of a planet's surface by the blanketing effect of the atmosphere, resulting from the greater opacity of the atmosphere in the infrared, relative to the visible, part of the spectrum.

Hadley Cell: Large-scale overturning of the lower atmosphere that extends from the equator to middle latitudes.

Homopause: The level at which the atmosphere ceases to be mixed and gravitational separation of species begins. Approximately 100 km altitude.

Ionosphere: Region of the atmosphere containing layers of charged particles; overlaps the upper stratosphere, the mesosphere, and the lower thermosphere.

Lapse Rate: The gradient of temperature with height.

Limb Scanning: A remote sensing technique in which the instrument views the atmosphere tangentially from a satellite.

Mesopause: The level separating the mesosphere from the thermosphere; approximately 90 km altitude.

Mesosphere: The region above the stratosphere, characterized by decreasing temperature with height; approximately 50 to 90 km altitude.

Meteorology: The study of localized atmospheric phenomena ("weather").

Remote Sensing: The technique of measuring atmospheric parameters (temperature, etc.) from a distance, usually from a satellite, using (e.g.) infrared radiometry.

Secondary Atmospheres: Exhaled from the solid body of the planet, as opposed to retained during formation.

Scale Height: The vertical distance in which an atmospheric parameter, usually pressure or density, falls by a factor of $1/e$.

Stratopause: The level separating the stratosphere and the mesosphere.

Stratosphere: The region above the troposphere, approximately 10 to 50 km altitude.

Superrotation: The situation where a region of the upper atmosphere rotates much more quickly than the solid planet.

Thermosphere: The highest region of the atmosphere, characterized by increasing temperature with height.

Trace Constituents: Species present in amounts of a few parts per million or less, often highly variable and chemically active.

Tropopause: The level separating the troposphere from the stratosphere.

Troposphere: The lowest region of the atmosphere, characterized by convective, overturning motions; approximately 0 to 10 km altitude.

Weighting Function: In atmospheric radiative transfer and remote sounding theory, a mathematical function describing the contribution of each height level to the intensity of thermally emitted radiation upwelling to space.

List of Works Cited

Andrews, D. G., Holton, J. R., Leovy, C. B. (1987), *Middle Atmosphere Dynamics*, New York: Academic.

Beer, T. (1974), *Atmospheric Waves*, Bristol: Adam Hilger.

Brasseur, G., Solomon, S. (1986), *Aeronomy of the Middle Atmosphere*, Dordrecht: Reidel.

Gill, A. E. (1982), *Atmosphere–Ocean Dynamics*, New York: Academic.

Henderson-Sellers, A. (1983), *The Origin and Evolution of Planetary Atmospheres*, Bristol: Adam Hilger.

Houghton, J. T., Taylor, F. W., Rodgers, C. D. (1986), *Remote Sounding of Atmospheres*, Cambridge: Cambridge University Press.

Kellogg, W. W., Schware, R. (1981), *Climate Change and Society*, Boulder, CO: Westview.

Kuhn, W. H. (1985), in: J. S. Levine (Ed.), *Photochemistry of Atmospheres*, New York: Academic.

Mason, B. J. (1974), *The Physics of Clouds*, Oxford: Clarendon.

NASA (1986), Reference Publication No. 1162, Washington: National Aeronautics and Space Administration.

Rees, M. H. (1989), *Physics and Chemistry of the Upper Atmosphere*, Cambridge: Cambridge University Press.

Tritton, D. J. (1988), *Physical Fluid Dynamics*, Oxford: Clarendon.

Further Reading

Gill, A. E. (1982), *Atmosphere–Ocean Dynamics*, New York: Academic.

Goody, R. M., Yung, Y. L. (1989), *Atmospheric Radiation*, New York: Oxford University Press.

Houghton, J. T. (1986), *Physics of Atmospheres*, Cambridge: Cambridge University Press.

Houghton, J. T., Taylor, F. W., Rodgers, C. D. (1986), *Remote Sounding of Atmospheres*, Cambridge: Cambridge University Press.

Levine, J. S. (Ed.) (1985), *Photochemistry of Atmospheres*, New York: Academic.

Rees, M. H. (1989), *Physics and Chemistry of the Upper Atmosphere*, Cambridge: Cambridge University Press.

Wayne, R. P. (1985), *Chemistry of Atmospheres*, Oxford: Clarendon.

ATMOSPHERIC STRUCTURE

RAYMOND G. ROBLE, *High Altitude Observatory, National Center For Atmospheric Research, Boulder, Colorado, U.S.A.*

Introduction 201
1. **Global Mean Effective Radiating Temperature** 202
2. **Mean Vertical Temperature Structure of the Atmosphere** . 203
3. **Static Structure of the Atmosphere** 204
3.1 Mean Particle Speed 206
3.2 Speed of Sound 207
3.3 Mean Collision Frequency 207
3.4 Mean Free Path 207
3.5 Mean Molecular Weight 208
4. **Physical and Chemical Processes Responsible for the Mean Atmospheric Structure** 208
4.1 Solar Radiative Output 209
4.2 Penetration of Solar Radiation 210
4.3 Aurora 211
4.4 Radiative Cooling 212
4.5 Global Average Heating and Cooling Rates 212
5. **Latitudinal Structure** 213
6. **Atmospheric Response to Solar Variability** 216
6.1 Thermosphere Response to Solar Variability 216
6.2 Middle-Atmosphere Response to Solar Variability 218
6.3 Lower-Atmosphere Response to Solar Variability 218
7. **Electrical Structure of the Atmosphere** 219
7.1 Air-Earth Conduction Current . 221
7.2 Electric Field 221
7.3 Space Charge 222
7.4 Other Aspects of Global Atmospheric Electricity 222
8. **Concluding Remarks** 222
Glossary 223
Works Cited 224
Further Reading 224

INTRODUCTION

The composition of a planetary atmosphere depends upon the way in which the planet and its atmosphere were originally formed, as well as on physical and chemical processes that continually add some gases to the atmosphere while removing others. Some examples of these processes include volcanic eruptions, chemical reactions at the ground and ocean, and the escape of gases from the top of the atmosphere into space.

The Sun and Earth's atmosphere have been evolving constantly with time. Stellar theory suggests that the early Sun was only 70% as luminous as it is today. Climate theory indicates that if the early Earth's atmosphere had its present composition, the Earth would have existed in permanent glaciation since its beginning and would not yet have thawed out. It has, therefore, been suggested that the early terrestrial atmosphere was poor in oxygen and rich in infrared-opaque gases so that an enhanced greenhouse effect could keep the mean surface temperature above freezing.

The present-day thermal structure of the atmosphere is also determined by how the present atmospheric composition absorbs solar radiation and reradiates it back to space. At today's composition, solar ultraviolet, visible, and infrared radiation are absorbed primarily by molecular oxygen (O_2), ozone (O_3), and water vapor (H_2O). Solar energy in the visible is also absorbed at the ground and oceans. This absorbed solar energy is balanced by infrared radiation to space from the radiatively active atmospheric gases, primarily carbon dioxide (CO_2), O_3, and H_2O vapor. Clouds, however, can influence this clear-sky radiative balance by reflecting solar radiation

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

back into space and also by altering the infrared radiational field. The present-day basic atmospheric structure that is maintained by this radiative balance is described in this article.

In addition to the long-term changes in atmospheric composition because of planetary evolution, there have also been more recent variations. Fluctuations in greenhouse gases (CO_2 and CH_4) have been measured in deep ice cores extracted from the Antarctic ice cap. These ice cores provide a record of atmospheric climate and CO_2 concentration that is representative of global changes over the last glacial-interglacial cycle (160 000 years). CO_2 concentrations as low as 180 parts per million by volume (ppmv) have been derived during the height of the glacial advances about 18 000 years ago (Lorius *et al.*, 1988). The present CO_2 concentration (350 ppmv) is changing rapidly, but this time in response to man's increasing activities (see AIR POLLUTION: COMPONENTS, CAUSES, AND CURES). A doubling of the CO_2 concentration is expected to occur by the end of the 21st century. The greenhouse radiative effect on climate from CO_2 is also being enhanced by increasing concentrations of methane (CH_4), tropospheric O_3, nitrous oxide (N_2O), and chlorofluorocarbons (CFCs). In addition, temperature changes associated with increasing greenhouse gases may also influence water vapor concentration, cloud amounts, and atmospheric structure. The atmospheric structure described in this article applies only to present-day conditions, and it is likely that changes in atmospheric structure will occur in future centuries. Past changes have occurred naturally but projected future changes appear to have a component of man's activities associated with them.

1. GLOBAL MEAN EFFECTIVE RADIATING TEMPERATURE

The Earth is the third planet in our solar system at a mean distance of 1.49×10^8 km from the Sun. It is enveloped by an atmosphere of gases that make our planet habitable by a delicate balance between absorbed solar radiation and radiation emitted to space. The Sun, which has an effective surface temperature of about 6000 K, emits most of its radiation as a blackbody in the wavelength range of 0.2–4.0 μm, including the ultraviolet, visible, and near-infrared wavelengths. The Earth-atmosphere system absorbs part of this energy and is heated by the absorbed radiation at a mean rate of $S(1-a)/4$, where S is the solar irradiance or power per unit area intercepted at the mean Earth-Sun distance (1365–1372 W/m^2), a is the planetary albedo, which is the fraction of solar irradiance reflected by the planet's surface and the atmosphere back to space ($30 \pm 3\%$), and the geometric factor 4 is the ratio of the Earth's surface area to the area of the Earth's disk. The absorbed heating must be balanced by the emission of long-wavelength (thermal or infrared) radiation from the Earth and its atmosphere to space. The rate of cooling is given by σT_e^4, where σ is the Stefan-Boltzmann constant ($\sigma = 5.669 \times 10^{-8}\ W/m^2\ K^4$) and T_e is the effective blackbody radiating temperature of the Earth-atmosphere system. The simplest model of the Earth's climate, considering the long-term average of the global and annual mean temperature (on a time scale greater than a year), is determined from the equilibrium relationship,

$$S(1-a)/4 = \sigma T_e^4. \tag{1}$$

Assuming $S = 1365\ W\ m^{-2}$, $a = 0.3$, yields $T_e = 255$ K (−18 °C). In the absence of an atmosphere T_e is the Earth's global mean surface temperature. However, the value of a that is used does depend on the atmosphere because scattering from clouds and aerosols as well as the properties of the Earth's surface are involved in determining its value.

It is the presence of certain atmospheric gases, such as water vapor, carbon dioxide, and others that are relatively inefficient absorbers of visible solar radiation, but effectively absorb the long-wavelength infrared radiation emitted from the surface, that provides an additional warming of the Earth. These gases absorb, then reradiate both back to the surface, producing an additional warming, and to space, providing the cooling necessary to maintain a radiative balance with incoming solar radiation. The present-day concentration of these so-called "greenhouse gases" raises the effective global mean surface temperature, T_e, from −18 °C to the current observed value of +15 °C. Equation (1) still gives T_e (assuming $a = 0.3 = \text{const}$), but the ef-

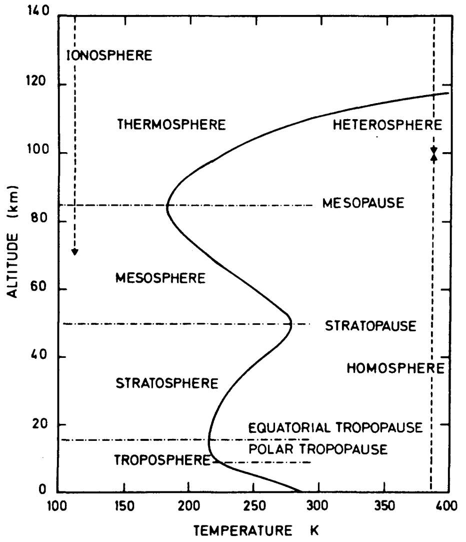

FIG. 1. Thermal structure of the atmosphere from the ground to 140 km. Various atmospheric regions are indicated, as well as the boundaries separating the regions.

fect of the trace gases is to raise T_e to higher values, warming the surface below. Internal heat generated by radioactivity within the planet is a negligible heat source for the atmosphere.

2. MEAN VERTICAL TEMPERATURE STRUCTURE OF THE ATMOSPHERE

In practice, the atmosphere is not in pure radiative equilibrium at a single mean temperature as described above, but varies in all three spatial dimensions as well as in time. Its most fundamental variation, however, is with respect to altitude. The atmosphere has, therefore, been divided scientifically into layers based on the vertical structure of the temperature field. These layers, the troposphere, stratosphere, mesosphere, thermosphere/ionosphere, and exosphere/magnetosphere are separated by boundaries known as the tropopause, stratopause, mesopause, and thermopause. The regions and boundaries, along with a representative vertical temperature profile, are illustrated in Figs. 1 and 2.

The troposphere lies below the first temperature minimum at around 10–15 km, and it is the region where most of the clouds and weather as we experience it exist. The stratosphere lies between this temperature minimum and the temperature maximum around 50 km. The temperature maximum occurs because ozone in this region absorbs solar ultraviolet radiation and heats the atmosphre. Ozone effectively shields the Earth's surface from this harmful radiation. Because the temperature increases with altitude, the atmosphere is stable, forming an inversion layer that confines weather process to the troposphere. Next comes the mesosphere that extends from the 50-km temperature maximum to a temperature minimum around 90 km, the coldest region in the Earth's atmosphere. The thermosphere and embedded ionosphere extend upward from 90 km to several hundred km where the temperature asymptotically approaches a maximum, called the exospheric temperature at the base of the exosphere (Fig. 2). The exosphere lies above the thermopause, and it is so rarefied that the vertical temperature structure remains isothermal because of rapid molecular heat transport. The thermosphere and exosphere are regions where the density is so low that satellites freely orbit the Earth. In recent years the region that includes the stratosphere, meso-

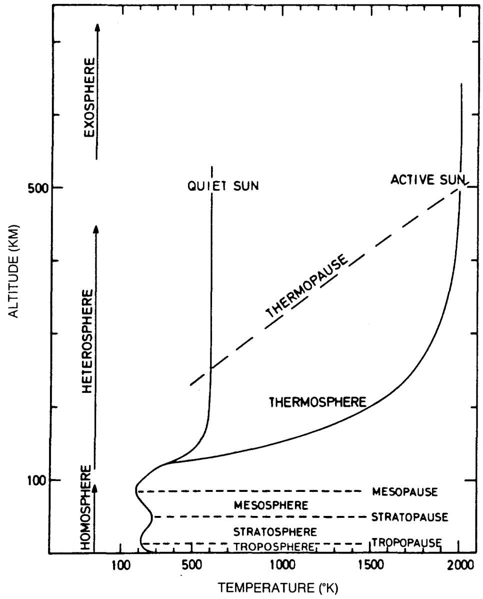

FIG. 2. Mean vertical temperature profile from the ground to the exosphere (from the *U.S. Standard Atmosphere 1976*).

sphere, and the lower thermosphere, between about 10- and 120-km altitude, has been referred to as the "middle atmosphere." This distinction between middle and the lower and upper atmospheres is now widely accepted although there is still frequent reference to the original classification.

The ionosphere (*q.v.*) is a weakly ionized plasma that is embedded within the mesosphere and thermosphere. It is produced by intense solar extreme ultraviolet (EUV) radiation at wavelengths less than 103 nm that ionize the neutral gases in these regions. The magnetosphere (*q.v.*) is a rarefied plasma that coexists with the neutral gases in the exosphere and extends deep into space where it eventually merges with the solar wind. The solar wind is a rarefied plasma that flows radially outward from the Sun into deep space and engulfs the Earth as it passes its orbit. The plasmas contained in both the ionosphere and magnetosphere are strongly controlled by the Earth's geomagnetic field and are subject to electric fields generated by the interaction of the solar wind with the Earth's geomagnetic field.

3. STATIC STRUCTURE OF THE ATMOSPHERE

Atmospheric statics is the study of the relationship between the thermodynamic variables of pressure, density, and temperature, p, ρ, and T, respectively. The atmospheric gases are usually treated as a thermodynamic ideal gas,

$$p=\rho RT, \tag{2}$$

where R is the gas constant, R^*/M, R^* is the universal gas constant, and M is the mean molecular mass.

The vertical profiles of pressure and density, given the vertical temperature profile, shown in Fig. 1, can be determined by the use of the ideal-gas law and the hydrostatic equation. The hydrostatic equation balances the vertical pressure gradient and gravity as

$$dp/dz=-\rho g, \tag{3}$$

where g and z are the acceleration of gravity and the geometric height, respectively. The acceleration due to gravity at sea level, g_0, is 9.806 65 m/s^2. The variation of gravity with altitude follows the inverse-square law of gravitation and for most model atmosphere computations can be determined from

$$g(z)=g_0r_0^2/(r_0+z)^2, \tag{4}$$

where r_0 is the effective radius of the Earth. The *U.S. Standard Atmosphere 1976* gives r_0 as 6356.766 km.

The ideal-gas equation (2) and the hydrostatic equation (3) can be combined to give an equation that describes the vertical atmospheric structure of the atmosphere as

$$\frac{dp}{p}=-\frac{gdz}{RT}=-\frac{dz}{H}, \tag{5}$$

where $H=RT/g$ is called the atmospheric scale height. It represents an altitude difference which corresponds to a reduction by $1/e$ in the pressure or total atmospheric density. The atmospheric scale height is roughly 7 ± 1 km in the lower atmosphere up to about 86 km, but increases rapidly in the thermosphere because of the large increase in temperature and decrease in mean molecular mass. Integrating Eq. (5) gives a simple relationship that describes the vertical structure as

$$p(z)=p_0\exp\left(-\int_{z_0}^{z}\frac{dz}{H(z)}\right), \tag{6}$$

where the subscript 0 denotes that the corresponding values are defined at some reference altitude, z_0, such as sea level. The total number density as a function of altitude can also be derived in a similar manner as

$$n(z)=\frac{n_0T_0}{T(z)}\exp\left(-\int_{z_0}^{z}\frac{dz}{H(z)}\right). \tag{7}$$

The *U.S. Standard Atmosphere 1976* gives the following values for the sea-level temperature, pressure, and density which have been standard for decades:

Temperature, 288.15 K, 15 °C, or 59 °F.
Pressure, 1013.25 mb, 760 mm of Hg, 29.92 in. of Hg, or 1.013 25×10^5 Pa.
Density, 1225.00 g/cm^3 or 0.076 474 lb/ft^3.

The relationship between geometric height, temperature, pressure, and density as given in the *U.S. Standard Atmosphere 1976* is presented in Table 1 and shown in Figs. 1–3. These values are a mean profile of atmospheric structure and should only be consid-

Table 1. Temperature, pressure, and density at various altitudes (from the *U.S. Standard Atmosphere 1976*).

Altitude	Temperature (K)	Pressure (mb)	Density (kg m^{-3})
0	288.15	1013.25	1.225
5	255.65	540.19	0.7361
10	223.25	264.99	0.4135
15	216.65	121.11	0.1947
20	216.65	55.29	0.0889
25	221.55	25.49	0.0400
30	226.51	11.97	0.0181
35	236.51	5.75	8.463×10^{-3}
40	250.35	2.87	3.995×10^{-3}
45	264.16	1.49	1.966×10^{-3}
50	270.65	0.79	1.027×10^{-3}
55	260.77	0.43	5.681×10^{-4}
60	247.02	0.22	3.097×10^{-4}
65	233.29	0.11	1.632×10^{-4}
70	219.58	5.22×10^{-2}	8.283×10^{-5}
75	208.40	2.38×10^{-2}	3.992×10^{-5}
80	198.64	1.05×10^{-2}	1.846×10^{-5}
85	188.89	4.45×10^{-3}	8.219×10^{-6}

ered as a representative profile since at any given time and location there may be considerable departures from the mean depending upon meteorological conditions.

The air in the standard atmosphere is assumed to be dry. The sea-level dry-air fractional-volume composition of the *U.S. Standard Atmosphere 1976* is given in Table 2. These constituents are generally inert, well mixed throughout the atmosphere by turbulence, and maintain this fractional volume composition up to about 86 km. Above this altitude diffusive separation dominates turbulent mixing, and each species will settle according to its own molecular weight with the heavier species settling out in the lower thermosphere and the lighter species dominating the higher altitudes as shown in Fig. 4.

In addition to these chemically inert species, there are also chemically active trace gases, and the concentrations of some of

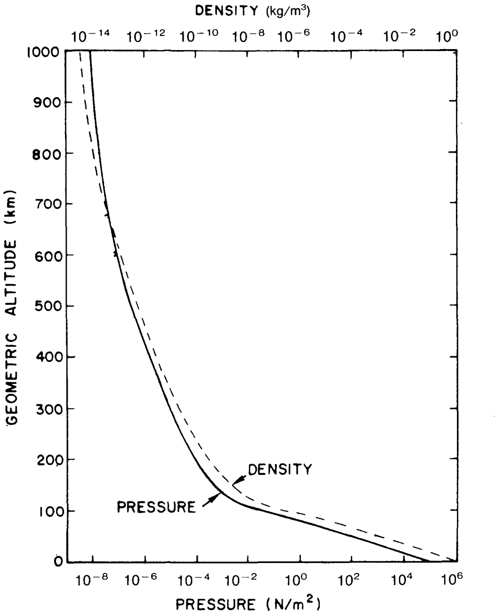

FIG. 3. Total pressure and mass density plotted as a function of geometric altitude (from the *U.S. Standard Atmosphere 1976*).

Table 2. Molecular weight and fractional-volume composition of sea-level dry air (from the *U.S. Standard Atmosphere 1976*).

Gas species	Molecular weight M (kg/kmol)	Fractional volume F (dimensionless)
N_2	28.013 4	0.780 84
O_2	31.998 8	0.209 476
Ar	39.948	0.009 34
CO_2	44.009 95	0.000 314
Ne	20.183	0.000 018 18
He	4.002 6	0.000 005 24
Kr	83.80	0.000 001 14
Xe	131.30	0.000 000 087
CH_4	16.043 03	0.000 002
H_2	2.015 94	0.000 000 5

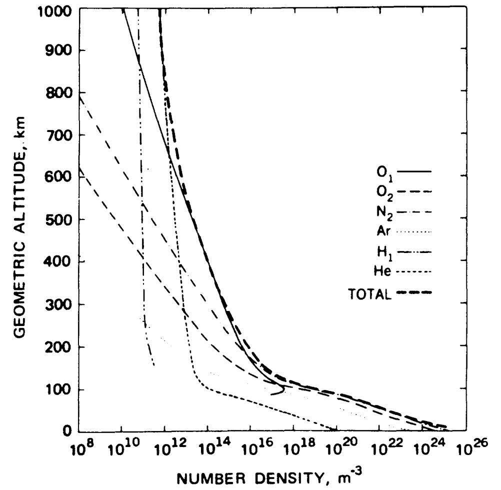

FIG. 4. Number density of individual atmospheric species and total number density as a function of geometric altitude (from the *U.S. Standard Atmosphere 1976*).

these gases near the Earth's surface are given in Table 3. These trace gases can vary widely depending upon geographic location, altitude, and time. Several of these gases are increasing as a result of man's activities.

The number density of individual species and the total number density of all species are shown as a function of altitude in Fig. 4. In the lower atmosphere up to the turbopause near 100 km molecular oxygen and nitrogen dominate at relative percentages of approximately 21% and 78%, respectively, with the remaining 1% attributed to the trace gases. Atomic oxygen, produced by photodissociation of O_2 by intense solar ultraviolet radiation in the upper atmosphere, is scarce in the lower atmosphere but becomes the dominant species above about 140 km. Helium becomes dominant above 600 km and finally atomic hydrogen, the lightest gas, which is also scarce in the lower atmosphere, becomes dominant in the exosphere above about 1500 km. The distribution of atmospheric species above about 100 km undergoes large diurnal, seasonal, solar cycle, and geomagnetic storm variations about the mean profile shown here. Below 100 km, these variations are considerably smaller but, nevertheless, are significant. The variations about this mean profile are discussed in a later section.

There are also many other properties that describe standard atmospheric structure, and some of these are listed below.

Table 3. Concentrations of various tropospheric trace-gas constituents near the Earth's surface (from the *U.S. Standard Atmosphere 1976*).

Constituent	Typical concentration, parts per billion by volume (ppbv)
N_2O	270
NO	0.5
NO_2	1
H_2S	0.05
NH_3	4
SO_3	1
CO	190
0_3	40

3.1 Mean Particle Speed

The mean particle speed V is the arithmetic average of the speeds of all the neutral air particles in a volume element. The analytic expression for V is given by

$$V=(8R^*T/\pi M)^{1/2}, \tag{8}$$

where $\pi=3.141\,5926$, R^* is the universal gas constant, T is the temperature, and M is the mean molecular weight at a specific altitude. The mean particle speed V is plotted in Fig. 5 as a function of altitude. The structure below 100 km properly reflects changes in T because

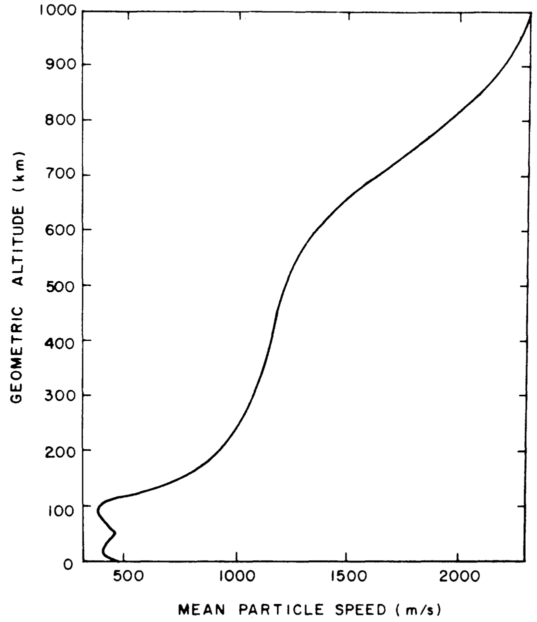

FIG. 5. Mean particle speed as a function of geometric altitude (from the *U.S. Standard Atmosphere 1976*).

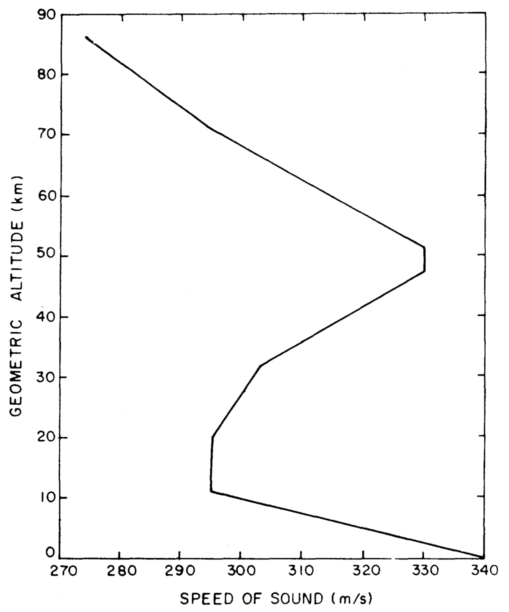

FIG. 6. Speed of sound as a function of altitude (from the *U.S. Standard Atmosphere 1976*).

M is well mixed by turbulence and remains nearly constant with altitude. However, the large increase in the mean particle speed above 100 km results from changes in both *T* and *M*, as molecular species dissociate into atomic species, throughout the thermosphere and exosphere.

3.2 Speed of Sound

The expression for the speed of sound C_s is similar to that of the mean particle speed,

$$C_s = (\gamma R^* T/M)^{1/2}, \tag{9}$$

where γ is the ratio of the specific heat at constant pressure to that at constant volume, 1.40 for air below about 86 km. Above that altitude γ can vary with changing atmospheric composition. The above equation for the speed of sound applies only when the sound wave is a small perturbation on ambient conditions. The attenuation that exists at sea level for high sound frequencies applies to successively lower frequencies as atmospheric pressure decreases and the mean free path increases. For this reason the concept of the speed of sound progressively loses its range of applicability above a certain altitude. The calculated speed of sound as a function of altitude below 86 km is shown in Fig. 6.

3.3 Mean Collision Frequency

The mean collision frequency ν, number of collisions per second, between neutral particles of the atmosphere is a function of pressure, temperature, and molecular weight, as well as of an effective mean collision diameter σ. It is an important parameter for a number of important physical and chemical atmospheric processes, such as collisional quenching of excited species in the high atmosphere. The expression for the mean collision frequency is

$$\nu = 4\sigma^2 np(\pi/MR^*T)^{1/2}, \tag{10}$$

where n is the number density of air and σ the effective collision diameter for an air molecule, 3.65×10^{-10} m^2. The mean collision frequency as a function of altitude in the *U.S. Standard Atmosphere 1976* is shown in Fig. 7.

3.4 Mean Free Path

The mean free path L is equal to the ratio of mean particle speed V to the mean collision frequency ν,

$$L = V/\nu. \tag{11}$$

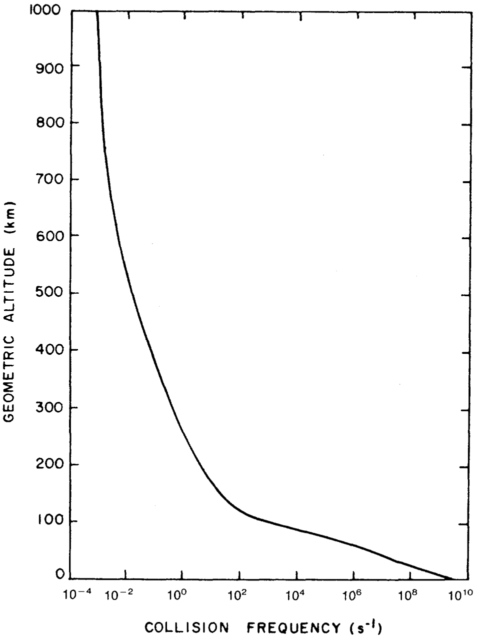

FIG. 7. Collision frequency as a function of geometric altitude (from the *U.S. Standard Atmosphere 1976*).

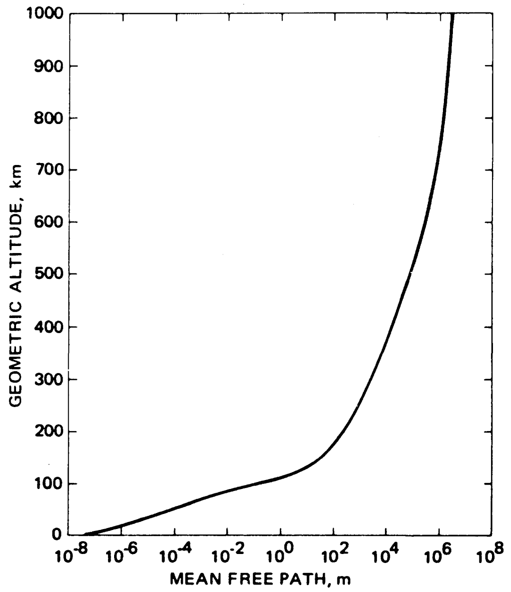

FIG. 8. Mean free path as a function of geometric altitude (from the *U.S. Standard Atmosphere 1976*).

This quantity represents the average travel distance between collisions of neutral particles in any direction for which the total change of number density is small over the mean-free-path length. It is an important quantity in the physical theory of gases. A plot of the mean free path as a function of altitude is given in Fig. 8. The Earth's exosphere is defined as the region above the surface where the mean free path is approximately equal to the local atmospheric scale height, and it occurs near 500 km. This altitude varies considerably with local time, season, and solar activity. Within the exosphere the concept of continuum fluid mechanics breaks down, and air molecules can assume ballistic trajectories or even escape the Earth's gravitational pull and flow outward into interplanetary space.

3.5 Mean Molecular Weight

The mean molecular weight M of a mixture of gases is defined by

$$M=\frac{\Sigma_i n_i M_i}{\Sigma_i n_i}, \tag{12}$$

where n_i and M_i are the number density and molecular weight, respectively, of the ith gas species and Σ_i indicates a sum over all species. Up to about 86-km altitude, turbulent mixing maintains the mean molecular weight at its sea-level value, $M=28.96$ kg/kmol. Above 86 km, diffusive separation and photodissociation of molecular species into atomic species occur, and thus the mean molecular weight decreases with altitude. The mean molecular weight as a function of altitude is shown in Fig. 9.

There are additional mean properties of the atmosphere, such as dynamic viscosity, kinematic viscosity, thermal conductivity, and others, that are not specified here but are given in the *U.S. Standard Atmosphere 1976.*

4. PHYSICAL AND CHEMICAL PROCESSES RESPONSIBLE FOR THE MEAN ATMOSPHERIC STRUCTURE

The physical and chemical processes responsible for establishing the mean structure of the atmosphere are many and varied depending upon the specific atmospheric region. Absorption of solar radiation by the Earth's atmosphere and surface provides the primary energy source, and it is balanced by infrared radiation to space and modulated by

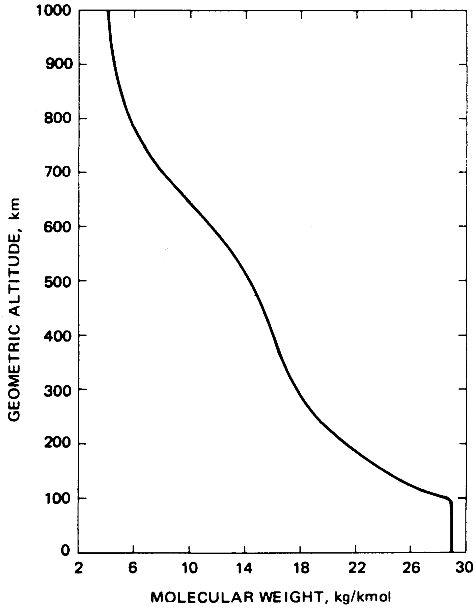

FIG. 9. Mean molecular weight as a function of altitude (from the *U.S. Standard Atmosphere 1976*).

global cloudiness. In this section the basic processes that establish the global mean atmospheric structure are briefly summarized. A more detailed description is given in the article ATMOSPHERIC PHYSICS.

4.1 Solar Radiative Output

The Sun is composed primarily of hydrogen and helium, along with smaller amounts of heavier elements such as iron, magnesium, aluminum, nickel, etc. The temperature in the interior of the Sun is believed to be as high as 2×10^7 K, maintained by a chain of nuclear reactions which convert hydrogen to helium. This energy is radiated to the upper surfaces of the Sun through a series of absorption and reemission processes. Most of the energy reaching the Earth's atmosphere originates from a relatively thin layer of the Sun of about 1000 km in thickness, called the photosphere. Radiation from the photosphere in the visible and infrared follows a blackbody spectrum at an effective surface temperature of about 6000 K as shown in Fig. 10. The visible and infrared portion of the solar spectrum is essentially a continuum emission. Absorption lines become increasingly superimposed on the continuum in the ultraviolet at wavelengths less than 300 nm where departures from the 6000 K blackbody spectrum occur. The longer ultraviolet wavelengths emanate from the Sun's chromosphere, which lies directly above the photosphere, and shorter extreme ultraviolet and x-ray radiation emanates from the Sun's corona, the tenuous outermost atmosphere that is very hot, on the order of several million degrees K. At wavelengths shorter than 140 nm, the emission by hot chromospheric and coronal lines begins to dominate the spectrum. Furthermore, these emission lines are highly variable depending upon the level of solar activity, as shown in Fig. 10. The UV portion of the solar spectrum varies on daily, 27-day solar rotation, and 11-yr sunspot cycles. Also at the extreme ultraviolet (EUV) and x-ray wavelengths, below 50 nm, there are strong enhancements during solar flares and other intense physical processes that occur in the Sun's atmosphere. The variable energetic solar emissions have a pronounced influence on the Earth's atmosphere above about 70 km.

Recent satellite measurements of the total solar radiative output show a variation over an 11-yr sunspot cycle, but it is quite small. The total radiative output of the Sun is about 0.09% larger during sunspot maximum than during sunspot minimum (Willson *et al.*, 1986). Two main components of total radiative output have been identified. Sunspots are

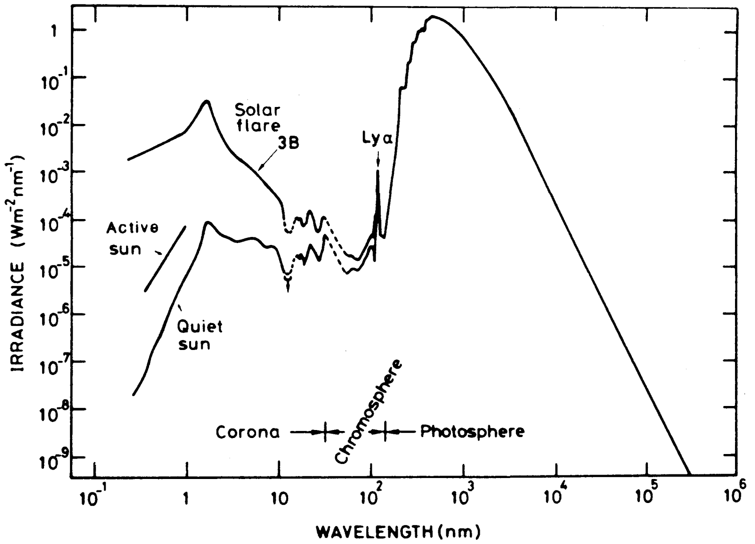

FIG. 10. Spectral distribution of the solar irradiance, and its variation with solar activity. The logarithmic wavelength scale emphasizes the contribution of x rays and extreme ultraviolet radiation (Brasseur and Solomon, 1986). (Copyright by Reidel Publishing Company.)

associated with a decrease in solar radiative output over a solar cycle but it is more than offset by an increase in total radiation from active region phenomena, such as from faculae and plages. Hence there is an increase in total radiative output from sunspot minimum to sunspot maximum. The overall impact of such a small variation in total radiative output on the weather in Earth's lower atmosphere is not well understood at the present time.

4.2 Penetration of Solar Radiation

The mixture of gases at the top of the atmosphere is exposed to the full electromagnetic spectrum of the Sun. As solar photons penetrate into the atmosphere, they undergo collisions with atmospheric molecules and are progressively absorbed and scattered. The probability of absorption by a molecule depends upon the nature of the molecule, its absorption cross section, and the wavelength of the incoming photon.

The Beer-Lambert law describes the absorption of a ray of incident intensity $I(\lambda)$ at wavelength λ, passing through an infinitesimally thin layer of an absorbing gas of thickness ds, as

$$dI(\lambda) = -\sigma_a(\lambda) n_a I(\lambda) ds, \tag{13}$$

where $\sigma_a(\lambda)$ is the effective absorption cross section at wavelength λ in cm^2 and n_a is the number density of the absorbing particle. Integrating Eq. (13) gives

$$I(\lambda) = I_0(\lambda) \exp\left(- \int_s \sigma_a n_a ds \right), \tag{14}$$

where the integral over s is along the path of penetrating radiation.

Solar radiation penetrates the atmosphere at an angle of incidence that depends upon the latitude ϕ, the solar declination angle δ, and the local solar hour angle t. The solar declination, which varies with seasons, is $\delta = \pm 23.5°$ for solstice conditions and 0° for equinox conditions. The local solar hour angle t is 0° for local noon and increases at a rate of 15°/hr in both directions from noon. The cosine of the local solar zenith angle (χ) is given by

$$\cos\chi = \cos\phi \cos\delta \cos t + \sin\phi \sin\delta. \tag{15}$$

For solar zenith angles less than about 70° where spherical and refraction effects are neglected,

$$ds = dz \sec\chi, \tag{16}$$

where ds is the distance along a slant path and dz is an increment in altitude.

If we assume an isothermal (constant temperature) atmosphere, the concentration $n(z)$ varies exponentially with altitude according to a scale height H:

$$n_a(z) = n_a(z_0) \exp[-(z - z_0)/H], \tag{17}$$

where $n_a(z_0)$ is the number density of species n_a at a reference level z_0; then the variation of monochromatic solar radiation with altitude caused by the absorbing gas becomes

$$I(z,\lambda) = I_0(\lambda) \exp[-\sec(\chi) \sigma_a(\lambda) n_a(z) H], \tag{18}$$

where $I_0(\lambda)$ is the intensity at the top of the atmosphere, and $I(z,\lambda)$ is the intensity penetrating down to altitude z. This equation describes the penetration of monochromatic solar radiation of wavelength λ into an atmosphere at a zenith angle χ, with an absorbing species a providing the attenuation. This theoretical derivation was first done by Sidney Chapman in 1931, and it conveniently explains the behavior of layers of ionization in the Earth's ionosphere or photodissociation or photoabsorption in the atmosphere. For multiple species the quantity in the exponential must be summed over all absorbing species a,

$$I(z,\lambda) = I_0(\lambda) \times \exp\left(-\sec\chi \sum_a \sigma_a(\lambda) n_a(z) H_a \right). \tag{19}$$

The optical thickness τ for vertical incidence from space down to an altitude z is defined as

$$\tau(\lambda,z) = \int_z^\infty \sigma_a(\lambda) n_a(z) dz, \tag{20}$$

and the corresponding atmospheric transmission $T(\lambda,z)$ is defined as

$$T(\lambda,z)=\exp[-\tau(\lambda,z)]. \qquad (21)$$

The major absorbing constituents of solar radiation in the atmosphere are N_2, O, O_2, and O_3. The principal absorption of these species occurs at ultraviolet wavelengths with absorption below 127 nm for N_2, 91.1 nm for O, 242.2 nm for O_2, and 1.14 μm for O_3. The depth of penetration, or optical depth, of vertically incident solar radiation as a function of wavelength is shown in Fig. 11. It is the depth at which the incident solar radiation is attenuated by the factor $1/e$. For wavelengths less than 100 nm, radiation is almost completely absorbed by O_2, O, and N_2 above 100-km altitude. Only x-ray radiation at wavelengths less than 1 nm can penetrate below 100 km into the middle atmosphere where it can ionize and dissociate the major species N_2 and O_2. Because these x rays are highly variable and energetic enough to ionize atmospheric species, they are responsible for certain ionospheric perturbations, such as during solar flares. For wavelengths greater than 100 nm, solar ultraviolet radiation can penetrate into the middle atmosphere and photodissociate certain atmospheric molecules. In particular, the solar Lyman-α line at 121.6 nm is very intense and is situated in an optical window where the O_2 cross section is low ($\sim 10^{-21}$ cm^2). Molecular oxygen is the major absorbing species from low wavelengths to near 200 nm. For wavelengths longer than 200 nm, ozone is the major absorber in the lower middle atmosphere. Although ozone absorbs weakly in the visible and even in the infrared, the greater part of the solar photons at wavelengths longer than 310 nm reaches the surface. Solar radiative transfer also includes scattering from gases (Rayleigh scattering), aerosols, and clouds. These affect the planetary albedo and are important to climate.

4.3 Aurora

The solar wind, which is a charged particle (plasma) outflow from the Sun, is another form by which energy is transferred from the Sun to the Earth's upper atmosphere, primarily the thermosphere and ionosphere. Because the solar wind is a plasma (an ionized medium where currents can flow), it is guided by solar magnetic fields that extend outward from the Sun. The electrically charged solar wind is mainly deflected by the

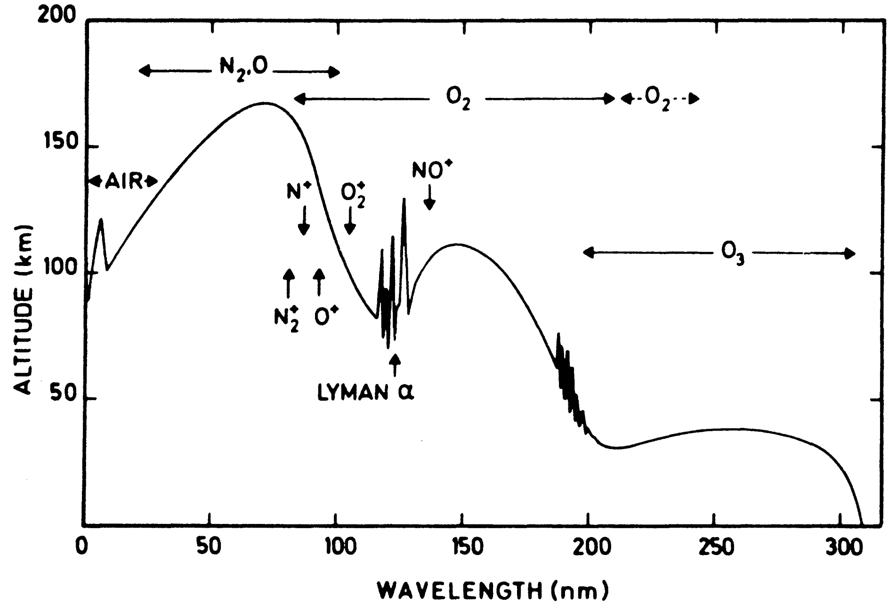

FIG. 11. Depth of penetration of solar radiation as a function of wavelength. Altitudes correspond to an attenuation of $1/e$. The principal absorbers and ionization limits are also indicated (Herzberg, 1965). (Copyright by Prentice-Hall Publishers.)

Earth's own magnetic field, and the bulk of the solar-wind plasma flows around the Earth's magnetosphere into interplanetary space. When the direction of the solar-wind magnetic field is southward, however, complex plasma processes allow a much more efficient coupling with the Earth's magnetic field. Under these conditions a transfer of solar-wind energy is allowed into the Earth's magnetosphere. The solar-wind particles are channeled by the Earth's magnetic field into the high-latitude ionosphere in the magnetic polar caps. The stream of energetic charged particles that flows into the thermosphere and ionosphere collides with the neutral gases and ionizes them, dissociates molecular species, and excites atoms and molecules to higher energy states. These excited atoms and molecules then relax back to their ground states by the emission of photons. The atmospheric glow from atoms above about 100-km altitude is known as the Aurora Borealis in the Northern Hemisphere and the Aurora Australis in the Southern Hemisphere. The energy transferred from the solar wind in the aurora is highly variable and greatly enhanced during geomagnetic storms and substorms. The power from aurora particle precipitation and the Joule dissipation of ionospheric electrical currents associated with the aurora can vary by about 2 orders of magnitude: 10^{10} W for very quiet geomagnetic conditions, 10^{11} W for moderate geomagnetic conditions, and 10^{12} W for brief periods during geomagnetic storms. The bulk of the energy and momentum transfer to the atmosphere in the aurora occurs in the thermosphere and ionosphere, and it is sufficiently large to cause large variations in the thermosphere and ionosphere basic temperature, composition, and wind structure during auroral events.

4.4 Radiative Cooling

The solar energy inputs, described in the previous sections, are balanced by infrared radiative cooling to space, primarily by the 15-μm band of CO_2, but also to a limited extent by the 9.6-μm band of O_3 and the 80-μm band of H_2O in the middle atmosphere and the 6.3-μm band of H_2O in the troposphere. Terrestrial radiation in the infrared region at wavelengths longer than 4 μm dominates the cooling to space. The thermal structure of the atmosphere depends upon an equilibrium between these heating and cooling processes as well as other phenomena, such as production of latent heat in the troposphere and the dissipation of atmospheric waves in the mesosphere and thermosphere. A detailed study of the radiative transfer in the Earth's atmosphere is quite complicated and requires a complete knowledge of the important bands and a compilation of a large body of infrared spectroscopic data. Infrared radiation is discussed in detail in the article on ATMOSPHERIC PHYSICS.

In recent years it has become evident that other radiatively active trace gases, such as methane (CH_4), nitrous oxide (N_2O), and the chlorofluorocarbons, released by anthropogenic activity can interfere with the present-day radiative balance of the atmosphere. Furthermore, the projected rate of increase of these gases into the 21st century indicates that they will become increasingly more important in establishing the mean structure of the entire atmosphere. A list of some of the important infrared atmospheric bands in the Earth's atmosphere is given in Table 4.

4.5 Global Average Heating and Cooling Rates

A calculation of the global average radiative heating- and cooling-rate components in the atmosphere from the surface to 100 km is shown in Fig. 12. The most important radiatively active gases in the atmosphere are O_2, O_3, CO_2, and H_2O vapor. Absorption of solar UV radiation by O_2 and O_3 is responsible for strong radiative heating in the lower thermosphere and in the region of the stratopause. This is balanced, in part, by infrared cooling by CO_2 and to a lesser extent by O_3. Infrared emission in the troposphere, due mainly to H_2O vapor, is partially offset by absorption of solar radiation by H_2O vapor in the near infrared. These are the main heating and cooling mechanisms responsible for the mean atmosphere structure to an altitude of about 100 km. In equilibrium, the differences between radiative heating and cooling rates are balanced by other physical processes such as turbulent transport or release of latent heat.

Above 100 km, the physical and chemical processes responsible for the mean atmospheric structure change dramatically. Virtually all solar photons at wavelengths less than

Table 4. Important infrared atmospheric bands for maintaining the structure of the Earth's atmosphere (World Meteorological Organization, 1982).

Trace gas	Band	Band center (cm^{-1})	Band intensity ($cm^{-2}\ atm^{-1}$) (296 K)
CO_2	ν_2 (15 μm)	667	≈220
CO_2	ν_3 (4.3 μ)	2348	≈2440
O_3	ν_3 (9.6 μm)	1041	312
H_2O	rotation	0–1650	1306
H_2O	ν_2 (6.3 μm)	640–2800	257
CH_4	ν_4 (7.66 μm)	1306	134
N_2O	ν_1 (7.78 μm)	1285	218
N_2O	ν_2 (17 μm)	589	24
$CFCl_3$	ν_4 (11.8 μm)	846	1813
$CFCl_3$	$\nu_1+2\nu_2$	1085, 2144	679
CF_2Cl_2	ν_8 (10.9 μm)	915	1161
CF_2Cl_2	ν_1 (9.13 μm)	1095	1141
CF_2Cl_2	ν_6 (8.68 μm)	1152	777

200 nm are absorbed by the major constituents of the thermosphere, O, O_2, and N_2. The infrared cooling mechanism operating in the thermosphere (5.3-μm emissions from NO, 15-μm emissions from CO_2, and 63-μm emissions from O) are relatively weak and not sufficient to balance the heating rates to maintain radiative equilibrium. The thermosphere is therefore cooled by downward molecular heat conduction above about 140 km and by downward eddy (turbulent) heat conduction to the vicinity of the mesopause where ir cooling becomes sufficiently large to radiate the excess energy to space. The cooling by molecular thermal conduction in the vertical is given by

$$C_m=\frac{1}{\rho C_p}\frac{\partial}{\partial z}\left(K_m\frac{\partial T}{\partial z}\right), \tag{22}$$

where z is the altitude, T is the temperature, K_m is the molecular thermal conductivity coefficient, ρ is the density, and C_p is the specific heat at a constant pressure. Downward thermal conduction implies that the temperature increases with altitude, giving a positive temperature gradient. Therefore, temperature increases with altitude throughout the thermosphere. The molecular thermal conductivity coefficient becomes so large in the upper thermosphere and the heating rate so small, because of the very rarefied atmosphere, that the temperature becomes isothermal at a value called the exospheric temperature.

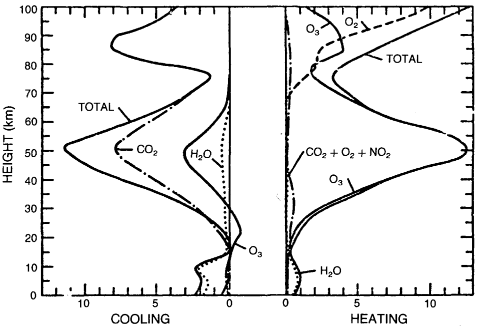

FIG. 12. Components of the global average radiative heating and cooling rates from 0 to 100 km (K/day) (London, 1980).

In the lower thermosphere and upper mesosphere, eddy heat conduction caused by turbulent mixing is an important heat transport mechanism. The cooling by eddy thermal conduction in the vertical is given by

$$C_e=-\frac{1}{\rho C_p}\frac{\partial}{\partial z}\left[K_e\left(\frac{\partial T}{\partial z}+\frac{g}{C_p}\right)\right], \tag{23}$$

where K_e is the eddy thermal conduction coefficient and g is gravity.

5. LATITUDINAL STRUCTURE

The basic vertical structure of the atmosphere was described in the previous sections. The next most important description of structure is its variation with latitude, since there are large seasonal variations in atmospheric properties.

The latitudinal distribution of atmospheric structure is established primarily by the latitudinal distribution of the net heating rate. The net heating rate is the difference between heating rates caused by absorption of solar radiation and the infrared cooling rate from the terrestrial atmosphere. A calculation of the zonally averaged net heating rate, as a function of altitude and latitude for solstice conditions, is shown in Fig. 13. There is net heating in the middle-latitude summer hemisphere and net cooling in the middle-atmosphere winter hemisphere. If the atmosphere were in radiative equilibrium, the temperature structure of the atmosphere would reflect this net heating pattern. The zonally averaged temperature structure of the atmosphere from the ground to 100 km as a function of latitude for solstice conditions is shown in Fig. 14. The bulk of the middle atmosphere is in solar radiative equilibrium, but there are certain atmospheric regions, such as the lower stratosphere and the upper mesosphere, that are far from radiative equilibrium. In these regions the temperature pattern differs considerably from the net heating pattern.

In the troposphere, the maximum temperatures occur in the summer hemisphere tropics with a general decrease in temperature toward both poles, the winter hemisphere being much colder than the summer hemisphere. This pattern follows the net heating rate structure in the troposphere. In the lower stratosphere, between 15 and 30 km, the latitudinal temperature structure reverses. Now the equatorial lower stratosphere is colder than either the summer or winter polar regions. The cold equatorial lower stratosphere occurs because of dynamical effects. The formation of clouds and precipitation in tropical regions is accompanied by the release of latent heat of vaporization providing the energy needed to drive upward motions near the tropopause. Upward motions cool adiabatically, and in the tropical regions above the clouds it is sufficiently strong to overcome radiational heating, thereby maintaining the cold tropical tropopause. Sinking motions in the polar regions produce adiabatic warming.

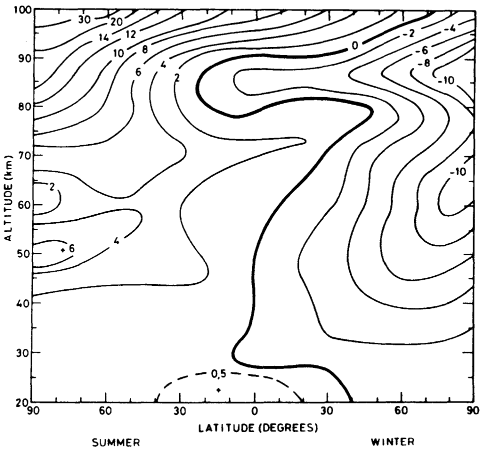

FIG. 13. Total net radiative heating rate associated with the absorption of ultraviolet radiation by ozone and molecular oxygen, and cooling by carbon dioxide (K/day) (London, 1980).

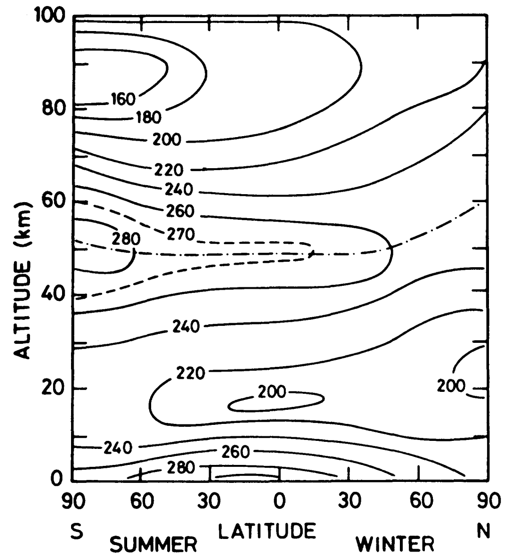

FIG. 14. Latitudinal temperature structure of the atmosphere from 0 to 100 km. The dash-dotted curve represents the height of the stratopause (London, 1980).

The large radiative heating rates in the summer hemisphere stratosphere and lower mesosphere are responsible for the temperature structure of these regions. These regions of the atmosphere are in near radiative equilibrium and they reflect the pattern of net heating. The temperature structure of the upper mesosphere, however, is opposite to what one would expect from the heating rates shown in Fig. 13. The summer upper mesosphere is cold, where the net heating rate is largest, and the winter upper mesosphere is warm where the net heating rate is negative, indicating strong cooling. The source of energy for maintaining the reverse latitudinal temperature gradient from what one would expect from radiative equilibrium appears to be momentum transfer from breaking gravity waves that are sufficiently large to drive a reverse circulation. The upper mesosphere thus acts much like a refrigerator, where momentum transfer from breaking gravity waves drives a circulation that results in the summer mesopause being colder than the winter mesopause.

The mean zonal wind distribution of the atmosphere from the ground to 100 km is shown in Fig. 15. In the troposphere the mean winds are directed eastward (westerly, in meteorological terms) with jets appearing at midlatitudes. The winter jet is about a factor of 2 larger than the summer hemisphere jet. In the middle atmosphere, the zonal winds are eastward (westerly) in the winter hemisphere and westward (easterly) in the summer hemisphere.

Above 100 km in the thermosphere, the basic latitudinal structure displays considerable variability because of the high variability associated with the EUV component of the Sun's radiation that is absorbed in this region. Large amounts of solar-wind energy are also deposited in the high-latitude thermosphere through auroral processes, sometimes locally exceeding the absorbed solar radiation. This makes the thermosphere a dynamically active region with large variations about its global mean state.

The latitudinal variations of temperature and mean mass at 300 km for equinox and solstice conditions are shown in Fig. 16. The zonally averaged temperature and mean mass both have maximum values in the summer hemisphere and minimum values in the winter hemisphere during solstice. At equinox the temperature and mean mass latitudinal variations are small. The enhancements at high latitude are caused by a small auroral heat source for quiet geomagnetic activity. The mean mass varies considerably during solstice because the composition of the atmo-

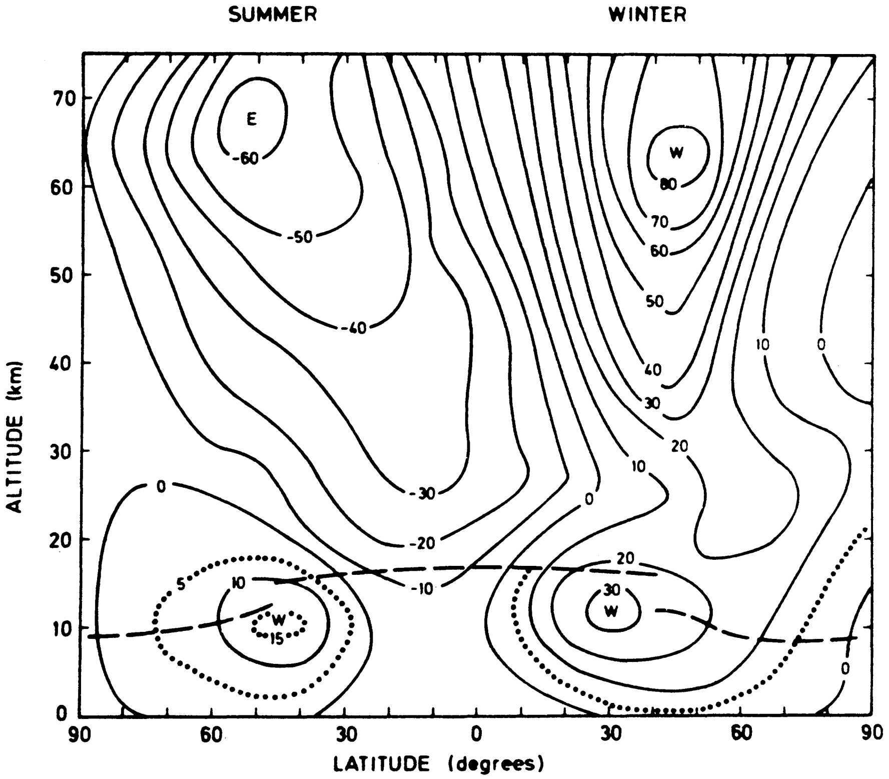

FIG. 15. Mean zonal wind distribution in the middle atmosphere. The contours represent zonal wind speed in m/s. Positive values are eastward (westerly in meteorological terms) and negative values are westward (easterly) (Murgatroyd, 1969). (Copyright by the Royal Meteorological Society.)

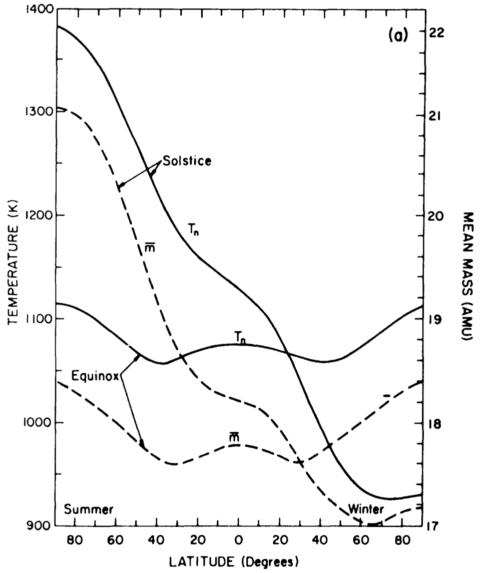

FIG. 16. Latitudinal distribution of the neutral gas temperature and mean mass at an altitude of 300 km in the thermosphere. Equinox and solstice conditions are shown (Roble, 1977).

sphere also varies. Atomic oxygen is the main species at this altitude and it has only a small latitudinal variation; however, both N_2 and 0_2 contribute to the mean mass. In the summer hemisphere the molecular species such as N_2 and 0_2 are enhanced whereas they are suppressed in the winter hemisphere. Because these species are heavier than atomic oxygen the enhancement causes the mean mass to increase to 22 in the summer hemisphere whereas in the winter hemisphere it is only 17.

6. ATMOSPHERIC RESPONSE TO SOLAR VARIABILITY

Solar variability affects each of the atmospheric regions in a different manner. The largest responses occur in the upper atmosphere and the effects become progressively smaller at lower altitudes.

6.1 Thermosphere Response to Solar Variability

The solar EUV and UV radiation at wavelengths less than 200 nm that affect the thermosphere cannot be measured from the ground because it is all absorbed in the upper atmosphere. However, it was shown about 10 years ago that the solar EUV and UV output is closely related to the magnitude of the solar decimeter radio emission at a wavelength of 10.7 cm, which can be measured from the ground with radio telescopes. Since the thermosphere responds dramatically to the solar heat source, the global mean temperature of the thermosphere can be expressed as a function of the solar radio emission. The notation for this function is $F_{10.7}$. During solar minimum the Sun is relatively quiet, and solar $F_{10.7}$ values are typically 80×10^{-22} W/m^2 Hz. Near solar maximum, values of $F_{10.7}$ are 160×10^{-22} W/m^2 Hz or greater with the largest value observed near 400×10^{-22} W/m^2 Hz. The magnitude of the solar EUV output roughly doubles between solar minimum and maximum, and there is also considerable day-to-day variation. Figure 17 illustrates how the global mean temperature profile of the thermosphere responds to the solar $F_{10.7}$ index. During the solar minimum typical exospheric temperatures are 600 K, increasing to 1200 K near the solar maximum. In addition to the temperature variations, the entire thermosphere expands outward with increasing solar heating. For example, the atmosphere density at 500 km can vary by almost 3 orders of magnitude over a solar cy-

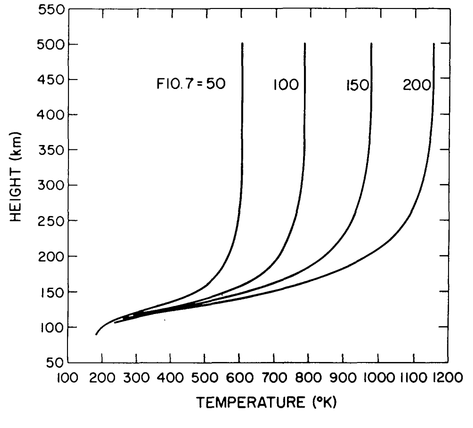

FIG. 17. Global mean neutral gas temperature in the thermosphere as a function of the solar decimeter radio emission at 10.7-cm wavelength ($F_{10.7}$). The solar decimeter radio emission $F_{10.7}$ ($\times 10^{-22}$ W/m^2 Hz) is an indicator of the solar extreme ultraviolet radiation output (Roble, 1977).

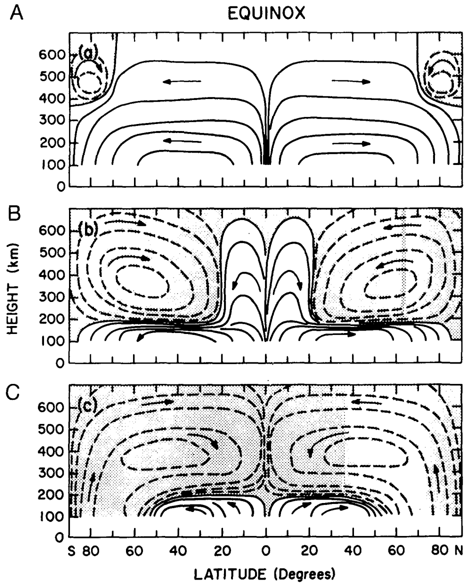

FIG. 18. Schematic of the zonal mean meridional circulation in the Earth's thermosphere during equinox conditions for various levels of auroral activity: (a) extremely quiet auroral activity (10^{10} W), (b) average auroral activity (10^{11} W), and (c) auroral storms (10^{12} W). The contours illustrate the mass flow and the arrows indicate the direction of the motion. Stippling indicates regions of reverse circulation (Roble, 1977).

cle, and this increase in thermospheric density can have an important influence on the drag experienced by orbiting satellites.

The circulation during equinox is shown in Fig. 18, where the mass flow is illustrated schematically, indicating the cellular motion of the thermosphere between 80 and 500 km. Figure 18(a) shows the circulation during very quiet geomagnetic conditions: upward motion occurs over the equatorial subsolar point with a poleward flow in both hemispheres to high latitudes where sinking motion occurs. The small reverse circulation at high latitudes is caused by a small auroral heat source. During average geomagnetic activity the size of the equatorial circulation is larger, as shown in Fig. 18(b). Below about 150 km the strong solar UV heating still maintains a poleward flow; however, above that altitude the flow is equatorward to about 20° latitude. During sizable geomagnetic storms, the equatorward circulation increases greatly, extending right down to the equator, as shown in Fig. 18(c). The circulation below 150 km is also affected by the auroral heating. Thus, it is evident that the mean motions within the thermosphere are in a constant state of agitation, depending on the magnitude of the auroral heat source that varies daily.

The situation is similar during solstice. There is a surplus of energy in the summer hemisphere and a deficit in the winter hemisphere, by comparison with the global mean values. Under geomagnetically quiet conditions, the asymmetry in solar heating drives a summer-to-winter circulation as shown in Fig. 19(a). In Fig. 19(b), the auroral heat source reinforces the solar-driven summer-to-winter circulation in the summer hemisphere but forces a reverse circulation in the winter hemisphere. The two cells converge in the midlatitude winter hemisphere. Below 150 km the summer-to-winter circulation is maintained at all latitudes. As the magnitude of the auroral heat source increases during geomagnetic storms the circulation is equatorward in both hemispheres above 300 km with an asymmetry at lower altitudes. Thus even during solstice conditions the auroral heat source has a strong influence in controlling the circulation and basic structure of the thermosphere (Roble, 1977).

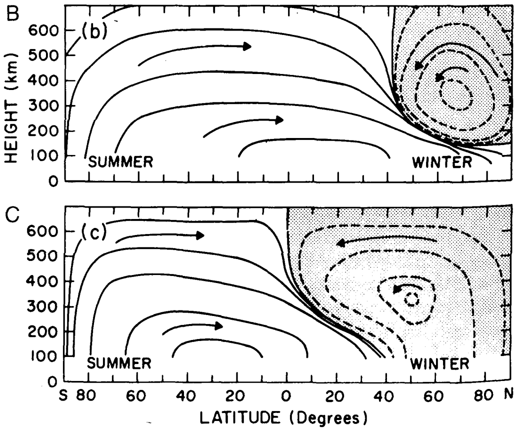

FIG. 19. Same as Fig. 18 but for solstice conditions (Roble,1977).

6.2 Middle-Atmosphere Response to Solar Variability

Solar UV radiation affects a variety of physical and chemical processes in the middle atmosphere. The response of ozone to solar UV variability has been detected by satellites. Near 70 km, it is found that ozone is immediately depleted by short-term increases in solar UV radiation. An apparent mechanism for the ozone depletion is an increase in odd hydrogen production as a result of an increase in the photodissociation of H_2O vapor by solar Lyman-α radiation. The odd hydrogen, defined as the sum of chemically active hydrogen species ($H+OH+HO_2$) catalytically destroys ozone. The response of mesospheric temperature to short-term solar UV variability has also been identified in data sets. There is an approximate 0.3 K increase for a 1% increase in solar UV radiation centered near 205 nm with a small phase lag (Keating *et al.*, 1987).

Temperature and ozone are oppositely correlated throughout the lower mesosphere and upper stratosphere. Temperature-sensitive chemical rate coefficients that control the amount of ozone present at any time are responsible for this effect; as temperature increases ozone decreases and vice versa. Ozone can also be destroyed by an increased photodissociation by solar UV radiation. In the stratosphere, ozone responds directly to changes in solar UV radiation with a phase lag of 0.5 days. Deeper in the atmosphere near 20 km it responds directly but it has a 2-day phase lag. Zonally averaged models of the middle atmosphere have also been used to demonstrate the response of the middle-atmosphere chemistry and dynamics to solar variability over the 11-yr sunspot cycle (Garcia *et al.*, 1984).

6.3 Lower-Atmosphere Response to Solar Variability

The lower atmosphere is a highly complex nonlinear system that is subject to a large component of random variability due to its internal dynamics. Even under conditions of constant external forcing, weather and climate will vary. Weather variability arises primarily as a result of instabilities of the mean flow of the atmosphere, and it is determined by the day-to-day evolution of the large-scale synoptic pressure systems that dominate the atmospheric circulation. Climate is simply the statistical average result of averaging the weather over a period of time.

There have been many attempts to correlate the weather with various properties of the Sun. Most correlations have involved sunspots, but others have considered solar-wind magnetic-sector boundary crossings, flares, etc. Generally, these correlations have a tendency to break down after a period of time. But recently, it has been discovered that there is a strong correlation between the solar cycle and high-latitude stratospheric temperature when the data are categorized according to the east or west phase of the quasibiennial oscillation (QBO) of the winds in the equatorial stratosphere (Labitzke and van Loon, 1988). When the QBO winds are westward the polar data are positively correlated with the sunspot cycle while those at middle and low latitudes are negatively correlated. The converse holds true for the eastward wind phase of the QBO. These effects have been found to hold throughout the troposphere and even into the mesosphere. No physical mechanism explaining the observed correlation has yet been found.

The troposphere has a large natural variability that can exist without any change in external forcing. The energy forcing associated with known solar variability is too small, by about 3 orders of magnitude, to perturb the enormous energy of the troposphere. Because of this, Sun-weather relationships have usually invoked some trigger mechanism that somehow taps the basic energy of the flow and forces an atmospheric response to solar activity.

There appears to be more evidence suggesting possible Sun-climate relationships (Reid and Gage, 1988). Climate energy-balance models all tend to show that the sensitivity of the global climate response to solar variations is an important function of the period of forcing, because of couplings between the ocean and atmosphere. The thermal lag in the models suggests that any climate response to solar irradiance variations on an 11-yr sunspot cycle is small but the response increases on longer time scales. A more effective response is obtained on the timescale of the solar Gleisberg cycle of sunspot activity ($\sim$80–90 yr) and from simulations of the Maunder minimum, an 80-yr period of little or no sunspots from 1645 to 1715 AD that corresponded with the little ice age in Europe and elsewhere.

7. ELECTRICAL STRUCTURE OF THE ATMOSPHERE

The atmosphere is an electrically conducting medium because of the presence of positive and negative ions and electrons. The principal sources of ionization in the atmosphere include

1. ionization by galactic cosmic rays that bombard the atmosphere,
2. photoionization by solar x rays and solar EUV radiation at wavelengths less than about 130 nm,
3. ionization by energetic electrons and protons that bombard the Earth in the aurora,
4. radiation from radioactive substances in soil, and
5. radioactive gases produced by the decay of 4.

In the atmosphere below about 80 km, the net production rate of small ions and electrons, q, is balanced by complex ion chemistry and can be approximated by assuming chemical equilibrium,

$$q = \alpha n_{\pm}^2, \qquad (24)$$

where α is the volume recombination coefficient and $n_{\pm}$ represents the number density of equal positive and negative ions, respectively. The number density of small positive and negative ions as a function of altitude at various latitudes is shown in Fig. 20. Galactic cosmic-ray ionization is the main ionization source for the production of the ion density profiles. Because the galactic cosmic-ray ions are charged they are influenced by the Earth's magnetic field to an extent that lower-energy cosmic rays are predominantly guided into high magnetic latitudes and only the more energetic cosmic rays penetrate to low and equatorial latitudes. As a result, there is a significant latitudinal variation in the vertical profile shapes.

The electrical conductivity σ of air is defined as

$$\sigma = \sum_i n_i e_i k_i, \qquad (25)$$

where n_i is the number of the ith ion species present, and e_i and k_i are the charge and mobility of the ith ion species, respectively. Two classes of ions are normally present in the atmosphere, small ions of charged molecular species and large, heavier ions of charged aerosols, cloud particles, dust, etc. Small ions, because of the low mass, have higher mobilities than larger ions. The mobility of an ion is its velocity per unit electric-field strength. Near the Earth's surface, $n^+ \cong 6 \times 10^8$ m^{-3}, $n^- \sim 5 \times 10^8$ m^{-3} (the difference due to the electrode effect, a charge separation process near the Earth's surface that resembles a similar process near an electrode of an electronic device), $k^+ \sim 1.3 \times 10^{-4}$ m^2/V s, and $k^- \sim 1.6 \times 10^{-4}$ m^2/V s, yielding a conductivity of about $2.5 \times 10^{-14}/\Omega$ m. The electrical conductivity increases rapidly with altitude, as shown in Fig. 21, for the following reasons:

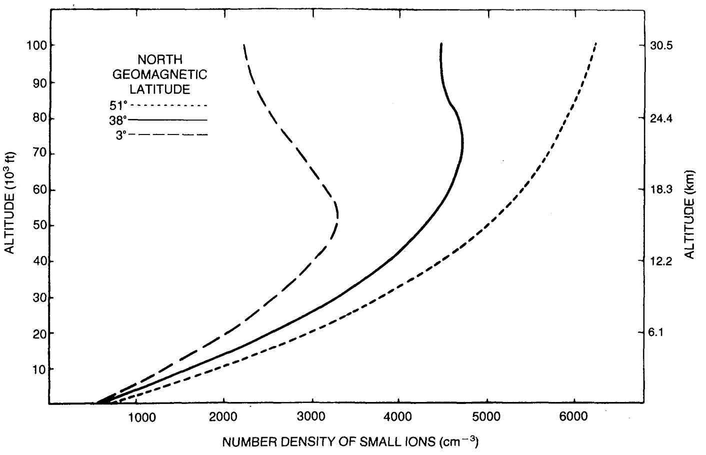

FIG. 20. Number density of small positive or negative ions as a function of altitude computed assuming equilibrium with the galactic cosmic-ray radiation. The variation with latitude illustrates the shielding effect of the Earth's magnetic field (from *Handbook of Geophysics and the Space Environment*, 1985).

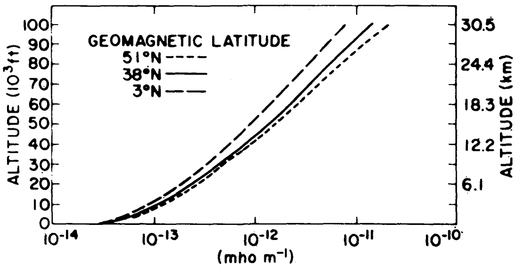

FIG. 21. Total electrical conductivity as a function of altitude for various latitudes (from *Handbook of Geophysics and the Space Environment*, 1985).

1. the mobility of ions increases as the neutral number density decreases,
2. the presence of numerous large aerosols lowers the conductivity of air near the ground, and
3. galactic cosmic-ray ionization, photoionization, and energetic particle ionization increase with altitude.

Investigators in the 18th century recognized that a weak electrification exists as a permanent property of the atmosphere even during fair weather. The Earth's surface was found to be charged negatively and the air charged positively, with a vertical electric field of about 100 V/m existing in the atmosphere near the surface in fair weather. The search for the cause of this electrification was a major investigation for many years. Because the air is electrically conducting, a continuous charging source had to exist. In the 1920s, the British physicist C. T. R. Wilson suggested that currents flowing from thunderstorms were the charging source responsible for maintaining the fair-weather electric field. This hypothesis was based on his observations that beneath a thundercloud negative charge is carried to the Earth's surface and above the thundercloud positive charge flows upward toward the ionosphere. Within the highly conducting ionosphere, the charge is rapidly distributed around the globe, and then it flows down to Earth through the atmosphere in fair-weather regions between thunderstorms as shown in Fig. 22. In the figure, region 0 represents the planetary boundary layer in the troposphere; region 1, the atmospheric region that includes thunderstorms; and region 2, the stratosphere and mesosphere. In the upper portion of the figure, current flows along magnetic-field lines in the ionosphere and magnetosphere. This global circuit can be compared to a giant leaky

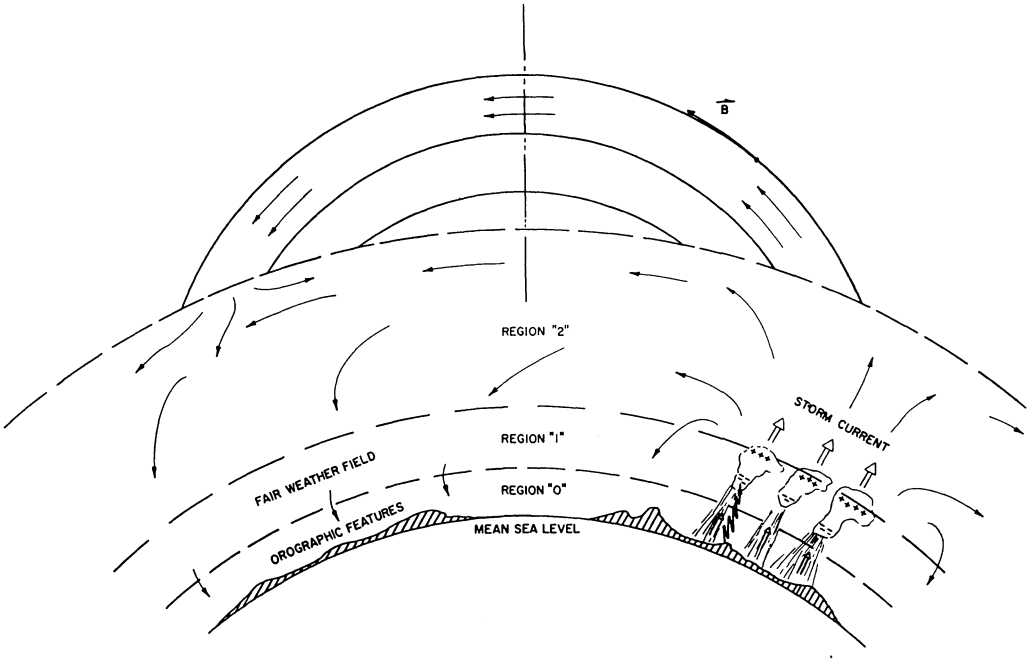

FIG. 22. Schematic of the Earth's global electric circuit (Hays and Roble, 1979). (Copyright by American Geophysical Union.)

spherical capacitor with the highly conducting Earth's surface and the ionosphere representing the plates, the weakly conducting lower atmosphere representing the dielectric in between, and thunderstorms maintaining the charge separation.

The simplest thunderstorm model assumes a quasistatic dipole charge distribution, with positive charge embedded in the cloud top and the negative charge in the cloud bottom. The dipole is immersed in an atmosphere where the electrical conductivity increases exponentially with altitude from the surface to the ionosphere. The 20–100 C of charge that are separated in a thunderstorm maintain large potential differences within spatial regions of the storm. The average current flowing from a thundercloud into the global circuit is typically 1–2 A. Since there are about 1000 thunderstorms operating globally at any given time, about 1–2 kA of current flows continuously into the global circuit. Even though the columnar resistance of the atmosphere is high ($1\times10^{+17}$ Ω/m^2), the columns are coupled in a parallel circuit to give a global resistance of about 250 Ω. Thus, a ground-to-ionosphere potential difference of 250 000 V is maintained by global thunderstorm activity. This potential difference is uniform around the globe because the conductivities of the ionosphere and ground are so high that currents flow freely in these regions to reduce horizontal potential differences.

7.1 Air-Earth Conduction Current

As a result of the potential difference between the Earth's surface and ionosphere, a current is constantly flowing to the Earth's surface. This current is called the fair-weather air-Earth conduction current and is typically $(1\text{–}2)\times10^{-12}$ A/m^2. Since the columnar resistance that controls this current varies depending upon the Earth's orography, higher current densities flow into mountainous areas compared with sea level. There is also a diurnal variation of the surface-ionosphere potential difference and air-Earth current density because of the diurnal variation of global thunderstorm activity. Maximum potential differences and air-Earth current densities occur near 1900 UT when the Sun is heating the South American and African continents and producing copious amounts of thunderstorms. The minimum values occur near 0400 UT when the Sun is heating the Asian sector where thunderstorm activity is lower. A typical diurnal variation is about 10%–20%. The air-Earth current in fair weather is constant with altitude.

7.2 Electric Field

In fair-weather areas the horizontal components of the electric field are small compared with the vertical component. Thus, the atmospheric electric field can be defined as

$$E = -dV/dz\,, \tag{26}$$

where V is the electric potential and z is the altitude. Since the ionosphere is charged positively with respect to the Earth's surface, the electric field is directed downward. Because the air-Earth current is constant with altitude and the electrical conductivity increases exponentially with altitude, the electric field, by Ohm's law, must decrease exponentially with altitude as shown in Fig. 23. Assuming an exponential increase in electrical conductivity with altitude, the basic equations for the atmospheric electric field can be expressed as

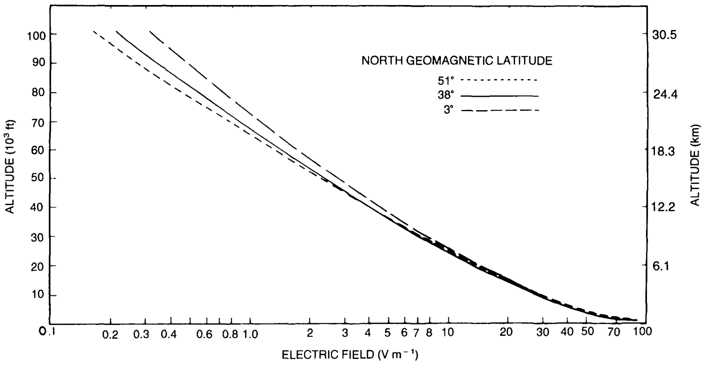

FIG. 23. Vertical electric field as a function of altitude calculated from the electrical conductivity shown in Fig. 21 using Ohm's law and an assumed vertical conduction current of 2.7×10^{-12} A/m^2 (from Jursa, 1985).

$$\sigma(z)=\sigma_0 \exp(2kz), \tag{27}$$

where σ is the electrical conductivity, k is the mobility, and z is altitude. In fair weather there is no charge source and the potential equation is

$$\frac{d^2V}{dz^2}+2k\frac{dV}{dz}=0. \tag{28}$$

Assuming that the potential at the Earth's surface is zero and that the total potential difference between the top of the atmosphere (height H) and the Earth's surface is V_∞, the solution of Eq. (28) yields

$$E=\frac{V_\infty 2k \exp(-2kz)}{1-\exp(-2kH)}=E_0 \exp(-2kz), \tag{29}$$

where E_0 is about 130 V/m at the Earth's surface.

7.3 Space Charge

The space charge density ρ is related to the electric field by

$$\frac{dE}{dz}=-4\pi\rho. \tag{30}$$

7.4 Other Aspects of Global Atmospheric Electricity

The basic atmospheric electric structure in fair weather is generally described by conduction currents flowing in accordance with Ohm's law. There are, however, other sources of current such as

1. convection current, where movement of air carrying charge produces a transfer of charge;
2. displacement current, caused by changes of the electric field with time, $j=\epsilon_0\, dE/dt$, where ϵ_0 is the permittivity of free space;
3. precipitation current, where ions of one sign attached to raindrops fall as rain; and
4. lightning current, where charge, either positive or negative but predominantly negative, is transferred to ground during impulsive lightning strokes.

Besides thunderstorms acting in the lower atmosphere, there are two other upper-atmosphere electrical generators that couple into the Earth's global electric circuit: the ionospheric wind dynamo, where winds flowing in the thermosphere near 100 km generate a horizontal potential difference in the ionosphere of 5–10 kV with 100 000 A flowing horizontally in the dayside ionosphere, and the solar-wind/magnetosphere dynamo that generates a 30–150 kV potential drop across magnetic conjugate polar caps with 1 000 000 A flowing primarily in the auroral zone. Just how all of these current systems in the Earth's global electric circuit interact and respond to solar and auroral variability presents an exciting research challenge in atmospheric science.

8. CONCLUDING REMARKS

In recent years, it has become evident that the trace gases that are being released by man's activities have a potential for altering the detailed structure of the atmosphere and causing possible future climate changes (Bolin *et al.*, 1986). Trace gases such as CO_2, CH_4, tropospheric O_3, N_2O, and CFCs are steadily increasing in the atmosphere. A doubling of the present-day concentration of several of these gases is predicted to occur by the end of the next century. According to the greenhouse theory of climate change, the radiative imbalance caused by trace gas increases will be restored to equilibrium by a warming of the surface-troposphere system. There are many uncertainties concerning trace-gas increases and also with modeling the climate response, but it is generally agreed that the global mean equilibrium surface temperature, due to an increase of CO_2 and other greenhouse gases, is likely to be in the range of 1.5–5.5 °C. This is about the same magnitude as the change of global temperature from the last glacial period to the present interglacial.

The stratosphere is expected to cool by about 10 K in response to a doubling of the CO_2 concentration, because of the enhanced infrared cooling to space. But the stratospheric concentrations of other trace gases, such as CH_4, N_2O, and CFCs, are also increasing at a rapid rate, and these gases are predicted to have an additional influence on the temperature structure, not only because of the effect of blocking thermal infrared radiation, but also because of their ability to destroy ozone through catalytic chemical reactions. Recent calculations have shown that the stratopause will cool from present-day values by −15 to −20 K and ozone will be depleted by −5% to −30% near 40 km by the end of the 21st century (Brasseur and Hitchman, 1988). While ozone depletions at specific altitudes, such as 40 km, are important in determining effects of circulation and temperature profiles, a more important quantity used for policy and determination of biological effects is the percentage depletion of total column ozone. With ozone increasing in the troposphere as a result of man's activities, the change in total column ozone in future years may be small.

The projected CO_2 and CH_4 doublings will also have an effect on the mesosphere and thermosphere. The mesosphere will cool by 10 K and the thermosphere will cool by 50 K. Thus, if the trace-gas releases continue at their projected rate it is expected that the entire atmospheric structure will change somewhat by the end of the 21st century. The present-day atmospheric structure is temporary and quite unique.

GLOSSARY

Chlorofluorocarbons: Man-made chemical compounds that release chlorine in the stratosphere, such as trichlorofluoromethane ($CFCl_3$), or F-11, and dichlorodifluoromethane (CF_2Cl_2), or F-12, and others.

Diffusion: Gas transport due to forces such as pressure gradient and gravity. Eddy diffusion in the atmosphere is the process of transport of gases due to turbulent mixing in the presence of a composition gradient. Molecular diffusion is the net transport of molecules which results from their molecular motions alone in the absence of turbulent mixing; it occurs when the concentration gradient of a particular gas in a mixture differs from its equilibrium value. Eddy diffusion is the most important mixing process in the lower atmosphere, while molecular diffusion becomes significant at the lower pressures of the upper atmosphere.

Electrode Effect: An electrical effect near the Earth's surface where the voltage impressed by the global electric circuit causes a space charge to develop similar to that near an electrode in an electronic vacuum tube.

Galactic Cosmic Rays: High-energy charged particles, primarily protons, that are accelerated by galactic processes and that bombard the Earth's atmosphere and maintain the atmosphere electrically conducting.

Geomagnetic Storm: A period of enhanced magnetic activity caused by intense currents in the auroral ionosphere as detected by ground magnetometers.

Gleisberg Cycle: In addition to the well-known 11-yr cycle in sunspot activity there is a weaker, approximately 80-yr cycle in sunspot activity known as the Gleisberg cycle.

Greenhouse Effect: Certain radiatively active gases in the atmosphere (water vapor, carbon dioxide, ozone) trap infrared radiation emitted by the surface thereby raising the surface temperature of the planet over that which would exist in the absence of an atmosphere.

Ionosphere: The weakly ionized tenuous envelope that surrounds the Earth between about 60- and 1000-km altitude. The charged particles of the region reflect radio waves and are strongly influenced by electric and magnetic fields.

Latent Heat: Energy released when water vapor condenses to liquid water or ice in a cloud.

Magnetosphere: A magnetic cavity in the solar wind that surrounds the Earth and effectively shields the Earth from the continual supersonic outflow of the Sun's ionized upper atmosphere. The charged particles of the region are strongly influenced by electric fields. The magnetosphere overlaps the ionosphere between 60 and 1000 km.

Maunder Minimum: A prolonged period of low sunspot activity that extended from 1645 to 1715 AD.

Mesosphere: That region of the atmosphere which lies above the stratopause (about 47–52 km) and below the mesopause (about 80–90 km) and in which temperature decreases with increasing height; this is the region in which the lowest temperatures of the atmosphere occur.

Odd Hydrogen: A chemical family of active hydrogen compounds such as hydrogen (H), hydroxyl radical (OH), and hydroperoxy radical (HO_2). These form the odd hydrogen or HO_x family ($HO_x = H + OH + HO_2$).

Photochemistry: Chemical processes influenced by sunlight.

Plasma: Gas containing a significant fraction of free electrons and ions.

Quasibiennial Oscillation: An oscillation of the east-west winds in the equatorial stratosphere with an approximate 26-month period.

Radiative Equilibrium: A condition where the thermal structure of the atmosphere is determined by a local balance between heating by absorption of solar radiation and cooling by infrared emission from trace gases.

Solar Cycle: The approximately 11-yr cycle of sunspots.

Solar Flare: A sudden plasma eruption on the Sun that emits a burst of particles that travel outward into interplanetary space.

Solar Wind: Supersonic plasma flow from the Sun.

Stratosphere: The atmospheric shell lying just above the troposphere, characterized by an increasing temperature with altitude. The stratosphere begins at the tropopause (about 10–15 km height) and extends to a height of about 50 km, where the lapse rate changes sign at the stratopause and the beginning of the mesosphere.

Thermosphere: Atmospheric shell extending from the top of the mesosphere to outer space. It is a region of steadily increasing temperature with height, starting at 70 or 80 km. It is below the exosphere and includes most of the ionosphere.

Troposphere: The lowest layer of the atmosphere, ranging from the ground to the base of the stratosphere (tropopause) at 10–15 km altitude depending on the latitude and meteorological conditions. About 70% of the mass of the atmosphere is in the troposphere. This is where most of the weather features occur and where the chemistry of the reactive anthropogenic species released into the atmosphere takes place.

Turbopause: The region of the atmosphere near 100 km where transport by turbulent mixing ceases and transport by molecular diffusion dominates.

List of Works Cited

Bolin, B., Doos, B. A., Jager, J., Warrick, R. A. (Eds.) (1986), *The Greenhouse Effect, Climate Change and Ecosystems,* New York: Wiley.

Brasseur, G., Hitchman, M. H. (1988), *Science* **240**, 634–636.

Brasseur, G., Solomon, S. (1986), *Aeronomy of the Middle Atmosphere,* New York: Reidel.

Garcia, R. R., Solomon, S., Roble, R. G., Rusch, D. W. (1984), *Planet. Space Sci.* **32**, 411–423.

Hays, P. B., Roble, R. G. (1979), *J. Geophys. Res.* **84**, 3291–3305.

Herzberg, L. (1965), *Physics of the Earth's Upper Atmosphere,* New York: Prentice-Hall.

Jursa, A. S. (Ed.) (1985), *Handbook of Geophysics and the Space Environment,* Springfield, VA: National Technical Information Service.

Keating, G. M., Pitts, M. C., Brasseur, G., De Rudder, A. (1987), *J. Geophys. Res.* **92**, 889–902.

Labitzke, K., van Loon, H. (1988), *J. Atmos. Terr. Phys.* **50**, 197–202.

London, J. L. (1980), *Proc. NATO Advanced Study Institute on Atmospheric Ozone,* FAA-EE-80-20, Springfield, VA: National Technical Information Service.

Lorius, C., Barkov, N. I., Jouzel, J., Korotkvich, Y. S., Kotyakov, V. M., Raynaud, D. (1988), *EOS (Madrid)* **69**, No. 26.

Murgatroyd, R. J. (1969), in: G. A. Corby (Ed.), *The Global Circulation of the Atmosphere,* London: Royal Meteorological Society.

Reid, G. C., Gage, K. S. (1988), *Secular Solar and Geomagnetic Variations in the Last 10,000 Years,* Norwell, MA: Kluwer Academic.

Roble, R. G. (1977), in: F. S. Johnson (Ed.), *The Upper Atmosphere and Magnetosphere,* Geophysics Study Committee, Washington, DC: National Academy.

U.S. Standard Atmosphere 1976, Washington, DC: U. S. G.P.O.

Willson, R. C., Hudson, H. S., Frohlich, C., Brusa, R. W. (1986), *Science* **234**, 1114–1116.

WMO (World Meteorological Organization) (1982), *The Stratosphere 1981: Theory and Measurements,* Report No. 11, Geneva, Switzerland: WMO.

Further Reading

Andrews, D. G., Holton, J. R., Leovy, C. B. (1987), *Middle Atmosphere Dynamics,* New York: Academic.

Barry, R. G., Chorley, R. J. (1982), *Atmosphere, Weather and Climate,* New York: Methuen.

Goody, R. M., Yung, Y. L. (1989), *Atmospheric Radiation: Theoretical Basis,* New York: Oxford University.

Holton, J. R. (1972), *An Introduction to Dynamic Meteorology,* New York: Academic.

Johnson, F. S. (Ed.) (1977), *The Upper Atmosphere and Magnetosphere,* Geophysics Study Committee, Washington, DC: National Academy.

Krider, E. P., Roble, R. G. (Eds.) (1986), *The Earth's Electrical Environment,* Geophysics Study-Committee, Washington, DC: National Academy.

Liou, K. N. (1980), *Introduction to Atmospheric Radiation,* New York: Academic.

Rees, M. H. (1989), *Physics and Chemistry of the Upper Atmosphere,* Cambridge, England: Cambridge University Press.

13
Atomic Spectrometry and Elemental Analysis

Bernhard Welz and Daniel L.G. Borges

13.1 Introduction 423
13.2 General Considerations 424
13.2.1 Atomic Spectra 424
13.2.2 Interferences 425
13.3 Optical Atomic Spectroscopy 426
13.3.1 Basic Principles 426
13.3.1.1 Physics of Atomization and Excitation 426
13.3.1.2 Atomic Spectra and Spectral Lines 427
13.3.1.3 Measurement of Radiation, Calibration, and Evaluation 428
13.3.2 Optical Emission Spectrometry 429
13.3.2.1 Excitation Sources 430
13.3.2.2 Spectrometers 436
13.3.2.3 Detectors 438
13.3.3 Atomic Absorption Spectrometry 439
13.3.3.1 Atomizers 440
13.3.3.2 Spectrometers for Line Source AAS 446
13.3.3.3 Spectrometers for High- Resolution Continuum Source AAS 452
13.3.4 Atomic Fluorescence Spectrometry 460
13.3.4.1 General Principles 460
13.3.4.2 Instrumentation 461
13.4 X-Ray Spectroscopy 463
13.4.1 Theoretical Principles 464
13.4.1.1 Emission of X-Rays 464
13.4.1.2 Absorption of X-Rays 466
13.4.1.3 X-Ray Fluorescence (XRF) 467
13.4.2 Instrumentation 467
13.4.2.1 X-Ray Fluorescence 467
13.4.2.2 Total-Reflection (TR) XRF 469
13.4.2.3 Particle-Induced X-Ray Emission (PIXE) 471
13.4.2.4 Portable Instrumentation 471
13.5 Outlook 471

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

Acknowledgments 472
Acronyms 472
References 473

13.1
Introduction

The early history of atomic spectroscopy is closely connected with the observation of the sunlight. Wollaston (1802) reported seeing dark lines in the sun's spectrum, but he could not give a satisfactory explanation for them. Fifteen years later, Fraunhofer designed the first grating ruling engines and succeeded in producing transmission gratings with up to 600 grooves per centimeter; he noted the same lines and proceeded to map the dark lines. As spectra produced by gratings, unlike prism spectra, are nearly linear with wavelength, Fraunhofer (1817) succeeded in calculating the wavelength of the dark lines with remarkable accuracy.

During the first half of the nineteenth century, a good deal of experimentation took place with colored flames generated by injecting salts into alcohol burners or oil lamps as excitation sources. Herschel (1823) and Talbot (1826) succeeded to identify alkali metals using this technique. A breakthrough occurred in 1855, when Bunsen designed an efficient burner for coal gas, in which fuel and air were mixed before the mixture was ignited. The new flame was much steadier and gave a higher temperature, making possible a much more efficient study of the emitted spectra.

Soon after, a close collaboration began between Kirchhoff, professor of physics, and Bunsen, professor of chemistry, both at the University of Heidelberg, Germany. Kirchhoff compared the Fraunhofer lines in the sun's spectrum to those obtained in the laboratory, and showed that they arose from the same elements. He explained the reversed appearance of the lines as due to a process of absorption as the emission rays passed through the cooler outer layer of the sun's atmosphere, which caused them to show up dark against the bright background. Kirchhoff (1860) also concluded that absorption occurs only at the same wavelengths as emission, a phenomenon called *resonance*.

Among the most important stages in the development of emission spectroscopy are the first measurements in the ultraviolet (UV) made by Mascart (1864) with a photographic plate, which have later been extended to the low-vacuum UV. Balmer (1885), with his equation for the Balmer series of hydrogen lines, started the search for an explanation for the origin of atomic spectra, which was continued by several

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

other researchers. It was the work of Bohr (1913), with his concept of the "astronomical" atom that placed the origin of spectra on a firm theoretical foundation. Later, Sommerfeld (1916) proposed elliptical orbits and the Schrödinger wave equation finally permitted calculation of the quantum states necessary to describe the energy positions of electrons in atoms. The development of X-ray spectroscopy is mainly based on the work of Barkla, Moseley, and Siegbahn, who received the Nobel price for physics in 1924.

13.2
General Considerations

13.2.1 Atomic Spectra

Atomic spectra are due to radiative transitions of electrons between the ground state and/or excited states of an atom. Three forms of radiative transition are possible: (i) *spontaneous emission* for the transition from a higher excited state to a lower state after the transfer of kinetic or electrical energy to the atom; (ii) *absorption of radiation* with a corresponding transition of a lower state to a higher excited state; and (iii) *induced emission* for the transition from a higher excited state to a lower state after the absorption of external radiation of corresponding frequency (the reversed process to absorption of radiation); the latter is generally referred to as *atomic fluorescence*. Atomic spectra are line spectra and are specific to the absorbing or emitting atoms (elements), that is, the spectra contain information on the atomic structure. An electron in an atom can only absorb and emit energy at well-defined wavelengths, which correspond to the energy difference between the two levels of the transition. Absorption spectra are generally simpler than emission spectra, as absorption usually starts from the ground state of the neutral atom according to selection rules, whereas all kinds of excited states can be reached in thermal and electrical excitation, resulting in a whole series of de-excitation pathways, and a correspondingly greater number of emission lines, as shown in Figure 13.1 for the spectrum of sodium.

Transitions of electrons of the outer shell are found in the spectral range between 100 nm and 1 mm, and are referred to as *optical spectroscopy*; transitions of electrons from inner (K, L, and M) shells are found in the spectral range between about 10 pm and 10 nm, and are referred to as *X-ray spectroscopy*.

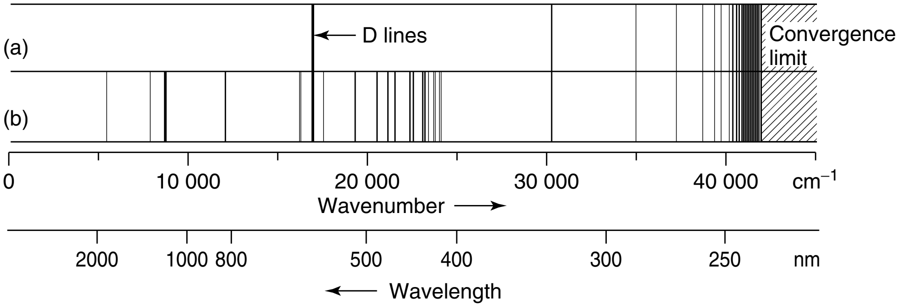

Fig. 13.1 The spectrum of the sodium atom; (a) in absorption and (b) in emission. (From Welz and Sperling (1999), Wiley-VCH, p. 67; reproduced with permission.)

The term *spectroscopy* – at least in atomic spectroscopy – refers to the visual evaluation of spectra using a dispersive system; however, it is also used as the general term for all spectroscopic measurement and observation procedures. The term *spectrometry* refers to the photoelectric measurement and evaluation of spectral radiation units using a dispersive spectral apparatus.

13.2.2
Interferences

The presence of concomitants (matrix) in the test sample can cause interferences in the determination of the analyte. The interference can influence both the attainable precision and the trueness. However, interference only leads to a measurement error if it is not eliminated, or if it is not taken into consideration in the evaluation process, by suitable measures. Such measures can be instrumental, such as background correction, or the choice of a suitable calibration technique, such as the analyte addition technique, or the use of chemical additives.

Interferences in spectrochemical analysis can be classified according to a number of aspects. The various categories are not mutually exclusive but rather complement each other in characterizing the interference. Frequently, the distinction is made between *specific* (for the analyte) and *nonspecific* interferences or between *additive* (the measured signal is increased or reduced by the same value independent of the analyte concentration or mass) and *multiplicative* interferences (the slope of the calibration curve is affected by concomitants). In the following, we shall largely adhere to International Union of Pure and Applied Chemistry (IUPAC) recommendations (IUPAC, 1978), according to which interferences are divided generally into *spectral* and *nonspectral interferences*.

Spectral interferences are due to the incomplete isolation of the radiation emitted or absorbed by the analyte from other radiation or radiation absorption detected and processed by the electrical measuring system. Spectral interferences may be due to (i) direct or partial overlapping of the analytical line with the emission or absorption line of another element; (ii) emission of radiation by concomitants that is not overlapping with the analyte line, but is not separated by the monochromator/polychromator; (iii) emission or absorption of the radiation of the analytical line by gaseous molecules; and (iv) radiation scattering caused by particles in the observation volume. Spectral interferences are always additive interferences; details are discussed with the specific techniques, where the interferences appear to a significantly different degree.

Nonspectral interferences are those that cause an influence on the number of analyte atoms (absolute or per time unit) in the observation volume, and thus on the measurement value. Nonspectral interferences are most conveniently classified according to the place or time of their occurrence. The most common types of nonspectral interferences are (i) *transport interferences*, that is, interferences that occur during the transport of a (liquid) sample from the sample container to the atomizer/excitation source due to the physical properties of the solution; (ii) *solute-volatilization interferences*, which are due to incomplete transfer of the sample into gaseous species, an interference that is typically limited to atomizers/excitation sources of relatively low temperature; (iii) *vapor-phase interferences*, which include all the equilibrium processes that may occur in the vapor phase, such as *dissociation, ionization,* and *excitation interferences*. If an interference cannot be classified because its cause is unknown or it is of complex nature, we speak of a *matrix effect*. Other interferences that are typical for specific atomizers and excitation sources are discussed in connection with these atomizers.

13.3
Optical Atomic Spectroscopy

13.3.1
Basic Principles

13.3.1.1 Physics of Atomization and Excitation

In the atomizer, the sample to be analyzed is dissociated into atoms; the *degree of dissociation* depends on the temperature of, and on chemical reactions within the atomizer. The *ground state* is the lowest energy state of an atom. By absorption of additional thermal, electrical, or optical energy, a neutral atom can be transferred into an excited state or it can be ionized. The *excited state* is unstable; the atom returns to a lower energy state after a lifetime of typically 10^{-9}–10^{-8} seconds. The transitions are limited by selection rules. *Ionization* refers to the complete separation of one or more electrons from an atom, so that the remaining atom becomes electrically charged. The degree of ionization is the ratio between the number of ions and that of the originally present number of atoms of the element under consideration.

If thermal equilibrium is established, the fraction of atoms N_j in an excited stage j can be calculated by statistical mechanics. The fraction of excited atoms at a certain energy level in comparison to the number of atoms in the ground state N_0 is given by the *Boltzmann distribution*

$$\frac{N_j}{N_0} = \frac{P_j}{P_0} e^{-E_j/kT} \qquad (13.1)$$

where P_j is the statistical weight of the selected excited state j, P_0 the statistical weight of the ground state, E_j is the energy of excitation, k is the Boltzmann constant, and T the absolute temperature. In Eq. (13.1), the exponent is inversely proportional to the absolute temperature, which means that the increase in the relative number of excited atoms N_j/N_0 is exponential with increasing temperature. Walsh (1955) calculated the ratio N_j/N_0 for a number of elements at various temperatures, as is shown in Table 13.1. For most of the elements and temperatures, the number of atoms in the ground state N_0 is virtually identical to the total number of atoms N.

Tab. 13.1 Temperature and wavelength dependence of the ratio N_j/N_0 (Walsh, 1955).

Element	Excitation energy eV	Wavelength nm	2000 K	N_j/N_0 3000 K	4000 K
Zn	5.80	213.9	7.29×10^{-15}	5.58×10^{-10}	1.48×10^{-7}
Ca	2.93	422.7	1.21×10^{-7}	3.69×10^{-5}	6.03×10^{-4}
Na	2.11	589.0	0.86×10^{-4}	5.88×10^{-4}	4.44×10^{-3}
Cs	1.46	852.1	4.44×10^{-4}	7.24×10^{-3}	2.98×10^{-2}

13.3.1.2 **Atomic Spectra and Spectral Lines**

Dispersion of the radiation emitted or absorbed by the atomizer/excitation source results in a spectrum that is characteristic of the emitting or absorbing species. Atomic spectra are line spectra; molecular spectra are band spectra that extend over a wider spectral range. A *spectral line* is defined as the monochromatic radiation that is emitted or absorbed at the transition of an atom or ion between different energy states E_1 and E_2. The wavelength λ of the transition is characteristic for the two energy states according to

$$\lambda = \frac{hc}{E_2 - E_1} \tag{13.2}$$

where h is Planck's constant and c is the speed of light. Spectral lines, the radiation of which can be absorbed by atoms in the ground state or that are generated by a transition into the ground state, are called *resonance lines.*

Atomic lines are not strictly monochromatic, but characterized by a distribution of the radiation intensity over wavelength that can be described by a profile function, as is schematically shown in Figure 13.2. The width of an atomic line is usually given as the full width at half maximum (FWHM), or briefly *half width.* The part of the line that is within the FWHM is called the *line core*; the parts that are outside are called the *line wings.* The *natural line width* is the smallest possible line width, which is caused by the residence time of the atom in the involved energy states and the resulting uncertainty of the measurement. The natural width of atomic lines is typically of the order of 0.01–0.1 pm, which is, however, not of practical significance due to the various broadening effects that occur under normal analytical conditions. The most important effects that finally determine the *physical line width* are Doppler and pressure broadening. The *Doppler broadening* is caused by the random movement of the emitting or absorbing atoms; it is proportional to the square root of the temperature and results in a symmetric broadening of the line. *Collision broadening* is caused by collision of the emitting or absorbing atoms with other neutral or charged particles; it depends on the pressure and temperature of the environment and results in an asymmetric broadening. Depending on the mass of the atom and the collision partner this effect may shift the line

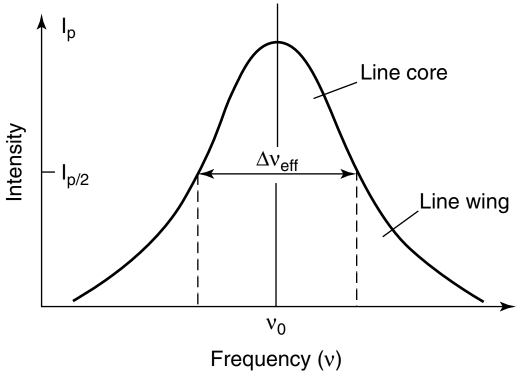

Fig. 13.2 Typical line profile and half width (FWHM, $\Delta\nu_{\text{eff}}$) of a spectral line. (From Welz and Sperling (1999), Wiley-VCH, p. 74; reproduced with permission.)

maximum to lower or higher wavelength, an effect called *Lorentz shift*.

The physical width of an atomic line depends on a variety of factors (besides individual differences in the hyperfine structure). Lines in the vacuum UV are approximately one order of magnitude narrower than lines in the beginning infrared. Lines emitted by low-pressure lamps, such as hollow cathode lamps (refer to Section 13.3.3.2, Radiation Sources for LS AAS), essentially only exhibit Doppler broadening, no Lorentz shifts, and typically have half widths between 0.3 and 1.5 pm. Lines emitted by a high-temperature plasma (refer to Section 13.3.2.1, Plasma Sources) are typically one order of magnitude broader, and the absorption lines in flames are in between with half widths of 1.5–5 pm.

13.3.1.3 **Measurement of Radiation, Calibration, and Evaluation**

The intensity of the emitted or absorbed radiation is measured using a *photoelectric detector* that converts the flux of photons into a flux of electrons using the outer or inner photo effect. The most common detectors are the photocathode detector, commonly denominated as photomultiplier tube (PMT) that uses the outer photo effect and the semiconductor detectors that use the inner photo effect. The latter ones come in a variety of array detectors that are increasingly used in modern equipment.

The measurement of signals in optical emission spectrometry (OES) and atomic fluorescence spectrometry (AFS) is typically done in emission or fluorescence intensity with the dimension counts per second (s^{-1}). In atomic absorption spectrometry (AAS), in contrast, measurement is done in absorbance, A, according to the Beer–Lambert law, which, for AAS, is usually written in the form:

$$A \equiv \log \frac{\Phi_0(\lambda)}{\Phi_{\mathrm{tr}}(\lambda)} = 0.43 N \ell \kappa(\lambda) \quad (13.3)$$

where $\Phi_0(\lambda)$ is the incident radiant power, $\Phi_{\mathrm{tr}}(\lambda)$ is the transmitted radiant power, N is the total number of free atoms, ℓ is the length of the absorbing layer, and $\kappa(\lambda)$ is the spectral atomic absorption coefficient. In addition to the absorbance A, which is typically used for steady-state signals, for example in flame atomic absorption spectrometry (FAAS) the time-integrated absorbance, A_{int} (which corresponds to the area under the absorption pulse), is used for the evaluation of transient signals, as they are typical for graphite furnace atomic absorption spectrometry (GF AAS). Whereas the absorbance is a dimensionless measure, the integrated absorbance has the dimension seconds.

None of the techniques used for atomic spectroscopy is an absolute technique, which means that, for quantitative determinations, all of them require calibration using appropriate *calibration standards*. These could be, in an ideal case, aqueous solutions with a known content of the analyte element, or, in the case of severe matrix effects, have to be reference materials with analyte content and matrix composition close to that of the test sample to be analyzed. A *calibration function* that describes the mathematical dependence of the measured value on the concentration or mass of the analyte has to be established for quantitative determination. The slope of the calibration function is termed *sensitivity*.

Table 13.2 shows the different measurement units, their dimension and the resulting units for the sensitivity of different atomic spectrometric techniques.

Tab. 13.2 Examples for the dimension of sensitivity in different atomic spectrometric techniques.

Technique	Measured quantity		Dimension	Concentration or mass[a]	Sensitivity
OES	Intensity	I	s^{-1}	$mg \cdot L^{-1}$	$L \cdot mg^{-1} \cdot s^{-1}$
FAAS	Absorbance	A	–	$mg \cdot L^{-1}$	$L \cdot mg^{-1}$
GF AAS	Integrated absorbance	A_{int}	s	ng	$s \cdot ng^{-1}$

[a] Other concentration or mass units might be used as well.

This comparison clearly demonstrates that "sensitivity" is not an appropriate parameter to compare the analytical performance of different systems due to the entirely different definitions. In addition, sensitivity is only one aspect, but noise is at least as important. Hence, the comparison of different systems should be on the basis of the signal-to-noise (S/N) ratio exclusively. Appropriate measures of the S/N ratio are the limit of detection (LOD), and the limit of quantification (LOQ), which are usually defined as 3 times and 10 times, respectively, the standard deviation of a blank, divided by the slope of the calibration function. Obviously, LOD and LOQ should be compared for real samples, not only for matrix-free calibration solutions. Another important parameter for OES is the signal-to-background ratio, and the figure for comparison is usually the background equivalent concentration (BEC), that is, the analyte concentration that gives the same signal as the background.

The measurement of the emission intensity, particularly in OES, is subject to relatively great day-to-day variations due to minor changes in the temperature of the excitation source and in detector sensitivity. The measurement of the absorbance in AAS, in contrast, is relatively independent of such fluctuations, firstly, as the number of atoms in the ground state is essentially independent on minor temperature changes (refer to Section 13.3.1.1) and the intensity of the radiation from the primary source has no influence on the sensitivity either, as the absorbance only depends on the ratio $\Phi_0(\lambda)/\Phi_{tr}(\lambda)$. Measurement quantities, such as the characteristic concentration, c_0, the concentration of an analyte that gives an absorbance reading of $A = 0.0044$ in FAAS, and the characteristic mass, m_0, the mass of an analyte that gives an integrated absorbance reading of $A_{int} = 0.0044$ seconds in GF AAS may therefore be used to test the proper function of the equipment and of analytical procedures on a day-to-day basis.

13.3.2
Optical Emission Spectrometry

OES, often referred to as atomic emission spectrometry (AES) – an acronym the use of which has been discouraged by IUPAC because it is reserved for Auger electron spectroscopy (AES) – is the oldest of the atomic spectrometric techniques. A great deal of research has been carried out in this area in the late nineteenth and the early twentieth century, including the design of ever more advanced mono- and polychromators. Modern quantitative methods of spectrochemical analysis using OES, however,

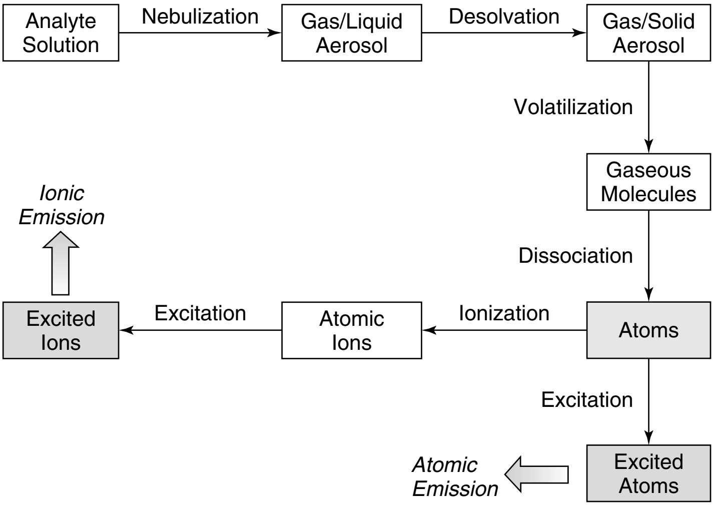
Analyte Solution
Nebulization
Gas/Liquid Aerosol
Desolvation
Gas/Solid Aerosol
Volatilization
Gaseous Molecules
Dissociation
Ionic Emission
Excited Ions
Excitation
Atomic Ions
Ionization
Atoms
Excitation
Atomic Emission
Excited Atoms

Fig. 13.3 The process of atomization and excitation for a liquid/dissolved sample.

did not begin until about 1925, when the first commercial optical equipment began to be manufactured in Europe and the United States. A photographic plate was used as detector in these instruments, a practice that continued for about half a century.

An optical emission spectrometer consists basically of an atomization and excitation source with a sample introduction system, a wavelength-dispersing system, such as a monochromator or polychromator, and a detector, amplifier, and readout system. The kind of samples that can be analyzed and the elements that can be determined depend mostly on the excitation source.

A liquid sample, when introduced into the excitation source, is submitted to several steps that are shown schematically in Figure 13.3. Initially, the sample aerosol, which is formed by a nebulizer, is desolvated, that is, the solvent is evaporated. The desolvated particles then are vaporized, that is, converted to gaseous species, and finally dissociated in the atomization process. The atoms formed in this process are in equilibrium with their respective excited (neutral) and ionic species, which may also undergo excitation (Boss and Freeden, 1999). In the cases of direct solid sample and gas analysis, some of the above stages might be omitted or altered.

13.3.2.1 Excitation Sources

The role of the excitation source in OES is to transfer the sample into the gaseous state, to dissociate molecular species, and to provide sufficient energy to excite the analyte atoms, promoting the emission of photons that will later be detected.

Electrical Arcs and Sparks Many analytical applications of OES produced their spectra by arc or spark excitation techniques. The historical development in this area is most difficult to document since almost from the start, after observations of the sun's spectrum, the attempt was to use high-energy sources. The first commercial instruments for OES that have been introduced in the first half of the twentieth century also used electrical arcs and sparks for the direct vaporization and excitation of solid samples. This kind of equipment is the working horse in most metallurgical laboratories even today. An *arc* is an electrical discharge that takes place between two or more conducting electrodes. The sample usually stands as one of the electrodes, either as a powder, a solid mixture, or a solution residue.

One of the most common arc types used in OES is the direct current (*dc*) *arc*. The dc arc consists of a continuous discharge of 1–30 A between a pair of metal or graphite electrodes, one of which contains the sample. The dc arc provides good detection power and generates relatively high amounts of neutral (excited) atoms, although it is also characterized by relatively poor precision and by an effect denominated *selective volatilization*, which occurs because the electrodes are only gradually heated by the arc. The consequence is that the most volatile materials vaporize first, which may result in an erroneously high result for volatile analytes. A *spectrochemical buffer* is frequently used to reduce or control the effects of selective volatilization; substances that are typically employed include graphite powder or alkali and alkaline–earth compounds (Ingle and Crouch, 1988; Mika and Török, 1974; Zhou, Zhou and Hou, 2005).

The alternating current (*ac*) *arcs* may be either operated at high (2–4 kV) or low (100–400 V) voltages. Unlike the dc arc, the ac arc is extinguished at the end of each half-cycle, which allows the sampling of different portions of the electrode at each cycle. Compared to dc arcs, ac arcs lead to better precision and more efficient sample volatilization, whereas the sensitivity is usually lower and the analysis of refractory materials may become more difficult (Ingle and Crouch, 1988).

Sparks are intermittent electrical discharges of a few microseconds that occur under application of high electrical potentials. Spark sampling usually consists of two stages: firstly, a low-energy discharge is produced to ionize the noble gas (usually Ar) in the sampling chamber; later, after the conducting plasma has been generated, a high-energy discharge (spark) is induced, causing vaporization of the sample in the location where the spark impacts, where temperatures as high as 50 000 K can be reached (Ingle and Crouch, 1988; Mika and Török, 1974). Spark sources are characterized by relatively high precision, as they are able to vaporize several distinct portions of the sample electrode in successive discharges. Spark-source OES instruments are, however, usually less sensitive than arc and discharge instruments, mostly because the high-peak currents and high-power density in sparks lead to a high degree of single and even multiple ionization of analyte species. In addition, the mass of sample that is vaporized in each cycle is relatively low. Despite those drawbacks, spark sources have been successfully applied to the analysis of samples, such as environmental, geological, and biological materials, but the main focus has been, and is, the analysis of metal alloys (Zhou, Zhou and Hou, 2005).

One of the limitations of arc and spark OES is the need of reference materials with a matrix composition and analyte content as close as possible to the sample for calibration due to the significant matrix effects encountered with these techniques. This requirement might increase the cost of analysis significantly.

Glow Discharges Glow discharges (GDs) have been used in OES since the beginning of the twentieth century. Usually, the sample to be analyzed was filled into a hollow cathode, which was placed inside a quartz tube through which an inert gas (usually Ar) was pumped at low pressure. A voltage potential was then applied between the anode (usually made of graphite) and the hollow cathode, causing electrons to migrate from the cathode to the anode. In their path they collide with Ar atoms, generating positively charged metastable argon species, which are accelerated to the sample holder, held at negative potential. Collision with the sample surface will result in the ejection of atoms and electrons from the sample, a process called *sputtering*. Sputtered atoms collide with metastable Ar atoms or high-energy electrons, and are excited. With relaxation, the photons emitted from the excited species produce a small plasma (GD), which is used for analytical purposes (Pisonero *et al.*, 2006; Angeli *et al.*, 2003; Jakubowski *et al.*, 2007). In the second half of the twentieth century glow discharge optical emission spectrometry (GD OES) became quite popular after Grimm (1968) had proposed several modifications of the setup. He used a flat cathode (the sample) and a ring anode, as shown in Figure 13.4, which significantly facilitated the preparation and change of samples. GD OES is ideally suited for surface analysis and depth profiling, as it produces a crater with a flat bottom, which is in contrast to other ablation techniques. Some typical applications of GD OES include the analysis of thin films (Pisonero *et al.*, 2006; Angeli *et al.*, 2003) and metal alloys (Galindo and Albella, 2008).

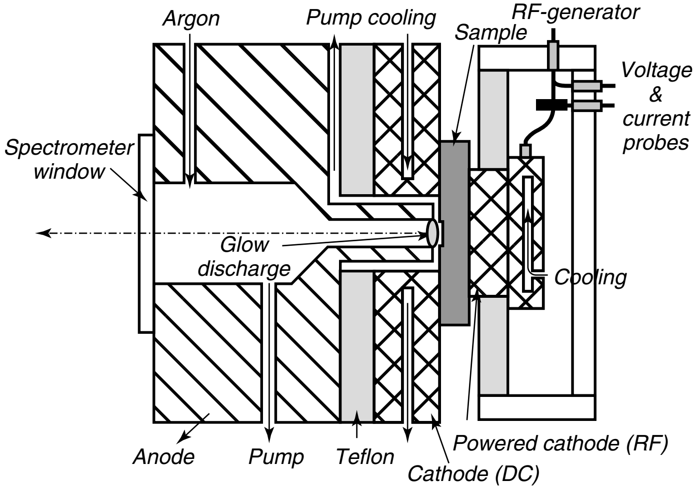

Fig. 13.4 Schematic design of a Grimm discharge lamp for GD OES. (From Hodoroba *et al.* (2001), p. 31; reproduced with permission from Elsevier.)

Flames The simplest excitation source is a *flame*, a mixture of at least two gases, a fuel and an oxidant, and the first burner designed for atomic spectroscopy was the burner of Bunsen. The work that gave principal impetus to flame excitation was that of Lundegårdh (1936). In his method, the sample solution was sprayed from a nebulizer into a condensing chamber and then into an air–acetylene flame. Since spectra produced by flames exhibit significantly less lines compared to those produced by arc and spark emission, simple devices for spectral isolation could be used. Barnes *et al.* (1945) reported a simple filter photometer for alkali metal determinations; instruments of this kind are still in use nowadays, mainly in clinical laboratories, and the technique is often referred to as *flame photometry*.

Flame OES is limited to the analysis of liquid or dissolved/digested samples, and a nebulizer is typically required for sample introduction into the flame. The nebulizer-burner unit might be similar to the design used for FAAS (refer to Section 13.3.3.1, Flame Atomizers) or it might be of the *total consumption* design that makes possible to use oxygen as the oxidant and hence to achieve higher flame temperatures. Among the advantages of flames as excitation sources are their stability and the relatively simple and inexpensive operation. However, the temperatures of flames are, in most cases, not sufficient to promote the excitation of a significant fraction of analyte atoms, as can also be deduced from the data in Table 13.1 (Ingle and Crouch, 1988). The use of flames for OES is restricted to a few isolated applications nowadays.

Plasma Sources One of the most widely used excitation sources in OES nowadays is the *inductively coupled plasma* (ICP), which has been investigated independently by Greenfield, Jones and Berry (1964) in the UK and Wendt and Fassel (1965) in the USA. The ICP is a partially ionized gas (usually argon) that is formed from an electrical discharge inside a quartz torch and maintained by an induction coil to which a radiofrequency (RF) power is applied. The schematic design of an ICP torch, which consists of three concentric cylinders, is shown in Figure 13.5. The main gas flow (usually at flow rates above $15\,\mathrm{l\,min^{-1}}$), which is the main flow responsible for plasma formation, is in the outer cylinder. A secondary gas flow (usually at flow rates of $1-1.5\,\mathrm{l\,min^{-1}}$) runs between the inner and the middle cylinder and is responsible for cooling the torch and maintaining a distance between the plasma and the torch walls. The *nebulizer gas* flows inside the inner cylinder, being responsible for the transport of the sample aerosol to the plasma itself. To initiate the plasma, a discharge takes place, generating Ar ions and electrons. An RF power is then applied to the metallic load coil located in the front part of the plasma torch, originating a magnetic field that maintains the electrons on the torch, promoting their collision with other neutral Ar atoms. The collisions generate more electrons and ionized gas species in a chain reaction, and the otherwise neutral gas becomes a plasma, constituted by neutral atoms, ions, and electrons. Owing to the high temperatures that can be attained in this system (up to about 10 000 K, depending on the plasma region), basically all elements can be excited and even ionized to a great extent in an ICP. The same ICP source can be used with a

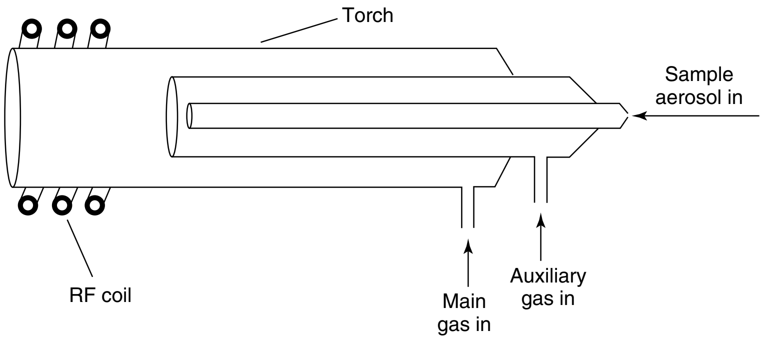

Fig. 13.5 Schematic design of an ICP torch; for details see text.

mass spectrometer, in a technique called *inductively coupled plasma mass spectrometry*, (*ICP-MS*) (Montaser and Golightly, 1992; Becker, 2007).

In inductively coupled plasma optical emission spectrometry (ICP OES), liquid samples are used almost exclusively. They are introduced into the plasma using a pneumatic nebulizer and a spray chamber in order to obtain an aerosol with very small droplet size. A wide variety of different nebulizers and spray chambers have been proposed for ICP OES (Nölte, 2003), but they are not discussed in detail here. The most commonly used one is the Meinhard nebulizer, a concentric nebulizer of very simple design, in combination with a Scott spray chamber, as shown in Figure 13.6. Another frequently used nebulizer is the *cross-flow* nebulizer, where a capillary with a high-pressure gas flow is positioned at a 90° angle with the capillary tip through which the

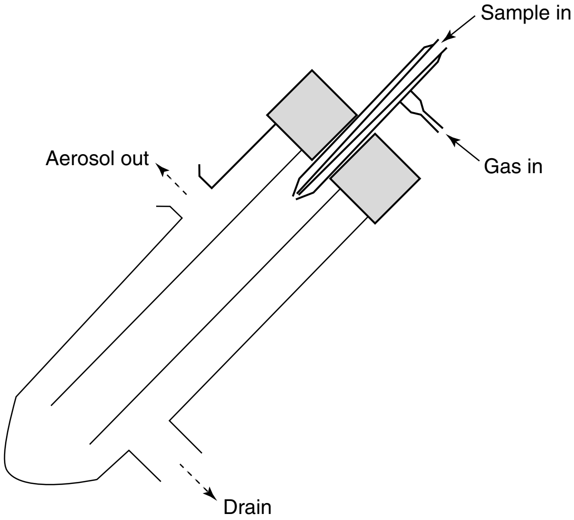

Fig. 13.6 Schematic design of a Meinhard concentric nebulizer in a Scott spray chamber for ICP OES.

sample flows. Both nebulizers have an uptake rate of about 2 ml min^{-1} and are typically operated in combination with a peristaltic pump for sample introduction in order to reduce the effect of viscosity of the solution on the flow rate. As a consequence of the requirement for very low volume droplets in order to maintain plasma stability, the sample introduction efficiency is around 2%, so that most of the aspirated sample is discarded through the drain (Boss and Freeden, 1999).

There are basically two different configurations possible for ICP OES, the *radially viewed plasma* (or side-on), where the ICP is positioned at a 90° angle with the detector, and the *axially viewed plasma* (or end-on), where the ICP and the detector are assembled in the same optical axis. The latter configuration provides higher sensitivity due to the longer path length that is exposed to the detector, although the risk of interference with this configuration is considerably higher (Boss and Freeden, 1999). The major problem in ICP OES, similar to arc and spark OES, is the extremely great number of emission lines that is produced at the high plasma temperatures, which obviously increases the risk of spectral interferences proportionally. Modern simultaneous ICP OES instruments offer computer programs to correct for spectral interferences, measuring the interferent emission at a suitable line, calculating and subtracting its contribution to the emission at the analytical line using a least-squares algorithm.

An alternative for sample introduction in ICP OES is the use of a graphite furnace specially designed to carry the sample aerosol from the graphite tube to the plasma source. This technique is known as *electrothermal vaporization* (ETV), and works under the same basic principles as the graphite furnace technique used in AAS (refer to Section 13.3.3.1, Electrothermal Atomizers). Among its advantages are the higher transport efficiency and the greater tolerance to matrix components, allowing even the analysis of solid samples directly or prepared as slurries. Most ICP OES instruments, however, have difficulties to handle transient signals, which are generated by ETV. As a consequence, the use of ETV devices became more popular in ICP-MS.

Another technique that is occasionally used in ICP OES is hydride generation (refer to Section 13.3.3.1, Atomizers for Chemical Vapor Generation) in order to separate the analyte from the matrix and introduce it into the plasma in the form of a gaseous compound. In this case, the plasma conditions have to be adapted in order to tolerate the hydrogen that is developed during the chemical reaction.

Other plasma sources have also been investigated as excitation sources in OES. The *direct current plasma* (DCP) was the first plasma source described and commercialized. In this system, the plasma is generated when a direct current is established between two fixed electrodes, and the sample aerosol is carried to the plasma region by an argon flow. The *microwave-induced plasma* (MIP) is also an important plasma source. In this case, the plasma is generated by application of microwave power, generated by a magnetron, to a cavity with an inert gas, usually Ar or He (Rosenkranz and Bettmer, 2000). Microplasmas with dimensions of 1 mm or less, usually microwave-induced, generated by high frequency or dc discharges, are a promising field of research in OES, and have been experiencing increasing popularity in the past few years (Foest, Schmidt and Becker, 2006; Broekaert and Siemens, 2004).

Lasers The most recent excitation sources in OES are lasers used in *laser-induced breakdown spectroscopy* (LIBS). In a very simplified definition, Sallé, Mauchien and Maurice (2007) described LIBS as a technique based on the focusing of high-power laser pulses onto a sample surface leading to the generation of a plasma composed of excited species, which emit radiation. Nd : YAG lasers are typically used at 1064 nm, but the results obtained with UV lasers are not significantly different (Fantoni *et al.*, 2008). The technique uses short-pulsed lasers (in the nanosecond range) to induce the dielectric breakdown of a gas located at the laser focal point. The microplasma that is generated as a consequence can reach temperatures as high as 20 000 K, which is sufficient to break chemical bonds present in the sample and to produce excited atoms and ions. The results obtained using LIBS depend on several factors, which include the surrounding gas, sample matrix, and laser parameters, such as wavelength, power density, and pulsing frequency (Rusak *et al.*, 1998; Kennedy, Hammer and Rockwell, 1997). The applications of LIBS for qualitative and quantitative chemical analysis range from depth profile studies (Das *et al.*, 2008) to the analysis of paints (Kim *et al.*, 2007) and meteorites (De Giacomo *et al.*, 2007). Fantoni *et al.* (2008) published a review article about semiquantitative and quantitative analysis using LIBS. Portable LIBS spectrometers have also been described and applied successfully to a number of elements (Harmon *et al.*, 2005; Mohamed, 2008; Goujon *et al.*, 2008).

13.3.2.2 **Spectrometers**

A spectrometer is an apparatus for the conduction and dispersion of the emitted radiation and a system for the measurement of spectral radiation quantities. We distinguish between *sequential spectrometers*, where the radiation of different wavelengths is measured one after the other, and *simultaneous spectrometers*, where the radiation of different wavelengths is measured at the same time, using more than one detector or a single space-resolving detector. The sequential spectrometer uses a *monochromator*; the simultaneous spectrometer uses a *polychromator*. Lenses and mirrors are used for the conduction and collimation of the radiation; prisms and gratings are used for their dispersion. Monochromators have one entrance and one exit slit; polychromators have one entrance slit and no exit slit in case a space-resolving detector is used, or multiple exit slits in case where more than one detector is used. Because of the great number of lines that are emitted in OES, particularly from high-temperature excitation sources, the quality of the spectrometer and the relative freedom from spectral interferences is directly related to its resolving power.

The most popular mounting used for monochromators in OES is the Czerny–Turner mounting, which is shown in Figure 13.7; the Czerny–Turner mounting offers the best correction for coma and astigmatism, and hence the lowest stray light level. Monochromator-based instruments are usually compact, show greater spectral flexibility and allow background correction to be performed more accurately compared to polychromators.

In the case of *polychromators*, we have to distinguish between three different

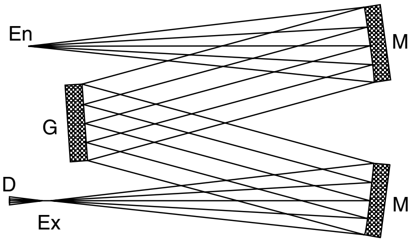

Fig. 13.7 Schematic design of a monochromator in Czerny–Turner mounting with flat grating. (From Welz and Sperling (1999), p. 62; reproduced with permission from Wiley-VCH.)

designs, the classical spectrograph, the conventional prism or grating polychromator, and the echelle grating polychromator. The *spectrograph* used a prism or a grating as the dispersive element, usually in Littrow, Ebert, or Czerny–Turner mounting, and it often had a focal length of several meters in order to obtain high resolution. It had no exit slit and a photographic plate or film as the detector, which recorded the entire spectral range of interest. Spectrographs have played a decisive role in research and routine analysis up to the 1970s, but are no longer in use because of their complexity and the time-consuming evaluation of the photographic plates.

In *conventional polychromators* with a prism or a grating and multiple exit slits, each exit slit is assigned to a specific wavelength (element) and equipped with its own detector, hence making simultaneous multielement determination possible. One of the most popular configurations used for polychromators in OES is the Paschen–Runge mount assembled in a *Rowland circle*, as shown in Figure 13.8. In this system, radiation from the excitation source is conducted toward a concave diffraction grating, and the resolved spectrum is directed to several exit slits located on the opposite side of the circle. An individual detector is associated to each exit slit. Conventional polychromator-based instruments usually have higher sample throughput, although there are restrictions with regard to the number of elements that can be determined, which have to be specified when the equipment is purchased. There are also limitations with respect to background correction, as the lines to be used for correction have to be defined in advance as well, which increases the risk of spectral interferences (Broekaert, 2005; Boss and Freeden, 1999). Simultaneous spectrometers with Rowland circle mounting and a series of space-resolving detectors instead of PMT obviously do not show the above limitations.

Many modern ICP OES instruments use an *echelle grating polychromator* for wavelength dispersion. Echelle gratings are used because of their much higher resolving power compared to conventional holographic gratings. According to the general grating formula

$$R = n\,O \tag{13.4}$$

where R is the resolution, n the number of grating grooves that are illuminated, and O is the grating order. The resolution of conventional holographic gratings, which are generally used in the first order, only depends on the ruling density and the grating size. Echelle gratings, in contrast, exhibit maximum intensity for grating orders around 100. This means that, although their ruling density is typically an order of magnitude lower than that of holographic gratings, their

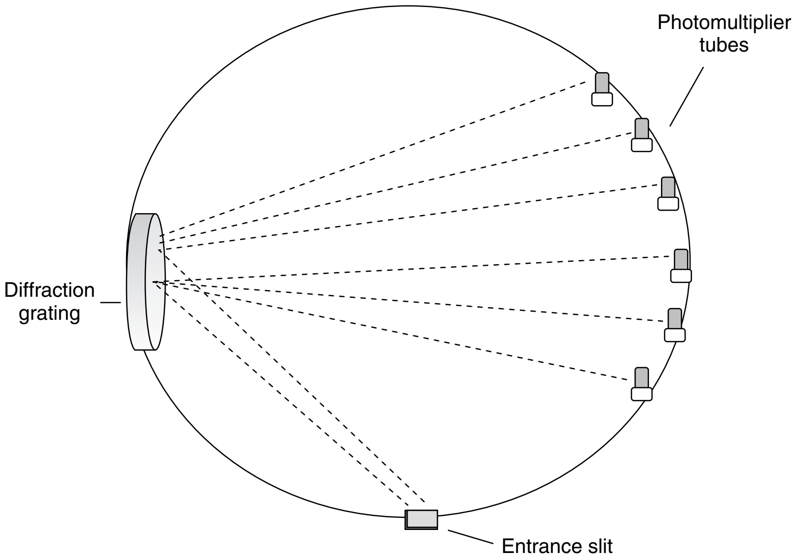

Fig. 13.8 Schematic design of a polychromator in Paschen–Runge mounting, assembled in a Rowland circle with concave grating.

resolution is still an order of magnitude higher. In this case, a second dispersive system, either a diffraction grating or a prism is used and positioned perpendicularly to the echelle grating. The echelle grating separates the polychromatic radiation by wavelength and overlapping spectral orders that are later separated or *cross dispersed* by the second optical component, generating a two-dimensional pattern. Detection of the bi-dimensional spectrum is usually achieved by a solid-state detector, such as a charge-coupled device (CCD) or charge-injection device (CID). A schematic representation of an echelle spectrometer is shown in Figure 13.9.

13.3.2.3 **Detectors**

The intensity of the radiation emitted by excited atoms or ions and separated by the wavelength-dispersing device is measured using a detector. For about half a century, photographic plates and films were used for that purpose in OES; they were increasingly replaced by PMTs in the second half of the twentieth century. In a PMT, an anode, a radiation-sensitive electrode (known as the *photocathode*) and a series of emission cathodes (known as *dynodes*), typically 10–12, are assembled and held at an increasingly positive potential. Radiation that passes the exit slit reaches the photocathode, which will emit electrons upon the impact of photons. The emitted electrons will be accelerated to the first dynode, which will release a greater number of secondary electrons that will be accelerated to the next dynode, where the process is repeated, resulting in a cascade effect. The multiplication of electrons along a PMT depends on the amplitude of the high voltage applied. At the end of the process, an electrical current is generated related to the number of photons that impacted the photocathode.

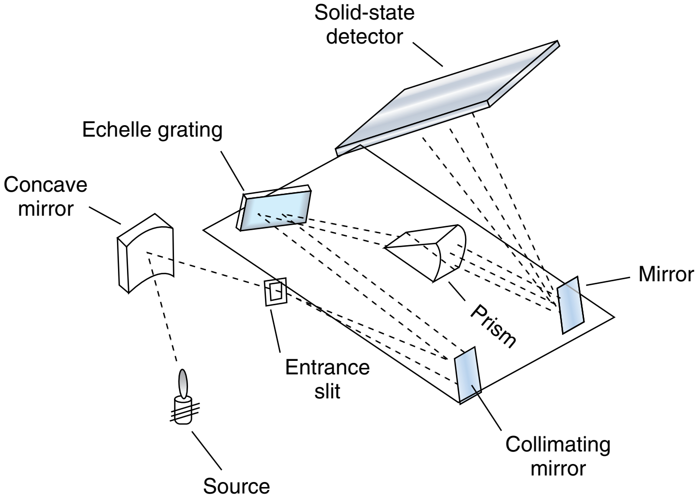

Fig. 13.9 Schematic design of an echelle spectrometer.

Modern echelle grating polychromators are equipped with solid-state detectors, such as CCDs or CIDs, which are based on the light-sensitive properties of solid silicon. In a CCD, the impact of photons generates electrons that are collected and accumulated by positively biased capacitors. The charge packets are transferred in discrete time increments by the controlled movement of potential wells to preamplifier and to the readout. In a linear CCD, the charge is moved in a stepwise fashion from element to element and is detected at the end of the line. A two-dimensional array CCD consists of an assembly of interconnected linear CCDs. The on-chip summing of charges in adjacent pixels along rows or columns is called *binning* (Laqua *et al.*, 1995). In the CID, photons striking the silicon wafer generate positive charges, and these charges are stored below negatively biased capacitor plates. The amount of stored charge is proportional to the intensity of the radiation and to the integration time. When a pixel is selected for readout, the row capacitor associated to it is maintained at a ground potential (zero volts), resulting in the injection of charge into the substrate and in an output signal (Ingle and Crouch, 1988).

13.3.3
Atomic Absorption Spectrometry

After the effectiveness of AAS for atomic spectrometric investigations had been demonstrated by Kirchhoff and Bunsen, the technique was used by other researchers as a qualitative method. Wood (1905) performed the definitive experiments that elucidated atomic absorption by resonance in gases. The absorption of the radiation from a mercury discharge lamp by mercury vapor, which appears as shadow on a screen, became a common experiment in physics lectures in colleges. Woodson (1939) applied the procedure to a quantitative method for the determination of mercury – the first application of atomic absorption to quantitative measurement. Besides these isolated examples, however, OES was preferred as an analytical technique during the first half of the twentieth century. This was to a great extent due to the fact that continuum radiation sources, as they have been used in the nineteenth century, were not really suited for AAS measurements. Particularly with a photographic plate as a detector, it was much easier to detect and measure a small emission, that is, a small increase in the opacity of the photographic layer than a small reduction of a strong emission over a narrow spectral range of a few picometers.

A turning point came in 1955 when Walsh reintroduced AAS to chemical analysis. He recognized the importance of a high-intensity, narrow linewidth source for high sensitivity and suggested the use of hollow cathode lamps (HCL). This way the high resolution that is required for atomic absorption measurements was provided by the line radiation source; in this case, a monochromator of medium resolution could be used, which only had to separate the analytical line from other lines emitted by the source. Walsh also proposed to modulate the HCL radiation and to use an amplifier tuned to the same modulation frequency (Russell, Shelton and Walsh, 1957), so that measurements could be made without interference from flame emission. This concept made AAS highly specific and selective for the element to be determined, but at the same time it limited the technique to the determination of one element at a time. The use of continuum radiation sources for AAS measurement could only be realized toward the end of the twentieth century due to the significant progress in the technology of components and in the design of spectrometers (Harnly, 1999; Heitmann *et al.*, 1996), as is discussed in Section 13.3.3.3.

An atomic absorption spectrometer consists of an atomizer with a sample introduction system and a spectrometer. The latter consists of a radiation source, a monochromator, and a detector, and it usually also contains a device for background correction. As there are fundamental differences between spectrometers for line source and continuum source AAS, these two spectrometer types are treated in separate sections.

13.3.3.1 **Atomizers**

The role of an atomizer in AAS, similar to the excitation source in OES (Section 13.3.2.1), is to transfer the sample into the gaseous state and to dissociate molecular species into ground-state atoms. Ionization of the analyte atoms has to be avoided, as ionic species have a different absorption spectrum and are therefore not available for atomic absorption measurement.

Flame Atomizers Walsh (1955) made use of a premixed, laminar-flow *air–acetylene flame*, similar to that of a Lundegårdh burner to produce atoms in the ground state. This flame, which has a temperature between 2200 and 2300 °C, and can be operated under reducing (fuel-rich) as well as oxidizing (fuel-lean) conditions, allowed the sensitive determination of some 35 elements. It still is the most popular flame used in AAS; however, it is not suitable for the determination of refractory elements and those with high affinity to oxygen. The introduction of the hot (~2700 °C) and reducing *nitrous oxide–acetylene flame*

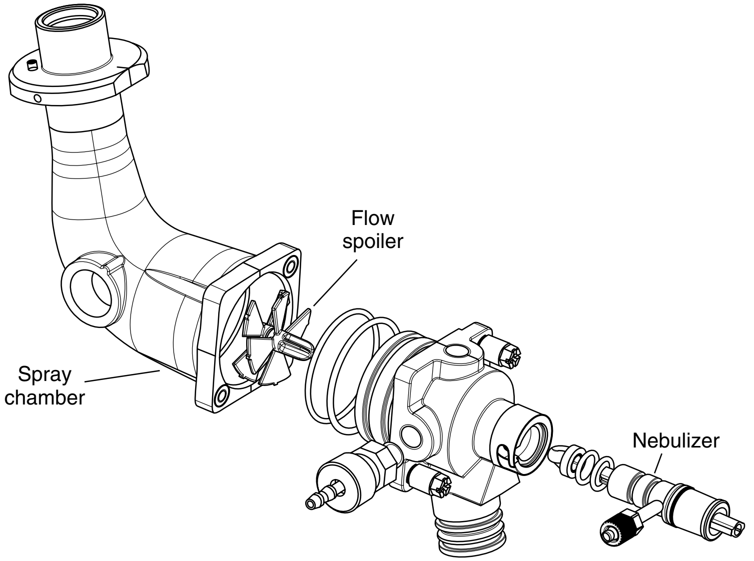

Fig. 13.10 Schematic design of a premix burner for FAAS with concrntric nebulizer, mixing chamber and impact devices. (With kind permission of Analytik Jena AG, Jena, Germany.)

(Willis, 1965) brought the solution for this limitation and added another 30 elements to the list of those that could be determined by AAS. Other gas combinations, such as the *air–hydrogen flame*, found only limited interest for special applications. Flames with oxygen as the oxidant cannot be used in a premix burner due to their high burning velocity; total consumption burners that have reached some popularity in OES have been found unsuitable for AAS (Welz and Sperling, 1999). Next to the temperature it is mostly the chemical environment, that is, the components of the flame gases, that plays a decisive role in the atomization process.

Samples are usually aspirated by a concentric *pneumatic nebulizer*, using the oxidant as the nebulizer gas; the uptake rate of these nebulizers is typically 5–10 ml min^{-1}. The generated aerosol is introduced into a mixing chamber, where it is mixed with the fuel gas and additional oxidant (Figure 13.10). The mixing chamber usually also contains impact devices, such as impact beads and flow spoilers, for postnebulization and separation of big droplets. The ''conditioned'' aerosol that finally reaches the flame typically has a maximum droplet size of <10 μm; only about 5% of the aspirated sample finally reaches the flame. Nevertheless, the amount of sample that is introduced into a flame in AAS is about an order of magnitude greater than that in ICP OES (refer to Section 13.3.2.1, Plasma Sources).

In the flame, the aerosol is submitted to several successive processes, including *desolvation* (evaporation of the solvent), *solute volatilization* (transfer of the dry particles into the gas phase), and *dissociation* of the molecules into atoms (refer to Section 13.3.2.1). The completeness of these three processes is decisive for the absence of nonspectral interferences; however, as the process of dissociation is an equilibrium process, it is sufficient that the same degree of dissociation is reached for the analyte in the sample and in the calibration solution to obtain freedom from interference. Because of its simplicity and freedom from spectral interferences, flame AAS was, to an extent, replacing flame emission spectrometry in the market, except for the flame photometers that were used mostly in clinical laboratories.

The major sources of interferences in flames are transport interferences due to differences in viscosity, surface tension, and so on, between the solution, which result in different aspiration rates and nebulization properties, as well as dissociation and ionization interferences. The former ones can be controlled using the analyte addition technique for calibration, whereas the latter ones can be removed by adding appropriate buffers or releasing reagents (Welz and Sperling, 1999).

Electrothermal Atomizers The term *electrothermal atomization* (*ETA*) refers to an atomizer that is made of a material of relatively high specific resistance, which is heated by application of a high current. Several ET atomizers have been described, differing mainly in the design and in the material from which the atomizer was made. Nowadays, essentially all of the commercially available ET atomizers are made of graphite, although other devices, such as tungsten coils, have also been used successfully (Berndt and Schaldach, 1988; Silva *et al.*, 1996). For simplicity reasons, graphite tube atomizers are discussed exclusively in this section. The GF atomizer has originally been proposed by L'vov (1961) and later been simplified by Massmann (1968); the first commercial atomizer of this type was introduced in 1970. The main advantages of this atomizer are that the entire sample portion that is introduced into the graphite tube is atomized typically within 1–2 seconds and the analyte atoms remain in the absorption volume much longer, resulting in an increase in sensitivity of 2–3 orders of magnitude, compared to flame atomizers (Welz and Sperling, 1999). Another important feature of this atomizer is that not only liquid and dissolved samples can be analyzed but also gaseous and solid samples directly (Welz *et al.* 2007b), resulting in further increase in sensitivity and greatly simplified sample preparation.

A GF atomizer consists of a cylindrical graphite tube, typically 20–25 mm long with a diameter of about 5 mm. A small orifice is usually present, through which the (liquid) sample is introduced. The tube is positioned between two cooled graphite electrodes, and acts as a resistance, reaching temperatures as high as 2700 °C. An inert gas (argon) constantly flows around the tube, protecting it from contact with the outside air and, consequently, from burning at high temperatures. An additional internal gas flow is used to eliminate the solvent and matrix vapors during thermal pretreatment of the sample. A *temperature program* is usually applied, consisting of four essential stages: *drying*, where a sufficiently high temperature is applied to eliminate the solvent; *pyrolysis*, where the intention is to eliminate the matrix components without vaporizing the analyte; *atomization*, where the temperature is increased rapidly in order to promote the formation of the maximum concentration of analyte atoms in the absorption volume. The internal gas flow is normally interrupted in this stage to increase the residence time of analyte atoms in the absorption volume. The last stage is a *cleaning* stage, where all matrix and analyte residues are eliminated prior to the introduction of the next sample.

While the temperature, heating ramp, and hold time of all stages of the temperature program require optimization, the two most critical parameters are pyrolysis and atomization. Firstly, a *pyrolysis curve* is established plotting the integrated absorbance signals obtained at a set atomization temperature against the pyrolysis temperature, which is increased in increments of typically 100 °C. This curve usually exhibits a plateau up to a maximum temperature above which the signal begins to drop, indicating analyte losses in the pyrolysis stage as shown in Figure 13.11(a). Secondly, using the maximum pyrolysis temperature at which no analyte losses occur, the atomization temperature is increased gradually, and at least for the more volatile elements a maximum is obtained, as shown in Figure 13.11(b).

During the first decade after its introduction, GF AAS was considered a technique that is subject to numerous difficult-to-control interferences, which was mostly due to the fact that it was largely unknown how to use the technique properly. Slavin, Manning and Carnrick (1981), based on research of L'vov (1978), introduced the stabilized temperature platform furnace (STPF) concept, which was the key to essentially interference-free GF AAS analysis. The most important components of the concept were (i) atomization of the sample from a graphite *platform* inserted into the graphite (instead of from the tube wall) in order to delay atomization until the tube has reached its final temperature; (ii) integration over the *peak area* of the transient signal (instead of using peak height) in order to compensate for matrix influences on the kinetics of atom release; (iii) use of *chemical modifiers* in order to avoid low-temperature analyte losses due to matrix components. The last topic has been treated in a large number of publications (Tsalev, Slaveykova and Mandjukov, 1990); a mixture of Pd and Mg nitrates, co-injected with the sample, was proposed as a *universal chemical modifier* (Schlemmer and Welz, 1986), and its use has been demonstrated for over 20 elements (Welz, Schlemmer and Mudakavi, 1992). In the meantime, *permanent chemical modifiers* are investigated increasingly, where noble metals, such as Pd, Ir, Ru, or Rh, and carbide-forming elements, such as W, Zr, Th, and Ta, or mixtures of both, are used as a kind of surface coating of the graphite platform to stabilize the analyte for a large number of consecutive determinations.

One aspect that was not included in the original STPF concept, but which turned out later to be another very decisive parameter, was the spatial isothermality of the graphite atomizer. In the original concept of Massmann, which has been adopted by all GF manufacturers in the following years, the graphite tube was contacted at its ends and heated longitudinally. As the contacts were cooled, this resulted in a significant temperature gradient along the graphite tube, which was the reason for condensation of matrix and analyte residues on the cool ends, causing severe background absorption and memory effects. *Transversally heated atomizers* solved this problem and should

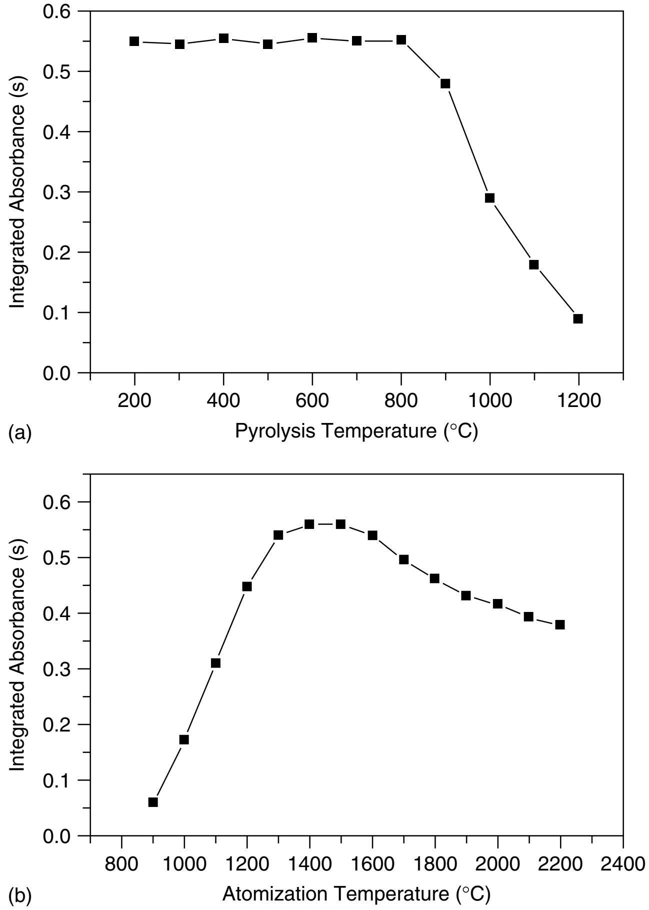

Fig. 13.11 Typical pyrolysis and atomization curves for GF AAS; (a) pyrolysis curve and (b) atomization curve; for details see text.

be preferred over longitudinally heated ones. The schematic design of such an atomizer tube is shown in Figure 13.12.

Atomizers for Chemical Vapor Generation

Chemical vapor generation (*CVG*) is actually a general term for several sample preparation and introduction techniques, where the analyte is converted into a gaseous compound using a chemical reaction. This gaseous compound may then be introduced into different atomizers for atomization and measurement of the absorbance. The two most important techniques are the cold vapor (CV) technique for the determination of mercury and the hydride generation (HG) technique for the determination of elements that form volatile hydrides, mainly As, Bi, Sb, Se, Sn, and Te. In the former technique, mercury (II) is reduced to the element using stannous chloride according to

$$Hg^{++} + Sn^{++} \rightarrow Hg^{0} + Sn^{4+} \quad (13.5)$$

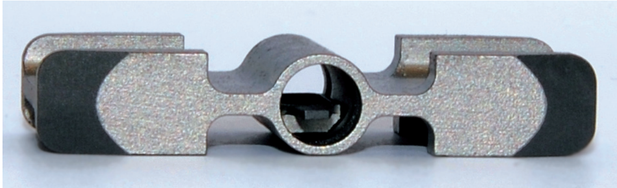

Fig. 13.12 Transversally heated graphite tube for GF AAS. (With kind permission of Analytik Jena AG, Jena, Germany.)

Other reducing agents, such as sodium tetrahydroborate (THB), may be used as well. As mercury is reduced to the element in this reaction, and as mercury has a significant vapor pressure already at room temperature, cold vapor atomic absorption spectrometry (CV AAS) does not need an atomizer. The mercury vapor can simply be transferred to an absorption cell using a flow of gas, where the absorbance can be measured, a procedure that has been described by Hatch and Ott (1968). Alternately, it can be collected by amalgamation on a precious metal, such as gold gauze, and liberated by heating in order to increase the sensitivity. Another way of preconcentrating mercury is in a graphite tube treated with a permanent modifier and atomization using GF AAS. This preconcentration also eliminates interferences due to different release kinetics of the vapor from the solution.

The formation of volatile hydrides of arsenic and a few other elements upon reaction with zinc in acid solution is known as the *Marsh and Gutzeit tests* since the nineteenth century. Holak (1969) proposed this technique for the determination of As by AAS. A major breakthrough came with the introduction of sodium THB as the reducing agent (Braman, Justen and Forebank, 1972), which increased the number of elements that could be determined to eight. The mechanism of hydride formation was, for a long time, believed to be due to "nascent hydrogen," a mechanism that has now been ruled out (D'Ulivo *et al.*, 2004a; D'Ulivo, Onor and Pitzalis, 2004b) In fact, it is due to direct interaction between the analyte and THB, resulting in hydroborane intermediates that will later decompose to generate the volatile species. Another recent advance in this technique is that the number of elements that form volatile compounds with THB has increased to more than 15 (Guo *et al.*, 2004; Sturgeon, Guo and Mester, 2005), although not all of these reactions might be of practical analytical importance.

Initially, the generated hydrides were introduced into a flame for atomization, which, however, did not result in sufficiently high sensitivity. For about two decades, the heated quartz tube atomizer (QTA), an electrically or flame-heated quartz tube has been the preferred atomizer for hydride generation atomic absorption spectrometry (HG AAS), mostly because of its high sensitivity. It took some time until the mechanism of atomization in the QTA was understood, as thermal dissociation could be excluded as the mechanism, because atoms of As, Se, and so on, could not exist at temperatures $<1000\,^{\circ}\mathrm{C}$ thermodynamically. Later, it has been shown that atomization occurs in a cloud of hydrogen radicals that is formed in the entrance part of the T-tube in a reaction between hydrogen that is generated in the reaction of the acid sample solution with THB and traces of oxygen (Dědina, 1988).

The QTA has been increasingly replaced by the GF atomizer over the past decade, mostly because preconcentration in a graphite tube, treated with a permanent modifier, increases the sensitivity and eliminates kinetic interferences due to the release of the hydride from the solution. The major advantage of HG AAS is the separation of the analyte from the matrix, which largely eliminates interferences in the atomization stage. The major problems are interferences in the liquid phase that impede the formation or liberation of the gaseous compound from the solution. These interferences are mostly caused by transition elements of Group VIII and Group I, which are reduced by THB, for example, to the finely dispersed metal that adsorbs and decomposes the hydrides (Welz and Melcher, 1984(a, b, c); Welz and Schubert-Jacobs, 1986). Another problem of CVG is that the analyte has to be present in ionic form and in a well-defined oxidation state, as, for example, only tetravalent selenium and tellurium form a hydride, and in the case of As and Sb the trivalent form reacts faster and gives better sensitivity. This means that CVG generally requires complete sample digestion and a prereduction step in order to give reliable results.

13.3.3.2 Spectrometers for Line Source AAS

In line source atomic absorption spectroscopy (LS AAS), the high resolution required to measure atomic absorption is provided by the radiation source that emits radiation only within the narrow spectral interval that can be absorbed by the analyte atoms. As there is practically only the spectrum of the analyte element emitted by the source, the only task of the monochromator is to isolate the analytical line from other lines emitted by the source.

Radiation Sources for LS AAS The line source used most frequently in AAS is the *hollow cathode lamp* (HCL), a lamp that has been described as an emission source as early as 1916 (Paschen, 1916). However, in order to become applicable for routine use in AAS, this lamp had to be modified significantly (Russell and Walsh, 1959, Jones and Walsh, 1960), and was optimized further in the following years (Manning and Vollmer, 1967). The schematic design of an HCL is shown in Figure 13.13. An HCL is a spectral lamp with a hollow, cylindrical cathode made of, or containing the element of interest; the anode is mostly made of tungsten or nickel. The glass cylinder of the lamp is filled with an inert gas, usually neon or argon, at a pressure of about 1 kPa; for elements that have their main analytical line in the UV range, the front window is mostly made of quartz. HCL often also contain some insulation in order to concentrate the discharge to the inside of the cathode. When a voltage of 100–200 V is applied across the electrodes, a GD takes place in the reduced-pressure gas atmosphere, as described for the GD excitation source (refer to Section 13.3.2.1, Glow Discharges). As a result, the emission spectrum of the element of interest is obtained, and the radiation beam can be directed toward the absorption volume

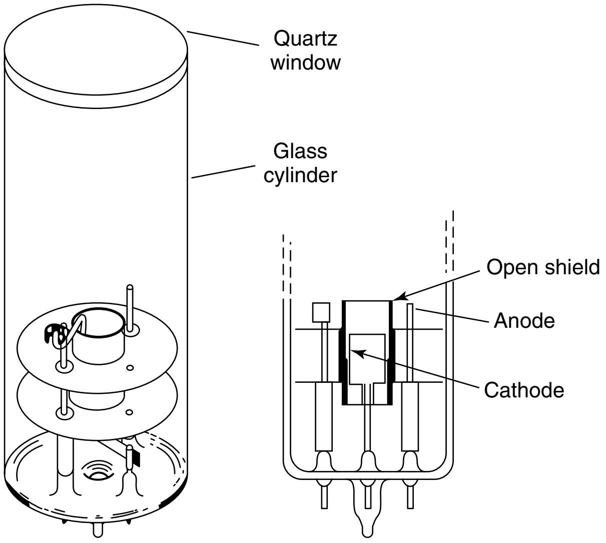

Fig. 13.13 Schematic design of a hollow cathode lamp for AAS. (From Welz and Sperling (1999), p. 107; reproduced with permission from Wiley-VCH.)

in the atomizer, where the absorption process takes place.

For some of the volatile elements, such as Hg, and those that have their main lines in the vacuum-UV range, particularly As and Se, the quality of HCL and their emission intensity are not always satisfactory. In these cases, *electrodeless discharge lamps* (EDLs) might be used with advantage. An EDL consists of a sealed quartz bulb filled with the element of interest (or a compound of it, such as an iodide) and an inert gas. The bulb is placed inside a metallic coil, to which an RF power is applied, generating an ICP discharge. This promotes volatilization of the element inside the quartz bulb; further dissociation and excitation will be induced by the plasma, resulting in the emission of the wavelengths characteristic for the element. The emission intensity of an EDL is typically 20–50 times higher than that of an HCL, resulting in better precision and lower LOD. The disadvantages of this lamp type are the need of an additional power supply and the long warmup times until they reach stable emission.

Other radiation sources, such as diode lasers (Zybin *et al.*, 2005; Hergenröder and Niemax, 1988), have also been investigated with some success, but their use remains restricted to research instruments.

Monochromators for LS AAS The role of the monochromator in LS AAS is to isolate the analytical line from all other lines emitted by the radiation source. Medium-resolution grating monochromators in Czerny–Turner (refer to Section 13.3.2.2) or the simpler Littrow mounting are used most frequently. A particularly important aspect is the *spectral bandwidth,* $\Delta\lambda$, which corresponds to the radiation interval that effectively reaches the detector. The spectral bandwidth is determined by the width of the entrance slit, the *geometric slit width*, s, which controls the amount of radiation that enters the monochromator, and by a characteristic of the dispersing device, the *reciprocal linear dispersion*, $d\lambda/dx$, (dimension: nanometers per millimeter) according to

$$\Delta\lambda = s\frac{d\lambda}{dx} \tag{13.6}$$

The reciprocal linear dispersion depends on the ruling density and the ruled area of the grating; the smaller the numerical value of the reciprocal linear dispersion, the wider can be the geometric slit width in order to obtain a given spectral bandwidth. A wide geometric slit width means that a great amount of radiation can enter the monochromator, resulting in a good S/N ratio. The spectral bandwidth is determined by the spectral environment of the analytical line; it should be chosen as wide as possible, and as narrow as necessary to exclude other lines emitted by the radiation source. Typical values to fulfill this requirement are between 0.2 and 2 nm.

Detectors for LS AAS The role of detectors in AAS is to proportionally convert optical radiation (photons) into an electrical signal; PMT are mostly used for that purpose, although solid-state detectors provide better S/N ratios (for details refer to Section 13.3.2.3).

Spectral Interferences and Background Correction in LS AAS Spectral interferences due to direct or partial *line overlap* are relatively scarce in AAS due to the much smaller number of absorption lines compared to the number of emission lines, particularly in high-temperature excitation sources. However, there are a few well-known examples, such as the overlap of the 213.859-nm secondary iron line with the main analytical line for zinc at 213.857 nm. There is essentially no way to correct for this kind of interference in LS AAS, which means that the determination of Zn in the presence of high iron concentration is always affected by a systematic error.

Other spectral interferences encountered in AAS are *radiation scattering* at particles in the absorption volume, an interference that is mostly limited to GF AAS, and *molecular absorption* due to nondissociated matrix components. The latter has to be classified into dissociation continua, that is, molecules are dissociated by the incident radiation, and electron excitation spectra. In the latter case, we again have to distinguish between electron excitation spectra of diatomic molecules, which exhibit a pronounced rotational fine structure, and those of tri-atomic molecules, the fine structure of which cannot be resolved, so that they appear as broadband spectra.

There is no way to measure only atomic absorption in the presence of any kind of background absorption. In LS AAS, it is necessary, first to measure total absorption and then background absorption, and to calculate atomic absorption by difference in a fast sequential process.

The first system that was proposed to correct for background absorption in AAS used a *deuterium lamp* as a second radiation source in addition to the line source (Koirtyohann and Pickett, 1965). The radiation of the two sources was passed alternately through the flame, typically at a frequency of 50–60 Hz. This system assumed that background absorption is spectrally continuous, at least within the section of the spectrum that is passing the exit slit (0.2–2 nm), whereas atomic absorption occurs only within a very narrow spectral interval (a few picometers). Under this premise, essentially only the background will attenuate the continuum radiation emitted from the deuterium lamp, as atomic absorption represents a fraction of <1% of the spectral interval. The line source, in contrast, will measure total absorption, that is, atomic and background absorption. A "subtraction" of the signal obtained from the D_2 lamp from the signal of the line source will provide the net atomic absorption signal; Figure 13.14 shows a schematic representation of this system.

Background correction with a deuterium lamp is relatively simple and does not add significantly to the cost of the instrument. However, it increases the noise level, as two radiation sources are used, hence deteriorating the S/N ratio and LOD. Although rapidly changing background signals, as they typically occur in GF AAS, could cause artifacts, the system works reasonably well in the case of continuous background absorption. However, the system fails to correct for any *structured background*, such as the electron transition spectra of diatomic molecules that exhibit rotational fine structure. In this case, the deuterium lamp makes an average of the background absorption over the spectral bandwidth, which is in most cases different from the actual background at the analytical wavelength, resulting in over- or under-correction. Deuterium background correction can actually even be the source of errors when the absorption line of a concomitant element is within the spectral bandwidth without overlapping with the analytical line. If this concomitant element is present at high concentration, it can absorb a significant portion of the continuous radiation without absorbing radiation of the line source, resulting in over-correction, which appears as a negative absorbance signal. The emission intensity of deuterium lamps decreases significantly above 350 nm, so that it

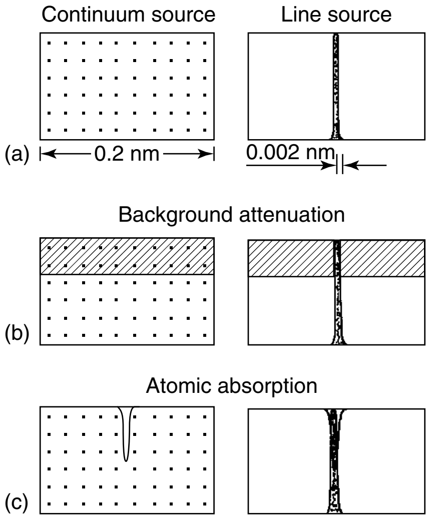

Fig. 13.14 Mode of function of deuterium lamp background correction; (a) the radiant intensity, represented schematically by dots, for the deuterium lamp is distributed over the entire width of the spectral band isolated by the slit (e.g., 0.2–2 nm), while for the line source it is limited to a few picometers; (b) broad-band background absorption attenuates the radiation emitted by both sources to equal degrees; (c) atomic absorption, which is limited to a few picometers, in the first approximation attenuates only the radiation from the line source. (From Welz and Sperling (1999), p. 122; reproduced with permission from Wiley-VCH.)

cannot be used for correction at higher wavelengths.

A more efficient, but also more complex approach for background correction makes use of the *Zeeman effect*, that is, the splitting of electronic levels of atoms (and consequently of spectral lines) under the influence of a magnetic field (De Loos-Vollebregt and de Galan, 1985). The spectral lines split into three components or groups of components: a π component of the same frequency as the original line, and two σ components (σ^+ and σ^-), with slightly higher and lower frequency, as shown in Figure 13.15. In addition to the splitting the components are also polarized, the π components are linearly polarized in a direction parallel to the magnetic field, while the σ components are circularly polarized in a direction perpendicular to the magnetic field. For *Zeeman-effect background correction* (ZBC), an alternating magnetic field is usually applied at the atomizer, and all measurements are made using the line profile emitted by the spectral lamp. The total absorbance (atomic and background) is measured with the magnetic field off, and the background signal is measured with the magnetic field on. In this case, the σ components of the analyte atoms are shifted out of the emission profile of the lamp, and the π components are usually eliminated using a polarizer. Alternately, the magnetic field can be applied parallel to the radiation beam; in this case, the π components become "invisible," so that no polarizer is necessary.

ZBC is almost exclusively used for GF AAS and can be successfully employed for most analytical applications, as it

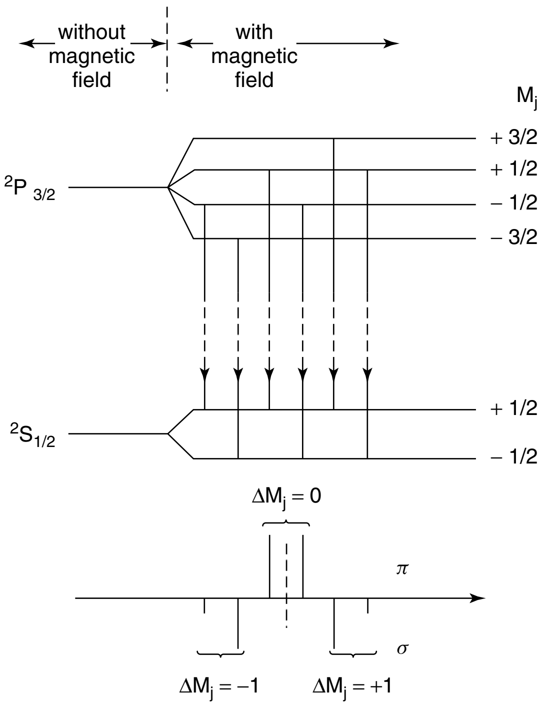

Fig. 13.15 Splitting of the energy levels in a magnetic field with the resulting splitting pattern for the electron transition $^2S_{1/2} \leftrightarrow {}^2P_{3/2}$. (From Welz and Sperling (1999), p. 97; reproduced with permission from Wiley-VCH.)

efficiently corrects for most types of background absorption, including structured background. The only notable exceptions are when the concomitant species, an atom or a diatomic molecule, is also affected by the magnetic field; in this case, the background without and with magnetic field is not the same, which inevitably results in background correction errors. In the case of a near-line coincidence with an absorption line of a concomitant element, the latter obviously also shows Zeeman splitting; in this case, the σ components of the concomitant element might overlap with the emission line even more severely than without magnetic field, causing an even more pronounced spectral interference. The same is true for the spectra of some diatomic molecules, such as the phosphorus suboxide PO (Becker-Ross, Florek and Heitmann, 2000), which also undergo changes in the magnetic field, causing background correction errors.

The sensitivity of AAS instruments with ZBC is lower when compared to other instruments, as the separation of the $\sigma^{\pm}$ components is usually not complete, which means that some atomic absorption will always be measured together with the background and "subtracted" from the total absorption, resulting in lower integrated absorbance signals. For higher analyte concentrations, this may result in the so-called *rollover effect*, which can also be noticed on transient signals that start to split and exhibit two maxima at high concentrations (Figure 13.16). More recently, instruments using a "three-field" mode have been introduced, where in addition to the "zero" field (magnet off) and the high field strength, an adjustable intermediate field can be used in order to deliberately reduce sensitivity and

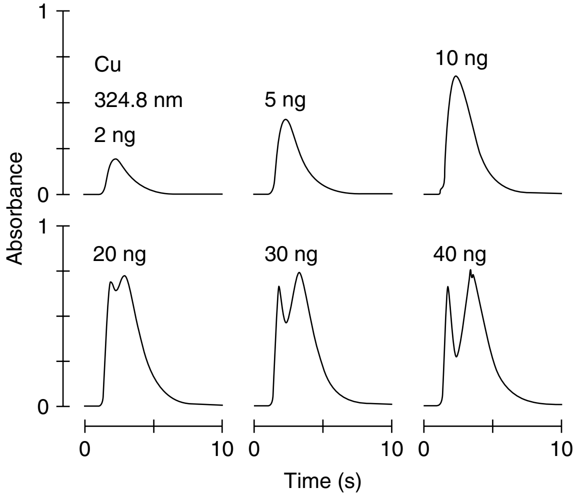

Fig. 13.16 Effect of rollover of the calibration curve on the signal shape for GF AAS; determination of increasing masses of copper in an alternating magnetic at the atomizer. (From Welz and Sperling (1999), p. 131; reproduced with permission from Wiley-VCH.)

increase linearity, making possible the determination of higher concentrations of the analyte without the need for dilution (Gleisner, Eichardt and Welz, 2003).

13.3.3.3 Spectrometers for High-Resolution Continuum Source AAS

Although LS AAS is an established, reliable, and robust analytical technique for a large number of applications, it has been questioned already in the late 1980s that if AAS could survive in the face of strong competition mainly from ICP OES and ICP-MS without a renovation of the basic concept (Hieftje, 1989). The approach that has been proposed was the use of a continuum radiation source, high-resolution spectral sorting devices and completely new detection systems. Although continuum radiation sources have been investigated as an alternative to line sources since the early 1960s (Welz *et al.*, 2005), it was only some 30 years later, when finally components became available, that made it possible to design equipment for high-resolution continuum source atomic absorption spectroscopy (HR-CS AAS) with a performance that was superior to that of LS AAS (Heitmann *et al.*, 1996, Harnly, 1999). The new approach includes a xenon short-arc lamp, a high-resolution double monochromator and a linear CCD array detector, as is shown schematically in Figure 13.17.

Radiation Source for HR-CS AAS The radiation source, a specially-designed high-pressure Xe short-arc lamp with a nominal power of 300 W is operating in a hot-spot mode, as shown in Figure 13.18(a). The hot spot with a diameter of about 0.2 mm reaches a temperature of about 10 000 K, which is the precondition for providing sufficient radiation intensity over the entire spectral range from 190 to 900 nm. Figure 13.18(b) shows that the emission

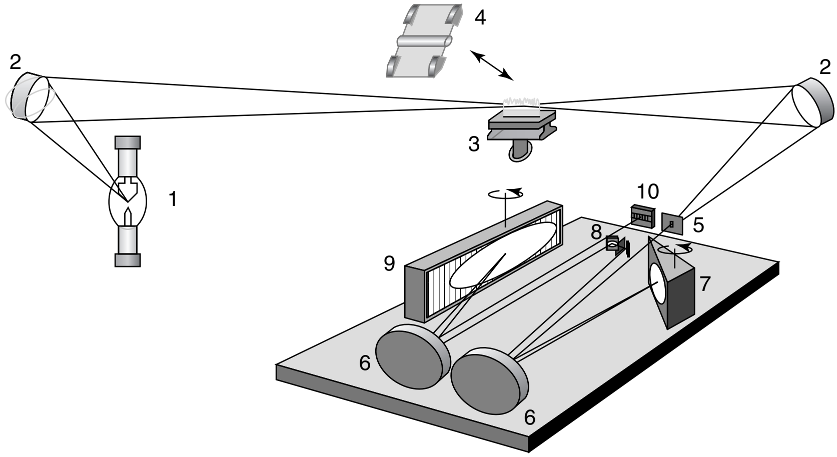

Fig. 13.17 Schematic design of an instrument for HR-CS AAS; 1 – xenon short-arc lamp; 2 – elliptical mirrors; 3 – flame atomizer; 4 – graphite tube atomizer; 5 – entrance slit; 6 – parabolic mirrors; 7 – prism; 8 – intermediate slit; 9 – echelle grating; 10 – CCD detector. (From Welz *et al.* (2003), p. 223; reproduced with permission from the Brazilian Chemical Society.)

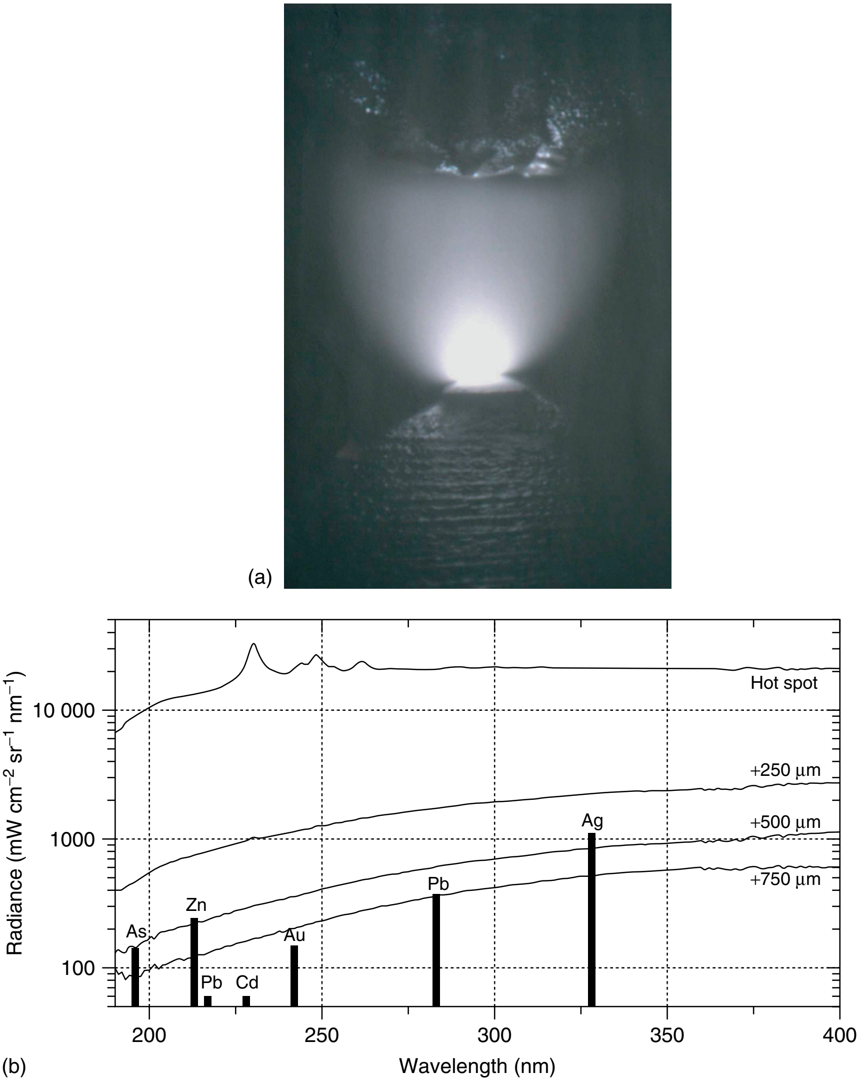

Fig. 13.18 High-pressure Xe short-arc lamp as radiation source for HR-CS AAS; (a) hot spot with a diameter of about 0.2 mm and a temperature of ~10 000 K; (b) wavelength-dependent spectral radiance in the hot spot in comparison to some selected emission lines from HCL. (From Welz *et al.* (2005), pp. 32–33; reproduced with permission from Wiley-VCH.)

intensity of this lamp in the UV range is typically 2 orders of magnitude higher than that of selected HCL and conventional continuum radiation sources. This high radiation intensity has a direct influence on the S/N ratio, and hence on the LOD, as in AAS the noise is inversely proportional to the square root of the radiation intensity. As the position of the hot spot is not stable over time, the radiation emitted by the Xe short-arc lamp is directed toward the atomizer by a set of mirrors, and it is actively stabilized by computer software to assure that the radiation beam is in a stable (central) position in the absorption volume.

Monochromator for HR-CS AAS The high-resolution double monochromator consists of a predispersing prism monochromator and a high-resolving echelle grating monochromator, both in Littrow mounting. The radiation enters the monochromator, is reflected by a parabolic mirror onto the Littrow prism, which has a mirror on its other side, so that the radiation passes the prism twice. The prism is rotated in a way that only that part of the spectrum that contains the analytical line will be reflected to pass the intermediate slit and enter the second monochromator. A second parabolic mirror directs the radiation toward the echelle grating, which provides the high resolution of the small spectrum interval selected by the intermediate slit, which is then reflected onto the detector without passing an exit slit. This double monochromator provides a resolution of $\lambda/\Delta\lambda \approx 140\,000$, corresponding to a pixel resolution better than 2 pm at 200 nm. This resolution is some 2 orders of magnitude higher than that of a conventional LS AAS monochromator, and also better than that of typical ICP OES mono- and polychromators.

Detector and Measurement Principle in HR-CS AAS After the radiation has been highly resolved by the echelle grating, it is directed to the detector, which is typically a UV-sensitive linear CCD array with 512 or 588 pixels, 200 of which are used for analytical purposes. All pixels are illuminated for typically 1–10 milliseconds and read out simultaneously, and as each pixel is equipped with an individual amplifier, the system actually operates with 200 individual detectors. Each pixel detects a small portion of the highly resolved spectrum around the analytical line, as shown in Figure 13.19 for the primary absorption line of cadmium. The center pixel measures the absorbance at the line core and the neighbor pixels measure the absorbance at the line wings. This means that in HR-CS AAS the measured absorbance becomes a function of the pixel width, whereas in LS AAS it is a function of the width (profile) of the line emitted by the radiation source. In contrast to LS AAS, however, there is the possibility in HR-CS AAS to add the absorbancies measured at several pixels and arrive at the *wavelength selected absorbance* (WSA), $A_{\Sigma n}$, for steady-state signals, and the *peak volume selected absorbance* (PVSA), $A_{\Sigma n,\mathrm{int}}$, for transient signals, where n is in both cases the number of pixels used for measurement (Heitmann *et al.* 2007). The number of pixels that are finally used for measurement has to be optimized, as each pixel also has a contribution to the noise. For narrow absorption lines, the optimum S/N ratio is usually obtained with three pixels, whereas for wider lines the sum of up to seven pixels might be used.

The availability of 200 independent and simultaneously operating detectors adds a third dimension, the spectral resolution, to the traditional absorbance-over-time measurement obtained by LS AAS, which means that three signal evaluation modes are available: first, the traditional absorbance over time can be visualized for each individual pixel of the detector; secondly, the absorbance as a function of the wavelength can also

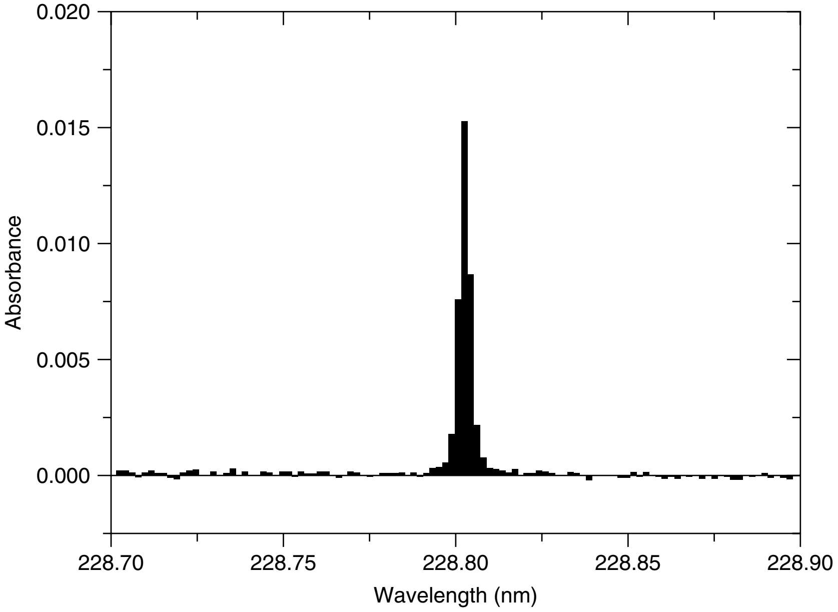

Fig. 13.19 Absorption spectrum in the vicinity of the Cd resonance line at 228.802 nm recorded simultaneously with 200 pixels of a CCD array detector. (From Welz *et al.* (2003), p. 223; reproduced with permission from the Brazilian Chemical Society.)

be visualized, showing all the spectral events in the vicinity of the analytical line; and thirdly, a time-resolved three-dimensional spectrum can be obtained, showing the entire highly resolved spectrum around the analytical line, as shown in Figure 13.20. With such information, any spectral event occurring in addition to analyte absorption, such as absorption by concomitant elements or structured molecular absorption can be detected, and appropriate measures can be taken in order to avoid or correct for spectral interferences (Welz *et al.*, 2005).

Background and Baseline Correction In HR-CS AAS, no additional instrumental devices, such as secondary radiation sources or magnetic fields, are necessary for background correction; any kind of background correction, and also the correction for baseline noise due to lamp flicker noise, and so on, is carried out with detector pixels that are not used to measure analyte absorption. This means that background measurement, background correction, and baseline correction are strictly simultaneous with the measurement of atomic absorption, which is in contrast to LS AAS, where these measurements and corrections are sequential.

Initially, we have to distinguish between continuous (within the spectral section that reaches the detector) and discontinuous absorption (refer to Section 13.3.3.2, Spectral Interferences and Background Correction in LS AAS). Any spectrally continuous increase or decrease of the radiant flux over time is measured identically at all pixels of the detector, and can be eliminated

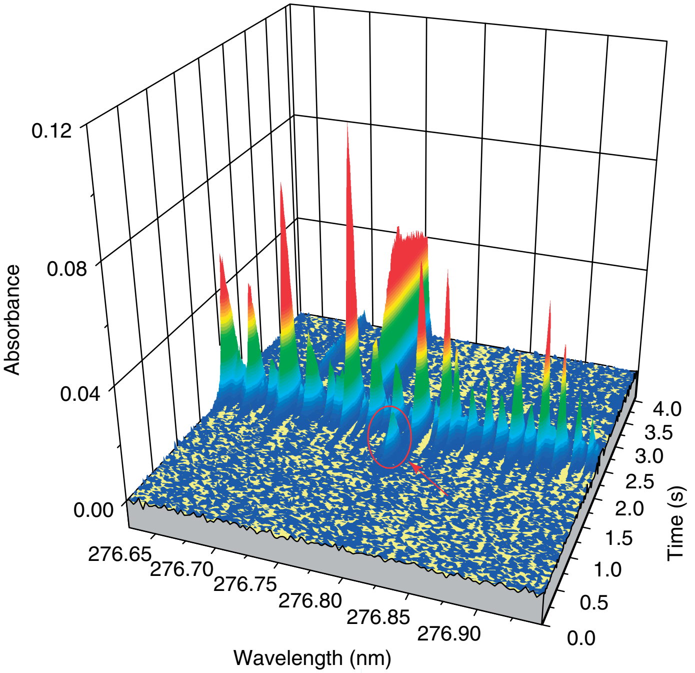

Fig. 13.20 Time- and wavelength-resolved absorption spectrum for a marine sediment recorded in the vicinity of the Tl absorption line at 276.787 nm using HR-CS GF AAS; the Tl absorption, which is marked by a circle, is temporally and spectrally resolved from the fine-structured molecular absorption and from an iron absorption line that appears only late in the atomization cycle.

using correction pixels. This includes any fluctuation in the lamp intensity (arcs are notoriously instable), but it also includes any continuous background, such as radiation scattering and molecular absorption without fine structure, as the entire radiation is passing through the atomizer. Any events affecting all pixels at the same time will be readily identified and corrected automatically, resulting in a smooth baseline and an absorption signal with an extremely low noise level, as shown in Figure 13.21 for the determination of Pb using high-resolution continuum source graphite furnace atomic absorption spectrometry (HR-CS GF AAS).

After correction for any continuous change in the radiant flux remains the discontinuous absorption of radiation. Analyte absorption is obviously discontinuous, as it occurs within a few picometers, and 3–5 pixels are typically sufficient for its measurement. If more than one absorption line of the same or of different elements is within the wavelength interval that reaches the detector, this absorption is measured as well and appears in the wavelength-resolved image. No measure is necessary if there is no spectral overlap between the analyte line and a concomitant line, as at high resolution only the signals in the center of the analytical line are

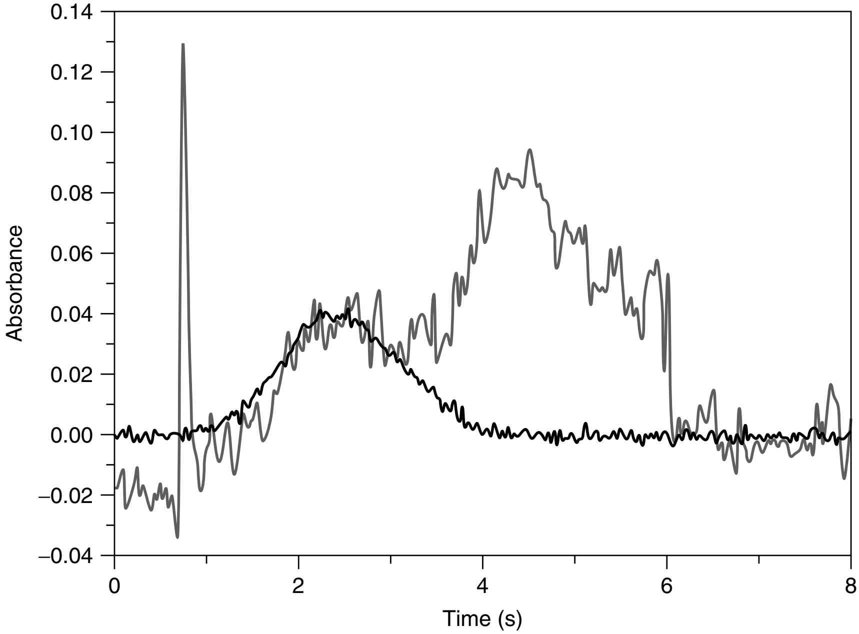

Fig. 13.21 Absorbance-over-time signal recorded for Bovine Muscle certified reference material at the Pb resonance line at 217.001 nm using one pixel only; *gray line*: without background correction; *black line*: after automatic correction for all spectrally continuous fluctuations of the intensity. (From Borges *et al.* (2006).)

considered. In the case of GF AAS there is an additional possibility, even in the case of direct line coincidence, to separate the absorbance in time, in case the analytes have significantly different volatility. Other sources of discontinuous absorption are electron excitation spectra of diatomic molecules, such as NO, PO, CS, and SiO, which exhibit pronounced rotation fine structure. These spectra can be easily recognized due to visualization of the spectral region in the vicinity of the analytical line. In this case, the same rules apply as in the case of a concomitant absorption line, i.e., no measure is necessary if there is no spectral or temporal overlap between the analyte line and the molecular absorption "lines."

In the case that analyte and concomitant atomic or molecular absorption overlap spectrally and cannot be separated in time, the software allows the user to correct for the interference using a least-squares algorithm. Figure 13.22 shows the well-known coincidence of the primary resonance line for Zn at 213.857 nm with a secondary Fe line at 213.859 nm, which cannot be separated spectrally, as the line width is greater than their distance. This interference cannot be corrected at all in classical LS AAS. Fortunately, Fe has a second absorption line at 213.970 nm within the spectral interval that reaches the detector, which can be used for correction. As the absorbance ratio between the two lines is constant and known, it is sufficient to measure the absorbance at the 213.970-nm line, calculate the absorbance at the 213.859-nm Fe line, and subtract this

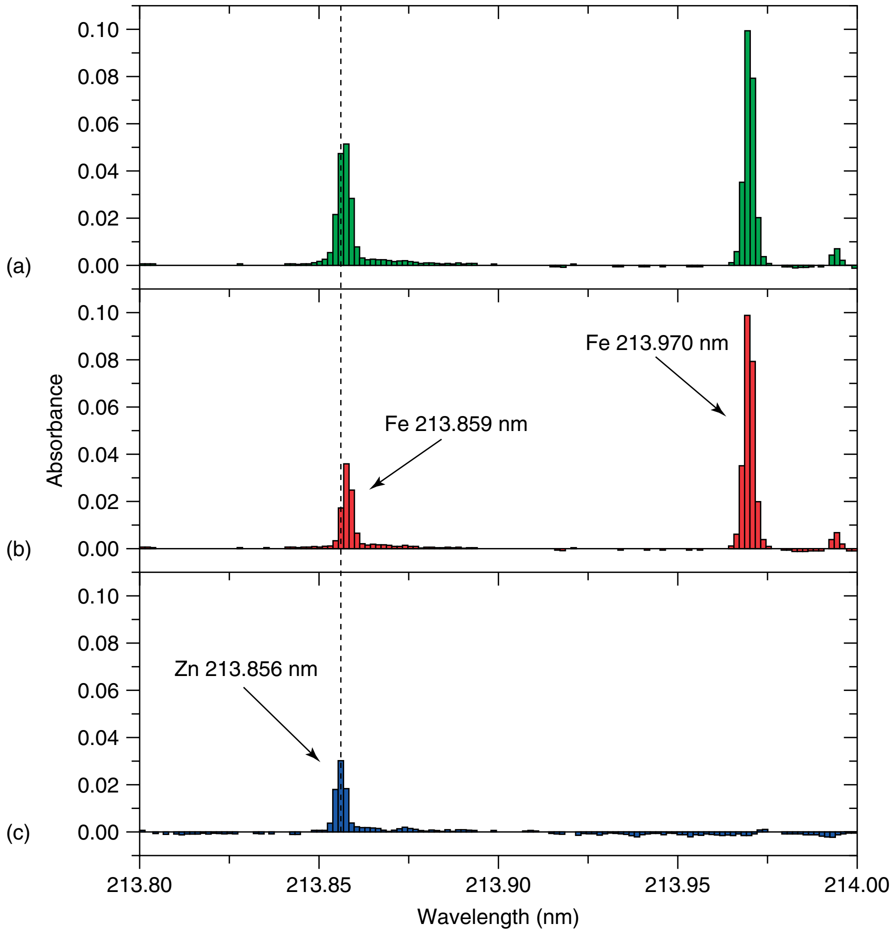

Fig. 13.22 Least-squares correction of the spectral interference due to direct-line overlap between the Zn resonance line at 213.856 nm and the secondary Fe absorption line at 213.859 nm: determination of 0.3 mg L^{-1} Zn in the presence of 1000 mg L^{-1} Fe; (a) integrated absorbance spectrum for Zn in iron matrix without correction; (b) spectrum for pure iron used for correction; (c) resulting spectrum after correction. (From Heitmann, U., ISAS – Institute for Analytical Sciences, Department of Interface Spectroscopy, Berlin, Germany, reproduced with permission.)

value from the total absorbance to get the net absorbance due to Zn.

The same algorithm can be used to correct for molecular spectra with rotational fine structure. In this case, a synthetic solution is prepared with a compound that generates the desired reference spectrum in the "atomization" cycle (such as $NH_4H_2PO_4$ to generate PO). This reference spectrum is stored in the software and "subtracted" from the sample spectrum (Welz *et al.*, 2005). This process can be repeated in case there is more than one diatomic molecule or concomitant atom causing spectral interference.

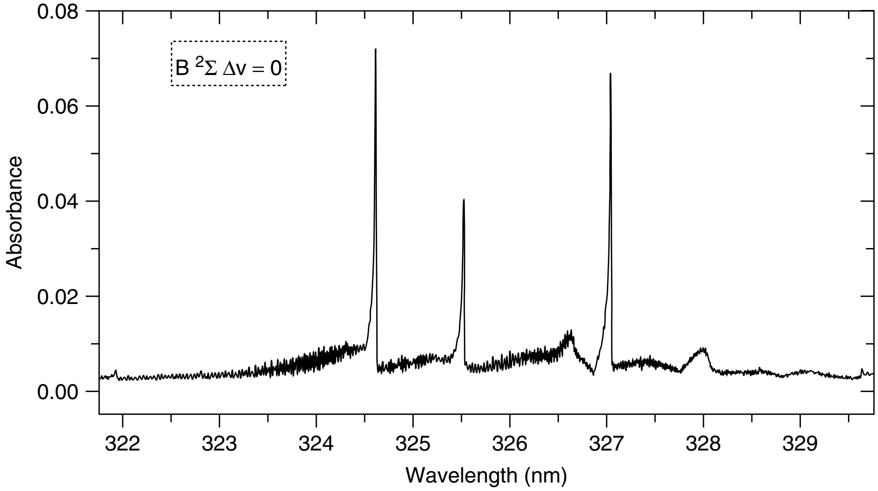

Fig. 13.23 Detail of the molecular spectrum of the phosphorus monoxide PO in the range between 322 and 329 nm, corresponding to the electronic transition X 2Π → B 2Σ. (From Welz *et al.* (2005), p. 195; reproduced with permission from Wiley-VCH.)

Advantages and Applications of HR-CS AAS There are several advantages of HR-CS AAS over conventional LS AAS, which obviously include the use of a single radiation source for all elements instead of a separate source for each analyte. The high emission intensity of the Xe short-arc lamp has no influence on the measured absorbance, which, according to Eq. (13.4) is proportional to the ratio of the incident and the transmitted radiant power; however, the emission intensity has an influence on the S/N ratio, and hence on the LOD, as the noise is inversely proportional to the square root of the radiant power. The visibility of the spectral environment around the analytical line at high resolution is providing a new dimension of information that has previously been unknown, and which can readily be used for method development and to avoid interferences. The truly simultaneous correction for all wavelength-independent fluctuations of the radiant power, including continuous background absorption, have been discussed in Section 13.3.3.3, Background and Baseline Correction, together with the correction capabilities for discontinuous background. Obviously, HR-CS AAS can be used for any application traditionally assigned to LS AAS, but it has also been shown to improve and facilitate the analysis of complex samples, especially using direct solid sampling (Welz *et al.*, 2007b). As the use of a continuous radiation source makes any wavelength between 190 and 900 nm available for analytical purposes, HR-CS AAS is not limited to atomic absorption, but can also be used for the determination of nonmetals, such as phosphorus, sulfur, and the halogens using the absorption spectra of diatomic molecules, such as PO, which is shown in part in Figure 13.23, CS, AlF, and so on (Huang *et al.*, 2006(a, b, c); Huang *et al.*, 2008; Heitmann *et al.*, 2006). The analytical applications of HR-CS AAS have been recently reviewed by Welz *et al.* (2007a).

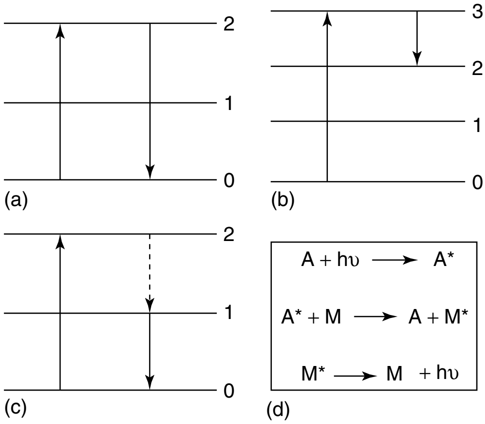

Fig. 13.24 Schematic presentation of important processes for AFS; (a) resonance fluorescence; (b) direct-line fluorescence; (c) stepwise-line fluorescence; (d) sensitized fluorescence. (Adapted from Greenfield (1994), p. 566, reproduced with permission of The Royal Society of Chemistry.)

13.3.4
Atomic Fluorescence Spectrometry

Atomic fluorescence was studied as early as 1902 by Wood (1902), and Nichols and Howes (1924) considered fluorescence in flames, but neither of these publications dealt with analytical applications. Winefordner and Vickers (1964) were the first to investigate the possibilities of using AFS as a practical analytical technique. Veillon *et al.* (1966) demonstrated the usefulness of continuum radiation sources for AFS, whereas West (1967) investigated EDL. Further information about the research activities during that period might be obtained from a review article by Winefordner and Mansfield (1967). The major advantage of AFS was the significantly better *S*/*N* ratio, and hence better LOD compared to FAAS. Nevertheless, AFS has never found general application as an analytical technique. Among the reasons for that might be that the best *S*/*N* ratios were obtained in low-temperature flames that were more subject to matrix effects. The main reason, however, was that the research in AFS was coinciding in time with the development of GF AAS and ICP OES, and none of the instrument manufacturing companies was prepared to invest in yet another technique. Nowadays, AFS has found a niche in combination with chemical vapor generation for the determination of mercury and some hydride-forming elements, such as arsenic and selenium.

13.3.4.1 General Principles

Atomic fluorescence is a radiational deactivation process, which occurs after the excitation of free atoms by the absorption of radiation of characteristic wavelength from an excitation source. Atomic fluorescence can occur by a number of different processes, but only four of them are considered to be of importance and are schematically represented in Figure 13.24 (Browner, 1974; Robinson, 1990; Greenfield, 1995).

Resonance fluorescence (Figure 13.24a) occurs when atoms absorb and reemit radiation of the same wavelength. Lines corresponding to transitions between excited electronic states and the ground state of an atom constitute most analytically useful examples of resonance fluorescence, since the transition probabilities for resonance transitions are

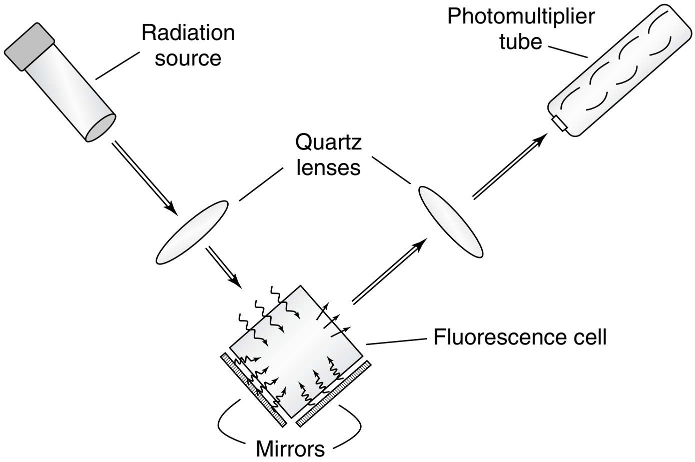

Fig. 13.25 Schematic design of an atomic fluorescence spectrometer.

usually much greater than those for other transitions.

Nonresonance fluorescence occurs when the exciting line and the observed fluorescent line are of different wavelengths. *Stokes direct-line fluorescence* results when the exciting and the fluorescent lines originate from optical transitions with a common upper level (Figure 13.24b). An atom is excited (usually from the ground state) to a higher excited state, and then undergoes a direct transition to a metastable level above the ground state, emitting a fluorescent line at longer wavelength, compared to that of the exciting radiation. When the upper level of the exciting and the emitted lines are different, *Stokes stepwise-line fluorescence* is said to occur (Figure 13.24c). Atoms excited by radiation lose part of their energy, usually by collisional deactivation, before emitting fluorescence radiation of longer wavelength. *Anti-Stokes fluorescence* (not shown here) occurs when the wavelength of the fluorescent line is shorter than that of the exciting line. The deficit in photon energy is usually supplied thermally, and the term *thermally-assisted* fluorescence can be used.

In addition to the above-discussed types of atomic fluorescence, another, although less usual type exists, called *sensitized fluorescence*. In this process, an atom or molecule excited by absorption of electromagnetic radiation (donor) transfers its excitation energy to the sample atom (acceptor) by collision. The acceptor then undergoes radiative deactivation, resulting in atomic fluorescence. The process is schematically represented in Figure 13.24(d).

13.3.4.2 **Instrumentation**

The basic setup of an AFS instrument consists in a radiation source, an atomizer, a spectral sorting device, a detector, and devices for signal processing and readout, as shown in Figure 13.25. The detection of the fluorescence radiation should always be misplaced in relation to the incident radiation, that is, for instance, be positioned at a 90° angle, in order to prevent that photons emitted by the radiation source reach the detector.

Atomizers for AFS The physical and chemical environments within an atomizer have a profound effect upon the magnitude of the fluorescent flux emitted from the atomizer in addition to its influence on atomization efficiency. In practice, the excited atom is almost inevitably subject to radiationless deactivation due to collisions with other species present in the atomizer. These so-called *quenching collisions* are characterized by a quantum yield, Y, which is the number of quanta emitted as fluorescence to the number of quanta absorbed and capable of producing fluorescence. Quantum yields were calculated for several analytes, gas mixtures, and temperatures (Sychra, Svoboda and Rubeška, 1975); for air–acetylene and air–hydrogen flames Y was found to be typically below 0.1; values close to $Y = 1$ could only be obtained in hot argon environment, which would make ET atomizers an ideal source for AFS.

A variety of atomizers have been used for AFS, including flames (Sychra, Svoboda and Rubeška, 1975), graphite furnaces (Massmann, 1968), and ICP (Greenfield, 1994); however, there are no instruments commercially available with such configurations. The only instrumentation that is available nowadays for AFS are specialized apparatus for the determination of mercury (Cai, 2000) using the cold-vapor technique, which does not require any atomizer, and for the hydride-forming elements (refer to Section 13.3.3.1, Atomizers for Chemical Vapor Generation). In the latter case special atomizers are used, such as small hydrogen diffusion flames, where the atomization occurs by a chemical process rather than a thermal one. The interferences that might be observed during generation and separation of the gaseous analyte species from the liquid phase are identical to those observed for AAS (refer to Section 13.3.3.1, Atomizers for Chemical Vapor Generation).

Radiation Sources The intensity of the fluorescence signal is directly proportional to the intensity of the incident radiation, that is, the excitation source has a major influence on the sensitivity that can be obtained by AFS. This dependence can be mathematically expressed by means of the simplified Eq. (13.8), where Φ_0 and Φ_F represent the incident and the fluorescence (emitted) radiant power, κ_0 corresponds to the atomic absorption coefficient, ℓ is the length of the absorption cell, and φ is the quantum efficiency of the fluorescence process (Robinson, 1990).

$$\Phi_F = \Phi_0(1 - e^{-\kappa_0 \ell})\varphi \qquad (13.7)$$

Significant research in AFS was therefore done for the development of high-intensity radiation sources, particularly electrodeless discharge lamps (Sychra, Svoboda and Rubeška, 1975). Modern instruments usually employ a line radiation source, such as a hollow cathode lamp or an electrodeless discharge lamp; these lamps are similar to those used for AAS (refer to Section 13.3.3.2, Radiation Sources for LS AAS). Other sources, such as specially designed mercury vapor lamps, might be used as well.

When lasers are used as an excitation source in AFS, the technique is called *laser-excited atomic fluorescence spectrometry* (*LEAFS*). Among the advantages that are associated with the use of lasers are the higher radiant power compared to conventional line sources, resulting in high sensitivity, and the extended linear dynamic range, which may cover up to 7 orders of magnitude, and the possibility to perform sequential multielement analysis, using oscillator-based lasers (Stchur *et al.*, 2001; Hou *et al.*, 1998). Disadvantages include the high cost and complexity of laser systems.

Spectral Sorting Devices and Detectors One of the advantages of AFS is the selective character of the fluorescence radiation to be monitored, that is, usually only the analyte element can absorb the radiation emitted by the primary source if a line source is used. This means that the monochromator might even be a simple nondispersive system, that is, a filter. The monochromator used in AFS has the sole function of isolating the analytical fluorescence wavelength from other wavelengths that might be emitted by concomitants or by the analyte. Lasers require more sophisticated optical devices, such as beam expanders and diffraction gratings (Hou *et al.*, 1998).

The detectors used in AFS are similar to those used in AAS and OES. The PMT is by far the most frequently used detection device, although solid-state detectors, such as the CCD, have also been employed (for details refer to Section 13.3.2.3).

Interferences Nonspectral interferences are obviously independent of the detection technique and specific for the sample introduction and atomization technique. Reference can therefore be made to similar atomizers used in AAS. Spectral interferences due to direct-line or near-line overlap is extremely unlikely with line radiation sources, as usually only the analyte can absorb the radiation and emit fluorescence. The most serious problem, which cannot be eliminated by improving spectral resolution or by modulating the radiation source, is radiation scattering at nonvolatilized aerosol particles, particularly in flame atomizers. This problem is most serious for resonance fluorescence and might be avoided using a nonresonance fluorescence line; this, however, is usually associated with a significant loss of sensitivity and therefore not generally applicable. Another, although less frequently encountered problem is broadband fluorescence emitted by molecules in the same spectral interval where atomic fluorescence is monitored. This interference also cannot be eliminated by modulating the radiation source, as the fluorescence is caused by the radiation source. Both interferences are much more serious if a continuum radiation source is used instead of a line source. The problem of thermal emission from the atomizer can be solved by modulating the primary radiation source and using a selective amplifier.

13.4
X-Ray Spectroscopy

X-ray spectroscopy (XRS), like optical spectroscopy, is based on the measurements of emission, absorption, fluorescence, scattering, and diffraction of electromagnetic radiation. X-rays are defined as radiation of short wavelength, typically between 10 pm and 10 nm (0.1–100 Å), which is generated by *deceleration* of high-energy electrons or by electron transitions in the *inner orbitals* of atoms (Jenkins and de Vries, 1970; Jenkins, 1988). Upon ejection of an inner electron from an atom, the remaining electrons in the outer orbitals will decay to lower-energy states to compensate for the electron loss and to fill in the vacancy in the inner orbital. These transitions are characterized by the emission of the excess energy in the form of X-ray photons with characteristic energies corresponding to the energy of the electron transition; therefore, each element has its own characteristic X-ray spectrum (Carr-Brion and Payne, 1970). These transitions are named after the orbital from where the original electron has been ejected. For instance, ejection of an electron from the innermost orbital generates "K" X-rays; ejection from the next orbital generates "L" X-rays, and so on, which is in accordance to the atom model of Bohr. Each emission is also associated to a determined transition probability, which is indicated by Greek letters (α, β, γ). For instance, a K-series spectrum, corresponding to ionization from a K-shell, consists of a strong K_α and a weaker K_β line due to a less-frequent electron transition (Müller, 1972).

13.4.1
Theoretical Principles

13.4.1.1 Emission of X-Rays

For analytical purposes, X-rays may be produced in different ways: (i) by bombardment of a metal target with a high-energy electron beam; (ii) by exposing a substance to a primary X-ray, generating a secondary X-ray fluorescence beam; (iii) by using a radioactive source that emits X-rays upon its decay; and (iv) by using synchrotron radiation. X-ray sources, like sources for optical radiation, often generate both, continuous and discontinuous (line) spectra. The continuous radiation is known as *white radiation* or *bremsstrahlung*, as it is the result of the deceleration of high-energy electrons.

The most usual X-ray sources are *X-ray tubes*, which are operated at controlled (high) voltages up to 100 kV applied between a cathode and an anode. In an X-ray tube, a hot filament, acting as a cathode, emits electrons that are accelerated in a high-voltage field to impact against a metal anode (target). The impact of the electrons with the anode generates heat and electromagnetic radiation in the form of X-rays, as a result of the deceleration of the electrons and the conversion of kinetic energy. Under certain conditions, only a continuous spectrum is obtained, as shown in Figure 13.26(a), which under other conditions may be superimposed by a line spectrum, as shown in Figure 13.26(b). The continuous X-ray spectra are characterized by an exactly defined short-wavelength limit, λ_0, which only depends on the acceleration voltage, but is independent of the target material. The maximum photon energy corresponds to an abrupt deceleration of the electron to the kinetic energy 0 in a single impact; the continuum at longer wavelengths corresponds to a stepwise deceleration in multiple collisions. The superimposed line spectra are due to electron transitions within the atom, and are characteristic for the element, as is discussed in the context of X-ray fluorescence (refer to Section 13.4.1.3). The line spectra are simple and consist of a few lines only.

Radioisotopes are also common excitation sources; however, they are some 4 orders of magnitude weaker compared to X-ray tubes. In this case, a γ-ray-emitting radioisotope, such as ^{241}Am, ^{109}Cd, ^{153}Gd, ^{155}Eu, or ^{145}Sm is used, and γ-rays induce the emission of

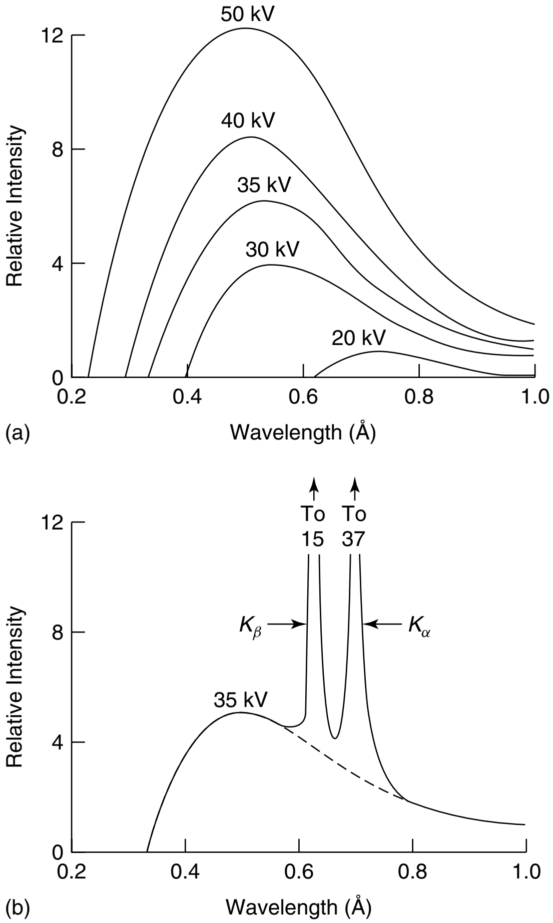

Fig. 13.26 Typical spectra emitted by X-ray tubes under different conditions; (a) continuous spectra emitted by an X-ray tube with a tungsten target using different acceleration voltages; (b) continuous spectrum with superimposed line spectrum emitted by an X-ray tube with a molybdenum target. (From Skoog *et al.* (1998), Brooks & Cole, P. 273; reproduced with permission from Cenage learning Inc.)

X-rays either by direct impact on the sample surface or by impact on a secondary target, which will then emit X-rays that are used to excite sample components (known as the *γ-X source*). Several advantages are ascribed to radioisotope excitation sources, such as high stability and low cost (Jenkins, 1988).

The acceleration of high-energy electrons in a high vacuum tube under the influence of a strong magnetic field results in the emission of the so-called *synchrotron radiation*, which can also be used as an excitation source in XRS. Synchrotron radiation can be considered as magnetic *bremsstrahlung*, and it can be used as a source either directly or after modifications, which include selection by an absorber and the use of diffraction crystals and mirrors. Owing to the high intensity of synchrotron radiation, the analytical sensitivity of X-ray fluorescence (XRF) is significantly increased, as well as the S/N ratio (Saisho, 1989).

13.4.1.2 Absorption of X-Rays

When an X-ray penetrates a thin layer of a solid substrate, its intensity is reduced due to absorption and scattering. Except for the lightest elements the latter effect can be neglected, at least in the wavelength range of strong absorption. Similar to the emission spectrum, the absorption spectrum of an element is simple and consists of a few lines only. The absorption spectrum of lead and silver, which are shown in Figure 13.27, consist of a few peaks only with the first one at 0.14 Å. A peculiarity of X-ray absorption spectra is their discontinuous increase with so-called absorption edges. One of the inner electrons of the atomic shell is expelled upon the absorption of an X-ray photon, resulting in an excited ion. In this process, the entire energy $h\nu$ of the radiation is distributed to the kinetic energy of the electron and the potential energy of the excited ion. The probability of absorption is highest when the photon energy corresponds exactly to the energy that is necessary to expel the electron from the atom, that is, if the kinetic energy of the expelled electron is close to zero. For higher photon energy the probability decreases gradually, and for lower photon energy abruptly, as the photon does not have enough energy to expel an ion from the K-shell any more, and the next peaks are reached only when the photon energy reaches a level that corresponds to the energy of the L electrons.

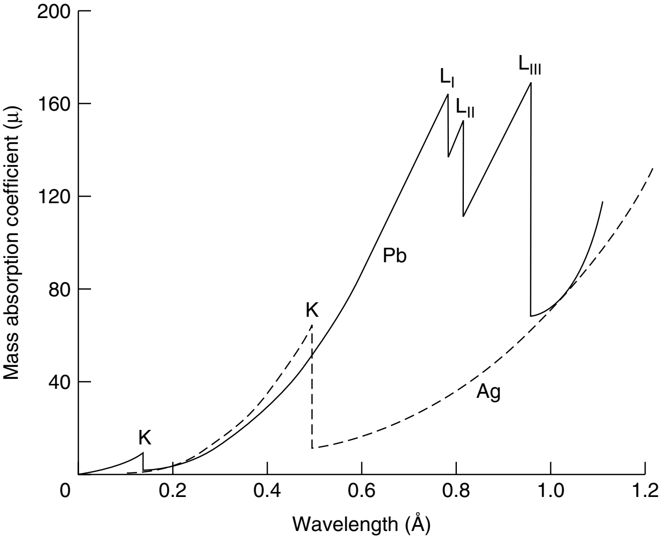

Fig. 13.27 X-ray absorption spectrum of silver and lead. (From Skoog *et al.* (1998), Brooks & Cole, P. 277; reproduced with permission from Cenage learning Inc.)

13.4.1.3 X-Ray Fluorescence (XRF)

The absorption of X-rays produces electronically excited ions that return into their ground state by electron transitions from higher energy levels. These transitions are associated by the emission (fluorescence) of characteristic wavelengths, which are identical to those that are generated by electron bombardment. However, the wavelengths observed in XRF are always longer than those of the corresponding absorption edge, as the complete liberation of the electron, that is, ionization of the atom, is necessary for absorption, whereas fluorescence only requires transition of an electron from a higher to a lower energy level. The intensity of the emitted X-ray photon will depend on the effectiveness of the de-excitation process within a given atom, and the *fluorescence yield* can be defined as the fraction of all electron transitions that are associated with the emission of an X-ray photon (Müller, 1972). The fluorescence yield increases dramatically with increasing atomic number, and is close to unity for elements with atomic numbers above 70.

13.4.2
Instrumentation

13.4.2.1 X-Ray Fluorescence

The most commonly used technique is XRF, which has a wide range of application in the analysis of solid materials and is characterized as a nondestructive technique. Commercially available XRF instruments can be classified into wavelength-dispersive and energy-dispersive instruments. *Wavelength-dispersive* (WD) instruments, which are schematically represented in Figure 13.28(a), generally use an X-ray tube as the source in order to compensate the energy losses that are encountered when an X-ray is first collimated onto a crystal and then dispersed into its wavelength components. Analyzing crystals made of a wide variety of materials have been used, including LiF, Si, pentaerythritol (PE), topaz, CaF_2, and many

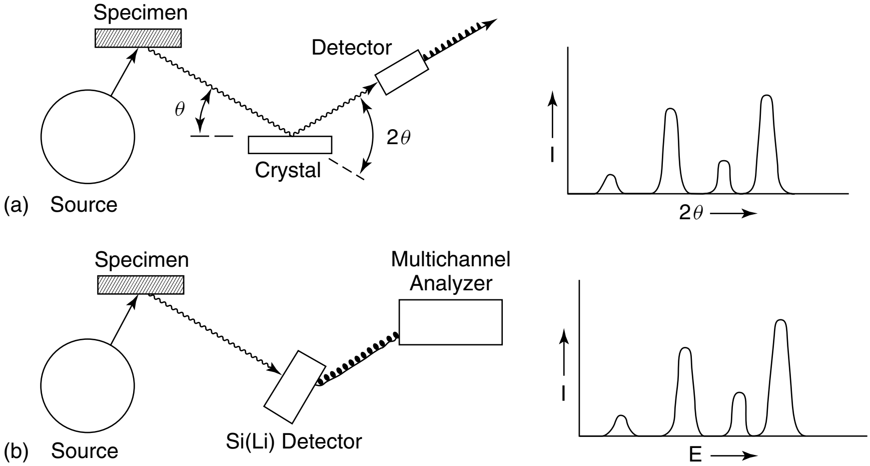

Fig. 13.28 Schematic design of typical XRF instruments; (a) wavelength-dispersive X-ray fluorescence equipment; (b) energy-dispersive X-ray fluorescence equipment. (From Jenkins (1988), p. 54; reproduced with permission from Wiley-Interscience.)

others, depending mainly on the atomic number of the element to be determined and on the specific application. We distinguish between sequential single channel and simultaneous multichannel instruments.

Single-channel instruments can be operated manually or automatically. The former ones can be used for the *quantitative determination* of a few elements. In this case, crystal and detector are adjusted at an appropriate angle (θ and 2θ) and the photons counted for a long enough time to produce a sufficiently high signal. Automatic instruments are better suited for *qualitative analysis*, where the entire spectrum has to be registered. For this purpose, analyzing crystal and detector are rotated synchronously maintaining their relative angular position (Lachance and Claisse, 1994).

Wavelength-dispersive *multichannel instruments* are big and expensive instruments that allow the simultaneous detection and determination of up to 24 elements. The individual channels, each of which consists of a crystal and a detector, are grouped radially around an X-ray source and a sample holder at an appropriate angle for a given analyte line. Each detector has its own amplifier, analyzer, and counter or integrator. The simultaneous determination of all elements may take between a few seconds and a few minutes. Instruments of this type are made for a specific application and offer little flexibility for other purposes.

In *energy dispersive* (*ED*) instruments, illustrated in Figure 13.28(b), the fluorescence radiation beam coming from the sample is immediately directed toward a semiconductor detector, which is coupled to several electronic devices that are capable of selecting the signal according to small differences in the energy, depending on their wavelength. In this case, a spectrum of voltage pulses is obtained, which is directly proportional to the spectrum of X-ray photon energies entering the detector, as shown in Figure 13.29 (Jenkins 1988; Lachance and Claisse, 1994).

The obvious advantages of ED systems are their simpler design and the absence of movable parts, which also simplifies their operation. The absence of collimators and diffracting crystals, and the short distance between sample and detector increase the signal intensity by about 2 orders of magnitude, which allows the use of much weaker (and less expensive) radiation sources (refer to Section 13.4.2.1). Although this then results in lower sensitivity compared to wavelength-dispersive instruments, their much lower price and their ability to measure all the elements within the range at the same time makes them an attractive alternative.

There are three main types of X-ray detectors to convert X-ray photon energy into voltage pulses. A gas flow *proportional counter* is a tube filled with a mixture of an inert gas (typically Ar) and a quenching gas (such as CH_4) with a wire along its radial axis. The tube wall is grounded and a high potential (usually around 1.8 kV) is applied to the wire. As the X-ray photons enter the tube, the inert gas is ionized and the electrons generated in the ionization process are accelerated toward the wire (anode), colliding with other neutral gas species and generating more ionized gas species in a cascade effect. A voltage pulse is therefore produced, which is proportional to the energy of the incident X-ray photon. The proportional counter is ideally suited for measurement of longer wavelengths (above 1.5 Å).

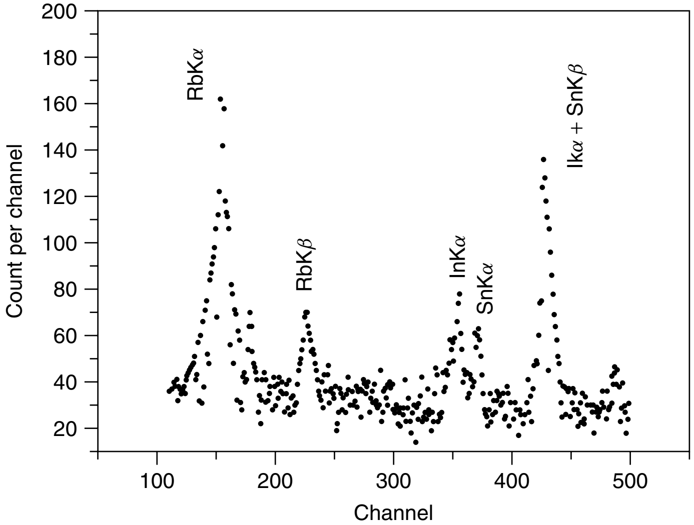

Fig. 13.29 Spectrum of an iron sample recorded with an ED XRF equipment. (From Ekinci and Íngeç (2008), p. 1120, reproduced with permission from Elsevier.)

A *scintillation counter* can be used for shorter wavelengths. In this detector, the incident X-ray photons impact against a phosphor substance, such as a NaI crystal doped with Tl, resulting in the emission of lower energy photons, which are then accelerated to a PMT, generating electrons in a cascade effect, similar to the PMT used in optical spectroscopy. The generated current is proportional to the energy of the incident X-ray photon. A variation of this detector is the *gas scintillation counter*, where a gas is used instead of a solid phosphor.

Semiconductors are also frequently employed in detectors, particularly in energy-dispersive instruments. The most usual is the *Si(Li) detector*, where a small Si cylinder, compensated by Li in order to increase the electrical resistivity, is used. In this case, a p–i–n-type diode is assembled and, upon impact, X-ray photons generate a number of electron-hole pairs and, consequently, a determined electric charge. Application of a high voltage to the Si disk results in transfer of the accumulated charge to a charge-sensitive preamplifier, producing a charge pulse, the voltage amplitude of which is proportional to the energy of the incident X-ray photon. Electronic noise is reduced by cooling the detector, usually with liquid nitrogen.

The major limitation of all XRF techniques is the need of solid reference materials – usually certified reference materials – with an analyte content and a matrix composition as close as possible to the composition of the sample for calibration in order to compensate for matrix effects.

13.4.2.2 Total-Reflection (TR) XRF

Total reflection XRF is a variation of energy dispersive X-ray fluorescence (ED XRF), which is based on the fact that at an incident angle below the critical angle, the primary radiation beam is completely reflected. The reflection angle is exactly the same as the incidence angle, and the intensity is also roughly the same, except for a small fraction of the radiation that may be refracted by the sample (Wobrauschek, 2007; Yoneda and Horiuchi, 1971). As at a low incidence angle the X-ray photons practically do not penetrate the sample substrate or support, the scattering of radiation, which is a major source of background in XRF, is remarkably reduced, which ultimately results in an improvement in the S/N ratio and, consequently, in the LOD. In addition, the sample itself is irradiated twice by the X-ray photons, as both the primary incident and the reflected beam strike the sample surface. The result is a highly sensitive XRF technique with LOD in the picogram per gram range for selected applications (Knoth and Schwenke, 1978; Wobrauschek, 2007).

The instrumental setup for total-reflection X-ray fluorescence (TR XRF) is not much different from that for ED XRF, except for the fact that the primary radiation beam has to be adjusted to enter the sample at a glancing angle of a few arc seconds. This can be achieved by placing reflection units and controlled apertures in the beam path between the source and the sample chamber. In addition, in TR XRF, the detector is positioned directly above the sample, as represented schematically in Figure 13.30. As in most ED XRF instruments, X-ray tubes, and Si(Li) detectors are commonly used.

Prior to TR XRF analysis, samples must be placed on a reflective, flat polished surface. Measurements can be performed in air, in a He atmosphere or in a vacuum chamber, which has the advantage of avoiding the scattering of primary X-ray photons from atmospheric gas molecules and is mandatory

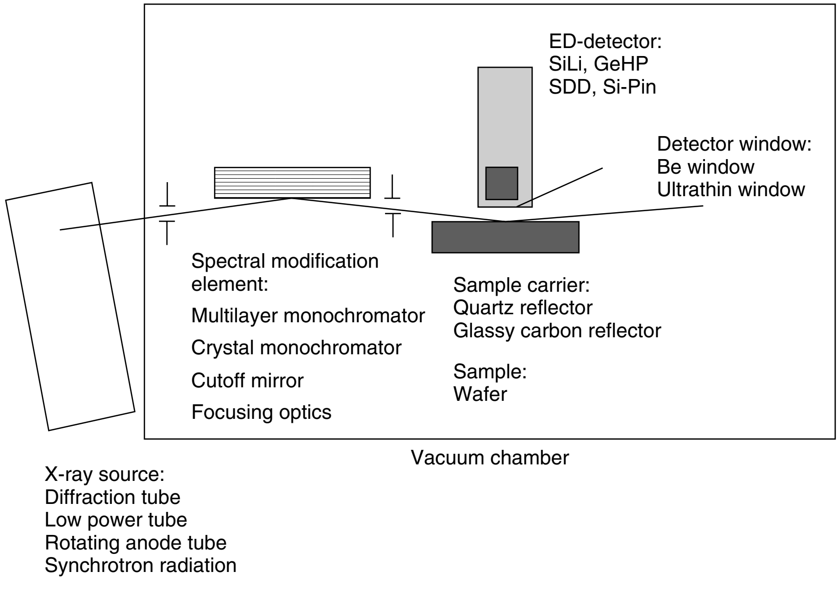
ED-detector:
SiLi, GeHP
SDD, Si-Pin
Detector window:
Be window
Ultrathin window
Spectral modification element:
Multilayer monochromator
Crystal monochromator
Cutoff mirror
Focusing optics
Sample carrier:
Quartz reflector
Glassy carbon reflector
Sample:
Wafer
Vacuum chamber
X-ray source:
Diffraction tube
Low power tube
Rotating anode tube
Synchrotron radiation

Fig. 13.30 Instrumental setup for TR XRF. (From Wobrauschek (2007), p. 293; reproduced with permission from Wiley.)

for the detection of low-atomic-number elements. Samples are more suitably analyzed as aqueous solutions, and solid samples should typically be digested prior to analysis (Klockenkämper and von Bohlen, 1999). The application of TR XRF for trace element determination covers a wide range of sample types; a complete review has been published by Wobrauschek (2007).

13.4.2.3 Particle-Induced X-Ray Emission (PIXE)

Particle (or proton) induced X-ray emission (PIXE) is a technique based on the emission of X-ray photons by sample constituents excited by collision with charged particles, usually protons. The sequence of events is identical to the one described for XRF, that is, emission of X-ray photons following the filling of electron vacancies by outer-shell electrons (Johansson and Campbell, 1988). Instrumentation for PIXE is rather similar to that for ED XRF, except for the excitation source. In PIXE, a proton beam with energies varying typically between 1 and 10 MeV and beam currents of a few microamperes is generated by means of a small particle accelerator. The proton beam is collimated, usually by a magnetic field, and the proton energy is stabilized and uniformed by passing through a slit, before entering the sample chamber, which is under vacuum. An Si(Li) detector is typically employed (Johansson and Campbell, 1988; Khan and Crumpton, 1981).

Several advantages are associated with the use of PIXE, such as the ability to perform fast multielement analysis, the high sensitivity for elements with atomic numbers above 11, the suitability for small samples (micro- or μ-PIXE) and the lower *bremsstrahlung*, due to the smaller deceleration experienced by protons upon collision with electrons (Khan and Crumpton, 1981). The application of PIXE ranges from the analysis of pigments in historic paintings (Neelmeijer *et al.*, 2000) to biological samples (Bertrand, Weber and Schoefs, 2003).

13.4.2.4 Portable Instrumentation

Portable (field) XRF instruments are commercially available and have been gaining importance for a range of on-site applications. The development of portable XRF instruments has become possible mostly due to the development of small X-ray tubes and to the use of radioisotopes as X-ray sources. Portable instruments have been mostly developed for "conventional" XRF, but instruments for PIXE (Pappalardo *et al.*, 2007) and TR XRF (Mages *et al.*, 2003), among others, have been described as well. Applications range from archaeometrical studies (Zarkadas and Karydas, 2004) to coating thickness evaluation (Carapelle *et al.*, 2007).

13.5
Outlook

Most of the instrumental techniques of atomic spectrometry can nowadays be considered established techniques, so that no dramatic innovations should be expected. Miniaturization is an interesting trend that not only reduces the size of the equipment but also often reduces sample consumption and measurement time, and improves LOD and LOQ. Microplasmas with dimensions of 1 mm or less, usually microwave-induced, generated by high-frequency or dc discharges, are a promising field of research in OES, and have been experiencing increasing popularity in the past few years (Foest, Schmidt and Becker, 2006; Broekaert and Siemens, 2004). Yu, Du and Wang (2007) proposed a miniature AFS system in a lab-on-valve for mercury determination, and Wormhoudt *et al.* (2005) described a determination of carbon in steel by LIBS using a microchip laser and miniature spectrometer. It is not easy to predict, however, if such systems will find entrance into routine laboratories, or if they will remain restricted to research labs.

The most recent development in atomic spectrometry, HR-CS AAS, has been received extremely well in the market, and will have a significant impact not only on classical LS AAS but also on alternate techniques, such as ICP OES and ICP-MS. The most obvious next step will be the development of simultaneous HR-CS AAS, which will make it even more competitive. Although this development depends on further progress in detector technology, it will, for sure, take less than a decade until this step becomes technically feasible.

Acknowledgments

The authors are grateful to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for scholarships and financial support.

Acronyms

AAS	atomic absorption spectrometry
ac	alternating current
AES	auger electron spectroscopy
AFS	atomic fluorescence spectrometry
BEC	background-equivalent concentration
CCD	charge-coupled device
CID	charge-injection device
CV	cold vapor (technique)
CV AAS	cold vapor atomic absorption spectrometry
CVG	chemical vapor generation
dc	direct current
DCP	direct current plasma
ED	energy dispersive
EDL	electrodeless discharge lamp
ED XRF	energy dispersive X-ray fluorescence
ETA	electrothermal atomization
ETV	electrothermal vaporization
FAAS	flame atomic absorption spectrometry
FWHM	full width at half maximum (half width)
GD	glow discharge
GF	graphite furnace
GF AAS	graphite furnace atomic absorption spectrometry
HCL	hollow cathode lamp
HG	hydride generation (technique)
HG AAS	hydride generation atomic absorption spectrometry
HR-CS AAS	high-resolution continuum source atomic absorption spectrometry
HR-CS GF AAS	high-resolution continuum source graphite furnace atomic absorption spectrometry
ICP	inductively coupled plasma
ICP-MS	inductively coupled plasma mass spectrometry
ICP OES	inductively coupled plasma optical emission spectrometry

IUPAC	International Union of Pure and Applied Chemistry
LEAFS	laser-excited atomic fluorescence spectrometry
LIBS	laser-induced breakdown spectroscopy
LOD	limit of detection
LOQ	limit of quantification
LS AAS	line source atomic absorption spectrometry
MIP	microwave-induced plasma
OES	optical emission spectrometry
PIXE	particle-induces X-ray emission
PMT	photomultiplier tube
PVSA	peak volume selected absorbance
QTA	quartz tube atomizer
RF	radio frequency
S/N	signal-to-noise (ratio)
STPF	stabilized temperature platform furnace
THB	tetrahydroborate
TR XRF	total-reflection X-ray fluorescence
UV	ultraviolet
WD	wavelength dispersive
WD XRF	wavelength dispersive X-ray fluorescence
WSA	wavelength selected absorbance
XRF	X-ray fluorescence
XRS	X-ray spectroscopy
ZBC	Zeeman-effect background correction

References

Angeli, J., Bengtson, A., Bogaerts, A., Hoffmann, V., Hodoroaba, V.D., and Steers, E. (2003) *J. Anal. At. Spectrom.*, **18**, 670.

Balmer, J.J. (1885) *Ann. Phys.*, **25**, 80.

Barnes, R.B., Richardson, D., Berry, J.W., and Hood, R.L. (1945) *Ind. Eng. Chem. Anal. Ed.*, **17**, 605.

Becker, J.S. (2007) *Inorganic Mass Spectrometry*, John Wiley & Sons, Ltd, New York.

Becker-Ross, H., Florek, S., and Heitmann, U. (2000) *J. Anal. At. Spectrom.*, **15**, 137–141.

Berndt, H. and Schaldach, G. (1988) *J. Anal. At. Spectrom.*, **3**, 709–712.

Bertrand, M., Weber, G., and Schoefs, B. (2003) *Trends Anal. Chem.*, **22**, 254.

Bohr, N. (1913) *Philos. Mag.*, **26**, 476.

Borges, D.L.G., Silva, A.F., Curtius, A.J., Welz, B., and Heitmann, U. (2006) *J. Anal. At. Spectrom.*, **21**, 763–769.

Boss, C.B. and Fredeen, K.J. (1999) *Concepts, Instrumentation and Techniques in Inductively Coupled Plasma Optical Emission Spectrometry*, Perkin-Elmer Co., Norwalk, USA.

Braman, R.S., Justen, L.L., and Forebank, C.C. (1972) *Anal. Chem.*, **44**, 2195–2199.

Broekaert, J.A.C. (2005) *Analytical Atomic Spectrometry with Flames and Plasmas*, Wiley-VCH Verlag GmbH, Weinheim, New York.

Broekaert, J.A.C. and Siemens, V. (2004) *Anal. Bioanal. Chem.*, **380**, 185–189.

Browner, R.F. (1974) *Analyst*, **99**, 617.

Cai, Y. (2000) *TrAC, Trends Anal. Chem.*, **19**, 62.

Carapelle, A., Fleury-Frenette, K., Collette, J.P., Garnir, H.P., and Harlet, P. (2007) *Rev. Sci. Instrum.*, **78**, 123109.

Carr-Brion, K.G. and Payne, K.W. (1970) *Analyst*, **95**, 977.

Das, D.K., McDonald, J.P., Yalisove, S.M., and Pollock, T.M. (2008) *Spectrochim. Acta, Part B*, **63**, 27.

Dedina, J. (1988) *Prog. Anal. Spectrosc.*, **11**, 251–360.

De Giacomo, A., Dell'Aglio, M., De Pascale, O., Longo, S., and Capitelli, M. (2007) *Spectrochim. Acta, Part B*, **62**, 1606.

De Loos-Vollebregt, M.T.C. and de Galan, L. (1985) *Prog. Anal. Spectrosc.*, **8**, 47–81.

D'Ulivo, A., Baiocchi, C., Pitzalis, E., Onor, M., and Zamboni, R. (2004a) *Spectrochim. Acta, Part B*, **59**, 471–486.

D'Ulivo, A., Onor, M., and Pitzalis, E. (2004b) *Anal. Chem.*, **76**, 6342–6352.

Ekinci, N. and Íngeç, M. (2008) *Appl. Radiat. Isot.*, **66**, 1117–1122.

Fantoni, R., Caneve, L., Colao, F., Fornarini, L., Lazic, V., and Spizzichino, V. (2008) *Spectrochim. Acta, Part B*, **63**, 1097–1108.

Foest, R., Schmidt, M., and Becker, K. (2006) *Int. J. Mass Spectrom.*, **248**, 87–102.

Fraunhofer, J. (1817) *Ann. Phys.*, **56**, 264.

Galindo, R.E. and Albella, J.M. (2008) *Spectrochim. Acta, Part B*, **63**, 422.

Gleisner, H., Eichardt, K., and Welz, B. (2003) *Spectrochim. Acta, Part B*, **58**, 1663–1678.

Goujon, J., Giakoumaki, A., Piñon, V., Musset, O., Anglos, D., Georgiou, E., and Boquillon, J.P. (2008) *Spectrochim. Acta, Part B*, **63**, 1091–1096.

Greenfield, S. (1994) *J. Anal. At. Spectrom.*, **9**, 565.

Greenfield, S. (1995) *TrAC, Trends Anal. Chem.*, **14**, 435.

Greenfield, S., Jones, I.L.W., and Berry, C.T. (1964) *Analyst*, **89**, 713.

Grimm, W. (1968) *Spectrochim. Acta, Part B*, **23**, 443.

Guo, X., Sturgeon, R.E., Mester, Z., and Gardener, G.J. (2004) *Anal. Chem.*, **76**, 2401–2405.

Harmon, R.S., De Lucia, F.C., Miziolek, A.W., McNesby, K.L., Walters, R.A., and French, P.D. (2005) *Geochem.: Explor. Environ., Anal.*, **5**, 21–28.

Harnly, J.M. (1999) *J. Anal. At. Spectrom.*, **14**, 137.

Hatch, W.R. and Ott, W.L. (1968) *Anal. Chem.*, **40**, 2085–2087.

Heitmann, U., Becker-Ross, H., Florek, S., Huang, M.D., and Okruss, M. (2006) *J. Anal. At. Spectrom.*, **21**, 1314–1320.

Heitmann, U., Schütz, M., Becker-Ross, H., and Florek, S. (1996) *Spectrochim. Acta, Part B*, **51**, 1095–1105.

Heitmann, U., Welz, B., Borges, D.L.G., and Lepri, F.G. (2007) *Spectrochim. Acta, Part B*, **62**, 1222–1230.

Hergenröder, R. and Niemax, K. (1988) *Spectrochim. Acta, Part B*, **43**, 1443–1449.

Herschel, J.F.W. (1823) *Trans. R. Soc. Edinb.*, **9**, 445.

Hieftje, G.M. (1989) *J. Anal. At. Spectrom.*, **4**, 117–122.

Hodoroba, V.D., Unger, W.E.S., Jenett, H., Hoffmann, V., Hagenhoff, B., Kayser, S., and Wetzig, K. (2001) *Appl. Surf. Sci.*, **179**, 30–37.

Holak, W. (1969) *Anal. Chem.*, **41**, 1712–1713.

Hou, X., Stchur, P., Yang, K.X., and Michel, R.G. (1998) *TrAC, Trends Anal. Chem.*, **17**, 532.

Huang, M.D., Becker-Ross, H., Florek, S., Heitmann, U., and Okruss, M. (2006a) *Spectrochim. Acta, Part B*, **61**, 181–188.

Huang, M.D., Becker-Ross, H., Florek, S., Heitmann, U., and Okruss, M. (2006b) *Spectrochim. Acta, Part B*, **61**, 572–578.

Huang, M.D., Becker-Ross, H., Florek, S., Heitmann, U., and Okruss, M. (2006c) *Spectrochim. Acta, Part B*, **61**, 959–964.

Huang, M.D., Becker-Ross, H., Florek, S., Heitmann, U., and Okruss, M. (2008) *Spectrochim. Acta, Part B*, **63**, 566–570.

Ingle Jr, J.D. and Crouch, S.R. (1988) *Spectrochemical Analysis*, Prentice Hall, New Jersey, USA.

IUPAC. (1978) *Spectrochim. Acta, Part B*, **33**, 247–269.

Jakubowski, N., Dorka, R., Steers, E., and Tempez, A. (2007) *J. Anal. At. Spectrom.*, **22**, 722.

Jenkins, R. (1988) *X-Ray Fluorescence Spectrometry*, Wiey-Interscience, New York.

Jenkins, R. and de Vries, J.L. (1970) *Practical X-Ray Spectrometry*, MacMillan, London.

Johansson, S.A.E. and Campbell, J.L. (1988) *P.I.X.E. – A Novel Technique for Elemental Analysis*, John Wiley & Sons, Ltd, Chichester.

Jones, W.G. and Walsh, A. (1960) *Spectrochim. Acta*, **16**, 249–254.

Kennedy, P.K., Hammer, D.K., and Rockwell, B.H. (1997) *Prog. Quant. Electron.*, **21**, 155.

Khan, M.R. and Crumpton, D. (1981) *CRC Crit. Rev. Anal. Chem*, **11**, 103.

Kim, T., Nguyen, B.T., Minassian, V., and Lin, C.T. (2007) *J. Coat. Technol. Res.*, **4**, 241.

Kirchhoff, G.R. (1860) *Philos. Mag.*, **20**, 1.

Klockenkämper, R. and von Bohlen, A. (1999) *J. Anal. At. Spectrom.*, **14**, 571.

Knoth, J. and Schwenke, H. (1978) *Fresen. Z. Anal. Chem.*, **291**, 200.

Koirtyohann, S.R. and Pickett, E.E. (1965) *Anal. Chem.*, **37**, 601–603.

Lachance, G.R. and Claisse, F. (1994) *Quantitative X-Ray Fluorescence Analysis – Theory and Application*, John Wiley & Sons, Ltd, Chichester.

Laqua, K., Schrader, B., Hoffmann, G.G., Moore, D.S., and Vo-Dinh, T. (1995) *Pure Appl. Chem.*, **67**, 1745–1760.

Lundegårdh, H. (1936) *Lantbruks-Hogskol. Ann.*, **3**, 49.

L'vov, B.V. (1961) *Spectrochim. Acta*, **17**, 761.

L'vov, B. (1978) *Spectrochim. Acta, Part B*, **33**, 153–193.

Mages, M., Woelfl, S., Óvári, M., and Jun, W.V.T. (2003) *Spectrochim. Acta, Part B*, **58**, 2129–2138.

Manning, D.C. and Vollmer, J. (1967) *At. Absorp. Newsl.*, **6**, 38–41.

Mascart, E. (1864) *Compt. Rend.*, **58**, 1111.

Massmann, H. (1968) *Spectrochim. Acta, Part B*, **23**, 215–226.

Mika, J. and Török, T. (1974) *Analytical Emission Spectroscopy Fundamentals*, Butterworths, London.

Mohamed, W.T.Y. (2008) *Opt. Laser Technol.*, **40**, 30–38.

Montaser, A. and Golightly, D.W. (1992) *Inductively Coupled Plasmas in Analytical Atomic Spectrometry*, Wiley-VCH Verlag GmbH, New York.

Müller, R.O. (1972) *Spectrochemical Analysis by X-Ray*, Plenum Press, New York.

Neelmeijer, C., Brissaud, I., Calligaro, T., Demortier, G., Hautojärvi, A., Mäder, M., Martinot, L., Schreiner, M., Tuurnala, T., and Weber, G. (2000) *X-Ray Spectrom.*, **29**, 101.

Nichols, E.L. and Howes, H.L. (1924) *Phys. Rev.*, **23**, 472.

Nölte, J. (2003) *ICP Emission Spectrometry–A Practical Guide*, Wiley-VCH Verlag GmbH, Wienheim.

Pappalardo, L., de Sanoit, J., Marchetta, C., Pappalardo, G., Romano, F.P., and Rizzo, F. (2007) *X-Ray Spectrom.*, **36**, 310–315.

Paschen, A. (1916) *Ann. Phys.*, **50**, 901–940.

Pisonero, J., Fernández, B., Pereiro, R., Bordel, N., and Sanz-Medel, A. (2006) *TrAC, Trends Anal. Chem.*, **25**, 11.

Robinson, J.W. (1990) *Atomic Spectroscopy*, Marcel Dekker, New York.

Rosenkranz, B. and Bettmer, J. (2000) *TrAC, Trends Anal. Chem.*, **19**, 138.

Rusak, D.A., Castle, B.C., Smith, B.W., and Winefordner, J.D. (1998) *TrAC, Trends Anal. Chem.*, **17**, 453.

Russell, B.J., Shelton, J.P., and Walsh, A. (1957) *Spectrochim. Acta*, **8**, 317.

Russell, B.J. and Walsh, A. (1959) *Spectrochim. Acta*, **15**, 883–885.

Saisho, H. (1989) *Trends Anal. Chem.*, **8**, 209.

Sallé, B., Mauchien, P., and Maurice, S. (2007) *Spectrochim. Acta, Part B*, **62**, 739.

Schlemmer, G. and Welz, B. (1986) *Spectrochim. Acta, Part B*, **41**, 1157–1165.

Silva, M.M., Krug, F.J., Oliveira, P.V., Nóbrega, J.A., Reis, B.F., and Penteado, D.A.G. (1996) *Spectrochim. Acta, Part B*, **51**, 1925–1934.

Skoog, D.A., Holler, F.J., and Niemann, T.A. (1998) *Principles of Instrumental Analysis*, Harcourt College Publishers, Philadelphia, USA.

Slavin, W., Manning, D.C., and Carnrick, G.R. (1981) *At. Spectrosc.*, **2**, 137–145.

Sommerfeld, A. (1916) *Ann. Phys.*, **51**, 1.

Stchur, R., Yang, K.X., Hou, X., Sun, T., and Michel, R.G. (2001) *Spectrochim. Acta, Part B*, **56**, 1565.

Sturgeon, R.E., Guo, X., and Mester, Z. (2005) *Anal. Bioanal. Chem.*, **382**, 881–883.

Sychra, V., Svoboda, V., and Rubeška, I. (1975) *Atomic Fluorescence Spectroscopy*, Van Nostrand Reinhold, London.

Talbot, W.H.F. (1826) *Brewster's J. Sci.*, **5**, 77.

Tsalev, D.L., Slaveykova, V.I., and Mandjukov, P.B. (1990) *Spectrochim. Acta Rev.*, **13**, 225–274.

Veillon, C., Mansfield, J.M., Parsons, M.L., and Winefordner, J.D. (1966) *Anal. Chem.*, **38**, 204.

Walsh, A. (1955) *Spectrochim. Acta*, **7**, 108–117.

Welz, B., Becker-Ross, H., Florek, S., and Heitmann, U. (2005) *High-Resolution Continuum Source AAS – The Better Way to Do Atomic Absorption Spectrometry*, Wiley-VCH Verlag GmbH, Weinheim.

Welz, B., Becker-Ross, H., Florek, S., Heitmann, U., and Vale, M.G.R. (2003) *J. Brazil. Chem. Soc.*, **14**, 220–229.

Welz, B., Borges, D.L.G., Lepri, F.G., Vale, M.G.R., and Heitmann, U. (2007a) *Spectrochim. Acta, Part B*, **62**, 873–883.

Welz, B. and Melcher, M. (1984a) *Analyst*, **109**, 569–572.

Welz, B. and Melcher, M. (1984b) *Analyst*, **109**, 573–575.

Welz, B. and Melcher, M. (1984c) *Analyst*, **109**, 577–579.

Welz, B., Schlemmer, G., and Mudakavi, J.R. (1992) *J. Anal. At. Spectrom.*, **7**, 1257–1271.

Welz, B. and Schubert-Jacobs, M. (1986) *J. Anal. At. Spectrom.*, **1**, 23–27.

Welz, B. and Sperling, M. (1999) *Atomic Absorption Spectrometry*, Wiley-VCH Verlag GmbH, Weiheim.

Welz, B., Vale, M.G.R., Borges, D.L.G., and Heitmann, U. (2007b) *Anal. Bioanal. Chem.*, **389**, 2085–2095.

Wendt, R.H. and Fassel, V.A. (1965) *Anal. Chem.*, **37**, 920.

West, T.S. (1967) *Endeavour*, **26**, 44.

Willis, J.B. (1965) *Nature*, **207**, 715.

Winefordner, J.D. and Mansfield, J.M. (1967) *Appl. Spectrosc. Rev.*, **1**, 1.

Winefordner, J.D. and Vickers, T.J. (1964) *Anal. Chem.*, **36**, 161.

Wobrauschek, P. (2007) *X-Ray Spectrom.*, **36**, 289.

Wollaston, W.H. (1802) *Philos. Trans. R. Soc. London*, **92**, 365.

Wood, R.W. (1902) *Philos. Mag.*, **3**, 128.

Wood, R.W. (1905) *Philos. Mag.*, **10**, 513.

Woodson, T.T. (1939) *Rev. Sci. Instrum.*, **10**, 308.

Wormhoudt, J., Iannarilli Jr, F.J., Jones, S., Annen, K.D., and Freedman, A. (2005) *Appl. Spectrosc.*, **59**, 1098–1102.

Yoneda, Y. and Horiuchi, T. (1971) *Rev. Sci. Instrum.*, **42**, 1069.

Yu, Y.-L., Du, Z., and Wang, J.-H. (2007) *J. Anal. At. Spectrom.*, **22**, 650–656.

Zarkadas, C. and Karydas, A.G. (2004) *Spectrochim. Acta, Part B*, **59**, 1611–1618.

Zhou, Z., Zhou, K., and Hou, X. (2005) *Appl. Spectrosc. Rev.*, **40**, 165.

Zybin, A., Koch, J., Wizemann, H.D., Franzke, J., Niemax, K. (2005) *Spectrochim. Acta, Part B*, **60**, 1–11 .

ATOMS

BENJAMIN BEDERSON, *Physics Department, New York University, New York, New York, U.S.A.*

	Introduction	285
1.	**Charge and Mass**	286
2.	**The Periodic Table**	286
3.	**Atomic States and Energy Levels**	288
4.	**Ionized Atoms**	289
5.	**Atomic Wave Functions**	289
6.	**Negative Ions**	290
7.	**Lifetimes, Transition Probabilities, and Matrix Elements**	291
8.	**Metastable States**	291
9.	**Atomic Properties**	292
9.1	Magnetic Moments	292
9.2	Magnetic Resonance	292
9.3	Hyperfine Structure	293
9.4	Atomic Parameters	293
9.4.1	Polarizability	293
9.4.2	Diamagnetic and Paramagnetic Susceptibilities	294
10.	**Atomic Signatures**	295
11.	**Summary**	296
	Further Reading	296

INTRODUCTION

The atom can be considered to be the basic building block of nature at all levels of matter above the nuclear. The intrinsic properties of the isolated atom in large measure determine the more complex properties of bulk matter, in whatever form this matter may take. It is therefore often the case that detailed understanding of atomic structure and other atomic properties is necessary in order to understand such matter in any state, i.e., ionized, neutral gas, or condensed matter. A basic goal of atomic physics is to obtain that understanding.

One could in principle completely categorize isolated atoms by their electronic and electromagnetic properties, that is, by a complete description of the dynamics of the orbiting atomic electrons and by the ways in which atoms interact with external particles and fields. The study of such properties belongs properly to what could be characterized as "fundamental" atomic physics. There is not a very large distance between this characterization and the many applied areas that depend upon it.

For the purposes of this article one can move a step away from the fundamental aspects to refer instead to the atomic properties that derive from atomic dynamics, that is, to those characteristics of atoms that are specific to a particular atomic species and state (these terms are discussed more fully below), and that reveal some important aspect of their structure. Such properties are generally studied by means of some selective technique, of which there are now many. Recent years have witnessed dramatic improvements in traditional laboratory instrumentation and in the addition of important new experimental techniques to the atomic physicist's bag of tricks. The principal such devices will be discussed in the following sections.

Even further removed from fundamental atomic concepts are the observables that are actually studied by atomic scientists—atomic *signatures*. Atomic signatures are closely related to, but not necessarily identical to, atomic properties. These latter are those characteristics of atoms that are determined from the application of the known laws of electricity and magnetism, applied to atoms through the rules of quantum mechanics (these are

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

briefly referred to in Sec. 4). For example, while atomic spectra can be considered to be signatures, atomic energy levels are atomic properties which are inferred from atomic spectra, but are not generally directly observable. In the discussion that follows, we start with a compendium of atomic properties, then proceed to atomic signatures, and then specialize our discussion to several particularly important applied atomic physics techniques.

The most useful weapon in the experimentalist's arsenal, utilized by physicists and chemists for well over a century, is optical spectroscopy. Even here recent advances in spectroscopic techniques, mainly attributable to lasers, have totally revolutionized the study of atomic physics, and of atomic properties in particular. This article will touch on such developments, although the reader is referred to the article on ATOMIC SPECTROSCOPY for a more complete discussion.

1. CHARGE AND MASS

The most important atomic (and chemical) property of an atom is the charge of its atomic nucleus, which determines the number of electrons that orbit it in the neutral atom and which plays the principal role in determining the dynamics of these electrons in their orbital motion around the nucleus. While the structure and abundances of nuclei are the domain of nuclear rather than atomic physics, once the nuclear charge is determined atomic properties follow. The nuclear charge is itself determined by the number of protons in the nucleus (the atomic number Z), while the total number of protons and neutrons, i.e., the number of nucleons (the atomic weight), is responsible for virtually all the atomic mass, except for the very small contribution of the orbiting electrons.

Atoms of a given element may possess different numbers of neutrons, such variants being referred to as *isotopes*. Thus, the familiar 92 atomic elements from hydrogen ($Z=1$) to uranium ($Z=92$) are constructed from the 92 naturally occurring members of the periodic table, with a number of transuranic elements thrown in for good measure. Apart from significant though small differences in their spectra, and of course their masses, isotopes of a given element possess very similar physical and chemical properties. (An important exception to this statement can occur because of the different statistics obeyed by even and odd isotopes, but these do not generally play a role in atomic, as opposed to condensed matter, physics.)

Because of the similarity in the chemical and physical properties of the isotopes of a given element, isotope separation is generally a nontrivial pursuit. As a consequence, isotope separation schemes tend to exploit nuclear rather than atomic isotopic properties. Despite this, while isotopic spectra of a given element are similar, their differences are significant enough so that lasers, with their intrinsically high spectroscopic resolution, can be fruitfully exploited in some instances to achieve isotope separation in useful quantities.

One of the greatest intellectual achievements of the twentieth century has been the rapid development of quantum mechanics in the 1920s. Quantum mechanics (*q.v.*) replaced Newtonian mechanics in describing the dynamical behavior of microscopic particles such as atoms. Among its earliest, dramatic successes was its ability to predict the chemical and physical properties of the atomic elements, starting with very few basic principles. From these the systemization of the *periodic table* followed. It is important to emphasize that were ordinary Newtonian mechanics to govern the dynamics laws of orbiting electrons in analogy to planetary motion, none of the ordered and generally predictable properties of the elements would exist.

(We do not attempt to discuss the subject of quantum mechanics in any detail here, as this would take us too far afield from our intent of discussing atoms and their properties, even though these are in fact determined in the first instance by the laws of quantum mechanics. The list of references contains the names of several excellent source books for thorough discussions of this subject.)

2. THE PERIODIC TABLE

The periodic table is a two-dimensional array of atoms arranged according to the number of electrons in the neutral atom. At first this systemization, by Mendeleev and others, was strictly empirical, based on the similar chemical and physical properties possessed by atoms when suitably arranged by columns and rows. Only after the advent of quantum mechanics and the subsequent understanding of the quantum many-body problem that resulted was the empirical basis of the periodic table replaced by a thorough understanding of this order, stemming from many-body electron orbit theory. It is sufficient for our purposes to note that electron orbits, unlike planetary motion, arrange themselves in shells as atomic number increases, with each shell being filled (or in some cases partially filled) before newly added electrons start to fill new shells. The number of electrons in the outermost unfilled shell or subshell determines the atomic *valence*, a quantity which can be positive or negative, depending on whether the last unfilled shell contains an excess or a deficiency of electrons. For example, the halogens (fluorine, chlorine, bromine, etc.) possess a valence of -1, since their last unfilled shell or subshell contains a deficiency of one electron to make it complete.

The simplest examples of shell structure are the column one and column eight elements. Column one, i.e., atoms possessing a single valence electron, consists of atomic hydrogen and the alkali metals, lithium, sodium, potassium, rubidium, cesium, and francium. Hydrogen, while having some characteristics in common with the alkali metals, is really unique in the periodic table. It is by far the most abundant element in the Universe, its chemical reactivity makes it a partner in a huge variety of chemical reactions, and it is a constituent of many simple molecules, including those of biological interest. And in addition to its ubiquity it is also of fundamental importance as the principal testing ground for theory, since it is the simplest of all atomic systems and therefore is the most amenable to calculation. Because of their single valence electron, column one atoms are chemically very active. Equally active are the halogens, fluorine, chlorine, bromine, iodine, and astatine. As noted above these lack one electron to make up a complete subshell. As a consequence, while the alkalis readily give up their valence electron, i.e., are easily ionized, the halogens have strong electron affinities, which chemists would refer to as being electronegative (see Sec. 7).

The rare-gas elements are obtained with the completion of the lowest nonzero angular momentum subshell for each principal quantum number, excepting helium, which fills the first shell. (This shell can only contain angular momentum states equal to zero.) These elements, helium, neon, argon, krypton, xenon, and radon, are normally chemically inert, and all occur in the gaseous state under normal conditions of temperature and pressure. In fact, in recent years it has been shown that these gases can be chemically reactive, under certain restrictive conditions. One consequence of this reactivity has been the development of a new class of lasers, the so-called "excimer" lasers, some of which rely on this reactivity to produce lasers with the useful characteristic of being broadband, i.e., capable of lasing over a relatively wide range of wavelengths.

The empirical ordering exhibited by the periodic table, including the shell structure hierarchy, follows from the approximate picture of ground-state electron configurations determined by solutions to the Schrödinger equation. According to these solutions each shell has associated with it a *principal* quantum number n. Convention designates the various shells as K, L, M, etc., in accordance with whether $n=1,2,3$, etc. The principal quantum number is the main determinant of the overall electron energy; it also corresponds to the number of nodes, i.e., the zeros in the radial part, of the electronic wave function. Within each shell energy states are classified into subshells determined by orbital and spin angular momenta and their projections along an atomic axis of symmetry. The lighter elements are constructed in a regular fashion by filling shells and subshells sequentially. For the heavier elements it can happen that outer shells or subshells are completed before all energy states with lower principal quantum number are filled, an example being the series of rare earth elements from lanthanum ($Z=57$) to hafnium ($Z=72$). Since the outer electron configurations for these elements are the same they possess similar chemical properties. The reader is referred to any standard book on modern physics for a complete description of the periodic table.

3. ATOMIC STATES AND ENERGY LEVELS

Without specifically stating so, the discussion in the above section assumed that the atomic properties which are systematized in the periodic table refer to the lowest energy state of the atom, called the "ground state." Left alone, unperturbed by external influence, this state corresponds to the electronic configurations for which the total electron energy of the system is at a minimum. Here is another dramatic difference between so-called "classical" mechanics, which would describe the electronic motion using Newton's laws under the action of electromagnetic forces, or the relativistic generalization of these, and quantum mechanics. Because accelerating charges radiate, Newtonian mechanics would mandate that all electronic orbital motion collapses the atom, that is, the atomic electrons would fall into the nucleus. That this does not happen is attributable to the fact that under specific conditions the rules of quantum mechanics permit so-called "stationary states," states in which, and only in which, orbiting electrons do not radiate (see ELECTRONIC STRUCTURE OF ATOMS AND MOLECULES).

The only truly stable stationary state is the ground state. However, above the ground state all atomic systems possess an infinite number of stationary excited states, which can be reached in a variety of ways, for example, by collisions with other atoms or molecules, by interactions with radiation fields, or simply by being placed in a high-temperature environment (which is another way of generating interactions with particles and fields). Apart from the ground state, which possesses an infinite lifetime, lifetimes of all other states are finite (see Sec. 9).

Generally these excited states are still bound states, that is, the atom remains electrically neutral. In addition, one or more of the atomic electrons may receive from an external source sufficient energy to remove it completely from the atom, which becomes ionized. Ionization phenomena are discussed in Sec. 5.

There is a distinction between atomic state and atomic energy level. It can happen that more than one state can possess the same energy, in the absence of some perturbing interaction. Such states are described as being *degenerate*, the degeneracy being removed by the interaction. While this distinction can be extremely important in a variety of phenomena, for the purposes of this article we can neglect this distinction between state and level. Just as is the periodic table, energy levels for a given atomic species are systematized. The arrangements of these levels in any atom are again governed by the laws of quantum mechanics, which mandates that these levels be characterized by a set of quantum numbers. No two atomic states can possess all the same such numbers—this is the famous "exclusion principle," without which all electrons in an atom would fall into the lowest orbit, and thus would have doomed our universe to an existence in which no interesting chemistry or biology could flourish. There are in general four quantum numbers for each involved electron which labels an atomic state. The most important one, from an energy point of view, is the *principal* quantum number, an integer usually designated by n, which plays the dominant role in determining the mean distance of the involved electron from the nucleus. All electrons in a given shell possess the same principal quantum number. A second quantum number, l, is related to the orbital angular momentum. l is also an integer; the orbital angular momentum is given by $\sqrt{l(l+1)}$, in units of $h/2\pi$, where h is Planck's constant, the unit of angular momentum in quantum mechanics. A third quantum number, m, corresponds to the projection of the angular momentum along an axis of symmetry of the atom, which is equal to $mh/2\pi$. The fourth quantum number is the (half integral) electronic spin projection along the same axis, equal to $\pm h/2\pi$. (Generally speaking each quantum number corresponds to a "degree of freedom" of a particle. For the electron there are three spatial degrees of freedom, plus one degree of freedom due to the intrinsic electron spin.) Because there is a magnetic interaction between the orbital and spin angular momenta for each electron's orbital motion and between electrons as well, various coupling schemes are possible; these determine both the methods of classification and the actual values of the electronic energy levels and spectra of the elements.

The role played by these quantum numbers in establishing the periodic table has already been discussed in Sec. 3. The excited atom's energy levels are determined by the ordering of these quantum numbers as they acquire higher values. As with the periodic table, the energy levels of an atom can be arranged in columns, each corresponding to different values of orbital angular momentum and electron spin projection, with each row corresponding to a different principal quantum number. A simple example is atomic lithium ($Z=3$)—an alkali atom which for most practical purposes is a one-electron atom, since the first two electrons form a closed shell, and therefore do not play an active role in most lithium interactions. The complete set of appropriate quantum numbers for each atomic electron is referred to as an "orbital," in analogy to the classical counterpart of electron motion.

4. IONIZED ATOMS

Ionized atoms play dominant roles in a wide variety of phenomena, many of which are discussed in detail in other articles of this Encyclopedia. The entire field of plasma physics deals with ionized gases, and ionized species dominate such disparate fields as gas lasers (see LASERS, GAS), gaseous discharges (see ELECTRIC DISCHARGES) including fluorescent light sources, upper atmospheric and solar atmospheric physics (see ATMOSPHERIC PHYSICS), flames, and combustion (*q.v.*). Many high-energy particle accelerators use ion beams as input; ion beams are also used in a wide variety of practical applications including ion implantation (*q.v.*) in solids and ion beam irradiation for biological and medical uses. Plasma etching (*q.v.*), one of the newest technologies for the manufacture of computer chips, also relies on gaseous discharges to generate ionic currents for deposition onto substrates. Mass spectroscopy (see SPECTROMETERS, MASS) almost universally requires ionic beams to measure the generally unique atomic signature of charge-to-mass ratio, the most powerful tool in the experimentalist's arsenal for assaying specific concentrations of particular atomic (and molecular) species (as opposed to optical spectroscopy, which, as discussed in Sec. 1, is the most powerful experimental tool for studying atomic structure—see ATOMIC SPECTROSCOPY).

The stage of ionization can be anything from single ionization (one electron removed) up to fully ionized (all electrons removed). Removal of the first electron is the easiest to accomplish with respect to the energy required. The "stablest" element is helium—as much like a tiny billiard ball as atoms get—and helium requires only about 24 eV (electronvolt) for single ionization, where 1 eV is the amount of kinetic energy acquired by an electron in accelerating through a potential difference of one volt. The removal of each successive electron is more difficult, i.e., requires more energy. Removing the second electron in helium, for example, requires about 54 eV. Removing the 92nd electron from atomic uranium requires more than a GeV!

5. ATOMIC WAVE FUNCTIONS

The most fundamental of all atomic properties is the atomic wave function. This quantity represents the solution of the Schrödinger equation, or its relativistic counterpart (the Dirac equation). In principle all properties of the isolated atom could be obtained were its wave functions precisely known, since it describes—as best as can be done quantum mechanically—the complete set of coordinates of the atomic electrons in both magnitude and phase, from which these properties can be derived. (Such a complete set is not solely confined to spatial coordinates; for example, the wave function also includes a description of electronic spin.) It is a fundamental property of quantum mechanics that the wave function cannot itself be observed directly. Mathematically the wave function can be expressed as a function ψ (psi) of position and spin coordinates of each of the N electrons in the atom, as a function of time. For an isolated atom in the absence of external forces, the wave function describes stationary states, for which the time dependence is periodic with a frequency that is proportional to the energy of the state. Although ψ, which is generally complex (representing magnitude and phase), is not observable, its absolute value squared represents the electron density in probabilistic terms. Thus the electron charge cloud (what one might imagine an atom to look like were one to observe it from a distance) is given by $\psi\psi^*$, the product of the wave function by its complex conjugate. "Pictures" of atoms showing these charge clouds have served as metaphors for atomic structure since the early days of quantum mechanics, although these are actually computed charge clouds rather than observed ones.

The only exactly solvable atomic wave equation is that of atomic hydrogen, since this is a two-body problem (electron + proton). Even helium is not exactly solvable since the Schrödinger equation is not separable in the coordinates of its two orbiting electrons. The beautiful diagrams that represent $\psi\psi^*$ for hydrogen derive from the fact that the angular portion of ψ consists of pure associated Legendre polynomials, while the principal quantum number n is represented simply by the number of nodes in the radial part of ψ.

Although helium and heavier atomic wave functions are not exactly solvable, modern computational techniques can generate these with considerable precision and reliability, although the going gets rougher as one progresses through the periodic table. One of the most important practical goals of computational physics is the generation of reliable wave functions for ground and excited states, and indeed there are available codes that can be employed by users for this purpose. How good are these? In the last analysis they must meet the test of agreement with experiment, that is, by their ability to predict measurable atomic properties—the atomic signatures of Sec. 11. Such comparisons show that computational techniques are very powerful, and are of course rapidly getting better.

Finally, one general remark: not only are atomic properties derivable from wave functions, they also serve as input data for more complex phenomena, including collisions between atomic systems and other particles, both nonreactive and reactive. In virtually all such processes computational methods require as input reliable wave functions. From this point of view, study of isolated atoms serves as an invaluable test of the reliability of atomic wave functions for use in such applications.

The reader can be referred to a variety of sources for the generation of appropriate wave-function algorithms, using various approximation techniques. The most useful of these employs variations on so-called Hartree–Fock models. Such models are based on the assumption that each electron moves in the field of the remaining electrons (and nucleus), with the entire ensemble being made self-consistent. For heavier atoms Hartree–Fock is augmented by inclusion of relativistic effects, whence the model is designated as Hartree–Fock–Dirac.

6. NEGATIVE IONS

Besides suffering a loss of one or more orbiting electrons, many atoms of the periodic table can exist in stable configurations with an extra electron. These are negative ions, which play such an important role in many gaseous phenomena, including the upper atmosphere. Most negative ions exist in the molecular rather than the atomic state. However, some of the most important elements in the periodic table possess negative ions. Perhaps the most important of these are atomic oxygen and hydrogen, which help govern the dynamics and the radiation properties of the upper atmosphere and the solar corona, respectively.

The properties of negative ions are quite different from those of positive ions. Positive ions are stable since the remaining electrons are even more tightly bound to the nucleus than in the neutral atom. Negative ions, on the other hand, tend to be loosely bound, since the electrons' attraction to the atom must rely on detailed atomic structure properties rather than on the simple Coulomb attraction of positive ions. Also, negative ions usually possess only a single energy level, rather than the infinity of levels possessed by positive ions. All these properties result in the negative ion being much less stable, with respect to ionization, than its positive counterpart. The degree of stability is governed by the negative ion's *binding energy*, or electron affinity, the energy required to remove the excess electron. This is the negative-ion analog to the ionization energy of the positive ion. Binding energies range from only a few millielectron volts for some atomic systems to, at most, a few electron volts. The largest atomic binding energies, as might be expected, are those of the halogens. (Some *molecular* systems possess even larger electron affinities, and are therefore very useful as sinks for free electrons, e.g., are excellent high-voltage insulators.) There are no known doubly charged negative ions, and as already noted, very few cases of systems possessing more than one bound negative-ion state.

7. LIFETIMES, TRANSITION PROBABILITIES, AND MATRIX ELEMENTS

All atomic states except the ground state possess finite lifetimes, even when totally unperturbed by external influence. The range of such lifetimes is immensely broad, ranging from fractions of nanoseconds to many years, depending upon the characteristics of the particular states involved. A more specific descriptor of energy level stability is the *transition probability*, describing the relative rate of an atom spontaneously decaying to another, lower state. Transition probabilities can be calculated by taking the quantum-mechanical matrix elements of the transitions, i.e., the moments of atomic wave functions over electron coordinates of states connected by the transition in question. They describe the rates at which transitions occur between energy levels that are connected by what would classically correspond to radiating multipole moments. The most common of such matrix elements is that of the electric dipole, the moment obtained by calculating the space average of the position vector of the radiating electron, taken over pairs of states that differ by a single unit of orbital angular momentum. This is the quantum analog of a classically oscillating electric dipole charge, which sets up a field similar to that of a radiating dipole antenna. (Magnetic dipole radiation can also occur, although with orders of magnitude lower probability.)

Higher-order electric and magnetic multipole moments also radiate, in analogy to higher-order classically oscillating multipole charges, though with transition probabilities that decrease with increasing multipole order. Accordingly atomic transition probabilities for electric quadrupole transition probabilities will also be orders of magnitude smaller than electric dipole probabilities.

Typical lifetimes of states which radiate through electric dipole transitions in the visible spectrum are in the range of 10^{-7}–10^{-8} s; inner-shell transitions generally lie in the uv or x-ray region, with correspondingly shorter lifetimes. It is for this reason that the design of uv and x-ray lasers is difficult, since population inversion—the normal requirement for laser operation—becomes increasingly more difficult as the lifetime of the upper state of the laser transition becomes shorter.

8. METASTABLE STATES

There exist significant numbers of excited states which possess extremely long lifetimes, that is, which are not connected to lower states by transition matrix elements of any significant strength. These states are *metastable*—they are not truly stable, since they lie above the ground state and can undergo transitions to lower states either by weakly connected spontaneous decay or through some external perturbation. Perhaps the most famous of all metastable states are the 2*s* states of atomic hydrogen ($n=2$, orbital angular momentum$=0$), which lie some 10.2 eV above the ground state, and which cannot decay since they possess the same orbital angular momentum, namely, zero, as the ground states (we refer to these states in the plural since while the *electronic* configurations of 1*s*, 2*s* are unique these are actually doublets because of the interaction of the electronic and nuclear magnetic moments). The unperturbed lifetimes of the 2*s* states are approximately $\frac{1}{7}$ sec.

Metastable states can occur for a variety of reasons, always involving a violation of appropriate, either strict or approximate, *selection rules*. Such rules categorize those changes of appropriate quantum numbers in a transition that are permitted by the rules of quantum theory. Important high-lying metastable states exist in the rare gases and in mercury—the latter state playing an important role in the efficiency of fluorescent mercury lamps. Many elements possess low-lying metastable states, wherein principal and orbital quantum numbers are the same as those of the ground state, but with electron spin orientations which forbid electric dipole transitions. Such metastable states, for example, in oxygen and nitrogen play a significant role in upper atmosphere dynamics.

9. ATOMIC PROPERTIES

In addition to atomic energy levels there are a variety of atomic properties that are derivable, in principle, from atomic wave functions and that play major roles in many practical situations. We briefly note the most important of these here and refer the reader to individual sections for more detailed discussions.

9.1 Magnetic Moments

The individual electron orbitals of atomic states which do not form complete shells or subshells generally possess net angular momenta. Likewise, the electron spins of unpaired electrons will also result in a net angular momentum. These several individual angular momenta, orbital and spin, will combine to form a net, overall angular momentum for the entire electron configuration of the particular atomic state. Precisely how this comes about depends in detail on the nature of the magnetic spin-orbit interaction—in light atoms, usually the unpaired electron spins first couple to form a net atomic spin S, and the orbital angular momenta couple to form a net orbital momentum of integral value (in units of Planck's constant) L, and the separate S and L momenta then couple to form the final atomic angular momentum J—the so-called L-S coupling scheme. In very heavy atoms, the coupling first occurs between l and s for each electron to form individual j's which then couple to form the final J. This is the j-j coupling scheme. Intermediate cases also occur.

In any case, both the intrinsic spin and the orbital momenta have magnetic fields associated with them, as is the case for any rotating charges. Further, these are *dipole* fields, that is, fields whose spatial dependence (external to the sources) are the same as those that would be produced by magnetic dipoles. The ratio of magnetic moment to the angular momentum that produces it is called the *gyromagnetic ratio*. The natural unit of magnetic moment for atomic systems is the *Bohr magneton*, which in terms of fundamental atomic constants is $eh/4\pi mc$, equal in SI units to 9.2723×10^{-24} J T^{-1}. In units of the Bohr magneton the magnetic moment of a p orbital ($l=1$) is 1. In the same units the spin angular momentum of the free electron is $\frac{1}{2}$, while the corresponding magnetic moment of the electron is approximately 1. Accordingly, while the gyromagnetic ratio for electron orbitals is 1, that of the electron spin is approximately 2—actually, 2.0023; quantum electrodynamic effects are responsible for the so-called "anomaly" in the g value.

Magnetic effects significantly alter atomic energy levels, generally on a scale of millielectron volts, and are referred to as *fine structure*. Such fine structure is also observable spectroscopically through the dependence of atomic spectral wavelength on external magnetic fields—the *Zeeman effect* (see ZEEMAN AND STARK EFFECTS). Such fields could be intentionally generated laboratory fields, or could be local fields resulting from environmental effects, which can therefore serve as a probe for the study of magnetic fields of astrophysical or geophysical interest—i.e., sunspots, aurorae, galactic fields.

9.2 Magnetic Resonance

These magnetic moments in the unperturbed atom are responsible for many important magnetic effects, perhaps the most important of these being ferromagnetism (*q.v.*) and paramagnetism (*q.v.*) in condensed matter. Electron paramagnetic resonance (*q.v.*) is a technique which relies on the absorption of radio frequency or microwave radiation to cause transitions among the Zeeman levels of a paramagnetic medium. This is a cousin to nuclear magnetic resonance (see MAGNETIC RESONANCE, NUCLEAR), which relies on the same principle to probe the local fields seen by precessing protons and other nuclei in condensed matter environments, notably those of biological and medical interest. Such resonance techniques can be used either as a fundamental tool to measure magnetic moments of atoms and nuclei in known magnetic fields, or, turned around, to employ the known magnetic moments of atoms and nuclei as magnetometers, either in instrumentation, or, as noted in the last paragraph, as a probe of natural environments (see MAGNETIC RESONANCE IMAGING).

The existence of atomic magnetic moments led to the development of magnetic resonance techniques, both in bulk matter and, notably, in atomic beam spectroscopy. This technique, which was pioneered by I. I. Rabi and his students at Columbia University in the 1930s and 1940s, relied in the first instance on the mechanical force which acts on a magnetic moment in the presence of a nonuniform magnetic field (a uniform field generates a *torque* on a magnetic dipole, but not a net force). This force was exploited even earlier by Stern and Gerlach to demonstrate the spatial quantization of atomic angular momentum. Rabi used this "Stern–Gerlach" force, generated by use of specially shaped magnetic polefaces, to deflect atomic beams. These generally consisted of alkali-metal atoms, which are easily produced in beams and which possess large magnetic moments in their ground states. When radio-frequency fields are applied that are in resonance with the precessing moments, the resulting torques cause the atoms to "flop," that is, change their spatial orientation. This in turn alters the mechanical force which acts on them past the radio-frequency region, causing a change in detector signal (see MAGNETIC RESONANCE SPECTROMETERS).

9.3 Hyperfine Structure

Atomic magnetic fields interact with, i.e., couple to, the far weaker magnetic moment of the atomic nuclei. This interaction energy results in significant modification of atomic energy levels, called *hyperfine structure.* Hyperfine structure is on a scale of microelectronvolts, as a general rule, rather than millielectronvolts (fine structure) and electronvolts (electronic energy levels). Equivalent interactions between the atomic electric quadrupole moment and the nuclear quadrupole moment result in additional energy level modifications (even higher-order multipole moment interactions occur, and can be observed in precision measurements).

The Rabi method of studying hyperfine structure via atomic beam magnetic resonance revolutionized our knowledge of the static properties of the atomic nucleus. Such a technique, which directly observes hyperfine transitions, is far more sensitive to hyperfine energy levels than is normal optical spectroscopy, which has to rely on very precise measurements of small differences in optical wavelengths to observe the same interaction. Only with the advent of high resolution laser sources has hyperfine spectroscopy using optical techniques become competitive with magnetic resonance spectroscopy (see MAGNETIC RESONANCES, DETERMINATION OF STRUCTURES BY). Again, as happens constantly throughout experimental physics, the method can be, and has been, turned around to be used as a primary frequency standard—the atomic clock. Commercial atomic clocks using cesium atomic beams are available; such standards are of immense value in a variety of aerospace applications.

Perhaps of even greater importance in the development of current technology, atomic beam magnetic resonance was the first application of the use of commercial (or homebuilt) tunable oscillators to the control of atomic oscillations. Subsequently, along with parallel developments in electron paramagnetic resonance techniques and electric resonance techniques for molecules in beams, the first maser was built, by Townes and Schawlow, in 1953. The maser (acronym for microwave amplification by the stimulation of electromagnetic radiation) concept led naturally to the development of the laser, named in analogy to the maser, in 1959 (see LASER PHYSICS).

9.4 Atomic Parameters

We summarize here a number of other important atomic properties related to atomic structure and derivable from atomic wave functions, either of single states or, through appropriate matrix elements, by moments that connect several such states. These play major roles in determining physical and optical properties of neutral gas-phase systems as well as of weakly ionized plasmas. (The behavior of strongly ionized plasmas is usually dominated by collective effects, although atomic properties even here turn out to be extremely important because of the radiation properties of nonhydrogenic impurities in high-temperature plasmas.)

9.4.1 Polarizability. The atomic polarizability describes the distortion of the atomic charge cloud caused by the application of an external electric field. The lowest-order distortion is the electric dipole polarizability, which represents the charge separation that occurs in the atom when it is placed in an external uniform electric field.

The dipole polarizability plays an important role in atomic collision processes, where the electric field is generated by the collision partners. For neutral gases the polarizability is responsible for an attractive force between neighboring atomic partners, called the *van der Waals interaction.* This results from the instantaneous electric dipole of an orbiting electron in one of the partners inducing a dipole in the other. The product of polarizabilities appears in the so-called "van der Waals constant" which determines the strength of this interaction. Polarizability also governs the dielectric constants of gases, which is attributable to the inducing of static or dynamic electric fields in bulk matter by an external field. It can control ion mobilities in gases. Generally the above manifestations of polarizability refer to static, or dc, applied electric fields. Dynamic polarizabilities refer to frequency-dependent induced moments. Dynamic polarizabilities determine the index of refraction for the transmission of radiation through gases.

Polarizabilities form two hierarchies in order of powers of the applied field, one related to departure from linearity of the induced moments with increasing electric field strength and the other related to the order of the multipole moment with the spatial dependence of the applied field. Thus, each induced multipole contains a series of atomic constants, each proportional to a power of the applied field, and likewise each power of the spatial dependence of the electric field has associated with it a polarizability multipole. Applied fields of what would be considered normal intensity before the advent of high-powered lasers are generally of strength insufficient to generate more than the first term in these expansions, namely, the electric dipole polarizability, the parameter that is generally tabulated in data compendia. However, high-intensity lasers can easily generate nonlinear effects in bulk media, and the higher-order moments which govern these are responsible for the various nonlinear optical gas-laser interactions, including harmonic generation, four-wave mixing, and phase conjugation (see OPTICS, NONLINEAR).

Polarizabilities can be calculated using appropriate atomic matrix elements, or directly from known wave functions using variational formulas. Conversely, experimental determinations of polarizabilities can be used as a sensitive test of accuracy of computed wave functions. Atomic states whose wave functions are spherically symmetric possess scalar polarizabilities, that is, the induced moments in an external field will always lie along the direction of the field. However, most atomic systems are not spherically symmetric, even in their ground states, and these will in general possess tensor polarizabilities, that is, the induced moments in an external field will not lie in the direction of the field. Such polarizabilities then are describable by two parameters for each member of the polarizability hierarchy, for example, the polarizabilities that describe the induced moments along the two symmetry axes of the asymmetric atomic charge cloud (atoms possess time-averaged cylindrical symmetry), or equivalently, an average polarizability and a polarizability anisotropy. Again just as with scalar polarizabilities these are in principle derivable from exact knowledge of the atomic wave functions or, equivalently, matrix elements, although the set of such matrix elements may require inclusion of contributions from the continuum—that is, unbound states above the ionization limit—and these are particularly difficult to compute.

9.4.2 Diamagnetic and Paramagnetic Susceptibilities. All atoms are *diamagnetic,* in the sense that externally applied magnetic fields will induce magnetic moments in the atom as a consequence of the orbital motion of the atomic electrons (see DIAMAGNETISM). The magnetic moment induced by diamagnetism generates, as a microscopic application of Lenz's law, its own magnetic field which is oriented opposite to the applied field, external to the atom. Atoms which possess intrinsic magnetic moments are, in addition, *paramagnetic,* that is, their moments will be spatially aligned in external magnetic fields according to the usual prescriptions for spatial quantization (see PARAMAGNETISM). For such atoms the paramagnetic effect will dominate over the generally much smaller diamagnetic effect. A bulk system consisting of paramagnetic atoms in the gaseous state will then possess a net magnetism determined by Maxwell–Boltzmann statistics. Such systems are also optically active. If a magnetic field is applied parallel to the direction of propagation of an incident polarized light beam, the plane of polarization is rotated, with the degree of rotation directly proportional to the strength of the magnetic field. This is the *Faraday effect* (see MAGNETOOPTICAL DEVICES). The proportionality constant is called the *Verdet constant.* The Faraday effect has a variety of applications in optical and waveguide devices (usually, however, using paramagnetic materials in the condensed phase). The Faraday effect can account for the rotation of polarization in large galactic clouds in the presence of galactic magnetic fields.

10. ATOMIC SIGNATURES

We discuss here the several diagnostic techniques that are used to identify, or detect, specific atomic systems and states. The prime diagnostic tool of all is, of course, atomic spectroscopy (*q.v.*). Since atomic energy levels are specific to individual elements and to isotopes of elements, spectroscopy can be used to identify these unambiguously and without great difficulty. Resonance fluorescence—stimulating the emission of atomic radiation by selective excitation using lasers—is of singular value in such identifications. However, there are other diagnostic methods that do not involve radiation, which add invaluable tools to the investigator interested in atomic identification.

The neutral, isolated atom is not an easy object to observe. Apart from its radiation signature its electric neutrality renders it almost impervious to observation. The relatively crude diagnostic of mass deposition—allowing atoms to deposit on a substrate and be weighed—can be, and has been, used. However, there are far more sensitive detection schemes available.

Apart from resonance fluorescence, sensitive atomic detection schemes rely upon the ionization of the atom first, followed by measurement of an electric current—a two-step process. Ionization is most easily accomplished by use of *surface ionization,* wherein an atom gives up its valence electron to a metal surface and evaporates as an ion. For this to be accomplished the atom must possess an ionization potential that is comparable to, and preferably less than, the work function of the metal. This condition is most readily satisfied by the alkali metals interacting with almost any refractory metal surface such as tungsten or platinum. Alkali metals can be detected with virtually 100% efficiency by this means. This work function can be increased by oxidation of the surface thereby significantly enhancing the number of elements that can be detected by surface ionization.

Equally efficient is ionization of high-lying metastable states, such as metastable hydrogen and the metastable states of the rare gases. Here the excess energy of the metastable state is converted into kinetic energy of the valence electron, which can be sufficient to ionize (the *Auger effect*—see AUGER SPECTROSCOPY). Both of these methods are restricted to relatively small classes of atoms, those with low ionization potentials or with high metastable energies. The most general method, albeit a brute force one, is the use of electron impact ionization. This method is generally used in mass spectrometry applications, and such devices are readily available commercially. It has the disadvantage of being relatively inefficient, in general ionizing only a fraction of 1% of the neutral atoms. Another brute force method, the use of *field emission* (see FIELD EMISSION AND FIELD IONIZATION), is very effective in ionizing atoms near conducting whiskers to which strong electric fields are applied. However, this method possesses limited applicability and is effective only over small areas. A more general and selective method is the use of *resonant ionization spectroscopy (RIS)* to ionize. Many schemes can be used for this purpose. An atom can be raised to a highly excited state by laser excitation, either in a one or two step process, with an external electric field then used to ionize. RIS efficiencies can approach and even exceed 100% (the same atom can be detected more than once!), and it is nondestructive, leaving the atom in the same state after as before detection.

11. SUMMARY

In this short article we have attempted to outline some of the principal characteristics of atomic systems and to link these to some important applications. The brief discussions herein presented cannot attempt to cover these in detail; the reader is referred to the many articles of this Encyclopedia that do treat them more comprehensively. Some general sources for seeking further information on the many topics lightly touched on here are listed in the references given below. Certain extremely important topics that belong properly within the study of atoms have barely been noted here, since they address so-called "fundamental," as opposed to "applied," topics. However, as has been mentioned more than once in this article, there is not a large separation between these two supposedly disparate areas. Yesterday's fundamental studies were the basis of today's practical applications, and this statement applies equally well for today's fundamental studies and tomorrow's applications.

The best source material that covers both of these aspects of atomic structure and properties is contained in the series of volumes generically entitled "Atomic Physics," which presents all the invited papers and selected contributed papers of the biannual conference series "International Conference on Atomic Physics" (ICAP). Included in recent volumes are papers on structure of very high-Z atoms, tests of nonconservation of parity in weak interactions by spectroscopic techniques, the development of very high precision time and frequency standards, energy levels and electron correlation effects in very highly excited (Rydberg) states, properties of such states in strong electric and magnetic fields, precision measurements of hyperfine and fine structure in atomic hydrogen, analogous studies in "exotic" atoms (see the article EXOTIC ATOMS AND MOLECULES), precision measurements of fundamental atomic constants (see CONSTANTS, FUNDAMENTAL), and precision tests of fundamental concepts in quantum mechanics using atomic physics techniques (i.e., Bell's inequality and the Einstein–Podolsky–Rosen "paradox"), to mention some of the important "hot topics" of current interest.

Further Reading

See Froese Fischer, C. (1977), *The Hartree–Fock Method For Atoms*, New York: Wiley, for discussion of atomic structure methods. General reference books for atomic structure and simple quantum mechanics: Weissbluth, M. (1978), *Atoms and Molecules*, New York: Academic (paperback edition); Bransden, H., Joachain, C. J. (1986), *Physics of Atoms and Molecules*, New York: Longman.

For applications to the important field of atomic collisions, see the five-volume series, Massey, H. S. W., McDaniel, E. W., Bederson, B. (Eds.) (1982), *Applied Atomic Collision Physics*, New York: Academic. Vol. 1: Atmospheric physics and chemistry; 2: Controlled fusion; 3: Gas lasers; 4: Condensed matter; 5: Special topics. The best general references for recent research in atomic physics, including many articles on atomic structure and properties, are the proceedings of the International Conference on Atomic Physics, held biannually. The latest such volume is Haroche, S., Gay, J. C., Grynberg, G. (Eds.) (1989), *Atomic Physics 11*, Singapore: World Scientific. Earlier volumes (10, 9) were also published by World Scientific; volumes 1–8 were published by Plenum.

For articles on currently active areas of research in atomic physics the reader is referred to the annual series, Bates, Sir D. R., Bederson, B. (Eds.) (1990), *Advances in Atomic, Molecular, and Optical Physics*, Vol. 28. New York: Academic.

Negative ions: The classic monograph on this topic is Massey, H. S. W. (1976), *Negative Ions*, 3rd ed. London: Cambridge University Press. See also Massey, H. S. W. (1979), in: Bates, Sir D. R., Bederson, B. (Eds.), *Advances in Atomic, Molecular, and Optical Physics*, Vol. 17. New York: Academic. The most recent summary of the state of negative-ion research is Bates, D. R. (1990), "Negative ions: Structure and spectra," in: Bates, Sir D. R., Bederson, B. (Eds.), *Advances in Atomic, Molecular, and Optical Physics*, Vol. 28. New York: Academic.

ATOMS AND MOLECULES, ELECTRONIC STRUCTURE OF

See ELECTRONIC STRUCTURE OF ATOMS AND MOLECULES

AUGER SPECTROSCOPY

REED R. CORDERMAN, *GE Aircraft Engines, Cincinnati, Ohio, U.S.A.*

	Introduction	298
1.	**The Auger Electron Process**	299
1.1	Auger Electron Current	299
1.2	Phenomenological Analysis of AES Data	301
1.3	Sensitivity Factors and Auger Intensities	302
2.	**Experimental Apparatus**	303
2.1	Vacuum System	303
2.1.1	General	303
2.1.2	Sample-Load-Lock and Ion-Gun Differential Pump	303
2.2	Electron Gun	305
2.2.1	Thermionic	305
2.2.2	Field Emission	305
2.3	Electron Kinetic-Energy Analyzer	305
2.3.1	Cylindrical Mirror Analyzer	305
2.3.2	Hemispherical Sector Analyzer	306
2.3.3	Electron-Gun–Electron-Kinetic-Energy Analyzer Orientation	307
2.4	Detectors	308
2.4.1	Electron Detectors	308
2.4.2	Additional Detectors	308
2.5	Ion Guns	308
2.5.1	Thermionic	308
2.5.2	Discharge	309
2.6	Sample Handling	309
2.6.1	Sample Stage	309
2.6.2	Sample Carousel	309
2.6.3	Sample Heating and Cooling	309
2.6.4	Sample Fracture	310
2.6.5	Sample Scribing	310
2.7	Data Acquisition System	311
2.7.1	AES Spectrometer Digital Control	311
2.7.2	SAM Data-Analysis Computer	311
2.7.3	Ancillary SAM Capabilities	311
3.	**Experimental Techniques and Data Analysis**	312
3.1	AES Experimental Techniques	312
3.1.1	Auger Data Acquisition	312
3.1.2	Auger Point Spectra	312
3.1.3	Auger Line Scans	312
3.1.4	Auger Mapping	312
3.2	AES Depth Profile Analysis	312
3.2.1	Sequential Versus Continuous Acquisition	312
3.2.2	Calibration of the Depth Scale	313
3.3	Depth Deconvolution Techniques	313
3.3.1	Factor Analysis	313
3.3.2	Depth Profiling of Multilayers—Logistic Functions	314
4.	**Application of AES to Select Systems**	314
4.1	Microelectronics	314
4.1.1	General	314
4.1.2	Microelectronic Device Diffusion Barriers	316
4.1.3	Semiconductor Surface Potential Distribution	316
4.2	Superconductors	316
4.3	Metallurgy	317
4.3.1	Grain-Boundary Segregation	317
4.3.2	Thin-Film Interdiffusion Studies	317
4.3.3	Historical Metallurgy	318
4.4	Other Applications	318
4.4.1	Corrosion	318
4.4.2	Insulating Samples and Mineralogy	318
4.4.3	Surface Extended Energy-Loss Fine Structure	319
4.4.4	Chemical Effects in AES	319
	Glossary	319
	Works Cited	320
	Further Reading	322

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

When a solid sample is bombarded with energetic electrons (1–10 keV), core electrons will be ejected from an energy level E_x for a depth of approximately 1–2 μm. The core hole E_x is then filled by an internal process in which electrons from energy level E_y fall into the core holes E_x. Energy balance is achieved either by x-ray photon emission (fluorescence) with energy approximately $E_y - E_x$ or by ejection from the solid sample of a third electron from energy level E_z. The third electron, called an Auger electron after Pierre Auger, who first observed such events in a cloud chamber (Auger, 1925), is thus ejected from the atom with an energy E_{Auger} given approximately by

$$E_{\text{Auger}} = E_z + E_y - E_x. \tag{1}$$

Because the energy levels E_z, E_y, and E_x are unique for each element, the Auger electron E_{Auger} will have a unique energy. Thus, if the Auger energy spectrum from 0 to $\approx$ 2 keV is measured, then all the elements present on the solid sample, except hydrogen and helium, may be identified (Seah and Briggs, 1983).

Auger electron spectroscopy (AES) is a surface-sensitive technique because of the intense inelastic scattering that occurs for electrons in the energy range 0 to $\approx$ 2 keV, illustrated schematically in Fig. 1. Only Auger electrons from the outermost $\approx$ 0.5–3 nm of a solid survive to be ejected and measured in the spectrum.

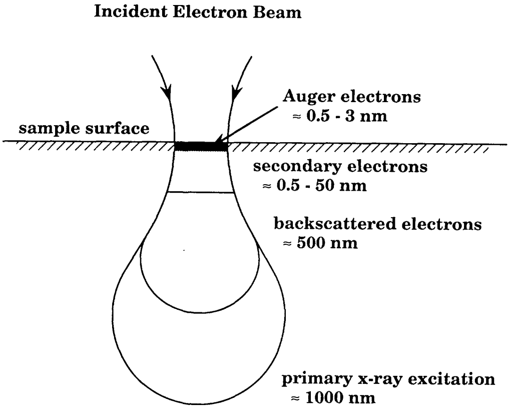
Incident Electron Beam
Auger electrons ≈ 0.5 - 3 nm
sample surface
secondary electrons ≈ 0.5 - 50 nm
backscattered electrons ≈ 500 nm
primary x-ray excitation ≈ 1000 nm

FIG. 1. Schematic diagram comparing Auger electron escape depths with emission depths for backscattered electrons and x rays (after Joshi, 1986).

Harris (1968a, 1968b, 1974) at General Electric was the first to realize that the small Auger peaks, which occur on a much larger background of secondary electrons ejected from the sample surface, could be separated from the background through the use of ac modulation and detection techniques. He used potential modulation of the analyzer to obtain the differentiated Auger electron spectrum, $d[EN(E)]/dE$ vs E, which has been used ever since (potential modulation is described in Sec. 3.1.1).

The addition of electron-beam steering optics and a secondary electron detector enables the now modified Auger electron spectrometer to obtain both Auger and secondary electron images of the sample under study. The modified apparatus is known as a scanning Auger microprobe (SAM).* A SAM apparatus provides an ultrahigh-vacuum environment in which to obtain Auger spectra of localized areas (typically $\approx$ 25 nm–1 μm), a high-brightness tightly focused electron gun, a high-transmission electron energy analyzer, and a computer system to acquire and manipulate the resultant data. A SAM can also function as a medium-resolution scanning electron microscope (SEM). While questions of chemical composition are easily answered using the AES-SAM, structural- and chemical-state information is much more difficult to extract.

The present article will review both the physical principles involved in the Auger electron process and phenomenological approaches to the analysis of Auger electron spectra. A detailed discussion of the experimental apparatus, which then leads into a description of experimental techniques such as *in situ* sample fracture and depth profiling, will be presented. Finally, several applications of AES to problems in microelectronics, metallurgy, mineralogy, etc., will be briefly discussed. The example applications discussed here are selected for their breadth across several different fields.

The present article is not an exhaustive review of the application of AES to specific systems; the recent bibliographic reviews [Fulghum *et al.* (1989) and Turner (1990, 1988, 1986)] provide more complete coverage. The reader is encouraged to pursue specific information from the many excellent review articles [Powell and Seah (1990); Nebesny, Maschhoff, and Armstrong (1989); Joshi (1986); Seah (1984); Werner and Garten (1984); Morgan (1983); Shimizu (1983); Zinner (1983); Stein and Joshi (1981); Wild (1981); and Hondros and Stoddart (1980)] and from the additional references given here.

---

*In the present article, SAM will refer to scanning Auger microprobe, not to be confused with Scanning Acoustic Microscope.

To start, Table 1 presents a summary of both the capabilities and applicability of Auger electron spectroscopy to the study of solid surface composition (Joshi, 1986).

1. THE AUGER ELECTRON PROCESS

A schematic diagram of the Auger emission process in a solid is presented in Fig. 2. The Auger process is a three-electron process. The incident electron beam strikes the sample with energy E_p, which must obviously be greater than the binding energy for the K (or L, M, etc.) core level E_K (or E_L, E_M, etc.) and ionizes the K (L, M) core level, Fig. 2(b). Due to the variation of the ionization cross section with energy, the incident electron-beam energy E_p must be greater than approximately $5E_K$ for efficient core-level ionization to occur (Briggs and Rivière, 1983). An electronic transition from an outer (lower-energy) level, for example, the L_1 level, to the core ionization hole relaxes the ionized atom. As a result, the energy $E_K - E_{L_1}$ becomes available and produces either characteristic x-ray photon or electron emission, Fig. 2(b). The first process is that of x-ray fluorescence while the second is Auger electron emission. The energy $E_K - E_{L_1}$ can be given to another electron in the same (L_1 in the present example), or less energetic (M), energy level, and that electron is ejected. Thus, the energy of the ejected Auger electron is

$$E_{KL_1L_{2,3}} = E_K - E_{L_1} - E^*_{L_{2,3}}, \tag{2}$$

where E_i are now the binding energies of the ith atomic energy levels [cf Eq. (1)] and $E^*_{L_{2,3}}$ is starred because it represents the binding energy of the $L_{2,3}$ level in the presence of a hole in level L_1, and therefore has an energy different from $L_{2,3}$ in its ground state. A doubly ionized final state results, Fig. 2(c). Note that the zero of energy is assumed to be at the Fermi level, since it is normal in AES to measure binding energies with respect to the Fermi level.

The Auger energy expressed by Eq. (2) is a function only of atomic energy levels. For each element in the periodic table, there is a unique set of Auger energies since no two elements have the same atomic binding energies (Davis *et al.*, 1976; McGuire, 1979). Thus, analysis of Auger energies leads immediately to elemental identification.

For heavy elements, the number of energetically possible Auger transitions for a given primary electron energy can become increasingly large as the number of atomic energy levels increases. However, the Auger transition probabilities favor only a few of the many possible transitions, so that the problem of elemental identification is feasible.

1.1 Auger Electron Current

The derivation of the Auger electron current is presented following the analysis of Seah (1983a). A total ionization $\sigma_{AX}(E_p)$ is caused by incidence of the primary electron beam, with energy E_p and current I_0, upon the sample. Additional ionization of atomic energy levels is caused by backscattered energetic electrons. If the number of backscattered electrons with sufficient energy to ionize atomic energy level X is given by

$$\int_{E_{AX}}^{E_p} \sigma_{AX}(E) n(E) dE, \tag{3}$$

where $n(E)$ is the true backscattered electron spectrum per unit incident electron, then the total ionization due to both the primary electron beam and the backscattered energetic electron may be written as

$$\sigma_{AX}(E) = \sigma_{AX}(E_p)[1 + r_M(E_{AX}, E_p, \alpha)], \tag{4}$$

where the backscattering term r_M is dependent upon the matrix M in which the A atoms are embedded and the angle α between the surface normal and the incident electron beam. The backscattering term r_M typically varies between 0.1 and 1.5 for 5-keV electrons incident at 30° from the surface normal (Shimizu and Ichimura, 1981; Ichimura and Shimizu, 1981).

The Auger electron created by the ionization process has a probability e^{-1} of traveling

Table 1. Auger electron spectroscopy—summary of capabilities (after Joshi, 1986).

Use	Estimated Analysis Time
(1) 0–3-nm surface elemental analysis for all elements except H and He. (2) Depth vs composition profiling and thin-film analysis. (3) ⩾20-nm lateral resolution surface chemical analysis. (4) Grain-boundary and other interface analyses facilitated by sample fracture. (5) Identification of phases in cross sections.	(1) Survey spectra: ≈5 min for a complete scan from 0–2000 eV. (2) Selected peak analyses: ≈15–30 min depending upon the number of peaks. (3) Depth profiling: ≈30–300 min or longer. (4) Auger elemental imaging: ≈30–300 min or longer.
Limitations	Examples of Applications
(1) Quantitative detection sensitivity is from 0.1 to 1.0 at.% (2) Accuracy of quantitative analysis limited (±30%) when calculated using sensitivity factors. Better quantification (±10%) when standards that closely resemble the samples are available. (3) Insensitive to H and He. (4) Electron-beam charging may limit the analysis of insulating materials. (5) Electron-beam damage may limit analysis of organic and biological materials.	(1) In-depth compositional variation of thin films used for microelectronic applications; thin films and coatings used for metallurgical surface modifications. (2) Analysis of grain-boundary chemistry to determine the role of precipitation and solute segregation on mechanical properties, corrosion, and stress corrosion cracking. (3) Analysis of surface contamination of materials as a result of thin-film processing, corrosion, wear, and catalysis.
Samples	Capabilities of Related Techniques
(1) Form: Solids with low vapor pressure [$\leqslant 1.33 \times 10^{-6}$ Pa (10^{-8} Torr) at 25 °C]. (2) Size: Limited by specific instrument; $2 \times 1 \times 1$ cm is typical; powder particle $\geqslant 1$ μm diameter may be analyzed. (3) Topography: Flat surfaces are preferable; rough surfaces may be analyzed either in small areas (≈1 μm^2) or averaged over large areas (≈200 μm^2). (4) Preparation: Frequently none; samples must be free of fingerprints, oils, and other high-vapor-pressure materials.	(1) X-ray photoelectron spectroscopy: provides compositional and chemical binding-state information; less destructive of organic and biological samples. (2) Ion scattering spectroscopy: provides unambiguous surface composition (3) Secondary-ion mass spectrometry: part-per-billion elemental detection sensitivity; surface composition information; depth profiling capability; sensitivity for all elements, including H and He. (4) Electron probe: analysis to 1 μm depth and lateral resolution; quantitative and nondestructive. (5) Analytical electron microscopy: chemical analysis in conjunction with high-resolution transmission electron microscopy.

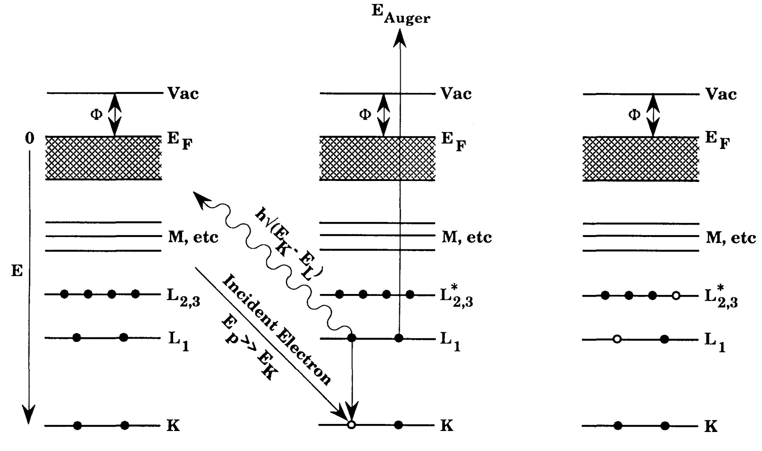

(a) Initial State (b) Excitation and Emission (c) Final State

FIG. 2. Schematic diagram of the Auger emission process in a solid (after Rivière, 1983).

a distance $\lambda_M(E_A)$ in the matrix material before being inelastically scattered and no longer appearing in the Auger peak at energy E_A. The distance $\lambda_M(E_A)$ is known as the inelastic mean free path. The flux of Auger electrons decays as $\exp[z/\lambda_M]$ as a function of distance z from the point of origin. The characteristic depth from which Auger electrons can be emitted is $\lambda_M(E_A)\cos\theta$ (commonly called the escape depth), where θ is the angle of emission relative to the surface normal. Values of λ are typically 2–10 atomic layers for Auger electrons with $E_A > 10$ eV, Fig. 3 (Seah and Dench, 1979), which accounts for the surface-sensitive nature of Auger electron spectroscopy [and x-ray photoemission spectroscopy (XPS) as well]. The curve shown in Fig. 3 is a compilation of experimental data derived from Auger transitions for many elements.

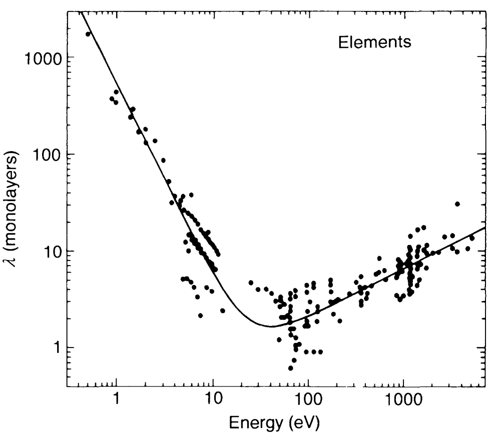

FIG. 3. The dependence of λ, the inelastic mean free path, on electron energy. The curve is a compilation of experimental data derived from Auger transitions for many elements (after Seah and Dench, 1979).

The emitted Auger electrons are detected by an electron kinetic-energy spectrometer with a transmission efficiency $T(E_A)$, and an electron detector of efficiency $D(E_A)$. Thus, the Auger electron current I_A may be written as

$$I_A = I_0\sigma_A(E_p)[1 + r_M(E_A,\alpha)]T(E_A)D(E_A) \times \int_0^\alpha N_A(z)\exp[-z/\lambda_M(E_A)\cos\theta]dz, \tag{5}$$

where $N_A(z)$ is the A-atom distribution with depth z into the sample surface. The current should be integrated over the appropriate angular entrance aperture of the electron spectrometer.

1.2 Phenomenological Analysis of AES Data

At the present time Eq. (5) is not used directly for the quantification of Auger data. Instead, Eq. (5) is used as a starting point and various simplifications are made (Hall and Morabito, 1979) to determine the mole fractions of the specimen constituents. The end result is a simplified expression which is generally applicable to the analysis of multicomponent solids.

Starting with the simplest multiconstituent solid, a homogeneous binary system, *AB*, the number of unknowns in Eq. (5) can be reduced by considering the ratio of intensities I_A/I_B for *AB* and the ratio of intensities for the pure element standards I^∞_A/I^∞_B, recorded on the same AES instrument,

$$\frac{X_A}{X_B}=F^A_{AB}\frac{I_A/I^\infty_A}{I_B/I^\infty_B}, \tag{6}$$

where X_A and X_B are the mole fractions of *A* and *B*, respectively, in the solid *AB*, and the Auger electron matrix factors F^A_{AB} are given by

$$F^A_{AB}(X_A\rightarrow 0)=\left(\frac{1+r_A(E_A)}{1+r_B(E_A)}\right)\left(\frac{a_B}{a_A}\right)^{1.5}, \tag{7}$$

$$F^A_{AB}(X_A\rightarrow 1)=\left(\frac{1+r_A(E_B)}{1+r_B(E_B)}\right)\left(\frac{a_B}{a_A}\right)^{1.5}, \tag{8}$$

where r_A and r_B are the backscattering factors for *A* and *B* in *AB* [see Eq. (4) above] and a_M is the atom size of *A* derived from $1000\rho_M N a^3_M = A_M$ where ρ_M is the density, *N* is Avogadro's number, and A_M is the mean atomic weight of the matrix. The ratio r_A/r_B is relatively energy independent (Shimizu and Ichimura, 1981), and so F^A_{AB} varies little across the composition range from *A* to *B* and may usefully be considered a constant for the *AB* system.

The above approach was originated by Hall and Morabito (1979), who calculated the matrix terms for the two extremes of Eqs. (7) and (8) for 4860 binary systems. In this way, they showed that in 86% of the binary systems

$$F^A_{AB}(X_A\rightarrow 0)=F^A_{AB}(X_A\rightarrow 1) \tag{9}$$

is valid to within 5%. If the matrix factors are ignored, then Eq. (6) reduces to

$$X_A=\frac{I_A/I^\infty_A}{\sum\limits_{i=1}^{n} I_n/I^\infty_n}. \tag{10}$$

1.3 Sensitivity Factors and Auger Intensities

In the absence of a large number of pure elemental standards, from which values for I^∞_A, I^∞_B, etc. may be obtained, the Auger signal from the sample may be normalized relative to both

1. the signals of all of the sample Auger peaks present in the spectrum, and
2. the signal of the pure sample element compared to the signal from a pure elemental target, typically silver.

The latter, shown in Fig. 4, are readily available in handbooks (e.g., Davis *et al.*, 1976). Thus,

$$X_A=\frac{I_A/S_A}{\sum\limits_{i=1}^{n} I_n/S_n}, \tag{11}$$

where I_A is the intensity of the Auger peak associated with element *A*, and S_A is the Auger

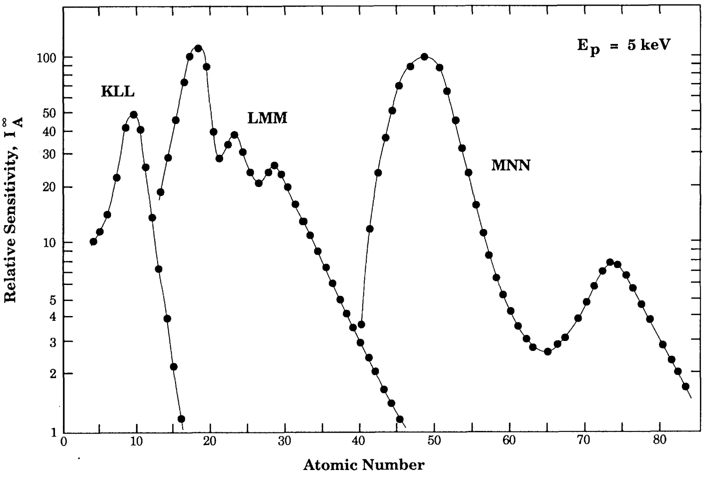

FIG. 4. Relative sensitivity factors for the elements from derivative AES spectra obtained at E_p=5 keV compiled by Davis *et al.* (1976).

sensitivity factor for element A at the particular E_p employed. The Auger sensitivity factors are used for I_A^∞, I_B^∞, etc., subject to the assumptions that

1. the peak shapes are the same in both the analysis and reference spectra,
2. the analyzers used for both spectra have the same resolution, and
3. the same modulation is used in both cases.

The choice of Auger intensities I_A and I_B is a matter of some debate (Nebesny *et al.*, 1989; Seah, 1983a). The use of peak-to-background (P/B) ratios in AES was first introduced as a means of correcting for small microtopographical variations on the surfaces of solid samples and has been assessed (El Gomati, 1988). More recently, Powell and Seah (1990) have critically reviewed the major steps in typical AES (and XPS) quantitative analyses and have commented on the major sources of error which lead to estimates of uncertainty in AES measurements.

2. EXPERIMENTAL APPARATUS

A photograph and a schematic diagram of a typical scanning Auger microprobe (SAM) are presented in Fig. 5. The stainless-steel ultrahigh vacuum chamber, which contains the electron gun and electron kinetic-energy analyzer, is visible on the left-hand side in Fig. 5(a), while the data-acquisition and analysis computer with associated displays and printer is seen on the right. The instrument control electronics are below. Each of the major subsystems of the instrument is discussed in greater detail below.

2.1 Vacuum System

2.1.1 General. Auger spectra are acquired in an ultrahigh-vacuum (UHV) environment. There are two main reasons for this. First, the Auger electrons, of energy $\approx$10–4000 eV, must not be scattered by the ambient gas to a degree sufficient to cause signal attenuation. The condition that the electron mean free path be much greater than the physical dimensions of the Auger spectrometer is met by a vacuum in the 1.33×10^{-3} to 1.33×10^{-4} Pa (10^{-5} to 10^{-6} Torr) range. The second reason for an UHV environment is far more stringent; the rate of sample surface contamination, from whatever source, must be slow compared to the Auger analysis time. Since Auger spectroscopy samples a depth of only a few atomic layers, typically the first one or two layers, and sensitivities of $\approx$ 0.5 at. % are routine, Auger spectra are extremely susceptible to surface contamination effects, which therefore mandates UHV conditions. Base spectrometer pressures of 6.7×10^{-9} to 2.7×10^{-8} Pa (5×10^{-11} to 2×10^{-10} Torr) are typical.

The three types of vacuum pumps commonly found on Auger spectrometers are sputter ion, turbomolecular, and titanium sublimation pumps. Each type of pump has both advantages and disadvantages peculiar to its operation (Rivière, 1983) with an Auger electron spectrometer. In general, sputter ion pumps of capacity 100–200 L s^{-1} are employed as the main or holding UHV pump. Their operation is somewhat feeble with respect to their capacity to pump both noble gases (especially helium) and hydrogen. Turbomolecular pumps are generally used to pump spectrometers from atmospheric pressure to 1.33×10^{-3} to 1.33×10^{-4} Pa (10^{-5} to 10^{-6} Torr), where sputter ion pumps become effective, and to differentially pump sputter ion guns (see Sec. 2.5). For long-term use, however, questions of bearing lifetime and especially mechanical vibration, a serious impediment to operation at high magnifications (e.g., $> 10\,000\times$), plague turbomolecular pumps. Titanium sublimation pumps are typically used at pressures $\leqslant 1.33 \times 10^{-4}$ Pa (10^{-6} Torr) to getter hydrogen, hydrocarbons, and other reactive gases (i.e., oxygen and nitrogen-containing gases). Sublimation pumps reduce the gas ingestion load that the sputter ion pump is exposed to as UHV conditions are attained. Detailed information about the production and measurement of ultrahigh vacuum conditions is presented in the article VACUUM TECHNIQUES AND INSTRUMENTATION.

2.1.2 Sample-Load-Lock and Ion-Gun Differential Pump. Most Auger spectrometers typically utilize a small (20–50 L s^{-1}) turbomolecular pump for two purposes. First, the pump serves to evacuate a small, auxiliary chamber adjacent to the main vacuum chamber and employed as a sample load lock, shown schematically in Fig. 5(b). A sample on

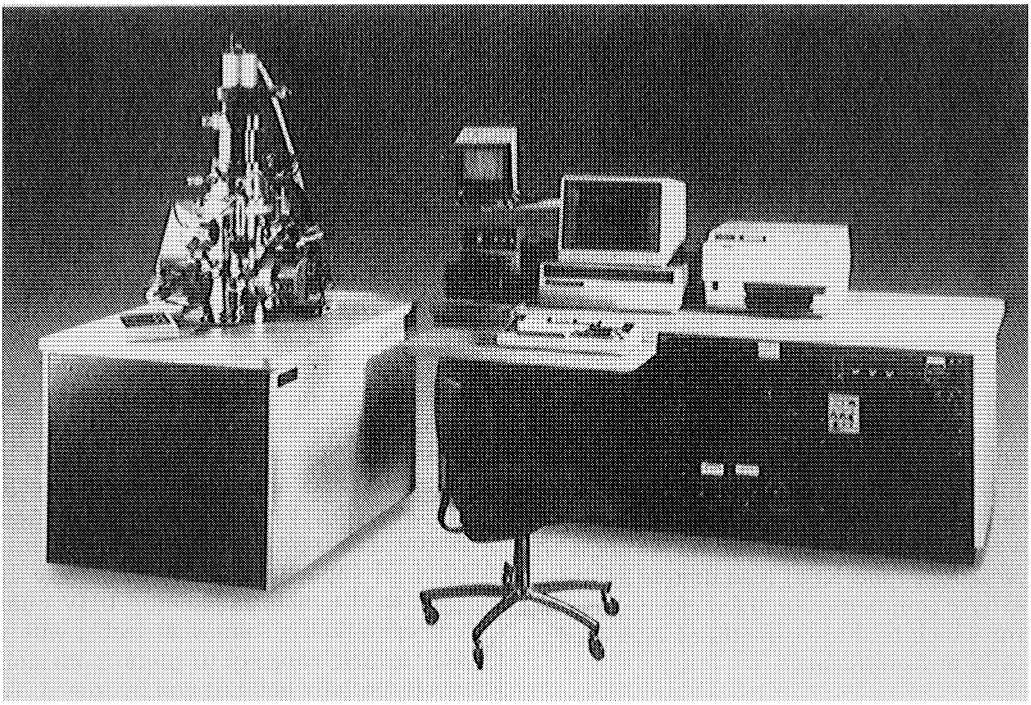

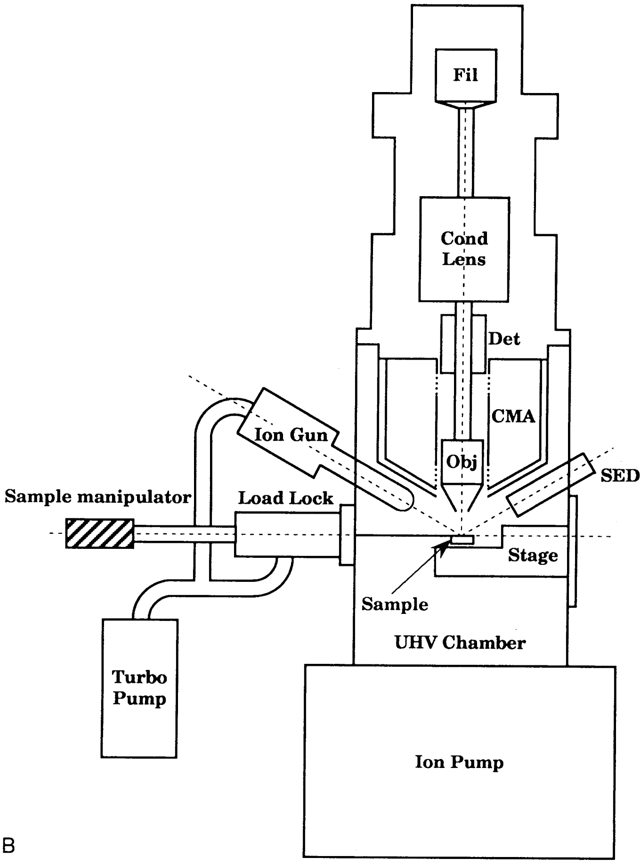

Fil
Cond Lens
Det
CMA
Ion Gun
Obj
SED
Sample manipulator
Load Lock
Stage
Sample
UHV Chamber
Turbo Pump
Ion Pump
B

FIG. 5. (a) A typical scanning Auger microprobe (SAM). The stainless-steel ultrahigh-vacuum chamber, which contains the electron gun and electron kinetic-energy analyzer, is visible on the left-hand side. The data-acquisition and analysis computer with associated displays and printer is seen on the right-hand side. The instrument control electronics are below. (b) Schematic diagram of a SAM. In the diagram, Fil is a LaB_6 filament, Cond Lens is a condenser lens, Det is an electron detector, CMA is a cylindrical mirror analyzer, Obj is an objective lens, SED is a secondary-electron detector, and Stage is a sample manipulation stage.

a carrier may be placed in the load lock, the load lock pumped to a relatively high vacuum 1.33×10^{-5} to 1.33×10^{-6} Pa (10^{-7} to 10^{-8} Torr), and the sample introduced into the spectrometer using a mechanical or magnetic linear motion feedthrough shaft. The load lock greatly increases the productivity of the Auger apparatus since it obviates the need to bake out and pump the system into the UHV range every time a sample must be changed. Thus, samples may be inserted into and removed from the AES spectrometer in minutes rather than days. The small turbomolecular pump also serves to pump differentially a sputter ion gun attached to the Auger apparatus for sputter depth profiling (see Sec. 2.5 below). The turbomolecular pump evacuates the sputter gas (typically argon) from the filament region of the sputter ion gun, thus relieving the gas load on the sputter ion pump.

2.2 Electron Gun

The electronic and mechanical stability of the electron gun is of utmost importance. The lack of stability, or drift, of the electron gun can severely limit the ability to perform Auger analysis of small samples ($\approx$ 0.1 μm) over long periods of time ($\approx$ 1 h) because the electron beam will drift away from the sample point of interest.

2.2.1 Thermionic. Thermionic emitters are typically employed in AES. Due to advances in design and reliability, filaments of lanthanum hexaboride (LaB_6), with its greater brightness ($\approx 10^6$ A cm^{-2} sr^{-1}) and lifetimes in excess of 500 h, are displacing those of tungsten (W) (brightness $\approx 10^4$ A cm^{-2} sr^{-1}) (Christou, 1976). Electron-beam diameters have decreased over 1000× during the ≈20 yr commercial development of Auger spectroscopy, as shown in Table 2. In purely electrostatic electron guns, spot size has decreased progressively from $\approx$ 400 to 0.5 μm, while a beam diameter of $\approx$ 0.1 μm is obtainable with electromagnetic guns prevalent today.

Table 2. Progress in AES beam diameter.

Date	Beam diameter (μm)	Beam current (nA)	Beam voltage (keV)
Early 1970s	400	2×10^4	2.5
Mid 1970s	40	5×10^3	5
Mid 1980s	0.5	10	10
Late 1980s	0.1	1	25

2.2.2 Field Emission. Field emission guns, which have even greater brightness ($\approx 10^7$ A cm^{-2} sr^{-1}) than LaB_6 thermionic emitters, have been employed only rarely in AES to date. However, field emission sources employed with electrostatic electron-gun optics exhibit beam diameters typical of thermionic guns with electromagnetic optics (Prutton *et al.*, 1982). Hence, as smaller and smaller beam diameters are required, the use of field emission guns is expected to increase.

2.3 Electron Kinetic-Energy Analyzer

The electron kinetic-energy analyzer is the heart of the Auger spectrometer. Because of the low electron kinetic energies involved ($\approx$ 10–4000 eV), and because magnetic fields are difficult to produce and control in UHV, all Auger spectrometers to date use electrostatic electron kinetic-energy analyzers exclusively. The two principal types are the cylindrical mirror analyzer (CMA) and the hemispherical sector analyzer (HSA).

Two different definitions of energy resolution are in current use. The absolute resolution ΔE is defined as the full width at half-maximum (FWHM) height of a particular Auger peak. The relative resolution is defined as $R=\Delta E/E_0$, where E_0 is the kinetic energy at the peak position. The Auger linewidth is independent of the linewidth of the incident electron beam. Therefore, the resolution requirements are determined by the linewidths of the Auger peaks themselves, which are typically several eV wide. In practice, it is desirable to have high sensitivity at high Auger energies ($>$ 1000 eV), where the production of Auger electrons is diminished due to branching into x-ray fluorescence relaxation channels, and high absolute resolution at low Auger energies ($<$ 150 eV). This is achieved by operation at constant relative resolution. For further information about electron kinetic-energy analyzers, see SPECTROMETERS, ELECTRON AND ION.

2.3.1 Cylindrical Mirror Analyzer. The cylindrical mirror analyzer is shown schematically in Fig. 6. Auger electrons enter at an angle α to the axis of cylindrical symmetry and pass between two coaxial cylinders of ra-

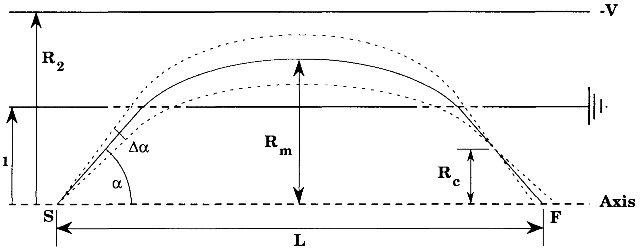

FIG. 6. Diagrammatic arrangement of a cylindrical mirror analyzer (CMA). R_c is the position of the minimum trace width [after Bishop, Coad, and Rivière (1972) and Rivière (1983)].

dii R_1 (inner) and R_2 (outer). The potential of the inner cylinder is at ground, while the potential of the outer cylinder is at $-V$. Electrons emitted from a source S on the axis with a particular energy E_0 are deflected by the outer cylinder potential through another aperture to focus on the axis at F. The relationship between V and E_0 is given by

$$\frac{E_0}{eV}=\frac{K}{\ln(R_2/R_1)}, \tag{12}$$

where K is a geometrical constant. For the special case of $\alpha = 42° 18'$, the CMA becomes a second-order focusing electron kinetic energy analyzer with $K = 1.31$ (Zashkvara *et al.*, 1966). It is desirable to achieve maximum sensitivity compatible with adequate resolution by operating the CMA with apertures that accept a range of angles about α, typically $\Delta\alpha=6°–10°$. For the angle $\alpha = 42° 18'$, the distance on the symmetry axis between the source and focus is $6.1R_1$; and the maximum distance off axis reached by the electrons, R_m, is given by

$$R_m=R_1\exp(1.3 \sin^2\alpha) \tag{13}$$

or, for $\alpha = 42° 18'$, $R_m= 1.8R_1$.

Scanning the potential $-V$ applied to the outer cylinder of the CMA gives the energy distribution of the electrons traversing through it. Because the transmission of the CMA varies as E_0, the energy distribution obtained is not $N(E)$, but rather $EN(E)$.

The half-width energy resolution for a CMA operated in the double-focusing mode, i.e., the entrance angle $\alpha = 42° 18'$, is given by

$$\frac{\Delta E}{E_0}=\frac{0.18w}{R_1}+1.39(\Delta\alpha)^3 \tag{14}$$

(Rivière, 1983) where w is the slit width (assumed to be equal at the entrance and exit), R_1 is the inner cylinder radius, and $\Delta\alpha$ is the angular aperture. For the CMA, the relatively wide angular apertures used (typically $\approx$ 6°) tend to limit the resolution obtainable. For example, a CMA built with an inner radius R_l of 3 cm and an angular aperture $\Delta\alpha$ of 6° will have a theoretical half-width resolution $\Delta E/E_0$ of 0.16% (1.6 eV for a 1000-eV Auger electron) for source sizes of 1–5 μm. Adjustable slits at the exit of the CMA are commonly employed to vary the energy resolution of commercial AES spectrometers over a broad range, e.g., from 0.3% to 2.0%.

The single-pass CMA is most effective in AES because of its very high transmission. This is the single most important property for an AES analyzer because the source area in AES using a highly focused electron beam is invariably much smaller than the acceptance area of the analyzer. With the small signal levels inherent in submicron Auger studies, high electron energy analyzer sensitivity is an absolute requirement. In this instance, the CMA is much preferred over the HSA because of the full 2π electron entrance cone and the high transmission ($\approx$ 10%) of the CMA compared to the HSA ($\approx$ 1%– 2%).

2.3.2 Hemispherical Sector Analyzer.

The hemispherical sector analyzer, also known as the concentric hemispherical analyzer, is more widely used in x-ray photoemission (or photoelectron) spectroscopy (see PHOTOEMISSION AND PHOTOELECTRON SPECTRA) than in Auger electron spectroscopy. The HSA is shown schematically in Fig. 7. Two hemispherical surfaces of inner radius R_1 and outer radius R_2 are positioned concentrically. Potentials V_1 and V_2 are applied to the inner and outer surfaces, respectively, and $V_2 > V_1$, so that a potential difference $\Delta V= V_2- V_1$ is established between the surfaces. While a continuous infinity of equipotential surfaces exists between the two surfaces, only the median equipotential surface is established with radius R_0. All other equipotential surfaces re-

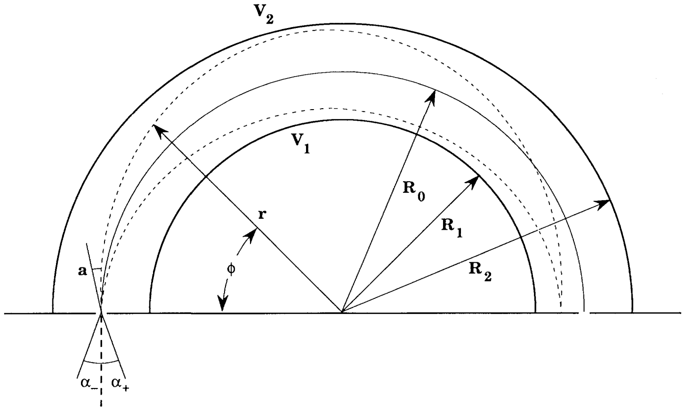

FIG. 7. Schematic diagram of a concentric hemispherical sector analyzer (HSA). If E is the kinetic energy of an electron traveling in an orbit of radius R_0, then the relationship between E and V is given by Eq. (16). ϕ and r are the angular and radial coordinates, respectively, of an electron of energy $E_0(=e\Delta V_0)$ entering the analyzer at an angle α to the slit normal (after Roy and Carette, 1971).

sult in radii greater or less than R_0. Entrance and exit slits are centered at radius R_0. In the ideal situation,

$$R_0=(R_1+R_2)/2. \tag{15}$$

The relationship between the deflecting potential ΔV and the ideal energy E for an electron trajectory that follows the equipotential R_0 is given by

$$e\Delta V=E(R_2/R_1-R_1/R_2). \tag{16}$$

The HSA, like the CMA, is double focusing. Early HSA Auger analyzers were operated by maintaining the potential along radius R_0 at ground potential and varying the specimen potential. Today, however, the situation is reversed—the specimen is grounded and the potentials of the analyzer hemispheres, now isolated from ground, are scanned.

The half-width energy resolution for an HSA operated in the double-focusing mode is given by

$$\Delta E/E_0=w/2R_0+\alpha^2/4, \tag{17}$$

(Rivière, 1983), where w is the slit width (assumed to be equal at the entrance and exit), R_0 is the interhemisphere equipotential cylinder radius, and α is the entrance angle into the HSA (equivalent to $\Delta\alpha$ of the CMA). For the HSA, it is desirable to use a large entrance angle for increased sensitivity, and it is common to use α so that $\alpha^2 \approx w/2R_0$. Thus, Eq. (17) reduces to

$$\Delta E/E_0=0.63w/R_0, \tag{18}$$

so that for a given R_0, there is an inverse relationship between the pass energy E and the slit width w for a certain absolute resolution. To illustrate, an HSA built with an interhemisphere equipotential cylinder radius R_0 of 10 cm and a pass energy E of 1000 eV will have a theoretical half-width resolution $\Delta E/E_0$ of 6.3×10^{-4} % (0.0063 eV for a 1000–eV Auger electron) for an entrance slit width of 1 μm. In practice, such high resolution is not obtained.

2.3.3 Electron-Gun–Electron-Kinetic-Energy Analyzer Orientation. There are two electron-gun–electron-kinetic-energy analyzer orientations in common use. In the first, the electron gun is coaxial with the electron kinetic-energy analyzer. This geometry has been used only with cylindrical mirror analyzers. Typically, the electron gun is located either inside the inner CMA cylinder or at a greater axial distance from the sample than the CMA.

Many of the samples to be analyzed by Auger spectroscopy are fractured *in situ* [i.e., within the scanning Auger microprobe (SAM) instrument]. This technique not only is the basis for many metallurgical investigations (corrosion, fatigue, etc.), but also provides an uncontaminated surface in an UHV environment for analysis. Because these fracture surfaces are quite rough (on a microscopic scale), an electron energy analyzer with a coaxial electron gun is useful because Auger electrons enter the CMA from the same sample area that is illuminated by the electron gun. More simply stated, the CMA sees no "shadowing" of Auger electrons emitted from a rough surface.

In the second geometry, the electron gun axis is located orthogonally (i.e. at 90°) to the electron kinetic-energy analyzer axis. In this case, severe shadowing can result, in which regions of the sample are shielded from the detector. Shadowing may limit the performance of the SAM during the analysis of fracture samples. The coaxial geometry offers the advantage of being able to analyze and map rough surfaces with much more signal collection than the orthogonal geometry.

2.4 Detectors

2.4.1 Electron Detectors. Multistage or continuous dynode electron multipliers (CDEM; also known as Channeltrons) are typically employed to detect Auger electrons. The CDEM may be biased to operate in either a pulse counting mode, for low counting rates (small-diameter electron beams), or an analog (continuous) mode, where the signal is large.

A relatively new development is the use of position sensitive detectors in electron spectrometry. The position-sensitive detector is constructed by combining a microchannel plate (MCP) assembly with a resistive anode. The MCP assembly may consist of a pair or triplet of plates, to obtain a peak gain of $\approx 5\times10^6$. The simultaneous detection of a range of energies reduces the measuring time required or, alternatively, increases the signal-to-noise ratio (van Oostrom, 1984).

2.4.2 Additional Detectors. Oftentimes, additional electron and x-ray detectors are added to the Auger electron spectrometer. These detectors permit several different types of analysis to be performed on the sample while it is in the spectrometer. Most frequently, a secondary electron detector (SED) is added to the Auger electron spectrometer. Addition of a secondary electron detector to the Auger spectrometer permits the Auger spectrometer to be operated as a medium-resolution secondary electron microscope (SEM), with great utility in locating specific features of the sample for further Auger analysis.

Less frequently, a semiconductor x-ray detector is added to the Auger electron spectrometer. This detector permits energy dispersive x-ray analysis (EDX) to be performed. Because x rays emanate from a balloon-sized region 1–2 μm in diameter, the number and identity of elements in the sample bulk may be rapidly identified. However, EDX analysis is not surface sensitive and may not reveal any information about the elements located at the sample surface.

2.5 Ion Guns

Practically all modern Auger electron spectrometers are equipped with a sputter ion gun, which allows for bombardment of the specimen surface by an energetic beam of rare-gas ions. Sputter ion guns are used both to clean and to erode the surface by energetic ion impact. Ion-bombardment cleaning is necessary to remove atmospheric gases chemisorbed and physisorbed on the specimen surface; sample fracture (Sec. 2.6.4) and surface scribing (Sec. 2.6.5) may also be employed. The sputter ion gun may also be used to obtain a compositional depth profile (i.e., the in-depth distribution analysis of chemical composition) of a specimen (Sec. 3.2). In this application, an impact or sputter crater forms on the specimen surface as energetic ion bombardment progresses. Both thermionic ionization and discharge sputter ion guns are in common use and are described in greater detail.

To obtain good depth resolution, the AES signal must originate from the flat portion of the sputter crater bottom (Zinner, 1983). Means to insure uniform erosion include defocusing the ion beam or rastering it over the sample surface. To exclude edge effects, only the central portion of the crater bottom is probed in AES analysis.

2.5.1 Thermionic. The simplest type of ion gun is similar in construction to an ionization gauge tube used to measure vacuum pressures. A hot-wire filament emits electrons into a positively biased wire-mesh grid cylinder, within which the electrons make several oscillations before being collected. A gas, typically argon, for sputtering the sample surface, is admitted to the gun at pressures between 1.33×10^{-2} and 1.33×10^{-3} Pa (10^{-4} and 10^{-5} Torr), and argon ions produced in the grid are accelerated by an extraction plate to a potential of $\approx$ 0.5–5 kV. This type of ion source can produce a maximum ion current density of $\approx 1\ \mu\text{A cm}^{-2}$ in a $\approx$ 200-μm spot at 5 cm from the gun exit aperture.

The advantages of this type of ion source are low cost, reliability, and ion-beam purity, since few energetic neutral atoms are produced in the ionizing region. A disadvantage of this type of ion gun is that either the AES spectrometer must be valved off during sputtering, since the ion pump must not be allowed to pump argon continuously, or differential pumping of the ion gun must be provided (Rivière, 1983).

2.5.2 Discharge. A second popular type of ion gun employs a gas discharge to produce ionization of the gas in the gun. Typically, the combined action of both electric and magnetic fields on argon at high pressure in the confined space of the ion-gun source produces a cold-cathode discharge from which the positive ions are extracted. Because the apertures in the ion gun are sufficiently small, differential pumping of the source is possible, allowing the pressure in the AES spectrometer to be maintained at a low level, e.g., 1.33×10^{-6} to 1.33×10^{-7} Pa (10^{-8} to 10^{-9} Torr), permitting Auger analysis to proceed during sputtering. More sophisticated ion guns of this type can produce an ion spot size of $\approx$ 0.1 mm diameter at 2–10 kV. This type of ion source can produce a maximum ion current density of $\approx 200\ \mu\text{A cm}^{-2}$.

In spite of the much greater current density of the discharge ion gun compared to the simple ionization ion gun, there are two disadvantages to this type of ion source. Ions inside the gun can strike and sputter other internal surfaces as well as being extracted into the ion beam. Therefore, arcing will occur when the accumulation of conductive material on insulating surfaces becomes sufficient to cause breakdown. Additionally, the ion beam is not pure and contains an appreciable proportion of energetic neutral atoms in addition to positive ions (Rivière, 1983).

2.6 Sample Handling

2.6.1 Sample Stage. The ability to manipulate the position of a sample in a precisely controlled manner after the sample has been inserted into the Auger spectrometer is essential. There are, in general, three reasons for this.

1. Because the shapes and sizes of specimens can vary widely, the optimum position for spectrometer energy resolution and sensitivity for one sample will not be the same as for the next sample.
2. It is often desirable either to move laterally or to rotate the sample during analysis in order to view and analyze a different region of the sample.
3. If other analysis techniques are available on the spectrometer, such as energy-dispersive x-ray analysis (EDX), it may be necessary to position the sample for analysis different from AES.

The considerations listed above suggest that at a minimum, 5 degrees of freedom in movement should be available. Typically, these involve the x, y, and z translation and both axial (z axis) and azimuthal (z tilt) rotation.

The temporal mechanical stability, or drift, of the sample stage or manipulator is of utmost importance. The drift of the sample manipulator (or the electron beam, see Sec. 2.2), can severely limit the ability to perform Auger analysis of small samples ($\approx 0.1\ \mu$m) over long periods of time ($\approx$ 1 h) because the sample stage will drift away from the electron beam.

2.6.2 Sample Carousel. A carousel-type sample holder, allowing multiple samples to be loaded into the vacuum chamber and stored under UHV conditions, is highly desirable in order to maximize the experimental throughput of the SAM. After loading several samples, the vacuum system can recover to its base pressure without further sample introductions [e.g., load samples in morning (or better, the previous evening) and analyze samples throughout the day].

2.6.3 Sample Heating and Cooling. Sample heating and cooling stages may also be useful. For heating stages which heat the entire sample and sample mount, temperatures are limited to typically 600 °C, with a long-term stability (several hours) of ± 2 °C. If the sample is thin, e.g., a foil or wire, then direct or indirect (i.e., filament) resistive heating of the sample may be used. However, the heating current should be kept as low as possible in order to reduce the effect of the magnetic fields produced on the low-energy Auger electrons. The heating stage should dissipate as little heat as possible to the Auger spectrometer, since radiational heating of other vacuum elements may result in increased outgassing, i.e., higher pressures in the Auger apparatus and possibly additional noise in the AES spectra.

Sample cooling stages are somewhat simpler in construction than heating stages and typically involve a thermally conductive pathway (copper block, heavy copper braided wire, etc.) between the sample stage and a liquid-nitrogen Dewar.

2.6.4 Sample Fracture. In the context of Auger spectroscopy, a clean surface is defined as the state of a surface in which AES does not detect any features other than those which constitute the bulk composition of the sample. Upon introduction of a sample into an AES spectrometer, initial observation will almost always reveal carbon, oxygen, and sometimes sulfur. Although ion bombardment is universally used for surface cleaning, either by itself or in combination with heating, ion bombardment may introduce artifacts, either chemical or topographical (or both), into the AES analysis of which the user should be aware. Preferential sputtering in alloys and compounds can lead to a change in the measured surface composition.

One method of producing a clean surface in an Auger spectrometer is by *in situ* fracture within the UHV environment. While many materials (metals, alloys) are too ductile at room temperature to cleave, fracture may occur at temperatures lower than the brittle-ductile transition temperature. The fracture path may then be transgranular (through individual grains) or intergranular (on grain boundaries); in either case, a surface free from contamination is produced. Note that the intergranular surface may contain impurities which have segregated from the bulk, which may then be further studied.

All practical fracture stages provide for liquid-nitrogen cooling of the sample to attempt to obtain a brittle fracture. Typically, the sample may be cooled to between 77 and ≈ 200 K. While many practical fracture stages use an impact hammer or chisel to fracture the sample, which places the sample in both tension and compression, tensile stages are available for doing quantitative fracture studies in the AES spectrometer UHV environment (e.g., from UHV Instruments, Inc., Buffalo, NY).

2.6.5 Sample Scribing. Another method of preparing a clean surface is by mechanical removal of the original surface and its associated contamination. A mechanism to scratch or scribe the surface, while not frequently found on commercial Auger electron spectrometers, can be a quite useful, simple, and inexpensive addition to the AES system. To study changes in the surface composition of alloys and compounds as a result of either ion bombardment or surface segregation, it is desirable to examine a clean surface that has a composition equal to the bulk composition. Two different methods have usually been employed to create such a surface: transgranular fracture (described above) and scribing (Betz, 1980; Betz, Marton, and Braun, 1980; Betz, Dudonis, and Braun, 1981).

The utility of sample scribing applies to three areas of Auger analysis: routine AES of samples without sputtering, AES of insulating thin films on conducting substrates, and direct measurement of preferential sputtering. By using a scriber to provide an atomically clean bulk surface rather than sputtering, the effects of preferential sputtering, in which different elements are removed from the surface at different rates, are avoided. Of secondary importance is potential contamination of the surface by low-level contaminants in the argon sputtering gas. This problem may be especially severe for highly reactive surfaces such as aluminum and titanium alloys. Auger analysis of insulating thin films on conducting substrates is often facilitated by scribing the surface, which provides a conducting pathway for the incident electrons to be dissipated from the sample.

Scribing may be used to provide a direct measurement of preferential sputtering. From the AES signal ratio for a scribed (or fractured) alloy AB surface, where R_{AB}^{bulk} is the ratio between the bulk concentrations of A and B, and $R_{AB}^{\text{sput sur}}$ is the ratio from the sputtered alloy AB surface of A and B, the component sputter yield ratio Y_A^c/Y_B^c can be obtained from

$$\frac{Y_A^c}{Y_B^c}=\frac{R_{AB}^{\text{bulk}}}{R_{AB}^{\text{sput sur}}}=\frac{c_A^{\text{bulk}}/c_B^{\text{bulk}}}{c_A^{\text{sput sur}}/c_B^{\text{sput sur}}}\,. \tag{19}$$

For the Ag–Pd system, nine alloys with $c_{\text{Ag}}=5\%$–95% were sputtered with 2-keV Ar ions under oblique incidence (75°) and also scribed and fractured (Betz, Marton, and Braun, 1980; Betz and Wehner, 1983). Surface enrichment of Pd with ion bombardment was observed using AES [Ag (351/355 eV) and Pd (326/330 eV) transitions]. The component sputtering yield ratio varied from 1.6 for Pd-rich alloys to 3.4 for Ag-rich alloys,

$$Y^c_{\mathrm{Ag}}/Y^c_{\mathrm{Pd}}=1.6 \text{ to } 3.4 \text{ as } c_{\mathrm{Ag}}=5\% \text{ to } 95\%. \tag{20}$$

2.7 Data Acquisition System

2.7.1 AES Spectrometer Digital Control. The AES electron optics may be operated under digital control to facilitate system automation, to simplify system operation, and to maximize the experimental throughput of the Auger spectrometer system. Digital control contributes ultimately to faster analyses and improved data accuracy and reproducibility. All electron-gun functions including beam parameters (beam voltage, emission current, condenser and objective lens voltages and currents), static beam positioning, scanning mode (i.e., multipoint/multiarea analyses, line scans, elemental maps, etc.), and scanning parameters may be controlled. The electron kinetic-energy analyzer and electron detector electronics and power supplies may also be digitally controlled. Predetermined data-acquisition parameters should be available in order to minimize setup time.

Auger spectra, $N(E)$ vs E, are typically converted to digital form as acquired. Once the digital Auger data have been collected, the data may be analyzed by various methods at a later time. Analog data, on the other hand, are subject to problems of uncertainty in peak position and peak area due to dependence upon instrumental time constants, modulation levels, and scanning speeds. If smoothing of the data is to be performed, it is much better to smooth digital data in a controlled manner rather than to rely upon the correct selection of scanning speed and time constant by the operator. One of the most severe disadvantages of analog data-collection methods is that it precludes data accumulation (i.e., signal averaging) to improve the signal-to-noise ratio (SNR) of a given Auger spectrum.

2.7.2 SAM Data-Analysis Computer. In addition to digital control of the AES spectrometer, additional computer software is required to analyze the data as it is acquired. The analysis computer also both serves to maximize the experimental throughput of the Auger spectrometer system and contributes ultimately to faster analyses and improved data accuracy and reproducibility. Some of the functions of the analysis computer are as follows:

1. spectral display and expansion;
2. smoothing;
3. derivative spectra;
4. peak integration and area measurement;
5. analysis of overlapping spectral features;
6. addition and subtraction of spectra;
7. peak maximum identification and location; and
8. noise spike removal.

Functions 1–3 are used in the vast majority of Auger spectra that are acquired. Function 4 is used quite frequently, while functions 5–8, though only occasionally required, are nonetheless desirable features of any data analysis package.

Associated with the AES data-analysis computer may be a full range of ancillary computer equipment such as color graphics display, color pen plotters, cameras, or archival storage media for the display, presentation, or storage of the digital AES data. For a review of computerized data acquisition and analysis, see DATA ACQUISITION SYSTEMS.

2.7.3 Ancillary SAM Capabilities. Typically the AES instrument provides a variety of ancillary signal processing features, which may be used occasionally either to calibrate and adjust the electron (and ion) optics or to obtain additional information regarding the experiment at hand. These measurements fall into three broad categories:

1. current measurement (absorbed by the sample, secondary electron, selected energy secondary electron, primary electron beam, and sputter ion beam);
2. secondary electron image manipulation (image polarity, contrast, brightness, offset, and gamma); and
3. specimen bias (positive/negative with respect to the instrument ground).

3. EXPERIMENTAL TECHNIQUES AND DATA ANALYSIS

3.1 AES Experimental Techniques

3.1.1 Auger Data Acquisition. As commonly practiced, the electron energy distribution, $N(E)$ vs E, is accumulated directly by pulse counting (for small primary electron currents, $I_p \leqslant 10^{-7}$ A) or current-to-digital conversion and measurement (for $I_p > 10^{-7}$ A). Plots of $dN(E)/dE$ vs E are produced by digital differentiation and reduce the effect of the high secondary electron background.

Older Auger spectrometers use an analog electronic method to generate $d[EN(E)]/dE$ vs E displays of Auger data. The electronic method involves imposition of a small alternating voltage on the outer cylinder of the CMA and synchronous detection of the in-phase component of the analyzer output using a phase-sensitive detector (i.e., lock-in amplifier). Because the transmission of a CMA varies as E_0, the electron kinetic energy, the distribution obtained as the outer cylinder potential V is swept is $EN(E)$, not $N(E)$. Modulation of V by a small sinusoidal signal with synchronous detection will extract $d[EN(E)]/dE$ vs E. The high secondary electron background will not be modulated and is thus eliminated from the resultant spectrum. Harris (1968a, 1968b, 1974) first used potential modulation of the analyzer to obtain differentiated Auger spectra.

Digital electronic control of the electron beam used in AES offers various choices in examining a given surface. The electron beam may be fixed on one particular point, or it may be rastered over many different points, lines, or small rectangular regions. A rastered electron beam can be used to obtain secondary electron images of the surface topography to complement AES information (Joshi, 1986).

3.1.2 Auger Point Spectra. Digital control of the electron gun permits Auger data from several different geometries and/or energy regions to be collected. Perhaps the most common are single or multiple Auger point spectra, in which the electron beam remains at one point, $I(x_0,y_0)$, on the sample for the duration of the Auger data acquisition. Similarly, Auger spectra at multiple data points, $I^n_{i=1}(x_i,y_i)$, where I_i is the Auger signal at one point and n is the number of points, may be acquired.

At each analysis location, the electron kinetic-energy analyzer may be scanned through the entire Auger kinetic energy range (typically $\approx$ 30–1000 or 2000 eV). Alternatively, several $\approx$ 20–30-eV energy windows may be recorded which include Auger transitions for only those elements of interest. For example, for GaAs, only energy windows at 1055–1085 eV and 1215–1245 eV, which bound the Ga and As *LMM* Auger transitions, respectively, need be recorded. A variation on single/multiple Auger point scanning is to acquire Auger data from single/multiple areas, typically rectangularly shaped with arbitrary width and length.

3.1.3 Auger Line Scans. Quite often Auger line scans are useful, especially when the sample is prepared such that a cross-sectional view of an interface is observable. Single (or multiple) vertical (or horizontal) lines, $I^n_{i=1}(x_0,y_i)$, of either single (or multiple) elements may be acquired. In this manner, a one-dimensional view of the change in elemental concentration across an interface may be constructed.

3.1.4 Auger Mapping. By rastering the electron beam across a sample, and at each raster location measuring the amount of an element present, $I^n_{i;\,j=1}(x_i,y_j)$, an image may be constructed in which the intensity of each raster point (pixel) represents the amount of the element present. Typically the Auger signal is collected for only one, two, or three energies at each (x_i,y_j) point. These energies include the peak of the Auger transition for those elements of interest and either one or two base-line energies from which to establish the background signal. In this manner, a two-dimensional view, or Auger map, of the change in elemental concentration across a surface may be constructed and is particularly useful during the early stages of analysis to distinguish significant from insignificant features of the surface.

3.2 AES Depth Profile Analysis

3.2.1 Sequential Versus Continuous Acquisition. The combination of an Auger electron spectrometer and an ion gun may be used to obtain a compositional depth profile (i.e., the in-depth distribution analysis of chemical composition) of a surface (Zinner, 1983). The depth profile may be obtained either sequentially or continuously. In the sequential technique, by alternately and repeatedly sputtering the surface, gating off the ion gun, and subsequently measuring the surface composition by Auger analysis, the chemical composition of the surface at each depth may be obtained. Alternatively, in the continuous technique, by continuous sputtering with simultaneous Auger analysis, the chemical composition of the surface may be obtained. Since the sputtering proceeds continuously, each depth, for each element, is slightly different than for each other element. Specimen rotation during sputtering has been shown to give improved interfacial resolution for multilayered samples (Zalar, 1985).

The data obtained during the depth profile experiment consist of Auger signal intensities of the detected elements, I_x, as functions of sputtering time, t. The data must then be converted to concentration c_x versus depth z data by an appropriate calibration of both the concentration and depth scales. The calibration of the concentration scale has been discussed above.

3.2.2 Calibration of the Depth Scale. The velocity of surface erosion is described by an instantaneous sputtering rate $\dot{z}=dz/dt$ which determines the mean eroded depth z as a function of sputtering time according to

$$z(t)=\int_0^t \dot{z}\,dt \qquad (21)$$

(Hofmann, 1983). The sputter rate $\dot{z}$ (m/s) is determined by

$$\dot{z}=(M/\rho N_A e)S j_P, \qquad (22)$$

where M is the atomic mass number; ρ is the density (kg m^{-3}); $N_A=6.02\times10^{23}$ (Avogadro's number); $e=1.6\times10^{-19}$ (electron charge); S is the sputtering yield (atoms/ion); and j_P is the primary-ion current density (A m^{-2}). For a constant sputtering rate, Eq. (21) shows a direct proportionality between sputtered depth and time, $z = \dot{z}\,t$. In this case the z axis (depth) is linear with the sputtering time and only one point is necessary to determine the calibration factor $\dot{z}$. From Eq. (22), $\dot{z}$ can be calculated by taking the sputtering rate S from literature data (Behrisch, 1981; Seah, 1981; and Oechsner, 1975) and measuring j_P with a Faraday cup.

A better method to obtain the sputtering rate $\dot{z}$ is to measure the time required to sputter through a layer of known thickness, e.g., an evaporated metallic layer or an oxide layer, for given ion-gun operation parameters. Anodized tantalum foils have proved useful for this purpose, since the thickness is easily controlled by the formation voltage. In addition, due to the sharp metal-oxide interface, oxidized tantalum foils provide a quick test method for the instrumental depth resolution. Thus, $\dot{z}$ is determined with knowledge of the Ta_2O_5 sputtering rate as compared to the material under study.

A third possibility is to measure the actual sputtered crater depth after depth profiling by conventional interferometry or stylus methods.

3.3 Depth Deconvolution Techniques

3.3.1 Factor Analysis. An old analytical technique, factor analysis (Malinowski and Howery, 1980), newly applied to Auger lineshape analysis of spectra selected from a compositional depth profile, can determine the presence and number of new chemical states formed at an interface (Gaarenstroom,1982). The procedure is superior to other methods of interpreting Auger line shape and depth profile data, such as visual inspection, spectral stripping, or peak area measurements. The range of information gained by employing factor analysis is broad: determination of the number of components, quantitative analysis from overlapping spectra, extraction of spectra for unidentified components, and speed of analysis ($\approx$ 20 times faster than spectral stripping and integration).

As an example, Fig. 8(a) shows a conventional Auger depth profile of Pt-zirconia, while Fig. 8(b) shows a chemical state depth profile of Pt-zirconia computed from Auger line-shape information. Spectra of the three individual components Pt, PtZr, and ZrO_2 are calculated by the factor analysis technique, and may be compared to spectra of reference materials in order to verify the chemical identity of the unknown (i.e., interfacial) component (PtZr in this example).

3.3.2 Depth Profiling of Multilayers—Logistic Functions. The composition versus depth distribution of a solid/solid interface as determined by Auger sputter depth profiling can be described by a logistic function of the form

$$Y=\frac{A+a(T-T_0)}{1+e^x}+\frac{B+b(T-T_0)}{1+e^{-x}}, \tag{23}$$

where Y is a measure of the surface elemental concentration of one of the components that defines the interface (Kirchhoff, Chambers, and Fine, 1986). T is a measure of the sputtering depth and is usually expressed as sputtering time assuming a constant sputtering rate. A is a measure of the preinterface surface concentration for one of the elements of the interface, and B is a measure of the postinterface surface concentration, where "pre" and "post" are taken in the sense of time. The empirical parameters a and b are introduced to account for additional time dependence of instrumental effects. T_0 is the apparent midpoint of the interface region, and x is a dimensionless, reduced time defined as $(T-T_0)/D$. The scaling factor D is the characteristic time for sputtering through the interface region. All measures of the interface width will be proportional to D. Asymmetry in the depth profile can be accommodated by allowing D to vary with time. A least-squares fitting program has been written which fits the above equation to the measured Auger spectral intensities to within the measurement error (approximately 1%). The statistics associated with the least-squares fit allow confidence limits to be placed on the measured widths of interface regions and on the asymmetry associated with each such region.

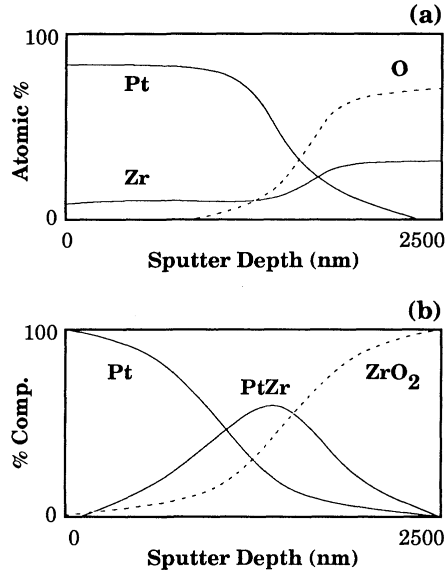

FIG. 8. (a) Conventional Auger depth profile of Pt-zirconia. (b) Chemical state depth profile of Pt-zirconia computed from Auger line-shape information (from Gaarenstroom, 1982).

4. APPLICATION OF AES TO SELECT SYSTEMS

A list of the applications of Auger electron spectroscopy to a variety of problems of current research interest is given in Table 3. Brief descriptions of each of these problems is presented in the following section.

4.1 Microelectronics

4.1.1 General. The role of analytical instrumentation is becoming increasingly critical as all aspects of microelectronic technology trend towards increasing circuit density. For a review of microelectronic devices and technology, see MICROELECTRONICS. Submicrometer feature sizes necessitate analytical techniques with comparable, or better, lateral and depth resolutions. A scanning Auger microprobe, because of its high lateral ($\approx$ 50–100 nm) and depth ($\approx$ 0.5 –3 nm) resolutions and inherent capability to function as a scanning electron microscope, is ideally suited to identify and correlate submicrometer features (SEM) with elemental analysis (AES) of these features.

By combining an Auger electron spectrometer with *in situ* ion-beam sputter etching, depth profiling on a very small scale [2-nm resolution in $\leqslant$ 100-nm depth (Helms, 1982)] may be performed. Several of the applications of scanning Auger microscopy (SAM) to microelectronics technology are as follows.

1. Wafer evaluation: AES may be used to determine that the wafer surface is free of contamination (polishing residues, etching residues, inclusions, particulates, films, surface haze, etc.).
2. Oxidation processing steps: oxide thick-

Table 3. Selected applications of Auger electron spectroscopy.

Application	Reference
(1) Microelectronics, including	Thomas and Joshi (1983); Helms (1982), Holloway (1980a, 1980b); Lowry and Hogrefe (1980); Olson *et al.* (1983); van Oostrom (1984)
(a) Depth vs composition profiling	
(b) Wafer evaluation	
(c) Oxidation processing	
(d) Wet chemical or plasma etching	
(e) Diffusion and ion implantation doping	
(f) Metallization	
(g) Packaging and bonding	
(h) Diffusion barriers	Grove *et al.* (1989); Huelsman and Fonstad (1985)
(i) Semiconductor surface potential distribution	Pantel *et al.* (1982); Pantel (1983)
(2) Grain-boundary segregation	Seah (1975, 1980, 1983b, 1983c); Thien *et al.* (1984)
(3) Superconductors	Suenaga and Jansen (1983); Kroeger, *et al.* (1988)
(4) Thin-film interdiffusion	Pamler (1987); Hall and Morabito (1976)
(5) Ti-6A1-4V powder for aerospace structural components	Wirth and Grundhoff (1981)
(6) Historical metallurgy	Mello (1986)
(7) Corrosion	McIntyre (1983); Castle (1986); Lumsden and Szklarska-Smialowska (1978); Schreifels *et al.* (1989)
(8) Mineralogy and insulating samples	Hochella *et al.* (1986); Wei (1988)
(9) Surface extended energy-loss fine structure	De Crescenzi (1985); Idzerda *et al.* (1985a) Natarajan, Abel, and Hoffman (1985)
(10) Chemical effects	Madden (1983); Haas, Grant, and Dooley (1972); Hegde (1984); Kamath *et al.* (1982)

ness, segregation of impurities to the oxide surface, and localized oxide defects in metal-oxide-semiconductor (MOS) gate oxides may all be examined.

3. Wet chemical or plasma etching: etch resistant residues may be identified; redeposition of materials on side walls of steps can be detected.
4. Diffusion or ion implantation doping: the minimum detectable concentration of dopants detectable by SAM is $\approx 0.1\%$ or $\approx 10^{19}$ cm^{-2}. While this is adequate to characterize the first step (i.e., low-temperature step in which dopant is deposited onto the surface) of a normal two-step diffusion process, AES is not sensitive enough, in general, to detect dopants in finished devices. Similarly, the thickness, uniformity, and concentration of the dopant in high-dose shallow ion implants can be measured.
5. AES is applicable to many problems which involve metallization systems.
   (a) The composition distribution through the thin film depth may be obtained, and reveals segregation and annealing effects. Impurity concentrations and the interface chemistry of both Ohmic contacts and Schottky barrier diodes may be determined.
   (b) Refractory metal silicide features and contact metallization may be monitored to determine if the refractory

metal has fully reacted with the silicon.
   (c) The stoichiometry of codeposited silicides may be determined.
   (d) Aluminum spiking through shallow emitter junctions may be observed.
   (e) Interlayer diffusion and the effectiveness of diffusion barriers can be determined.
6. Packaging and bonding: the surface chemistry of pad and mating lead materials determines their bondability, strength, and electrical integrity.

Rather than to enumerate the literally thousands of journal articles in which AES and SAM are used to characterize microelectronic devices [by 1983 the number of papers published per annum for AES exceeded 1000 (van Oostrom, 1984)], the reader is referred to the review articles on applications of AES in microelectronics by Thomas and Joshi (1983), Helms (1982), Holloway (1980a, 1980b), Lowry and Hogrefe (1980), Olson *et al.* (1983), and van Oostrom (1984), which include extensive references.

4.1.2 Microelectronic Device Diffusion Barriers. As features of very-large-scale integrated microelectronic devices evolve into submicron dimensions, highly reliable diffusion barriers are necessary for metal contacts to shallow junctions. Several complementary analytical techniques—Auger electron spectroscopy (AES), Rutherford backscattering spectroscopy, secondary neutral mass spectroscopy, and secondary-ion mass spectroscopy—have all been used to investigate a reliable contact structure composed of sputtered W as the diffusion barrier and a thermally formed $TiN_y/TiSi_z$ contacting layer (Grove *et al.*, 1989). Similarly, refractory metal diffusion barriers for III-V semiconductor Ohmic contacts have been studied using Auger techniques (e.g., Huelsman and Fonstad, 1985).

4.1.3 Semiconductor Surface Potential Distribution. A contactless measurement technique known as Auger voltage contrast (AVC) for the observation of *p-n* junctions and regions using a scanning Auger microprobe has been described (Pantel *et al.*, 1982). By detecting the energy shift in Auger peaks related to Fermi-level changes in a semiconductor, internal potential variations as low as 20 meV can be measured with high lateral resolution. The source of potential contrast is clearly separated from other secondary emission contrast sources, such as topography, crystallinity, surface contamination, etc. Thus, AVC imaging of a doped semiconductor surface reveals *p-n* doped regions with high resolution and clarity, while these regions are not revealed by secondary electron microscopy.

The AVC technique may also be used for electrical depth profiling of shallow *p-n* junctions (Pantel, 1983). Depth profiling is obtained either directly, using an angle-lapped semiconductor surface, or indirectly, using sputtering. Note that while the sensitivity of Auger spectroscopy for the detection of chemical species is 10^{19}–10^{20} cm^{-3}, the AVC measurements correspond to a concentration of electrically active dopants of $\approx 1\times10^{15}$ cm^{-3}. For shallow junctions ($\approx$ 40 nm), sputter ion milling may be used to determine *p-n* junction depth. The main advantage of the AVC technique over traditional Rutherford backscattering, neutron activation, staining, pulsed capacitance, spreading resistance, or secondary-ion mass spectroscopy techniques is excellent depth resolution and information on the electrical activity of dopants.

4.2 Superconductors

Auger spectroscopy has long been used to study fracture surfaces and grain-boundary segregation processes (Seah, 1980, 1983b, 1983c). Alloyed superconducting Nb_3Sn-bronze wire filaments were fractured under ultrahigh-vacuum conditions $\approx 4\times10^{-7}$ Pa ($\approx 3\times10^{-10}$ Torr) and the structure and elemental composition of the individual wire filaments were examined (Suenaga and Jansen, 1983).

Comparison of Auger electron spectra obtained from intergranular and transgranular areas exposed on fracture surfaces of sintered $YBa_2Cu_3O_{7-\delta}$ specimens indicates that most grain boundary surfaces are deficient in oxygen and rich in copper compared to the bulk (Kroeger, *et al.*, 1988). No evidence of segregation of impurities to grain boundaries was observed. The results suggest that the grain-boundary layer is nonsuperconducting and a likely contributor to the problem of low critical current densities in $YBa_2Cu_3O_{7-\delta}$ materials. For a review of superconductivity, see SUPERCONDUCTIVITY, LOW TEMPERATURE and HIGH TEMPERATURE SUPERCONDUCTIVITY.

4.3 Metallurgy

4.3.1 Grain-Boundary Segregation. The classic example of catastrophic intergranular failure in engineering materials involves the disintegration of a turbine rotor from the Hinkley Point Power Station (United Kingdom) steam turbine generator during a routine overspeed test in 1969 (Kalderon, 1972). Scanning electron photomicrographs of a fracture surface from the failed rotor material revealed that the fracture had closely followed the prior austenite grain-boundary path with little deformation or energy expending processes. In other words, the rotor failed through brittle fracture. The fracture behavior is not characteristic of the particular alloy steel in general, since other rotors made from this material were quite safe.

Auger examination (Fig. 9) of a fracture surface from the failed rotor material showed that the fracture surface contained $\approx$ 0.45 monolayer of phosphorus at the grain boundaries. Thus, the catastrophic turbine rotor failure resulted from the segregation of the phosphorus impurity in the steel to the grain boundaries. This demonstrates the degree of care that must be taken, from casting to casting, to control the level of impurities in engineering materials.

4.3.2 Thin-Film Interdiffusion Studies. Bulk diffusion information can be extracted from Auger depth profile data of thin double-layer films (Pamler, 1987). In the case of a diffusivity which is independent of concentration, the one-dimensional diffusion equation can be written as

$$\frac{\partial c}{\partial t}=D\frac{\partial^2 c}{\partial z^2}\,. \tag{24}$$

The film consists initially of two separate layers with concentrations c_1 and c_2 (100 at.% and 0 at.%, respectively). In theory (Shewmon, 1963), the layers should be infinitely thick, but in practice, it is sufficient only that the thin-film thickness be greater than the diffusion length. After locating the origin $z = 0$ at the interface, the post-annealed composition profile $c(z,t)$ can be described by the complementary Gaussian error function to give

$$c(z,t)=\frac{1}{2}\left(\operatorname{erfc}\frac{z}{2\sqrt{Dt}}\right)\,. \tag{25}$$

Hall and Morabito (1976) proposed a simple deconvolution procedure to extract the diffusion coefficient D for the case that both annealed and unannealed profiles are error functions:

$$D=\frac{1}{4\pi t}(G_t^2-G_0^2), \tag{26}$$

which takes into account both the slope G_t of the annealed sample and the slope G_0 of the unannealed sample. The gradient,

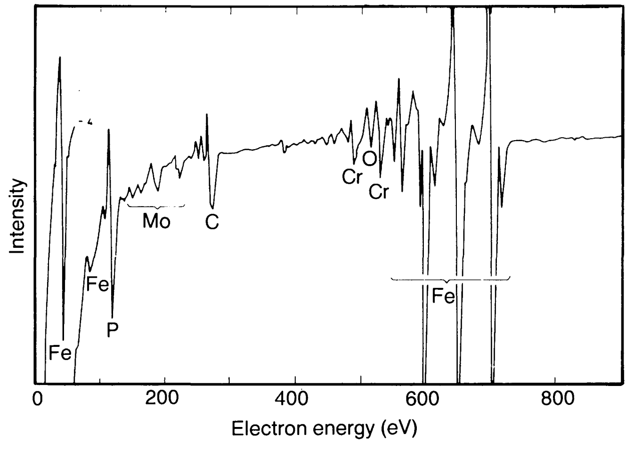

FIG. 9. Auger electron spectrum from the grain-boundary surface of the failed Hinkley Point power plant turbine rotor material exhibiting $\approx$ 0.45 monolayer of phosphorus segregated at the grain boundary (after Seah, 1975).

$G(z)=(z_{0.895}-z_{0.105})^{-1}$,

is the reciprocal of the difference in distance between the 89.5% and 10.5% points on a linear-probability scale (Hall, 1953) and is valid as long as the slope is linear at the interface ($z = 0$). More sophisticated analyses are available for situations in which the diffusivity should be considered as a function of concentration and when grain-boundary diffusion occurs in addition to bulk diffusion (Pamler, 1987).

4.3.3 Historical Metallurgy. Complementary analytical techniques were used to show that a panel of Ghiberti's "Doors of Paradise" of the Baptistry in Florence was originally gilded by fire or Hg-gilding (Mello, 1986). Auger spectroscopy provided an atomic layer-by-layer surface-composition depth profile. Metallographic microscopy showed that when amalgam was spread over the bronze substrate it accumulated in depressions on the metal surface. The Au layer was particularly porous and penetrated into the deep surface pits; this was confirmed by SEM. A depth profile of the outer surface showed that Hg was not on the external layer but in the outer portion of the Au film, so that it is unlikely to be from external contamination. An electron microprobe study of lower strata within the metal showed a little Au-Cu diffusion and the outer surfuce of the Au layer at minus 3–4 μm, in agreement with other Hg-gilded samples.

4.4 Other Applications

4.4.1 Corrosion. Corrosion processes which involve the formation or rupture of passivating thin films on metal surfaces have long been studied by AES (McIntyre, 1983, and Castle, 1986). Protective chromate, phosphate, nitrite, and molybdate films on iron formed by corrosion inhibitors added to aerated distilled water were first studied by Auger depth profile analysis in 1978 (Lumsden and Szklarska-Smialowska, 1978). More recently, an Auger depth profile study of corrosion-inhibiting films formed using synthetic cooling-water conditions has been published (Schreifels *et al.*, 1989). Similarly, AES has also been used to study the effects of fuel additives on the hot corrosion of industrial gas turbine blade alloys (Schweitzer *et al.*, 1982).

4.4.2 Insulating Samples and Mineralogy. The largest drawback to the use of Auger electron spectroscopy on insulating surfaces is sample charging (Wei, 1988). Certain beam conditions and tilt angles can prevent charging, however.

Insulators under electron-beam bombardment can remain neutral under certain beam conditions (Bruining, 1954). No charging will occur at the crossover voltages E_{p1} and E_{p2} where the primary beam current (I_p) is equal to the backscattered electron current (I_γ) plus the secondary electron current (I_δ) or $I_p=I_\gamma+I_\delta$ (Hochella *et al.*, 1986). Figure 10 shows secondary electron emission (I_δ) as a function of primary electron-beam energy (E_p) for an insulator. E_{p1} and E_{p2} typically differ by a few thousand electronvolts, and E_{p1} is usually only a few tens of volts (Dawson, 1966), too low to be useful in AES. Most Auger work on insulators is performed near E_{p2}, which is typically between 2 and 5 keV.

Tilting the sample relative to the primary beam axis permits the collection of unshifted Auger spectra at higher E_p than can be achieved with an untilted specimen. Tilting the sample results in an increase in both the backscattered and secondary emission efficiency for any particular beam energy (Kanter, 1961).

Hochella *et al.* (1986) demonstrated that the instrument and sample conditions under which a reasonable signal-to-noise ratio can be achieved for small spot AES/SAM of silicate, oxide, and sulfide mineral surfaces without charging or significant sample degradation are as follows:

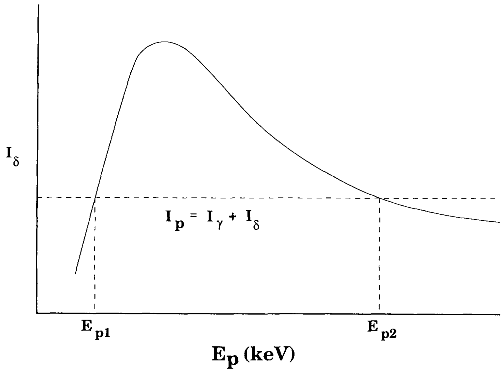

FIG. 10. Secondary-electron emission (I_δ) as a function of primary electron-beam energy (E_p) for an insulator (see text) (after Hochella *et al.*, 1986).

1. beam voltage typically 3 keV and no greater than 5 keV;
2. beam currents between approximately 1 and 20 nA for < 1-μm beam diameters;
3. tilt angle typically between 40° and 60°;
4. flat to semirough surface topology; and
5. light C, CO, CO_2 atmospheric contamination, sometimes resulting in a weak C Auger peak in the AES spectrum.

Scanning Auger microscopy is suitable for the semiquantitative probing of < 1-μm grains in rocks as well as performing surface studies with very high lateral resolution [Hochella *et al.* (1986)].

4.4.3 Surface Extended Energy-Loss Fine Structure. Auger electron spectrometers may be used, without modification, to obtain bond length information for certain types of systems. Surface extended energy-loss fine structure (SEELFS) is a new surface analytical technique [De Crescenzi (1985); Idzerda *et al.*, (1985a); Natarajan, Abel, and Hoffman (1985)] which provides structural information such as bond lengths and nearest neighbor coordination number. SEELFS is quite similar to EXAFS (extended x-ray absorption fine structure), except that electron energy-loss spectra are measured using an Auger electron spectrometer from approximately 50 to 500 eV above the core excitation edge of the material being studied.

While SEELFS may be used to study bulk metals and compounds [e.g., Cu (Idzerda *et al.*, 1985b) and GaAs (Terauchi *et al.*, 1985)], due to the surface sensitivity of the electron detection technique, SEELFS is more advantageously employed to study low atomic number submonolayer adsorbates on metals [e.g., Si-C (Natarajan, Abel, and Hoffman, 1985), Ni-O (De Crescenzi *et al.*, 1983), and Ti-O (Morar and Park, 1983)]. SEELFS is especially useful to study other absorbate/substrate systems which are inaccessible to current EXAFS techniques because of overlapping extended fine structure (Morar and Park, 1983).

4.4.4 Chemical Effects in AES. AES can be used as a valence-band spectroscopy that is surface sensitive, particularly when one or more valence electrons are involved in the Auger process (Madden, 1983). Experimentally observed changes in Auger line shapes have proved useful in identifying chemical states of elements at surfaces, particularly carbon, sulfur, nitrogen, and oxygen. Figure 11 illustrates changes in carbon *KVV* line shapes for three different carbon environments [molybdenum carbide (Mo_2C), silicon carbide (SiC), and graphite (C)] ranging from carbidic to graphitic (Haas, Grant, and Dooley, 1972). Note that both peak energy shifts and line-shape changes can affect quantitative analysis significantly. Similarly, analysis of Auger line shapes has been used to study the interaction of CH_3OH with oxygen precovered Cu (Hegde, 1984; Kamath *et al.*, 1982).

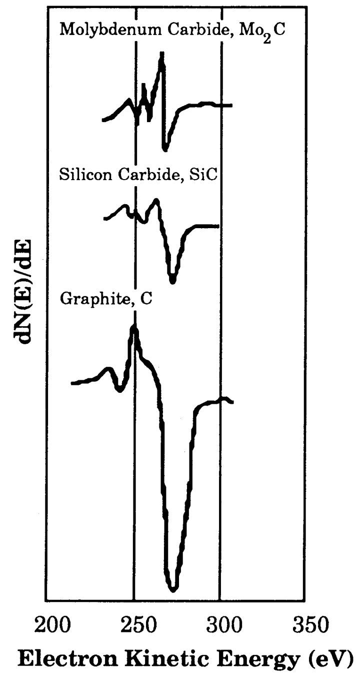

FIG. 11. Changes in carbon *KVV* line shapes for three different carbon environments ranging from carbidic to graphitic (after Haas, Grant, and Dooley, 1972).

GLOSSARY

Auger Line Scan: A one-dimensional view of the change in elemental concentration across an interface may be constructed by sweeping an electron beam in a line across the interface and measuring the elemental Auger electron peak intensity at multiple points on the line.

Auger Map: By rastering the electron beam across a sample surface, and at each raster point measuring the elemental Auger electron peak intensity, an image may be constructed in which the intensity of each raster point represents the amount of the element present. In this manner, a two-dimensional view, or Auger map, of the change in elemental concentration across a surface may be constructed.

Auger Sensitivity Factor: A set of relative normalization factors used in place of pure elemental standards to determine atomic concentrations from the Auger peak intensities of unknown samples.

Backscattering Term: A factor which accounts for sample ionization by backscattered energetic electrons produced in the sample by the primary electron beam.

Component Sputtering Yield Ratio: For a homogeneous alloy *AB*, the component sputtering yield ratio is the ratio of the bulk ratio of *A* to *B* to the ratio of *A* to *B* observed at a sputtered surface.

Cylindrical Mirror Analyzer (CMA): An electron kinetic-energy analyzer in which electrons pass between two coaxial cylinders of different radii. For the special case in which the entrance angle is $\alpha = 42^\circ\ 18'$, the CMA becomes a second-order focusing electron kinetic-energy analyzer.

Depth Profile: A compositional depth profile of a sample is the in-depth distribution analysis of chemical composition in the vicinity ($\approx 1\ \mu\text{m}$) of the sample surface.

Factor Analysis: A technique for Auger line-shape analysis of spectra selected from a compositional depth profile in which the presence and number of new chemical states formed at an interface may be determined.

Hemispherical Sector Analyzer (HSA): An electron kinetic-energy analyzer in which electrons pass between two hemispherical surfaces of different radii. The HSA is a second-order focusing electron kinetic-energy analyzer when electrons exit through the aperture of the exit slit and when the electron trajectory is within the limits imposed by the hemispherical surfaces.

Inelastic Mean Free Path: The probability-e^{-1} distance that Auger electrons created by the ionization process travel in the matrix material before being inelastically scattered and no longer being collected by the detection optics.

Logistic Function: A function involving an exponential and shaped like the letter "S."

Scanning Auger Microprobe (SAM): The addition of electron beam steering optics and a secondary electron detector enables the now modified Auger electron spectrometer to obtain both Auger and secondary electron images of the sample under study. The modified apparatus is known as a scanning Auger microprobe.

Spectral Stripping: An iterative process of subtracting known elemental Auger spectra from an unknown Auger spectrum to resolve small Auger peaks in the unknown Auger spectrum.

Surface Extended Energy-Loss Fine Structure (SEELFS): The electron analog of EXAFS (extended x-ray absorption fine structure), in which electron energy-loss spectra are measured using an Auger electron spectrometer from approximately 50 to 500 eV above the core excitation edge of the material being studied.

List of Works Cited

Auger, P. (1925), *J. Phys. Radium* **6**, 205–208.

Behrisch, R. (1981), in: R. Behrisch (Ed.), *Sputtering by Ion Bombardment I*, Springer-Verlag: New York, Chap. 1.

Betz, G. (1980), *Surf. Sci.* **92**, 283–309.

Betz, G., Wehner, G. K. (1983), in: R. Behrisch (Ed.), *Sputtering by Particle Bombardment II*, Springer-Verlag: New York, Chap 2.

Betz, G., Marton, J., Braun, P. (1980), *Nucl. Instrum. Methods* **168**, 541–545.

Betz, G., Dudonis, J., Braun, P. (1981), *Surf. Sci.* **104**, L185–L188.

Bishop, H. E., Coad, J. P., Rivière, J. C. (1972–73), *J. Electron Spectrosc. Relat. Phenom.*, **1** 389–401.

Briggs, D., Rivière, J. C. (1983), in: D. Briggs and M. P. Seah (Ed.), *Practical Surface Analysis by Auger and X-ray Photoelectron Spectroscopy*, Chichester: Wiley, Chap. 3.

Bruining, H. (1954), *Physics and Applications of Secondary Electron Emission*, London: Pergamon.

Castle, J. E. (1986), *Surf. Interface Anal.* **9**, 345–356.

Christou, A. (1976), *J. Appl. Phys.* **47**, 5464–5466.

Davis, L. E., MacDonald, N. C., Palmberg, P. W., Riach, G. E., Weber, R. E. (1976), *Handbook of Auger Electron Spectroscopy*, Eden Prairie, MN: Perkin-Elmer Corporation.

Dawson, P.H. (1966), *J. Appl. Phys.* **37**, 3644–3645.

De Crescenzi, M. (1985), *Surf. Sci.* **162**, 838–846.

De Crescenzi, M., Antonangeli, F., Bellini, C., Rosei, R. (1983), *Phys. Rev. Lett.* **50**, 1949–1952.

Fulghum, J. E., McGuire, G. E., Musselman, I. H., Nemanich, R. J., White, J. M., Chopra, D. R., Chourasia, A. R. (1989), *Anal. Chem.* **61**, 243R–269R.

Gaarenstroom, S. W. (1982), *J. Vac. Sci. Technol.* **20**, 458–461.

El Gomati, M. M. (1988), *Vacuum (GB)* **4/5**, 337–340.

Grove, C. L., Gregory, R. B., Hance, R. L., Sun, S. W., Kelly, N. (1989), *J. Vac. Sci. Technol. A* **7**, 1596–1600.

Haas, T. W., Grant, J. T., Dooley, G. J. (1972), *J. Appl. Phys.* **43**, 1853–1860.

Hall, L. D. (1953), *J. Chem. Phys.* **21**, 87–89.

Hall, P. M., Morabito, J. M. (1976), *Surf. Sci.* **54**, 79–90.

Hall, P. M., Morabito, J. M. (1979), *Surf. Sci.* **83**, 391–405.

Harris, L. A. (1968a), *J. Appl. Phys.* **39**, 1419–1427.

Harris, L. A. (1968b), *J. Appl. Phys.* **39**, 1428–1431.

Harris, L. A. (1974), *J. Vac. Sci. Technol.* **11**, 23–28.

Hegde, M. S. (1984), *Proc. Indian Acad. Sci. Chem. Sci.* **93**, 373–387.

Helms, C. R. (1982), *J. Vac. Sci. Technol.* **20**, 948–952.

Hochella, M. F., Jr., Hamis, D. W., Turner, A. M. (1986), *Amer. Mineral.* **71**, 1247–1257.

Hofmann, S. (1983), in: D. Briggs and M. P. Seah (Ed.), *Practical Surface Analysis by Auger and X-ray Photoelectron Spectroscopy*, Chichester: Wiley, Chap. 4.

Holloway, P. H. (1980a), *Appl. Surf. Sci.* **4**, 410–444.

Holloway, P. H. (1980b), in: T. L. Barr and L. E. Davis (Ed.), *Applied Surface Analysis ASTM STP 699*, Philadelphia: American Society for Testing and Materials, pp. 5–23.

Hondros, E. D., Stoddart, C. T. H. (1980), *Anal. Proc. R. Soc. Chem.* **17**, 467–471.

Huelsman, A. D., Fonstad, C. G. (1985), *J. Electrochem. Soc.* **132**, 2989–2992.

Ichimura, S., Shimizu, R. (1981), *Surf. Sci.* **112**, 386–408.

Idzerda, Y. U., Williams, E. D., Einstein, T. L., Park, R. L. (1985a), *Bull. Am. Phys. Soc.* **30**, 310, Abstract No. DI6.

Idzerda, Y. U., Williams, E. D., Einstein, T. L., Park, R. L. (1985b), *Surf. Sci.* **160**, 75–86.

Joshi, A. (1986), in: ASM Handbook Committee (Eds.), *Materials Handbook*, Vol. 10, Metals Park: American Society for Metals, 9th ed., pp. 549–567.

Kalderon, D. (1972), *Proc. Inst. Mech. Eng.* **186**, 341–377.

Kamath, P. V., Yashonath, S., Srivasan, A., Rao, C. N. R. (1982), *Appl. Surf. Sci.* **10**, 559–566.

Kanter, H. (1961), *Phys. Rev.* **121**, 677–681.

Kirchhoff, W. H., Chambers, G. P., Fine, J. (1986), *J. Vac. Sci. Technol. A* **4**, 1666–1670.

Kroeger, D. M., Choudhury, A., Binestad, J., Williams, R. K., Padgett, R. A., Coghlan, W. A. (1988), *J. Appl. Phys.* **64**, 331–335.

Lowry, R. K., Hogrefe, A. W. (1980), *Solid State Technol.* (Jan.), 71–75.

Lumsden, J. B., Szklarska-Smialowska, Z. (1978), *Corrosion* **34**, 169–176.

Madden, H. H. (1983), *Surf. Sci.* **126**, 80-100.

Malinowski, E. R., Howery, D. G. (1980), *Factor Analysis in Chemistry*, New York: Wiley-Interscience.

McGuire, G.E. (1979), *Auger Electron Spectroscopy Reference Manual*, New York: Plenum.

McIntyre, N. S. (1983), in: D. Briggs and M. P. Seah (Ed.), *Practical Surface Analysis by Auger and X-ray Photoelectron Spectroscopy*, Chichester: Wiley, Chap. 10.

Mello, E. (1986), *Gold Bull.* **19**, 123–126.

Morar, J. F., Park, R. L. (1983), *J. Vac. Sci. Technol. A* **1**, 1043–1046.

Morgan, A. E. (1983), *Nucl. Instrum. Methods Phys. Res.* **218**, 401–408.

Natarajan, C., Abel, P. B., Hoffman, R. W. (1985), *J. Vac. Sci. Technol. A* **3**, 1309–1311.

Nebesny, K. W., Maschhoff, B. L., Armstrong, N. R. (1989), *Anal. Chem.* **61**, 469A–481A.

Oechsner, H. (1975), *Appl. Phys.* **8**, 185.

Olson, R. R., Palmberg, P. W., Hovland, C. T., Brady, T. E. (1983), in: D. Briggs and M. P. Seah (Ed.), *Practical Surface Analysis by Auger and X-ray Photoelectron Spectroscopy*, Chichester: Wiley, Chap. 6.

van Oostrom, A. (1984), *Vacuum (GB)* **34**, 881–892.

Pamler, A. (1987), *Appl. Phys. A* **42**, 219–226.

Pantel, R., Arnaud D'Avitaya, F., Ged, Ph., Bois, D. (1982), *Scanning Electron Microsc.* **2**, 549–557.

Pantel, R. (1983), *J. Appl. Phys.* **43**, 650–652.

Powell, C. J., Seah, M. P. (1990), *J. Vac. Sci. Technol. A* **8**, 735–763.

Prutton, M., Browning, R., El Gomati, M. M., Peacock, D. (1982), *Vacuum* **32**, 351–357.

Rivière, J. C. (1983), in: D. Briggs and M. P. Seah (Ed.), *Practical Surface Analysis by Auger and X-ray Photoelectron Spectroscopy*, Chichester: Wiley, Chap. 2.

Roy, D., Carette, J.-D. (1971), *Can. J. Phys.* **49**, 2138–2159.

Schreifels, J., Labine, P., Gailey, R., Goewert, S., O'Brien, M., Jost, S. (1989), *Corrosion (Houston)* **45**, 420–428.

Schweitzer, K. K., Sickinger, A., Johner, G., Graber, U., Scmittt-Thomas, Kh. G. (1982), *Thin Solid Films* **95**, 73–83.

Seah, M. P. (1975), *Surf. Sci.* **53**, 168–212.

Seah, M. P. (1980), *Philos. Trans. R. Soc. (London) Ser. A* **295**, 265–278.

Seah, M.P. (1981), *Thin Solid Films* **81**, 279–287.

Seah, M. P. (1983a), in: D. Briggs and M. P. Seah (Eds.), *Practical Surface Analysis by Auger and X-ray Photoelectron Spectroscopy*, Chichester: Wiley, Chap. 5, 181–216.

Seah, M. P. (1983b), in: D. Briggs and M. P. Seah (Eds.), *Practical Surface Analysis by Auger and X-ray Photoelectron Spectroscopy*, Chichester: Wiley, Chap. 7, 247–282.

Seah, M. P, (1983c), *J. Microsc. Spectrosc. Electron. (France)* **8**, 177–191.

Seah, M. P. (1984), *Vacuum* **34**, 463–478.

Seah, M. P., Briggs, D. (1983), in: D. Briggs and M. P. Seah (Eds.), *Practical Surface Analysis by Auger and X-ray Photoelectron Spectroscopy*, Chichester: Wiley, Chap. 1.

Seah, M. P., Dench, W. A. (1979), *Surf. Interface Sci.* **1**, 2–11.

Shewmon, P. G. (1963), *Diffusion in Solids*, New York: McGraw-Hill.

Shimizu, R. (1983), *Jpn. J. Appl. Phys.* **22** (Pt. 1), 1631–1642.

Shimizu, R., Ichimura, S. (1981), *Quantitative Analysis by Auger Electron Spectroscopy*, Osaka: Toyota Foundation Research Report I-006, No. 76-0175.

Stein, D. F., Joshi, A. (1981), in: R. A. Huggins, R. H. Bube, and D. A. Vermilyea (Eds.), *Annual Review of Materials Science*, Palo Alto, CA: Annual Reviews, Chap. 11.

Suenaga, M., Jansen, W. (1983), *Appl. Phys. Lett.* **43**, 791–793.

Thien, V., Schmitz, F., Slotty, W., Voss, W. (1984), *Fresenius Z. Anal. Chem.* **319**, 646–654.

Terauchi, H., Sekimoto, S., Sano, N., Kat, H., Nakayama, N. (1985), *Appl. Phys. Lett.* **46**, 148-149.

Thomas, S., Joshi, A. (1983), in: Society for Advanced Materials and Process Engineering (Ed.), *Materials and Processes—Continuing Innovations*, Science of Advanced Materials and Process Engineering Series, Vol. 28. Azvsa, CA: SAMPE, pp. 752–764.

Turner, N. H. (1986), *Anal. Chem.* **58**, 153R–165R.

Turner, N. H. (1988), *Anal. Chem.* **60**, 377R-387R.

Turner, N. H. (1990), *Anal. Chem.* **62**, 113R-125R.

Wei, W. (1988), *J. Vac. Sci. Technol.* **6**, 2576–2581.

Werner, H. W., Garten, R. P. H. (1984), *Rep. Prog. Phys.* **47**, 221–344.

Wild, R. K. (1981), *Vacuum* **31**, 183–194.

Wirth, G., Grundhoff, K.-J. (1981), *Z. Metallkd.* **72**, 673–678.

Zalar, A. (1985), *Thin Solid Films* **124**, 223–230.

Zashkvara, V. V., Korsunskii, M. I., Kosmachev, O. S. (1966), *Sov. Phys. Tech. Phys.*, **11**, 96–99.

Zinner, E. (1983), *J. Electrochem. Soc.* **130**, 199C–222C.

Further Reading

Briggs, D., Seah, M. P. (Eds.) (1983), *Practical Surface Analysis by Auger and X-ray Photoelectron Spectroscopy*, Chichester: Wiley.

Casper, L. A. (Ed.) (1986), *Microelectronic Processing: Inorganic Materials Characterization*, Washington, D. C.: American Chemical Society, pp. 118–143.

Seah, M. P. (1980), *J. Vac. Sci. Technol.* **17**, 16–24.

Seah, M. P. (1986), *Microscopic Methods in Metals*, Berlin: Springer-Verlag, pp. 219–248.

Walls, J. M. (Ed.) (1988), *Methods of Surface Analysis*, New York: Cambridge University.

AUTOMOTIVE PROPULSION SYSTEMS

CHARLES A. AMANN AND TANVIR AHMAD, *General Motors Research Laboratories, Warren, Michigan, U.S.A.*

	Introduction	323
1.	**Thermodynamic Cycle**	323
1.1	Air-Standard Cycle	324
1.2	Fuel-Air Cycle	324
2.	**Mechanical Design**	325
2.1	Mechanical Configuration	325
2.2	Strokes per Cycle	326
2.3	Fuel System	327
2.4	Ignition System	327
2.5	Forced Induction	328
3.	**Performance Characteristics**	329
4.	**Transmission**	330
4.1	Manual Transmission	330
4.2	Automatic Transmission	331
4.3	Continuously Variable Transmission	332
5.	**Vehicle Characteristics**	333
5.1	Power Requirement	333
5.2	Fuel Economy	333
6.	**Environmental Aspects**	334
6.1	Sources of Tailpipe Emissions	334
6.2	Tailpipe-Emission Control	336
6.3	Efficacy of Controls	336
7.	**Fuels**	337
7.1	Liquid Hydrocarbons	337
7.2	Gaseous Fuels	338
7.3	Alcohols	338
7.4	Emissions Prospects of Alternative Fuels	339
8.	**Alternative Engines**	339
8.1	Diesel Engine	339
8.2	Spark-Ignition Stratified-Charge Engine	340
8.3	Rotary Engine	341
8.4	Gas Turbine	341
8.5	External-Combustion Engines	342
9.	**Electric Vehicles**	343
9.1	Battery-Electric Vehicle	343
9.2	Engine/Electric Hybrid Vehicle	343
9.3	Fuel-Cell Vehicle	344
	Glossary	344
	Works Cited	346
	Further Reading	346

INTRODUCTION

The traditional automotive propulsion system consists of the engine and the drive train, which together comprise the power train.

The engine is a machine that converts chemical energy stored in a fuel into useful mechanical work. By far the most common engine in automobiles is a multicylinder reciprocating mechanism that consumes a spark-ignited mixture of gasoline and air. The compression-ignited diesel engine, burning diesel fuel, dominates the heavy-duty truck field and also enjoys limited application in passenger cars and light-duty trucks.

The drive train includes a transmission and a differential mechanism. The transmission facilitates variation of its input/output speed ratio, usually in finite steps, to match the performance characteristics of the engine to the requirements of the vehicle. It also provides for operation of the vehicle in reverse. The differential allows the driving wheels on opposite sides of the vehicle to rotate at different speeds, as necessary during a turning maneuver.

In the automobile, the engine is most often located at the front of the vehicle, with power being transmitted to either the two rear wheels or the two front wheels. Alternatively, a rear-mounted or midship-mounted engine has been used with rear-wheel drive, and four-wheel drive has also been employed.

1. THERMODYNAMIC CYCLE

Because the energy conversion accomplished in the engine involves thermodynamics, it is helpful to understand the basic thermodynamic cycle after which the engine is patterned.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1.1 Air-Standard Cycle

Operation of the spark-ignition gasoline-fueled engine commonly used in automobiles is based on the Otto cycle. The air-standard Otto cycle, i.e., an Otto cycle in which the working fluid is approximated by air having a constant specific heat, is represented on a pressure-volume plane in Fig. 1. At the beginning of the cycle (point 1), the air is at its lowest pressure and temperature, and the cylinder volume is at its maximum. If the common slider-crank mechanism is used to effect the cycle, this means that the piston is at bottom dead center (BDC). As the piston rises toward top dead center (TDC) at 2, the cylinder air is compressed reversibly and adiabatically. From 2 to 3, heat is added at constant volume, increasing cylinder pressure and temperature. From 3 to 4, the cylinder air undergoes reversible adiabatic expansion. From 4 to 1 the air is cooled to its original state.

The area under a curve on a pressure-volume diagram is proportional to mechanical work. The area under 1-2 represents compression work done on the cylinder gas; that under 3-4 represents expansion work done by the gas. The difference, i.e., the enclosed area of the diagram, represents the net work done by the cycle. Its ratio to the heat added from 2 to 3 is the indicated thermal efficiency of the cycle, which is a function of the compression ratio ($V_1/V_2=r$) and the ratio of specific heats of the working fluid ($\gamma=1.4$ for a perfect diatomic gas). The thermal efficiency of the air-standard Otto cycle can be shown to be $\eta = 1-r^{1-\gamma}$. It is plotted in Fig. 2.

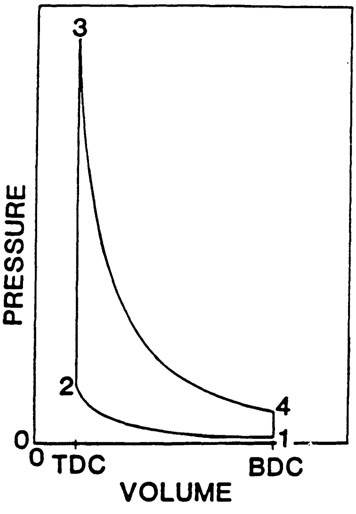

FIG. 1. Pressure-volume diagram for the air-standard Otto cycle.

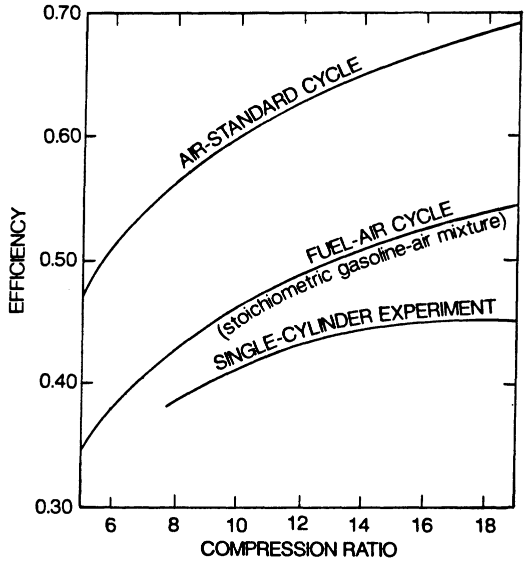

FIG. 2. Thermal efficiency as a function of compression ratio for the air-standard Otto cycle, for the fuel-air Otto cycle with a stoichiometric mixture of gasoline and air, and as determined experimentally from the indicated power of a single-cylinder engine.

1.2 Fuel-Air Cycle

The next level of thermodynamic approximation of the engine is the fuel-air Otto cycle. It recognizes that from 1 to 2, the working fluid is a mixture of air and fuel. From 2 to 3, the added heat is generated internally through combustion. From 3 to 4, the working fluid consists of combustion products. From 4 to 1, the products must be expelled from the cylinder to accommodate a fresh fuel-air charge. Throughout the cycle, specific heat varies with the temperature, pressure, and composition of the cylinder gas. During combustion the temperature is high enough to cause dissociation of the chemical species. The thermal efficiency of the fuel-air cycle, which accounts for these effects, is also shown in Fig. 2 for a stoichiometric (chemically correct) mixture. For mixtures richer than stoichiometric, the fuel-air cycle efficiency falls below that shown. Conversely, as the mixture is leaned from stoichiometric, the fuel-air cycle efficiency progressively increases, approaching the air-standard cycle as an upper limit.

2. MECHANICAL DESIGN

A broad variety of engine configurations has seen automotive use, including versions operating on both four-stroke and two-stroke cycles. The fuel system and the ignition system are key aspects of engine operation. Although most automotive engines are naturally aspirated, some use forced induction.

2.1 Mechanical Configuration

The normal embodiment of the Otto cycle incorporates the slider-crank mechanism illustrated in the cylinder cross section of Fig. 3. The slider element is the piston, acted upon by the gas pressure within the cylinder working space above it. The crankshaft, which extends the length of the engine and is normal to the plane of Fig. 3, contains offset cranks for the individual cylinders. The connecting rod, which joins the piston pin to the crankpin, transforms the reciprocating motion of the piston into the rotary motion of the crankshaft. Engine power is delivered to the transmission input by the crankshaft. Features of the slider-crank mechanism evident in Fig. 3 are the crankshaft counterweight opposite the crankpin for maintaining engine balance and grooves in the piston for mounting expandable piston rings that seal the working-space pressure and control the lubricating-oil film on the cylinder wall.

The engine structure of Fig. 3 is comprised of three assemblies, stacked vertically. The block contains the cylinder bore in which the piston reciprocates. Coolant passages surrounding the bore are also seen, as is the camshaft, discussed below. (In the past a few automotive engines have seen use in which the liquid-cooling system is abandoned in favor of air cooling. In this case the exteriors of the cylinder and head are heavily finned, and the fins are subjected to a directed cooling-air stream for removal of excess heat.) Atop the block is the cylinder head, which contains poppet valves for controlling the flow of mixture into the cylinder and exhaust from it. The head contains additional coolant passages (not shown), as well as the intake and exhaust ports that communicate with the intake and exhaust manifolds, which fasten onto the head. The third element is the oil pan, which fastens onto the bottom of the block and contains the oil reservoir. The rotating end of the slider-crank mechanism splashes oil from this reservoir onto the cylinder walls. Crankshaft journal bearings are normally lubricated with pressurized oil conducted through passages drilled in the crankshaft. Lubrication of the piston-pin bushing is usually accomplished with splashed oil, although in some engines the oil is delivered under pressure through a hole running the length of the connecting rod.

The intake valve, mounted in the head, is shown in Fig. 3. Behind it is an exhaust valve. Both are held closed by springs until the valve-actuation mechanism applies sufficient force to overcome the spring load. In this engine, a camshaft rotates in the block at half crankshaft speed. The intake and exhaust cam lobes each raise and lower their respective tappets, which move push rods extending from the block into the head. There the upper end of each push rod bears against one end of a centrally pivoted rocker arm, the opposite end of which oscillates downward and upward to open and close the valve at the proper time.

Alternatively, the camshaft may be located in the head, providing an overhead cam

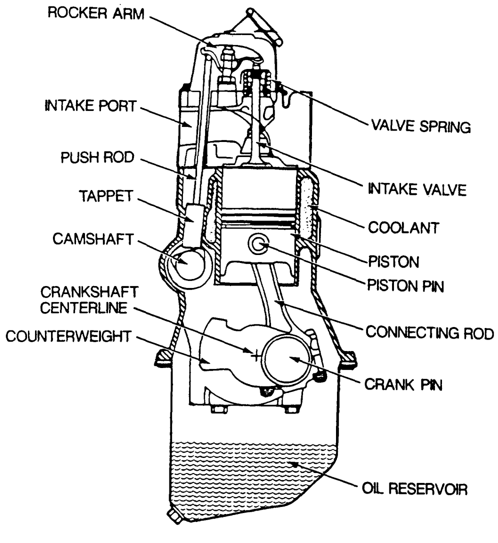

FIG. 3. Cross section through the cylinder of a four-stroke gasoline engine.

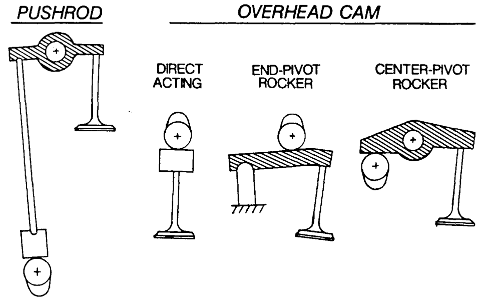

FIG. 4. Various overhead-cam arrangements for the automotive four-stroke engine.

(OHC). The camshaft is then driven from the crankshaft through a belt, a chain, or a gear train. In Fig. 4, several OHC arrangements are shown schematically alongside the pushrod arrangement. In some OHC engines, one camshaft is used for the intake valves and a second for the exhaust valves, resulting in a dual overhead cam (DOHC) engine.

Although only a single cylinder is shown in Fig. 3, the modern automotive engine normally uses a multiplicity of cylinders. Most common are four cylinders arranged in line, six cylinders either in line or in a "vee" arrangement, and eight cylinders in a vee. Variations include the in-line three-, five-, and eight-cylinder engines, horizontally opposed four- and six-cylinder engines, and vee engines with ten, twelve, and even sixteen cylinders.

2.2 Strokes per Cycle

Most automotive engines employ four complete piston strokes: compression, expansion, exhaust, and intake. The latter two strokes are not part of the ideal Otto cycle but have proven to be an effective addition for exchanging burned products for a fresh fuel-air charge. This four-stroke cycle is executed within the engine geometry of Fig. 3.

As an alternative, the Otto cycle can be completed in just two strokes of the piston, eliminating the two extra strokes dedicated to gas exchange. A variety of two-stroke engine arrangements exists. The two most common types are diagrammed in Fig. 5.

In the crankcase-scavenged configuration, the intake and exhaust processes are effected through ports in the cylinder wall. The fresh charge is first drawn into the crankcase through a check valve by the upward motion of the piston during its compression stroke. As the piston approaches TDC, the cylinder contents are ignited with a spark plug. As the piston descends on the expansion stoke, its upper edge first begins to uncover the exhaust port, allowing escape of the pressurized combustion products. Shortly thereafter, uncovering of the intake port begins and the downward piston motion forces the charge in the crankcase through a transfer passage into the cylinder. A portion of this incoming charge escapes directly through the exhaust port as it scavenges out the burned gas from the previous cycle, but the balance is trapped for compression and subsequent combustion.

This crankcase-compression version precludes the customary oil reservoir in the crankcase, leading to the use of rolling-element bearings that are more tolerant of sparse lubrication than the journal bearings normally found in four-stroke engines. If the fuel is added to the air prior to its entry into the crankcase, a substantial fraction of it es-

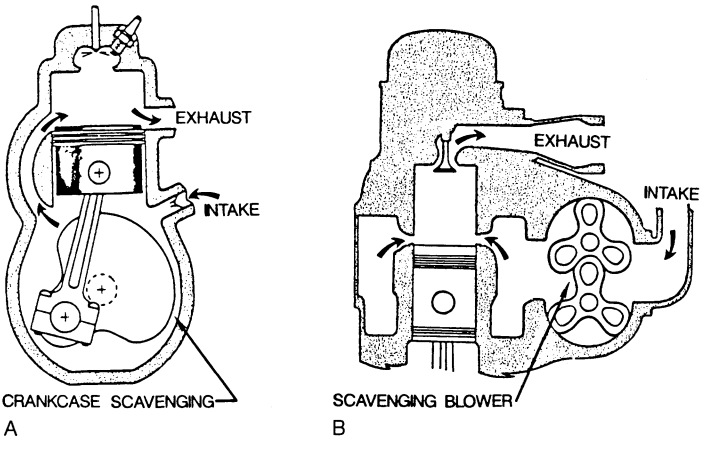

FIG. 5. Two-stroke engine arrangements: (a) crankcase-scavenged, (b) uniflow with blower scavenging.

capes unburned into the exhaust system during scavenging. This is intolerable if control of exhaust emissions is critical. In that case, introduction of fuel into the cylinder must be delayed until both ports have been closed. This can be accomplished by injecting the fuel directly into the cylinder.

In an alternative (not shown) to crankcase scavenging that allows oil in the crankcase, the incoming charge is compressed by an engine-driven blower and introduced through the intake port in the same manner as with crankcase scavenging.

The uniflow configuration, also shown in Fig. 5, uses an engine-driven scavenging blower to force the cylinder charge through intake ports surrounding the cylinder, but exhaust is accomplished through one or more poppet valves in the head. The unidirectional upward flow of the fresh charge offers more effective scavenging, and use of cam-actuated exhaust valves provides greater flexibility in timing the gas-exchange events.

2.3 Fuel System

In automotive engines, for many years the fuel was introduced into the air by means of a carburetor located at the entrance of the intake manifold. The simple carburetor is shown schematically in Fig. 6. A float-actuated inlet valve maintains a constant fuel level in the float bowl. An open-ended fuel tube emanating from the bottom of the bowl terminates in the throat of a venturi through which the engine air is drawn. With atmospheric pressure above the fuel in the float bowl and a depression in the venturi throat that increases as engine airflow increases, fuel exits the fuel tube in approximate proportion to engine airflow.

Although this simple sketch illustrates the central feature of the carburetor, many refinements (not shown) are required in a practical implementation. These include provisions for counteracting the surface tension of the fuel, inclusion of a special idling circuit, a means for enriching the mixture at heavy engine loads, and a means for temporary enrichment during sudden increases in demanded engine load.

Also shown in Fig. 6 are a choke valve upstream of the venturi and a throttle valve downstream of the venturi. Partially closing the choke during cold starts enriches the mixture fed to the engine to compensate for the loss in fuel evaporation within the intake manifold at low temperatures. Regulation of the throttle valve controls engine output. As the throttle is closed, the increasing pressure loss across it depresses intake-manifold pressure so that the compression stroke starts from a subatmospheric pressure. The resultant reduction in mass of the fresh charge trapped in the cylinder decreases engine output. Throttle-valve position is adjusted to match driver needs by means of the driver-actuated accelerator pedal.

As exhaust-emission standards have been tightened, the traditional carburetor is no longer capable of controlling air-fuel ratio within the necessary limits. That carburetor was originally replaced by a variation in which fuel flow was modulated by a pulsing solenoid valve. This approach has been largely supplanted by one of two forms of fuel injection. With single-point injection, one injector located above the throttle plate supplies several, if not all, cylinders. With multipoint fuel injection, an injector located in each intake port supplies the fuel needs of its own cylinder. The amount of fuel injected per stroke is controlled by an on-board computer.

2.4 Ignition System

In the automotive gasoline engine, the ignition system provides a spark for igniting the cylinder mixture at the proper time during the cycle in a reliable and repeatable manner. The traditional battery-ignition system,

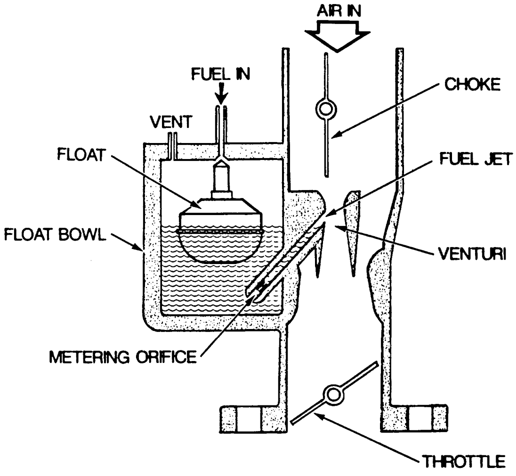

FIG. 6. Elements of the simple carburetor.

illustrated in Fig. 7, includes a battery, ignition switch, and coil. The primary side of the coil is grounded through mechanically actuated breaker points and a condenser, connected in parallel. The secondary side of the coil is grounded through a distributor switch and the individual spark plugs. In a four-stroke engine, the breaker-point cam and the distributor arm rotate at half crankshaft speed to provide a spark every other engine revolution.

When the breaker points are closed, current flows to the primary windings of the coil, building up a magnetic field. When the points are opened, the field collapses and the condenser is charged by the resulting induced current. When the condenser discharges back into the primary winding, the magnetic field quickly breaks down completely. This induces a high-voltage current in the secondary winding that is routed to the appropriate cylinder by the distributor, causing a spark to jump the spark-plug gap and ignite the mixture.

The time that elapses between occurrence of the spark and completion of combustion varies with engine speed and load. In the traditional ignition system of Fig. 7, this characteristic is accommodated by changing the phasing of breaker-point opening with respect to the crankshaft. This is accomplished with a speed-sensitive centrifugal mechanism and a diaphragm responding to intake-manifold vacuum.

The modern electronic ignition system is an improved embodiment of the same principles. It employs a pulse generator, which serves as a maintenance-free electronic circuit breaker, in place of the breaker points. The mechanical distributor may be replaced by a maintenance-free electronic version. The speed- and vacuum-sensitive signals formerly used to establish spark timing are replaced by commands from the engine control module (ECM). The ECM contains tabulations or algorithms for desired spark timing as a function of a broad range of measured operating parameters, including speed, manifold pressure, barometric pressure, coolant temperature, and throttle position.

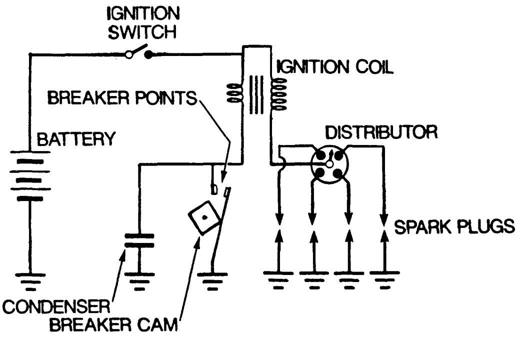

FIG. 7. Diagram of a battery-ignition system.

2.5 Forced Induction

Because engine power is dependent on mass airflow rate, power can be increased by pressurizing the intake manifold for increased intake-air density. This is accomplished either by supercharging or turbocharging.

In supercharging, an engine-driven blower is used to raise intake-air density. A common blower is the type illustrated in the uniflow two-stroke configuration of Fig. 5, in which air is pumped into the engine by the intermeshing lobes of two counter-rotating rotors. When such an engine is operated at partial engine load, a subatmospheric pressure is desired in the intake manifold and the blower becomes superfluous. Under these circumstances the blower must be either bypassed or disconnected to avoid its parasitic power drain, which would otherwise decrease engine thermal efficiency.

The turbocharger involves a centrifugal compressor and an exhaust turbine (usually of the radial-inflow type) mounted on a common high-speed shaft. The high-temperature gas exiting the exhaust manifold passes through the turbine, causing it to spin the compressor. The engine intake charge passes through that compressor, experiencing an increase in pressure on its way to the cylinders. For operation at part load, the turbocharger speed falls and the engine throttle valve, located upstream of the compressor, decreases cylinder inlet pressure to the desired level.

An intercooler (or aftercooler) is sometimes located between the compressor and the engine to lower the temperature of the compressed charge, hence increase its density, so that a greater charge mass can be inducted for increased power. The lower temperature also reduces the tendency for combustion knock to occur. The intercooler coolant is usually ambient air.

3. PERFORMANCE CHARACTERISTICS

The performance of the automotive engine is tied to the thermodynamic cycle on which it is based. One characteristic of normal interest is the relationship between its output, measured in terms of either power or torque, and its speed. Another is the effectiveness with which it converts fuel energy into power, normally measured in terms of either thermal efficiency or brake specific fuel consumption.

The actual pressure-volume diagram for a four-stroke Otto-cycle engine operated at part load is shown in Fig. 8 and may be contrasted to the air-standard full-load diagram of Fig. 1. In the actual engine, compression starts from subatmospheric pressure at *A*. Spark ignition occurs before TDC at *e*, with combustion continuing to point *f*. Clearly, heat addition does not occur at constant volume in a real engine. Neither the compression nor the expansion process is adiabatic, as it is in the ideal cycle. During both the compression and expansion strokes, there is a net heat loss to the coolant. Gas begins to leave the cylinder as the exhaust valve starts opening at *eo*. The exhaust valve remains open during the exhaust stroke (*CD*) and does not close completely until after TDC at *ec*. The intake valve starts to open before TDC at *io*, remains open during the intake stroke (*DhA*) and past BDC, closing during the compression stroke at *ic*.

The mean height of area *AgeBfC*, obtained by dividing the enclosed area by the stroke length *CD*, defines the indicated mean effective pressure (IMEP). The indicated power (P_i) developed at the piston during the compression and expansion stroke is

$$P_i = \bar{p}_i D(N/k),$$

where $\bar{p}_i$ equals IMEP, D is the total piston displacement, N is the crankshaft rotational speed, and k is the number of revolutions per cycle (equal to 2 for the four-stroke cycle, 1 for the two-stroke cycle). For a fixed air-fuel ratio and with optimized ignition timing, indicated power in a given engine tends to vary directly with the mass airflow rate.

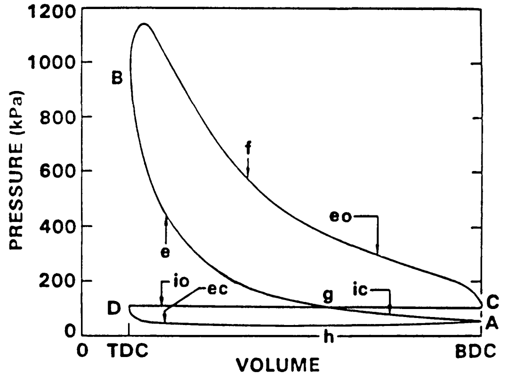

FIG. 8. Actual pressure-volume diagram for a four-stroke engine operating at part load.

The effect that variations from the ideal fuel-air cycle have on the efficiency of an actual engine, based on its indicated power at full throttle, is shown by the lowest curve in Fig. 2. This curve is from experimental data on a single-cylinder engine having a displacement of 0.66 L (Caris and Nelson, 1959). In this engine, maximum efficiency occurred at a compression ratio of about 17. Other experiments suggest that the benefit of high compression ratio diminishes as individual-cylinder displacement is decreased (Muranaka *et al.*, 1987). The principal factors causing the loss in efficiency in the real engine, compared to the fuel-air cycle, are heat loss to the coolant and departure from constant-volume combustion.

The output brake power (P_b) actually delivered by the engine is less than the indicated power by the power lost to mechanical friction plus that expended in pumping the working fluid through the engine. The associated pumping work is represented by area *CgDhA* in Fig. 8. As the throttle in the intake manifold is closed to decrease engine output, intake-manifold absolute pressure decreases, intake-stroke pressure *DhA* assumes a lower position on the diagram without a corresponding effect on exhaust-stroke pressure *CgD*, and the pumping work increases.

The mechanical efficiency accounting for the friction and pumping losses, $\eta_m (= P_b / P_i)$, reaches about 90% for a modern automotive engine at full load but approaches zero at the idle condition. Typical wide-open-throttle curves of indicated power, brake power, and mechanical efficiency are shown as a function of engine speed in Fig. 9.

Engine brake torque (τ) is given by $\tau = P_b / 2\pi N$. It tends to be proportional to the product of the air consumed per cycle and the me-

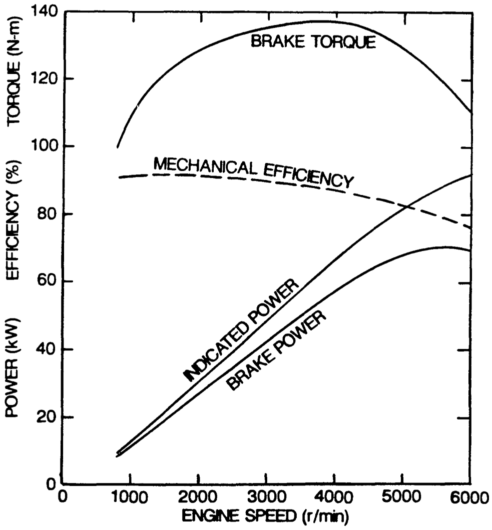

FIG. 9. Full-load characteristics of a typical gasoline engine.

chanical efficiency. A typical wide-open-throttle torque curve is included in Fig. 9.

The ratio of fuel-consumption rate to brake power, known as the brake specific fuel consumption (BSFC), is a frequently used measure of the effectiveness of energy conversion that is customarily expressed in units of g/kW h. Brake thermal efficiency (η_b), which is dimensionless, is equal to the reciprocal of the product of BSFC and the gravimetric heating value of the fuel, expressed in appropriate units.

A typical engine performance map is shown in Fig. 10. The upper boundary traces the maximum brake torque, and the isopleths show the variation in brake thermal efficiency. Maximum efficiency occurs within an island located at a fairly high fraction of engine load and an intermediate speed range. Above this island, efficiency falls because the mixture is enriched for maximum power. Below the island, efficiency falls primarily because mechanical friction and pumping losses consume an increasing proportion of the indicated power developed. Contributing to the fall in efficiency at high speeds is the associated reduction in mechanical efficiency. At low speeds, a factor tending to decrease brake thermal efficiency is increased heat loss, a result of there being more time for heat transfer.

4. TRANSMISSION

The transmission is located between the engine and the differential. ln the absence of losses, power, which is the product of rotational speed and torque, is conserved across the transmission. Thus, the transmission offers a means of trading a reduction in its output-shaft speed, relative to its input-shaft speed, for a multiplication of torque.

Geared transmissions may be classified according to whether shifting is accomplished manually or automatically. When the vehicle is standing, the engine operates at its idling speed, but the driving wheels of the vehicle are stationary. Consequently, with either transmission some provision must be included to accommodate this infinite input/output speed ratio.

4.1 Manual Transmission

For passenger cars, the manual transmission normally incorporates from three to six finite speed ratios, with five currently being the most common. These speed steps are effected by transmitting the power through various combinations of driver-selected gear pairs.

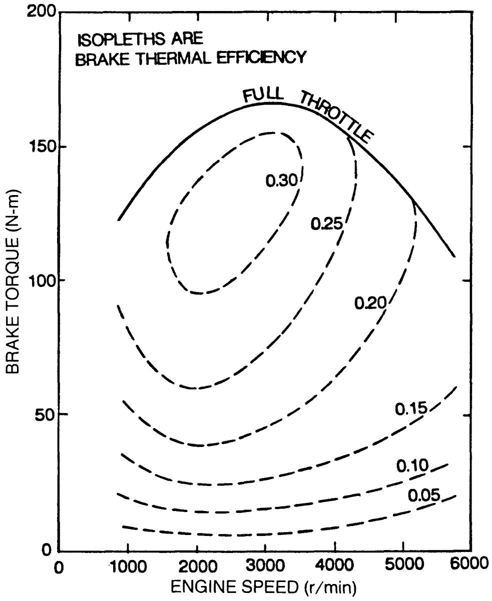

FIG. 10. Performance map of a typical gasoline engine.

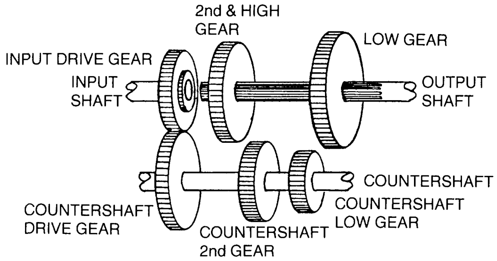

FIG. 11. Principles of a three-speed manual transmission, shown in neutral. In low, the low gear on the splined output shaft is slid axially along it into engagement with the countershaft low gear. In second, the two second gears are similarly engaged instead. In high, the gear at the left end of the output shaft is locked to the input drive gear to provide direct drive.

The principle of such a transmission system is illustrated in Fig. 11. The input and output shafts are coaxial and parallel to a countershaft. The input drive gear is permanently meshed with the countershaft drive gear, causing both shafts to rotate concurrently. The axial positions of the remaining gears on the countershaft are fixed, but the gears on the output shaft are internally splined so they can be moved axially along matching splines on the periphery of the output shaft. The position of the gear-selector lever determines which pair of gears on the countershaft and output shaft is engaged. Since the working diameters of meshed gears are proportional to the number of teeth they contain, the input/output shaft-speed ratio is $(N_i/N_{cd})(N_c/N_o)$, where N is the number of gear teeth, and the subscripts i, cd, c, and o denote the input drive gear, countershaft drive gear, and the countershaft and output-shaft gears that have been selected for engagement, respectively. For direct drive (input/output speed ratio equal to 1), teeth are provided on the gear at the left end of the output shaft so that it can be slid along the shaft into locked engagement with matching teeth added to the input drive gear.

To manage the infinite speed ratio in a standing vehicle, the transmission is shifted into neutral, which disengages all output-shaft gears from gears on the countershaft. When the transmission is shifted into reverse, an extra gear is interposed between the appropriate countershaft and output-shaft gears to reverse the rotation of the output shaft.

During shifting from one speed ratio to another, the driver depresses a clutch pedal. This opens a friction clutch between the engine crankshaft and the transmission-input shaft to sever that connection until the desired set of transmission gears has been engaged by manipulating the gear-selector lever. The clutch consists of two parallel coaxial circular plates, to one of which is attached a friction material. When the clutch pedal is released, these plates are pressed together by springs to lock the engine to the transmission input.

4.2 Automatic Transmission

The automatic transmission incorporates the same fundamental functions as the manual version, but there are some significant differences in implementation. For instance, the typical manual transmission employs spur gears in a countershaft arrangement, as illustrated in Fig. 11, but the automatic transmission usually employs planetary gears. The speed ratio across a planetary gear set, which is illustrated in Fig. 12, depends on which of its three elements—sun, planets, or ring gear—is prevented from rotating, i.e., grounded. Grounding is accomplished with clutches that are automatically engaged and disengaged by a shift-scheduling controller that works with such input signals as engine speed, throttle position, and intake manifold pressure.

There is a difference in the way manual and automatic transmissions normally interact with the engine. With the manual transmission, the typical driver releases the accelerator pedal (closes the throttle) while the clutch is disengaged for gear shifting. With the automatic transmission, however, the

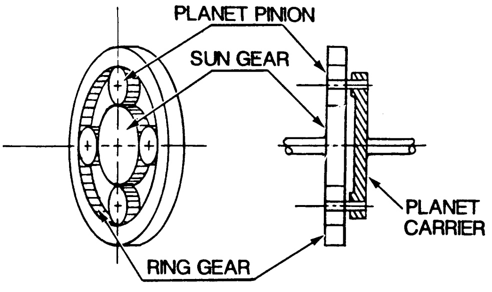

FIG. 12. Planetary gear set, used in most automatic transmissions.

shift is effected automatically without a correspondingly severe, temporary drop in engine power.

In the automatic transmission, the clutch of the manual transmission is typically replaced by a torque converter such as illustrated in Fig. 13. The basic torque converter consists of a torus filled with circulating automatic transmission fluid and containing three sets of blading—pump (or impeller), turbine, and stator (or reactor). The engine rotates the pump element, which pumps fluid radially outward. As that fluid then flows radially inward through the turbine, it transmits torque through the turbine blading to the gearbox input.

The stator blades are mounted on the transmission case through a one-way clutch. When a standing vehicle starts to move, the clutch is engaged and the stator blades are stationary. They redirect the fluid flowing through them en route to the pump inlet in a way that provides multiplication of torque across the torque converter. When the vehicle wheels are locked, this torque multiplication is normally in the range of 2.0–2.5, the input/output speed ratio across the torque converter is infinite, and both the power transmitted and the efficiency of transmission are zero. As the vehicle begins to move, the output/input torque ratio across the converter falls toward unity, the turbine/pump speed ratio rises toward unity, and the power transmitted and the efficiency of the torque converter both increase. As the turbine speed nears the pump speed, the changing reaction torque on the stator releases the one-way clutch through which it is grounded to the case and allows it to rotate (free-wheel). Torque multiplication ceases, and the pump and turbine act as a fluid coupling that transfers engine torque to the gearbox input through the action of the circulating transmission fluid on the blading of the torque converter.

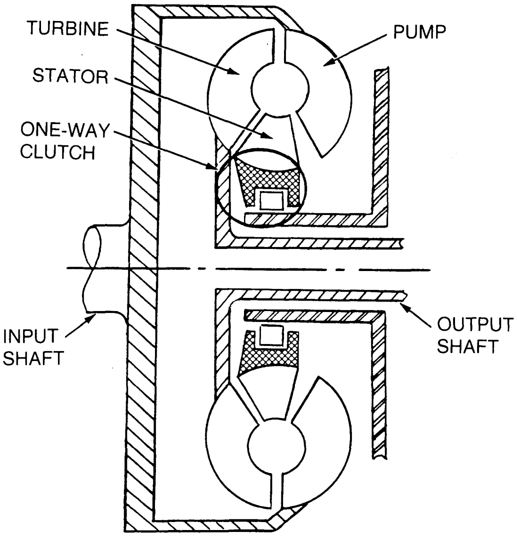

FIG. 13. Basic torque converter.

The turbine speed never quite reaches the pump speed. When increased power is demanded, this speed difference (slip) increases. To overcome the power loss associated with slip, many transmissions incorporate a clutch that mechanically joins the turbine to the pump during driving at constant speed and modest acceleration and deceleration. This clutch is released when the driver demands a large increase in power, allowing the torque converter to multiply input torque.

The torque-converter clutch is normally engaged only in the highest gear ratio(s), i.e., lowest input/output speed ratio(s). Engagement replaces the fluid joint of the torque converter with a solid mechanical connection. This defeats the damping characteristics of the fluid joint, allowing torque impulses from the crankshaft to be transmitted into the drive train, where they excite objectionable vibrations in the vehicle in the lowest gear(s).

Because of the torque-multiplication capability of the converter, the typical automatic transmission needs from one to two fewer speed ratios in its gearbox. Thus, today most automatic torque-converter transmissions have three or four forward speeds, compared to four or five in the manual transmission.

4.3 Continuously Variable Transmission

While common manual and automatic transmissions vary the input/output speed ratio in finite geared steps, over a limited range of speed ratios the continuously variable transmission (CVT) accomplishes such speed variation as a continuum. Current CVTs transmit torque through a specially designed belt or chain connecting driving and driven sheaves. One-half of each sheave can be moved axially a small distance with respect to its mating half, effectively varying the sheave working diameter. Thus the input/

output speed ratio can be decreased by increasing the diameter of the driving sheave while concurrently decreasing the diameter of the driven.

The torque-transmission capability of such a system currently limits its use to small cars with correspondingly low-output engines. Because of ratio-range limitations of the belt or chain, such additions as a two-speed gearbox or a torque converter have been considered to match the performance characteristics of existing transmissions.

5. VEHICLE CHARACTERISTICS

Vehicle performance and the distance traveled on a given quantity of fuel depend on the interaction between power-train characteristics and the power requirement of the vehicle.

5.1 Power Requirement

The power required (P_r) to propel a vehicle is

$$P_r = [C_r mg \cos\alpha + C_d A(\rho V^2/2) + m_e g a + mg \sin\alpha]\, V,$$

where C_r is the rolling-resistance coefficient of tires, m is the vehicle mass, m_e is the effective vehicle mass, including rotational inertias, C_d is the aerodynamic drag coefficient, A is the vehicle frontal area, V is the vehicle velocity, a is the vehicle acceleration, g is the gravitational constant, and α is the slope of road grade. On the right-hand side of this equation the first term within the brackets accounts for the rolling resistance of the tires, the second represents the aerodynamic drag, the third is the power consumed in accelerating the vehicle mass, and the fourth accounts for the extra burden when climbing a hill.

In Fig. 14 the power required by a typical car to overcome rolling resistance and the additional increment attributable to aerodynamic drag are shown shaded. The upper boundary of the shaded region is the constant-speed level-road power requirement. The dashed curve above it reflects the additional power required to climb a 5% grade. Above these power-required curves are the wide-open-throttle power-available curves in each gear of a five-speed transmission. The difference between the power available and the power required is the power reserve available for acceleration. At a given road speed, say 60 km/h, it is evident that less reserve power is available in the highest gears (lowest input/output speed ratio). Additional power reserve is made available by downshifting to a lower gear.

5.2 Fuel Economy

Not apparent from Fig. 14 is the fact that when driving at a given speed in a low gear, fuel economy suffers. This is illustrated in Fig. 15, where the level-road requirement is superimposed on the engine torque-speed map. It is evident that the higher the gear, the more favorable the brake thermal efficiency. However, again, the torque reserve is least in the highest gears, necessitating more frequent downshifting. For any point along one of the operating lines in Fig. 15, the fuel economy in km/L can be computed from

$$\text{km/L} = d/(\text{BSFC})V,$$

where d is the specific gravity of the fuel, BSFC can be calculated from the thermal efficiency at the point of interest and the heating value of the fuel, and vehicle velocity is obtained by dividing the engine speed on the abscissa by the engine-speed/vehicle-velocity

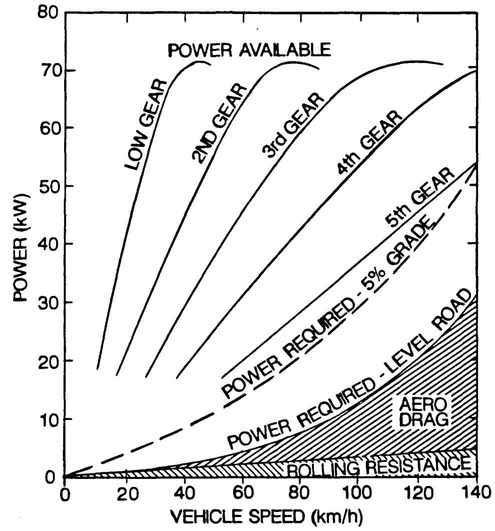

FIG. 14. Vehicle power requirement, and power available with a five-speed manual transmission.

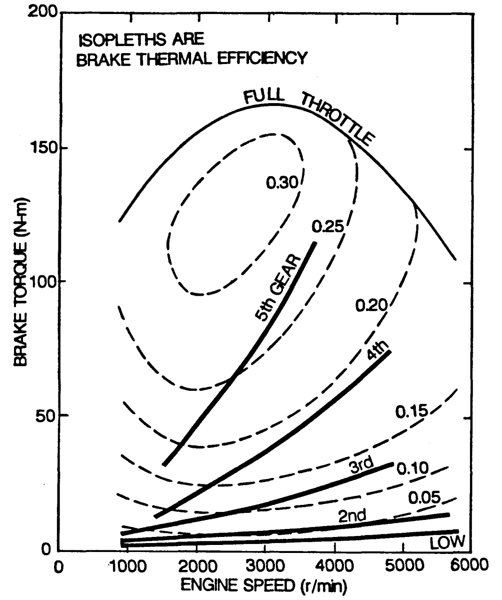

FIG. 15. Engine torque requirement at road load in various gears, superimposed on an engine map.

ratio provided by the drivetrain when the transmission is operated in the indicated gear.

6. ENVIRONMENTAL ASPECTS

At the turn of the century the fledgling automobile was welcomed as a relief from the problems of solid-waste disposal, odor, and flies associated with the ubiquitous horse. A half-century later the deteriorating air quality in southern California, fostered by climate, topography, and population density, initiated the era of environmentalism. Los Angeles smog correlated with atmospheric ozone concentration, and ozone formation was traced to the warm-weather action of sunlight on nitrogen dioxide and certain hydrocarbons mixed in proper proportions (see AIR POLLUTION: COMPONENTS, CAUSES, AND CURES). Automobile exhaust was identified as one of the anthropogenic sources of nitrogen oxides (NO_x) and hydrocarbons (HC). Legislated control of emissions from automotive vehicles and other sources followed.

An early improvement was to route hydrocarbon vapors in the crankcase, originally vented to the atmosphere, into the engine intake instead through a positive crankcase ventilation (PCV) system. This move eliminated about 20% of the HC emitted by an uncontrolled engine.

Initial regulation of tailpipe emissions limited only HC and carbon monoxide (CO), a toxic gas. About 60% of the HC from an uncontrolled car and all of the CO was coming from the tailpipe. Ultimately NO_x, which appears exclusively in the exhaust gas, was added to the list of regulated tailpipe emissions. The dominant oxide of nitrogen is nitric oxide (NO), but this later oxidizes in the atmosphere to nitrogen dioxide (NO_2). Consequently, all oxides of nitrogen in the exhaust are considered to have the molecular weight of nitrogen dioxide in measuring emissions gravimetrically for regulatory purposes.

The final 20% of the HC from an uncontrolled car comprised evaporative losses from the fuel system, notably the carburetor float bowl and the fuel tank. Vapor from the fuel system is now captured in a canister of activated charcoal for subsequent consumption by the engine.

6.1 Sources of Tailpipe Emissions

Hydrocarbon emissions exhausted from the engine consist primarily of unburned fuel. Lubricating oil can also contribute, although that is a minor source in an engine maintained in good mechanical condition.

If the mixture is rich, insufficient oxygen is available to burn all the fuel, so HC emissions would be expected. Even at stoichiometric and leaner mixtures, however, when the theoretically required amount of oxygen is present, unburned HC is found in the exhaust gas.

Originally, such unburned HC was believed to result primarily from the quenching of the cylinder flame as it approached the walls of the combustion chamber. Later work has suggested that under most conditions, this quenched air-fuel mixture diffuses into the hot burned gas and oxidizes. Regardless of the extent to which this occurs, the chamber walls are contributors to HC emissions through two established mechanisms. First, the oil film on the cylinder walls absorbs molecules of unburned fuel. These molecules are desorbed as the pressure in the cylinder falls during the expansion stroke, at a time when most of the mixture volume in the chamber has already been consumed by the flame. Second, porous combustion-chamber deposits formed on the cylinder head and piston crown also absorb fuel molecules, releasing them too late for their complete oxidation.

Another source of HC emissions is fuel trapped in crevices. That fuel escapes the normal combustion process because the flame is unable to penetrate into a crack. The most significant crevice is formed between the piston periphery and the cylinder wall, above the uppermost piston ring. Fresh air-fuel mixture is forced into this crevice as the piston ascends on the compression stroke, and additional unburned mixture enters early during the expansion stroke as the cylinder pressure continues to rise toward its maximum. Once cylinder pressure begins to fall, this unburned crevice gas starts to reenter the combustion chamber, but too late to take part in the normal combustion process.

One more source of unburned HC is flame extinction, which is most likely to be significant if the mixture is lean or is heavily diluted with residual burned gas from an earlier cycle. In such cases, the mixture ahead of the flame that is traveling across the chamber is subjected to a falling pressure and temperature as the piston descends on the expansion stroke. The last part of this unburned mixture to be consumed by the flame is also exposed to the cooling effect of the chamber walls. These circumstances are unfavorable to combustion, and the flame may simply be extinguished before the last small fraction of air-fuel mixture is consumed.

All of these unburned molecules of fuel have the opportunity to oxidize into innocuous carbon dioxide and water vapor as they mix with the hot burned products of combustion, but due to a combination of inadequate mixing and insufficient time to complete the oxidation process, a portion escapes into the exhaust system as HC emissions. The qualitative trend of these exhaust-HC emissions with air-fuel ratio is illustrated for gasoline in Fig. 16, where the vertical dashed line denotes the stoichiometric mixture ratio. The HC emission rises at an increasing rate as the mixture is enriched from stoichiometric because of the shortage of oxygen. HC persists on the lean side of stoichiometric and ultimately rises as the mixture is leaned, despite the increased presence of oxygen, because the lower cylinder temperatures accompanying lean combustion are less favorable for post-flame oxidation, and eventually because of the increased incidence of flame quenching before the flame has completed its travel across the chamber.

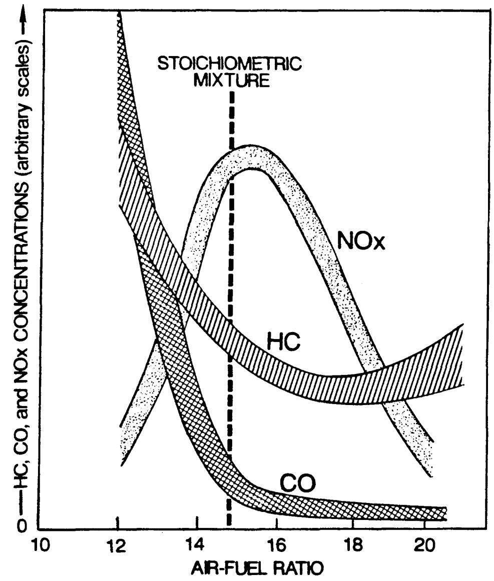

FIG. 16. Effect of air-fuel ratio on trends in engine-out emission concentrations. Fuel is gasoline.

Carbon monoxide also trends upward in Fig. 16 as the mixture is enriched, again because there is insufficient oxygen present for complete conversion of the fuel carbon to carbon dioxide. CO emissions are seen to be low in lean mixtures, but not zero. This product of incomplete combustion is to be expected as long as partial oxidation of HC is occurring late in the cycle and in the exhaust port.

The NO_x trend with mixture ratio is also shown in Fig. 16. NO_x, primarily nitric oxide, is an equilibrium constituent of air heated to a high temperature. The maximum combustion temperature is essentially reached at the stoichiometric ratio. Progressing richer from stoichiometric, temperature is falling because of incomplete combustion of fuel and because fuel carbon and hydrogen are competing with nitrogen for the short supply of oxygen present. Consequently, the NO_x concentration falls. On the lean side, dilution with excess air also lowers the temperature, but there is an excess of available oxygen. As the mixture is leaned from stoichiometric, the tradeoff between falling temperature and increasing availability of oxygen causes NO_x production to peak at a slightly lean mixture ratio.

6.2 Tailpipe-Emission Control

Nationwide regulation of tailpipe emissions began in the U.S. in 1968. Emissions are measured while driving the car on a chassis dynamometer over a prescribed transient schedule representative of urban operation. The dynamometer loads the driving wheels as a function of car speed and acceleration in a manner simulating driving on an actual road.

Fuel economy is evaluated on a second driving schedule, representative of highway driving. Fuel-economy measurements on the urban and highway schedules are combined to establish a vehicle fuel economy for regulatory purposes, but the highway schedule is not used in assigning emissions levels.

In the early years of emission regulation, tailpipe emissions were decreased by such measures as lowering compression ratio, retarding ignition timing, and departing from established air-fuel ratio schedules. When NO_x was added to HC and CO as a regulated emission in 1973, EGR (exhaust gas recirculation, the recycling of a fraction of the engine exhaust back into the intake manifold) was introduced to lower combustion temperature through dilution of the fuel-air charge. These techniques compromised fuel economy and drivability.

In 1975 the oxidizing catalytic converter was introduced on a broad scale. This converter consists of a sheet-metal container inserted in the exhaust system that exposes the engine exhaust gas to minute quantities of catalytic materials, typically platinum and palladium, which are capable of catalyzing the oxidation of HC and CO. The catalyst is dispersed on the surfaces of either pellets comprising a bed through which the exhaust gas flows, or a monolith structure of parallel passages through which the exhaust gas flows. Given the presence of oxygen and a temperature above 300–400 °C, such a catalyst can eliminate over 90% of the entering HC and CO. In some systems, an engine-driven air pump has been added to provide the required oxygen. Engineering the system to ensure that the catalyst is heated quickly to its functional temperature is important for achieving low tailpipe emissions. The availability of the catalytic converter eased the requirements on engine-out emissions, so engines could be retuned for better fuel economy and drivability.

In 1981 the federal NO_x standard was lowered to the point where its catalytic reduction became necessary. This led to addition of rhodium to the converter, making it a three-way device controlling HC, CO, and NO_x.

Catalytic reduction of NO_x becomes ineffective in the presence of excess oxygen, but oxidation of HC and CO cannot occur in the absence of oxygen. Consequently, air-fuel ratio had to be controlled within a very narrow window about the stoichiometric ratio to achieve high conversion efficiency of all three regulated emissions. This requirement is beyond the capability of the traditional carburetor.

The need for precise mixture regulation introduced the era of electronic engine control. An on-board computer tells the fuel system how much fuel is needed to provide the proper air-fuel ratio. The computer calculates this requirement either from stored tables of engine characteristics as a function primarily of engine speed and intake-manifold absolute pressure, or from the signal of an intake-airflow meter. An oxygen sensor in the exhaust stream, which responds too slowly to be used for real-time fuel control, signals the computer if the mixture ratio strays from the stoichiometric value. This facilitates a continuing calibration of the fuel-control system to guard against long-term drift.

In order to regulate the fuel rate electronically, its flow is controlled by a pulse-width-modulated solenoid. The solenoid-controlled fuel orifice may be incorporated in an electronic carburetor, but the growing preference is for replacing the carburetor with fuel injection. Two principal options are used. With single-point injection, the fuel nozzle is located at the entrance to the intake manifold, where the carburetor is normally placed. With multipoint injection, typically port injection, a separate nozzle is located in the intake port of each cylinder.

6.3 Efficacy of Controls

U.S. emissions standards have been lowered over time to the point where a new 1990 car must emit 96% less HC and CO, and 76%

less NO_x, than a 1960s precontrol car did. New U.S. cars are designed to satisfy emission standards at 80 500 km and are subject to recall for correction in the case of high failure rate.

From estimates by the Environmental Protection Agency (EPA) for 1970 and 1987 model-year cars in use at 80 500 km, aggregate HC and CO emissions were actually decreased by 80%, and NO_x by 70%, over that 17-year time interval after accounting for the 48% increase in vehicle miles traveled; yet over that same time span, the EPA composite average ozone and CO concentrations decreased by only about one-third (Atkinson *et al.*, 1990). Ozone depends on both HC and NO_x, but is not linearly related to them. The relationship between urban CO and CO sources is more direct, however.

There are a number of other reasons why average urban ozone and CO have not decreased in proportion to the reduction in tailpipe emissions from new cars. For example, for all of the listed emissions, the automobile is not the only source. Moreover, the atmospheric inventory depends on all cars on the road, not just those from a specified model year. Much of the in-use fleet consists of older cars built when emissions standards were less stringent, thus allowing them to emit more than newer cars. In this connection, the average age of the U.S. in-use fleet has increased steadily over recent years (Motor Vehicle Manufacturers Association, 1989).

Some cars on the road are improperly maintained or have been misfueled with leaded gasoline, which deactivates the catalytic converter and increases tailpipe emissions markedly. Just one such car can emit as much HC, CO, and NO_x as many properly maintained cars.

The possibility also exists that the test used to assess vehicle emissions is not representative of real-world driving. One known factor affecting total HC emissions has been the increase in gasoline vapor pressure over the past two decades (National Institute for Petroleum and Energy Research, 1989), which increases evaporative emissions. As a result, gasoline vapor pressure is destined for regulation.

7. FUELS

The performance, fuel economy, and emissions of the automotive propulsion system all depend on the fuel consumed. Gasoline derived from petroleum has long been the dominant automotive fuel, although petroleum-derived diesel fuel has been required for the compression-ignition diesel engine. Recently, interest has arisen in broadened use of gaseous fuels or alcohols because of anticipated environmental advantages.

7.1 Liquid Hydrocarbons

Commercial gasoline is a mixture of many different liquid-hydrocarbon compounds. It has an average atomic H/C ratio of 1.5–2.0, with a typical value of 1.85 and a lower heating value of 43.5 MJ/kg. The individual hydrocarbon compounds found in gasoline are classified by family.

Alkanes, or paraffins, are saturated hydrocarbons that have a chemical composition of C_nH_{2n+2}. In a normal-paraffin molecule, the carbon atoms are arranged in a straight chain. Isomers of paraffins are more compactly arranged molecules having carbon-containing branches attached to the main carbon chain. Naphthenes are cycloparaffins in which at least part of the carbon chain is arranged in the form of a closed ring. They have the chemical formula C_nH_{2n}.

In normal combustion of a homogeneous charge, as the flame progresses across the chamber, the unburned mixture ahead of it is compressed to a high pressure and temperature. If this unburned-mixture fraction is at these conditions too long before being consumed by the flame, it ignites spontaneously, causing a sudden rise in cylinder pressure and a noise known as combustion knock. For a specified fuel, knock limits the usable compression ratio, hence thermal efficiency.

The octane rating of a fuel marks its resistance to knock. Long-chain normal paraffins have comparatively low octane ratings, but a more compact isomer having the same chemical formula has a higher octane rating. By definition, an octane number of 100 is assigned to isooctane, an isomer containing eight carbon atoms. Normal-heptane, comprising a straight seven-carbon chain, has an octane number of zero. Commercial gasolines fall between these extremes. Two standardized tests are used to assess the knock resistance of gasoline, one giving the research octane number (RON) and the other the motor octane number (MON). In the U.S., the average of these, $(R+M)/2$, generally falls in the range 85–95.

For more than 60 years, tetraethyl lead was added to gasoline in small quantities to enhance its resistance to knock. That practice is being phased out in the U.S., both to avoid adding lead to the atmosphere and to preserve the exhaust catalyst, which is poisoned by lead.

Olefins resemble paraffins but are unsaturated, evidenced by a double bond between one pair of carbon atoms in the chain. Olefins are of the form C_nH_{2n}. Diolefins contain two double bonds and have the formula C_nH_{2n-2}. Double bonds characterize an unsaturated molecule that is more prone to gum formation than a saturated molecule.

In the aromatic family, which has the formula C_nH_{2n-6}, at least some of the carbon atoms are arranged in rings. Aromatics are unsaturated but have high octane ratings. Benzene, an aromatic, is considered toxic.

Like gasoline, diesel fuel is a mixture of hydrocarbons, but is less volatile. Because ignition in the diesel engine is caused by compression of the charge rather than a spark, a good diesel fuel should be autoignition prone. Thus, a good diesel fuel knocks too readily for use in a spark-ignition engine, and conversely, gasoline makes a poor diesel fuel. Cetane number is a measure of the propensity of a fuel to ignite when sprayed into air compressed in a standardized test engine. On average, U.S. automotive diesel fuels have a cetane number of about 45.

7.2 Gaseous Fuels

Spark-ignition engines can be adapted to operate on gaseous fuels. Those commonly used are natural gas and liquefied petroleum gas (LPG).

Natural gas is extracted from underground like petroleum. It varies somewhat in composition but is usually about 90% methane. Methane (CH_4) is the lightest member of the paraffin family. It has a high octane rating but is difficult to carry on board a vehicle. Normally it is stored at high pressure (165–200 atm) and called compressed natural gas (CNG). Even at these high pressures its volumetric energy density is so poor compared to gasoline that vehicle range is severely curtailed. Energy density is improved by storing it as a liquid (LNG), but the liquid is cryogenic, introducing a new set of problems for vehicular use.

LPG is a by-product of petroleum refining and natural gas production and, as such, is not as abundantly available as natural gas. LPG used as an engine fuel is mostly propane, a normal paraffin having the formula C_3H_8. It, too, has a high octane rating. Its volumetric energy density is higher than that of natural gas, but not as high as that of gasoline.

Hydrogen is the least dense of the gaseous fuels. After making appropriate engine changes, it can be burned in a spark-ignition engine. It is not presently available in sufficient quantity at a reasonable price to be used broadly as an engine fuel, and it has many handling and storage problems. As a compressed gas, its volumetric energy density is lower than that of natural gas. If liquefied, it is cryogenic. It has been stored in a metal hydride, being released when the hydride is heated, but this storage technique carries a heavy weight penalty.

7.3 Alcohols

The lowest-molecular-weight alcohols, methanol (CH_3OH) and ethanol (C_2H_5OH), are used as engine fuels. The volumetric energy densities are approximately 1/2 and 2/3 that of commercial gasoline for methanol and ethanol, respectively. Both alcohols have high octane ratings and high latent heats. The high latent heat cools the charge in the intake manifold during evaporation of the alcohol, enabling the engine to induct more air for increased output power. If the pure alcohols are used as fuels, this cooling effect combines with a low vapor pressure to make starting very difficult in even modestly cool weather. For this reason, a blend of 85% methanol and 15% gasoline, known as M85, has been advocated by some to ease the starting problem.

Methanol is made primarily from natural gas, although it can be made from biomass sources such as wood. Ethanol is biomass-sourced, being made from sugar cane in Brazil and from corn in the U.S. Under existing technology, the high cost of ethanol makes it unattractive in the U.S. as a primary fuel. However, it has been blended into gasoline in proportions up to 10% to enhance octane rating. This blend is commonly called gasohol. Methanol has also been blended into gasoline, but at a lower percentage. This requires addition of a cosolvent to avoid phase separation. If blends containing higher percentages of alcohol in gasoline, or if neat alcohols, are considered for use in automotive propulsion, the fuel system must be revised because of the damaging effect that alcohol, especially methanol, has on many commonly used materials.

7.4 Emissions Prospects of Alternative Fuels

Interest in alternatives to commercial gasoline and diesel fuel has been fostered by the failure of many urban areas to meet federal air-quality standards. Fuels receiving primary attention are methanol made from natural gas and natural gas itself.

It has been suggested by some that methanol would reduce smog in urban areas because engine emissions of unburned fuel should be mainly methanol, which has a relatively low reactivity in the chemical reactions responsible for smog compared to many constituents in gasoline exhaust. Unfortunately, however, the concentration of formaldehyde in the products of methanol combustion is many times that from gasoline combustion, and formaldebyde is very reactive. So far, formaldehyde has proven stubbornly resistant to catalytic elimination at high mileage. The ultimate result of the use of methanol fuel, or the more practical M85, on urban air quality remains unsettled.

During any transition to methanol as an automotive fuel, limited availability favors the flexible- or variable-fuel vehicle capable of operating on M85, neat gasoline, or any mixture between these extremes. The vapor pressure of a blend of 10% methanol in commercial gasoline is higher than the vapor pressure of either the gasoline itself or M85, creating concerns about meeting more stringent evaporative-emission standards. In addition, in short-trip and winter driving, upper-cylinder wear has proven more severe with methanol and its blends than with gasoline.

Some suggest that it is foolish to convert natural gas to methanol and suffer significant energy loss in the conversion process when natural gas itself is a suitable fuel for spark-ignition engines. The primary unburned-fuel constituent in the exhaust is methane, which has such a low reactivity that for regulatory purposes it is often excluded from the exhaust-hydrocarbon inventory. Moreover, the formaldehyde emission is low, and evaporative emissions are eliminated. Natural gas has a major problem with on-board storage, however. It is customarily stored as a compressed gas at 165–200 atm, but even at such high pressures its energy density is such that vehicle range is cut to 20%–25% of its value on gasoline if the vehicle space allotted to fuel storage is fixed. For this reason, natural gas is more often favored for light- or medium-duty trucks, where it is easier to find space for additional tankage, than for passenger cars.

8. ALTERNATIVE ENGINES

Despite the dominance of the spark-ignition engine in the automotive field, alternative engines exist and either are being used to a lesser extent or are the subjects of research and development in the interest of improved fuel economy and/or lower exhaust emissions.

8.1 Diesel Engine

The cycle of events employed by the diesel engine duplicates that of the spark-ignition engine. Both the four-stroke and the two-stroke cycle are used, although the former predominates. The fundamental differences between spark-ignition and diesel engines center on the methods of charge preparation and ignition.

In the diesel engine, fuel is not introduced upstream of the intake valve. Instead it is sprayed into the combustion space through a high-pressure injection nozzle late during the compression stroke, shortly before the start of combustion is desired. Ignition occurs spontaneously at some point(s) in the fuel spray, without help from a spark plug, as a result of both the elevated compression temperature and the distinctive properties of diesel fuel. Combustion continues during injection of the remainder of the fuel. To ensure compression ignition of the fuel, the diesel engine uses a higher compression ratio than the spark-ignition engine, typically in the range 17–21.

The fuel economy of a diesel-powered vehicle, expressed as the distance traveled per unit volume of fuel consumed, exceeds that of a comparable vehicle with a spark-ignited gasoline engine for three reasons: the diesel employs a higher compression ratio than the knock-limited ratio tolerated by the traditional gasoline engine, typical diesel fuel contains about 13% more energy per unit volume than gasoline, and diesel-engine load is controlled by varying the fuel injected into the cylinder air rather than by throttling the intake air. The last feature replaces the part-load pumping penalty of the traditional automotive engine with the thermodynamic advantages of a lean average air-fuel ratio.

In contrast to the homogeneous air-fuel charge of the spark-ignition engine, the diesel operates with a stratified charge, the air-fuel ratio varying from zero at the injector-nozzle exit to infinity in the air surrounding the spray plume. Exposure of the very fuel-rich mixture in the core of a spray plume to the high combustion temperatures of the adjacent burning mixture promotes the formation of soot, minute particles comprised primarily of carbon atoms. Most of this soot oxidizes as combustion progresses, but a significant quantity escapes with the exhaust gas. In the process it attracts unburned HC. The exhaust soot and adsorbed HC, lubricating oil escaping from the cylinder, and sulfates formed from sulfur in the fuel, together with the water molecules associated with these sulfates, collectively comprise the exhaust particulate matter (PM). PM is a regulated exhaust emission in the U.S. Although the exhaust HC of the diesel is fairly low without aid from an exhaust catalyst and the CO is very low, simultaneous compliance with PM and NO_x standards is a challenge. For example, EGR (exhaust gas recirculation) is an effective means of lowering NO_x at part load, but its introduction increases PM.

Broadly speaking, there are two types of diesel engine in use: the indirect-injection (IDI) and the direct-injection (DI) versions. They are contrasted in Fig. 17.

In the IDI version, fuel is injected into an antechamber connected to the main chamber above the piston by a narrow passageway. Bulk air motion established in the antechamber fosters mixing of the air with the fuel spray. Upon ignition, the burning mixture expands through the passageway into the air-filled main chamber, where combustion is completed. The IDI engine is the type used in most diesel passenger cars because of advantages in emissions and noise production. It normally has a compression ratio greater than 20 to help compensate for the increased heat loss associated with the high surface/volume ratio of its combustion chamber. The antechamber also includes a glow plug for aid in starting.

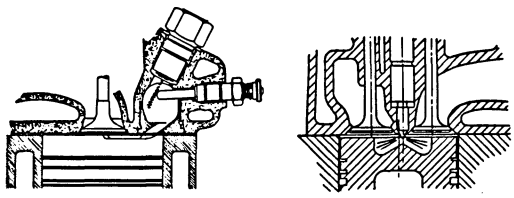
INDIRECT INJECTION DIRECT INJECTION

FIG. 17. Two types of diesel engines.

In the DI version the fuel is injected directly into the cylinder. In the version of Fig. 17 the piston contains a deep bowl. Swirl induced in the intake air is intensified through conservation of angular momentum as the piston ascends on the compression stroke, creating bulk air motion in the bowl that promotes mixing of air and fuel. Such a DI diesel is being pursued for passenger cars because it has the potential for a 10%–15% increase in fuel economy compared to the IDI engine, but noise and emissions compliance have impeded its broad application.

In a variation on the DI engine of Fig. 17, the bowl is eliminated from the piston crown, providing a more quiescent field of air. Mixing is then promoted by using very high fuel-injection pressures, typically combined with a multiplume spray pattern. DI diesels of this type are common in heavy-duty trucks.

8.2 Spark-Ignition Stratified-Charge Engine

Spark-ignition versions of the compression-ignition stratified-charge diesel engine have been the subject of considerable attention over the years.

A divided-chamber stratified-charge engine bearing a family resemblance to the IDI diesel has seen limited production in passenger cars. A rich mixture is carbureted into the antechamber, a lean mixture into the main chamber. Spark ignition of the rich mixture sends a flaming torch of burning gases into

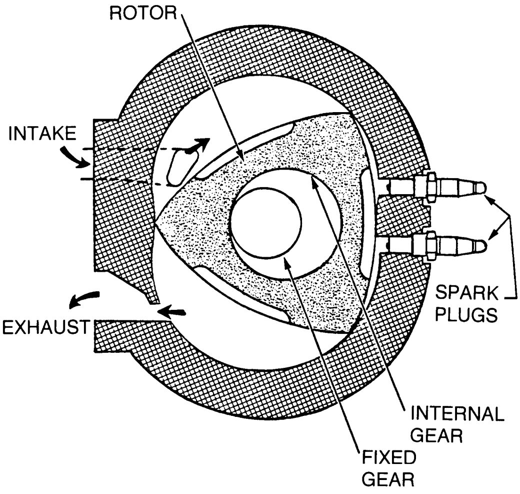

FIG. 18. Rotary engine.

the main chamber for ignition of the lean mixture. This approach has not shown sufficient advantage to earn widespread acceptance.

Two types of spark-ignited DI engines have been the subject of extensive research. Members of this class are frequently called DISC (direct-injection stratified-charge) engines. In the late-injection type, fuel is injected into the cylinder late during the compression stroke as in the diesel, and the spark plug has typically ignited the mixture before injection is completed. In the early-injection type, fuel is injected early during the compression stroke, and fuel injection is typically completed before the spark plug ignites the mixture. Both approaches have demonstrated fuel-economy advantages, but both have difficulty matching the exhaust emissions of competitive engines.

8.3 Rotary Engine

Many rotary engines with intermittent combustion have been invented over the years to avoid the vibration associated with the conversion from reciprocating to rotary motion accomplished by the slider-crank mechanism in the conventional reciprocating engine. However, only the one due to Wankel that is illustrated in Fig. 18 has achieved significant success commercially. A three-sided member rotates eccentrically within a specially shaped housing. Three chambers are enclosed between the flanks of the rotor and the interior of the housing. They change in volume as the central element rotates. Each of these variable-volume chambers is equivalent to a cylinder. The fresh mixture is inducted, and the burned gas exhausted, through ports in the housing. Two spark plugs are shown fixed in the housing, and the charge is swept past them by the rotor. The use of two spark plugs in a single combustion space reflects an effort to compensate for the undesirably long and narrow combustion chamber inherent with this configuration. The engine is compact and light, comparatively quiet, and vibration free, but generally has been unable to match the fuel economy of a comparable reciprocating engine.

8.4 Gas Turbine

The automotive gas turbine, diagrammed in Fig. 19, duplicates the compression, combustion, and expansion events of the reciprocating engine, but on a continuous rather than an intermittent basis. Intake air is compressed by a centrifugal compressor. The compressed air passes through a regenerative heat exchanger for preheating by the exhaust gas before entering the burner. There combustion of a fuel spray takes place on a continuous basis at an essentially constant pressure. The gas leaving the burner is at the peak temperature of the cycle, a parameter strongly influencing the thermal efficiency. This hot compressed gas expands first through a gasifier turbine that is connected on a common shaft to the compressor and provides the energy necessary to drive it. Expansion continues through a power turbine that is connected through a transmission to the vehicle-drive wheels. The hot gas leaving the power turbine, now at near-atmospheric

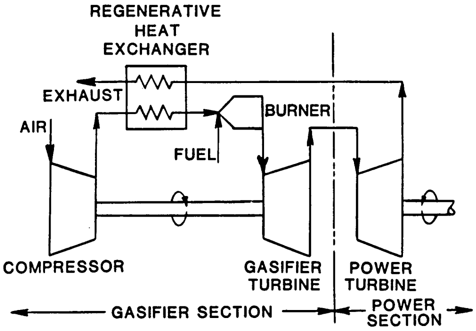

FIG. 19. Schematic diagram of a two-shaft regenerative gas turbine.

pressure, rejects heat to the compressor-discharge air in the heat exchanger before being exhausted to the atmosphere.

The regenerative heat exchanger usually takes the form of a porous matrix in a large-diameter thin disk. The opposite face areas of the disk are divided into high-pressure and low-pressure sectors by rubbing seals fixed to the housing and radiating from the disk center. As the disk slowly rotates on its axis, any given element of the matrix alternately passes through each of these two sectors. As the matrix element rotates through the low-pressure sector, hot gas exiting from the power turbine passes through it to the exhaust system, heating the matrix element in the process. When that matrix element rotates past the seals into the high-pressure sector, cooler compressor-discharge air passes through it in the opposite direction, extracting heat from the matrix en route to the burner.

The presence of the regenerator distinguishes the automotive gas turbine from its aircraft counterpart. The regenerator saves fuel by preheating the burner inlet air so that less fuel is required to achieve the peak cycle temperature. It cools the engine exhaust to a level more acceptable for disposal on streets and highways. It also helps to silence the exhaust noise.

The usual two-shaft automotive gas turbine has been illustrated in Fig. 19, but single-shaft versions have also been built experimentally in which the gasifier and power turbines have been combined into a single unit. This simplifies the engine but demands an efficient continuously variable transmission to accommodate the difference between the engine-shaft speed, which is related to the instantaneous power demand of the driver, and the vehicle speed, which is not proportionately related.

Past automotive gas turbines have been unable to match the fuel economy of existing piston engines, one reason being the limitation placed on peak cycle temperature by existing high-temperature alloys. Research is now directed toward use of structural ceramics in the hot sections of the engine in an effort to remove that barrier. Obstacles yet to be overcome include proof of reliable cost-competitive ceramic components, achievement of satisfactory efficiencies in the miniature turbomachinery that follows from the low engine airflow requirement attending high peak cycle temperature, and development of a burner that can meet low NO_x requirements at high operating temperature.

8.5 External-Combustion Engines

The foregoing engines all involve internal combustion, i.e., combustion in through-flowing air. The steam engine and the Stirling engine utilize external combustion, with the heat generated being transferred into the working fluid through a heat exchanger. The working fluid then circulates in a closed cycle.

In the steam engine, the working fluid is water in both its liquid and vapor phases. The steam engine works on the Rankine cycle. Water is pressurized by a pump, then passed into a steam generator or boiler, where it is heated to its boiling point, vaporized, and superheated, all at essentially constant pressure. The hot steam then passes through an expander of either the reciprocating or turbine type, although the former has proven more effective in automotive sizes. The expander extracts work from the steam, after which the low-pressure steam passes to the condenser. There the steam is condensed back into a liquid at constant pressure for passage to the pump and recirculation through the cycle. Condensation is accomplished by passing ambient air through the condensing heat exchanger.

The preferred working fluid for the Stirling engine is hydrogen, although helium has been used with some sacrifice in thermal efficiency. In contrast to the steam engine, no phase change is involved. In the ideal Stirling cycle, the gas is first compressed isothermally, then heated at constant volume (isochorically) to a still higher pressure, next expanded isothermally, and finally cooled to its original state isochorically. The heat rejected from the gas during isochoric cooling is stored in a regenerator for subsequent transfer into the gas during its isochoric heating. It can be shown that this ideal cycle has as high a thermal efficiency as can be achieved when operating between specified upper and lower temperature limits. In implementing this Stirling cycle in hardware, reciprocating pistons are employed, but the heat addition is accomplished through continuous combustion.

Being external-combustion engines, both of these alternatives can be operated on a broad spectrum of fuels. Because they are free of explosive combustion, both avoid the combustion noise of reciprocating internal-combustion engines. However, both of these closed-cycle engines share some common drawbacks for automotive application, although the more efficient Stirling engine comes the closer to meeting recognized needs. Both are large and heavy. Upon starting a cold engine, both require an annoying waiting period until the working fluid reaches an acceptable operating temperature. Especially in light-load operation, the brake efficiencies of both suffer because of the energy requirements of the large number of engine-required auxiliaries. The steam engine is poorly suited to operation in winter climates because of the freezing problem. The Stirling engine has an unsolved problem with leakage of the hydrogen working fluid.

9. ELECTRIC VEHICLES

Because of problems with local air quality in many urban areas, interest is strong in the electric vehicle. Leading candidates are the battery-electric, in which the conversion from fuel-stored energy into electricity is accomplished at a remote location, and fuel cells, in which that conversion takes place on board the vehicle.

9.1 Battery-Electric Vehicle

It has been estimated that use of an electric car in place of one powered by an internal-combustion engine would reduce HC and CO emissions by one to two orders of magnitude. On the other hand, a significant reduction in NO_x is unlikely, although the NO_x would usually be emitted at an electric generating plant remotely located from any urban center. In trade, a substantial increase in emissions of particulate matter and sulfur oxides is anticipated, even with expected environmental controls, because over half of the electricity generated in the U.S. will involve combustion of coal on into the 21st century.

Energy efficiency is high from the battery through the control system and electric motor to the drive wheels, but that figure is not of great significance. When the efficiency is reckoned from the primary-energy source through the steam power plant and its electric generator, the electricity delivery system, the battery charger, and into the battery, studies have suggested no significant energy advantage for the battery-electric vehicle.

Because of presently limited battery capabilities, the utility of the battery-electric vehicle is restricted. A maximum range between charges on the order of 150 km is typical. Range falls quickly if battery energy is used to provide heating or air conditioning, and it also falls with ambient temperature and with battery age. As the battery charge is consumed, vehicle performance deteriorates. When repeatedly deep discharged to utilize maximum range, battery life suffers. While an empty gasoline tank can be refilled in a matter of minutes, it takes hours to recharge a battery. Fast recharging also depreciates battery life. For this reason, the battery-electric approach is presently best suited to commuter cars and urban delivery vans.

Although energy cost for a battery-electric car can be competitive, battery replacement cost presently tilts economics away from this option. Future substantial increases in gasoline prices could reverse this balance. Clearly, what is needed is an improved battery.

Research and development are being conducted on many types of batteries. Some batteries are superior to others in terms of power density (W/kg), a measure of performance capability. Others are best in terms of energy density (W h/kg), indicative of range capability. All would benefit from the ability to accept more charge/discharge cycles before requiring replacement.

A candidate promising one of the best combinations of energy, power and life is the sodium/sulfur battery, which operates at about 350 °C with the elements in a molten state. Such high-temperature batteries employing reactive materials involve unresolved questions about safety in the event of serious accidents and about freezing after long periods out of service, especially in winter climates.

9.2 Engine/Electric Hybrid Vehicle

The engine/electric hybrid, which adds a combustion engine to the battery-electric system, can be used to overcome both the range handicap and the normally restricted performance capability of the battery-electric vehicle. The engine and the battery-electric system can be combined in a variety of arrangements.

The simplest involves addition to the battery-electric vehicle of a small engine driving a generator to recharge the batteries as energy is extracted from them. In such a range extender, vehicle range is determined by the size of the fuel tank rather than the energy capacity of the battery pack. Performance is still limited by the energy-delivery rate of the batteries, however.

At the expense of increased complexity, the engine can be linked to the drive wheels along with the electric motor, while still driving a generator to keep the batteries charged. While extending vehicle range as in the above case, this arrangement also enhances performance capability from a given battery pack by augmenting the propulsive power available.

A third arrangement is to link both the engine and the electric motor to the drive wheels as above, but to eliminate the engine-driven generator for battery recharging. This option also has the potential for increased range over the simple battery-electric vehicle, with range again being determined by the size of the fuel tank. A smaller engine is often proposed with this arrangement, allowing the engine to operate nearer full load, where it has a higher thermal efficiency. The shortfall in performance is then made up by the power available from the battery-electric system. Although this arrangement may offer an advantage in fuel economy over a vehicle of equal performance powered solely by an engine, some utility is sacrificed in the general case. For example, on a prolonged upgrade with partially discharged batteries, the electric motor may no longer be able to compensate for the performance shortfall of the undersized engine.

Many variations exist in the details of implementing the above schemes. In addition, hybrids have been demonstrated in which an engine is coupled with either mechanical-energy storage in a flywheel or hydraulic/pneumatic-energy storage in a hydraulic accumulator rather than in a battery. All of these possibilities share the burden of increased complexity, weight and cost, and of reduced carrying capacity because of the increased space requirement of the hybrid system. For the latter reason, such hybrid systems are often considered for vans and small trucks rather than for the severely space-restricted passenger car.

9.3 Fuel-Cell Vehicle

The fuel cell is not as developed as the battery for vehicular application. It offers a potential solution to the range problem of the battery-electric option. Currently, the leading approach is to break down methanol in an on-board reformer, extracting hydrogen. This fuel is then reacted in the fuel cell with oxygen from the air to make electricity for the electric drive motor(s). Because such a system is unable to follow the fast transients characterizing typical automotive power requirements, it must be combined with batteries for energy storage. The batteries also allow recovery of vehicle kinetic energy through regenerative braking. The energy density of such a system is presently too poor for serious consideration in automobiles, but the technology is relatively immature, leaving room for future improvement.

GLOSSARY

Brake Power: The power delivered by an engine at its output shaft.

Brake Specific Fuel Consumption (BSFC): The mass of fuel consumed per unit of brake power delivered.

Differential: The mechanism incorporated in the drive train to allow the vehicle drive wheels to rotate at different speeds, as when turning a corner.

Drive Train: The components connecting the engine output shaft and the drive wheels, normally including a transmission and a differential.

Dynamometer: A device, usually an electric generator, capable of loading an engine (an engine dynamometer) or vehicle (a chassis dynamometer) at various torques and speeds for test purposes. An electric dynamometer capable of operating also as a motor to rotate the engine or the vehicle drive wheels when the engine is not delivering power is called a motoring dynamometer.

Exhaust Gas Recirculation (EGR): Recycling of a fraction of the engine exhaust gas back into the intake manifold, done to control the emission of nitrogen oxides.

Exhaust Manifold: A branched conduit that collects the exhaust gas from the individual cylinders and delivers it to the exhaust pipe.

Four-Stroke Engine: A reciprocating piston engine that completes a thermodynamic cycle in any given cylinder during four strokes of its piston (two revolutions of the crankshaft).

Indicated Power: The power developed by the cylinder gas during the compression and expansion strokes of the piston.

Intake Manifold: A branched conduit that delivers the engine intake air to the individual cylinders. The air delivered may also contain fuel, depending on the type of engine and its fuel system.

Intercooler: A heat exchanger located between a supercharging compressor or a turbocharger compressor and the engine, used to cool the compressed cylinder charge. The cooling medium may be either engine coolant or ambient air.

Journal Bearing: A lubricated sleeve in which a shaft or pin rotates or oscillates.

Knock: In a spark-ignited homogeneous fuel-air charge confined by an enclosure, an abnormal combustion phenomenon in which part of the unburned charge remote from the spark plug and normally the last to be reached by the advancing flame front ignites spontaneously.

Lean Mixture: Fuel and air mixed in proportions such that there is an excess of air present compared to that required for complete combustion of the fuel.

Mechanical Efficiency: The ratio of brake power to indicated power, a measure of losses attributable to friction and pumping.

Multipoint Fuel Injection: Spraying of fuel into the engine air in the intake system at a multiplicity of points, normally at the entrance port to each cylinder.

Octane Number: A numerical rating given to a gasoline to indicate its resistance to combustion knock and determined by one of several standardized test procedures.

Overhead Cam (OHC): Indication that cams operating the intake and exhaust valves are mounted above the cylinder head, requiring that the camshaft be driven from the crankshaft by a belt, a chain, or gears.

Oxidizing Catalytic Converter: A device in the exhaust system through which the exhaust gas flows. It contains minute amounts of special metals that catalyze the oxidation of both unburned hydrocarbons and carbon monoxide.

Power Train: The combination of the engine and the drive train.

Pumping Loss: The parasitic loss associated with pumping engine air into the cylinder and exhaust gas out of the cylinder during the intake and exhaust strokes.

Rich Mixture: Fuel and air mixed in proportions such that there is an insufficiency of air present compared to that required for complete combustion of the fuel.

Scavenging: Removing burned gas from the cylinder during the exhaust process to make room for fresh charge for the next cycle.

Single-Point Fuel Injection: Spraying of fuel into the engine air at a single point at the entrance to the intake manifold. (In vee engines, one injector may be used for each bank of cylinders.)

Stratified Charge: A cylinder charge in which the local ratio of fuel to air spatially varies significantly within the chamber, normally by intent.

Supercharger: An engine-driven compressor upstream of the intake manifold used to compress the air above ambient pressure for increased engine power.

Thermal Efficiency: The dimensionless ratio of the output energy from an engine to the fuel input energy.

Three-Way Catalytic Converter: A device in the exhaust system through which the exhaust gas flows, that device containing minute amounts of special metals that catalyze the oxidation of both unburned hydrocarbons and carbon monoxide, and that simultaneously reduce the oxides of nitrogen.

Throttle-Body Fuel Injection: Another commonly used name for single-point fuel injection, denoting that the injector is mounted in the element housing the throttle plate.

Torque Converter: An element in the transmission that includes a centrifugal hydraulic pump on its input shaft, a radial-inflow hydraulic turbine on its output shaft, and at least one row of stator blades between them. It transmits power from the engine crankshaft to the vehicle drive wheels, smoothly accommodating a continuously variable speed ratio across it and multiplying engine torque when the speed of the output shaft is much less than that of the input shaft.

Transmission: An element placed in the drive train to facilitate changing the ratio of the engine rotational speed to the rotational speed of the vehicle drive wheels.

Turbocharger: A unit mechanically separate from the engine within which a turbine driven by engine exhaust spins a compressor mounted on a common shaft, that compressor delivering compressed air to the engine for increased power output.

Two-Stroke Engine: A reciprocating piston engine that completes a thermodynamic cycle in any given cylinder during two strokes of its piston (one revolution of the crankshaft).

List of Works Cited

Atkinson, D., Cristofaro, A., and Kolb, J. (1990), *Energy and the Environment in the 21st Century*, Cambridge, MA: MIT Energy Laboratory.

Caris, D. F., Nelson, E. E. (1959), *SAE Trans.* **67**, 112–124.

Motor Vehicle Manufacturers Association (1989), *Facts & Figures '89*, Detroit, MI.

Muranaka, S., Tagaki, Y., Ishida, T. (1987), *SAE Trans.* **96**, 4.526–4.536.

National Institute for Petroleum and Energy Research (1989), *Motor Gasolines, Summer 1988*, NIPER-158 PPS 8911.

Further Reading

Evans, R. L. (Ed.) (1986), *Automotive Engine Alternatives*, New York: Plenum. Collection of papers on various engine types, including conventional spark-ignition engine, gas-fueled diesel, turbocompound diesel, Stirling, stratified charge, and rotary.

Ferrari, N. A. and Tester, J. (Ed.) (1990), *Proceedings—Energy and the Environment in the 21st Century*, Cambridge, MA: The MIT Press. Collection of papers providing an overview of energy consumption and environmental impact.

Heywood, J. B. (1988), *Internal Combustion Engine Fundamentals*, New York: McGraw-Hill. Advanced text on engines and engine technologies.

Newton, K., Steeds, W., Garrett, T. K. (1983), *The Motor Vehicle*, London: Butterworths. Book discussing engines and transmissions.

Obert, E. F. (1973), *Internal Combustion Engines and Air Pollution*, New York: Intext & Row. Basic text discussing engine operation and testing, cycle thermodynamics, fuels, spark-ignition and diesel engines, emissions, and lubrication.

AVIONICS

See AERONAUTICS

BEAMS, PARTICLE, INTENSE

See INTENSE PARTICLE BEAMS

BETATRONS

DONALD KERST, *Physics Department, University of Wisconsin, Madison, Wisconsin, U.S.A.*

Introduction 347
1. **Physical Principles** 348
2. **Equations of Motion** 350
3. **The Radiation-Loss Energy Limitation** 353
4. **The Beam Current Limit** 353
5. **Injection** 355
6. **The Target, Beam Extraction** 356
7. **Field Errors and Their Effects** 357
8. **Applications** 359
9. **Chronological Developments** 362
Glossary 362
Works Cited 363
Further Reading 363

INTRODUCTION

The early methods of producing high-energy electron beams for x-ray applications required generating the full voltage to give the final electron energy and then applying this voltage to a vacuum tube into which a source of electrons at the negative electrode injected the particles. Energies or voltages of one or two million electron volts were possible, but the techniques for holding such high voltages on a vacuum tube were very difficult to extend to higher voltages. At the same time the energies desired for exploring the energetic transformations within the nucleus of atoms were an order of magnitude higher, and such voltages could not be reached with the electron accelerators of the 1930s.

Higher voltages were also needed for the medical application to x-ray therapy since it was realized that there would be a buildup of secondary ionization, the "dose" responsible for biological action, below the surface of the body at a depth which increased with x-ray energy. High skin dose imposed an undesirable limit on the dose which could be delivered to the tissue to be treated deep below the skin if voltages below one million were used. Furthermore, if electrons of sufficient energy could be produced, these electrons could be used directly on deep tissue instead of using the secondary electron ionization produced by x rays. The energy loss of an electron in penetrating waterlike tissue is about 2 MeV per centimeter provided that the electron is relativistic, that is, provided its energy is well over the value of the electron's rest mass, 511 kV. Consequently a 20-MeV beam of electrons would penetrate about 10 cm and create an ionization dose along the way. Since the radiation therapist wants to deliver the dose to such depths, electron energies adjustable from 10 to 40 MeV would be desirable for either x-ray production or for direct electron beam application.

For radiographic examination of thick metal objects, such as iron castings, to discover flaws, the thickness that can be penetrated by x rays or gamma rays depends upon the atomic number of the metal's elements and on the energy of the x-ray photons. There is an optimum energy due to the decrease of Compton absorption with increasing x-ray energy and the increase of pair production with increased energy. For iron the most penetrating photon energy is around 25 MeV, and for lead it would be only about 3 MeV.

It is evident that for the applications mentioned it would be desirable to have beams of electrons of several tens of MeV. Since the electrons would be highly relativistic at these energies, the cyclotron could not be used. The cyclotron was very well adapted to the acceleration of ions to these energies in the 10-MeV range which are far below the rest mass of the ions used. Thus the betatron method of electron acceleration, which can operate at classical and at relativistic energies, had immediate application when it was developed in 1940.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

The later development of the synchrotron principle by McMillan (1945) and Veksler (1945) made possible the acceleration of electrons by radio frequency means to energies far higher than betatrons can achieve. Betatron acceleration has been applied for experiments with mesons at meson-producing energies of about 300 MeV, but the loss of electron energy by centripetal acceleration radiation can become about 10% of the final energy there, and it rises rapidly with energy. Hence the synchrotron principle of radio frequency acceleration is used to take relativistic particles to higher energies since this radiation loss is conveniently supplied by synchrotron acceleration. It is for this reason that centripetal acceleration radiation is called "synchrotron radiation," and this radiation now has found extensive applications as a light source with polarization and high intensity over a spectral range not accessible by other means.

Experiments and studies of betatron acceleration date back to times well before any of these applications. Slepian, Wideröe, Walton, Breit and Tuve, and Steenbeck, to name a few, are among the early workers. Slepian imagined a closed iron transformer core with electrons circulating around it like the windings of the secondary coil of a transformer. The electrons were to be bent into the circulating orbit by the bending action of permanent magnets placed around the core with their magnetic field increasing with radius to provide for holding both low-energy electrons at small radii and later for more energetic accelerated electrons at a somewhat higher energy (Kerst, 1946). It is interesting to note that the evolution of betatrons eventually led to just such a structure after alternating-gradient focusing was discovered.

The experiments of Wideröe and especially of Breit and Tuve and Walton were directed toward nuclear disintegration attempts with the beams. All of the applications mentioned were evident as possibilities for electron beams by the late 1930s when experiments guided by detailed theoretical work resulted in a small model operating at 2 MeV which produced an x-ray intensity equivalent to the gamma rays from 1 g of radium (Kerst, 1941; Kerst and Serber, 1941). Yields from betatrons have been sufficient for many practical purposes, but electrostatic repulsion between particles in the orbit limits the current which can be held by focusing forces. By injecting at higher energies, and hence at higher magnetic fields, magnetic focusing forces are increased and the space charge limit can be raised. The possibility of holding larger currents by the addition of a toroidal field has often been attempted, and it has ultimately been accomplished with the aid of alternating gradient or strong focusing.

1. PHYSICAL PRINCIPLES

To increase the energy of a particle in an orbit of radius r in a magnetic field the particle must experience an electric field in the direction of its motion. The magnetic field can do no work on a particle because the force which it provides is at right angles to the motion. However, a transformer action providing an increase of a certain number of electron volts per turn comes not from the field at the orbit, but from the rate of change of the flux, Φ, linking the orbit which is due to the rate of change of the vector potential at the orbit. In a betatron a transformer core within a circular orbit provides the voltage, $d\Phi/dt$, volts per turn. Figure 1 shows the structural forms and the accelerating chamber which is placed between the poles of the magnet. The iron magnetic circuit must be laminated because of the time-varying field, although air-core betatrons can also be made. The inductance represented by the coil system is usually in a resonant circuit with a large capacitor to store the energy. Some betatrons are pulsed with the capacitor's energy discharged unidirectionally through the coils as in Figs. 1(c) and 6(b).

The particles circulating in orbit between the poles experience an electric field

$$E=\frac{1}{2\pi r}\frac{d\Phi}{dt}\ \text{V/m}$$

and the vector potential associated with the linked flux is then

$$A=\frac{\Phi}{2\pi r}\ \text{Wb/m.}$$

Since the field at the orbit is

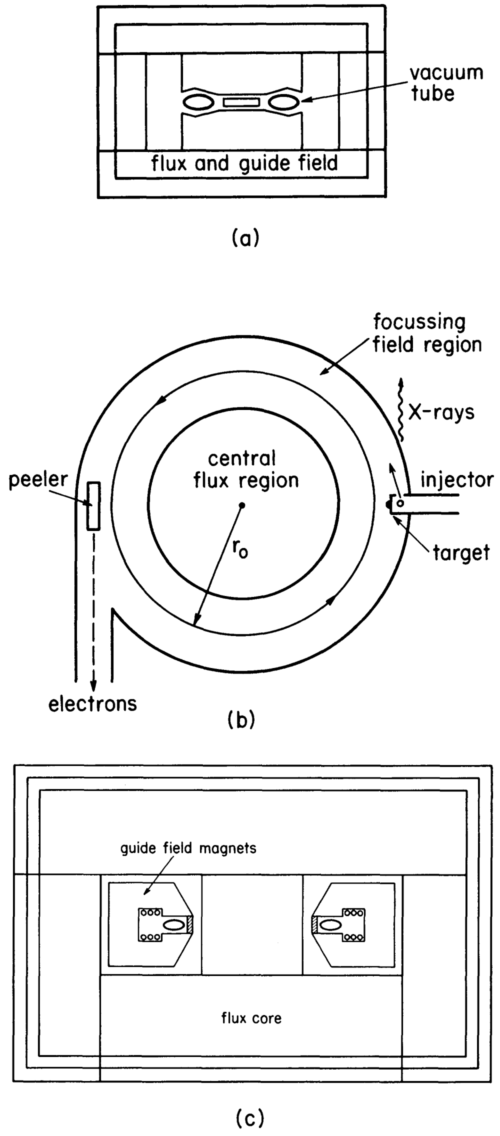

FIG. 1. (a) Magnet configuration of a 22-MeV betatron with the vacuum tube shown between the poles. (b) Toroidal vacuum tube. A resistive metallic interior coat prevents wall charge or *IR* voltage drops from deflecting the injected beam. (c) Configuration of unidirectionally pulsed magnets. The biased flux core links the orbit which circulates in a ring of focusing magnets.

$$\mathbf{B}=\mathrm{curl}(\mathbf{A}),$$

it is evident that in Fig. 1(c) the central portion of flux in the flux core will produce no magnetic field at the orbit because this core's vector potential has a $1/r$ dependence and thus its curl is zero and its contribution to B at the orbit is then zero.

However, the vector potential of the necessary focusing magnets in Fig. 1(c) at the orbit must have a nonvanishing curl(**A**) in order to give the focusing field **B**, but some electric field due to any orbital flux linkage from the focusing magnets is added to the electric field due to the accelerating flux of the central core producing the total electric field acting on the particles.

The momentum of a charged particle moving in an orbit of constant radius r is

$$p=eBr$$

and if this momentum was given to it by an electric field from a changing flux linkage within its orbit, then its rate of change of momentum was

$$\frac{dp}{dt}=eE.$$

Substituting

$$\frac{e}{r}\frac{dB}{dt}=\frac{e}{2\pi r}\frac{d\Phi}{dt}$$

and integrating, we find the flux condition

$$2\pi\, r_0^2 B=\Phi-\Phi_0.$$

We have assumed that the particle is started with $B=0$ and that the orbit maintained a fixed radius, and we find that the flux change within the orbit must be twice that for a uniform field within the orbit area.

The path of an electron in typical betatrons which require 1–4 ms to accelerate can be as long as 1000 km. Evaluation of beam loss by scattering on residual gas in the vacuum chamber and experimental observations have shown that a vacuum pressure of less than 10^{-5} Torr is required for these long paths. This vacuum is not difficult to obtain, but there must be focusing forces to keep the beam together stably. That there are such forces, under certain conditions, can be seen by a simple argument. For axial (vertical) focusing note that the centripetal force due to the bending magnetic field is directed inward radially at right angles to the magnetic line of force. Thus if the lines of force, which are vertical in the plane of the orbit, bend inward above and below this midplane, the force on the particle will acquire a component directed back toward the midplane when it strays out of that plane. This is the necessary restoring force to cause the particle to oscillate about the midplane. This field line curvature occurs when the strength of the magnetic field decreases in magnitude as the radial distance from the axis increases. This is the same vertical focusing action occurring in cyclotrons. If we describe this field dependence as

$$B \approx r^{-n},$$

then

$$n \geqslant 0$$

for the necessary decrease of magnetic field with radius for axial focusing.

The requirement on n for radial focusing can be seen by noting that the centripetal force to keep a particle in a circular orbit decreases as $1/r$. Thus if the magnetic field which supplies this force decreases less rapidly than $1/r$, a particle displaced to a radius too large will find the magnetic force larger than needed to remain at the large radius and its path will curve inward to smaller radii. Likewise, a particle at too small a radius will find that the field is not strong enough to hold it at this radius. Hence n must be less than 1, and there is a fixed equilibrium radius where the centripetal force is equal to the magnetic force. Thus

$$0 \leqslant n \leqslant 1.$$

This is visualized in Fig. 2 where the required centripetal force to hold a particle in an orbit of radius r_0 is shown along with the radial dependence of magnetic force needed to accomplish focusing. The difference between the curves is available for focusing.

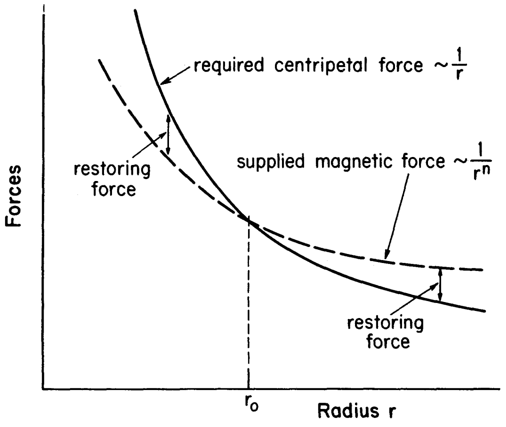
required centripetal force $\sim \frac{1}{r}$
restoring force
supplied magnetic force $\sim \frac{1}{r^n}$
restoring force
Forces
r_0
Radius r

FIG. 2. Radial forces. The required centripetal force to maintain a constant radius, $1/r$, is exceeded by the magnetic force for radii greater than r_0. Thus return toward r_0 is assured providing focusing.

The invention of alternating gradient focusing some years after the construction of the first betatrons removed this requirement on n and allowed magnetic field with much larger gradients, $n \gg 1$, but with n alternating in sign along the orbit. This led to the possibility of far smaller apertures and magnets and thus a great saving in cost.

It is now necessary to use the equations of motion to show the conditions for the fields to produce stable orbits.

2. EQUATIONS OF MOTION

For a circularly symmetric magnetic structure the vector potential describing the field has only a ϕ component. The relativistic Hamiltonian is

$$H = c\{m^2c^2 + P_r^2 + P_z^2 + [P_\phi/r - eA(t,r)]^2\}^{1/2} \tag{1}$$

and Hamilton's equations give

$$\frac{dH}{dP_r} = \frac{dr}{dt} = \frac{c^2}{H} P_r, \tag{2}$$

$$\frac{dH}{dr} = -\frac{dP_r}{dt} = -\frac{c^2}{H}\left(\frac{P_\phi}{r} - eA\right)\left(\frac{P_\phi}{r^2} + e\frac{dA}{dr}\right), \tag{3}$$

$$\frac{dH}{dP_z} = \frac{dz}{dt} = \frac{c^2}{H} P_z, \tag{4}$$

$$\frac{dH}{dz} = -\frac{dP_z}{dt} = -\frac{c^2}{H}\left(\frac{P_\phi}{r} - eA\right)\left(e\frac{dA}{dz}\right). \tag{5}$$

P_ϕ, P_r, and P_z are the canonical momenta with the ϕ component being a constant of the motion because of the lack of dependence of the problem on ϕ. P_r and P_z are the radial and axial momenta, respectively.

Taking $P_\phi = 0$, the classical and the relativistic cases will be treated separately, and then the case with $P_\phi \neq 0$ will be described. A more general treatment is in Kerst and Serber (1941).

When the particle's energy is far less than its rest mass energy, we have the classical case with $H \approx mc^2$ in (1), and (2) becomes $dr/dt = P_r/m$ and (3) becomes $dP_r/dt = -e^2A(dA/dr)/m$ giving

$$\frac{d^2r}{dt^2}=-\left(\frac{e}{m}\right)^2A\left(\frac{dA}{dr}\right) \tag{6}$$

for the radial acceleration. If the field is arranged so that there is a position where $dA/dr=0$, say at r_0, then that would be the radius for a fixed orbit with no radial acceleration. Furthermore if dA/dr is negative at $r\leqslant r_0$ and positive for $r\geqslant r_0$, then the radial acceleration, d^2r/dt^2, is always directed back toward r_0, an equilibrium orbit. Thus radial focusing is provided by such a field with the vector potential having a minimum at r_0.

The conditions just stated in (6) would describe an oscillatory motion with an angular frequency ω satisfying

$$\frac{d^2x}{dt^2}=-\omega^2x, \tag{7}$$

where

$$x=r-r_0$$

and

$$\omega^2x=\left(\frac{e}{m}\right)^2A\,\frac{dA}{dr}\,.$$

But since $dA/dr=0$ and A is minimum at r_0, expansion about r_0 gives

$$\frac{dA}{dr}=\frac{d^2A}{dr^2}\,x$$

and thus

$$\omega^2=\left(\frac{e}{m}\right)^2A\,\frac{d^2A}{dr^2}\,. \tag{8}$$

To evaluate (8) in terms of the magnetic field shape note that the curl equation gives $B_z=A_\phi/r+dA_\phi/dr$ and thus

$$\frac{dB_z}{dr}=\frac{d^2A}{dr^2}+\frac{dA/dr}{r}-\frac{A}{r^2}\,.$$

Then with $dA/dr=0$ at the equilibrium orbit and $\Omega=eB_z/m$, the circling angular velocity is

$$\omega_r=\Omega\left[1+r\frac{dB_z/dr}{B_z}\right]^{1/2}.$$

If the radial field dependence is $1/r^n$, then the familiar form results:

$$\omega_r=\Omega(1-n)^{1/2}, \tag{9}$$

with n less than 1 giving a real result.

For a differential equation like (7) with an adiabatically time-varying coefficient on the right-hand side which represents a stiffening spring, the amplitude of oscillation will vary as $1/\omega^{1/2}$ and this shows that the radial oscillation "damps" as the acceleration of the particles to higher energy progresses:

$$-\frac{\delta a}{a}=\frac{\delta B}{2B}=\frac{\delta E}{4E}, \tag{10}$$

where a is the amplitude of the orbital oscillation and E is the energy gain associated with the rising field B.

Finally on the equilibrium orbit $B_z=A/r_0$ and the accelerating electric field is

$$-\frac{1}{2\pi r}\frac{d\Phi}{dt}=-\frac{dA}{dt} \tag{11}$$

so

$$\Phi-\Phi_0=2\pi\, r_0^2B_z, \tag{12}$$

the equilibrium orbit flux condition with the possibility of a curl-free initial reverse bias flux Φ_0 through a portion of the iron within the orbit. Such a bias allows central iron flux to change twice as much as it could if it were unidirectional.

About this equilibrium orbit with damped radial motion of stray particles, axial motion out of the plane of the orbit should also be focused and damped. As mentioned the outward bulging of the magnetic field lines in the orbital plane provides this force with a component back toward the plane. The resulting axial acceleration is [like (6)]

$$\frac{d^2z}{dt^2}=-A\left(\frac{e}{m}\right)^2\frac{dA}{dz}\,. \tag{13}$$

In this case dA/dz is zero on the orbital plane by symmetry for positive and negative z.

Since $dA/dz=B_r$ and curl($\mathbf{B}$)=0 in the vacuum and $A=rB_z$, $dB_z/dr=dB_r/dz=B_r/z$, near $z=0$ (13) becomes

$$\frac{d^2z}{dt^2}=\left(\frac{eB_z}{m}\right)^2\left(\frac{r_0}{B_0}\right)\frac{dB_z}{dr}\,z \tag{14}$$

or

$$\frac{d^2z}{dt^2}=-\Omega^2nz \tag{15}$$

showing that the axial angular frequency is

$$\omega_z=\Omega n^{1/2}. \tag{16}$$

(16) and (9) substantiate that $0 \leqslant n \leqslant 1$. The time increase of B in (14) shows that the amplitude of axial oscillation is "damped" by the same factor as is the radial motion (10) for nonrelativistic particles.

For the relativistic case we ignore mc in the Hamiltonian (1) and find Hamilton's equations with small P_r and P_z to be

$$\frac{dr}{dt}=\frac{c}{eA}P_r,$$

$$\frac{dP_r}{dt}=-ce\frac{dA}{dr},$$

$$\frac{dz}{dt}=\frac{c}{eA}P_z,$$

$$\frac{dP_z}{dt}=-ce\frac{dA}{dz}.$$

The radial acceleration equation is

$$\frac{d^2r}{dt^2}+\frac{1}{A}\frac{dA}{dt}\frac{dr}{dt}=-\frac{c^2}{A}\frac{dA}{dr}$$

and the axial, or z, equation has the same form with z substituted for r. The reasons for dA/dr and dA/dz vanishing on the equilibrium orbit are the same as for the classical case giving a generic differential equation with x for $(r-r_0)$ or for z and $(d^2A/dx^2)x$ for dA/dx:

$$\frac{d^2x}{dt^2}+\frac{1}{A}\frac{dA}{dt}\frac{dx}{dt}=-\frac{c^2}{A}\frac{d^2A}{dx^2}x \tag{17}$$

with

$$\frac{d^2A}{dr^2}=\frac{dB_z}{dr}+\frac{A}{r^2}$$

for radial motion where $dA/dr=0$ and

$$\frac{d^2A}{dz^2}=-\frac{dB_r}{dz}=-\frac{dB_z}{dr}$$

for axial motion.

Since $A=Br_0$ on the equilibrium orbit, the right-hand side of (17) which is $-(\omega_x^2)\,x$ shows that

$$\omega_r^2=\left(\frac{c}{r}\right)^2\left(1+\frac{r}{B_z}\frac{dB_z}{dr}\right) \quad \text{or} \quad \Omega^2(1-n)$$

and that

$$\omega_z^2=-\left(\frac{c}{r}\right)^2\frac{r}{B_z}\frac{dB_z}{dr} \quad \text{or} \quad \Omega^2 n$$

as in the classical case if B has an r dependence $1/r^n$.

The constant rotational angular velocity Ω at the speed of light does not produce the spring-stiffening "damping" effect seen at classical energies but instead the differential equation (17) shows a damping term which is proportional to dx/dt producing an exponential amplitude factor on the solution for times of steady field rise, if dA/dt or dB/dt is constant and positive, of

$$\exp\left(-\frac{dA}{dt}\frac{t}{2A}\right).$$

This results in an orbit shift toward the equilibrium orbit and away from the injector at relativistic energies of

$$-\frac{\delta x}{x}=\frac{dB}{2B}=\frac{dE}{2E} \tag{18}$$

radially or axially in contrast to that for classical energies (10).

These amplitude changes result in a very small beam width when high energy is reached. This is very valuable for sharp radiographic purposes or beam manipulations. Together with influences of space charges and space currents of the beam, this damping made possible injection from a particle source very near the orbit in early accelerators.

A further transient motion after injection occurs if P_ϕ is not zero, that is, if the particle is injected tangentially but at an energy which is correct for the initial radius r_i but wrong for the equilibrium radius r_0,

$$P_\phi=(P_i-eA_i)r_i,$$

where P_i is mechanical angular momentum, and P_ϕ must be a constant in time. Thus if A changes with time,

$$d(P_ir_i)=d(er_iA_i).$$

But P_i on the instantaneous circle is

$$er_iB_i=er_i\left(\frac{A_i}{r_i}+\frac{dA}{dr}\right).$$

Combining we have $d(r^2dA_i/dr)=0$ giving

$$-\frac{dr_i}{dt}=\frac{\dfrac{\partial}{\partial t}\left(\dfrac{dA}{dr}\right)}{\dfrac{\partial^2A}{\partial r^2}+\dfrac{2}{r}\dfrac{\partial A}{\partial r}}.$$

Neglecting the small last term in the denominator near r_0 and noting that

$$\frac{\partial}{\partial t}\left(\frac{\partial A}{\partial r}\right)=\frac{\partial}{\partial t}\left(\frac{\partial^2 A}{\partial r^2}(r_i-r_0)\right)$$

gives

$$\delta r_i/(r_i-r_0)=-\delta B/B=-\delta E/E \tag{19}$$

for the shift in the instantaneous circle off-energy orbit radius with field or energy change.

These transient motions of injected particles all serve to move the particles to the equilibrium orbit and away from the injector as the energy rises.

3. THE RADIATION-LOSS ENERGY LIMITATION

A limitation of the betatron method of magnetic induction acceleration is that, with electrons, centripetal acceleration radiation loss of energy becomes so great at several hundred MeV that an iron-core flux with the ordinary rise time of a few milliseconds cannot sustain the loss rate. For example, a 300-MeV betatron with a 4-ms rise time has approximately 10% energy loss during acceleration. The loss in electron volts per second is

$$dE/dt=4.22[E\ (\text{MeV})]^4/[R\ (\text{m})]^2.$$

This radiation has a continuous spectrum described by Schwinger (1949) which makes energetic electrons very useful as a radiation source in the ultraviolet to soft x-ray region. However, the maximum in the spectral intensity occurs at the $(E/mc^2)^3$ harmonic of the orbit's revolution frequency. Consequently electron energies of interest are in a range up to 1000 MeV. This radiation, which is a serious obstacle for a betatron, was overcome elegantly by the invention of the synchrotron radio frequency accelerating method by McMillan (1945) and Veksler (1945). The radiation is consequently called synchrotron radiation, and the high-energy limit for electron synchrotrons is orders of magnitude higher than the limit for electron betatrons.

It can be said that synchrotron acceleration uses a very small amount of flux linking the orbit for induction acceleration but uses it again and again by reversing it when the particle is not at the accelerating gap of an rf-flux-containing cavity. Hence the particle experiences the electric field of a rising flux linkage only at the time it crosses the gap. The guide or focusing field configuration at the orbit can be the same as that for a betatron, but the synchrotron has in addition automatic ϕ or azimuthal phase focusing which maintains the equilibrium energy. With enough rf flux in the cavity the radiation loss can be supplied to energies more than an order of magnitude higher than electron betatron energies.

4. THE BEAM CURRENT LIMIT

If the electrostatic repulsion of the charge in the accelerated beam is sufficient to overcome the magnetic focusing forces, then more particles cannot be held in the beam. This space-charge limit can be estimated assuming the beam to be cylindrical with radius a. The electrostatic repulsion of a beam of I amperes for electrons of speed v at the edge of the beam is then

$$F_E=e\,I/(2\pi\epsilon_0 av)$$

and the axial force supplied by the magnetic field gradient is

$$F_m=evanB_z/r. \tag{20}$$

With the injection field B_i these are equal for a beam current limit of

$$I=2\pi\epsilon_0 a^2v^2nB_i/r_0. \tag{21}$$

This estimates limiting ring currents of 0.8 A for the betatrons which have been used for therapy and radiography, and this agrees approximately with observations. Since injection in such applications is at about one-third the speed of light, the final ring current rises to 2 or 3 A in the process of gaining speed.

This limit on current is altered when relativistic energies are approached since then the self-magnetic field of the beam adds a pinching force $F_m=v^2\mu_0\epsilon_0F_E$. The sum of the two forces is

$$F_E-F_m=F_E(1-\beta^2) \tag{22}$$

which vanishes if the speed of light is approached for the injection energy. Thus the limits on space charge and space current can be very high for high-energy injection. If the beam is not a straight cylinder, or if it has conducting walls with image currents and charges in the vicinity, then the cancellation is not quite as shown by (22). However, using the repulsion calculated by (22) balanced by the focusing force from (20), the limiting beam current is

$$I=\frac{2\pi\epsilon_0 nmc^3a^2\beta^3}{er^2(1-\beta^2)^{3/2}}$$

or

$$I=\frac{8.53n\beta^3a^2}{r^2(1-\beta^2)^{3/2}}\ \text{kA}. \tag{23}$$

This can be very large if the beam radius a fills the vacuum tube aperture at injection and if β is near its maximum of 1.

To raise the limit farther an additional magnetic field B_ϕ can be applied along the path of the particles. This accelerator is the "modified betatron" of Sprangle and Kapetanakos (1978), Rostoker (1980), Chernin and Sprangle (1982), and Kapetanakos *et al.* (1983). The effect of this field is to convert a velocity leading away from the beam bundle to a spiraling motion of the particles. This multiplies the limiting current in (23) by $(B_0/B_z)^2$. At the low field at injection time it is easy to make this ratio very large, thus raising the limit into the thousands of amperes. As acceleration progresses, B_ϕ is not to be increased as much as B_z since it is needed only at injection energies. If it could be increased proportionally with B_z, the difficulty to be described next would be avoided.

The B_ϕ field operating on the centroid of a beam traveling along (not across) the ϕ direction produces a drift velocity,

$$\mathbf{V}_d=-\mathbf{F}\times\mathbf{B}_\phi/eB_\phi^2.$$

F is any force such as a focusing force on the beam which has a component orthogonal to B_ϕ. This drift is like the drift of a particle in crossed electric and magnetic fields, $\mathbf{V}=\mathbf{E}\times\mathbf{B}/B^2$, but instead of the E field's force we have F composed of the magnet's focusing force on the beam centroid, plus wall image current repulsion, minus wall image charge attraction, so $F=F_{\text{focus}}-F_{\text{images}}$. The image force for very large currents can overwhelm the focusing force at low injection energies. But the current and charge image forces approach self-cancelation at relativistic energies and thus the magnet's focusing forces would dominate after some acceleration. Therefore, the drift velocity reverses after some energy is reached during acceleration for beams started in what is called the "high current range." There is a "low current range" in which the magnet's focusing forces exceed the wall images' defocusing attraction at all times and there would then be no reversal of drift during acceleration.

The beam centroid drift tends to follow along the wall if the beam is not in the center of the chamber since that is the $\mathbf{F}\times\mathbf{B}_\phi$ direction. To evaluate this drift circulation, or "bounce" frequency, first consider a low current case with negligible image forces. We had by (9) and (16)

$$\omega_r=\Omega(1-n)^{1/2}\ \text{and}\ \omega_z=\Omega n^{1/2}$$

or with $n=\frac{1}{2}$, for example, we have

$$F_{\text{focus}}=m\omega_r^2x\ \text{or}\ m\omega_z^2z;$$

then $V_d=m\omega_r^2x/eB_\phi$ or

$$\omega_{\text{drift}}=n^{1/2}\Omega n^{1/2}B_z/B_\phi.$$

If $n\neq\frac{1}{2}$, then the angular velocity of the drift is

$$\omega_{\text{drift}}=\pm[n(1-n)]^{1/2}\Omega B_z/B_\phi$$

and thus the drift frequency can be slow compared to the orbital revolution time when $B_z<B_\phi$.

We can now see what will happen if the current is raised to the high current regime where the wall image forces pull the centroid of the beam toward the wall and overcome the magnet's focusing forces. To do this consider the meaning of n and $1-n$. They are ordinarily used as the power on $1/z$ or $1/x$, respectively, for the displacement dependence of the magnetic focusing force which brings the beam back toward the center of the chamber. Now consider n to describe the power law for the combination of image and focusing magnet forces acting on the beam centroid as a whole in the z direction, and $1-n$, call it n', to be a similar exponent on $1/x$ which describes the whole force on the beam in the x or r direction. Then

$$\omega_{\text{drift}}=\pm[n\,n']^{1/2}\Omega B_z/B_\phi.$$

It is then evident that if these two exponents, n and n', are both positive, as in a conventional betatron or a modified betatron in the low current regime, the drift angular velocity is real giving a stable drift around the tube wall. It would be real also if n and n' were both negative and a stable drift would occur in the opposite direction. This is the case of the high current regime for image forces overwhelming the magnetic focusing forces and pulling the beam toward the wall. The beam is still contained but "bouncing" in the opposite direction around the wall.

Instability and an accompanying drift to the wall can occur when the low-energy injected beam is in the high current regime with n and n' thus both negative. Then subsequent acceleration increases magnetic focusing and decreases image forces leading to the low current regime. n and n' thus must turn positive but at slightly different energies causing n and n' to be opposite in sign at some energy thus making the drift angular velocity imaginary. This means instability and the beam is then lost.

Loss from this instability can be avoided by starting in the "low current regime" which is the final energy state. This conservative design choice does not gain from the factor B_ϕ^2/B_z^2 but uses the limit given in Eq. (23) which is the same as that for the conventional betatron—except that a is the chamber minor radius and except that injection of the order of 1 MeV is used (Kapetanakos *et al.*, 1983).

It is evident that the beam bounce drift instability could be moved to much higher currents if the magnetic field's focusing force could be increased. This can be accomplished by adding alternating-gradient or strong-focusing windings around the chamber. Roberson *et al.* (1983) call this the "stellatron" since the windings are similar to stellarator windings. With the strong-focusing modified betatron the limit on circulating current enables experiments to extend into the multi-kiloampere range. Nevertheless the added toroidal field brings in new beam trapping possibilities such as inductive trapping by raising B_ϕ during injection.

5. INJECTION

Space charge limits on the current which can be held in accelerators may be technically difficult to reach. From the single-particle equations of motion the shift away from the injecting structure during acceleration was evaluated. Our very high-energy accelerators follow such single-particle dynamics. Design is usually guided by Liouville's theorem which puts the limit of current on the phase space density achievable within the source or injector. This is invariant throughout acceleration. The actual limit allowed also depends on parameter changes which cause instabilities or resonances and which are influenced by charge density. The limit is not simply the space charge limit of the focusing forces which has just been discussed.

With the earliest betatrons, observations showed that orbit shifts, greater than those predicted by single-particle dynamics, aided particle capture into orbits that allowed acceleration to full energy. Subsequent measurements revealed that injecting a large beam current promoted capture. We can expect some such behavior since the electromagnetic energy associated with a circulating current is removed from the kinetic energy of the beam as the beam fills the chamber. This energy is proportional to the square of the number of turns the injected stream makes and thus some of the particles are trapped on a different instantaneous circle just by their energy change due to the incoming beam. Other beam-produced processes which can act to change untrapped orbits to trapped orbits immediately after injection have to do with the buildup of space charge, the sacrifice of space charge, the resistance of the wall coating, and the generation of radio frequency fields.

Purposeful parameter changes during injection are used very successfully to move particles to orbits which do not lead back to the injector structure—the main obstacle. Almost any field transient at that time will help, provided that the injector sprays an abundance of electrons in the right directions. A so-called orbit contractor is used on some betatrons to subtract a pulse of flux from the central flux at injection time, just as the injected beam current does. The contractor is very helpful in improving capture if a small beam is injected. However, if an intense beam is injected, the orbit contractor does not capture much more. In the latter case the reverse ampere turns of the beam itself have produced the contraction. See Kerst *et al.* (1950) and Kerst (1959).

The experience with betatrons has been to develop capture processes which put only a fraction of the injected particles into good orbits. The industrial betatrons approach the space charge limit at injection by using guns which can provide excess electrons that can be sacrificed. Not much current is needed at their injection energy to reach the space charge limit, and hence simple hot filament electron guns suffice. High current kiloampere range modified betatrons using a rising toroidal field for inductive charging or runaway plasma electrons for injection also trap by processes which sacrifice many electrons (Ishizuka *et al.*, 1985). However, injection technique with cold emission diodes seeks to minimize stray electrons and to put a directed beam on the proper orbit (Golden *et al.*, 1988).

The plasma runaway electron source for reaching high currents had been suggested by Budker (1956). He also pointed out that very strong focusing would result if charge neutralization occurred. Several groups experimented subsequently with this method and results are detailed in the CERN International Conference on High Energy Accelerators proceedings for 1959 and in Reynolds and Skarsgard (1960). At that time Linhart's group at CERN, which built five experimental betatrons, and Reynolds were able to inject and accelerate about 1 A to 2 MeV at a suitable pressure, although a conduction current orders of magnitude larger existed at the start of acceleration. These were what are now called "modified betatrons" in that they had toroidal fields. However, the later experiments of Rostoker and Kapetanakos produced circulating currents in their high-current modified betatrons of several hundred amperes which lasted about 100 ms during acceleration in the MeV range. The beam currents sometimes decayed and in some cases abruptly disappeared. Various instabilities associated with large beam currents or with field inhomogeneities are expected to cause such beam loss. Nevertheless, the plasma source used by Rostoker's group gave much more current than the plasma sources of the earlier similar betatrons. When strong focusing was added to these recent already modified betatrons, then with the plasma source the accelerated beam reached a ring current of 1 kA at 10 MeV (Ishizuka *et al.*, 1988), and with diode injection of directed electrons it reached about 1 kA at 16 MeV (Kapetanakos, 1989 and 1990). Figure 3 shows this betatron with its million volt injector, and vertical and toroidal coils. The alternating gradient strong focusing coils are not shown.

6. THE TARGET, BEAM EXTRACTION

To bring the beam to the target, extra flux is driven through the orbit's center to expand the orbit's radius to the target radius, or a field bump at some azimuth is pulsed on to distort the orbit toward the target. Practical application for radiography requires a target which, for the smallest source size, is inside the vacuum tube. This location benefits from the focusing and damping of orbital oscillations resulting in a beam size of a millimeter or less when high energy is reached. Furthermore, if a target smaller than the beam size is inserted, then electrons which pass above the

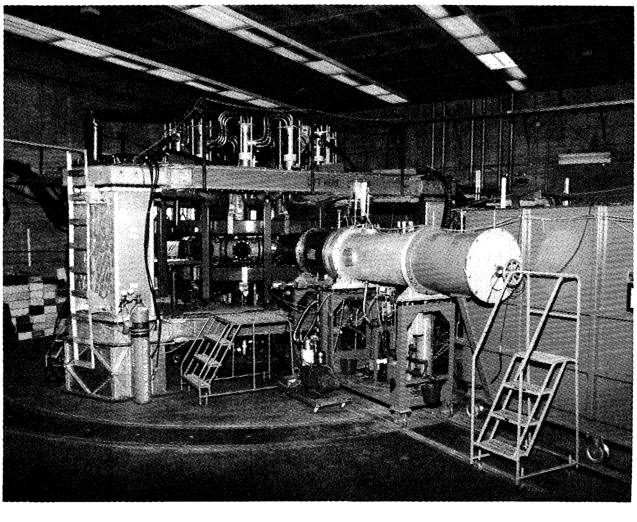

FIG. 3. High-energy injected modified betatron of the U.S. Naval Research Laboratory.

target on one pass will hit the target subsequently since the phase of their axial orbital oscillations should differ on each pass, and the source can thus be no bigger than the target. The radiographic definition possible with such targeting is unmatched making a betatron or synchrotron the choice for high definition work.

With a small target protruding on the injector structure, the obstacles interfering with injection are minimized. It is estimated that when the electron beam grazes the target, it scatters in less than 0.004 cm, and that more than half of the beam would scatter out before producing thick target radiation as indicated in Figs. 4 and 5. At the electron energy which produces the most penetrating x-ray spectrum for iron (20–30 MeV), the efficiency of x-ray production is more than 60% and target heating is not a problem.

Extraction of the electron beam for nuclear experiments or for electron beam therapy has been done by scattering off of the target or by expanding the beam into a magnetic shielding channel. In both cases a fine beam does not emerge. There is always a septum effect which spreads the beam extracted by a channel or "peeler."

There is a resonant process useful for beam extraction from accelerators in general which can be exemplified by the case of a magnetic peeler in a betatron where $n=\frac{3}{4}$ near the extraction radius. By Eq. (16) one-half of a radial oscillation will occur between passages of the peeler location. The magnetic field gradient caused by the peeler perturbation excites a half-integral resonance causing an exponential growth of radial oscillation and rapid steps into the shielded region of the peeler.

Modern high-energy accelerators handle beams far more efficiently than early betatrons did by properly designed pulsed inflectors or extractors which preserve the emittance of the beam and thus the particle density in phase space.

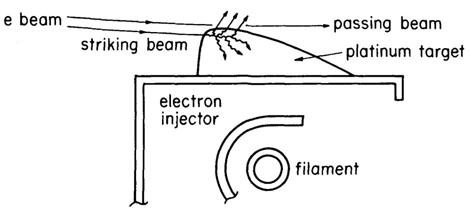

FIG. 4. Multiple and single scattering out of the target tip grazed by a slowly expanding orbit.

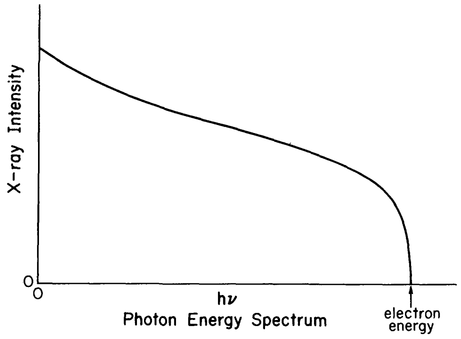

FIG. 5. Thin target photon intensity of bremsstrahlung spectrum.

7. FIELD ERRORS AND THEIR EFFECTS

Magnetic field distortion from the symmetry and uniformity implied by the equations of motion is caused by misplaced conductors, eddy currents in the conductors, or nearby objects in the magnet's fringing field, and phase shifts in alternating current magnets or the equivalent time lag for pulsed fields with some iron in the circuit.

The problem is to establish tolerances on allowed orbit distortion which results from field errors and then to evaluate the limits of structural error which must be met.

When a particle of momentum $p=eB_z r_0$ passes through a distortion δB of length L, it is given a perpendicular impulse $eL\delta B=\delta p$ so its angle of deflection is $\alpha=\delta p/p=(L/r_0)\delta B/B_z$. This deflection leads to a sinusoidal oscillation of amplitude a for which the angle at impulse is $(a/r_0)\times(1-n)^{1/2}=\alpha$, and thus the tolerance condition on length of error L and δB would be

$$(a_{\max})(1-n)^{1/2} \geqslant \delta B\, L/B. \qquad (24)$$

Here $1-n$ refers to radial motion and axial field strengths. For axial motion and radial field errors use n.

Since the particle returns to the field bump after one revolution, it has executed a fraction ω_r/Ω of a betatron radial oscillation at which time it receives a second impulse. The sum of all sine oscillations produced can be evaluated as done in optics for the Cornu spiral. The result can be described as a steady orbit displacement (because the average field strength along the orbit has been changed), a forced oscillation driven by repeated impulses which forms a new closed orbit which makes an angle $\alpha/2$ with the unperturbed closed orbit, and a free oscillation (Kerst, 1959). The aperture available would be decreased by the excursions of the forced oscillation of the distorted equilibrium orbit plus the free oscillation between passages, but the steady displacement can be offset by injecting at a different time. If the available aperture has width D, it will be decreased by the closed orbit excursion,

$$\delta D = \frac{\delta B}{B}\frac{\Omega}{\omega}\frac{L}{2}\tan\left[\frac{\pi}{2}\left(\frac{\omega}{\Omega}-Q\right)\right], \tag{25}$$

where Q is the integer less than the number of orbital oscillations once around the magnet if the case is an accelerator with stronger focusing than a conventional betatron. Note that resonance amplitude increase occurs when ω/Ω is integral. The tolerable δB and L can be determined with Eq. (25).

Sometimes resonances are to be passed due to parameter changes in an accelerator having more than one orbital oscillation per revolution, such as a modified betatron. In this case the tactic to avoid resonance blowup is to pass through the resonances quickly. However, the Cornu spiral for passage through a resonance leaves a residual vector of free oscillation having magnitude

$$A_f \approx \frac{2L}{\pi^{3/2}\left[\frac{d}{dt}\left(\frac{1}{\omega}\right)\right]^{1/2}}\left(\frac{\Omega}{\omega}\right)^2\frac{\delta B}{B}. \tag{26}$$

The speed of passage and the tolerances can be estimated for each resonance passed and the residual vectors can be summed (Kerst, 1983). Random relative phases between the residuals should lead to growth proportional to the square root of the number of resonances passed through, but programmed passages might be arranged to cancel the final resultant vector thereby succeeding in getting through a series of hazardous resonances.

In an ac magnet these errors are not easily seen with a search coil; but with a sliver of high permeability iron within a coil at the time of no saturation, close to zero field or injection time, a sharp voltage peak results. Relative phase shifts between two such peaking strips as a function of position in the field can be determined. From this the length L and the magnitude of δB for Eq. (24) are found.

Zero field strength has also been detected by collimated electron beam vacuum tubes which pass the beam only at the time when there is no field. The detector can be moved about over the pole face to test simultaneity of zero field.

Over an iron pole face a phase shift error can be corrected by a coil on the patch of length L or by a coil wound on the bundle of iron carrying the flux to that position. All that is needed is a small current which is either dc or picked up from a turn on the magnetic core and phase-shifted.

In the case of a pulsed guide field magnet, such as used in a flux-biased betatron, the field starts from its remnant value and rises in unison all over the pole unless some pole region has a different eddy current loss delaying its rise. Since the remnant field is approximately

$$B_{\text{rem}} = B_c(\text{air gap})/(\text{iron length}),$$

where B_c is the coercive force, the iron return paths for flux to all sectors of the pole face should have the same length so that they all start from the same remnant field magnitude, or else patch correcting coils would be needed to bring all regions of the guide field into synchronism.

In practice an accelerator with any beam can use its beam magnitude to help trim fields. For example, a method used on the first Cornell synchrotron is especially instructive. With long-wavelength betatron oscillations it is the first azimuthal harmonic of the field error that disturbs the orbit most. Consequently if correction coils are placed on the poles which produce a field distribution proportional to the sine of the azimuth ϕ, and if another set of coils produces corrections proportional to the cosine of ϕ, then a first harmonic correction can be moved to any angle around the orbit by controlling the magnitudes of the coil currents. The corrections can be used to cancel error effects by optimizing the beam.

8. APPLICATIONS

When electron energies of several million electron volts became available from betatrons, the order of magnitude increase of energy was the right range for nuclear binding research, x-ray radiography in iron, deep tumor therapy, and electron beam therapy.

The lowest binding energies are for beryllium and deuterium in the 2-MeV region, but the threshold for photo-disintegration of most of the elements in the periodic table is from 8 and upward to carbon near 19 MeV. The photoneutron reaction is usually prevalent giving a radioactive product. However, betatrons are not as good as cyclotrons for producing abundant radioactive samples for tracers or for therapy. One of the characteristics of the radiation is that the x-ray spectrum is continuous giving very few photons at the peak electron energy (Fig. 5). Although disintegration thresholds can be determined because the electron energy can be very accurately adjusted, the yield rises slowly with energy due to the few photons at the high-energy end of the spectrum. It requires a great effort to detect the true threshold to the precision achievable in the electron's energy. A pulse of extra flux is supplied within the orbit when the field at the orbit or the flux within the orbit reaches that for a precise energy. This expands the orbit radius at the desired energy and all the electrons are brought to the same target radius, and therefore the same energy, where a spectrum with the same high-energy limit is produced.

A fruitful use of the electron beam extracted through a magnetic channel is for the tagging of photons which subsequently produce a nuclear reaction. This is done by capturing, in an energy analyzer, the electron which has produced the photon that caused an observed coincident event. The photon energy is then the precisely known original energy of the electron minus its analyzed energy. The system is sometimes called a photon monochromator because only a very small section of the photon spectrum is used.

The betatrons for these studies were similar to the 20- to 25-MeV therapy or radiographic machines. It was clear that meson-producing energies could be reached without too much synchrotron radiation loss. 100 MeV was too low for making up the rest mass of the muon or the pion meson although there were some betatrons made and used for nuclear experiments in this range. 300 MeV or more was chosen for the meson-producing accelerators of the period. Several synchrotrons were made and one betatron for 320 MeV was finished in 1948 (Fig. 1). This structure had a separate biased flux core through the ring of field, or guiding, magnets. The ring of field magnets was very similar to the magnets of some synchrotrons, and its coils were connected in parallel with the core flux coil so that the equilibrium radius flux-field relation would be maintained throughout the acceleration pulse. The precision of fulfilling the flux relation, and hence the constancy of the orbit radius, depended on making the coil resistance low enough so that its resistive voltage drop was small compared to the inductive voltage on the coils. The six orbit magnets weighed 60 tons total and the flux magnet weighed 230 tons. The six pulses per second wave form is shown in Fig. 6(b). If flux biasing is used as shown in the figure, a choke coil is needed for keeping the pulse out of the dc generator which supplies the bias winding on the closed iron core. Since the flux magnetic circuit is all iron, very few bias

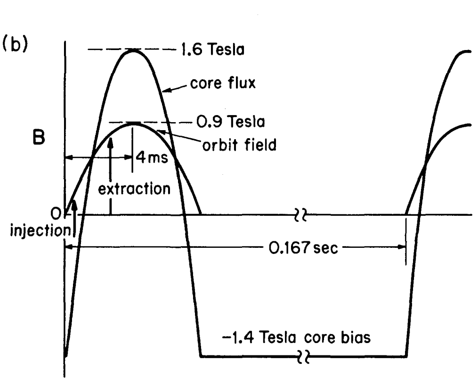

FIG. 6. (a) Accelerating field for sinusoidal commercial operation at 180 Hz. Three-fourths of the cycle is not used. (b) Six pulses per second unidirectional flux-biased wave form of the 320-MeV Illinois betatron.

ampere turns are needed. An alternate choice for biasing is to bias the field coils. In this case the bias current must be much larger.

The electron synchrotrons of the day tended to yield less output than the 320-MeV betatron. They used betatron action at injection with a very small flux core which soon saturated, and with the beam captured, synchrotron acceleration by radio frequency flux took over. In fact the first test of synchrotron acceleration was performed on a small 4.5-MeV betatron made by the University of Illinois group and given to the Woolwich arsenal in Britain where Goward and Barnes (1946) were able to penetrate the resistive wall coating with a radio frequency electric field and to continue the acceleration to 8 MeV.

Much later small FFAG betatrons played a role in the development of synchrotron acceleration manipulations. The purpose was to learn how to stack successively accelerated groups of particles in a ring current of sufficient magnitude for colliding beam experiments (Terwilliger *et al.*, 1957). The small betatron acted as a storage ring with beams started by betatron acceleration and then picked up by the rf cavity and brought to the stacking energy. It was necessary to determine the interference of successive group accelerations on the density of phase space in the stack. Although the stacking process could be modeled on a digital computer, the experiments provided confidence that one could carry out the required physical manipulations for synchrotron stacking and storing of beams.

Another small MURA FFAG betatron of the spiral sector type with all six sectors of the same polarity was applied to the establishment of experience with the wide variety of resonances which an alternating gradient focusing accelerator must avoid, to practice passage through the transition energy when synchrotron acceleration is employed, and to carry out the operation of producing a functioning physical structure. It accelerated only to one-third of the electron rest mass, but that was sufficient to exceed the transition energy (Haxby *et al.*, 1959). High-energy high-intensity sector-focused cyclotrons now use this spiraling pole structure to approach the rest mass energy.

Radiography of metal structures was the first commercial use of betatrons (Fig. 7). In

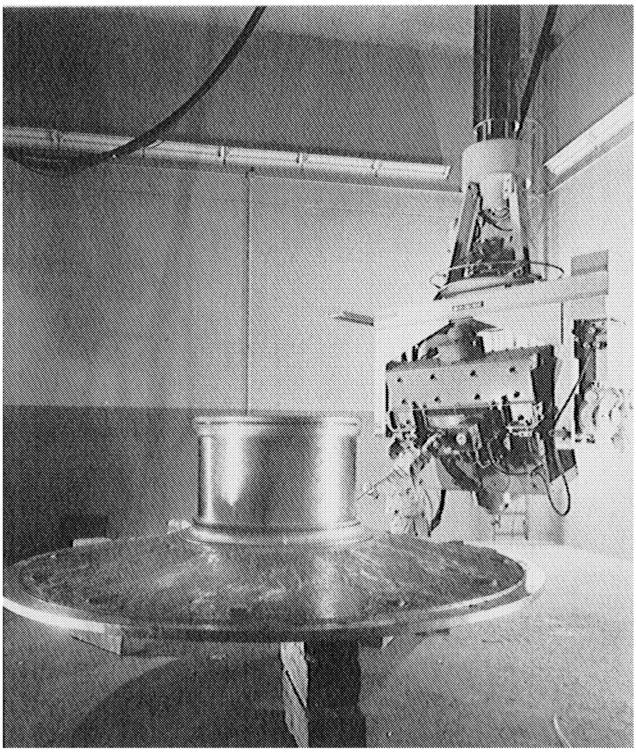

FIG. 7. Industrial radiography betatron. (From ATC Betatron Corp.)

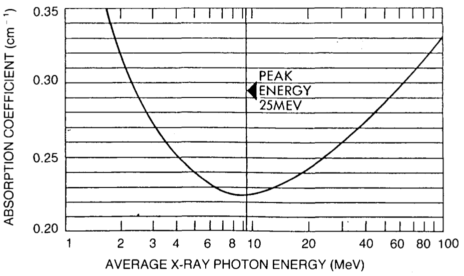

FIG. 8. Absorption coefficient for radiography of iron. (From ATC Betatron Corp.)

1943 they were installed in arsenals and at Los Alamos. The decrease of the Compton-effect absorption coefficient with photon energy is offset by the rise of electron pair production causing the most penetrating radiation to require about 25-MeV electrons (Fig. 8). If necessary fifteen to twenty inches of steel can be x-rayed in fifteen minutes to an hour.

As mentioned earlier the focal spot is vanishingly small because the beam naturally has small size and especially because the target can be a pointlike obstacle. A 0.002-cm platinum tip on the injector serves as a target. The high resolution possible with such a target has made betatrons especially suitable for certain high-resolution radiographic requirements.

Similar betatrons mounted differently became electron beam and x-ray beam deep therapy radiation sources (Fig. 9). Figure 10 shows how the ionization distribution in tissuelike material builds up beneath the surface for x rays from 20-MeV electrons. Since the tolerance of the patient for treatment is often limited by skin reactions at lower voltages, the minimizing of surface dose as well as the deeper penetration at voltages over 20 MeV has brought improved treatment possibilities.

A somewhat heroic effort was made to establish the physical means and techniques to adapt the order of magnitude increased radiation energy which became available in the 1940s for patient use (Quastler *et al.*, 1949). Collimation, dose measurement, beam flattening, aiming, and clearing away stray electrons required new methods of solution at 20 MeV. The requirement for beam flattening results from the target radiation being predominantly within an angle of mc^2/E of the forward direction of the electrons of energy E. Lengthwise cigar-shaped absorbers are needed to reduce the central intensity thereby flattening the dose distribution in the patient. Dose measurements were aided by passage of the applied radiation directly through foils which became radioactive and integrated the dose to the patient. Ultimately betatrons were used to develop a calorimetric method which made it possible to base radiation dose at these energies on the calorie to which intermediate standards could be referred.

Treatment by the accelerated particles themselves has some advantages, and therapeutic betatrons are arranged to provide both an x-ray and an electron beam. The ionization distribution along the path of an electron beam in tissue-like material is shown in Fig. 10. At relativistic energies electrons in tissue lose about 2 MeV/cm so a 22-MeV beam as

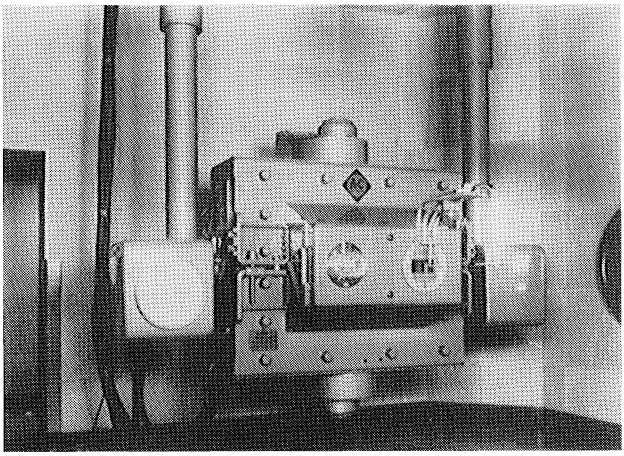

FIG. 9. Medical betatron used for deep therapy at 25 MeV.

shown in the figure would go as deep as 11 cm. It is evident that scattering has deflected many electrons before that depth. The finite range is advantageous because deeper tissue is not damaged. However, the multiple scattering and radiation straggling is much greater for electrons than it is for heavier particles, for example, for protons, so that a more sharply defined ionization distribution is possible with protons. It is necessary to use protons having an order of magnitude more energy to penetrate sufficiently.

9. CHRONOLOGICAL DEVELOPMENTS

The first description of an accelerator using magnetic induction seems to be Slepian's patent dated 1922. He described an accelerating flux core surrounded by permanent magnets for guiding electrons emitted from a filament. The bending field was to be greater at larger radii to accommodate the highest-energy electrons in the chamber. This is the wrong way for the gradient to focus axially, but it is interesting that direct current FFAG magnets can solve the focusing problem and a functioning betatron was made this way by Jones *et al.* (1956).

Experimenters who made attempts to accelerate electrons later in the 1920s are Breit and Tuve, Wideröe, and Walton. The full transient theory of single particle motion was not determined until much later (Kerst and Serber, 1941). This led to the design of an injection method capable of trapping injected particles. Quite a few other contributions to theory and experiments are summarized in a review by Kerst (1946).

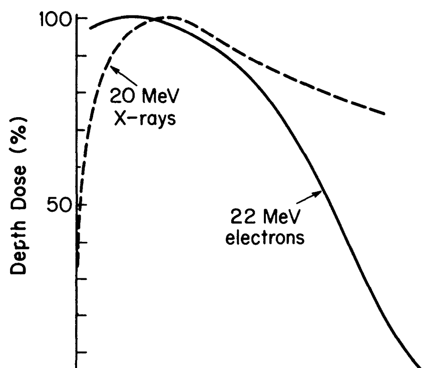

FIG. 10. Buildup of ionization below the skin for 22-MeV x rays, and range limitation of 22-MeV electrons inside the body. Depth dose curves.

The earliest commercial model was that of the Allis Chalmers company shown in Fig. 7 in its industrial support and in Fig. 9 as a therapeutic model. These operated at 25 MeV giving about 160 R/min at 1 m distance. Swiss Brown Boveri produced a 35-MeV unit developed by Wideröe and it was widely used for therapy. Siemens of Germany and General Electric of U.S.A. and the Japanese industries have also provided betatrons for therapy and radiography. Budker in the Soviet Union produced betatrons for commerce. The linear electron accelerator has replaced most of the betatrons, but there are still applications using the fine focal spot size that a betatron provides.

Interest in high current betatrons stems mainly from Budker (1956) with the series of experiments in the years immediately following. These efforts involved plasma injection which, as we have seen, awaited Rostoker's 1985 experiments for successful high current generation. The high currents in vacuum without the possibility of problems due to plasma were reached soon after at the U.S. Naval Research Laboratory by Kapetanakos's group. Almost all of these experiments used an added toroidal field, but the later work was greatly aided by the application of strong focusing or alternating gradient focusing.

In the interim in the Soviet Union Pavlevskii *et al.* (1977) did injection and capture experiments producing 100-A currents and brief acceleration sufficient to guarantee a good beam.

GLOSSARY

Bounce: The cyclic drift of the centroid of a beam traveling in the direction of a toroidal magnetic field. The drift is due to the forces perpendicular to this field. Its velocity is $\mathbf{V} = -\mathbf{F} \times \mathbf{B}/eB^2$.

CERN: European Organization for Nuclear Research.

Depth Dose: The ion density produced by x rays or charged particles penetrating tissue as a function of depth below the surface. It is usually expressed as percentage of surface dose.

Drift: See Bounce.

FFAG: Fixed field alternating gradient accelerator.

Injector: The source of particles sent into orbit for acceleration. An ion or electron gun or a boosting accelerator.

MURA: Midwestern Universities Research Association. A corporation of institutions for the advancement of high-energy facilities.

Peeler: A device to extract particles from the magnetic field in which they are in orbit—for example, a magnetic shield which prevents an entering particle from experiencing the accelerator's bending field.

Stacking: The process of building a large current in orbit by bringing successive small beams to adjacent orbits of similar energy. Phase space density invariance of injected beams determines the current possible in the final available phase space.

List of Works Cited

Budker, G. I. (1956), in *Proceedings of the CERN Symposium on High Energy Accelerators and Pion Physics*, Geneva, Switzerland: CERN Scientific Information Service.

Chernin, D., Sprangle, P. S., *Particle Accelerators*, **12**, 101–106.

Golden, J., Len, L. K., Deniz, A. V., Mathew, J., Smith, T. J., Loschialpo, P. F., Chang, J. H., Dialetis, D., Marsh, S. J., Kapetanakos, C. A. (1988), in: W. Bauer, W. Schmidt (Eds.), *Proceedings of the Seventh International Conference on High Power Particle Beams, "Beams '88,"* Karlsruhe, Germany: Karlsruhe Kongresszentrum.

Goward, F. K., Barnes, D. E. (1946), *Nature* **158**, 413.

Haxby, R. O., Laslett, L. J., Mills, F. E., Peterson, F. L., Rowe, E. M., Wallenmeyer, W. A. (1959), in: L. Kowarski (Ed.), *Proceedings of the CERN International Conference on High Energy Accelerators and Instrumentation, Geneva, 1959*, Geneva, Switzerland: CERN Scientific Information Service.

Ishizuka, H., Leslie, G., Mandelbaum, B., Fisher, A., Rostoker, N. (1985), *IEEE Trans. Nucl. Sci.* **NS-32**, 2727–2729.

Ishizuka, H., Prohaska, R., Fisher, A., Rostoker, N. (1988), in: W. Bauer, W. Schmidt (Eds.), *Proceedings of Seventh International Conference on High Power Particle Beams, "Beams '88,"* Geneva, Switzerland: CERN Scientific Information Service.

Jones, L. W., Terwilliger, K., Haxby, B. (1956), *Rev. Sci. Instrum.* **27**, 651–652.

Kapetanakos, C. A., Sprangle, P., Chernin, D., Marsh, S. J., Haber, I. (1983), *Phys. Fluids* **26**, 1678–1681.

Kapetanakos, C. A. (1989), *Bull. Am. Phys. Soc.* **34**, 1988.

Kapetanakos, C. A. (1990), *Phys. Rev. Lett.* **64**, 2374.

Kerst, D. W. (1941), *Phys. Rev.* **60**, 47.

Kerst, D. W. (1959), in: S. Flugge (Ed.), *Handbuch der Physik*, Vol. XLIV, *Nuclear Instrumentation I*, Berlin: Springer, pp. 193–217.

Kerst, D. W. (1983), in: R. J. Briggs and A. J. Toepfer (Eds.), *Proceedings of the Fifth International Conference on High Power Particle Beams, "Beams '83,"* Springfield, VA: National Technical Information Service, CONF-830911.

Kerst, D. W., Serber, R. (1941), *Phys. Rev.* **60**, 53.

Kerst, D. W., Adams, G. D., Koch, H. W., Robinson, C. S. (1950), *Rev. Sci. Instrum.* **21**, 462–480.

McMillan, E. (1945), *Phys. Rev.* **68**, 143.

Pavlevskii, A. T., Kuleshov, G. A., Sklizkov, G. V., Aysin, Y. A., Gerasimov, A. I. (1977), *Sov. Phys. Tech. Phys.* **22**, 218.

Quastler, H., Adams, G. D., Almy, A. M., Dancoff, S. M., Hanson, A. O., Kerst, D. W., Koch, H. W., Lanzl, L. H., Laughlin, J. S., Riesen, D. E., Robinson, C. S., Jr., Austin, V. T., Kerley, T. G., Lanzl, E. F., McClure, G. Y., Thompson, E. A., Skaggs, L. S. (1949), *Am. J. Roentgenol. Radium Ther.* **61**, 591–625.

Reynolds, P., Skarsgard, H. M., (1959), *J. Nucl. Energy Part C* **1**, 36–39.

Roberson, C. W., Mandelli, A., Chernin, D. (1983), *Phys. Rev. Lett.* **50**, 507.

Rostoker, N. (1980), *Comments Plasma Phys.* **6**, 91.

Schwinger, J. (1949), *Phys. Rev.* **75**, 1912.

Sprangle, P., Kapetanakos, C. A. (1978), *J. Appl. Phys.* **49**, 1–6.

Terwilliger, K. M., Jones, L. W., Pruett, C. H. (1957), *Rev. Sci. Instrum.* **28**, 987–997.

Veksler, J. (1945), *J. Phys. Acad. Sci. USSR* **9**, 153.

Further Reading

(1956), *Proceedings of CERN Symposium on High Energy Accelerators and Pion Physics*, Geneva: CERN Information Service.

(1959), *Proceedings of International Conference on High Energy Accelerators and Instrumentation, CERN, European Organization for Nuclear Research*, Geneva: CERN Information Service.

Kerst, D. W. (1946), *Nature* **157**, 90.

23
Biological Effects of Electromagnetic and Particle Radiation

Jürgen Kiefer

23.1 Introduction 745
23.2 Radiation Types and Sources 745
23.3 Interaction with Biosystems at the Physical Level 748
23.3.1 Ionizing Radiation 748
23.3.1.1 Electromagnetic Radiation 748
23.3.1.2 Particle Radiation 748
23.3.2 Optical Radiation 749
23.3.3 Radio Frequency and Microwaves 750
23.4 Deposition of Energy 750
23.4.1 Ionizing Radiation 750
23.4.2 UV and Visible Radiation 754
23.4.3 Infrared 755
23.4.4 Radio Frequency and Microwaves 755
23.5 Effects on Molecules of Biological Importance and Cell Organelles 755
23.5.1 Water 756
23.5.2 Photochemistry and Radiation Chemistry of DNA 757
23.5.3 Chromosomal Aberrations 757
23.5.4 Membranes 759
23.5.5 Effects of Heat Treatment 760
23.5.6 Specific Molecular Aspects of Radio Frequency and Microwave Actions 760
23.6 Cellular Effects 761
23.6.1 Experimental Endpoints of Cellular Radiation Action 761
23.6.1.1 Survival 761
23.6.1.2 Mutations 764
23.6.1.3 Neoplastic Transformation 764
23.6.2 Modifications of Cellular Radiation Effects 765
23.6.2.1 The Temporal Pattern of Exposure 765
23.6.2.2 Chemical Protectors and Sensitizers 766
23.6.2.3 Repair Processes 767
23.6.2.4 Radiation Quality 768

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

23.6.2.5 "Nonclassical" Effects in Cellular Radiobiology: Bystander Effect, Genomic Instability, Low Dose Sensitivity, and Adaptive Response 769
23.6.3 Heat Effects at the Cellular Level 770
23.6.4 Specific Cellular Effects of Radio Frequency and Microwaves 771
23.7 Organ and Whole-Body Effects 771
23.7.1 Acute Effects 772
23.7.1.1 Skin 772
23.7.1.2 Eye 773
23.7.1.3 Nervous System and Other Organs 774
23.7.1.4 Radiation Syndromes and Lethality 774
23.7.2 Radiation and Progeny 776
23.7.2.1 Fertility 776
23.7.2.2 Teratogenic Effects 776
23.7.2.3 Genetic Effects 777
23.7.3 Radiation Cancerogenesis 779
23.7.3.1 Ultraviolet Radiation 779
23.7.3.2 Ionizing Radiation 780
23.7.4 Tissue and Whole-Body Effects of Infrared Radiation 785
23.7.5 Tissue and Whole-Body Effects of Radio Frequency and Microwaves 785
23.8 Applied Radiation Biology 786
23.8.1 Molecular Techniques 786
23.8.1.1 Action Spectroscopy 786
23.8.1.2 Molecular Weight Determination by "Radiation Inactivation" 787
23.8.2 Radiation Disinfection and Sterilization 788
23.8.3 Therapeutic Applications 789
23.8.3.1 Nonmalignant Diseases 789
23.8.3.2 Cancer Therapy 790
23.8.4 Biodosimetry 792
23.8.4.1 Fluorescence *In situ* Hybridization (FISH) for Chromosomal Translocations 792
23.8.4.2 ESR Dosimetry 792
23.9 Radioecology, Radiation Protection, and Guidelines 793
23.9.1 Radioecology 793
23.9.1.1 Nonionizing Radiations 793
23.9.1.2 Ionizing Radiation 793
23.9.2 Principles of Radiation Protection Regulations 798
23.9.2.1 Ionizing Radiations 798
23.9.2.2 Nonionizing Radiations 800
Glossary 802
References 804

23.1
Introduction

The field to be treated in this article is wide and sometimes controversial. Radiation types encompass a large range of energies, and the interaction mechanisms differ considerably. Radiation biology *sensu strictu* deals with the effects of ionizing and ultraviolet (UV) radiations, and it is in this field where most information is available. Consequently, the article concentrates particularly on this aspect, also because the approaches used for the other radiation types were largely adopted from radiobiology.

The treatment follows the "classical way" from physics via chemistry to biology. The time course of events spans more than 30 orders of magnitude (see Figure 23.1), which illustrates the difficulty involved in solving the the task. Section 23.9.2 gives a few examples of radiation application, and protection principles are summarized in Section 23.9.2. Completeness is not intended and is impossible to achieve. Only fundamental problems are described, in a way that they can be understood also by the nonspecialist reader. For further reading, there are a few more general books covering more or less the whole field, for example Hall and Giaccia (2005) and Kiefer (1990). Reference reviews for special areas are given in each chapter.

23.2
Radiation Types and Sources

This section is restricted to those sources that present a possible biological hazard. Technical details have to be omitted here because of space limitations.

Under "optical radiations" are summarized those kinds that can be analyzed by means of optical instruments like prisms, gratings, and lenses. They are subdivided in terms of wavelength as follows:

1. UV: ≈100–380 nm
2. Visible (VIS): 380–800 nm
3. Infrared (IR): 800 nm to ≈100 μm

The UV region is often further characterized as

1. UV-A: 315–380 nm
2. UV-B: 280–315 nm
3. UV-C: 200–280 nm
4. Vacuum UV: <200 nm

Similar classifications have been suggested for the IR region:

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

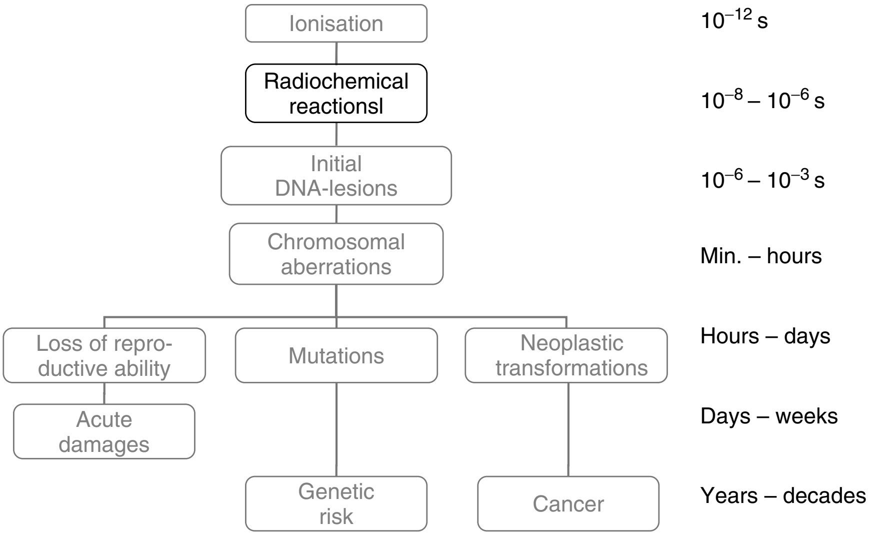
Ionisation
Radiochemical reactionsl
Initial DNA-lesions
Chromosomal aberrations
Loss of repro-ductive ability
Mutations
Neoplastic transformations
Acute damages
Genetic risk
Cancer
10^{-12} s
$10^{-8} – 10^{-6}$ s
$10^{-6} – 10^{-3}$ s
Min. – hours
Hours – days
Days – weeks
Years – decades

Fig. 23.1 The stages of biological action of radiation and the approximate timescales.

1. IR-A: 0.78–1.4 μm
2. IR-B: 1.4–3 μm
3. IR-C: 3–1000 μm

The most important natural source of optical radiation is the Sun, whose emission spectrum closely resembles that of a blackbody with a temperature of about 6000 K. Its radiation is attenuated by the atmosphere. In the short-wavelength part, ozone (O_3) absorption causes not only a reduction in fluence but also a shift to longer wavelengths, which is very important for the maintenance of life since the damaging part of the spectrum is largely eliminated on the surface of the Earth. Water and carbon dioxide are mainly responsible for the absorption in the IR region. The energy fluence rate outside the atmosphere (commonly called the *solar constant*) is 1394 $W\,m^{-2}$, which is reduced to about 1/10th at sea level depending on geographical latitude. Technical UV sources include gas-discharge lamps (mercury, hydrogen, and noble-gas lamps), incandescent sources, and fluorescent lamps. Examples of emission spectra can be found in the World Health Organization report (WHO, 1982). Lasers present an exceptional case, also with regard to possible hazards (see Section 23.9.2.2).

Radiofrequency (RF) and microwave (MW) radiations are characterized by the electric and magnetic field strengths at the place of interactions. Ideally, the two vectors are in phase and perpendicular to each other. In this case, the "power intensity" is given by their product, usually expressed in watts per square meters. One has to note, however, that the simple situation is only found at some distance from the emitting source and that great distortions may occur upon interaction with matter. High-power sources are somewhat arbitrarily defined as those that can produce a mainbeam power density of at least 1 $W\,m^{-2}$ at a distance of 100 m. They include broadcast transmitters, tracking and acquisition radar, and communication systems such as satellite communications earth terminals. Low-power sources (<1 W m^{-2} at 100 m distance) include traffic radar and MW relay systems as well as MW ovens. Particularly, the last example may be of public concern because of its ubiquitous use in many countries. Although these appliances are enclosed and leakage is not intended, it cannot always be avoided. With an assumed level of 50 W m^{-2} at a distance of 50 mm, the power density will fall to less than 15 W m^{-2} at 30 cm and to about 0.1 W m^{-2} at 1 m. Traffic radar leads to an exposure of 10 mW m^{-2} at distances of 10 m.

Mobile communication by cellular phones has increased tremendously in recent years. There are two frequency bands used, both in the MW range, namely around 900 and 1800 MHz. Typical emissions by base stations lie in the order of 10–50 W but the field strengths decrease rapidly with distance. The signals are received and amplified by the phone so that the main source of exposure is from the handheld mobile phone with values up to 2 W near the head of the user. The emission powers of TV transmitters are considerably higher than that of base stations for mobile communication.

Natural ionizing radiation originates either from outer space or from terrestrial sources. Space radiation consists essentially of protons, helium ions, and a few heavier nuclei of which iron plays the most important role. Primary cosmic radiation does not reach the surface of the Earth. The particles are, depending on their energy, deflected by the magnetic field, or they react with the constituents of the atmosphere to create cosmogenic radionuclides and other secondary products, mainly mesons, neutrons, and electrons. The charged components are captured in the two Van Allen belts at heights of about 2.8 (mainly electrons) and up to 12 Earth radii (electrons and protons). Since the magnetic field is not exactly collinear with the Earth's axis and the dipole field is distorted, the height of the radiation belts varies with geographical locations with the lowest altitude in the South Atlantic ("South Atlantic Anomaly"). The intensity of cosmic rays is also modulated by solar activity, being highest at solar minimum and vice versa. The most important cosmogenic radionuclides are ^{14}C, ^{7}Be, ^{22}Na, and ^{3}H. Terrestrial radiation sources are the so-called primordial radionuclides, which survived from the formation of our planet because of their long half-life, and their daughter products. The most important are ^{40}K, ^{87}Rb, ^{238}U, ^{235}U, and ^{232}Th. The last three are starting products of decay series whose members contribute also to natural radiation. The more relevant among them are ^{226}Ra and ^{222}Rn. Particularly, the latter and its immediate daughter products are now considered to be the most important part of the natural radiation burden.

The cosmic component varies with altitude while the terrestrial contribution depends especially on soil composition. Some areas in India (Kerala coast), Brazil (monazite sands), and China display radiation levels that are at least one order of magnitude higher than the average. The relative contribution from natural sources to human radiation burden is discussed in Section 23.9.2.

Technical sources of ionizing radiations include nuclear energy production plants, medical applications in diagnosis and therapy, the use of γ irradiation, mainly for sterilization purposes, and of X-rays for quality control in industry.

The unit of activity is the Becquerel (Bq):

$$1\text{Bq} = 1\ \text{s}^{-1} \tag{1}$$

The older unit – no longer to be used in official or legal documents – was the Curie (Ci), originally based on the activity of 1 g of ^{226}Ra and later defined by

$$1\text{Ci} = 3.7 \times 10^{10}\ \text{Bq} \tag{2}$$

For radioactive gases like ^{222}Rn (radon) and ^{220}Rn (thoron), "working levels" (WL) are still widely used as units of activity concentrations, although they also are not part of the now-accepted international system. Without further derivation, the relations are given here, which are different for the two nuclides mentioned:

$$1\text{WL}(^{222}\text{Rn}) = 100\text{pCi}\,\text{l}^{-1} = 3.7\ \text{Bq}\,\text{l}^{-1} \tag{3}$$

and

$$\begin{aligned}1\text{WL}(^{220}\text{Rn}) &= 7.5\ \text{pCi}\,\text{l}^{-1}\\ &= 0.2775\ \text{Bq}\,\text{l}^{-1}\end{aligned} \tag{4}$$

23.3
Interaction with Biosystems at the Physical Level

23.3.1
Ionizing Radiation

23.3.1.1 Electromagnetic Radiation

When electromagnetic radiation traverses matter it is attenuated in an exponential manner:

$$I/I_0 = \exp(-\mu x) \tag{5}$$

where I_0 is the intensity at the entrance, I is the intensity after passing through a layer of thickness x, and μ is the attenuation coefficient. Equation (5) applies only to monochromatic radiation and homogeneous material; in the case of composite matter and radiation it has to be replaced by appropriate integral expressions. It is customary to use the mass attenuation coefficient μ/ρ instead of μ, with ρ being the density of the absorber. Attenuation is brought about by scattering (without energy loss) and energy transfer to the exposed medium. With regard to radiation action, the distinction must be made between "energy transferred" and "energy (locally) absorbed," which are taken into account by introducing the energy transfer coefficient μ_k and the energy absorption coefficient μ_{en}, respectively. The first signifies the complete energy transferred to secondary particles in the form of kinetic energy by the initial ionization event, while the second includes only the fraction that is locally absorbed.

All photon interactions (photoabsorption, Compton scattering, and pair formation) lead to the formation of electrons (and positrons that play, however, a less important role since they combine readily after short distances). They are responsible for the actual energy deposition in the medium; the initial processes contribute directly only to a very small extent to the total number of ionizations.

23.3.1.2 Particle Radiation

Ions The main interaction process in terms of energy deposition in exposed matter is the collision with atomic electrons. The energy loss per distance traveled is given by the Bethe–Bloch formula. Energy loss is commonly given as "mass stopping power," which is related – but not equal – to linear energy transfer (LET) (Section 23.5.4). The ICRU has compiled data for protons and heavier ions (ICRU, 2000, 2005). For ions with the same velocity, the energy loss may be calculated from proton data with the effective charge as scaling parameter.

Electrons Electrons play an essential role in the ultimate deposition of energy with all kinds of ionizing radiation since they are always formed in the ionization process. There are three types of interactions that are important in the present context: collisions with shell electrons, bremsstrahlung production, and emission of Cherenkov radiation. In terms of energy deposition, the first process is the most relevant. Bremsstrahlung production becomes important only at higher energies. It must be taken into account for dose calculations since the generated electromagnetic radiation may transport a great fraction of the originally transferred energy out of the exposed body (see also the discussion of "dose" and "kinetic energy released per mass" (KERMA) in Section 23.5). Cherenkov radiation is generated if the speed of a charged particle passing through matter exceeds that of light in the medium. This criterion sets a lower limit below which this process does not occur (about 500 keV for water). In terms of energy loss, it is usually not important, but it may lead to the generation of UV light, which may then be responsible for a certain biological radiation damage.

Neutrons The most important interaction with regard to biological damage by neutrons is the collision with atomic nuclei in the exposed medium. As would be expected from classical collision dynamics, protons are the most relevant constituents here; other nuclei contribute significantly to energy transfer only with higher neutron energies. The actual deposition in the medium is via these charged secondary particles that are displaced by the collision, a situation similar to the that of electrons with electromagnetic radiation. Nuclear processes initiated by neutron interactions do not contribute to a large extent in common energy ranges.

Mesons π^- mesons have, for some time, been used in radiation therapy because of their favorable depth dose distribution but this technique is no longer applied and been superseded by proton and carbon ion therapy. Mesons are, however, an important part of secondary cosmic radiation. The main interaction of π^- mesons occurs with atomic nuclei in the exposed biological material, resulting in highly excited states. The energy gained is deactivated by nuclear spallation, resulting in a number of charged fragments that deposit their energy over short distances with high ionization density. Interactions with oxygen and nitrogen play the most important role in biological matter.

23.3.2
Optical Radiation

Optical radiation is absorbed selectively by certain compounds, called *chromophores*, in the exposed medium. Energy deposition from optical radiation hence depends entirely on the absorption properties of the exposed medium.

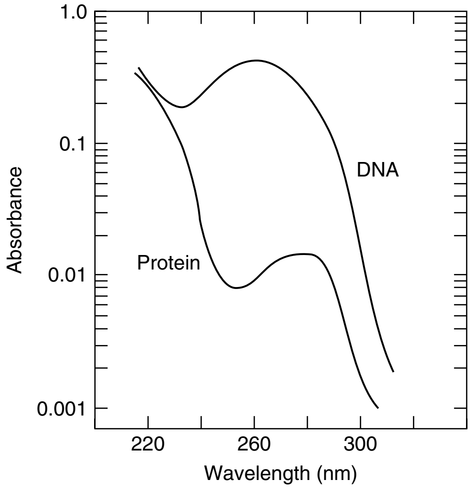

Fig. 23.2 Absorption spectra of DNA and proteins (bovine serum albumin) (after Harm, 1980).

Spectra for materials of biological interest are given in Figure 23.2.

With intense laser pulses two quantum excitations may take place, which can lead to photoionization. Whether they are biologically relevant is not yet clear. Excited molecules – particularly in the triplet state – are the starting points for photochemical reactions (see Section 23.6). If the absorption takes place in short times and is locally concentrated as with lasers, heat effects become the most prominent type of damage.

IR radiation causes excitation of vibrational and rotational molecular states, which is dissipated by heat production. This kind of damage, which is less localized and specific than that caused by UV and visible radiation, may then act as starting point of further tissue destruction.

23.3.3
Radio Frequency and Microwaves

RF and MW interact with matter by inducing electric and magnetic fields. A perfect dipole does not remove any energy, and the impinging radiation is not attenuated. In a medium of finite conductivity – as is all biological material – charge motion in the form of free ions (conductive loss) and molecular rotations (dielectric loss) is caused. The energy absorbed is finally dissipated as heat, leading to local or global temperature increase. The relative contributions depend on the electric properties of the medium, which in turn are frequency dependent. In principle, it should be possible to calculate the energy absorption for any biological object, but the practical restrictions are such that only very simple geometries can be adequately treated. Consequently, one has to rely on experimental determinations.

23.4
Deposition of Energy

23.4.1
Ionizing Radiation

As pointed out in the previous section, energy deposition by ionizing radiations is exclusively due to charged particles created in the exposed medium via the primary interaction processes. In most cases these are electrons, but with neutrons incident, protons and even other nuclei may also contribute. In the case of charged-particle irradiation, energy is deposited by their primary interactions and those of the secondary species. Both undergo further interactions by which the initial energy distribution is modified. This process is called *slowing down.* The resultant spectra may be calculated by Monte Carlo simulations taking into account the cross-sections of all relevant process for the medium under investigation. After a certain depth, a steady state is reached, called *secondary-particle equilibrium,* which is characterized as a situation where in a given mass element each particle leaving it is replaced by an incoming one of exactly the same kind and energy. Secondary-particle equilibrium can never be attained close to interfaces. The theory of slowing down is well established for electrons in water (see, e.g. Kiefer, 1990 for further references); more recently, the computations were also extended to materials of biological interest.

The biological activity of different kinds of ionizing radiations depends on the spatial pattern of energy deposition. A macroscopic parameter that is used to describe it is the LET, which is defined as the energy deposited *locally* per path length traveled. LET applies to the *charged* components only, since they are almost exclusively responsible for energy deposition. The term *locally* requires some specification. It depends obviously on the size of the site under consideration. If it is larger than the range of the most energetic particle, all interactions are included and LET becomes numerically equal to the stopping power, which gives the *energy loss* per unit distance of the incoming particle. It has become customary to use energy restrictions (in electronvolts) for the LET, indicated by suffixes. Thus, LET_{100} means that only those energy deposits are included where the energy of the electrons formed does not exceed 100 eV, corresponding to a range of about 5 nm in water.

Since energy deposition is always due to the interaction of a whole spectrum of primary and secondary particles (even in the case of monoenergetic ions) there is always a *distribution* of LETs. They may be characterized by certain average values, the simplest one being the "track" or "number average" $\overline{L}_T$, defined as the expectation value of the distribution $f(L)\,dL$:

$$\overline{L}_T = \int Lf(L)\,dL \tag{6}$$

Here, each event is counted with equal weight. If one is interested in the relative contribution to the total energy deposition, the "dose" or "energy average" $\overline{L}_D$ is used:

$$\overline{L}_D = \frac{\int L^2 f(L)\,dL}{L_T} \tag{7}$$

where each event is weighted by its energy deposition. The denominator is just a normalization factor. The two averages are only equal if the variance of the distribution vanishes.

LET is usually measured in kiloelectronvolt per micrometer, but megaelectronvolt square centimeters per gram is also widely used. Examples of LET distributions are shown in Figure 23.3.

The most important quantity to describe energy deposition by ionizing

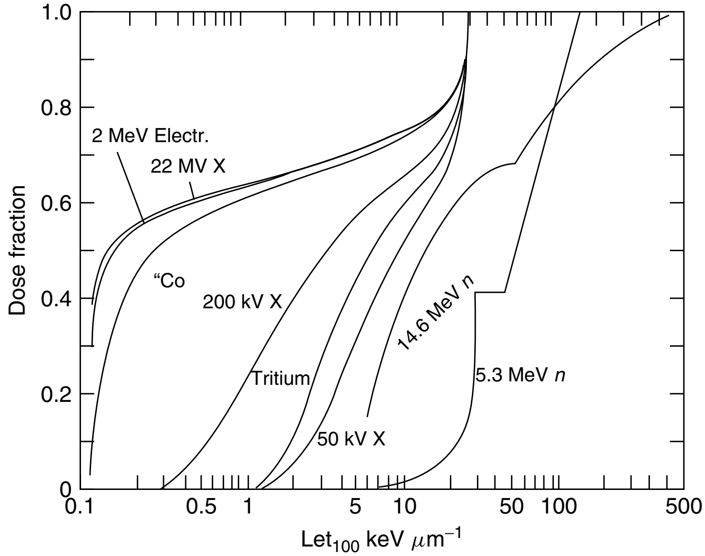

Fig. 23.3 LET distributions of several radiation types in water (after ICRU, 1970).

radiations is “dose,” which is defined as absorbed energy per mass. In sites of microscopic dimensions, which are relevant for biological targets, however, it loses its well-defined meaning because of the statistical nature of interactions (“microdosimetry” Rossi and Zaider, 1996). To account for this, another quantity has been introduced termed as *specific energy* z, which is the microscopical counterpart of dose. It is a *stochastic* quantity that is characterized by a distribution function $f(z)\ dz$ whose actual form depends on, among other things, the size of the mass element, the kind of particle, the particle's energy, and the medium exposed. Similarly, as with LET (see above), certain averages may be used for shorthand characterization, namely the track average or frequency mean $\bar{z}_F$:

$$\bar{z}_F = \int z f(z) dz \tag{8}$$

and the energy average or dose mean $\bar{z}_D$:

$$\bar{z}_D = \frac{1}{\bar{z}_F} \int z^2 f(z) dz \tag{9}$$

Generally, both the distribution and the expectation values depend on dose. The microdosimetric approach just introduced allows a clear and meaningful definition of this quantity: *Dose is the expectation value of specific energy* and hence equal to $\bar{z}_F$. Macroscopically speaking, one can also say that it is the mean energy absorbed per mass element; this used to be the classical definition.

The microdosimetric counterpart to LET is “lineal energy” y, defined as the energy deposited *by single traversals* in the given site (assumed to be spherical) divided by the average path length.

Mainly for reasons of practical measurement, another “dose” quantity had formerly been widely used, namely “charge liberated per unit mass,” which is now called *exposure*. By convention, only charged particles of one sign are counted, and the medium in which it is determined has to be air. Exposure is thus not really a dose but rather a characterization of the radiation field at a given position.

The international unit for dose is joules per kilogram with the special name "gray" (Gy). An older, no longer sanctioned, unit is the radians:

$$1\text{Gy} = 100\text{ rad} \tag{10}$$

The new unit for exposure is coloumb per kilogram; no special name has been suggested. The old one was the roentgen (R):

$$1\text{R} = 2.58 \times 10^{-4}\text{C}^{-1}\text{kg} \tag{11}$$

For a given type of radiation and a specified medium, conversion factors between dose and exposure can be given (Table 23.1).

Dose D, particle fluence ϕ, and track average LET L_T are related in a simple way:

$$D = L_T\phi/\rho \tag{12}$$

where ρ is the density of the medium.

The quantity "dose" as defined above refers to the energy actually absorbed in the medium and does not necessarily equal the energy lost by the interacting particle or quantum. The secondary particles produced by the primary collisions may gain sufficient kinetic energy to travel considerable distances before they are finally stopped. This situation is taken into account by another quantity called *KERMA*, which has the same dimensions as dose but relates to the energy loss of the incoming species. For photons, the difference between KERMA and dose is reflected in the coefficients μ_K and μ_{en}, respectively. KERMA and dose are only equal under the conditions of secondary-particle equilibrium *and* in the absence of bremsstrahlung losses. It is clear from this that they are always different at interfaces.

In an exposed body, the dose varies with penetration depth. With photon radiations, energy is deposited by electrons with ranges depending on the quantum energy. The dose maximum is thus shifted toward larger depth with increasing energy (Figure 23.4). This behavior plays an important role in the radiation therapy of deep-seated tumors (see Section 23.8.3.2)

Tab. 23.1 Relation between dose and exposure for various substances and photon energies (after ICRU, 1979).

Photon energy [keV]	Water [Gy/(C kg^{-1})]	Bone [Gy/(C kg^{-1})]	Muscle [Gy/(C kg^{-1})]
10	35.0	140	34
50	34.7	134	34.3
100	36.9	56.1	36.5
200	37.4	37.9	37.0
400	37.5	36.1	37.0
600	37.5	35.9	37.0
1000	37.5	35.8	37.0
2000	37.5	35.6	37.0

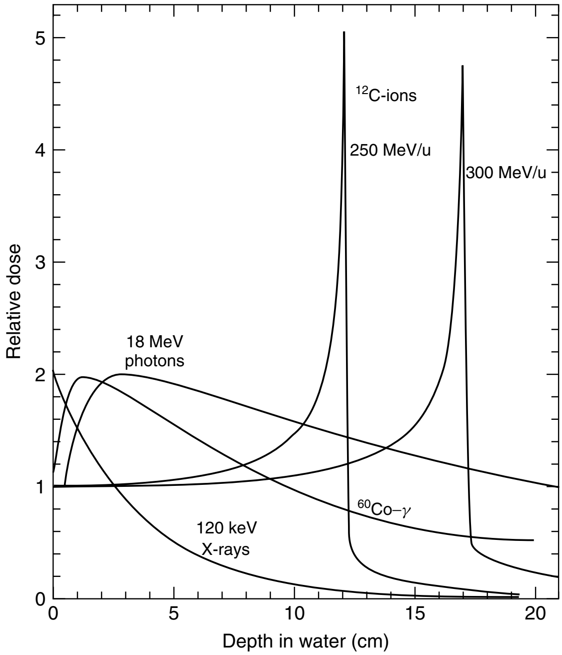

Fig. 23.4 Depth dose distributions for various types of ionizing radiations (GSI Darmstadt, with permission).

Depth dose distributions of protons and heavier nuclei are characterized by very sharp maxima ("Bragg peak"), which is one of the reasons for their increasing application in radiation therapy.

23.4.2
UV and Visible Radiation

Energy deposition by UV and visible radiation depends on the absorption properties of the molecular and atomic constituents of the exposed medium. Photobiological effects can only be caused if there is an overlap between the source emission and the tissue absorption spectrum for the wavelength range considered. This constitutes, however, only a *necessary* but not a sufficient condition for photobiological actions. Only if essential structures are altered via photochemical reactions will biological effects be seen. They may be caused by a direct destruction of the absorbing species or via secondary processes where light absorption creates intermediate reactive species, which in turn then may damage other constituents. This second case is called a *photosensitized reaction*. An important example is the formation of singlet oxygen, which is the mediator of the "photodynamical action." Primary chromophores may be identified by measuring an "action spectrum" (see Section 23.9).

Because of the dependence of energy deposition on the specific absorbing properties, "dose" as used for ionizing radiation cannot be defined in a reasonable way for UV and visible light although the term is still widely but wrongly used. The quantity normally applied is "energy fluence" (impinging energy per unit area, measured in joules per square meter), often called *dose*. In some cases, particularly in photochemistry, the "quantum fluence" (number of photons per unit area, unit m^{-2}) is also used. In some instances, particularly in conjunction with radiation protection, energy fluences are weighted by applying "sensitivity function" (see Section 23.9.2.2 for details). As the units used are identical (J/m^2) there is a definite danger of misinterpretation, and careful cross-checking is advisable.

23.4.3
Infrared

Energy deposition by IR is determined by the absorption properties of the exposed tissue, with local heating being the main result. It is not clear at present whether nonthermal effects on molecules and cells play a significant role, although they cannot be generally excluded.

23.4.4
Radio Frequency and Microwaves

Energy is deposited by the induction of currents where freely movable charge carriers are available and by the induction of molecular rotations.

The absorption of RF and MW radiation is related to the complex permittivity ε^* of the material to be considered:

$$\varepsilon^* = \varepsilon'(\nu) + j\varepsilon''(\nu) \tag{13}$$

The power absorbed, P, is given by

$$P = 2\pi\nu\varepsilon_0\varepsilon'' E^2 V \tag{14}$$

Here $\varepsilon'(\nu)$ and $\varepsilon''(\nu)$ are the real and the imaginary part of the complex permittivity, respectively, both being dependent on frequency ν, ε_0 is the electric field constant, E the electric field strength at the point of interaction, V the volume, and i the imaginary unit ($\sqrt{-1}$).

The coefficients have been measured for a number of materials of biological interest in the frequency range from 10 Hz to 20 GHz (Gabriel, Gabriel, and Corthout, 1996a; Gabriel, Lau, and Gabriel, 1996b, 1996c)

The quantity used to describe the amount of interaction is "specific absorption rate" (SAR), measured in watts per kilogram.

Depending on the geometrical dimensions of the exposed body, resonance phenomena must be taken into account. Although the energy absorbed will ultimately be dissipated in the form of heat, leading to local or global temperature increases, this is not necessarily the only cause of biological effects. Changes in charge distribution at membranes and possible resonances at the molecular level also have to be taken into account, at least theoretically. It is, however, not at all clear whether they play a role at subthermal powers.

23.5
Effects on Molecules of Biological Importance and Cell Organelles

Biological systems contain a plethora of molecules at different levels of complexity. All of them may be altered by radiation action. It is obviously necessary to make a choice here, which is guided either by the relative abundance of a particular species or by its biological importance. The discussion is, therefore, restricted to water, as the most abundant molecule type, and deoxyribonucleic acid (DNA), as the carrier of genetic information. This selection is, of course, somewhat arbitrary. Similar deliberations may be made with regard to subcellular structures. Chromosomal aberrations are dealt with because of their fundamental and also practical importance, as are membranes because they are often implicated as targets for radiation action and also since they constitute the interface between the biological system and its environment.

For the quantitative description of radiation-induced molecular alterations in photochemistry the quantum yield Q has been introduced, which is defined as the mean number of altered molecules per quantum absorbed. Its radiation-chemical counterpart is the G value, originally used as the number of altered entities (molecules or atoms) produced per 100 eV absorbed energy. With the introduction of the new SI units this has been changed to molar concentration of species produced per unit dose. The numerical relationship is given by the following expression:

$$G/(100\ \mathrm{eV})^{-1} = 9.64 \times 10^6 \frac{c/(\mathrm{mol\ dm^{-3}})}{(D/Gy)^{\rho/(\mathrm{kg\ dm^{-3}})}} \qquad (15)$$

where c is the molar concentration and ρ the density of the medium.

23.5.1
Water

Biological systems contain up to 80% water, so it must be considered as a major primary target for interaction with radiation. Its interaction with optical radiation depends on the absorption spectrum, which shows complete transparency between 200 and 800 nm. No effects are thus to be expected in this region. In the far UV, excitations may lead to radical formation, as with ionizing radiation (see below), while IR absorption causes heat production. The photochemistry of water does not play an important role under normal circumstances. Radiation-chemical processes, on the other hand, are considered to be responsible for a great part of biological damage at the cellular level, at least with sparsely ionizing radiation. The initial products formed are H and OH radicals and electrons. The latter interact electrostatically with surrounding water dipoles to form a transient hydration shell, which is indicated by the suffix "aq." Because of interradical reactions, the initial yields are always higher than the final ones, even in ultrapure water.

The hydrated electron possesses interesting properties. Because of the shielding by the hydration shell, its reaction probability with solutes is smaller than that for the other species; this results in a comparatively long lifetime (about 600 μs in pure water). The positive potential created by the surrounding water dipoles leads to a quantization of the possible energy states and hence to light absorption. The specific properties described have made the hydrated electron a key product in radiation chemistry.

The primary radicals react with each other to form the molecular products, molecular hydrogen (H_2) and hydrogen peroxide (H_2O_2). Oxygen, if present, is an important reaction partner that must be taken into account when discussing the radiobiological oxygen effect (Section 23.7). The probability of interradical reactions depends, of course, on the proximity of their initial formation and is hence increased in the tracks of high-LET radiations. The biological importance of radiation-induced water radicals lies in the fact that they react readily with solutes in their vicinity and are thus able to damage DNA or other critical biomolecules. This *indirect radiation effect* plays a significant role in dilute solutions and even in cells, where its existence can be demonstrated by the protective action of radical scavengers (see Section 23.6.2).

23.5.2
Photochemistry and Radiation Chemistry of DNA

DNA is the carrier of cellular genetic information. This justifies the special interest in radiation-induced alterations of this molecule, which may ultimately lead to loss of viability, induction of mutations, and initiation of cancer.

The absorption spectrum of DNA is shown in Figure 23.2, where it is seen that in the far UV range the region around 260 nm is expected to be most relevant for damage production. The most important products are pyrimidine dimers, pyrimidine adducts, single-strand breaks (SSBs), and crosslinks with adjacent proteins. Dimers are most important in microorganisms; their role in higher systems is less clear. They are the unique substrates for photoreactivation (see Section 23.6.2.3) and are also subject to other repair processes. Pyrimidine adducts have recently been implied to be the starting points for mutations, but this requires further substantiation. The other types play a lesser role in the far UV but may become more relevant with longer wavelengths. SSBs are produced only with low yield directly upon UV irradiation but are introduced during repair processes (Kiefer, 2007)

The radiation chemistry of DNA is less transparent than its photochemistry. Very many effects have been identified; some of the more relevant ones are SSBs, double-strand breaks (DSBs), base damage, base loss, denatured zones, intramolecular cross-links, and DNA–protein cross-links. Most of them have only been identified in aqueous solutions; data for cellular systems are scarce. DSBs have received the greatest attention, and it has been claimed that they are the main causes of cell inactivation. In dilute DNA solutions they are produced with a linear-quadratic dependence on dose, intracellularly in strict proportionality. (This statement has recently been called into question; clarification must await further experimentation.) DSB may be repaired (see Section 23.6.2.3) although not necessarily correctly. Presumably there is no single damage type that can be held responsible for the biological action of ionizing radiation. DNA damage may also be produced indirectly via the reaction of water radicals (see above) where OH seems to be the most important species (von Sonntag, 2007).

The dependence on LET has only been studied in some detail for strand breaks. It is seen that the yield of SSB decreases while that for DSB increases. This finding is important since it may serve to explain the higher biological effectivity of high-LET radiations. It is also taken as an indirect proof of the relevance of DSB.

23.5.3
Chromosomal Aberrations

The genetic material of higher cells is organized in supramolecular structures consisting of DNA and proteins, which are called *chromosomes*. They may be seen visually during the division process in the microscopic picture after appropriate staining. Radiation absorption can cause aberrations, broadly classified into two categories: numerical and structural. In the first class the number of chromosomes, which is a constant for a given species (e.g. 46 in man), is changed. Loss of a chromosome is normally lethal; a supernumerary one is often related to severe genetic diseases, the best known being Down’s syndrome (commonly called “mongolism”), which is caused by an extra chromosome 21 (“trisomy 21”). The role of radiation in its etiology is not quite clear. Generally it appears that radiation is a weak agent for the production of numerical chromosome aberrations.

Structural aberrations are more important. Some of them are diagrammatically depicted in Figure 23.5. They may be formed as a direct consequence of one initial event (“break”) but in most cases they are caused by the interaction of two primary lesions. Not all of them are directly visible – for example, if only the position of a piece is changed as in inversions or translocations. This does not, however, mean that they are harmless; quite to the contrary, it is now well established that translocations are connected with a number of cancer types. Practically, *dicentrics (dics)* play the most important role. They are caused by the interaction of two “breaks” in separate chromosomes and subsequent rejoining, leading to a chromosome with two centromeres. They are lethal alterations since they cannot be disentangled

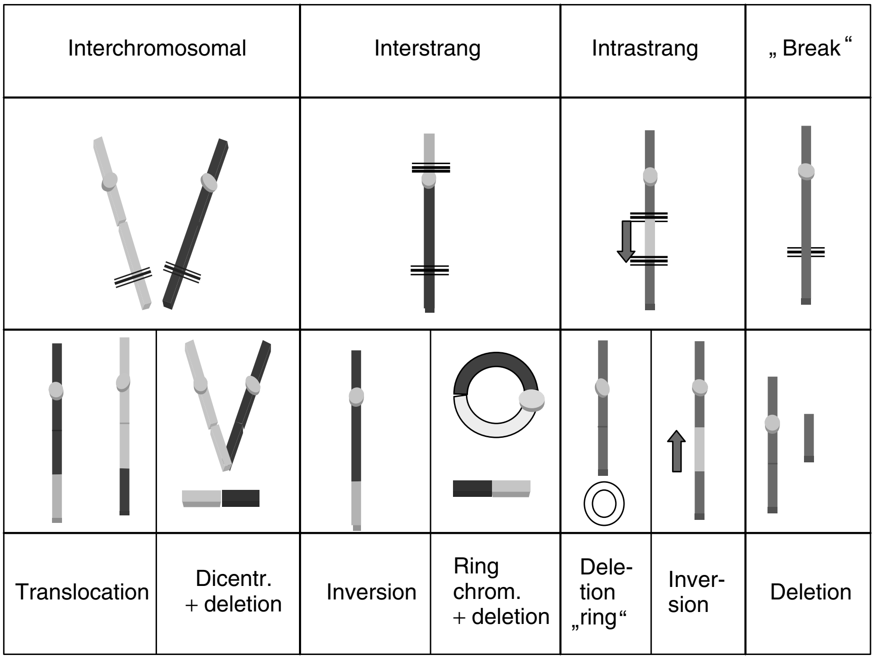

Fig. 23.5 Schematic representation of some chromosome aberrations (after Savage, 1978).

in division. More importantly, they can be recognized with comparative ease and thus serve as a convenient indicator of radiation damage. They even play a role in practical radiation protection. If lymphocytes of the peripheral blood are extracorporally stimulated into division, dicentrics can be scored and their relative number can be used as "biological dosimeter."

Radiation-induced breaks if not rejoined leave chromosomal fragments that are also replicated but usually not correctly distributed at mitosis. Quite often these fragments are encapsulated in membranes, thus forming "micronuclei." They can be easily scored in interphase and are also used for biological dosimetry (Fenech, 2007; Hayashi, 2007)

If replication is stalled, for example, under the influence of mutagens, chromatids may be exchanged between the sister chromosomes. These "sister chromatids exchanges (SCEs)" are thus sensitive indicators of disturbances of replication. Their detection requires special staining techniques (Wojcik and Obe, 2007).

Modern immunofluorescence staining techniques allow the detection and scoring of translocations and even inversions and other intrachromosomal alterations. Monoclonal antibodies coupled to fluorescent dyes can be raised against human chromosomes or parts of them so that every chromosome can be selectively "painted." Translocations are then visualized by color changes along the chromosomal structures. Studies have shown that they persist indeed much longer than dicentrics. Their dose dependence is similar to that of dicentrics and their detection limit lies in the same range (about 100 mGy for γ-rays).

Compared to other cellular endpoints (like loss of viability or mutations), chromosome aberrations constitute a relatively early event that can be assessed a few hours after exposure and may be less subject to postirradiation modification. They are induced by both UV and ionizing radiation. Dicentrics and translocations are formed according to a linear-quadratic relationship to dose with low-LET radiations. This functional dependence changes to straight linear for high-LET radiations.

As chromosomal aberrations can be seen only in mitosis they do not really reflect the original damage because they are the results of interactions at DNA level and possibly of repair and msirepair. Cells in the G_1, S, and G_2 phases of the cell's cycle may be stimulated into mitosis by fusion with mitotic cells. This technique of "premature chromosome condensation (PCC)" offers the possibility to study the formation of chromosomal aberrations in more detail.

Radiation exposure leads to genomic instability in many cell lines. New types of chromosomal alterations may appear many cell generations after exposure and persist in the progeny. It is speculated that this phenomenon may be an important step in radiation carcinogenesis.

A survey of recent developments in the field of chromosomal changes is given by Obe and Vijayalaxmi (2007).

23.5.4
Membranes

Membranes not only act as interfaces between the cell and its exterior or between different compartments, but also contain a number of enzymes whose functions depend critically on their integrity. The most sensitive components are the lipids. Radiation-induced radicals react readily with oxygen to form peroxides with high yields. The most interesting feature of this process is that the yield increases with lower dose rates, which is contrary to most other radiation reactions. Membrane damage does not seem to play a role in reproductive cell death but may be relevant in "interphase death," which is found, for example, in lymphocytes and spermatogonia.

23.5.5
Effects of Heat Treatment

Exposure to IR as well as to RF and MWs increases the temperature of the treated specimen. A biological effect may be caused by heat damage to critical constituents of the system. This possibility has to be carefully excluded if a claim is made for "specific actions." Hyperthermia in the sense of controlled heating has already made its way into cancer therapy. There is no doubt that cells are killed at elevated temperature (see Section 23.6.3), but the search for the target molecule or structure has been elusive. Proteins appear to be one of the prime candidates since they can be denatured, that is, change their molecular structure, at temperatures around 45 °C, with an accompanying loss in enzyme activity. There are, however, large differences, and the rate of inactivation depends not only on the particular protein but also on the presence of other molecules in the immediate vicinity, which may exert protective or sensitizing influences. If ordinary water is replaced by deuterium oxide (D_2O), cells and enzymes become more heat resistant, while they are sensitized in the presence of SH compounds (i.e. chemicals containing a free sulfhydric group), in contrast to the case for ionizing radiation where the same chemicals act as protectors.

Nucleic acids are comparatively heat resistant. DNA undergoes "melting," that is, unfolding of the double-stranded structure, in the temperature region 85–95 °C, depending on the relative proportion of guanine–cytosine pairs. It has been shown that chromosomal aberrations are formed in mammalian cells if they are heated to 45 °C during the DNA-synthesis phase, but this does not prove a direct action on DNA and suggests rather an interference with the synthesis process.

Membranes may play a more important role. They undergo phase transitions at certain temperatures, which can be demonstrated by electron spin resonance (ESR) of incorporated spin labels or by Raman spectroscopy. Also, the activity of membrane-bound proteins is changed after moderate heat treatment. These effects may be translated into impairment of cellular functions, considering the eminent role these structures play not only for transport processes but also for energy metabolism and the general integrity of the complex interrelationships in the cell's life. In spite of a great number of investigations, there is yet no consensus about "critical targets;" details of the ongoing discussion are found in Hahn (1982).

23.5.6
Specific Molecular Aspects of Radio Frequency and Microwave Actions

Experimental situations to address this question have to be designed in such a way that heat effects are excluded. This is not easy on a macroscopic scale and is virtually impossible at the microscopic or molecular level. Molecular vibrations or rotations may lead to local temperature differences that in turn cause structural alterations.

Thermal effects are well understood, current interest thus concentrates on "athermal biological effects." There is presently no accepted mechanism on how they may be caused, and the experimental evidence is quite ambiguous. Most studies deal with genotoxic actions, that is, effects on DNA. Recent reviews (Vijayalaxmi and Obe 2004; SSK, 2006) come to the conclusion that there is no convincing evidence for genotoxic effects of RF or MW fields.

23.6
Cellular Effects

All biological systems are composed of cells. The human body contains about 10^{14} of these elementary units. Most radiation actions on the organism as a whole can be traced back to cellular effects. This holds true for UV and ionizing radiations, but not exclusively for IR, RFs, and MWs. Nevertheless, a thorough discussion of cellular radiation effects is needed to pave the way for understanding tissue and organ malfunction.

23.6.1
Experimental Endpoints of Cellular Radiation Action

23.6.1.1 Survival

Survival in the present context is defined as the ability of a cell to divide indefinitely. It is experimentally determined by the "colony-formation assay," where single cells are plated on suitable nutrient media and incubated at the appropriate temperature. Cells are scored as "survivors" if they grow and form visible colonies. This method, which has been used for a long time with microorganisms like bacteria, yeast, and algae, has also been adapted for mammalian cells in the 1950s and is now a standard technique in many laboratories. In most cases so-called established cell lines are employed, derived from a variety of animal and human tissues. Nearly all of them have undergone transformation, that is, they have the properties of tumor cells, even if they were originally taken from healthy organs. There are only a few exceptions to this rule.

Mammalian cells in culture do not all grow to colonies, even in the absence of damaging agents. The "plating efficiency" is normally quite high (between 70 and 90%) for established cell lines but may be as low as only a few percent for freshly excised tissue.

Survival in the sense of colony-forming ability (cfa) as defined earlier is usually plotted on a logarithmic scale versus radiation dose (or any other physical parameter related to it) on a linear scale. Figure 23.6 gives typical examples of commonly obtained shapes. In curve A, survival decreases with increasing dose in an exponential fashion, which is quantitatively characterized by the dose D_0 reducing cfa to $1/e$. In the majority of cases "shouldered" survival curves are found (types B or C). If the terminal part is again exponential (curve B), the "shoulder width" may be given either by the intercept of its back-extrapolation with the ordinate, which is then called

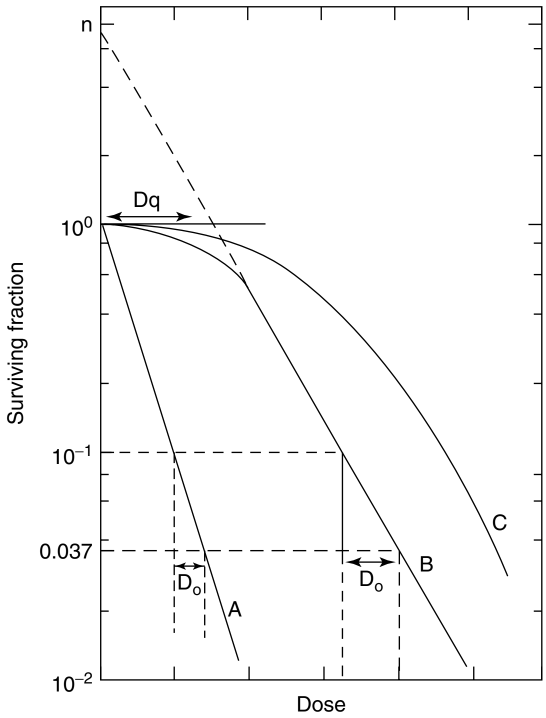

Fig. 23.6 Typical survival curve shapes.

the extrapolation number n, or by the quasithreshold dose D_q, which is obtained from the intercept with the 100% line as shown. Continuously bending curves (type C) are often found especially with mammalian cells. They may be satisfactorily approximated – at least in the initial part – by the expression:

$$y = \exp(-\alpha D + \beta D^2) \qquad (16)$$

commonly referred to as the *linear-quadratic* or "α/β relationship"; y stands for survival, D for dose, and α and β are the system-characterizing parameters.

For low-LET ionizing radiations it has been found for all cellular systems that the sensitivity as given, for example, by the D_0 values depends within certain margins on the mean DNA content per average chromosome. It is calculated by dividing the total mass of DNA per cell by the number of chromosomes. A useful rule of thumb is

$$D_0/\mathrm{Gy} = 3 \times 10^{-16}/(\text{kg of DNA per average chromosome}) \qquad (17)$$

which neglects, of course, differences due to repair deficiencies and may be assumed to be correct within about half an order of magnitude. No such simple relationship exists for other kinds of radiation. Table 23.2 lists the X-ray and UV sensitivities of a few cell types. The apparently simple shapes of survival curves prompted numerous attempts at theoretical interpretations, but there is yet no general agreement (see, e.g. Kiefer, 1990 for further references).

Tab. 23.2 D_0 values and cell nucleus factors for various systems.

Name	Species	DNA/cell[a] (kg)	Chromosome number	DNA/ chromosome[a] (kg)	D_0(UV)[b] (J m^{-2})	D_0(X-ray)[a] (Gy)
T_1 phage	Virus	8.1×10^{-20}	1	8.1×10^{-20}	41	2600
Escherichia coli B/r	Bacteria	1.9×10^{-17}	1	1.9×10^{-17}	85	30
Bacillus subtilis	Bacteria	7.7×10^{-18}	1	7.7×10^{-18}	50	33
Charomyces cerevisiae	Yeast	2.8×10^{-17}	36	7.8×10^{-19}	1150	150
Chlamydomonas	Alga	1.4×10^{-16}	16	8.8×10^{-18}	2500	24
Hamster cells	Mammal	1.5×10^{-14}	22	6.8×10^{-16}	50	1.4
Hela cells	Human	1.2×10^{-14}	78	1.5×10^{-16}	108	1.4

[a] After Underbrink, Sparrow, and Pond (1968).
[b] After Kiefer and Wienhard (1977).

The present preference of the linear-quadratic formula for mammalian cells is rather based on practicability than on theoretical considerations.

With UV, the effectivity to cause loss of cfa depends on wavelength. The action spectrum for mammalian cells in the far UV range coincides largely with the absorption of DNA. Deviations are seen in the region above 320 nm, which suggests that other chromophores, and possibly sensitized reactions, may also contribute here.

The inactivation by ionizing radiations depends also on the radiation type. The effectiveness of charged particles can be given as the inactivation cross-section σ_i. In the lower LET range, σ_i increases almost linearly with all systems. Around 100 keV/μm a supralinear rise is found (corresponding to relative biological effectiveness values greater than 1; see Section 23.6.2) before a transient plateau region is reached. For very high LETs (attainable only by very heavy ions), the unique relationship breaks down, and each ion has to be represented by a separate curve. Here LET is no longer a useful parameter. This is due to the influence of long-range δ electrons, as discussed in more detail elsewhere (Kiefer, 2008). In most cases – most obviously for mammalian cells – σ_i is smaller than the cross-sectional area of the cell nucleus, assumed to be the critical target. This means that a cell cannot be killed by the traversal of single particle through its nucleus, which is very surprising, taking into account the large amounts of energy that are deposited – far exceeding the D_0 values found with low-LET radiations. This finding has implications also for the production of mutant or transformed cells as discussed subsequently. The reason for this behavior cannot yet be fully explained.

The loss of cfa is commonly caused by mitotic failure, that is, the cells die when they attempt to divide. There is another mode of cell death, termed *apopotosis.* This constitutes a physiological process for cells to be eliminated because of age or loss of function. It may be important for cell killing at low radiation doses.

It is commonly assumed that cells surviving a radiation exposure and that are able to form a colony do not carry any remaining damage, except for mutations. This is not generally the case. If cells from colonies grown from irradiated cells are cultivated further they display very often a decline in cfa, which depends on the first dose. This ''delayed reproductive death'' is presumably related to ''chromosomal instability'' (see Section 23.6.2.5).

23.6.1.2 Mutations

Mutations are changes in genetic information that are stably transmitted from one cell generation to the next. On a cellular level they are expressed either as resistance to certain toxic drugs or biochemical deficiencies, resulting in the obligatory dependence on certain nutrients. At the molecular level a mutation represents a change or a complete loss of the function of a particular protein. Since in most cells the chromosome complement is diploid, that is, there are *two* genetic loci for a given protein, a mutation in only one chromosome is usually masked because the information for the protein is provided by the remaining undamaged one. Special test systems have to be developed to overcome this problem, which in classical genetics is called *recessivity*.

All mutation tests require the mutant cell to grow into a colony to be scorable. This means that cell killing, which occurs simultaneously, has to be adequately taken into account. For this it is usually assumed that the survival probability is the same for the mutant and the unmutated cell. This presumption is nearly impossible to prove but must be made for practical reasons. The ''mutant yield,'' that is, the total number of mutants in a cell population, not corrected for cell killing, invariably approaches very low values with high doses – not because they are not mutagenic but because all mutants are killed before they can be detected. It is dangerous to extrapolate from the action of high doses to that of low exposure levels – a situation well known in radiation protection but not always sufficiently appreciated.

Mutation induction curves may depend on dose linearly or (in most cases) in a linear-quadratic manner. UV is considerably more mutagenic than X-rays. The action spectrum for mutant induction resembles very closely that for cell inactivation and hence also that for the absorption of DNA, which is to be expected. The identification by action spectroscopy of nucleic acids as the carrier of genetic information preceded the biochemical experiments by nearly a decade, but remained almost unnoticed in the scientific community.

With ionizing radiations, mutations may also appear delayed after several cell generations. The mutation types may be quite different from those which are seen soon after exposure, a phenomenon linked to delayed reproductive death and genomic instability (Section 23.6.2.5).

23.6.1.3 Neoplastic Transformation

Cancer induction is generally considered to be one of the most important hazards of radiation. Although it is clearly a systemic disease, some of the initial stages can be studied at the cellular level by *in vitro* techniques. They require the use of untransformed cells, which are not readily available since most of the established laboratory strains are essentially tumor cells. They may be taken from fresh preparations (e.g. hamster embryonic tissue), but there are also a few permanent lines, mainly from mice, that grow like untransformed cells.

Whether they are true representatives of "normal" cells is still under debate.

Untransformed mammalian cells grow attached to the surface of the culture dish in a thin unicellular layer ("monolayer"). After transformation into tumor cells they build up to densely packed aggregates, called *foci,* and can be easily recognized after staining. That they really are tumor cells has been proved by injecting them into isogenic animals that do not react immunologically to them, where they inevitably cause tumors while the "nonfocal" cells do not. This comparatively simple approach allows assessment of the cancerogenic potential of radiation. Figure 23.7 shows the transformation "yield," that is, the number of transformants per cell at risk after X-ray or neutron exposure. After an initial rise it declines again with higher doses, which is due to the killing of potential transformants. The situation is quite similar to mutation induction. This graph demonstrates impressively the problems with which one is confronted if risk estimates for low doses have to be derived from measurements at high doses, as is frequently the case in practical radiation protection. In *in vitro* systems, cell killing can be separately assessed and corrected for by plotting "transformants per survivor," but this is hardly possible in tissues and organs.

The probability of transformation after a given dose increases considerably if "tumor promoters" are added to the culture medium. This is taken as evidence that tumor formation is a multistep process that may be modified by special conditions after radiation exposure.

23.6.2
Modifications of Cellular Radiation Effects

23.6.2.1 The Temporal Pattern of Exposure

As pointed out in the general introduction, the temporal evolution of biological radiation effects may span more than 30 orders of magnitude from the initial physical event until its final clinical expression. It is obvious that the time during which a dose is applied must play an important role. Generally, it is found that cell killing with sparsely ionizing radiation is reduced if a dose is split

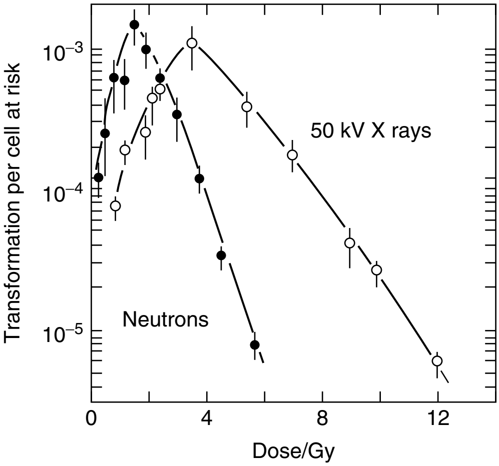

Fig. 23.7 Yield of transformed cells (that is, number of transformants per cell at risk) after X-ray or neutron irradiation (after Han and Elkind, 1979).

into two or more fractions (“split-dose effect”) with an interval comparable to the generation time of the cell type under investigation. A similar result is obtained if the dose rate is lowered and thereby the exposure is spread over a longer time period. In actively growing cells there is an additional influence of the cell growth phases having different radiation sensitivities, so that a rather complex pattern evolves, which, however, does not invalidate the general statement. With mutation and transformation as experimental endpoints, the influence of dose rate has not been studied to the same extent. There appears to be agreement that there is a reduction similar to cell killing if the dose rate is lowered, but there are also reports in the literature claiming an actual increase if growing cell populations are exposed for extended periods of weeks or months. There are only very few dose–rate studies with nonionizing radiations so that a general statement is not possible, although it appears that also in this case reduced killing is found with lower dose rates. Dose–rate effects are absent with densely ionizing radiations. In some instances, even a higher effectiveness of long-term exposures has been claimed, but this requires further confirmation.

The influence of very high dose rates has so far not received much attention. They may be important with laser sources where multiphoton excitations may occur.

23.6.2.2 Chemical Protectors and Sensitizers

Optical Radiations Photosensitized reactions play an important role for the damage of cells by near UV and visible radiation. In most cases, the excitation of the chromophore is transferred to oxygen to form singlet oxygen, which is the damaging species. This mechanism is called *photodynamic action.* The point of attack appears to be mostly membranes, but other targets cannot be excluded. Important examples of photosensitizers include riboflavin, acridine orange, acriflavin, bengal red, and thiopyramine.

Two substances have also recently gained practical importance for the treatment of certain diseases; psoralen and hematoporphyrine. The first interacts with DNA and on light absorption causes interstrand crosslinks in a two-step process, thus leading to cell killing. It is widely used for the treatment of psoriasis and (less frequently) vitiligo. In practice, the derivative 8-methoxypsoralen (8-MOP) is mostly used. The effective wavelength region is in the UV-A around 365 nm. Hematoporphyrin has been claimed to be a potent anticancer agent since it binds with some selectivity to tumor cells. It is used in a preparation, not very well defined chemically, called *hematoporphyrine derivative* (HpD), which has two absorption bands around 400 and 600 nm. Excitation leads to the formation of singlet oxygen, which causes cell killing mainly via membrane damage.

Ionizing Radiation The damaging action of sparsely ionizing radiations to cells can be reduced by the presence of chemical radioprotectors during exposure. There are essentially two classes: radical scavengers and the so-called SH substances. In the first case, the action is via the suppression of the indirect effect (see Section 23.5); typical representatives are alcohols like ethanol, butanol, and glycerol but also dimethylsulfoxide (DMSO). They all react readily with OH radicals to form less aggressive products. SH compounds exert their protective action via the donation of a hydrogen atom to a site whence it was abstracted by radiation, either by direct absorption or by primary radicals. Cells contain already naturally protective substances of which glutathion is the most important. The practical relevance of radioprotectors is small since appreciable effects require rather high, already toxic concentrations. They are considered as possible adjuvants in radiation therapy for the sparing of healthy tissues.

The most potent radiosensitizer is molecular oxygen. If cells are exposed in its absence to X or γ-rays, the doses necessary to achieve a given effect have to be increased by a factor between 2 and 3. The general shape of the dose–effect curves is not altered; oxygen acts in a "dose-modifying" way, that is, the curves are superimposable by application of an appropriate dose-scaling factor. This is commonly called the *oxygen enhancement ratio* (OER), defined as

$$\mathrm{OER} = \frac{\text{dose in the absence of } O_2}{\text{dose in the presence of } O_2} \quad (18)$$

for equal effect

The radiobiological oxygen effect is considered to play an important role in radiotherapy since many tumors contain regions with low oxygen contents. To overcome this problem, agents that act as sensitizers similarly to oxygen without being metabolized so that they may be evenly distributed in the tumors have been developed. The most important examples are metronidazole (also known as *flagyl*) and misonidazole. Their sensitizing action is less than that of O_2 and their applicability *in vivo* is limited by severe side reactions.

The oxygen effect is due to a reaction with transient radiation-induced intermediates in DNA by which they are stabilized and no longer accessible to H donation by intrinsic SH protectors. This is clearly a radiochemical process and has nothing to do with the physiological role of oxygen in cellular respiration. Support for this notion comes from the time dependence. In order to sensitize, oxygen has to reach the sensitive site within less than a millisecond, which is much too fast for any physiological intervention.

23.6.2.3 **Repair Processes**

There are quite a number of processes by which radiation-induced damage to genetic material can be repaired. The molecular mechanisms are quite well known – at least in bacteria, but more and more also in mammalian cells – but space does not permit description of them in detail. *Photoreactivation* is specific for the *in situ* monomerization of pyrimidine dimers and hence only relevant for UV-induced damage. The protein photophospholyase binds to the altered site in a dark reaction. Upon light absorption (effective wavelength region 350–500 nm) the dimer is split and the protein released. Photoreactivation has been demonstrated in bacteria, lower eukaryotes, birds, and amphibians but not without doubt in higher mammals. *Excision repair* is a multistep process in which the damage is first excised from one strand of the DNA, and the resulting gap is filled by repair–replication where the opposite undamaged strand is used as a template. It is operative in all cell types including humans. Excision repair was shown to work preferentially in active genes via a coupling with transcription. Apart from this, "general" excision is also operative. *Postreplication repair* operates if damaged DNA enters replication. The lesions stop the formation of new strands and the resulting gaps are filled either by recombination from the old complementary strand (bacteria) or by translesion synthesis (mammalian cells). In this way it is ensured that at least the newly formed strands are free of lesions although the lesions are not removed from the old ones. In bacteria the process is error-free, that is, it does not lead to mutations, while it is thought to be error prone in mammalian cells since the translesion synthesis may cause the insertion of wrong, that is, noncomplementary, nucleotides.

If the lesion density is too high, repair–replication would lead to DSBs because of overlapping gaps. In this case translesion synthesis occurs also in bacteria, which is normally prevented by a "proofreading" activity ensuring that only correctly matching nucleotides are inserted. Under special circumstances this safeguarding mechanism is "switched off" to allow DNA replication to proceed, although at the expense of possible errors. This process has been termed *SOS repair*. It is only induced if daughter strand gaps are present that cannot be filled by recombination, and is repressed again after completion. Obviously SOS repair is error prone and thought to be the main mechanism for mutation induction in bacteria.

SSBs do not present a problem for the cell under normal circumstances since they are formed also during ordinary replication and can be sealed by special enzymes called ligases. If they are produced by ionizing radiation the chemical configuration may be different so that processing by nucleases may be necessary before final ligation.

DSBs can also be repaired. They were formerly assumed to be lethal but it is known now that they are also the origin of mutations that are formed through misrepair. In mammalian cells, DSBs are handled by two repair systems, namely "nonhomologous endjoining (NHEJ)" and "homologous recombination (HR)." The first one is prone to misrepair while the second is essentially error free.

Repair processes are genetically determined. Many of the genes responsible have been identified, often cloned, both in bacteria and lower eukaryotes like yeast as well as in mammalian cells. Hereditary diseases causing repair deficiencies are also known in man; most of them are inherited recessively so that their prevalence in the population is quite small. Nevertheless, they pose a severe problem for the individuals affected, leading to death at an early age in many cases. This fact demonstrates the importance of repair processes for the integrity of life. UNSCEAR (2000) has discussed the issue of repair in some detail; an older, but still very useful, intoduction was provided by Friedberg, Walker, and Siede (1995).

23.6.2.4 **Radiation Quality**

The effectiveness of ionizing radiations in causing biological damage depends on the spatial pattern of energy deposition at a microscopic or submicroscopic level. This means that dose is not the only determinant to describe a certain effect level in a given system. Examples have been given before for the various endpoints. To account for this dependence on radiation quality, the term *relative biological effectiveness* (RBE) has been introduced. It is defined as

$$\mathrm{RBE} = \frac{D(\text{reference radiation})}{D(\text{test radiation}),\ \text{isoeffect}} \quad (19)$$

that is, it gives the ratio of doses that cause a given effect with a test radiation compared with a reference source. For the latter, ^{60}Co γ-rays are normally used but sometimes 250 kV radiation is also used. Even for a given system, RBE is not constant but depends on the effect level. The reason for this is that the dose–effect curves do not have the same shape with different radiation qualities. One finds a shouldered curve for X- or γ-rays but an exponential type for densely ionizing radiations. RBE obviously increases for low doses and may approach very high values. This makes the determination of the initial part of the curve crucial. A constant limiting RBE would be found if there was an initial exponential component. Since in radiation protection one is usually concerned with low doses, the problem is of great practical relevance. RBE increases with LET up to a maximum around 200 keV μm^{-1}, from where it declines. The reason for this fall is presumably the existence of saturation processes for the formation of primary lesions with very high ionization densities. All cellular endpoints have similar RBE/LET dependences although they differ quantitatively with different parameters, systems, and effect levels. A typical example for mutation induction is shown in Figure 23.8.

23.6.2.5 "Nonclassical" Effects in Cellular Radiobiology: Bystander Effect, Genomic Instability, Low Dose Sensitivity, and Adaptive Response

Classical radiation biology is governed by the "hit hypothesis" (which led to the "target theory") where it is assumed that cellular radiation effects are caused by photon or particle energy deposition in the cell nucleus. Present technical achievements made it possible to target cell nuclei by single particles to check the hypothesis. The experiments led to unexpected results: cells can also be inactivated or mutated if they did not receive a hit but if only neighbors were exposed.

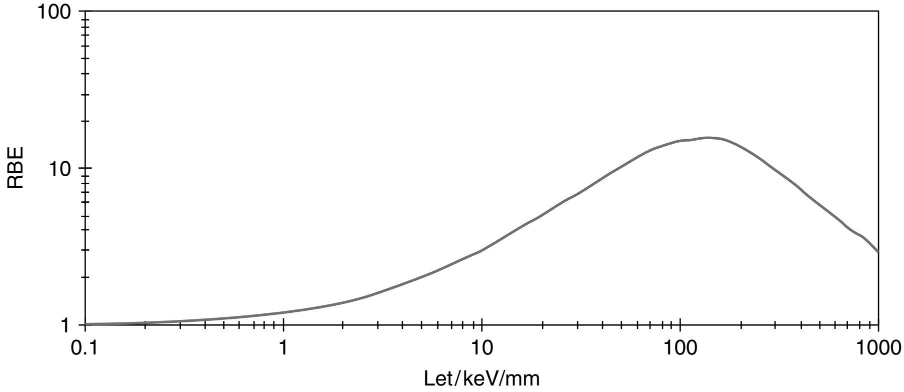

Fig. 23.8 Relative biological effectiveness for mutation induction in mammalian cells as a function of LET (after Kiefer, 2002).

This phenomenon has been termed *bystander effect*. It plays a role particularly at low doses and is not yet fully understood. The current assumption is that hit cells secrete "factors" that are taken up by other cells. It has been shown that the "bystander effect" is also related to genomic instability and delayed reproductive death.

The response of cells to low doses of ionizing radiation has received increased attention also in another respect. It is now clear that gene activity and its regulation can be influenced by X-rays as well as by UV. Studies using single-cell observations suggest that mammalian cells may be more sensitive to killing by low doses of X-rays than anticipated from an extrapolation from the overall survival curve. At the same time, it has been reported by several groups that the yield of chromosomal aberration after high doses is reduced if the cell received "priming doses" in the order of a few hundred milligrays several hours before. This "adaptive response" (which is not found in all cells) depends on an active cellular metabolism and may be related to inducible repair processes, but there is presently no proof for this supposition.

Whether the phenomena described here play a role in the evaluation of radiation hazard is a matter of intense discussion as mirrored for instance in a chapter "Nontargeted and delayed effects of exposure to ionizing radiations" in UNSCEAR (2009).

23.6.3
Heat Effects at the Cellular Level

It is commonly known that cells can be killed by exposing them to high temperatures. This forms the basis for heat sterilization. There are, however, drastic differences between species and strains; thermophilic prokaryotes like *Sulfolobus*, able to sustain growth even at 97 °C, may be mentioned as an extreme example. The more recent interest in thermal effects on cells comes from the application in the therapy of tumors. This section will, therefore, concentrate on mammalian cells only.

Thermal damage shows survival curve shapes similar to those found with UV or ionizing radiation and split-dose recovery similar to X-rays, but there is an additional feature not found with ionizing radiation or UV called *thermotolerance*. This is a transient phenomenon (and hence not due to the selection of genetically heat-resistant substrains) that can be induced by holding cells continuously at elevated but sublethal temperatures (around 42.5 °C) or by a fractionated treatment for shorter times at higher temperatures (up to 47 °C, depending on duration). It has been shown that even after 4-hours cells are more resistant than expected just from the recovery from sublethal damage (see Hahn, 1982). Thermotolerance is most clearly expressed in about 12–48 hours and declines after 72 hours, although not completely. This behavior is of great fundamental and practical importance, particularly in tumor thermotherapy, since in clinical practice treatments are always given in fractionated schedules. The mechanism is not known. It seems that two steps – induction and development – are required, and it has been shown that the second one requires protein synthesis. Heat treatment induces the synthesis of very specific stress proteins ("hsp" for heat stress proteins) but it has not yet been demonstrated unequivocally that they are involved in the expression of thermotolerance. Membranous structures like the cytoskeleton may also participate.

Heat sensitivity depends on a number of factors, which, in most cases cannot be easily separated, for example, cell cycle state, pH, oxygen tension, or glucose concentration. Although it has often been claimed, there is no evidence that tumor cells are *generally* more heat sensitive than their normal counterparts, but it may well be the case under specific conditions. The cytotoxic action of other agents – and ionizing radiations have to be especially mentioned here – is increased in combination with treatment at higher temperatures.

It does not seem that heat induces mutations directly but it may modify synthesis or repair processes in such a way that mutants are produced. This has been found in a number of instances, but more data are required. The issue is even more controversial with regard to neoplastic transformations: some researchers have found an increase, others a decrease in transformation rates with heat treatment. Obviously the particular cell strain and the special experimental protocol are critically important.

23.6.4
Specific Cellular Effects of Radio Frequency and Microwaves

Because of the rising use of mobile communication the influence of MWs has been studied extensively in recent years. In spite of this, no clear picture emerges. Most papers deal with cytogenetic parameters that were already discussed in Section 23.5.6. As pointed out, genotoxic actions with intensities below current International Commission on Nonionizing Radiation Protection (ICNIRP) recommendations cannot be unequivocally demonstrated. Mutation induction was only rarely investigated and not found. There are only very few experiments on neoplastic transformation, which also yielded mostly negative results (Meltz, 2003). Interestingly, cell inactivation was only very rarely included in the study protocol. The present state may be summarized by stating that the great majority of authors were not able to demonstrate cellular effect of MW and RF radiations below currently recommended limits in a reproducible manner.

23.7
Organ and Whole-Body Effects

It has become customary – essentially based on experience with ionizing radiation – to classify organismic damage into two groups, namely, *acute* and *late* effects. Although it is not always applicable in a clear-cut manner, this concept proves to be useful for other types of radiations also and will be followed here.

Acute damage may start, depending on dose, very shortly after exposure and is terminated within weeks or months. In most cases, its degree can be functionally related to dose, and it does not develop below certain threshold doses that depend, of course, on the particular endpoint and may be sometimes quite low. For this kind of behavior the term *nonstochastic* has been coined.

Late effects may take from months to years, or even generations in the case of genetic hazards, to become evident, and severity cannot be defined in a quantitative way. Only the *probability* of their occurrence is a function of dose. They are, therefore, summarized under the heading "stochastic damage." It is commonly assumed that there are no threshold doses although there is no general agreement about this at least for cancer induction, and this supposition is not inherent to the concept. An important difference between nonstochastic and stochastic damage is that the former is, within quantitative variations, predictable in an exposed individuum while the latter occurs in a random manner and is only governed by statistical rules.

For UV and ionizing radiation acute effects can be related to the inactivation of cells in critical tissues, but for other types of electromagnetic radiation the reaction is not so simple. Late effects are linked to the cellular parameters, mutation (for the genetic risk) and transformation (for cancer induction).

Tissue and whole-body effects have been most extensively studied with ionizing radiations and UV. In these cases links to cellular and molecular alterations can be shown so that a framework of explanation may be established. With other radiation types the situation is less clear. To facilitate presentation, the next sections describe in more detail what is known with ionizing radiations and UV while IR, RF, and MW are treated in separate subsections at the end.

23.7.1
Acute Effects

The impairment of cell division is the reason for most acute radiation effects. It is clear from this that the organs expected to be particularly failure-prone are those that depend critically on functioning of cell replacement. These are all surfaces (external and internal) and also the blood-forming system. They constitute "renewal systems" that are all functionally structured as follows: mature cells responsible for proper organ function have only a limited life span in the body, which varies widely. Cell loss is compensated by the entry of newly formed cells from the proliferating "stem cell" compartment into the process of differentiation during which the ability to divide is gradually lost and the specific function attained. This process requires time so that acute cell loss cannot immediately be compensated. The proliferative fraction is the main target for radiation action, which explains why even "acute" effects develop with a certain time lag.

23.7.1.1 **Skin**

The skin is the entry organ for all kinds of external radiation and hence is particularly at risk. The sensitive part is the *stratum germinativum*, consisting of proliferating epithelial cells. In man it lies at a depth of about 150–200 μm and is shielded by other epidermal layers whose absorption properties determine the amount of optical radiation reaching the proliferative compartment. Figure 23.9 gives the relevant information. As a function of dose the following stages of skin damage may be recorded: erythema (reddening, "sunburn"), inflammation, ulceration, and necrosis. Apart from this, skin cancer may develop later (see following text). The minimal erythema dose depends on wavelength and the time of recording. The action spectrum differs also

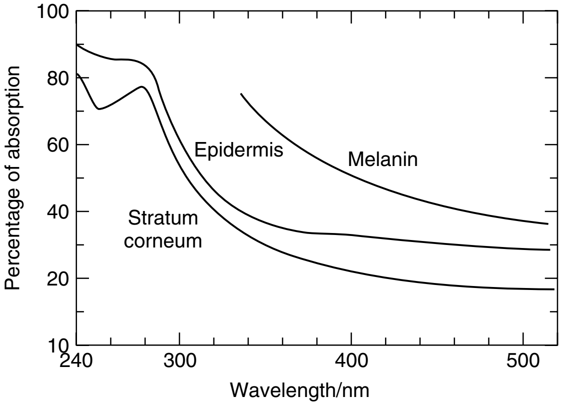

Fig. 23.9 Optical absorption of skin constituents in the UV and visible region (after Parrish *et al.*, 1978).

for different degrees of damage with an increased efficiency around 320 nm for causing more severe effects.

The skin also represents an important immune barrier of the body. Animal experiments show that this protective property is reduced after UV exposure because of damage to Langerhans cells, which comprise about 3% of the epithelial population and play an important role in the immune system. There may also be additional hormonal factors. The wavelength dependence of this effect has not yet been established, and it is not clear whether immune suppression occurs in man.

The skin is also very sensitive to ionizing radiations. It can even be reached by less penetrating kinds like β or even α particles. Effects and their severity depend on dose and the affected area. An abridged summary (after UNSCEAR, 1988) is as follows:

- Erythema develops in 50% of cases after an X-ray dose of 5.7 Gy to a field of $10 \times 10\ \text{cm}^2$.
- Epilation and abnormal hair growth are seen after about 1.5 Gy with 50% probability and may be permanent after more than 7 Gy.
- Desquamation occurs with more than 12 Gy.
- Effects on sebaceous glands also have to be considered. Reductions in size can be recorded after 3 Gy, and severe atrophy is found after 15 Gy.

23.7.1.2 **Eye**

UV effects on parts of the eye depend on the transmission properties, which are depicted in Figure 23.10, and the fluences applied. With lower levels, the main symptoms are photokeratitis and conjunctivitis, which appear a few hours after exposure and last usually not longer than about five days without leaving permanent damage. The threshold fluence for photokeratitis is lowest at 270 nm ($50\ \text{J/m}^2$). Lens damage in the form of cataracts (opacity) may be produced by UV-B and by extended exposures to UV-A. At 300 nm the threshold for transient opacities was found to be about $1500\ \text{J/m}^2$. It is not clear at present whether retinal damage may be caused in humans by excessive fluences in the UV

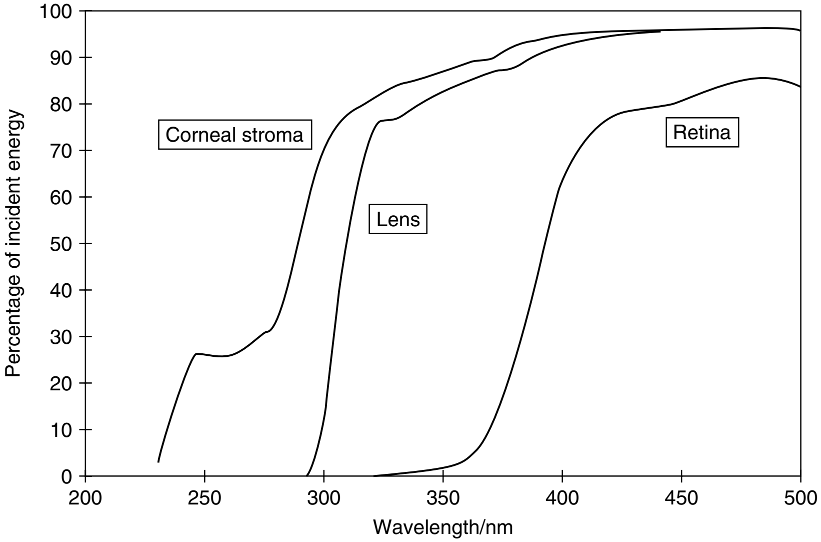

Fig. 23.10 Percentage of energy incident on the cornea that impinges on the anterior surface of various ocular media (after Kinsey, 1948).

region. Acute effects by ionizing radiation include (threshold dose in parentheses) (from UNSCEAR, 1988) erythema of the sensitive conjunctiva (2 Gy), hyperemia and hyperpigmentation of the lid (4–6 Gy), keratitis (4–10 Gy), and epilation of the eyebrows and necrosis (>10 Gy). Apart from these, cataract formation has to be taken into account, which is not clearly an acute effect since it may develop, depending on exposure level, a long time after irradiation. The doses required may be quite low, and rather large RBE values for neutrons have been reported. The eye therefore must be considered as a critical target organ even under circumstances of "normal" working conditions.

23.7.1.3 Nervous System and Other Organs

Direct effects on the nervous system have been found only with very high doses of ionizing radiation and may then lead to very early death as part of the radiation syndrome. Epithelial parts of other organs are also sensitive to the action of penetrating ionizing radiation. Injuries in the mucosa of the mouth and throat cause inflammation, swelling, and necrosis. The salivary glands react to 6–10 Gy within 48 hours. Pneumonitis of the lung can be recorded after one to three months after about 8 Gy.

23.7.1.4 Radiation Syndromes and Lethality

The human body reacts to radiation in a systemic way by developing a number of interrelated symptoms that are commonly called *radiation syndrome.* Its time course and expression depend on the dose level. With more than 50 Gy (always whole-body exposure) there is a direct effect on the central nervous system (which is not related to cell killing) leading to death within 48 hours after periods of lethargy and coma. In the dose range of 10–15 Gy the small intestine is most affected; death will occur with a probability of 90–100% within about two weeks, during which the persons suffer from severe diarrhea, fever, and electrolyte imbalance. With lower dose levels the critical organ is the blood-forming system, mainly the bone marrow. The symptoms for a midline dose range of 2–3.5 Gy are shown in Figure 23.11. It is obvious from this that the radiation syndrome shows a rather complex pattern and that a number of organs are affected.

Damage to the blood-forming system shows up in a reduction of the cellular constituents in the peripheral blood. This is most pronounced for the white blood cells, and less so with erythrocytes because of their comparatively long lifetime (110 days in man). The lymphocyte count is reduced to about 50% after 1–2 Gy within 48 hours. This causes severe impairment of the immune response and favors infections that, in many cases, constitute the ultimate cause of radiation-induced death. Thrombocytes are also reduced in number, thus increasing the danger of hemorrhages and internal bleeding. Even if the bone marrow is the critical organ with intermediate doses, this must not be taken to mean that damage to other organs does not contribute. Apart from the special effects described earlier, the alimentary tract, especially the small intestine, has to be mentioned. Even a "mild" syndrome is always accompanied by diarrhea, nausea, and gut malfunction.

If the doses are not too high the stem-cell system is able to recover, and there will be no remaining acute damage. There is (fortunately) considerable uncertainty about the dose–response curve for radiation-induced death in humans. The mean lethal dose observed within two months, abbreviated $LD_{50,60}$, depends on the health status, age, and amount of medical aftercare. According to recent experience, $LD_{50,60}$ ranges from 2.5 Gy (without medical care) to about 5 Gy and may be sometimes as high as 9 Gy if successful marrow transplantation can be performed. All these values apply to low-LET radiations. Very little is known in this respect for densely ionizing radiation, even for neutrons.

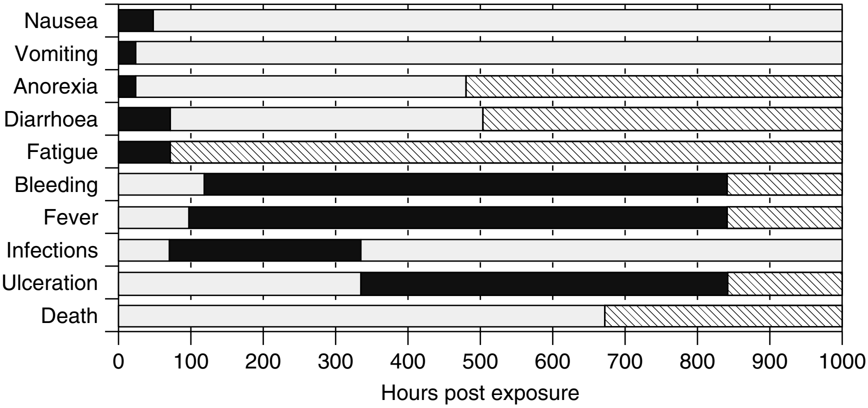

Fig. 23.11 Symptoms of radiation syndrome after whole-body exposure to 2–3.5 Gy (after UNSCEAR, 1988). Filled bars: 50–100% of all cases, hatched bars: 5–50% of all cases.

23.7.2
Radiation and Progeny

23.7.2.1 Fertility

Germ cell production in males represents a good example of a renewal system. The stem cells – type A spermatogonia – are particularly sensitive to ionizing radiation and may not only be killed at mitosis but also suffer interphase death. Moderate reduction in fertility, which is, however, fully reversible, may be brought about by doses as low as 0.15 Gy. Prolonged sterility is caused by more than 6 Gy.

In women the situation is different since the mature germ cells are derived from follicles that are not replenished via a renewal system. There are about 2 million of them at birth and about 8300 left at an age of 45 years. During the reproduction period about 400 mature oocytes are formed; the remainder degenerate by a process called *atresia*. Doses up to about 4 Gy cause temporary sterility, higher ones permanent sterility. Even after lower doses fertility is not fully recoverable.

23.7.2.2 Teratogenic Effects

The embryo, being a vigorously proliferating tissue, constitutes a very sensitive target for radiation action, and teratogenic effects are a matter of grave concern. Embryogenesis in humans may be roughly subdivided into three phases. Before the fertilized egg is nested in the ovary (preimplantation phase), radiation leads to failure of nidation, absorption of the fertilized egg, and hence premature termination of pregnancy. Animal experiments suggest that doses between 1 and 2 Gy are sufficient. There are no data for humans.

Most effects on the growing embryo are characterized by threshold doses (which may be quite low) but there is also the possibility of tumor induction (Table 23.3). Radiation action in embryogenesis displays a distinct temporal pattern (Figure 23.12). Organogenesis, especially its earlier parts (second to sixth week of gestation), is the most critical period in terms of possible teratogenic effects of radiation exposure. The most sensitive organs are those derived early in development from very few determined cells. The central nervous system, eye, skeleton, and spine belong to this category. Severe malformations were found in children exposed *in utero* by the atomic bombs in Japan 1945.

At later stages of development (weeks 8–16) irradiation may cause functional

Tab. 23.3 Threshold doses for radiation effects on embryonic development (modified after ICRP, 2003).

Effect	Critical period (weeks of gestation)	Estimated threshold dose to uterus (mGy)
Intrauterine death	0–2	1000
Malformations	2–6	50–100
Mental retardation	8–16	300
Tumor induction	Complete period	40–60 (doubling dose, no threshold)

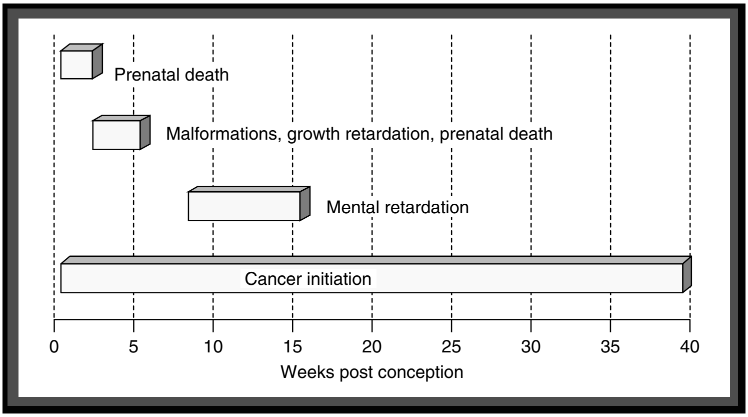

Fig. 23.12 Prenatal radiation effects.

impairment of the brain, leading to mental retardation. Threshold doses are higher in this case.

Tumors may be initiated at any time during embryonic and fetal development, leading to the expression of cancer in early childhood. The evidence is less clear here. No statistically significant increases of childhood cancers after exposure *in utero* were recorded in Hiroshima and Nagasaki but an extensive study of children whose mothers underwent X-ray diagnosis for obstetric reasons while pregnant showed increased numbers of childhood cancers (mainly but not exclusively leukemia). This Oxford Survey of Childhood Cancer (Bithell and Stewart, 1975) suffers from uncertain dosimetry but gives sufficient evidence for an increased rate of childhood cancers. There are, of course, no dose thresholds; the doubling dose may be estimated to be around 50 mGy to the uterus but with large uncertainties (Wakeford and Little, 2003).

The later stages in embryo development (fetal period) appear to be comparatively resistant to the action of ionizing radiation in contrast to the action of many teratogenic chemicals.

23.7.2.3 **Genetic Effects**

Germ cell mutations are transmitted to the progeny of exposed parents and hence present a serious potential hazard to future generations. To define it in quantitative terms, however, is very difficult for various reasons. Most mutations are recessive, that is, they are only phenotypically expressed if present on both homologous chromosomes. The probability for this increases, of course, with their proportion in the total gene pool. On the other hand, not all mutations will persist indefinitely in the population since they may be eliminated because of reduced fitness of the affected individuals, so that eventually equilibrium may be reached. Our present knowledge about genetically caused diseases is still rudimentary in spite of considerable progress in human genetics. A statistically increased rate of genetically caused diseases due to the exposure of the parents before conception has so far not been shown in any population. Quantitative data on genetic effects of radiation in man are thus not available so that one must refer to animal experiments. They have been performed extensively, mainly with mice, involving many millions of animals, providing a large database.

The now accepted approach to quantify the results is the "doubling dose method." It is based on the determination of the dose that increases the rate of spontaneous mutations by a factor of 2 (hence the name). UNSCEAR (2001) considered all available data extensively and derived an estimate of the doubling dose based on the frequencies of radiation-induced mutations in mice and spontaneous mutations in man. Its value is slightly less than 1 Gy for low-dose rate low-LET radiations like X- or γ-rays; for risk estimation it is rounded to 1 Gy.

Table 23.4 summarizes the results obtained for chronic X- or γ-irradiation.

There is no easy direct relationship between the (genetic) risk for a disease and the frequency of induced mutations. The "mutational component," that is, the relative increase in disease frequency (relative to the baseline incidence) per unit relative increase in mutation rate (relative to the spontaneous rate), depends both on the type of mutations and the disease. This complex problem is extensively discussed in UNSCEAR (2001).

Multifactorial genetic diseases form the great majority. They were not always appropriately taken into account in past analyses. They are observed as congenital abnormalities at birth (about 6% spontaneous rate) but are also present in about 65% of the population and only rarely directly observable. With the advent of more sophisticated methods in human genetics it became clear that many diseases like, for example, cancer, diabetes, or coronary heart disease have genetic traits. The mutational component is considered to be low so that the expected risk is of the same order of magnitude as for autosomal and X-linked diseases in spite of the high frequency in the population.

Careful examinations of the bomb survivors in Hiroshima and Nagasaki using a number of approaches did not reveal any significant increase in germ

Tab. 23.4 Risk estimates for radiation-induced genetically caused diseases based on a doubling dose of 1 Gy (chronic low dose rate low-LET radiations) (UNSCEAR, 2001).

Disease class	Baseline frequency per million livebirths	Risk per gray per million progeny	
		First generation	Second generation
Mendelian			
Autosomal dominant and X linked	16 500	750–1500	1300–2500
Autosomal recessive	7500	0	0
Chromosomal	4000	–	–
Multifactorial			
Congenital abnormalities	60 000	2000	2400–3000
Chronic	650 000	250–1200	250–1200
Total	**738 000**	**3000–4700**	**3930–6700**

cell mutations compared with unexposed control groups. This must not be taken to mean that the genetic risk to humans is negligible, since the groups involved were not large enough. The estimates based on the methods described above are fully in line with this conclusion.

23.7.3
Radiation Cancerogenesis

23.7.3.1 Ultraviolet Radiation

With UV radiation the only tissue at risk in terms of cancer formation is the skin. The majority of tumors are of the nonmelanoma type and comparatively easy to cure if treated at any early stage. Indirect information on humans may be obtained by comparing skin tumor incidences for different geographical latitudes where the UV-B component of sunlight varies. Such a compilation is depicted in Figure 23.13 for nonmelanoma skin cancers, demonstrating clearly the increased cancerogenic potency of short-wavelength UV.

All types of skin cancer show a worldwide dramatic increase. This is not only the case for the comparatively harmless basal and squamous cell carcinoma but also for the very aggressive melanoma. Figure 23.14 summarizes data obtained in British Columbia illustrating this alarming development. It is certainly not due to the reduction in stratospheric ozone but presumably to changed leisure habits and possibly also the increased use of sunbeds and other tanning devices.

Ozone depletion will add to this situation. Evidence is mounting that this is not a theoretical possibility only. It can now be seen worldwide, although the most dramatic effect has been registered in Antarctica. Future trends are difficult to predict and depend essentially on the use of chlorofluorocarbon compounds. It has to be kept in mind, however, that atmospheric chemistry operates at long timescales. A comprehensive review of all problems involved can be found in SSK, 1996.

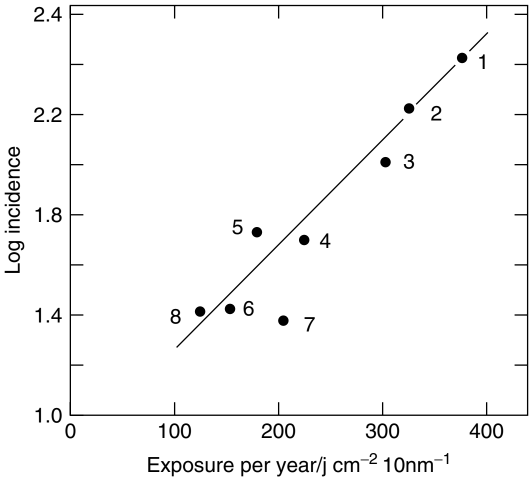

Fig. 23.13 Skin cancer as a function of UV-B exposure in different countries (after Gordon and Silverstein, 1980): 1, Australia; 2, Texas; 3, South Africa; 4, Nevada; 5, Canada; 6, England; 7, Germany; 8, Scotland.

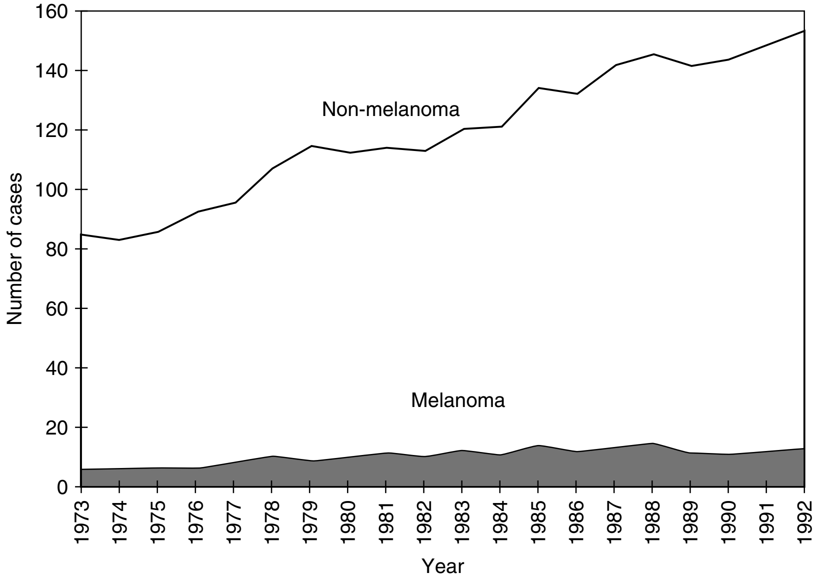

Fig. 23.14 Increase in nonmelanoma and melanoma skin cancer in British Columbia from 1973 to 1992 (Philips *et al.*, 1996).

23.7.3.2 **Ionizing Radiation**

With ionizing radiations no organ is spared from possible tumor formation although the probabilities vary. So do the latency times, that is, the period that elapses between exposure and the clinical manifestation. They span a wide range, which is considerably larger than the mean lifetime of common experimental animals like mice or rats. This means that for a quantitative assessment one can only have recourse to retrospective human data. They are available from a number of groups that received radiation either occupationally, accidentally, or as part of medical treatment. UNSCEAR continuously reviews all available studies and provides actual updates (the last one in UNSCEAR, 2008).

The most important source is still the follow-up of survivors of Hiroshima and Nagasaki, commonly called the *Life Span Study* (LSS). They constitute a cohort of about 100 000 people for which complete medical records are available. The determination of dose presents a problem. Approximate values were initially calculated in 1965 from the then-accessible bomb data and the evaluation of the exact location of each single victim relative to the epicenter. The doses thus obtained were termed *T65D* ("tentative doses 1965"). The calculations gave a higher neutron component for Hiroshima, and the comparison of leukemia cases in both cities provided thus an estimate for neutron RBE, which formed the basis for the recommendation of quality factors in radiation protection (see Section 23.9.2). A reevaluation led to a revision of the original estimates with the main results that the doses were slightly (by about a factor of 1.5) lower than previously thought and, more seriously, that the previously postulated difference in neutron contribution is only very small and insufficient

Tab. 23.5 Recent major epidemiological studies on cancer risk by ionizing radiations.

	Time period	No. of people	Person years	Reference
LSS incidence solid tumors	1958–1998	105 427	2 764 730	Preston *et al.*, 2007
LSS mortality solid tumors	1950–1997	86 572	3 062 046	Preston *et al.*, 2003
LSS incidence leukemia	1950–1987	93 696	2 778 000	Preston *et al.*, 1994
Workers solid tumors and leukemia	variable (1943–1997)	407 391	5 192 710	Cardis *et al.*, 2005

for a reliable RBE determination. This system is referred to as *DS*86. Recently another revised system has been established (DS 02) (Young and Kerr, 2005) which had, however, only minor influence on risk estimates (Table 23.5).

There are two ways to analyze existing data, namely, in terms of *absolute* and *relative* risk. In the first case it is assumed that radiation induces cancers on top of those occurring spontaneously; in the latter it is postulated that radiation increases the already prevailing risk by a dose-dependent factor, which means that the tumor types expected to have the highest numbers after radiation are those that already have a high prevalence in the population. Tumor rates (TR) may then be formally expressed as

$$TR = TR_{sp} + g_a(D) \quad \text{(absolute risk model)} \qquad (20)$$

or

$$TR = TR_{sp}[1 + g_r(D)] \quad \text{(relative risk model)} \qquad (21)$$

where TR_{sp} represents the spontaneous tumor rate, which depends, of course, on age, sex, tumor type, and other factors, and g_a and g_r are the absolute and relative risk factors, respectively. They are also functions of age, sex, and so on, but particularly of dose. The kind of relationship to be applied is still under debate, especially whether a linear or a linear-quadratic dose dependence is more appropriate. It is not possible to arrive at unambiguous decisions for the time being, so that linearity is commonly assumed as a prudent method of hazard estimation.

The time course of tumor appearance differs between leukemia and solid tumors. While leukemia has only comparatively short latency periods of 8–15 years, solid tumors manifest themselves only after much longer periods of time and tend to accumulate later in life. Figure 23.15 displays yearly rates of radiation-induced tumors for different periods of observation. The total number of deaths for the entire time 1950–1990 amounted to 334 caused by solid tumors and 87 due to leukemia (out of a population of 86 000).

Radiation-induced solid tumors seem to follow roughly the age behavior of spontaneous tumors so that the model of relative risk is applicable if allowance is made for the fact that there is some reduction in excess relative risk with

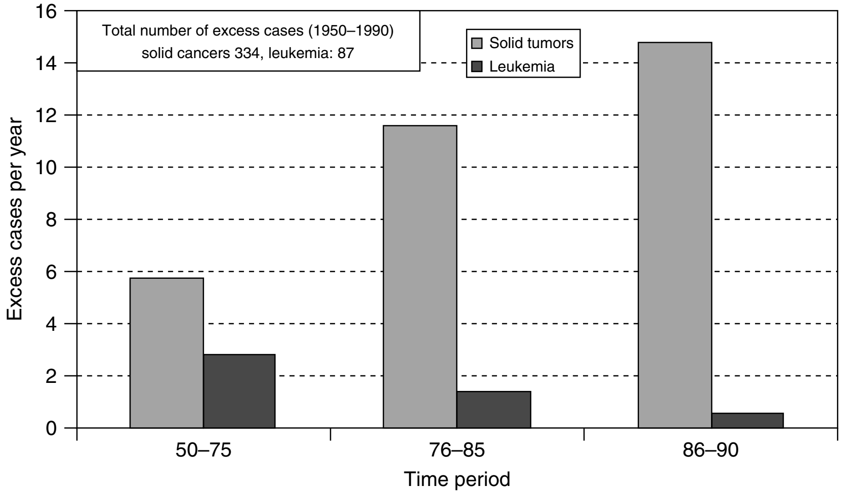

Fig. 23.15 Excess tumor cases per year for different time periods after exposure (after data from Pierce *et al.*, 1996).

increasing attained age. Leukemia, on the other hand, shows up comparatively early and does not depend very much on attained age. The rate of new cases decreases with time and it may be assumed that now, more than 50 years after exposure, no new cases are to be expected. "Excess relative risk" is thus not a really appropriate indicator for leukemia.

The long-term follow-up of the Japanese bomb survivors has seen not only revisions in dose determinations but also reappraisals of the medical findings. Because of the long latency periods, malignancies other than leukemia can only be properly assessed after many decades.

These data are now available, which led to a substantial increase in risk estimates. A matter of special concern is the finding that children exposed at young age are considerably more susceptible to radiation-induced tumors than previously assumed. This conclusion could not be reached before because most cancers develop only late in life. These stages are now seen in those people who were very young at the time of bombing.

Figure 23.16a gives incidence "excess relative risks" (relative risk – 1) for a number of cancer sites. Figure 23.16b shows the respective mortality estimates.

More recently, tumor mortality has also been studied in people occupationally exposed to radiation. In principle, it could be expected to obtain more reliable data as doses are continuously monitored and the health status is usually also regularly checked. The doses incurred, however, are rather small so that the statistical power is expected to be limited. This is borne out by a recent worldwide pooled study (Cardis *et al.*, 2005). Nevertheless, it was possible to obtain risk estimates for solid tumors and leukemia. They are slightly higher than those from the LSS in Hiroshima and Nagasaki but well within the error margins (Figure 23.17). It is noteworthy that the workers are exposed to radiation continuously at lower dose rates. There is no indication that the effect

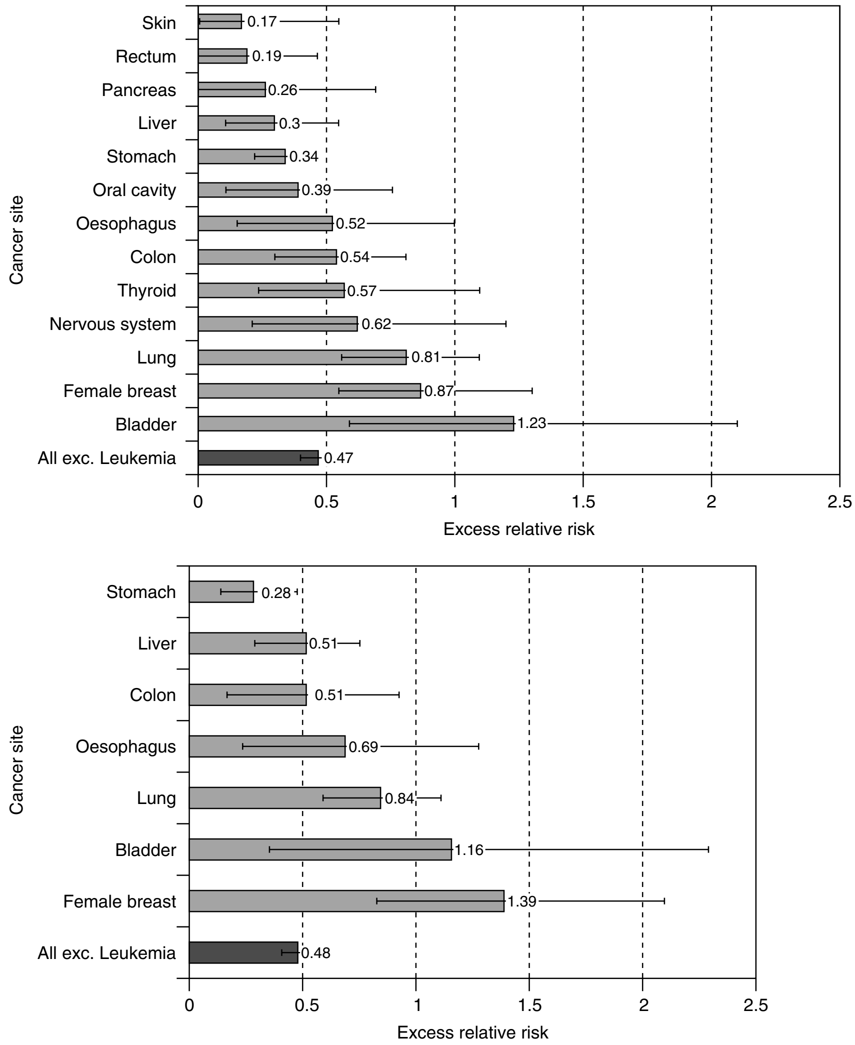

Fig. 23.16 (a) Excess relative incidence risks (RR) estimates for solid cancer at different sites (Preston *et al.*, 2007). The error bars indicate 95% confidence intervals (b) as in Figure 23.16a but for mortality (Preston *et al.*, 2003).

level is reduced because of this exposure condition.

The American BEIR Committee (BEIR, 2005) analyzed all available data to estimate the lifetime risk. Figure 23.18 summarizes some of their results. It gives the percentage risk for cancer incidence and mortality of solid cancers and leukemia after an exposure to 0.1 Gy. There is a marked difference between males and females, the latter being clearly more sensitive. The total risks for incidence after a dose of 0.1 Gy are 0.9% for men and 1.4% for women. The respective figures for fatal cancers are 0.5 and 0.7%.

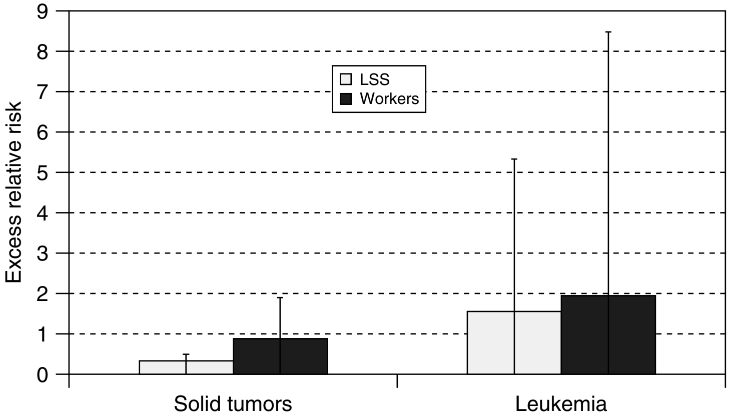

Fig. 23.17 Comparison of mortality excess relative risks from epidemiological studies in occupationally exposed people (‘‘workers’’) and the Life Span Study (LSS) of survivors in Hiroshima and Nagasaki (Cardis *et al.*, 2005). The Japanese data were reanalyzed by Cardis *et al.*, (2005).

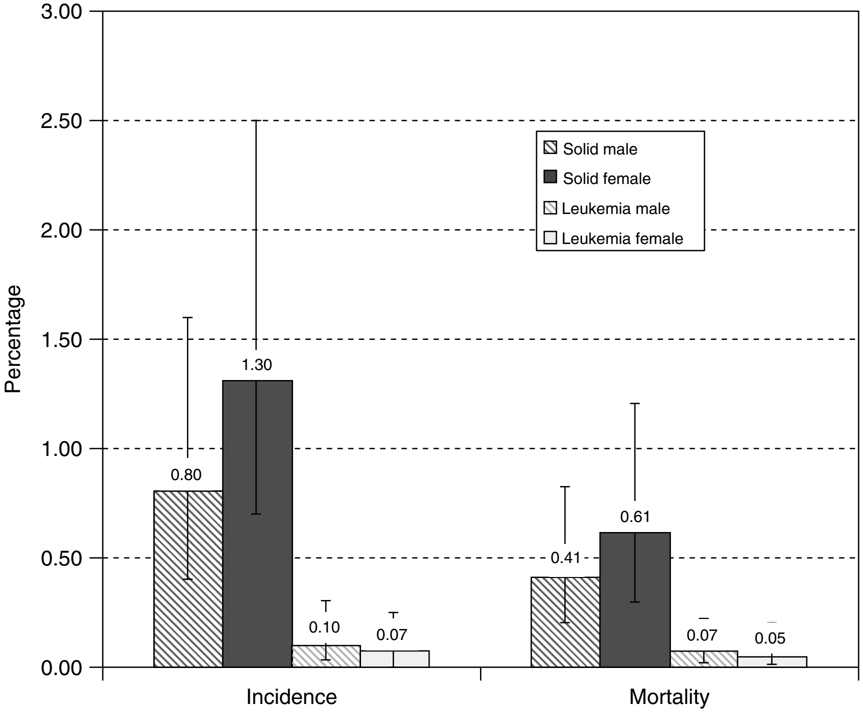

Fig. 23.18 Lifetime risks after an exposure to 0.1 Gy of ionizing radiation (after data from BEIR, 2005).

Another important source of radiation-induced tumor initiation in humans is the epidemiology of uranium mine workers. They are exposed to α particles of ^{222}Ra and its daughter products, which cause lung cancer. There is quite a number of studies on this topic, which were summarized and analyzed in the BEIR VI report (BEIR, 1999). Radon plays a significant role in the normal environment also (see Section 23.9.1.2, Environmental Radiation Impact).

23.7.4
Tissue and Whole-Body Effects of Infrared Radiation

Because of the weak penetration ability of IR, the organs most at risk are the eye and the skin. Penetration depth depends on wavelength: IR-A penetrates deeper (to several mm) than IR-B or IR-C.

The most important eye damage is cataract formation, which has often been reported as a result of long-term exposure to comparatively weak IR sources. With the widespread use of lasers this hazard has to receive special attention. The most effective wavelength region appears to be IR-A. Although there appears to be common agreement that a threshold exists for cataract formation, it has so far not yet been clearly defined. The exposure limits currently recommended (Section 23.9.2.2) are assumed to be on the "safe" side. With higher exposures retinal burns may also be found.

The effects on the skin are due to local heating and depend on the penetration depth, which is a function of wavelength. There may be considerable differences with skin type and pigmentation. With low exposures, erythema-like reddening is seen; higher levels may lead to edema, blisters, and burns.

As a delayed effect "erythema *ab igne*" may be seen. It appears at skin areas that were repeatedly exposed to higher temperatures not sufficient to cause burning (ICNIRP, 2006). In very rare cases, carcinomas may also be an aftermath of this condition.

23.7.5
Tissue and Whole-Body Effects of Radio Frequency and Microwaves

Possible or suspected carcinogenic actions of RF and MW radiation have recently gained considerable attention in public discussion because of the widespread use of mobile communication. Organ and whole-body effects may be due to two phenomena: heat production and electrical or magnetic interferences. Although there is a wealth of literature reports, no clear picture has yet emerged, much less a coherent theoretical interpretation. There is no sound evidence for carcinogenic, mutagenic, or teratogenic action. Impairment of fertility can be caused by temperature increase in the testis since this organ is known to be particularly heat sensitive. The same might apply to the embryo but there is no proof so far that, for example, diathermia may lead to malformations. Local overheating can also lead to cataracts in the eye but the SAR values necessary are far beyond present protection limits (Elder, 2003).

A large epidemiological study on cancer and mobile communication is presently under way in a number of European countries ("INTERPHONE study"). As a common design is used, it is planned to pool all results to provide a large database. Parts have already been published (Berg *et al.*, 2006). No associations between brain tumors and mobile phone use could be found. A negative finding was also reported for acoustic neurinoma in Japan (Takebayashi *et al.*, 2006).

Induced electrical currents can interfere with nerve action. Changes in electroencephalograms have been reported but the techniques employed are subject to criticism. Very high fields can cause respiratory tetanus, that is, the inability to breathe due to muscle contraction, and even to uncoordinated asynchronous heart contractions. It is conceivable that lower levels may change behavioral performance; this is indicated by some animal studies, but studies with human subjects are inconsistent.

A recent critical review may be found in Valberg, van Deventer, and Repacholi (2007).

It should be noted that special conditions prevail for electronic equipment on or in the body, like hearing aids or cardiac pacemakers.

23.8
Applied Radiation Biology

There are quite a number of radiation techniques in the life sciences; only a few can be briefly reviewed here. Diagnostic radiology and the use of radioisotopes as tracers are completely omitted because radiation *action* is not the essential determinant in these cases. The range of applications extends from molecular biology to medical therapy and food preservation; examples will be given. Only basic principles can be discussed; detailed prescriptions would be out of scope.

23.8.1
Molecular Techniques

23.8.1.1 Action Spectroscopy

UV and visible radiation can only exert a biological action if it is absorbed by a suitable chromophore. The dependence of a given biological effect on wavelength can, at least in principle, be used to identify the responsible component by means of its absorption spectrum. This is the basic idea of "action spectroscopy." Despite its impressive simplicity it is not free from pitfalls and requires careful consideration if erroneous conclusions are to be avoided. There are a number of necessary conditions which are not always easy to fulfill:

1. The fluence dependence of the effect should be similar for all wavelengths, that is, the response curves should be superimposable if the fluence is appropriately scaled by a constant factor for each wavelength. If this is not the case, the reaction is most likely caused by different mechanisms.
2. The quantum yield Q should be the same at all wavelengths.
3. The absorption spectrum of the chromophore in the system must be known, or it must be determined that it does not differ from that in its pure form.
4. Absorption and scattering of substances not taking part in the primary reaction must be known if they cannot be neglected.
5. The chromophore absorption should not be too high, to avoid saturation effects.
6. The effect should be independent of fluence rate since this often varies widely with different wavelengths ("reciprocity").

Response curves may have very different shapes, sometimes well defined in mathematical terms but mostly not. To exclude any possible bias, action spectroscopy should always be based on the *same* effect, and the isoeffect fluences should be compared.

If one is dealing with the same mechanism over the whole range (condition 1) and if the quantum yield Q is constant (condition 2), it can be safely assumed that the same effect is caused by the same number of primary alterations. Their number depends on the quantum yield and the number of quanta absorbed; thus, when comparing the same effect at two different wavelengths,

$$Q\sigma_{\text{abs}}(\lambda_1)\phi(\lambda_1) = Q\sigma_{\text{abs}}(\lambda_2)\phi(\lambda_2) \tag{22}$$

where σ_{abs} is the absorption cross-section and ϕ the quantum fluence. Since $Q =$ constant, one obtains

$$\phi(\lambda_1)/\phi(\lambda_2) = \sigma_{\text{abs}}(\lambda_2)/\sigma_{\text{abs}}(\lambda_1) \tag{23}$$

The actual task is to determine the quantum fluence necessary to produce a given effect as a function of wavelength. Since σ_{abs} is proportional to the extinction coefficient $\in$, one may also write

$$\phi(\lambda_1)/\phi(\lambda_2) = \varepsilon(\lambda_2)/\varepsilon(\lambda_1) \tag{24}$$

If the measured quantity is not the quantum fluence but rather the energy fluence F_E, the relation becomes

$$F_E(\lambda_1)\lambda_1/F_E(\lambda_2)\lambda_2 = \varepsilon(\lambda_2)/\varepsilon(\lambda_1) \tag{25}$$

In practical action spectroscopy, $1/F(\lambda_1)\lambda_1$ is usually normalized to a reference wavelength and plotted versus wavelength. If the above conditions are fulfilled, the resultant curve should be equal to the chromophore's extinction spectrum.

23.8.1.2 Molecular Weight Determination by "Radiation Inactivation"

Nearly all techniques for the determination of molecular weights of biological macromolecules require the separation of the particular component. Apart from the experimental effort involved they cannot be used in those cases where, for example, an enzyme loses its specific activity if it is removed from its supporting matrix. This is very often the case with membrane-bound proteins. Their molecular weight may be estimated from the inactivation doses upon γ- or X-ray exposure. In most cases an exponential relationship is found between enzyme activity and dose:

$$y = e^{-kD} \tag{26}$$

where y is the remaining activity, D the dose, and k the "inactivation coefficient." If one assumes that the indirect effect (see Section 23.5) can be neglected and that a fixed energy absorbed per molecule suffices for inactivation, one may also write

$$y = e^{-mD/\varepsilon} \tag{27}$$

which forms the basis for the application. Here m is the molecular mass and $\in$ the "critical energy," which has been found to be 66 eV for γ-rays with energies >1 MeV. Over about two orders of magnitude there is a good proportionality between analytical (obtained by conventional techniques) and radiological molecular weight. The upper limit of the applicability lies around 10^8 g/mol (Kempner, 1988; Beauregard *et al.*, 1987).

23.8.2
Radiation Disinfection and Sterilization

Microbial contaminations are a major source of concern in food and drinking water production and in other branches of industry. In many cases they are fought by the application of chemicals which, however, may also pose a health hazard. Although radiation could serve as a convenient and generally less hazardous alternative, its acceptance – at least as far as human consumption is concerned – meets with considerable reservations in public, which are very often based on misconceptions. Sterilization of technical products, on the other hand, is already standard practice.

The discussion in this section will be restricted to UV and ionizing radiation. IR and MWs can also be used under certain circumstances but their action is exclusively through heat production, and a treatment of the problems of heat sterilization is clearly out of scope of this chapter.

UV is in principle a very convenient and comparatively inexpensive tool to kill contaminating microorganisms. Its applicability is severely limited, however, by its low penetration power. It can therefore only be applied for surface decontamination and with fluids transparent to the wavelength range applied. Although pure water does not absorb in the region between 200 and 300 nm, in most cases dissolved or suspended constituents prohibit the use of UV for the processing of drinking water. By constructing appropriate exposure devices where the effective penetration depths can be kept short, this problem could eventually be circumvented. The cost factor has to be taken into account, but it must be weighed against the possible benefits, which lie in the avoidance of toxic and unpleasant chemicals like chlorine.

The most effective wavelengths as deduced from the action spectrum of cell inactivation are around 260 nm, making the low-pressure mercury lamp (which emits essentially only 254 nm) the source of choice. Its applicability is, however, limited by the comparatively low irradiances available so that high-pressure mercury lamps also have to be taken into consideration. Since they emit a much broader spectrum, photochemical reactions with chemical contaminants must be considered.

The use of ionizing radiation is currently heavily debated, particularly with regard to food irradiation. It is feared that it may cause unforeseen hazards to the general population, but the arguments are rarely based on solid scientific facts. Some of the points raised are as follows.

1. *Nutritive value may be lost.*
   It is true that certain vitamins like A, E, thiamin, pyridoxine, and ascorbic acid are to a certain extent destroyed by irradiation. Fatty acids undergo changes that may influence flavor and odor. Foods with high water content may change their texture after irradiation. These statements are certainly true, but they must be judged against the effects of other ways of food processing, such as heat sterilization or the addition of chemical preservatives.

2. *Radiation may produce toxic products.*
   Doses necessary for a sizable reduction of microbial contamination do cause radiation-chemical processes. Animal experiments on a

very large scale have shown, however, that harmful effects are not to be expected. The toxicity involved is certainly less than that of "acceptable" chemical preservatives.

3. *Irradiated food may become radioactive.* This is the argument that causes the most widespread fears, but it is nevertheless wrong. Since nuclear reactions that may lead to induced radioactivity have clear energy thresholds, they can be easily avoided by imposing upper limits, which are 5 MeV for photon and 10 MeV for electron irradiations. Under these conditions there is no induced radioactivity in foodstuff.

There are, however, a few real disadvantages that must not be overlooked:

- The doses normally applied are not sufficient to kill viruses.
- Bacterial toxins produced before sterilization are not destroyed, but this is true also for most other treatments.
- Irradiation may change the microflora in such a way that only the most resistant organisms survive. Their regrowth could lead to "atypical" spoilage, which is not readily recognized, and the food may erroneously be considered safe.

The last three points do not appear to be strong enough to ban food irradiation as an easy and comparatively cheap and safe way to reduce considerable losses by contamination and spoilage. As a consequence a special committee of the European Community recommended food irradiation as an allowed means of preservation (Urbain, 1986). The European Union (EU) publishes yearly reports about food irradiation in its member states, the last one available is for 2005 (European Union, 2007). In the United States, wheat and wheat flour may be irradiated with doses up to 0.5 kGy, pork up to 1 kGy, dry enzyme preparations up to 10 kGy, and dry vegetable seasonings up to 30 kGy. With these doses, reduction but never a complete eradication of the contaminants is achieved. Higher doses may, of course, be applied for the sterilization of technical products that are not for human consumption. Current information is provided by the FDA website (http://www.cfsan.fda.gov/~dms/opairrad.html).

For more detailed discussions, the reader is referred to (Molins 2001; Sommers and Fan 2006).

23.8.3
Therapeutic Applications

23.8.3.1 Nonmalignant Diseases

Ionizing radiations are nowadays mainly used for cancer treatment; but they are also applied to the treatment of nonmalignant diseases, for example, arthritis or macular degeneration (see Meyer, 2002; Seegenschmiedt, Micke, and Willich, 2004). There is, of course, always the risk of cancer induction to be taken into account but as the patients involved are usually of older age this is not considered to outweigh the expected benefits.

IR radiation serves as a convenient method for heating certain parts of the body, but the action is not specific for its radiative nature and hence will not be discussed. Radio waves and MWs are also used to increase the temperature at deeper sites in the body ("diathermy"), which has also found applications in cancer therapy. The beneficial action of electric and magnetic fields has sometimes been claimed, but it still waits for solid experimental evidence and is, therefore, not treated.

UV light is commonly believed to improve health, although there is no basis for such a general statement. The wide distribution of tanning studios reflects a psychological rather than a physiological impact. It is quite clear that their frequent use presents a considerable danger to acquire skin cancer. This is particularly the case for juveniles. ICNIRP has, therefore, published a strong warning against the indiscriminate use of tanning appliances (ICNIRP, 2004a).

There are, however, a few cases where UV treatment is indicated. Rickets, which is caused by vitamin D deficiency, does not play an important role nowadays and is mentioned only for historical reasons.

There are, however, a number of special cases where phototherapy can be successfully applied:

- Hyperbilirubinemia is a condition sometimes occurring in newborns. It is characterized by an increased blood bilirubin level, a decay product of hemoglobin. It may be treated by exposure to blue light, which is thought to stimulate the further decomposition of this metabolic product.
- The UV treatment of skin tuberculosis is today only of historical interest. More important is the treatment of two particular skin diseases: vitiligo and psoriasis. The first one expresses itself in the form of nonpigmented circumscribed areas, while the other is typified by sharply defined red-covered spots caused by a misregulation of epithelial cell divisions. It is estimated to be found in 1–5% of the population. The classical way to treat these diseases was to deliver series of UV-B exposures; nowadays photochemotherapy is preferred. A photosensitizer is given, either by local injection or orally, and the skin is exposed to UV having a wavelength coinciding with the maximum of the sensitizer's action spectrum. Tar preparations were formerly used, but today psoralens are exclusively applied. The best results appear to be obtained with 8-MOP. The most used wavelengths are around 360 nm, that is, in the UV-A region; hence the name PUVA (=psoralen + UV-A) was given to the method. Since PUVA has been shown to cause mutations in *in vitro* systems, a possible carcinogenic action must be taken into account. However, so far there has been no evidence of skin tumors from observations on a large number of patients. Nevertheless, further careful supervision is warranted, particularly in view of the very long latency periods involved.

23.8.3.2 **Cancer Therapy**

The "classical" radiation therapy of tumors uses quantum or electron radiation provided commonly by linear accelerators. The energy is adjusted in such a way that the dose maximum lies at the tumor site. By careful treatment planning using multiple beams and beam shaping by custom-made absorbers, the tumor volume can be delineated and the healthy tissue spared to a great extent. This technique has now reached a high degree of sophistication.

Intensity-modulated radiation therapy (IMRT) is a new approach, which has the potential to revolutionize classical radiation therapy (Cash, 2006). With the help of imaging techniques like computed tomography (CT), positron emission tomography (PET) and magnetic resonance tomography (MRT) a three-dimensional picture of the tumor and its surrounding tissue is constructed forming the basis for treatment planning. The images are coupled to the multileaf collimator at the head of the linear accelerator, which is moved around the patient in such a way that a true conformal dose delivery is possible. The X-ray beam is subdivided into numerous small "beamlets" whose intensities are adjusted in such a way that the tumor and more resistant parts receive higher doses than the surrounding healthy tissue. In this way one can achieve a high degree of sophistication in individual exposure schemes to reach local control of the tumor with (at least theoretically) maximum sparing of nearby critical structures. IMRT appears to be particularly applicable to prostate and head and neck cancers. It is assumed that in this way the advantages of particle therapy may be approached without the need for expensive particle accelerators that are not available in standard clinics.

IMRT requires a correct positioning of the patient during treatment and a close spatial agreement with the treatment plan. Movements of the patient or of internal organs, for example, by breathing, can lead to substantial shifts in the dose distribution so that parts of the tumor may be missed and healthy organs overdosed. This problem may be overcome by image guided radiotherapy (IGRT) (Dawson and Sharpe, 2006) where images of the tissue containing the tumor are recorded immediately before or during treatment and local dose distributions adjusted. IGRT is still in the experimental stage and requires considerable effort of physicians and medical physicists but it offers the potential of remarkable improvements.

More recently, particle therapy has also gained some importance. The most widely used is based on protons. Since they deliver most of their energy just before the end of their range, very good dose discrimination can be achieved. There is typically a dose factor of 5 between the so-called *plateau* and the *peak* regions. As the dose maximum is very sharp, small tumors can be very effectively treated with an optimal sparing of the surrounding areas. This technique has been applied especially to ocular melanomas; several thousand patients have so far been treated with excellent results. Not only could the tumor be locally cured, but also eyesight could be saved to a large extent in most cases. While in the beginning, treatments were performed with accelerators built for research in physics (the forerunner being Boston, Massachusetts, U.S.A.), more recently "dedicated" machines have been established in specialized hospitals.

Protons possess a very favorable dose distribution, but their biological effectiveness is not much different from that of γ- or X-rays as a consequence of their comparably low LET. To deliver an even higher "biological" dose and overcome complications like the "oxygen effect," heavier ions are required. Studies along this line started many years ago in Berkeley, California, USA, but had to be terminated because of the shutdown of the accelerator. Meanwhile two new studies were started, one in

Chiba, Japan, where a dedicated heavy-ion accelerator was built, and the other in Darmstadt, Germany, which made use of the existing heavy-ion accelerator and is still in the preclinical study phase. A new method for treatment planning has here been developed where the beam energy is modulated in such a way that the treatment volume matches the tumor as closely as possible ("raster scan"). The technology developed in Darmstadt has been transferred to a new specialized clinical center in Heidelberg (Germany), which will presumably open in summer 2009 (Schulz-Ertner, Jakel, and Schlegel, 2007).

A recent review lists a total of more than 56 000 patients who have been treated so far by particle therapy, with the majority being treated by protons 49 000. There are presently 26 medical proton facilities worldwide in operation and three using carbon ions. Quite a number is in the planning or building state. Reviews about clinical experiences are provided by Schultz-Erkner, Jakel, and Schlegel (2007), Schultz-Erkner and Tsujii (2007), and Greco and Wolden (2007). The Particle Therapy Co-Operative Group (PTCOG) operates a website where all current developments are listed (http://ptcog.web.psi.ch/).

23.8.4
Biodosimetry

23.8.4.1 Fluorescence *In situ* Hybridization (FISH) for Chromosomal Translocations

Chromosomal aberrations are discussed in Section 23.5.3. Apart from their fundamental relevance they are an important tool of biodosimetry. If lymphocytes taken from exposed individuals are cultivated *in vitro*, aberrations can be scored and used as an indicator of dose. Classically, dicentric chromosomes ("dics") are used for this purpose since they are comparatively easy to recognize. The lowest dose limit resolvable is about 0.1 Gy with X- or γ-rays. Dics interfere with further cellular proliferation; 50% of the cells carrying them will die at every division. This means that these "unstable" aberrations will rapidly fade out so that reliable dose estimates are impossible after months or years. Translocations, on the other hand, are stably transmitted to the progeny, allowing in principle biodosimetry at any time, even after decades. Technical problems of their scoring prevented their use until the so-called fluorescence *in situ* hybridization (FISH) for chromosomal translocations technique was introduced. Studies have shown that translocations persist indeed much longer than dics. Their dose dependence is similar to that of dics, and the detection limit lies in the same range (about 100 mGy for γ-rays). The issue of biodosimetry by chromosomal aberrations is discussed by Sevan'kaev *et al.* (2006) and by several authors in Obe and Vijayalaxmi (2007).

23.8.4.2 ESR Dosimetry

Ionizing radiation induces the formation of free radicals in the exposed tissue, but their lifetime is usually very short. This is not the case if they are created in a solid matrix where their mobility is severely restricted. In this case they may persist for very long periods and may thus be detected by ESR, for example, in teeth or bone. The method has practical applications both in radiation protections and in radiotherapy. It can also be applied to irradiated food (Sevan'kaev *et al.*, 2006; Regulla, 2005).

23.9
Radioecology, Radiation Protection, and Guidelines

Possible hazards by radiation constitute a matter of great public concern. Although it was originally concentrated on "radioactive radiation" (this misnomer is used here only as a quotation, obviously meaning "ionizing radiation") it has recently been extended even into the field of RF and MW radiation. Possible dangers have long been recognized and regulations established to protect those people professionally exposed to radiation. This is at least true for ionizing radiation for which the regulation system is most advanced. Because of this the discussion will concentrate mainly on "radiological protection" since the recommendations for nonionizing radiations are developed to a large extent on the basis of a similar philosophy. Before dealing with the matter, it is appropriate to discuss briefly the impact of radiation sources in the environment.

23.9.1
Radioecology

23.9.1.1 Nonionizing Radiations

Radioecological problems involving nonionizing radiation are restricted to photochemical processes in the atmosphere. Ozone (O_3) plays a central role here. It is photochemically formed by light absorption in the wavelength range around 240 nm. The maximum of its absorption spectrum coincides roughly with that of nucleic acids; compared to O_2 it is shifted toward longer wavelengths. This means that ozone formation is favored in the upper layers of the atmosphere, while its destruction prevails in lower regions since both reactions are part of a complete steady state. A reduction of the ozone belt would lead to a shift of the solar spectrum on the Earth's surface toward shorter wavelengths. It is estimated that a decrease by 10% would double the fraction of solar radiation at 290 nm (Caldwell *et al.*, 2007; Rowland, 2006).

Estimating the biological risk is not easy. There are a number of studies on the effect of increased short-wavelength UV in plant communities and marine ecosystems. The results are inconclusive so far; it appears that there is a high buffering capacity for varying amounts of UV light. The expected increase of skin tumors in humans appears to be the most important risk.

23.9.1.2 Ionizing Radiation

With regard to human radiation hazards, a distinction must be made between external and internal exposure. In the first case only electromagnetic radiations, neutrons, and a few energetic electron sources are of any practical concern. Internal exposure is brought about by the incorporation of radionuclides via ingestion with food or inhalation. Environmental radioactivity can enter the food chain in a number of ways. Most important is the uptake from water and soil by plants and animals or directly in the drinking water. Here all radionuclides have to be taken into account, and α emitters play a particular role since they deposit energy with high LET.

For the assessment of doses, one must consider all the transfer probabilities from the original source to the final deposition in the body. Additionally the dose distribution inside the organs affected, either directly by incorporation or indirectly by exposure from a distance, has to be computed. To facilitate the solution, the path of radioactivity in the environment is broken down into a number of transfer compartments. The whole web can be treated as a system of coupled linear differential equations. The most important parameters are the transfer coefficients, which give the fraction of a given compound transported from one compartment to another. Transfer coefficients depend only on the chemical nature and are, therefore, equal for different isotopes of the same element. They can be used to estimate the expected radioactivity in food or milk if the extent and kind of environmental contamination are quantitatively known.

To assess the impact on human populations, food consumption has to be taken into account. This cannot be done in an exact way since personal lifestyle and regional habits may lead to wide variations. The fate and radiological importance of incorporated radioactivity depends on the chemical nature of the compound as well as on the physical properties. The length of expected exposure is determined by the effective half-life $T_{(1/2)\mathrm{eff}}$, which is composed of the physical decay ($T_{(1/2)D}$) and the biological excretion probability ($T_{(1/2)B}$) according to

$$1/T_{(1/2)\mathrm{eff}} = 1/T_{(1/2)D} + 1/T_{(1/2/)B} \tag{28}$$

Most radionuclides are not distributed equally in the human body but show certain organ specificities. A well-known example is iodine, which is selectively enriched in the thyroid. Organ specificity must be taken into account for hazard evaluation. On the basis of physiological, chemical, and physical data, it is possible to calculate organ and whole-body doses that can be expected after the incorporation of a given amount of a certain radionuclide. A few examples are listed in Table 23.6.

Environmental Radiation Impact

Natural sources There are wide variations in doses received by the general public depending on occupation, terrestrial environment, and even housing. Natural sources are still most important. Cosmic space radiation contributes significantly even at sea level but the doses increase, of course, with altitude. It may be estimated that the dose rate at 11 km (typical cruising altitude of passenger aircraft) is about 6 μS per hour (see Section 23.9.2.1, Basic Principles, Quantities, and Units, for definition of Sv). Environmental radioactivity from long-lived radionuclides surviving from the formation of the earth ("primordial radionuclides") contributes both to exposure from the outside (only γ radiation) or to internal irradiation via ingestion or inhalation.

Table 23.7 lists average values as estimated in a recent UNSCEAR report (2000). Data are given as "effective doses" in mSv. The world average of the total is 2.4 mSv but there may be very wide variations because of different geological conditions. There are areas (e.g. in India or Brazil) where the natural background is one order of magnitude larger than the world mean.

Radon (^{222}Ra) and its immediate daughter products contribute obviously the largest share of environmental radiation hazards. They originate from the underground and building materials and may concentrate in housings. It is well known from uranium miners who are exposed to high concentrations that the

Tab. 23.6 Dose factors for conversion of radioactivity taken up by ingestion in effective whole-body or organ doses. Calculations for adults (>17 years of age) and for new born children (< 1 year of age). Data from ICRP (2002).

Nuclide	Whole body ($Sv\ Bq^{-1}$)		Organ ($Sv\ Bq^{-1}$)		Affected organ
	Adults	Children	Adults	Children	
^{3}H	1.8×10^{-11}	6.4×10^{-11}	–	–	–
^{14}C	5.8×10^{-10}	1.4×10^{-9}	6.2×10^{-10}	2.0×10^{-9}	stomach
^{32}P	2.4×10^{-9}	3.1×10^{-8}	8.2×10^{-9}	1.7×10^{-7}	Bone marrow
^{60}Co	3.4×10^{-9}	5.4×10^{-8}	1.2×10^{-8}	8.7×10^{-8}	Colon
^{90}Sr	2.8×10^{-8}	2.3×10^{-7}	1.8×10^{-7}	1.5×10^{-6}	Bone marrow
^{125}I	1.5×10^{-8}	5.2×10^{-8}	3.0×10^{-7}	1.0×10^{-6}	Thyroid
^{129}I	1.1×10^{-7}	1.8×10^{-7}	2.1×10^{-6}	3.7×10^{-6}	Thyroid
^{131}I	2.2×10^{-8}	1.8×10^{-7}	4.3×10^{-7}	3.7×10^{-6}	Thyroid
^{137}Cs	1.3×10^{-8}	2.1×10^{-8}	–	–	–
^{210}Po	1.2×10^{-6}	2.6×10^{-5}	1.3×10^{-5}	1.8×10^{-4}	Kidney
^{226}Ra	2.8×10^{-7}	4.7×10^{-6}	1.2×10^{-5}	1.6×10^{-4}	Bone lining
^{232}Th	2.3×10^{-7}	4.6×10^{-6}	1.2×10^{-5}	1.3×10^{-4}	Bone lining
^{238}U	4.5×10^{-8}	3.4×10^{-7}	7.1×10^{-7}	6.9×10^{-6}	Bone lining
^{239}Pu	2.5×10^{-7}	4.2×10^{-6}	8.2×10^{-6}	7.4×10^{-5}	Bone lining
^{241}Am	2.0×10^{-7}	3.7×10^{-6}	9.0×10^{-6}	8.3×10^{-5}	Bone lining
^{252}Cf	9.0×10^{-8}	5.0×10^{-6}	2.9×10^{-6}	1.1×10^{-4}	Bone lining

Tab. 23.7 Estimated per capita annual effective dose equivalent from natural sources in areas of normal background (UNSCEAR, 2000).

Source	World average annual effective dose (mSv)	Typical range (mSv)
Cosmic radiation	0.4	0.3–1.0
Terrestrial external	0.5	0.3–0.6
Inhalation (mainly Radon)	1.2	0.2–10
Ingestion	0.3	0.2–0.8
Total	**2.4**	**1–10**

inhalation causes lung cancer (BEIR, 1999). Extensive new epidemiological studies in Europe (Darby, Hill, and European Collaborative Group on Residential Radon and Lung Cancer, 2003) and the United States (Field *et al.*, 2006) show that radon also plays an important role in residential dwellings and may contribute significantly to lung cancer morbidity and mortality in nonsmokers.

Man-made radiation sources Figure 23.19 surveys the contribution of nonnatural radiation sources to the

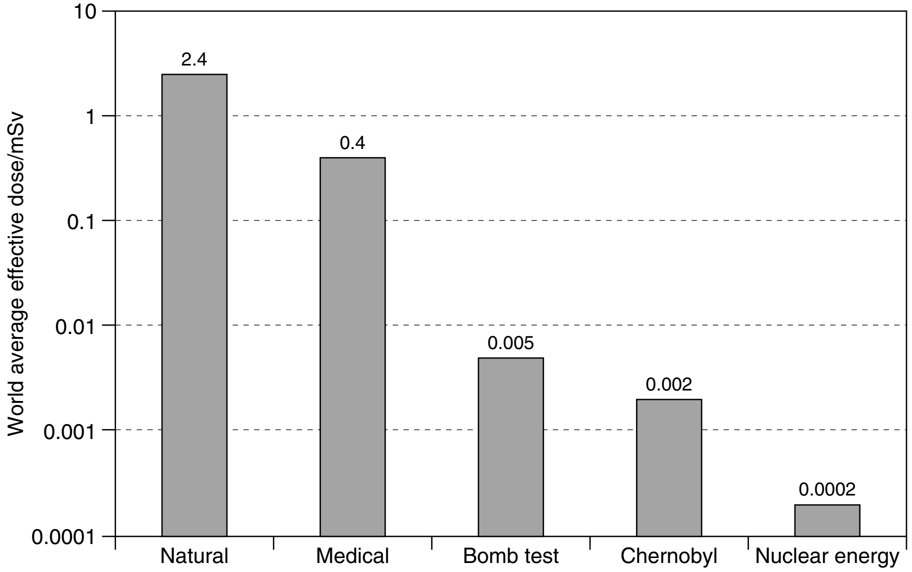

Fig. 23.19 The contribution of man-made radiation to the global radiation budget (data from UNSCEAR, 2000).

exposure of human populations at a global scale. Medical applications (essentially only radiodiagnostics) are most important although there are very wide variations between countries. Japan, Germany, and the United States give values of around 1–2 mSv per year, and there is a tendency for further increases.

Atmospheric testing of nuclear weapons was banned and stopped in 1963, but there are still remains of the radioactivity released at those times. The world average of the contamination by the Chernobyl accident in 1986 is now relatively low but the level depends very much on the distance and is still considerably larger in the countries originally affected (see next section). Nuclear energy production with presently about 450 plants in operation worldwide gives only a very small proportion, but this is, of course, only valid for normal undisturbed working conditions.

Medical exposures constitute an important branch of the health system in developed countries. There is no question of the justification but it must be acknowledged that medical exposures also constitute a certain risk. The doses depend on the particular examination and the technique employed. Computed X-ray tomography provides impressive images but the dose per examination is about 10 times higher than with conventional techniques. Table 23.8 lists a few examples. Doses are given as ''effective doses'' although this concept is not really appropriate but it may serve as a convenient means of comparison. World averages of frequencies of diagnostic procedures are also given, as well as their relative contribution to total population dose. Computed X-ray tomographies are still comparatively rare (0.6%) but they

Tab. 23.8 Effective doses of X-ray diagnostic examinations, their frequencies, and relative contribution to population dose

Examination type	Effect. dose (mSv)[a]	Frequency (world average) (%)[b]	Percentage of population dose[b]
Chest radiography	0.02–0.05	27	3
Chest fluoroscopy	~0.4	11	10
Limbs and joints	0.01–0.1	17	0.8
Lumbar spine	0.8–1.8	5	7
Thoracis spine	0.5–0.8	1	1
Cervical spine	0.5–0.8	3	0.7
Hip and pelvis	0.5–1	3	0.5
Head	0.03–0.1	6	0.5
Abdomen	0.6–1.1	5	2
Gastro-intestinal	10–18	5	17
Urography	2.5–7	2	3
Mammography	0.2–0.6	2	0.9
Angiography	10–20	0.6	6
Interventional		0.3	4
Head CT	2–4	0.6	34
Spinal cord CT	3–10		
Thorax CT	6–10		
Abdomen CT	10–25		
Others		11.5	9.6

[a] From Veit, Griebel, and Bauer (2001).
[b] Taken from UNSCEAR (2000) (covering the time period 1991–1996).

amount to 34% of the population dose due to medical diagnostic applications. It can already be foreseen that their frequency will rise in future, particularly in developed countries where they are often regarded as a convenient means for a general health check. They are, however, not without risk. With a mean effective dose of 8 mSv a single CT examination corresponds to the exposure on about 80 transatlantic flights.

Impact of the Chernobyl accident The accident in the Soviet reactor of Chernobyl on April 26, 1986, created a widespread radioactive contamination affecting more than 100 000 people in Belarus, Ukraine, and parts of the Russian Federation. Because of uncertain dose assessment, firm epidemiological data are not readily available, but a few conclusions can now be drawn:

1. Acute radiation syndrome was diagnosed in 134 recovery workers who were close to the reactor site. Twenty-eight of them died during the first four months. Eleven more deaths were confirmed in the years 1987–1998 but the causes were not always clearly radiation related (UNSCEAR, 2000).
2. Contrary to common perceptions, there is *no* increase in childhood leukemia, both in the Ukraine and in Belarus. There are indications of a slight increase of leukemia in the group of liquidators but the epidemiological evidence is not

compelling because of methodological problems. It cannot be excluded that cancer rates may rise in future years but this statement is based only on estimates of low-dose risks, which are not proven. In any case, the expected increase is presumably too low to be detectable by epidemiological surveys (Cardis *et al.*, 2006)

3. A dramatic increase in thyroid cancers was found in children and young adolescents caused by the uptake of ^{131}I. In the period 1986–2002 there were 2010 cases in Belarus, 2344 in Ukraine, and 483 in the Russian Federation, giving a total of 4837. As the background rate is rather low, one can assume that most of them are radiation related. Fifteen patients are known to have died; most of the others could be successfully treated. Based on previous experience, one must expect that the number of thyroid cases in those individuals exposed as children will continue to rise (Cardis *et al.*, 2006).
4. There is no evidence for radiation-related stillbirths, miscarriages, and malformations in newborns. It has to be acknowledged, however, that the sequelae of the disaster led to widespread psychological consequences in the whole population, which may have serious health impacts (UNSCEAR, 2000).

23.9.2
Principles of Radiation Protection Regulations

Radiation protection as a system of recommended measures has evolved essentially to protect people occupationally exposed to ionizing radiations from possible harmful effects. It was not the original intention to apply it for the general population, but it is now widely used in this context and constitutes an important issue in political discussion. Since radiation can be measured with high precision and sensitivity and because of the detailed knowledge of biological radiation effects, radiation protection may without exaggeration be considered to be the best-founded approach to the assessment of environmental hazards.

23.9.2.1 Ionizing Radiations

Basic Principles, Quantities, and Units For the purpose of radiation protection, a distinction is made between nonstochastic (formerly also named *deterministic*, more recently (ICRP, 2007) renamed as *tissue damage*) and stochastic damages. The first group comprises effects where the severity increases with dose and where a threshold dose exists (however small it may be). Examples are all acute effects but also radiation-induced developmental malformation as a result of *in utero* exposure. Stochastic damage is characterized by the fact that a degree of severity cannot be given and that only the *probability* of the occurrence is a function of dose. It is generally assumed that in this case there are *no* threshold doses although this is not necessarily inherent to the concept. For protection purposes, one assumes a linear dependence between the probability of the effect and dose (“linear-no-threshold (LNT) concept,” LNT hypothesis). Stochastic effects are germline mutations constituting the genetic risk and cancer.

The physical quantity "dose" is not sufficient for hazard evaluation since the extent of damage depends also on radiation quality, as discussed in Section 23.6.2.4. This was taken into account by introducing a new quantity named *equivalent dose* or more recently "radiation weighted dose" H, used exclusively for radiation protection measurement. It is related to dose D in the following manner:

$$H = w_R D \tag{29}$$

where w_R is the radiation weighting factor that is based on relative biological effectiveness for stochastic effects. Table 23.9 summarizes the recommended values. The International Commission of Radiological Protection (ICRP) in its most recent recommendations (ICRP, 2007) made few adjustments that are also given in the table. If D is measured in Gy, the unit of radiation weighted dose is "Sievert" (Sv). The old unit is "rem" (1 rem = 0.01 Sv).

To convert partial body exposures into whole-body equivalents, the quantity "effective dose" has been introduced. It applies only to stochastic effects and is not intended to be used for risk estimation in specific individual exposure conditions. The effective dose is obtained by multiplying the organ or tissue radiation weighted doses by "tissue weighting factors" w_T, which relate to the organ sensitivities with regard to genetic (in the case of gonads) or cancer risk. They are averaged over the whole population (i.e. all age groups); recommended values are listed in Table 23.10. Organs not specifically mentioned are summarized as "remainder tissues." According to the most recent recommendations (ICRP, 2007) they are composed of 14 tissues: adrenals, extrathoracic region, gall bladder, heart, kidneys, lymphatic nodes, muscle, oral mucosa, pancreas, prostate (male), small intestine, spleen, thymus,

Tab. 23.10 Weighting factors w_T for tissues at risk (after ICRP, 1991, 2007).

Tissue or organ	ICRP 60 (1991)	ICRP 103 (2007)
Gonads	0.20	0.08
Bone marrow	0.12	0.12
Colon	0.12	0.12
Lung	0.12	0.12
Stomach	0.12	0.12
Urinary bladder	0.05	0.04
Breast	0.05	0.12
Liver	0.05	0.04
Oesophagus	0.05	0.04
Thyroid	0.05	0.04
Skin	0.01	0.01
Bone surfaces	0.01	0.01
Brain		0.01
Salivary glands		0.01
Remainder	0.05	0.12
Sum	**1.0**	**1.0**

Tab. 23.9 Radiation weighting factors. The most recent ICRP recommendations (ICRP, 2007) give a functional dependence for the neutron weighting factor versus neutron energy.

Radiation type	ICRP 26 (1977)	ICRP 60 (1991)	ICRP 103 (2007)
Photons, electrons	1	1	1
Protons	10	5	2
Neutrons	10	5–20	5–20
α particles, heavier ions	20	20	20

and uterus/cervix. The w_T is used for the average weighted dose of all the remainder organs or tissues. The sum of all weighting factors is unity; they thus give the relative sensitivity with respect to stochastic effects.

Limitation It cannot be overemphasized, particularly for stochastic damage, that there is no "safe" amount of radiation. Consequently, the former usage of the phrase "tolerance doses" has been abandoned. The ICRP states the overriding principle in radiation protection in a rather categorical way by saying that (ICRP, 1977).

1. no practice shall be adopted unless its introduction produces a positive net benefit;
2. all exposures should be kept as low as reasonably achievable, and economic and social factors taken into account; and
3. the dose equivalent to individuals shall not exceed the limits recommended for the appropriate circumstances by the Commission.

The second point is known commonly as the "ALARA principle," which implies that the limits mentioned in Table 23.11 have to be considered as *upper limits* and by no means as "tolerable."

Radiation protection thus rests on three fundamental principles: justification, optimization, and risk limitation.

Since with partial body irradiation, rather high values might result for organs with a low risk factor for stochastic damage, separate limits for nonstochastic damage have also been defined. They are intended to protect from acute damages where threshold doses exist.

Tab. 23.11 Yearly dose limits (ICRP, 1991).

Body, organ or tissue	General public (mSv)	Persons occupationally exposed to radiation (mSv)
Whole-body, effective dose	1	20
Lens of the eye	15	150
Skin, hands, or feet	50	500

Limits are not only given for irradiation from the outside but also for incorporated radioactivity. They are, of course, specific for each particular radionuclide. By use of the conversion factors given in Table 23.8, annual limits of intake (ALI) can be calculated that correspond to the stochastic limit under the assumption that the nuclide in question is the only radiation source. In the case of multiple exposures, the total impact can be calculated according to

$$\sum \frac{w_{Ti} H_{Ti}}{H_{Wb,l}} + \sum \frac{I_j}{ALI_j} \leq 1 \qquad (30)$$

Here H_{Ti} is the dose equivalent of the ith tissue, w_{Ti} is the corresponding weighting factor, $H_{Wb,l}$ is the limit dose equivalent for whole-body exposure, I_j is the yearly intake of the jth radionuclide, and ALI_j is the corresponding annual limit of intake.

The presently recommended limit values are listed in Table 23.11.

23.9.2.2 Nonionizing Radiations

As indicated earlier, the system for radiation protection is most advanced for ionizing radiations. The ICRP is the most influential body in this field in setting out general rules and recommendation for protection measures in this field. There is an analog organization for nonionizing radiations, namely, the ICNIRP that deals – among others – with guidelines for the protection against UV, IR, RF, and MW radiations.

Protection from UV radiation in the wavelength range from 180 to 400 nm (excluding coherent sources like lasers) was treated by ICNIRP in 2004 (ICNIRP, 2004b). Recommended exposure limits are given in terms of "weighted energy fluences" in J/m^2. The weighting function shown in Figure 23.20 follows essentially the action spectrum for erythema induction, normalized to a wavelength of 270 nm. As the unit (J/m^2) is the same as for unweighted radiation, one has to be careful when comparing values from different sources. The eyes and the skin constitute the organs at risk. For the first it is stated that the exposure should not exceed 30 J/m^2 spectrally weighted radiation and also the unweighted exposure should not be larger than 10^4 J/m^2 to avoid damage.

Skin sensitivity depends on the "skin type" with "type I" characterized by a light complexion and subject to sunburn even with little exposure, being the most sensitive. As a guideline it is recommended that the weighted exposure should not exceed 30 J/m^2 to minimize long-term risks such as skin cancer. These limits are intended to be used for the working population and do not apply for therapeutic treatments. It is also clear that extensive sunbathing and/or frequent use of sunbeds leads to overexposure, increasing the risk of skin cancer. The World Health Organization (WHO, 2003) as well as (ICNIRP, 2004a) address this question, which is of increasing concern for public health, in special publications. An extended foundation document on UV protection of workers has been recently published by ICNIRP in cooperation with WHO and

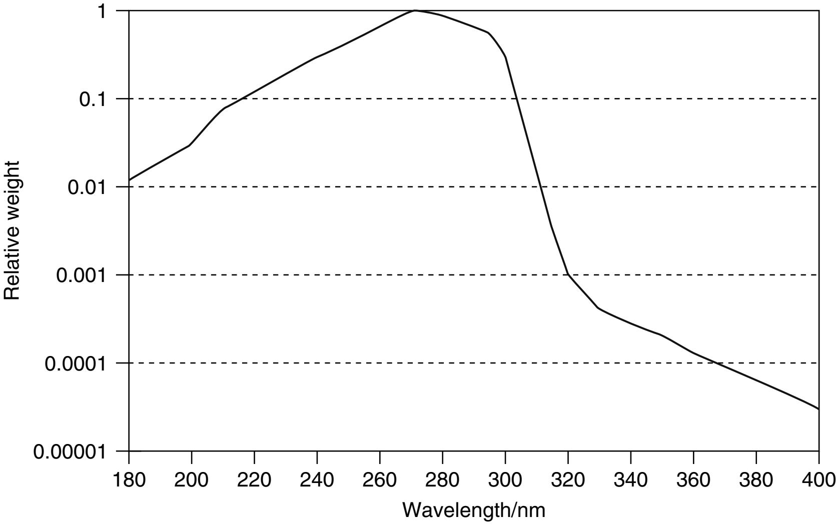

Fig. 23.20 UV weighting function (ICNIRP, 2004b)

the International Labour Organisation (ILO) (Vechia *et al.*, 2007).

The special properties of lasers, namely, the coherence and particularly the high intensities, warrant a more detailed consideration of protection measures, which is outside the scope of this article. The organs most at risk are again the eye and the skin. Protection measures and exposure limits are discussed in ICNIRP (2000).

ICNIRP has also issued a statement about protection from extended IR exposure (ICNIRP, 1997, 2006) that should be consulted for more details.

Exposure limits for RF and MW are based essentially on thermal effects as there is no convincing evidence for biological or health effects below the limits for thermal effects. ICNIRP (1998) proposed the values listed in Table 23.12

Glossary

Action spectroscopy: determination of the wavelength dependence of a biological effect. May be used to identify primary chromophores.

ALARA principle: principle of radiation protection to keep all exposures "as low as reasonably achievable".

Apoptosis: "Programmed cell death": disintegration of a dying cell characterized by a sequence of distinct metabolic processes.

Ataxia telangiectasia: Recessive hereditary disease in humans characterized by extreme sensitivity to ionizing radiations.

Bacteriophage: viruses of bacteria.

Cell cycle: Sequence of events between two subsequent cell divisions.

Chromophores: Light absorbing species responsible for a photobiological effect.

Chromosomal aberrations: change in the number of chromosomes (numerical aberration) or in their structure (structural aberrations). Radiation produces mainly the latter.

Chromosome: organizational unit of genetic material in eukaryotes, consists of deoxyribonucleic acid and protein, can be seen in the microscope in mitosis.

Colony-forming ability: the ability of a cell to divide indefinitely. In radiobiology usually equated with survival.

Deoxyribonucleic acid: molecular carrier of genetic information. Commonly abbreviated as DNA.

Dicentric chromosomes: chromosomal aberration where two chromosomes are joined in such a way that there are two

Tab. 23.12 Recommended exposure limits for RF and MW irradiations ("Specific absorption rates (SAR)") (ICNIRP, 1998).

	Frequency range	Whole-body limit (W/kg)	Localized limit head and trunk (W/kg)	Localized limit limbs (W/kg)
Occupational	100 kHz–10 MHz	0.4	10	20
	10 MHz–10 GHz	0.4	10	20
General public	100 kHz–10 MHz	0.08	2	4
	10 MHz–10 GHz	0.08	2	4

centromeres (attachment sites for the spindle during mitosis).

Dimer: UV photoproduct in DNA. Formed between two pyrimidine bases in the same strand.

DNA: abbreviation for *deoxyribonucleic acid,* the macromolecule that carries the genetic information in cells.

Dose equivalent: Formerly, the product of the (physical) dose and the quality factor for different radiation types. It is now replaced by "equivalent dose."

Dose: quantity of absorbed ionizing radiation. Defined as energy absorbed per unit mass. Units: Gray, rad.

Double-strand breaks (DSBs): rupture of the phosphate–sugar backbone of DNA in both strands opposite to each other. DSB are assumed to be the most important lesions caused by ionizing radiations.

Equivalent dose: see "dose equivalent".

Erythema: reddening of the skin ("sunburn").

Escherichia coli: bacterium mostly used in molecular biology.

Eukaryotes: cells possessing a nucleus and chromosomes.

Exposure: quantity defined as the total charge (of one sign) per unit mass liberated by ionizations. Measured in C/kg.

Genome: the total genetic material of a cell.

Gray: SI unit for dose of ionizing radiation. Equal to 1 J/kg. Abbreviation: Gy.

Hyperthermia: Increase of body temperature above normal levels. Used in cancer therapy.

ICNIRP: International Commission for Nonionizing Radiation Protection.

ICRP: International Commission on Radiological Protection, an international nongovernment organization that reviews the scientific literature and issues recommendations for radiation protection. Regularly publishes recommendations in the *Annals of the ICRP.*

Linear energy transfer (LET): energy locally imparted per unit distance. The term *locally* requires specification, for example, by "cutoff energies" for liberated electrons. Unrestricted LET is numerically equal to the stopping power.

Melanoma: malignant invasive skin tumor, starting from melanocytes as precursor cells.

Microdosimetry: Subdiscipline concerned with energy depositions in sites of microscopic dimensions.

Mitosis: The cell division process in eukaryotes.

Mutation: Stably inherited change of genetic information.

Necrosis: Disintegration of a dying cell.

Organogenesis: the phase during embryonic development when the organs are initiated. In humans from about day 8 to day 60 of pregnancy. Sensitive period for teratogenic effects by ionizing radiation.

Phage: bacteriophage.

Photochemotherapy: combination of light and specific photosensitizers for the treatment of diseases.

Photodynamic effect: photobiological effect caused by singlet oxygen formed in sensitizing molecules upon light absorption.

Photoreactivation: Monomerization of UV-induced pyrimidine dimers through light absorption.

Quality factor: Weighting factor to account for the relative biological effectiveness of different kinds of ionizing radiation. Used in radiation protection to determine the equivalent dose.

rad: old unit for the quantity "dose". 1 rad = 0.01 Gy.

Radicals: short-lived aggressive chemical species formed by radiation absorption, which have unpaired electrons in their outer orbitals.

Recombination: exchange of parts of genetic material between DNA molecules or chromosomes.

Relative biological effectiveness (RBE): a factor defined as the ratio of doses to give the same effect with different kinds of ionizing radiations compared to X- or γ-rays.

rem: old unit for the quantity "equivalent dose" and "effective dose." 1 rem = 0.01 Sv.

Roentgen: old unit for the quantity "exposure." Abbreviation R, $1\ \mathrm{R} = 2.58 \times 10^{-4}$ C/kg.

Sievert: SI unit for equivalent dose and effective dose. Abbreviation: Sv.

Stem cells: undifferentiated cells capable of division in organs.

Stopping power: energy lost by a partricle per unit distance.

Syndrome: complex combination of clinical symptoms.

Teratogenesis: formation of malformations during embryonic development.

Translocations: exchange of parts between different chromosomes without loss of genetic material.

Virus: noncellular entity consisting of nucleic acid and a protein coat. Cannot proliferate without the cell's machinery. Most viruses are specific for a given cell type.

References

Beauregard, G., Maret, A., Salvayre, R., and Potier, M. (1987) The radiation inactivation method as a tool to study structure-function relationships in proteins. *Methods Biochem. Anal.*, **32**, 313–43.

BEIR (1999) *Committee on the Biological Effects of Ionizing Radiations The Health Effects of Exposure to Indoor Radon (BEIR VI)*, National Academy of Sciences, Washington, DC.

BEIR (2005) *Committee on the Biological Effects of Ionizing Radiations Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII – Phase 2. Committee to Assess Health Risks from Exposure to Low Levels of Ionizing Radiation*, National Research Council, National Academy Press, Washington, DC.

Berg, G., Spallek, J., Schüz, J., Schlehofer, B., Böhler, E., Schlaefer, K., Hettinger, I., Kunna-Grass, K., Wahrendorf, J., Blettner, M., Interphone Study Group, Germany (2006) Occupational exposure to radio frequency/microwave radiation and the risk of brain tumors: Interphone Study Group, Germany. *Am. J. Epidemiol.*, **164**, 538–48.

Bithell, J.F. and Stewart, A.M. (1975) Pre-natal irradiation and childhood malignancy: a review of British data from the Oxford Survey. *Br. J. Cancer.*, **31**, 271–87.

Caldwell, M.M., Bornman, J.F., Ballare, C.L., Flint, S.D., and Kulandaivelu, G. (2007) Terrestrial ecosystems, increased solar ultraviolet radiation, and interactions with other climate change factors. *Photochem. Photobiol. Sci.*, **6**, 252–66.

Cardis, E. *et al.* (52 authors) (2005) Risk of cancer after low doses of ionising radiation: retrospective cohort study in 15 countries Brit. *Med. J.*, **331**, 77–83.

Cardis, E. *et al.* (30 authors) (2006) Cancer consequences of the Chernobyl accident. *J. Radiat. Prot.*, **26**, 127–40.

Cash, J.C. (2006) Changing paradigms: intensity modulated radiation therapy. *Semin. Oncol. Nurs.*, **22**, 242–48.

Darby S.C., Hill D.C., and European Collaborative Group on Residential Radon and Lung Cancer (2003) Health effects of residential radon: a European perspective at the end of 2002. *Radiat. Prot. Dosimetry*, **104**, 321–29.

Dawson, L.A. and Sharpe, M.B. (2006) *Lancet Oncol.*, **7**, 848–58.

Elder, J.A. (2003) Ocular effects of radiofrequency energy. *Bioelectromagnetics*, (Suppl 6), S148–S161.

European Union (2007) *Report from the Commission on Food Irradiation for the year 2005 (2007/C 122/03) Official Journal of the European Union 2. 6.*

Fenech, M. (2007) Cytokinesis-block micronucleus assay: a comprehensive "cytome" approach for measuring chromosomal instability, mitotic dysfunction andd cell death simultaneously in one assay, in *Chromosomal Alterations: Methods, Results and Importance in Human Health*, (eds G. Obe and Vijayalaxmi), Springer, pp. 241–56.

Field, R.W., Krewski, D., Lubin, J.H., Zielinski, J.M., Alavanja, M., Catalan, V.S., Klotz, J.B., Letourneau, E.G., Lynch, C.F., Lyon, J.L., Sandler, D.P., Schoenberg, J.B., Steck, D.J., Stolwijk, J.A., Weinberg, C., and Wilcox, H.B. (2006) An overview of the North American residential radon and lung cancer case-control studies. *J. Toxicol. Environ. Health A.*, **69**, 599–631.

Friedberg, E.C., Walker, G.C., and Siede, W. (1995) *DNA repair and mutagenesis*, ASM Press, Washington DC.

Gabriel, S., Gabriel, C., and Corthout, E. (1996a) The dielectric properties of biological tissue I. Literature survey. *Phys. Med. Biol.*, **41**, 2231–49.

Gabriel, S., Lau, R.W., and Gabriel, C. (1996b) The dielectric properties of biological tissue II. Measurements in the frequency range 10 Hz to 20 GHz. *Phys. Med. Biol.*, **41**, 2251–69.

Gabriel, S., Lau, R.W., and Gabriel, C. (1996c) The dielectric properties of biological tissue III. Parametric models for the dielectric spectrum of tissues. *Phys. Med. Biol.*, **41**, 2271–93.

Gordon, D., Silverstein, H. and Thorington, L.. (1976, 1980) quoted in *Photochem. Photobiol.*, **32**, 117–29.

Greco, C. and Wolden, S. (2007) Current status of radiotherapy with proton and light ion beams. *Cancer*, **109**, 1227–38.

Hahn, G.M. (1982) *Hyperthermia and Cancer*, Plenum, New York.

Hall, E.J. and Giaccia, A.J. (2005) *Radiobiology for the Radiologist*, 6th edn, Lippincott Williams & Wilkins.

Han, A. and Elkind, M.M. (1979) Transformation of mouse CH3/10T1/2 cecells by single and fractionated doses of X-rays and fission-spectrum neutrons. *Cancer Res.*, **39**, 123–30.

Harm, W. (1980) *Biological Effects of Ultraviolet Radiation*, Cambridge University Press, Cambridge.

Hayashi, M. (2007) In vivo rodent micronucleus assay, in *Chromosomal Alterations: Methods, Results and Importance in Human Health* (eds G. Obe and Vijayalaxmi), Springer, pp. 257–70.

ICRP (2003) *Publication 90:Biological Effects after Prenatal Irradiation (Embryo and Fetus)*, Elsevier.

ICRP International Commission on Radiological Protection (2007) Recommendations of the International Commission on Radiological Protection. ICRP Publication 103. *Ann. ICRP*, **37**(2–4), Amsterdam, Boston, Jena: Elsevier.

International Commission on Non-Ionizing Radiation Protection (ICNIRP) (1997) Guidelines of limits of exposure to broad-band incoherent optical radiation (0.38 to 3 μm). *Health Phys.*, **73**, 539–54.

International Commission on Non-Ionizing Radiation Protection (ICNIRP) (1998) Guidelines for limiting exposure to time-varying electric, magnetic and electromagnetic fields (up to 300 GHz), *Health Phys.*, **74**, 494–515.

International Commission on Non-Ionizing Radiation Protection (ICNIRP) (2000) Revision of the guidelines on limits of exposure to laser radiation of wavelengths between 400 nm and 1.4 μm. *Health Phys.*, **79**, 431–40.

International Commission on Non-Ionizing Radiation Protection (ICNIRP) (2004a) Health issues of ultraviolet tanning appliances used for cosmetic purposes. *Health Phys.*, **84**, 119–27.

International Commission on Non-Ionizing Radiation Protection (ICNIRP) (2004b) Guidelines on limits of exposure to ultraviolet radiation of wavelengths between 180 nm and 400 nm (incoherent optical radiation). *Health Phys.*, **87**, 172–86.

International Commission on Non-Ionizing Radiation Protection (ICNIRP) (2006) ICNIRP statement on far infrared radiation exposure. *Health Phys.*, **91**, 630–45.

International Commission on Radiological Protection (ICRP) (1977) Publication 26: Recommendations of the International Commission on Radiological Protection. *Ann. ICRP*, **1**(3).

International Commission on Radiological Protection (ICRP) (1991) Publication 60 : 1990 Recommendations of the International Commission on Radiological Protection. *Ann. ICRP*, **21**, 1–3.

International Commission on Radiological Protection (ICRP CD1) (2002) *Database of Dose Coefficients: Workers and Members of the Public CD-ROM*, Elsevier.

International Commission on Radiation Units and Measurements (ICRU) (1970) *Report 16: Linear Energy Transfer*, ICRU, Washington, DC.

International Commission on Radiation Units and Measurements (ICRU) (1979) *Report 30: Quantitative Concepts and Dosimetry in Radiobiology*, ICRU, Washington, DC.

International Commission on Radiation Units and Measurements (ICRU) (2000) *Report 63: Nuclear Data for Neutron and Proton Radiotherapy and for Radiation Protection*, ICRU, Washington, DC.

International Commission on Radiation Units and Measurements (ICRU) (2005) Report 73: Stopping of ions heavier than helium. *J. ICRU*, **5**(1).

Kempner, E.S. (1988) Molecular size determination of enzymes by radiation inactivation. *Adv. Enzymol. Relat. Areas Mol. Biol.*, **61**, 107–47.

Kiefer, J. (1990) *Biological Radiation Effects*, Springer, Berlin.

Kiefer, J. (2002) Mutagenic effects of heavy charged particles. *J. Radiat. Res. (Tokyo)*, **43**(Suppl), S21–S25.

Kiefer, J. (2007) Effects of ultraviolet radiation on DNA, in *Chromosomal Alterations: Methods, Results and Importance in Human Health* (ed G., Obe and Vijayalaxmi), Springer, pp. 39–51.

Kiefer, J. (2008) The physical basis for the biological action of heavy ions. *New J. Phys.*, **10**(075004), 15.

Kiefer, J. and Wienhard, I. (1977) Biologische Wirkungen ultravioletter Strahlung, in *Ultraviolette Strahlen* (ed J., Kiefer), de Gruyter, Berlin, pp. 445–556.

Kinsey, V.E. (1948) Spectral transmission of the eye to ultra-violet radiation. *Arch. Ophthalmol.*, **39**, 508–13.

McKinlay, A.F. and Diffey, B.L. (1987) A reference action spectrum for ultraviolet induced erythema in human skin CIE Research Note 6, pp. 17–22.

Meltz, M.L. (2003) Radiofrequency exposure and mammalian cell toxicity, genotoxicity, and transformation. *Bioelectromagnetics*, (Suppl 6), S196–S213.

Meyer, J.L. (ed) (2002) *The Radiation Therapy of Benign Diseases Frontiers of Radiation Therapy and Oncology*, Vol. 35, Karger, Basel-Freiburg-Paris-London-New York.

Molins, R. (ed) (2001) *Food irradiation – Principles and applications*, Wiley Interscience, New York.

Obe, G. and Vijayalaxmi (eds) (2007) *Chromosomal Alterations: Methods, Results and Importance in Human Health*, Springer.

Parrish, J.A., Anderson, R.R., Urbach, J., and Pitts, D. (1978) *UV-A: Biological Effects of Ultraviolet Radiation with Emphasis on Human Responses to Longwave Ultraviolet*, Plenum, New York.

Pierce, D.A., Shimizu, Y., Preston, D.L., Vaeth, M., and Mabuchi, K. (1996) Studies of mortality of atomic bomb survivors. Report 12. Part I. Cancer: 1950–1990. *Radiat. Res.*, **146**, 1–27.

Philips, N.A., Gallagher, R.P., Coldman, A.J., and Band, P.R. (1996) Skin cancer incidence: what has happened and what can we expect, in *Environmental UV-radiation, risk of skin cancer and primary prevention, Veröffentlichungen der Strahlenschutzkommission, Band 34*, Stuttgart: G. Fischer Verlag, pp. 323–32.

Preston, D.L., Kusumi, S., Tomonaga, M., Izumi, S., Ron, E., Kuramoto, A., Kamada, N., Dohy, H., Matsuo, T., and Mabuchi, K. (1994) Cancer incidence in atomic bomb survivors: Part III. Leukemia, lymphoma and multiple myeloma. 1950–1987. *Radiat. Res*, **137**((),Suppl), S68–S97.

Preston, D.L., Ron, E., Tokuoka, S., Funamoto, S., Nishi, N., Soda, M., Mabuchi, K., and Kodama, K. (2007) Solid cancer incidence in atomic bomb survivors: 1958–1998. *Radiat. Res.*, **168**, 1–64.

Preston, D.L., Shimizu, Y., Pierce, D.A., Suyama, A., and Mabuchi, K. (2003) Studies of mortality of atomic bomb survivors. Report 13: solid cancer and noncancer disease mortality: 1950–1997. *Radiat. Res.*, **160**, 381–407.

Regulla, D.F. (2005) ESR spectrometry: a future-oriented tool for dosimetry and dating. *Appl. Radiat. Isot.*, **62**, 117–27.

Rossi, H.H. and Zaider, M. (1996) *Microdosimetry and its Applications*, Springer, Berlin.

Rowland, F.S. (2006) Stratospheric ozone depletion. *Philos. Trans. Royal Soc. (London) B Biol. Sci.*, **361**, 769–90.

Savage J.R.K. (1978) Some thoughts on the nature of chromosomal aberrations and their use as a quantitative endpoint for radiobiological studies, in *6th Symposium Microdosimetry*, Harwood, London pp. 39–54.

Schulz-Ertner, D., Jakel, O., and Schlegel, W. (2006) Particle radiation therapy using proton and heavier ion beams. *Semin. Radiat. Oncol.*, **16**, 249–59.

Schulz-Ertner, D. and Tsujii, H. (2007) Particle radiation therapy using proton and heavier ion beams. *J. Clin. Oncol.*, **25**, 953–64.

Seegenschmiedt, M.H., Micke, O., and Willich, N. (2004) Radiation therapy for nonmalignant diseases in Germany. *Strahlentherapie und Onkologie*, **180**, 718–30.

Sevan'kaev, A., Khvostunov, I., Lloyd, D., Voisin, P., Golub, E., Nadejina, N., Nugis, V., Sidorov, O., and Skvortsov, V. (2006) The suitability of FISH chromosome painting and ESR-spectroscopy of tooth enamel assays for retrospective dose reconstruction. *J. Radiat. Res. (Tokyo)*, **47**(Suppl A), A75–A80.

Sommers, C.H. and Fan, X. (eds) (2006) *Food Irradiation Research and Technology*, Blackwell Publishing, Ames, IA.

SSK (1996) Strahlenschutzkommission, in *Environmental UV-radiation, risk of skin cancer and primary prevention*, Veröffentlichungen der Strahlenschutzkommission, 34, G. Fischer-Verlag, Stuttgart.

SSK (2006) Strahlenschutzkommission Wirkung hochfrequenter Felder auf das Genom: Genotoxizität und Genregulation-Wirkung hochfrequenter Felder auf das Genom: Genotoxizität und Genregulation (in German, English translation in preparation) Stellungnahme der Strahlenschutzkommission.

Takebayashi, T., Akiba, S., Kikuchi, Y., Taki, M., Wake, K., Watanabe, S., and Yamaguchi, N. (2006) Mobile phone use and acoustic neuroma risk in Japan. *Occup. Environ. Med.*, **63**, 802–7.

Underbrink, A.G., Sparrow, A.H., and Pond, V. (1968) Chromosomes and cellular radiosensitivity II: Use of interrelationship among chromosome volume, nucleotide content and D0 of 120 diverse organisms in predicting radiosensitivity. *Radiat. Bot.*, **8**, 205–38.

United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) (1988) *Sources, Effects and Risks of Ionizing Radiation*, United Nations, New York.

UNSCEAR (2000) United Nations Scientific Committee on the Effects of Atomic Radiation, in *Sources and Effects of Ionizing Radiation*, United Nations, New York.

United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) (2001) *Hereditary Effects of Radiation*, United Nations, New York.

United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) (2009) *Effects of ionizing radiation*, Vol. 2, United Nations, New York, in print.

Urbain, W.M. (1986) *Food Irradiation*, Academic Press, Orlando, FL.

Valberg, P.A., van Deventer, T.E., and Repacholi, M.H. (2007) Workgroup report: base stations and wireless networks-radiofrequency (RF) exposures and health consequences. *Environ. Health Perspect.*, **115**, 416–24.

Vecchia, P., Hietanen, M., Stuck, B.E., Deventer, Evan., and Niu S. (eds) (2007) *Protecting Workers from Ultraviolet Radiation*. International Commission on Non-Ionizing Radiation Protection In Collaboration with:International Labour Organization World Health Organization ICNIRP 14/2007.

Veit, R., Griebel, J., and Bauer, B. (2001) Patientenexposition in der Röntgendiagnostik, in *Ganzkörperbestrahlung – Strahlenschutz substanzen – Strahlenexposition in der Diagnostik* (Hrsg.: eds D. Gottschild and Chr. Reiners), Strahlenschutz in Forschung und Praxis Band 44, Verlag Urban & Fischer, München-Jena, pp. 95–101.

Vijayalaxmi and Obe G. (2004) Controversial cytogenetic observations in mammalian somatic cells exposed to radiofrequency radiation. *Radiat. Res.*, **162**, 481–96.

Von Sonntag, C. (2007) DNA lesions induced by ionizing radiation, in *Chromosomal Alterations: Methods, Results and Importance in Human Health* (ed G., Obe and Vijayalaxmi), Springer, pp. 21–38.

Wakeford, R. and Little, M.P. (2003) Risk coefficients for childhood cancer after intrauterine irradiation: a review. *Int. J. Radiat. Biol.*, **79**, 293–309.

World Health Organization (WHO) (1982), *Nonionizing Radiation Protection*, WHO Regional Office for Europe, Copenhagen.

World Health Organization (WHO) (2003) *Artificial Tanning Sunbeds – Risks and Guidance*, WHO, Geneva.

Wojcik, A. and Obe, G. (2007) Sister-chromatid exchanges, in *Chromosomal Alterations: Methods, Results and Importance in Human Health* (eds G. Obe and Vijayalaxmi) Springer, pp. 271–284.

Young, R.W. and Kerr, G.D. (eds) (2005) *Reassessment of the Atomic Bomb Radiation Dosimetry for Hiroshima and Nagasaki, Dosimetry System 2002,* The Radiation Effects Research Foundation, Hiroshima.

24

Biological Effects of Sound and Ultrasound

Wesley L. Nyborg

24.1 **Introduction 811**
24.2 **Biological Effects from Sonically Induced Temperature Elevation 812**
24.2.1 Basic Theory 812
24.2.2 Safety Concerns for Diagnostic Ultrasound 813
24.2.3 Thermal Therapy Using Ultrasound 813
24.3 **Biological Effects from Cavitation at Low and Intermediate Levels 814**
24.3.1 Introduction 814
24.3.2 Activity at Low Sound Levels 815
24.3.2.1 Linear Theory of Bubble Vibration 815
24.3.2.2 Scattering: Ultrasound Contrast Agents 816
24.3.3 Activity at Intermediate Sound Levels 817
24.3.3.1 Heating 817
24.3.3.2 Rectified Diffusion 818
24.3.3.3 Acoustic Radiation Forces on Free Gas Bubbles and UCAs 818
24.3.3.4 Attraction of Particles to a Vibrating Bubble or UCA 819
24.3.3.5 Microstreaming Associated with Vibrating Bubbles and UCAs 821
24.4 **Biological Effects from Nonthermal Noncavitational Mechanisms 822**
24.4.1 Introduction 822
24.4.2 Acoustic Radiation Force 822
24.4.3 Acoustic Streaming and Microstreaming 825
24.4.4 Other Nonthermal Noncavitational Effects 826
24.5 **Nonlinear Cavitation Activity 829**
24.5.1 Introduction 829
24.5.2 Bubble Dynamics Under High Amplitude Conditions 829
24.5.3 Effects of Cavitation in Biological Suspensions 830
24.5.4 Effects of Cavitation in Tissues 831
24.6 **Discussion 833**
Glossary 833
References 834
Further Reading 836

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

24.1
Introduction

Although living systems, including humans, are exposed to sound (including ultrasound) in the natural environment, the main effects of the exposure (other than sensory) comes from human activities and man-made devices. With the advent of electronic technology in the early 1900s, it became possible to generate ultrasound conveniently and economically. It was soon found in laboratory experiments that ultrasound, in the frequency range 10 kHz to a few megahertz, was capable of producing a wide range of effects in physical, chemical, and biological systems. Reviews and selected papers from the early work are contained in publications by El’piner (1964), Dunn and O’Brien (1976), Wells (1977), and Nyborg (1991, 1999). (The present chapter is essentially a revision of the latter two references, which includes advances since 1999.)

Naturally, there was a great deal of curiosity as to how ultrasound produced these effects. Under some circumstances, temperature elevation was shown to be the cause. As ultrasound propagates through biological tissue, absorption occurs (NCRP, 1992); that is, acoustic energy is converted into heat. The temperature rises to an extent that depends on the nature of the sound field as well as on characteristics of the tissue. In some experimental situations with mammals, the temperature elevation is high enough to produce structural or functional change, either beneficial or adverse. Thermal mechanisms for biological effects of sound are taken up in Section 24.2.

In other applications, temperature elevation is negligible. Chief among nonthermal mechanisms is *cavitation*, an activity that involves bubbles or cavities containing gas or vapor. This topic has taken on greatly increased significance during the last decade because of the practice, now widespread in examinations with diagnostic ultrasound, of injecting gas-containing microspheres into the blood as contrast agents to improve the images. There are many facets to cavitation activity; gas bubbles exhibit a surprising variety of behavior in a sound field. This topic is discussed in Sections 24.3 and 24.5.

Other aspects of sound by which action occurs include steady (time-averaged) forces, torques, and steady circulations set up in various media. Radiation forces and torques occur also in electromagnetic fields, but the acoustic counterparts

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

have features that are unique to sound fields. These are taken up in Section 24.4, along with some intriguing phenomena for which the mechanism has not been determined, though believed to be neither heat nor cavitation.

An understanding of the mechanisms for biological effects of sound are important in connection with such widely varying medical applications as stimulation of tissue repair and destruction of malignant tissue. Applications to cellular biotechnology include disintegration of cells, cell separation, cell fusion, and, most recently, mediation of gene therapy by making cells temporarily permeable. In addition, there is also considerable need for increased understanding of mechanisms in order to evaluate risks involved in medical applications. This is especially true for applications of diagnostic ultrasound to examinations of the fetus, since these are now employed very widely. Today, most children have been exposed to ultrasound before birth.

24.2
Biological Effects from Sonically Induced Temperature Elevation

24.2.1
Basic Theory

In a fluid-like medium, sound energy is converted irreversibly into heat by viscous dissipation and relaxation processes (Pierce, 1994). In a continuous-wave sound field of moderate pressure amplitude, where the pressure at a point varies above and below the ambient pressure by an amount that varies sinusoidally with time, the time-averaged rate (q_v) at which heat is generated per unit volume is often given adequately by the expression

$$q_v = \alpha {p_0}^2/\rho_0 c \tag{1}$$

Here, p_0 is the pressure amplitude, α is the absorption coefficient in nepers per unit length, ρ_0 is the density of the fluid, and c is the speed of sound in the fluid (Nyborg, 1981; Cavicchi and O'Brien, 1984). In a plane traveling wave, energy is transported along its direction of propagation at a rate I per unit area given by

$$I = 0.5p_0^2/\rho_0 c \tag{2}$$

Combining Eqs (1) and (2), one obtains, simply,

$$q_v = 2\alpha I \tag{3}$$

For a plane wave traveling along the x direction, the pressure amplitude (p_0) is proportional to $e^{-\alpha x}$ if scattering is negligible, as has been shown to be true for liver and other soft tissues (NCRP, 1986). From Eq. (2), the intensity is then given by

$$I = I_0 e^{-2\alpha x} \tag{4}$$

where I_0 is the intensity at $x = 0$. One never has perfect plane waves, but Eqs (2)–(4) are often good approximations if the sound field resembles a traveling wave and is not too nonuniform.

In mammalian tissues, heat is transported not only by conduction but also by convection in the flowing blood, that is, by *perfusion*. Knowing q_v for a given situation where ultrasound is propagating through perfused mammalian tissue, the temperature elevation can be approximated by means of appropriate solutions to a widely used *bioheat transfer equation* from Pennes (1948). This can be written (NCRP, 1992) as

$$\frac{\partial T}{\partial t} = \kappa \nabla^2 T - \frac{\Delta T}{\tau} + \frac{q_v}{c_v} \qquad (5)$$

where T is the temperature and ΔT its elevation above the base temperature; the tissue is characterized by a thermal diffusion constant κ, a perfusion time constant τ, and a heat capacity per unit volume c_v. Basic solutions have been obtained for a point source that begins generating heat at $t = 0$; also, for spherical, line, and discoid sources, as well as for ultrasound beams propagating through perfused absorbing media. These have proved useful in arriving at safety criteria for diagnostic ultrasound and, also, in planning therapeutic or surgical applications of ultrasound.

24.2.2
Safety Concerns for Diagnostic Ultrasound

At frequencies in the megahertz range used in most medical ultrasound, acoustic absorption coefficients for mammalian tissues are high enough to allow temperature elevations produced in medical applications to be significant, even in some applications of diagnostic ultrasound. In tissues where cells divide rapidly, as in the mammalian embryo or fetus, even a small temperature rise can produce damage if prolonged sufficiently. For a wide range of teratogenic effects, the time t in minutes required for exposure to temperature T (in degrees Celsius) to produce an effect has been equated, approximately, to

$$t = 4^{43-T} \qquad (6)$$

when T is not greater than 43 °C (Miller and Ziskin, 1989; NCRP, 1992). According to Eq. (6), damage may occur after 4 minutes of exposure to a temperature of 42 °C, and after 16 minutes of exposure to 41 °C. For more recent analysis of possibilities for thermal risk to the fetus, see Miller, Miller, and Church (2005). Among fetal tissues, the absorption coefficient is highest in bone, and the potential for heating increases with gestational age. Safety criteria derived from algorithms for calculating estimates of the maximum temperature rise that would be produced by ultrasonic exposure of bone and other tissues have been formulated during the past few years (AIUM/NEMA, 1992; NCRP, 1992). These criteria are expressed in terms of *thermal indices,* which are becoming familiar to users of diagnostic ultrasound equipment.

24.2.3
Thermal Therapy Using Ultrasound

First introduced in the 1940s and 1950s, ultrasound, focused or unfocused, has received increasing attention in recent years as a mean of modifying tissue in selected parts of the body; special emphasis has been given to its capability for elevating tissue temperature in a controlled manner. Applications of unfocused ultrasound to physical therapy are discussed in reviews by Lehmann and de Lateur (1990) and McDiarmid, Ziskin and Michlovitz (1996). As reported by these authors, heating produced by ultrasound may lead to increase in collagen extensibility, alterations in blood flow, and increases in enzymatic activity of skeletal muscle. By using frequencies ranging from 1.0 to 3.0 MHz and intensities in the range 1.0–3.0 W cm^{-2}, it is possible to increase significantly the temperature of tissue at depths of 5 cm or more. Possibilities can be seen for an example by using Eqs (1)–(4). At a frequency of 1 MHz the absorption coefficient α of muscle tissue is, roughly, 0.1 Np cm^{-1} (NCRP, 1986). From Eq. (4), the intensity of a 1 MHz beam entering muscle tissue with an initial value of 1.0 W cm^{-2} will be 0.37 W cm^{-2} at a distance of 5 cm and, according to Eq. (3), heat will be deposited there at the rate of 0.074 J cm^{-3} s^{-1}. Taking the heat capacity to be 4.2 J cm^{-3} for the tissue, one concludes, neglecting effects of heat conduction and perfusion, that the tissue temperature will be increased at the rate of 0.0176 °C s^{-1}, or 1.06 °C min^{-1}. (Neglecting effects of heat conduction is reasonable for exposures with a broad beam, e.g. with a cross-sectional area of the order of 10 cm^2, applied for only a few minutes).

When using focused ultrasound, a much higher intensity can be applied to a designated small region of tissue. In techniques developed for medical applications using high *intensity focused ultrasound* (HIFU), temperatures ranging from 60 to 90 °C are reached. In a general review of current practice, ter Haar (2004) described applications of FUS to thermal ablation of tissue in the prostate, liver, and other organs. It is reported by Wu (2006a) that HIFU is widely employed in China, where 3500 patients with solid tumors were treated with this modality in 20 hospitals during the period December 1997 to March 2001. A problem in treating internal tumors is to ensure that the entire malignant region is raised to the required temperature. It is shown by Hynynen and McDannold (2006) that the temperature can be monitored successfully by the use of magnetic resonance imaging (MRI) and exploiting the thermal sensitivity of the water proton resonance frequency shift. Many articles describing recent findings from research on biological effects produced by focused ultrasound are contained in a conference proceedings volume (Clement, McDannold, and Hynynen, 2006).

24.3
Biological Effects from Cavitation at Low and Intermediate Levels

24.3.1
Introduction

It was found during the 1920s and 1930s, in the earliest experiments on physical, chemical and biological effects of ultrasound in liquid media, that many of the effects were caused by ‘‘cavitation,’’ an activity occurring when small cavities (‘‘bubbles’’) filled with gas or vapor were present. Since then, this activity has been extensively studied. The physical response of a bubble to ultrasound can be, in one extreme, simple sinusoidal pulsation or, in the other, complex motion including violent collapse. The effect on a biological cell from exposure to ultrasound in the presence of bubbles can vary from mild perturbation to complete destruction. In mathematical analysis of mild bubble response to a sound field, linear equations are used, analogous to those employed in treating the classic ‘‘mass-on-a-spring’’ problem in introductory physics. For the more complex response, the equations required are nonlinear and are dealt with by use of computational techniques. The subject has been reviewed for various purposes by Flynn (1964), Coakley and Nyborg (1978), Neppiras (1980), Apfel (1981), Atchley and Crum (1988), Young (1989, 2005), and Leighton (1994, 1996).

24.3.2
Activity at Low Sound Levels

24.3.2.1 Linear Theory of Bubble Vibration

Theory for cavitation is based primarily on a model in which a small spherical body of gas, the bubble, exists in a homogeneous liquid, away from disturbing boundaries. It is set into volume (breathing) oscillation by the time-varying acoustic pressure (p) associated with a sound field that is set up in the liquid. (The acoustic pressure is defined as the difference between the instantaneous pressure at a point in the field and the static ambient pressure.) If p varies sinusoidally in time with frequency f, and if the pressure amplitude (p_0) is not too high, the radial displacement (ξ) of the bubble (increase in bubble radius) will also vary sinusoidally in time with frequency f and with ξ proportional to p. Under these conditions the equation of motion is a linear differential equation:

$$m\frac{\partial^2 \xi}{\partial t^2} + b\frac{\partial \xi}{\partial t} + k\xi = F(t) \qquad (7)$$

By analogy to the mass-on-a-spring problem, the parameters m and k are the effective mass and spring constant, respectively, while b is proportional to the damping constant. If there is a film, shell, or collection of impurities at the surface of the gaseous body, these constants may be affected; in the preparation of *contrast agents,* discussed in Section 24.3.2.2, they are altered deliberately. In this section, the bubble is considered to be a "free bubble," consisting simply of homogeneous gas surrounded by homogeneous liquid. For a bubble whose equilibrium radius is R_0, in a liquid of density ρ_0, the effective mass is given by

$$m = 4\pi \rho_0 {R_0}^3 \qquad (8)$$

this is three times the mass of liquid that would be contained in a sphere equal in volume to the bubble. When effects of surface tension are ignored, the effective stiffness constant, k, for the bubble is proportional to the polytropic index, γ, being given by

$$k = 12\pi \gamma R_0 P_a \qquad (9)$$

where P_a is the ambient hydrostatic pressure. For a free air bubble, γ is 1.4 if the changes in volume occur adiabatically. The effective driving force function $F(t)$ is proportional to the local acoustic pressure in the fluid (specifically, the pressure that would exist in the absence of the bubble); taking this pressure to be $p_0 \cos(\omega t)$, where ω is the angular frequency $2\pi f$, the driving function is found to be

$$F(t) = -4\pi {R_0}^2 p_0 \cos(\omega t) \qquad (10)$$

Like its mass-on-a-spring analog, the oscillating bubble responds with increased displacement amplitude when the angular frequency is in the vicinity of the resonance value:

$$\omega = \omega_0 = (k/m)^{1/2} \qquad (11)$$

Corresponding to ω_0 is the resonance frequency $f_0 = \omega_0/2\pi$. Choosing convenient units, one obtains from Eqs (7), (8), and (11) a simple formula for f_0, applicable in the lower ultrasonic frequency range to an air bubble in water at 20 °C and standard atmospheric pressure, namely,

$$f_0 = 3.3/R_0 \qquad (12)$$

Here f_0 is in kilohertz when R_0 is in millimeters. For example, f_0 is about 33 kHz for a bubble of radius 0.1 mm. However, this simple formula fails at megahertz frequencies and must be modified to take into account surface tension. For a free bubble of 1 μm radius, the computed f_0 in water, assuming adiabatic conditions, is 3.3 MHz if surface tension is neglected and 4.7 MHz if air–water surface tension for 20 °C is assumed.

When the frequency of a driving ultrasound field is equal to the resonance frequency of the bubble, the ratio (ξ_0/R_0) of the radial displacement to the bubble radius is given by the expression $p_0/\delta P_a$, where δ (equal to $b/m\omega_0$) is a damping constant that depends on properties of the liquid. For an air bubble in water, resonant at a frequency of 1 MHz, δ has been calculated to be 0.14; hence, for these conditions, a relative pressure amplitude (p_0/P_a) will produce a relative change in radius (ξ_0/R_0) equal to $7p_0/P_a$. For example, if the ratio of p_0 to atmospheric pressure is 0.1%, the ratio of the oscillatory amplitude of the bubble to its radius will be 0.7%. This is, of course, a much greater change in radius than would be produced in a solid or liquid sphere.

24.3.2.2 **Scattering: Ultrasound Contrast Agents**

In the situation discussed above, a gas-filled bubble in a liquid where there is an ultrasound field of given frequency f, will pulsate at the same frequency and will then itself become a source of ultrasound, which radiates outward from the bubble. The latter is thus an agent that is capable of *scattering* acoustic energy from the original field. The usefulness of this capability in medicine was discovered in the late 1960s and early 1970s by investigators using diagnostic ultrasound to study blood flow in the circulatory system. In their studies, the blood had been made visible mostly by ultrasound scattered from erythrocytes and other cells. However, they found that when suitable bubbles were present in the blood they scattered far more energy than the cells and greatly increased the contrast between blood and the background tissue (Gramiak, Shah, and Kramer, 1969; Ziskin *et al.*, 1972). These findings created considerable interest, and led to much use of small bubbles as contrast agents. It was necessary to find means of stabilizing the bubbles by adding suitable material to their surfaces to avoid rapid dissolution by surface tension. By 1990, commercial products had become available and detailed investigations of their acoustical properties were being carried out (Bleeker, Shung, and Barnhart, 1990). Since then, many more types of bubble-based ultrasound contrast agents (UCAs) have been commercially produced; 16 are described by Bouakaz and de Jong (2007). The UCAs must be small, a few micrometers in diameter, in order to pass through narrow capillaries. The majority are encapsulated with flexible shells or films, 10–250 nm in thickness, which prevent the gas from dissolving too rapidly, and, also, affect the dynamics. Fluorocarbons of high molecular weight are often used as the gas, since these dissolve relatively slowly.

24.3.3
Activity at Intermediate Sound Levels

When UCAs of uniform size and specified characteristics became available commercially, it became possible to carry out ultrasound exposures under a degree of control over bubble population, which was much greater than with free bubbles. Both *in vitro* and *in vivo* studies have shown that the presence of either free (uncoated) bubbles or UCAs increases the likelihood of biological effects, unless the pressure amplitude is fairly low. In obtaining the experimental results discussed in the following sections, the sound levels were such that the concepts of linear acoustics are useful, although the assumptions are not fully applicable.

In dealing with phenomena of acoustics, which occur at sound levels that are somewhat higher than those at which linear acoustics applies, it is common to use a method of successive approximations in which quantities such as the displacement, velocity, and pressure are each represented as a series of terms of decreasing order of magnitude. For example, the acoustic pressure might be written as

$$p = P_a + p_1 + p_2 + \cdots \tag{13}$$

Here, the zero-order term P_a is the pressure, which exists in the absence of sound, while the first-order term p_1 is the pressure in the "linear acoustics" approximation, and varies periodically in time with frequency f. The second-order term is found to be a constant times the product of two first-order quantities and consists of two parts: one is "second-harmonic," being periodic with frequency $2f$, and the other is time-independent. It is the latter, time-independent second-order quantities that are particularly relevant in discussing biological effects of ultrasound at intermediate sound levels. The quantities q_v and I, for which expressions are given in Eqs (1) and (2), are time-averaged second-order quantities, each being proportional to the square of p_0, the amplitude of the first-order pressure. Later sections deal with radiation pressure, radiation force, and acoustic streaming, all of which are treated as time-independent quantities of second order.

24.3.3.1 Heating

Gas bubbles of appropriate size can be very effective in converting acoustical energy into heat. The energy per unit volume extracted from a plane ultrasound wave of intensity I traveling through a suspension of n bubbles per unit volume is given by the product $n\sigma I$, where σ is the total single-bubble cross-section for absorption and scattering. For a free resonant air bubble in water, σ is given approximately (Coakley and Nyborg, 1978) by

$$\sigma = 2900\,\pi R_0{}^2 \tag{14}$$

thus, the area σ is 2900 times the cross-sectional area (πR_0^2) of the bubble. Equating the expression for q_v in Eq. (3), to the expression $n\sigma I$, one finds that the attenuation coefficient is equal to $n\sigma/2$. Taking the radius R_0 at 1 MHz to be 3 μm and considering about 90% of the attenuation to be caused by absorption, one obtains $3.7 \times 10^{-4} n$ as the absorption coefficient α in Np/cm for the bubble suspension. Choosing n as 1000 cm^{-3}, corresponding to an average of one resonant bubble per cubic millimeter and a volume fraction of air equal to 1.1×10^{-7}, one obtains 0.37 Np cm^{-1} for α, a value larger than the measured absorption coefficient for most mammalian soft tissues at 1 MHz.

For UCAs, the vibrations induced by ultrasound are restrained by the films or shells at their surfaces, but they nevertheless cause high absorption. Marsh *et al.* (1997) measured the attenuation (most of which was absorption) of ultrasound in highly concentrated solutions of a commercial contrast agent, using frequencies in the range 1.0–10 MHz and obtained values up to about 9 Np cm^{-1}. Such values are higher than reported absorption coefficients for any soft tissues (NCRP, 1986; Duck, 1990). The presence of UCAs in tissue being exposed to ultrasound will increase the temperature rise produced; this can be helpful in therapeutic applications and, on the other hand, needs to be recognized in safety considerations.

24.3.3.2 Rectified Diffusion

It was observed in early experience with ultrasound that small free gas bubbles in water may become larger after exposure to ultrasound. For single bubbles, this phenomenon is now understood as resulting from the process of gas diffusing into and out from a pulsating bubble. As a bubble expands and contracts in response to the time-varying pressure in a sound field, gas diffuses inward and outward. The net inward flow during a cycle is, in general, nonzero; this is partly because the area of the gas–liquid interface varies and partly because the gas-concentration gradient on the liquid side of the interface varies during the cycle (Hsieh and Plesset, 1961). The net inward flow is often positive, causing the bubble to grow in a process called *rectified diffusion*. Theories have been developed for the process, and experiments have been done to test them, especially in water and other aqueous media. Crum (1984) has reviewed the state of knowledge on conditions for growth and on the rate of growth.

24.3.3.3 Acoustic Radiation Forces on Free Gas Bubbles and UCAs

Here the term "bubble" is used to represent either a free bubble or a UCA, unless noted otherwise. In a standing-wave field, any gas bubbles that are smaller than the resonance size tend to collect at maxima of the pressure amplitude, while larger ones move to the minima. This behavior, attributed to acoustic radiation forces (ARFs), is readily explained by considering the force exerted by surface stresses on a highly compressible body in a field over which the acoustic pressure varies sinusoidally in time, with spatially varying amplitude but with constant phase in the vicinity of the body. The instantaneous force is given by $-V\nabla p$, where V is the instantaneous volume of the body, ∇ is the gradient operator, and p is the instantaneous acoustic pressure near the bubble. The acoustic pressure varies in time with frequency f, while V is assumed to be the sum of a constant part V_0 and a part v that varies in time with the same frequency f.

The time-averaged force over a cycle, that is, the ARF, is then equal to $\langle -v\nabla p\rangle$, where $\langle\rangle$ indicates the temporal average. Since v and p are taken to be first-order quantities, the radiation force is of second order (as indicated earlier). This time-averaged product (of two sinusoidally varying quantities) is equal to the product of the two amplitudes times the cosine of the phase difference between them. From these considerations, it is found that for a small bubble the radiation force is in the direction of increasing pressure amplitude, for a large one it is in the opposite direction, and for a resonant bubble it is zero. These aspects of radiation force on bubbles are important in designing arrangements for exposing cell suspensions to ultrasound when free bubbles are involved. If the arrangement is such that standing waves are established with periodically spaced maxima and minima of the pressure amplitude, bubbles smaller than resonance size will move toward the pressure maxima and become active there, perhaps growing to resonance size, while the larger bubbles will go to the pressure minima and become less active there. (Some of these may then shrink and be returned to a pressure maximum.) Hence, the effects on cells will occur mostly at the pressure maxima unless means are employed to rotate the vessel or stir the suspension.

In a traveling wave, the situation is unlike that for a standing wave; here, the ARF on a gas bubble is always in the direction of sound propagation and is maximum for bubbles of resonance size (Coakley and Nyborg, 1978; NCRP, 2002, Chapter 4).

Since UCAs have become available, much has been learned about their acoustical characteristics, as well as their potential for biomedical applications. When UCAs are present in a liquid through which an ultrasound beam is passing, they move in the direction of the beam, as expected from theory for radiation force in a traveling wave (Dayton, Allen, and Ferrara, 2002). This force has implications for applications in which UCAs are present in the blood of a patient who is being examined with diagnostic ultrasound. Experiments with small animals have shown that when an ultrasound beam passes across a blood vessel, any UCAs in the blood will be forced toward the distal side of the vessel. This capability may be useful in drug delivery or other applications where it is desired to bring a UCA in contact with a binding site on the epithelium.

If two or more free bubbles are near one another in a sound field, they will often move toward each other and coalesce to form a single larger bubble. It is this process and that of rectified diffusion (Section 24.3.3.2) that are believed to account for the growth of free gas bubbles in a sound field. The attraction that leads to coalescence can be explained in terms of radiation forces (Coakley and Nyborg, 1978). For UCAs, the forces between bubbles are of the same nature as for free bubbles. The attractive forces bring resonant UCAs together and aggregates form. Dayton *et al.* (1997) have studied the behavior of UCA aggregates in flow fields while exposed to ultrasound, finding conditions under which deaggregation occurs temporarily, followed by reaggregation.

24.3.3.4 **Attraction of Particles to a Vibrating Bubble or UCA**

It is shown in Section 24.4.2 that any small vibrating object in a suspension of particles will create highly localized fields of radiation force that act on neighboring particles. In the vicinity of a vibrating bubble, any particle nearby is acted on by a radiation force that is directed toward the bubble if the particle is more dense than the surrounding fluid. Hence, particles in solution tend to collect at the surfaces of vibrating bubbles. This conclusion can be reached by considering the situation of a small particle of volume v near a vibrating spherical bubble of rest radius R_0 as a special application of Eqs (21a and b) in Section 24.4.2 (Nyborg and Miller, 1982; NCRP, 2002). The following equation is obtained:

$$F_{\mathrm{rad}} = vD\frac{\partial}{\partial r}\langle \mathrm{KE}\rangle \tag{15}$$

where D is defined in Eq. (21b) and $\langle \mathrm{KE}\rangle$ is the time-averaged kinetic energy density in the liquid immediately outside the bubble. (The term in Eq. (21a) involving $\langle \mathrm{PE}\rangle$, the time-averaged potential energy, is negligible in this region.) The alternating flow here is essentially that of an incompressible fluid at points whose distance from the bubble center is small compared to the acoustic wavelength. The bubble surface is considered to be vibrating sinusoidally at frequency f with displacement amplitude ξ_0, and velocity amplitude U_0, equal to $2\pi f\xi_0$. The corresponding oscillatory velocity of the liquid at a distance r from the bubble center is radially directed, with amplitude u_0, given by $U_0R_0^2r^{-2}$. The time-averaged kinetic energy density $\langle \mathrm{KE}\rangle$ in the fluid is $0.25\rho_0u_0^2$, where ρ_0 is the density of the liquid. If $(\rho - \rho_0)$ is much less than ρ_0, where ρ is the density of the particle, the combination $D\rho_0$ can be replaced by $(\rho - \rho_0)$ and one obtains from Eq. (15)

$$F_{\mathrm{rad}} = -v(\rho - \rho_0)U_0{}^2R_0{}^4r^{-5} \tag{16}$$

the minus sign indicating that the force is directed toward the bubble. Considering the liquid to be Newtonian with viscosity coefficient η, and that the Stokes law applies, the particle will move toward the bubble, acquiring the limiting velocity U_{lim}, equal to $F_{\mathrm{rad}}/6\pi\eta a$. Taking the particle to be spherical with radius a and the volume v thus equal to $4\pi a^3/3$, one obtains

$$U_{\mathrm{lim}} = (2/9\eta)a^2(\rho - \rho_0)U_0{}^2R_0{}^4r^{-5} \tag{17}$$

When considering the particles to be biomolecules, such as protein molecules, it is convenient to characterize them in terms of the *sedimentation coefficient* s, a quantity used in centrifugation measurements and other applications. For a particle driven through a viscous fluid, the quantity s, in units of seconds, is the ratio of its terminal speed to the acceleration associated with the driving field (IUPAC, 1997). In the present application, the acceleration is acoustical in origin. In a liquid of Newtonian viscosity η where Stokes Law applies, it is found for a spherical particle of radius a that s is given by

$$s = 2a^2(\rho - \rho_0)/9\eta. \tag{18}$$

Combining Eqs (17) and (18), one obtains, simply,

$$U_{\mathrm{lim}} = sU_0{}^2R_0{}^4r^{-5} \tag{19}$$

As an example, suppose a UCA with a radius R_0 of 5.0 μm vibrates at a frequency of 1 MHz with a displacement amplitude of 0.50 μm. The velocity amplitude U_0 is then 3.14 m s^{-1}. A protein molecule for which s is 4×10^{-13}, s is suspended in the fluid outside the bubble at a distance of 0.5 μm from its surface so that the distance r from the bubble center is 5.5 μm. According to Eq. (19), the radiation force will cause the molecule to move toward the surface with velocity U_{lim} equal to 0.62 μm s^{-1}. It thus appears that all such molecules contained in a shell of thickness 0.5 μm outside the bubble can be drawn toward it and reach its surface in less than a second. The bubble acts as a miniature centrifuge!

There are no known experiments with particle suspensions to which the above theory has been applied, in which the particles are of the size of proteins. However, the phenomenon of particle collection near vibrating bubbles has been observed and investigated with larger particles, such as platelets and erythrocytes (Nyborg and Miller, 1982; Miller, 1988). Although not yet been explicitly demonstrated, it is probable that the attraction of suspended substances to vibrating UCAs is involved in the gene transfer techniques described recently by Miller (2006), Tachibana and Tachibana (2006), and Wu (2006b).

24.3.3.5 Microstreaming Associated with Vibrating Bubbles and UCAs

It is pointed out in Section 24.4.3 that most sound fields set up in fluids give rise to acoustic streaming, that is, to time-independent circulation of the fluid. When bubbles are present in a cell suspension exposed to ultrasound, small-scale acoustic streaming (often called *microstreaming*) takes place in the immediate vicinity of each bubble and causes viscous stress to be exerted on nearby cells. Elder (1959) made detailed studies of the motion near vibrating bubbles of near-resonant size at 10 kHz. In his experiments, the bubble was situated near a solid surface. He found that the streaming pattern might take on any of four different forms, depending on the purity of the liquid and other factors. He also made determinations of the streaming speeds and, under some conditions, found order-of-magnitude agreement with approximate theory of the kind described in Section 24.4.3.

Following these findings, many studies were made on the biological implications of microstreaming associated with small gas bodies; a thorough review of available information on this topic was published by Miller (1987). To obtain this information, investigators had employed a variety of arrangements: (i) a single bubble of near-resonance size, (ii) an array of gas bodies formed by trapping air in small holes drilled in a solid surface, (iii) an array of gas bodies trapped in holes of Nuclepore membranes, and (iv) intracellular gas-filled channels in plant tissues. With these arrangements, cellular and molecular alterations were observed at relatively low values of the pressure amplitude; the effects included hemolysis, release of protein from bacteria, degradation of (high molecular weight) DNA in solution; aggregation of human platelets; release of adenosine triphosphate (ATP) from erythrocytes; and mechanical disruption of plant cells. Theory and experiment showed that the magnitude of the velocity gradient required to break a rodlike molecule, such as DNA, decreases as the length increases.

The bubble-associated streaming field is highly localized; a cell in suspension must be brought into near-contact with an active bubble in order for the stress to be effective. The attractive force discussed in Section 24.3.3.4 tends to bring about cell–bubble contacts, as does the outer part of the microstreaming flow itself. Calculations for cells in suspension indicate that in a cell–bubble ‘‘contact’’ the time interval during which the cell is subjected to high stress is short, of the order of milliseconds or less. Usually, a cell must undergo several contacts before breaks occur in the membrane, which allow cell contents to be released. However, for a vibrating gas-filled channel contained within plant tissue, the situation is very different. Here, the stresses produced by continuous vibration are applied steadily to membranes adjoining the channel, and membrane failure occurs at relatively low values of the stress.

In nearly all *in vitro* ultrasound experiments and applications involving activity of small gas bodies prior to about 1995, the gas bodies were "free" bubbles. Uncertainties arose because investigators were usually not able to specify the numbers, sizes, or positions of free bubbles present during experiments. When UCAs became commercially available, investigators were presented with new possibilities, by having bubbles of controlled size and characteristics for use in experiments and applications. As part of a recent symposium, Bouakaz and de Jong (2007) describe UCAs, listing characteristics of over a dozen types now available commercially, Blomley, Claudon, and Cosgrove (2007) discuss their applications to ultrasound imaging, Miller (2007) and Dalecki (2007) describe, respectively, information on biological effects obtained from *in vitro* and *in vivo* exposures to ultrasound in the presence of UCAs, and Nyborg (2007) discusses mechanisms for effects. There is much interest in findings that cell membranes can be altered reversibly, thus facilitating transfection, during exposure to ultrasound of low or moderate acoustic amplitude when UCAs are present. The mechanism probably includes aspects of microstreaming and radiation pressure. It is concluded by Wu (2002) from theoretical considerations that these are likely under some conditions and that under others, higher amplitudes are required, for which the mechanisms are discussed in Section 24.5. Miller (2007) reaches similar conclusions from experiments of a special kind, made possible by availability of UCAs. Here, cells to be exposed are arranged as a monolayer on Mylar film that is positioned in a suitable liquid so that the film is horizontal and the cells are below the film. UCAs are then introduced below the monolayer and allowed to rise until they are in contact with the cells, after which ultrasound is applied from below. Effects are observed later by examining the monolayer with microscopy.

24.4
Biological Effects from Nonthermal Noncavitational Mechanisms

24.4.1
Introduction

There have been numerous investigations in which changes were produced by sound in biological systems, where there was no significant temperature elevation and no indication of cavitation. Instead, the mechanism was identified as a manifestation of ARF or torque, or as acoustic streaming.

24.4.2
Acoustic Radiation Force

When sound interacts with an object, the temporal average of the force exerted on it is usually not zero and is called the *acoustic radiation force*. In Sections 24.3.3.3 and 24.3.3.4, the forces acting on a gas bubble and on a suspended particle near a vibrating gas bubble were discussed. Here, the topic is taken up somewhat more generally.

One application of ARF is to the calibration of acoustic equipment used for generating sound beams of defined characteristics. When a beam of ultrasound travels through water and impinges on an immersed object that absorbs all of the incoming energy, the radiation force F_{rad} exerted on the object is given simply by

$$F_{\text{rad}} = W/c \tag{20}$$

where W is the total acoustic power in the beam and c is the speed of sound in water. Taking c to be 1500 m s^{-1} and the power to be 1 W, one obtains 7×10^{-4} N (equivalent to the weight of 68 mg) for F_{rad}. Equation (20) is the working equation for radiation balances that are used to measure the total power in a beam (NCRP, 1986).

Equation (20) can be used to calculate the steady force F_{rad} on any object, or any portion of tissue that absorbs (and does not reflect or scatter) ultrasound, if W is equated to the power absorbed. Using this relationship, Mihran, Barnes, and Wachtel (1990) explained changes in the action potential of myelinated frog sciatic nerve generated by electric stimuli, which occurred when 500 μs pulses of 2 MHz ultrasound were focused on a small area of the nerve trunk; the threshold value of F_{rad} was determined to be 0.2μN.

In other studies, Dalecki *et al.* (1997a) applied 5 ms pulses of focused 1.2 MHz ultrasound to frog heart and found that if these were applied during the systole (contraction) phase the blood pressure in the aorta might decrease. Tests indicated that radiation force was the cause; a threshold value of about 1 mN was obtained for F_{rad}.

Many studies have been made of effects produced by exposing cell suspensions to ultrasound in which ARF is found to be involved. In the analysis of results, use is sometimes made of an equation developed by Gor'kov (1962), which applies rather generally to small particles, such as biological cells. According to this equation, the radiation force on a sphere of volume v and density ρ in a liquid of density ρ_0 is

$$F_{\text{rad}} = vD\nabla\langle KE\rangle - v(1-\theta)\nabla\langle PE\rangle \tag{21a}$$

$$D = 3(\rho_s - \rho_0)/(2\rho_s + \rho_0) \tag{21b}$$

where θ is the ratio of the compressibility of the particle to that of the surrounding medium. Here, $\langle \text{KE}\rangle$ and $\langle PE\rangle$ are the time-averaged densities of the kinetic energy and potential energy, respectively, in the ultrasound field. One important application of Eqs (21a and b) is to the force on small particles in a stationary wave. There $\nabla\langle PE\rangle$ is equal and opposite to $\nabla\langle KE\rangle$ so that F_{rad} can be written as $C\nabla\langle KE\rangle$, where C is a constant equal to $v(D+1-\theta)$. If the cell is more dense and less compressible than the surrounding fluid, as is usually true for cells in saline or in body fluids, then C is positive and the radiation force is in the direction of $\nabla\langle KE\rangle$. Since the kinetic energy density KE is proportional to the square of the velocity amplitude, this means that cells move toward velocity "antinodes," that is, to maxima of the velocity amplitude. Experiments with cell suspensions exposed to standing waves support this conclusion; cells collect in bands one-half wavelength apart, as expected.

Similar banding has been observed *in vivo*. Observations of blood vessels during application of ultrasound, under special conditions where standing waves were established, have shown that a series of evenly spaced clear regions may appear in the otherwise opaque interior of the vessel. The red cells are packed into periodically spaced bands, the distance between centers of the bands being about one-half of a wavelength. When the ultrasound is turned off, the cells typically disperse and the blood flow continues normally (Dyson, Woodward, and Pond, 1971).

An acoustic intensity of about 1 W cm^{-2} has been found to be sufficient to produce the banding in vessels where the blood is moving slowly. Such intensities are commonly produced in therapy devices and are sometimes encountered in the focal regions of beams used in Doppler techniques. The possibility of adverse effects from the banding can be avoided or reduced by minimizing the time during which the beam is held in a single location.

An application of Eqs (21a and b) to the attraction of particles by a vibrating bubble has already been discussed in Section 24.3.3.4.

A somewhat related application of Eqs (21a and b) is to the force between two particles, such as two biological cells, in a sound field. This situation is somewhat more complex than the bubble–particle interaction since the scattered field of a particle such as a biological cell (whose compressibility is much less than that of a bubble) is a combination of mono- and dipolar fields. It is found that the force between particles is a function of the angle between the particle–particle axis and the direction of the local acoustic particle velocity. If the angle is zero or π radians, the particles repel each other; if it is $\pi/2$ they attract. Because of this, particle doublets tend to form whose axes are perpendicular to the local acoustic particle velocity. Further, these doublets may combine with other doublets or singlets to form larger groups, all with the same alignment. So far this aggregation process has been observed only *in vitro* (Nyborg, 1989).

Acoustical methods have proven useful for moving or manipulating cells. Bardsley *et al.* (1990) have found that techniques involving the use of ARF are advantageous for concentrating cells in bringing about cell fusion. Wu (1991) developed theory for “acoustic tweezers” involving the radiation force field created by superposing two 3.5 MHz focused beams projected in opposite directions; he showed experimentally that, with this arrangement, latex particles of 270 μm diameter in water could be trapped at the focus. Spengler and Coakley (2003) showed that with 3.17 MHz standing waves in a chamber only a single half-wavelength in depth, latex particles 25 μm in diameter were trapped at a central plane. Although confined to a small region by the ARF, the particles were found free to combine with each other in patterns determined by van der Waals force and electrostatic interactions. Hence, it was concluded that the ultrasound trap technique was a promising one for facilitating studies of aggregation patterns. Coakley *et al.* (2004) followed up on these findings. They employed a 3.17 MHz half-wavelength standing-wave trap together with video recording to study aggregation processes involving (i) erythrocytes, with and without lectin, (ii) chondrocytes, and (iii) neural cells. Haake *et al.* (2005) report further development of methods for use of ultrasound in manipulation of cells, and provide references showing that this is an active area.

Quite a different kind of application for ARF is found in techniques now employed or proposed for determining elastic properties of tissues. An example is described by Nightingale *et al.* (2001) in which breast lesions are detected by a method ("remote palpation") analogous to manual palpation. Here, diagnostic ultrasound is used to measure displacements of tissue components produced by pulses of focused ultrasound projected into breast tissue. Other methods for determining tissue properties by use of ARF are described in a review by Parker *et al.* (2005). These techniques, included under the rubric of *elastography*, have already become an important part of medical imaging procedures.

24.4.3
Acoustic Streaming and Microstreaming

Small-scale acoustic streaming associated with vibrating gas bubbles is discussed in Section 24.3.3.5. Here, we take up the subject of streaming more generally, although only briefly. Relevant aspects of the topic have been reviewed by Nyborg (1965, 1978, 1998), Rooney (1981), and Miller (1987). When a sound field exists in a fluid, acoustic boundary layers are set up near surfaces of walls or immersed objects. Near these surfaces the tangential component $u_{\tan}$ of the acoustic particle velocity changes rapidly. At a solid wall where a nonslip condition applies, $u_{\tan}$ drops abruptly to zero in a short distance (δ) called the *acoustic boundary thickness*. At the surface of a "semisolid" object, such as a biological cell, $u_{\tan}$ is not zero, but differs considerably from its value in the fluid; again, an acoustic boundary layer of thickness δ exists near the surface in which there is a high gradient of $u_{\tan}$. The quantity δ depends on the viscosity coefficient η and density ρ_0 of the fluid, as well as the frequency f, as follows:

$$\delta = (\eta/\pi\rho_0 f)^{1/2} \tag{22}$$

For water at room temperature at a frequency of 1 MHz, the boundary-layer thickness is 0.56 μm and is thus comparable to a wavelength of visible light. It is the existence of boundary layers that is responsible for the generation of acoustic streaming near surfaces of cells or other objects when there is relative motion between the fluid and the object. In the boundary layer, the acoustic particle velocity varies rapidly with distance from the boundary and also varies, though usually more slowly, in directions parallel to the boundary. It has been shown that in the boundary layers established by a continuous sound field, the elements of fluid are subjected to steady torque; it is the action of this torque that produces the steady circulatory motion known as acoustic streaming in the outer fluid.

Theory for the acoustic streaming has been developed for a number of basic situations. We shall not attempt to summarize the theoretical findings here, but instead refer the reader to the reviews cited above. In the much-investigated topic of streaming near a fixed cylinder in an oscillating fluid (which presents essentially the same streaming situation as an oscillating cylinder in a quiescent fluid), it is found that the streaming flow is characterized by high velocity gradients in a thin boundary layer near the cylinder surface. If the cylinder diameter is not too small and the oscillatory amplitude is not too great, the acoustic streaming ("dc") boundary-layer thickness is approximately equal to δ, the "ac" boundary-layer thickness, given in Eq. (22). An approximate expression for the viscous stress S at the cylinder surface is

$$S = 2\pi \eta f \xi_0{}^2 / R_0 \delta \qquad (23)$$

where ξ_0 is the amplitude of the fluid oscillation relative to the cylinder and R_0 is the cylinder radius. Experiments have been done with vibrating tungsten wires for which it is found that time-averaged stresses of the order of 10^2 Pa or more are easily produced in the small-scale streaming (often called *microstreaming*) near the vibrator. Exposing cells and large biomolecules to the action of a vibrating wire has brought about a variety of effects, including lysis of erythrocytes, altered phagocytic index for leucocytes, clumping of platelets, and deactivation of enzymes (Rooney, 1981; Williams, 1983). More recently, Wu (2006b, 2007) found that exposing suspensions of Jurkat lymphocytes to microstreaming of this kind can cause them to become temporarily permeable to 500 kDa molecules.

Analysis of conditions in the half-wavelength traps discussed in Section 24.4.2, and related arrangements, shows that small-scale acoustic streaming plays an important role, in addition to the radiation force. Kuznetsova and Coakley (2004) describe detailed findings for a 1.5 MHz chamber and include an extensive list of references to other studies where microstreaming was involved.

Eddying motions of this kind can also be produced by ultrasound within blood vessels or in the interior of cells if the contents are sufficiently fluid. In general, microstreaming is generated in fluids adjacent to vibrating surfaces if the vibration pattern is sufficiently nonuniform. In their reviews, Williams (1983) and Miller (1987) describe experiments involving acoustic microstreaming produced in marine eggs, muscle fibers, plant cells, and blood vessels, leading to disturbances of various kinds. In the exposures of blood vessels (which were done *in vivo*), vigorous microstreaming led to thrombogenesis, apparently because of shearing action on platelets.

In a series of publications by Stavros and Dennis (1993), Nightingale *et al.* (1995), and Nightingale, Kornguth, and Trahey (1999), it is demonstrated that radiation force and acoustic streaming can be employed clinically for determining the nature of breast lesions. If the lesion under investigation is simply a fluid-filled cyst, focused ultrasound will produce flow, which can be detected with Doppler ultrasound. If no flow occurs, the lesion must be recognized as solid.

24.4.4
Other Nonthermal Noncavitational Effects

A challenge for investigators of mechanisms comes from findings that some therapeutic effects of ultrasound occur under conditions where the temperature rise is minimal, there is no evidence of cavitation activity, and the mechanism has not yet been determined. Insight into possibilities comes from earlier studies. For example, Schnitzler (1969) and Ravitz and Schnitzler (1970) did experiments in which highly localized ultrasound was applied to frog muscle fibers. A metal cone tapered to a small rounded tip with radius of curvature 20 μm acting as a point source was set into vibration at a frequency of 85 kHz with the tip in contact with a muscle fiber. Vibrating the tip with displacement amplitude of 2 μm for 1 minute produced disruption within mitochondria on the interior of an adjacent fibril, as determined by electron microscopy. (A wedge-shaped tip acting as a line source was also used.) Increasing the amplitude and time to 5 μm for 10 minutes led to severe damage. The effect was less for fibrils at a greater distance from the vibrator. In other experiments, the permeability of cells to sodium ions was reversibly increased by exposure to the localized vibration at an amplitude of 1 μm for an hour. The above results were obtained under conditions where there was little temperature rise and no indication of cavitation. It was reasonably concluded that the localized mitochondrial damage and the reversible increase in membrane permeability resulted from fields of time-averaged stress produced by the ultrasound. These stresses may have led to strain or microstreaming in the muscle fiber.

Related experiments were subsequently done by Frost (1974) and by Frost and Nyborg (1973) in which a localized ultrasound field was set up in a similar manner in a soft viscoelastic solid material, with an arrangement permitting determination of stress/strain fields. The latter was accomplished by choosing birefringent epoxy for a solid sample and using a polarizing microscope. A localized 85 kHz ultrasound field was established in the epoxy by the vibrating wedge-shaped tip of a tapered horn; this small tip, acting as a line source, was pressed against one edge of the sample by a static force F^*, which was superposed on the oscillatory force that produced the vibration. By observing the isochromatics as the vibration continued, for various values of the source amplitude, it was found that strain was generated for which the orientation was outward from the tip; that is, it was in the radial direction, like that produced by the static force F^*. Also, the magnitude of the vibration-produced strain (averaged over a cycle), like the strain produced by the static force F^*, was found to vary inversely with distance from the source. For both the F^*-produced and the vibration-produced strain, it was found that as the vibration continued the strain magnitude varied with time approximately according to a parallel spring-dashpot (Kelvin-Voigt) model for a viscoelastic medium. At low amplitudes, the cycle-averaged strain magnitude at a given point after a given time was found to be approximately proportional to the square of the vibration amplitude, but at higher amplitudes the relationship became linear, then nonlinear again.

Since it became known that localized ultrasound of low or moderate level can produce significant changes in viscoelastic media, including biological tissue, it was natural to explore possibilities for the broader beams of ultrasound that were commonly available. Pilla *et al.* (1990) directed a 1.5 MHz ultrasound beam, about 4 cm^2 in cross-sectional area, at the hind limbs of rabbits in each of which the tibular shaft had been cut. The ultrasound was applied in pulses, each of 200 μs duration, delivered at the rate of 1000 s^{-1}; the duty factor was thus 0.2. The intensity during each pulse, averaged over the 4 cm^2 area, was 150 mW cm^{-2}; the time-averaged (and spatially averaged) intensity was therefore 30 mW cm^{-2}. Tests made at various times after the treatment showed that, although the ultrasound did not affect appreciably the final mechanical strength, the healing process occurred more quickly in the treated fibula than in the sham-treated controls.

Since the 1990 study many others have been done, with similar findings; some of these studies were with laboratory animals and some were done as clinical trials (NCRP, 2002, Chapter 3). Here the conditions used were the same *low intensity pulsed ultrasound* (LIPUS) conditions as were employed by Pilla *et al.* (1990). In a clinical trial, Heckman *et al.* (1994) treated fractures of 33 patients and found that the time for (clinical and radiographic) healing was 86 ± 6 days for patients treated with ultrasound, and was 114 ± 10 days for controls. In an animal study, Cook *et al.* (2001) evaluated the repair of osteochondral defects (3.2 mm diameter and 5.0 mm deep) in the patellar groove of 106 rabbits; histological tests showed that daily 20-minute applications of ultrasound caused continuous improvement with most of the benefit occurring in the first 12 weeks.

Parvizi *et al.* (2002) considered mechanisms for the healing action of LIPUS, noting, for analogy, that clinicians commonly apply mechanical force to the skeleton in various ways when treating skeletal disorders. Also, chondrocytes, known to be critical in the response to fractures, have been shown to respond to mechanical forces. These authors carried out *in vitro* experiments in which cultured rat chondrocytes were exposed to LIPUS, finding that the internal concentration of calcium was significantly increased by the exposure. They concluded that it was signaling from this increase in calcium concentration that led to the increased synthesis of proteoglycan (critical for fracture repair) by chondrocytes, which had been observed in previous experiments.

Zhou *et al.* (2004) found that exposures to LIPUS, from 6 to 11 minutes in duration, of cultured fibroblasts from human foreskin, resulted in an increased number of the cells. They also obtained evidence that the ultrasound promotes the cell proliferation via activation of integrin; in this they support findings of Schmidt *et al.* (1988) that integrin receptors at the cell surface mediate the response of a cell to mechanical stress.

Much has now been learned about the effects exposure to LIPUS produces on cells that are important to the healing process. It is expected that these effects come about because of mechanical stress applied by the ultrasound. However, the specific nature of the stress is still unknown. The nonthermal noncavitational mechanisms discussed in Sections 24.4.2 and 24.4.3 are obvious candidates. For any of these, it must be recalled that under LIPUS conditions, the ultrasound is delivered in pulses about 200 μs in duration delivered periodically at a frequency of about 1 kHz.

Acoustic streaming, probably of small scale, would be a possible source of stress if there is fluid present in the tissue region exposed.

For any medium, radiation force might be involved in any of various ways. Radiation force might be delivered by a beam passing through an absorbing medium; the force per unit volume (cf. Eqs (3) and (20)) would be $2\alpha I/c$, where α is the absorption coefficient for the medium, I is the intensity and c the speed of sound. In another situation, the force given by Eqs (21a and b), might be exerted on a single particle in any sound field for which the potential and kinetic energy densities are known. In still another situation (for which the equations are given elsewhere (NCRP, 2002, Section 3; Nyborg, 1989), the force is between two small particles in a sound beam, as discussed briefly in

Section 24.4.2; here, the force is attractive if the line connecting the particles is perpendicular to the beam direction, and is repulsive if the two directions are parallel.

24.5
Nonlinear Cavitation Activity

24.5.1
Introduction

In Sections 24.3.2 and 24.3.3, the gas bubble is considered to vibrate in a volume mode, maintaining a spherical or quasispherical shape during its oscillation. Much valuable information is obtained from theory for a spherical bubble. However, in aqueous media exposed to ultrasound in the lower-frequency range (e.g. below 1 MHz) at pressure amplitudes in common use, free bubbles seldom remain spherical. Even at modest amplitudes, surface waves develop on the bubbles and the associated microstreaming (see Section 24.3.3.5) is altered. At higher amplitudes, the waves become highly distorted and microbubbles are entrapped, as when “white caps” appear on a lake or ocean during a windy day. As a result of these processes, the original free bubble may fragment and disappear. Surface waves may also develop on UCAs when in an ultrasound field of high amplitude; these may lead to fragmentation of the UCA shell, and release of the enclosed gas as free bubbles. Although the activity of a free gas bubble may not continue long in an ultrasound field of high amplitude, it may be of a violent nature which is destructive or, if controlled, is useful in various applications. Much attention has been given to the analysis of bubble dynamics under high amplitude conditions. In the following section, some relevant theories are briefly discussed.

24.5.2
Bubble Dynamics Under High Amplitude Conditions

In Section 24.3.2.1, an equation of motion, Eq. (7), is given for a gas bubble that applies when the oscillation amplitude is small. Small-amplitude theory, based on solutions of this equation, is very useful in that it leads to convenient analytical expressions. For most of the phenomena discussed in Section 24.3.3, such expressions have been developed and have proven sufficient for many purposes. However, the basic equations that govern bubble motion are nonlinear, and at high amplitudes the consequences of this nonlinearity cannot be avoided. Reviews dealing with nonlinear equations for cavitation and their solutions have been published by Flynn (1964), Neppiras (1980), Apfel (1981), Young (1989, 2005) and Leighton (1994, 1996). As given by Neppiras (1980), one form of the governing equation is (using the dot to represent differentiation with respect to time)

$$R\ddot{R} + (3/2)\dot{R}^2 = (1/\rho_0)\{P_{\text{eff}}(R_0/R)^{3\gamma} - 2\sigma/R - [P_a - p_0\sin(\omega t)]\} \tag{24}$$

where R is the instantaneous bubble radius, P_{eff} is equal to $P_a + 2\sigma/R_0$, and σ is the surface tension at the gas–liquid interface; γ, R_0, P_a, and p_0 have the same meanings as in Section 24.3.2. Equation (24) applies to a sphere containing a perfect gas surrounded by inviscid liquid in which the acoustic pressure is p_0

$\sin(\omega t)$. Equation (24) reduces to Eq. (7) (neglecting damping and surface tension) when linearizing approximations are made, namely, neglecting the term $(3/2)\dot{R}^2$, replacing R with $R_0 + \xi$, and assuming the radial displacement ξ is much less than R_0.

Except at small values of the pressure amplitude p_0, solutions of Eq. (24) have quite a different character from the sinusoidal functions of time that satisfy Eq. (7). As the applied pressure amplitude p_0 is increased, the bubble response may dramatically change. If p_0 becomes great enough to cause enough bubble growth during an expansion phase so that the maximum value of R exceeds R_0 by a critical factor of about 2 or more, the conditions for *inertial* (formerly called *transient*) *cavitation* will be met. Holland and Apfel (1990) arrived at a practical criterion for diagnostic ultrasound. Under these conditions, the inward momentum of the liquid, acquired during the contraction phase, causes the bubble to collapse. During a collapse event, the inward speed of the bubble surface can become supersonic and give rise to shock waves, both within the bubble and in the surrounding liquid. After a collapse, the bubble may rebound and undergo further cycles of expansion and contraction, or it may fragment because of the surface phenomena referred to earlier (Section 24.5.1). During a very short time interval, when the radius is near its minimum, the pressure in the bubble can reach hundreds of megapascals and the temperature thousands of kelvins. Experimental evidence for the existence of high temperatures and their implications for biological effects of ultrasound are discussed in Sections 24.5.3 and 24.5.4.

Curves of radius versus time can be complex and critically dependent on parameters of the bubble and the sound field. In a sound field of frequency f, the curves can exhibit harmonics of frequency $2f$, $3f$, and so on, even for modest values of the pressure amplitude. At somewhat higher levels, the motion becomes aperiodic; subharmonics of frequency $f/2$, $f/3$, and so on, occur as do ultraharmonics at frequencies $3f/2$, and so on. Under some conditions, the solution becomes indeterminate, allowing multiple values for the radius. Such indeterminacy is an aspect of *deterministic chaos*, a development in theory that is generally relevant to nonlinear systems; the investigations by Lauterborn and associates on the implications of chaos theory for bubble behavior have been reviewed by Young (1989).

24.5.3
Effects of Cavitation in Biological Suspensions

Many experiments have been done in which ultrasound is applied to suspensions of cells or macromolecules and evidence is found for a change of some kind. The subject has been reviewed by El'piner (1964), the NCRP (2002), Miller (1987), Williams (1983), ter Haar (2004), and Miller (2007). In these *in vitro* experiments the mechanism for observed effects is almost always some form of cavitation, involving either free bubbles or UCAs. Ultrasound is employed widely for sterilization and for applications where it is desired to release cell contents. Following the advent of UCAs, which allow much better control of the cavitation process, there has been increasing interest in *sonoporation*, a technique that makes cells temporarily permeable, thus facilitating transfection (Wu, 2006b).

In special situations, when controlled bubble populations and low pressure amplitudes are used, the effects of bubble activity can be ascribed to microstreaming and related phenomena, which are discussed in Sections 24.3.3.4 and 24.3.3.5. At the higher amplitudes, which are commonly used, the action is much more complex, as described in Section 24.5.2.

24.5.4
Effects of Cavitation in Tissues

Organized tissues present contrasts to typical *in vitro* systems in the nature and extent of cavitation activity that occurs in them during exposure to ultrasound. Cell culture media are usually well supplied with freely mobile cavitation nuclei, whereas the same is not true in tissues. Evidence for cavitation nuclei in normal body fluids has not been reported, and the gaseous bodies that exist in some nonfluid tissues are restricted in their translational motion.

Much information on the response to ultrasound of tissue containing gas-filled spaces has been obtained by experiments with plants and insects. The subject of cavitation in organized tissues has been reviewed by Carstensen (1987), Miller (1987), and Frizzell (1988). Certain plant tissues make convenient materials for study because they are relatively transparent, contain gas-filled intracellular channels, and can be viewed with an optical microscope while they are being exposed to ultrasound. Some insects have proved to be attractive animal models, since gas-filled tracheae and branching tubes permeate the tissue.

When a leaf of the water plant *Elodea* is viewed during application of ultrasound, a great deal of motion is seen within cells (in the large vacuoles that these cells possess), especially cells near the larger gas-filled channels. This motion is evidently acoustic microstreaming, brought about by the vibration of cell walls that accompanies volume oscillation of the gas in the intracellular channels. At low values of the pressure amplitude the motion is slow, but at large amplitudes it becomes vigorous enough to cause cell death, probably by rupture of the plasma membrane. Miller (1987) presents evidence that the rupture is caused by shearing stress associated with the microstreaming. From experiments at different frequencies, resonances are observed, approximately as expected from calculations of resonance frequencies for the gas-filled channels.

In experiments with larvae of the fruit fly *Drosophila,* it was found by Carstensen (1987) that larvae are killed by exposure to pulsed ultrasound if the incident intensity during the pulse exceeds about 10 W cm^{-2}, corresponding (for a plane traveling wave in water) to a pressure amplitude of about 0.6 MPa. This was found true even with pulse repetition rates and pulse durations approaching those used in diagnostic ultrasound. From studies with fertilized *Drosophila* eggs, it was found that the sensitivity of the organism increased greatly at the stage of development when the trachea and tubules filled with air. These results have implications for safety of pulsed diagnostic ultrasound when it is applied to tissues containing small air spaces.

In mammalian tissues the lung alveoli provide an abundant source of small gas-filled spaces; effects of acoustic waves on the lung are discussed later. In other tissues, there is little evidence for cavitation activity produced by ultrasound, unless the amplitude is high *or unless bubbles have been inserted into the tissue*. Since UCAs became available, the latter possibility has become feasible as part of common practice; this changes the potential significantly for cavitation activity. Dalecki (2007) has reviewed existing information on bioeffects produced by ultrasound in tissues containing blood into which UCAs have been injected. These bioeffects include hemolysis (Dalecki *et al.*, 1997b), effects on the vasculature (Miller and Gies, 2000) and effects on the heart (Dalecki, *et al.*, 1997a). Besides deleterious effects, UCAs show promise for enhancing the dissolution of clots *in vivo* (Birnbaum *et al.*, 1998).

An intriguing finding, with recognized implications, is that phagocytic cells will engulf UCAs and that the latter will respond actively to ultrasound even after being phagocytosed (Dayton *et al.*, 2001). This phenomenon has added new capabilities to medical imaging (Forsberg *et al.*, 1999), and also provides excellent capabilities for research on UCA–cell interactions (Miller and Quddus, 2002).

Some of the effects obtained in tissues when UCAs are present occur at modest ultrasound levels under which noninertial forms of cavitation are involved, while others occur at higher levels and require the more violent activity characteristic of inertial cavitation. In their review, Hynynen and McDannold (2006) describe medical procedures currently used or considered, involving HIFU, in which cavitation sometimes plays an important part. Particularly noteworthy was the demonstration that, if UCAs are present, the blood–brain barrier can be opened by the use of HIFU, so that drugs can be administered as needed. Also, when UCAs are used, it is found that the desired temperature can be reached with the use of lower intensities; rationale for this was developed by Holt and Roy (2001). In another application, Vaezy *et al.* (1999) found that HIFU can be used to stop active bleeding from vascular injuries; while heating is the primary mechanism, useful roles are also played by acoustic streaming and cavitation.

If medical applications of ultrasound were ranked in order of the amplitudes employed, acoustic lithotripsy would be at the top of the list. An acoustic lithotripter is a device for destroying stones in the kidney or gallbladder by focused ultrasound of very high amplitude.

Coleman and Saunders (1989) showed that extracorporeal shock lithotripters generate short pulses (“shocks”) in water in which peak positive pressures extend up about 100 MPa and peak negative pressures to 10 MPa. These pulses are less than a microsecond in duration and, in clinical practice, are delivered at pulse repetition frequencies up to 40 Hz. In a surgical procedure, it is typical to apply 1000 shocks, or more. Cavitation appears to be one of the principal means of action for lithotripter pulses, but there are many questions to be answered about when and where the action will occur. Recalling results discussed above for plant and insect tissue, it would seem reasonable that the mammalian lung, whose function requires air-filled spaces, would respond to ultrasound in a special way, as has been proven to be true. Delius *et al.* (1987) carried out experiments in which lithotripter pulses were applied *in vivo* to canine gallbladder, in which human gallstones had been implanted. Besides obtaining 90% success in destruction of the gallstones, they found lung hemorrhages in 80% of the animals.

Electron microscopy of hemorrhaged tissue revealed small breaks in the walls of blood vessels. Further information on lung damage was obtained by Hartman *et al.* (1990) in experiments where adult and fetal mouse lung were exposed *in vivo* to pulses from a lithotriptor. In adult, lung hemorrhage was produced when the acoustic pressure in the pulse was 2 MPa or more; however, in fetal lung, which evidently contains little or no air, there was no observable damage, even for acoustic pressures as high as 20 MPa. Further biophysical information on applications of acoustic lithotripsy may be found in NCRP (2002).

24.6
Discussion

Since the first experiments with ultrasound (and intense audible sound), much has been learned about its potential for altering biological structures and processes, and about its mechanisms for producing effects. Thus, useful information has been obtained and made available to users of ultrasound in diagnostic medicine so that harmful thermal effects can be minimized. At the same time, advances have been made in techniques for delivering heat in a prescribed manner to sites where it is needed. New methods, including the use of MRI, enable the determination and control of temperatures in therapeutic procedures much more accurately than before.

The role of small gas bodies in biomedical applications of ultrasound has undergone a transformation since UCAs became commercially available. Investigators are now able to have microbubbles of controlled size and characteristics at their disposal. Among the promising applications that these make possible (besides their intended use in improving ultrasound image contrast) is in sonoporation, that is, controlled modification of cell membrane permeability, in order to facilitate drug transfer.

In another advance, LIPUS procedures have become well established and are increasingly used for the treatment of bone fractures. Since the intensity levels used are modest, LIPUS is a form of ‘‘soft’’ ultrasound therapy for which the mechanisms are yet to be learned. Neither heat nor cavitation appears to be important. Serious consideration is being given to the interaction of ARFs, modulated at an audible frequency, with chondrocytes and other structures that are important to the healing process and are known to be stimulated by mechanical stress.

From these examples, it is seen that much has been learned about biological effects that ultrasound can produce; however, important questions remain. Continued research into the mechanisms, and further development of applications, show promise of being very rewarding.

Glossary

Acoustic radiation force: time-averaged force exerted by a sound field.

Acoustic streaming: time-averaged flow produced by a sound field.

Bubble: a small body of gas surrounded by liquid.

Cavitation: the mechanical response of one or more bubbles in a sound field.

Free bubble: a bubble with no film or shell at its surface.

LIPUS: acronym for low-intensity ultrasound.

UCA: acronym for ultrasound contrast agent.

Ultrasound contrast agent: a bubble designed for use with diagnostic ultrasound; it has a thin film or shell at its surface.

References

AIUM/NEMA (1992) *Standard for Real/Time Display of Thermal and Mechanical Acoustic Output Indices on Diagnostic Ultrasound Equipment*, Laurel, Maryland, American Institute of Ultrasound in Medicine.

Apfel, R. (1981) in *Methods of Experimental Physics* (P.D. Edmonds), Vol. 19, Chapter 7, Ultrasonics, Academic Press, New York.

Atchley, A. and Crum, L.A. (1988) in *Ultrasound: Its Chemical, Physical and Biological Effects* (K.S. Suslick), Chapter 1, New York, VCH Publishers.

Bardsley, D.W., Liddell, J.E., Coakley, W.T., and Clarke, D.J. (1990) *J. Immunol. Methods*, **129**, 41–47.

Birnbaum, Y., Luo, H., and Nagai, T., *et al.* (1998) *Circulation*, **97**, 130–34.

Bleeker, H.J., Shung, K.K., and Barnhart, J.L. (1990) *J. Acoust. Soc. Am.*, **87**, 1792–97.

Blomley, M., Claudon, M., and Cosgrove, D. (2007) *Ultrasound Med. Biol.*, **33**, 180–86.

Bouakaz, A. and de Jong, N. (2007) *J. Ultrasound Med.*, **33**, 187–96.

Carstensen, E.L. (1987) *Ultrasound Med. Biol.*, **13**, 597–606.

Cavicchi, T.J. and O'Brien, W.D. (1984) *J. Acoust. Soc. Am.*, **76**, 1244–45.

Clement, G.T., McDannold, N.J., Hynynen, K. (eds) (2006) *Therapeutic Ultrasound. 5th International Symposium on Therapeutic Ultrasound*. AIP Conference Proceedings, Vol. 829, AIP, Melville, New York.

Coakley, W.T. and Nyborg, W.L. (1978), in *Ultrasound: Its Applications in Medicine and Biology* (F.J. Fry), Part I, Chapter 2, Elsevier, New York.

Coakley, W.T., Bazou, D., Morgan, J., Foster, G.A., Archer, C.W., Powell, K., Borthwick, K.A.J., Twomey, C., and Bishop, J. (2004) *Colloids Surf. B Biointerfaces*, **34**, 221–30.

Coleman, A.J. and Saunders, J.E. (1989) *Ultrasound Med. Biol.*, **15**, 213–27.

Cook,S.D., Samantha,L.S., Popich-Patron,L.S., Ryaby, J.P., Deryk, G.J., and Barrack, R.L. (2001) *Clin. Orthop. Relat. Res.*, **391S**, s231–s43.

Crum, L.A. (1984) *Ultrasonics*, **2**, 215–23.

Dalecki, D. (2007) *Ultrasound Med. Biol.*, **33**, 205–13.

Dalecki, D., Raeman, C.H., Child, S.Z., and Carstensen, E.L. (1997a) *Ultrasound Med. Biol.*, **23**, 275–85.

Dalecki, D., Raeman,C.H., Child,S.Z., Cox, C., Francis,C.W., Meltzer,R.S., and Carstensen, E.L. (1997b) *Ultrasound Med. Biol.*, **23**, 307–13.

Dayton, P.A., Allen, J.S., and Ferrara, K.W. (2002) *J. Acoust. Soc. Am.*, **112**, 2183–92.

Dayton, P.A., Chomas, J.E., Lum, A.F.H., Allen, J.S., Lindner, J.R., Simon, S.I., and Ferrara, K.W. (2001) *Biophys. J.*, **80**, 1547–56.

Dayton, P.A., Morgan, K.E., Klibanov, A.L., Brandenburger, G., Nightingale, K.R., and Ferrara, K.W. (1997) *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, **44**, 1264–77.

Delius, M., Enders, G., Heine, G., Stark, J., Remberger, K., and Brendel, W. (1987) *Ultrasound Med. Biol.*, **13**, 61–67.

Duck, F.A. (1990) *Physical Properties of Tissue*, Academic Press, London.

Dunn, F., O'Brien, W.D. Jr (eds) (1976) *Ultrasound Biophysics*, Benchmark Papers in Acoustics, Vol. 7, Dowden, Hutchinson and Ross, Stroudsburg, PA.

Dyson, M., Woodward, B., and Pond, J.B. (1971) *Nature (London)*, **232**, 572–73.

Elder, S.A. (1959) *J. Acoust. Soc. Am.*, **31**, 54–64.

El'piner, I.E. (1964) *Ultrasound: Physical, Chemical and Biological Effects*, Consultants Bureau, New York.

Flynn, H.G. (1964) in *Physical Acoustics* (W.P. Mason), Vol. **1B**, Chapter 9, Academic Press, New York.

Forsberg, F., Goldberg, B.B., Liu, J.B., Merton, D.A., Rawool, N.M., and Shi, W.T. (1999) *Radiology*, **210**, 125–32.

Frizzell, L.A. (1988) in *Ultrasound: Its Chemical, Physical and Biological Effects* (K.S. Suslick), Chapter 7, VCH Publishers, New York.

Frost, H.M. (1974) PhD Thesis, University of Vermont.

Frost, H.M. and Nyborg W.L. (1973) *Ultrasonics International 1973 Conference Proceedings*, pp. 81–88 .

Gor'kov, L.P. (1962) *Sov. Phys. Dokl.*, **6**, 773–75.

Gramiak, R., Shah, P.M., and Kramer, D.H. (1969) *Radiology*, **92**, 939–48.

Haake, A., Neild, A., Kim, D.H., Ihm, J.E., Sun, Y., Dual, J., and Ju, B. (2005) *Ultrasound Med. Biol.*, **31**, 857–64.

Hartman, C., Child,S.Z., Mayer, R., Schenk, E., and Carstensen, E.L. (1990) *Ultrasound Med. Biol.*, **16**, 675–79.

Heckman,J.D., Ryaby,J.P., McCabe, J., Frey, J.J., and Kilcoyne, R.F. (1994) *J. Bone Joint Surg. Am.*, **76**, 26–34.

Holland, C.K. and Apfel, R.E. (1990) *J. Acoust. Soc. Am.*, **88**, 2059–69.

Holt, R.G. and Roy, R.A. (2001) *Ultrasound Med. Biol.*, **27**, 1399–412.

Hsieh, D.Y. and Plesset, M.S. (1961) *J. Acoust. Soc. Am.*, **33**, 206–15.

Hynynen, K. and McDannold, N. (2006), in *Emerging Therapeutic Ultrasound* (eds J. Wu, W.L. Nyborg), Chapter 5, World Scientific Co. Pte. Ltd, Singapore.

McNaught, A.D. and Wilkenson, A. (1997) *IUPAC Compendium of Chemical Terminology*, The Gold Book, 2nd edn., Blackwell Science .

Kuznetsova, L.A. and Coakley, W.T. (2004) *J. Acoust. Soc. Am.*, **116**, 1956–66.

Lehmann, J.F., de Lateur, B.J. (1990) in *Therapeutic Heat and Cold*, 4th edn (J.F. Lehmann), Williams & Wilkins, Baltimore.

Leighton, T.G. (1994, 1996) *The Acoustic Bubble*, Academic Press, London.

Marsh, J.A., Hall, C.S., Hughes, M.S., Mobley, J., and Miller, J.G. (1997) *J. Acoust. Soc. Am.*, **101**, 1155–61.

McDiarmid, T., Ziskin, M.C., Michlovitz, S.L. (1996), in *Thermal Agents in Rehabilitation*, 3rd edn (S.L. Michlovits), F.A. Davis Company, Philadelphia.

Mihran, R.T., Barnes, F.S., and Wachtel, H. (1990) *Ultrasound Med. Biol.*, **16**, 297–309.

Miller, D.L. (1987) *Ultrasound Med. Biol.*, **13**, 443–70.

Miller, D.L. (1988) *J. Acoust. Soc. Am.*, **84**, 1378–87.

Miller, D.L. (2006), in *Emerging Therapeutic Ultrasound* (eds J. Wu, W.L. Nyborg), Chapter 3, World Scientific Co. Pte. Ltd, Singapore.

Miller, D.L. (2007) *Ultrasound Med. Biol.*, **33**, 197–204.

Miller, D.L. and Gies, R.A. (2000) *Ultrasound Med. Biol.*, **26**, 307–13.

Miller, M.W., Miller, H.E., and Church, C.C. (2005) *J. Therm. Biol.*, **30**, 400–9.

Miller, D.L. and Quddus, L. (2002) *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, **49**, 1094–102.

Miller, M.W. and Ziskin, M.C. (1989) *Ultrasound Med. Biol.*, **15**, 707–22.

NCRP (1986) *Biological Effects of Ultrasound: Mechanisms and Clinical Implications*. Report No. 74. National Council of Radiation Protection and Measurements, Bethesda, MD.

NCRP (1992) *Exposure Criteria for Medical Diagnostic Ultrasound: I. Criteria Based on Thermal Mechanisms*. Report No. 113. National Council on Radiation Protection and Measurements, Bethesda, MD.

NCRP (2002) *Exposure Criteria for Medical Diagnostic Ultrasound: II. Criteria Based on All Known Mechanisms*. Report No. 140. National Council on Radiation Protection and Measurements, Bethesda, MD.

Neppiras, E.A. (1980) Acoustic cavitation. *Phys. Rep.*, **61**, 159–251.

Nightingale, K.R., Kornguth, P.J., and Trahey, G.E. (1999) *Ultrasound Med. Biol.*, **25**, 75–87.

Nightingale,K.R., Kornguth,P.J., Walker,W.F., McDermott, B.A., and Trahey, G.E. (1995) *Ultrasound Med. Biol.*, **21**, 745–51.

Nightingale, K.R., Palmeri, E.L., Nightingale, R.W., and Trahey, G.E. (2001) *J. Acoust. Soc. Am.*, **110**, 625–34.

Nyborg, W.L. (1965), in *Physical Acoustics* (W.P. Mason), Vol. **2B**, Chapter 11, Academic Press, New York.

Nyborg, W.L. (1978) in *Ultrasound: Its Applications in Medicine and Biology* (F.J. Fry), Part I, Chapter 1, Elsevier, New York.

Nyborg, W.L. (1981) *J. Acoust. Soc. Am.*, **70**, 310–12.

Nyborg, W.L. (1989) *Ultrasound Med. Biol.*, **15**, 93–99.

Nyborg, W.L. (1991) Biological effects of sound and ultrasound, in *Encyclopedia of Applied*

Physics (G.L. Trigg), Vol. 2, VCH Publishers, New York, pp. 403–20.

Nyborg, W.L. (1998) in *Nonlinear Acoustics* (eds M.F. Hamilton and D.T. Blackstock), Chapter 7, Academic Press, San Diego.

Nyborg, W.L. (1999) Biological effects of sound and ultrasound, in *Encyclopedia of Applied Physics*, Update 2, WILEY-VCH Verlag GmbH, pp. 371–77.

Nyborg, W.L. (2007) *Ultrasound Med. Biol.*, **33**, 224–32.

Nyborg, W.L. and Miller, D.L. (1982) *Appl. Sci. Res.*, **38**, 17–24.

Parker, K.J., Taylor, L.S., Gracewski, S., and Rubens, D.J. (2005) *J. Acoust. Soc. Am.*, **117**, 2705–12.

Parvizi, J., Parpura, V., Greenleaf, J.F., and Bolander, M.E. (2002) *J. Orthop. Res.*, **20**, 51–57.

Pennes, H.H. (1948) *J. Appl. Physiol.*, **1**, 93–122.

Pierce, A.D. (1994) *Acoustics: An Introduction to its Physical Principles and Applications*, Acoustical Society of America, Woodbury, New York.

Pilla,A.A., Mont,M.A., Nasser,P.R., Khan,S.A., Figueiredo, M., Kaufman, J.J., and Siffert, R.S. (1990) *J. Orthop. Trauma*, **4**, 246–53.

Ravitz, M.J. and Schnitzler, R.M. (1970) *Exp. Cell Res.*, **60**, 78–85.

Rooney, J.A. (1981) in *Methods of Experimental Physics* (P.D. Edmonds), Vol. 19, Chapter 6, Ultrasound, Academic Press, New York.

Schmidt, C., Pommerenke, H., Dürr, F., Nebe, B., and Rychly, J. (1988) *J. Biol. Chem.*, **273**, 5081–85.

Schnitzler, R.M. (1969), PhD Thesis, University of Vermont.

Spengler, J.F. and Coakley, W.T. (2003) *Langmuir*, **19**, 3635–42.

Stavros, A.T., Dennis, M.A. (1993) in *Percutaneous Breast Biopsy* (eds Parker. S.H., Jobe., W.E.), Raven Press, Ltd, New York, pp. 111–15.

Tachibana, K. and Tachibana, S. (2006), in *Emerging Therapeutic Ultrasound* (eds J. Wu., W.L. Nyborg), Chapter 4, World Scientific Co. Pte. Ltd, Singapore.

ter Haar, G.R. (2004), in *Physical Principles of Medical Ultrasonics* (eds C.R. Hill, J.C. Bamber, G.R. Haar), Chapter 13, John Wiley & Sons, Ltd, Hoboken.

Vaezy, S., Marti, R., Mourad, P., and Crum, L. (1999) *Eur. J. Ultrasound*, **9**, 79–87.

Wells, P.N.T. (1977) *Biomedical Ultrasonics*, Academic Press, London.

Williams, A.R. (1983) *Ultrasound: Biological Effects and Potential Hazards*, Academic Press, New York.

Wu, F. (2006a), in *Emerging Therapeutic Ultrasound* (eds J. Wu, W.L. Nyborg), Chapter 8, World Scientific Co. Pte. Ltd, Singapore.

Wu, J. (1991) *J. Acoust. Soc. Am.*, **89**, 2140–43.

Wu, J. (2002) *J. Acoust. Soc. Am.*, **111**, 1460–64.

Wu, J. (2007), *Shear Stress in cells generated by ultrasound*, in: Progress in Biophysics and Molecular Biology **93**, 363–373, Elsevier, Amsterdam.

Wu, J. (2006b), in *Emerging Therapeutic Ultrasound* (eds J. Wu. and W.L. Nyborg), Chapter 6, World Scientific Co. Pte. Ltd, Singapore.

Young, F.R. (1989) *Cavitation*, McGraw-Hill, London.

Young, F.R. (2005) *Sonoluminescence*, CRC Press, London.

Zhou, S., Schmelz, A., Scufferlein, T., Li, Y., Zhao, J., and Bachem, M.G. (2004) *J. Biol. Chem.*, **279**, 54463–69.

Ziskin, M.C., Bonakdarpour, A., Weinstein, D.P., and Lynch, P.R. (1972) *Invest. Radiol.*, **7**, 500–5.

Further Reading

Crum, L.A., Daniels, S., ter Haar, G.R., and Dyson, M. (1987) *Ultrasound Med. Biol.*, **13**, 541–54.

Daniels, S., Blondel, D., Crum, L.A., ter Haar, G.R., and Dyson, M. (1987) *Ultrasound Med. Biol.*, **13**, 527–39.

Hill, C.R., Bamber, J.C., and ter Haar, G.R. (2004) *Physical Principles of Medical Ultrasonics*, John Wiley & Sons, Ltd, Chichester.

Miller, D.L. (1981) *Ultrasonics*, **19**, 217–24.

Nyborg, W.L. (1984) *J. Acoust. Soc. Am.*, **80**, 1133–39.

Nyborg, W.L. (2006), in *Emerging Therapeutic Ultrasound* (eds J. Wu and W.L. Nyborg), Chapter 2, World Scientific Co. Pte. Ltd, Singapore.

ter Haar, G.R., Daniels, S., and Morton, K. (1986) *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, **UFFC-33**, 162–64.

ter Haar, G.R., Duck, F., Starritt, H., and Daniels, S. (1989) *Phys. Med. Biol.*, **34**, 1533–42.

Ziskin, M.C., Michlovitz, S.L. (1986), in *Thermal Agents in Rehabilitation* (eds S.Z. Michlovitz and S.L. Wolfe), Chapter 7, F.A. Davis Company, Philadelphia.

BIOMAGNETISM AND GEOMAGNETIC FIELD DETECTION BY ORGANISMS

RICHARD P. BLAKEMORE, *Department of Microbiology, University of New Hampshire, Durham, New Hampshire, U.S.A.*

Introduction 421
1. **Geomagnetic Parameters as Sensory Cues** 424
2. **Biological Adaptations and Geomagnetism** 425
2.1 Biogenic Magnetic Substances. 426
2.2 The Bacterial Paradigm; A Biomagnetic Needle 427
2.3 The Quest for Biomagnetic Needles 428
3. **Some Examples of Biological Interactions with the Geomagnetic Field** 430
3.1 Prokaryotes 430
3.2 Eukaryotic Microbes 435
3.3 Mollusks 435
3.4 Honeybees 435
3.5 Amphibians 437
3.6 Birds 437
3.7 Primitive Fishes 439
3.8 Bony Fishes 439
3.9 Mammals 440
4. **What is the Status of Biological Magnetic Sensing?** 441
4.1 Biological Compasses 441
4.2 Magnetic Maps 442
4.3 Magnetism-Based Clocks 443
5. **The Ongoing Study of Biological Geomagnetic Detectors** 443
Acknowledgments 443
Glossary 444
Works Cited 445
Further Reading 450

INTRODUCTION

Life on Earth has, since its inception some 3.5 billion years ago, been constantly bathed in geomagnetic information of a reasonably steady and uniform nature. Speculations have existed since the mid-1800s that organisms might use geomagnetism in orientation or navigation. Thus, it is surprising that scientific investigation of a possible geomagnetic basis for the long-recognized, but inexplicable, migratory and homing feats by some animals (e.g., a "magnetic sense") did not occur until the 1940s (Yeagley, 1951). By the late 1960s, however, students of animal behavior had accumulated data consistent with this possibility (see Yeagley, 1951; Barnothy, 1964; Brown, 1962; Brown *et al.*, 1960; Brown *et al.*, 1964; Merkel and Wiltschko, 1965).

Many early investigations addressing geomagnetic responsiveness of organisms were suggestive and not universally accepted. Some were conducted with very strong (10× earth strength, or more) inhomogeneous fields produced with bar magnets and super imposed on the local magnetic environment (Brown, 1962; Yeagley, 1947, 1951; Keeton, 1971; Southern, 1972a). Others were correlative studies (Southern, 1972b) or involved attempts to reduce or null the geomagnetic field such as with metal cages (Merkel and Wiltschko, 1965). Although their reports comprise much of the seminal literature describing magnetic influences on animal behavior, the effects noted by these and other early workers either had little relevance to the strength or configuration of the geomagnetic field or were not reliably reproduced—in some cases even by the same experimenters.

That animals perceive and respond to geomagnetism was probably first universally accepted around 1970, largely as a result of the studies of Lindauer and Martin (1972) with honeybees and those of Keeton (1974) (see Larkin and Keeton, 1976) with homing pigeons. Certainly by 1975 the biology-geomagnetism connection had been convincingly made even though a mechanistic basis for animal orientation or navigation using the "weak" geomagnetic field remained a com-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

plete mystery. One possibility, *that organisms might respond to or sense the geomagnetic field by virtue of being permanently magnetized or through magnetic tissues* (biomagnetism), although obvious in retrospect, was not seriously considered until after 1975. Keeton (1972), for instance, in reviewing pigeon homing, summarized several hypothetical models of bird orientation. Included were mechanisms based upon magnetically induced electrical fields (Faraday effect), accumulation of Coulombic electrostatic charges by flying birds, piezoelectric properties of feathers, nuclear magnetic resonance, and mechanical displacement as with accelerometers (otoliths or hair cells) but with no mention of permanently magnetic materials *per se.*

It was into this setting that the magnetotactic bacteria (Blakemore, 1975) made an unexpected entry. As a microbiology graduate student at the University of Massachusetts, Amherst, the author was microscopically examining muds which had been collected and brought from Woods Hole, MA, by his major advisor, Professor E. Canale-Parola. He observed swimming bacteria constantly moving with a bias toward one side of the mud drops being examined (Fig. 1). These bacteria reoriented, making arcs or U-turns, all swimming in the same (magnetically oriented) direction when a small magnet was brought near. Upon examining these "north-swimming organisms" by electron microscopy (Fig. 2), the author was again surprised to find that each bacterial cell possessed intracellular, iron-rich crystals which he speculated might be a magnetic material such as magnetite. [Pioneering studies of H. Lowenstam (1962a) describing biogenic magnetite comprising a hardened surface coating of mollusc radular denticles were unknown to him.] Shortly after the magnetotactic bacteria were reported (Blakemore, 1975), honeybees (Gould *et al.*, 1978) and birds (Walcott *et al.*, 1979), previously shown by ethologists to have geomagnetic sensibilities, were examined with sensitive cryogenic rock magnetometers to determine if they, too, might contain permanently magnetized material. These searches must have produced some heady moments!

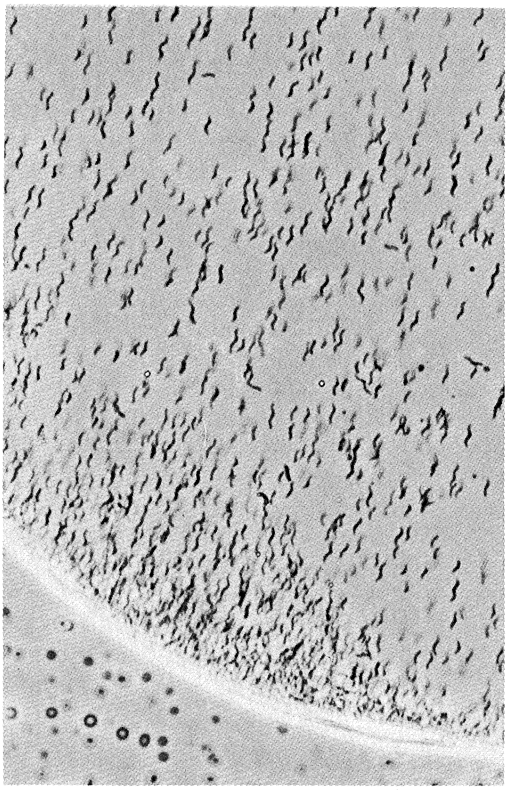

FIG. 1. Swimming dipping needles. Cells of the bacterium *Aquaspirillum magnetotacticum* each possess an intrinsic magnetic dipole moment [μ approximately $(1-3)\times10^{-12}$ erg G^{-1}; $(1-3)\times10^{-3}$ pA m^2] and consequently align in magnetic fields. Their swimming direction is thereby magnetism-biased (i.e., they are magnetotactic). Shown here are cells from a laboratory culture photographed with electronic flash to arrest their motion as they swim and preferentially accumulate at the edge of a water drop in response to an applied magnetic field of μT intensity. Each cell measures approximately 0.32×4.0 μm.

In the ensuing decade and a half, the list of magnetic organisms steadily increased (Table 1). Thus, magnetotactic bacteria appear to have influenced the entire field of biomagnetism by providing a novel, and simple but credible, paradigm for the basis of geomagnetic sensitivity in organisms. In 1975 it became possible to refer in biology to permanently magnetized cells—biomagnetism had acquired factual status.

For magnetotactic microbes, the connection between cell magnetism and cell alignment in the Earth's field has been conclusively established (see below). However, the fact that large, multicellular organisms such as bees or pigeons may have tissues as much as 10 or 20 times more magnetic than background does not necessarily equate their biomagnetism with a geomagnetic responsive-

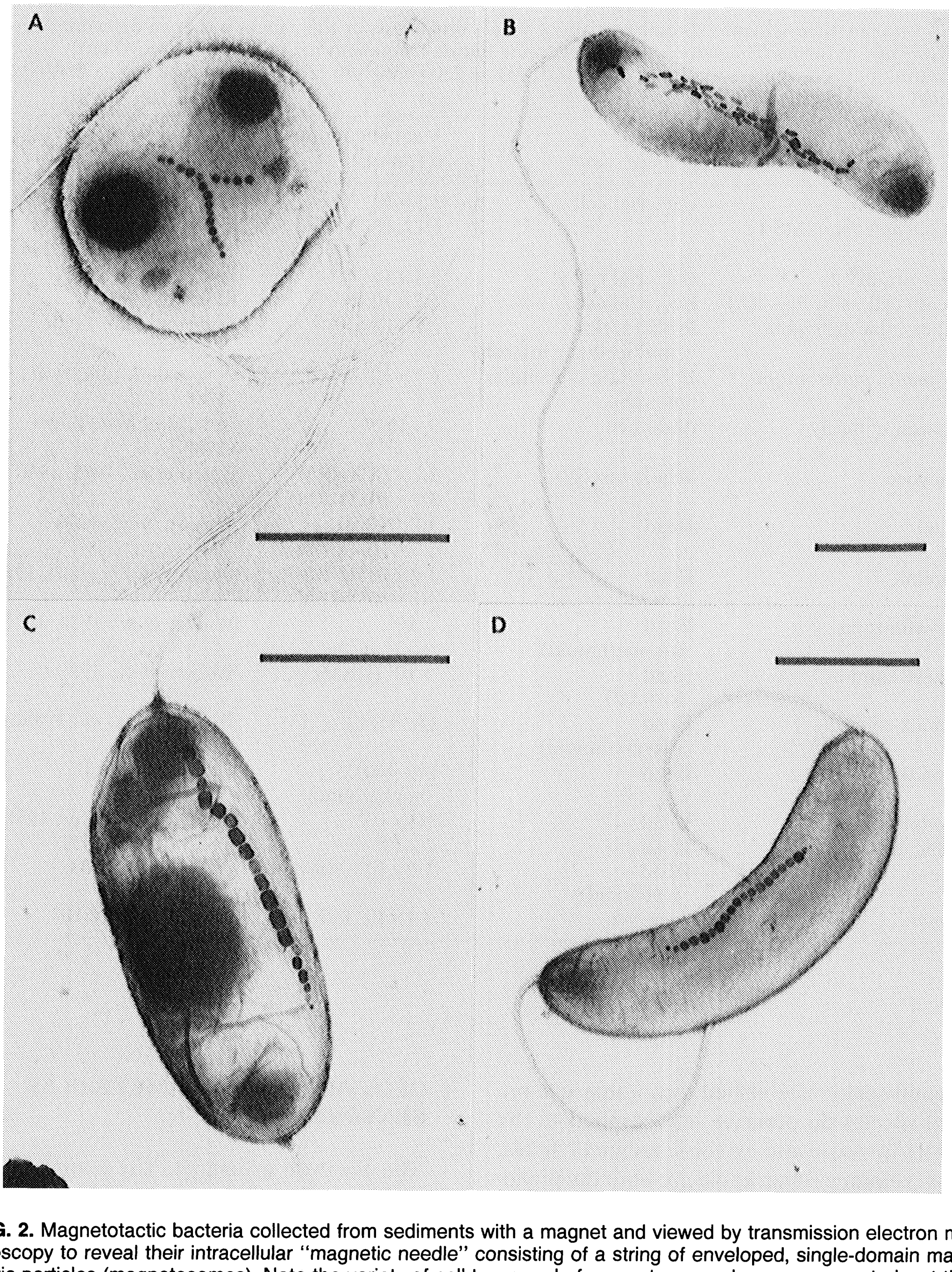

FIG. 2. Magnetotactic bacteria collected from sediments with a magnet and viewed by transmission electron microscopy to reveal their intracellular "magnetic needle" consisting of a string of enveloped, single-domain magnetic particles (magnetosomes). Note the variety of cell types and of magnetosome shapes represented, yet the consistency of magnetosome size and shape within each cell. Cells align in ambient magnetic fields by means of their intrinsic dipole moment imparted by the magnetosome string and swim in the field by means of their flagella. Bars = 1 μm.

ness. The greatest remaining challenge of biomagnetism research in higher organisms, in fact, may be separating out what appears to be an abundance of magnetized material in animals from that seemingly minute fraction which may be involved in magnetic sensing (Kirschvink, 1989). However, studies of conditioned and unconditioned behavior have

Table 1. Magnetic organisms. Biomagnetism of whole organisms or their parts has been determined for a variety of life forms. The values cited refer to the natural remanent magnetization (NRM) or saturation-induced remanent magnetization (sIRM) as determined in almost every case with a SQUID. See text (Sec. 2.3) for details.

Organism	Organ/Tissue	Magnetic moment μ (pA m^2)	Reference
Aquaspirillum magnetotacticum MS-1	Bacterial cell	1.3×10^{-3} $(1–2.4)\times10^{-4}$	Frankel *et al.*, 1979 Rosenblatt *et al.*, 1982; Moskowitz *et al.*, 1988
Bacterium MV-1	Bacterial cell	2.1×10^{-4}	Sparks *et al.*, 1990
Coccoid cell as in Fig. 2(A)	Bacterial cell	6.2×10^{-4}	Kalmijn, 1981
Multicellular aggregate	Prokaryotic multicellular microbe	$(0.5–1)\times10^{-3}$	Rodgers *et al.*, 1990
Anisonema platysomum	Eukaryotic unicellular microbe	6.7×10^{-1}	Torres de Araujo *et al.*, 1986
Monarch butterfly	Abdomen	9×10^{-4}	Jones and MacFadden, 1982
Honeybee	Whole insect	1.2×10^{3} (NRM) 2.7×10^{3} (sIRM)	Gould *et al.*, 1978, 1980
Pigeon	Head	10^{3} (NRM) 1.5×10^{4} (sIRM)	Walcott *et al.*, 1979
Bobolink	Head	1.9×10^{4} (NRM) 4.2×10^{4} (sIRM)	Beason and Nichols, 1984
Yellowfin tuna	Head (dermethmoid)	$>10^{3}$	Walker *et al.*, 1984, 1985
Sockeye salmon	Head (orbital)	$>10^{3}$ (sIRM)	Walker *et al.*, 1988
Chinook salmon	Head (dermethmoid)	$(3–9)\times10^{2}$	Kirschvink *et al.*, 1985b
Green turtle	Head	$(10–100)\times$ background	Perry *et al.*, 1981
Woodmouse	Head	1.3×10^{3}	Mather and Baker, 1981
Pacific dolphin	Head	2×10^{4}	Zoeger *et al.*, 1981
Human	Head (sphenoid)	1.6×10^{4} (sIRM)	Baker *et al.*, 1983
Human	Adrenals	3.4×10^{3} (sIRM)	Kirschvink, 1981b

unequivocally established that numerous animal species do perceive and respond to the earth's intrinsic and extrinsic magnetic fields. What remains remarkable are both the phylogenetic diversity of organisms with geomagnetic responsiveness, and their degree of magnetic sensitivity. Remaining enigmatic are *how* animals sense the different field parameters and, in most instances, exactly *what*, in their normal lives, they use magnetic sensing for. The remainder of this chapter addresses these questions, and provides selected examples illustrating geomagnetic-biomagnetic interactions among phylogenetically diverse life forms.

1. GEOMAGNETIC PARAMETERS AS SENSORY CUES

The Earth is a magnet. The geomagnetic field does, in fact, have a major, although not perfect, dipolar character. This *intrinsic* predominant, dipolar field (*H*) has a dipole moment value (μ) of 8×10^{22} A m^2 (=J T^{-1}) at its center. Earth's dipolar field has an intensity of roughly 50 μT at its surface and is directed northward globally (the field direction is, by convention, the direction in which the north-seeking end of a compass needle points). Except at the magnetic equator, the field has a vertical component and is inclined from the horizontal. It is directed outward (upward) in the southern hemisphere and increasingly inward (downward) as one moves north of the magnetic equator (Fig. 3). At any particular locale, however, the geomagnetic field is neither as uniform nor constant as one might expect (given the dimensions of the magnet!). The field has a gradient of intensity increasing by several nT/km as one moves poleward such that the totally horizontal field at the magnetic equator is 29 μT and the totally vertical field at the North magnetic pole is >80 μT. Iron deposits and other crustal features contribute to permanent 100–1000-nT (1 nT is 1 γ of the older literature) local spatial anomalies, usually of somewhat greater intensity occurring as "peaks" against the generally more even "magnetic topography" of the background field.

The Earth also has *extrinsic* field components. Diurnal heating and cooling of the globe cause north-south shifts of the ionosphere (magnetosphere) producing daily 30–50-nT variations against the 50-μT background geomagnetic field. Magnetic storms caused by influx of ions from solar flares impose additional, but unpredictable, 10–10 000-nT fluctuations in the field. The minute temporal and spatial variations in the intensity or direction of the geomagnetic field, although insufficient to deflect a magnetic compass needle visibly, appear to be important factors in animal orientation behavior and navigation. As discussed below, they influence honeybee dancing, pigeon homing, and bird migration, and appear to correlate with locations of live whale strandings, for instance.

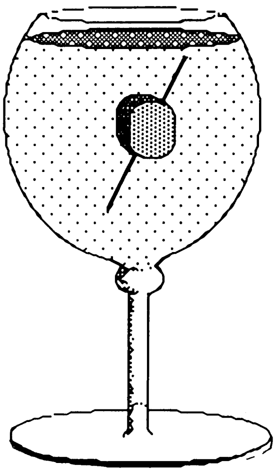

FIG. 3. An iron wire thrust through cork which is subsequently pared away to make the whole neutrally buoyant will, after being stroked with lodestone to magnetize it, always incline in the direction of magnetic dip (after W. Gilbert, 1600).

Potentially, organisms might detect the geomagnetic field direction, its intensity, intensity gradient in space or time, inclination, declination, or combinations of these parameters. Magnetism-based biological *compasses* respond to the overall dipolar field direction (although not all biological compasses are magnetic and at least two types exist among those which are). Small spatial anomalies in the Earth's field provide "magnetic topography" appropriate for a *navigational map,* and small diurnal or lunar fluctuations may enable temporal sensing thereby driving or entraining *biological clocks.*

2. BIOLOGICAL ADAPTATIONS AND GEOMAGNETISM

The sensory repertoire of birds, bees, and aquatic animals includes geomagnetic cues. Behavioral evidence suggests that in some animals, magnetic and other (visual, olfactory, vestibular, pineal) sensory systems or organs may be integrated. Except for the magnetotactic microbes, however, the magnetic information used and mechanism(s) by which organisms extract it from their surroundings remain unknown or equivocal. Larger, eukaryotic organisms, unlike microbes, cannot be passively oriented by earth-strength fields and therefore require neuronal processing of primary magnetic stimuli. They also respond to multiple and hierarchically arranged cues (physiographic, wind/air, light, celestial, odor, magnetic) in homing, migration, or communication (Southern, 1988; Gould, 1980a, 1982; Wiltschko *et al.,* 1981; Wiltschko and Wiltschko, 1989; Moore, 1980), of which geomagnetism may be of relatively low priority or preference.

In the early days of biomagnetic effects research the overall dipolar geomagnetic field was considered a "weak" field—perhaps too weak for biological sensing, critics argued. It is now more commonly considered to be "strong." This begs the question of just what a weak field may be. Birds (Southern, 1972b, 1988; Keeton *et al.,* 1974; Walcott, 1978, 1980) and especially honeybees (Lindauer, 1977; Walker and Bitterman, 1989a, 1989c) respond to nT fluctuations of the background field and may use normal 30–50-nT diurnal geomagnetic variations as *Zeitgebers*—environmental stimuli that serve to "set" a biological clock—or to calibrate solar or stellar compasses or other sensory systems. Magnetic storms, which produce 10–1000-nT fluctuations against the background field intensity of 50 μT, have resulted in disastrous bird losses when they have occurred during long-distance pigeon races (Keeton, 1974). Thus, the criterion for weak field has been set to a very low value by living organisms themselves. Ironically, animals possess sensitivity to these very small spatial and temporal perturbations of the geomagnetic field, yet when getting down to the actual business of moving around outdoors, they may select preferred cues over magnetic information. This has provided interest and challenge to the task of unraveling magnetic from other sensory capacities in eukaryotes.

Several mechanisms by which organisms might obtain and utilize geomagnetic information in their life's activities have been suggested. Of these, two presently have experimental support, one having been proven in the bacteria. The first involves electro-orientation (Faraday effect) in which an organism moving with respect to the geomagnetic field induces an electric field to which it is sensitive (Kalmijn, 1981). This is a "*dH/dt*" type of system in which the organism detects changes in the field with respect to its changing position or with time. The other (an "*H*" system in which absolute or relative field parameters are detected) recognizes that organisms or their components may be permanently magnetic ("biomagnetism") and thereby may couple magnetically to the geomagnetic field. Geomagnetic orientation via biomagnetism has been proven in the case of magnetotactic bacteria and is highly likely in at least some other organisms.

It has recently become surprisingly evident that there is a common basis for biomagnetism among diverse life forms, this property being conferred by an ability to produce and retain ultrafine magnetic biominerals of a particle size, particle number, interparticle configuration, and particle orientation or placement optimal for geomagnetic sensing.

2.1 Biogenic Magnetic Substances

Magnetic biominerals are known to be produced by unicellular bacteria, a multicellular prokaryote, and unicellular and multicellular (animal) eukaryotes. Table 2 summarizes currently known magnetogenic microbes, i.e., those which form magnetic biominerals (Blakemore and Frankel, 1989). Some magnetogenic microbes such as GS-15, *Shewanella*, and possibly *Desulfovibrio* produce magnetic minerals extracellularly in the culture medium. They are of little relevance to the subject of this chapter. Magnetogens which retain the magnetic biomineral phase intracellularly are magnetic. All magnetic magnetogens which have been described are motile, and they are *magnetotactic*; their motility is directed by magnetism. In addition to those which synthesize the iron oxide magnetite, magnetotactic bacteria have recently been discovered which produce intracellular

Table 2. The magnetogens. Diverse microbes produce magnetic iron biominerals, presumably as a result of dissimilatory iron reduction. Those (1–6) with intracellular particles of oxide or sulfide are magnetotactic and respond to geomagnetism. References to particular examples cited are (1) Maratea and Blakemore, 1981; (2) Moench, 1988; (3) Bazylinski *et al.*, 1988; (4) Torres de Araujo *et al.*, 1986; (5) Farina *et al.*, 1990; Rodgers *et al.*, 1990; (6) Heywood *et al.*, 1990; (7) Lovley *et al.*, 1987; (8) Myers and Nealson, 1988; (9) Bell *et al.*, 1987; (10) Freke and Tate, 1961.

Biomineral	Intracellular particles	Extracellular particles
Magnetic iron oxide (magnetite)	1. *Aquaspirillum magnetotacticum* MS-1 2. *Bilophococcus magnetotacticus* 3. Marine isolate MV-1 4. Magnetotactic "alga" *Anisonema*	7. Freshwater isolate GS-15 8. *Shewanella putrefaciens* strain MR-1 9. Mixed freshwater bacteria
Magnetic iron sulfide (greigite)	5. Prokaryotic multicellular aggregate 6. Uncharacterized rod-shaped bacteria	10. Anaerobic mixed microbial population containing *Desulfovibrio*?

magnetic iron sulfides (Farina *et al.*, 1990; Mann *et al.*, 1990; Heywood *et al.*, 1990) including greigite (Fe_3S_4) and the nonmagnetic mineral pyrite (FeS_2) or, perhaps, pyrrhotite (Fe_7S_8). Diverse magnetogens appear to produce iron biominerals by means of a common biochemical process which may involve cell respiration with iron reduction (dissimilatory iron reduction). A relationship between iron biomineralization and cell respiratory metabolism in magnetogenic microbes, if indeed it exists, is not expected to be true of animals because the latter are obligated to use O_2 in mitochondrial respiration.

Magnetite is one of the most prevalent biogenic minerals (Lowenstam, 1981). It has a characteristic ^{57}Fe Mössbauer resonance spectrum (Fig. 4) consisting of two overlapping six-line spectra and reflecting its mixed-valence state (Frankel *et al.* 1979, 1983). Magnetite can exist as bulk crystalline material such as in rocks, or as ultrafine particles. The tiniest grains, <40 nm, are superparamagnetic (SPM). Although they are permanently magnetized, their direction of magnetization is constantly buffeted by thermal forces, kT (commonly associated with Brownian motion), into one of the several crystallographically preferred or "easy" axes of magnetization. Thus, SPM particles maintain a constant direction of magnetization only when under the influence of an external, aligning magnetic field. Within large volumes of magnetite, magnetic domains (crystallographic regions in which unpaired spins are directed similarly) form which are separated by domain walls. This multidomain (MD) character is typical of magnetite particles larger than approximately 100 nm along a major axis (shape is important; somewhat longer but anisotropic particles may not take on MD character). Excluding anisotropy, for reasons having to do with the width of domain walls relative to the entire crystal dimension (see Frankel, 1982, 1984), particles larger than SPM (40 nm) and smaller than MD (<100 nm) exist as single domains (SD) of magnetite. SD grains have the saturating magnetization value, J_s, characteristic of the mineral ($480\ \mathrm{G\,cm^{-3}} = 4.8\times10^{5}\ \mathrm{A\,m^{-1}} = 4.8\times10^{5}\ \mathrm{J/T\,m^{3}}$ for magnetite); that is, they are each stably, spontaneously, and fully magnetized to the limit for the mineral and are, consequently, most efficient magnets.

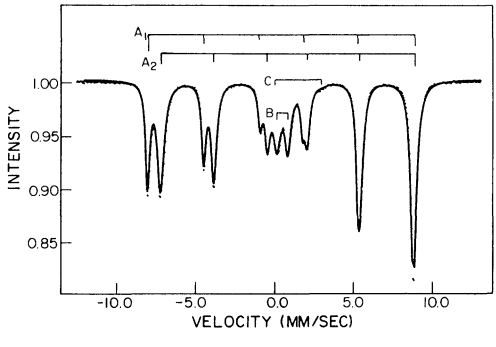

FIG. 4. The ^{57}Fe Mössbauer resonance spectrum of cells of the magnetic bacterium *Aquaspirillum magnetotacticum* indicates the presence of several iron species. Fe_3O_4 has a distinctive spectrum consisting of two overlapping six-line spectra denoted by lines A_1 and A_2. Spectral lines A_1 correspond to ^{60}Co gamma-ray absorption by Fe^{3+} in tetrahedral lattice sites of the mineral, whereas A_2 absorption lines are due to Fe^{2+} and Fe^{3+} in octahedral lattice sites. Spectral lines B are from a high-density hydrous iron oxide (ferrihydrite) and those marked C indicate the presence of some ferrous iron in the cells [from Frankel *et al.* (1983) and Frankel *et al.* (1985)].

2.2 The Bacterial Paradigm; A Biomagnetic Needle

The magnetic moment μ of a 50-nm SD magnetite particle, which is the product of its volume V and J_s, is 5.9×10^{-14} erg G^{-1} cm^{-3} (or electromagnetic unit, emu; see Glossary for equivalent SI unit). The particle magnetic energy tending to align it in the earth's field (i.e., the product of its magnetic moment μ and the geomagnetic field H) is roughly 3×10^{-14} erg. This is equivalent to the thermal energy ($kT=4.1\times10^{-14}$ erg) tending to randomize its orientation in water. Thus, despite having a permanent direction of magnetization, a single SD magnetite crystal will not align well in the geomagnetic field at room temperature. However, as predicted by the "chain of spheres" model of Jacobs and Bean (1955), SD grains constrained to lie in linear strings magnetically orient north-to-south and their moments are therefore additive. When connected in long strings within cells as in Fig. 2, SD grains comprising magnetosomes make superb magnetic needles. They align in the field (dragging their surrounding biomass with them) according to the Langevin function as in Fig. 5 (see Blakemore and Frankel, 1981; Frankel, 1982, 1984; Kalmijn, 1981); and when the ratio of their magnetic

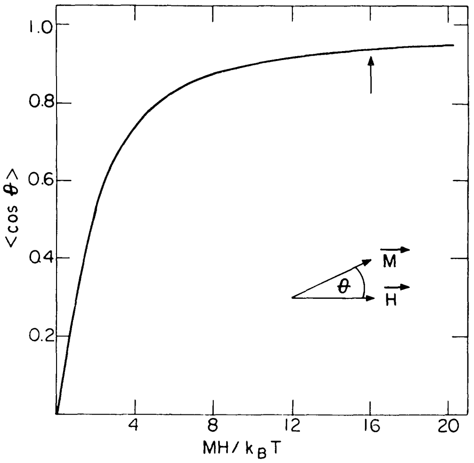

FIG. 5. The Langevin function: $\langle \cos\theta \rangle = L(MH/kT)$, plotted as a function of the ratio of magnetic (μ·**H** or **M**·**H**) to thermal (kT) energy. The arrow indicates the orientation of magnetotactic bacteria having a calculated cell dipole moment of 1.3×10^{-12} emu (1.3×10^{-3} pA m^2) in the geomagnetic field at ambient temperature (Frankel and Blakemore, 1980).

(μH) to thermal (kT) energies has a value of approximately 10 (as is the case for bacterial cells, which commonly have some 20 stringed particles), 80%–90% of the cells in a population are aligned in an earth-strength field of 50 μT (see Figs. 5 and 6 and Kalmijn, 1981). Neither more magnetite grains per cell nor a stronger ambient geomagnetic field would bring about significantly greater cell alignment in the field direction.

Consequently, magnetotactic bacteria in nature appear to be optimized for geomagnetic orientation in terms of

1. the value of their cell magnetic moment, i.e., number of magnetosomes;
2. the SD nature of their magnetosomes;
3. the arrangement of their magnetosomes in a string within the cell; and
4. the orientation of the string with respect to the axis of cell motility.

As stated by Professor E. Purcell, "This is a case of the compass steering the ship." Because their particles are SD and are arranged in a string, magnetic bacteria can be remagnetized (Kalmijn and Blakemore, 1978; Blakemore *et al.*, 1980; Mizota and Maeda, 1983), but not demagnetized. Magnetic moments of microorganisms have been evaluated in a variety of ways (see Sec. 3.1 and Table 1), yielding quite consistent values in the range $(0.1–1)\times10^{-12}$ emu $[(0.1–1)\times10^{-3}$ pA m$^2]$.

2.3 The Quest for Biomagnetic Needles

Animals containing magnetic biominerals frequently have small quantities per unit biomass. Moreover, in the case of magnetite, for which the SD dimension limits are known, the crystal size is limited (by unknown biological means) to the SD size range. SD magnetite grains are from 10 to 100 nm, and therefore are near the resolution limit (20 nm) of light microscopy. Light microscopy is not a reliable means of detecting even the largest magnetosomes. The Perl reaction involving ferricyanide can be used to stain bulk material for iron deposits but no histochemical stain specific for magnetite exists. How, then, may one determine the presence and form of minute quantities of minute magnetic biomineral crystals in biological samples? Electron- and x-ray-diffraction spectra, as well as the ^{57}Fe Mössbauer spectrum (Fig. 4), of magnetite are distinctive. However, when confronted with limited amounts (in some cases only several crystals) of material *in situ*

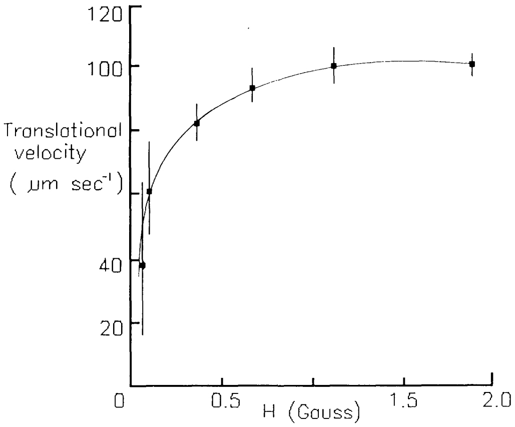

FIG. 6. The magnetic moment of magnetotactic bacteria from natural mud samples can be evaluated by a Langevin analysis. Cells directed by an ambient field of varying intensity up to 2.0 G (2×10^{-4} T) will orient progressively well as the field is increased, such that their forward speed (i.e., their translational velocity as opposed to absolute cell velocity) increases to the maximum cell swimming speed. As shown, cells in populations in natural habitats are 80%–90% aligned at the Earth's field intensity of 0.5 G (50 μT): they are optimized for geomagnetic interaction (see also Kalmijn, 1981).

or after tissue extraction, investigators of animal magnetism have too little material for techniques such as Mössbauer spectroscopy. This leaves as useful tools cryogenic magnetometry by means of superconducting quantum interference devices (SQUIDs), and the electron microscopical techniques of energy-dispersive x-ray analysis (EDXA), high-resolution transmission electron microscopy (HRTEM), and electron diffraction. The usefulness of magnetometry is related directly to the degree of care exercised in the skill of tissue dissection aimed at localizing biological magnetic material. Electron microscopy provides necessary confirmation of the nature of biomagnetic materials but at an extremely different degree of scale, the successful transition between these two scales requiring both skill and fortune.

Obviously, the most useful property in a search for biogenic magnetite is its magnetism. For instance, magnetite has a characteristic Curie temperature (858 K) at which it loses magnetism. Josephson-junction-based cryogenic rock magnetometers (Goree and Fuller, 1976; Fuller, *et al.*, 1985) provide sufficient sensitivity to determine the intensity of weak, natural remanent magnetization (NRM) of relatively large (100 cm^3) biological samples such as entire insects, bird heads, or portions of organs. The techniques are noninvasive and essentially nondestructive (of tissues). The presence of magnetite can raise the sample NRM above background by a factor of tens. Application of a strong ($>$ 300 mT) magnetizing field to a sample containing ferromagnetic material saturates its magnetic moment to the maximum value (i.e., that of its saturation isothermal remanent magnetization, or sIRM value). When magnetized in small incremental steps, resulting sIRM values for small, strongly magnetic, and homogeneously sized SD particles plot in a narrow range of magnetizing field values (see, for instance, Kirschvink *et al.*, 1985b; Walker *et al.*, 1985; Moskowitz *et al.*, 1988). The value at which a magnetizable substance maximally saturates provides information concerning its form and composition; MD magnetite saturates at about 90 mT, whereas SD magnetite continues to gain remanence to about 300 mT, while magnetizing fields greater than 1000 mT are required to saturate hematite and magnetic alloys containing nickel or cobalt. Additional information can be obtained by measuring remanence during progressive alternating current (ac) or alternating field demagnetization of the sample. Resulting coercivity spectra (plots of sample acquisition and loss of remanence) can indicate magnetic domain size and the particle-size distribution (see Cisowski, 1981; Walker *et al.*, 1985, 1988; Moskowitz *et al.*, 1988). It is possible to demagnetize MD but not SD particles or even assemblages of SD particles if they occur in certain distributional arrangements. Linear strings of SD particles, as in the arrangement of bacterial magnetosomes (Fig. 2), allow for strong particle-particle magnetic interactions. The resulting anisotropic character of the ensemble resists demagnetization. By carrying out ac demagnetization in the presence of a constant dc biasing field, one can determine the sample anhysteretic remanent magnetization (ARM). The resistance of (weak field) ARM to ac demagnetization is greater than that of (strong field) sIRM if the particles are SD or very small MD. Conversely, large MD particles have the opposite behavior. This different resistance to demagnetization of MD and SD particles is the basis of the Lowrie-Fuller test (Lowrie and Fuller, 1971; Fig. 7). Thus, merely determining that the median

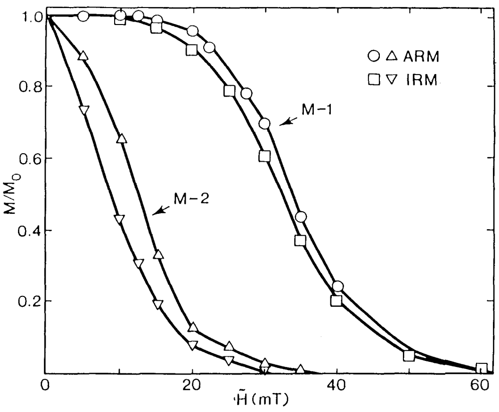

FIG. 7. The Lowrie-Fuller test applied to intact cells (M-1) and to extracted strings of magnetosomes (M-2) of *Aquaspirillum magnetotacticum*. The data are consistent with the presence of SD magnetite grains in a string arrangement as discussed in the text. (From Moskowitz *et al.*, 1988.)

destructive demagnetizing field of ARM is greater than that of sIRM is sufficient to establish the presence of SD particles in biological tissue. The manner in which ARM varies with variation in the dc biasing field also provides information regarding possible grain-grain interactions (i.e., reflecting the degree of separation between magnetic particles). As discussed by others (Kirschvink *et al.*, 1985b; Walker *et al.*, 1984, 1988; Moskowitz *et al.*, 1988), Cisowski (1981) showed that the shape of the ac demagnetization curve when compared to that of gain in sIRM should be symmetrical about the 50% magnetization point for uniformly dispersed magnetic particles. Particle-particle magnetic interactions, on the other hand, cause the intersection of these curves to fall below the 50% magnetization point. This technique has been useful in deducing that SD particles detected in fish (see Sec. 3.8) are arranged in strings as they are in the magnetotactic bacterium *Aquaspirillum magnetotacticum*.

Properly shielded and in clean environments, cryogenic magnetometers can provide sensitivity sufficient to detect several ng of magnetite. An instrument at Caltech, situated in a background noise of 5×10^{-12} A m^2, is sensitive to 0.1-ng magnetite in a sample volume of roughly 3000 cells (Kirschvink, 1982; Walker *et al.*, 1985).

The successful hunt for geomagnetically meaningful biological magnetism, then, should involve preliminary behavioral verification that an animal senses geomagnetism (preferably through conditioning but not merely through correlative studies), followed by magnetometry to localize magnetic tissues (carried out in concert with histological examination for iron deposits). Subsequently, conventional dissection (in unconventionally "clean" surroundings), chemical fixation, and electron microscopy allow fine-scale searches at the ultrastructural level. Most investigators hunting "biomagnetic needles" have concentrated on innervated tissues. Ultimately, influences of geomagnetism on nerve impulse transmission or other electrophysiological evidence of neuronal involvement will be necessary to elucidate completely the nature of the magnetic sense in higher organisms.

3. SOME EXAMPLES OF BIOLOGICAL INTERACTIONS WITH THE GEOMAGNETIC FIELD

Sufficient information has been gleaned from a number of living systems such that a "comparative biology" of geomagnetic responsiveness is emerging (Blakemore *et al.*, 1988). The variation typical of biology is also evident from consideration of biomagnetic-geomagnetic interactions exhibited by the phylogenetically diverse organisms described in the remainder of this section. A common theme throughout, however, appears to be the presence in geomagnetically responsive organisms of cell, tissue, or organ magnetic susceptibility—a property conferred upon widely different organisms by ultrafine magnetic particles similar to those within magnetotactic bacteria.

3.1 Prokaryotes

Except for the eukaryotic euglenoid discussed below, known magnetotactic microbes are prokaryotic. Great morphological diversity exists among these bacteria, and they are rather cosmopolitan in their environmental distribution. However, only several of the dozens of easily discernible types have been isolated and grown in axenic culture (Blakemore, 1982). Magnetotactic bacteria which have been phylogenetically analyzed belong to the α-subgroup of the proteobacteria; they are Gram-negative eubacteria (Eden *et al.*, 1991). No magnetotactic (or magnetogenic) archaebacteria have been described. None have been shown to exist autotrophically. All magnetotactic bacteria examined to date have been Gram-negative and facultatively or obligately anaerobic or facultatively or obligately microaerophilic. No strictly aerobic magnetotactic (or magnetogenic, for that matter) bacteria have been described.

Magnetotactic bacteria containing intracellular, membrane enveloped, iron-rich, structured particles (Blakemore, 1975) were shown by means of pulse remagnetization (Kalmijn and Blakemore, 1978) to be permanently magnetized cells. Axenically cultured cells (Blakemore *et al.*, 1979) of a microaerophilic, freshwater spirillum, *Aquaspirillum magnetotacticum* strain MS-1 (Maratea and Blakemore, 1981), were examined by Mössbauer spectroscopy which revealed that their particles, missing from nonmagnetotactic cells of the same species, were magnetite (Frankel *et al.*, 1979). Balkwill *et al.* (1980) referred to the *A. magnetotacticum* cell magnetic apparatus as the "magnetosome." More recent ultrastructural evidence indicated that each element of the magnetic apparatus of this organism consists of a crystalline mineral core enveloped within a lipid bilayer membrane (Gorby *et al.*, 1987) and the term "magnetosome" should now more appropriately be used to denote *each* membrane-bound crystalline particle (Fig. 8). Magnetic properties including NRM, sIRM, and ARM values and results of the Lowrie-Fuller test applied to *A. magnetotacticum* cells and their purified magnetosomes (Fig. 7) have been reported (Denham *et al.*, 1980; Moskowitz *et al.*, 1988).

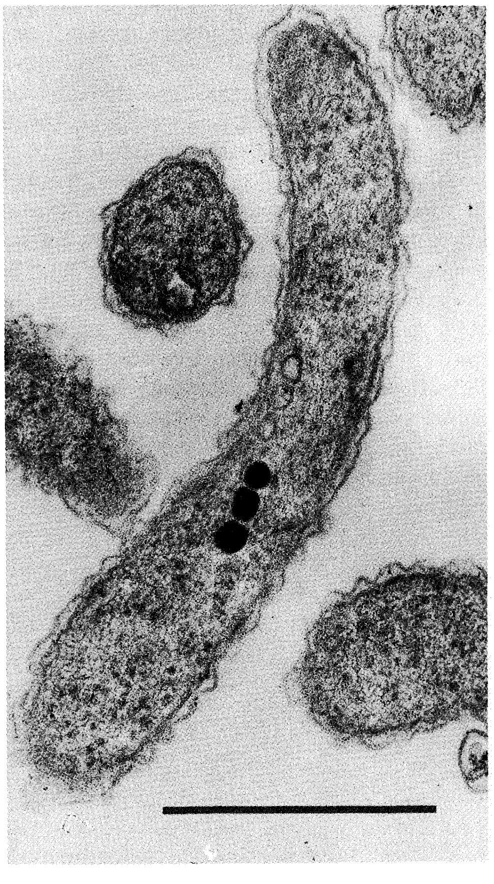

FIG. 8. Thin section through a portion of the magnetotactic bacterium *Aquaspirillum magnetotacticum* and viewed by transmission electron microscopy. Intracellular magnetosomes are evident as are triple-track (lipid bilayer) intracytoplasmic membranous magnetosome envelopes occurring as vesicles devoid of magnetite cores. Bar = 0.5 μm. (From Gorby *et al.*, 1987.)

The magnetic moments of magnetic microbes have been evaluated in a variety of interesting ways. These have included

1. electron microscopical determination of the volume of magnetite present, having a known value of J_s (Frankel *et al.*, 1979; Frankel and Blakemore, 1980; Kalmijn, 1981);
2. determination of the radius of U-turns and circular motions made by swimming cells in response to reversal of the ambient magnetic field and as a function of the applied field strength (Bean, 1979, 1989);
3. evaluation of cell translational velocity as function of aligning field intensity (e.g., a "biological Langevin analysis"; Kalmijn, 1981; Rodgers *et al.*, 1990);
4. pulse remagnetization (see Fig. 9; Kalmijn and Blakemore, 1978, and unpublished data);
5. evaluation of magnetic-field-dependent cell alignment as determined from elastic light scattering and magnetic-field-dependent cell birefringence (see Fig. 10; Rosenblatt *et al.*, 1982a, 1982b); and
6. use of cryogenic magnetometers (Moskowitz *et al.*, 1988, 1989).

Values of cell magnetic moments obtained by these diverse means are consistent among bacteria (Table 1), roughly $(0.1\text{–}1)\times 10^{-3}$ pA m^2.

The quantity, SD particle size, and arrangement of magnetosomes within the bacterial cell are ideal for geomagnetic sensing through biomagnetism and passive cell orientation in the geomagnetic field (Frankel and Blakemore, 1980, 1989; Blakemore, 1982; Blakemore and Frankel, 1981; Frankel, 1982, 1984). Studies of natural populations from around the world have revealed global selection for cell types with appropriate magnetic polarity for directing them away from surface waters (Blakemore *et al.*, 1980; Frankel *et al.*, 1981). The fact that no aerobic magnetotactic bacteria have been observed is consistent with an ecological role for magnetotaxis in directing cells downward toward less-oxygenated zones where they thrive. Two other strains of freshwater magnetic spirilla, MS-2 (isolated by N. Blakemore from the Durham,

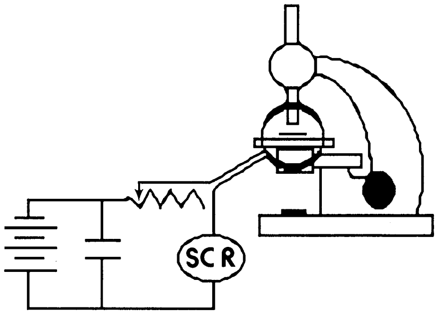

FIG. 9. Pulse remagnetization of bacterial cells (Kalmijn and Blakemore, 1978) was carried out while viewing them by means of a light microscope from which all ferrous metal components near the sample had been removed or replaced with nonferrous metal components. A 4-cm-diam single-turn Helmholtz pair of copper coils was fitted to surround the sample of bacteria on the microscope stage. A 4-μF capacitor connected to the coil in series with a variable resistor (critically damped RCL circuit) was discharged (triggered by a fast-switching silicon-controlled rectifier, SCR) to obtain a monophasic single-current pulse (up to 1350 A) having a time constant of 1.2 μs. The pulsing apparatus was placed within two room-sized Helmholtz pairs by means of which the local geomagnetic field was compensated or controlled. Magnetic pulses delivered antiparallel to the swimming bacteria were intense enough (200–600 G) and brief enough to effect their repolarization. Remagnetized cells in a constant ambient field immediately made U-turns and swam thereafter in the opposite direction. No cells became demagnetized.

NH, water-treatment plant) and a freshwater spirillum similar to MS-1 and MS-2 (T. Matsunaga, personal communication), have been isolated and cultured axenically, as has a magnetotactic, marine, helical to vibrioid strain, MV-1, which, unlike strains MS-1 (Blakemore *et al.*, 1984) and MS-2, produces magnetosomes anaerobically (Bazylinski *et al.*, 1988). These organisms are all microaerobic, tolerating O$_2$ to various low degrees, but none produce magnetite under aerobic conditions.

Magnetosomes of the marine strain MV-1, which appear as rectangular parallelepipeds in horizontal projection, are elongated hexagonal prisms (Sparks *et al.*, 1990) and crystallographically similar to particles from certain, as yet uncultured, magnetotactic bacteria [Fig. 11(A), 11(B)] found in sediments (Mann *et al.*, 1984b; Matsuda *et al.*, 1983; von Dobeneck *et al.*, 1987; Vali *et al.*, 1987; Petersen *et al.*, 1986; Towe and Moench, 1981). Magnetosomes of other species (Fig. 2), including those of *A. magnetotacticum*, which appear equant, with slightly rounded corners in horizontal projection, are cubo-octahedral [Fig. 11(C); Mann *et al.*, 1984a]. Still other magnetic bacteria collected directly from sediments possess anisotropic magnetosomes having a tear-drop or bullet shape (Fig. 2), often slightly arched (Blakemore *et al.* 1989) but based upon a hexagonally prismatic morphology (Mann *et al.*, 1987a, 1987b). Lattice imaging with HRTEM has revealed that SD particles, regardless of their morphology, when arranged in strings, have preferred crystallographic orientation such that the {111} faces of adjacent crystals are perpendicular to the string axis. This arrangement places the crystallographic easy axis of particle magnetization along the axis of the string, resulting in particle magnetic moments being

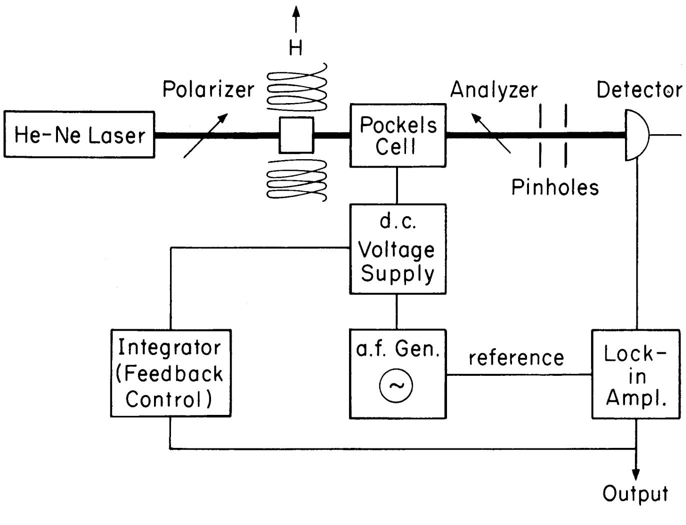

FIG. 10. Apparatus used by Rosenblatt *et al.* (1982a, 1982b) to measure the magnetic moment of *Aquaspirillum magnetotacticum* cells based upon their field-dependent optical birefringence. The bacterial cell suspension in a cuvette was placed within an electromagnetic coil all within a Mumetal enclosure to control bacterial magnetic alignment.

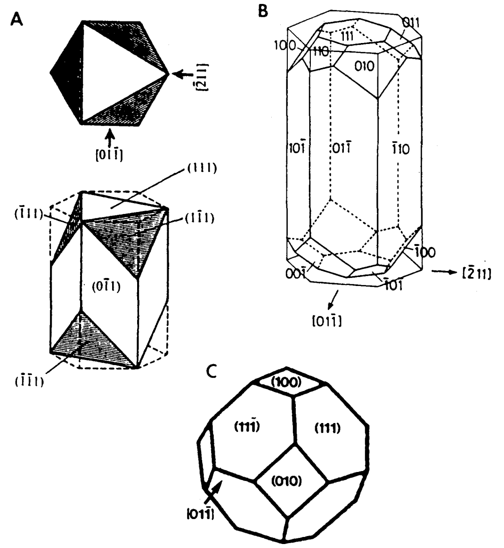

FIG. 11. Bacterial magnetosomes contain single-crystal mineral cores which are single magnetic domains. Magnetite grains which have been examined occur as truncated hexagonal prisms as in (A) (from Matsuda *et al.*, 1983) and (B) (from Mann *et al.*, 1984b), or as truncated octahedrons as in (C) (from Mann *et al.*, 1984a). Other types of crystal morphologies and mineral cores exist (see text). The crystal dimensions are roughly 40–100 nm along a major axis.

additive and in confirmation of the Jacobs and Bean "chain of spheres" model of magnetic SD interaction.

The crystal morphology of bacterial magnetite (but not that of sulfidic types of) magnetosomes varies in what is apparently a species-specific manner. This implies genetic control over magnetosome crystal morphology as well as biocontrol of their general arrangement within the cell (e.g., "stringed" vs clustered in arrays). Some biogenic forms of magnetite have not been synthetically produced. Biogenesis of fine-grain magnetite by *Aquaspirillum magnetotacticum* proceeds via partial reduction of a ferrihydrite precursor (Frankel *et al.*, 1983). Intracellular bacterial magnetite is formed, therefore, in a controlled process. It is tempting to speculate that the magnetosome membrane (Fig. 8), which is a true biological membrane consisting of a lipid bilayer in *A. magnetotacticum* (Gorby *et al.*, 1987), may control aspects of biomineralization including the maximum (SD) particle-size limit, arrangement in the cell, and crystal morphology. Consistent with this is the observation that extracellular magnetite produced in the absence of membranous structures by organism GS-15 (Lovley *et al.*, 1987) lacks the uniformity of crystal size and morphology exhibited by the biologically controlled forms of this mineral in magnetotactic bacteria and is not uniformly SD (Moskowitz *et al.*, 1989). Magnetite present within several higher organisms has been shown to exist within membranous vesicles (see Secs. 3.6 and 3.8). The fact that the somewhat elusive (Balkwill *et al.*, 1980) magnetosome membrane has not been detected in some magnetic bacteria (Bazylinski *et al.*, 1988) suggests caution in assuming that it has an important role in mineral biogenesis, however. More research is needed.

A northern hemisphere downward-swimming multicellular magnetotactic prokaryote from sulfidic, brackish water has been described (Rodgers *et al.*, 1990). A similar organism in Brazil (Esquivel *et al.*, 1983), also swam downward in that hemisphere. How individual flagellated cells comprising this organism are coordinated to bring about its translational motility in the magnetic field direction remains a mystery. Individual cells each contained magnetosomes consisting of the iron sulfide pyrrhotite (Farina *et al.*, 1990) or a mixture of pyrite (a nonmagnetic biomineral) and greigite (Mann *et al.*, 1990). The sulfides were detected using EDXA and their mineral structure was confirmed by electron diffraction and lattice imaging with HRTEM. The magnetic moment of an entire multicellular organism as determined by the method of Kalmijn (1981) was estimated to be 10^{-3} pA m^2 (Table 1). Magnetic iron sulfides have also been detected within unicellular prokaryotes found as cohabitants with this multicellular form (Heywood *et al.*, 1990; R. Blakemore, N. Blakemore, and R. Frankel, unpublished data). The process of magnetic iron sulfide biomineralization by these magnetogens is completely unknown.

There is strong selection pressure on natural populations of magnetotactic bacteria for downward swimming: virtually all cells swim downward. That this is a consequence of the interaction between cell magnetism and geomagnetism is evident both from study of natural populations and laboratory observations (Blakemore, 1982; Blakemore *et al.*, 1980; Kirschvink, 1980; Spormann and Wolfe, 1984; Torres de Araujo *et al.*, 1990). Wendy O'Brien in the author's laboratory incubated six small, natural enrichment cultures within solenoids capable of inverting only the vertical component of the local (northern hemisphere) geomagnetic field. The numbers of cells in each culture (as determined by direct microscopic counts) which migrated in the field direction were compared with those directed opposite to it before and after energizing three of the six coils (Fig. 12, vertical arrow). Within several weeks, south-and-down-seeking cells dominated the total population of magnetotactic bacteria in sediment samples within coils energized to reverse the vertical component (FR) but not in control samples in normal fields (FN). The population density of all magnetotactic bacteria decreased by several orders of magnitude (most likely as a result of the disruptive effect of O_2 introduced during the sampling process) but with strongest selection against upward directed cells. Obviously, the bacterial population responded within a few generations to magnetic-field inversion by acquiring reversed magnetic polarity, thereby preserving the downward directed trait. This new population of cells having reversed polarity probably derived from

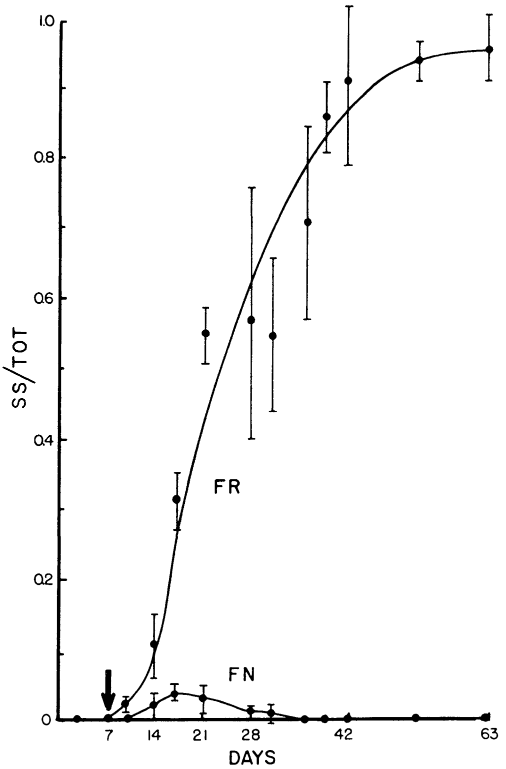

FIG. 12. Cell remagnetization without repolarization. Entire populations of magnetotactic bacteria as found in muds in the Northern Hemisphere, when incubated within solenoids to invert the vertical component of the local geomagnetic field, within several weeks acquired reversed magnetic polarity. Data shown indicate the mean values and ranges determined from direct cell counts in triplicate of north- and south-swimming bacterial cells in samples from triplicate cultures incubated in inverted (FR) and normal (FN) geomagnetic fields. The ratio of south-swimming to total magnetotactic cells is plotted against time. The vertical arrow indicates the point at which three of the six solenoids were energized (three were left as unperturbed controls) to invert the local field (FR). Populations of magnetotactic bacteria responded to a field inversion and preserved their downward swimming behavior.

1. death of upward-directed cells, and
2. favorable selection of the rare "improperly" magnetized cells present in any population resulting in their proliferation.

Magnetic repolarization requires either strong field gradients (Blakemore *et al.*, 1980) or magnetic pulses (Kalmijn and Blakemore, 1978; Mizota and Maeda, 1983) and was not a possible factor in these experiments. Thus, natural selection accomplished what could be managed by physicists only with a fair amount of energy: cell remagnetization in the absence of magnetic repolarization.

Despite this fascinating natural selection for downward swimming, magnetotaxis is not an essential trait. Nonmagnetotactic, nonmagnetogenic mutants of *Aquaspirillum magnetotacticum* grow at rates and to final cell yields of the magnetic parent strain in laboratory culture (laboratory cultures typically lack O_2 gradients, sediments, and other features which, in nature, maintain selection pressure for magnetotaxis). Moreover, magnetotactic bacteria in natural habitats can override their magnetic orientation with other taxes such as aerotaxis (Blakemore, 1982; Spormann and Wolfe, 1984). However, magnetic orientation will provide a constant bias to cell motion should other directional influences wane sufficiently to allow its expression (e.g., if chemical gradients are disrupted or move, or if cells or sediments containing them are physically dispersed as by scientists' wading feet, or wave or current action). In this regard, bacteria respond to multiple sensory inputs and are not so unlike pigeons (Sec. 3.6).

3.2 Eukaryotic Microbes

Torres de Araujo *et al.* (1986) reported magnetotaxis at geomagnetic field strengths in a heterotrophic, unicellular euglenoid, *Anisonema platysomum* Skuja from a Brazilian mangrove swamp. Each cell possessed numerous iron-rich crystalline particles having electron-diffraction patterns with d spacings characteristic of magnetite. Each $(80\text{–}180)\times(40\text{–}50)\text{-nm}^2$ particle was similar in size to the previously reported SD particles within magnetotactic bacteria. Each appeared to be of the anisotropic, hexagonally prismatic type of particle morphology peculiar to certain magnetotactic bacteria [see Fig. 2(B)]. On the basis of cell motion analysis in a reversing magnetic field, each "algal" cell was estimated to have a magnetic moment (Table 1) of 6.7×10^{-10} emu (6.7×10^{-1} pA m^2), or 600–700 times that of a representative magnetotactic bacterium. Assuming the particles were magnetite, each particle magnetic moment was estimated from its volume to be approximately 1.5×10^{-13} emu. From consideration of its total magnetic moment, each *Anisonema* cell was estimated to possess approximately 3000 such particles. Electron microscopy revealed cells densely endowed with strings of anisotropic, tooth-shaped magnetosomes. Thus, by maintaining the magnetite particle size in the magnetically efficient SD size range while increasing the number of strings of particles per cell, these eukaryotic unicellular organisms, although much larger than magnetotactic prokaryotes, and consequently having much greater coefficients of viscous drag, had roughly similar efficiency of (passive) orientation in the geomagnetic field due to their enhanced magnetism.

3.3 Mollusks

Some of the earliest reported biomagnetic effects were those of Frank Brown, Jr., who, working in Woods Hole, Massachusetts, in the 1960s, investigated responses to magnetism of a variety of invertebrates. Among these were studies indicating that mud snails, *Nassarius* (*Illyanassa*) *obsoleta*, were magnetically sensitive, showing responses correlating with lunar periodicities (Brown, Webb, and Brett, 1960) with maximal responses in mean daily snail orientation occurring near new moon and minimal responses at full moon. Despite these workers' efforts to address effects of geomagnetic strength fields (Brown, Webb, and Barnwell, 1964), it is unfortunate that their exhaustive studies were conducted by imposing very strong (10× Earth strength) inhomogeneous magnetic fields obtained with bar magnets upon the uniform geomagnetic field. This makes interpretation of these pioneering studies, although probably correct in their most important conclusions, difficult at best. Additional behavior studies with electromagnetic coil systems to provide uniform fields and analysis by magnetometry and electron microscopy of the invertebrates they studied might prove very interesting. Similarly, additional studies of a very interesting magnetic response in the sea slug *Tritonia*, more recently described by Lohmann and Willows (1987), are called for by their behavioral data.

Permanently magnetized biological material was first found in chitons. These primitive marine mollusks possess radular teeth capped with the iron biominerals magnetite, goethite, and lepidocrocite (Lowenstam, 1962a, 1962b, 1967; Towe and Lowenstam, 1967). The role of magnetite relates to its density and abrasion-resistance because these animals graze algae from rocks. However, strongly magnetized chiton teeth (in addition to other sources of biogenic magnetite) have been implicated in magnetization of marine sediments (Kirschvink and Lowenstam, 1979). Lowenstam (1962a) proposed that magnetite might serve to orient chitons with respect to the geomagnetic field. Tomlinson *et al.* (1980) noted seasonal alignment of chitons in captivity and in the sea in artificial and geomagnetic fields, but this has not been meaningfully related to the creature's biology.

3.4 Honeybees

Honeybee (*Apis mellifera*) scouts communicate directional information regarding a distant feeding site to forager bees in a hive by means of their waggle dance which they perform in the dark on normally vertical combs. This ritualized dance includes a straight run along the comb during which the bee waggles its abdomen (von Frisch, 1967). The run direction is related to the vertical (gravity vector) by an angle essentially equal to that between the sun's azimuth and the direction to the food source (e.g., up the comb if the food site is on an axis between the hive and sun, down if it is directly opposite the sun from the hive, and at appropriate intermediate angles with respect to vertical for sites at corresponding angles between these extremes). Martin and Lindauer (1977; see also, Lindauer and Martin, 1972) demonstrated that scouts consistently incorporated into their waggle dance a 15°–20° angle of misdirection (*"Mißweisung"*) between the vertical gravity vector and the inclined geomagnetic vector. Forager bees thus (mis)informed accurately compensated for this error when interpreting dance information relating the angle between solar azimuth and the line to the feeding site. Geomagnetism appeared to influence this important social communication behavior. *Mißweisung* disappeared over a period of several weeks if bees were made to dance in the dark on horizontal rather than vertical comb surfaces (in which case the bees danced along the cardinal points of the compass) and it disappeared entirely in a nulled geomagnetic field (Martin and Lindauer, 1977; see also, Kirschvink, 1981a).

Martin *et al.* (1989) showed that compensation of the geomagnetic field to within 2% of its natural intensity significantly reduced the normal tempo of scout bee waggle dances. This effect was even more pronounced if an oscillating (5 Hz) 90–125-μT sinusoidal magnetic field was imposed in an east-west direction after the initial compensation. The slowed dance tempo resulted in communication of misinformation regarding the distance to the feeding location, an error corresponding to more than 250 m in the particular experiments reported. Thus, alterations of earth-strength magnetic fields misdirected bees (through altered *Mißweisung*) and also (through altered dance tempo) caused them to communicate errors in distance information regarding location of a food source. These and other results clearly indicated that bees sensed the overall geomagnetic field.

Bees also sense temporal fluctuations in the geomagnetic field. Lindauer and Martin (1972) obtained evidence that normal diurnal fluctuations in the geomagnetic field influenced *Mißweisung*, and Walker and Bitterman (1989a, 1989c) have trained bees to respond to such fluctuations (discussed below). Martin (discussed in Neumann, 1988) has suggested that natural rhythms in geomagnetic and astrophysical magnetic fields are so regular, and detected so well by bees, that they use these directly in time-keeping functions, thereby obviating a role for an endogenous circadian clock. While this has been contested (Neumann, 1988), bee sensitivity to temporal magnetic fluctuations has not. Honeybees do, in fact, possess a time-averaged sun compensation system; e.g., after being deprived of sunlight for several hours, they can extrapolate the sun's rate of movement from when it was last observed (Gould 1980b). But unlike birds (see below), bees do not appear to use geomagnetic information as a backup orientation system on cloudy days. Rather, if on a cloudy day their hive is moved to a new location but of similar appearance with respect to a bait and major local landmarks such as a line of trees, they then refer in their waggle dance to (an informationally incorrect) solar azimuth which would be as predicted if they somehow remembered the sun's course as it would have been at the original site (Dyer and Gould, 1981).

Other effects of altered geomagnetic field on bee physiology and flying behavior have been noted. For instance, worker bees showed 40% reduced foraging activity (and, consequently, increased lifespan) in the presence of a static, inhomogeneous 80–145 μT field (Martin *et al.*, 1989), and comb-building by bees (de Jong, 1982) has been shown to have a geomagnetic field-dependent orientation. Free-flying (but not stationary) honeybees have been conditioned to discriminate the presence and absence of local magnetic anomalies (time-varying 0.01–10 Hz; 1–100 μT) superimposed on the background geomagnetic field (Walker and Bitterman, 1989c). As predicted if the basis of their geomagnetic sensitivity were through a mechanism involving permanent ferromagnetic material, flying bees with stainless-steel magnetized wires glued to their anterodorsal abdomens (Walker and Bitterman, 1989b) were significantly less able to be trained to discriminate between food targets (one having an associated anomalous magnetic field) as compared to controls with, or without, copper wires.

The mechanism of bee geomagnetic sensitivity and response is unknown. However, bees have single magnetic domain (SD) and superparamagnetic (SPM) materials of high coercivity (or else showing anisotropic character) located mostly in their abdomens (Gould *et al.*, 1978). Natural remanent magnetization (NRM) values of $(0.34\text{–}1.2)\times10^{-6}$ emu $[(0.3\text{–}1.2)\times10^{3}$ pA m$^2]$ and sIRM of $(1.7\text{–}2.7)\times10^{-6}$ emu per bee have been measured (Gould *et al.*, 1978; Gould *et al.*, 1980). This would correspond to roughly 2×10^{-8} g of Fe_3O_4. Apropos of this, a band of iron-rich oenocyte cells containing abundant particles consisting of (nonmagnetic) hydrous iron hydroxides was located in the abdomens of worker bees (Kuterbach *et al.*, 1982). These would not contribute to the magnetic remanence values measured by Gould *et al.* (1978). However, only 1/3 of the total hydrous iron oxide estimated to be present would need to be reduced to produce the quantity of Fe_3O_4 needed to account for the observed magnetic moment. Kuterbach *et al.* (1982) noted that this quantity of Fe_3O_4 would not have been detected by their methods.

Pulse remagnetization of foraging bees has reportedly been carried out and the results considered "positive" (Kirschvink, 1989). However, bees having sIRM values of 2×10^{-6} emu, when subjected to a very strong (210 mT) 60-Hz alternating demagnetizing field (Gould *et al.*, 1980), lost remanence to background levels $(0.01\times10^{-6}$ emu). This result suggests that most, if not all of the permanently magnetized material within bees is in the form of noninteracting SD or MD particles [i.e., it does not exist as strings of SD particles as found in bacteria, tuna, or salmon (see below) which resist demagnetization due to anisotropy]. Demagnetized bees showed no reduced ability to orient in the geomagnetic field as evidenced from analysis of their waggle dance (Gould *et al.*, 1980). Because bees may use different magnetic information as a compass from that which might be used as a map or in assessing time, it would be interesting to determine whether after attempts to demagnetize them, they subsequently danced on horizontal comb in the cardinal points of the compass or showed evidence of altered diurnal rhythmicity in behavior.

Natural magnetic remanences significantly above background values have been measured in other insect species (Table 1 and references therein). Whether this indicates the presence of biogenic magnetic material has not been adequately addressed. The wings of butterflies were magnetically very "dirty" due presumably to adventitious magnetic material (Jones and MacFadden, 1982), although head/thorax magnetism developed during pupation when nothing could enter the insect. No link between the presence of magnetite and ability to orient in the earth's field has been firmly established for any insect, although the results of Walker and Bitterman (1989a, 1989b) are consistent with this possibility.

3.5 Amphibians

Results obtained with male Eastern Red Spotted Salamander newts (*Notophthalmus viridescens*) provide fascinating evidence that these animals have more than one type of geomagnetic response. The data (Phillips, 1986) are consistent with two magnetic compass responses, one axial, one polar, depending upon whether the newts are appropriately induced by water temperature history to display either shoreward positional alignment or homing behavior.

3.6 Birds

The literature addressing magnetic influences on bird behavior is older and more voluminous than that for any other animal group, perhaps as a consequence of longstanding human interest in bird migration and pigeon homing (Keeton, 1974). The fact that the scientific literature extends back more than 40 years (Yeagley, 1951) attests to the relative ease with which bird magnetic responsiveness can be demonstrated (e.g., by attaching magnets to them); the fact that it is voluminous attests to the fact that simple statements have rarely been made (or uncontested). Birds (and bees) are behaviorally sophisticated animals which respond to a variety of hierarchically arranged sensory inputs including those of a solar, celestial, visual, olfactory, and magnetic nature (Southern, 1988; Moore, 1980; Gould, 1980a, 1982; Wiltschko *et al.*, 1981; Kiepenheuer, 1984). Different field parameters, including the overall dipolar geomagnetic field direction, the dip angle, change in dip angle, and anomalies in field uniformity in space or time are used by birds of various species or ontogenetic stages and at seasonally different times. Finally, whether individual birds exhibit behavior which can be correlated with geomagnetic sensory input can depend upon their previous experience (Lednor and Walcott, 1988), i.e., birds learn.

Convincing early experiments with bar magnets (Keeton, 1971, 1972) or paired, energized coils (Walcott and Green, 1974; Visalberghi and Alleva, 1979) attached to pigeon heads revealed that homing is disrupted (e.g., pigeon vanishing bearings are altered, although they may subsequently home correctly) by the resulting anomalous field which deprives the released bird of access to normal dipolar geomagnetic information. But these disturbances can be demonstrated best in young birds and on overcast days (Keeton, 1974) when reference to solar cues is abandoned [although even under sun, young birds cannot orient properly if fitted with disruptive magnets on their first flight (Keeton, 1971)]. Restrained pigeons cannot be trained by otherwise suitable conditioning methods to use magnetic information of geomagnetic intensity (Kriethen and Keeton, 1974), suggesting that bird motion is important for magnetic sensing. The first migratory excursions made by young birds may depend for success upon magnetic stimuli (Southern, 1972b, 1988; Alerstam and Högstedt, 1983; see also Bingman, 1984) whereas birds appear to learn subsequently to use other cues. For instance, although pigeons under sun show preference for using their solar compass, and nocturnal migrants a stellar compass, to be able to do so naive birds first calibrate it with their magnetic compass (Wiltschko *et al.* 1981; Wiltschko *et al.*, 1983). These and other results indicate that the bird magnetic sense is innate and may be a primary reference from which celestial (solar, stellar) compasses are calibrated in the manner of imprinting at an early age.

Analysis of premigrational restlessness (*Zugunruhe*) of caged European robins (Wiltschko and Wiltschko, 1972; Wiltschko, 1972) indicated that these birds do not respond to the absolute compass bearing but, instead, detect and use as "north" the angle of magnetic dip (e.g., the minimum angle between the gravity vector and the vertical component of the geomagnetic field). Similar findings were obtained in field studies with homing pigeons (Walcott and Green, 1974) who showed that birds with head coils energized to produce upward-directed fields (e.g. "north up" or "Nup") had, on cloudy days, vanishing bearings directed opposite to home and also to those of "Sup" pigeons with coils directing the field downward. This use of axial rather than polar north is identical to the manner in which Eastern Red Spotted Salamander newts use geomagnetic information (Phillips, 1986). Outdoors, however, European robins use (as yet unidentified) cues integrated with magnetic information in migration (Sandberg *et al.*, 1988). During nocturnal migration, they use their magnetic compass as a primary reference to select the seasonally appropriate migration direction and then switch to a stellar compass for maintaining accurate orientation (Wiltschko and Wiltschko, 1975, 1978). The fact that their axial compass does not discriminate magnetic north from south appears not to be relevant because they are not transequatorial migrants (for them, axial north always coincides with polar north).

Pigeons have magnetic tissues (Walcott *et al.*, 1979). Walcott *et al.* detected significant remanence in the dura and Presti and Pettigrew (1980) reported high sIRM values [$(10^{-6}-1.5)\times10^{-5}$ emu] in heads and necks of birds. However, these results have not been consistently reproduced (Walcott and Walcott, 1982). Presti and Pettigrew reported orange and black tissue deposits (magnetite is the only black iron oxide) which were not localized but diffuse throughout the head and neck muscles. Tissues provided a magnetometry signal consistent with the presence of about 10 ng of 70-nm SD magnetically interacting (stringed) magnetite cubes. The authors speculated on the presence of an axial magnetic detector based upon a muscle spindle (stretch or compression) type receptor. Attempts by others to detect significant remanences in diverse migratory birds (Ueda *et al.*, 1982) were unsuccessful, undoubtedly because of the high background associated with the SQUID employed.

The bobolink (*Dolichonyx oryzivorus*) migrates over the longest transequatorial route of any bird species. Behavioral responses of wild bobolinks and those reared in captivity correlated with reversal of the geomagnetic field (Beason and Nichols, 1984). The birds' preferred headings were analyzed in a planetarium equipped for reversal of the geomagnetic field enabling magnetic north to be dissociated from celestial north. In contrast to the behavior exhibited by European robins, (Wiltschko, 1972) these transequatorial migrants responded to polar rather than axial north. NRM and sIRM values of frozen or chemically fixed bird heads were 1.9×10^{-5} emu (1.9×10^{4} pA m^2) and 4.17×10^{-5} emu, respectively. Similar sIRM values (5×10^{-5} emu) as for entire heads were obtained with tissue from the post-nasal cavity. Heads continued to gain remanent magnetization with exposure to strong inducing fields of greater than 100 mT, up to a maximum at 300 mT. These results are consistent with the interpretation that the heads contained SD or small multidomains (pseudosingle domains) of magnetite. The Curie temperature was not determined nor was the Lowrie-Fuller test applied, nor has Mössbauer characterization been carried out because of the limited amount of material per bird. However, dissection and histochemical staining and x-ray spectra obtained with a scanning-tranmission electron microscope (STEM) revealed iron-rich deposits situated in sheaths around the olfactory nerve and in bristles projecting into the nasal cavity and intimately associated with nerve tissue. If the olfactory-associated iron is indeed biogenic as opposed to adventitious, and if it is associated with magnetic sensing, this finding may have important implications for studies by others implicating olfactory cues in bird orientation (see discussion by Beason and Nichols, 1984). The amount detected is theoretically (Yorke, 1979, 1981) within the range required to detect changes of less than 1% in the geomagnetic field. Thus, bobolinks may have enough sensitivity for a magnetic map sense and much more than that needed for their polar compass.

3.7 Primitive Fishes

Electroreception is known in chondrichthian (cartilaginous) fishes including the agnathans such as the anadromous lamprey *Lampetra tridentata* (Bodznick and Northcutt, 1981) but especially in elasmobranchs (sharks, skates, and stingrays), which use this sense in prey detection (Kalmijn, 1977, 1978, 1982). Chondrostean (primitive bony) fishes also possess electroreceptive capabilities (Fields and Lange, 1980). Cardiac decelerations and other electrophysiological evidence of sensitivity evoked by voltage gradients as low as 0.01 μV/cm (dc up to 8 Hz) were obtained in skates (Kalmijn, 1971) and fields estimated to be 0.2 μV/cm elicited a conditioned escape response in the ratfish, *Hydrolagus colliei* (Fields and Lange, 1980). These latter holocephalan creatures from the deep benthos possess ampullar organs presumably similar in function to the ampullae of Lorenzini, the highly sensitive electroreceptors of elasmobranchs. The electroreceptive ability of fish in seawater provides a theoretical basis for considering that they might also be geomagnetically responsive through the Faraday effect (Rommel and McCleave, 1972, 1973; Akoev *et al.*, 1976; Kalmijn, 1971, 1981, and references therein). This possibility was pursued by A. Kalmijn with V. Kalmijn, who conditioned captive stingrays (*Urolophus halleri*) to respond to reversal of the horizontal geomagnetic component. These animals, and perhaps other electroreceptive marine species, may thus detect the geomagnetic field by detecting induced voltage gradients in either active (swimming) or passive (drifting in currents) situations (Kalmijn, 1984). Field studies addressing the remaining question of whether induction or other means of geomagnetic orientation (sharks also contain magnetite) operate in free-roaming sharks have not been carried out.

3.8 Bony Fishes

The migratory feats of anadromous teleost fishes have made them especially good targets for inquiries of geomagnetic responsiveness (Quinn, 1982). Juvenile sockeye salmon (*Oncorynchus nerka*) without conditioning showed directional preferences in tank experiments which (in the absence of visual cues) were geomagnetically biased (Quinn *et al.* 1981; Quinn and Brannon, 1982). Similarly, evidence for geomagnetic-based orientation or navigation has been obtained for Atlantic salmon, *Salmo salar* (Rommel and McCleave, 1973), and chum salmon, *O. keta* (Quinn and Groot, 1983). Eels, *Anguilla rostrata* (Rommel and McCleave, 1972, 1973) and *A. anguilla* (Tesch, 1974), and the nonanadromous salmonid, the rainbow trout, *Salmo gairdneri* (Chew and Brown, 1989), also showed unconditioned orientational biases correlating with the Earth's magnetic-field direction.

Walker (1984) observed unconditioned avoidance responses to unanticipated magnetic fields presented to actively swimming, captive, juvenile yellowfin tuna (*Thunnus albacares*). He subsequently succeeded in training these fish to discriminate changes in Earth-strength magnetic fields reproducibly. The training methods used indicated neural processing of magnetic information, most likely by a magnetite-based transduction mechanism (Walker, 1984). Walker *et al.* (1984) inventoried tuna tissue and organ samples for magnetic remanence using a SQUID situated in facilities yielding a background moment of merely 5 pA m^2. Greatest values of remanence (>1000 pA m^2) were detected in eye tissue and sinuses within dermethmoid bones of the fish head. Alternating field demagnetization and sIRM data were consistent with the presence of approximately 8.5×10^7 single-domain magnetite particles in the dermethmoid tissue. The isolated crystals had characteristics of 45×38 nm^2 (SD) magnetite grains as determined by microscopic, electron microprobe, and x-ray-diffraction methods. The particles could provide sufficient torque for geomagnetic alignment only if they aggregated appropriately (e.g., preferably in strings) such that the product of M (the summed moments of stringed particles) with H (the geomagnetic field) exceeded kT by about a factor of 2–10 (see Frankel and Blakemore, 1980; Frankel, 1984) *and* if, in these large organisms, they were components of a sensory device equipped for neuronal amplification (Kirschvink and Gould, 1981). Walker *et al.* (1984) suggested that strings of 20–60 such particles would be sufficient for the fish to resolve magnetic-field intensity differences of 1–100 nT.

Walker *et al.* (1988), using magnetometry, obtained evidence consistent with the presence in sockeye salmon of interacting SD Fe_3O_4 grains, particularly in the eye and ethmoid tissue. Eye tissue was found to lack SD particles through its reaction in a modified Lowrie-Fuller test. Mann *et al.* (1988) confirmed the presence in sockeye ethmoid tissue of organically enveloped SD cubo-octahedral single crystals of Fe_3O_4 arranged in strings (median number per string of 21–25 particles) and arranged with their $\{111\}$ faces adjacent and lying perpendicular to the string axis. In each of these respects the magnetic material from salmon was identical to that described previously within cells of the bacterium *A. magnetotacticum*! Values of the ratio of magnetic (MH) to thermal (kT) energies for strings clustered bimodally: one at 2.7–5.3 and another at 6.6–9.5. The authors suggested that this might be indicative of a dual role for stringed magnetite: both as a compass and as a magnetic intensity detector. A rationale for magnetite-based orientation as developed by others (Yorke, 1979, 1981; Bean, 1979; Frankel and Blakemore, 1980; Frankel, 1982, 1984) is in support of their conclusions as to the suitability of this material in its quantity, size, and arrangement for geomagnetic sensing.

Single-domain magnetic particles have also been detected in chinook salmon, *O. tshawytscha* (Kirschvink *et al.*, 1985b). Alternating field demagnetization and sIRM data were consistent with magnetically interacting SD magnetite grains arranged in strings within dermethmoid cartilage tissue. Indeed, magnetic particles extracted from tissues were in strings, possibly with attendant organic material as a sheath or envelope. Thus, the magnetic material detected in this salmonid appeared similar in nearly all respects to that from sockeye salmon and yellowfin tuna (and the bacterium *Aquaspirillum magnetotacticum*).

The ability of some fishes to sense experimental fields of geomagnetic character is clearly established. Of the several mechanisms by which diverse species may do so (Faraday effect, permanent magnetism), models incorporating biogenic magnetite have been supported by the bulk of published work. While this does not prove their validity, it does provide a broad database upon which an inferential link between magnetite-based reception and animal geomagnetic responsiveness may eventually be further forged or broken.

3.9 Mammals

Zoeger *et al.* (1981) studied dead, stranded Pacific dolphins (*Delphinus delphis*) in which they identified strongly magnetized tissue (magnetic moment 2×10^{-8} A m^2, 20 times that of the Earth's field). Magnetism was localized to innervated dura mater tissues of the animal's skull (not consistently located from animal to animal). Alternating field demagnetization of the tissue NRM at low field strength indicated that the material contained anisotropic but magnetically "soft" (easily demagnetized or remagnetized) MD magnetite.

Other specimens appeared to lack this but possessed instead what appeared to be superparamagnetic or near-superparamagnetic SD magnetic grains. Zoeger *et al.* (1981) speculated that despite its magnetic "softness" the material might function as a magnetic detector, by deforming in the field, for instance.

After reviewing the U.S. Geological Survey East Coast aeromagnetic data and the location of 212 confirmed cetacean live strandings within the same region, Kirschvink *et al.* (1986) concluded that whales beach themselves at locations where weakly magnetic local anomalies representing minima in intensity and having lower than background gradients intersect the coastline. Their study corroborated observations of Klinowska (1985, 1986), who concluded that live strandings in Great Britain have tended to occur where local minima or "valleys" in the total intensity field cross the coastline or during geomagnetic disturbances, and that whales actively avoid areas of locally higher magnetic intensity.

A report (Baker, 1980) that British college students displaced tens of kilometers were geomagnetically influenced in their choice of selecting the direction toward home was unable to be repeated in the United States (Gould and Able, 1981) or corroborated by numerous others elsewhere (see Baker, 1987a). Baker (1987a) reviewed available relevant data (excluding that of a perpetrated hoax) and concluded that published data conflict with published interpretations of the data—thus keeping alive the controversy concerning whether humans use a nonvisual or magnetic sense in orientation information. Magnetic remanences of 2.3×10^{-6} emu g^{-1} (2.3×10^{3} pA m^{2} g^{-1}) in pineal tissue and approximately 1.6×10^{-5} emu g^{-1} in sphenoid-ethmoid tissue (roughly $10\times$background) have been measured in human bones, and histological staining for ferric iron revealed a continuous iron-rich layer situated 5 μm below the surface of the sphenoid sinus (Baker *et al.*, 1983). Kirschvink earlier (1981b) had reported magnetic remanence associated with human adrenal tissue. Additional papers of interest are found in Kirschvink, Jones, and MacFadden (1985a).

4. WHAT IS THE STATUS OF BIOLOGICAL MAGNETIC SENSING?

The examples cited above are sufficient to illustrate the diverse manner in which organisms may obtain and use geomagnetic information. In addition to its possible use in primary sensing (compasses, maps, clocks) it is possible that geomagnetism may be used by organisms to calibrate clocks or maps based primarily upon celestial, olfactory, or other cues. Biomagnetism appears to occupy a central place in current conjecture regarding such sensing.

4.1 Biological Compasses

Evidence consistent with avian compasses based upon solar, stellar, and magnetic information has been obtained. To show that geomagnetism is used by birds has not, however, been easy. The avian magnetic compass is probably innate and may serve to calibrate celestial compasses through imprinting during the first weeks or months of life; after this it may no longer be a primary reference. Moreover, nocturnally migrating birds appear to use their magnetic compass to choose overall flight direction initially and subsequently switch to stellar compass cues for navigation. Magnetic biocompasses can be sensitive to cardinal direction of the geomagnetic dipolar field (polar compass), to the angle of dip (axial compass), or possibly (Kiepenheuer, 1984) to apparent dip angle (change in dip with respect to a fixed reference such as gravity). The inclination-sensitive types do not allow the animal to discriminate magnetic north from south. At least some animals appear to have both polar and axial magnetic compasses. This could be of particular value to transequatorial migrants (bobolinks) but has also been found in Eastern Red Spotted Salamander newts.

Two postulated biocompass mechanisms have persisted. One is active electroorientation through the Faraday effect (Rommel and McCleave 1972, 1973; Akoev *et al.*, 1976; Kalmijn, 1971, 1981, and references therein, 1984; Jungerman and Rosenblum, 1980), which requires that an animal possess numerous, readily detectable, highly specialized and sensitive electroreceptors. It is a dH/dt type of system in which changes occasioned by the animal's movement with respect to the geomagnetic field are necessary (Kalmijn, 1984). It would work at a required level of sensitivity for geomagnetic compass information only if the return loops had diameters of millimeter scale, and if these paths were in a medium of low electrical resistance such as seawater (Jungerman and Rosenblum, 1980). Thus, it is difficult to conceive of this as a mechanism for birds or even for freshwater fishes, nor would it explain sensitivity to small temporal or spatial geomagnetic field fluctuations that some animals appear to sense.

The second animal biocompass mechanism, as yet unproven but given credibility by the bacterial paradigm (Frankel and Blakemore, 1980, 1989) of SD intracellular ferrimagnetic particles (see also Yorke, 1979, 1981; Bean, 1979; Frankel, 1982, 1984; Kirschvink, 1982, 1983, 1989; Kirschvink and Gould, 1981; Kirschvink, Jones, and MacFadden, 1985a), has been incorporated into diverse animal biocompass models. These include muscle spindle stretch receptors (Presti and Pettigrew, 1980) and proposals by Kirschvink and Gould (1981), which include a "membrane short model," "torque detectors," "superparamagnetic (sympathetic satellite, linear and lattice array) detectors," and "hybrid detectors." A navigational biocompass consisting of stringed or aligned SD particles is proven for magnetotactic bacteria and conceptually valid for animal compass sensing. For organisms other than the bacteria, a conclusive link between the presence of a biogenic mineral phase and geomagnetically influenced compass behavior has not been made.

Consequently, it would be at least premature to focus on ferrimagnetism to explain animal geomagnetic sensitivity to the exclusion of still other fascinating proposals even though the latter are unsupported by experimental data and some seem physical impossibilities given the constraints imposed by conditions within cells or organisms. Such alternatives have included postulated biological (room temperature) superconductors (Cope, 1973), optical pumping (Leask, 1977), Lorentz forces in cytoplasm (Ruhenstroth-Bauer *et al.*, 1987), and biradical chemical reactions induced by anisotropic hyperfine interactions coupled to external magnetic fields (Schulten and Windemuth, 1986). Integration of magnetic detectors with other sensory systems including gravity detectors (Kiepenheuer, 1984; Wiltschko, 1980), the visual system (Semm *et al.*, 1984; Leucht, 1984), or the vestibular organs (Semm *et al.*, 1984) is a distinct possibility. Interesting implications of such mixed sensing are

1. that animals might possess redundant compasses while minimizing structural complexity,
2. that animals might, in effect, "see" or "smell" magnetic stimuli registered as perturbations in these other sensory perceptions, and
3. that disruption of one sense or organ by an experimenter (as has been done in vestibular or olfactory studies with pigeons) might concomitantly prevent magnetic sensing.

4.2 Magnetic Maps

The nature of bird navigational maps (Gould, 1982; Moore, 1980; Ossenkopp and Barbieto, 1978) is incompletely understood and controversial. Olfactory information may comprise the basis for a navigational map outside the local familiar area (Wallraff and Neuman, 1989) whereas nonolfactory, visual information (sun compass, local topography) may be used in pilotage behavior in familiar territory (Wiltschko and Wiltschko, 1989). That the "magnetic topography" can be important to birds is evidenced by the inability of faithful birds to home normally or at all when released for the first time at certain magnetically anomalous locales (Walcott, 1978, 1980; Kiepenheuer, 1984), particularly if they are mature birds not raised there (Lednor and Walcott, 1988). Compass orientation of European robins was unaffected by field intensity differences of 10% (Wiltschko, 1972). Thus, the "map sense," which appears to depend upon minor differences in intensity or field gradients, would appear to involve magnetic reception mechanisms (Yorke, 1981) different from that involved as a compass (see Yorke, 1979). Walker *et al.* (1986) have presented a mathematical model for a compass which would not require a second frame of reference other than the geomagnetic field. Additional experimental studies of the sort conducted by Walcott *et al.* (1988), who studied homing in magnetized and demagnetized pigeons, are needed.

4.3 Magnetism-Based Clocks

There are a number of biological examples in which magnetism appears to play a role in time-keeping. Lunar periodicities have been correlated with the magnetically influenced behavior of invertebrates including snails (Brown, Webb, and Brett, 1960), the sea slug *Tritonia* (Lohmann and Willows, 1987), and a moth (Baker, 1987b). Bees and birds are also sensitive to normal, 30–50 nT, diurnal magnetic fluctuations and to pulses associated with unpredictable magnetic storms (Keeton, Larkin, and Windsor, 1974; Southern 1972b; Martin and Lindauer 1977). The suggestion (Neumann, 1988) that bees might use fluctuating, diurnal, magnetic periodicities directly in time-keeping functions has interest and appears conceptually valid in the light of their extreme geomagnetic sensitivity. No study has revealed just what such a putative magnetic-field-dependent clock might consist of.

5. THE ONGOING STUDY OF BIOLOGICAL GEOMAGNETIC DETECTORS

It is somewhat ironic that 15 years after their discovery, the magnetotactic bacteria remain as the only biological system for which there is a proven mechanism for the geomagnetic response even though the basic feature of their mechanism (intracellular strings of SD biogenic magnetite) figures into most extant models for geomagnetic compass and map sensing by animals. Why this should be so is obvious; animals cannot be "torqued" into position by the geomagnetic field and so would require not only magnetic detector(s) but neuronal amplification and transmission of transduced information prior to its involvement at an appropriate effector organ. A successful hunt would require coordinated behavioral, anatomical, and cell-tissue-organ (electro)physiology studies not representative of the way modern science usually functions (or, at least, is funded). Several lines of experimental evidence could provide direct evidence of a putative magnetic sensor, including

1. surgical removal of suspect tissue from trained animals,
2. analysis of the effect of disrupting an associated sensory (visual, olfactory, vestibular, pineal) system or organ, and
3. electrophysiology in altered magnetic fields.

These will not be trivial pursuits. The reader is reminded that prokaryotes are more equivalent to subcellular organelles than to eukaryotic cells themselves. Thus, the search is for a bacterial-type magnetic apparatus ("magnetic needle") dimensionally equivalent to a few submitochondrial-sized particles possibly within a single animal cell buried in a haystack comprised of an overwhelming mass of tissue! If this is the case, not even magnetometry will disclose the location of such a detector. Frankel (1982) indicated that merely three 50-nm magnetosomes would be needed per cell (e.g., for $\mu H/kT$ to >2) to bring about bacterial rotation in the geomagnetic field within an orientationally meaningful time interval. Models such as the "membrane short model" of Kirschvink and Gould (1981), as presented, require a single, SD magnetite grain of subcellular dimensions. Anatomists, then, may hope that the magnetic sense of animals is so vital that it has been constructed with much redundancy while physiologists would welcome a distinct, localized detector functioning independently of other senses. If animals do indeed use the basic bacterial paradigm for magnetic sensing, then unless it has evolved into a larger, more complex structure than has previously been imagined, locating the detector in the "background" of cell and tissue mass without destroying it may be unlikely. Quite possibly, the use of genetic probes developed from molecular elements of the bacterial compass may prove more effective than tissue disruption in locating the putative animal magnetic detector.

ACKNOWLEDGMENTS

The long-term interest, encouragement, support, and contributions of Dr. Richard Frankel and Nancy Blakemore to studies of magnetogenic bacteria in the author's laboratory are gratefully acknowledged. These studies have been made possible principally by support from the Office of Naval Research and the National Science Foundation. However, funding from the University of New

Hampshire Research Office, the National Geographic Society, and BioMagnetech Corporation is also gratefully acknowledged.

GLOSSARY

Anhysteretic Remanent Magnetization (ARM): Magnetic remanence imparted to a sample by applying an alternating field which is then reduced to zero (ac demagnetization) in the presence of a constant, steady dc field.

Biomineralization: The production of mineral phases within living organisms. Diverse prokaryotic organisms may produce intracellular or extracellular biominerals, including magnetic substances, in which case the organism is said to be *magnetogenic*. Magnetogens producing intracellular biominerals may be *magnetotactic*; that is, they respond in their swimming direction to changes in the ambient geomagnetic field.

Biomagnetism: The property exhibited by cells or organisms of possessing an intrinsic permanent magnetic dipole character due to the presence of an endogenous (in contrast to adventitious) magnetic substance. In all known cases biomagnetism is conferred by the presence of a magnetic biomineral. Two types of these are presently known: magnetic iron oxides (widespread among life forms) and magnetic iron sulfides (confined to certain bacteria).

Curie Temperature: The temperature above which exchange coupling (forces aligning atomic dipoles in a parallel fashion, despite kT) in a ferromagnetic material is lost and the substance becomes paramagnetic (e.g., it loses its permanently magnetized character). For iron the temperature is 1043 K (770 °C) and for magnetite it is 858 K (585 °C).

Dissimilatory Iron Reduction: The process in which organisms oxidize cell energy sources using Fe^{3+} in lieu of, or in addition to, O_2 as a terminal electron acceptor.

emu: "Electromagnetic unit"; $1 \text{ emu} = 1 \text{ erg G}^{-1} = 1\ \mu\text{J T}^{-1} = 10^9 \text{ pA m}^2$.

Eukaryote: An organism belonging to one of the three major, distinct lines of evolutionary descent and consisting of, or having, nucleated cells. Included are traditional groups referred to as fungi, algae, plants, animals, etc., but not prokaryotes (bacteria).

Ferromagnetism: The property shared by iron, cobalt, nickel, gadolinium, dysprosium, and alloys of these, of forming magnetic domains within bulk material as a consequence of exchange coupling. Exchange coupling interactions (a quantum effect) serve to align atomic dipoles in a common direction despite the randomizing tendency of the atoms due to kT. "Hard" ferromagnetic materials have high permeability and retain significant magnetization upon removal of a strong magnetizing field. "Soft" ferromagnetic materials lose permanent magnetism to a considerable degree after removal of a magnetizing field.

Gamma: An older term referring to units of geomagnetism. $1\gamma = 1 \text{ nT} = 10^{-5} \text{ G}$.

Greigite: A magnetic iron biomineral produced by some magnetogenic bacteria and having the empirical formula Fe_3S_4.

Magnetite: A ferrimagnetic mixed-valence iron mineral of the inverse spinel series which can be a biomineral produced either intracellularly or extracellularly by prokaryotic and eukaryotic organisms. It has the empirical formula, $Fe_2O_3{\cdot}FeO$, more commonly expressed Fe_3O_4.

Magnetization (J_s): Magnetic moment (μ) per unit volume (V).

Magnetogen: A microorganism which produces a magnetic (iron) biomineral. Some magnetogenic bacteria produce such minerals intracellularly, others extracellularly. The minerals include magnetic iron oxide (magnetite) or magnetic iron sulfide (greigite). Known magnetogens appear to be dissimilatory iron reducers.

Magnetosome: A biogenic magnetic particle found within an organism or its cells and consisting of a crystalline, magnetic, mineral core, usually within an envelope. In the one case examined to date, the envelope comprises a lipid bilayer-membrane vesicle. Magnetosomes of magnetotactic microbes are commonly arranged in an intracellular string of from several to hundreds of particles per cell.

Magnetotactic: A behavior exhibited by organisms in which their swimming or motion is directed by a magnetic field. The term is conventionally applied to microorganisms and to fields of geomagnetic intensity.

Natural Remanent Magnetization (NRM): The magnetic remanence of a natural substance (rock or cell) in the absence of other magnetizing or demagnetizing fields.

Prokaryotes: Organisms existing as independent cells or cell aggregates but lacking tissue grade of organization. Prokaryotic cells lack true nuclei and differ in other fundamental ways from eukaryotic cells. Prokaryotes include the eubacteria and archaebacteria.

Pyrite: A nonmagnetic iron sulfide biomineral produced by some magnetogenic bacteria and having the empirical formula FeS_2.

Pyrrhotite: A magnetic iron sulfide biomineral which may be produced by some magnetogenic bacteria and which has the empirical formula Fe_xS_{1+x}; e.g., Fe_7S_8.

Saturation Isothermal Remanent Magnetization (sIRM): The magnetic remanence left in a sample of ferromagnetic material after brief exposure to a steady field of up to 10 000 G (800 000 A m^2).

Superconducting Quantum Interference Device (SQUID): A cryogenic (4 K) magnetometer containing one or more loops of superconducting material, each interrupted by a thin insulating barrier to form a "Josephson tunnel junction," and used to detect small amounts of magnetic flux on the order of 2×10^{-15} T m^2 such as in magnetized rocks or cells.

Zeitgeber: An environmental stimulus that serves to "set" a biological clock.

List of Works Cited

Akoev, G. N., Ilyinsky, O. B., Zadan, P. M. (1976), "Responses of electroreceptors (Ampullae of Lorenzini) of skates to electric and magnetic fields," *J. Comp. Physiol.* **106**, 127–136.

Alerstam, T., Högstedt, G. (1983), "The role of the geomagnetic field in the development of birds' compass sense," *Nature (London)* **306**, 463–465.

Baker, R. R. (1980), "Goal orientation by blindfolded humans after long-distance displacement: possible involvement of a magnetic sense," *Science* **210**, 555–557.

Baker, R. R. (1987a), "Human navigation and magnetoreception: the Manchester experiments do replicate" *Animal Behav.* **35**, 691–704.

Baker, R. R. (1987b), "Integrated use of moon and magnetic compasses by the heart-and-dart moth *Agrotis exclamationis*," Animal Behav. **35**, 94–101.

Baker, R. R., Mather, J. G., Kennaugh, J. H. (1983), "Magnetic bones in human sinuses," *Nature (London)* **301**, 78–80.

Balkwill, D. L., Maratea, D., Blakemore, R. P. (1980), "Ultrastructure of a magnetotactic spirillum," *J. Bacteriol.* **141**, 1399–1408.

Barnothy, M. F. (1964), *Biological Effects of Magnetic Fields*, New York: Plenum.

Bazylinski, D. A., Frankel, R. B., Jannasch, H. W. (1988), "Anaerobic magnetite production by a marine, magnetotactic bacterium," *Nature (London)* **334**, 518–519.

Bean, C. P. (1979), "Single domain particles and magnetotatic (*sic*) bacteria," *Biophys. J.* **25**, 216a.

Bean, C. P. (1989), "Motion of magnetotactic bacteria in a rotating magnetic field," *Biophys. J.* **54**, 258a.

Beason, R. C., Nichols, J. E. (1984), "Magnetic orientation and magnetically sensitive material in a transequatorial migratory bird," *Nature (London)* **309**, 151–153.

Bell, P. E., Mills, A. L., Herman, J. S. (1987), "Biogeochemical conditions favoring magnetite formation during anaerobic iron reduction," *Appl. Environ. Microbiol.* **53**, 2610–2616.

Bingman, V. P. (1984), "Night sky orientation of migratory pied flycatchers raised in different magnetic fields," *Behav. Ecol. Sociobiol.* **15**, 77–80.

Blakemore, R. P. (1975), "Magnetotactic bacteria," *Science* **190**, 377–379.

Blakemore, R. P. (1982), "Magnetotactic bacteria," *Annu. Rev. Microbiol.* **36**, 217–238.

Blakemore, R. P., Frankel, R. B. (1981), "Magnetic navigation in bacteria," *Sci. Am.* **254**, 58–65.

Blakemore, R. P., Frankel, R. B. (1989), "Biomineralization by magnetogenic bacteria," in: R. K. Poole and G. M. Gadd (Eds.), *Metal-Microbe Interactions*, Oxford: IRL, Chap. 6, pp. 85–98.

Blakemore, R. P., Maratea, D., Wolfe, R. S. (1979), "Isolation and pure culture of a freshwater magnetic spirillum in chemically defined medium," *J. Bacteriol.* **140**, 720–729.

Blakemore, R. P., Frankel, R. B., Kalmijn, A. J. (1980), "South-seeking magnetotactic bacteria in the southern hemisphere," *Nature (London)* **286**, 384–385.

Blakemore, R. P., Short, K. A., Bazylinski, D. A., Rosenblatt, C., Frankel, R. B. (1984), "Microaerobic conditions are required for magnetite formation within *Aquaspirillum magnetotacticum*," *Geomicrobiology J.* **4**, 62–66.

Blakemore, R. P., Blakemore, N. A., Frankel, R. B. (1988), "Bacterial biomagnetism and geomagnetic field detection by organisms," in: A. A. Marino (Ed.), *Modern Bioelectricity*, New York: Marcel Dekker, Chap. 2.

Blakemore, R. P., Blakemore, N. A., Bazylinski, D. A., Moench, T. T. (1989), "The magnetotactic bacteria," in: J. T. Staley, M. P. Bryant, N. Pfennig, and

J. G. Holt (Eds.), *Bergy's Manual of Systematic Bacteriology*, Vol. 3. Baltimore: Williams and Wilkins, pp. 1882–1888.

Bodznick, D., Northcutt, R. G. (1981), "Electroreception in lampreys: evidence that the earliest vertebrates were electroreceptive," *Science* **212**, 465–467.

Brown, F. A., Jr. (1962), "Responses of the planarian, *Dugesia*, and the protozoan, *Paramecium*, to very weak horizontal magnetic fields," *Biol. Bull.* **123**, 264–281.

Brown, F. A., Jr., Webb, H. M., Brett, W. J. (1960), "Magnetic response of an organism and its lunar relationships," *Biol. Bull.* **118**, 382–392.

Brown, F. A., Jr., Webb, H. M., Barnwell, F. H. (1964), "A compass directional phenomenon in mud-snails and its relation to magnetism," *Biol. Bull.* **127**, 206–220.

Chew, G. L., Brown, G. E. (1989), "Orientation of rainbow trout (*Salmo gairdneri*) in normal and null magnetic fields," *Can. J. Zool.* **67**, 641–643.

Cisowski, S. (1981), "Interacting vs noninteracting single-domain behavior in natural and synthetic samples," *Phys. Earth Planet. Inter.* **26**, 56–62.

Cope, F. W. (1973), "Biological sensitivity to weak magnetic fields due to biological superconductive Josephson junctions?" *Physiol. Chem. Phys.* **5**, 173–176.

De Jong, D. (1982), "The orientation of comb-building by honeybees," *J. Comp. Physiol.* **147**, 495–501.

Denham, C. R., Blakemore, R. P., Frankel, R. B. (1980), "Bulk magnetic properties of magnetotactic bacteria," *IEEE Trans. Magn.* **MAG-16**, 1006–1007.

Dyer, F. C., Gould, J. L. (1981), "Honey bee orientation: a backup system for cloudy days," *Science* **214**, 1041–1042.

Eden, P. E., Schmidt, T. M., Blakemore, R. P., Pace, N. R. (1991), "Phylogenetic analysis of *Aquaspirillum Magnetotacticum* using polymerase chain reaction–amplified 16S rRNA-specfic DNA," *Int. J. Syst. Bacteriol.* **41**, 324–325.

Esquivel, D. M. S., Lins de Barros, H. G. P., Farina, M., Aragao, P. H. A., Danon, J. (1983), "Microorganisms magnetotactiques de la region de Rio de Janiero," *Biol. Cell.* **47**, 227–234.

Farina, M., Esquivel, D. M. S., Lins de Barros, H. G. P. (1990), "Magnetic iron-sulphur crystals from a magnetotactic microorganism," *Nature (London)* **343**, 256–258.

Fields, R. D., Lange, G. D. (1980), "Electroreception in the ratfish (*Hydrolagus colliei*)," *Science* **207**, 547–548.

Frankel, R. B. (1982), "Magnetotactic bacteria," *Comments Mol. Cell. Biophys.* **1**, 293–310.

Frankel, R. B. (1984), "Magnetic guidance of organisms," *Annu. Rev. Biophys. Bioeng.* **14**, 85–103.

Frankel, R. B., Blakemore, R. P. (1980), "Navigational compass in magnetic bacteria," *J. Magn. Magn. Mater.* 15–18, 1562–1564.

Frankel, R. B., Blakemore, R. P. (1989), "Magnetite and magnetotaxis in microorganisms," *Bioelectromagnetics* **10**, 223–237.

Frankel, R. B., Blakemore, R. P., Wolfe, R. S. (1979), "Magnetite in freshwater magnetotactic bacteria," *Science* **203**, 1355–1356.

Frankel, R. B., Blakemore, R. P., Torres de Araujo, F. F., Esquivel, D. M. S., Danon, J. (1981), "Magnetotactic bacteria at the geomagnetic equator," *Science* **212**, 1269–1270.

Frankel, R. B., Papaefthymiou, G. C., Blakemore, R. P., O'Brien, W. (1983), "Fe_3O_4 precipitation in magnetotactic bacteria," *Biochim. Biophys. Acta* **763**, 147–159.

Frankel, R. B., Papaefthymiou, G. C., Blakemore, R. P. (1985), "Mössbauer spectroscopy of iron biomineralization products in magnetotactic bacteria," in: J. L. Kirschvink, D. S. Jones, and B. J. MacFadden (Eds.), *Magnetite Biomineralization and Magnetoreception in Organisms: A New Biomagnetism*, New York: Plenum, Chap. 13.

Freke, A. M., Tate, D. (1961), "The formation of magnetic iron sulphide by bacterial reduction of iron solutions," *J. Biochem. Microbiol. Technol. Eng.* **3**, 29–39.

Fuller, M., Goree, W. S., Goodman, W. L. (1985), "An introduction to the use of SQUID magnetometers in biomagnetism," in: J. L. Kirschvink, D. S. Jones, and B. J. MacFadden (Eds.), *Magnetite Biomineralization and Magnetoreception in Organisms: A New Biomagnetism*, New York: Plenum, pp. 103–153.

Gorby, Y. A., Beveridge, T. J., Blakemore, R. P. (1987), "Characterization of the bacterial magnetosome membrane," *J. Bacteriol.* **170**, 834–841.

Goree, W. S., Fuller, M. (1976), "Magnetometers using RF-driven SQUIDS and their applications in rock magnetism and paleomagnetism," *Rev. Geophys. Space Phys.* **14**, 591–608.

Gould, J. L. (1980a), "The case for magnetic sensitivity in birds and bees (such as it is)," *Am. Sci.* **68**, 256–267.

Gould, J. (1980b), "Sun compensation by bees," *Science* **207**, 545–547.

Gould, J. L. (1982), "The map sense of pigeons," *Nature (London)* **296**, 205–211.

Gould, J. L., Able, K. P. (1981), "Human homing: an elusive phenomenon," *Science* **212**, 1061–1063.

Gould, J. L., Kirschvink, J. L., Deffeyes, K. S. (1978), "Bees have magnetic remanence," *Science* **201**, 1026–1028.

Gould, J. L., Kirschvink, J. L., Deffeyes, K. S., Brines, M. L. (1980), "Orientation of demagnetized bees," *J. Exp. Biol.* **86**, 1–8.

Heywood, B. R., Bazylinski, D. A., Garratt-Reed, A., Mann, S., Frankel, R. B. (1990), "Controlled biosynthesis of greigite (Fe_3S_4) in magnetotactic bacteria," *Naturwissenschaften* **77**, 536–538.

Jacobs, I. S., Bean, C. (1955), "An approach to elongated fine-particle magnets," *Phys. Rev.* **100**, 1060–1067.

Jones, D. S., MacFadden, B. J. (1982), "Induced magnetization in the monarch butterfly, *Danaus plexippus* (Insecta, Lepidoptera)," *J. Exp. Biol.* **96**, 1-9.

Jungerman, R. L., Rosenblum, B. (1980), "Magnetic induction for the sensing of magnetic fields by animals—an analysis," *J. Theor. Biol.* **87**, 25–32.

Kalmijn, A. J. (1971), "The electric sense of sharks and rays," *J. Exp. Biol.* **55**, 371–383.

Kalmijn, A. J. (1977), "The electric and magnetic sense of sharks, skates and rays," *Oceanus* **20**, 45–52.

Kalmijn, A. J. (1978), "Experimental evidence of geomagnetic orientation in elasmobranch fishes," in: K. Schmidt-Koenig and W. T. Keeton (Eds.), *Animal Migration, Navigation, and Homing: Symposium Held at the University of Tuebingen, August 17–20, 1977*, Berlin: Springer-Verlag, pp. 135–142.

Kalmijn, A. J. (1981), "Biophysics of geomagnetic-field detection," *IEEE Trans. Magn.* **MAG-17**, 1113–1124.

Kalmijn, A. J. (1982), "Electric and magnetic-field detection in elasmobranch fishes," *Science* **218**, 916–918.

Kalmijn, A. J. (1984), "Theory of electromagnetic orientation: a further analysis," in: L. Bolis, R. D. Keynes, and S. P. H. Maddrell (Eds.), *Comparative Physiology of Sensory Systems*, New York: Cambridge University, pp. 525–560.

Kalmijn, A. J., Blakemore, R. P. (1978), "The magnetic behavior of mud bacteria," in: K. Schmidt-Kenig and W. T. Keeton (Eds.), *Animal Migration, Navigation and Homing: Symposium Held at the University of Tuebingen, August 17–20, 1977*, New York: Springer-Verlag, pp. 354–355.

Keeton, W. T. (1971), "Magnets interfere with pigeon homing," *Proc. Natl. Acad. Sci.* **68**, 102–106.

Keeton, W. T. (1972), "Effects of magnets on pigeon homing," in: S. E Galler, K. Schmidt-Koenig, G. J. Jacobs, and R. E. Belleville (Eds.), *Animal Orientation and Navigation*, Washington, DC: U.S. GPO, NASA SP-262, pp. 579–594.

Keeton, W.T. (1974), "The mystery of pigeon homing," *Sci. Am.* **231**, 96–107.

Keeton, W. T., Larkin, T. S., Windsor, D. M. (1974), "Normal fluctuations in the earth's magnetic field influence pigeon orientation," *J. Comp. Physiol.* **95**, 95–103.

Kiepenheuer, J. (1984), "The magnetic compass mechanism of birds and its possible association with the shifting course directions of migrants," *Behav. Ecol. Sociobiol.* **14**, 81–99.

Kirschvink, J. L. (1980), "South seeking magnetic bacteria," *J. Exp. Biol.* **86**, 345–347.

Kirschvink, J. L. (1981a), "The horizontal magnetic dance of the honey bee is compatible with a single domain ferromagnetic magnetoreceptor," *BioSystems* **14**, 193–203.

Kirschvink, J. L. (1981b), "Ferromagnetic crystals (magnetite?) in human tissue," *J. Exp. Biol.* **92**, 333–335.

Kirschvink, J. L. (1982), "Birds, bees and magnetism: a new look at the old problem of magnetoreception," *Trends Neurosci.* **5**, 160–167.

Kirschvink, J. L. (1983), "Biogenic ferrimagnetism: a new biomagnetism," in: S. Williamson *et al.* (Eds.), *Biomagnetism: An Interdisciplinary Approach*, New York: Plenum.

Kirschvink, J. L. (1989), "Magnetite biomineralization and geomagnetic sensitivity in higher animals: an update and recommendations for future study," *Bioelectromagnetics*, **10**, 239–259.

Kirschvink, J. L., Gould, J. L. (1981), "Biogenic magnetite as a basis for magnetic-field sensitivity in animals," *BioSystems* **13**, 181–201.

Kirschvink, J. L., Lowenstam, H. A. (1979), "Mineralization and magnetization of chiton teeth: paleomagnetic, sedimentologic, and biologic implications of organic magnetite," *Earth Planet. Sci. Lett.* **44**, 193–204.

Kirschvink, J. L., Jones, D. S., MacFadden, B. J. (Eds.) (1985a), *Magnetite Biomineralization and Magnetoreception in Organisms: A New Biomagnetism*, New York: Plenum.

Kirschvink, J. L., Walker, M. M., Chang, S-B. R., Dizon, A. E., Peterson, K. A. (1985b), "Chains of single-domain magnetite particles in chinook salmon, *Oncorhynchus tshawytscha*," *J. Comp. Physiol. A* **157**, 375–381.

Kirschvink, J. L., Dizon, A. E., Westphal, J. A. (1986), "Evidence from strandings for geomagnetic sensitivity in cetaceans," *J. Exp. Biol.* **120**, 1–24.

Klinowska, M. (1985), "Cetacean live stranding sites relate to geomagnetic topography," *Aquat. Mammals* **11**, 27–32.

Klinowska, M. (1986), "Cetacean live stranding dates relate to geomagnetic disturbances,"*Aquat. Mammals* **11**, 109–119.

Kreithen, M. L., Keeton, W. T. (1974), "Attempts to condition homing pigeons to magnetic stimuli," *J. Comp. Physiol.* **91**, 355–362.

Kuterbach, D. A., Walcott, B., Reeder, R. J., Frankel, R. B. (1982), "Iron-containing cells in the honeybee (*Apis mellifera*)," *Science* **218**, 695–697.

Larkin, T. S., Keeton, W. T. (1976), "Bar magnets mask the effect of normal magnetic disturbances on pigeon orientation," *J. Comp. Physiol.* **110**, 227–231.

Leask, M. J. M. (1977), "A physiochemical mechanism for magnetic-field detection by migratory birds and homing pigeons," *Nature (London)* **267**, 144–145.

Lednor, A. J., Walcott, C. (1988), "Orientation of homing pigeons at magnetic anomalies: the effects of experience," *Behav. Ecol. Sociobiol.* **22**, 3–8.

Leucht, T. (1984), "Responses to light under varying magnetic conditions in the honeybee, *Apis mellifica*," *J. Comp. Physiol.* **154**, 865–870.

Lindauer, M. (1977), "Recent advances in the orientation and learning of honeybees," *Proc. XV Int. Congr. Entomol.* **15**, 450–460.

Lindauer, M., Martin, H. (1972), "Magnetic effects on dancing bees," in: S. R. Galler, K. Schmidt-

Koenig, G. J. Jacobs, and R. E. Belleville (Eds.), *Animal Orientation and Navigation*, Washington, DC: U.S. GPO, NASA SP-262, pp. 559–567.

Lohmann, K. J., Willows, A. O. D. (1987), "Lunar-modulated geomagnetic orientation by a marine mollusk," *Science* **235**, 331–334.

Lovley, D. R., Stolz, J. F., Nord, G. L., Jr., Phillips, E. J. P. (1987), "Anaerobic production of magnetite by a dissimilatory iron-reducing microorganism," *Nature (London)* **330**, 252–254.

Lowenstam, H. A. (1962a), "Magnetite in denticle capping in recent chitons (polyplacophora)," *Geol. Soc. Am. Bull.* **73**, 435–438.

Lowenstam, H. A. (1962b), "Goethite in radular teeth of recent marine gastropods," *Science* **137**, 279–280.

Lowenstam, H. A. (1967), "Lepidocrocite, an apatite mineral, and magnetite in teeth of chitons (Polyplacophora)," *Science* **156**, 1373–1375.

Lowenstam, H. A. (1981), "Minerals formed by organisms," *Science* **211**, 1126–1131.

Lowrie, W., Fuller, M. (1971), "On the alternating-field demagnetization characteristics of multidomain thermoremanent magnetization in magnetite," *J. Geophys. Res.* **76**, 6339–6349.

Mann, S., Frankel, R. B., Blakemore, R. P. (1984a), "Structure, morphology, and crystal growth of bacterial magnetite," *Nature (London)* **310**, 405–407.

Mann, S., Moench, T. T., Williams, R. J. P. (1984b), "A high-resolution electron microscopic investigation of bacterial magnetite: implications for crystal growth," *Proc. R. Soc. London Ser. B* **221**, 385–393.

Mann, S., Sparks, N. H. C., Blakemore, R. P. (1987a), "Ultrastructure and characterization of anisotropic magnetic inclusions in magnetotactic bacteria," *Proc. R. Soc. London Ser. B* **231**, 469–476.

Mann, S., Sparks, N. H. C., Blakemore, R. P. (1987b), "Structure, morphology and crystal growth of anisotropic magnetite crystals in magnetotactic bacteria," *Proc. R. Soc. London Ser. B* **231**, 477–487.

Mann, S., Sparks, N. H. C., Walker, M. M., Kirschvink, J. L. (1988), "Ultrastructure, morphology and organization of biogenic magnetite from sockeye salmon, *Oncorhynchus nerka*: implications for magnetoreception," *J. Exp. Biol.* **140**, 35–49.

Mann, S., Sparks, N. H. C., Frankel, R. B., Bazylinski, D. A., Jannasch, H. W. (1990), "Biomineralization of ferrimagnetic greigite (Fe_3S_4) and iron pyrite (FeS_2) in a magnetotactic bacterium," *Nature (London)* **343**, 258–261.

Maratea, D., Blakemore, R. P. (1981), "*Aquaspirillum magnetotacticum* sp. nov., a magnetic spirillum," *Int. J. Syst. Bacteriol.* **31**, 452–455.

Martin, H., Korall, H., Forster, B. (1989), "Magnetic field effects on activity and ageing in honeybees," *J. Comp. Physiol.* **164**, 423–431.

Martin, H., Lindauer, M. (1977), "Der Einfluss der Erdmagnetfelds und die Schwerorientierung der Honigbiene," *J. Comp. Physiol.* **122**, 145–187.

Mather, J. G., Baker, R. R. (1981), "Magnetic sense of direction in woodmice for route-based navigation," *Nature (London)* **291**, 152–155.

Matsuda, T., Endo, J., Osakabe, N., Tonomura, A. (1983), "Morphology and structure of biogenic magnetite," *Nature (London)* **302**, 411–412.

Merkel, F. W., Wiltschko, W. (1965), "Magnetismus und richtungsfinden zugunruhiger rotkehlchen (*Erthacus rubecula*)," *Vogelwarte* **23**, 71–77.

Mizota, M., Maeda, Y. (1983), "Single-domain particles in magnetotactic bacteria," *Annu. Rep. Res. React. Inst. Kyoto Univ.* **16**, 144–150.

Moench, T. T. (1988), "*Bilphococcus magnetotacticus* gen. nov. sp. nov., a motile, magnetic coccus," *Antonie van Leeuwenhoek* **54**, 483–496.

Moore, B. R. (1980), "Is the homing pigeon's map geomagnetic?" *Nature (London)* **285**, 69–70.

Moskowitz, B. M., Frankel, R. B., Flanders, P. J., Blakemore, R. P., Schwartz, B. B. (1988), "Magnetic properties of magnetotactic bacteria," *J. Magn. Magn. Mater.* **73**, 273–288.

Moskowitz, B. M., Frankel, R. B., Bazylinski, D. A., Jannasch, H. W., Lovley, D. R. (1989), "A comparison of magnetite particles produced anaerobically by magnetotactic and dissimilatory iron-reducing bacteria," *Geophys. Res. Lett.* **16**, 665–668.

Myers, C. R., Nealson, K. H. (1988), "Bacterial manganese reduction and growth with manganese oxide as the sole electron acceptor," *Science* **240**, 1319–1321.

Neumann, M. F. (1988), "Is there any influence of magnetic or astrophysical fields on the circadian rhythm of honeybees?" *Behav. Ecol. Sociobiol.* **23**, 389–393.

Ossenkopp, K-P., Barbeito, R. (1978), "Bird orientation and the geomagnetic field; a review," *Neurosci. Biobehav. Rev.* **2**, 255–270.

Perry, A., Bauer, G. B., Dizon, A. E. (1981), "Magnetite in the green turtl," *EOS* **62**, 850.

Petersen, N., von Dobeneck, T., Vali, H. (1986), "Fossil bacterial magnetite in deep-sea sediments from the South Atlantic Ocean," *Nature (London)* **320**, 611–615.

Phillips, J. B. (1986), "Two magnetoreception pathways in a migratory salamander," *Science* **233**, 765–767.

Presti, D., Pettigrew, J. D. (1980), "Ferromagnetic coupling to muscle receptors as a basis for geomagnetic field sensitivity in animals," *Nature (London)* **285**, 99–101.

Quinn, T. P. (1982), "An experimental approach to fish compass and map orientation," in: J. D. McCleave, G. P. Arnold, J. J. Dodson, and W. H. Neill (Eds.), *Mechanisms of Migration in Fishes*, New York: Plenum, pp. 113–123.

Quinn, T. P., Brannon, E. L. (1982), "The use of celestial and magnetic cues by orienting sockeye salmon smolts," *J. Comp. Physiol.* **147**, 547–552.

Quinn, T. P., Groot, C. (1983), "Orientation of chum salmon (*Onchorynchus keta*) after internal and external field alteration," *Can. J. Fish. Aquat. Sci.* **40**, 1598–1606.

Quinn, T. P., Merrill, R. T., Brannon, E. L. (1981), "Magnetic-field detection in sockeye salmon," *J. Exp. Zool.* **217**, 137–142.

Rodgers, F. G., Blakemore, R. P., Blakemore, N. A., Frankel, R. B., Bazylinski, D. A., Maratea, D., Rodgers, C. (1990), "Intercellular structure in a many-celled magnetotactic prokaryote," *Arch. Microbiol.* **154**, 18–22.

Rommel, S. A., Jr., McCleave, J. D. (1972), "Oceanic electric fields: perception by American eels?" *Science* **176**, 1233–1235.

Rommel, S. A., Jr., McCleave, J. D. (1973), "Sensitivity of American eels (*Anguilla rostrata*) and Atlantic salmon (*Salmo salar*) to weak electric and magnetic fields," *J. Fish. Res. Board. Can.* **30**, 657–663.

Rosenblatt, C., Torres de Araujo, F. F., Frankel, R. B. (1982a), "Light scattering determination of magnetic moments of magnetotactic bacteria," *J. Appl. Phys.* **53**, 2727–2729.

Rosenblatt, C., Torres de Araujo, F. F., Frankel, R. B. (1982b), "Birefringence determination of magnetic moments of magnetotactic bacteria," *Biophys. J.* **40**, 83–85.

Ruhenstroth-Bauer, G., Rüther, E., Reinertshofer, Th. (1987), "Dependence of a sleeping parameter from the N-S or E-W sleeping direction," *Z. Naturforsch.* **42c**, 1140–1142.

Sandberg, R., Pettersson, J., Alerstam, T. (1988), "Shifted magnetic fields lead to deflected and axial orientation of migrating European robins at sunset," *Animal Behav.* **36**, 877–887.

Schulten, K., Windemuth, A. (1986), "Model for a physiological magnetic compass," in: G. Maret, J. Kiepenheuer, and N. Boccara (Eds.), *Biophysical Effects of Steady Magnetic Fields*, New York: Springer-Verlag, pp. 99–106.

Semm, P., Nohr, D., Demaine, C., Wiltschko, W. (1984), "Neural basis of the magnetic compass: interactions of visual, magnetic, and vestibular inputs in the pigeon's brain," *J. Comp. Physiol.* **155**, 283–288.

Southern, W. E. (1972a), "Magnets disrupt the orientation of juvenile ring-billed gulls," *Bioscience* **22**, 476–479.

Southern, W. E. (1972b), "Influence of disturbances in the earth's magnetic field on ring-billed gull orientation," *Condor* **74**, 102–105.

Southern, W. E. (1988), "The earth's magnetic field as a navigational cue," in: A. A. Marino (Ed.), *Modern Bioelectricity*, New York: Marcel Dekker, pp. 35–74.

Sparks, N. H. C., Mann, S., Bazylinski, D. R., Lovley, D. R., Jannasch, H. W., Frankel, R. B. (1990), "Structure and morphology of magnetite anaerobically produced by a marine magnetotactic bacterium and a dissimilatory iron-reducing bacterium," *Earth Planet. Sci. Lett.* **98**, 14–22.

Spormann, A. M., Wolfe, R. S. (1984), "Chemotactic, magnetotactic, and tactile behavior in a magnetic spirillum," *FEMS Microbiol. Lett.* **22**, 171–177.

Tenforde, T. S. (1989), "Electroreception and magnetic reception in simple and complex organisms," *Bioelectromagnetics* **10**, 215–221.

Tesch, F.-W. (1974), "Influence of geomagnetism and salinity on the direction choice of eels," *Helgol. Wiss. Meeresunters.* **26**, 382–395.

Tomlinson, J., Reilly, D., Ballering, R. (1980), "Magnetic radular teeth and geomagnetic responses in chitons," *The Veliger* **23**, 167–170.

Torres de Araujo, F. F., Pires, M. A., Frankel, R. B., Bicudo, C. E. M. (1986), "Magnetite and magnetotaxis in algae," *Biophys. J.* **50**, 375–378.

Torres de Araujo, F. F., Germano, F. A., Goncalves, L. L., Pires, M. A., Frankel, R. B. (1990), "Magnetic polarity fractions in magnetotactic bacterial populations near the geomagnetic equator," *Biophys. J.* **58**, 549–555.

Towe, K. M., Lowenstam, H. A. (1967), "Ultrastructure in development of iron mineralization in the radular teeth of *Cryptochiton stelleri* (Mollusca)," *J. Ultrastruct. Res.* **17**, 1–13.

Towe, K. M., Moench, T. T. (1981), "Electron-optical characterization of bacterial magnetite," *Earth Planet Sci. Lett.* **52**, 213–220.

Ueda, K., Kusunoki, M., Kato, M., Kakizawa, R., Nakamura, T., Yaskawa, K., Koyama, M., Maeda, Y. (1982), "Magnetic remanences in migratory birds," *J. Yamashina Inst. Ornith.* **14**, 166–170.

Vali, H., Forester, O., Amarantidis, G., Petersen N. (1987), "Magnetotactic bacteria and their magnetofossils in sediments," *Earth Planet. Sci. Lett.* **86**, 389–400.

Visalberghi, E., Alleva, E. (1979), "Magnetic influences on pigeon homing," *Biol. Bull.* **156**, 246–256.

von Dobeneck, T., Petersen, N., Vali, H. (1987), "Bakterielle Magnetofossilien," *Geowiss. Unserer Zeit.* **5**, 27–35.

von Frisch, K. (1967), *The Dance Language and Orientation of Bees*, Cambridge, MA: Harvard University.

Walcott, C. (1978), "Anomalies in the earth's magnetic field increase the scatter of pigeon's vanishing bearings," in: K. Schmidt-Koenig and W. T. Keeton (Eds.), *Animal Migration, Navigation and Homing: Symposium Held at the University of Tuebingen, August 17+20, 1977*, Berlin: Springer-Verlag, pp. 143–151.

Walcott, C. (1980), "Homing-pigeon vanishing bearings at magnetic anomalies are not altered by bar magnets," *J. Exp. Biol.* **86**, 349–352.

Walcott, C., Green, R. P. (1974), "Orientation of homing pigeons altered by a change in the direction of an applied magnetic field," *Science* **184**, 180–182.

Walcott, B., Walcott, C. (1982), "A search for magnetic-field receptors in animals," in: F. Papi, H.G. Wallraff (Eds.), *Avian navigation*, Berlin: Springer-Verlag, pp. 338–343.

Walcott, C., Gould, J. L., Kirschvink, J. L. (1979), "Pigeons have magnets," *Science* **184**, 180–182.

Walcott, C., Gould, J. L., Lednor, A. J. (1988), "Homing of magnetized and demagnetized pigeons," *J. Exp. Biol.* **134**, 27–41.

Walker, M. M. (1984), "Learned magnetic field discrimination in yellowfin tuna, *Thunnus albacares*," *J. Comp. Physiol. A* **155**, 673–679.

Walker, M. M., Bitterman, M. E. (1989a), "Conditioning analysis of magnetoreception in honeybees," *Bioelectromagnetics* **10**, 261–275.

Walker, M. M., Bitterman, M. E. (1989b), "Attached magnets impair field discrimination by honeybees," *J. Exp. Biol.* **141**, 447–451.

Walker, M. M., Bitterman, M. E. (1989c), "Honeybees can be trained to respond to very small changes in geomagnetic-field intensity," *J. Exp. Biol.* **145**, 489–494.

Walker, M. M., Kirschvink, J. L., Chang, S-B. R., Dizon, A. E. (1984), "A candidate magnetic sense organ in the yellowfin tuna, *Thunnus albacares*," *Science* **224**, 751–753.

Walker, M. M., Kirschvink, J. L., Dizon, A. E. (1985), "Magnetoreception and biomineralization of magnetite: fish," in: J. L. Kirschvink, D. S. Jones, and B. J. MacFadden (Eds.), *Magnetite Biomineralization and Magnetoreception in Organisms: a New Biomagnetism*, New York: Plenum, pp. 417–437.

Walker, M. M., Kirschvink, J. L., Kobayashi-Kirschvink, A. (1986), "Mathematical model for magnetic navigation by animals," in: G. Maret, J. Kiepenheuer, and N. Boccara (Eds.), *Biophysical Effects of Steady Magnetic Fields*, New York: Springer-Verlag, pp. 207–211.

Walker, M. M., Quinn, T. P., Kirschvink, J. L., Groot, C. (1988), "Production of single-domain magnetite throughout life by sockeye salmon, *Oncorhynchus nerka*," *J. Exp. Biol.* **140**, 51–63.

Wallraff, H. G., Neumann, M. F. (1989), "Contribution of olfactory navigation and nonolfactory pilotage to pigeon homing," *Behav. Ecol. Sociobgiol.* **25**, 293–302.

Wiltschko, W. (1972), "The influence of magnetic total intensity and inclination on directions preferred by migrating robins (*Erithacus rubecula*)," in: S. R. Galler, K. Schmidt-Koenig, G. J. Jacobs, and R. E. Belleville (Eds.), *Animal Orientation and Navigation*, Washington, DC: U.S. GPO, NASA SP-262, pp. 569–578.

Wiltschko, W. (1980), 'The earth's magnetic field and bird orientation," *Trends Neurosci.* 1980, 140–144.

Wiltschko, W., Wiltschko, R. (1972), "Magnetic compass of European robin," *Science* **176**, 62–64.

Wiltschko, W., Wiltschko, R. (1975), "The interaction of stars and magnetic field in the orientation system of night migrating birds," *Z. Tierpsychiol. Tierernaehr.* Futtermittelkd. Beih. **37**, 337–355.

Wiltschko, R., Wiltschko, R. (1978), "Relative importance of stars and the magnetic field for accuracy of orientation in night-migrating birds," *Oikos* **30**, 195–206.

Wiltschko, R., Wiltschko, W. (1989), "Pigeon homing: olfactory orientation—a paradox," *Behav. Ecol. Sociobiol.* **24**, 163–173.

Wiltschko, R., Nohr, D., Wiltschko, W. (1981), "Pigeons with a deficient sun compass use the magnetic compass," *Science* **214**, 343–344.

Wiltschko, W., Wiltschko, R., Keeton, W. T., Maddon, R. (1983), "Growing up in an altered magnetic field affects the initial orientation of young homing pigeons," *Behav. Ecol. Sociobiol.* **12**, 135–142.

Yeagley, H. L. (1947), "A preliminary study of the physical basis of bird navigation I," *J. Appl. Phys.* **18**, 1035–1063.

Yeagley, H. L. (1951), "A preliminary study of the physical basis of bird navigation II," *J. Appl. Phys.* **22**, 746–760.

Yorke, E. D. (1979), "A possible magnetic transducer in birds," *J. Theor. Biol.* **77**, 101–105.

Yorke, E. D. (1981), "Sensitivity of pigeons to small magnetic field variations," *J. Theor. Biol.* **89**, 533–537.

Zoeger, J., Dunn, J. R., Fuller, M. (1981), "Magnetic material in the head of the common pacific dolphin," *Science* **213**, 892–894.

Further Reading

Able, K. P. (1980), "Mechanisms of orientation navigation and homing," in: S. A. Gauthreaux, Jr. (Ed.), *Animal Migration Orientation and Navigation*, New York: Academic, pp. 283–373.

Adler, H. E. (Ed.) (1971), "Orientation: sensory basis," *Ann. N.Y. Acad. Sci.*, Vol. 188.

Bolis, L., Keynes, R. D., Maddrell, S. P. H. (Eds.), *Comparative Physiology of Sensory Systems*, New York: Cambridge University.

Dubrov, A. P. (Ed.) (1978), *The Geomagnetic Field and Life: Geomagnetobiology*, translated from Russian edition of 1974, New York: Plenum, 318 pp.

Fessard, E. (Ed.) (1974), *Handbook of Sensory Physiology*, New York: Springer-Verlag.

Galler, S. R., Schmidt-Koenig, K., Jacobs, G. J., Belleville, R. E. (Eds.) (1972), *Animal Orientation and Navigation*, Washington, DC: U. S. GPO, NASA SP-262.

Gauthreaux, S. A., Jr. (Ed.) (1981), *Animal Migration Orientation and Navigation*, New York: Academic, pp. 283–373.

Hodgson, E. S., Mathewson, R. F. (Eds.) (1978), *Sensory Biology of Sharks, Skates and Rays*, Washington, D.C.; U.S. GPO.

Kirschvink, J. L., Jones, D. S., MacFadden, B. J. (Eds.) (1985), *Magnetite Biomineralization and Magnetoreception in Organisms: A New Biomagnetism*, New York: Plenum.

Marino, A. A. (Ed.) (1988), *Modern Bioelectric-*

ity, New York: Marcel Dekker, 1050 pp.

Maret, G., Kiepenheuer, J., Boccara, N. (Eds.) (1986), *Biophysical Effects of Steady Magnetic Fields,* New York: Springer-Verlag.

McCleave, J. D., Arnold, G. P., Dodson, J. J., Neill, W. H. (Eds.) (1982), *Mechanisms of Migration in Fishes,* New York: Plenum.

Papi, F., Wallraff, H. G. (Eds.) (1982), *Avian Navigation: Proceedings,* Berlin: Springer-Verlag.

Presman, A. S. (1970), *Electromagnetic Fields and Life,* New York: Plenum, 336 pp.

Schmidt-Koenig, K., Keeton, W. T. (Eds.), *Animal Migration, Navigation, and Homing: Symposium Held at the University of Tuebingen, August 17–20, 1977,* Berlin: Springer-Verlag.

Sheppard, A. R., Eisenbud, M. (1977), *Biological Effects of Electric and Magnetic Fields of Extremely Low Frequency,* New York: New York University.

Tenforde, T. S. (Ed.) (1979), *Magnetic Field Effects on Biological Systems,* New York: Plenum, 336 pp.

Weinberg, H., Stroink, G., Katila, T. (Eds.) (1985), *Biomagnetism: Applications and Theory,* New York: Pergamon.

Williamson, S., *et al.* (Eds.) (1983), *Biomagnetism: An Interdisciplinary Approach,* New York: Plenum.

It should be noted that Professor Heinz Lowenstam's studies of chiton magnetite and of other biominerals (Lowenstam, 1981) became the foundation of a new and currently exciting field of scientific inquiry: that of biomineralization. Recently several books including material on magnetic biominerals have appeared: Mann, S., Webb, J., Williams, R. J. P. (Eds.) (1989), *Biomineralization: Chemical and Biochemical Perspectives,* New York: VCH Publishers, 541 pp.; K. Simkiss, Wilbur, K. M. (Eds.) (1989), *Biomineralization,* New York: Academic, 337 pp.; Lowenstam, H. A., Weiner, S. (Eds.) (1989), *On Biomineralization,* New York: Oxford University, 324 pp.; and Frankel, R. B., Blakemore, R. P. (Eds.) (1991), *Iron Biominerals,* New York: Plenum,435 pp.

BIOMAGNETISM, MEDICAL ASPECTS

SAMUEL J. WILLIAMSON, *Department of Physics, New York University, New York, New York, U.S.A.*

	Introduction	453
1.	**Biomagnetic Fields**	453
2.	**Instruments and Their Use**	454
2.1	Flux-Gate Sensors	455
2.2	Induction Coils	455
2.3	SQUID Sensors	456
2.3.1	Detection Coils	456
2.4	Measuring a Field Pattern	457
3.	**Electric Current Sources**	458
3.1	Current Dipole Model	458
3.2	Primary and Secondary Currents	459
3.3	Application to the Human Head	459
4.	**Neuromagnetism**	459
4.1	Action Potential of an Isolated Neuron	460
4.2	Activity of Neural Tissue	461
4.3	Sensory-Evoked Processes	461
4.3.1	Auditory Cortex	461
4.3.2	Somatosensory Cortex	462
4.3.3	Visual Cortex	463
4.3.4	Cognitive Processes	463
4.4	Spontaneous Rhythms	463
4.5	Clinical Applications	464
5.	**Magnetic Field of the Eye**	464
6.	**Cardiomagnetism**	465
6.1	Conduction System	465
7.	**Remanent Magnetism**	466
7.1	Pneumomagnetism	467
7.1.1	Clearance	467
7.1.2	Relaxation	468
8.	**Biosusceptibility**	468
9.	**Summary**	469
	Glossary	469
	Works Cited	470
	Further Reading	471

INTRODUCTION

Magnetism of the human body was discovered nearly thirty years ago, when Baule and McFee (1963) detected the weak magnetic field of electrical currents in the heart. Now, with superconducting field sensors, it is possible to monitor fields from many organs of the body. In this way the study of *biomagnetism*—magnetic fields associated with biological activity—has made important contributions in several areas of basic and clinical research. Biomagnetic fields arise from other organisms as well. The smallest are single cells such as macrophages of the lung, which engulf particles that lodge on the alveolar surface. The field produced by these particles provides evidence of the dynamics of the cell's internal cytoskeleton, and the viability of the cells. In studies of other species, such as the homing pigeon and honey bee, there is evidence that magnetic particles biologically produced in the body serve as a "compass" to assist the host in navigation (Kirschvink, 1983). Effects produced by the application of magnetic fields to biological systems are not considered here. (See BIOMAGNETISM AND GEOMAGNETIC FIELD DETECTION BY ORGANISMS.)

1. BIOMAGNETIC FIELDS

Biomagnetic fields arise from three classes of sources: One is *electric current,* associated with the movement of ions within cells and throughout the extracellular space of tissue. Electric currents are responsible for the magnetic field of the heart, brain, muscle, and some other active organs of the body. *Magnetic particles* are a second source. They produce a steady magnetic field because of the type and geometrical arrangement of their constituent atoms. They occur as contaminants in the human body, as well as in other organisms, and their field can be detected when particles are magnetized in a common direction by application of a strong magnetic

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

field. Examples include airborne particles lodged in the lung as well as ingested pieces of metal in the stomach that come from opening a can of food. The third source is the *magnetic susceptibility* of body tissue. When a magnetic field is applied, the small adjustments of electron orbits in molecules create their own magnetic field. Because the predominant body constituent is water, normal tissue is diamagnetic; however, certain diseases lead to such a large accumulation of iron-bearing ferritin and hemosiderin in the liver that the organ becomes paramagnetic.

The strongest biomagnetic fields are associated with magnetic contaminants in the body, but even near the chests of welders and foundry workers exposed to high levels of magnetic dust they reach only 10^{-5} of the earth's field. The weakest of the biomagnetic fields commonly measured are associated with electrical currents flowing within small populations of active neurons in the brain. They are about 10^{-9} of the earth's field. Figure 1 summarizes typical amplitudes and frequency regions of activity for a variety of biomagnetic fields of interest. Field strengths are commonly given as the magnetic induction *B*, expressed in units of teslas. A typical strength for brain activity near the scalp is about 100 fT (100×10^{-15} T) and for cardiac activity near the chest 10 pT (10×10^{-12} T). An overview of this subject is given by tutorials and research papers of a recent conference (Williamson *et al.*, 1989), and fundamental aspects are covered by an edited textbook (Williamson *et al.*, 1983). The approach taken here is to summarize the methods by which these very weak fields can be detected and then to explain how these measurements provide unique information about physiological processes.

2. INSTRUMENTS AND THEIR USE

Three types of magnetic sensors have proven useful in biomagnetic applications: flux-gate magnetometers, induction coils, and superconducting quantum interference devices (SQUIDS) (see SQUIDS; MAGNETIC INSTRUMENTATION). Examples of how their sensitivity varies with the frequency of a magnetic signal are given in Fig. 2. An induction coil used for biomagnetic measurements (not shown) has achieved an equivalent noise level

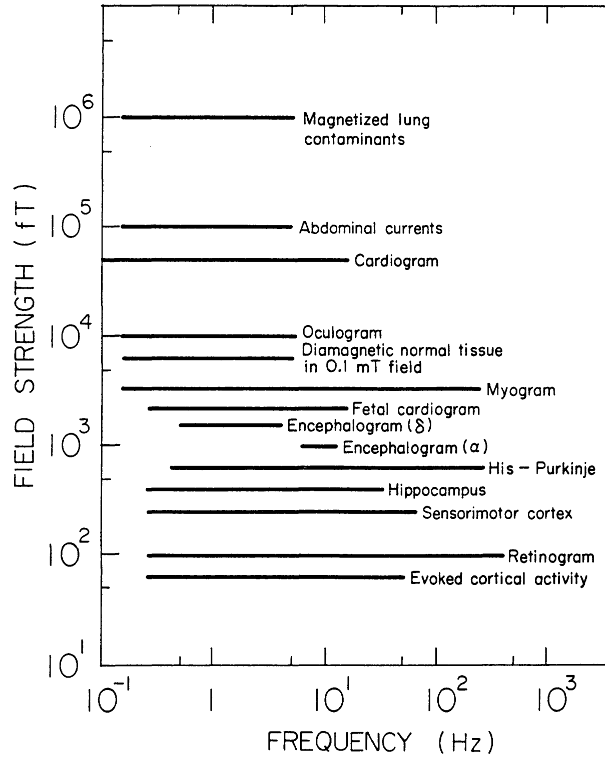

FIG. 1. Typical field strength (in femtotesla) and bandwidth of activity for various biomagnetic phenomena. Lung contaminants refer to exposure in a foundry or welding shop. Abdominal currents may arise after drinking cold water. The cardiogram characterizes the peak strength when ventricular muscle is activated. The oculogram refers to the steady field of the eye. The myogram is the peak strength from stressed skeletal muscle. The fetal magnetocardiogram is measured just outside the mother's abdomen. The encephalogram characterizes low-frequency δ and α rhythms. His–Purkinje signals of the cardiac conduction system are measured directly over the chest. The hippocampal signal is produced when a subject experiences an unexpected sensory stimulus in a train of stimuli. The sensorimotor cortex signal characterizes the peak field when a person voluntarily moves a finger. The retinogram is the change in field near the eye when the level of illumination changes abruptly, and the measure for evoked cortical activity corresponds to the brain's response to a strong visual stimulus.

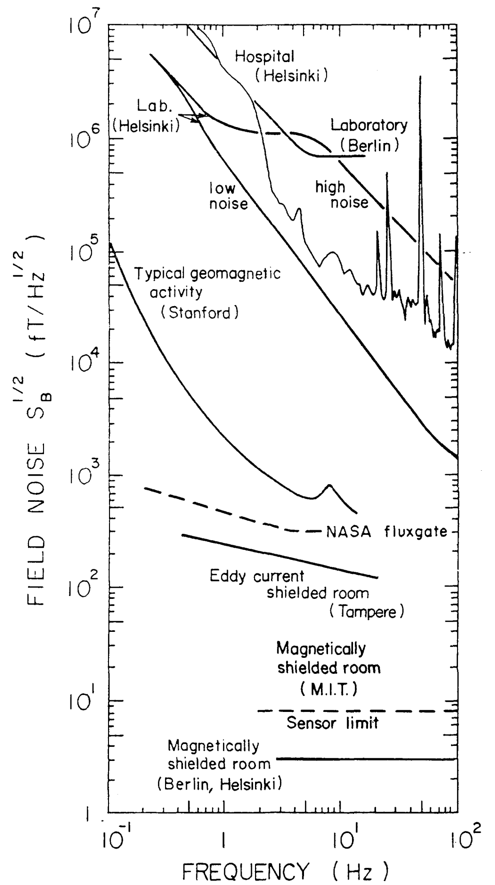

FIG. 2. Magnetic-field noise levels in representative environments compared with sensor noise, each expressed as the rms magnetic field within a 1-Hz bandwidth at each frequency. Representative environmental noise spectra are shown for for a hospital in Helsinki; a laboratory in Berlin; conditions of high and low noise in a laboratory in Helsinki; outdoors at Stanford University, California; within a room shielded from eddy currents by welded aluminum plates in Tampere, Finland; and within a five-layer magnetically shielded room in Berlin. Dashed lines illustrate typical sensor noise for a NASA flux-gate magnetometer measured within a magnetically shielded room (below 5 Hz the room cannot shield all of the environmental noise) and an MIT SQUID sensor (where the shielded room was effective to below 2 Hz).

of 600 fT Hz$^{-1/2}$ at 10 Hz, and this decreases inversely with increasing frequency. By comparison, a special feature of the flux gate and SQUID is their ability to respond to very-low-frequency signals, including dc. The relative advantages of each type of sensor will be considered in turn.

2.1 Flux-Gate Sensors

The heart of a flux-gate sensor is a core of magnetic material with high permeability, so that its magnetization readily responds to the application of a weak field of interest (see MAGNETOSTATICS). Great sensitivity can be achieved with this device, because an ac technique is commonly used which applies an alternating "modulation field" to the core to drive the magnetization periodically around its hysteresis loop. In the presence of a dc field of interest, transitions between the two polarities of magnetization are made slightly earlier when the ac field is in one direction than the opposite direction, and this asymmetry in response can be detected electronically with high sensitivity. The corresponding field sensitivity is limited by inherent magnetic noise in the core.

2.2 Induction Coils

An induction coil was used by Baule and McFee when they first succeeded in detecting a magnetic field of the human body—the heart's field. The coil itself consists of many turns of copper wire, which produces an induced voltage when an ac field passes through the loops. A sensitive, low-noise amplifier enhances the voltage that appears between the ends of the wire. Because this sensor relies on the phenomenon of magnetic induction, it is suitable only for comparatively rapidly varying fields. Its sensitivity can be further enhanced by using a core of high-permeability magnetic material. Sensitivity is limited either by the fluctuating voltage associated with thermal noise of the resistance of the winding (Johnson noise) or by magnetic noise of the core. This instrument was sufficient for Cohen (1968) to detect the α rhythm of the brain, although its sensitivity is comparatively poor at such low frequencies (near 10 Hz).

Quite a different kind of inductive sensor has been developed for magnetic studies of action potentials in isolated neurons having large axons (Wikswo, 1983). Since the field from a single neuron's axon is too weak to detect with a sensor placed several centimeters away, advantage was gained by moving the sensor closer to the source and employing a geometry that maximizes the magnetic flux linking the turns of wire. This is achieved with a small toroidal coil, only millimeters in size (Fig. 3), with the axon threaded through the hole in the toroid. Sensitivity is enhanced if the ring-shaped former is made of a high-permeability ferrite material. The biological current of interest is directed along the length of the axon, and so the corresponding magnetic field threads all of the individual turns of wire of the coil. A SQUID sensor is not required, because the action potential is a rapidly moving excitation, so the frequency content of the signal is many hundreds of hertz, sufficiently high that the voltage induced in the toroidal coil can be detected by a very low-noise amplifier operating at room temperature.

2.3 SQUID Sensors

By comparison with the flux gate and induction coil, the SQUID sensor offers many advantages in addition to greatly enhanced sensitivity (see SUPERCONDUCTING DEVICES). The SQUID wins on several counts: highest sensitivity in the bandwidth of most interest, a usable bandwidth extending down to dc frequencies, and relatively small size for its sensing element. The last-named feature permits many sensors to be mounted close together, with little magnetic coupling, so that details of a field pattern can be obtained simultaneously (see SQUIDS). On the other hand, a SQUID sensor is more expensive than either the induction coil or flux gate. Moreover, aside from a few special cases where specialized refrigeration has been developed (Buchanan, Paulson, and Williamson, 1988), it requires liquid helium to keep the SQUID superconducting.

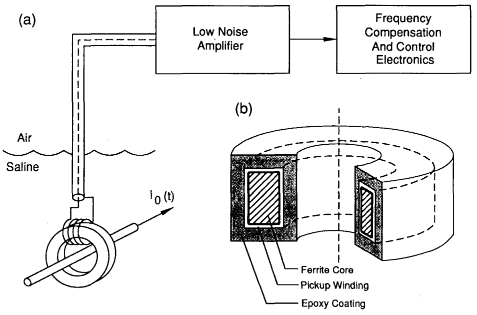

FIG. 3. System for monitoring the magnetic field of an action potential. The specimen is threaded through the hole in the core of the toroidal coil, and the signal following amplification is corrected for the effects of inductance and capacitance in the circuit. (Provided by J. P. Wikswo, Jr.)

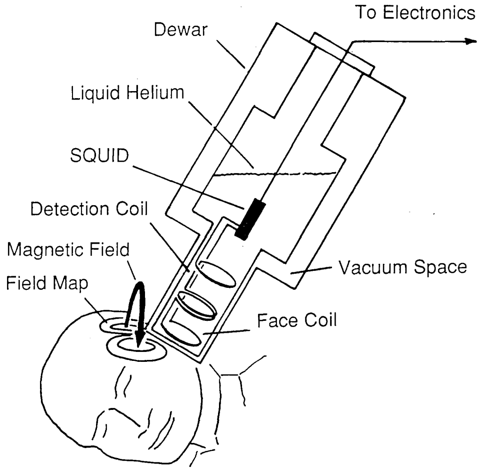

FIG. 4. Typical arrangement for biomagnetic measurements with a SQUID sensor, here for recording the magnetic field of the brain.

2.3.1 Detection Coils. A SQUID is not used directly to detect the field of interest. As illustrated in Fig. 4, there is an advantage to employing a superconducting "detection" coil to do that. One reason is that a large coil couples more signal energy into the SQUID. Another important reason is that the geometry of the detection coil can help to reject unwanted noise from the environment. Most noise sources are relatively distant, so their fields are more uniform than those of interest. Therefore a detection coil formed of two counterwound coils of the same size connected in series, but separated axially by a short distance, discriminates against such sources of uniform field but retains sensitivity to a source that is close by. It is called a *first-order gradiometer.* A uniform field from a distant noise source applies positive magnetic flux in one coil and negative flux of the same magnitude in the other, thus producing net cancellation. By contrast, a source very near the end coil produces a much larger flux in it than in the more distant coil, so its signal is registered with only a small reduction in amplitude. This function of a gradiometer is called *spatial discrimination.* The first-order gradiometer has proved useful in measurements made at sites remote from sources of magnetic noise (Zimmerman and Frederick, 1971).

The detection coil shown in Fig. 4 can be considered two first-order gradiometers arranged back-to-back. It does not respond to a uniform field because neither gradiometer does. Also, it does not respond to an axial field whose strength varies linearly along the axis of the coil. This is because the positive signal in one gradiometer is cancelled by the negative signal in the inverted gradiometer. This provides better discrimination against relatively close noise sources. Therefore the *second-order gradiometer* can be effective in an unshielded laboratory environment, where field gradients are high because of steel in the building or noise sources are simply stronger. With this, even the magnetic field of the human brain can be monitored without shielding (Brenner, Williamson, and Kaufman, 1975).

Spatial discrimination is also achieved with another coil geometry: a planar gradiometer. The simplest is a "figure-8" formed from a single length of wire. In use, it is oriented with the plane of the figure-8 parallel to the surface of the body, so that the component of the field normal to the skin produces the magnetic flux. A uniform field produces no net flux: Only when the field in one loop differs from the field in the other is there net flux. Because the loops are close together, the same subtraction is also carried out with the biomagnetic field of interest. The signal is greatest when the sensor is placed directly over an electric current, oriented with loops on opposite sides of the current. This is because the magnetic field encircling the current passes upward through one loop and downward through the other. The magnetic flux in the circuit has the same sign for each loop, thus producing a maximum output. This feature was exploited in studies of human subjects to deduce the ionic current patterns in the arms and legs of human subjects that produce a steady magnetic field (Cohen *et al.*, 1980). The patterns can be distorted by injury, such as a broken bone. The relative advantages of the figure-8 gradiometer are presently being evaluated for determining the location of electrical activity within the human brain (Kajola *et al.*, 1989). One attractive feature of this planar geometery is that techniques of thin-film deposition could be used to fabricate them with high precision. SQUID sensors made by similar techniques might also be deposited on the same substrate, for simplicity and reliability.

2.4 Measuring a Field Pattern

To determine where a magnetic source lies, its magnetic-field pattern must be measured over the relevant area of the body. If a time-dependent signal is of interest, as for the heart's activity, the sensor is moved from one location to another in succession to record the signal. Brain activity is not so repetitive, but sensory activity can be evoked reproducibly by applying an appropriate stimulus (e.g., appearance of a pattern of light, touch of the skin, or auditory tone) and recording the response with the stimulus as a time reference. Usually the responses are so weak in comparison with ongoing brain activity that measurements at a common location are averaged together to bring the signal-to-noise ratio to an acceptable level. Measuring at a large number of positions is a laborious procedure, so there is great interest in developing multisensor systems to do it at many locations simultaneously. Figure 5 shows an example of an array of detection coils for a 37-sensor system used to monitor brain activity. The diameter of each coil is 2 cm, and their centers are separated by 3.5 cm.

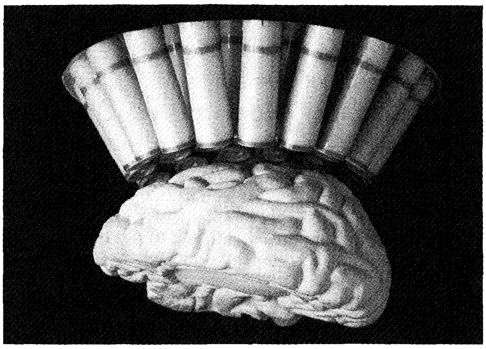

FIG. 5. Array of 37 detection coils positioned for comparison above a representation of the human cortical surface. Each coil is wound as a second-order gradiometer, and only the face coil and middle compensation coil are shown. The axes of the coils are canted so that they point approximately toward the center of the head, and that the dominant signal comes from the radial component of the brain's field. (Provided by Biomagnetic Technologies, Inc.)

Sequential placement of a sensor over the body cannot be employed to measure a steady field pattern, such as that produced by a magnetized lung or by long-lasting currents in the brain. The environmental field has such large gradients that they will likely produce a signal that dominates the one of interest. A common way to avoid this problem is to mount the sensor rigidly in place and move the subject up to the sensor, and then back. By doing this repeatedly and moving different areas of the body up to the sensor, the pattern of steady field can be mapped.

3. ELECTRIC CURRENT SOURCES

An electric current is produced by movement of electric charge. In biological tissue the charge carriers are ions, such as sodium, potassium, and calcium. These are allowed to cross cell membranes by proteins that open channels for specific ions, so that they diffuse to the region of lower concentration. In addition, membranes contain active biological "pumps" that carry specific ions to the region of higher concentration. The pumps for Na^+, K^+, and Ca^{2+} in neurons and cardiac muscle cells are examples. Because of these features, detailed current patterns at the cellular level may be quite complicated. However, it is not essential to take them all into account, because such fine details are not revealed by measurements of a magnetic field made several centimeters away. Indeed, transmembrane currents—while of keen interest to the electrophysiologist—do not produce an appreciable magnetic field. One reason is the symmetry of the radial flow: Current on one side of the cell is opposed by current on the other side, and so their magnetic fields tend to cancel each other. It is the accompanying flow of ions along the axis of the cell that is the principal contributor of magnetic fields outside the body. In tissue where elongated cells are aligned parallel to each other (e.g., muscle of the heart and pyramidal cells of the brain's cortex, called the "gray" matter), there is a net directed flow of current, and this gives rise to a measurable magnetic field.

3.1 Current Dipole Model

The magnetic field of a short length of current can be computed by the law of Biot and Savart (see ELECTRODYNAMICS, CLASSICAL), which says that the magnetic field describes circles centered on the axis of the current, and its strength varies as the inverse square of the distance from the current. For a population of aligned cells, the total field is simply a summation of contributions over the volumes of all cells. This is equivalent to computing the product of current and the path length along which it flows within the cells. The source is therefore characterized by a direction (the average direction of the current) and a strength (the product of the current and length). Therefore the source is a vector. It is known as a *current dipole*, $\mathbf{p}_i$, and its magnitude p_i has the units of ampere–meter. The current dipole is a valid model for a patch of active tissue so long as the spatial extent of the tissue is much smaller than the distance to the field sensor. Typical current dipole moments describing measured activity of the human brain range from 1 to 20 nanoampere–meters (1–20 nA m). This simplified view of a localized biological source plays a central role in biomagnetism, because it provides a means of quantifying the strength of activity. The term "current dipole" describes the fact that this active region of tissue provides current from the head of the dipole to the surrounding conducting volume and receives the returning current at its tail. To the outside world it appears as a source of current I slightly separated by a distance l from an equivalent sink (a "negative" source, $-I$), so the moment is defined as $p_i = Il$. The name was coined by analogy with a charge dipole, where a positive charge Q is slightly separated from a negative charge of equal magnitude $-Q$, and the electric dipole moment is defined as $p_e = Ql$.

It is important to recognize that volume current flowing through a medium of uniform conductivity produces no magnetic field. One way to show this is to integrate the Biot–Savart law over the entire volume when the source is a current dipole. The computation shows that the net field is zero. Therefore, by superposition, any source that can be described by a configuration of current dipoles produces no magnetic field from its volume current in the extracellular medium. The sole sources of field are the current dipoles themselves.

3.2 Primary and Secondary Currents

An active region of tissue that is described by a current dipole, or configuration of current dipoles in different positions and orientations, is called a *primary source.* It is a simplified representation of the more complex configuration of underlying cellular current flow. The primary source sets up a pattern of volume current within the surrounding conducting tissue. In dealing with biological applications, it is essential to take into account the effect of boundaries between tissues of differing conductivity (lung, heart, bones, etc.). They distort the pattern of volume current, and this influence may contribute to the observed magnetic field.

This effect can be taken into account conceptually and mathematically by imagining an equivalent situation: Space is considered to have uniform conductivity, and the effect of each boundary is represented by an appropriate distribution of current sources across the area of the boundary, oriented perpendicular to it so as to reproduce the same volume-current pattern as in the physical situation. These current sources at the boundaries are called *secondary sources* (Geselowitz, 1970). Where the volume current flows into a region of lower conductivity in the physical situation, the secondary source points toward the region of higher conductivity in the equivalent situation, and vice versa. The strength of the secondary source at each point on a boundary is the product of the electric potential and difference in conductivities across the boundary. Therefore if the distribution of electric potential throughout the medium is computed for a given configuration of primary sources, the secondary sources at each boundary can also be computed. The field anywhere outside the primary-source region is predicted by applying the Biot–Savart law to sum contributions from both primary and secondary sources.

3.3 Application to the Human Head

The characteristics of secondary sources have profound implications for the field patterns that are produced by neuronal currents in the human brain. For instance, if the head is considered a spherical core surrounded by a concentric shell of lower conductivity representing the skull, secondary currents at the inner and outer surface of the skull are radially oriented. Therefore, the fields they create outside the head are purely tangential to the scalp underneath. In other words, by measuring the radial component of the field only the primary sources are detected. This greatly simplifies the analysis. On the other hand, for the spherical geometry, the field of a radially oriented primary source is exactly cancelled by the field of the secondary currents.* In other words, a radially oriented primary current is "magnetically silent." Only the tangential component of a primary current can be detected magnetically. Paradoxically, this may be an advantage in brain research, because it means that the field pattern across the scalp has a recognizable symmetry for a localized neuronal source. As illustrated in Fig. 6, the field emerges from one area of the scalp and enters at another area, with the source lying under the midpoint. The span between the locations of these field extrema, in comparison with the radius of the head, indicates the depth of the source (Williamson and Kaufman, 1987). The span increases with depth, reaching half the circumference when the source approaches the center of the sphere. Of course, this limiting case is an example of a radial source, so the field strength at the extrema also goes to zero in this limit.

4. NEUROMAGNETISM

The creation of magnetic fields as a consequence of neural activity is called *neuromagnetism.*

There are three broad areas of investigation that have attracted greatest attention. The first to be considered focuses on the properties of an isolated nerve cell. The second

---

*The easiest way to see this is to recognize that the geometry is cylindrically symmetric about the dipole's axis, so the field lines would be circles centered on this axis. By Ampère's law, the field integrated about a circle outside the sphere is proportional to the net current passing through the area of the circle. But the net current is zero, because the total current, including the dipole's, passing through the area in one direction is the same as that passing through in the other direction.

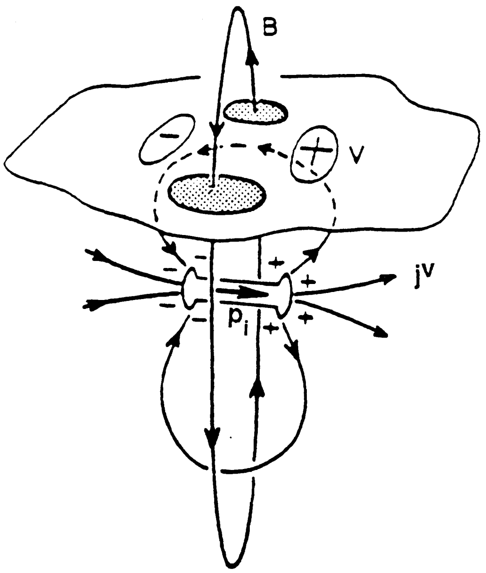

FIG. 6. Current dipole $\mathbf{p}_i$ in the conducting volume of the brain, with the associated volume current density $\mathbf{j}^V$ emerging from the head of the dipole and reentering at the tail. Ovals schematically indicate the positions of the two extrema in the scalp potential (V) and the two extrema of the radial magnetic field (B).

concerns activity of neuronal tissue, such as slices of cerebellum or muscle. The final area comprises processes of the human brain. They include sensory-related activity that is produced in response to a stimulus such as a sound or light pattern; spontaneous activity, with the α rhythm as the best known; and clinical applications, especially for epilepsy.

4.1 Action Potential of an Isolated Neuron

The business of neural tissue is to process information and transmit it from one place in the body to another. Transmission is the simplest process, so it merits initial consideration. To convey information over long distances in the body requires a rapid means, and the *action potential* serves this function. It propagates without distortion along the axon of a nerve cell. Because of its high speed, the entire chain of events as it passes by a particular location on the axon is completed within a few milliseconds. The sequence forming the action potential begins with depolarization of the transmembrane potential, commonly as a result of multiple inputs onto a neuron from neighboring neurons. The standing potential of −80 mV (inside potential relative to outside) rapidly decreases to nearly zero, or even becomes positive. This opens channels in the membrane through which sodium ions enter the cell, where their initial concentration is lower than outside. The electric field produced by the local increase of intracellular positive charge accelerates positive ions within the axon (predominantly potassium) to flow in the direction of propagation, away from the initial site of excitation [Fig. 7(a)]. This is known as the depolarization front. Slower responding potassium channels then open to permit an outflow of these ions, and the decrease in positive charge following the depolarization front causes a backward-directed intracellular current. This is the repolarization front. The action potential is therefore accompanied by two counterdirected intracellular currents. Once the action potential has passed, metabolically-driven "pumps" in the membrane slowly reestablish the original concentration of ions by passing sodium outward and potassium inward.

It is possible with a toroidal coil and low-noise amplifier to measure the magnetic fields accompanying the depolarization and repolarization fronts as they pass by. Comparing the transmembrane potential that can be deduced from this field with the potential directly measured with electrodes is a stringent test of the underlying theory for neuromagnetic fields. Studies of the crayfish axon show that the computed and measured wave forms closely agree with only one adjustable parameter (Roth and Wikswo, 1985). From the fact that the input data were measurements of a current and a corresponding voltage, it can be understood that the parameter corresponds to the intracellular electrical conductivity. Thus this comparison of electric voltage and magnetic field not only provided verification of the underlying theory for neuromagnetic sources but also gave an improved value for the conductivity. These techniques are being extended to characterize the function of a damaged nerve in the extremities of the human body, such as the arm, where a coil can be clamped around the nerve to monitor its response to stimulation.

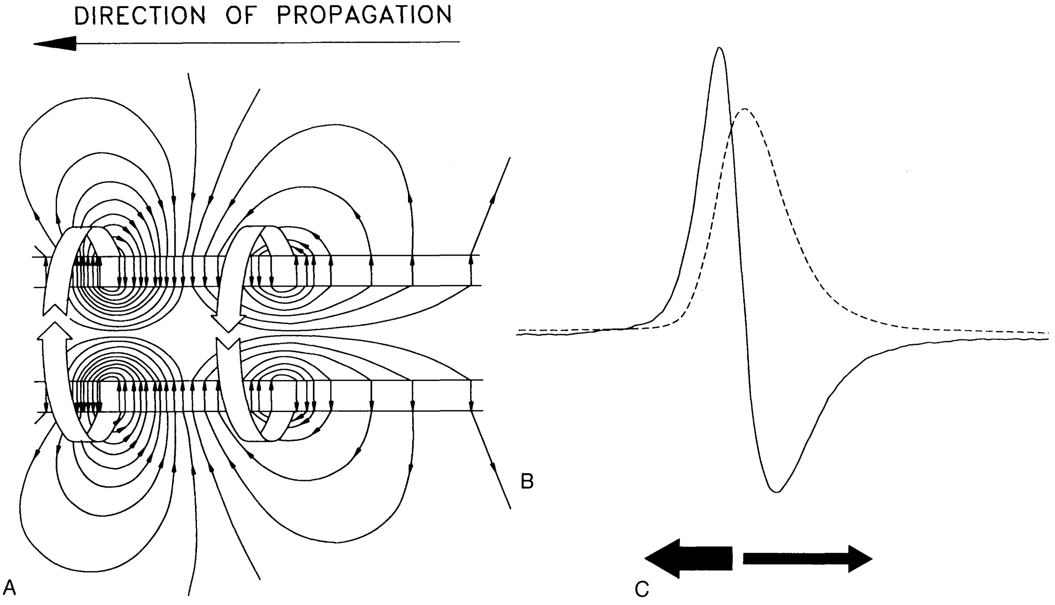

FIG. 7. Currents associated with the action potential of a nerve axon moving toward the left. (a) Curved thin lines indicate current patterns of the intracellular and extracellular space; straight segments show the direction of the transmembrane current (thickness of the membrane is greatly exaggerated compared with the diameter of the axon); and broad arrows encircling the axon show the direction of magnetic field for the depolarization front (left) and repolarization front (right). (b) Corresponding variation of transmembrane potential (dashed) and the axial intracellular currents (solid). (c) Depolarization (broad arrow) and repolarization (thin arrow) current dipoles representing axial currents. (Provided by J. P. Wikswo, Jr.)

4.2 Activity of Neural Tissue

Neurons in the brain are not isolated but establish intimate contact with one another. For this reason there is much merit in studies of tissue removed from the brain of an animal model or invasive studies of animals themselves. The work of Okada, Lauritzen, and Nicholson (1988) shows that it is possible in this way to study phenomena such as spreading cortical depression, where transmembrane potentials in cells become depressed sequentially from one location to another across the tissue. These studies demonstrate that electrode measurements within the tissue are consistent with magnetic measurements outside. This work also illustrates the interrelationship between such basic studies and clinical applications, for there is some evidence that spreading depression may be related to migraine in humans.

4.3 Sensory-Evoked Processes

The greatest effort in neuromagnetism has been devoted to studies of the human sensory systems, which are located in specialized areas of the cerebral cortex, the thin layer of gray tissue covering the brain. This is motivated in part by the notion that electrophysiological studies on animals reveal functional mappings across the sensory cortex, and these mappings suggest underlying strategies that have evolved to most effectively process sensory information. Indeed, the ability of magnetic studies to locate precisely activity evoked in the human auditory cortex has revealed just such maps.

4.3.1 Auditory Cortex. The auditory cortex in humans forms a portion of the horizontal, upper surface of the temporal lobe and is located about 5 cm above the ear canal. One functional sequence across this surface revealed magnetically is a "tonotopic" organization (Romani, Williamson, and Kaufman, 1982) for responses to an amplitude-modulated tone, where tones of higher frequencies evoke neuronal activity deeper beneath the scalp [Fig. 8(a) and 8(b)] (see ACOUSTICS, PHYSIOLOGICAL). The progression can be systematically mapped along a line, where the cumulative distance from one response to the next across the cortex increases as the logarithm of the frequency [Fig. 8(c)]. A similar progression is found for the source of strong neuronal activity following, by 100 ms, the onset of a tone burst. Moreover, tones of greater intensity do not enhance neuronal response strengths, as might be expected. Instead, neuronal activity is shifted to different locations across thecortex with increasing strength. This "amplitopic" organization extends along a line that is approximately at right angles to the tonotopic sequence (Pantev *et al.*, 1988). While neuronal response strength is not affected by tone frequency or intensity (at least for easily heard tones), it is affected by a person's attention. If attention is diverted by a distraction, the neuronal response to a tone decreases in strength. Such applications of neuromagnetic techniques to auditory studies have been particularly fruitful in revealing these as well as many other aspects of functional organizations in humans that had not been known (Hari and Lounasmaa, 1989).

4.3.2 Somatosensory Cortex. Somatic cortical responses can be evoked with electrical or mechanical stimulation of the skin. Neuronal activity appears within the central sulcus, a fissure that begins at a point just anterior to the vertex and runs downward and forward on both sides of the head to meet the upper surface of the temporal lobe. Systematic magnetic studies, in agreement with electrical studies, show the classic somatotopic representation where the location of the cortical response shifts systematically across the cortex with application of stimuli to different areas of the body. The spatial resolution obtained magnetically is sufficient to reveal a slight shift of activity across the central sulcus when a person voluntarily moves a finger. Neuronal activity is first seen in the motor cortex, forming the anterior bank, in preparation for finger movement. As the finger begins to move, the balance of neuronal activity shifts 6 mm posteriorly into the anterior bank as information from the proprioceptors in the joints and muscles is relayed to the somatosensory cortex (Okada, Williamson, and Kaufman, 1982). Therefore it is possible to study activity of this sensory–motor feedback loop noninvasively in humans. Activity in the secondary somatosensory cortex can also be observed (Hari and Kaukoranta, 1985).

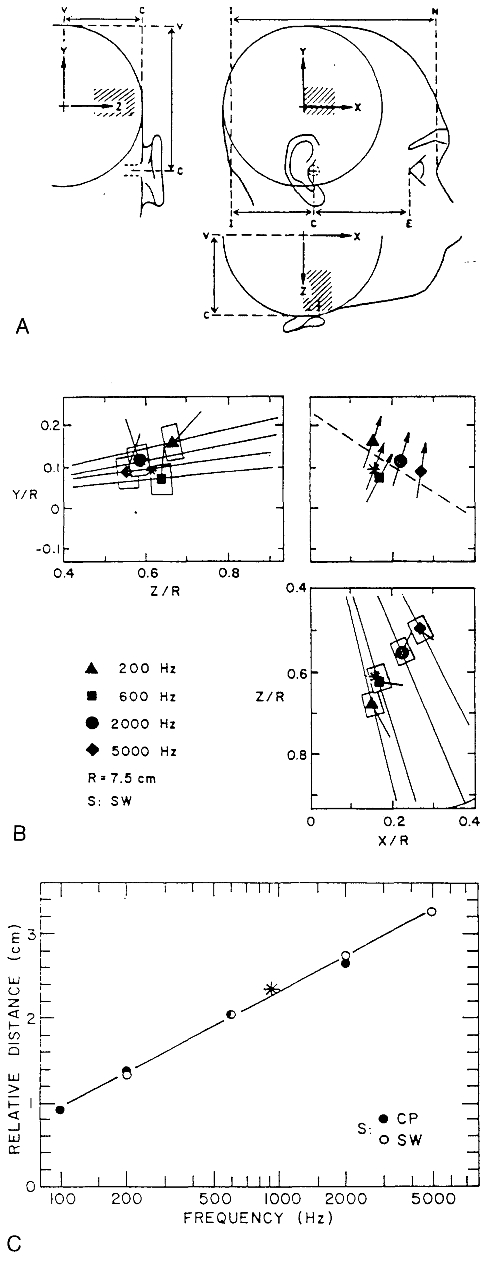

FIG. 8. (a) Three cross sections of the head, depicting the sphere used to model it for auditory studies. (b) Enlarged versions of the shaded sections showing deduced positions of neuronal activity evoked by modulated tones of the indicated frequencies. Boxes indicate the uncertainties for individual source positions. (c) Cumulative distance from one neuronal position to the next, for each of two subjects whose data are indicated by different symbols. The origin is arbitrarily set for 20 Hz. The star indicates a response from SW for rapidly presented click stimuli, where the peak in the power spectrum lies at 900 Hz (Romani, Williamson, and Kaufman, 1982).

4.3.3 Visual Cortex. Although the first sensory evoked response was detected in the visual cortex (Brenner, Williamson, and Kaufman, 1975), this modality has been less studied than the others. One reason, since overcome, was the difficulty of presenting a visual stimulus without also producing a magnetic artifact from the display. An early application provided evidence that a simple reaction time matches a series combination of a sensory response time, which varies with the type of stimulus, and a fixed motor response time of about 115 ms (Williamson *et al.*, 1978). The visual response time can vary from as little as 85 ms to well over 150 ms in response to contrast reversal of a grating pattern when the stimulus is increased in spatial frequency, decreased in contrast, or decreased in luminance. These kinds of studies probe the sequence of timing across various cortical areas. Other studies are beginning to characterize the sequence of subsequent activity in higher visual areas (Lu *et al.*, 1991), some comparing source locations with anatomical features provided by magnetic resonance images.

4.3.4 Cognitive Processes. A subject attracting increasing attention is how higher levels of brain functions are carried out. Memory processes are an example. One structure involved with declarative memory processes is the hippocampus, which is shaped like a pair of ram's horns forming a portion of the innermost surface of the lower right and left temporal lobes. There is considerable interest in finding techniques to elucidate its functions. One such approach, based on many sophisticated investigations of scalp potentials, is a paradigm where a sequence of common stimuli are presented with an occasional stimulus that differs in some physical aspect. Following these "odd" stimuli, one or more long-latency components appear that are not exhibited for the expected stimuli. It was possible magnetically to deduce the location of two of these components, for a visual task, and both were very deep regions within the left and right hemispheres, corresponding to the hippocampal formation (Okada *et al.*, 1983). This suggests that other deep structures may also be accessible to neuromagnetic study.

4.4 Spontaneous Rhythms

It is surprising that the strongest electric and magnetic activity of the brain is in a sense least understood. The well-known α rhythm is generated by a neural circuit involving both the thalamus deep within the brain and the cerebral cortex. This initiates within cortical pyramidal cells a series of bursts of oscillatory activity within the bandwidth of about 8–12 Hz. However, the function of this rhythm is yet to be explained. Magnetic studies are providing the first evidence that α rhythm in humans may be localized, with the strongest features originating within the parieto-occipital fissure. This fissure is the boundary between the classic visual and somatosensory areas of the brain.

Moreover, sources of other spontaneous rhythms may also be localized magnetically by employing a "resonance" technique. Periodic stimulation of an appropriate area of the skin can produce a greatly enhanced somatosensory response if the frequency is properly chosen. When stimulation is abruptly terminated, the response slowly dies away, similar to that of a resonant system. In this way the field pattern can be mapped over the scalp, since the resonant response is quite reproducible. One source of the resonant response is found in the central sulcus at the somatosensory projection area for the portion of the body being stimulated. Other sources are found in the parieto-occipital fissure (Narici *et al.*, 1990). Such investigations are only beginning to shed light on the functions served by these spontaneous rhythms.

Even if the specific function of a spontaneous rhythm is unknown, the signal itself can be used to advantage. One example of research into cognitive functions provides evidence that the visual cortex is involved in mental imagery. When a person is asked to compare an abstract figure with others previously shown in the same portion of the visual field, there is overall suppression of the α rhythm over the back of the head as the person mentally compares the images. However, there is much stronger suppression of the field near the midline at the back of the head, where a field would arise from sources within the visual cortex. Thus, preferential suppression of neuronal activity within the visual cortex implies that it is involved in the mental comparison task (Kaufman *et al.*, 1990). A control experiment in which the person is shown a word and asked to find a rhyme to it does not result in prolonged suppression of the α rhythm over the visual cortex. Instead, there is prolonged suppression over an anterior temporal area, implying that other functions of the brain are involved in this mental task. Thus the source localization provided by neuromagnetic studies is revealing distinct regions of spontaneous activity that are modulated differently by specific cognitive functions. By monitoring the spatial distribution of suppression, it will be possible to identify strategies that individuals employ when dealing with various mental tasks.

4.5 Clinical Applications

The most successful clinical application of neuromagnetism is in locating the site of focal epilepsy preparatory to surgical removal. A large percentage of epileptics who are under drug therapies do not respond appropriately to the therapy, and they become candidates for surgical removal of the epileptic region of tissue. Lesions of the brain, perhaps due to physical injury, may be sources of epilepsy but often focal epilepsy occurs without any manifestation shown in x-ray or magnetic resonance images (Sutherling *et al.*, 1987). Figure 9(a) illustrates one case where two foci alternate in producing epileptic discharges (Barth *et al.*, 1984). Field maps carefully recorded during the discharges reveal two different field patterns from which their respective sources were deduced, using two current dipoles to represent them. The deduced locations on an x-ray image lie at the boundary of a large lesion [Fig. 9(b)].

Other clinical applications of neuromagnetism are on the horizon. There is evidence, for instance, that auditory evoked magnetic fields provide an objective measure of tinnitus. This disease produces "ringing" in the ear, perceived by some patients with tinnitus as wide-band noise and by others as one or more tones of specific frequencies, or as a combination of the two. The ratio of the peak field at about 180 ms following the onset of a tone to the peak field at about 100 ms latency is markedly weaker for patients with tinnitus (Hoke *et al.*, 1989). In the case of a soldier who suffered tinnitus as a result of a nearby explosion, the ratio was seen to increase as the symptoms of tinnitus diminished. In another application, the ability to monitor dc fields from the brain has prompted research into the relationship between slowly varying fields and migraine. Such fields may very well arise from spreading cortical depression, but more research is needed to demonstrate the relationship.

5. MAGNETIC FIELD OF THE EYE

The aligned photoreceptors of the eye's retina produce a steady intracellular current directed toward the front of the eye. When a person changes direction of view, the eyes' rotation markedly alters the pattern of electric potential across the scalp. The corresponding change in magnetic field detected near the eyes is called the *magneto-oculogram.* A particularly interesting feature, magnetically speaking, is that the current dipole source that explains the corresponding magnetic-field pattern for one eye is about 10 000 times greater than expected from the primary source. The reason is that the pigment epithelium containing the photoreceptors has much lower electrical conductivity (by a factor of about 10 000) than the surrounding tissue. To account for current continuity across the epithelial boundaries, secondary sources must be similarly enhanced compared with primary sources. The secondary sources provide the dominant contribution to the measured field (Katila *et al.*, 1981). Changes in the strength of the magneto-oculogram are clearly detected when the eyes adapt to different illumination levels. Changes in the field near the eye are also observed when a person blinks. This is attributed to a change in the pattern of volume current, which introduces a new secondary source at the interface between the eyeball and eyelid.

These are not the only magnetic phenomena that the eye exhibits. Changes in the strength of the magneto-oculogram are produced when the eye responds to a brief flash of light. This is called the *magnetoretinogram,* by analogy with its well-known electrical counterpart, the electroretinogram. Rapid variations in field strength take place over a period of several hundred milliseconds. By comparison with the electroretinogram, individual features can be attributed to activity in photoreceptors and Müller cells of the retina. One advantage of magnetic recordings is that

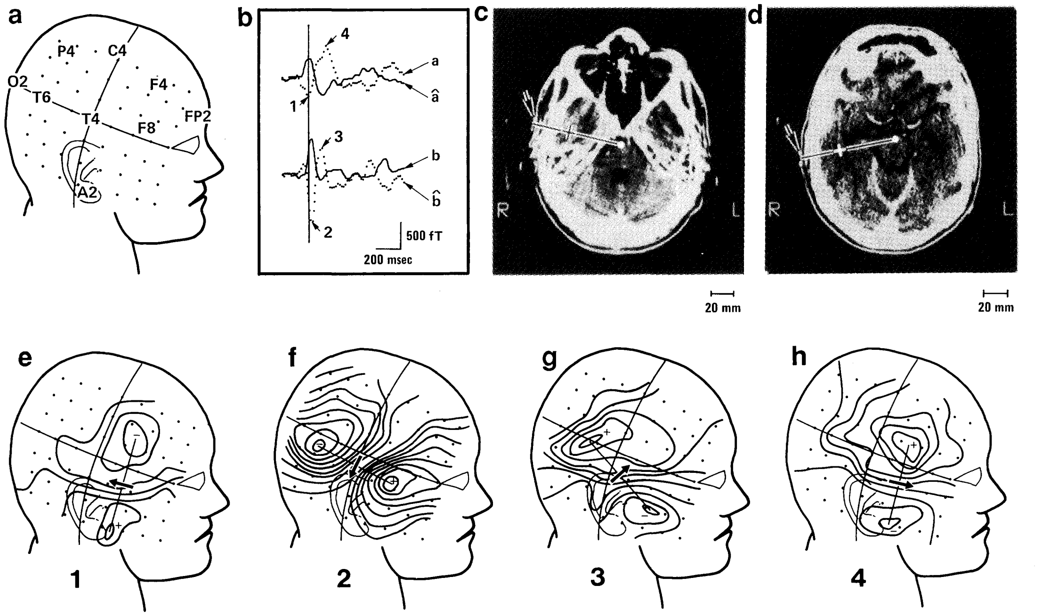

FIG. 9. Sequential activity of an epileptic discharge, obtained from field measurements across the scalp at positions indicated by the dots in (a). Averaged magnetic data (b) and corresponding isofield contours (e)–(h) reveal four alternating spike complexes. Two of them (1 and 3) arise from neuronal activity indicated by the arrows at one location [(e) and (g)] and the other two (2 and 4) from a second location [(f) and (h)]. Computed tomography sections through the head [(c) and (d)] show both sources to be located at the edge of a large area of scar tissue in the right temporal lobe. (Barth *et al.*, 1984; copyright 1984 by the AAAS.)

an electrode need not be placed on the cornea of the eye to register responses, a feature that is particularly beneficial to patients with sensitive eyes.

6. CARDIOMAGNETISM

A biomagnetic application that had attracted the greatest attention during the 1970s was characterizing cardiac activity. Measurements of the magnetocardiogram (MCG) at various positions over the torso revealed different wave forms than did the electrocardiogram (ECG) because the two recordings emphasize different features of the underlying pattern of current. The wave form during one cardiac cycle begins with the *P* wave, which arises from reduction (depolarization) of the transmembrane voltage of muscle forming the walls of the atria. It is followed by a much stronger, spikelike *QRS* wave from depolarization of the ventricles. The sequence concludes with a weaker and slower *T* wave from repolarization of the ventricles.

However, visual inspection of the wave forms has shown few obvious advantages for the MCG compared with the ECG in revealing cardiac abnormalities. Since the ECG is a well-established procedure involving much less expensive apparatus, medical interest in cardiomagnetism has not developed much for this specific application. More elaborate statistical studies have revealed, however, that signatures of the MCG provide information on certain disorders that are not as well revealed by the ECG, and vice versa (Nakaya, Mori, and Nomura, 1989). There may well be clinical benefits to be derived from the MCG in such applications.

6.1 Conduction System

More attention has been focused recently on exploiting the capability of magnetic measurements to locate disorders by applying the current dipole model to describe sources ofabnormal signals. While some attention has been directed to infarcts of cardiac muscle, primary emphasis has been placed on evaluations of the conduction system. This system

(Fig. 10) provides signals to coordinate activity of the atria and ventricles. The pacemaker at the sinus node establishes the basic cardiac rhythm. Its output initiates the depolarization of membranes of working cells in the walls of the right and left atria, as expressed by the *P* wave. In response, the muscle begins to contract. A delay of about 40 ms in the timing signal is provided by slow conduction through the atrial–ventricle node (A-V node) to allow the atrial muscles sufficient time to force blood into the ventricles. Then the signal quickly passes through a conduction path known as the His bundle, which bifurcates to form the left and right bundle branches. These in turn convey the excitation to Purkinje fibers that form a spreading network across the working muscles of the ventricles. There the excitation propagates as a wave front, moving upward through left and right ventricles, producing the *QRS* wave of the cardiogram as the muscles begin to contract to force blood out of the heart. This is followed by a period of repolarization of the working muscles, as indicated by the *T* wave.

Several types of disorders may occur in the conduction system. A delay or "block" may form at any level in the pathway; an "accessory" conducting path may be formed that bypasses the A-V node and directly connects one of the atria and the corresponding ventricle; and "reentry" may occur in muscle, where a pathway feeds the excitation back to another location. Any of these may disrupt the normal excitation cycle and cause an arrhythmia or increased frequency (tachycardia).

The conventional treatment for such disorders is surgery to bypass a block or disrupt an accessory pathway. To minimize the time needed to locate the site of disfunction during surgery, it is important to deduce where it may be as accurately as possible beforehand. The standard procedure consists of inserting a catheter through a vein into the heart and monitoring electrical activity at various positions, as the catheter is followed by x-ray images. In special cases, the catheter may even be used to affect treatment, by passing a strong current to ablate the portion of tissue.

Recent studies indicate that many of these disorders of the conduction system may be localized magnetically. A "high-resolution" procedure is carried out by measuring the cardiac signal at closely spaced intervals of 3–4 cm across the chest while recording within a wide bandwidth extending up to 300 Hz or more. In this way, excitation of the His bundle can be monitored, and localization of various types of arrhythmogenic tissue is possible by fitting a current dipole's field pattern to the measured pattern at select moments of time (Fenici *et al.*, 1989). Validation of this interpretation comes from comparisons with invasive electrophysiological studies, endocardial catheter mapping, and in some cases the results of surgical operations. For instance, Fig. 10 illustrates a possible site for preexcitation of the left ventricle via a conducting path from the left atrium (denoted as ECP). Certain atrioventricular accessory pathways may be localized magnetically with an accuracy of about 2 cm if the source is modeled as a current dipole (Nenonen *et al.*, 1990). But to achieve such accuracy it appears essential to employ a realistic model of the torso, as illustrated in Fig. 11, to account for the secondary sources associated with its surface. Infarcts within cardiac muscle may also be identified by using a current dipole model to represent the disruption of electrical activity.

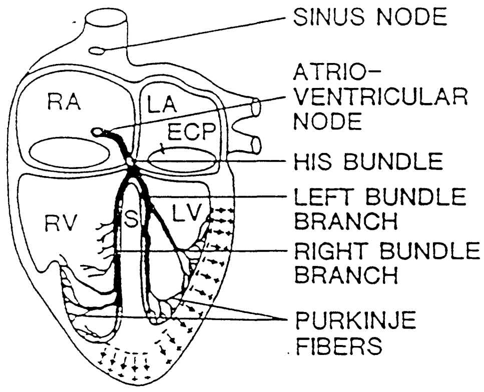

FIG. 10. Conduction system of the human heart. An accessory pathway between left atrium and ventricle is denoted by ECP. (Provided by T. Katila.)

7. REMANENT MAGNETISM

Normal human body tissue is nonmagnetic, in the sense that it has no remanent moment in the absence of an applied magnetic field. However, a remanent moment may be

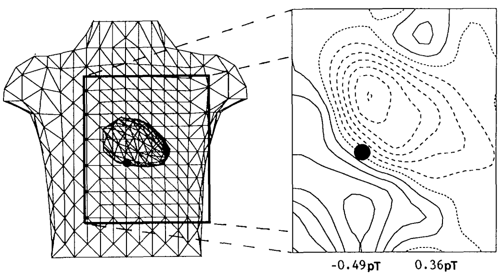

FIG. 11. Localization of the accessory pathway (dark circle) between atrium and ventricle in a patient (Nenonen *et al.*, 1990). Isofield contours in the right panel illustrate where the field emerges from the chest (solid contours) and reenters (dotted contours). The wire frame figure depicts the realistic torso and heart models employed in the computer calculation. (Provided by G. Stroink.)

detected when magnetic materials are ingested with food or are introduced by other means. Indeed, with a SQUID sensor it has been possible to locate a small portion of a magnetized acupuncture needle that had been broken off in a patient's back (Ilmoniemi *et al.*, 1988). It is also possible to detect biologically produced magnetic materials such as magnetite in the pigeon's head by using a SQUID magnetometer (Walcott, Gould, and Kirschvink, 1979). Nevertheless, the phenomena that have attracted greatest interest in studies of humans are associated with magnetic particles that become lodged within the lung.

7.1 Pneumomagnetism

Magnetic particles in sufficient quantity in the lung can be magnetized to produce a field that is detectable outside the body (Cohen, 1973). This is a pronounced effect in workers who have been engaged in specific occupations where they were exposed to magnetic dust, as in foundries, welding shops, and shipyards. If a strong magnetic field, of a fraction of a tesla or so, is applied to the volume of the lung for a few minutes and then terminated, a remanent field can be detected outside the chest. Flux-gate sensors are often sufficient for this purpose. Researchers originally hoped that the field pattern measured across the chest would provide information about the distribution of material within the lung. However, the inherent ambiguities in the solution for this inverse problem have discouraged subsequent development in this approach. Nevertheless, two other characteristics of the magnetic field continue to attract attention: clearance and relaxation.

7.1.1 Clearance. If the process of magnetizing the lung is repeated over many days, and the remanent field is measured shortly after the field is turned off, the remanent field is found to decrease from one week to the next. This has been interpreted as indicating that particles are being cleared from the lung, and the rate of decrease may be a measure of the effectiveness of clearance processes. The phenomenon can also be applied for research purposes if a small amount of magnetite dust (considered harmless in small quantities) is breathed by a volunteer subject. Figure 12 indicates that clearance is much slower in longterm smokers than in nonsmokers. This implies that other foreign materials are also likely to remain longer in smokers' lungs, thus putting smokers at greater risk to physi-

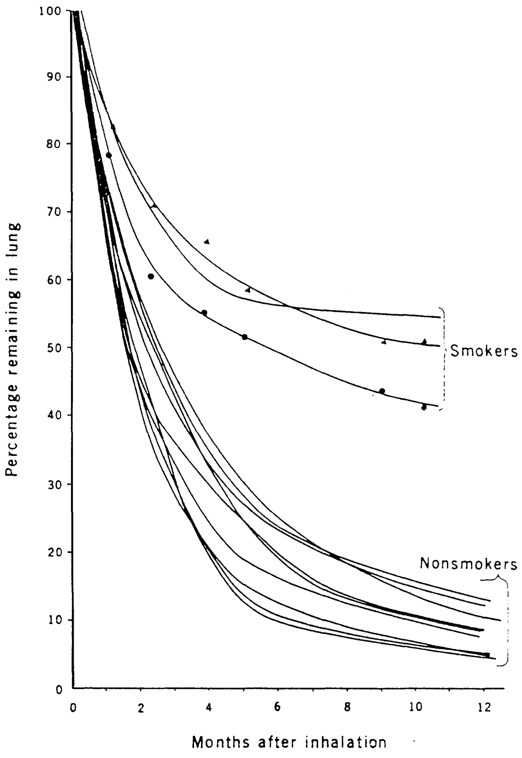

FIG. 12. Decrease of the remanent moment near the chest of smokers and nonsmokers who had inhaled fine magnetite dust. (From Cohen, Arai, and Brain, 1979; used with permission of the American Association for the Advancement of Science.)

ological insults of air pollutants in addition to the cigarette smoke itself.

7.1.2 Relaxation. The second feature of interest is *relaxation*, a steady reduction in the lung's field once the original magnetizing field is reduced to zero. It is now recognized that relaxation is signaled by particles that have been captured by macrophages (Brain *et al.*, 1984), specialized cells released on the alveolar surface that serve as the lung's garbage trucks. Macrophages engulf foreign material, thereby keeping it separate from normal body tissue, and the cells with their burdens are ultimately carried up the ciliary escalator of the bronchi to be swallowed, or taken into the body by other mechanisms. Relaxation occurs because the internal cytoskeleton of a macrophage is in continual motion, and this rotates magnetic particles so that their magnetic moments no longer point in a common direction (Valberg and Feldman, 1987). *In vitro* studies show that application of a material that is toxic to macrophages eliminates relaxation. This has encouraged a wide variety of magnetic studies to investigate the influence of diverse substances on relaxation rates.

8. BIOSUSCEPTIBILITY

One additional biomagnetic effect is proving beneficial in the clinic: measurement of the magnetic susceptibility of the liver in patients with iron overloads. Water, the major constituent of the body, responds to an applied magnetic field by producing a much weaker magnetic field in the opposite direction (diamagnetism). Deoxygenated hemoglobin in blood is paramagnetic; its responding field lies in the same direction as the applied one. As a consequence, the magnetic susceptibility of blood in the veins is 7% less than the diamagnetic susceptibility of pure water. Patients with the diseases hemochromatosis or thalassemia major are unable to recycle iron from dead red blood cells into the formation of new cells. Instead, iron is stored as components of the magnetic proteins ferritin and hemosiderin, which accumulate in the spleen, liver, and eventually other organs. Accumulation within the heart leads to disease that proves fatal. One method to combat this accumulation is administration of iron chelation drugs into the blood, so that iron is captured and eliminated in the stool. However, to follow the progress of treatment, it is important to monitor the changing iron concentration in the liver. This can be done with chemical analysis of biopsy tissue removed by insertion of a needle; however, many patients refuse this because of the pain involved.

Farrell *et al.* (1980) developed a magnetic technique to measure the susceptibility of liver tissue *in vivo*. A Dewar flask containing a SQUID sensor is placed a few centimeters above the patient, who is reclined on a special bed (Fig. 13). The Dewar also contains a superconducting magnet that produces a strong field in the region within a few centimeters below the SQUID's detection coil. Between the Dewar and the patient is a flexible bag containing water, with a connection to a water reservoir. The patient is raised by the bed, displacing some of the water, and the SQUID output is a measure of the difference between the magnetic susceptibilities of water and tissue. To obtain a quantitative value for liver susceptibility, the geometry of liver and intervening tissue must be taken into account. This geometry can be obtained from an ultrasonic scan *in situ*.

Normal liver tissue is diamagnetic, but patients with large iron overloads will exhibit a strongly paramagnetic susceptibility. Figure 14 shows the good correlation between iron

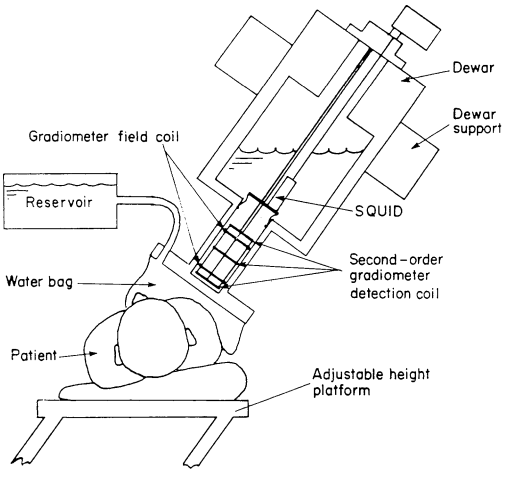

FIG. 13. Arrangement for recording the magnetic field produced by the susceptibility of a nearby region of the body when the person is moved toward the detection coil.

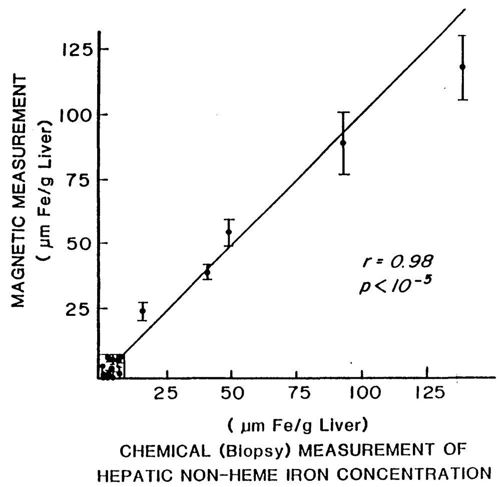

FIG. 14. Correlation of liver iron concentration obtained from magnetic measurements (vertical axis) and from chemical analysis of tissue provided by biopsy (horizontal axis) (Brittenham *et al.*, 1982).

concentrations determined magnetically and by chemical biopsy. There is reason to believe that the former is more representative of liver iron concentrations, because the volume assayed is several cubic centimeters, whereas the volume of tissue obtained with a needle extraction is orders of magnitude less. There is optimistic hope that with modest improvements in sensitivity and control over the measurements, iron deficiency may also be detected. This would prove a major benefit for the differential diagnosis of several diseases.

9. SUMMARY

Advances in biomagnetic studies during the past twenty years have been made possible by the continued development of the superconducting sensor (SQUID) and in some cases by the commercial appearance of large arrays of SQUID sensors. Basic research in understanding how to interpret magnetic-field patterns of the brain, heart, lung, and liver has provided a framework for ongoing clinical and occupational health studies. There are three sources of biomagnetic fields: ionic electric current, which is important in the brain, nerve, and heart; materials with permanent magnetic moments that are taken up from the environment, as in respiration; and the magnetic susceptibility of tissue whose atomic electrons respond to an applied field by producing their own, much weaker, field. The latter is an important signal indicating the progress of diseases that produce iron overload in the liver.

A diverse range of biological topics are within reach of biomagnetic techniques. In some cases, as for neuromagnetism and pneumomagnetism, the applications are making fundamental contributions to our understanding of physiology, with corresponding benefits for clinical applications. In others, such as cardiomagnetism, the technique is a complement to existing procedures, and ultimate acceptance rests on the balance between benefit and relative costs of competing technologies. New applications of magnetic measurements continue to be found, and the ever-increasing sensitivity of SQUID detectors ensures that this will continue to be true for many years.

GLOSSARY

Action Potential: Propagating excitation along a neuronal axon.

Current Dipole: Short element of current, representing intracellular current flow within a single cell or population of cells. Serves as the simplest model for a source of magnetic field.

Detection Coil: Superconducting coil connected as the input to a SQUID magnetic field sensor.

Flux Gate: A magnetic-field-measuring instrument that relies on an ac system to detect the response of its magnetic core.

Induction Coil: A coil of wire used to indicate the strength of an ac field by the strength of the voltage induced in it.

Macrophage: Mobile cell on the lung's alveolar surface that scavenges foreign materials.

Neuron: Nerve cell.

Primary Source: Current distribution representing active tissue, often modeled as a current dipole or arrangement of current dipoles.

Secondary Source: Current distribution along the locations of boundaries between media of differing electrical conductivity, to represent their effect on the extracellular current when the body is represented as a medium of uniform conductivity.

SQUID: Superconducting quantum interference device, an extremely sensitive device for measuring magnetic fields.

List of Works Cited

Barth, D. S. (1984), "Neuromagnetic Measurements of Epilepsy in Animal and Man," Ph.D. thesis, Department of Psychology, University of California at Los Angeles.

Barth, D. S., Sutherling, W., Engel, J., Jr., Beatty, J. (1984), "Neuromagnetic Evidence of Spatially Distributed Sources Underlying Epileptiform Spikes in the Human Brain," *Science* **223**, 293–296.

Baule, G. M. and McFee, R. (1963), "Detection of the Magnetic Field of the Heart", *Am. Heart J.* **66**, 95–96.

Brain, J. D., Bloom, S. B., Valberg, P. A., Gehr, P. (1984), "Correlation between the Behavior of Magnetic Iron Oxide Particles in the Lungs of Rabbits and Phagocytosis," *Exp. Lung Res.* **6**, 115–131.

Brenner, D., Williamson, S. J., Kaufman, L., (1975), "Visually Evoked Magnetic Fields of the Human Brain," *Science* **190**, 480–482.

Brittenham, G. M., Farrell, D. E., Harris, J. W., Feldman, E. S., Danish, E. H., Muir, W. A., Tripp, J. H., Bellon, E. M., (1982), "Magnetic Susceptibility Measurement of Human Iron Stores," *New Engl. J. Med.* **307**, 1671–1675 .

Buchanan, D., Paulson, D., Williamson, S. (1988), "Instrumentation for Clinical Applications of Neuromagnetism," in: R. Fast (Ed.) *Advances in Cryogenic Engineering*, New York: Plenum, pp. 97–106.

Cohen, D. (1968), "Magnetoencephalography: Evidence of Magnetic Fields Produced by Alpha-Rhythm Currents," *Science* **161**, 784–786 .

Cohen, D. (1973), "Ferromagnetic Contamination in the Lungs and Other Organs of the Human Body," *Science* **180**, 745–748.

Cohen, D., Arai, S. F., Brain, J. B. (1979), "Smoking Impairs Long-Term Dust Clearance from the Lung," *Science* **204**, 514–517.

Cohen, D., Palti, Y., Cuffin, B. N., Schmid, S. J., (1980), "Magnetic Fields Produced by Steady Currents in the Body," *Proc. Natl. Acad. Sci. USA* **77**, 1447–1451.

Farrell, D. E., Tripp, J. H., Zanzucchi, P. E., Harris, J. W., Brittenham, G. M., Muir, W. A. (1980), "Magnetic Measurement of Human Iron Stores," *IEEE Trans. Magn.* **MAG-16,** 818–823 .

Geselowitz, D. B. (1970), "On the Magnetic Field Generated Outside an Inhomogeneous Volume Conductor by Internal Current Sources," *IEEE Trans. Magn.* **MAG-6,** 346–347.

Hari, R., Kaukoranta, E. (1985), "Neuromagnetic Studies of the Somatosensory System: Principles and Examples," *Progr. Neurobiol.* **24**, 233–256.

Hari, R., Lounasmaa, O. V. (1989), "Recording and Interpretation of Cerebral Magnetic Fields," *Science*, **244**, 432–436.

Hoke, M., Feldmann, H., Pantev, C., Lütkenhöner, B., Lehnertz, K. (1989b), "Objective Evidence of Tinnitus in Auditory Evoked Magnetic Fields," *Hearing Res.* **37**, 281–286.

Ilmoniemi, R. J., Williamson, S. J., Kaufman, L., Weinberg, H., Boyd, A. (1988), "Method for Locating a Small Magnetic Object in the Human Body," *IEEE Trans. Biomed. Eng.* **35**, 561–564.

Kajola, M., Ahlfors, S., Ehnholm, G. J., Hällström, J., Hämäläinen, M. S., Ilmoniemi, R. J., Kiviranta, M., Knuutila, J., Lounasmaa, O. V., Tesche, C. D., Vilkman, V. (1989), "A 24-Channel Magnetometer for Brain Research," in: S. J. Williamson, M. Hoke, G. Stroink, M. Kotani (Eds.), *Advances in Biomagnetism*, New York: Plenum, pp. 673–676.

Katila, T., Madiewshi, R., Poutanen, T., Varpula, T., Karp, P. J. (1981), "Magnetic Fields Produced by the Human Eye," *J. Appl. Phys.* **52**, 2565–2571.

Kaufman, L., Schwartz, B., Salustri, C., Williamson, S. J. (1990), "Modulation of Spontaneous Brain Activity during Mental Imagery," *J. Cog. Neurosci.* **2**, 124–132.

Kirschvink, J. L. (1983), "Biogenic Ferrimagnetism: A New Biomagnetism," in: S. J. Williamson, G. Romani, L. Kaufman, I. Modena (Eds.), *Biomagnetism: An Interdisciplinary Approach*, New York: Plenum, pp. 501-531.

Lu, S., Hämäläinen, M. S., Hari, R., Ilmoniemi, R. J., Lounasmaa, O., Sams, M., Vilkman, V. (1991), "Seeing Faces Activates Three Separate Areas Outside the Occipital Visual Cortex in Man," *Neuroscience*, in press.

Nakaya, Y., Mori, H., Nomura, M. (1989), " Clinical Value of Magnetocardiographic Mapping" in: S. J. Williamson, M. Hoke, G. Stroink, M. Kotani (Eds.), *Advances in Biomagnetism*, New York: Plenum, pp. 403–408.

Narici, L., Pizzella, V., Romani, G., Traversa, R., Rossini, P. (1990), "Evoked Alpha and Mu Rhythm in Humans: A Neuromagnetic Study," *Brain Res.*, in press.

Nenonen, J., Purcell, C., Horacek, B. M., Stroink, G., Katila, T. (1991), "Magnetocardiographic Functional Localization Using a Current Dipole in a Realistic Torso," *IEEE Trans. Biomed. Eng.* **38**, 658–664.

Okada, Y. C., Kaufman, L., Williamson, S. J. (1983), "The Hippocampal Formation as a Source of the Slow Endogenous Potentials," *Electroenceph. Clin. Neurophysiol.* **55**, 417–426.

Okada, Y. C., Lauritzen, M., Nicholson, C. (1988), "Magnetic Field Associated with Spreading Depression: A Model for the Detection of Migraine," *Brain Res.* **412**, 185–190.

Okada, Y. C., Williamson, S. J., Kaufman, L. (1982), "Magnetic Field of the Human Sensorimotor Cortex," *Intern. J. Neurosci.* **17**, 33–38.

Pantev, C., Hoke, M., Lütkenhöner, B., Lehnertz, K. (1988), "Influence of Stimulus Intensity on the Location of the Equivalent Current Dipole in the Human Auditory Cortex," in: K. Atsumi, M. Kotani, S. Ueno, T. Katila, S. J. Williamson, (Eds.), *Biomagnetism '87*, Tokyo: Tokyo Denki University Press, pp. 146–149.

Romani, G. L., Williamson, S. J., Kaufman, L. (1982), "Biomagnetic Instrumentation," *Rev. Sci. Instrum.* **53**, 1815–1845.

Roth B., Wikswo J., Jr. (1985), "The Magnetic Field of a Single Axon," *Biophys. J.* **48**, 93–109.

Sutherling, W., Crandall, P., Cahan, L., Barth, D. (1988), "The Magnetic Field of Epileptic Spikes Agrees with Intracranial Localization in Complex Partial Epilepsy," *Neurology* **38**, 779–786.

Valberg, P. A., Feldman, H. A. (1987), "Magnetic Particle Motions within Living Cells," *Biophys. J.* **52**, 551–561.

Walcott, C., Gould, J. L., Kirschvink, J. L. (1979), "Pigeons Have Magnets," *Science* **205**, 1027–1029.

Wikswo, J. P., Jr. (1983), "Cellular Action Currents," in: S. J. Williamson, G. Romani, L. Kaufman, I. Modena (Eds.), *Biomagnetism: An Interdisciplinary Approach*, New York: Plenum, pp. 173–207.

Williamson, S. J., Kaufman, L., Brenner, D. (1978), "Latency of the Neuromagnetic Response of the Human Visual Cortex," *Vision Res.* **18** 107–110.

Williamson, S. J., Kaufman, L. (1987), " Analysis of Neuromagnetic Signals," in: A. S. Gevins, A. Rémond (Eds.), *Methods of Analysis of Brain Electrical and Magnetic Signals*, Amsterdam: Elsevier, Chap. 14, pp. 405–448.

Williamson, S. J., Hoke, M., Stroink, G., Kotani, M. (Eds.) (1989), *Advances in Biomagnetism*, New York: Plenum, 771 pages.

Williamson, S. J., Romani, G. L., Kaufman, L., Modena, I. (Eds.) (1983), *Biomagnetism: An Interdisciplinary Approach*, Vol. 66 of NATO Advanced Science Institutes Series. Series A.: Life Sciences, New York: Plenum.

Zimmerman, J. E., Frederick, N. V. (1971), "Miniature Ultrasensitive Superconducting Magnetic Gradiometer and its Use in Cardiography and other Applications," *Appl. Phys. Lett.* **19**, 16–19.

Further Reading

Brittenham, G. M., Farrell, D. E., Harris, J. W., Feldman, E. S., Danish, E. H., Muir, W. A., Tripp, J. H., Bellon, E. M. (1982), "Magnetic Susceptibility Measurement of Human Iron Stores," *New Engl. J. Med.* **307**, 1671–1675. The basic paper describing how magnetic measurements determine the amount of liver iron overload.

Hari, R., Lounasmaa, O. V (1989), "Recording and Interpretation of Cerebral Magnetic Fields," *Science* **244**, 432–436. How functions of the human auditory cortex can be mapped from MEG studies.

Nenonen, J., Purcell, C., Horacek, B. M., Stroink, G., Katila, T. (1990), "Magnetocardiographic Functional Localization Using a Current Dipole in a Realistic Torso," *IEEE Trans. Biomed. Eng.* **38**, 658–664. An illustration of the importance of using a realistic model to represent the body's torso when localizing disfunctions in the cardiac conduction system.

Sutherling, W., Crandall, P., Cahan, L., Barth, D. (1988), "The Magnetic Field of Epileptic Spikes Agrees With Intracranial Localization in Complex Partial Epilepsy," *Neurology* **38**, 779–786. An example of how the site of epileptic spikes is identified.

Williamson, S. J., Hoke, M., Stroink, G., Kotani, M. (Eds.) (1989), *Advances in Biomagnetism*, New York: Plenum, 771 pages. Comprehensive overview of biomagnetic applications, beginning with a series of tutorials for nonexperts.

BIOMASS ENERGY

Donald K. Walter,* *Severna Park, Maryland, U.S.A.*

	Introduction	473
1.	**Energy Consumption**	473
2.	**The Resource**	474
2.1	Conventional Wood and Wood Wastes	474
2.2	Agricultural Crops, Wastes, and Residues	474
2.2.1	Crops	474
2.2.2	Wastes	474
2.2.3	Excess Corn	474
2.3	Municipal Wastes	474
2.3.1	Residential Wastes	475
2.3.2	Commercial, Institutional, and Industrial Wastes	475
2.3.3	Sludge	475
2.3.4	Recycling	475
2.3.5	Biogas	475
2.4	Energy Crops	475
2.4.1	Terrestrial Crops	475
2.4.2	Aquatic Crops	476
2.5	International Considerations	476
3.	**Conversion Technologies**	477
3.1	Preparation	477
3.2	Thermochemical Conversion	477
3.2.1	Combustion	477
3.2.2	Gasification	479
3.2.2.1	Reactors	480
3.2.2.2	Utilization	480
3.2.3	Liquefaction	480
3.3	Biochemical Conversion	481
3.3.1	Fermentation	481
3.3.1.1	Future Ethanol Production	481
3.3.1.2	Pretreatment	482
3.3.1.3	Hydrolysis	482
3.3.1.4	Xylose Fermentation	482
3.3.1.5	Lignin	482
3.3.2	Anaerobic Digestion	482
4.	**Biomass Energy and the Environment**	483
4.1	Agricultural Sector	483
4.2	Forestry Sector	483
4.3	Greenhouse Gases	483
4.4	Acid Gases	485
4.5	Other Emissions	485
	Glossary	485
	Works Cited	486
	Further Reading	487

INTRODUCTION

Biomass converted to fuel (biofuels) has had a significant role in the supply of energy throughout history. Until the middle of the 19th century, it was the predominant source of energy providing heat, cooking, and light. During the late 19th century, the industrial revolution demanded more energy-dense fuels to produce steam. Coal filled that need until it was supplanted by petroleum, particularly as the demand for liquid transportation fuel arose in the early 1900s. During the 1930s and 40s, natural gas, then considered a waste product from oil wells, entered the marketplace as a source of fuel and chemical feedstock. With the increase in environmental awareness in the 1960s, biomass began a resurgence and today provides about 4% of the United States' energy needs. In addition, it supplies as much as 30% of some lesser developed countries' energy needs. With further utilization of existing and development of the resource and/or new improved conversion technologies, it can be a major energy supply source that has positive impact on domestic energy supply, the environment, the forest products industry, farm crops, and local job markets.

1. ENERGY CONSUMPTION

Today the United States has three principal sources of primary energy: coal, petroleum,

---

*Mr. Walter is the Director of the Biofuels and Municipal Waste Division, U.S. Department of Energy, Washington, D.C. Any opinions expressed are those of the author and not those of the Department of Energy.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Table 1. Energy consumption by source in the United States (in EJ).

Source	Consumption as primary energy
Oil[a]	35.79
Natural gas[a]	19.60
Coal[a]	19.82
Nuclear electricity[a]	5.98
Hydroelectric[a]	2.76
Other[a]	0.30
Biomass[b]	3.69

[a] From Energy Information Administration (1988).

[b] From Energy Information Administraton (1989) and National Wood Energy Association (1988).

and gas. Other sources are nuclear energy, hydropower, biomass, geothermal, and other renewables such as wind and photovoltaics. Table 1 lists the consumption of energy by source. Further reading on energy consumption in the United States can be found in Energy Information Administration (1988, 1989) and National Wood Energy Association (1988).

2. THE RESOURCE

There is a broad range of materials which make up the biomass resource. Table 2 lists biomass resources by resource measurement category and current consumption.

2.1 Conventional Wood and Wood Wastes

This resource includes the standing forests of the United States and the wastes from various forestry and wood product operations. It has a moisture content of 50% when green and physical characteristics dependent on the source. Examples of wood waste include the limbs and slash left in the forest during logging, sawdust from woodworking and sawmill operations, and bark and sludge from paper making. The principal energy use today is to cogenerate steam and electricity for the paper and lumber industries. The second largest use is for home heat in the form of logs, wood waste, and pellets. Wood averages about 19.5 MJ/kg on a moisture- and ash-free basis.

2.2 Agricultural Crops, Wastes, and Residues

This is a diverse resource including residues from food crops, wastes from food processing, and manures. They are more moist than wood and are more diverse physically.

2.2.1 Crops. Crops such as plant oils from seeds (sunflower, rapeseed, safflower, etc.) can be used as diesel fuel, particularly for on-farm vehicles. Engine trials indicate severe coking of diesel injectors unless the oils are modified by transesterification with methanol or ethanol. Then they become suitable diesel substitutes. Certain species such as guayule and euphorbia produce latexes that can be converted to liquid fuels.

2.2.2 Wastes. Wastes such as corn cobs and stover, wheat straw, and rice husks are used to generate heat for drying, steam for industry, and electricity for the grid. Manures and food-processing wastes are used to produce biogas for boiler and engine fuel.

2.2.3 Excess Corn. Excess corn and other grains, molasses, and food-processing wastes are used to produce fuel ethanol.

2.3 Municipal Wastes

These are the wastes of residential, commercial, and institutional activities and the mixed industrial wastes that are discarded in

Table 2. Biomass resources (in quads).

Resource	Total	Accessible	Reserves	Current
Forests & wood wastes	442	20	12.5	2.5
Agricultural wastes & manures	10.5	5.2	1.1	0.2
Municipal & industrial wastes	2.4	1.8	1.6	0.2
Recycling	0.8	0.4	0.4	0.2
Energy crops	55	28.0	9.2	0.0
Aquatic crops	5	1	0	0

sanitary landfills. Municipal wastes are burned directly for heat and/or electricity, used to produce refuse-derived fuel (RDF), or recycled. Municipal waste, as are most wastes, is a unique fuel. Typically, it is delivered free. Further, a disposal or tipping fee is paid to the owner or operator of the energy plant. This fee is based on the cost of disposing of the waste by other means. For example, the cost of disposing of waste in a specific Long Island community is $150 per ton (1988) whether sent to the local waste-to-energy plant or hauled off the island. The negative cost of waste as a fuel is increasing as the environmental costs of landfill as an alternative disposal system are recognized and internalized.

2.3.1 Residential Wastes. Residential wastes are heterogeneous with roughly 55% organic matter (mixed paper, food and yard wastes, plastic, textiles, and urban wood), 25% moisture, and 20% inorganics (iron, glass, aluminum, other metals, dirt, and grit). They average about 21.8 MJ/kg on a moisture- and ash-free basis.

2.3.2 Commercial, Institutional, and Industrial Wastes. Commercial, institutional, and industrial wastes tend to be more uniform, drier, and of high combustible content. However, they are closer to wood and average about 19.5 MJ/kg on a moisture- and ash-free basis.

2.3.3 Sludge. Sludge from sewage treatment plants is included in the municipal wastes. Sewage sludge is 92%–95% moisture as produced, and depending on the source may be dewatered mechanically to 20%–40% solids. The energy content is about 25.5 MJ/kg on a moisture- and ash-free basis.

2.3.4 Recycling. There are several components of municipal solid waste that are recyclable, including paper and plastics, aluminum and iron, and glass. The best recyclable component is aluminum beverage containers. They have a positive market value in part because each kilogram conserves 220 MJ when compared to production from ore. For the last few years, aggressive recovery activities by aluminum companies and a purchase price of $600–$1000 per ton have resulted in 55%–60% of all aluminum beverage containers being source separated and recycled. By tonnage, waste paper has been the principal recycled material. Manufacturing waste newsprint conserves about 4.6 MJ/kg. However, because of oversupply, collectors have had to pay to dispose of separated newsprint on the East Coast for the year 1989. Other paper such as corrugated boxes requires as much energy to recycle as to produce from virgin material. Further reading on the energy aspects of several recyclable materials can be found from United States Department of Energy (1982).

2.3.5 Biogas. Biogas is recovered from landfills and sewage treatment plants. It is used as a boiler fuel, as a fuel for engine generator sets, as natural gas after the carbon dioxide is scrubbed, and in a few instances compressed as an automotive fuel.

2.4 Energy Crops

A significant increase in the availability of biomass resources will require the development of new and improved crops. These crops can be suited for terrestrial or aquatic environments. Energy crops are at varying stages of development.

2.4.1 Terrestrial Crops. Terrestrial energy crops can be either trees or grasses depending on the conversion process and the land to be utilized.

There are two broad categories of trees, coniferous or softwood and deciduous or hardwood. Softwoods are used commercially for pulp and dimension lumber. Research has improved their growth from 2.5 to 7.5 metric tons per hectare and their fiber yield. Hardwoods have been used for high value specialty purposes such as furniture and veneers. The natural supply has been sufficient to meet demand and no improvement was undertaken until their value as energy crops was recognized. Hardwoods are preferred as an energy crop since certain species grow faster than softwoods and they coppice when cut. Further, second and third growth cycles are proving to be more productive than the first cycle. Because of the broad diversity of land forms, soils, and climate, several species are required. Six species have been selected for development from 140 surveyed based on growth rate, survivability in adverse conditions, nutrient requirements, and ability to coppice. The principal genus in development is *Populus* (poplar, cottonwood, etc.). The *Populus* species are candidates for growth in every region of the United States. The other species are more limited in range. Black locust is a nitrogen-fixing tree with range limited to the midwest and southeast. Silver maple has been selected for the midwest, great plains, and northeast; American sycamore for the southeast, sweet gum for the southeast, and eucalyptus for subtropical regions. In nature, these trees produce up to 7.5 dry metric tons per hectare per year (DT/H). However, by selection of fast-growing clones, hybridization, and biotechnological manipulation, growth rates of over 30 DT/H have been reached and an average United States growth of 25 DT/H is reasonably achievable. Since the nutrients concentrate in the leaves and energy wood is harvested in the winter while dormant, fertilizer requirements are minimal. Since the trees are closely spaced (2 m or less), the leaves and branches shade the ground and weeds are controlled.

As with wood species, there are two broad categories of agricultural crops, the annuals which die and require replanting each year, and the perennial crops which retain viable root mass each year. The perennials are preferred as energy crops since annual reestablishment is not required and the living root system reduces erosion. Selection of target species is still ongoing in the United States; however, the thick-stemmed perennial grasses (e.g., napier grass) are favored. The annuals such as sorghum are being considered because of their fast growth rates. Energy cane (a variety of sugar cane) has produced 150 DT/H in Puerto Rico.

The availability of land and the potential for competition between food and fiber and energy and chemical uses are of concern. However, world economies are learning to grow food. Two major importers of food in the past, India and China, now export food. Exports of large United States surpluses are decreasing. The federal government is subsidizing farmers to keep land out of production. Twelve million hectares of land are in the Conservation Reserve Program which removes highly erodible land from agriculture practice. Trees are approved for planting on these lands to stabilize the soil. The federal government pays to keep 20 to 30 million hectares of land fallow to reduce surplus crops. The Agricultural Research Service has identified a need for alternative crops for 60 million hectares of current farm land by 2010. Further, the Department of Agriculture has identified 62 million hectares of land, now in pasture and range, that are suited for farmland. This total acreage at an average of 25 DT/H could provide 58 EJ (55 quads) of energy per year.

2.4.2 Aquatic Crops. There are three potential aquatic biomass resources: the emergents or marsh crops such as cattails, macroalgae such as giant kelp grown in the open sea, and microalgae.

Emergents have a resource base limited primarily to the marsh land of the upper midwest. They are difficult to harvest and to match to a conversion process which must have a year-round supply of feedstock. Macroalgae are difficult to grow and harvest although they have some valuable energy-crop characteristics such as the 45% sugar content of kelp.

Microalgae are single-cell organisms that live in either fresh or saline water, grow rapidly, and under stress fix large quantities of lipid (plant oil). These oils are typically hydrocarbons in the C_{18} to C_{22} range and are similar to the plant oils from seeds. The same transesterification will be needed to convert them to suitable diesel fuels.

2.5 International Considerations

Worldwide, there is broad interest in the use of biomass as an energy resource. Many of the developing countries generate as much as 30% of their primary energy from biomass. In many cases the resource is overused and actions are necessary to reverse the trend through improved efficiency and replanting. For example, inefficient stoves in the sub-Saharan Desert region are being replaced and new trees planted to halt desertification of the area.

The major food-raising countries all have an interest in the improved use of agricultural and wood wastes. Under the auspices of both internal and international agencies there is increasing use of bagasse from sugar cane and rice husks for the cogeneration of steam and electricity for refinery operation. Many countries are considering the infrastructure and legal changes necessary to supply electricity from wastes to the nations' electric grid. The Brazilians have embarked on a major program to use ethanol from cane sugar as a gasoline replacement.

The industrialized nations are more advanced than the United States in the productive use of municipal solid waste. While we use perhaps 20% for energy and recycling other countries make use of up to 75% of their waste.

Worldwide there is increasing interest in the growth of energy crops. As noted, Brazil produces ethanol from cane sugar. In addition, they have a great deal of expertise in fast-growing eucalyptus trees. Now they are engaging in research to produce ethanol from those trees. England, Sweden, Canada, Finland, and the United States cooperate under the auspices of the International Energy Agency. Belgium maintains a bank of over 3000 *Populus* species clones. In fact, several of the United States commercial plots have been planted with clones from that bank. India and the United States have a bilateral agreement in part to develop energy crops. The Indians are the preeminent researchers in species of *Populus* for alkaline soil conditions. However, statistics are not readily available for the worldwide potential for energy crops.

3. CONVERSION TECHNOLOGIES

There are two broad categories for the conversion of biomass to a fuel form, thermochemical and biochemical. The former uses heat and is typified by fast reaction times (seconds to minutes) and harsh conditions (high temperatures from 300 °C to over 1000 °C and pressures of up to 200 atm). The products tend to be variable. The biochemical systems use enzymes (natural catalysts) and are typically slow (reaction times of hours to days) and mild (temperatures from room to the boiling point of water, and atmospheric pressure). The products are specific.

3.1 Preparation

The design and operation of any biofuels system depends on the form of the feedstock. In general, smaller and more uniform particles favor conversion. However, obtaining such fuel can be expensive. The simplest form of preparation is drying, and the most complex is the separation of municipal solid waste (MSW) into various components. A recent innovation is a combustion system that can burn whole trees.

The preparation of biomass has two aims: provision of a more uniform fuel and separation of the fuel into components. Wood generally is used as chunk wood (5- to 15-cm pieces), hog fuel (2.5 to 5 cm), and sawdust. Finer particles may be compressed into pellets or briquettes. Screens are used to remove rocks and other inerts, and magnets to remove tramp metal such as saw teeth, barbed wire, etc. MSW is used as received or processed to produce refuse-derived fuel (RDF), materials for recycling, and inerts. MSW is prepared by screens of various designs for size separation, hammermills, and flail mills for size reduction and a series of technologies to separate specific components, such as magnets to recover ferrous metals. Refuse-derived fuel and magnetic metals have been recovered successfully. The technology to recover glass and aluminum has been developed. MSW separation is an art unto itself. Further reading can be found in Diaz *et al.* (1982) and Hasselriis (1984).

3.2 Thermochemical Conversion

As biomass is heated, the long chains of molecules that comprise cellulose, hemicellulose, and lignin begin to break into simpler molecules and become modified. The final form of those molecules depends on the rate at which heat is applied, on the length of heat application, on the final temperature, on the external oxygen available, and to a lesser extent on the pressure. High heat rates, short residence times, and limited oxygen can favor the retention of long-chain molecules that are liquid at standard temperature and pressure. As reaction time increases, heat rate decreases, final temperature increases, and oxygen levels increase, the molecules become shorter until with sufficient time and oxygen, the fuel is completely oxidized.

3.2.1 Combustion. The most common technology for the use of biomass as a fuel today is combustion. There are thousands of facilities that burn biomass and MSW to provide industrial steam and heat, electricity, and district heating. In addition, over 2 million homes use wood stoves as a principal means of heat.

During combustion, fuel is first dried to drive off free water, then the long chains of molecules are broken until essentially the carbon reacts with oxygen to form carbon dioxide. The general chemical equation is

$$C_x(H_2O)_y + xO_2 \rightarrow xCO_2 + yH_2O. \qquad (1)$$

In an ideal combustion process only the stoichiometric amount of oxygen is added. In practice, excess air (oxygen) must be added to maximize the oxidation of the carbon and minimize the formation of carbon monoxide. Figure 1 is a typical graph of carbon monoxide versus oxygen. Since these graphs are a function of the specific fuel and conditions of combustion, no units of measurement have been added. As the amount of excess air is increased, the combustion process is cooled, carbon monoxide production increased, and overall efficiency reduced. A modern biofuels combustion system can reduce carbon monoxide emissions to 30 parts per million (ppm) and achieve efficiencies in the 65%–75% range. Here, efficiency is defined as the steam produced per unit of total-energy input. For comparison a modern natural gas plant is 80% efficient, fuel oil 85%, and coal 90%. The differences are principally related to the amount of water produced per joule of fuel burned both from the conversion of fuel hydrogen to water and from free and fuel-bound water.

The success of the combustion of biomass is a function of the design of the system. The majority of biomass combustion facilities have a grate to retain fuel for complete combustion and to distribute part of the combustion (underfire) air. Grates vary from a simple slab of metal with small air distribution holes (pinhole grate) for sawdust to complex reciprocating grates for raw MSW. The Energy Research and Development Administration (1977) provided sketches of a series of MSW grate systems. In addition to the primary air delivered through the grate, secondary air is delivered above the grate to promote turbulence and to ensure complete combustion.

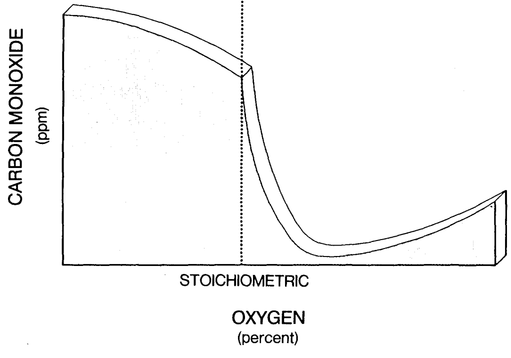

FIG. 1. Carbon monoxide vs oxygen.

The grate is contained in a firebox which may have walls constructed from either refractory fire brick or a membrane of welded tubes (waterwall). In the latter case the tubes are filled with water which removes both radiant and sensible heat. Waterwall systems have integrated superheaters, economizers, and feedwater heaters to recover steam. Refractory systems have a separate waste heat boiler. Because radiant heat of the fire is captured, waterwalled systems tend to be more efficient than refractory systems; however, the refractory systems tend to have lower maintenance and repair costs. Further reading on design and a more complete description of boilers can be found in Robinson (1986), Chap. 13.

The quality (temperature and pressure) of the steam recovered depends on its end use and the overall system function. For example, industrial processes frequently require 1 MPa, saturated steam while electric generation is most efficient at high steam pressures and degrees of superheat. Particularly in MSW fired units, steam temperature is technically limited by the wall temperature of the superheater tubes. As the tubes are heated, they are subjected to two distinct corrosion patterns. First, there is a low-temperature peak of corrosion at about 150 °C followed by a high-temperature peak at around 450 °C. Corrosion at high temperatures is not well understood although it is most likely caused by a complex sulfur, iron, and chlorine reaction. Overall pressure is not a limiting factor. See Plumley (1977) for diagrams and further discussion on corrosion regimes.

An emerging thermal technology is the fluid bed combustor. The biomass fuel is combusted in a chamber that is partially full of sand. Air is injected below the bed and the sand acts as a fluid. The large mass of turbulent hot sand ensures very rapid heat transfer and complete combustion of the fuel. Heat is recovered by in-bed and above-bed tubes, waterwall tubes, and integrated boiler banks. If limestone is substituted for sand, the bed also captures chlorides and sulfur oxides. Fluid bed combustion units require less excess air than other biomass combustion designs and are more efficient in the production of steam. There are two variants of the fluid bed combustor. The atmospheric fluid bed uses low air velocities to retain sand in the combustion chamber. The circulating fluid bed uses high enough air velocities to elutriate sand. Hot sand and char are removed in a hot cyclone and recycled to the main bed. There are a significant number of fluid bed combustion units operating in the wood products industry, most providing hot air for drying and curing products.

Pollution-control systems for combustion systems include particulate and acid-gas controls. Table 3 has proximate and ultimate analyses for wood, MSW waste, and coal. Because of low-acid gas-forming components, wood-burning systems require only particulate control. Modern MSW combustion facilities generally are required to have sophisticated acid-gas controls. The acid-gas controls available include wet scrubbers, dry scrubbers, and direct lime injection into the combustion chamber. Wet scrubbers condense acid gases with a water spray. They leave an acid solution that is difficult to manage. Dry scrubbers spray a slurry of lime into the combustion gas stream and collect dry reagent in a bag house or electrostatic precipitator. They reduce chlorine emissions by 90% and sulfur by 70%. Direct injection of lime into the combustion chamber costs 10% that of dry scrubbing with 10% degradation of efficiency of removal. The control of acid gases in fluid beds is achieved by the selection of lime or limestone as the bed material.

Table 3. Proximate and ultimate analysis (percentage by weight). From Domalski *et al.* (1987).

	Wood	MSW	Coal
Proximate analysis (as received)			
Moisture	50.0	25.1	13.1
Volatile	40.3	43.8	27.9
Fixed	9.7	6.1	51.1
Ash	0.02	25.0	7.9
Ultimate analysis (dry)			
Carbon	48.9	39.5	74.3
Hydrogen	5.9	5.0	4.9
Oxygen	43.3	29.5	7.8
Nitrogen	0.03	0.5	1.4
Chlorine	0.01	0.4	NA
Sulfur	0.01	0.1	3.7
Ash	0.02	25.0	7.9

There are several particulate control systems including cyclones, baghouses, and electrostatic precipitators. Cyclones act as a centrifuge. They are relatively inefficient and not used in modern pollution-control systems. Baghouses pass combustion gas through finely woven bags that capture particulates. As a layer of particulate builds, efficiency of capture increases. When the pressure drop across the bag exceeds a set limit the bag is taken off line and cleaned with a reverse flow of clean air. They are favored in modern plants since they are more efficient in the removal of fine respirable particulate. Electrostatic precipitators generate an electric charge on particulates before passing the combustion gas between plates with the opposite charge to attract and hold the particulate until the unit is taken off line and the plates are cleaned. Both baghouses and electrostatic precipitators are more than 99% efficient.

3.2.2 Gasification. As the available oxygen is reduced in a combustion system, the carbon is no longer completely oxidized and combustible gases containing carbon monoxide, hydrogen, and other higher hydrocarbons are produced. Continued decrease of oxygen finally results in unreacted char that is mostly carbon and ash. The gasification reaction is endothermic; therefore, an external source of heat is required. The source of heat may be the combustion of char with air or oxygen, or an external heat source. When char is burned with air in the reactor, the product gas is diluted with nitrogen and has a heating value of 1.8–7.2 MJ per standard cubic meter (MJ/SCM) (low-energy gas). If pure oxygen is used, the diluting nitrogen is removed and a product gas with a heating value of 10.9–14.5 MJ/SCM (medium-energy gas) is produced. Finally, if the reactor is heated externally, the production of hydrocarbons such as methane is favored over carbon dioxide. The gas produced has a heating value of 12.7–19.9 MJ/SCM (medium-energy gas). The chemical reaction is

$$C_x(H_2O)_y + \text{heat} \rightarrow yCO + yH_2 + (x-y)C. \quad (2)$$

Since there is no way to exclude oxygen and free water from the reactor completely, a series of other reactions occur, and Eq. (2) is modified. At least part of the CO is shifted to hydrogen via the water-gas shift reaction [Eq. (3)], and if partial combustion is the entire source of heat, the carbon char is burned to carbon dioxide [Eq. (4)]. Other reactions and incomplete reactions combine to produce higher hydrocarbons such as methane and long-chain molecules that are solid or liquid at atmospheric conditions:

$$CO + H_2O \leftrightarrow CO_2 + H_2 \text{ (water-gas shift)}, \quad (3)$$

$$C + O_2 \rightarrow CO_2. \quad (4)$$

3.2.2.1 Reactors. There are a series of types of gasification reactors. In the updraft gasifier, a descending bed of feedstock is first dried, then gasified, and the char is burned (either in air or oxygen) to produce a rising stream of hot gas to drive the reactions. The downdraft gasifier partially combusts the surface of a bed of feedstock. An induced draft pulls the hot product gas down through the glowing bed of coals. As a result the majority of the long-chain molecules are cracked and tars are reduced to produce gas. This product gas is cleaner and more adaptable to the internal combustion and gas turbine engines. Also, the fluid bed reactor has been adapted to gasification. The high thermal mass of the bed with its rapid heat transfer and relatively long residence time is well adapted to gasification. Fluid bed reactors can operate with the source of external heat from partial combustion of the char in the bed, from fire tubes in the bed, or by external heating of the bed material. Further reading on gasification can be found in Reed and Das (1988).

3.2.2.2 Utilization. Low-energy gas cannot be transported economically over distances of more than a kilometer. Therefore, its use is mostly viewed as a fuel for a boiler or, where tar and ash are low, as an engine fuel. Most of the 40 or so applications to date have been as a boiler fuel although at least one gasifier–engine-generator system has been permitted. In addition to fuel uses, medium-energy gas is a suitable chemical feedstock. Again, low tars and ash are desirable characteristics. In addition, the CO/H_2 ratio may be adjusted by the water-gas shift reaction [Eq. (3)]. For example, if methanol (CH_3OH) is the end product, then the CO/H_2 ratio must be adjusted to 1/2.

3.2.3 Liquefaction. As reaction time and final temperature are decreased, more long-chain molecules survive when compared to gasification. These molecules condense at standard conditions to a viscous, oxygenated fuel called biocrude. Biocrude might be considered as a substitute for fuel oil but because of instability and corrosiveness, will more likely be upgraded to other liquids such as gasoline. One of the early, large-scale attempts to produce a liquid biofuel used MSW refined to a fine (less than 30 mesh) powder which was pyrolyzed while entrained in a hot gas stream. The unreacted char was removed and burned to provide the reaction gas stream. The liquid product was condensed to a viscous, relatively unstable, corrosive oil that was to be used as a partial replacement for fuel oil in a utility boiler. The combustible gas produced was used for drying the feedstock and for reactor heat. The risk of the technology and the lack of government funds to support further testing led the industrial partner to abandon the technology when there was a major refractory failure in the reactor. The high level of processing and the limited use of the product oil were contributing factors. A second system was constructed based on the high-pressure liquefaction of wood in a carrier of hot wood oil. This system successfully produced many barrels of biocrude, but was uneconomic.

Any of the reactors discussed under gasification could be adapted to liquefaction. However, only two systems are in active research. The first is based on fine wood in a carrier of product oil at high temperature and pressure. The biocrude oil produced is relatively low in oxygen and can be hydrotreated to produce first run (i.e., 72 octane) gasoline. The chemistry of the technology is complex and varies with the feedstock. The second technology uses a vortex reactor that transfers heat rapidly from a hot wall to fine biomass. Unreacted wood and char are removed by a cyclone and recycled. The product gas is passed through a zeolite catalyst to produce gasoline. More development is necessary before either technology can be fully assessed for commercial viability. Current models indicate that when fully developed the technology could produce gasoline for a wholesale cost of $1.00 per gallon.

3.3 Biochemical Conversion

Biochemical conversion takes advantage of enzymes to convert biomass to other more useful products. The two processes in use which are significant to energy are fermentation to produce ethanol and anaerobic digestion to produce methane. In the fundamental operation of life processes, genes, which are segments of DNA, produce enzymes which convert organic molecules to useful products. Enzyme production is self-regulated. A specific promoter initiates the production of enzyme. When either enough product is available or no more feedstock is present a repressor gene turns off the production of enzyme. For example, in human nutrition, organisms in the mouth produce alpha-amylase, an enzyme which reduces starch to sugar. Food more difficult to digest is reduced to sugar in the stomach by dilute acid. The sugar is absorbed in the bloodstream. Further reading on biotechnology and enzyme production can be found in Watson *et al.* (1983).

3.3.1 Fermentation. Fermentation processes have been used for thousands of years to produce beverage alcohol. More recently with the oil crises of 1973 and 1978, ethanol has been produced as a transportation fuel. Starch is reduced to glucose by commercial alpha-amylase enzymes [Eq. (5)]. The glucose is then fermented by a yeast such as *Saccharomyces cerevisiae* to ethanol and carbon dioxide [Eq. (6)]:

$$(C_6H_{10}O_5)_n + nH_2O + \text{yeast} \rightarrow nC_6H_{12}O_6, \quad (5)$$

$$C_6H_{12}O_6 \rightarrow 2C_2H_5OH + 2CO_2 + \text{heat}. \quad (6)$$

The conversion of glucose to ethanol is not a simple rupturing of appropriate atomic bonds in the glucose atom. The pathway was first described by Embden and Meyerhoff (Solar Energy Research Institute, 1981). There are 12 distinct steps that are catalyzed by different enzymes. First, a phosphate group is added, then that compound is changed to fructose and another phosphate group is added. This compound is cleaved into two different isomers of three carbon compounds, each with a phosphate group. One of the isomers (dihydroxyacetone phosphate) is modified by another enzyme into the second (glyceraldhyde 3-phosphate). A second phosphate group is added and a hydrogen ion removed. Next, one of the phosphate groups is replaced by a hydroxyl (OH) radical. Then the phosphate group is moved to a new location in the atom and a hydroxyl radical and a hydrogen ion are removed from different locations in the molecule. The second phosphate group is removed to produce pyruvate, which is reduced to carbon dioxide and acetaldehyde. Finally, a hydrogen ion is added to acetaldehyde to produce ethanol. Further information on this reaction and microbiology, chemistry, and ethanol production is provided in Solar Energy Research Institute (1981).

A beer with 8%–10% alcohol is produced. Any remaining unconverted organic material is filtered and the water removed by distillation to the azeotrope of a mixture of 93% ethanol and 7% water. The remainder of the water is separated by chemical alteration of the combined boiling points or by selective absorption.

The current fuel ethanol market uses starch or sugar as its feedstock. In the United States over 85% of the fuel-grade ethanol is made from corn starch. In 1988, about 330 bushels of excess corn were converted to 850 million gallons of fuel-grade ethanol. The remainder is made from other grains, molasses, and industrial waste. In Brazil the majority is made from sugar. However, the United States and Brazilian industries require subsidies to survive. The United States industry is supported by a rebate of 6¢ of the 9¢ federal gasoline excise tax if 10% ethanol is blended with the gasoline (equivalent to a 60¢ per gallon of ethanol rebate) or a 60¢ per gallon of ethanol income tax credit (rarely used). In addition, there are state tax incentives equivalent to 10¢–20¢ per gallon of ethanol in 20 states. In Brazil, where all gasoline is a blend with 20% ethanol and a significant number of new cars are designed for pure ethanol, the industry is supported by a tax on gasoline.

3.3.1.1 Future Ethanol Production. Sugar and starch prices are based on food and feed value. They are unlikely to produce economic ethanol until the current price of gasoline doubles. Cellulosic crops such as wood or grasses could provide an economic feedstock. These might be wastes or specific energy crops (see earlier section on terrestrial crops).

3.3.1.2 Pretreatment. To use crops, their three major components (cellulose, hemicellulose, and lignin) must be separated and the former two hydrolyzed to sugar. The paper industry hydrolyzes lignin with a strong alkaline solution to leave the cellulose and hemicellulose fibers intact. For ethanol production, the hydrolysis of the hemicellulose into five-carbon sugars (primarily xylose) is preferable. A number of technologies including weak acid, organosolv, ammonia–freeze-dry explosion, and steam explosion are used to hydrolyze the hemicellulose.

3.3.1.3 Hydrolysis. Cellulose is crystalline and difficult to hydrolyze. Acid digestion will reduce the cellulose molecule to glucose; however, control is difficult and the acid will degrade glucose to furfural. Acid hydrolysis was used in the United States and Canada to produce industrial ethanol in World War II. That practice stopped as soon as ethylene became available from oil after the war. The Russians still operate over 40 acid hydrolysis plants to convert wood to ethanol.

Also, cellulose can be hydrolyzed enzymatically. Certain yeasts produce a consortium of enzymes that degrade cellulose specifically to glucose. The enzyme endoglucosidase breaks the long- chain cellulose molecule at interior bonds. Exoglucosidase removes a disaccharide from the end of the cellulose chain. The third, beta-glucosidase, completes the reduction of the disaccharide to hexose which may then be fermented to ethanol. Betaglucosidase is the rate-limiting enzyme. Finally, hexose is fermented to ethanol via Eq. (6).

3.3.1.4 Xylose Fermentation. The hemicellulose that was hydrolyzed to xylose in the pretreatment step represents 25%–35% of the total feedstock. Its use is essential if the overall ethanol production system is to be economic. There is no known organism or enzyme that economically will ferment xylose at efficiencies comparable to glucose. However, an enzyme, xylose isomerase, produced by *E. Coli* naturally converts xylose to a different five-carbon sugar isomer, xylulose. Xylulose is fermented to ethanol by commercial enzymes. Normally, *E. Coli* does not produce large quantities of xylose isomerase. Biotechnology has been used to modify the appropriate repressor gene of *E. Coli* to increase xylose isomerase production. Today 70% of the hemicellulose can be fermented to ethanol. Incremental improvements are expected to increase that to 90%:

$$(C_5H_8O_4)_n + nH_2O \rightarrow nC_5H_{10}O_5, \qquad (7)$$

$$3C_5H_{10}O_5 + \text{enzymes} = 5C_2H_5OH + 5CO_2. \qquad (8)$$

The actual conversion path is from 3 xylulose to 1 fructose and 1 glyceraldhyde 3-phosphate. These then follow the pathway briefly described above for glucose fermentation.

3.3.1.5 Lignin. Lignin, the third component of plant matter, is not readily biodegradable, and therefore cannot be readily converted to ethanol at this time. However, it is a useful fuel that can be used to produce steam and electricity to power the overall plant. Today, lignin that is hydrolyzed in paper making is burned for energy, to recover chemicals, and for disposal. In addition, lignin can be converted to an aryl ether thermochemically. The process is one of hydrogenation and then reaction with either methanol or ethanol to produce methyl (MAE) or ethyl aryl ether (EAE). MAE and EAE are high-octane liquids that can be used in transportation fuel blends.

3.3.2 Anaerobic Digestion. Anaerobic digestion is a common occurrence in nature. It is likely to have been the source of today's natural gas. As biomass died in swamps and was submerged 50 to 250 million years ago, it was digested into methane and carbon dioxide. Some of the gas was trapped and, as time passed, water dissolved the carbon dioxide, leaving natural gas.

Like ethanol fermentation, anaerobic digestion is a complex process that relies on a host of organisms. First, carbohydrates are hydrolyzed to simple organic acids. These are converted to acetic acid by acetogenic bacteria:

$$(C_6H_{10}O_5)_n + 7nH_2O + \text{enzymes} \rightarrow 6nCH_3OOH. \qquad (9)$$

Then the strict anaerobic methanogens reduce acetic acid to methane and carbon dioxide:

$$CH_3OOH + \text{enzymes} \rightarrow CH_4 \text{ and } CO_2. \qquad (10)$$

Also in the methane production pathway is the reduction of carbon dioxide with hydrogen via

$$CO_2+4H_2 \rightarrow CH_4+2H_2O. \qquad (11)$$

A description of these pathways was given in National Research Council (1977). Chapter 19 of Robinson (1986) contains design parameters.

4. BIOMASS ENERGY AND THE ENVIRONMENT

Biofuels impact on many environmental sectors including the agricultural, forestry, greenhouse, energy conversion, and energy use sectors. The overall impact is a function of the specific fuel pathway; for example, the impact of an energy crop is different from the impact of a corn crop. Similarly, the impact of a combustion system is different from that of a fermentation process.

4.1 Agricultural Sector

The principal concerns of the agricultural sector are erosion and chemical releases to the environment. Erosion is caused by the lack of a stabilizing vegetation layer on the land. In general, it is associated with an annual row crop such as corn. Energy crops such as fast growing trees or perennial grasses leave viable root systems to hold soil in place. Currently, trees are approved for the purpose of controlling erosion on United States land which has been taken out of service under the Soil Conservation Reserve program. Chemical pollution results in surface runoff and ground-water contamination. Excess fertilizer causes algae blooms which cause oxygen depletion in receiving waters. Herbicides and pesticides may be poisonous to aquatic life. Energy crops require smaller applications of fertilizer and chemicals. For example, while an annual farm crop requires annual fertilizer applications of 260 kg/hectare, energy trees require 225 kg/hectare in the first year, and 110 kg/hectare in the second year, and no further applications until harvest. In addition, some fast-growing trees (e.g., black locust) are nitrogen-fixing species that might be intermingled to provide fertilizer to a plot. Bacteria living symbiotically in root structures actually provide for nitrogen fixation. Biotechnology may be able to transfer these characteristics to other fast-growing energy species. Herbicide application is required annually for agricultural crops but only in the first and second years after establishment for an energy crop. Then the crown closes and shade provides weed control. Pesticide applications are not well understood but the first energy tree clone biotechnologically altered for disease control has been planted in test plots. Overall, energy crops are expected to improve the agricultural environment.

4.2 Forestry Sector

The principal concern in the conventional forestry sector is the air pollution caused by the open burning of slash and waste developed in the harvesting of a timber stand. The development of uses and collection systems for these materials plus environmental regulation will lead to their recovery and use as energy feedstocks. In addition, the sawdust and wastes of the forest product industry increasingly are being used as energy sources rather than being landfilled or incinerated with poor combustion control. Today, the forest products industry is over 50% self-sufficient in energy production. Within 20 years, it expects to be 100% self-sufficient.

4.3 Greenhouse Gases

Incoming short-wave solar radiation is absorbed by the earth, clouds, and the atmosphere and reflected from the earth and clouds, and backscattered by air. Long-wave surface and atmospheric radiations are emitted to space. Some of this radiation is absorbed by CO_2, ozone, and water vapor from which it is reflected back to earth. Several gases are increasing in concentration in the atmosphere and may upset the normal balance of radiation so as to cause the earth's temperature to rise. The greenhouse gases of concern are CO_2, CH_4, NO_x, ozone, and CFC's (chlorofluorocarbons such as Freon). Biofuels can affect the release of the first three and generally are not involved in the generation of the last two. Figure 2, adapted from Green (1989), displays the global average earth energy balance. The buildup of CO_2 and NO_x is directly related to the combustion of fossil fuels while the increase in CH_4 is related to food production and waste disposal (see AIR POLLUTION: COMPONENTS, CAUSES, AND CURES).

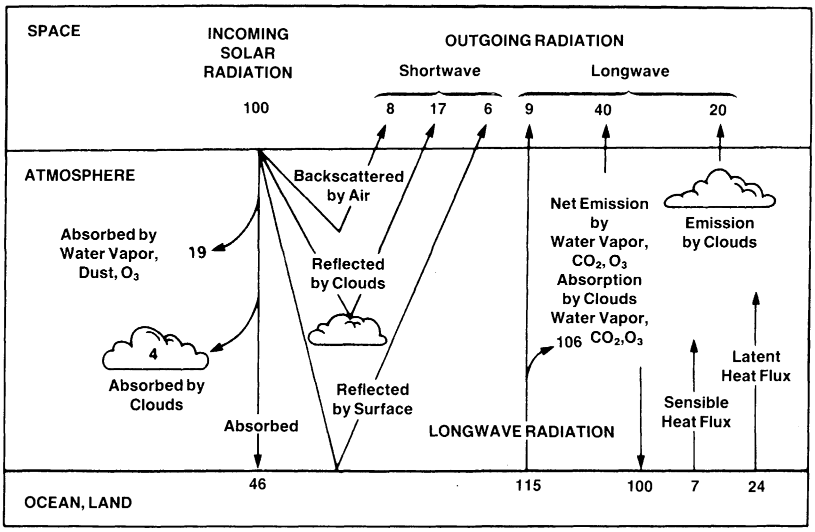
SPACE
INCOMING SOLAR RADIATION
100
OUTGOING RADIATION
Shortwave
8
17
6
Longwave
9
40
20
ATMOSPHERE
Backscattered by Air
Absorbed by Water Vapor, Dust, O3
19
Reflected by Clouds
4
Absorbed by Clouds
Reflected by Surface
Absorbed
Net Emission by Water Vapor, CO2, O3
Absorption by Clouds Water Vapor, CO2, O3
106
Emission by Clouds
Latent Heat Flux
Sensible Heat Flux
LONGWAVE RADIATION
OCEAN, LAND
46
115
100
7
24

FIG. 2. Earth's energy balance.

Carbon dioxide is produced during the combustion of any fuel. Two important fuel characteristics are the heating value per pound and the hydrogen-carbon ratio. Figure 3 is adapted from Green (1989) and gives these values for several common fossil and biofuels. The higher the hydrogen-carbon ratio, the less carbon dioxide produced per unit of energy. Thus the best fuel is pure hydrogen and the worst, anthracite coal. The quantity of carbon dioxide in the atmosphere has been increasing since the beginning of the industrial revolution and currently is about 350 ppm (up from 280 ppm in 1850). The uncontrolled burning of biomass fuel releases both carbon dioxide and methane. However, when a biomass waste is burned, no new net carbon dioxide is added to the atmosphere since the biomass will decay in a relatively short period. Further, when an energy crop is grown for energy, the carbon dioxide absorbed in the growth is released in the utilization process. Thus the carbon dioxide is recycled.

Methane is present in significantly lesser

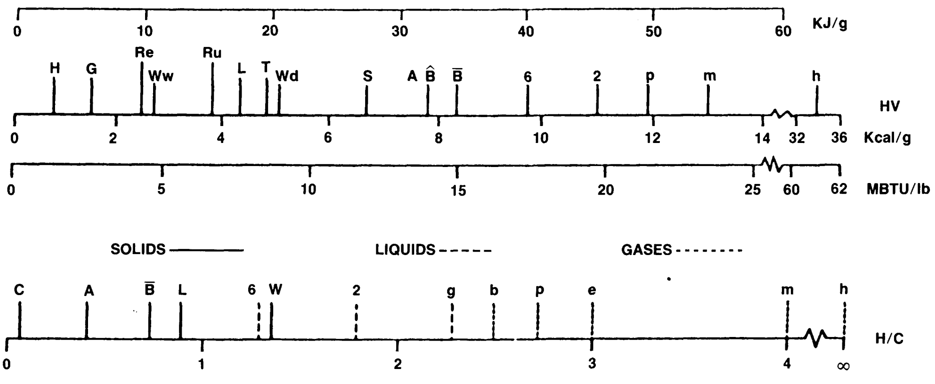

LEGEND:
Upper Scale, HEATING VALUE; Lower Scale, HYDROGEN/CARBON RATIO.
Symbols denote the following: A–anthracite, B–Low volatile bituminous, G–Garbage, H–High volatile bituminous, h–Hydrogen, L–Lignite, M–Methane, Re–Refuse, Ru–Rubbish, S–Sub-bituminous, T–Trash, W–Wood, Wd–Dry wood, Ww–Wet wood, 2–No. 2 oil, 6–No. 6 oil.

FIG. 3. Fuel characteristics.

amounts than CO_2; however, methane is a much more reactive gas than CO_2 in the upper atmosphere. While expert opinion varies, recently the EPA has published the relative reactivity of methane as being 30 times that of carbon dioxide as a greenhouse gas. Further reading on greenhouse gas issues can be found in Walter *et al.* (1989). The specific impact of greenhouse methane is particularly important in the utilization of wastes. Municipal waste decays anaerobically in landfills and in certain circumstances biomass may also decay anaerobically. The equations of this decay are the same as Eqs. (9) and (10), that is, for every molecule of CO_2, a molecule of CH_4 is produced. Each molecule of methane is estimated to have 30 times [United States Environmental Protection Agency (1990)] the greenhouse effect of a molecule of carbon dioxide. Therefore, productively using organic waste not only reduces the relative CO_2 effect from 31 to 2, but also displaces CO_2 that would have been produced from fossil fuel.

4.4 Acid Gases

Another environmental issue is the release of acid gases and more specifically sulfur oxides into the atmosphere. The conversion processes generally collect much of the acid-gas component of the primary feedstock, e.g., refining crude oil to motor fuel collects the majority of sulfur that is naturally present in oil. Despite this, gasoline still contains about 0.4% sulfur and diesel fuel up to 3% sulfur. In contrast, bioalcohols are sulfur free.

The direct combustion processes oxidize the available sulfur compounds to SO_x. This produces acid rain which, in certain locations, is believed to be the cause of damage to forests and lakes. The amount of acid rain produced by any combustion process is proportional to the amount of fuel-bound sulfur. Table 3 lists the fuel-bound sulfur for biomass, municipal waste, oil, and coal. For comparison, the chlorine contents of biomass and municipal waste are also listed since it produces HCl when burned. Both sulfur and chlorine emissions are being controlled by wet or dry scrubbers, particularly in the case of municipal waste combustion.

Nitrogen oxides are produced both from fuel nitrogen and from the oxidation of nitrogen in the air during combustion. The amount of NO_x produced by the fuel is proportional to the nitrogen in the fuel. Table 3 lists the fuel nitrogen for biomass, municipal solid waste, oil, and coal. Nitrogen oxidized from the air is a function of the temperature of combustion. It is lower for biomass and municipal solid waste which tend to burn at lower temperatures because of relatively high moisture contents.

4.5 Other Emissions

The principal products of combustion are carbon dioxide and water. Sulfur oxides, nitrogen oxides, and hydrogen chloride are discussed above. Carbon monoxide and other trace gases also are produced. These are controlled to meet current emission standards and are no higher than the similar emissions from coal and oil plants. In addition, particulates are discharged from the stack. Compared to the combustion of coal and No. 6 oil, particulates are controlled to the same or more stringent standards, particularly for municipal waste.

GLOSSARY

Accessible Resource: A subset of the total resource base that can be acquired by current technology or technology that will be available in the near future.

Anaerobic Digestion: A biological reaction that occurs in the absence of air.

Azeotrope: A mixture of two liquids such that the vapor at the boiling point has the same composition as the liquid, and thus the boiling point does not change as the evaporation proceeds.

Biogas: The gas produced by anaerobic digestion. Stoichiometrically it is 50% methane and 50% carbon dioxide.

Biomass: Any natural organic material or waste of recent origin excluding fossil fuels.

Carbohydrate: A compound made up of carbon, hydrogen, and oxygen with the hydrogen and oxygen in the atomic ratio of 2 to 1.

Cellulose: One of the four main components of plant matter. It is a long, crystalline, straight-chain carbohydrate primarily made up of glucose. The chemical formula of cellulose is complex but may be simplified to $(C_6(H_2O)_5)_n$.

Coppice: The regrowth of a tree from the stump.

Disaccharide: Two monosaccharides bound together with an OH^- radical removed from one and a H^+ radical from the other. The most common is sucrose which has one glucose and one fructose molecule combined. The chemical formula is $C_{12}H_{22}O_{11}$.

Enzyme: A natural biochemical catalyst. They are responsible for specific chemical reactions and are not consumed during the reactions.

Extractive: One of the four main components of plant matter. They are various chemicals and components such as sap. They carry nutrients throughout the tree, and provide disease and pest resistance. Of the extractives, sugar and starch are the main components of interest for biochemical conversion processes and resin is the main component of interest for thermal conversion processes.

Hemicellulose: One of the four main components of plant matter. It is a long, straight chain polysaccharide primarily composed of xylose (a five-carbon sugar). The chemical formula of hemicellulose is complex but may be simplified to $(C_5(H_2O)_4)_n$.

Hexose: A monosaccharide with six carbon atoms.

Hydrolysis: The process of splitting an atom by the addition of water with one hydrogen molecule incorporated into one fragment and the hydroxyl radical into the other.

Lignin: One of the four principal components of plant matter. It is an amorphous polymer primarily composed of phenol molecules.

Monosaccharide: A simple carbohydrate. The two most common are glucose and fructose, both of which have the formula $C_6H_{12}O_6$; however, the atoms are organized in different ways and they have different chemical characteristics. The corresponding pentose is xylose with the formula $C_5H_{10}O_5$.

Pentose: A monosaccharide with five carbon atoms.

Photosynthesis: The system by which plants use sunlight to convert CO_2 and water to plant matter and oxygen. Although the actual chemical composition of plant matter is very complex, frequently it is simplified as $C_x(H_2O)_y$. The values of x and y vary, although typical values are $x=1$, $y=0.72$ to 1. The general formula for photosynthesis is $nCO_2+xH_2O+\text{sunlight} \rightarrow C_n(H_2O)_x+nO_2$.

Polysaccharide: More than two linked monosaccharide molecules.

Quad: A measure of energy equal to 1 quadrillion Btu or 1.054 EJ.

Reserve: That subset of the accessible resource that can be economically and legally used to yield useful energy.

Resource Measurement Categories: Standardized definitions from the mineral industry to describe a resource. Further reading on the categories and quantities of renewable resources can be found from the Meridian Corporation (1989).

Short Rotation Hardwood: A fast-growing wood species that coppices and is grown on excess agricultural land. Short rotation wood was developed for energy but is being commercialized as a fiber and energy resource.

Starch: A folded chain of monosaccharides with the general formula of $(C_x(H_2O)_{(x-1)})_n$.

Sugar: A simple carbohydrate. There are three basic classes: monosaccharide, disaccharide, and polysaccharide.

Total Resource Base: The combination of undiscovered and identified, subeconomic and economic concentrations of a resource.

List of Works Cited

Domalski, E. S., Jobe, T. L., Milne, T. A. (1987), *Thermodynamic Data for Biomass Materials and Waste Components*, New York: American Society of Mechanical Engineers.

Energy Information Administration (1988), *Annual Energy Review*, Washington, DC: U.S. EIA, Report No. DOE/EIA-0384(88).

Energy Information Administration (1989), *Estimates of Biofuels Consumption in the United States During 1987*, Washington, DC: U.S. EIA , Report No. CNEAF/NAFD 89-03.

Energy Research and Development Administration (1977), *Waste-to-Energy Systems: An Overview*, Washington, DC: U.S. Department of Energy, Report No. CONS/2103-6. National Technical Information Service, Springfield, VA.

Green, A. E. S. (1989), *Overview, Summary and Afterword*, Vol. 7. New York: American Society of Mechanical Engineers.

National Wood Energy Association (1988), *State Biomass Statistical Summary*, Arlington, VA: National Wood Energy Assoc.

Solar Energy Research Institute (1981), *Ethanol Fuel: Use, Production and Economics,* Golden, CO: Solar Energy Research Institute.

United States Environmental Protection Agency (1990), *Facts: Ethanol as a Fuel,* Washington, DC: U.S. GPO.

Further Reading

Diaz, L. F., Savage, G. M., Golueke, C. G. (1982), *Resource Recovery from Municipal Solid Wastes,* Boca Raton, FL: CRC.

Hasselriis, F. (1984), *Refuse Derived Fuel Processing,* Boston, MA: Butterworth.

Meridian Corporation (1989), *Characterization of Energy Resources and Reserves,* Alexandria, VA: Meridian Corp.

Neenan, B., Feinberg, D., Hill, A., McIntosh, R., Terry, K. (1986), *Fuels from Microalgae: Technology Status, Potential, and Research Requirements,* Golden, CO: Solar Energy Research Institute.

Plumley, A. (1977), *Present Status and Research Needs in Energy Recovery from Wastes, Proceedings on the 1976 Conference,* New York: American Society of Mechanical Engineers.

Reed, T. R., Das, A. (1988), *Handbook of Downdraft Gasifier Engine Systems,* Golden, CO: Solar Energy Research Institute.

Robinson, W. D. (1986), *The Solid Waste Handbook, A Practical Guide,* New York: Wiley.

Solar Energy Research Institute (1989), *Biofuels Technical Information Guide,* Golden, CO: Solar Energy Research Institute.

National Research Council (1977), *Methane from Human, Animal and Cellulosic Wastes,* Washington, DC: National Academy of Sciences.

United States Department of Energy (1982), *A Review of the Comparative Energy Use in Materials Potentially Recoverable from Municipal Solid Waste,* Washington, DC: U.S. GPO, Report No. CE-20167-12.

Walter, D. K., Murphy, W. E., Richards, R. (1989), *Biofuels and Greenhouse Mitigation,* Vol. 7. New York: American Society of Mechanical Engineers.

Watson, J. D., Tooze, J., Kurtz, D. T. (1983), *Recombinant DNA, a Short Course,* New York: Freeman.

BIOMECHANICS OF LIVING ORGANISMS

RICHARD SKALAK, *Department of Applied Mechanics & Engineering Sciences, University of California, San Diego, California, U.S.A.*

	Introduction	489
1.	**Whole-Body Dynamics**	491
1.1	Applications of Whole-Body Dynamics	491
1.1.1	Crashworthiness	491
1.1.2	Aerospace Biomechanics	491
1.1.3	Gait Analysis	491
2.	**Mechanical Properties of Biological Tissues**	492
2.1	Mechanical Properties of Bone	492
2.2	Mechanical Properties of Soft Tissues	493
3.	**Biomechanics of Joints**	495
3.1	Anatomy of Diarthroidal Joints	495
3.2	Mechanical Properties of Cartilage	496
3.3	Lubrication Mechanisms in Joints	497
4.	**Muscle Mechanics and Physics of Body Movements**	499
4.1	Skeletal Muscle	499
4.2	Heart Muscle	500
4.3	Smooth Muscle	500
4.4	Coupling of Muscles and Skeleton	501
5.	**Mechanics of the Lung**	501
5.1	Structure of the Lung	501
5.2	Zones of the Lung	502
5.3	Elasticity and Surface Tension in the Lung	502
6.	**Cochlear and Vestibular Mechanics**	503
6.1	Structure of the Inner Ear	503
6.2	Frequency Discrimination	503
6.3	Vestibular System Mechanics	504
7.	**Swimming and Flying**	504
7.1	Swimming of Microorganisms	504
7.2	Modes of Swimming of Fish	505
7.3	Flight of Birds	506
7.4	Modes of Insect Flight	507
8.	**Scaling and Limitations in Biomechanics**	507
8.1	Size and Power Limitations	507
8.2	Scaling in Terrestrial Locomotion	508
8.3	Scaling in Aquatic Locomotion	508
8.4	Scaling in Flight	509
8.5	Scaling in the Cardiovascular and Respiratory Systems	509
	Glossary	510
	Works Cited	510
	Further Reading	511

INTRODUCTION

Biomechanics is defined as the application of mechanics to problems in biology and physiology. The mechanics implied in this definition are the classical fields of rigid-body dynamics, solid mechanics, and fluid mechanics. In the applications of each of these fields to biological problems, nonlinear as well as linear theories are often required. In representing body motions not only rigid-body dynamics may be required but also dynamics of jointed, hinged, and flexible structures. In the applications of solid mechanics to soft tissues, the strains are often large and the constitutive equations are nonlinear. Biological fluids, particularly blood, are non-Newtonian and thixotropic. Both biological fluids and solids are generally viscoelastic. Thus, while linear theories of elasticity and Newtonian fluids are often useful as a first approximation, nonlinear field theories are often required.

If the definition of mechanics is expanded to include statistical mechanics, thermodynamics, heat, and mass transfer, then the area of application of biomechanics can be correspondingly expanded. However, for the purposes of this article the more classical definition will be assumed and only a limited

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

number of areas of application will be mentioned, mostly in relation to the human body. In biological tissues, temperature is often an important variable. Further, many mechanical properties are modulated by biochemical changes. These important aspects, however, will not be generally covered in this article since there are special articles on various biophysical topics in this Encyclopedia. They include CELLULAR BIOMECHANICS, BIOMEDICAL ENGINEERING, BIORHEOLOGY, MEMBRANE BIOPHYSICS, PROSTHETICS AND ARTIFICIAL ORGANS, AEROSPACE MEDICAL PHYSICS, and BIOLOGICAL EFFECTS OF SOUND AND ULTRASOUND. The reader is referred to these articles.

The term *biomechanics* and research specifically directed to this field have become widespread only in the last two or three decades. Looking back in history one can identify investigations and ideas that can be classified as biomechanics although they may not have been called by this name at the time. Such early efforts tended to be sporadic and due to a few outstanding individuals who were often physicists as well as physicians. One can find instances of application of ideas of mechanics to biology which paralleled or even preceded the development of new principles in mechanics. For example, it could be said that Leonardo da Vinci (1452–1519) as an engineer and artist appreciated the mechanical aspects of the muscle and skeletal systems of the human body and to some extent the mechanics of the heart, heart valves, and blood flow. Galileo (1564–1642) was a student of medicine before he became a physicist. He discovered and used the pendulum to measure the pulse rate. The discovery of circulation of blood is attributed to William Harvey(1578–1658). Harvey studied at Padua while Galileo was a professor there. During this period Santorio (1561–1636), also a professor of medicine at Padua, made careful measurements of human body weight and showed it decreased even in periods of no obvious intake or efflux. René Descartes (1596–1650) wrote a theory of physiology based on mechanical principles. However, it was not very realistic. A classical treatment of muscular and skeletal dynamics was given by Giovanni Borelli (1608–1679). He also considered the mechanics of swimming, flying, and the contractions of the heart. Robert Hooke (1635–1703), after whom Hooke's law is named, also worked on biomechanics. Leonhard Euler (1707–1783) wrote one paper on wave propagation in the arteries among his many works on fluid mechanics. Thomas Young, after whom Young's modulus is named, also discussed the theory of blood flow. In fact, the concept of Young's modulus came from the study of elasticity of artery walls. Young was a physician concerned with astigmatism and color vision as well. The French physician Jean-Louis-Marie Poiseuille (1779–1869) established Poiseuille's law of viscous flow in a circular tube out of his interests in blood flow. In this case the experimental law (1841) preceded the solution of the Navier–Stokes equations. The theoretical formula was derived some years later by Hagenbach (1859).

Another physician who contributed to many aspects of the mechanics of physiology was Hermann Von Helmholtz (1821–1894). In biomechanics he is best known for his studies of hearing. Helmholtz represents a kind of scientist who was equally at home in physiology and in the latest physics of the times. Such comprehensive individuals seem to be more rare after 1900 as scientific knowledge rapidly expanded on many fronts. In the period from 1900 to about 1960 applications of mechanics to the development of such engineering products as aircraft and automobiles greatly advanced in theory and practice. But the development of applications of mechanics within medicine and physiology did not share equally in this development.

Beginning about 1960 there developed a widespread interest in applications of mechanics to biology. At the present time there is again a close integration of these diverse fields most often by a cooperation of individuals trained in different disciplines, e.g., biology and physics, or medicine and engineering. The symposium on biomechanics held in 1970 (Fung *et al.*, 1972) indicates the status of the field at that time. Since 1973 the American Society of Mechanical Engineers has held annual symposia on biomechanics. There is also a Journal of Biomechanics at this time and many additional meetings and societies concerned with applications in bioengineering which include biomechanics discussions. One area in which applications have become highly developed is orthopedics. A new Journal of Orthopedic Research has been recently instituted. A series of books by Fung (1981, 1984, 1990) and the *Handbook of Bioengineer-*

ing (Skalak and Chien, 1987) indicate the scope of current work in biomechanics.

1. WHOLE-BODY DYNAMICS

In considering the motion of an entire human being, an adequate model can be obtained for many purposes by considering each major segment as a rigid body connected by appropriate joints and hinges. Each segment must obey the laws of motion of a rigid body which may be expressed in any convenient coordinate system. At each of the hinges or joints of the body the reactions on the two segments connected must be equal and opposite and the transmitted moments along hinge axes are set equal to zero or they may depend on a spring and/or dashpot behavior (see Fig. 1).

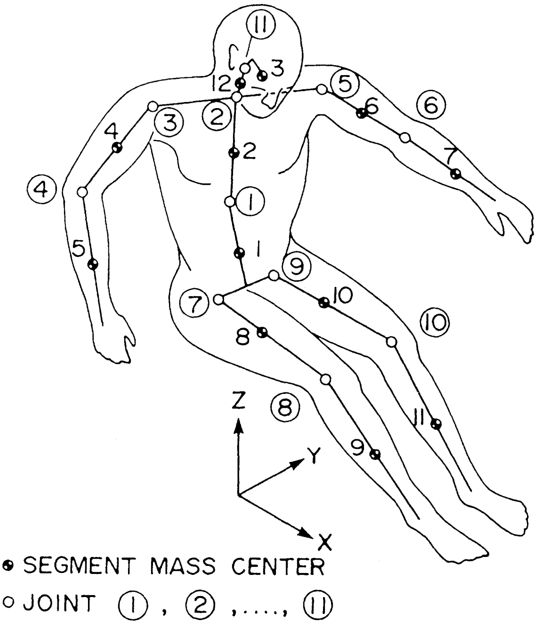

FIG. 1. Three-dimensional model of aircraft occupant consisting of 12 rigid segments. The model has 29 degrees of freedom. Resistances at each of the body joints are modeled using both a nonlinear torsional spring and a viscous damping element. (From Laananen, 1986, by permission.)

1.1 Applications of Whole-Body Dynamics

Models of the human body consisting of linkages of rigid subunits are useful in studying crashworthiness, gait analysis, and stresses developed during automatic ejection from aircraft.

1.1.1 Crashworthiness. In the case of crashworthiness, the motions of the various parts of the body are well represented and computer models predict such things as head contact with a front windshield or sliding under a seat belt during severe deceleration. The analysis of the injury or trauma to be expected from such crash forces requires a more detailed model and analysis including an appropriate representation of the skeletal system and the soft tissues (see Nahum and Melvin, 1985). In the problem of head injury, for example, the skull may be represented as an elastic shell. The brain and other soft tissues are much less rigid than the skull and may be represented as a viscoelastic solid of low modulus as a first approximation. Detailed studies indicate that injury to the brain is most easily caused by sudden angular accelerations which induce shear stresses that in turn may rupture blood vessels. The combination of such studies with neurophysiological function is a current area of research (Viano, 1988).

1.1.2 Aerospace Biomechanics. The behavior and performance of the human body in aerospace environments offers a variety of interesting problems. The rigid-body linkage model is useful in estimating forces produced by emergency ejection for aircraft. Models for this purpose have been highly developed and analyzed in detail. In space the absence of gravity limits the motions which a freely floating person can perform. The linkage model is also useful in this case.

1.1.3 Gait Analysis. Whole-body models are also useful in studying gait analysis. The details of the motion of each part of the body during walking and running have been analyzed from cinematographic records and by measurement of forces by devices flush with the walking surface. This equipment is similar to the devices used in measuring forces on model aircraft in wind tunnels. Gait analyses are useful in diagnosing pathologies of the joints and skeletal system. They are also useful in evaluating the functioning and efficiency of prostheses.

2. MECHANICAL PROPERTIES OF BIOLOGICAL TISSUES

2.1 Mechanical Properties of Bone

Bone is a composite material, made up of a matrix and a mineral crystal phase, principally hydroxyapatite, which is hard, dense, strong, and brittle. These fibers are surrounded by the softer collagenous matrix. Mammalian bones have a specific gravity of about 2.0 and have roughly equal parts of one-third mineral, one-third collagenous matrix, and one-third water. These proportions vary with species, age, and location. A typical long bone is a hollow structure with a dense bone shell made up of compact or cortical bone. The cavity is usually partly filled with a less-dense form of bone called cancellous or trabecular bone. Cancellous bone is made up of a meshwork of small plates with many spaces between them like a sponge. The individual plates are of similar strength and composition to cortical bone.

In vivo bone is a living tissue made up of highly organized cells and a blood supply which permeates the entire structure. Bone may remodel itself under stress by adding material at its outer surface or by changes in its compositon and density throughout its volume. These functions are carried out by special cells called osteoblasts which build up new bone and osteoclasts which cause its degradation. The cells and fibers in bone are mainly oriented along principal stress trajectories which makes the mechanical properties of bone anisotropic. Among the tissues of the body, bone is relatively solid, rigid, elastic, and brittle. In tension, bone fractures at an extension of about 2%. The ultimate strength of bone depends on the direction of stress with respect to the axis of the bone. Bone is stronger in the direction of its longitudinal axis. It is generally taken to be transversely isotropic, meaning that properties are the same for all directions perpendicular to the long axis. The ultimate strength of cortical bone has been determined for samples from various animals tested in different directions. Mean values for the ultimate strength of human bone are given by Reilly and Burstein (1975). In the direction of the long axis the ultimate strengths of human bone in tension and compression are 133 and 193 MPa, respectively. Perpendicular to the long axis the ultimate strengths in tension and compression are found to be 51 and 133 MPa, respectively. Shear strengths of bone are reported as derived from torsion tests on circular specimens whose long axis coincides with that of the bone. For such tests the mean shear ultimate strength is 68 MPa. A more comprehensive discussion of the elasticity and strengths of bone may be found in the chapter by Cowin *et al.* (1987).

Testing of bone has been carried out in standard procedures of tension, compression, bending, and shear testing statically and at various strain rates. Data by Carter and Hayes (1977) show that both elastic modulus and ultimate strength increase with the strain rate but only to a power of about 0.06. Viscoelastic and fatigue properties of bone are also discussed by Cowin *et al.* (1987). Fatigue data indicate that *in vitro*, bone accumulates fatigue damage so that the stress amplitude for failure decreases sharply at the order of 10^6 cycles. However, *in vivo* it is expected that bone repairs and remodels continuously so that the *in vitro* fatigue testing does not fully reflect the performance of bone *in vivo*.

In the range of most physiological stresses, bone acts as an anisotropic (transversely isotropic) elastic material. Cowin *et al.* (1987) showed that linear stress-strain relations for the elastic range require 12 nonzero material constants of which 9 are independent. In terms of the usual Young's modulus in tension or compression, shear modulus, and Poisson's ratio, the 12 elastic parameters can be summarized as 3 Young's moduli in the principal directions, 3 shear moduli, and 6 Poisson's ratios. The Young's modulus for human bone in the longitudinal direction is about 17 000 MPa. In the transverse directions, the Young's modulus is about 11 500 MPa. The shear moduli are of the order of 3300 MPa. The Poisson's ratio varies widely in different experiments but generally lies within the range of 0.3–0.5. The properties of bone vary widely with the particular bones, age, sex, species, and conditions of stress. Astronauts in space and patients who are for some reason immobilized lose mineral content of the bone over a period of days or weeks (see AEROSPACE MEDICAL PHYSICS). The influence of stress on bone remodeling has been discussed by Cowin (1988) and the modulation of formation of bone during growth by the stress field has been discussed by Carter (1988).

Additional discussion of the structural function of bone has been given by Fung (1984). In a macroscopic sense the form of bones can often be identified as tending toward an optimal structural arrangement. The trabeculi in cancellous bone tend to be oriented along lines of principal stress. Further, the hollow form of bone is a very efficient structural shape for stiffness and bending rigidity. Computer programs utilizing remodeling laws based on strain energy density show realistic developments comparable to observed forms (Huiskes, 1988).

In the testing of bone, it is important that it be kept wet at all times. Otherwise the properties begin to change rapidly. The properties of bone may be determined by standard tension and compression tests but also by ultrasonic testing (Cowin *et al.*, 1987).

Bone also exhibits interesting electromechanical effects. It is piezoelectric and it has been suggested that there may be an electrochemical effect as an important part of bone remodeling under stress. Bending of a bone may produce an internal fluid flow which in turn may produce a streaming potential. These electromechanical effects are not discussed here in detail but more information may be found in a chapter by Johnson and Katz (1987).

2.2 Mechanical Properties of Soft Tissues

Most soft tissues of the body have characteristically a highly nonlinear stress-strain curve. At low stresses the tension modulus is usually quite small. After an extension of 20%–50%, the stress-strain curve turns upward and the tension modulus is much larger just prior to failure. Figure 2 shows some test results for skin which exhibit this typical nonlinear behavior. Many tissues of the body are anisotropic. For example, the arterial walls are stiffer in a circumferential direction as compared to the longitudinal elasticity.

For most applications, it is sufficiently accurate to assume that soft tissues are incompressible. The bulk or volumetric elastic modulus of most soft tissues is several orders of magnitude greater than the shear or Young's modulus. An exception to the assumption of incompressibility is in the lung which undergoes large volume changes during breathing. However, these changes occur in the air space, and the soft tissues which make up the walls of the alveoli are themselves relatively incompressible. Another exception is in the analysis of ultrasound or shock waves. In these phenomena the compressibility of the

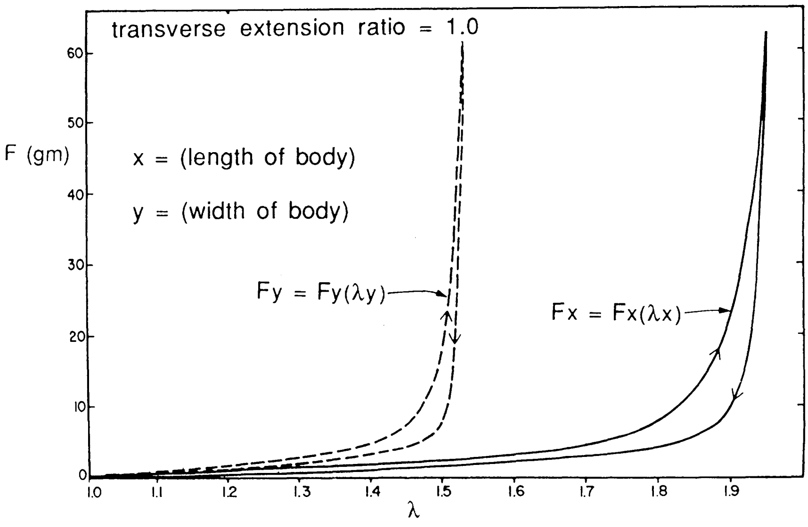

FIG. 2. Load-strain relations of a square (13.5×13.5 cm) specimen of rabbit abdominal skin. The tissue is stretched in one direction while the transverse dimension is kept constant. Solid line: stretch along the body length. Dashed line: stretch along the body width. (Adapted from Lanir and Fung, 1974, by permission.)

tissue, although high, is essential to understanding the physics of the stress propagation.

Most soft tissues also exhibit some degree of history dependence and viscoelasticity, but these are relatively secondary effects. If a tissue is removed from the body and tested in a cyclical manner it is generally found that the first few cycles do not repeat each other exactly, but there is a gradual increase of strain. After several cycles the stress strain curve becomes repeatable but not exactly the same on loading or unloading. The unloading curve generally shows a smaller stress at a given strain resulting in a closed loop in the stress-strain diagram as shown in Fig. 2. The area within this loop represents energy lost by hysteresis. Fung (1984) suggests that this behavior be called pseudoelasticity. The loading limb of the curve may be described as nonlinear elastic and the unloading curve is also nonlinear elastic but with a somewhat different stress-strain relation. Each of the loading and unloading curves may be described in terms of a strain energy function W per unit initial volume. A convenient set of stresses and strains for this description are the Green's strain tensor E_{ij}, and the corresponding Kirchhoff stress components (see Fung, 1984). The strain energy function W for an elastic material is assumed to be a function only of the strain components E_{ij}. Then, the Kirchhoff stresses S_{ij} are given by the derivatives of W with respect to E_{ij}, as is usual in finite deformation theory of elasticity.

If the tissue is considered as incompressible, then isotropic pressure p must be added to the stress components. This pressure p is not determined by the strains but can be found from the equations of motion and boundary conditions in solving a particular problem.

The form of strain energy function which gives the characteristic steepening nonlinearity of soft tissues is usually an exponential function of the strain components as suggested by Fung (1984). In general, it will not be sufficient to do only uniaxial tension or compression tests to define the strain energy function, but tests with more than one stress component different from zero are required.

For tissues which are naturally a thin sheet such as skin or arterial wall, it is convenient to test and describe the stress-strain relations in terms of tensions per unit length of a cut edge (dyn/cm). Such tensions may be regarded as the integral of the Cauchy stress through the thickness of the material. In testing of a plane sheet of tissue such as skin, the anisotropy is revealed by applying loads in different directions, as illustrated in Fig. 2. Lanir (1987) lists about 25 different stress-strain relations which have been used to describe tests on skin.

Soft tissues also show some viscoelastic behavior in the sense that the stress-strain curves do depend to some extent on strain rates. However, this is a very weak dependence and as a result hysteresis curves such as shown in Fig. 2 are relatively independent of frequency over a large range of frequencies of the order of 0.1–100 Hz.

The viscoelastic behavior of soft tissues may be described in terms of relaxation functions as suggested by Fung (1984). Typical forms for the relaxation functions are found by assuming simple models such as a Kelvin, Voigt, or standard solid model. For large deformations, describing nonlinear stress-strain relations, the variables in the relaxation functions are taken to be the Kirchhoff stress tensor and Green's strain tensor. Complete tensor descriptions have not been fully developed and validated for any particular tissues. However, one- and two-dimensional experiments using reduced relaxation functions fit the experimental data well. The relaxation functions of biological tissues are relatively independent of frequency over a wide range of frequencies which may be interpreted physically as indicating that the appropriate model is a large number of Kelvin bodies in series, each element having a slightly different time constant (Fung, 1984).

The typical steepening of the soft-tissue stress-strain curve at high strains is physically due to the presence of collagen fibers which are in a coiled form in most soft tissues. When these fibers are stretched so they uncoil and orient in the direction of principal tension, they become very stiff and account for the stiffening indicated in Fig. 2 and most other tissues. One tissue in which collagen is arranged in a parallel straight fiber arrangement is tendon, which consists primarily of collagen only. This produces a much stiffer and stronger material than soft tissues in which less collagen is present and is woven in a loose and random manner. The ultimate strength of tendon is about 60 MPa, while that of skin is about 15 MPa. Further, the strain at failure is at about 10% elongation of tendon and 110% for skin. The stress-strain for tendon is much closer to that of an elastic brittle material than the highly nonlinear form for skin (Fig. 2).

All of the above discussion and most reports in the literature of the properties of soft tissues are, of course, on excised specimens which no longer have their blood supply, innervation, or metabolism. It is clear that since a normal person does not change dimensions or elastic properties radically day to day, the normal body function restores properties to a hemostatic condition after most stresses below a point of tissue damage or trauma. This regenerative capacity is not represented in physical testing and it should be kept in mind in interpreting tests that there may be differences between excised specimens and the behavior tissue *in vivo*.

Another interesting aspect, closely related to the above, is the recently explored question of residual stress in soft tissues *in vivo* (Fung, 1984). For purposes of this discussion the residual stresses are defined as those remaining in a tissue or organ after removal from the body and placed in a state of no external loading. For example, if a section of artery or the entire heart is excised from an animal and placed in an open container of saline then essentially there is no pressure difference between the inside and outside surfaces so these tissues are said to be unloaded. However, there still may be some residual stresses present. This is most dramatically revealed in the case of the artery by cutting a small ring off the artery which remains circular as long as the ring is intact. But if the ring is cut through at one point, it opens out. The angle drawn between the cut ends and a point midway between them on the far side of the artery is called the opening angle. It may be anywhere from a negative angle (closing up) to over 180° (bending backwards) so that the curvature is actually reversed. The residual stresses implied by this experiment are probably functioning *in vivo* to produce a much more uniform stress distribution in the arterial wall than would occur if the wall were purely elastic with zero residual stress when unloaded (Fung, 1984).

It should also be appreciated that many tissues are normally under some stress *in vivo*. This may be partially residual stress but may also be the normal state with or without residual stress. For example, if a section of artery is excised, it shrinks in length by 10%–20%. This indicates that it is stretched and under stress longitudinally *in vivo* but this is not a residual stress. (The same is true of nerves.) In the circumferential direction, arteries are normally under tension due to the pressure which rarely if ever drops to zero. The circumferential stress due to pressure is in addition to the residual stress.

The human skin is under some tension *in vivo*. This, however, is not due to residual stress but is a result of the growth and metabolism of the skin itself. Scar tissue tends to shrink abnormally and produce locally high tensions. This is more of the nature of a local residual stress.

Lung tissue normally operates in a state of isotropic tension but this is in the nature of *in vivo* stress rather than a residual stress. The heart wall undergoes large fluctuations of stress due to the active contraction of the tissue. However, when excised and unloaded, a ring cut transversely from the ventricle shows residual stresses when the ring is cut open similar to that of the arterial wall. In any stress analysis, the *in vivo* and residual stress levels should be taken into account to have a realistic modeling of the stress distribution *in vivo*.

3. BIOMECHANICS OF JOINTS

3.1. Anatomy of Diarthroidal Joints

The skeletal joints which allow relative motion of the bones they connect are called diarthroidal or synovial joints. This includes the hip, knee, elbow, shoulder, ankle, wrist, and finger joints. In such joints the bones themselves are not in contact, but are covered at their contact ends by a thin layer of a special tissue, namely, articular cartilage. This is a fibrous, porous tissue which is usually between 0.5 and 2 mm in thickness. The cartilage is bathed in a special fluid called synovial fluid, and the joint is enclosed in a fibrous capsule which retains this fluid. The combination of cartilage and synovial fluid results in an extremely low coefficient of friction in joints, of the order of 0.05 to 0.003. This is much lower than is typically achieved in machine designs.

A schematic of the human knee joint anatomy is shown in Fig. 3. The contact surfaces are generally curved so that the contact area is much less than the cartilage surface area at a given position, and it changes as the joint moves. Tendons and ligaments connecting the two bones at a joint aid in its stabilization and give it a tensile strength. The capacity of the cartilage and synovial fluid to carry any tension is negligible. The meniscus of the knee is a softer tissue than cartilage and plays a role in stabilization of the knee movement.

Cartilage consists of about 25% solid material and 75% fluid. The principal solid phase is collagen fibers (65%) and proteoglycans (25%). The pores of this matrix are extremely fine, of the order of the 2–6.5 nm. For this reason the fluid in cartilage is not easily squeezed out. The cartilage is a porous medium but with quite low permeability. There are some living cells distributed in cartilage called chondrocytes. These cells can repair and reproduce cartilage, but do so at a very slow rate, and wear or damage to cartilage is often, in effect, irreparable.

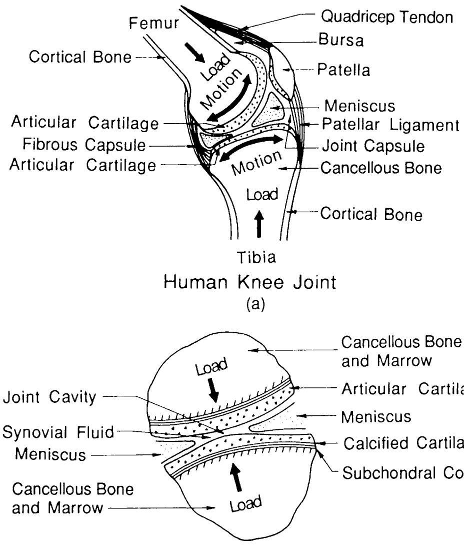

Human Knee Joint
(a)

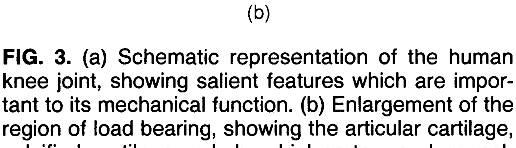

Load Bearing Region
(b)

FIG. 3. (a) Schematic representation of the human knee joint, showing salient features which are important to its mechanical function. (b) Enlargement of the region of load bearing, showing the articular cartilage, calcified cartilage, subchondrial cortex, and cancellous bone. The calcified cartilage and subchondrial cortex function as a shell, providing backing to the articular cartilage, and are supported by cancellous bone. The cancellous bone–bone marrow system provides the shock-absorbing mechanism for the diarthroidal joint. (Adapted from Mow and Mak, 1987, by permission.)

Synovial fluid is a special clear fluid which is generated by a special tissue called the synovium. Synovial fluid is derived in part from blood plasma but with an important addition of hyaluronic acid protein added. This is a very large macromolecule which results in a non-Newtonian behavior of the synovial fluid. It is viscoelastic and exhibits shear thinning and a normal stress effect. Synovial fluid is important to the production of a low coefficient of friction in joints. If it is replaced by saline, the coefficient of friction may rise by several orders of magnitude. Further details and references can be found in the chapter by Mow and Mak (1987).

3.2 Mechanical Properties of Cartilage

The importance of cartilage to the performance of joints has led to extensive experimental and theoretical work concerning its properties and behavior. It is possible to cut disks or elongated tension specimens of the cartilage from the surfaces of large joints. Such specimens can be tested as if they consisted of a single, solid material. However, to have reproducible results, the tests need to be conducted while the sample is submerged in fluid. It is found that cartilage behaves macroscopically as a viscoelastic solid. In a simple tension test there is just one stress component, in the direction of tension, and one corresponding relaxation function. In this case, the Green's strain for the direction of extension is

$$E_{11}=\tfrac{1}{2}\,(\lambda_1^2-1), \tag{1}$$

where λ_1 is the stretch ratio, i.e., the final length divided by the initial length of the specimen. The Kirchhoff stress is

$$S_{11}=F/A\lambda_1, \tag{2}$$

where F is the tensile force, A is the initial cross-sectional area, and λ_1 is the extension ratio. By applying a sudden stretch and measuring the relaxation of force with time, the relaxation function for a cartilage specimen can be measured directly, as shown in Fig. 4.

In order to evaluate the possible exudation of fluid from cartilage under stress and the effect of fluid movement on the properties of

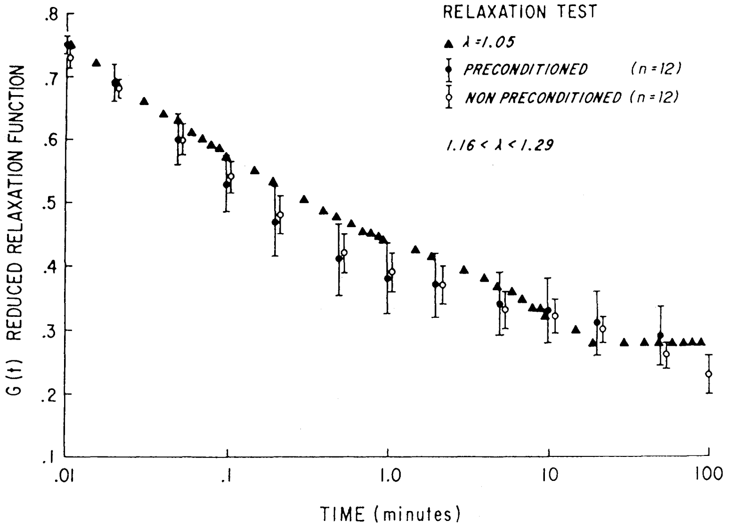

FIG. 4. The reduced relaxation function of articular cartilage at a lower stretch ($\lambda = 1.05$) and at higher stretches, $\lambda = 1.16$–1.29. Preconditioning refers to performing a number of cycles of stress and strain before a test, until the initial response is stabilized; *n* is the number of specimens tested. (From Woo *et al.*, 1979, by permission.)

cartilage it is necessary to establish experimentally the permeability of the cartilage as a porous matrix. The apparent permeability k_a of porous medium is defined by Darcy's law,

$$V = k_a \nabla p, \qquad (3)$$

where the V is the velocity of fluid flow computed as the discharge divided by the gross area, k_a is the apparent permeability, and ∇p is the spatial pressure gradient. An average value of permeability of cartilage is 0.58×10^{-15} m^4/N s (McCutchen, 1962). When sufficiently high pressure gradients are applied to cartilage, the pressure causing the flow may also cause appreciable compression of the solid structure of the cartilage, which then, in turn, modifies the permeability. Cartilage permeability may also be decreased by compression not associated with pressure, for example, by squeezing a specimen between porous plates. The two effects have been expressed in an empirical formula by Mow and Lai (1980),

$$k_a = A \exp(-\alpha \epsilon), \qquad (4)$$

where A is a parameter dependent on the pressure drop across the specimen, α is a second constant also dependent on pressure drop, and ϵ is the strain which is imposed on the cartilage in advance of testing the fluid flow through it; ϵ is called the clamping strain. The permeability decreases with increasing clamping strain and with increasing pressure drop.

For a complete treatment of the mechanics of cartilage a biphasic model is assumed (Mow and Lai, 1980). In this theory both the fluid and the solid matrix of the cartilage are assumed to be individually incompressible. The relative motion of the fluid and solid gives rise to a drag force on the fluid, which in turn is applied also to the solid with reverse sign. The constituent equation for the solid is assumed to be that of a viscoelastic material, and the fluid is taken to be a viscous incompressible fluid, but on a macroscopic basis, a stress associated with the change of volume per unit of cartilage volume is incorporated. The possibility of volume change arises from the possible relative motion of the solid and fluid. The biphasic theory has been applied to compression and relaxation tests and shown to reproduce the deformation and stress data quite accurately. Thus, the properties of cartilage are fairly well defined, but the mechanics of joint lubrication are not entirely clear because the roles played by various possible lubrication mechanisms are not definitively known.

3.3 Lubrication Mechanisms in Joints

Several different mechanisms of lubrication of joints have been proposed theoretically and demonstrated to some extent experimentally. It appears that no one mechanism is operative in all cases and that a variety of lubricating effects may take place under different conditions of movement and loading.

The mechanism of lubrication of the knee or hip joints may be quite different after a period of standing still as opposed to a repetitive sequence of steady running or walking. Various mechanisms of joint lubrication that have been proposed in the literature are outlined below and some of the experimental results are indicated.

A logical suggestion for the mechanism of joint lubrication is the kind of viscous hydrodynamic action which is achieved by the use of lubricating oil in machine bearings of either rotating or sliding type. This mechanism is known as hydrodynamic lubrication. This kind of lubrication requires the presence of a continuous film of fluid between the parts of the joint, which is tapered to some extent. It also requires a continuous relative sliding motion and a continuous supply of fluid. It is estimated that during level walking the fluid film thickness which would be required in the knee would be of the order of 10^{-3} mm. This is less than the surface roughness of the cartilage surface so hydrodynamic lubrication seems unlikely under these conditions. For faster motions it remains a possibility.

A second kind of mechanism is the direct sliding of one cartilage surface on the other. This is known as boundary lubrication. Among artificial surfaces, Teflon on Teflon is an example of fairly low boundary lubrication capacity. The frictional coefficient for Teflon-coated surfaces is of the order of 0.05–0.1. For cartilage it is suggested that there is a lubricating glycoprotein fraction (LGF) which provides a low boundary friction. This component is not the hyaluronic acid, but a single polypeptide chain of about 250 000 molecular weight. This molecule adheres strongly to the cartilage surface. It appears that *in vivo* after a long period of standing still some such boundary lubricating mechanism would be required.

Basically, the two mechanisms available for joint lubrication are boundary lubrication and hydrodynamic lubrication. However, the fact that cartilage is porous, elastic, and fluid filled leaves open some possibilities of exchange of fluid between the lubricating gap and the cartilage and also for a role for the deformability of the cartilage itself. Even in the case of materials which are not porous, such as rubber, there can be an effect of the deformability of the surface on lubrication phenomena. The pressures generated by hydrodynamic lubrication can deform a soft boundary such as a rubber sheet, which in turn influences the distribution of pressures and eventually the friction coefficient. This mechanism of deformation of the cartilage may also play a role in joint lubrication. A related idea is called weeping lubrication in which the normal load applied to a joint is assumed to compress the cartilage and thereby cause exudation of a fluid film which then can act as a lubricant. This is similar to a bearing which is pressure fed oil continuously. Another possible mechanism is called squeeze film lubrication. In its pure form, squeeze film lubrication occurs when two rigid surfaces with a viscous fluid between them are pressed together. A transient fluid motion ensues in which the viscous liquid is forced out of the space between the approaching rigid surfaces. This viscous flow can generate large pressures and reduce the normal velocity of the plates to very small values. A tangential force applied at the same time will then experience a very low friction coefficient. Nonlinear and viscoelastic fluids will have enhanced squeeze film lubrication. This mechanism may be operative when a person lands after jumping off some small height. During the sudden load, squeeze film action is likely.

Another suggestion is called boosted lubrication. In this mechanism it is suggested that as two cartilage surfaces approach each other, water is squeezed out from the contact region into the cartilage. This would leave a gel layer, which could support greater loads than the normal synovial fluid. A more complete discussion of these suggestions and references may be found in the chapter by Mow and Mak (1987).

In summary, there is no single mechanism of joint lubrication which is known to be operative in all conditions *in vivo*. Both boundary and hydrodynamic lubrication appear to be involved in different situations. The involvement of the flow through the porous media of the cartilage and its deformation may play roles that are not yet fully or accurately identified. A possibility of both exudation and absorption of lubricating fluid has been discussed by Lai and Mow (1979). A theoretical analysis of a porous indent sliding along a layer of cartilage shows that the approaching compression may force fluid out of the cartilage, and on passing a given area a fluid may be reabsorbed by the cartilage. This kind of action indicates some of the weeping-type action and some reabsorption. This has been termed a self-generating mechanism of joint lubrication.

4. MUSCLE MECHANICS AND PHYSICS OF BODY MOVEMENTS

Throughout the animal kingdom, almost all active force generation is provided by muscles whose basic capacity is to generate tensile force or contraction, rather than thrust or compression. Body movements are usually the result of muscle contraction coupled to skeletal elements or softer tissues via tendons. Many different geometric arrangements of muscle, tendons, and bones give the large variety of resultant motions that animals can perform. The entire system is under neural control, both for stimulation of muscle and for the return of sensory signals indicating position. Visual, balance, and touch information supplied to the brain serves to generate exquisite control in time and space.

Mammalian muscles are grouped into three types: skeletal muscle, heart muscle, and smooth muscle. Skeletal muscles are those that produce voluntary body motions. Heart muscle has some unique characteristics and is controlled by its own pacemaker. Smooth muscles are generally not under conscious control. They occur in the intestines, arteries, ureter, and many other organs. They are called smooth muscles because under the microscope, they appear to be homogenous, as opposed to skeletal muscles which show characteristic striations at the microscopic level (see Fig. 5).

4.1 Skeletal Muscle

The basic force-generating elements of all mammalian muscles are myosin and actin fibers whose interaction produces a sliding motion of these interdigitating filaments as shown in Fig. 5. Groups of actin-myosin fibers are organized into myofibrils with a very regular spacing along the length called the sarcomere length and a regular hexagonal pattern in cross section. Groups of muscle fibers form larger units called fascicles which have distinct connective tissue sheaths. A number of fascicles form a muscle (Fig. 5) which has a thicker sheath and is connected by tendons at each end to bones or other tissues.

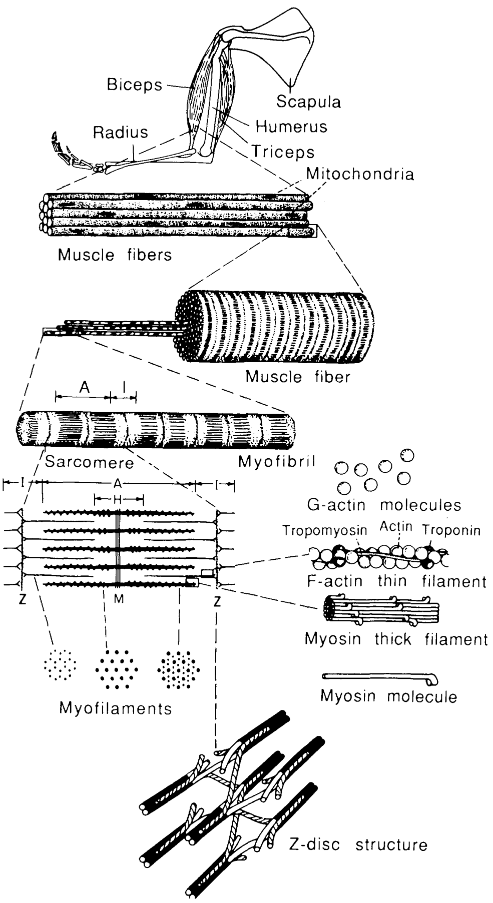

FIG. 5. Structure of skeletal muscle and arrangement of agonist (biceps) and antagonist (triceps) muscles of the arm. (Adapted from McMahon, 1984, by permission.)

There is an extensive literature on the anatomy, physiology, and biophysics of muscles (e.g., *Handbook of Physiology,* 1979, Sec. 10). The details of the interaction of myosin heads which rock back and forth to move along the actin filaments is still an area of active research. This is the level at which the consumption of biochemical energy in the form of ATP (adenosine triphosphate) takes place and is transduced to mechanical work and motion. The resulting macroscopic behavior has been described by various phenomenological theories, of which the most famous and most widely used is Hill's equation (1938). Improvements and more detailed models have made Hill's equation obsolete to a certain extent, but it captures some basic aspects of muscle behavior.

Hill's equation may be written as

$$(v+b)(T+a)=b(T_0+a), \tag{5}$$

where T is the tension force in the muscle, v is the velocity of contraction, and a and b are constants. This equation relates the velocity v to the tension T when the tension is suddenly reduced from an isometric, tetanized state, characterized by an initial length L_0 and tension T_0. The tetanized state means the frequency of electrical pulse stimulation is greater than a critical value (about 100 Hz), which produces the maximum tension at the fixed length L_0. The main effect contained in Hill's equation is that the velocity of contraction increases as the tension T decreases. The main physiological result is that the useful power generated tends to be approximately constant in the middle range of velocities. For more comprehensive theories, see the *Handbook of Physiology* (1979), Sec. 10; Fung (1981); and McMahon (1984).

4.2 Heart Muscle

Heart muscle has sarcomeres made up of actin and myosin filaments interdigitating as in skeletal muscle. The main differences between heart and skeletal muscle are in the organization of the fibers and in the manner in which they are excited and supplied with nutrients. The muscle fibers of the heart have more branching of their structures and capillaries and mitochondria are more abundant. The capillaries deliver oxygen and nutrients to the tissue which must be supplied continuously. The mitochondria are intracellular units which manufacture ATP, which is the molecule used as fuel in the actin-myosin contraction.

Another difference between heart and skeletal muscle is in the excitation control. Heart muscle is excited by a wave of electrical polarization which causes all fibers of the heart to contract nearly synchronously. In skeletal muscle, the response may be graded by having a limited number of fibers activated depending on the frequency and degree of stimulation.

The mechanical properties of a single heart muscle fiber in the activated state (during systole) is similar to skeletal muscle in that Hill's equation (10) is a good first approximation to its behavior. In the resting state, a significant difference is found. In the heart, for the resting state (during diastole) the remaining stress is appreciable and is important to appropriate refilling of the heart with blood. In skeletal muscle, the tension in the relaxed state is very small and the degree to which it may be extended by small tensions is larger than in the heart. Further details on stress-strain behavior of heart muscle may be found in Fung (1981).

The macroscopic arrangement of heart muscle fibers is also very different from skeletal muscle. In skeletal muscle usually all fibers of a muscle have more or less the same orientation in space to produce a contractile force in a particular direction. Heart muscle is wound in approximately helical layers around the chambers of the heart. Along a line normal to surface of the heart, the orientation of the muscle fibers varies continuously through the thickness of the heart wall by 90° to 120°. The anatomy of the heart muscle is extensively illustrated by Streeter (1979) and models incorporating this geometry are developed in Glass and Hunter (1990).

4.3 Smooth Muscle

Smooth muscles differ from skeletal and heart muscles in that the smooth muscles do not have the regular arrangement of actin and myosin which form the sarcomere striations. The contractile elements in smooth muscle are also actin and myosin, but these fibers are arranged in a more random manner. The actin filaments are attached to centers called dense bodies and form a network more or less aligned in the principal length direction of the fiber. The myosin filaments have a concentration in smooth muscle which is less than 10% of that in skeletal muscle. The contraction of smooth muscle is comparatively slow (1–3 s) but may maintain some contractile force continuously.

Two types of smooth muscle are distinguished as multi-unit smooth muscle and visceral smooth muscle. Multi-unit smooth muscle consists of fibers that act independently and are stimulated by the sympathetic nervous system. Multi-unit smooth muscles are found, for example, in arterial walls. Visceral smooth muscle fibers are more organized and form a network or connected sheets which contract as a unit. This type of smooth muscle is found in the intestinal walls, the ureters, uterus, bile ducts, and vas deferens. Visceral smooth muscles may undergo spontaneous rhythmic contractions with a period of the order of 1 min. Further details of smooth muscle properties may be found in the *Handbook of Physiology* (1980), Sec. 2, Vol. 2, and Fung (1981).

4.4 Coupling of Muscles and Skeleton

Muscles are generally connected to skeletal elements via tendons which are comprised of stiff fibers of collagen. Tendons insert directly into the bones and operate effectively only in tension. Thrusting and retracting movements are produced by having two sets of muscles and tendons operating on opposite sides of a joint which acts as a fulcrum. An example, the elbow, is shown in Fig. 5. The biceps muscle (agonist) acts to close the angle of the elbow; the triceps (antagonist) acts to straighten out the arm.

In most muscles, the tendon and muscle fibers have more or less parallel directions and the tendons are similar at the two ends [Fig. 6(a)]. However, other arrangements occur in which muscle fibers connect at distinct angles to the tendon. These pennate arrangements are illustrated in Fig. 6(b). A bipennate arrangement occurs inside the claw of the crab (Alexander, 1983).

The leverages and linkages of muscles, bones, and tendons in various animals are extensive and varied. A famous and remarkable early (1681) work "On the Movements of Animals" by G. A. Borrelli has recently been translated into English (Maquet, 1989). Modern works, which clarify and update much of the speculation of Borelli, are books by Alexander (1983) and McMahon (1984).

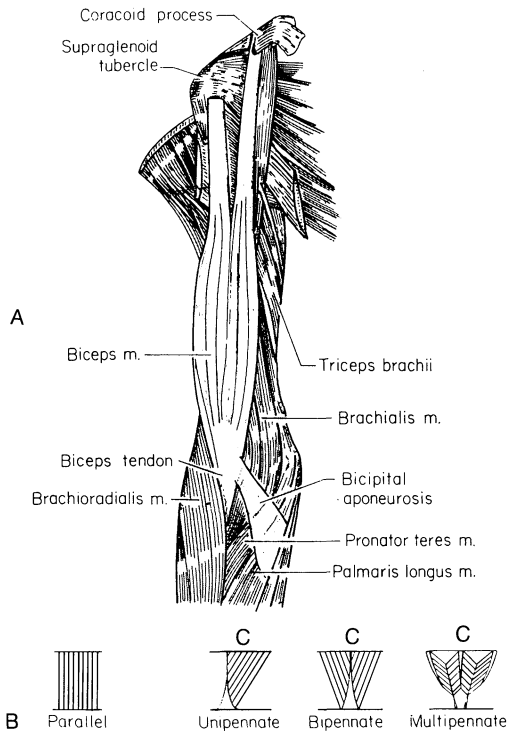

FIG. 6. (a) Muscles of the upper arm. (b) Categories of muscle fibres and tendons arrangements. In (b), the line labeled C in each case is a midline between the two ends of the muscle. (From McMahon, 1984, by permission.)

5. MECHANICS OF THE LUNG

The mechanics of the lung is a particularly interesting subject as it involves the flow of blood on one side and/or air on the other side of membranes only about 1 μm thick arranged in a complex and remarkably efficient manner. Several subtopics may be defined in the mechanics of the lung, each of which is an extensive subject. These may be broadly indicated as blood flow mechanics, air flow and diffusion, gas exchange, elasticity and surface tension, control of breathing, lymph flow, and oedema formation. For discussions of blood flow and air flow, the reader is referred to the article BIORHEOLOGY. The present article will cover some mechanical effects of elasticity and surface tension. Other topics may be found in the books by Fung (1984), West (1977), and Guyton (1986); in the *Handbook of Physiology* (1981), Sec. 4; and in the *Handbook of Bioengineering* (1987).

5.1 Structure of the Lung

Most of the volume of the lung is made of small terminal air sacs called alveoli which are polyhedral in shape and have one wall open, the mouth through which gases enter and leave, mostly by diffusion at this level. The dimensions of alveoli are about 150 μm and there are about 300 million of them in the human lung.

The walls of each alveolus may be described in an idealized manner as consisting of two parallel membranes connected by a regular array of posts. The blood flow takes place between the membranes and the air is on the outside. An increase in blood pressure can expand the space between membranes and reduce the resistance to blood flow (see Fung, 1984). An increase in the alveolar air pressure reduces the blood flow space, and if the alveolar air pressure exceeds that of the blood, the blood flow is cut off. This leads to a variety of interesting possibilities which will be discussed in the next section.

Each alveolus faces onto an air passageway or bronchiole which is in turn joined to a larger bronchial airway and so on until the main bronchus is reached. Alveoli are like expandable leaves on a tree whose twigs and branches are the hollow bronchi. The blood flow system has two more trees in it: one of arterial blood vessels and one of venous blood vessels which tend to run close to the arterial tree until the finest level supplying and draining the alveolar walls.

Lungs are divided into several lobes, each of which is something like a balloon filled with alveoli. The exterior of each lobe carries a thin pleural membrane which is separated from the chest wall by a very thin fluid layer. This allows the lobes of lung to slide relative to the chest wall.

5.2 Zones of the Lung

In a person in an upright position, gravity produces an appreciable pressure difference in blood pressure from the apex to the base of the lung. The variation in the air pressure throughout the lung is much less because of the much lower density of air. The result is that three different zones can be distinguished:

zone 1, $p_A > p_a > p_v$;
zone 2, $p_a > p_A > p_v$;
zone 3, $p_a > p_v > p_A$;

where p_a and p_v stand for the arterial and venous blood pressures and p_A is the alveolar air pressure. In zone 1, which may occur at the apex of the lung, the blood flow is cut off. In zone 3, the capillary bed of the alveolar walls will be fully open to blood flow. In zone 2, there will be some blood flow, but there will be a narrowing of the blood vessels. In this condition it is found that the flow is independent of venous pressure, but depends only on the difference $p_a - p_A$. This condition is sometimes likened to a waterfall because the flow over a waterfall depends only on the level of the water upstream (arterial pressure) and the height of the lip (alveolar air pressure) but not on the tailwater elevation to which it falls (venous pressure).

5.3 Elasticity and Surface Tension in the Lung

The alveolar walls of the lung, together with the airways, blood vessels, and connecting septa, form a continuous three-dimensional structure somewhat like a foam. Also like a foam, each alveolar wall is normally under some tension. The lung is in fact inflated like a balloon. But since the air pressure in the lung is, on the average, only equal to atmospheric pressure, the inflation pressure difference required from inside to outside must be provided by having the pleural space (thin fluid layer) around the lung at subatmospheric pressure. Remarkably, this is provided by nature by appropriate osmotic and lymph flow effects. This negative pressure is of the order of 6 cm H_2O below atmospheric pressure.

The elastic recoil of the lung is due in part to the stretching of the alveolar walls, but an even larger part of the elasticity observed is attributable to surface tension. The interface between the solid tissue of the alveolar wall and the air in an alveolus is coated with a thin layer of fluid containing a surfactant manufactured by the cells of the lung. This surfactant reduces the surface tension to a level of about 5–40 dyn/cm, well below that of an air-water interface (70 dyn/cm). This reduction of surface tension reduces the work of breathing and makes the lung less susceptible to collapse of alveoli by surface tension. Alveoli do not collapse by emptying one alveolus into the other because the surface tension and alveolar air pressure act on both sides of each alveolar membrane. The free edges of the mouths of the alveoli are kept from collapsing by a ring of elastic fibers which circle each opening and are connected to nets of similar connective tissue fibers in the alveolar walls.

6. COCHLEAR AND VESTIBULAR MECHANICS

The inner ear contains two systems which transduce mechanical events into neural signals. The cochlear apparatus (Fig. 7) responds to sound waves impinging on the ear drum while the vestibular apparatus senses linear acceleration, angular motion, and orientation with respect to the direction of gravity. Each of these systems has complex mechanical responses which aid in coding the neural signals sent to the brain.

6.1 Structure of the Inner Ear

In humans, the pressure fluctuations of sound reaching the ear drum set the small bones of the middle ear [malleus, incus and stapes (Fig. 7)] into vibration which in turn generates a mechanical wave motion through the spiral of the cochlea. Frequency discrimination is established primarily in the cochlea. The extreme range of human hearing is about 20 Hz to 20 kHz. Certain animals such as the elephant can respond to lower frequencies which propagate well through a jungle. Some species, such as bats, whales, and dolphins, which use echo ranging for prey and obstacle detection, regularly use frequencies in the range of 100–200 kHz.

The main components of the vestibular system are the semicircular canals, which sense angular motion, and the utricle and saccule, which sense linear acceleration and gravity. The three rings of the circular canals are at right angles to each other which allows for three-dimensional sensing. While these structures are located in the inner ear in close proximity to the cochlea, they are mechanically distinct and transduce neural signals to separate sets of nerve fibers.

6.2 Frequency Discrimination

Acoustic signals are transduced in the cochlea by so-called hair cells attached to the basilar membrane which runs the length of the spiral of the cochlea. The basilar membrane tapers along the spiral channel and its amplitude of vibration varies along its length, depending on the frequency of excitation (see ACOUSTICS, PHYSIOLOGICAL). On both sides of the basilar membrane there is a fluid-filled space and the motion of this tapered membrane-fluid system has been extensively analyzed (see Steele, 1987). The fluid motion bends the fine "hairs," thus stretching the cell membrane and initiating the mechanoelectrical transduction. There is also a contribution to the tuning (localization of frequency response) by an active process which utilizes electrical potential differences to amplify selected frequencies.

The details of the structure of the inner ear vary greatly in different animal species. In bats, whales, and dolphins, who use very high (100–200 kHz) echo-ranging signals, the cochlea is a heavier, bony structure with very stiff elements throughout. In lizards, snakes, and turtles, the cochlea has a simpler structure, but the tuning due to the transducing cells (hair cells) is still present to a large extent. The mechanism of this tuning is a subject of continuing current research.

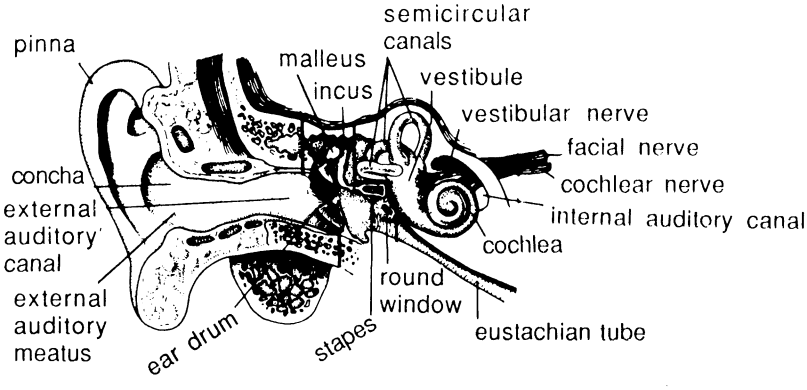

FIG. 7. Cross section of human ear showing outer, middle, and inner ear. (Adapted from Yost and Nielson, 1977, by permission.)

6.3 Vestibular System Mechanics

The three semicircular canals of the inner ear each have an enlarged portion called the ampulla which contains a gelatinous partition called the cupula. Mechanical deformation of the cupula is transduced by neural cells into electrical signals which are sent via axons to the brain. During angular acceleration, the fluid (endolymph) lags behind the motion of the bony walls which results in a relative motion of the fluid and its bony housing. This relative motion produces deformation of the cupula which is then transduced to neural impulses (see AEROSPACE MEDICAL PHYSICS). The fluid mechanics and frequency response of this system are discussed by Van Buskirk and Grant (1987). They also developed a theory for the otolith organs, which are linear acceleration sensors. These units contain a thin layer of calcium carbonate crystals (otoconial membrane) which rests on a soft gelatinous membrane (cupular membrane). The calcium carbonate crystals have a specific gravity of 2.71 and their inertia causes a shear deformation of the cupular membrane during linear acceleration parallel to the membrane plane. Two such units (the utricular and saccular maculae) in different planes allow sensing of linear acceleration in any direction. These deformations of the cupula membranes are transduced to neural impulses by cellular structures of the nerve endings.

7. SWIMMING AND FLYING

Observations and theories of swimming and flying of fish, birds, and insects have become analytical and quantitative only in the last 30 to 40 years. The aerodynamic theories and fluid mechanics developed to design airplanes and ships had little impact on the theory of biological counterparts until about 1950. Since then, many papers and books have reflected considerable progress (see Lighthill, 1975; Pedley, 1977; Alexander, 1983).

Different mechanisms of swimming naturally divide according to the type of motion involved and the size of the animal, or more accurately the Reynolds number range in which the motion takes place. The usual Reynolds number, R, used is

$$R = UL/\nu, \tag{6}$$

where U and L are the velocity and characteristic length of the animal, and ν is the kinematic viscosity of the external fluid (usually air or water). Because of their small size, the Reynolds numbers of sperm and microorganisms are generally in the range of 10^{-3}. In this range all inertial effects are negligible and analyses of swimming modes are problems in viscous flow (Stokes flow). Birds and fish generally operate at much higher Reynolds numbers (10^2–10^6). Here inertial effects are important and, in fact, usually essential, as in the production of lift by a wing. Some of the corresponding theories for low and high Reynolds numbers and various types of propulsion (cilia, flagellate, undulatory, flapping, clapping) are outlined in the following sections. A pictorial summary of swimming modes from Lighthill (1975) is shown in Fig. 8.

7.1 Swimming of Microorganisms

Some protozoa and sperm cells swim by creating an undulatory motion of a long thin tail or flagellum. A two-dimensional theory of such motion at low Reynolds number was first developed by Taylor (1952). Later theories (e.g., Blake, 1974) treat the three-dimensional flow field by considering the flagellum to be replaced by a distribution of stokeslets (point forces). Flagella produce propulsion by actively propagating waves of lateral (or rotatory) displacement rearwards along their length. The resistive viscous (stokeslet) force developed has a useful, propulsive component in the direction of motion.

Another type of microorganism motion utilizes a large number of cilia, which are like fine, short hairs covering a surface. Cilia beat in a rhythmic, coordinated fashion so that waves of displacements are propagated rearwards as envelopes of the outer tips of the cilia. Models of such motions are also based on use of stokeslet distributions, but in this case, taking the boundary into account by the use of appropriate images in the boundary surface (Blake, 1974). Ciliary propulsion is limited to small organisms because it is comparatively slow and because the structural requirements for a large cilium would be severe. By contrast, undulatory propulsion can be effective over the full range of Reynolds

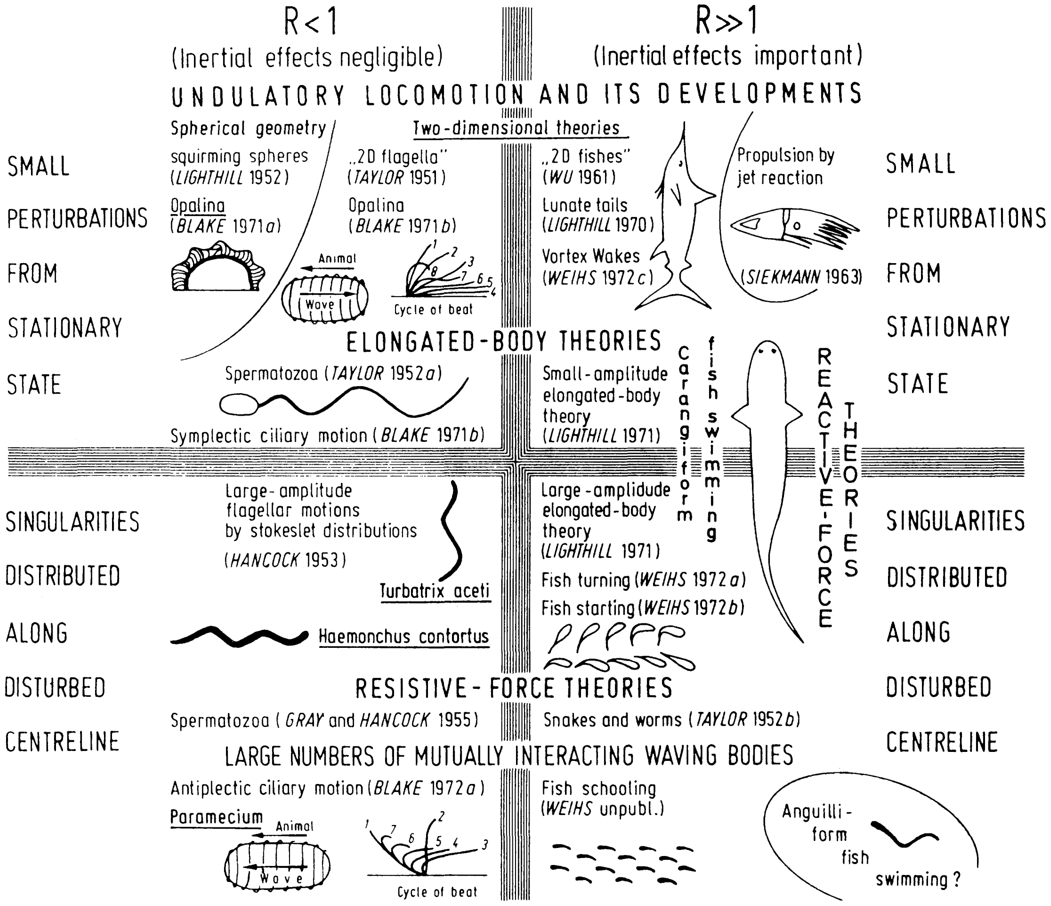
R<1
(Inertial effects negligible)
R>>1
(Inertial effects important)
UNDULATORY LOCOMOTION AND ITS DEVELOPMENTS
Spherical geometry
Two-dimensional theories
SMALL PERTURBATIONS FROM STATIONARY STATE
squirming spheres (LIGHTHILL 1952)
Opalina (BLAKE 1971a)
"2D flagella" (TAYLOR 1951)
Opalina (BLAKE 1971b)
Animal
Wave
Cycle of beat
"2D fishes" (WU 1961)
Lunate tails (LIGHTHILL 1970)
Vortex Wakes (WEIHS 1972c)
Propulsion by jet reaction
(SIEKMANN 1963)
ELONGATED-BODY THEORIES
Spermatozoa (TAYLOR 1952a)
Symplectic ciliary motion (BLAKE 1971b)
Small-amplitude elongated-body theory (LIGHTHILL 1971)
carangiform fish swimming
REACTIVE-FORCE THEORIES
SINGULARITIES DISTRIBUTED ALONG DISTURBED CENTRELINE
Large-amplitude flagellar motions by stokeslet distributions (HANCOCK 1953)
Turbatrix aceti
Haemonchus contortus
Large-amplidude elongated-body theory (LIGHTHILL 1971)
Fish turning (WEIHS 1972a)
Fish starting (WEIHS 1972b)
RESISTIVE-FORCE THEORIES
Spermatozoa (GRAY and HANCOCK 1955)
Snakes and worms (TAYLOR 1952b)
LARGE NUMBERS OF MUTUALLY INTERACTING WAVING BODIES
Antiplectic ciliary motion (BLAKE 1972a)
Paramecium
Animal
Wave
Cycle of beat
Fish schooling (WEIHS unpubl.)
Anguilliform fish swimming?

FIG. 8. Subdivisions of hydrodynamic theory relevant to aquatic animal locomotion. (From Lighthill, 1975, by permission.)

numbers, although the source of propulsive force changes from viscous to inertial as Reynolds numbers increase.

7.2 Modes of Swimming of Fish

The range of Reynolds numbers for a small fish swimming slowly to a blue whale at top speed may vary from 10^3 to 10^9. The propulsive mechanisms in this range are various, but all depend on unsteady inertial forces. Viscous effects are limited to a thin boundary layer which produces the predominant component of drag. The drag is overcome by propulsive forces which may be computed by inviscid flow theory.

For long thin fish, like the eel, the common form of propulsion is by creation of an undulatory wave which is propagated rearward along the length of the body. This is called anguilliform propulsion. The amplitude of the motion typically increases rearward and the tail makes much wider lateral excursions than the head. Lighthill (1975) gives a qualitative explanation of the propulsive force generated based on the relation

$$w = W(V-U)/V, \tag{7}$$

where U is the mean forward velocity of the fish, V is the velocity of propagation of the wave rearward, W is the lateral velocity of the fish, and w is the effective transverse velocity of the tail. The thrust arises due to the inertial reaction of the water being set in motion primarily as a two-dimensional flow in planes normal to the undulating axis. U is typically about $0.6V$, for efficient forward propulsion. The lateral momentum of the fluid generated when the tail passes a given point is proportional to w and the force reaction on the fish is equal to the rate of change of momentum proportional to Uw. The rate of work done is then proportional to UwW. Part of this work goes into vortex wake production and is lost in the fluid. The remainder is useful work of propulsion (which is also lost to the fluid through viscous drag).

Anguilliform propulsion does not require that the entire length of the fish participates in the lateral undulatory motion, but does require that the amplitude increase to a maximum value at the tail. In many fish, the lateral undulatory motion is restricted to the latter half or third of the fish's length. This is called the carangiform mode of propulsion. The effective part of the fish in producing force is then the tail which may be extended (vertically) and shortened (lengthwise) for efficiency. However, the basic rearward propagating wave motion is preserved even if the undulating region is less than a full wavelength. For high speed and efficiency, the ultimate form of the tail of a fish evolves to a crescent moon shape called the lunate tail typical of sharks, marlin, and tuna. Aquatic mammals such as dolphins, whales, and mermaids also have lunate tails, but these are flukes in the horizontal plane which undergo undulatory motion vertically. A lunate tail generates a vortex at each point that it reverses its transverse velocity, W. This leaves a Kármán vortex street in the fluid behind the fish which represents lost energy, but is necessary to produce forward thrust.

There are a variety of other mechanisms for swimming which are used by smaller classes of fish. Some so-called armored fish, like the sea horse, swim by dorsal and pectoral fin motions. Jellyfish and squid use an intermittent jet propulsion. Rays and skates use a large horizontal area with a flapping motion resembling the flight of a bird. Sea turtles have limbs grown into flippers which also swim by flapping like a bird's wings.

7.3 Flight of Birds

Most birds fly by a flapping motion of their wings which produces both lift and forward thrust. The up and down strokes are not symmetric either in time or in inclination of the wings. The down stroke is longer and takes place with the plane of wing more nearly horizontal. The combined downward motion of the wing and forward motion of the bird gives lift and some thrust. During the upward stroke, the wing is also tilted upwards and typically the wing tip is drawn in and back relative to the body. Both lift and thrust are smaller in the upstroke. Many birds are able to separate the primary feathers of their wing tips during the downstroke which enhances lift as in leading-edge slots of high-lift airplane wings. The tails of most birds are used primarily for control, but may also be used for braking on landing and for more efficient soaring. Most soaring birds glide at a low angle within a rising column or bubble of thermal currents. During soaring, no active lift is generated, but altitude may be gained or maintained by the upward motion of the air.

Hovering flight of the humming bird may be regarded as an adaptation of flapping forward flight in which the motion of the wings is more or less in a horizontal plane. The body of the bird is nearly vertical to allow the movement to and fro in a horizontal plane, and the wings are rotated rapidly at the end of each stroke so that the same edge is leading and the angle of attack is positive during both forward and backward strokes. Many larger birds which do not hover because the power requirements would be excessive nevertheless use a more nearly hovering mode during

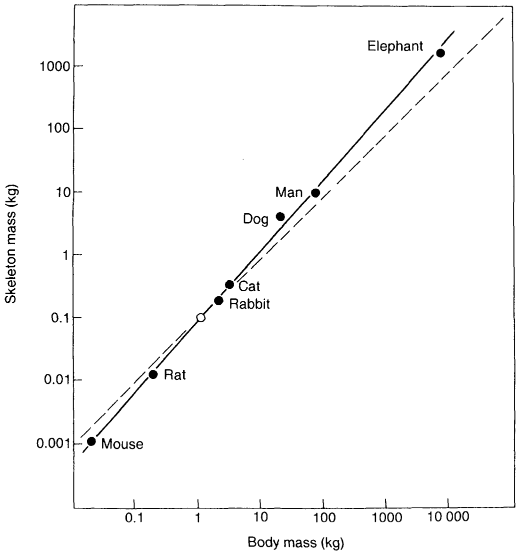

FIG. 9. Log-log plot of mass of mammalian skeletons vs body mass. The solid regression line has a slope of 1.13. The dashed line has a slope of 1.0. (From Schmidt-Nielsen, 1977, by permission.)

landing and briefly at steep takeoff. The wings beat more nearly horizontally and twist to maintain lift.

7.4 Modes of Insect Flight

Insect flight preceded that of birds in an evolutionary sense, dating back 3×10^8 years whereas birds evolved about 1.6×10^8 years ago. Evolution of flight mechanisms is summarized by Lighthill (1975) and studied in detail by Smart and Hughes (1972).

Many insects use a flapping motion for forward flight that resembles that described above for birds. The wings are rotated to maintain a positive angle of attack during both down and up strokes. Insects hovering may use a wing motion similar to that of humming birds, but other systems have also evolved. The dragon fly uses a complex coordinated motion of four wings. Another remarkable system involves clapping; the wings are brought together at the end of each cycle. This is the clap. The wings then separate by rotation about a line along the common trailing edge, like opening a closed book. Air sucked into the space between the unfolding wings generates a circulation and hence lift on the rest of the downstroke. This motion was first analyzed by Weis-Fogh (1973) studying the hovering of a small wasp. It is fundamentally different from other motions producing lift described above.

8. SCALING AND LIMITATIONS IN BIOMECHANICS

Although the size of animals measured in terms of total body mass ranges from approximately 10^{-16} kg for some bacteria to 10^5 kg for a whale, the range of size of a particular kind of animal is limited by constraints of material strength, energy generation capabilities, and physical laws, including geometric factors strongly. Moreover, within broad categories of animals, parameters such as metabolic rates and top speed vary in systematic ways with size and for discernible reasons (see McMahon and Bonner, 1983). Scaling laws are extensively discussed in the Symposium volume edited by Pedley (1977) and the book by McMahon (1984). Some of the salient results are outlined below.

8.1 Size and Power Limitations

One of the apparently simplest scalings is the estimation of skeletal dimensions versus body weight. But the realities are not so simple as the theory and illustrate the kind of pitfalls possible. If different sized animals were simply scaled up equally in all dimensions, the mass of the skeleton M_s would be a constant fraction of the total body mass M_0. But, in fact, a larger fraction of the body mass is occupied by skeleton in larger animals. This is a consequence of the fact that the strength of bone per unit area is approximately constant in all mammals. Then to support an increased mass, the cross-sectional area A of supporting bones should be in proportion to total mass M_0. The length of any bone L is roughly proportional to ${M_0}^{1/3}$. Hence the skeletal mass should be proportional to ${M_0}^{1.33}$. But the reality is (Schmidt-Nielsen, 1977) (see Fig. 9)

$$M_s=0.1{M_0}^{1.13}. \tag{8}$$

This shows that static load of total mass is not the controlling factor. Impact loading controls, and even then the accelerations that different animals invoke vary greatly. As Schmidt-Nielsen (1977) remarks, if an elephant were to jump like a mouse, he might break his bones on take-off and totally collapse on impact. Nevertheless, the strength of bone limits the total size (mass) of terrestrial animals by virtue of the strength of bone being relatively a constant. The heaviest terrestrial animal was the dinosaur baluchitherium weighing about 30 tons. If evolution had developed stronger materials than bone and higher power capabilities per unit weight than muscle, then much bigger, faster, and more powerful mobile entities could evolve. The huge open-mine steam shovels, ocean liners, nuclear submarines, jet planes, and spacecraft are existing examples, although they are not yet self-reproducing.

The limitations of power generation spring from two sources. One is metabolic rate which is a measure of power-generating capacity. The other is muscle properties; most muscles have similar maximum stress and maximum strain-rate capabilities.

Resting or basic metabolism rates in warm-blooded animals appear to be a function of heat loss to surroundings. Hence, basic metabolism w_0 (in watts) might be expected to vary in proportion to surface area, which in turn varies approximately as $M_0^{2/3}$. But data from many animals show that

$$w_0 = KM_0^{0.75}, \qquad (9)$$

where K is a constant. The difference between $\frac{2}{3}$ and 0.75 has been attributed to the fact that animals are not strictly geometrically similar (McMahon, 1973).

8.2 Scaling in Terrestrial Locomotion

In higher vertebrates the maximum stress that muscle can develop is about 300–400 kN m^{-2}, independent of the size of the animal. The maximum strain (shortening) is about 0.3, also relatively constant. Hence the work performed per contraction per unit volume of muscle is relatively constant independent of body size. By contrast, the frequency of contraction generally increases with decreasing body size so that the power (watts) delivered by a unit volume of muscle is higher in small animals and their peak metabolism is greater per unit volume. Incidentally, the volume of a muscle is very nearly constant during any contraction.

Data on various four-legged animals show that the metabolic cost (as measured by O_2 consumption) of moving a unit mass of the animal a unit distance increases dramatically as the size of the animal decreases. This specific metabolic cost varies as $M_0^{-0.4}$. An interesting aspect of muscle behavior is that metabolic energy is consumed both when muscles perform positive work and when they perform negative work, as in braking an impact. In level running, there is no change in kinetic energy, potential energy, or elastic strain energy over a complete stride. But metabolic work is done in accelerating and decelerating limbs relative to the body. Energy is also consumed in the imperfect elasticity of the system. In running, the body bounces like a mass on a spring, but the legs are imperfect springs and energy is lost in each rebound.

The quantitative relations of maximum speed and power consumption as a function of size are complicated by variations of geometry, skeletal mass, stride frequency, and gait. A theory based on geometric similarity and the muscle properties mentioned above was given by Hill (1950). The maximum running speed was found to be independent of size and the metabolic power at top speed varied as $M_0^{2/3}$. Within certain limits, these predictions are approximately correct. A greyhound and a horse have nearly the same top speed, around 18 m/s, but extrapolations to larger and much smaller animals fails.

A more elaborate analysis has been given by McMahon (1984) based on elastic similarity. The basic idea of elastic similarity is that the elastic deflection of a structure under its own weight should be proportional to its length. This leads to a law of transverse dimensions (say diameter d) being proportional to $L^{3/2}$. This in turn leads to realistic predictions of the variations of galloping frequencies, rise of center of mass during a stride, joint angle excursions, and power consumption. However, a coherent theory for prediction of top speeds for the full range of existing animal sizes is not yet assured.

8.3 Scaling in Aquatic Locomotion

The main determinant of similarity of fluid motion in swimming is the Reynolds number, Eq. (11). At low Reynolds numbers the flagellate type of swimming has been analyzed by Lighthill (1975) to show that there is a maximum efficiency which is achieved when the ratio of forward speed U to the rearward wave speed V is about 0.3. The peak efficiency, however, is low, around 0.1. In all flagellate microorganisms the cross section and molecular structure is similar. Since the propulsive power per unit length is then more or less constant, the swimming velocity U is relatively constant, independent of total length.

In high–Reynolds-number fish swimming, the maximum speed increases with length or size of the fish because of the nature of the drag and power available. The drag D is proportional to $\frac{1}{2}\rho U^2$ times surface area, A, and a drag coefficient, C_D. The metabolic power available is increased with body mass M_0 to some power b between about 0.8 and 1.0. Assuming M_0 is proportional to L^3 and surface area S is proportional to L^2, where L is the length of the fish, equating P to UD yields

$$U = kL^{\beta}, \qquad (10)$$

where k is a constant and β varies from about 0.11 to 0.60 depending on the variation of the drag coefficient C_D with Reynolds number. Wu (1977) shows that Eq. (10) can be fitted to a variety of sizes of fish under maximum or lower states of effort. In all cases, the speed of swimming increases with size.

One consequence of the above scaling law is that the specific energy cost S (energy expended per unit of mass moved a unit distance) decreases with size. The theory and data show that (Wu, 1977)

$$S = S_0 M_0^{-\gamma} \tag{11}$$

where S_0 is a constant and γ varies from 0.11 to 0.34 for different cases.

Another aspect of scaling concerns the frequency ω (rad/s) of beating of the tail motion. In carangiform propulsion, Bainbridge (1958) states that the reduced frequency ($\omega L/U$) is approximately constant and equal to 10 independent of length L. Wu (1977) interprets this as an implication that the hydrodynamical efficiency is independent of size. Due to the variation of U with L, ω itself is proportional to $L^{-\alpha}$, where α is of the order of 0.88. The beat frequency of birds' wings also decreases as size increases.

8.4 Scaling in Flight

Scaling in flight mechanisms is more complex than in swimming because in addition to propulsion to overcome drag, a lift must be provided simultaneously. The scaling of lift and power limits the maximum wing span of birds. Either more power or strength of structure per unit weight would allow larger birds.

The fact that large birds fly faster than small ones may be understood in terms of wing loading w defined as total weight gM_0 divided by wing area S. S increases proportionately to $M_0^{2/3}$ so wing loading w increases as $M_0^{1/3}$. To maintain higher wing loadings, higher speeds are required to operate at the best lift/drag ratio. The trend (of scattered data for many birds) is that the flight speed U varies as $M_0^{1/6}$.

The mass of the pectoralis major muscle of birds is about 15% of total body mass M_0. This enters into considerations of power and maximum size of birds. Available power is proportional to muscle mass times contraction (flapping) frequency. As the wing span of a bird increases, the relative wing mass increases since the wings act as cantilever beams for which the stress increases rapidly with span; in fact, to keep peak bending stress constant, wings must thicken and total mass rises more quickly than available power which suffers as size increases because larger size means lower flapping frequency. Tucker (1977) estimates all these factors in a quantitative formula and concludes that a wing span of about 15 m is a realistic upper limit. For flapping flight, the maximum climb angle goes down with increasing wing span until it becomes completely impractical.

8.5 Scaling in the Cardiovascular and Respiratory Systems

The mean arterial blood pressure of mammals from the size of a mouse to a horse is very nearly the same. Further, the muscle stress during contraction is also relatively constant. It follows that hearts of different sizes are more or less geometrically similar. Measurements of many animals show that the volumes associated with the heart and lungs vary directly with the total body mass. West (1977) has also shown that alveolar surface area varies directly with body mass.

The heart rate of small animals is much faster than for large ones. This may be understood in terms of the metabolic rate variation given in Eq. (9) and the linear variation of heart size with body mass M_0. The stroke volume (blood volume ejected per stroke) is proportional to heart size and hence to M_0. The mean blood flow required is proportional to metabolic rate and hence to $M_0^{0.75}$. Hence the heart rate (beats per minute) must vary as $M_0^{-0.25}$. Measurements on various animals confirm this prediction. For similar reasons, the respiratory frequency (breathing cycles per minute) also varies as $M_0^{-0.25}$ (see McMahon, 1984).

The variation of heart rate with body size has another interesting result concerning wave propagation in the arterial system. The speed of a pressure wave in an artery is given approximately by the Korteweg-Moens formula

$$C = (Eh/2\rho R)^{1/2}, \tag{12}$$

where E and h are the Young's modulus and thickness of the arterial wall, R is the artery radius, and ρ is the density of the blood. It is found that C is relatively constant at about 4 m/s in mammals. The increasing heart rate with decreasing body size indicated above results in a constant ratio of wavelength of pressure waves at the fundamental heart rate to the length of the main arteries. The mean distance from the heart to capillaries is approximately 1/4 of a wavelength in all animals. This is a favorable value from the standpoint of the mechanical impedance that the arterial system presents to the heart (McMahon, 1984).

Another interesting aspect of heart rate concerns life span. In terrestrial mammals, life span has been shown to vary approximately as $M_0^{0.25}$. The variation of heart rate as $M_0^{-0.25}$ then results in the conclusion that all mammals live for the same total number of heart beats per lifetime. No particular reason for this result is known.

In conclusion, it may be said that scaling laws have some validity and many broad trends of size variations follow roughly their predictions. But individual variations of form and functional adaptation make scaling only an approximate guide in specific cases.

GLOSSARY

Actin: A type of fiber found in most cells; a principal component in muscle.

Alveolus: The smallest subdivision in a lung; a small air sac.

Anguilliform Propulsion: The mode of swimming by causing a lateral wave motion, like an eel.

ATP: adenosine triphosphate.

Carangiform Propulsion: A mode of swimming in which only the tail and hinder part of the fish is oscillated.

Chondrocytes: Cells which produce cartilage.

Cochlea: The spiral apparatus in the inner ear which transduces sound to neural impulses.

Collagen: A fibrous material found in tendons and many other connective tissues.

Cortical Bone: The part of a bone that is more dense and usually at the surface.

Diarthroidal Joints: Flexible joints such as the elbow or knee.

Diastole: The portion of the heart cycle during which the heart muscle is relaxed.

Elastic Similarity: The proportions which arise if the elastic deflection of a structure is a constant fraction of its size.

Fascicles: Subdivisions of a muscle.

Flagella: Long thin structures of microorganisms that are actively oscillated to produce propulsion.

Hydroxyapatite: A hard, brittle component of bone.

Lunate: Shaped like a crescent moon.

Myosin: A fiber that produces contractile force in muscle by its interaction with actin.

Pectoralis Major: The main muscle used by birds on the downstroke of wing flapping.

Pennate: A shape of a muscle resembling a pennant.

Pleural Space: A very thin, fluid layer surrounding the lung.

Sarcomere: A microscopic unit of sliding filaments of a muscle.

Smooth Muscles: Muscles which have a homogeneous microscopic appearance and contract slowly.

Squeeze Film: Lubrication that derives from squeezing out a fluid from between two surfaces.

Synovial Fluid: The lubricating fluid found in joints.

Systole: The portion of the heart cycle during which the heart is actively contracting.

Tetanized: The state of a muscle stimulated to its maximum contraction.

Trabecular Bone: A porous form of bone composed of a network of bony struts and plates.

Vestibular System: The semicircular canals and organs which provide rotational sense and balance.

List of Works Cited

Alexander, R. M. (1983), *Animal Mechanics, 2nd ed.*, Oxford: Blackwell Scientific.

Bainbridge, R. (1958), *J. Exp. Biol.* **35**, 109–133.

Berne, R. M. (1979), *Handbook of Physiology*, Bethesda: American Physiological Society, Sec. 2, Vol. I.

Blake, J. R. (1974), *J. Theor. Biol.* **45**, 183–203; 52, 67–82.

Bohr, D. F., Somlyo, A. P., Sparks, H. V. (Eds.) (1980), *Handbook of Physiology*, Bethesda: American Physiolological Society, Sec. 2, Vol. II.

Carter, D. R. (1988), in: R. Skalak and C. F. Fox, (Eds.), *Tissue Engineering*, New York: Alan R. Liss, pp. 173–180.

Carter, D. R., Hayes, W. C. (1977), *J. Bone Jt. Surg.* 59A, 954–962.

Cowin, S. C., Van Buskirk, W. C., Ashman, R. B. (1987), in: R. Skalak and S. Chien (Eds.), *Handbook of Bioengineering*, New York: McGraw-Hill, Chap. 2.

Cowin, S. C. (1988), in: R. Skalak and C. F. Fox (Eds.), *Tissue Engineering*, New York: Alan R. Liss, pp. 181–188.

Fung, Y. C. (1981), *Biomechanics: Mechanical Properties of Living Tissues*, New York: Springer-Verlag.

Fung, Y. C. (1984), *Biodynamics: Circulation*, New York: Springer-Verlag.

Fung, Y. C. (1990), *Biomechanics: Motion, Flow, Stress, and Growth*, New York: Springer-Verlag.

Fung, Y. C., Perrone, N., Anliker, M. (Eds.) (1972), *Biomechanics: Its Foundations and Objectives*, Englewood Cliffs: Prentice-Hall.

Glass, L., Hunter, P. J. (Eds.) (1990, in press), *Theory of the Heart*, New York: Springer-Verlag.

Guyton, A. C. (1986), *Textbook of Medical Physiology*, Philadelphia: Saunders.

Hill, A. V. (1938), *Proc. R. Soc. London, Ser. B* **126**, 136–195.

Hill, A. V. (1950), *Sci. Prog.* **38**, 209–230.

Huiskes, R. (1988), in: R. Fitzgerald (Ed.) *Non-Cemented Total Hip Arthroplasty*, New York: Raven.

Johnson, M. W., Katz, J. L. (1987), in: R. Skalak and S. Chien (Eds.), *Handbook of Bioengineering*, New York: McGraw-Hill, Chap. 3., P. Tong, C. M. Ni, A. King (Eds.),

Laananen, D. H. (1986), in: P. Tong, C. M. Ni, A. King (Eds.), *Symposium on Vehicle Crashworthiness Including Impact Biomechanics*, New York: American Society of Mechanical Engineers, pp. 47–56.

Lai, W. M., Mow, V. C. (1979), *Annu. Rev. Fluid Mech.* **11**, 247–288.

Lanir, Y. (1987), in: R. Skalak and S. Chien (Eds.), *Handbook of Bioengineering*, New York: McGraw-Hill, Chap. 11.

Lanir, Y., Fung, Y. C. (1974), *J. Biomech.* **7**, 29–34; **7**, 171–182.

Lighthill, M. J. (1975), *Mathematical Biofluiddynamics*, Philadelphia: SIAM.

Maquet, P. (Translator) (1989), Giovanni Alfonso Borelli, *On the Movement of Animals* (1681), New York: Springer-Verlag.

McCutchen, C. W. (1962) *Wear* **5**, 1–17.

McMahon, T. A. (1973), *Science* **179**, 1201–1204.

McMahon, T. A. (1984), *Muscles, Reflexes, and Locomotion*, Princeton: Princeton University.

McMahon, T. A., Bonner, J. T. (1983), *On Size and Life*, New York: Scientific American Books.

Mow, V. C., Lai, W. M. (1980), *SIAM Rev.* **22**, 275–317.

Mow, V. C., Mak, A. F. (1987), in: R. Skalak and S. Chien, (Eds.), *Handbook of Bioengineering*, New York: McGraw-Hill, Chap. 5.

Nahum, A. M., Melvin, J. (1985), *The Biomechanics of Trauma*, Norwalk: Appleton-Century-Crofts.

Peachy, L. (Ed.) (1983), *Handbook of Physiology*, Bethesda: American Physiological Society, Sec. 10.

Pedley, T. J. (Ed.) (1977), *Scale Effects in Animal Locomotion*, London: Academic.

Reilly, D. T., Burstein, A. H. (1975), *J. Biomech.* **8**, 393–405.

Schmidt-Neilsen, K. (1977), *Scale Effects in Animal Locomotion*, London: Academic, pp. 1–21.

Skalak, R., Chien, S. (Eds.) (1987), *Handbook of Bioengineering*, NewYork: McGraw-Hill.

Smart, J., Hughes, N. F. (1972), *Symp. R. Entomol. Soc. London* **6**, 143–155.

Steele, C. R. (1987), in: R. Skalak and S. Chien (Eds.), *Handbook of Bioengineering*, New York: McGraw-Hill, Chap. 30.

Streeter, D., Jr. (1979), in: R. M. Berne and N. Sperelakis (Eds.), *Handbook of Physiology*, Sec. 2, *Cardiovascular System*, Vol. 1, *The Heart*, Bethesda: American Physiological Society, pp. 61–112.

Taylor, G. I. (1952), *Proc. R. Soc. London, Ser. A* **211**, 225–239.

Tucker, V. A. (1977), in: T. J. Pedley (Ed.), *Scale Effects in Animal Locomotion*, London: Academic, pp. 497–509.

Van Buskirk, W. C., Grant, J. W. (1987), in: R. Skalak and S. Chien (Eds.), *Handbook of Bioengineering*, New York: McGraw-Hill, Chap. 31.

Viano, D. C., Lau, V. (1988), *J. Biomech.*, **21**, 387–399.

Weis-Fogh, T. (1973), *J. Exp. Biol.* **59**, 169–230.

West, J. B. (Ed.) (1977), *Bioengineering Aspects of the Lung*, New York: Marcel Dekker.

Woo, S. L. Y., Lubock, P., Gomez, M. A., Jemmot, G. F., Kuci, S. C., Akeson, W. H. (1979), *J. Biomech.* **12**, 437–446.

Wu, T. Y. (1977), in: T. J. Pedley (Ed.), *Scale Effects in Animal Locomotion*, London: Academic, pp. 203–232.

Yost, W. A., Nielson, D. W. (1977), *Fundamentals of Hearing: An Introduction*. New York: Holt, Rinehart and Winston.

Further Reading

Alexander, R. M. (1988), *Elastic Mechanisms in Animal Movement*, Cambridge: Cambridge University.

Fung, Y. C. (1981), *Biomechanics: Mechanical Properties of Living Tissues*, New York: Springer-Verlag.

Fung, Y. C. (1984), *Biodynamics: Circulation*, New York: Springer-Verlag.

Fung, Y. C. (1990), *Biomechanics: Motion, Flow, Stress, and Growth*, New York: Springer-Verlag.

Lighthill, M. J. (1975), *Mathematical Biofluiddynamics*, Philadelphia: SIAM.

McMahon, T. A. (1984), *Muscles, Reflexes, and Locomotion*, Princeton: Princeton University.

Skalak, R., Chien, S. (Eds.) (1987), *Handbook of Bioengineering*, New York: McGraw-Hill.

BIOMEDICAL ENGINEERING

JOSEPH D. BRONZINO, *Trinity College/The Hartford Graduate Center, Hartford, Connecticut, U.S.A.*

	Introduction	513
1.	**Evolution of the Modern Health Care System**	514
2.	**What is Biomedical Engineering?**	515
3.	**Present and Future Research Directions in Biomedical Engineering**	519
3.1	Systems Science and Integrated Biology	519
3.2	Cardiovascular Assist Devices	521
3.2.1	Cardiac Pacemakers	521
3.2.2	Defibrillation Techniques	522
3.2.3	Artificial Heart and Cardiac Assist Devices	523
3.3	Biomaterials	525
3.3.1	Metals	525
3.3.2	Ceramics	525
3.3.3	Polymers	527
3.4	Rehabilitation Engineering and Prosthetic Devices	528
3.5	Instrumentation and Sensors	529
3.6	Medical Imaging Technology	531
3.6.1	Radiography	532
3.6.2	Nuclear Medicine	535
3.6.3	Ultrasonic Imaging	538
3.6.4	Magnetic Resonance Imaging	540
3.7	Biomedical Computing	540
4.	**Status of Biomedical Engineering Education**	542
	Glossary	546
	Works Cited	547
	Further Reading	548

INTRODUCTION

In the 20th century, technological innovation has progressed at such an accelerated pace that it has permeated almost every facet of our lives. This is especially true in recent developments in the field of medicine and the delivery of health care services. Although the art of medicine has a long history, the evolution of a health care system capable of providing a wide range of positive therapeutic treatments in the prevention and cure of illnesses is a decidedly new phenomenon. Of particular importance in this evolutionary process has been the establishment of the modern hospital as the center of a technologically sophisticated health care system. In the process of centralizing the support and utilization of medical technology, it became apparent that specialization of a large number of health care professionals was also necessary. In the United States, the result has been the emergence of a technologically dominated health care system focused around the activities of regional hospitals serviced by a technologically sophisticated staff.

Since technology has had such a dramatic impact on medical care, engineering professionals have become intimately involved in many medical ventures. As a result, the discipline of *biomedical engineering* has emerged as an integrating medium for two dynamic professions, medicine and engineering, assisting in the struggle against illness and diseases by providing materials, tools, and techniques (such as imaging and artificial intelligence) that can be utilized for research, diagnosis, and treatment by health care professionals. Thus, biomedical engineers serve as relatively new members of a health care delivery "team" seeking new solutions for the difficult health care problems confronting our society. The purpose of this article, therefore, is to provide a broad overview of technology's role in shaping our modern health care system, to discuss the basic aspects of the field of biomedical engineering, to review some of the contributions made by biomedical engineers including cardiovascular assist devices, med-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ical imaging, and diagnostic support systems, and to present the status of biomedical engineering education.

1. EVOLUTION OF THE MODERN HEALTH CARE SYSTEM

Before 1900, medicine had little to offer the average citizen, since its resources consisted mainly of the physician, his education, and his "little black bag." In general, physicians seemed to be in short supply, but the shortage had rather different causes than more current crises in the availability of health care professionals. Although the costs of obtaining medical training were relatively low, the demand for doctors' services was also very small, since many of the services provided by the physician could also be obtained from experienced amateurs in the community. The home was typically the site for treatment and recuperation, and relatives and neighbors constituted an able and willing nursing staff. Babies were delivered by midwives, and those illnesses not cured by home remedies were left to run their natural, albeit frequently fatal, course. The contrast with contemporary health care practices, in which specialized physicians and nurses located within the hospital provide critical diagnostic and treatment services, is dramatic.

The changes that have occurred within medical science originated in the rapid developments that took place in the applied sciences (chemistry, physics, engineering, microbiology, physiology, pharmacology, etc.) at the turn of the century. This process of development was characterized by intense interdisciplinary cross-fertilization, which provided an environment in which medical research was able to take giant strides in developing techniques for the diagnosis and treatment of disease. For example, in 1903, Willem Einthoven, the Dutch physiologist, devised the first electrocardiograph to measure the electrical activity of the heart. In applying discoveries in the physical sciences to the analysis of a biological process, he initiated a new age in both cardiovascular medicine and electrical measurement techniques.

New discoveries in medical science followed one another like intermediates in a chain reaction. However, the most significant innovation for clinical medicine was the development of x rays. These "new kinds of rays," as their discoverer W. K. Roentgen described them in 1895, opened the "inner man" to medical inspection. Initially, x rays were used to diagnose bone fractures and dislocations. In the process, this "modern technology" became commonplace in most urban hospitals. Separate departments of radiology were established, and their influence spread to other departments throughout the hospital. By the 1930s, x-ray visualization of practically all organ systems of the body had been made possible through the use of barium salts and a wide variety of radio-opaque materials.

X-ray technology gave physicians a powerful tool that, for the first time, permitted accurate diagnosis of a wide variety of diseases and injuries. Moreover, since x-ray machines were too cumbersome and expensive for local doctors' clinics, they had to be placed in health care centers or hospitals. Once there, x-ray technology essentially triggered the transformation of the hospital from a passive receptacle for the sick poor to an active curative institution for all members of society.

For economic reasons the centralization of health care services became essential because of many other important technological innovations appearing on the medical scene. However, hospitals remained institutions to dread, and it was not until the introduction of sulfanilamide in the mid-1930s and penicillin in the early 1940s that the main danger of hospitalization, i.e., cross infection among patients, was significantly reduced. With these new drugs in their arsenals, surgeons were permitted to perform their operations without the prohibitive morbidity and mortality due to infection. Furthermore, even though the different blood groups and their incompatibility were discovered in 1900 and sodium citrate was used in 1913 to prevent clotting, full development of blood banks was not practical until the 1930s when technology provided adequate refrigeration. Until that time, "fresh" donors were bled and the blood transfused while it was still warm.

Once these surgical units were established the employment of the available technology assisted in further advancing the development of complex surgical procedures. For example, the Drinker respirator was introduced in 1927 and the first heart-lung bypass in 1939. By the 1940s, medical procedures

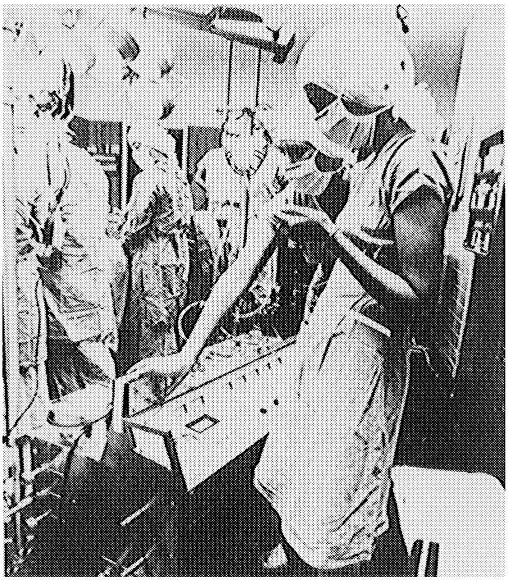

FIG. 1. The operating room scene during the early 1950s illustrates the invasion of medical technology into the clinical theater of operation. (Courtesy of Hartford Hospital.)

heavily dependent upon medical technology such as cardiac catheterization and angiography (the use of a cannula threaded through an arm vein and into the heart with the injection of radio-opaque dye for the x-ray visualization of lung and heart vessels and valves) were developed. As a result, accurate diagnoses of congenital and acquired heart disease (mainly valve disorders due to rheumatic fever) became possible, and a new era of cardiac and vascular surgery was established.

Following World War II, the development of medical devices accelerated rapidly and the medical profession benefited greatly from this rapid surge of "technological finds" (Fig. 1). Consider the following examples:

1. Advances in electronics made it possible to map the subtle electrical behavior of the fundamental unit of the central nervous system—the neuron—as well as to monitor various physiological parameters, such as the electrocardiogram (ECG), of patients in intensive care units.
2. Nuclear medicine—an outgrowth of the atomic age—emerged as a powerful and effective approach in detecting and treating specific physiological abnormalities.
3. Diagnostic ultrasound based on sonar techniques became so widely accepted that ultrasonic studies are now part of the routine diagnostic workup in many medical specialties.
4. "Spare parts" surgery also became commonplace. Technologists were encouraged to provide prosthetic devices, such as artificial heart valves and artificial blood vessels, and the artificial heart program was launched to develop a replacement for a defective or diseased human heart.
5. Computers similar to those developed to control the flight plans of the Apollo capsule were used to store, process, and cross-check medical records, to monitor patient status in intensive care units, and to provide sophisticated statistical diagnosis of potential diseases correlated with specific sets of patient symptoms (Fig. 2).
6. Development of the first computer-based medical instrument, the computerized axial tomography (CAT) scanner, revolutionized clinical approaches to noninvasive diagnostic imaging procedures which now include magnetic resonance imaging (MRI) and positron emission tomography (PET) as well.

The impact of these discoveries and many others was profound. The health care system consisting primarily of the "horse and buggy" physician was gone forever. It was replaced by a clinical staff operating primarily in hospitals which began to change to accommodate the new medical technology. The "modern hospital," therefore, in its contemporary, familiar form is essentially less than sixty years old (Crichton, 1970; Knowles, 1973; Bronzino, 1977, 1986).

2. WHAT IS BIOMEDICAL ENGINEERING?

Many of the problems confronting health professionals today are of extreme interest to engineers because they involve the design and practical application of medical devices and systems—processes that are fundamental to engineering practice. These medically related design problems can range from very complex large-scale constructs, such as the design and implementation of automated clinical laboratories, multiphasic screening

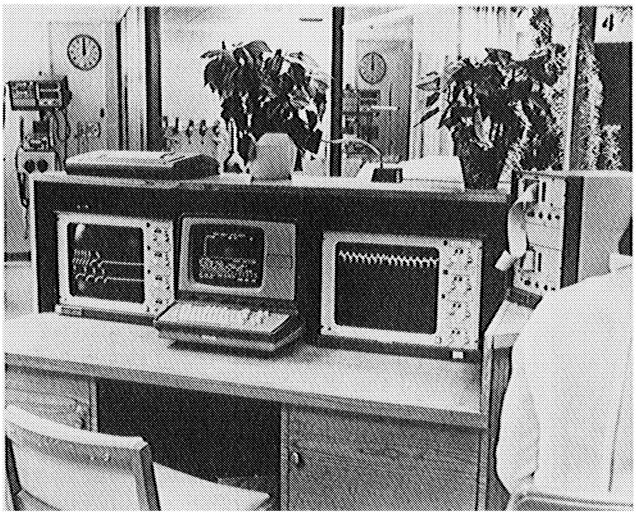

FIG. 2. Computer-based monitoring systems scan the vital signs of patients in hospital intensive care units. (Courtesy of Electronics for Medicine, Inc., White Plains, NY.)

facilities (i.e., centers that permit many clinical tests to be conducted), and hospital information systems, to the creation of relatively small and "simple" devices such as recording electrodes and transducers that may be used to monitor the activity of specific physiological processes in either a research or clinical setting. They encompass the many complexities of remote monitoring and telemetry, including the requirements of emergency vehicles, operating rooms, and intensive care units. The American health care system, therefore, encompasses many problems that represent challenges to certain members of the engineering profession called biomedical engineers.

Since biomedical engineering involves applying the concepts, knowledge, and approaches of virtually all engineering disciplines (e.g., electrical, mechanical, and chemical engineering) to solve specific health-care-related problems, the opportunities for interaction between engineers and health professionals are many and varied.

Biomedical engineers may become involved, for example, in the design of a new medical imaging modality or development of new medical prosthetic devices to aid the handicapped. Although what is included in the field of biomedical engineering is considered by many to be quite clear, many conflicting opinions concerning the field can be traced to disagreements about its definition (Plonsey, 1973; Johns, 1975). For example, consider the terms *biomedical engineering*, *bioengineering*, and *clinical* (or *medical*) *engineering* which are defined in the *Bioengineering Education Directory* (Pacela, 1990). While Pacela defines "bioengineering" as the broad umbrella term used to describe this entire field, bioengineering is usually defined as a basic-research–oriented activity closely related to biotechnology and genetic engineering, that is, the modification of animal or plant cells, or parts of cells, to improve plants or animals or to develop new microorganisms for beneficial ends. In the food industry, for example, this has meant the improvement of strains of yeast for fermentation. In agriculture, bioengineers may be concerned with the improvement of crop yields by treatment of plants with organisms to reduce frost damage. It is clear that bioengineers of the future will have a tremendous impact on the quality of human life; the full potential of this specialty is difficult to imagine. Typical pursuits include the following:

- development of improved species of plants and animals for food production;
- invention of new medical diagnostic tests for diseases;
- production of synthetic vaccines from clone cells;
- bioenvironmental engineering to protect human, animal, and plant life from toxicants and pollutants;
- study of protein-surface interactions;
- modeling of the growth kinetics of yeast and hybridoma cells;

- research in immobilized enzyme technology;
- development of therapeutic proteins and monoclonal antibodies.

The term *biomedical engineering* in this author's opinion has the most comprehensive meaning. Biomedical engineers apply electrical, chemical, optical, mechanical, and other engineering principles to understand, modify, or control biological (i.e., human and animal) systems. When a biomedical engineer works within a hospital or clinic, he or she is more properly called a *clinical engineer.* However, this theoretical distinction is not always observed in practice; many professionals working within U.S. hospitals today continue to be called biomedical engineers. In the field of biomedical engineering there are seven major career areas which include application of engineering system analysis and modeling (computer simulation) to biological problems; measurement or monitoring of physiological signals; diagnostic interpretation via signal processing techniques of bioelectric data; therapeutic and rehabilitation procedures and devices; prosthetic devices for replacement or augmentation of bodily functions; computer analysis of patient related data; and medical imaging, i.e., the graphic display of anatomical detail or physiological function. Typical pursuits include the following:

- design of instrumentation for human physiology research;
- monitoring astronauts and maintenance of life in space;
- research in new materials for implanted artificial organs;
- development of new diagnostic instruments for blood analysis;
- computer modeling of the function of the human heart;
- writing software for analysis of medical research data;
- analysis of medical device hazards for the U.S. government;
- monitoring the physiological functions of animals;
- development of new diagnostic imaging systems;
- design of telemetry systems for patient monitoring;
- design of biomedical sensors for measurement of human physiological systems variables;
- research on artificial intelligence (AI) (*q.v.*) and development of expert systems for diagnosis of diseases;
- design of closed-loop control systems for drug administration;
- modeling of the physiological systems of the human body;
- design of instrumentation for sports medicine;
- development of new dental materials;
- design of computers and communication aids for the handicapped;
- research in pulmonary fluid dynamics (see BIORHEOLOGY);
- study of the biomechanics of the human body.

The above list is not intended to be all-inclusive. There are many other applications that utilize the talents and skills of the biomedical engineer. In fact, the list of activities of biomedical engineers depends upon the medical environment in which they work. This is especially true for the "clinical engineers," biomedical engineers employed in hospitals or clinical settings. Clinical engineers are essentially responsible for all the high-technology instruments and systems used in hospitals today; for the training of medical personnel in equipment use and safety; and for the design, selection, and use of technology to deliver safe and effective health care (Fig. 3).

Hospitals that have established centralized clinical/biomedical engineering departments to meet these responsibilities, therefore, use clinical engineers to provide the hospital administration with an objective opinion of equipment function, purchase, application, overall system analysis, and preventive maintenance policies. With the in-house availability of such talent and expertise, the hospital is in a far better position to make more effective use of its technological resources (Jacobs, 1975; Bronzino, 1977, 1986).

In the past, industry has been reluctant to make large investments in biomedical engineering research and development. The lack of a clearly defined and predictable market for successful technologies is the major culprit here. Competent clinical engineers, however, as part of the health care system, can help create a more unified and predictable market. By providing health professionals

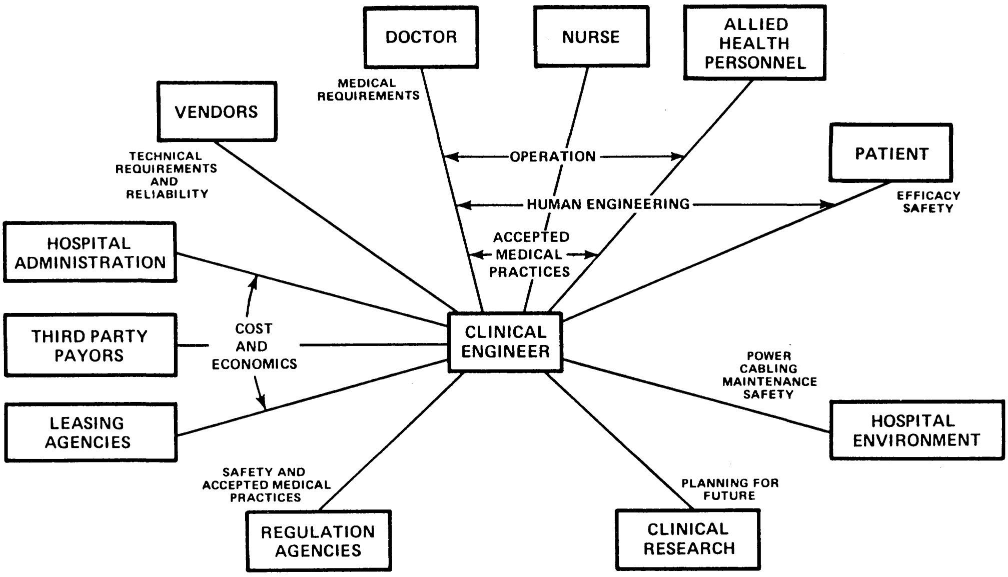
DOCTOR
NURSE
ALLIED HEALTH PERSONNEL
VENDORS
MEDICAL REQUIREMENTS
PATIENT
TECHNICAL REQUIREMENTS AND RELIABILITY
OPERATION
HUMAN ENGINEERING
EFFICACY SAFETY
HOSPITAL ADMINISTRATION
ACCEPTED MEDICAL PRACTICES
THIRD PARTY PAYORS
COST AND ECONOMICS
CLINICAL ENGINEER
POWER CABLING MAINTENANCE SAFETY
LEASING AGENCIES
HOSPITAL ENVIRONMENT
SAFETY AND ACCEPTED MEDICAL PRACTICES
PLANNING FOR FUTURE
REGULATION AGENCIES
CLINICAL RESEARCH

FIG. 3. Block diagram illustrating the wide range of interactions of the clinical engineer.

with needed assurance of safety, reliability, and efficiency of new and innovative equipment, they can, in the process, identify poor-quality and ineffective equipment much more readily. These activities can lead to a faster and broader acceptance of new technology and provide a natural incentive for greater industry involvement, a step that is an essential prerequisite to widespread utilization of any technology (Neuman *et al.*, 1989).

Typical pursuits of clinical engineers, therefore, include the following:

- supervision of a hospital clinical engineering department;
- design, modification, or repair of sophisticated medical instruments or systems within a hospital;
- evaluation of a new noninvasive monitoring system during a clinical research program;
- prepurchase evaluation and planning for new medical technology;
- cost-effective management of a medical equipment repair service;
- safety testing of medical equipment;
- risk management and liability control within the hospital;
- coordination of outside services and vendors;
- supervision of biomedical engineering technicians;
- providing training for medical personnel in the safe and effective use of medical devices and systems;
- custom modification of medical devices for clinical research.

Of all of the definitions discussed above, the term *biomedical engineering* provides the broadest meaning (Fig. 4). Biomedical engineering is an interdisciplinary branch of engineering heavily based both in engineering and in the life sciences. It ranges from theoretical, nonexperimental undertakings to state-of-the-art applications. It can encompass research, development, implementation, and operation. Accordingly, like medical practice itself, it is unlikely that any single person can acquire expertise that encompasses the entire field. As a result, there has been an explosion of biomedical engineering specialists to cover this broad spectrum of activity. Yet because of the interdisciplinary nature of this activity, there is considerable interplay and overlapping of interest and effort between them. For example, biomedical engineers engaged in the development of biosensors may interact with those interested in prosthetic devices to develop a means to detect and use the same bioelectric signal to power a prosthetic device. Those engaged in automating the clinical chemistry laboratory may collaborate with those developing expert systems to assist

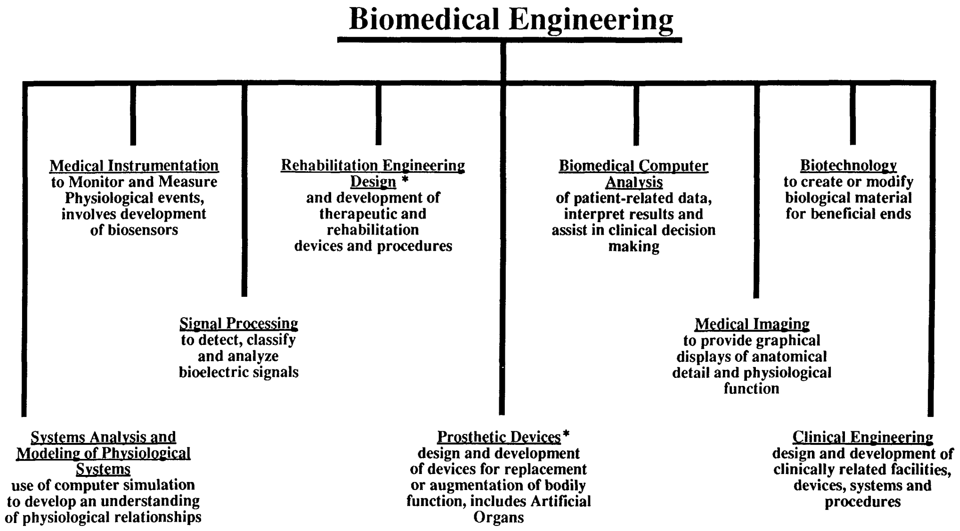
Biomedical Engineering
Medical Instrumentation to Monitor and Measure Physiological events, involves development of biosensors
Rehabilitation Engineering Design * and development of therapeutic and rehabilitation devices and procedures
Biomedical Computer Analysis of patient-related data, interpret results and assist in clinical decision making
Biotechnology to create or modify biological material for beneficial ends
Signal Processing to detect, classify and analyze bioelectric signals
Medical Imaging to provide graphical displays of anatomical detail and physiological function
Systems Analysis and Modeling of Physiological Systems use of computer simulation to develop an understanding of physiological relationships
Prosthetic Devices* design and development of devices for replacement or augmentation of bodily function, includes Artificial Organs
Clinical Engineering design and development of clinically related facilities, devices, systems and procedures

* Both of these activities involve the fields of Biomechanics and Biomaterials

FIG. 4. Schematic diagram illustrating the various fields of activity within the discipline of biomedical engineering.

clinicians in making clinical decisions based upon specific laboratory data. The possibilities are endless.

Perhaps a greater potential benefit occurring from the utilization of biomedical engineers is the identification of problems and needs of our present health care delivery system that can be solved using existing engineering technology and systems methodology. Consequently, the field of biomedical engineering offers hope in the continuing battle to provide high-quality health care at a reasonable cost. If properly directed towards solving problems related to preventative medical approaches, ambulatory care services, and the like, biomedical engineers can provide the tools and techniques to make our health care system more effective and efficient.

3. PRESENT AND FUTURE RESEARCH DIRECTIONS IN BIOMEDICAL ENGINEERING

Biomedical engineering research has taken place in a number of important areas in the basic life sciences and clinical medicine (Table 1). Modifying the categories of biomedical research presented in a recent review by Neuman *et al.* (1989), we have the following:

1. systems science and integrated biology,
2. cardiovascular assist devices,*
3. biomaterials,
4. rehabilitation engineering and prosthetic devices,
5. instrumentation and sensors,
6. medical imaging technology, and
7. biomedical computing.

In keeping with this classification, let us review some of the research activities in each category.

3.1 Systems Science and Integrated Biology

The application of engineering systems analysis and modeling to biological problems is perhaps one of the earliest recognized research areas in biomedical engineering. The quantification of physiological relationships and the building of mathematical models, i.e., simulations, based upon these relationships have not only been central to the field of biomedical engineering, but have been signifi-

---

*Other assist devices such as mechanical drivers of limbs, etc., are covered in Sec. 3.4.

Table 1. Research areas for biomedical engineering.[a]

1. Basic medical research technologies
    A. Structural properties (anatomy)
        1. Component (imaging techniques)
        2. Biomechanics (physical properties)
        3. Chemical composition (microscopic, gross)
    B. Functional anatomy (physiology)
        1. Physical factors
        2. Metabolic mechanisms
        3. Control mechanisms
        4. Integrated functions
    C. Pathophysiology (pathology)
        1. Etiology of disease processes
        2. Sequential course of ailments
        3. Causes of complications
2. Clinical technologies
    A. Clinical diagnosis (applications of basic techniques)
        1. Anatomical alterations (imaging methods)
        2. Functional factors (physical, chemical evidence)
        3. Data acquisition and analysis (computers)
    B. Therapeutic alternatives
        1. Drug delivery systems
        2. Biomaterials
        3. Artificial sensors, tissues, and organs
        4. Surgical support
            a. Anesthesia application and monitoring
            b. Least invasive surgery (e.g., angioplasty)
    C. Disease prevention
        1. Lifestyles (i.e., diet, stress, exercise, etc.)
        2. Environmental aspects
            a. Air
            b. Water
            c. Food
            d. Noise
        3. Well-patient care and monitoring

[a] Taken from Neuman *et al.* (1989).

cant aids to life scientists in their effort to understand basic physiological mechanisms better.

The relevance of simulation to biological research lies in its utility, which suggests to biological researchers the most efficient, meaningful, and creative avenues of investigation. Knowledge about a particular biological system is usually pieced together from facts gathered in a variety of laboratories, using many different experimental procedures in many different environments; in some instances attention is given to microscopic experimentation and in other instances to macroscopic experimentation. It is not surprising then, that comprehensive understanding does not necessarily correlate well with the number of "facts" available about the system being studied.

As the various small bits of information are synthesized into a meaningful composite, the level of complexity of the description becomes greatly increased. The language of mathematics, assuming the data under consideration are quantitative in nature, allows the biological investigator to contemplate complex interrelationships that appear to be beyond comprehension when viewed casually. In turn, modern high-speed computing machinery can manipulate such mathematics often with dramatic effect which improves understanding of the biological system under study. The mathematical description (and any resultant solutions) forms a compact, exact language that can easily be communicated to others.

Successful simulation demands that the experimenter have a certain talent in model building and a grasp of the physical principles that are pertinent to his field of research. It is, therefore, important that a continual interaction take place between experimentation and simulation (Fig. 5 suggests several modes of interaction). The ultimate objective is to increase the knowledge of a particular system. If we start a study by doing simulation there are several pathways that lead to "knowledge." After a simulation is performed, the questions "Has the simulation been a creative experience?" or "Does the solution prove new insight into the system's performance?" must be answered. If the answer is "yes," then appropriate experimentation is in order. If the answer is "no" the simulation activity was not completely successful and alterations in the model are required. In a similar fashion, the next step is either more experimentation or more simulation.

Today, new modeling approaches, such as parallel processing and neural networks, are opening new horizons in enabling life scientists to understand complex cognitive function as well as the interaction of processes occurring at the cellular or even subcellular levels. At the same time further advances in computing technology may allow investigators to reduce the need of using animals in a wide variety of experiments.

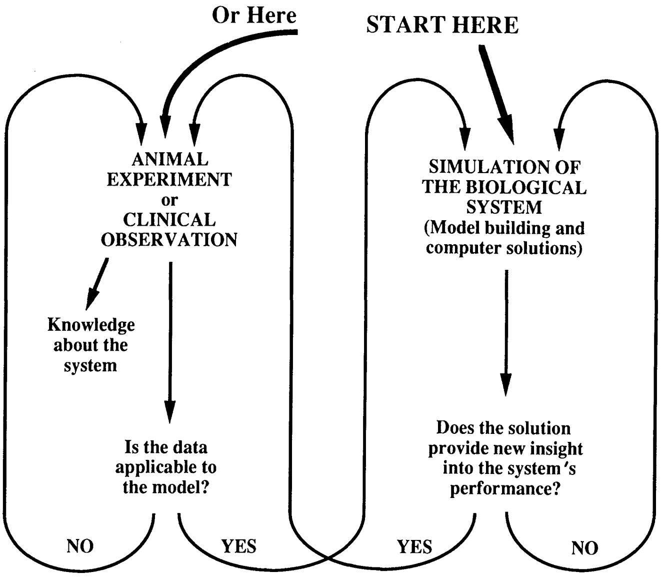
Or Here
START HERE
ANIMAL EXPERIMENT or CLINICAL OBSERVATION
Knowledge about the system
Is the data applicable to the model?
NO
YES
SIMULATION OF THE BIOLOGICAL SYSTEM (Model building and computer solutions)
Does the solution provide new insight into the system's performance?
YES
NO

FIG. 5. Block diagram illustrating the interaction between simulation and experimentation.

3.2 Cardiovascular Assist Devices

Biomedical engineering has played an important role not only in increasing our understanding of the cardiovascular system, but also in the development of diagnostic and therapeutic devices that are routinely used in the care of cardiac patients. Since this is such an important area of research and involves many biomedical engineering disciplines, such as biomaterials and instrumentation, let us explore it in more detail.

3.2.1 Cardiac Pacemakers. The use of artificial cardiac pacemakers is well established in medical practice. Since its development in 1960, over 500 000 pacemakers have been implanted into patients in the United States alone, making this the most common of all cardiovascular therapeutic procedures. The most common indication for a pacemaker is very slow heart rate or bradycardia due to a failure of the natural pacemaker or a component of the conduction system. Patients with bradycardia are prone to loss of consciousness as their hearts cannot speed in response to stress and their brains do not receive adequate blood flow.

Modern pacemakers can sense the heart's own electrical activity and pace appropriately. For patients with sinus atrium (SA) node dysfunction the pacemaker can pace the atrium directly through an atrial electrode and allow the heart's own conduction system to function normally. For patients with a blockage of impulse conduction between the atria and ventricles, a condition called heart block, the heart's own atrial activity can be sensed by the pacemaker and a stimulus sent directly to the ventricle. This type of physiological pacing allows the patient's own SA node to set the rate. Many patients have widespread conduction system disease. At various points in their conduction system, problems may occur and flexibility is needed to control the situation. These patients may initially experience SA node dysfunction but subsequently develop heart block. To deal with this problem of change in the status of the heart, biomedical engineers developed a dual-chamber programmable pacing system. After implantation, the functions of these units can be controlled and reprogrammed by an external device using radio-frequency commands. The functions that can be controlled are the selection of the chambers to be paced, the chambers to be sensed, and the mode of response. Furthermore, all of the above functions and battery status are presently monitored by telemetry.

Typically, a pacemaker unit will be installed in a patient and programmed by a cardiologist (Fig. 6). The status of the unit will be monitored by telemetry, increasingly via the telephone, on a monthly basis, and periodic reprogramming of the pacemaker will take place as needed. Present pacemakers are so

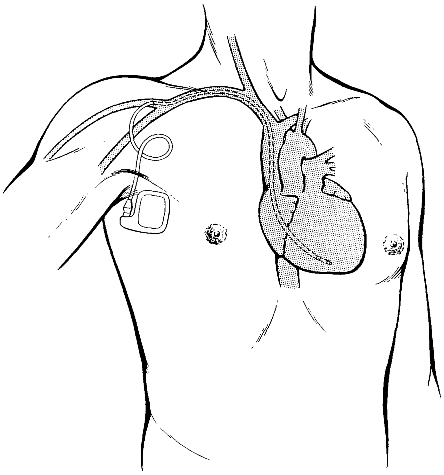

FIG. 6. Sketch illustrating the placement of a cardiac pacemaker. (Courtesy of Medical Systems Division, General Electric Company, Milwaukee, WI.)

durable they even can be reused. In the future, it is almost certain that more-advanced pulse generators will be developed to provide even more flexibility.

3.2.2 Defibrillation Techniques. The normal heart normally contracts in a rhythmic fashion. In the diseased heart, especially in the case of coronary artery disease when the heart muscle does not receive adequate blood flow through the blocked coronary arteries, rhythm disturbances occur. These are usually limited to occasional extra beats called ventricular extrasystoles, but can progress to the potentially lethal rhythm called ventricular fibrillation. In this state, there is no orderly beat and the ventricle is overstimulated electrically to the point where mechanical activity ceases and there is no effective cardiac pumping. Death occurs in a few minutes if the condition persists (Fig. 7).

In 1962, Dr. Bernard Lown of Harvard, a cardiologist, developed the direct current (dc) defibrillator in which a capacitor is charged to the desired dc voltage and then discharged through large metal paddles placed on the patient's chest. This method is still being used today.

Today the defibrillator is carried as part of the ambulance equipment in most areas. The paramedical team, which arrives at the home

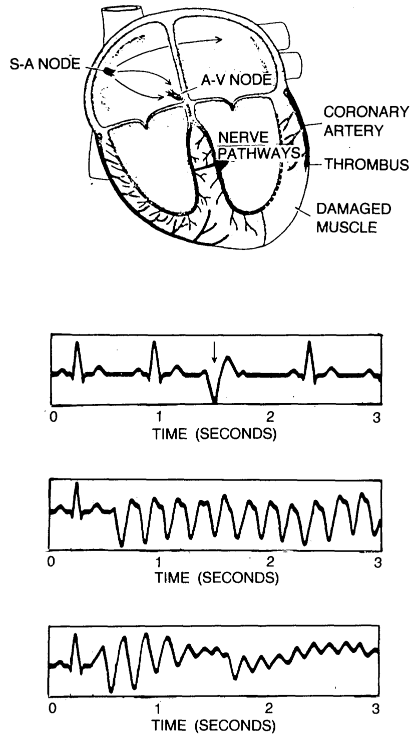

ELECTRICAL ACTIVITY OF DAMAGED HEART is erratic because nerve impulses can no longer flow smoothly from the atrioventricular node through the pathways that feed impulses to the ventricular muscles. Most heart-attack patients develop one or more of the arrhythmias shown below.

FIRST SIGN OF ARRHYTHMIA in the heartbeat, following a coronary heart attack, usually takes the form of a premature ventricular beat (arrow).

MORE SERIOUS ARRHYTHMIA is tachycardia, or fast heartbeat, in which ventricular impulses occur at two or three times the normal rate. If not halted, they can cause death.

FATAL ARRHYTHMIA known as fibrillation can develop from tachycardia or when premature beats fall within a critical part of the T wave.

FIG. 7. Electrical activity of damaged heart. Redrawn from Lown (1968).

of a patient who is thought to be suffering from a heart attack, promptly attaches electrocardiogram (ECG) leads to the patient, observes the ECG, and sends a copy of the signal by radio to the physician at the hospital base. If ventricular fibrillation is seen, cardiopulmonary resuscitation is started and, by remote direction of the physician, the defibrillator paddles are placed on the patient's chest. Ventricular fibrillation is again confirmed by analysis of the ECG signal sensed by the paddles.

One paddle is placed over the upper chest and the other over the lower chest. The discharge switches are located in the handles of the paddles and both must be pressed at the same time to allow the defibrillator to discharge. Within seconds, the ECG trace hopefully reveals a return to a normal rhythm. The patient can then be safely transported to the hospital for further care. This system of portable monitoring, telemetry, and defibrillation is a major reason that patients who have heart attacks now survive in greater numbers.

At present there also exist automatic, implantable defibrillators. In patients with very unstable rhythm which is only partially controlled by medication, this type of device can be used to allow them to leave the hospital and resume a more normal lifestyle. This new electronic device can even be programmed to monitor cardiac electrical activity and to recognize ventricular fibrillation. If ventricular fibrillation is sensed, an internal discharge of the device can be triggered.

3.2.3 Artificial Heart and Cardiac Assist Devices. It has always been clear to many researchers that many patients with heart disease could not always be aided by medication or surgical repair of their damaged hearts. The heart valves could be replaced with artificial devices but the heart muscle or myocardium was often so damaged that it required replacement or the patient would die. As a result, two separate avenues of research to the problem were begun.

First, research into cardiac transplantation was initiated. Kidney and cornea transplants from other humans had proven so successful that cardiac transplantation was studied. Although many problems such as rejection, donor organ procurement, and infection remain, considerable progress has been made and this approach has now become an accepted part of cardiac therapy.

The second area of research in the area of cardiac replacement was the development of the artificial heart. This effort was subdivided into two areas: the left ventricular assist device and total cardiac replacement (Bernard and LaFarge, 1969). Both approaches require close cooperation between the biomedical engineer and the surgeon. The relative successes that have been achieved to date are testament to the accomplishments possible from these joint ventures.

The first temporary use of a mechanical device to sustain life while the heart was unable to function was the pioneering effort of J. H. Gibbon, Jr. and C. W. Lillehei at the University of Minnesota (Allen and Lillehei, 1957). In 1953 their group began the use of the techniques of cardiopulmonary bypass, enabling the current practice of open-heart surgery to evolve. The device they developed consisted of a pump to maintain blood flow and an oxygenator to permit gas exchange when the heart and lungs were removed from the path of circulation.

The modern heart-lung machine typically consists of a double-roller pump which propels the blood forward by a squeezing action and either a bubble or membrane oxygenator. The roller pump is quite effective as a generator of flow, but prolonged use causes mechanical damage to the red blood cells (hemolysis). This damage can be tolerated for the short periods (several hours) which are typical in open-heart surgery but not for prolonged periods as the damaged red cells release hemoglobin into the bloodstream which eventually causes kidney failure. To permit support for longer periods of time, pumps that compress the blood between the walls of flexible chambers in a way similar to the natural process are more desirable. As new types of pumps are developed they are being utilized.

It is important to note that the materials in these devices are sensed as foreign to the body. While they are not rejected as unacceptable biological tissue (a problem that occurs in the case of heart transplants), the body does attempt to form clots on their surface. Such clotting is to be avoided since it could cause the mechanism to fail. Patients with these devices must therefore be given drugs called anticoagulants to prevent clotting, and the use of anticoagulants has to continue as long as the mechanical device is sustaining circulation. Care must be taken, however, to avoid overuse of anticoagulants since spontaneous bleeding in such vital organs as the brain or the gastrointestinal tract can take place.

Both the artificial heart and the left ventricular assist devices are extensions of earlier efforts to develop the heart-lung machine for open-heart surgery. In many patients, the left ventricle is damaged during an episode of coronary artery blockage. The left ventricle, which is the pump for the entire body, can be revascularized by coronary artery bypass, but often requires time to heal. As the body must be served by the ventricle immediately after surgery, there can be no rest period. What many patients require, therefore, is temporary assistance for the left ventricular function. The development of a left ventricular assist device (LVAD) was led by Dr. John C. Norman and colleagues (Norman and Huffman, 1974; Norman, 1975a, 1975b). Such assist devices are designed to provide partial or total support of the circulation.

The second and more highly publicized effort has been in the development and use of the *total artificial heart*. These devices are designed to replace completely the damaged heart and to support the circulation on a permanent basis. Unlike the LVAD where the patient's own heart can recover and the support device can be removed, the artificial heart is implanted in place of the patient's own heart, which is then discarded.

In 1963, the National Advisory Heart Council recommended giving priority to this type of research, as well as the necessary federal funding, focusing on such areas of research as biomaterials, pumping mechanisms, and control systems.

In 1969, sufficient progress had been made to permit Dr. Denton Cooley of the Texas Heart Institute in Houston to implant an artificial heart in a patient awaiting a heart transplantation whose own heart had failed. The device sustained the patient for 64 h until a suitable donor heart was found and the heart transplant took place. A similar type of temporary artificial heart was again used in a similar circumstance by Dr. Cooley in 1981 to sustain a patient for 54 h. In both these cases, the use of the artificial heart was described as temporary, i.e., providing assistance until a heart transplant could be performed. The work of Dr. William C. DeVries and Dr. Robert K. Jarvik, supported initially by Dr. W. J. Kolff at the University of Utah, and initially and subsequently by the Humana Heart Institute in Louisville, Kentucky, has led to the use of a *permanent artificial heart* to sustain a patient (DeVries *et al.*, 1984; DeVries, 1986).

On 2 December 1987 Dr. Barney Clark became the first legal recipient of a permanent artificial heart—the Jarvik-7—after the U.S. Food and Drug Administration approved the device in 1982 (Fig. 8). Dr. Clark suffered from cardiomyopathy, a disease that causes the heart muscle to deteriorate and has no known cure. He showed no chance of recovery through any heart surgery and was very close to death—conditions required by the FDA before approval to implant the artificial heart could be given. At 61, he was considered too old to receive a transplanted heart, even if a donor had been available.

The Federal Food and Drug Administration has given the University of Utah Medical Center permission to implant the Jarvik-7 heart and its driving system in seven patients. However, while the implantation of the Jarvik-7 heart continues, the design team of

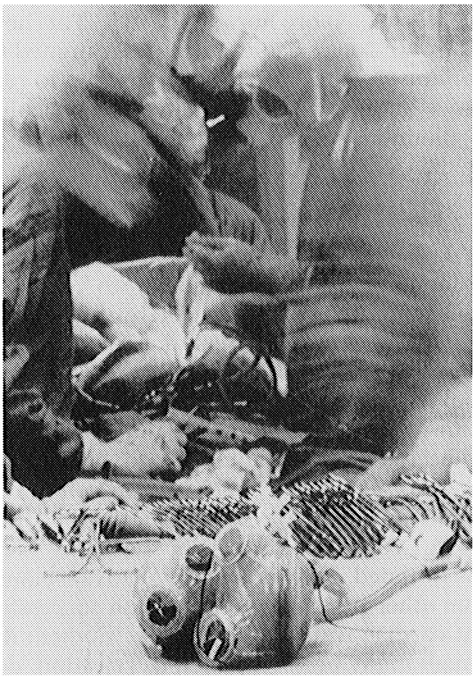

FIG. 8. The Jarvik heart being implanted.

Symbion, the firm in Salt Lake City headed by Dr. Robert Jarvik, works on the successor model, the Jarvik-8. The major problems of the Jarvik-7—clotting, size, durability, and "mass" production—are being addressed in the redesign.

Both the total artificial heart and the LVAD face similar engineering problems (Galioto, 1986). Materials for their construction must be durable as they are truly life sustaining. Their mechanisms must safely pump blood but should not damage the delicate cellular elements of the red blood cells. Blood can be protected by nonocclusive pumps and anticoagulants but the patient's own long-term responses to such devices must be borne in mind at all times. Most important, these devices should allow for a reasonable and tolerable lifestyle.

3.3 Biomaterials

The use of surgical implants and invasive medical therapies has expanded greatly in recent decades. Devices ranging in complexity from chronic infusion catheters and lens implants to total hip prostheses, to implantable pacemakers/defibrillators, heart-lung machines, and hemodialyzer systems are now routinely used. These devices are employed on a vast scale: tens to hundreds of thousands are used each year in the United States alone. On the medical horizon are numerous devices (such as the artificial heart discussed above) of increasing sophistication and complexity. Common to all these devices is the use of biomaterials.

What are these biomaterials and how do they differ from other engineering materials? They are metals, plastics, and ceramics, some of which are merely purified forms of industrial materials. Some are new formulations with specific properties found to be advantageous in medical applications. Others are surface modifications of, or bulk additives to, conventional materials. Still others are processed natural (biological) materials (Eberhart, 1989).

Research efforts in biomaterials include understanding the biological and physical properties of naturally occurring materials, such as bone, connective tissue, keratin, etc., and creating new artificial materials or generating "natural" materials with similar properties to their natural counterparts. Since biomaterials are used to replace or augment natural tissues, it is also important to understand and control the biology of the interface between artificial and natural biomaterials in the body.

There are three main classifications of artificial biomaterials: metals, ceramics, and polymers. Each class has specific properties that make it especially suitable for some function in the body, but in many cases it cannot perform today as well as its natural counterpart. It is hoped that through research, new and improved materials can be developed that meet or even exceed the capabilities of the natural substances (Neuman *et al.*, 1989).

3.3.1 Metals. A major component of bone and joint prostheses are metals, used because of their strength and ability to absorb energy before fracture. Table 2 presents properties of some commonly used biomaterial metals. Their major drawback, however, is associated with the propensity of the elements responsible for their resistance to corrosion (chromium, nickel, cobalt) to induce a cell-mediated hypersensitivity. Today, work is under way to utilize titanium-based alloys because there has yet to be any record of hypersensitivity response from these materials.

3.3.2 Ceramics. The class of materials called "ceramic" includes a variety of substances that are nonmetallic and inorganic in nature, and generally consist of clay or other silicates. Although ceramics are the oldest manmade materials and plaster of Paris ($CaSO_4$) has been used for casts in fracture immobilization for many years, the use of ceramics for biomedical applications only started in the mid-1960s (Saha, 1989).

Compared to the large number of different ceramic materials that are used for industrial purposes, only a few classes of these materials have been found suitable for implantation and clinical use, in spite of the fact that they are highly inert and elicit a minimum foreign-body reaction. These classes include

1. oxide ceramics and ceramics of calcium phosphate salts,
2. various types of glass or enamel, and
3. crystalline or glassy forms of carbon and its compounds (Heimke, 1986).

The oxide ceramics are known by the metals from which they are formed. Alumina

Table 2. Properties of commonly used biomedical metals.[a]

	Stainless steel No. 316L	Titanium	MP35 N	Elgiloy	Platinum	Nitinol/Tinel	
Composition	Fe, Cr, Ni, Mo, Si, Mn	Ti	Co 35%, Ni 35%, Cr 20%, Mn 10%	Co 40%, Cr 20%, Fe 16%, Ni 15%, Mo 7%, Mn 2%, Co 0.001%, Be 0.04%	Pt	50 Ni, 50 Ti	
Application	Implant case; electrode lead	Implant case	Electrode	Electrode	Electrode	Special (shape memory)	
Density (g/cm^3)	7.9	4.5	8.3	8.2	21.4	6.5	
Electrical resistivity ($\mu\Omega$ cm)	74	48–60	101	99.5	9.8	100 80	(high-temp.) (low-temp.)
Thermal conductivity (W/cm °C)	0.15	0.16	...	0.13	0.73	0.18 0.086	(high-temp.) (low-temp.)
Specific heat (cal/g °C)	0.5	0.52	...	...	0.12	0.84	
Thermal expansion coefficient (10^{-6}/°C)	16	9.2–9.9	15.7	15.8	8.8	11 8.6	(high-temp.) (low-temp.)

---

[a] From Tarjan and Gold (1988).

(Al_2O_3) is the most commonly used and most biocompatible ceramic material for orthopedic implants. Other oxides of calcium and phosphorus, such as hydroxyapatite [$Ca_{10}(PO_4)_6(OH)_2$] and hydroxylapatite [$Ca_4H(PO_4)_3OH$] have also been used. Among these, hydroxyapatite (HA) is of particular interest as its chemical composition is very similar to that of the mineral content of bone.

The glasses that have been studied for implantation are based on silica (SiO_2) and sometimes contain a small amount of other crystalline phases. Bioglass, developed by Hench and co-workers (1988), is perhaps the most common among these types of ceramics.

The carbons that are primarily used in prosthetic devices include low-temperature isotropic (LTI) pyrolitic carbon (Pyrolite), biotite carbon, and carbon or graphite fibers. LTI pyrolitic carbon has been used in the DeBakey heart valve since 1969 and has also been used to make coatings for other valves, vascular access devices, and dental implants. Carbon fibers are being used to replace or augment ligaments and tendons. Carbon-fiber-based composites are also being developed to manufacture orthopedic implants.

During the last twenty years, ceramics have also been used as a material for total joint-replacement implants as well as a coating on metal implants for improved biological fixation with the host bone (Christel, Meunier, and Lu, 1986).

Total joint replacement is one of the great orthopedic advances of this century. Sir John Charnley introduced the metal-on-plastic total hip replacement in the mid-1950s, and since then this orthopedic procedure has revolutionized the treatment of painful arthritic hip disorders. This procedure is highly successful with rates of success between 80% and 98%, depending upon the type of implant, surgical technique, patient selection, and, particularly, follow-up time. In long-term follow-up studies, it has been found that aseptic loosening is the significant cause of failure of total hip replacement. This failure rate is estimated to be 7%–24% for a 5–10 yr follow-up (Russotti, Coventry, and Stouffer, 1988). The enormity of the problem is evident from the fact that in the United States alone, over 120 000 total hip replacement arthroplasties are performed each year. Worldwide, over 2.5 million people are estimated to have artificial hip joints.

To improve the long-term results of total hip replacement (THR), a large number of new prosthesis designs have been proposed. These include use of a low-modulus metal such as titanium alloy to reduce the stress shearing effect on bone. Ceramics have also been used to make the whole or parts of a THR and in coating on metal implants to obtain a chemical bond between the implant and surrounding bone.

In the future, it is anticipated that ceramics will become a major component for hermetic sealing of implantable microelectronic devices and sensors because of a combination of their properties. Of major importance is their transparency to electromagnetic radiation, as well as their impermeability to water. It has been suggested that sealing processes using laser technology will solve the problem of packaging implantable electronic devices without damage and that modification of the surface charge on such ceramic packages will help to control their thrombogenicity, thus allowing their use in packages for sensing applications in blood.

3.3.3 Polymers. Polyurethanes are generally formed by the addition of a polyglycol to an isocyanate. Typically, the isocyanate is part of an aromatic molecule that gives rigidity or hardness to this portion of the polymer chain, while the polyglycol is the part of an aliphatic molecule that produces a less rigid or soft portion of the polymer chain. Most polyurethanes, therefore, have hard and soft segments and are called segmented polyurethanes (McMillin, 1989).

Polyurethanes can be readily tailored for many biomedical applications by either changing the chemical used in any of the components or by changing the length of the prepolymer or extended chains of the polyurethane. Thus, polyurethanes are in themselves a wide class of materials. Ether-based polyurethanes are known to be less susceptible to hydrolytic cleavage than ester-based polyurethanes and have therefore been used for most biomedical applications. Polyurethanes are usually strong, flexible, transparent, or translucent, tissue compatible, and reasonably blood compatible, and are used for a variety of applications requiring these properties, such as peristaltic pump tubing, parenteral solution tubing, balloon pumps, and catheters.

Much of the research on biomedical polyurethanes focuses on a relatively few commercial polymers. For competitive reasons, the chemical composition of these polyurethanes are sometimes trade secrets. The known compositions of some urethanes are shown in Table 3.

Polymers of biological origin such as collagen and elastin appear to be a logical solution for implant applications. The major problem is how to remove their antigenic properties while maintaining their desirable properties. It is important to make progress in using materials of biological origin as scaffolds for the reconstruction of certain tissues. Probably the major advance in this area will be in the use of synthetic polymers as scaffolding structures that will degrade, and therefore be repopulated by body cells, to construct the appropriate collagen network to give structural integrity to the system as a whole.

A major limitation in the use of synthetic polymers as biological materials is their tendency to fatigue. More fatigue-resistant polymers with isotropic mechanical properties need to be developed. These materials are needed as the basis of the diaphragms for pumps in total artificial heart and heart assist systems, as well as polymeric valves that will replace the artificial heart valves used today.

3.4 Rehabilitation Engineering and Prosthetic Devices

Engineering devices have been applied to helping disabled individuals return to more useful lives. The design of prosthetic devices and artificial organs frequently involves mechanical principles, and therefore these areas lie partly in the province of biomechanics.

Biomechanics has traditionally been concerned with the analysis of macroscopic biological structures such as bones, blood vessels, other soft tissues, and blood and other biological fluids. In the past, computer modeling techniques have been applied to help understand not only normal function but also the pathophysiology of various diseases and conditions. For example, computer simulations have provided great insights into the mechanics of lifting as it applies to lower back pain, optimization of athletic performance, mechanics of bone healing, and transport processes in various organ systems. Biomechanics has also led to the development of procedures in diagnostic medicine, such as gait analysis, and the design of various prosthetic devices ranging from joint replacements to artificial organs. This field has contributed to the design and optimization of

Table 3. Composition of some polyurethane biomedical elastomers. Modified from McMillan (1989).

	Urethane components[a]						
Trade name/supplier	PTMEG	MDI	ED	BD	PD	HMDI	PDMS
Biomer/Ethicon Inc., Somerville, NJ	×	×	×				
Surethane/Cardiac Control Systems, Palm Coast, FL	×	×	×				×
Pellethane/Dow Chemical Co., Midland, MI	×	×		×			
Corplex/Corvita, Miami, FL	×	×		×			(×) post treatment
Toyobo TM5/Toyobo Co., Osaka, Japan	×	×			×		
Tecoflex HR/Thermedics, Inc., Woburn, MA	×			×		×	
Cardiothane-51/Kontron Institute, Everett, MA	×	×		×			×

[a] PTMEG, polytetramethylene ether glycol; MDI, methylene diisocyanate; ED, ethylenediamine; BD, 1,4-butanediol; PD, propylenediamine; HMDI, hydrogenated MDI; PDMS, polydimethylsiloxane.

man-machine interfaces, such as improved sports equipment, safety of automobiles and aircraft, and specialized equipment for space exploration.

As physiologists and other biological scientists increasingly turn their attention to studies at the cellular and molecular levels, it becomes clear that mechanics has important applications in these areas as well. Atherosclerosis remains a major clinical problem, and only preliminary exploration of the effects of blood fluid mechanics on endothelial cells has been done. Other cellular and molecular aspects of biomechanics remain relatively unexplored. There are many phenomena at the microscopic level in which mechanics is likely to play a role. Among these are

1. atrophy and hypertrophy of muscle,
2. regulation of blood pressure,
3. the origin of atherosclerosis,
4. microscopic processes in bone healing, and
5. bone remodeling and osteoporosis.

The roots of rehabilitation originate in the historical use of prosthetic devices (wooden limb substitutes) and orthotic devices (crutch-like aids) (see PROSTHETICS AND ARTIFICIAL ORGANS). In modern times, visual orthoses (eyeglasses) have become ubiquitous. Technological advances in medical treatment were developed during World War II at the same time as technological applications to weaponry leaped forward. Thus, society believed that the people maimed by the war and saved by medicine had to be helped to lead fuller lives.

After the war, the Veterans Administration (VA) was founded and began to apply engineering solutions to some of the physically disabling problems which patients in VA hospitals presented—principally in the areas of wheelchair development and prosthetic limbs. Around 1970, several *Rehabilitative Engineering Centers* were funded by the U.S. government to increase the application of engineering techniques in the solution of problems related to rehabilitation. However, it was not until about 1980 that the idea developed, in major centers of rehabilitation, that engineers could be useful in direct service to patients along with the more usual deliverers of patient care.

Recently, the idea that people with technical training in engineering can help in the patient care process has spread outside of the major rehabilitation hospitals to smaller hospitals with rehabilitation departments. The Rehabilitative Engineer, trained in engineering and biology, has a capacity to act as the interface between the patient, the rehabilitative team, and technology.

Although many orthotic and prosthetic devices have been developed to help individuals regain some form of functionality, many opportunities remain for future research. In the future, rehabilitation research will probably be more in the direction of utilizing existing but nonfunctional biological structures, such as a paralyzed limb, to regain function rather than by employing external orthoses. Functional electrical stimulation of intact nerves and muscles will allow patients to control contraction of paralyzed muscles. To control such motor prostheses effectively, it will be necessary to obtain a better understanding of the body's natural control systems and to determine what constitutes minimal processing to regain particular functions. Sensory input to such systems will be important, and research into making use of intact but disconnected natural sensors, as well as the development of implantable artificial sensors, will continue to be important research areas.

The computer has already played an important role in helping individuals with disabilities to be viably employed and to perform tasks that were once thought to be impossible. The interface between the patient and the computer, however, still remains crude. Research is needed to develop more effective interfaces, as well as processing algorithms for these interfaces. Advances in the neural sciences should help to establish connections between a patient's nervous system and external electronic devices. By understanding natural signal processing in the nervous system, it should be possible for computers to emulate this human function and to control both internal body structures and external devices.

3.5 Instrumentation and Sensors

Instrumentation is involved throughout the field of biomedical engineering. This certainly becomes clear when one reviews the number and diversity of biomedical quantities measured by biomedical instruments in Table 4. In practice, however, the transducers which convert signals in one form of energy into signals in another form of energy are incorporated in an instrumentation package or system. The most general instrumentation system then consists of the transducer or biosensor, a signal processing component which amplifies and filters the signal and performs some preprocessing to permit the measured quantity to be stored, to be displayed, to activate some device, or to be processed further.

The biosensor, therefore, serves as the interface between the biological system being measured and the electronic system of the in-

Table 4. Biomedical electronic measurements.[a]

Measured quantity	Principle or devices used	Biomedical applications
	A. Physical parameters	
Electrical potential	Voltage	Nerve and muscle activity, ECG, EEG, EMG, EOG, etc.
Impedance	*Z* bridge	Skin condition, respiration rate, peripheral blood flow
Temperature	Thermistor, *p-n* junction, infrared	Inflammation, cancer location, blocked circulation
Sound	Microphones	Heart and breathing sounds, blood pressure measurement, fluid flow
Light	Transmission, refraction, absorption	Physiology research, clinical laboratory, gas and ion sensors
Magnet	Permeability, field	Blood flow, heart activity, proximity
Force and pressure	Piezoresistance, capacitance deformation	Pressure of blood, CSF, urine, etc.; locomotion
Displacement, velocity, acceleration	Capacitive, magnetic, laser, ultrasound, acceleration force	Body movement, physiology research, orthopedics
Flow: blood, fluid, air	Electromagnetic, optical ultrasonic Doppler, pressure drop, streaming potential	Blood, air, body fluid; heart, lung, kidney function
Volume	Displaced volume, dye dilution	Physiology research, cardiac output
Time	Delay, response time	Alertness, nerve signal transmission, signal transmission in heart and other organs
	B. Chemical parameters	
Gas concentration	Gas-sensitive electrodes, chromatography, chemical analysis	Blood gas (H^+, O_2, CO_2, O_2 saturation), respiration, perfusion, anesthesia, hospital operations, clinical laboratory
Humidity	Dew-point temperature, resistance and capacitance change due to condensation or absorption	Pediatric, respiratory research, intensive care
Ions	Ion-selective electrodes, ISFET, chemical analysis	Clinical laboratory, intensive care unit, acid-base balance
Biomolecules	Enzyme electrodes, chemical analysis	Clinical laboratory, research laboratory, patient monitoring

---

[a] Modified from Ko (1988).

strumentation. Work during this century has been devoted to developing new types of sensors of physical and chemical variables and incorporating these sensors into instrumentation systems. The application of technologies such as microelectronics and optoelectronics to biomedical sensor problems has shown that imaginative solutions are possible. For example, a few years ago transducers for measuring biomedical pressures were based on handmade, unbonded strain gauge technology. These devices were expensive and had to be individually calibrated for the electronic system with which they were used. Today, semiconductor electronics allows mass produced biomedical pressure transducers to be sold at less than 5% of the price of the earlier devices. This makes it possible to have disposable pressure sensors for clinical applications at a cost that would have been impossible in the past (Neuman *et al.*, 1989).

Future research in sensors will certainly follow corresponding improvements in technology, as well as the feasibility of developing bioanalytical sensors for complex biochemical substances that are found in the body in only very small concentrations. Substances such as neurotransmitters, tumor markers, and local and systemic hormones are important analytes for future biomedical sensors. The availability of such sensors would make it possible to measure these biochemicals directly and continuously within the body itself.

It is important to note that creative ways to present and summarize large amounts of patient-related data for busy physicians and nurses need to be incorporated into modern instrumentation systems. Furthermore, providing instrumentation packages to monitor patients in their homes will become more and more important as health care becomes decentralized and the need for larger data samples becomes important. Coupled with these efforts is the need to develop communication systems to tie such monitoring devices into central facilities.

3.6 Medical Imaging Technology

Developing noninvasive tools that can "look into" the human body while posing minimal risks for the patient has been given an extremely high priority in the post 19th century scientific era of medical research and practice (see IMAGING TECHNIQUES, BIOMEDICAL). Diagnostic and therapeutic techniques that remove the need for invasive (surgical) exploration often both reduce patient risk and lower the costs of health care. However, until recently the health care profession "black bag" has contained only a relatively small number of noninvasive diagnostic tools that provided images of the body. None were available until the end of the 19th century.

We now know that x rays

1. are part of the electromagnetic spectrum, that have relatively short wavelengths,
2. are generated by collision of atomic particles,
3. travel in straight lines in all directions from the point of origin, and
4. are capable of penetrating many forms of matter opaque to ordinary light.

Therefore, the classical medical imaging technique, radiology, uses high-energy electromagnetic waves or x rays as a tool to "look" into the patient to view certain anatomical structures. During the early part of the 20th century additional improvements in the field of radiography involved advances in technology related to particular components of the system including intensifying screens and rotating anode tubes.

It is interesting to note that after 1930, profound improvements in radiology were achieved by developing procedures for selectively opacifying regions of interest in the body. These developments were achieved through a variety of approaches, such as the injection of dyes, often invasive to the body, to facilitate visual representation of otherwise invisible organs.

Since 1960, through the marriage of computers and sophisticated imaging processing techniques, medical imagery has been able to pursue totally new directions. Images of the internal structures of the body are now created that are often of startling clarity, significantly enhancing the ability of the clinician to arrive at an appropriate diagnosis.

Although modern medical imaging takes many forms, systems of most practical interest today include the following: nuclear medicine, diagnostic ultrasound, computerized tomography, and magnetic-resonance imaging. These imaging devices enable doctors to avoid exploratory surgery and still watch vital organs at work, identify blockages and growths, and detect the warning signs of possible future disorders. Since these techniques, for the most part, do not involve penetrating the skin, they have been classified as noninvasive. However, patients may face other consequences of their use (e.g., the possible destruction of body tissue caused by the radioactive materials used in nuclear medicine, or the increase in tissue temperature as a result of high-frequency ultrasound). Bearing this in mind, it may be more precise to consider the relative degree of direct surgical intervention that certain devices entail. Thus the term noninvasive will be used here to refer only to specific techniques developed in radiology, nuclear medicine, and ultrasound.

Major advances in medical diagnostic procedures were made during the 1980s that greatly enhanced image quality and expanded the types of tissues that could be visualized with good resolution. Even with this advanced technology and the integration of computer and image processing techniques into medical imaging devices, problems and opportunities remain for biomedical engineering researchers. These include instrumentation not only for visualizing internal structures but for assessing their function.

3.6.1 Radiography. Radiography, the classical medical imaging technique routinely applied to nearly all hospitalized patients whether or not there is a direct need, uses high-energy electromagnetic waves or photons (x rays), as a tool to "look" through the patient to obtain information on anatomical structures (Fig. 9). The propagation of x rays, characterized by wavelengths that are shorter than the atomic diameters of the penetrated matter, is dominated by corpuscular ray properties rather than by the wave character of light. The velocity of propagation is essentially constant for all kinds of biological tissue; thus the refractive index of all such structures is unity. As a consequence, the only mechanisms of interaction are absorption and scattering; no reflection or diffraction takes place. So, unlike ultrasonic waves, x rays travel in straight lines through the body, unaffected by tissue interfaces, and provide undistorted images of high resolution (see BIOMEDICAL USES OF RADIATION).

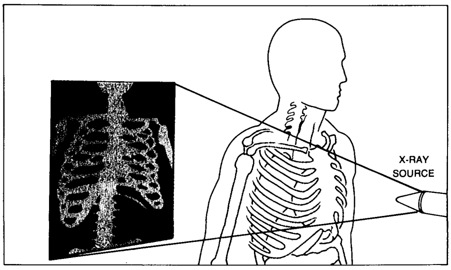

CONVENTIONAL X-RAY PICTURE is made by allowing the X rays to diverge from a source, pass through the body of the subject and then fall on a sheet of photographic film.

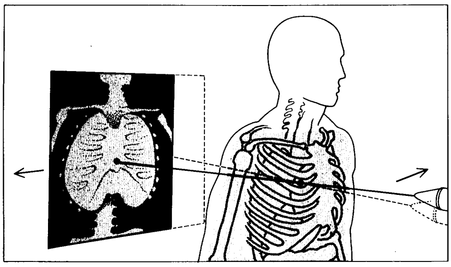

TOMOGRAM is made by having the X-ray source move in one direction during the exposure and film in the other direction. In projected image only one plane in body remains stationary with respect to moving film. In the picture all other planes in body are blurred.

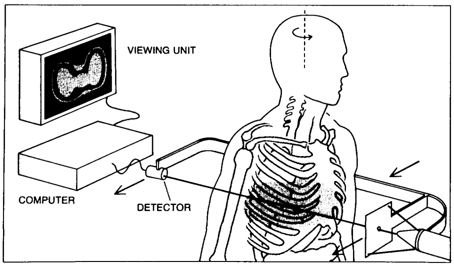

RECONSTRUCTION FROM PROJECTIONS is made by mounting the X-ray source and an X-ray detector on a yoke and moving them past the body. The yoke is also rotated through a series of angles around the body. Data recorded by detector are processed by a special computer algorithm, or program. Computer generates picture on a cathode-ray screen.

FIG. 9. Image reconstruction from x-ray projections.

The term radiography is sometimes misleading because it applies to various x-ray techniques. Its original definition referred only to the recording of x-ray projection images on photographic film. Use of a fluoroscopic screen as the x-ray detector (instead of film), called fluoroscopy, provided the ability for continuous x-ray monitoring. The use of TV systems for electronically monitoring the fluoroscopic screen is called x-ray television. Because all of these systems, including film recording, use fluoroscopic screens as image converters, and all of them are based on the same physical principle of imaging (i.e., the projection of body absorption onto a single plane), the terminology is illogical and arbitrary. So, with the development of an imaging principle called computed tomography (CT) (Ledley, 1976), a renaming took place, in which all possible imaging techniques using x rays are called radiography (Nagel, 1988).

Since its introduction, computed tomography has undergone continuous refinement and development. Significant improvements have been made in

1. x-ray tube performance,
2. reduction of scan times,
3. reduction of image reconstruction time,
4. reduction and, in some cases, elimination of image artifacts,
5. reduction of patient dose.

Present CT scanners are usually integrated units consisting of three major elements (Fig. 10):

1. *the scanning gantry*, which takes the readings in a suitable form and quantity for a picture to be reconstructed;
2. *the data handling unit*, which converts these readings into intelligible picture information, presents this picture information in a visual format, and provides various manipulative aids to enhance the image and thereby assist the doctor in forming a diagnosis, and
3. *a storage facility*, which enables the information to be examined or reexamined at any time subsequent to the actual scan (Walmsley, 1979).

The objective of the scanning system is to obtain enough information to reconstruct an image of the cross section of interest. All the information obtained must be relevant and accurate, and there must be enough independent readings to reconstruct a picture with sufficient spatial resolution and density discrimination to permit an accurate diagnosis. The operation of the scanning system is therefore extremely important.

During the past decade, the CT scanner has undergone several design changes. The earliest generation of gantries used a system known as "traverse and index." In these systems, the tube and detector were mounted on a frame and a single beam of x rays traversed the slice linearly, thereby providing absorption measurements along one profile. At the end of the traverse, the frame indexed through 1° and the traverse was repeated. This procedure continued until 180 single traverses were made and 180 profiles were measured. Using the "traverse and index" approach, the entire scanning procedure took approximately 4–5 min. The images it provided had excellent picture quality and hence high diagnostic utility. However, this approach had several disadvantages. First of all it was relatively slow, resulting in relatively low patient throughput. Second, streak arti-

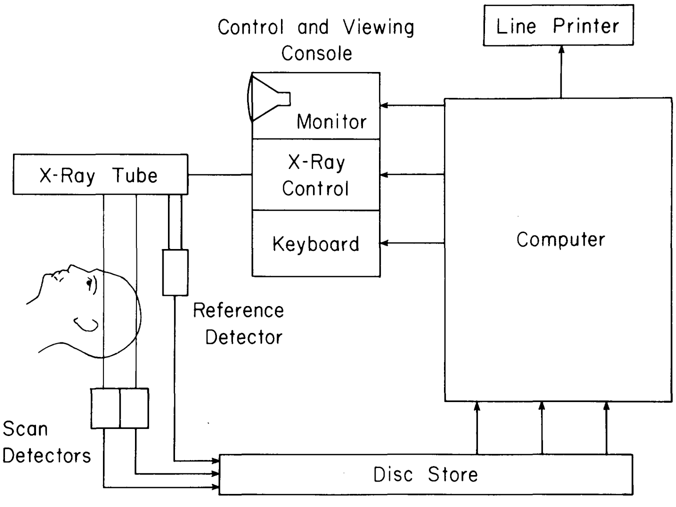

FIG. 10. Basic components of a CT scanner.

facts, i.e., lines in the image, were common, and although these artifacts normally were not diagnostically confusing, the image quality did suffer. Finally, perhaps the biggest drawback was that since the patient had to remain absolutely still during the scan, its use was restricted solely to brain scans.

In an attempt to solve these problems, several generations of gantries have evolved over the past 15 years (Fig. 11). Initially new scanning procedures included a translation step as in the first generation but permitted the scanning gantry to go around the patient (180°) in 10° steps. With use of this 10° fan beam, it is possible to obtain a full set of 180° profiles by making only 18 traverses. At the rate of approximately 1 s for each traverse, scanning systems of this type usually operate in the range of 18–20 s while still retaining picture quality.

However, very shortly a major redesign ensued and a third generation of gantries evolved consisting of a pure *rotational system.* In these systems the source/detector unit, with a pulsing, highly collimated wide-angle (typically 20°–50°) fan beam and a multiple detector array, is rotated 360° around the patient. This single 360° smoothly rotating movement decreases scan time to approximately 3 s, increases appreciably the reliability of the data since it is taken twice, and increases the quality of the reconstructed image. A by-product of this design is that the pulsing x-ray source can be synchronized with physiological parameters enabling

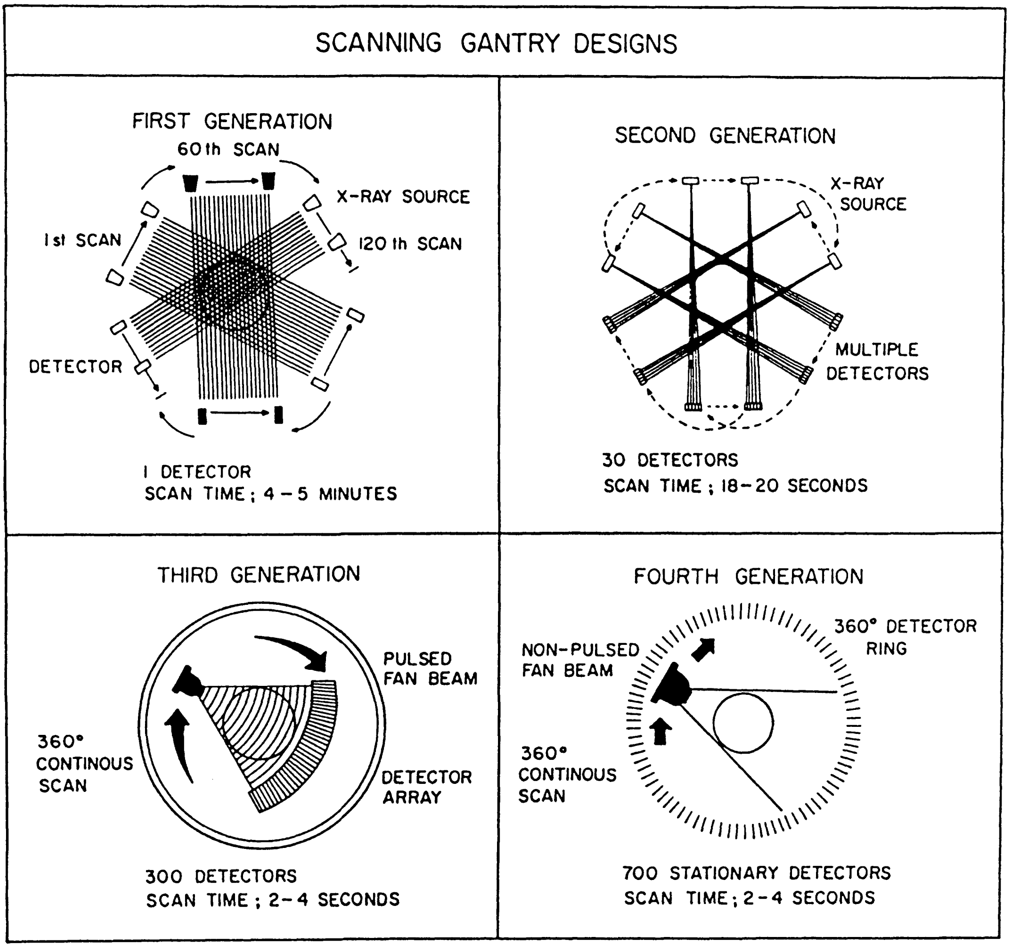
SCANNING GANTRY DESIGNS
FIRST GENERATION
60th SCAN
X-RAY SOURCE
1st SCAN
120th SCAN
DETECTOR
1 DETECTOR
SCAN TIME; 4 – 5 MINUTES
SECOND GENERATION
X-RAY SOURCE
MULTIPLE DETECTORS
30 DETECTORS
SCAN TIME; 18–20 SECONDS
THIRD GENERATION
PULSED FAN BEAM
360° CONTINOUS SCAN
DETECTOR ARRAY
300 DETECTORS
SCAN TIME; 2–4 SECONDS
FOURTH GENERATION
360° DETECTOR RING
NON-PULSED FAN BEAM
360° CONTINOUS SCAN
700 STATIONARY DETECTORS
SCAN TIME; 2–4 SECONDS

FIG. 11. Four generations of CT scanners.

rhythmically moving structures, such as the heart, to be accurately imaged.

The main advantages of this type of system are its simplicity and its speed; however, its fixed geometry system with a fan beam, usually established for the largest patient, makes it inefficient for smaller objects, in particular head scanning. Consequently a fourth generation gantry system consisting of a continuous ring of fixed detectors (usually numbering between 500 and 700) was developed. In this system the x-ray source rotates as before, but the detectors remain stationary, arranged in a full circle around the gantry. The fan beam is increased slightly so that the detectors on the leading and trailing edge of the fan can be continuously monitored and thereby adjust the data in case of shifts in detector performance (Maravilla and Pastel, 1978).

Recently the bulky, noisy electromechanical rotation mechanism has been eliminated and, in its place, a magnetic induction system is employed to rotate the source. This has enabled designers to attain a long sought goal: the 1-s scan.

The whole purpose of the scanning procedure is to take thousands of accurate absorption measurements through the body. Taken at all angles through the cross section of interest, these measurements provide an enormous amount of information about the composition of the section of the body being scanned.

In conjunction with the operation of the scanning and detection system, the data obtained must be processed rapidly to permit producing and viewing an image or scan as quickly as possible. The first stage of this computation process is to analyze all of this raw data and convert it into the set of profiles, normally 180 or more, and then convert these profiles into information which can be displayed as a picture and used for diagnosis. Here we have the heart of the operation of a CT scanner, that element which makes CT totally different from conventional x-ray techniques and most other imaging techniques. The algorithms or computer programs for reconstructing tissue x-ray absorption presently fall into one of the following categories: simple back projection (sometimes called summation), Fourier transforms, integral equations, and series expansions. The choice of an algorithm for a CT scanner depends upon its speed and accuracy (Nagel, 1988).

To facilitate the utilization of these computational processes significant advances had been made in the development of data acquisition and conversion systems by the late 1970s. One of the most important of these developments was the design of the array processor which has become an integral part of the data acquisition and conversion systems used in the most recent CT systems.

Once the image is displayed and initially reviewed, it is often important that this information be retained. Consequently, present CT systems have a variety of storage media such as floppy disks or magnetic tape available. The type of storage facility employed depends to a large extent on the length of time the information is required to be held.

The availability of various storage media, therefore, provides a means of recording, storing, and viewing patient scans independently of the scanning system's operational viewing unit. Consequently, health professionals are provided with around-the-clock access to existing patient scans, irrespective of the in-use status of the basic scanner. With present CT systems, patient records can be stored on floppy disks or magnetic tape in less than 1 min. These magnetic storage devices can serve as a means of permanent storage of patients' records or may be erased and then used again.

Although there have been technological improvements of methods for generating x rays during this century, for the most part medical x-ray sources are basically the same today as they were in the early years. There is a need to develop new methods of generating x rays, including tunable, steerable, and coherent beams. Improved detectors are also needed. Photographic film will no longer be used for detection and archiving in medical imaging systems of the next century. High-resolution, two-dimensional detectors with high sensitivity and capable of rapid sequence imaging and energy resolution are needed. New methods of practical and efficient picture archiving, image communication, and three-dimensional image display will also be important areas for significant work to be done.

3.6.2 Nuclear Medicine. Nuclear medicine has been defined as the application of radioactive materials (tracers) to medical diagnosis and treatment (see NUCLEAR MEDICINE). It is a classic example of a medical discipline that has embraced and utilized concepts developed in the physical sciences. Conceived as a "joint venture" between the clinician and physical scientist, nuclear medicine has evolved into a science with its own body of knowledge and techniques. In the process, the domain of nuclear medicine has grown to include studies pertaining to

1. the creation and proper utilization of radioactive tracers (or radiopharmaceuticals),
2. the design and application of appropriate nuclear instrumentation to detect and display the activity of these radioactive elements, and
3. the determination of the relationship between the activity of the radioactive tracer and specific physiological processes (Wagner, 1975).

The gamma camera system, one of the most basic instruments for diagnostic nuclear medicine (Spencer and Hosain, 1986), converts the radioactive emissions into an image (Fig. 12). Interfacing this device with a digital computer enables clinicians to perform data collection, storage, and analyses.

The gamma camera systems are large and heavy and cannot be moved easily. They are regarded as stationary cameras. However, relatively mobile versions are available with computer interface especially for cardiac studies. Gamma cameras often have capabilities for obtaining whole-body images by linearly moving the camera head or the patient bed. Often called scanning cameras, they have replaced rectilinear scanner systems. A major innovation in gamma camera design has been the addition of an extra camera head with rotational capabilities which is used with a computer system known as single-photon-emission computed tomography (SPECT) for carrying out optional tomographic work. The technique is known as computer-assisted reconstruction tomography (CT) scanning. Tomograms represent the images of isolated cross-sectional slices of the body. They facilitate identification of abnormalities that are otherwise difficult or impossible to identify in a two-dimensional image covering overlying and underlying tissues (Spencer and Hosain, 1986).

With the discovery of artificial radionuclides, the "modern era" of nuclear medicine began. The availability of these radioactive elements increasingly encouraged pioneers in this discipline to employ radioactive tracer techniques to gain information regarding biological and physiological systems. One of the first physiological systems to be studied in both animal and man was the metabolism of phosphorus. Using the cyclotron-produced tracer ^{32}P, several investigators observed the

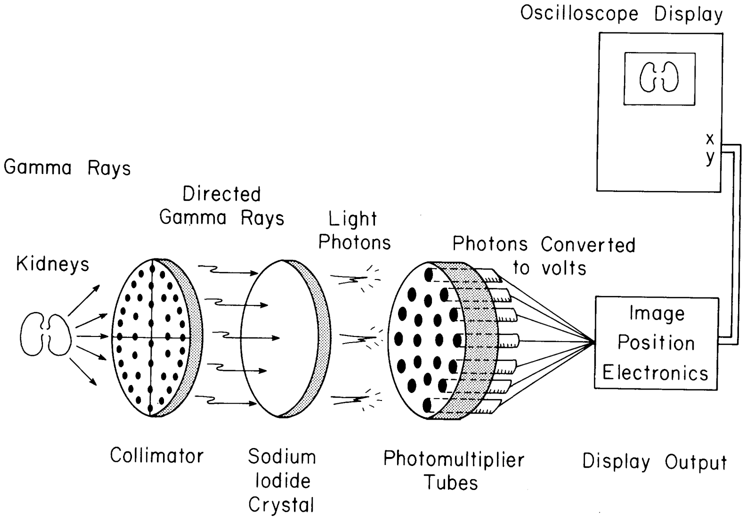

FIG. 12. Basic steps in obtaining an image in nuclear medicine.

great efficiency with which the body absorbed inorganic phosphorus and preferentially utilized it in rapidly multiplying tissues, such as those associated with malignancies. In 1936, John Lawrence, recognizing the therapeutic implications of these findings, acted. He used ^{32}P in the treatment of leukemia, and in the process inaugurated the therapeutic employment of artificial radionuclides (Wagner, 1975).

Although the early uses of radioactive materials in medicine were chiefly in radiation therapy, today most radioactive materials are used to provide useful diagnostic information. These radionuclides can be monitored within the patient (*in vivo*) or in various bodily fluids removed from the patient (*in vitro*). For example, it is possible to inject a patient with a radioactive substance that is "taken up" by a particular organ such as the kidney. By placing a detector over the kidneys, the amount of radioactive material accumulated by the kidney can actually be measured. This uptake test allows the clinician to monitor the activity of specific organs and determine whether they are functioning properly. Another example of an *in vivo* study involves imaging the distribution of a radionuclide within an organ. This can be very important, especially since it has been demonstrated that abnormal tissue tends to accumulate more or less of the radionuclide administered than the normal tissue surrounding it. In this way, *in vivo* measurement can help the clinician delineate the presence of these tissue abnormalities.

The *in vitro* category, on the other hand, includes tests made outside the body. These tests are used to study various chemical, as well as physiological, processes. Today, *in vitro* tests are capable of detecting extremely small amounts of various hormones and chemicals in the blood and have been used to determine

1. insulin dosage in diabetics,
2. whether individuals are immune to specific diseases such as hepatitis, and even
3. the proper dosage of digitalis required by cardiac patients.

Recent developments in this area have indicated that this is an area of application that will continue to grow in importance over the next decade.

The major thrust of the current application in nuclear medicine lies primarily in the use of radioactive tracers in both *in vivo* and *in vitro* studies to evaluate various physiological systems within a patient. The key to success in this application is the specificity of the radioactive material utilized (Bronzino, 1977). That is, the more closely the radionuclide can be tied to a specific function, then the easier it is to examine the dynamic physiological system under study. Clearly, it is beyond the scope of this text to provide a complete outline of all the possible applications presently being utilized. However, some of the most common clinical applications of nuclear technology in studying major organs are summarized in Table 5. Obviously, these applications are only representative of the type of information available using nuclear instrumentation techniques.

Although the initial notion of developing radioactive "magic bullets" capable of seeking out and destroying diseased tissue in the human body has not been realized, it is increasingly evident that the techniques embodied in current nuclear medicine (NM) clearly allow the acquisition of quantitative information not available by other means (Spencer and Hosain, 1986). The utilization of radionuclides makes possible the measurement and visualization of the regional function. As a result, it is possible to use nuclear technology to determine both the site and activity (rate of change) of essential biological processes, thereby permitting abnormalities, i.e., regions of dysfunction, to be detected.

Since most physiological processes within the body are usually in a dynamic state insofar as they exhibit continual change, radioactive tracer techniques have enabled the clinician to monitor accurately the activity of a variety of different physiological systems. These systems range from the pulmonary and cardiovascular systems, to renal function, to the microvascular circulation of the eye. Today, several of these procedures are widely used in clinical medicine, whereas others are conducted only in research laboratories. As has been the case in other fields of medicine, procedures that are presently being viewed as "ancillary" may become an essential part of the routine medical examination conducted in the future.

Although NM will continue to be an important imaging modality, improved radiation

Table 5. Clinical application of nuclear medicine imaging technology.

1. Brain scanning
   a. Detection of intracerebral-space-occupying disease: primary or secondary neoplasms, abscesses, arteriovenous malformations
   b. Evaluation of cerebrovascular disease
   c. Detection and localization of intracranial injury after trauma
2. Lung scanning
   a. Diagnosis and management of pulmonary embolism
   b. Evaluation of regional function in emphysema and other forms of chronic obstructive lung disease
   c. Evaluation of pulmonary venous hypertension
3. Thyroid scanning
   a. Evaluation of thyroid size, position, and function
   b. Functional assessment of thyroid nodules
   c. Diagnosis of functioning metastatic lesions in patients with known carcinoma
4. Liver scanning
   a. Evaluation of liver size, configuration, and position
   b. Preoperative search for metastatic disease when a primary neoplasm exists elsewhere
   c. Functional evaluation of cirrhosis and other diffuse parenchymal processes
5. Spleen scanning
   a. Evaluation of spleen size, shape, and position
   b. Evaluation of spleen function
   c. Detection of intrasplenic-space-occupying disease
6. Pancreas scanning
   a. Carcinoma of the pancreas
   b. Pancreatitis
   c. Miscellaneous pancreatic problems such as pseudocysts

sources and detectors with higher resolution need to be developed. The development of nuclear probes for use in the operating room during surgery would also be of extreme assistance to the surgeon to visualize structures during complex operations.

3.6.3 Ultrasonic Imaging. The application of ultrasound as a noninvasive diagnostic tool is presently widely accepted within the medical community (see IMAGING TECHNIQUES, BIOMEDICAL). As a matter of fact, diagnostic ultrasound studies are often the method of choice because they provide information otherwise not available. The term ultrasound is used to describe sound waves (mechanical vibrations) that are beyond the range of human hearing, that is, above 20 000 Hz. Since sound waves of any frequency are essentially vibratory phenomena, their transmission from one point to another requires the presence of matter. As a result, substances best suited for transmission of sound waves are those with a significant number of molecules. Consequently, solids are better transmitters of sound waves than gases, in which there are fewer molecules. In a vacuum where there are no molecules present, the conduction of sound waves becomes impossible.

Since the composition of each substance varies, these ultrasonic or high-frequency sound waves travel through each of them at different velocities. Thus it is possible to investigate the internal composition of many materials simply by studying the transmission of ultrasound through them. In addition, just as audible sound under correct conditions produces echoes that can be heard, ultrasound, also under the correct conditions, produces echoes that can be detected with appropriate instruments, and then processed and displayed. The application of these basic concepts of ultrasound transmission and echo-ranging led to the development of nondestructive testing techniques in industry and eventually to the utilization of ultrasonic techniques in the conduction of noninvasive testing in medicine today. However, although the discovery of ultrasound preceded the discovery of x rays by 12 years, x rays found almost immediate clinical application, while the use of ultrasound did not become widespread until much later. The primary reason for this delay is that the application of ultrasound had to await the development of appropriate technologies to detect the presence of echoes and display the information in a meaningful manner.

Real-time ultrasonic examinations are being used diagnostically to characterize normal and abnormal anatomical structures accurately in many areas of the body (Peura, 1986). Ultrasonic images of the fetus, for example, are especially useful in fetal examinations since there is no electromagnetic radiation danger, and they can be used to guide a needle to harvest ovary material for pathological tests. Today, it is estimated that from one-half to two-thirds of the pregnant women in the United States have at least one fetal ultrasound scan, while in Great Britain and Germany fetal examination by ultrasound is required by law.

Fetal ultrasound scans are used to view the fetus and to determine whether the fetus is normal. Fetal anatomical features including the width of the fetal head, size of brain ventricles to monitor possible hydrocephalus, crown-to-rump length to determine the age of the fetus, spinal examination to check for abnormalities, the femur length to correlate with biparietal diameter to assess proper physical development of the fetus, and detectable fetal anomalies can be determined.

Real-time ultrasound has been used to perform interventional therapeutic techniques on the fetus *in utero.* Fetal transfusions have been performed using real-time ultrasonic visualization. Also, ultrasonic guidance has been used to drain hydrocephalus, posterior urethral valves, and chylothorax *in utero.*

The introduction of ultrasonic examinations of the heart (Fig. 13) has had great impact on the field of cardiology since the present real-time two-dimensional ultrasound images or echocardiograms provide an almost unlimited number of cross-sectional image planes. Furthermore, using real-time scanners, additional spatial resolutions of the valve morphology as well as the cardiac chambers is now possible. The recent addition of Doppler flow measurements to echocardiography imaging also holds great promise for the future, since the cardiologist will be able not only to visualize the movement of cardiac structures but also to measure the blood-flow velocity in the heart chambers and great vessels as well.

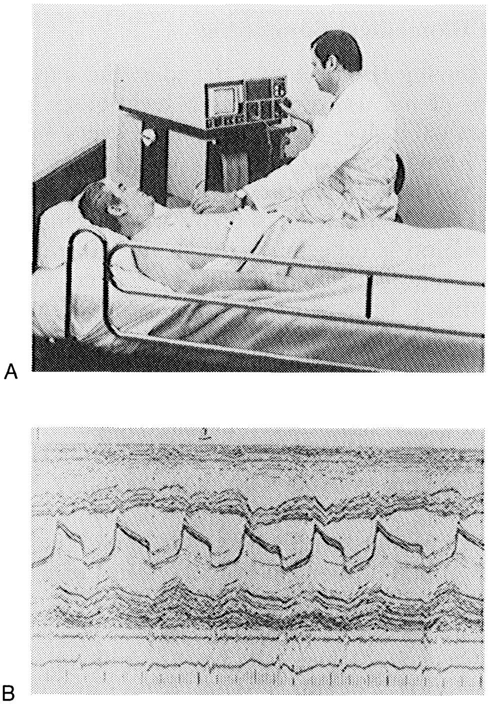

FIG. 13. (A) Diagnostic ultrasound being used to assess cardiac function. (B) Once trained, the physician can interpret the ultrasonic images shown here to detect the presence of any abnormalities. (Courtesy of Pecker Corporation, Cleveland, OH.)

Superficial structures, located less than 5 cm from the skin surface, can also be imaged in real time using ultrasound techniques. Since depth penetration is not a requirement, high-frequency transducers with resultant submillimeter anatomic resolution are used, thereby permitting very small lesions (2–3 mm) to be observed. Superficial anatomical structures which can be imaged include the thyroid, carotid arteries, eye, scrotum, breast, and subcutaneous tissue.

Ultrasonic examinations of the eye are important especially when an ophthalmologist cannot visually examine the eye (in situations in which opaque material such as blood is present or when the patient has cataracts). It would be unfortunate to have a situation in which a cataract was removed only to find that the patient had a retinal detachment or a large melanoma. The major application of ultrasonic examinations of the eye is in the evaluation of nonmetallic foreign bodies, retinal detachment, and intraocular tumors. The availability of high-frequency transducers makes evaluating the eye for periorbital and intraorbital disorders possible.

The future of the medical applications of diagnostic ultrasound is tied to the progress of electronic and computer technology. As electronic and computer circuits become more powerful for the same hardware volume and cost, more sophisticated ultrasonic imagers will be developed and new medical applications will be found. Furthermore, the future of diagnostic ultrasonic applications in medicine is contingent on the development of improved digital processing of the ultrasonic data. New image processing algorithms will handle and quantify the large amounts of ultrasonic data. In present-day systems, once the data are stored in relatively inexpensive digital memory devices, they can be manipulated to improve the image quality and to make simple calculations of tissue dimensions. Further work remains in terms of improving the image quality, extracting information about the nature of the tissue, and making accurate calculations of tissue dimension, including volumes, from three-dimensional measurements.

Perhaps one of the most exciting and challenging areas of ultrasonic imaging deals with tissue characterization. Wells (1977) reviewed the early work in this area and indicated that the ultrasonic signal (or signature) can distinguish between normal or abnormal tissue. A number of ultrasound parameters have been investigated, including attenuation, scattering, and velocity of propagation. These parameters now need to be tested to determine, e.g., from an appropriate ultrasonic scan of the heart, whether a myocardial infarction has taken place, and if so, how severe it was. Similar questions need to be posed for benign and cancerous tissue.

Another area of future development deals with the development of specialized transducers that would be placed inside the body for better imaging of a specific organ. For instance, approximately 30% of all men aged 50 or over have prostate cancer. It has been proposed that a high-resolution ultrasonic transducer operating at 10 MHz, when placed in the rectum, could provide high-resolution ultrasound images of prostate cancer. This should give more definitive results than those available with the normal digital examination for prostate cancer.

3.6.4 Magnetic Resonance Imaging. Tissue variations can be measured by methods other than those using x rays. For example, protons, neutrons, high-frequency sound waves, or nuclear magnetic resonance can also be employed. Each of these energy-containing sources can be used to generate sets of measurements or projections at different angles across the body, and, in a similar manner, a picture can be reconstructed from them.

In recent years, attention has focused on body-scanning devices capable of providing information on specific chemicals as well as structural information about soft tissues. Employing magnetic resonance imaging (*q.v.*) (MRI) techniques instead of x-ray radiation, several systems have produced images that permit the detection of healthy bone, organs, and tissue, as well as areas of chemical and structural irregularity (Lauterbur, 1973). This method is noninvasive, penetrates bony structures without attenuation, does not use ionizing radiation, and appears to be without hazard. As a result, MRI has become an essential and well-established technique for analytic, structural, and dynamic investigations of matter in many disciplines, especially chemistry and physics, and in recent years has made significant inroads in the field of medicine (Taylor, 1986).

Although MRI is still a relatively new technology, it has already changed the definition of optimal diagnostic care in several clinical areas, particularly the brain, spinal cord, and musculoskeletal system. Thus, many radiologists, neurosurgeons, neurologists, and orthopedists are incorporating MRI scanning into their practices, either as a single diagnostic test or as a complement to other modalities such as computed tomography (CT).

3.7 Biomedical Computing

During the past two decades, the expansion of the use of computer technology has been explosive, involving almost every facet of human activity. Today computers are nearly everywhere, and yet these devices are still considered to be in their infancy. This observation is especially true in the utilization of computer technology in the clinical environment. In recent years, the introduction of computerized systems into the medical environment has greatly accelerated, and there is every indication that this will continue. In the future, it is almost certain that computer systems will be even more extensively utilized by health care systems than they are today. Major areas that have been and will continue to be important include patient care and the development of medical expert systems to assist clinicians in the diagnosis and treatment of their patients. As a result, these evolving computer technologies will have significant impact on U.S. health care procedures and policies into the 21st century.

The digital computer is clearly the linchpin for modern state-of-the-art information processing systems. This is especially true in modern health care institutions. The techniques actually used to collect, store, retrieve, and analyze data on the status of a patient, however, have been until recently quite slow and inefficient, even rudimentary. Since the quality of patient management and care depends on the rapid processing of accurate information about the patient's status, comprehensive computer-based health care information management systems have been increasingly used for these purposes. In spite of the progress to date, it is still the case that in many hospital and other health care facilities, patient management and care could be further enhanced by the adoption of current hardware and software data processing systems. Moreover, in the future, the types of information processing tasks that could be carried out by health care professionals will be even further expanded as new applications of computer-based technologies are developed.

Since the early 1960s, health care technologies based on the digital computer have become more and more pervasive. Advances in almost all types of medical research have come to depend heavily on computer-based technologies, either through the acquisition of accurate data from experimentation or the analysis of that data. Efficient financial management of health care facilities has also become increasingly dependent on high-speed data collection and retrieval of patient-related information. Only with such systems is it possible to achieve economic management of hospital inventories of drugs and equipment and provide accurate accounting of patient care costs and billing.

Since the 1970s, patient care computing systems have been developed to provide comprehensive information about the health care status of the patient (Fig. 14). These systems utilize computer hardware and software to collect, retrieve, and present clearly a wide variety of patient-related information (Jenkins, 1978; Bronzino, 1982). The development of such patient care systems has significantly altered and improved the way in which health care professionals utilize information about the patient's health care status. For example, patient care computing systems do the following:

1. provide clinicians with access to sophisticated and often widely dispersed resources,

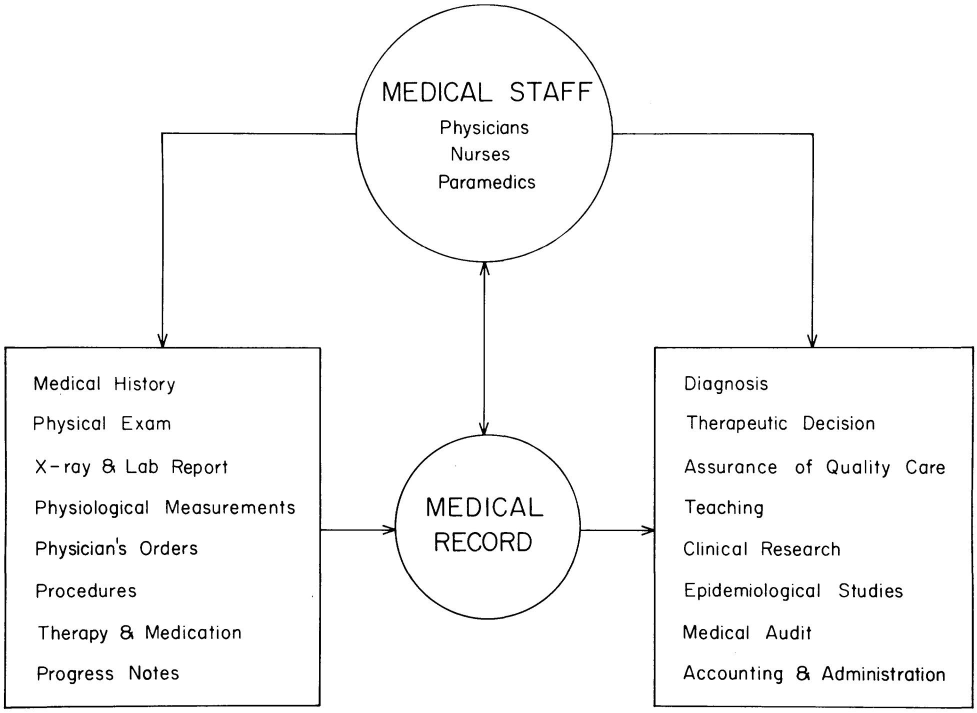
MEDICAL STAFF
Physicians
Nurses
Paramedics
Medical History
Physical Exam
X-ray & Lab Report
Physiological Measurements
Physician's Orders
Procedures
Therapy & Medication
Progress Notes
MEDICAL RECORD
Diagnosis
Therapeutic Decision
Assurance of Quality Care
Teaching
Clinical Research
Epidemiological Studies
Medical Audit
Accounting & Administration

FIG. 14. Block diagram of medical information systems.

2. simplify the assembly of patient-related data for review by all types of health care professionals, and
3. assist health care professionals in making fundamental decisions about patient health care strategies.

As current systems are more widely adopted, the scope of the services they provide expands, and patient care computing will enhance both clinical performance and patient care.

Specific patient-care-related computer applications can be (approximately) allocated to the following categories (Bronzino, 1982):

- clinical laboratory analysis;
- acquisition of patient data and the evolution of automated multiphasic testing systems;
- computer-based medical-record data collection, storage, and information retrieval systems for patient management;
- automated patient monitoring;
- diagnostic support systems (such as electrocardiographic interpretive systems);
- medical imaging;
- artificial intelligence and expert systems.

The application of computers in biology and medicine has been important, and there have been many advances in applying computer technology to problems in the basic life sciences and clinical medicine. Even so, problems remain where computer technology can have a significant effect.

One class of these problems revolves around record keeping in clinical medicine. A hospital chart today does not look much different from what it did ten or twenty years ago, with the exception that it is thicker because of additional diagnostic procedures and patient monitoring data available. Trying to extract information from one of these records remains a problem for anyone other than those intimately involved in the patient's care and the production of the record in the first place. There is little standardization, and storage and retrieval techniques are not at the frontiers of technology. Yet the medical chart is becoming increasingly important in the patient's care, as well as for medical-legal reasons. Change in this feature of clinical medicine has been slow and will probably continue to be slow until such time that truly useful systems are developed and clinicians become more comfortable with computer technology. It is anticipated that these conditions will evolve in the 1990s, and opportunities for developing new methods of medical record keeping on a national and perhaps even an international basis will be best at the beginning of the next century. Such computer systems will have to incorporate advances in computer science such as artificial intelligence and expert systems, high-density data storage, more direct input and output from human beings (such as computer recognition of speech), and a combination of alphanumeric and pictorial data.

Computers will continue to play an important role in medical instrumentation, as they have begun to do in the late 1980s. No longer will instruments be concerned only with data gathering. Signal processing, including specific pattern recognition, self-calibration, preventative maintenance, data transmission, and multiple instrument control, will continue to offer research opportunities. As instrumentation and computer technology both become more complex, it will be important to develop software to keep the technology within reasonable bounds for human users and to optimize the application of the hardware to medical problems.

While computer engineers are just beginning to find the value of electronic systems for parallel processing as a means of speeding up complex computer operations, biological systems have been using this technique all along. Complex pattern recognition routines play an important role in biochemical reactions, such as between enzyme and substrate or antigen and antibody. Although investigators have explored the possibility of harnessing these complex processes into computational hardware, this field still remains at the frontier of technology, and many problems will need to be solved before techniques can be demonstrated and applied.

4. STATUS OF BIOMEDICAL ENGINEERING EDUCATION

The first undergraduate programs in biomedical engineering were initially organized as options in the traditional engineering undergraduate departments (such as electrical and mechanical engineering). As the field of biomedical engineering matured, separately accredited programs in biomedical engineering emerged. Today there are 18 organized undergraduate programs which have been accredited by the Accreditation Board for Engineering and Technology (ABET). A list of the presently accredited programs with the year of initial accreditation is shown in Table 6.

A significant number of undergraduate "option" programs, however, have not moved to separate department status. Since growth in enrollments and in the demographic projections for college-age youth has nearly ceased, it is expected that few, if any, new departments will be formed (Pilkington *et al.*, 1989).

The first formal programs in biomedical engineering began at the master's level, the program at Drexel which began in 1959 being a prime example; Johns Hopkins and the University of Pennsylvania began Ph.D. programs shortly thereafter. Graduate programs, because they are either research oriented or in some cases clinically oriented, are very much cast in the "image" of the particular faculty forming the program. In this respect, graduate programs often have different directions, and thus offer very different areas of specialization to the student. For example, some institutions have a basic research effort in physiological system modeling, others emphasize instrumentation and devices, while still others may have a clinically based program or one that emphasizes rehabilitation engineering. This diversity is appropriate because no one program can possibly provide the depth of education and training required to be competent in all fields (Pilkington *et al.*, 1989).

Table 6. Bioengineering group accredited programs (IEEE Lead Society, with AIChE, ASAE, ASME, and NICE). (Programs in this group are accredited according to the program criteria for bioengineering and similarly named engineering programs. Year of accreditation is shown.[a])

Bioengineering
- Arizona State University, 1985
- University of Illinois at Chicago, 1976
- University of Pennsylvania, 1982
- Texas A&M University, 1977

Biomedical engineering
- Boston University, 1983
- Brown University, 1973
- University of California at San Diego, 1987
- Case Western Reserve University, 1977
- Duke University, 1972
- University of Iowa, 1985
- Johns Hopkins University, 1983
- Louisiana Tech University, 1978
- Marquette University, 1983
- Northwestern University, 1982
- University of Pennsylvania, 1986
- Rensselaer Polytechnic Institute, 1972
- Tulane University, 1981
- Wright State University, 1988

[a] From Pilkington *et al.* (1989).

In Table 7, the actual number of students

Table 7. Total engineering and biomedical engineering.[a]

Full-time enrollments (total for indicated degree)									
Fall (yr)	B.S.-BME	B.S.-Eng.	Percent	M.S.-BME	M.S.-All	Percent	Ph.D.-BME	Ph.D.-All	Percent
1975	1610	231 300	0.70	341	26 004	1.31	233	11 281	2.07
1976	1954	257 800	0.76	302	25 516	1.18	210	10 963	1.92
1977	2355	289 300	0.81	452	26 107	1.73	294	12 359	2.38
1978	2654	311 200	0.85	451	25 360	4.78	243	12 321	1.97
1979	2724	340 500	0.80	464	24 349	1.91	279	13 461	2.07
1980	3252	365 100	0.89	481	28 571	1.68	256	14 465	1.77
1981	3420	387 600	0.88	459	30 679	1.50	287	15 472	1.85
1982	3478	403 400	0.86	510	32 709	1.56	287	16 442	1.75
1983	6747	406 144	0.92	679	37 769	1.80	300	18 540	1.62
1984	3701	394 635	0.94	693	37 718	1.84	283	19 559	1.45
1985	3518	384 191	0.92	641	38 499	1.66	348	21 494	1.62
1986	3627	369 520	0.98	660	42 664	1.55	423	24 227	1.75
Average			0.81			1.61			1.93

[a] From Pilkington *et al.* (1989).

enrolled both in engineering and in biomedical engineering (BME) for the B.S., M.S., and Ph.D. is shown. It will be noted that at the undergraduate level, the total engineering enrollment peaked in 1983, and that biomedical engineering represents 1% of the total number. On the other hand, the percentage of BME students has remained fairly steady in recent years on both the graduate and undergraduate levels, and on this basis the expectation is for BME to follow engineering enrollments as a whole.

Finally, in Table 8 are listed the biomedical engineering undergraduate and graduate programs by state (Browneller, 1986a–f). It will be noted that the field of biomedical engineering is represented in almost every state and has indeed become an important component of the engineering profession.

Table 8. Bioengineering education. Types of programs offered: B.S./B.A., Bachelor of Science and Bachelor of Arts; M.S., Master of Science; Ph.D., Doctor of Philosophy; M.D./Ph.D., M.D. and Ph.D. combined.

State	B.S./B.A.	M.S.	Ph.D.	M.D./Ph.D.
Arizona				
Arizona State Univ.	×			
Univ. of Arizona, Tucson	×	×	×	
Arkansas				
Univ. of Arkansas, Fayetteville		×	×	
California				
California State Univ., Sacramento		×		
Stanford Univ.		×	×	
Univ. of California, Berkeley	×	×	×	
Univ. of California, Davis		×	×	
Univ. of California, Los Angeles	×	×	×	
Univ. of California, Santa Barbara		×	×	
Colorado				
Colorado Technical College	×			
Univ. of Colorado, Boulder	×			
Connecticut				
Hartford Graduate Center		×		
Trinity College	×			
Univ. of Bridgeport	×			
Univ. of Connecticut		×	×	
District of Columbia				
Catholic Univ. of America	×	×	×	
George Washington Univ.	×	×	×	
Hawaii				
Univ. of Hawaii, Manoa		×	×	
Illinois				
Northwestern Univ.	×	××	×	
Univ. of Illinois, Chicago		×	×	×
Univ. of Illinois, Urbana/Champaign	×	×	×	
Indiana				
Purdue Univ.	×	×	×	
Iowa				
Iowa State Univ., Ames		×	×	
Univ. of Iowa, Iowa City	×	×	×	
Kansas				
Kansas State Univ.	×	×	×	

(continued)

Table 8. *(continued).*

State	B.S./B.A.	M.S.	Ph.D.	M.D./Ph.D.
Maryland				
The Johns Hopkins Univ.			×	×
Massachussetts				
Boston Univ.	×	×	×	
MIT		×	×	×
Northeastern Univ.	×			
Western New England College	×			
Michigan				
Michigan State Univ.	×			
Michigan Technological Univ.	×			
Univ. of Michigan		×	×	
Wayne State Univ.		×	×	
Minnesota				
Univ. of Minnesota		×	×	
Nevada				
Univ. of Nevada, Reno[a]				
New Hampshire				
Dartmouth College (Thayer School of Engineering)		×	×	×
Univ. of New Hampshire	×			
New Jersey				
Fairleigh Dickinson Univ.		×		
New Jersey Institute of Technology		×		
Rutgers Univ.		×	×	×
New Mexico				
Univ. of New Mexico	×			
New York				
Columbia Univ.	×	×	×	
Hofstra Univ.	×			
New York Institute of Technology	×			
Rensselaer Polytechnic Institute	×	×	×	
Univ. of Rochester	×	×	×	×
Syracuse Univ.	×			
Ohio				
Air Force Institute of Technology		×		
Case Western Reserve Univ.	×	×	×	×
Ohio State Univ.		×	×	×
Univ. of Akron		×	×	
Univ. of Toledo		×	×	
Wright State Univ.		×	×	
Oklahoma				
Univ. of Oklahoma		×	×	
Pennsylvania				
Carnegie Mellon Univ.	×	×	×	
Drexel Univ.		×	×	
Univ. of Pennsylvania	×	×	×	×
Rhode Island				
Brown Univ.	×	×		
South Dakota				
South Dakota State Univ.				

(continued)

Table 8. Bioengineering education. Types of programs offered: B.S./B.A., Bachelor of Science and Bachelor of Arts; M.S., Master of Science; Ph.D., Doctor of Philosophy; M.D./Ph.D., M.D. and Ph.D. combined. *(continued.)*

State	B.S./B.A.	M.S.	Ph.D.	M.D./Ph.D.
Texas				
Southern Meth. Univ.	×	×	×	
Texas A&M Univ.	×	×	×	
Univ. of Texas, Arlington		×	×	
Univ. of Texas, Austin		×	×	
Univ. of Texas Health Science Center		×	×	
Vermont				
Univ. of Vermont	×	×		
Virginia				
Univ. of Virginia		×	×	×
Virginia Commonwealth Univ.		×		
Virginia Polytechnic Institute & State Univ.		×	×	
Washington				
Walla Walla College	×			
Univ. of Washington, Seattle		×	×	×
West Virginia				
West Virginia Univ.		×	×	
Wisconsin				
Marquette Univ.	×	×	×	
Milwaukee School of Engineering	×		×	
Univ. of Wisconsin, Madison		×	×	
Wyoming				
Univ. of Wyoming	×	×	×	

---

[a] Details not received in time for publication.

GLOSSARY

Aliphatic Molecule: A molecule belonging to that series of compounds characterized by open chains of carbon atoms rather than by rings.

Artificial Intelligence: The science of making machines do things that would require intelligence if done by men.

Atherosclerosis: A form of arteriosclerosis characterized by a variable combination of changes to the intima of arteries, not arterioles, consisting of the focal accumulation of lipids, complex carbohydrates, blood and blood products, fibrous tissue, and calcium deposits; associated with changes in the media of the arteries.

Digitalis: The dried leaves of the *Digitalis purpurea* plant which are powdered into tablets or capsules. Cardiotonic glycosides, especially digitoxin and digoxin, are obtained from various species of the Digitalis plant. Use of these glycosides increases cardiac output and results in diuresis.

Etiology: The study of the causes of disease.

Hemodialyzer: Instrument for removal of chemical substances from the blood by passing it through tubes made of semipermeable membranes.

Hemolysis: The destruction of red blood cells with the liberation of hemoglobin which diffuses into the fluid surrounding them.

Hermetic: Airtight.

Hydrolytic Cleavage: The splitting of a molecule into two or more simpler ones through the addition of water.

Isotropic: 1. Possessing similar qualities in every direction. 2. Having equal refraction.

Left Ventricular Assist Device (LVAD): Implantable device used to assist in those instances where the major contractual muscle of the heart, the left ventricle, is not performing satisfactorily.

Morbidity: 1. State of being diseased. 2. The number of sick persons or cases of disease in relationship to a specific population.

Morphology: Science of structure and form without regard to function.

Mortality: 1. State of being mortal. 2. The death rate; ratio of number of deaths to a given population.

Multiphasic Screening: A battery of clinical tests to which a patient is subjected in order to evaluate his or her medical status.

Orthotic Device: A device for straightening or correcting a deformity or disability.

Osteoporosis: Increased porosity of bone, seen most often in the elderly.

Periorbital: Surrounding the socket of the eye.

Picture Archiving: The use of computers to digitize and store images.

Prosthetic Devices: Devices designed to replace missing parts (limbs, etc.)—artificial substitutes.

Rehabilitation Engineering: A subspecialty of biomedical engineering concerned with the development of devices to assist the disabled.

SA (Sino-Atrial) Node: Node at junction of superior vena cava and right cardiac atrium, regarded as the site of initiation of the heartbeat.

Scintillation: The photoeffect caused by collision of particles emitted by radioactive substances and a detector surface.

Streak Artifacts: Blurs which occur on computed tomographic (CT) images.

Sulfanilamide (Para-aminobenzenesulfonamide): White, slightly bitter, crystalline substance derived from coal tar, the parent of the azo family of dyes. Formerly widely used in the treatment of a number of infections, due to its toxic effects its use has been superceded by more effective and less toxic sulfonamides.

Tomography: Any of several noninvasive special techniques of roentgenography designed to show detailed images of structures in a selected plane of tissue by blurring images of structures in all other planes.

Thrombogenuity: Producing or tending to produce a clot.

Ultrasound: Inaudible sound in the frequency range of approximately 20 kHz to 10 GHz. Ultrasound waves have different velocities in tissues of different densities and elasticity. This property permits the use of ultrasound in outlining the shape of various tissues and organs in the body.

List of Works Cited

Allen, P., Lillehei, C. W. (1957), *Minn. Med.* **40**, 672–676.

Bernard, W. F., LaFarge, C. G. (1969), in: F. W. Hastings (Ed.), *Artificial Heart Program Conference, 1969*, Washington, DC: U.S. GPO.

Bronzino, J. D. (1977), *Technology for Patient Care*, St. Louis, MO: Mosby.

Bronzino, J. D. (1982), *Computer Applications for Patient Care*, Reading, MA : Addison-Wesley.

Bronzino, J. D. (1986), *Biomedical Engineering and Instrumentation: Basic Concepts and Application*, Boston, MA : PWS.

Browneller, P. A. (1986a), *J. Clin. Eng.* **11**, 39–53.

Browneller, P. A. (1986b), *J. Clin. Eng.* **11**, 153–172.

Browneller, P. A. (1986c), *J. Clin. Eng.* **11**, 239–250.

Browneller, P. A. (1986d), *J. Clin. Eng.* **11**, 297–307.

Browneller, P. A. (1986e), *J. Clin. Eng.* **11**, 377–393.

Browneller, P. A. (1986f), *J. Clin. Eng.* **11**, 469–480.

Christel, P., Meunier, A., Lu, A. J. C. (Eds.) (1986), *Biological and Biomechanical Performance of Materials*, New York: Elsevier.

Crichton, M. (1970), *Five Patients—The Hospital Explained*, New York: Knopf.

DeVries, W. (1986), in: J. D. Bronzino, V. Smith, and M. L. Wade (Eds.), *Technology and Medicine*, Hartford, CT: Trinity College.

DeVries, W. C., Anderson, J. L., Joyce, L. D., Anderson, F. L., Hammond, E. H, Jarvik, R. K., Kolff, W. J. (1984), *N. Engl. J. Med.* **310**, 273–278.

Eberhart, R. C. (1989), *IEEE Eng. Med. Biol. Mag.* **8** (2), 7.

Galioto, I. M. (1986), in: J. D. Bronzino (Ed.), *Biomedical Engineering and Instrumentation*, Boston, MA: PWS, pp. 79–105.

Heimke, G. (1986), in: A. F. von Recum (Ed.), *Handbook of Biomaterials Evaluation*, New York: Macmillan, pp. 38–54.

Hench, L. L. (1988), in: P. Ducheyne, J. E. Lemons (Eds.), *Bioceramics: Material Characteristics vs. In Vivo Behavior*, Vol. 523. Boca Raton, FL: CRC, pp. 54–71.

Jacobs, J. E. (1975), *IEEE Trans. Biomed. Eng.* **22**, 100–106.

Johns, R. J. (1975), *IEEE Trans. Biomed. Eng.* **22**, 107–110.

Knowles, J. (1973), *Sci. Am.* **229**, 128–137.

Ko, W. (1988), in: J. Kline (Ed.), *Handbook of Biomedical Engineering*, New York: Academic, pp. 3–71.

Lauterbur, P. (1973), *Nature (London)* **242**, 190–191.

Ledley, R. S. (1976), *Comput. Biol. Med.* **6**, 239–246.

Lown, B. (1968), *Sci. Am.* **219** (1), 19–27

Maravilla, K., Pastel, M. (1978), *Comput. Tomog.* **2**, 137–144.

McMillin, C. R. (1989), *IEEE Eng. Med. Biol. Mag.* **8** (2), 30–36.

Nagel, J. (1988), in: J. Kline (Ed.), *Handbook of Biomedical Engineering*, New York: Academic, pp. 243–314.

Neuman, M. R., Brill, A. B., Gibbons, C. F., Greatbatch, W., Mates, R., Rushmer, R. (1989), *IEEE Eng. Med. Biol. Mag.* **8** (3), 18–26.

Norman, J. C. (1975a), *Cardiovas. Dis. Bull. Tex. Heart Inst.* **2**, 259–264.

Norman, J. C. (1975b), *Cardiovas. Dis. Bull. Tex. Heart Inst.* **2**, 425–437.

Norman, J. C., Huffman, F. N. (1974), in: H. I. Russek (Ed.), *Cardiac Disease: New Concepts in Diagnoses and Therapy*, Baltimore, MD: University Park.

Pacela, A. (1990), *Bioengineering Education Directory*, Brea, CA: Quest.

Peura, R. A.(1986), in: J. D. Bronzino (Ed.), *Biomedical Engineering and Instrumentation*, Boston, MA: PWS, pp. 347–386.

Pilkington, T. C., Long, F. M., Plonsey, R., Webster, J., Welkovitz, W. (1989), *IEEE Eng. Med. Biol. Mag.*, **8** (3) 9–17.

Plonsey, R. (1973), *Eng. Educ.* **64**, 117–179.

Russotti, G. M., Coventry, M. B., Stouffer, R. N. (1988), *Clin. Orthop. Relat. Res.* **235**, 141–147.

Saha, S. (1989), *IEEE Eng. Med. Biol. Mag.* **8** (2), 37–39.

Spencer, R. P., Hosain, F. (1986), in: J. D. Bronzino (Ed.), *Biomedical Engineering and Instrumentation*, Boston, MA: PWS, pp. 279–346.

Tarjan, P. P., Gold, R. D. (1988), in: J. Kline (Ed.), *Handbook of Biomedical Engineering*, New York: Academic, pp. 123–148.

Taylor, K. D. (1986), in: J. D. Bronzino (Ed.), *Biomedical Engineering and Instrumentation*, Boston, MA: PWS, pp. 387–430.

Wagner, H. N., Jr. (1975), *Nuclear Medicine*, New York: HP.

Walmsley, B. (1979), in: L. Kreel, R. E. Steiner (Eds.), *Medical Imaging*, Buckinghamshire, England: HMM, pp. 15–22.

Wells, P.N.T. (1977), *Biomedical Ultrasonics*, London: Academic.

Further Reading

Anbar, M. (1987), *Computers in Medicine*, Rockville, MD: Computer Science.

Boretos, J. W., Eded, M. (1984), *Contemporary Biomaterials*, San Francisco: Noyes.

Bronzino, J. D. (1986), *Biomedical Engineering and Instrumentation*, Boston, MA: PWS.

Christensen, D. A. (1988), *Ultrasonic Bioinstrumentation*, New York: Wiley.

Feinborg, B. N. (1986), *Applied Clinical Engineering*, Englewood Cliffs, NJ: Prentice-Hall.

Fung, Y. C. (1981), *Biomechanics: Mechanical Properties of Living Tissue*, New York: Springer-Verlag.

Gedder, L. A. (1986), *Cardiovascular Devices*, New York: Wiley.

Kline, J. (1988), *Handbook of Biomedical Engineering*, New York: Academic.

Macovski, A. (1983), *Medical Imaging Systems*, Englewood Cliffs, NJ: Prentice-Hall.

Park, J. B. (1979), *Biomaterials: An Introduction*, New York: Plenum.

Webster, J. G. (1988), *Encyclopedia of Medical Devices and Instrumentation*, New York: Wiley.

BIOMEDICAL IMAGING TECHNIQUES

See IMAGING TECHNIQUES, BIOMEDICAL

BIOMEDICAL USES OF RADIATION

WILLIAM R. HENDEE, *American Medical Association, Chicago, Illinois, U.S.A.*

Introduction 549
1. **Overview** 549
1.1 Radiation as a Diagnostic Probe 549
1.2 Radiation as a Therapeutic Agent 550
2. **Properties of X and Gamma Radiation** 551
2.1 Production 551
2.2 Interactions 552
2.3 Detection 553
2.3.1 Gas-Filled Detectors 555
2.3.2 Solid Detectors 555
3. **X-Ray Imaging** 556
3.1 Fluoroscopy 557
3.2 Radiography 557
3.3 Angiography 558
3.4 Computed Tomography 558
4. **Nuclear Imaging** 560
4.1 Radiopharmaceuticals 560
4.2 Scintillation Camera 561
4.3 Single Photon Emission Computed Tomography 562
4.4 Positron Emission Tomography 562
5. **Radiation Therapy** 563
5.1 High-Energy Teletherapy 563
5.2 Treatment Planning 564
5.3 Brachytherapy 565
5.4 Particle Therapy 566
6. **Future Prospects** 567
Glossary 567
Works Cited 568
Further Reading 569

INTRODUCTION

Feelings of fear and dread are a typical initial reaction of most persons to the expression "radiation." Yet the expression simply means that energy is being transferred from one location to another. Often this transfer leads to useful applications that are highly beneficial to humans. Among applications of great benefit are the uses of radiation for the diagnosis and treatment of human disease and injury. This article examines these uses, and explores future possibilities that may add to the beneficial applications of radiation in medicine.

1. OVERVIEW

1.1 Radiation as a Diagnostic Probe

The detection and diagnosis of disease and injury has always been a significant challenge to physicians. Before the end of the past century, this challenge frequently was insurmountable, because the physician was confined primarily to interpreting external signs for evidence of anatomic, physiologic and metabolic difficulties within the patient. Vague indicators such as temperature, pulse and respiratory rates, and visual and muscular reflexes were often all the physician had available to determine what may be troubling the patient. Exploratory surgery to identify the cause of a patient's suffering was a common but risky procedure. On other occasions sympathetic understanding and emotional support were all the physician could offer as the patient struggled to recover from a malady.

The discovery of x rays in 1895 by Wilhelm Röntgen, a physicist at the University of Wurzburg, provided the first opportunity to examine the interior of the body in a noninvasive manner. Unlike lower-energy forms of electromagnetic radiation such as visible and ultraviolet light, x rays (and gamma rays emitted by radioactive nuclei) have the ability to penetrate biological tissues. As a beam of x rays traverses the body, it is absorbed to varying degrees depending on the anatomic structures in the path of the beam. The x-ray beam

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

exiting from the body varies in intensity from one location to another, depending on the structures through which each part has passed. By directing the x-ray beam onto a fluorescent screen or x-ray film, the varying x-ray intensity can be transformed into a pattern of differing light intensity or optical density. This pattern forms a two-dimensional image of the three-dimensional structures in the body through which the x-ray beam has passed. One of the first x-ray images captured on film is shown in Fig. 1.

Today x-ray imaging, referred to as radiography when the image is captured on film or digital storage device, and fluoroscopy when the image is displayed dynamically on a fluorescent screen, is widely employed in medicine. Over the years it has been joined by several other techniques to display images of internal anatomy and physiology. The creation of cross-sectional images of patient anatomy from x-ray projections obtained at different orientations and processed by computer is referred to as computed tomography. In nuclear medicine (*q.v.*), images depict the distribution of radioactive materials inside the patient's body after they have been administered orally, intravenously or by breathing a radioactive gas. Ultrasonography employs beams of ultrasound to reveal the internal anatomy of patients and the flow of blood within arteries and veins (see BIOLOGICAL EFFECTS OF SOUND AND ULTRASOUND). Magnetic resonance imaging (*q.v.*) produces images from weak radiofrequency signals elicited from patients by placing them in an intense magnetic field. All of these imaging technologies are invaluable aids in the detection and diagnosis of disease. Each has contributed in a substantial way to the dramatic improvements in the health and well-being of people that have occurred over this century.

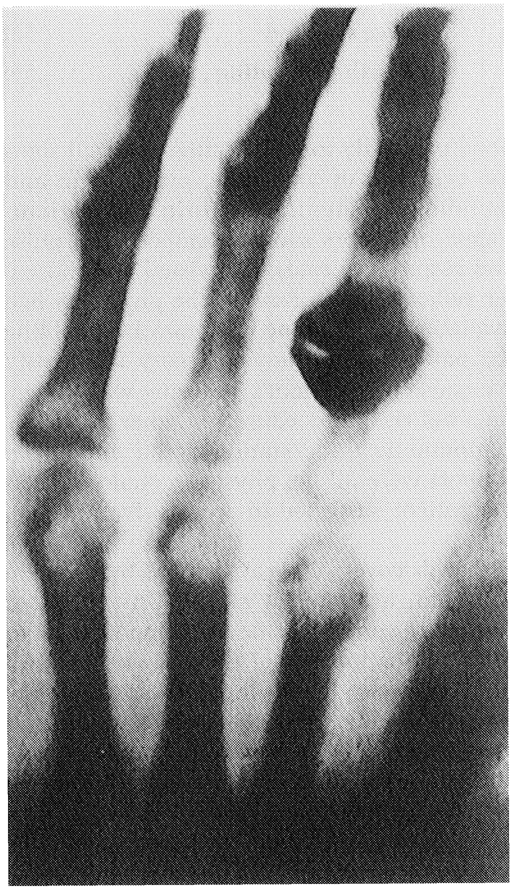

FIG. 1. Early radiograph of the hand taken by Röntgen with his wife as the subject. The enlargement on one finger is a ring. (Courtesy of Deutches Röntgen-Museum, Remschied-Lennep, West Germany.)

1.2 Radiation as a Therapeutic Agent

As a beam of x or gamma rays traverses the body, it gradually loses intensity by transferring energy from the beam to the atoms comprising the tissues along the path of the radiation. Beams of particles such as electrons and protons exhibit a similar behavior. This dissipation of energy causes the ejection of electrons from the atoms, resulting in a concentration of ions along the path of the radiation. Hence, the radiation is categorized as ionizing radiation.

An ionizing event can injure or kill a cell if it occurs in an atom contained within a critical molecule of the cell. This effect, termed the "direct effect of radiation," is responsible for some of the biological effects of radiation. More often, however, the ionizing event occurs in an atom that is not an integral part of an essential cellular structure. Cell injury or death can still occur, however, because the event may create reactive products called free radicals that in turn attack cellular structures. This mechanism of radiation action on cells is referred to as the "indirect effect of radiation." The direct and indirect mechanisms of radiation action on cells are responsible for the damaging consequences of exposure to radiation, including the induction of cancer (carcinogenicity), development of abnormalities in persons exposed before birth (teratogenicity), and production of genetic effects in offspring of exposed individuals (mutagenicity) [Hendee, in press (A)] (see BIOLOGICAL EFFECTS OF ELECTROMAGNETIC AND PARTICLE RADIATIONS).

Cancer is a disease in which certain cells have been transformed so that they are growing and multiplying at a much faster rate than normal. If these cells can be destroyed, the patient often has a good chance of recovering from the cancer. If the cells are disseminated throughout the body, perhaps because the cancer has spread (metastasized) from its initial (primary) site, the preferred treatment usually is to try to kill the cells with a cellular poison. This approach is referred to as chemotherapy, and the specialists who employ it are called medical oncologists. If the cells are confined to a specific region of the body, often they can be removed surgically, or destroyed by concentrating radiation in the region. The latter technique is referred to as radiation therapy, and the physicians who utilize it are called radiation oncologists. A frequent approach to cancer treatment is to employ radiation to destroy the primary site of the cancer (sometimes after most of the cancer has been removed surgically), followed by chemotherapy in an effort to kill cancer cells that have spread to distant sites.

In the early years following the discovery of x rays, radiation was used to treat a wide spectrum of conditions, including arthritis, muscular aches and pains, tuberculosis, and cancer. Over the years, most of these conditions have succumbed to better methods of treatment. However, radiation continues to be used for treatment of cancer, with excellent results for many forms of cancer when the disease is confined to a primary site, or when distant metastases can be treated with cellular poisons.

2. PROPERTIES OF X AND GAMMA RADIATION

2.1 Production

X rays and gamma rays are electromagnetic radiations with energies greater than those possessed by other components of the electromagnetic spectrum (Fig. 2). X rays are produced as a result of electron transitions within atoms, whereas gamma rays are released as atomic nuclei undergo transitions from higher to lower energy states. In medicine, x rays are produced in x-ray tubes in which electrons that have been accelerated to high energy bombard a metal target where they interact with target atoms to produce x rays. The x rays are both characteristic x rays emitted during interactions of the incident electrons with electrons of the target atoms, and *bremsstrahlung* released during interactions of the incident electrons with target nuclei. Each of these types of interactions is discussed below. Medically useful gamma rays are released from radioactive atoms selected to yield radiations of desired energies.

As electrons accelerated through a potential difference of several thousand volts impinge on the metal target of an x-ray tube, they interact with the electrons and nuclei of the target atoms. Interactions with electrons cause the electrons to be ejected from the target atoms, including some from the electron shells nearest the nucleus. As the vacancies left by these ejected inner electrons are filled by other electrons cascading down from higher levels, x rays are released. These x rays are called characteristic x rays because their energies are characteristic of the electron energy levels of the particular metal used for the target. In a typical x-ray beam used for medical diagnosis, 10% or so of the x-ray intensity

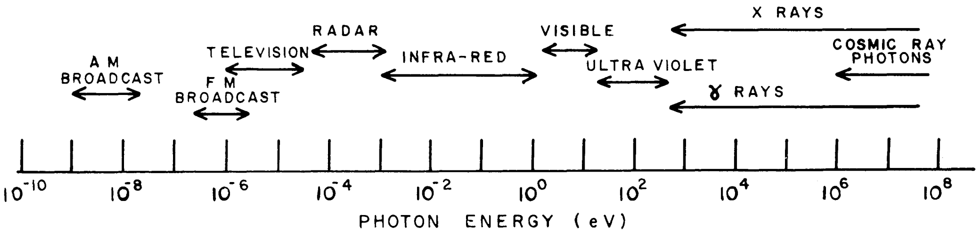

FIG. 2. The electromagnetic spectrum.

is contributed by characteristic x rays (Hendee, 1979).

High-energy electrons impinging on a target also interact with the nuclei of the target atoms. These interactions result in a slowing of the electrons and a release of energy in the form of x rays. This radiation is termed "braking radiation" or *bremsstrahlung*. In a diagnostic x-ray beam, over 90% of the x-ray intensity is due to bremsstrahlung (Hendee, 1979). In the higher-energy x-ray beams used for therapy, the x-ray intensity is almost entirely due to bremsstrahlung.

For the x-ray beams used for diagnosis, a relatively simple x-ray tube can be used to produce x rays. The device consists of a filament that is heated to high temperature to release electrons by thermionic emission, and a voltage of several thousand volts to accelerate the released electrons toward a metal target. X rays are produced as the electrons are absorbed in the target. However, less than 1% of the electron energy is released as x rays, and over 99% is dissipated as heat (Curry, Dowdey, and Murry, 1990). Often the metal target is embedded in a rotating anode so that the intense heat can be dissipated over a larger volume of material. The filament and anode are enclosed within an evacuated glass envelope to prevent molecules of air from scattering the electrons on their way to the target.

X rays used for radiation therapy are much more energetic than those employed in radiologic diagnosis. To prevent their emission in undesired directions, heavy shielding is required around the x-ray source. Also, the electrons that produce them must be accelerated to very high energies, on the order of several million electron volts. (An electron volt is the kinetic energy acquired by an electron accelerated through a potential difference of one volt.) Electrons accelerated to such high energies produce x rays more efficiently compared with those accelerated to several thousand electron volts in diagnostic x-ray tubes. Hence, stationary targets, sometimes embedded in a large mass of copper and cooled by a circulating fluid, are used in radiation therapy.

For x-ray beams of several million electron volts, it is not feasible to provide acceleration all at once by subjecting the electrons to a single voltage difference between electrodes. Instead, the acceleration is provided in a continuous fashion in an accelerating tube known as a waveguide that is part of a linear accelerator (see ACCELERATORS, LINEAR). In a traveling-wave linear accelerator, electrons are accelerated by riding the crest of an electrical field that sweeps down the entire length of the waveguide. At the end of the waveguide the electrons are either extracted as a high-energy electron beam, or directed onto a metal target to produce x rays. In a standing-wave linear accelerator, electrons are accelerated by injecting them into a waveguide in which the electrical field is present in a standing-wave pattern. Most linear accelerators used today in radiation therapy are the standing-wave type (Hendee, 1981).

2.2 Interactions

X and gamma rays are electromagnetic radiations, termed photons, that carry no electronic charge and hence do not interact through charged particle (Coulombic) interactions with matter. Instead, they are absorbed by various mechanisms, of which three are especially important for the range of photon energies employed in diagnostic imaging and radiation therapy. These three types of interactions are the photoelectric effect, Compton interaction, and pair production.

Photoelectric interactions occur primarily with relatively low-energy x and gamma rays such as those used for medical diagnosis. In these interactions, the photon interacts usually with an inner electron of an atom by transferring its entire energy to the electron. In this manner the x or gamma ray disappears, and in its place is a high-speed electron that has escaped from the atom as a consequence of receiving the photon energy. The kinetic energy E_k of the electron is

$$E_k = h\nu - E_B,$$

where $h\nu$ is the energy of the x or gamma ray, and E_B is the binding energy of the electron, defined as the energy required to remove the electron from the attractive influence of the nucleus. In general, the likelihood of a photoelectric interaction increases rapidly with increasing atomic number of the absorbing material, and decreases rapidly with increasing energy of the x or gamma ray. These trends are depicted in the attenuation curves shown in Fig. 3, which also reveal discontinuities

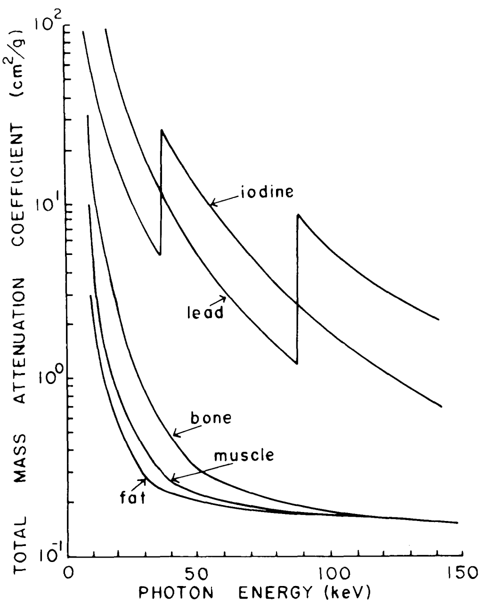

FIG. 3. The attenuation coefficient of x rays in lead and iodine, plotted as a function of x-ray energy.

where the absorption curves for iodine and lead rise rapidly at the specific energy of 33 keV for iodine and 88 keV for lead. This "resonance energy," termed an absorption edge, occurs at an energy where the photon has just enough energy to eject an electron from the innermost electron energy level (the *K* shell) of the absorbing atom.

Higher-energy x rays in diagnostic radiology, and almost all x rays employed in radiation therapy, tend to interact with matter by a process known as the Compton interaction. In this process, the x ray transfers part of its energy to an outer electron of an atom, with the x ray scattering in a new direction with reduced energy. The electron is ejected from the atom with a kinetic energy essentially equal to that received from the photon, because the binding energy of the outer electron is negligible. The probability of a Compton interaction decreases gradually with increasing energy, and does not depend on the atomic number of the absorbing material. It does, however, depend on the electron density (number of electrons per gram) of the absorber. Hydrogen is a particularly efficient absorber of x rays by the Compton process because it contains no neutrons and hence has almost twice as many electrons per unit mass as do other elements. Compton interactions produce scattered radiation that interferes with the portrayal of fine detail in diagnostic images unless it is removed by use of a mechanical grid in diagnostic radiology or a pulse-height analyzer in nuclear medicine. Compton interactions are preferred in radiation therapy because they do not produce excessive absorption of energy in tissue components (e.g., bone) of high atomic number.

For photons of energy greater than about 1 MeV, a third type of interaction, pair production, can occur. This process involves an interaction of an x or gamma ray in the presence of a nucleus of the absorbing material. The photon disappears, and a pair of electrons, one negative and one positive, is created. The kinetic energy of the negatron-positron pair equals the energy of the photon, less 1.02 MeV of energy which is expressed as the mass of the pair of particles. The positron exists only until its kinetic energy has been expended, at which time it interacts with an electron. The interaction causes the particles to be annihilated, with their mass expressed as two "annihilation photons" of 0.51 MeV each that emerge in opposite directions from the interaction site. The likelihood of pair production increases with increasing photon energy, and it is a common interaction for x and gamma rays of several MeV.

2.3 Detection

Energetic radiations such as x and gamma rays, and charged particles such as electrons and protons, can induce cellular damage in tissues that ultimately may affect the well-being of exposed individuals. In spite of this damaging effect, ionizing radiations are not detectable by any of the human senses. Consequently, the presence of radiation can be measured only through indirect methods that take advantage of some property of the radiation that affects materials in a manner that is detectable by humans. In most cases this property is the ability of radiation to ionize and excite atoms in an absorbing medium.

The first detector of x rays was the fluorescent material in Röntgen's laboratory that glowed during his experiments of electrical conduction through gases. Studies of this fluorescence caused Röntgen to speculate that a new form of radiation, which he called "x rays," was generated by his experiments. As the x rays interacted by depositing energy in the fluorescent material, electrons were raised to higher energy levels. As these electrons returned to more stable levels, visible light was emitted which could be seen as fluorescence.

The property of x-ray-induced fluorescence was used by Edison to construct fluorescent screens to provide x-ray images of human anatomy, and to enhance the sensitivity of techniques to record x-ray images on photographic film. For several decades, fluorescent screens of materials such as (Zn,Cd)S were employed in medicine to display dynamic x-ray images. These images were very dim and could be viewed reliably only in a darkened room. The ability to see fine detail in the images was compromised by these requirements, because the persons viewing the images had to be dark-adapted, and hence were forced to use rod vision rather than the cone vision available under conditions of brighter illumination. Also the method was time-consuming and tedious.

In the 1950s efforts were directed to developing a way to increase the brightness of the fluoroscopic x-ray image so that dark-adaptation could be eliminated and the visibility of detail improved. The product of these efforts was the x-ray image intensifier. This device amplified image brightness by converting the x-ray induced light image to a pattern of electrons that could be accelerated and then converted back into a light image. The image intensifier is illustrated in Fig. 4. X rays from the patient impinge on a fluorescent screen where they are converted into light. The light is projected onto a photoelectric surface that emits electrons in response to the intensity of incident light. The electrons are accelerated through a potential difference of several thousand volts onto a second, smaller fluorescent screen. Emerging from this screen is a light image that is much brighter than the original light pattern from the input fluorescent screen. The ratio of the brightness of the two screens, known as the brightness gain of the intensifier, is a product of two factors: the minification gain accomplished by decreasing the size of the image, and the electronic gain achieved by accelerating the electrons to high energy before the output image is produced. Brightness gains of 3000–4000 are typical of modern image intensifiers, and they alleviate the need for dark-adaptation and the use of rod vision.

For several years, the input screen of the image intensifier was composed of materials identical to those used in fluorescent screens. In the 1960s, however, CsI replaced (Zn,Cd)S as the most common fluorescent material. Today almost all fluoroscopy is performed with image intensifiers employing CsI screens. The output image can be viewed directly through an optical system that collects the light and projects it into an orientation convenient for observation. More commonly, most fluoroscopic systems employ a television camera coupled to the output screen so that the image can be viewed on one or more television monitors.

Fluoroscopic images are very useful for examining rapidly changing anatomic pro-

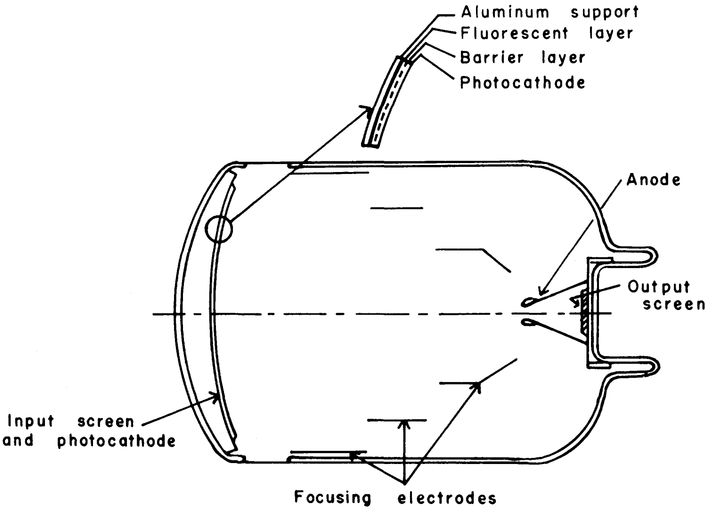

FIG. 4. An x-ray image intensifier.

cesses. However, they do not provide permanent images. To obtain a permanent record of an x-ray examination, images must be captured on photographic film or some type of electronic storage medium. In some examinations, the x-ray pattern emerging from the patient is imposed directly on a film. In most cases, however, the film is exposed primarily by light released during interactions of x rays with fluorescent screens positioned on both sides of the photographic film. These intensifying screens greatly enhance the sensitivity of the recording process and yield acceptable images in much shorter examination times that require less exposure of the patient to radiation. For many decades $CaWO_4$ was the fluorescent material most commonly employed in intensifying screens. In the 1970s rare-earth fluors were introduced as a replacement for $CaWO_4$. Rare-earth intensifying screens are more efficient absorbers of x rays, and today most intensifying screens employ these materials in place of $CaWO_4$ (Mackay, 1984).

Fluorescent screens and photographic film are excellent methods for recording images of the pattern of x rays emerging from a patient during an x-ray examination. They are not entirely satisfactory, however, for routine quantitative measurements of the amount of radiation present at some location near a source of radiation. For these measurements, two other types of detectors have been developed. These types of detectors, referred to as gas-filled detectors and solid detectors, are discussed in detail in companion articles on DETECTORS, RADIATION; DETECTORS, SEMICONDUCTOR; and DETECTORS, SCINTILLATION. Hence, only an overview of gas-filled and solid detectors is provided here.

2.3.1 Gas-Filled Detectors. Radiations such as x and gamma rays are termed "ionizing radiations" because they ionize any material through which they pass. If the material is a gas, the ejected electrons and residual positive ions can be collected by charged electrodes to yield a signal that depicts the passage of radiation through the gas. In some cases, the signal can be interpreted not only that radiation was present, but also how much radiation was present.

Conceptually, the simplest type of gas-filled detector is the ionization chamber. This device consists of two electrodes operated at a potential difference of a few hundred volts. The electrodes are suspended in a gas, often air, inside a detector whose walls are similar in atomic composition (but not physical density) to the gas. As radiation interacts with the walls of the chamber, electrons are released that traverse and ionize the gas inside the chamber. The resulting electrons constitute a tiny electrical signal signifying that radiation has interacted with the detector.

With an air-filled ionization chamber, measurements usually are expressed in units of coulombs/kilogram to signify the amount of charge collected per unit mass of air inside the chamber. This unit specifies the "radiation exposure" present at the location of the chamber. In past years, a different unit, the roentgen (R) was used to describe the "exposure" of the chamber to radiation. A person exposed to the radiation source at the location of the chamber for the same length of time was described as receiving a certain "exposure" to radiation expressed in roentgens. With adoption of the International System (SI) of Units, the roentgen has fallen into disfavor. Today the unit coulombs/kilogram, rather than roentgen, is preferred for descriptions of radiation exposure.

The ionization chamber is a useful device for measuring relatively intense beams of radiation. For small amounts of radiation, however, the ionization chamber is less useful because the current or voltage drop is extremely weak. For these measurements, a device that amplifies the signal electronically in the detector is more useful. This approach is used in a number of devices, including portable proportional and Geiger–Mueller (GM) survey meters, to provide a relatively simple and inexpensive way to detect the presence of radiation.

2.3.2 Solid Detectors. Gas-filled detectors are useful in a variety of applications to detect and measure the amount of radiation present at a particular location at some moment in time. For x and gamma rays, however, these detectors are relatively inefficient because not many of the photons are stopped by the detectors. Absorption of x and gamma rays could be improved substantially by using solid detectors of greater physical density. However, the solid material must be carefully chosen so that a signal is generated only when radiation interacts with the detector.

Certain semiconducting solids have this property, and detectors that employ them are known as semiconductor radiation detectors.

A semiconductor is a material with properties intermediate between those of an insulator and a conductor. Under normal circumstances the most energetic electrons in a semiconductor are confined to the energy level known as the "valence" band, and the material functions as an insulator. However, the valence band is separated from the higher-level "conduction" band by only a few electron volts of energy. Hence electrons can be moved from the valence band into the conduction band by absorption of small amounts of energy such as those delivered during exposure of the semiconductor to radiation. During exposure, the number of electrons moved into the conduction band can be measured by placing a potential difference across the material and collecting the resulting current. In this manner, the semiconductor serves as a radiation detector. Semiconducting materials frequently employed as radiation detectors include germanium and silicon. These materials are usually "doped" with a small amount of lithium to improve their detection characteristics, and are known as Ge(Li) and Si(Li) detectors.

Scintillation detectors present an alternative approach to the design of solid detectors of radiation. In these devices, a solid insulating material such as sodium iodide is used in which a few electrons are raised to a higher energy band during exposure to radiation. By adding a small amount of impurity (usually thallium) to the NaI crystal, some of the electrons return to the lower energy band through an intermediate energy state. This return path results in the release of visible light that can be measured with one or more photomultiplier tubes. Combinations of NaI(Tl) crystals and photomultiplier tubes are known as scintillation detectors. These detectors are widely used for imaging in nuclear medicine (Fig. 5).

Another type of scintillation detector is useful for measuring the amount of radiation delivered in radiation therapy. This type of detector, known as a thermoluminescent detector (TLD), can be exposed to radiation, and later heated to release an amount of light proportional to the energy absorbed in the detector. The most common TLDs are $Li_2B_4O_7$ and LiF. These materials have a composition that resembles soft tissue, so the signal depicts the approximate energy that would have been absorbed in tissue at the location of the detector. In this manner, the detector is said to measure the "dose" of radiation absorbed in soft tissue, where absorbed dose is expressed in grays (Gy), with 1 gray = 1 joule of energy absorbed per kilogram of absorbing material. An earlier unit of absorbed dose, the rad, is still frequently encountered in the radiologic literature. An absorbed dose of 1 rad equals 0.01 joule of energy absorbed per kg of absorbing material. That is, 1 rad = 0.01 Gy.

3. X-RAY IMAGING

An x-ray beam emerging from an x-ray tube presents a relative uniform intensity of x rays across its area. This uniform intensity is

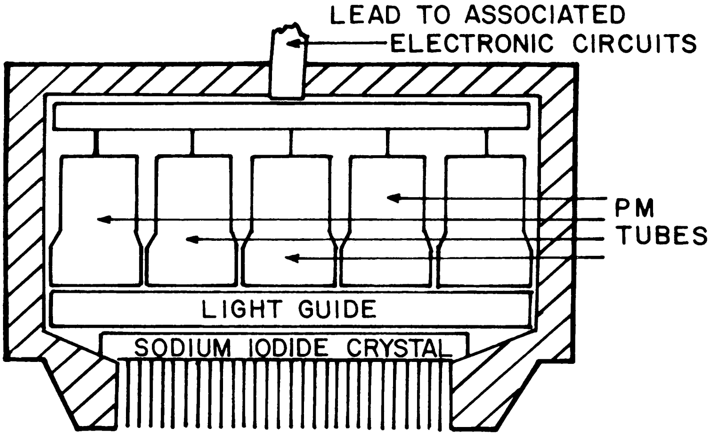

FIG. 5. A scintillation detector used for nuclear imaging.

modulated by the types of tissues encountered as the x-ray beam traverses the patient. X rays that traverse bone are attenuated to a greater extent than those that pass only through soft tissue, because bone has a higher atomic number and a greater physical density than soft tissue. Fat transmits slightly more x rays than muscle because its atomic number and physical density are lower. X rays are transmitted readily through air-filled structures such as lung because the physical density is very low. At times highly attenuating substances known as contrast agents are added to the patient to enhance the absorption of x rays. For example, barium-containing substances are frequently administered orally [barium swallows for upper gastrointestinal (GI) studies] or rectally (barium enemas) to improve the visualization of the esophagus, stomach, small and large intestine and rectum. Contrast agents also can be administered intra-arterially and intravenously to improve visualization of the circulatory system, a specialty of diagnostic imaging known as angiography. Contrast agents also are useful in visualizing the kidneys and urinary tract (pyelography), spinal column (myelography), and a number of other anatomic structures.

Because of its differential attenuation in various tissues, the x-ray beam emerging from the patient is no longer uniform in intensity. Instead, it possesses an intensity pattern that conveys information about the composition of the body through which the x-ray beam has passed. As discussed earlier (and in the companion articles on IMAGING TECHNIQUES, BIOMEDICAL and X-RAY IMAGING: RADIOGRAPHY AND TOMOGRAPHY), this information can be captured as a dynamic image on a fluoroscopic screen (fluoroscopy), or as a permanent image on film (radiography).

3.1 Fluoroscopy

In fluoroscopy, the image is captured on a fluoroscopic screen and viewed as a dynamic display of anatomy. This technique is especially useful for studies where the anatomy is constantly in motion, such as those of the gastrointestinal tract and heart. However, even in fluoroscopic studies of rapidly-changing anatomy, images must be recorded to permit later review and consultation and provide a permanent record of the examination. Permanent images may be captured in a variety of ways. For example, inserting a cassette containing intensifying screens and film momentarily into the x-ray beam between the patient and the image intensifier provides a "spot film" of patient anatomy at a particular moment. Alternatively, the light image from the output screen of the intensifier may be deflected momentarily into a 100-mm "photospot" camera positioned near the output screen. High image frame rates required for studies of the heart can be achieved by taking a motion picture of the fluoroscopic image with a 35-mm camera coupled to the intensifier output screen. This technique is known as cineradiography or cineangiography. Most fluoroscopic systems in use today employ a television camera and monitor to view the image. In these systems, the image may be captured on video tape or disk for later playback and analysis. If the television signal is converted from analog into digital form, the resulting images can be stored on computer disk or tape for subsequent retrieval and display.

3.2 Radiography

Virtually everyone has had radiographs made to assess a possible fracture or to evaluate the lungs and chest. Each year many people have a radiographic examination to identify a possibly fractured vertebra, or to characterize problems in the urinary or GI tract. Women are encouraged to have periodic x-ray examinations (mammography) of the breast to rule out the presence of breast cancer. Radiography is a very common diagnostic technique, and medical diagnosis would be severely handicapped if it were not available.

Until recently, radiographs were obtained by exposing film directly to x rays or indirectly to light emitted from intensifying screens exposed to x rays. Today, radiography is in transition from a film-based diagnostic specialty to one that will eventually depend, at least in part, on the use of electronic detectors. Electronic detectors are important for radiology because electronic signals can be easily digitized for processing and display by computer. The movement of radiography from an analog to a digital mode of operation is of great interest because it addresses a major problem facing medical institutions today.

In most institutions the x-ray film fileroom is regarded as a major weak link in the chain of services provided for patients. Films are easily lost, misplaced and borrowed, and not infrequently are unavailable when needed for medical diagnosis and follow-up. Improving this situation by maintaining a computer file of all diagnostic images would be a major improvement over retaining images in analog form. However, a digital process for capturing or translating radiographic images is a precondition for a computerized information file.

With the exception of radiography and fluoroscopy, diagnostic imaging techniques used today in medicine are largely digitally based. Adding digital radiography to this array of digital devices permits consideration of a digital imaging network (DIN), often described as a picture archiving and communications system (PACS). In the PACS approach, digital images can be acquired, stored, retrieved, transmitted, and displayed at great convenience throughout an institution. A few institutions have made a firm commitment to evolution of a PACS approach to diagnostic imaging. Many problems remain to be solved, however, not the least of which are the cost of transition and the challenge of converting existing films to digital form [Hendee, in press (B)].

3.3 Angiography

One of the most important applications of x-ray imaging is for the study of the circulatory system. From studies of the heart and great vessels to analyses of the circulation in the brain and extremities, x-ray imaging is relied upon for the depiction of normal anatomy and the identification of abnormal conditions, including circulatory malformations, dysfunctional heart valves, coronary artery disease, the presence of atherosclerotic plaques and blood clots, abnormal circulation caused by injuries and diseases such as cancer, and a wide spectrum of other conditions. Under most circumstances, blood-containing vessels are not distinguishable from muscle and other soft tissues displayed in x-ray images. To help the vessels stand out in an x-ray image, they must be filled with a contrast agent. The agent must have the same physical density and osmolarity as blood; however, it can contain a substance of high atomic number to increase the absorption of x rays. Iodine-containing materials are the preferred substances for studies of the circulatory system, primarily because the absorption edge of 33 keV for iodine (Fig. 3) is about equal to the average energy of x rays in a typical diagnostic x-ray beam. X-ray studies of arteries and veins with iodine-containing contrast agents is referred to as angiography.

3.4 Computed Tomography

For the first seven decades of this century, the medical applications of x rays improved slowly but steadily. In the early 1970s, however, a revolutionary approach to x-ray imaging was introduced. This approach was x-ray transmission computed tomography (CT).

One of the major disadvantages of conventional x-ray imaging is that a three-dimensional volume of tissue is represented as a two-dimensional image. The person viewing the image is forced to interpret the third dimension of depth by knowing the anatomy and noticing which structures seem to lie in the shadows of others. Computed tomography reduces this difficulty by presenting images of thin slices of tissue with little overlapping information. Consequently, the visibility of structures of similar composition is improved, and the images are easier to interpret. Visualization of low-contrast information is improved also because only a thin slice of tissue is exposed to x rays, and scattered radiation is reduced.

The first CT scanners commercially introduced employed a tightly collimated, pencil-like x-ray beam that scanned across the patient (Fig. 6). Moving in synchrony on the opposite side of the patient was a scintillation detector that measured the intensity of x rays transmitted through the patient. Measurements of x-ray transmission were made at mm increments along each of several scan paths oriented at slightly different angles with respect to the patient. After several thousand measurements had been obtained, the data were subjected to a reconstruction algorithm that computed a cross-sectional distribution of x-ray attenuation coefficients for the anatomic section of interest. These coefficients could be sorted into numerical ranges, with each range assigned a shade of gray for display as a cross-sectional image. Coefficients in contiguous sections could be interpreted in

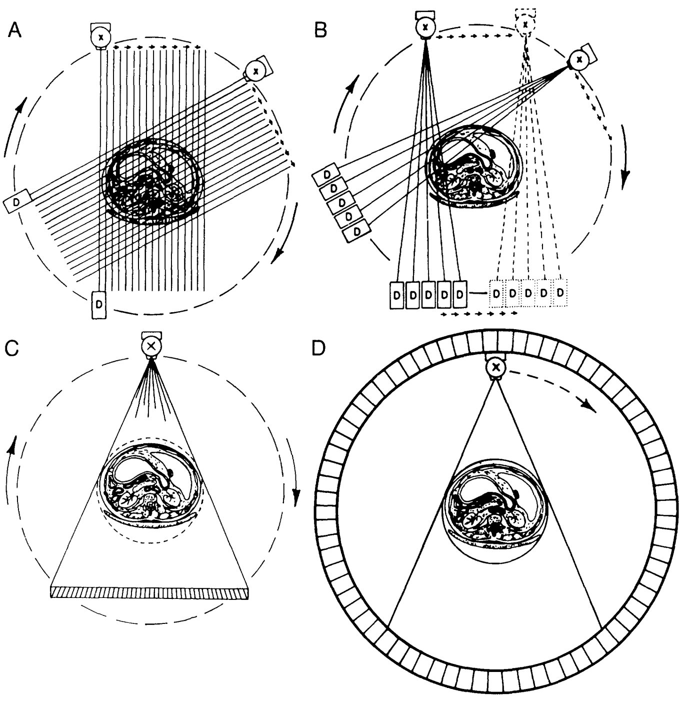

FIG. 6. Different geometries used for data acquisition in x-ray computed tomography. (A) First generation; (B) second generation; (C) third generation; and (D) fourth generation.

similar fashion to yield coronal and sagittal images. Representative CT images are shown in Fig. 7.

The major difficulty with early CT scanners was that several minutes were required to obtain the transmission data from which images were reconstructed. Within a few months imaging times were reduced to several seconds by employing a fan-shaped x-ray beam and a bank of detectors on the opposite side of the patient to record the transmission data. This second-generation CT scanner also quickly became outdated as a third generation scanner was developed that employed purely rotational motion of the x-ray tube and bank of detectors. With this approach, imaging times as short as 1–2 s could be achieved. An alternate approach, known as a fourth

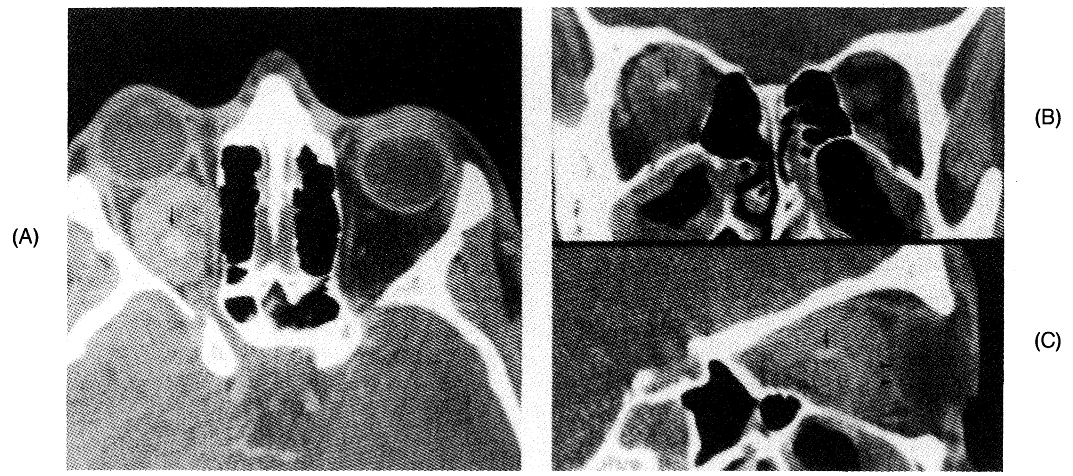

FIG. 7. (A) Cross-sectional (transverse), (B) coronal, and (C) sagittal CT images of an optic nerve glioma (astrocytoma). The cancer cannot be discriminated from the optic nerve and exhibits calcifications (arrows) and compression of the globe (arrowheads). [From Zonneveld (1988).]

generation scanner that was competitive with third generation scanners, employed a circular array of stationary detectors and an x-ray tube that rotated around the patient inside the detector array (Fig. 6). Some third and fourth generation scanners employ fast scintillation detectors, usually bismuth germanate crystals, while others use ionization chambers filled with a high atomic number gas such as xenon.

Computed tomography revolutionized the field of diagnostic imaging and opened the discipline to the development of similar techniques employing other types of radiations, including gamma rays (SPECT and PET imaging) and magnetic resonance imaging (*q.v.*) (MRI) (Hendee, 1983). Today CT is a well-established and mature technology that is found in almost every hospital and outpatient imaging center of modest or greater size. Computed tomography has been of particular benefit to neurology and neurosurgery, and it is difficult to remember some of the diagnostic difficulties encountered frequently by these specialties before CT became available.

4. NUCLEAR IMAGING

Many atomic nuclei are unstable because they contain unusual configurations of neutrons and protons, or because they possess more energy than required to exist in the state of least energy. In the latter case, the nucleus can gain stability by releasing its excess energy. The most common mechanism for releasing excess energy is by its emission as one or more gamma rays, electromagnetic radiations that resemble x rays except that they originate from nuclear rather than electronic transitions. Gamma-ray emission is one of the modes of radioactive decay available to unstable atoms. Because gamma rays can penetrate tissue in a manner similar to x rays, they can be detected outside the body after their emission from radioactive materials present inside the body. The procedure of administering a radioactive substance to a patient, and then using radiation detectors to measure the gamma rays emitted from the patient, is known as the medical specialty of nuclear medicine.

Nuclear medicine originated in the 1920s and 1930s during studies of the behavior of certain radiopharmaceuticals (chemical compounds containing radioactive atoms) in specific organs such as the thyroid and kidneys. A major impetus to nuclear medicine occurred after World War II when nuclear reactors became available for the wide-scale production of radioactive atoms (radionuclides) that were potentially useful in diagnostic medicine. Two developments in the early 1960s, the scintillation camera and the technetium-99m generator, further advanced nuclear medicine as a diagnostic specialty. In the 1970s the computer was incorporated into nuclear imaging to permit studies of dynamic physiologic processes with radiopharmaceuticals. These studies were enhanced in the 1980s by application of computed tomographic techniques to analyses of the distribution of radiopharmaceuticals in the body. As nuclear medicine moves into the 1990s, the technology of positron emission tomography promises to be another milestone in the evolution of nuclear medicine. Today nuclear medicine is one of the most sophisticated of the imaging technologies, and is essential to the detection and diagnosis of a variety of diseases and injuries.

4.1 Radiopharmaceuticals

The key to success in nuclear imaging is the availability of radiopharmaceuticals that distribute in the body in ways that reveal the static and dynamic features of certain organs and regions. Desirable characteristics of a radiopharmaceutical are that it should

1. be relatively inexpensive and easy to prepare,
2. remain intact chemically after administration,
3. concentrate rapidly in the organ of interest,
4. emit gamma rays of easily measurable energy,
5. emit little particulate radiation,
6. be eliminated rapidly from the organ after analysis, and
7. have a relatively short half life for radioactive decay.

Several radiopharmaceuticals satisfy these criteria for the study of specific organs and tissues in the body (Sorenson and Phelps, 1987).

The nuclide employed most frequently for nuclear imaging is a radioactive isotope of the element technetium known as technetium-99*m* (${}^{99}\mathrm{Tc}^{m}$), where 99 defines the isotopic form of technetium, and "m" depicts a metastable energy state (i.e., a higher-energy state with a lifetime significantly longer than the typical nuclear time scale of 10^{-13} s). This nuclide emits gamma rays of 140 keV, an almost ideal energy for measurement with a scintillation camera. Little particulate radiation is emitted with the gamma rays. The half life of technetium-99*m* (the time required for half of the atoms in a sample to decay) is 6 h, long enough to conveniently synthesize a radiopharmaceutical containing ${}^{99}\mathrm{Tc}^{m}$, and short enough to limit the absorbed dose delivered to tissue containing the nuclide. It can be readily incorporated into a variety of compounds with desirable physiologic properties. A major advantage of technetium is its production by decay of molybdenum-99 (${}^{99}\mathrm{Mo}$), an easily produced radioactive nuclide with a half life of 66 h. The molybdenum-99 can be placed on an ion-exchange column, and the ${}^{99}\mathrm{Tc}^{m}$ removed periodically in the laboratory by washing the column with a saline solution. Many nuclear medicine laboratories receive one or more of these "technetium generators" each week from a commercial supplier. Often a generator is referred to as a "technetium cow," and removal of the technetium, usually twice a day, is referred to as "milking the cow." An image of the liver obtained following intravenous administration of a colloidal suspension of technetium sulfide labeled with technetium-99*m*, is shown in Fig. 8.

At least 85% of all nuclear medicine studies are performed with technetium-99*m* incorporated into various chemical compounds for analysis of lung physiology, cardiac dynamics, liver integrity, kidney function, the presence of cancer metastases in bone and various organs, and a number of other conditions. Other radioactive nuclides used in specific applications include iodine-123 and iodine-131 for studies of the thyroid, and thallium-204 for detection of impaired heart muscle. The use of radiopharmaceuticals for medical diagnosis and evaluation is essential to the care of millions of patients each year, and the quality of their health care would be severely compromised if radiopharmaceuticals were not available.

4.2 Scintillation Camera

Almost all nuclear medicine images are obtained with a scintillation camera. This device includes a radiation detector consisting of a large-diameter (up to 20 inches) but very thin ($\frac{1}{4}$–$\frac{1}{2}$ in.) scintillation crystal of sodium iodide that is optically coupled to several (up to 70 or so) photomultiplier tubes. When a gamma ray emerging from the patient's body interacts with the crystal, light is emitted which is captured by the photomultiplier tubes. This process induces a signal in each tube that varies with the total amount of light emitted (and therefore the energy of the gamma ray), and with the distance of the tube from the interaction site in the crystal. By analyzing the signals from the photomultiplier tubes, the site of interaction in the crystal can be determined, as well as whether the interaction represents absorption of a primary gamma ray or one scattered into the crystal during a Compton interaction in the patient. By using a pulse-height analyzer to reject signals induced by scattered gamma rays, the anatomic detail presented in the image can be preserved.

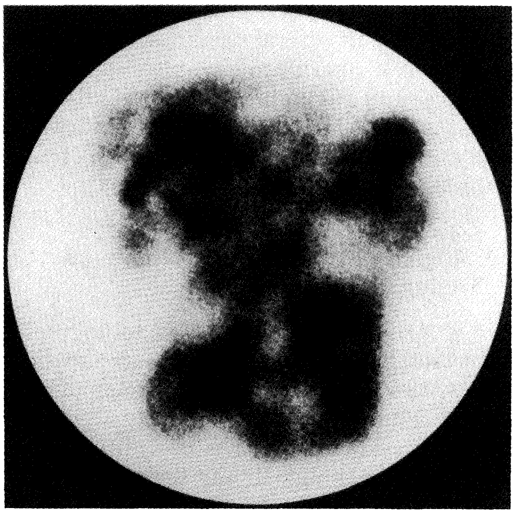

FIG. 8. A TcS liver image revealing areas of reduced uptake indicative of cancer metastases. [From Phillips and Bernardino (1988).]

Knowing where the gamma-ray interactions occur in the crystal is useful only if the interaction sites can be correlated with the origin of the gamma rays in the patient. This correlation is achieved by placing a mechanical collimator on the patient side of the scintillation crystal. The type of collimator most often used is the parallel multihole collimator. With this collimator, a gamma-ray interaction in the crystal can be correlated with the origin of the gamma ray directly below the interaction site. In this manner the distribution of interactions across the crystal reflects the distribution of radioactivity in a region of the patient directly below the crystal. Other types of collimators that are useful in limited applications include the pin-hole collimator (particularly helpful in thyroid imaging) and the converging collimator (used primarily for cardiac studies).

Nuclear images reveal the static distribution of radioactivity in a particular organ or region of the body, or the transient passage of radioactivity through the organ or region. Nuclear images are captured on film as a permanent record, often in a multiformat camera that places several sequential images on one large sheet of photographic film. Data processing by computer is an integral feature of nuclear imaging, and permits quantitative measurement and display of the distribution of radioactivity in selected regions of interest over time. Production of radioactive pharmaceuticals, and acquisition and interpretation of nuclear images requires the involvement of specialists in several disciplines, including chemistry or pharmacology, physics or engineering, and medicine. Consequently, nuclear medicine is a true multidisciplinary specialty that represents the leading edge of high technology in medicine (Saha, 1984).

4.3 Single Photon Emission Computed Tomography

The principles of computed tomography have been applied to nuclear imaging to yield the useful imaging technique of single photon emission computed tomography (SPECT). The expression "single photon" is included in the title to distinguish the technique from Positron Emission Tomography (PET) in which images are produced by collecting annihilation photons emitted and collected in pairs. In the SPECT approach, images of the distribution of radioactivity in the patient are obtained from several orientations, usually with a rotating camera that acquires data from several angles during the examination. These projection data are combined by computer to yield cross-sectional images of the distribution of radioactivity. Application of the SPECT approach to nuclear imaging is particularly useful in studies of the heart where cross-sectional displays of radioactivity in the myocardium reveal damage to the heart muscle caused by diminished blood flow in the coronary arteries. Another application of growing importance is the use of SPECT to study the distribution of radioactivity in the brain, and how the distribution is affected by conditions such as Alzheimer's dementia, epilepsy, and stroke. Additional applications are under development, and SPECT is today an important part of the services offered by any nuclear imaging facility.

4.4 Positron Emission Tomography

Atomic nuclei with a ratio of neutrons to protons that is too low for stability tend to decay by the emission of positively charged electrons. These electrons are referred to as positrons. In the immediate vicinity of their origin, positrons unite with electrons to produce annihilation photons, electromagnetic radiations of 0.51 MeV that leave the annihilation site in opposite directions. If a positron-emitting pharmaceutical is administered to a patient, the annihilation photons can be detected with scintillation detectors positioned on opposite sides of the patient. By operating the detectors in coincidence (i.e., an interaction is recorded only if it occurs in both detectors simultaneously), the origin of the photons can be localized to the cylinder of tissue between the two detectors. If the patient were examined with coincidence detectors operated at different angles, the origin of the photons could be localized to a small volume of tissue within the patient.

In a positron emission tomographic (PET) imaging device, the patient is surrounded by a ring of detectors, with each detector operated in coincidence with a few detectors on the opposite side of the patient. Annihilation photons are detected by this network of coincidence detectors to yield a tomographic image of the distribution of the positron-emitting pharmaceutical across a slice of tissue through the patient. The usefulness of PET imaging is being explored for a number of applications, including the detection of cancer and the study of heart disease. But the application generating the greatest interest is the use of PET for studying the physiology of the brain.

Several investigators are studying the usefulness of PET for fundamental investigations of mental activity in response to various internal and external stimuli, and for evaluation of a number of mental disorders ranging from dementia and depression to schizophrenia and sociopathic behavior (Wong, 1990). In addition, PET is being used to examine conditions such as autism and dyslexia. In these studies, images of the rate of utilization of positron-labeled glucose derivatives, or of the concentration of other positron-containing agents, are examined to determine how well they delineate the condition under investigation. Examples of PET images of the utilization of deoxyglucose labeled with the positron emitter carbon-11 are shown in Plate 1 for an individual subjected to various external stimuli.

5. RADIATION THERAPY

Radiologic diagnosis is accomplished by exposing the body to radiation delivered by a distant source (x-ray imaging), or by measuring radiation from a source of radiation present inside the body (nuclear imaging). Similarly, radiation therapy of cancer is conducted by exposing the body to an external source of radiation (teletherapy, where "tele" is from the Greek for *distant*), or by placing one or more radiation sources inside the body (brachytherapy, where "brachy" is from the Greek for *short*). Although most applications of radiation therapy employ x or gamma rays, beams of energetic electrons sometimes are used, and other types of particles are being explored experimentally.

Before WWII, teletherapy was performed almost entirely with x rays of intermediate energy (between 100 and 400 keV). These x rays were not very penetrating, and high doses were required to superficial tissues in order to deliver enough radiation to the cancer. The highest doses often were administered to the skin, and inflammation and burning of this sensitive tissue often occurred during treatment. In fact, skin reddening (erythema) was used as an indicator of the effectiveness of the treatment and the ability of the patient to withstand it.

5.1 High-Energy Teletherapy

At the close of WWII, nuclear reactors became available for the production of radioactive nuclides that might be useful in medicine. One of the first nuclides to be explored was cobalt-60 (^{60}Co) for use in radiation therapy. Each decay of a ^{60}Co nucleus releases two gamma rays that are high enough in energy (1.17 and 1.33 MeV) to deliver the maximum dose a few millimeters below the skin rather than on the skin surface. This "skin-sparing effect" of ^{60}Co greatly reduced the problems of inflammation and burning encountered with x rays of lower energy. For this reason, and because the gamma rays are more penetrating than conventional x rays, ^{60}Co teletherapy units were widely adopted for radiation therapy in the 1950s and 1960s. However, the skin-sparing effect also removed a familiar landmark with which the radiation oncologist monitored the patient's treatment. In place of this landmark, indirect methods of dose estimation became very important to the delivery of radiation treatments. These methods required more elaborate dosimetry and treatment planning, and many physicists became involved in clinical radiation therapy. Although a few physicists had worked in radiation therapy in earlier years, the advent of high-energy therapy with ^{60}Co required many more, and gave rise to the specialty of medical physics.

Although ^{60}Co units were a major advance in treating cancer with radiation, they had some disadvantages as well. Unlike an x-ray tube, a ^{60}Co source could not be "turned off," since it is always radioactive. Hence, elaborate means were required to shield the source when not in use, and to move it automatically into position during treatment so that a beam of unshielded radiation could be directed towards the patient. This massive shielding made ^{60}Co units bulky and difficult to maneuver into desired positions around the patient. Also, the ^{60}Co source was relatively large, and the edges of the radiation beam were not as well-defined as needed for precise separation of treated and untreated tissues. Cobalt-60 decays with a half life of 5.3 yrs, and the high-activity sources used for teletherapy had to be replaced every few years, a precarious operation for all involved. Finally, the intensity of radiation delivered by a ^{60}Co source was limited by the activity of the source, and was never as high as might be desired for the treatment of some types of cancer. These limitations resulted finally in replacement of most ^{60}Co units with linear accelerators in the late 1960s and 1970s.

Linear accelerators provide a versatile and dependable method for delivering high-energy x rays (a few MeV) in a sharply defined beam to cancers deep within the body. The x-ray beam can be directed towards the cancer from several directions so that the absorbed dose is concentrated in the cancer, while the dose to overlying tissues is spread over several entrance pathways. The technique of using x-ray beams at several orientations is referred to as a multiple-field treatment, and permits delivery of higher doses to the cancer without major complications caused by the irradiation of normal tissues. In addition, tissues such as the spinal cord, esophagus and rectum that are particularly sensitive to radiation often can be spared from direct exposure to radiation (Withers and Peters, 1988). Many linear accelerators also permit extraction of the electron beam for treatment of certain cancers by electrons rather than by x rays.

5.2 Treatment Planning

The delivery of high doses of radiation to cancers deep within the body is a challenging problem. Enough radiation must be delivered to kill all (or nearly all) of the cancer cells in patients for whom there is a reasonable prospect of cure. In some patients this prospect is remote at best, possibly because the cancer has already spread to one or more essential organs, or perhaps because the cancer is present in a critical organ and has already done so much damage that the patient cannot survive. Radiation plays a vital role in the treatment of these patients as well, because it can prevent further spread of the cancer, and often shrink its volume, thereby preventing or alleviating some of the agony that often accompanies cancer. This use of radiation is called palliative therapy. In a typical radiation therapy facility, about half of the patients are seeking a cure, and the others are undergoing palliative therapy.

With all radiation therapy patients, enough radiation must be delivered to the cancer to accomplish the objective of treatment. At the same time, great care must be taken to not expose normal tissues, especially those particularly sensitive to radiation, to an amount of radiation that will initiate severe consequences. Achieving this balance is the challenge of radiation therapy, and it is why the high-energy x rays, sharply defined beams, and versatile orientations provided by the linear accelerator are major advantages. It is also why a full course of radiation treatment requires several weeks. Normal tissues can tolerate more radiation if it is delivered gradually rather than all at once. This challenge is also why careful and precise planning of radiation treatments is essential.

Radiation treatments are planned in a series of steps that require the work of physicians, physicists, and technologists. First the patient is examined by the radiation oncologist, who may request additional imaging studies to complement those in the patient's file. The oncologist provides information about the patient's anatomy, including the location of the cancer and diagnostic images relevant to the case, to the medical physicist who enters the information into a computer. The physicist's responsibility is to generate a treatment plan, usually in consultation with the oncologist, that delivers an adequate amount of radiation to the cancer and preserves normal tissues to the maximum extent possible. The treatment plan is constructed on a computer in which the physical characteristics of a variety of radiation beams are simulated from data obtained by direct measurements on the treatment unit. These simulated data are positioned in the computer at the orientations proposed for the treatment beams, and composite curves are constructed to display the summation of radiation doses at each location in the patient. Analysis of the composite curves allows the oncologist and physicist to judge whether the proposed treatment plan is satisfactory.

Often a number of treatment plans are constructed for a patient before one is identified that satisfies the criteria for a satisfactory plan. The accepted plan is then implemented on a treatment-planning simulator that resembles the radiation therapy unit except that

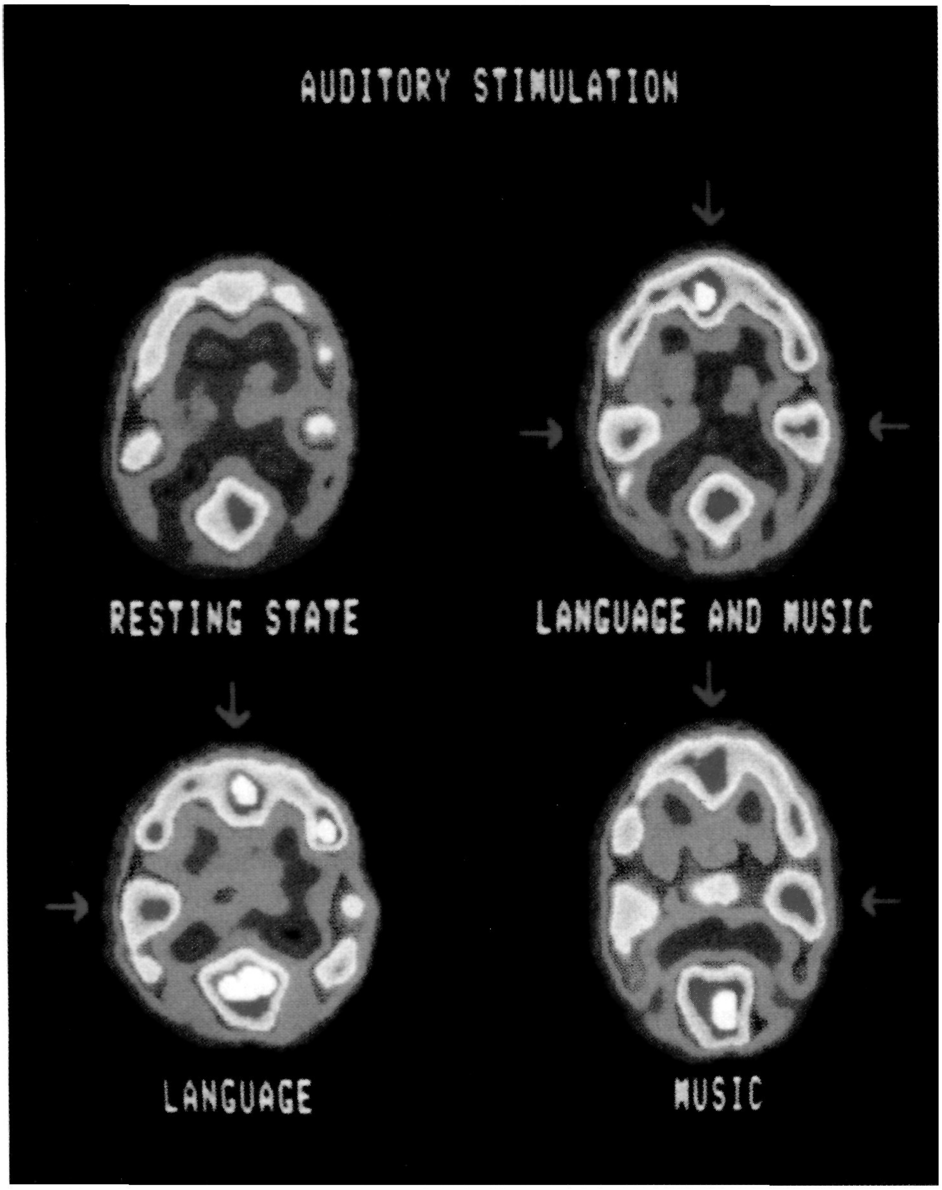
AUDITORY STIMULATION
RESTING STATE
LANGUAGE AND MUSIC
LANGUAGE
MUSIC

Vol. 2 – PLATE 1. PET images of the utilization of deoxyglucose in the brain during different auditory stimuli. (Courtesy of E. Hoffman, University of California, Los Angeles.)

the patient is exposed to diagnostic x rays rather than to the high-energy x rays used for treatment. Images obtained with the simulator are used to verify the acceptability of the treatment plan and to refine it where necessary. Once a plan judged acceptable by the oncologist and physicist has been constructed by iteration between the computer and simulator, treatment of the patient can begin. A computer-generated treatment plan for a radiation therapy patient is shown in Fig. 9.

5.3 Brachytherapy

For a small source of radiation, the intensity of radiation decreases as the square of the distance from the source, a property known as the "inverse square law." By placing small sources of gamma rays within tissue to be treated, large doses can be delivered to cancer cells in the immediate vicinity, with much smaller doses administered to tissues farther away. Of course, the tissue to be treated must be accessible for implantation of the sources. Regions of the body where gamma-ray sources can be implanted include the skin, oral cavity, vagina, uterus, and prostate. By far the most common application of the technique of "short-distance" or brachytherapy is its use for treatment of cancer of the uterus and vagina.

For many years radium was almost the only radioactive nuclide employed for brachytherapy, and the technique still is sometimes referred to as radium therapy. However, radium is almost never used today, because cesium-137 sources in the form of capsules and needles are safer to use and store. Also used occasionally are small seeds of iodine-125 for treatment of prostate cancer. These seeds are permanently implanted with a pistol-like impeller, where they deliver radiation as they decay with a half life of 60 days.

Careful dosimetry and treatment planning is as important for brachytherapy as it is for teletherapy. In treatment of cancer of the uterus, for example, a high dose of radiation is delivered over a period of a few days with a source applicator that is removed from the patient after the desired dose has been reached. The preferred time for removal is

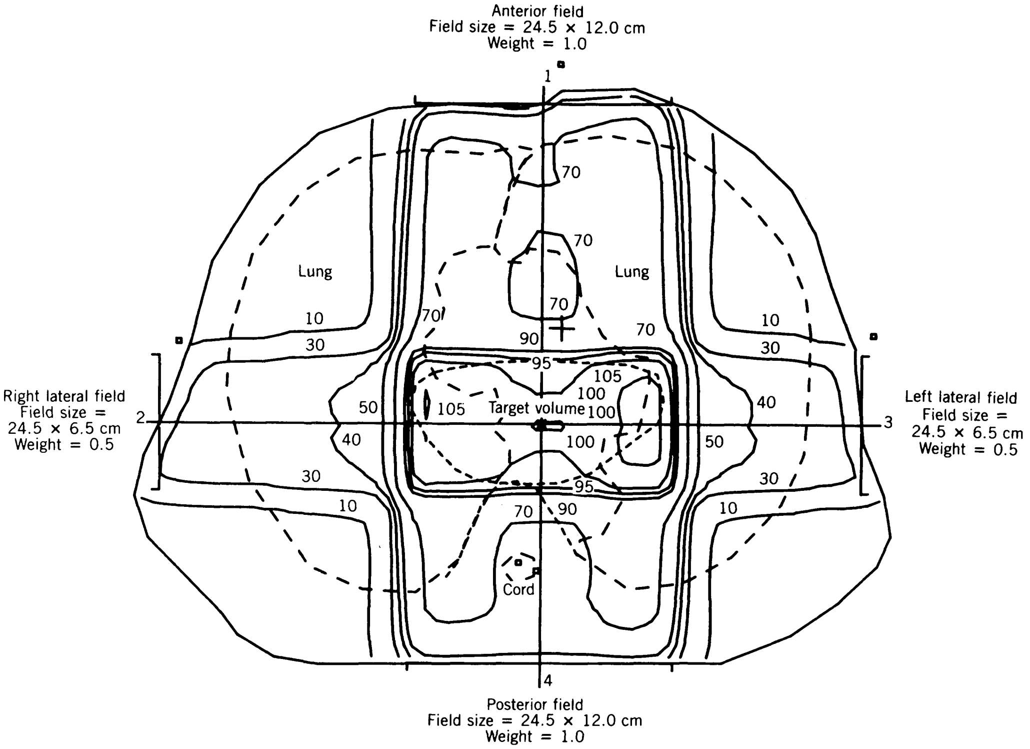
Anterior field
Field size = 24.5 x 12.0 cm
Weight = 1.0
1
70
70
Lung
Lung
10
30
70
70
90
70
10
30
95
105
100
Target volume
100
105
50
40
40
100
50
30
95
30
10
70
90
10
Cord
2
3
4
Right lateral field
Field size =
24.5 x 6.5 cm
Weight = 0.5
Left lateral field
Field size =
24.5 x 6.5 cm
Weight = 0.5
Posterior field
Field size = 24.5 x 12.0 cm
Weight = 1.0

FIG. 9. A computer-generated treatment plan for a patient with cancer of the esophagus. [From Mohan (1988).]

determined from computer-generated computations of the dose distribution estimated from images obtained with the applicator in place in the patient. After removal of the applicator, most patients are subjected to additional radiation from an external beam to treat cancer cells that may have spread to surrounding lymph nodes. This radiation is then added by computer to the dose delivered by the applicator to yield a composite dose distribution for the entire course of therapy.

5.4. Particle Therapy

Electron beams from a linear accelerator are particularly useful for treating relatively superficial cancers. These beams deliver a relatively uniform dose of radiation over a depth of a few centimeters corresponding to the range (penetration) of the electrons in tissue. The range of the electrons varies directly with the energy of the electrons, which can be varied within limits with most accelerators. Beyond the electron range, the radiation dose decreases very rapidly, and tissues beyond the treated region receive very little radiation. The dosimetry of electron beams is more complex than that for x and gamma rays and represents a continuing challenge to medical physicists.

Neutrons also are being studied for their possible use in radiation therapy. For example, neutron beams from cyclotrons are under investigation in a few institutions around the world for their effectiveness in treating a wide variety of cancers (Cohen and Awschalom, 1982). One potential advantage of neutrons is that they are more effective than x rays in killing poorly oxygenated cells that may be present in the central region of cancers. It is the survival of these cells that may be responsible for the occasional recurrence of cancers after their treatment with radiation. Preliminary results from the clinical trials of neutrons are promising, but definitive data have not yet been collected for most cancer sites.

Heavy charged particles (charged particles with a mass greater than that of an electron) also have been investigated for their possible uses in radiation oncology. These particles offer excellent treatment characteristics because of the manner in which they deliver energy to tissue. As a beam of heavy charged particles enters a patient, it produces only light ionization along its path, and the surrounding tissue receives a relatively low dose of radiation. Near the end of the trajectory of the charged particles, however, the rate of energy deposition increases markedly, creating a concentration of radiation dose at depth in a region known as the Bragg peak. The depth of the peak can be altered by varying the particle energy, and the width of the peak region can be changed by inserting a "range modulator" that continuously varies the amount of external absorber in the path of the charged particle beam. In this manner, the Bragg peak can be shaped to encompass cancer cells located at a specific depth and occupying a particular space in a patient. Beyond the Bragg peak, the dose from the charged particle beam rapidly falls to near zero. The characteristics of charged particles are illustrated in Fig. 10, in which the dose-delivery properties of helium ions are compared with those of 24 million volt x rays.

Heavy charged particles are described as high LET (linear energy transfer) radiations, because they produce dense ionization and deliver high radiation doses in the region of the Bragg peak. High LET radiation is especially effective in killing cells in the center of a cancer that are deprived of oxygen (so-called hypoxic cells) and that are relatively resistant to destruction by x and gamma rays. The oxygen enhancement ratio (OER) is the quotient of the radiation dose needed to produce a given effect (e.g., cell killing) in the absence of oxygen, divided by the dose required to produce the same effect in the presence of oxygen. Beams of charged particles have OERs near unity, whereas x and gamma rays

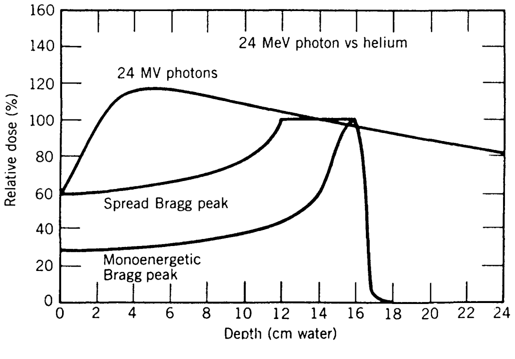

FIG. 10. Comparison of depth dose distributions for 24-MV x rays, monoenergetic, and spread Bragg peak helium beam. [From Chen (1988).]

exhibit OERs usually in the range of 2–3. Hence, only $\frac{1}{2}$–$\frac{1}{3}$ as much dose may be required from charged particles compared with x and gamma rays to reduce the fraction of surviving cells to a prescribed level. Many investigators believe that the biological advantages of heavy charged particles are even more important than their physical characteristics with respect to the application of charged particles to radiation oncology.

Heavy charged particles that offer potential advantages for radiation oncology include protons, helium ions, and accelerated nuclei of carbon, neon, silicon and argon. The earliest exploration of heavy charged particles for radiation oncology was the pioneering work of scientists at Berkeley (Tobias, Anger, and Lawrence, 1952). Since 1975, several hundred cancer patients have been entered into clinical trials with heavy charged particles at the Lawrence Berkeley Laboratory, and explorations of charged particles are currently underway in several other institutions, including Harvard University, University of Uppsala, United Kingdom Atomic Energy Authority, and the Institute of Clinical and Experimental Oncology in Moscow (Raju, 1980). Cancers being treated include malignant tumors of the brain, advanced cancers of the head and neck region, esophagus, stomach and pancreas, sarcomata arising in bone, choroidal melanomas of the eye, and juxtaspinal tumors (Chen, 1988). In addition, several nonmalignant diseases have been treated, including benign tumors of the pituitary and arteriovenous malformations (Verhey and Munzenrider, 1982).

6. FUTURE PROSPECTS

The use of ionizing radiation as a diagnostic probe to reveal abnormal conditions within the patient is an invaluable tool in medicine. In recent years it has been supplemented by important diagnostic advances such as ultrasonography and magnetic resonance imaging. Additional nonradiation technologies such as magnetic source imaging and electrical impedance tomography offer promise for the future. Although these advances complement the use of radiation for the production of diagnostic images, they are unlikely to replace more traditional imaging techniques such as radiography, fluoroscopy, and nuclear medicine. The most probable scenario for the near future is an increased digitization of traditional imaging techniques, and a gradual but continual movement towards integrated diagnostic imaging services networked through picture archiving and communications systems.

For decades self-professed clairvoyants have predicted the replacement of conventional radiation therapy by the use of cellular poisons, monoclonal antibodies labeled with radioactivity, and other "magic bullets" that will target cancer cells selectively for destruction while producing little damage to normal cells. These predictions have yet to be realized, principally because substances have not been developed that select only cancer cells with adequate sensitivity and specificity. Furthermore, many cancers are discovered only after they have grown and spread to such an extent that they produce clinical symptoms in the patient. These patients are less desirable candidates for the "magic bullet" approach to treatment, even if one were available. It seems safe to assume that radiation therapy will continue as an effective approach to cancer therapy for several years to come.

GLOSSARY

Absorber: Any material that absorbs or lessens the intensity of ionizing radiation.

Absorption: The process by which the number of particles or photons entering a body of matter is reduced or attenuated by interaction with the matter.

Alpha Particle: A positively charged particle ejected spontaneously from the nuclei of some radioactive elements. It is identical to a helium nucleus that has a mass number of 4 and an electrostatic charge of $+2$. It has low penetrating power and short range.

Attenuation: The process by which a beam of radiation is reduced in intensity when passing through some material. It is the combination of absorption and scattering processes.

Beta Particle: A charged particle emitted from a nucleus during radioactive decay, with a mass equal to $\frac{1}{1837}$ that of a proton. A negatively charged beta particle is identical to an electron. A positively charged beta particle is called a positron.

Decay, Radioactive: The decrease in the amount of any radioactive material with the passage of time, due to the spontaneous emission from the atomic nuclei of either alpha or beta particles, often accompanied by gamma radiation.

Detector: A material or device that is sensitive to radiation and can produce a response signal suitable for measurement or analysis. A radiation detection instrument.

Electromagnetic Radiation: A traveling wave motion resulting from changing electric or magnetic fields. Familiar electromagnetic radiations range from x rays (and gamma rays) of short wavelength, through the ultraviolet, visible, and infrared regions, to radar and radio waves of relatively long wavelength. All electromagnetic radiations travel in a vacuum with the velocity of light.

Gamma Ray: High-energy, short-wavelength electromagnetic radiation (a gamma packet of energy) emitted from the nucleus. Gamma radiation frequently accompanies alpha and beta emissions.

Geiger-Mueller Counter: A radiation detection and measuring instrument. It consists of a gas-filled tube containing electrodes, between which there is an electrical voltage but no current flowing. When ionizing radiation passes through the tube, a short, intense pulse of current passes from the negative electrode to the positive electrode and is measured or counted. The number of pulses per second measures the intensity of radiation.

Gray: A unit of absorbed dose in the International System of Units (SI), equal to 1 joule per kilogram.

Ionization: The process of adding one or more electrons to, or removing one or more electrons from, atoms or molecules, thereby creating ions. High temperatures, electrical discharges, or nuclear radiations can cause ionization.

Ionization Chamber: An instrument that detects and measures ionizing radiation by measuring the electrical current that flows when radiation ionizes gas in a chamber, making the gas a conductor of electricity.

Ionizing Radiation: Any radiation capable of displacing electrons from atoms or molecules, thereby producing ions. Examples: alpha, beta, gamma, and x rays, neutrons, and ultraviolet light. High doses of ionizing radiation may produce severe skin or tissue damage.

Photon: A quantum (or packet) of energy emitted in the form of electromagnetic radiation. Gamma and x rays are examples of photons.

Positron: Particle equal in mass, but opposite in charge, to the electron; a positive electron.

Rad: Acronym for radiation absorbed dose. A basic unit of absorbed dose of radiation.

Radiation: Particles (alpha, beta, and neutrons) or photons (gamma) emitted from the nucleus of an unstable (radioactive) atom as a result of radioactive decay.

Radioactivity: The spontaneous emission of radiation, generally alpha or beta particles, often accompanied by gamma rays, from the nucleus of an unstable isotope.

Radioisotope: An unstable isotope of an element that decays or disintegrates spontaneously, emitting radiation. Approximately 5000 natural and artificial radioisotopes have been identified.

Roentgen: A unit of exposure to ionizing radiation.

Scintillation Detector or Counter: The combination of phosphor, photomultiplier tube, and associated electronic circuits for counting light emissions produced in the phosphor by ionizing radiation.

Sievert: A unit, in the International System of Units (SI), of dose equivalent which is equal to 1 joule per kilogram.

Survey Meter: Any portable radiation detection instrument especially adapted for inspecting an area to establish the existence and amount of radioactive material present.

X Rays: Penetrating electromagnetic radiation (photon) having a wavelength that is much shorter than that of visible light.

List of Works Cited

Chen, G. T. Y. (1988), in: J. G. Webster (Ed.), *Encyclopedia of Medical Devices and Instrumentation,* New York: Wiley.

Cohen, L., Awschalom, M. (1982), *Ann. Rev. Biophys. Bioeng.* **11**, 359–390.

Curry, T. S., Dowdey, J. E., Murry, R. C. (1990), *Christensen's Introduction to the Physics of Diagnostic Radiology,* Philadelphia: Lea & Febiger.

Hendee, W. R. (1979), *Medical Radiation Physics*, 2nd ed. Chicago: Mosby-Yearbook Medical Publishers.

Hendee, W. R. (1981), *Radiation Therapy Physics*, Chicago: Mosby-Yearbook Medical Publishers.

Hendee, W. R. (1983), *The Physical Principles of Computed Tomography*, Boston: Little, Brown.

Hendee, W. R. [in press (A)], "Overview of Radiation Effects and Control," *Occup. Med. State-of-the-Art Rev.*

Hendee, W. R. [in press (B)], "Radiologic information overload and management," *Am. J. Roent.*

Mackay, R. S. (1984), *Medical Images and Displays: Comparisons of Nuclear Magnetic Resonance, Ultrasound, X-rays and other Modalities*, New York: Wiley.

Mohan, R. (1988), in: J. G. Webster (Ed.), *Encyclopedia of Medical Devices and Instrumentation*, New York: Wiley.

Phillips, V. M., Bernardino, M. E. (1988), in: C. E. Putnam and C. E. Ravin (Eds.), *Textbook of Diagnostic Imaging*, Philadelphia: Saunders.

Raju, M. R. (1980), *Heavy Particle Radiotherapy*, New York: Academic.

Saha, G. B. (1984), *Fundamentals of Nuclear Pharmacy*, 2nd ed. New York: Springer.

Sorenson, J. A., Phelps, M. E. (1987), *Physics in Nuclear Medicine*, 2nd ed. New York: Grune & Stratton.

Tobias, C. A., Anger, H. O., Lawrence, J. H. (1952), *Am. J. Roent.* **67**, 1–27.

Verhey, L., Munzenrider, J. (1982), *Ann. Rev. Biophys. Bioeng.* **11**, 331–357.

Withers, H. R., Peters, L. J. (Eds.) (1988), *Innovations in Radiation Oncology*, New York: Springer.

Wong, D. F. (1990), in: J. H. Anderson (Ed.), *Innovations in Diagnostic Radiology*, New York: Springer.

Zonneveld, F. W. (1987), *Computed Tomography of the Temporal Bone and Orbit*, Baltimore: Urban and Schwarzenberg.

Further Reading

Anderson, J. H. (Ed.) (1990), *Innovations in Diagnostic Radiology*, New York: Springer.

Hendee, W. R. (1979), *Medical Radiation Physics*, 2nd ed. Chicago: Mosby-Yearbook Medical Publishers.

Johns, H. E., Cunningham, J. R. (1983), *The Physics of Radiology*, 4th ed. Springfield, IL: Thomas.

Karzmark, C. J. (1984), *Med. Phys.* **11**, 105–128.

Saha, G. B. (1984), *Fundamentals of Nuclear Pharmacy*, 2nd ed. New York: Springer.

Sorenson, J. A., Phelps, M. E. (1987), *Physics in Nuclear Medicine*, 2nd ed. New York: Grune & Stratton.

Webb, S. (Ed.) (1988), *The Physics of Medical Imaging*, Philadelphia: Hilger.

Webster, J. G. (Ed.) (1988), *Encyclopedia of Medical Devices and Instrumentation*, Vols. 1–4. New York: Wiley.

BIOPHYSICS

Denis L. Rousseau and Lynn W. Jelinski, *AT&T Bell Laboratories, Murray Hill, New Jersey, U.S.A.*

Introduction 571

1. **Building Blocks of Biological Systems and their Hierarchical Assembly** 572

1.1 Simple Building Blocks 572

1.1.1 Nucleic Acids and the Central Dogma of Molecular Biology 573

1.1.2 Proteins 575

1.1.3 Lipids and Biological Membranes 576

1.1.4 Carbohydrates 578

1.2 Assemblies 578

1.2.1 Molecule-Molecule Interactions and Assemblies 578

1.2.1.1 Protein-Substrate Interactions and Catalysis 578

1.2.1.2 Protein-DNA Interactions 578

1.2.1.3 Protein-Protein Interactions 579

1.2.1.4 Lipid-Protein Interactions 581

1.2.2 Organelles 581

1.2.2.1 Pumps and Gradients: Bioenergetics in the Mitochondrion 581

1.2.2.2 The Nucleus 582

1.2.3 Cells 583

1.2.4 Organs, Organisms, and Systems 583

2. **Techniques of Biophysics** 584

2.1 Structural and Spectroscopic Techniques 584

2.1.1 X-Ray Crystallography 584

2.1.2 Nuclear Magnetic Resonance (NMR) 586

2.1.2.1 Nuclear Interactions 587

2.1.2.2 Nuclear Relaxation 588

2.1.2.3 Solid-State NMR 588

2.1.2.4 Two- and Three-Dimensional NMR 588

2.1.3 Electron Paramagnetic Resonance (EPR) 589

2.1.4 Vibrational Spectroscopy (IR, Raman) 589

2.1.5 Optical Techniques 590

2.1.5.1 Optical Absorption 590

2.1.5.2 Circular Dichroism 591

2.1.5.3 Fluorescence Spectroscopy 591

2.1.6 Extended X-Ray Absorption Fine Structure (EXAFS) 592

2.2 Electrophysiology 592

2.3 Imaging Techniques 594

2.3.1 Positron Emission Tomography (PET) 594

2.3.2 Magnetoencephalography (MEG) 594

2.3.3 Computed Tomography (CT) 595

2.3.4 Magnetic Resonance Imaging (MRI) 595

2.3.5 Scanning Tunneling and Atomic Force Microscopy (STM and AFM) 596

3. **An Example: Biophysics of the Visual System** 596

3.1 Morphological Structure 596

3.2 The Primary Event in Visual Excitation 597

3.3 Electrical Response Induced by Molecular Conformation Changes 600

3.4 Conversion of Electrical Signals into Visual Images: Motion and Color 601

3.5 The Primary Visual Cortex 602

3.6 Summary 603

Glossary 604

Works Cited 607

Further Reading 608

INTRODUCTION

Biophysics is the synthesis of chemistry, physics, and biology. Biophysicists elucidate the chemical and physical processes of biological molecules and systems through the application of physical techniques. Since biophysics is a diverse and broad field spanning the extremes of science, often what one considers biophysics is related to one's point of

digital Encyclopedia of Applied Physics

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

view. For example, a physicist may apply the principles of chaos to the cardiac cycle, whereas a chemist's interests may be along the lines of understanding how a small molecule such as oxygen binds to hemoglobin.

Biological molecules and systems are inherently complex. The quote by P. W. Anderson, "more is different" (Anderson, 1982), underscores the idea that new physics and new phenomena will emerge from the study of complicated molecules and systems that would go undiscovered by the study of simple systems alone. It is precisely this complexity that provides the intellectual challenge in biophysics.

Key questions in biophysics concern the way that biological information is encoded, processed, stored, and used, as well as the structure and function of biological molecules. The goal is to understand how biology implements, e.g., assembly, recognition, energy transduction, transport, distributed information coding, sensing, and perception. Biological information is stored and processed at several different levels, ranging from "simple" biological molecules such as DNA and strings of amino acids, to complicated organs such as the brain. Given the complexity of biophysics and the enormous range of science that it encompasses, an introductory article on biophysics cannot possibly cover all aspects of the field. Recognizing this dilemma, we have chosen to introduce the reader first to the building blocks of biology, then to the physical techniques used to study them, and finally to a detailed example of the application of biophysics to the visual system.

In this article, the building blocks are presented according to their biological hierarchy, beginning with the simplest molecules of life, the nucleic acids and the amino acids, and following their assembly into DNA (deoxyribonucleic acid) and proteins. These larger molecules do not function in isolation, but rather are incorporated into membranes and organelles such as the nucleus or the mitochondria. These organelles are the functional elements of cells, which are themselves assembled into organs. Each organ carries out its specific function in the organism. Finally, the organisms themselves form systems, such as bacterial colonies. Biophysics embodies the study of each level in this hierarchy as well as the information transfer between levels.

Many physical techniques provide unique windows into the biological structure and function of these various levels of hierarchy. Recognizing that each of these physical techniques cannot be covered in reasonable depth in this article, we describe them only briefly and tabulate their typical applications.

We have chosen a single example, the visual system, to illustrate the scope of biophysics. We will follow this system, beginning with its molecular assembly, to the rods and cones, and finally up to visual perception and processing of visual information. The visual system will thus be used to highlight selected techniques of biophysics.

1. BUILDING BLOCKS OF BIOLOGICAL SYSTEMS AND THEIR HIERARCHICAL ASSEMBLY

1.1 Simple Building Blocks

Biological systems are built from relatively few basic molecular units. The many diverse functions of complete molecules, organelles, and organs result from the composition, order, and arrangement of these small building blocks. It is for this reason that a major focus in biophysics has been to elucidate the structure of biological molecules in order to determine how a given structure leads to a specific biological function. To appreciate the contributions biophysical studies have made in the past and can make in the future on the understanding of biological molecules and processes, it is necessary to be familiar with these fundamental molecules of biology and to know how they are combined and assembled into complex systems that carry out the functions that are necessary for the survival of the organism. We first describe the simple building blocks and how they are combined to form the basic molecules of biology. We then show how these molecules of biology are assembled into living structures. Many fine reference books are available where in-depth information on biomolecules may be obtained. See, e.g., Alberts *et al.* (1983) and Stryer (1988).

1.1.1 Nucleic Acids and the Central Dogma of Molecular Biology. The existence and continuity of life and its diversity depend on the order of four simple bases which are the functional elements of DNA. These four bases (nucleotides) are joined to form a polynucleotide by a polymeric sugar-phosphate (see Sec. 1.1.4) backbone. The entire assembly is called a biopolymer. Owing to this very simple structure, illustrated in Fig. 1, polynucleotides are the most highly ordered class of biological macromolecules. In DNA there are four nitrogenous bases: adenine (A) and guanine (G), which are purine derivatives, and cytosine (C) and thymine (T), which are pyrimidine derivatives. The structure of DNA is formed by attaching one of these bases to a deoxy ribose sugar through a glycosidic linkage. The deoxy ribose is connected to phosphate groups at its 3′ and 5′ positions forming a backbone for the macromolecule. Specific base pairing (A to T, G to C) (see Sec. 1.2, Fig. 5) through hydrogen bonding permits DNA to exist in a double helix of complementary polymeric strands, although many single-stranded forms of polynucleotides also exist. The order (primary structure) of the base pairs determines the end products of protein synthesis and the arrangement (secondary structure) determines the efficiency of the synthetic process. Ribonucleic acid (RNA) is closely related to DNA. The sugar of RNA differs from that of DNA by having a hydroxyl group at the 2′ position rather than a proton. The bases of RNA are adenine, guanine, cytosine, and uracil instead of the thymine which is present in DNA.

The process by which the genetic information which is stored in the DNA is translated into a protein sequence has been termed by

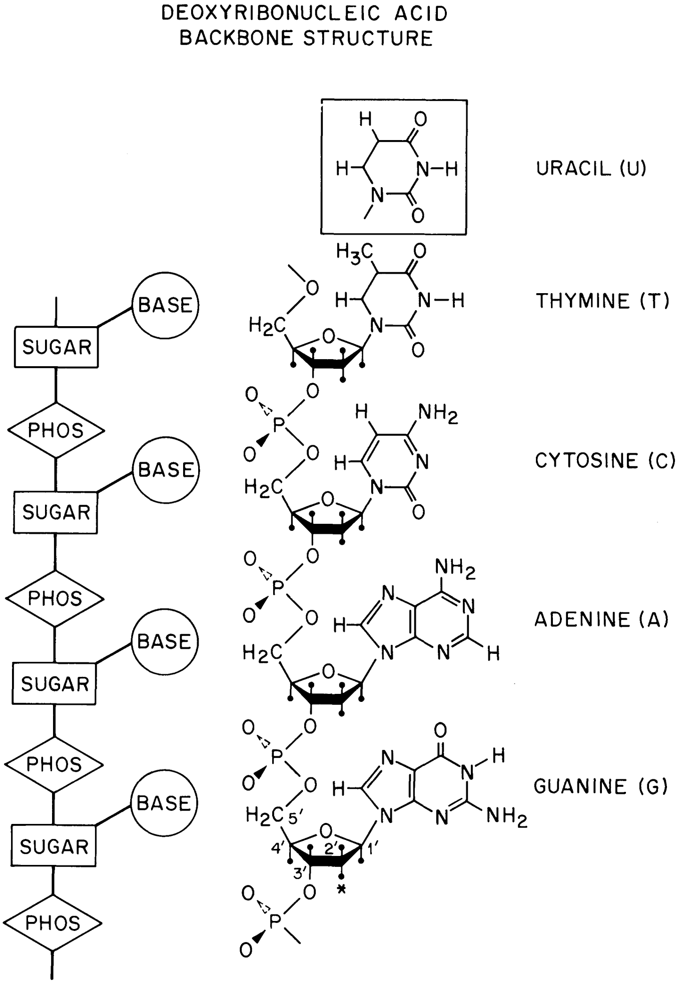

FIG. 1. Structure of polynucleotides. The black dots on the ribose (sugar) units of the molecular structure of deoxyribonucleic acid (DNA) represent hydrogen atoms. In ribonucleic acids (RNA) the 2′ position on each ribose, one of which is marked with an asterisk, has an OH group on it. Thymine, cytosine, adenine, and guanine are the bases in DNA and uracil replaces thymine in RNA.

Francis Crick as "the central dogma of molecular biology." It involves a series of steps known as replication, transcription, and translation (Watson, 1977). This process, illustrated in Fig. 2, involves the flow of information from DNA of the gene to a series of RNAs, to the formation of protein sequences in an organelle within the cell called a ribosome. In all eukaryotic species (all plants and animals other than bacteria and certain algae) the DNA of the gene is sequestered in a membrane-bound organelle (the nucleus) and is wrapped around a protein structure (a histone). It is the interaction between DNA and the histone that allows the very long strands of DNA, nearly 1 m long in the human genome, the full set of chromosomes, to fit into the very small nucleus. Biophysical studies using electron microscopy have played an essential role in identifying and determining the organization of the strands of DNA in the various organelles. Furthermore, the structure of DNA first reported by Watson and Crick (1953) was the triumph that ushered in the era of molecular biology. Biophysical studies have played a central role in determining the complex series of events which comprise the "central dogma." The techniques of x-ray crystallography and electron microscopy have been essential in determining the structures of DNA, the many RNAs that are part of the information transfer, and the protein-polynucleotide complexes which are crucial elements of the pathway.

The transcription process (Fig. 2) involves transcribing the genetic information in the DNA to RNA. In this process three different types of RNA are formed: messenger RNA (mRNA), which contains the genetic message, ribosomal RNA (rRNA), which is a component of the ribosomes, a set of oganelles in which protein synthesis takes place, and transfer RNA (tRNA), which transports amino acids to the ribosomes for incorporation into proteins. The "workbench" for protein synthesis is the ribosome where the mRNA translates its message into the protein sequence. Translation of genetic information into the protein sequence depends on the genetic code (i.e., a mapping from the sequence of bases in RNA to specific amino acid residues). The discovery of this code is one of the great achievements of molecular biology (it can be considered its Rosetta Stone). It was found that a sequence of three bases codes for a specific amino acid. For example, AUG codes for methionine and GCU, GCC, GCA, and GCG all code for alanine. Thus, the protein sequence information is completely stored in the primary structure of the polynucleotides. These sets of three bases that code for specific amino acids are known as "codons."

All the proteins needed by the organism are synthesized on the ribosomes. The mechanism by which a polypeptide, generated in the ribosome, folds into a biologically active protein is very complex and is currently a

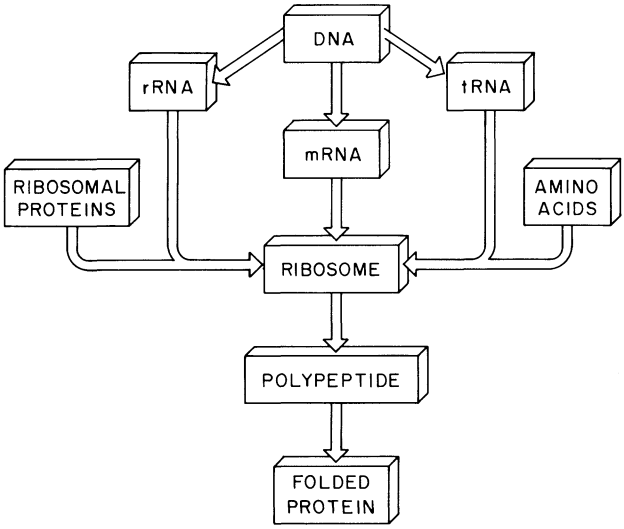

FIG. 2. The flow of genetic information from DNA to assembled proteins.

topic of intensive experimental and theoretical biophysical studies. Indeed, the protein folding problem, to be discussed in more detail below, has been termed the second half of the genetic code. Another area of extensive study involves the controlling mechanisms by which protein synthesis is initiated and terminated.

1.1.2 Proteins. Proteins, the product of genetic translation, form the foundation, structure, and machinery of life. In order for proteins to carry out their diverse functions they have structures that are far more complex than the polynucleotides. Furthermore, whereas the polynucleotides have only 4 bases, proteins are comprised of 20 amino acids, each with different characteristics. Proteins with vastly different properties may be generated, depending upon how the amino acids are combined in a polypeptide chain. Many proteins function as enzymes, i.e., they catalyze a specific chemical reaction. There is an enormous number of chemical reactions that are constantly occurring in any living system. In the absence of the catalytic properties of enzymes the progress of these reactions would be so slow that life could not exist. Enzymes speed up these reactions by 1000-fold or more. To have the specificity needed to function in only selected chemical reactions, both the diversity of the amino acid properties and the final three-dimensional structure are fully utilized. For example, some proteins are membrane-bound (see Sec. 1.1.3). These proteins have hydrophobic amino acid residues on their exterior that allow them to interact with the lipids of the membrane. Enzymes exhibit high specificity, since they possess very specific active sites each of which will react with only a single type of molecular group. This function is achieved by forming a structure, called an active site, that contains a specific set of amino acids according to their size, charge, and hydrophobicity. Certain proteins also incorporate special ligands or chromophores within their structures, such as the retinal group of rhodopsin or the heme group of hemoglobin.

Proteins display several levels of structural organization. The first level is the primary structure, which is the sequence of the amino acid residues comprising the protein (see Fig. 3). The amino acid residues are joined by peptide bonds between the amine group of one amino acid and the carboxylic acid group of another. The secondary structure involves local interactions between the adjacent amino acids of the chain. The most frequently found secondary structures are alpha helices (Fig. 3) and beta sheets (so named because the α helix was the first structure identified and the β sheet was the second). As their names imply, in the alpha-helical structure the polypeptide chain is coiled into a helix, whereas in the beta-sheet configuration the polypeptide chains are arranged into a pleated sheet-like configuration. The configuration that a given amino acid sequence adopts is determined by both the amino acid sequence of that segment and its overall environment. The tertiary structure of a protein involves the long-distance interactions in the protein and the spatial arrangement of the helices and sheets with respect to each other. Many forces are involved in determining the tertiary structure, including hydrogen bonding, van der Waals interactions, and electrostatic forces. The complexity and long-range nature of these forces makes prediction of the tertiary structure of a protein from its amino acid sequence extremely difficult. This is currently a very active area of biophysics research. Its full understanding will lead to a knowledge of the interactions that hold proteins together as well as a tool for predicting structures and thereby tailoring proteins for specific needs. The final level of organization of proteins lies in the quaternary structure. This is the relationship between independent polypeptide chains in proteins which have multiple subunits, such as hemoglobin. In the case of hemoglobin small changes in the relationship of the subunits can significantly affect the functioning of the local active site of the protein. The way long-range structure can influence local functional properties is another area of intense biophysical research.

In the past, inferences concerning the influence of particular amino acids on protein function could not be tested readily. However, since the development of recombinant DNA technology, site-directed mutagenesis has allowed the replacement of one amino acid by another, almost at will. Thus, the influence of specific amino acids can be tested directly. This development has significantly changed the field of biophysics by allowing many qualitative inferences to be tested spe-

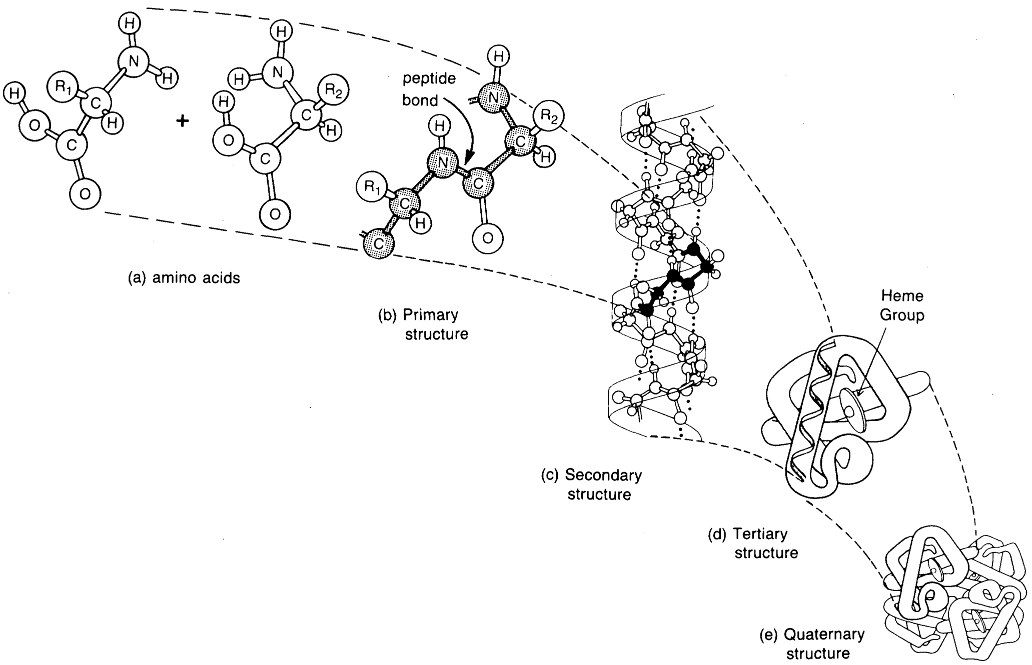

FIG. 3. The levels of protein structure. Amino acids (a) form a peptide bond (b) through the condensation of an amine group with a carboxylic acid. The order of the amino acids in the chain constitutes its primary sequence. The secondary structure (c) pertains to the formation of specific ordered but local structures such as helices, β sheets, or loops. The tertiary structure (d) is the three-dimensional relationship between amino acids, helices, and sheets that are far from each other in the sequence. For those proteins which have more than one subunit, the arrangement of the subunits is referred to as the quaternary structure (e). The illustration in the figure refers to hemoglobin in which there is a heme group in each of the four subunits. The assembled tetramer comprises the hemoglobin molecule. [Adapted from Becker (1986), p. 86.]

cifically and often quantitatively (e.g., see Nagai *et al.*, 1987).

1.1.3 Lipids and Biological Membranes. A unique feature of living systems is that they are always operating out of equilibrium. This aspect of living systems is necessary so that the many ionic and molecular gradients that are used to generate chemical energy are maintained. The gradients are maintained by lipid bilayer membranes that encapsulate cells and organelles. These membranes are loaded with proteins that shuttle various ions and molecules across the membrane to maintain the needed gradients. These proteins may passively transport specific ions and molecules across the membrane by opening and allowing them to pass in the direction of the concentration gradient, or the proteins may actively translocate ions or molecules across the membrane by utilizing energy to transport ions against a gradient. The understanding of how such proteins function is currently a topic of extensive theoretical and experimental research.

Lipids are the structural elements of membranes. Most lipids are rod-like molecules, one end of which is hydrophobic and the other end of which is hydrophilic (see Fig. 4). Consequently, it is energetically favorable for lipids to assemble into a bilayer structure. In this bilayer all of the hydrophilic sections of the lipid are on the outside and the hydrophobic tails are on the inside, thereby forming a membrane which is about 75 Å thick. The spontaneous permeability of molecules and ions through the layered membrane barrier is very small. Such membranes enclose cells (plasma membranes) and enclose organelles within eukaryotic cells.

The most prevalent biological lipids are phospholipids. The structure of a typical phospholipid, phosphatidylcholine, is shown in Fig. 4. It consists of fatty acid hydrocarbon tails and a polar "head group" made up of

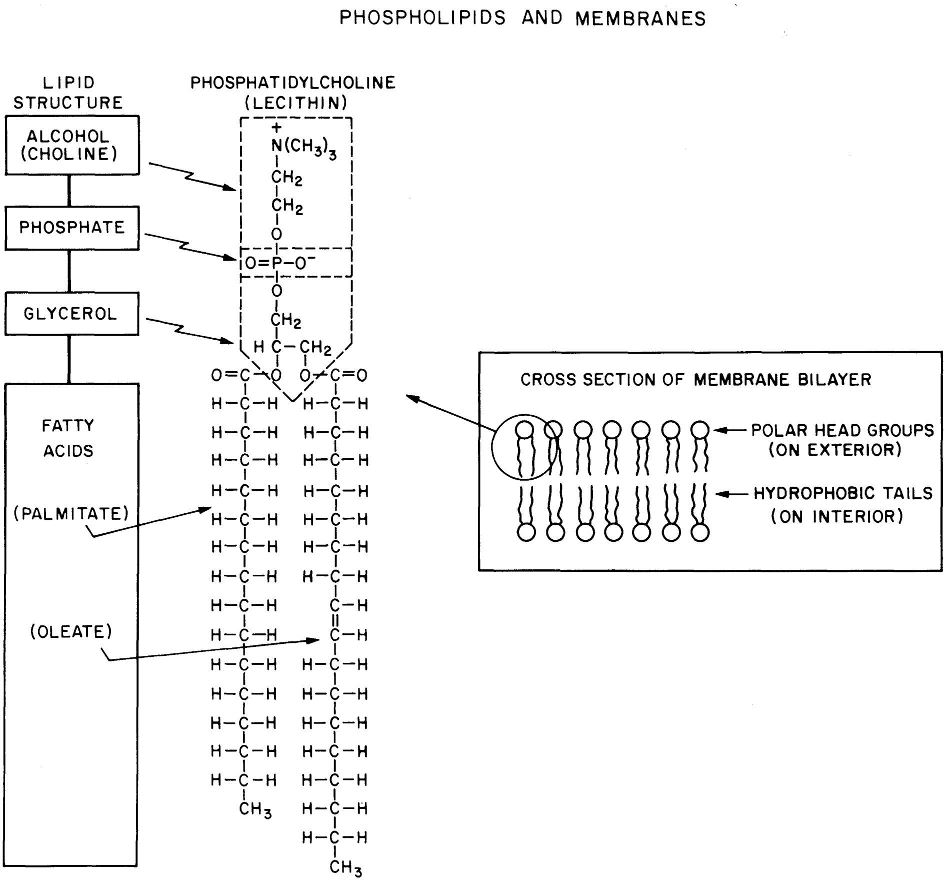
PHOSPHOLIPIDS AND MEMBRANES
LIPID STRUCTURE
ALCOHOL (CHOLINE)
PHOSPHATE
GLYCEROL
FATTY ACIDS
(PALMITATE)
(OLEATE)
PHOSPHATIDYLCHOLINE (LECITHIN)
N(CH3)3
CH2
CH2
O
O=P-O⁻
O
CH2
H C-CH2
O=C-O O-C=O
H-C-H
C-H
CH3
CROSS SECTION OF MEMBRANE BILAYER
POLAR HEAD GROUPS (ON EXTERIOR)
HYDROPHOBIC TAILS (ON INTERIOR)

FIG. 4. Structural features of phospholipids. Lecithin (a phosphatidylcholine) is illustrated as an example. Biological membranes consist of lipid bilayers in which the head groups are on the exterior and the hydrophobic tails are on the interior.

glycerol, phosphate, and an alcohol. Different alcohols are present in various lipids. Although the structural element of the membrane is the lipid molecule, biological membranes contain high levels of protein. Even myelin, which is an insulator around nerve cells, contains about 20% protein. Other membranes such as the inner mitochondrial membrane contain 75% protein. These proteins carry out extensive biological functions.

Whereas membranes serve as the structural elements, the proteins associated with them carry out many important biological functions. Proteins in membranes serve as channels, pumps, gates, receptors, structural anchors, identity markers, and enzymes. For example, in the mitochondrial inner membrane the chemical energy resulting from the burning of fuels, such as carbohydrates, is used to generate ATP (adenosine triphosphate), the energy currency molecule. ATP supplies the energy for chemical reactions, protein synthesis, and the generation and maintenance of molecular and ionic gradients throughout the organism. Diffusional properties are key features of the membrane, and are necessary for the proper function of many proteins. Within the plane of the membrane, lipid diffusion is very fluid. This is termed lateral diffusion and has been studied extensively by biophysical techniques, especially fluorescence. The rate of lateral diffusion of lipids in membranes is in the range of 2 μm/s. Some proteins have diffusion rates in membranes that are similar to that of lipids whereas others have much slower rates. For example, the diffusion rate of the photoreceptor protein molecules (rhodopsin) in the membrane is ~ 1 μm/s, very similar to that of lipids (Stryer, 1988). The flip-flop (transverse diffusion) of lipids or proteins from one side of the membrane to the other is very slow for lipids (rate about one per several hours) and has not been detected for proteins. This is an essential feature for the functional aspects of proteins in lipid membranes since many of the functions depend on the directionality of the protein in the membrane. This is true in energy generation, e.g., in which the production of ATP is driven by a proton gradient across a membrane.

1.1.4 Carbohydrates. The final class of biomolecules is that of the carbohydrates. Common carbohydrates include sugars, starches, and cellulose. Indeed, these biomolecules are so abundant that they make up most of the organic matter on earth. Carbohydrates, as the name implies, consist of carbon, hydrogen, and oxygen, with the hydrogen and oxygen in the atomic ratio 2:1, and serve many different roles. Sugar, stored as starch in plants and as glycogen in animals, is the fuel used in metabolism in the generation of energy. Cell walls of plants are composed of cellulose, another polymer of carbohydrates. Finally, as discussed above, the main structural element of nucleotides is a pentose unit, a five-carbon sugar. In nucleotides a base is attached to the pentose which is then attached to a phosphate. This forms the basic structural unit of all nucleic acids as well as a component of important coenzymes (e.g., NADH), metabolic regulators (e.g., GMP which will be discussed in a subsequent section), and energy storage molecules such as ADP and ATP.

1.2 Assemblies

DNA, proteins, and membranes self-assemble to form a hierarchy of structures, leading from the individual molecules to subcellular organelles, to cells, and finally to organs and the organism (see Fig. 5). Biophysics is concerned with both the individual and collective properties of these assemblies, as well as with the physical aspects of the assembly process. Biophysical studies have contributed greatly to our present understanding of the basic biology and function of these systems.

1.2.1 Molecule-Molecule Interactions and Assemblies

1.2.1.1 Protein-Substrate Interactions and Catalysis. Enzymes are proteins that carry out catalytic functions (see CHEMICAL REACTIONS; CATALYSIS). Generally they bind a substrate such as oxygen or a peptide, and chemically transform the substrate by oxidation, reduction, addition, or cleavage. Enzymes demonstrate highly selective recognition and binding as they carry out their catalytic functions. These reactions are in fact so selective that biocatalysis is used in some industrial processes, particularly involving the production of optically active materials. Understanding the recognition and binding events is an area of great research interest in biophysics.

The following example illustrates how genetic engineering and rate constant measurements can be used to uncover insights into protein-substrate recognition and binding events. Estell *et al.* (1986) studied the steric and hydrophobic effects involved when the enzyme subtilisin carries out its catalytic function of cleaving peptides. Using genetic engineering, they produced different mutant enzymes by replacing the conserved glycine of residue 166 with 12 different nonionic amino acids. They found that the catalytic efficiency toward small hydrophobic substrates was actually *increased* by as much as 16-fold over the nonmodified ("wild-type") enzyme when the glycine in position 166 was replaced with hydrophobic amino acids. Although the enzyme has evolved to be an efficient catalyst for a broad range of molecules, this result shows that the catalytic efficiency toward a *specific* molecule can be improved by genetic engineering. Protein engineering is being increasingly used to establish the basic biophysics of enzyme specificity (Carter and Wells, 1987).

The production and evaluation of bio-mimicking molecules is a completely different approach to understanding biological catalysis (Rebek, 1987; Cram, 1988). Here, the idea is to synthesize molecules that possess the minimum requirements for molecular recognition. Bio-mimicking molecules are often designed to test hypotheses based on our understanding of the biophysics of recognition and catalytic events.

1.2.1.2 Protein-DNA Interactions. The binding of proteins to DNA is essential in the

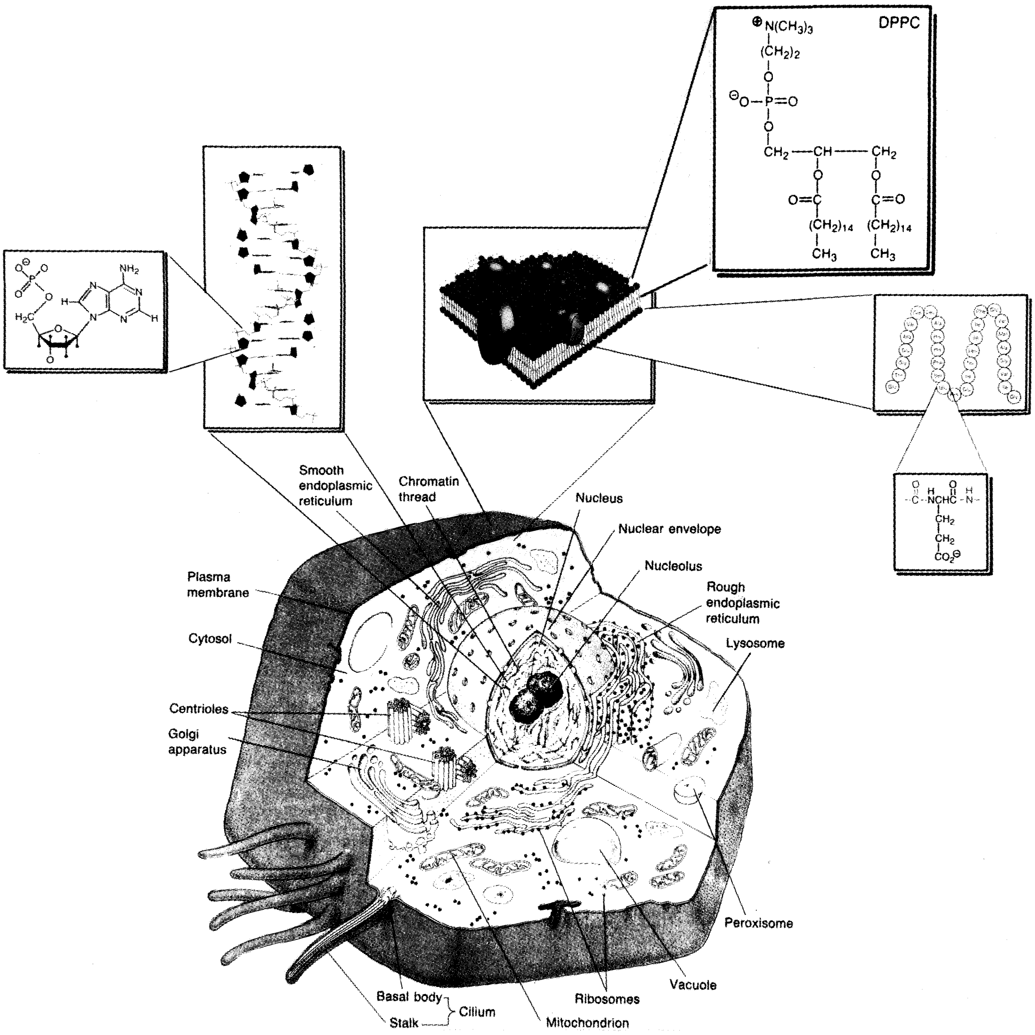
DPPC
Smooth endoplasmic reticulum
Chromatin thread
Nucleus
Nuclear envelope
Nucleolus
Rough endoplasmic reticulum
Lysosome
Plasma membrane
Cytosol
Centrioles
Golgi apparatus
Peroxisome
Vacuole
Basal body
Stalk
Cilium
Ribosomes
Mitochondrion

FIG. 5. Diagram of a "typical" non-neuronal animal cell showing various organelles. The DNA (upper left), contained in the nucleus, is shown at the molecular and double helix levels. The membrane (upper right) and its associated proteins are shown at the molecular and supramolecular levels. [Adapted from Mason (1983), p. 52.]

function of organisms. For example, DNA that has been damaged by environmental chemicals, ultraviolet light, or ionizing radiation is continually being repaired by a series of enzymes that recognize the damaged DNA and repair it in a stepwise fashion. Furthermore, the binding of various proteins to DNA is thought to control the genetic expression of various gene products, ultimately turning on or turning off the gene. Some DNA-binding molecules are flat and they intercalate between the base pairs; others interact via hydrogen bonding, electrostatic, or van der Waals interactions (Dervan, 1986). Since the base-pair sequence controls the three-dimensional shape of DNA, including the helical twist, groove shape, and bending, the design and testing of sequence-specific DNA-binding molecules is an active area of biophysical research.

1.2.1.3 Protein-Protein Interactions. One of the most well-known examples of protein-protein interactions is how proteins recognize foreign substances. Antibodies are protein molecules, produced by the body's defense system, that recognize and bind specific target molecules. The target molecule, an antigen, can be either a foreign chemical substance or protein. The principles of antibody-antigen recognition are widely applied in diagnostic kits and in immunocytochemical staining of cells. Biophysical studies of antibody specificity have been advanced by x-ray crystallographic studies. One such study is of an antibody bound to its protein antigen, hen egg white lysozyme, a protein that cleaves the polysaccharides of bacterial cell walls. In this structure the two proteins, the antibody and lysozyme, interact at two almost completely flat surfaces with small indentations where amino acids from one protein penetrate the structure of the other (Marrack, 1987). For antigenic sites on another protein, myohemerythrin, x-ray crystallographic studies and work with small peptides suggest that the primary event, recognition of highly exposed side chains on myohemerythrin, is coupled to protein movements, thereby allowing the antibody to interact with otherwise buried side chains of the protein (Getzoff *et al.*, 1987). Thus, one proposed antibody-protein interaction mechanism involves recognition, induced fit, and binding, a mechanism which bears similarities to the induced-fit mechanisms suspected in enzyme-substrate and DNA-protein interactions.

Another well-known example of protein-protein interactions involves assembly of the structural protein, collagen. The type I collagen molecule is composed of two distinct types of chains: two $\alpha 1(\mathrm{I})$ chains and one $\alpha 2$ chain, that have the triplet amino acid repeat (gly-x-y), where x and y are often proline and hydroxyproline. (Here the α designator is purely historical and does not refer to the type of helix.) The presence of glycine as every third residue of the chains allows the chains to pack in a triple-helical structure [Fig. 6(a)]. Because of steric constraints, only the glycine residues occupy positions within the core of the molecule, where they participate in interchain hydrogen bonds that stabilize the helix. The side chains of other amino acids occupy positions on the surface of the molecules. The collagen molecule is approximately 3000 Å long and 15 Å in diameter. Its unusual rod-like shape makes it a favorite for biophysical light-scattering measurements.

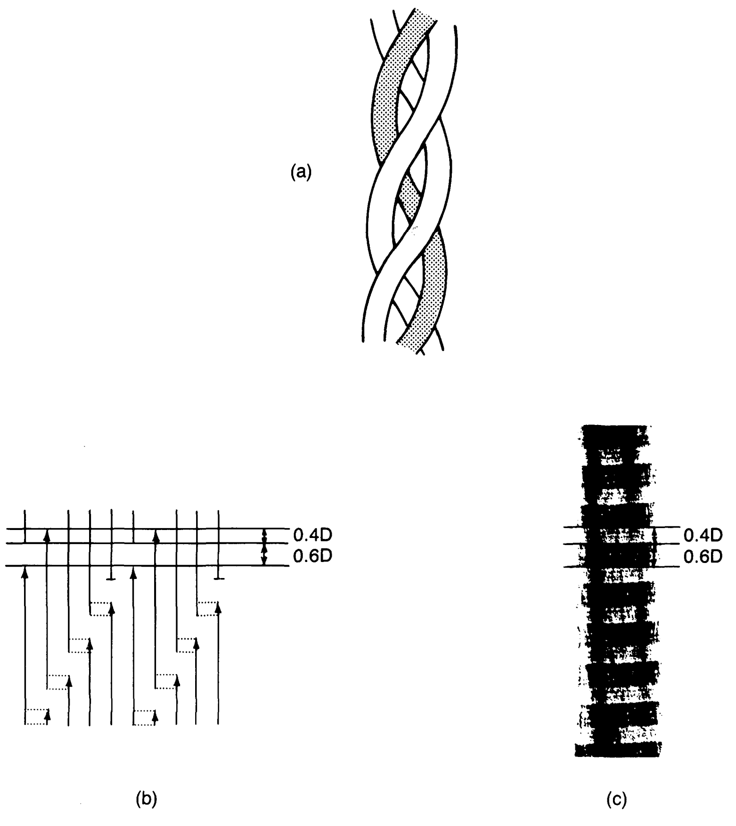

FIG. 6. An illustration of protein-protein assembly. (a) The type I collagen molecule is composed of two $\alpha 1(\mathrm{I})$ chains and one $\alpha 2$ chain (shaded) that form a triple-helical structure. (b) These triple helices pack so that the side chains (not shown) occupy the surfaces of the molecule. The molecules assemble in a highly regular fashion to form well-organized fibrils. They pack with a *D* stagger of 668 Å or 234 residues, forming gap and overlap regions. Probable intermolecular cross-link regions are indicated for some molecules by dotted lines. (c) An electron micrograph of reconstituted rat tail tendon collagen negatively stained with 2% sodium phosphotungstate. The dark areas correspond to the gap regions and the light areas to the overlap regions. [Adapted from Jelinski and Torchia (1980).]

The collagen molecules pack in a highly regular fashion, staggered by the D repeat of 668 Å, or 234 residues [Figs. 6(b) and 6(c)]. (The D repeat describes the side-by-side stagger between the heads of adjacent collagen molecules.) The biophysics underlying this highly regular protein-protein assembly was uncovered, in part, by analysis of the distributions of interacting residues along the axial projection of the molecule. Fourier analysis of the collagen $\alpha 1(\mathrm{I})$ and $\alpha 2$ chain sequences showed that the distribution of amino acids having charged or large hydrophobic side chains has regularities which are related to the D repeat. These regularities in the sequence explain various features of the three-dimensional packing of collagen molecules into fibrils (Piez and Trus, 1978).

1.2.1.4 Lipid-Protein Interactions. Although many proteins and enzymes reside inside cells in the cytosol, there are also integral membrane proteins. These proteins are actually part of the membrane and are specific to the function of the membrane. Because of lipid-protein interactions, integral membrane proteins have a very different structure than water soluble proteins. Instead of a hydrophilic exterior, they usually consist of a bundle of alpha helices with hydrophobic amino acid residues on the exterior. The protein structure is stablized by the interaction of these residues with the hydrophobic tails of the lipids on the interior of membranes. Membranes support a wide range of functions, including respiration, vision, photosynthesis, cell-cell recognition, and nerve impulse transmission. The middle inset in Fig. 5 shows a schematic representation of a membrane containing various membrane-bound proteins. Some of these proteins span the entire membrane while others face either side of the membrane. Since the lipid component is an electrical and chemical insulator, the proteins provide the main means for moving ions, molecules, or protons across the membrane. Although the protein content varies from membrane to membrane, it is generally very high.

Channels are membrane proteins that contain pores spanning the membrane. Channels have many purposes, including the regulation of gradients of ions such as Na^+, K^+, and Ca^{++}, maintaining intracellular pH, and transporting metabolites. Channel proteins are highly specific. For example, the acetylcholine channel, which functions in neural transmission, undergoes a conformational change and opens only after two molecules of the neurotransmitter, acetylcholine, have bound to the extracellular surface of the protein. Once open, it passes sodium and potassium ions.

Many theoretical and experimental biophysical studies are concerned with the diffusion of the lipid component in the presence of the protein obstructions; others are concerned with the motion of the proteins themselves. Nuclear magnetic resonance (NMR) spectroscopy has been particularly useful in probing the microscopic details of the dynamic structure of membrane lipids and proteins (Smith and Oldfield, 1984).

1.2.2 Organelles. Organelles are subcellular assemblies that carry out specified functions. Cells contain many different organelles (as shown in Fig. 5). Since it is beyond the scope of this article to describe all organelles, we will concentrate on two: the mitochondrion and the nucleus.

1.2.2.1 Pumps and Gradients: Bioenergetics in the Mitochondrion. Biological systems need energy to carry out their basic functions of protein synthesis, muscle contraction, and ion transport. They get this energy from hydrolysis of the molecule adenosine triphosphate, or ATP. ATP stores a large amount of free energy, which is liberated when its phosphate bonds are hydrolyzed to produce ADP (adenosine diphosphate) or AMP (adenosine monophosphate). This liberated energy is used to drive otherwise unfavorable reactions. Biological systems would "run down" once all of the ATP was utilized unless there were mechanisms to drive the "uphill" synthesis of new ATP molecules. *Bioenergetics* is concerned with how energy is transduced from the oxidation of substances or by the absorption of light to drive the synthesis of ATP.

Most energy transduction takes place in specialized energy-transducing membranes in organelles such as chloroplasts or mitochondria (Fig. 7) (Nicholls, 1982). The organelles within the cell also have membranes; e.g., the mitochondria have two membranes, an outer membrane that is very permeable, and an inner membrane, where respiration takes place. These membranes have many distinct protein assemblies including two which

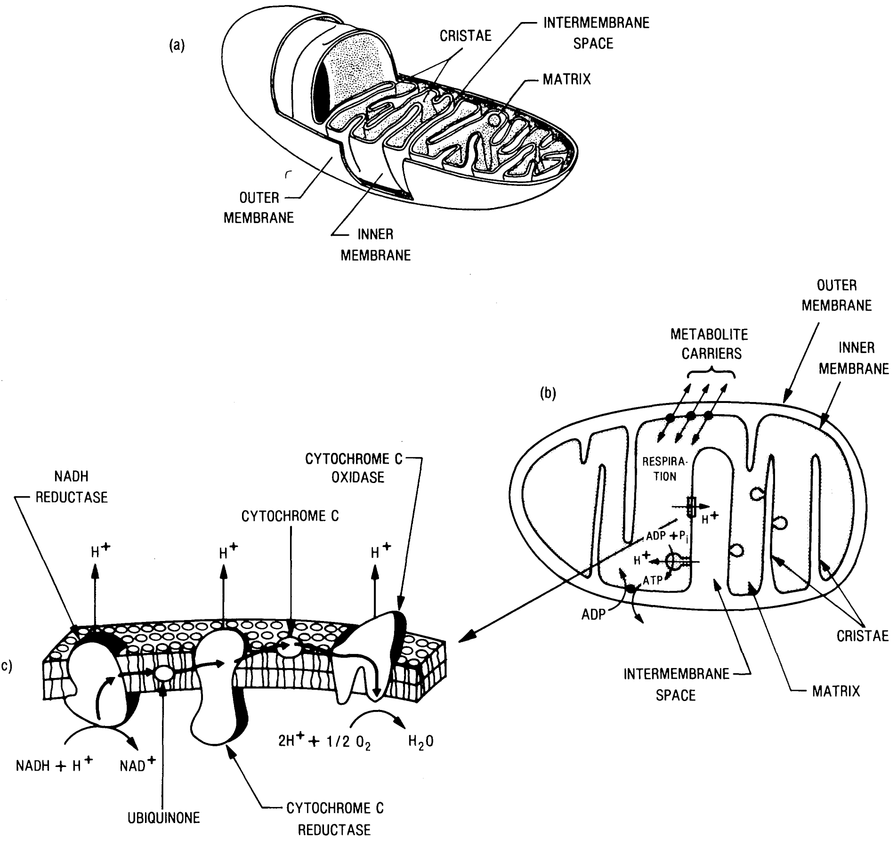
(a)
CRISTAE
INTERMEMBRANE SPACE
MATRIX
OUTER MEMBRANE
INNER MEMBRANE
(b)
OUTER MEMBRANE
METABOLITE CARRIERS
INNER MEMBRANE
RESPIRATION
H+
ADP +Pi
H+
ATP
ADP
CRISTAE
INTERMEMBRANE SPACE
MATRIX
(c)
NADH REDUCTASE
CYTOCHROME C OXIDASE
CYTOCHROME C
H+
H+
H+
2H+ + 1/2 O2
H2O
NADH + H+
NAD+
UBIQUINONE
CYTOCHROME C REDUCTASE

FIG. 7. (a) Schematic representation of a mitochondrion. The mitochondrion is an organelle typically 0.7–1 μm in length. [Adapted from Stryer (1981), p. 308.] (b) A cross-sectional view of a mitochondrion, showing various energy-transducing reactions. [Adapted from Nicholls (1982), p. 6.] (c) A close-up view of the protein assemblies involved in respiration and electron transfer.

are involved in the generation of ATP. One is responsible for ATP synthesis, the "uphill" reaction. The other, the "downhill" reaction, uses respiration as its energy source. A proton gradient across the membrane, the proton electrochemical potential $\Delta\mu_{H^+}$, provides the link between these two protein assemblies. The coupling between these pumps is shown in Fig. 7(b). ATP synthesis is driven by the proton gradient (bottom part of figure). ATP is continuously removed into the cytosol to provide energy for cellular function, while the proton gradient is continuously replenished (top part of figure) by respiration. Consequently, the proton gradient is an interconvertible source of free energy.

The details of ATP synthesis and of mitochondrial respiration are far beyond the scope of this article. Figure 7(c) provides a glimpse into this complexity; it shows some of the different proteins in the membrane and the electron transfer events that comprise the proton pump. Electrons enter (on the left) via NADH (the reduced form of nicotinamide adenine dinucleotide, or NAD^+), one of the major electron carriers in the oxidation of metabolic fuels. Electrons are transferred from NADH (dark lines and arrows) as protons are pumped across the membrane. The electrons are transferred along this chain of proteins until they reach the final enzyme, cytochrome *c* oxidase, where the electrons are used to reduce oxygen to water, with the concomitant pumping of another H^+. The membrane potential thus produced is used to drive the synthesis of ATP.

1.2.2.2 The Nucleus. A eukaryotic cell, by definition, contains a nucleus (Fig. 5). The nucleus is surrounded by a membrane called the nuclear envelope and houses most of the cell's DNA. We have already seen (Sec. 1.1) that DNA winds around proteins called histones. The DNA-histone assembly is called a *nucleosome;* nucleosomes are the repeating units of chromatin. The structure of chromatin has been elucidated, in part, by electron microscopy and x-ray and neutron diffraction. The nucleolus is also a highly organized compartment of the nucleus, and is the part of the nucleus that is most readily apparent when a cell is observed under a light microscope. The nucleolus contains RNA and proteins and functions to synthesize and assemble ribosomal RNA.

1.2.3 Cells. Cells have an amazing diversity of function, size, and shape. Some organisms such as the bacterium *E. coli* and certain algae are comprised of single cells. Other cells have highly specialized functions. For example, red blood cells carry oxygen, muscle cells participate in locomotion, and nerve cells form the brain's complex electrical signalling system. Nerve cells differ somewhat from the "typical" eukaryotic cell that we have already examined. Figure 8 shows a typical neuronal cell with several of its synapses. The dendrites bring information toward the cell body and the axons carry it away. Although neuronal cells have all of the genetic material that other cells have, they do not divide. Information from other cells is communicated via chemicals that are released at the synapses (Stevens, 1979). Neurotransmitter molecules in the synaptic vesicles are released, they diffuse across the synaptic gap, and they bind to protein receptors of another cell, where they generally induce a channel opening or closing.

Biophysics is concerned with essentially every level of cellular function, including bioenergetics, photosynthesis, cellular division, differentiation, adhesion, electrophysiology, neural transmission, and chemotaxis.

1.2.4 Organs, Organisms, and Systems. Groups of cells form organs, the most complex organ being the brain. The brain is composed of about a billion neurons, and a single neuron can be connected to thousands of others (Nauta and Feirtag, 1979). Central issues in biophysics concern the ways that information is organized, stored, encoded, and utilized in the brain.

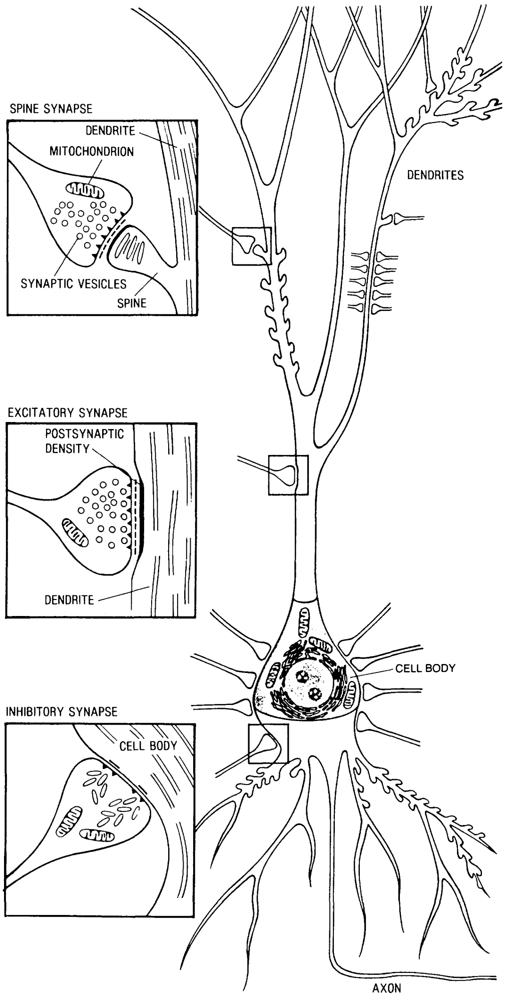

FIG. 8. Drawing of a typical neuron in the brain, showing three different types of synapses. [Iversen (1979).]

Organs in the body communicate with each other through a chemical signaling system mediated by hormones. The organism is of course comprised of organs; the human organism communicates with other organisms through its senses of hearing, feeling, seeing, tasting, and smelling. Biophysics of organisms and systems of organisms concerns areas such as perception, speech synthesis, and population behaviors. Here the tools generally are sophisticated measurement apparatus and statistical methodology. Certain insect colonies—honey bees, army ants, and fungus-growing termites—form superorganisms, where food, information, and aid pass freely among the members (Seeley, 1989). Such systems of organisms provide important models for mechanisms of genetic success.

2. TECHNIQUES OF BIOPHYSICS

2.1 Structural and Spectroscopic Techniques

Major advances in our understanding of biophysical phenomena have taken place as new techniques are developed. Typically, when a new technique is developed it is first applied to systems of interest to physicists, chemists, and physical chemists. Following its refinement it is then applied to more complex biological systems. This dependency on technique results primarily from the need to uncover structure to determine the molecular basis of the sought after function. In molecular biophysics it is the molecular structure that is determined, whereas in neurobiophysics it is the electrical impulse characteristics that lead to an understanding of the neural circuitry. Table 1 summarizes various spectroscopic techniques that are important in biophysical studies.

2.1.1 X-Ray Crystallography. X-ray crystallography ushered in the modern era of molecular biology. By x-ray diffraction the positions of the nuclei may be determined (see CRYSTALLOGRAPHY) and the structure of the macromolecule is thereby revealed from which inferences concerning the function of the molecule may be made. Comparison of two related structures will often reveal the structural basis for their differences in function. The early x-ray experiments on fibers of DNA by Franklin and Gosling (1953) and Wilkens and Randall (1953) allowed Watson and Crick (1953) to propose the molecular structure of nucleic acids. This work serves as the foundation of modern molecular biology. Later, the protein myoglobin was crystallized by Kendrew *et al.* (1960), giving the first glimpse of the three-dimensional structure of a protein. Even though the crystal structures of several hundred proteins have been determined, many proteins do not crystallize.

Macromolecular crystal structure determination, although very powerful, remains a very difficult task (Moffat, 1984). The prerequisite is a good quality crystal, which for some molecules cannot be obtained. When a suitable crystal is produced a set of x-ray scattering amplitude data is obtained. This is typically a set of many thousands of diffraction points. Next is the most difficult part of the determination, namely, solving the so-called phase problem. This problem results from the fact that each spot in the diffraction pattern is Fourier transformed, resulting in a wave with a given amplitude and phase. The amplitude is proportional to the number of electrons on each nucleus and the phase is the relationship of waves from different spots in the Fourier transform. The solution to the phase problem is to locate one position in the unit cell. The technique to do this is termed "isomorphous replacement" and involves replacement of one atom of the molecule by a heavy atom such as mercury. Once the phases have been determined the electron density map may be generated and the structure refined by various computer algorithms.

Table 1. Spectroscopic and structural techniques of molecular biophysics.

Technique	Excitation source	Typical wavelength (cm)	What is being probed	Typical information gained	Limitations
X-ray diffraction	X rays	10^{-8}	Diffraction from crystal lattice	Two- or three-dimensional arrangement of atoms of structural repeats such as helices	Molecules must form ordered structures or single crystals

(continued)

Table 1. *(continued).*

Technique	Excitation source	Typical wavelength (cm)	What is being probed	Typical information gained	Limitations
Neutron diffraction	Neutrons	10^{-8}	Diffraction from crystal lattice	Two- or three-dimensional arrangement of atoms or structural repeats such as helices	Molecules must form ordered structures or single crystals
EXAFS	X rays	10^{-8}	Inner-shell electrons of metals	Nearby atomic distances from metal centers	Model-dependent; provides distances associated only with metal absorption centers
NMR	Radio-frequency waves	10	Nuclear spin	Primary through quaternary structure, binding strength, molecular motion	Nucleus must be NMR-active; low sensitivity; overlap of peaks in complicated materials
EPR	Microwaves	1	Electron spin	Spin states, binding sites, dynamics	Molecule must have a paramagnetic reporter group attached to it
IR absorption	Infrared radiation	10^{-3}	Bond vibrations	Identification of structural moieties; monitor conformational changes	IR absorption of water dominates spectra of biological specimens; strong overlap of lines in complicated molecules
Raman and resonance Raman scattering	Ultraviolet and visible light	10^{-5}	Bond vibrations	Identification of structural moieties; monitor conformational changes and changes in local environment	Resonance Raman requires that a chromophore (e.g., retinal) be present in molecule; it then probes environment near this site
UV–vis absorption	Ultraviolet and visible light	10^{-5}	Electronic transitions	Electronic structure related properties of chromophore and changes in local environment	Molecular-level structural information is limited
CD	Polarized light	10^{-5}	Electronic transitions; sensitive to chirality	Relative amounts of helices, β sheets, random coil, degree of chirality in proteins; monitor conformational changes and local environment in biomolecules	Information is not very specific

(continued)

Table 1. Spectroscopic and structural techniques of molecular biophysics *(continued).*

Technique	Excitation source	Typical wavelength (cm)	What is being probed	Typical information gained	Limitations
Fluorescence	Ultraviolet and visible light	10^{-5}	Electronic transitions	Environment surrounding fluorophore, intramolecular distances via energy transfer molecular motion	Requires a fluorophore; not very specific if molecule contains more than one fluorophore; sensitive to quenching
STM	Electrons	10^{-5}	Work function of the surface	Surface landscape; atomic arrangement of atoms on surface, also information about band structure	Works best with conducting samples
AFM	Attractive and repulsive interatomic forces		Weak interatomic forces between surface atoms and tip	Topography of surface	Difficult data interpretation at present
TEM, SEM	Electrons	10^{-5}	Ultrastructure of metal-shadowed materials	Long- and short-range morphological features; organization of structures	Samples must be dried and fixed unless freeze-fracture techniques are used

The structures of many water-soluble proteins have been solved by the above techniques. However, to gain an understanding of proteins which play critical roles in cellular processes, crystallographers are now turning their attention to membrane-bound proteins. Since the structures of these proteins depend on interactions with lipids in membranes, in crystalline preparations the lipids are replaced by detergents. Great success has been achieved already with the photosynthetic reaction center from which a full crystallographic determination has been made (Deisenhofer *et al.*, 1985). Interestingly, many of the distances and relationships found in the crystallographic determination were in agreement with previous inferences made by other biophysical measurements.

A drawback of conventional crystallographic studies has been in the need to study only the static conformation of equilibrium forms of crystals. Other techniques, discussed below, permit the study of motional properties of molecules, properties which are known to be important for biological function. With the use of synchrotron radiation, in which high intensity pulses of x rays are generated, it is becoming possible to carry out time-resolved x-ray structural studies. This will allow the full set of changes in a molecule to be followed as it carries out its function. Another advantage that is being realized in the use of synchrotron radiation is that much smaller crystals may be used than in the past, thereby alleviating some of the difficulties presented by the need for large crystals when conventional instrumentation is employed.

2.1.2 Nuclear Magnetic Resonance (NMR).

Developed by Purcell and Bloch in 1946, NMR spectroscopy has been used extensively for biophysical problems (see NUCLEAR MAGNETIC RESONANCE). It is an exceptionally powerful technique and has been used, e.g., to establish motions in lipids, to determine protein structures, to follow *in vivo* metabolism, and to diagnose disease for medical applications (Jelinski, 1984).

2.1.2.1 Nuclear Interactions. The NMR phenomenon arises because most nuclei have a nuclear spin whose quantized energy levels become nondegenerate when placed in a magnetic field. The energy necessary to elicit transitions between these energy levels is in the radio-frequency regime. Thus, an NMR experiment consists of placing the sample in a highly homogeneous magnetic field and applying a radio-frequency pulse of exactly the correct resonant frequency. This excites the nuclei, which undergo a free-induction decay, providing a signal (voltage) as a function of time. This time-domain signal is amplified, digitized, and Fourier-transformed to produce a spectrum consisting of amplitude vs frequency.

NMR-active nuclei of biophysical interest include ^{1}H, ^{23}Na, ^{31}P, ^{39}K, ^{15}N, ^{13}C, and ^{17}O, although ^{2}H, ^{113}Cd, and ^{19}F, among others, can be used for labeling and tracer experiments. Although the interactions of these NMR-active nuclei with the main magnetic field produces the NMR phenomenon, most of the information content in an NMR spectrum arises from more subtle interactions, including the chemical shift, dipolar coupling, scalar, or *J* coupling, quadrupolar interactions, and relaxation effects.

The chemical shift reflects the electronic screening about the nucleus. For example, the four different types of protons in lactic acid [$CH_3CH(OH)COOH$] have slightly different chemical shifts because their electronic environments are not identical. The chemical shift tensor describes the direction and magnitude of the chemical shift with respect to the principal axis system of the molecule. If the molecule is tumbling isotropically in solution, the chemical shift tensor is averaged to its isotropic value. The isotropic chemical shifts of lactic acid can be used to identify its presence, to quantify the amount present, to monitor the *p*H of its environment, and perhaps to assess its interaction with other substances. Figure 9 shows how a ^{31}P NMR spectrum of the muscle of a living rat can provide information about the *p*H via the chemical shift of the inorganic phosphate peak and about energy metabolism via the intensities of the peaks from ATP (note that the α, β, and γ phosphorus peaks of ATP are well resolved).

The dipole-dipole interaction, a direct coupling effect, is also orientation dependent. Its strength is inversely proportional to the third power of the distance between the interacting nuclei, so that relaxation mechanisms based on the dipolar interaction are exceptionally useful for providing distance or structural information. Spectral broadening caused by the dipole-dipole interactions is theoretically averaged to zero for molecules that tumble isotropically in solution. Molecules which are unable to reorient freely give rise to broad features in the baseline of ^{31}P NMR spectra and are generally not observed.

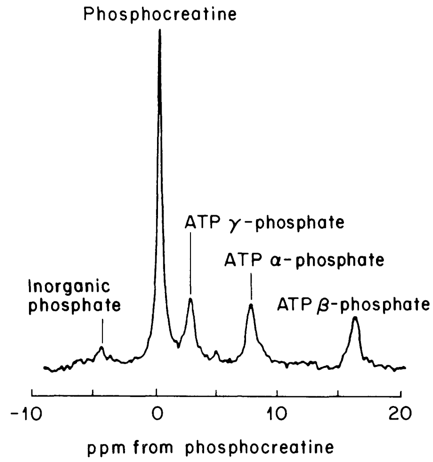

FIG. 9. ^{31}P NMR spectrum of the muscle of a live laboratory rat. Peaks corresponding to the major metabolites are labeled. [From Jelinski (1984).]

The scalar, or *J*, coupling is an indirect interaction mediated by the bonding electrons. Under high-resolution conditions, this produces a several-Hz splitting of the resonance lines. The multiplicity of the splitting is related to the spin of the nucleus and to the number of interacting spins. More importantly for biophysical studies, the magnitude of the splitting depends on the geometry of the molecule. For peptides and proteins, e.g., the relationship between dihedral angle and the size of the coupling constant between the N–H and the α-CH protons is well established and is used extensively to establish peptide geometries.

Quadrupolar interactions are usually dominant for nuclei with spin quantum number greater than $\frac{1}{2}$. For biophysical studies, these most commonly include ^{39}K, ^{23}Na, ^{17}O, and ^{2}H. As with the dipole-dipole interaction, quadrupolar broadening interactions for nuclei tumbling isotropically in solution are averaged to zero. For solids the quadrupolar interaction produces well-defined line shapes that can be analyzed to provide information about the rate and angular range of motion.

2.1.2.2 *Nuclear Relaxation.* There are two types of nuclear relaxation: spin-lattice and spin-spin. Spin-lattice relaxation pertains to relaxation induced by the lattice, or environment of the spin. For example, when a sample is placed in a magnetic field, a certain amount of time is required for the spins to align themselves either with or against the magnetic field. Although the population of spins in these energy states is determined by the Boltzmann distribution, the rate at which these populations are established is determined by the spin-lattice relaxation time T_1. Fluctuating local fields, most often dipolar or quadrupolar, provide a mechanism through which the spin can couple with its lattice. Spin-lattice relaxation can be used to uncover the details of molecular motion and molecular interactions.

Spin-spin relaxation, or T_2 relaxation, arises from fluctuating local fields set up by mutual spin flips. When motion is fast, the T_1 and T_2 are equal. They are unequal, however, in the regime of slow motions, where the T_2 provides valuable information on protein and DNA binding events.

The enhancement in intensity of a spin transition in one nucleus that occurs when a nearby nucleus is irradiated with a radio-frequency pulse at its resonant frequency is the nuclear Overhauser enhancement, or NOE. It arises from purely dipolar relaxation and has an inverse sixth-power dependence on distance. The NOE is widely used to establish tertiary structures of proteins and for exploring helices, bends, and interactions in DNA.

2.1.2.3 *Solid-State NMR.* Although the chemical shift anisotropy and the dipole-dipole interaction in solids produce substantial broadening, they are nevertheless valuable sources of orientation-dependent information, particularly if the molecule under study is involved in restricted molecular motions. It is possible, however, to eliminate these broadening mechanisms to produce high-resolution spectra of solids. These techniques involve magic-angle spinning, which reduces the chemical shift anisotropy pattern to its isotropic average, and dipolar decoupling, which removes the effects of dipolar broadening. (The "magic angle," or 54.7°, is the angle where the expression $3\cos^2\theta - 1$ is equal to zero. Both the chemical shift and the dipolar tensors have a $3\cos^2\theta - 1$ dependence.) High-resolution spectra of solids have been used in numerous studies of bones, teeth, and high-molecular-weight DNA, and in establishing the biophysics of the visual system (see Sec. 3).

Broad-line spectra of quadrupolar nuclei (obtained without magic-angle spinning and dipolar decoupling) have also been used with selective labeling techniques to study the visual system. Such experiments have also been applied to structural proteins such as collagen and elastin.

2.1.2.4 *Two- and Three-Dimensional NMR.* Two-dimensional (2D) NMR spectra can be obtained by adding an additional time period, the evolution time, to the standard pulse sequence used to obtain a one-dimensional spectrum. A number of spectra (perhaps 256) are acquired that differ from one another by the length of the evolution time. If subjected to a two-dimensional Fourier transform, a two-dimensional spectrum is obtained. Frequencies that were unchanged during the evolution period appear on the diagonal of the 2D data matrix, since $\omega_1 = \omega_2$. If two spins interacted with each other during the evolution time, perhaps by J coupling or by NOE, depending on the pulse sequence, $\omega_1 \neq \omega_2$, and off-diagonal elements appear.

Two-dimensional NMR spectra of proteins, DNA, and RNA have completely revolutionized structural analysis. Two-dimensional experiments are especially useful for structure determinations in proteins that do not crystallize, although at present they are limited to proteins that have less than about 140 amino acid residues. A 2D structure determination involves first assigning all of the resonances, a time-consuming and often difficult process made possible by an entire battery of 2D experiments. The structure is determined computationally using distance constraints imposed from 2D-NOE measurements. The structure of basic pancreatic trypsin inhibitor (BPTI), a 58–amino-acid protein, was the first to be analyzed in this manner. Since then, the structures of a number of proteins have been established. Much of what we know about DNA structure, including the way that bases are stacked one on top of the other, the sugar pucker (distortion of the five-membered sugar ring), and propeller twists (nonplanarity within pairs of hydrogen-bonded bases), have been established by 2D NMR experiments.

Two-dimensional NMR can be extended to 3D and 4D NMR by performing experiments with additional evolution times. When Fourier-transformed, these spectra produce a multidimensional data matrix where different planes correspond to correlations.

2.1.3 Electron Paramagnetic Resonance (EPR). The basic concepts for electron paramagnetic resonance (*q.v.*) (EPR; also called electron spin resonance, or ESR) are similar to those of NMR (Kosman, 1984). However, EPR is used to probe the environment around unpaired spins, whereas NMR is most useful for diamagnetic samples. When placed in a magnetic field, the electron spins exist in two energy states. The frequency corresponding to the difference in energy between these two states is in the microwave regime. Thus, the resonant condition is attained by microwave irradiation of the sample. One of the parameters of interest (analogous to the chemical shift in NMR) is the *g* factor. The *g* factor represents the ratio of the magnetic moment to the angular momentum and has a value of 2.002 319 for a free electron. The *g* value is sensitive to the chemical environment surrounding the unpaired spin. Furthermore, like the chemical shift anisotropy, it exhibits an anisotropy from which the extent of molecular motion can be derived.

Many insights into the environments at the active site of enzymes or in lipid bilayers have been obtained by a technique known as spin labeling. In this method, a small probe molecule containing a stable free-radical species (such as a nitroxyl) is attached to the biomolecule of interest. This probe molecule serves as a reporter of the environment and dynamics in its immediate vicinity.

2.1.4 Vibrational Spectroscopy (IR, Raman). Vibrational spectroscopy probes the frequencies of the molecular vibrations (see MOLECULAR SPECTROSCOPY). The energy of molecular vibrations is in the infrared region of the electromagnetic spectrum ($\sim$0.1 eV). Vibrational excitations may be measured by the direct absorption of light (infrared absorption spectrum) or by light scattering (Raman scattering). A serious difficulty in the infrared absorption technique results from the absorption by water which is a very intense absorber and overlaps with the vibrational modes of biological macromolecules. Because of this, extremely thin and highly concentrated samples must be used for infrared absorption measurements. In contrast, water is a very weak scatterer in the Raman spectrum and therefore does not present any difficulties. However, except for special conditions such as resonance enhancement, to be discussed below, the intensity in the Raman spectrum is quite weak and so high concentrations (at least 5–10 mM) are necessary. Crystalline or solution samples may be studied by either vibrational spectroscopic technique.

Vibrational spectroscopy may be used to study all classes of biomolecules (Rousseau and Ondrias, 1984; Alben and Fiamingo, 1984). In lipids, vibrational modes involving the head group (phosphate and C–N stretching modes) are present in Raman spectra, and several modes of the hydrophobic tail are also present. These modes have provided a great deal of information concerning the thermotropic transitions in lipid membranes. In proteins, amide modes of the polypeptide backbone are present in Raman scattering and infrared absorption spectra. Modes from some of the amino acid residues are also present. Many studies of proteins have been carried out and useful information has been obtained concerning protein secondary structure, residue exposure to solvent, and protein unfolding. Raman spectra of polynucleotides are also rich in information. Modes are present from the sugar phosphate backbone as well as from the individual bases. The spectra are sensitive to the degree of order of the polynucleotide, to the structure (*A*, *B*, or *Z*), and to the extent of base pairing.

To obtain a Raman spectrum the samples are exposed to incident laser light with wavelengths typically in the visible spectral region, and the spectrum of scattered light is recorded. The vibrational excitation is detected by the shift in wavelength of the scattered light. If the incident wavelength is in coincidence with a real electronic transition of the molecule being studied the intensity is enhanced and the effect is referred to as resonance Raman scattering. Two biological chromophores have been extensively studied by resonance Raman scattering: the heme macrocycle (Spiro, 1988) and retinal (Lugtenburg *et al.,* 1988), the visual pigment. The heme group consists of an iron atom coordinated to an aromatic porphyrin ring. Heme proteins are ubiquitous in the biological world and function in a variety of capacities ranging from multifunctional enzymes, to oxygen transport and storage proteins, to electron transfer proteins. A great deal of understanding of how the protein interacts with the heme group has been made by resonance Raman spectroscopy. Retinal is the biologically active center of the visual system. Resonance Raman studies of retinal-containing proteins have allowed for the identification of the various intermediates present in the visual system, as discussed in greater detail in Sec. 3 of this article. An important aspect of resonance Raman scattering as applied to these two chromophores (heme and retinal) is that it allows for the study of the biologically active site of the molecule without interference from the protein matrix which surrounds it. This selectivity makes this technique extremely powerful for the study of biological macromolecules containing chromophores.

When ultraviolet light sources are used, the advantages of resonance enhancement may be extended to aromatic amino acid residues and to polynucleotide bases (Asher, 1988). This is a growing field and although it has some difficulties, such as the need for high average laser power, it is very promising because information about specific amino acids and nucleotide bases can be obtained.

Many spectroscopic and crystallographic studies of biomolecules are carried out on molecules under equilibrium conditions. However, most biological reactions take place under nonequilibrium conditions. In addition there are often many different fleeting metastable intermediates which must be studied to determine a mechanism for a particular binding or catalytic process. For such studies it is necessary either to freeze trap a particular intermediate or do time-resolved spectroscopy. Both techniques have been widely applied to the study of biological reactions by vibrational spectroscopy. Most extensively studied have been the photoexcited intermediates in rhodopsin and bacteriorhodopsin (Lugtenburg *et al.,* 1988) and transient states following ligand photolysis in heme proteins (Rousseau and Friedman, 1988). Some of the results of the rhodopsin studies will be discussed in Sec. 3. Time-resolved vibrational studies have been carried out on heme proteins by using resonance Raman scattering from the picosecond time scale to the millisecond time scale. This work, along with freeze trap studies, has provided a determination of the relaxation pathway in hemoglobin and myoglobin following ligand photodissociation (Rousseau and Friedman, 1988).

2.1.5 Optical Techniques. In this section we consider several optical techniques that utilize visible and ultraviolet excitation. These include optical absorption, circular dichroism, and fluorescence. There are many excellent review articles which give in-depth descriptions of these techniques. In this section we present briefly some of the types of problems that are addressed by optical techniques.

2.1.5.1 Optical Absorption. The technique of optical absorption is one of the most widely used in biological research because it is a simple, inexpensive, and reliable probe of many macromolecules (Cantor and Schimmel, 1980). Indeed proteins, nucleic acids, and many small prosthetic groups have characteristic absorption spectra. Consequently, absorption spectroscopy is used at two levels. First, it is used as a way of characterizing and assaying a biological molecule or assembly. Second, it is used as a spectroscopic research tool.

In the optical absorption technique the incident light excites an electronic transition. These transitions are sensitive to molecular conformation and therefore, the absorption spectrum is a very rapid way of assessing the conformation of macromolecules. Spectra may be obtained from isolated molecules, organelles, organs, or intact animals. This broad range of applications makes it a powerful technique for characterizing materials. In addition, the light intensity used is generally weak so that no changes are induced in the system being studied. Finally, with the many different light sources that are available (e.g., lasers), time-resolution studies may be made as well. Such studies have been the basis for establishing the photocycle in rhodopsin, the visual pigment (Birge, 1981).

Proteins have two main optical absorptions: that of the polypeptide backbone and that of the amino acid residues. The π electrons of the peptide bond extend over three atoms: the nitrogen, the carbon, and the associated carbonyl oxygen. Two transitions involving these electrons are detected in the 190–220 nm region. The intensities and wavelengths of these transitions are sensitive to the secondary structure of the protein, displaying significant differences between alpha-helical, beta-sheet, or random coil structures. The aromatic amino acids phenylalanine, tyrosine, and tryptophan also have electronic transitions in the UV region. In the 230–300 nm region of the absorption spectrum these transitions are sensitive to changes in protein structure. For example, the electronic transitions are sensitive to whether an aromatic residue is in the interior of a protein in a hydrophobic environment or is exposed to the solvent and is thereby highly solvated. This sensitivity has been used to advantage in studies of hemoglobin in which its optical absorption spectrum has been used to monitor and characterize changes in the quaternary structure of the protein (Perutz, 1979). The optical transition in tyrosine is a sensitive reporter of *p*H, since the electrons on the oxygen of the -OH group affect the electronic structure of the ring through conjugation effects.

Prosthetic groups within proteins have been studied extensively by optical absorption techniques. The spectra of these groups are often sensitive to the protein environment in which they are embedded. The absorption spectrum of the heme group has played a very important role in characterizing this enormous class of proteins. Studies of heme proteins have ranged from investigating specific properties of isolated heme proteins, to observing the electronic structure when assembled in organelles, to characterizing metabolic states in the brain and monitoring the saturation of oxygen by blood in the tissues.

The absorption spectra of nucleic acids is dominated by the purine and pyrimidine bases. They produce a composite absorption band centered at about 260 nm. Although complicated by the overlap of the absorption band from each base, the intensity and the position of the absorption maximum is sensitive to *p*H and interactions between bases.

2.1.5.2 Circular Dichroism. Nearly all biological molecules are optically active. Optical activity results from the low symmetry of biological molecules and produces some measurable effects. Circular dichroism is the most widely used optical activity effect (Cantor and Schimmel, 1980; Maestre, 1984). In this technique a comparison is made between the absorption of left- and right-circularly polarized light. Because of optical activity the amount of absorption is different for these two polarizations.

Circular dichroism is used extensively to determine secondary structure in proteins. In the region from 190 to 250 nm, where the optical absorption is dominated by transitions from the polypeptide backbone, a strong CD spectrum is observed which is very sensitive to the secondary structure, as is evident in Fig. 10. By using various deconvolution techniques, the amount of secondary structure may be determined from the CD spectrum of a protein.

2.1.5.3 Fluorescence Spectroscopy. When a molecule absorbs light, the energy must eventually be reemitted in some form. Frequently this relaxation will take the form of fluorescence, a reemission which is shifted to somewhat lower energy than the excitation frequency. In biological molecules the wavelength, polarization, and lifetime of the fluorescence are each useful in characterizing the properties of the molecule (Stryer, 1978; Cantor and Schimmel, 1980; Schlessinger and Elson, 1982; Yguerabide and Yguerabide, 1984).

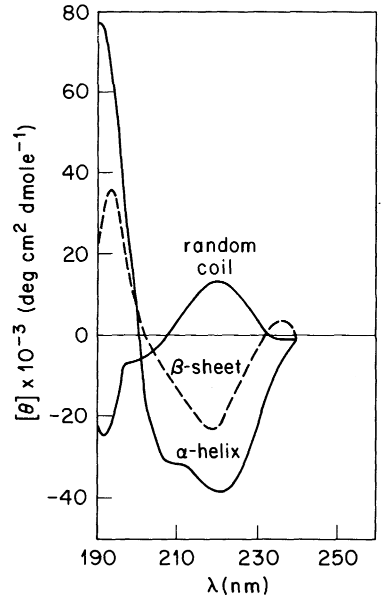

FIG. 10. CD curves showing characteristic patterns for random coil, β-sheet, and α-helix secondary structures. [Adapted from Cantor and Schimmel (1980), p. 427.]

In proteins, the tryptophan residue is the most common intrinsic chromophore studied by fluorescence. Excitation at about 280 nm leads to emission at about 350 nm. The wavelength of the fluorescence is sensitive to the environment surrounding the tryptophan, shifting significantly, e.g., when a tryptophan which is buried in the hydrophobic region of a protein becomes exposed to the solvent. The lifetime of the emission is also sensitive to solvent exposure and therefore can be used to separate the contributions from interior and exterior tryptophans.

Tumbling motions of molecules may be measured by polarization anisotropy. For a molecule that is fixed in space, the polarization of the fluorescence is a function of the angle between the absorbing and the emitting dipole. However, in a molecule that is tumbling on a time scale that is faster than the fluorescence emission time the anisotropy of the emitted polarization becomes zero. Thus, the polarization anisotropy can be used to measure the relationship between the two dipoles and it can be used to measure the tumbling frequency of molecules. Rotational and diffusional properties of membranes have been determined by following the time course of the anisotropy of the polarization of fluorescent probes placed in membranes.

Distances between two molecular sites may be determined by taking advantage of fluorescence excitation transfer. To carry out such measurements the fluorescence intensity and lifetime of a fluorophore are measured. These measurements are compared to those made when an acceptor molecule is placed at some nearby site. The energy transfer between the sites is proportional to $1/r^6$, where r is the distance between the sites. This theory for energy transfer involving the dipole-dipole interaction was developed by Förster. This technique has been used to measure distances in a wide variety of nucleic acids and proteins (Stryer, 1978).

2.1.6 Extended X-Ray Absorption Fine Structure (EXAFS). Atoms absorb light in the x-ray region by the photoionization of inner shell electrons. The absorption spectrum has a series of edges in it corresponding to each of the inner shell electronic orbitals. An absorption spectrum of the x rays reveals variations in intensity in the 50–100-eV region past the edge inflection. This fine structure in the absorption spectrum (EXAFS) results from elastic backscattering of the photoejected electrons with the electron density of nearby atoms. (For a good review see Scott, 1984.) As a result the EXAFS spectrum is sensitive to the distance and type of atom in the neighborhood of the atom producing the edge absorption. In biological applications absorption edges above 4000 eV are primarily studied to avoid the absorption by protein and water. The most commonly studied biological molecules are metalloproteins and metalloenzymes. EXAFS only became readily feasible for biological molecules since the development of synchrotron radiation which gives about two orders of magnitude more intensity than the conventional Cu K_α line.

EXAFS gives selective information around the metal center being studied. This turns out to be especially useful in biological applications because if there is a metal at the active center it can be probed without interference from other atoms in the molecule. The technique is especially powerful when the identity of atoms which are bound to a metal center is known and the distances must be determined. EXAFS measurements can be used to determine distances accurately in such cases and it also has the advantage that single crystals are not necessary as in x-ray crystallography. EXAFS measurements have been performed on nearly every type of metalloprotein and have been most successful on studies of iron-sulfur proteins, heme proteins, copper proteins, and molybdenum enzymes.

2.2 Electrophysiology

Much of what we know about neurobiology has come from electrophysiological measurements, which combine chemical and electrical information about cells. The neuron (Fig. 8) has a dual signalling system, since it uses both chemicals (neurotransmitters; see Sec. 1.2.1) and electrical impulses (action potentials) to transmit its information.

There is a voltage difference across the membranes of neuronal cells referred to as the membrane potential (Miles, 1969). This potential is a consequence of an ionic imbalance across the cell membrane. This imbalance is generated by two main features of the cell membrane: the sodium/potassium pump and the fact that the leakages of sodium and potassium ions back through the membrane are different. Consider a membrane with equal concentrations of Na^+ and K^+ ions on each side. The sodium/potassium pump transports three Na^+ ions out of the cell and two K^+ ions into the cell for each molecule of ATP hydrolyzed. This process creates a membrane potential. Moreover, the leakage permeabilities of these two ions back through the membrane are very different. K^+ ions go through the membrane much more easily than do Na^+ ions. Because of the permeability difference, the K^+ ions leak out and a significant membrane potential is developed.

In addition to the force of the concentration gradient, the ionic flow across the membrane is also dependent on the voltage gradient across the membrane. To understand the balance between these forces the concentration gradient is converted to a potential by the Nernst equation

$$V = (RT/FZ)\log(C_o/C_i).$$

In this expression R is the universal gas constant, F is the Faraday, Z is the valence of the ion, and C_o and C_i represent the concentration of some ionic species on the outside and on the inside of the membrane, respectively. Assuming that there is a channel for a given ion, the net flow of that ion through the membrane will be zero if the membrane potential is the same as the Nernst potential. The ion will flow out of the cell if its Nernst potential is less than the membrane potential and into the cell if it is greater than the membrane potential.

The resting potential of a cell membrane is the membrane potential when the net ionic flow is balanced. For typical nerve cells this potential is usually about -60 mV. When a nerve cell is excited, channels are opened or closed so that the flow of ions through the membrane changes and therefore the membrane potential changes. Signal propagation down a nerve cell is caused by the opening of Na^+ channels. When they are excited by an electrical pulse Na^+ ions flow into the cell (its Nernst potential is about $+50$ mV), causing the membrane potential to become depolarized, i.e., a positive shift in its potential. In some cases there are current flows in the other direction causing the cell to have a more negative membrane potential. This is referred to as hyperpolarization and will be seen in the example of the visual system as an essential feature in the visual information transduction process.

When a cell receives a depolarizing pulse it generates a wave, or action potential, which propagates along the axon of the cell until it synapses on the next cell in the circuit. The wave itself is very complicated, involving both sodium and potassium channels. At that point, information can be transmitted electrically to nearby cells or it can be transmitted chemically. In chemical transmission, the nerve terminal releases neurotransmitter into the synaptic gap. These neurotransmitter molecules diffuse across the gap and bind to receptor channel proteins on the surface of the next cell in the circuit. Once the molecules have bound, the channels open and ions flow, changing the intracellular potential. Thus, information is transmitted both by electrical and chemical means.

To characterize the properties of nerve cells the voltage across the cell or the current flow in or out of the cell is measured by placing electrodes in the cell. For example, by placing electrodes at several places along an axon the action potential may be followed as the voltage impulse passes down the cell. A common procedure used to study electrical and chemical properties of cells is to apply a voltage clamp. This is done by placing two electrodes in a cell: one measures the voltage and the other is used to inject current. By controlling the current which is passed into the cell the voltage may be fixed at predetermined values. The amount of current needed to hold the voltage at a certain value depends on the rate at which various ions can flow through the membrane. By changing the concentrations of each ion in the solution it becomes possible to determine which ions carry the transmembrane current.

Patch clamping is a method used to study the individual channels in the membrane. In this technique, a glass micropipette having a tip with a very narrow diameter (1–3 μm) is filled with a conducting salt solution. This micropipette is pressed onto the cell surface, where it forms a gigaohm seal with the membrane. The salt solution in the micropipette forms part of a circuit with the patch of membrane spanning the tip of the micropipette. Current flows through the single membrane channels, and the electrical potential across the membrane can be clamped at different voltages to provide information about the channel behavior. Different configurations possible for patch clamping are shown in Fig. 11. Whole-cell recording can be performed by applying enough suction to rupture the cell membrane or inside-out or outside-out patches can be formed. One can also obtain extracellular recordings by placing the electrode in a well-defined extracellular position. Information about the functioning of the cellular channels is obtained by applying known neurotransmitters, specific channel blockers, and solutions of known ionic concentration.

2.3 Imaging Techniques

Although they grew out of biophysical research, many imaging techniques have now found wide applicability in the medical community. Among these are positron emission tomography (PET), magnetoencephalography (MEG), computed tomography (CT), and nuclear magnetic resonance imaging (MRI). In addition to their medical utility, these imaging techniques continue to serve as valuable tools for biophysics research, particularly for studying *in vivo* function. Using these noninvasive techniques, it is possible to obtain images of pulsatile blood flow, to visualize the difference between normal and cancerous tissue, and to monitor activity patterns in the brain.

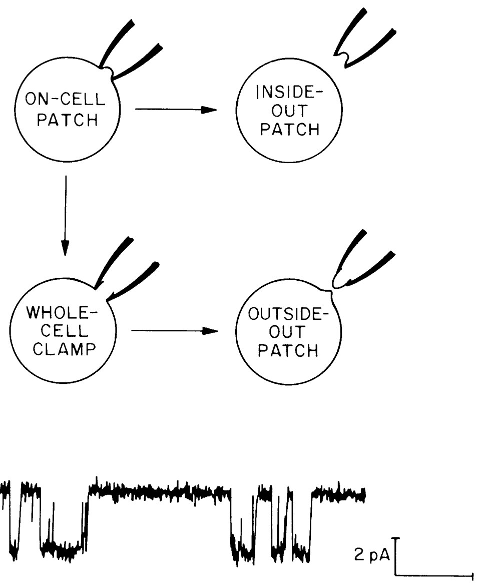

FIG. 11. Top: Various configurations for patch clamping. In the on-cell patch the electrode makes a tight seal with the outside of the intact cell. By pulling the electrode away from the cell, a patch of membrane remains on the electrode with its inside surface facing out. By applying suction to the on-cell patch the membrane is broken open and the voltage of the cell may then be clamped to a new value. By pulling the electrode away from the cell, in the latter case, the membrane reseals exposing the outside surface. Bottom: Patch recordings through a single protein channel, showing the open and closed states of the channel. [Adapted from Shepherd (1988), pp. 33 and 34.]

In contrast to these biomedical imaging techniques which produce macroscopic information, scanning tunneling microscopy (STM) and atomic force microscopy (AFM) produce atomic-scale information. Electron microscopy bridges the size gap, providing mainly supramolecular images.

2.3.1 Positron Emission Tomography (PET).

PET images provide information about oxygen utilization and metabolic disorders such as Alzheimer's disease, epilepsy, schizophrenia, Parkinson's disease, or stroke (Dagani, 1988).

In PET, the patient or animal is injected with radioisotopically labeled compounds such as glucose, ^{18}F-containing dopamine, or other labeled molecules. Because they are radioactive, these particular compounds emit positrons as they flow through the blood stream. When positrons collide with electrons, they annihilate and produce gamma rays that fly off from each other in opposite directions. PET scanners work by detecting these gamma rays with an array of detectors. The image is reconstructed by combining the information from the circular array of gamma ray detectors. Because the radiolabeled chemicals have very short half-lives, they must be prepared freshly in a nearby cyclotron.

2.3.2 Magnetoencephalography (MEG).

When neurons in the brain discharge, the tiny electrical currents produce magnetic fields which extend out beyond the skull (Rose *et al.*, 1987) (see BIOMAGNETISM, MEDICAL APPLICATIONS). These fields, 10^9 times weaker than Earth's magnetic field, can be measured with ultrasensitive detectors called SQUIDs (superconducting quantum interference devices). Multisensor SQUID systems can measure this magnetic field simultaneously at different locations around the head, thus mapping regions of neuronal activity in the brain. MEG has been used to show that a tone map exists across the human auditory cortex. It has also been used in cognitive studies and to locate the center of an epileptic discharge.

2.3.3 Computed Tomography (CT). Computed tomography uses x rays to make images. As in x-ray shadowgrams, the contrast in CT is dominated by the density of bones and tissues. Since these materials cast shadows in normal x-ray images, they obscure whatever is behind them. This problem is circumvented in CT by using an x-ray source that can rotate in a circle. This provides multiple pictures of the object from many different angles. All of these images are combined by the computer to produce an image of a slice of the body. These slices can also be used to construct three-dimensional images; these three-dimensional pictures are useful for finding the exact size, shape, and location of tumors, and for pinpointing the precise location for beaming radiation for radiation therapy. They can also be used in reconstructive surgery to define the shape and size of artificial joints.

2.3.4 Magnetic Resonance Imaging (MRI). Magnetic resonance imaging (*q.v.*) is an extension of NMR (see Sec. 2.1.2), wherein frequency information is encoded into spatial information (Jelinski *et al.*, 1989; Morris, 1986). The sample is placed in a homogeneous magnetic field and is subjected to radio-frequency pulses to achieve the resonance condition, as previously described for NMR spectroscopy. However, in MRI the frequency information is translated into spatial information by subjecting the sample also to field *gradients* in the x, y, and z directions. This produces images, or "slices" of the sample; one such slice through a plane of the brain containing the eyes is shown in Fig. 12.

Since MRI is a noninvasive technique that does not use ionizing radiation, it has found wide utility in clinical applications. Although MRI is routinely used for medical applications, much biophysical research is also being performed, particularly in the area of imaging of function.

MRI is most often used for imaging the protons in water, since water is ubiquitous in biological tissue. The contrast in an MRI image arises from several sources, and the contrast can be "tuned" for the desired effect by adjusting various experimental parameters. The amount of water present in the sample is responsible for only part of the contrast. The other aspects of contrast arise from the differences in relaxation times between water in various tissues. These relaxation times, called the spin-lattice (T_1) and spin-spin (T_2) relaxation times (see Sec. 2.1.2), are different for different tissues. These differences arise primarily from differences in the way that water interacts with the proteins and other macromolecules in the tissues. Thus, the molecular basis for relaxation time differences is an essential research area for MRI.

Although proton imaging of water is most widely used, other biologically important molecules such as lactate and nuclei such as sodium can also be imaged. MRI can also be used to provide information about blood flow. Some of the biophysical applications regarding the imaging of function require that a fourth dimension also be obtained. This fourth dimension is the chemical shift, wherein the chemistry occurring within a

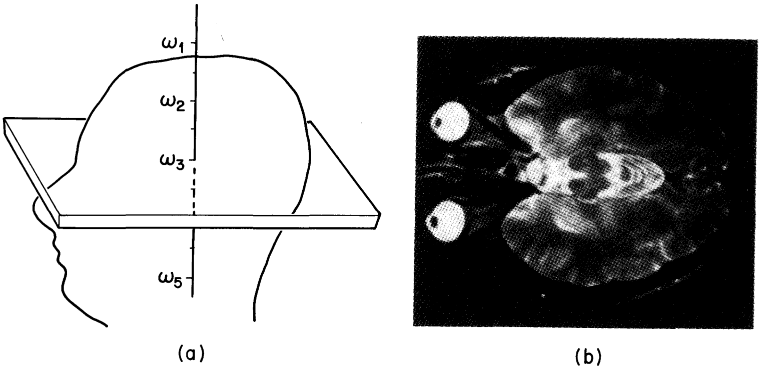

FIG. 12. Magnetic resonance image through a plane of a brain containing the eye. [Jelinski *et al.* (1989).]

particular volume element ("voxel") of the image is monitored.

2.3.5 Scanning Tunneling and Atomic Force Microscopy (STM and AFM). Binnig and Rohrer were awarded the Nobel Prize in physics for their invention of the scanning tunneling microscope (STM), which made it possible to obtain images with atomic-scale resolution of conducting surfaces (Hansma and Tersoff, 1987) (see SCANNING TUNNELING MICROSCOPY). Since its invention, STM and the related atomic force microscopy (AFM) have been applied to biological molecules such as leucine (Gould *et al.*, 1988), lipids (Zasadzinski *et al.*, 1988), and DNA (Amrein *et al.*, 1988), as well as to semiconductors and other inorganic and organic materials of nonbiological origin. The STM works in air and under water, as well as in vacuum. Figure 13 illustrates the basic operation of an STM instrument. The tip of the STM is made by etching a wire (usually tungsten) to produce a small radius of curvature. Tunneling occurs off the last atom or atoms on the surface of the tip. The tip is scanned across the sample by means of piezoelectric *x*, *y*, and *z* translators, which produce carefully controlled, atomic-scale steps. In the constant-current mode, the tunneling current is monitored as the sample is scanned, and a feedback circuit adjusts the height of the tip to maintain this current constant. The image is then read out as the movement of the tip. In the constant-height mode, the tip is kept at a constant height, and the image consists of the variations in the current as the tip passes over atomic features on the surface. Since the tunneling current varies exponentially with the distance between the tip and the sample, a small change in the distance translates into a very large change in the tunneling current, thus accounting for the sensitivity of this technique.

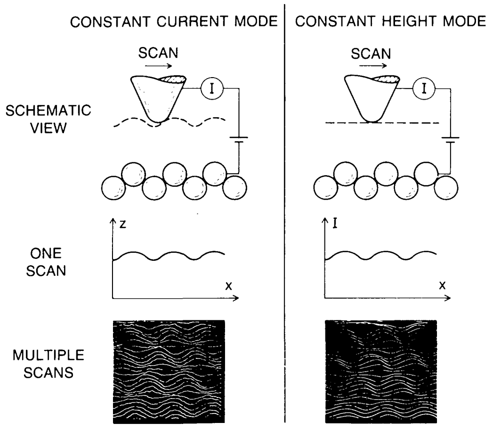

FIG. 13. Schematic view of the operation of a scanning tunneling microscope in the constant-current and constant-height modes. [Hansma and Tersoff (1987).]

The AFM works in a fashion similar to the STM (Binnig *et al.*, 1986), except here the sample need not be conductive or mounted on a conductive grid. The tip of the AFM is attached via a spring to a cantilever beam, and the deflection of the beam is monitored as the tip scans over the sample. Forces are generated by interatomic attractions and repulsions as the tip passes over various regions of the sample. The AFM can work in constant-force mode, where a feedback circuit adjusts the height of the tip from the sample to keep the force constant.

3. AN EXAMPLE: BIOPHYSICS OF THE VISUAL SYSTEM

In the last section of this article we focus on the visual system, which provides a specific example of how biophysical studies have uncovered some of the mysteries of nature. Information is received, converted into chemical information, and then into electrical information so that it may be transmitted, processed, and stored. This process of information transduction in vision is one of the most interesting and complex areas of biophysics. The machinery of this process spans nearly all areas of biophysics and essentially all biophysical techniques have been used to study the visual system.

3.1 Morphological Structure

Determination of the molecular and electrophysiological basis for the information transduction in the visual system has been made possible by understanding the physical map of the visual system from the many electron microscopic studies of the cells of the retina. Identification of these cells has allowed the pathway for the information flow to be determined. As may be seen in Fig. 14, light strikes the photoreceptor cells after first passing through the neurons which connect

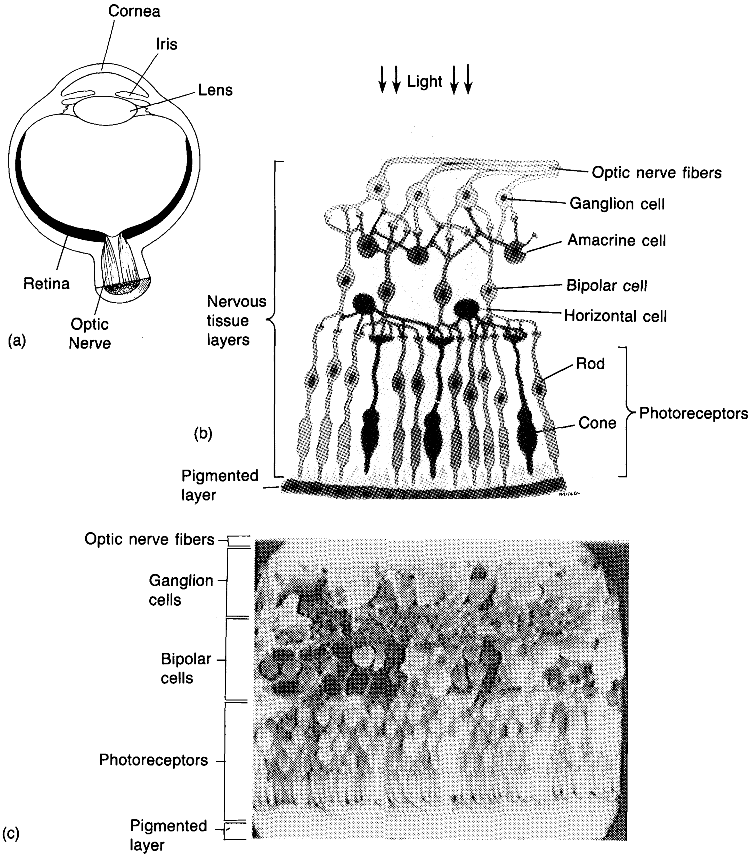
Cornea
Iris
Lens
Retina
Optic Nerve
(a)
Light
Nervous tissue layers
(b)
Pigmented layer
Optic nerve fibers
Ganglion cell
Amacrine cell
Bipolar cell
Horizontal cell
Rod
Cone
Photoreceptors
Optic nerve fibers
Ganglion cells
Bipolar cells
Photoreceptors
Pigmented layer
(c)

FIG. 14. Major structures of the eye. (a) Schematic view of the eye. (b) Schematic view of the retina. (c) Electron micrograph of the retina showing the same cells as those shown in the schematic view. [Adapted from Mason (1983), and from Kessel and Kardon (1979).]

the photoreceptors to the optic nerve. Identification of the photoreceptors from electron microscopy is very striking as is evident in Figs. 14 and 15 where the much larger rods and the smaller cones are clearly observed. Color sensitivity results from the cones. There are three types of cone cells with absorption maxima in the red, green, or blue. The same chromophore, retinal, is used in each of these, but its optical properties are modified by the protein which surrounds it, accounting for the differential color sensitivity.

The rods, which provide monochromatic detection of dim light, have been the most extensively studied photoreceptor cells. As shown in Fig. 15, the rods are complicated cells in which a section called the outer segment contains the photoreceptors, or disks (Figs. 15 and 16), and another section, called the inner segment, contains the cellular apparatus for generating energy and making molecules necessary for the functioning of the outer segment.

In addition, the inner segment contains a synaptic terminal which connects the rod to other cells of the nervous system. The electrical signals generated in the photoreceptor cells are processed by several layers of neurons (see Fig. 14) prior to passage through the optical nerve to the lateral geniculate nucleus, and finally to the visual region of the cerebral cortex (visual cortex).

3.2 The Primary Event in Visual Excitation

Within the disks (Fig. 16) of the outer segments of the rod cells is a membrane-bound protein called rhodopsin. It consists of a chro-

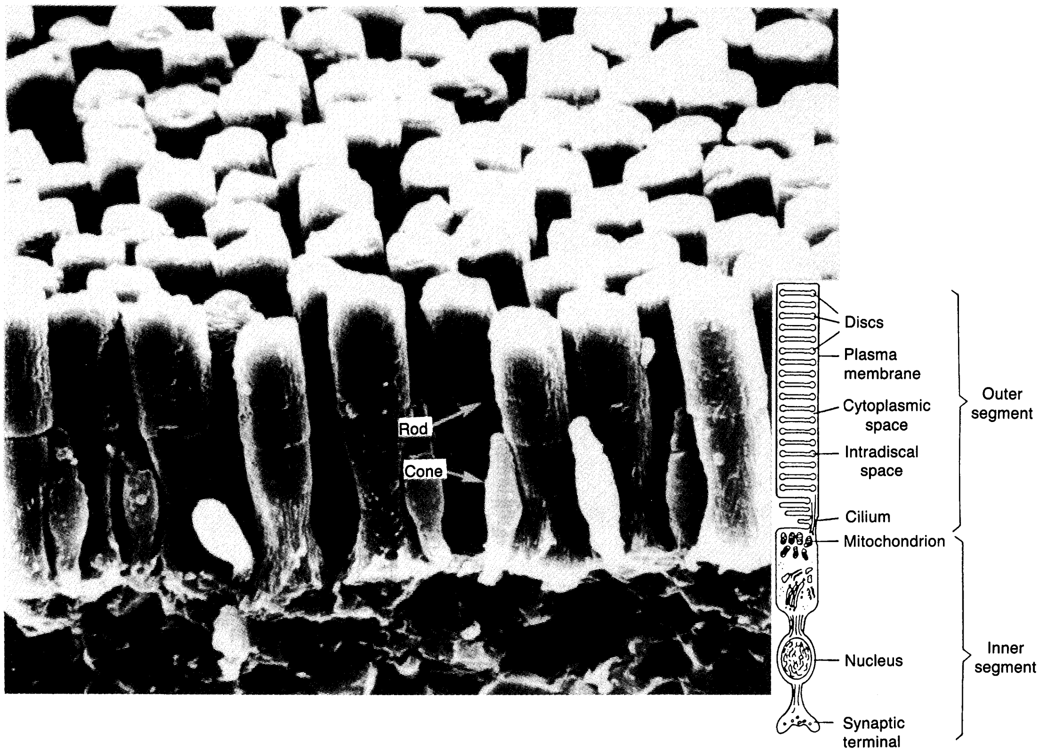
Rod
Cone
Discs
Plasma membrane
Cytoplasmic space
Intradiscal space
Cilium
Mitochondrion
Nucleus
Synaptic terminal
Outer segment
Inner segment

FIG. 15. Electron micrograph of visual photoreceptors (rods and cones). The schematic drawing shows the structure of the rod cell. [From Schnapf and Baylor (1987) and O'Brian (1982).]

mophore, retinal, surrounded by a protein with several helical segments passing through the membrane wall of the disk. Retinal, one of the active elements of rhodopsin, is a polyene, an alternating chain of single and double bonds, attached to an ionene ring (Fig. 17). It is derived from all-*trans*-retinol (vitamin A). The delocalized electrons in the polyene chain give rise to a very strong optical absorption transition. Many experimental and theoretical studies of this transition have been made to understand how binding retinal to the protein (opsin) can shift the natural retinal transition from the near ultraviolet (380 nm) to the visible region (500 nm) for light absorption in the photoreceptor (Birge, 1981). This large shift of the optical transition of retinal is called the "opsin shift."

Recent solid-state NMR measurements showed differences in the ^{13}C chemical shift at positions 11, 12, and 13 in the comparison between the free retinal and the protein-bound form. These data confirm the point charge model for the opsin shift, i.e., the hypothesis that the shift is a consequence of a

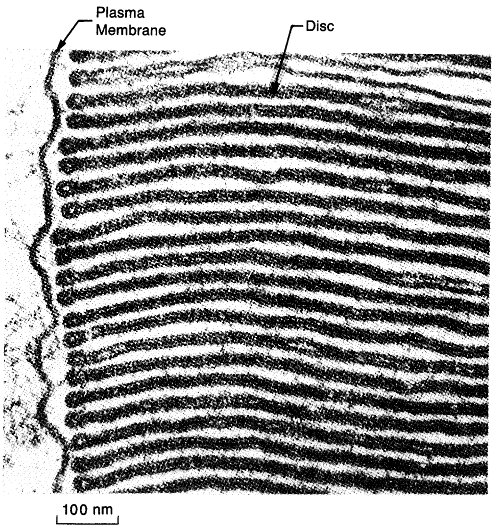
Plasma Membrane
Disc
100 nm

FIG. 16. Electron micrograph of disks in rod outer segment. [From Dowling (1967) and reproduced in Rodeick (1973).]

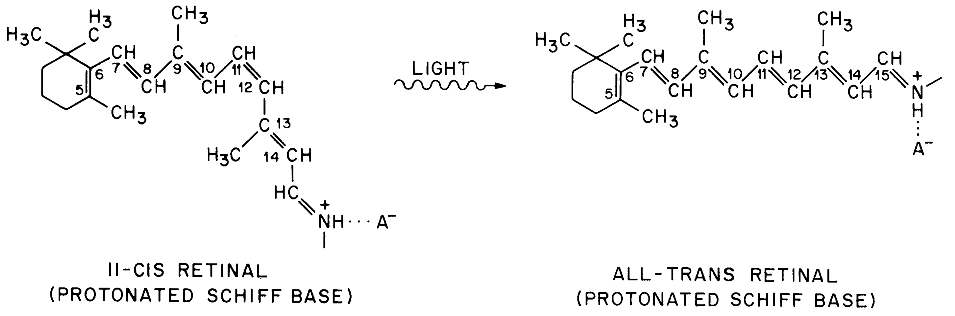

FIG. 17. Retinal, the visual pigment, depicted in its 11-*cis* (left) and its all-*trans* (right) forms. In both cases the aldehyde group is replaced by a protonated Schiff base linkage, as in most of the physiological forms.

negative charge located near positions 11, 12, and 13 (Lugtenburg *et al.*, 1988). Site-directed mutagenesis has established that this negative charge arises from glutamic acid residue 113 (Zhukovsky and Oprian, 1989). NMR data have also confirmed the orientation of the ring with respect to the polyene chain as being 6-*S*-*cis*.

Many different techniques have led to an understanding of the structure of rhodopsin and to the process which eventually leads to the production of free retinal. Our current picture of rhodopsin is summarized in Fig. 18 (Stryer, 1987). A number of biophysical studies contributed to the early understanding of the structural properties of the membrane-bound protein (Chabre, 1985). For example, neutron scattering studies demonstrated that 50% of the protein is embedded in the membrane bilayer (Saibil *et al.*, 1976). Similarly, small-angle scattering studies showed that rhodopsin consists of a hydrophobic region sandwiched between two hydrophilic ends (Osborne *et al.*, 1978). Infrared measurements of proton/deuteron exchange demonstrated that there are large regions of the protein that are not exchangeable, consistent with a hydrophobic region passing through a membrane (Osborne and Nabedryk-Viala, 1977). Circular dichroism measurements demonstrated that about 50% of the protein is alpha-helical in character (Shichi and Shelton, 1974). This latter finding, combined with other data, was taken as evidence that the alpha helices form part of the transmembrane structure. In summary, the biophysical measurements led to a view that rhodopsin is a transmembrane protein with alpha helices passing through the membrane and with hydrophilic sections on both sides of the membrane. These conclusions are supported by the amino acid sequence analysis in which seven hydrophobic sections were identified that would be expected to form alpha helices (Hargrave *et al.*, 1983). In addition, Henderson and Unwin (1975) built a three-dimensional structure of a related protein, bacteriorhodopsin, from high-resolution electron microscopic measurements of two-dimensional crystalline sheets. (Bacteriorhodopsin is a membrane-bound energy-transducing protein found in the purple membrane of *Halobacterium halobium.* Its active phototransduction element is also retinal.) The structural analysis done by Henderson and Unwin demonstrated that it, too, has seven transmembrane helical segments. Fluorescence energy-transfer measurements in-

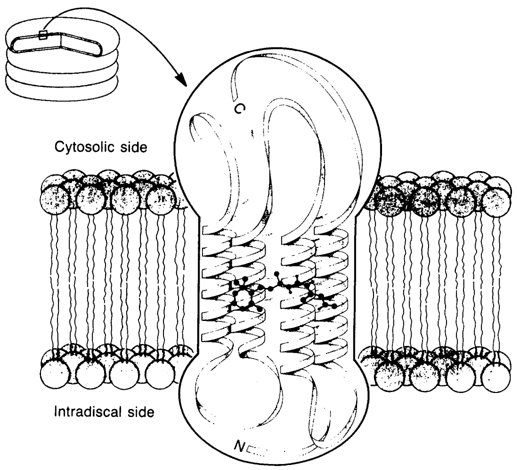

FIG. 18. Schematic drawing of a patch of membrane from a disk of a rod cell showing the proposed structure of rhodopsin. [From Stryer (1988).]

dicated that retinal lies in the center of the hydrophobic region of the protein (Stryer *et al.*, 1982).

The absorption of light by 11-*cis* retinal of rhodopsin brings about a structural change of the chromophore, causing it to isomerize to an all-*trans* form (Fig. 17). The large difference between the conformation of the all-*trans* retinal and that of the 11-*cis* retinal induces a series of conformational changes in the protein which surrounds it. There are several intermediate structures of the retinal associated with the evolving structure of the protein. Many physical techniques have been applied to study these intermediates to determine their photophysics, structure, and kinetics (Wald, 1968; Ebrey and Honig, 1975; Birge, 1981; Chabre, 1985). In the final step retinal is released by the protein and is only reconstituted with it after an enzyme converts the expended all-*trans* form back to the 11-*cis* isomer for reincorporation.

The process leading to the change in the rhodopsin structure and the associated release of the retinal have been very extensively studied. A molecular-level basis for these first steps in the information transduction process, known as the photocycle (Fig. 19), has been established. Information has been obtained both from direct studies of rhodopsin and from studies of bacteriorhodopsin, the analog of rhodopsin obtained from photosynthetic bacteria. Isomerization of the 11-*cis* retinal to the all-*trans* form has been confirmed by several studies. However, it was thought for some time that a *cis-trans* isomerization could not occur on the picosecond time scale. Molecular-dynamics calculations demonstrate that this isomerization can occur on the picosecond time scale (Birge, 1981). Furthermore, resonance Raman studies comparing model compounds and bathorhodopsin (the latter which forms on the picosecond time scale) indicate the presence of a strained all-*trans* isomer with a protonated Schiff base (Palings *et al.*, 1987). (A Schiff base is formed by the facile condensation of the aldehyde group of retinal with the amine terminus of a nearby lysine residue.) As expected for the very fast transition from rhodopsin to bathorhodopsin, changes in the protein are very small upon formation of bathorhodopsin. In contrast, infrared absorption studies have shown that there are larger changes in the protein in the next step, i.e., the conversion of bathorhodopsin to lumirhodopsin (Ganter *et al.*, 1988). The retinal remains a protonated Schiff-base all-*trans* isomer. However, the environment around the Schiff-base linkage changes, apparently in preparation for its deprotonation in metarhodopsin II. The change in the opsin shift presumably results from the changes in the electrostatic environment in this region of the retinal. Further relaxation to metarhodopsin I occurs on the 50- μs time scale. In this form, retinal in the protein has the structure of a fully relaxed, protonated Schiff-base all-*trans* isomer. Deprotonation occurs for the formation of the meta II intermediate at about 20 ms. This latter intermediate is referred to as R^*, the photoexcited state which is the enzymatically active form of the protein. This form is stable for about 1 min after which the Schiff base of the all-*trans* retinal is hydrolyzed and the retinal separates from the opsin. It is later reconstituted after enzymatic conversion to the 11-*cis* isomer.

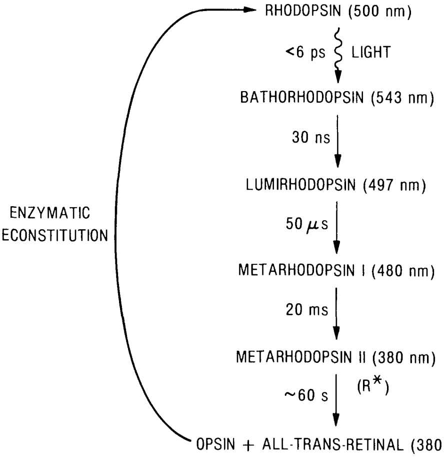

FIG. 19. Intermediate states in the thermal relaxation of rhodopsin following photoexcitation.

3.3 Electrical Response Induced by Molecular Conformation Changes

The primary event in the information transduction process discussed above results in a conformational transition in rhodopsin, i.e., the formation of R^*. Whereas this step was solved primarily by spectroscopic techniques, the next step, that of converting this structural information to an electrical impulse, was solved primarily by electrophysio-

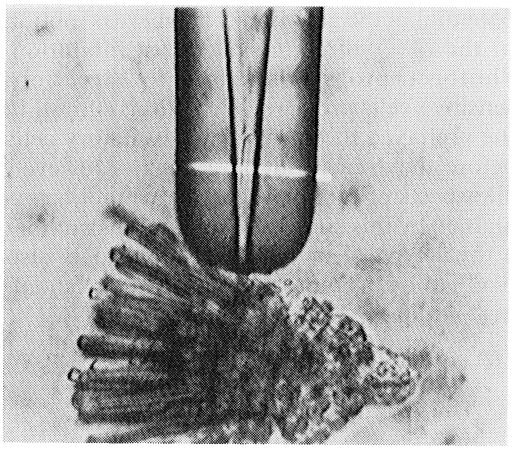

FIG. 20. Suction electrode method for measuring photoexcitation of a single rod cell. A seal between the electrode and the cell membrane allows for the removal of the remaining section of retina. Measurements of the single cell can then be made as a function of optical and chemical environment. [From Lewis and Del Priore (1988), from a photograph supplied by D. A. Baylor of Stanford University.]

logical measurements (Fig. 20) (Pugh and Cobbs, 1986; Stryer, 1987; Schnapf and Baylor, 1987). These were difficult experiments which in the end led to a very elegant model of how a structural change can be highly amplified so that single-photon events impinging on the retina can generate electrical signals which are processed by the brain.

Electrophysiological measurements showed that upon exposure to light the ionic permeability of the rod outer segment decreased. The conductance channels that are blocked by light absorption were found to be sodium channels. Normally, cyclic-GMP keeps the membrane sodium channels open. By blocking of these channels the sodium level in the photoreceptor cell decreases as there is also a sodium/potassium pump that is constantly pumping sodium out of the cell. The effect of this is a hyperpolarization, i.e., the potential inside the cell becomes more negative. This hyperpolarization, illustrated in Fig. 21 for several different light intensities, is transmitted to the synapse in the inner rod segment and the optical information is thereby passed to other cells for processing.

This mechanism of the energy transduction is so powerful that a detectable signal is generated by the absorption of a single photon by the photoreceptor. The amplification responsible for this sensitivity occurs during the chemical steps involved in the sodium channel closing. The first step in this process is the interaction of the light-modified rhodopsin (R^*) with a molecule called transducin, a signal-coupling protein. A subunit dissociation occurs in the transducin due to the conformational change in the rhodopsin. A single rhodopsin molecule can induce the subunit dissociation in 500 transducin molecules. These in turn activate an enzyme which hydrolyzes cyclic-GMP. No longer cyclic, GMP is thus unable to maintain the sodium channels in an open state, so the sodium influx stops and the cell hyperpolarizes. Each activated enzyme is capable of cleaving several thousand cyclic-GMP molecules. A single photon absorbed in the photoreceptor blocks the entry of a million sodium ions and results in a measurable nerve impulse!

3.4 Conversion of Electrical Signals into Visual Images: Motion and Color

Whereas the response of a rod cell to a single photon is large enough to be recorded, the response of a single cone cell is perhaps 100-fold lower and too small to be measured. However, the cone response is about four

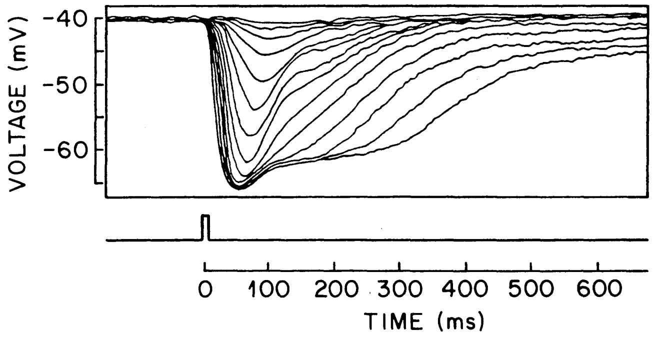

FIG. 21. Hyperpolarization of a photoreceptor cell as a function of the strength of the optical flash. Each flash is a factor of 2 more intense than the previous. [From Schnapf and Baylor (1987).]

times faster than the rod response. These different light-level and temporal sensitivities account for sensitive but monochrome night vision mediated by the rods as well as motion and color detection at high light levels by the cones (Schnapf and Baylor, 1987). There are three types of cone cells that have absorption maxima in the blue, green, and red regions of the visible spectrum. Color discrimination of light of a particular wavelength is computed by sensing the ratios of the responses of the three types of cone cells toward the light.

So far we have seen that the lens of the eye forms a pattern of light on the photoreceptors, and how the light impinging on these photoreceptors is first converted into molecular/chemical information, which is then transduced and amplified into an electrical signal. We examine how the nerve impulses, now encoding the amount of light received, are processed by the brain to interpret visual images, sensitive to form, movement, and color. The photoreceptors, oddly enough, are not on the surface of the retina; rather they line the back of the retina (see Fig. 14). The chemical and electrical signals generated by the photoreceptors are passed through several cell layers to the ganglion cells, which pass this information via the optic nerve to the brain.

The ganglion cells in certain species are capable of detecting the direction of motion. Certain cells respond to movement in only one direction. At its simplest level, motion in the preferred direction produces an action potential in the ganglion cell, whereas motion in the other direction produces inhibition. Photoreceptors connected to direction-sensitive cells are arranged so that stimuli in the preferred direction reach excitatory cells before they reach inhibitory cells. However, all aspects of motion detection cannot be explained by this simple model; it turns out that a third type of synaptic mechanism is also likely to be present, called "silent" or "shunting" inhibition (Poggio and Koch, 1987).

3.5 The Primary Visual Cortex

The primary visual cortex, also known as area 17, is located at the back of the brain in humans. It receives signals from the optic nerves after the signals have passed through the lateral geniculate nucleus (Fig. 22). The function of the lateral geniculate nucleus is to serve as a type of relay mechanism. Information from the left half of the retinas of both eyes passes through the left lateral geniculate nucleus, and then goes to the left-hand side of the brain. Thus, the left-hand side of the brain receives inputs from both eyes. The same is true for the right-hand side of the brain. However, within the lateral geniculate nucleus, cells in layers 1, 4, and 6 receive inputs from one eye, and the cells in layers 2, 3, and 5 receive inputs from the other eye. The segregation of left and right visual information is preserved as the signals pass into the visual cortex.

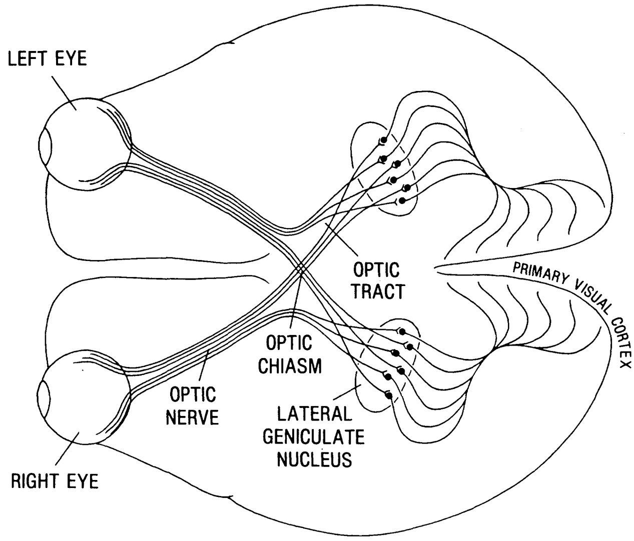

FIG. 22. Visual pathway is traced schematically in the human brain, seen here from below. The output from the retina is conveyed, by ganglion-cell axons bundled in the optic nerves, to the lateral geniculate nuclei; about half of the axons cross over to the opposite side of the brain, so that a representation of each half of the visual scene is projected on the geniculate of the opposite hemisphere. Neurons in the geniculates send their axons to the primary visual cortex. [From Hubel and Wiesel (1979).]

The primary visual cortex is a layered structure, wherein the six layers (layers I through VI) consist of dense and less-dense regions of cells. Early cell-staining work suggested that information coming into the cortex is processed in a very localized manner. That is, through synaptic connections, a signal can traverse the entire depth of the cortex in about four cells, whereas laterally, information is confined.

Basic experiments aimed at understanding the function of particular cells, whether in the cell layers of the eye, in the lateral geniculate nucleus, or in the cortex, consist of microelectrode recordings of single nerve fibers. Often carried out on monkeys, they are performed by presenting the animal with a visual stimulus, say a particular color, a dot, or a moving line. The responses to all these stimuli are recorded, and the cell is classified as to its function. Results from experiments of this sort showed that the receptive fields of the cells in the eye and in the lateral geniculate nucleus can be generally described as having "center-surround" behavior, i.e., a central circular excitatory field surrounded by an inhibitory one. They thus serve primarily to make comparisons between light levels in their receptive fields. Within the cortex the light-dark information is transformed into detection of line segments. Certain cells respond most strongly to lines or bars tilted at defined orientations. At this point the image is being operated on bit-by-bit, in a very localized fashion. At some point, the information must be analyzed to produce the perceived visual image.

The left and right visual information is combined within the cortex. This does not happen in the region that receives its input directly from the lateral geniculate nucleus, nor, in monkey, within the simple cells, which continue to preserve monocular information in what are called "ocular dominance" columns. The information from the two eyes is finally combined in subsequent projections.

Different regions of the primary visual cortex project into different regions of the brain. For example, the deepest layer, layer VI, projects back to the lateral geniculate nucleus, whereas layers II and III project to other regions of the cortex. Layer V projects to a visual center in the midbrain. It is likely that depth and color perception occur at these higher cortical areas. The study of these higher levels of perception and stereopsis rely heavily on psychophysics experiments and are somewhat out of the realm of biophysics.

The foregoing discussion has centered around the neural wiring of the visual system and the electrophysiological experiments that have established ocular dominance and sensitivity to motion and light intensity. Although genetic information dictates which neurons are initially wired together, these existing synapses are refined and new connections are made as the brain develops (Kalil, 1989). A series of very important biophysics experiments established many of the developmental aspects of the visual system. Torsten Wiesel and David Hubel in the 1960s sutured shut one of the eyes of newborn kittens and showed that activity of both eyes was necessary for the development of ocular dominance columns. Later work by Stryker and Harris (1986) showed through pharmacological blocking experiments that coincident nerve impulses, or action potentials, were essential for normal development. Work concerning the mechanism underlying these findings is of current research interest, and is presenlly focused on the protein molecules (receptors) and neurotransmitters involved in strengthening synaptic pathways (Kalil, 1989).

3.6 Summary

Experiments on the visual system have covered the full cycle, from the molecules responsible for the primary event in photoreception, to the membranes and organelles (rods and cones) which sense this information. Light absorption, converted to chemical information, is transmitted by electrical signals. These electrical signals convey information about form and motion to higher areas of organization in the brain, where perception takes place. Yet neural wiring in the brain is not a strict function of morphology and genetics; it is use-dependent. Hence, from the vantage point of this high level of organization, we are forced again to study the molecules, this time to uncover the mechanisms underlying synaptic strengthening.

This example of the visual system typifies the diversity and span of biophysics. Full understanding of this one aspect of biology— here the sense of vision— requires research at all levels of structure and function, including molecules, molecule-molecule interactions, assemblies, organelles, cells, and organs. Progress or a breakthrough in just one of these areas invariably opens up new lines of inquiry, which ripple through the entire fabric of the subject. Thus in biophysics, as in other areas of sciences, ideas and knowledge build upon an interplay between the results from various techniques and between the different levels of investigation.

GLOSSARY

Acetylcholine: A small molecule which is a neurotransmitter.

Action Potential: Transient electrical signal caused by alterations of the electrical properties of the membrane.

Adenine: A purine base found in nucleotides. Its one-letter code is A.

Adenosine Diphosphate (ADP): The product which results after one phosphate group has been hydrolyzed from ATP.

Adenosine Monophosphate (AMP): The product which results after two phosphate groups have been hydrolyzed from ATP.

Adenosine Triphosphate (ATP): A nucleotide containing three phosphate units. Hydrolysis of the phosphates provides energy for numerous biological functions.

***α* Helix:** A common structural motif found in proteins. It is stabilized by a hydrogen-bond network formed between the proton on the amino group of one amino acid and the carbonyl group of an amino acid three units away.

Amino Acids: Primary building blocks of proteins. They all contain a carboxyl group and an amino group, and one of 20 side chains. The amino acids condense through their carboxy and amino groups to form the polypeptide chains of proteins.

Antibody: A large protein molecule that recognizes a specific target molecule. Antibodies form the basis of the immune response.

Antigen: A foreign chemical substance or protein which is the target molecule recognized by an antibody.

Axon: The conduit which carries information from the neuron.

Bacteriorhodopsin: A membrane-bound, energy-transducing protein found in the purple membrane of *Halobacterium holobium.* Its active phototransduction element is retinal, the same chromophore present in the visual protein, rhodopsin.

Bases: The primary building blocks of DNA and RNA. These are nitrogen-containing ring compounds classed as either purines or pyrimidines. In DNA the purines are adenine and guanine; the pyrimidines are cytosine, and thymine. In RNA, thymine is replaced by uracil.

***β*-Pleated Sheet:** A common structural motif found in proteins. It usually contains an amino acid such as glycine or alanine which allows the polypeptide chains to be fully extended rather than coiled as in the *α* helix. Large stretches of this structure form a pleated sheet-like structure in which the polypeptide chains are either parallel or antiparallel to each other. Stabilization occurs by hydrogen bonding between an amino group on one chain and a carbonyl group on an adjacent chain.

Biophysics: An interdisciplinary field of science in which physical techniques and physical measurements are applied to biological problems. It combines physics, chemistry, and biology.

Cell: The simplest unit capable of reproduction in living organisms. It contains cytoplasm in which there are many different organelles and molecules and is surrounded by a membrane.

Channels: Proteins spanning a membrane and containing a pore, through which specific ions are allowed to pass.

Chemotaxis: Chemically directed movement of cells.

Chloroplast: An organelle found in plants which is responsible for photosynthesis.

Chromatin: The assembly of nuclear DNA and histones, which contains the complete genetic information for the cell.

Codon: The three-base code in DNA and RNA that specifies a particular amino acid.

Coenzymes: Small molecules associated with enzymes that are essential for enzymatic activity.

Cones: Photoreceptor cells largely responsible for color vision.

Cortex: The outer region of the brain, concerned with higher perceptual, cognitive, and motor functions.

Cytosine: A pyrimidine base that is found in nucleotides. Its one-letter code is C.

Cytosol: The intracellular fluid.

DNA—Deoxyribonucleic Acid: A biopolymer composed of a sugar (deoxyribose)- phosphate backbone which has bases (adenine, guanine, cytosine, and thymine) attached to it. DNA often forms a double helix and is the genetic material of all cells.

Dendrites: The fibers (processes) that carry information to the neurons.

Deoxyribose: A five-carbon sugar that occurs in DNA.

Depolarization: The process by which the cell's membrane potential becomes less negative.

Differentiation: The process during cell development, by which a cell acquires specialized characteristics.

Electrophysiology: The technique used to study the currents and electrical potentials associated with cells.

Enzyme: Proteins that catalyze specific biochemical reactions or transformations.

Eukaryotic: All plants and animals other than bacteria and certain algae. All of the cells in eukaryotes contain several organelles and have nuclei. The nuclei contain most of the cell's DNA.

Excitation: The neural process by which the electrical activity of a cell is increased.

Gene: A highly specific stretch of DNA along a chromosome that codes for a particular product.

Genetic Code: The mapping between the sequence of bases in DNA and specific amino acid residues.

Genome: Full set of chromosomes in a cell. The genome completely specifies the organism.

Glycosidic Linkage: A bond between a base and a sugar, found in nucleotides.

Guanine: A purine base that is found in nucleotides. Its one-letter code is G.

Guanosine Monophosphate (GMP): A nucleotide essential for the visual process.

Heme: An iron atom surrounded by a porphyrin macrocyclic ring, responsible for such diverse functions as oxygen storage and electron transfer.

Hemoglobin: A tetrameric heme-containing protein. It stores and transports oxygen.

Histone: A protein assembly inside the nucleus around which the DNA is wound.

Hyperpolarization: The process by which the cell's membrane potential becomes more negative.

Immunocytochemical: A technique in which cells are stained with antibodies which recognize specific chemicals in the cell.

Inhibition: The neural process by which the electrical activity of a cell is decreased.

Lateral Geniculate Nucleus: An electrical relay system which is part of the visual process.

Lipid: A molecule containing a hydrophobic tail and a hydrophilic head group. A fatty acid is a lipid.

Membrane Potential: The voltage difference across the cell's plasma membrane.

Membrane: A lipid bilayer which contains numerous proteins. It serves to isolate organelles or cells from their surroundings and also maintains chemical gradients and electrical potentials.

Messenger-RNA (*m*-RNA): The RNA which contains the genetic message, ultimately translated in the ribosome.

Mitochondrion (pl. Mitochondria): An organelle found in eukaryotic cells, responsible for ATP generation and respiration.

Neuron: A nerve cell. It receives, conducts, and transmits electrical signals.

Neurotransmitter: Small molecules that are released in the synapse and are responsible for chemical transmission of information in neurons.

Nucleic Acids: Linear polymers of nucleotides, linked by the phosphate of one unit and the sugar of the next. Examples are DNA and RNA.

Nucleolus: The part of the nucleus that contains RNA and proteins and functions to synthesize and assemble ribosomal RNA.

Nucleosome: A DNA-histone assembly.

Nucleotide: A basic building unit of nucleic acids. A nucleotide consists of a purine or pyrimidine base, a five-carbon sugar, and one or more phosphate groups.

Nucleus: The central organelle of eukaryotic cells. It contains a complete copy of the organism's DNA, and some of the machinery for genetic replication.

Ocular Dominance Columns: Regions of the visual cortex that are specifically associated with one eye or the other.

Organelles: Subcellular components that carry out highly specific functions. Examples include ribosomes, chloroplasts, and mitochondria.

Patch Clamping: Electrophysiological technique in which the electrode forms a high impedance "patch" with a portion of a membrane, enabling one to monitor channel activity.

Phospholipid: The most ubiquitous biological lipid consisting of a fatty acid hydrocarbon tail and a polar head group.

Photosynthesis: The biological transduction of light energy into chemical energy.

Photosynthetic Reaction Center: Membrane-bound assembly of several proteins responsible for photosynthesis. It is the first membrane-bound protein whose crystal structure was solved.

Polypeptide Chain: A linear polymer of amino acids.

Primary Structure: The linear sequence of monomer units that forms a nucleic acid or a polypeptide.

Procaryotes: Bacteria and certain algae whose cells contain simple internal structures.

Protein: A linear chain of amino acids, folded into a highly specific three-dimensional shape. Other than water, proteins are the most abundant molecules of the cell and take part in all aspects of cellular function and structure such as catalysis, transport, immune response, etc.

Pumps: Specific proteins responsible for maintaining ionic and *p*H gradients across the membrane.

Purine: A nitrogen-containing ring compound, derivatives of which form the bases adenine and guanine found in nucleotides.

Pyrimidine: A nitrogen-containing ring compound, derivatives of which form the bases thymine, cytosine, and uracil found in nucleotides.

Quaternary Structure: The relationship between subunits in multiple-subunit proteins, such as hemoglobin.

Random Coil: The random state of protein structure.

Receptors: Proteins that have specific binding sites for target molecules. Once the receptor has bound its target molecule, it undergoes a structural or chemical change, enabling it to control a specific biological function.

Replication: The process of forming identical copies of DNA from existing DNA.

Respiration: The biological oxidation of small molecules in which oxygen is the ultimate electron acceptor. Respiration generates the high-energy molecule ATP.

Resting Potential: The potential across the membrane of an electrically inactive (not spiking) cell.

Retinal: The chromophore of the visual system. Its structure is related to vitamin A and it undergoes a conformational change when it absorbs light.

Rhodopsin: A membrane-bound protein of the visual system. It contains retinal as its prosthetic group.

Ribose: A five-carbon sugar found in RNA.

Ribosome: An organelle in which protein synthesis takes place. The genetic message contained in RNA is converted into a specific string of amino acids on the ribosome.

RNA—Ribonucleic Acid: A biopolymer composed of a sugar (ribose)-phosphate backbone which has bases (adenine, guanine, cytosine, and uracil) attached to it.

Rod: A photoreceptor cell largely responsible for night vision.

Secondary Structure: In nucleic acids, the arrangement of bases which determines the efficiency of the replication process. In proteins, the secondary structure is generated by local interactions between adjacent amino acids of the polypeptide chain. The most frequently found protein secondary structures are α helices and β sheets.

Substrate: A molecule which is acted upon by an enzyme.

Synapse: The junction formed between neurons.

Tertiary Structure: The three-dimensional arrangement of helices and sheets with respect to each other. It involves both long- and short-range amino acid interactions.

Thymine: A pyrimidine base that is found in nucleotides. Its one-letter code is T.

Transcription: The process by which the genetic message in DNA is converted into various RNAs, including messenger-RNAs, transfer-RNAs, and ribosomal-RNAs.

Transfer-RNA (*t*-RNA): The RNA responsible for transporting specific amino acids to the ribosome for their ultimate incorporation into polypeptides.

Translation: The process in which the message contained in RNA is converted into the specified amino acid sequence.

Uracil: A pyrimidine base that is found in nucleotides. Its one-letter code is U.

Voltage Clamp: Electrophysiological technique in which the membrane potential is controlled by injecting current into the cell.

List of Works Cited

Alben, J. A., Fiamingo, F. G. (1984), in: D. L. Rousseau (Ed.), *Optical Techniques in Biological Research,* New York: Academic, pp. 133–179.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J. D. (1983), *Molecular Biology of the Cell,* New York: Garland.

Amrein, M., Stasiak, A., Gross, H., Stoll, E., Travaglini, G. (1988), *Science* **240**, 514–516.

Anderson, P. W. (1972), *Science* **177**, 393–396.

Asher, S. (1988), *Annu. Rev. Phys. Chem.* **39**, 537–588.

Becker, W. M. (1986), *The World of the Cell,* Menlo Park, CA.

Binnig, G., Quate, C. F., Gerber, Ch. (1986), *Phys. Rev. Lett.* **56**, 930–933.

Birge, R. R. (1981), *Annu. Rev. Biophys. Bioeng.* **10**, 315–354.

Cantor, C. R., Schimmel, P. R. (1980), *Biophysical Chemistry, Part II: Techniques for the Study of Biological Structure and Function,* San Francisco: Freeman.

Carter, P., Wells, J. A. (1987), *Science* **237**, 394–399.

Chabre, M. (1985), *Annu. Rev. Biophys. Biophys. Chem.* **14**, 331–360.

Cram, D. J. (1988), *Science* **240**, 760–767.

Dagani, R. (1988), *Chem. Eng. News* **66**, 26–29.

Deisenhofer, J., Epp, O., Miki, K., Huber, R., Michel, H. (1985), *Nature (London)* **318**, 618–624.

Dervan, P. (1986), *Science* **232**, 464–471.

Dowling, J. E. (1967), in: J. M. Allen (Ed.), *Molecular Organization and Biological Function,* New York: Harper & Row.

Ebrey, T. G., Honig, B. (1975), *Q. Rev. Biophys.* **8**, 129–184.

Estell, D. A., Graycar, T. P., Miller, J. V., Powers, D. B., Ng, P. G., Wells, J. A. (1986), *Science* **233**, 659–663.

Franklin, R. E., Gosling, R. G. (1953), *Acta Crystallogr.* **6**, 673–677.

Ganter, U. M., Gartner, W., Siebert, F. (1988), *Biochemistry* **27**, 7480–7488.

Getzoff, E., Geysen, H. M., Rodda, S. J., Alexander, H., Tainer, J. A., Lerner, R. A. (1987), *Science* **235**, 1191–1196.

Gould, S., Marti, O., Drake, B., Hellemans, L., Bracker, C. E., Hansma, P. K., Keder, N. L., Eddy, M. M., Strucky, G. D. (1988), *Nature (London)* **332**, 332–334.

Hansma, P. K., Tersoff, T. (1987), *J. Appl. Phys.* **61**, R1–R24.

Hargrave, P. A., McDowell, J. H., Curtis, D. R., Wang, J. K., Juszczak, E., Fang, S. L., Rao, J. K. M., Argos, P. (1983), *Biophys. Struct. Mech.* **9**, 235–244.

Henderson, R., Unwin, P. N. T. (1975), *Nature (London)* **257**, 28–32.

Hubel, D. H. Wiesel, T. N. (1979), *Sci. Am.* **241**(3), 150–162.

Iversen, L. L. (1979), *Sci. Am.* **241**(3), 134–139.

Jelinski, L. W. (1984), *Chem. Eng. News* **62**, 26–47.

Jelinski, L. W., Behling, R. W., Tubbs, H. K., Cockman, M. D. (1989), *Am. Biotech. Lab.* **7**, 34–41.

Jelinski, L. W., Torchia, D. A. (1980), *J. Mol. Biol.* **138**, 255–272.

Kalil, R. E. (1989), *Sci. Am.* **261**, 76–85.

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., Shore, V. C. (1960), *Nature* **185**, 422–427.

Kessel, R. G., Kardon, R. H. (1979), *Tissues and Organs: A Text Atlas of Scanning Electron Microscopy,* New York: Freeman.

Kosman, D. J. (1984), in: D. L. Rousseau (Ed.), *Structure and Resonance Techniques in Biological Research,* New York: Academic, pp. 99–244.

Lewis, A., Del Priore, L. V. (1988), *Phys. Today* **41**(1), 38–46.

Lugtenburg, J., Mathies, R. A., Griffin, R. G., Herzfeld, J. (1988), *TIBS* **13**, 388–393.

Maestre, M. F. (1984), in: D. L. Rousseau (Ed.), *Optical Techniques in Biological Research,* New York: Academic, pp. 291–341.

Marrack, P. (1987), *Science* **235**, 1311–1313.

Mason, E. B. (1983), *Human Physiology,* Menlo Park, CA: Benjamin/Cummings.

Miles, F. A. (1969), *Excitable Cells,* London: William Heineman Medical Books.

Moffat, K. (1984), in: D. L. Rousseau (Ed.), *Structural and Resonance Techniques in Biological Research,* New York: Academic, pp. 363–435.

Morris, P. G. (1986), *NMR Imaging in Medicine and Biology,* Oxford: Clarendon.

Nagai, K., Luisi, B., Shih, D., Miyazaki, G., Imai, K., Poyart, C., DeYoung, A., Kwiatkowsky, L., Noble, R. W., Lin, S.-H., Yu, N.-T. (1981), *Nature (London)* **329**, 858–860.

Nauta, W. J. H., Feirtag, M. (1979), *Sci. Am.* **241**(3), 88–111.

Nicholls, D. G. (1982), *Bioenergetics*, New York: Academic.

Osborne, H. B., Nabedryk-Viala, E. (1977), *FEBS Lett.* **84**, 217–220.

Osborne, H. B., Sardet, C., Michel-Villaz, M., Chabre, M. (1978), *J. Mol. Biol.* **123**, 177–206.

Palings, I., Pardoen, J. A., van den Berg, E., Winkle, C., Lugtenburg, J., Mathies, R. A. (1987), *Biochemistry* **26**, 2544–2556.

Perutz, M. E. (1979), *Annu. Rev. Biochem.* **48** 327–386.

Piez, K. A., Trus, B. L. (1978), *J. Mol. Biol.* **122**, 419–432.

Poggio, T., Koch, C. (1987), *Sci. Am.* **256**(5), 46–52.

Pugh, Jr., E. N., Cobbs, W. H. (1986), *Vision Res.* **26**, 1613–1643.

Rebek, J., Jr. (1987), *Science* **235**, 1478–1484.

Rodeick, R. W. (1973), *The Vertebrate Retina*, New York: Freeman.

Rose, D. F., Smith, P. D., Sato, S. (1987), *Science* **238**, 329–335.

Rousseau, D. L., Friedman, J. M. (1988), in: T. G. Spiro (Ed.), *Biological Applications of Raman Spectroscopy. Vol. 3: Resonance Raman Spectra of Heme and Metalloproteins*. New York: Wiley, pp. 133–215.

Rousseau, D. L., Ondrias, M. R. (1984), in: D. L. Rousseau (Ed.), *Optical Techniques in: Biological Research*, New York: Academic, pp. 65–132.

Saibil, H., Chabre, M., Worcester, D. (1976), *Nature (London)* **262**, 266–270.

Schlessinger, J., Elson, E. L. (1982), in: G. Ehrenstein and H. Lecar (Eds.), *Methods of Experimental Physics. Vol. 20: Biophysics*. New York: Academic, pp. 197–227.

Schnapf, J. L., Baylor, D. A. (1987), *Sci. Am.* **256**(4), 40–47.

Scott, R. A. (1984), in: D. L. Rousseau (Ed.), *Structural and Resonance Techniques in Biological Research*, New York: Academic, pp. 295–362.

Schichi, H., Shelton, E. (1974), *J. Supramol. Struct.* **2**, 7–16.

Seeley, T. D. (1989), *Am. Sci.* **77**, 546–553.

Shepherd, G. M. (1988), *Neurobiology*, 2nd ed., New York: Oxford Univ.

Smith, R. L., Oldfield, E. (1984), *Science* **225**, 280–288.

T. G. Spiro (Ed.) (1988), *Biological Applications of Raman Spectroscopy. Vol. 3: Resonance Raman Spectra of Heme and Metalloproteins*. New York: Wiley.

Stevens, C. F. (1979), *Sci. Am.* **241**(13), 54–65.

Stryer, L. (1978), *Annu. Rev. Biochem.* **47**, 819–846.

Stryer, L. (1981), *Biochemistry*, 2nd ed. New York: Freeman.

Stryer, L. (1987), *Sci. Am.* **257**(1), 32–40.

Stryer, L. (1988), *Biochemistry*, 3rd ed., New York: Freeman.

Stryer, L. L., Thomas, D. D., Maeres, C. F. (1982), *Annu. Rev. Biophys. Bioeng.* **11**, 203–222.

Stryker, M. P., Harris, W. A. (1986), *J. Neurosci.* **6**, 2117–2133.

Wald, G. (1968), *Nature (London)* **219**, 800–807.

Watson, J. B. (1976), *Molecular Biology of the Gene*, Menlo Park, CA: Benjamin.

Watson, J. D., Crick, F. H. C. (1953), *Nature (London)* **171**, 737–738.

Wilkins, M. H. F., Randall, J. T. (1953), *Biochem. Biophys. Acta* **10**, 192–193.

Yguerabide, J., Yguerabide, E. E. (1984), in: D. L. Rousseau (Ed.), *Optical Techniques in Biological Research*, New York: Academic, pp. 181–290.

Zasadzinski, J. A. N., Schneir, J., Gurley, J., Ellings, V., Hansma, P. K. (1988), *Science* **239**, 953–1068.

Zhukovsky, E. A., Oprian, D. D. (1989), *Science* **246**, 928–930.

Further Reading

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J. D. (1983), *Molecular Biology of the Cell*, New York: Garland.

Cantor, C. R., Schimmel, P. R. (1980), *Biophysical Chemistry*, Parts I, II, and III. San Francisco: Freeman.

Kandel, E. R., Schwartz, J. H. (1985), *Principles of Neuroscience*, 2nd ed. New York: Elsevier.

Stryer, L. (1988), *Biochemistry*, New York: Freeman.

Watson, J. B., Hopkins, N. H., Roberts, J. W., Steitz, J. A., Weiner, A. M. (1987), *Molecular Biology of the Gene*, 4th ed., Vols. I and II. Menlo Park, CA: Benjamin/Cummings.

21
Biorheology

Rudolf Dietrich Bauer and Tilmann Volk

21.1 Introduction 671
21.2 Rheological Concepts 671
21.2.1 Basic Constitutive Equations 672
21.2.2 Viscoelastic Models 672
21.2.2.1 Transient Tests 673
21.2.2.2 Oscillatory Tests 674
21.2.3 Pressure–Volume Relationships 675
21.2.4 Pressure–Flow Relationships 675
21.2.4.1 Laminar Flow 675
21.2.4.2 Turbulent Flow 676
21.2.4.3 Oscillatory Flow 677
21.3 Rheology of Body Fluids 678
21.3.1 Viscosity of Non-Newtonian Fluids 678
21.3.2 Instrumentation 679
21.3.3 Rheology of Blood 680
21.3.3.1 Macrorheology 680
21.3.3.2 Microrheology 681
21.3.4 Rheology of Viscoelastic Body Fluids 684
21.4 Hemodynamics 684
21.4.1 Heart Mechanics 685
21.4.1.1 Events of the Cardiac Cycle 686
21.4.1.2 Rheological Properties of the Heart 686
21.4.2 Rheological Approaches to Arteries 689
21.4.2.1 Quasistatic Mechanical Properties of Arteries 690
21.4.2.2 Dynamic Mechanical Properties of Arteries 692
21.4.3 Pressure and Flow in the Vascular System 693
21.4.3.1 Systemic Arteries 694
21.4.3.2 Pulmonary System 695
21.4.3.3 Arterial Dynamics 695
21.4.3.4 Pressure and Flow in Veins 699
21.5 Pneumodynamics 700

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

21.5.1 Functional Anatomy of the Lungs 701
21.5.1.1 Airways 701
21.5.1.2 Lung Volumes 701
21.5.1.3 Ventilation 702
21.5.2 Mechanics of Breathing 702
21.5.2.1 Mechanics of Inspiration 703
21.5.2.2 Mechanics of Expiration 703
21.5.3 Rheology of the Respiratory Tract 704
21.5.3.1 Rheology of the Lung 704
21.5.3.2 Rheology of the Chest 706
21.5.3.3 Rheology of the Lung–Chest System 706
21.5.4 Air Flow in the Respiratory Tract 707
21.5.4.1 Intrapulmonary Pressure 707
21.5.4.2 Distribution of Airway Resistances 707
21.5.4.3 Pressure–Air Flow Relationship 708
Glossary 708
Acronyms 709
References 710
Further Reading 710

21.1
Introduction

The science of the deformation and flow of matter was termed *rheology* by E. C. Bingham on the occasion of the formation of the American Society of Rheology in 1929. During the First International Congress on Rheology, held in Holland in 1948, A. L. Copley suggested the term *biorheology* for the application of rheology to biological systems. Later on, further terms were introduced, such as *hemorheology* – also suggested by Copley – for the branch of biorheology that is, concerned with blood and blood vessels, or *electrorheology* – suggested by G. W. S. Blair. As in rheology, the fundamental problem in biorheology is the formulation of constitutive equations. The basis for the mathematical description of the rheological behavior of biological systems is a detailed knowledge of anatomical structures and organic functions. Against this background, close interdisciplinary cooperation is an essential prerequisite for the successful application of rheological methods to mechanical problems in biology.

This review of biorheology deals with several problems inherent in hemorheology, hemodynamics, and pneumodynamics. In the first section, an introduction to rheological concepts is given, which includes a description of pressure–flow relationships for steady and pulsatile flow. In the second section, the rheology of body fluids, in particular that of blood in macrovessels and microvessels, is outlined. The basic techniques for studying blood rheology are also presented. The third section deals with the rheological properties of the myocardium and arterial mechanics, in particular the viscoelastic properties of the arterial wall, the propagation of pulse waves within the arterial system, and the pressure–flow relationships within the venous system. Finally, in the fourth section, the rheological behavior of the lung–chest system and the problems inherent in the airway resistance are discussed.

21.2
Rheological Concepts

An important task in biorheology is to establish mathematical relationships between deformation or flow of biological material and those processes associated with changes in dimensions and time.

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

21.2.1
Basic Constitutive Equations

Equations that define the relationships between stresses and dynamic variables, for example, strains, or, in the case of fluids, rates of strain, are called *constitutive equations.*

In the classical theory of elasticity, Hooke's law relates the stress σ applied to an elastic material to the change in extension, the strain ε:

$$\sigma = E\varepsilon \tag{1}$$

where the elastic modulus E is a proportionality constant that defines the elastic properties of the material. In the classical theory of hydrodynamics, Newton's law is a constitutive equation relating shear stress τ – which is the frictional force F over the area A – applied to a viscous fluid, to the shear rate γ – which is identical to the velocity gradient dv/dr within the fluid:

$$\tau = \eta\gamma \tag{2}$$

where the viscosity η is a proportionality constant that defines the viscous properties of the fluid, v is the velocity of the fluid, and r the radial coordinate.

In the case of infinitesimal changes in stress and strain, the behavior of many solids and fluids is well approximated by these laws. Marked deviations, however, can be observed when the changes are not infinitesimal, or when they apply to materials that combine elastic and viscous properties; solids and fluids exhibiting both elastic and viscous behavior are commonly termed viscoelastic.

21.2.2
Viscoelastic Models

Constitutive equations of viscoelastic solids and fluids are much more complicated than Eqs (1) and (2). Constitutive equations that permit a complete specification of the rheological behavior of biological material are still not available. A suitable way of describing the different types of behavior of materials, however, is to consider the analogy of the behavior of mechanical models. One approach to viscoelasticity is to describe elasticity and viscosity in terms of elastic springs and viscous dashpots, respectively, and to combine these elements to describe more or less complicated viscoelastic bodies.

The model most analogous to an ideal elastic material is the spring. The stress acting on a spring and the resulting strain are always in phase. Furthermore, the compressed spring can store energy, which can be used to do work when the spring is released. The mechanical model analogous to an ideal viscous fluid is the dashpot comprising a piston moving in oil. The dashpot is an energy-dissipating element. It cannot support any stress, which means that not the strain but the shear rate is directly proportional to the applied stress.

The simplest mechanical model of a viscoelastic solid is the parallel arrangement of a spring and a dashpot, known as the *Voigt (or Kelvin) model*, shown in Figure 21.1; it is a model of a solid, since the maximum extension of this model is limited by the elastic properties of the spring and the force applied. The simplest mechanical model of a viscoelastic fluid is a spring and a dashpot

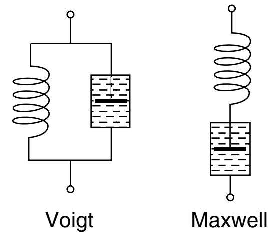

Fig. 21.1 Viscoelastic elements. Left: the Voigt element; right: the Maxwell element.

in series, known as the *Maxwell model* (Figure 21.1). This is a model of a fluid because any force applied leads to a continuous extension or flow of the model. More complex material that does not behave as an ideal viscoelastic solid or an ideal viscoelastic fluid can be described in terms of combinations of these basic two-element models.

21.2.2.1 **Transient Tests**

Fundamental rheological parameters can be measured using creep testing and by the use of the viscoelastic models. When a constant stress σ is applied to a Voigt solid, the spring cannot be extended instantaneously because of the viscous resistance of the dashpot arranged in parallel. As time proceeds, however, the strain gradually increases to a maximum value, which depends on the applied force and the elasticity of the spring (Figure 21.2a). When the stress is removed, the energy stored by the Voigt solid is also released and the system recovers its initial length. This behavior is termed *creep* and can be described by the following equation:

$$\varepsilon(t) = \sigma J(t) \tag{3}$$

where $\varepsilon(t)$ is the time-dependent strain, and $J(t)$ the creep compliance.

The quotient of the viscosity of the dashpot and the elasticity of the spring is a measure of the time required to extend the spring, which is retarded by the dashpot; this time is thus known as the *retardation time*.

When a Maxwell fluid is subjected to a constant strain ε, as shown in Figure 21.2b, there is an initial jump in the stress, reflecting the elastic extension of the spring. As time passes, the spring contracts and pulls the piston out of the

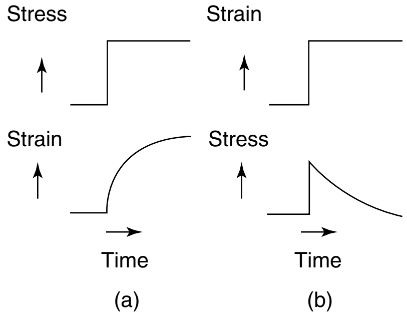

Fig. 21.2 Time dependence of stress and strain in (a) creep and (b) relaxation.

dashpot, so that the stress relaxes. When the spring reaches its initial length, the flow stops and the stress become zero. The relationship between the time-dependent stress and the constant strain is called *stress relaxation,* and is given by

$$\sigma(t) = \varepsilon G(t) \tag{4}$$

where $G(t)$ is the relaxation modulus.

In the case of the Maxwell model, the quotient of the viscosity of the dashpot and the elasticity of the spring is called the *relaxation time,* since it is a measure of the time required for stress relaxation.

21.2.2.2 **Oscillatory Tests**

In addition to these transient tests, dynamic tests are often used, during which the stress applied to material under investigation may be varied sinusoidally. If the material exhibits linear viscoelastic properties, the strain is also found to change sinusoidally, but there will be a phase shift between stress and strain. The relationship between stress and strain can be described by expressing the periodically varying stress and the relaxation modulus as complex quantities:

$$\sigma^*/\varepsilon = G^* = G' + jG'' \tag{5}$$

where σ^* is the complex stress, G^* the complex modulus, whose real part G' is the storage and whose imaginary part G'' the loss modulus, and $j = \sqrt{(-1)}$. For a description of the mechanical properties of viscoelastic solids, for example, blood vessels, the term elastic modulus is used instead of storage modulus. The loss modulus G'' is the product of the viscosity η and the angular frequency $\omega (= 2\pi f, f =$ frequency):

$$G'' = \omega\eta \tag{6}$$

G' is the in-phase relationship between stress and strain and is therefore associated with the elastic behavior and the energy storage and release of the material. G'' is the 90° out-of-phase relationship between stress and strain and is thus associated with the loss of energy in the form of heat due to the viscous resistance of the material.

When the storage and the loss moduli of a Maxwell model, and those of a Voigt model, are plotted against angular frequency, curves as shown in Figure 21.3 are obtained. In both models, G' increases with frequency, while G'' rises to a maximum at a frequency equal to the inverse of the relaxation and the retardation time, respectively. In most cases, biological materials deviate markedly from this simple behavior, so that the frequency dependence of the moduli is much more complicated.

Equations (1)–(6) are, strictly, valid only for linear viscoelastic materials. To overcome these restrictions, Frank (1920) introduced a differential definition of the elastic modulus E^*, relating an infinitesimal change in stress ($d\sigma$) to that in the corresponding strain ($d\varepsilon$):

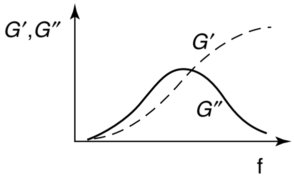

Fig. 21.3 Response of viscoelastic elements to sinusoidal stress. G': storage modulus; G'': loss modulus; and f: frequency.

$$E^* = \frac{d\sigma}{d\varepsilon} \tag{7}$$

This modulus, which is often referred to as *tangent elastic modulus*, permits a description of nonlinear viscoelastic solids when very small changes in stress are applied and the mean stress is varied over a wide range. This procedure is widely used for the examination of the mechanical properties of biological solids, for example, of the blood vessel wall (Section 21.4.2.2).

The elastic elements of the Maxwell and Voigt models are originally assumed to be Hookean and the viscous elements Newtonian. For the representation of nonlinear viscoelastic solids and liquids, these assumptions are no longer maintained, and merely the arrangement of the components of these models are termed Maxwell and Voigt.

21.2.3
Pressure–Volume Relationships

An alternative approach to the problem of characterizing the mechanical properties of structures, such as heart, lungs, or blood vessels, is the introduction of relationships between changes in pressure (dP) and changes in volume (dV) instead of those of stress and strain. Quantities that are frequently used are compliance C:

$$C = \frac{dV}{dP} \tag{8}$$

and the reciprocal of compliance, elastic stiffness E':

$$E' = \frac{1}{C} = \frac{dP}{dV} \tag{9}$$

Since compliance and elastic stiffness depend on the initial volume of the body, normalizations of C and E' are often employed, which are called *volume distensibility* C^*:

$$C^* = \frac{1}{V}\frac{dV}{dP} \tag{10}$$

and volume elasticity κ (the reciprocal of the volume distensibility):

$$\kappa = \frac{dP}{dV}V \tag{11}$$

21.2.4
Pressure–Flow Relationships

The flow rate i is defined as fluid volume dV flowing through the cross-sectional area A of a vessel in time dt:

$$i = \frac{dV}{dt} \tag{12}$$

As can be seen from a consideration of dimensional similitude, the flow rate can also be written as:

$$i = Av_m \tag{13}$$

where v_m is the mean velocity of the fluid in the vessel.

The flow rate in a vessel is proportional to the pressure difference ΔP over the length of the vessel, and may be written, in accordance with Ohm's law:

$$i = \frac{\Delta P}{W} \tag{14}$$

This is also known as *Darcy's law*. The proportionality constant W is the viscous or frictional resistance to flow.

21.2.4.1 Laminar Flow

When the flow of a Newtonian fluid in a very long, straight, rigid, cylindrical vessel is laminar, the frictional resistance can be calculated from the vessel dimensions and the viscosity of the fluid.

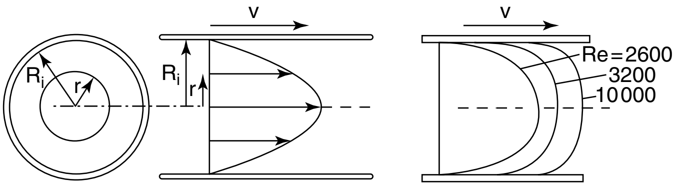

Fig. 21.4 Velocity profiles in a rigid, straight tube of a circular cross-section for steady (middle) and turbulent flow (right) with increasing Reynolds numbers (*Re*). R_i is the internal radius, r is the radial coordinate, and v is the velocity.

The fluid layers within the vessel are arranged in cylindrical form (Figure 21.4). The contact surface between two cylinders of radius r and length l is $2\pi rl$; thus, the viscous or frictional force F between these cylinders is, in accordance with Eq. (2), $F = 2\pi r\eta l dv/dr$. For the velocity v of the fluid to be steady, the frictional force has to be counterbalanced by the pressure difference ΔP over the length l acting on the cross-sectional area $r^2\pi$ of the cylinder:

$$\Delta Pr^2\pi = -2\pi r\eta l\frac{dv}{dr} \qquad (15)$$

Integrating and applying the boundary condition that the fluid rests on the vessel wall ($r = R_i$ = internal radius) gives the following expression:

$$v = \Delta P(R_i^2 - r^2)/(4\eta l) \qquad (16)$$

Thus, the velocity profile is parabolic, as shown in Figure 21.4, irrespective of the dimensions of the vessel and the properties of the fluid. With this type of flow, the velocity of the fluid is a maximum in the center of the vessel.

The flow rate i is obtained by integrating Eq. (16) once from $r = 0$ to $r = R_i$:

$$i = \Delta PR_i^4\pi/8\eta l \qquad (17)$$

This is known as *Hagen–Poiseuille's law.*

According to Eq. (14), the frictional resistance to flow is

$$W = 8\eta l/R_i^4\pi \qquad (18)$$

where W is proportional to the length of the vessel and to the viscosity of the fluid, and inversely proportional to the fourth power of the internal radius. It should be noted that this resistance to flow is always the smallest possible for given vessel dimensions and viscosity. The resistance to flow is far greater in all other types of flow than in the Hagen–Poiseuille type. To obtain a lower limit, Eq. (18) is often used for the estimation of resistances to flow in arteries, although none of the assumptions necessary for the evaluation of Eq. (17) is applicable to the circulation.

21.2.4.2 Turbulent Flow

In turbulent flow, the velocity components have random turbulent fluctuations superimposed on their mean values, so that it is not possible to characterize the motion of individual elements mathematically. The change from laminar to turbulent flow is determined by the entrance of the vessel, wall roughness, and Reynolds number *Re*; this number is defined by a dimensionless combination of the mean velocity v_m, the viscosity η, the density ρ of the fluid, and the internal diameter D_i of the vessel:

$$Re = D_i v_m \rho / \eta \tag{19}$$

The critical Reynolds number for the transition from laminar to turbulent flow is 2000–2200. However, if sources of disturbance such as roughness, vibration, or sharp corners are eliminated at the entrance of the vessel, transition can occur at Reynolds numbers as high as 40 000–50 000.

As shown in Figure 21.4, too, the velocity profile in turbulent flow is virtually uniform near the center of the vessel, with a very steep velocity gradient near the wall. Furthermore, the velocity profile depends on the Reynolds number; when Re increases, the velocity distribution becomes more and more uniform. The flow rate is no longer directly proportional to the pressure difference ΔP, but increases almost proportionally to the square root of ΔP over the vessel length. Hence,

$$\Delta P = K\, i^2 \tag{20}$$

where K is a proportionality constant.

The formation of eddies within a tube, for example, at branches, does not necessarily lead to true turbulence, since such disturbances may disappear again some distance downstream. This type of flow is called *transitional.*

21.2.4.3 **Oscillatory Flow**

The equations of motion for a viscous Newtonian fluid are the Navier–Stokes equations. In general vector form, these equations are

$$\frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{v} \cdot \nabla)\boldsymbol{v} = -\frac{1}{\rho}\mathrm{grad}P + \frac{\mathbf{F}}{\rho} + \frac{\eta}{\rho}\nabla^2 v \tag{21}$$

where ν is the fluid velocity, ρ and η are density and viscosity, P is the pressure of the fluid, $\mathbf{F}$ is an external force, for example, gravity, and ∇ is the nabla and ∇^2 the Laplace operator.

Exact solutions of the Navier–Stokes equations can only be obtained for the flow in situations permitting reduction of these equations to linear differential equations. A description of pulsatile flow in rigid vessels requires the application of various approximations.

1. The velocities must be small, so that the convective terms, $(\nu\cdot\nabla)\ \nu$, can be neglected.
2. All velocities except those in the axial direction are very small.
3. There is no pressure gradient in the radial direction.
4. The fluid velocity is independent of the axial position, that is, the total fluid mass within the vessel is accelerated and decelerated simultaneously.
5. Forces acting from the outside, such as gravity, can be neglected.

If these assumptions apply, the Navier–Stokes equations reduce to

$$\frac{\partial v_z}{\partial t} = \frac{1}{\rho}\frac{\partial P}{\partial z} + \frac{\eta}{\rho}\left(\frac{\partial v_z}{\partial r^2} + \frac{1}{r}\frac{\partial v_z}{\partial r}\right) \tag{22}$$

where v_z is the velocity in the axial direction, z the axial, and r the radial coordinate. Furthermore, assuming pulsatile flow under steady-state conditions, both the pressure gradient $\partial P/\partial z$ and the axial fluid velocity v_z can be written as components of Fourier series:

$$-\frac{\partial P}{\partial z} = P_z \exp(j\omega t) \quad \text{and}$$
$$v_z = V_z \exp(j\omega t) \tag{23}$$

P_z is the complex amplitude of the pressure gradient and V_z the complex amplitude of the velocity in the axial direction at vessel radius r. Under the boundary condition that the fluid does not slip on the wall ($v_z = 0$ at $r = R_i$), and using Eq. (23), Eq. (22) becomes

$$v_z = \frac{P}{j\omega\rho}\frac{J_0((r/R_i)\alpha_w\sqrt{-j})}{J_0(\alpha_w\sqrt{-j})}\exp(j\omega t) \tag{24}$$

As abbreviation, the dimensionless number $\alpha_w = R_i\sqrt{\omega\rho/\eta}$, which characterizes kinematic similarities in the oscillating fluid, is introduced. J_0 is a Bessel function of the first kind and zero order. Equation (24) gives the distribution of the fluid velocity in a rigid vessel as a function of the radius and angular frequency.

Using Eq. (24), the axial velocities of fluid layers and their distribution within the vessel can be calculated. For values of α_w below 1, all velocity profiles are virtually parabolic throughout pulse cycle. With increasing α_w, the velocity distributions become more and more uniform near the center, the velocity amplitudes decrease, and the velocity lags behind the pressure; this is shown in Figure 21.5 for three different α_w values. Furthermore, as shown in this figure, reversal of velocity begins in the layers near the wall, and the phases of the velocities near the center lag behind those of the marginal layers. At very high α_w values, the central layers oscillate like a solid body, so that the steepest velocity gradients are near the wall. An application of these computations for pulsatile flow in arteries is given in Section 21.4.3.3, Velocity Profiles in Arteries.

21.3
Rheology of Body Fluids

21.3.1
Viscosity of Non-Newtonian Fluids

Under conditions of unidirectional, laminar flow and constant temperature, Newton's law (Eq. (2)) can be used to define the viscosity η:

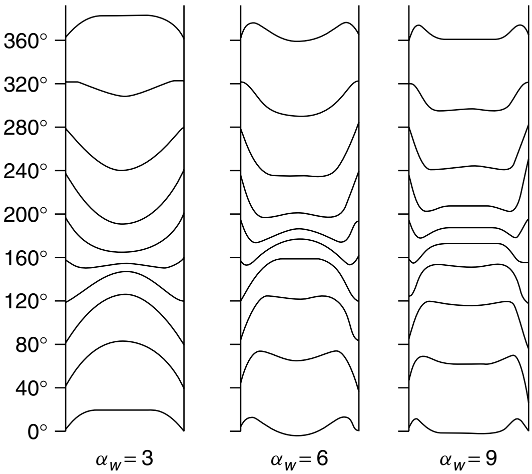

Fig. 21.5 Velocity profiles resulting from a sinusoidal pressure gradient in a rigid tube at three different values of α_w, as calculated with Eq. (24). Profiles are given at intervals of 40°. In each case, the zero value of velocity is indicated by a horizontal line in the left margin. The magnitudes of the velocities are multiplied (from left to right) by 1, 4, and 8, respectively.

$$\eta = \tau/\gamma \tag{25}$$

For homogeneous or Newtonian fluids, η is a constant of the fluid, which is independent of shear stress and shear rate and, thus also of the tube dimensions and the velocity of the fluid.

In the case of inhomogeneous fluids, the relationship between τ and γ is not the simple one proposed in Eq. (25); the viscosity of such fluids depends not only on temperature but also on shear rate and/or shear stress:

$$\eta = \eta(\tau, \gamma) \tag{26}$$

Fluids exhibiting variable viscosity are called *non-Newtonian*. In particular, suspensions of rigid particles in a homogeneous solution are numbered among the non-Newtonian fluids. The rheological behavior of a suspension is determined by the viscosity of the continuous phase and by the nonlinear viscoelastic properties of the suspended particles. Therefore, the Newtonian constitutive equation cannot represent the behavior of fluids exhibiting a variable η. If this fact is ignored and Eq. (25) is applied to a non-Newtonian fluid, the so-called *apparent viscosity* (η_{app}) is determined.

The particles impede the free movement of adjacent laminas of the streaming suspension. This gives rise to the formation of a central cell column with a thickness that is in the order of the dimension of the particles and results in an increase in apparent viscosity. The degree of obstruction of the laminar flow depends on the size, shape, number, and deformability of the suspended particles.

21.3.2
Instrumentation

When one investigates the rheology of biological fluids, specially designed equipment is required. The measurement of viscous or viscoelastic properties of liquids is carried out on viscometers that permit analysis of stress acting on a fluid for a given shear rate. The principle of such instruments is that of two parallel plates at a given distance y apart, with a fluid between them (Figure 21.6). A tangential force F acts on the top plate in such a way that this plate is moved relative to the resting bottom plate at a constant velocity v. The marginal fluid layer adhering to the top plate has the velocity v, while the velocity of the marginal fluid layer adhering to the bottom plate is zero. The velocities of the fluid layers between these two marginal layers increase linearly from the bottom to the top, so that a linear velocity gradient $-dv/dy$, referred to as *shear rate* γ, is developed. The force per unit area F/A is known as *shear stress* τ. According to Eq. (25), the plot of F/A vs $-dv/dy$ characterizes the viscosity of the fluid.

The parallel plate system is not a suitable viscometer, since the fluid cannot

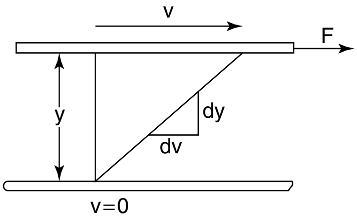

Fig. 21.6 Velocity distribution in a steady-state laminar flow between a fixed ($v = 0$) and moving plate (v).

be kept in place. Many instruments have been developed which overcome these difficulties. Viscometric devices designed for the measurement of the viscous properties of blood, blood plasma, and other body fluids, which are generally based on this two-plate principle and are capable of keeping the fluid in place, can be placed in one of three categories (Figure 21.7): cone and plate, cone in cone, and Couettetype (or narrow-gap concentric cylinder) viscometers. Within all such viscometers, the fluid under test is kept in the space between the cone and the plate, the cone and the cone, or the two cylinders. One unit is rotated and the drag transmitted by the fluid to the other unit is measured by the angular deflection of this unit or by the force required to return it to its original position.

A second type of viscometer is the capillary tube viscometer, which is often used in the measurement of macro- and microrheological properties of blood.

A third type, in which the resistance of a fluid to a body moving through it is measured, has also been employed in various forms of falling-sphere viscometers.

21.3.3
Rheology of Blood

Whole blood is a non-Newtonian fluid, since it is a suspension of formed elements – including red cells (erythrocytes), white cells (leucocytes), and platelets – in plasma. Plasma consists mainly of proteins, at an average concentration of 7% of the total volume, which are dissolved in a saline solution. In contrast to the results of earlier experiments, plasma is now described as a Newtonian fluid, so that it is a continuous suspension medium for the formed elements. Normally, at 37 °C the viscosity of plasma is 1.5–1.8 mPa s, that is, 1.5–1.8 times as great as that of water ($\eta_{\text{water}} = 1$ mPa s).

Apart from other parameters, the apparent viscosity of blood increases when the concentration of the formed elements increases. The ratio of blood cells to total blood volume is known as the *hematocrit*. For human blood, the hematocrit is normally about 45%, and the viscosity in large vessels at 37 °C is about 4–6 mPa s. The normal value of the hematocrit permits optimal oxygen transport within the vascular system. If the hematocrit increases – as in polycythemia – the potential advantage for increased oxygen transport capacity is offset by the increase in blood viscosity, thus reducing – at a given pressure – the flow to the tissues (Eq. (17)).

21.3.3.1 Macrorheology

The non-Newtonian behavior of blood is characterized by an increase in apparent viscosity when the shear stress (or the shear rate) decreases (Figure 21.8). The dependency of the apparent viscosity on shear stress is smaller when the hematocrit is lower. The high apparent viscosity under creep flow is due to cell aggregation, that is, the attraction of one erythrocyte for another, which under low

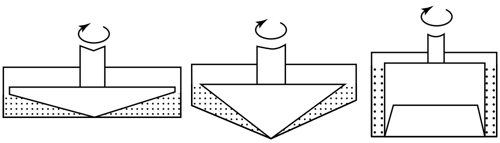

Fig. 21.7 Viscometers. Left: cone and plate; middle: cone in cone; right: Couette type.

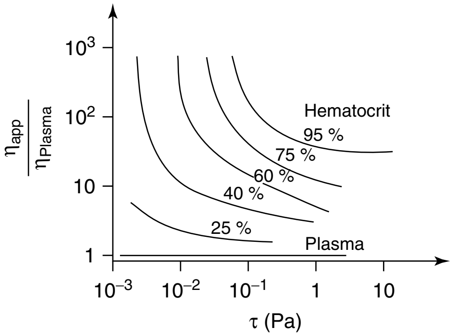

Fig. 21.8 Apparent viscosity (η_{app}) of human blood related to the viscosity of plasma (η_{plasma}) as a function of shear stress (τ). At each curve the hematocrit is given.

shear stress results in the formation of rolls of cells (rouleaux). When shear increases, the cells disaggregate, and the rouleaux disperse; this results in a decrease in apparent viscosity.

The aggregation of blood cells is accomplished under the influence of large proteins, for example, fibrinogen and α_2 globulins, the so-called agglomerines. These proteins support the aggregation of erythrocytes which impede their movement, thus the blood flow. At low shear rates, any increase in fibrinogen concentration leads to an increase in the apparent blood viscosity; the viscosity of a suspension of erythrocytes in serum, which is plasma lacking fibrinogen, is lower than that of a suspension of erythrocytes in plasma that has the same hematocrit.

21.3.3.2 **Microrheology**

In tubes with internal diameters greater than about 1 mm, whole blood behaves like a Newtonian fluid. Only in the microcirculation does the anomalous rheological behavior of blood become evident. In Figure 21.9 the relationship between the apparent viscosity and the hematocrit of red cell suspensions, streaming with relatively high velocities through tubes of various diameters, is shown. At tube diameters of 1 mm or more, the apparent viscosity increases markedly with increasing hematocrit of the suspension. This increase in viscosity is less pronounced for smaller tube diameters. In tubes with diameters of 11 µm, the apparent viscosity is almost independent of the hematocrit, and equals that of the solution.

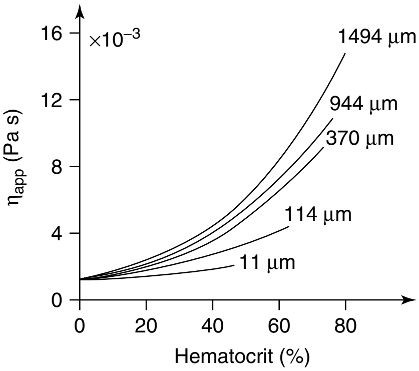

Fig. 21.9 Apparent viscosity (η_{app}) of suspensions of erythrocytes in physiological saline solution at 25 °C as a function of hematocrit. At each curve, the internal diameter of the perfused tube is given (Modified from Haynes, 1962).

For the same hematocrit, the apparent viscosity of the suspension decreases as the tube size becomes smaller. This rheological behavior of blood is due mainly to the deformability of the red cells, since these cells are the major formed elements in blood, occupying more than 95% of this volume. At rest, the shape of the erythrocytes is generally that of a biconcave discoid, the diameter of which is about 8 μm. Because of the fluid content of the erythrocytes and the high flexibility of their membrane, streaming blood behaves like an emulsion of liquid droplets rather than a suspension of rigid corpuscles. The shear stresses acting on the membranes cause rotation of the cells, which leads to deformation of the erythrocytes and to a sliding of the membrane relative to the viscous fluid within the cell. Thus, some force acting at right angles to the axis of the vessel is exerted on red cells near to the wall and carries the cells toward the axis. This radial migration results largely in the deformability of erythrocytes – as the deformability becomes less, so does the migration velocity – and the distribution of the shear rate in the fluid, which is high near the wall and low near the axis. Therefore, the radial migration velocity is greatest near the wall and decreases as the cells move toward the axis.

As a result of this migration, there is a higher concentration of blood cells near the axis than at the wall. The reduction in the number of red cells within the marginal fluid layers creates an almost cell-free zone of low viscosity, while in the middle of the vessel a central column is formed, in which the flexible erythrocytes are packed close together. Because of this process, the shear rate within the marginal zone further increases, while that in the vicinity of the axis is decreased, resulting in the formation of a nonsheared plug of cells, which moves at a relatively high velocity in a marginal zone of low viscosity. The narrower the tube, the greater the shearing force acting on the cell-free layers in the vessel diameter and the more pronounced the reduction in apparent viscosity. Therefore, the apparent viscosity of the blood is reduced. This decrease in apparent viscosity as the vessel diameter becomes smaller is referred to as the *Fåhraeus–Lindqvist effect* after the scientists who first reported it (Fåhraeus and Lindqvist, 1931). This effect starts at tube diameters of about 300 μm; the minimum value of apparent viscosity is reached at diameters of 8–9 μm. In these vessels, the apparent viscosity is only about 10–15% greater than the viscosity of plasma. In smaller vessels, the apparent viscosity increases again.

Dynamic Hematocrit The mean transit time of blood cells is reduced relative to the marginal plasma zone, since the central cell column streams faster through the blood vessel than the marginal layers. Therefore, the *dynamic* hematocrit in tubes exhibiting a marked cell-free marginal zone is smaller than the hematocrit in larger vessels (Fåhraeus effect; Fåhraeus, 1929). A proportion of the plasma leaks out of the capillary near the arterial end and returns in the distal zones. This extracapillary flow can help to explain the shorter residence time of blood cells in capillaries. The Fåhraeus effect is thought to be partly responsible for the reduction in the apparent viscosity in narrow tubes. It should be noted, however, that the relative volumes of blood cells and plasma moving through the microcirculation per unit time do not differ from the relative volumes defined by the *static* hematocrit.

Blood Flow in Capillaries It is not possible to give a uniform description of the systemic microcirculation, since there are considerable differences due to the specific vascular organization of each individual organ. Majno (1965) proposed a classification of capillaries on the basis of the continuity of endothelial cells into three main types: continuous, fenestrated, and discontinuous. Continuous capillaries are found in the skin, muscle, and central nervous system. The endothelial cells of capillaries in the intestine and endocrine glands and the glomeruli and the peritubular capillaries in the kidney manifest sievelike intracellular circular fenestrae having a diameter of 4 nm. Discontinuous capillaries with gross intercellular gaps exist in the sinusoids of the liver, spleen, and bone marrow. A further peculiarity that markedly influences shear stress and shear rate is seen in lung capillaries; blood is considered to flow around the alveoli in a space between two surfaces separated by posts having an irregular arrangement.

In spite of these substantial differences in detail, the passage of a suspension of relatively large particles at high concentrations is common to all capillaries. Therefore, capillary blood flow is generally discussed on the basis of typical capillary elements, namely continuous tubes ranging in diameter between 10 and 3 μm. At diameters down to about 8 μm, red blood cells can pass along a capillary without any distortion, but they have to flow in single file surrounded by a *zone* of plasma. In this case, the shape and deformability of the red cells do not noticeably influence capillary flow. Furthermore, at a diameter of about 10 μm, the apparent viscosity of blood is at its lowest and is independent of the hematocrit, since red cells can flow freely through this vessel, and an increase in the number of blood cells results only in a reduction in the distance between the cells within the single file, with no change in shear stress or shear rate.

In capillaries having a diameter of less than 8 μm, the erythrocytes must undergo deformation in order to pass through the lumen and normal red cells may adopt a variety of shapes, for example, parachute, drop, or arrow-head. Normal cells are capable of considerable distortion without any change in either surface area or volume. When the cells move along small capillaries, a lubricating layer of plasma is formed between the cell membranes and the vessel wall, which reduces the frictional forces between these surfaces. In order to keep the surfaces separate and to deform the cells adequately, high pressures must be generated locally. Therefore, shear stress increases at the same shear rate, so that, with decreasing tube diameter, apparent viscosity also increases. This is the reverse of the Fåhraeus–Lindqvist effect, which becomes apparent when the tube diameter is smaller than the diameter of the red cells.

Rheological Behavior of Pathological Blood Abnormal cells that are hardened or almost spherical cannot be deformed so easily, and therefore cannot pass undamaged through the smallest vessels. The deformability of erythrocytes further depends on the concentration of calcium ions in the plasma. When the blood contains less oxygen, calcium ions can enter the red cells in excess. This increase results in greater stiffness of the erythrocytes and, thus, in an increase in blood viscosity. In disorders such as sickle cell anemia, an increase in viscosity due to decreased deformability of red cells is also found. The spleen and kidneys of patients suffering from this type of disease are gradually destroyed by repeated blockage of the microcirculation.

21.3.4
Rheology of Viscoelastic Body Fluids

In contrast to blood or blood plasma, a large variety of biological fluids such as respiratory, gastric, and cervical mucus, sputum, and saliva are highly complex materials having a pronounced gel-like structure. Such fluids possess viscoelastic properties. With respect to both physiological and pathological processes, the rheological behavior of these biological materials is very important, since the viscoelastic properties of such fluids play a fundamental role in the clearance of both normal and pathological material.

For the investigation of sputum, saliva, and mucus, all types of viscometers have been widely used. In many cases, the viscoelastic model has been employed as a convenient starting point, and experimental results have been interpreted in terms of storage and loss modulus, as well as relaxation or retardation times (Section 21.2.2).

Sputum exhibits highly variable complex non-Newtonian rheological behavior. The characteristic relaxation time of sputum is in the order of several minutes; it is probably the most elastic of body fluids, to the extent that it may be termed *superelastic*. The viscoelastic properties already of normal sputum, however, are changed significantly by alterations in pH or concentration of substances. For instance, the values reported for viscosities of sputum range from 0.01 to 10^5 Pa s.

Saliva is viscoelastic, with similar properties to sputum; its storage modulus is strongly related to the mucin concentration.

The range of variation in the viscoelastic properties of normal mucus is also considerable, apparently with random variation in pH and concentration of substances being responsible.

Such wide variations point out the difficulty of comparing values obtained with different viscometers and different ways of collecting viscoelastic body fluids. These variations emphasize the need for rigid standardization of procedures to obtain representative data.

21.4
Hemodynamics

The circulatory system consists of two pumps – the cardiac ventricles – and a closed-loop system of vessels – the arteries, microvasculature, and veins – all filled with blood. The primary role of the cardiovascular system and the blood is to ensure the survival of all cells of the body, which at all times must be provided with materials rich in oxygen and metabolic substrates and have their metabolic wastes eliminated. The heart also permits the transport of hormones (which act as chemical signal transmitters), immunologically active substances, and heat between various regions of the body. The rate of flow of the blood is adapted to the conveyance of nutrients and wastes – which must be transported fastest – and the respiratory gases, oxygen and carbon dioxide.

A fundamental aim of hemodynamics is to describe the rheological behavior of both the heart and the blood vessels. To be able to do this, a detailed knowledge of the anatomical structures and functions of all parts of the circulatory system is essential.

21.4.1
Heart Mechanics

The weight of an average adult male heart varies between 280 and 340 g, that of the female heart between 230 and 280 g. With the exception of the valves and the fibrous "skeleton" separating the atria and ventricles and surrounding the tricuspid and mitral as well as the smaller aortic and pulmonary valve orifices, the heart consists mainly of muscle. There are two kinds of muscle fibers within the heart, specialized muscle cells arranged in the form of nodes and fibers, which mediate the rhythmic excitation of the heart, and the working muscle fibers of the atria and the ventricles, the myocardium.

By virtue of the rhythmic contraction of the heart and the function of the four heart valves, the blood is propelled in the mean unidirectionally through the blood vessels. As shown in Figure 21.10, the uniform direction of the blood in the various cavities of the heart and blood vessels is made possible by the atrioventricular valves between the atria and the ventricles, the mitral valve on the left and the tricuspid valve on the right of the heart, and the semilunar valves between the ventricles and the main arteries, the aortic and the pulmonary valves, which prevent the backflow of blood. From the left atrium the blood enters the left ventricle. The contraction of the left ventricle forces the blood into the aorta and the systemic arteries, which convey it to the organs of the body. From the peripheral tissues and organs, the blood returns through the systemic veins and the two venae cavae to the right atrium and ventricle. The contraction of the latter expels the blood through the pulmonary artery into the vessels of the lungs, from which it returns through the lung veins into the left

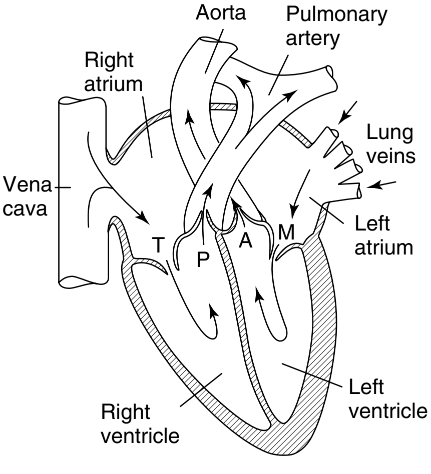

Fig. 21.10 Blood flow through the heart. Atrioventricular valves: *M*, mitral valve; *T*, tricuspid valve. Semilunar valves: *A*, aortic valve; *P*, pulmonary valve.

atrium and ventricle, thus completing the circulation.

21.4.1.1 Events of the Cardiac Cycle

Both atria are thin-walled muscular chambers while the ventricles have thick muscular walls – especially the left ventricle that has about twice the thickness and three times the mass of the right. Within one cardiac cycle, systole (contraction of the heart musculature) alternates rhythmically with diastole (relaxation of the ventricles).

Systole comprises two phases, the preejection period (PEP, line 1 in Figure 21.12) and the ejection period (EP, curve 2 in Figure 21.12). When the heart begins to contract, the ventricular pressure exceeds the pressure in the atria, which leads to the closing of the atrioventricular valves. During PEP, all four valves of the heart are closed, and the ventricular pressure increases rapidly, while the ventricular volume remains unchanged (isovolumetric contraction). When the ventricular pressures exceed the pressures in the roots of the aorta and the pulmonary artery, the semilunar valves open and the EP begins. The myocardial fibers shorten and about half of the blood volume contained in the ventricles is expelled into the aorta and pulmonary artery. The pressure within the ventricle reaches a maximum and then decreases. When the ventricular pressures become a little lower than the pressures in the aorta and the pulmonary artery, a very brief backflow of blood into the ventricles leads to closure of the semilunar valves, and diastole begins.

The diastole also comprises two phases: the isovolumetric relaxation (line 3 in Figure 21.12) and the ventricular filling phase (curve 4 in Figure 21.12). During the relaxation all valves are closed and the ventricular pressure falls rapidly. This phase ends when the ventricular pressure falls below the pressures in the atria. The atrioventricular valves now open, and the ventricular filling phase begins. Toward the end of the diastole, contraction of the atria helps to fill the ventricles. At the beginning of the next ventricular contraction, the atrioventricular valves are closed again, and the next cycle begins.

21.4.1.2 Rheological Properties of the Heart

For a description of the passive and the active properties of the heart, the following variables are available: ventricular pressure, ventricular volume and wall thickness, blood flow in the roots of aorta and pulmonary artery, and time. On the basis of these variables, the forces operative within the ventricular wall and the contraction velocities of the myocardium can be determined. The mechanical behavior of the intact heart is best described in terms of its stress–strain relation, rather than its force–length relation. With such parameters it is possible to make valid comparisons of the mechanical behavior of ventricles of different shapes and sizes.

Stress–Strain Relationships The ventricular wall must develop stress that counterbalances the ventricular pressure. Using simultaneous recordings of pressure and ventricular geometry, for example, by means of ultrasound echocardiography, computed tomography (CT), or magnetic resonance imaging (MRI) it is possible to quantify ventricular wall stress. The complex configuration of the ventricle, however, precludes an accurate determination of all the dimensions and stresses of the heart musculature without simplifying assumptions. For this purpose, the ventricles have usually been approximated by a sphere, and the myocardium is treated as an isotropic and homogeneous medium with linear elastic properties. If one assumes that the thickness of the ventricular wall h is uniform and small compared with the inner and outer radii, stress can be calculated by

$$\sigma = \frac{PR}{2h} \tag{27}$$

This equation is known as *Laplace's law*, and is widely used by many investigators for determining ventricular wall stress because of its simplicity and a lack of experimental evidence validating more sophisticated stress formulas (see Mirsky, 1979). Despite the fact that Laplace's law can lead to erroneous results if used in thick-walled left ventricles or in hypertrophic hearts – Laplace's law generally gives an overestimation of midwall stress – this simple equation reveals fundamental and important relationships, which are also true for the heart *in vivo*:

1. The greater the radius of the ventricle, the greater the ventricular wall stress at the same ventricular pressure. Therefore, dilation of the heart leads to an increase in wall stress, that is, in the load of the myocardial fibers.
2. The greater the wall thickness, the smaller the stress within the ventricular wall at the same ventricular pressure. Hypertrophy of the heart leads to a reduction in the stress within the ventricular wall.

The mechanical behavior of the heart muscle as a viscoelastic material which is rhythmically stretched and compressed can be described, according to Eq. (7), by the tangent elastic modulus. If the distribution of stress within the heart and all dimensions are known, the tangent elastic modulus at any point of the myocardium can be determined. Since our knowledge of ventricular stress is incomplete, the mechanical properties of the heart muscle are often represented in terms of the simplest possible analog models. In the late 1930s, Hill showed that the mechanical behavior of activated skeletal muscle could be described by means of a two-element model consisting of a contractile element (CE) representing the contractile processes, and an elastic element in series (series elastic element (SE)) representing the elastic properties of muscle fibers (Hill, 1938). In contrast to skeletal muscle, active myocardium develops a significant restoring force when stretched. Thus, in a model of the heart muscle, an additional elastic element (parallel elastic element (PE)) is needed, which can be arranged in parallel either to CE (Voigt model) or to the combination of CE and SE in series (Maxwell model), as shown in Figure 21.11. In both models, CE is responsible for generating force and shortening the myocardial fibers; during rest it is assumed that CE is freely extensible and exhibits no appreciable viscoelastic properties. SE and PE are assumed to behave as passive nonlinear springs whose lengths are determined by the forces.

Various mechanisms of the contraction of the heart have been proposed in terms of the Maxwell and Voigt models. The use of these models for the interpretation of the rheological properties of the

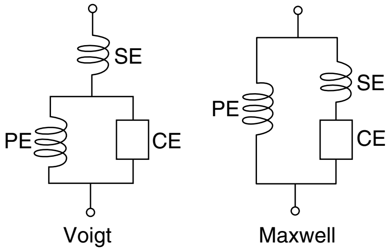

Fig. 21.11 Models of the heart muscle. Left: the Voigt model; right, the Maxwell model. PE: parallel elastic element; SE: series elastic element; and CE: contractile element.

heart muscle must, however, be treated with caution, since the stress–strain relationship of SE is time dependent, which implies that viscous elements may have to be included in these models. Furthermore, SE depends on the instantaneous state of contraction of the myocardium, which results in a nonsteady behavior of this element. The most appropriate model of the heart that represents actual performance best has not yet been established.

Pressure–Volume Relationships In general terms, the dependency of the time course of ventricular pressure $P(t)$ on ventricular volume V can be described as

$$P(t) = f(E', V) + g\left(W_v, \frac{dV}{dt}\right) + h\left(M', \frac{d^2V}{dt^2}\right) \quad (28)$$

where f, g, and h are symbols of functions. This equation shows that the pressure depends not only on the elastic stiffness (E', see Eq. (9)) and the volume of the heart, but also on the viscous resistance (W_v) of the ventricular wall, the rate of change of volume, the effective mass (M', see also Eq. (36)), and the acceleration of blood during ejection. The faster the filling of the ventricles and/or the ejection of blood, the greater the influence of the second and third terms on the ventricular pressure. Under normal conditions, however, the influence of these two terms is very small as compared with the first term.

Passive Properties of the Heart The passive behavior of the intact heart muscle can be determined only when the heart is at rest, that is, during diastole. For the nonactive heart, there is a progressive increase in pressure with increasing volume; this nonlinearity is due mainly to the increase in the tangent elastic modulus of the myocardial fibers with increasing wall tension.

The elastic stiffness of the ventricle determines the blood volume which can be stored within the ventricle during diastole. The stiffness depends on the thickness of the ventricular wall: thick, hypertrophic walls are stiffer than thinner walls; the left ventricle has a greater elastic stiffness than the right. Furthermore, the elastic stiffness of the heart depends on dimensions such that larger ventricles are stiffer than smaller ones at the same filling pressure; dilation of the ventricle increases its elastic stiffness.

Some actions do not markedly change the passive pressure–volume relationship of the left ventricle. For instance, the application of norepinephrine, which leads to an increase in the ability of the myocardium to contract and to develop force (which is sometimes referred to as *contractility*), or paralysis of the myocardium by means of potassium ions, has little influence on the passive mechanical behavior of the heart. However, during hypoxia, that is, decreased oxygen delivery to the heart muscle, elastic stiffness is increased due to incomplete relaxation of the muscle fibers. After myocardial infarction, the elastic stiffness of the myocardium is also increased, since fibrous tissue is incorporated into the heart musculature.

Active Properties of the Heart Suga, Kitabatake, and Sagawa (1979) found that for a given contractile state of the heart, instantaneous ventricular pressure is strongly correlated to instantaneous ventricular volume. On the basis of these findings, the authors proposed an equation for the instantaneous pressure–volume relationship of the active heart:

$$P(t) = E(t)[V(t) - V_0] \qquad (29)$$

where $E(t)$ is the quotient of the instantaneous pressure by the instantaneous ventricular volume, and V_0 an experimentally determined correction volume that corresponds approximately to the ventricular volume at zero pressure. The variation of $E(t)$ with time reflects the contraction of the myocardium (Figure 21.12). During the ejection period, $E(t)$ increases and reaches a maximum at the end of systole. It has been found that end-systolic (ES) maximum of $E(t)$ can be used for the characterization of the contraction state of the myocardium: ES increases with enhancement of the contractility of the heart. Generally speaking, this statement by the above-mentioned authors seems quite promising for the characterization of the mechanical properties of the active heart.

21.4.2
Rheological Approaches to Arteries

In common with all soft biological tissues, arteries are composed of highly distensible materials that exhibit nonlinear viscoelastic behavior. Under the conditions of cyclic loading and unloading, the combined effects of elasticity and viscosity result in hysteresis loops in the stress–strain diagram. In general, the cylindrical arterial walls show anisotropic behavior. Shear strains, however, are much smaller than the strains in the longitudinal, circumferential, and radial directions, which means that arteries are elastically symmetrical or orthotropic, and the three orthogonal strains are the

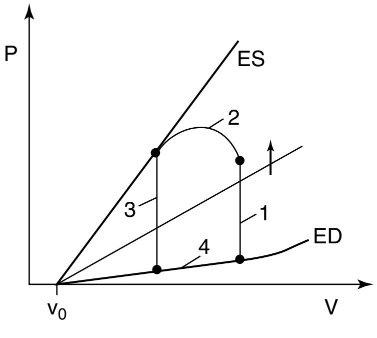

Fig. 21.12 Schematic representation of the ventricular pressure–volume (P–V) relationship. V_0, correction volume; ED, end-diastolic relation curve; and ES, end-systolic relation line. Lines 1–4 correspond to the four phases of the cardiac cycle.

only important ones. These orthogonal strains are interrelated, since the arterial wall is incompressible.

21.4.2.1 **Quasistatic Mechanical Properties of Arteries**

For the computation of elastic moduli of arteries, the differentially defined elastic modulus (Eq. (7)) must be used. For complete characterization of the arterial mechanics, the elastic moduli in the longitudinal, circumferential, and radial directions must be determined. Arteries *in situ*, however, are normally tethered in a prestretched state, so that the length of an arterial segment is independent of the pressure within the artery. The most important modulus for an analysis of pressure–volume relationships and the propagation of pressure and flow pulses within the arterial system is the circumferential elastic modulus, which is defined as the quotient of circumferential wall stress and circumferential strain.

The midwall circumferential wall stress (σ_c) of a cylindrical arterial segment is given by

$$\sigma_c = \frac{PR_i}{h} \tag{30}$$

where P is the pressure within the vessel minus the ambient pressure, known as the *transmural pressure*, R_i is the internal radius, and h is the wall thickness.

The circumferential strain ε_c is the ratio of the change in circumference of the segment ($=2dR_i\pi$) to the circumference of the segment ($=2R_i\pi$):

$$\varepsilon_c = \frac{dR_i}{R_i} \tag{31}$$

Simplifying $d\sigma_c = R_i dP/h$, the circumferential elastic modulus can be written as

$$E_c = \frac{dP}{dR_i}\frac{R_i^2}{h} \tag{32}$$

From this equation it can be seen that the wall thickness–radius relationship is one factor in determining the mechanical properties of blood vessels. Throughout the arterial system, the quotient of wall thickness and radius is virtually constant from the aorta down to the iliac artery, ranging between 0.12 and 0.14. In the peripheral arteries, this quotient increases, reaching a value of about 1.0 in the terminal arteries.

In addition to this geometrical factor, the mechanical behavior of an artery is determined by the viscoelasticity of its components, and their relative content within the wall. The histological structure of arterial walls may be divided into three layers: (from inside to outside) the intima, the media, and the adventitia. With respect to mechanical behavior, the media and the adventitia are of special interest since these layers contain the three components that above all determine the wall mechanics: the passive tissues elastin and collagen in the media and adventitia, and the vascular smooth muscle as the active component in the media. All three types of tissue show nonlinear viscoelastic behavior; the magnitudes of elasticity and viscosity, however, are very different. For low stresses, the elastic modulus of elastin is 0.3–0.6 MPa and that of collagen is 100 MPa, while the modulus of the vascular smooth muscle ranges between 10 and several hundred kPa. All three types of tissue are viscous; under pulsatile conditions, however, the viscosity of elastin and collagen exceeds that of the vascular smooth muscle.

There is a good correlation between the quasistatic elastic properties of arterial walls and the relative content of elastin and collagen within the wall. In the human ascending and thoracic aorta, the relation of collagen to elastin ranges from 0.5 to 0.7; in the middle abdominal aorta, this quotient increases abruptly in favor of collagen to about 1.5 and shows a further increase toward the periphery. Except for the coronary arteries, which have a collagen–elastin quotient of about 3 due to their small content of elastin, this quotient ranges from 1.5 to 2.5 for all larger arteries. According to this distribution, the ascending and thoracic aorta are very soft, while the quasistatic elastic modulus of the peripheral arteries increases with increasing distance from the heart.

On the basis of the distribution and the relative content of collagen, elastin, and vascular smooth muscle within the wall, it is not possible to give a reliable description of both the dependence of the elastic modulus on wall stress and the mechanical behavior of an artery under dynamic conditions, since these relationships are determined mainly by the functional arrangement of the various components within the vessel wall. In Figure 21.13, the pressure–radius relationship of a cylindrical arterial segment is shown. The volume changes applied to the artery were extremely slow (quasistatic conditions). According to Eq. (7), the elastic modulus of an artery corresponds to the slope of the stress–strain curve. The slope $d\sigma/d\varepsilon$ – which is equal to the slope of the pressure–radius curve – increases with increasing pressure; the nonlinearity of the viscoelastic properties of the artery can clearly be seen. The distensibility of the artery at very low stress is due to the high distensibility of the initially relaxed and then gradually stretched elastin tissue fibers, a behavior that is referred to as *recruitment*. In the physiological state of distension, the fibers of the collagen tissue are, because of their greater initial lengths, still unstretched and therefore exert no influence on the elastic modulus of the artery. Only at very high stress is the artery distended to such an extent that the very stiff collagen fibers act to protect the vessel against overstretching at high blood pressure.

The form of the curve during elongation and shortening results in the formation of a *hysteresis loop*. This hysteresis is more pronounced in the state of strong contraction induced by norepinephrine than in relaxation of the

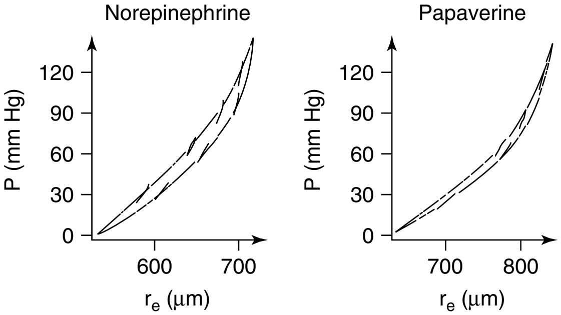

Fig. 21.13 Quasistatic pressure – external radius (P–r_e) curves of rat abdominal aorta under norepinephrine and papaverine. At different pressure levels, small sinusoidal pressure changes of 1 Hz were superimposed.

vascular smooth muscle induced by papaverine. In principle, the hysteresis can be explained by the combined effect of elasticity and viscosity of the vessel wall on the basis of the Voigt model (Section 21.2.2). To take into account the nonlinear behavior of the arterial wall, a specific strength of damping at each stress level and a spring with appropriately curved characteristics have to be taken into consideration.

21.4.2.2 Dynamic Mechanical Properties of Arteries

Since arteries *in vivo* are permanently subjected to periodic extensions of variable duration, an analysis of changes in stress and strain over a wide frequency range is of great importance in hemodynamics. Under dynamic conditions, the elastic modulus of the arterial wall increases. This is shown by Figure 21.13, in which small sinusoidal stress changes of 1 Hz were superimposed at different stress levels on the ascending and descending limbs of the quasistatic loops. The quotient of the dynamic and quasistatic elastic modulus depends on vascular smooth muscle tone and the content of muscle within the vessel wall. For larger arteries and relaxed smooth muscle, this quotient ranges from 1.3 to 2.0, while for muscular arteries and strong contraction of smooth musculature it can exceed a figure of 10. The transition between quasistatic and dynamic conditions takes place already at frequencies below 0.2 Hz.

To date, the mathematical description of the rheological behavior of blood vessels has been possible only under simplifying assumptions. For example, linear behavior of the arterial wall is usually assumed for small amplitudes of stress and strain. In this case, the equations given in Section 21.2.2.2 can be applied to the blood vessel wall. For sinusoidal oscillations of stress and strain, the elastic modulus can be defined as a complex modulus E^* (Section 21.2.2.2), the real part of which is represented by the dynamic modulus E_d and the imaginary part by the loss modulus $\omega\eta_w$:

$$E^* = E_d + j\omega\eta_w = E\cos\phi + jE\sin\phi \tag{33}$$

where ϕ is the phase angle between sinusoidal stress and strain changes. In accordance with Eq. (6), the loss modulus is the product of angular frequency ω and wall viscosity η_w.

An example of the interrelation between E_d, circumferential wall stress, and frequency of the stress changes under strong smooth-muscle contraction (induced by norepinephrine) and during smooth-muscle relaxation (induced by papaverine) is shown in Figure 21.14. From these diagrams, it can be seen that E_d increases markedly with increasing wall stress, while for a given wall stress, E_d is virtually independent of frequency. A comparison of the E_d values obtained for relaxed and strongly contracted smooth muscle shows that, for a given wall stress, E_d is virtually independent of the smooth-muscle tone.

The coefficient of wall viscosity increases with increasing wall stress – as does the dynamic modulus. Within a given wall stress range, the wall viscosity is not influenced significantly by the vascular smooth-muscle tone either. In contrast to the dynamic modulus, the wall viscosity decreases markedly with increasing frequency; the behavior of the wall viscosity is called – in the area of polymer rheology – *thixotropy* or *pseudoplasticity*.

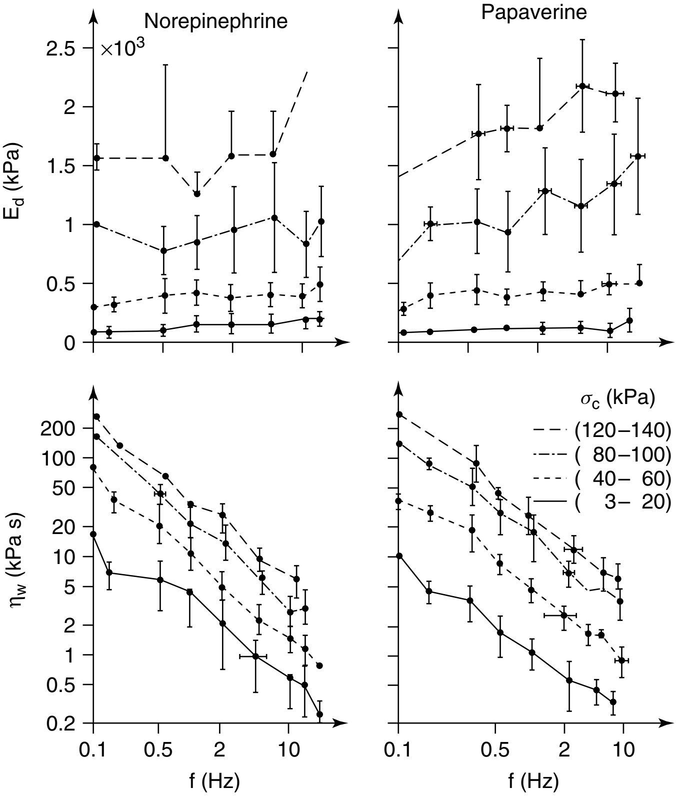

Fig. 21.14 Dependence of dynamic elastic modulus (E_d) and wall viscosity (η_w) of the rat abdominal aorta on frequency (f) for different ranges of circumferential wall stress (σ_c) under norepinephrine and papaverine. Symbols and bars represent means ± standard errors of the means (SEM).

The rheological behavior, as shown in Figure 21.14, is, in principle, common to all arteries. The values of E_d and $\omega\eta_w$, however, differ for large, medium-sized, and small arteries. The lowest values of both magnitudes are seen in the major arteries. The dynamic modulus of the thoracic aorta of a healthy young man, for example, is about 0.1 MPa at very low stress, and 0.4–0.5 MPa at a stress corresponding to the normal blood pressure range. The loss modulus of this artery is about 10–15% of the dynamic modulus. Medium-sized and small arteries show both greater dynamic and greater loss moduli. In the muscular arteries, however, the loss moduli increase more than the dynamic moduli. In general, vessels having a larger smooth-muscle component exhibit higher values of wall viscosity.

21.4.3
Pressure and Flow in the Vascular System

Pressure and flow in arteries fluctuate, since ejection of blood from the ventricles into the aorta and the pulmonary artery is intermittent. In addition to these fluctuations, the description of the relationship between pressure and flow within the arterial system is greatly complicated – in contrast to technical flow in tubes – by the rheological properties of arteries. First, as already described, the vessel walls are extensible and exhibit nonlinear viscoelastic behavior. Second, the stiffness of the arteries increases from the aorta toward the smallest arteries, which is due to the increase both in wall thickness–radius relationship – commonly referred to as *tapering* – and in elastic modulus. Third, the elasticity of the arterial wall depends on age. All these properties play an important role in hemodynamics.

21.4.3.1 **Systemic Arteries**

The events in the arterial system resulting from the intermittent ejection of blood by the ventricles can be considered with respect either to the mean values of pressure, flow, and volume (or diameter) or to the pulsations of these magnitudes. In the healthy adult, the mean aortic pressure (MAP) is about 100 mm Hg and the mean flow in the ascending aorta – which is the cardiac output (CO) minus the flow rate entering the coronary arteries (5% of CO) – is about $5\,\mathrm{l\,min^{-1}}$. Besides fluid viscosity, within every subsequent part of the circulation the resistance to flow is determined by both the dimensions of each single vessel and the number of vessels arranged in parallel. Owing to the relatively large diameters, and thus small resistances (Eq. (18)), of the large, medium-sized, and small arteries, the mean pressure decreases in this part of the systemic circulation only from 100 to about 90 mm Hg (Figure 21.15). The greatest resistance to flow is experienced in the smallest arteries, the arterioles, and the capillaries; these vessels are therefore often referred to as *resistance vessels*. Although the diameters of the arterioles are markedly greater, the pressure drop over these vessels is about twice that over the capillaries. This is due to the smaller number of arterioles arranged in parallel, and their greater lengths as compared with the capillaries. Within the systemic circulation of the dog, arterioles are estimated to number about 4×10^7 and capillaries about 1.2×10^9.

In total, the mean pressure of blood falls to less than 10 mm Hg by the time it returns to the right atrium of the heart.

In accordance with Eq. (14), the total frictional resistance within the systemic circulation can be calculated by the difference between the MAP and the mean right atrial pressure (MRAP) divided by the CO. Since the resistance to flow is exerted mainly by the peripheral vessels, the total circulatory resistance is called *total peripheral resistance* (TPR). Hence,

$$\mathrm{TPR} = (\mathrm{MAP} - \mathrm{MRAP})/\mathrm{CO} \tag{34}$$

In the human systemic circulation, normal TPR is about $1.5 \times 10^2\ \mathrm{MPa\,s\,m^{-3}}$.

According to Eq. (13), the mean velocity v_m of the fluid flowing along a vessel – or a system of vessels arranged in parallel – is inversely proportional to the (total) cross-sectional area(s) of the vessel(s):

$$v_m \simeq 1/A \tag{35}$$

The larger the A, the smaller the v_m. In the human ascending aorta, the mean velocity of blood is about

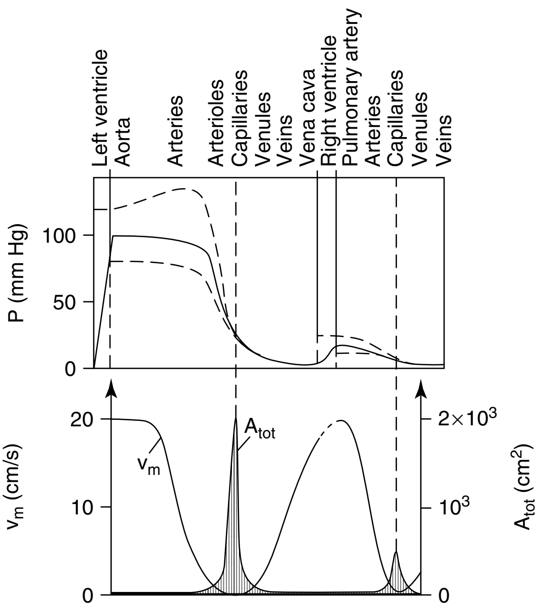

Fig. 21.15 Schematic representation of mean pressure and pressure variations (P), mean velocity (v_m), and total cross-sectional area (A_{tot}) in various parts of the human circulatory system.

0.25 m s^{-1}. Each ramification of the aorta and its branches toward the periphery results in an increase in the total cross-sectional area of the arteries arranged in parallel, since the sum of the cross-sectional area of the branches is always greater than that of the stem. As shown in Figure 21.15, the mean velocity decreases from the ascending aorta to the peripheral vessels. The largest cross-sectional area is found in the capillaries, and in a resting man is about 500 times that of the aorta. Thus, the mean velocity of blood decreases in the systemic capillaries to about 0.5 mm s^{-1}.

21.4.3.2 **Pulmonary System**

The same course of blood pressure and blood velocity is found in the pulmonary circulation, although the pressure levels and resistances are considerably lower than in the systemic circulation (Figure 21.15). The mean pressure in the pulmonary artery is about 15 mm Hg and in the pulmonary veins is about 7 mm Hg. Since the output of the right ventricle equals that of the left ventricle, the TPR of the pulmonary circulation is only about 10% of that of the systemic circulation. The total cross-sectional area of the perfused pulmonary capillaries is 100–150 times as large as that of the pulmonary artery, so that in the pulmonary capillaries the mean velocity of blood is greater than in the systemic capillaries, amounting to about 2 mm s^{-1}.

21.4.3.3 **Arterial Dynamics**

Because of the intermittent ejection of blood into the aorta and pulmonary artery, pulse waves are generated which are propagated in the systemic and pulmonary systems, respectively. The pulse waves are formed by the combined effects of elasticity of the vessel wall and mass inertia of the accelerated blood. Because of inertia, the blood ejected, for example, by the left ventricle resists acceleration. This leads to a local increase in pressure, and thus to elastic distension of the first segment of the aortic wall, so that part of the ejected blood is stored while the rest flows through the aorta. When the pressure decreases, the distended elastic segment contracts and the stored blood is now accelerated and driven into the second aortic segment in which part of the blood is again stored while the rest is in motion. In the following segments the processes of storage and acceleration are repeated. These events take place simultaneously and are propagated continuously along the systemic, and also the pulmonary, arteries as pulse waves having a certain velocity known as *pulse wave velocity*.

The total energy transported by a pulse wave is made up of the potential energy of the distended segment and the kinetic energy of the moving blood. Within a given segment, these two kinds of energy are equal at any given instant, and are perpetually being transformed, one into the other.

For a complete analysis of the relationships between the pulsations of pressure, flow, and vessel diameter, two sets of equations are needed: the equations of motion and continuity of the blood and the equations of motion and continuity of the vessel wall. By means of these equations, the distribution of the velocity of blood within arteries, as well as the propagation of pressure and flow pulses within the arterial system, can be calculated.

Velocity Profiles in Arteries With regard to mean flow in the circulation, critical values of the Reynolds number (Eq. (19)) are exceeded in the ascending and descending aorta and in the initial part of the pulmonary artery. Because of these high Reynolds numbers – and the pulsatile nature of the flow – true turbulence is most likely to occur in these arteries during each pulse cycle – also under normal conditions. In peripheral arteries and in veins, Reynolds numbers are much smaller than the critical value, so that the flow in these vessels is not turbulent.

The pulsatility of flow in arteries – and also in veins – complicates the description of the behavior of flow. In Section 21.2.4.3, a solution to the equations of motion of fluid (Navier–Stokes equations) is given. From this solution, the velocity profiles of oscillating fluid and their dependency on the dimensionless value α_w can be seen (Figure 21.5). In man, the α_w values for the first harmonic of the natural pulses range from 17 in the ascending aorta to 1 in the peripheral leg arteries. Therefore, the velocity profile is not parabolic but more uniform near the vessel axis and also in medium-sized and small arteries in which the flow is not turbulent. Velocity profiles determined in human arteries *in vivo* by means of pulsed ultrasound Doppler devices (e.g., Anliker *et al.*, 1977) confirm these theoretical considerations.

Arterial Pulse Waves For a description of the propagation of waves in the arterial system, both motion of fluid and elasticity of the vessel walls must be taken into account. Since, as mentioned above, the effects of friction do not play too prominent a role in major and medium-sized arteries, the laws of wave propagation can be represented as follows.

In the absence of resistance to flow ($\eta = 0$), Eq. (22) reduces to

$$-\frac{\partial p}{\partial z} = M' \frac{\partial i}{\partial t} \tag{36}$$

when Eq. (13) is employed; $M'(= \rho/A)$ is the effective mass of the fluid per unit of length. Pulsatile changes of the cross-sectional area A of the vessel are not considered.

The flow rate entering a segment of the elastic vessel is the sum of the fluid which – due to the pressure elevation – is stored in this segment per unit of time, and the flow rate leaving this segment. This relationship is expressed by the equation of continuity:

$$-\frac{\partial i}{\partial z} = C' \frac{\partial P}{\partial t} \tag{37}$$

where C' (Eq. (8)) is the compliance per unit of length of the vessel. Differentiating and combining Eqs (36) and (37) result in the wave equations:

$$\frac{\partial^2 P}{\partial t^2} = c^2 \frac{\partial^2 P}{\partial z^2} \quad \text{and} \quad \frac{\partial^2 i}{\partial t^2} = c^2 \frac{\partial^2 i}{\partial z^2} \tag{38}$$

where c is the pulse wave velocity, which is given as $c = (M'C')^{-1/2}$. According to d'Alembert, the solutions of Eq. (38) are

$$P(z, t) = f_1(z - ct) + f_2(z + ct) \tag{39}$$

and

$$i(z, t) = Z^{-1}[f_1(z - ct) - f_2(z + ct)] \tag{40}$$

Z is the characteristic impedance of the arterial wall, which is defined as the quotient of pressure and flow of a wave propagated in one direction:

$$Z = P/i \tag{41}$$

The functions f_1 and f_2 represent pressure pulses of any arbitrary shape, but without discontinuities. $f_1(z - ct)$ is a pressure wave propagated antegrade, that is, in the positive direction of the z axis, and $f_2(z + ct)$ is a pressure wave propagated retrograde, that is, in the negative direction of the z axis. The total pressure is, therefore, the sum of all antegrade and retrograde waves running through the vessel, while the total flow rate is the difference between all antegrade and retrograde waves. In principle, this explains the discrepancies between the pulse contours of pressure and flow pulses in the arterial system.

Generally, arteries can be regarded as hydraulic wave transmission lines. The characteristic impedance of an arterial segment is determined by the inertia and viscosity of the blood, and the viscoelastic properties of the arterial wall. If fluid and wall frictions are neglected, the characteristic impedance of a homogeneous arterial segment can be calculated by means of the equation:

$$Z = \rho c/A \tag{42}$$

where ρ is the density of blood, c the pulse wave velocity, and A the internal cross-section of the artery. In young men, the average pulse wave velocity along the aorta is about 5 m s^{-1}; in the limb arteries c increases from 6 to 8 m s^{-1}, and in the distal leg arteries to even more than 10 m s^{-1}. Because of the increase in pulse wave velocity and the decrease in the cross-sectional area, a strong increase in the characteristic impedances from the aortic root toward the smallest arteries results. This increase is, in part, stepwise, for example, at every branching point of an artery, in part continuous, which is commonly referred to as *tapering*. Reflection occurs at a site at which the characteristic impedance changes. The arterial system therefore represents an inhomogeneous wave transmission line exhibiting multiple internal reflection sites. Thus, the primary pulse wave generated at the aortic root by the contraction of the left ventricle undergoes multiple reflections within the arterial system. At each site and instant, ante- and retrograde running pressure and flow waves are superimposed one upon the other. The different behavior of pressure and flow waves with respect to reflections and the different signs of retrograde pressure and flow waves are responsible for the different contours of pressure and flow pulses within the arterial system.

Models of the Arterial System A major topic in the field of hemodynamics is the development of models that can be used to explain the genesis of the pulse contours of pressure and flow in the various regions of the arterial tree. In the 1960s and early 1970s, a variety of models, represented by real hydraulic, electric analog, or theoretical systems, were developed. Recently, the demand for models of the arterial system has been renewed because of the widespread use of Doppler ultrasound devices that allow noninvasive, transcutaneous recording of blood flow velocity. Today, Doppler ultrasound recordings are routinely performed in a great number of blood vessels, in particular, in those of the head and brain, the method being termed transcranial Doppler ultrasound. Efforts are being made to introduce models as diagnostic tools; these models are required to simulate, and thus determine quantitatively, the influence of disturbances – such as aneurysms, stenoses, or occlusions – on the pulse contours. For this purpose, computer models are particularly convenient, since such models can readily be adapted to any subject. In view of the extreme complexity of the arterial system, any model must necessarily be simplified to a greater or lesser degree. As an example, the simulation of arterial pulses by means of an inhomogeneous tube model (Wetterer, Bauer, and Busse, 1977) is given. Natural pressure and flow pulses recorded at different sites in the arterial system, and pulses calculated at corresponding sites in the theoretical model, are shown in Figure 21.16. The pressure pulses have been recorded as sphygmograms, the flow pulses by means of a Doppler

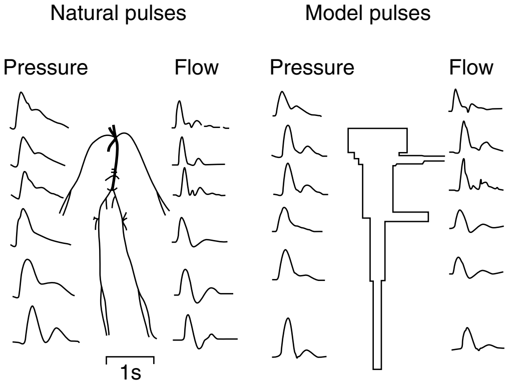

Fig. 21.16 Left: pressure and flow pulses recorded at various sites of the human arterial system. The recording sites are, from top to bottom, carotid artery, brachial artery, radial artery, abdominal aorta, femoral artery, and posterior tibial artery. Right: pulses of a theoretical model as shown schematically in the middle. Pulses are calculated for those sites of the model, which correspond to the recording sites of the natural pulses.

ultrasound device. The model consists of an arterial main stem representing the aorta and its continuation down to the foot arteries and two side branches, one representing the arteries supplying the head and the arms and the other the arteries branching from the abdominal aorta. As seen in Figure 21.16, there is good agreement between the pressure and flow pulses calculated by means of the model and the natural pulses, so that this relatively simple model is, in principle, suitable to explain the genesis of pulse contours in the arterial system.

21.4.3.4 **Pressure and Flow in Veins**

From the point of view of hemodynamics, the venous system differs essentially from the arterial system. When the body is in horizontal position, the venous pressure is generally low, decreasing from about 15 mm Hg in the venules to about 5 mm Hg in the right atrium. During inspiration, the pressures in the intrathoracic veins and the right atrium may even reach subatmospheric values. In the upright position, the pressure in the veins of the foot may, for hydrostatic reasons, be more than 90 mm Hg; while walking this pressure decreases markedly because the deep veins lying within the muscles of the legs are rhythmically compressed, and the valves in the peripheral veins ensure blood flow toward the heart (muscle pump). Furthermore, the walls of veins are thinner and their distensibility is much greater than those of arteries.

In principle, the same laws are applicable to the description of propagation of pulse waves within the venous and the arterial system. Because of the high distensibility and the small wall thickness–radius relationship, however, in the horizontal position the pulse wave velocity in veins is much smaller than in arteries (Section 21.4.3.3, Arterial Pulse Waves). In the vena cava, for example, the pulse wave velocity is about 1 m s^{-1} at a transmural pressure of about 4 mm Hg. In contrast to the arterial system, the pulsations of pressure and flow in the central veins are due to the action of the right heart, while pulsations within the peripheral veins are caused by respiration.

A description of the pressure–flow relationship in veins is complicated by the fact that veins may collapse since their soft thin walls are unable to withstand even small negative transmural pressures. *In vivo*, the pressure in large and medium-sized veins oscillates about a value close to atmospheric pressure, so that conditions in which the pressure can drop below zero may well prevail. As shown in Figure 21.17, collapsible veins lose their circular cross-section when the transmural pressure approaches zero.

The pressure–flow relationship of collapsible vessels has often been investigated by means of a laboratory device, first introduced by Holt (1941). In this device, a soft rubber tube is stretched between two rigid pipes within an airtight chamber that applies an adjustable external pressure to the tube. The pressure (P_1) in front of the rubber tube can be varied by adjusting the level of fluid in a supply tank, while the pressure at the end of the tube (P_2) can be changed by means of an outlet resistance. Relationships between pressure difference $\Delta P (= P_1 - P_2)$ and the flow i through the tube are shown in Figure 21.18 for three different constant external pressure levels. At low flow, ΔP increases with increasing flow. In this range, the rubber tube is collapsed throughout its whole length. As soon as P_1 exceeds the external pressure, the first part of

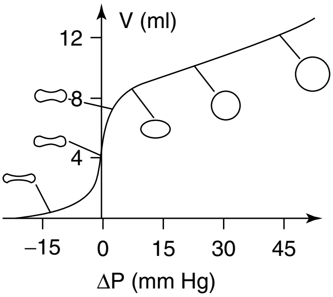

Fig. 21.17 Typical cross-sections of a collapsible tube shown at various points of the volume–transmural pressure ($V-\Delta P$) relationship.

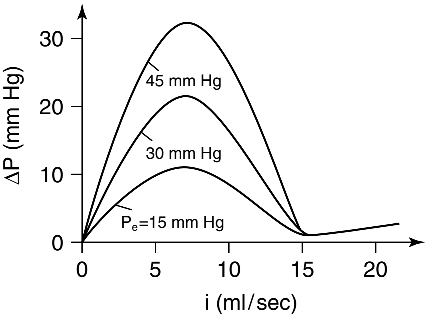

Fig. 21.18 Three transmural pressure–flow ($\Delta P-i$) relationships for a collapsible tube. External pressure (P_e) is a parameter (Modified from Katz, Chen, and Moreno, 1969).

the tube is distended and ΔP decreases, since P_2 becomes greater. A further increase in flow results in an expansion of the next part of the tube and thus a further decrease in ΔP. Finally, the tube is completely distended over its whole length so that, in accordance with Eq. (14), an increase in flow leads to an increase in the pressure difference over the rubber tube. As shown in Figure 21.18, each external pressure exhibits its own $\Delta P - i$ curve. It should be noted that instabilities resulting in self-excited oscillations of pressure and flow can occur in the region of negative slope of the $\Delta P - i$ curves.

Generally, the rheological behavior of veins can be described by means of hydraulic models. However, the applicability of such *in vitro* investigations to the venous system is limited, since *in vivo* the external pressure – and in the central veins also the blood pressure – alternates periodically. Furthermore, the venous wall contains smooth musculature and can therefore change its diameter actively over a wide range.

21.5
Pneumodynamics

Cells generally require a permanent oxygen supply and carbon dioxide removal to maintain their metabolic processes. In mammals almost all oxygen is collected, and carbon dioxide given off, by the blood passing through the capillaries of ventilated lungs. The transport of respiratory gases can be divided into four main components:

1. convective transport of gases between the atmosphere and the lung alveoli;

2. diffusive transport of O_2 and CO_2 between the alveoli and the pulmonary capillary blood;
3. convective transport of O_2 and CO_2 in the circulation to and from cells, respectively; and
4. diffusion of CO_2 and O_2 between the systemic capillary blood and the cells in the tissue.

For a discussion of pneumodynamics, the mechanical properties of the chest and lungs and the gas flow within the airways are of major importance.

21.5.1
Functional Anatomy of the Lungs

21.5.1.1 Airways

The right and left lungs are divided into three and two lobes, respectively. The air inspired at the nose and/or the mouth passes the pharynx, glottis, and larynx to the trachea. The trachea bifurcates into left and right main bronchi supplying the left and right lungs. The main bronchi in turn bifurcate into lobar, then segmental bronchi; these in turn divide successively to form the terminal bronchioles, which are the smallest air tubes without alveoli. The respiratory element distal to a terminal bronchiole is called an *acinus*. The terminal bronchioles branch to form respiratory bronchioles, which already have alveoli in their walls. The respiratory bronchioles further divide into the alveolar ducts which lead to the atria. Finally the air sacs open into the atria; the walls of the air sacs are made up of alveoli, while the walls of all the airways smaller than the terminal bronchioles contain only alveoli. All told, the adult human lungs contain 300–600 million alveoli. The terminal bronchioles have a layer of smooth musculature, the contraction state of which modifies the amount of gas going into and out of the alveoli.

According to Weibel (1963), the tracheobronchial tree is numbered in successive generations of branching airways from the trachea (generation 0) down to alveolar sacs (generation 23). The first 16 generations down to the respiratory bronchioles constitute the conduction zone, the function of which is to conduct the respiratory gases to and from the respiratory zone of the lungs. Since this part of the respiratory tract cannot take part in gas exchange it forms what is known as the *anatomical dead space*, the volume of which is about 150 mL. The last seven generations from the respiratory bronchioles to the alveoli make up the respiratory zone of the respiratory tract.

The total cross-section of the airways increases progressively as the bronchi divide. The cross-sectional area of the trachea is about 2.5 cm^2, while the total alveolar surface is between 50 and 100 m^2. The distance the respiratory gases have to be transported is in inverse proportion to the total cross-sectional area of the airways. The distance from the beginning of the trachea to the end of the terminal bronchioles is about 25 cm, while that from the beginning of the respiratory bronchioles to the most distal alveoli is only about 5 mm.

21.5.1.2 Lung Volumes

The lung volumes can be subdivided as follows. The tidal volume (TV) is the air that enters and leaves the lungs in a respiratory cycle. On normal breathing it is about 500 mL. Only about 350 mL, however, reaches the respiratory zone, the rest remaining in the dead space.

The maximum amount of air that can be inspired from the end-inspiratory position is termed inspiratory reserve volume (IRV); the maximum volume that can be expired from the end-expiratory level is termed expiratory reserve volume (ERV). The residual volume (RV) is the amount of gas remaining in the lungs after a maximal expiration.

A number of lung capacities are also defined: The vital capacity (VC) is the maximum volume of gas that can be arbitrarily ventilated, and is given by VC = TV + IRV + ERV. Only these volumes can be measured using a spirometer. In the normal man, the VC is about 5 L.

The functional residual capacity (FRC) is the amount of gas remaining in the lungs at the resting expiratory level. For the most part, the FRC acts as a buffer against extreme changes in alveolar partial pressures of O_2 and CO_2 on each breath. Finally, the total lung capacity (TLC) is the volume of gas that the lungs contain at the end of a maximal inspiration.

21.5.1.3 **Ventilation**

The normal rate of respiration of adults at rest ranges from 10 to 18, with a mean of 15, breaths per minute. This basal frequency is influenced by many conditions, including age, sex, size or height, work and rest, sleep, change in position, and temperature. At 15 breaths per minute and a TV of 0.5 L, the volume voided by the lungs is 7.5 L each minute; this flow rate is termed the minute volume, or tidal ventilation. Since only 350 mL of the TV reach the alveoli, the alveolar ventilation, which defines the fresh gas entering the respiratory zone per minute, is $5.25\ \mathrm{l\ min^{-1}}$.

The expansion of the lungs is due to lengthening and widening of the fine branches of the airways in the respiratory zone. The most distensible parts are the alveolar ducts, while the alveoli do not change much in volume. Thus, not all regions of the normal lungs dilate uniformly during inspiration. The center of the lungs containing the bronchi and some parenchymatous tissue is only slightly distensible. Distensibility increases in the direction of the surface of the lungs reaching a maximum 25 to 30 mm below this surface. Furthermore, the different parts of the lungs do not expand uniformly during inspiration, so that they do not have the same ventilation. In the upright position, the lower lobes expand more and thus ventilate better than the upper ones. This behavior is mainly due to gravitational effects, and will be discussed later.

During normal quiet breathing, TV only suffices to fill the respiratory tract down as far as the terminal bronchioles, while only a very small portion of the inspired air passes into the respiratory zone. The distance from the terminal bronchioles to the alveoli is covered by diffusion; the velocities of the molecules in the respiratory gas are so great, and the distances are so short, that the gas covers these distances in only a fraction of a second.

21.5.2
Mechanics of Breathing

The lungs – and also other thoracic viscera, for example, the heart – are subjected to passive changes caused by the expansion and retraction of the thoracic wall. For inspiration to occur, alveolar pressure must be less than atmospheric pressure; for expiration to occur, alveolar pressure must be greater than atmospheric pressure. Inspiratory flow of air can be produced by lowering the alveolar pressure to below atmospheric pressure – this is termed natural or *negative* pressure breathing – or by raising the pressure in the external environment above the alveolar pressure – termed *positive* pressure breathing. Normal inspiration is accomplished by the active contraction of the inspiratory muscles, which enlarges the thorax thus decreasing alveolar pressure to below atmospheric pressure. Active muscle contraction is needed to provide the force necessary to overcome the elastic recoil of the lungs and chest, the viscosity of both the tissues of the lungs and chest, and the air flowing through the conducting airways.

The thorax consists of the thoracic vertebrae, the ribs together with their cartilages, and the sternum. In man, the ribs increase in length from the 1st to the 7th, which is the longest, and then decrease to the 12th, which is the shortest. Each rib is attached to the spinal column at two points: the head to the body and the tubercle to the transverse process of the vertebrae. Because of this anatomical arrangement, the ribs can rotate only around an axis joining these two points. During normal breathing the change in the shape of the thoracic cavity is caused by contraction and relaxation of the diaphragm – resulting in a lengthening and shortening of the chest cavity – and the muscles of inspiration, which elevate, and the muscles of expiration, which contract the chest cage.

21.5.2.1 Mechanics of Inspiration

The diaphragm is the principal inspiratory muscle, most fibers of which form a dome-shaped structure descending to insertions on the lower ribs and the sternum. When the diaphragm contracts, the thoracic cavity enlarges mainly downward, but the contraction of its costal fibers may also elevate the lower ribs, thus increasing the circumference of the thorax. On normal quiet breathing, the diaphragm moves downward about 1 cm; when inspiration is maximal, however, its downward movement may be as much as 10 cm. A paralyzed diaphragm behaves as a passive membrane, and thus ascends during inspiration. This is known as *paradoxical* respiration.

The respiratory muscles that raise the ribs can be divided into principal and accessory muscles. The external intercostal muscles are principal muscles and extend from the lower edge of one rib to the upper end of the rib below. When the first five ribs are elevated, the anteroposterior diameters of the chest are increased; the sixth to the tenth ribs, however, effect a considerable increase in transverse diameters as well. In forced inspiration, the upper ribs and the sternum are elevated by the contraction of the accessory muscles, in the main sternocleidomastoid and scaleni. During quiet breathing, these muscles are virtually nonactive.

21.5.2.2 Mechanics of Expiration

During quiet breathing, expiration is largely passive by virtue of the elastic forces of the lungs and thorax. It is only during rigorous breathing that the expiratory muscles are called into action. The main muscles of expiration are those of the abdominal wall: the transversus, the internal and external oblique, and the rectus muscles. When these contract, the pressure in the abdominal wall is raised. The pressure forces the diaphragm up into the thorax and – if the glottis is kept open – gas is expelled from the lungs. In addition, ribs are depressed by the contraction of the internal intercostal muscles and the triangular muscle of the sternum, thus decreasing the volume of the thoracic cavity.

21.5.3
Rheology of the Respiratory Tract

The lungs and the thoracic cage form a functional unit that can do its work only in the intact state. For an understanding of the system, however, an analysis of the passive properties of the two constituent components of the lung–chest system is desirable.

21.5.3.1 Rheology of the Lung

The lungs can be considered as viscoelastic structures having an elastic tendency to contract and therefore pull away from the chest wall. This elastic recoil is caused by two different factors. First, in the lung parenchyma there are elastin fibers that are stretched by lung inflation and thus have a tendency to shorten. Second, much of the elastic recoil is due to surface tension of the fluid lining the alveoli. It has been calculated that the elastic fibers in the lungs account for one-third of the recoil, and the sum of the recoil forces due to surface tension accounts for two-thirds.

The surface tension at a gas–water interface produces forces that tend to reduce the area of the interface. The pressure P within a bubble depends on the surface tension T and the radius R of the curvature of the bubble, and is given by

$$P = \frac{2T}{R} \tag{43}$$

From this equation it can be seen that for a constant T the pressure increase when the radius of the curvature decreases. When two or more bubbles are connected in parallel, the bubbles having the smaller radii empty into the larger, so that instabilities may occur.

The problem of maintaining bubbles in parallel is critical in the lungs where millions of alveoli exist side by side. In alveoli, however, the surface tension is not a constant, but decreases when the radius of an alveolus becomes smaller. This is due to the presence of a surface active material, known as *surfactant*, which is secreted by special cells of the alveolar epithelium. The surface tension is lowered in proportion to the concentration of the surfactant molecules at the interface. During expiration, when the radii, and thus the internal areas, of the alveoli diminish, the surfactant molecules are packed more densely, thus decreasing surface tension and promoting the stability of the alveoli. On the whole, the surfactant lowers surface tension in the alveoli, so that the compliance of the lungs is increased and the work of breathing is reduced. Furthermore, the surfactant helps to keep the alveoli dry; surface tension forces tend not only to collapse the alveoli but also to suck fluid into the alveolar space from the pulmonary capillaries, so that surfactant prevents the transudation of fluid by reducing these forces.

Another mechanism that contributes to the stability of the alveoli is termed *interdependence*. Each alveolus represents a unit in an interconnected mesh.

Therefore, small units that might have a tendency to collapse are supported and stabilized by surrounding parenchyma.

The outer surface of the lungs is covered by visceral pleura. This is in intimate contact with the inner surface of the thoracic cage covered with parietal pleura. The pleural space between these two pleurae is filled with a thin film of fluid. Movement of the thorax leads to movement of the lungs through the linkage of this film of fluid. The pressure in the pleural space which is required to keep the lungs expanded to normal size is subatmospheric; it is usually called *negative intrapleural pressure*. This pressure changes during breathing, ranging from 5 to 8 mm Hg during expiration and from 10 to 15 mm Hg during inspiration. Even in maximum expiration, however, the lungs are still subject to elastic stress.

The pressure–volume curves of the lungs during inflation and deflation differ. As shown in Figure 21.19, at any given pressure the lung volume is larger during deflation than during inflation. Because of this hysteresis, the pressure developed by the lung at a given volume strongly depends on the prior inflation history. The hysteresis is due to the viscoelastic properties of the lung tissues. The area of the loop depends on TV; it becomes progressively greater as TV is increased.

A widely used measure of the mechanical properties of the lungs and chest wall is compliance, that is, the slope of the $P-V$ curve (Eq. (8)). When the viscosity of the lung tissue is not taken into account, compliance is given as the average between compliance during inspiration and that during expiration. Over the range of normal breathing, the compliance of the isolated human lungs is about 0.2 L (cm water)$^{-1}$. At high expanding pressures, the lungs are stiffer, and their compliances are smaller, as can be seen from the decreasing slope of the $P-V$ curve in Figure 21.19.

Owing to the weight of the lungs, in an upright position the alveoli at the apex of the lungs have a larger volume than those at the base. The greater degree of expansion of the alveoli in the upper parts results in a more negative intrapleural

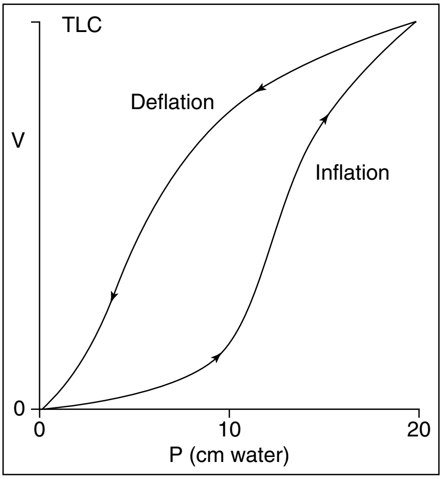

Fig. 21.19 Pressure–volume ($P-V$) curve of the excised lung. TLC Total lung capacity .

pressure which increases steadily down the lung by about 1 cm of water per 3 cm of vertical height. The dependent parts of the lungs are relatively compressed in their resting state, but expand better on inspiration than the upper parts. In consequence, the ventilation is greater in the base of the lungs than in the apex.

21.5.3.2 Rheology of the Chest

An isolated lung will always tend to contract to its minimal volume. In contrast, when the thoracic cage is opened, for example, by putting air into the pleural space – known as *pneumothorax* – the thorax tends to expand to a volume about 1.5 L greater than FRC. Since the thoracic cage contains muscular structures, passive mechanical properties can be considered only when the muscles are relaxed. The chest wall, like all other tissues, exhibits viscoelasticity, but for most purposes the viscous behavior of the thorax itself can be neglected. In analogy to pulmonary compliance, the compliance of the intact thoracic cage is defined as the change in lung volume per unit change in pressure gradient between atmosphere and intrapleural space. The value of static compliance of the chest is about 0.2 L (cm water)$^{-1}$, thus equaling that of the isolated lung.

21.5.3.3 Rheology of the Lung–Chest System

As already mentioned, when the intrapleural pressure is raised to atmospheric pressure, the lungs collapse and the chest wall expands. Thus, under normal conditions, the thoracic cage is pulled inward, while the lungs are pulled outward. Mechanically, the lung–chest system can be represented by two Voigt models or – neglecting the viscosities – by two springs arranged in parallel, possessing different resting lengths. In the intact state the lungs pull the chest wall and diaphragm inward, so that the combined resting length of the whole model is intermediate between the individual resting lengths.

The relationships between pressure and volume can be seen from the static $P-V$ curves of the lung–chest system shown in Figure 21.20. For the determination of this $P-V$ curve, the subject inspires a certain volume from a spirometer and then relaxes against the closed airway while the airway pressure is measured. In Figure 21.20, the $P-V$ curve of the isolated chest wall and that of the isolated lungs are also shown. For any given volume, the sum of the transpulmonary pressure – which is defined as the difference in pressure between the inside and the outside of the lung – and the passive chest wall recoil pressure is equal to the total pressure within the respiratory system (relaxation pressure).

At FRC, the relaxation pressure is atmospheric, so that the tendency of the chest wall to expand and the elastic recoil of the lung are exactly balanced. During inspiration the elastic recoil of the lung increases, while the chest wall recoil decreases; thus the relaxation pressure becomes positive. The chest wall reaches its equilibrium position at about 75% of VC. A further increase in lung volume requires positive pressure to increase the volume of the thoracic cage. At volumes below FRC, the relaxation pressure is subatmospheric. In accordance with Eq. (8), the compliance is inversely proportional to the pressure difference. Therefore, at a given volume change, the compliance of the lung–chest system (C_{tot}) can be calculated from the

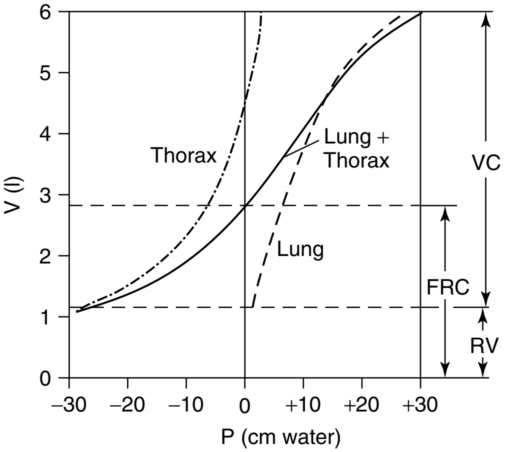

Fig. 21.20 Quasistatic pressure–volume (P–V) curves of the respiratory system (–), the lung (--), and the thorax (-··-).

compliance of the lung (C_1) and that of the thorax (C_t) by using the following expression:

$$\frac{1}{C_{\mathrm{tot}}} = \frac{1}{C_1} + \frac{1}{C_t} \tag{44}$$

Thus, the compliance of the lung and chest of a healthy young man is approximately 0.1 L (cm water)$^{-1}$.

21.5.4
Air Flow in the Respiratory Tract

The air passes in and out of the lungs as a result of rhythmic changes in pressure difference between the pressure at the nose or that at the mouth and the alveolar pressure. In accordance with Eq. (14), airway resistance W_a is defined as $W_a =$ [(nose or mouth pressure) – (alveolar pressure)]/air flow rate. Thus, air flow depends on the extent of pressure difference and airway resistance.

21.5.4.1 Intrapulmonary Pressure

During quiet breathing the alveolar pressure becomes slightly negative with respect to mouth pressure, normally about −1 mm Hg, and during expiration the alveolar pressure rises less than +1 mm Hg. During forced inspiration, alveolar pressure may drop to as low as −100 mm Hg, while during forced expiration (with glottis closed) the intrapulmonary pressure can be increased to as much as 150 mm Hg. The maximum expiratory effort is far greater than that during inspiration; expiration is used for voice production and elimination of foreign particles from the respiratory tract by coughing.

21.5.4.2 Distribution of Airway Resistances

About half of the total airway resistance during quiet breathing is in the nasal region, while there is less resistance to breathing through the mouth. A typical value for the total airway resistance measured during nose breathing would be 4 cm water s/L, while that during mouth breathing would be 2 cm water s/L. This total resistance can be partitioned into a series of resistances, each representing a section of the air-conducting system. On mouth breathing near FRC, the glottis, which is the narrowest part of the airway, accounts for more than 40% of the total airway resistance. From the remainder, the major site of resistance is the medium-sized bronchi, while the very small bronchioles contribute only about 20% to the total resistance. Even though each bronchiole has, according to its small radius (Eq. (18)), a high resistance to air flow, the number of these tubes in parallel is so enormous that their total resistance is lower than that of the larger air tubes.

21.5.4.3 **Pressure–Air Flow Relationship**

Air flow in the respiratory system is determined by the same factors that govern the flow of viscous fluid in tubes or blood vessels, so that the considerations of Section 21.2.4 can, in principle, also be applied to the respiratory tract. Because of the low viscosity of air and the relatively high velocities, turbulent air flow may occur in the trachea and larynx, but never in the bronchioles. In most of the bronchial tree, flow is not turbulent but transitional; if eddies are formed at branches, they are carried downstream some distance before disappearing. Fully developed laminar air flow may only occur in the smallest airways distal to the bronchioles where the Reynolds numbers are 1 or lower.

Taking resistance to include two components, one for laminar and the other for turbulent flow, the total pressure drop ΔP along the airway is, in accordance with Eqs (14) and (20), dependent on both the flow rate $\dot{V}$ and its square:

$$\Delta P = K_1 \dot{V} + K_2 \dot{V}^2 \tag{45}$$

The proportionality constants K_1 and K_2 were obtained by fitting this equation to the experimental data of each subject. For the normal respiratory tract in man, the values of these constants have been found to be $K_1 = 2.4$ and $K_2 = 0.3$ when ΔP is given in centimetre water and $\dot{V}$ in liters per second (Mead and Agostoni, 1964).

Although most fits obtained using Eq. (45) are adequate, the interpretation of the two terms as laminar and turbulent components is more or less arbitrary. Another empirical equation describes the pressure difference between the mouth and the alveoli over a wide range of flow rates as well:

$$\Delta P = K \dot{V}^n \tag{46}$$

where n has a value ranging from 1 with laminar to 2 with turbulent flow, so that n is a useful indicator of the nature of the flow. For the normal human respiratory tract, the constants are found to be $n = 1.3$ and $K = 2.4$ when ΔP is given in cm water and $\dot{V}$ in liters per second. It has been shown by Pedley, Schroter, and Sudlow (1970) that if nonparabolic flow profiles in branching tubes are taken into account, a theoretical expression can be derived in which $n = 3/2$, and K is dependent on density and viscosity of the gas.

Glossary

Aggregation: unspecific reversible conglomeration of erythrocytes or other cells within a saline solution in which proteins are dissolved, for example, blood plasma.

Alveolus: a small cavity, as one of the terminal thin-walled chambers of the lung, surrounded by networks of capillaries through whose walls the gaseous exchange take place.

Aneurysm: a circumscribed dilatation of an artery or a blood-containing tumor, which stands in open connection to the lumen of an artery.

Computed tomography (CT): a medical imaging method employing tomography where digital geometry processing is used to generate a three-dimensional image of the internals of an object from a large series of two-dimensional X-ray images taken around a single axis of rotation.

Doppler Ultrasound Device: an instrument for the measurement of blood velocity, based on the Doppler effect. This effect states that the frequency of sound reflected from moving blood cells is altered in proportion to the velocity of the cells.

Echocardiography: a diagnostic procedure that utilizes the reflection of ultrasound pulses to visualize the heart in a noninvasive manner.

Hypertrophy: excessive development of the size, but not the number, of single cells of an organ due to increased load.

Infarction: necrotic changes in cells resulting from the engorgement of an artery supplying these cells.

Magnetic resonance imaging (MRI): a noninvasive radiology technique that uses magnetism radio waves, and a computer to render images of the inside of an object. It is primarily used in medical imaging to demonstrate pathological or physiological attentions of living tissues.

Norepinephrine: a hormone secreted by both the adrenal medulla and the sympathetic nerve endings in response to sympathetic stimulation; it causes constriction of the blood vessels.

Papaverine: a benzylisoquinoline alkaloid of opium; it causes peripheral vasodilation by inhibiting phosphodiesterase, an enzyme that catalyzes the breakdown of cyclic 3′,5′-adenosine monophosphate, so that the intracellular level of this substance increases.

Parenchyma: the actual cellular substance of a gland or organ, excluding the connective tissue.

Pulsed Doppler Device: an instrument that combines echo ranging and the Doppler effect for the determination of the velocity of blood.

Sphygmogram: an uncalibrated pressure pulse curve obtained by means of a displacement transducer applied, at appropriate pressure, to the skin above an artery.

Spirometer: a device for measuring respiratory volumes. Classically, it is a counterpoised displacement bell chamber floating in a water tank. A stylus fixed to a counterweight is made to write the volume displacement movements of the chamber. The bell contains a breathable gas (air or oxygen). A tube connects the mouth to the chamber. When a test subject breaths in and out of the chamber, the bell rises and falls, thus permitting quantitation of several pulmonary parameters.

Stenosis: a narrowing of an artery or vein.

Acronyms

CE	Contractile element
EP	Ejection period
ERV	Expiratory reserve volume
FRC	Functional residual volume
IRV	Inspiratory reserve volume
PE	Parallel elastic element
PEP	Preejection period
Re	Reynolds number
RV	Residual volume
SE	Series elastic element

TLC	Total lung capacity
TRP	Total peripheral resistance
TV	Tidal volume
VC	Vital capacity

References

Anliker, M., Casty, M., Friedli, P., Kubli, R., and Keller, H. (1977) *Cardiovascular Flow Dynamics and Measurements* (eds N.H.C. Hwang and N.A. Norman), Chapter 1, University Park, Baltimore.

Fåhraeus, R. (1929) *Physiol. Rev.*, **9**, 241–74.

Fåhraeus, R. and Lindqvist, T. (1931) *Am. J. Physiol.*, **96**, 562–68.

Frank, O. (1920) *Z. Biol.*, **71**, 255–72.

Haynes, R.H. (1962) *Biophys. J.*, **2**, 95–103.

Hill, A.V. (1938) *Proc. R. Soc. London Ser. B*, **126**, 136–95.

Holt, J.P. (1941) *Am. J. Physiol.*, **134**, 292–300.

Katz, A.I., Chen, Y., and Moreno, A.H. (1969) *Biophys. J.*, **9**, 1261–79.

Majno, G. (1965) *Handbook of Physiology, Section 2: Circulation* (ed W. Hamilton.), Vol. 3., Chapter 64, American Physiological Society, Bethesda.

Mead, J. and Agostoni, E. (1964) *Handbook of Physiology, Section 3: Respiration* (eds W.O. Fenn and H. Rahn), Vol. 1., Chapter 1, American Physiological Society, Bethesda.

Mirsky, I. (1979) *Handbook of Physiology, Section 2: The Cardiovascular System* (eds R.M. Berne, N. Sperelakis, and S.R. Geiger), Vol. 1., Chapter, American Physiological Society, Bethesda.

Pedley, T.J., Schroter, R.C., and Sudlow, M.F. (1970) *Respir. Phys.*, **9**, 387–405.

Suga, H., Kitabatake, A., and Sagawa, K. (1979) *Circ. Res.*, **44**, 238–49.

Weibel, E.R. (1963) *Morphometry of the Human Lung*, Academic, New York.

Wetterer, E., Bauer, R.D., and Busse, R. (1978) *Cardiovascular and Pulmonary Dynamics* (ed. M.Y. Jaffrin), INSERM-Euromech 92, 1977, Vol. 71, p. 17, INSERM, Paris.

Further Reading

Bauer, R.D., Busse, R. (eds) (1978) *The Arterial System – Dynamics, Control Theory and Regulation*, Springer, Berlin.

Bergel, D.H. (ed) (1972) *Cardiovascular Fluid Dynamics*, Academic Press, London.

Caro, C.G., Pedley, T.J., Schroter, R.C., Seed, W.A. (1978) *The Mechanics of the Circulation*, Oxford University Press, Oxford.

Charm, S.E., Kurland, G.S. (1974) *Blood Flow and Microcirculation*, Wiley, New York.

Ferry, J.D. (1970) *Viscoelastic Properties of Polymers*, Wiley, New York.

Flügge, W. (1975) *Viscoelasticity*, Springer, Berlin.

Fung, Y.C., Perrone, N., and Anliker, M. (eds) (1972) *Biomechanics – Its Foundations and Objectives*, Prentice-Hall, Englewood Cliffs.

Gabelnick, H.L. and Litt, M. (eds) (1973) *Rheology of Biological Systems*, Thomas, Springfield, IL.

Kenner, T., Busse, R., and Hinghofer-Szalkay, H. (eds) (1982) *Cardiovascular System Dynamics – Models and Measurements*, Plenum, New York.

McDonald, D. (1974) *Blood Flow in Arteries*, Arnold, London.

Noordergraaf, A. (1978) *Circulatory System Dynamics*, Academic Press, New York.

Shepherd, J.T. and Vanhoutte, P.M. (1979) *The Human Cardiovascular System – Facts and Concepts*, Raven, New York.

West, J.B. (ed) (1977) *Bioengineering Aspects of the Lung*, Dekker, New York.

Wetterer, E. and Kenner, T. (1968) *Grundlagen der Dynamik des Arterienpulses*, Springer, Berlin.

BUS SYSTEMS AND COMPUTER INTERFACING

PETER CLOUT, *Vista Control Systems, Inc., Los Alamos, New Mexico, U.S.A.*

	Introduction	639
1.	**Basic Subfunctions of a Bus**	640
1.1	Data	640
1.2	Addressing	640
1.2.1	Geographical Addressing	640
1.2.2	Logical Addressing	640
1.2.3	Broadcast Addressing	642
1.2.4	Addressing Conventions	642
1.3	Control	642
1.3.1	Timing and Synchronization	642
1.3.2	Transfer Mode	642
1.3.3	Device Control	643
1.3.4	Status	643
1.3.5	Interrupts	643
1.4	Arbitration	644
1.4.1	Bus Arbitration	644
1.4.2	Network Arbitration	645
1.4.2.1	Token Arbitration	645
1.4.2.2	CSMA/CD	645
1.5	Cache Coherency	645
1.6	Error Detection and Correction	645
1.7	Bus Device Fault Handling	646
2.	**Electrical and Mechanical Considerations**	646
2.1	Signal Propagation	646
2.2	Mechanical Design	646
2.3	Cooling	646
3.	**Software Considerations**	646
3.1	Device Access Software Protocol	646
3.2	Real-Time Executive and Development System	647
4.	**Bus Requirements**	647
4.1	Single-Processor Bus	647
4.2	Multiprocessor Bus	647
4.3	Computer Peripheral Bus	648
4.4	I/O Bus	648
4.5	Intercomputer Bus, Networks	648
4.6	Instrumentation Bus	649
5.	**Current Standards Categorized**	649
6.	**System Design Considerations**	649
6.1	Desires and Technical Capabilities of the Staff Available	649
6.2	Technical Culture	649
6.3	Availability of Boards and Support Software	649
6.4	Longevity of the Bus	650
6.5	Compatibility	650
	Glossary	650
	Works Cited	651
	Further Reading	651

INTRODUCTION

It was over 20 years ago that buses first became available and used in computers and laboratory data acquisition. The announcement of the Digital Equipment Corporation PDP-11/20 computer introduced the revolutionary concept of a single bus in a computer for all communication between the individual circuit boards that made up the computer. This one bus, the Unibus™, was the single device that connected, physically, logically, and electrically, the CPU, main memory, and I/O boards that made up a fully operational computer. The PDP-11 also introduced the idea that the I/O subsystem of a computer could be treated as memory and thus addressed with the same instruction set.

At about the same time, a number of high-energy physics laboratories in Europe published the specifications for a standard data-acquisition bus system, CAMAC. CAMAC only addresses the I/O requirements of a computer for data acquisition and so the combination of CAMAC and a PDP-11 was a good match. It is interesting to note that the Unibus role has been relegated to a secondary I/O bus and will soon no longer be manufactured, while CAMAC is still being manufactured for and used in new data-acquisition systems.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

These two buses have been superseded in both capability and performance. However, the current investment in board designs, know-how, and, in the case of CAMAC, software investment keeps the use of these buses effective.

Another bus that had its origins at about the same time is the IEEE 488 bus. This started as a bus to connect up to 15 instruments and was developed initially by the Hewlett–Packard Company. The bus definition was taken over by the IEC and an IEEE committee to develop the bus design further as an international standard. IEEE standard 488 was published in 1975. As with all active standards, further documents were published extending its scope.

Since this early work, there has been constantly increasing activity in designing and using existing and new buses. At the same time, the application areas for buses have increased in size and number with the capabilities. We have reached the position where there is a confusing number of buses described and advertised in the trade journals. This article is an attempt to make sense of this confusion and thus help in the process of making a choice for a new project. We describe all the possible functions of a bus and detail the different implementations where applicable in a general way. The requirements for different applications are described together with software considerations. Current and emerging bus standards are broadly categorized so that a subset can be chosen from this information for further detailed study. Finally, some other considerations to recognize are described. Because of the diversity of buses, manufacturers, and software support, detailed descriptions are impossible in an article of this scope, and the reader should refer to the standards listed in Table 1 and the current trade press.

In principle, any bus for which the specifications are available can be used for computer interfacing but in practice one is often limited by other project constraints. These might be performance, computer choice, availability of board designs either commercially or in-house, or higher-level management decisions.

1. BASIC SUBFUNCTIONS OF A BUS

This section defines and discusses the various kinds of information that are transmitted on a bus and the possible capabilities of a bus. The application and design of an individual bus will define how these kinds of information are transferred and which capabilities are supported.

1.1 Data

The whole purpose of a bus is the transfer of data. All the buses transfer binary data with the significance of the bits clearly defined. The buses differ in the amount of data transferred in each bus cycle or message, and this is normally a multiple of 8 bits.

1.2 Addressing

The bus master, in initiating a data transfer on the bus, must define the address of the device or register to transfer the data from or to. The bus master is the single bus device that initiates the bus cycle. When there is the possibility of more than one bus master, the potential masters of the bus must arbitrate for which will be the single master of the next bus cycle.

The addressing scheme of the bus is an important property of the bus and will reflect the applications for which it is designed. It is interesting to note that a device on the bus which only takes part in bus transactions as a master requires no bus address.

1.2.1 Geographical Addressing. Geographical addressing is implemented when the position on a bus where the device is installed defines the primary address. The geographical address will define the base address for the address space of the device. With geographical addressing, each device slot on the bus has a fixed amount of address space to use. The size of this address space for each device is defined by the specification of the bus.

1.2.2 Logical Addressing. In logical addressing the base address of a device is independent of the position of the device on the bus and is instead determined either by switches or by a register on the device. Logical addressing allows the address space of the bus to be divided up freely according to the needs of the devices and the application. This

Table 1. Currently available and newer buses compared.

Bus name	Age[a]	Date of standard	Address width	Speed[b]	Interlocks?	Distance[c]	Arbitration[d]	Cache coherency	Physical form[e]	Notes	IEEE Standard number/Ref.
CAMAC	M	'69–'83	13	M	N	S/M/L	P	N	B&C	Original standards published by ESONE and NIM committees	583,595,596,675 683,726,758, Clout (1982)
Unibus	O	'70	18	M	Y	S	P	N	B	Digital Equipment Corp.	
Q-bus	M	'75	18/22	M	Y	S	P	N	B	Digital Equipment Corp.	
GPIB	M	'75–'87	4	L	N	S	P	N	C	Hewlett–Packard Corp.	488,728,P981
S-100	M	'83	16/24	M	Y	S	P	N	B	16-bit data	696
Multibus I	M	'83	24	M	Y	S	P	N	B		796
Futurebus	I	'87	32	VH	Y	S	P&F	Y	B		896
Fastbus	I	'86	32	VH	Y	S/M	P	N	B&C		960
STD bus	M	'87	16/24	M	N	S	P	N	B	8-bit data. General specification	961
Versabus	M	'85	32	H	Y	S	P	N	B	Predecessor to VME	P970
PC bus	M	'83	20	M	Y	S	N	N	B		P996
STE bus	Y	'87	20	M	N	S	P	N	B		1000
VME	Y	'87	32	H	Y	S	P&F	N	B		1014
VSB	Y	'88	32	H	Y	S	P	N	B		1096
Nubus	Y	'87	32	H	Y	S	P	N	B		1196
Multibus II	I	'87	32	H	Y	S	P	N	B		1296
IBM Microchannel	Y	'87	32	H	Y	S	P	N	B	Supports PC and AT bus	
EISA	Y	'88	32	H	Y	S	P	N	B	Half of the pins =AT bus, superset	
STD-32 bus	I	'89	32	H	N	S	N	N	B	Ziatech Inc.	
Futurebus+	D	'90	64	EH	Y	S	P&F	Y	B		P896.1
Ethernet	M	'85	48	M	N/A	M	C	N	C		802.3
G-64	M	'79	16	M	Y	S	P	N	B	G-96 has 25-bit address	GESPAC (1985)
Token Ring/Bus	M	'85	48	L–M	N/A	M–L	T	N	C		802.5,802.4
MIL Std 1553B	M	'78	5	L	N	M	N	N	C		Titus (1985)

---

[a] Age: D=Design; I=Infancy; Y=Youth; M=Maturity; O=Old age.

[b] Speed: L= <1 Mbyte/s; M= <10 Mbyte/s; H= <100 Mbyte/s; VH= <1 Gbyte/s; EH= <10 Gbyte/s.

[c] Distance: S=Crate or rack; M=Room or building; L=Campus.

[d] Arbitration: P=Priority; F=Fair; C=CSMA/CD; T=Token ring; N=None, polled bus.

[e] Physical form: B=Bus; C=Cable.

form of addressing also allows a device to have two or more base logical addresses.

1.2.3 Broadcast Addressing. Broadcast addressing allows a master to address multiple slaves in a single bus transaction. Different broadcast addresses may be assigned to classes of devices. Broadcast addressed bus transactions are normally data transfers to the bus slaves.

1.2.4 Addressing Conventions. In order to make the use of a bus simpler for both software and hardware, certain addressing and register conventions are often mandated. These conventions might be, for example, the relative addresses for control and status registers as well as the bit allocations and use of those registers.

Some buses separate control and data registers into different address spaces. An obvious example would be the error registers of a computer memory board which clearly cannot be part of the memory address space.

A particular bus specification can use more than one addressing scheme. Consider the problem of establishing the device addresses of a particular system on powering up the system. This is not a problem with geographical addressing so long as the devices are correctly configured on the bus. With logical addressing, the device base addresses could be set by switches, but this is prone to error both in setting up and in repairing the system. A much more flexible scheme is to write the base logical addresses to the devices in the system initialization phase. This is only possible if there is a device address that survives power-up of the system and only a geographical address guarantees a unique power-up device address.

1.3 Control

The control signals of a bus orchestrate the bus transactions and communication. The bus protocol that these control signals implement has to be implemented in a different way for computer networks. In a computer network the control protocol is implemented in the arbitration and transmission protocol and the control bits of the defined message structure. Bus arbitration is dealt with separately in the next section as it is a significant topic.

1.3.1 Timing and Synchronization. The timing signals on a bus are used to synchronize the transaction. There are two basic types of bus transaction, a synchronous or an asynchronous transaction. Synchronous transactions have their timing set by the bus master within the rules of the bus specification. The transaction slave must respond within the time given. Thus the bus master will establish the bus address lines and possibly the bus data and control lines, and then set a timing line to indicate to the bus slaves that the address and data lines are correct and that the address lines should be decoded and acted on. The addressed slave then has a fixed time to respond.

Asynchronous transactions are ones where the master starts the bus transaction and the slave completes the timing of the transaction. The first part of the bus cycle is the same as the synchronous bus cycle but the master then keeps its timing line set until either the slave sets a second timing line to indicate that its response is complete or a timeout occurs. The master would then accept data, in the case of a bus read, and clear its timing line. When the slave sees the master timing line clear, it can remove its response from the bus and clear its timing line.

Thus asynchronous bus transactions can adapt to the performance of the individual devices rather than have the synchronous bus timing set by the worst-case device performance. To balance this advantage, while data only travel one way on the bus, the data transfer synchronization signal has to make a round trip between the master and the slave, adding to the bus timing. Asynchronous protocols also need a timeout mechanism to prevent bus lockup when a device does not respond for whatever reason.

1.3.2 Transfer Mode. In order to implement more than one transfer mode, the bus master must have a mechanism for defining the required transfer mode on the bus. A minimum set of types of bus transfers is read and write. Additional transfer types are the following:

1. Read–modify–write. This mode allows uninterrupted modification of a device register or memory location by one bus master. The definition of this type of cycle provides an interlock mechanism on the bus.

2. Block data transfer. This mode allows a block of data to be transferred at the cost of a single bus arbitration cycle. Each cycle of data transfer is accompanied by an address. The cost of block data transfers is an increase of the maximum bus access latency for the other masters on the bus. When designing systems, care must be taken to ensure that no device will block access to the bus for an unacceptable time.
3. Pipeline data transfer. This mode allows a block of data to be transferred at the cost of a single bus arbitration and address cycle. This transfer mode is only applicable to bus designs that place the time-multiplexed address and data onto the same physical lines of the bus. This transfer increases the data transfer performance at the cost of increasing the bus access latency for the other masters on the bus.
4. Arbitration locked access. This mode allows a bus master to keep the use of the bus for a number of bus cycles. This access mode may not require any dedicated bus lines but rather represent a new bus protocol using existing lines.
5. Increment. This mode is used in data-acquisition systems to accumulate spectra and histograms. This is normally implemented within the memory accumulating the spectra rather than as a defined bus cycle.

In some bus designs, the address and data are time-multiplexed onto the same physical lines of the bus. This can result in a considerable savings in connector pins and bus logic at the expense of the extra time taken for a separate address and data cycle for each bus transaction. This overhead can be significantly reduced by the use of pipelined data transfers.

1.3.3 Device Control. Each bus requires some basic device control lines. At a minimum, a reset signal is required to set all device logic into a known state when the system is initialized. Other bus lines might be these:

1. A busy line, set whenever a bus transaction is in progress.
2. A power-low line to indicate an imminent power failure. This would allow devices to prepare for a safe shutdown and possible transparent start-up after power is restored.

1.3.4 Status. The status lines are basically the response of the bus slave to the bus cycle of the master. One might be defined to indicate a status return from the slave to an enquiry from the master while another might be an indication to the bus master that the transaction was a valid bus request of the slave.

1.3.5 Interrupts. Interrupts are really a mechanism for a device that is usually a bus slave to initiate bus transactions. When a bus slave asserts a bus interrupt, it is indicating to one or more bus masters that it needs attention. Thus the CAMAC standard gave interrupts the name look-at-me's or LAM's. From the master's point of view, a LAM is an interruption to its current thread of execution.

Interrupts can be supported in a bus design at a single level or at multiple levels. Because it is not possible to nest the responses to a series of interrupts from a device during the master's response to an interrupt, further interrupts must be blocked by the bus master. Thus, with a single level of interrupts, an interrupt of lesser importance can block the rapid processing of an important interrupt. This can be overcome if a priority interrupt scheme is defined both for the bus and for the processor which responds to the interrupts as bus master. The interrupt latency is the time between assertion of an interrupt and the start of the processing of the response. In comparing interrupt latency times care must be taken that the same end point for the time is used. Possible end points are a bus interrupt response cycle, a CPU interrupt response cycle, or the first CPU instruction which responds to the specific interrupt. Outstanding higher-priority interrupts and the current processing of other interrupts at the same priority will increase the latency while lower-priority interrupts in any state will not significantly affect the interrupt latency.

For the response to an interrupt to proceed, the interrupting device must be identified in some way to the master responding to the interrupt. For buses using logical addressing, the device will present a vector or some other identification to the interrupt response code in the master as part of an interrupt cycle. This vector can be derived from switches or a register written as part of the system initialization phase. The interrupt response protocol must select a single, highest-priority, interrupting slave to respond to the interrupt cycle. This arbitration can use the general bus arbitration described below (Sec. 1.4) or a separate, possibly simpler and less expensive, arbitration scheme. This is normally the grant chain. It is interesting to note that the Q-bus™ supports four interrupt levels and thus has four interrupt request lines, but it has only a single interrupt cycle grant chain, thus requiring that boards be placed in interrupt priority order along the bus. Although some systems will work with the boards placed in another order along the bus, this level of the interrupt response is defined in the microcode or logic of the computer and misplacing the boards can give hardware errors when two interrupts occur at about the same time.

For buses that support geographical addressing, it is possible that the bus design includes a mechanism for the bus master responding to the interrupt to identify the interrupting stations.

1.4 Arbitration

This area of bus design is by far the most contentious. It is also the most intellectually demanding and many bus board designs have had to be reworked in this area due to poor guidance from the bus specification or poor design and incomplete testing. During a bus cycle, one or more potential bus masters, including the current bus master, may become ready to use the next cycle of the bus. Equally, while the bus is idle, two or more potential bus masters may simultaneously become ready to use the bus. The arbitration scheme of a bus must

1. resolve requests to use the next cycle of the bus from one or more potential bus masters;
2. ensure that the system designer is able to maximize system performance by setting arbitration priorities.

It is about this latter point that much discussion occurs. Should the priority scheme be a fair one, so that at each arbitration each bus master has an equal chance of getting the bus, or should it be a set priority so that a low-priority bus master may be blocked from the bus for a considerable time? Can the priority be set by the task within the potential bus master that requires the use of the bus? Can one mix these schemes?

1.4.1 Bus Arbitration. The overall arbitration circuit of a bus can be either a central arbitration circuit, to which every potential bus master makes a request and one receives a grant, or distributed, where the circuit making the decision is divided between the potential bus masters. The key is the location of the logic that decides if a particular potential bus master next uses the bus. If that circuit is on a single board for all the bus masters, then the implementation is a centralized one. If the arbitration circuit is on each potential bus master, the arbitration scheme is a distributed arbitration scheme even though there may be some centralized bus logic to time the arbitration cycle. The distributed arbitration logic can be designed to be serial or parallel. An example of a serial arbitration scheme is the request/grant chain. In this scheme each priority level has a bussed request line which is monitored by the bus logic. The grant line for each priority level goes in and out of each board on separate bus pins and it is only passed on if that board has no outstanding request at that priority level. A disadvantage of this scheme is that the grant signal has to propagate down the bus, through the arbitration circuit of each potential bus master, until it reaches the potential bus master requesting the bus. This time is a function of the priority of the requesting master rather than the arbitration need.

In a parallel arbitration scheme, all potential bus masters carry out the arbitration algorithm in parallel, thus increasing the speed of arbitration. The arbitration algorithm can be designed such that the time taken is a function of the number of potential bus masters needing access to the bus, thus speeding the process for the majority of arbitration cycles where there is only one requesting bus master.

An example of a parallel arbitration scheme, taken from the S-100 and Fastbus standard, is one where each master on the bus is assigned a unique priority level to use during the arbitration cycle. During the arbitration cycle each master requesting the bus asserts its priority level on the arbitration bus with the following rule: if a master finds that any of its arbitration priority bits differs from the value of the corresponding bit on the arbitration bus, that master inhibits the placing of its data on all arbitration bus lines of lower significance. This process may well go through several cycles and will be complete when one requesting master detects that the value on the arbitration bus is equal to its own internal priority level. In practice, there will need to be some delays built into the scheme which reflect the bus propagation delays.

1.4.2 Network Arbitration. Network arbitration is by nature distributed because of the extended distances covered. For baseband systems, there can be no dedicated arbitration lines or signal paths. On broadband systems, one or more channels of the frequency spectrum can be dedicated to network access arbitration. Two network arbitration algorithms are presently widely used, token ring and CSMA/CD. In a network, all nodes are potential bus masters, where a bus master only writes to the destination node, the slave, or to a group of nodes defined by a group address.

1.4.2.1 Token Arbitration. Token ring is the network equivalent of the request/grant chain. The difference is that the token, or permission to transmit on the network, is continuously circulated on the network between the nodes of the network. The minimum average access latency is half the token rotation time while the maximum access latency is the token rotation time plus the maximum message transmission time multiplied by the number of nodes on the network. It is this defined maximum access time that attracts many to this network arbitration scheme. Clearly, the token arbitration algorithm has to take account of nodes being added and removed from the network and for tokens being lost for whatever reason. The protocol to detect and replace lost tokens defines the absolute maximum access time.

1.4.2.2 CSMA/CD. The origin of the Collision Sense Multiple Access with Collision Detect (CSMA/CD) is a radio-based computer network, Aloha Net, developed by the University of Hawaii. The network scheme was developed into Ethernet by the Xerox Corporation, which holds patents on the network. In this scheme, each station or node of the network is able to

1. detect if any other node is transmitting;
2. while it is transmitting, detect if another node is also transmitting, a collision.

With this capability, a node that wishes to transmit simply starts transmission if no other node is detected as transmitting. However, two nodes some distance apart could start before the signal from one node can reach the other. If this happens, then both nodes will detect the collision and thus abort that transmission. Both nodes then enter a back-off and retry algorithm which is designed to minimize future collisions.

With CSMA/CD, bus or network average access latency is a function of the loading of the network. It can be effectively zero for a lightly loaded network and theoretically infinite in a very extreme case.

1.5 Cache Coherency

A cache is a local copy of recently used instructions or data that is stored in a local, fast memory on the basis that, once having been used, it is very likely to be used again. Thus, in a multiprocessor system with some common main memory, it is possible that some data have one or more copies stored in local caches as well as in the main memory. What then happens if another processor writes a new value to the main memory? Clearly, it is unacceptable that the other processors continue to use the old data stored in their local cache. Thus, some mechanism is needed to detect updates to the data in the main memory and invalidate the local cache copy. Only the most recent bus standards incorporate a cache coherency protocol.

1.6 Error Detection and Correction

High-performance buses used to build computer systems operate in a well-controlled environment. As a result, they generally have very low error rates due to electrical noise and thus either have no error-detection facility or have some level of simple parity checking. On the other hand, computer networks are quite prone to errors, especially those using common carrier facilities, and as a result extensive error-detection and correction facilities are needed. This correction of detected transmission errors can be within the message, using error-correcting codes, or by a retransmission scheme, or both.

1.7 Bus Device Fault Handling

Devices on the bus will fail. If the system using the bus is designed to be fault tolerant, other devices on the bus must, at least, be able to continue operation and possibly take over the functions of the failed module while that module is found and replaced. At a minimum, this puts the requirement on the bus design for live removal and insertion of devices and the initialization of newly inserted devices in an otherwise running system. A bus that uses a request/grant chain for bus arbitration will not tolerate the removal of any device from the bus without some back-up protocol.

2. ELECTRICAL AND MECHANICAL CONSIDERATIONS

2.1 Signal Propagation

The key to high performance is to maximize the actual signal bandwidth of the chosen transmission medium. This is done by several means:

1. The choice of the medium.
2. The electrical signal standards for driving the medium and receiving the signal. Here the cost and circuit board area taken by the chosen technology are important in local bus systems.
3. The use of medium bandwidth. This is less critical in local bus systems where the cost of the connections to the backplane is not great. However, for computer networks, the arbitration and network management overhead can become an important factor.

2.2 Mechanical Design

For local bus systems, the mechanical design defines the circuit board area available for the logic of a device and thus the cost of providing a board for a given function. The mechanical design also defines the connector scheme, and here, for local bus standards, the trend has been away from printed-circuit edge connectors to manufactured connectors mounted on the circuit board. This trend is because the manufactured connectors have greater reliability resulting from tighter quality control, and their use eliminates the need for gold plating in the circuit board production process. The standards here are the Eurocard mechanics, IEEE Standard 1011, and the DIN connector standard, IEC 603-2.

2.3 Cooling

High-performance systems invariably use considerable power and thus require well-designed cooling. A bus system providing defined, uniform cooling will simplify the board designer's and system builder's task considerably as well as adding to the reliability of the system.

3. SOFTWARE CONSIDERATIONS

In the process of designing and building a system there will be many choices to be made between the available components on the market that meet the basic system performance requirements. One of the factors that can help to eliminate some of these choices is the cost of completing and maintaining the system once the components have been acquired. This section addresses the software component of these costs.

3.1 Device Access Software Protocol

A bus specification describes how data are written and read by a bus master, usually a processor. These data are, to the bus, just words made up of bits. At the other end of the system is a stream of computer instructions that are defined by a high-level language. The questions here are these:

1. Are there any other considerations in a bus design that need to be taken into account in order to simplify the application software?
2. Are there any other related standards needed to simplify the use of the bus, to bridge between the low-level bus protocol and the high-level language application program?

Where the devices on the bus are mapped into the memory space of the bus, just as is processor main memory, the software problems are those set by the operating system or executive of the processor. On the other hand, those buses that are designed solely as I/O or instrumentation buses are likely candidates for standardization. There are two areas of standardization:

1. Device command and data protocols. These will allow devices with similar functions to use the same software.
2. High-level language extensions to add access to the devices on the bus. These extensions can take the form of new statements added to the language or a set of subroutines or functions with which to link the program.

Examples of the first are the function definitions of CAMAC, and Fastbus, Recommended Practice for Code and Format Conventions for Use with IEEE Standard 488, and the emerging ISO standards for network software (ISO, 1984). Examples of the second software standardization are the Subroutines for CAMAC and Real-Time Basic for CAMAC.

3.2 Real-Time Executive and Development System

Careful consideration should be given to the choice of the real-time executive and, in particular, the software development system that will allow the design, coding, and testing of the software to complete the system. Not all systems on the market support all the combinations of CPU board, I/O controller, and network controller. A basic decision is between independent development on the system or cross-development using another computer system. The former choice is less expensive but is usually less productive for the programmers compared to the latter choice.

4. BUS REQUIREMENTS

In this section we consider the requirements for particular applications of a bus. Where a bus is required to meet a number of these requirements, they will need to be combined. Table 2 summarizes the functional requirements.

4.1 Single-Processor Bus

The requirement for a processor bus is to provide the connection between the CPU and the memory. The address width of the bus must be at least the physical addressing capability of the CPU, and the performance of the bus should not be the limiting system performance factor. The data width of the bus is a system design option where the tradeoff is between cost and performance. With RISC architectures for processors, the requirement for memory access bandwidth increases dramatically because of both the increased number of instruction bytes required per clock cycle and the increased clock frequencies. The provision of cache memories on the processor boards can reduce the required bandwidth by a factor between 10 and 100, depending on the application.

A processor bus has no requirement for interrupt support and, as memory is always a slave in bus transactions, a single-processor bus needs no bus arbitration.

4.2 Multiprocessor Bus

The requirements added to support multiprocessing are those that support more than one master on the bus accessing the memory. In addition, on the assumption that the processors are cooperating as part of a single sys-

Table 2. Requirements for different bus functions.

Bus function required	Interlocks	Arbitration[a]	Distance[b]	Cache coherency
Single-processor bus			S	
Multiprocessor bus	Y	F	S	Y
Peripheral bus		P	S/M	
I/O bus		P	S/M/L	
Intercomputer bus		F	M/L	
Instrument bus		P	M	

[a] Arbitration: P=Priority; F=Fair.
[b] Distance: S=Crate or rack; M=Room or building; L=Campus.

tem, there will need to be a mechanism for interlocking access to common data structures.

The new requirements are thus a bus arbitration scheme, a cache coherency feature, and a read–modify–write or similar bus cycle to implement a processor interlock instruction such as test-and-set.

Depending on the application, the bus arbitration scheme needs to be fair, for general data processing, or priority based for deterministic real-time systems.

4.3 Computer Peripheral Bus

Computer peripherals, such as disk, magnetic tape, and terminal controllers, have different requirements from those of the processor bus. The peripherals will use direct memory access for the data transfer and will thus require the same addressing capability as the CPU–memory bus. This address width can be provided by the bus interface with mapping registers, thus mapping an assigned region of the computer peripheral bus address space to the appropriate region of the full address space.

There will be the need for a program running in the computer to define the I/O operation to the peripheral and then read the status of the completed operation. This is often provided by the peripheral having registers which can be read and written in the memory space of the computer. Another technique is to establish a block of memory which defines the operation required and then point the peripheral to the block. This technique is used in computer systems having high I/O bandwidth requirements.

Unlike main memory, computer peripherals do not work at the same speed as the CPU and thus interrupts are needed to notify the CPU of the completion of an I/O operation. Equally, the peak data-transfer bandwidth is usually at least a factor of 10 less than the memory bandwidth. The arbitration required is usually a priority arbitration scheme in order that the needs of the computer I/O subsystem are filled and it is kept busy.

4.4 I/O Bus

The I/O bus is here assumed to be the computer's real-time connection to a device, experiment, industrial plant, etc. These connections may measure temperature, pressure, voltage, position, intensity, status, etc., and also control similar parameters. The requirements for this bus are a reasonable address width so that each electronic module can have a number of data and control registers, and a data width that will allow all measurements to be read in a single bus operation.

For this application, direct memory access for the modules to the computer memory is not a requirement but could find application in some systems. This direct memory access facility is then implemented by the controller of the I/O bus, which is processor dependent. Where the I/O bus is also the processor bus, direct memory access facilities can be implemented on the I/O modules.

An I/O bus does require an interrupt facility for all but the simplest, non-real-time applications. This will allow the computer to respond to events in a predictable manner without requiring the computer to continually look for the possibility of an event, a process called polling.

Some application may require that some of the I/O boards in the system be remote or are required to be, for example, in a high-voltage enclosure. In these cases, a standard is required where crates can be connected with communications, normally serial to reduce costs, implemented with fiber optics.

4.5 Intercomputer Bus, Networks

The requirement here is the ability to pass blocks of data between the computers that make up the network. For reasons of reliable data transmission, that is, the assurance that the data were received and were correct, and the need to share one physical link for many uses, other protocols are usually layered on top of the basic message passing system.

The Open Systems Interconnect [OSI (1984)] reference model defines seven layers of protocol for the interconnected computer system. The lowest two layers are the physical layer and the data-link layer. Between these layers implemented on different computers, logical links are established over the single physical link. The OSI model allows the protocols and interfaces between the layers to be independent of the data link used. Thus, different data links are available depending on the distance between the computers and the data-transmission rate required. The shortest distance, within a computer room, can use a bus with many wires and a simple low-level protocol. Within a building, the cabling costs become more important and thus network connections use either a single coaxial cable connection or one or two twisted pairs. Campus-wide networks also use coaxial or twisted-pair cable. Both the building-wide connections and the campus-wide connections transmit data at a raw rate of 1 to 10 MHz. Wide-area connections emphasize low-cost transmission media at the expense of the cost of the devices that connect the transmission media to the computer. From these considerations it will be seen that, in general, the needs of a network are simple, from the bus point of view, and thus most buses can support intercomputer communication. However, outside the computer room the emphasis is on the lowest cost of cabling for a given data rate.

4.6 Instrumentation Bus

It is generally assumed that instruments here are self-powered, free-standing instruments. For communication they need a bus that is a simple cable to connect them together and connect them to the computer or other controller. In the past, many instruments have had a simple terminal interface but the simplicity of the connection often had to be traded for complexity in the communication protocol.

5. CURRENT STANDARDS CATEGORIZED

Table 1 shows the key characteristics of the currently available and newer buses. This table concentrates on the general characteristics of interest to the system designer and is intended to enable the system designer to narrow the choice for detailed study.

6. SYSTEM DESIGN CONSIDERATIONS

All this information is input to the key decisions involved in designing and building a system. The system requirements must be weighed against the choices and decisions made based on a broad picture of how the system will be implemented. There are, however, other considerations to be recognized and evaluated. These are discussed in the remainder of Sec. 6.

6.1 Desires and Technical Capabilities of the Staff Available

Work with any of these buses will involve gaining some understanding of the operation of the buses and the devices that are connected to them. At least there will be a need to establish testing facilities so that faulty modules can be identified and returned for repair. The higher-performance buses are more complex and generally difficult to trouble-shoot. To balance this, recent designs have emphasized the self-testing of modules. Thus, the ability to assemble, test, and maintain the system during its life by the staff available is an important consideration.

Another consideration is the opinions of the staff that will be building the system and maintaining it during its life.

6.2 Technical Culture

This covers the previous experience of the group of people involved, and both the type of support they expect and how this matches with the support that the vendors are able to give. Some people are used to integrating their systems and even designing and building many components, while others require service and are able to pay for it. Equally, the use of a particular bus in other parts of the same organization will mean that there are people available for consultation who understand the application.

6.3 Availability of Boards and Support Software

The newer buses do not have the broad range of board designs and support software available, while the older buses require fewer, if any, designs of special boards and are easier to use. If one is planning to use a relatively new bus, one is making the decision based on a judgment of the future support for the bus. This introduces some risk into the decision, and this risk must be recognized.

6.4 Longevity of the Bus

Many systems are built to have a long operational life, and, should they need upgrading because of obsolete buses or computers, there will be many years of refinements and additions to be reengineered in the replacement system. Often, the difficulty introduced by this continual improvement of the system makes a necessary upgrade very difficult. Good technical judgment is needed here to ensure that the support for the technical choices made will be available for the life of the system.

6.5 Compatibility

It is an unfortunate aspect of many of the bus standards that the specification leaves choices for the board designer that can make different boards that meet the bus standard incompatible with each other. The system designer therefore has to identify these aspects of the chosen standard and ensure that the components purchased are compatible with each other.

GLOSSARY

Address Space: A range of addresses, normally contiguous and with each address defining a single byte.

Baseband: Where the transmission medium is dedicated to a single channel of communication and thus the carrier frequency of the signal on the medium is also the fundamental frequency transmitted.

Bus: That which provides the electrical communication between components of a system. Devices connected to a bus must conform to signal standards and protocols. A bus can also provide electrical power and mechanical support.

Bus Cycle: A nondivisible data-transfer operation in the bus.

Bus Master: The device on the bus that initiates the bus cycle. When there is the possibility of more than one bus master, the potential masters of the bus must arbitrate for which will be the single master of the next bus cycle.

Bus Transaction: A bus cycle or a message.

CPU: Central processing unit, the section of a computer that fetches, decodes, and executes instructions.

Executive: Operating system software that provides the basic arbitration of and access to the computer hardware resources, the CPU, and the I/O devices.

False: Boolean state represented by a voltage in logic systems.

IEC: International Electrotechnical Committee, an international standardization body.

IEEE: Institute of Electrical and Electronics Engineers, an international professional society based in New York.

I/O: Input/output.

Message: A nondivisible data-transfer operation in the bus or network.

Multicast: The sending of information to a group of devices defined by a common group address.

Negative Logic: Logic implemented with the true state being represented by the less positive voltage.

Parallel: Information for one logical bus transaction is provided at one time on a separate electrical connection for each bit of information.

Positive Logic: Logic implemented with the true state being represented by the more positive voltage.

Real Time: The time in the world to which the computer is connected.

Real-Time System: A system that is driven by and responds in a predictable manner to external events.

RISC: Reduced instruction set computers, a central processor design which restricts the instruction set and the addressing modes to the simpler ones to allow the implementation of the CPU on a single chip with the emphasis on instruction execution in about one clock cycle.

Serial: Information for one logical bus transaction is provided in time sequence on one or a few electrical connections.

True: Boolean state represented by a voltage in logic systems.

List of Works Cited

Clout, P. N. (1982), Los Alamos publication LA-UR 92-2718.

ISO 7498-1984, Information processing systems, Open systems interconnection, Basic reference model.

GESPAC S.A. (1985), Specification Manual for the G-64 Bus, 18 Chemin des Aulx, 1228 Geneva, Switzerland.

Titus, J. (1985), *EDN* April 11, 61–72.

Twenty-six IEEE Standards listed in Table 1.

Further Reading

Andrews, W. (1989), *Comput. Des.* **28** (21), 78–96.

Borrill, P. L. (1989), *IEEE Spectrum* **40** (7), 34–37.

Brand, J. H., de Graaf, L. (1989), *EDN* **34** (23), 255–258.

Dawson, W. K., Dobinson, R. W. (1987), *Comput. Stand. Interfaces* **6**, 403–425.

Elsmore, T., Shialds, W. R. (1984), *IEEE Micro* **4** (4), 72–80.

Simpson, D. (1989), *Syst. Integration* **22** (11), 85–90.

CAD

See COMPUTER-AIDED DESIGN AND ENGINEERING

CAE

See COMPUTER-AIDED DESIGN AND ENGINEERING

CALIBRATION AND MAINTENANCE OF TEST AND MEASURING EQUIPMENT

JERRY L. HAYES, *Hayes Technology, Ramona, California, U.S.A.*

	Introduction	1
1.	**Test and Measuring Equipment Calibration**	2
1.1	Objectives of Calibration	2
1.1.1	What is Calibration?	2
1.1.2	Why Calibrate?	2
1.1.3	When to Calibrate	3
1.1.4	Where to Calibrate	3
1.1.5	How to Calibrate	4
1.1.6	Risks in Testing/Calibration	4
1.1.7	Factors Influencing Testing and Calibration Integrity	7
1.1.7.1	Measurement Uncertainty versus Testing/Calibration Specifications	7
1.1.7.2	Testing/Calibration Specifications	8
1.1.7.3	Other Factors Influencing Testing/Calibration Risks	8
1.1.7.4	Interrelating the Seven Risk Factors	10
1.2	Designs for Calibration	11
1.2.1	Determining Calibration Requirements	11
1.2.2	Selecting Calibration Equipment	11
1.2.3	Establishing Calibration Traceability	12
1.2.4	Designing Calibration Procedures	13
1.2.5	Defining Calibration Servicing Intervals	14
2.	**Test and Measuring Equipment Maintenance**	15
2.1	Objectives of Maintenance	15
2.2	Designs for Maintenance	15
2.2.1	Defining Maintenance Requirements	15
2.2.2	Selection of Maintenance Equipment	15
2.2.3	Designing Maintenance Procedures	16
2.2.4	Defining Maintenance Servicing Intervals	16
3.	**Summary**	16
	Acknowledgments	17
	Glossary	17
	Works Cited	17
	Further Reading	18

INTRODUCTION

Calibration and maintenance of measuring devices have been practiced since measuring devices were first developed. The Egyptians used master length standards based on the length of the Pharaoh's right foot during the construction of the pyramids, with a penalty of death to whoever failed to recalibrate their linear measuring instruments on schedule. Today's calibration and maintenance systems are designed to encompass all the phenomena encountered in modern science and even to anticipate measurements not yet reduced to practical application. Standards and measuring instruments have grown in complexity along with the world's technology. While the current penalties for failure to calibrate them on time are more likely to end in litigation than death, there are still occasions where the latter is a possible consequence of using uncalibrated critical hardware.

The topic of calibration and maintenance of standards and measuring equipment is broad in scope. It covers all the measurable phenomena encountered in science. Test and measuring equipment and the standards to which they are referenced are active in the life cycle of all scientific efforts. It is the collection of equipment that generates and detects the parameters of processes and devices to determine their attributes. It performs research

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

measurements, confirms designs, makes production and process monitoring measurements, executes product test measurements and inspections, performs calibration measurements, injects signals for troubleshooting, and makes maintenance measurements. (Because the term will be used often in this article, the acronym TME will be used also to denote this composite population of test and measuring equipment.)

Typical TME populations encompass spectrometers, standard gas samples, standard reference materials, digital multimeters, ac voltage sources, Zener voltage standards, spectrum analyzers, signal generators, pulse standards, coordinate measuring machines, gage blocks, interferometers, force machines, load cells, mass standards, accelerometers, shake tables, acoustic reciprocity standards, frequency counters, maser frequency standards, rf power meters, microwave sources, power calorimeters, fiber-optic reflectometers, light sources, thermal black bodies, platinum resistance thermometers, freezing points of metals, etc.

This article has not been written for the metrologist or calibration expert; it is aimed at the practicing physicist, engineer, scientist, or technologist. These are individuals who may have had extensive contact with measurement applications, but minimal involvement with TME maintenance and calibration. The article is intended to provide a straightforward, practical reference on the concepts and applications of TME calibration and maintenance. However, it also speaks to the subject of tests and measurements, since calibration is, itself, a test and measurement process. Because of this, the term "testing/calibration" will be used at times to remind the reader when the two processes are essentially inseparable. For a quick overview, the early stages of the article provide the what, why, when, where, and how of the calibration and maintenance process, with more detail appearing in later sections.

The adequacy of decisions and designs made from measurement data is crucial to the success of experiments and to the quality and reliability of products and systems. Both calibration and maintenance are essential elements of today's technology. Calibration is critical to the control of uncertainty growth of TME. Lack of such control leads to wrong measurements, wrong decisions, wrong experiments, and wrong designs. Maintenance is critical to prevent operational failures and the danger, cost, and disruption that failure causes.

1. TEST AND MEASURING EQUIPMENT CALIBRATION

1.1 Objectives of Calibration

1.1.1 What is Calibration? There are a variety of definitions of calibration. In its simplest form it is the comparison of two measurement devices or systems, one of known uncertainty and one of unknown uncertainty, to estimate the correct value of the unknown and to determine its uncertainty. The initial calibration of TME is usually performed when and where it is manufactured. This involves scribing scales, adjusting circuits, setting digital indicators, and estimating characteristic uncertainties and long-term stabilities. This new population of TME then is used in lifetime applications where recalibration is necessary to assure that it continues to operate within its performance specifications. Recalibration is performed to detect and correct any insidious out-of-specification conditions that may exist through performance decay. These calibrations typically provide a good/bad or go/no-go test of the parameters of the TME. If the measuring equipment is found to perform to within assigned specifications (a good or "go" condition), it passes calibration (sometimes adjusted or corrected, sometimes not, depending on local calibration practices) and is returned to service. If it fails calibration (a bad or no-go condition), it is repaired and/or corrected (adjusted) to restore it to within its performance specifications.

1.1.2 Why Calibrate? The initial calibration of TME provides a basis for the performance claims made when it is sold to the user. The parameters generated or measured by the devices (length, flow rate, voltage, intensity, power, etc.) are referenced to standards of measurement that stem from a common source. If they are not, there is no reliable basis for their use. Subsequent measurements made without such references would bear no definable relationship to those made with them.

Recalibration is essential to maintain control of the measurement processes made by the TME. It assures a continuous common reference and retains the performance benefits gained by the initial calibration. With rare exception, the uncertainty concerning the values of TME attributes increases with usage and time. The gradual growth in uncertainty beyond designated limits typically goes undetected by users. Recalibration is essential to assure reference to a common measurement base, such as that afforded by an unbroken chain of calibrations extending from the TME through institutional, company, and national standards to international standards at the Bureau International des Poids et Mesures (BIPM), located on the outskirts of Paris, France. For simplicity and want of a better term, this unbroken chain of calibrations has become known as traceability. Traceability plus calibration is essential to measurement control and international commerce. It is specified in many national and international procurement and quality assurance specifications as a necessary element of any research, design, and production activity and especially for final test, alignment, and calibration of products and instruments prior to delivery. Traceable calibrations also assure that measurements made to specified uncertainties at one part of a process or location can be reproduced and depended upon at another part of the process or location.

The validity of research measurements depends upon accurate estimates of the phenomena being studied and requires sound knowledge of the uncertainties involved. Calibration of the TME used in research, as in other applications, defines its uncertainty and controls the growth of that uncertainty to known limits. It aids the researcher in assuring that research results reflect actual changes in the phenomena and are not due to inaccurate estimates of measurement uncertainties. The need for more reliable and traceable measurements and calibrations has grown as experiments, designs, and processes have become more precise and demanding.

1.1.3 When to Calibrate. One of the most significant decisions regarding calibration is when to do it. If done too often it is a waste of time and money—remember, the TME is frequently found to be within specification. Yet, if done too seldom, it is a waste of experiment or product—the measurement uncertainties cause wrong decisions based on misleading measurement results. The objective is to recalibrate at the optimum interval to get a balance between the cost to calibrate and the cost not to calibrate. Unfortunately, there is currently too little known about how to weigh these factors accurately enough to use cost as the criterion for determining time intervals between calibration.

Present objectives focus on setting intervals based on a target of a certain percentage of TME expected to be found within specification upon recalibration. The larger this percentage, the smaller the chance that wrong measurements are being made by the entire population of TME. While some users seek to keep the percentage above 95% to assure tight measurement quality controls, others allow as little as 70% within specification at the end of the calibration interval. The percentage is a quality management decision usually reflecting some overall quality policy, assuming the significance of the target is understood by the decision makers. Usually, the percentage target is selected by default and 85%–90% within specification is accepted as a comfortable figure. Better ways to select these targets are being explored.

The process of determining recalibration intervals is a difficult mathematical and statistical process, requiring accurate and sufficient data taken during the calibration process. How to treat this data and select appropriate calibration intervals is discussed in greater depth later in this article and in the references (Castrup, 1989a).

1.1.4 Where to Calibrate. Calibrations have been performed traditionally in the confines of a carefully controlled laboratory environment. While this practice is still true of primary or reference standards laboratories, calibration laboratories are involved increasingly in doing calibrations in place at the site where TME is used. On-site calibrations have been made possible by better environmental controls at use points, better temperature compensation circuitry or software being used in newer TME, and automation of calibration systems. With on-site calibration,

1. the user retains custody of the equipment with minimum down time and operational disruptions;
2. the TME suffers less in-transit stress;
3. simpler, less expensive calibrations are pos-

sible because they can be tailored to the user's express needs at the point of use.

However, there are still many instances where in-place calibration is unsuitable because of

1. environmental considerations (the calibration equipment is too environmentally sensitive),
2. size considerations (the calibration equipment is too bulky), or
3. cost considerations (more calibration systems may be needed to handle the dispersed workload.)

Development continues on smaller, less sensitive, computer-aided, automated calibration systems that will promote on-site calibration. The computer has been an especially valuable tool in the control of calibration systems and in the collection and processing of measurement and calibration data.

1.1.5 How to Calibrate. High-quality, cost-efficient calibrations depend on calibration methods that achieve a balance between a brief calibration process that would check only a few points of an instrument over a limited spectrum versus an elaborate calibration that would check hundreds of points over a wide spectrum of parameter values.

Calibration procedures are the documents most commonly used to accomplish the "how" of calibration. Enough points on each range need to be checked at properly selected frequencies to confirm that the equipment is operating within performance specifications. Wherever possible, the TME should be calibrated in terms of how it is used. If only a limited portion of a range is used, calibration should be limited to only those attributes. Furthermore, all attributes of the TME may not need calibration or examination at each servicing interval. Proper examination of calibration and maintenance data can reveal where the preponderance of out-of-specification incidents occur. Calibration procedures and servicing interval systems can be tailored to accommodate these situations at significant savings.

1.1.6 Risks in Testing/Calibration. Using TME that is operating in an out-of-specification state invites serious consequences. Such usage increases the risk of making wrong decisions, such as whether experimental results derive from real phenomena or are an artifact of measurement errors or whether to accept or reject products being tested. The whole purpose of selecting adequate TME for a given task is based on minimizing the risk of making wrong decisions from measurement data (Castrup, 1980). The more uncertain the measured value, the greater the risk of wrong decisions. Wrong decisions cost time, money, lives, and reputations.

As an example, the ball component of a ball bearing is to be measured to confirm that it was manufactured to specifications. The specifications require the ball to be round to within a given plus or minus value—in this case a microinch. It is determined that a special dimensional measuring machine will be needed to test the ball. The question is raised of what uncertainty should be specified for the measuring machine. This raises yet another question as to how much risk, how much uncertainty, the process can endure. The least risky approach would be to select a machine with minimum uncertainty—say, to plus or minus one tenth of a microinch (10 to 1 specification to uncertainty ratio). Unfortunately, that calls for an extremely expensive machine. For half the price, a machine with twice the uncertainty, or 0.2 microinch (5 to 1 ratio) can be obtained. Which machine should be bought? Will the higher risk of using the cheaper machine create offsetting production costs through wrongly rejected bearing parts because of increased measurement uncertainties? Will wrongly accepted bearing parts lead to in-use failures and resulting warranty or litigation costs that far exceed the savings incurred by using the cheaper machine? Figure 1 shows the relationships involved.

The "Center Spec" represents the desired diameter of the ball. The bell-shaped curve straddling the "Center Spec" represents the production process distribution of balls being manufactured, where most of the balls being produced have sizes that are expected to cluster near the center spec with fewer and fewer expected to be found beyond the "Perf Spec." This is known in statistics as a "normal" distribution. The "+Perf Spec" represents the performance specification upper limit, or +1 microinch greater than the "Center Spec." The "−Perf Spec" represents the lower limit of −1.0 microinch less than the "Center Spec." The upper and lower "Test Spec" typify a test procedure where the test specifications have

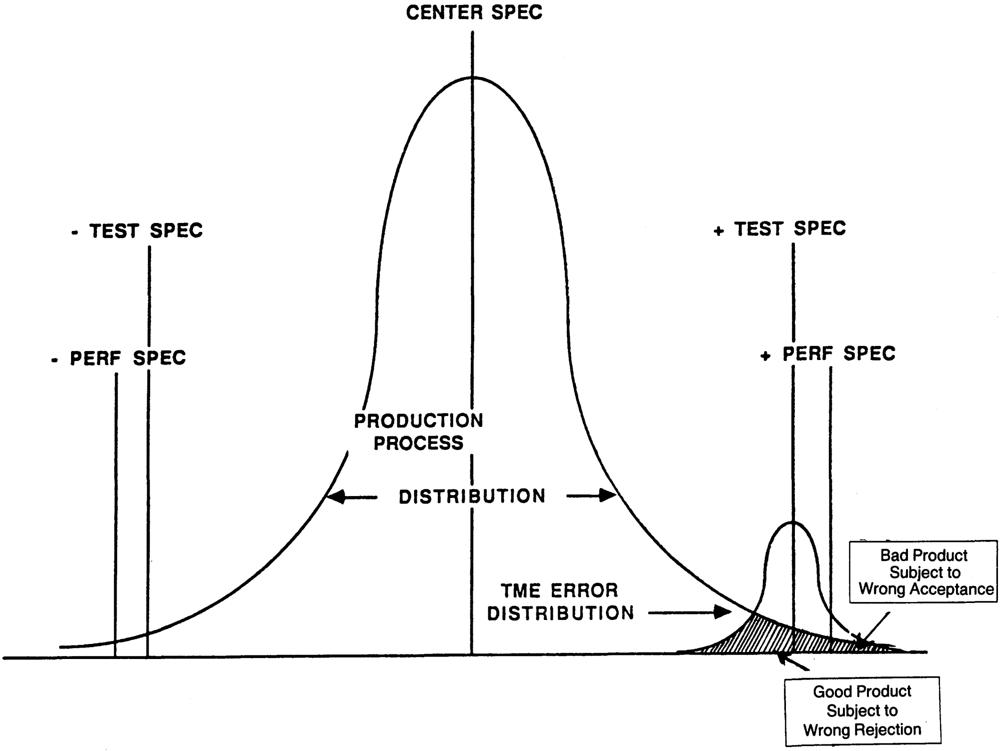

FIG. 1. Relationship of production yield, specifications, and testing errors.

been offset slightly from the performance specifications to compensate for unknown testing errors. Setting the test specifications inside the performance specifications reduces the risk of accepting nonconforming, or bad, measured values, but increases the risk of rejecting good ones. Setting the test specifications outside the performance specifications reduces the risk of rejecting conforming, or good, values, but increases the risk of accepting bad ones. The curve at the right-hand side of the figure, which represents the "TME Error (uncertainty) Distribution," has been made small to depict the size of TME uncertainty in relation to the "Production Process Distribution" (in the case of the example, it is either one tenth or two tenths the size, respectively). When the spread of the "TME Error Distribution" grows relative to the spread of the "Production Process Distribution," the risk of making wrong testing decisions grows. The words in the boxes point out the areas where those decisions take place and are most critical. As can be seen, the larger the "TME Error Distribution," the greater the quantity of wrong test decisions. As the "TME Error Distribution" increases, "Good product subject to wrong rejection" increases at a far faster rate than "Bad product subject to wrong acceptance." Thus, in the ball bearing example case, were the TME uncertainty allowed to grow to nearly the size of the "Production Process Distribution," the "Good product subject to wrong rejection" would be a sizable portion of the entire production of balls and would run a high risk of wrong rejection by the measuring machine, while at the same time, many of the balls beyond the "Perf Spec" would be wrongly accepted. Controlling the proper size of the "TME Error Distribution" in relation to the "Production Process Distribution" is one of the principal elements in controlling the risks of measurement, testing, and calibration (Hayes, 1955). Periodic calibration of the selected machine would assure that its uncertainty and the amount of its wrong test data was controlled.

Thus, in the example given earlier, the decision of which measuring machine to buy depends upon the impact that an uncertainty of 0.2 microinch versus 0.1 microinch would have on wrong testing decisions. Later in this article, material will be presented that quantifies these relationships, which are known as uncertainty ratios or accuracy ratios, i.e., the ratio between the size of the product test limit and the size of the TME uncertainty (Ferling, 1984).

Before leaving this subject, the reader should be aware of the pitfalls of misusing or miscalculating the uncertainties of testing/calibration processes. As was mentioned previously, the correct estimate of the magnitude of the uncertainty of a measurement is of great importance to the quality of the test being made. Many unsuspecting people select TME for a measuring process without a real awareness of the uncertainty being introduced. They observe the manufacturer's published specifications and presume wrongly that the advertised "accuracy" is transferred unchanged to the measurement process.

The terms "uncertainty," "accuracy," and "error" tend to be used interchangeably. Error is the reciprocal of accuracy and indicates a specific deviation from a "true" value with no indication of the degree of confidence to which that figure may be known. Uncertainty is a more meaningful term that has an error component and a "sureness" component. It depicts the combination of parameter generation and measurement errors in a test or calibration process and the degree of "sureness" in estimating the magnitude of each of those errors.

A properly stated uncertainty of measurement includes both components—the estimated error component (the ±0.1 microinch in the foregoing example) and another component which states the probability that the error component has been correctly defined (expressed as one or more standard deviations). Thus, a complete uncertainty statement would read, "When measuring one microinch, there is an uncertainty of ±0.1 microinch at three standard deviations." Thus, the statement expresses an error and a "degree of sureness" as to how well that error is known. Three standard deviations is roughly equivalent to providing an assurance that the error will be within the limits stated at the time of measurement with 0.997 probability. Had the estimate been made to, say, one standard deviation, the error would be within the limits given at the time of measurement with 0.67 probability. (An even more accurate and useful statement would also include the rate of uncertainty growth with time, or the time period over which the uncertainty statement applies, such as one week, six months, one year, etc.)

To determine the overall uncertainty of a testing/calibration process requires that all of the error components and their companion estimation components be combined in a consistent fashion. Some controversy currently exists as to the precise way these errors should be combined. Some experts say to add the errors algebraically without regard for the estimation uncertainties in order to get a "worst case" combination. This clearly makes sense where no estimation values are available or where the estimator is very insecure about the error estimate. Yet, other experts argue that it is best to sum the errors in quadrature [by the root sum of squares (RSS) method] and present a statement for the combined errors for a specified number of standard deviations. (See ERROR ANALYSIS.)

The balance of opinion has been shifting most recently to use of the RSS method. All things being equal, the RSS method gives smaller estimates of error than does algebraic summing. For example, given a measurement system comprised of, say, three equal-error components, an algebraic summation would yield an uncertainty estimate 1.73 times larger than the RSS combination would yield. This number would grow to 2.64 were seven such error components combined. While RSS combining could mean higher risks, it is considered by many to be the more realistic method. A second benefit of applying the RSS method is that it provides a consistent statistical basis for estimating the confidence level associated with stated uncertainty limits.

Table 1 provides an example from the calibration of two rf voltage standards. At 10 MHz, the RSS method results in an intercomparison uncertainty of 53 parts per million at one standard deviation for three years. The algebraic method yields an uncertainty of 116 parts per million at this frequency, the number of standard deviations is undefined, and the estimated limit of time on the estimate is three years.

More research is being done to determine the most mathematically correct methods of combining uncertainty estimates. National and international standards bodies [American So-

Table 1. Uncertainties in voltage source intercomparisons by RSS and algebraic combination methods.

Error source	Error estimates in parts per million (all random errors @1 s.d.) @10 MHz	@50 MHz
ac-dc difference	8	11
Influence of noise	25	185
Skin effect	24	65
Connector alignment	6	38
Comparator uncertainty	18	95
Long-term instability (3 yr)	35	150
RSS combination	53	269
Algebraic combination	116	544

ciety of Testing Materials (ASTM), Institute of Electrical and Electronic Engineers (IEEE), and the Bureau International des Poids et Mesures (BIPM)] are seeking consensus methodology to resolve current disagreements. Whatever method is used, the important thing is that the error sources be listed, the uncertainty of estimation stated, and the method of combination documented. Then others will be able to interpret correctly the uncertainty statement for replication and for their own applications.

1.1.7 Factors Influencing Testing and Calibration Integrity

1.1.7.1 Measurement Uncertainty versus Testing/Calibration Specifications. Accuracy or uncertainty ratios are a representation of the relationship between the measurement uncertainty of the testing/calibration process and the performance, test, or calibration specifications of the device under test or calibration. For simplification, call this the ratio between the measurement uncertainty and the test specification. What happens when this ratio changes? This is equivalent to a situation in Fig. 1 where the "TME Error Distribution" is expanded to the size of the "Production Process Distribution" or contracted to being barely visible. Figure 2 represents the relationship of these values and is based on the following typical set of conditions:

1. the device under test came from a production process distribution that is normal (Gaussian);

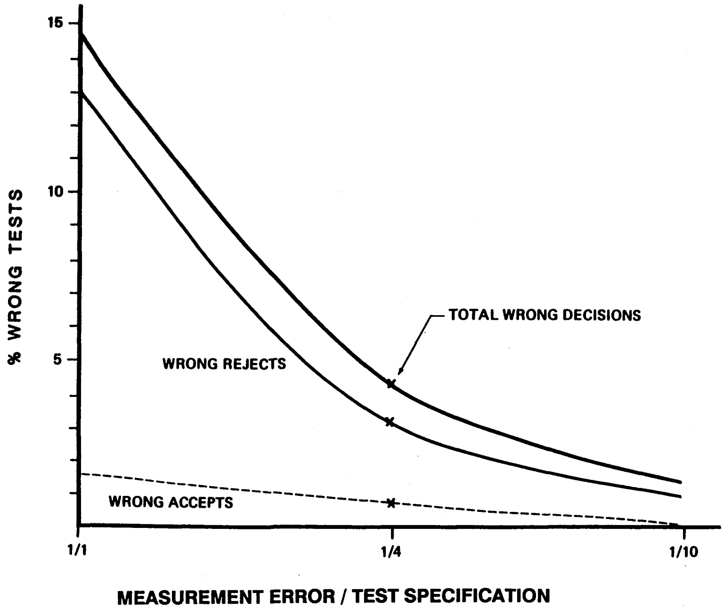

FIG. 2. Effect of varying measurement errors.

2. the performance specifications have been set at 2σ confidence limits of the production process distribution;
3. the measurement uncertainty has a normal (Gaussian) distribution with limits at two standard deviations or at 2σ confidence.

Figure 2 shows the effect of increasing or decreasing the measurement error in relation to the test or performance specification. At the point where the error is as great as the test specification (1/1 ratio), nearly 15% of the tests conducted result in wrong acceptances or rejections. At a ratio of 1/4, the percentage of wrong tests has been reduced dramatically to less than 5%. Below this point, the rate of reduction in the percentage of wrong tests slows as the ratio approaches 1/10 or less. These figures typify the impact of diminishing measurement error on test or calibration risks. At ratios greater than 1/4 (1/3, 1/2, 1/1.5), the percentage of wrong tests increases sharply. If one is required to make measurements under such conditions and these higher risks cannot be tolerated, compensation schemes should be used, as will be discussed later.

Another underlying assumption here is that the measurement error represents the uncertainty that characterized the measurement process at the time the measurement was made, not at some earlier time when calibration was performed. Thus, the uncertainty statement should include estimates of error growth over the time between calibrations. Stated another way, the longer the TME is into its calibration interval, the more uncertain its measurement.

1.1.7.2 Testing/Calibration Specifications. The control of testing/calibration risk also depends upon the relationship between performance specifications and the test or calibration specifications used to confirm performance compliance. What happens as the relationship of specifications is changed? Refer again to Fig. 1. The small bell-shaped curve to the right of the figure represents the total combined uncertainty of the TME and test system used to test a typical product. Note that the "Test Spec" is located inside the "Perf Spec" limit. The location of the "Test Spec" in relation to the "Perf Spec" has a significant impact on the number of wrong test decisions that a given test process will produce.

These wrong test decisions are made up of wrong acceptances (sometimes called buyer's risk) and wrong rejections (sometimes called seller's risk). Wrong acceptances hurt the buyer and damage the seller's reputation; wrong rejections hurt the seller and raise costs to the buyer. If the TME uncertainty is large enough to create unwanted testing risks, the "Test Spec" can be moved inside or outside the performance specification to reduce the unwanted risk. By looking at the shaded areas in Fig. 1, one can see that the further the "Test Spec" is inside the "Perf Spec," the greater the amount of "Good product subject to wrong rejection" and the smaller the amount of "Bad product subject to wrong acceptance." In this instance, the seller is taking a much larger risk than the buyer to compensate for the TME uncertainty. Were the seller willing to pass the testing risk on to the buyer at some risk to his reputation, the seller would locate the "Test Spec" outside the "Perf Spec." Refer now to Fig. 3. It shows the effect of varying test specifications in relation to performance specifications for a given TME error distribution. Figure 3 is based on the same conditions as Fig. 2 and reflects the percentage of wrong tests that result when the location of the test specification is moved.

Note that the "Wrong Rejects" increase to 10% as the "Test Spec" is located further and further inside of the "Perf Spec," while the "Wrong Accepts" diminish to below 1%. As the "Test Spec" is moved to the outside of the "Perf Spec," wrong rejections go to zero while wrong acceptances increase to a peak of about 3%. Thus, one method available to control testing/calibration risk is to reposition test and calibration specifications, especially where TME uncertainties are too large and create unwanted risks (Eagle, 1954; Grubbs and Coon, 1954).

1.1.7.3 Other Factors Influencing Testing/Calibration Risks. While uncertainty ratios and tolerance positioning strongly affect testing/calibration risks, other factors have significant influence as well. They are as follows.

1. *The quality of units yielded from the production process.* The production quality variable has one of the strongest influences on testing/calibration risk. This variable is reflected in Fig. 1 as the "Production Process Distribution." The more units that are produced which lie within the performance specifications, the higher the quality

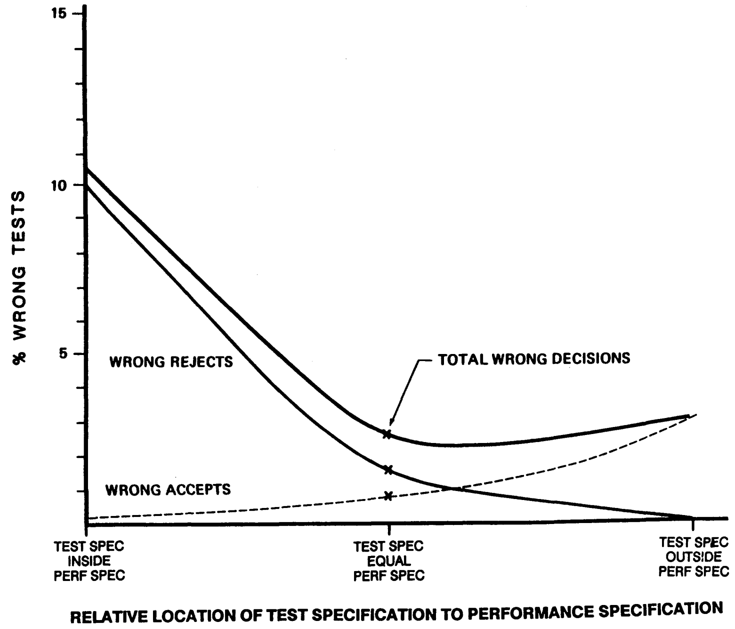

FIG. 3. Effect of varying test specifications.

and the lower the measurement risk. This is true whether the product is a ball bearing or an item of TME. This quality indicator is usually expressed in terms of statistical confidence limits. The higher the confidence limit is, the greater the amount of product within the limit and the higher the quality. A 3σ limit would indicate that 99.7% of yielded product was being produced within the stated performance specifications; a 2σ limit, 95.5%; a 1σ limit, 67%.

2. *Length of test intervals.* The longer the time between tests on a population of products or devices, the greater the likelihood of making wrong test decisions. For example, the within-specification probability for most simple electronic devices usually decays exponentially with time. Thus, devices in the population move toward and beyond their performance specifications with time. The majority of wrong test decisions are made on devices operating near their performance specifications. Thus, the risk of making wrong test decisions increases with the elapse of time between tests due to performance decay of the devices under test.
3. *Length of calibration intervals.* The longer the time between calibrations of TME, the larger the uncertainty associated with its measurements. The performance of the TME decays in the same manner as the devices under test. Thus, TME uncertainty grows with time as does the risk of making wrong measurement decisions.
4. *Number of independent tests or calibrations performed on the same device or TME.* It is not uncommon for devices to be tested independently at several points in a production process, then tested further following stress induced to simulate usage and environmental conditions. The more tests that are performed on devices or TME, the greater their exposure to wrong test decisions. Marginally bad devices are eliminated in successive tests, which reduces the opportunities to make wrong acceptances, while more and more marginally good devices are exposed to wrong rejections. Thus, increasing the number of independent tests performed increases the probability of wrong rejection and decreases the probability of wrong acceptance (provided the induced stresses do not create impending failures—yet another risk). This impact also applies to the number of test points and parameters that are tested or calibrated. Testing/calibration decisions regarding device or TME acceptability are normally based on a policy that if one test point is beyond specifications, the entire device or TME is considered to be beyond specifications. Thus, the impact of testing the device at multiple test points and mul-

tiple parameters is similar to that of performing repeated tests. All three decrease wrong acceptances and increase wrong rejections. (Ferling, 1987).

1.1.7.4 Interrelating the Seven Risk Factors. Because of the complex mathematics and considerable computational facilities involved, it is very difficult to give exact figures for the mutual interaction of the seven factors affecting test and calibration integrity (Hayes, 1988). However, Fig. 4 provides a generalized representation, showing the combined impact of the seven interrelationship factors. They are presented as follows.

1. The *y* axis depicts the typical spread of wrong test decisions that result from the interactions, ranging from less than 1% up to greater than 10%. It identifies separately the wrong acceptances and wrong rejections that make up total wrong decisions, since the two elements react differently.
2. The *x* axis shows four factors with their typical ranges in parentheses as follows:
   (a) "% TME In-Tolerance": the amount of TME in a testing population found in-tolerance at the end of their calibration intervals (i.e., when submitted for calibration).
   (b) "Test Intervals": the length of time between scheduled operational tests of devices during their life cycle.
   (c) "Calibration Intervals": the length of time between scheduled calibration of the TME population during its life cycle.
   (d) "Test Accuracy Ratios": the ratios between the test or performance specification of a device under test or TME and the uncertainty of the TME or calibration system used to test or calibrate it.
3. Arrows on the curves include three additional factors as follows.
   (a) "More Tests": the number of independent tests and/or independent parameters involved in the test or calibration (multiple test points can be regarded as extrapolations of "more tests").
   (b) "Higher UUT Quality": the quality of units under test in a production process (usually expressed as a percentage expected or found to be within

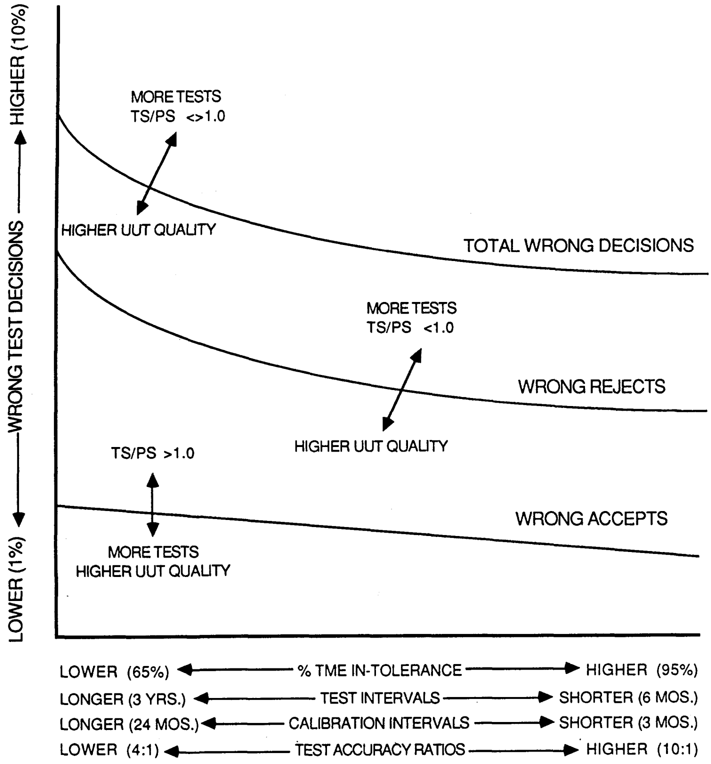

FIG. 4. Effect of test decisions on testing quality.

specified confidence or probability limits stated in σ's or standard deviations).

(c) "TS/PS": the ratio between the location of the test specification (TS) to the performance specification (PS) of a device under test or TME. At 1.0 the TS is located equal to the PS; at < 1.0, the TS is located inside the PS; at > 1.0, outside.

A careful review of these interrelationships will give the reader a better sense of the combined impact of the factors on testing/calibration risks and the need to control all seven to achieve targeted risk levels. Efforts have been under way since the early 1950s to quantify these relationships. Several combinations have been worked out (Castrup, 1989b), but a mathematical model that smoothly interrelates all seven factors has yet to be completed, programmed, and made operational on desktop or work-station-type computers.

1.2 Designs for Calibration

A properly designed calibration will fulfill the objective of retaining measurement control by detecting out-of-specification TME with known risks and restoring it to acceptable limits. Armed with an awareness of the factors that create risks in measurement and the what, why, when, where, and how of calibration, the design of successful calibrations can now be undertaken.

1.2.1 Determining Calibration Requirements. This first state of calibration design is aimed at determining if an item of TME actually requires periodic calibration. While almost all of it will, the following are situations where it will not.

1. It is not used to make quantitative measurements or provide known outputs.
2. It is obvious to the user when the TME fails to operate within performance specifications.
3. It makes measurements or provides outputs, but is monitored by other calibrated items of TME during use.

Once the TME requiring calibration has been identified, its performance specifications need to be pinned down. Assure that the parameters, ranges, and stated uncertainties are known and ambiguities eliminated. Then, find out which of the parameters and which of the ranges within the parameter are actually being used. (Minimizing the number of parameters and ranges calibrated reduces costs and the risk of wrongly rejecting adequate TME.) Next, establish the degree of uncertainty that can be tolerated based on the uncertainty ratios discussed earlier. If the allowable application uncertainties are larger than the TME performance specification, these can be used to expand the TME's calibration tolerances, which will, in turn, result in longer calibration intervals. Next, the environment in which the TME will be used is crucial for determining if any environmental compensation circuits need testing.

The final step involves selecting an adequate number of calibration points on each required range and parameter. The objective is to select enough points to assure that there are no hidden "holes" in the parameter where a discontinuity might exist. Meeting this objective calls for a balance between the minimum number of points to characterize a range and the time and added risk of not selecting more points. A minimum of two points is needed for the simplest of linear parameters, and from five to several hundred points where complex, nonlinear parameters are involved. A rule of thumb is that the greater the number of points selected, the less the risk of making wrong acceptances, but the greater the risk of making wrong rejections. A reasoned balance should be found for this important design decision (U.S. Navy Metrology Engineering Center, 1983).

1.2.2 Selecting Calibration Equipment. Calibration equipment that has sufficient parameters, range, and calibration point precision and sensitivity capabilities to match the defined TME requirements needs to be selected. Also, consideration should be given to the use of computer-aided calibration systems. These are helpful where the number of calibration points required create tedious measurement processes that often result in human error. Automated calibration systems are especially useful to process high quantities of TME using lower-skilled personnel and to collect and process large volumes of data.

To assure controlled testing risks, calibration equipment whose own uncertainties are significantly smaller than the TME performance specifications should be selected. For new equipment, the standard deviation of the uncertainty claim should be obtained from the manufacturer; otherwise, an estimate will be needed. Confidence limits of at least two standard deviations are suggested. For existing calibration equipment, its tolerance limits and the respective percentages in tolerance from calibration records should be used. Calibration equipment uncertainties smaller than one-fourth of the TME uncertainty have become the target for most calibration designs. (Larger TME uncertainties require that correction data, tolerance adjustments, repeated measurements, or shortened calibration intervals be used to reduce the influence of calibration errors. Figure 4 may be of help in suggesting avenues of escape from excessive risks caused by calibration equipment uncertainties.)

The site where the calibration is to be performed may have an influence on the equipment selected. Calibration equipment for in-place and on-site work requires more transportable, rugged, environmentally compensated, and easily operated systems than those found in the typical standards or calibration laboratory. Consideration should be given also to the difficulties that might be encountered in providing calibration support to the calibration equipment about to be selected. Often, acquisition of a new calibration capability creates a need for new standards capabilities to keep pace and to maintain measurement control through the traceability chain.

1.2.3 Establishing Calibration Traceability. Earlier, mention was made of the importance of maintaining common international measurement references. This is needed for measurements of all parameters for the worldwide exchange of goods, products, and technology. Transferring these common references in a controlled manner throughout the world to millions of individual measurements made by calibration equipment and TME is the objective of traceability. This goal requires a measurement support system of international, national, and grass-roots proportions.

Typical support hierarchies begin at the international level at the Bureau International des Poids et Mesures (BIPM) in France, where international standards of measurements are maintained. Then, their units of measurement are disseminated through reference standard artifact exchanges or comparisons with national standards laboratories. These laboratories, in turn, exchange standard artifacts with a larger number of government and commercial standards laboratories within each country. These standards laboratories then may exchange standard artifacts with an even larger number of locally operated calibration facilities, where calibration equipment transfers the units of measurement to their final destination during TME calibration.

The chain of traceability is completed at this state; it assures that measurements made on one side of the world mean, within uncertainty limits, the same thing as on the other side of the world. It also means that measurements made during research in one facility are relevant during design and production at another facility and when subsequently tested by a customer at yet another facility.

A typical example of a traceability chain would involve, for instance, a digital voltmeter as the item of TME requiring calibration support. It would be calibrated, typically by an automated calibration system comprised of a voltage source monitored by a more accurate digital multimeter with both source and meter controlled by a computer/controller, which also collects and processes calibration data, makes out-of-specification decisions, and establishes uncertainty estimates. (The multimeter would have built-in reference standards of voltage and resistance that the several internal circuits of the multimeter would use to update and correct the worst of the uncertainty drifts so as to extend calibration intervals.) During periodic calibrations, the multimeter's voltage references and complete voltage circuits would be calibrated against ac voltage transfer standards and dc voltage references, such as a bank of Zener diodes and ratio comparators with sensitive detectors. These Zener references would then be calibrated and certified by comparison with a bank of Josephson effect junctions, submerged in a special cryogenic bath, serving as voltage references and operating at microwave frequencies. (At cryogenic temperatures, these junctions resonate at frequencies known to a few parts in a billion of the international standards of time and frequency. They produce extremely stable and predictable voltage levels from which primary and national voltage standards are maintained. See CONSTANTS, FUNDAMENTAL.) As added protection, the Josephson junctions are intercompared internationally with BIPM and other national standards laboratories through a voltage measurement assurance program (MAP), which uses monitoring and data exchange schemes and intercomparison techniques. The MAP confirms measurement control of the references over long time intervals. Thus, traceable control and dissemination of the measurement unit of voltage are assured throughout the support hierarchy. The traceability and control of other measurement units are maintained through similar artifacts and processes.

To accomplish credible traceability, every effort must be made to transfer the measurement units down the chain with minimum increase in uncertainty. Because the artifacts normally are fixed in value, the use of uncertainty ratios between links in the chain is not usually applicable. Typically, at these stages in the process, corrections are included in certificates of calibration, and measurement uncertainties are combined as they pass along the links of traceability. Accuracy ratio protection is not required for these intercomparisons as long as the combined uncertainties are finally included at each progressive stage until they are included in the calibration equipment's estimate of uncertainty.

Somewhere along the chain and for convenience, the process usually changes from a one-to-one comparison of fixed-value standard artifacts to go/no-go (tolerance comparison) type calibrations. When this happens, alternative action must be taken to ensure that uncertainty growth is controlled and meaningful traceability is retained. Accuracy/uncertainty ratios have typically been the vehicle used to assist in traceability control for tolerance comparison calibrations.

Increasingly, there are processes where measurements are being linked to reference standard artifacts that embody physical constants. These have come to be called "intrinsic" standards; that is, they intrinsically embody the international value of the parameter they describe without direct reference to national or international laboratories. Some of the units of length, voltage, time, frequency, and temperature are represented by these intrinsic standards. Intrinsic standards do not follow the traditional traceability process of passing measurement units on through the use of standard artifacts. These standards are considered by some to be attractive because they allow the owning laboratory more freedom to maintain standards of measurement over longer periods of time. Their use also avoids the necessity of sending them to national or international standards laboratories for intercomparisons. However, they are expensive, require highly skilled personnel to operate, and require MAPs (data exchange schemes and intercomparison techniques) with national laboratories to assure that the equivalent of traceability is being achieved. MAPs can provide a significant quality benefit. They create a positive auditing function on all interacting laboratories (local, national, and international), which helps to detect human or equipment errors at any point in the traceability system that might go unnoticed otherwise.

Standard reference materials issued by some national laboratories are considered to be a form of intrinsic standard. These are specially prepared artifacts that embody a selected characteristic (hardness, spherical size, particle size, gas, chemical composition, etc.) and that disseminate traceable measurements from the national reference standards level to all other levels in the traceability hierarchy.

Sometimes, traceability is not possible because of a lack of capability at national or international levels. This problem appears most often where research efforts and discoveries have moved into new measurement parameters and ranges. Such inadequacies necessitate the temporary designation of an interim reference standard at a manufacturer's facility, a government laboratory, or a university. Other interim solutions include a round-robin intercomparison of an interim reference standard artifact that substitutes until a more rigorously devised national standard can be completed. The earliest possible detection of traceability gaps and the reporting of these to national standards agencies helps to minimize these traceability voids.

1.2.4 Designing Calibration Procedures. The design and development of calibration procedures is at the heart of competent and useful calibrations. Wrong calibration methods will readily negate even the best calibration equipment selection methods and calibration traceability schemes. The following are the elements of properly documented calibration procedures.

1. *Style.* The style of writing is intended to communicate with calibration personnel. It takes into account their level of education, their general experience in the field, and their experience in dealing with calibration processes.
2. *Identity of TME.* The item of TME is clearly identified as to model number, nomenclature, and any special configuration or production sequence.
3. *Identity of calibration equipment.* The required calibration equipment is identified, along with substitute items, hook-up diagrams, and details of the calibration configuration.
4. *Testing conditions.* Testing conditions are specified, including safety precautions, cable routing, switch positions, computer status, and interface settings.
5. *Test steps.* Sequential and sufficient test steps are listed for each measurement parameter that the calibration requirements effort has identified. Functional tests are also identified that confirm the operational adequacy of the TME without regard for tolerance tests.
6. *Pass-fail specifications.* The individual specifications and uncertainties of the TME are defined, including specific pass-fail criteria for each test point.
7. *Data collection.* The specific data to be taken have been identified (for on-site and future analysis), and data forms have been designed to collect it efficiently and accurately. Care must be exercised to take only enough data to execute the calibration, yet to take enough to allow for calibration interval assessment, parameter failure analysis, repair diagnostics, and part replacements, and management information such as time and material data.

There are a variety of sources for design guidance. TME and calibration equipment manufacturers often prepare calibration procedures for the instruments they produce. Government facilities prepare and distribute a large variety of procedures and guidance documents (U.S. Navy Metrology Engineering Center, 1984). The Government-Industrial Data Exchange Program (GIDEP) of the U.S. Navy at Corona, California, maintains a library of calibration procedures from many military and industrial sources. U.S. trade and technical societies, such as the American Society of Testing Materials (ASTM), American National Standards Institute (ANSI), Instrument Society of America (ISA), and the Institute of Electrical and Electronic Engineers (IEEE), have repositories of selected calibration procedures as do counterpart documentation standards institutions in most countries. Whatever the source of the procedure, it should meet the requirements discussed in this section to assure the integrity of the calibration process.

1.2.5 Defining Calibration Servicing Intervals. The design of a calibration support system would be incomplete without some established means of determining how often to calibrate TME and calibration equipment. A calibration performed only once establishes a one-time reference of uncertainty. Recalibration is necessary to detect uncertainty growth and to maintain the uncertainty within limits specified. A properly selected interval assures, to a known risk target, that TME is receiving recalibration at the proper time.

For a very few devices, the interval is based on stresses such as the number of uses or passes, or on wear attributable to abrasion or physical distortion. On most TME, the interval is based on elapsed time since last calibration, reflecting the tendency of its uncertainty to drift beyond established specifications as a function of time. Mathematical models have been developed that describe this. The exponential and gamma (Γ) models fit the behavior of most TME (other models include Weibull, log-normal, random walk, and drift, which relate to a significant minority of items of TME). These models establish the rate at which the TME moves out of specification and predict the time at which a given percentage of TME will be found to remain within specification. Stated another way, the models establish the functional relationship between time elapsed since the last calibration and the probability that an individual TME, selected randomly from its population, will be out of specification at the next calibration.

The design objective for a calibration interval determination system is to generate the proper calibration interval in the shortest possible time. The quality and quantity of data obtained by the calibration procedure have a major impact on the timeliness and accuracy of the designated intervals. Approximately 30 bits of data are required before reasonably confident intervals can be selected. To achieve that much data requires 30 separate calibrations of a single item of TME or, because that may take several years, the calibration of 30 items of TME of nearly identical configuration. Thus, for such practical reasons, data are acquired from TME by serial number, but compiled by identical or similar model number and manufacturer to accelerate the interval decision process. Because of this, most rigorous systems designate intervals at the manufacturer/model number level of identity (U.S. Navy Metrology Engineering Center, 1986; Castrup, 1989b).

Figure 4, presented earlier, reflects the effect that intervals have on percentage of wrong test decisions. Any interval setting system needs to select a target for the amount of out-of-specification TME that the process or facility can tolerate. This target will have a major impact on the length of calibration intervals.

Calibration intervals are listed in very few published-for-release documents. U.S. military services have extensive interval listings for their inventory of TME (U.S. Army Test, Measurement and Diagnostic Equipment Support Group, 1991; U.S. Air Force Aerospace Guidance and Metrology Center, 1991; U.S. Navy Metrology Engineering Center, 1991), as do industrial activities with military contracts.

2. TEST AND MEASURING EQUIPMENT MAINTENANCE

2.1 Objectives of Maintenance

TME maintenance focuses on technical activities intended to keep TME in satisfactory working condition and to avoid catastrophic failures. Unlike calibration, which is designed to control uncertainty growth beyond specified limits and to detect insidious failures unnoticeable by an operator, maintenance is designed to avoid accelerated wear-out and catastrophic operational failures.

The objective of maintenance is to assure that there will be no breakdowns of TME systems and that they can continue to operate safely. Typically, the cost to maintain is traded off against the cost to suffer TME breakdowns and increased safety risks. To quote some currently popular terminology, "If it ain't broke, fix it before it is." That succinctly represents the objective of maintenance.

While maintenance is an independent function, for convenience much of it is performed during calibration. Typically, maintenance intervals are longer than calibration intervals. Accordingly, many TME managers schedule maintenance to be done at every second or third scheduled calibration.

2.2 Designs for Maintenance

2.2.1 Defining Maintenance Requirements. The requirements for maintenance are usually defined in manufacturer manuals where specific activities are indicated to keep TME operable. Other requirements are acquired from data taken during calibrations and repairs, and from user complaints made to repair/maintenance personnel. These requirements attempt to define the circuits, parts, mechanisms, and devices whose failure could be avoided by detecting diminished capability, fluid loss, dirt and grease accumulation, environmental stresses, and wear. In addition, TME use should be reviewed to find out the experience level of users, their opinions regarding TME functional reliability, the environment in which it is used, and whether maintenance can be divided between the user and the maintenance facility.

2.2.2 Selection of Maintenance Equipment. Typically, maintenance performed during the calibration process uses much of the same equipment used for calibration. In addition, special facilities are needed for cleaning, lubricating, and stress testing for safety hazards and imminent failures.

Some items categorized as TME may require only maintenance and no calibration, yet require special tests or actuations to confirm operability of emergency circuits and actuator equipment for more complex tasks or non-uncertainty-related measurement capabilities, such as indications of presence or absence of signal, pressure, flow, etc.

As with calibration equipment, the site where maintenance is to be performed will have an influence on the equipment selected.

Much routine maintenance can be conducted on-site where the TME is located.

2.2.3 Designing Maintenance Procedures. Clearly written and logically sequenced procedures are essential to successful maintenance operations. Where these procedures are scheduled in conjunction with calibration operations, they should be integrated to follow the flow of the calibration process; however, many maintenance operations should precede calibration to assure functional adequacy of the TME before subjecting it to more time-consuming calibrations. Maintenance procedures should have the same characteristics previously outlined for calibration procedures (Sec. 1.2.4). The better and more clearly written these procedures, the less costly the continued maintenance operations will be. A small investment in well-prepared procedures will pay large dividends in the long run.

2.2.4 Defining Maintenance Servicing Intervals. One of the more difficult design problems is to develop a system that determines the most desirable time to perform maintenance. Done too frequently, maintenance is a waste of time; done too infrequently, it results in earlier failures with costly losses to both TME and the operations in which it is used. Many owners schedule TME maintenance at multiples of the TME calibration interval—a practical approach based on the fact that typical calibration intervals are shorter than maintenance intervals. As more knowledge accrues about calibration interval systems and calibration risk targets, basing maintenance intervals on calibration intervals may prove to be unwise. Calibration intervals are based on partial failures of TME, known more accurately as "soft" failures, whereas functional failures are known as "hard" failures. Therefore, preventive maintenance intervals lend themselves to somewhat different mathematical models for servicing interval estimates than do calibration intervals.

Ideally, maintenance interval analysis should stand alone and be based on mathematical and statistical correlation of historical functional failure data that focus on types of maintenance performed, time between maintenance, failed components/parts, and time between failures. From these data, mean-time-between-failure (MTBF) figures can be developed for each model-numbered TME. These figures reflect the reliability index of the TME and can be related to a MTBF target of the TME population. Thus, MTBF reliability targets can be established for which appropriate maintenance intervals can be designated.

3. SUMMARY

Control of tests and measurements is essential to the quality of development, design, manufacturing and industrial processes. Calibration of test and measurement equipment is a critical measurement control tool to assure that measurements are traceable to a common reference and to restrict test/measurement risks. Seven factors affect measurement risk. These factors include the initial quality of the product, the tolerances used in testing the product (as they relate to its design specifications), the uncertainty of the measurement equipment used in the tests (as it relates to testing tolerances), the number of tests performed, the time elapsed between tests, the percent of measurement equipment out of tolerance when used for tests, and the time elapsed between measurement equipment calibrations.

Well-designed test and calibration programs provide monitoring and control of the risk factors. Measurement assurance processes are a partner of calibration and help detect loss of measurement control. Maintenance programs assure that test and measurement equipment operates with minimum failures. They prevent untimely and costly operational disruptions. Collectively, these processes assure the cost-effective control of test and measurements throughout technology.

Future trends in calibration and maintenance include more intensive efforts to develop calibration systems that reduce measurement uncertainties and improve efficiencies. Measurement science research projects will continue to be of paramount importance to define, develop, and produce reference standards that keep pace with technology. Greater use of built-in and external computer systems will assure improved control and operation of calibration and maintenance systems. Special software and artificial intelligence technology will provide major assistance in configuring calibration systems, providing calibration and maintenance procedures, and analyzing and controlling measurement risks.

ACKNOWLEDGMENTS

The author wishes to express his deepest appreciation for the additions, suggestions, and encouragement of Dr. Arthur O. McCoubrey of the National Institute of Standards and Technology, Gaithersburg, Maryland, and Dr. Howard T. Castrup of Integrated Sciences Group, San Dimas, California.

GLOSSARY

Accuracy: See measurement uncertainty.

Accuracy Ratio: See measurement uncertainty ratio.

Attribute: The quantity, parameter, or characteristics of products or devices being tested or of TME being calibrated (voltage, pressure, viscosity, etc.).

Calibration: The process of measuring and testing attributes of TME to estimate the error or measurement uncertainty of the TME. For the majority of TME, the error is then compared to calibration specification limits; if found beyond specification, corrective repair and/or adjustment and recalibration is performed; if not found beyond specification, compensation or adjustment is optional. For fixed-value-type TME, such as reference standards, the attribute value and calibration system measurement uncertainty is determined and reported in a calibration certificate.

Calibrational Interval: The maximum allowable period of time for TME to remain in use before being calibrated. Sometimes referred to as "calibration period."

Error: The amount to which an attribute's measured value differs from its international reference or "true" value.

Maintenance: A process that cleans, adjusts, services, measures, diagnoses, and restores TME to such satisfactory working conditions as to avoid early wear-out and catastrophic operational failure.

Maintenance Interval: The maximum allowable period of time for TME to remain in use before being maintained. Sometimes referred to as "maintenance period."

Measurement Uncertainty: The magnitude of limits that bound a measured value and that contain its "true" value with a specified probability.

Measurement Uncertainty Ratio: The ratio of TME calibration tolerance limits (normally, TME uncertainty limits) to calibration equipment uncertainty limits, or, the ratio of a device's testing tolerance limits to the uncertainty limits of the TME testing it.

Test and Measurement Equipment (TME): Items of test, calibration, diagnostic, monitoring, simulating, and measuring equipment used in research, design, manufacturing, process monitoring, testing, and calibration operations.

Testing/Calibration Performance Limits or Tolerances: *Two sided:* Assigned values, symmetrically located on each side of the nominal value of a device or TME attribute, that represent the maximum acceptable limits for the value of the attribute. *Single sided*: An assigned value (upper or lower) that represents the maximum (or minimum) acceptable limit for the value of a device's attribute.

Testing: The process of measuring attributes of devices with TME and comparing the measured values with the devices' performance limits. If the device is found beyond performance limits, corrective repair and/or adjustment and retesting is performed.

Traceability: The ability to relate individual, working-level measurement results, through a contiguous sequence of measurements and intercomparisons, to nationally and internationally accepted measurement standards (e.g., reference standards, standard reference materials, intrinsic/quantum-mechanical standards, etc.).

List of Works Cited

Castrup, H.T. (1980), in: *Proceedings, NCSL 1980 Workshop and Symposium*, Boulder, CO: National Conference of Standards Laboratories.

Castrup, H. T. (1989a), in: *Proceedings, NCSL 1989 Workshop and Symposium, Denver, CO,* Boulder, CO: National Conference of Standards Laboratories.

Castrup, H. T. (1989b), *NCSL Recommended Practice*, RP-1, Boulder, CO: National Conference of Standards Laboratories.

Eagle, A. R. (1954), *Ind. Quality Control*, March, 10–15.

Ferling, J. A. (1984), "The Role of Accuracy Ratios in Test and Measurement Processes," presented at the 1984 Measurement Science Conference, Long Beach, CA (unpublished).

Ferling, J. A. (1987), "Calibration Intervals for Multi-Function Test Instruments, a Proposed Policy," presented at the 1987 Measurement Science Conference, Irvine, CA (unpublished).

Grubbs, F. E., Coon, H. J. (1954), *Ind. Quality Control*, March, 15–20.

Hayes, J. L. (1955), *Factors Affecting Measurement Reliability*, Technical Memorandum No. 63-106. Corona, CA: U. S. Naval Ordnance Laboratory.

Hayes, J. L. (1988), "Decisions that Influence Testing Policy," presented at the 1988 Measurement Science Conference, Long Beach, CA (unpublished).

U. S. Air Force Aerospace Guidance and Metrology Center (1991), *TMDE Intervals, Calibration and Repair*, Newark, OH (Technical Order 33K-1-100).

U. S. Army Test, Measurement and Diagnostic Equipment Support Group (1991), *Calibration and Repair Requirements for the Maintenance of Army Material*, Huntsville, AL (Technical Bulletin 43-180).

U. S. Navy Metrology Engineering Center (1983), *Technical Requirements for Calibration of Test and Monitoring Systems (TAMS)*, Corona, CA (NAVAIR 17-35TR-1).

U. S. Navy Metrology Engineering Center (1984), *Requirements for Preparation and Submission of Instrument Calibration Procedures*, Corona, CA (NAVAIR 17-35TR-4).

U. S. Navy Metrology Engineering Center (1986), *Technical Requirements for Calibration Interval Establishment for Test and Monitoring Systems (TAMS)*, Corona, CA (NAVAIR 17-35TR-5).

U. S. Navy Metrology Engineering Center (1991), *U. S. Navy Metrology Requirements List*, Corona, CA (NAVAIR 17-35-MTL-1, SPAWARS P4734-310-0001, NAVSEA OD 45845, USMC TI-4733-15/13).

Further Reading

Caldwell, D. H. Ferling, J. A. (1989), in: *Proceedings, NCSL 1989 Workshop and Symposium, Denver, CO,* Boulder, CO: National Conference of Standards Laboratories.

Castrup, H. T. (1980), in: *Proceedings, NCSL 1980 Workshop and Symposium,* Boulder, CO: National Conference of Standards Laboratories.

Castrup, H. T. (1989a), in: *Proceedings, NCSL 1989 Workshop and Symposium, Denver, CO,* Boulder, CO: National Conference of Standards Laboratories.

Castrup, H. T. (1989b), *NCSL Recommended Practice,* RP-1, Boulder, CO: National Conference of Standards Laboratories.

Eagle, A. R. (1954), *Ind. Quality Control*, March, 10–15.

Grubbs, F. E., Coon, H. J. (1954), *Ind. Quality Control*, March, 15–20.

Hayes, J. L. (1955), *Factors Affecting Measurement Reliability,* Technical Memorandum No. 63-106. Corona, CA: U. S. Naval Ordnance Laboratory.

Klaassen, K. B., van Peppen, J. C. L. (1986), *J. Phys. E* **19**, 577–578.

Kuskey, K. (1984), in: *Proceedings, 1984 Measurement Science Conference, Long Beach, CA.*

Larsen, J. (1988), in: *Proceedings, NCSL Workshop and Symposium, Washington, D.C.*, Boulder, CO: National Conference of Standards Laboratories.

Lorenzen, D. A., Amano, E. A., Robinson, K. (1990), *Training Information Directory*, Boulder, CO: National Conference of Standards Laboratories. See the "Text and Reference Material" (throughout each section).

Mann, N., Schafer, R., Singpurwalla, N. (1974), *Methods for Statistical Analysis of Reliability and Life Data*, New York: Wiley.

Marguglio, B. W. (Ed.) (1977), *Quality Systems in the Nuclear Industry (and in Other High-Technology Industries),* ASTM Special Technical Publication 616. Philadelphia: American Society for Testing Materials, Chaps. 6 and 14.

McCoubrey, A. O. (1983), "Measures and Measurement Systems," in: *Encyclopedia Americana*, Danbury, CT: Grolier.

McCoubrey, A. O. (Ed.) (1985), *Precision Measurement and Calibration: Electricity,* NBS Special Publication 705. Gaithersburg, MD: National Institute of Standards and Technology.

Rice, G. O. (1983), *NCSL Laboratory Manager's Guidebook*, Boulder, CO: National Conference of Standards Laboratories.

U. S. Navy Metrology Engineering Center (1983), *Technical Requirements for Calibration of Test and Monitoring Systems (TAMS),* Corona, CA (NAVAIR 17-35TR-1).

U. S. Navy Metrology Engineering Center (1983), *Guide for Establishment of A Local Calibration Program*, Corona, CA (NAVAIR 17-35HBK-3-1).

U. S. Navy Metrology Engineering Center (1984), *Requirements for Preparation and Submission of Instrument Calibration Procedures*, Corona, CA (NAVAIR 17-35TR-4).

U. S. Navy Metrology Engineering Center (1986), *Technical Requirements for Calibration Interval Establishment for Test and Monitoring Systems (TAMS*), Corona, CA (NAVAIR 17-35TR-5).

U. S. Navy Metrology Engineering Center (1991), *U. S. Navy Metrology Requirements List*, Corona, CA (NAVAIR 17-35-MTL-1, SPAWARS P4734-310-0001, NAVSEA OD 45845, USMC TI-4733-15/13).

Numerous authors, selected standards on test and calibration processes from the lists of the American Society of Testing Materials, American National Standards Institute, Institute of Electrical and Electronic Engineers, Instrument Society of America, British Standards Institute, and similar institutions in major countries.

CARBON MATERIALS

OTTO J. VOHLER, *formerly SIGRI GmbH, Meitingen, Germany*
FERDINAND VON STURM, *Research Center KFA, Jülich, Germany*
ERHARD WEGE, *formerly SIGRI GmbH, Meitingen, Germany*

	Introduction	21
1.	**Diamond**	22
1.1	Properties and Structure	22
1.2	Natural Diamond	23
1.3	Diamond Synthesis	23
1.4	Diamond Thin Films	24
1.5	Applications	24
2.	**Graphite**	24
2.1	Single-Crystal Graphite	25
2.2	Natural Graphite	26
2.2.1	Mining and Properties	26
2.2.2	Uses	26
2.3	Artificial Graphite and Carbon with Layered Structure	27
2.3.1	Structure and Pyrolysis	27
2.3.2	Carbon and Graphite Artifacts	28
2.3.2.1	Manufacture	28
2.3.2.2	Properties	31
2.3.2.3	Applications	34
2.4	Monogranular Carbon and Graphite Products	36
2.4.1	Carbon Fibers	36
2.4.2	Glassy Carbon	37
2.4.3	Pyrocarbon and Pyrographite	37
2.4.4	Graphite Intercalation Compounds	38
2.4.5	Graphite Foils	38
	Works Cited	39
	Further Reading	39

INTRODUCTION

Carbon, the sixth element of the periodic table, has an atomic mass of 12.011 15 ±0.000 05. It consists of 98.89% of the isotope ^{12}C, 1.11% of the isotope ^{13}C, and several unstable isotopes with nuclide masses between 9 and 16. The isotope ^{14}C, which is formed in the atmosphere by the reaction $^{14}\mathrm{N}(n,p)^{14}\mathrm{C}$, is used in archeology for age determination and for tracing of chemical compounds; the half-life of this β emitter is about 5570 years (Gmelin, 1967, 1968).

Carbon is the thirteenth most frequently occurring element. The earth's crust, including the lithosphere, hydrosphere, and atmosphere, contains about 2.9×10^{16} t of carbon, which equals 0.087 wt % or 0.152 at. %, mostly as compounds and only a small portion in elementary form. The carbon content of sea water, atmosphere, and biosphere is about 3.8×10^{13}, 3.8×10^{11}, and 2.8×10^{11} t, respectively.

Carbon exists in several allotropic modifications (Fig. 1). *Chaoite*, the most frequent member of the *carbyne* group, was found in small quantities together with graphite in the rocks of the Nördlinger Ries, Federal Republic of Germany, which were transformed by meteoritic impact (Goresey, 1968). Small quantities of other members of the carbyne group are formed together with graphite by condensing carbon vapor generated from laser-heated carbon on substrates, and also in meteorites. Carbyne carbon forms are also found mingled with graphite material in natural carbon deposits all over the world. Like diamonds, they are thermodynamically unstable under ordinary conditions, but have a kinetic barrier for transformation to graphite. A common feature of the numerous members of this group is the presence of triple bonds (*sp*). The crystals have trigonal symmetry. A few of these forms, like chaoite, for example, are very hard; others are as soft as graphite or even softer (Whittaker, 1979; Whittaker *et al.*, 1983).

Graphite is characterized by its layered lattice structure, but only a few natural graphites and hot-worked pyrographite samples are nearly perfect crystals. Most graphitic materials deviate more or less from the perfect crystal structure. The differentiation between

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

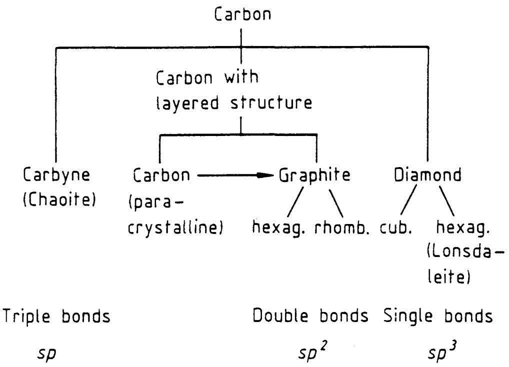

FIG. 1. Allotropic carbon modifications.

graphite and carbon, a term for less ordered species, e.g., metallurgical coke, carbon black, or charcoal, is only for the sake of convenience.

Diamonds are most frequently cubic. The hexagonal diamond (lonsdaleite) was first found as a product of diamond synthesis but meanwhile has also been discovered in meteorites. Graphite is the stable phase at standard conditions; diamond can be transformed into graphite by heating above 2000 K. At atmospheric pressure, graphite vaporizes without melting at ca. 3895–4020 K. By use of a laser heating technique, a graphite–liquid–vapor equilibrium pressure of 10.7 ± 0.2 MPa at a temperature of about 4000 K has been found (Haaland, 1976). The density of liquid carbon is about 1.37 g/cm^3. The graphite–diamond–liquid triple point amounts to ca. 4020 ± 50 K and $(11.0\text{–}14.0) \times 10^3$ MPa. The heat of sublimation depends on the vapor composition. It amounts to 711, 825, and 788 kJ/mol for monatomic, diatomic, and triatomic molecules, respectively. At 4000 K, about 80% of carbon vapor consists of C_3 molecules.

Fullerenes are distinct carbon clusters characterized by a cage structure. C_{60} (buckminsterfullerene) and C_{70} have been identified in soot and condensed carbon vapor. Other stable members of the fullerene group should be C_{20}, C_{24}, C_{28}, C_{32}, C_{36}, C_{50} (Kroto, 1988), C_{76}, C_{84}, C_{90}, and C_{94}.

The fullerenes are obviously important intermediates in the formation process of soot. The structure of C_{60}, buckminsterfullerene, is shown in Fig. 2.

The variations in bonding of the carbon atoms inaugurate a broad spectrum of carbon species, including conductors and insulators having extraordinarily high or low hardness and stiffness. Carbon materials accordingly offer a broad range of properties that are exploited in numerous applications.

1. DIAMOND

For many centuries, diamond has been appreciated as a valuable gem and investment. The first center of mining and cutting was India, which was replaced by Brazil (18th century), South Africa (19th century), and most recently Zaire. The output amounts to ca. 50 megacarats (1 carat=0.2 g), which is not sufficient to satisfy the increasing need for industrial diamonds. Diamonds have been synthesized since the mid–20th century [General Electric Co., ASEA (now ASEA Brown Boveri AG)]. The estimated capacity for synthetic diamond production exceeds 100 megacarats.

1.1 Properties and Structure

Diamond is characterized by its hardness (Mohs hardness 10) and luster. The mineral is usually colorless, but may be pink, light blue, or even black (carbonado) as a result of incorporated impurities. The density is 3.515 g/cm^3 (that of graphite is 2.266 g/cm^3). The lattice of the cubic modification is face-centered with the space group $Fd3m\text{–}O_h^7$ and the lattice constant $a = 0.356\,683$ nm (Fig. 3).

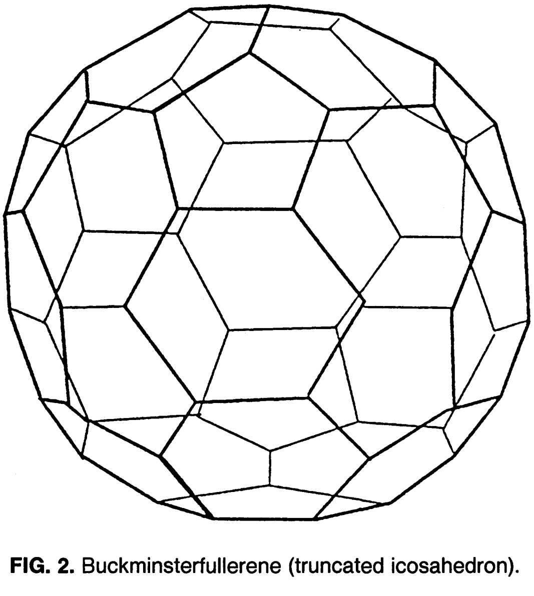

FIG. 2. Buckminsterfullerene (truncated icosahedron).

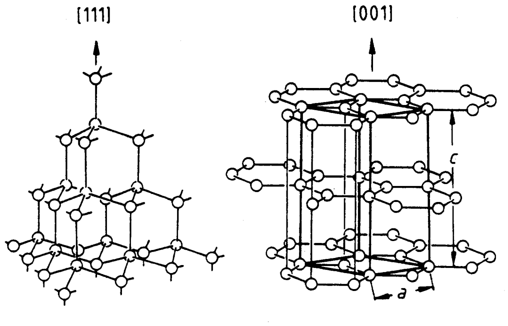

FIG. 3. Lattices of the cubic diamond and the hexagonal graphite crystals.

Each carbon atom has four nearest neighbors at a distance of 0.1544 nm and is linked by sp^3 hybrid bonds; the bonding angle of the covalent bonds is 109.5°. Distinction must be made between type-I and type-II diamonds (and several subgroups) (see DIAMONDS AND DIAMONDLIKE MATERIALS), which have different absorption spectra. The electrical resistivity at room temperature is about 10^{18}–10^{20} Ω cm (type I) and 10–10^3 Ω cm (type II) with larger deviations for contaminated or doped species. Type-II crystals are typical semiconductors; the resistivity decreases with temperature according to $\log \rho \sim 1/T$. Diamonds are excellent conductors of heat. The thermal conductivity at room temperature amounts to ca. 10 W cm^{-1} K^{-1} (type I) and ca. 20 W cm^{-1} K^{-1} (type II), which is about 3 times greater than that of copper. The specific heat is 6.20 (300 K), 13.55 (500 K), and 21.60 (1000 K) J g^{-1} K^{-1}, and the coefficient of linear thermal expansion (between 0 and 100 °C) is about 1×10^{-6} K^{-1}. Crystals cleave well parallel to the (111) plane. The cutting hardness is different for different crystal faces, which must be taken into consideration when cutting. The coefficient of friction depends on the purity of the surfaces. With increasing degasification, the coefficient rises from ca. 0.1 to 0.6 (Field, 1979). Natural diamonds contain up to 0.2% nitrogen (types Ia and Ib), 10 ppm Al, and 0.25 ppm B (types IIa and IIb). Diamond is resistant to most chemical substances at normal temperatures. At higher temperatures, it is attacked by oxygen, oxidizing agents, and carbide-forming elements. At temperatures of ca. 2000 K and above, diamond converts into graphite both internally and on the crystal surfaces. The rate of phase transformation increases rapidly with temperature. See DIAMOND AND DIAMONDLIKE MATERIALS.

1.2 Natural Diamond

The sources of diamonds are probably ultrabasic rocks, especially eclogite, consisting essentially of omphacite and pyrope. The famous blue ground of the Kimberlite slots is a decomposed eclogite rock containing ca. 1–2 carats of diamonds per ton of rock material. Placers—e.g., Lüderitz Bay, Namibia—are also important deposits.

While about one-quarter of the mined diamonds can be used for gems, the greater part is used in industry (bort, carbonado).

1.3 Diamond Synthesis

Graphite can be converted into diamond by first dissolving it in melted metal and then crystallizing at pressures above equilibrium. At pressures above 12.5×10^3 MPa and temperatures in the 3000-K range, graphite spontaneously collapses into polycrystalline diamonds (Bundy *et al.*, 1973). Figure 4 shows the phase and reaction diagram with the pressure–temperature regions used for synthesis.

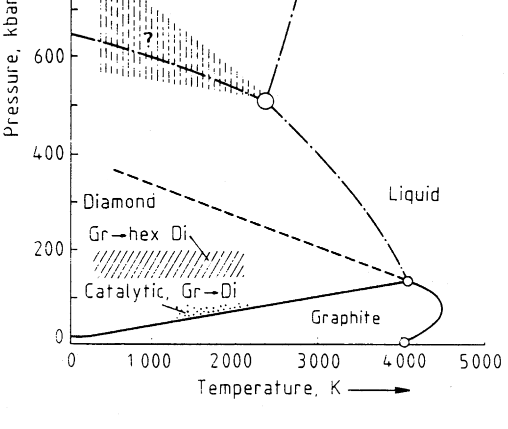

FIG. 4. Carbon phase diagram (Bundy *et al.*, 1973).

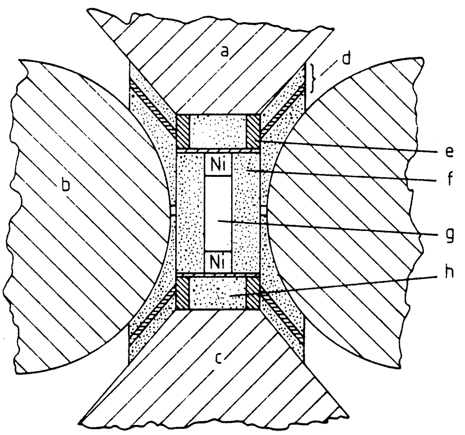

FIG. 5. Belt-type pressure apparatus (Bundy *et al.*, 1973). *a*, top piston; *b*, cylinder; *c*, bottom piston; *d*, gasket assembly; *e*, current ring; *f*, sample holder; *g*, graphite; *h*, insulating pill.

For synthesis, special high-pressure and temperature-resistant apparatus have been developed. The "belt"- and "multipiston"-type apparatus are the most commonly used. The belt-type apparatus is illustrated in Fig. 5. It consists of two truncated conical pistons and a prestressed cylinder assembly, which compose a reaction chamber. The circular gaps between the sections are filled with conical gaskets made of pyrophyllite and steel. The gaskets act as seals, pressure transmitters between the segments, and electrical and thermal insulators. The reaction sample contains stacked graphite and metal plates, or a mixture of graphite powder and nickel powder or other group-VIII metals, and is heated by resistance heating. The current is supplied via pistons and current rings. The lifetime of this apparatus with a diameter of up to 40 mm is ca. 100–1000 cycles when properly operated. After cooling, the diamonds are separated by dissolving the metal matrix with acids. The average size of synthetic diamonds is ca. 0.05–0.5 mm. For the production of semiconductors, doping of the crystals with boron is possible. Smaller diamonds also can be produced by shock waves (see Fig. 4).

1.4 Diamond Thin Films

A wide variety of methods based on activated chemical vapor deposition is proposed for the generation of diamond thin films. Hydrogen gas containing a few percent of methane, ethane, or other hydrocarbon compounds is activated by thermal plasma, ion-beam, laser-beam, microwave plasma, and other techniques at pressures of about 0.01–0.1 bar and temperatures of about 800–1300 K. The difficulties of diamond nucleation and growth are the short lifetime of the excited carbon atoms and the suppression of the formation of nondiamond carbon. Nucleation kinetics, film growth, and film properties depend strongly on the deposition conditions. The deposition rate can be as high as 20 μm/h; it seems to be possible to increase this rate. Diamond thin films contain, besides diamond (sp^3 bond), variable amounts of graphite (sp^2 bond), even carbynes (*sp* triple bond), and hydrogen.

Because of their outstanding hardness, it is proposed that cutting tools be coated with diamond thin films. Other expected applications are optical films and windows for x-ray equipment, semiconductor components, and temperature sinks for semiconductor devices (Angus, 1990).

1.5 Applications

Because of its hardness and toughness, diamond is well suited for all cutting and grinding operations. About 35% of industrial diamonds are used in grinding and honing tools, 20% for saw edges, 20% for drilling heads, 20% for loose grinding, lapping, and polishing powders, and about 5% for drawing dies, bearings, and turning tools. The powdery diamond usually must be bound into a metal or resin matrix, which is fixed in a support. Single diamonds are used for drawing dies, bearings, and tactile devices, as well as for diamond jewelry.

2. GRAPHITE

Graphite has been used for the production of refractory crucibles since at least the 15th century, and for pencils since about 1550. It is named after the Greek word γραφειν (to write). The history of graphite has been closely associated with the development of electricity. Sir Humphry Davy, in 1800, was probably the first to use carbon electrodes for his studies on the electric arc. The increasing demand for electric arc-light carbons after about 1870 gave rise to several processes for the manufacture of artificial carbon (Carré, Brush, Siemens). At the beginning of the 20th century, important electrolytic and electrothermic processes using carbon or graphite electrodes became available, e.g., aluminum electrolysis (Hall, Héroult), alkali-chlorine electrolysis (Castner *et al.*), calcium carbide (Willson), electric steel (Héroult), and silicon carbide (Acheson). See also GRAPHITE.

2.1 Single-Crystal Graphite

The ideal crystal lattice of hexagonal graphite is illustrated in Fig. 3. Each carbon atom has three nearest neighbors within the layer plane at a distance of 0.1421 nm and a bonding angle of 120°. Within the layer planes, the bond energy due to the trigonal sp^2 hybridization of the bonding orbitals is far greater than that of the nonhybridized orbitals, which form weak π bonds vertical to the layer planes. The layer or basal planes of the hexagonal modification are stacked in the sequence *ABAB*..., and those of the rhombohedral modification are in the sequence *ABCABC*... . Natural graphite normally contains both modifications, in some cases up to 30% of the rhombohedral modification. The proportion of this modification may be increased by milling and shearing and is decreased by thermal annealing. The space groups of the graphite modifications are D_{6h}^4–$P6_3/mmc$ and D_{3d}^5–$R\bar{3}m$, respectively.

The interlayer spacing depends on the degree of disorder and amounts to 0.3354 nm in an ideal crystal, 0.344 nm for crystals with turbostratic structure (random stacking sequence), and even more for less-ordered structures. One indicator of the disorder is the width of the x-ray diffraction lines. Most theories explaining the broadening are based on the assumption that aligned layer stacks only correlate in parallelism of the layers. The broadening is then due to translation and rotation of the layers (Franklin, 1951).

Other models only accept translation parallel to the layer planes, caused by clusters between the layers and larger vacancies within the layers (Ruland, 1968). These defects yield to deformations of the layers and deviations from the ideal *AB* sequence, respectively.

Because of the layered structure, all vectorial properties are to a large extent controlled by direction. Some properties of single graphite crystals, their dependence on the crystal direction, and, for purposes of comparison, some common properties of diamond are shown in Table 1. The anisotropy of thermal expansion is particularly remarkable since the crystal expands in the *c* direction and contracts in the *a* direction at temperatures below 400 °C. Parallel to the basal planes, graphite is a metallic conductor whose resistivity increases with rising temperature. In the *c* direction, graphite behaves like a semiconductor. The thermal conductivity strongly decreases with decreasing temperature. At room temperature, the thermal conductivity in the *a* direction is better than that of copper. The properties shown in Table 1 strongly depend on the type of samples; therefore the original publications or manuals should be consulted for a deeper understanding (Kelly, 1981; Reynolds, 1968).

Table 1. Properties of graphite and diamond (single crystals).

	Graphite		
	c direction	*a* direction	Diamond[a]
Density, g/cm^3	2.266		3.515
Coefficient of linear thermal expansion (15–150 °C), 10^{-6} K^{-1}	28.3	-1.5	0.8
Thermal conductivity, W cm^{-1} K^{-1}	0.04–0.06	10–15	20
Resistivity, Ω cm	1	50×10^{-6}	10^{20}
Magnetic susceptibility, 10^{-6} cm^3 g^{-1}	-21	-0.3	
Elastic modulus, GPa	36	1000	1000
Mohs hardness	9	0.5	10

[a] The properties of diamond crystals vary considerably.

2.2 Natural Graphite

2.2.1 Mining and Properties. It is currently supposed that organic matter was the origin of most natural graphites, with the probable exception of the Sri Lanka deposits. Graphite inclusions in plutonic rocks can be traced back to the accumulation of insoluble carbon within the residual melt. The decomposition of inorganic compounds, e.g., carbide, carbonyl, and cyano compounds, is less important for the genesis of graphite deposits.

Crystallinity, refractoriness, and other properties of natural graphite from different deposits may vary enormously as a result of differences in the precursor materials and the conditions of metamorphosis. In general, natural graphite is a ductile soft mineral that easily cleaves parallel to the basal layer. The flaky grades (plumbago or vein graphite) show a typical metallic luster; the microcrystalline "earthy" grades have a duller black appearance. The most common accompanying minerals are quartz, mica, and feldspars; accessory minerals may include pyrite, magnetite, pyrrhotite, and calcite.

Natural graphite is mined in open-pit and underground mines. Crude ores with a carbon content sufficiently high for industrial use are only crushed, dried, graded, and, if necessary, milled. In general the carbon content is only about 30% to 50%, and graphite must be enriched, preferably by flotation processes.

Carbon enrichment by flotation yields up to about 96% C for flake graphites. The mineral components of graphite may be further reduced if the graphite is purified by treatment with hydrofluoric acid, melted sodium carbonate, or chlorine and fluorine at temperatures up to 3000 °C. The ash content can be reduced by these processes to less than 50 mg/kg.

2.2.2 Uses. Natural graphites vary widely in flakiness, refractoriness, toughness, durability, bulk density, etc., and meet the requirements of various processing industries: 24% are used in foundries, 24% in the steel industry, 15% for the production of crucibles, 12% in the electrical field, 8% in the refractory industry, 5% for lubricants, 4% for pencils, and 3.5% for batteries.

Graphite crucibles are used in foundries for melting steel, nonferrous metals, and precious metals. Bricks for open-hearth furnaces, ladle plugs, and tap spouts are other examples of the use of natural graphite in metallurgy. When compared with other refractory materials, graphite-containing grades have generally better thermal conductivity, thermal shock resistance, and resistance against corrosion.

In the production of dry cells and storage cells, graphite powder is mixed with the active mass to increase the electrical conductivity. The conductivity of natural graphite powder is also used in plastics (against static charge buildup), explosives (against ignition by friction), and conductive lacquers.

Natural graphite is also used as a lubricant. Fine-grained graphite powders often having a particle size less than 1 μm are suspended in liquid carrier media, such as oils, greases, alcohols, or water. Most self-lubricating bearings and packings contain natural graphite as a lubricant; pencils function through the gliding and lubrication of their graphite component. Other applications that use these properties are the smoothing of the surface of molds to facilitate the removal of castings after cooling, blackwashes, and additives decreasing the friction during the compaction of powders.

Natural graphite is also used for carburization, deoxidation, and the covering of metal melts. Plastics, ceramics, porcelain, and concrete are colored with graphite, particularly with microcrystalline grades, and it also serves as a pigment in temperature-resistant paints. Well-crystallized flake graphites are a precursor for graphite intercalation compounds, expanded graphite, and graphite foils (see Secs. 2.4.4 and 2.4.5).

Purified flake graphite has been proposed as a moderator and reflector material for nuclear fission reactors. To some extent, natural graphite is used for the production of fuel elements for high-temperature nuclear reactors.

The world production of natural graphite amounts to about 500×10^3 t. The most important countries mining graphite are the following:

- flaky graphite grades: USSR, China, Malagasy, Sri Lanka, Germany, and Norway;
- earthy graphite grades: Korea, Mexico, and Austria.

2.3 Artificial Graphite and Carbon with Layered Structure

2.3.1 Structure and Pyrolysis. Artificial graphite is produced by heating carbon compounds or carbon powder to 2500–3000 °C. By this treatment the structural order is increased, depending on the structure of the precursor materials. One distinguishes between soft (or graphitizing) carbons and hard (or nongraphitizing) carbons by the fact that the preordering of the first type is distinctly better than that of the second. Examples of the two types are petroleum coke and charcoal, respectively. Usually, solid carbon material heated above ca. 2500 °C is designated as an artificial graphite, irrespective of its real crystal structure. Other methods for the production of artificial graphite are CVD (chemical vapor deposition) processes or the condensation of carbon vapor (pyrographite), crystallization of carbon melts at pressures above 15 MPa, precipitation from oversaturated metal melts (kish graphite), the decomposition of carbide crystals, and nitrogenation of calcium carbide, as well as the electrolytic decomposition of carbonate melts. All these processes are economically less significant than thermal graphitization.

All carbon-containing materials are, in principle, suited as a raw material for the production of carbon and graphite. In the first step, the raw material is decomposed by pyrolysis, and in the second step, the intermediate carbon product is graphitized by further heat treatment. During pyrolysis, the intermediate material may be solid, liquid, or gaseous; it is obvious that the mobility of the molecules determines the degree of microstructural ordering.

It is, therefore, customary to distinguish between solid-, liquid-, and gas-phase pyrolysis, although transitions between the pure reaction types exist. Examples of *solid-phase pyrolysis* are chars and vitreous or glassy carbon. Because of the structural disorder and the small average crystal size of their precursors, the ability to transform these materials into highly ordered graphitic structures is low. The precursor material of carbon fibers, e.g., thermally stabilized polyacrylonitrile or pitch, also does not melt during pyrolysis, but an ordered microstructure is achieved by stretching the fibrous material below the pyrolysis temperature and preserving this degree of order by the above-mentioned stabilization treatment.

An example of *liquid-phase pyrolysis* is petroleum coke, which is the most important raw material for the manufacture of carbon and graphite products. Brooks and Taylor (1968) first discovered the formation of optically anisotropic spherulites (*mesophase*) during the pyrolysis of petroleum pitch, coal tar pitch, and softening coal. During heat treatment, pitch gradually forms an isotropic melt and develops small spherulites. The spherulites are nematic liquid crystals composed of large lamellar aromatic molecules. The isotropic phase is finally exhausted, and the spherulites touch each other and coalesce. Formation and structure of mesophase are essentially governed by three limiting conditions:

1. The shape of primary molecules must favor the formation of liquid crystals;
2. the system must have sufficient fluidity;
3. the reaction or condensation rate must be smaller than the ordering rate.

Other factors influencing the structure of the mesophase that eminently affects structure and properties of coke materials are the amount of nonmelting constituents and/or impurities (Marsh and Walker, 1979).

The chemistry of the pyrolytic conversion of hydrocarbon compounds to solid carbon involves numerous chemical reactions. Generalized schemes have been developed by studying simpler models (Fitzer, Müller, and Schäfer, 1971). The major reactions involved are

1. bond cleavage and formation of free radicals,
2. molecular rearrangement,
3. thermal polymerization,
4. aromatic condensation, and
5. elimination of side chains and hydrogen.

A general scheme is shown in Fig. 6. Through the formation of volatile products during pyrolysis and structural transformation, pores and shrinkage cracks are generally developed. Carbon must be heated to ca. 3000 °C to be transformed into graphite. The degree of graphitization, which is an indication for a perfect graphite crystal, the spatial distribution of the crystallites, and the amount and distribution of porosity determine the properties of the

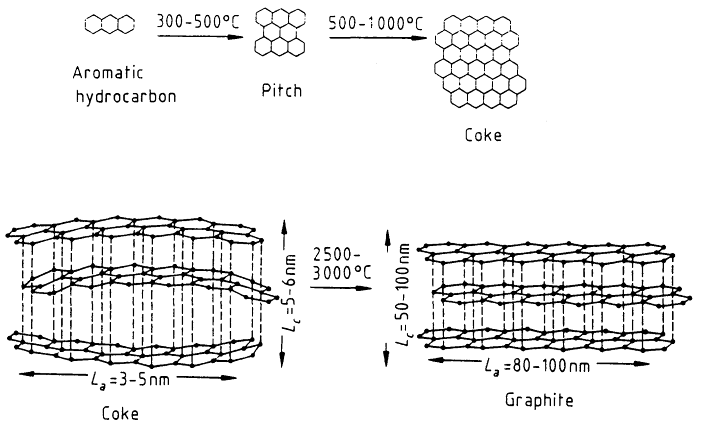
300-500°C
Aromatic hydrocarbon
Pitch
500-1000°C
Coke
L_c=5-6nm
2500-3000°C
L_c=50-100nm
L_a=3-5nm
Coke
L_a=80-100nm
Graphite

FIG. 6. Reaction scheme for carbonization and graphitization. L_a and L_c are the apparent crystallite sizes in the *a* and *c* directions, respectively.

final product. Lattice disorder as well as porosity can be reduced by high-pressure treatment at temperatures above 2500 °C.

Gas-phase pyrolysis is characterized by thermal cracking of gaseous hydrocarbon compounds and deposition of carbon on a substrate. Examples are pyrocarbon (see Sec. 2.4.3), carbon films, and carbon black. Stress annealing of pyrocarbon at high temperature yields nearly perfect graphite crystals.

2.3.2 Carbon and Graphite Artifacts

2.3.2.1 Manufacture. Most methods for the manufacture of carbon artifacts consist of a two-stage process. In the first step, raw materials, especially coke, are produced by pyrolysis of carbon-containing substances, e.g., pitches and coals. In the second step, the coke, together with a binder, is processed to carbon in a manner similar to the manufacture of ceramic materials. The carbon can be transformed into graphite by subsequent heat treatment (Ishikawa and Nagaoki, 1984; Blackman, 1970).

Petroleum coke, pitch coke, metallurgical coke, anthracite, and carbon black are commonly used as raw materials. Coal tar pitches are the most commonly used binder material; less common are petroleum pitches and resins. Mixtures of these materials are shaped by extrusion, vibration molding, die molding, or isostatic molding and are baked at temperatures of ca. 1000 °C in an inert or reducing atmosphere and may be graphitized by further heating to 3000 °C. The manufacturing process is shown in Fig. 7.

Raw Materials. The range of raw materials extends from well-ordered *natural graphites* and *needle cokes* with nearly completely two-dimensional ordered layer stacks to less-ordered materials such as chars and certain

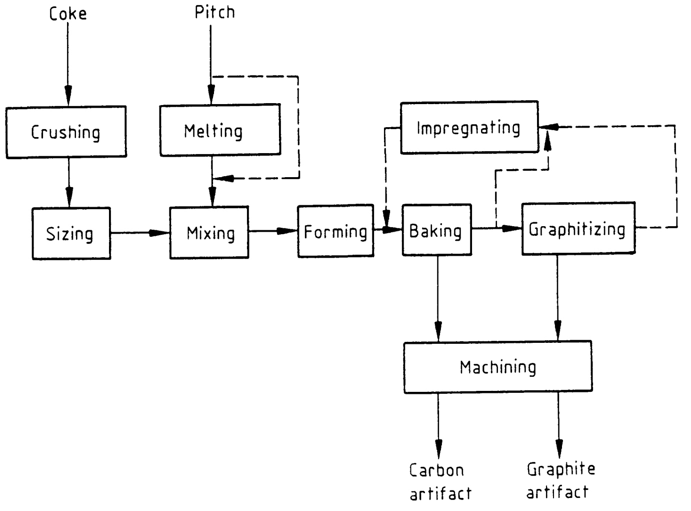
Coke
Pitch
Crushing
Melting
Impregnating
Sizing
Mixing
Forming
Baking
Graphitizing
Machining
Carbon artifact
Graphite artifact

FIG. 7. Carbon and graphite manufacturing, flow diagram.

types of *carbon blacks*. The properties of carbon and graphite products and their performance depend to a large extent on the properties of the raw material. Therefore the selection of raw materials is an important step in the manufacture of carbon and graphite.

The purpose of *binder materials* is to wet the surface of carbon grains, give mixtures of high plasticity when mixed with grains, and give high-yield carbon residue after pyrolysis. Coal tar pitches, the residues from the distillation of crude tar produced while coking bituminous coal, are most suited. Petroleum pitches processed from decant oils and aromatic extracts have also been proposed as binding and impregnating materials. Thermosetting resins are also reliable binders and impregnating agents, although their use is limited to the production of smaller carbon and graphite shapes because of higher costs. Phenolic and furan resins are frequently used.

Grinding and Sizing. The bulk density of carbon and graphite shapes depends on the porosity of the coke grains and the packing density of these grains (Mrozowski, 1956). These variables are controlled by crushing and grinding the coke and sizing the milled product. Large grains are generally used for the production of large shapes and smaller ones for small bodies. The maximum grain size is ca. 15–20 mm for large electrodes and less than 20 μm for spectroscopic carbon grades. The microstructure of the artifacts improves with decreasing grain size.

Mixing. Mixing is generally performed at temperatures at least 50 K above the softening temperature of the pitch binder. The proportion of the filler material and binder determines the fluidity of the mixture, which also depends on the grain size of the filler material and the forming procedure. The average binder content is about 15%–30%. A green mixture consists primarily of filler, binder, and some air bubbles.

Forming. Elongated shapes are usually formed by *extrusion*. The cross section may be circular, rectangular, or more complicated, e.g., for finned tubes. The cross section ranges from a few square millimeters (battery carbon, spectroscopic carbon) to nearly 1 m^2 (electric furnace electrode). The hot carbon mixture is charged into the cylindrical chamber (mud cylinder) of a conventional hydraulic press, evacuated, and compacted. The mixture is forced through a forming dye by an advancing ramhead or piston. The mud chamber is up to ca. 1.5 m in diameter and up to 3 m in length, and the pressing force up to 60 MN. Smaller bodies, as well as some of the anode blocks used in the electrolysis of aluminum, are frequently produced by *molding*. Because of the smaller amount of binder, the shrinkage of these bodies can be better controlled, and forming to the final size is possible ("press to size"). Pressures as high as 200 MPa (2 kbar) are necessary for compacting.

In *isostatic molding* the compacting forces are transmitted by a fluid acting equally from all directions. Accordingly, the density gradient is smaller and the uniformity of the pressed form is better. The dies (bags) consist of an elastomeric material, e.g., silicone rubber. They are usually evacuated during or after being filled with the pressing powder and are attached to the pressure tank (dry bag) or are freely removable (wet bag). The pressure tank is up to 1 m in diameter and more than 2.5 m in height. The pressure is up to 200 MPa (2 kbar). During compacting, all anisometric particles are aligned perpendicular to the direction of force, yielding a quasi-isotropic body.

During the last decade, *vibration compacting* has steadily increased, particularly for the production of carbon electrodes used for aluminum electrolysis and the production of other large shapes containing coarse-grained carbon. To achieve sufficient compaction, the pressing mixture must be loaded with top weights and then evacuated. The vibration frequency is ca. 1000–2000 cycles per minute. *Jar molding* and *tamping vibration molding* methods with a low frequency, but large amplitude, are used to form large-diameter carbon electrodes.

Other forming methods are less common. Carbon shapes with a high bulk density can be produced by explosion impact or by implosion into a mold. Spin molding is used especially for mixtures containing clay for the production of graphite crucibles.

Baking. In the baking process, the binder material of the shaped artifact is pyrolyzed, and a binder coke is formed that joins together any coke particles. Volatile products are generated, which must be removed from the carbon shape. The process is governed by the following steps:

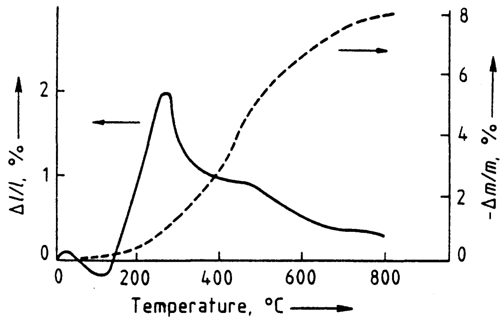

FIG. 8. Variation in length and mass of pitch-bonded carbon shapes during baking.

1. the heat transfer from the outside into the shape,
2. the pyrolysis reactions, and
3. the removal of volatile products from the shape

(Hüttinger, 1972). In addition to the mass loss, the shapes undergo complicated volume changes during baking that depend on the type and amount of the binder, the heating rate, and the volume of the body (Fig. 8). To prevent deformation and oxidation during baking the carbon shapes are enveloped by selected packing materials.

In general, all furnaces are suited for baking carbon bodies, provided the temperature is adjustable. Each furnace also must have devices for collecting and purifying the volatile pyrolysis products, e.g., electrostatic precipitators or incinerators. Most frequently *ring furnaces* are used, consisting of 10–30 identical brick chambers or pits being heated with oil or natural gas. A baking or heating cycle is shown in Fig. 9. Several types of *single-chamber furnaces* and *tunnel kilns* are also used for baking. The baking cycle time can be reduced by applying pressure during heat treatment (pressure baking).

Graphitization. Graphitization is the transformation of disordered carbon material into better-ordered graphitic structures; the heat treatment of carbon shapes at temperatures above ca. 2500 °C is referred to by this term. During this heat treatment, the average interlayer distance decreases, and the height and width (L_a and L_c, defined in the caption of Fig. 6) of the ordered areas increase. The attainable order or degree of graphitization depends on the structure of the precursor material. Structural ordering is a relatively rapid process (Fig. 10). All volume properties, such as bulk density, resistivity, strength, etc., depend not only on the maximum temperature, but also on the heating rate. The heating rate of larger specimens may be limited by insufficient thermal shock resistance. The actual rate is shown in Fig. 11. During heat treatment, the impurities within the artifacts react with the carbon to form carbides, which mostly decompose and vaporize at higher graphitizing temperatures. At temperatures of ca. 1600–1700 °C, the sulfur contained in the carbon bodies is promptly emitted, thereby increasing the body's volume (puffing). To avoid the formation of cracks, puffing must be reduced by adding inhibitors.

Graphitization is performed almost exclusively with Acheson and Castner furnaces. In both furnace types, the heating is accomplished by means of electric current passed through the bodies. Inductive and radiation heating is used only for the production of specialities.

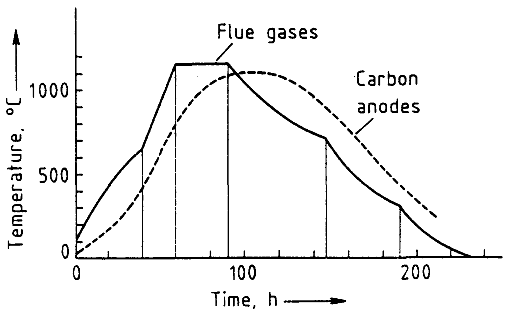

FIG. 9. Ring furnace heating cycle.

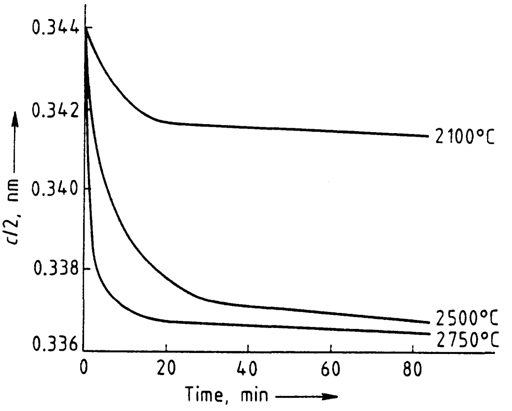

FIG. 10. Change of interlayer spacing during graphitization.

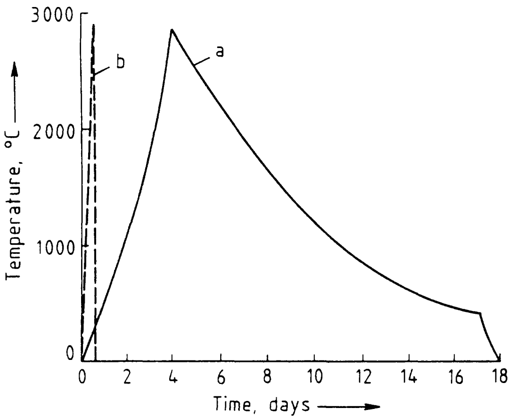

FIG. 11. Temperature cycles of (curve *a*) the Acheson furnace and (curve *b*) the Castner furnace.

The *Acheson furnace* consists of a horizontal bed of refractory materials, furnace heads carrying the current load to the furnace, and long sidewalls made of concrete blocks, steel sheets, or steel grids. The bottom is cooled by air and loaded with a layer of insulating materials, e.g., granular silicon carbide, metallurgical coke, sand, or sawdust. The same materials are used for insulating the sidewalls and the surface of the furnace. Carbon bodies that are to be graphitized are stacked between the furnace heads, usually horizontally and transversely to the flow of the electric current. The bodies are separated from each other and surrounded by a granular resistor medium, e.g., coke–graphite mixtures. The maximum capacity of Acheson furnaces is about 100 t, and the electric load 0.1–1 MW/t. The energy consumption of larger furnaces amounts to 3–4 kW h per kg graphite material.

Castner furnaces do not contain granular resistor materials. One or more carbon artifacts are clamped between head electrodes, these forming the path for the heating current (lengthwise or longitudinal graphitization). One or both ends of the column are subjected to a mechanical or hydraulic pressure device to maintain a tight connection during the heating cycle. Insulation is similar to that of the Acheson furnace. Much higher heating rates are possible, and heat losses are accordingly smaller. The heating time is less than 10 h, and the energy consumption is about 2–3 kWh per kg of graphite material.

Purification. Coke and binder materials used for the production of graphite artifacts contain various quantities of impurities that may prevent some applications. By thermal, thermochemical, and chemical purification the ash content can be decreased considerably, e.g., as low as 1 mg/kg for spectroscopic carbons.

Machining. Graphite is readily machinable. Lubricants or cutting fluids are not necessary. It can be sawed, turned, bored, drilled, planed, shaped, and milled. Surface roughness decreases in this sequence from ca. 25–125 to 1–6 μm. A fine-grained grade is always smoother as a result of the smaller pores revealed through cutting. The same machines are generally used as in metal working.

Impregnation and Surface Coating. About 25% of the volume of carbon or graphite artifacts consists of pores that are sometimes detrimental to the bulk properties. The pores may be filled at least partially or closed by impregnation with a solid material whereby strength, conductivity, and resistance against oxidation are improved. Pitch is the most frequently used impregnating agent. Graphite grades used at lower temperatures may be impregnated with curable resins, particularly phenol-aldehyde and furan resins. The tolerated application temperature of these grades is at the most ca. 200 °C. The pores of carbon and graphite bodies may also be filled with carbon formed by CVD processes. Coatings containing refractory materials, e.g., carbides, silicides, borides, or nitrides, protect carbon artifacts against oxidation. Coatings are brushed or sprayed onto the surface, and the deposited layer is sintered, or the coating is applied by flame or plasma spraying or CVD techniques.

2.3.2.2 Properties. Artificial graphite and carbon bodies differ more or less from the graphite single crystal. This difference is due to structural disorder, imperfect alignment of the crystallites, and porosity. Other parameters influencing the properties of graphite artifacts are the size and size distribution of the coke grains, the shape of the grains, the pore size distribution, the interaction between the components, and the content of impurities. Each parameter can be changed through the selection of raw materials and production methods. For instance, needle coke consists of highly ordered and aligned crystallites, regular petroleum coke is less ordered, and carbon-black products are not ordered at all.

Physical Properties. The variation of properties of carbon and graphite products is

Table 2. Properties of carbon and graphite products (with grain).

	Bulk density ($g\ cm^{-3}$)	Open porosity (%)	Young's modulus ($kN\ mm^{-2}$)	Flexural strength ($N\ mm^{-2}$)	Resistivity ($\Omega\ \mu m$)	Thermal conductivity ($W\ m^{-1}\ K^{-1}$)	Coefficient of thermal expansion ($10^{-6}\ K^{-1}$)
Carbon block for furnace linings	1.55	17	11	10	50	3	3.5
Carbon anode	1.52–1.60	15–18	12–15	10–20	25–35	3–5	2.5–3.5
Graphite electrode							
Normal	1.55–1.65	20–25	6–8	7–10	8–10	150–160	0.7–1.4
UHP (ultrahigh power)	1.65–1.75	18–20	7–9	10–15	5–8	160–180	0.5–1.1
Graphite die for continuous casting	1.70–1.75	15–18	10–12	30–40	10–15	120–140	1.2–1.6
Resin-impregnated graphite for chemical application	1.85	0	15	35–40	15	120–130	3.5
Glassy carbon	1.55	0	25	70	50	5	3.5
Carbon fiber (high modulus)	1.95	0	400	3000	1	>100	−1.0

illustrated in Table 2. The mechanical properties are similar to those of ceramic materials, whereas the thermal conductivity and electrical resistivity correspond to those of metals (Pacault, 1965).

The unique combination of refractory and metallic properties is useful for electrothermal and electrolytic processes performed at higher temperatures. In progressively loaded graphite artifacts, a rising number of noncritical microcracks is generated, propagating preferentially normal to the applied stress and causing relief of internal stress. The propagation of the cracks is restricted by pores or structural elements that are oriented transversely to the direction of propagation. When a critical crack length is reached through linking of microcracks, macrofailure will occur. The Weibull approach can be used to give a reasonable prediction of failure. The fracture toughness K_c of common graphite grades is about 1–15 MN/m$^{3/2}$ (Jenkins, 1974).

Graphite is a semimetal (see ELECTRONIC STRUCTURE OF SOLIDS) in which the highest filled valence band overlaps the lowest empty conduction band (Spain, 1973). There is no simple relationship between the conductivity of single graphite crystals and that of polygranular artifacts. The resistivity (with grain) of the latter is more than 20 times greater than that of a single crystal (in the *a* direction), and it decreases with temperature up to ca. 350 °C for graphite grades and up to ca. 1000 °C for carbon grades. The sign of the temperature coefficient of resistivity changes above these temperatures because increased scattering of the charge carriers compensates their rising number in the conduction band. The thermal conductivity of carbon and graphite artifacts is determined by phonon scattering at the crystal boundaries. Conductivity is at its maximum at about room temperature. More informative for users of carbon and graphite artifacts may be the change of these properties with the temperature. Strength and Young's modulus of graphite artifacts reach maxima at temperatures of ca. 2500 °C, where they are 50%–100% larger than at room temperature. Above 1400 °C, the specific strength (strength/density) of graphite is better than that of metals and most other refractory materials. The coefficient of emissivity amounts to 0.8–0.9.

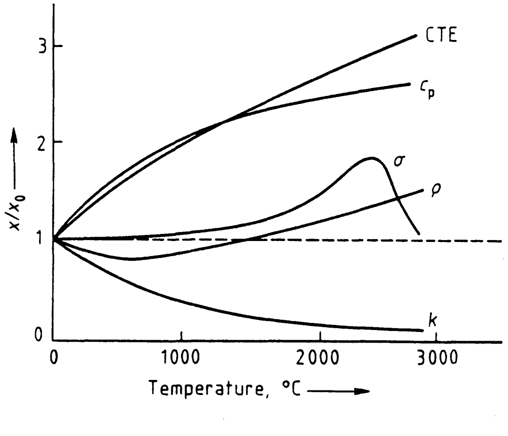

FIG. 12. Relative properties of polygranular graphite grades vs temperature (schematic). CTE, coefficient of thermal expansion; c_p, specific heat capacity; σ, tensile strength; k, thermal conductivity; ρ, electrical resistivity.

Table 3. Average thermal-stress-resistance factor R' (W/m). $R' = k\sigma/(1-\mu)E\alpha$, where k=thermal conductivity, μ=Poisson's number, E=Young's modulus, α=coefficient of thermal expansion, and σ=tensile strength.

Graphite	50 000
Titanium carbide	1 000
Magnesia	50
Alumina	1

The temperature dependence of strength and several other properties is illustrated in Fig. 12. Types used at high temperatures are often exposed to sudden temperature changes or large temperature gradients, both including mechanical stresses. Table 3 shows the favorable thermal-stress-resistance factor R' of graphite compared to that of several refractory materials.

Chemical Properties. Carbon and graphite products are resistant to most corrosive substances, except strong oxidants, substances forming intercalation compounds, and carbide-forming substances. At higher temperatures, graphite is, in fact, less reactive to oxygen than most metals. But since the reaction products CO_2 and CO are volatile, no protective oxide film can be formed. The reactivity depends primarily on the degree of crystalline order. At lower temperatures, the reaction is not limited to the outer surface of the artifact but includes practically the entire inner surface, because of pore diffusion. With increasing temperature, the depth of the reaction zone decreases. In air, the average loss of small graphite artifacts in the temperature range between 500 and 600 °C is ca. 1% per day. Impurities such as alkali metals, vanadium, and lead accelerate the reaction rate considerably. The burn-off may be decreased by impermeable protective coatings, e.g., carbides or silicides, or by the addition of substances which inhibit oxidation.

2.3.2.3 Applications. *Electrodes for Arc Furnaces.* About 90% of graphite products are electrodes used for the production of electric arc furnace steel. Graphite electrodes up to 700 mm in diameter consist of cylindrical sections up to 2700 mm in length, which are joined with tapered, threaded nipples. The electrical load is up to 30 A/cm^2, and the surface temperature increases along the length of the electrode up to 2300–2800 °C (Weng and Seldin, 1972).

Electrodes for Electrothermal Reduction Furnaces. Electrothermal reduction furnaces are used for the production of calcium carbide, ferroalloys, phosphorus, silicon, and refractory oxides. Because of the relatively low current density of ca. 6–10 A/cm^2, adequate results can be achieved with carbon electrodes, particularly Söderberg electrodes (selfbaking electrodes). Söderberg electrodes consist of a sheet metal casing that is filled with a pasty mixture of electrode raw materials. The lower part of the electrode is baked by the excess furnace heat. Because of their higher electric conductivity, semigraphite electrodes are used, especially for the production of silicon.

Electrodes for Molten Salt Electrolysis. Carbon anodes and cathodes are used for the electrolytic production of aluminum. The anodes are partially submerged into a cryolite melt containing several percent of alumina. During electrolysis, the anodically produced oxygen reacts with the anode material and CO and CO_2 are formed. The ratio CO_2:CO determines the current efficiency. Because of the operation temperature of 950–970 °C, the parts of the anode extending out of the melt are attacked by oxygen in the air. Consumption greatly depends on the properties of the carbon anodes, especially their permeability and reactivity (Keller and Fischer, 1981). The average consumption of carbon anodes is ca. 0.47 kg per kg of Al (theoretically 0.33 kg per kg of Al). Current supply rods are connected to the anode by carbon-containing cements or metal solders. The edge length of the anodes is up to 1500 mm.

The bottom and sidewalls of the Héroult–Hall cells, which protect the pot and the ceramic insulation against the cryolite melt, also consist of carbon material. Carbon blocks that form the bottom have longitudinal grooves in which cathodic bus bars are inserted. For the winning of alkali, alkaline-earth, and rare-earth metals and fluorine, graphite anodes are used.

Electrodes for Aqueous Electrolysis. During the past several years, titanium anodes have been increasingly substituted for graphite anodes. Accordingly the use of graphite anodes for chlorine, chlorate, and hydrochloric acid electrolysis, for the production of manganese dioxide and hydrogen peroxide, for electrolytic pickling and polishing processes, and for protection against cathodic corrosion is decreasing.

Linings, Casting Molds. Because of their temperature resistance and their resistance to most metal melts and slags, carbon and graphite blocks are widely used as a lining for metallurgical furnaces. For use in blast furnaces, resistance to alkali attack and wear is required. For each refractory material, a critical temperature, which must not be exceeded during furnace operation, has been defined. To obtain sufficient lifetime from the lining carbon, grades with high critical temperature should be used (refractory approach), or the cooling of the lining must be increased by using semigraphite or graphite grades (thermal approach) (Herron and Beechan, 1981).

Because of its thermal conductivity, thermal shock resistance, and low wettability by molten metals and slags—the wetting angle for most metals is 120°–140°—graphite is an outstanding material for foundry molds, ingot molds, and crucibles for the casting of aluminum, magnesium, iron, steel, copper, brass, bronze, and other metals. Continuous casting devices for steel generally consist of materials containing natural graphite and corundum. For nonferrous metals, high-density and high-strength graphite grades are used. Railway wheels are die cast in graphite dies. Graphite molds can even be applied for precision casting.

Graphite is also used for the production of cemented carbide and diamond tools, such as die, boat, crucible, and auxiliary devices for the production of semiconductor materials, and for the refining of metal by feeding gases into the melt. Graphite rods, tubes, or slotted hollow cylinders have a uniform resistance and are preferred as heating elements. Graphite heating elements must be used in an inert atmosphere or under vacuum, keeping in mind that the vapor pressure of carbon is 0.5 kPa at 3000 °C.

Process Equipment. One outstanding property of carbon and graphite is their resistance to corrosion. Graphite also has a relatively high thermal conductivity. For use as heat exchangers and chemical equipment, permeability is considerably reduced by impregnation, especially with thermosetting resins. The permeability of impermeable graphite grades is less than 10^{-6} cm^2/s. The temperature at which it can be used is, of course, restricted by the decomposition temperature of the resins, but the limit of 200 °C is adequate for most chemical processes. Examples of impervious carbon and graphite grades are absorbers, columns, heat exchangers, pumps, valves, pipes, fittings, towers, tower packings, and rupture disks.

Nuclear Reactors. Carbon has a low absorption cross section for thermal neutrons and a favorable moderation coefficient. Since Fermi's Chicago pile reactor, graphite has been used successfully as a moderator and reflector material for gas-cooled reactors. The operation temperature is limited by the nature of the cooling media and is far lower for CO_2-cooled Magnox reactors than for helium-cooled high-temperature reactors (HTR). The neutron absorption cross section is particularly affected by such impurities as boron and rare-earth metals, which are reduced by purification. During neutron irradiation, graphite undergoes remarkable dimensional changes, which depend on structural disorder and the coefficient of anisotropy. The degree of damage can be considerably reduced through selection of raw materials and processing conditions (Nightingale, 1962; Delle *et al.*, 1983). The dimensional changes of the structural elements of graphite during the lifetime of an HTR are now small enough to be compensated by structural design.

Mechanical Engineering Applications. The self-lubricating characteristic of carbon is a fundamental property. The coefficient of static friction is 0.1–0.25 at normal humidity. Carbon and graphite are, therefore, widely used as mechanical seals and bearings if the use of lubricants is not possible because of contamination (food), contact by aggressive substances (chemical industry), or operation at high temperatures. Examples are radial and axial bearings and bushings, seal rings, and vanes.

Miscellaneous Electrical Applications. Carbon and graphite are classical materials for electric contacts and electrodes. Arc-light carbons had a comparatively short history as a street illuminant. They are now almost exclusively used in film projectors and as arc cutting and welding rods. Welding carbons are preferred for spot-welding sheet metal.

For spectroscopic emission analysis, the sample is vaporized and the atoms are excited in an arc between two carbon electrodes. These electrodes tolerate only little contamination, which is also true for crucibles and boats used in atomic absorption spectroscopy. Graphite tool electrodes are juxtaposed with a workpiece in an electric-discharge machine (EDM), both immersed in a dielectric liquid. The workpiece is eroded by a series of electric pulses. The size of EDM electrodes, which can be easily machined, ranges between a few millimeters and more than 1 m.

Graphite electrodes are also used as anodes for mercury-vapor rectifiers and power tubes. Sparking can be avoided if the electrode consists of high-density graphite grades that were heat treated under vacuum at temperatures above 2000 °C. Graphite grids for high-power electron tubes improve the power handling capacity.

Carbon is applied to various types of primary batteries (battery carbon) and fuel cells. Carbon acts as a corrosion-resistant conductor, as an electrocatalyst in air-consuming electrodes, as a catalyst carrier, and as an intercalate-forming active mass (Watanabe *et al.*, 1988). In Leclanché cells, the carbon rod is surrounded by a depolarizing mixture containing MnO_2 and graphite powder or carbon black. The rod is often impregnated with waxes or fats to prevent evaporation of the electrolyte and permit the diffusion of gases. Leakproof cells are provided with additional seals.

Stacked carbon plates or disks are used as pressure-sensitive pile resistors. Examples are voltage regulators for sewing machines and illumination equipment. The high contact resistance of carbon granules, their invariability against atmospheric conditions, and the reproducibility of the resistance–pressure curve make carbon granules a satisfactory material for microphones.

Carbon and graphite are important materials for use as electrical contacts (Holm, 1967). Carbon contacts have an excellent arcing resistivity; they neither weld nor form insulating oxide films. Examples of stationary contacts are switches and circuit breakers for sliding contacts, as well as brushes and current collectors for trolley shoes and wheels. Pantograph shoes are sliding contacts that transfer electric energy from a wire or rail to a moving device (vehicle). Carbon shoes do not wear out trolley cables, wear slowly, have adequate resistance to impact, and cause little radio interference. The shoes are adjusted to special pantographs that maintain nearly constant contact pressure. Examples are electric locomotives (engines), tramways, trolley buses, and mobile cranes.

In most electric motors and generators, the revolving part of the machine and the stationary part are electrically connected by brushes. Brushes must be able to "absorb" the roughness of the commutator or slip ring surface, control film formation on the metallic counterpart, and absorb current pulses that result during current reversal (commutation) (Shobert, 1965).

2.4 Monogranular Carbon and Graphite Products

Carbon and graphite products described in Sec. 2.3 consist of an agglomeration of single carbon particles and can be regarded as a polygranular structure. During recent years, industrially important products have been manufactured from essentially monogranular carbon materials (Fitzer, 1980).

2.4.1 Carbon Fibers. Carbon fiber materials are produced and used as yarns, short filaments, fabrics, or felt. The commercial production of the fibrous materials is based on the carbonization of polymer fiber precursors, in the course of which the fiber form has to be preserved. Therefore thermoplastic fibers have to be converted into a nonmeltable intermediate before carbonization. With this stabilization process additional cross-links are formed, and linear polymers are converted into ladder polymers.

The most important precursor materials for the production of carbon fibers are polyacrylonitrile (PAN) fibers and pitch fibers, particularly fibers converted into mesophase pitch (MPP) before spinning; isotropic pitch fibers are also used, especially for low-end applications. Cellulosic materials (rayon fibers) are used less frequently.

PAN and pitch fibers are first heated up to about 300 °C—generally under tension, in order to avoid shrinkage—and then contacted with cyclization agents, e.g., oxygen. By this treatment, cyclization of the thermoplastic chains to naphthydrine ring systems occurs, and conjugated double bonds are formed by dehydrogenation. The resulting stabilized intermediate preserves the original fiber structure, is infusible, and can be pyrolyzed analogously to polygranular carbon grades (Jenkins and Kawamura, 1976; Fitzer and Heine, 1986).

The properties of the carbon fibers and their behavior depend strongly on the type of precursor and its degree of orientation along the fiber axis. Carbon fibers are outstanding in stiffness and strength. Their stiffness is based on the high C–C bonding energy within the aromatic ring systems ($\sim$600 kJ/mol), whence the maximum preferred orientation in direction of the fiber axis yields the highest stiffness. Typical mechanical properties of the most important commercial PAN- and pitch-based carbon fibers are given in Table 4. With MPP-based fibers Young's moduli even up to 800 kN/mm^2 can be achieved, i.e., 80% of the theoretical C–C bonding strength.

Because of their superior stiffness, combined with high strength and low weight, carbon fibers are predominantly used as reinforcement of duroplasts and thermoplasts in the production of so-called advanced or high-performance composites. *Carbon-fiber-reinforced plastics* (CFRP) are superior to metals and other composite materials (*q.v.*), especially in respect to the specific-strength data (related to the low weight). Therefore high-tech applications such as aerospace and aircraft structural materials give the highest technical and economic benefit. Other application fields are sports goods, engine parts, and the automotive industry, a very promising field for the future.

Carbon fibers are also used for the reinforcement of carbon matrices. These *carbon-fiber-reinforced carbon shapes* (C/C composites) combine the unique high temperature resistance of polygranular carbon with the

Table 4. Mechanical properties of commercial carbon fibers.

	Density (g/cm^3)	Tensile strength (kN/mm^2)	Young's modulus (kN/mm^2)	Breaking strain (%)
Pan-based				
HT (high strength)	1.75	1.95	1.95	1.60
HM (high modulus)	3	2	2	1
Pitch-based				
Anisotropic MPP (mesophase based)	200	400	400	70
Isotropic	1.5	0.6	0.8	3.5

superior strength of carbon fibers: high-tech fiber ceramics with high tensile strength (100–200 N/mm^2) and fracture toughness ($K_c = 20$–$30\ MN/m^{3/2}$). Important applications are aerospace structures, nozzles, aircraft brake disks, high-temperature insulation, linings, and heating elements. To a lesser extent carbon fibers are also used as reinforcement of metals and ceramics. Because of the superiority of the carbon fibers, the demand is increasing rapidly.

2.4.2 Glassy Carbon. Glassy carbon is formed by pyrolysis of thermosetting resins, e.g., phenol, furan, polyimide, and polyphenylene resins (Jenkins and Kawamura, 1976). Glassy carbon is impermeable to liquids and gases and its fracture exhibits glossy shell-like planes. The highly disordered structure consists of a network of branched and twisted ribbons and microfibrils, which are interconnected at several points, with an average height and length of ca. 3 and 10 nm, respectively. The ribbons surround micropores of 5–10 nm diameter.

Because of its cross-linked microstructure, glassy carbon is isotropic and stronger and harder than polygranular carbon grades. The oxidation rate is considerably lower, since the pores are not accessible to reagents. These properties are advantageous for laboratory equipment, e.g., crucibles, boats, bowls, trays, electrodes for electroanalysis, and tubes for evaporation and smelting, as well as for the decomposition of materials for chemical and spectroscopic analysis. Resistance-heated tubes are used in atomic absorption spectroscopy, molds for glass forming, electrodes for organic electrochemical synthesis, and other shapes for numerous other uses. Because of its compatibility with human tissue, glassy carbon has also been successfully used for implants, e.g., bile ducts, and as stimulating electrodes for cardiac pacemakers.

2.4.3 Pyrocarbon and Pyrographite. Pyrocarbon and pyrographite are products of the decomposition of gaseous carbon-containing compounds, a process known as chemical vapor deposition (CVD) (*q.v.*). Parameters influencing this process are the type and concentration of the carbon donor, the temperature of pyrolysis, contact time, and gas flow (Bokros, 1969). Hydrocarbon compounds are favored as reaction gases, especially methane, propane, and propylene. At low temperature and low concentration, the deposited pyrocarbon layers have a laminar structure. At high concentration and a low temperature gradient, the gas is supersaturated and optically isotropic layers are formed. High temperature and low partial pressure promotes the formation of granular and columnar microstructures. The pyrocarbon is deposited on substrates that have the contour of the desired final body. The deposition rate is 10–2000 $\mu m/h$. The bulk density depends primarily on the deposition temperature; it is lowest at temperatures of ca. 1400 °C (1.3 g/cm^3) and increases to 2.23 g/cm^3 at 2700 °C. Through structural variation, it is possible to adjust the coefficient of thermal expansion of the layer to that of the substrate, thereby minimizing stresses caused by temperature gradients.

Highly oriented pyrographite (HOPG) is obtained by thermal or stress annealing of pyrolytic carbons (Moore, 1973). This process decreases the proportion of disordered layers considerably. Interlayer spacing, which is the distance between the single graphite layer planes, can be as low as 0.3354 nm. Accordingly, the properties of this type of pyrographite are similar to a single graphite crystal. HOPG plates are used as flat and bent x-ray monochromators, as neutron monochromators, and for the polarization of high-energy photons. Nuclear fuel particles for high-temperature nuclear reactors are coated with less-ordered pyrolytic carbons impermeable to fission products at high operating temperatures (coated particles) (Delle *et al.*, 1979). The carbon is deposited in a fluid bed, and the deposition conditions are altered so that the coating consists of several layers differing in microstructure and properties.

2.4.4 Graphite Intercalation Compounds. At relatively low temperatures, graphite forms intercalation compounds (*q.v.*) with numerous substances. Graphite oxide with the ideal formula $C_8O_2(OH)_2$ was the first compound investigated. The following rules are valid for these intercalation compounds:

1. Only highly ordered graphite or graphitizable materials yield homogeneous intercalation compounds;
2. the atomic arrangement within the carbon layers is preserved;
3. the intercalated substance is regularly distributed between the carbon layers; and
4. most compounds can be reconverted into graphite.

The reagents capable of intercalation with graphite can be divided into two classes: electron donors (alkali metals, alkaline-earth metals, lanthanides, metal compounds containing hydrogen or polar molecules, and aromatic compounds) and electron acceptors (halogens, halides, oxyhalides, and acids). The number of electrons in the conduction band changes accordingly, eventually reducing the electrical resistivity considerably. Because of reagent penetration, the graphite crystal swells in the *c* direction and the identity period *I*, the height of the repeat unit in the layer sequence (in nanometers), changes according to $I_n = d_i + (n - 1)0.3354$, where 0.3354 nm is the interlayer spacing of graphite, d_i is the period of a first-stage intercalation compound, and n is the stage number. "First-stage" compound implies that all intervals between the carbon layers are occupied. In a second-stage compound, each intercalated layer is separated by two carbon layers, etc. A defined sublattice is associated with each stage conserving the tridimensional structure (Ubbelohde and Lewis, 1960; Hérold, 1979).

Graphite intercalation compounds have only limited stability. Compounds containing electron donors are easily oxidized when exposed to air—some are even pyrophoric—and compounds containing electron acceptors decompose easily or are strongly hygroscopic. The use of graphite intercalation compounds is, therefore, very limited. This is not true for graphite fluoride, CF_x, since this compound is adequately stabilized by the formation of sp^3 bonds.

The proposed application of graphite intercalation compounds as low-resistance conductors has especially been studied with SbF_5 and $SbCl_5$ intercalants (Wu and Vogel, 1982). At room temperature, the electrical resistivity is as low as 3×10^{-8} Ω m, and perhaps even lower.

2.4.5 Graphite Foils. Highly ordered graphite flakes with stack heights of at least 75 nm are used as a raw material. The precursor flakes are usually treated with an oxidizing agent, e.g., mixtures of sulfuric acid and nitric acid at ca. 100 °C. After the graphite intercalation compound is formed—in this example graphite hydrogen sulfate—the particles are thoroughly rinsed with water and then rapidly heated to ca. 1000 °C. The gas pressure of the vaporized and trapped intercalants partially contributes to the exfoliation of the stacked layers. The exfoliated particles (expanded graphite) can be molded to sheets, laminates, and other useful shapes in the absence of a binder. During compression, the particles interlock and form nonrigid bonds.

Properties. The large coefficient of anisotropy is due to the nearly complete alignment of the layer planes perpendicular to the compression force. Sheets are good thermal and electrical conductors parallel to the surface, and semiconductors in perpendicular planes. Graphite foils have considerable flexibility; they can, for example, be wrapped around cylinders. Like other graphite grades, graphite foils are resistant to temperature, aging, and most corrosive media. Graphite foils have a smooth surface, they are impermeable to fluids, under most conditions they are resistant to wear, and the coefficient of friction is about 0.08–0.1. They are not wetted by melts of glass, ceramic, or metal.

Applications. Flexible graphite is an extraordinarily good packing and sealing material, suited to replace asbestos. Examples are cylinder-head gaskets reinforced with metallic inserts and stuffing box seals consisting of flexible graphite tapes wrapped around the shaft of a pump or valve. Preshaped lamellar rings are used as gaskets, e.g., for exhaust system piping. Other applications are linings, blackwash, and heat shields often in combination with carbon felt sheets.

List of Works Cited

Angus, J. C. (Ed.) (1990), Workshop of the Science and Technology of Diamond Thin Films, in: *Carbon* **28**, 743.

Blackman, L. C. F. (1970), *Modern Aspects of Graphite Technology*, London: Academic Press.

Bokros, J. C. (1969), in: P. L. Walker (Ed.), *Chemistry and Physics of Carbon*, Vol. 5, New York: Marcel Dekker, p. 1.

Brooks, J. D., Taylor, G. H. (1968), in: P. L. Walker (Ed.), *Chemistry and Physics of Carbon*, Vol. 4, New York: Marcel Dekker, p. 243.

Bundy, F. P., Strong, H. M., Wentdorf, Jr., R. H. (1973), in: P. L. Walker, P. A. Thrower (Eds.), *Chemistry and Physics of Carbon*, Vol. 10, New York: Marcel Dekker, p. 213.

Delle, W., Koizlik, K., Nickel, H. (1979), *Graphitische Werkstoffe für den Einsatz in Kernreaktoren*, Part 1, München: Verlag K. Thiemig.

Delle, W., Koizlik, K., Nickel, H. (1983), *Graphitische Werkstoffe für den Einsatz in Kernreaktoren*, Part 2, München: Verlag K. Thiemig.

Field, J. E. (Ed.) (1979), *The Properties of Diamond*, London: Academic Press.

Fitzer, E., (1980), *Angew. Chem.* **92**, 375–386.

Fitzer, E., Heine, M. (1986), in: *Ullmann's Encyclopedia of Industrial Chemistry*, Vol. A 11, Weinheim: VCH Verlagsgesellschaft, p. 42.

Fitzer, E., Müller, K., Schäfer, W. (1971), in: P. L. Walker (Ed.), *Chemistry and Physics of Carbon*, Vol. 7, New York: Marcel Dekker, p. 237.

Franklin, R. (1951), *Acta Crystallogr.* **4**, 253–261.

Gmelins Handbuch der Anorganischen Chemie (1967), System no. 14, Carbon, Sec. B, Part 1, Heidelberg: Springer.

Gmelins Handbuch der Anorganischen Chemie (1968), System no. 14, Carbon, Sec. B, Parts 2 and 3, Heidelberg: Springer.

Goresey, A. el, Donnay, G. (1968), *Science* **161**, 363–364.

Haaland, D. M. (1976), *Carbon* **14**, 357–361.

Hérold, A. (1979), in: F. Lévy (Ed.), *Intercalated Layered Materials*, Dordrecht: D. Reidel Publ.

Herron, R. H., Beechan, C. R. (1981), *Am. Ceram. Eng. Sci. Proc.* **3**, 1126–1140.

Holm, R. (1967), *Electric Contacts Handbook*, Berlin: Springer Verlag.

Hüttinger, K. J. (1972), *Chem. Ing. Tech.* **44**, 849–853.

Ishikawa, T., Nagaoki, T. (1984), *Recent Carbon Technology*, Cleveland: JEC Press.

Jenkins, G. M. (1974), in: P. L. Walker, P. A. Thrower (Eds.), *Chemistry and Physics of Carbon*, Vol. 11, New York: Marcel Dekker, p. 189.

Jenkins, G. M., Kawamura, K. (1976), *Polymeric Carbons—Carbon Fibre, Glass and Char*, Cambridge: Cambridge University Press.

Keller, F., Fischer, W. K. (1981), in: J. E. Andersen (Ed.), *Proceedings of Technical Sessions: Light Metals 1982*, New York: AIME, p. 729.

Kelly, B. T. (1981), *Physics of Graphite*, London: Applied Science Publ.

Kroto, H. (1988), *Science* **242**, 1139–1145.

Marsh, H., Walker, P. L. (1979), in: P. L. Walker, P. A. Thrower (Eds.), *Chemistry and Physics of Carbon*, Vol. 15, New York: Marcel Dekker, p. 229.

Moore, A. W. (1973), in: P. L. Walker, P. A. Thrower (Eds.), *Chemistry and Physics of Carbon*, Vol. 11, New York: Marcel Dekker, p. 69.

Mrozowski, S. (1956), *Proceedings of the First and Second Conferences on Carbon*, Baltimore, MD: Waverly, p. 195.

Nightingale, R. E. (Ed.) (1962), *Nuclear Graphite*, New York: Academic Press.

Pacault, A. (Ed.) (1965), *Les Carbones*, Vols. 1 and 2, Paris: Masson et Cie.

Reynolds, W. N. (1968), *Physical Properties of Graphite*, New York: Elsevier.

Ruland, W. (1968), in: P. L. Walker (Ed.), *Chemistry and Physics of Carbon*, Vol. 4, New York: Marcel Dekker, p. 1.

Shobert, E. I., II (1965), *Carbon Brushes*, New York: Chemical Publ.

Spain, I. L. (1973), in: P. L. Walker, P. A. Thrower (Eds.), *Chemistry and Physics of Carbon*, Vol. 8, New York: Marcel Dekker, p. 1.

Ubbelohde, A. R., Lewis, F. A. (1960), *Graphite and its Crystal Compounds*, Oxford: Clarendon Press.

Watanabe, N., Nakajima, T., Tonhara, H. (1988), *Graphite Fluorides*, Amsterdam: Elsevier.

Weng, T.-L., Seldin, E. J. (1972), *Carbon* **15**, 391–398.

Whittaker, A. G. (1979), *Carbon* **17**, 21–24.

Whittaker, A. G., Neudorfer, M. E., Watts, E. J. (1983), *Carbon* **21**, 597–599.

Wu, T. C., Vogel, F. L. (1982), *Carbon* **20**, 225–230.

Further Reading

Blackman, L. C. F. (1970), *Modern Aspects of Graphite Technology*, London: Academic Press.

Fialkov, A. S. (1979), *Graphitic Materials* (in Russian), Moscow: Energia.

Field, J. E. (Ed.) (1979), *The Properties of Dia-*

mond, London: Academic Press.

Ishikawa, T., Nagaoki, T. (1984), *Recent Carbon Technology,* Cleveland: JEC Press.

Jenkins, G. M., Kawamura, K. (1976), *Polymeric Carbons—Carbon Fibre, Glass and Char,* Cambridge: Cambridge University Press.

Kelly, B. T. (1981), *Physics of Graphite,* London: Applied Science Publ.

Mantell, C. G. (1968), *Carbon and Graphite Handbook,* New York: Interscience Publ.

Marsh H. (1989), *Introduction to Carbon Science,* London: Butterworth.

Tschalych, J. F. (1961), *Herstellung von Kohle- und Graphitelektroden,* Leipzig: VEB-Verlag für Grundstoffindustrie.

Ubbelohde, A. R., Lewis, F. A. (1960), *Graphite and its Crystal Compounds,* Oxford: Clarendon Press.

ERRATUM

CARBON MATERIALS (Vol. 3, pp. 21–40). OTTO J. VOHLER, FERDINAND VON STURM, and ERHARD WEGE.

Page 37. The entries in Table 4 were incorrectly arranged. The correct form follows.

Table 4. Mechanical properties of commercial carbon fibers.

	Density (g/cm^3)	Tensile strength (kN/mm^2)	Young's modulus (kN/mm^2)	Breaking strain (%)
Pan-based				
HT (high strength)	1.75	3	200	1.5
HM (high modulus)	1.95	2	400	0.6
Pitch-based				
Anisotropic MPP (mesophase based)	1.95	2	400	0.8
Isotropic	1.60	1	70	3.5

CARTOGRAPHY AND GEOGRAPHIC INFORMATION SYSTEMS*

DAVID M. MARK, *National Center for Geographic Information and Analysis, State University of New York at Buffalo, Buffalo, New York, U.S.A.*

	Introduction	41
1.	**Cartography**	43
1.1	History of Cartography	43
1.2	Map Projections	44
1.2.1	Properties of Map Projections	44
1.2.2	Classification of Map Projections	45
1.3	Reference Maps	50
1.3.1	Topographic Maps	50
1.3.2	Road Maps	54
1.3.3	Charts	55
1.3.4	"Base" Maps	55
1.4	Thematic Maps	56
1.4.1	Isoline Maps	56
1.4.2	Choropleth Maps	56
1.4.3	Proportional Point Symbols	57
1.5	Computers and Cartography	59
1.6	Visualization in Geographic Computing	59
1.7	Summary	59
2.	**Geographic Information Systems**	59
2.1	Data Models, Data Structures, and Algorithms	59
2.1.1	Data Models	59
2.1.2	Data Structures	60
2.1.3	Algorithms	60
2.2	Geographical Analysis and Spatial Statistics	61
2.3	User Interfaces and Human Factors	61
2.3.1	Forms of Human-Computer Interaction	61
2.3.2	Who are the "Users"?	62
2.4	Social, Economic, Institutional, and Legal Issues	62
3.	**Summary and Prospects**	63
	Glossary	63
	Works Cited	64
	Further Reading	65
	Journals	65
	Books	65

INTRODUCTION

This article provides an overview of cartography, the art and science of maps and mapping. It also reviews geographic information systems (GISs), a "new cartography" that has developed over the last three decades from cartography as well as from other disciplines.

Definition of cartography. The British Cartographic Society has proposed two definitions of cartography (Harley, 1989, p. 2):

"Cartography is the art, science and technology of making maps."

"Cartography is the science and technology of analyzing and interpreting geographic relationships, and communicating the results by means of maps."

They state that the former is "for use in communication with the general public," whereas the latter is for professional or practicing cartographers. Harley (1989) has spoken strongly against such a dichotomy, which even if accepted begs the question for audiences of researchers and practitioners from other disciplines and professions, such as the majority of readers of this encyclopedia.

Since cartography is a rather technical term, it is reasonable for its practitioners and their associates to provide definitions. But

---

*This article represents a contribution of the National Center for Geographic Information and Analysis, supported by a grant from the National Science Foundation (NSF) (No. SES-88-10917); support by the NSF and the National Center for Geographic Information and Analysis during the writing of this article is gratefully acknowledged.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

note that the word "map," which is an essential component of both of the above definitions, is not defined. As might be expected, cartographers have at times attempted to redefine the term "map" to fit their technical needs. However, "map" seems to be a basic-level word in the English language, and its definition really lies with the native speakers of that language [see the works of Rosch (1973, 1978) and of Lakoff (1987) for a discussion of "basic-level" terms and concepts].

Recently, Vasiliev and others (1990) have examined the meaning of "map" using the conceptual framework of radial categories (see Lakoff, 1987). Data gathered by Vasiliev and her co-authors suggests that the prototypical map is presented on a flat or folded piece of paper, and depicts part of the Earth's surface. Features of the Earth are represented on the map by symbols that have been crafted by a human operator, and the map bears a scale, a north arrow, and a legend. Almost any English-speaking observer will consider such a thing to be a map. Other maps and maplike things are related to such a prototypical map by departures of various kinds. For example, a globe does not have the flatness property, and so a few subjects consider that globes are not maps. An image on a computer screen may be judged less of a map because it is not on paper, and has been termed a "virtual map" because it does not have a "permanent tangible reality" (Moellering, 1984), whereas prototypical maps do. At a greater distance from the prototype, most people do not consider a printed verbal script of how to get from one place to another to be a map; likewise, a digital data file containing coordinates from which a graphic map could be plotted is not a map to most people. However, they certainly have some similarities to maps, and some academics consider them to be maps. Much closer to the border between maps and non-maps are air photographs and satellite images. Since they are not crafted, selected, and drafted, some people do not consider them to be maps, whereas others do. Such maplike forms are treated elsewhere in the *Encyclopedia*. The first part of this article will concentrate primarily on prototypical maps.

Over the last century or two, cartography has diverged in a number of ways. Two fundamentally different types of maps are *reference* maps, which attempt to portray concrete, measurable features of the Earth's surface as accurately as possible, and *thematic* maps, which portray georeferenced data, often rather abstract, in a way that allows people to comprehend patterns and distributions. These kinds of maps are sufficiently different that they will be treated in separate sections below.

Maps have traditionally fulfilled at least three roles. First, maps were used as a sort of analog data base, as a device for cataloging geographic information, for storing information about relative locations of things. A second function, related to the first, is to support navigation and wayfinding; one can speculate that the very first maps were drawings in the sand or mud, intended to guide others to sources of water, other resources, or danger. And thirdly, maps have served as what would now be called visualization aids, helping map users to grasp spatial patterns and to make conjectures about spatial relationships and processes.

Technological change in the second half of the twentieth century has led to a change in the role of maps. Computers will in the future dominate in the area of spatial data storage; geographic information systems (GISs) have this as a central part of their utility. And on-board microcomputers already dominate in aircraft navigation, and can be expected to play an increasing role in vehicle navigation for automobiles as well. Thus, as cartography and GIS diverge, cartography's chief role may well be as a visualization aid.

Geographic Information Systems. Geographic information systems represent an important area of computer technology that is strongly related to cartography and mapping. Currently, a geographic information system is usually defined to be a computer system for the storage, retrieval, manipulation, and display of geographic data. But technically, such a definition would include all computer mapping software. Experts seem to agree that the essence of GISs is the ability to conduct simple analyses, especially the logical (Boolean) overlay of two or more spatial distributions. Integration of analytical techniques into a GIS framework is an important element in the research agenda for geographic information systems. But, even if cartography is narrowly defined, graphic map display is a common form of GIS output, and so the links to cartography are inescapable. Furthermore, one of the several roots, or academic "ancestors," of GISs is the computer-assisted cartography that developed in the 1950s and 1960s.

Relation to Academic Disciplines. Cartography as a profession stands in its own right. However, cartography is rarely seen as an academic discipline, in the form of a "Department of Cartography" at a college or university. Part of the essence of maps is their depiction of geographic phenomena and their distributions and locations; part of their essence is their graphic representation *per se*; and part of the essence of many maps is their relation to precise measurements and plotting of geographic objects. Thus it is not surprising that cartography finds a home in several disciplines. In the United States, Canada, and many other countries, most cartographers are found, and trained, in geography departments. However, there are cartography programs in surveying engineering, civil engineering, geodetic sciences, and other fields. In many other countries, a place in the engineering sciences dominates, and cartography and related courses are even found in computer science departments.

The closely related field of GISs is becoming a fundamental research tool in many disciplines, especially in the natural sciences. And despite the fact that most GIS courses are taught within geography departments, GIS is being taught in many academic settings, including those mentioned above, as well as geology and earth sciences, landscape architecture, environmental sciences, forestry, and many others. This trend toward disciplinary diversification of GIS training and research can be expected to increase, especially in academic institutions without geography departments.

1. CARTOGRAPHY

1.1 History of Cartography

Maps have been used by human beings for a very long time. Perhaps the oldest known map to have survived to the present day is one from central Italy that is believed to have been made about 5000 B.C. (Abler, Adams, and Gould, 1971, p. xix). The map was carved on a large rock overlooking a neolithic village. Thrower (1972) showed maps on clay tablets, dating from Mesopotamia, about 2500 B.C. This is about the same time and place as that from which earliest written versions of human natural language survive, but this probably does not indicate a simultaneous development of written language and of mapping. Rather, it suggests that when clay tablets were developed as a new technology for the representation and storage of information, that technology was used to represent both linguistic and graphic information. Similarly, the printing press was used for mass production of printed maps very soon after its use for text.

Geographical knowledge blossomed in ancient Greek times, and probably was accompanied by maps. However, few maps from that period have survived. One noteworthy accomplishment of the period was the first accurate estimate of the size of the Earth, made by Eratosthenes around 250 B.C.; Eratosthenes used the difference in the angle between the noonday sun at two sites of known distance apart to estimate the circumference of the Earth to within about 1%. But later, Europe lagged behind in cartographic development. Relatively modern-looking reference maps were produced in China around 1100 A.D., and the first known paper map was made in China around 1155 A.D., some three centuries before paper maps were made in Europe (see Fig. 1). Geographically accurate maps were also being made in the Arabic world at this time. In contrast, European maps of this period were as much cosmographic statements, typically centered on Jerusalem and with east at the top, as they were attempts to accurately portray known geographic facts (see Fig. 2). From ancient Greek times until the Renaissance, the cardinal directions and latitude could be determined in clear weather from the observation of the sun and stars; however, in cloudy conditions, only "dead reckoning" was available before the widespread adoption of the magnetic compass.

Between 1300 and 1600 A.D., European cartography developed rapidly, in conjunction with the great navigators of Spain, Portugal, England, and other European cultures. Mercator's world map projection, published in 1569, was a landmark in the refinement of rules for transforming locations from the globe to the plane (see Sec. 1.2). Many of the innovations that began what we might call "modern" cartography were in fact developed by someone who would commonly be considered to be a physicist and astronomer: Ed-

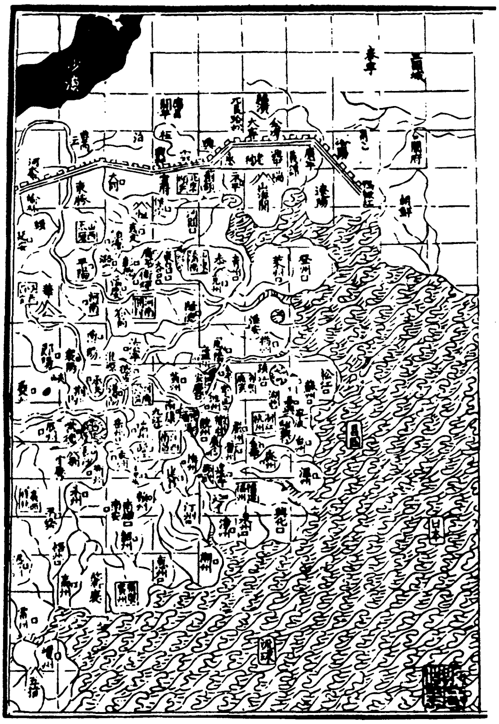

FIG. 1. Atlas map of eastern China, showing part of the Gobi (solid black area), the Great Wall, and the China Sea.

mund Halley, who worked with Isaac Newton in the late 17th and early 18th centuries in England. Halley's work at the Greenwich observatory included applied work for the British Navy. Within this context, Halley produced a generalized map of world wind patterns (1686) (Fig. 3) and another generalized world map of the divergence between magnetic compasses and true north (1701) (see Fig. 4) (Thrower, 1972, pp. 65–68). The wind map is generally considered to be the first meteorological chart, and perhaps also the first thematic map; the magnetic map, using isogones, is the first known published use of isolines to portray statistical surfaces.

From 1700 on, mapping progressed steadily and continuously toward its present state. Perhaps the greatest innovations since 1700 are those involving the digital computer. For a much more detailed discussion of the history of cartography and its relation to the development of science and culture, the reader is referred to Thrower's (1972) book.

1.2 Map Projections

Most maps are presented on flat (planar) media such as pieces of paper, or computer display screens. (The chief exceptions are globes.) However, it is impossible to flatten out a portion of the surface of a sphere without some distortion or stretching. A *map projection* is a formula or geometric construction procedure for transferring locational information from the spheroid which approximates the Earth's surface to a plane. Most map projections either preserve some particular geometric property of interest exactly (while distorting other aspects), or represent a compromise which tries to minimize several different distortions simultaneously, without completely eliminating any of them.

1.2.1 Properties of Map Projections. One property of considerable interest in mapping is the preservation of relative areas. If regions of equal area on the Earth's surface always have equal areas on some particular map projection, regardless of shape or location, then that projection is referred to as an *equal-area map projection*. Conventional cartographic wisdom suggests that equal-area projections should be used as base maps for the plotting of most thematic information, so that the areas represented are not distorted. But many commonly used world maps are not equal-area, and often present polar areas at relatively larger scale than the equatorial areas; this has led to systematic distortions of people's "mental maps."

A second property, also of great interest, is the representation of angles and shapes. If all

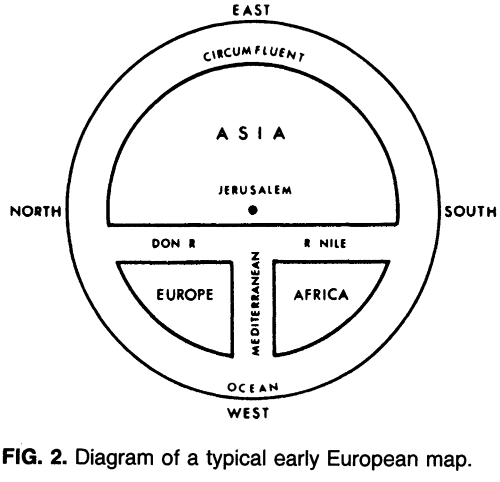

FIG. 2. Diagram of a typical early European map.

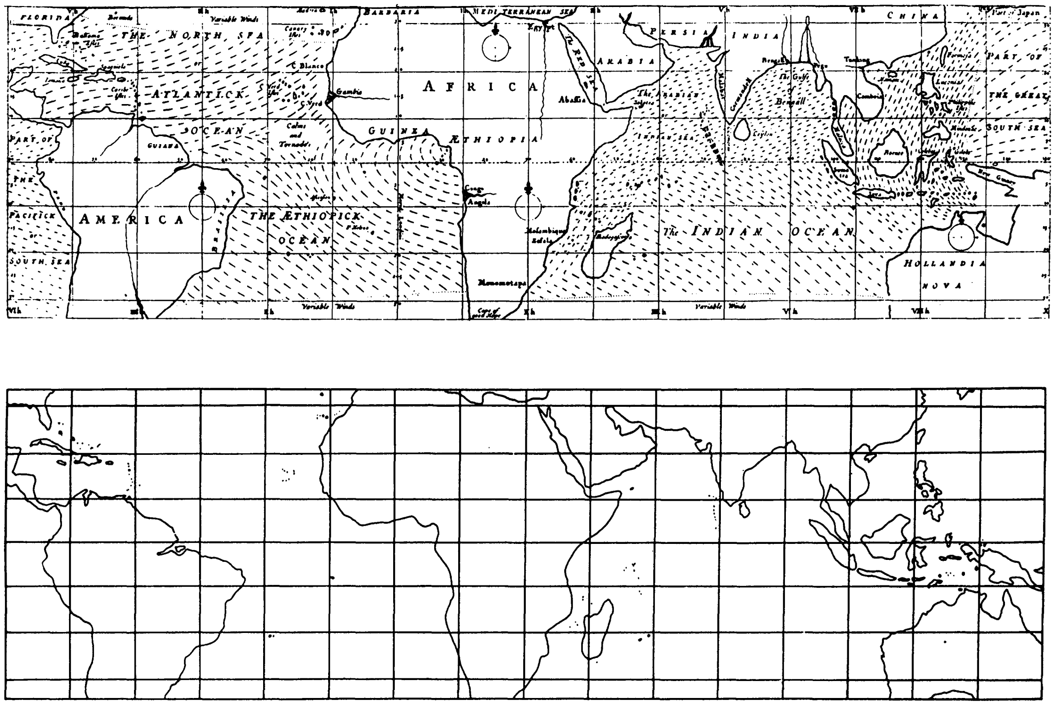
FLORIDA
THE NORTH SEA
Variable Winds
BARBARIA
MEDITERRANEAN SEA
Egypt
PERSIA
INDIA
CHINA
Part of Japan
C. Blanco
Gambia
C. Verd
ATLANTICK
AFRICA
The Red Sea
ARABIA
Abassia
The Arabian
Bengall
PART OF
THE GREAT
OCEAN
GUIANA
Calms and Tornados
GUINEA
ÆTHIOPIA
Ceylon
Borneo
SOUTH SEA
PART OF
THE
PACIFICK
AMERICA
THE ÆTHIOPICK OCEAN
Congo
Angola
Madagascar
The INDIAN OCEAN
Java
New Guinea
SOUTH SEA
Monomotapa
HOLLANDIA NOVA
Variable Winds
Cape of good Hope

FIG. 3. Chart of the trade winds by Edmund Halley (1686). From the *Philosophical Transactions* (above); (below) a map on the same scale and projection to indicate a modern delineation of the world's coastlines, for comparison.

angles on the Earth's surface are preserved on the map, then that map is said to be a *conformal map projection.* Conformal maps are useful for some navigation tasks, and furthermore, if all angles are preserved, then so are the shapes of relatively small forms or outlines. (Larger shapes may still be distorted on a conformal map by changes in the local map scale.)

A map that was both equal-area and conformal would be highly desirable. However, it was proved by the French mathematician Tissot that such a projection cannot exist. The properties "equal-area" and "conformal" are mutually exclusive, because of the necessity of some distortion as the geometry on the surface of a sphere or spheroid is transferred to the plane.

Other map properties could be of interest in choosing a projection. The shortest surface distance between two points on the surface of a sphere is defined by a segment of a *great circle.* (A great circle is a circle formed by the intersection of the surface of a sphere with a plane through its center.) A map on which all great circles appear as straight lines would be very valuable in navigation. Another property would be preservation of relative distances among points, but this is impossible in general; however, projections which preserve distances to one or two central points of interest can be constructed.

1.2.2 Classification of Map Projections.

Map projections are called "projections" because many of them could, in principle, be constructed by shining a light through a transparent globe, casting an image onto some other surface, and then flattening that surface. A *developable surface* is one which can be flattened without distortion; commonly used developable surfaces are cones, and two special cases of the cone, namely the plane and the cylinder. Map projections originally were classified according to the nature of the developable surface involved in their construction. Still today, the terms *cylindrical, conic,* and *planar* (or *azimuthal*) are used for three groups of map projections.

Three of the *azimuthal projections* could, in principle, be constructed by the projection method. The planar surface is placed tangent

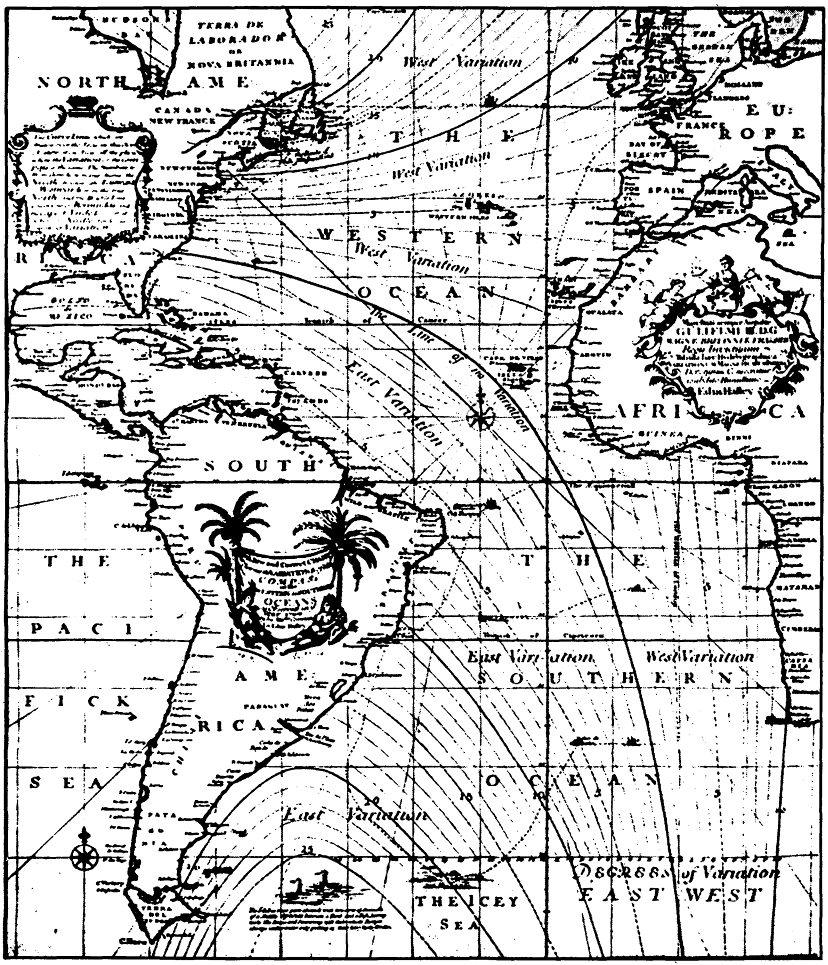
TERRA DE LABRADOR & NOVA BRITANNIA
NORTH AME
CANADA NEW FRANCE
West Variation
THE
West Variation
WESTERN
West Variation
OCEAN
Line of no Variation
East Variation
SOUTH
THE
PACI
FICK
SEA
AME
RICA
EU
ROPE
FRANCE
SPAIN
AFRICA
THE
East Variation
West Variation
SOUTHERN
OCEAN
East Variation
THE ICEY SEA
of Variation
EAST
WEST

FIG. 4. Isogonic map of the Atlantic by Edmund Halley.

to the sphere at some point, which will form the center of the resulting map. Then the light source is placed along the line from the tangent point through the center of the sphere. If the light source is placed at infinite distance, the result is the *orthographic projection* (Fig. 5). This projection has no important mathematical properties, but it "looks good." Since it appears as would a picture of the globe, it is useful as an illustration, or as a location key map at a hemispherical scale. If the light source is on the surface of the sphere at a point directly opposite the tangent point, then the *stereographic projection* results (Fig. 6). The stereographic is the only conformal projection in the azimuthal family. The stereographic projection is not commonly used in cartography, but is important in geology and crystallography, for plotting orientation data. (The property of conformality allows angles to be measured from the diagram.) The third interesting place to put the light source is at the center of the sphere. This produces the *gnomonic* projection (Fig. 7), which looks strange, but on which all great circles plot as straight lines, and *vice versa.* Thus the gnomonic is very important in air or ship navigation for planning routes over long distances.

Two other common azimuthal projections cannot be created by projecting with a light source, but can only be constructed mathematically. An *azimuthal equidistant* projection preserves relative distance to the center point; the polar case would be easily approx-

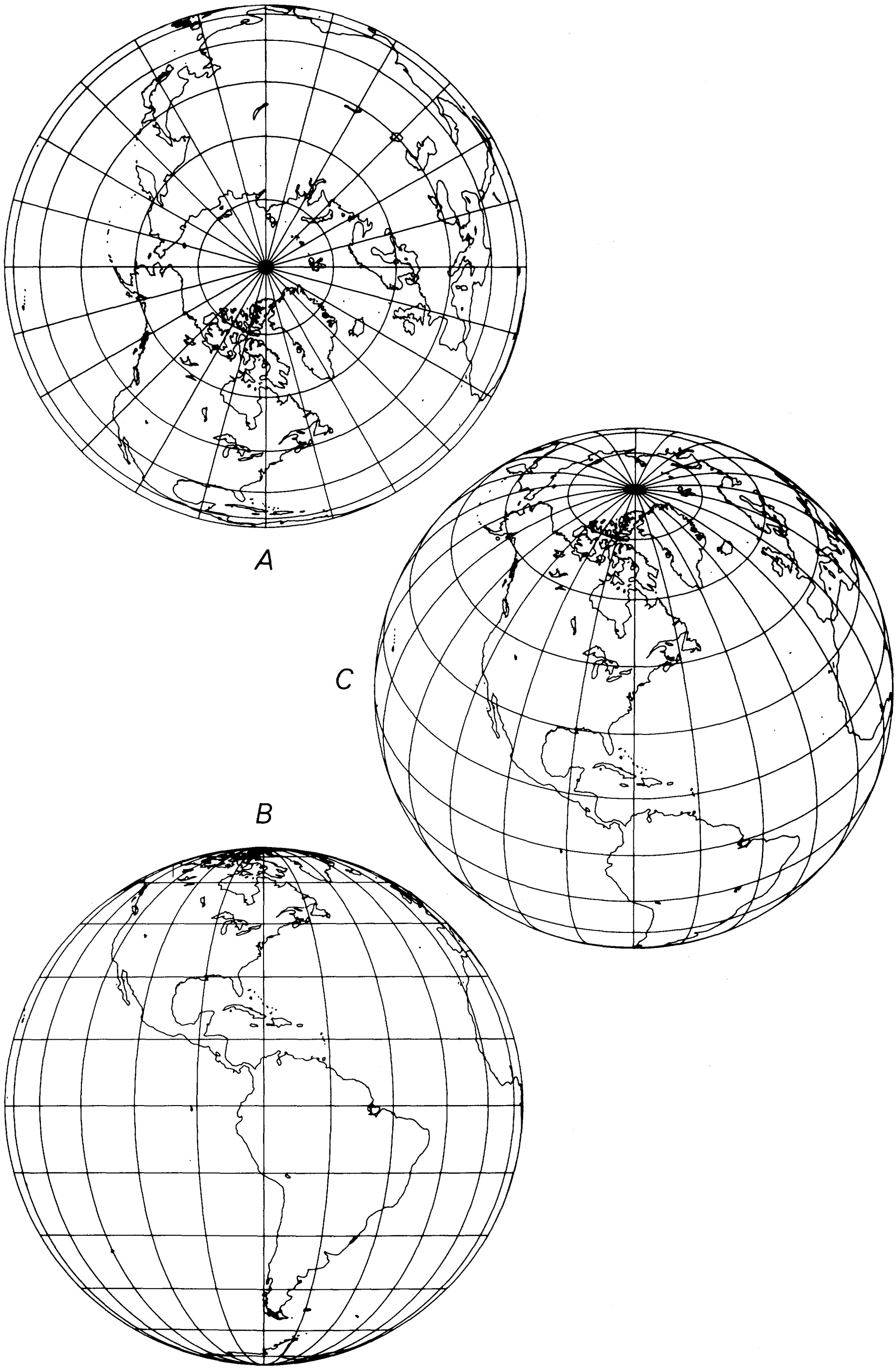

FIG. 5. Orthographic projection. (A) Polar aspect. (B) Equatorial aspect, approximately the view of the Moon, Mars, and other outer planets as seen from the Earth. (C) Oblique aspect, centered at lat. 40° N., giving the classic globelike view.

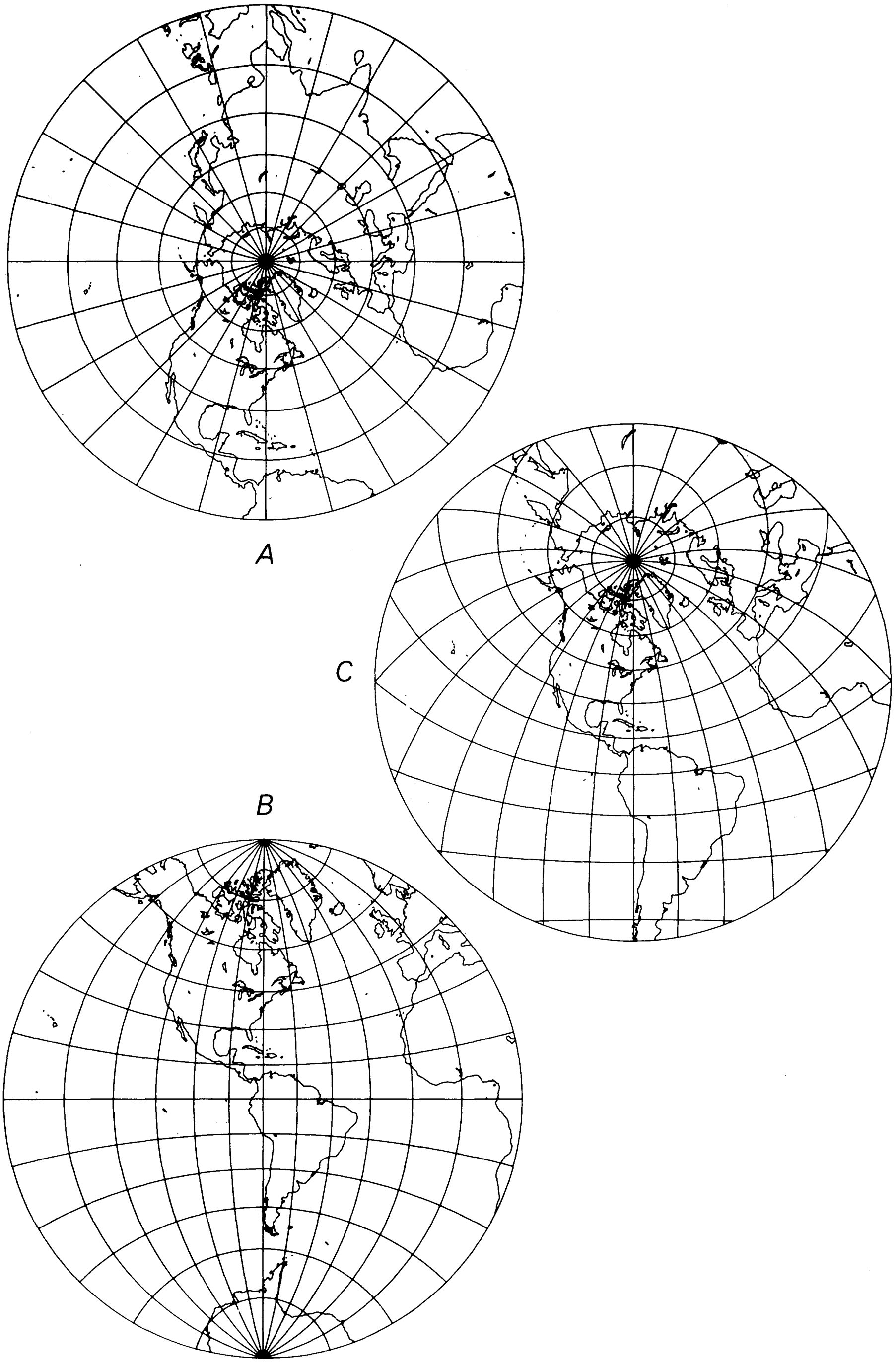

FIG. 6. Stereographic projection. (A) Polar aspect; the most common scientific projection for polar areas of Earth, Moon, and the planets, since it is conformal. (B) Equatorial aspect; often used in the 16th and 17th centuries for maps of hemispheres. (C) Oblique aspect, centered on lat. 40° N. The stereographic is the only geometric projection of the sphere which is conformal.

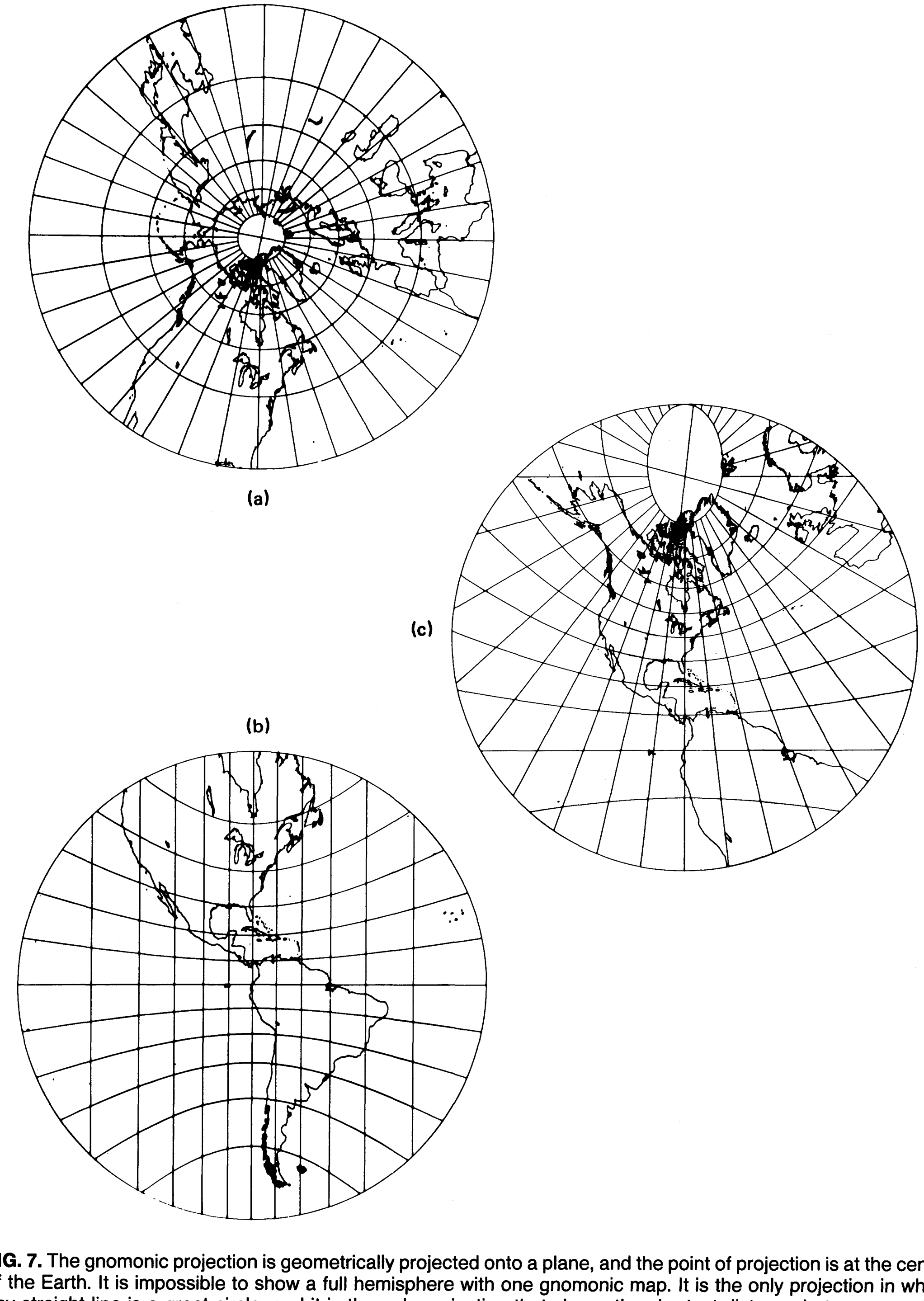

(a)

(c)

(b)

FIG. 7. The gnomonic projection is geometrically projected onto a plane, and the point of projection is at the center of the Earth. It is impossible to show a full hemisphere with one gnomonic map. It is the only projection in which any straight line is a great circle, and it is the only projection that shows the shortest distance between any two points as a straight line. Consequently, it is used in seismic work because seismic waves travel approximately great circles. The gnomonic projection is used with the Mercator projection for navigation.

imated, since the latitude could easily be converted to ground distance to the pole and scaled to the radial distance from the center of the projection (Fig. 8). An *azimuthal equal-area* projection would be slightly more complicated, since the radii must be scaled so that the areas of interlatitude bands on the map are proportional to the areas of those bands on the globe (Fig. 9).

Cylindrical projections are normally constructed with the cylinder tangent to the sphere at the equator, and are plotted with the equator extending horizontally across the centers. The most famous of these is Mercator's projection, which Gerhardus Mercator first published in 1569, and which is conformal. Mercator's projection is unique in that lines of constant compass bearing (sometimes termed "rhumb lines" or "loxodromes") are straight lines on the map (see Fig. 10). Recall that lines of longitude converge on the globe, but they appear as parallel straight lines on a cylindrical projection tangent to the equator. Thus the east-west scale necessarily increases from the equator toward the poles. The conformal property is achieved by increasing the north-south scale (i.e., by increasing the spacing of lines of latitude) at the same rate that the straightening of the meridians increases the east-west scale. A navigator could draw a straight line between origin and destination on a map using the Mercator projection, measure the bearing on the map, and then reach the destination by traveling on that bearing (barring intervening barriers, and after making appropriate magnetic corrections). Transverse Mercator projections result when some great circle other than the equator is used as the point of tangency. The universal transverse Mercator (UTM) divides the earth into 60 north-south bands (and two polar circles), and constructs a transverse Mercator projection centered on the central meridian of each 6° zone. UTM coordinates are an important standard coordinate system for topographic mapping (especially by the military) and for integrating data from disparate sources into a geographic information system.

Conic projections involve a developable surface in the form of a cone. In their simplest forms, the cone is tangent to the sphere along some small circle (often a line of latitude). In an important variation, the cone intersects the sphere along two circles. Perhaps the best known of these projections is *Lambert's conformal conic*, which is the *de facto* standard map projection for maps of the 48 contiguous United States. With the two standard parallels at 33° N and 45° N latitude, Lambert's conformal conic projection distorts areas and distances by less than ½% over the 48 contiguous states, while perfectly preserving angles because of its conformal nature (Fig. 11). Other projections in the conic family include polyconic projections, which in effect combine several conic segments, either piecewise or through a continuous change in the apex angle of the cone.

Finally, there are many important projections that cannot be cast in the traditional form of a projection onto a developable surface. These are called *mathematical* or, occasionally, *"conventional" projections*, and include such popular projections as the sinusoidal projection, the homolographic (or Mollweide's), the homolosine (Goode's), and others. Goode's projection is interrupted, meaning that continents are mapped onto components of the projection while oceans are divided, and the resulting world maps are important in atlases when land-based data are presented.

For further readings and information on map projections, see Snyder (1982).

1.3 Reference Maps

One of the basic functions of maps is to provide locational information within recognizable frames of reference. Maps that are produced primarily to serve such a function are often called *reference maps*. Possible synonyms include "general maps," and the entire class has occasionally been named after one of its more prominent members, the topographic map. What these have in common is an attempt to represent to a great extent physical features of the world, such as roads, rivers and lakes, shore lines, railways, towns and villages, parks, etc. Name labels for such features also are fundamental, since they are almost always essential in the process of relating map features to the real-world objects they represent.

1.3.1 Topographic Maps. The term "topographic" map is often used as a synonym for "contour map," and although contours are not absolutely required on a topographic map,

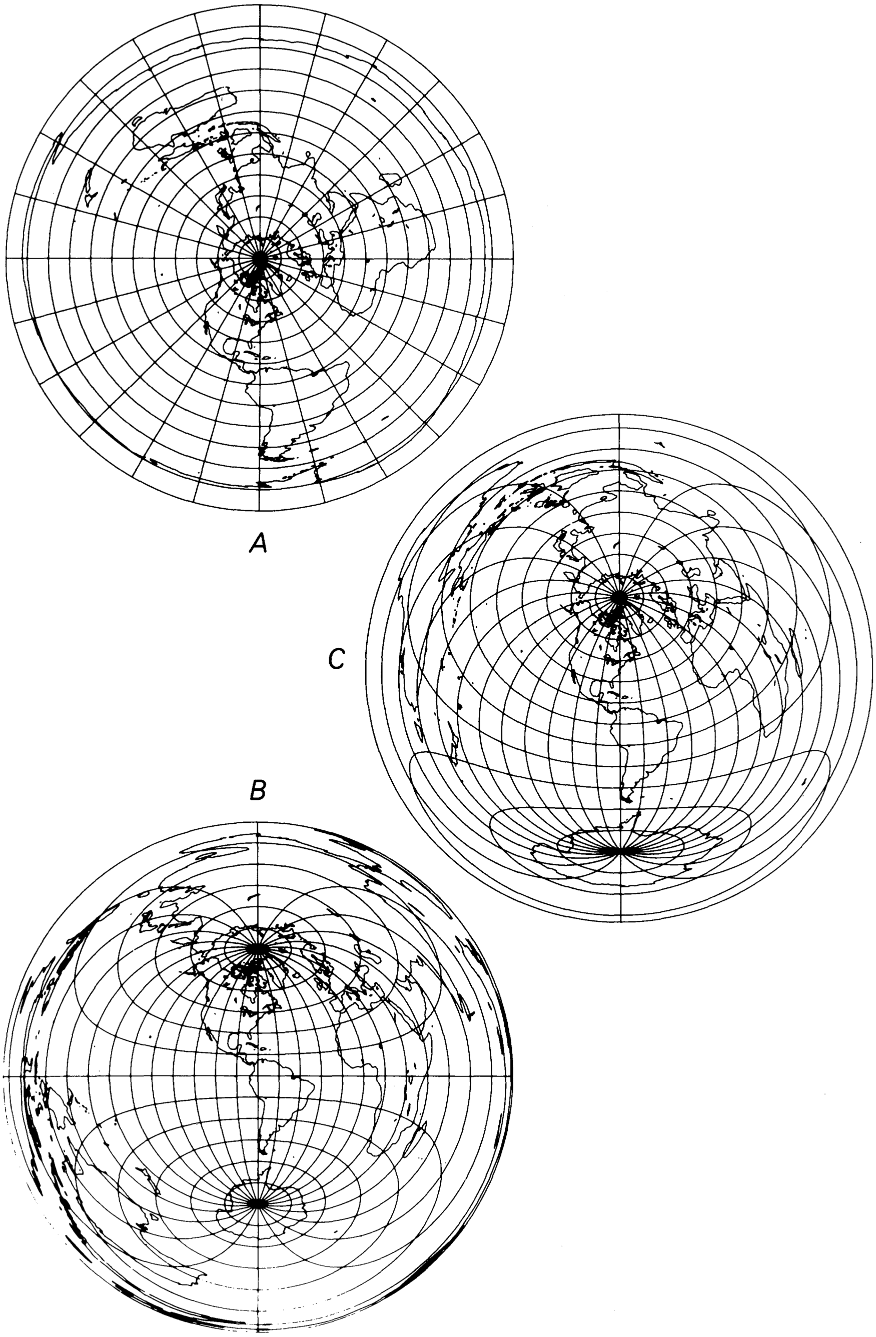

FIG. 8. Azimuthal equidistant projection. (A) Polar aspect extending to the South Pole; commonly used in atlases for polar maps. (B) Equatorial aspect. (C) Oblique aspect centered on lat. 40° N. Distance from the center is true to scale.

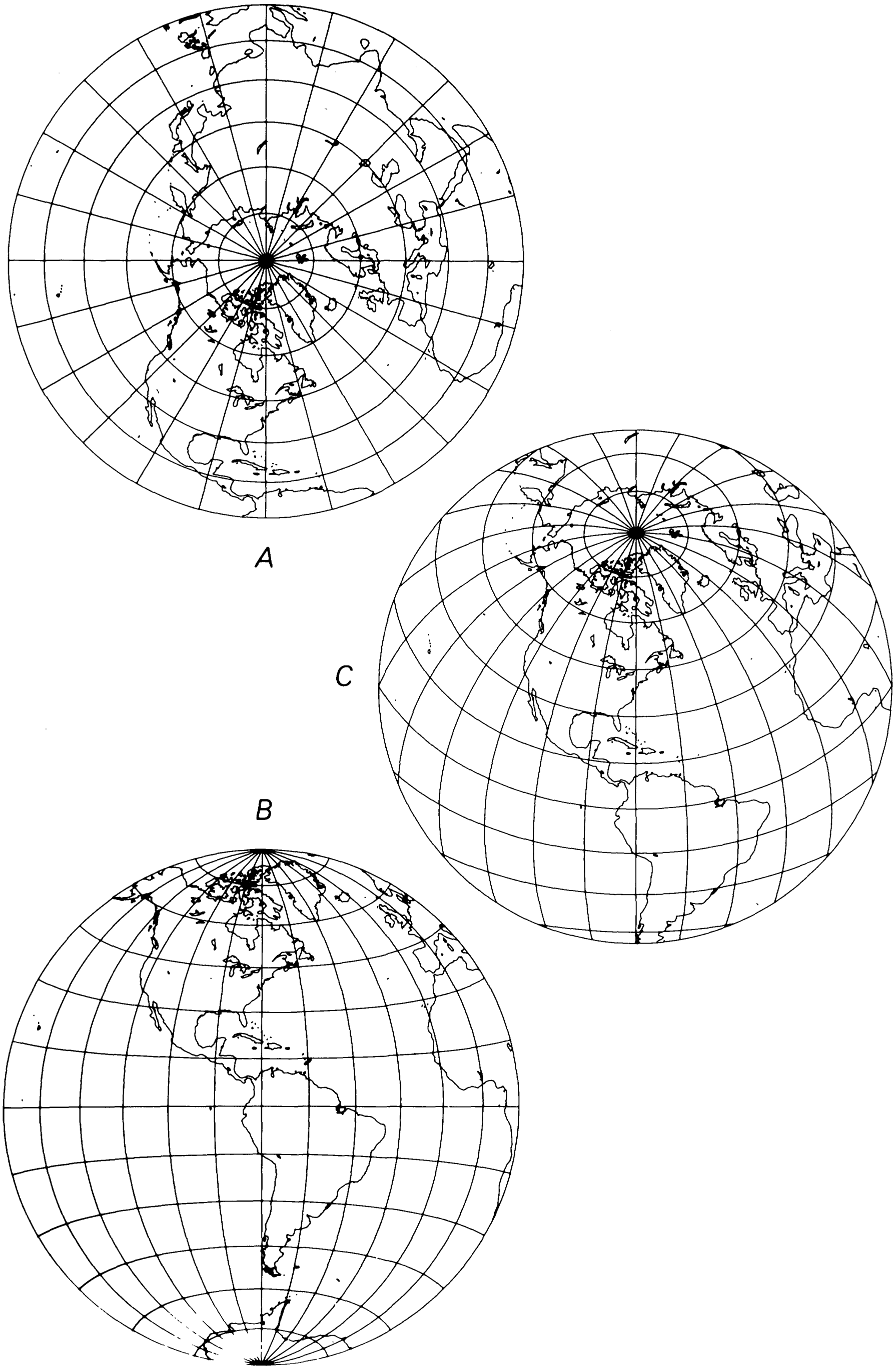

FIG. 9. Lambert azimuthal equal-area projection. (A) Polar aspect showing one hemisphere; the entire globe may be included in a circle of 1.41 times the diameter of the equator. (B) Equatorial aspect; frequently used in atlases for maps of the Eastern and Western hemispheres. (C) Oblique aspect; centered on lat. 40° N.

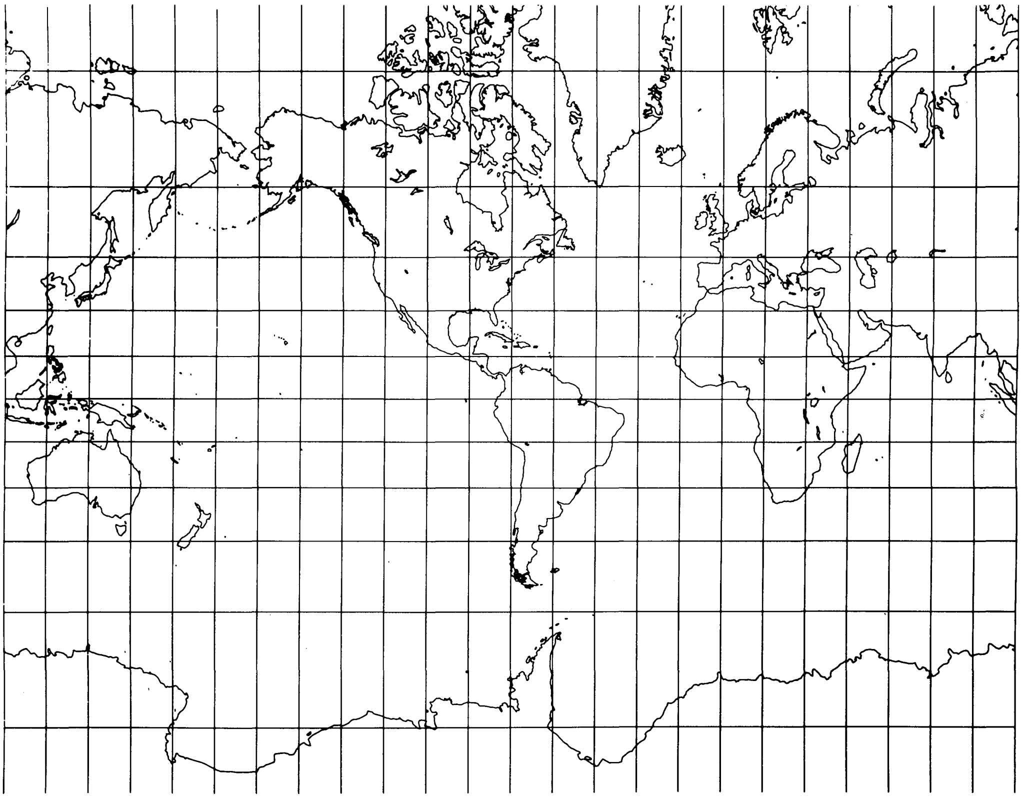

FIG. 10. The Mercator projection. All angles are shown correctly; therefore small shapes are true, and it is called conformal. Since rhumb lines are shown straight on this projection, it is very useful in navigation. It is commonly used to show equatorial regions of the Earth and other bodies.

they usually are a predominant theme. Contours are lines joining places of equal elevation above some datum, and have been the standard way of depicting elevation information on maps for almost 300 years. Topographic maps usually are dominated by cartographic features that represent visible entities on the Earth's surface: streams, lakes, and other hydrographic features; transportation features such as roads, railways, power lines, etc.; buildings and built-up areas; and political and institutional boundaries (see Fig. 12). Topographic maps are commonly produced by governmental agencies, either civilian or military, and most countries have a standard series of scales. In the U.S., the most detailed standard scale published by the U.S. Geological Survey (USGS, the civilian agency responsible for topographic mapping) is 1:24 000. At this scale, 1 in. on the map represents 2000 ft on the ground; 1 km on the ground would occupy about 4.2 cm on the map. In Canada, 1:25 000 scale mapping is conducted only in the most heavily settled southern regions, while for the rest of the country, the most detailed scale is 1:50 000. In Australia, 1:100 000 scale represents the most detailed national coverage, whereas in most European countries, maps at scales of 1:10 000 or larger often are available.

Topographic maps are used in forest management, military planning, engineering, and earth science applications. They are also used by hikers and to support a variety of outdoor activities. The digital equivalent of the elevation data depicted on such a map is called a *digital elevation model* (DEM), and occasionally is referred to as a "digital *terrain* model" (DTM). The USGS distributes 30-m-resolution grids for areas equivalent to their 1:24 000 scale map series. Digital cartographic data equivalent to other information on the topographic maps, such as hydrography and transportation, are distributed as "digital line

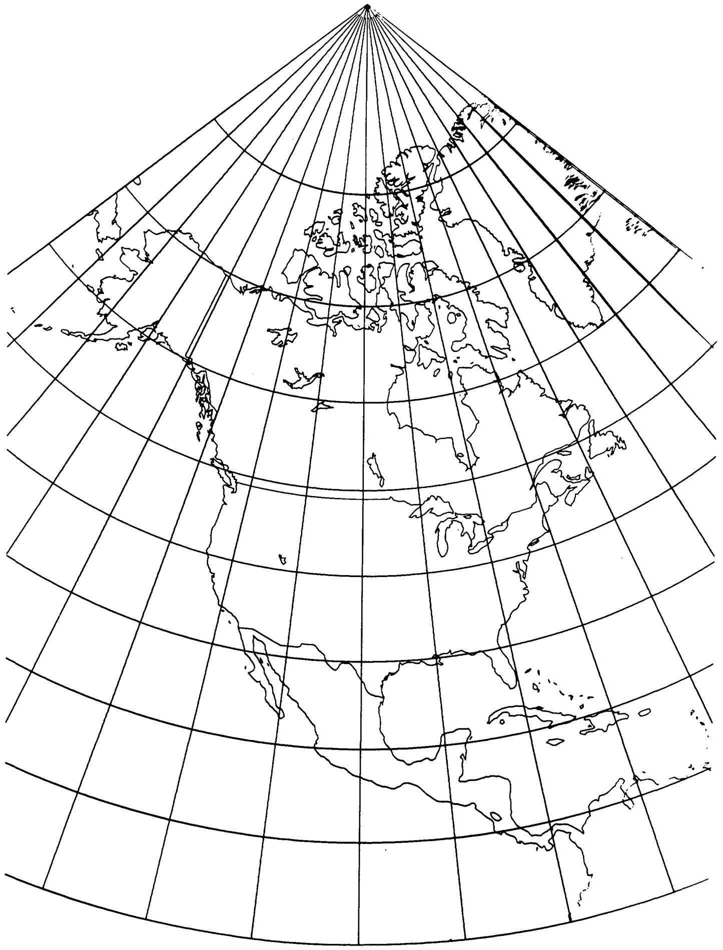

FIG. 11. Lambert conformal conic projection, with standard parallels 20° and 60° N. North America is illustrated here to show the change in spacing of the parallels. When used for maps of the coterminous United States or individual states, standard parallels are 33° and 45° N.

graphs" (DLGs). Similar digital topographic data are now available in many countries.

1.3.2 Road Maps. Road maps are perhaps the most familiar and most widely distributed maps in North America and Europe. They are an almost essential accessory in an automobile, and annual sales are very great. Unlike topographic maps, most road maps are produced by private-sector firms, and most are sold directly to the public. Exceptions are automobile clubs that provide maps to their members. Some state (provincial) and local governments provide free road maps, but this is much less common than it was 10 or 20 years ago.

Although road maps serve many purposes, their main function is to provide navigation assistance to drivers. Roads are obviously among the most important features on these maps. For city road maps (sometimes called "town plans" in British English), the roads and road names predominate, whereas for nonurban areas, names of populated places share the design spotlight with the roads. Other information is provided to serve as landmarks and to indicate potential trip destinations (parks, airports, shopping centers, etc.). Since the map cannot perform its primary function if the destination cannot be located on the map, a premium is placed on

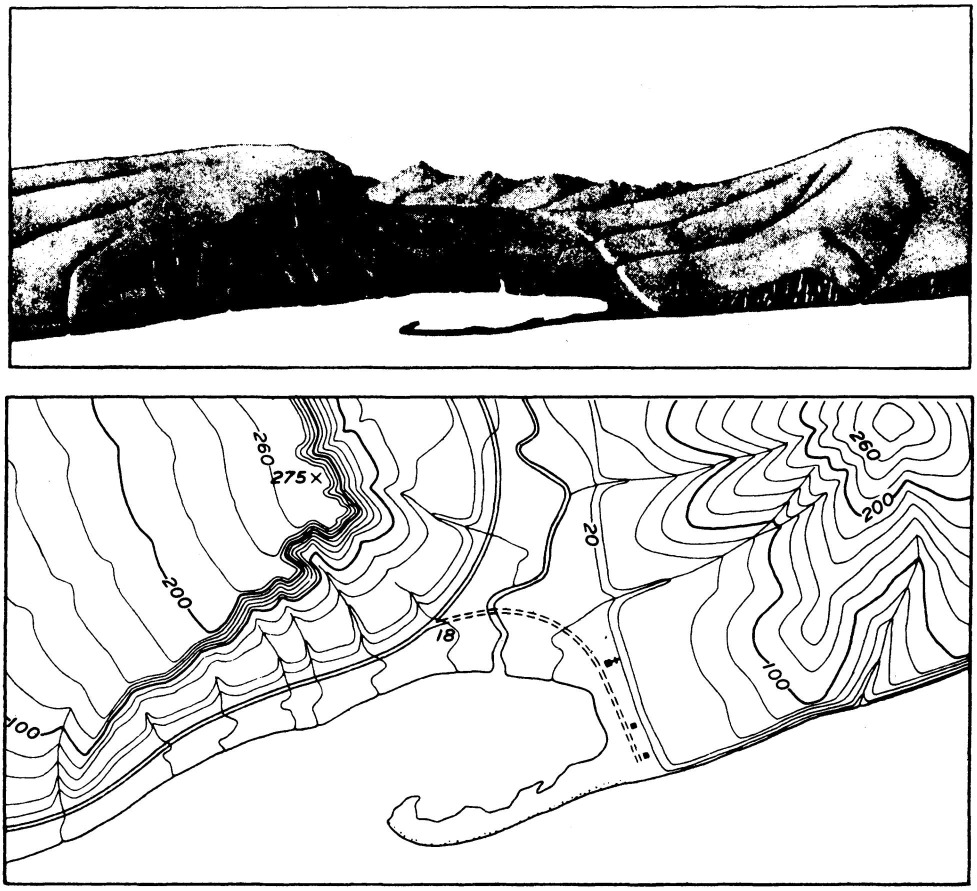

FIG. 12. In the early twentieth century the U.S. Geological Survey used this pair of oblique (above) and plan (below) views on the back of its topographic maps to explain to users the representation of terrain by contour lines (in the plan view). Plan view is also a planimetric map.

fitting as many name labels on the map as possible while retaining clarity.

In the late 1980s, several computer-based in-car vehicle navigation aid (VNA) systems were tested, and some reached the market. Such systems might eventually replace printed road maps for many situations, and can potentially be linked to traffic computers to advise motorists of construction zones, accidents, and other kinds of traffic congestion, even routing vehicles along roads to avoid such difficulties.

1.3.3 Charts. Although "chart" is an archaic synonym for any kind of map, the word is commonly used in modern English to refer to specialized maps designed to support navigation by ships and aircraft. (Note, however, that the cognate words in some other languages, such as *carte* in French and *Karte* in German, are the general words for "map" in those languages.) Ship and aircraft navigation problems are similar to each other, and distinct from road navigation, in that travel is much less restricted, and often involves physically unrestricted movement in almost any direction through or on a fluid medium containing occasional obstacles or other vehicles. Although charts are distinguished from other reference maps by purpose and have specialized information for maritime (e.g., water depths, buoys) and airborne (e.g., towers, radio beacons) navigators, they otherwise are very similar to other kinds of reference maps.

1.3.4 "Base" Maps. The term base map is used to refer to a reference map which serves as a background or reference frame for thematic information. A thematic map (see Sec. 1.4) of, say, mean January temperatures for Australia will mean very little to most map users if it is simply plotted over a latitude-longitude graticule. Most users would find it much easier to interpret the pattern if it were plotted on a base map showing the Australian coastline, the state boundaries, a few major cities, and perhaps some rivers. Base maps are designed according to principles similar to the ones that govern the production of other reference maps, except that information density is much lower to reduce interference with the thematic data. With the widespread availability of computer mapping systems, especially based on microcomputers, the use of preprinted base maps is not as important as it was a decade ago.

1.4 Thematic Maps

A thematic map portrays the geographic distribution of one or a small number of variables. Often, the phenomena shown are not directly visible in the real world, but rather are statistical in nature, determined by measurements, surveys, censuses, etc. Examples of topics for thematic maps include precipitation, gravity anomalies, crime rates, industrial production, and election results. Many atlases are collections of thematic maps.

Because many phenomena shown on thematic maps are not visible in the real world, the cartographer has a much greater range of possibilities in the way the data are represented. For example, election results, gravity anomalies, and travel times do not have color, and so in principle any color-coding system could be used on maps of these phenomena.

There is a wide range of techniques available for thematic mapping, again perhaps because the phenomena being mapped are often abstract in their geographical expression, and thus have no "natural" visual expression. In the remainder of this section, three important techniques for presenting *quantitative* information on thematic maps are presented: *isoline maps, choropleth maps,* and *proportional area symbols.* Other techniques, such as area cartograms, dot maps, and various mapping techniques for nominal data, can be found in standard cartographic and map-reading texts (e.g., Robinson, Sale, Morrison, and Muehrcke, 1984; Muehrcke, 1986; Campbell, 1984).

1.4.1 Isoline Maps. Isolines are commonly used to portray phenomena which can be represented as single-valued functions of location in geographic space. (The term *isoline* is used here in preference to the more etymologically sound *isopleth* and *isarithm,* because many cartographers use isopleth and isarithm to refer to different types of data—there is no accepted classical-root–based term that includes both.) Such data are sometimes referred to as *fields.* An isoline map is composed of a set of lines, along which the phenomenon in question has a constant value. An example is shown in Fig. 13. Although the lines are often described as *joining* points (places) of equal value, it probably is more useful to think of the isoline as *dividing* areas below some value from areas above that value. A contour map (see above) is a familiar example, and "contour" is sometimes (incorrectly) used as a synonym for isoline. The earliest known isoline map, portraying lines of equal magnetic compass declination, was included in a report by Edmund Halley, published in 1701 (Thrower, 1972, pp. 67–68). Thus, isoline maps are among the oldest forms of thematic maps.

Isoline maps can now be produced efficiently and objectively by computer programs running on microcomputers, workstations, and larger computers. Often, the programs offer the user a choice from among several interpolation methods. Isolines are often named for the phenomenon represented: iso*bars* for pressure, iso*therms* for temperature, iso*pachs* for thickness, iso*chrones* for time, etc. The name "contour" is not considered to be a synonym for "isoline," but should be considered to be the name for isolines depicting topographic elevations.

1.4.2 Choropleth Maps. The choropleth method shows quantitative data for statistical units by filling each unit with a uniform color or pattern (see Fig. 14). The term choropleth should only be used when quantitative data are collected for a set of mutually exclusive and collectively exhaustive zones, such as countries, states, counties, or census districts. Then a scheme is developed for determining a color or pattern to represent the value of the statistic for each zone. Usually, this involves the classification of the zone values into about 5–7 categories, although "unclassed" choropleth maps, in which each zone's color or

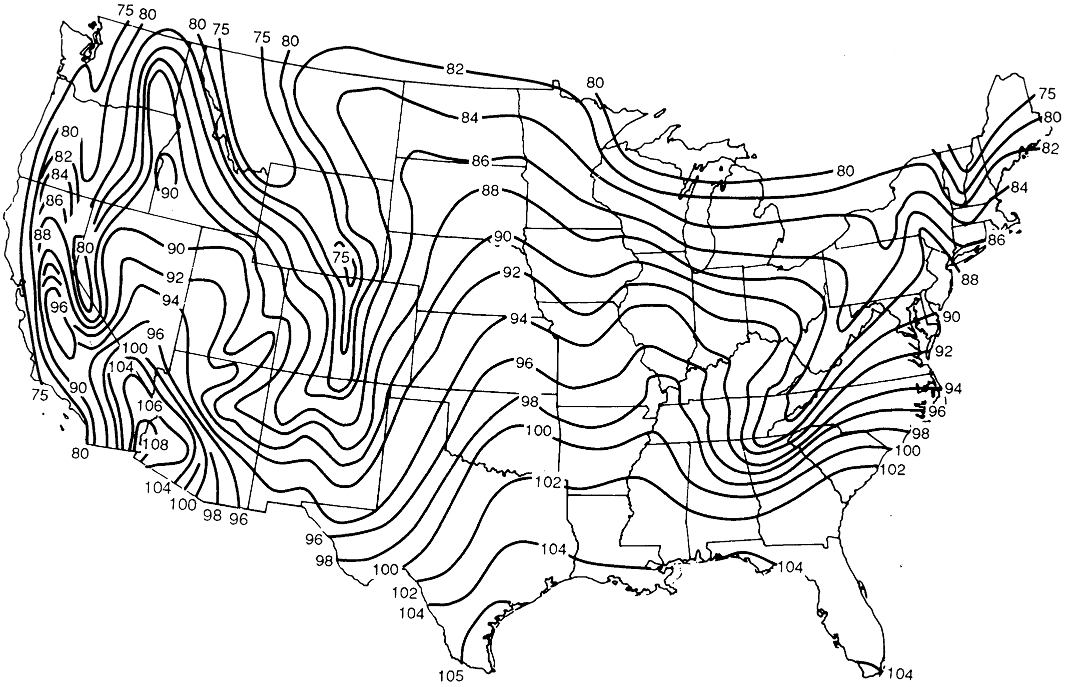

FIG. 13. Example of an isoline map. Average mid-summer noon apparent temperatures in the U.S. adjusted for relative humidity, wind, and isolation. [After R. G. Steadman, "The Assessment of Sultriness," *Journal of Applied Meteorology*, **18**, no. 7, 861–884 (1979), with supplemental data added by the National Climatic Center; U.S. Department of Commerce, National Oceanic and Atmospheric Administration, Environmental Data and Information Service, *Heat Stress*, Environmental Information Summaries C-19 (Asheville, NC: National Climatic Center, 1980), p. 4.]

pattern is a function of the zone's exact value, have also been constructed. (Such unclassed maps only became a practical possibility with the advent of computer-driven plotters in the 1960s; conventional cartographic wisdom still prefers "classed" choropleth maps.)

It is important in choropleth mapping that the quantity being mapped have the property of being ***intensive***. For physical objects, properties such as color or density are intensive, and would apply to parts if the object were cut apart; this is in contrast to an extensive property such as weight, where the value of the whole is divided among the parts. If an extensive geographic statistic, such as number of people, number of crimes, or number of cases of a disease, is plotted using the choropleth technique, then the resulting map would be more a function of the zones than of the phenomenon. A cynical police administrator could reduce the mapped intensity in a district merely by dividing it into two parts, and thus such extensive data should never be mapped by choropleth. Instead, the data should be transformed into intensive data by dividing by land area, base population, or some other appropriate item. If per capita crime rates are mapped, then splitting a district will either leave both halves in the same category as the original, or will result in the rate rising in one subzone and declining in the other. Unfortunately, no mapping program known to the author keeps such "data-type" distinctions, or is capable of detecting and eliminating such misuse; thus, if users are not aware of this, invalid and perhaps misleading maps will result.

1.4.3 Proportional Point Symbols. Often, quantitative data are available for features which are essentially points. Examples might be industrial production by plant or city, population figures by populated place, or mineral production by mine or well. If the data are of ratio type (real numbers, with "0" actually designating the absence of the at-

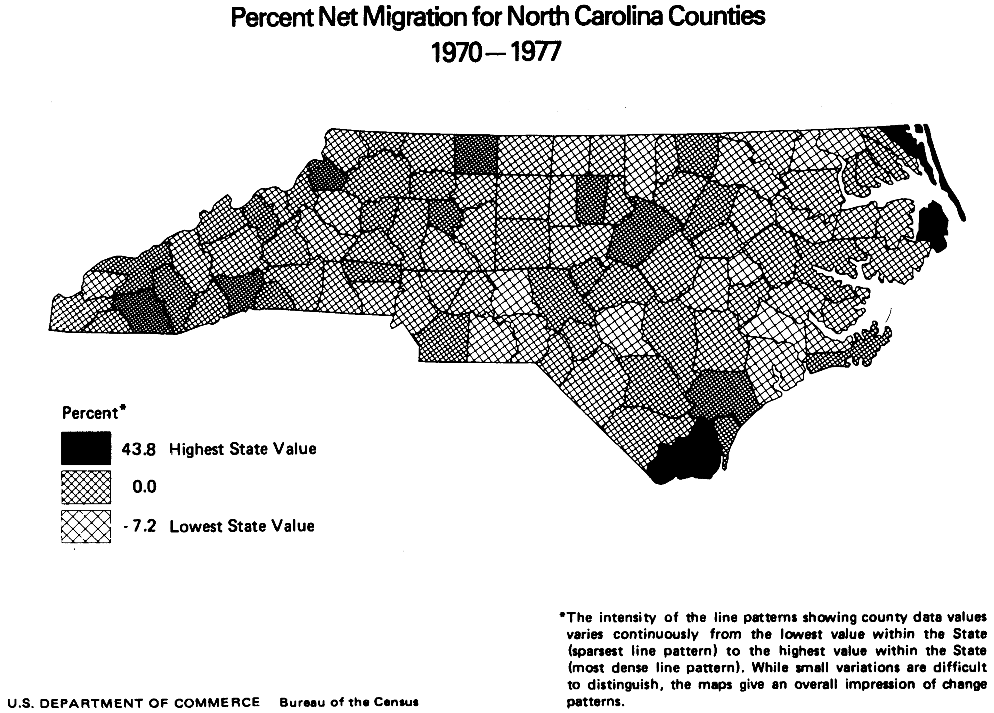

Percent Net Migration for North Carolina Counties 1970–1977

*The intensity of the line patterns showing county data values varies continuously from the lowest value within the State (sparsest line pattern) to the highest value within the State (most dense line pattern). While small variations are difficult to distinguish, the maps give an overall impression of change patterns.

U.S. DEPARTMENT OF COMMERCE Bureau of the Census

FIG. 14. Continuous-tone, nonclassed choropleth map showing percent net migration, by county. North Carolina, 1970–1977. (From U.S. Bureau of the Census, *Current Population Reports*, Series P-26, No. 77-33, September 1978, p. 9).

tribute being considered), *proportional point symbols* can be a powerful and effective component of a thematic map.

One of the most commonly used forms of proportional point symbols is the proportional circle. Geometrically, it would appear that the area of the circle should be scaled to the quantity being represented, and a circle with 3 times the radius would represent a quantity 9 times as large (3^2). However, psychophysical research conducted in the 1960s by Flannery showed that people systematically underestimate the "sizes" of larger circles compared to smaller ones. Empirically, Flannery found that scaling radii to the 0.57 power of size to be represented (rather than the 0.5, as suggested by the area principle), leads to circles whose "sizes" are perceived "correctly" by typical users. Such an adjustment for "perceptual biases" is now firmly entrenched in cartographic practice (cf. Robinson *et al.*, 1984, pp. 292–294; Campbell, 1984, pp. 300–302). This is a good example of the application of a "communication" theory of mapping: The most important thing is placing an image of geography in the mind of the map user; thus any biases introduced during map reading should be anticipated by building their inverses into map construction procedures.

Many other symbols besides circles have been used. In addition to other regular geometric forms, it is possible to use pictorial symbols, such as trees (for forestry), people (for personnel numbers), or ships (for naval strength), scaling their area according to the quantities to be represented. However, these symbols have not been investigated for perceptual "biases" similar to those discussed above for circles. It is perhaps more common for these symbols to choose a constant symbol size and then to reproduce it several to many times to indicate different quantities. And of course type size itself is often used to indicate city size; the label can be considered to be a special sort of graded proportional symbol.

1.5 Computers and Cartography

As recently as 1980, an article or section in computer-assisted cartography would have had to begin by making the case for the computer's role in map-making. However, during the 1980s, computers became so pervasive that it is hardly necessary to provide such a justification. Computers are essential to much of cartography today.

Much of computer mapping in the 1960s and 1970s used the computer as "electronic paper"; effort was almost entirely directed toward producing by computer the sorts of hard-copy graphics that had been developed and perfected in manual cartography. But even though Craig Fields challenged researchers at a 1977 Harvard conference to look "beyond electronic paper," techniques possible on the computer but impossible or impractical on a static print medium have yet to receive much attention from cartographers. Interactive maps (connected to GIS or similar databases) can be designed quite differently, placing only essential information on the graphic, and allowing the user to request additional information by commands or pointing. Similarly, animation and related techniques have great potential to show time-varying spatial data, such as atmospheric phenomena, earthquake activity, land-use change, etc. Such techniques, associated with the newly named field of *scientific visualization* (see next section), will almost certainly gain in importance in the cartography of the future.

1.6 Visualization in Geographic Computing

During the 1980s, the scientific research community "discovered" the power of graphic representation of data (see COMPUTER GRAPHICS). Whereas geography, geology, and many natural sciences have routinely used maps and graphics as tools for data exploration for many years, this is not true of all disciplines. In many areas, especially physics, simulations using supercomputers produced data at unprecedented rates and volumes, such that it was essentially impossible for human minds to comprehend them in numerical form. Similarly, space-borne earth observation platforms also were recording observations at unprecedented rates. In response to this situation, the U.S. National Science Foundation and the computer science community developed the "Visualization in Scientific Computing" (ViSC) report and initiative; see McCormick *et al.* (1987) for the report. Earth science applications figure prominently in that report.

1.7 Summary

Cartography is the science, art, and craft of making, evaluating, and using maps. Maps have been a very important part of culture for a very long time, and remain important today. Reference maps are used to represent locations of concrete objects on the Earth's surface, and thematic maps show more abstract or statistical data. Many techniques are available for thematic mapping, and not all can be applied to any data. Computers are becoming more and more important in mapping, and continue to provide challenges for cartographers to take full advantage of new technologies.

2. GEOGRAPHIC INFORMATION SYSTEMS

In their glossary, Star and Estes define a *geographic information system* as "the complete sequence of components for acquiring, processing, storing, and managing spatial data" (Star and Estes, 1990, p. 267). Often, the ability to overlay two data layers, in a procedure analogous to overlaying two paper maps on a light table, is viewed as the quintessential GIS function. Critical issues for GISs include data models, data structures, and algorithms; geographic analysis and spatial statistics; user interfaces and human factors; and social, economic, institutional, and legal issues. These will be reviewed in the following sections.

2.1 Data Models, Data Structures, and Algorithms

2.1.1 Data Models. A geographic information system may be thought of as a computerized model of some aspects of the geographical world (cf. COMPUTER DATABASES). The term *data model* is used to refer to a model of relevant aspects of the real world, which has been abstracted from that world in such a way as to retain those properties which are relevant to the current problem domain, and yet which can be implemented on a computer. Data modeling for GISs is thus an exercise in geography, plus forestry, earth science, or whatever domain the GIS is to be applied to. To some extent, a generic GIS thus is unlikely, unless it is an implementation of the union of all possible application-domain data models. However, different application domains have enough in common that generic commercial GISs are possible and useful. Nevertheless, any particular commercial GIS is better for some classes of applications than for others.

2.1.2 Data Structures. The term *data structure* is usually used to refer to the details of how the data model is implemented on a computer. One dichotomy is between raster and vector GISs. In the *raster* family of geographic information systems, geographic space is partitioned into a large number of small cells, usually of a uniform size and shape; then, for each cell, geographic attributes are recorded. In most such systems, cells are square, and often are referred to as *pixels*, a contraction of "picture elements"; in image processing, a single complete row of the image is referred to as a raster, from which this family of data structures gets its name. Raster data structures dominate in remote sensing. The essence of the raster approach is a data model based on "spatial occupancy": space is partitioned into cells, independently of theme, attributes, or objects, and then theme, attributes, or objects are observed and recorded for each cell. One problem with the raster approach is that often there is a great deal of redundancy, with many neighboring cells having identical values. Specialized methods such as quadtrees have been developed to make raster data handling more space efficient, or to give quicker response time to spatial search queries. A quadtree is a data model that divides an image into quadrants, and recursively divides each of these into subquadrants as long as they contain variation.

An alternative data modeling approach is to observe the themes, attributes, objects, or phenomena in geographic space, and then to include location as part of their representation on the computer. This defines the family of data structures commonly referred to as *vector* data structures, which often represent objects as digital versions of lines on maps (hence, "vector"). Most commercial GISs use this approach to represent geographic information. Some systems take an explicit map-like approach, conceptualizing all data as "layers," conceptually equivalent to digital maps. In other systems, the emphasis is on the representation of individual geographic objects, which can be constructed from spatial primitives that may be shared among compound objects. Of course, layers and objects can often be supported as different user views of the same database. Ideally, the system should support whatever view of the data corresponds to the way the user conceptualizes the data.

2.1.3 Algorithms. An algorithm is a sequence of steps which, when carried out in order, lead to the solution. A recipe for baking a cake is a good example of an everyday algorithm. Algorithms are essential to computing. Some of the more important algorithms correspond to basic GIS procedures. Map overlay (Boolean or logical combination of two layers) is one of the most basic GIS procedures. Another is "point-in-polygon"; the object is to determine in which of a set of plane polygons some specified coordinate pair falls—point-in-polygon would be used in a vector GIS to determine the attribute at a specified point, by first determining the geometric polygon that the point falls in, and then looking up the attribute in a database. Other basic GIS procedures include windowing (retrieving all objects within a user-defined subarea or "window") and various display procedures.

Algorithms are intimately intertwined with data structures. Algorithms for overlaying two layers are extremely simple for raster data sets using identical pixels, somewhat more complicated for specialized raster structures such as quadtrees, and very different and much more involved for vector data structures. Even within the same data structure, several different algorithms may be available to solve the same problem. One may be easiest to program, another fastest for average or typical data but occasionally very slow, and another fastest in a worst-case situation. In commercial GISs, algorithms are sometimes proprietary, making it difficult for users to know how answers to their commands or queries are being achieved.

2.2 Geographic Analysis and Spatial Statistics

Up to now, GISs have been used mainly in an inventory or archival role or in conceptually simple Boolean overlay analyses. However, much of the potential of GIS lies in combining the ability to handle large and complex data sets with sophisticated modeling and analysis techniques. Many analytical procedures were developed by geographers and other spatial analysts in the 1960s and 1970s, but were never widely applied because of problems with data availability and with applying the methods to very large data sets.

Geographic information systems offer the potential to bring these spatial analysis methods to users. Commercial GISs now often have "macrolanguages" or "tool boxes" that allow users to build and apply their own models. Soon, it is expected, vendors will begin to supply ready-made macros for common or user-requested analytical procedures. Network analysis methods can be applied to find minimum paths through networks of links and nodes, and can be incorporated into location–allocation models. Spatial decision support systems represent an important class of spatial analysis procedures that have much in common with GISs, but differ from typical GISs in the high capacity for user interaction, testing, and comparing of various scenarios, and ease of user-defined analytical models.

Spatial statistics are critical in at least two ways. In one case, they can be used to remove spatial effects in statistical models and to determine correlations and regressions that have had such spatial effects removed. [Otherwise, two autocorrelated or coherent spatial distributions will often have an apparently significant yet spurious positive (if peaks by chance coincide) or negative (if peaks are far apart) correlations.] Controlling for spatial autocorrelation effects is difficult in standard statistical packages and impractical in current GISs, but might be easier to incorporate into a GIS framework in the future. [See Griffith (1987) for an overview.]

Spatial statistics are of current practical importance in the area of error modeling and assessment (see ERROR ANALYSIS). Error in digital geographic data has often been ignored, but it has an obvious and important impact on the use of such data in decision making and in scientific research. Furthermore, GIS modeling may amplify errors or damp them out; sensitivity of GIS procedures to error, and error propagation through GISs, are important research topics. Recently, a book edited by Goodchild and Gopal (1989) has surveyed some of the critical issues.

2.3 User Interfaces and Human Factors

Any computer system is only as useful as its user interface, since that is the only way that the users can use it. In fact, Matthew McGranaghan of the University of Hawaii recently claimed in a conference presentation that, for the end user, the user interface *is* the system, since that is the only part that the user "sees." The interface must allow the user to pose commands and queries, and to get results of those commands and queries in a form consistent with what they had in mind when they asked. This requires that the user interface and the user be employing similar concepts of the problem and of the objects and operators involved.

Human-computer interaction (HCI) is the term used for a subfield of "human factors," which is concerned with evaluating how well people can use computers. It is part of engineering. *Computer-human interaction* (CHI) is the term most often used in computer science for a very similar subdiscipline. CHI represents the applied side of "cognitive science," a field which studies how humans conceptualize and reason about phenomena.

2.3.1 Forms of Human-Computer Interaction. The early days of computing required users to enter programs and data through the medium of paper tape. Many people in the GIS field began their computing experience with the next medium, punch cards (or "Hollerith cards"). Almost all computing in the late 1960s and early 1970s was "batch mode." Punch cards are almost obsolete, at least in North America and Europe, and at present and in the near future, five forms of human-computer interaction are important:

- **Command line interpreter.** This is the familiar method in which the user types commands at some prompt, and the computer then executes the command. If required parameters or arguments for the command are missing, the computer may prompt the user for them, or may give the user an error message. Almost always, each command is executed when complete (or when the user hits "return" or "enter"), and then the computer "forgets" it, requiring complete specification of the next command, even if many of the parameters or arguments are the same. Most systems require a "verb-object" syntax similar to commands or queries in English.

- **Form fill-in.** In this approach, the computer specifically prompts the user for each parameter or argument needed for the current command or query.
- **Menu selection.** Menus can be permanent on-screen features, or the now well-known "pull-down" menus of the Macintosh and many other systems. Either the cursor keys or a pointing device (such as a "mouse") is used to select from a menu of possible commands or actions. This method, and especially the pull-down form, has a certain element of direct manipulation, and often is combined with that method.
- **Direct manipulation.** In the direct manipulation approach, the user employs a pointing device (usually a mouse) to select, "grasp," and "move" objects on the screen. Such movements may rearrange directory structures, select files, execute procedures, redesign screen layouts, or create and edit images or text. Direct manipulation and menu selection also more easily lend themselves to "object-verb" syntax. (For example, on a Macintosh the user can first select a file icon, and then choose some command or operation to apply to it, such as "open.")
- **Natural language.** "Natural language" is a term used to refer to human languages such as Spanish, English, Hopi, Japanese, etc. (The modifier "natural" is inserted to distinguish them from computer languages such as Pascal, FORTRAN, C, etc.) Typed natural language is not a very practical method for HCI; however, with the advent of real-time speech-understanding systems, spoken natural language is likely to become a very important form of system interaction for GIS. Ideally, users will be able to express their needs through a combination of speech and gestures (pointing at items on CRTs, using a mouse, track ball, or touch-screen technology).

2.3.2 Who Are the "Users"? A final set of questions that are an important part of the research agenda for GIS involves the differences among current and potential users of GISs. What level of geographic knowledge and training should be assumed for the users? A system designed for "spatially aware professionals," such as geologists, urban planners, or foresters, could borrow concepts from cartography and from those disciplines. An ideal system for such users would probably be too complicated for a "citizen-scientist" to walk up to in a public library and use to enquire about an environmental or planning issue. But the "walk-up-and-use" system would probably be tedious for the professional to use. Furthermore, some of the simplification needed for the hypothetical walk-up-and-use system would be compensating for low computer knowledge and skills. How should level of computer knowledge and related skills interact with level of geographic knowledge in order to create an ideal user interface? And how does frequency of use influence the design? A person who uses a system 30–40 h per week as a GIS technician or operator would soon learn even the most cryptic commands; on the other hand, a person with high geographic and computer skills might still need a simple system, if he or she only used it for half an hour at a time, twice a month. Finally, a programmer's interface might be quite different from the interface for "end users."

This section presents more questions than answers. That reflects the author's opinion of the state of user interface research in GISs late in 1990. In the next decade, we can expect dramatic improvements in this area. Specialists in related disciplines will play an important role in tailoring the interface design to their field's view of the geographic world.

2.4 Social, Economic, Institutional, and Legal Issues

The introduction of any new technology raises a number of social, economic, institutional, and legal issues. Geographic information systems are no exception. The ability to use home addresses as a method for combining disparate databases leads to new potentials for modeling and planning on the one hand, and invasion of privacy on the other. GISs are being used to transfer census data to postal delivery zones, improving potential for targeted commercial mailings, often considered "junk mail" by recipients. GIS-related technology in redistricting can bring "gerrymandering" (manipulation of electoral districts to maximize the chances for selected candidates) to a whole new level of sophistication, but low-cost solutions can give the same power to manipulate to all political factions and special interests.

Much effort is currently being expended in attempts to determine how people actually use geographic information in decision making, and how the availability of computerized systems (GISs) changes that use. Then, it would be very useful to know the value that the information has in promoting better decisions, or in reducing risk associated with "wrong" decisions. One difficulty in using a cost-benefit analysis approach is that benefits often are unanticipated, and may even fall outside the organization's mandate.

Another set of concerns reflects the conceptual and technical issues of institutions sharing information. If two organizations or departments are not competitors, they might both gain from a sharing of information. However, neither organization may be willing to rely on the other for information critical to its own mission, and thus they might duplicate the data collection in order to remain in control. The situation is further complicated by the fact that organizations might develop different data models for what are essentially the same phenomenon. Thus, roads may be viewed as networks of links with capacities by a garbage-collection entity or a delivery fleet, but as polygonal regions with pavement types and ages by the road maintenance department. Lastly, different proprietary GISs often use different data structures, and exchange of data between different systems may be complicated.

In the area of legal issues, decision making always involves some risk of litigation if a "wrong" decision is reached. If a GIS is used together with commercial market research data to decide where to locate a retail outlet, and if the outlet then fails because the number of potential customers was much less than predicted, who would be liable—the GIS vendor, the data vendor, or the consulting firm that used both to come up with the recommendation? On the other hand, it may be more important in a trial to be able to document the decision-making procedures than to justify the result *per se*, and GISs may be extremely helpful in documenting how the decision was reached.

3. SUMMARY AND PROSPECTS

Maps will continue to play an essential role in the analysis of spatial data in the earth sciences, biology, the social sciences, and other areas. As computer systems such as GISs take over the data-storage function that maps have traditionally performed, graphic maps will remain important as visualization tools. Paper maps will continue to be important, but real-time mapping, involving animation, three-dimensional effects, and intense human-computer interaction, will gain in importance.

Geographic information systems (GISs) have much in common with maps and mapping, and indeed have cartography as a chief conceptual technical ancestor. However, as GIS matures, analytical capabilities will continue to grow, and mere storage, retrieval, and display of map data will receive less emphasis. Remote sensing and other advanced data acquisition techniques will, combined with GISs, computer mapping, and geographical visualization, lead to powerful tools for decision makers, planners, and scientists alike.

GLOSSARY

Chart: In modern English, a specialized map designed to support navigation by ships and aircraft.

Choropleth Map: A map which shows quantitative data for statistical units by filling each unit with a uniform color or pattern.

Conformal Projection: A map projection in which all angles on the Earth's surface are preserved on the map.

Data Model: A model of relevant aspects of the real world, which has been abstracted from that world in such a way as to retain those properties which are relevant to the current problem domain, and yet which can be implemented on a computer.

Data Structure: A detailed description of how a data model is implemented on a computer.

Developable Surface: A term in map projections for a surface that can be flattened without distortion; includes cones, planes, and cylinders.

Digital Elevation Model (DEM): A digital representation of topographic elevation, usually as a regularly spaced grid of heights, but also represented in other ways.

Equal-Area Map Projection: A map projection such that regions of equal area on the Earth's surface always have equal areas on the map projection, regardless of their shapes or locations.

Geographic Information System (GIS): A computer program or system of programs used for acquiring, storing, retrieving, analyzing, managing, and displaying geographic data and associated processes.

Isoline Map: A map which shows a phenomenon that can be considered to be a single-valued function of location in geographic space, using lines connecting points of equal value for the phenomenon. A contour is a special case of the isoline.

Map Projection: A formula or geometric construction procedure for transferring locational information from the spheroid which approximates the Earth's surface to a plane.

Pixel: In image-processing and raster GISs, a single element or cell in a digital image. It is a contraction of "picture element."

Quadtree: A spatial data model that divides an image into quadrants, and recursively divides each of these into subquadrants, keeping the structure either as a recursive tree of out-degree four, or through index numbers for subquadrants. Typically, in image analysis, the stopping criterion for recursive subdivision is uniformity of image values.

Raster Data Structure: In GISs, a data structure which represents geographic information as images composed of large numbers of small cells (pixels), usually of a uniform size and shape, and records geographic attributes for each cell. A "raster GIS" is one which employs this data structure.

Reference Map: A map which attempts to portray concrete, measurable features of the Earth's surface as accurately as possible for general purposes of locational reference and measurement.

Scientific Visualization: Techniques for the visual communication of scientific information to researchers, using computer graphics, animation, etc.

Thematic Map: A map which portrays the geographic distribution of one or a small number of variables to allow people to comprehend geographic patterns and distributions.

Vector Data Structure: In GISs, a data structure which represents geographic objects as digital versions of lines on maps, using sequences of coordinate pairs along boundaries and other lines. A "vector GIS" is one which employs this data structure.

Vehicle Navigation Aid (VNA): An in-car system that keeps track of the location of a vehicle and provides navigation assistance to the driver, either through maps or other means.

List of Works Cited

Abler, R., Adams, J. S., Gould, P. (1971), *Spatial Organization: The Geographer's View of the World*, Englewood Cliffs, NJ: Prentice-Hall.

Campbell, J. (1984), *Introductory Cartography*, Englewood Cliffs, NJ: Prentice-Hall.

Goodchild, M. F., Gopal, S., (Eds.) (1989), *The Accuracy of Spatial Databases*, London: Taylor & Francis.

Griffith, D. A. (1987), *Spatial Autocorrelation: A Primer*, Washington, DC: Association of American Geographers.

Harley, J. B. (1989), *Cartographica* **26**, 1–20.

Lakoff, George (1987), *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*, Chicago: University of Chicago Press.

McCormick, B. H., DeFanti, T. A., Brown, M. D. (1987), "Visualization in Scientific Computing (ViSC)," in *Association for Computing Machinery, SIGGRAPH Newsletter* **26**, 6.

Moellering, H. (1984), "Real Maps, Virtual Maps, and Interactive Cartography," in: G. Gaile and C. J. Willmot (Eds.) *Spatial Statistics and Spatial Models*, Boston, MA: Reidel.

Muehrcke, P. (1986), *Map Use: Reading, Analysis, and Interpretation* (2nd ed.), Madison, WI: JP Publications.

Robinson, A. H., Sale, R. D., Morrison, J. L., and Muehrcke, P. C. (1984), *Elements of Cartography* (5th ed.), New York: Wiley.

Rosch, E. (1973), "On the Internal Structure of Perceptual and Semantic Categories," in: T. E. Moore (Ed.), *Cognitive Development and the Acquisition of Language*, New York: Academic, pp. 111–144.

Star, J., Estes J. (1990), *Geographic Information Systems: An Introduction*, Englewood Cliffs, NJ: Prentice-Hall.

Synder, J. P. (1982), *Map Projections Used by the U. S. Geological Survey*, Geological Survey Bulletin No. 1532 (2nd ed.), Washington, DC: U.S. GPO.

Thrower, N. J. W. (1972), *Maps and Man: An Examination of Cartography in Relation to Culture and Civilization*, Englewood Cliffs, NJ: Prentice-Hall.

Vasiliev, I., Freundschuh, S. M., Mark, D. M., Theisen, G. D., McAvoy, J. (1990), "What is a Map," *Cartogr. J.* **27**, 119–123.

Further Reading

Journals

Cartographica (published by the University of Toronto Press; associated with the Canadian Cartographic Association).

Cartography and Geographic Information Systems (published by the American Congress on Surveying and Mapping; called *The American Cartographer* before 1990).

International Journal of Geographical Information Systems (published by Taylor & Francis in the United Kingdom).

Books

Burrough, P. A. (1986), *Principles of Geographical Information Systems for Earth Resources Assessment*, Oxford: Clarendon.

Campbell, J. (1984), *Introductory Cartography*, Englewood Cliffs, NJ: Prentice-Hall.

Muehrcke, P. (1986), *Map Use: Reading, Analysis, and Interpretation* (2nd ed.), Madison, WI: JP Publications.

Robinson, A. H., Sale, R. D., Morrison, J. L., and Muehrcke, P. C. (1984). *Elements of Cartography* (5th ed.), New York: Wiley.

Snyder, J. P. (1982), *Map Projections used by the U.S. Geological Survey*, Geological Survey Bulletin No. 1532 (2nd ed.), Washington, DC: U.S. GPO.

Star, J., Estes, J. (1990), *Geographic Information Systems: An Introduction*, Englewood Cliffs, NJ: Prentice-Hall.

Thrower, N. J. W. (1972), *Maps and Man: An Examination of Cartography in Relation to Culture and Civilization*, Englewood Cliffs, NJ: Prentice-Hall.

CATALYSIS

GARY L. HALLER, *Department of Chemical Engineering, Yale University, New Haven, Connecticut, U.S.A.*

DANIEL E. RESASCO, *Sun Oil Company, Marcus Hook, Pennsylvania, U.S.A.*

Introduction 67
1. **Kinds of Catalysis** 68
1.1 Homogeneous Catalysis 68
1.2 Enzyme Catalysis 68
1.3 Heterogeneous Catalysis 68
2. **History** 69
3. **Adsorption of Gases on Solid Surfaces** 70
3.1 Types of Adsorption 71
3.2 Adsorption Isotherms 71
3.3 Dissociative Adsorption 72
3.4 Thermal Desorption 72
4. **Types of Catalytic Materials** 73
4.1 Supported and High-Area Catalysts 73
4.1.1 Supported Metal Catalysts 73
4.1.2 Supported Nonmetal Catalysts 73
4.1.3 Amorphous Oxide Catalysts 73
4.1.4 Zeolites 74
4.2 Unsupported Catalysts 74
4.3 Single-Crystal Catalysts 74
5. **Preparation and Characterization of Heterogeneous Catalysts** 75
5.1 Preparation 75
5.1.1 Incorporation of the Active Components on the Support 75
5.1.1.1 Impregnation 75
5.1.1.2 Adsorption from Solution 75
5.1.1.3 Precipitation 75
5.1.2 Drying and Calcination 75
5.1.3 Thermal Activation 75
5.2 Characterization 76
6. **Kinetics of Heterogeneous Catalytic Reactions** 77
6.1 Elementary Steps 77
6.1.1 Unimolecular Reactions on Surfaces 78
7. **Mechanisms of Heterogeneous Catalysis** 78
8. **Chemical Enhancement/Poisoning of Catalytic Activity** 79
8.1 Catalyst Promoters 79
8.2 Catalyst Poisons 80
9. **Heterogeneous Catalytic Reactors** 80
10. **Industrial Catalytic Processes** 81
10.1 Pollution Control 81
10.2 Selective Oxidation 81
10.3 Cracking 81
10.4 Reforming 82
11. **Summary** 82
Glossary 83
Works Cited 83
Further Reading 84

INTRODUCTION

A *catalyst* is a substance that increases the rate of a chemical reaction but is not consumed or permanently altered by the reaction. It will be obvious that increasing the rate of desired chemical reactions will have economic benefits for society. While it is not easy to quantify these precisely, it has been estimated that between 25% and 30% of all manufactured goods in the U.S.A. involve a catalyzed step, which suggests that the economic impact on our daily lives is very large indeed. Industrially, catalyzed reactions are very much dominated by heterogeneous catalysts: catalysts that are in a phase different from that of the reactants, most commonly a solid catalyst acting on fluid (gas or liquid) reactants. Heterogeneous catalysts have several advantages over other kinds of catalysts which are usually classified as either enzyme or homogeneous. Let us briefly discuss each kind before we return to our main topic which is heterogeneous catalysis.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. KINDS OF CATALYSIS

1.1 Homogeneous Catalysis

As can be divined from the name, we expect that a homogeneous catalyst will be uniformly distributed in the reactant phase. There are many examples of gas-phase catalyzed reactions which are nearly uniform on a macroscale (meters to hundreds of meters) in stratospheric chemistry. Consider one that is relevant to the discussion of destruction of ozone in the stratosphere. Ozone is formed by UV radiation, which photolyzes molecular O_2 into atoms that then react with molecular O_2 to form the triatomic molecule O_3. In the absence of catalysis, the main reaction that converts O_3 back into molecular O_2 is the direct reaction with oxygen atoms:

$$O+O_3 \rightarrow 2O_2. \tag{1}$$

The formation and destruction of O_3 would establish some steady-state concentration which would only change if either the formation or destruction rates were changed. Such a steady state is established, but it results in a much lower concentration than would result from the two reactions so far mentioned, O_2 photolysis and reaction of oxygen atoms with O_3. This is the result of homogeneous catalysis of the latter reaction by oxides of nitrogen. Nitrous oxide, N_2O, is formed at ground level by natural biological activity. Because this molecule is rather inert in the troposphere, it diffuses into the stratosphere where it can be converted by UV radiation into nitric oxide, NO, a more active form which acts as a homogeneous catalyst for reaction (1) above:

$$NO+O_3 \rightarrow NO_2+O_2, \tag{2}$$

$$NO_2+O \rightarrow NO+O_2. \tag{3}$$

What should be noted here is that at steady state the rate of NO consumed in reaction (2) will be equal to the rate of NO production in reaction (3), and a similar statement can be made about NO_2. Thus NO (or NO_2) can be viewed as the catalyst here. However, there is a net consumption of both types of odd oxygen, i.e., atoms and O_3. Also, it should be appreciated that the rate of O_3 destruction by the set of reactions (2) and (3) is much faster than the direct reaction (1). This is a common fact of catalysis, i.e., more reaction steps which represent a more complex set than the uncatalyzed reaction can still be faster than the uncatalyzed reaction.

As an example of a commercial homogeneously catalyzed reaction, we can consider the Monsanto process for methanol carbonylation to acetic acid. Acetic acid is produced continuously by reacting methanol and CO in a liquid-phase homogeneous catalytic reactor using $[Rh(CO)_2I_2]^-$ as a catalyst. In the Monsanto process acetic acid is separated from the catalyst by distillation. The Monsanto process operates at low temperatures and pressures relative to earlier commercial processes.

1.2 Enzyme Catalysis

Enzymes are biological catalysts and might also be considered homogeneous since they are often uniformly distributed at least on the cellular level. However, their size (on the order of microns) puts them somewhere between homogeneous and heterogeneous catalysts. Also, the fact that they are biologically active proteins (amino acid polymers) which are sensitive to *p*H, temperature, solvent, etc., as is true of almost all biological material, distinguishes them from most heterogeneous catalysts, but less so from homogeneous catalysts, which are often metal organic complexes which might have similar thermal stability to that of enzymes.

An important commercial application of enzymes is as additives to synthetic detergents. This additive is protease or another hydrolase (an enzyme that catalyzes hydrolysis of various kinds) which breaks down a variety of fabric stains and aids in the solubilization of soil.

1.3 Heterogeneous Catalysis

Heterogeneous catalysts will be our main topic of discussion, but are introduced here primarily to allow a quick comparison of properties with homogeneous and enzyme catalysts. These catalysts are almost always solids and usually metals (e.g., Pt) or inorganic oxides (e.g., alumina). They are often combined, e.g., very small particles (of order of 10 Å) of Pt on high-area (of order of 100 m^2/g) alumina. The largest dollar use of heterogeneous catalysts at the present time is for the application to pollutant control of automobile exhaust. These catalysts are a combination of Pt, Pd, and Rh supported on a ceramic matrix (and often have other important components, e.g., ceria, BaO, etc.). These catalysts can effectively convert a wide variety of unburned hydrocarbons and CO to CO_2 and H_2O while also converting NO to molecular N_2. This application can be used to illustrate the advantages of heterogeneous catalysts over their homogeneous counterparts. First among these is the cost of separation of the reactants from the catalyst, which can be minimal for heterogeneous catalysts since the gas passes over the solid catalyst, reacts, and is displaced by incoming reactant (exhaust gas). Second is the fact that reactors can be designed to process huge flows. This is usually quantified in terms of a space velocity, defined as the volume of gas in units of the reactor volume that can be processed in a given time. Typically an auto exhaust catalyst has a hourly space velocity of 100 000, meaning that a volume of gas 100 000 times the size of the reactor is processed each hour, something that would not be possible for most homogeneous catalysts, which normally are dispersed in a liquid. Third, there is the harsh environment in which the catalyst must be stable. Typical exhaust catalysts experience temperature excursions that can reach 1000 °C and, by law, have to last 50 000 miles (soon to be increased to 100 000 miles). Most enzymes would have long been reduced to a char by the time the exhaust catalyst reaches its steady-state operating temperature of a few hundred degrees Celsius.

Finally, we should dispel the mistaken conception that enzymes catalyze reactions at rates that man-made heterogeneous catalysts cannot attain. The origin of this presumption is that we ask heterogeneous catalysts to operate at the same temperature at which enzyme action will be optimized, which is around ambient or somewhat above. But this is not a fair comparison because we should allow both catalysts to operate where they are near their optimum. Consider CO oxidation. For a stoichiometric mixture of CO and O_2, this will be about 600 K for Pd, when more than a quarter of all CO molecules that strike the Pd surface react. If we use a foil or gauze, the rate at which the CO molecules reach the Pd surface is limited primarily by molecular velocities over which we have not too much control. There are enzymes that catalyze this reaction, but they will be at a triple disadvantage if operated at their optimum conditions for a high rate. This will be a lower temperature, but the activation energies are comparable; thus the rate will be lower for reasons of thermal activation. Next, the enzyme will have to operate in some solvent through which the diffusion to the surface will be slower than in the gas phase. Finally, because of the large size of the enzyme, the number of active sites per unit area (defined later) will be small relative to that of the Pd surface where it appears that most atoms are active. However, the ground on which heterogeneous catalysts cannot compete with enzymes is specificity or selectivity, e.g., to hydrogenate a given double bond among many in a complex molecule.

2. HISTORY

Catalysts have been called the philosophers' stone of the alchemist, and perhaps for this reason the mysterious nature of catalysis was retained long after physical chemists began to study it (Burwell, 1983). In 1835 Berzelius surveyed several different kinds of chemical reactions—homogeneously catalyzed hydrolysis, heterogeneously catalyzed oxidation, and both homogeneously and heterogeneously catalyzed decomposition of H_2O_2—and designated the common agent in these reactions a catalyst. Unfortunately, he attributed the actions of these agents to a "catalytic force," an idea that did not disappear until the beginning of the twentieth century.

The oldest technological use of catalysis is surely related to the enzymatic catalysis involved in the fermentation for wine making and beer brewing. Even homogeneous catalysis, as practiced in the acid-catalyzed hydrolyses needed for the manufacture of soaps, predates industrial applications of heterogeneous catalysis since, apparently, the first technological application of a heterogeneous catalyst was as a spontaneous igniter of H_2 in the Döbereiner's Tinder Box, an early version of the modern flint lighter that was commercialized in the early part of the nineteenth century. Other industrial applications of heterogeneous catalysis for chemicals production were introduced in the second half of the nineteenth century with the oxidation of HCl to Cl_2 on $CuCl_2$ (the Deacon process) and the Pt-catalyzed oxidation of SO_2 to SO_3 in the manufacture of sulfuric acid.

The development of heterogeneous catalysis both scientifically and for industrial applications remained largely a European activity at the beginning of the twentieth century. Ostwald and Sabatier received Nobel Prizes for work on catalysis in 1909 and 1912, respectively, and Sabatier's work on catalytic hydrogenation had wide industrial applications. Haber's and Bosch's Nobel Prizes derived from the discovery and development, respectively, of the technical basis for the synthesis of ammonia from its elements, a process commercialized in 1913. It would be possible to outline almost all of the modern catalytic concepts by reviewing work on this catalytic reaction alone. One would not wish to discontinue the name dropping of Nobel Prizes for catalysis without mention of Langmuir, for several reasons. It is worth noting that Langmuir, while making very significant contributions to industrial catalysis, must be given primary credit for advancing new concepts and techniques in catalysis, e.g., the Langmuir isotherm and Langmuir kinetics, and that he did this work in an industrial laboratory while trying to make a better light bulb for General Electric.

Important modern applications of heterogeneous catalysis have occurred in the petroleum and petrochemical industry, and these were largely to be developed in the U.S. Since this brief historical sketch is necessarily very selective, we will focus on one class of catalysts, solid acids. Such catalysts are used for polymerization, alkylation, cracking, isomerization, and other important refinery reactions. Solid acid catalysts were the first kind of catalysts to be introduced into petroleum refining. Polymerization was the first commercial catalytic process introduced to the petroleum industry and initially utilized solid phosphoric acid formed by heating H_3PO_4 on kieselguhr (siliceous earth, SiO_2), a catalyst developed by V. N. Ipatieff at Universal Oil Products. However, catalytic cracking (the conversion of heavier petroleum hydrocarbons into lighter fragments, particularly in the gasoline boiling range) is the most important acid-catalyzed reaction and remains the largest volume use of heterogeneous catalysts yet today. Early but limited use of an $AlCl_3$ catalyst occurred in 1915. However, the modern process was developed and commercialized by the Houdry Process Corp. (with Socony-Vacuum, now Mobil, and Sun Oil Co.) in the decade of the 1930s with 1936 usually taken as the year of commercialization. The initial commercial catalysts were clays rich in silica and alumina and low in Na, Fe, Ca, and Mg impurities. By 1940 these were replaced by synthetic silica-alumina cracking catalysts that eliminated the variability in clay properties and produced less coke. A major discontinuity in technology resulted from the commercialization of zeolite cracking catalysts by Socony-Mobil Oil Corp. based on the work of Plank and Rosinsky. A zeolite is a crystalline silica-alumina with micropores and an ion-exchange capacity associated with each Al^{3+} in the tetrahedral framework. Plank and Rosinsky stabilized the zeolite (a faujasite structure called *Y*) by ion exchange of rare-earth metals and by distributing the zeolite crystallites in a matrix of silica-alumina. Such catalysts were commercialized in 1961.

Finally, let us return to the fact that catalysis is by definition a kinetic phenomenon in order to mention some names passed over up to this point. Bodenstein, a student of Ostwald, must be mentioned here, although his interpretation of the kinetics, i.e., the mechanism of catalysis involving multilayers that inhibited reaction, was discarded by Langmuir in favor of partial monolayers of chemisorbed molecules. We also must mention E. K. Rideal and H. S. Taylor, who published in 1919 the first comprehensive book on catalysis, *Catalysis in Theory and Practice* (1919). Finally, we cannot close without recognizing P. H. Emmett's work on the kinetics of ammonia synthesis and the BET (Brunauer, Emmett and Teller; see below) method for determining surface area.

3. ADSORPTION OF GASES ON SOLID SURFACES

For heterogeneous catalysis the term *adsorption* describes the accumulation of any particular substance on the surface of a solid, without penetration into the bulk, in a concentration which is higher than that in the bulk. In the particular case of adsorption on solid surfaces from the gas phase, the term refers to the process of a gas molecule (adsorbate) sticking on the solid surface (adsorbent) as a consequence of a collision. The term *desorption* describes the reverse process, i.e., the transfer of adsorbed molecules from the surface to the gas phase.

From a thermodynamic point of view, adsorption is almost always an exothermic process. This can be rationalized considering that the thermodynamic criterion for a spontaneous process requires that the free energy (ΔG) be negative. For most adsorption processes the entropy change (ΔS) is negative as the molecule proceeds from a disordered state in the gas phase to a more ordered state on the surface. Because $\Delta G = \Delta H - T\Delta S$, it follows that ΔH must be negative and its absolute value must be greater than $T\Delta S$.

3.1 Types of Adsorption

The magnitude of ΔH reflects the strength of adsorption and can be used to distinguish two different types of adsorption phenomena. These types are *physisorption*, or physical adsorption, and *chemisorption*, or chemical adsorption.

Physisorption can be considered as a condensation process. The ΔH values for this type of adsorption are about 2–5 kcal/mol, which is in the range of the heats of liquefaction. Physisorption is nearly independent of the nature of the solid surface, and because it is nonspecific, multilayers can form. The forces involved in the physisorption process are of the van der Waals type. The physisorbed molecule is not associated with any particular site on the surface. Thus, the physisorption can be used as a measure of the overall surface area of any solid. On the other hand, the values of ΔH for chemisorption processes are much higher and usually exhibit a much broader range than physisorption. Typical ΔH values for chemisorption range from 15 to 50 kcal/mol, which are of the same order of chemical bond energies. This indicates that chemisorption indeed involves the formation of surface chemical bonds. Consequently, adsorbates on the surface are generally considered to occupy localized *adsorption sites*. The nature of these sites can vary greatly and, in many cases, their exact identity and structure are unknown. An unoccupied site is symbolized by an asterisk (*), while an occupied site is symbolized by A^*, where A is the adsorbed molecule or molecular fragment.

3.2 Adsorption Isotherms

An adsorption isotherm is the constitutive relation which gives the amount of adsorbate on the surface in equilibrium with the pressure of the particular gas being adsorbed, at constant temperature.

The adsorption equilibrium is represented by

$$A + {*} \Leftrightarrow A^* \tag{4}$$

with an equilibrium constant $K_A = [A^*]/[\mathrm{A}][*]$, where the brackets imply concentration. If it is assumed that the rate of adsorption is proportional to the number of gas molecules colliding on vacant sites, then it can be written

$$r_{\mathrm{ads}} = k_{\mathrm{ads}} P_A (1 - \theta_A), \tag{5}$$

where r_{ads} is the adsorption rate, k_{ads} the adsorption rate constant, P_A the partial pressure of the adsorbing gas, and θ_A the fraction of occupied sites, or fractional coverage. Similarly, it can be assumed that the rate of desorption should be proportional to the fractional coverage. Thus,

$$r_{\mathrm{des}} = k_{\mathrm{des}} \theta_A. \tag{6}$$

At equilibrium, the two rates must be equal. Thus,

$$k_{\mathrm{ads}} P_A (1 - \theta_A) = k_{\mathrm{des}} \theta_A. \tag{7}$$

Consequently,

$$K_A = \theta_A / P_A (1 - \theta_A) \tag{8}$$

or

$$\theta_A = K_A P_A / (1 + K_A P_A). \tag{9}$$

This equation is the well-known Langmuir isotherm. Typical adsorption isotherms which have the Langmuir form are illustrated in Fig. 1.

When two different molecules, A and B, adsorb simultaneously on the surface and compete for the same adsorption sites, Eq. (9) changes. The corresponding isotherm becomes

$$\theta_A = K_A P_A / (1 + K_A P_A + K_B P_B). \tag{10}$$

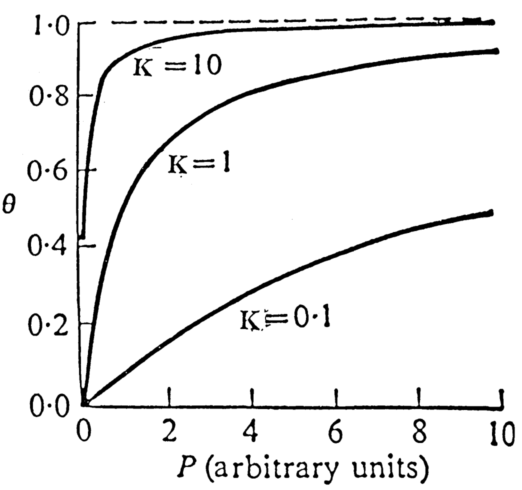

FIG. 1. Langmuir-type adsorption isotherms for different values of the equilibrium adsorption constant, *K* (Bond, 1987).

3.3 Dissociative Adsorption

Dissociative adsorption refers to the process by which the gas molecule dissociates on the surface during the adsorption. The chemisorption of hydrogen on metals is an example of dissociative adsorption. This process can be represented by

$$H_2 + 2^* \Leftrightarrow 2H^*. \qquad (11)$$

Figure 2 illustrates potential-energy curves as a function of distance from a metal surface for H_2 molecules and H atoms. A H_2 molecule, far from the surface, is taken as a reference for the zero potential energy. Both curves exhibit minima at particular distances from the surface, r_P and r_C. The depths of these minima reflect the strength of the adsorption processes. On the basis of the great differences typically observed for these depths, the adsorption of H_2 can be called physisorption while the adsorption of H can be called chemisorption. Accordingly, the values ΔH_P and ΔH_C, indicated in the figure, represent the heats of physisorption and chemisorption, respectively.

3.4 Thermal Desorption

The rate of desorption can be written in terms of an Arrhenius expression:

$$-d\theta/dt = \nu\theta^n \exp(-E_d/RT), \qquad (12)$$

where ν is the frequency factor, θ the surface coverage, n the kinetic order, and E_d the activation energy for desorption. By increasing the temperature of the sample with a linear heating program and monitoring the variation of θ with time it is possible to evaluate E_d and n.

It can be shown that when the sample chamber is evacuated at a pumping rate much higher than the desorption rate, the variation in surface coverage with time is proportional to the pressure inside the chamber. As the temperature varies linearly with time, plots of desorption rates (pressure) as a function of temperature (time) can be readily obtained. From the analysis of the variation in positions and shapes of the peaks in those plots, the kinetic parameters can be determined. Figure 3 illustrates temperature-programmed desorption (TPD) profiles for the desorption of hydrogen from tungsten. Four different peaks are clearly distinguished corresponding to different binding states, each one having different kinetic parameters. For example, β_1 corresponds to a first-order desorption and the position of the peak does not change with varying coverage. On the other hand, β_2 corresponds to a second-order desorption in which the peak shifts to lower temperatures as the coverage increases. Decomposition of the TPD curves into peaks requires appropriate assump-

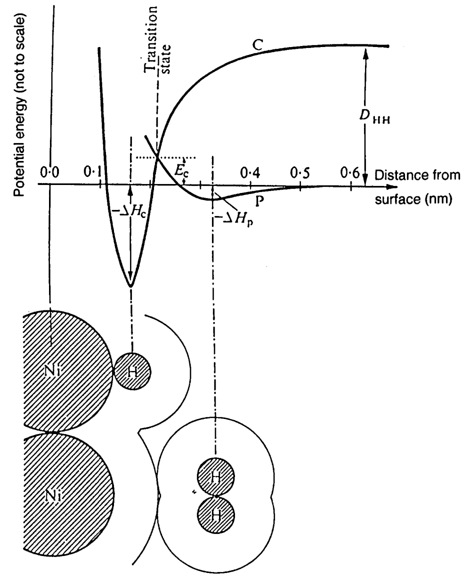

FIG. 2. Potential-energy curves for the adsorption of hydrogen on a metal surface (Bond, 1987).

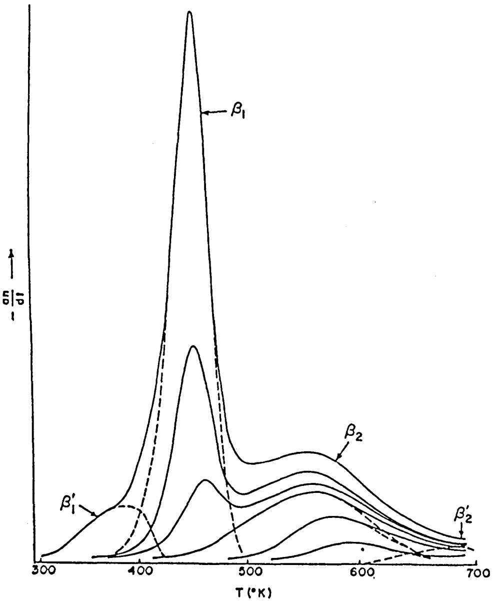

FIG. 3. A TPD sequence for successively larger amounts (curves from bottom to top) of H_2 adsorbed at 300 K on W (100) (Tamm and Schmidt, 1969).

tions. However, if both the coverage and rate of temperature change are varied over a wide range, the whole curve can be modeled and the frequency factor, kinetic order, and activation energy can be extracted without *a priori* assumptions.

4. TYPES OF CATALYTIC MATERIALS

Heterogeneous catalysts may adopt various forms. In many cases, such as in most petrochemical industrial processes, the catalytically active material is stabilized on a high–surface-area *support*. In other cases, unsupported and massive catalysts are used. For fundamental research purposes, catalytic chemists and surface physicists often use well-characterized, single crystals as models of catalysts.

4.1 Supported and High-Area Catalysts

When the catalytic reaction is not limited by mass transport to or from the surface, the rate of reaction is proportional to the surface area of the active phase. This can be maximized by making the particles of the active material as small as possible. To stabilize this small size it is very advantageous to support them on a high–surface-area material.

There is a wide range of materials which can be used as catalyst supports. The most commonly used are

- high-area amorphous silica (SiO_2), such as silica gels;
- alumina (Al_2O_3) in any of several metastable crystalline forms (γ, η, θ);
- titania (TiO_2) in its two most common crystalline forms, rutile and anatase;
- mixed oxides such as SiO_2-Al_2O_3;
- zeolites and natural clays;
- carbon;
- low-area monolithic supports made from metals or ceramic materials (cordierite, mullite, etc.)

In some cases, the high–surface-area materials (zeolites, aluminas, etc.) may themselves be catalytically active. These are called high-area catalysts. In other cases, both the supported active material and the support participate in the catalytic reaction. This type of catalyst is called "bifunctional."

4.1.1 Supported Metal Catalysts. A large number of catalytic reactions use a metal as the active material. In general, *d* orbitals are involved in the adsorbate–metal bonding. Consequently, most catalytically active metals belong to the *d* transition series. Transition metals are used in oxidation, hydrogenation/dehydrogenation, reforming, hydrocracking, and synthesis reactions.

4.1.2 Supported Nonmetal Catalysts. Supported nonmetal catalysts have found a large number of industrial applications. Among them, transition-metal oxides, sulfides, and chlorides have been the most widely used. Transition-metal oxides are used in a number of chemical processes, such as selective and total oxidation, oxidative dehydrogenation, ammoxidation, and water-gas shift. Sulfides are used in hydrodesulfurization of petroleum. Among other reasons, the sulfur commonly present in crude oils has to be removed before entering the reforming processes because the catalysts used in that process are metallic and are poisoned by sulfur.

4.1.3 Amorphous Oxide Catalysts. Single oxides generally exhibit a degree of acidity which is too low for catalytic purposes. But binary oxide compounds may drastically increase the acidity and make the material catalytically active. This is the case of the silica-alumina system. The origin of the acidity is the bridging hydroxyl between Si and Al which is formed for charge balance. One can picture the formation of this site as the removal of tetrahedral Si, which is formally +4, and replacement by Al in the same tetrahedral position, but the Al is formally +3. There is a wide distribution of acidity of these sites in amorphous silica-alumina, and it has been shown that the acidity is affected by both the bond angles of the Si-OH-Al linkage and the distribution of ligands in the coordination spheres of Si and Al.

4.1.4 Zeolites. Zeolites are highly crystalline aluminosilicate materials of high surface area and well-defined structure. Some of these materials can be found as natural minerals. Examples of these natural zeolites are the faujasites and mordenite. In the last few decades, a great variety of synthetic zeolites have been prepared and used as adsorbents and catalysts. The basic unit of a faujasite is the sodalite cage tetrahedral units. This cubo-octahedral unit, shown in Fig. 4(a), is composed of SiO_4^{4-} and AlO_4^{5-}. The -4 charge on SiO_4^{4-} is balanced by bonding to four other tetrahedral Si or Al ions. The extra negative charge in the AlO_4^{5-} unit is balanced by positive ions, such as Na^+,K^+,Ca^{2+}, etc. From the sodalite unit different zeolites can be obtained. When the sodalite units are connected by their square faces they form a highly porous and regular array. This is called *A*-zeolite [Fig. 4(b)]. The size of the opening depends on the cations balancing the negative charge, i.e., 0.3 nm for K^+, 0.4 nm for Na^+, 0.5 nm for Ca^{2+}. On the other hand, when the sodalite units are connected by their hexagonal faces, a structure containing large cages joined by smaller openings results. Depending on the silica/alumina ratio in the material, the resulting zeolites are called *X* (silica/alumina 1.5) or *Y* (silica/alumina 1.5 to 3).

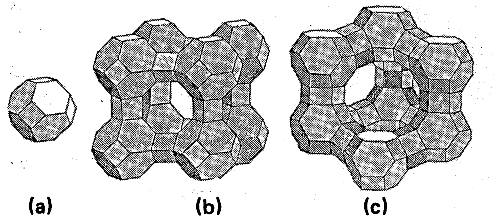
(a) (b) (c)

FIG. 4. Zeolite structure (Bond, 1987): (a) Cubo-octahedral unit of faujasite zeolites; (b) *A*-zeolite structure; (c) *X*- and *Y*-zeolite structures.

Zeolites exhibit a great variety of possibilities as catalytic materials. They can be designed with precise control of acidity and structure. The density and strength of acid sites can be varied by exchanging the zeolites with different cations. The dimensions of the pore openings are of the same order as molecular sizes, and so shape and size selectivity can also be readily affected by proper cation exchange.

4.2 Unsupported Catalysts

Although in most cases, it is more convenient to have the active species of a catalyst supported on a high–surface-area material, sometimes unsupported catalysts can be used. For example, the Raney nickel catalysts are unsupported Ni prepared from Ni-Al alloys by leaching the Al out with NaOH to make a porous metal. This catalyst is used to hydrogenate vegetable oil and in the production of fine chemicals.

Also, unsupported colloidal platinum of relatively uniform size (about 1.5 nm) has found interesting applications, particularly as a model catalyst. However, the most widely used unsupported metal catalyst is not in the form of particles, but metallic gauzes. This type of catalyst is used, for example, in nitric acid plants for the oxidation of ammonia. In this and most similar applications, the reaction temperature is very high, and essentially every molecule that strikes the catalyst surface reacts. Thus, high area, which is always accomplished by porosity, will not improve the conversion because the transport into the pores would limit the rate for such fast reactions.

4.3 Single-Crystal Catalysts

In many fundamental studies, single-crystal model catalysts have been used. The advantage of these catalysts is that the composition and structure of the surface can be readily determined and controlled. They can be analyzed under ultrahigh-vacuum conditions, minimizing the presence of impurities or undesired adsorbates. In addition, they can be studied by physical characterization techniques such as x-ray photoelectron spectroscopy (XPS), Auger-electron spectroscopy (AES), low-energy electron diffraction (LEED), or electron energy-loss spectroscopy (EELS), which can determine surface composition, symmetry, and molecular vibrations. Single-crystal surfaces are identified in terms of the Miller indices of the exposed planes. Figure 5(a) shows the arrangement of atoms on the surfaces of a face-centered-cubic metal for the particular indices (100), (110), and (111). In other cases the surface exhibits defects such as those illustrated in Fig. 5(b): steps, kinks, ledges, vacancies, etc.

5. PREPARATION AND CHARACTERIZATION OF HETEROGENEOUS CATALYSTS

5.1 Preparation

In the preparation of supported catalysts the following sequence of steps is usually followed: (a) incorporation of the active components on the support, (b) drying, and (c) thermal activation.

5.1.1 Incorporation of the Active Components on the Support. There exist several practical methods for incorporating the precursors. The most widely used are the following.

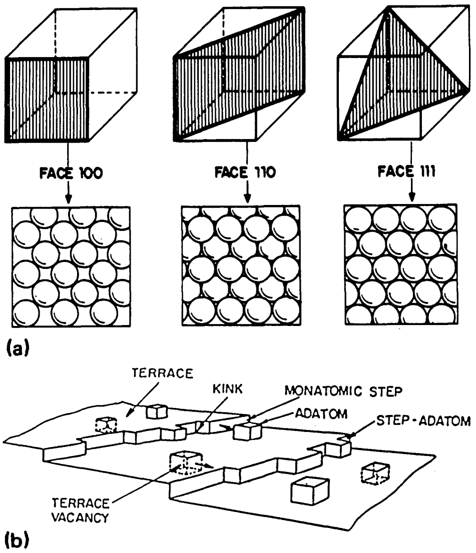

FIG. 5. (a) Ideal fcc metal surfaces for (100), (110), and (111) planes (Le Page, 1987). (b) Typical defects commonly present on metal surfaces (Boudart and Djega-Mariadassou, 1984).

5.1.1.1 Impregnation. The support is impregnated with a solution (usually aqueous) containing the ions of the active species. The solvent is evaporated, and the precursors deposited. After the proper thermal treatments, small particles are formed. The size of these particles strongly depends on the concentration of the impregnating solution. When the volume of the solution used does not exceed the pore volume, the method is called "incipient wetness impregnation."

5.1.1.2 Adsorption from Solution. The oxide supports usually have a relatively high surface density of hydroxyl groups. Depending on the nature of the oxide and the solution in which it is immersed, these hydroxyl groups can be exchanged with charged species from the solution. In this way, cation adsorption takes place when surface H^+ are exchanged with cationic species, e.g.,

$$\mathrm{MOH}+C^+\rightleftarrows\mathrm{MO}C+\mathrm{H}^+. \tag{13}$$

The analogous anion exchange involves the adsorption of H^+ and the anionic species. This process can be described as

$$\mathrm{MOH}+H^+ +A^-\rightleftarrows\mathrm{MOH}_2A. \tag{14}$$

In either case, the precursors are incorporated in a highly dispersed state. After the thermal treatment, the resultant particles deposited on the support are very small.

5.1.1.3 Precipitation. The solution containing the active species, usually in the form of nitrate, chloride, or sulfate, is contacted with an alkali solution to cause precipitation of the active species in the form of a metal oxide or carbonate on the support.

5.1.2 Drying and Calcination. During the drying step the precursors crystallize on the surface of the support. The rate of drying must be carefully controlled. When drying is too fast nonuniform distribution of the active phase may occur. During the calcination the crystallized salt is decomposed and dispersed oxide particles are formed which tend to wet or bind to the support surface.

5.1.3 Thermal Activation. The thermal activation is the final step in producing the active species on the support surface. Depending on the nature of the catalyst being prepared, the thermal activation can be an oxidation in O_2 (oxide catalysts), a reduction in H_2 (metal catalysts), a treatment in mixtures of H_2S/H_2 (sulfided catalysts), HCl/O_2 (oxychloride catalysts), etc.

5.2 Characterization

The importance of catalyst characterization can be analyzed from several points of view. In the first place, to understand the microscopic processes and reaction mechanisms taking place on the surface of heterogeneous catalysts, a detailed knowledge of their structure, composition, and properties is absolutely necessary. Also, from the practical point of view, in the development of effective and reproducible catalyst preparation methods, structure and composition of the catalysts during the different stages of the process must be known. Some of the most commonly used characterization techniques are listed in Table 1.

Bulk composition can be determined by dissolution and analysis of the solution by atomic absorption (AA) spectroscopy. But the determination of surface composition is less straightforward since it strongly depends on the surface sensitivity of the particular analytical technique being used. Two of the most widely used techniques for surface analysis are x-ray photoelectron spectroscopy (XPS) and Auger-electron spectroscopy (AES). In both techniques electrons are ejected from the sample after being excited by a photon or an electron source, respectively. In the case of XPS the kinetic energy of primary photoelectrons is measured, while for AES, that of secondary Auger electrons is measured. From the intensities of the observed peaks and with differences in excitation cross-sectional values for each element taken into account, a surface composition can be derived by both techniques. Because of the variation of the elastic mean free path of electrons in solids as a function of kinetic energy, the surface sensitivity of XPS is significantly lower than that of AES.

In the case of supported metal catalysts, the fraction of exposed metal atoms is a very important parameter that can be determined by selective chemisorption. Hydrogen, CO, and oxygen selectively chemisorb on most

Table 1. Most common techniques employed in catalyst characterization.

Technique	Information Provided
Gas adsorption	Surface area Exposed metal area
Mercury porosimeter	Pore size distribution
Temperature–programmed reduction (TPR)	Reducibility of metal particles Metal-support interactions Metal-metal interactions
Differential thermal analysis (DTA)	Heats of adsorption
Infrared spectroscopy (IR)	Nature of adsorbed species and reaction intermediates
X-ray diffraction (XRD)	Crystallinity Phases present Particle size
Electron microscopy (TEM, SEM, STEM)	Particle sizes and shapes Texture
Electron spectroscopies (XPS, AES)	Surface composition Electronic interactions Oxidation states
X-ray absorption (EXAFS, XANES)	Coordination number Oxidation states Electronic interactions
Atomic absorption (AA)	Bulk composition

metals at room temperature. As chemisorption occurs into a monolayer, the exposed metal fractions can be immediately determined after the measurement of saturation uptakes, provided that the metal content and the adsorption stoichiometry are known. In most metals, hydrogen adsorbs with a stoichiometry H/M of about 1/1. On the other hand, CO frequently exhibits a range of stoichiometries. For example, Fig. 6 illustrates an IR spectra of CO adsorbed on Rh. The band at 2070 cm^{-1} corresponds to linearly adsorbed CO (stoichiometry 1/1), the doublets at 2101 and 2031 cm^{-1} are due to gem dicarbonyl species (stoichiometry 2/1 with Rh oxidized to Rh^+), while the broad band at about 1870 cm^{-1} corresponds to multiply bonded bridge form (stoichiometry 1/2, 1/3, or even 1/4).

On the other hand, physisorption of an inert gas, such as N_2, Ar, or Kr, at low temperatures can be used to determine the total surface area of the catalysts. Physisorption is a nonselective process and therefore takes place over the entire solid surface. The BET method (Brunauer-Emmett-Teller model for the relation between amount adsorbed and pressure) provides reproducible values of overall surface areas and is universally used.

Direct observation of metal particles can be achieved by transmission electron microscopy (TEM). A typical TEM micrograph is shown in Fig. 7, for a Pt-Sn/Al_2O_3 reforming catalyst. The metal particles appear as dark spots over a light background. This technique allows one to obtain not only an average particle size, but also particle size distributions. Particles of the order of 0.5–1 nm in size can be observed.

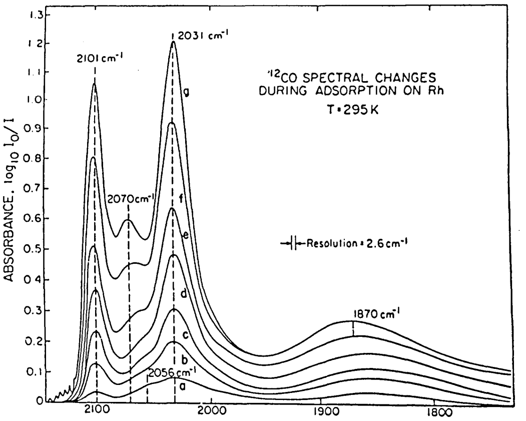

FIG. 6. IR spectra for CO adsorbed on Rh at increasing surface coverages (Yates *et al.*, 1979).

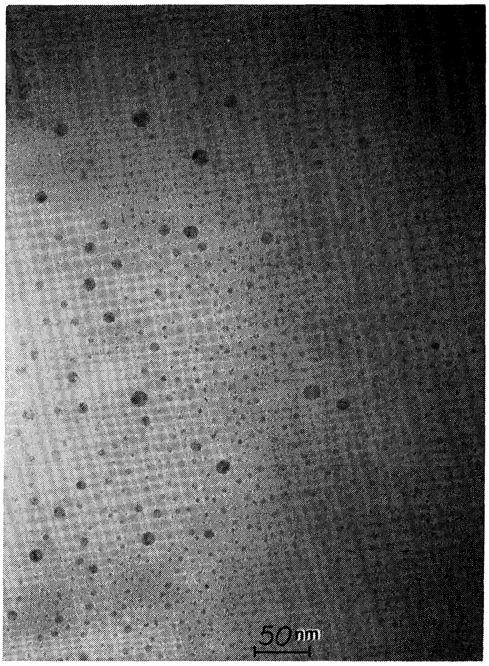

FIG. 7. Electron micrograph of a supported bimetallic Pt-Sn reforming catalyst. The metallic particles appear as well-defined dark spots, contrasting with the light background of the oxide support (Srinivasan *et al.*, 1991).

6. KINETICS OF HETEROGENEOUS CATALYTIC REACTIONS

6.1 Elementary Steps

The catalytic reactions on solid surfaces generally comprise the following sequence of steps:

1. diffusion of reactants to the surface,
2. adsorption,
3. surface reaction,
4. desorption,
5. diffusion of products from the surface.

Any one of these steps may be slow enough to become rate limiting. If steps 1 or 5 limits the rate, the reaction is said to be transport limited. If any of the other steps involving chemistry limit the rate, this chemical step would be called the rate-determining step (RDT; Boudart and Djega-Mariadassou, 1984).

6.1.1 Unimolecular Reactions on Surfaces. If a reactant A adsorbs and reacts on a catalytic site (*), the following set of elementary steps can be written:

1. adsorption; $A + * \rightarrow A*$
2. desorption; $A* \rightarrow A + *$
3. surface reaction; $A* \rightarrow$ Products.

If the surface reaction is much more rapid than the rates of adsorption and desorption, the reaction rate will be determined by the rate of adsorption. On the other hand, if the surface reaction is slow compared with adsorption-desorption, then steps 1 and 2 can be considered in quasiequilibrium and the reaction rate will be determined by step 3. Thus,

$$r = k_S \theta_A, \tag{15}$$

where k_S is the surface reaction coefficient and θ_A is the fractional coverage in equilibrium with the concentration of A in the gas phase. Then

$$r = k_S K_A P_A / (1 + K_A P_A). \tag{16}$$

At low concentrations of A in the gas phase, $1 \gg K_A P_A$. Then Eq. (16) becomes

$$r = k_S K_A P_A \tag{17}$$

and the reaction order is one.

The surface reaction coefficient, k_S, can be written in an Arrhenius form as

$$k_S = A_0 \exp(-E_S/RT), \tag{18}$$

A_0 being the preexponential factor and E_S the activation energy for the surface reaction, and K_A as

$$K_A = \exp(-\Delta G/RT)$$
$$= \exp(\Delta S/R) \exp(-\Delta H_A/RT), \tag{19}$$

where ΔG_A, ΔS_A, and ΔH_A are, respectively, the changes in free energy, entropy, and enthalpy for the adsorption of A. Then Eq. (17) can be written as

$$r = A_0 P_A \exp(-E_S/RT) \exp(\Delta S/R)$$
$$\times \exp(-\Delta H_A/RT), \tag{20}$$

which can be rearranged to give

$$r = A' \exp[-(E_S + \Delta H_A)/RT]. \tag{21}$$

Then, the activation energy determined experimentally would not be the true activation energy, E_S, but rather an apparent activation energy,

$$E_{\text{app}} = E_S + \Delta H_A. \tag{22}$$

As described in Sec. 3, in most cases $\Delta H_A < 0$. Then, E_{app} can be expressed in terms of the absolute value of the heat of adsorption, λ_A:

$$E_{\text{app}} = E_S - \lambda_A. \tag{23}$$

On the other hand, at high concentrations of A, $1 \ll K_A P_A$. Then

$$r = k_S. \tag{24}$$

In this case, the reaction order is zero and $E_{\text{app}} = E_S$. Table 2 summarizes rate expressions and apparent activation energies for bimolecular reactions for different heats of adsorption of reactants.

7. MECHANISMS OF HETEROGENEOUS CATALYSIS

There are different levels of sophistication with regard to description of mechanisms of heterogeneous catalysis (Haller and Delgass,

Table 2. Rate expressions and apparent activation energies for bimolecular reactions.

$\Delta H_{\text{ads}}(A)$	$\Delta H_{\text{ads}}(B)$	r	E_{app}
Moderate	Moderate	$\dfrac{k_{SA}k_{SB}K_AK_BP_AP_B}{(1+K_AP_A+K_BP_B)^2}$	
High	Low	$\dfrac{k_{SA}k_{SB}K_BP_B}{K_AP_A}$	$E_S-\lambda_B+\lambda_A$
Low	Low	$k_{SA}k_{SB}K_AK_BP_AP_B$	$E_S-\lambda_B-\lambda_A$
Low	Low (inhibitor X strongly absorbed)	$\dfrac{k_{SA}k_{SB}K_BP_BP_AP_B}{(K_XP_X)^2}$	$E_S-\lambda_B-\lambda_A+2\lambda_X$

1986). The most elementary mechanism is a sequence of elementary steps that accomplishes the overall reaction. As an example, we can consider the oxidation of CO catalyzed by noble metals,

$$CO+{*}\leftrightarrow{*}CO, \tag{25}$$

$$O_2+2{*}\rightarrow 2{*}O, \tag{26}$$

$${*}CO+{*}O\rightarrow CO_2+2{*}. \tag{27}$$

Here, we have assumed that CO is *associatively* and *reversibly* absorbed and O_2 is *dissociatively* and *irreversibly* adsorbed at the reaction temperature. Irreversibility implies a negligible rate of desorption at the given temperature; there will always be a finite probability of desorption which will become rapid with sufficient increase in temperature. It is likely that the second step in this mechanism is more complicated than indicated since there is good evidence that O_2 initially adsorbs as a molecular *precursor* to dissociative chemisorption. This is an elementary mechanism, i.e., a hypothesis of how the reaction occurs, but it leaves out a lot that we know or would like to know. At the next level we would like to know something about the details of the bonding of the intermediates to the surface and the energetics of this bonding. For example, from infrared experiments we know that the predominant chemisorbed species is a linear bonded CO on small Pt particles, e.g., Pt–CO, bonded through carbon, but that on Pd particles CO is likely to be bridge bonded, e.g., Pd_2CO. Note that we are suggesting the beginning of structure about the site where CO is chemisorbed, but it is not very definite, and at temperatures where this reaction occurs at maximum rate, this detail might not be important since the rate of diffusion is rapid and there may not be a particular bonding structure of the CO intermediate. This may be one reason why this reaction appears to be *structure insensitive,* i.e., occurs at the same rate on different crystal planes of Pt, Pd, Rh, etc., or different metal particle sizes. Not all reactions are so structure insensitive. For example, CH_3–CH_3 hydrogenolysis to CH_4 is *structure sensitive,* i.e., it occurs at different rates on different metal crystal planes or on different particle sizes. It should also be appreciated that in the above elementary mechanism the sites for CO and for oxygen-atom chemisorption, while both are represented by *, are not the same. On the high-density Pt surfaces, we know from LEED that the oxygen atom sits in a threefold hollow of Pt atoms.

Perhaps the most sophisticated mechanism description that we can aspire to is one in which not only do we know the structure of intermediates and the sites to which they are bonded as well as the bond energies, but we also have comparable information about the activated complex. If we have such detailed information about the motion of the atoms as they form the activated complex from reactants and decompose into the products, along with the potential energies between the activated complex and the surface, we can say that we understand the *dynamics* of the reaction. This implies that we know how the energy between the top of the barrier to reaction (the activated complex) and the ground state of the product (in the gas phase if it is not chemisorbed, the case for the CO_2 product for the above reaction) is distributed between the surface and the product; that we know how that energy carried away by the product molecule is distributed between translation, rotation, vibration, and electronic excitation; that we know the angular dependence of the desorption trajectories, etc. Such a complete mechanism is not known for any reaction, but considerable progress toward this goal has been made for chemisorption of some simple molecules, e.g., CH_4 on metals, and for the CO oxidation on Pt that we have been discussing.

8. CHEMICAL ENHANCEMENT/ POISONING OF CATALYTIC ACTIVITY

8.1 Catalyst Promoters

A catalyst promoter is any substance which, although it has no catalytic activity by itself, when added to a catalyst improves the activity, selectivity, or stability. The nature of promoter action can vary widely. Some examples of catalyst promoters are given in Table 3. The improvement of catalytic properties by the addition of a promoter may be due to

- increase in activity of the catalyst by modification of the electronic or geometric structure of the active site;

Table 3. Some catalyst promoters and their functions.

Promoter	Function
Re	Decreases catalyst coking in reforming processes
S	Decreases hydrogenolysis in reforming processes
HCl	Increases acidity of the support in reforming catalysts
K	Increases selectivity towards longer-chain hydrocarbons in CO hydrogenation reactions
K_2O	Increases rate of N_2 dissociation in ammonia synthesis on Fe
KCl	Increases reaction rate and stability of $CuCl_2$ catalysts in oxychlorination reactions
Al_2O_3	Prevents sintering of Fe in ammonia synthesis catalysts and stabilizes activity
Cl_2	Increases selectivity in ethylene oxidation on Ag catalysts

- increase in catalytic activity by creation of new sites;
- increase in selectivity towards desired products by selective deactivation of undesired paths;
- interaction of the promoter with the reaction intermediates modifying their energetics;
- stabilization of particular surface species.

8.2 Catalyst Poisons

A catalyst poison is any substance which makes the catalyst lose activity (Butt and Petersen, 1988). In general, a poison is an adsorbate which reacts permanently with the active site (strong adsorption). A poison can be introduced in the catalyst itself or in the reactant feed. Metal catalysts are very sensitive to poisons. Metal poisons can be metallic or nonmetallic. In both cases, the degree of toxicity depends on the ability of the *d* orbitals of the catalyst to interact with those of the poison. For example, the presence of small quantities of sulfur may rapidly deactivate a catalyst by formation of surface sulfides. Also, oxygenated compounds (air, water, carbon dioxide) may act as poisons for metal catalysts. Poisoning of acid catalysts can be explained in terms of the basicity of the poison. For example, a strongly basic organic molecule, such as quinoline, has a much higher toxicity for acid catalysts than aniline, a weaker basic molecule.

A special type of poisoning which is very frequently encountered in catalytic reactions is called self-poisoning. In this case, the poisoning is caused by the reaction itself. The reactants, the products, or the reaction intermediates may act as precursors of poisoning. The most common type of self-poisoning is known as *coking*. This is a general term which describes the deposition of carbonaceous deposits onto the catalyst surface, causing a decrease in activity. However, not all of the carbonaceous deposits have the same deactivating effects. Some authors distinguish between "soft coke" (high H/C ratio) and "hard coke" (low H/C ratio). The former can be considered as a strongly dehydrogenated hydrocarbon deposit while the latter has a pseudographitic structure.

In continuous operation, the activity loss due to coke formation has to be overcome by an increase in the operating temperature. When this temperature is too high, the catalyst must be regenerated by burning off the coke. Coke may not only affect the activity of a catalyst, but it may also affect selectivity. In general, structure-insensitive reactions, e.g., dehydrogenation, hydrogenation, are less affected by coke than structure-sensitive reactions, e.g., hydrogenolysis.

9. HETEROGENEOUS CATALYTIC REACTORS

There exist several criteria by which we classify reactors. For example, according to the mode of operation they can be classified as batch, semibatch, or continuous. According to the flow pattern used, they can be of the plug flow type, perfect mixing type, or intermediate. Plug flow (characterized by the fact that the flow of fluid is so ordered that no element of fluid mixes with another element ahead or behind it) or perfect mixing (mixing of elements of fluid on the molecular level) flow patterns are called "ideal" in the sense that the flow can be completely described on a molecular level. Although they cannot exactly describe the behavior of real reactors, they make possible a simple and relatively accurate mathematical treatment that approximates many real reactors.

A large number of heterogeneous catalytic reactors have flow patterns which can be described as plug flow. Among them we can find fixed-bed and moving-bed reactors. Fixed-bed reactors are the catalytic reactors most commonly found in industrial operations. They cover a large variety from packed-bed reactors, such as those used for naphtha reforming, to catalytic wall reactors, such as the catalytic converters used for controlling automobile exhausts. Among the fixed-bed type, we can also mention the metal gauze reactors, such as those used for the oxidation of ammonia in nitric acid plants. Another type of fixed-bed reactor is the trickle bed. As opposed to the previously mentioned reactors, in this case the reactants are in both the liquid and the gas phase, e.g., a liquid olefin being hydrogenated on a solid catalyst.

Moving-bed reactors find many applications in those cases in which the catalyst loses activity very rapidly and has to be continuously reactivated. In moving-bed reactors, the catalyst is continuously drawn off, regenerated, and reinjected. An example of these reactors is the Thermofor (bead moving bed) catalytic cracking process in petroleum refining.

Fluidized-bed reactors have a very complicated flow pattern and cannot be described as plug-flow reactors. They are frequently used to improve heat or mass transfer and also when the catalyst has to be continuously regenerated. Examples of these type of reactors are found in oxidation, oxychlorination, ammoxidation, and catalytic cracking processes.

10. INDUSTRIAL CATALYTIC PROCESSES

Catalysts play an essential role in many chemical, petrochemical, and petroleum refining processes. Some industrially important catalytic processes are summarized in Table 4.

10.1 Pollution Control

The abatement of nitrogen oxides (NO_x), carbon monoxide (CO), and unburned hydrocarbons is carried out on catalytic converters. These catalytic devices are used for both industrial operations (stationary sources) and automobile exhausts (mobile sources). The conditions to be met by these catalytic devices are indeed rigorous. They should be able to remove carbon monoxide, hydrocarbons, and NO_x at very high spatial velocities, reaching conversions close to 100%. They should resist cyclic variations in temperature and gas composition, maintaining their activity for at least 50 000 miles or five years, without regeneration. Since emissions are higher when the engine is cold, the catalyst should be effective at the lowest possible temperature.

10.2 Selective Oxidation

Catalytic reactions involving hydrocarbons and oxygen are frequently found in industrial processes. They can be classified into total oxidation, selective (partial) oxidation, ammoxidation, oxychlorination, oxidative coupling, and oxidative dehydrogenation. Among the numerous selective oxidation reactions found in the petrochemical industry, we can mention the production of (a) ethylene oxide by oxidation of ethylene over promoted silver catalysts at temperatures below 250 °C, (b) acrolein by oxidation of propylene over bismuth molybdate catalysts, (c) maleic anhydride by oxidation of butadiene over vanadium oxide catalysts, etc.

10.3 Cracking

The conversion of large molecules present in crude oil into smaller hydrocarbons in the gasoline range is called cracking. Cracking processes were originally noncatalytic, but after the development of highly efficient catalysts, most of them were rapidly switched to the catalytic alternative. The cracking reac-

Table 4. Some industrially important catalytic processes.

Process	Main reactions	Catalysts used
Air pollution control	Oxidation of CO, unburned hydrocarbons NO_x reduction	Pt-Pd-Rh/monolithic supports
Ammonia synthesis	$1/2N_2+3/2H_2 \rightarrow NH_3$	Fe promoted with alkali, alumina, etc.
Cracking	Cracking of larger paraffins, naphthenes, or olefins to smaller olefins Dealkylation of alkylaromatics Isomerization of olefins Condensation of alkylaromatics	Solid acids (zeolite crystals in silica-alumina matrix)
Oxidation	Total oxidation Selective (partial) oxidation Ammoxidation, oxychlorination Oxidative coupling Oxidative dehydrogenation	Platinum, bismuth molybdate, vanadium oxide, copper chloride, multicomponent oxides, chromia-alumina
Polymerization	Small olefins to polymers (polyethylene, polypropylene, etc.)	Transition-metal salts (Ti, V, Cr, Mg, etc.) and metal alkyls of Al
Reforming	Dehydrogenation of cyclohexane Isomerization of *n*-alkanes Dehydroisomerization of alkylcyclopentanes Dehydrocyclization of alkyls Hydrogenolysis of alkanes	Pt promoted with Re, or Ir, supported on chlorided alumina

tions involve the rupture of C–C bonds and mainly proceed through surface carbenium-ion intermediates on acid sites. The improvement of the activity of the catalysts over the years, as a consequence of intense research work in the field, has helped to reduce the size of the cracking reactors. The cracking process is carried out at temperatures about 550 °C in a riser-tube reactor. In that tube, the feed contacts the catalyst for a short time (about 5 s). During that time, high conversions are reached while the catalyst rapidly loses its activity through coke formation. Subsequently, the catalyst is separated and regenerated in a fluidized-bed reactor where the coke is burned off. The cracking process does not produce gasoline with a sufficiently high octane number to meet the market demands. To obtain a higher octane number the concentration of aromatics and branched alkanes must be increased. This is accomplished in a catalytic process known as naphtha reforming.

10.4 Reforming

The reforming catalysts contain small amounts of platinum, or a combination of platinum with rhenium, tin, or iridium, supported on an acidic alumina carrier. In these catalysts, the support plays a decisive role since it participates in the reactions, providing the acid sites on which some of the reactions occur. The hydrogenation/dehydrogenation reactions occur on the metallic sites, while the molecular rearrangement takes place on the acid sites. The catalysts are therefore called *bifunctional.* The reforming process is carried out in a series of fixed-bed reactors at temperatures about 450–500 °C and pressures between 10 and 35 atm.

11. SUMMARY

Heterogeneous catalysis is an interdisciplinary field involving the efforts of chemical engineers, material scientists, chemists, physicists, etc. This is so because it is an applied science which is advanced by an understanding of chemical reactions at the molecular level on well-defined surfaces (e.g., single crystals) and which is practiced in reactors that are from fractions of a meter to tens of meters in length, e.g., a refinery. We have briefly surveyed some of the fundamentals and outlined a sampling of applications, mostly related to petroleum processing.

It should be repeated that catalysis is by definition a chemical kinetic phenomenon and, as such, must be considered in any chemical reaction. It is very likely that catalysis will have been introduced either in a Darwinian way in nature or by empirical optimization by man. In either case, it will be the goal of the catalytic scientist both to try to understand the fundamental chemistry and to improve on the existing catalyst. It is generally agreed that these two goals are strongly coupled, i.e., if we understand a catalytic reaction on a molecular mechanistic level, we will then have the basis for designing an improved catalyst.

GLOSSARY

Adsorption: Accumulation of any particular substance on the surface of a material without penetration into the bulk of the material, in a concentration which is higher than that in the bulk fluid. The term associative adsorption indicates that the molecule adsorbs without undergoing decomposition. Dissociative adsorption indicates that the molecule dissociates on the surface upon adsorption.

Adsorption Site: A particular location on the solid surface where the adsorption is more probable to occur.

Bifunctional Catalysis: Catalysis that involves different functions on different sites which are physically separated: a metallic function such as hydrogenation/dehydrogenation, and an acid function such as olefin structural isomerization.

Catalyst: A substance which accelerates a chemical reaction but is not consumed or permanently changed during the process.

Chemisorption: Adsorption process involving enthalpy changes of the order of chemical bond energies.

Coke: A carbon-rich deposit formed on the surface of a catalyst by hydrocarbon reactants/products. It does not have a well-defined stoichiometry or molecular weight.

Desorption: Evolution of an adsorbate from the surface to the gas phase.

Elementary Step: A reaction step written as it takes place at the molecular level.

Physisorption: Adsorption process involving enthalpy changes of the order of liquefaction heats.

Plug Flow: Ordered flow characterized by the fact that for any given element of fluid, no element ahead of or behind it becomes mixed with it.

Precursor State: Intermediate state in the adsorption process in which the adsorbing molecule has lost one degree of freedom compared to the gas phase but does not yet occupy a localized adsorption site.

Reversible or Irreversible Adsorption: These terms distinguish adsorption processes in which the adsorbate desorbs (reversibly) or stays on the surface (irreversibly) upon lowering of the gas partial pressure over the solid surface.

Support: A generally inert material of high surface area used to stabilize high dispersions (small particle size) of the active phase. In some cases it may participate in the reaction (see bifunctional catalysts).

Structure Sensitivity: Effect of particle size or crystallographic plane of the catalyst on the turnover number (TON) of a particular reaction.

Turnover Number (TON): Specific rate of a catalytic reaction expressed as the number of molecules reacted per active site in unit time.

List of Works Cited

Bond, G. C. (1987), *Heterogeneous Catalysis: Principles and Applications,* Oxford: Clarendon.

Boudart, M., Djega-Mariadassou, G. (1984), *Kinetics of Heterogeneous Catalytic Reactions,* Princeton, NJ: Princeton Univ. Press.

Butt, J. B., Petersen, E. E. (1988), *Activation, Deactivation and Poisoning of Catalysts,* New York: Academic.

Burwell, R. L., Jr. (1983), in: B. H. Davis and W. P. Hettinger, Jr. (Eds.), *Heterogeneous Catalysis: Selected American Histories,* ACS Symposium Series 222, Washington, DC: American Chemical Society.

Haller, G. L., Delgass, W. N. (1986), in: C. F. Bernasconi (Ed.), *Techniques of Chemistry,* Vol. VI, *Investigation of Rates and Mechanisms of Reactions,* New York: Wiley.

Le Page, J. F. (1987), *Applied Heterogeneous Catalysis: Design, Manufacture, Use of Solid Catalysts,* Paris: Editions Technip, IFP Publications.

Rideal, E. K., Taylor, H. S. (1919), *Catalysis in Theory and Practice,* New York: MacMillan.

Srinivasan, R., Rice, L. A., Davis, B. H. (1991), *J. Catal.* **129**, 257–268.

Tamm, P. W., Schmidt, L. D. (1969), *J. Chem. Phys.* **51**, 5352–5363.

Yates, J. T., Jr., Duncan, T. M., Worley, S. D., Vaughan, R. W. (1979), *J. Chem. Phys.* **70**, 1219–1224.

Further Reading

Gasser, R. P. H. (1985), *An Introduction to Chemisorption and Catalysis,* Oxford: Clarendon. An excellent introduction to surface science as it applies to heterogeneous catalysis.

Delgass, W. N., Haller, G. L., Kellerman, R., Lunsford, J. H. (1979), *Spectroscopy in Heterogeneous Catalysis,* New York: Academic. A detailed description of the application of seven of the most-used spectroscopies.

Richardson, J. T. (1989), *Principles of Catalyst Development,* New York: Plenum. An introduction to practical catalyst development.

Satterfield, C. N. (1980), *Heterogeneous Catalysis in Practice,* New York: McGraw-Hill. A comprehensive text on industrial heterogeneous catalysis.

Tamaru, K. (1978), *Dynamic Heterogeneous Catalysis,* London: Academic. A specialist review of transient techniques for analysis of heterogeneous catalysis mechanisms.

Somorjai, G. A. (1981), *Chemistry in Two Dimensions: Surfaces,* Ithaca, NY: Cornell Univ. Press. A comprehensive review of surface science.

CATASTROPHE THEORY

ROBERT GILMORE, *Department of Physics and Atmospheric Science, Drexel University, Philadelphia, Pennsylvania, U.S.A.*

Introduction 85
1. **What It Is** 87
1.1 The Program of Catastrophe Theory 88
1.2 Three Theorems from Elementary Calculus 89
1.2.1 Implicit Function Theorem 89
1.2.2 Morse Lemma 89
1.2.3 Thom Splitting Lemma 90
1.3 Thom Classification Theorem 90
1.4 Thom's List of Elementary Catastrophes 91
1.5 Why a List of Perturbations is Required 93
1.6 Geometry of the Fold and the Cusp 94
1.6.1 Geometry of the Fold Catastrophe 94
1.6.2 Geometry of the Cusp Catastrophe 95
1.6.3 Bifurcation Sets for the Three-Dimensional Catastrophes 97
1.7 Perturbations of Gradient Dynamical Systems 97
2. **Why It Exists** 99
2.1 A Simple Example 99
2.2 General Procedure 101
2.3 A More Complicated Example 102
3. **How It Works** 104
3.1 Catastrophe Conventions 105
3.2 Catastrophe Flags 108
3.2.1 Modality 108
3.2.2 Sudden Jumps 109
3.2.3 Inaccessibility 109
3.2.4 Sensitivity 109
3.2.5 Hysteresis 109
3.2.6 Divergence of Linear Response 109
3.2.7 Time Dilation (Critical Slowing Down, Mode Softening) 110
3.2.8 Anomalous Variance 110
3.3 The Dangers of Design Optimization 110
3.4 Elementary Catastrophes in Nonlinear Dynamics 114
Appendix: A Brief History of Catastrophe Theory 116
Glossary 117
Works Cited 119
Further Reading 119

INTRODUCTION

Push on something. It will move. Push just a little bit harder and it will move just a little bit more. This incremental response to incremental stresses is very typical. But under rare conditions a small increase in the level of stress will produce a large and dramatic response. Such a response is called a "catastrophe." This kind of behavior has been summarized succinctly in the phrase "the straw that broke the camel's back."

Although this phenomenon occurs under rare conditions, it is also "typical." That is, although it is unlikely that any *particular* straw will break the camel's back, it is certain that after enough straw has been loaded, *some* straw *will* break the camel's back.

Situations in which gradually increasing stress leads to gradually increasing response, followed by a sudden catastrophic jump to a qualitatively different response state, are all too common. Many examples can be given:

1. Under gradually increased loading a bridge sags to a greater degree, followed by a sudden collapse under the last bit of stress (Zeeman, 1977; Poston and Stewart, 1978; Gilmore, 1981).
2. As temperature is gradually decreased at

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

constant pressure, a gas gradually contracts until a certain temperature is reached, at which it condenses to its liquid (or solid) state, with a sudden and very large change in its volume (Zeeman, 1977; Poston and Stewart, 1978; Gilmore, 1981).

3. Slow variation of the aileron trim settings (wing flaps) of an airplane leads to slow change in the attitude of the plane, until a certain threshold value is passed, at which point there occurs a large change in aircraft attitude (Gilmore, 1981).
4. When a glass tube filled with a helium–neon or carbon dioxide–neon mixture is set inside a cavity formed by two highly reflecting mirrors and a gradually increasing current is passed through it, the amount of incoherent light emitted gradually increases. After a current threshold is passed, the light which is emitted increases rapidly in intensity and coherence: the laser has turned "on" (Poston and Stewart, 1978; Gilmore, 1981).
5. The amount of sunlight falling on the Earth varies gradually over about 1000 years due to variation of the Earth's orbital parameters. Sudden dramatic changes leading to the occurrence and disappearance of the Ice Ages occur during this period and seem to be precipitated only by this gradual variation of the Earth's orbital parameters (Gilmore, 1981).
6. It is even possible to conceive of small causes producing big effects in biological, economic, social, and political systems (Thom, 1975; Zeeman, 1977; Poston and Stewart, 1978).

Catastrophes are widespread, occurring throughout all fields in the scientific and engineering disciplines, and even beyond. Moreover, the mathematical description of a catastrophe follows the same procedure and draws from the same restricted set of functions, independent of whether the catastrophe occurs in the area of physics, chemistry, structural engineering, aircraft dynamics, climate dynamics, etc. The mathematical description of catastrophes, involving mathematical functions called elementary catastrophes, provides a language in terms of which discontinuous phenomena can adequately be described.

It is easy to imagine how to describe systems in which a small push gives rise to a small response. One expects the mathematics of continuity, the calculus of Newton, to be applicable. One also expects that linearization about the local state (linear response function) will give a quantitative estimate of how much the system will respond to a small push.

But what of the mathematics of discontinuity? How does one describe catastrophes? Is it necessary to give up the ideas of continuity?

Roughly speaking, the state of a system is an equilibrium—in fact, a stable equilibrium. By changing some external parameters called control parameters (stress, loading), the equilibrium is displaced. A small change in these parameters usually results in a small displacement of the equilibrium. Sometimes small parameter changes result in the appearance of new equilibria or the disappearance of old equilibria. It is the latter instance in particular that can lead to a catastrophic sudden jump. A systematic study of catastrophes is closely related to a systematic study of equilibria, and especially the appearance and disappearance of equilibria.

Families of functions depending on (control) parameters are called catastrophes when the number of equilibria they possess changes as the parameters are varied. There are only a small number of catastrophes. A very small number have been used to model sudden jumps in physical systems. We will study the mathematics of catastrophes in two steps:

1. We first study functions representing situations in which the number of equilibria is about to change. This occurs when two or more equilibria occur at the same point (become degenerate).
2. We then study the effects of perturbations on the degenerate equilibria. In fact, we identify the simplest perturbation which can reproduce the effects of the most general perturbation.

That both these programs can be carried out successfully is remarkable.

A major obstacle in applying the mathematics of catastrophe theory to physical systems is in identifying the underlying catastrophe. A multiplicity of phenomena occur in the presence of a catastrophe (catastrophe flags). The occurrence of any one of these is an indication that others are present and can be found, and that a catastrophe is ultimately responsible for all. The catastrophe flags are easy to recognize and provide an inordinate amount of information about the underlying catastrophe. This information includes the type of catastrophe, a rough indication of where the sudden jump may occur, and—most important—how to avoid it if that is a desirable objective.

This article is organized into three parts. Section 1 describes what catastrophe theory is, Sec. 2 describes how the catastrophe functions are constructed, and Sec. 3 describes how catastrophe theory is applied to the description of phenomena that occur in the science and engineering disciplines.

In Sec. 1 we describe a progression of three theorems of elementary calculus. These theorems describe local standard forms for functions in the neighborhood of a point. The first, the implicit function theorem, tells us that a function can be replaced by its linear approximation when its slope is nonzero at a point. The second theorem, the Morse lemma, tells us that under suitable conditions a function can be well approximated by a quadratic form in the neighborhood of an equilibrium. The third result, the Thom splitting lemma, describes what happens when the "suitable conditions" required above are not satisfied. In this case two or more equilibria occur very close together (are degenerate), making it possible for a small perturbation either to split or annihilate the equilibria. This change in the number of equilibria is a "catastrophe." Tables containing Thom's list of elementary catastrophes and a complete list of all elementary catastrophes are provided. This is followed by a discussion of the geometric properties of the very simplest catastrophe functions. This section provides a clear answer to one of the questions posed above: it is possible to describe discontinuous phenomena without giving up the ideas of continuity. In fact, this mathematics of discontinuity is an essential part of the calculus of Newton.

In Sec. 2 we compute some explicit catastrophe functions. The first example starts with a family of functions depending on one state variable and two control parameters. Following an algorithmic procedure, we reduce this to a standard form in the neighborhood of its most degenerate equilibrium. This degenerate equilibrium is then perturbed in order to determine how these critical points can be created and annihilated as a function of the control parameters. The algorithm developed to effect this reduction to normal form is then summarized and applied to a more complicated family of functions.

In Sec. 3 we address the question of how to apply the mathematics of catastrophe theory to real world processes that exhibit discontinuous phenomena. Two widely adopted conventions are first described. These are assumptions about the mathematical characterization of the equilibrium state of a physical system: whether it is determined by a local minimum or the global minimum of a potential, and the conditions under which a jump from one minimum to another occurs. The mathematics tells us only where, how many, and what type of equilibria a catastrophe function possesses; the convention isolates the physically important equilibria. When a physical system exhibits a catastrophe, a multiplicity of phenomena occur. It is useful to be able to recognize them—particularly if one wants to avoid the physical catastrophe (e.g., bridge collapse). These phenomena are called catastrophe flags. The presence of any one is an indication that the others are present. Their recognition provides a great deal of information about the underlying catastrophe. The use of catastrophe conventions and catastrophe flags is illustrated in the context of an important example in Sec. 3.3. This illustrates unexpected dangers which may arise in the design of structures following standard optimization criteria employed to reduce costs. In Sec. 3.4 we indicate how the elementary catastrophes may make their appearance within the broader program of catastrophe theory, in the field of dynamical systems theory. That is, the fold catastrophe appears in the guise of saddle-node bifurcations while the cusp catastrophe appears in the guise of pitchfork and Hopf bifurcations.

In a short Appendix we outline the early turbulent and confusing history of catastrophe theory, which was at one point heralded as the greatest advance in mathematics since the development of the calculus by Newton.

1. WHAT IT IS

In Sec. 1 of this article we describe the enormous mathematical program called catastrophe theory and the much smaller and more manageable mathematical program called elementary catastrophe theory (Sec.

1.1). This latter program is not only the starting point for the study of the larger program, but also a continuation of a very important and rather simple program in elementary calculus. This is the program of determining and classifying the standard, or canonical, forms that functions can assume in any of their neighborhoods.

The first two stages in this program are well known at an intuitive level (Sec. 1.2). These are the implicit function theorem and the Morse lemma, which are the mathematical justifications underlying the approximation of a function by a linear function at a point where its slope is nonzero, or by a quadratic form at an equilibrium (where the slope is zero). The third stage, the Thom splitting lemma, describes what happens when two or more equilibria become degenerate. Under this condition a perturbation can either split the equilibria or annihilate them. Change in the number of equilibria is closely related to the occurrence of sudden jumps in physical systems.

Functions that describe sudden jumps, or changes in the number of equilibria of a system, are called catastrophes. There is a small number of elementary catastrophes. They are classified in Sec. 1.3 and presented in Tables 1 and 2 (Sec. 1.4). Since the number of catastrophes is small, the properties of each can be studied in detail (Sec. 1.5); the results are then directly applicable to any physical system described by that mathematical function. The geometric properties of the two simplest of the elementary catastrophes are studied in detail in Sec. 1.6. In Sec. 1.7 we make quantitative the statement that a small push will displace an equilibrium by just a little bit by computing the linear response function for a large class of physical systems, and showing that this function diverges precisely when a sudden jump is imminent.

1.1 The Program of Catastrophe Theory

Catastrophe theory is a program. The purpose of the program is to determine how the qualitative properties of solutions of equations change as the parameters that appear in the equations change (Gilmore, 1981).

It often happens that small changes in the values of parameters that appear in equations produce only small quantitative changes in the solutions of the equations. However, there may be parameter values for which a small change, either in parameter values, initial conditions, or boundary conditions, produces a large quantitative change in the solutions to these equations. Large quantitative changes in solutions describe qualitative changes in the behavior of the system. Catastrophe theory is concerned with determining the parameter values at which qualitative changes occur in solutions of equations described by parameters (Thom, 1975; Zeeman, 1977; Poston and Stewart, 1978; Gilmore, 1981).

This is an ambitious and difficult program. For example, for systems of equations of the form

$$F_\alpha(x,x',t;c)=0, \tag{1}$$

where F_α is a set of functions; x is an n-vector, $x=(x_1, x_2,...,x_n)\in R^n$, called a state vector; c is a k-vector, $c=(c_1, c_2,...,c_k)\in R^k$, called control parameters; and $'=d/dt$, there are no general results. When the set of equations is restricted to the simpler form of coupled nonlinear first-order ordinary differential equations (also called dynamical systems) of the form

$$x_i'=f_i(x,t;c), \tag{2}$$

very little can be said in general. Many results are known when $n=2$ and f is independent of t. A few results are known when $n=2$ and the forcing term is periodic, $f(x,t;c)=f(x,t+T;c)$. Much less is known when $n=2$ and f is not periodic. The case $n>2$ invites a lifetime of work.

When the forcing function in the dynamical system equations (2) is independent of time and can be written as the gradient of some potential,

$$f_i=-\partial V(x;c)/\partial x_i, \tag{3}$$

then the system

$$x_i'=-\partial V(x;c)/\partial x_i \tag{4}$$

is called a gradient dynamical system. For such systems many results are available.

The qualitative properties of a gradient dynamical system can be constructed by investigating the phase-space portrait of its flow. This can be done by plotting the value of the potential as a function of the phase-space coordinates x_i. The phase-space flow is "downhill" on the potential function. It is easily determined in the neighborhood of each equilibrium, or critical point, independent of the stability of the equilibrium. The local flow portraits around each critical point can then be pasted together to determine a global phase portrait. The potential

$$V(x,y;a,b)=\tfrac{1}{4}x^4+\tfrac{1}{2}ax^2+bx+\tfrac{1}{2}y^2 \tag{5}$$

is illustrated in Fig. 1, together with the phase-space portrait in the state-variable space.

Small changes in the control-parameter values (a,b) typically produce small changes in the location of the critical points. In turn, this produces only small quantitative changes, and therefore no qualitative change, in the phase-space portrait. Qualitative changes will only occur when changes in the control-parameter values result in changes in the number of critical points. This number can change only when two or more critical points coalesce and annihilate or, viewed from the other direction, two or more critical points are created in phase space and then move apart from each other as the control parameters are varied.

Elementary catastrophe theory is the study of how the critical points of a potential, $V(x;c)$, move about, coalesce and annihilate each other, or are created and disperse from each other, in state space $x\in R^n$ as the control parameters $c\in R^k$ are varied.

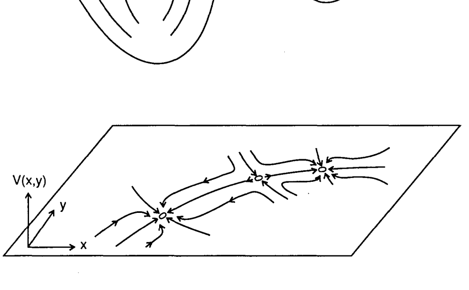

FIG. 1. The dynamics of a gradient system are governed by a potential. The potential of Eq. (5) is shown, together with the flows to and from the stable and unstable equilibria. These are projected down into the *x*-*y* plane.

1.2 Three Theorems from Elementary Calculus

Elementary catastrophe theory is found at the intersection of two lines of mathematical development, one old, one new. On the one hand, it is the latest development in the quest, in elementary calculus, for standard local forms for functions. On the other hand, it is the first result in the quest, in catastrophe theory, for canonical representations of functions that show qualitative changes when control parameters are varied.

Elementary catastrophe theory is the third in a series of reduction-to-standard-form theorems in elementary calculus. The three developments are the implicit function theorem, which depends on the first derivatives of a function; the Morse lemma, which depends on the second derivatives of a function; and the Thom splitting lemma, which depends on the third (and higher) derivatives of a function. Each of these results provides a standard, or canonical, form for a function in the neighborhood of a point.

We summarize these results now.

1.2.1 Implicit Function Theorem. The implicit function theorem tells us that if the slope of a function is nonzero at a point, the function can be represented locally by a linear approximation to that function. In a rough sense, it tells us that it is justified to linearize a function about a point at which its derivative is nonvanishing.

Implicit Function Theorem: Let $f(x)=f(x_1, x_2,\ldots,x_n)$ be a function with nonzero gradient at x_0:

$$\nabla f|_{x_0}\neq 0. \tag{6}$$

Then it is possible to find a new coordinate system, $y=(y_1,y_2,\ldots,y_n)$, $y=y(x)$, so that

$$f=y_1. \tag{7}$$

That is, f is equal, after a smooth change of coordinates, to y_1.

1.2.2 Morse Lemma. The Morse lemma takes over where the implicit function theorem leaves off. Suppose the gradient of a function does vanish at a point—what then? Such a point is called an equilibrium, or critical point. Provided that the function has enough "curvature" at the critical point, it can be represented locally by a quadratic form. In a rough sense, the Morse lemma tells us that it is justified to represent a function at an equilibrium by a quadratic form, provided none of the eigenvalues vanish.

Morse Lemma: Let $f(x)=f(x_1,x_2,\dots,x_n)$ be a function with equilibrium at x_0 and nonsingular stability matrix at x_0:

$$\begin{array}{ll} \nabla f|_{x_0}=0, & \text{equilibrium,} \\ \det[\partial^2 f/\partial x_i \partial x_j]|_{x_0}\neq 0, & \text{nonsingular.} \end{array} \qquad (8)$$

Then there is a smooth change of coordinates, $x'=x'(x)$, so that

$$f=\sum \lambda_i (x'_i)^2, \qquad (9)$$

where λ_i are the eigenvalues (all nonzero) of the stability matrix.

A critical point satisfying the conditions (8) is called variously a Morse critical point or an isolated critical point.

The quadratic form (9) can be put into a canonical form by rescaling the coordinates:

$$y_i=|\lambda_i|^{1/2} x'_i. \qquad (10)$$

Under this scale transformation the function at equilibrium assumes Morse canonical form:

$$\begin{array}{l} f=M_i^n, \\ M_i^n=-y_1^2-\cdots-y_i^2+y_{i+1}^2+\cdots+y_n^2. \end{array} \qquad (11)$$

The quadratic forms (11) are called Morse saddles. The Morse saddle M_0^n has a minimum at $y=0$, while M_n^n has a maximum at $y=0$. The remaining Morse saddles M_i^n, $i\neq 0,n$, have equilibria at $y=0$ that are neither maxima nor minima.

1.2.3 Thom Splitting Lemma. The Thom splitting lemma takes over where the Morse lemma leaves off. Suppose the stability matrix of a function is singular at an equilibrium. Then one or more eigenvalues (λ_i) vanish. What then? The Thom splitting lemma tells us that there is a smooth change of coordinates, $x'=x'(x)$, where $x'_1,\dots,x'_l$ are tangent to the eigenvectors with vanishing eigenvalues at the critical point, and $x'_{l+1},\dots,x'_n$ are tangent to the eigenvectors with nonvanishing eigenvalues at the critical point, so that the function can be broken down into two parts. One part, associated with the nonzero eigenvalues, is simple and can be put into Morse canonical form. The other part, associated with the vanishing eigenvalues, is interesting and has all its second derivatives equal to zero at the critical point. This non-Morse function is the principal object of study in elementary catastrophe theory.

Thom Splitting Lemma: Let $f(x)=f(x_1,x_2,\dots,x_n)$ be a function with equilibrium and singular stability matrix at x_0:

$$\begin{array}{ll} \nabla f|_{x_0}=0, & \text{equilibrium,} \\ \det[\partial^2 f/\partial x_i \partial x_j]|_{x_0}=0, & \text{singular.} \end{array} \qquad (12)$$

If the stability matrix has exactly l vanishing eigenvalues, then there is a smooth change of coordinates, $x'=x'(x)$, so that

$$\begin{array}{l} f(x)=f_{\mathrm{NM}}(x'_1,\dots,x'_l)+M_i^{n-l}(x'_{l+1},\dots,x'_n), \\ \partial^2 f_{\mathrm{NM}}/\partial x_i \partial x_j|_{x_0}=0, \quad 1\leqslant i,j\leqslant l. \end{array} \qquad (13)$$

The Thom splitting lemma can be proven by the methods of elementary calculus. It tells us that when the Morse lemma is not applicable, the function can be split into two functions, a "good" function in $n-l$ coordinates which can be put into Morse canonical form and a "bad," or non-Morse, function of l variables which bears further scrutiny. It tells us nothing about the non-Morse function except that its Taylor series expansion about the critical point begins with at least third-degree terms.

We emphasize here that the three results of elementary calculus, the implicit function theorem, the Morse lemma, and the Thom splitting lemma, depending on first, second, and third derivatives, are local in nature. The theorems do not provide an estimate for the size of the neighborhood for which the statement of the result is true.

1.3 Thom Classification Theorem

For a typical function, $f(x)$, the gradient at a random ("typical") point will be nonvanishing, so that the implicit function theorem is applicable. There are, however, typically isolated points at which the gradient vanishes. At such points the stability matrix is typically nonsingular, so that the Morse lemma is applicable. How, then, does it come about that the machinery of elementary catastrophe theory becomes useful?

When the function depends on control parameters c as well as state variables x, $f=f(x;c)$, then the eigenvalues of the stability matrix at a critical point, $x_0=x_0(c)$, depend on the control-parameter values: $\lambda_i=\lambda_i(c)$. As a result, there may be choices of the control-parameter values that annihilate one or more of the eigenvalues.

As a result, the structure of the non-Morse function in Eqs. (13) will depend on control parameters. The Thom classification theorem, which is outside the scope of calculus (elementary or otherwise), provides a further resolution of the non-Morse function into two functions. One of these, the catastrophe germ, depends only on the l state variables $y_1,y_2,\ldots,y_l$ and summarizes the nature of the singularity at the non-Morse critical point. The other function, the universal perturbation, is a function of both the l state variables and k control parameters. This function summarizes what can happen to the singularity, or degenerate critical point, under the most general possible ("universal") perturbation.

Thom Classification Theorem: Let $f_{\mathrm{NM}}(y;c)=f(y_1,\ldots,y_l;c_1,\ldots,c_k)$ be a non-Morse function of l state variables and k control parameters. Then there is a smooth change of coordinates so that

$$f_{\mathrm{NM}}(y;c)=\mathrm{Cat}(l,k). \tag{14}$$

The elementary catastrophe function, $\mathrm{Cat}(l,k)$, is the sum of two terms:

$$\mathrm{Cat}(l,k)=\mathrm{CG}(l)+\mathrm{Pert}(l,k). \tag{15}$$

The catastrophe germ, $\mathrm{CG}(l)$, depends only on the l state variables. All its second partial derivatives vanish at the critical point. The universal perturbation depends on the k control parameters as well as the l state variables. The dependence of $\mathrm{Pert}(l,k)$ on the control-parameter values is linear. For "most" choices of control-parameter values (all but a set of measure zero) the function $\mathrm{Cat}(l,k)$ has isolated critical points.

The Thom classification theorem, like the three results of elementary calculus described in the previous section, is local in nature. The theorem does not provide an estimate of the size of the neighborhood for which the statement of the theorem is true.

1.4 Thom's List of Elementary Catastrophes

Thom's original classification theorem provided a list of the elementary catastrophes (Thom, 1975). A slightly expanded version of this list is provided in Table 1 (Arnol'd, 1981, 1986). This list contains the canonical catastrophe functions for $k<6$ and therefore (cf. Sec. 2.3) $l<3$. This list consists of the classification of the function following the beautiful convention introduced by Arnol'd (1981), the original descriptive name, when it exists (Thom, 1975; Zeeman, 1977; Poston and Stewart, 1978), values for k and l, the catastrophe germ, $\mathrm{CG}(l)$, and the universal perturbation, $\mathrm{Pert}(l,k)$. Thom's original list contained only the seven members with $k \leqslant 4$ (dimension of spacetime) for unsupportable historical reasons.

The catastrophe functions listed in Table 1 are elementary in the sense that all coefficients in the catastrophe germ can be assigned canonical values. There are no free parameters; every coefficient in the catastrophe germ can be given canonical numerical values such as $\pm 1,0$ by a coordinate change. For example, a term of the form $-3x^4$ in the A_{-3} catastrophe could be transformed to the canonical

Table 1. All catastrophes up to control-parameter dimension five are elementary and are listed below by dimension of control-parameter space.

k	l	Classification	Name	$\mathrm{CG}(l)$	$\mathrm{Pert}(l,k)$
1	1	A_2	Fold	x^3	a_1x^3
2	1	$A_{\pm 3}$	Cusp	$\pm x^4$	$a_1x^2+a_2x^2$
3	1	A_4	Swallowtail	x^5	$a_1x^2+a_2x^2+a_3x^3$
3	2	D_{-4}	Elliptic umbilic	x^2y-y^3	$a_1x+a_2y+a_3x^2$
3	2	D_{+4}	Hyperbolic umbilic	x^2y+y^3	$a_1x+a_2y+a_3x^2$
4	1	$A_{\pm 5}$	Butterfly	$\pm x^6$	$a_1x+a_2x^2+a_3x^3+a_4x^4$
4	2	D_5	Parabolic umbilic	$\pm(x^2y+y^4)$	$a_1x+a_2y+a_3x^2+a_4y^2$
5	1	A_6	Wigwam	x^7	$a_1x+a_2x^2+a_3x^3+a_4x^4+a_5x^5$
5	2	D_{-6}	Second elliptic umbilic	x^2y-y^5	$a_1x+a_2y+a_3x^2+a_4y^2+a_5y^3$
5	2	D_{+6}	Second hyperbolic umbilic	x^2y+y^5	$a_1x+a_2y+a_3x^2+a_4y^2+a_5y^3$
5	2	$E_{\pm 6}$	Symbolic umbilic	$\pm(x^3+y^4)$	$a_1x+a_2y+a_3xy+a_4y^2+a_5xy^2$

form $-x'^4$ (or $-\frac{1}{4}x'^4$) by an appropriate scale change $x'=\lambda x$.

Thom's list of elementary catastrophes is not a complete list. The complete list is provided in Table 2. There are two infinite series of elementary catastrophes and one finite series. One infinite series, the cuspoids $A_{\pm k}$, depends on only one state variable. The other infinite series, the umbilics $D_{\pm k}$, depends on two state variables. The exceptional elementary catastrophes, $E_{\pm 6}$, E_7, E_8, depend on two state variables (Arnol'd, 1981, 1986; Poston and Stewart, 1978; Gilmore, 1981).

There is a remarkable correspondence between the classification theory of elementary catastrophes and the classification theory for Lie algebras all of whose roots (in the root space diagram) have the same length (Arnol'd, 1981, 1986; Gilmore, 1981). The correspondence is as follows. The phase-space portraits of the elementary catastrophes with maximum number of isolated critical points can be summarized by drawing the flow from each equilibrium to any of the others to which a flow is possible. The phase-space portraits so obtained are exactly the Dynkin diagrams which classify all the simple Lie algebras whose roots have equal length; these are A_{n-1}, D_n, and the exceptional simple Lie algebras E_6, E_7, and E_8. This nomenclature for simple Lie algebras has accordingly been adapted to the classification of elementary catastrophes. For Lie algebras the subscript (e.g., 8 for E_8) denotes its rank; for elementary catastrophes the subscript denotes the number of isolated (complex) critical points generated by an arbitrary perturbation of the function. This is the maximum number of real critical points into which the non-Morse critical point splits under a general perturbation.

Tables 1 and 2 differ in a subtle way, indicating that they are responses to somewhat different questions. The question to which Table 1 responds is: "Up to what control-parameter dimension are all catastrophes elementary, and what are they?" The question to which Table 2 responds is: "For each control dimension $k=1,2,\ldots$, what are the elementary catastrophes?"

The difference between the two tables indicates that for control dimension $k\geqslant 6$ there are catastrophes that are elementary and those that are not, while for $k<6$ all catastrophes are elementary. We will explore what happens at $k=6$ to generate nonelementary (modal) catastrophes as well as elementary catastrophes in Sec. 2.3. Briefly, the result is as follows. To annihilate l eigenvalues of the stability matrix requires $k\geqslant l(l+1)/2$ control parameters. A linear transformation can be used in an attempt to provide canonical values for cubic terms in the Taylor series expansion

Table 2. There are three series of elementary catastrophes. The cuspoids A_k depend on one state variable while the umbilics D_k and the exceptional catastrophes E_k, $k=6,7,8$, depend on two state variables. The subscript k indicates the maximum number of real Morse critical points the catastrophe splits into under an arbitrary perturbation. The number of control parameters required in the universal perturbation is $k-1$.

Symbol	Catastrophe germ	Universal perturbation
$A_{\pm k}{}^{a}$	$\pm x^{k+1}$	$\sum_{j=1}^{k-1} a_j x^j$
$D_{\pm k}$	$x^2y \pm y^{k-1}$, k even $\pm(x^2y+y^{k-1})$, k odd	$\sum_{j=1}^{k-3} a_j y^j + \sum_{j=k-2}^{k-1} a_j x^{j-(k-3)}$
$E_{\pm 6}$	$\pm(x^3+y^4)$	$\sum_{j=1}^{2} a_j y^j + \sum_{j=3}^{5} a_j xy^{j-3}$
E_7	x^3+xy^4	$\sum_{j=1}^{4} a_j y^j + \sum_{j=5}^{6} a_j xy^{j-5}$
E_8	x^3+y^5	$\sum_{j=1}^{3} a_j y^j + \sum_{j=4}^{7} a_j xy^{j-4}$

[a] $A_{+k}=A_{-k}$ if k is even.

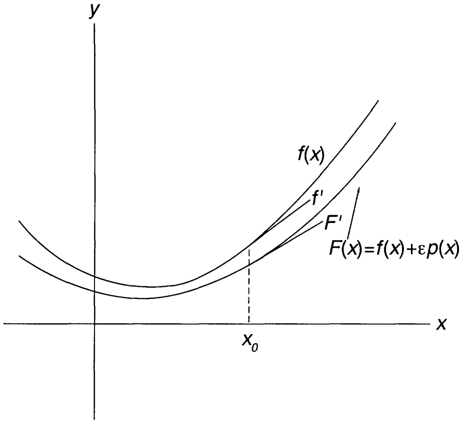

FIG. 2. At a point where the implicit function theorem is applicable, perturbation of a function does not produce a qualitative change in the function.

of the non-Morse function. When the number of cubic coefficients $[(l+3-1)!/(l-1)!3!]$ exceeds the number of degrees of freedom in the $l \times l$ linear transformation (l^2), then it is not possible to assign canonical values to all cubic terms in the Taylor series expansion of the non-Morse function, and the resulting catastrophe germ cannot be elementary. This first occurs for $l=3$ ($\Rightarrow k=6$). For $k \geqslant 6$ there are catastrophes that are elementary and those that are not. For $k<6$ all catastrophes are elementary.

1.5 Why a List of Perturbations is Required

The first two canonical form theorems of elementary calculus are clean and simple. If the function has certain properties at a point, then the canonical form in the neighborhood of the point is provided by the statement of the theorem (implicit function theorem, Morse lemma). By contrast, the third result is not nearly so clean cut. The Thom splitting lemma tells us that we can decompose a function at a non-Morse critical point into the sum of two functions, one Morse, the other interesting. The classification theorem provides a list of the interesting functions by number of state variables (l) and control parameters (k). Why is it that the classification theorem, in addition to providing a list of canonical forms for catastrophe germs, in the spirit of the implicit function theorem and the Morse lemma, also provides a list of canonical perturbations? The reason is that perturbation of the canonical linear form which is provided by the implicit function theorem does not change its qualitative properties. The same is true for the canonical quadratic form which is provided by the Morse lemma. However, perturbation of the canonical singularity CG(l) provided in Tables 1 and 2 produces dramatic changes in its qualitative properties. Different perturbations produce different qualitative changes. The canonical perturbation, Pert(l,k), of each catastrophe germ is the "smallest" function, in the sense of number of control parameters required, which incorporates all distinct qualitative changes produced by all possible perturbations of the catastrophe germ.

We illustrate these statements in Figs. 2–4. In Fig. 2 we show a function, $f(x)$, which satisfies the conditions of the implicit function theorem at x_0. Under a perturbation, $\varepsilon p(x)$, where $p(x)$ is a well-behaved function and ε is a small parameter, the new function, $F(x) = f(x) + \varepsilon p(x)$, also satisfies the conditions of the implicit function theorem at x_0 for ε sufficiently small. Therefore, perturbation of $f(x)$ at x_0 does not change its qualitative properties (perturbation "commutes" with the implicit function theorem).

In Fig. 3 we show a function, $f(x)$, which satisfies the conditions of the Morse lemma at x_0. Under a perturbation, $\varepsilon p(x)$, the new function, $F(x) = f(x) + \varepsilon p(x)$, no longer appears to satisfy the conditions of the Morse lemma at x_0, since typically $p'(x_0) \neq 0$. However, $F'(x_0) = \varepsilon p'(x_0)$ is small for small ε so

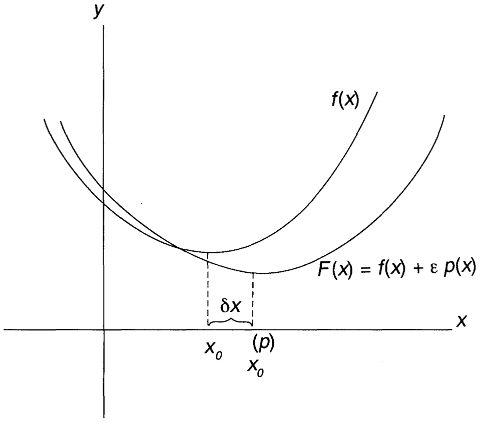

FIG. 3. At a point where the Morse lemma is applicable, perturbation of a function does not produce a qualitative change in the function.

that the implicit function theorem is on the verge of not being applicable. It is more useful to regard the perturbation as moving the location of the critical point to a nearby point, $x_0^{(p)} = x_0 + \delta x_0$. At this perturbed critical point the conditions for the Morse lemma are fulfilled. Thus, a small perturbation at an equilibrium produces only a small displacement of that equilibrium. Not only that, but the curvature or, more generally, the Morse saddle type of the canonical form remains unchanged. As a result, perturbation of a function that satisfies the conditions for the Morse lemma does not produce a qualitative change of the function in the neighborhood of the critical point.

The situation is quite different for a catastrophe germ. In Fig. 4 we plot the catastrophe function A_2: $f(x;a_1)=\frac{1}{3}x^3+a_1x$ for three values of the control parameter a_1. For $a_1=0$ the catastrophe germ $f(x;a_1=0)=x^3/3$ has a doubly degenerate critical point at $x=0$. The perturbation with $a_1<0$ splits this doubly degenerate critical point into two isolated critical points at $x_-=-(-a_1)^{1/2}$ and $x_+ =+(-a_1)^{1/2}$. The perturbation with $a_1>0$ removes the critical point altogether. These are the only two qualitatively distinct things that can occur to a doubly degenerate critical point under an arbitrary perturbation. These correspond to scattering of the solutions of $\nabla f(x;a_1)=0$ from the real axis to the imaginary axis as a_1 increases through zero, with the point of double degeneracy occurring at $a_1=0$. It is a remarkable result that the simple perturbation $P(1,1)=a_1x$ encapsulates all distinct possibilities under generic perturbations.

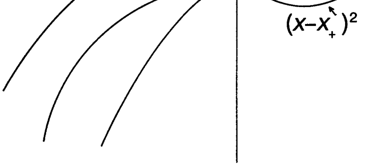

FIG. 4. At a point where a function has a degenerate critical point, so that neither the implicit function theorem nor the Morse lemma is applicable, perturbation of a function produces a qualitative change in the function. In the case shown, perturbation of the function x^3 either annihilates the critical points or splits them into two nondegenerate critical points.

It should now be clear why the catastrophe germs listed in the classification theorem must be accompanied by a list of universal perturbations while the implicit function theorem and the Morse lemma are not encumbered by such baggage. The canonical linear and quadratic forms are invariant under perturbation: perturbation produces no qualitative change. However, the catastrophe germs undergo a wide spectrum of distinct qualitative changes under perturbation. The perturbation functions listed are those of minimal control-parameter dimension which are capable of reproducing the entire spectrum of distinct qualitative changes induced by the most general perturbation.

It is a remarkable result that the control parameters appear linearly in these perturbations.

1.6 Geometry of the Fold and the Cusp

In this section we review the properties of the two simplest elementary catastrophes, the fold catastrophe A_2 and the cusp catastrophe $A_{\pm 3}$. Since the cusp catastrophe A_{+3} occurs more frequently in physical applications than A_{-3} (which is not bounded below), we study specifically the properties of A_{+3}. The properties of A_{-3} are related by appropriate sign changes. We also review a restricted set of properties of the co(ntrol)-dimension three catastrophes A_4, $D_{\pm 4}$.

For the fold and the cusp we study the following properties:

1. typical functions in the family of functions as well as the bifurcation set;
2. location of the critical points;
3. values of the function at the critical points;
4. curvature of the function at the critical points.

We present only the bifurcation set for the three catastrophes A_4, $D_{\pm 4}$.

1.6.1 Geometry of the Fold Catastrophe.

The fold catastrophe is

$$A_2:\ f(x;a)=\tfrac{1}{3}x^3+ax. \qquad (16)$$

The canonical properties of this function are shown in Fig. 5. In Fig. 5(a) we show members of this family with $a>0$, $a=0$, $a<0$. The bifurcation set is the set of points in the control parameter space at which there is a qualitative change in the nature of the function. This occurs when two or more critical points become degenerate. For the fold catastrophe this consists of the single point $a=0$, at which there is a doubly degenerate critical point at $x=0$.

The location of the critical points, the solution of $\nabla f(x;a)=0$, is shown in Fig. 5(b). The critical points, $x_\pm(a)=\pm(-a)^{1/2}$, have a standard 1/2 power-law dependence on the control parameter a. Note that the critical points $x_\pm(a)$ exist only for $a\leqslant 0$ and that the graph of $x_\pm(a)$ as a function of a is a smooth manifold embedded in the space R^1 (state space) $\times R^1$ (control parameter space). This is a general result. The fold catastrophe derives its name from the shape of its critical set $\nabla f(x;a)=0$, which looks like a curve folded over itself.

The value of the function at the critical points,

$$f_c(x_\pm(a);a)=\pm\tfrac{2}{3}(-a)^{3/2}, \tag{17}$$

is plotted in Fig. 5(c). This curve has a canonical 3/2 power-law dependence. This graph is not generally a manifold.

The curvature of the function,

$$f''(x_\pm(a);a)=\pm 2(-a)^{1/2}, \tag{18}$$

is shown in Fig. 5(d). Although the critical curvature happens to be a manifold in the present case, this is not generally true for the remaining catastrophes.

1.6.2 Geometry of the Cusp Catastrophe.

The cusp catastrophe is

$$A_{+3}\colon\ f(x;a,b)=\tfrac{1}{4}x^4+\tfrac{1}{2}ax^2+bx. \tag{19}$$

The canonical properties are shown in Fig. 6. In Fig. 6(a) we show the control-parameter plane $R^2=(a,b)$, various points in this plane, and the function $f(x;a,b)$ evaluated at these points. Within the cusp-shaped region the

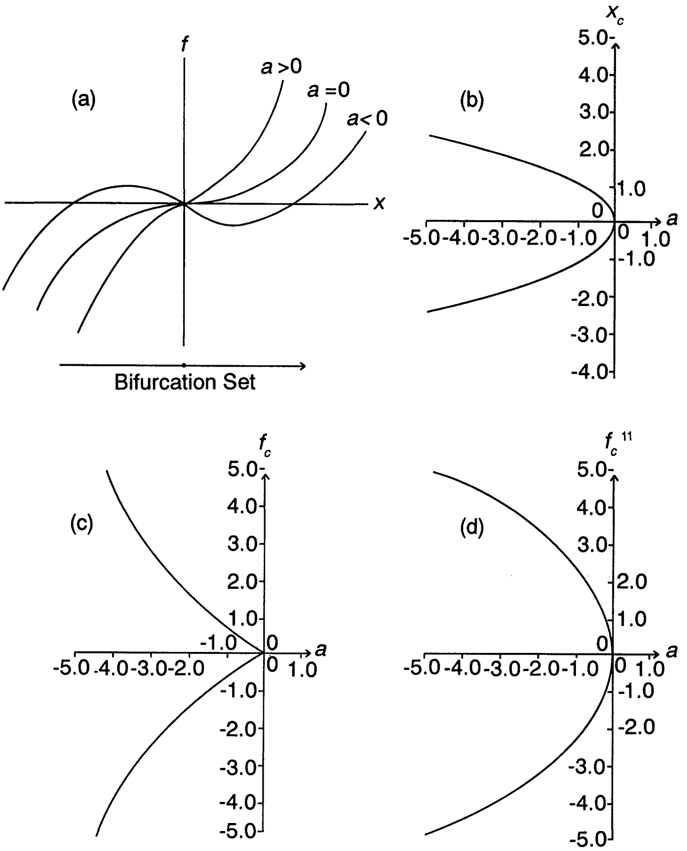

FIG. 5. (a) Members of the fold family of functions $f(x;a)=\frac{1}{3}x^3+ax$ for various values of the control parameter a. (b) Location of the critical points as a function of a. (c) Value of the function at its critical points. (d) Curvature of the function at its critical points.

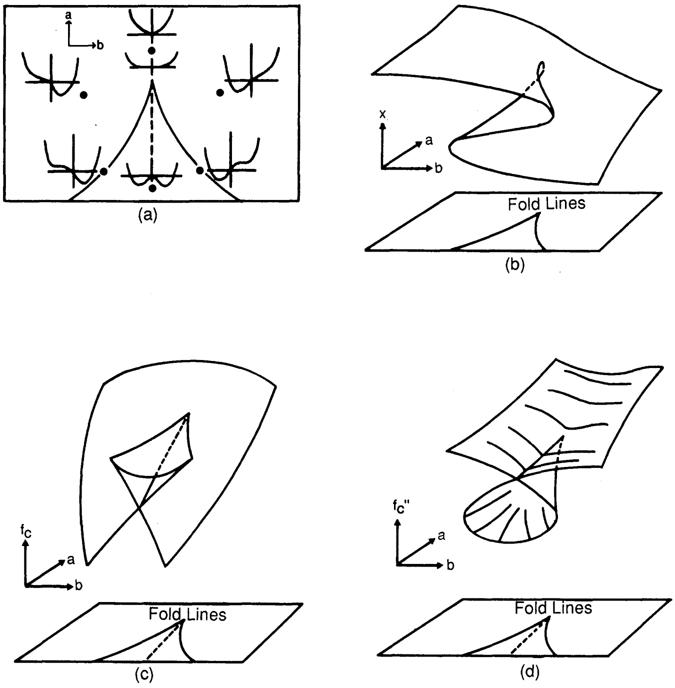

FIG. 6. (a) Members of the cusp family of functions $f(x;a,b)=\frac{1}{4}x^4+\frac{1}{2}ax^2+bx$ for various values of the control parameters *a,b*. (b) Location of the critical points as a function of position over the control parameter plane. (c) Value of the function at its critical points. (d) Curvature of the function at its critical points.

function has three isolated critical points, two minima separated by a local maximum. Outside the cusp-shaped region the function has only a single minimum. These two regions in the control plane parametrize two qualitatively distinct types of functions. Any path from one region to the other must pass through the cusp-shaped curve, along which there is a doubly degenerate critical point (triply degenerate at the tip of the cusp). This degeneracy occurs when the local maximum collides with one of the two minima. Entering the larger region complementary to the cusp-shaped region, the two degenerate critical points annihilate each other in a catastrophe which is of type A_2.

The cusp-shaped bifurcation set is determined by the condition that a critical point $[\nabla f(x;a,b)=0]$ is degenerate $[\nabla^2 f(x;a,b)=0]$:

$$\begin{aligned} &\text{critical point:} \quad f'(x;a,b)=x^3+ax+b=0,\\ &\text{degenerate:} \quad f''(x;a,b)=3x^2+a=0. \end{aligned} \tag{20}$$

From these two equations we compute the semicubical parabola

$$\begin{aligned} a&=-3x_c^2,\\ b&=2x_c^3, \end{aligned} \tag{21a}$$

along which a critical point is degenerate. The projection of this space curve into the control parameter plane is

$$(a/3)^3+(b/2)^2=0. \tag{21b}$$

This is the bifurcation set shown in Fig. 6(a).

In Fig. 6(b) we show the critical point(s), $x_c(a,b)$, as a function of the control parameters *a,b*. These points lie on the critical manifold or catastrophe manifold $\nabla f(x;a,b)=0$. Outside the cusp-shaped region there is a single critical point. Over the cusp-shaped region there are three. The middle critical point is the local maximum which separates the two minima. Moving toward the edge of the cusp-shaped region, two of the critical points move together. They collide on the bifurcation set and annihilate each other beyond the bifurcation set. The graph $x_c(a,b)$ in R^1 (state space) $\times R^2$ (control parameter space) is a smooth two-dimensional manifold. The locus of points on this manifold where the tangent is "vertical" is the semicubical parabola (21). From another point of view, the cusp-shaped bifurcation set in R^2 is the projection into R^2 of the fold in the manifold $x_c(a,b)$ in $R^1\times R^2$. The singularity in this catastrophe lies not in the catastrophe manifold itself, which is smooth, but in the projection of this two-dimensional manifold down into the two-dimensional control parameter space. In general, the graph of $\nabla f(x;c)=0$, with $x \in R^n$ and $c \in R^k$, is a smooth k-dimensional manifold embedded in $R^n \times R^k$. The only singularity occurs in the projection of this k-dimensional manifold into the k-dimensional space of control parameters.

In Fig. 6(c) we present the critical function, the value(s) of the function at the critical point(s). This graph is not a manifold because of the sharp corners and self-intersections. The two lower pieces of the graph are the values of the function at the two minima. Where these pieces intersect, the minima are equally deep (Maxwell set, $a<0$, $b=0$, Sec. 3.1). The remaining piece of this graph, which looks like the seat of an Art Moderne chair, is the value of the function at the local maximum. The creases at which the values at the local maximum and minimum join have canonical power-law dependence familiar from the behavior of the Gibbs free energy of a function exhibiting a second-order phase transition (Gilmore, 1981). This canonical power-law dependence is that of the fold catastrophe, namely, 3/2.

The critical curvature, $f''(x_c(a,b);a,b)$, or curvature of the function at its critical point(s), is shown in Fig. 6(d). The curvature is positive at the local minima and negative on the intermediate local maximum. Although there are no creases as in Fig. 6(d), this graph is not a manifold because of the self-intersection. Over the bifurcation set the critical curvature vanishes because the second derivatives vanish and the tangent is "vertical."

1.6.3 Bifurcation Sets for the Three-Dimensional Catastrophes. The geometry of the fold and the cusp was relatively easy to visualize because their graphs could be embedded in low-dimensional spaces: $R^1 \times R^1$ for the fold and $R^1 \times R^2$ for the cusp. Higher-dimensional catastrophes are more difficult to visualize. The catastrophe A_4 should be viewed in $R^1 \times R^3$ while the catastrophes $D_{\pm 4}$ should be viewed in $R^2 \times R^3$. However, the bifurcation sets for these three catastrophes are relatively simple to visualize, since they are embedded in the control-parameter space R^3. These three catastrophes are

$$A_4:\quad f(x;a,b;c)=\tfrac{1}{5}x^5+\tfrac{1}{3}ax^3+\tfrac{1}{2}bx^2+cx,$$

$$\begin{aligned} D_{+4}:\quad & f(x,y;a,b,c)=x^2y+\tfrac{1}{3}y^3+a(y^2-x^2)\\ &\qquad +bx+cy,\\ D_{-4}:\quad & f(x,y;a,b,c)=x^2y-\tfrac{1}{3}y^3+a(y^2+x^2)\\ &\qquad +bx+cy. \end{aligned} \tag{22}$$

In each instance the three-dimensional control-parameter space is partitioned into open regions by two-, one-, and zero-dimensional manifolds, components of the bifurcation set on which two, three, and all four of the critical points are degenerate. Within each open region the critical points are isolated; their number and type are unchanged by a sufficiently small perturbation. The number of critical points can change only when passing from one open region to another through the bifurcation set. The bifurcation sets for these three catastrophes are shown in Fig. 7. Shown in each figure is the number of critical points possessed by the catastrophe function in each of the open regions in its control parameter space. The catastrophes A_4, D_{+4} can each have zero, two, or four nondegenerate critical points while D_{-4} can have only two or four nondegenerate critical points.

1.7 Perturbations of Gradient Dynamical Systems

The qualitative properties of a gradient dynamical system are determined by the number, saddle type, and distribution of its critical points. If the critical points are isolated, then the dynamical system is structurally stable against perturbations. If one or more critical points are degenerate, the system is structurally unstable—a perturbation will produce a qualitative change in the properties of the system by splitting or annihilating the degenerate critical points.

As a result it is sufficient to use perturbation theory to describe the effect of a perturbation on a structurally stable system. In the case of a structurally unstable system it is useful to reduce the degenerate critical point to canonical form (a catastrophe) and then discuss the effect of a perturbation by using the catastrophe germ's universal perturbation.

To illustrate the effect of a perturbation in the structurally stable case, we consider a family of potentials, $V(x;c)$, depending on n state variables and k control parameters. As-

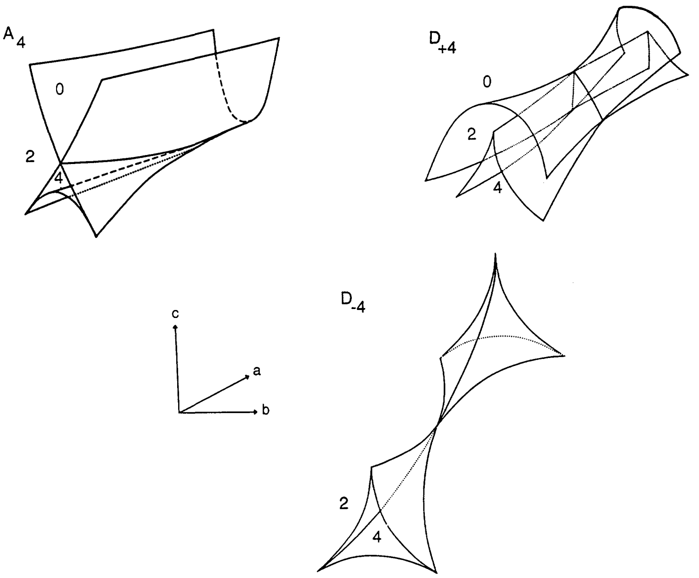

FIG. 7. Bifurcation sets for the three elementary catastrophes [Eq. (22)] of control dimension three as a function of the three control parameters *a,b,c*. The number of isolated critical points is shown in each of the open regions into which the control-parameter space is divided by the bifurcation set.

sume x_0 is a critical point for control-parameter value c_0. What happens under a perturbation $c_0 \to c' = c_0 + \delta c$? We expect that the critical point x_0 will move to a nearby point: $x_0 \to x' = x_0 + \delta x$. The displacement of the critical point as a function of the change in control-parameter value is determined by expanding $V(x_0+\delta x;c_0+\delta c)$ in a Taylor series about x_0, c_0:

$$
\begin{aligned}
V(x_0+\delta x;c_0+\delta c)& \\
&= V(x_0;c_0) + V_i\delta x_i + V_\alpha \delta c_\alpha + \tfrac{1}{2}V_{ij}\delta x_i \delta x_j \\
&\quad + V_{i\alpha}\delta x_i \delta c_\alpha + \tfrac{1}{2}V_{\alpha\beta}\delta c_\alpha \delta c_\beta \\
&\quad + \text{higher-order terms.}
\end{aligned} \tag{23}
$$

The coefficient $V_i(x_0;c_0)=0$, since x_0 is assumed to be a critical point for $c=c_0$. The value of δx is computed by solving $\nabla V(x_0+\delta x;c_0 + \delta c)=0$. To lowest order (linear), we find

$$V_{ij}\delta x_j + V_{i\alpha}\delta c_\alpha = 0. \tag{24}$$

If the stability matrix V_{ij} is nonsingular,

$$\partial x_i / \partial c_\alpha = -(V^{-1})_{ij} V_{j\alpha}, \tag{25}$$

where (V^{-1}) is the matrix inverse of the nonsingular stability matrix: $(V^{-1})_{ij}V_{jk} = \delta_{ik}$. That is, a small change in control-parameter value produces a small change in the location of the critical point, as long as the stability matrix is nonsingular. The matrix (25) is the linear response function for the potential at the equilibrium x_0: it describes how much the equilibrium is displaced by a small change in the control parameters.

To second order the value of the potential at the displaced critical point is

$$
\begin{aligned}
V(x_0+\delta x;c_0+\delta c)& \\
&= V(x_0;c_0) + V_\alpha \delta c_\alpha \\
&\quad + \tfrac{1}{2}[V_{\alpha\beta} - V_{\alpha i}(V^{-1})_{ij}V_{j\beta}]\delta c_\alpha \delta c_\beta.
\end{aligned} \tag{26}
$$

The stability matrix at the displaced critical point, $V_{ij}(x_0+\delta x;c_0+\delta c)$, is related to the stability matrix at the original critical point, $V_{ij}(x_0;c_0)$, by

$$
\begin{aligned}
&V_{ij}(x_0+\delta x;c_0+\delta c) = V_{ij}(x_0;c_0) + P_{ij\alpha}\delta c_\alpha, \\
&P_{ij\alpha} = V_{ij\alpha}(x_0;c_0) - V_{ijk}(V^{-1})_{kl}V_{l\alpha}.
\end{aligned} \tag{27}
$$

As a result, for sufficiently small perturbations the Morse saddle type cannot change if the stability matrix is nonsingular.

This application of perturbation theory, and the analytic results constructed in Eqs. (25)–(27), evaporate when the stability matrix V_{ij} becomes singular. Under these conditions the evolution of the dynamical system under change in the control-parameter values ("perestroika") is computed by expressing the potential, $V(x;c)$, in the neighborhood of a non-Morse critical point by an appropriate catastrophe, following the perturbation through the well-defined bifurcation sets, and then resuming the perturbation treatment sufficiently far from the critical point degeneracy.

2. WHY IT EXISTS

The reduction of a family of functions to the sum of a non-Morse function and a Morse *i*-saddle at a non-Morse critical point, which is guaranteed by the Thom splitting lemma, can be carried out by the methods of elementary calculus. It will therefore not be reviewed here.

The reduction of a non-Morse function in the neighborhood of a degenerate (non-Morse) critical point cannot be carried out by the methods of elementary calculus. It will therefore be discussed in Sec. 2 of this article. This reduction to canonical form can be accomplished following a simple algorithm. The algorithm is illustrated in Sec. 2.1 in the context of a simple example: the reduction of a two-parameter family of functions of a single state variable to canonical form in the neighborhood of its most degenerate critical point. The steps involved in this algorithm are summarized in Sec. 2.2, and applied in Sec. 2.3 to an important example in which the canonical form is not at all evident.

The algorithm involves two main procedures. The first is the determination of the most degenerate critical point, and the catastrophe germ at this point. This is accomplished by finding critical points, using the control parameter degrees of freedom to eliminate the leading terms in the Taylor series expansion of the function in the neighborhood of that critical point, and then using a nonlinear change of coordinates to eliminate the higher-degree terms in the Taylor series expansion. What is left over, between the eliminated terms of low and high degree, is the catastrophe germ, CG(l), describing the degenerate critical point.

The second procedure in this algorithm is the determination of a universal perturbation. This follows the steps described above in more or less reverse order. First, an arbitrary function (perturbation) is added to the catastrophe germ. Then the high-degree terms are eliminated by a nonlinear change of coordinates. Finally, as many as possible of the low-degree terms are eliminated by a rigid translation of the coordinate system. The small number of terms remaining from the original general perturbation form the universal perturbation, Pert(l,k), of the catastrophe germ CG(l).

The mathematics used in these examples appears naively simple. However, this simplicity should not hide the fact that the mathematical approach used here is correct in spirit and can be (has been) made rigorous with sufficient effort (Poston and Stewart, 1978; Arnol'd, 1981; Gilmore, 1981).

2.1 A Simple Example

We carry out the reduction of a family of functions with a triply degenerate critical point to the canonical form of the cusp catastrophe $A_{\pm 3}$. We begin by assuming $f(x;c_1,c_2)$ is a family of functions depending smoothly on one state variable, x, and two control parameters, c_1 and c_2. We wish to determine the qualitative properties of any member of this family and, in particular, to determine how these qualitative properties change as the control parameter values change.

To facilitate this study, we expand $f(x;c_1,c_2)$ about some point x_0:

$$f(x;c_1,c_2)=\sum_{n=0}^{\infty} f_n(x_0;c_1,c_2)(x-x_0)^n. \qquad (28)$$

The Taylor series expansion of this function is shown in line 1 of Fig. 8. The remainder of the discussion of this example will refer to subsequent lines of this figure. At points x_0 where $f_1 \neq 0$ the implicit function theorem is applicable and nothing qualitatively interesting will happen. We therefore search over the range of the state variables to find a critical point. At that critical point $f_1=0$. Since we are interested in the qualitative properties (shape) of the potential, we readjust our y axis so the function vanishes at the critical point. As a result of using these two degrees of freedom—choice of origin in both ordinate and coordinate—the first two terms in the Taylor series expansion of f vanish. The expansion is shown in line 2 of Fig. 8.

There remain the two degrees of freedom represented by the two control-parameter values c_1, c_2. It might be expected that we can use these two degrees of freedom to annihilate one, or at most two, of the remaining coefficients in the Taylor series expansion, but that

Object	Line / Procedure	Coefficients of x_n						
		x^0	x^1	x^2	x^3	x^4	x^5	x^6
Find Canonical Germ	1. Taylor Expansion	f_0	f_1	f_2	f_3	f_4	f_5	f_6
	2. Adjust Ordinate; Locate Critical Point	0	0	f_2	f_3	f_4	f_5	f_6
	3. Exploit Control Parameter Degrees of Freedom			0	0	f_4	f_5	f_6
	4. Smooth Change of Coordinates					±1	0	0
Find Canonical Perturbation	5. Add Arbitrary Perturbation	ε_0	ε_1	ε_2	ε_3	$\pm 1+\varepsilon_4$	ε_5	ε_6
	6. Smooth Change of Coordinates	ε_0	ε_1	ε_2	ε_3	±1	0	0
	7. Shift Origin	0	ε_1	ε_2	0	±1		

$$f(x{:}C_1,C_2) = \underbrace{\pm x^4}_{CG(1)} + \underbrace{\varepsilon_1 x^1 + \varepsilon_2 x^2}_{\mathrm{Pert}(1,2)}$$

FIG. 8. The steps used to reduce a two-parameter family of functions to its canonical form at the most degenerate possible critical point are illustrated in lines 1–4. The steps used to determine the universal perturbation are shown in lines 5–7.

we cannot generically (some deep mathematics is required at this point: cf. Poston and Stewart, 1978; Arnol'd, 1981, 1986; Gilmore, 1981) annihilate more than two coefficients. If coefficients of some high powers (f_5 of x^5, f_7 of x^7) are annihilated, then the coefficient f_2 of x^2 remains nonzero, the Morse lemma is applicable, and nothing qualitatively interesting will happen. The most interesting things happen when the control parameter degrees of freedom are used to annihilate the leading nonzero terms in the Taylor series expansion. In the present case the following two interesting possibilities arise:

(i) $f_2=0,\ f_3\neq 0,$

(ii) $f_2=0,\ f_3=0.$

The second case is more interesting than the first, which can already be encountered in functions depending on one state variable and a single control parameter.

We assume therefore that the two control-parameter degrees of freedom can be used to annihilate the two leading nonzero terms in the Taylor series expansion given on line 2 of Fig. 8. The resulting Taylor series expansion shown on line 3 of Fig. 8 begins with a term of degree 4.

This is as far as we can go using the simple "linear" degrees of freedom represented by choice of origin in the ordinate and coordinate and the two control-parameter degrees of freedom. However, we have not yet exploited any "nonlinear" degrees of freedom. That is, there is the possibility to perform a smooth nonlinear change of coordinates to remove some of the higher-degree terms in the Taylor series expansion. To do this we seek a nonlinear change of variables,

$$x'=A_1x+A_2x^2+\cdots, \tag{29}$$

which will eliminate as many of the higher-degree terms as possible from the remaining Taylor series expansion (line 3, Fig. 8). It is possible to find a nonlinear change of variables that eliminates the Taylor tail, the tail of the Taylor series expansion (the term x^4 is determinate; more powerful mathematics is needed at this point also, cf. Poston and Stewart, 1978; Arnol'd, 1981, 1986; Gilmore, 1981). The first two coefficients in the nonlinear change of variables above are $A_1 = |f_4|^{1/4}$, $A_2 = \pm f_5/4|f_4|^{3/4}$. The expansion is analytic and the sum converges locally. In the new coordinate system the transformed function is $\pm x^4$. The Taylor series after this nonlinear change of variables is shown on line 4 of Fig. 8.

This procedure produces the catastrophe germ of the two-parameter family of functions $f(x;c_1,c_2)$:

$$f(x;c_1,c_2) = \pm x^4 \quad \text{(catastrophe germ)}. \tag{30}$$

The second step in the process of reducing a non-Morse function to canonical form is computation of the universal perturbation of minimum dimension. To do this, we repeat the procedure described above in almost the reverse order. We begin by adding an arbitrary perturbation

$$\epsilon(x) = \sum_{i=0}^{\infty} \epsilon_i x^i \tag{31}$$

to the catastrophe germ. The Taylor series expansion of the perturbed function is presented on line 5 of Fig. 8.

A nonlinear change of variables can once again be called upon to eliminate the Taylor tail. The nonlinear transformation (29) fails to converge unless the base term from which the elimination is made has a large coefficient. Thus, trying to eliminate the Taylor tail of terms above x, x^2, or x^3 will fail because the coefficients of these terms, ϵ_1, ϵ_2, ϵ_3, are small and not bounded away from zero. However, since the coefficient $(\pm 1 + \epsilon_4)$ of x^4 is large and bounded away from zero when ϵ_4 is small, all terms above x^4 may be eliminated by a nonlinear change of coordinates. To state this in another way: Any terms in the Taylor tail that may be eliminated in the initial construction of the catastrophe germ (line 3 to line 4) can be eliminated from the perturbation (line 5 to line 6). The result of this nonlinear change of variables is shown in line 6 of Fig. 8.

As a final step, we can choose a new origin of ordinates (rigid vertical shift) to eliminate ϵ_0 and of coordinates (rigid horizontal shift) to eliminate one of the three remaining terms. Since the linear and quadratic terms cannot always be eliminated but the cubic term can be, we choose to eliminate this term. The final result,

$$f(x;c_1,c_2) = \pm x^4 + \epsilon_1 x^1 + \epsilon_2 x^2, \tag{32}$$

is shown in line 7 of Fig. 8.

We summarize briefly the steps taken to answer the two equations implicit in the statement of the classification theorem presented at the beginning of this section:

1. How to construct the catastrophe germs:
   - (a) Locate a critical point and move the origin of coordinates to that critical point.
   - (b) Eliminate the leading Taylor series coefficients by using the control parameters.
   - (c) Remove the Taylor tail by a smooth nonlinear change of variables.
2. How to construct the universal perturbation:
   - (d) Add an arbitrary perturbation.
   - (e) Remove the Taylor tail by the same smooth nonlinear change of variables.
   - (f) Move the origin of coordinates to eliminate unnecessary terms.

2.2 General Procedure

The general procedure for reducing a non-Morse function to the sum of a catastrophe germ and a universal perturbation has been illustrated by example in the previous section. In some respect, the procedure followed in the reduction follows in spirit the procedure used in the study of electrodynamics. We often find that the near-field regime (low-degree Taylor series terms) is tractable, as is the far-field regime (Taylor tail). The really interesting part is between these two regimes (intermediate-field regime ~catastrophe germ).

We summarize the general procedure below and refer to Fig. 9. The following steps are taken to compute the catastrophe germ of a function $f(x;c)$ with $x \in R^n$ and $c \in R^k$:

1. Computation of the catastrophe germ CG(l):
   - (a) The function is expanded in a Taylor series. The origin of ordinates and coordinates is rigidly (linearly) shifted to a critical point. This eliminates the zeroth- and first-degree terms in the Taylor series expansion.
   - (b) The control parameter degrees of freedom are used to eliminate as many of the remaining low-degree Taylor series coefficients as possible. No more than k coefficients can typically be annihi-

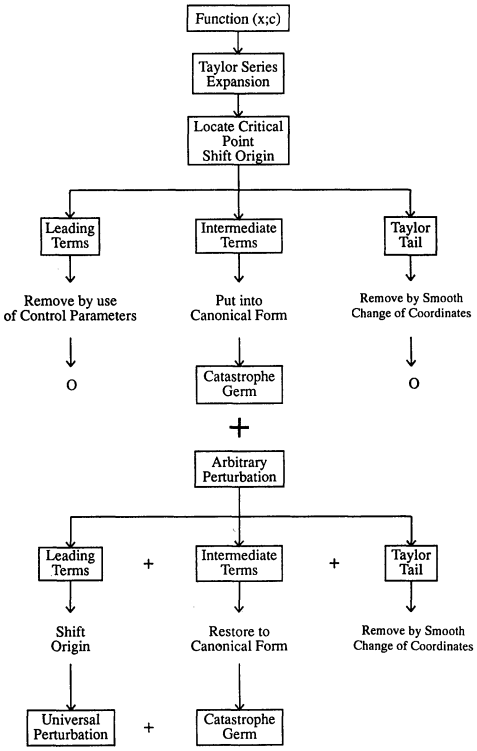
Function (x;c)
Taylor Series Expansion
Locate Critical Point Shift Origin
Leading Terms
Intermediate Terms
Taylor Tail
Remove by use of Control Parameters
Put into Canonical Form
Remove by Smooth Change of Coordinates
0
Catastrophe Germ
0
+
Arbitrary Perturbation
Leading Terms
+
Intermediate Terms
+
Taylor Tail
Shift Origin
Restore to Canonical Form
Remove by Smooth Change of Coordinates
Universal Perturbation
+
Catastrophe Germ

FIG. 9. General procedure for computing the catastrophe germ and the universal perturbation of a family of functions.

lated (by transversality: cf. Poston and Stewart, 1978; Arnol'd, 1981, 1986; Gilmore, 1981).

(c) A smooth nonlinear change of variables is introduced to eliminate as many terms in the Taylor tail as possible. Whether all terms above some finite degree, or only most terms, can be eliminated is determined algorithmically (determinacy algorithm: cf. Poston and Stewart, 1978; Arnol'd, 1981, 1986; Gilmore, 1981).

After the catastrophe germ has been computed, the steps followed above are taken in reverse order to compute the universal perturbation of the catastrophe germ:

2. Computation of the universal perturbation $\mathrm{Pert}(l,k)$:
   (d) An arbitrary perturbation is added to the germ. The smooth transformation used to eliminate the Taylor tail in step (c) above is used again to eliminate the same terms of the perturbed germ and reduce the coefficients in the germ to the same canonical values.
   (e) A shift of ordinate and coordinate is introduced to eliminate the constant term in the perturbation and l (state space dimension of the germ) of the remaining terms of the perturbation.

This procedure reduces a family of functions to canonical form in the neighborhood of a degenerate critical point.

2.3 A More Complicated Example

In this closing section of Sec. 2 we tackle a more complicated example. This is done for two reasons:

1. to illustrate the methodology described in the previous section for a multidimensional state-variable space.
2. to illustrate exactly why the classification of elementary catastrophes terminates at control dimension $k=5$.

The particular example we consider is the reduction to standard form of the six-control-parameter family of functions defined on a three-dimensional state-variable space,

$$f(x;c)=f(x,y,z;c_1,c_2,c_3,c_4,c_5,c_6). \qquad (33)$$

To effect this reduction we follow the steps described in the previous two sections. The computation of the catastrophe germ is illustrated in Fig. 10(a). The computation of the universal perturbation is shown in Fig. 10(b). The steps 1–5 used to compute the germ below are indicated by the corresponding numbers over the arrows, which show the effect of these steps, in Fig. 10(a). Steps 6–9, used to compute the universal perturbation, are similarly keyed to the arrows in Fig. 10(b).

1. The Taylor series expansion for this function is

$$f(x;c)=\sum_{\substack{i\geqslant 0\\ j\geqslant 0\\ k\geqslant 0}} \frac{i!j!k!}{(i+j+k)!} f_{ijk}(x-x_0)^i \times (y-y_0)^j(z-z_0)^k. \qquad (34)$$

   This Taylor series is illustrated schematically in Fig. 10(a), where we have grouped together all terms of the same degree and suppressed factorials.
2. We search over state-variable space for a

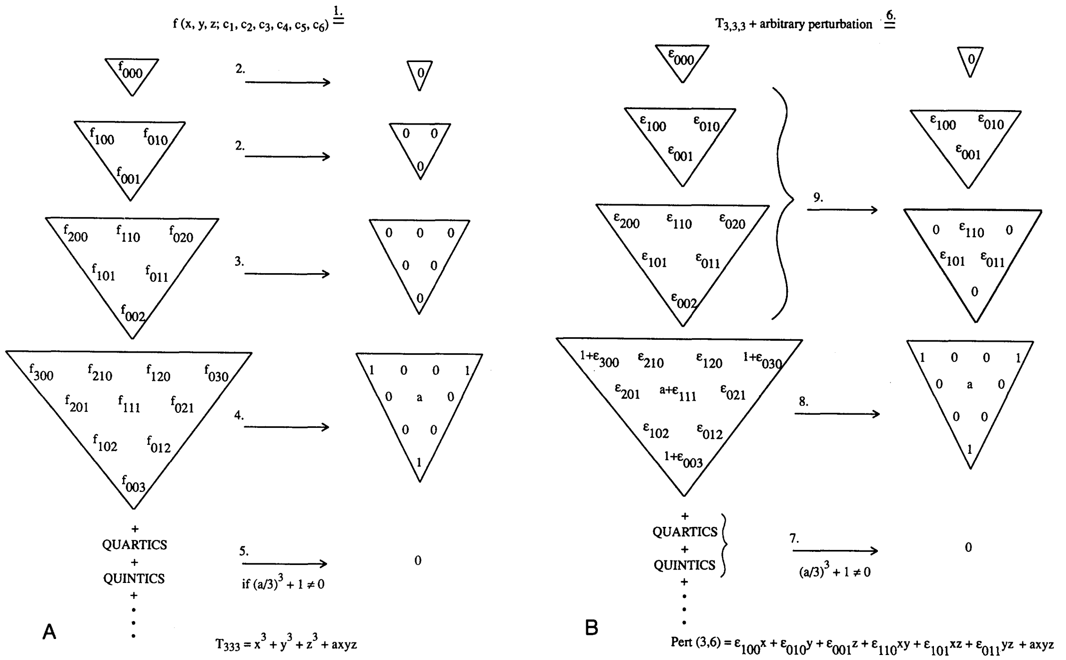
f (x, y, z; c1, c2, c3, c4, c5, c6) 1. =
f000
f100 f010 f001
f200 f110 f020 f101 f011 f002
f300 f210 f120 f030 f201 f111 f021 f102 f012 f003
2.
2.
3.
4.
5.
if (a/3)^3 + 1 ≠ 0
+ QUARTICS + QUINTICS +
0
0 0 0
0 0 0 0 0 0
1 0 0 1 0 a 0 0 0 1
0
A
T333 = x^3 + y^3 + z^3 + axyz
T3,3,3 + arbitrary perturbation 6. =
ε000
ε100 ε010 ε001
ε200 ε110 ε020 ε101 ε011 ε002
1+ε300 ε210 ε120 1+ε030 ε201 a+ε111 ε021 ε102 ε012 1+ε003
9.
8.
7.
(a/3)^3 + 1 ≠ 0
0
ε100 ε010 ε001
0 ε110 0 ε101 ε011 0
1 0 0 1 0 a 0 0 0 1
B
Pert (3,6) = ε100x + ε010y + ε001z + ε110xy + ε101xz + ε011yz + axyz

FIG. 10. The procedure used to compute the catastrophe germ and universal perturbation for a six-parameter family of functions depending on three state variables. The numbers over the arrows are keyed to the steps taken in the text.

critical point, then move the origin of coordinates to that point. This removes the constant and linear terms from the Taylor series expansion.

3. The six control-parameter degrees of freedom can be used to eliminate up to six of the remaining terms in the Taylor series expansion. If any of the quadratic terms remain, the catastrophe ultimately constructed appears in the list given in Table 2. However, if we use these six control-parameter degrees of freedom to eliminate all six second-degree terms from the Taylor series expansion, something new appears. Accordingly, we make this choice. The resulting expansion begins with the third-degree terms.
4. Using a homogeneous linear transformation we can attempt to put the cubic terms into some canonical form. Since a linear transformation on the state-variable space is a 3×3 matrix with 3^2 degrees of freedom and there are ten cubic coefficients, it is possible to give canonical values to only nine of these ten coefficients. Symmetry dictates the choice

$$\sum_{i+j+k\geqslant 3} f_{ijk}x^iy^jz^k = x^3+y^3+z^3+axyz \quad (+\text{ higher-degree terms}). \tag{35}$$

This linear transformation maps nth-degree terms ($n\geqslant 3$) into terms of the same degree. The parameter, a, which appears in this function cannot be given a canonical value, and is the reason why this catastrophe is not elementary.

5. We then search for a smooth nonlinear change of variables which will eliminate as much of the tail of the Taylor series expansion as possible. If

$$(a/3)^3+1\neq 0, \tag{36}$$

the cubic terms are determinate. This means that a smooth transformation exists which transforms away all terms in the Taylor series expansion of degree in excess of some finite value, in this case three. The resulting catastrophe germ is

$$f(x;c) = T_{3,3,3} = x^3+y^3+z^3+axyz. \tag{37}$$

This catastrophe germ is shown at the bottom of Fig. 10(a).

6. To compute the universal perturbation we add a general function of three state variables,

$$\epsilon(x,y,z) = \sum_{\substack{i\geq 0\\ j\geq 0\\ k\geq 0}} \epsilon_{ijk} x^i y^j z^k, \tag{38}$$

   to the catastrophe germ $T_{3,3,3}$. All coefficients ϵ_{ijk} are assumed to be small. The Taylor series of the resulting perturbed function is shown in Fig. 10(b).
7. Working backwards, we find a linear transformation which transforms the cubic terms to the form (37).
8. If the original function $T_{3,3,3}$ is determinate $[(a/3)^3+1\neq 0]$, the perturbed function must also be determinate $[(a'/3)^3+1\neq 0,\ a'=a+\epsilon_{111}]$ for a sufficiently small perturbation. This means that a smooth transformation exists which transforms away all terms in the Taylor series expansion of the perturbed germ of degree greater than 3.
9. Finally, we make a rigid displacement of coordinate system to eliminate the constant term ϵ_{000} in the perturbation and three of the remaining nine linear and quadratic terms. Since it is always possible to eliminate the three terms x^2, y^2, z^2, the universal perturbation is

$$\begin{aligned}\mathrm{Pert}(l=3,k=6) = {} & \epsilon_{100}x+\epsilon_{010}y+\epsilon_{001}z\\ & +\epsilon_{110}xy+\epsilon_{101}xz\\ & +\epsilon_{011}yz+axyz. \end{aligned}\tag{39}$$

   The seventh term, $axyz$, has been added because it is not possible to reduce the coefficient of xyz in the catastrophe germ $T_{3,3,3}$ to a standard value. The universal perturbation is shown at the bottom of Fig. 10(b).

The net result of this calculation is a reduction of the function $f(x;c)$ given in Eq. (33) to canonical form:

$$\begin{aligned} f(x;c) &= T_{3,3,3} + \mathrm{Pert}(3,6)\\ (33) &= (37) + (39) \end{aligned}\tag{40}$$

We emphasize once more that for $k<6$ control parameters it is possible to assure that any catastrophes that are encountered are elementary, but for $k=6$ (and $k>6$) this is no longer possible, as $T_{3,3,3}$ shows.

3. HOW IT WORKS

Elementary catastrophe theory consists of a collection of theorems about the canonical forms of functions in the neighborhood of degenerate critical points. Catastrophes in the real world consist of processes that exhibit discontinuities and sudden jumps. It is clear that the mathematics provides the right language for describing the physical processes. But how? The mathematics describes potentials that depend on control parameters; processes depend on time. It is tempting to allow control parameters to depend on time, but then the catastrophe functions are no longer strictly potentials. And what of the dynamics describing a physical system? If many equilibria are available, which does the system choose? Finally, how can we recognize when a physical process can be modeled by a mathematical catastrophe function?

These practical questions are the subject of Sec. 3 of this article. The difficulty of relating the dynamics of a system to a potential that describes it is discussed in Sec. 3.1. In most instances the system state is governed either by a local or the global minimum of a potential, and sudden jumps from one state to another occur when the control parameters pass through an appropriate component of the bifurcation set. Although the two conventions presented (Maxwell, Delay) are not complete, they serve well in a preponderance of situations.

When a catastrophe occurs in some physical process, a multiplicity of related phenomena occur. These phenomena are called catastrophe flags. They are treated in Sec. 3.2. The identification of any is a clear indication that the others are present and will be recognized if sought. These flags may be used to identify the mathematical catastrophe function that most accurately describes the discontinuous physical process.

In Sec. 3.3 we illustrate the use of catastrophe theory in determining the sensitivity of failure modes of a complex structure to hidden or unforeseen imperfections. This is done by investigating the catastrophe germ for an optimized system and then determining the universal perturbation for that germ. In the particular example considered, that of a propped cantilever, strong coupling of two "soft" collapse modes yields a "soft" collapse direction and a "hard" direction, one in which a severe and unexpected collapse is possible. Further, analysis of the universal perturbation for that collapse mode reveals that the mode is extremely sensitive to imperfections in either the design or fabrication stages of construction. Finally, the use of catastrophe theory to suggest tests to locate the actual (as opposed to designed) failure load, as well as ways to increase the load at which failure occurs, are suggested.

Since elementary catastrophe theory is the first step in the program of catastrophe theory, all its results and phenomenology will necessarily be present in subsequent developments in this program. The next stage in this program following elementary catastrophe theory is the study of dynamical systems, sets of coupled first-order ordinary differential equations. We show in Sec. 3.4 how the fold and the cusp catastrophe appear in dynamical systems theory under the guise of three elementary bifurcations: the saddle-node bifurcation (A_2), the pitchfork bifurcation (A_3), and the Hopf bifurcation (A_3).

3.1 Catastrophe Conventions

Elementary catastrophe theory is the first tentative step in the program of catastrophe theory. It was motivated above by our desire to describe the qualitative properties of gradient dynamical systems of the form (4). The qualitative dynamics of such systems are determined by a potential function. The local flow in the neighborhood of each nondegenerate critical point (Morse i-saddle) is canonical and the separate flows can be pasted together to construct a global flow.

As the study of elementary catastrophe theory evolved, it became clear that the key to our understanding of non-Morse critical points was the determination of how the critical points moved about, coalesced and disappeared, or were born degenerate and moved apart as the control-parameters changed. The dependence of critical points on control-parameter values, as described by the canonical decomposition (15), provides a complete resolution to the mathematical question posed by elementary catastrophe theory. However, it does not address a corresponding physical question. That is, how does the dynamics depend on the control parameters? Once the control parameters are allowed to vary in time, the system (4) is no longer a gradient system—the potential is time dependent. As a result, knowledge of the shape of the potential is not sufficient to determine the state of the system unless the equations of motion are known.

In the absence of a detailed set of equations of motion, the state of the physical system described by an elementary catastrophe must be determined by some infusion of intuition. This is done by adopting one of two standard assumptions. These are the Delay convention and the Maxwell convention (Zeeman, 1977; Poston and Stewart, 1978; Gilmore, 1981).

Delay Convention: The state of a system is determined by a (local) minimum of a potential. As the control parameters change, the state remains at the local minimum as long as that minimum exists. When the local minimum ceases to exist (at a bifurcation set) the system state jumps to a new local minimum.

Maxwell Convention: The state of a system is determined by the global minimum of a potential. As the control parameters change, the state remains at the minimum as long as that minimum remains the global minimum of the potential. When the minimum ceases to be the global minimum, the system state jumps to the new global minimum.

We stress again that these assumptions about the behavior of a physical system are required only by our lack of knowledge of the system's equations of motion.

The difference between these two assumptions is illustrated in Fig. 11 for the cusp catastrophe. When $a<0$ the potential may have one minimum or two. We assume the system begins in the unique minimum of the potential. As the control parameter b is changed, a new minimum is created in a fold catastrophe. The original minimum eventually becomes metastable with respect to the new minimum, and finally the original minimum disappears in a fold catastrophe. According to the Delay convention, the system jumps to the new minimum only when the original minimum is annihilated. According to the Maxwell convention, the jump occurs when the two minima become equally deep (at $b=0$). These two conventions are illustrated on the canonical critical-value surface for the cusp in Fig. 11(b). The jumps occur at the parts of the critical-value graph

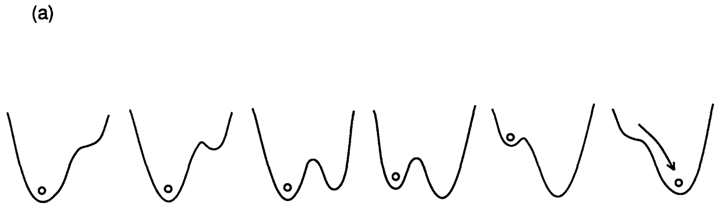

Delay Convention

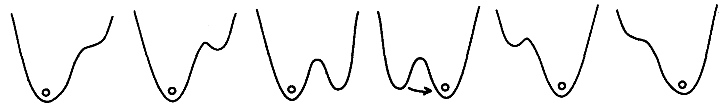

Maxwell Convention

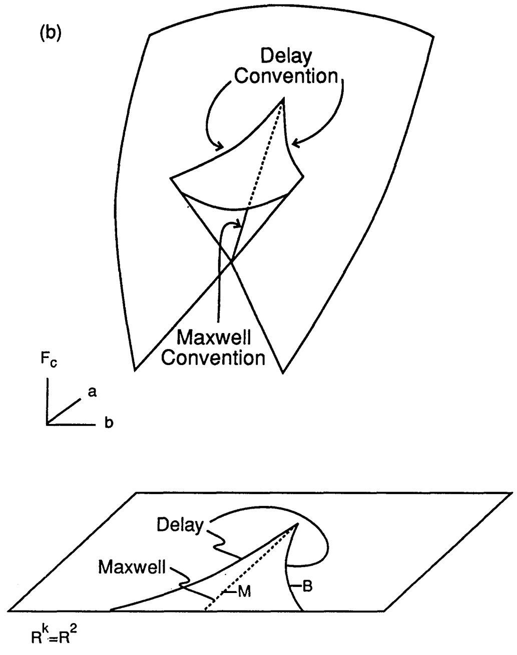

FIG. 11. (a) The Delay convention and the Maxwell convention determine when the jump from one local minimum to another takes place as the control parameters are varied. (b) The two bifurcation sets correspond to pieces of the critical surface which prevent the surface from being a manifold. Under the Delay convention, jumps occur at the creases in the surface, where a local minimum and maximum become degenerate. Under the Maxwell convention, jumps occur at the intersection of two of the leaves of this surface, where the two minima become equally deep.

$$f = f(x_c(a,b);a,b),$$

which prevent the surface from being a manifold. The Delay convention requires jumps to occur at the creases which represent the collision of a local minimum and maximum and which lie over the cusp-shaped curve (21). The Maxwell convention requires jumps to occur at the self-intersection which represents the surface on which the two minima are equally deep and which lies over the half-line $b=0$, $a<0$.

The bifurcation sets for the Delay and Maxwell conventions are sets in the control-parameter space at which jumps to new states take place. These sets are determined by local and global considerations for the two conventions. The bifurcation set for the Delay convention consists of that subset of the bifurcation set for the catastrophe which involves a Morse 0-saddle. For catastrophes more complicated than $A_{\pm 3}$, some components of the bifurcation set involve Morse saddles M_i^n and M_{i+1}^n, $i > 0$. On these components of the bifurcation set the two saddles M_i^n and M_{i+1}^n are degenerate. These components of the catastrophe's bifurcation set are not components of the bifurcation set determined by the Delay convention, since local minima (M_0^n) must be present. The bifurcation set for the Maxwell convention is determined by nonlocal arguments. Once a point on the Maxwell set has been located, the remainder of the set may be constructed by integrating a set of equations of Clausius–Clapeyron type (Gilmore, 1981).

The appropriate choice of convention must balance the rate at which the control-parameters are varied against the noise level of the system—specifically, the probability that a fluctuation in energy will occur in excess of the barrier height separating a local minimum from other minima, including the global minimum. When the noise level is low, it is usually safe to adopt the Delay convention. When large fluctuations occur more rapidly than the control parameters are swept, the Maxwell convention is more suitable.

From Fig. 11 it can be observed that the bifurcation set for the Maxwell convention "lies inside" the bifurcation set for the Delay convention. This is a general feature of catastrophes and has the following physical interpretation. If a system whose phase transition properties are usually described by the Maxwell convention is handled very gently, the jump from a metastable state to a global minimum may not occur at the bifurcation set for the Maxwell convention. The transition may be delayed—but it cannot be delayed beyond the point at which the local minimum ceases to exist—the bifurcation set for the Delay convention. The outer limits beyond which metastability cannot occur for a physical system are called spinodal curves (or surfaces). The spinodal curves surrounding the bifurcation set for the Maxwell convention consist of the bifurcation set for the Delay convention. The spinodal curves for the symmetry-restricted catastrophe

$$A_{+5}\colon f(x;a,c)=\tfrac{1}{6}x^6+\tfrac{1}{4}ax^4+\tfrac{1}{2}cx^2 \tag{41}$$

are shown in Fig. 12.

When the equations of motion for the system are known explicitly, there is no need to make an assumption (i.e., adopt a convention) about the system's behavior. For example, the system state may be governed by a probability distribution, $P(x,t)$, which satisfies a Fokker–Planck equation

$$\partial P/\partial t=\nabla(P\nabla V)+\nabla^2(DP), \tag{42}$$

where V has a time dependence due to the time dependence of some control parameters:

$$V(x;t)=V(x;c(t)) \tag{43}$$

and D is a diffusion constant. Then the probability distribution for the system state can be determined explicitly as a function of time. In this case there is no well-defined bifurcation set—there is an occupation probability for the two minima (and surrounding regions) and that probability changes smoothly in time.

It has been shown that the Fokker–Planck equation can be used to interpolate between the two extreme limits represented by the Delay convention and the Maxwell convention (Gilmore, 1981).

We remark, finally, that the problem of system state is not necessarily resolved by adopting the Delay convention, as appears to be the case in Fig. 11. In that figure there is one remaining minimum when the metastable minimum disappears in a fold catastrophe, so the system has a unique minimum to jump to. But what happens when there are two or more remaining minima that the system can jump to? Does the system jump to the deepest minimum, the nearest minimum, or yet another minimum? On this the Delay

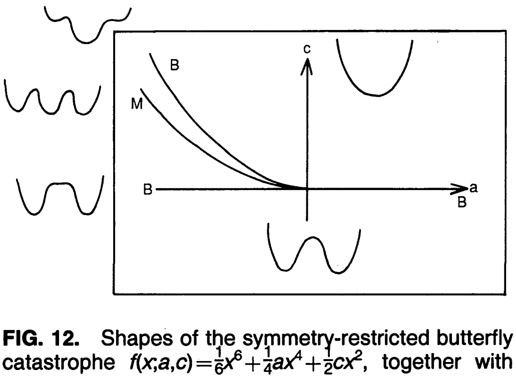

FIG. 12. Shapes of the symmetry-restricted butterfly catastrophe $f(x;a,c)=\frac{1}{6}x^6+\frac{1}{4}ax^4+\frac{1}{2}cx^2$, together with bifurcation sets under the Delay and Maxwell conventions. Spinodal curve for the Maxwell set is the bifurcation set for the Delay convention.

convention has nothing to say. In short, the conventions are incomplete. They can be very useful but they are not substitutes for detailed knowledge of the equations of motion.

3.2 Catastrophe Flags

An important reason for studying the elementary catastrophes is that their properties are canonical. Therefore, predictions based on the properties of an elementary catastrophe in a physical system are canonical. This means that their properties need be studied only once. They can then be applied directly, without change, to any physical system described by that catastrophe. As a result, it is important to be able to determine the presence and type of a catastrophe underlying the behavior of a physical system.

This may be done either deductively or inductively. In the first procedure there is full knowledge of the equations of motion describing the system. When this is the case it is generally possible to make whatever predictions are desired by computing solutions of these equations. It is then possible to predict the behavior of the system without any knowledge of the elementary catastrophes. This, in fact, has been a long-standing operational mode in the physical sciences.

There are, however, instances in which the equations of motion are not known. Then the presence of a catastrophe cannot be deduced (from the equations). It may, however, be induced from the behavior of the system. When a catastrophe is present, the system undergoes (or may undergo) a qualitative change in its behavior. This qualitative change is often accompanied by a number of other phenomena. For any particular system these may be more or less difficult to recognize. The point is that the observation of any will imply the presence of all and that if we look hard enough we will find all these correlated phenomena.

Eight phenomena typically occur when a catastrophe is present. These are called catastrophe flags: in a sense the catastrophe "waves flags" to attract our attention. Of the eight, five are classical: they occur when there is a qualitative change in the system. The remaining three also occur when there is a qualitative change, but they may also be observed before a phase change occurs. This is very important when the phase change represents the transition of a structure from one state to another (i.e., collapse of a bridge). The five classical catastrophe flags are modality, sudden jumps, inaccessibility, sensitivity, and hysteresis (Zeeman, 1977; Poston and Stewart, 1978). The three diagnostic catastrophe flags are divergence of linear response, time dilation, and anomalous variance (Gilmore, 1981). These are now discussed and illustrated in Fig. 13 for the cusp catastrophe.

3.2.1 Modality. In the neighborhood of a catastrophe the system can exhibit two or

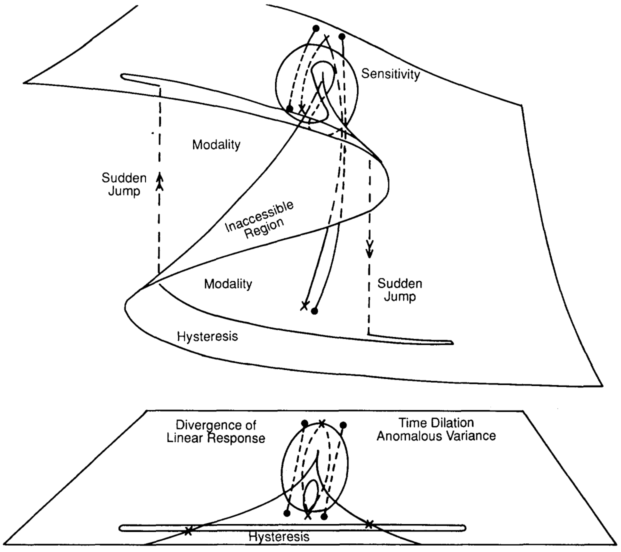

FIG. 13. Eight phenomena (flags) associated with the presence of a catastrophe.

more distinct types of behavior (near a critical point: liquid, gas; near a triple point: solid, liquid, gas). The upper and lower surfaces of the cusp catastrophe manifold represent two distinct types of locally stable behavior in Fig. 13. The intermediate part of this manifold represents an unstable mode of behavior which nevertheless may be observed with sufficient care.

3.2.2 Sudden Jumps. The system may suddenly jump from one mode of behavior to another mode of behavior (liquid, high density, to gas, low density) as the control parameters (intensive thermodynamic variables: temperature and pressure) are varied. These jumps represent the transition from one local minimum to a global or another local minimum. The location of the sudden jumps may or may not be adequately described by one of the two conventions presented in Sec. 3.1.

3.2.3 Inaccessibility. Separating the two or more local minima (M_0^n) responsible for "modalities" is at least one Morse saddle (M_1^n). The Morse saddles M_i^n ($i > 0$) are all dynamically unstable. In the neighborhood of such a saddle the system, if perturbed, will move quickly to a locally stable state. As a result, modality will always be accompanied by inaccessible regimes as represented by the middle sheet of the cusp catastrophe manifold.

3.2.4 Sensitivity. If the system is brought from the single-mode regime to the multimode regime the question arises: what is the final state of the system? For most initial conditions and processes, the final state is robust against perturbations. That is, if the final state lies on the upper sheet of the cusp catastrophe manifold, then a perturbation of either initial conditions and/or processes will leave the final state on the upper sheet. However, there is a small set of initial conditions and/or processes for which this is not true. This sensitivity to (some) initial conditions is shown in Fig. 13. For the process $db/dt=0$, $da/dt<0$ the initial condition with $b=\epsilon>0$ will result in a final state on the lower sheet while the nearby initial condition $b=-\epsilon<0$ lies on the upper sheet. A similar phenomenon occurs for processes that are almost identical, beginning from appropriate initial conditions. We remark that for the elementary catastrophes there is sensitive dependence on *some* initial conditions while in the study of chaos there is sensitive dependence on *all* initial conditions.

3.2.5 Hysteresis. A sudden jump from one modality to another may occur at the bifurcation set for the Delay or Maxwell convention, or anywhere in between. If the process is reversed, the jump back to the original modality may occur at the same control-parameter value as the initial jump (Maxwell convention) or at different control parameter values (any other convention). Unless the Maxwell convention is observed, jumps between the different modalities will exhibit hysteresis. Hysteresis, while not necessarily a part of the catastrophe scenario, can usually be observed with sufficient care even when the Maxwell convention suggests itself (e.g., superheating and supercooling in a liquid–gas phase transition).

The five classical catastrophe flags just described are of little use to the structural engineer looking for the limits of stability of a large structure (bridge, building) or a naval engineer testing a ship's stability. In many instances, transition from one modality to another is not only a catastrophe, but a disaster as well (collapsed bridge, building; capsized ship). For this practical reason it is essential to have a set of catastrophe flags that identify the presence of a catastrophe, and can be used to map out approximately the bounds of the catastrophe (i.e., bifurcation sets) without actually having to enter the multimode regime, or at least make the transition between modes. The three diagnostic catastrophe flags accomplish just this.

3.2.6 Divergence of Linear Response. The response, δx_i, of an equilibrium in state-variable space to a small change in the control parameters, δc_α, is given by

$$\delta x_i=\chi_{i\alpha}\delta c_\alpha, \tag{44}$$

where the linear susceptibility tensor $\chi_{i\alpha}$ is given explicitly by Eq. (25) for a system governed by a potential. As a local bifurcation set is approached, at least one of the eigendirections has a very large response. This is the direction in which two or more critical points are approaching each other. This direction is an eigenvector of the stability matrix, V_{ij}, in Eq. (25). The eigendirection, and the dependence of the diverging eigenvalue (λ_i^{-1}) on the control parameters c_α, can be used to map out the bifurcation set without actually penetrating it (Gilmore, 1981). In the liquid–gas phase transition divergence of linear response manifests itself as the divergence of the compressibility as the critical point is approached from either the single-phase or the multiphase region. Divergence of the compressibility may also be seen, with greater difficulty, as the spinodal lines are approached from within the multiphase region.

3.2.7 Time Dilation (Critical Slowing Down, Mode Softening). In the neighborhood of a nondegenerate critical point the equations of motion for a dissipative gradient system are

$$dx_i/dt = -V_{ij}x_j. \quad (45)$$

For a dissipative system the relaxation to equilibrium following a perturbation occurs on a time scale that is the reciprocal of the smallest eigenvalue of the stability matrix V_{ij}. For a conservative system the spectrum of oscillation frequencies, ω_i, consists of the square roots of the eigenvalues of V_{ij}. Upon approach to a bifurcation set one or more of the eigenvalues of V_{ij} approaches zero. Relaxation to equilibrium slows down (critical slowing down) as the bifurcation set is approached. The dependence of the vanishing eigenvalue on the control parameters can be used to determine where in control-parameter space the bifurcation set occurs. The associated eigenvector can be used to locate where in state space the critical points become degenerate. A similar analysis is possible when one (or more) of the frequencies of a conservative system approaches zero (mode softening). Time dilation is well known to experimentalists. The time required for equilibration near the critical point in the liquid–gas phase transition grows very rapidly as the critical point is approached, making experiments on the equilibrium properties of fluids near their critical points very difficult and time consuming.

3.2.8 Anomalous Variance. When the state of a physical system is described by a Morse critical point, motion about the equilibrium is confined by a quadratic potential well. On approach of another critical point the quadratic potential becomes flattened and motion about the equilibrium becomes less confined. This reduction in localization, or deconfinement, is called anomalous variance. In a gas–liquid system anomalous variance is dramatically observed at the critical point as critical opalescence. This occurs when fluctuations in the size of liquid droplets in the gas phase and gas bubbles in the liquid phase can occur at all length scales, including length scales comparable to optical radiation, producing an anomalously strong scattering of light.

Although catastrophe theory provides a qualitatively correct description of physical properties of a fluid near its critical point, it does not provide a quantitatively correct description of these properties and must be replaced by a better descriptive mechanism for the purposes of making accurate predictions of physical behavior.

3.3 The Dangers of Design Optimization

Large structures are built from many smaller parts. Cost is almost always a major factor in the design and construction of a complex structure, such as a bridge. As a result, a design philosophy has been widely adopted. This is the philosophy of design optimization. The basic idea is to design each component to meet but not exceed its specifications. For example, if an elevated roadway designed to support a load L is to be supported by ten cantilevers, each cantilever should be designed to support some fraction of that load. It makes no sense to design some cantilevers to be stronger (and more expensive) than others: what good will they do when the others have already collapsed, taking the roadway with them?

This is the philosophy of design optimization. It has hidden dangers because complex structures may have hidden or unexpected collapse modes. Further, these modes may have exceptionally sensitive dependence to imperfections in the fabrication/construction stages. We illustrate these ideas and dangers by treating a simple example. This is a propped cantilever designed to support a load F up to some critical load, F_c.

The cantilever is shown in Fig. 14. We assume it has unit length and a force F is applied vertically at the top of the cantilever. The mass of the beam is assumed small compared to the loading force F.

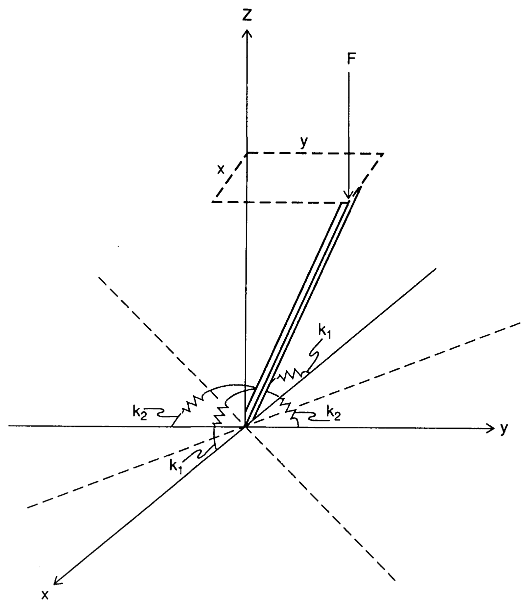

FIG. 14. The cantilever consists of a vertically propped beam supported by springs of spring constant k_1 providing restoring forces in the x–z plane and springs of spring constant k_2 providing restoring forces in the y–z plane. The length of this beam is scaled to 1. Displacements from the vertical are described by the (x,y) coordinates of the top of the beam. Solid axes (x,y) are axes along which a "soft" A_{+3} catastrophe occurs. Dashed axes ($x=\pm y$) are axes along which a "hard" A_{-3} catastrophe occurs.

We assume the collapse modes of this cantilever involve deflection in the x-z plane through an angle θ_1 from the vertical, deflection in the y-z plane through an angle θ_2 from the vertical, or any combination of these motions. We also assume that deflection is counteracted by springs with spring constants k_1, k_2 providing restoring forces in the x-z and y-z planes, respectively. The potential describing this cantilever is

$$V(\theta_1,\theta_2;F)=(2/2)k_1\theta_1^2+(2/2)k_2\theta_2^2+Fz, \quad (46)$$

where z is the height of the top of the cantilever.

It is useful to express this potential as a function of the displacement coordinates $x=\sin\theta_1$ and $y=\sin\theta_2$ of the top of the cantilever. The potential, up to quartic terms, is

$$V(x,y;F)=k_1x^2(1+x^2/3)+k_2y^2(1+y^2/3)$$
$$+F(1-x^2-y^2)^{1/2}. \quad (47)$$

For displacements in the x-z plane with $y=0$,

$$V(x,0;F)=F+(k_1-F/2)x^2+(k_1/3-F/8)x^4. \quad (48)$$

This can immediately be recognized as a symmetry-restricted cusp catastrophe, A_{+3}. A similar potential describes displacements in the y-z plane $x=0$. Since the potential $V(x,y;F)$ exhibits a bifurcation at $k_1-F/2=0$ and $V(0;y;F)$ exhibits one at $k_2-F/2=0$, it makes sense to design the cantilever with equal spring constants $k_1=k_2=F_c/2$, where F_c is the critical load to be supported by the cantilever.

Or does it?

The potential for the optimized cantilever under an arbitrary displacement is

$$V(x,y;F)=F+\tfrac{1}{2}(F_c-F)(x^2+y^2)$$
$$+(F_c/6-F/8)(x^4+y^4)-\tfrac{1}{4}Fx^2y^2. \quad (49)$$

The last term represents a strong coupling between the two displacement directions. This coupling term has severe consequences for the collapse behavior of the cantilever.

As the force F is increased, the potential $V(x,y;F)$ remains locally stable until the critical load F_c is reached. At this load, the fourth-degree term $F_c(x^4-6x^2y^2+y^4)/24$ allows sudden collapse of the cantilever in particular directions. This is not what is expected from investigating the planar potentials $V(x,0;F)$ and $V(0,y;F)$: for these potentials the coefficient of the quartic term is positive at the critical load, so that only a small in-plane displacement occurs after the critical load is exceeded. In other words, the postbuckling behavior of the cantilever in either the x-z plane or the y-z plane is "soft" in the sense that after the bifurcation the stable equilibrium is located near the original stable equilibrium $(x,y)=(0,0)$ in the state-variable space.

In cylindrical coordinates $x=r\cos\phi$ and $y=r\sin\phi$ the potential is

$$V(r,\phi;F)=F+(k-F/2)r^2$$
$$+(k/3)r^4(\cos^4\phi+\sin^4\phi)-(F/8)r^4. \quad (50)$$

The coefficient of r^4 is largest $(=+1)$ for displacements along the $\pm x$ or $\pm y$ axes ($\phi=n\pi/2$, n integer) and smallest $(=1/2)$ for displacements along the axes $x=\pm y$ ($\phi=\pi/4$

$+n\pi/2$, n integer). Along these two different sets of displacement directions the potentials are

$$V(r,\phi=0;F)=F+(k-F/2)r^2+(k/3-F/8)r^4, \tag{51a}$$

$$V(r,\phi=\pi/4;F)=F+(k-F/2)r^2+(k/6-F/8)r^4. \tag{51b}$$

The first potential, representing displacements in the x-z or y-z planes, is a catastrophe of type A_{+3}, with "soft" postbuckling behavior. That is, for loads in excess of the critical load the displacement from vertical increases rapidly but continuously. The second potential, representing coupled displacements ($x=\pm y$), is a catastrophe of type A_{-3}, with "hard" postbuckling behavior. For loads in excess of the critical load the only stable position is horizontal—completely collapsed.

These observations may appear academic. After all, the cantilever remained vertical and supported loads F below the critical load. However, we have not yet considered imperfection sensitivities. These are encountered in the fabrication stages (of the springs, for example) and the construction stages in every large project. The most general perturbation, representing arbitrary imperfections for the cantilever, are obtained by investigating the germ of the potential, which is $(x^4-6x^2y^2+y^4)/24$. The universal perturbation of this catastrophe germ is the eight-parameter function

$$\mathrm{Pert}(x,y;a_{ij})=\sum a_{ij}x^iy^j, \quad i\leqslant 2,\ j\leqslant 2,\ (ij)\neq(00). \tag{52}$$

It is truly remarkable that every imaginable imperfection in fabrication of the individual components (nonuniformity in density of material from which the beam is constructed, unwanted bends in the beam, air bubbles, etc.) as well as every imaginable imperfection in construction (springs along the x axis with different spring constants, springs along x and y axes slightly out of perpendicular, rest height not quite vertical, loading force not quite vertical, etc.) can be represented by a universal perturbation with only eight parameters.

However, even eight parameters will present a complicated problem of analysis. In the present case we can do even better. The failure eigendirections for the soft A_{+3} catastrophe are the x and y axes $(\pm x,0)$, $(0,\pm y)$. Under this catastrophe [Eq. (51a)] the original unperturbed stable solution $(x,y)=(0,0)$ becomes unstable for $F>F_c$, giving rise to new stable equilibria $(\pm[6(F/F_c-1)]^{1/2},0)$, $(0,\pm[6(F/F_c-1)]^{1/2})$. Since the new equilibria are relatively near the old one and are stable, this structural failure is a "soft" bifurcation. This "soft" bifurcation is shown in Fig. 15(a).

The other failure eigendirection, $x=\pm y$, is a "hard" bifurcation and a completely different story. In this case [Eq. (51b)] the new equilibria exist for $F<F_c$ at $(\pm[6(1-F/F_c)]^{1/2}, \pm[6(1-F/F_c)]^{1/2})$. These equilibria are saddles of type M_1^2. If for $F<F_c$ a perturbation moves the system state (x,y) to the neighborhood of any of these saddles, there is the possibility of sudden collapse of the cantilever. This "hard" bifurcation is shown in Fig. 15(b).

As a result of these considerations, in order to study the severe failure modes of the propped cantilever, and in particular the imperfection sensitivity of these modes, it is sufficient to study the imperfection sensitivity of the A_{-3} catastrophe [Eq. (51b)]. Under the most general perturbation, the potential describing failure modes in the weak eigendirection is

$$V(r;b,F)=-\tfrac{1}{24}F_cr^4+\tfrac{1}{2}(F_c'-F)r^2+br, \tag{53}$$

where $F_c' = F_c + a$, and a is one of the two control parameters for the cusp catastrophe. The other parameter, b, describes symmetry-breaking perturbations. The load $F_f = F(F_c',b)$ at which the propped cantilever collapses is determined by the values of F_c', b at which the locally stable equilibrium becomes degenerate with either of the two saddles it separates. This locus is the semicubical parabola (21b), which gives for the failure load

$$F_f=F_c'-\tfrac{1}{2}(3bF_c^{1/2})^{2/3}. \tag{54}$$

This simple expression reveals that the failure load is severely decreased by any symmetry-breaking imperfection ($b\neq0$). The severity is indicated by the power-law dependence: 2/3. An imperfection leading to a small asymmetry in either plane $x=\pm y$ will lead to an unexpectedly large reduction in the cantilever's failure load. The imperfection sensitivity

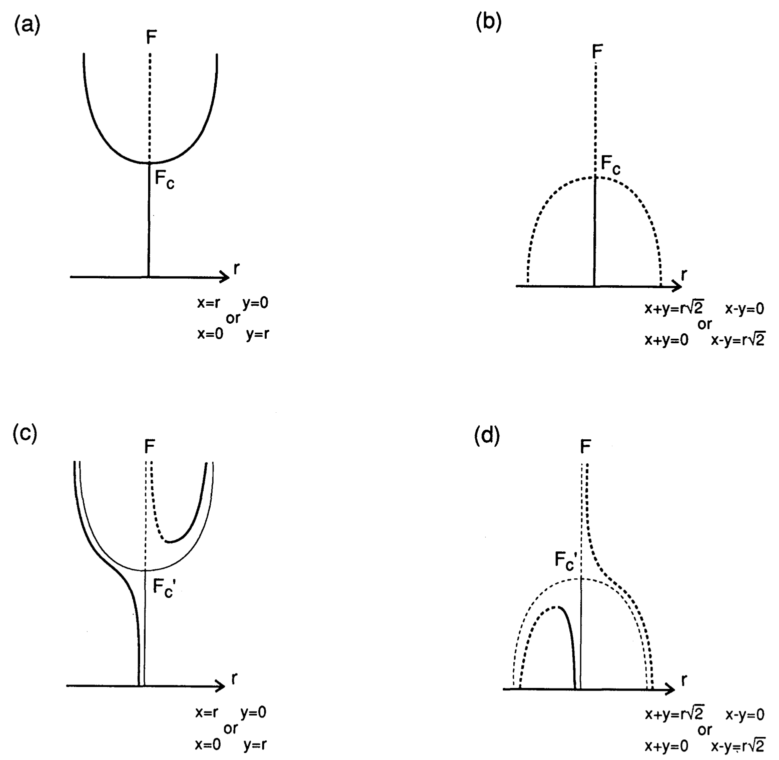

FIG. 15. In all figures stable equilibria (Morse saddles M_0^2) are shown by solid curves and unstable equilibria (M_1^2) are shown by dashed curves. (a) Equilibrium displacement for the "soft" A_{+3} catastrophe along the axes $(x,y=0)$ or $(x=0,y)$. Prebuckling path $(F<F_c)$ is $(x,y)=(0,0)$. Beyond the critical load, $F>F_c$, the postbuckling path is parabolic, with $r\sim(F-F_c)^{1/2}$. (b) Equilibrium displacement for the "hard" A_{-3} catastrophe along the directions $x=\pm y$. Prebuckling path $(x,y)=(0,0)$ is stable for $F<F_c$. There are no stable equilibria for $F>F_c$. (c) The most general perturbation of a cusp catastrophe consists of a linear plus a quadratic term: $(a/2)r^2+br$. The quadratic term $(a/2)r^2$ shifts the critical load: $\frac{1}{2}(F-F_c)r^2+(a/2)r^2+br\rightarrow\frac{1}{2}(F-F_c')r^2+br$, where $F_c'=F_c-a$. The linear term breaks the symmetry. For the "soft" catastrophe A_{+3} the perturbation displaces the equilibrium as shown. (d) For the "hard" catastrophe A_{-3} the perturbation can dramatically reduce the load at which failure occurs, as shown.

of the "soft" and "hard" bifurcations are shown in Figs. 15(c) and 15(d).

Having analyzed the problem, can we do anything about it? The first step is to determine how severe the problem is. To determine the failure load, we make use of the catastrophe flags. The first five (modality, sudden jumps, inaccessibility, sensitivity, hysteresis) are observable when the cantilever collapses. This is precisely the problem we would like to avoid. The remaining three catastrophe flags (divergence of linear response, critical slowing down, anomalous variance) are diagnostics designed precisely for the purposes at hand. That is, the system is subjected to gradually increasing loading. The following data are then recorded:

1. the displacement, r, from equilibrium;
2. the normal-mode frequency, ω, for small oscillations;
3. the amplitude, A, for small oscillations about equilibrium.

These three observables depend in a canonical way on the control parameters $F-F_c'$ and b. Testing is carried out gently and terminated before the collapse of the cantilever. By extrapolating the equilibrium displacement $r(F)$, the normal-mode frequency $\omega(F)$, and the amplitude of small oscillations $A(F)$, the failure load F_f and the control-parameter values F_c' and b at the failure load can be determined (Gilmore, 1981). Such extrapolations are shown in Fig. 16.

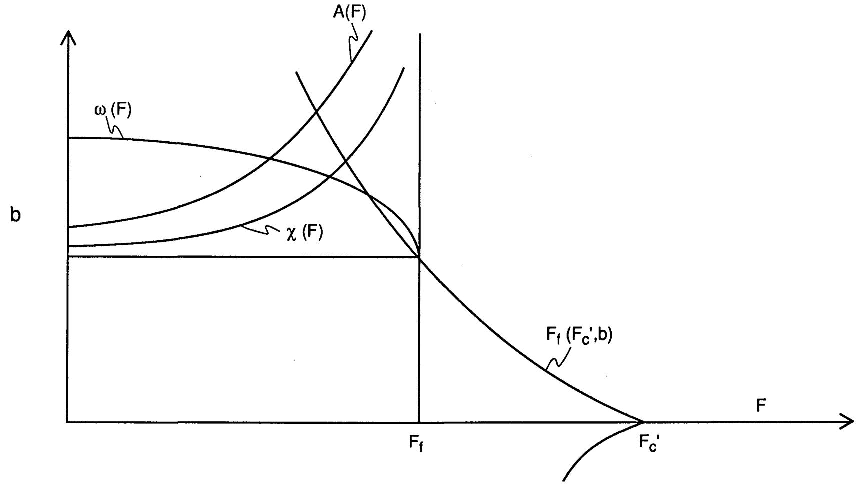

FIG. 16. The failure load $F_f(F_c,b)$ has a 2/3 power-law dependence on the symmetry-breaking parameter b. The value of the load at which collapse occurs (vertical line at F_f) can be located by exploiting the catastrophe flags. This is done by plotting the frequency of small oscillations, $\omega(F)$ (critical slowing down), by determining the linear susceptibility, $\chi(F)$ (divergence of linear response), and by measuring the amplitude of small equal-energy oscillations, $A(F)$ (anomalous variance), all as functions of increasing load F. The amplitude diverges and causes collapse before the failure load F_f is reached.

Once the symmetry-breaking parameter b has been determined, small adjustments to the cantilever can be made to reduce the magnitude of b. Such adjustments may possibly include the nonintuitive steps of decreasing the strength of some component (e.g., reducing one of the spring constants) in order to increase the strength of the compound system.

Much larger compound structures have many more degrees of freedom, and consequently many more failure modes. While some failure modes are clear-cut, others may be very subtle—depending on unexpected coupling between "soft" failure modes, as the example above illustrates. Characterization of the imperfection sensitivity is algorithmic: for any catastrophe germ (the potential of the perfect optimized structure at the failure load) there is a universal perturbation. This is finite dimensional provided that the germ is determinate. It is important to understand the universal perturbation in order to catalog all possible failure modes of the system, both obvious and not, in order to prevent catastrophes from occurring in such complex structures.

3.4 Elementary Catastrophes in Nonlinear Dynamics

Elementary catastrophe theory is the first result in the program of catastrophe theory. As a result, it stands at the base of the pyramid of bifurcations to be expected in nonlinear dynamical systems (see CHAOTIC PHENOMENA). Any bifurcations encountered in elementary catastrophe theory will be encountered in nonlinear systems with greater complexity.

To illustrate how the elementary catastrophes are encountered in more complex nonlinear systems, we survey here how the two simplest elementary catastrophes appear in nonlinear systems at the next level of complexity in the hierarchy of nonlinearity. Such systems are low-dimensional periodically driven dynamical systems or, equivalently, autonomous (time independent) dynamical systems with one more dimension.

In treating two-dimensional periodically driven dynamical systems and three-dimensional autonomous systems it is typical to explore the bifurcation diagram by sweeping one control parameter and locating periodic and aperiodic orbits. When a single control parameter is varied, it is typical to encounter among the catastrophes only the fold. If, however, there is a symmetry present, then the symmetry-restricted cusp catastrophe $[V(x;a)=\frac{1}{4}x^4+\frac{1}{2}ax^2=V(-x;a)]$ may also be encountered.

The fold catastrophe appears in bifurcation diagrams for autonomous dynamical systems as saddle-node bifurcations. These are shown in Fig. 17. In this figure there are several control parameter values at which new stable (solid curve) and unstable (dashed curve) solutions appear. If the system is periodically driven the saddle-node pair consists of a stable periodic orbit and a saddle orbit with one unstable direction. Both orbits have the same period.

Under a broad range of conditions the saddle remains unchanged as the control-parameter values are further increased, but the node undergoes additional bifurcations. Such bifurcations are often of the period-doubling variety. The period-doubling bifurcation is essentially a symmetry-restricted cusp catastrophe. Its relation to the cusp is shown in Fig. 18. As a control-parameter value is increased past some threshold, a stable periodic orbit loses its stability. In a transverse section (Poincaré plane) the unstable point is surrounded by two stable nodes. As the Poincaré plane is swept around the unstable periodic orbit, the two nodes will rotate around the direction of propagation of the Poincaré section either an integer number of times or a half-integer number of times. If the two nodes rotate an integer number of times around the saddle that is between them [Fig. 18(a)], the two nodes belong to two separate periodic orbits, both of which have the same period as the unstable periodic orbit from which they bifurcated. On the other hand, if the two nodes rotate a half-integer number of times around the saddle that they straddle [Fig. 18(b)], they belong to a single orbit whose period is twice the period of the unstable periodic orbit from which they bifurcated.

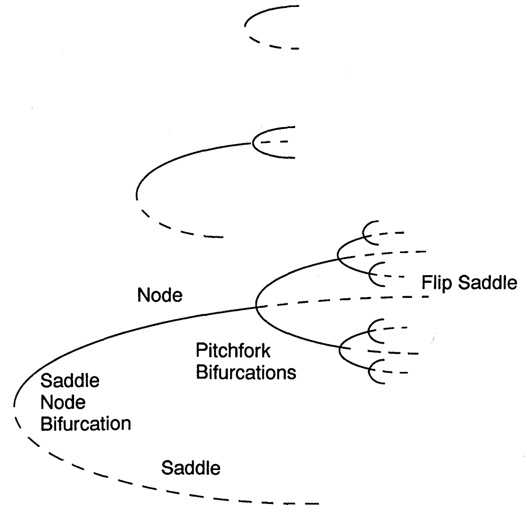

FIG. 17. Fold and symmetry-restricted cusp (also called pitchfork) bifurcations are commonly encountered in nonlinear dynamical systems. At a fold or saddle-node bifurcation a stable orbit (solid lines) and a saddle orbit (dashed lines) are created. At a pitchfork bifurcation, period doubling may occur, leading to creation of orbits of twice the period of the original orbit, which loses its stability.

Saddle-node bifurcations and period-doubling cascades are general features of autonomous dynamical systems. The pitchfork bifurcation with an integer number of twists is atypical in autonomous dynamical systems. However, if there is a symmetry, as occurs, for example, in the Lorenz equations, a period-doubling cascade cannot occur beginning at a symmetric orbit (Gilmore, 1981). As a result, the first pitchfork bifurcation occurs at a symmetric orbit of period T, creating an asymmetric pair of orbits of the same period. Subsequent pitchfork bifurcations from the asymmetric periodic orbits then generate period-doubling cascades.

A simple Hopf bifurcation can also be viewed as a symmetry-restricted cusp catastrophe (Gilmore, 1981). In the neighborhood of a Hopf bifurcation the dynamical equations can be written in polar coordinates as

$$\begin{aligned} dr/dt &= \lambda r \pm r^3, \\ d\theta/dt &= \omega. \end{aligned} \tag{55}$$

During the bifurcation the equation of motion for θ remains essentially unchanged. The bifurcation is therefore described by the radial equation, which can be written in gradient form:

$$\begin{aligned} dr/dt &= -\partial V/\partial r, \\ V(r;\lambda) &= -\tfrac{1}{2}\lambda r^2 - (\pm\tfrac{1}{4}r^4). \end{aligned} \tag{56}$$

A typical Hopf bifurcation is illustrated in Fig. 19. A section of the bifurcation surface in a plane containing the control parameter axis has the standard pitchfork shape.

Since the saddle-node, period-doubling, symmetry-breaking, and Hopf bifurcations are all related to elementary catastrophes, all

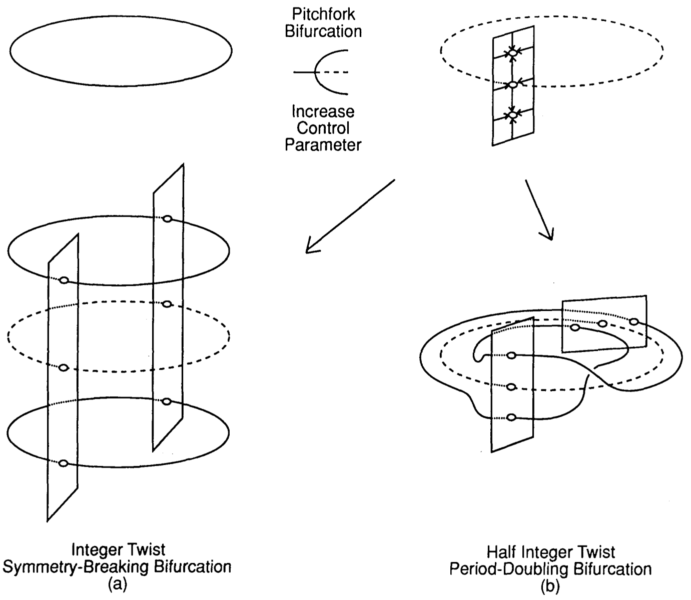

FIG. 18. When a periodic orbit loses its stability in a pitchfork bifurcation, either two new orbits of the same period may be created or one new orbit of twice the period may be created, depending on the local torsion about the original orbit (integer or half-integer).

the phenomenology associated with the elementary catastrophes is present and observable for these bifurcations as well. For example, one can observe modality, sudden jumps, inaccessible regions, sensitivity, hysteresis, divergence of linear response, time dilation, and anomalous variance in all these bifurcations. Furthermore, the behavior of each bifurcation has the same canonical form as that of the corresponding catastrophe (e.g., power-law dependence).

APPENDIX: A BRIEF HISTORY OF CATASTROPHE THEORY

Catastrophe theory burst upon public consciousness in the mid-1970s with a series of articles in widely available sources: *Scientific American, Science, London Times, Time,* and *The New York Times.* The articles themselves were bimodal. At one extreme the *London Times* heralded it as the "main intellectual movement of the century" while the *Science* article announced "The emperor has no clothes."

The theory of singularities of mappings was intensively developed by Whitney, Thom, and Mather in the decade 1955–1965. Thom presented his formulation of the theory of singularities in his eagerly awaited work *Stabilité Structurelle et Morphogénese* (1972). This book was an enigma in both form and substance. It was largely inaccessible to the mathematics community because it was written in the language of biologists, and inaccessible to the biological community due to its presentation of mathematical concepts which seemed to be deep and mysterious.

Yet it held the promise of describing discontinuous phenomena in a systematic way. Zeeman crowned the subject with the florid name "catastrophe theory" and took up the challenge of presenting these mathematical ideas to the larger community of scientists. When challenged to show concrete applications in the real world, he responded with a series of articles of stunning originality which explored the range of possible applications of this new subject.

These applications eventually worked their way into public consciousness through the popular journals. A reaction to the value of catastrophe theory set in. This reaction was

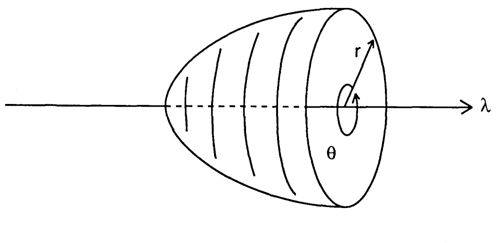

FIG. 19. A Hopf bifurcation can be viewed as a symmetry-restricted cusp catastrophe in the radial direction.

partly as a result of the overblown claims made in its name, partly as a result of the neglect offered to the workers who created the field of dynamical systems theory: Poincaré, Andronov, Pontryagin, Smale, etc., and those currently working in this field. The result was a multiyear public dialogue on the merits of this subject of discontinuities using arguments now long forgotten and best left unearthed, a dialogue of which the public eventually tired.

The appearance of the monographs by Poston and Stewart (1978) and Gilmore (1981) made it clear that this was a subject of substance, which had to be taken seriously, one capable of providing a useful language for the description of discontinuities at both a qualitative and a quantitative level.

GLOSSARY

Anomalous Variance: A catastrophe flag. Amplitude of motion about an equilibrium becomes increasingly large as a degenerate critical point is approached.

Bifurcation: A qualitative change in the properties of a system.

Bifurcation Set: Set of values in control-parameter space at which qualitative changes occur.

Catastrophe: Mathematical—A family of functions, depending on control parameters, in which the number of equilibria changes as the control parameters are varied. Physical—A sudden, discontinuous change in the state of a system.

Catastrophe Flags: Phenomena that occur when a catastrophe (sudden jump) is present.

Catastrophe Germ: A function which describes a degenerate critical point.

Catastrophe Theory: The program for determining how the qualitative properties of the solutions of equations change as parameters appearing in these equations change.

Clausius–Clapeyon Equations: A set of equations which determine the Maxwell set of a catastrophe.

Control Parameters: A set of parameters $c=(c_1,c_2,\ldots,c_k)$ which appear in the equations that describe a physical system.

Critical Curvature: The curvature d^2f/dx^2 of a function evaluated at its critical points $\nabla f=0$.

Critical Point: An equilibrium: a point at which motion does not occur.

Critical Set: The k-dimensional manifold $\nabla_x f(x;c)=0$ for $x\in R^n$ and $c\in R^k$.

Critical Slowing Down: A catastrophe flag. Relaxation to equilibrium takes increasingly long times as a degenerate critical point is approached.

Critical Values: Values of a function at its critical points.

Cusp Catastrophe: The function $A_3(x)=x^4/4+ax^2/2+bx$. So named because the critical set $\nabla_x A_3(x;a,b)=0$ exhibits a cusp on projection to the a-b control parameter plane.

Cuspoids A_{n-1}: A class of functions depending on one state variable x and $n-2$ control parameters. The family of functions $A_{n-1}(x)=x^n/n+\Sigma_{i=1}^{n-2}a_ix^i$ has up to $n-1$ isolated critical points.

Degenerate Critical Point: An equilibrium ($\nabla V=0$) at which the stability matrix $(\partial^2 V/\partial x_i\partial x_j)$ is singular. This requires two or more critical points to be arbitrarily close to each other (degenerate).

Delay Convention: The system state remains in a local minimum until that minimum ceases to exist.

Divergence of Linear Response: A catastrophe flag. Linear response coefficients increase very rapidly as a degenerate critical point is approached.

Dynamical System: A set of coupled first-order ordinary differential equations which may be nonlinear: $x_i' = f(x,c;t)$, where x is a state vector and c are control parameters.

Elementary Catastrophe: A function of one or two state variables (x,y) and one or more control parameters. These functions are used to describe systems in which small causes can produce large effects.

Equilibrium: A point at which all forces vanish; a critical point.

Exceptional Elementary Catastrophes: E_6, E_7, E_8 depend on two state variables and have up to 6, 7, 8 critical points, respectively.

Flow: The motion of a point in phase space under the equations of motion for the system.

Fokker–Planck Equation: An equation that describes the evolution of a probability distribution in the presence of diffusion and driving forces.

Fold Catastrophe: The function $A_2(x) = x^3/3+ax$. So named because the critical set exhibits a fold.

Generic: Typical. Occurs with "probability one."

Gradient Dynamical System: A dynamical system in which the time-independent forcing term is the gradient of a potential function: $f_i(x,c) = -\partial V(x,c)/\partial x_i$.

Hopf Bifurcation: Bifurcation in which a stable fixed point becomes unstable and emits a stable periodic orbit (limit cycle).

Hysteresis: Transitions from one state to another do not occur at the same values of the control parameters when the control parameters are changed in the opposite direction.

Implicit Function Theorem: A theorem of elementary calculus. If a function has a nonzero slope at a point, $\nabla f \neq 0$ at x_0, then it can always be approximated by a linear function at x_0.

Inaccessibility: An unstable physical state separating two stable physical states.

Isolated Critical Point: An equilibrium ($\nabla V=0$) at which the stability matrix ($\partial^2 V/\partial x_i \partial x_j$) is nonsingular.

Maxwell Convention: The system state is always in the deepest minimum.

Maxwell Set: The set of control parameter values at which a function has two or more equally deep minima.

Metastable: Stable to small perturbations but not to large perturbations.

Modal Catastrophe: A catastrophe germ which depends on one or more parameters.

Modality: Distinct types of behavior that a system can exhibit under identical or nearly identical conditions (e.g., liquid–gas states for a fluid).

Mode Softening: A catastrophe flag. Oscillation frequencies approach zero as a degenerate critical point is approached.

Morse *i*-Saddle: The quadratic function $M_i^n(x) = -x_1^2 - \cdots - x_i^2 + x_{i+1}^2 + \cdots + x_n^2$.

Morse Lemma: A normal-form theorem like the implicit function theorem. If a function of n variables has a nonsingular stability matrix (det $\partial^2 f/\partial x_i \partial x_j$) at an equilibrium x_0, it can always be approximated by a quadratic function at x_0. Such a function is called a Morse saddle.

Nongeneric: Atypical. Occurs with "probability zero."

Perestroika: Change in the properties of a dynamical system due to changes in its control-parameter values.

Perturbation: A function which is everywhere small.

Phase Space: A space in which the coordinates of a point define the state of the system.

Phase-Space Portrait: A trajectory (or trajectories) in phase space which describes the evolution of a physical system from one (or more) initial conditions.

Pitchfork Bifurcation: Bifurcation in which a stable fixed point becomes unstable and two new stable fixed points are emitted. Corresponds to a single-well potential developing into a double-well potential.

Quadratic Form: A function of n variables $x=(x_1,x_2,\ldots,x_n)$ which can be written x^tMx, where M is a symmetric $n \times n$ matrix.

R^n: Real n-dimensional space.

Saddle-Node Bifurcation: Bifurcation in which two fixed points, one saddle and one node, are created.

Sensitivity: The final state of a system may change under small perturbations of either the initial conditions or processes applied to the system.

Stability Matrix: The matrix of mixed second partial derivatives of a function: $\partial^2 f/\partial x_i \partial x_j$.

State Vector: A set of parameters $x=(x_1, x_2,\ldots,x_n)$ which describes the state of a system.

Structural Stability: Invariance of qualitative properties under a perturbation.

Taylor Tail: Tail of a Taylor series expansion: all terms beyond a certain degree.

Thom Classification Theorem: A list of elementary catastrophes depending on one or two state variables and up to five control parameters.

Thom Splitting Lemma: A normal form theorem for functions. If a function has an equilibrium ($\nabla f=0$) at which the stability matrix $\partial^2 f/\partial x_i \partial x_j$ has l vanishing eigenvalues, the function can always be written as the sum of two functions, $f(x) = f_{\mathrm{NM}}(y_1,\ldots,y_l) + M_i^{n-l}(y_{l+1},\ldots,y_n)$, where $f_{\mathrm{NM}}(y_1,\ldots,y_l)$ is a function with vanishing first and second derivatives in the l variables corresponding to the vanishing eigenvalues and M_i^{n-l} is a Morse i-saddle in the remaining $n-l$ variables.

Time Dilation: A catastrophe flag. System response takes increasingly long times as a degenerate critical point is approached.

Umbilics D_{n-1}: A class of functions depending on two state variables and $n-2$ control parameters. The family of functions $D_{n-1}(x,y)$ has up to $n-1$ isolated critical points.

Universal Perturbation: Simplest function (i.e., lowest dimension, or number of control parameters) that describes all possibilities that can result when a degenerate critical point is perturbed.

List of Works Cited

Arnol'd, V. I. (1981), *Singularity Theory,* Cambridge, England: Cambridge University.

Arnol'd, V. I. (1986), *Catastrophe Theory,* 2nd ed., Berlin: Springer-Verlag.

Gilmore, R. (1981), *Catastrophe Theory for Scientists and Engineers,* New York: Wiley.

Poston, T., Stewart, I. N. (1978), *Catastrophe Theory and Its Applications,* London: Pitman.

Thom, R. (1975), *Structural Stability and Morphogenesis,* New York: Benjamin–Addison–Wesley.

Zeeman, E. C. (1977), *Catastrophe Theory: Selected Papers (1972–1977),* Reading, MA: Addison–Wesley.

Further Reading

Anon. (1973), *Manifold* **14**, 6–13.

Anon. (1975), *Times Higher Educational Supplement,* Dec. 5, p. 12.

Arnol'd, V. I. (1974), *Proc. Int. Cong. Math. Vancouver 1974,* pp. 19–39.

Croll, J. (1976), *New Sci.* **17**, 630–632.

Fussbudget, H. J., Znarler, R. S. (1979), *Math. Intell.* **2**, 56–59.

Guckenheimer, J. (1978), *Math. Intell.* **1**, 15–20.

Saunders, P. T. (1980), *An Introduction to Catastrophe Theory,* Cambridge, England: Cambridge University.

Stewart, I. N. (1975), *New Sci.* **68**, 447–464.

Stewart, I. N. (1977), in: *Encyclopedia Britannica,* Special Supplement to the Yearbook.

Sussman, H. J., Zahler, R. S. (1978a), *Synthèse* **37**, 117–126.

Sussman, H. J., Zahler, R. S. (1978b), *Behav. Sci.* **23**, 383–389.

Thom, R. (1975), in: *Dynamical Systems—Warwick 1974,* Lecture Notes in Mathematics 468, Berlin: Springer-Verlag, pp. 366–372.

Thompson, J. M. T., Hunt, G. W. (1974), *Eng. Optim.* **1**, 99–110.

Thompson, J. M. T. (1982), *Instabilities and Catastrophes in Science and Engineering,* New York: Wiley.

Whitney, H. (1955), *Ann. Math.* **62**, 374–410.

Woodcock, A., Davis, M. (1980), *Catastrophe Theory,* New York: Avon.

Zahler, R. S., Sussman, H. J. (1977), *Nature* **269**, 759–763 (Correspondence Dec. 1, Dec. 29, 1977).

Zeeman, E. C. (1975), in: *Dynamical Systems–Warwick 1974,* Lecture Notes in Mathematics 468, Berlin: Springer-Verlag, pp. 373–383.

Zeeman, E. C. (1976), *Sci. Am.* **234**, 65–83.

CATHODOLUMINESCENCE

JOHN W. STEEDS, *H. H. Wills Physics Laboratory, University of Bristol, Bristol, England*

Introduction 121
1. **Different Forms of Cathodoluminescence** 122
2. **Absorption Processes** 124
3. **Emission Processes** 125
3.1 Introduction 125
3.2 Excitons and Excitonic Processes 125
3.3 Donor-Acceptor Pair, Free-to-Bound, and Band-to-Band Emission 127
3.4 Terminology 128
3.5 Deep Center Emission 128
3.6 Phonon Coupling and the Configuration-Coordinate Model 129
3.7 Summary 130
4. **Comparison between Cathodoluminescence and Photoluminescence** 130
5. **Physical Applications of Cathodoluminescence** 131
5.1 Introduction 131
5.2 Quenching of Excitonic Emissions at Dislocations in InP 132
5.3 Y^0 Emission from Dislocations in ZnSe 132
5.4 Colored Dislocations in Natural Diamond 134
5.5 Dislocation-Related Luminescence in Si 134
5.6 Dislocation Velocity and Misfit Dislocations 134
5.7 Abruptness of Semiconductor Quantum-Well Interfaces 135
5.8 Electron-Beam Pumping of Lasers 136
5.9 Investigation of Radiation Damage 136
Glossary 137
Works Cited 138
Further Reading 139

INTRODUCTION

Luminescence spectroscopy is well established as a technique for nondestructive evaluation of materials. An introduction to the literature on the subject is given in the suggestions for Further Reading at the end of this article. By the use of an incident electron beam ("cathode rays") as the source of excitation a very wide range of materials can be studied, including semiconductors, ceramics, geological minerals, polymers, and biological and medical samples. Because the beam is easily focused to a small exploratory probe, the technique is not only very versatile but also capable of providing information with high spatial resolution.

Luminescence of inorganic materials is frequently concerned with low concentrations of point defects, either singly or in small clusters. Phosphors are one exception to this generalization, since high concentrations of additives are commonly required; another is the free-exciton emission from pure quantum-well structures. By contrast, luminescence of organic material is generally produced by intrinsic molecular units of the structure. Unfortunately, electron irradiation is a serious cause of damage to organic material, although the rate of damage varies widely.

When point defects are responsible for luminescence, they can be either intrinsic (vacancies or interstitial atoms) or extrinsic (impurities or dopants). Best understood, at the moment, is the luminescence associated with isolated point defects in rather pure materials. As the technique is very sensitive, extremely low concentrations of impurity can

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

be detected in ultrapure material, even at levels below parts in 10^9 in silicon, for example. The position of an added atom in the lattice is very important. Completely different results can be obtained from a foreign atom depending on whether it is in an interstitial or substitutional site. If the atom is in an interstitial position there may be a variety of sites, each with a different local symmetry and each leading to a different optical effect. Two-atom complexes of dissimilar foreign atoms are also well understood in a number of instances. However, as the number of atoms involved in a luminescence center increases, our understanding of its behavior rapidly decreases. It seems that it is generally the case that as atoms cluster together the associated luminescence properties, if any, are lost, well before any precipitate is formed which might be directly visible by a modern transmission electron microscope.

There are, of course, two extremes of approach to the investigation of materials by cathodoluminescence. One seeks to catalog and characterize rather than explain; the other seeks a detailed electronic and atomic understanding of the light-emitting process. While both undoubtedly have their place, it is perhaps timely to warn those unfamiliar with the technique that the second approach can be extremely laborious, and several common emission lines have been studied with some persistence over periods of ten years or more without yet fully revealing their secrets. Moreover, it is probably the case that cathodoluminescence is a less potent tool for securing fundamental understanding than is photoluminescence. A comparison between the two techniques, cathodoluminescence and photoluminescence, is included in what follows.

1. DIFFERENT FORMS OF CATHODOLUMINESCENCE

In view of the very wide range of applications of cathodoluminescence in different fields of scientific research, it is not surprising that a comparably wide range of equipment has been developed to record the results with different objectives in mind. It is, therefore, very important to be aware of the nature of the apparatus used to generate the results of interest and its operating conditions before attempting to assess them. It is not difficult to find examples where completely different results can be obtained from the same region of a semiconductor specimen examined under different extremes of operating conditions.

Although wide differences of equipment exist, the one common point of examination by cathodoluminescence is that the energy of the electrons used for excitation of the specimen under study is very much greater than the band-gap energy. This characteristic is in clear distinction with photoluminescence, a technique contrasted with cathodoluminescence later in this chapter, and it confers on cathodoluminescence its wide range of application.

The first important distinction between cathodoluminescence systems is the nature of the instrument itself. Four broad categories exist. These are topographic cathodoluminescence, and scanned beam cathodoluminescence performed either in a scanning electron microscope (SEM), in a transmission microscope (STEM), or in an ultrahigh-vacuum chamber used for surface science or analysis.

Although topographic cathodoluminescence can be performed in a scanning electron microscope (Lang, 1980), it is more normally carried out in a simple system which is only capable of bathing the specimen in a broad beam of electrons of relatively ill-defined energy (Hagni and Karakus, 1989). The image will often be recorded on color film with or without the use of filters, although spectrometry is sometimes also performed on the light generated.

The most common form of cathodoluminescence analysis of semiconductors is that performed in a scanning electron microscope adapted for that purpose. The cathodoluminescence images are generated with relatively high spatial resolution by a scanned beam, and although many manufacturers provide attachments for cathodoluminescence experiments, many individually developed systems are also in use. It follows that the performance of this form of instrument varies considerably from one installation to another.

For higher spatial resolution studies a few cathodoluminescence systems have been installed on scanning transmission electron microscopes. These are at present individually developed systems (for examples, see Petroff *et al.*, 1978; Pennycook, 1982; Berger and Brown, 1984; Roberts, 1981; and Myhajlenko, 1984). The high spatial resolution is achieved by the use of thin specimens, albeit at a significant loss of signal. Much attention has therefore been given to efficient collection and detection of the light emitted.

For those interested in light emission from surfaces, cathodoluminescence detectors have been added to surface science equipment (Brillson *et al.*, 1988).

The second important consideration, having established the category of instrument used for generating cathodoluminescence information, is that of operating conditions. These can vary widely and may be linked to the system in use. One important parameter of operation is that of electron energy. In scanning electron microscopes and scanning transmission electron microscopes, this energy can vary widely and is used to control the range (penetration depth) of the electrons. The energy of the beam is deposited in a pear-shaped volume in a bulk sample, and the electron range is given approximately by the expression

$$R=(k/\rho)(E_b)^{\alpha}, \qquad (1)$$

where k and α are material- and electron-energy-dependent constants, ρ is the density of the material, and E_b is the beam energy. For a discussion of this point the reader is referred to the review article of Yacobi and Holt (1986) or to the original work of Everhart and Hoff (1971) or Kanaya and Okayama (1972).

Thus, relatively low-energy electrons give shallow penetration, while high-energy electrons can be used to excite at greater depths or to penetrate surface coatings so as to reach underlying cathodoluminescent layers. The simple control over range of excitation is a very attractive feature of cathodoluminescence. The details of the microscopic processes involved in absorption of energy from the incident electron beam are apparently not very important. The straggling of the beam through large-angle deflections is largely the result of phonon scattering, but this does not contribute directly to energy loss to an appreciable extent. The two major components of energy loss are plasmon excitation and valence-electron excitation with a smaller contribution from core-electron excitation. The loss processes are reviewed, for example, by Reimer (1984). However, there is no evidence to suggest that hot electrons created during energy loss from the high-energy beam lead ultimately to differences between cathodoluminescence (CL) and photoluminescence (PL) spectra. It is a very good first approximation to regard the electron beam as a controllable source for production of excess carriers in a chosen location of the material under study.

A second important operating parameter is the beam current (excitation level). This may well be determined by the experimental conditions since poor light emitters or inefficient collection and detection systems call for high beam currents. One result of the use of high beam currents is that appreciable specimen heating may occur. However, this is a relatively unimportant effect compared with spectral changes which occur as the excitation level is changed. Luminescence spectra are often highly nonlinear. Defect-related luminescence saturates when the defects are all excited, but the level of excitation to cause this depends on the defect density, which can vary enormously. At high excitations super-radiance of intrinsic luminescence can occur through population inversion. At low temperatures excitons can bind together to form multiple bound-exciton complexes or even condense into excitonic droplets at sufficiently high excitations. As a result, great care must be exercised in comparing results obtained from different systems, possibly operated at very different excitation levels.

A third important parameter is the specimen temperature. This is particularly the case for rather pure semiconductors where the so-called edge emission is predominant over deep-level emission (see Sec. 3). Enormous changes occur in the band-edge emission of such materials in the range 2–50 K, say, and indeed the spectral details may often be used to determine the specimen (or else the electron) temperature very accurately. As manufacturers of liquid-helium stages are sometimes rather optimistic about the performance of their products, temperatures quoted in the literature should be treated with caution unless substantiated by direct evidence from luminescence spectra. At a more obvious level there are usually considerable changes between band-edge luminescence at room and liquid-helium temperatures as a result of thermal ionization of shallow acceptors and donors at ambient temperatures.

Some cathodoluminescence systems are equipped to generate monochromatic maps. Clearly, materials generating spectra rich in detail will produce a wide range of monochromatic images, many of which may bear little relationship to the integrated cathodoluminescence image of a simpler system. In some cases it is also important to bear in mind the spectral range over which the "monochromatic" image is collected. A few systems are also equipped for time resolution of images (Bimberg *et al.*, 1985), and the images obtained depend strongly on the time scale over which they were generated.

Two further points should be noted. Cathodoluminescence images from bulk specimens in scanning electron microscopes may also differ markedly from those obtained from thin specimens in scanning transmission electron microscopes. A number of factors contribute to these differences, and it is well known that spectral details change considerably as the specimen thickness is reduced. The proximity of nearby free surfaces affects different emission processes to different extents (some processes can become completely quenched), the effective carrier diffusion length is reduced, and internal stress fields may be altered and internal electrical fields may be generated. The final point, not well understood at present, is concerned with beam-dependent changes induced during examination. These effects are evidently column vacuum dependent and specimen dependent. A number of independent factors can be identified. These include carbon or ice (cooled specimens) contamination effects, hydrogen passivation as a result of hydrocarbons in the system, and displacement or ionization damage of the sample. In addition, well-aligned electron-optical columns with line of sight from specimen to source are also liable to ion damage from the electron gun. This little-appreciated problem is widely overlooked and has been found to be a serious problem in work in this laboratory. The effect can be monitored using nuclear track emulsions in the specimen position which are developed after exposure (Tastrak CR39 has been used for this purpose; see Henshaw, 1982). It can be overcome by introduction of a dog's-leg into the column, although this step is not popular with manufacturers who have spent years developing well-aligned optics.

2. ABSORPTION PROCESSES

The absorption of energy from the incident electron beam occurs by a variety of different processes such as core-electron excitation, plasmon excitation, and phonon scattering. We are only interested in those processes which lead to the creation of electron-hole pairs in the near-ultraviolet, visible, or near-infrared parts of the spectra. These pairs may take the form of free mobile carriers or bound excitons of considerable spatial extent (see Sec. 3) in semiconductors, or of individual localized traps in insulators.

The details of what happens next depend on whether the excited material is a semiconductor or an insulator. In the case of insulators the carriers which are created diffuse until they either combine radiatively or nonradiatively. It is common to associate a lifetime with each of these decay modes, referred to as τ_R and τ_{NR}, respectively, and an overall lifetime τ given by

$$\frac{1}{\tau}=\frac{1}{\tau_R}+\frac{1}{\tau_{NR}}. \tag{2}$$

The nonradiative behavior may be associated with Auger transitions or phonon cascades; nonradiative transitions may occur at internal or surface sites. It follows that any nonradiative process which is particularly efficient will completely dominate the nonradiative lifetime. This may be an internal center, sometimes called a killer center, or surface recombination, which can be very efficient. Clearly, the nonradiative recombination of thin films is likely to be dominated by the surface recombination rate, and a quantity called the surface recombination velocity with units of cm/s is widely used as its measure. Surface passivation treatments such as hydrogen passivation exist to reduce the surface recombination velocity. The radiative lifetime is also a quantity with several potential components. The first and most obvious component is that of free minority carriers before radiative recombination. However, radiative transitions also occur from carriers which become trapped at defect centers. Alternatively, the electrons and holes may combine to make either free excitons before radiative recombination, or else excitons that themselves become trapped by defect centers, later recombining radiatively. It therefore becomes important to distinguish between the capture time of excited carriers and the recombination time after entrapment. The former is inversely related to a carrier diffusion length which can vary greatly from one semiconductor to another. For example, GaAs normally has a minority carrier diffusion length of a few micrometers, while the diffusion length in relatively pure silicon can be several hundred micrometers.

In summary, it is evident that the terms "diffusion length" and "lifetime" cover a great deal of complex detail, and that the precise behavior of the excitation after creation of electron-hole pairs will be strongly temperature and impurity dependent. They are, nevertheless, very useful in solving diffusion equations in investigating the effects of carrier injection by the electron beam. The various recombination routes, radiative and nonradiative, all compete with each other as recombination channels for the excitation which the electron beam creates. Thus cathodoliminescence spectroscopy is quite different from x-ray spectroscopy in electron microscopy where each of the core shell absorptions is independent and the effects are additive. A particularly efficient emission center can completely dominate the cathodoluminescence spectrum.

The situation in insulators is considerably less complicated. The excitation is either directly absorbed by a radiative center, or "activator" as it is called in the case of phosphors, or else transferred to a radiative center by the absorbing center ("promoter" or "sensitizer" in the case of phosphors) (Garlich, 1958; Blasse and Bril, 1970).

3. EMISSION PROCESSES

3.1 Introduction

When light emission occurs there are a number of different processes to distinguish. By far the most important are those giving rise to the relatively sharp and distinctive features on which we shall concentrate. The greatest variety of effects is encountered in the case of semiconductors which have band gaps corresponding to visible or near-infrared emission. It is important to bear in mind the distinction between direct gap materials not requiring phonon assistance and indirect gap materials where phonon assistance is required to achieve photon emission during electron-hole annihilation. Table 1 lists some common materials with the nature of their band gaps and the associated energies at room temperature.

Table 1. Natures and magnitudes of the band gaps of different materials.

Type of material		Type of band gap	E (room temperature) in eV
	Al_2O_3, quartz		~9
Rock salt	LiF	Direct	13.6–13.7
	NaCl		~8
Group IV	Diamond	Indirect	5.48
	Silicon	Indirect	1.110
	Germanium	Indirect	0.664
6H cubic	α SiC	Indirect	2.86
	β SiC	Indirect	2.2
Group III-V	Cubic BN	Indirect	6–8
	GaN	Direct	3.44
	GaP	Indirect	2.26
	AlAs	Indirect	2.15
	GaAs	Direct	1.42
	InP	Direct	1.351
Group II-VI	MgO	Direct	7.70 (85 K)
	ZnS	Direct	3.68
	ZnO	Direct	3.35
	ZnSe	Direct	2.7
	CdS	Direct	2.5
	ZnTe	Direct	2.3
	CdSe	Direct	1.75
	CdTe	Direct	1.5

3.2 Excitons and Excitonic Processes

For semiconductors, at low enough temperatures, an electron in the conduction band will bind with a hole in the valence band to form a so-called exciton. To a reasonable first approximation this free-exciton formation can be treated by Bohr theory to yield values for the binding energy and exciton diameter as follows:

$$E_X = -\frac{m_r^* e^4}{2(4\pi\epsilon_0)^2\hbar^2\epsilon^2} = -13.6\left(\frac{m_r^*}{m_0}\right)\frac{1}{\epsilon^2}\ \text{(eV)}, \tag{3}$$

$$a_x = \frac{\hbar}{(2m_r^*|E_X|)^{1/2}}\,. \tag{4}$$

m_r^* is the reduced mass given by

$$\frac{1}{m_r^*}=\frac{1}{m_e^*}+\frac{1}{m_h^*} \tag{5}$$

with m_e^* and m_h^* averaged effective electron and hole masses, respectively, appropriate to the relevant energy gap. Note that the indirect-gap materials are likely to have particularly anisotropic effective masses (as, for example, in Table 2). e is the electronic charge, $\hbar$ Planck's constant, ϵ_0 the vacuum permittivity, ϵ the static dielectric constant of the material, and m_0 the electron rest mass.

Note that the free-exciton binding energy is closely related in the same approximation to the donor (E_D) and acceptor (E_A) ionization energies, as follows:

$$E_D=(1+\sigma)E_X, \tag{6}$$

$$E_A=E_D/\sigma, \tag{7}$$

where

$$\sigma=m_e^*/m_h^*. \tag{8}$$

Table 2 gives representative values of some of the parameters, energies, and radii which appear in these equations for a few common semiconductors. These energies are useful in estimating the temperatures at which excitons become thermally unstable and donors or acceptors become ionized.

Free-exciton emission (X) is rather weak and rarely observed in cathodoluminescence. There are three reasons for this. First, the exciton field becomes mixed with the photon field present in the crystal generating hybridized states called polaritons. Second, the associated momentum not only leads to considerable line broadening but also gives rise to problems about momentum conservation when a photon is emitted. Third, and most important, is that the free excitons are relatively easily trapped at low temperatures by a variety of centers, notably neutral donors (D^0), neutral acceptors (A^0), and isoelectronic traps (an isoelectronic trap is composed of two neighboring substitutional impurities with the same combined valence as the host atoms which they replace) as well as at present unidentified defects. The large size of the excitons, which comes from the large dielectric constants and small reduced masses of semiconductors, enhances the ease of entrapment. Once trapped, the excitons, lacking momentum, give very sharp emission lines from relatively pure materials at low temperatures. The binding energies (E_{BX}) of the excitons to neutral donors and acceptors are found experimentally to be approximately proportional to the respective donor or acceptor ionization energy (E_i),

$$E_{BX}=a+bE_i, \tag{9}$$

where the constants a and b in the equation (known as Haynes's rule) are characteristic of the semiconductor concerned. As donors are more easily ionized than acceptors (Table 2), acceptor-bound excitons survive to higher temperatures than donor-bound excitons. In fact, the binding energies of excitons to donors and acceptors varies slightly from one impurity or dopant to another, but these variations are normally too small to detect by cathodoluminescence spectroscopy. Materials studied by cathodoluminescence are not normally exceptionally pure or lightly doped, and the operating temperatures are usually not particularly low, more often closer to 30 K than 4 K, so that the resulting spectra are relatively broad. The approximate formulas given above are particularly appropriate under these circumstances.

Table 2. Some typical experimental values for quantities relating to the free-exciton binding energy and its binding energy to neutral donors and acceptors, together with the calculated exciton radius.

Material	m_e^*/m	m_{lh}/m	E_X (meV)	ϵ	E_{D^0X} (meV)	E_{A^0X} (meV)	a_X (nm)
ZnSe	0.16	0.6	−19	8.66	−(2–6)	−(8–12)	3.6
GaAs	0.0665	0.087	−4.2	12.53	−1	−(2.4–2.7)	17.5
InP	0.079	0.12	−5.12	12.61	−1	−(3.5–4)	14
Si	0.9163 (long.) 0.1905 (trans.)	0.153	−14.3	11.7	−(5–5.6)	−(4.2–6.0)	5.2

There is a large variety of more complicated transitions involving excitons, some of which may not be resolvable in cathodoluminescence experiments for the reasons already mentioned. A few of the more important of these will be discussed. First, excitons may bind together (self-trapping) to form excitonic molecules (biexcitons), triexcitons, or, in general, bound multiple-exciton complexes (BMECs). Second, more complex transitions involving bound excitons can occur. When a donor (acceptor) is simultaneously excited, say from its 1*s* to a 2*s* or 2*p* state as the exciton collapses emitting a photon (of reduced energy), the spectral lines are called two-electron (hole) satellites. Examples of two-electron satellites (TES) will be discussed later (Sec. 5). At high concentrations where the impurity or dopant atom separations become comparable with or smaller than the exciton size, excitons will tunnel from one site to another and are no longer trapped. Although these free excitons (FE) will not usually annihilate directly because of the problem about momentum conservation, they may undergo inelastic transitions from 2*s* or 2*p* states down to the ground (1*s*) state, and the associated light emission has been observed in cathodoluminescence spectra from ZnSe (Sec. 5) and are referred to here as FE-(1*s*-2*s*).

The theory of weakly bound excitons, found in semiconductors, was first developed by Wannier and Mott, and so the quasiparticles are often referred to as Wannier or Wannier-Mott excitons. In insulators the binding energy is much higher, the excitons are of atomic dimensions, and they are known as Frenkel excitons in recognition of Frenkel's pioneering work on the theory relevant to this situation.

3.3 Donor-Acceptor Pair, Free-to-Bound, and Band-to-Band Emission

Another very common form of luminescence, with photon energies only slightly lower than those of the excitons, is known as donor-acceptor pair (DAP) emission. In this case an electron on a neutral donor annihilates with a hole on a neutral acceptor emitting a photon of appropriate energy. Although of roughly similar energy, the underlying mechanism is very different and, as a result, a quite different spectral width and shape is observed. The important point about this form of emission is that the participating donor and acceptor atoms are initially in positive and negative charge states, and hence possess an attractive Coulomb energy. The magnitude of this energy depends on the separation between the atoms, and this energy is added when the photon is emitted. Photons emitted from near neighbors will have higher energies than photons from distant atoms, but there is only a relatively small number of potential near-neighbor sites. It follows that the emitted photon has an energy given by

$$h\nu=E_g-(E_A+E_D)+\frac{e^2}{4\pi\epsilon_0\epsilon r}, \tag{10}$$

where E_g is the gap energy and r is the interatomic separation. Relative to the excitonic peaks, a broad peak is predicted with its maximum related to the average separation between the donors and acceptors. The peak will be asymmetric in energy about the maximum with a relatively short tail on the near-neighbor high-energy side and a long tail on the distant-neighbor low-energy side. In addition, the peak position will depend on the excitation level, shifting to higher energy at higher excitations because the distant pairs have low recombination rates and are easily saturated.

In order to observe DAP emission it is necessary to have an appreciable population of neutral donors and acceptors. Even when present in the semiconductor the donors will become thermally ionized at relatively low temperatures even though neutral acceptors remain. When this situation prevails, free electrons from the conduction band will annihilate with holes on the neutral acceptors, giving rise to the so-called free-to-bound (F-B) emission. In most semiconductors the electron is free and the hole is bound to an acceptor, thus neutralized, and this emission is therefore generally referred to as eA^0. The photon energy is given by

$$h\nu=E_g-E_A. \tag{11}$$

Note that a further characteristic of DAP emission is thereby created. As the free-to-bound emission is of higher energy, temperature-dependent experiments will reveal an additional peak on the high-energy side of the DAP peak with increasing temperature. At high enough temperatures the DAP will vanish, being replaced by eA^0 emission. Note that this increase in emission energy with rise in temperature is in the opposite sense from the band-gap variation with temperature, which decreases. Where time-dependent emission can be studied a further characteristic of DAP emission is available. At short delay times near-neighbor emission predominates given an increased peak energy, while at longer times the peak shifts to lower energy.

Finally, at even higher temperatures the acceptors become ionized, and direct cross-gap recombination occurs (band-to-band emission).

3.4 Terminology

The emission processes described so far all have emission energies close to that of the band gap. In older literature this form of emission was called "band-edge" emission to distinguish it from lower-energy "deep-level" or "deep-center" emission. More recent literature refers instead to "near-gap" emission and reserves this term for the sharp high-energy spectral lines of well-resolved low-temperature spectra. In the case of cathodoluminescence the distinction between band-edge and near-gap emission is not so clear-cut. An important, if crude, measure of the quality of a semiconductor specimen is the ratio of the maximum peak height of near-gap or band-edge emission to that of the deep-level emission. High-quality materials are completely dominated by the former, while band-edge emission may be almost entirely absent in the case of material of low quality.

3.5 Deep Center Emission

Many deep centers are nonradiative. Even when radiative, deep levels generally give rather broad featureless emission which is not capable of providing much detailed information about the material under study. However, there are a number of notable exceptions which are well researched and richly informative. Much of the luminescence of insulators fits into this category with equivalent centers occurring in the case of semiconductors. Of course, it follows that any center emitting in the visible region of the spectrum from an insulator must correspond to a deep level in the insulator band gap. Much studied in the past have been the vacancy-related color or *F* (Frenkel) centers in alkali halide crystals. Transition-metal impurities and rare earths are two other common forms of deep centers in insulators. This field has been recently comprehensively reviewed in the book by Henderson and Imbusch (1989).

The reason why the transition-metal and rare-earth impurities give sharp characteristic deep-level luminescence is because the atoms contain inner incomplete electronic shells which are relatively little affected by the atomic environment. The emission observed is therefore closely related to the atomic emission, modified, however, because of the crystal fields to which the atoms are subjected. The symmetry of the site occupied by the atom is therefore critical, and the peaks observed can be labeled as allowed transitions between levels with specific symmetry properties. Group theoretical labels are customarily attached. Very similar behavior is observed by deep-level emission from transition metals in GaAs and InP, but since the band gaps are relatively small the deep-level emission occurs rather far into the red end of the spectrum, calling for a lead sulfide or equivalent detector.

F centers are relatively straightforward to treat theoretically, and hence the luminescent spectra can be analyzed in some detail. The simplest *F* center is that of an electron trapped at a single anion vacancy. An elementary treatment of this problem for alkali halides gives the *F* band emission energy as

$$E_F = 1.13/a^2 \quad (\mathrm{eV}), \tag{12}$$

where a (nm) is the lattice constant of the material. Experimental results plotted as ln E_F vs ln$a$ give a reasonably straight line with a best fit of

$$E_F = 0.97a^{-1.772}. \tag{13}$$

Sharp deep-level emission is also an important characteristic of irradiated silicon. In particular, two centers have been widely studied and are relatively well understood. One of these, called the *G* center, has been the subject of considerable recent research and is now known to result from a complex formed from carbon on an interstitial site together with carbon on a substitutional site (Davies *et al.*, 1987; Davies *et al.*, 1989). The second is the *C* center, which has been identified as a carbon-oxygen complex (Chappell *et al.*, 1988).

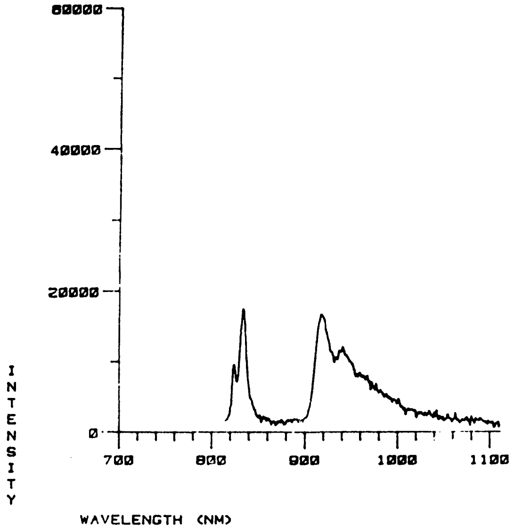

FIG. 1. Cathodoluminescence of GaAs containing copper. The broad copper-related band starts at 900 nm and has reasonably strong phonon coupling. First and second phonon replicas occur at approximately 940 and 960 nm (courtesy of Dr. S. Bailey). Temperature is ~30 K.

3.6 Phonon Coupling and the Configuration-Coordinate Model

The phonon coupling associated with each of the above emission processes provides a further means of distinguishing between them. In general terms the larger the extent of the wave function associated with a center, the weaker the phonon coupling. Excitons, which are extensive, couple very weakly with the lattice. Donor-acceptor pairs, which approach atomic dimensions, couple more strongly. Deep levels couple more strongly still. The *G* and *C* centers in silicon involve such large distortions that the strong local modes induced essentially overcome the indirect nature of the material and hence give rise to very strong emission. In order to discuss these differences it is helpful to introduce the idea often referred to as the configuration-coordinate model. In this simplified model we use the Born-Oppenheimer principle to separate electronic and ionic motion. We try to represent with one parameter, *Q*, called the configuration coordinate, what should in reality be described by a very large number of parameters. This parameter describes the state of distortion at an optical center and is related to the mean outward displacement of the neighboring atoms. We further assume that there is just one mode of lattice distortion, the breathing mode, with angular frequency ω, and that all static and vibrational distortions are of this type. We assume that the distortions can be described by a harmonic oscillator; and although excited states of the center will generally have different values of *Q*, we assume that the breathing mode is unaffected by excitation. Although very useful, this harmonic model misses the parity-breaking vibronic (dipole) transitions which may, in fact, be important.

If the crystal is in its ground state, we can write its energy as

$$V=\tfrac{1}{2}M\omega^2(Q-Q_0)^2, \tag{14}$$

where Q is the configurational coordinate (lattice normal coordinate), Q_0 is the average configuration coordinate in the ground state, and M is an equivalent ionic mass.

On excitation of the electron system an exactly similar vibration is assumed, shifted upwards in energy, and displaced to some new average configuration coordinate Q_0' as a result of the excitation. According to the Born-Oppenheimer principle there is no change of lattice arrangement during transition ($\Delta Q=0$), and vertical transitions occur in this diagram. Excitation from the ground state leads to an excited state with vibrational energy. On relaxing to the energy minimum in the excited state, phonon emission leads to a return to the ground state with vibrational energy. As a result there is an energy shift between absorption and emission which is known as the Stokes shift. The strength of the phonon coupling to the photon emission is determined by the horizontal shift ΔQ ($=Q_0'-Q_0$) between the two parabolas. Large shifts correspond to strong phonon coupling. In dimensionless terms this shift is measured by a parameter known as the Huang-Rhys factor (S) given by

$$S\hbar\omega=\tfrac{1}{2}M\omega^2\Delta Q^2. \tag{15}$$

In terms of S, the Stokes shift is given by $(2S-1)\hbar\omega$. The strength of the phonon replication may be calculated for different values of S. Examples of transitions with small S values can be found in the discussion of results for ZnSe presented in Sec. 5. The copper deep level in Fig. 1 has a relatively large S value. This behavior is easily understood in the qualitative terms introduced earlier. The most weakly bound states have the largest dimensions and hence couple most weakly to the lattice. Electrons couple more weakly to the lattice than holes for the materials in Table 2 because of the ratio m_h^*/m_e^*; electrons in GaAs or InP couple more weakly than in ZnSe for the same reason.

Questions also arise about the details of the phonon coupling. In ZnSe, InP, and GaAs, which are direct-gap materials, the strongest coupling is with zone-center longitudinal-optical (LO) phonons. This can be understood in physical terms as the result of effective coupling to the macroscopic electrostatic field associated with the LO phonons (Dean *et al.*, 1978). By way of contrast, the strongest phonon replicas in the luminescence from silicon involve transverse-optical (TO) phonons. As silicon is an indirect semiconductor it is clear that phonon coupling is required for momentum conservation, and the band structure of silicon reveals at once that the phonons must come from close to the *X* point (Madelung, 1982). However, while in germanium (also indirect) LO and TA (transverse-acoustic) coupling is forbidden by symmetry (Lax and Hopfield, 1961), this is not the case for silicon. A detailed examination of the phonon coupling of light emission from silicon has been given (Smith and McGill, 1976).

3.7 Summary

In summary of this section, the characteristics of particular spectral details of interest can generally be used to determine the origin of the cathodoluminescence emission. Apart from the emission energy, peak shape, and peak width, three other characteristics are of great value in seeking an identification, namely, the temperature dependence, excitation level dependence, and time dependence of the peak. Unfortunately, none of these last three experiments is easy to perform. Each can be affected by beam-induced changes, and temperature or excitation level experiments should be concluded by returning to the initial conditions to check whether any time-dependent changes have occurred. In addition, the temperature range and control is likely to be rather limited in cathodoluminescence experiments. Time resolution is perhaps the most promising approach, and Bimberg *et al.* (1985) have obtained some impressive results by the use of this technique. In particular, excitonic and DAP lines are easily distinguished by their very different transition times. However, very short time scales of picoseconds or nanoseconds are likely to be involved, and so considerable modification of conventional electron-optical equipment is required to exploit the technique.

4. COMPARISON BETWEEN CATHODOLUMINESCENCE AND PHOTOLUMINESCENCE

Although superficially similar, there are, in fact, considerable differences between these two different methods for generating emission spectra from solids. One of the major differences is the electron energies normally used in cathodoluminescence in comparison with the photon energies used in photoluminescence. The high electron energy means that cathodoluminescence is very versatile and can be applied both to insulators with large band gaps and to small band-gap semiconductors. Another attractive feature of the high voltage is the ease of control of electron range by altering the operating voltage. However, there is no equivalent to the technique of tuning the photon energy to near the band gap of the semiconductor being studied or to an absorption peak of an insulator. This technique, called photoluminescence excitation spectroscopy (PLE), provides some of the richest spectral detail and permits the measurement of Stokes shifts. The range of excitation can also be controlled in photoluminescence by this technique since the absorption coefficient varies very rapidly with change of wavelength near an absorption edge.

An additional price to pay in performing cathodoluminescence as compared with photoluminescence is the much greater likelihood of beam-induced changes to the specimen. This can occasionally be advantageous when the electron beam is used to create luminescence centers which did not exist before irradiation. A case in point is in the production of *G* centers in silicon (Canham *et al.*, 1987). As other radiation-induced centers become better understood in the future, this aspect of cathodoluminescence might become quite useful.

One of the big advantages of photoluminescence on very pure materials is the ease with which experiments can be performed with the sample immersed in liquid helium below the λ point. However good the liquid-helium-cooled stages become in electron microscopes, such low temperatures are unlikely to be attained, and there will always be some difficulty in deducing the actual specimen temperature.

Although cathodoluminescence has the obvious advantage of small lateral spot sizes, this is not generally effective because of the large electron range in bulk specimens and the subsequent carrier diffusion. Only in the case of scanning transmission electron microscopes is the benefit of high spatial resolution realized, and even then only after investing a considerable amount of effort into the design of the system. Once achieved, however, there is the added and in some cases essential information provided by the ability to image directly and identify extended crystal defects such as dislocations, stacking faults, grain boundaries, or precipitates.

In summary, for samples of small dimensions or fine microstructure which are not of the highest purity and perfection, cathodoluminescence is a relatively quick, relatively high spatial resolution technique for providing, often in a nondestructive fashion, valuable information not otherwise easily obtainable. Moreover, when the facility exists on a scanning electron microscope it can be available as a routine examination procedure of samples under study.

5. PHYSICAL APPLICATIONS OF CATHODOLUMINESCENCE

5.1 Introduction

Cathodoluminescence is a technique with a long history. In this section I have picked a few topics from recent research which have made an important impact on either pure or applied physics. For references to the earlier literature the section on Further Reading at the end of this chapter should be consulted. This includes recent reviews emphasizing the relevance of cathodoluminescence to the study of ceramics, geological materials, organic materials, and in medical science. I restrict my examples largely to semiconductors.

The first aspect of recent cathodoluminescence which deserves to be emphasized relates to the rather obvious point that luminescence centers are not necessarily homogeneously distributed. While few would doubt this conclusion, it is most important to have concrete evidence about the spatial distribution of active centers. Cathodoluminescence has begun to provide such information on an impressive scale. It has revealed depth dependences of active centers in cross sections through epitaxial layers, marked inhomogeneities in bulk materials, and epilayers (Chu *et al.*, 1981; Chin *et al.*, 1984), sometimes associated with dislocation cell boundaries, sometimes as a result of processes occurring during growth without the presence of extended crystallographic defects (Warwick and Booker, 1983). Luminescence variations associated with dislocations, stacking faults, so-called oval defects in molecular-beam epitaxy (MBE) layers, grain boundaries, and homostructure and heterostructure interfaces have all been revealed (Bailey and Steeds, 1989a,b; Steeds, 1989b).

Three examples will be discussed as illustrations of the nature of the information which can be obtained. Two were studied by cathodoluminescence in a STEM, the third by cathodoluminescence topography in a SEM. They are each concerned with dislocations; the first is the change of cathodoluminescence observed near a single grown-in dislocation in liquid encapsulation (LEC) InP substrates annealed at a high temperature (1023 K); the second is the creation of an unusual deep level (called Y_0) at dislocations in epitaxial layers of ZnSe; the third is the observation of colored dislocations in natural diamond. Each is believed to be the result of impurity segregation to the dislocation, and none is well understood at present. The study of luminescence effects arising from impurity segregation to dislocations is in its infancy. Nevertheless, it offers a very attractive means of examining the gettering and diffusion effects associated with dislocations which have been realized in general terms for a long time, but which until now have largely defeated the attempts made to obtain detailed experimental information about them.

5.2 Quenching of Excitonic Emission at Dislocations in InP

The spectrum shown in Fig. 2(a) is typical for thin samples of good quality InP at the temperature of acquisition. The peak labeled D^0X is excitonic in character and comes mainly from donor-bound excitons. The slightly lower energy peak labeled eA^0 is in reality a mixture of DAP and free-to-bound emission at this temperature. It has a longitudinal-optical phonon replica displaced about 40 meV on the low-energy side. The relative heights of the exciton and eA^0 peaks are affected by the specimen thickness, the exciton peak being more strongly quenched as the specimen thickness drops below 1 μm. Figure 2(b) shows a spectrum acquired with the electron beam centered on the threading dislocation illustrated in Fig. 3. As proof of the apparent exciton quenching, monochromatic images were formed in each of the major peaks D^0X, eA^0, and the deep level C. Dramatically different spatial distributions of intensity were obtained (Fig. 4). It should be noted that the black spot on the exciton map at the position of the dislocation is not related to dislocation black spots often observed in SEM cathodoluminescence experiments (normally on GaAs). The latter results are obtained in integrated cathodoluminescence at room temperature and are also very much broader than the spot in Fig. 4. When the experiment was repeated on similar LEC InP which had not undergone the high-temperature heat treatment, similar behavior was not observed, and there was no dramatic change of intensity of any peak near the dislocations.

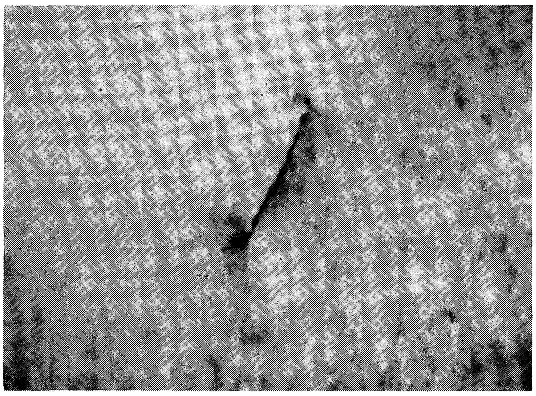

FIG. 3. The dislocation under investigation in Figs. 2 and 4; its length is approximately 0.5 μm.

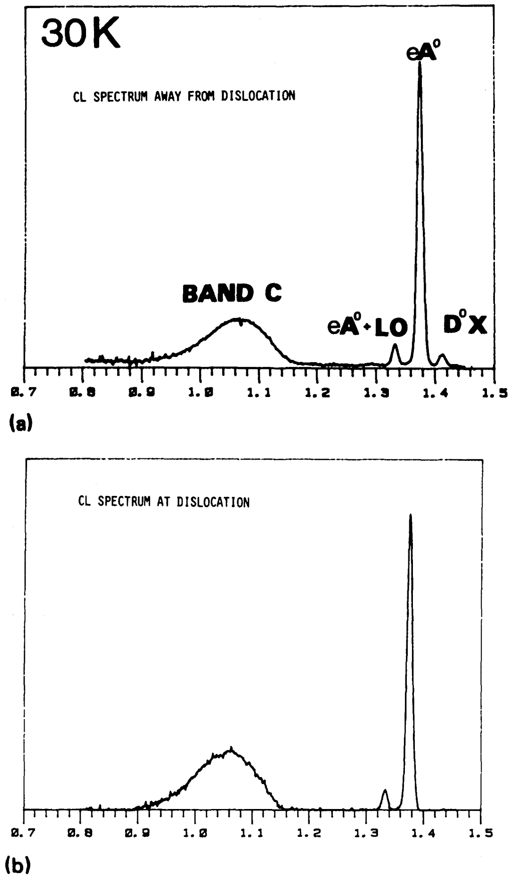

FIG. 2. Cathodoluminescence of InP with the electron beam directed at (b) a grown-in dislocation in an electron transparent region and (a) an adjoining dislocation-free region. The horizontal scale is measured in eV. Note the absence of a D^0X peak in the lower curve. [Reproduced from Myhajlenko *et al.* (1984) by courtesy of the Institute of Physics.]

5.3 Y^0 Emission from Dislocations in ZnSe

The second observation concerns dislocation emission in ZnSe. Thick (2–5 μm) epilayers were grown by metal organic chemical vapor deposition (MOCVD) on GaAs substrates. A typical low-temperature spectrum from a ~1-μm-thick specimen studied by cathodoluminescence in a STEM is shown in Fig. 5. As the spectrum is rich in detail the various features of the spectrum will first be described. The complex near-edge spectrum (2.65–2.8 eV) can be interpreted as coming from four distinct processes. The highest energy peak, labeled D^0X, comes from donor-

(chlorine) bound excitons. For regions of the specimen with high chlorine concentrations, excitons tunnel from one donor to another and undergo inelastic scattering, giving rise to the second highest energy peak [FE-(1*s*-2*s*)]. In terms of the discussion of phonon coupling by Eq. (15) in Sec. 3, the bound-exciton emission in this figure has a very small *S* value, the TES emission has a slightly higher value, while the FE-(1*s*-2*s*) emission has a still higher *S* value, but considerably less than that of the copper emission in Fig. 1.

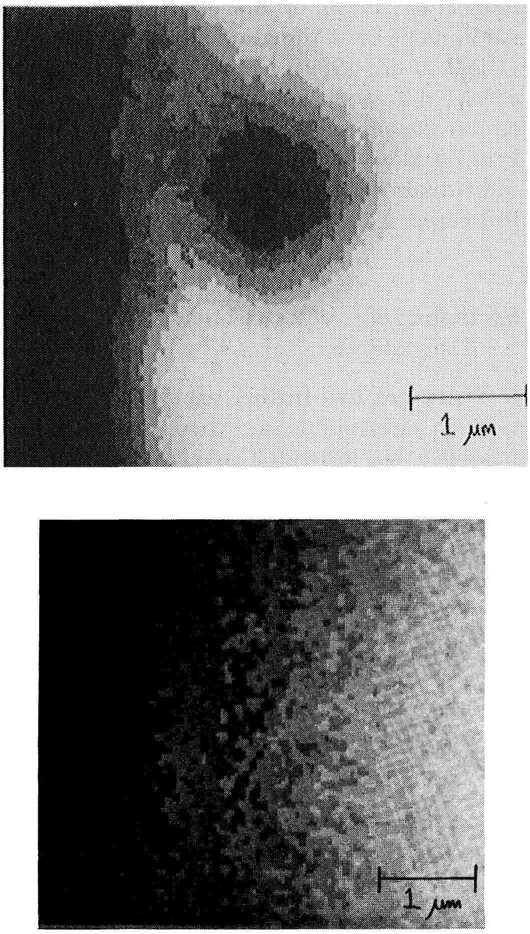

FIG. 4. Monochromatic cathodoluminescence images centered on the dislocation in Fig. 3. The upper image was formed with D^0X emission, and the black circular region is centered on the dislocation. The lower image was formed in the eA^0 peak of Fig. 2. The increase of intensity from left to right of the images is a consequence of the increase in thickness of a wedge-shaped specimen. [Reproduced from Myhajlenko *et al.* (1984) by courtesy of the Institute of Physics.]

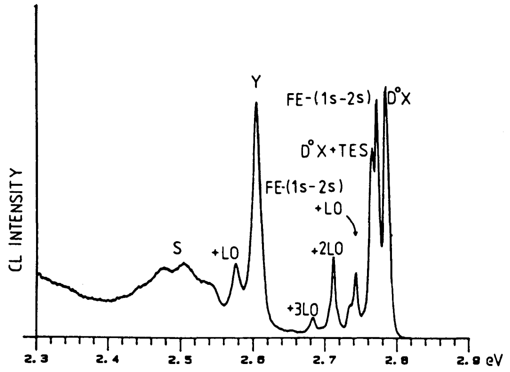

FIG. 5. Cathodoluminescence spectrum of a ~1-µm-thick undoped epitaxial layer of ZnSe. An explanation is given in the text for most of the peaks. (Reproduced by courtesy of Dr. J. L. Batstone.)

The monochromatic maps formed in these two highest energy peaks sometimes reveal marked anticorrelations, the one being bright where the other is dark and vice versa, as would be expected since they represent extremes of chlorine concentration variation. The third peak, which appears as a low-energy shoulder to the FE-(1*s*-2*s*) peak, is a two-electron satellite of excitons bound to neutral donors. It has the same spatial distribution as D^0X. There follows a series of longitudinal-optical phonon replicas of the FE-(1*s*-2*s*) and TES transitions of ever lower-energy emission as the number of phonons increases. The excitons are quite strongly bound in ZnSe (Table 2) and hence quite compact, so that the phonon coupling is stronger than in GaAs or InP (Fig. 2) but still relatively weak compared with DAP emission (small Huang-Rhys factor). However, there are evidently some anomalies in the series of replicas. Structure in the first phonon replica is the result of separate contributions from the FE-(1*s*-2*s*) and the TES processes. However, the TES emission is relatively weakly phonon coupled, and its contribution to the second phonon replica is not detectable. The second phonon replica from the TES peak has a second contribution from an independent process, as is indicated by its anomalous height and its peculiar shape. This conclusion was unambiguously confirmed by forming monochromatic maps in the TES peak and each of the three phonon replicas. The four images show very similar spatial distributions of intensity, but the images formed in the second and third phonon replicas show additional intense sources of light not present in the two higher energy images. On examining the regions from which this intense emission occurred by transmission electron microscopy small precipitates were observed, sometimes with associated punched-out loops. The origin of this emission is not yet clear.

The final peak in this spectrum to be discussed is that labeled Y^0 in Fig. 5. This peak varied considerably in intensity from one sample to another, suggesting impurity involvement, and has some very surprising properties (Dean, 1984). Although a deep level, the peak is both very narrow and only relatively weakly phonon coupled in comparison with what might have been expected. The complex S in Fig. 5 shows more characteristic deep-level behavior. Further confirmation of the excitonic nature of the Y^0 emission was obtained from temperature variation experiments which revealed the same emission energy dependence as that of the near-edge excitonic lines (Batstone and Steeds, 1986). Dean (1984) concluded that the emission was associated with dislocations in ZnSe, and cathodoluminescence experiments conclusively established this fact (Myhajlenko *et al.*, 1984). However, there remains a good measure of uncertainty about the details of the emission. On the one hand, there is evidence of impurity involvement with the dislocations; on the other hand, much of the detailed Y^0 behavior fits remarkably well the dislocation exciton model of Bozhokin *et al.* (1982) and developed by Lelikov *et al.* (1989) for dislocations in indirect-gap materials.

5.4 Colored Dislocations in Natural Diamond

The third example, that of cathodoluminescence topography of dislocations in natural diamond, is not quite so recent (Hanley *et al.*, 1977). Nevertheless, it effectively illustrates what can be achieved in a relatively simple fashion and leads to another story about possible impurity involvement with dislocation luminescence in a much wider band-gap material. In this case thick specimens were studied at room temperature in an SEM operated at 30 kV, with electron penetration of about 10 μm. Both blue-violet and green-yellow dislocations were observed, as illustrated in Plate 1.

5.5 Dislocation-Related Luminescence in Si

Each of the above studies has depended on the high spatial resolution easily achievable by cathodoluminescence. A related issue, hotly pursued at present, is that of the dislocation-related luminescence observed in silicon (Sauer *et al.*, 1985). According to different points of view, some of the emission lines are the consequence of dislocation excitons (Lelikov *et al.*, 1989), or alternatively from transition-metal contamination of the dislocations (Higgs *et al.*, 1989). Unfortunately, the long carrier diffusion distance and the high dislocation densities in the samples which have been studied prevents the examination of individual well-characterized dislocations in this case.

5.6 Dislocation Velocity and Misfit Dislocations

There are two further aspects of dislocations in semiconductors where cathodoluminescence has provided invaluable information. One of these is the measurement of dislocation velocities in bulk GaAs at high temperatures by the use of cathodoluminescence in an SEM to reveal the dislocations (Maeda *et al.*, 1983). The nondestructive nature of this method makes it very attractive for this purpose. The second is the use of cathodoluminescence to reveal misfit dislocations at heterostructure interfaces in epilayers (Petroff *et al.*, 1980). There is considerable current interest in the critical thicknesses of strained layers which result in the introduction of misfit dislocations. Use of cathodoluminescence imaging not only provides a convenient nondestructive method for observing misfit dislocations at buried interfaces (e.g., Watanabe *et al.*, 1989; Fitzgerald *et al.*, 1989), but by using monochromatic imaging it is possible to detect misfit dislocations at different levels in a complex multiple-interface structure independently of each other, if different regions of the structure emit in different parts of the spectrum (Steeds and Bailey, 1987).

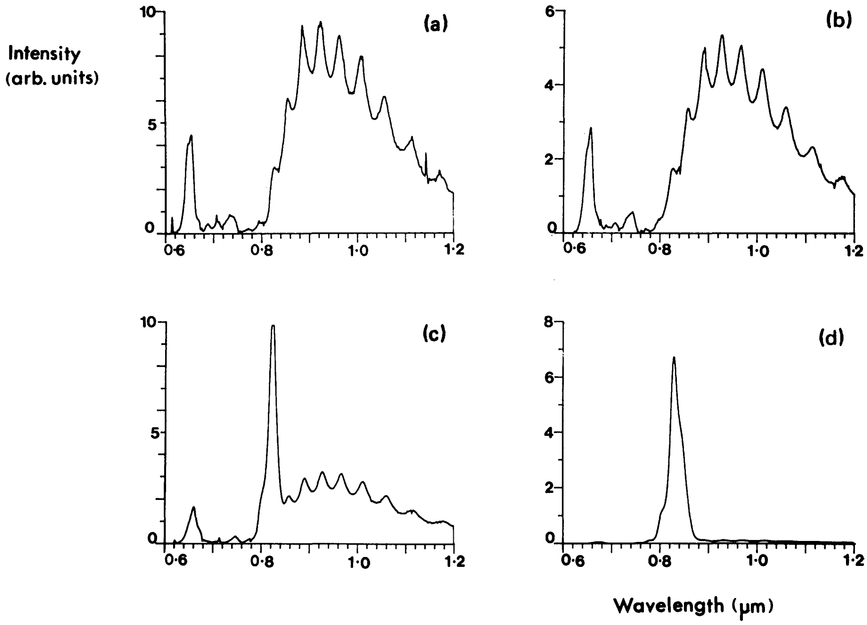

FIG. 6. Cathodoluminescence spectra obtained from a plan-view GaAs/AlGaAs double-heterostructure laser at increasing electron-beam pumping levels in the sequence (a)–(d). The arbitrary units of the vertical scales differ in each separate part of the figure. [From Steeds (1989), reproduced by courtesy of Les Editions de Physique.]

5.7 Abruptness of Semiconductor Quantum-Well Interfaces

Another problem of considerable topical interest is the flatness of interfaces grown by MBE. Cathodoluminescence imaging provides a convenient approach to investigation of this problem if single quantum-well specimens are available with sufficiently flat interfaces to be consistent with the spatial resolution available. Most of the experiments have been performed on GaAs single quantum wells with AlGaAs barriers. Layer growth is followed by reflection high-energy electron-diffraction (RHEED) oscillations, and on completion of growth of both GaAs and AlGaAs layers a pause of 2 min or so is introduced so as to reduce the interface roughening (Fukunaga and Nakashima, 1986). Growth-interrupted samples showed fine structure in the single quantum-well emission spectra. Each substructure peak was interpreted as coming from regions with flat interfaces and extending distances which were large compared with the two-dimensional exciton size. Adjoining peaks were believed to come from regions one monolayer different in thickness. The relative heights of the peaks from a single quantum well were found to vary with time on the picosecond scale. Short times gave stronger emission from the narrower portions of the well, but this decreased at longer times as the excitation diffused into broader regions (Tu *et al.*, 1988). Monochromatic cathodoluminescence images formed in the substructure peaks from a particular well were found to reveal intensity variations on the 2–3-μm scale, and these were interpreted as island sizes over which the interfaces were flat and uniformly spaced (Petroff *et al.*, 1987; Bimberg *et al.*, 1987). Later experiments by Bimberg *et al.* (1988) showed the dependence of the scale of the variations on the growth temperature and time, and spectrally resolved cathodoluminescence images revealed the diffusion of excitation across the islands. Other work on this topic produced monochromatic images with a high degree of anticorrelation as would be required for regions of different thickness (Steeds, 1989a, reporting experiments of Wang). However, these results probably cannot be understood in quite such simple terms (Ourmazd and Warwick, 1989; Warwick, 1989). The island sizes deduced by the cathodoluminescence imaging are not consistent with the considerably smaller extent of monolayer flat regions measured directly by the high-resolution imaging mode of transmission electron microscopy. Moreover, measurements at Bell Laboratories on the same samples as those used by Petroff *et al.* (1987) indicate that the substructure peaks vary considerably in energy separation from one position to another across a wafer. It has to be concluded that the large-scale structures of monochromatic cathodoluminescence images represent regions of different average well width, and that these average values will not necessarily correspond in integral numbers of monolayers in the quantum well.

5.8 Electron-Beam Pumping of Lasers

There is some interest in electron-beam pumping of lasers (Potts *et al.*, 1988; Colak *et al.*, 1987). Thin samples of heteroepitaxial layers may be prepared for STEM cathodoluminescence examination by selective removal of the substrate by use of an appropriate etchant. The samples are then studied in plan view with the incident electron beam penetrating through the epilayers. On account of the large refractive index difference at the top and bottom surfaces of the sample, strong internal reflection occurs, and this can be used to induce local superlinearities of emission and to investigate local changes of behavior, relevant to lasing properties. An example is shown in Fig. 6. The spectra were obtained from a conventional GaAs/AlGaAs double heterostructure laser prepared for STEM examination as described and pumped in a transverse rather than a longitudinal mode (Steeds, 1989a, reporting work of Myhajlenko in Bristol). The GaAs, prepared by MOCVD growth, revealed rather a broad emission spectrum crossed by a succession of thin-film interference fringes in the region from 800 to 1200 nm. With increasing pumping current in the sequence (a)–(d) the interference fringes shifted as the refractive index of the GaAs changed with increasing carrier density. From Fig. 7, which presents these results graphically, the AlGaAs emission at about 650 nm increases linearly with the pumping current, but the GaAs emission, initially linear [(a),(b)], becomes strongly superlinear for higher levels of pumping [(c),(d)].

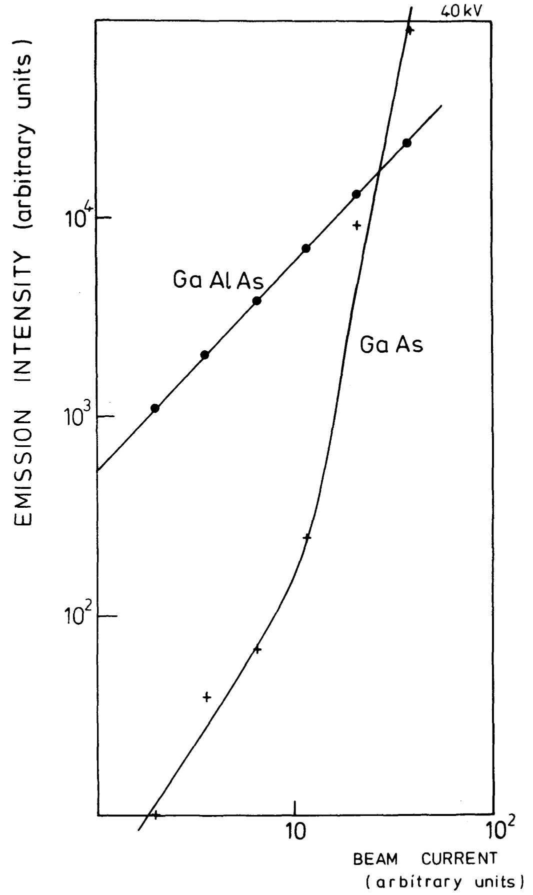

FIG. 7. Dependence of the emission intensity of the AlGaAs and GaAs peaks in Fig. 6 as a function of the electron-beam current [from Steeds (1989), reproduced by courtesy of Les Editions de Physique].

5.9 Investigation of Radiation Damage

The final example of recent applications of cathodoluminescence to semiconductors concerns near-threshold investigation of displacement damage in silicon. A modern 300-kV transmission electron microscope is probably the ideal means of undertaking such studies once the possibility of an accompanying ion beam is totally eradicated. The source has a well-defined energy (down to a few eV), has a continuously variable accelerating energy, and can be directed at any chosen region of a specimen, with a wide range of beam currents, with the specimen in any chosen crystallographic orientation, and at any temperature in the range 20–1300 K. It has recently been shown that silicon implanted with car-

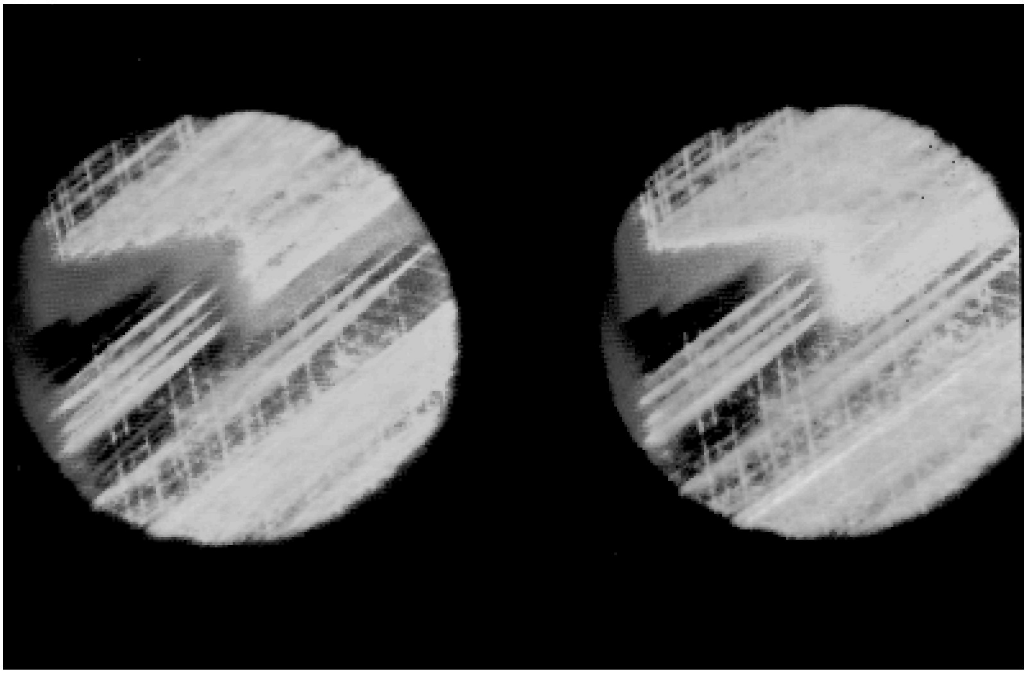

Vol. 3 – PLATE 1. Cathodoluminescence topographs showing two types of visible luminescence from dislocations in natural diamond. The dislocations become visible locally in diamond regions where nitrogen impurity, which produces blue luminescence from the diamond matrix, has a low concentration. Individual violet-emitting dislocations are juxtaposed among the yellow-green slip bands. The topographs were photographed through a polarization filter which was oriented in the left image so as to reduce the intensity of the principal violet-emitting lines relative to the unpolarized emission from the yellow-green slip bands. The diameter of the field is approximately 0.3 mm with an electron-beam energy of 35 kV. (Unpublished work of Professor A. R. Lang, FRS, published by permission.)

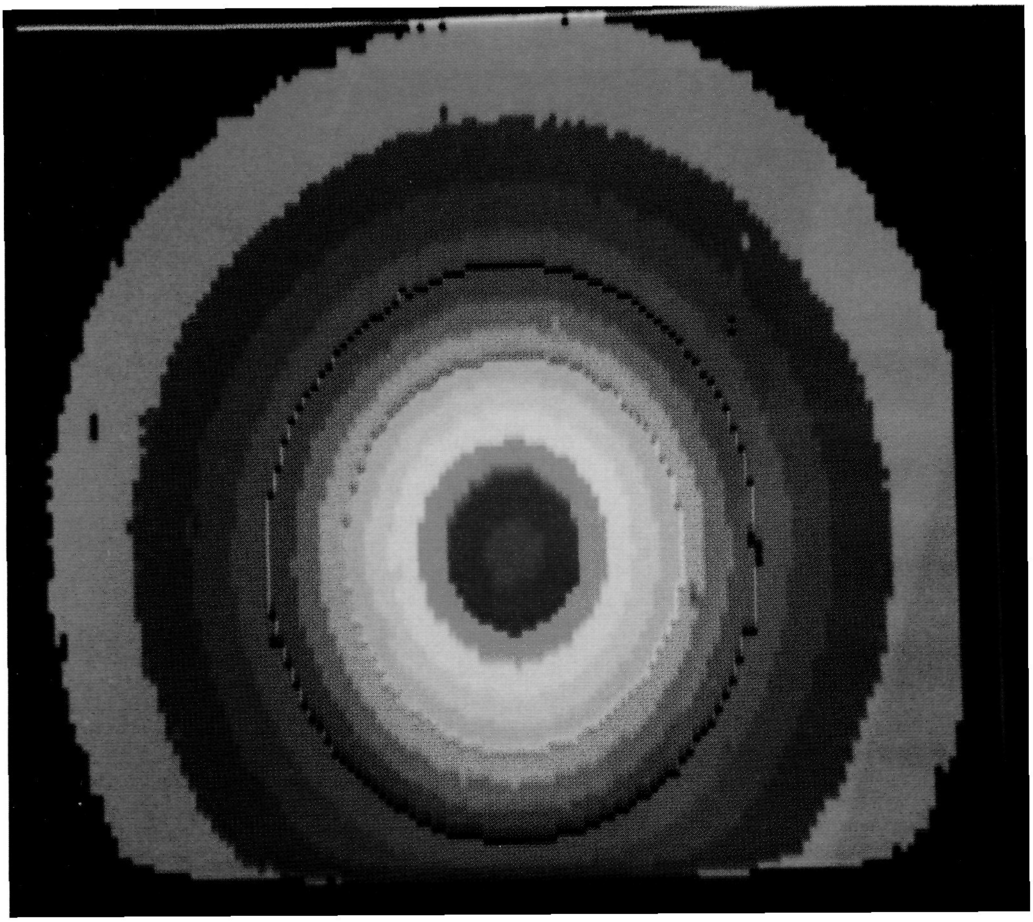

Vol. 3 – PLATE 2. False color representation of the variation of intensity of the G emission line with position in the silicon sample which was locally irradiated. Red indicates high intensity (in the irradiated region) and black low intensity with shorter wavelengths in false color representing lower intensities and long wavelengths higher intensities. (Unpublished work of F. Johnson, reproduced by permission.)

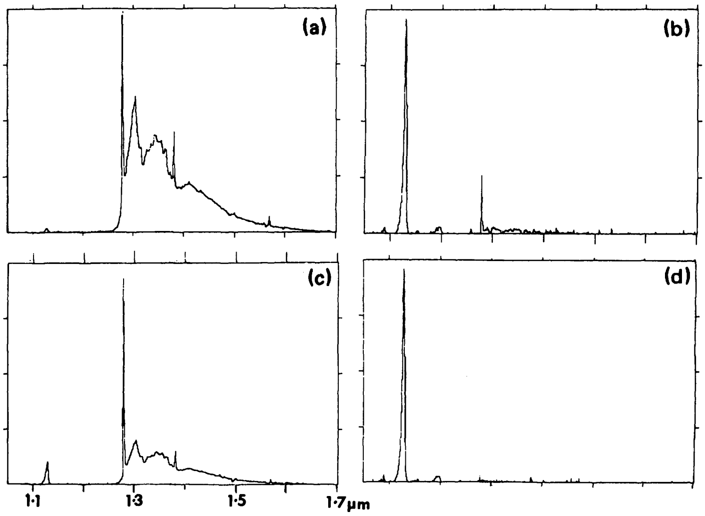

FIG. 8. Variation of cathodoluminescence spectra from electron-irradiated silicon as a function of distance from the irradiated region. The emission from the irradiated region is dominated by the *G* emission spectrum with a sharp peak near 1.3 μm (a). At a distance of about 250 μm (d) the spectrum is characteristic of free excitons in silicon, and the strongest peak near 1.1 μm is the TO phonon replica. (Unpublished work of F. Johnson, reproduced by permission.)

bon can be converted into efficient light-emitting diodes (emitting at 1.3 μm) by a process involving electron irradiation (Canham *et al.*, 1987). The *G* center involved in the emission is now well understood, as mentioned earlier (Sec. 4). We used our 300-kV TEM to irradiate float-zone silicon, containing carbon as an impurity, in the vicinity of the displacement threshold at about 100 kV. Figure 8 shows the variation of cathodoluminescence spectra with distance from a 15-μm-diam electron-irradiated area of 4000-Ω cm float-zone silicon containing $\sim 10^{16}$ cm^{-3} of carbon. The *G* center with strong emission near 1.3 μm was still detectable when the electron-beam-exciting cathodoluminescence was incident about 250 μm away from the irradiated region [Fig. 8(d)] because of long carrier diffusion lengths in silicon. However, the spectrum was then dominated by the characteristic free-exciton emission center with a transverse-optical coupled peak at 1.130 μm. The STEM cathodoluminescence system was used to produce false color images representing the intensity variation of the *G* center emission with distance of the incident beam from the electron-irradiated region (Plate 2).

GLOSSARY

A^0X, D^0X, and X: Conventional descriptions of excitonic (X) emission processes. A^0 represents a neutral acceptor, D^0 a neutral donor. A^0X (D^0X) indicates photon emission from an exciton bound to a neutral acceptor (donor).

Activator: An element added to a phosphor to enhance its light-emission properties.

Auger Transition: A transition in which an excited electronic state returns to its ground state by imparting its energy to a second electron which is emitted.

Band-Edge Luminescence (Edge Emission): The light emission from an excited semiconductor of photons with energy close to that of the band-gap energy.

Carriers: A collective name for mobile charged particles which may be electrons or holes or both.

Deep-Level (Deep Center) Emission: Emission from crystal defects which create energy levels within the band gap, well separated from either the conduction or valence-band edges. Some deep levels emit light when excited, but many do not. They cannot be treated by hydrogenic theory.

Edge Emission: Same as band-edge luminescence.

Excitons: Bound electron-hole pairs. They may vary in size from atomic dimensions to rather large, weakly bound pairs with radii of some tens of atomic spacings which can be treated, to a reasonable approximation, by Bohr theory.

Excitonic Droplets: A low-temperature condensation of large numbers of individual excitons into liquidlike objects within an indirect-gap semiconductor.

Free to Bound (F-B): The conventional description of the emission line resulting from the annihilation of a free electron (hole) and a bound hole (electron). eA^0 is the particular F-B transition where a free electron combines with the hole bound to a neutral acceptor (A^0).

FE-(1*s*-2*s*): An emission line associated with the transition of a free exciton from a 2*s* (or 2*p*) to a 1*s* state.

Interstitial Sites: Positions in a crystal structure which lie between the sites occupied by the atoms which comprise the ideal structure.

Misfit Dislocations: Dislocations at the interface between two dissimilar materials with different lattice constants which help to accommodate the mismatch between the atomic structures.

Multiple-Bound-Exciton Complexes: Just as atoms can sometimes bind together to form diatomic, triatomic, etc., molecules, so can excitons bind together to form complexes which contain two, three, four, or more excitons.

Passivation: Surface treatments which reduce the nonradiative surface recombination of carriers.

Phonon: A mode of distortion of the lattice during thermal vibration.

Photoluminescence: Luminescence excited by photons (light) which are sufficiently energetic to excite optical centers in a crystal or to lift electrons out of the valence band into the conduction band.

Plasmon: A quantized oscillation of the conduction-electron gas relative to the positive ion lattice.

Polaritons: Particlelike entities caused by the mixing between photons and excitons of similar wave number in a semiconductor.

Quantum Well: A thin planar two-dimensional layer of a semiconductor embedded in a semiconductor of larger band gap. The thickness of the embedded layer is less than the electron coherence length.

Sensitizer: An element added to a phosphor to enhance its ability to absorb incident radiation and transfer it to an emitting center.

Substitutional Sites: Sites in a crystal structure normally occupied by atoms of the host material but capable of replacement by impurities or dopant atoms.

List of Works Cited

Bailey, S. J., Steeds, J. W. (1989a), *Proc. Mater. Res. Soc.* **138**, 503.

Bailey, S. J., Steeds, J. W. (1989b), in: Cullis and Hutchison (Eds.), *Microscopy of Semiconducting Materials 1989*, Inst. Phys. Conf. Ser. No. 100, Bristol, England: Institute of Physics, pp. 777–782.

Batstone, J. L., Steeds, J. W. (1986), Proceedings of the 44th Annual Meeting of the Electron Microscopy Society of America (unpublished), pp. 818–819.

Berger, S. D., Brown, L. M. (1984), in: Doig (Ed.), *Electron Microscopy and Analysis 1983*, Inst. Phys. Conf. Ser. No. 68, Bristol, England: Institute of Physics, pp. 115–118.

Bimberg, D., Christen, J., Fukunaga, T., Nakashima, H., Mars, D. E., Miller, J. N. (1987), *J. Vac. Sci. Technol. B* **5**, 1191–1197.

Bimberg, D., Christen, J., Fukunaga, T., Nakashima, H., Mars, D. E., Miller, J. N. (1988), *Superlattices Microstruct.* **4**, 257–263.

Bimberg, D., Munzel, H., Steckenborn, A., Christen, J. (1985), *Phys. Rev. B* **31**, 7788–7799.

Blasse, G., Bril, A. (1970), *Philips Tech. Rev.* **31**, 304–332.

Bozhokin, S. V., Parshin, D. A., Kharchenko, V. A. (1982), *Sov. Phys. Solid State* **24**, 800–803.

Brillson, L. J., Viturro, R. E., Shaw, J. L., Richter, H. W. (1988), *J. Vac. Sci. Technol. A* **6**, 1437–1445.

Canham, L. T., Barraclough, K. B., Robbins, D. (1987), *Appl. Phys. Lett.* **51**, 1509–1511.

Chappell, S. P., Claybourn, M., Newman, R. C., Barraclough, R. G. (1988), *Semicond. Sci. Techol.* **3**, 1047–1051.

Chin, A. K., Caruso, R., Young, M. S. S., von Neida, A. R. (1984), *Appl. Phys. Lett.* **45**, 552–554.

Colak, S., Khurgin, J., Seemungel, W., Hebling, A. (1987), *J. Appl. Phys.* **62**, 2633–2639.

Chu, Y. M., Darby, D. B., Booker, G. R. (1981), in: A. G. Cullis, D. C. Joy (Eds.), *Microscopy of Semiconducting Materials 1981*, Inst. Phys. Conf. Ser. No. 60, Bristol, England: Institute of Physics, pp. 331–338.

Davies, G., Lightowlers, E. C., Newman, R. C., Oates, A. S. (1987), *Semicond. Sci. Techol.* **2**, 524–532.

Davies, G., Brian, H., Lightowlers, E. C., Barraclough, K., Thomaz, M. F. (1989), *Semicond. Sci. Technol.* **4**, 200–206.

Dean, P. J. (1984), *Phys. Status Solidi A* **81**, 625–646.

Dean, P. J., Venghaus, H., Pfister, J. C., Schaub, B., Marine, J. (1978), *J. Lumin.* **16**, 363–394.

Everhart, T. E., Hoff, P. H. (1971), *J. Appl. Phys.* **42**, 5837–5846.

Fitzgerald, E. A., Watson, G. P., Proano, R. E., Ast, D. G., Kirchner, P. D., Pettit, G. D., Woodall, J. M. (1989), *J. Appl. Phys.* **65**, 2220–2237.

Fukunaga, T., Nakashima, H. (1986), *Jpn. J.*

Appl. Phys. **25**, L856–858.

Garlick, G. F. J. (1958), in: S. Flügge (Ed.), *Encyclopedia of Physics*, Vol. 26, *Light and Matter II*, Berlin: Springer, p. 1–128.

Hagni, R. D., Karakus, M. (1989), *MRS Bull.* **14**, 54–59.

Hanley, P. L., Kiflawi, I., Lang, A. R. (1977), *Phil. Trans. R. Soc. London, Ser. A* **284**, 329–368.

Henderson, B., Imbusch, G. F. (1989), *Optical Spectroscopy of Inorganic Solids*, Oxford: Clarendon.

Henshaw, D. L. (1982), *Phys. Technol.* **13**, 266–273.

Higgs, V., Lightowlers, E. C., Davies, G., Schaffler, F., Kasper, E. (1989), *Semicond. Sci. Technol.* **4**, 593–598.

Kanaya, K., Okayama, S. (1972), *J. Phys. D* **5**, 43–58.

Lang, A. R. (1980), in: M. Schlenker *et al.* (Ed.), *Imaging Processes and Coherence in Physics*, Vol. 112 of Lecture Notes in Physics, Berlin: Springer-Verlag, pp. 533–544.

Lax, M., Hopfield, J. J. (1961), *Phys. Rev.* **124**, 115–123.

Lelikov, Yu. S., Rebane, Yu. T., Shreter, Yu. G. (1989), in: Roberts (Ed.), *Structure and Properties of Dislocations in Semiconductors 1989*, Inst. Phys. Conf. Ser. No. 104, Bristol, England: Institute of Physics, pp. 119–129.

Madelung, O. (Ed.) (1982), *Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology, New Series*, Group 3, Vol. 17a. *Physics of Group IV Elements and III-V Compounds*, Berlin: Springer-Verlag.

Maeda, K., Sato, M., Kubo, A., Takeuchi, S. (1983), *J. Appl. Phys.* **54**, 161–168.

Myhajlenko, S. (1984), in: Doig (Ed.), *Electron Microscopy and Analysis 1983*, Inst. Phys. Conf. Ser. No. 68, Bristol, England: Institute of Physics, pp. 111–114.

Myhajlenko, S., Batstone, J. L., Hutchinson, H. J., Steeds, J. W. (1984), *J. Phys. C* **17**, 6477-6492.

Ourmazd, A., Warwick, C. A. (1989), private communication.

Pennycook, S. J. (1982), in: Gorringe (Ed.), *Electron Microscopy and Analysis 1981*, Inst. Phys. Conf. Ser. No. 61, Bristol, England: Institute of Physics, pp. 55–58.

Petroff, P. M., Cibert, J., Gossard, A. C., Dolan, D. J., Tu, C. W. (1987), *J. Vac. Sci. Technol. B* **5**, 1204–1208.

Petroff, P. M., Lang, D. V., Strudel, J. L., Logan, R. A. (1978), in: O. Johari (Ed.), *Scanning Electron Microscopy*, Vol. 1. Chicago, IL: SEM Inc., pp. 325–332.

Petroff, P. M., Logan, R. A., Savage, A. (1980), *J. Microsc.* **118**, 255–261.

Potts, J. E., Smith, T. L., Cheng, H., Yang, B., Wessels, B. W. (1988), *J. Cryst. Growth* **86**, 935–941.

Reimer, L. (1984), *Transmission Electron Microscopy*, Berlin: Springer-Verlag.

Roberts, S. H. (1981), in: A. G. Cullis, D. C. Joy (Eds.), *Microscopy of Semiconducting Materials 1981*, Inst. Phys. Conf. Ser. No. 60, Bristol, England: Institute of Physics, pp. 377–380.

Sauer, R., Weber, J., Stolz, J., Weber, E. R., Kusters, K. H., Alexander, H. (1985), *Appl. Phys.* **36**, 1–13.

Smith, D. L., McGill, T. C. (1976), *Phys. Rev. B* **14**, 2448–2455.

Steeds, J. W. (1989a), *Rev. Phys. Appl. Colloq.* C6, Suppl. 6, 65–72.

Steeds, J. W. (1989b), in: Roberts (Ed.), *Structure and Properties of Dislocations in Semiconductors 1989*, Inst. Phys. Conf. Ser. No. 104, Bristol, England: Institute of Physics, pp. 199–202.

Steeds, J. W., Bailey, S. J. (1987), in: P. G. Merli (Ed.), Proceedings of the XVI Congress on Electron Microscopy, Bologna, 1987 (unpublished), pp. 107–110.

Tu, C. W., Miller, R. C., Petroff, P. M., Kopf, R. F., Deveaud, B., Damen, T. C., Shah, J. (1988), *J. Vac. Sci. Technol. B* **6**, 610–612.

Warwick, C. A. (1989), in: Goodhew and Elder (Eds.), *EMAG—MICRO 89*, Inst. Phys. Conf. Ser. No. 98, Bristol, England: Institute of Physics, pp. 447–450.

Warwick, C. A., Booker, G. R. (1983), in: A. G. Cullis (Ed.), *Microscopy of Semiconducting Materials 1983*, Inst. Phys. Conf. Ser. No. 67, Bristol, England: Institute of Physics, pp. 321–326.

Watanabe, Y., Kadota, Y., Okamoto, H., Ohmachi, Y. (1989), *Jpn. J. Appl. Phys.* **28**, 16–20.

Yacobi, B. G., Holt, D. B. (1986), *J. Appl. Phys.* **59**, R1–23.

Further Reading

Bastard, G., Delalande, C., Guldner, Y., Voisin, P. (1988), *Adv. Electron. Electron Phys.* **72**, 1. (PL, semiconductors.)

Brocker, W., Pfefferkorn, G. (1979), *Scanning Electron Microscopy II*, p. 125. (Biology and medicine; CL, other materials.)

Chemla, D. S. (1985), *J. Lumin.* **30**, 502. (PL, semiconductors.)

Dean, P. J. (1982), *Prog. Cryst. Growth Charact.* **5**, 89. (PL, semiconductors.)

Dean, P. J., Herbert, D. C. (1979), in: K. Cho (Ed.), *Excitons*, Vol. 14 of Topics in Current Physics, Berlin: Springer-Verlag, Chap. 3. (Excitons.)

Dexter, D. L., Knox, R. S. (1965), *Excitons*, New York and London: Interscience. (Excitons.)

Hagni, R. D., Karakus, M. (1989), *MRS Bull.* **14**, 54. (Ceramics; CL, other materials.)

Henderson, B., Imbusch, G. F. (1989), *Optical Spectroscopy of Inorganic Solids*, Oxford: Clarendon. (PL, insulators.)

Kearsley, A., Wright, P. (1988), *Microscopy and Analysis*, No. 7. (CL, other materials.)

Klick, C. C., Schulman, J. H. (1957), in: F. Seitz and D. Turnbull (Eds.), *Solid State Physics*, Vol. 5, New York: Academic, p. 97. (Configurational-coordinate model.)

Knox, R. S., (1963), in: F. Seitz and D. Turnbull (Eds.), *Solid State Physics*, Suppl. 5, New York and London: Academic. (Excitons.)

Lohnert, K., Kubalek, E. (1983), in: A. G. Cullis (Ed.), *Microscopy of Semiconducting Materials 1983*, Inst. Phys. Conf. Ser. No. 67, Bristol, England: Institute of Physics. (CL, semiconductors.)

Lumb, M. D. (1978), in: M. D. Lumb (Ed.), *Luminescence Spectroscopy*, London and New York: Academic, Chap. 2. (PL, organics.)

Markham, J. J. (1959), *Rev. Mod. Phys.* **31**, 956. (Configurational-coordinate model.)

Marshall, D. J. (1988), *Cathodoluminescence of Geological Minerals*, Boston, MA: Unwin Hyman. (CL, other materials.)

Pankove, J. I. (1971), *Optical Processes in Semiconductors*, New York: Prentice-Hall. (PL, semiconductors.)

Reynolds, D. C., Collins, T. C. (1981), *Excitons: Their Properties and Uses*, New York and London: Academic. (Excitons.)

Schiller, C., Boulon, M. (1975), *Philips Tech. Rev.* **35**, 239–246. (CL, semiconductors.)

Spivak, G. V., Petrov, V. I., Antoshin, M. K. (1986) *Sov. Phys. Usp.* **29**, 364–380. (CL, semiconductors.)

Voos, M., Leheny, R. F., Shah, J. (1982), in: M. Balkansk (Ed.), *Handbook on Semiconductors*, Vol. 2, Chap. 6. (PL, semiconductors.)

Yacobi, B. G., Holt, D. B. (1986), *J. Appl. Phys.* **59**, R1–R24. (CL, semiconductors.)

CCDs

See CHARGE-COUPLED DEVICES

CELLULAR BIOMECHANICS

RICHARD SKALAK, *Department of Applied Mechanics and Engineering Sciences, University of California, San Diego, La Jolla, California, U.S.A.*

Introduction 141
1. **Cell Anatomy and Constituents** 142
1.1 Cytoplasm Properties 143
1.1.1 Actin 143
1.1.2 Myosin 143
1.1.3 Microtubules 144
1.1.4 Intermediate Filaments 144
1.1.5 Nucleus 145
1.1.6 Other Constituents 145
1.2 Cell Membrane Properties 145
1.2.1 Lipid Bilayer 146
1.2.2 Intramembranous Proteins 146
1.2.3 Cytoskeletal Components 147
1.2.4 Area Elasticity 147
1.2.5 Shear Elasticity and Viscosity 148
1.2.6 Bending Elasticity 148
1.3 Physical Processes 148
1.3.1 Polymerization 148
1.3.2 Osmotic Pressure 149
1.3.3 Active Contraction 149
1.3.4 Tubulin Molecular Motors and Laser Tweezers 149
1.3.5 Adhesion and Cell Junctions 151
2. **Mechanics of Blood Cells** 151
2.1 Red Blood Cells 152
2.1.1 Membrane Properties and Cell Shape 152
2.1.2 Hemoglobin Properties 152
2.1.3 Pipette and Flow-Chamber Tests 152
2.1.4 Flickering 154
2.1.5 Capillary Flow 154
2.1.6 Rouleau Formation 155
2.2 White Blood Cells 155
2.2.1 Passive Properties: Pipette and Poking Tests 155
2.2.2 Pseudopod Formation 156
2.2.3 Capillary Flow 157
2.2.4 Sieving Tests 157
3. **Cell Motility** 157
3.1 Flagella and Cilia 157
3.1.1 Flagella Mechanics 158
3.1.2 Cilia Motions 158
3.2 Amoeba Motility 159
3.2.1 Two-Phase Models 159
3.3 Leukocytes and Fibroblasts 159
3.3.1 Sequence of Locomotion Cycle 160
3.3.2 Mechanics of Contraction 160
4. **Cytokinesis and Growth** 160
4.1 Mechanics of Cell Division 161
4.1.1 Contractile Ring Models 162
4.1.2 Chromosome Control 162
4.2 Morphogenesis and Growth 162
4.2.1 Role of Adhesion 163
4.2.2 Pattern Formation 164
Glossary 165
Works Cited 165
Further Reading 167

INTRODUCTION

Cellular biomechanics is defined herein as the analysis of the passive and active mechanical behaviors of individual living cells by the use of continuum mechanics field theories. This includes classical disciplines of solid mechanics, fluid mechanics, and viscoelasticity, but also requires additional concepts such as osmotic pressure, gelation, active contraction, and growth. All living tissues are made up of individual cells plus extracellular constituents produced by cells. These extracellular products may be the main constituents of a tissue such as bone, cartilage, or tendon, but even in these cases, the tissues are maintained by cellular activity. Growth or atrophy in response to changes of stress is mediated by cellular processes.

The identification of the cell as the ubiquitous living unit of all tissues was a concept that was established only in the 19th century.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Since then extensive microscopic and biochemical investigation has established the morphology and function of the great variety of internal structures of many different cell types. The active processes of growth and locomotion have also been observed in detail. However, the mathematical description of deformation and active motion of cells using continuum mechanics theory has been extensively developed only in the past two decades. This is due in part to the complexity of most cells, which have a variety of internal structures such as nucleus, fibrous structures, gels, vesicles, and membranes. Moreover, the properties of these structures individually are not known, and so the predictive power of analysis is limited. Nevertheless, the range of cellular deformations and active motions being elucidated by rational field theories is rapidly expanding.

Historically, the red blood cell has probably received the most attention because it is a relatively simple cell and its rheology is important to mechanics of blood flow (see BIORHEOLOGY). The molecular mechanics of muscle contraction has also been extensively studied. In recent years, many other cells such as leukocytes, endothelial cells, fibroblasts, and even neurons have been tested and described in quantitative terms. Theories of active motions and cell locomotion are less well developed than interpretations of passive mechanical properties. In many cases, the qualitative nature and source of mechanical driving force of active motions and growth have not been identified in detail. In the subsequent sections, some of the main areas of presently developed analyses are outlined.

1. CELL ANATOMY AND CONSTITUENTS

All animals and plants are formed by cells which are small semiautonomous units enclosed in a thin membrane (Fig. 1). Some microbes consist of a single cell, but higher animals contain billions of cells plus extensive extracellular components that cells produce. However, the individual cells of higher animals can live, grow, and multiply outside the body in appropriate cell culture media.

The cell membrane is a thin (5-nm) bilayer of phospholipids which separates the interior of the cell from the external world. It is one of the distinguishing features of living matter. Cells may also have internal units enclosed in membranes, such as the nucleus and other organelles. Primitive cells which do not have a separate nucleus, such as bacteria, are called procaryotes. Higher forms which have a nucleus containing the reproductive DNA and RNA molecules are called eucaryotes. In this article, most cells considered will be eucaryotes, except perhaps the red blood cell which starts out life with a nucleus but dispenses with it before joining the bloodstream.

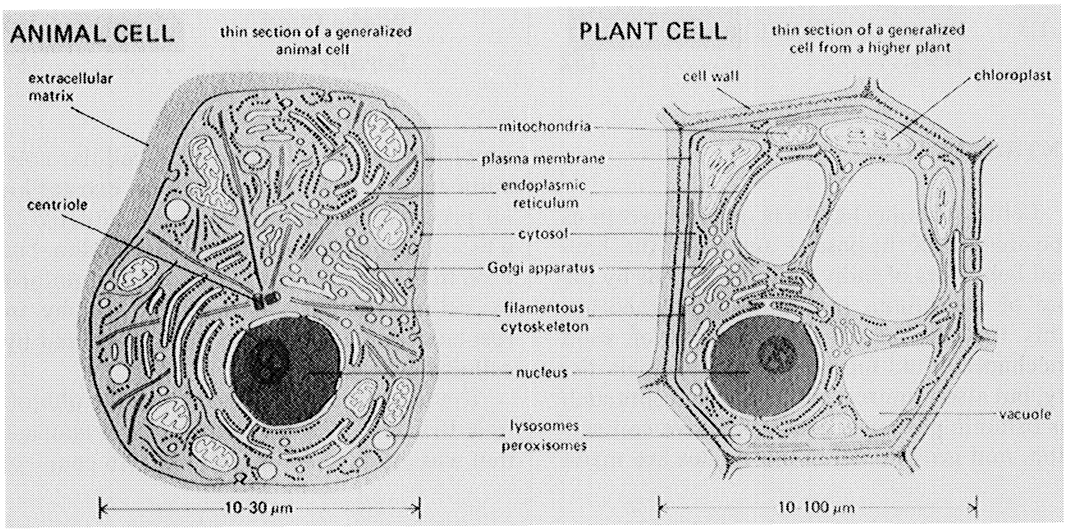
ANIMAL CELL
thin section of a generalized animal cell
extracellular matrix
centriole
mitochondria
plasma membrane
endoplasmic reticulum
cytosol
Golgi apparatus
filamentous cytoskeleton
nucleus
lysosomes
peroxisomes
10-30 μm
PLANT CELL
thin section of a generalized cell from a higher plant
cell wall
chloroplast
vacuole
10-100 μm

FIG. 1. Schematic anatomy of animal and plant cells. (From Alberts *et al.*, 1989, by permission.)

Figure 1 shows generalized sketches of animal and plant cells. Besides the nucleus, there are various organelles enclosed by membranes including mitochondria, Golgi apparatus, endoplasmic reticulum, and lysosomes. The spaces between these organelles are filled with cytosol, a complex solution containing various proteins and fibrous structures. These are mostly actin, tubulin, and intermediate filaments. A comprehensive treatment of eucaryote anatomy, biology, and biochemistry can be found in the book by Alberts *et al.* (1989).

Plant cells (Fig. 1) have a cell wall which is external to the plasma membrane and is composed of cellulose, up to 10 μm thick. Plant cells also have vacuoles which are fluid filled and act as space fillers to some extent.

1.1 Cytoplasm Properties

From the standpoint of material properties, the interior of the cell is a heterogenous composite consisting of a mixture of solid, gel, and fluid components. The overall behavior is generally viscoelastic. For some purposes, some cells can be represented as uniform viscoelastic solids. For detailed description of internal motions or local deformations, the individual components need to be recognized. The subdivision most frequently needed first is to describe the cell as a thin membrane filled with a viscoelastic fluid or solid, depending on the cell involved. Next, the cytoplasm may be represented as a two-phase material, for example, an actin gel filled with a thin fluid. For particular cells, for example, white blood cells, the nucleus needs to be recognized as a separate entity with its own membrane, forming a much stiffer unit than the cytoplasm. The degree of subdivision depends on the phenomena under consideration. The principal types of constituents are discussed in the paragraphs below.

It is also important to recognize that living tissues may undergo rapid transformations from fluid to gel, or internal stress generated by molecular interactions as in muscle contraction. These active motions are under complex and delicate biochemical, neural, and mechanical controls. This feature of living matter means that cell mechanical properties are not fixed, but the constituents may change state, form, and internal stress in response to small biochemical changes or neural impulses (see Stracher, 1983).

1.1.1 Actin. Actin is the most abundant cytoskeletal protein and forms the basis of many cytoskeletal structures such as pseudopodia, intestinal villi, hair cells of the ear, and the cortical shell, about 0.5 μm thick, that is found just under the cell membrane of many cells. Actin is also one of two principal proteins of all muscles. Actin is the so-called thin filament of skeletal muscle along which the thick filament (myosin) moves to produce muscle contraction. Actin also forms the less organized structure of smooth muscle and so-called stress fibers which occur in many cells such as leucocytes, fibroblasts, and endothelial cells.

An important aspect of actin as a cell constituent is that it occurs in two forms. As a structural element it is a fiber called F-actin and as a monomer which is easily carried in liquid suspension it is called G-actin. The polymerization of G-actin to F-actin and the inverse process are closely regulated and essential processes in many cell motions and growth. G-actin occurs in most cells at a concentration above that at which F-actin fibers would form spontaneously, but are prevented from doing so by capping proteins, mainly profilactin. When F-actin filaments form, the polymer is asymmetric in the sense that one end (called the plus end) has a higher rate of growth than the other (minus end). Assemblies of actin filaments form stress fibers and many other cell structures (Stossel, 1982). Such gels are held together by a variety of actin-binding proteins such as filamin, tropomyosin, and α-actinen. Myosin can act to contract an actin gel. This is the basis of active amoeba movements. A typical actin gel will be only 1% or 2% actin, with pores of the order of 20–100 nm which are filled with an aqueous solution of various constituents which may include a supply of G-actin for further growth of the gel. In most cells, growth of the actin cortical layer or protrusions is controlled by special proteins on the inner surface of the cell membrane. The plus ends of actin filaments terminate and grow at the membrane (see Ishikawa *et al.*, 1986).

1.1.2 Myosin. Myosin is a specialized molecule which is found to some extent in all cells and all animals. Together with actin, it forms the basis of muscle contraction. A myosin molecule is about 150 nm long, consisting of a long thin (about 2-nm) coiled section at the end of which are located two heads. These heads have an affinity to actin and can perform a rocking motion by which the myosin moves along an actin filament, towards its plus end. This action is the molecular basis of all muscle contraction and force generation. It is a process that consumes energy, or, more accurately, converts chemical energy into work. The energy comes from hydrolysis of adenosine triphosphate (ATP).

In skeletal muscle, myosin molecules are arranged in bundles which are the so-called thick filaments. The myosin heads are oriented in opposite directions on each end of the filament so that when the myosin heads move along the opposing actin filaments, a contraction results (see BIOMECHANICS OF LIVING ORGANISMS, Fig. 5). Smooth muscle has a more random arrangement of actin and myosin filaments, but it is again their interaction which produces contraction. Myosin may be assembled into filaments and more or less oriented systems as needed within a cell, such as in the formation of a contractile ring when a cell divides: an actin-myosin ring girdles a single spherical cell and, by steady contraction, divides it into two daughter cells. Myosin is also found in stress fibers which contract during cell locomotion, and in adhesion bands of epithelial cells (Alberts *et al.*, 1989).

1.1.3 Microtubules. Microtubules form the basis of a second system of structural support and cellular motion, quite different from and intermingled with the actin-myosin system. Microtubules are hollow tubes formed of tubular arrangements of the molecule tubulin. There are 13 molecular subunits around the circumference of a microtubule arranged in a very regular helical pattern. The resulting tube is a stiff, straight hollow cylinder having a diameter of about 25 nm and a typical length of 10 μm; the microtubule length may be 200 μm.

Microtubules are found in two kinds of arrangements. One forms the core of cilia and flagella. The other is a more distributed form in the cytoplasm of many cells. In many cells, microtubules are found as single tubes emanating from specialized centers called centrosomes usually near the cell nucleus. The arrangement of microtubules resembles pins in a pin cushion. In the cytoplasm, microtubules are often not stable in their length or position. Microtubules have directional structure, like actin, and have plus ends which extend by polymerization rapidly and minus ends which grow slowly. Microtubules are most often found in a state of dynamic instability in which the length of each tubule either increases or decreases. Microtubules may be capped by certain proteins which then stabilize the microtubules. Uncapped microtubules can depolymerize rapidly. Selective capping can guide the arrangement of microtubules.

The microtubules in the cytoplasm of many cells, including neural axons, form a guidance and transport system of vesicles and other organelles. Vesicles are moved along microtubules by interactions with the proteins dynein and kinesin at a rate of up to 2 μm/s. Microtubules have also been implicated in directional control of cell locomotion (Euteneuer and Schliwa, 1986). Usually a cell moves in the direction in which most microtubules are pointing. Many aspects of microtubule dynamics are discussed in Soifer (1986).

1.1.4 Intermediate Filaments. A third system of cytoskeletal elements is formed by the intermediate filaments, so named because they are intermediate between the thick and thin filaments of muscle cells. They form a connected network, especially around the nucleus of most cells, but form a stable, open network throughout the cell. This network has been described as ropelike or cablelike. It connects to most of the other structures in the cell.

Intermediate filaments are made up of four types of proteins, all of which have a common central portion. The first type is an acidic keratin, which is the principal protein in epithelial cells, hair, nails, feathers, scales, and rhinoceros horn. The second type is vimentin, which is found in muscle and other cells. The third is neurofilament, which occurs in neurons. The fourth is nuclear lamin, which surrounds the nucleus in many cells.

Intermediate filaments appear to be structural elements, i.e., they provide mechanical support of other cell structures and stiffen the cells against deformation by external forces. They are not generally formed and decomposed rapidly like actin and myosin and are not involved in contractile processes. The large number of variations of intermediate filaments appears to be associated with different cells and connections. The variety of intermediate filaments can be seen in the book edited by Wang *et al.* (1985).

1.1.5 Nucleus. In eucaryote cells, the nucleus is a separate compartment with its own plasma membrane enclosing the DNA and associated protein synthesis apparatus. The membrane of the nucleus is a double layer. Its two bilayers are connected at pores through which a traffic of proteins into and out of the nucleus is regulated. Each pore is surrounded by specialized granules that control transport through the pores. The opening of the pores is only about 9 nm.

The outer membrane of the nucleus is continuous with the endoplasmic reticulum. This is a thin space enclosed by two bilayer membranes. The nucleus has two layers of intermediate filaments associated with it, one inside and one outside the double nuclear membrane. The inner intermediate filament layer is relatively thin and compact, but the outer layer is like a loosely woven nest surrounding the nucleus (Alberts *et al.*, 1989). These structures and the closely packed DNA and associated proteins make the nucleus a much stiffer entity than the cytoplasm outside the nucleus or the cell as a whole. This is evident, for example, in leukocytes in a blood smear. The nucleus is still rounded although the rest of the cell is quite flattened.

The nuclear envelope keeps the DNA and associated proteins separated from the rest of the cell cytoplasm. DNA transcription and RNA processing goes on inside the nucleus. However, during cell division, the nuclear membrane and surrounding structures are broken down; then division takes place and two new nuclear membranes are formed in the daughter cells. During cell division, the cell is like a procaryote which does not have a separate nuclear compartment.

1.1.6 Other Constituents. Eucaryotic cells have several specialized membrane compartments in addition to the nucleus. Most of these function as sites of production of particular proteins. From the standpoint of total number and volume, mitochondria are most prominent. These ellipsoidal compartments are 2–3 μm long and are the principal site of ATP production. This is the fuel for muscles as well as many other energy-consuming reactions. From the standpoint of membrane surface, the endoplasmic reticulum is most important. It may have 10 or 20 times the surface area of the outer plasma membrane of the cell. The endoplasmic reticulum encloses only a thin, lamellar space. The outer surface of its membrane contains ribosomes which manufacture lipid for membranes and intramembranous proteins.

Another important compartment is the Golgi apparatus which is a stack of disk-shaped compartments which processes lipid and other proteins which are dispatched in membrane vesicles. There are also a number of smaller organelles and vesicles found in different cells with special functions. These include delivery of enzymes and signaling proteins.

The mechanical properties of the individual compartments described above are generally not known. In large deformations of the cell, these internal structures probably behave as more or less rigid bodies. The surrounding cytosol forms a viscoelastic fluid, and the specialized organelles may be regarded as solid particles in suspension in the cytosol. When actin gel forms or microtubular structures stabilize, a portion or all of the cell may be regarded as a viscoelastic solid. The time scale and degree of deformation will enter into the choice of an appropriate model. Various cases are discussed in subsequent sections below.

1.2 Cell Membrane Properties

All cells are enclosed by a lipid bilayer membrane which acts as a barrier to entry and exit of all components. By virtue of a variety of special proteins which float in or on the lipid bilayer, cell membranes perform many active functions in transport of ions, transducing mechanical and biochemical stimuli, generating neural signals, and initiating immune responses.

From a mechanical standpoint, the lipid bilayer itself is a two-dimensional liquid sheet. It resists any change in its surface area, but offers little resistance to deformation otherwise. In most cells, the lipid membrane is attached to the underlying cytoskeleton via intramembranous proteins which are firmly anchored in the lipid bilayer. The term cell membrane may include some layers of cytoskeleton as well as lipid membrane. This is

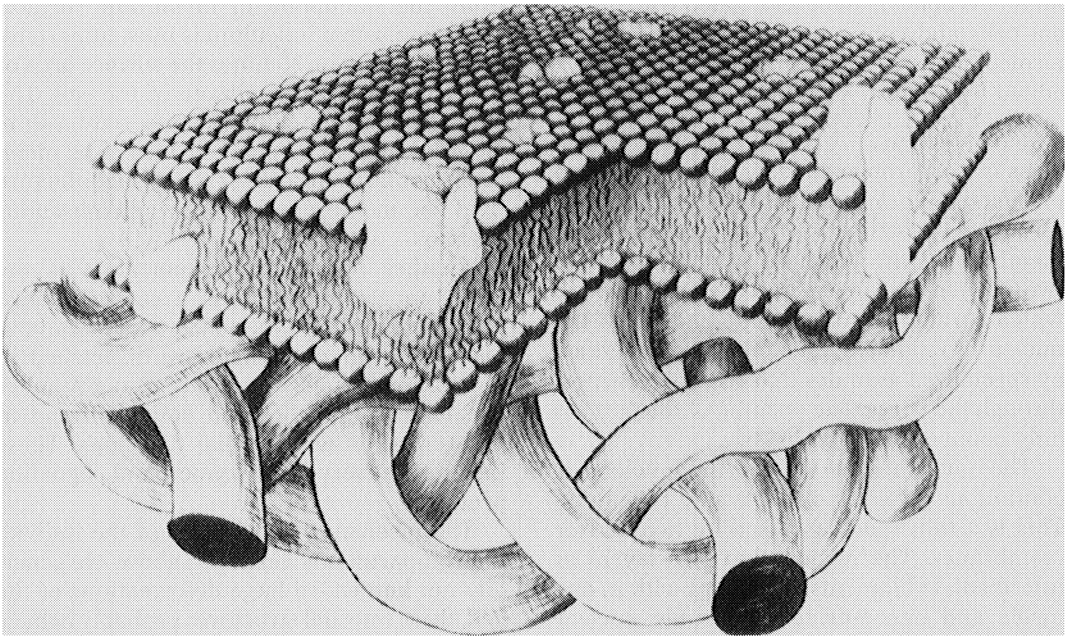

FIG. 2. An idealized sketch of the red blood cell membrane. The lipid bilayer is fluid and is supported by the spectrin network (cytoskeleton) which provides shear elasticity and strength. (From Evans and Skalak, 1980, by permission.)

the case for the red blood cell in which a network layer of spectrin is attached to the cytoplasmic side of the lipid bilayer (Fig. 2). The red blood cell membrane properties usually reported are those of the composite of lipid bilayer and spectrin cytoskeleton (see, e.g., Hochmuth, 1987).

1.2.1 Lipid Bilayer. The lipid bilayer, which is the principal membrane component of all cells, is a structure that forms spontaneously when amphipathetic molecules are placed in an aqueous environment. Amphipathetic means there is a polar head group at one end which is hydrophilic and fatty acid chains on the other end which are hydrophobic. Cell membranes are made up of phospholipids which have two hydrophobic hydrocarbon tails. The molecules of the bilayer orient themselves with the hydrophilic heads pointing outwards on each face of the membrane and the hydrophobic tails on the inside of the membrane. This structure is remarkably stable, although it is quite fluid with regard to motions of the molecules and shear deformations in the plane of the membrane. The hydrophobic and hydrophilic forces hold the thickness of the bilayer to within a narrow range. The polar head groups on each face are then like a monolayer which has a fixed concentration of molecules (about 2.5×10^6 molecules/μm^2). The compressibility of this layer is quite low, and so it resists any attempt to change its surface area. The fluidity of each leaflet of the bilayer results in rapid diffusion of lipid molecules in the plane of the bilayer. The lateral diffusion coefficient is about 10^{-8} cm^2/s. On the other hand, interchange of lipid molecules between the two leaflets is a very slow process. The cell membrane retains an asymmetry throughout its life. There are several different phospholipids in cell membranes. Those which have choline in their head group are found mostly in the outer leaflet. This and other differences result in different surface charges on the two surfaces. Another difference of inner and outer faces is that glycolipids are found only on the outer face. Cholesterol also occurs in both leaflets of the cell membrane. It moves readily from one layer to the other, and so it does not contribute to asymmetry but does reduce the viscosity of the bilayer.

1.2.2 Intramembranous Proteins. The structural functions of providing a seal and fluidity of a cell membrane are provided by the lipid bilayer, but most biochemical functions are carried out by membrane proteins which float in the lipid layer. Some such proteins have been shown to attach to only one leaflet of the membrane or the other. Some intramembranous proteins cross the entire membrane and have functional parts on both faces.

Membrane proteins have fixed orientations with respect to being located on the interior or exterior faces of the cell membrane. They are usually free to diffuse in the plane of the membrane and have a rotational mobility. In lateral diffusion, the diffusion coefficient will usually be half or less than that of lipid because they are much larger molecules. For many purposes, intramembranous proteins are not left free to diffuse, but are restricted to some part of the cell, like neurotransmitter receptors at a synapse. They may be organized in patches, like gap junctions which serve as connections and channels of communication (via small ions) between neighboring cells. They may also be organized in strands, such as tight junctions between endothelial cells which serve to confine blood proteins within the circulation. Patches of membrane proteins are involved in vesicle formation. The charge carried by such a patch of protein on the cytoplasmic side of a cell membrane readily induces vesicle formation by bending the membrane.

1.2.3 Cytoskeletal Components. The lipid of cell membranes is usually free to diffuse and flow in the plane of the membrane independently of the cell cytoplasm beneath it. The membrane is often tethered to cytoskeletal elements via the intramembranous proteins so that the lipid bilayer and the cytoskeletal layer cannot separate in the direction normal to the membrane surface. The cytoskeleton and lipid bilayer then function as a unit in respect to mechanical properties and deformations.

The cytoskeletal structure and properties are most extensively known for the red blood cell membrane. This is due in part to the importance of the red blood cell in blood rheology and in part because the red cell membrane cytoskeleton is readily isolated and has a relatively definite structure (Fig. 2). The spectrin network stabilizes the lipid membrane of the red blood cell into its characteristically smooth shape. The shear elasticity and viscosity of the red blood cell membrane are due almost entirely to the cytoskeleton (see Cokelet *et al.*, 1980).

In most other cells, other than red blood cells, the extent of the cytoskeleton attached to the cell membrane is much less well defined, but is a thicker (about 0.5–1 μm) layer, often consisting of an actin gel. This cortical layer may be under active tension due to myosin, and the surface of the cell is usually wrinkled. Thus, the mechanical properties of the plasma membrane plus the cytoskeleton are dominated by the cytoskeleton. The anchorage of the cytoskeleton of actin to a cell membrane is not so completely known as for the red blood cell. The actin filaments are known to terminate with their plus ends at the cell membrane in many cases and to have special proteins to anchor them to the plasma membrane.

1.2.4 Area Elasticity. The lipid bilayer of cell membranes has a number density of phospholipids (molecules per unit area) which is quite stable. In order to expand the surface area of a cell membrane, isotropic tensile stresses must be applied in the plane of the membrane to overcome the intermolecular attractions and hydrophobic forces. In such an isotropic stretching of a membrane, the area elastic modulus K is defined by

$$K = \frac{\Delta T}{\Delta A / A}, \qquad (1)$$

where ΔT is an increment of isotropic tension (i.e., equal in all directions in the plane of the membrane), and ΔA is the increment of area in a portion of the membrane having initially the area A. Equation (1) is a two-dimensional analog of the three-dimensional bulk or volumetric modulus of a solid. The units of T and K are dyn/cm. The membrane tension T is the integral, over the thickness of the membrane, of stress in the membrane (dyn/cm^2).

For the red blood cell, K has been measured by pipette tests (see Evans and Skalak, 1980) to be about 450 dyn/cm. This value is attributed primarily to the lipid bilayer itself rather than the cytoskeleton which is readily expanded in surface area.

An important characteristic of lipid bilayers is the fact that pores open and the membrane may rupture if the area strain ($\Delta A/A$) is greater than 0.01 to 0.02 (1% or 2% increase in area). The isotropic tension required for such rupture is of the order of 10–20 dyn/cm.

In the red blood cell, large deformations are possible with essentially no change of surface area because of the initial disk shape. In other cells, the plasma membrane is often wrinkled to begin with and contains many pits or attached vesicles. The cell can then deform without increasing the plasma bilipid membrane area by unwrinkling the membrane, like the bellows of an accordion.

1.2.5 Shear Elasticity and Viscosity. Since a bilayer lipid membrane is essentially a fluid, it does not exhibit appreciable elasticity in shear, but does have a definite shear viscosity, which is of the order of 10^{-5} P cm (Waugh, 1982). Most cells have a cytoskeleton attached which has a shear elasticity and shear viscosity which are larger than and predominate over that of the bilayer lipid membrane. For red blood cells, the membrane surface shear elasticity has been derived from pipette tests (see Evans and Skalak, 1980). The shear elastic modulus is defined by

$$\mu = T_{xy}/2\epsilon_{xy}, \tag{2}$$

where T_{xy} is the shear resultant in the membrane and ϵ_{xy} is the shear strain. For the red blood cell, μ is about 7×10^{-3} dyn/cm. This includes the contribution of the cytoskeleton. For other cells, the comparable value of shear elasticity is much higher, but not well established because the bulk properties of the cell interfere with obtaining separate membrane values easily.

The shear viscosity of the red blood cell membrane has been estimated from pipette and other tests (see Hochmuth, 1987). The shear viscosity η of a membrane is defined by

$$\eta = T_{xy}/2\dot{\epsilon}_{xy}, \tag{3}$$

where T_{xy} is the membrane shear stress which causes the shear rate $\dot{\epsilon}_{xy}$. For the red blood cell, η is about 10^{-5} dyn s/cm (P cm or surface poise) (Waugh, 1982). Most of this viscosity is due to the cytoskeleton.

1.2.6 Bending Elasticity. A lipid bilayer has a small resistance to changes in curvature, like a liquid crystal, due to crowding and expansion on opposite faces of membrane when the curvature is changed. Because the two leaves of the membrane are free to slide over each other, the bending stiffness is small. In most cells, the cytoskeleton stresses completely dominate the lipid bilayer in regard to determination of overall shape of the cell. At the scale of small vesicles, the control of curvature by local molecular forces predominates. In the red blood cell, the membrane bending stiffness is important to explain its resting shape, but is usually negligible under large deformations. The bending stiffness B for a membrane is defined by

$$B = M/K, \tag{4}$$

where M is an isotropic bending moment per unit length (dyn cm/cm), which bends a flat membrane into a uniform sphere of curvature $K = 1/R$, where R is the sphere radius. For the red blood cell membrane, including the cytoskeleton, B is about $(1\text{–}2)\times10^{-12}$ dyn cm.

The moduli defined by Eqs. (1)–(4) are expressed in terms of simple experiments for clarity. In a general dynamic deformation the several elastic and viscous effects are superimposed and the deformations may be large. The simple forms of Eqs. (1)–(4) must then be expanded to proper tensor forms, including the large deformation geometry appropriately (see Evans and Skalak, 1980).

1.3 Physical Processes

To account for the active motions of cells, the study of their biomechanics must include various active processes. Active motions cannot be explained on the basis of passive properties and response to external forces. Rather, they involve changes in internal forces and/or rearrangements of charges or mass.

The principal sources of active motions are contractile forces (as in a muscle), polymerization, and osmotic pressures. Effects of cell adhesion and membrane fusion may also be components of active processes.

1.3.1 Polymerization. Changes of cell shape and properties are generally accomplished either by active contraction or by polymerization. The main molecules involved in polymerization are actin and tubulin. Actin gelation occurs at the leading edge of pseudopods and lamella of locomoting cells such as leukocytes, fibroblasts, and endothelial cells. Gelation results in a change of properties from a viscoelastic fluid to a viscoelastic solid. Solation is the inverse process which breaks down the gel into a fluid state. In the process of locomotion, portions of a cell may repeatedly cycle through gelation and solation.

The polymerization of tubulin leads to the formation of microtubules, but not to the formation of a gel. The microtubules generally remain as stiff, straight rods. Their length, number, and orientation may change continuously by polymerization and disassembly. They are prominent in neural growth cones of actively extending axons. It is possible that the process of polymerization may do work in thrusting a pseudopod or other extension of the cell forward. Hill (1981) has shown that it is thermodynamically possible for a polymerization process to do work, but such forces have not been measured or clearly identified in cell movements.

1.3.2 Osmotic Pressure. Osmotic pressure differences can drive fluid across a semipermeable membrane, according to Starling's law:

$$j=K(p_1-p_2-\pi_1+\pi_2), \tag{5}$$

where p_1, p_2 are fluid pressures on either side of the membrane; π_1, π_2 are the corresponding osmotic pressures; and K is a permeability. If a red blood cell is placed in a solution of lower osmolarity than blood plasma, water enters the cell according to Eq. (5) until the internal pressure rises and the flux j drops to zero. This kind of swelling is a possible process in cell motion. Oster *et al.* (1982) have proposed a model of the extension of the acrosome of a sperm by osmotic pressure. In this case there is an influx of fluid which drives the acrosome extension. It is not clear whether or not osmotic pressure is involved in cell locomotion in which repeated cycles of forward movement take place. It is possible that the gelation and solation of actin leads to local osmotic pressure changes because the effective osmolarity varies as the state of the actin and its associated proteins change. Oster (1989) has suggested this as the driving mechanism of lamellipod extension.

1.3.3 Active Contraction. The interaction of actin and myosin results in an active contraction which is the basis of most active motion and force generation observed in both muscle and nonmuscle cells. The heads of the myosin molecules pull themselves along the actin filaments by a rocking motion like a ratchet. This forces a relative motion of the actin and myosin filaments, sliding lengthwise. The geometric arrangement of actin and myosin in a skeletal muscle is highly organized and compact. In smooth muscle it is less organized. In nonmuscle cells which contract, the actin network may be quite random and temporary. For example, the actin-myosin contractile ring which forms during cell division is rapidly assembled during mitosis and then disassembled after cell division is completed.

Microtubules and intermediate filaments are not generally associated with active contraction, but microtubules are involved in active motions in cilia, flagella, and transport of vesicles as described in the next section.

1.3.4 Tubulin Molecular Motors and Laser Tweezers. The discovery and identification of microtubule-based motion and transport in cells has been elucidated only in the last 20 years. Current status of the field may be seen in detail in the two volumes on cell movement edited by Warner *et al.* (1989). Microtubules are involved in two kinds of active motions. One is the bending type of motion of flagella and cilia. The other is active transport of vesicles along microtubules. Although these appear to be quite different kinds of motion, they are associated with the same kinds of interactive molecules, namely, dynein and kinesin.

In flagella and cilia there is a central structure called the axoneme made up of nine pairs of microtubules arranged in a circle and a central pair, all of which are interconnected into a stable structure (Fig. 3). The dynein molecules are attached to the axoneme at regular intervals and have inner and outer arms that interact with neighboring microtubules. The activation of the dynein molecules tends to slide neighboring microtubules longitudinally past one another. The asymmetric distribution of this motion and its periodic control so as to produce bending waves observed in flagella are necessary, but the details are not entirely clear. There are protozoa in which flagella perform three-dimensional helical rotary motions. In these there is a rotation of the outer microtubules with respect to the central element.

The motion of vesicles along microtubules within the cytoplasm of various cells is accomplished by the same kind of dynein molecules as are found in cilia and flagella. However, in

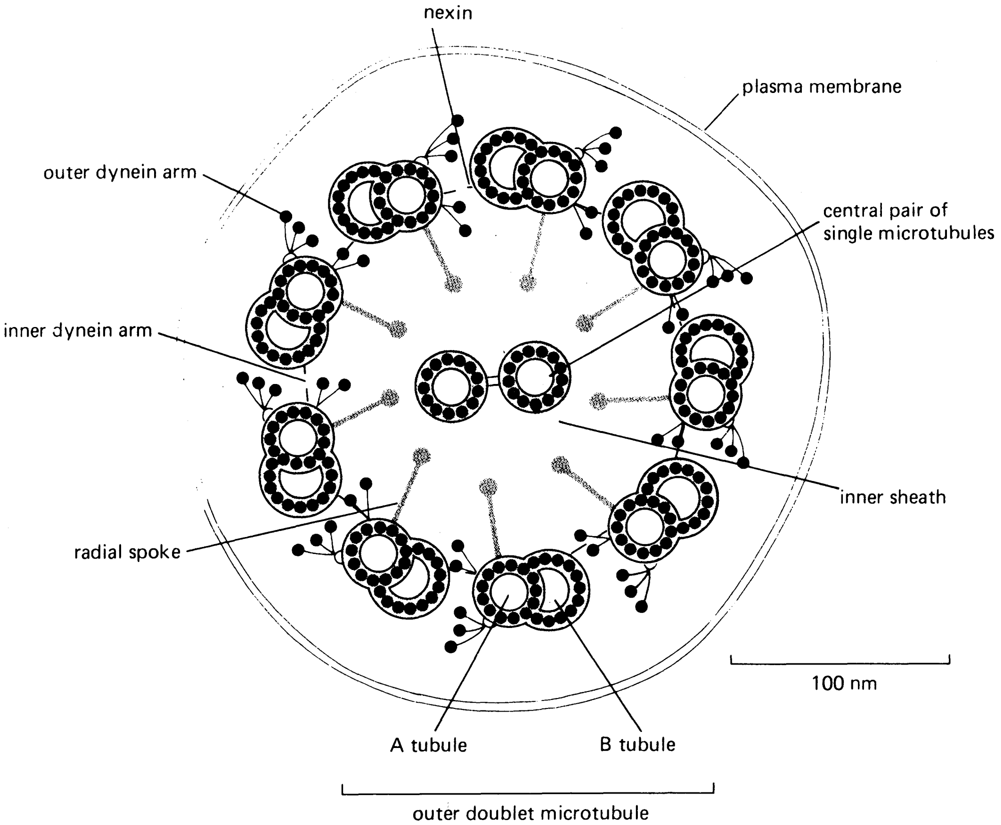

FIG. 3. A schematic sketch of the cross section of a cilium. The various projections from the microtubules occur at regular intervals along the cilium. (From Alberts *et al.*, 1989, by permission.)

this case, the dynein is attached to the membrane of the vesicle or other particle that is transported. The relative motion of the components is due to a walking or ratchetlike motion of the dynein arms along the oriented structure of the microtubules. This resembles, to some extent, the motion of myosin heads along actin filaments.

Another large molecule, kinesin, is also capable of moving vesicles and small particles along microtubules. Kinesin moves vesicles toward the minus (slow-growing) end of a microtubule. Dynein produces motion toward the plus end. Transport of vesicles along microtubules is the basis of so-called fast axonal transport of vesicles in both directions along axons.

The transport of kinesin- and dynein-coated beads along microtubules has been tracked with a precision of 1–2 nm (Sheetz, 1990) by the use of so-called laser tweezers. Small particles in a sharply focused microscopic laser beam can be tracked and manipulated within a closed chamber (Buican *et al.*, 1989). The basic principle is that diffracted photons exert a force on the diffracting particle due to their change of momentum. In a gradient of light intensity a net force can be exerted. This effect centers a particle in the laser beam and can be used to move or track the particle in a controlled manner.

An interesting demonstration of microtubule motor activity can be performed *in vitro* by anchoring the motor molecules, either dynein or kinesin, to a glass surface. Then microtubules in suspension can be seen to be propelled longitudinally along their axis when they settle. The motor action requires the presence of ATP which is hydrolized and serves as the energy source of all active motions associated with interactions with microtubules.

1.3.5 Adhesion and Cell Junctions. The cells of most tissues are connected to each other and to various extracellular matrices produced by cells. The junctions between cells serve as important biochemical barriers or communicating channels. Cell junctions also serve as mechanical connections to prevent opening under stress. In most tissues, the major stress-bearing components are extracellular, as in tendon and bone. These tissues, nevertheless, have living cells interspersed in the extracellular matrices which maintain the tissue in a functional state. Some external appendages such as hair, nails, hoofs, horns, and feathers have no living cells in them. Epithelial cells, which form the outer layer of the skin and line the intestines, are unusual in that there is little external matrix so that any stress applied to them is carried through the cells via junctions connected to the internal fiber systems of the cells. However, they are underlain by a system of extracellular collagen fibers which provide most of the strength of skin.

Cell junctions may be classified as occluding, communicating, and anchoring types. Occluding junctions are called tight junctions and are formed by bonds between lines of molecules which are intramembranous proteins of adjacent cells. Such junctions prevent or retard passage of molecules between cells. The tight junctions of endothelial cells which line all blood vessels keep the protein efflux from the blood to a low level.

Communicating junctions are of two types. One is gap junctions in which matching intramembranous proteins from adjacent cells form a channel through which small molecules (mol. wt. up to about 1000) may pass freely. These junctions are under some control and are shut down if one cell is injured or dies. Another type of communication between cells is through synapse, as in neural cell junctions. Here there is no direct transfer of mass; transmitter molecules are released into the intercellular space which they cross by diffusion and then are received by special receptors on the neighboring cells.

Anchoring junctions also have special structures residing in and on the plasma membrane of the cell. Cytoplasmic plaques called desmosomes protrude through the cell membranes of adjacent cells and form a bond like a spot weld. Cells are attached to an extracellular matrix by half of a desmosome called a hemidesmosome.

Blood cells are unique in that they normally function as separate entities rather than establishing firm bonds to other cells. The adhesion of blood cells and proteins is important to the rheology of blood, to clotting mechanisms, to leukocyte function, and in pathologic conditions. Normally, red blood cells are kept separate by the shear stresses in flowing blood, but if blood is brought to rest, the red cells quickly aggregate. Leukocytes normally flow without adhering to other cells, but adhere to the vascular endothelium when stimulated by a chemoattractant such as is released by an injured tissue. In this case the leukocytes will actively penetrate the endothelial wall and migrate toward the injury. This process of locomotion also requires transient adhesion. The process of phagocytosis, for example, when a leukocyte ingests a bacterium, also requires adhesion of the leukocyte and bacterium. The physical basis of cell-to-cell adhesion is developed in the book edited by Bongrand (1988).

2. MECHANICS OF BLOOD CELLS

Blood is a suspension of several different types of cells in a plasma which is clear Newtonian fluid. In its bulk behavior, blood is non-Newtonian and exhibits a shear-thinning behavior due to the presence of the blood cells. Most of the cells in blood are red blood cells (erythrocytes), normally occupying a volumetric fraction (hematocrit) of about 40%–50% of the total blood volume. White blood cells (leukocytes) are normally much less numerous than erythrocytes (about 1:600) and occupy about 0.3% of the total volume. Platelets are more numerous than white cells, but are smaller and occupy about 0.5% of the total volume.

The bulk rheology of blood is determined by the properties of the plasma and red blood cells primarily because the volumetric fraction of leukocytes and platelets is small. Leukocytes are more important in capillary flow where they may block a vessel completely. Platelets are important to the formation of clots and respond to injury very quickly. Otherwise, the rheological effects of platelets are negligible (see BIORHEOLOGY).

2.1 Red Blood Cells

The study of the physical properties and behavior of red blood cells is more extensive than for any other cell probably because it is an important cell to blood rheology and hematology and because it is a simple, discrete, and readily accessible cell.

Human red blood cells are disks about 8 μm in diameter and 2.5 μm thick with a dimple in each face [Fig. 4(a)]. At rest, the surface is a smooth axisymmetric shape with an average surface area of about 130 μm^2 and a volume of 98 μm^3 (Fung, 1981). The surface area is about 30% greater than that required to enclose the same volume in spherical form. Hence the disk shape can be and is distorted in flowing blood to a considerable extent without any change in the surface area or volume. The red blood cell can be stretched to about double its resting diameter without injury.

2.1.1 Membrane Properties and Cell Shape. The properties of the red cell membrane usually reported pertain to the bilipid plasma membrane plus the cytoskeleton of spectrin on the inner membrane surface (Fig. 2). For a general deformation, the principal membrane stresses T_1, T_2 in the plane of the membrane may be approximated by (Evans and Skalak, 1980)

$$T_1=\frac{\mu}{2}(\lambda_1^2-\lambda_2^2)+K(\lambda_1\lambda_2-1)+2\eta\dot{\lambda}_1/\lambda_1, \quad (6)$$

$$T_2=\frac{\mu}{2}(\lambda_2^2-\lambda_1^2)+K(\lambda_1\lambda_2-1)+2\eta\dot{\lambda}_2/\lambda_2, \quad (7)$$

in which λ_1 and λ_2 are the principal extension ratios (final length of a line element divided by initial length) in the two principal axes of deformation; $\dot{\lambda}_1/\lambda_1$ and $\dot{\lambda}_2/\lambda_2$ are the strain rates in the principal directions. The coefficients μ and K are the shear and area elastic moduli and η is the membrane viscosity, as discussed in the preceding section.

For a complete description of the membrane properties, the behavior in bending must be specified also. Assuming the bending moment is isotropic, it can be approximated in principal axes of the membrane curvature by

$$M_1=M_2=B(K_1+K_2)/2, \quad (8)$$

where K_1 and K_2 are the changes in the curvatures from the unstressed state and B is the bending stiffness discussed in the preceding section.

The use of the stretch ratios λ_1, λ_2 and the curvature changes K_1, K_2 imply that there is a known unstressed state of the membrane. However, such information is not definitely known. The normal biconcave shape of the red blood cell may or may not have residual stresses in it. If the unstressed shape of the cell is assumed to be spherical, then it can be shown that a shape very close to the observed red blood cell shape can be achieved by deflating the uniform sphere to the appropriate volume. If the unstressed shape of the cell is assumed to be the normal disk shape [Fig. 4(a)], then during osmotic swelling, the shape eventually approaches a sphere [Fig. 4(b)]. The shapes shown in Fig. 4(b) are minimum strain-energy shapes within the constraints of constant surface area and specified volumes.

If the initial, stress-free curvature of the cell is altered, or the ratio of the shear modulus times the square of the cell diameter to the bending modulus is changed, then a variety of different rest shapes may be obtained (Stokke *et al.*, 1986). *In vitro*, a large variety of shapes are observed (e.g., echinocytes having many spicules or stomatocytes which are cup-shaped). These shapes are usually the result of biochemical modifications of the membrane. These changes usually influence the stress-free curvature (also called the spontaneous curvature).

2.1.2 Hemoglobin Properties. The hemoglobin solution inside the red blood cell has been shown to be a Newtonian fluid with a viscosity of about 6.2 cP, i.e., about five times the viscosity of blood plasma or water and about twice the apparent viscosity of whole blood at high shear rates (see Cokelet *et al.*, 1980). The density of hemoglobin is a few percent higher than that of plasma which is sufficient to be able to measure blood velocity ultrasonically. For purposes of red blood cell and blood rheology studies, the hemoglobin is well approximated by an incompressible Newtonian fluid.

2.1.3 Pipette and Flow-Chamber Tests. The mechanical properties of red blood cells have been derived from various microscopic tests, such as pipette, flow-chamber, and siev-

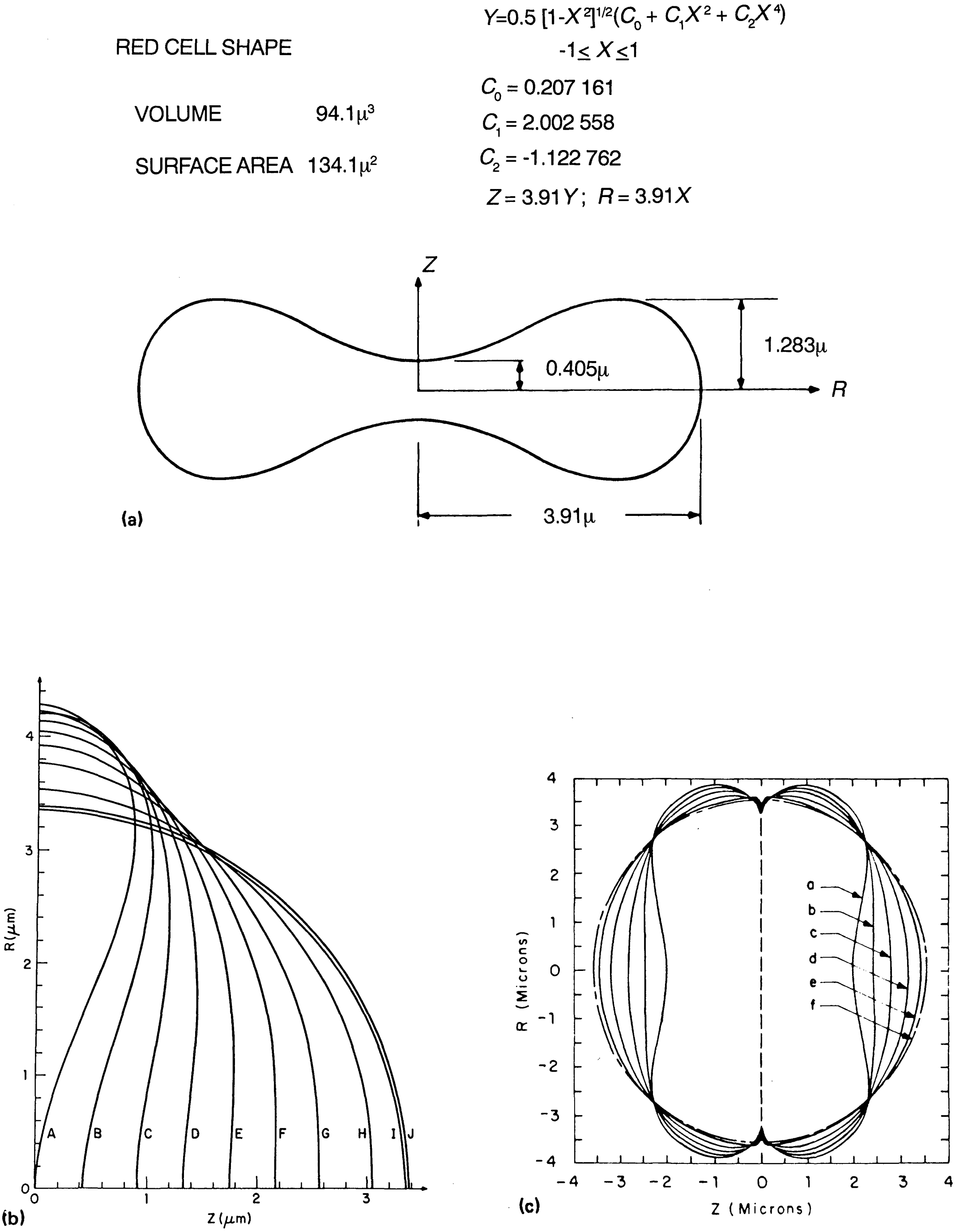

FIG. 4. (a) Red blood cell shape assumed for free suspension, at rest. (From Skalak, 1980, based on Evans and Fung, 1972, by permission.) (b) Computed shapes of various stages of swelling of a red blood cell. Curve *B* was taken to be the unstressed shape. Curve *A* is computed for subnormal cell volume. The surface area is constant until the state *I*. (From Zarda *et al.*, 1977, by permission.) (c) Computed two-cell rouleau shapes assuming initial shapes as in (a). The successive shapes (a)–(f) correspond to increasing adhesive energies. The dashed line is the limiting sphere theoretically reached at infinite adhesive energy. (From Skalak, 1984, by permission.)

ing experiments. In pipette experiments, a portion of the cell is drawn into the tip of the pipette, which is about 2 μm in diameter. In this experiment the portion of the cell outside the pipette retains its normal disk shape and the pressure difference required to draw a tongue of cell membrane into the pipette is interpreted in terms of the shear elastic modulus of the membrane (see Evans and Skalak, 1980). This experiment may also be used to estimate the bending stiffness of the membrane from the pressure required to buckle the membrane around the tip of the pipette (Evans, 1983).

The pipette experiment performed on a swollen red cell can be used to derive the area elastic modulus of the membrane by aspirating enough of the cell so that the remainder is spherical. By further increasing the pressure, the membrane is stretched in isotropic tension and the surface area increment can be calculated from the increase in the length of the membrane tongue inside the pipette (see Evans *et al.*, 1976).

The viscosity of red blood cell membranes can be estimated from time-dependent pipette experiments (Chien *et al.*, 1978) and by stretching a cell which is fixed to a glass slide (Hochmuth *et al.*, 1979). The time of recovery of the normal red cell shape is about 0.3 s, which, together with the shear elastic modulus, allows an estimate of the membrane viscosity.

Another type of microscopic experiment is the so-called flow chamber in which fluid is passed between closely spaced parallel plates (gap width about 125 μm). If red cells are allowed to settle in such a chamber, some adhere to the wall. Then, by pumping fluid through the gap, a known fluid shear stress can be applied to the cell. The experiment can be observed directly through a microscope. From measurements of the stretching and recovery time of the cell, the elastic shear modulus and membrane viscosity can be estimated.

2.1.4 Flickering. When red blood cells are observed in a suspension without any macroscopic flow, the cells do not appear to be entirely stationary, but the membrane has a small random motion which is termed flickering. Some components of the motion are slow enough and over a sufficient area to be detected visually through a microscope. Flickering has been reported as early as 1890. It is attributed to thermal fluctuations, like Brownian motion of small particles in a fluid. The red blood cell is too large and deformable to exhibit Brownian motion as a whole, but the membrane is sufficiently flexible to be transiently deformed by the thermal molecular motions of the surrounding liquids. It is restored to its equilibrium shape primarily by its bending elasticity. For small deformations of the dimple of the red blood cell membrane, the shear and stretching energies are of higher order so bending modes predominate.

In the analysis of flickering it is assumed that all modes are independently excited and that there is an equipartition of energy in all modes. The mean bending strain energy in each mode is set equal to $k_B T$, where k_B is Boltzmann's constant and T is the absolute temperature. Quantitative estimates of the bending stiffness of the red blood cell membrane were made by Brochard and Lennon (1975). Their results are lower than those found by more direct measurements. Refinements of the method (see Schneider *et al.*, 1984) may explain the discrepancy.

2.1.5 Capillary Flow. The diameters of most capillary blood vessels are about the same as or even a little smaller than the diameter of red blood cells. Hence red blood cells are normally quite deformed when flowing in capillaries. Due to the red cell flexibility, the increase in pressure gradient due to the presence of the cells is a factor of two or three times that for the same velocity of plasma alone. The factor depends on the velocity of flow because the cells are deformed to a greater extent at high velocity (see Skalak *et al.*, 1989). In bulk flow, as in large arteries, the apparent viscosity of blood is also shear-rate dependent due to the deformability of the red cells (see BIORHEOLOGY).

In small capillaries (< 7-μm diameter), red blood cells usually move in single file and tend to be more or less axisymmetrically located. As the diameter of the vessel and the hematocrit increase, a nonaxisymmetric arrangement with two or more rows tends to form. When two rows form, the cells tend to be arranged in an alternating pattern which is sometimes called a zipper arrangement (see Cokelet, 1987). An arrangement with two rows of cells arranged with each pair of cells in the same cross section of the vessel is unstable (Sugihara-Seki and Skalak, 1989). The alternating arrangement (like a Kármán vortex street) is stable. At larger vessel diameters, (> 15-μm diameter) the arrangement of cells is observed to be irregular. In capillary flow, the red blood cells generally move more rapidly than the mean velocity in the vessel because they occupy the central portion of flow where the velocity is greatest. This is called the Fåhraeus effect which is detectable in vessels up to about 200 μm in diameter (Cokelet, 1987). There is another result, largely due to the Fåhraeus effect, which is called the Fåhraeus-Lindqvist effect. This refers to the fact that the apparent viscosity of blood decreases as the vessel diameter decreases below about 300 μm (see BIORHEOLOGY).

2.1.6 Rouleau Formation. Whenever blood is brought to rest, the red blood cells aggregate, typically forming cylindrical stacks called rouleaux because they resemble rolls of coins. When two red blood cells touch, they slide or roll until the two disk profiles are quite accurately aligned axisymmetrically [Fig. 4(c)]. Depending on the strength of the adhesive molecular species, the cells will be more or less deformed. The spacing between adjacent cells is uniform and is proportional to the molecular weight of the adhesive species.

Red blood cells normally carry a small negative charge so that work must be done to draw the cells together during adhesion. This energy is supplied by the chemical potential decrement of the adherent species. An unusual aspect of adhesion of cells, in contrast to adhesion of most inert materials, is that the adhesive energy per unit area for a particular cell and concentration of adhesive molecules is not constant. During the forcible separation of two cells the adhesion energy increases as the contact area decreases. This behavior is attributed to the mobility of the molecules and the proteins within the cell membranes. The ligand molecules attached to the cell surface outside of the contact area can move to the contact area by diffusion. In addition, when two cells are gradually pulled apart, the tension developed in the ligands tends to slide them along tangentially to the membrane and sweep them into the contact area. The concentration of bonds between two red blood cells by this mechanism has been demonstrated by fluorescent measurements by Vayo (1986). A theoretical treatment including effects of movable bonds on leukocytes has been given by Tözeren *et al.* (1989). Various cell adhesion phenomena are discussed in the book edited by Bongrand (1988).

2.2 White Blood Cells

The rheology of leukocytes is less definitively explored and is much more complex than that of red blood cells. The presence of a nucleus and many organelles makes the passive properties viscoelastic and nonlinear (Chien *et al.*, 1984). In addition, the possibilities of active motion and rearrangements of the cell interior can produce cell locomotion and phagocytosis which do not occur in red blood cells.

2.2.1 Passive Properties: Pipette and Poking Tests. The passive properties of leukocytes can be probed by micropipette aspiration and by observing their recovery after deformation (Fig. 5). White blood cells are much stiffer than red blood cells and recover from a deformation much more slowly. Whereas a red blood cell will recover its biconcave shape in about 1/3 of a second, a white blood cell deformed into an ellipsoidal shape may take 60–100 s to recover its spherical shape. In free suspension, the fact that it does recover a spherical shape has been interpreted to indicate that the cortical shell of the cell is under a small active tension. The cortical shell is a layer of actin gel primarily, about 0.5 μm thick, just below the plasma membrane. The contractile stress is due to a small amount of myosin present and acts like a small surface tension.

Most of the cytoplasm of a leukocyte is in a constant state of Brownian motion and is a viscoelastic fluid in its bulk behavior. It has been modeled as a Maxwell fluid (Dong *et al.*, 1988) for small deformations and as a Newtonian fluid (Evans and Young, 1989) for large deformations. The apparent viscosity for large deformations is about 2000 dyn s/cm^2. Due to the cortical tension, there is a critical pressure in a pipette aspiration experiment below which the cell will not fully enter the pipette.

For small deformations, the passive behavior of the leukocyte can be represented as a homogeneous sphere of a viscoelastic solid, usually called a standard solid, made up of a Maxwell element (a spring and a dashpot in

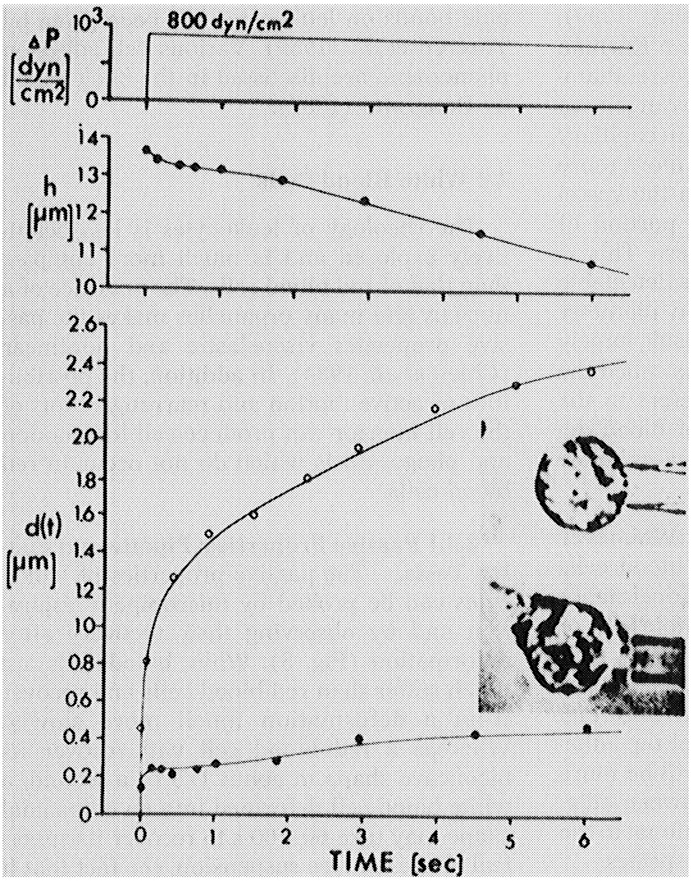

FIG. 5. Micropipette aspiration pressure ΔP and deformation $d(t)$ for a neutrophil which was aspirated at the main cell body (open circles) and at a small protopod (solid circles). The length h is the major axis of the cell. (From Chien *et al.*, 1984, by permission.)

series) in parallel with another spring (which represents the cortical shell). This model is convenient to compare results from cells of different types or various biochemical reagents (Sung *et al.*, 1982).

Another kind of test which has been applied to leukocytes as well as other cells is an indentation experiment dubbed "cell poking" (Zahalak *et al.*, 1990). In this technique a small flat-ended cylinder, a few microns in diameter, is pressed into the surface of a cell which is attached to a rigid substrate. The force required for indentation is calculated from the deflection of a slender glass beam which supports the indenter. A very fine (0.1-μm) control of the indenter is achieved by a piezoelectric driver. The results on leukocytes show a greater stiffness by a factor of 2 to 10 over that derived from micropipette experiments. So far, the reasons for this discrepancy are not known.

2.2.2 Pseudopod Formation. The pseudopods of leukocytes are typical cell extensions consisting of thin rods or sheets that are actively protruded from the cell body. In leukocytes, it may be seen in electron microphotographs that the interiors of pseudopods have a uniform appearance with very few, if any, organelles or granules. The pseudopods consist mainly of actin gel, and it has been established by fluorescence studies that pseudopods grow by addition of actin monomers to the actin gel polymer at the tip, i.e., at the cell membrane.

A theory of pseudopod formation has been given by Zhu and Skalak (1988) in which the driving force for the pseudopod extension is actin polymerization at the tip of the pseudopod. It is implied that the polymerization process can push the cell membrane outwards. This also acts to draw G-actin by convection and diffusion from the cell body to the tip of the pseudopod. The process of pseudopod formation is under biochemical control, but full details are not known. Chemoattractants sensed by the cell membrane trigger a series of internal responses. The level of calcium ion appears to be one of the chief regulators of actin polymerization.

The retraction of a pseudopod appears to be a different process than its formation. The solation of the gel takes place at the base of the pseudopod where it joins the cell body. A contractile force in the cortical shell of the cell is required to pull the cell back into spherical shape.

Pseudopod formation is the basis of phagocytosis in which a leukocyte engulfs a bacterium or other foreign object, like a small plastic bead. The pseudopods are thin sheets in this case and adhesion plays a role in stabilizing the leukocyte contact to the bacteria or particle. When the pseudopod sheet completely surrounds the bacteria, the leukocyte membrane fuses so that the bacteria is wrapped in a cell membrane and completely internalized after which it can be killed and digested.

2.2.3 Capillary Flow. The flow of leukocytes in capillaries is quite different from that of red cells which are readily and quickly deformed by the stresses in normal blood flow. Leukocytes pass freely through capillaries which are larger than about 8-μm diameter, but move more slowly than red cells because they are larger and the Fåhraeus effect is smaller. As a result, red blood cells often catch up to leukocytes in capillaries in a so-called train formation.

If a leukocyte encounters a narrow capillary, it may block it completely while it is slowly deformed by the driving pressure. This may take a minute or more. This is one source of the apparently random variations of velocity found in capillary flows.

Another distinction of leukocyte behavior is that, in small venules, the leukocytes may adhere to the wall or roll along slowly. They can even adhere and actively move upstream against the flow. Such adhesion and motion are part of the processes of defense against bacteria or cell injury by leukocytes.

2.2.4 Sieving Tests. A clinical test of blood cell deformability can be based on the time it takes to pass a fixed volume of blood through a fine porous filter at a fixed pressure, or by measuring the pressure for a fixed flow rate. A typical filter is a thin polycarbonate sheet, about 13 μm thick, in which pores about 5 μm in diameter have been produced by radiation and etching. Such tests were originally thought to measure red blood cell deformability, but it has been shown experimentally and theoretically (Skalak *et al.*, 1987) that the results are more nearly indicative of the number and condition (activated or nonactivated) of the white blood cells present. The stiffer white cells tend to clog the pores and cause a disproportionate rise in pressure. The early phases of such tests can be used as a measure of red blood cell deformability before white cells are trapped in significant numbers.

3. CELL MOTILITY

Many different kinds of cells during some phase of their life actively move from one location to another. Some, like fibroblasts, may be motile most of the time, while others, like endothelial cells, may relocate only to compensate for injury or other changes. Cells of the body usually require a solid substrate or neighboring cells to react against to produce their own motion and cannot swim in a fluid. On the other hand, micro-organisms are usually adapted to swimming in free fluid.

The driving mechanisms of cell motility are of two kinds. In cells of the body, locomotion is the result of rearrangement of the internal cytoskeleton by gelation, solation, and contraction. Adhesion and release is also necessary. In free-swimming micro-organisms and in cilia the microtubule structures remain fixed in place relative to other cell structures and in composition, but perform oscillatory deformations that produce movement. These different modes of motility are discussed in more detail in the following sections.

3.1 Flagella and Cilia

The basic driving mechanism or motor is the same in flagella and cilia. The active element is an arrangement of microtubules in an axoneme consisting of nine pairs of tubes in a circle plus one central pair (Fig. 3). Flagella, like the tail of a sperm, usually act singly and have a wavelike motion. Cilia beat back and forth with different configurations in the forward and backward strokes. Cilia occur in huge numbers, such as $10^9/\text{cm}^2$ in the human respiratory tract, and act in concerted waves.

3.1.1 Flagella Mechanics. The motion of flagella, viewed in successive frames of a video, looks like a large-amplitude sine wave which propagates toward the rear (Fig. 6). This backward motion produces the forward propulsive force. The entire mechanical system of elastic forces and active bending of the flagella and the external viscous fluid flow have been analyzed in considerable detail (Holwill, 1989). It has been shown that to overcome the elastic flexural stiffness and viscous forces, the active bending moment must be distributed along the length of the flagella. In fact, the activation of dynein needs to be coupled to the bending movement of its neighbors in a mechanical way to produce a propagating wave. The asymmetry required to produce a bending motion from the relative sliding motion provided by dynein activity is provided by activation of only one-half of the circular cross section. Another requirement is that free sliding of tubules relative to each other is prevented. This is provided by radial spoke elements.

The details of the control system that produces a traveling wave and a definite beat frequency are not entirely clear. The maximum frequency is in the range of 50–100 Hz. This suggests that biochemical and mechanical events occurring in the movable dynein arms must be locally coupled and propagate like falling dominoes. One theory suggests that the active bending moment is turned off when the curvature reaches a predetermined level (Brokaw, 1989). The frequency and even the direction of the wave propagation are sensitive to calcium levels. The action of calcium has been identified as operating through the regulatory protein calmodulin (Otter, 1989).

The fluid mechanics of the external flow and forces generated by flagella motion are developed on the basis of drag coefficients for the forces exerted on an element of length *ds* of the flagella in the directions normal and tangential to *ds* (Lighthill, 1975). Taking components of these forces in the direction of motion and integrating along the length of the flagella yields the appropriate propulsive force. In the steady state, this propulsive force is balanced by the viscous drag of the cell body which is also computed by viscous, low-Reynolds-number flow theory.

3.1.2 Cilia Motions. The beating of cilia is coordinated so that waves are propagated in one direction, like waves on a wheat field in the wind. In the trachea, the direction of wave propagation is toward the mouth. In microorganisms that use cilia for propulsion, the wave direction is toward the rear. The beating of each cilium goes through a cycle of about 0.1–0.2 s in duration. The power stroke in the direction of propagation takes place quickly with the cilium nearly straight and fully extended. During the return stroke, the cilia is bent so it lies close to the surface.

Organisms that swim by cilia motion are usually in the range of 20–2000 μm in length. The number of cilia of a single protozoan may be 200–10^6. Individual cilia range from about 5 to 15 μm in length. The scaling of organisms that swim by cilia motion has been studied by Sleigh and Blake (1977). It is shown that the efficiency of cilia propulsion decreases as L^{-1} where L is the body length. This is due, in part, to the fact that very large numbers of cilia are required as the size of the animal increases and the ratio of cilia length to body length decreases. The theoretical upper limit of efficient size is about 350 μm. Larger ciliated animals swim with low efficiency. On the

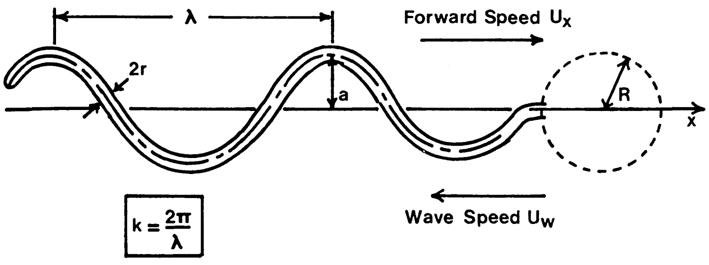

Parameters used to describe undulatory propulsion

FIG. 6. Schematic of undulating flagella motion indicating opposite directions of wave motion and forward speed. (From Holwill, 1977, by permission.)

other hand, the ratio of wave amplitude to wave length of most cilia waves is found to be near a theoretical optimum (Holwill, 1977).

3.2 Amoeba Motility

The locomotion of an amoeba generally proceeds in a series of cyclic stages rather than a continuous motion. The typical time for a cycle is in the range of 1–10 min during which the cell may elongate, put out extensions, and move forward perhaps half a body length. A central feature of amoeboid movement is the protrusion of a pseudopod or lamella from the leading edge in the direction of motion. The physical basis of the motive force for such protrusion is not entirely clear. One kind of theory postulates an actin-myosin interaction which always produces contraction of the actin network and can thus generate internal flows (Dembo, 1989). Another force that may be involved is osmotic pressure (Oster and Perelson, 1987). In any case, the solation and gelation of an actin network is part of the cyclic process. The polymerization of the actin itself may be a source of propulsive force (Zhu and Skalak, 1988). These various theories are outlined in the following sections.

3.2.1 Two-Phase Models. Two-phase models consist of an actin gel or network component which is considered to be solid or a very viscous fluid, plus a fluid phase which occupies the pores of the network. In some regions the network may be dense and solid. In other places the network may be so sparse that the cytoplasm is essentially fluid. In any case, the two phases interpenetrate and may have some relative motion. This produces a viscous drag force on the network. This force may also act in reverse; a contracting network can move fluid along via this viscous drag. In addition to the solation, gelation, and contraction of the network, mass transport can take place by diffusion and by pressure-driven convection forcing fluid through the network.

A set of equations and boundary conditions has been proposed by Dembo (1989) and solved numerically for a model of internal motion in *Amoeba proteus*. The basic equations are the conservation of mass of actin and of the fluid, equations of motion of the actin network and the fluid, and an equation on a control substance which may be concentration of calcium, but is not positively identified. The distribution of the control factor is determined by a diffusion and convection equation as it is carried by the fluid phase. A rate of automatic decay of the control substance is also postulated. The control substance regulates the contractile stress generated by the network. At the forward end of the amoeba, the concentration of the control substance is high. The gelation or solation of the network is assumed to depend on the density of the network. It is postulated that there is an equilibrium density toward which the network evolves with a time constant of about 12 s.

Numerical results (Dembo, 1989) show a close approximation of the steady internal flow observed in *Amoeba proteus* when it is fixed to a substrate. This is a so-called fountain flow in which fluid flows forwards in the center of the amoeba, spreads out at the anterior end, and returns with a sheath (endoplasmic layer) of actin gel which moves steadily toward the rear of the cell. In Dembo's numerical solutions, it is shown that the kind of reaction-diffusion system of equations for interpenetrating networks is prone to instability of static fields, but can produce steady internal motions.

3.3 Leukocytes and Fibroblasts

The projection of pseudopods of leukocytes and fibroblasts is a cyclic process in time, but appears to be a more homogeneous process than in amoebae with respect to variations through the thickness of the projection. Fountain type of flow is not observed in the sense of a core region of fluid moving forward and a shell of network moving backward. There is another kind of steady motion observed called treadmilling in which the front edge of the cell is stationary but actin filaments move steadily backwards, away from the cell membrane. The implication is that actin in solation (G-actin) percolates through the actin gel (F-actin) and polymerizes at the membrane surface. The treadmilling of actin gel in a pseudopod is established by fluorescent markers and photobleaching techniques (Wang, 1985). The driving force for the treadmilling motion, or for pseudopod extension, is not entirely clear. One possibility is that the pressure at the tip of a pseudopod is elevated by local osmotic effects. The cleavage of actin and profilin of the G-actin complex by enzymes at the cell membrane could be the cause of increased osmolarity. This possibility is discussed by Oster and Perelson (1987). The role of osmotic pressure has also been developed in acrosomal extensions (Oster and Perelson, 1987). The fact that raising external osmolarity inhibits acrosomal protrusion favors the role of osmotic pressure.

Another possible source of protrusion force is the polymerization process itself. If the gelation is controlled at the membrane, the addition of actin molecules to the existing network would force a relative motion of separation of the membrane and the existing network. This polymerization force could thrust the network backwards as in treadmilling or push the membrane forwards as in pseudopod extension. Equations governing such pseudopod growth and numerical solutions are discussed by Zhu and Skalak (1988). Hill and Kirschner (1982) have discussed the theoretical possibilities of work done by polymerization, but it has yet to be demonstrated directly experimentally.

3.3.1 Sequence of Locomotion Cycle. The typical cycle of cell locomotion starts with protrusion of a pseudopod or lamellapod. The second step is attachment of the pseudopod to the substrate or to the surrounding cells or extracellular matrix. If there is no attachment, the pseudopod will be retracted. If the pseudopod is attached, a third and last phase ensues which is a contraction of the body of the cell and to some extent of the pseudopod. The net result is to pull the body of the cell forward over the pseudopod. During this process, there may be previous attachments that are now at the rear of the cell. Sometimes long, thin strands of actin filaments stretch from the old attachments to the cell body. These are the so-called stress fibers, although the stress in them is not known and they are more likely to be stretched passively rather than actively. The breaking of posterior attachments and reabsorption of the stress fibers into the cell body completes the cycle. The cell is then ready to start a new protrusion. The breaking of the posterior bonds suggests that cell adhesion is time dependent so that during cell contraction, the weakest bonds break, appropriately at the rear. The contraction probably extends over most of the cortical shell of the cell by actin-myosin interaction. Points of attachment have been shown to connect directly to interior actin filaments so that contractile stress can be transmitted to the boundary (Alberts *et al.*, 1989).

3.3.2 Mechanics of Contraction A bipolar myosin molecule (one with an active head on each end) can slide two actin filaments past each other indefinitely if ATP is supplied. This action can contract an actin network. When myosin reaches a junction of the actin network, it may shift to a new actin filament and continue contraction. *In vitro* gels can be contracted to a small fraction of their original extent. Another process which makes this possible is that the cross-linking molecules which form the actin network junctions dissociate and reassociate repeatedly. The temporary release of junctional bonds means that myosin has the opportunity to contract the gel further over a period of time much more than it could without breaking network bonds.

If the opposing forces are sufficient, contraction is stopped, but a steady force may persist. This is the condition that likely occurs in the cortical shell of cells which are in a static position, such as the free-floating leukocyte with a spherical shape or an endothelial cell in a layer which is confluent and not altering its geometry in response to changing stress fields. These kinds of stresses are a type of prestress or residual stress at the cellular level. Various aspects of the mechanisms and regulation of cell motility are discussed in the book edited by Ishikawa *et al.* (1986).

4. CYTOKINESIS AND GROWTH

Cytokinesis refers to the process of cell division which begins by the formation of a cleavage furrow which then contracts and divides a cell in two (Fig. 7). This mechanical event is preceded by several steps during which the internal structures of the cell, particularly the nucleus and its chromosomes, are divided into two sets which are sequestered at opposite ends of the cell. The mechanics of these early phases involve microtubular guidance, but are not yet analyzed in quantitative terms of forces and resistances. The cleavage phase has been described to some extent by mathematical models (Akkas, 1990).

Growth takes place by cell division followed by an increase of size of individual

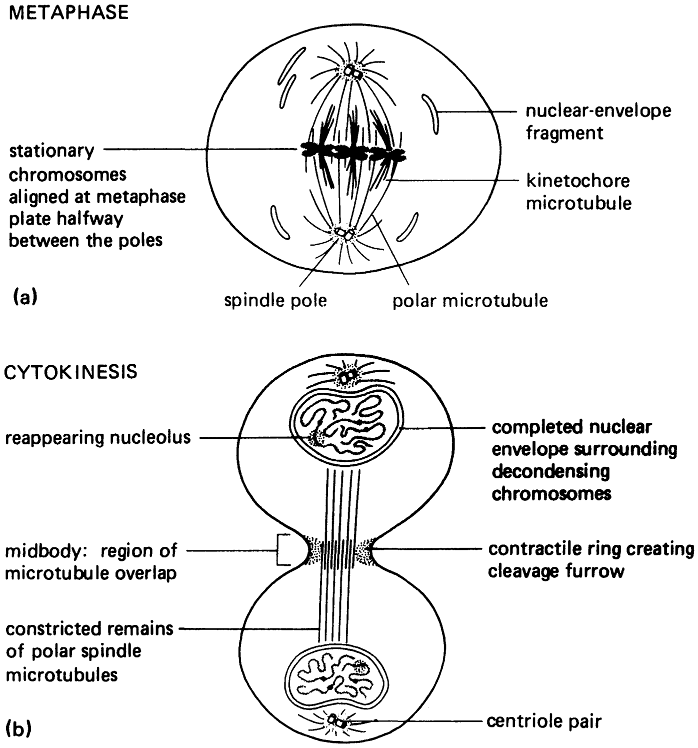
METAPHASE
stationary chromosomes aligned at metaphase plate halfway between the poles
nuclear-envelope fragment
kinetochore microtubule
(a)
spindle pole
polar microtubule
CYTOKINESIS
reappearing nucleolus
completed nuclear envelope surrounding decondensing chromosomes
midbody: region of microtubule overlap
contractile ring creating cleavage furrow
constricted remains of polar spindle microtubules
(b)
centriole pair

FIG. 7. (a) Schematic sketch of the metaphase stage of cell division showing the alignment of chromosomes on a plane halfway between the two spindle poles. (b) Schematic sketch of the cytokinesis stage of cell division showing cleavage near completion. (From Alberts *et al.*, 1989, by permission.)

cells. The arrangement of large numbers of cells into tissues and organs is called morphogenesis. The form and functioning of all tissues depends on the appropriate behavior at the cellular level. All development is driven by the interactions of individual cells. This interactive process is extremely complex. It is basically controlled by the genetic code which gives each cell its instructions for manufacture of proteins, enzymes, and extracellular components. The instructions for tissue formation, size, and shape are less explicit and depend on interactive response between cells and between the organism and the external environment. Mechanical rearrangement processes may utilize differential adhesions, cell locomotion, cell differentiation, and cell metamorphism from one form to another. For example, stem cells of the bone marrow give rise to red blood cells, leukocytes, and platelets.

4.1 Mechanics of Cell Division

The mechanics of cell division usually refers to two main processes, namely, the movement of chromosomes (during anaphase) and the subsequent cleavage of the cell into two cells (cytokinesis) (Fig. 8). However, many additional mechanisms are operative during mitosis which are well known in that they are clearly identified and analyzed as to the form and movement of the components involved. But the forces and the specific control mechanisms that bring about the observed results are not known in detail. Most accounts of the mechanics of cell division contain no numerical values for forces, viscosities, or other properties, but simply describe the events that take place. The number of different components involved is large, including internal membranes that disassemble and reassemble, chromosomes that divide, microtubules, centrosomes, actin, myosin, organelles, and endoplasmic reticulum, to name just some of the principal constituents. The concerted relative movements in space and time of these components will have to be described by a multiphase, multiply interactive media theory with both Brownian motion (random events) and highly specific and symmetric events all combined, while keeping track of the location, size, and properties of hundreds of entities.

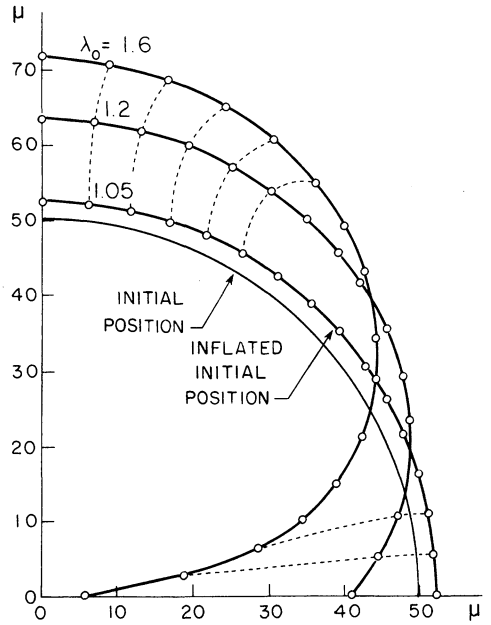

FIG. 8. Computed shapes of cell undergoing cytokinesis. The curves show one-quarter of the field of Fig. 7(b). The dotted lines are particle paths. (From Akkas, 1980, by permission.)

Nothing like this has been attempted. The mechanics theory applied to date is to models of one or two components only.

4.1.1 Contractile Ring Models. When a cell divides, it elongates slightly from a spherical shape and then a furrow appears, girdling the cell around the smaller diameter. Just below the membrane in the cleavage furrow, a contractile ring is formed which is made up of actin and myosin filaments. This contractile ring is diffuse at its inner boundary and may be only 0.1 μm thick and 10 μm wide. There are about 5000 F-actin filaments, each about 15 nm thick, in a typical contractile ring. The connection of the contractile ring to the cell membrane has not been demonstrated, but observation shows that as the contractile ring contracts, the membrane remains closely apposed to it.

Models of cell division assume that the contractile ring exerts a circumferential tension and that the rest of the cell behaves in a passive manner, as a viscoelastic fluid. The contraction of the contractile ring proceeds from the full diameter of the cell down to essentially zero and the ring then disappears (Fig. 8). The origin of the contractile ring is most probably from actin and myosin already present in the cortical shell. It appears that cytokinesis is preceded by a rounding of the cell which may indicate a general contraction of the cortical shell which is somehow enhanced at the equator by the influence of the two mitotic asters which are then separated by the cleavage.

In the modeling shown in Fig. 8, it is essential that the cortical shell be able to expand during cytokinesis because the surface area of the two daughter cells is greater than that of the original cell. The total volume is assumed to remain constant during cytokinesis. The cytoplasmic actin gel of the cortical shell can be readily expanded by slipping or releasing and reforming cross-linking proteins. The plasma bilipid layer of the membrane cannot increase its area more than 1% or 2% without rupture. It supplies the apparent increase in surface area by unfolding of wrinkles and protrusions on the surface of the cell.

4.1.2 Chromosome Control. In the early stages of cell division during so-called metaphase, Fig. 7(a), the chromosomes are aligned on the equatorial plane which later becomes the cleavage plane. The positions of the chromosomes appear to be controlled by microtubules which are the principal active element in the internal rearrangements during mitosis. Early in mitosis a bipolar structure of microtubules forms called the mitotic spindle. Microtubules radiate from two centers, called spindle poles. Proteins called kinetochores attach chromosomes to microtubules from the spindle poles. The two sets of chromosomes are pulled to the spindle poles by microtubule shortening. The poles themselves are pushed apart by sliding of microtubules which overlap, Fig. 7(b). Once two new nuclei are formed, microtubules are largely disassembled and cytokinesis proceeds. The magnitudes of forces and drags in these complex movements have not been evaluated to date.

4.2 Morphogenesis and Growth

Growth implies an increase of size which may or may not be accompanied by a change in form or topology. If only size changes, without any change of shape, the growth is said to be self-similar or isotropic, i.e., the growth is equal, percentagewise, in every direction and dimension. In many cases, a change of shape takes place without adding any new types of tissue nor by changing topology, but by unequal rates of growth in different directions. Such growth often results in different dimensions, a, b, both being functions of time and being related by $a=kb^m$. This is called allometric growth, where m is an allometric exponent and k is a constant. Allometric growth can transform a sphere into an ellipsoid, but not into a torus, which requires a topologic change.

Changes in topology and the diversification of groups of cells into different tissues are implied in the term morphogenesis. In various cases, morphogenesis and growth may proceed independently. In the frog, the fertilized egg (about 1 mm in diameter) divides repeatedly to become about 30 000 cells in about 10 h without an increase in mass. At this stage the embryo is called a gastrula which represents a considerable stage of morphogenesis although there is no growth in the sense of adding mass. Before gastrulation, the embryo already has developed some differentiation in that the external cells have tight junctions so the embryo is effectively sealed. Internal to the spherical embryo, cells are connected by gap junctions which may function as channels of communication.

The process of gastrulation is a major event which occurs in the development of all embryos. Gastrulation starts from a blastula which is a sphere with a distinct surface layer of epithelial cells. This epithelial layer is first invaginated and then proceeds through the interior of the cell to form, topologically, a torus. The central tube evolves into the gut. The interior of the torus evolves muscles, nerves, and other organs. The gastrula is topologically a torus, but its outer form is still more or less spherical with a thin tube down the middle (Akkas, 1990).

The next step in embryological development is formation of the neural tube (Fig. 9). The external layers of the dorsal ectoderm thicken and then fold to form a tube that later develops into the spinal cord and brain. The tube formation involves asymmetric growth and adhesions as outlined in the next section.

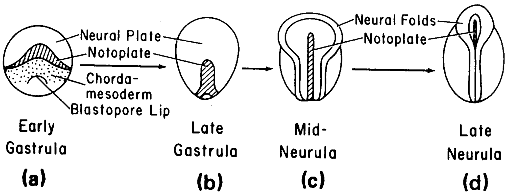

FIG. 9. Drawings of the neural plate and its folding from early gastrula to late neurula stages in the newt. (a) Posterior surface view of whole embryo. (b)–(d). Subsequent stages showing neural tube formation. (From Mittenthal and Jacobsen, 1990, by permission.)

4.2.1 Role of Adhesion. Cell-to-cell adhesion is a central mechanism to the coherence of tissues and to morphogenesis in several ways. Without adhesion, tissues would become liquid masses without resistance to indiscriminate mixing.

Adhesion not only holds tissues together, but may be an active force in sequestering similar cells. In a blastula, cells with lower adhesiveness will form the exterior epithelial layer. This idea comes from the observation of droplets of different liquids. A droplet of lower surface tension will tend to envelope a droplet of higher surface tension. In cellular sheets or masses, there is no appreciable surface tension of the kind found on liquid droplets (except in the lung where there is a moist air-tissue interface). In cellular sheets, the connection of the cytoskeletons of neighboring cells via desmosomes allows stress to be transmitted from cell to cell. Then the contractile mechanism (actin-myosin) within each cell can produce a tension in a sheet of cells, which on the macroscopic scale acts like a surface tension.

The contractile stress in a sheet of cells can give rise to a bending moment and curvature of the sheet if the plane in which the tension acts is not at the mid-depth of the sheet. A cortical tension in the outer half of an epithelial cell layer can give rise to a tendency for invagination as in gastrulation.

Differentials in adhesiveness between various parts of a sheet of cells can give rise to differential stretching in the plane of the sheet. Anisotropic tension fields can change the shape of cells involved. A compression in one direction may occur with elongation in the perpendicular direction. In this situation, another phenomenon may occur which is a rearrangement of cells by a change of neighbors. To illustrate the idea crudely, a square patch of 16 cells, oriented in 4 rows of 4 cells each, could be rearranged into two rows of 8 cells each. Such rearrangements occur during gastrulation and other morphogenetic events. This allows large changes in the form of a tissue without subjecting individual cells to the same degree of distortion.

4.2.2 Pattern Formation. Patterns of the structure and coloring of various organs come about through the interactive behavior of individual cells. Any information that leads a cell to perform an appropriate differentiation or pattern is called positional information. Tracking of cell groups from embryo to birth indicates that positional information is available early in development and is retained to much later stages.

Positional information is already present in the fertilized egg in the frog because it is not a uniform sphere, but has a yolk displaced from the center toward one pole which ultimately becomes the tail of the embryo. After gastrulation, further positional information is somehow recorded so that appropriate areas develop the forelimbs and hindlimbs and, internally, the various organs, glands, and circulatory and nervous systems.

One theory of the nature of positional information involves concentration gradients of a trigger or control substance called a morphogen. If such a substance is generated in one area, say in the head, and it is distributed through the embryo by diffusion, then positional information can be derived from the concentration of the morphogen at some time in the life of a cell. It appears that once positional information assignment is given to a group of cells, they retain the memory of their position indefinitely and develop into the appropriate type, size, and arrangement of cells. By transplanting parts of a wing bud, the local imprinting of developmental history is shown to be associated with the transplanted cells, rather than the local area to which they are transplanted (see Connelly *et al.*, 1981).

The existence of positional information and memory of it is quite clearly established, but the identification of the nature of the positional information and the manner of its interpretation is not yet entirely deciphered. For the limbs of the chick embryo, retinoic acid has been identified as a morphogen. But in most cases, the substance or nature of the positional signal is not clear. Similarly, it is clear that there must be a role for the expression of the genetic code in the recording and interpretation of positional information. It appears likely that positional information involves interpretation of the genetic code in such aspects as what parts or products should be suppressed and which parts should be utilized or expressed. The choice of the part or state of the genetic code interpretation is a likely form of cell memory of its positional information. But these details are not precisely known.

An interesting aspect of the use of positional information is a concept of successive refinement or local differentiation developed as positional information is utilized. Thus, as a limb bud grows, it develops fingers in the correct numbers and positions based on local relative positional information. In lower forms which reproduce limbs readily, a principle of intercalation has been suggested, which is that development proceeds so as to make positional values continuous. This provides local control to produce preset arrangements and sizes. In this way it is possible for a cell, by communication with its immediate neighbors, to perform the correct differentiation or tissue production to make the whole animal in all its complex parts follow a preset plan. In recent literature, various growth patterns have been analyzed by comparison to idealized fractal forms (see Ishizaka *et al.*, 1987). But the direct link of the genetic code to fractal instructions is not yet entirely clear.

In many cases, the results of interpretation of the genetic information depend on external factors being present. To begin with, water and nutrients must be available, but stresses due to gravity, muscles, and visual and audible signals all influence development. Thus both genetic and epigenetic (environmental) influences are necessary for normal development. Although the end results are attributable to individual cell behavior, theories of morphogenesis are usually based on continuum field theories and phenomenological equations. An extensive exposition and discussion of pattern formation, in both space and time, as controlled by reaction-diffusion equations is given in the book by Murray (1989).

GLOSSARY

Actin: A type of fiber found in most cells. One of the two principal components of muscle.

ATP: Adenosine triphosphate.

Axoneme: A tubular cylindrical structure which is the main element of flagella and cilia.

Blastula: An early stage of embryonic development consisting of a spherical mass of cells enclosed in an ectodermal layer.

Centrosomes: Small specialized structures in a cell from which microtubules extend.

Contractile Ring: A band of actin-myosin which girdles a cell and divides it by contracting.

Cytokinesis: The stage of cell division in which the parent cell is cleaved in two.

Cytoskeleton: The fibrous, solid components of interior cell structure.

Desmosome: A specialized connecting spot between two cells which joins fibers from each cell for stress transmission.

Dynein: A large molecule capable of producing force and motion along microtubules.

Ectoderm: The specialized outer layer of cells in a developing embryo.

Epigenetic: Pertaining to factors external to an animal or embryo which influence its development.

Erythrocyte: A red blood cell.

Eucaryote: A cell with a separate nucleus as in mammalian cells.

Flagellum: The tail of a sperm or microorganism which undulates to propel it in swimming.

Flickering: A small local motion of the red blood cell membrane due to Brownian motion.

Gastrulation: A process of embryonic development in which a spherical blastula evolves a central tube (gut).

Intermediate Filament: A fiber found in most cells forming ropelike networks.

Kinesin: A large molecule capable of producing force and motion along microtubules.

Leukocyte: A white blood cell.

Metaphase: A phase of the process of cell division in which chromosomes are aligned on a central plane.

Microtubule: A hollow cylindrical fiber composed of tubulin, found in most cells.

Mitosis: The process of cell division.

Morphogen: A substance whose distribution imparts positional information to cells in an embryo.

Morphogenesis: The development of a characteristic form or structure of a growing body.

Myosin: An active molecule which develops contractile force in muscle and other cells.

Procaryote: A cell without a separate membrane-coated nucleus, as in bacteria.

Pseudopod: An extension of a cell in the form of a rod or sheet, actively protruded.

Rouleau: A stack of red blood cells.

Stress Fibers: Strands of actin in cells or their extensions.

Treadmilling: A molecular motion in which molecules are added to one end of a fiber while they are removed from the other end.

Vesicle: A small, circular, membrane-covered particle (~ 60-nm diameter).

List of Works Cited

Akkas, N. (Ed.) (1987), *Biomechanics of Cell Division,* New York: Plenum.

Akkas, N. (1980), *J. Biomech.* **13**, 977–988.

Akkas, N. (Ed.) (1990), *Biomechanics of Active Movement and Deformation of Cells,* New York: Springer-Verlag.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J. D. (1989), *Molecular Biology of the Cell,* 2nd ed. New York: Garland.

Bongrand, P. (Ed.) (1988), *Physical Basis of Cell-Cell Adhesion,* Boca Raton, FL: CRC.

Brochard, F., Lennon, J. F. (1975), *J. Phys. (Paris)* **36**, 1035–1047.

Brokaw, C. J. (1989), in: F. D. Warner, P. Satir, and I. R. Gibbons (Eds.), *Cell Movement, Vol. I: The Dynein ATPases; Vol. II: Kinesin, Dynein, and Microtubule Dynamics,* New York: Liss, pp. 267–279.

Buican, T. N., Neagley, D. L., Morrison, W. C., Upham, B. D. (1989), in: G. C. Salzman (Ed.), *New Technologies in Cytometry,* SPIE Proceedings Vol. 1063. Bellingham, WA: SPIE, 190–197.

Chien, S., Sung, K.-L. P., Skalak, R., Usami, S., Tozeren, A. (1978), *Biophys. J.* **24**, 463–487.

Chien, S., Schmid-Schoenbein, G. W., Sung, K.-L. P., Schmaltzer, E. A., Skalak, R. (1984), in: H. Meiselman, M. A. Lichtman, and P. L. La Celle (Eds.), *White Cell Mechanics,* New York: Liss.

Cokelet, G. R. (1987), in: R. Skalak and S. Chien (Eds.), *Handbook of Bioengineering,* New York: McGraw-Hill, pp. 14.1–14.17.

Cokelet, G. R., Meiselman, H. J., Brooks, D. E. (Eds.) (1980), *Erythrocyte Mechanics and Blood Flow,* New York: Liss.

Connelly, T. G., Brinkley, L. L., Carlson, B. M. (1981), *Morphogenesis and Pattern Formation*, New York: Raven.

Dembo, M. (1989), *Biophys. J.* **55**, 1053–1080.

Dong, C., Skalak, R., Sung, K.-L. P., Schmid-Schoenbein, G. W., Chien, S. (1988), *J. Biomech. Eng.* **110**, 27–36.

Euteneuer, V., Schliwa, M. (1986), in: D. Soifer (Ed.), *Dynamic Aspects of Microtubule Biology*, New York: New York Academy of Sciences, pp. 867–886.

Evans, E. A. (1983), *Biophys. J.* **43**, 27–30.

Evans, E. A., Fung, Y. C. (1972), *Microvas. Res.* **4**, 335–347.

Evans, E. A., Skalak, R. (1980), *Mechanics and Thermodynamics of Biomembranes*, Boca Raton, FL: CRC.

Evans, E. A., Yeung, A. (1989), *Biophys. J.* **56**, 151–160.

Evans, E. A., Waugh, R., Melnik, L. (1976), *Biophys. J.* **16**, 585–595.

Fung, Y. C. (1981), *Biomechanics: Properties of Living Tissues*, New York: Springer-Verlag.

Hill, T. L. (1981), *Proc. Natl. Acad. Sci.* **78**, 5613–5617.

Hill, T. L., Kirschner, M. W. (1982), *Proc. Natl. Acad. Sci. U.S.A* **79**, 480–484.

Hochmuth, R. M. (1987), in: R. Skalak and S. Chien (Eds.), *Handbook of Bioengineering*, New York: McGraw-Hill, pp. 12.1–12.17.

Hochmuth, R. M., Worthy, P. R., Evans, E. A. (1979), *Biophys. J.* **26**, 101–114.

Holwill, M. E. J. (1977), in: T. J. Pedley (Ed.), *Scale Effects in Animal Locomotion*, New York: Academic, pp. 233–242.

Holwill, M. E. J. (1989), in: F. D. Warner, P. Satir, and I. R. Gibbons (Eds.), *Cell Movement, Vol. I: The Dynein ATPases*; *Vol. II: Kinesin, Dynein, and Microtubule Dynamics*, New York: Liss, pp. 199–208.

Ishikawa, H., Hatano, S., Sato, H. (Eds.) (1986), *Cell Motility: Mechanism and Regulation*, New York: Liss.

Ishizaka, S., Kato, Y., Takaki, R., Toriwaki, J. (Eds.) (1987), *Science and Form*, Tokyo: KTK Scientific.

Lighthill, M. J. (1975), *Mathematical Biofluiddynamics*, Philadelphia: Society for Industrial and Applied Mathematics.

Mittenthal, J. E., Jacobson, A. G. (1990), in: N. Akkas (Ed.), *Biomechanics of Active Movement and Deformation of Cells*, New York: Springer-Verlag, pp. 295–401.

Murray, J. D. (1989), *Mathematical Biology*, New York: Springer-Verlag.

Oster, G. F. (1989), in: W. Stein and F. Bronner (Eds.), *Cell Shape: Determinants, Regulation and Regulatory Role*, New York: Academic.

Oster, G. F., Perelson, A. S. (1987), *J. Cell Sci.* **8**, 35–54.

Oster, G. F., Perelson, A. S., Tilney, L. G. (1982), *J. Math. Biol.* **15**, 259–265.

Otter, T. (1989), in: F. D. Warner, P. Satir, and I. R. Gibbons (Eds.), *Cell Movement, Vol. I: The Dynein ATPases; Vol. II: Kinesin, Dynein, and Microtubule Dynamics*, New York: Liss, pp. 281–298.

Schneider, M. B., Jenkens, J. T., Webb, W. W. (1984), *J. Phys. (Paris)* **45**, 1457–1472.

Sheetz, M. P. (1990), *Cytometry*, Suppl. 4, p. 6.

Skalak, R. (1980), in: G. R. Cokelet, H. J. Meiselman, and D. E. Brooks (Eds.), *Erythrocyte Mechanics and Blood Flow*, New York: Liss, pp. 149–164.

Skalak, R. (1984), in: Y. C. Fung, E. Fukada, and W. Junjian (Eds.), *Biomechanics in China, Japan and U.S.A.*, Beijing: Science, pp. 108–119.

Skalak, R., Soslowsky, L., Schmalzer, E., Impelluso, T., Chien, S. (1987), *Biorheology* **24**, 35–52.

Skalak, R., Ozkaya, N., Skalak, T. C. (1989), *Annu. Rev. Fluid Mech.* **21**, 167–204.

Sleigh, M. A., Blake, J. R. (1977), in: T. J. Pedley (Ed.), *Size Effects in Animal Locomotion*, New York: Academic, pp. 243–256.

Soifer, D. (Ed.) (1986), *Dynamic Aspects of Microtubule Biology*, New York: New York Academy of Sciences.

Stokke, B. T., Mikkelsen, A., Elgsaeter, A. (1986), *Euro. Biophys. J.* **13**, 203–218; **13**, 219–233.

Stossel, T. P. (1982), *Philos. Trans. R. Soc. London Ser. B* **299**, 275–289.

Stracher, A. (Ed.) (1983), *Muscle and Nonmuscle Motility*, Vols. 1 and 2. New York: Academic.

Sugihara-Seki, M., Skalak, R. (1989), *Biorheology*, **26**, 261–277.

Sung, K.-L. P., Schmid-Schoenbein, G. W., Skalak, R., Schuessler, G. B., Usami, S., Chien, S. (1982), *Biophys. J.* **39**, 101–106.

Tözeren, A., Sung, K.-L. P., Chien, S. (1989), *BioPhys. J.* **55**, 479–493.

Vayo, M. (1986), *Theoretical and Experimental Analysis of Shear Disaggregation of Red Cell Rouleaux*, Doctoral dissertation, New York: Columbia University.

Wang, E., Fischman, D., Liem, R. K. H., Sun, T.-T. (Eds.) (1985), *Intermediate Filaments*, New York: New York Academy of Sciences.

Wang, Y.-L. (1985), *J. Cell Biol.* **101**, 597–602.

Warner, F. D., Satir, P., Gibbons, I. R. (Eds.) (1989), *Cell Movement, Vol. I: The Dynein ATPases; Vol. II: Kinesin, Dynein, and Microtubule Dynamics*, New York: Liss.

Waugh, R. E. (1982) *Biophys. J.* **38**, 29–37.

Zahalak, G. E., McConnaughey, W. B., Elson, E. L. (1990), *J. Biomech. Eng.* **112**, 283–294.

Zarda, P. R., Chien, S., Skalak, R. (1977), *J. Biomech.*, **10**, 211–221.

Zhu, C., Skalak, R. (1988), *Biophys. J.* **54**, 1115–1137.

Further Reading

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J. D. (1989), *Molecular Biology of the Cell*, 2nd ed. New York: Garland.

Bongrand, P. (1988), *Physical Basis of Cell-Cell Adhesion*, Boca Raton, FL: CRC.

Evans, E. A., Skalak, R. (1980), *Mechanics and Thermodynamics of Biomembranes*, Boca Raton, FL: CRC.

Skalak, R., Chien, S. (Eds.) (1987), *Handbook of Bioengineering*, New York: McGraw-Hill.

Warner, F. D., Satir, P., Gibbons, I. R. (Eds.) (1989), *Cell Movement, Vol. I: The Dynein ATPases; Vol. II: Kinesin, Dynein, and Microtubule Dynamics*, New York: Liss.

CERAMICS, MATERIAL PROPERTIES OF

James F. Shackelford, *Division of Materials Science and Engineering, University of California, Davis, California, U.S.A.*

Introduction 169
1. **Ceramic Science and Technology** 169
2. **Categories of Ceramic Materials** 170
2.1 Silicate Ceramics 170
2.2 Nonsilicate Oxide Ceramics ... 170
2.3 Nonoxide Ceramics 171
2.4 Glass–Ceramics 171
2.5 Ceramic Composite Systems... 172
2.5.1 Cermets 172
2.5.2 Ceramic-Matrix Composites ... 172
3. **Ceramic Processing** 173
3.1 Traditional Methods 173
3.2 Recent Advances 173
4. **Mechanical Properties** 174
4.1 Brittle Fracture 174
4.1.1 Elastic Constants 174
4.1.2 Fracture Strength 174
4.1.3 Fracture Mechanics 175
4.2 Static Fatigue 177
4.3 Creep Deformation 177
4.4 Thermal Shock 178
4.5 Viscoelastic Deformation of Glassy Phases 179
5. **Optical Properties** 180
5.1 Refraction 180
5.2 Reflection 180
5.3 Transmission 181
6. **Electrical Properties** 182
6.1 Conductivity of Traditional Ceramics 182
6.2 Superconducting Ceramics.... 182
6.3 Dielectrics 183
6.4 Ferroelectrics and Piezoelectrics 183
7. **Magnetic Properties** 185
7.1 Ferrimagnetism 185
7.2 Superconducting Magnets..... 186
Glossary 187
Works Cited 187
Further Reading 187

INTRODUCTION

Ceramic materials are the source of some of our most ancient cultural artifacts as well as some of the most advanced technologies of the late 20th century (Kingery and Lense, 1985). Ceramics are traditionally defined as the products of firing nonmetallic minerals at high temperatures. In the field of materials science and engineering, these ceramic products are often defined by what they are not, viz., a ceramic is a nonmetallic, inorganic solid material. This latter definition distinguishes the ceramic from metals or polymers, the other two basic categories of engineered materials. A brief overview of ceramic science and technology will be followed by a systematic discussion of the various categories of these important materials.

1. CERAMIC SCIENCE AND TECHNOLOGY

The scientific foundation of ceramic technology is widely considered to have been most effectively outlined by Kingery (1960). This classic treatment was updated and expanded by Kingery, Bowen, and Uhlmann (1976). The nature of ceramic materials can be initially appreciated by considering a typical oxide, such as Al_2O_3. As an oxide compound of metallic aluminum, the ceramic has two principal advantages, viz., chemical stability in a wide variety of severe environments and a significantly higher melting point than the unoxidized metal (2020 °C compared to 660 °C). As such, Al_2O_3 is a widely used refractory, i.e., a high-temperature-resistant material. The superior chemical and temperature-resistant properties of ceramics are countered by tra-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ditionally inferior mechanical behavior, especially a characteristic brittleness. Structural applications of ceramics have been severely limited as a result. Substantial progress has been made, however, in the past two decades in producing ceramic materials with increased strength and resistance to fracture. In Sec. 4.1.3, the important concept of fracture toughness will be introduced as a central parameter in the expanded role of structural ceramics. Perhaps the most celebrated development in this area has been the research on candidate materials for the "ceramic engine," in which adequate mechanical properties are available to permit metal replacement in the design of an automotive power plant. In this configuration, the inherent temperature resistance of the ceramics allows significantly higher engine operating temperatures and correspondingly higher Carnot efficiencies. In Sec. 3.2, the advances in ceramic-powder processing technology necessary to produce this new generation of structural materials will be briefly reviewed.

An important feature of many ceramic materials, especially the silicates, is an ability to be formed easily in a noncrystalline state. The resulting vitreous or glassy materials are the subject of separate articles in the *Encyclopedia* and will not, in general, be discussed here. An exception will be the glass–ceramics, a relatively new family of ceramic materials in which the product shape is formed in the glassy state and is then devitrified or crystallized by an appropriate thermal treatment. Relatively complex product shapes are possible due to the glass-phase forming step, and good mechanical properties result from the fine-grained, low-porosity microstructure produced by the devitrification step. Excellent thermal shock resistance is produced by the use of low-thermal-expansion coefficient compositions, e.g., lithium aluminosilicates.

As with traditional silicate glasses, many crystalline ceramics have applications based on light transmission. This and other optical properties will be discussed in Sec. 5. Ceramic materials also have numerous, traditional applications based on their electronic and magnetic properties. Ionic and covalent bonding produces characteristically low electrical conductivity in traditional ceramics and resulting applications as dielectrics. Various ceramic compositions have commercially important magnetic behavior (see MAGNETIC MATERIALS). The ferrites are common examples. Perhaps no more dramatic development has occurred in applied physics during the past decade than the discovery of superconducting behavior in certain oxide ceramics (Strongin *et al.*, 1987) (see HIGH-TEMPERATURE SUPERCONDUCTIVITY). These "high-temperature" superconductors raise the possibility of significantly new technologies based largely on their associated magnetic behavior. Thin-film devices appear to be the most likely initial area of development. These recent advances will be discussed in Secs. 6 and 7, along with more traditional electrical and magnetic properties.

2. CATEGORIES OF CERAMIC MATERIALS

Ceramic materials can be conveniently categorized by chemical composition and atomic- and microscopic-scale structure. The first three categories are chemical in nature, with the last two based on structure.

2.1 Silicate Ceramics

Silicon and oxygen together account for approximately 75% of the elemental composition of the Earth's crust. As a result, the SiO_2-based silicates are abundant and economical raw materials. Many of the traditional ceramics fall within this category. Pottery and related whiteware ceramics have been commercial products since roughly 4000 B.C. Common brick, tile, and sewer pipe constitute the structural clay products. This wide range of silicate ceramics reflects the diversity of the silicate minerals available to local manufacturing plants. Table 1 summarizes the general compositions of typical silicate ceramics.

2.2 Nonsilicate Oxide Ceramics

Table 2 gives examples of various nonsilicate oxide ceramics. Magnesia (MgO) is a traditional refractory material widely used in the steel industry. Many of the nonsilicate oxides, however, are among the more advanced structural ceramics. Many are commercially available as essentially single-compound compositions, although impurity levels can range from a few parts per million (ppm) for so-called "advanced ceramics" or "fine ceramics" to several weight percent in the

Table 1. Compositions of some silicate ceramics.

Ceramic	Composition[a] (wt %) SiO_2	Al_2O_3	K_2O	MgO	CaO	Others
Silica refractory	96					4
Fire-clay refractory	50–70	45–25				5
Mullite refractory	28	72				...
Electrical porcelain	61	32	6			1
Steatite porcelain	64	5		30		1
Portland cement	25	9			64	2

[a] Approximate, indicating primary components. Impurity levels can vary significantly.

industrial refractories. (Although both are commonly used, the term "advanced ceramics" is gaining wider favor than "fine ceramics" to label these contemporary materials.) Chemical separation and subsequent processing to produce relatively high-purity advanced ceramics results in a substantially more costly product compared to the silicate ceramics made from locally available and generally impure minerals. It is useful to note that even the advanced ceramics are sometimes referred to by traditional mineral names, e.g., "corundum" for Al_2O_3.

Table 2. Some nonsilicate oxide ceramics.

Primary composition[a]	Common product names
Al_2O_3	Alumina, alumina refractory
MgO	Magnesia, magnesia refractory, magnesite refractory, periclase refractory
$MgAl_2O_4$ (= $MgO \cdot Al_2O_3$)	Spinel
BeO	Beryllia
ThO_2	Thoria
UO_2	Uranium dioxide
ZrO_2 (stabilized[b] with CaO)	Stabilized (or partially stabilized) zirconia
$BaTiO_3$	Barium titanate
$NiFe_2O_4$	Nickel ferrite

[a] Some products, such as the industrial refractories, may have several weight percent oxide additions and impurities.

[b] Pure ZrO_2 has a phase transformation at 1000 °C in which the crystal structure change produces a catastrophic volume change. A 10-wt % CaO addition produces a cubic crystal structure stable to the melting point (~2500 °C), making "stabilized" zirconia a highly useful refractory. Lesser CaO additions can produce a two-phase microstructure, with cubic zironica being one of the phases. This "partially stabilized" zirconia has superior fracture toughness (Sec. 4.1.3).

2.3 Nonoxide Ceramics

Some nonoxide ceramics, such as silicon carbide, have been common industrial materials for several decades (Hampshire, 1986). SiC is widely used as a furnace-heating element and as an abrasive material. Other nonoxide ceramics, such as silicon nitride, have been the focus of substantial research and development in the past two decades as candidates for advanced structural ceramics. As with the use of mineral names mentioned in the preceding section, commercial names are sometimes used to identify these materials, e.g., "carborundum" for SiC. Table 3 lists various examples of nonoxide ceramics.

2.4 Glass–Ceramics

Glass–ceramics were among the early advanced ceramics (McMillan, 1979). As the name implies, they begin as relatively ordinary glassware (see GLASSES), and, by a care-

Table 3. Some nonoxide ceramics.

Primary composition[a]	Common product names
SiC	Silicon carbide
Si_3N_4	Silicon nitride
TiC	Titanium carbide
TaC	Tantalum carbide
WC	Tungsten carbide
B_4C	Boron carbide
BN	Boron nitride
C	Graphite

[a] Some products may have several weight percent additions or impurities.

Table 4. Compositions of some glass–ceramics.

Glass–Ceramic	Composition (wt %)							
	SiO_2	Li_2O	Al_2O_3	MgO	ZnO	B_2O_3	TiO_2 [a]	P_2O_5 [a]
Li_2O-Al_2O_3-SiO_2 system	74	4	16				6	
MgO-Al_2O_3-SiO_2 system	65		19	9			7	
Li_2O-MgO-SiO_2 system	73	11		7		6		3
Li_2O-ZnO-SiO_2 system	58	23			16			3

Source: Data from McMillan (1979).
[a] Nucleating agents.

fully controlled heat treatment, over 90% of the material crystallizes. The average crystallite grain size is characteristically between 0.1 and 1 μm. A small amount of residual glassy phase remains at the grain boundaries, helping to create a pore-free microstructure. As will be seen in Sec. 4.1, this typical glass–ceramic microstructure lends itself to superior mechanical impact resistance. The characteristic compositions of typical glass–ceramics shown in Table 4 are associated with low-thermal-expansion-coefficient materials and resulting superior thermal shock resistance, a topic discussed further in Sec. 4.4.

The benefits of the uniformly fine-grained microstructures of glass–ceramics can be ascribed indirectly to the compositions in Table 4. One should note that TiO_2 and P_2O_5 are nucleating agents. A fine dispersion of these components provides a high density of heterogeneous nucleation sites for the glass devitrification process.

The most important commercial glass–ceramic system in Table 4 is the Li_2O-Al_2O_3-SiO_2 system. The low thermal expansion coefficient is produced by the presence of crystallites of β-spodumene ($Li_2O \cdot Al_2O_3 \cdot 4SiO_2$) with a small expansion coefficient; or β-eucryptite ($Li_2O \cdot Al_2O_3 \cdot SiO_2$), which actually has a negative expansion coefficient.

2.5 Ceramic Composite Systems

Ceramic composites can be subdivided into two categories, viz., traditional cermets, in which ceramic and metallic phases are dispersed on the microscopic scale, and the relatively recent development of ceramic–matrix composites (CMCs), in which various dispersed phases are used to improve the mechanical behavior of structural ceramic materials.

2.5.1 Cermets. Ceramic–metal composite systems, known as cermets, became the source of intense research during and just following World War II for the purpose of developing materials incorporating the most desirable characteristics of both ceramics and metals (Tinklepaugh and Crandall, 1960). A special focus was the substitution for critical alloys in aircraft engines. Although largely unsuccessful for that purpose, the general combination of ceramic and metallic components on a microstructural scale has been the basis of numerous, useful materials for cutting tools, nuclear fuels, and electronic devices (Upadhyaya, 1984). Some typical cermets are listed in Table 5.

2.5.2 Ceramic–Matrix Composites. Ceramic–matrix composites (CMCs) are relatively recent developments in the area of advanced materials (see COMPOSITE MATERIALS). A primary driving force for their development is superior high-temperature resistance, e.g., high-efficiency engine designs (Lemkey *et al.*, 1988). Some typical ceramic–matrix composite systems are given in Table 6. As will be discussed in Sec. 4.1.3, a key focus of research on CMCs is their potential for improved fracture toughness. CMCs fall into two categories: fiber-reinforced and particulate- and whisker-containing. The fiber-reinforced configuration

Table 5. Various cermets.

Ceramic–Metal
Al_2O_3-Cr
B_4C-Al
TiC-Ni
WC-Co
WC-TiC-Co

Source: Data from Tinklepaugh and Crandall (1960) and Upadhyaya (1984).

Table 6. Various ceramic–matrix composite (CMC) systems.

Fiber–matrix	Particulate–matrix	Whisker–matrix
Nb/$MoSi_2$	TiC/Al_2O_3	SiC/Al_2O_3
C/C		
C/SiC		
SiC/Al_2O_3		
SiC/SiC		
SiC/Si_3N_4		
SiC/Li-Al-silicate (glass-ceramic)		

Source: Data from K. K. Chawla, University of Illinois, Urbana-Champaign; A. K. Dhingra, E. I. duPont de Nemours & Co., Inc.; A. J. Klein, ASM International; S. T. Buljan, A. E. Pasto, and H. J. Kim, GTE Laboratories, Inc.

represents a more challenging processing technology, but with the possibility of non-catastrophic failure if mechanical behavior is fiber-dominated. (In regard to the mechanical failure of engineered materials, the term "catastrophic" implies rapid failure without significant prior warning. In the case of fiber-reinforced CMCs, fiber pull-out is a time-consuming mechanism that provides this prior warning.) To date, the mechanical advantages of fiber-reinforced CMCs have been limited to relatively low-temperature applications. Particulate and whisker CMCs are easier to process but tend to fail catastrophically by a matrix-controlled mechanism. It is useful to note that many properties of composite systems can be described as averages of the properties of the individual component phases. It is equally important to note, however, that the specific nature of the "average" is a sensitive function of the microstructural geometry of the composite and the nature of the fiber–matrix interface. The interfacial issue is discussed further in Sec. 4.1.3.

3. CERAMIC PROCESSING

As noted at the outset, ceramic materials represent some of our most ancient as well as some of our most advanced materials. As a result, this discussion of ceramic processing will be divided into a section dealing with traditional methods and another dealing with recent technologies needed to develop the more modern materials. It is important to appreciate that such a discussion is due to the general effect of processing on materials properties. For example, properties as diverse as both mechanical strength and optical transparency can be seriously degraded by residual porosity, a common feature of traditional ceramic products.

3.1 Traditional Methods

Ceramic processing is both an ancient art and a relatively young applied science (Reed, 1988). The processing technology for ceramic materials is generally affected by the combined characteristics of mechanical brittleness and relatively high melting points. For example, fusion casting is similar to metal casting but is limited in application by high melting points. Some low-porosity refractories are formed in this way, but at a relatively high cost. More common for traditional clayware ceramics is slip casting. This room-temperature casting involves a "slip," or clay–water mixture poured into a porous mold. Water absorption by the mold produces a relatively rigid powder that can be removed from the mold. After a drying step in which the remaining absorbed water is driven off at moderate temperatures, a structurally viable product is produced by firing at elevated temperatures (typically above 1000 °C). Much of the strength of the fired piece is due to solid-state diffusion. Various phase transformations can also accompany the firing step, especially the production of glassy phases in many silicates. Sintering, defined as the bonding of powder particles by solid-state diffusion, is a component of the firing step following slip casting. The production of many nonsilicate oxide ceramics and nonoxide ceramics is done directly by the sintering of a pressed powder compact. The typically high melting points of these materials make sintering a widespread processing technique.

3.2 Recent Advances

Hot pressing is basically sintering under an applied mechanical load. Similarly, hot isostatic pressing (or "HIPing") is sintering under an applied, isostatic pressure from an inert gas. These technologies are increasingly popular for advanced ceramics, especially in providing fully dense products with superior mechanical properties. Sol–gel processing is among the most rapidly developing technologies for ceramic processing. The method provides the formation of uniform, fine particulates of high purity at relatively low temperatures. Such powders can subsequently be sintered to high density with correspondingly good mechanical properties. The essential feature is the formation of an organometallic solution. The dispersed-phase "sol" is then converted into a rigid "gel" which, in turn, is reduced to a final composition by various thermal treatments. In chemical-vapor deposition (see CVD), a source material is vaporized and thermally decomposed or reacted with a gas or vapor phase to produce a reaction product deposited onto a substrate. The CVD process corresponds to a molecular-scale buildup of a thin-film configuration. It has the capability of forming high-density, uniform films on complex-shaped surfaces. Substantial work is being done on the development of a wide variety of novel processing techniques for ceramic materials. As an example, self-propagating high-temperature synthesis (SHS) uses a highly exothermic energy of reaction to produce the necessary sintering temperature for the final ceramic product (Munir, 1988). Once initiated, a combustion wave can pass through a powder, providing both a chemical reaction and the densification of the compact. A simple example is the formation of titanium nitride by the reaction between titanium powder and nitrogen gas.

4. MECHANICAL PROPERTIES

The various mechanical properties discussed in this section are not an exhaustive listing but do represent the major factors in selecting a durable material for structural applications under a variety of service conditions.

4.1 Brittle Fracture

Perhaps the most distinctive characteristic of traditional ceramics is their tendency to exhibit brittle fracture (Evans, 1984). Such failure with the absence of significant ductility has represented a primary obstacle to be overcome in the application of the newer structural ceramics. In the sections that follow, the specific discussion of fracture is preceded by treatment of the elastic constants associated with the deformation prior to failure. The discussion of fracture strength is followed by the equally important topic of fracture toughness, the key property associated with the resistance of the material to flaw-induced failure (Ashby and Jones, 1980). This important factor in the application of structural ceramics comes from the general subject of fracture mechanics.

4.1.1 Elastic Constants. Unlike the cases for relatively ductile metals and polymers, the deformation of ceramics is generally elastic in nature up to the point of fracture. The mechanical behavior of a structural ceramic in this regime can be adequately characterized by two elastic constants, the elastic (or Young's) modulus and Poisson's ratio. The linear relationship between the applied stress σ and the resulting strain ϵ is given by Hooke's law:

$$\sigma = E\epsilon, \tag{1}$$

where E is the elastic (or Young's) modulus. Values of the modulus of elasticity for a variety of ceramics are given in Table 7. Poisson's ratio ν is defined as the negative of the contraction perpendicular to the extension caused by a tensile stress. Specifically,

$$\nu = -\epsilon_x/\epsilon_z, \tag{2}$$

where the strain in the x direction is perpendicular to that in the z direction, the direction of the tensile stress. Values of Poisson's ratio are given in Table 7.

4.1.2 Fracture Strength. The fracture strength can be defined simply as the tensile stress value at which fracture suddenly occurs following the elastic deformation of the ceramic. As a practical matter, a more reproducible fracture strength is obtained in a bending mode than in pure tension. For simple three-point bending, the flexural strength, FS (or modulus of rupture, MOR) is given by

$$\mathrm{FS} = \mathrm{MOR} = 3FL/(2bh^2), \tag{3}$$

where F is the force applied to a rectangular-bar-shaped specimen, L the separation distance between the two lower supports (each of which provides a reaction force, $F/2$), b the width of the specimen, and h the height. Typical values of FS or MOR are given in Table 8.

Table 7. Modulus of elasticity and Poisson's ratio for some ceramics.

	E (MPa)	v
Mullite (aluminosilicate) porcelain	69×10^3	0.25
Steatite (magnesia aluminosilicate) porcelain	69×10^3	
Superduty fire-clay (aluminosilicate) brick	97×10^3	
Alumina (Al_2O_3) crystals	380×10^3	0.26
Sintered alumina (~5% porosity)	370×10^3	
Alumina porcelain (90–95% alumina)	370×10^3	
Sintered magnesia (~5% porosity)	210×10^3	
Magnesite (magnesia) brick	170×10^3	
Sintered spinel (magnesia-aluminate) (~5% porosity)	238×10^3	
Sintered stabilized zirconia (~5% porosity)	150×10^3	0.23–0.32
Sintered beryllia (~5% porosity)	310×10^3	0.26
Dense silicon carbide (~5% porosity)	470×10^3	0.19
Bonded silicon carbide (~20% porosity)	340×10^3	
Hot-pressed boron carbide (~5% porosity)	290×10^3	
Hot-pressed boron nitride (~5% porosity)	83×10^3	

Source: Kingery, Bowen, and Uhlmann (1976), and data from Smothers (1985, 1986).

The magnitudes of the values in Table 8 correspond to those for tensile strength, as the failure mode in bending is tensile (along the outermost edge of the sample opposite the applied force). It is important to note that, for ceramics, the magnitude of fracture strength in a compressive mode is often substantially higher than in tension. Differences as much as an order of magnitude can be appreciated from the simple Griffith crack model. Assuming the presence of elliptical cracks at the surface and/or in the interior of the material, the Griffith model predicts the amount of stress concentration at a given crack tip to be

$$\sigma_m/\sigma \cong 2(c/\rho)^{1/2}, \qquad (4)$$

where σ_m is the highest stress at the tip of such a crack, σ the applied tensile stress, c the crack length, and ρ the radius of the crack tip. As the crack tip radius can be on the order of an interatomic spacing, the stress intensification can be quite large. A compressive load tends to close, not open, the Griffith flaws and consequently does not diminish the inherent strength of the material.

4.1.3 Fracture Mechanics. The term *fracture mechanics* refers to the general analysis

Table 8. Flexural strength (modulus of rupture) for some ceramics.

	MOR (MPa)
Mullite (aluminosilicate) porcelain	69
Steatite (magnesia alumino silicate) porcelain	140
Superduty fire-clay (aluminosilicate) brick	5.2
Alumina (Al_2O_3) crystals	340–1000
Sintered alumina (~5% porosity)	210–340
Alumina porcelain (90%–95% alumina)	340
Sintered magnesia (~5% porosity)	100
Magnesite (magnesia) brick	28
Sintered spinel (magnesia-aluminate) (~5% porosity)	90
Sintered stabilized zirconia (~5% porosity)	83
Sintered beryllia (~5% porosity)	140—280
Dense silicon carbide (~5% porosity)	170
Bonded silicon carbide (~20% porosity)	14
Hot-pressed boron carbide (~5% porosity)	340
Hot-pressed boron nitride (~5% porosity)	48–100

Source: Kingery, Bowen, and Uhlmann (1976).

Table 9. Typical values of fracture toughness (K_{IC}) for various ceramics.

Material	K_{IC} (MPa $m^{1/2}$)
Partially stabilized zirconia (PSZ)	9
Electrical porcelain	1
Alumina (Al_2O_3)	3–5
Magnesia (MgO)	3
Cement/concrete, unreinforced	0.2
Silicon carbide (SiC)	3
Silicon nitride (Si_3N_4)	4–5
SiC whiskers in Al_2O_3	8.7
SiC fibers in SiC	25
SiC whiskers in reaction-bonded Si_3N_4	25

Source: Data from GTE Laboratories, Waltham, MA; Ashby and Jones (1980); Klein (1986).

of failure of structural materials with preexisting flaws. The most widely used single parameter from this field is the fracture toughness K_{IC}, the critical value of the stress–intensity factor at a crack tip necessary to produce catastrophic failure. In general, the value of fracture toughness is given by

$$K_{IC}=Y\sigma_f(\pi a)^{1/2}, \qquad (5)$$

where Y is a dimensionless geometry factor on the order of 1, σ_f is the overall applied stress at failure, and a is the length of a surface crack (or one-half the length of an internal crack). It should be noted that K_{IC} is associated with plane-strain conditions, and that the subscript I refers to "mode I" (or uniaxial tensile) loading and the subscript C to critical (or catastrophic) failure. The similarity to the Griffith crack model equation is obvious. The fracture-toughness expression of Eq. (5) is, however, not limited to purely brittle materials. In fact, one of the primary uses of fracture-toughness data for structural ceramics is the comparison with values for the common structural metals. Typical values of fracture toughness for a variety of ceramic materials are given in Table 9.

One should note that, for traditional ceramics, K_{IC} is typically ≤ 5 MPa $m^{1/2}$. In contrast to these relatively low values, typical aluminum alloys have K_{IC} values of between 25 and 45 MPa $m^{1/2}$ and pressure-vessel steels have values on the order of 170 MPa $m^{1/2}$. Values of 5 MPa $m^{1/2}$ for Al_2O_3 and Si_3N_4 are associated with careful processing, e.g., the elimination of residual porosity by hot pressing or hot isostatic pressing. Such resulting fracture toughness values represent substantial improvement over the values of ≤ 1 MPa $m^{1/2}$ for many conventionally sintered ceramics.

In order to make structural ceramics more competitive for applications traditionally dominated by metallic materials, K_{IC} values greater than 5 MPa $m^{1/2}$ are needed. One technique has involved the purposeful introduction of microcracks which serve the purpose of blunting the tip of an advancing crack. The Griffith expression [Eq. (4)] indicates that the larger effective tip radius dramatically reduces the local stress at the crack tip. The value of 9 MPa $m^{1/2}$ for partially stabilized zirconia (PSZ) in Table 9 is due to another microstructural mechanism known as transformation toughening and illustrated in Fig. 1. In this case, improved toughness is due to second-phase particles of tetragonal zirconia in a matrix of cubic zirconia. A propagating crack creates a local stress field that induces a transformation of tetragonal zirconia particles to the monoclinic structure in that vicinity. The slightly larger specific volume of the monoclinic phase causes an effective compressive load locally and, in turn, the "squeezing" of the crack shut.

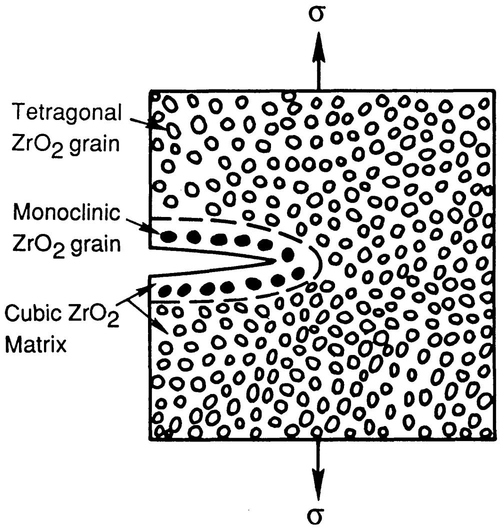

FIG. 1. Transformation toughening of partially stabilized zirconia (PSZ) involves the stress-induced transformation of tetragonal grains to the monoclinic structure, which has a larger specific volume. The result is a local volume expansion at the crack tip, squeezing the crack shut.

As noted in Sec. 2.5.2, many of the material properties of composite systems can be predicted as averages of the properties of the individual components (see COMPOSITE MATERIALS). For this property averaging to be in effect, however, the dispersed and matrix phases must be adequately bonded together. For example, the interface between the matrix and a discontinuous fiber-reinforcing phase must be strong enough to transmit a mechanical load from one phase to the other. In ductile–matrix composites, a high interfacial strength is desirable to maximize the overall composite strength, as failure generally originates in or along the reinforcing fibers. In ceramic–matrix composites, however, failure generally originates in the matrix phase. In order to maximize the fracture toughness of these materials, it is desirable to have a relatively weak interfacial bond allowing fibers to pull out. A crack initiated in the matrix can then be deflected along the fiber–matrix interface. Figure 2 illustrates the mechanism of fiber pull-out for improving fracture toughness. As can be seen in Table 9, some of the fiber-reinforced ceramic–matrix composites have fracture-toughness values approaching those of some conventional metallic alloys.

4.2 Static Fatigue

Analogous to the "fatigue" of metals (loss of strength over time due to cyclic loading), ceramics can display a loss of strength over time but without cyclic loading. This static fatigue is the result of crack growth due to chemical reaction with the environment. The classic example is crack extension in silicates by the breaking of Si–O bonds by reaction with H_2O molecules from atmospheric moisture. The mechanism involving reaction with atmospheric moisture confines this effect to the vicinity of room temperature.

4.3 Creep Deformation

Creep can be defined as plastic deformation occurring at a relatively high temperature (above about one-half the absolute melting point) under constant load over a long time period (see RHEOLOGY). A schematic creep curve is shown in Fig. 3. After the initial elastic deformation at $t \cong 0$, there are three stages of creep deformation: the primary stage characterized by a decreasing strain rate, the secondary stage corresponding to a constant strain rate, and the final (tertiary) stage in which the strain rate increases. The secondary-stage (steady-state) strain rate is a useful parameter to characterize the overall creep behavior of the ceramic and represents a balance between the increased ductility mechanisms of the primary stage and the increased resistance to deformation in the tertiary stage.

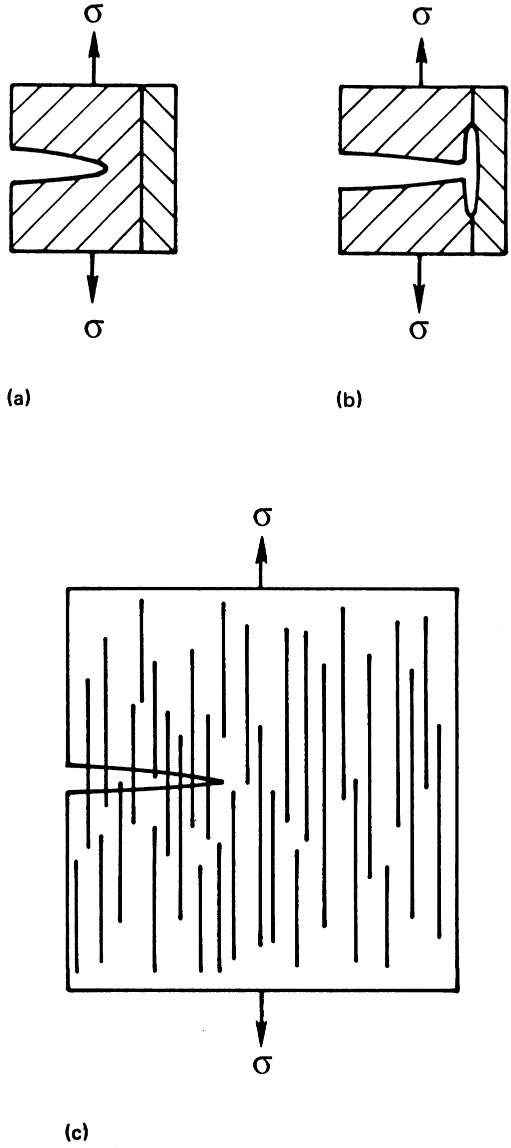

FIG. 2. For a close-up view of a ceramic-matrix composite, (a) a matrix crack approaching a fiber is (b) deflected along the fiber-matrix interface. For the overall composite, (c) the increased crack path length due to fiber pull-out significantly improves fracture toughness.

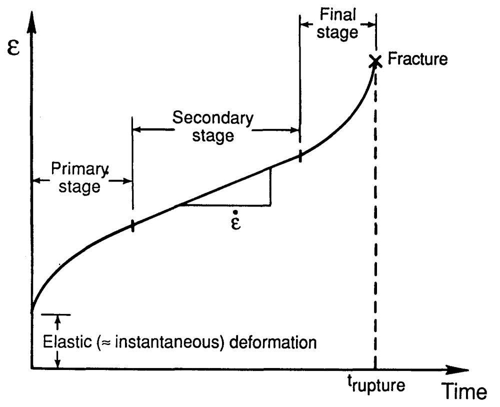

FIG. 3. Creep curve indicating the plastic strain occurring over time for a ceramic stressed at high temperatures (above about one-half the absolute melting point).

Another key parameter for characterizing creep behavior is the overall time to creep rupture.

Diffusional mechanisms (e.g., dislocation climb) are often less important in the creep deformation of ceramics than in metals, largely because of the greater complexity of diffusion in ceramic compounds. As a result, grain-boundary sliding (see GRAIN BOUNDARIES) frequently plays a dominant role in the creep of ceramics. In some relatively impure refractory ceramics, a substantial silicate glassy phase may form at the grain boundary, allowing for a creep mechanism of grain-boundary sliding by viscous deformation of the glassy phase. Such deformation is generally undesirable for a material intended for high-temperature structural applications.

The general dependence of the steady-state creep rate ($\dot{\epsilon}$) on both stress (σ) and absolute temperature (T) can be expressed as

$$\dot{\epsilon}=C\sigma^{n}e^{-Q/RT}, \qquad (6)$$

where C is a constant, n a constant typically between 3 and 8, Q the activation energy for the creep deformation mechanism, and R the universal gas constant. The preexponential term in Eq. (6) is essentially temperature independent, and the presence of the σ^n term gives the name "power-law" creep to the overall expression. Creep-rate data for some common ceramics at a fixed temperature are given in Table 10.

4.4 Thermal Shock

The routine use of structural ceramics at elevated temperatures in combination with inherent brittleness creates special concern about thermal shock, defined as fracture (partial or complete) as a result of a temperature change (usually sudden cooling). The mechanism of thermal shock involves two more fundamental thermal properties:

1. thermal expansion and
2. thermal conductivity.

The linear coefficient of thermal expansion α is given by

$$\alpha=(1/L)dL/dT, \qquad (7)$$

with L being specimen length and T temperature. The thermal conductivity k is defined by Fourier's law of heat conduction:

$$k=-\frac{dQ/dt}{A(dT/dx)}, \qquad (8)$$

where dQ/dT is the rate of heat transfer across an area A due to a temperature gradient dT/dx. Thermal expansion and thermal conductivity data for various ceramics are given in Table 11.

Figure 4 illustrates schematically a common mechanism of thermal shock for structural ceramics. The rapid cooling of the surface of a slab of material initially at an elevated temperature leads to surface tensile stresses. With the inevitable presence of Griffith flaws at the surface, this surface tensile stress creates the potential for brittle fracture.

Table 10. Creep-rate data for various polycrystalline ceramics.

Material	$\dot{\epsilon}$ at 1300 °C, 12.4 MPa [mm/(mm h) $\times 10^6$]
Al_2O_3	1.3
BeO	300
MgO (slip cast)	330
MgO (hydrostatic pressed)	33
$MgAl_2O_4$ (2–5 μm)[a]	263
$MgAl_2O_4$ (1–3 mm)[a]	1
ThO_2	1000
ZrO_2 (stabilized)	30

Source: Kingery, Bowen, and Uhlmann (1976).

[a] The grain-boundary sliding mechanism causes a substantially greater degree of creep deformation in the finer-grained $MgAl_2O_4$ specimen.

Table 11. Mean thermal expansion coefficients and thermal conductivity for various ceramics.

	Linear coefficient of thermal expansion, 0–1000 °C [mm/(mm °C) × 10^6]	Thermal conductivity [J/(s m K)]	
		100 °C	1000 °C
Mullite ($3Al_2O_3\cdot 2SiO_2$)	5.3	5.9	3.8
Porcelain	6.0	1.7	1.9
Fire-clay refractory	5.5	1.1	1.5
Al_2O_3	8.8	30	6.3
Spinel ($MgO\cdot Al_2O_3$)	7.6	15	5.9
BeO	9.0	219	20
MgO	13.5	38	7.1
ThO_2	9.2	10	2.9
UO_2	10	10	3.3
ZrO_2 (stabilized)	10	2	2.3
SiC	4.7	...	...
TiC	7.4	25	5.9
B_4C	4.5	...	...

Source: Kingery, Bowen, and Uhlmann (1976).

The ability of a given structural ceramic to withstand a given temperature change depends on a complex combination of thermal expansion, thermal conductivity, overall geometry, and the inherent brittleness of the material.

4.5 Viscoelastic Deformation of Glassy Phases

As noted in the discussion of creep deformation, glassy grain-boundary phases in otherwise crystalline ceramics can play a critical role in mechanical behavior (Shackelford, 1987). A key parameter in describing the mechanical deformation of a glassy phase is the glass transition temperature T_g, defined as the temperature below which the material tends to behave as a true glass (a rigid solid) and above which the material tends to behave as a supercooled liquid (see GLASSES). Below T_g, the glass exhibits elastic deformation, and above T_g, the supercooled liquid exhibits viscous deformation. It should be noted that, unlike the melting point, T_g represents a narrow transition temperature range rather than a single temperature.

There are two viscosity levels that are used to define characteristic temperatures in the viscous deformation range in traditional glass-making technology. These can be useful reference points for determining the deformation behavior of a secondary glassy phase. The annealing point is the temperature at which the viscosity is equal to $10^{13.4}$ P, where P stands for the unit poise [$=1$ g/(cm s) $=0.1$ Pa s]. The practical significance of the annealing point is that internal stresses can be relieved in about 15 min at this viscosity level. The glass transition temperature occurs around the annealing point. The softening point is the

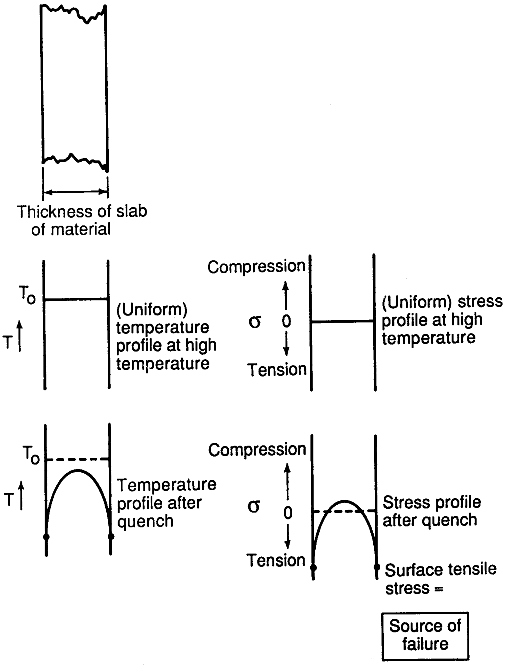
Thickness of slab of material
T_0
T
(Uniform) temperature profile at high temperature
Compression
σ
0
Tension
(Uniform) stress profile at high temperature
T_0
T
Temperature profile after quench
Compression
σ
0
Tension
Stress profile after quench
Surface tensile stress =
Source of failure

FIG. 4. Thermal shock resulting from temperature gradients created by a finite thermal conductivity. Rapid cooling produces surface tensile stresses.

Table 12. Viscosity values for various glassy phases.

Glass	Annealing point (°C)	Softening point (°C)
Na_2O-CaO-SiO_2	514	696
Na_2O-B_2O_3-SiO_2	544	780
Na_2O-Al_2O_3-SiO_2	712	915
$3B_2O_3$-$97SiO_2$	1020	1530
SiO_2	1084	1580

Source: Data from Corning Glass Works, Corning, NY

temperature at which the viscosity is equal to $10^{7.6}$ P, a value at which the material is easily deformable. Commercial glass compositions are formed into product shapes in a viscosity range near this point. The annealing point represents a temperature at which a glassy phase can begin to have a deleterious effect on the mechanical behavior of a structural ceramic, and the softening point represents a temperature at which the mechanical behavior can be seriously degraded. Table 12 gives values of the annealing point and softening point for various glassy phase compositions.

5. OPTICAL PROPERTIES

Traditional ceramic products, such as the pottery and whitewares of Sec. 2.1, are classic examples of opaque materials. It is important to note, however, that on an atomic scale these materials are often inherently transparent. Their opacity is the result of light scattering mechanisms associated with microscopic-scale structure. Following a discussion of refraction and reflection in microstructurally transparent materials, the mechanism of opacity in porous materials will be described. It is important to note that many applications of structural ceramics depend upon their optical properties, specifically their ability to refract, reflect, or transmit visible light (Kingery, Bowen, and Uhlmann, 1976). Although we may associate traditional ceramics with opacity, optical transmission is possible by the elimination of light-scattering porosity. The transparency of contemporary window glass results from the absence of such porosity. Both ceramic and glass materials are inherently good transmitters of visible light. The degree of transmission is a microstructural issue. Many of the advances in ceramic processing described in Sec. 3.2 have been driven by the superior optical quality derived from fine-grained, pore-free microstructures.

5.1 Refraction

The refraction of light by a transmitting ceramic is given by the refractive index n, equal to the ratio of the speed of light in vacuum to that in the transmitting medium. Table 13 gives values of n for typical ceramics and glassy phases. One may note that typical values are in the vicinity of 1.5 to 2.0.

5.2 Reflection

The technological importance of reflection involves two factors:

1. the fraction of light reflected at the ceramic surface and
2. the nature of the reflection in relation to the average surface topography.

The fraction of light reflected at a ceramic surface is given by the reflectance R and is a simple function of the index of refraction by Fresnel's formula (see ELECTROMAGNETIC RADIATION; ELECTROMAGNETIC WAVE PROPAGATION):

$$R=\left(\frac{n-1}{n+1}\right)^2 . \tag{9}$$

Eq. (9) is strictly valid for normal incidence but is a good approximation over a range of incidence angles. Clearly, Eq. (9) indicates that high-n materials are of high reflectivity or "glossy" in appearance.

Table 13. Refractive index for various ceramics.

Materials	Average refractive index
Quartz (SiO_2)	1.55
Mullite ($3Al_2O_3 \cdot 2SiO_2$)	1.64
Orthoclase ($KAlSi_3O_8$)	1.525
Albite ($NaAlSi_3O_8$)	1.529
Corundum (Al_2O_3)	1.76
Periclase (MgO)	1.74
Spinel ($MgO \cdot Al_2O_3$)	1.72

Source: Kingery, Bowen, and Uhlmann (1976).

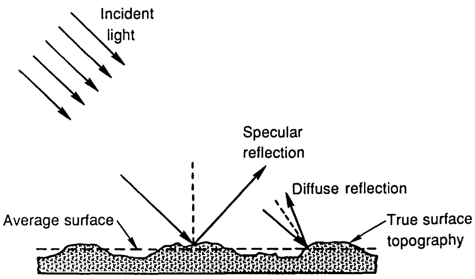
Incident light
Specular reflection
Diffuse reflection
Average surface
True surface topography

FIG. 5. Specular reflection occurs relative to the "average" surface, and diffuse reflection occurs relative to locally nonparallel surface elements.

The general appearance of a given material is strongly affected by the relative amounts of specular reflection (defined as reflection relative to the "average" surface) and diffuse reflection (defined as relative to local surface irregularity). Figure 5 illustrates schematically these two types of reflection. Microscopically smooth surfaces tend to provide a high degree of specular reflection (another contributor to "glossy" appearance) and "rough" surfaces tend to provide a high degree of diffuse reflection.

5.3 Transmission

As discussed in the two preceeding sections, the overall quality of the optical behavior of a given ceramic depends upon the combination of reflection and refraction of light, as summarized in Fig. 6. The angle of reflection equals the angle of incidence θ_i, and the angle of refraction θ_r is related to the angle of incidence by the index of refraction, i.e.,

$$n = (\sin\theta_i)/(\sin\theta_r). \tag{10}$$

In general, the refraction of light on the microscopic scale leads to the scattering of light and a loss of image transmission quality, as indicated in Fig. 7, which illustrates how scattering can occur at a single pore. Transparency can be defined as the ability to transmit a clear image in a pore-free material. Translucency means that a diffuse image is transmitted, and opacity means total loss of image transmission.

In addition to the scattering effects described above, color is an important optical property based on a true absorption mechanism, viz., the selective absorption of certain wavelength ranges within the visible spectrum due to electron transitions in transition-

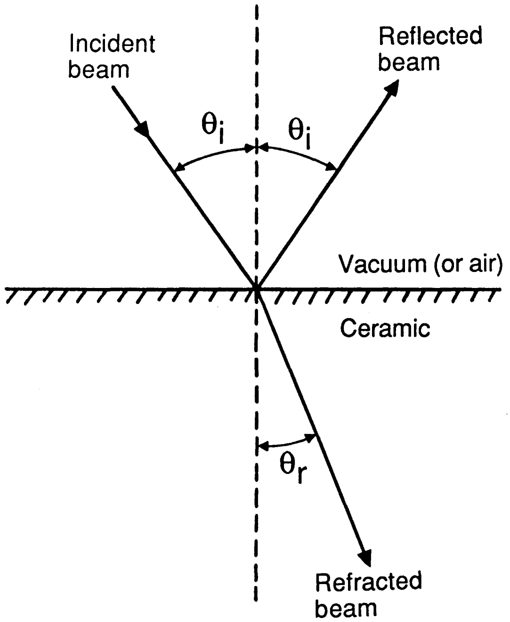
Incident beam
Reflected beam
θi
θi
Vacuum (or air)
Ceramic
θr
Refracted beam

FIG. 6. Reflection and refraction of light at the surface of a transparent ceramic.

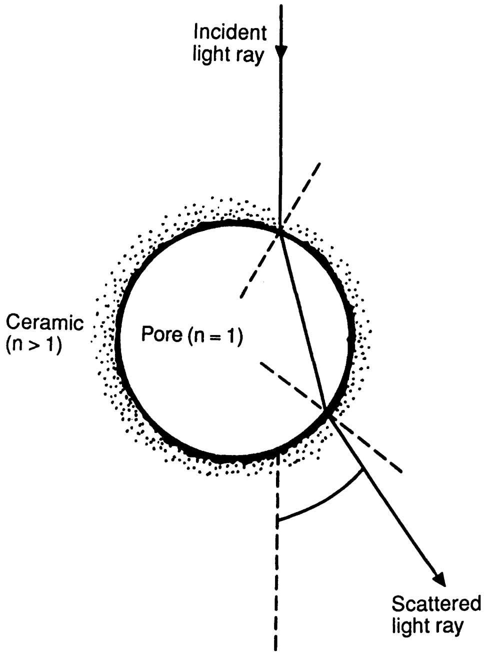
Incident light ray
Ceramic (n > 1)
Pore (n = 1)
Scattered light ray

FIG. 7. Light scattering is the result of local refraction at interfaces of second-phase particles or pores. The case for scattering by a pore is illustrated here.

metal ions. The transmitted color is the resulting combination of the wavelength ranges not absorbed. Colors provided by various metal ions are summarized in Table 14.

6. ELECTRICAL PROPERTIES

Until 1986, a discussion of the electrical properties of ceramics primarily involved characteristic insulating behavior and related dielectric properties (Kingery, Bowen, and Uhlmann, 1976). Since that time, the discussion necessarily includes the superconducting behavior of a select family of oxides (Strongin *et al.*, 1987).

6.1 Conductivity of Traditional Ceramics

In general, the electrical conductivity (see INSULATORS AND SEMICONDUCTORS, CONDUCTIVITY IN), σ, of any material is the sum of the contributions of both negative and positive charge carriers, viz.,

$$\sigma = n_n q_n \mu_n + n_p q_p \mu_p, \tag{11}$$

where n is the density of charge carriers, q the charge carried by each, and μ the carrier mobility. The subscripts n and p refer to negative and positive carriers, respectively. For typical ceramics, the conductivity can be due to electrons, electron holes, or ions. In any case, the magnitude of conductivity in these solids with typically strong ionic and covalent bond character is quite low, e.g., in the range of 10^{-10} to 10^{-13} Ω^{-1} m^{-1}. Exceptions to this rule of thumb include ZnO semiconducting "varistors" (providing overvoltage protection for sensitive equipment) and "fast-ion" conducting Naβ-Al_2O_3, a candidate for high-energy-density battery applications (Buchanan, 1986). The ultimate exception to the rule is given in the next section.

Table 14. Coloration due to transition metal ions.

Configuration	Ion	Color
d^0	Ti^{4+}	Colorless
	V^{5+}	Faint yellow to colorless
	Cr^{6+}	Faint yellow to colorless
d^1	Ti^{3+}	Violet-purple
	V^{4+}	Blue
	Mn^{6+}	Colorless
d^2	V^{3+}	Yellow-green
d^3	Cr^{3+}	Green
d^4	Cr^{2+}	Faint blue
	Mn^{3+}	Purple
d^5	Mn^{2+}	Light yellow
	Fe^{3+}	Faint yellow
d^6	Fe^{2+}	Blue-green
	Co^{3+}	Faint yellow
d^7	Co^{2+}	Blue-pink
d^8	Ni^{2+}	Brown-purple
d^9	Cu^{2+}	Blue-green
d^{10}	Cu^{+}	Colorless

Source: Data from Sigel, G. H., Jr. (1977).

6.2. Superconducting Ceramics

For a select group of materials known as superconductors, the electrical resistivity drops to zero at a critical temperature T_c. The discovery in 1986 that a $(La,Ba)_2CuO_4$ ceramic exhibited a critical temperature T_c of 35 K (more than 10 K higher than the previous record for a Nb_3Ge alloy) stimulated an unprecedented interest in ceramic materials (Strongin *et al.*, 1987) (see HIGH-TEMPERATURE SUPERCONDUCTIVITY). In 1987, $YBa_2Cu_3O_7$ (the 1-2-3 superconductor, named after the ratio of cations) was found to have a T_c of 95 K, well above the temperature of liquid nitrogen (77 K). By 1988, a Tl-Ca-Ba-Cu-O ceramic was found that exhibited a T_c of 125 K. Considerable evidence has been found for superconducting behavior in limited regions of samples of other ceramic oxides at even higher temperatures. The possibility of a room-temperature superconductor is, of course, a source of considerable interest.

The mechanism for superconductivity in metals is well established, viz., the "Cooper pairs" of electrons of opposite spin. Although superconductivity in high-T_c superconductors also involves paired electrons, the nature of the conduction mechanism is not as well understood. Specifically, electron pairing does not seem to result from the same type of synchronization with lattice vibrations. What has become apparent is that copper–oxygen planes in the crystal structure are the pathways for the supercurrent. In the 1-2-3 superconductor, that current is carried by electron holes. Other ceramic oxides have been developed in which the current is carried by electrons, not holes.

In either case, the promise of superconductors produced by the dramatic increase in T_c has encountered an obstacle in the form of another important material property, viz., the critical current density, defined as the current flow at which the material stops being superconducting. Metallic superconductors used in applications such as magnets in large-scale particle accelerators have critical current densities on the order of 10^{10} A/m^2. Such magnitudes have been produced in thin films of ceramic superconductors, but bulk samples typically give values about one-hundreth that size. Unfortunately, the limitation in current density becomes more severe for increasing values of T_c. The diminished critical current density is caused by the penetration of the surrounding magnetic field into the material, creating an effective resistance due to the interaction between the current and mobile magnetic flux lines. The problem does not exist in metallic superconductors because the magnetic flux lines are not mobile at such low temperatures. Above the commercially important temperature of 77 K, the effect (known as "flux lattice melting") becomes significant and is increasingly so with further increases in temperature. As a practical consequence, T_c values above that of the 1-2-3 material may be of little benefit. Even for this material, immobilization of magnetic flux lines may require a thin-film configuration or some special form of microstructural control. (In general, a primary challenge in the commercial development of high-T_c superconductors is the fabrication of usable product shapes from the inherently brittle ceramics.)

As the applications of superconductors generally involve their magnetic behavior, a further discussion of these materials will be given in Sec. 7.2.

6.3 Dielectrics

A parallel-plate capacitor is composed of an insulator (dielectric) between two conducting electrodes (see DIELECTRIC PROPERTIES OF INSULATORS). The charge density D built up in the insulator is proportional to the electric field strength E, according to the relation

$$D=\epsilon_0 \kappa E, \tag{12}$$

where ϵ_0 is the electric permittivity of a vacuum and κ the dielectric constant, representing the factor by which the capacitance of the system is increased by inserting the dielectric in place of a vacuum. For a given dielectric, there is a limiting voltage gradient, termed the dielectric strength, at which an appreciable current flow (or "breakdown") occurs and the dielectric fails. Representative values of dielectric constant and dielectric strength for various ceramics are given in Table 15.

Table 15. Dielectric constant and dielectric strength for various ceramics.

Material	Dielectric constant[a] κ	Dielectric strength[b] (kV/mm)
Al_2O_3 (99.9%)	10.1	9.1
Al_2O_3 (99.5%)	9.8	9.5
BeO (99.5%)	6.7	10.2
Cordierite	4.1–5.3	2.4–7.9

Source: Data from Smothers (1985).
[a] At 10^3 Hz.
[b] Average rms values at 60 Hz.

6.4 Ferroelectrics and Piezoelectrics

The nature of ferroelectric and piezoelectric behavior in ceramics can be conveniently represented by barium titanate, $BaTiO_3$. Above 120 °C, $BaTiO_3$ has the cubic crystal structure of the mineral perovskite, $CaTiO_3$. Upon cooling below 120 °C, $BaTiO_3$ undergoes a phase transformation to a tetragonal modification, as shown in Fig. 8.

Below the critical temperature (T_c) of 120 °C, $BaTiO_3$ is ferroelectric. This behavior is directly related to the asymmetric nature of the tetragonal structure shown in Fig. 8(b). The offset between the overall center of positive charge for the cation distribution and the overall center of negative charge for the anion distribution amounts to a permanent electrical dipole in the tetragonal $BaTiO_3$ unit cell. This dipole structure allows for a large polarization of the material in response to an applied electrical field. The specific nature of the polarization (P)–electric field (E) relationship is shown in Fig. 9.

The origin for the P-E axis system (in which zero polarization occurs under zero applied field) corresponds to a balance of equal volumes of oppositely oriented domains in a random microstructure. The domains can be defined as microscopic-scale crystallite regions with substantial polarization corresponding to the crystal structure of Fig. 8(b). Under an applied field, unit-cell dipole orientations roughly parallel to the applied field direction

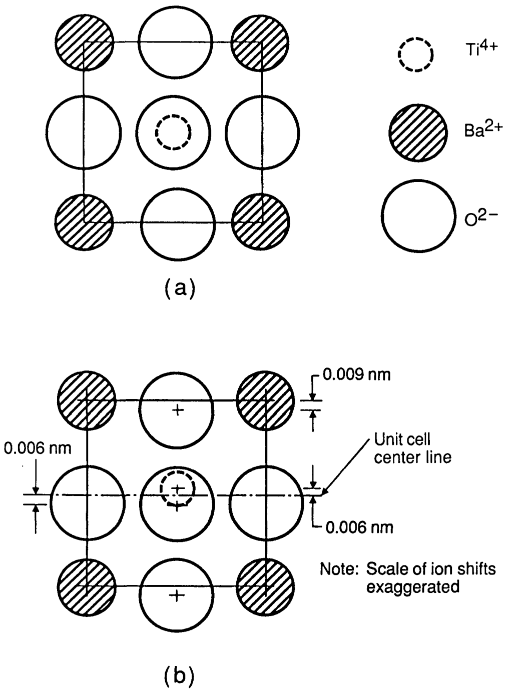

FIG. 8. (a) Front view of the cubic $BaTiO_3$ structure. (b) Below 120 °C, a tetragonal modification of the structure occurs. The net result is an upward shift of cations and a downward shift of anions.

are favored, and such orientations "grow" at the expense of other, less favorably oriented ones. The specific mechanism of domain-wall motion is simply the small shift of ion positions within unit cells, resulting in the net change of orientation of the tetragonal *c* axis. Such domain-wall motion results in the sharp rise in polarization for low-field application, termed spontaneous polarization (curved dashed line in Fig. 9). The straight dashed line from the origin represents the modest amount of polarization (induced dipole) that would occur in a paraelectric material, e.g., the distortion of the symmetrical cubic structure under the applied field. The solid line in Fig. 9 summarizes the hysteresis loop that results when the electrical field is repeatedly cycled. Key parameters that quantify the hysteresis loop are the saturation polarization P_s, the polarization due to maximum domain growth and determined by extrapolating to zero field; the remanent polarization P_r, the polarization remaining upon field removal; and the coercive field E_c, the field necessary to reduce the polarization to zero. Remanent polarization is due to the fact that simply removing the field does not return the domain structure to equal volumes of opposing polarization. The coercive field is needed to achieve this result.

The characteristic hysteresis loop gives ferroelectricity its name by analogy to the similar shape of the induction (*B*)–magnetic field (*H*) plots for ferromagnetic materials (Fig. 10). As the prefix "ferro-" implies, ferromagnetic materials generally contain iron, although ferroelectrics rarely contain iron as a significant constituent.

The most common applications of ferroelectrics stem from a closely related phenomenon, piezoelectricity. The prefix "piezo" comes from the Greek word for pressure, indicating that piezoelectrics give an electrical response to mechanical pressure application. Conversely, electrical signals make them pressure generators. Both functions are directly related to the asymmetric crystal structure of Fig. 8(b). The piezoelectric effect (in which an applied stress produces a measurable voltage across the material) is the result of the crystallographic dipole dimensions being affected by the stress-induced distortion of the ionic positions within a given unit cell. Similarly, the reverse piezoelectric effect involves a thickness change in the material due to a change in the magnitude of the polarization, in turn

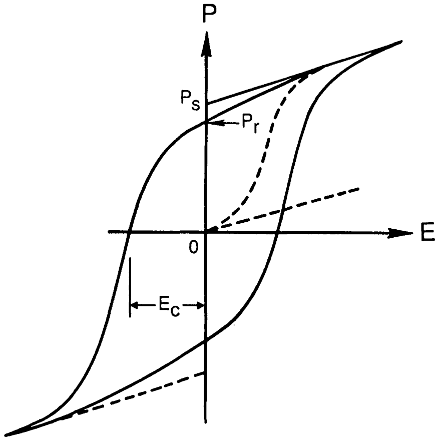

FIG. 9. A ferroelectric hysteresis loop is the result of an alternating electric field. A dashed line indicates the initial spontaneous polarization. Saturization polarization (P_s) is the result of maximum domain growth (extrapolated back to zero field). Upon actual field removal, some remanent polarization (P_r) remains. A coercive field (E_c) is required to reach zero polarization (equal volumes of opposing domains).

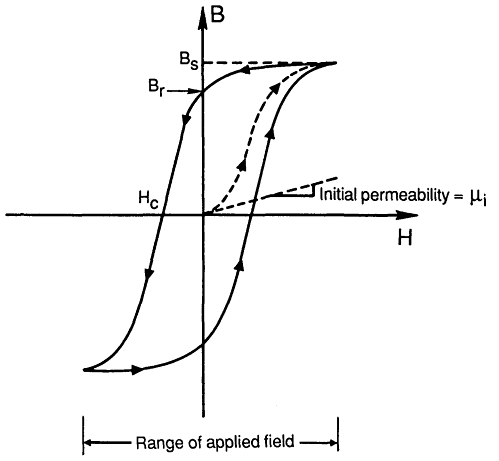

FIG. 10. The *B-H* plot hysteresis loop for a ferromagnetic material indicates a large rise in *B* during initial magnetization (shown by the dashed line). The induction reaches a large, "saturation" value (B_s) upon application of sufficient field strength. Much of that induction is retained upon removal of the field (B_r = remanent induction). A coercive field (H_c) is required to reduce the induction to zero.

caused by an applied voltage. By constraining the thickness change, the applied voltage produces a mechanical stress. A common application for a piezoelectric transducer is as an ultrasonic transmitter/receiver (see PIEZOELECTRIC DEVICES). In this case, voltage oscillations in the megahertz range produce (or sense) ultrasonic waves of that frequency.

The functioning of a piezoelectric transducer is dependent on a common crystallographic orientation throughout the material. A straightforward means of ensuring this is with a single crystal. Quartz (SiO_2) is an example and was widely used as a piezoelectric transducer after World War II. $BaTiO_3$ has a higher piezoelectric coupling coefficient k (= fraction of mechanical energy converted to electrical energy) than does SiO_2. The k value for $BaTiO_3$ is approximately 0.5 compared to 0.1 for quartz. As $BaTiO_3$ cannot be conveniently manufactured in single-crystal form, the material is processed in a powder form, and individual particles are aligned in a single crystallographic orientation under a strong electrical field. The powder is then sintered. The resulting polycrystalline material is said to be electrically poled. During the 1950s, this technology allowed $BaTiO_3$ to become the predominant piezoelectric transducer material. In the 1960s, the solid solution $Pb(Ti,Zr)O_3$, or PZT, became more widely used because of a substantially higher T_c (commonly in excess of 200 °C). In the 1970s, the addition of La_2O_3 in solid solution with PZT produced an optically transparent material, PLZT, which has found application as a high-frequency electro-optic shutter (Buchanan, 1986).

7. MAGNETIC PROPERTIES

As with electrical properties, the discussion of the magnetic behavior of ceramics has changed fundamentally since 1986. Before that time, the discussion concentrated primarily on the nature of ferrimagnetism (Buchanan, 1986). Since 1986, the magnetic behavior of the ceramic superconductors must be included (see MAGNETIC MATERIALS).

7.1 Ferrimagnetism

In general, we can define an induction **B** as a vector quantity whose magnitude is the flux density surrounding a magnetic field source. The induction is related to the magnetic field strength **H** by

$$\mathbf{B}=\mu\mathbf{H}, \tag{13}$$

where μ is the permeability of the solid within the magnetic field.

As noted in Sec. 6.4, ferromagnetic materials, such as pure iron, exhibit a hysteresis loop in the plot of induction B versus magnetic field strength H. Figure 10 is the magnetic analog for the ferroelectric hysteresis loop of Fig. 9, and, as pointed out in Sec. 6.4, serves as the source of the term "ferroelectric." The mechanism for ferromagnetism is fully analogous to that for ferroelectricity, viz., the saturation induction B_s is the result of complete domain growth, the remanent induction B_r is the value of induction remaining upon removal of the field, and the coercive field H_c is that needed to reduce the induction to zero. In this magnetic system, the domain is a region of aligned electron spins. The magnetic moment due to electron spin is given in units of the Bohr magneton μ_B, which has a value of 9.27×10^{-24} A m^2. A nonzero net magnetic moment is the result of unpaired electron spins that are present in materials such as the transition metals (e.g., Fe).

Table 16. Magnetic moment of various transition-metal ions.

Ion	Moment (μ_B)[a]
Mn^{2+}	5
Fe^{2+}	4
Fe^{3+}	5
Co^{2+}	3
Ni^{2+}	2
Cu^{2+}	1

[a] $\mu_B = 1$ Bohr magneton$=9.27\times10^{-24}$ A·m^2.

Traditional magnetic ceramics have a magnetic hysteresis loop similar to that in Fig. 10, but their crystal structure leads to some antiparallel spin pairing, causing the net magnetic moment to be less than that possible in metals. This similar phenomenon is distinguished from ferromagnetism by the slightly different spelling ferrimagnetism. The historical model of a magnetic material is magnetite, Fe_3O_4, in which one-third of the iron ions are in the Fe^{2+} charged state and two-thirds in the Fe^{3+} state. The trivalent ions are equally distributed between tetrahedral (fourfold coordination) and octahedral (six-fold coordination) sites, and their magnetic moments are antiparallel, leading to the cancellation of their contribution to the net magnetic moment of the crystal. The net moment is, then, provided solely by the divalent ions. The number of Bohr magnetons contributed by various transition-metal ions is summarized in Table 16. There are eight divalent ions per unit cell in magnetite, giving a predicted magnetic moment of $32\mu_B$, in good agreement with the experimental value of saturation induction.

The dominant examples of commercial ceramic magnets are the ferrites, which have the same crystal structure as magnetite (the so-called "inverse spinel" structure). A list of representative commercial ferrites is given in Table 17. Another group of magnetic ceramics is the garnets, based on the crystal structure of the natural gem garnet, $Al_2Mg_3Si_3O_{12}$. Yttrium iron garnet (YIG) is an example. Both ferrites and garnets are soft magnets, i.e., have a relatively small area within the hysteresis loop. The magnetoplumbites are, conversely, hard magnets with a characteristically high coercive field. These hexagonal materials with the composition $M_{0.6}Fe_2O_3$ (M =divalent cation) find applications as inexpensive permanent magnets. Another hard magnetic material is γ-Fe_2O_3, widely used as a fine-grained film in recording tape and both flexible and rigid computer disks.

7.2 Superconducting Magnets

As discussed in Sec. 6.2, ceramic superconductors have been available since 1986 (Strongin *et al.*, 1987). As pointed out in the discussion of their electrical properties, the encouraging rise in critical temperature for these materials has been contrasted with the discouraging limitation of critical current density. There is a third, major materials property for superconductors, viz., the critical magnetic field above which the material stops being superconducting. Unlike the critical current density, the critical magnetic field increases along with critical temperature for the high-T_c materials. Unfortunately, the limited current density (due to magnetic field penetration) remains as a primary obstacle, especially for large-scale applications such as power transmission and transportation. As pointed out earlier, the most promising applications

Table 17. Some commerical ferrite compositions.

Name	Composition	Comments
Magnesium ferrite	$MgFe_2O_4$	
Magnesium–zinc ferrite	$Mg_xZn_{1-x}Fe_2O_4$	$0<x<1$
Manganese ferrite	$MnFe_2O_4$	
Manganese-iron ferrite	$Mn_xFe_{3-x}O_4$	$0<x<3$
Manganese-zinc ferrite	$Mn_xZn_{1-x}Fe_2O_4$	$0<x<1$
Nickel ferrite	$NiFe_2O_4$	
Lithium ferrite	$Li_{0.5}Fe_{2.5}O_4$	Li^+ occurs in combination with Fe^{3+}, producing a configuration $(Li_{0.5}Fe_{0.5})^{2+}Fe_2{}^{3+}O_4$

of these new materials appear to be in thin-film devices, such as Josephson junctions.

GLOSSARY

Ferrimagnetism: As with ferromagnetism, the phenomenon of a sharp rise in induction with applied magnetic field. Distinguished from ferromagnetism by the presence of antiparallel spin pairing.

Ferroelectric: Material that exhibits a sharp rise in polarization under an applied electrical field.

Fracture toughness: Critical value of the stress–intensity factor at a crack tip necessary to produce catastrophic failure.

Piezoelectric: Material that produces a measurable voltage as a result of an applied stress (the piezoelectric effect) and undergoes a thickness change due to a change in applied voltage (the reverse piezoelectric effect).

Refractory: High-temperature-resistant material.

Sintering: Bonding of powder particles by solid-state diffusion.

Sol–Gel Processing: Technique for forming ceramics of high density at a relatively low temperature by means of an organometallic solution.

Transformation Toughening: Mechanism for enhanced fracture toughness involving a stress-induced phase transformation.

Vitreous: Noncrystalline or glassy phase.

List of Works Cited

Ashby, M. F., Jones, D. R. H. (1980), *Engineering Materials: An Introduction to Their Properties and Applications*, Elmsford, NY: Pergamon.

Buchanan, R. C. (Ed.) (1986), *Ceramic Materials for Electronics: Processing, Properties, and Applications*, New York: Dekker.

Evans, A. G. (Ed.) (1984), *Fracture in Ceramic Materials*, Park Ridge, NJ: Noyes.

Hampshire, S. (Ed.) (1986), *Non-oxide Technical and Engineering Ceramics*, New York: Elsevier.

Kingery, W. D. (1960), *Introduction to Ceramics*, New York: Wiley.

Kingery, W. D., Lense, E. (Eds.) (1985), *Ancient Technology to Modern Science*, Columbus, OH: American Ceramic Society.

Kingery, W. D., Bowen, H. K., Uhlmann, D. R. (1976), *Introduction to Ceramics*, 2nd ed. New York: Wiley.

Klein, A. J. (1986), *Adv. Mater. Proc.* **2** (9), 26–27, 30–33.

Lemkey, F. D., Fishman, S. G., Evans, A. G., Strife, J. R. (1988), *High Temperature/High Performance Composites*, Pittsburgh, PA: Materials Research Society.

McMillan, P. W. (1979), *Glass-Ceramics*, 2nd ed. New York: Academic.

Munir, Z. A. (1988), *Am. Ceram. Soc. Bull.* **67**, 342–349.

Reed, J. S. (1988), *Introduction to the Principles of Ceramic Processing*, New York: Wiley.

Shackelford, J. F. (1987), in: S. Saito (Ed.), *Fine Ceramics*, Tokyo: Ohmsha.

Sigel, G. H., Jr. (1977), in: M. Tomozawa and R. H. Doremus (Eds.), *Glass I*, New York: Academic.

Smothers, W. (Ed.) (1985), *Ceramics '86*, Columbus, OH: American Ceramic Society.

Smothers, W. (Ed.) (1986), *Ceramics '87*, Columbus, OH: American Ceramic Society.

Strongin, M., Terwilliger, R., Adams, P. D. (Eds.) (1987), *High-Temperature Superconductivity, Reprints from Phys. Rev. Lett. and Phys. Rev. B, Jan.–June 1987*, New York: American Physical Society.

Tinklepaugh, J. R., Crandall, W. B. (Eds.) (1960), *Cermets*, New York: Reinhold.

Upadhyaya, G. S. (Ed.) (1984), *Sintered Metal-Ceramic Composites*, New York: Elsevier.

Further Reading

Ashby, M. F., Jones, D. R. H. (1980), *Engineering Materials: An Introduction to Their Properties and Applications*, Elmsford, NY: Pergamon.

Buchanan, R. C. (Ed.) (1986), *Ceramic Materials for Electronics: Processing, Properties, and Applications*, New York: Dekker.

Evans, A. G. (Ed.) (1984), *Fracture in Ceramic Materials*, Park Ridge, NJ: Noyes.

Kingery, W. D., Bowen, H. K., Uhlmann, D. R. (1976), *Introduction to Ceramics*, 2nd ed. New York: Wiley.

Reed, J. S. (1988), *Introduction to the Principles of Ceramic Processing*, New York: Wiley.

Shackelford, J. F. (1988), *Introduction to Materials Science for Engineers*, 2nd ed. New York: Macmillan.

CHAOTIC PHENOMENA

HEINZ GEORG SCHUSTER, *Institut für Theoretische Physik, Universität Kiel, Kiel, Germany*

Introduction 189
1. **History** 190
2. **Nonlinear Dynamical Systems** 191
2.1 Examples 191
2.1.1 Periodically Driven Pendulum. 191
2.1.2 Bénard Experiment 191
2.2 Lorenz Equations 192
2.3 Poincaré Section 192
2.4 Mechanism for Chaos 193
3. **Strange Attractors and their Characterization** 194
3.1 Attractors of Dynamical Systems 194
3.1.1 Lorenz Attractor 194
3.2 Characterization of Chaotic Motion 195
3.2.1 Liapunov Exponents 196
3.2.2 Hausdorff Dimension 196
3.2.3 Spectrum of Attractor Dimensions 196
4. **Routes to Chaos** 197
4.1 Period-Doubling Route 197
4.1.1 Universality and Feigenbaum Constants 198
4.1.2 Pitchfork Bifurcation and Functional Fixed-Point Equation 199
4.1.3 Experimental Observation of Period Doubling 199
4.2 Intermittency 199
4.2.1 Mechanisms for Intermittency 200
4.2.2 Experimental Observation of Intermittency 201
4.3 Ruelle–Takens–Newhouse Route 201
4.4 Comparison of Three Routes to Chaos 202
5. **Experiments and Applications** 202
5.1 Reconstruction of an Attractor from a Measured Time Series . 202
5.2 Correlation Integral 204
5.3 Applications 205
5.3.1 Lorenz Model 205
5.3.2 Taylor Instability 205
5.3.3 Medical Data 206
6. **Conservative Systems** 207
6.1 Integrable Systems 208
6.2 Kolmogorov–Arnol'd–Moser Theorem 208
6.3 Coexistence of Regular and Irregular Motion 209
6.4 Unstable Planetary Motion 210
7. **Chaos in Quantum Systems?** . 210
7.1 Classical versus Quantum Systems 210
7.2 A Quantum Particle in a Stadium 211
7.3 Explicitly Time-Dependent Quantum Systems 212
8. **Outlook** 213
Glossary 213
Works Cited 213
Further Reading 214

INTRODUCTION

Daily experience shows that for many systems small changes in the initial conditions lead to small changes in the outcome. If we turn the steering wheel by a small angle the course of our car will only gradually deviate from its previous route. But there are other cases: For a coin standing on end, a slight touch will suffice to make it fall and to determine the side on which it will fall. In recent years it has become clear that sensitive dependence on initial conditions is not exceptional but typical of many systems and leads to chaotic, i.e., irregular, behavior in time.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

In the following we shall consider systems whose time dependence is deterministic, i.e., there exists a prescription in terms of either difference or differential equations for calculating their future behavior from given initial conditions. One could assume naively that deterministic motion is rather regular and far from being chaotic because successive states evolve continuously from each other. But it was already discovered at the turn of the century by the mathematician Poincaré (1892) that certain mechanical systems could display chaotic motion. Unfortunately, this was considered by many physicists as a mere curiosity, and it took another 70 years until in 1963 the meteorologist E. N. Lorenz found that even a simple set of three differential equations which describe in a crude sense the flow of air in the earth's atmosphere, i.e., the problem of weather forecasting, can lead to completely irregular trajectories. Mainly through the availability of high speed computers it has now become clear that chaotic behavior is abundant in nature (see Table 1) and has far-reaching consequences in many branches of science.

The observed chaotic behavior in time is due neither to external sources of noise (there are none in Lorenz's equations) nor to an infinite number of degrees of freedom (Lorenz's system has only three degrees of freedom) nor to the uncertainty associated with quantum mechanics (Lorenz's system is purely classical). The actual source of irregularity is the nonlinearities in the equations of motion which amplify small but practically unavoidable errors in the initial conditions so strongly that the long-time behavior of the system becomes erratic and unpredictable.

Table 1. Some nonlinear systems which display deterministic chaos.

Forced pendulum
Fluids near the onset of turbulence
Lasers
Nonlinear optical devices
Josephson junctions
Chemical reactions
Classical many-body systems (three body problem)
Particle accelerators
Plasmas with interacting nonlinear waves
Biological models for population dynamics
Stimulated heart cells

Lorenz called this sensitive dependence on initial conditions the butterfly effect because the outcome of his equations, which describe the flow of air in the earth's atmosphere, could be changed by a butterfly flapping its wings.

Thus the question arises whether the appearance of chaotic motion implies the end of predictability for most nonlinear systems or whether one can still learn something from a chaotic signal (Ford, 1983). As we will seen in the following, fortunately the latter is the case; however, one has to pose the right questions.

The approach taken here is to review the general characteristics of chaotic systems and then to classify and quantify their behavior in more detail. We begin with a historical introduction followed by more a general introduction to nonlinear dynamical systems. Next we introduce the concept of a strange attractor and review different routes by which a dynamical system could develop chaotic behavior if one external control parameter is changed. The remaining sections are devoted to the occurrence of chaos in classical systems and the behavior of quantum systems whose classical counterpart displays chaos. We conclude with several applications of chaotic motion analysis and an outlook on further developments.

1. HISTORY

Chaos was first discovered in mechanical systems by the great French mathematician H. Poincaré (1892), who was stunned by the complexity of trajectories traced by three bodies interacting by gravitational forces (the classical three-body problem). About sixty years later Kolmogorov (1954), Arnol'd (1963), and Moser (1973) proved, in what is now called the KAM theorem, that motion in classical mechanics is neither completely regular nor completely irregular, but that the type of trajectory depends sensitively on the initial conditions. Another breakthrough came when, in 1963, E. N. Lorenz found the first example of chaos in dissipative systems. In contrast to systems of classical mechanics where energy is conserved, in dissipative systems energy is dissipated by friction, and motion is kept up only by external driving forces. A typical dissipative system is a damped pendulum which is driven by an external torque. Lorenz's discovery not only enlarged the number of systems which could display chaos but also opened the door for the discovery of qualitatively new phenomena, like universal routes to chaos and strange attractors.

In the 1970s Ruelle and Takens (1971) and Newhouse *et al.* (1978) suggested a transition to chaotic motion which was radically different from an earlier proposal by Landau (1944). Landau considered "turbulence in time" as the limit of an infinite sequence of instabilities (Hopf bifurcations) each of which creates a new basic frequency, such that the motion becomes more and more complicated. However Ruelle, Takens, and Newhouse showed that after only two instabilities the trajectory could move onto a so-called strange attractor displaying chaotic motion. A strange attractor is a generalization of simple attractors such as fixed points or limit cycles and allows for a classification of chaotic systems.

A more recent route to chaos has been found by Grossmann and Thomae (1977), Feigenbaum (1978), and Coullet and Tresser (1978). They considered a simple nonlinear difference equation (logistic equation) that has been used to describe the time evolution of populations in biology, and found that the population oscillated between distinct values (fixed points) whose number doubles again and again if an external parameter is increased. This continues until the number of (unstable) fixed points becomes infinite at a finite parameter value where the time variation of the population becomes irregular. Feigenbaum (1978) has shown, and this was a major achievement, that these results are not restricted to this special model, but are in fact universal and hold for a large variety of physical, chemical, and biological systems.

Almost in parallel to this discovery, Manneville and Pomeau (1979) found a third route to chaos via intermittency. Intermittency means that a signal which behaves regularly (or laminarly) in time becomes interrupted by statistically distributed periods of irregular motion (intermittent bursts). The average number of these bursts increases with the variation of an external control parameter until the motion becomes completely chaotic.

These discoveries triggered an explosion of theoretical and experimental research, part of which will be reported in the following sections.

2. NONLINEAR DYNAMICAL SYSTEMS

Nonlinear dynamical systems are generally described by first-order differential equations

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}, \lambda_1, \lambda_2, \ldots); \quad \mathbf{x} = (x_1, \ldots, x_d) \tag{1}$$

(the overdot denotes time derivative) which are autonomous, i.e., **F** does not contain the time explicitly, and nonlinear, i.e., **F** is a nonlinear function of the variables $x_1, \ldots, x_d$. These equations can lead to chaotic behavior in time if one or several external control parameters λ_1, $\lambda_2, \ldots$ are varied.

2.1 Examples

In Table 2 we describe two examples of systems which display deterministic chaos and comment on the methods by which chaos could be detected experimentally.

2.1.1 Periodically Driven Pendulum. The equation of motion for the angle θ of the pendulum shown in the upper part of Table 2 contains a friction term $\gamma\dot{\theta}$, a periodic driving torque $A\cos(\omega t)$, and the characteristic nonlinearity $\sin\theta$. The dot denotes the time derivative and ω is the frequency of the driving force. By introducing new variables x, y, z, the equation of motion of the driven pendulum can be brought into the canonical form of Eq. (1). The external control parameters are A and ω, and for special values of both, the time dependence of the angle simply "looks" chaotic as shown in the right column of Table 2.

2.1.2 Bénard Experiment. In this experiment a fluid layer in a "matchbox" is heated from below in a gravitational field. The heated fluid at the bottom rises to the top and cold liquid descends but both motions are opposed by viscous forces. For small temperature differences ΔT viscosity wins, the liquid remains at rest, and heat is transported by heat conduction. This state becomes unstable for larger values of ΔT, and a state of stationary convection rolls develops which is in turn unstable and leads to chaotic motion in time if ΔT is further increased. Here ΔT is the control parameter and chaotic motion is detected in the power spectrum

$$p(\omega) = |x(\omega)|^2, \tag{2}$$

Table 2. Detection of chaos in simple systems.

System	Equation of Motion	Indication
Pendulum 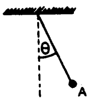	$\ddot{\theta}+\gamma\dot{\theta}+g\sin\theta=A\cos\omega t$ $x=\theta,\ y=\dot{\theta},\ z=\omega t$ $\dot{x}=y$ $\dot{y}=-\gamma y-g\sin x+A\cos z$ $\dot{z}=\omega$	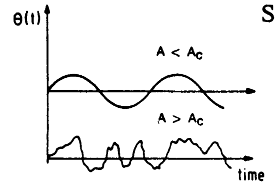 Signal
Bénard experiment 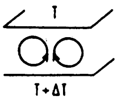	$\dot{x}=-\sigma x+\sigma y$ $\dot{y}=rx-y-xz$ $\dot{z}=xy-bz$	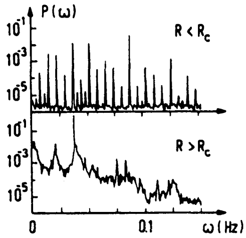 Power Spectrum

where $x(\omega)$ is the Fourier transform of the fluid velocity $x(t)$:

$$x(\omega)=\lim_{T\to\infty}\int_0^T dt\, e^{i\omega t}x(t). \tag{3}$$

For regular motion $p(\omega)$ consists of discrete lines indicating distinct basic frequencies with which the system oscillates, whereas chaotic motion which is completely aperiodic is indicated by a broad continuum in $p(\omega)$ which extends to very low frequencies.

2.2 Lorenz Equations

The Bénard experiment can be crudely described by the following famous equations of motion which were first derived by Lorenz (1961) to describe meteorological phenomena:

$$\begin{aligned}\dot{x}&=-\sigma x+\sigma y,\\ \dot{y}&=-xz+rx-y,\\ \dot{z}&=xy-bz.\end{aligned} \tag{4}$$

A derivation of these equations from the Navier–Stokes equation for fluid motion is beyond the scope of this contribution and we will merely comment on the physical meaning of the different variables. The variable x in Eq. (4) is proportional to the circulatory fluid flow velocity, y characterizes the temperature difference between ascending and descending fluid elements, and z describes the deviation of the vertical temperature profile from its equilibrium value. The variables σ and b are dimensionless constants which characterize the system, and r is proportional to the temperature difference ΔT which is the control parameter. Again the Lorenz equations are of the form of Eq. (1).

2.3 Poincaré Section

The natural space in which the time behavior of a dynamical system could be visualized is phase space, which is spanned by the variables

$$\mathbf{x}(t)=[x_1(t),\ldots,x_d(t)] \tag{5}$$

that describe the state of the system. Figure 1 shows a two-dimensional projection of the phase-space variable

$$x_1(t)=\theta(t),\quad x_2(t)=\dot{\theta}(t),\quad x_3(t)=\omega t \tag{6}$$

for a damped driven pendulum. We distinguish the following different attractors, i.e., bounded regions in phase space to which the trajectory becomes attracted in the course of time:

- Fixed point: The trajectory comes finally to rest in a point in phase space, $\theta=0$, $\dot{\theta}=0$, if the pendulum swings back to its resting position with decreasing angle and angular velocity $\dot{\theta}$.
- Limit cycle: The trajectory becomes at-

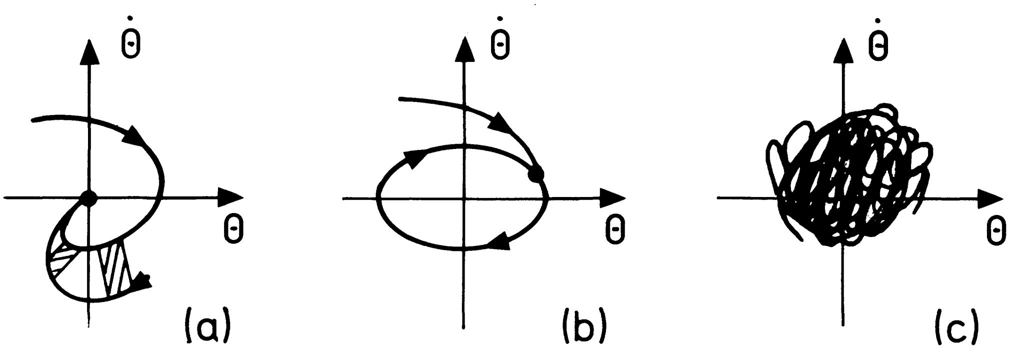

FIG. 1. Schematical drawing of different attractors which could occur in the phase space of a driven pendulum: (a) fixed point; (b) limit cycle; (c) strange attractor.

tracted to a simple closed loop in phase space if the pendulum is externally driven in such a way that the external torque compensates friction, and stabilizes regular pendulum motion, without leading to turnovers.

- Strange attractor: The trajectory performs irregular motion in a bounded region of phase, i.e., on a so-called strange attractor. This is because the driving torque is strong enough to generate turnovers of the pendulum. At the summit $\theta=\pi/2$ a tiny touch is enough to determine whether the angle moves to the right or left, i.e., one sees at least qualitatively that we have there sensitive dependence on initial conditions which leads to error amplification and finally to chaotic motion because at each turnover the pendulum passes the point $\theta=\pi/2$ again and again, i.e., we have so-called "backfolding" which amplifies errors enormously and makes the system unpredictable.

Following a method introduced already by Poincaré we have plotted in Fig. 2 the points in which the trajectory cuts the x axis. We see that a limit cycle corresponds to a single point, a two-cycle to two points, and chaotic motion to a sequence of successive points $x_0, x_1, x_2, \ldots$. This method to study the trajectories of a d-dimensional flow described by Eq. (1) by cutting it with a $(d-1)$-dimensional hyperplane is called Poincaré section, and it is often applied to visualize three-dimensional flows in the two-dimensional plane.

2.4 Mechanism for Chaos

If in Fig. 2 the sequence of points is irregular, we know that the system displays chaos, because these points belong to a trajectory. It is therefore possible and convenient to visualize the basic mechanism for chaotic motion via a one-dimensional map

$$x_{n+1}=f(x_n), \quad n=0,1,2,\ldots \tag{7}$$

which generates a sequence of chaotic points. In order to do so the map should have the

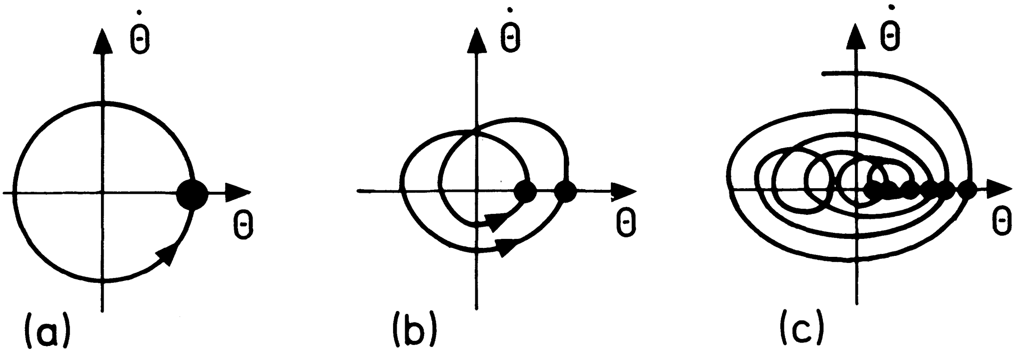

FIG. 2. Poincaré section of trajectories with an axis. (a) Limit cycle; (b) two-cycle; (c) chaos. Successive crossing points are denoted by x_n, $x_{n+1}, \ldots$. Note that there is only one crossing point for a circle.

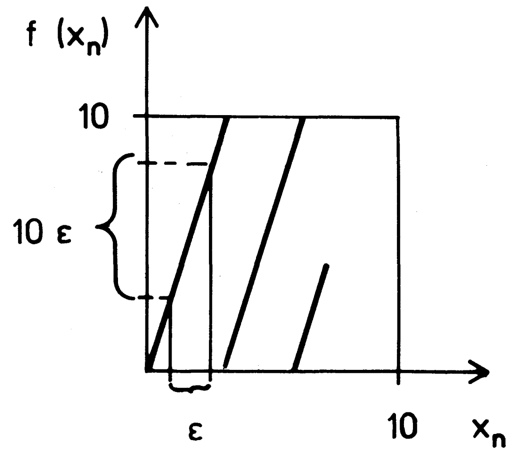

FIG. 3. The map $x_{n+1}=(10x_n)\,\mathrm{mod}10$.

property of error amplification and backfolding that are needed to generate chaos.

Figure 3 shows that the map

$$x_{n+1}=(10x_n)\mathrm{mod}10, \quad n=0,1,2,..., \tag{8}$$

amplifies an error ϵ by a factor of 10 and folds the iterates back to the interval [0, 10]. If we apply this map to an irrational number, say π, we obtain the following sequence of iterates:

$$\begin{aligned} x_0&=\pi &&=3.14159...,\\ x_1&=(31.4159...)\mathrm{mod}10 &&=1.4159...,\\ x_2&=(14.159...)\mathrm{mod}10 &&=4.159..., \end{aligned} \tag{9}$$

i.e., we see that the map shifts the sequence of decimals to the left and cuts off all but one digit before the decimal point. In this fashion it "pumps" the irregular sequence of decimals of π in front of the decimal point and makes them visible. This is all that a chaotic system does: It makes the chaos which exists already in the data of the initial conditions (x_0 for our example) macroscopically visible. If we assume for example that x_0 is a measured variable which is known up to three digits after the decimal point, we could write

$$x_0=a_0\cdot a_1a_2a_3???, \tag{10}$$

where the question marks indicate the unknown decimals. After four iterations with our map from Fig. 3 we obtain

$$x_4=?.???, \tag{11}$$

i.e., the fourth iterate is completely unknown; we cannot make any prediction about it; the system displays chaos.

3. STRANGE ATTRACTORS AND THEIR CHARACTERIZATION

3.1 Attractors of Dynamical Systems

Attractors of dissipative dynamical systems are bounded regions in phase space to which the trajectory becomes attracted in the course of time. The simplest attractor is a fixed point on which the trajectory comes to rest. More complicated attractors are limit cycles, or tori, on which the trajectory spirals around. In dissipative chaotic systems the trajectory often moves onto a so-called strange attractor, a concept that was introduced by Newhouse *et al.* (1978). We will explain its properties by way of an example.

3.1.1 Lorenz Attractor. The trajectory which is generated by Eq. (4) of the Lorenz model which has been explained in Sec. 2.2, i.e.,

$$\begin{aligned} \dot{x}&=-\sigma x+\sigma y,\\ \dot{y}&=-xz+rx-y,\\ \dot{z}&=xy-bz, \end{aligned} \tag{12}$$

leads for $r=28$, $\sigma=10$, $b=8/3$, and $\mathbf{x}(t=0)=[0,0.1,0]$ to the trajectory shown in Fig. 4. One sees that

1. the motion is confined to a bounded region in phase space;
2. the motion is erratic, i.e., the trajectory makes a few loops to the right, then a few more to the left, etc.

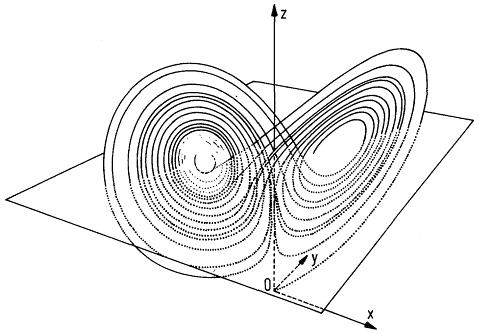

FIG. 4. The Lorenz attractor, after a computer calculation by Lanford (1977).

Numerical simulations show that there exists also a sensitive dependence on initial conditions, i.e., if instead of $\mathbf{x}(0)=[0,0.1,0]$ an adjacent initial condition is chosen, the new solution soon deviates from the old one and the numbers of right and left loops become different.

Since volume elements in phase space of dissipative systems shrink (see Fig. 1), the volume of the attracting bounded region contracts to zero. This means that the flow of the three-dimensional Lorenz system generates a set of points whose dimension is less than three. One might think that the dimension of the object spanned by the trajectory should then be two or one. However, there is no chaotic motion in systems of two dimensions or less. The fact that there are no trajectory crossings (otherwise a specific initial condition could lead to different trajectories) restricts the attractors in two dimensions to limit cycles or fixed points. The solution to this problem is that the attractor of the Lorenz system has a dimension (more precisely a Hausdorff dimension, defined in Sec. 3.2.2) which is noninteger and lies between 2 and 3; its precise value is $D=2.06$. This means that the attractor of the Lorenz system is indeed an object which deserves the name strange attractor. The general definition of a strange attractor is as follows:

1. It is an attractor, i.e., a bounded region of phase space to which all sufficiently close trajectories from the so-called basin of attraction are attracted asymptotically for long enough times. The basin of attraction is the collection of all points in phase space whose trajectories move in the limit of infinite time onto the attractor (see Fig. 5). Furthermore, the attractor itself should be indecomposable; i.e., the trajectory should visit every point on the attractor in the course of time. A collection of isolated fixed points is not a single attractor.
2. The property which makes the attractor strange is the sensitive dependence on the initial conditions; i.e., despite the contraction in volume, distances between points on the attractor need not shrink in all directions, and points which are arbitrarily close initially become exponentially separated at the attractor for sufficiently long times.

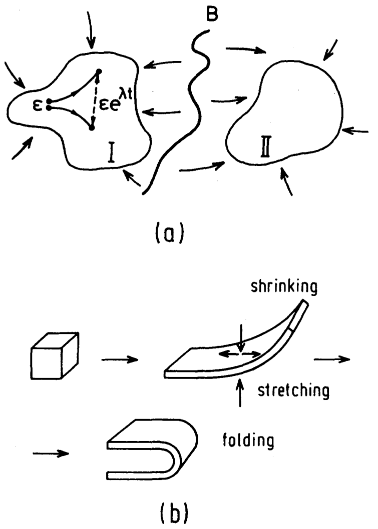

FIG. 5. (a) Two strange attractors I and II with different basins of attraction separated by a boundary *B*. (b) Deformation of a volume element on a strange attractor with increasing time.

All strange attractors that have been found up to now have fractional Hausdorff dimensions.

A strange attractor arises typically when the flow contracts the volume element in some directions, but stretches it along the others. To remain confined to a bounded domain, the volume element is folded at the same time. By analogy to the broken linear maps in Sec. 2 Eq. (7), this stretching and backfolding produces a chaotic motion of the trajectory at the strange attractor (see Fig. 5).

3.2 Characterization of Chaotic Motion

Let us now consider the question of how chaotic motion on a strange attractor could be characterized in more detail. There are in fact two essential quantities which can be computed locally for every point on the attractor:

1. the local density of points, or, more precisely, the local dimensionality, and
2. the local Liapunov exponents which measure the local instability, i.e., how fast and in what directions adjacent trajectories which start locally will diverge in the course of time.

By averaging these quantities over long times, i.e., over the whole attractor, one obtains several global numbers by which the attractor can be described.

3.2.1 Liapunov Exponents. For chaotic systems the difference between nearby trajectories $\epsilon(0)$ separates exponentially in time, i.e., $\epsilon(t)=\epsilon(0)e^{\lambda t}$. The averaged rates of separations in different directions of a dimensional phase space are the Liapunov exponents $\lambda_1,\ldots,\lambda_d$ shown in Fig. 6.

This leads for chaotic systems that are described by autonomous differential equations

$$\mathbf{x}=\mathbf{f}(\mathbf{x}) \tag{13}$$

to the following expression for the Liapunov exponents λ_i:

$$\lambda_i=\text{modulus of an eigenvalue of } \Lambda \tag{14}$$

where the matrix Λ is given by

$$\Lambda=\lim_{t\to\infty}\frac{1}{t}\log\left[\hat{T}\int_0^t M(t')dt'\right].$$

Here $M(t)$ is the matrix

$$M_{ij}(t)=\frac{\partial f_i(\mathbf{x})}{\partial x_j};\quad \mathbf{x}=\mathbf{x}[t,\mathbf{x}(0)] \tag{15}$$

and $\hat{T}$ orders the matrices $M(t)$ according to increasing times.

Figure 7 shows the largest Liapunov exponent $\lambda_{\max}$ of the Lorenz model for $\sigma=10$, $b=8/3$, as a function of the parameter r. Note that $\lambda_{\max}$ varies widely with the control parameter and becomes zero very often. This means that Lorenz model only displays chaos $\lambda_{\max}>0$ for specific values of r and chaotic regions are densely interwoven in parameter space with nonchaotic regions $\lambda_{\max}\leqslant 0$.

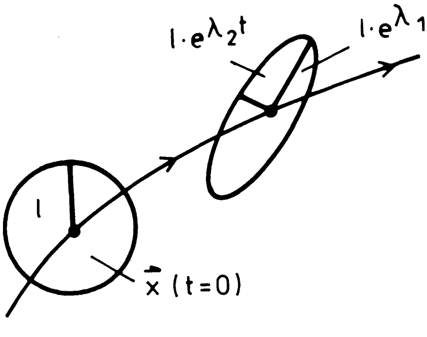

FIG. 6. Deformation of a sphere on a strange attractor in phase space as a function of time. The surface of the sphere consists of trajectory points.

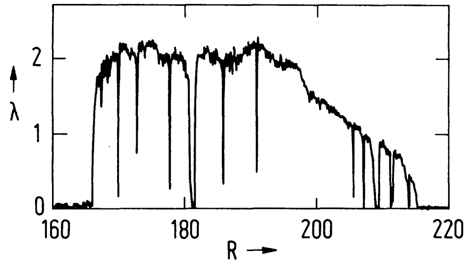

FIG. 7. The largest Liapunov exponent of the Lorenz model [Eqs. (4) for $\sigma=10$, $b=8/3$] as a function of the parameter $r=R$ (after K. Schmidt, private communication).

3.2.2 Hausdorff Dimension. Another important quality by which a strange attractor can be characterized is its fractal or Hausdorff dimension. The Hausdorff dimension D of a set of points is defined as follows: If for a set of points in d dimensions the number $N(l)$ of d-spheres of diameter l needed to cover the set increases like

$$N(l)\propto l^{-D}\quad\text{for } l\to 0 \tag{16}$$

then D is called the Hausdorff dimension of the set. A typical set of points with a noninteger Hausdorff dimension is the Cantor set shown in Fig. 8.

3.2.3 Spectrum of Attractor Dimensions. In order to characterize attractors that have a very inhomogeneous density in phase space, the notion of a Hausdorff dimension has been generalized to a spectrum $\rho(\alpha)$ of local dimensions α. The definition of a local dimension is explained in Fig. 9. The distribution $\rho(\alpha)$ of α's is obtained by measuring α for a large set of points on the attractor. One often plots instead of $\rho(\alpha)$ the function

$$f(\alpha)=\lim_{l\to 0}\frac{\log\rho(\alpha)}{\log l} \tag{17}$$

which for many strange attractors has the characteristic shape shown in Fig. 9. It can be shown that the value $\alpha_{\max}$ where $\rho(\alpha)$ [and $f(\alpha)$] have their maximum gives just the Hausdorff dimension of the attractor, i.e., $\alpha_{\max}=D$.

4. ROUTES TO CHAOS

It has been found experimentally effective for many systems to reduce the description of their dynamics to one- or two-dimensional maps by performing Poincaré sections. It has also been observed that only some general features of these maps, such as the existence of a single maximum, determine how chaos emerges if an external control parameter is tuned. Various routes to chaos have been found which differ in the way in which the signal behaves before becoming completely chaotic. Interestingly enough, all these routes have been verified experimentally, and they show a fascinating universal behavior which means that basic properties such as critical exponents near the transition to chaos depend only on some global features of the map, such that many physical systems (a whole class) display the same behavior at the transition to chaos. The beauty of nature reveals itself only after asking the right questions, and it seems that for dynamical systems the question about universal features at the transition to chaos has led to the discovery of a beautiful unifying pattern behind different phenomena. In the following we will discuss the three most prominent routes, although more routes have been found.

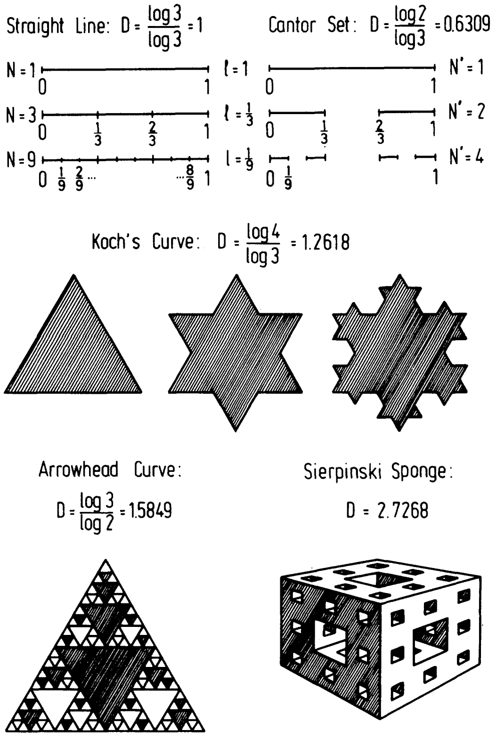

Straight Line: $D = \frac{\log 3}{\log 3} = 1$
Cantor Set: $D = \frac{\log 2}{\log 3} = 0.6309$
N = 1
N = 3
N = 9
l = 1
$l = \frac{1}{3}$
$l = \frac{1}{9}$
N' = 1
N' = 2
N' = 4
0
$\frac{1}{3}$
$\frac{2}{3}$
1
Koch's Curve: $D = \frac{\log 4}{\log 3} = 1.2618$
Arrowhead Curve: $D = \frac{\log 3}{\log 2} = 1.5849$
Sierpinski Sponge: D = 2.7268

FIG. 8. Hausdorff dimension of a straight line and of some typical self-similar point sets, so-called fractals (drawn after Mandelbrot, 1982). It is understood that the ramifications continue ad infinitum. Koch's curve is a line of infinite length that encloses a finite area, i.e., it is the perimeter of the shaded figure.

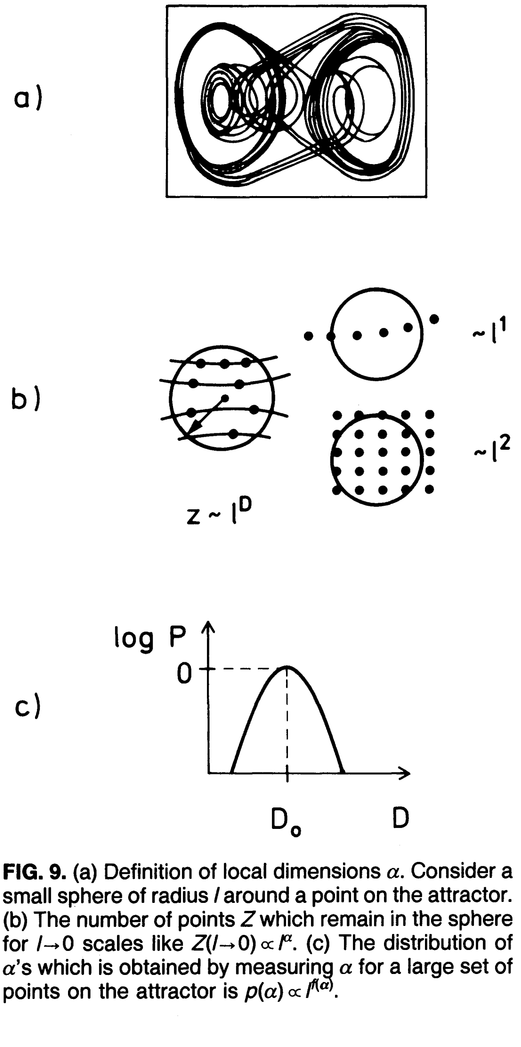

a)
b)
$\sim l^1$
$\sim l^2$
$z \sim l^D$
c)
log P
0
D_0
D

FIG. 9. (a) Definition of local dimensions α. Consider a small sphere of radius l around a point on the attractor. (b) The number of points Z which remain in the sphere for $l \to 0$ scales like $Z(l \to 0) \propto l^{\alpha}$. (c) The distribution of α's which is obtained by measuring α for a large set of points on the attractor is $p(\alpha) \propto l^{f(\alpha)}$.

4.1 Period Doubling Route

This route to chaos was discovered by studying the sequence of points generated by the logistic map

$$x_{n+1} = f(x_n) = r x_n (1 - x_n) \tag{18}$$

shown in Fig. 10. This map was already introduced by P. F. Verhulst in 1845 to simulate the

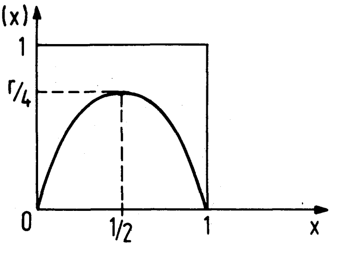

FIG. 10. The quadratic map $f(x)$ [Eq. (18)] on the unit interval.

growth of a population in a closed area. The number x_{n+1} of individuals of a species is proportional to the number in the previous year x_n and to the remaining area which is diminished proportionally to x_n, i.e., $x_{n+1}=rx_n(1-x_n)$, where the parameter r depends, e.g., on the fertility.

4.1.1 Universality and Feigenbaum Constants. In order to provide an overview over the dynamic behavior of the logistic map from Eq. (18), its iterates $x_0, x_1, \ldots, x_n$ and its Liapunov exponent

$$\lambda=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\log|f'(x_n)|$$

are plotted in Fig. 11. We distinguish a period-doubling regime for $1\leqslant r\leqslant r_\infty$ where the iterates move for $r<r_1$ to a fixed point, for $r_1<r<r_2$ to a cycle of period two, for $r_2<r<r_3$ to a cycle of period $4=2^2$, etc., until at r_∞ the cycle period becomes infinite. In this regime the Liapunov exponent is almost always negative and only becomes zero at the bifurcation points r_n where a cycle changes from period $2^{n-1}$to 2^n. The distances of the bifurcation points r_n to the accumulation point $r_\infty=3.569\,9456$ shrink geometrically as

$$r_n=r_\infty-\text{const}\times\delta^{-n} \quad \text{for } n\gg 1, \tag{19}$$

and the distances d_n of the point of a 2^n-cycle that is closest to the cycle point $x=\frac{1}{2}$ (see Fig. 12) have constant ratios:

$$d_n/d_{n+1}=-\alpha \quad \text{for } n\gg 1. \tag{20}$$

The so-called Feigenbaum constants δ and α have the values

$$\delta=4.669\,201\,6091\ldots.$$

$$\alpha=2.502\,907\,8750\ldots. \tag{21}$$

For $r_\infty<r<4$ the Liapunov exponent is mostly positive, indicating chaotic behavior that is interrupted by nonchaotic periodic windows. Although one can also find universal behavior in this region, the most important finding is that for $r_\infty<r<4$ chaotic and regular behavior are strongly intermingled on the

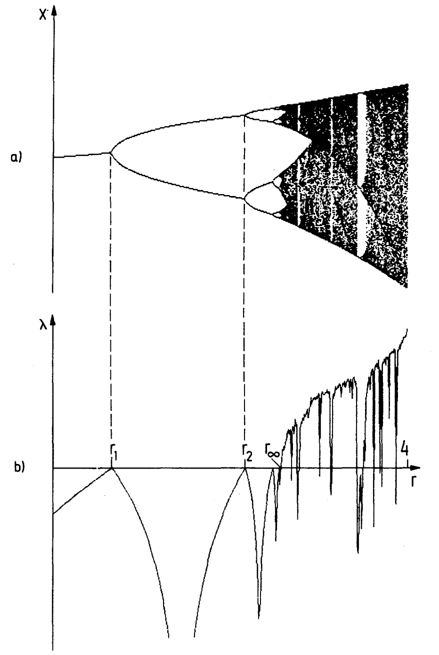

FIG. 11. (a) Iterates of the logistic map, (b) Liapunov exponent λ (after W. Desnizza, private communication).

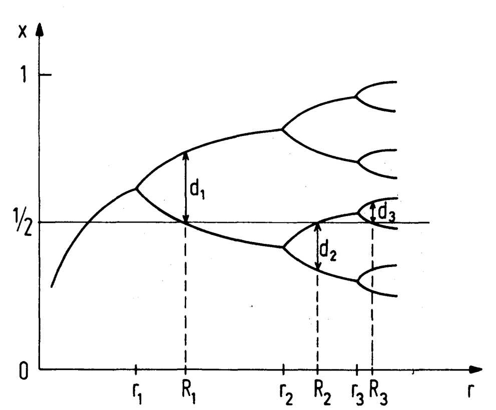

FIG. 12. Distances d_n of the fixed points closest to $x=1/2$ for superstable 2^n-cycles (schematically). Note that $x=1/2$ is an element of the cycle.

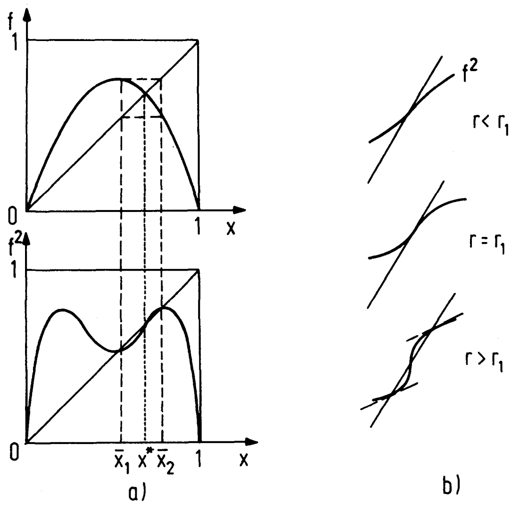

FIG. 13. (a) $f(x)$ and $f^2(x)=f[f(x)]$ for $r>r_1$. (b) Generation of two new stable fixed points in f^2 via a pitchfork bifurcation. (The bifurcation diagram looks like a pitchfork; see Table 4.)

parameter axis and the set of r values has a finite length for both types of behavior.

4.1.2 Pitchfork Bifurcation and Functional Fixed-Point Equation. The elementary mechanism which leads to period doubling is the pitchfork bifurcation shown in Fig. 13. Here a fixed point x^* of $f(x)$, i.e., $f(x^*)=x^*$, becomes unstable in favor of a cycle of period two. The points x_1 and x_2 of the two-cycle follow each other in time, i.e., $x_2=f(x_1)$ and $x_1=f(x_2)$. Since after two iterations the two-cycle closes we have $f(f(x_1))=x_1$ and $f(f(x_2))=x_2$, i.e., the elements of the two-cycle are fixed points of the second iterate $f^2(x)=f(f(x))$. This means that for the two-cycle, $f^2(x)$ plays the same role as $f(x)$ plays for a single fixed point, and so on, for the points of the 2^n-cycle are fixed points of $f^{2^n}(x)$. A comparison shows that the rescaled second iterate $(-\alpha)f^2(x/(-\alpha))$ at r_2 looks similar to the first iterate $f(x)$ at r_1, i.e., we have $(-\alpha)f^2_{r_2}(x/(-\alpha))\approx f_{r_2}(x)$ where the index denotes the values of r in $f_r(x)=rx(1-x)$. Although the details are beyond the scope of this article, it should be noted that this self-similarity has been used by Feigenbaum to relate α rigorously to the fixed-point function $g(x)$ of the so-called doubling operator T:

$$Tg(x)=-\alpha g(g(x/-\alpha))=g(x). \qquad (22)$$

This fixed-point equation can be solved by expanding $g(x)$ into a power series around $x=0.5$, and Feigenbaum (1978) obtained

$$\begin{aligned} g(x)=1&-1.527\,63(x-0.5)^2\\ &+0.104\,815(x-0.5)^4\\ &+0.026\,7057(x-0.5)^6\ldots\,. \end{aligned}$$

It follows from Eq. (22) that for a normalized solution with $g(0)=1$ the Feigenbaum constant α is given by

$$\alpha=1/g(1)=2.502\,807\,876\ldots\,. \qquad (23)$$

This rather elegant relation of the Feigenbaum constant α to a special value of the general fixed-point equation (22) establishes the universality of α because its value depends only on the boundary conditions for the solution of Eq. (22), i.e., on the order of the maximum of $g(x)$ (quadratic, quartic, etc.) at $x=0.5$.

The other Feigenbaum constant δ is the largest eigenvalue of the linearized doubling operator $L\{h(x)\}$. It can be computed from the eigenvalue equation

$$\begin{aligned} L\{h(x)\}=&-\alpha\{g'[g(x/-\alpha)]h(-x/\alpha)\\ &+h[g(-x/\alpha)]\}=\delta h(x). \qquad (24) \end{aligned}$$

Since the eigenfunction $h(x)$ and the eigenvalue δ depend, according to Eq. (24), only on the universal quantities $g(x)$ and α, it follows that δ itself also has a universal character.

4.1.3 Experimental Observation of Period Doubling. The period-doubling route has been discovered in many experiments. A typical example is shown in Fig. 14. Here the nonlinear element is the capacitor diode whose voltage-charge relation is nonlinear. Since at each doubling of the period of the signal its frequency is reduced to one half, one sees an increasing number of low-frequency peaks in Fig. 14.

4.2 Intermittency

The intermittency route to chaos is characterized by the appearance of a signal which alternates randomly between long regular phases and relatively short irregular bursts. A typical example for an intermittent signal is shown in Fig. 15. With increasing control parameter the lengths of the laminar regions shrink and the chaotic regimes become longer

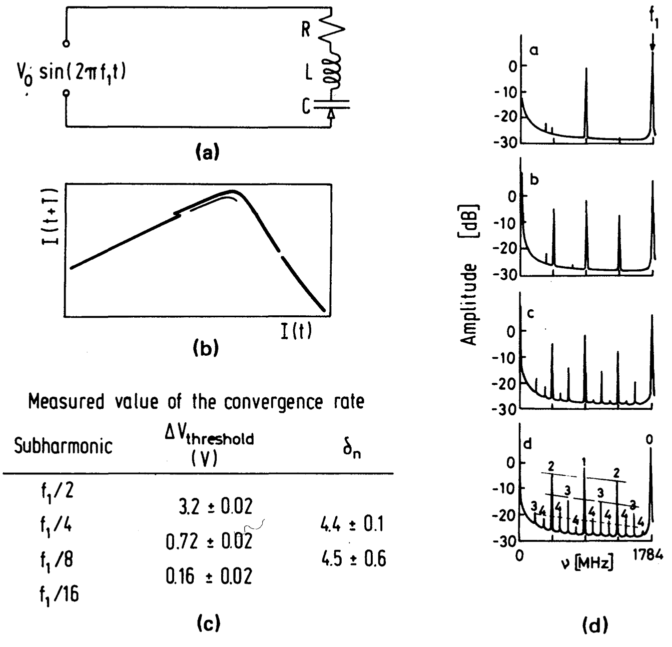

Measured value of the convergence rate

Subharmonic	$\Delta V_{\text{threshold}}$ (V)	δ_n
$f_1/2$		
	3.2 ± 0.02	
$f_1/4$		4.4 ± 0.1
	0.72 ± 0.02	
$f_1/8$		4.5 ± 0.6
	0.16 ± 0.02	
$f_1/16$		

FIG. 14. (a) Circuit for the driven nonlinear RCL oscillator. (b) The observed current $I(t+T)$ vs $T(t)$ yields a one-dimensional map with a single maximum. (c) Determination of δ from the values of the control parameter V_0. (d) *a–c*: Subharmonics in the power spectrum for increasing V_0; *d*: comparison with Feigenbaum's theory (horizontal lines). (After Linsay, 1981.)

and longer. This offers a continuous route from regular to chaotic motion.

4.2.1 Mechanism for Intermittency. A mechanism for the intermittency route in terms of one-dimensional Poincaré maps has first been given by Manneville and Pomeau (1979). Infinitely long laminar regions correspond to period-one cycle oscillations, i.e., a fixed point in the Poincaré map. If the control parameter is increased this fixed point becomes unstable. Because there are three ways in which a fixed point can become unstable, Pomeau and Manneville distinguished three types of intermittency (I, II, III) shown in Table 3. In all three cases the signal leaves the unstable fixed point only slowly, i.e., there is a sort of memory of the fixed point which leads to the typical regular laminar signals that increase slowly with time n. Above a certain value of the laminar signal the motion becomes chaotic until the signal becomes reinjected to the vicinity of the unstable fixed point, where it becomes again laminar. The theory of Pomeau and Manneville explains only the characteristic behavior in the laminar region, e.g., how the mean length $\langle l \rangle$ of these regions, averaged over long time scales, varies with the control parameter; but it makes no statements about the chaotic regime. The different scaling behavior together with the

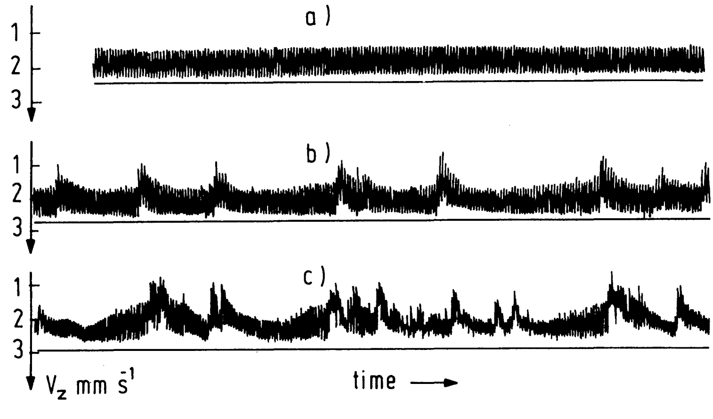

FIG. 15. Intermittency for a Bénard experiment: time dependence of the vertical velocity component measured in the middle of a Bénard cell. It changes with increasing Rayleigh number from periodic motion (a) via intermittent motion (b) to chaos (c) (after Bergé *et al.*,1980).

Table 3. Characteristic properties of different types of intermittency. The different types of intermittency can be distinguished by their Poincaré maps, their forms of the laminar signal, and their distributions $P(l)$ of laminar lengths.

Type	Poincaré map	Laminar Signal	Distribution $P(l)$
I	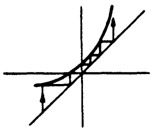 $x_{n+1}=x_n+x_n^2+\epsilon$	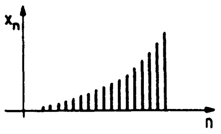 increases monotonically	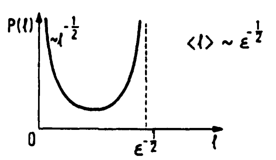
II	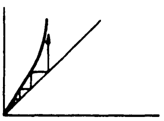 $r_{n+1}=(1+\epsilon)r_n+ur_n^3$ $\theta_{n+1}=\theta_n+\Omega$	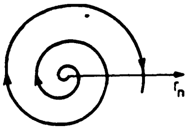 spirals	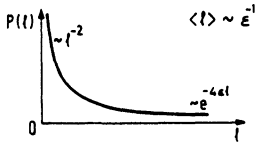
III	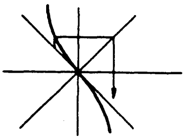 $x_{n+1} = -(1+\epsilon)x_n - ux_n^3$	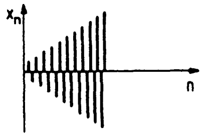 alternates	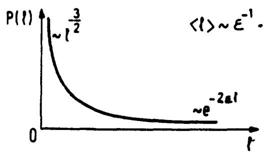

distribution $P(l)$ of laminar lengths is also shown in Table 3.

4.2.2 Experimental Observation of Intermittency. Figure 16 shows that the intermittency route can be seen in the same experimental setup which displays period doubling (see Fig. 14), albeit for different values of the control parameter.

4.3 Ruelle–Takens–Newhouse Route

This route to chaos can be visualized most easily by displaying the Fourier spectrum of the signal. Figure 17 shows the power spectrum of the fluid velocity at a special point in a Bénard cell (see Table 2). With increasing temperature difference the following states are observed:

1. periodic movement with one frequency and its harmonics,
2. quasiperiodic motion with two incommensurate frequencies, and their linear combinations,
3. nonperiodic chaotic motion with some sharp lines,
4. chaos.

This experiment was in fact performed after Newhouse, Ruelle, and Takens (1979) suggested this route to chaos. Its characteristic features are as follows. As the external control parameter is changed the system undergoes first two Hopf bifurcations which are characterized by the facts that first a fixed point becomes unstable in favor of a limit cycle on which the trajectory moves with a frequency ω_1. The second Hopf bifurcation generates

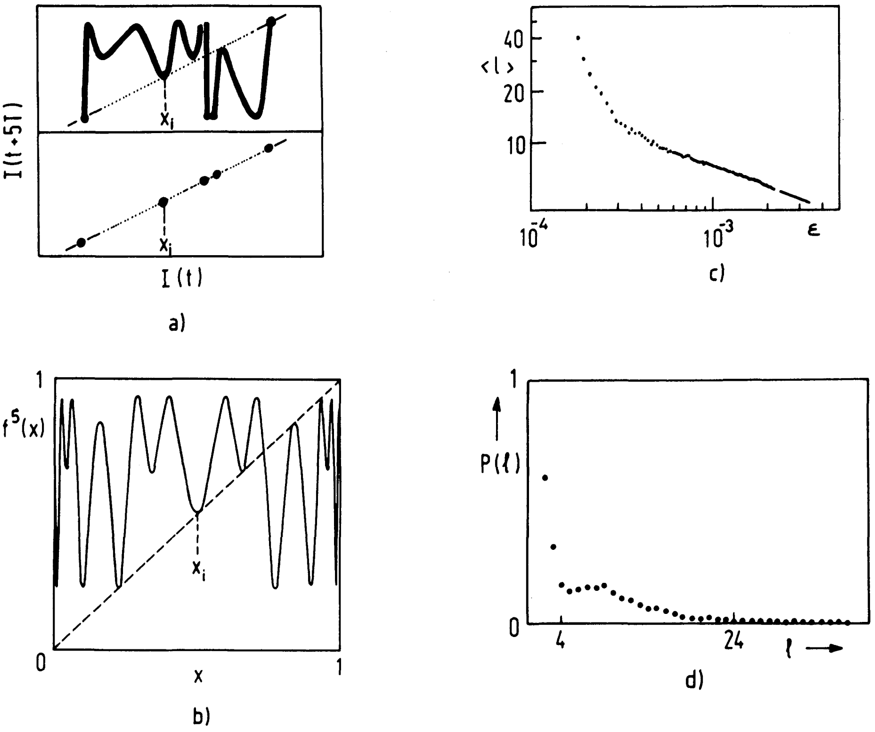

FIG. 16. Intermittency in the nonlinear RCL oscillator: (a) $I(t+5T)$ vs $I(t)$ which corresponds to the fifth iterate of the logistic map at tangency which is shown in (b). (c) The measured averaged length of the laminar regions scales like $\langle l \rangle \propto \epsilon^{-0.43}$ (where $\epsilon \sim V_0 - V_c$) [V_0 is the external control parameter and V_c is its value at the critical point where $I(t+5T)$ vs $I(t)$ is at tangency with the bisector], which is in reasonable agreement with the prediction of Manneville and Pomeau, $\langle l \rangle \propto \epsilon^{-0.5}$. (d) $P(l)$ vs laminar lengths l (in units of 5 T) for $\epsilon = 2.5 \times 10^{-4}$. (After Jeffries and Péres, 1982.)

another frequency ω_2 (which is generally incommensurate with ω_1, i.e., the ratio ω_2/ω_1 is irrational), and the trajectory moves on a torus (see Fig. 18). Further increase of the control parameter yields a "three-torus" with three independent frequencies, which, however, decays with finite probability onto a strange attractor. This breakup of the torus can be visualized in the experimentally observed Poincaré section shown in Fig. 19.

4.4 Comparison of Three Routes to Chaos

Table 4 summarizes the three different routes to chaos which we have discussed up to now. In the bifurcation diagram the behavior of stable (s) and unstable (u) solutions is shown as a function of the control parameter r.

This table, however, should only be considered as a first approximation to the true variety of transition scenarios. While it is natural to focus on common features, it would be premature to make sweeping generalizations about routes to chaos, and it should be emphasized that the range of dynamical behavior observed is quite large.

5. EXPERIMENTS AND APPLICATIONS

Having observed a seemingly chaotic signal $x(t)$, one wants to know what information it contains about the strange attractor.

5.1 Reconstruction of the Attractor from a Measured Time Series

Before one can extract the relevant variables λ_i and $f(\alpha)$ mentioned above in Sec. 3.2, one must in the first place reconstruct the trajectory $\mathbf{x}(t)$ of the system in phase space. This problem has been solved by Takens

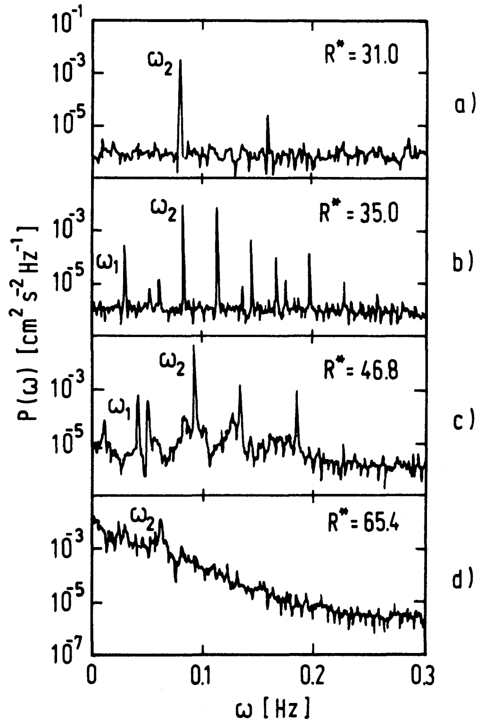

FIG. 17. Power spectrum of the convection current for a Bénard experiment (after Swinney and Gollub, 1978). With increasing (relative) Rayleigh number $R^* = R/R_c$ (where R_c is the critical value of the Rayleigh number, above which convective rolls develop in a Bénard cell), the following states are observed: (a) periodic movement with one frequency and its harmonics, (b) quasi-periodic motion with two incommensurate frequencies and their linear combinations, (c) nonperiodic chaotic motion with some sharp lines, (d) chaos.

(1981), and we illustrate his method by means of a simple example. Suppose that the true trajectory is a circle, i.e.,

$$\mathbf{x}(t) = [x_1(t), x_2(t)] = (\sin t, \cos t), \tag{25}$$

where the parameter t denotes time in dimensionless units. If we had only measured the variable $x_1(t) = \sin t$ we could reconstruct the second component $x_2(t)$ of the vector $\mathbf{x}(t)$ in phase space by performing a phase shift in $x_1(t)$, i.e.,

$$x_2(t) = \cos t = \sin(t + \pi/2) = x_1(t + \tau);$$

$$\tau = \pi/2 \tag{26}$$

Takens has shown that this method works also for the reconstruction of vectors in higher dimensional phase space, i.e.,

$$[x_1(t) \ldots x_m(t)]$$
$$= \{x_1(t), x_1(t+\tau) \ldots x_1[t+(m-1)\tau]\}, \tag{27}$$

yielding a strange attractor which has the same topological features as the "true" attractor (but it need not "look" the same).

Figure 20 shows a reconstruction of (a projection of) the Rössler attractor (Rössler, 1976), which is generated by the system

$$\begin{aligned} \dot{x} &= -y, \\ \dot{y} &= x + ay, \\ \dot{z} &= b + z(x - c), \end{aligned} \tag{28}$$

from about 6×10^5 points for different choices of the delay time τ. Although for an infinite amount of noise-free data, τ could be chosen almost arbitrarily (Takens, 1981), it can be seen from Fig. 20 that for a finite time series the quality of the reconstruction depends on τ. If τ is too small, $x(t)$ and $x(t+\tau)$ become practically indistinguishable and one obtains a linear dependence that is not present for the coordinates of the real trajectory. It is, therefore, reasonable to choose the decay time of the autocorrelation function $C(t)$ of the signal x_n for τ:

$$C(t) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N-1} x_n x_{n+1} \equiv \langle x_0 x_t \rangle, \tag{29}$$

$$C(t) \approx \tfrac{1}{2} C(0),$$

which ensures that $x(t)$ and $x(t+\tau)$ become linearly independent; but other choices for τ have also been proposed (Fraser and Swinney, 1986).

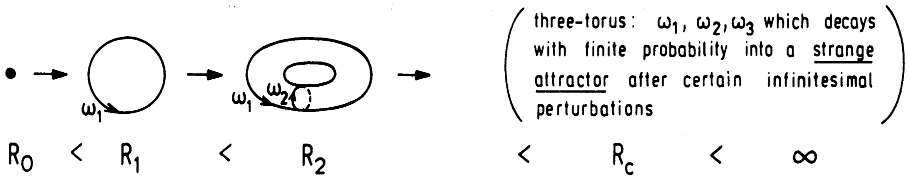
ω_1
ω_1
ω_2
$R_0 < R_1 < R_2$
three-torus: $\omega_1, \omega_2, \omega_3$ which decays with finite probability into a strange attractor after certain infinitesimal perturbations
$< R_c < \infty$

FIG. 18. The Ruelle–Takens–Newhouse route to chaos.

Table 4. Summary of three main routes to chaos.

Feigenbaum	Manneville-Pomeau	Ruelle–Takens–Newhouse
Pitchfork bifurcation	Tangent bifurcation	Hopf bifurcation
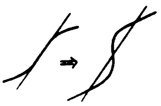	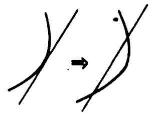	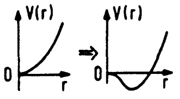
	Bifurcation diagrams	
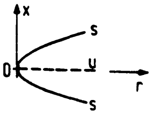	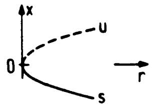	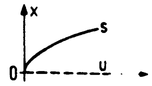
	Main phenomena	
Infinite cascade of period doublings with universal scaling parameters	Intermitted transition to chaos. The laminar phase has a duration $(r - r_c)^{-1/2}$	After three bifurcations, strange attractor "probable."
	Experiments	
Bénard experiment Taylor experiment Driven nonlin. oscill. Chemical reactions Optical instabilities	Bénard experiment Josephson junction Chemical reactions Lasers	Bénard experiment Taylor experiment Nonlin. conductors

5.2 Correlation Integral

Having reconstructed $\mathbf{x}(t)$ one could proceed to calculate the spectrum of dimension and the Liapunov exponents as indicated in Sec. 3. But for the calculation of fractal dimensions there exists the problem that in order to compute $p(l) \sim l^{\alpha}$ one has to cover the whole attractor with little spheres of diameter l and then count how the number of points in them change with l. In order to avoid counting the many empty boxes Grassberger and Procaccia (1983) introduced the correlation integral as a rather elegant method to compute fractal dimensions, and it has found wide application. The correlation integral $C(l)$ measures the probability of finding two points of an attractor in a box of size l:

$$
\begin{aligned}
\sum_i p_i^2 &= \text{the probability that two points of the attractor lie within a cell } l^d \\
&\simeq \text{the probability that two points at the attractor are separated} \\
&\quad\ \text{by a distance smaller than } l \\
&= \lim_{N\to\infty} \frac{1}{N^2}\{\text{number of pairs } ij \text{ whose distance } |\mathbf{x}_i - \mathbf{x}_j| \text{ is less than } l\} \\
&= \lim_{N\to\infty} \frac{1}{N^2} \sum_{ij} \theta\,(l - |\mathbf{x}_i - \mathbf{x}_j|) = C(l) = \text{correlation integral.}
\end{aligned}
\tag{30}
$$

Here $\theta(z)$ denotes the step function with $\theta(z>0)=1$ and $\theta(z<0)=0$.

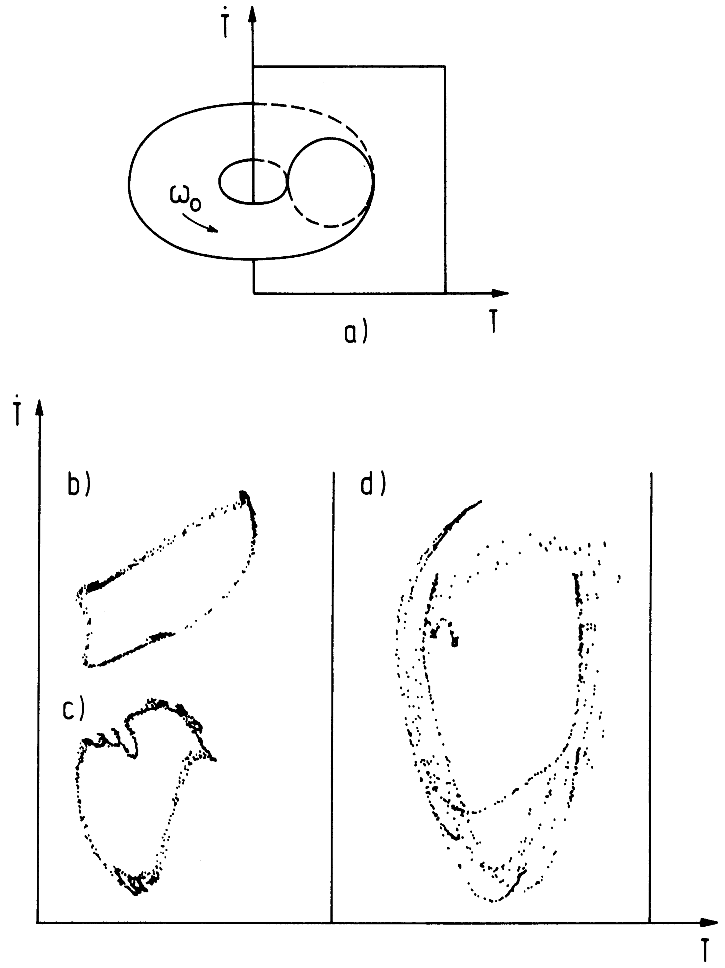

FIG. 19. Poincaré sections for the Bénard experiment: (a) schematic section through torus; (b)–(d) experiments showing with increasing Rayleigh number a transition from quasiperiodic motion (b) to substructures indicating the destruction of the torus (c) and then to a strange attractor (d). (After Dubois, Bergé and Croquett, 1982.)

Since the probabilities p_i scale like $p_i \sim l^{\alpha_i}$, where α_i are the local dimensions, it can be shown that the correlation integral scales like

$$C(l) \sim l^{-D_2} \quad \text{for } l \to 0, \tag{31}$$

where the correlation dimension D_2 yields a lower bound to the Hausdorff dimension D, i.e., $D_2 \leqslant D$.

5.3 Applications

5.3.1 Lorenz Model. Figure 21 shows how Takens's reconstruction works together with the correlation integral method to yield from about 10^4 points the correlation dimension D_2 for the Lorenz attractor.

5.3.2 Taylor Instability. Another example where Takens's method works has been observed experimentally in Taylor instability. The Taylor instability occurs in a fluid layer between an inner cylinder rotating with an angular velocity Ω and a stationary outer cylinder (see Fig. 22). For small Ω, angular momentum fed to the inner cylinder is transported outside by viscosity. Above a critical angular velocity Ω_c', this state becomes unstable, and momentum is transported by annular convection cells. At still higher Ω's, periodic and multiply periodic oscillations of these

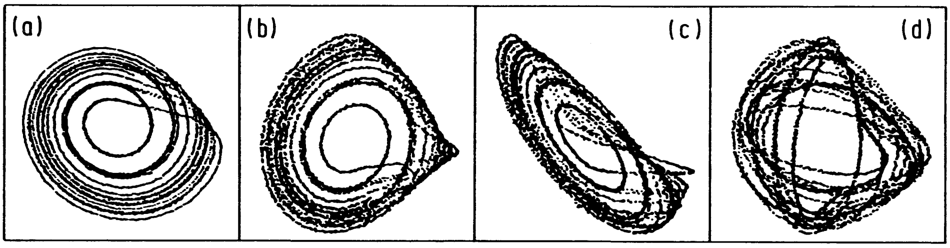

FIG. 20. Reconstruction of the Rössler attractor for $a=0.15$, $b=0.20$, $c=10.0$, $x_0=(10.0, 0, 0)$ from a time series: (a) x, y coordinates of the "true" attractor obtained by numerical integration of Eq. (27); (b), (c), (d) reconstructions for $\tau=0.23$, 0.39, 3.26, measured in units of the average orbital time, respectively. (After Fraser and Swinney, 1986.)

occur which merge into chaos after two Hopf bifurcations.

The following results in Fig. 23 have been obtained by reconstructing the phase space for a Taylor experiment from a series of the radial velocity measured at a special point in the Taylor cylinder

$\{v(t_k),\ldots,v(t_k+m\tau)\}$ with $t_k=k\tau_0$,

$k=0,1,2,\ldots$ $(\tau_0<\tau)$:

1. The Poincaré section shows the breakup of a torus.
2. The largest Liapunov exponent λ (obtained from separation of nearby orbits in five-dimensional phase space) becomes positive for $\Omega>\Omega^*$. This proves experimentally the existence of a strange attractor.
3. The Hausdorff dimension D and D_2 increase slowly with Ω/Ω_c. This shows that there are only a few relevant degrees of freedom even at Ω values that are 30% above the critical value $\Omega^*\approx 12\Omega_c$ at the onset of chaos.

5.3.3 Medical Data.

Figure 24 shows the logarithm of the correlation integral $C(l)$ as a function of $\log l$ and phase-space dimension n as computed from data points of a magnetoencephalogram for (a) a normal and (b) an epileptic patient. The correlation dimension, which is given by the average slope of these curves, is smaller for the epileptic patient. This is a first hint of more regular brain waves which have a higher tendency to culminate

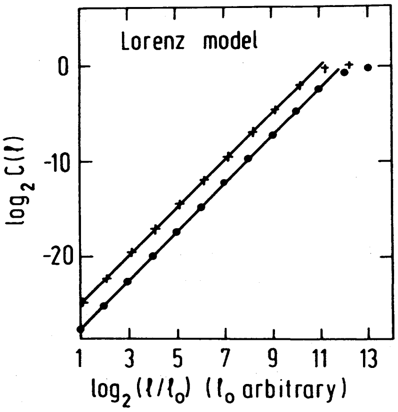

FIG. 21. $\log C(l)$ vs $\log l$ for the Lorenz model. The lower line is obtained directly from the three-dimensional time series $\{x(t_i), y(t_i), z(t_i)\}$, whereas the upper line originates from the reconstructed series $\zeta(t_i)=\{x(t_i), x(t_i+\tau), x(t_i+2\tau)\}$. The slopes of both curves are the same, i.e., the correlation dimension $D_2=2.05$ obtained by both methods is the same, as stated above (Grassberger and Procaccia, 1983).

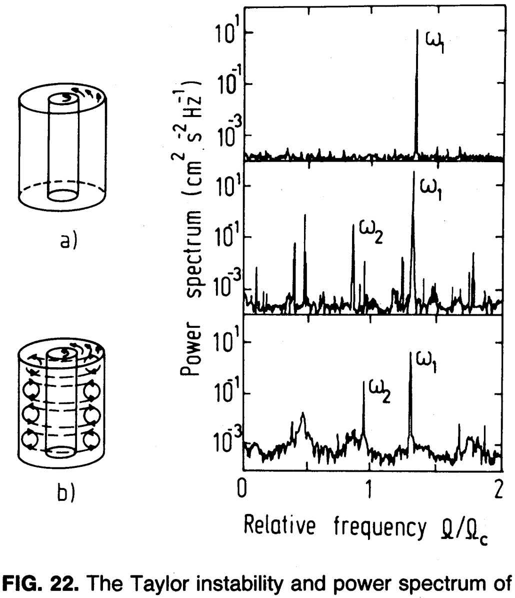

FIG. 22. The Taylor instability and power spectrum of the radial velocity measured at a point in the Taylor cylinder (after Swinney and Gollub, 1978).

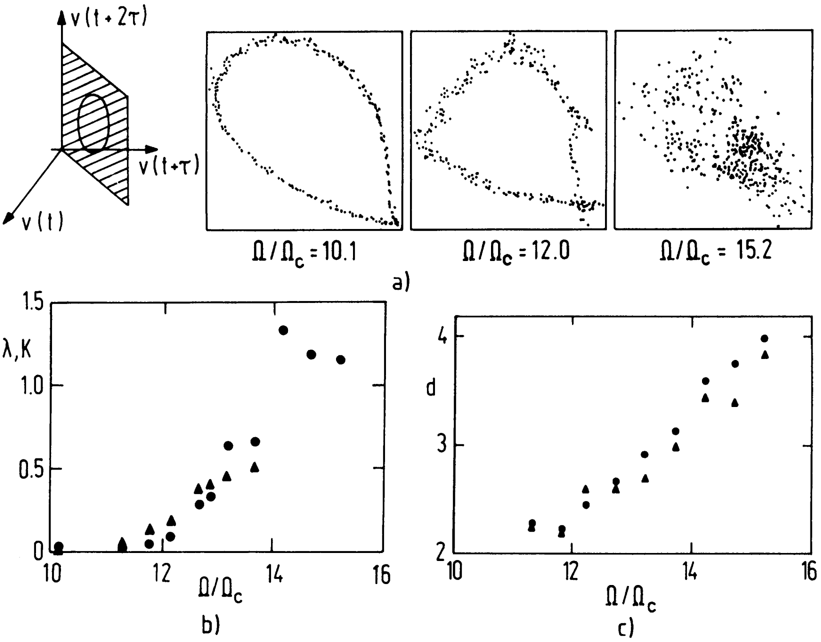

FIG. 23. Experimental properties of a strange attractor which occurs in a Taylor experiment: (a) Plane of the Poincaré section and break up of the torus with increasing Ω. (b) K-entropy (▲) and largest Liapunov exponent (●) vs Ω/Ω_c. (c) Hausdorff dimension D (●) and correlation dimension D_2 (▲) vs Ω/Ω_c. (After Brandstäter *et al.*, 1983.)

in regular bursts of an epileptic seizure than in a normal patient with more irregular neural activity. This last example shows that rather esoteric looking measures of chaos have quite practical medical applications.

6. CONSERVATIVE SYSTEMS

Conservative systems differ from dissipative systems by the fact that volume elements in phase space are conserved. Figure 25 shows the different behavior of trajectories in phase space in the vicinity of a fixed point for both types of dynamical systems. Conservative systems encompass all dynamical systems of classical mechanics, i.e., their time evolution is governed by Hamilton's equation for the coordinates **q** and momenta **p**:

$$\dot{q}_i=\frac{\partial H}{\partial p_i},\quad \dot{p}_i=-\frac{\partial H}{\partial q_i},\quad i=1,\ldots,d, \tag{32}$$

and $H=H(\mathbf{p},\mathbf{q})$ is the Hamilton function which specifies the energy E of the system. The simplest example for a conservative system is a harmonic oscillator, whose Hamilton function reads (after proper normalization)

$$H=\tfrac{1}{2}(p^2+\omega^2q^2), \tag{33}$$

where ω is the oscillator frequency. The corresponding trajectories in phase space are ellipses (since $H=E=$const.) as shown in Fig. 25(b).

The fact that volumes do not change in conservative systems implies immediately that they display (in contrast to dissipative systems) no attracting regions in phase space, i.e., no attracting fixed points, no attracting limit cycles, and no strange attractors. Nevertheless, in conservative systems one also finds chaotic regions in phase space, but they are not attractive and can be densely interwoven with regular regions.

For some time attention has shifted from the calculation of individual orbits to consid-

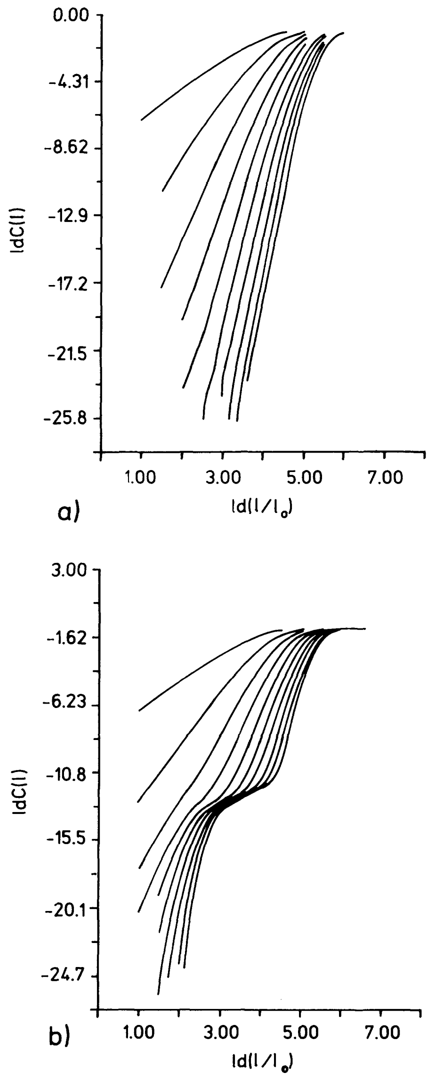

FIG. 24. Correlation integral $C(l)$ vs $\log(l)$ for (a) a normal subject and (b) an epileptic patient based on data from magnetoencephalograms. (After Saermark *et al.*, 1988.)

eration of the qualitative properties of families of orbits as shown in Fig. 26.

6.1 Integrable Systems

A conservative system is called integrable if one can find a transformation of the coordinates **q** and momenta **p** such that in the new variables $\boldsymbol{\theta}$, **I** the equations of motion can be integrated as

$$I_i = \text{const}_i,$$

$$\theta_i = \omega_i t + \delta_i, \quad i = 1 \ldots n, \tag{34}$$

i.e., the motion in phase space can be described by angular variables on a torus as shown in Fig. 27. In the following, we will confine ourselves to systems with $n=2$ degrees of freedom, but most results can be extended $n>2$. Figure 27 shows the motion of an integrable system with two degrees of freedom (i.e., with a four-dimensional phase space) on a torus. Closed orbits occur only if

$$n\Delta\theta_2 = 2\pi m, \text{ i.e., } \omega_2/\omega_1 = m/n = \text{rational};$$

$$m,n = 1,2,3\ldots . \tag{35}$$

For irrational frequency ratios, the orbit never repeats itself but approaches every point on the two-dimensional manifold infinitesimally closely in the course of time. In other words, the emotion is ergodic on the torus.

6.2 Kolmogorov–Arnol'd–Moser Theorem

The Kolmogorov–Arnol'd–Moser theorem states that under certain technical conditions the motion of an integrable system which is perturbed externally remains stable. This theorem is important, for example, for such questions as the stability of the solar system or the stability of trajectories in a particle accelerator in the long time limit. More precisely the Kolmogorov–Arnol'd–Moser (KAM) theorem makes a statement about the dynamical behavior which occurs if an integrable system whose frequency ratio ω_1/ω_2 is close to an irrational value is perturbed. The theorem states that under certain technical conditions, those tori whose motion is sufficiently irrational that

$$\omega_1/\omega_2 - m/s > K(\epsilon)/s^{2.5}, \tag{36}$$

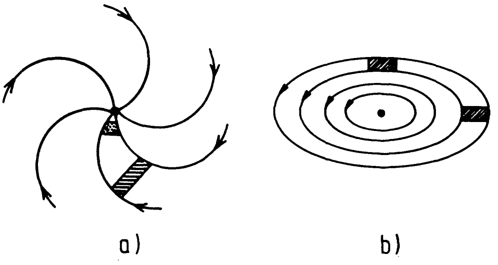

FIG. 25. (a) In dissipative systems trajectories are attracted to a fixed point, and volume shrinks. (b) In conservative systems the points rotate around an elliptic fixed point; volume is conserved.

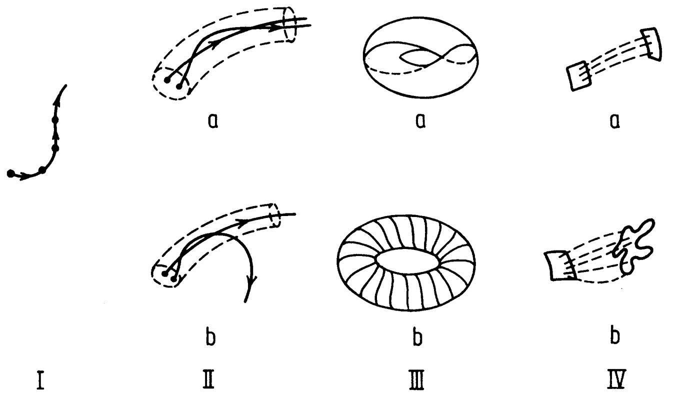

FIG. 26. Problems of increasing globality in classical mechanics. I. Step by step integration of the equations of motion. II.(a) Local stability; (b) Local instability. III. Topological nature of complete trajectories: (a) Periodic motion on a torus; (b) Motion on a torus with irrational frequency ratios. IV. Types of flow in phase space: (a) nonmixing; (b) mixing. (After Balescu, 1975.)

where m and s are mutually prime integers, are stable under a perturbation whose strength is proportional to ϵ. $K(\epsilon)$ is a constant with $K(\epsilon \to 0) = 0$. It is important to note that the set of initial conditions for which Eq. (36) holds has a finite volume in phase space, i.e., there is a finite probability that the motion remains stable under perturbation.

6.3 Coexistence of Regular and Irregular Motion

If ω_1/ω_2 is rational, the original torus decomposes into smaller and smaller tori. Some of these newly created tori are again stable according to the KAM theorem. But between the stable tori there appear unstable fixed points. Figure 28 shows that, near a hyperbolic fixed point H, the motion becomes unstable and orbits are driven away from it, in contrast to the stable rotational motion around an elliptic fixed point.

The stable (W_s) and unstable (W_u) lines which lead to or emanate from H behave highly irregularly since

1. they cannot intersect themselves (otherwise the motion on a trajectory in phase space would not be unique for a given set of initial conditions);
2. but W_u can intersect W_s at a so-called homoclinic point (see Fig. 29).

According to the so called Lamba lemma [see, e.g., Guckenheimer and Holmes (1983), p. 248] the curves W_u and W_s form an extremely complex network which characterizes chaotic motion.

It is shown in Fig. 30 that stable and unstable regions are densely interwoven. So-called elliptic fixed points are stable and surrounded by stable KAM tori, whereas hyperbolic fixed points are unstable and in their vicinity the motion becomes chaotic.

Whereas for systems with two degrees of freedom the irregular orbits are trapped between stable KAM tori, this is no longer the case if the number of degrees of freedom exceeds two. Then so-called Arnol'd diffusion (Arnol'd, 1964) of trajectories occurs between tori as shown in Fig. 31.

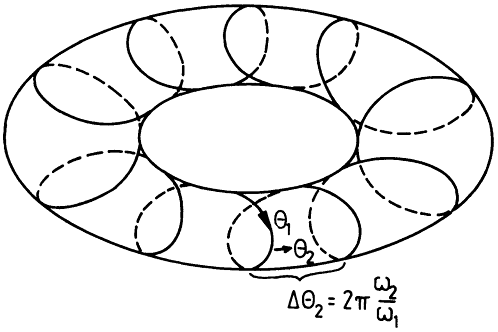

FIG. 27. Torus in phase space.

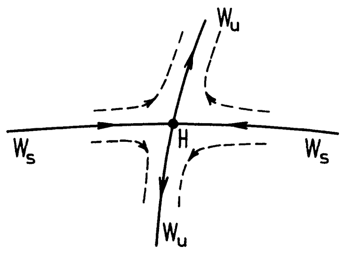

FIG. 28. Hyperbolic fixed point H with stable (W_s) and unstable (W_u) lines.

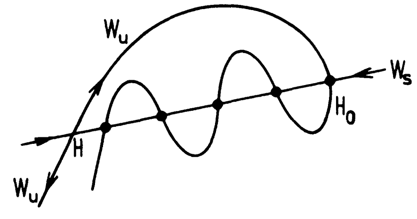

FIG. 29. Homoclinic points H_0 are the intersections of W_u and W_s.

6.4 Unstable Planetary Motion

An example of the coexistence of regular and irregular trajectories in a Hamiltonian system is given by the motion of an asteroid around the sun perturbed by Jupiter. This three-body problem is nonintegrable, and according to Eq. (36) we expect that the asteroid motion becomes unstable if the ratio of the unperturbed frequency of the asteroid motion ω and the angular frequency of Jupiter ω_J becomes rational. Figure 32 illustrates that, in fact, gaps occur in the asteroid distribution f for rational ω/ω_J. On the other hand, the existence of stable asteroid orbits ($f \neq 0$) can be considered as a confirmation of the KAM theorem.

7. CHAOS IN QUANTUM SYSTEMS?

The existence of chaotic motion in classical conservative systems naturally leads to the question of how this irregularity manifests itself in the corresponding quantum systems. In a broader context, one might inquire about the nature of the solutions to the wave equations that arise, e.g., in plasma physics, optics, or acoustics, whose ray trajectories, e.g., in geometric optics, are stochastic.

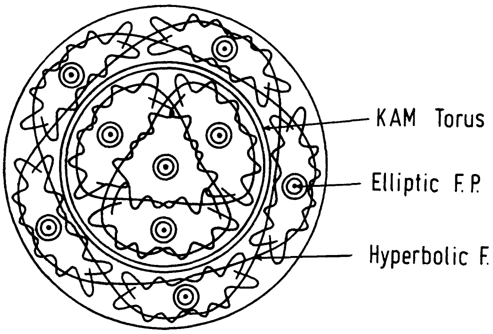

FIG. 30. Regular and irregular motion in the phase space of a nonintegrable system.

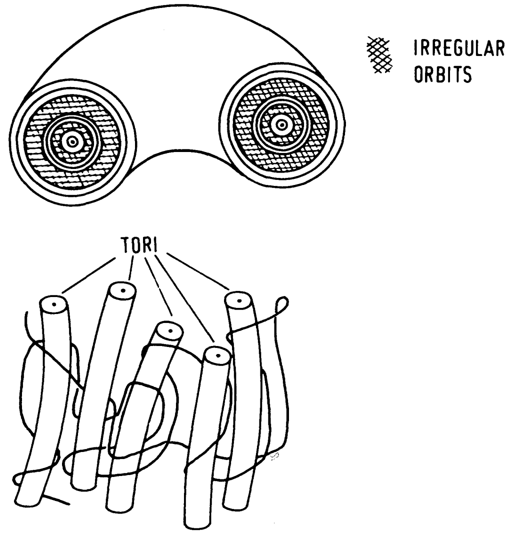

FIG. 31. Arnol'd diffusion for Hamiltonian systems with more than two degrees of freedom (schematically).

The question about the behavior of quantum systems whose classical limit exhibits chaos has been posed since the early days of quantum mechanics (Einstein, 1917) because it raises the problem of how to quantize a system which executes nonperiodic motion. At that time, periodic systems were quantized via the Bohr-Sommerfeld quantization rule $\oint p\, dq = nh$, where h is Planck's constant and $\oint$ denotes integration over a closed trajectory, which occurs for periodic motion. Since the discovery and the establishment of wave mechanics, we know how to proceed if we wish to learn about the time evolution of any quantum system: solve the time-dependent Schrödinger equation

$$\hat{H}\psi = -\frac{\hbar}{i}\frac{\partial}{\partial t}\psi, \qquad (37)$$

where $\hat{H}$ is the Hamilton operator of the system, ψ is its wave function, and $\hbar = h/2\pi$.

7.1 Classical versus Quantum Systems

In order to develop some intuition for the changes which will arise if we pass from a classically chaotic system to its quantum mechanical version, we recall several major differences between classical and quantum systems:

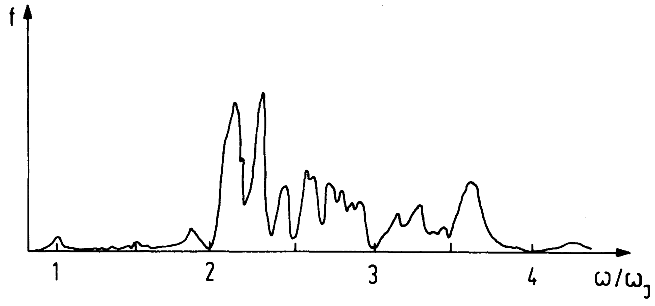

FIG. 32. Fraction f of asteroids in the belt between Mars and Jupiter as a function of ω/ω_J. (After Berry, 1978.)

1. In contrast to classical mechanics (where a statistical description is only necessary if the system becomes chaotic in time), quantum mechanics allows only statistical predictions. Although the Schrödinger equation is linear in ψ and can, e.g., be solved exactly for a harmonic oscillator with the result that ψ depends regularly on time (i.e., there is no chaotic time behavior), this does not mean that the motion is completely deterministic, since $|\psi(\mathbf{x}, t)|^2$ is only the probability density to find an electron at a space-time point $(\mathbf{x}, t)$.
2. Because of Heisenberg's uncertainty principle

   $$\Delta p \Delta q > \hbar/2 \tag{38}$$

   there are no trajectories in quantum mechanics [if one measures q with precision Δq, one disturbs the momentum p by Δp according to Eq. (38)]. Therefore, the characterization of chaos based on the exponentially fast separation of nearby trajectories becomes useless for quantum systems.
3. The uncertainty principle Eq. (38) implies also that points in $2n$-dimensional phase space within a volume $\hbar^n$ cannot be distinguished, i.e., the phase space becomes coarse grained. This means that regions in phase space in which the motion is classically chaotic (see Fig. 31), but which have volumes smaller than $\hbar^n$, are not "seen" in quantum mechanics; and for the corresponding quantum system, we expect a regular behavior in time. Thus the finite value of Planck's constant tends to suppress chaos. On the other hand, the limit $\hbar \to 0$ becomes difficult (for quantum systems which have a classical counterpart which displays chaos) because if $\hbar$ becomes smaller, more irregular structures will appear.

7.2 A Quantum Particle in a Stadium

Although it was made plausible in the previous section that the finite value of Planck's constant tends to suppress chaos, one nevertheless expects some difference between a quantum system having a classical counterpart that displays chaos and a quantum version of an integrable classical system having regular trajectories. To cast some light on this problem, McDonald and Kaufmann (1979) calculated numerically the wave functions and spectra of a free particle in a stadium (i.e., a rectangular region of length R which is closed at two opposite ends by semicircles with radius a) and in a circular disk by solving the Schrödinger equation for a free particle in two dimensions,

$$\nabla^2\psi = E\psi, \tag{39}$$

with the boundary condition $\psi(x, y) = 0$ at the "walls."

In the classical limit most (not all) trajectories of a free moving particle in a stadium (but not in a circle) are chaotic because the curved boundary reflects the particle in such a way that nearby trajectories separate exponentially in time. The results of McDonald and Kaufmann for the quantum situation are summarized as follows (see Fig. 33):

1. The eigenfunctions of the stadium problem show irregular nodal curves [where $\psi(x, y) = 0$] in contrast to the regular curves for the circle.
2. The distribution $N(\Delta E)$ of the eigenvalue spacings ΔE for the circle shows a maximum at $\Delta E = 0$, i.e., there is a high probability of level degeneracies, and one finds level clustering. It has been proved by Berry and Tabor (1977) that for integrable systems $N(\Delta E) \propto \exp(-\Delta E \times \text{const})$. For the stadium $N(\Delta E)$ has a maximum at $\Delta E \neq 0$,

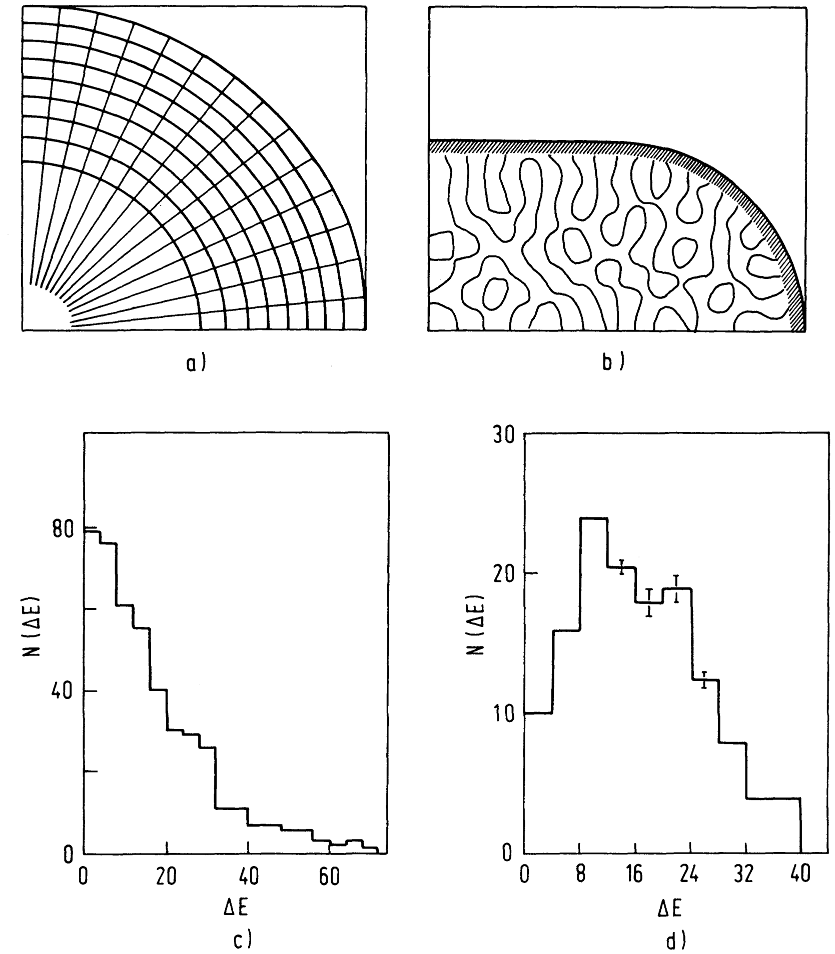

FIG. 33. Nodal curves [$\psi(x,y)=0$] for one quadrant of the (odd-odd parity) eigenfunctions (a) in a disk and (b) in a stadium (with dimensions $R=a$). Distribution $N(\Delta E)$ of (odd-odd parity) energy level spacings (c) for a circular boundary and (d) for a stadium boundary. (After McDonald and Kaufmann, 1979.) Note that $\Delta E=E_{j+1}-E_j$ is the spacing between neighboring levels; j increases with energy.

i.e., there is level repulsion. This level repulsion seems to be a characteristic feature of quantum systems whose classical limit shows chaos. It is related to the fact that no symmetries exist in these systems, i.e., there are no degeneracies (and no selection rules which prevent mutual interaction of the levels) such that $\lim_{\Delta E\to 0}(N(\Delta E))=0$. Several theoretical explanations for this phenomenon have been offered, and an interesting connection to random matrix theory (which is used to explain level repulsion in nuclear spectra) has been suggested (Berry, 1983).

The stationary Schrödinger equation in two dimensions is formally equivalent to the equation for the eigenfrequencies of a flat microwave resonator. This has recently been used by Stöckmann and Stein (1990) to confirm the computations shown in Fig. 33 by a microwave experiment.

7.3 Explicitly Time-Dependent Quantum Systems

Let us finally mention that the investigation of quantum chaos is not only of academic interest. Besides the stationary problems mentioned above, there exists also the large class of explicitly time-dependent systems such as atoms or molecules in a time-dependent radiation field. An understanding of, for example, the quantum ionization behavior of these systems which are again not integrable classically is of uttermost interest for laser chemistry.

8. OUTLOOK

Let us now take a glance at future possible developments by indicating several topics not dealt with in previous sections.

First of all, there is the problem of chaotic motion in spatially coupled nonlinear systems, such as coupled heart cells, chemical reactions (Vidal and Pacault, 1984) where the diffusion term is included [i.e., $\mathbf{c}=\mathbf{F}(\mathbf{c})+k\nabla^2\mathbf{c}$ instead of Eq. (1)], and the Navier–Stokes equations which could describe the motion of a turbulent fluid in time and space (Ruelle, 1983). Some pertinent questions are: How do these nonlinear oscillators influence each other? Do they synchronize? Does there exist something like spatial chaos? What is the influence of spatial motion on temporal chaos? What are the dimensions of strange attractors if one approaches fully developed turbulence?

Another major area is the question of the chaotic behavior of quantum systems with dissipation, such as lasers of Josephson junctions. It is also interesting to note that in quantum systems with many particles the question of chaos is related to the fundamental problem of the "arrow of time" (Misra and Prigogine, 1980).

This list is, of course, by no means complete. From the mathematical point of view, the very nature of a random number is still an issue of interest (De Long, 1970), and we have not discussed the possible role that the close coexistence of chaos and regular motion could play for the formation of structures in biology (Hess and Markus, 1984). It should also be noted that cellular automata (i.e., discrete approximations to partial differential equations in which all variables—time, space, and the signal—only take integer values) promise to become an important tool for answering many of the questions mentioned above (Wolfram, 1985).

The major conclusion, however, should be clear: Since nature is nonlinear, one has always to reckon with chaos.

GLOSSARY

Bifurcation: A dynamical system loses as a function of an external control parameter its stability and moves to one or several new stable states.

Critical Exponent: Near a transition from order to chaos tuned by a control parameter r, a measurable quantity x displays a power-law behavior $x \propto (r - r_c)^{\alpha}$ where α is the critical exponent and r_c is the transition point.

Ergodicity: The motion of a dynamical system in phase space is called ergodic if the trajectory covers a finite region of phase space densely. For ergodic systems time averages over dynamical variables can be replaced by phase-space averages.

Hopf Bifurcation: A fixed point becomes unstable in favor of a limit cycle if a control parameter is increased. The radius of the limit cycle increases smoothly from zero with the control parameter, whereas the frequency of the limit cycle jumps to a finite value at the transition point, i.e., each Hopf bifurcation introduces a new frequency into the system.

Navier–Stokes Equations: Spatio-temporal differential equations which determine the velocity profile of a liquid as a function of external pressure and boundary conditions.

List of Works Cited

Arnol'd, V. I. (1963), *Russ. Math. Surveys* **18**, 5–25.

Arnol'd, V. I. (1964), *Sov. Math. Dokl.* **5**, 581–593.

Balescu, R. (1975), *Equilibrium and Nonequilibrium Statistical Mechanics*, New York: Wiley.

Bergé, P., Dubois, M., Manneville, P., Pomeau, Y. (1980), *J. Phys. (Paris) Lett.* **41**, L344–L354.

Berry, M. V. (1978), *Topics in Nonlinear Dynamics*, Am. Inst. Phys. Conf. Proc. Vol. 46, New York: American Institute of Physics.

Berry, M. V. (1983), in R. H. G. Hellemann, G. Joos (Eds.), *Les Houches Summer School on Chaotic Behavior in Deterministic Systems*, Amsterdam: North-Holland.

Berry, M. V., Tabor, M. (1977), *Proc. Roy. Soc. A* **356**, 375–394.

Brandstäter, A., Swift, J., Swinney, H. L., Wolf, A., Farmer, D., Jen, E., Crutchfield, P. (1983), *Phys. Rev. Lett.* **51**, 1442–1445.

Coullet, P., Tresser, J. (1978), *J. Phys. (Paris), Colloq.* **39**, 5–17.

De Long, H. (1970), *Randomness and Gödel's Incompleteness Theorem*, Reading, MA: Addison Wesley.

Dubois, M., Bergé, P., Croquette, V. (1982), *J. Phys. (Paris) Lett.* **41**, L295–L301.

Einstein, A. (1917), *Verh. Dtsch. Phys. Ges.* **19**, 82.

Feigenbaum, M. (1978), *J. Stat. Phys.* **19**, 25–39.

Ford, J. (1983), *Phys. Today* **36**, 40–52.

Fraser, A. M., Swinney, H. L. (1986), *Phys. Rev. A* **33**, 1134–1140.

Grassberger, P., Procaccia, I. (1983), *Phys. Rev. Lett.* **50**, 346–349.

Grossman, S., Thomae, S. (1977), *Z. Naturforsch.* **32A**, 1553–1561.

Guckenheimer, J., Holmes, P. (1983), *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields,* Berlin: Springer.

Hess, B., Markus, M. (1984), *Synergetics–From Microscopic to Macroscopic Order,* Berlin: Springer.

Jeffries, C., Pérez, J. (1982), *Phys. Rev. A* **26**, 2117–2122.

Kolmogorov, A. N. (1954), *Dokl. Akad. Nauk USSR* **98**, 525–533.

Landau, L. D. (1944), *Dokl. Akad. Sci. USSR* **44**, 311–320.

Lanford, O. E. (1977), *Lecture Notes in Mathematics* **615**, Berlin: Springer.

Linsay, P. S. (1981), *Phys. Rev. Lett.* **47**, 1349–1352.

Lorenz, E. N. (1963), *J. Atmos. Sci.* **20**, 130–145.

Mandelbrot, B. B. (1982), *The Fractal Geometry of Nature,* New York: Freeman.

Manneville, P., Pomeau, Y. (1979), *Phys. Lett.* **75A**, 1–2.

McDonald, S. W., Kaufmann, A. N. (1979), *Phys. Rev. Lett.* **42**, 1189–1191.

Misra, B., Prigogine, I. (1980) *Suppl. Progr. Theor. Phys.* **69**, 101–114.

Moser, J. (1973), *Stable and Random Motion,* Princeton, New Jersey: Princeton Univ. Press.

Newhouse, S., Ruelle, D., Takens, F. (1978), *Comm. Math. Phys.* **64**, 35–46.

Poincaré, H. (1892), *Les Methodes Nouvelles de la Mèchanique Celeste,* Paris: Gauthier Villars.

Rössler, O. E. (1976), *Phys. Lett.* **57A**, 397–398.

Ruelle, D, Takens, F. (1971), *Comm. Math. Phys.* **20**, 167.

Ruelle, D. (1983), *Physica* **D7**, 40–44.

Saermark, K., Lebech, J., Bak, C. K., Sabers, A. (1988), *Springer Series in Brain Dynamics, Vol. 2,* Berlin: Springer.

Stöckmann, H. J., Stein, J. (1990), *Phys. Rev. Lett.* **64**, 2215–2218.

Swinney, H. L., Gollub, J. (1978), *Phys. Today* **31** (8), 41–49.

Takens, F. (1981), *Lecture Notes in Mathematics* **898**, Berlin: Springer.

Vidal, C., Pacault, A. (1984), *Nonlinear Phenomena in Chemical Dynamics,* Berlin: Springer.

Wolfram, S. (1985), *Phys. Rev. Lett.* **54**, 735–738.

Further Reading

Bergé, P., Pomeau, Y., Vidal, Ch. (1984), *L'Ordre dans le Chaos,* Paris: Hermann.

Chirikov, B. V. (1980), "A Universal Instability of Many Dimensional Oscillator Systems," *Phys. Rep.* **52**, 463–492.

Guckenheimer, J., Holmes, P. (1983), *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields,* Berlin: Springer.

Hellemann, R. H. G., Joos, G. (Eds.) (1983), *Les Houches Summer School on Chaotic Behavior in Deterministic Systems,* Amsterdam: North-Holland.

Lichtenberg, A. J., Liebermann, M. A. (1983), *Regular and Stochastic Motion,* Berlin: Springer.

Peitgen, H. O., Richter, P. H. (1986), *The Beauty of Fractals,* Berlin: Springer.

Schuster, H. G. (1989), *Deterministic Chaos,* 2nd ed., Weinheim: VCH.

CHARACTERIZATION AND ANALYSIS OF MATERIALS

JOHN P. SIBILIA, *Allied-Signal Inc., Morristown, New Jersey, U.S.A.*

Introduction 215
1. **Microscopy** 216
1.1 Optical Microscopy and Image Analysis 216
1.2 Scanning Electron Microscopy/Electron Probe Microanalysis 217
1.3 Transmission Electron Microscopy 218
2. **Spectroscopy** 219
2.1 Molecular Spectroscopy 219
2.1.1 Infrared Spectroscopy 219
2.1.2 Ultraviolet, Visible, and Near Infrared Spectroscopy 220
2.1.3 Raman Spectroscopy 221
2.1.4 Nuclear Magnetic Resonance Spectroscopy 221
2.1.5 Mass Spectrometry 222
2.2 Atomic Spectroscopy 223
2.2.1 Atomic Absorption Spectroscopy 223
2.2.2 Inductively Coupled Plasma Spectroscopy 225
2.2.3 Flame Emission and Emission Spectrographic Analysis 225
3. **Chromatography** 225
3.1 Gas Chromatography 226
3.2 Liquid Chromatography 226
3.3 Ion Chromatography 227
3.4 Gel Permeation Chromatography 227
4. **Thermal Analysis** 228
4.1 Thermogravimetry 228
4.2 Differential Scanning Calorimetry 229
4.3 Thermomechanical Analysis 230
4.4 Dynamic Mechanical Analysis 230
5. **X-Ray Analysis** 230
5.1 X-Ray Fluorescence Spectroscopy 231
5.2 X-Ray Diffraction Analysis 231
5.2.1 X-Ray Powder Diffraction 231
5.2.2 X-Ray Diffraction of Polymers 232
6. **Surface Analysis** 233
6.1 Scanning Auger Microscopy 234
6.2 X-Ray Photoelectron Spectroscopy 234
6.3 Secondary-Ion Mass Spectrometry 235
6.4 Rutherford Backscattering 236
6.5 Electron Energy-Loss Spectroscopy 237
6.6 Scanning Tunneling and Atomic Force Microscopy 237
6.7 Low-Energy Electron Diffraction and Reflection High-Energy Electron Diffraction 238
Glossary 238
Acronyms and Initializations 239
Further Reading 239

INTRODUCTION

The science of materials analysis and characterization is an integral part of the procedure of developing new materials. More accurate, more precise, and faster response instruments and methodologies are always being sought to aid in the understanding of material behavior and in the development of new materials themselves.

In recent years advances in computer hardware and software, electronic circuitry, sensing devices, vacuum systems, and many other technologies have manifested themselves in the development of analytical instrumentation beyond the wildest dreams of scientists

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

only one generation ago. Again, in recent years, significant advances have been made in materials development. Novel polymers with enhanced mechanical, chemical, electrical, high-temperature, and barrier properties have been developed. Light, composite plastics have found their way into the aerospace, automotive, and housing-construction industries. Great strides have also been made in the development of metals and metal alloys, with improved mechanical, magnetic, fatigue and corrosion resistance properties. New ceramic materials with excellent high temperature properties and improved strength and toughness are also beginning to appear in commerce. Dramatic breakthroughs in new superconducting materials have also been made and the biosciences have made dramatic strides in the last two decades.

In this article a variety of instruments, techniques, and methodologies will be described which have played a major role in the development and commercialization of new polymers, biomaterials, metals, and ceramics. Each methodology will be described in order to acquaint the reader with the principles of the technique, sample requirements, and how it is applied to characterization and analysis. Where possible, specific examples and in some cases the limitations of the techniques will be discussed. Since the article is limited in its scope, it will not be possible to go into extensive depth in any particular technique. However, references will be cited where the reader will be able to follow up in more detail on any of the subjects described.

The sections are written in such a manner as to familiarize a novice with the first principles of each technique and how that technique is applied to materials analysis and characterization. The more experienced investigator can use the article as a quick review with an opportunity to examine the techniques in more depth through utilization of the numerous reference citations.

Anyone who practices in the field of characterization and analysis understands that most problems involve a multidisciplinary approach in order to provide the most efficient solutions. Therefore, while each technique will be described in a "stand alone" manner, in reality a combination of many techniques is used in materials characterization and analyses.

1. MICROSCOPY

Optical microscopy is one of the oldest characterization tools known. In many cases it is one of the first examinations made on a material, since it is an extension of what the human eye itself sees. However, the resolution in optical microscopy is limited by the wavelength of light. Greater magnification can only be achieved by electron microscopy and newer techniques such as scanning tunneling and atomic force microscopy. The latter two techniques, as well as other "surface microscopic" techniques, will be covered in Sec. 6.

1.1 Optical Microscopy and Image Analysis

The optical microscope obtains information about a material either through transmitted or reflected light (see OPTICAL MICROSCOPY). Magnification from 10× to 1400× can be achieved with modern optical microscopes. The limit of resolution is about 0.2 μm or the order of the wavelength of visible light. Cameras can be adapted to light microscopes which will allow for permanent storage of images; or, in the case of video cameras, real-life observations can be made by groups of people at the same time. Cameras are usually equipped with polarizing devices and bright-field, dark-field, and temperature control units which allow observations on specimens from -20 to 350 °C. Temperature studies are extremely useful in determining phase transitions in materials, the most common of which is the melting point.

Transmission optical microscopy is used to examine thin (of the order of microns) specimens which can be prepared in a variety of ways. Polymer films or fibers, powders, biological or thin petrographic sections, dispersions, and emulsions are all routinely examined. Reflected-light microscopy is used primarily on flat specimens. Polished and etched metallurgical or ceramic samples are studied for the nature and variability of grain sizes in the specimens.

All microscopic techniques lend themselves to image analysis. A real microscopic image or a photograph can be scanned with a video camera. The camera scans the image and generates an analog signal. The variation in the signal voltage corresponds to intensity variations in the photograph or sample being

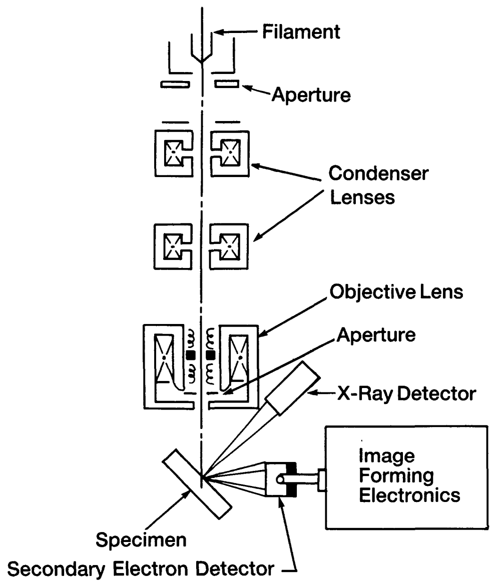

FIG. 1. Schematic drawing of a combined scanning electron microscope–microprobe analyzer.

scanned. The data can then be used to calculate particle sizes and shapes or any structural elements in the material which exhibit intensity variability.

1.2 Scanning Electron Microscopy/Electron Probe Microanalysis

All solid materials can be studied by scanning electron microscopy (SEM). Sample sizes are normally limited to < 10 cm. Information about surfaces or the surface topography is usually obtained by SEM techniques since the depth of field is far greater than that achieved in either optical or transmission electron microscopy (TEM). Resolution is about 3 nm, or about two orders of magnitude greater than with the optical microscope and one order of magnitude less than with TEM (see ELECTRON MICROSCOPY.)

In Fig. 1 is given a schematic diagram of an SEM with an electron probe microanalysis (EPMA) attachment. The EPMA attachment is indicated as the x-ray detector. The higher resolution of an SEM as compared to an optical microscope is achieved by use of an electron beam, generated by a heated filament and focused in a vacuum chamber by electromagnetic lenses. The beam is rastered over the specimen in synchronism with the beam of a cathode-ray-tube display screen. The brightness of the cathode-ray tube is modulated by the signal from a scintillator which in turn has collected electrons that have been scattered from the sample surface. The result is an image on the cathode-ray tube which represents the surface topography of the specimen.

In Fig. 2 is an SEM image of the fracture face of an optical fiber. Note the small (~1 μm) impurity inclusion which is responsible for the failure of the fiber.

When materials are bombarded by high-energy electrons (10 keV–50 eV) such as those generated in SEM, x-ray fluorescence radiation is produced by the elements in the specimen being bombarded. The elements are subsequently identified and their concentrations measured by distinguishing characteristic x-ray wavelengths. Detectors equipped with analyzing crystals can be used to detect the characteristic wavelengths based on the use of

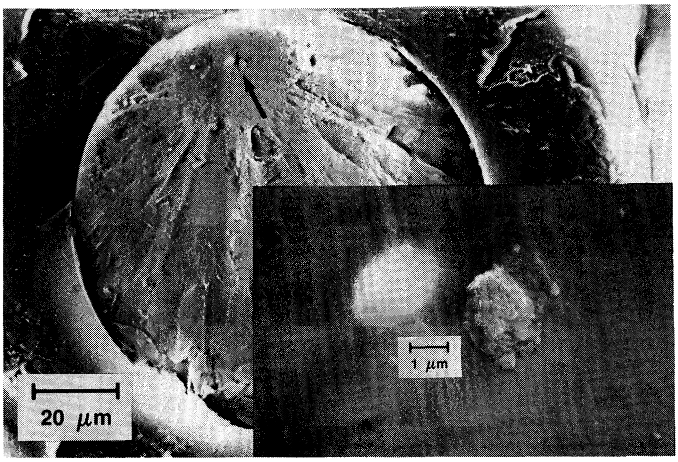

FIG. 2. Optical fiber fracture face showing inclusions responsible for failure.

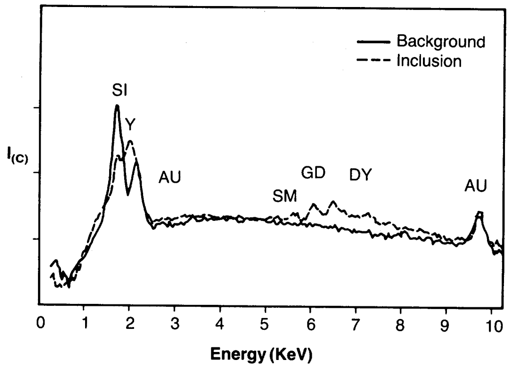

FIG. 3. Energy-dispersive x-ray spectra indicating that the inclusions shown in Fig. 2 consist of rare earth elements, and are probably in the form of complex oxides.

Bragg's law. Energy-dispersive detectors can also be used, giving a signal amplitude which is proportional to the quantum energy of the radiation and can be separated out by a suitable electronic gate. When an energy-dispersive detector is used, the intensity of the x-ray signal is plotted as a function of the energy. In Fig. 3 are EPMA energy-dispersive x-ray spectra of the elemental composition of the optical fiber (shown in Fig. 2) background and the inclusion. Note that the inclusion contains the elements Y, Sm, Gd, and Dy and is most likely a mixed oxide of the compounds.

The SEM and EPMA techniques are applied to a wide variety of materials. For example, they have been used to study the distribution of glass fibers in glass-polymer composites, the microstructure of ceramics, the nature of small spot impurities in electronic materials, the nature of fracture surfaces in composites, and the magnetic domain structures in amorphous metals.

1.3 Transmission Electron Microscopy

High resolution transmission electron microscopy (TEM) instruments have resolutions finer than 0.2 nm. Sample specimens are usually <3 mm in diameter and <0.2 μm in thickness. In TEM a collimated beam of high-energy electrons (100–400 keV) in a vacuum chamber is allowed to pass through a thin specimen. The transmitted beam can be used to project an image of the specimen on a viewing screen or photographic plate. If the transmitted beam is allowed to raster and a scintillation detection system is used then a scanning transmission electron microscope (STEM) image is obtained. When the high-energy electron beam impinges on the specimen, some of the electrons are diffracted and can be imaged on a fluorescent screen. In this way information about the crystal and compositional structure of material can also be obtained.

Bright-field and dark-field TEM images are used to study the microstructures of ceramics, metal alloys and semiconductors. Diffracted electrons (see ELECTRON DIFFRACTION) can be imaged to create an electron diffraction pattern of the material. Such a pattern is characteristic of the crystal structure at the microscopic level. Information about the orientation of microstructural features can also be obtained from electron diffraction patterns. In Fig. 4 is shown a TEM micrograph and corresponding electron-diffraction pattern of a device used as a hetero interface in the semiconductor industry. It is produced by epitaxially depositing GaAlAs layers on a GaAs substrate. Note how TEM can provide a clear measure of thickness and thickness variation of the layers. Figure 5 shows the TEM micrograph of a

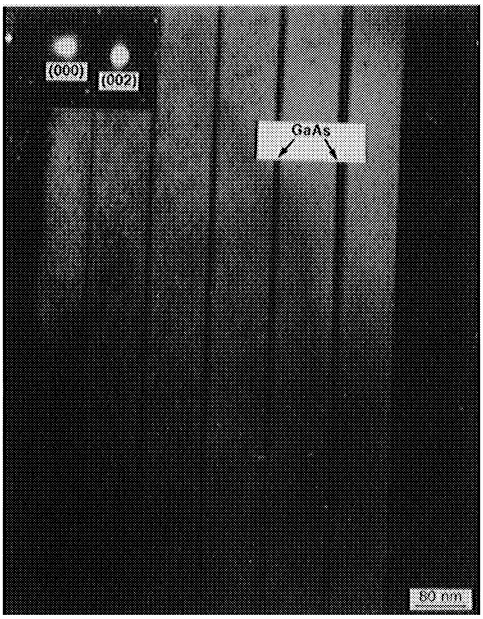

FIG. 4. Dark-field transmission electron micrograph of alternating layers of GaAs (dark)/GaAlAs (light) epilayers deposited by molecular beam epitaxy onto a GaAs substrate. Inset shows the corresponding electron diffraction pattern.

crystallized microstructure of a metal alloy. It consists of Mo/Fe/B precipitates in a Ni-rich matrix.

In general, TEM is used to study the morphology and microstructure of materials. It is used to obtain information about the particle size and particle size distribution of powders, the dispersion of one polymer phase in another, the nature of grain boundaries in metals, metal alloys, and ceramics, and the relative orientations of microstructural features.

2. SPECTROSCOPY

Some spectroscopies, similar to optical microscopy, represent some of the earliest techniques in the characterization and analysis of materials. Spectroscopic techniques can be divided into two categories. The molecular spectroscopies are those such as infrared, ultraviolet, visible, Raman, microwave, electron paramagnetic, and nuclear magnetic resonance spectroscopies which provide information about the structure of molecules. All of these spectroscopies are based on the interaction of electromagnetic radiation with molecules. The way in which the incident radiation is modified by a molecule gives specific information about the structure of the molecule (see MOLECULAR SPECTROSCOPY).

The second type of spectroscopy, atomic spectroscopy (*q.v.*), involves the breakdown of molecules in a material into their individual atoms with subsequent qualitative and quantitative analyses of the elemental contents of the material. These spectroscopies include atomic absorption (AA), emission spectrographic, flame emission, and inductively coupled plasma emission spectroscopies.

Mass spectrometry (see SPECTROMETERS, MASS) does not fall neatly into either of these categories. It can be used as both an elemental analytical technique and a molecular analytical technique. In most cases mass spectrometric analyses involves a partial breakdown of molecules through electron or ion impact and an analysis of the resulting mass fragments. A spark source can also be used on metals to provide elemental analysis. For organic materials analysis, information from mass-spectrometric techniques is similar to that obtained from the molecular spectroscopies since it provides knowledge about the molecular structure of materials. In the current work, mass spectrometry is included in the section on molecular spectroscopy since a wider number of applications appear in this context.

2.1 Molecular Spectroscopy

2.1.1 Infrared Spectroscopy Infrared (IR) spectroscopy is based on the fact that atoms in molecules vibrate at specific frequencies which are characteristic of the atoms involved, the nature of their bonding, and the nature of the molecule in general. Most of these frequencies occur in the 200- to 4000-cm^{-1} region of the spectrum, i.e., the infrared region. When infrared radiation is passed through an organic or inorganic substance, those specific incident frequencies that match the frequencies of the substance are absorbed. Therefore, if infrared frequencies are allowed to be passed through material and the resultant frequencies are scanned and detected individually, a plot of absorption intensity vs infrared frequency can be obtained. This plot is unique to a given material and can be used as a "fingerprint" of the material for identification purposes. Characteristic frequencies,

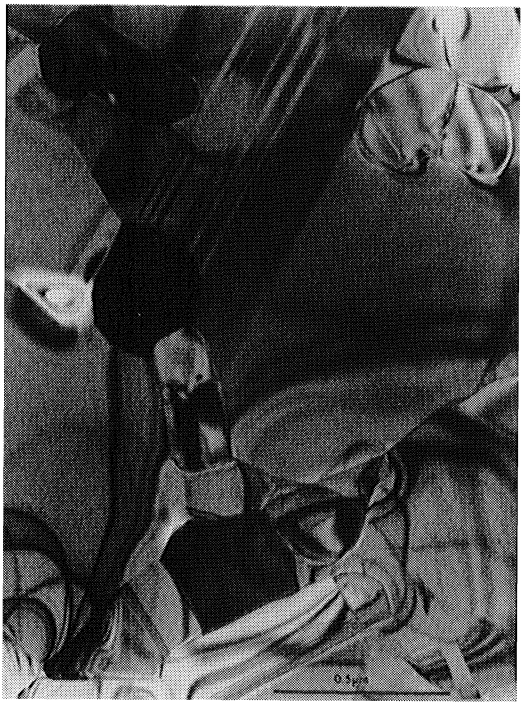

FIG. 5. Alloy 5006 microstructure showing Mo/Fe/B precipitates in a Ni-rich matrix.

usually expressed as "wave numbers" (cm^{-1}), can also give detailed information about the structural nature of the material. Quantitative measurements of the intensity can serve to give a measurement of the amount of a chemical species or amount of a compound present in a mixture. Samples are usually analyzed in the milligram range but samples as small as 50 pg can be analyzed with special techniques. Materials in the solid, liquid, or gaseous state can be analyzed. Infrared spectroscopy is used routinely to perform both quantitative and quantitative analyses on organics, inorganics, polymers, and biomaterials.

In polymers, functional groups, conformations, orientation, molecular weight, degradation, impurities, and many other characteristic features can be studied. Infrared spectroscopy has also been used to study the reaction kinetics of organic molecules, configurational structures, the nature of hydrogen bonding, and the nature of hydration in inorganics.

In Fig. 6 is given a series of spectra which illustrate an application of infrared spectroscopy to study the nature of thermal degradation in an organic fluoropolymer. When the material is heated in a vacuum at 300 °C, an absorption peak occurs which is indicative of cracking of the polymer chain with subsequent formation of a perfluorovinyl end group.The changes are quite different when the material is heated in the presence of oxygen as indicated in the bottom curve in Fig. 6. This type of information is useful in determining what happens to polymeric materials when they are subjected to degrading temperatures.

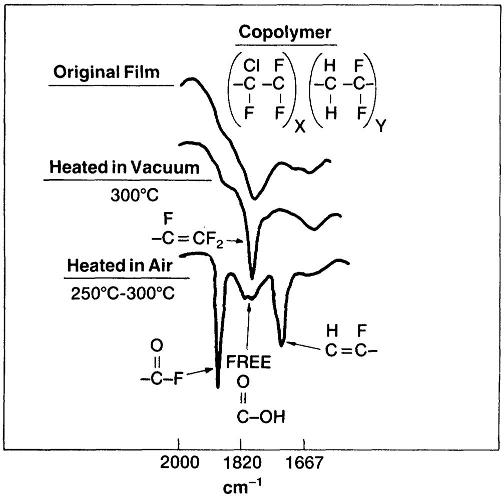

FIG. 6. Infrared spectra of thermally degraded halocarbon film.

2.1.2 Ultraviolet, Visible, and Near Infrared Spectroscopy. Spectroscopy in the broad region of the electromagnetic spectrum from about 2500 nm to <200 nm, which includes ultraviolet (UV), visible, and near-infrared (NIR) radiation, is used for both qualitative and quantitative analyses of organic, inorganic, and polymeric materials. Materials in the solid, liquid, or gaseous state can be analyzed. In some cases nanogram quantities of materials can be analyzed.

The UV region ($\sim$400 to 180 nm) and the visible region ($\sim$750 to 400 nm) provide information about the electronic transitions which occur in molecules. In particular, the techniques give information about the nature of conjugated double bonds or chromophoric groups in molecules. Generally speaking, the larger the region over which π-electron conjugation can occur, the longer the wavelength of the absorption and the more intense becomes the absorption. For example, simple alkenes absorb at about 200 nm while large dye molecules absorb very strongly in the visible region. Therefore UV–visible spectroscopy is particularly suited for analyzing aromatics and dyes at the ppm level. However, since the absorptions in the UV–visible region tend to be very broad, structural information is sometimes difficult to obtain.

The NIR region of the spectrum, like the infrared region, provides information about the molecular structure of materials based on an analysis of vibrational modes in a molecule. Absorption bands in the NIR usually arise from combinations and overtone bands which arise from fundamental vibrations in the infrared region of the spectrum. Therefore, the NIR absorptions are relatively weak compared to the IR regions. However, the region can be extremely useful for analyzing compounds containing hydrogen, since vibrations involving hydrogen atoms occur at high vibrational frequencies and extend into the near-infrared region of the electromagnetic spectrum. For example, the NIR region can be used to obtain information about bonded and unbonded -NH and -OH in compounds such as amines, amides, oximes, alcohols, and lactams.

2.1.3 Raman Spectroscopy. Raman spectroscopy, like infrared spectroscopy, can detect vibrational frequencies in a molecule. However, while infrared spectroscopy is an absorption phenomenon, Raman spectroscopy is a scattered-light phenomenon.

When an intense beam of light enters a sample, scattering can occur in all directions, with the frequency of the scattered light being the same as that of the incident light (ν_0). This effect is known as Rayleigh scattering. Another type of scattering that can occur simultaneously with the Rayleigh scattering is Raman scattering. It occurs at frequencies higher and lower than ν_0 and at much lower intensities. The differences $\Delta\nu$ between the incident and scattered frequencies are equal to the vibrational frequencies of the material. Therefore, Raman spectroscopy, like infrared provides information about the various functional groups in a molecule. However, since the selection rules governing the allowable vibrational transitions are different for the two techniques, some frequencies may be observed in the Raman spectrum which do not appear in the infrared and vice versa.

Raman spectra are usually obtained on solid or liquid materials. One advantage of Raman spectroscopy over infrared is that the spectra of materials in aqueous solutions, such as biological, organic, and inorganic compounds can be readily obtained. It is difficult to obtain usable spectra in the infrared since water absorbs too strongly, thus obscuring the spectra of dissolved species.

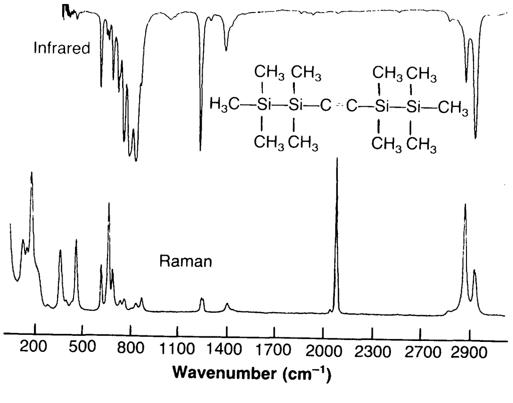

FIG. 7. Infrared and Raman spectra of bis(pentamethyldisilanyl) acetylene.

In Fig. 7 are given the infrared and Raman spectra of an organic silicone compound. The very intense infrared band at 1246 cm^{-1} is indicative of the CH_3 group on a Si atom, while the very intense band at 2092 cm^{-1} in the Raman spectra is indicative of the $-C{\equiv}C-$ group in the molecule. Note that this frequency is essentially nonexistent in the IR spectrum.

2.1.4 Nuclear Magnetic Resonance Spectroscopy. Nuclear magnetic resonance spectroscopy (NMR) (see MAGNETIC RESONANCE, NUCLEAR) is a much newer molecular structure technique than either infrared or Raman spectroscopy. However, in a relatively short time it has proven itself to be a major tool particularly in the structural analysis of organic, polymeric, and biological materials. In high-resolution NMR, structural analysis is accomplished on materials dissolved in suitable solvents. Concentrations greater than 1% are normally used.

When a material containing a nucleus with a magnetic moment is placed in a strong external magnetic field (H_0), the magnetic moment of the nucleus will begin precessing about H_0 with a frequency $\nu=\gamma H_{\text{eff}}/2\pi$, where γ is the gyromagnetic ratio of the nucleus and H_{eff} is the effective magnetic field which the nucleus "sees." When a strong radio-frequency pulse at a frequency close to ν and perpendicular to H_0 is applied, the angle of the magnetic moment with respect to H_0 will change, inducing a resonance signal which can be detected. Since H_{eff} is sensitive to the immediate environment of the nucleus in a molecule, nuclei in different environments will resonate at different frequencies. The difference between the frequency of a particular resonance peak and the frequency of a reference peak when divided by the frequency corresponding to the external magnetic field for the particular nucleus under study is called the "chemical shift." It is expressed in parts per million (ppm). Therefore, a plot of intensity of the signal as a function of chemical shift in ppm will give characteristic information about the functional groups and overall structure of a molecule.

Nuclei commonly studied in NMR are ^1H, ^{13}C, ^{19}F, ^{29}Si, and ^{31}P, although any nucleus with a magnetic moment can give rise to an

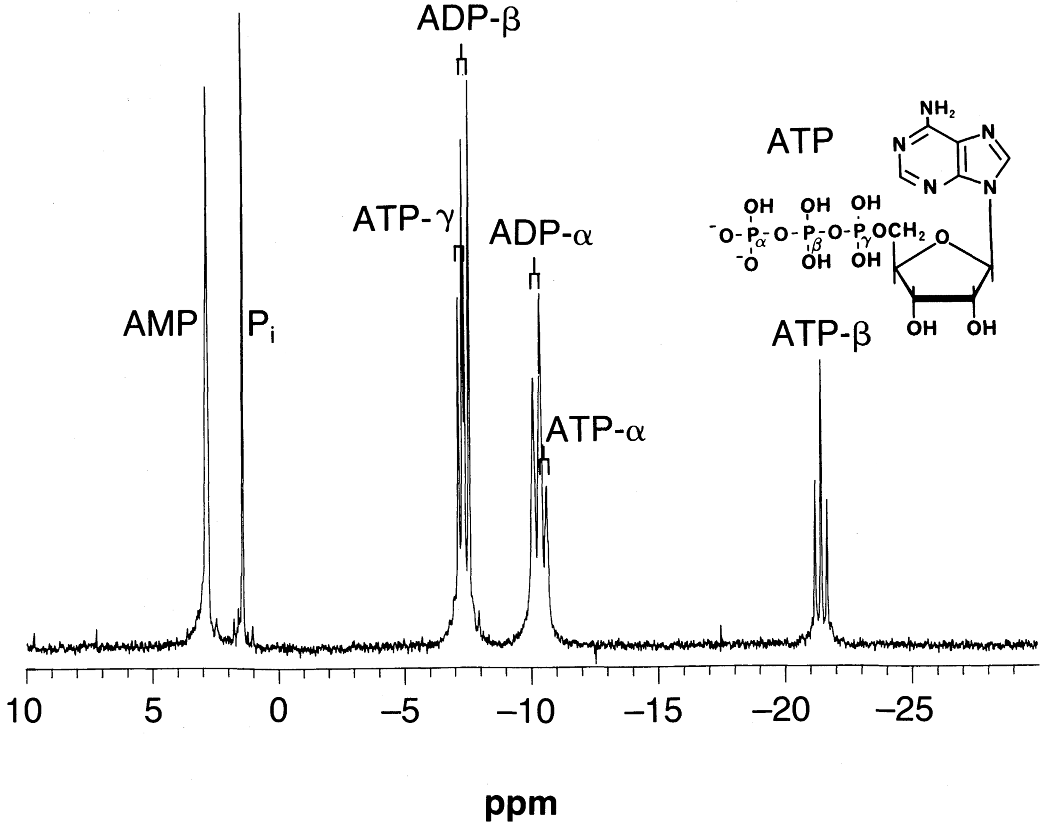

FIG. 8. ^{31}P, 162 MHz nuclear magnetic resonance spectrum of impure ATP.

NMR effect. In Fig. 8 is reproduced a ^{31}P NMR spectrum (ppm vs intensity) of adenosine triphosphate (ATP), an important component of living systems. Adenosine diphosphate (ADP), adenosine monophosphate (AMP), and inorganic phosphate are also detected in the particular sample studied. Resonance signals from the various types of phosphorus atoms are indicated. It has been more difficult to obtain detailed structural information on solids by NMR techniques. The basic principles which apply to solution NMR are applicable to solid-state NMR. However, while molecules are free to tumble and rotate in solution, which gives rise to very narrow lines, their restricted motion in solids can result in very broad noninformative spectra. The nuclei in a solid "see" the effects of the different local magnetic fields of adjacent nuclei. This causes broadening of the spectra. However, when the solid sample is rotated at very high speeds (>200 000 rpm) and at an angle of 54° 44′ (the magic angle), the solid-state interaction effects can be largely reduced and relatively narrower line spectra can be obtained. This technique, together with improvement in instrument performance, has enabled solid-state NMR to yield a wealth of new information.

Solid-state NMR has been successfully applied to the study of polymers, polymer blends, ceramics, and inorganic zeolites. In Fig. 9 is a solid-state ^{29}Si NMR spectrum of a zeolite. It clearly shows that the silicon atoms in the SiO_4 tetrahedra have different environments, i.e., different associations with the aluminum (Al) atoms in the zeolite.

2.1.5 Mass Spectrometry. Mass spectrometry is a structural technique which can be applied to the analysis of organic, organometallic, inorganic, and ionic compounds as well as polymers, ceramics, metals, and alloys. It is possible to obtain information on materials with molecular weights as high as

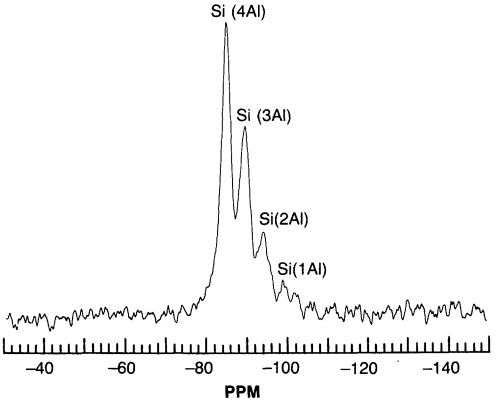

FIG. 9. ^{29}Si, 39.7 MHz solid-state nuclear magnetic resonance spectra of a zeolite.

10 000 g mol^{-1}. Liquid, gaseous, and solid samples can be analyzed.

The common mode of operation of a mass spectrometer is to produce gas-phase ions-from a sample which can then be separated according to their mass-to-charge ratio. Separations are made under high vacuum by means of electrical and magnetic fields. Gas-phase ions can be produced by using electron impact, chemical ionization, and fast-atom bombardment techniques. Other excitation techniques such as plasma discharge ionization and spark excitation are used for metals and ceramics analyses.

The most common applications of mass spectrometry are to determine elemental metal contents, to identify the structures of new materials, and to confirm the structures of known substances. Two examples are given. In Table 1 is given a computer printout of a high- resolution mass spectrum with mass measurements accurate to 0.1 millimass unit. One peak of the unknown compound had a measured mass of 251.1543 which is consistent with the empirical formula $C_{14}H_{21}NO_3$. These data along with other supporting data confirmed the organic structure as indicated. Another example involves a polymeric structure. An important piece of information in understanding the properties of polymeric materials is to know the molecular weight and molecular-weight distribution of the material. The flow characteristics, viscosity, solution properties, etc., depend on the molecular weight and molecular-weight distribution. In Fig. 10 is given the fast-atom-bombardment mass spectrum (FAB-MS) plotted as the intensity of detector response against the mass-to-charge ratio (m/z), for a polyether. In the FAB-MS technique, the surface of a liquid solution which contains the analyte is bombarded by a beam of energetic neutral atoms such as xenon. Secondary ions typical of the solvent and analyte are sputtered from the liquid surface and analyzed by the mass spectrometer. This technique is extremely useful for analyzing nonvolatile, thermally labile, and high molecular-weight compounds. From Fig. 10, both the average molecular weight and the distribution of molecular weights of the polyether can clearly be ascertained.

In many cases mass spectrometry is coupled with separation techniques such as gas or liquid chromatography. These combination techniques, GC-MS and LC-MS, have been demonstrated to be some of the most powerful methods for analysis, materials identification, and structure elucidation.

2.2 Atomic Spectroscopy

2.2.1 Atomic Absorption Spectroscopy. Atomic absorption (AA) spectroscopy is a technique for determining the concentration of metals in solution. It allows for analysis of only one element at a time. A solution of the metal to be quantitatively analyzed is first vaporized into a flame. An emission source containing the element to be analyzed emits radiation characteristic of the element which passes through the vaporized analyte element in the flame. Absorption of the specific radiation occurs which is related to the concentration of the analyte which has been vaporized into the flame. Vapor may also be formed by converting the metal to a volatile hydride as in the case of As or by cold vaporization as in the case of Hg. If a graphite furnace is used as the vaporization source, the method is known as flameless AA and the analysis is two orders of

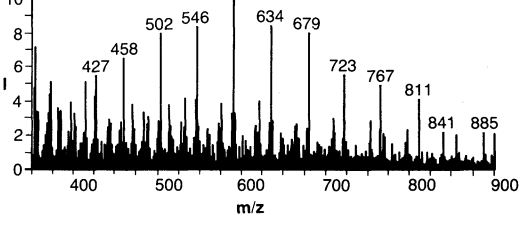

FIG. 10. Molecular-weight distribution of a commercial polyether sample by positive-ion fast-atom bombardment mass spectrometry.

Table 1. High-resolution mass spectrum computer printout.

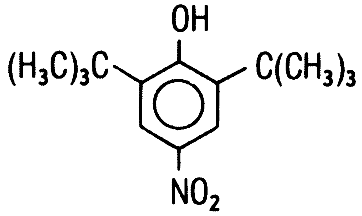

```
ELEMENTAL COMPOSITION           DATA: 00515F #  25          BASE M/E:    55
05/19/80 13: 52: 00 +  6: 40    CALI: 00515F #  25          RIC:      36800.
SAMPLE: CW YELLOW 00515
- (/S)
MIN INTEN:      72.             DEF: 30 MMU/100 AMU

MIN     53       0. 00        0.                  2. 0
MAX    393                  #  0   30. 0   100 20. 0  C200. H400. N4. O5
      MASS        % RA    INTEN.     MMU   PPM BONDS  FORMULA

 208. 1055       6. 53      486.  -19. 6    95  9. 0  C16. H16
                                   -7. 0    34  9. 5  C15. H14. N
                                    5. 4    26 10. 0  C14. H12. N2
                                   18. 0    87 10. 5  C13. H10. N3
                                   16. 7    80 10. 0  C15. H12. O
                                   29. 3   141 10. 5  C14. H10. N. O
                                  -26. 8   129  5. 0  C10. H16. N4. O
                                  -28. 1   136  4. 5  C12. H18. N. O2
                                  -15. 6    75  5. 0  C11. H16. N2. O2
                                   -3. 0    15  5. 5  C10. H14. N3. O2
                                    9. 5    46  6. 0  C9. H12. N4. O2
                                   -4. 3    21  5. 0  C12. H16. O3
 236. 1283      19. 66     1464.  -28. 1   119  9. 0  C18. H20
                                  -15. 6    66  9. 5  C17. H18. N
                                   -3. 0    13 10. 0  C16. H16. N2
                                    9. 5    40 10. 5  C15. H14. N3
                                   22. 1    94 11. 0  C14. H12. N4
                                    8. 1    35 10. 0  C17. H16. O
                                   20. 7    88 10. 5  C16. H14. N. O
                                  -24. 1   102  5. 0  C13. H20. N2. O2
                                  -11. 5    49  5. 5  C12. H18. N3. O2
                                    0. 9     4  6. 0  C11. H16. N4. O2
                                  -12. 9    55  5. 0  C14. H20. O3
                                   -0. 3     1  5. 5  C13. H18. N. O3
                                   12. 2    52  6. 0  C12. H16. N2. O3
                                   24. 8   105  6. 5  C11. H14. N3. O3
                                   23. 4    99  6. 0  C13. H16. O4
                                   16. 2    69  2. 0  C7. H16. N4. O5
 251. 1543       1. 05       78.  -25. 6   102  8. 5  C19. H23
                                  -13. 0    52  9. 0  C18. H21. N
                                   -0. 5     2  9. 5  C17. H19. N2
                                   12. 0    48 10. 0  C16. H17. N3
                                   24. 6    98 10. 5  C15. H15. N4
                                   10. 7    43  9. 5  C18. H19. O
                                   23. 3    93 10. 0  C17. H17. N. O
                                  -21. 6    86  4. 5  C14. H23. N2. O2
                                   -9. 0    36  5. 0  C13. H21. N3. O2
                                    3. 5    14  5. 5  C12. H19. N4. O2
                                  -10. 3    41  4. 5  C15. H23. O3
                                    2. 1     9  5. 0  C14. H21. N. O3
                                   14. 7    59  5. 5  C13. H19. N2. O3
                                   27. 3   109  6. 0  C12. H17. N3. O3
                                   26. 0   103  5. 5  C14. H19. O4
```

magnitude more sensitive than when a flame is used. In Table 2 is given a comparison of the sensitivities of both flame and flameless AA techniques. These analyses are applied widely to metals, metal alloys, and trace metals in polymers and ceramic materials.

2.2.2 Inductively Coupled Plasma Spectroscopy. Inductively coupled plasma (ICP) spectroscopy, unlike AA, is an emission technique. Its chief advantage over AA is its ability to analyze many elements at one time. In ICP a solution of the analyte is atomized into a stream of a gas such as argon which passes through a radio frequency generator to create an argon plasma and excited atoms of the analyte. The excited atoms emit characteristic radiation specific to their atomic structures which in turn can be detected by photomultiplier tubes.

The limits of detection in ICP when photomultiplier tubes are used are similar to AA except for the refractory elements (e.g., boron, zirconium, and phosphorus) which are more readily determined by means of the hotter plasma source. When ICP is coupled with a mass spectrometer detection system (ICP-MS), detectabilities can be improved by several orders of magnitude. Again like AA, the method is used for trace metal analysis in polymers, metal matrixes, ceramics, and chemicals, and for determining composition of metal alloys.

2.2.3 Flame Emission and Emission Spectrographic Analysis. Flame emission spectroscopy involves excitation of atoms in a flame with subsequent emission of radiant energy specific to the elements in the flame. In emission spectrographic analyses excitation occurs through an electrical arc or a spark to produce emission from specific elements exposed to the arc or spark. In both cases the emitted frequencies from all the elements exposed are passed through a monochromator and then into a detection system which can be either photographic or electronic. Both methods allow for analysis of many elements simultaneously. However, problems in analysis can arise from overlapping frequencies when many elements are present. Arc or spark emission methods are less sensitive than AA techniques. The flame emission technique is particularly suited to the analysis of alkali and alkaline earth metals. In Table 3 are listed the detection limits for the flame method for alkali and alkaline earth metals.

Emission methods are useful for analysis of metals, elements in metal alloys, polymers, ceramics, clinical materials, and soils. In Table 4 are given the detection limits for metals by emission spectrography.

3. CHROMATOGRAPHY

Chromatography (*q.v.*) is an important part of the technology of the analysis and characterization of materials. It is the science of separating components in a mixture, identifying the components, and quantitatively measuring the amount of each component. Chromatographies are dynamic in nature. They consist of a flow of liquid or gas through a medium which has the ability to separate components and measure their concentrations with a suitable detector. Chromatography is a broad field and overlaps the field of separation sciences. However, only four types of chromatography will be covered in this section.

Table 2. Atomic absorption detection limits.

	Flame (mg/L)	Flameless (mg/L)
Aluminum	0.1	0.003
Antimony	0.2	0.003
Arsenic	0.002[a]	0.001
Barium	0.1	0.002
Beryllium	0.005	0.0002
Cadmium	0.005	0.0001
Chromium	0.05	0.001
Copper	0.02	0.001
Gold	0.1	0.001
Iron	0.03	0.001
Lead	0.1	0.001
Manganese	0.01	0.0002
Mercury	···	0.0002[b]
Nickel	0.04	0.001
Potassium	0.01	···
Silicon	0.02	0.001
Silver	0.01	0.0002
Sodium	0.002	···
Tin	0.8	0.005
Vanadium	0.2	0.004
Zinc	0.005	0.000 05

[a] By gaseous hydride method.
[b] By cold vapor technique.

3.1 Gas Chromatography

Gas chromatography (GC) is used to analyze gases or those liquids and solids which can be heated to a volatile state without decomposition occurring.

Samples as small as one microliter can be analyzed at detection levels, in some cases as low as parts per billion.

In GC, a sample is vaporized into a stream of inert gas such as argon, nitrogen, or helium, and allowed to pass through a column containing a separation medium. As components in the gas are preferentially absorbed and desorbed from the separation medium or "packing," less tightly bound molecules will elute first from the column followed by those components which were more tightly bound. A variety of detectors such as thermal conductivity, flame ionization, and electron capture are commonly used to monitor the compounds as they elute from the column. The nature of the separation medium, length of the column, type of detector, flow rate of the gas, and column temperature can all be adjusted to effect the desired separations and level of detection of the components. In many cases, where a material is insufficiently volatile to be transported through a column, the material can be "derivatized" to form a more volatile species. In Fig. 11 is given a gas chromatogram of a mixture of amino acids which have been derivatized to give volatile components.

Gas chromatography is a major analytical tool in organic and bio-organic analyses, especially when used in conjunction with mass spectrometry. It is a key tool in the analysis of pollutants in our air, water, and soil. Gas chromatography is also relied upon heavily in determining impurities in plastics, especially when such impurities might migrate from food packaging plastics.

Table 3. Detection limits of alkali and alkaline earth metals by flame emission spectrometry.

Element	Limit (ppm)
Barium	0.03
Calcium	0.005
Cesium	0.005
Gallium	0.01
Indium	0.03
Lanthanum	0.1
Lithium	0.000 03
Potassium	0.001
Rubidium	0.002
Sodium	0.0001
Strontium	0.004

Table 4. Emission spectrographic detection limits.

Element	Limit (μg)	Element	Limit (μg)
Aluminum	0.05	Manganese	0.05
Barium	0.5	Molybdenum	0.05
Beryllium	0.002	Nickel	0.1
Boron	0.1	Palladium	0.02
Calcium	0.01	Silicon	0.2
Chromium	0.05	Silver	0.01
Cobalt	0.05	Tin	0.1
Copper	0.02	Titanium	0.02
Gold	0.2	Vanadium	0.05
Iron	0.1	Zirconium	0.05
Lead	0.2		

3.2 Liquid Chromatography

Liquid chromatography (LC) is particularly useful for analysis of liquids and solutions of solids when such materials cannot be analyzed by gas chromatography because of their thermal instability. Like gas chromatography, it can be coupled with mass spectrometry. Sensitivities are in the ppm range, sometimes extending to the ppb range. The carrier phase in LC is a liquid which is forced under pressure through a suitably packed column. Early column materials were silica and alumina. Other columns consist of a support material to which is bound a material in which the solute components have some degree of solubility. Separation is effected by a partitioning of solute molecules between the liquid mobile phase and the solid stationary phase. Detection systems are more limited than those available in gas chromatography. Ultraviolet absorption and index of refraction are two commonly used detection systems.

Liquid chromatography can be used for analysis of over 80% of known organic compounds, whereas gas chromatography is limited to analysis of only about 20%. Therefore, although gas chromatography is a very rapid and reliable analytical technique, it is limited in scope because of the requirement that samples be volatile. On the other hand, liquid chromatography is much more flexible and can lend itself to analysis of such nonvolatile species as macromolecules, ionic species, and biological and natural product materials.

3.3 Ion Chromatography

Ion chromatography (IC) is used for qualitative and quantitative analyses of ionic species in complex mixtures. Liquids and solids can be analyzed after dissolution in deionized water. Gases which produce ionic species in water can also be analyzed. Under some conditions ionic species can be detected and measured at the parts per billion level.

Ion exchange columns are used to effect the chromatographic separations with sodium bicarbonate/sodium carbonate as an eluent for anions and dilute nitric acid as an eluent for cations. An electrical conductivity detector is most commonly used while ultraviolet/visible and amperometric detectors can also be used.

It is possible to separate and analyze transition metals by forming complexes with acid such as oxalic, citric, tartaric, and isobutyric. The selectivity and sensitivity of the method is dependent on the type of metal, its oxidation state, the complexing agent, and the *p*H. In general, ion chromatography allows for the separation and detection of a wide variety of ionic species. Some of these are F^-, Br^- NO_3^-, SO_3^{-2}, SO_4^{-2}, PO_4^{-3}, AsO_4^{-3}, SeO_4^{-2}, the alkali and alkaline earth metals, transition and lanthanide metals, ammonium and low molecular weight amines, quaternary ammonium compounds, weak organic acids, aliphatic and aromatic sulfonic acids, carbohydrates, and amino acids. Ion chromatography has been extremely useful in determining ionic concentrations in pickling baths, plating baths, and electrolysis solutions.

3.4 Gel Permeation Chromatography

Gel permeation chromatography is a technique which separates molecules by size. It is particularly useful for determining the molecular weight and molecular-weight distributions of polymeric materials. Low-molecular-weight species such as oligomers and additives can also be separated and analyzed by gel permeation chromatographic techniques.

Size separation is accomplished by passing a solution of a solid or liquid polymer through a column containing a semiporous or gel-like packing. Cross-linked polystyrene gel or porous silica is commonly used as a packing material. Separation is effected when smaller molecules are trapped in the pores or openings in the gel structure and larger molecules pass more quickly around the pores through channels in the packing material. Since the packing materials can contain a distribution of pore sizes, the polymer molecules will likewise be separated into a distribution based on their sizes. Larger molecules will elute first since their paths have been least restricted, while the smaller, more readily trapped molecules will take a longer time to elute from the column. Refractive index detectors are most commonly used. However, ultraviolet absorption, infrared absorption, and light scattering detectors are also used.

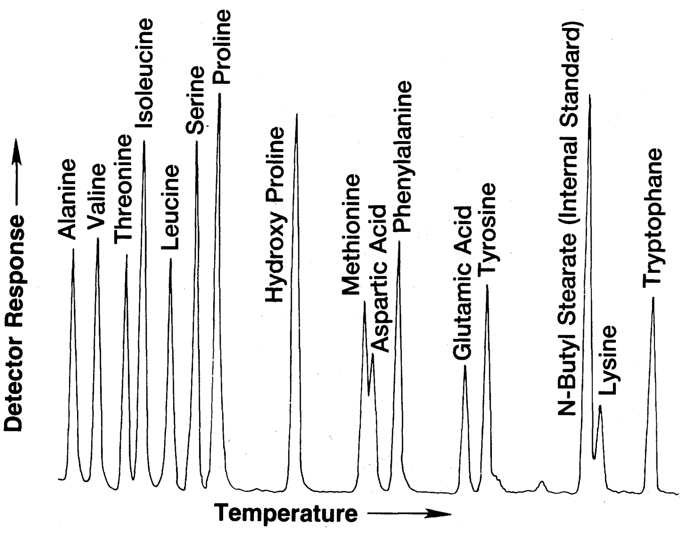

FIG. 11. Gas chromatogram of *n*-butyl esters of trifluoroacetylated amino acids.

In Fig. 12 are given the gel permeation chromatograms of two different styrene vinylbenzyl chloride copolymers. The elution times can be calibrated with molecular weights by first eluting polymers of known molecular weights. Note in the figure that the polymers are of different molecular weights. One polymer has a peak elution time of about 20 min corresponding to a peak molecular weight of about 700 000, while the other polymer has a peak elution time of about 23 min corresponding to a peak molecular weight of about 100 000. Some differences in the molecular-weight distributions of the polymers are also noted. The molecular weight and molecular-weight distribution are important properties of polymers which affect both their processibility and their end-use properties.

4. THERMAL ANALYSIS

Thermal analysis is a general term for a series of techniques all involving the measurement of changes which occur in materials when subjected to heat. The simplest of these is thermogravimetry (TG), which measures the weight change of a material as it is heated. In addition to TG there are other techniques which include differential scanning calorimetry (DSC), thermomechanical analysis (TMA), and dynamic mechanical analysis (DMA). Many other techniques such as thermal conductivity, dielectric thermal analysis, and dilatometry (which is a vital tool used by metallurgists for determining phase transitions), while also important in the characterization of materials, will not be included in the present article. The reader is referred to more general references for further detail and a discussion of these and other related subjects.

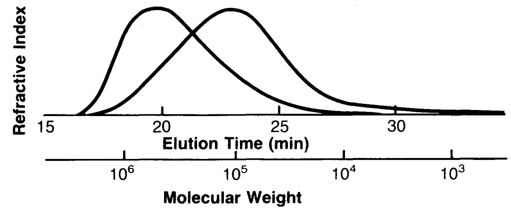

FIG. 12. Gel permeation chromatograms of styrene vinylbenzyl chloride copolymers.

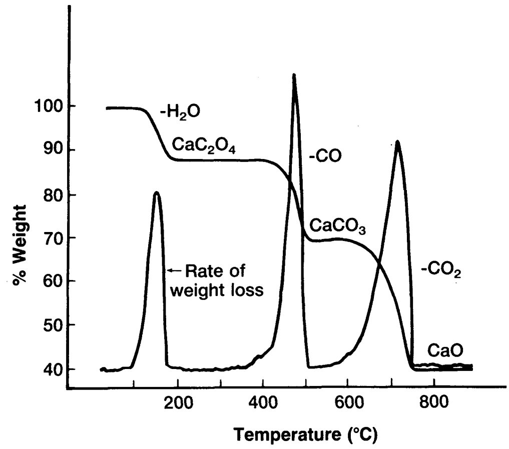

FIG. 13. Calcium oxalate monohydrate ($CaC_2O_4 \cdot H_2O$). TG in nitrogen (10 °C/min heating rate).

4.1 Thermogravimetry

Thermogravimetry (TG) is a technique to measure automatically the weight change as a material is heated either isothermally or in a programmed manner. The sample is placed on a very sensitive balance which is contained in an oven. Weight changes in the sample are recorded as the temperature of the oven is increased.

Thermogravimetry is particularly useful for understanding the thermal stability of materials. However, many other applications such as residue analysis, volatile contents, materials composition, and purity analysis are conducted on a routine basis. Materials such as polymers, metals, and ceramics can be heated in a variety of atmospheres (including vacuum and reactive gases) up to 2400 °C. If a TG instrument is coupled with a mass spectrometer, the composition of the volatiles emitted during degradation and volatilization can be detected. Sample amounts as small as 1 mg can be studied by this technique.

In Fig. 13 is given an example of a thermogravimetric analysis of a chemical compound, calcium oxalate monohydrate. The dehydration and decomposition is clearly seen to take place in three steps when the sample weight is plotted against the temperature of the sample at a heating rate of 10 °C/min. The last transition occurs about 700 °C, the final state of the sample at 800 °C being CaO.

If a magnet is placed around a sample which is suspended from a balance and if the sample is ferromagnetic, then a characteristic of the material called the Curie temperature can be obtained as the sample is heated. The Curie temperature defines a change in state of the material from ferromagnetic to paramagnetic. When a sample is originally placed above the magnetic field, an apparent weight gain is noted as the sample is pulled down into the magnetic field. At this point the instrument is set so that the observed weight is recorded as 100%. As the temperature is raised, the sample loses its ferromagnetism and an apparent "weight loss" begins to occur. In Fig. 14 is shown the "apparent" weight changes which occur as an amorphous metal alloy is heated up to temperatures of 900 °C. At about 400 °C, the Curie temperature, a sharp loss in apparent weight is noted. Other magnetic changes also occur in the material as devitrification or crystallization of the amorphous metal occurs at temperatures above 400 °C.

4.2 Differential Scanning Calorimetry

Differential scanning calorimetry (DSC) involves heating a sample along with a thermally inert reference material and measuring the relative heat changes that occur in the sample during the heating process. The most common heat change normally observed is the melting transition of a crystalline material which results in an absorption of heat or energy relative to the inert standard. It is observed as an endothermic peak in the thermogram. The reverse phenomenon, crystallization, which occurs upon cooling a crystallizable sample from the melt, manifests itself as a release of heat and is observed as an exothermic peak.

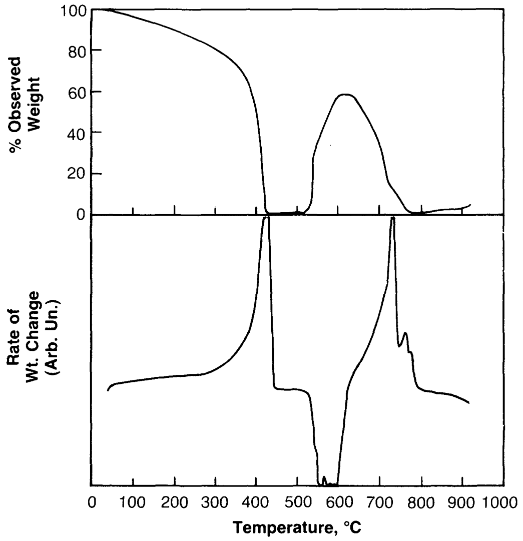

FIG. 14. Thermogravimetry of amorphous $Fe_{78}B_{17}Si_5$ in a magnetic field.

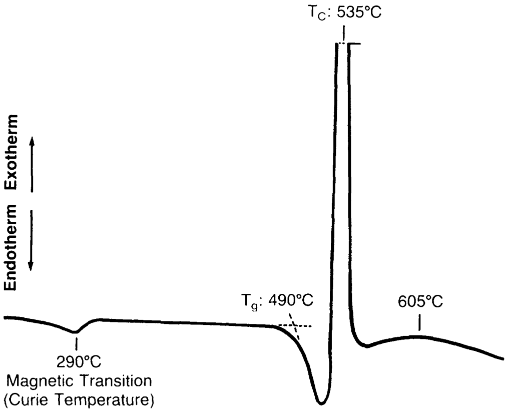

FIG. 15. Differential scanning calorimetric curve of an amorphous metal alloy heated in argon (20 °C/min heating rate).

Materials can undergo many other heat-sensitive changes which can be detectable by DSC. Loss of volatiles, the glass transition, multiple crystal phase transitions, degradation processes, melting of impurities, and melting of blends and alloys are other commonly observed phenomena of the DSC technique. All materials undergo changes when heated; therefore, the DSC technique is widely used and generally applicable.

In Fig. 15 is the thermogram of an amorphous metal alloy. The Curie temperature, previously described in the section on thermogravimetry, is accompanied by an endothermic heat change which occurs in this material at 290 °C. The temperature of the onset of large-scale molecular mobility, the glass transition temperature (T_g), is observed at 490 °C. Crystallization is evident with the sharp exotherm at 535 °C and the broader exotherm at 605 °C, indicating multiple or complex crystallization for this material.

The DSC technique is widely used in the characterization of polymers, composites, and blends. In Fig. 16 is given the thermogram of poly(ethylene terephthalate). The glass transition (T_g), crystallization (T_c), and melting (T_m) temperatures are all clearly indicated. In some cases the amount of heat absorbed during the melting process of a semicrystal-

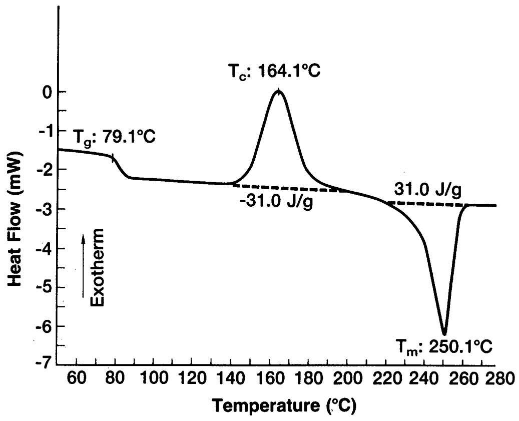

FIG. 16. Differential scanning calorimetry of poly(ethylene terephthalate) in argon (20 °C/min heating rate).

line polymer can be used as a measure of the crystallinity of the polymer when it is compared to the heat of fusion of the 100% crystalline material. However, care must be taken in this analysis since materials can crystallize during the heating process. In this case the melting endotherm does not represent the original crystallinity of the material.

4.3 Thermomechanical Analysis

Thermomechanical analysis (TMA) can give information, like dilatometry, about the dimensional changes which occur in a material when it is heated. Coefficients of expansion and shrinkage performance, two important characteristic of materials can readily be determined by TMA techniques.

Samples can be in the form of fibers, rods, plaques, bars, and even powders. In the TMA experiment, a sample is in contact with a rod-like probe whose motion can be detected as the sample is heated and undergoes contraction or expansion. The motion of the rod is detected and translated into an electrical signal by means of a linear variable differential transformer. The temperature range in TMA is from −150 to 1500 °C. The technique is used whenever information is needed about the dimensional stability of a material in a particular temperature range.

4.4 Dynamic Mechanical Analysis

The dynamic mechanical analysis technique provides information about the viscoelastic properties of materials as a function of temperature and time. A constant oscillating stress, at a frequency in the range of 0.01–200 Hz, is applied to the sample at strains of 0.03%–1.5% while the temperature is raised in a controlled manner. Samples are usually of the order of about 25 to 5000 μm thick and can be in the form of films, foils, sheets, or fibers.

A number of viscoelastic parameters can be obtained from DMA. One parameter is the dynamic (elastic) modulus (E') of a material. The dynamic modulus is an important measure of the stiffness of a polymer. It is important to know this parameter, e.g., for designing polymer-graphite composites. In Fig. 17 are given the plots of $\log E'$ as a function of temperature for graphite composites with three different polymer matrices. A second parameter, the dynamic loss (viscous) modulus E'' can also be obtained. The dissipation or damping factor, $\tan\delta = E''/E'$, is used as a measure of the damping characteristic of a material.

5. X-RAY ANALYSIS

There are a variety of techniques which utilize x rays to characterize and analyze materials. X-ray fluorescence spectroscopy (XRF) is a nondestructive qualitative and quantitative technique for elemental analysis. X-ray diffraction analysis (XRD) is used to determine the molecular structure of materials. These two techniques will be discussed in the following sections.

There are a variety of other x-ray techniques, such as x-ray absorption and small-angle x-ray scattering, which will not be included in the current article but which are

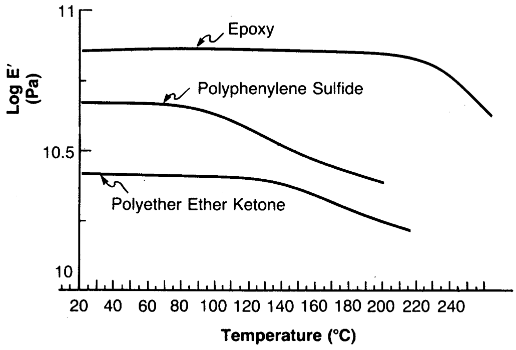

FIG. 17. Dynamic mechanical analysis of graphite fiber composites.

sufficiently important to be included as further reading in the references.

5.1 X-Ray Fluorescence Spectroscopy

X-ray fluorescence (XRF) (see INTERACTION OF SOLIDS WITH PARTICLES AND RADIATION) is a relatively simple, nondestructive technique for elemental analysis. Samples can be liquid or solid, and as little as milligram quantities is sufficient for analysis in many cases. The XRF technique is based on the principle that if a material is bombarded by a high-energy source, such as electrons or photons, some of the electrons in the atoms of the material will be ejected. As other electrons fill the vacated energy levels, quanta of radiation, characteristic of the atom, are emitted. Such quanta, in the form of x rays, can be detected by either a wavelength-dispersive x-ray fluorescence (WDXF) spectrometer or an energy-dispersive x-ray fluorescence (EDXF) spectrometer. In WDXF a crystal is used to separate the energies and a detector such as a scintillation counter is used to measure the intensity of the x rays as a function of energy, since different energies are associated with different wavelengths and thus different angles. In EDXF, a solid-state detector is used which has an inherent ability to separate the energies without any mechanical movement. The development of computers and accompanying software have been instrumental in improving the accuracy and precision of the data. The same techniques are applied at the microscopic level and are described in the section on electron probe microanalysis.

In Fig. 18 is given a plot of the lower limit of detection (LLD) as a function of atomic number for a WDXF detector and two different EDXF detectors. It is possible to detect and analyze all elements down to boron with the lower limit of detection, in some cases, down to several ppm.

5.2 X-Ray Diffraction Analysis

X-ray diffraction (*q.v.*) is one of the most important tools in characterizing the crystalline nature of materials. If a beam of monochromatic x radiation is directed at a crystalline material, then diffraction of the x-ray beam can occur. The relationship between the wavelength of the x-ray beam (λ), the angle of diffraction 2θ, and the distance between each set of atomic planes of the crystal lattice d is given by the well-known Bragg equation $n\lambda = 2d \sin\theta$, where n represents the order of diffraction. From this equation the interplanar distances can be calculated. By an analysis of the intensity of the x-ray scattering from the various crystalline planes, it is possible to reconstruct the total molecular structure of the crystalline material.

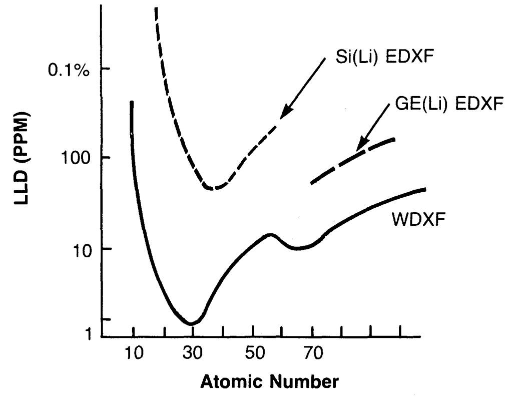

FIG. 18. Lower limits of detectability of elements with various x-ray fluorescence detectors.

For the purposes of the current article, x-ray diffraction analysis is divided into three areas: single-crystal analysis, powder diffraction techniques, and analysis of partially crystalline, partially oriented polymer materials. Single-crystal x-ray diffraction analysis is now a mature, well-defined methodology. Crystal structure of most metal alloys, inorganics, ceramics, and even complex organic and biochemical compounds can be determined in relatively short times (see CRYSTALLOGRAPHY). These techniques will not be described in this article but references are provided for further reading on the subject.

5.2.1 X-Ray Powder Diffraction. X-ray powder diffraction is used to obtain information about the structure, composition, purity, and state of polycrystalline materials. Materials to be analyzed can be powders, bulk solids, films, or ribbons. Sample sizes should be at least of the order of grams, although in some cases information can be obtained on milligram quantities. The powder x-ray pattern of a crystalline material, like its infrared spectra, can be used as a fingerprint, unique to that material alone. The Joint Committee on Powder Diffraction Standards has published the diffraction patterns of some 50 000 compounds. An unknown compound can be identified by comparing its interplanar spacings and intensities with those in the powder diffraction standards file. When elemental data on the material is added, such as that obtained from atomic absorption (AA) or emission or x-ray fluorescence (XRF) analysis, the number of patterns involved in the search can be reduced. The composition of compounds in mixtures of up to nine compounds can be determined for compounds present at concentrations greater than 1%. Computer methods have greatly facilitated these comparisons.

In Fig. 19 is given the powder diffraction pattern of an "unknown" inorganic material. A computer search reveals two stick diagrams, whose combined pattern matches the unknown. In this case the material in question is found to be a mixture of Al_2O_3 and AlN.

In addition to making identification of crystalline materials, powder diffraction techniques can also provide information on the crystallite size, the degree of crystal perfection, the degree of crystallinity in poorly crystallized metals and ceramics, and many other crystal state characteristics. In Fig. 20 is given the diffraction pattern of zirconia which has been partially stabilized with yttria. By use of an on-line least-squares program, the diffraction pattern has been deconvoluted to yield information about the various phases that are present.

5.2.2 X-Ray Diffraction of Polymers. X-ray diffraction analysis of polymers requires special attention because of the many unique features of polymers. Although there are many techniques such as IR, NMR, and DSC which can provide information about molecular ordering or crystallinity in polymers, XRD can be considered the definitive technique because it directly measures the three-dimensional relationship between molecules in polymeric crystals. XRD can provide information about the degree of crystallinity, crystallite sizes, degree of crystal perfection, orientation

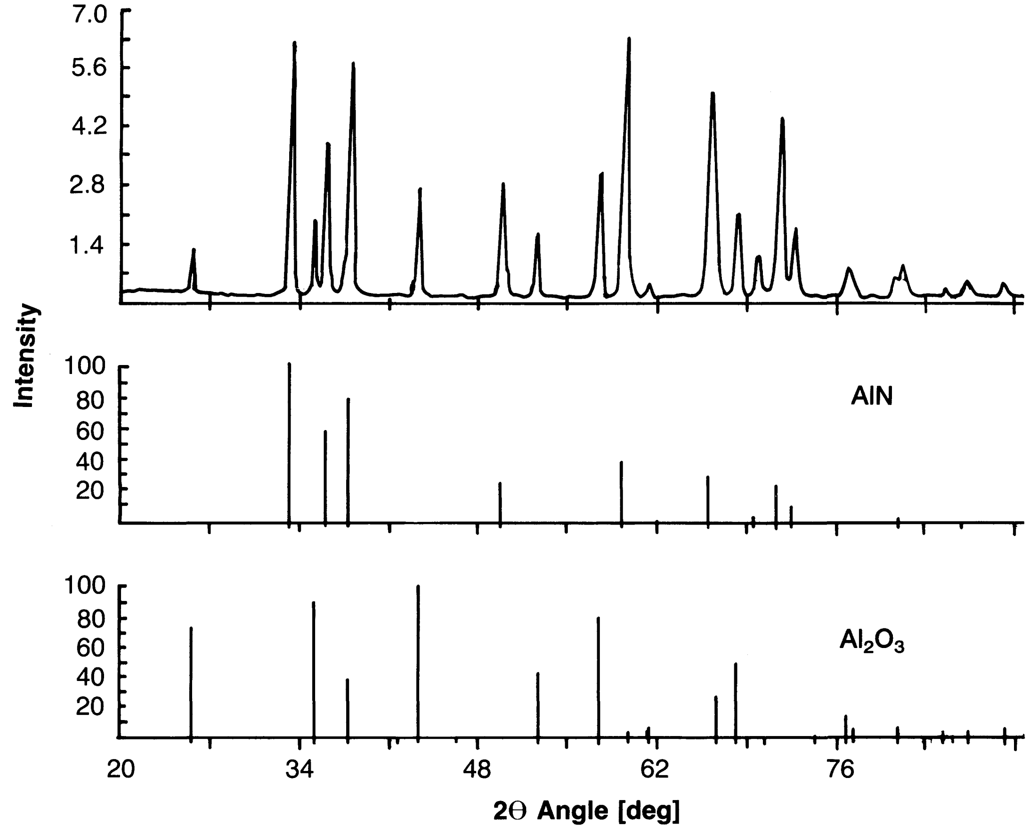

FIG. 19. Results of computer-assisted x-ray diffraction phase identification of a two-component system.

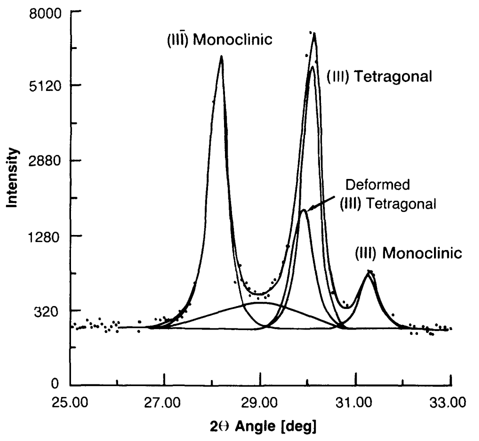

FIG. 20. Deconvolution of x-ray diffraction patterns of tetragonal and monoclinic phases in a $Zr(Y)O_2$ ceramic.

of the crystallites, and phase composition in polymorphic materials. Information has also been obtained in the amorphous state in both amorphous and semicrystalline polymers by XRD techniques. Materials to be examined by XRD can be in the form of fibers, films, plaques, pellets, powders, and even gels.

In Figs. 21(a) and 21(b) are given the diffraction patterns from an oriented fiber of Nylon 6. The photographic insert illustrates that the fiber is partially oriented and partially crystalline. By deconvolution of the patterns in Figs. 21(a) and 21(b), it is possible to determine the phase contents of the α and γ phases, the amount of amorphous phase (AM), the overall crystallinity, and the orientation of the different phases.

6. SURFACE ANALYSIS

The surface properties of materials play a large role in their performance. Oxidation and corrosion on metal surfaces can initiate bulk failure in the metal or metal alloy. Surface properties such as smoothness, adhesion, printability, wettability, friction, and wear are important for products such as molded plastics, plastic films, ceramics, and composite materials used in the automotive and aircraft industries. Surface properties of components are extremely important in the electronics and optical electronics fields. Since the importance of surface properties is clearly recognized, then attention must be focused on the means to measure and understand surface properties and surface phenomenon.

In this section a variety of surface analytical techniques will be described and their utilization discussed. Auger electron microscopy, x-ray photoelectron spectroscopy, Rutherford backscattering, electron energy-loss spectroscopy, secondary-ion mass spectrometry, scanning tunneling microscopy, atomic force microscopy, and electron diffraction techniques will be discussed. Again the reader is directed to additional readings since surface science is a broad field encompassing many techniques and disciplines, only some of which will be covered in this article.

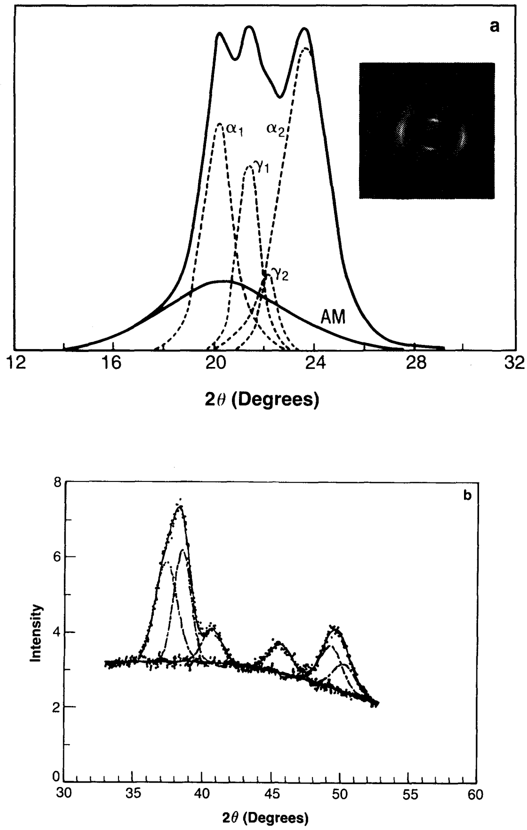

FIG. 21. Equatorial scan of an oriented nylon-6 fiber. (a) The pattern for 2θ from 15° to 30° resolved into α, γ, and amorphous components. The α form has two reflections at ~20.3° and ~23.5°; the γ has two reflections at ~21.3° and ~22.5°. (b) The higher angle equatorial scan of an α-rich fiber resolved into six individual reflections. Inset: The corresponding diffraction photograph of the fiber.

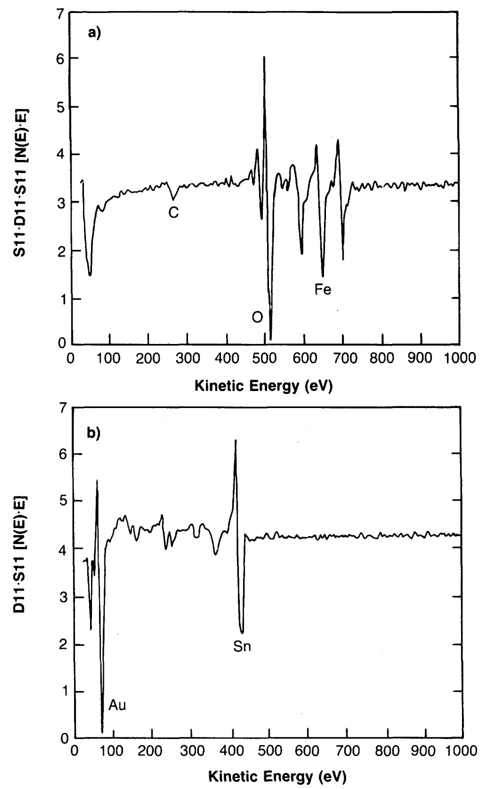

FIG. 22. Scanning Auger microanalysis of blue tarnish on a metal preform (a) initially and (b) after Ar ion sputtering.

6.1 Scanning Auger Microscopy

Scanning Auger microscopy (SAM) is a technique for determining the elemental composition of surfaces. Surfaces can be probed to depths of 1–5 nm, with spot analyses on the order of 100 nm. Sample sizes can range from a few millimeters up to about a square centimeter.

The source in an Auger spectrometer is a finely focused electron beam contained along with the sample in a high-vacuum chamber. When the electron beam impinges on a material, core-level electrons can be ejected from atoms in the material creating a core hole. An electron from a higher energy level can then fall into the core vacancy. The energy of this transition can be released by characteristic x rays or by the ejection of a second (Auger) electron. The Auger electron then passes into a detection system which determines the number of electrons and their kinetic energies. Metals and ceramics are readily analyzed by Auger microscopy; but the high-energy electron source causes degradation of polymeric materials, making analysis of these materials problematical by this method. Depth profiling can be accomplished by ion sputtering away the surface and determining the elemental composition at various sputtering times or depths into the sample.

One of the advantages of SAM over other surface techniques is its high spatial resolution. An example of this is shown in Figs. 22(a) and 22(b). The problem was to determine the cause of tarnishing of a Au/Sn alloy which was bonded to a Au-plated substrate containing iron. An Auger elemental scan of the tarnished area [Fig. 22(a)] revealed that C, O, and Fe were the only elements present on the surface. After argon-ion sputtering of the surface, the contaminants were removed and only a clear Au/Sn surface was revealed [Fig. 22(b)]. It was concluded that the tarnish was due to Fe which had migrated from the substrate to form an oxide layer of about 100 nm in thickness.

Another application of SAM is elemental mapping. Contamination of surfaces in the semiconductor industry is a serious problem. An example where SAM was able to identify the cause of such a contamination is given in Fig. 23. During fabrication of semiconductor wafers, a dark stain was noted on the aluminum bonding pads used to bond wire leads to the internal circuitry. An elemental fluorine map of the bonding pad area clearly showed the presence of fluorine as a contaminant on the bonding pad surface.

6.2 X-Ray Photoelectron Spectroscopy

X-ray photoelectron spectroscopy (XPS), also named electron spectroscopy for chemical analysis (ESCA), is a technique for probing the surfaces of materials at depths of 2–5 nm (see PHOTOEMISSION AND PHOTOELECTRON SPECTRA). It differs from Auger spectroscopy in that it is possible to determine the oxidation or valence states of atoms as well as detect their presence.

The excitation source in XPS is x-ray photons rather than electrons as in Auger spectroscopy. When such photons impinge on the surface of a material, photoelectrons in turn are emitted from the surface atoms of the material. These photoelectrons originate from discrete electronic energy levels associated

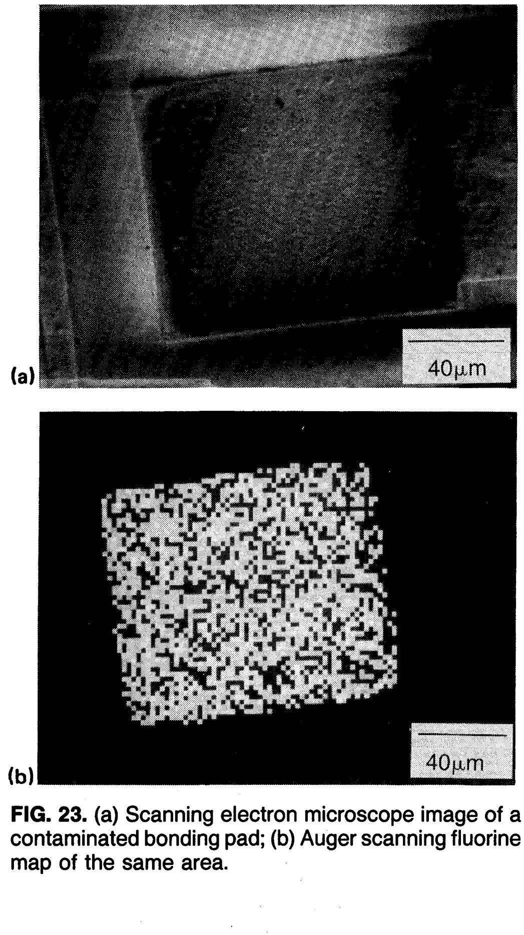

FIG. 23. (a) Scanning electron microscope image of a contaminated bonding pad; (b) Auger scanning fluorine map of the same area.

with these atoms that are unique and characteristic of the atoms. Since these energy levels can be dependent on the oxidation state and/or the electronic state of the atoms, the emitted photoelectrons can in turn give information about the local electronic environment. For example, it is possible to detect the difference between Mo in such compounds as Mo^0, MoS_2, and MoO_3, and the difference in elements in mixed oxidation states such as F_3O_4 (Fe^{2+} and Fe^{3+}) and Pb_3O_4 (Pb^{2+} and Pb^{4+}).

While the high energy of the source electrons in Auger spectroscopy prevents analysis of organic materials, materials such as organic polymers are readily analyzed by XPS. In addition XPS is applicable to metals, alloys, inorganics, and ceramics. In Fig. 24 is given an example of how XPS is used to solve an adhesion problem. The problem involved delamination of a fluoropolymer film from an epoxy-impregnated backing-board composite. The delaminated area reveals the presence of nitrogen and an alteration in the strength of the fluorine signal from that of the bonded area. These data suggested that an amine curing agent migrated from the backing to react with the fluoropolymer at the interface to cause delamination. In Fig. 25 is given an example of how XPS is used to solve a surface wetting problem. The wettability of Pb–In–Ag foils was found to change upon aging. Two major changes were noted upon aging. The lead oxide layer was thicker in the aged material and In, as evidenced by the $3d_{5/2}$ transition line, had migrated to the surface. Both these factors can affect the surface wettability.

6.3 Secondary-Ion Mass Spectrometry

Secondary-ion mass spectrometry (SIMS) can provide surface information on areas ranging from several square millimeters down to the submicron level. Any solid material can be studied with sample size usually limited to slightly less than 1 cm^2 and extending to as small as several micrometers.

In contrast to Auger spectroscopy where excitation is caused by electrons, and XPS where excitation is caused by x-ray photons, SIMS uses energetic ions as the excitation source. Ions such as He^+, Ar^+, and Xe^+ or metal ions can be used which in turn cause ejection of secondary ions from the sample surface. The ejected ions are detected and analyzed by a quadrupole mass spectrometer.

When noble gas ions are used the probe diameters are between about 1 mm and 10

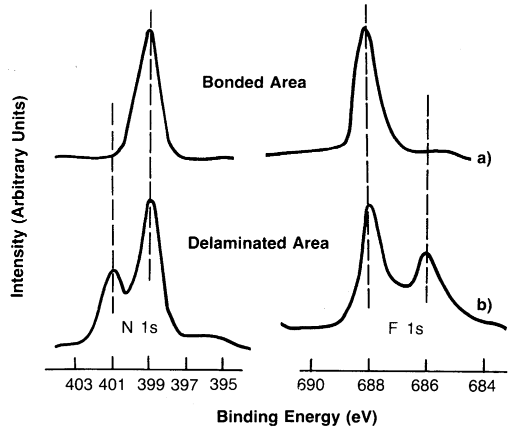

FIG. 24. The N 1*s* and F 1*s* spectra associated with (a) bonded and (b) delaminated areas of a composite board.

μm. However, if a liquid metal ion (LMI) source such as Cs or Ga is used, probe diameters < 50 nm may be achieved.

Both conducting and nonconducting materials can be studied by the SIMS techniques. Metals, ceramics, inorganic catalysts, and polymers can all be studied by SIMS. In Fig. 26 is an example where SIMS is used to detect contaminants on an integrated-circuit chip which could not be detected by Auger spectroscopy. SIMS was able to show that the contaminated surface had a much higher concentration of Na, Li, K, and B atoms, which affected the electrical performance of the chip.

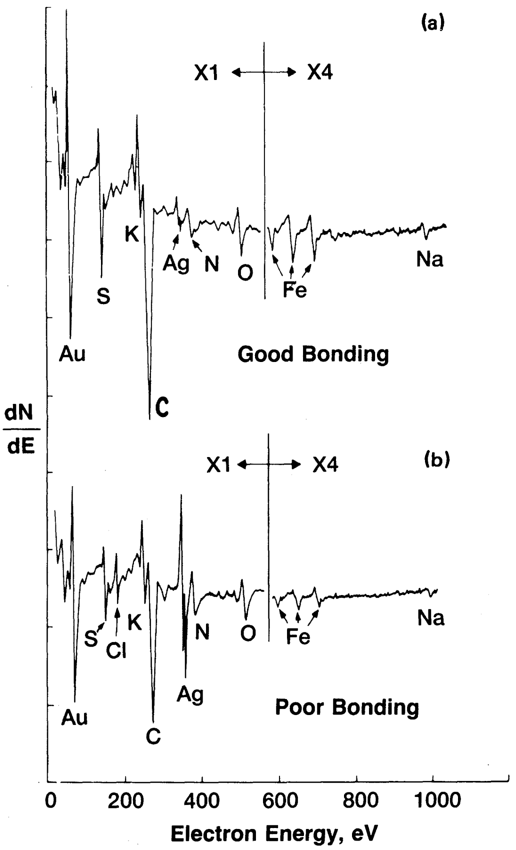

FIG. 26. Secondary-ion mass spectrometry spectra of deposited SiO_2 layers for (a) control sample and (b) contaminated sample. (Reprinted from PHI Application Note, 1979 with permission of the Perkin Elmer Corporation.)

6.4 Rutherford Backscattering

When a beam of energetic ions strikes a solid surface, the change in energy of the backscattered ions is characteristic of the nature of the solid. The intensity of the scattered beam, determined by a suitable detector, is related to the number of atoms doing the scattering and to the depth of penetration, which can be varied by altering the velocity of the exciting ions. Therefore, the method can be quantitative.

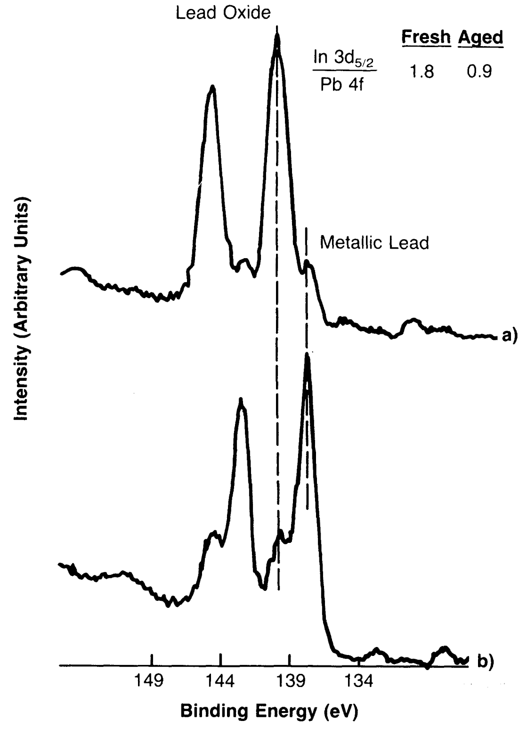

FIG. 25. Typical Pb 4*f* x-ray photoelectron spectra for (a) aged and (b) fresh Pb–In–Ag foil.

He^+ is the most generally used ion although H^+ may be used if deeper penetration into the sample is desired; or C^+, N^+, or O^+ can be used as source ions if greater mass separation is needed. The method can be used to determine the identity of a surface contaminant if the atomic mass of the contaminant is greater than that of the substrate. The technique has been useful in the identification of contaminants of heavy atoms on the surface of solid-state devices, such as might occur after etching, anodizing, or chemical treating.

Rutherford backscattering is also a good technique for determining the thickness of thin films deposited on surfaces, as might occur in chemical vapor deposition or plasma deposition processes. Such measurements have been made for gold, copper, aluminum, and silicon deposited on graphite surfaces. When a surface has been chemically altered, e.g., by oxidation or nitriding, the thickness and composition of the surface layer can be determined.

In general, Rutherford backscattering is applicable to the analysis of impurities which exist in a thin layer either on or within a few hundred nanometers of the surface, when the impurities are of greater atomic mass than the substrate, and when the amount of impurities is not too great.

6.5 Electron Energy-Loss Spectroscopy

In electron energy-loss spectroscopy (EELS), an energetic beam of electrons passes through a thin sample and interacts with it. The loss of energy of the transmitted electrons provides information about the chemical, physical, and electrical nature of the sample. EELS can be accomplished by the addition of an electron energy analyzer after the sample in a conventional STEM or TEM. Sample requirements are similar to those for conventional TEM.

A typical EELS spectrum will consist of three regions. They are a zero-loss region (~ 0 eV), a low-loss region (< 50 eV), and a core-loss region (> 50 eV).The unscattered transmitted beam accounts for most of the zero-loss region, and plasmons and interband and intraband transitions of the outer atomic electrons account for the low-loss region. The core-loss region is the most useful analytical portion of the spectrum and arises from transitions of the core electrons to states above the Fermi level. For lighter elements EELS has an advantage over energy-dispersive x-ray fluorescence spectroscopy. The elemental composition can be quantified by measuring the intensity of the edge signals. In addition the shape of the energy-loss spectral edge will depend upon the chemical and crystallographic environment of the atoms in the sample. For example, different types of carbon, i.e., graphite, diamond, and amorphous carbon, are readily distinguishable by EELS.

6.6 Scanning Tunneling and Atomic Force Microscopy

Scanning tunneling microscopy (STM) (*q.v.*), atomic force microscopy (AFM), and a number of other similar techniques are used to study the morphology and the microstructure of surfaces at the atomic level. The STM technique is based on the use of a tungsten or platinum stylus which is passed over the surface of a relatively flat conductive material. If a potential difference is maintained between the stylus and the specimen and the stylus is scanned very close to the surface, a current will be generated since electrons from the specimen can "tunnel" their way across the gap between the surface of the specimen and the tip of the stylus. As the stylus is traversed over the surface in a constant-current mode, it will experience upward and downward motions corresponding to the surface topography at the atomic level. The motion of the stylus can be sensed by a transducer, the signals from which can be converted into a contour map of the surface. There is also a constant-z mode in which the vertical position of the stylus is maintained constant and the varying current is monitored. In Fig. 27 is shown a contour STM map of the surface of graphite. The resolution is such that a single-atom defect in the surface is clearly shown.

In atomic force microscopy a sharpened stylus, attached to a cantilever, is scanned across a surface. As the stylus is scanned across the sample, the electronic forces between the stylus tip and the surface atoms cause a deflection of the cantilever. These deflections can be monitored in a number of ways, such as by interferometry or by use of a laser beam. A map of the deflections results in a map of the surface topography or morphology at the atomic level. The AFM technique can also be modified to respond to a sample's magnetic field (magnetic force microscopy) or to electric charges on the surface (electrostatic force microscopy).

STM, AFM, and related techniques are used to study the atomic structure on crystal surfaces, cleaved crystal surfaces, and thin films of semiconductors, metals, or biological specimens which have been deposited on suitable substrates.

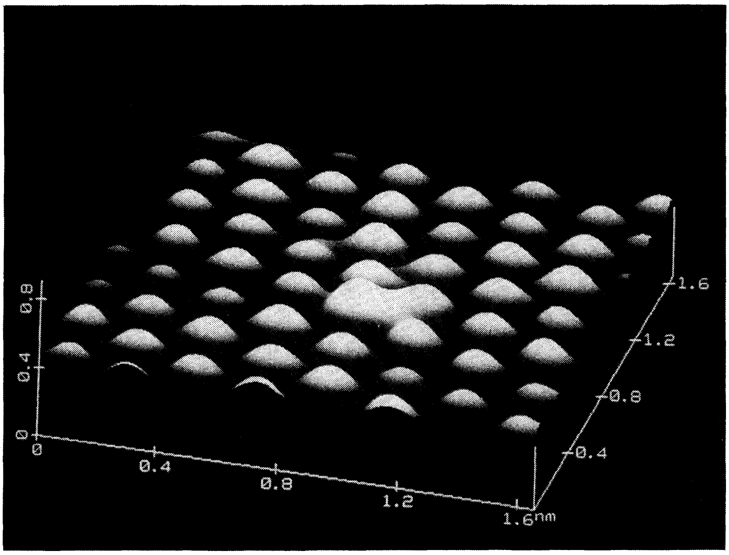

FIG. 27. A single-atom defect in graphite demonstrating the resolution of subtle atomic structures achievable with STM. (Reprinted from Nanoscope II 1990 Bulletin, with permission of Digital Instruments, Inc.)

6.7 Low-Energy Electron Diffraction and Reflection High-Energy Electron Diffraction

Low-energy electron diffraction (LEED) is based on the principle that slow electrons are mainly reflected by the first few atomic layers. Electrons with energies in the range of 10 to 500 eV impinge on a sample at normal incidence and the backscattered electrons which have been diffracted from the first few atomic layers can form a diffraction pattern on a fluorescent screen. In reflection high-energy electron diffraction (RHEED), the high-energy (~ 10 keV) beam impinges on the sample at a glancing angle and the reflected diffracted electrons are collected by a detector such as a photographic plate. The electron diffraction patterns obtained by LEED and RHEED are obtained on solid samples and are essentially the same as those obtained in transmission electron microscopy where electron energies of several hundred keV are used (see ELECTRON DIFFRACTION).

Surfaces must be carefully prepared by polishing, cleaving, or depositing such that appropriate crystal faces and orientations are obtained. Surfaces must also be free of all contaminants. Procedures such as annealing in vacuum, using an oxidizing or reducing atmosphere, or ion beam etching can be used to clean surfaces. Auger spectroscopy is often used in combination with LEED to monitor cleanliness of the surface. LEED provides similar information to RHEED; however, in some cases one technique is more applicable than the other. For example, LEED can determine atomic positions (to 0.01 nm) more readily than RHEED, while three-dimensional surface structures are more readily determined by RHEED.

In general, LEED and RHEED are used to identify surface compositions by measurement of lattice parameters, determine surface crystallography and microstructure, and measure dynamic processes such as surface crystal growth. The techniques have been applied to semiconductors, metals, alloys, catalysts, graphite, mica, and alkali halides. The location and arrangements of atoms on the surface of crystals is very important to the understanding of phenomena such as absorption, vapor deposition, and chemical and catalytic activity.

GLOSSARY

Atomic Spectroscopy: The field of science which involves the qualitative and quantitative elemental analysis of materials on the basis of the frequencies of light that they emit or absorb.

Electron Microscopy: A series of high-resolution microscope techniques based on the principle of the interaction of electrons with materials.

Image Analysis: The analysis of the particle shapes and sizes or morphology of material surfaces accomplished either directly or from a photograph, based on the difference in light contrast in the various features of the sample.

Molecular Spectroscopy: The field of science dealing with the molecular structure of materials as determined from the interaction of the material with electromagnetic radiation.

Morphology: The manner in which molecules are arranged in a material to form larger species such as crystallites, lamellae, grains, etc.

Surface Analysis: The science of examining surface features of materials at probing depths down to 1 nm and at atomic level resolutions.

Thermal Analysis: A series of techniques which can measure the weight loss, heat change, and mechanical property changes which occur in a material when it is heated under controlled conditions.

ACRONYMS AND INITIALIZATIONS

AA:	Atomic Absorption Spectroscopy
AFM:	Atomic Force Microscopy
DMA:	Dynamic Mechanical Analysis
DSC:	Differential Scanning Calorimetry
EELS:	Electron Energy Loss Spectroscopy
EPMA:	Electron Probe Microanalysis
ESCA:	Electron Spectroscopy for Chemical Analysis
FAB-MS:	Fast Atom Bombardment Mass Spectrometry
GC:	Gas Chromatography
GPC:	Gel Permeation Chromatography
IC:	Ion Chromatography
ICP:	Inductively Coupled Plasma Spectroscopy
IR:	Infrared Spectroscopy
LC:	Liquid Chromatography
LEED:	Low-Energy Electron Diffraction
LLD:	Lower Limit of Detection
MS:	Mass Spectrometry
NIR:	Near-Infrared Spectroscopy
NMR:	Nuclear Magnetic Resonance Spectroscopy
RHEED:	Reflection High-Energy Electron Diffraction
SAM:	Scanning Auger Microscopy
SEM:	Scanning Electron Microscopy
SIMS:	Secondary-Ion Mass Spectrometry
STEM:	Scanning Transmission Electron Microscopy
STM:	Scanning Tunneling Microscopy
TEM:	Transmission Electron Microscopy
TG:	Thermogravimetry
TMA:	Thermomechanical Analysis
UV:	Ultraviolet Spectroscopy
WDXF:	Wavelength Dispersive X-Ray Fluorescence
XDXF:	Energy Dispersive X-Ray Fluorescence
XPS:	X-Ray Photoelectron Spectroscopy
XRD:	X-Ray Diffraction Analysis
XRF:	X-Ray Fluorescence Spectroscopy

Further Reading

Alexander, L. E. (1969), *X-Ray Diffraction Methods in Polymer Science*, New York: Wiley.

Barrett, C. S., Cohen, J. B., Faber, J., Jr., Jenkins, R., Leyden, D. E., Russ, J. C., Predecki, P. K. (1986), *Advances in X-Ray Analysis*, New York: Plenum.

Benninghoven, A., Giber, J., Laslo, J., Riedel, M., Werner, H. W. (Eds.) (1982), *Secondary Ion Mass Spectrometry*, New York: Springer Verlag.

Bever, M. B. (Ed.) (1986), *Encyclopedia of Materials Science and Engineering*, Oxford: Pergamon; Cambridge, MA: MIT Press, Vols. 1–8.

Briggs, D. (Ed.) (1972), *Handbook of X-Ray and Ultraviolet Photoelectron Spectroscopy*, London: Heyden.

Briggs, D., Seah, M. P. (1983), *Practical Surface Analysis by Auger and X-Ray Photoelectron Spectroscopy*, New York: Wiley.

Cautle, J. E. (Ed.) (1982), *Atomic Absorption Spectrometry*, New York: Elsevier.

Eddington, J. W. (1976), *Practical Electron Microscopy in Materials Science*, New York: Van Nostrand Reinhold.

Egerton, R. F. (1986), *Electron Energy Loss Spectroscopy in the Electron Microscope*, New York: Plenum.

Ferry, J. D. (1980), *Viscoelastic Properties of Polymers*, New York: Wiley.

French, A. D., Gardner, K. H. (Eds.) (1980), *Fiber Diffraction Methods*, Washington, DC: American Chemical Society.

Fuggle, J. C. (1981), in Brundle, C. R., Baker, A. D. (Eds.), *Electron Spectroscopy: Theory, Tech-*

niques and Applications, London: Academic Press, Vol. 4.

Griffiths, P. R., de Haseth, J. A. (1986), *Fourier Transform Infrared Spectrometry*, New York: Wiley.

Harris, R. K. (1983), *Nuclear Magnetic Resonance Spectroscopy*, Massachusetts: Pitman.

Hemminger, N., Hohne, G. (1984), *Calorimetric Fundamentals and Practice*, Basel: Verlag Chemie.

Keattch, C. J., Dollimore, D. (1975), *An Introduction to Thermogravimetry*, New York: Heyden.

Kolthoff, I. M., Elving, P. J. (Eds.) (1965), *Treatise on Analytical Chemistry*, New York: Interscience, Vol. 6, Chap. 64.

Lagally, M. G. (1986), "Low-Energy Electron Diffraction" in *Metals Handbook*, Vol. 10, *Materials Characterization*, 9th ed. Metals Park, OH: American Society for Metals.

Lipkowitz, K. B., Boyd, D. B. (Eds.) (1990), *Reviews in Computational Chemistry*, New York: VCH.

Long, D. A. (1977), *Raman Spectroscopy, Theory and Practice*, New York: Plenum.

Mann, C. K., Vickers, T. J., Gulick, N. M. (1974), *Instrumental Analysis*, New York: Harper and Row.

Mackintosh, W. D. (1974), in P. F. Kane and G. B. Larabee (Eds.), *Characterization of Solid Surfaces*, New York: Plenum Press, Chap. 16.

Mason, C. W. (1983), *Handbook of Chemical Microscopy*, New York: Wiley.

Olesik, J. W. (1991), *Anal. Chem.*, **63**, 12A–21A.

Newbury, D., Joy, D., Echlin, P., Fiori, C., Goldstein, J. (1986), *Advanced Scanning Electron Microscopy and X-Ray Microanalysis*, New York: Plenum.

Sibilia, J. P. (Ed.) (1988), *A Guide to Materials Characterization and Chemical Analysis*, New York: VCH.

Sibilia, J. P., Paterson, A. R. (1965), *J. Polym. Sci.* **8**, 41–57.

Snyder, R. L., Condrate, R. A., Johnson, P. F. (Eds.) (1985), *Materials Science Research* Vol. 19, *Advances in Materials Characterization* II, New York: Plenum.

Spencer, M. (1982), *Fundamentals of Light Microscopy*, Cambridge: Cambridge University Press.

Tertian, R., Claisse, F. (1982), *Principles of Quantitative X-Ray Fluorescence Analysis*, Philadelphia: Heyden.

Thompson, M., Walsh, J. N. (1983), *A Handbook of Inductively Coupled Plasma Spectrometry*, Glasgow: Blackie and Son, Limited.

Tromp, R. M., Hamers, R. J., Demuth, J. E. (1986), *Science* **243**, 304–309.

Turi, E. A. (Ed.) (1981), *Thermal Characterization of Polymeric Materials*, New York: Academic Press.

Watson, J. T. (1985), *Introduction to Mass Spectrometry*, New York: Raven.

Wendlandt, W. W. (1985), *Thermal Methods of Analysis*, New York: Wiley.

16

Chemical Analysis

Kenneth A. Rubinson

16.1 Introduction 529
16.2 Choosing the Analytical Methods 529
16.2.1 Standards and Validation 530
16.2.2 Precision, Bias, and Accuracy 531
16.2.3 Interferences and Representative Samples 531
16.2.4 Destructive and Nondestructive Analyses 531
16.2.5 Sample Classification by Size and Analyte Level 532
16.2.6 Sensitivity and Detection Limit 533
16.2.7 Imaging, Hyphenated, and Multidimensional Techniques 537
16.3 Assay Methods Based on Known Specific Chemical Reactions 538
16.3.1 Gravimetric Methods 538
16.3.2 Titrations 539
16.3.3 Analyses Based on Chemical Kinetics 539
16.4 Spectrometric Assay Methods 540
16.4.1 Fundamentals 540
16.4.1.1 Energy in Spectrometry 541
16.4.2 General Types of Spectrometry 542
16.4.2.1 Emission Spectrometry 542
16.4.2.2 Absorption Spectrometry 542
16.4.2.3 Fluorescence/Phosphorescence Spectrometry 544
16.4.3 Specific Spectrometric Methods 545
16.4.3.1 Gamma Spectrometry: Neutron Activation Analysis 545
16.4.3.2 Gamma Spectrometry: Mössbauer Spectrometry 546
16.4.3.3 X-ray Spectrometry: X-ray Fluorescence, X-ray Photoelectron Spectrometry, and Particle-Induced X-ray Emission 547
16.4.3.4 Ultraviolet, Visible, and Near-Infrared Spectrometry 549
16.4.3.5 Vibrational Spectrometry: Infrared Spectrometry and Raman Spectrometry 551
16.4.3.6 Microwave Methods: Microwave Spectrometry, Electron Paramagnetic Resonance 552

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

16.4.3.7 Radiofrequency Spectrometries: Dielectric Absorption, Nuclear Magnetic Resonance 554
16.5 Mass Spectrometry 558
16.5.1 Mass Spectrometers 559
16.5.2 Multidimensional Mass Spectrometry: MS^n 560
16.6 Electrochemical Methods 561
16.6.1 Potentiometry 562
16.6.2 Conductimetry 562
16.6.3 Methods that Use Oxidation and Reduction 563
16.6.3.1 Coulometry 563
16.6.3.2 Amperometry and Voltammetry 564
16.6.3.3 Anodic (and Cathodic) Stripping Voltammetry 565
16.7 Separations and Chromatography 566
16.7.1 Liquid Chromatography 567
16.7.1.1 Column Chromatography 567
16.7.1.2 Separations of Chiral Molecules 568
16.7.1.3 Thin-Layer Chromatography 569
16.7.1.4 Field-Flow Fractionation 569
16.7.2 Gas Chromatography 570
16.7.3 Supercritical Fluid Chromatography 571
16.7.4 Electromigration 571
16.8 Surface-Sensitive Methods 571
16.8.1 Electron Spectrometry 572
16.8.1.1 Photoelectron Spectrometry 572
16.8.1.2 Auger Spectrometry 573
16.8.2 Secondary-Ion Mass Spectrometry 573
16.8.3 Rutherford Backscattering 574
16.9 Thermal Methods of Analysis 574
16.9.1 Thermogravimetric and Evolved-Gas Analyses 575
16.9.2 Differential Scanning Calorimetry; Differential Thermal Analysis 575
Acronyms of Analytical Techniques and Instruments 576
Glossary 577
References 579
Further Reading 579
Further Reading: Textbooks 579

16.1
Introduction

Chemical analysis is the study and practice of learning about the elemental and molecular compositions of materials. The materials analyzed include all the natural and synthesized materials of the contemporary world as well as perhaps less obvious materials such as the components of flames or the materials at the atomic layer that indicates the interface among liquids, solids, and gases. The practice of chemical analysis involves all of the following: deciding on the various techniques that should be used; proving that the techniques do, indeed, provide the expected information under the conditions present; and demonstrating that the chemical content that is found does, in fact, answer the germane questions. In other words, contemporary chemical analysis invokes the use of the principles of chemistry that describe chemical behavior and structures, the applications of physics to design the instruments used for measurements, and the tools of statistics and logic to weave the technology to societal desires.

16.2
Choosing the Analytical Methods

The great majority of materials are chemically complicated; this means that they generally are composed of a large number (up to millions) of different elements and molecular species. Materials may be solids, liquids, gases, solutions, glasses, flames, and other forms of matter such as rubbers and gels. When one does an analysis to determine the amounts of species and their identities, the work implicitly includes the necessity of proof. Chemical analysis is, then, the study and practice of learning about materials – their elemental and/or molecular composition – and the tools of statistics and logic to determine the level of certainty in the results. Chemical analysis can also involve determining the spatial distribution of chemical species on the surface of or within a material.

A specific element or compound that is being assayed is called an *analyte*. To determine the presence of one or more analytes and the levels at which they are present requires a series of methods of

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

Tab. 16.1 Primary general applicability of analytical assay methods.

Method	Element	Multielement	Molecule	Surface	Polymer
Absorption spectrometry (UV–visible)	x	x	x		x
Anodic stripping voltammetry	x	x			
Atomic absorption spectrometry	x				
Auger electron spectrometry	x	x		x	
Electron microprobe	x	x		x	
Electron paramagnetic resonance	x	x	x		x
Electrochemical methods	x	x	x		
Emission spectrometry (UV–visible)	x	x			
Fluorescence (UV–visible)			x	x	x
Gas chromatography			x		x
Infrared, Raman spectrometry			x	x	x
Liquid chromatography	x	x	x		x
Mass spectrometry (MS)	x	x	x	x	x
Gas chromatography/MS			x		
Plasma MS	x	x			
Mössbauer spectrometry	x			x	
Neutron activation analysis	x	x			
Nuclear magnetic resonance			x		x
Thermal methods			x		x
Titrations	x		x		x
X-ray diffraction			x	x	x
X-ray fluorescence spectrometry	x	x			
X-ray photoelectron spectrometry	x	x		x	

analysis; this series is called the *methodology* or *protocol of the analysis*.

In Table 16.1, some of the assay techniques that are used for various types of analyses are indicated. This classification is quite rough, and should be understood to be neither exclusive nor complete.

16.2.1
Standards and Validation

Chemical analysis is usually carried out by indirect measurements such as mass, time, volume, and voltage. Such values are converted into measurement of content through calibration. In this sense, to calibrate means to ascertain the relationship between the content of the sample and the response found with the assay method.

As a result, an extremely important requirement is to demonstrate that any analytical procedure measures what the analyst says it measures on a specific type of sample. This is called *validation of the analytical method*. Validation generally means that the methodology must first produce correct results on accepted standards. Chemical standards are pure substances, mixtures, solutions, gases, or materials such as alloys or biological substances that are used to calibrate and determine the validity of each step of a methodology. These steps are sampling, sample preparation for the assay, choice of assay method(s), and statistical treatment of the data to determine the level of certainty in the result. This set of techniques making up the chemical analysis constitute the metrology of chemical analysis, which is covered in greater detail in a separate article in the Encyclopedia of Applied Physics.

16.2.2
Precision, Bias, and Accuracy

We can describe the difference between the precision and the accuracy of results in chemical analysis with the following analogy. Consider a target in which the bull's-eye at the center is analogous to the true result of an analysis. In shooting at the target, we might find that we have a number of hits clustered far off of the bull's-eye. The shooting is, then, precise, but has a bias. The precision of the results from a series of samples of the same material indicates how close they are to each other, that is, the reproducibility. The value of the precision does not describe the bias of the methodology, which is only determined with the aid of standards. Only if the results are closely clustered (labeled as *low imprecision*), and when the bias can be eliminated, does the average of the results lie in or near to the bull's-eye (the ''truth''). The analysis is then said to be *accurate*. Presenting a statistical uncertainty in the measured result is an essential part of the process, for example, 2.53 ± 0.04 ng.

16.2.3
Interferences and Representative Samples

The necessity for proof of a measurement of some analyte in, for example, a semiconductor or the air over a city results from the fact that no method of analysis, whether based on specific chemical reactions or those of applied physics, is free from interferences. Interferences (or interferents) are elements and compounds that are not only being measured but are also present in the material (in this context called the *matrix*) that cause, to some extent, a change in the measurement of the analyte. In other words, the measured quantity of analyte alone (say, in pure water) may not be equal to the same quantity measured in the presence of an interferent (say, in seawater).

The second problem is that the sample that is measured may not truly represent the material that is being analyzed. If one is analyzing for certain organic compounds in, say, a lake, then a decision must be made whether to sample from the mud at the bottom, the sand at the shore, and/or water at different locations and depths. It is highly likely that the organic compounds are not evenly distributed, and some care must be taken on how to report the results.

16.2.4
Destructive and Nondestructive Analyses

Another point to consider in any kind of analysis is whether or not the sample is to be destroyed. For instance, it is doubtful whether we would want to dissolve a priceless Egyptian statuette in acid in order to determine the brass formulation. To keep a sample intact, or cause no unacceptable change in its properties, it is necessary to use nondestructive analytical methods. The material may be modified on the atomic or molecular level to some extent, but otherwise the modification is undetectable. To cause as small a change as possible in the original sample, nondestructive tests are carried out with assay methods that are also used for microanalyses or trace analyses. One method that is useful for both micro- and trace analyses is neutron activation analysis (NAA). Two other microanalytical techniques are the electron microprobe and X-ray photoelectron spectroscopy (XPES, also called *electron spectroscopy for chemical analysis* or *ESCA*). In addition, a number of spectroscopic methods that only require collecting light reflected from the surface can be used to test nondestructively, for example, the composition of paints.

A larger number of techniques require destruction of the sample. Any methodology that requires the digestion with acid or other homogenizing methods must be classified as a destructive analytical methodology. A middle ground in classification arises when a small enough sample is needed for analysis so that sufficient material can be collected, say, from drilling a small pit in an obscure place on a valuable object. The method used for analysis might be destructive, but the sample is so small that the object is essentially intact.

16.2.5
Sample Classification by Size and Analyte Level

The size of a laboratory sample should not be confused with the concentration level of analyte in a sample. A descriptive classification scheme for these two properties is presented in Figure 16.1. The sample masses are divided into macro and micro with a vague boundary. The concentrations range from major through minor to trace

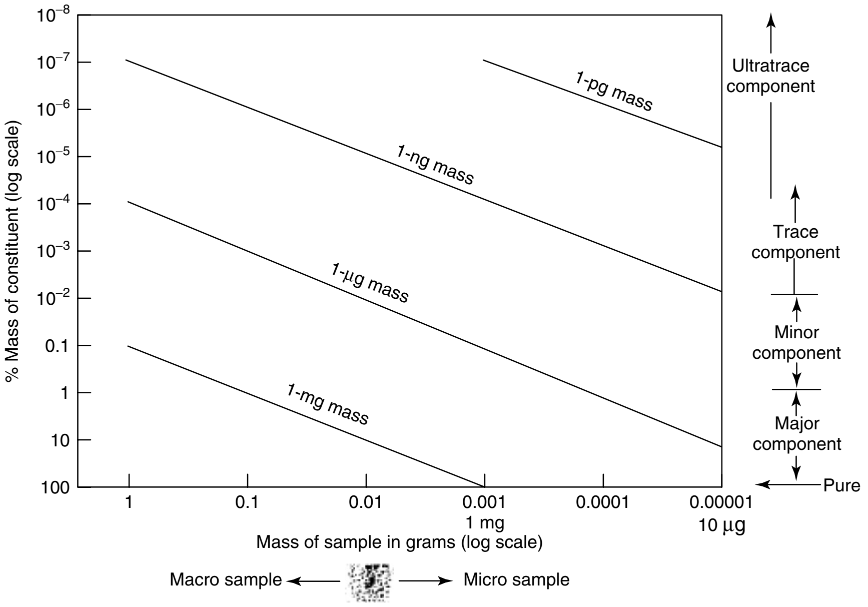

Fig. 16.1 Illustration showing the names given to the ranges of sample size and ranges of analyte content. The horizontal axis has a logarithmic scale of sample size. On the vertical axis is a logarithmic scale of the percentage by weight of a constituent in a sample. The diagonal lines correspond to the mass of the analyzed component. Samples are classified by the relative amount of an element or compound in it. The component can range from major (1–100%), through minor (0.01–1%) and trace (less than 0.01% = 100 ppm), to ultratrace (in the range of parts per 10^9 or less). (Modified from Sandell, 1959.)

and ultratrace. The figure caption contains further details.

When reading the analytical literature of trace and ultratrace analysis, everyone should be aware of a dichotomy in the nomenclature parts per billion (ppb) and parts per trillion (pptr). One group (primarily in the United States) ppb is parts in 10^9, and pptr is parts in 10^{12}, while for another group (primarily Europe) ppb is parts in 10^{12}, and pptr is parts per 10^{18}.

When the content of a component decreases into the parts per million (ppm) range and lower, eliminating interferences from the sample matrix becomes more difficult. As a result, disagreements among laboratories may increase until meaningful exchanges of information on content are not possible. An illustration of the trend is shown in Figure 16.2.

During analyses of trace components, every reagent, such as an acid used to dissolve a solid, must contribute as little interference as possible. Even the containers made of glass or plastic can contaminate solution samples. Such contributions of interfering species is one of the reasons for the trends shown in Figure 16.2. Analyses at the lowest concentration levels require that the laboratory spaces in which samples are handled be clean and filled with filtered air, and that reagents are highly pure and are stored in special containers. At these levels, even the containers that carry the samples to the laboratory must be clean and tested for suitability for the specific sample and for the assay to be run.

16.2.6
Sensitivity and Detection Limit

Two terms that are closely associated with measurements in trace and microanalyses are the sensitivity and the detection limit. The definitions of these terms in a chemical context are similar to our everyday language. If an instrument or other assay methods is more sensitive, it is able to measure smaller changes in content or concentration. *Sensitivity* is defined as the change in response for a given change in concentration or amount; a larger number means the method is more sensitive.

The detection limit is a numerical indication of the level (either concentration or mass) below which we cannot ascertain whether some analyte is present or not. Detection limits of methodologies are extrapolated from experiments done under optimized conditions. (See Table 16.2 for

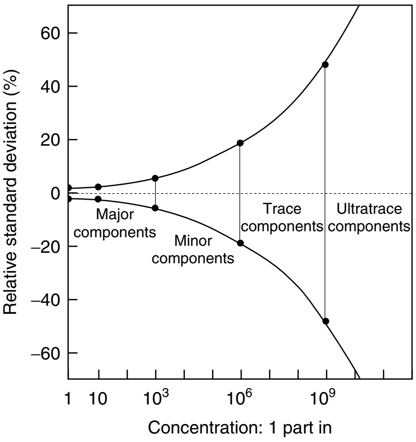

Fig. 16.2 When comparing results among analysts and among different laboratories, a tendency exists to exhibit an increase in the uncertainty of the results with decreasing concentrations of analyte. The horizontal axis is a logarithmic plot of relative concentration (ppb is parts in 10^9, pptr is parts in 10^{12}); the vertical axis is the percent relative standard deviation of the results, a common measure of uncertainty. This plot shows a general trend that has been demonstrated, not a result for a specific analyte or method. (Derived from Horwitz, 1982.)

Tab. 16.2 Characteristic concentrations.[a] and detection limits[a] of analytical techniques for representative elements in parts per 10^9. (Abbreviations and conditions are explained in the key.)

Element	Characteristic concentrations MAS	Detection limits DPP	Detection limits ASV	Characteristic concentrations AAS	Characteristic concentrations ETA–AAS	Detection limits ICP–ES	Detection limits ICP/MS	Detection limits NAA	Detection limits XRF
Ag	15	60	0.005	20	0.2	7	0.04	0.1	1300
Al	3	0.03	0.5	30	2	3	0.1	1	1600
As	15	1	0.5	100	2	60	0.4	1	300
Au	10	2	0.5	100	0.5	30	0.08	2	600
Ba	–	–	–	200	10	0.2	0.02	3	100
Br	20	0.03	–	–	–	–	–	0.2	600
Ca	20	15	–	20	1	0.02	5	200	100
Cd	15	0.5	0.005	15	0.05	2	0.07	10	200
Cl	–	0.3	–	–	–	–	–	2	200
Co	15	1	0.05	70	1	3	0.01	10	30
Cr	15	7	–	60	0.5	2	0.02	–	200
Cs	–	–	–	200	–	–	0.02	200	200
Cu	10	0.5	0.01	60	1	2	0.03	2	300
F	–	–	–	–	–	–	–	20	1000
Fe	20	1	0.1	60	1	7	0.2	–	300
Ge	3	–	–	1000	–	15	0.08	3	–
Hg	25	–	0.2	2200	100	60	0.08	3	300

I	–	0.03	–	–	–	–	0.01	0.2	100
In	15	0.5	0.01	300	10	60	0.01	0.1	–
K	–	–	–	20	0.5	60	–	10	30
La	60	–	–	–	–	3	0.01	1	200
Mg	15	15	–	3	0.03	0.2	0.1	10	600
Mn	1	3	–	30	0.2	1	0.04	0.05	100
Mo	30	6	–	300	5	3	0.08	60	600
Na	–	–	–	6	2	0.5	0.06	1	–
P	1	–	–	–	–	60	–	60	600
Pb	25	1	0.01	200	0.5	7	0.02		600
Pt	200	–	–	600	20	30	0.08	10	600
Rb	–	–	–	100	–	–	0.02	100	–
S	3	–	–	–	–	–	–	–	300
Si	10	–	–	2000	7	10	10	60	100
Sn	10	1	0.02	1000	0.4	30	0.03	200	20
Sr	–	–	–	100	1	0.1	0.02	60	300
Ti	10	3	–	100	–	2	0.06	60	100
Tl	20	3	0.01	200	3	200	0.05	–	–
U	30	3	–	–	1	30	0.02	1	100

(continued overleaf)

Tab. 16.2 (*continued*)

Element	Characteristic concentrations MAS	Detection limits DPP	Detection limits ASV	Characteristic concentrations AAS	Characteristic concentrations ETA–AAS	Detection limits ICP–ES	Detection limits ICP/MS	Detection limits NAA	Detection limits XRF
V	15	–	1	700	1	7	0.03	0.1	100
W	60	–	–	–	–	7	0.06	1	–

[a] As a result of historical practices, the detection limits and characteristic concentrations are calculated in a variety of ways depending on the assay method involved. The detection limits and characteristic concentrations apply to ideal samples. This means that the samples have no untoward matrix effects and no interferences. The values reported here are for samples directly assayed by the various methods without any preconcentration, which means without making the analyte more concentrated by, for example, evaporating solvent prior to the measurement. For realistic samples, detection limits are about 100 times higher than those noted here.

Key

A blank value (–) means either not a method for trace (less than ppm) analysis of the element or not reported.

MAS: Molecular absorption spectrometry in conjunction with colorimetric reagents. The values correspond to an absorbance of 0.025 for a 1-cm path length with a sample consumption of 1 ml.

DPP: Differential pulse polarography, assuming a sample consumption of 10 ml.

ASV: Anodic stripping voltammetry, with a sample size of 10 ml and a plating time of 10 minutes. Various media are used for electrodes. Ni and Co are determined by cathodic stripping of their oxides on Pt.

AAS: Atomic absorption spectrophotometry, with flame atomization and pneumatic nebulizer. The values correspond to an absorbance of 0.0025 with a 10-cm path and a 1-ml sample.

ETA–AAS: Electrothermal atomization–atomic absorption spectrometry. The values correspond to an absorbance of 0.025 with a 50-µl sample.

ICP–ES: Inductively coupled plasma–emission spectrophotometry. Values are for a pneumatic nebulizer.

ICP/MS: Inductively coupled plasma/mass spectrometry. This is with 10-seconds integration. All ions determined in the positive-ion mode.

NAA: Neutron activation analysis, with a sample volume of 100µl, thermal neutron flux of 10^{12} n cm^{-2} s^{-1}, and irradiation time of 10 hours followed by immediate counting.

XRF: X-ray fluorescence spectrometry, with a dispersive instrument for a sample of 1 ml over a surface area of 7–9 cm^2.

Sources: Commission on Microchemical Techniques and Trace Analysis (1982) and Morrison (1979). Manufacturer's literature for ICP/MS.

examples of analyses of the elements.) As a result, detection limits for more typical samples and conditions are factors of 10 to hundreds greater than those in the table. However, the values in the table apply more broadly. Reading across the table, the smaller the value of the detection limit generally means that the associated method will be more effective for that element at any concentration. Another term that is similar to detection limit is the characteristic concentration. More details are in the table notes.

16.2.7

Imaging, Hyphenated, and Multidimensional Techniques

In a simple color photograph, the picture can be considered to be a two-dimensional distribution (or map) of colors and their levels of brightness at each point or pixel (short for picture element). A similar map can be made for the response of a chemical measurement on the surface of a material. For example, the different responses of separate zinc and tin crystallites in an alloy could be mapped in two dimensions and a picture of the microscopic structural distributions recorded and analyzed. Another example would be to determine a full color analysis (a spectrum) for each pixel of the image of a material surface. Both these examples are called *spectral imaging*, and the latter, with the full spectra recorded, is also called *hyperspectral imaging*. Additionally, by recording a series of adjacent slices parallel to the surface, a three-dimensional map of the content can be constructed.

Hyphenated methods – also called *integrated* or *combination methods* – provide powerful tools by combining different analytical techniques. Examples are combining a separation of sample components (Section 16.7) by, for example, gas chromatography (GC) followed by identification and quantitation of each separated component using, for example, mass spectrometry (MS) (Section 16.5). The total process is called *gas chromatography–mass spectrometry* and (*GC–MS*) or *GCMS*.

A technique is multidimensional when a similar technique is used two or more times in sequence. In an example of separations, for instance, where different chemical components of a sample are separated in time, ‘‘slices’’ of this output are taken and each further separated with different conditions. Multidimensional methods differ from hyphenated ones in that the second level of analysis is similar in kind to the first. One example of two-dimensional chromatographies has samples that can physically be run on a square plate, first in one dimension and then perpendicular to that. Another example is two-dimensional GC, which is often abbreviated GC–GC or GC×GC, where each time slice of the first separation is relaunched into a second separation. A final example is MS, where molecules from samples are separated by their masses, and each of the masses produced can be chemically broken down by colliding it with inert gas atoms, and the fragments then analyzed by MS. This set of operations is called *two-dimensional MS*, or *two-dimensional* MS^2. In addition, the process can be repeated yet again on the second set of fragments, which results in the analysis of the fragments of the fragments: the overall measurement is labeled MS^3.

Finally, the multidimensional analysis can be part of a hyphenated or of an imaging method. A hyphenated method that has a multidimensional part includes GC–MS^2. Also, an image of organic molecules on a surface might be obtained by scanning the surface with a laser to desorb the molecules and observe the location-dependent material by MS or MS^2.

One common facet of imaging, hyphenated, and multidimensional methods is that they produce huge amounts of data. Fast computers and efficient algorithms are needed to process and make sense of this flood of data.

16.3
Assay Methods Based on Known Specific Chemical Reactions

The three types of assay methods described in this section – gravimetric, titration, and kinetic – are based on various specific chemical reactions that occur between a specific chemical form of the analyte and a chemical reagent that is added. In most cases, the reactions are done in a homogeneous solution, although some kinetic methods may be based on gas-phase processes. On the whole, these techniques are not inherently highly specific. An exception to the relative lack of specificity occurs when assaying for the total acid content of a solution by determining the quantity of hydrogen ion, H^+. This assay is relatively straightforward by titration.

16.3.1
Gravimetric Methods

Gravimetric analysis, as its name suggests, depends on a measurement of weight to determine the quantity of analyte in a sample. Because of the capabilities of weighing technology, the working precision of gravimetric analyses might be in the range of 1 part in 10 000 and, in special cases, 1 part in 100 000. However, the chemical purity and inhomogeneity of weighed materials limit the precisions of gravimetric analyses to 0.1–0.3%.

There are a number of different gravimetric assay methods. One of these is electrodeposition or electrogravimetry. This technique involves the electrochemical conversion of metal ions in solution into the metallic form which plates out onto an electrode. (Electrodes are electrical conductors such as platinum, gold, graphite, or electrically conducting polymers.) The amount of metal present is then found by the difference in weight of the electrode before and after the plating operation. Electrogravimetry is generally used to determine the quantity of a specific, known metal – for example, cadmium or zinc – in a macrosample. It can be highly precise (better than ±0.02%) and accurate. However, for routine determinations, when somewhat lower precision is needed (in the range ±0.2 to ±1%), the method has been supplanted significantly by spectroscopic methods such as atomic absorption (AA) (see Section 16.4.3.4).

A second gravimetric method involves measuring a sample's changes in mass upon heating at a controlled rate in a controlled atmosphere. This is called *thermogravimetry* (*TG*). Uses for TG include determining the water content of numerous types of samples and determining the carbon and ash content of coal. This technique is described further in Section 16.9.

A third technology for gravimetric determinations is to ‘‘weigh'' changes in material deposited on a vibrating surface. A piezoelectric material (one that deforms with the application of a voltage across it) is vibrated at an ultrasonic frequency which is a resonant vibration frequency of the system. (The resonant vibration frequency is the frequency at which the solid would vibrate freely if hit sharply.) Since the resonant frequency changes when mass is added or removed from the surface of the device, the measurement of the frequency change provides a measure of mass. A number of different methods are used to produce the vibrating surface, a surface acoustic wave (SAW). Among the materials used is a quartz crystal that distorts under the influence of a voltage across it, which is used in a quartz crystal microbalance (QCM) that can be used both in air and submerged in a liquid.

16.3.2
Titrations

For a titration assay, a volume of a standardized solution (one of known concentration of the reagent) is added to a solution containing an unknown amount of analyte in order to ascertain the quantity of analyte in the solution. When the standardized reagent solution is added to the analyte solution, the correct term to describe the process is a volumetric titrimetric assay. The chemistry of the reaction between the analyte and the titrant must be relatively specific and well understood.

All titrations require a technique to indicate the equivalence point – the point when the standardized titrant has just finished reacting with all of the analyte that was present. At the equivalence point, the volume of the titrant (the solution added) can be related to the quantity of the analyte. However, to discern the equivalence point some physical change must be observed. The center of that change is the end point of the titration. Ideally, the end point and the equivalence point coincide. The measurement of the end point may be done by any technique that is sensitive to the disappearance of the analyte and the changes in the solution that occur because of that fact.

A number of different types of reactions fulfill the requirements for titrations. These are as follows:

1. reactions of acids with bases;
2. reactions in which one species is oxidized (one or more electrons removed) and one reduced (one or more electrons added) – these are called *oxidation–reduction reactions*;
3. reactions in which a precipitate is formed from two or more soluble species;
4. reactions in which a coordination complex is formed (a coordination complex is a soluble species formed when two separate, soluble ions or molecules bind tightly together in solution); and
5. reactions in which tightly bound stable polymer pairs (the analyte polymer and the reactant polymer) or colloids are formed. Colloids are stable, small agglomerations of molecules that do not settle out of solution.

Titration is not a method used for trace analysis. In general, the lower limit for the use of titration for precise determinations is for solutions containing approximately one-tenth of a milligram of the analyte. The amount and precision of titrant addition are not a problem. Contemporary automatic titrators can add tens of microliters with volumes known to better than a microliter. Overall, imprecision in the range of ±0.1% is possible with moderate care using well-calibrated titrants.

16.3.3
Analyses Based on Chemical Kinetics

Kinetics is the area of chemistry in which the rates of reactions between reagents are studied. Kinetic methods of analysis are those in which analyte content is found by measuring the rate of a chemical reaction. The analyte can have one of the following two different chemical actions:

1. The analyte reacts with the reagent and is transformed in the reaction. This is called the *direct kinetic method*.
2. The analyte acts as a catalyst for a reaction. This is called the *catalytic method*. (A catalyst is a material that speeds the rate of chemical reaction between other reactants but is not itself used up.)

Experimentally, there is little difference between the direct and catalytic methods. The rates of reaction must be determined under a set of standard, reproducible conditions. The rates are determined by monitoring the change in the reagent or product concentrations over time. Often, spectroscopic methods are chosen to follow the reaction, for example the increase or decrease in concentration of a colored material. The time dependence is then related to the initial concentration of the analyte. The great majority of kinetic assays are run in solution. Usually good temperature control is required.

There are four possible reasons for developing kinetic assays.

1. Kinetic assays can be extraordinarily specific. Assays based on enzyme reactions are examples. Often, only a few species that are similar in chemical structure to the analyte will interfere.
2. Kinetic assays also can be quite sensitive and, thus, useful for trace analysis. For example, in a catalytic kinetic assay, the analyte as catalyst can produce many times its concentration in some product.
3. Kinetic assays can be used when the rate of a reaction is so slow that waiting for equilibrium to be reached would be impractical. The rate of reaction itself can be observed relatively quickly.
4. A useful assay reaction might be irreversible: equilibrium cannot be attained. Nevertheless, the rate of reaction can be observed to quantify the analyte.

Since all chemical and biochemical clinical and diagnostic methods that use enzymes as reagents are kinetic methods, the majority of all chemical and instrumental analyses are, in fact, carried out using kinetic methods since millions of clinical analyses are run every day.

16.4
Spectrometric Assay Methods

16.4.1
Fundamentals

The measurement of the absorption and emission of light by materials is called *spectrophotometry*. Spectrophotometry is usually shortened to spectroscopy or spectrometry. The latter term is used here. In spectrometry, the words absorption and emission have the same meaning that they have in everyday use. Absorption means to take up, and emission means to give off. In spectrometry, the word ‘‘light’’ is used as a general term for electromagnetic radiation (simply because it is less cumbersome than ‘‘electromagnetic radiation’’), and we use light as a generic term here.

Spectrometry can be used to assay both the elemental and molecular content of materials. The instruments that are used in spectrometry are referred to as *spectrophotometers* or *spectrometers*, the former associated more with light that is visible to our eyes. The term spectrometer is more common and is used here. The two fundamental variables that are measured in all spectrometric assays are the wavelength or frequency of the radiation and the amount of power at the wavelength. This is called the *spectrum* of the material in some range of wavelength.

Light energy will be absorbed (and/or emitted) by a specific chemical species in a sample only over certain wavelength ranges; at other wavelengths, the species will be transparent (or not emitting). The wavelength at which the energy is absorbed or emitted is a characteristic property of the specific chemical species and not the amount of material. This wavelength property gives spectrometry a measure of chemical specificity. In uncomplicated cases, changes in measured power – or, alternatively, ‘‘brightness’’ or light intensity – are proportional to changes in the amount of the species that interacts with (or produces) the light at the characteristic wavelength(s). As a result, measurement of the power at these characteristic wavelengths allows quantitative analysis of that analyte. This is the basis for all analytical spectrometry.

Because of the different spectra for different species, instead of separating analytes physically, the wavelengths of light associated with each may be measured separately, if the wavelengths do not overlap. Atomic emission spectrometry (AES) utilizes this property. In other cases, the spectra of components in a mixture do overlap. However, the total spectrum can often be separated into contributions from each species present, and the mixture components quantitated without separating them physically.

16.4.1.1 Energy in Spectrometry

An atom or molecule can take up energy from and give off energy to its surroundings in the following three forms: (i) as light; (ii) as heat; and (iii) as kinetic energy (the velocity) of any emitted particles such as electrons.

Different names are given to the spectrometries depending on the form of energy that impinges on the atoms or molecules and the form(s) that leave. These general names are listed in Table 16.3. The first two entries are emission and absorption. In the third, the luminescence can be divided into two types: fluorescence and phosphorescence. On an experimental basis, these can be separated by the differences in time over which emission takes place. When a phosphorescent material is illuminated, the emission may persist for an appreciable time after the illumination is removed. The fluorescence emission ceases virtually immediately. Phosphorescence is usually observed in the ultraviolet–visible range. Fluorescence is used analytically in the infrared, visible, ultraviolet, and X-ray regions. The last but one entry in Table 16.3 relates to electron emission. This is called the *photoelectric effect*, and is the basis for a number of electron spectroscopies, which are most important in the analysis of surfaces.

Tab. 16.3 The transformations of light energy interacting with atoms or molecules.

Energy in as	Energy out as	Spectrometry
Heat	Light	Emission
Light	Heat	Absorption
Light	Light	Phosphorescence, fluorescence
Light	Moving electrons	Photoelectron spectroscopies
Bonding energy	Light	Chemiluminescence

16.4.2
General Types of Spectrometry

Experimental spectrometry can be classified into three general types depending on the components of the spectrometers and their relative geometries. These three classifications are emission spectrometry (ES), absorption spectrometry, and fluorescence spectrometry.

16.4.2.1 Emission Spectrometry

A generalized representation of the equipment used in ES is shown in Figure 16.3. Emission from the sample itself provides the light, and the amount of light emitted at an analyte's characteristic wavelengths is a function of the analyte concentration. The illustration shows the light originating from a luminous flame, but gas discharges, such as are seen in a neon sign, can be used as well. In a few special cases, the energy is introduced from chemical-bond rearrangements; this is chemiluminescence, see the last entry in Table 16.3.

The transducer (a less general term is the detector) converts the light into a signal that we can relate to the concentration of the analyte. A common property of such transducers is that they respond to a range of wavelengths. For example, a transducer for the visible region responds to light of a wide range of colors – much wider than the emissions from specific analytes. Each analyte's emissions are produced at a set of relatively narrow ranges of wavelengths. To measure this set of emissions, each of these narrow ranges of wavelengths must be separated so they can be measured individually. For convenience, each narrow range of wavelength may be referred to as a *single wavelength*,that is, monochromatic, and the device used to isolate these single wavelengths of the spectrum is a monochromator. In ES, the monochromator is placed between the emitting sample and the transducer. Alternatively, after the wavelengths are spread out (dispersed), they can be measured simultaneously (see Figure 16.4).

16.4.2.2 Absorption Spectrometry

The second general spectral method involves measuring the amount of light at various wavelengths that pass through a sample. The geometry of the equipment is illustrated in Figure 16.4, where a separate radiant source generates light. Overall, the instrument measures the amount of light absorbed by the analyte at each wavelength. Both scanning and multichannel types are used.

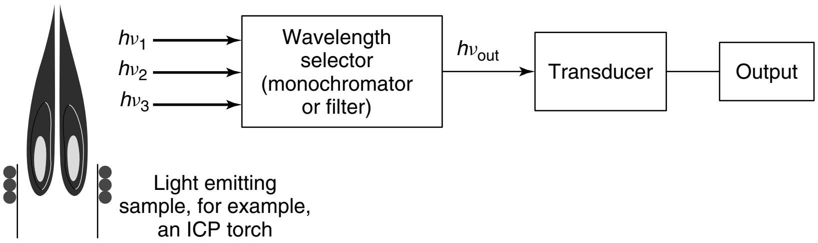
$h\nu_1$
$h\nu_2$
$h\nu_3$
Wavelength selector (monochromator or filter)
$h\nu_{out}$
Transducer
Output
Light emitting sample, for example, an ICP torch

Fig. 16.3 Generalized pictures of equipment designed to measure light emission. Shown is a scanning spectrometer. The light from the source is collected and passed through a monochromator, which only allows a specific wavelength to pass through. The transducer converts the light energy at that wavelength to an electrical signal, which is amplified and displayed as the instrument output. The wavelengths are scanned and associated intensities collected over time.

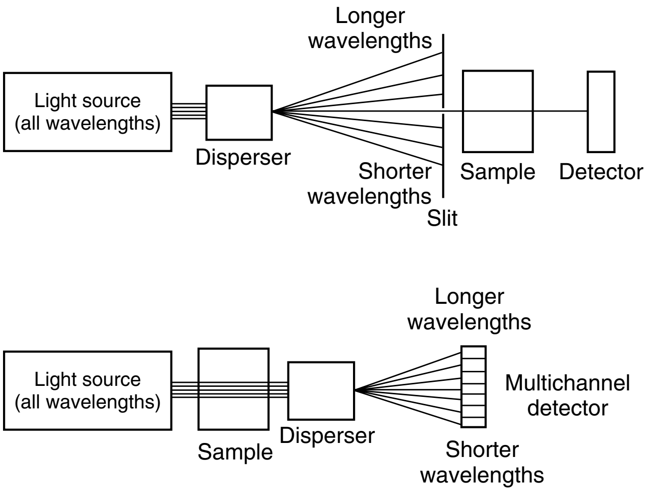

Fig. 16.4 Generalized picture of scanning instrument designed to measure sample absorption or transmittance. (a) A scanning spectrometer where the source light passes into a monochromator, which allows only a specific wavelength to pass through. The light passes through a sample that absorbs a part of the energy, and that which remains is converted to an electrical signal in the transducer. The monochromator is not needed if the light source is generated at a single wavelength such as a specific radiofrequency or a laser light emission. (b) A polychromator (multichannel) system where all the wavelengths are collected at once by a multichannel detector.

Most sources produce light with unwanted wavelengths in addition to the ones that are desired. (The exceptions to this general characteristic are radiofrequency sources and lasers.) Passing the light through either a monochromator or a filter selects the wavelength that is desired for the assay. The picture shows the monochromator between the source and the sample. However, in some instruments, the monochromator is placed between the sample and the transducer.

Absorbance and Concentration The amount of light absorbed by a sample can be reported as the absorbance of a sample. Experimentally, it can be shown to be directly proportional to, a constant that is a property of the material itself as well as the wavelength at which the measurement is being made; *b*, the length of the path through which the light travels in the sample; and *c*, the concentration of the material that absorbs the light. Algebraically, the relation is written as

$$A = \epsilon bc \tag{16.1a}$$

This equation describes Beer's law. When the concentration c is in (moles per liter), then ϵ will have the units (liters per mole centimeter), and is called the *molar extinction coefficient* or *molar absorptivity*. The value of b, the path length, must be expressed in units consistent with the units of ϵ. Labeling the units in Equation 16.1a, we get

$$A(\text{unitless}) = \epsilon(l(\text{mol}^{-1}\text{cm}^{-1}))b(\text{cm})c(\text{mol})^{-1} \tag{16.1b}$$

where l stands for liter, and a mole equals 6.022×10^{23} particles. A rough measure of the values of in the visible range with light passing through the sample's 1-cm

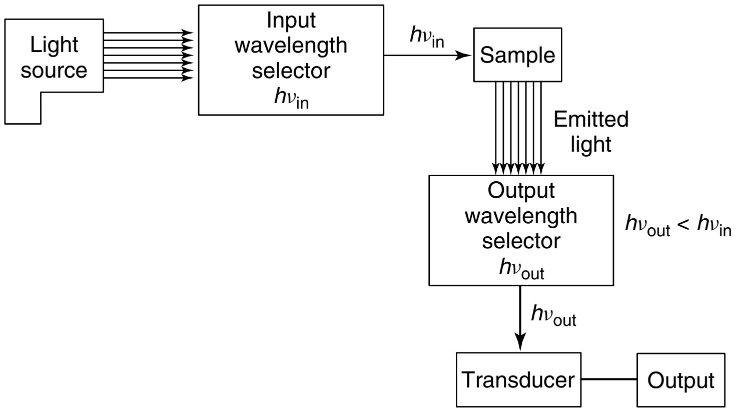

Fig. 16.5 Generalized picture of the equipment designed to measure sample luminescence in the wavelength regions from infrared through X-ray. The source light is collected and passed through a monochromator where a specific wavelength passes and impinges on the sample. The light emitted from the sample is collected and passed through a second monochromator set at a different wavelength. The light energy at that wavelength is converted to an electrical signal in the transducer. This is amplified and displayed as the instrument output. Various types of filters may be used in place of the monochromators.

path length are weakly colored, 10–100; moderately colored 100–10 000; intensely colored 10 000–200 000 in units (liters per mole centimeter).

16.4.2.3 **Fluorescence/Phosphorescence Spectrometry**

When light impinges on a sample and is absorbed, the energy can be reemitted in all directions as light of longer wavelengths. This emission of radiation is called *luminescence*. Measurement of the luminescence from analytes is done with the general instrument geometry shown in Figure 16.5. A light source for the appropriate energy region is required. The incident light is monochromatic after passing through the first monochromator. The light that comes from the sample is measured at an angle which is not in line with the source–sample axis. This angle is often 90°, but that is not the only angle possible. Figure 16.6 shows an example of both the spectrum of absorption and the spectrum of an analyte's reemission.

For chemical analysis, the division of luminescence into its two types – fluorescence and phosphorescence – is not too important. If the luminescence is present and is proportional to the analyte content, it can be used to quantify the substance. However, if a time-dependent spectrometric method is used, the light due to phosphorescence and that due to fluorescence may be separated. On this practical basis, phosphorescent reemission generally can last a much longer time after the light source no longer provides radiation to the phosphorescing sample. For both, measuring the quantity of analyte involves measuring the intensity of the emitted light at one or more wavelengths.

The organizing principle of the remainder of this section is simply the decreasing wavelength of the region of the spectrum that is used: γ-ray, X-ray, ultraviolet, visible, infrared, microwave, and radiofrequencies.

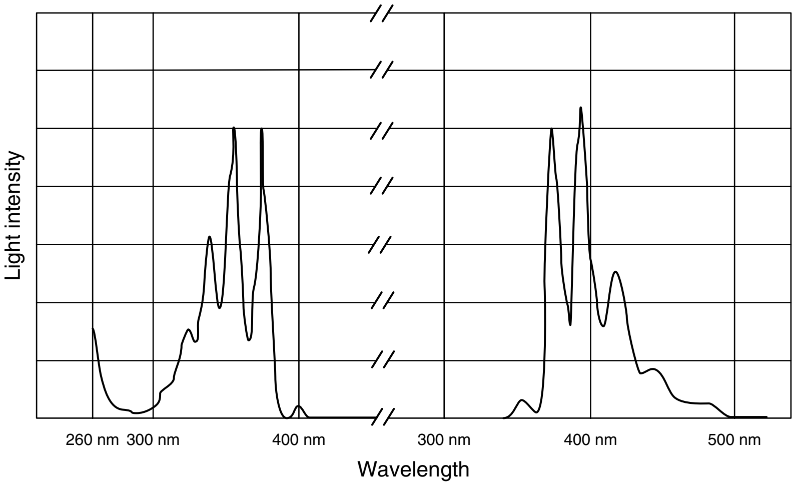

Fig. 16.6 Illustration of the absorption and reemission of radiation by molecules. The energy is absorbed with a wavelength dependence illustrated at the left. Absorbed energy is reemitted with a characteristic spectrum as shown on the right. The emission spectrum is at lower energy than the absorption spectrum; some energy is lost as heat. Sometimes light may be reemitted without changes in the spectrum, but merely with a decrease in intensity.

16.4.3
Specific Spectrometric Methods

16.4.3.1 Gamma Spectrometry: Neutron Activation Analysis

Unstable atomic nuclei that undergo spontaneous nuclear transformations with a concomitant emission of energy are called *radioactive*. Owing to the high energies of the electromagnetic radiation and the high kinetic energies (energy of motion) of the particles emitted upon nuclear decay, individual decay events are measured and counted. The emitted radiation, both electrons (e^-) and photons (the quantum of light), is dangerous physiologically and relatively difficult to block. As a result, analytical methods using the measurement of the radioactive decay have, for the most part, been developed for trace- or microanalyses.

Most radioactive isotopes used in chemical analyses are synthesized. This means that by an appropriate nuclear transformation, a nucleus is converted from a stable form to an unstable, radioactive form. Such transformations can be accomplished for a wide variety of elements by means of a reaction between a neutron and some other elementary particle with a stable atomic nucleus.

Sample irradiation with neutrons followed by measurement of the radioactivity characteristic of the atomic species present is called *neutron activation analysis* (*NAA*). The neutron source is a nuclear reactor in which the sample is placed. The radioactivity produced is characteristic in energy for each type of radioactive nucleus. A typical spectrum can be seen in Figure 16.7. The number of counts of each wavelength is used to quantitate the elemental contents of the material.

Virtually all NAAs are carried out by comparing the radioactivity induced in a sample with that of a standard. To make precise comparisons, the samples must have the same geometry. When these conditions are satisfied, the relative intensities

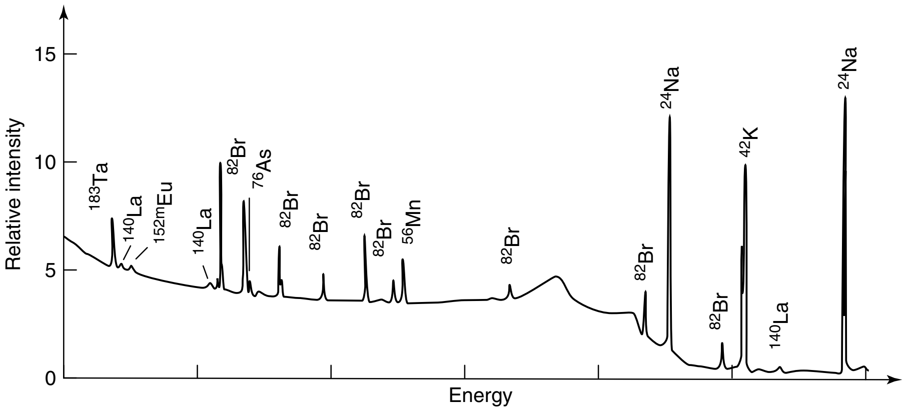

Fig. 16.7 Illustration of the γ-ray spectrum of a neutron-activated sample of tobacco leaf. A large number of different elements can be determined simultaneously in this manner. In this spectrum the lowest peak is 312.7 keV and the highest is 1732.9 keV. (Data from Ahmad *et al.*, 1979.)

of the characteristic radiation of the sample and standard are simply proportional to the total masses of the elements in each. In that way, an analysis utilizing γ-emissions is similar to any other ES.

An optimally sensitive and accurate NAA requires that isotopes be formed which have half-lives in a middle range of a few minutes to a few years. (The half-life is the time for half of the nuclei present to decay. For example, after two half-lives, a quarter of the original is left.) A list of the detection limits of NAA of some representative elements is presented in Table 16.2.

16.4.3.2 **Gamma Spectrometry: Mössbauer Spectrometry**

Mössbauer spectrometry is a γ-radiation spectrometry in which different, narrow ranges of wavelengths are characteristic of specific nuclear isotopes. The absorption and emission of radiation arise from spectroscopic levels of atomic nuclei and lie in the "hard X-ray/soft γ-ray" region of the spectrum. The precise wavelength depends on the chemical form of the atom.

A majority of elements of the periodic table have at least one isotope which is expected to be measurable by Mössbauer spectrometry. However, relatively few elements have been systematically studied and are, thus, useful in analysis. The isotopes that are used are listed in Table 16.4. One of the most commonly investigated elements is iron. A Mössbauer spectrum is illustrated in Figure 16.8.

Tab. 16.4 Most commonly studied Mossbauer isotopes and γ-emission energies).

Isotope	***E* (keV)**
^{57}Fe	14.39
^{99}Ru	90
^{119}Sn	23.88
^{121}Sb	37.15
^{125}Te	35.48
^{129}I	27.75
^{129}Xe	39.58
^{193}Ir	73.08
^{197}Au	77

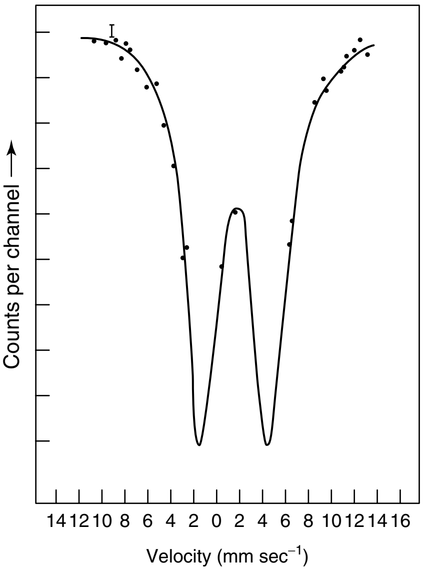

Fig. 16.8 An ^{57}Fe Mössbauer spectrum showing the two lines typical for iron. (This doublet – that is, two lines – results from a single line split by the charge distribution on the nucleus of the atom.)

Mössbauer spectra may be obtained in either absorption or emission, but usually are run in absorption. Emission spectra are usually used to investigate surfaces. The samples must be solids. The spectral energy levels are highly sensitive to the oxidation state of the atom and the chemical environment around the Mössbauer nucleus.

The absorption, which results from the Mössbauer effect, is quite small. Under normal experimental conditions, usually, only a few percent of the energy irradiating a concentrated solid sample is absorbed in the narrow-wavelength range. As a result, spectra are collected and averaged over times from minutes to days in order to see the absorption bands clearly among the noise.

Mössbauer spectrometry is used primarily for qualitative analysis of materials containing a known element in an unknown chemical form. Quantitation can be done, but its precision is subject to a number of assumptions. Only one isotope/element can be investigated at a time.

16.4.3.3 **X-ray Spectrometry: X-ray Fluorescence, X-ray Photoelectron Spectrometry, and Particle-Induced X-ray Emission**

Radiation in the X-ray region of the electromagnetic spectrum arises from transitions within the quantized energy levels of the most strongly held electrons of atoms. After an excitation, each element emits at characteristic wavelengths that can be used to identify the elements present in a sample (see Figure 16.9). The wavelengths of X-ray emission lines follow a regular pattern with the nuclear charge. The higher the nuclear charge is, the higher are the energies of the respective transitions.

After correcting for substantial background radiation, the intensities of the X-ray fluorescence are used to quantitate the amounts of the analytes that are present. In common with other luminescence methods, a number of elements may be determined during one experiment. X-ray fluorescence is used to determine the elemental composition of materials

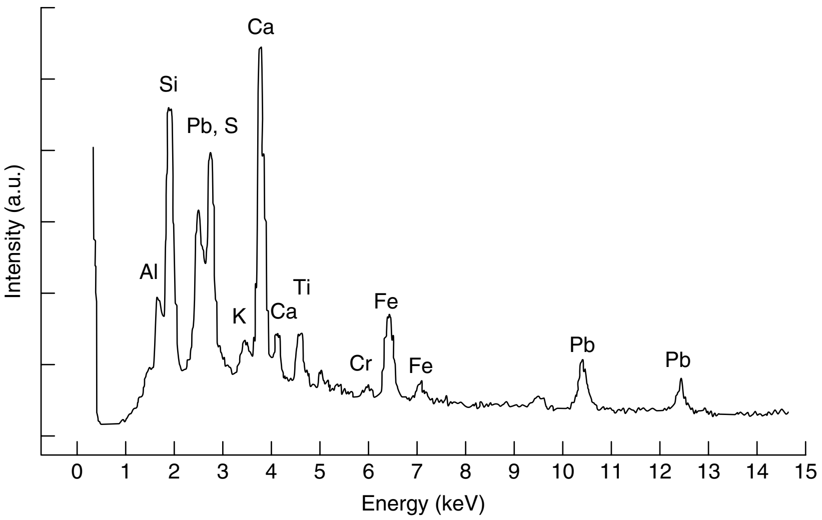

Fig. 16.9 Energy-dispersive X-ray spectrum of air particulates excited by X-rays generated by bombarding aluminum with electrons. Each point is the output summed over time (in counts of X-ray photons) in each of a number of channels of a computer memory. The peaks characteristic of the elements present and are labeled. The concentration of each element can be calibrated. (Redrawn from Wawros, Talik, and Pastuszka, 2003.)

since the X-ray spectra change little with changing chemical forms of elements. One of the method's main strengths is that it can be used for analysis of solids without sample dissolution.

The chemical elements which may be quantified by X-ray fluorescence are those heavier than neon, that is, sodium upward. Also, since almost all elements produce X-rays under the experimental conditions, the "contrast" is not especially high. As a result, X-ray fluorescence generally is limited to higher concentration levels – greater than about 0.1%, as can be seen from the detection limits listed in Table 16.2.

X-rays are relatively strongly absorbed by materials; on average, about 90% of the energy is absorbed within about 0.2 mm (see Table 16.7 in Section 16.8 below). Because of this strong absorption (and also because of possible surface contamination), samples must be prepared with reproducible physical properties.

This complicated spectral interference is one of the principal reasons that carefully prepared standards are important for high-precision (±1%) quantitative spectrometry. However, mathematical methods have been worked out to enable matrix corrections to be determined with fewer standards or to use some of the nonfluorescent, scattered radiation to calibrate the absorption coefficient. Contemporary instruments have both the hardware and software built in to do the corrections automatically after measuring the intensities of the peaks of all the elements present.

Another possible process that can occur when a material is irradiated with X-rays is the ejection of electrons from the atoms with energies characteristic for each element. These are called *photoelectrons,* and the ejection is called the *photoelectric effect.*

The depth from which these electrons can escape is only a few atomic diameters. As a result, these electrons can be used for surface analysis. The analytical technique is called *XPS* or *XPES*) and is described further in Section 16.8 on surface-sensitive techniques.

In contrast to this shallow probe, high-energy particles – with kinetic energies in the million electron volt range – penetrate into the bulk of a sample and cause the emission of both particles and X-ray fluorescence. The X-rays, since they can escape from the bulk, can be used to analyze the atomic species in the first half-millimeter or so of the material. This is particle-induced X-ray emission (PIXE). If the accelerated particles are protons, the technique is proton-induced X-ray emission, also called *PIXE*.

16.4.3.4 **Ultraviolet, Visible, and Near-Infrared Spectrometry**

Analytical methods using luminescence and absorption spectra in the ultraviolet through near-infrared regions (100–2000-nm wavelength), including the visible-light region, are among the most popular and broadly used in chemical analysis. Among the reasons for this popularity is that a large number of analyses of this type were developed before instrumental techniques were available, and the instruments allowed greater sensitivity, lower detection limits, and better precision. Experimentally atomic and molecular assays are often quite different. Atomic spectrometry is used for elemental analysis of analytes. An important condition of such an analysis is to break down the sample into its atomic components. On the other hand, an important condition of analytical molecular spectrometry is to leave the molecules intact.

Measurements are made of the relative amount of light absorbed (absorption) or emitted (emission) or the amount of light reemitted (fluorescence) by components of a sample. Essentially all elements can be assayed by AES or atomic absorption spectrometry (AAS). Also, essentially all molecules can be assayed by molecular absorption and/or fluorescence in the UV–visible or near-infrared spectral region. In this spectral range, atomic analysis is most commonly done on the atoms in the gas phase (usually in a flame or plasma) while molecular analytes tend to be dissolved in a liquid solvent.

Atomic Analysis Atomic spectra are strongly influenced by the atomic environment, and reproducibility in the spectra can best be obtained by producing isolated atoms (atomizing) in the gas phase. AA and atomic emission can be used for all levels of analyte from major to ultratrace by using a variety of methods to prepare and atomize the samples. Table 16.5 lists some sources such as flames, furnaces, plasma torches, and electric arcs that are used to heat and atomize nonvolatile samples.

AE or AES is inherently a multielement technique. The emission source must possess an extremely stable temperature to obtain the best precision. Also, hotter sources generally produce more complete atomization. As a result, in ES, a plasma source is now preferred to a flame source since plasma sources are both hotter than flames and more homogeneous in relative temperature. The major drawback to ES is difficulties with interferences. Nevertheless, quantitation of ±1% is possible with good control.

Unlike AES, AAS is inherently a method to assay a single element at a time since a light source producing a wavelength specific for one element is used. However,

Tab. 16.5 Sources of free atoms.

Method used to produce isolated atoms	Used in
Direct-current arc between two carbon rods in an inert atmosphere such as in argon	Emission
Radiofrequency-induced plasma in argon or helium (inductively coupled plasma)	Emission, fluorescence
Sample heated rapidly inside a hollow graphite tube (graphite furnace)	Absorption
Acetylene/air flame	Absorption, emission
Acetylene/oxygen flame	Emission

the offsetting benefit is that there are nearly no problems with interferences from other elements, which is a common problem in emission. Precision in AA assays better than ±1% is possible.

Molecular Analysis It is difficult to make generalizations about molecular analysis in this wavelength region because the spectroscopic properties of molecules vary so greatly. The basis for the analyses is Beer's law stated in Equation 16.1a. The value of A, the absorbance, can be measured below the 0.001 range with some contemporary instruments. Then, the lowest concentrations detectable depend on the molar absorptivity ε and the length of the light path (limited to about 10 cm for routine liquid analysis). With mirrors, the light's path for gases can be many meters by using multiple reflections through a relatively short cell. Precision in such absorption analyses depends more on the methodology than on the instruments. For instance, measurable variations in the amount of light passed can be seen with slight changes in angular position (order of a few degrees) of a sample. However, with rigid geometries or reproducible placement of samples, variations in results can be as good as ±0.1% for each sample.

Few generalizations can be made about molecular analyses with fluorescence except that the analytes must fluoresce. The practical limit of detection for highly fluorescent molecular species is in the picogram (10^{-12} g) range. That lower limit is practically routine since fluorescence from individual molecules can be measured, although such measurements are generally not done in the context of quantitating the fluorescent molecules.

For both absorption and fluorescence analyses, more than one molecular species can be analyzed in a mixture if standards for the analytes are available. In this case, spectral measurements are taken at a number of different wavelengths, and the concentrations of each of the analytes can be calculated from those data and the spectral properties of the standards. Fluorescence and absorption are two extremely popular methods for monitoring chromatographic separations (see Section 16.7) and quantifying the analytes separated chromatographically, and a majority of such detectors are multichannel systems that measure the entire spectrum continuously. Fluorescence is also used to monitor capillary electrophoresis (Section 16.7.4) as a part of DNA sequencing. Four different colors are used for the four different nucleotide bases.

16.4.3.5 **Vibrational Spectrometry: Infrared Spectrometry and Raman Spectrometry**

Infrared (IR) spectrometry and Raman spectrometry are used to probe the characteristic vibrations of molecular species in solids, liquids, and gases, as well as the vibrations of atoms and ions in crystals, glasses, and other condensed phases. The frequencies of vibrations of molecules correspond to the frequencies of infrared radiation. However, Raman spectrometry shows these vibrational frequencies as a set of lines separated from the illuminating laser line by those individual vibrational frequencies. (Experimentally, Raman spectrometry might more correctly be classified with visible spectrometries.) Infrared spectrometry for analysis is usually run as an absorption spectrometry (as Figure 16.4) while Raman is experimentally a fluorescence method (as Figure 16.5).

Vibrational spectra appear with numerous characteristic bands over the infrared wavelength range from 2.5 to 100 μm. (In chemistry, more commonly, the line energies are given in wave numbers, abbreviated as cm^{-1}, the inverse of the wavelength in centimeters. The infrared is from ~ 100 to $4000\,\mathrm{cm}^{-1}$.) Individual vibrational absorption bands can often be associated with specific chemical groups such as carbonyls, C=O, organic chlorides, –C–Cl, or ethers, –C–O–C–. As for all quantitative spectrometries, the wavelengths of the bands depend on properties of the molecules while the intensities can be used to determine concentrations. Quantitation to ±1% can be achieved.

Vibrational spectra are often complicated. Seldom can every band be assigned as originating from vibrations of specific groups in the molecule. However, an incomplete understanding of the spectra does not detract from their usefulness in both qualitative and quantitative analysis.

Since no two molecular species have the same vibrational spectrum – with every band matching in peak position (wave number), intensity, and widths of the absorption bands – infrared spectra can be used to identify molecular components of samples by matching the spectrum of an unknown with a library spectrum. (A library spectrum is one obtained from a pure species of known identity – from a standard.) An example of such spectrum matching is shown in Figure 16.10. The matching can be done by lining up spectra as shown in the figure or with computerized pattern-matching techniques.

Infrared spectrometry together with MS (and gas chromatography with mass spectrometry (GC/MS)) and nuclear magnetic resonance (NMR) form the basis for contemporary organic chemical qualitative analysis: identifying the molecular structure of unknown compounds and the compositions of mixtures.

Infrared spectra can be obtained on samples as small as a few micrograms although it is far easier to work with milligram samples. In this manner, infrared spectra are used in qualitative analysis of solids, liquids, and gases. In general, quantitation can be done relatively easily when individual bands are completely separated. With the best contemporary instruments, infrared spectra with precisions as good as for UV–visible spectra are obtained: easily ±1% and below. Raman spectral quantitation depends strongly on the properties of the sample and the stabilities of the laser

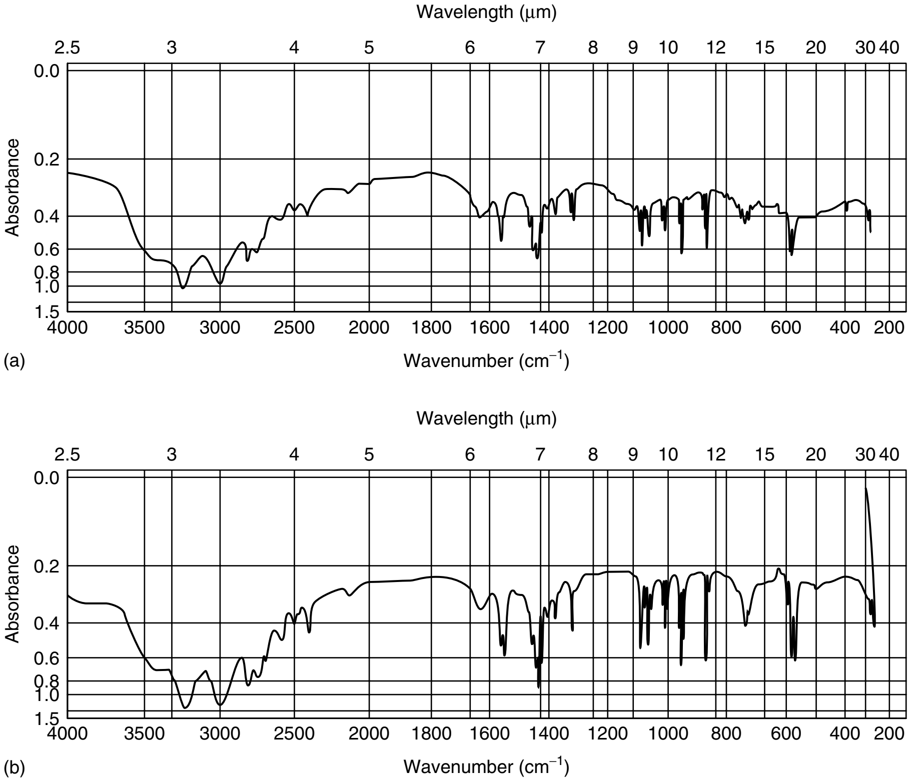

Fig. 16.10 Infrared spectra can be used to identify materials. The (a) is of some unidentified solids collected from a boiler. The (b) is of the tars formed on heating HCl salts of a corrosion inhibitor. There is one-to-one correspondence of all the features in the spectra. Since IR spectra are highly sensitive to structure, the identity of the two samples is quite certain.

light sources. As a result, precision can also be in the ±1% range.

Raman spectrometry tends, in general, to be limited to solids and relatively concentrated solutions. However, the intensity of the spectrum can be enhanced for the gas phase using a far more complicated experimental setup. This is the technique called *coherent anti-Stokes Raman spectrometry* (*CARS*), which is used to investigate, for example, the compositions of flames. Also, Raman spectral intensities can be enhanced by adsorbing to certain specially prepared metal surfaces, a method called *surface-enhanced Raman spectrometry* (*SERS*). Neither of these methods is as commonly used as simple Raman scattering.

16.4.3.6 Microwave Methods: Microwave Spectrometry, Electron Paramagnetic Resonance

Microwave Spectrometry In the microwave region of the spectrum, gas-phase molecules absorb and emit radiation due to changes in the energies of molecular rotations. The frequencies of the narrow spectral absorption lines are able to be measured with high accuracy, and can be used to identify the presence of numerous small molecules and the individual isotopic species such as $^{12}C{=}^{16}O$ and $^{13}C{=}^{16}O$. Rotational spectrometry is not a widespread analytical technique, but it has been used to identify a number of chemical species in astronomy.

Electron Paramagnetic Resonance (EPR) Spectrometry Electron paramagnetic resonance (EPR) spectrometry (or electron spin resonance (ESR) spectrometry) is an absorption spectrometry in the microwave region – 2–200 GHz – for which the sample is placed in a magnetic field. Experimentally, the frequency is fixed and the magnetic field is scanned. The magnetic field required is directly proportional to the microwave frequency. It is in the range around 0.3 T (3000 G) when the frequency is 10 GHz and 1 T when the frequency is 30 GHz. In EPR spectrometry, magnetic field positions are listed as "*g* values." The *g* values of a large number of compounds are near 2.0.

The primary requirement for a chemical entity to be measurable by EPR is that it contains a net electron spin. Such compounds include organic radicals (also called *free radicals*) such as $CH_3^{\bullet}$ and metal ions in specific oxidation states. As an example of the metals, cobalt as Co^{3+} has no EPR spectrum, but as Co^{2+} it does. As a result, EPR may be used to determine some oxidation states and quantities of a number of metal ions. However, not all paramagnetic metal ions exhibit EPR spectra. Some do so only in special chemical environments and the spectra of others may only appear at cryogenic temperatures. For analytical purposes, EPR spectra are run on solids and on liquid and frozen solutions.

A number of atomic constituents can be identified from the number of lines that appear in a characteristic spectrum. These lines are due to the interaction of the electron spin with the spin of nuclei in the molecule. Figure 16.11 illustrates the EPR spectrum of vanadium in its + 4 oxidation state.

As a result of the techniques used to increase instrument sensitivity, EPR spectra are presented as derivative spectra such as in Figure 16.11. The spectra are plots of the derivative of the energy absorbed

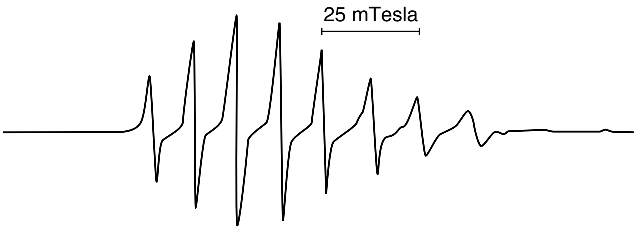

Fig. 16.11 Electron paramagnetic resonance spectrum of vanadium in its + 4 oxidation state in water solution. The eight individual lines result from the interaction of the electron with the nuclear spin. This is called *nuclear hyperfine interaction*. The spectrum is a mathematical derivative (the slope) of the absorption spectrum, and the different heights of the lines are due to the tumbling of the molecule.

versus a linear scale of magnetic field. The peaks (both positive and negative) of EPR spectra are at the points where a regular absorption spectrum has a maximum slope (either positive or negative, respectively). Precision in quantitation by EPR tends to be in the range ±5%. The detection limits are reached around 10^{12} spins for organic radicals (around 10^{-12} moles) and a factor of 10–100 higher for metal ions.

16.4.3.7 Radiofrequency Spectrometries: Dielectric Absorption, Nuclear Magnetic Resonance

Dielectric Absorption Absorption of radiofrequency electromagnetic energy by materials can occur through the interaction with the electric field or the magnetic field of the electromagnetic radiation. The former is the basis for dielectric absorption, and the latter is a property of NMR. The frequencies and magnitudes of dielectric absorption are used to characterize polymers and are the basis for some instrumental methods of moisture determination. However, in general, the absorption bands are not specific enough to be used alone to identify and quantify different molecular species.

Nuclear Magnetic Resonance (NMR) NMR spectrometry is one of the most powerful methods for identifying the chemical structures of organic – including biological and polymer – compounds. The reasons for this utility are primarily threefold. First, NMR spectra for nuclei – such as carbon (^{13}C) and proton (^{1}H) – are sensitive to an atom's chemical type, for example, $-CH_3$, $-CH_2-$, $=CH_2$, $-COH$ differ significantly for both the carbons and hydrogens. Second, the frequencies at which each of the chemically different nuclei appear in the spectra are characteristic for each. For example, hydrogens in methyl groups, $-CH_3$, appear consistently in one range of the spectra. And third, the chemical structures and magnetic resonance properties of atoms immediately adjacent to the resonant groups also influence the spectra. For example, the spectrum of a methyl group attached to a $-CH_2-$ group, that is, CH_3-CH_2-, differs in an obvious way from a methyl group attached to a $-CHCl-$ group although they both show up in the same ''methyl region.'' As a result, it is possible to ascertain molecular structures from NMR spectra alone. The spectrum of a compound does change its appearance somewhat with different magnitudes of the magnetic field. However, these changes are well understood, and usually present no problems in interpretation.

NMR spectra are most easily obtained on species in solution, but more advanced techniques allow useful spectra of solids to be measured.

Because of the sensitivity to small chemical differences, the spectra can be complex, with large numbers of lines overlapping. An example is shown in Figure 16.12. Simple NMR spectra can be used in quantitation, and internal standards are recommended. Multidimensional NMR is still qualitative in the chemical sense; it is used more for identification. The atoms most commonly investigated in chemical systems by NMR are hydrogen, carbon, phosphorus, and fluorine. The NMR phenomenon is limited to specific isotopes of the heavier elements: ^{13}C, ^{31}P, and ^{19}F.

Obtaining NMR spectra involves measuring the interaction of radiofrequency radiation with a sample material that resides in a strong magnetic field (see, *inter alia*, Nuclear Magnetic Resonance in the Encyclopedia of Applied Physics). The radiation used is in the 100-MHz to 1

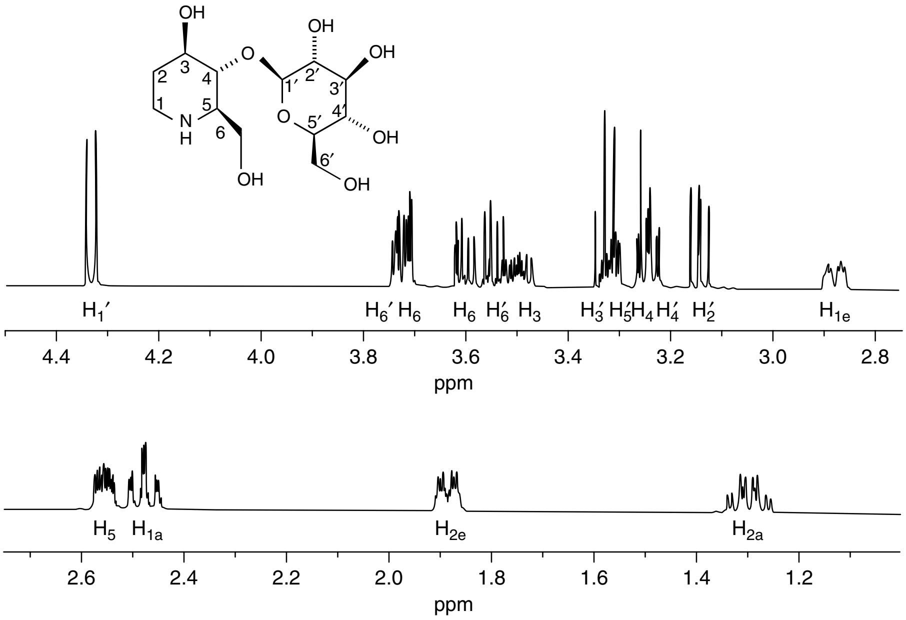

Fig. 16.12 A ^1H nuclear magnetic resonance spectrum of the molecule pictured. The assignments of the features in the spectrum to the various protons are noted below the spectra. The numbers refer to the carbon atoms to which the protons are bound. The numbers of the ring on the right are primed. By convention, the protons are not drawn but are understood to be attached. The bonds drawn as broadening lines indicate the bonds directed out of the plane of the paper. The dotted lines indicated bonds directed below the plane of the paper. The horizontal axis is numbered in ppm, which refers to parts per million of the radiofrequency used. (From Derome, 1987. Reproduced by permission.)

GHz range: the lower range is the frequency of FM radio in the United States and the higher end is near a common frequency for cell phones. The magnetic fields are large; the most sophisticated instruments use some of the largest constant, homogeneous magnetic fields that can be generated. Essentially all contemporary instruments collect the data as the sample's response to a train of high-power radiofrequency pulses that are a few microseconds long. The data obtained are mathematically transformed (a Fourier transform) to produce a spectrum such as shown in Figure 16.12. This method of data collection has the effect that all lines of the NMR spectrum are measured simultaneously. This manner of operation – so-called Fourier transform nuclear magnetic resonance (FT-NMR) – increases the sensitivity and improves the detection limit of the experiment. However, even with the best contemporary instruments, NMR is not really a trace method. To obtain proton spectra at least 10 μ g is required, and for carbon more than 1 mg of a species is needed. These limits can be lowered somewhat by using miniature coils surrounding a cell on the order of 1 μl in volume. The analyte must be relatively concentrated.

NMR of Solids To obtain optimum NMR spectra, the analyte molecules should be

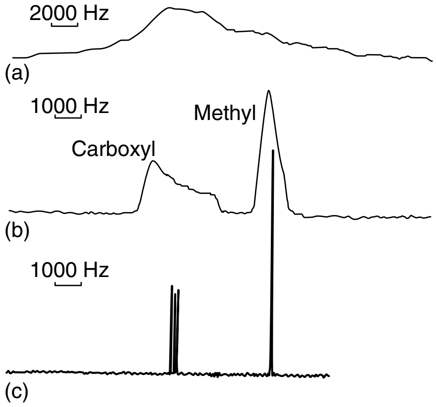

Fig. 16.13 Solid-state spectra of calcium acetate hemihydrate, $Ca(CH_3CO_2)_2 \cdot \frac{1}{2}H_2O$. (a) The solid-state spectrum; (b) the result when the various nuclear spin-nuclear spin interactions are removed by specific radiofrequency irradiation; and (c) when the sample is spun at the magic angle as well. (From Harris, 1983. Reproduced by permission.)

tumbling rapidly relative to the magnet. Part of the reason for this is to average out the interactions with neighboring molecules. In solids, neither the tumbling nor environmental averaging occurs within a sample, and the spectra tend to be broad and nearly featureless, with little chemical information (see Figure 16.13). However, moderate spectra can be obtained through two fundamental tricks. One is to spin the sample at a specific angle – 54°44′, called the *magic angle* – at a rate in excess of 200 KHz. This is called *magic-angle spinning* (*MAS*). The second is to average the magnetic interactions of the surroundings using strong radiofrequency fields at specific frequencies related to the resonance frequencies of the H nuclei present in the sample. This is the technique of proton decoupling. A spectrum is shown in Figure 16.13. A relatively large sample is needed. More commonly used instruments require tens of milligrams, while instruments with microcoils need a thousandfold less.

Two-Dimensional NMR A simple, one-dimensional NMR spectrum (such as shown in Figure 16.12) shows the magnitude of resonance (vertical) as it changes with a scale related to frequency (horizontal). The horizontal measure, ppm, is parts per million of the radiofrequency: 1 ppm equals 500 Hz for the 500-MHz instrument used to obtain the spectrum. A 2D-NMR spectrum, such as that shown in Figure 16.14, shows the resonance changes related to two frequency scales. In this case, both the horizontal scale (labeled on the bottom in ppm) and the vertical scale (labeled v_1) are identical to the ppm scale of the spectrum in Figure 16.12. The 2D spectrum shown is drawn as a contour plot. The usual 1D spectrum runs along the diagonal from lower left to upper right. The peaks that occur off the diagonal (which appear at the intersections of the vertical and horizontal positions of peaks on the diagonal) are called *cross peaks*. Where cross peaks occur, they show which of the peaks of the 1D spectrum are magnetically connected: their nuclei interact magnetically and are, thus, spatially close together. Two-dimensional NMR allows dissection of complex spectra into their component contributions, and because of the localized nature of the magnetic interactions, provides structural information. The results are qualitative in a chemical analysis sense.

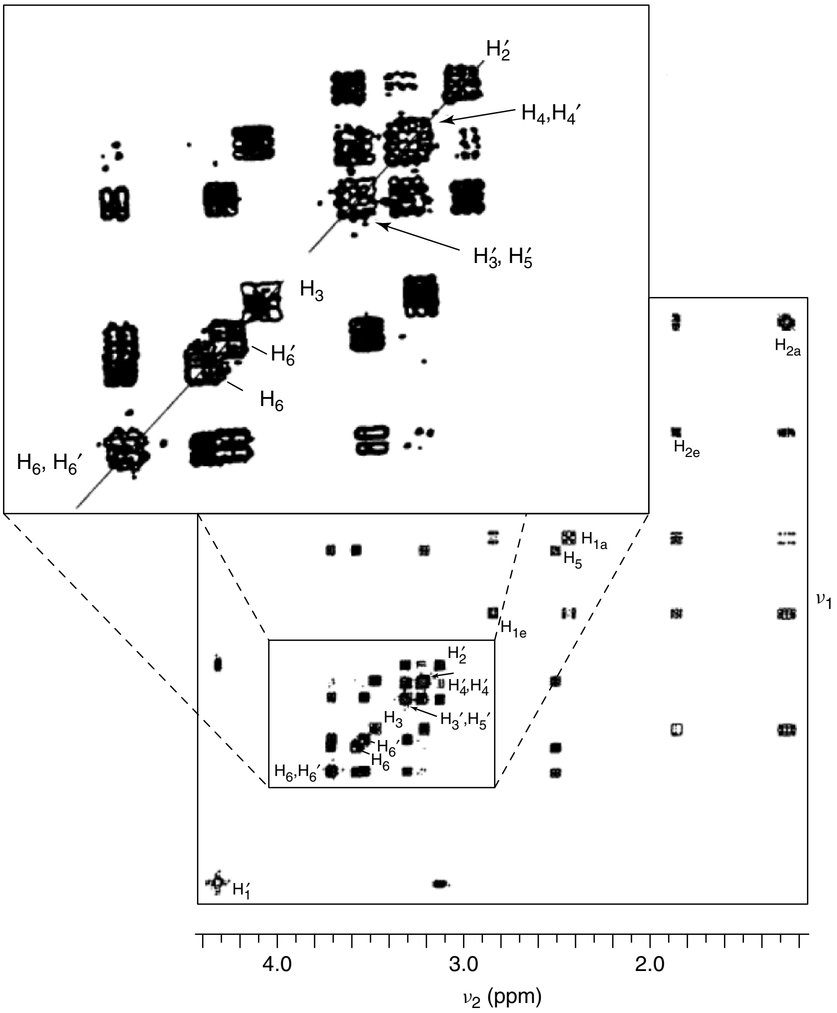

Fig. 16.14 The output from one type of 2D NMR as a contour plot. This is a spectrum of the same chemical as used for the 1D NMR in Figure 16.12. The 1D-NMR spectrum lies on the diagonal from lower left to upper right. The off-diagonal peaks lie on the vertical and horizontal locations of the two peaks that are magnetically connected. (From Derome, 1987. Reproduced by permission.)

Multidimensional NMR Other advanced methods – called *collectively multidimensional NMR* – involve three- and four-dimensional NMR spectroscopy. As for two-dimensional NMR, each added dimension involves a different range of frequencies and indicates how peaks of different nuclei are connected in space. Three- and four-dimensional NMR can be used to unravel far more complicated spectra to provide information that can be used to determine reliably the chemical structures of proteins up to molecular masses of about 30 000 Da. The methods require that the protein be partly enriched in ^{13}C and ^{15}N and require days of data collection on the most sophisticated instruments and extensive, intricate data analysis.

16.5
Mass Spectrometry

MS is the name for a collection of different techniques that are used to measure the relative numbers and the masses of ions that are in the gas phase. The methods of MS are among the most versatile and powerful tools of chemical analysis. This broad usefulness is due in part to having a wide choice of capabilities to transform atoms and molecules for analysis into gas-phase ions, and the sensitivity of ion detectors. Ions from hydrogen to as large as proteins can now be brought into the gas phase with varying amounts of excess energies, that is, energies beyond that needed to vaporize them. This energy causes the molecules to break up and/or ionize. Small amounts of excess energy – soft ionization – leaves the molecules relatively intact. Larger amounts cause fragmentation in amounts proportional to the excess. There are benefits to both having the molecules intact and causing fragmentation.

Once the ions are formed in the gas phase, their separation is determined by the ratio of the mass of the ion to its electric charge; this is written as m/z, the mass-to-charge ratio. A mass spectrum of the fragments of a small molecule is shown in Figure 16.15.

Analysis of organics for both qualitative (identification) and quantitative (the amount present), the method of choice is now the combination of GC (see Section 16.7) with a mass spectrometer as detector of separated species (GC/MS). The chromatography section of this instrument is used to separate the individual species of a sample, and the mass spectrometer allows identification from the pattern of the mass spectrum. Computerized methods of pattern identification are used. Similar mass spectral techniques are employed with liquid chromatography (LC) when separating, identifying, and quantifying less-volatile compounds. MS can be used to quantitate nanogram amounts of

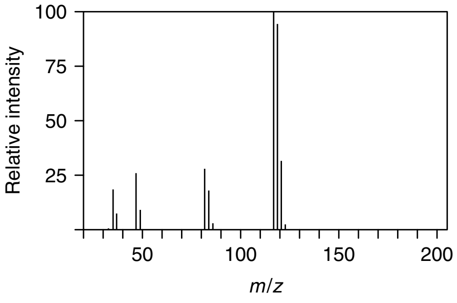

Fig. 16.15 A mass spectrum of carbon tetrachloride, CCl_4. The four peaks at the highest m/z are all the isotopic forms of CCl_3^+ :^{12} $C^{35}Cl_3$,^{12} $C^{37}Cl_3$,^{12} $C^{37}Cl_3$,^{12} $C^{37}Cl^{35}Cl_2$, $^{12}C^{37}Cl_2^{35}Cl$,^{12} $C^{37}Cl^{35}Cl_2$, with a small fraction contributed by the ^{13}C set as well.

molecules and atoms. With internal standards, precisions of ±0.5% are achievable.

16.5.1
Mass Spectrometers

Mass spectrometers, regardless of the design of the various parts, are constructed to separate ions of gas-phase molecules and atoms according to their masses. All mass spectrometers are constructed to carry out four operations as the following:

1. introduction of the sample as a gas;
2. ionization of the molecules in the gas;
3. separation by mass; and
4. detection of the separated species.

All the operations are done under high vacuum – generally around 10^{-9} Pa (1 Pa = 0.0075 torr = 0.0075 mm Hg.) The first two steps occur in the ion source (see Figure 16.16). (Some techniques of sample introduction inherently produce ionized molecules, which combines steps 1 and 2.) The third step is done by mass analyzers. The fourth step involves a transducer, usually an electron multiplier.

Bringing the atoms and molecules into the gas phase can be done in numerous ways; heating is the simplest. Others involve impact on the material with other particles, such as argon atoms, which dislodges molecules from the surrounding matrix of a sample; this is fast atom bombardment (FAB). If ions are used to dislodge the analyte, the ions expelled are called *secondary ions*, and the technique is called *secondary-ion mass spectrometry* (*SIMS*) (see Section 16.8.2). Ions of an analyte may also be formed by using a fast pulse of an ultraviolet laser to ablate a light-absorbing, low molecular mass matrix holding the analyte molecules. The gas-phase ions then pass into mass spectrometer. This is matrix desorption/ionization mass spectrometry (MALDI), which can bring polymers and proteins into the gas phase for analysis.

The methods for vaporization/ionization noted above require that the sample be placed into a high-vacuum chamber. A number of sources start with the samples at atmospheric pressure where they are brought into the gas phase and then pass

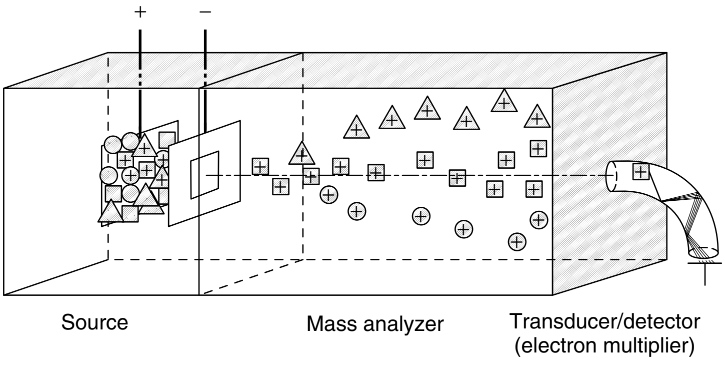

Fig. 16.16 Diagram of the major features of a mass spectrometer. The sample molecules travel from the left, where they are introduced, to the right. Shown in the source are three types of ions with different masses indicated by circles, triangles, and squares. They are separated in the mass analyzer, which may be a magnetic sector, a quadrupole mass filter, time-of-flight delay path, one of many forms of ion traps, or a small, highly precise cyclotron (FT-MS). The arrival of the ions is detected after the separation; here the detector is shown at the right.

into the vacuum for separation through a series of orifices with progressively lower pressures. The general method is called *atmospheric pressure ionization* (*API*). The most commonly used member of these atmospheric sources has a solution containing the sample pass through a hot (temperatures up to 10 000 K) plasma torch where the sample is atomized (broken into atoms) and the atoms ionized. Coupling the inductively coupled plasma with mass spectrometry (ICP/MS) is an extremely powerful method used for trace and ultratrace elemental analyses where nanograms per milliliter of most elements are easily quantitated (see Table 16.2)

In another external sample ionizer, a solution containing analytes can be ejected as an aerosol through a small, highly charged needle into warm air. The electric charges on the droplets cause the aerosol to break up into smaller and smaller droplets eventually yielding gas-phase ions such as proteins, which are drawn into the spectrometer. This source is called an *electrospray* and the general method is called *electrospray ionization* (*ESI*). The electrospray itself can be used as a solvent source to sample surfaces from which the solvent together with the analytes it picks up from the surface are drawn into the mass spectrometer. In that form it is called *desorption electrospray ionization* (*DESI*). A "dry" method of atmospheric ionization for small, relatively stable organics causes the molecules to be vaporized and ionized directly from a surface by placing the sample into a beam of excited helium atoms in front of the pickup orifice. An example of such an analysis is cocaine from paper money. The method is called *direct analysis in real time* (*DART*). These methods are only semiquantitative.

If ions are not produced by the above processes themselves, the molecules can be ionized in a number of different ways. One of the two most common methods is hitting the gas-phase molecules with accelerated electrons: this is electron-impact (EI) ionization. In the second method, chemical ionization (CI), a reagent gas such as methane (CH_4) is ionized by the accelerated electrons and, through subsequent reactions, forms a molecule ion, CH_5^+. The CH_5^+ then donates a proton, H^+, to the analyte to form a molecular ion with a mass one unit higher than before the protonation. (The extra mass mass unit is the mass of the added proton.) A molecule tends to break down into molecular fragments less with CI than with EI.

16.5.2
Multidimensional Mass Spectrometry: MS^n

It is somewhat like a picture puzzle to ascertain the identity of a molecule from the mass spectrum alone, especially if some material is not well purified. For instance, the fragments CO and C_2H_4 both have nominal mass of 28. However, if these fragments are further broken down, their identities become more certain; CO breaks into C and O while C_2H_4 produces CH_3 and CH fragments, among others. This analytical process can be carried out with a tandem mass spectrometer, such as is illustrated diagrammatically in Figure 16.17. In effect, a second mass spectrum is run on the breakdown products of a single peak of the first mass spectrum. MS/MS is an immensely powerful technique for identifying molecular species.

Such processes can be continued: measurement/breakup/measurement/breakup and so forth. This repetitive process is abbreviated MS^n as in MS–MS–MS, and so on, and can be observed with mass spectrometers based on various ion trap

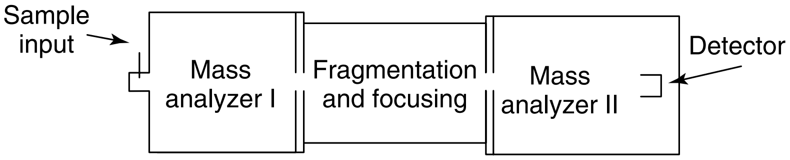

Fig. 16.17 A block diagram of one type of tandem mass spectrometer. The ions travel from the source through the first mass analyzer. Those that pass through this first mass spectrometer enter a region filled with a gas and surrounded by ion-focusing structures. The fast moving ions collide with the gas molecules and fragment. The ion fragments pass into the second mass analyzer, where they are separated by mass and detected.

designs. The ions are held in a vacuum in a small space inside the ion trap where by radiofrequency fields of specific geometries and frequencies, they are both held and manipulated to break up and then ejected to a detector.

16.6
Electrochemical Methods

In this section, some of the analytical instrumental methods used to measure electrical currents, charge, and/or electric potentials that originate from chemical reactions are presented. The analytes determined with electrochemical techniques must be contained in solutions containing unreacting ions called *spectator ions.* Many atomic and molecular species can be assayed from major to ultratrace levels. The subsections below deal with electrochemical methods in the following order: potentiometry, conductimetry, coulometry, amperometry, and voltammetry.

Briefly, electric potential differences (voltages, potentials) can arise in chemical systems due to differences in concentrations of ions and due to differences in the abilities of different species to give off or take up electrons (oxidation and reduction, respectively). The techniques based on measurements of these potentials when minimal current is allowed to flow are those of potentiometry.

When the content is quantified by measuring the changes in the electrical resistance of solutions of ions (electrolyte solutions), the techniques are those of conductimetry.

A potential can be applied to an electrode in a solution causing electrons to be exchanged with electroactive analytes (able to be oxidized or reduced) in the solution. When the analyte loses an electron it is said to be oxidized; when the analyte gains an electron, it is said to be reduced. We can relate the magnitude of the current and the time it flows to the amount of analyte present. The product current × time is the total charge transferred. The amount of the charge transferred for oxidation or reduction is the basis of a number of different techniques called *coulometric methods.*

One group of techniques combines the simultaneous measurement of a potential and current as they change with time. These are called the *voltammetric methods.* They are, perhaps, experimentally the most elegant and yet simple methods for trace and ultratrace analyses.

A straightforward electrochemical technique that does not fit well into the section's organization is electrogravimetry. It was discussed with other gravimetric methods in Section 16.3.1.

16.6.1
Potentiometry

The concentrations of analytes that are ions or change ion concentrations can be measured in solution with potentiometric methods, that is, those where the electric potentials of chemical reactions (their voltages) are obtained. The concentrations are related to the voltage logarithmically; for example, the voltage change for a 100-fold increase in concentration is twice the voltage change for a 10-fold increase in concentration.

Such a measurement requires a potential to which the measured one can be referred – a reference potential. Common references are the saturated calomel electrode (SCE) and the silver–silver chloride electrode. It is difficult to measure a cell potential to better than about $\pm 1\,\text{mV} = \pm 0.001$ V, which represents an uncertainty in concentration greater than 4%. This is the limit of precision for potentiometry. If more precision is needed, potentiometry can be used to detect end points of titrations (Section 16.3.2). The electrodes used for such measurements are designed to respond to one or a few ions specifically. The generic name for these devices is ion-selective electrodes. A well-known example is the pH electrode, which responds to the amount of H^+ in the solution. The operation of this electrode depends on the response of a surface that binds specifically to the analyte ion; this is an ion-selective surface. The oxidation state of the H^+ remains unchanged by the measurement. For potentiometric measurements, the current flow is kept as small as possible. For example, a common pH meter will measure the potential when the current flowing is on the order of 100 pA.

To find the analyte concentration, the voltage in the presence of the analyte must be calibrated with a solution containing a standard, known concentration of the same analyte. There are two ways to do this. First, the concentration is found by comparing the voltages from the standard and the analyte and recognizing the logarithmic relationship to find the analyte concentration. Second, we can extrapolate from a set of voltages measured after adding a series of known amounts of the analyte into the original solution after its potential has been measured. This latter technique is known as *standard addition* or, alternatively, *spiking*. Both these techniques can also be used for all electrochemical and spectroscopic analyses.

16.6.2
Conductimetry

Experimental values of conductivities (the amount of current an ionic solution can carry when a voltage is applied) almost always are found using rigidly fixed electrodes placed in the analyte solution. One type of conductivity cell is illustrated in Figure 16.18. Quantitation using conductance measurements can be done simply by calibrating the instrument with known concentrations of the ion-forming analyte and then measuring the conductivities of the solutions to be analyzed. The use of conductimetry is limited to analytes that are ions in solution. Conductimetry is used both to measure multiple species together, such as for monitoring contamination in streams and rivers, and for a single species such as continuous monitoring of an acid concentration.

Quantitative conductivities can be determined to about $\pm 1-2\%$ if the resistance between the electrodes is in the range of 1000Ω and the temperature is thermostated to within $\pm 0.1^\circ$. With proper standards and temperature control, special

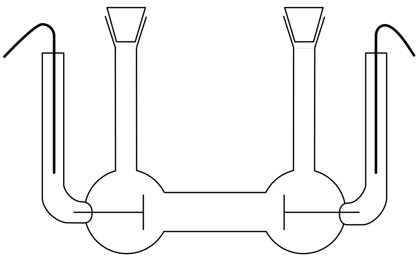

Fig. 16.18 Illustration of a conductivity cell. Usually the electrodes are platinum foil covered with a finely divided platinum on the surface. From its color, it is called *platinum black*, and it provides a large surface area – much larger than the dimension of the electrode. Samples fill the cells so that the geometry is the same for all runs.

high-precision conductimetry can be done with errors in the range of 0.1%. However, such precision is not routine.

16.6.3
Methods that Use Oxidation and Reduction

The remainder of the electrochemical methods discussed here depends on oxidation and/or reduction of the analytes with measurement of the amount of electric current that flows during the process. Either the current that flows or the total charge required for the oxidation/reduction process is measured. They are directly proportional to the analyte content. The types of materials that can be assayed include both elements and molecular species. When molecules contain organic groups that can be reduced electrochemically, they can be assayed by these methods. A number of electroactive groups are listed in Table 16.6.

16.6.3.1 **Coulometry**

A coulometric assay relies on measuring the total charge that is needed to complete an electrochemical oxidation or reduction. This total charge is relatable directly to the total amount of material electrolyzed. It is relatively easy to measure a total charge of millicoulombs over periods of minutes. A general calculation shows this to be representing about 1μ g of analyte.

The apparatus that is used is designed to keep a constant potential at the electrode where the analytical reaction (oxidation or reduction) occurs. The apparatus is a general piece of electrochemical equipment called a *potentiostat*, and it can also be used for electrogravimetry, coulometry, and the various forms of amperometry.

Coulometric assays are carried out in stirred solutions; at one time or another, all the analyte must come in contact with the electrode, and stirring hastens the process. An alternative to stirring is to use only a

Tab. 16.6 Bond types that can be reduced for electrochemical measurements.

C–C	C–N	C–O	C–S	C–X[a]
C=C	C=N	C=O	C=S	
C≡C	C≡N			
	N–N	N–O	N–S	
	N=N	N=O		
		O–O	O–S	O–X
			S–S	S–X
Condensed benzenoid rings				
Some heterocyclic rings				

[a] X = halogen

Source: *Chem. Eng. News* (1968), March 18, p. 96.

thin layer of solution so that the inherent diffusion of analyte is sufficient to result in a uniform concentration throughout the solution volume within a few minutes without stirring.

16.6.3.2 **Amperometry and Voltammetry**

Amperometric analyses are achieved by measuring the electrical current that passes into a solution and causes oxidation or reduction of the analytes at a fixed potential. The assay depends on quantitative measurements of the limiting current, i_L, which is proportional to the concentration of the species being oxidized or reduced at the electrode (see Figure 16.19). A relatively powerful but specialized use for fixed-potential amperometry is in LC. An electrode is set at a potential that will cause oxidation or reduction of the analytes, and the electrode is placed so that the analytes pass over it after they are separated on an LC column. The resulting increase and decrease in current as a band of material passes by can be used to quantify the amounts of each separated analyte. This is called *liquid chromatography with electrochemical detection* (*LC/EC*). Nanogram amounts of electroactive compounds can be determined inexpensively compared to other methods with comparable limits of detection.

A second, general set of amperometric methods involves measuring the changes in the current (as amperes or amps) as the voltage is changed; hence the name *voltammetry*. These voltammetric methods include polarography (a voltammetric method using mercury as its electrode), differential pulse voltammetry, and anodic stripping voltammetry (ASV). Figure 16.19 shows representative data of the current for an electroactive species as the applied potential varies from zero to more negative potentials. At first, little current flows since the analyte does not react in this voltage range. Then, the current increases relatively rapidly with voltage, and finally a plateau is reached where there is little further increase in current over the projected baseline as the potential is changed further. This is the limiting current, i_L. The curve shown in Figure 16.19 results from

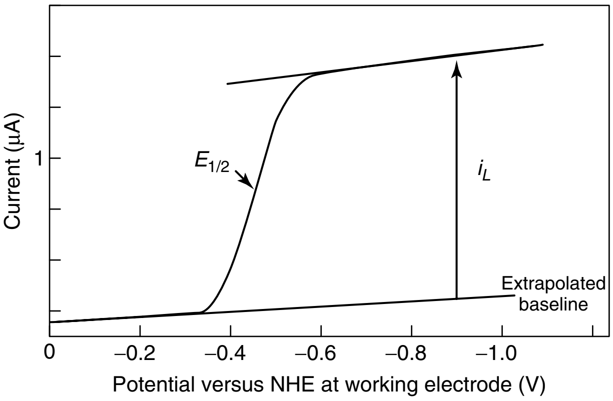

Fig. 16.19 Illustration of the output of one type of voltammetric experiment. The current flow is plotted versus electrode potential. At the left, only a background current is seen. Then, the electroactive species begins to react, finally reaching the limiting current. The midpoint voltage is characteristic of the analyte. The limiting current after the subtraction of the projected background is proportional to its concentration.

the simplest case of voltammetry, where the potential at the electrode changes with a constant increase or decrease with time. This is a voltage scan.

An increase in the flow of current can be measured more precisely by a careful combination of pulsing the potential together with selective data acquisition. The background current and the current produced when electrons are transferred after a voltage jumps are collected at specific times before and after the voltage jumps, superimposed on a more slowly changing voltage. One method that works this way is differential pulse voltammetry. (With a small mercury drop as the electrode, it is called *differential pulse polarography* (*DPP*)). Results of such determinations appear as seen in Figure 16.20. By comparing the area under the curve or just its height with the results from standards, one obtains the analyte content. DPP is a method suitable for trace analysis of both atomic and molecular species. The detection limits for various elements appear in Table 16.2.

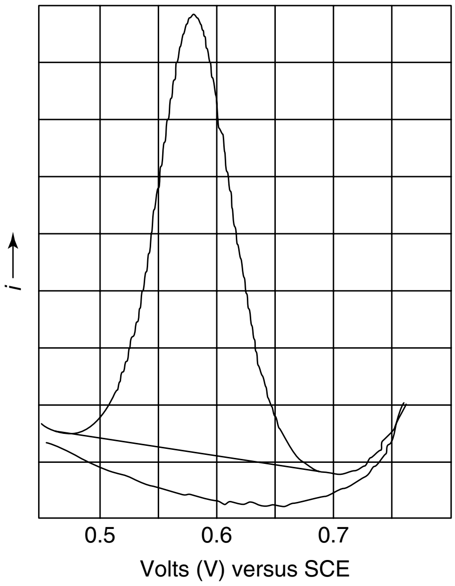

Fig. 16.20 Illustration of a differential pulse voltammogram and, underneath, the background current. The scale is about 5 nA per division. The shape of the curve is the slope of a curve such as seen in Figure 16.19. (From Rubinson and Rubinson, 2000.)

16.6.3.3 Anodic (and Cathodic) Stripping Voltammetry

ASV is a technique that can be used to assay a number of metals and nonmetals at the trace and ultratrace levels. The technique contains a step that first concentrates the analyte, which allows far lower detection limits for many metals than without such a step (compare DPP and ASV in Table 16.2). For example, for metals, the ions are first reduced to their metallic forms, which dissolve in the small mercury drop electrode (the mercury/metal solution is called an *amalgam*). The concentrated, dissolved metals are subsequently reoxidized, and the resulting current provides a measure of their concentrations. Differential pulse techniques give the best detection limits. An example of the type of data obtained is shown in Figure 16.21. When the analyte is first oxidized to concentrate it and then subsequently reduced for the assay, the technique is called *cathodic stripping*; both are used primarily for elemental analysis.

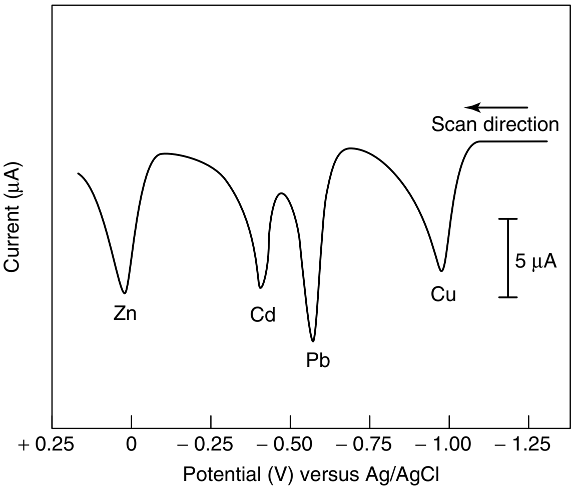

Fig. 16.21 An anodic stripping voltammogram of a solution 2.5 ppm in copper and zinc and 5 ppm in lead and cadmium. The sample was reduced and concentrated as an amalgam and then scanned as shown. The peak areas are proportional to the concentrations. (From Rubinson and Rubinson, 2000.)

16.7
Separations and Chromatography

To isolate quantities of pure substances involves the science and art of separations. Separation methods are based on the differing chemical properties of the components of a sample. In contemporary chemical analysis, separation is synonymous with chromatography. Figure 16.22 diagrammatically shows one type of chromatographic instrument in which the separation occurs within a hollow tube – the chromatographic column – containing the adsorbent or stationary phase that has a flowing fluid in contact with it. The stationary phase in column chromatography consists of either the column surface itself or particles of a solid material to which the analyte molecules bind. The analyte is placed at the beginning of the column and washed through with the liquid eluent. The individual components elute at the other end at earlier or later times depending on their differences in binding to the stationary phase. The analytes that spend more time on the stationary phase elute later.

The reasons for the wide applicability of chromatographic methods involve the variety of conditions that can be used to separate components of materials. For instance, we can use different mobile phases: gases, liquids, and supercritical fluids for LC, GC, and supercritical phase chromatography, respectively. Further flexibility arises from having many types of stationary phases (minerals and numerous kinds of polymers alone or with surfaces modified with molecules having varying properties). In addition, the geometries of the stationary phases can differ (inside columns of various sizes, on plates) or the scale of the separation can vary (large for preparative, smaller for analytical).

In contemporary analysis, the components that have been separated are detected as they elute from the stationary phase by measuring changes in any of a number of different chemical or physical properties. Virtually every instrumental and wet analytical chemical method has been used to detect and identify the components separated chromatographically. Four examples among these are electrical conduction, light absorption, ion production in flames, and atomic or molecular mass. The response of the detector with time as an eluted band passes – that is, a rise and fall (or

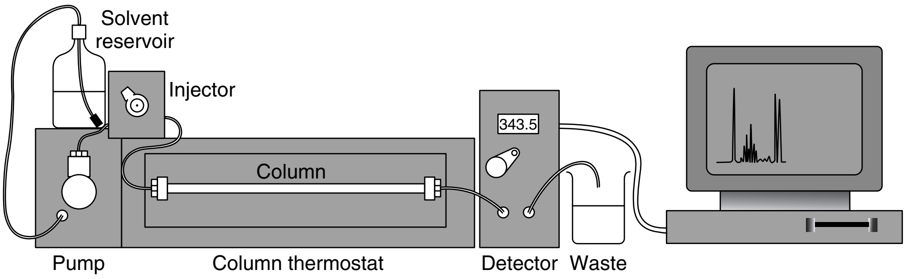

Fig. 16.22 Diagram of the components of a column liquid chromatograph. The sample of known volume is injected onto the column where the components are separated as they flow at a rate fixed by a pump. They elute and are detected, where a plot versus time is the chromatogram, as seen on the screen. Another chromatogram is shown in Figure 16.23.

fall and rise) in the property from the baseline level – produces a series of peaks as shown in Figure 16.23. This chart of detector response versus time is called a *chromatogram*.

16.7.1
Liquid Chromatography

16.7.1.1 Column Chromatography

Within that broad definition of LC are a myriad of techniques, and many of them are associated with more than one name. The names focus on many different aspects of LC methods. For instance, one of the primary classification schemes of LC is by the physical shape of the solid support such as column chromatography, thin-layer chromatography (TLC) (with the solid support forming a layer on a plate), capillary LC (with the inner surface of the capillary as the solid support), and flat-bed chromatography. Classification is also based on the efficiency of the separations such as high-performance liquid chromatography (HPLC) or high-performance thin-layer chromatography (HPTLC). Sometimes names of LC methods identify solutes which are separated and detected such as ion chromatography (done with column methods). One of the reasons that LC has become ubiquitous in analytical laboratories is the development of HPLC. The solid phase is composed of carefully packed particles that are between 10 and 2 μm in diameter. The size of the particles makes it necessary to pump the mobile phase through the column at pressures up to thousands of pounds per square inch (the commonly used, nonmetric pressure measurement, $1\,\text{lbf in.}^{-2} \approx 6895\,\text{Pa}$). HPLC pumps keep a highly precise eluent flow rate regardless of the conditions in the column. As a result, after calibration with standards, the positions of the peaks in time can be used for identification, and the peak heights and peak areas can be used to quantify analyte species without needing frequent recalibration.

Usually one of four basic types of chromatographic supports is used in LC. The general type of chemical interaction that occurs between the solid support and the solutes gives each separation a different name. The classifications are denoted as follows:

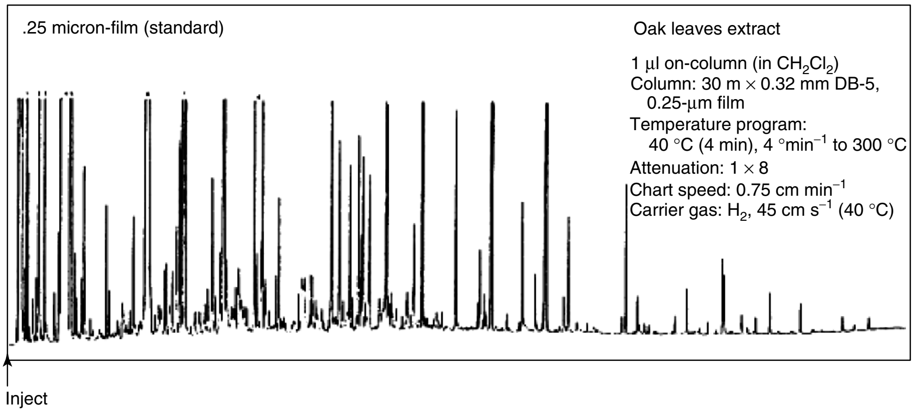

Fig. 16.23 Gas chromatogram of the volatile components of oak leaves. This was done on a capillary column. As complex as this appears, statistical arguments suggest that more than one component is under many if not most of the peaks. (Courtesy of J&W Scientific, Inc., Rancho Cordova, California.)

1. Normal phase or adsorption. The stationary phase is highly polar (i.e., large electric dipoles at the surface) and the eluent is somewhat less polar.
2. Reversed phase. The stationary phase is less polar than the eluent.
3. Ion exchange. The stationary phase has ionic charges on it and the separation is based on differences in charge.
4. Gel filtration, also called *gel permeation* or *exclusion*. The stationary phase has pores into which the analytes can migrate. Separation is by size.

Less commonly used are some solid supports that are designed to interact with specific chemical functional groups, such as thiols (–SH) or diols (–COH–COH–), that are on the solutes being separated. With a similar strategy of specific interactions, solid phases can be made with molecules that bind to one specific protein and can pull it out of a mixture. Chromatography with such protein-specific groups on the stationary phase is called *affinity chromatography*.

Running a chromatographic separation can take up to hours. In many cases, the process can be speeded up at least 10-fold if the analyte can withstand the conditions needed, which involve higher pressure to provide faster flow, higher temperatures, and shorter columns. However, the speed also requires detection methods that respond faster and possibly to smaller amounts of material. Nevertheless, this is an important trend.

Quantitation in column LC with automatic sample injection and with the use of internal standards can approach ±0.1%. Precision of ±1% is relatively straightforward with current instruments.

16.7.1.2 **Separations of Chiral Molecules**

There exist numerous molecules, particularly most biological ones, which, while having exactly the same compositions and same atoms attached to each other, nevertheless have different structures. Their differences are similar to the differences between right and left hands. As is true for hands, the molecular structures are mirror images of each other. These are chiral molecules. Significant progress has been made in recent years in both GC and HPLC separations of chiral molecules, that is, separating the right-handed isomers from the left-handed isomers.

The approaches are threefold. In one, a reaction is carried out binding a chiral reagent to the analytes, and the now-different reaction products separated with an ordinary chromatographic method. In the second, the analytes are chromatographed on a chiral stationary phase, and, since the two different structures are retained differently, they are separated. The third approach is a variation on the second; the stationary phase is not chiral, but the mobile phase has a chiral species added to it. This chiral species interacts reversibly with the chiral analytes, which allows a separation of the various products to occur on a nonchiral column. All three methods have been developed for HPLC. However, only the first two are used for GC chiral separations.

16.7.1.3 **Thin-Layer Chromatography**

In TLC, the adsorbent particles comprising the stationary phase are laid down as a layer a fraction of a millimeter thick attached to a solid backing plate of aluminum, glass, or polymer. Separations carried out on these stationary phases are called *TLC*.(Unfortunately, sometimes thick-layer chromatography, with layers 0.5–2 mm thick and primarily used for preparations, is also abbreviated TLC.) The primary advantage of TLC is that a number of samples and standards can be analyzed simultaneously side by side on the same plate. Using column chromatography, individual samples and standards must be analyzed sequentially. One other advantage of TLC is that samples that are difficult to resolve may be developed in two different solvents run in perpendicular directions: a two-dimensional separation. In this way, two or more species that run together in one solvent system may be separated from each other in the perpendicular direction in a second solvent system. One disadvantage of TLC is that quantitation tends to be less precise than in column chromatography. The greatest precision on intact plates is obtained with calibrated imagers.

16.7.1.4 **Field-Flow Fractionation**

When liquids flow without turbulence through a tube, the liquid at the walls is immobile, and that in the center flows the fastest. Since any suspended particles flow along with the fluid, the particles that reside in the center flow along the fastest and those nearer the wall more slowly. This is the basis for field-flow fractionation (FFF), which is carried out with the fluid flowing through a closed thin rectangular flow channel. FFF is the name of a set of techniques that use different kinds of forcing fields applied perpendicular to the flow direction across the thin direction to separate solutes of high molecular weight. The small separation set up by the field is, in effect, amplified by the flow of the fluid. The forcing fields can be set up by centrifuging (sedimentation), or by applying a voltage (electrical), or by applying a temperature difference (thermal). They find applications in the separation of macromolecules such as proteins, synthetic polymers, biological cells, and microscopic beads of various materials. Detection of the fractions is done with the same methods as for other LC methods: spectroscopic, electrochemical, and chemical.

16.7.2
Gas Chromatography

GC is an analytical method which has been studied and used intensively for more than 60 years in solving a wide variety of problems. Most classes of compounds that can be carried in the gas phase have been studied, and good separation conditions found for them.

GC equipment is quite different from that for HPLC. GC requires precise control of the flow of gases instead of liquids. The column, almost always, is significantly longer and/or narrower. Packing materials and stationary phases differ. The detectors differ both in construction and operation.

GC may be arbitrarily divided and named in terms related to the types of columns and column packings used. If the internal diameter of the column is in the range of 0.2–0.5 mm, this is called *capillary-column gas chromatography*, or *capillary GC*. With only a few exceptions, GC now means capillary GC when used in the context of analyses.

The analytes dissolve into and evaporate from the stationary phase multiple times while passing through the column. For analysis of organics on capillary columns, the material of the stationary phase generally is one of a number of different high-boiling-point liquids or stable solids that is coated on the inside surface of the capillary. With both these types of packings, the technique is called *gas chromatography* (*GC*). However, the latter is often called *gas–liquid chromatography* (*GLC*).

GC is routinely used for analyses of gases and for organic compounds with boiling points below about 450 °C. With boiling points in that range, the compounds have vapor pressures that are high enough so they will be carried in the gaseous mobile phase while the temperature is still low enough not to decompose either the analytes or the column materials. As a result, high-boiling materials are, perhaps, better separated by HPLC.

On the other hand, the GC column may be cooled during a separation of low-boiling materials such as methane or hydrogen. For such materials, GC can be run conveniently under conditions down to dry ice temperature, −80 °C.

The most broadly applicable detector for GC is a mass spectrometer (Section 16.5). With GC/MS, it is possible to identify most peaks that are of adequate size with good probability through the use of pattern-recognition methods. GC/MS is the method of choice for general organic chemical analysis, that is, when the identities of many different species are desired.

Other detectors, such as the electron-capture detector (ECD), are more sensitive and more specific than MS, and the specificity is often a benefit rather than a drawback. The ECD, which is relatively specific for halogens (F, Cl, Br, and I), is one of the most sensitive known; its detection limit is approximately 10^{-13} g s^{-1} of material passing into it. Even with capillary GC, in which the total sample may be in the microgram level, it is possible to obtain ppm determinations.

Numerous other types of detectors exist. Their use depends on the types of materials being analyzed and the sizes of the samples that are available. Two of the most common are the thermal conductivity detector, where the eluting gas flows over a hot wire and the temperature changes when eluents pass through, and the flame ionization detector. The latter works by burning the output material and causes gaseous ions of the separated components to form as they arrive, and the ions' quantities are measured. However, GC/MS is so powerful and flexible that MS has become the general detector of choice.

16.7.3
Supercritical Fluid Chromatography

In the chemistry of molecules that can be liquids at some temperature, there exists a temperature above which the substance will not condense to form a liquid-gas boundary at any pressure. This temperature is called the *critical temperature*. Above this temperature, the material is neither a liquid nor a gas; it is conveniently called a *supercritical fluid*. The mobile phase of supercritical fluid chromatography (SFC) is of this type. The most commonly used fluids are carbon dioxide and various lower molecular weight fluorocarbons. The equipment used for SFC is similar to GC, but the pressure throughout the column and the detector must be kept high enough to keep the conditions supercritical.

SFC is useful for certain classes of less-volatile analytes. Among these are molecules with hydrocarbon chains 12–50 carbons long such as the simple hydrocarbons themselves and detergents and fatty acids with such longer chains. Methods of separations and detection are well developed for these classes of compounds.

16.7.4
Electromigration

Electromigration is the movement of particles suspended in solution (most often water) in an electric field. The main electromigration technique that appears in chemical analysis is electrophoresis. Separations of charged particles from the size of individual ions of atoms to particles tens of micrometers in diameter are carried out by electrophoresis.

Electrophoresis separates various analytes based on their differences in rates of migration across the applied voltage. More highly charged particles migrate faster than less charged ones of the same size, and smaller particles migrate faster than larger ones with the same charge. The various analytes that are separated appear as zones as in other separation methods, and the name zone electrophoresis is used synonymously with electrophoresis alone. Electrophoresis is carried out either in a tube filled with solvent, or on a porous support – such as a plate with a gelled solution on it. The gel is, in fact, porous to the molecules which separate as they migrate through it. Gel electrophoresis is one of the primary techniques of biomacromolecule analysis. A popular technology involves carrying out electrophoretic separations in capillary tubes tens of micrometers in diameter – the diameter of hair. This produces high-resolution separations of small ionic species in micro samples. Perhaps the best known use of electrophoresis is the separation of sets of short DNA strands, which provides the experimental basis of most of contemporary genetic analysis.

16.8
Surface-Sensitive Methods

The difference between ‘‘regular’’ spectrometry and surface spectrometry is due not to the thinness of the sample, but to the thinness of the layer from which the analytical information is obtained. For example, in order to measure properties of a surface layer only, we can excite the spectral emission with particles, such as high-velocity electrons and ions. These can only penetrate through short distances (less than a few nanometers) of a solid. The name for the effective distance varies; two of these names are the penetration depth and the mean free path (see Table 16.7). The same limitation to penetration holds for the particles – such as electrons – expelled from the atoms residing in the interior.

With this idea in mind and from the information in Table 16.7, the following general statement can be made. To limit the spectrometric measurements to surface layers requires only that either energy is added using a bombarding charged particle or atom or a charged particle or atom is given off from the sample. Surface analyses are usually carried out under high vacuum, 10^{-10} Pa and less. ($1\ \text{Pa} = 9.87 \times 10^{-6}\ \text{atm} = 7.50 \times 10^{-3}$ torr.)

Other surface investigatory methods have been discussed in earlier sections although the surface analytical capabilities were not specifically emphasized. These methods are SIMS, and ultraviolet, visible, infrared and Raman, and X-ray spectrometry, done either at low angles of incidence or with total internal reflection methods. These methods can all be quantitative with precisions about ±20%.

Surface analyses heave two significant points of difficulty. Both result from the thinness of the sampled region – the "surface" of the assayed material. First, the quantity of sample is always small, and, as a result, the methods all involve microanalytical techniques. Second, the sample preparation methods are crucial to obtain a meaningful analysis. For instance, if a thin layer of oil about 5–10 *atoms* thick covers a metal sample, surface analysis may show only organic components in the sample.

16.8.1
Electron Spectrometry

16.8.1.1 Photoelectron Spectrometry

Electrons can be caused to be emitted from atoms and molecules upon irradiation with light. This is photoemission, and the electrons emitted are called *photoelectrons*. Each electron emitted has a kinetic energy (E_k) which is related to the wavelength of the exciting light and to the bonding energy of the electron that is emitted. Weakly held electrons have higher kinetic energies when emitted; strongly held ones have lower kinetic energies. An increase in irradiation intensity only increases the number of emitted electrons and not the individual kinetic energies.

When ultraviolet radiation is used, the spectrometry is called *ultraviolet photoelectron spectrometry* (*UPS* or *UV-PES*). If X-ray radiation is the source of ionizing energy, then the spectrometry is called *XPES* or *XPS*. Another name for XPS is ESCA. The two originally were separate classifications but are now joined together because a bright enough source of light tunable through both regions is available as synchrotron radiation.

Tab. 16.7 Approximate sampling depths of radiation (photons) and particles.

Name	Energy (eV)	Penetration depth (nm)
Infrared	0.5	100
Visible–UV	2–3	20
X-ray	1000	1000
Electron	2	0.5
Electron	1000	2
Ion	1000	2
Ion (proton)	10^7	10 000

For analysis of chemical content, however, it is XPS that is more often used, and its sensitivity tends to increase with atomic number of the elements. The strongly held electrons that are excited by the radiation are only slightly affected by chemical bonding. As a result, the values of the electron energies can be used to characterize the composition of the sample surface layer, but only major differences in chemical type can be seen. For example, sulfur in the form of sulfide and sulfate can be discriminated, but differences in the chemical environment of, for example, a sulfide, do not register.

In contrast, UPS causes electrons to be ejected from energy levels that change significantly with changes in molecular bonding. As a result, UPS is used more to investigate the types of chemical bonding present than for analyzing chemical makeup.

16.8.1.2 **Auger Spectrometry**

Strictly defined, the Auger electrons are also photoelectrons. However, it is convenient to separate them into their own category for a number of reasons. Irradiation with X-rays is comparatively inefficient in producing Auger electrons. Irradiation with accelerated electrons (a few hundred electron volts or so) is far more effective, and, thus, greater analytical sensitivity for the Auger method is obtained with electron excitation.

Auger spectrometry is primarily used to measure the elemental composition of a surface. Auger lines can be obtained for all elements except H and He. Its sensitivity is such that in the most favorable cases, less than 0.01 monolayer can be detected. Also, Auger spectra are characterized by a relatively high intensity, since the emitted electrons are able to be detected with excellent sensitivity. Further, the probability of Auger electron emission is high for elements lighter than zinc, making Auger spectrometry a complementary method to XPS. Auger spectrometry has become the method of choice for surface composition studies while XPS (ESCA) is used more for characterizing the chemical form of surface species.

If the electron beam which excites the Auger emission is focused onto a small (fraction of a micrometer) spot, the Auger electrons will be emitted only from the position of the spot on the surface. An Auger electron energy analysis can be used to quantify the elements at the one small part of the surface. The name given to this type of assay is electron microprobe analysis.

If the exciting electrons are focused to a small point and are scanned in a raster pattern over the sample (as done by the electron beam of a non-digital television screen to produce a picture), then a spatial distribution of the emissions can be made. For instance, either the characteristic X-rays or Auger electrons of a specific element can be monitored to obtain a picture of the spatial distribution of an element. An enlarged picture of the surface is generated on a television screen by simultaneously scanning the larger screen and a small area of the sample in synchrony. If the intensities of the Auger emissions for a specific element are recorded as a range of light intensity to produce a picture with good contrast, it will show the locations of the analyte. The Auger electrons without energy discrimination provide the basis for imaging in scanning electron microscopy (SEM).

16.8.2
Secondary-Ion Mass Spectrometry

If an accelerated charged particle impinges on a surface, it is stopped within a few atomic layers. One effect of this impact is to cause particles to be ejected from the surface. This is the process of sputtering. The emitted particles can be atoms, molecules, and molecule fragments. These may be analyzed with a mass spectrometer to determine information about the composition of the surface. The technique is called *secondary-ion mass spectrometry–SIMS*.

Of the sputtered material, only those that are ionized by the sputtering process can be accelerated directly into the mass spectrometer. As a result, complete analysis is problematic. However, quantitation and characterization can be more complete if all species are ionized. The neutral sputtered species can be ionized with UV light after their escape from the surface; this is photoionization. SIMS with photoionization has the potential to be far more reliable for quantitation.

Layer analysis can be carried out by obtaining a SIMS spectrum followed by ablation of the surface and another SIMS measurement. Alternate measurement and ablation can be reconstructed to view the chemical analysis of adjacent layers. Calibration of the ablation can provide an absolute depth scale.

16.8.3
Rutherford Backscattering

Rutherford backscattering involves, in effect, bouncing high-energy particles (usually helium ions with a few million electron volts of kinetic energy) off the nuclei of the surface atoms. Some of the energy of the incoming particles is transferred to the stationary atoms, and the amount of the energy transferred depends on the masses of the target atoms. As a result, the surface elemental composition can be probed by measuring the kinetic energy remaining with the probe particles after the collision. The technique is independent of the chemical bonding of the surface atoms, which results in this being a quantitative assay method. Absolute values of the coverage can be obtained without calibration of each sample since the scattering properties of each type of nucleus remain constant over the different samples. Less than a monolayer at the surface can be quantified if the surface layer is composed of heavier atoms on a solid composed of much lower weight atoms.

The mass resolution is not linear with the atomic mass of the analyte, and depends on the quality of the instrument. For instance, with 2-MeV ^{4}He probe particles, unit atomic masses can be resolved for atoms lighter than about chlorine, and, at the high end of atomic mass – near 200 amu – only masses differing by about 20 amu can be resolved.

16.9
Thermal Methods of Analysis

Energy added in the form of heat can cause changes in the structure or chemical content of materials. Examples of these changes are boiling (vaporization), melting (fusion), changes in crystalline form, transformation to and from a glass (glass transition), loss of water (dehydration), and chemical decomposition. These changes of structure occur at or above a characteristic temperature or range of temperature. Quantitation of the temperatures at which these changes occur and the amount of heat needed to cause these changes can be used to characterize or identify numerous materials. Among these are cements, ceramics, polymers, and plastics.

A large number of different techniques of thermal analysis have been devised including measurements of changes of electrical and magnetic characteristics, changes of elastic force, as well as changes in mass while a sample is heated. However, for chemical analysis, only a few are commonly used, and these are described briefly below.

16.9.1
Thermogravimetric and Evolved-Gas Analyses

Thermogravimetric analysis (TGA), or TG, involves measuring the change in mass of a substance as it is heated. TGAs involve recording the mass of a sample as its temperature is increased; mass is lost as decomposition occurs and volatile products escape, each process beginning at a characteristic temperature. Usually the temperature is raised at a steady rate, and the sample is kept in an atmosphere of flowing inert gas. Changes that occur in the sample that do not involve changes in mass, such as melting, cannot be studied by TG.

Characterizing the gaseous products of the decomposition as a function of sample temperature, a method called *evolved-gas analysis* is complementary to TG. The losses in mass from the sample may be due to loss of solvent or to breakdown of molecules in the material. Relatively detailed analyses can be done on sub-milligram samples. The analysis generally involves GC and/or MS of the volatiles.

16.9.2
Differential Scanning Calorimetry; Differential Thermal Analysis

The basis for calorimetric analyses is the measurement of the rise in temperature of a material for a given amount of heat added. The ratio of the temperature change to the heat added (T/Q) per unit weight is called the *heat capacity*. The heat capacity depends on the identity of the material and on the temperature at which the heat is added. Differential scanning calorimetry (DSC) works on the principle of keeping a constant rate of temperature change in a sample by regulating the amount of heat that is added. The rate of heat added is

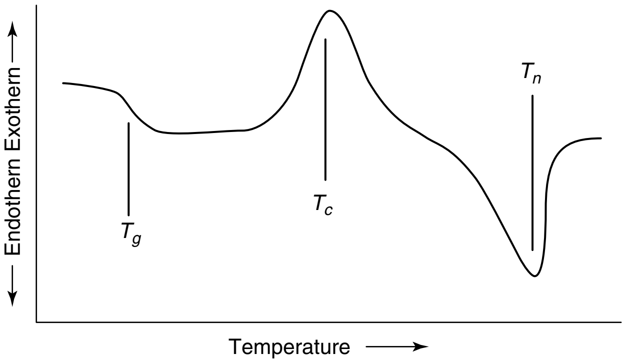

Fig. 16.24 Schematic differential scanning calorimetry thermogram of a polymeric material. The horizontal axis is a linear temperature scale, and vertical axis indicates the derivative of the amount of heat needed to obtain that temperature increase. The line would be flat if a material used the same amount of heat per degree to raise the temperature over the whole temperature range. The three deviations from the baseline (a flat line) are, from left to right, the glass transition, an exothermic (produces heat) chemical reaction, and endothermic (requires heat be added) melting.

plotted versus temperature as shown in Figure 16.24. This is called a *thermogram*.

An older method is differential thermal analysis (DTA) for which the heat is added at a constant rate, and the rate of temperature change is measured. The difference in temperature between a sample and standard is plotted versus temperature. The information obtained in DTA is similar to DSC, but they differ in experimental technique.

Acronyms of Analytical Techniques and Instruments

AA: Atomic absorption (spectrometry).
AAS: Atomic absorption spectrometry.
AE: Atomic emission (spectrometry).
AED: Atomic emission detector.
AES: Atomic emission spectrometry.
AES: Auger electron spectrometry.
AF: Atomic fluorescence.
ASV: Anodic stripping voltammetry.
ATR: Attenuated total reflection.
CARS: Coherent anti-Stokes Raman scattering (spectrometry).
CD: Circular dichroism spectrometry.
CI: Chemical ionization.
CI/MS: Chemical ionization mass spectrometry.
CP/MAS-NMR: Cross polarization magic-angle spinning nuclear magnetic resonance.
CW-NMR: Continuous-wave nuclear magnetic resonance.
DCP: Direct-current plasma.
DME: Dropping-mercury electrode.
DPP: Differential pulse polarography.
DSC: Differential scanning calorimetry.
DTA: Differential thermal analysis.
DTG: Differential thermogravimetry.
EAA: Electrothermal atomic absorption.
ECD: Electron-capture detector.
EDX: Energy-dispersive X-ray (spectrometers).
EDXF: Energy-dispersive X-ray fluorescence.
EGA: Evolved-gas analysis.
EI: Electron impact, electron ionization.
ELCD: Electrolytic conductivity detector.
EMP: Electron microprobe analysis.
ENDOR: Electron nuclear double resonance spectrometry.
EPR: Electron paramagnetic resonance (spectrometry).
ES: Emission spectrometry.
ESR: Electron spin resonance (spectrometry) (EPR).
ESCA: Electron spectrometry for chemical analysis.
ETA: Electrothermal atomization (electric furnace).
EXAFS: X-ray absorption fine structure.
FAA: Flame atomic absorption.
FAB: Fast atom bombardment.
FAE: Flame atomic emission.
FD: Field desorption.
FFF: Field-flow fractionation.
FFT: Fast Fourier transform.
FIA: Flow injection analysis.
FID: Flame ionization detector.
FT-IR: Fourier transform infrared (spectrometry).
FT-MS: Fourier transform mass spectrometry.
FT-NMR: Fourier transform nuclear magnetic resonance.
GC: Gas chromatography.
GC-IR: Gas chromatography with infrared detection.
GC-MS: Gas chromatography with mass spectrometry.
GDMS: Glow-discharge mass spectrometry.
GFAA: Graphite-furnace atomic absorption.
GPC: Gel permeation chromatography.

HPLC: High-performance liquid chromatography.

HPTLC: High-performance thin-layer chromatography.

IC: Ion chromatography.

ICAP: Inductively coupled argon plasma.

ICP: Inductively coupled plasma.

ICP/MS: Inductively coupled plasma/ mass spectrometry.

IC: Ion chromatography.

INAA: Instrumental neutron activation analysis.

IR: Infrared (spectrometry).

ISE: Ion-selective electrode.

LAMMS: Laser microprobe mass spectrometry.

LC: Liquid chromatography.

LC/EC: Liquid chromatography/electrochemical detection.

LC/MS: Liquid chromatography/mass spectrometry.

LEED: Low-energy electron diffraction.

MAS-NMR: Magic-angle spinning nuclear magnetic resonance.

MCD: Magnetic circular dichroism (spectrometry).

MIP: Microwave induced plasma.

MS: Mass spectrometry.

NAA: Neutron activation analysis.

NHE: Normal hydrogen electrode.

NIR: Near infrared.

NMR: Nuclear magnetic resonance.

NPD: Nitrogen phosphorus detector.

NQR: Nuclear quadrupole resonance (spectrometry).

ORD: Optical rotatory dispersion (spectrometry).

PAS: Photoacoustic spectrometry.

PD: Plasma desorption (in mass spectrometry).

PID: Photoionization detector.

PIXE: Particle-induced X-ray emission.

PMT: Photomultiplier tube.

QCM: Quartz crystal microbalanceRBS: Rutherford backscattering (spectrometry).

RI: Refractive index.

RIS: Resonance ionization spectrometry.

RP: Reversed phase (in liquid chromatography).

SAM: Scanning Auger microscopy (spectrometry).

SAW: Surface acoustic wave (sensor)SCE: Saturated calomel electrode.

SEM: Scanning electron microscopy.

SERS: Surface-enhanced Raman scattering.

SFC: Supercritical fluid chromatography.

SHE: Standard hydrogen electrode.

SIMS: Secondary-ion mass spectrometry.

SNMS: Sputtered neutral mass spectrometry.

STM: Scanning tunneling microscopy.

TCD: Thermal conductivity detector.

TG: Thermogravimetry (see TGA).

TGA: Thermogravimetric analysis.

TLC: Thin-layer chromatography.

TLC: Thick-layer chromatography.

TMS: Tetramethylsilane.

TOF/MS: Time-of-flight mass spectrometry.

TS: Thermospray.

UV: Ultraviolet (light) spectrometry.

vis: Visible (light) spectrometry.

WDX: Wavelength dispersive X-ray (spectrometer).

XRF: X-ray fluorescence.

XPS: X-ray photoelectron spectrometry (see ESCA).

XPES: X-ray photoelectron spectrometry (see ESCA).

XANES: X-ray absorption near-edge spectrometry.

Glossary

Accuracy: A qualitative measure of the closeness with which an analytical result corresponds to the true content of a material.

Analyte: The species for which one is analyzing.

Assay: As a noun, a chemical analysis to determine the level of a specific analyte. As a verb, to carry out such an analysis.

Bias: The quantitative amount that a result with high precision varies from the ''true content'' as a result of the analysis shortcomings.

Calibrate: To determine the bias and precision of the methodology used for an analysis or to determine an instrument's response to some known quantity of analyte.

Contaminant: Any material that is picked up during an analytical procedure that may cause the sample not to be characteristic of the original material from which the sample was taken.

Destructive Method: An analytical method that requires some obvious portion of a material to be removed and destroyed in the process of analysis. The opposite of a nondestructive method.

Detection Limit (Limit of Detection): The limit below which a particular instrument or analytical method cannot be used to observe the presence or absence of some analyte in a sample.

Electromagnetic Radiation: Depending on the frequency of oscillation, also called radio waves, infrared, visible, and ultraviolet light (or, simply, light), X-rays, and γ *rays*. Electromagnetic radiation is also described as oscillating orthogonal electric and magnetic fields that travel through space.

Interferent(s): A component or components (not the analyte) in a sample that cause a measure of the analyte content to be greater or less than it would be in the absence of the interfering components.

Matrix: The part of the material to be analyzed that is not of interest in an analysis.

Micro Analysis: An analysis of a sample that weighs less than approximately 10 mg.

Nondestructive Method: An analytical method that, after analysis, leaves the material being analyzed with no apparent changes.

Potential: Electric potential difference. Voltage. The difference in electric potential energy per unit charge.

Parts per Million: A relative measure of content. One ppm is one part in 10^6 of the sample such as grams per gram.

Parts per Billion: A relative measure of content. In the United States, one ppb is one part in 10^9 of the sample such as nanograms per gram. In Europe, one ppb is one part in 10^{12} such as nanograms per kilogram.

Parts per Trillion: A relative measure of content. In the United States, one pptr is one part in 10^{12} of the sample. In Europe, one pptr is one part in 10^{18}.

Precision, Imprecision: A measure of the scatter of experimentally determined values of content. The highest precision corresponds to the least scatter. Not the same as accuracy.

Sample, Laboratory Sample: The part of the original material that is used to produce an analytical result. It should accurately represent the contents of the original material.

Spectrum: A graph of the changes in the amount (more precisely, power) of electromagnetic radiation that is absorbed or emitted by a sample versus the wavelength, frequency, or energy of the radiation. Also, instrument response versus mass in mass spectrometry.

Sensitivity: The ability of an analytical method to measure changes in content or concentration. Greater sensitivity means that smaller changes give a measurable difference in response.

Standard: Substances that are used to ascertain the accuracies and precisions of analytical methods.

Standard Addition: A method of running standards in which a measurement is made on a sample and then repeated after known quantities of the analyte are added to the sample. Used for complicated mixtures.

Trace Analysis: Analyses for components that are less than 0.01% (100 ppm) of a material but more than ~1 part in 10^9.

Ultra-microanalysis: An analysis of samples of total mass in the range below about 10μg.

Ultratrace Analysis: Analyses for components of a sample that are at levels below ~1 part in 10^9.

Verification: Demonstration that an analytical procedure measures what it is claimed to measure.

Voltage: See Potential. Electric potential difference.

References

Ahmad, S. *et al.* (**1979**) *J. Radioanal. Chem.*, **54**, 331–341.

Commission on Microchemical Techniques and Trace Analysis (**1982**) *Pure Appl. Chem.*, **54**, 1565–1577.

Derome, A.E. (**1987**) *Modern NMR Techniques For Chemistry Research*, Pergamon, Oxford.

Harris, R.K. (**1983**) *Nuclear Magnetic Resonance Spectroscopy*, Pitman, London.

Horwitz, W. (**1982**) *Anal. Chem.*, **54**, 67A–76A.

Morrison, G.H. (**1979**) *CRC Crit. Rev. Anal. Chem.*, **33**, 287–320.

Rubinson, K.A. and Rubinson, J.F. (**2000**) *Contemporary Instrumental Analysis*, Prentice Hall, Upper Saddle River.

Sandell, E.B. (**1959**) *Colorimetric Determination of Trace Metals*, 3rd edn, Interscience, New York.

Wawros, A., Talik, E., and Pastuszka, J.S. (**2003**) *Microsc. Microanal.*, **9**, 349–358.

Further Reading

Worsfold, P.J., Townshend, A., and Poole, C.F. (**2004**) *Compendia Encyclopedia of Analytical Science*, 2nd edn, Academic Press, New York.

Further Reading: Textbooks

Rubinson, K.A. and Rubinson, J.F. (**2000**) *Contemporary Instrumental Analysis*, Prentice Hall, Upper Saddle River.

Skoog, D.A., West, D.M., Holler, F.J., and Crouch, S.R. (**2006**) *Principles of Instrumental Analysis*, Brooks/Cole, Florence.

CHEMICAL KINETICS

ROBERT W. CARR, *Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, Minnesota, U.S.A.*

Introduction 345
1. **Elementary Concepts** 346
1.1 Empirical Kinetics 346
1.2 Elementary Reactions 347
1.3 Kinetic Data 348
1.4 Mechanism 348
1.5 Microscopic Kinetics 349
1.6 The Effect of Temperature 350
1.7 Microscopic Reversibility and Detailed Balance 351
1.8 Thermodynamic Restrictions on Kinetics 352
2. **Potential Energy** 352
3. **Bimolecular Collision Dynamics** 356
3.1 Simple Collision Theory 356
3.2 Collision Dynamics 357
3.3 Classical Trajectory Calculations 359
4. **Transition-State Theory** 360
4.1 Conventional Transition-State Theory 360
4.2 Thermodynamic Formulation 361
4.3 The Transmission Coefficient 361
4.4 Variational Transition-State Theory 362
5. **Unimolecular Reactions** 362
5.1 The Lindemann Mechanism 362
5.2 Theoretical Models 363
5.3 Chemical Activation and Photoactivation 365
6. **Association Reactions** 367
6.1 Termolecular Combination of Atoms 368
6.2 Combination of Polyatomic Radicals 368
7. **Molecular Reaction Dynamics** 370
7.1 Molecular Beams 370
7.2 Energy Disposal in Chemical Reactions 372
Glossary 372
Works Cited 374
Further Reading 375

INTRODUCTION

Chemical kinetics is the description of rates of chemical transformations, the rates at which chemical species interact with one another to produce new chemical species. Along with the study of rates, chemical kinetics investigations usually seek to understand how chemical transformations occur. At the molecular level, a description of how the atomic motions facilitate the conversion of one species into another is the fundamental question that ultimately leads to a detailed understanding of chemical reactions.

Chemical reactions (*q.v.*) are ubiquitous in nature. They are essential for life, and they also occur in nonliving environments. Reactions are found in the gas phase, in liquids, and in solids. For example, gas-phase reactions occur in planetary atmospheres, including that of the Earth, and in flames and combustion. The well-known topics of urban air pollution (*q.v.*) and stratospheric ozone depletion have at their heart a very large number of gas-phase chemical reactions involving man-made pollutants that account for the production of harmful substances and, in the case of the stratosphere, the removal of ozone. Liquid-phase reactions span the range from biological reactions, such as enzyme catalysis and cellular processes, to industrial applications, such as the manufacture of polymers and organic and inorganic chemicals. Examples of solid-state reactions may be found in cooking food and in explosives. Many more examples could be cited. Chemical reactions are also of enormous practical importance. The chemical industry depends upon chemical reactions to convert raw materials or other

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

feedstocks into higher value substances. The modern standard of living would be quite different without sufficient knowledge of chemical reactions for success of the chemical industry.

The rates of reactions span an enormous range of time scales, from geologic to 10^{-15} s, the former for natural processes in the Earth's crust, and the latter for the fastest motions of atomic nuclei. Laboratory investigations span a somewhat shorter range, from a few days or so, depending on the patience of the experimenter, to a few femtoseconds. The latter has only recently become possible through the development of ultrashort pulsed laser sources that make time-resolved spectroscopic investigations on such a short time scale possible.

There is a trinity of levels for approaching problems in chemical kinetics. The first of these is empirical measurements of overall rates of change. This approach goes back to the beginnings of kinetics in the last century, and is still practiced today. A more detailed level consists of examining rates of change of individual molecular events. Such changes are referred to as elementary reactions. The most detailed level, called molecular dynamics (*q.v.*), seeks to understand how individual quantum states evolve during a reaction.

This article will cover the three approaches to kinetics. However, chemical kinetics is such an obviously broad field that complete coverage would be impossible in the limited space available. To narrow the scope, only gas-phase reaction kinetics will be covered, although the section on empirical kinetics is not restricted, the principles there being applicable to liquids and solids as well. The choice of gas-phase reactions is appropriate since it is only in the gas phase that the enormous wealth of detail that comes from being able to examine the chemical consequences of individual collisions, and of molecules in isolation, is accessible.

1. ELEMENTARY CONCEPTS

1.1 Empirical Kinetics

Overall chemical reactions, or stoichiometric reactions, describe the stoichiometric relationships between the reactants, those chemical species that are brought into contact initially, and products, those that are formed as a result of the contacting. They contain no information on reactive intermediates, such as free radicals, ions, or other species that may be present during the transformation but that are not found among the stable products. The stoichiometric relationship is conventionally expressed as

$$aA+bB+\cdots=mM+nN+\cdots, \tag{1}$$

where the lowercase letters are the stoichiometric coefficients and the uppercase letters are the chemical species. The stoichiometric coefficients prescribe the relative number of moles of each reactant producing the observed molar abundance of the products. This can be viewed as an algebraic equation and rearranged to $-aA-\cdots+mM+\cdots=0$, which can then be written in the compact notation

$$\sum \nu_i A_i=0, \tag{2}$$

where the ν_i are the stoichiometric coefficients, and the A_i are the chemical species. The stoichiometric coefficients are usually taken as being negative for reactants and positive for products. The summation runs over all chemical species.

The rates of consumption of reactants and formation of products, in moles n per unit time, in a closed system are related by the stoichiometry:

$$\begin{aligned} R&=(-1/a)(dn_a/dt)\\ &=(-1/b)(dn_b/dt)\cdots\\ &=(1/m)(dn_m/dt)\cdots,\\ R&=(1/\nu_i)(dn_i/dt). \end{aligned} \tag{3}$$

Consequently, for a self-consistent set of rates, the differentials dn/dt should always be divided by the appropriate stoichiometric coefficients. The above definition of rate makes it an extensive property. To make the rates independent of system size, the concentration, $[A_i]=n_i/V$, where V is the system volume, should be used. In nonconstant volume systems, the intensive rate has two terms,

$$\begin{aligned} r=R/V&=(\nu_i V)^{-1}(d/dt)(V[A_i])\\ &=(1/\nu_i)(\partial[A_i]/\partial t+[A_i]\partial\ln V/\partial t). \end{aligned} \tag{4}$$

At constant volume this reduces to

$$r=\nu_i^{-1}(d[A_i]/dt). \tag{5}$$

In open systems, the definition of rate must include terms for convection and diffusion. Most kinetics experimentation is designed to exclude transport effects since these may obscure the principal objectives of the work. Hence the description of open systems will not be elaborated upon here.

The functional dependence of the rate upon system parameters is called the rate expression. Important parameters are the concentrations of the reactants and (sometimes) the products, and the temperature. Other quantities may also play a role. In gas-phase reactions the total pressure is often important, while in liquid solutions ionic strength and dielectric constant are frequently found to influence the rate. Rate expressions are empirically determined, and in principle may take any form that fits the data:

$$r=f(T,A_1,A_2,\ldots,A_i\ldots,\text{other quantities}). \quad (6)$$

A common form is

$$r=k(T)[A_i]^{\alpha_1}[A_2]^{\alpha_2}\cdots[A_i]^{\alpha_i}\cdots=k(T)\Pi[A_i]^{\alpha_i}, \quad (7)$$

where $k(T)$ is called the rate constant or rate coefficient; the latter is preferable since $k(T)$ is not a constant, but in a properly determined rate expression is a function of temperature only. In such an expression the α_i's are the order of the reaction with respect to each chemical species. They are numbers that may take any value given by the results of the experiments, including negative numbers. Since they are empirically determined, they are not necessarily integers, although they often are found to have whole-number values. The overall order of the reaction is given by the sum of the individual orders,

$$\text{overall order}=\sum \alpha_i. \quad (8)$$

Frequently, the order is not defined. For example, the rate expression for the reaction between hydrogen and bromine, $H_2+Br_2=2HBr$, is

$$\tfrac{1}{2}d[HBr]/dt=(a[H_2][Br_2]^{1/2})/(b+[HBr]/[Br_2]), \quad (9)$$

where a and b are coefficients that are functions of temperature. Initially, before much HBr has accumulated, so that $[HBr]/[Br_2]\ll b$, the reaction appears to be first order in H_2 and one-half order in Br_2, while by the time enough HBr has accumulated that $[HBr]/[Br_2]\gg b$, the reaction appears to be first order in H_2, three-halves order in Br_2, and of order minus one with respect to HBr. Thus the order is observed to change during the course of the reaction. This behavior is by no means limited to this particular reaction, but is possible for any empirically determined rate expression. It should be kept in mind that the goodness of rate expressions is limited to the range of experimental conditions used, and should not be trusted outside this range.

The experimental determination of kinetic rate expressions is a subject that has been treated at length in the literature (Freiss *et al.*, 1961, 1968; Espenson, 1981). A survey of methods would be too lengthy for this article.

1.2 Elementary Reactions

The overall reactions described above are composed of a list of elementary reactions called the mechanism. An elementary reaction is a molecular event that normally involves chemical transformation of the reactant(s) into new chemical species, and frequently also involves collisional transfer of energy without chemical change. The reactants may be atoms, molecules, free radicals, or ions. The molecularity of an elementary reaction is defined as the number of molecules involved in the event. In unimolecular reactions a single reactant undergoes a spontaneous transformation

$$A\rightarrow\text{product(s)}$$

while bimolecular and termolecular reactions involve interactions of two and three species, respectively,

$$A+B\rightarrow\text{product(s)},$$

$$A+B+C\rightarrow\text{product(s)}.$$

Reactions of higher molecularity are not observed because the probability of obtaining simultaneous or nearly simultaneous interaction between four or more species is very small. The kinetic order of an elementary reaction is equal to its molecularity, and is thus limited to the positive integral values 1, 2, and 3, in contrast to the order of an overall reaction. Furthermore, the rate expression for an elementary reaction can be written by inspection. For example, the rate of the reaction between Cl atoms and methane, $Cl + CH_4 \rightarrow HCl + CH_3$, is expressed as

$$-d[Cl]/dt = k(T)[Cl][CH_4], \tag{10}$$

which is kinetically second order overall and first order each with respect to Cl atoms and methane. Elementary reactions are fundamental descriptions of chemical change, in contrast with the empirical nature of overall reactions. The rate coefficient can be described by rate theories, and since it is unique to the reaction between two specific species, such as Cl and CH_4, it can be used with confidence in any overall reaction in which that reaction occurs. Accordingly, much effort has been devoted to obtaining rate coefficients of elementary reactions, not only for use in describing overall reactions, but for the insight they provide in understanding how chemical transformations occur.

1.3 Kinetic Data

The kinetics of a reaction are determined by measuring concentrations of the reactants as functions of time. It is desirable to make similar measurements of products and, if possible, of reactive intermediates. Elementary reactions are studied by arranging experimental conditions so that the reaction being observed is the only kinetically important process, other reactions having negligible effect on the rate of the one being measured. Considerable effort is usually spent on experimental design to ensure isolation of the elementary reaction, and accuracy of the rate constants. Physical methods for *in situ* determination of concentrations without perturbing system composition are most desirable. Optical spectroscopies and mass spectrometry are common methods, but other physical measurement techniques may be used. Chemical analysis (*q.v.*) and chromatographic methods are usually only suitable for analysis of the final reaction mixture because they are slow and require removal of material. Reaction times that can be dealt with range from the very long, days or more, to as short as femtoseconds with the use of femtosecond pulsed lasers. Modern optical and electronic technology make investigations of highly reactive species, such as atoms and free radicals, on microsecond to nanosecond time scales possible. Fast reactions that require mixing of reagents are limited to milliseconds or perhaps slightly shorter times by mixing rate limitations. Experimental methods are treated in detail in several texts (Freiss *et al.*, 1961, 1963; Melville and Gowenlock, 1964; and Bamford and Tipper, 1969).

1.4. Mechanism

The list of elementary reactions that comprise the mechanism of a given overall reaction must yield the correct stoichiometric relationship for the overall reaction, with cancellation of reactive intermediates, when the reactions are regarded as algebraic relationships. Important energy-transfer steps must be included. Also, the net rate given by the mechanism must agree with the empirical rate expression. Writing the rate equation for each elementary reaction in the mechanism gives a set of coupled differential equations. When an analytical solution to this set of equations can be obtained, it must correspond with the empirical rate expression. Otherwise, if the rate coefficients of the elementary reactions are known, numerical simulation must yield the experimentally observed concentration-time relationships for the reactants. More than one mechanism can usually be found to satisfy these criteria. To distinguish between mechanistic possibilities normally requires careful experimental diagnostics. The nature of the experimentation is to attempt to disprove certain steps, or to obtain evidence in support of them. This gives rise to the frequently quoted caveat that a mechanism can never be proved, only disproved. Nevertheless, with modern experimental methods for direct, time-resolved detection of short-lived intermediates, it is possible to compile a compelling case in favor of a mechanism.

The analysis of complex overall reactions is a problem in the solution of simultaneous differential equations. The equations are usually coupled through the chemical species, which may occur in more than one rate equation. In homogeneous systems with uniform composition, it is a set of ordinary differential equations. If spatial variations of concentration exist, then a set of partial differential equations must be solved. In some cases, analytical solutions exist. There is extensive documentation of these in the literature (Szabo, 1961). However, in the vast majority of cases, analytical solutions cannot be obtained, and approximate methods must be sought.

An important method that has been much used in the past is the steady-state approximation. This can be understood by considering the sequence of irreversible reactions

$$A \rightarrow B \rightarrow C$$

for which the rate expressions are $-d[A]/dt = k_1[A]$, $d[B]/dt = k_1[A] - k_2[B]$, and $d[C]/dt = k_2[B]$, and for which analytical solutions $[A](t)$, $[B](t)$, and $[C](t)$ exist. It can be shown that in the limit as $k_1/k_2 \rightarrow 0$ the solution reduces to that for the set of equations $-d[A]/dt = k_1[A]$, $d[C]/dt = k_2[B]$, and $k_1[A] - k_2[B] = 0$. The three differential equations become two differential equations and an algebraic equation that can be obtained by setting $d[B]/dt = 0$ in the original set of differential equations. Since $[B]$ is now time invariant it is in a steady state. In practice k_1/k_2 does not approach zero, although the value of the ratio may be quite small when B is a reactive intermediate, such as a free radical. In such cases $d[B]/dt \cong 0$, and setting $d[B]/dt = 0$ is called the steady-state approximation. In the mechanism of a complex reaction it is common practice to set the time derivatives of all reactive intermediates equal to zero, converting, in each case, the differential equation to an algebraic equation. In this way a mixed set of differential and algebraic equations that has an analytical solution may be obtained.

When the mechanism consists of a set of linear differential equations the Laplace transform method can be applied (Jenson and Jeffreys, 1963; see also FOURIER AND OTHER MATHEMATICAL TRANSFORMS). This is of somewhat limited use since most mechanisms have reaction rate terms that are nonlinear. In these cases it is possible to linearize the equations, but the linearization, which is usually done by Taylor-series expansion, is only accurate near the origin of the expansion. Systems of first-order kinetic equations can also be solved by matrix methods (Jenson and Jeffreys, 1963). With the increasing use of digital computers, numerical methods (*q.v.*) have become more prevalent. Small sets of equations can be efficiently solved by Euler methods, or Runge-Kutta methods. The fourth-order Runge-Kutta method is particularly useful and accurate. A problem frequently encountered is called stiffness. This occurs when widely varying time constants are present. To calculate rapidly changing species concentrations accurately, small time steps are required, but these mean that computation of slowly changing compositions consumes large amounts of computer time. Fortunately, efficient algorithms employing variable time steps have been developed. These permit numerical integration of large, stiff mechanisms on modern, fast main-frame computers in very short times. Mechanisms describing atmospheric chemistry, and combustion reactions, including flames, have been developed that consist of upwards of 100 reactions. Clearly, with such large sets of coupled differential equations, numerical simulation is required.

1.5 Microscopic Kinetics

Microscopic kinetics seeks to understand the role of individual quantum states (see QUANTUM MECHANICS). The goal here is to observe the effect of molecular internal energy levels and of translational energy on the state and course of a reaction, and to see what energy levels of the products are populated. Studies aiming at this kind of a description are frequently called state-to-state kinetics. At this level of detail it is possible to gain insights into chemical reactivity that cannot be realized from research on the kinetics of overall reactions and conventional studies of elementary reactions. The use of crossed molecular beams, where each reactant is formed into a molecular beam intersecting the other in a well-defined region, and spectroscopy, particularly with tunable lasers, has permitted great progress to be made in this area. Experiments designed to select the initial states of the reactants are normally performed. In such work one is dealing with non–thermal-equilibrium energy distributions.

A bimolecular reaction between state-selected reactants leading to one pair of product states would be represented as

$$A(n_a,v_a,J_a,\mathbf{u}_a) + B(n_b,v_b,J_b,\mathbf{u}_b) = C(n_c,v_c,J_c,\mathbf{u}_c) + D(n_d,v_d,J_d,\mathbf{u}_d). \quad (11)$$

The quantities n_i, v_i, and J_i represent the electronic, vibrational, and rotational energy of each species, and $\mathbf{u}_i$ is the velocity. Although there has not yet been a report of a completely state-selected reaction, many examples of partially state-selected reactions can be found in the literature. Common cases are those where reactant(s) are observed in a single internal energy state, but have a Maxwell–Boltzmann distribution of translational energies, or have been velocity selected in a molecular-beam experiment, but have a distribution of internal energy states.

The connection between reaction rate at the microscopic level and the elementary reaction level can be made if the rate of the state-to-state process is written conventionally. For convenience, define A_i to be chemical species A in quantum state i, and B_j the same for species B. Then the state-to-state reaction rate is

$$r_{ij}=k_{ij}[A_i][B_j]. \tag{12}$$

In a situation where many states are populated, the total rate will be given by the sum of the individual rates,

$$r=\sum\sum r_{ij}. \tag{13}$$

If the population distributions are written as $[A_i]/[A]=f_i(\epsilon)$ and $[B_j]/[B]=f_j(\epsilon)$, then the rate becomes

$$r=[A][B]\sum\sum k_{ij}\,f_i(\epsilon)f_j(\epsilon), \tag{14}$$

and in systems where thermal equilibrium is maintained a Boltzmann distribution, $f(\epsilon)=g\exp(-\epsilon/kT)$, can be used:

$$r=[A][B]\sum\sum k_{ij}g_ig_j \times\exp[-(\epsilon_i+\epsilon_j)/kT]. \tag{15}$$

In systems where the rate of reaction is slow compared with the rate of collisional redistribution of energy, the Boltzmann distribution will be an excellent approximation. However, when the reaction rate is faster than redistribution the Boltzmann distribution may become significantly perturbed as reaction depletes states that have large microscopic rates.

Since the rate of a bimolecular reaction is written as

$$r=k(T)[A][B], \tag{16}$$

it can be seen that the thermal rate coefficients $k(T)$ can be associated with the microcanonical rate coefficients by the equations

$$k(T)=\sum\sum k_{ij}\,f_i(\epsilon)f_j(\epsilon),$$

$$k(T)=\sum\sum k_{ij}g_ig_j\exp[-(\epsilon_i+\epsilon_j)/kT]. \tag{17}$$

1.6 The Effect of Temperature

The rate coefficient $k(T)$ depends only upon temperature. The most common form of the temperature dependence is the Arrhenius equation,

$$k(T)=A\exp(-E_a/RT), \tag{18}$$

an empirical equation in which the preexponential factor A and the activation energy E_a are temperature-independent parameters, and R is the gas constant. The historical development of this important equation in the late 19th century has recently been summarized (Laidler, 1987). The Arrhenius equation predicts that a plot of $\ln k(T)$ vs $1/T$ (an Arrhenius plot) will be linear. The slope of the plot defines the activation energy according to the relationship

$$E_a=-R[d\ln k(T)/d(1/T)]. \tag{19}$$

In the vast majority of kinetics investigations over the past century the linearity of such plots has been repeatedly verified. More recently, it has been recognized that when data are available over a wide temperature range, Arrhenius plots are curved. Moreover, for some kinds of reactions, most notably elementary association reactions of the type $A+B=C$, the temperature dependence is definitely non-Arrhenius. The reasons for this will be discussed in Sec. 6.

A sketch of $k(T)$ vs T for the Arrhenius equation is given in Fig. 1, which clearly shows that the preexponential factor can be associated with the value of the rate coefficient as $T\to\infty$. The inflection point is easily shown to be $T_i=E_a/2R$. Activation energies typically have values that range from the vicinity of zero or a few kJ/mol to 400 kJ/mol, or in a few cases somewhat higher. For $E_a=200$ kJ/mol, $T_i=12\,000$ K, while for $E_a=20$ kJ/mol, $T_i=1200$ K. Thus, unless relatively low activation energy reactions are carried out at high temperatures, rate constants will normally be measured on the rapidly rising portion of Fig. 1.

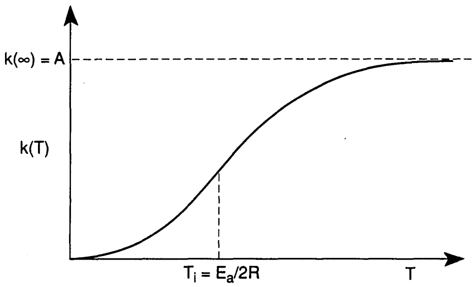

FIG. 1. Temperature dependence of the rate coefficient, $k(T)$, according to the Arrhenius equation, $k(T) = A \exp(-E_a/RT)$. T_i is the inflection point, and $k(\infty)$ is the rate coefficient as $T \rightarrow \infty$.

Theories of chemical kinetics usually predict temperature-dependent preexponential factors of the form $A = BT^n$, where n is normally a small number, which may be either positive or negative. Arrhenius plots are frequently found to be linear because the temperature dependence of the exponent is so much stronger than the temperature dependence of the preexponential factor. When experimental error is taken into account, curvature due to the preexponential is only detectable when data are available over an extended temperature range. An example of this is given in Fig. 2 for the elementary reaction of the OH radical with methane, an important combustion reaction. The activation energy is only 1060 J/mol, making the temperature dependence of the exponent less important, and the rate data extend from about 250 K to about 2000 K. It can easily be seen that over a temperature range of only 50–100 K, which is not uncommon in experimental studies of reactions, curvature would go undetected.

The activation energy is also temperature dependent, although the dependence is very weak. There is a statistical mechanical argument, due to Tolman (Fowler and Guggenheim, 1939), showing that the activation energy is given by the equation

$$E_a = \langle E^* \rangle - \langle E \rangle, \tag{20}$$

where $\langle E^* \rangle$ is the average energy of those molecules that can react, and $\langle E \rangle$ is the average energy of all of the molecules. The criterion for a molecule to be able to react is that it must have energy greater than the critical energy, a threshold energy below which molecules cannot react. Since $\langle E^* \rangle$ and $\langle E \rangle$ have similar temperature dependences, E_a is only weakly temperature dependent. It is worth emphasizing that Tolman's equation provides an interpretation of the activation energy.

1.7 Microscopic Reversibility and Detailed Balance

The principle of microscopic reversibility derives from the fact that the equations of motion for atoms and molecules are invariant under time reversal. Thus if in the final state after a molecular collision (reaction) all the momenta are reversed, including those of internal energy states, then the system will exactly retrace its path in the reverse direction to form a state similar to the original initial state, but with reversed momenta. This will be true of all of the energy states of a reacting molecule or pair of molecules. Application of microscopic reversibility to a macroscopic system consisting of a statistically large assembly of molecules leads to the conclusion that all chemical reactions are reversible. Furthermore, at chemical equilibrium, where all species concentrations are stationary in time, the rates of reaction in the forward and

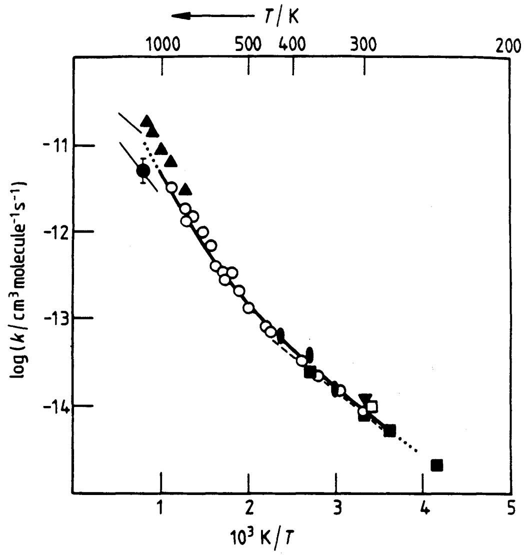

FIG. 2. Arrhenius plot of the rate coefficients for the reaction $OH + CH_4 \rightarrow H_2O + CH_3$. Symbols are data points, the thick solid line is a fit using the equation $\log(k/\text{cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}) = -20.24 + 3.08 \log(T/K) - 1.01 \times 10^3 K/2.303T$, and the other solid and dotted or dashed lines are fits to subsets of data. [Reproduced from Smith (1980) with permission of the author.]

reverse directions are equal. This leads to the principal of detailed balance. Consider the elementary reaction

$$A+B \underset{k_r}{\overset{k_f}{\rightleftarrows}} C+D,$$

where the reverse arrows indicate the reversibility of the reaction, and k_f and k_r are the rate coefficients for the forward and reverse reaction, respectively. It is readily shown that at chemical equilibrium,

$$k_f/k_r=K(T), \tag{21}$$

where $K(T)$ is the chemical equilibrium constant, defined in Eq. (22). Since the rate coefficients are independent of composition, this relationship is also true when the reaction is not at equilibrium, and holds over the entire course of the reaction. In a complex reaction, detailed balance applies to each elementary reaction in the mechanism. It rules out some types of mechanisms, such as the cyclic process

$$\begin{tikzcd} & A \arrow[dr] & \\ C \arrow[ur] & & B \arrow[ll] \end{tikzcd}$$

which cannot occur as written, although it is possible to envision starting, for example, with pure A and reacting until a cyclic equilibrium is obtained. Instead, for these three reactions each step must be reversible.

1.8 Thermodynamic Restrictions on Kinetics

The conclusions about reversibility of chemical reactions reached from considerations of microscopic reversibility are also available from thermodynamic considerations. The well-known relationship between the equilibrium constant, $K(T)$, and the standard Gibbs free-energy change,

$$K(T)=\exp[-\Delta G^0/RT], \tag{22}$$

combined with the ideal solution relationship between chemical equilibrium composition and the equilibrium constant,

$$K(T)=\Pi[A_i]^{\nu_i}, \tag{23}$$

where $[A_i]$ is the concentration of the ith chemical species and ν_i is the stoichiometric coefficient, shows that a reaction would only be irreversible if $\Delta G^0 \rightarrow -\infty$. In general, then, all reactions are reversible, although in cases where the Gibbs free-energy change is large and negative the reaction may appear to go to completion in the sense that only very small quantities of reactant remain at chemical equilibrium.

In the case of overall reactions the relationship between forward and reverse rate constants and the equilibrium constant given above for elementary reactions must be replaced by the equation

$$k_f/k_r=K(T)^n, \tag{24}$$

where n is a number that may have any positive value (Denbigh, 1964). This relationship holds for overall reactions that have rate expressions of the form $r=k_f(T)\Pi[A_i]^{\alpha_i}-k_r(T)\Pi[A_k]^{\alpha_k}$ where subscript i is for reactants and subscript k is for products. For empirical rate expressions having other forms this expression should not be used.

2. POTENTIAL ENERGY

At the most fundamental level, one can attempt to understand the factors that control chemical reactivity and energy transfer in terms of the forces between atoms and molecules. Since the force **F** as a function of interparticle separation **r** is given by

$$\mathbf{F}(r)=-dV(\mathbf{r})/d\mathbf{r}, \tag{25}$$

the problem becomes that of determining the potential energy $V(r)$. The rate coefficients of elementary reactions and the microscopic rate coefficients of state-to-state processes depend upon the potential-energy function. For reactions involving ions it is normally found that long-range forces dominate. Interactions between ions can be described by the Coulomb potential

$$V(r)=z_1z_2e^2/r, \tag{26}$$

while for ion-molecule reactions there are two cases, ion-induced dipole interactions, where the potential for the interaction with a neutral molecule for atom) having polarizability α is

$$V(r)=-\alpha(ze)^2/2r^4, \tag{27}$$

and ion-dipole interactions where the potential depends upon both the interparticle distance and the angle θ between the direction of the dipole, μ, and r,

$$V(r,\theta) = -ze\mu \cos\theta / r^2. \tag{28}$$

For interactions between neutrals, both long- and short-range forces have to be taken into account. To describe binary interactions empirical potentials have been extensively used to avoid the difficult problem of determining potentials from first principles. In the case of nonreactive collisions, the Lennard-Jones potential

$$V(r) = 4\epsilon[(\sigma/r)^{12} - (\sigma/r)^6] \tag{29}$$

provides a reasonably realistic representation. The r^{-6} term represents attraction due to dispersion forces, while the r^{-12} term gives a steeply rising short-range repulsion. The parameters ϵ and σ are defined in Fig. 3. The potential energy of diatomic molecules is frequently represented by the Morse function

$$V(r) = D_e[\exp(-2\beta r) - 2\exp(-\beta r)] \tag{30}$$

sketched in Fig. 4. The features of this function are a long-range attraction to describe the forces driving the formation of the chemical bond, and a short-range repulsion required by the Pauli exclusion principle. D_e is a dissociation energy (see Fig. 4), $r = R - R_e$ is the displacement from the potential minimum, and $\beta = \nu(\mu/2D_e)^{1/2}$, where ν is the frequency of the fundamental vibration, and μ is the reduced mass, $\mu = m_1 m_2/(m_1 + m_2)$.

For polyatomic molecules, the electronic potential is a function of more than one internuclear distance. It cannot be graphically represented in the plane of $V(r)$ and r, as for diatomics, but requires a space of higher dimensionality. For an N-atom polyatomic, $3N$ Cartesian coordinates are required to specify the position of each atom. The potential, however, depends upon the relative positions of the atoms, which for the arbitrary pair x and y are given by the vector $\mathbf{R}_{xy}$. Thus, only $N-1$ vector distances and $3N-3$ Cartesian coordinates are required, three for each vector, the remaining three coordinates fixing the position of the molecule in space, which is of no interest with respect to the potential.

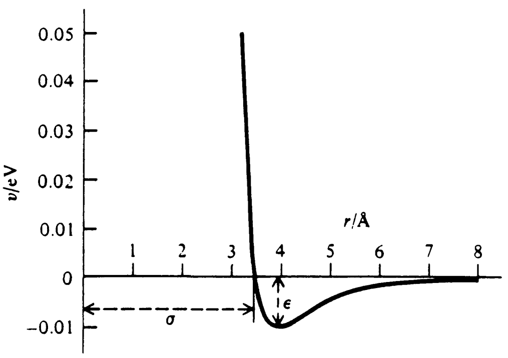

FIG. 3. The Lennard-Jones potential for argon. [Reprinted from Levine (1983), with permission of McGraw-Hill, Inc.]

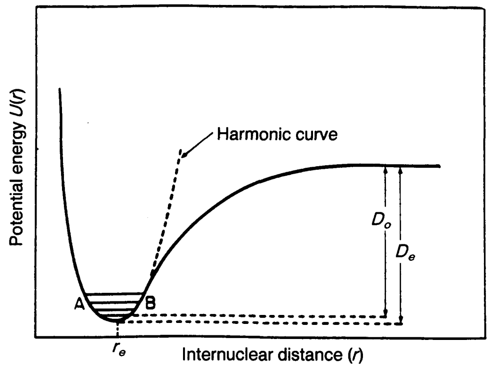

FIG. 4. The potential-energy curve for a diatomic molecule. [Reproduced from Robinson and Holbrook (1972), p. 8, with permission of the copyright holder.]

As an illustrative example we will consider a triatomic molecule, ABC. Choosing $\mathbf{R}_{AB}$ and $\mathbf{R}_{BC}$ specifies $\mathbf{R}_{AC}$ since $\mathbf{R}_{AC} = \mathbf{R}_{AB} - \mathbf{R}_{BC}$, and six coordinates are required. An alternative description would be to specify three scalar distances, either R_{AB}, R_{BC}, and R_{AC}, or R_{AB}, R_{BC}, and the bond angle Θ, as shown in Fig. 5, and three coordinates for the orientation of the plane containing atoms A, B, and C. Since the potential is independent of orientation, a graphical representation requires only four dimensions. Most mortals have difficulty visualizing four dimensions, so the representation is simplified by keeping one variable, usually the bond angle, constant. Then the potential can be presented as a surface which is a

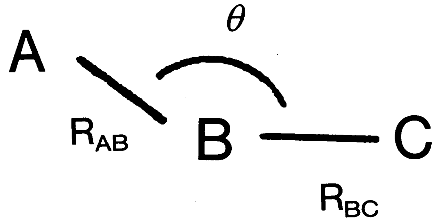

FIG. 5. Triatomic molecule A-B-C, with bond lengths R_{AB}, R_{BC}, and bond angle Θ.

function of R_{AB} and R_{BC} at a specified bond angle. Figure 6 shows a projection of such a surface onto the R_{AB}–R_{BC} plane, with contour lines of constant electronic potential. The minimum in the lower left-hand corner is the stable configuration for the bond angle, and the plateau in the upper right-hand corner gives the energy of the separated atoms. Starting at the minimum and moving parallel to either R_{AB} or R_{BC} describes dissociation of the triatomic molecule into an atom and a diatomic molecule. The trace of the surface on a cut perpendicular to the R_{AB} axis at large R_{AB} gives the potential curve for diatomic molecule *B–C*. A cut perpendicular to R_{BC} gives the *A–B* potential.

In a chemical reaction between an atom and a diatomic molecule,

$$A+BC=AB+C,$$

the triatomic molecule *ABC* formed by the close interaction resulting in reaction is unstable. This can be understood by examining the potential-energy surface, which can be described by using the same coordinates as for the stable triatomic molecule, but which has a different shape. Figure 7 shows a typical surface for this type of reaction. It can be seen that the minimum of Fig. 6 is replaced by a feature that resembles a mountain pass. A trace starting in the valley floor at either large R_{AB} or R_{BC} and proceeding along the minimum potential passes through a maximum which clearly shows the instability of *ABC*, which may dissociate into an atom and a diatomic molecule in either direction. The trace is a reaction path along the minimum potential energy. The potential-energy barrier is related to the activation energy, and plays an important role in determining the magnitude of the rate coefficient. Motion on the surface that is orthogonal to the reaction path corresponds to a bound system, as the potential energy increases in either direction. It can be seen that motion orthogonal to the reaction path corresponds to the symmetric stretch vibration, while motion along the reaction path would correspond to the antisymmetric stretch in a bound system. The bending vibration is not represented, since the bond angle is fixed.

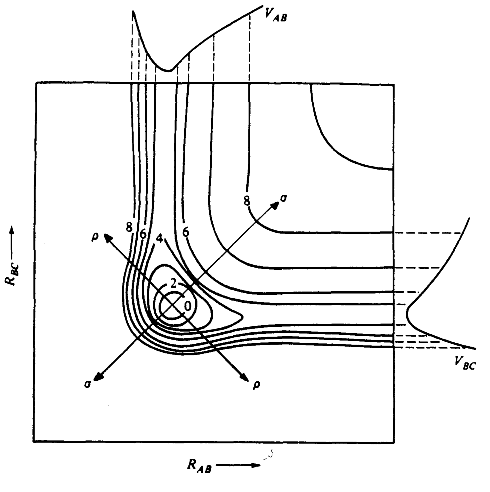

FIG. 6. The potential-energy surface for a stable, linear, triatomic molecule. Motion along directions ρ and σ represents asymmetric and symmetric stretch vibrations, respectively. (Weston and Schwarz, 1972, p. 34. Reprinted by permission of Prentice-Hall, Englewood Cliffs, NJ.)

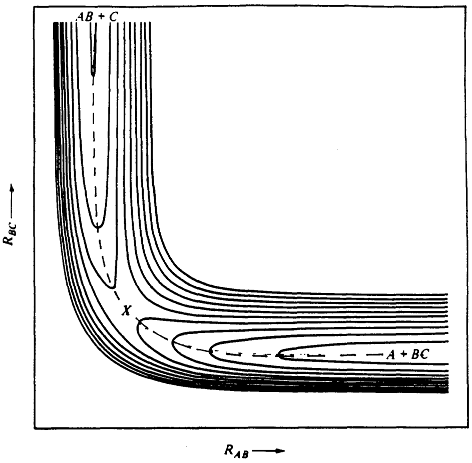

FIG. 7. Potential-energy surface for the reaction $AB+C \rightarrow A+BC$. The reaction path is indicated by the dashed line. The maximum along the reaction path is indicated by the *X* at the saddle point. (Weston and Schwarz, 1972, p. 36. Reprinted by permission of Prentice-Hall, Englewood Cliffs, NJ.)

A useful construction is the reaction-path profile, shown in Fig. 8. This is essentially an energy-vs-distance projection of the trace of the minimum-energy path for the *ABC* reaction, although the distance cannot be interpreted in terms of specific bond lengths, but should be seen as a somewhat vaguely defined reaction coordinate. The reaction-path profile is a convenient illustration of how potential energy changes as reaction proceeds. It is also frequently used to represent more complex polyatomic reactions, although in these cases

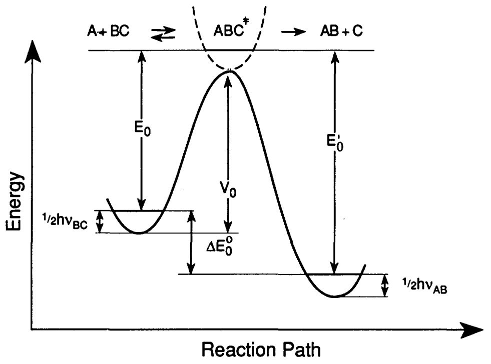

FIG. 8. Reaction-path profile corresponding to a surface of the type shown in Fig. 6.

the interpretation is strictly qualitative since the reaction-path profile exists on a hypersurface.

Electronic potentials for polyatomic molecules which have not been constrained for presentation in three dimensions require higher dimensionality, and are termed potential-energy hypersurfaces. There will be many hypersurfaces, each corresponding to an electronic state of the molecule, or an electronic state of the reacting system. Reactions may occur on a single hypersurface, in which case they are said to be electronically adiabatic, or they may occur with transitions between hypersurfaces (nonadiabatic). Chemical reactions most commonly occur on the ground-state hypersurface, although in excited-state kinetics, photochemistry, and nonadiabatic reactions, electronically excited hypersurfaces must be considered.

The electronic potential hypersurface plays a critical role in determining the rate and mechanism of a bimolecular reaction, as is apparent from the triatomic example above. It is enormously useful to know the surface both for understanding how the reaction proceeds, and for *a priori* calculation of rate parameters. While information about the potential for stable molecules may be experimentally obtained from spectroscopy, this has not been feasible in the past for reactions since the number of species in the midst of reacting at any instant is usually extremely small, and undetectable by suitable spectroscopies. This situation is currently changing, however, due to the application of sensitive laser-based spectroscopies. For example, a simple procedure for the inversion of femtosecond temporal spectra to yield the potential has recently been described (Bernstein and Zewail, 1990).

The determination of potential-energy hypersurfaces is most commonly approached by quantum-mechanical calculation. The *ab initio* approach requires solution of Schrödinger's equation

$$\mathscr{H}\psi=E\psi, \tag{31}$$

where the electronic potential is in the Hamiltonian operator,

$$\mathscr{H}=\sum -\hbar^2/2m_i\nabla_i^2+V(\mathbf{R}_{12},\ldots), \tag{32}$$

and it is convenient to employ the Born-Oppenheimer principle to separate electronic and nuclear motion,

$$\psi=\psi_e(\mathbf{r},\mathbf{R})\psi_n(\mathbf{R}). \tag{33}$$

Since exact solutions are not possible for the polyatomic systems of interest, variational methods, employing trial wave functions, have been developed extensively (Hirst, 1985). The experience of years of work on this problem has resulted in computational methods that usually give accurate structural information, but the situation with respect to energetics is less satisfactory. The self-consistent-field method is known to give unreliable energetics, while multiconfiguration and configuration-interaction approaches can give activation energies of simple reactions that are in error by only a few kJ/mol. There are presently surfaces of high accuracy for the reaction $H+H_2=H_2+H$ (Liu, 1973; Siegbahn and Liu, 1978; Truhlar and Horowitz, 1978), which is the simplest bimolecular reaction between neutral species in the sense that it involves only three protons and three electrons. Potential-energy surfaces have been calculated for many other reactions. A recent summary of progress over the last 40 years (Karplus, 1990) noted advances in the power of computers and in calculational methodology, so that "...It is now possible to do *ab initio* calculations that are of sufficient accuracy to answer chemically important questions for reasonably complex molecules and for potential surfaces describing their reactions."

The difficulty of *ab initio* calculations led to the development of semiempirical and empirical methods. Principal among these are the London-Eyring-Polanyi-Sato (LEPS) (Sato, 1955) and bond-energy–bond-order (BEBO) (Johnston, 1966) methods. The former is based upon the London equation for H_3, which in turn is based upon the Heitler-London theory for H_2. The latter is based upon conservation of bond order during a concerted bimolecular atom transfer reaction. Bond order n is defined by the relationship $r=r_s-0.26 \ln n$, where r is bond length, and r_s is the length of a single bond. The empiricism used in this approach does not come from kinetics, but from spectroscopy. Furthermore, BEBO does not give the complete surface, only the trace along the minimum potential-energy reaction path.

3. BIMOLECULAR COLLISION DYNAMICS

3.1 Simple Collision Theory

Since bimolecular reactions require an encounter between the two reactants, the upper limit of the reaction rate in the gas phase is the collision rate. For reaction between species A and species B, each considered to be structureless, spherical masses interacting according to the "hard-sphere" potential [$V(r)=0$, $r>d_{AB}$; $V(r)=\infty$, $r=d_{AB}$], the collision rate is

$$Z_{AB}=\pi d_{AB}^2(8kT/\pi\mu)^{1/2}n_A n_B, \tag{34}$$

and for reaction between two A species it is

$$Z_{AA}=\pi d_{AA}^2(4kT/\pi m)^{1/2}n_A^2. \tag{35}$$

The collision diameters are $d_{AB}=(d_A+d_B)/2$ and $d_{AA}=d_A$, where d_A and d_B are the respective diameters of species A and B, μ is the reduced mass of the collision pair, n_A and n_B are number densities, k is Boltzmann's constant, and T is the absolute temperature. The diameters d_A and d_B are normally taken from determinations via gas viscosity measurements, or other transport properties. The appropriateness of using this kind of data for reactive collisions is questionable, since the interactions leading to reaction might reasonably be expected to be "closer" than the physical interactions governing transport properties. The derivation of these equations can be found in any textbook of physical chemistry, and most chemical kinetics texts.

The collision rates can be expressed in conventional kinetics notation,

$$-dn_A/dt=k_{\text{coll}}n_A n_B \tag{36}$$

and the association $k_{\text{coll}}=\pi d_{AB}^2(8kT/\pi\mu)^{1/2}$ made. If $d_{AB}=4\times10^{-8}$ cm, and $\langle v\rangle=(8kT/\pi\mu)^{1/2}=4\times10^5$ cm s^{-1}, then $k_{\text{coll}}=2.5\times10^{-10}$ cm^3 molecule^{-1} s^{-1}. If this is taken as the bimolecular reaction rate constant, then at a pressure of one atmosphere, the one-half time (time taken for one-half of the reactants to be converted into products) is only 3×10^{-10} s! It is worth noting that πd_{AB}^2 and πd_{AA}^2 are collision cross sections, σ, so that $k_{\text{coll}}=\sigma\langle v\rangle$.

Of course, k_{coll} is the upper limit of the reaction rate constant because any potential barrier that must be surmounted will cause the reaction to be slower, and because there may be orientational restrictions. For example, a real molecule is not a sphere, but has chemical bonds that are directional. Some orientations of A and B as they collide may be favorable for the breaking and forming of bonds (reaction), while others may be unfavorable, and result in no reaction. The energetic requirement is usually treated by assuming that the probability of reaction upon collision is zero if the relative energy of the reactants along the line of centers is less than a critical energy, E_c, while the reaction probability is unity for all energies greater than E_c. Since the probability of having $E>E_c$ is given by $\exp(-E_c/RT)$, and a simple correction for orientation can be made by incorporating a temperature independent factor $p<1$, the simple collision theory (SCT) bimolecular rate constant is

$$k_{\text{SCT}}=k_{\text{coll}}p\exp(-E_c/RT). \tag{37}$$

Since $E_a=Rd\ln k(T)/d(1/T)$, according to SCT $E_a=E_c+\frac{1}{2}RT$ and the Arrhenius preexponential factor is $A=p(8\pi kT/\mu)^{1/2}d^2_{AB}\,e^{1/2}$. To predict k_{SCT} a value of E_c is needed. It is reasonable to obtain E_c from V_0 plus zero-point energy, but we have seen that calculation of potentials is difficult and frequently inaccurate. The common procedure has been to make comparisons of A_{SCT} with A_{expt} when an experimental value of E_a is available. Taking $p=1$, and typical collision diameter, A_{SCT} is of order 10^{-10} cm^3 molecule^{-1} s^{-1}, with very little variability with size or structure of the reactants. Values of A_{expt}, on the other hand, while in reasonable agreement with A_{SCT} for some cases, are most frequently smaller, sometimes by orders of magnitude. While some of this may be accounted for if the collision diameters from transport properties are too large to describe reactions, the correct diameters to use are generally not available. A modified SCT, in which orientation and collision energetics are taken into account, has been presented (Smith, 1980). This predicts probability factors p between 0.1 and 0.01, but fails to account for the smaller p's, 0.001 or less, that are frequently required to reconcile SCT with experiment.

The simple collision theory is an early treatment that contains a number of assumptions and approximations. Among these are the oversimplified interaction potential, no provision for nonadiabatic reactions, and the position that only energy of relative translation along the line of centers is important. In view of these, its inadequacy is perhaps not surprising. In the 1930s transition-state theory, a statistical theory of reaction rates, was developed and became highly successful. Collision dynamics was largely ignored in chemical kinetics, in spite of the physically satisfying picture of reactions it makes possible, until the 1950s when experiments on reactive scattering in crossed molecular beams were initiated.

3.2 Collision Dynamics

Most of the underlying concepts in modern approaches to collision dynamics can be understood from the classical treatment of collisions between structureless particles with spherically symmetric interaction potentials. Although this situation can only describe elastic scattering, and is only applicable to two-body collisions of atoms, it yields analytical solutions that provide a basis for understanding inelastic and reactive scattering of more complex species. For a specified potential, $V(r)$, the outcome of an elastic collision depends only upon the initial relative velocity g and the impact parameter b. The collision can be represented as the motion of a particle of mass μ, the reduced mass of the collision pair, with respect to a fixed scattering center. The impact parameter is the closest approach of μ to the scattering center that would occur in the absence of the potential. Figure 9 sketches this situation for a potential that is attractive at long range and repulsive at short range, for example, the Lennard-Jones potential. The deflection of the trajectory, χ, is given by

$$\begin{aligned}\chi &= \pi - 2\theta_m \\ &= \pi - 2b \int_{r_m}^{\infty} r^{-2}\left(1-\frac{b^2}{r^2}+\frac{V(r)}{E_0}\right)^{-1/2} dr, \end{aligned} \tag{38}$$

where $E_0 = (1/2)\mu g^2$. By considering the flux of particles scattered into solid angle $d\Omega$ as a function of b, a differential scattering cross section $\sigma(E,\chi)$ can be derived,

$$\sigma(E,x) = \frac{b}{\sin\chi}\left|\frac{db}{d\chi}\right| . \tag{39}$$

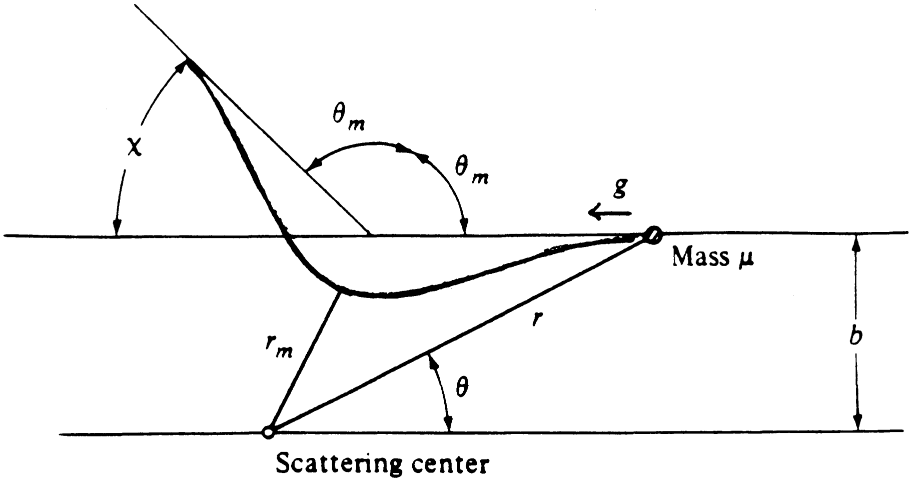

FIG. 9. Representation of a binary collision as the center-of-mass motion of a hypothetical particle of mass μ with respect to a fixed scattering center. (Weston and Schwarz, 1972, p. 47. Reprinted by permission of Prentice-Hall, Englewood Cliffs, NJ.)

By integrating over the deflection χ, the total scattering cross section can be obtained,

$$\sigma(E)=2\pi\int_0^\pi \sigma(E,\chi)\sin\chi\, d\chi. \tag{40}$$

Since the long-range part of the potential goes asymptotically to zero as $b\rightarrow\infty$, $\chi(b)\rightarrow 0$, and $\sin\chi\rightarrow 0$, so $\sigma(E,\chi)$ [and also $\sigma(E)$]$\rightarrow\infty$. This shortcoming of the classical treatment is removed by quantum mechanics, which predicts finite cross sections due to interference effects when the motion of the scattering particles is treated by wave equations. There are other unsatisfactory aspects of the classical picture of scattering which are also taken care of by quantum mechanics (Levine and Bernstein, 1987). In practice, there will be some minimum experimentally measurable deflection that is determined by the apparatus. The classical treatment can be adapted by setting this as a cutoff value of χ in the cross-section equations. Both $\sigma(E,\chi)$ and $\sigma(E)$ then become finite. The total cross section can then be expressed in terms of the impact parameter,

$$\sigma(E)=2\pi\int_0^{b_{\max}} b\, db, \tag{41}$$

where $b_{\max}$ is the largest value of b that produces a detectable deflection.

Another aspect of the elastic scattering picture that turns out to be useful in describing association reactions without a potential barrier comes out of considerations of angular momentum conservation. The centrifugal energy associated with a pair of particles can prevent them from having a close collision for large impact parameters. The centrifugal energy acts as a low repulsive barrier that is placed at relatively long range. The effective potential seen by the pair is then the sum of the electronic potential and the centrifugal contribution,

$$V_{\text{eff}}=V(r)+\tfrac{1}{2}\mu b^2g^2/r^2. \tag{42}$$

The effective potential is sketched in Fig. 10 for a specific b and g. A family of V_{eff} curves exists, depending on b and g. The long-range maximum has been used in some models to locate the position of the transition state in association reactions occurring without an electronic barrier. Note that the position of V_{eff} is energy dependent.

It is worthwhile to examine further the connection between cross sections and rate constants at this point. The hard-sphere collision-rate coefficient, shown above to be equal to the product of the cross section πd^2 and the average speed, can be readily generalized to any collision cross section. Since collision cross sections are found to be functions of the relative speed g, a microscopic rate coefficient may be expressed for collisions with specified relative speed,

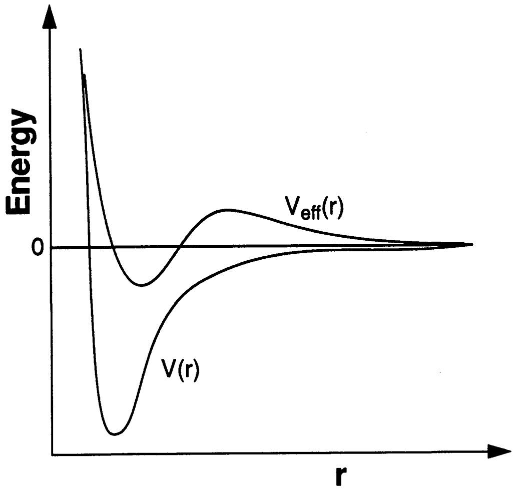

FIG. 10. The effective potential compared with the electronic potential for a binary collision.

$$k_{\text{coll}}(g)=\sigma(g)g, \tag{43}$$

and the thermal collision rate coefficient is obtained by thermally averaging over all relative speeds, $P(g)$ being the Maxwell-Boltzmann distribution:

$$k_{\text{coll}}=\int_0^\infty \sigma(g)gP(g)dg. \tag{44}$$

We are interested in reactive collisions, not just the total collision rate, which lumps together elastic, inelastic, and reactive collisions. The thermal rate constant can be obtained if $\sigma(g)$ is replaced with a reactive cross section, $\sigma_r(g)$:

$$k(T)=\int_0^\infty \sigma_r(g)gP(g)dg. \tag{45}$$

Although σ_r is taken here to be a function only of g it may also depend upon vibrational, rotational, or electronic energy. The reactive cross section is of central importance in any collision theory treatment of bimolecular reactions. Although at first glance it appears that cross sections could be determined by deconvolution of the integral in Eq. (45) if measurements of $k(T)$ were available, Fig. 11 illustrates that only limited information on cross sections can be obtained from thermal kinetics data. The product $gP(g)$ is plotted vs g in the figure, along with an assumed, but realistic, cross section that is a monotonically increasing function of energy, and has a threshold, ϵ_c, below which the cross section is zero. It can be seen that the integrand, $\sigma_r(g)gP(g)$, is only able to sample the cross section in the vicinity of the threshold because of the rapidity with which $P(g)$ dies out. Reactive cross sections are obtained from special experiments, particularly from scattering in crossed molecular beams, which is discussed in Sec. 7.

3.3. Classical Trajectory Calculations

Further development of collision theories aimed at deducing analytical expressions for cross sections and rate coefficients has been usurped by the advent of classical trajectory calculations. This method uses the classical equations of motion, Hamilton's equations, to calculate classical collision trajectories. Implementation of this approach was facilitated by the increasing availability of high-speed digital computers, particularly because many trajectories must be obtained to calculate cross sections and rate coefficients, which are statistically averaged quantities. The validity of this approach depends upon whether classical mechanics provides an adequate representation of what, strictly speaking, is a quantum phenomenon. There is a body of evidence supporting the position that classical trajectories give satisfactory results for many reactions. The limitations are being explored by quantum-mechanical scattering calculations.

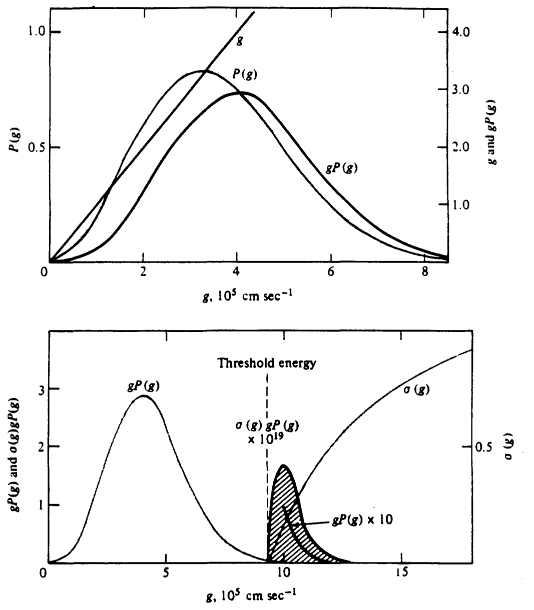

FIG. 11. Factors contributing to the rate coefficient given by the equation $k(T)=\int gP(g)\sigma(g)dg$. The rate constant is given by the shaded area. The rapid decrease of $gP(g)$ with increasing g leads to small contributions to $k(T)$ from high-velocity collisions, making it difficult to obtain information on the cross section from thermal rate data. (Weston and Schwarz, 1972, p. 86. Reprinted by permission of Prentice-Hall, Englewood Cliffs, NJ.)

The quantities required to perform a trajectory calculation are the potential, the impact parameter, the initial kinetic energy, the vibrational energy and phase, and the rotational energy and orientation (rotational phase). Of these, the potential is of central importance. The difficulties of calculating an accurate potential-energy hypersurface have already been discussed. In a few cases good *ab initio* surfaces are available. In other cases approximate *ab initio* or semiempirical surfaces have to be used. If necessary, the potential barrier may be scaled to provide agreement with the experimental activation energy. When a surface has been selected, it must be fitted by an analytical function to obtain V as a function of position coordinates.

The next step is to choose the initial conditions for each trajectory. These are the values of b, E_0, and the internal energies. Monte Carlo procedures are used for this selection in order to approximate closely the correct statistical distribution of initial states within a relatively small set of trajectories. The minimum number of trajectories, N, that must be solved for the results to have statistical significance is of the order of 1000 or more. Examination of the trajectories shows which ones result in chemical reaction and which ones result in elastic or inelastic scattering. Figure 12 shows a reactive trajectory for an $H+H_2$ collision. The probability of reaction P_r, as a function of initial conditions, can then be obtained,

$$P_r(E_0,J,v,b)=\lim_{N\to\infty}\frac{N_r(E_0,J,v,b)}{N(E_0,J,v,b)}\,. \tag{46}$$

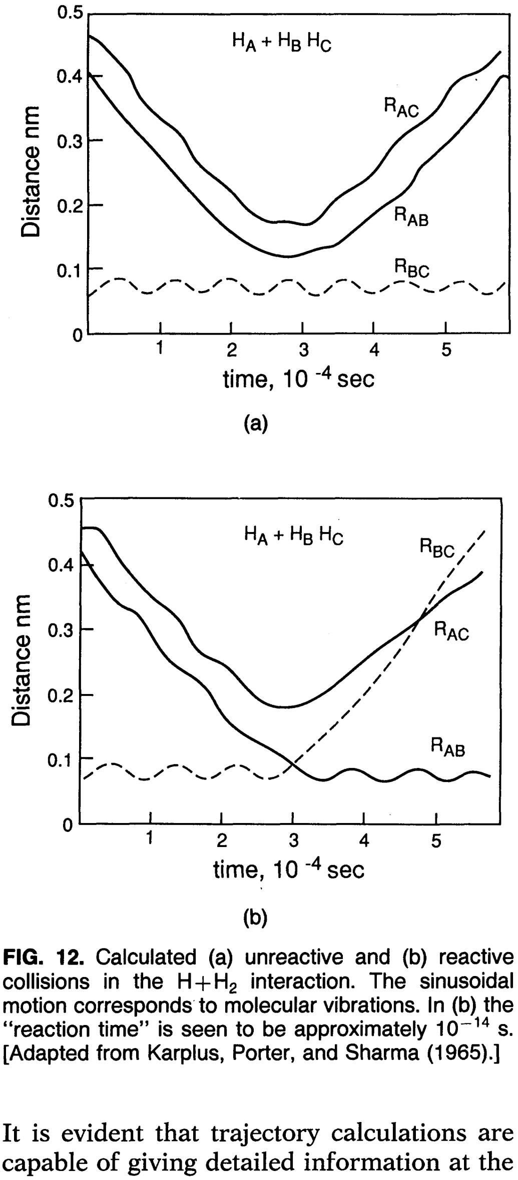

(a)

(b)

FIG. 12. Calculated (a) unreactive and (b) reactive collisions in the $H+H_2$ interaction. The sinusoidal motion corresponds to molecular vibrations. In (b) the "reaction time" is seen to be approximately 10^{-14} s. [Adapted from Karplus, Porter, and Sharma (1965).]

It is evident that trajectory calculations are capable of giving detailed information at the microscopic, or state-to-state, level. Reactive cross sections can be obtained by integrating over the impact parameters,

$$\sigma_r(E_0,J,v)=2\pi \int_0^{b_{\max}} P_r(E_0,J,v)b\, db, \qquad (47)$$

and microscopic thermal rate coefficients for individual vibration-rotation levels can be calculated by averaging over the Maxwell-Boltzmann distribution,

$$k(J,v)=\int_0^{\infty} \sigma(J,v)E_0\, f(E_0)dE_0. \qquad (48)$$

Finally, the thermal rate coefficient for the elementary reaction can be obtained by summing over the individual vibration-rotation states,

$$k(T)=\sum_{J,v} k(J,v). \qquad (49)$$

4. TRANSITION-STATE THEORY

4.1 Conventional Transition-State Theory

The transition-state theory (TST) was developed in the 1930s, primarily by Eyring and co-workers. A good account of the early work can be found in the book *Theory of Rate Processes* (Glasstone *et al.*, 1941). It is a statistical theory, using concepts of statistical thermodynamics to obtain an expression for the rate coefficient of an elementary reaction, and has been successfully applied to many different types of chemical reactions. The theory can be conveniently thought of as the solution to a barrier-crossing problem, where the barrier is an electronic potential maximum on the reaction path, as described in the section on potential energy. The reacting species that exists at the barrier maximum is called the transition state. The rate of crossing of the barrier can be calculated from the product of the concentration of transition states, $[A_i^{\ddagger}]$, the frequency of crossing the barrier, ν^*, and a transmission coefficient, κ :

$$r=\kappa\nu^*[A_i^{\ddagger}]. \qquad (50)$$

The derivation of the expression for the rate coefficient, which can be found in any textbook on kinetics, leads to the expression

$$k(T)=\kappa(kT/h)\frac{Q^{\ddagger}/V}{\Pi Q_i/V}\exp(-\epsilon_0/kT). \qquad (51)$$

In this equation, k is Boltzman's constant, h is Planck's constant, Q/V is the molecular partition function, and ϵ_0 is the difference in zero-point energy between the transition state and the reactants. The product in the denominator runs over all the reactant species, which in practice is usually either one or two, and the superscript double dagger (‡) denotes the transition-state. The theory assumes that there is a dividing surface between reactants and products, which locates the transition state, and beyond which there is no turning back to reactants. That is, all reactants that reach the transition-state configuration cannot return to the reactant side of the barrier, but must continue on to form products. Since some fraction is expected to turn back, TST gives an upper limit to the rate constant. Conventional TST locates the dividing surface at the barrier maximum. The reacting system is further assumed to be in thermal equilibrium. If the reaction rate is fast enough to deplete the population of reactant states above the barrier and cause deviations from a Boltzmann distribution, then the statistical arguments which lead to the incorporation of the partition functions become invalid. As a rule of thumb, if $E_a/RT \geqslant 10$, deviations from thermal equilibrium will not be severe enough to cause a problem. Finally, it is assumed that motion along the reaction path is separable. This permits factoring out a one-dimensional translational partition function that corresponds to motion of separating fragments away from one another. Thus $Q^{\ddagger}$ has one less internal degree of freedom than a stable molecule.

To evaluate the TST expression, it is necessary to calculate the partition functions. The reactants present no problem, for the usual assumption of separability into translational, rotational, and vibrational degrees of freedom is normally satisfactory. If the vibrational frequencies, moments of inertia, and electronic states are known, the computation is readily done. The evaluation of $Q^{\ddagger}$ is problematic, however. The electronic potential in the region of the transition state is only available for a few simple reactions, and we have seen that it is difficult to obtain. Furthermore, spectroscopic information is not available. Since the requisite molecular parameters are not generally calculable or measurable, they must be estimated by comparison with stable molecules. This provides sufficient latitude that agreement between theory and experiment can usually be easily obtained. It has been pointed out that since the lifetime of a transition state may be shorter than characteristic times for vibration and rotation, it is not clear whether its internal motions can be treated the same way that stationary states of stable molecules are treated (Smith, 1980). This is an issue that has not yet been addressed. Finally, TST does not provide for determination of ϵ_0, which must be obtained from other sources, either a potential surface or an experimental activation energy. (It is important to realize that ϵ_0 is not equivalent to E_a, or to V_0, but that these quantities are closely related.)

4.2 Thermodynamic Formulation

An alternative formulation of TST employs thermodynamic relationships, leading to the expressions

$$k(T)=\kappa e[kT/h(c^0)^{-\Delta\nu}]\exp(\Delta S^{0\ddagger}/R) \times\exp(-E_a/RT), \quad (52)$$

$$k(T)=\kappa[kT/h(c^0)^{-\Delta\nu}]\exp(1-\Delta n^{\ddagger}) \times\exp(\Delta S^{0\ddagger}/R)\exp(-E_a/RT), \quad (53)$$

the first of which applies to reactions in liquid solutions, and the second to reactions in the gas phase. In these equations, c^0 is a standard state concentration, conventionally taken to be 1 mol/L (solutions) or 1 molecule/cm^3 (gas phase), e is the base of natural logarithms, $\Delta n^{\ddagger}$ is 1 − (number of reactant molecules), and $\Delta S^{0\ddagger}$ is the entropy of activation, the difference between the entropy of the transition state and the entropy of the reactants. It can be seen that these are in the form of the Arrhenius equation, except that the coefficient of the $\exp(-E_a/RT)$ factor, which can be associated with the Arrhenius preexponential factor, is temperature dependent. It is common practice to ignore this, and make direct comparisons between theory and experiment whenever experimental values of $k(T)$ are available. Application of Eqs. (52) and (53) then allows calculation of $\Delta S^{0\ddagger}$. The numerical values of $\Delta S^{0\ddagger}$, which are found to encompass both negative and positive values, can be used as a guide to the general structural features of the transition state by the use of third-law molecular entropies. Johnston (1966) has summarized these ideas. It should be emphasized that a unique transition-state structure cannot be obtained, since a wide range of reasonable vibration frequencies and moments of inertia can be found that will fit a given $\Delta S^{0\ddagger}$.

4.3 The Transmission Coefficient

The transmission coefficient is a factor to correct for quantum-mechanical effects. For reactants that have total energies less than the classical barrier height, there is a probability that reaction may occur by tunneling, and if there is a surface of another electronic state nearby, there is a probability that reaction can occur by transition to that surface (nonadiabatic reaction). When tunneling is important, $\kappa > 1$, and when a nonadiabatic reaction occurs, $\kappa < 1$.

Analytical expressions for the contribution of tunneling to the transmission coefficient have been obtained for model barriers, such as the Eckert potential and the truncated parabola. For barriers of arbitrary shape, numerical calculations must be done. These are one-dimensional calculations, however, and tunneling is not a one-dimensional phenomenon. This is apparent from inspection of Fig. 6, which makes it clear that reaction paths other than the one of minimum energy will pass over higher and wider parts of the surface. Johnston (1966) has discussed this problem and tabulated some results. More recent multidimensional tunneling calculations have been done (Marcus and Coltrin, 1977), but they involve considerable effort. Tunneling will be important for light atoms, low temperatures, and thin, high barriers. In kinetics, tunneling usually only needs to be considered in reactions where the reaction coordinate involves motion of a H atom.

Electronically nonadiabatic reactions usually occur at the avoided crossing of two potential-energy curves (surfaces). The Landau-Zener model (Eyring *et al.*, 1944) gives the probability of transition between the two states in terms of the minimum-energy gap between them. It is capable of qualitatively explaining the low Arrhenius preexponential factors associated with reactions in which electron spin is not conserved, such as

$$N_2O(^1\Sigma^+) \rightleftharpoons N_2(^1\Sigma_g^+) + O(^3P),$$

which has a preexponential factor of about $10^{-11}\ s^{-1}$. The expected magnitude for a fully spin-allowed reaction would be $\geqslant 10^{13}\ s^{-1}$. The Landau-Zener treatment is a one-dimensional, classical model, and should not be considered accurate. Nikitin (1974) has given a treatment of nonadiabatic processes in chemical reactions.

4.4 Variational Transition-State Theory

As a result of dividing-surface recrossings, conventional TST leads to rate coefficients that are too large. The approach that has been taken to this problem is to relocate the dividing surface, and therefore the transition state, in a region where recrossings are minimal, since in general it may not be possible to find a location where recrossings are absent. Procedures for accomplishing this are classified as variational transition-state theories (VTST) (Truhlar and Garrett, 1980), since they have in common the minimization of some quantity. Canonical VTST minimizes the rate coefficient,

$$dk(T)/dq^\ddagger = 0, \tag{54}$$

where $q^\ddagger$ is the coordinate corresponding to the reaction path. This maximizes the Gibbs free energy. Microcanonical VTST minimizes the microcanonical rate coefficients, $k(E)$, since the number of recrossings at a given dividing surface location is most likely energy dependent:

$$dk(E)/dq^\ddagger = 0. \tag{55}$$

This defines a different transition state for each energy, and is expected to be an improvement over canonical VTST, which ignores the energy dependence of the recrossing.

5. UNIMOLECULAR REACTIONS

5.1 The Lindemann Mechanism

A unimolecular reaction is a spontaneous chemical change of an isolated molecule. The early history of unimolecular reactions centered on the question of how molecules obtain the requisite energy for spontaneous reaction, since it was thought that if collisions were responsible, the reaction would be bimolecular. The radiation hypothesis (Perrin, 1919), which claimed that absorption of blackbody radiation was the activation mechanism, was shown to be incorrect. F. A. Lindemann, at a 1922 meeting of the Faraday Society, presented a mechanism showing how collisional activation could account for first-order kinetics (Lindemann, 1922). The Lindemann mechanism is still the underlying basis for understanding unimolecular reaction kinetics. It is expressed by

$$A + M \underset{k_d}{\overset{k_a}{\rightleftharpoons}} A^* + M,$$

$$A^* \xrightarrow{k_E} \text{products}.$$

Reactant molecules A become activated by collisions with other species, or another A, here lumped together as M. The activated molecules A^* either react to form products, or become collisionally deactivated. The mechanism is an oversimplification, for it treats activation and deactivation as a one-step process, called "strong collisions," whereas multiple collisions, termed "weak collisions," are often required for an accurate description of unimolecular reactions. Also, the rate coefficient k_E for the unimolecular reaction is a strong function of energy. The steady-state approximation, applied to $[A^*]$, leads to the rate expression

$$\frac{d[P]}{dt}=\frac{k_E k_a[M][A]}{k_d[M]+k_E}=k_{\text{uni}}[A], \qquad (56)$$

and the experimentally observed first-order rate coefficient k_{uni} may be associated with $k_E k_a[M]/(k_d[M]+k_E)$, which correctly predicts that the unimolecular rate constant depends upon $[M]$, or equivalently, total pressure. When $k_d[M] \gg k_E$, the rate coefficient becomes independent of pressure, $k_{\text{uni}} \rightarrow k_E \times k_a/k_d \equiv k_\infty$, where k_∞ is the limiting high-pressure rate coefficient. At low pressures, when $k_d[M] \ll k_E$, $k_{\text{uni}} \rightarrow k_a[M] \equiv k_0$, k_{uni} becomes proportional to pressure at the low-pressure limit, and the kinetics become second order. A sketch of the pressure dependence is shown in Fig. 13, which is called a falloff plot.

If the energy dependence of k_E is taken into account, then the general expression for k_{uni} must be summed over all energies. Vibrational and, to a certain extent, rotational energy are responsible for unimolecular reactions. In molecules with enough energy to break bonds, vibrational states usually form a virtual continuum, so that summation can be replaced by integration, and the Lindemann expression for k_{uni} is replaced by

$$k_{\text{uni}}=\omega \int_{E=E_c}^{E=\infty} \frac{k_E P(E)dE}{\omega+k_E}. \qquad (57)$$

In this equation ω is the collision frequency $(\omega=k_d[M])$, $P(E)$ is the energy distribution function of the reactants, normally taken to be a Boltzmann distribution, and k_E is referred to as the microcanonical rate coefficient. The lower limit of the integral is the critical energy, a threshold energy below which reaction cannot occur. The strong-collision assumption is carried forward into this expression. Weak collisions can be empirically accounted for by replacing ω with $\beta_c\omega$, where β_c is a collision efficiency $0<\beta_c\leqslant 1$. Collisional energy transfer in unimolecular reactions has been reviewed (Tardy and Rabinovitch, 1977; Quack and Troe, 1981; Oref and Tardy, 1990).

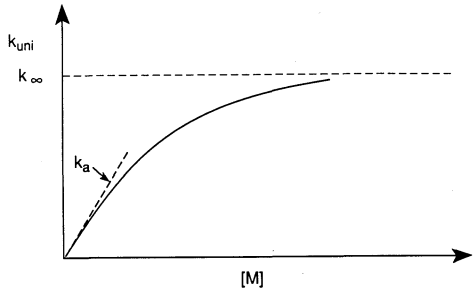

FIG. 13. Pressure dependence of k_{uni} according to Eq. (56).

5.2 Theoretical Models

Equation (57) and its modification to account for collisional inefficiency can be used for calculation of k_{uni} if a model for k_E is provided. The Rice-Ramsperger-Kassel (RRK) model, developed in the late 1920s and early 1930s (Kassel, 1932), is a statistical approach using combinational probabilities to describe randomization of vibrational energy among the s equal-frequency oscillators of the model reactant molecule. Reaction occurs when the reaction coordinate (one of the oscillators) contains energy $nh\nu$ greater than E_c. The model is successful in fitting experimental falloff data provided values of E_a and A are available. However, it is found that to obtain a satisfactory fit, s must be treated as a variable parameter, and its value usually turns out to be only one-half to two-thirds of the number of oscillators in the molecule. The model thus appears to give a contradiction between the assumption of energy randomization and the prediction that some of the oscillators are ineffective for intramolecular energy transfer.

A dynamical theory (Slater, 1959) assumes that reaction occurs when the harmonic normal-mode vibrations come into phase in such a way that the reaction coordinate extends by a critical amount. The theory was successful in application to cyclopropane isomerization, but was not able to account for some other reactions. For example, it predicts that the observed falloff curve for 1,2-dideuterocyclopropane should be shifted along the pressure axis with respect to cyclopropane, whereas no shift is found experimentally (Schlag and Rabinovitch, 1960).

The most successful unimolecular rate theory is the Rice-Ramsperger-Kassel-Marcus (RRKM) theory, an extension of RRK by incorporation of TST to replace the combinatorial statistics. These ideas (Marcus and Rice, 1951; Marcus, 1952), were further developed, primarily by Marcus and by Rabinovitch and co-workers. The RRKM theory is clearly laid out in two textbooks dealing with the theory of unimolecular reactions (Robinson and Holbrook, 1972; Forst, 1973). The book of Robinson and Holbrook is particularly useful as a tutorial in application of the theory. The incorporation of TST is made clear by the mechanism

$$A+M \underset{k_d}{\overset{k_a}{\rightleftarrows}} A^*(E)+M,$$

$$A^*(E) \underset{k_{-E}}{\overset{k_E}{\rightleftarrows}} A^\ddagger(E),$$

$$A^\ddagger(E) \xrightarrow{k_r} \text{products},$$

in which a distinction is made between activated molecules, $A^*(E)$, those that have enough energy to react, and transition states, $A^\ddagger(E)$, those that are at the critical configuration, or dividing surface. The energy dependence of the microcanonical rate coefficient is taken into account, but the strong-collision approximation is still made. Equation (57) can be used for k_{uni} with the appropriate expression for k_E and weak collisions can be taken into account by inclusion of a collision efficiency. The microcanonical rate coefficient in RRKM theory is given by

$$k_E = Q_i^\ddagger \sum G(E^\ddagger)/Q_i h N(E^*) \ , \qquad (58)$$

where Q_i are partition functions for inactive degrees of freedom of the transition state and the activated molecule, h is Planck's constant, $\Sigma G(E^\ddagger)$ is the total number of states of the active degrees of freedom of the transition state, and $N(E^*)$ is the internal state density of the activated molecules at energy E^*. Inactive degrees of freedom are those that do not couple with the reaction coordinate, and active degrees of freedom are those that exchange energy with it. Overall rotations are usually considered inactive, and taking all vibrational degrees of freedom to be active usually leads to satisfactory agreement between theory and experiment. Improved treatments of the adiabatic rotations are available (Waage and Rabinovitch, 1970).

The calculation of sums and densities of states can be done exactly on a digital computer by direct counting using the Beyer-Swinehart algorithm (Beyer and Swinehart, 1973; Stein and Rabinovitch, 1973). State sums and densities can be large numbers. For example, the relatively modestly sized molecule CH_3CN has about 10^6 states cm^{-1} at an energy of 350 kJ/mol. For computational efficiency an approximation due to Whitten and Rabinovitch (1963) is frequently used,

$$G(E^\ddagger) = (E+aE_z)^s \Big/ s! \prod h\nu_i \ . \qquad (59)$$

In Eq. (59), E is the total energy, E_z is the vibrational zero-point energy, a is an empirical correction factor that can be calculated from the vibrational frequencies, s is the number of vibrational degrees of freedom, and the product runs over all the normal vibrational frequencies. State densities at the same energy can be calculated by differentiation, $N(E^*) = dG(E^\ddagger)/dE$. This approximation is quite accurate when the nonfixed vibrational energy content is greater than an amount approximately equal to the zero-point energy, but can be appreciably in error at lower energies. In practice, direct count is used only up to energies where the Whitten-Rabinovitch approximation has acceptable accuracy.

Figure 14 shows an RRKM calculation for the thermal isomerization of methylcyclopropane compared with experiment (Topor and Carr, 1973). The calculation empoys the experimental activation energy, as the theory provides no guidance on energetics, and vibrational frequencies fitted to the experimental entropy of activation. The latter essentially provides a fit at the high-pressure limit, since it is readily shown that at the high-pressure limit, the RRKM expression for k_{uni} reduces to the TST rate coefficient. The vibrational frequency model is not unique, but falloff is

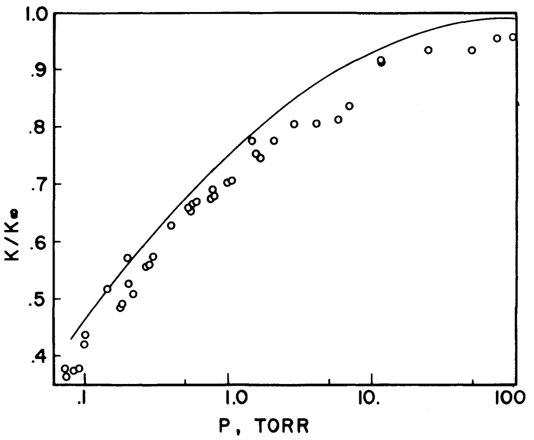

FIG. 14. Reduced (k/k_∞) unimolecular falloff plot for methylcyclopropane isomerization. Symbols are data points; the line is an RRKM calculation. Reprinted with permission from Topor and Carr (1973), p. 764.

insensitive to transition-state parameters so that any fit using reasonable values of frequencies is satisfactory. It can be seen that there is good agreement in the falloff without any further adjustment. Slightly better accord could be obtained by introducing $\beta_c < 1$, which shifts the RRKM curve to higher pressures. Figure 14 illustrates typical performance that can be expected from RRKM theory.

Troe has parameterized falloff curves to express the broadening that is observed at intermediate pressures in comparison with Lindemann-Hinshelwood predictions. This has led to a computationally efficient and reasonably accurate general procedure for calculating falloff curves. A summary of this approach has been given (Gardiner and Troe, 1984).

An alternative approach to the computation of k_E for use in situations where there is no maximum of the electronic potential, and hence there is an ambiguity in the location of the transition state, has been put forward (Quack and Troe, 1974; 1975). Termed the statistical adiabatic channel model, it takes angular momentum conservation explicitly into account as certain vibrational degrees of freedom become transformed into product overall rotations, by correlating states of reactants with states of products during passage through the transition region. The microscopic reaction coordinate is thus an adiabatic channel. A potential is constructed for each channel, involving an empirical looseness parameter. The calculations are somewhat more involved than RRKM, and agreement between calculation and experiment is good.

5.3 Chemical Activation and Photoactivation

The central assumptions of RRKM theory are that there is free-energy exchange between active degrees of freedom, that activation and deactivation occur via strong collisions, and that the activated molecules have random lifetimes and decay exponentially. These assumptions can be tested by employing the experimental techniques of chemical activation and photoactivation.

Chemical activation employs bimolecular association reactions to produce an adduct that contains enough internal energy to undergo a subsequent unimolecular reaction:

$$A+B\rightleftarrows AB^*\rightarrow \text{products},$$

$$AB^*+M\rightarrow AB+M.$$

The reverse of the association reaction may always occur, since the AB^* adduct is formed with at least enough energy to redissociate. If the reaction forming products has a lower critical energy than the reverse association reaction, then this path can dominate. Furthermore, AB^* is said to be chemically activated, since its energy distribution function is not characteristic of a thermally activated AB undergoing reaction to products. It is a distribution at the temperature of the association reaction, which may be considerably lower than the temperature needed to drive AB to products thermally, and which has been shifted to higher energy by an amount equal to the exothermicity of the AB^*-forming reaction. Figure 15 illustrates the energetics by means of a reaction-path diagram. The mean thermal energy of the adduct will be $\approx skT$, which may in some cases be only a few percent of the exothermicity of the addition reaction. Thus chemical activation may be used to probe the energy dependence of the unimolecular reaction. The second step in the above mechanism is the collisional deactivation of AB^*. In chemical activation reactions, the relative yield of the stabilized AB and the unimolecular reaction product is pressure dependent.

Photoactivation is analogous to chemical activation. Here, photon absorption produces an excited electronic state of the reactant, which must be selected to be a species that

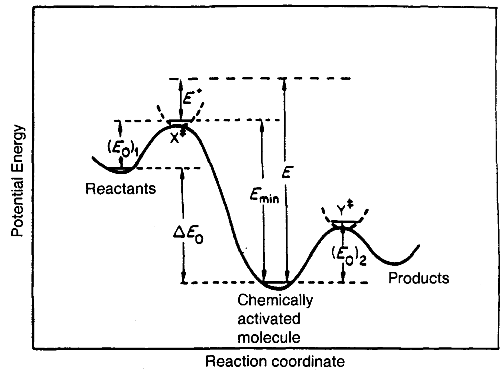

FIG. 15. Reaction-path profile for chemical activation by an exothermic reaction: $X^‡$ represents the transition state for the activating reaction and $Y^‡$ the transition state of the reaction to form products. [Reproduced from Robinson and Holbrook (1972), p. 269, with permission of the copyright holder.]

undergoes rapid intramolecular relaxation to high vibrational levels of the ground state. Photoactivation possesses the advantage that the energy of the photoactivated reactant can be varied by selecting the excitation wavelengths.

One of the first experimental tests of RRKM theory employed the addition of H and D atoms to 1-butene and *cis-* and *trans-*2-butene, which produces, in each case, a chemically activated *sec*-butyl radical capable of dissociating to form propylene and a methyl radical. The thermochemistry differs in each of the six reactions, and by varying temperature the range of energies was increased. Measurement of the unimolecular rate coefficients for these nearly monoenergetically excited radicals showed that they increased with increasing energy, as predicted by the model, and also were in very good quantitative agreement with values calculated by RRKM theory (Rabinovitch *et al.*, 1963).

The random lifetime assumption can be tested by determining the unimolecular rate constant in a chemical activation reaction over a wide range of pressure. If the decay of AB^* is exponential, the unimolecular rate constant will be independent of pressure, but if the decay is nonexponential, it will not be. In the H+2-butene reaction system, pressurized with H_2, the rate constant was independent of pressure from 55 to 203 atm, or up to a collision frequency of 5×10^{12} s^{-1} (Placzek *et al.*, 1966; Oref *et al.*, 1971). It can be concluded that energy randomization in *sec*-butyl radicals must occur on a time scale shorter than 2×10^{-13} s. This result is typical of polyatomic molecules, since a number of experiments have now been done, nearly all of which show that energy randomization occurs within the period of a few to a few tens of molecular vibrations. Organometallics having a relatively heavy metal atom bonded to lighter organic groups may be an exception (Sakai *et al.*, 1990). In these molecules the low-frequency vibrations associated with metal-ligand bonds may not couple strongly enough with portions of the molecule having higher-frequency vibrations to provide pathways for rapid intramolecular energy transfer.

Nonrandom decomposition can be diagnosed in experiments where isotopically labeled chemically activated intermediates are formed. For example, in the reaction

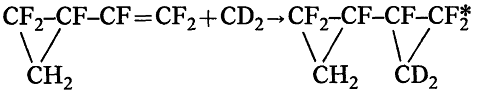

$$\underset{CH_2}{\underset{\diagdown\;\diagup}{CF_2-CF}}-CF=CF_2+CD_2\rightarrow\underset{CH_2}{\underset{\diagdown\;\diagup}{CF_2-CF}}-\underset{CD_2}{\underset{\diagdown\;\diagup}{CF-CF_2^*}}$$

the energy released upon addition of CD_2 to the double bond is initially localized there. Later, it may be distributed throughout the molecule. The reaction is unimolecular elimination of either CD_2 or CH_2, and the deuterium labeling permits one to distinguish which end of the molecule has reacted. Equal elimination of CH_2 and CD_2 denotes randomization, whereas any preferential elimination of CD_2 is evidence of nonrandom decomposition. In this case, 3.5% nonrandom decomposition was found, and the first-order rate constant for intramolecular energy relaxation was determined to be 1.1×10^{12} s (Rynbrandt and Rabinovitch, 1971). In other similar tests of randomization, small amounts of nonrandom reaction, amounting to a few percent, have been found.

Chemical and photochemical activation have been extensively employed for investigations of intermolecular energy transfer. As noted above, the ratio of reaction products to stabilized adducts (D/S) is pressure dependent. The strong-collision mechanism predicts that a plot of D/S vs the reciprocal of the collision frequency will be linear, whereas for weak collisions such a plot will be approximately linear for small values of D/S, but will be noticeably concave upward for $D/S > 10$ to

20. Figure 16 shows experimental data for the decomposition of methylcyclobutane (MCB), chemically activated by insertion of CH_2 into a C–H bond of cyclobutane, to propylene (P) and ethylene (McCluskey and Carr, 1976). Propylene is a decomposition product and methycyclobutane is the stabilized adduct so that $D/S = P/MCB$. It is clear that in order to diagnose weak collisional behavior, data must be taken at higher D/S. The line in the figure is a theoretical calculation, using RRKM theory in a steady-state master-equation formulation. Master-equation calculations are commonplace in the analysis of chemical and photochemical activation data. A collisional transition-probability model is required for these calculations. Simple analytical models are used since there is as yet almost no direct experimental or theoretical information on transition probabilities. The calculation presented in Fig. 16 employed equally spaced energy levels, called a stepladder model, with the spacing used as a fitting parameter. The average amount of energy removed per collision is 30 kJ/mol at an average MCB* energy of 483 kJ/mol, which is about 230 kJ/mol above the critical energy for the decomposition.

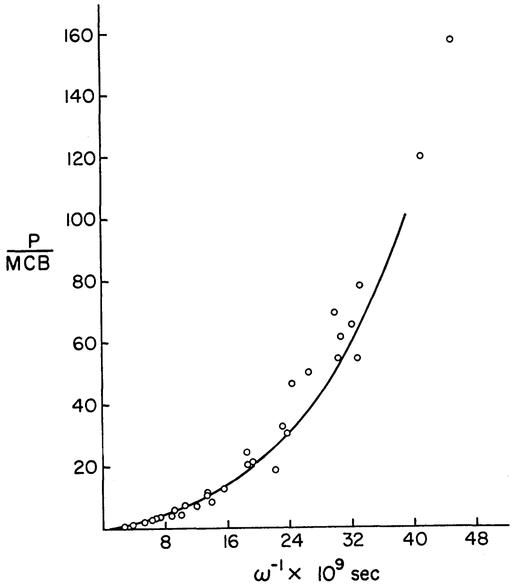

FIG. 16. The dependence of the ratio of decomposition P to collisional stabilization MCB on collision frequency ω for chemically activated methylcyclobutane MCB*. Reprinted with permission from McCluskey and Carr (1976), p. 1395. Copyright 1976, American Chemical Society.

Recent developments in unimolecular reactions center on direct detection of highly vibrationally excited molecules, primarily by application of laser spectroscopies (Crim, 1984; Reisler and Wittig, 1986). Time-resolved determination of unimolecular decay rates of molecules excited to very narrow energy distributions above the critical energy have become possible. These "almost state-resolved" experiments in dense vibrational manifolds are permitting unimolecular reactions to be examined in unprecedented detail, with direct determination of k_E and of reaction-product quantum states. Direct studies of collisional energy transfer have been pioneered by Troe and co-workers (Quack and Troe, 1981; Hippler, 1985) and by Barker (Barker, 1984).

6. ASSOCIATION REACTIONS

Association reactions are a class of reactions initiated by a bimolecular step, but which cannot be described by passage over a single potential barrier, as is done in the usual TST or collision-theory treatment of bimolecular reactions. They include atomic and free-radical combination reactions, as well as addition of atoms and free radicals to molecules. The association of atoms, and the reverse dissociation of diatomics, occurs without passage over a barrier, as the potential of the separating atoms goes asymptotically to the dissociation energy as shown in Fig. 4. The same situation holds for association of free radicals, and in these two cases a problem arises with respect to the placement of the dividing surface if TST is to be applied. In general, a chemically activated intermediate is formed in association reactions, although whether it is observed depends upon its lifetime and the total pressure. It can be seen that an association reaction is not a single elementary step, but a sequence that cannot be separated, since the subsequent steps all follow from the initial bimolecular event. The steps of an association reaction should therefore be treated as a whole. Pilling (1989) has recently reviewed association reactions of atoms and radicals.

6.1 Termolecular Combination of Atoms

The formation of diatomic molecules by combination of atoms is kinetically third order, and is commonly referred to as termolecular, although the reaction actually occurs by two bimolecular steps that must take place within about a picosecond of one another. This can be seen by the following, called the energy-transfer mechanism:

$$A+A \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} A_2^*,$$

$$A_2^*+M \xrightarrow{k_2} A_2+M.$$

A steady-state kinetic treatment of A_2^* leads to the rate expression

$$\frac{d[A_2]}{dt}=\frac{k_1k_2[A]^2[M]}{k_{-1}+k_2[M]}. \qquad (60)$$

The rate coefficient k_{-1} can be associated with the vibration frequency of the diatomic, $\approx 10^{13}\,\mathrm{s}^{-1}$, since there are no other vibrational degrees of freedom to share the energy released on formation of the bond, and redissociation will occur within a single vibrational period. Also, $k_2[M]$ may be associated with the collision frequency, assuming strong collisions. The value of $k_2[M]$ at 1 atm is about $10^{10}\,\mathrm{s}^{-1}$. Thus at virtually all pressures of interest $k_{-1} \gg k_2[M]$, and the rate expression becomes

$$\frac{d[A_2]}{dt}=\frac{k_1k_2}{k_{-1}}[A]^2[M], \qquad (61)$$

from which it is seen that the combination is kinetically third order. Pressures of the order of 1000 atm would be required to observe deviation from third-order kinetics. If collision theory is used to estimate k_1 and k_2, then $k_{\text{third}}=k_1k_2/k_{-1}$ is approximately 10^{-32} cm^6 $\mathrm{molecule}^{-2}$ s^{-1}.

Atoms may also combine by a mechanism, called the bound-complex mechanism, in which an atom forms a van der Waals cluster with M, followed by reaction of AM with another atom:

$$A+M \underset{k_4}{\overset{k_3}{\rightleftarrows}} AM,$$

$$AM+A \xrightarrow{k_5} A_2+M.$$

The rate of formation of diatomic molecules is given by

$$d[A_2]/dt=k_5[AM][A]. \qquad (62)$$

Since the atoms are normally present in a large excess of M species, the AM complexes can be considered to be in chemical equilibrium,

$$K_e=[AM]/[A][M], \qquad (63)$$

and the rate is third order,

$$d[A_2]/dt=k_5K_e[A]^2[M]. \qquad (64)$$

In this case, $k_{\text{third}}=k_5K_e$ can be evaluated by using equilibrium statistical mechanics for K_e, and computing k_5 by the collision theory. For weakly bound AM complexes, k_{third} is approximately 10^{-32} cm^3 $\mathrm{molecule}^{-1}$ s^{-1}, but for more strongly bound complexes, k_{third} will be larger. Data for first-row atoms with atomic and diatomic M species are normally found to have rate coefficients within a factor of 2 of the above estimates. Collected data for the combination of iodine atoms illustrates the much wider range of third-order rate-constant values possible for highly polarizable species (Johnston, 1966).

Measurements of the temperature dependence of k_{third} show that values decrease with increasing temperature. This is a common occurrence for association reactions with no potential barrier. The negative temperature coefficient can be rationalized for the bound-complex mechanism in terms of the temperature dependence of K_e, but is difficult to explain by the energy-transfer mechanism. Arrhenius plots of k_{third} have positive slopes, which according to the Arrhenius equation yield a negative value of E_a. Thus, termolecular atom combinations are frequently said to have negative activation energies, even though it is difficult to attach any physical significance to such a concept.

6.2 Combination of Polyatomic Radicals

The combination of radicals can be described by the mechanism

$$R\cdot+R\cdot \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} R_2^*,$$

$$R_2^*+M \xrightarrow{k_2} R_2+M,$$

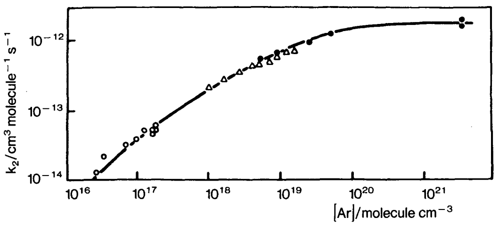

FIG. 17. The dependence of the second-order rate coefficient for the combination reaction $CH_3+O_2+Ar \rightarrow CH_3O_2+Ar$ on concentration of Ar. $k=k_2k_1[Ar]/(k_{-1}+k_2[Ar])$, where the quantities on the right-hand side are defined in Eq. (65). [Reprinted from Pilling (1989), p. 272. Copyright © 1989, John Wiley & Sons, Inc., with permission of John Wiley & Sons.]

which is formally identical to the energy-transfer mechanism of atom combination, with the rate of formation of R_2 given by

$$\frac{d[R_2]}{dt}=\frac{k_2k_1[R\cdot]^2[M]}{k_{-1}+k_2[M]}. \qquad (65)$$

In this case, however, $R_2^{\ddagger}$ is a polyatomic species, and the energy released by formation of the chemical bond between the two $R\cdot$ species can be distributed among its internal degrees of freedom, giving it a significantly longer lifetime than $A_2^{\ddagger}$. For this reason, $k_2[M]$ can no longer be neglected with respect to k_{-1}, and the pressure dependence of $d[R_2]/dt$ may be observable at conventional reaction pressures. It is correct to describe $R_2^{\ddagger}$ as a chemically activated intermediate, and to treat k_{-1} as a unimolecular rate constant. The relative magnitude of $k_2[M]$ and k_{-1} depends upon the pressure and on the lifetime of $R_2^{\ddagger}$, which is a function of the number of oscillators, the vibrational frequency pattern, and the excitation energy, or equivalently, the temperature. Figure 17 shows the pressure dependence of the combination of methyl radicals and O_2, from which it is seen that the pressure-dependent region is between a few Torr and 1 atm. As the combining species become larger (an increase in the number of oscillators) the falloff shifts to lower pressures.

A distinguishing feature of association reactions is that they proceed through the formation of an intermediate complex. This is due to the electronic potential, which has a deep well, in contrast with reactions occurring on surfaces such as depicted in Figs. 7 and 8, where the maximum corresponds to an unstable transition state with a lifetime typically less than a picosecond. Intermediate complexes larger than diatoms can have considerably longer lifetimes than this. The surfaces on which intermediate complexes are formed may have a single entrance and exit channel, such as in Fig. 4, which is correct for diatomic molecules. A curve similar to Fig. 4 may also represent the reaction path for polyatomics for which there is only one energetically accessible reaction. A reaction path profile for a two-channel reaction is sketched in Fig. 18. Here, the possibility of an entrance-channel barrier, with a transition state $X_l^{\ddagger}$, is represented by the solid line, and no barrier by the horizontal dashed line. In the forward direction, the entrance barrier is higher than the exit barrier, but the other case of a lower entrance barrier can be obtained simply by considering C and D to be the reactants. In this case we have the typical chemical activation situation of Fig. 15. The mechanism is given by

$$A+B \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} Y^* \underset{k_{-2}}{\overset{k_2}{\rightleftarrows}} C+D,$$

$$Y^*+M \overset{k_3}{\rightarrow} Y+M.$$

The overall rate coefficient is defined by $k_{\mathrm{exp}}=[A]^{-1}[B]^{-1}d[A]/dt$. If the exit barrier is high enough that k_{-2} is very small, a steady-state treatment of Y^* leads to

$$k_{\mathrm{expt}}=\frac{k_1(k_2+k_3[M])}{k_{-1}+k_2+k_3[M]}, \qquad (66)$$

which predicts that k_{expt} is pressure dependent. Several cases can be distinguished depending upon the relative magnitudes of the

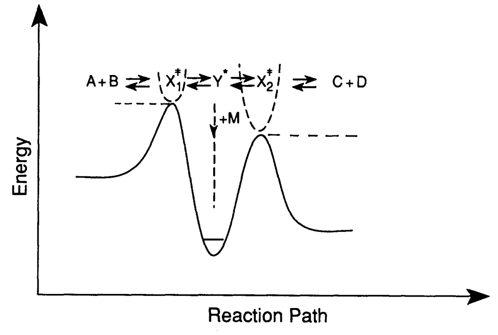

FIG. 18. Reaction-path profile for a two-channel association reaction.

rate coefficients of the elementary processes: For example if $k_3[M] \gg k_2, k_{-1}$, then $k_{\mathrm{expt}} = k_1k_2/(k_{-1}+k_2)$, independent of pressure, and further if $k_{-1} \gg k_2$, $k_{\mathrm{expt}} = k_1k_2/k_{-1}$, or if $k_{-1} \ll k_2$, $k_{\mathrm{expt}} = k_1$. Other cases can be worked out starting with $k_3[M] \ll k_2$. Values of k_{expt} can be estimated from theory by applying TST to k_1, and RRKM theory to k_{-1} and k_2.

It is worth emphasizing that although the reaction of A with B involves a bimolecular step (initially), there is only one case where application of TST alone leads to a correct interpretation of k_{expt}. Reactions that appear elementary should always be closely examined to make sure that they do not proceed by complex formation. The reaction $OH+CO = CO_2+H$, which is important in combustion, provides a case in point. It was originally thought to be an elementary atom-transfer reaction, although the temperature dependence from low- and high-temperature experiments was not in agreement. It is now known that Arrhenius plots for this reaction are strongly curved, being nearly independent of temperature at 300 K and lower temperatures, suggesting virtually no barrier, but increasing strongly with increasing temperature above 1000 K. This behavior cannot be explained if $OH+CO$ occurs as a direct reaction which can be modeled by TST. It is readily and quantitatively reproduced by postulating that the reaction involves the formation of an intermediate HOCO species. There is independent evidence from low-temperature spectroscopy in matrices for the existence of stable HOCO.

7. MOLECULAR REACTION DYNAMICS

Molecular reaction dynamics is the branch of kinetics dealing with elementary processes from the standpoint of the motions, or dynamics, of atomic and molecular motions during chemical reaction. These are most usefully investigated by examining quantum states. It would be of interest to know how microscopic rate coefficients depend upon the velocities and internal states of the reactants, and what the product velocities and internal state distributions are. A complete state-to-state description would provide an enormously detailed picture of a reaction. The accumulated results from this type of investigation would lead to deep understanding of the factors that control chemical reactivity, and to detailed tests of rate theories. Levine and Bernstein have recently published an excellent text on molecular reaction dynamics (Levine and Bernstein, 1987).

In practice, state-to-state kinetics usually falls somewhat short of these lofty aspirations. The amount of experimental effort that would be required for a complete state-to-state description of a single reaction in which the reactants are in thermal equilibrium with the surroundings would be formidable, with a number of experimental techniques most likely having to be brought to bear. Furthermore, it may not be possible to achieve complete initial-state selection in many experiments, and where possible, final-state detection sensitivity may require compromises. Most molecular-dynamics experiments that have been reported have required some concessions along these lines. Equations (17) give the relationship between microscopic rate coefficients and rate coefficients of elementary reactions. Although, in principle, the latter could be calculated from these equations, it is impractical to do so. In spite of these shortcomings, many very interesting experiments have been done, most of them with only partial state resolution. The resultant body of data has been instrumental in advancing our understanding of how chemical reactions occur.

The trajectory calculation approach discussed in Sec. 6 is capable of providing the detailed picture sought, but it must be remembered that the accuracy of these calculations depends critically upon the potential-energy surface. Since quantum-mechanical calculations of surfaces are always approximate, considerable effort must be put into evaluating the goodness of a trajectory calculation. Computational methodology has advanced to the point where good confidence can be placed in a number of surfaces.

7.1 Molecular Beams

Bimolecular reactions can be studied in crossed molecular beams. Two narrow beams of atoms or molecules are caused to intersect in a vacuum chamber. The beams are now most usually produced by supersonic nozzle expansions, although the earlier technique of effusive beams may also be used. In the individual beams, collisions are virtually absent, and the atoms or molecules move in essentially one dimension, given by the direction of the beam. However, in the volume defined by the intersection, collisions between the species of each beam occur. The forces that govern the collisional interaction cause the atoms and molecules to be scattered. A detector that can be moved about the scattering volume detects the scattered species. The angular dependence of the scattering provides information from which differential cross sections can be obtained. The detector most commonly used is the quadrupole mass filter, which provides species identification and can distinguish between reactive and unreactive scattering. Crossed–molecular-beam investigations can provide information about many aspects of bimolecular collision dynamics. The velocity dependence of the cross section, the effect of internal energy states of the reactants on the course of the reaction, and the dependence of product velocities, internal energy states, and angular momentum on initial states can all be investigated.

Molecular-beam investigations have revealed that for certain reactions, the reactive cross sections can be quite large, of the order of 100 Å^2. In these cases, there is a high probability of reaction at large impact parameters due to the dominance of long-range attractive forces, and the scattering in the center-of-mass coordinate system is predominantly in the forward direction. This behavior is observed in the reactions of alkali-metal atoms, M, with halogens, X_2, where a halogen atom is transferred to the metal,

$$M+X_2 \rightarrow MX+X;$$

this is termed a stripping reaction, since it appears that the metal strips an atom from X_2 as it passes by. In other reactions, having smaller cross sections in the vicinity of 10 Å^2, the products recoil backward in the center-of-mass system. These are said to occur by a rebound mechanism. The observed dynamics in these cases is a result of the low reaction probability for all but a range of small impact parameters. Reactive collisions are thus close interactions in which the products rebound from the repulsive wall of the potential.

In the above types of reactions, the anisotropic angular distribution of the products due to the bimolecular collision dynamics can be observed because the lifetime of the transition state is very short, of the order of 10^{-12}–10^{-13} s or less. If the lifetime is longer than the characteristic time for rotation, the angular distribution becomes one of forward-backward symmetry. Longer lifetimes are usually associated with bound collision complexes. Thus, molecular beams can be employed to study the unimolecular decay of complexes formed in bimolecular collisions. Quantities such as the relative probabilities of decay into different products, angular distributions of the products, the internal and translational energies of the products, and the dependence of these on the reactant energies can all be investigated. A very detailed picture of a unimolecular reaction may be obtained, permitting rigorous tests of unimolecular rate theories.

Orientation effects have been studied by producing beams of oriented molecules in a combination of inhomogeneous and homogeneous electric fields. Molecules with permanent dipole moments can be aligned parallel to the homogeneous field **E**, after being rotationally state selected in the inhomogeneous field. The direction of orientation can be reversed by reversing **E**, thus presenting one end or the other of the oriented molecule to the other beam. In the reaction of CH_3I with Rb atoms to form RbI and CH_3 radicals, interaction of Rb with the "I-atom end" of the molecule results in a high reaction probability, whereas interaction with the "CH_3-group end" results in no reaction when the atom is within a conical region determined by the C–H bond angle of the methyl group.

The effect of internal energy of the reactants can be studied in molecular-beam experiments by exciting one of the reactants. This was first demonstrated by Brooks and coworkers, who studied the reaction of K atoms with HCl (Odidorne *et al.*, 1971),

$$K+HCl=KCl+H,$$

in which the HCl was excited to $v=1$ by employing a pulsed HCl laser. It was found that the reactive cross section was larger by more than two orders of magnitude when the HCl was in $v=1$, compared with $v=0$. Furthermore, it was found that translational energy was less effective than vibrational energy in facilitating this reaction. Subsequent research has shown other cases where vibrational energy is more effective than translational energy in promoting reaction, and others where translational energy is the dominant factor, vibrational energy having little influence. These findings have been rationalized in terms of the position of the barrier maximum on the potential-energy surface.

Laser spectroscopies are now being used for quantum-state-specific detection of reaction products in crossed–molecular-beam experiments. In the reaction

$$D+H_2 \rightarrow HD+H,$$

measurements of differential cross sections for the HD product in $J=1$, 2, and 3 of $v=1$ have recently been made. These are extremely sensitive to subtle details of the potential-energy surface, and are expected to lead to refinement of the potential for this reaction. The prospects are that the interplay between theory and experiment will be exploited further in the future, leading to improvements in the calculation of accurate potentials.

7.2 Energy Disposal in Chemical Reactions

Energy disposal addresses how the energy brought to the reaction by the reactants, and in exothermic reactions by thermochemistry, is distributed among the products. There are several experimental methods for probing energy distributions.

One of the earliest and most successful methods is infrared chemiluminescence (IRC) (see CHEMILUMINESCENCE). IRC is the spontaneously emitted IR radiation from reaction products produced in vibrationally excited states. In small enough molecules that the emission, dispersed by a monochromator, can be resolved into individual vibration-rotation bands, the internal energy distribution of the products can be determined. The relatively long radiative lifetimes in the IR require low background pressures and large reaction chambers in order to avoid relaxation of the distribution by gas-phase and wall collisions, and to ensure that the nascent distribution, which contains the information on dynamics, is observed. This technique was pioneered by Polanyi and co-workers in studies of atoms reacting with diatomic molecules (Polanyi and Schreiber, 1974),

$$A+BC \rightarrow AB(v,J)+C.$$

Comparison of the vibrational and rotational energy of AB with the total internal energy change accompanying the reaction permits the relative translational energy of AB and C to be determined by difference for each v,J state formed, and a map of total-energy disposal is obtained. In the reaction of F atoms with H_2, which is exoergic by 130 kJ/mol, the fraction of total reaction energy that goes into vibration, rotation, and translation is 0.66, 0.08, and 0.26, respectively. Furthermore, there is a population inversion of the vibrational states, which is the basis of the HF chemical laser.

A drawback of IRC is that the population in $v=0$ is not observable. This may be circumvented by probing the product energy distribution with laser-induced fluorescence (LIF). In LIF, a pulsed, tunable, narrow-band laser is used to excite a molecular species into a fluorescent electronic state. The fluorescence intensity, which is related to the population of the state that is excited, is measured. In cases where the molecule of interest does not fluoresce, the multiphoton ionization technique can be used to good advantage.

Laser spectroscopies have been extensively employed to probe the population of product states in chemical reactions. Levine and Bernstein (1987) cover many examples of this type of experiment in bimolecular reactions, while Crim (1984) and Reisler and Wittig (1986) review recent developments in state-resolved experimentation in unimolecular reactions. Recent texts on gas-phase kinetics also give some coverage of the many experimental approaches to product distributions, as well as some important results.

GLOSSARY

Activation Energy: An energy defined by the equation $E_a \equiv -Rd\ln k(T)/d(1/T)$, where R is the gas constant, $k(T)$ is the rate coefficient, and T is the absolute temperature. It is also given by the difference between the average energy of those reactant molecules that have enough energy to react, and the average energy of all reactant species.

Arrhenius Equation: An empirical equation giving the temperature dependence of the rate coefficient $k(T)$, $k(T)=A\exp(-E_a/RT)$. A is the Arrhenius preexponential factor.

Association Reaction: Association of two molecules to form an energized intermediate complex which may be collisionally stabilized, redissociate to the reactants, or undergo unimolecular reaction(s) to form new product(s).

Chemical Activation: The formation of a highly internally excited molecule by either an exothermic addition reaction, or an endothermic addition reaction proceeding over a potential barrier.

Chemical Dynamics: The study of chemical reaction kinetics at the quantum-state–resolved (or partially quantum-state–resolved) level.

Collision Theory: Any theory that treats reactions as occurring via the dynamics of collisions, as opposed to statistical considerations.

Cross Section: The effective target area presented by two colliding species in either elastic scattering, inelastic scattering, or reaction.

Elementary Reaction: A chemical reaction viewed as it occurs as an isolated molecular event. Spontaneous reaction of a single molecule is a unimolecular reaction, a reaction requiring interaction of two molecules is bimolecular, and one involving three molecules is termolecular.

Empirical Kinetics: The determination of rate expressions for complex overall reactions by laboratory measurements of rates of change of reactants and products. No information on short-lived, reactive intermediates is obtained.

Mechanism: The list of elementary reactions that make a non-negligible contribution to the rate of an overall reaction.

Microcanonical Rate Coefficient: A rate coefficient for an elementary reaction at a specified total energy. Thermal rate coefficients $k(T)$ are thermally averaged microcanonical rate coefficients.

Microscopic Kinetics: The kinetics of quantum state-to-state processes.

Microscopic Reversibility: The principle that all reactions are reversible, and that therefore cyclic equilibria are not permitted.

Molecularity: The number of molecules involved in a specified elementary reaction.

Order: For reactions that have a rate expression of the form $r = k(T)\Pi[A_i]^{\alpha_i}$, where $[A_i]$ is the concentration of the ith chemical species, the order is defined as $\Sigma\alpha_i$.

Overall Reactions: A complex reaction written to show the stoichiometric relationship between reactants and products, but containing no information on the elementary reactions involved, or on any reactive intermediates.

Photochemical Activation: The formation of highly internally excited ground electronic-state molecules by photoexcitation to an excited electronic state, followed by a radiationless transition to the ground state.

Potential-Energy Surface: The electronic potential of a reacting system as a function of the relative positions of the atoms. For the three-atom system, the potential can be represented in three dimensions, but for any larger number of atoms, a potential-energy hypersurface in a higher-dimensional space is required.

Rate: The time derivative of a chemical species concentration in a reacting system.

Rate Coefficient: Proportionality coefficient appearing in rate expressions. Properly determined rate coefficients $k(T)$ are functions of temperature only.

Rate Expression: An equation expressing the dependence of the rate upon reactant (and product) concentrations, as well as any other parameters that influence the rate.

Steady-State Approximation: An approximation made in the kinetic analysis of a mechanism, whereby the net rate of formation of a highly reactive intermediate is set equal to zero, thus converting a differential equation into an algebraic equation.

Stoichiometric Reaction: An overall reaction.

Transition State: The assembly of atoms that exists at the dividing surface between reactants and products in a chemical reaction. This concept is used in transition-state theory.

Transition-State Theory: A statistical theory of reaction rates in which the rate of passage over a potential barrier is calculated by assuming dynamic equilibrium between the transition state and the reactants.

Variational Transition-State Theory: A version of transition-state theory in which the dividing surface is placed either to minimize the Gibbs free energy of the transition state (canonical variational transition-state theory) or to minimize the microcanonical rate coefficient (microcanonical variational transition-state theory).

List of Works Cited

Bamford, C. H., Tipper, C. F. H. (Eds.) (1969), *Comprehensive Chemical Kinetics*, Vol. 1. Amsterdam: Elsevier.

Barker, J. R. (1984), *J. Phys. Chem.* **88**, 11–18.

Bernstein, R. B., Zewail, A. H. (1990), *Chem. Phys. Lett.* **170**, 321–328.

Beyer, T., Swinehart, E. F. (1973), *Commun. ACM* **16**, 379.

Crim, F. (1984), *Annu. Rev. Phys. Chem.* **35**, 657–691.

Denbigh, K. (1964), *The Principles of Chemical Equilibrium*, Cambridge, England: Cambridge University Press, pp. 442–447.

Espenson, J. H. (1981), *Chemical Kinetics and Reaction Mechanisms*, New York: McGraw-Hill.

Eyring, H., Walter, J., Kimball, G. E. (1944), *Quantum Chemistry*, New York: Wiley, p. 326.

Forst, W. (1973), *Theory of Unimolecular Reactions*, New York: Academic.

Fowler, R. H., Guggenheim, E. A. (1939), *Statistical Thermodynamics*, Cambridge, England: Cambridge University Press.

Freiss, S. L., Lewis, E. S., Weissberger, A. (Eds.) (1961,1963), *Investigation of Rates and Mechansims of Reactions*, Vol. VIII, Parts I and II, *Technique of Organic Chemistry*, New York: Interscience.

Gardiner, Jr., W. C., Troe, J. (1984), in W. C. Gardiner, Jr. (Ed.), *Combustion Chemistry*, New York: Springer-Verlag.

Glasstone, S., Laidler, K. J., Eyring, H. (1941), *Theory of Rate Processes*, New York: McGraw-Hill.

Hippler, H. (1985), *Ber. Bunsenges. Phys. Chem* **89**, 303–309.

Hirst, D. M. (1985), *Potential Energy Surfaces*, London: Taylor and Francis.

Jenson, V. G., Heffreys, G. V. (1963), *Mathematical Methods in Chemical Engineering*, London: Academic.

Johnston, H. S. (1966), *Gas Phase Reaction Rate Theory*, New York: Ronald.

Karplus, M. (1990), *J. Phys. Chem.* **94**, 5435–5436.

Karplus, M., Porter, R. N., and Sharma, R. D. (1965), *J. Chem. Phys.* **43**, 3259–3287.

Kassel, L. S. (1932), *Kinetics of Homogeneous Gas Reactions*, New York: Van Nostrand Reinhold.

Laidler, K. J. (1987), *Chemical Kinetics*, 3rd ed., New York: Harper and Row, pp. 39–48.

Levine, R. D., Bernstein, R. B. (1987), *Molecular Reaction Dynamics and Chemical Reactivity*, New York: Oxford University Press, pp. 78–91.

Lindemann, F. A. (1992), *Trans. Faraday Soc.* **17**, 598–599.

Liu, B. (1973), *J. Chem. Phys.* **58**, 1925–1937.

Marcus, R. A., Rice, O. K. (1951), *J. Phys. Colloid Chem.* **55**, 894–907.

Marcus, R. A. (1952), *J. Chem. Phys.* **20**, 359–364.

Marcus, R. A., Coltrin, M. E. (1977), *J. Chem. Phys.* **67**, 2609–2613.

McCluskey, R. J., Carr, R. W. (1976), *J. Phys. Chem.* **80**, 1393–1398.

Melville, H. W., Gowenlock, B. G. (1964), *Experimental Methods in Gas Reactions*, London: Macmillan.

Nikitin, E. E. (1974), *Theory of Elementary Atomic and Molecular Processes in Gases*, Oxford: Clarendon.

Odidorne, T. J., Brooks, P. R., Kasper, J. V. (1971), *J. Chem. Phys.* **55**, 1980–1982.

Oref, I., Schuetzle, D., Rabinovitch, B. S. (1971), *J. Chem. Phys.* **54**, 575–578.

Oref, I., Tardy, D. C. (1990), *Chem. Rev.* **90**, 1407–1445.

Perrin, J. (1919), *Ann. Phys. (Paris)* **11**, 1–108.

Pilling, M. J. (1989), *Int. J. Chem. Kinet.* **21**, 267–291.

Placzek, D. W. Rabinovitch, B. S., Dorer, F. H. (1966), *J. Chem. Phys.* **44**, 279–284.

Polanyi, J. C., Schreiber, J. L. (1974), in: H. Eyring, D. Henderson and W. Jost (Eds.), *Physical Chemistry: An Advanced Treatise*, Vol. VIa, New York: Academic, Chap. 6.

Quack, M., Troe, J. (1974), *Ber. Bunsenges. Phys. Chem.* **78**, 240–252.

Quack, M., Troe, J. (1975), *Ber. Bunsenges. Phys. Chem.* **79**, 170–183.

Quack, M., Troe, J. (1981), *Int. Rev. Phys. Chem.* **1**, 97–147.

Rabinovitch, B. S. Kubin, R. F., Harrington, R. E. (1963), *J. Chem. Phys.* **38**, 405–417.

Reisler, H., Wittig, C. (1986), *Annu. Rev. Phys. Chem.* **37**, 307–349.

Robinson, P. J., Holbrook, K. A. (1972), *Unimolecular Reactions*, New York: Wiley-Interscience.

Rynbrandt, J. D., Rabinovitch, B. S. (1971), *J. Phys. Chem.* **75**, 2164–2171.

Sakai, Y., Iyer, R. S., Rowland, F. S. (1990), *J. Phys. Chem.* **94**, 3368–3370.

Sato, S. (1955), *J. Chem. Phys.* **23**, 592–593; **23**, 2465–2466.

Schlag, E. W., Rabinovitch, B. S. (1960), *J. Am. Chem. Soc.* **82**, 5996–6000.

Siegbahn, P. Liu, B. (1978), *J. Chem. Phys.* **68**, 2457–2465.

Slater, N. B. (1959), *Theory of Unimolecular Reactions*, Ithaca, NY: Cornell University.

Smith, I. W. M. (1980), *Kinetics and Dynamics of Elementary Gas Reactions*, London: Butterworths, pp. 81–84, p. 201.

Stein, S. E., Rabinovitch, B. S. (1973), *J. Chem. Phys.* **58**, 2438–2445.

Szabo, G. (1961), in C. H. Bamford and C. F. H.

Tipper (Eds.), *Comprehensive Chemical Kinetics*, Vol. 2. Amsterdam: Elsevier, Chap. 1.

Tardy, D. C., Rabinovitch, B. S. (1977), *Chem. Rev.* **77**, 369–408.

Topor, M. G., Carr, R. W. (1973), *J. Chem. Phys.* **58**, 757–768.

Truhlar, D. G., Horowitz, C. J. (1978), *J. Chem. Phys.* **68**, 2466–2476.

Truhlar, D. G., Garrett, B. G. (1980), *Acc. Chem. Res.* **13**, 440–448.

Waage, E. V., Rabinovitch, B. S. (1970), *Chem. Rev.* **70**, 377–387.

Weston, R. E., Schwarz, H. A. (1972), *Chemical Kinetics*, Englewood Cliffs, NJ: Prentice-Hall.

Whitten, G. Z., Rabinovitch, B. S. (1963), *J. Chem. Phys.* **38**, 2466–2473.

Further Reading

Benson, S. W. (1960), *The Foundations of Chemical Kinetics*, New York: McGraw-Hill.

Bunker, D. L. (1966), *Theory of Elementary Gas Reaction Rates*, New York: Pergamon.

Eyring, H., Lin, S. H., Lin, S. M., *Basic Chemical Kinetics*, New York: Wiley-Interscience.

Gardiner, W. C., Jr. (1969), *Rates and Mechanisms of Chemical Reactions*, New York: Benjamin.

Kondrat'ev, V. N. (1964), *Chemical Kinetics of Gas Reactions*, Reading, MA: Addison-Wesley.

Laidler, K. J. (1987), *Chemical Kinetics*, New York: Harper & Row.

Laidler, K. J. (1969), *Theories of Chemical Reaction Rates*, New York: McGraw-Hill.

Levine, R. D., Bernstein, R. B. (1987), *Molecular Reaction Dynamics and Chemical Reactivity*, London: Oxford University Press.

Mulcahy, M. F. R. (1973), *Gas Kinetics*, New York: Halsted.

Pilling, M. J., Smith, I. W. M. (Eds.) (1987), *Modern Gas Kinetics*, Cambridge, MA: Blackwell.

Pratt, G. L. (1969), *Gas Kinetics*, New York: Wiley.

Smith, I. W. M. (1980), *Kinetics and Dynamics of Elementary Gas Reactions*, London: Butterworths.

Steinfeld, J. I., Francisco, J. S., Hase, W. L. (1989), *Chemical Kinetics and Dynamics*, Englewood Cliffs, NJ: Prentice-Hall.

Weston, R. E., Schwarz, H. A. (1972), *Chemical Kinetics*, Englewood Cliffs, NJ: Prentice-Hall.

CHEMICAL REACTIONS

Ralph E. Weston, *Department of Chemistry, Brookhaven National Laboratory, Upton, New York, U.S.A.*

Introduction 377
1. Gas-Phase Reactions 378
1.1 Bimolecular Reactions 378
1.1.1 Reactions in Molecular Beams 378
1.1.2 Atom and Radical Abstraction Reactions 379
1.1.3 Ion–Molecule Reactions 379
1.1.4 Atom and Radical Recombinations 380
1.2 Unimolecular Reactions 380
1.2.1 Isomerizations 381
1.2.2 Bond Rupture Reactions 381
1.3 Some Practical Applications of Gas-Phase Reactions 381
1.3.1 Combustion Chemistry 381
1.3.2 Air Pollution: Photochemical Smog 383
1.3.3 Homogeneous Catalysis: Destruction of Ozone in the Stratosphere 385
2. Reactions in the Liquid Phase and in Solution 387
2.1 Ionic Reactions 387
2.2 Acids and Bases: Proton Transfer Reactions 388
2.3 Oxidation–Reduction Reactions 389
2.4 Catalysis by Enzymes 392
2.5 Polymerization Reactions 394
3. Reactions Involving the Solid Phase 395
3.1 Solid–Solid Reactions 396
3.2 Gas–Solid Reactions 396
3.2.1 Tarnishing of Metals 396
3.2.2 Chemical Vapor Deposition of Thin Films on Semiconductor Substrates 396
3.2.3 Vapor Etching of Semiconductors 397
3.2.4 Adsorption of Gases on Solids. 398
3.2.5 Heterogeneous Catalysis 401
3.2.5.1 Catalytic Processes Used in the Petroleum and Petrochemical Industry 402
3.2.5.2 Heterogeneous Catalysis in the Heavy Chemical Industry 403
4. Electrochemical Reactions ... 403
4.1 Electrochemical Cells 403
4.1.1 Batteries 404
4.2 Electrolysis 406
4.2.1 Electroplating 406
4.2.2 Electrolytic Production of Hydrogen and Deuterium 407
4.2.3 Electrolysis in Molten Salts ... 407
4.3 Corrosion of Metals 408
Glossary 409
Works Cited 410
Further Reading 410

INTRODUCTION

A chemical reaction can be defined as any process in which the chemical bonds, and hence the valence electrons, of one or more chemical species undergo rearrangement. Thus, this definition includes reactions such as

1. an electron transfer reaction,

$$Fe^{3+} + {}^{*}Fe^{2+} \rightarrow {}^{*}Fe^{3+} + Fe^{2+}$$

(where the asterisk indicates a radioactive label),

2. a geometrical isomerization,

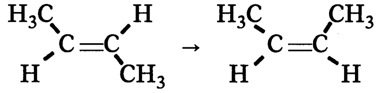

3. or a metathesis,

$$Cl + H_2 \rightarrow HCl + H.$$

A substantial fraction of the entire science of chemistry involves the study, in one way or

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

another, of chemical reaction—chemists seek to determine what chemical species react with each other, the rates at which the reactions take place, the energetics of the process, and, ultimately, the detailed reaction mechanism at the atomic level. The ultimate goal of such studies is to provide a basis from which the analogous properties of other hypothetical reactions can be predicted without the necessity of carrying out the actual measurements of reaction properties. The chemical equation expressing a chemical reaction, i.e., reactants → products, indicates the stoichiometry of the reaction, but may give a very misleading and oversimplified impression of the actual reaction mechanism. Many elementary reactions, involving intermediates that are not among the reactants or products, may be taking place simultaneously. For example, the reaction of hydrogen and bromine to form hydrogen bromide is described by the net reaction

$$H_2+Br_2 \rightarrow 2HBr,$$

but the actual mechanism consists of the elementary reactions

$$Br_2 \rightarrow 2Br,$$

$$Br+H_2 \rightarrow HBr+H,$$

$$H+Br_2 \rightarrow HBr+Br,$$

$$H+HBr \rightarrow H_2+Br,$$

$$Br+Br \rightarrow Br_2.$$

Throughout this article, the emphasis, whenever possible, will be on elementary reactions.

Many of the features of chemical reactions are treated under other headings in this *Encyclopedia*. For example, general characteristics of reaction rates and mechanisms are discussed in CHEMICAL KINETICS. Specific types of reactions are treated in greater detail under headings such as COMBUSTION, POLYMER REACTIONS, PHOTOCHEMISTRY, etc. Therefore, this article can only attempt to catalog major types of chemical reactions, whenever possible emphasizing those related to important technical applications.

1. GAS-PHASE REACTIONS

Gas-phase reactions play a very important role in the understanding of all chemical reactions. They are the simplest, in the sense that the reactants are isolated and are not perturbed by interactions with adjacent solvent or surface molecules. In principle, the simplest possible chemical reaction is that of a single isolated molecule, a unimolecular reaction. In fact, the understanding of unimolecular reactions followed that of bimolecular reactions, as is discussed below. As will be discussed later in this section, many reactions of practical importance take place in the gas phase.

1.1 Bimolecular Reactions

A bimolecular reaction, as its name implies, involves two reactants, one or both of which may be atoms, radicals, ions, or molecules. The reactants may simply combine to form a larger molecular entity, i.e.,

$$A+A \rightarrow A_2,$$

or can undergo a metathesis of the form

$$A+BC \rightarrow AB+C,$$

where A , B, and C may be atoms or larger groups. The rate of this reaction is given by the expression

$$\begin{aligned} -d[A]/dt &= -d[BC]/dt \\ &= d[AB]/dt = d[C]/dt \\ &= k[A][BC]. \end{aligned} \tag{1}$$

In this equation, t is time, the brackets indicate concentrations or pressures, and k is the rate constant or rate coefficient. The basic mechanism for such a reaction is easily understood, involving the approach of the reactive species to some distance at which chemical forces leading to product formation overcome the bonding forces in the original reactant molecules. For the sake of simplicity, such an encounter is called a "reactive collision," although it is somewhat ill-defined. A reaction may be either exothermic (heat evolved) or endothermic (heat absorbed), but in either case there is usually an energy barrier that must be surmounted before reaction takes place (see CHEMICAL KINETICS).

1.1.1 Reactions in Molecular Beams. Atomic beams were produced as early as 1911, and the techniques were extensively applied in atomic physics during the 1920s, earning Otto Stern a Nobel Prize for Physics in 1943. However, crossed-molecular-beam studies of chemical reactions were not carried out until the 1950s (Bull and Moon, 1954; Taylor and Datz, 1955). The experimental difficulties are exemplified by the fact that the density of product species is many orders of magnitude less than that of the reactants.

From the standpoint of understanding the detailed mechanism of a bimolecular reaction, one would like to study the following process:

$$A\ (v_{\mathrm{rel}}) + BC(v, J) \rightarrow AB(v', J') + C(v'_{\mathrm{rel}}). \quad (2)$$

Here, the atomic species A, with collision velocity relative to BC of v_{rel}, reacts with the diatomic molecule BC in internal energy states specified by rotational and vibrational quantum numbers J and v. The products separate with a relative velocity v'_{rel}, and with internal energy states of AB again specified.

Usually, chemical reactions occur under much less well-defined conditions, with a broad thermal distribution of both velocity and internal energy. Molecular beams permit an approach to the ideal reaction written above (Bernstein, 1982). First of all, the reaction takes place under single-collision conditions, ensured by working at very low pressures, and the relative velocity of the reactants can be controlled by the use of mechanical velocity selectors. The usually broad thermal distribution of internal energy states of the molecule BC can be greatly reduced by using a supersonic jet expansion that converts internal energy into translational energy, and effectively lowers the vibrational and rotational temperatures. Excitation of specific v, J levels can be achieved by using appropriate infrared lasers. To probe the internal energy states of the products, sensitive and selective detection methods are required; these include infrared chemiluminescence, laser-induced fluorescence, multiphoton ionization, bolometric detection, surface ionization, mass spectrometric analysis, etc. Another mechanical velocity selector, or a time-of-flight mass spectrometer, is used to determine product velocity over a range of scattering angles. By taking advantage of the principle of energy conservation, precise measurements of the relative kinetic energy of fragments can then be used to determine their internal energy states, if the energy of reaction is known.

The 1986 Nobel Prize for Chemistry was awarded to D. R. Herschbach, Y. T. Lee, and J. C. Polanyi for their contributions to the understanding of molecular reaction dynamics, and both Herschbach and his one-time postdoctoral student Lee used molecular-beam techniques to achieve this goal.

1.1.2 Atom and Radical Abstraction Reactions. A large number of bimolecular gas-phase reactions of the type $A + BC \rightarrow AB + C$ are metathetical reactions in which A and C are atoms or free radicals, and BC and AB are molecules. A free radical is an atom or molecular fragment that has at least one unpaired electron. Since chemical bonds usually are formed by paired electrons, the radical is an inherently reactive species. This type of reaction is also called an "abstraction" reaction, since the A species "abstracts" an atom or molecular fragment from the molecule BC. The energetics of this type of reaction are shown schematically in Fig. 1. It should be noted that even when the reaction is energetically "downhill," there is often an energy barrier to be surmounted for the reaction to occur. This barrier is closely related to the activation energy E_a, which can be derived from the temperature dependence of the rate constant, often expressed in the form first used by Arrhenius in 1889:

$$k(T) = A(T)\exp(-E_a/RT), \quad (3)$$

where T is the absolute temperature and R is the gas constant. The term $A(T)$ is known as the pre-exponential factor. Methods for estimating both $A(T)$ and E_a from molecular and thermodynamic properties have been extensively developed (Benson, 1976).

As the general chemical equation above indicates, the reactant free radical forms a stable molecule, but produces another radical in the process. This can lead to a chain reaction, exemplified by the one historically the first to be studied:

$$Br + H_2 \rightarrow HBr + H, \quad (4)$$

$$H + Br_2 \rightarrow HBr + Br. \quad (5)$$

Such a process can proceed indefinitely, until the chain is broken by the intervention of some other reaction that consumes one of the radical species.

1.1.3 Ion–Molecule Reactions. In reactions involving uncharged species, as discussed above, the forces between the reactants

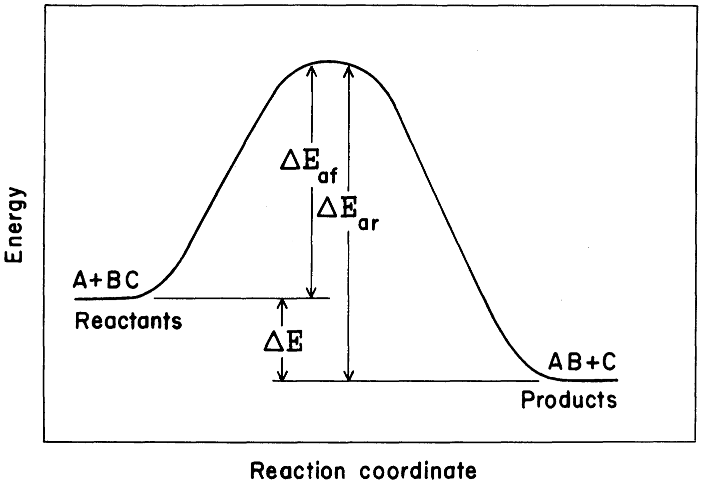

FIG. 1. Energy vs reaction coordinate for the hypothetical reaction $A+BC \rightarrow AB+C$, showing the transition state at the top of the energy barrier. ΔE is the energy of reaction, which in this case is exoergic; E_{af} and E_{ar} are the energy barriers for the forward and reverse reactions, respectively. They are closely related to the activation energies (Sec. 1.1.2).

are very short range (0.1–1.0 nm) chemical forces. Another class of reactions important in radiation chemistry, mass spectrometry, electric discharges, and the upper atmosphere involves ions. The interaction of a charged molecule or atom with an uncharged species involves much longer range electrostatic forces than does the interaction of two uncharged species. If the uncharged reactant is a dipolar molecule, such as HCl or NH_3, there is an attractive potential with an inverse dependence on the square of the distance between the two reactants. Even in the case of a nonpolar molecule, the dipole induced by the approaching ion will lead to a potential with a $1/r^4$ dependence. As a result of these attractive forces, the rate constants for ion–molecule reactions of the abstraction type are much larger than those for neutral species.

1.1.4 Atom and Radical Recombinations. In the discussion of chain reactions above, the termination of the chain was briefly mentioned. One obvious termination step involves the bimolecular reaction in which two atoms or radicals recombine. These reactions are exothermic by an amount corresponding to the bond dissociation energy, and typically have very small or zero activation energies. The only barrier to reaction is that imposed by the conservation of angular momentum, and the rate constants can be estimated reasonably well using a hard-sphere collision model. The exothermicity of the reaction results in a highly energetic product, which, unless stabilized by a simultaneous or subsequent collision with another species, has a finite probability of dissociation. In fact, the recombination process is just the microscopic reverse of a unimolecular dissociation discussed in Sec. 1.2.

1.2 Unimolecular Reactions

Although in principle a unimolecular reaction is the simplest possible chemical reaction (as it involves a single molecule), the understanding of such processes actually came later than the understanding of bimolecular reactions. The puzzle is: If an isolated molecule can react, how can it exist as a stable species? The clue to the answer is found in the observation that unimolecular reactions proceed more rapidly as the temperature is increased, indicating that some sort of excitation or activation of the reactants is required. After experiments showed that this excitation could not result from the absorption of radiation, it was recognized that collisions with other molecules were the source of the required energy. This collision mechanism for activation was proposed by Lindemann in 1922, and a simplified version is as follows:

$$A+M \rightarrow A^*+M \quad \text{(activation by collision),} \tag{6}$$

$$A^*+M \rightarrow A+M \quad \text{(deactivation by collision),} \tag{7}$$

$$A^* \rightarrow \text{products} \quad \text{(spontaneous reaction of activated molecule).} \tag{8}$$

In this mechanism, A is the reactant molecule, A^* is an activated reactant molecule, and M may be either another molecule of reactant or of any other species present in the reaction mixture. Although the first two steps are actually bimolecular reactions, at sufficiently high pressures a quasiequilibrium is reached so that the concentration of activated molecules A^* is maintained and the rate-determining step is its spontaneous reaction. At lower pressures, the activation step becomes the rate-limiting process. Thus, the apparent rate constant is a function of total pressure. This pressure dependence can be used to determine the relative energy transfer efficiencies of various added species M. More realistic models for unimolecular reactions take into account the energy dependence of the rate constants for the elementary reactions.

1.2.1 Isomerizations. Many unimolecular reactions are isomerizations, i.e., changes in molecular structure with no change in the elemental formula of the reactant. Examples are

$$\text{cyclopropane } (C_3H_6) \rightarrow \text{propene } (CH_2{=}CH\text{–}CH_3), \qquad (9)$$

$$CH_3CN \rightarrow CH_3NC. \qquad (10)$$

Note that such reactions may involve bond rupture and new bond formation with the reactant molecule.

1.2.2 Bond Rupture Reactions. A much larger number of unimolecular reactions result in the formation of atoms or radicals by fragmentation of the reactant. Thus, they may be considered the microscopic reverse of the bimolecular reactions considered in Sec. 1.1.4. For example,

$$H + H \rightarrow H_2^* \quad \text{(formation of activated molecule by recombination)}, \qquad (11)$$

$$H_2^* \rightarrow H + H \quad \text{(dissociation of activated molecule)}. \qquad (12)$$

The activation energies of such processes are closely related to the enthalpies of dissociation into the fragments (bond dissociation energies), and in fact, experimental determinations of such activation energies are frequently used to derive bond energies.

1.3 Some Practical Applications of Gas-Phase Reactions

1.3.1 Combustion Chemistry. Many fuels are solids or liquids at room temperature, but at the high temperatures of combustion, gas phase reactions are all-important. The burning of organic material is the source for over 90% of the energy we use. The objectives of combustion chemistry are to provide the knowledge needed to carry out this conversion of chemical energy to thermal energy as efficiently as possible and, simultaneously, to minimize the production of deleterious byproducts such as soot, carbon monoxide, nitrogen oxides, and sulfur oxides. The data base for this understanding is provided by a determination of the complicated reaction mechanisms of combustion and measurements of the rate constants of the elementary reactions involved in the mechanism (Miller and Fisk, 1987).

The methods of investigation rely upon the usual techniques of chemical kinetics (*q.v.*), with the special methods demanded by the fact that reactions must be studied at high temperatures (1000–2000 K). High-temperature experiments are necessary because there are numerous examples of reactions for which data obtained in experiments at lower temperatures cannot be extrapolated to combustion temperatures with useful accuracy. Shock tubes are often used to obtain temperatures as high as 5000 K. In the shock tube, a diaphragm separating a low-pressure reaction mixture from a high-pressure driver gas (such as helium) is suddenly ruptured and the resulting shock wave initiates reaction by heating and compressing the reaction mixture. An alternative approach, especially suitable for reactions initiated by an atom or radical species, is to use the shock wave only for heating, with the reactive species produced by flash photolysis of a suitable substrate. High-temperature flow systems are also used.

Theoretical approaches are also important for investigating combustion chemistry. At the macroscopic level of understanding, sophisticated methods for solving coupled differential equations are used to model the kinetics of the combustion process, including the chemical reactions, the convective fluid motion, and the molecular diffusion of the major chemical species. On the molecular level, quantum-mechanical calculations of the potential energy surface governing the interaction of the reactants are made using either semiempirical or *ab initio* methods. The potential energy surface thus obtained can be used either for classical trajectory calculations of the reaction rate constant, or for a simpler statistical mechanical method known as transition-state theory. Once the rate constants for elementary reactions have been measured or calculated, another major step is required to relate these to the processes actually taking place in a flame. Not only are chemical reactions taking place, but fuel and oxygen are flowing into the flame, while heat and reaction products flow out of the flame. Using laser-based spectroscopy as a diagnostic tool, it has recently become possible to measure concentration distributions of some of the major reactants and products within the flame, as well as the temperature profile.

The simplest hydrocarbon fuel, and one that has been extensively studied, is methane (CH_4). Nevertheless, even for this simple molecule, the reaction mechanism of combustion is fairly complicated, involving many steps (Fig. 2). The determination of the corresponding reaction mechanism for a larger fuel molecule, such as octane (C_8H_{18}), or for aromatic hydrocarbons, such as benzene (C_6H_6), is a formidable problem. However, the problem is somewhat simplified by the fact that many of the elementary reactions in the overall mechanism are also found in mechanisms for smaller fuel molecules. Thus a kind of "bootstrap" approach can be used. The general steps involved in the combustion of a saturated hydrocarbon such as methane are as follows:

1. Abstraction of a hydrogen atom from the hydrocarbon by an atomic or radical reactant such as H, O, OH, HO_2, CH_3O, or CH_3O_2, to form a hydrocarbon radical.
2. Alternatively, formation of a hydrocarbon radical by dissociation of a hydrocarbon molecule.
3. Formation of oxygenated species by reactions of hydrocarbon radicals with oxygen atoms or molecules. The oxygenated species are further oxidized, with CO as the final product of the sequence.
4. Recombination and disproportionation of hydrocarbon radicals to form larger hydrocarbons, larger hydrocarbon radicals, or unsaturated hydrocarbons.
5. Simultaneously occurring, a number of reactions that do not involve organic species at all, with reactants such as H, O, OH, HO_2, H_2, O_2,CO, and H_2O.

Two of the most important reactions in the combustion scheme are the chainbranching step

$$H+O_2 \rightarrow OH+O \qquad (13)$$

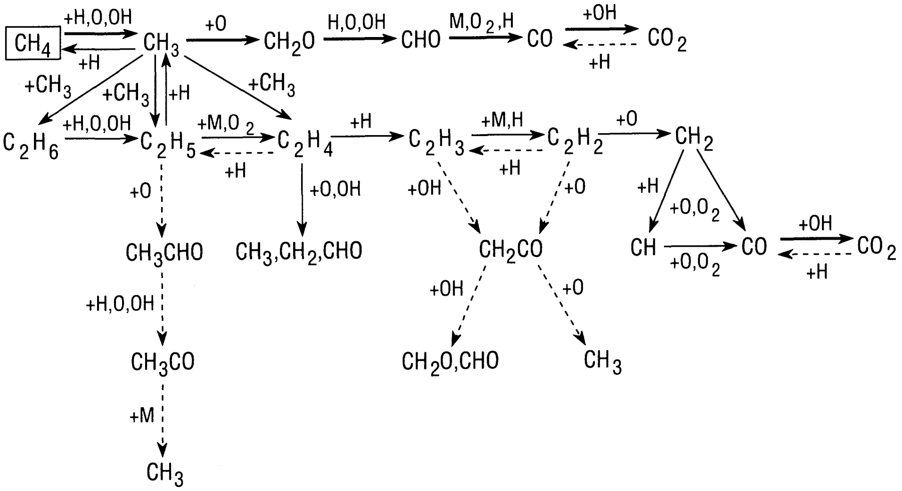

FIG. 2. A diagram that shows the mechanism of combustion of methane in a stoichiometric methane/air flame at atmospheric pressure. Only reactions involving carbon-containing species are indicated. Thickness of arrows indicates the importance of the reaction step, with the least important reactions indicated by dashed arrows. [Reprinted from Miller and Fisk (1987). Published 1987 by the American Chemical Society.]

and the step that converts CO to the final oxidation product CO_2:

$$OH + CO \rightarrow H + CO_2. \qquad (14)$$

Both of these reactions have been extensively studied, experimentally and theoretically, and there is substantial agreement between the rate constants obtained by various methods.

Of course, it is an oversimplification to consider only hydrocarbons as constituents of fuels, which in the real world also may contain significant amounts of oxygenated species, sulfur-containing compounds, and nitrogen-containing compounds. The latter two classes of compounds are important as sources of the nitrogen and sulfur oxides involved in air pollution. Major research efforts are directed towards an understanding of the additional reactions of these nonhydrocarbon species, as well as the combustion of alternative fuels such as methyl and ethyl alcohols.

1.3.2 Air Pollution: Photochemical Smog. Major cities of the world that are blessed with abundant sunshine are often cursed with urban air pollution that has come to be known as "photochemical smog," a specific form of air pollution first recognized in the Los Angeles area during the mid-1940s, but not unique to that locality. The factors that contribute to this problem are abundant sunshine, frequent temperature inversions that effectively place a lid on the lower atmosphere, a geography in which winds from the Pacific are blocked by mountain ranges to the east of the urban area, and a transportation system that relies on large numbers of private automobiles (see AIR POLLUTION: COMPONENTS, CAUSES, AND CURES). Other cities that share some of these characteristics may also suffer from photochemical smog. The term "smog" (=smoke+fog) was invented to describe London air pollution and, as the name implies, particulate matter is an important constituent of smog. Nonetheless, homogeneous gas-phase reactions also play a significant role in the generation of photochemical smog (Finlayson and Pitts, 1976; Okabe, 1978; Wayne, 1985).

The laboratory study of these reactions has relied extensively on the use of "smog chambers"—very large (thousands of liters) reaction vessels that can be filled with a mixture of gases to simulate a polluted atmosphere and then illuminated with sunlight or a powerful arc lamp as its surrogate. The resulting chemical changes can be measured by taking samples of the reaction mixture as reaction proceeds, or in a more sophisticated approach, spectroscopic methods can be used to follow changes in reactant and product concentrations continuously. Once specific reactions have been identified, they can be studied in isolation, using more conventional laboratory techniques. Direct studies of the polluted atmosphere rely on air samples taken at many locations, or on remote sensing methods such as long-path infrared absorption spectroscopy. Aircraft may be used either to collect air samples at various altitudes, or to carry analytical instrumentation aloft for *in situ* measurements. Theoretical modeling utilizes measured or estimated reactant concentrations and rate coefficients, together with data describing air mass motion to predict the spatial and temporal distribution of pollutants.

Measurements of pollutant concentrations in the atmosphere show a pronounced diurnal variation, which can be explained by the rates of dark reactions compared with those initiated by sunlight, and by the variation in traffic load during the day (Fig. 3). Typically, nitric oxide (NO) concentrations increase during the night and early morning, with nitrogen dioxide (NO_2) lagging somewhat behind. By the middle of the day there are high concentrations of nitrogen dioxide and other oxidants such as ozone (O_3) and PAN (peroxyacetyl nitrate). The latter species is in equilibrium with NO_2 and the peroxyacetyl radical and, therefore, can serve as a reservoir for nitrogen dioxide. The source of NO is the reaction of nitrogen and oxygen from the air at the high temperatures of the internal combustion engine, especially diesel engines. Other constituents of automobile exhaust are carbon monoxide and dioxide, partially oxidized and unburned hydrocarbons, and sulfur dioxide from sulfur-containing fuels. The reaction of NO with oxygen in the atmosphere is too slow to account for the formation of NO_2, but the rapid reaction with ozone could be significant. However, as we have seen, the ozone concentration peaks long after that of NO and NO_2. Evidently, other oxidants, such as hydroxyl and organic peroxy radicals, formed from constituents of automobile exhaust, must be important:

$$HO_2 + NO \rightarrow OH + NO_2, \qquad (15)$$

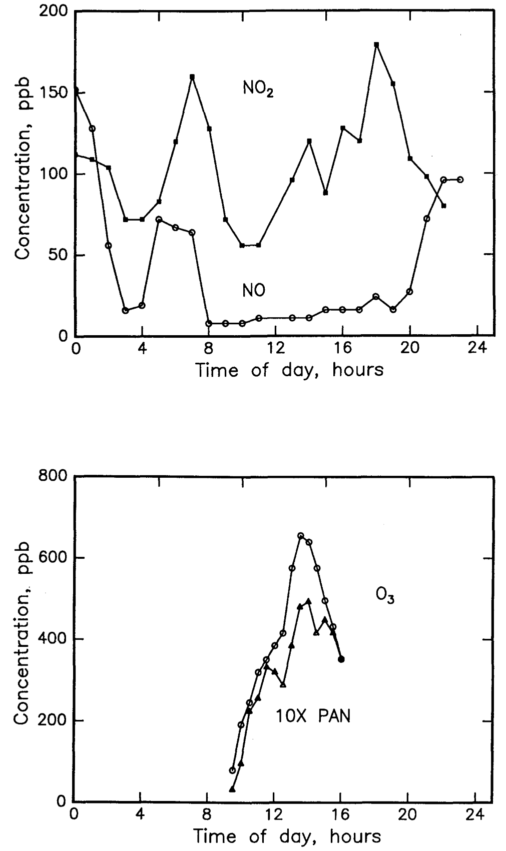

FIG. 3. Diurnal variation of some primary and secondary pollutants in Pasadena, California, on 25 July 1973. (Data from Finlayson and Pitts, 1976.)

$$CH_3O_2 + NO \rightarrow CH_3O + NO_2. \quad (16)$$

The radical initiators may be formed by reactions such as

$$NO + NO_2 + H_2O \rightarrow 2HONO \text{ (nitrous acid)}, \quad (17)$$

$$HONO + \text{light} \rightarrow OH + NO \quad (18)$$

and

$$HCHO + \text{light} \rightarrow H + HCO, \quad (19)$$

$$HCO + O_2 \rightarrow CO + HO_2, \quad (20)$$

$$H + O_2 + M \rightarrow HO_2 + M, \quad (21)$$

$$\text{Net: } HCHO + 2O_2 + \text{light} \rightarrow 2HO_2 + CO. \quad (22)$$

Ozone buildup is hindered by its reaction with NO until most of the free NO has reacted. Then it can be formed by the sequence

$$NO_2 + \text{light} \rightarrow NO + O, \quad (23)$$

$$O + O_2 + M \rightarrow O_3 + M. \quad (24)$$

In the reactions above, M indicates any collision partner that removes enough energy to stabilize the product molecule (cf. Sec. 1.2). It should be noted that the direct photodissociation of oxygen and subsequent reaction of O atoms to form ozone does not take place in the troposphere because the short-wavelength ($\lambda < 200$ nm) UV radiation required for this dissociation does not penetrate through the upper atmosphere. In addition to providing a reservoir for NO_x, PAN is of primary importance as a powerful lachrymator. Its formation involves aldehydes:

$$OH + CH_3CHO \rightarrow CH_3CO + H_2O, \quad (25)$$

$$CH_3CO + O_2 \rightarrow CH_3CO \cdot O_2, \quad (26)$$

$$CH_3CO \cdot O_2 + NO_2 \rightarrow CH_3CO \cdot O_2 \cdot NO_2 \text{ (PAN)}. \quad (27)$$

The reduction in visibility caused by photochemical smog is due mainly to particulate matter, principally organic aerosols formed in secondary reactions. Final products that have been identified include organic nitrates, carboxylic acids and their esters, carbonyl compounds, alcohols, and polymers of peroxides. Inorganic nitrates and sulfates, the latter originating from sulfur-containing fuels, are also found. The mechanism for formation of the organic constituents is not yet clear; however, it seems likely that unsaturated hydrocarbon precursors are oxidized by OH and peroxy radicals.

The deleterious effects of photochemical smog are manifold. Respiratory problems are caused by the oxidants such as ozone and nitrogen dioxide, and the lachrymatory effect of PAN has already been mentioned. Major cities now experience fairly frequent pollution alerts triggered by high ozone levels. The reduction in visibility from aerosols leads to a greater probability of automobile and aircraft accidents. The oxidant species also have a harmful effect on vegetation with serious implications for the agricultural economy. California, as befits its historic role in the understanding of photochemical smog, has taken the lead in measures to control harmful emissions from automobiles, gas stations, dry cleaning establishments, paint solvents, etc. Needless to say, the economic repercussions of such controls are themselves considerable, and it remains to be seen just how extensively they can be applied to the U.S. as a whole.

1.3.3 Homogeneous Catalysis: Destruction of Ozone in the Stratosphere. The triatomic allotrope of oxygen, ozone (O_3), is an important constituent of the earth's upper atmosphere, despite the fact that its maximum partial pressure (at an altitude of about 25 km) is of the order of 10^{-7} bar, corresponding to a mixing ratio of about 10 ppm (Wayne, 1985; Cicerone, 1987). Ozone absorbs biologically deleterious solar radiation between the wavelengths 240 and 320 nm. In addition to experiments showing that such short-wavelength ultraviolet radiation destroys DNA, there is strong epidemiological evidence for a correlation between exposure to solar UV and the incidence of skin cancer, including a frequently fatal form, melanoma. In absorbing the short-wavelength UV radiation, ozone plays an important role as a heat source. The recognition within the past two decades that the ozone concentration could be reduced by man-made pollutants has led to an active research effort to study the reactions involving this relatively unstable species.

The steady-state model proposed in 1930 by Chapman to account for the ambient stratospheric ozone concentration involves only allotropes of oxygen:

$$O_2+h\nu\rightarrow O+O, \tag{28}$$

$$2(O+O_2+M\rightarrow O_3+M)\quad (M=\text{ any molecule}), \tag{29}$$

$$O_3+h\nu\rightarrow O_2+O, \tag{30}$$

$$O+O_3\rightarrow O_2+O_2. \tag{31}$$

Laboratory measurements of the rate constants for these reactions could be combined with known solar UV flux and pressures of major atmospheric constituents to model the ozone concentration in the stratosphere. By about 1964, comparison of such model calculations with stratospheric measurements showed that the measured concentrations were significantly lower than those predicted by the Chapman model. It was apparent that an additional sink for O_3 or O must exist, and it was suggested by Crutzen in 1970 that the nitrogen oxides NO and NO_2 (collectively designated NO_x) could serve as catalysts to remove ozone itself. In this specific example, the proposed mechanism is

$$NO+O_3\rightarrow NO_2+O_2 \tag{32}$$

$$NO_2+O\rightarrow NO+O_2 \tag{33}$$

$$\overline{O+O_3\rightarrow O_2+O_2\text{ (net reaction)}.} \tag{34}$$

A major source of NO_x is nitrous oxide, N_2O, which results from biological sources, some of them man-made. The nitrous oxide reacts with electronically excited oxygen atoms in the ^1D state, produced by photodissociation of ozone or oxygen:

$$O(^1D)+N_2O\rightarrow NO+NO. \tag{35}$$

Reactions (32) and (33) constitute a "chain reaction," one that can go through many cycles before the catalyst is removed by some competitive process. The number of cycles per catalyst molecule is called the "chain length." In this particular example, the most important reaction for removal of NO_2 is

$$NO_2+OH+M\rightarrow HNO_3+M, \tag{36}$$

and in fact, nitric acid serves as a reservoir for the nitric oxides, which are regenerated by the relatively slow photolysis of the nitric acid. To complicate this scheme, the hydroxyl radical itself can enter into *another* catalytic cycle by reactions such as

$$OH+O_3\rightarrow HO_2+O_2 \tag{37}$$

$$HO_2+O_3\rightarrow OH+O_2+O_2 \tag{38}$$

$$\overline{2O_3\rightarrow 3O_2\text{ (net reaction)}.} \tag{39}$$

Coupling between different catalytic cycles like this, together with the large number of reservoir species that must be considered, leads to a very complicated set of chemical reactions (Wayne, 1985).

Harold S. Johnston recognized in 1971 that the NO_x catalytic cycle could be exacerbated by the injection into the stratosphere of large amounts of NO_x from the fleet of supersonic transport aircraft being planned for construction in the U.S. As a result of this insight, an extraordinary scientific effort was set into motion. As one participant said a few years later: "More has been learned about the stratospheric chemistry of ozone in the past few years than in all preceding time."

An even more significant anthropogenic threat to stratospheric ozone is now known to be the family of organic chemicals called chlorofluorocarbons or CFC's, such as CF_2Cl_2. (These are often called "Freons," actually a

DuPont trade name.) Their practical uses as aerosol propellants, plastic foaming agents, and refrigerants have led to an exponential growth in the manufacture of these compounds over the past few decades. The concentration of these compounds in the atmosphere was found to indicate that roughly the total amount that had been manufactured was still intact, if very dilute (Lovelock, Maggs, and Wade, 1973). This evident inertness even led to the suggestion that these compounds would be valuable as atmospheric tracers. However, Molina and Rowland (1974) realized that because of the chemical inertness of CFC's, the only escape route was diffusion into the stratosphere, where the short-wavelength solar UV has not yet been filtered out by ozone, followed by photodissociation:

$$CF_2Cl_2 + \text{light} \rightarrow CF_2Cl + Cl. \qquad (40)$$

The chlorine atoms then enter into a catalytic cycle:

$$Cl + O_3 \rightarrow ClO + O_2 \qquad (41)$$

$$\underline{ClO + O \rightarrow Cl + O_2} \qquad (42)$$

$$O + O_3 \rightarrow 2O_2 \quad (\text{net reaction}). \qquad (43)$$

This cycle is coupled to the NO_x cycle by reactions such as

$$ClO + NO_2 + M \rightarrow ClONO_2 + M, \qquad (44)$$

which *remove* active chlorine species. Dramatic confirmation of this Cl atom cycle was obtained in 1987, based on simultaneous high-altitude measurements of ozone and ClO in the Antarctic. These measurements were part of a major effort to understand the ozone depletion in this region revealed by both airborne and land-based measurements over the preceding two decades. Not only was there a slow decrease of about 30% from 1959 to 1986, but there was a striking drop in ozone in the Antarctic spring of each year. It became apparent that the chemical processes are closely coupled to large-scale atmospheric movements, and the darkness prevailing during the Antarctic winter. A stream of air—the polar vortex—circles Antarctica at this time, trapping air that reaches temperatures as low as −90 °C, cold enough to form clouds even in the very dry stratosphere. These clouds provide a mechanism for heterogeneous reactions that both sequester NO_x and convert chlorine reservoirs such as HCl (hydrochloric acid) and $ClONO_2$ (chlorine nitrate) into active species of chlorine. As Fig. 4 shows, high-altitude measurements with instruments aboard a NASA aircraft demonstrate the simultaneous increase in ClO and decrease in O_3 that would be predicted from reactions (41) and (42).

This brief and simplified description of the atmospheric ozone problem indicates the utility of basic information about chemical reactions, obtained in laboratory experiments, for understanding large-scale phenomena of great practical importance. In the present example, as a result of this understanding, the Montreal protocol of 1987 binds a number of countries to an agreement to cut CFC emissions in half by the end of the century. Major producers of

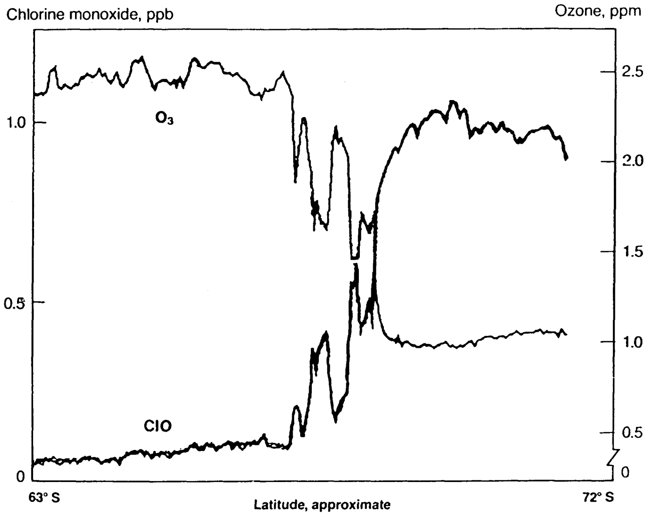

FIG. 4. Concentrations of ozone and chlorine monoxide as measured by instruments aboard a NASA research plane flying from Punta Arenas, Chile (53° S), to 72° S. The ozone "hole" is shown by the precipitous drop in ozone concentration. [Reprinted from Zurer (1988). Published 1988 by the American Chemical Society. Figure used by permission of Prof. J. G. Anderson.]

CFC's are at the present time vigorously searching for suitable substitutes.

2. REACTIONS IN THE LIQUID PHASE AND IN SOLUTION

The effect of a solvent on a chemical reaction can range from trivial to profound. In the latter category are many ionic reactions, which take place readily in polar solvents, but occur in the gas phase only under extreme conditions. The basic reason for this is made apparent by a comparison of heats of ionization in the gas phase with those in water:

$$\mathrm{HCl(g)} \rightarrow \mathrm{H^+(g)} + \mathrm{Cl^-(g)} \quad \Delta H^0 = 1360\ \mathrm{kJ/mol}, \tag{45}$$

$$\mathrm{HCl(aq)} \rightarrow \mathrm{H^+(aq)} + \mathrm{Cl^-(aq)} \quad \Delta H^0 = -104\ \mathrm{kJ/mol}. \tag{46}$$

The large difference between the enthalpies of ionization results from the highly exothermic solvation of the ions compared with a small enthalpy of solvation of the undissociated HCl.

Another important difference between reactions in the gas phase and those in the liquid phase is that in a bimolecular reaction, the reacting species must diffuse towards each other by a series of encounters with solvent molecules until they reach a distance at which a charge transfer or atomic rearrangement can take place. Of course, an important exception to this situation is a reaction involving the solvent itself. This will be discussed below, particularly for the case of acid–base reactions in water.

2.1 Ionic Reactions

Although difficult to define precisely in a condensed phase, the mean free path of a molecule or ion is small compared to molecular dimensions, in contrast to the situation in the gas phase. The random Brownian motion of a solute through a solvent can be treated with Fick's laws of diffusion, leading to a rate constant for binary encounters of species A and B (Weston and Schwarz, 1972):

$$k_D = 4\pi(D_A + D_B) d_{AB}(N^0/1000). \tag{47}$$

In this expression, D_i is the diffusion coefficient for species i, d_{AB} is the encounter distance, and the last term includes Avogadro's number N^0 to give a rate constant in units of $M^{-1}\ s^{-1}$. Reasonable values for the diffusion coefficients and d_{AB} lead to a rate constant of about $5 \times 10^9\ M^{-1}\ s^{-1}$. Thus, an upper limit is set on the rate at which binary encounters can take place, and a reaction occurring with a rate constant of this order of magnitude is said to be diffusion-controlled. Of course, other factors such as energy barriers or geometric factors will often reduce the rate constant to a lower value. Examples of reactions that appear to be diffusion-controlled are

$$\mathrm{I} + \mathrm{I} \rightarrow \mathrm{I_2}\ \text{in}\ \mathrm{CCl_4} \quad (13, 8.2), \tag{48}$$

$$\mathrm{OH} + \mathrm{OH} \rightarrow \mathrm{H_2O_2}\ \text{in}\ \mathrm{H_2O} \quad (6, 5). \tag{49}$$

The numbers in parentheses are the calculated and observed rate constants, respectively, in units of $10^9\ M^{-1}\ s^{-1}$.

Reactions between ions can also be treated by the diffusion equations, provided the electrostatic potential energy is taken into account. This introduces into Eq. (47) for the rate constant an additional factor

$$-z_A z_B r_0/d_{AB}[1 - \exp(z_A z_B r_0/d_{AB})],$$

where z_i is the ionic charge. The constant r_0 depends on the dielectric constant of the solvent and the temperature; in water at 25 °C the value is 7.1×10^{-8} cm. With this modification, the diffusion model is again found to be appropriate for many reactions such as those of solvated electrons in water with both positive and negative ions. Also, all reactions of hydrated protons (hydronium ions) with strong bases that have been studied have proven to be diffusion limited.

The situation becomes more complicated if the ionic solution cannot be assumed to be infinitely dilute. In this case the electrostatic potential of the entire ionic atmosphere must be taken into account. This leads to a dependence of the rate constant on the ionic strength of the solution, defined as

$$\mu = 0.5 \sum_i c_i z_i^2, \tag{50}$$

where c_i is the concentration and z_i is the charge of the ion. The sum is over all ionic species present. In terms of the rate constant at infinite dilution k^0, the rate constant is given by

$$\log k(\mu) = \log k^0 + [1.02\, z_A\, z_B\, \mu^{1/2}/(1+\mu^{1/2})] \quad (51)$$

in water at 25 °C. It can be seen from this expression that reactions between ions of like charge are facilitated by an increase in ionic strength, whereas the opposite effect is observed for reactions between ions of opposite charge (Fig. 5). There may also be small effects due to specific interactions such as ion pair formation, or due to interactions of ions with neutral but dipolar molecules.

2.2 Acids and Bases: Proton Transfer Reactions

In 1923, Brönsted and Lowry introduced a theory of acid–base reactions in which an *acid* was defined as any species capable of donating a proton to a *base*, a species capable of accepting a proton. Thus, the general reaction can be written as

$$AH + B \rightleftharpoons A + BH. \quad (52)$$

AH and A represent an acid and its conjugate base, while B and BH represent a base and its conjugate acid. A more concrete example is provided by the ionization of a strong acid such as hydrochloric acid, in water:

$$HCl + H_2O \rightarrow H_3O^+ + Cl^-. \quad (53)$$

Notice that this expression tacitly assumes that the proton exists in water as the hydronium ion, H_3O^+. In fact, further solvation may take place, but this formulation at least conveys the desired impression that the bare proton itself does not exist in aqueous solution. The solvent, water, has unique properties in that it can act as both acid and base, according to the equilibrium

$$H_2O(\text{acid}) + H_2O(\text{base}) \rightleftharpoons H_3O^+(\text{acid}) + OH^-\ (\text{base}). \quad (54)$$

The equilibrium constant for this reaction, expressed as the ion product

$$K_w = [H_3O^+][OH^-],$$

is 1×10^{-14} at 25 °C. Somewhat surprisingly, this autoionization reaction is apparently not diffusion controlled, as the observed rate constant is $5\times10^{-7}\ M^{-1}\,s^{-1}$. In this example, water is weak both as an acid and as a base; reactions involving either proton transfer from a strong acid or to a strong base do tend to be diffusion controlled. Rapid proton transfer reactions such as these can be studied by the technique of nuclear magnetic resonance (NMR) spectroscopy. This utilizes the fact that in a magnetic field, the separation of the energy levels (i.e., the transition frequency) of a magnetic nucleus such as a proton depends on the chemical environment of the proton. This environmental effect is known as the "chemical shift." Thus, in the acid–base reaction (52), the chemical shift for H will be different in AH and BH, and two peaks will be observed in the NMR spectrum. As the rate increases, the spectral features become broadened, and finally, if the exchange is rapid enough, a new narrow peak that is a statistically weighted average of those for AH and BH is observed.

Brönsted and Pedersen (1924) studied the decomposition of nitramide,

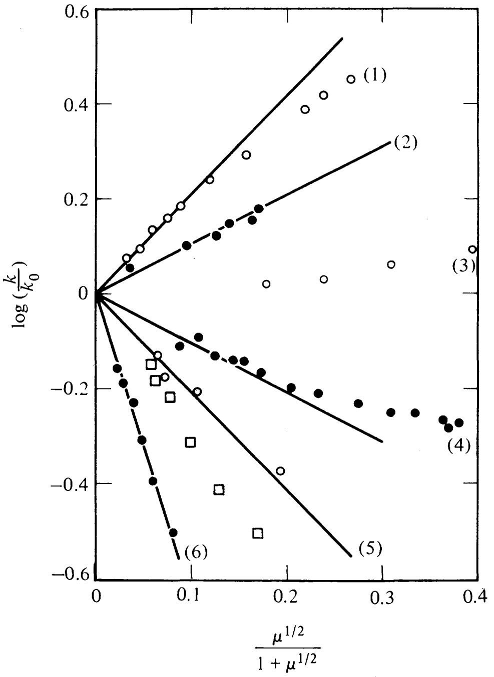

FIG. 5. Effect of ionic strength (μ) on rate constants. The ordinate is the logarithm of the rate constant at ionic strength μ relative to that at $\mu=0$. The reactions are (1) $BrCH_2COO^- + S_2O_3^{2-}$; (2) $e^-(aq) + NO_2^-$; (3) $H_3O^+ + C_{12}H_{22}O_{11}$ (inversion of sucrose); (4) $H_3O^+ + Br^- + H_2O_2$; (5) $OH^- + Co(NH_3)_5Br^{2+}$ [ionic strength varied with NaBr (circles) and Na_2SO_4 (squares)]; (6) $Fe(H_2O)_6^{2+} + Co(C_2O_4)_3^{3-}$. The lines are calculated according to Eq. (51). [From Weston and Schwarz (1972), p. 170. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, NJ.]

$$H_2N_2O_2 \rightarrow H_2O + N_2O, \qquad (55)$$

catalyzed by various bases B. They found a correlation between the base dissociation constant and the rate constant for the reaction. (Actually, they observed an *inverse* correlation with the acid dissociation constant of the conjugate acid, which is related to the base dissociation constant by $K_{HB}K_B = K_w$.) Brönsted further noted that a similar correlation applied to many acid-catalyzed reactions as well as other base-catalyzed reactions. The dependence of the rate constants on equilibrium constants can be expressed as follows:

$$(k_{HA}/p) = A(K_{HA}\,q/p)^{\alpha}$$

for acid catalysis and

$$(k_B/q) = B(K_B\,p/q)^{\beta} \qquad (56)$$

for base catalysis. In these expressions, A and B are constants, and p and q are statistical factors, i.e., the number of equivalent protons on the acid and the number of equivalent sites for accepting protons on the base, respectively. The rate constants k_i are related to the acid and base dissociation constants for the reactions

$$H_2O + HA \rightleftharpoons H_3O^+ + A^-, \quad K_{HA},$$

and

$$H_2O + B \rightleftharpoons HB^+ + OH^-, \quad K_B. \qquad (57)$$

An example of the dependence is shown in Fig. 6, where data for the acid-catalyzed dehydration of acetaldehyde hydrate are plotted. It is seen that the correlation between rate constant and dissociation constant is valid over several orders of magnitude, for the three general classes of compounds for which data are shown. Reactions that obey these rate laws are considered subject to either general acid or general base catalysis. In some other reactions, for example, if the coefficient α is large, the effective catalysis by the hydronium ion will mask acid catalysis by other, weaker, acids. This is a case of specific hydronium ion catalysis, and can be detected by noting that the rate does not depend on the concentration of buffer solutions of HA and A^- in which the concentration of H_3O^+ is kept constant. Other types of acid and base catalysis may not follow the Brönsted rate laws; for example, if K_{HA} is very large, the reaction may be diffusion limited.

2.3 Oxidation–Reduction Reactions

These reactions, often called redox reactions, include some of the most important biochemical reactions, such as those involved in photosynthesis. An oxidation reaction was originally defined as one in which a substance combined with oxygen, and a reduction was one in which oxygen was removed from a compound, as, for example, in the reaction of a metal ore such as Fe_2O_3 with a reducing agent to give the elemental metal. These definitions have been expanded to include all reactions in which the oxidation number of an element is changed. In a simple binary ionic compound, the oxidation number of an element is equal to the number of electrons gained or lost when the compound was formed. In a monatomic ion, the oxidation number is

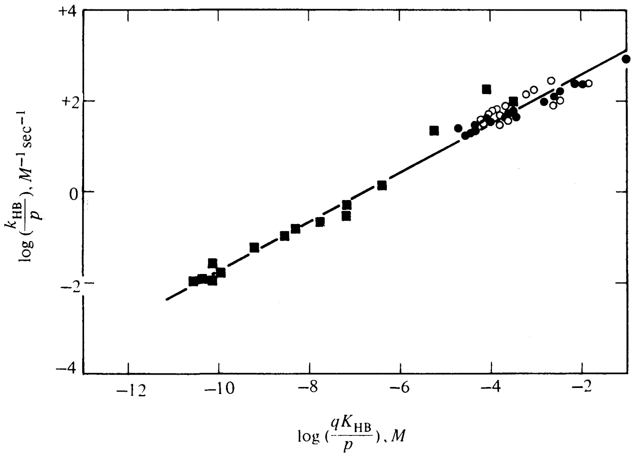

FIG. 6. The Brönsted relation [Eq. (56)] between acid dissociation constant K_{HA} and rate constant k_{HA} for the acid-catalyzed dehydration of acetaldehyde hydrate. Open circles, carboxylic acids; closed circles, substituted benzoic acids; filled squares, phenols. [From Weston and Schwarz (1972), p. 193. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, NJ.]

equal to the charge on the ion. The oxidation numbers of elements in covalent compounds are defined in a more formalistic way, but they still represent shifts of electron density. With these additional rules, an oxidation is now defined as an algebraic increase in the oxidation number (corresponding to a loss of electrons), and a reduction corresponds to a gain of electrons and an algebraic decrease in the oxidation number.

The simplest form of an oxidation–reduction reaction, at least in principle, is provided by the electron transfer process exemplified by

$$\mathrm{Fe}^{3+} + {}^{*}\mathrm{Fe}^{2+} \rightarrow \mathrm{Fe}^{2+} + {}^{*}\mathrm{Fe}^{3+}, \qquad (58)$$

where the asterisk indicates an isotopic label that makes it possible to follow such a reaction. The ferrous–ferric exchange was one of the first electron transfer reactions to be studied in detail (Silverman and Dodson, 1952), and it has the obvious attraction of a system in which no net chemical change takes place and there is no energy of reaction. The reaction follows a bimolecular rate law, but is not as simple as is indicated above, because the ferrous and ferric ions form coordination complexes in aqueous solution. Even in the absence of added ligand species, the ions will be complexed by six water molecules from the solvent. The study of a large number of electron transfer reactions involving transition-metal complexes indicates that they can be roughly divided into "inner sphere" and "outer sphere" processes (Sykes, 1967). Outer sphere reactions are those in which the reaction path does not involve the sharing of a ligand between the two metal atom centers. This mechanism must apply when the rate of ligand substitution on both centers is slower than the rate of the electron transfer process, as in the example:

$$\begin{aligned} &{}^{*}\mathrm{Ru(NH_3)}_6^{2+} + \mathrm{Ru(NH_3)}_6^{3+} \\ &\quad \rightarrow {}^{*}\mathrm{Ru(NH_3)}_6^{3+} + \mathrm{Ru(NH_3)}_6^{2+}. \qquad (59) \end{aligned}$$

The initial step in these reactions is the approach of the two reactants to a distance at which the electron transfer can take place, and this step is subject to the same considerations as those applicable to the reaction of two ions discussed in Sec. 2.1. The time scale for electronic motion ($\sim 10^{-15}$ s) is very rapid compared with that for nuclear motion (10^{-11} to 10^{-13} s), and according to the Franck–Condon principle, internuclear distances and nuclear momenta do not change during an electronic transition (see, for example, Berry, Rice, and Ross, 1980). The theoretical problem to be faced in considering electron transfer reactions is how they can be relatively slow when electronic motion is so fast. The answer lies in the fact that nuclear rearrangements *must* occur in these reactions: metal–ligand bond lengths depend on the metal atom charge, and the dipolar water molecules of the solvent must rearrange themselves to minimize the potential energy after the electron transfer has taken place (Fig. 7). A detailed theory that takes these factors into account has been developed and widely applied (Marcus, 1963; Sutin, 1983). Without going into the detailed formulation of this theory, some of the important factors can be indicated here. These include the equilibrium constant for the reaction (unity if there is no net chemical reaction), and contributions from the energy changes due to the solvent reorganization and the difference in metal–ligand bond length. The solvent reorganization energy can be calculated from a classical electrostatic model, while the inner shell energy depends on the force constant for metal–ligand vibrations and the change in bond length. It has recently become possible to estimate this latter term by measuring the metal–ligand bond lengths of complexes in solution, using EXAFS. The bond-length differences are very small; for example, in the hexa-aquo ferrous–ferric system, the difference is about 10 pm. Nevertheless, when the other factors entering the expression for the electron transfer rate constant

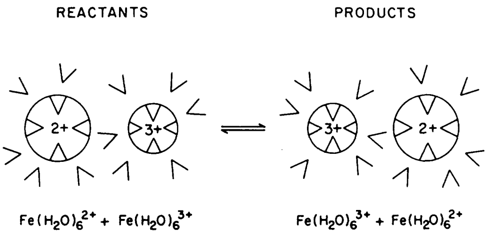
REACTANTS
PRODUCTS
2+
3+
3+
2+
$\mathrm{Fe(H_2O)_6^{2+} + Fe(H_2O)_6^{3+}}$
$\mathrm{Fe(H_2O)_6^{3+} + Fe(H_2O)_6^{2+}}$

FIG. 7. A schematic representation of the change in the equilibrium nuclear configuration of the inner-coordination shells of reactants and the change in orientation of surrounding solvent molecules after charge transfer. The circles represent the inner coordination shells of the metal complexes. The wedges represent water molecules. [From Sutin and Creutz (1983). Reproduced by permission of the Division of Chemical Education, American Chemical Society.]

are taken into account, a good correlation is found between these bond-length differences and the exchange rate constant (Brunschwig *et al.*, 1982).

The Marcus theory also leads to a relationship between the rate constants for self-exchange reactions of the type exemplified in Eqs. (58) and (59), and that of the cross reaction obtained by combining these two reactions:

$$\mathrm{Fe(H_2O)_6^{3+} + Ru(NH_3)_6^{2+} \rightarrow Fe(H_2O)_6^{2+} + Ru(NH_3)_6^{3+}}. \qquad (60)$$

The rate constant for the cross reaction k_{12} is approximately given by

$$k_{12} = (k_{11}k_{22}K_{12})^{1/2}, \qquad (61)$$

where k_{11} and k_{22} are the rate constants of the two self-exchange reactions, and K_{12} is the equilibrium constant for the cross reaction. The validity of this expression is demonstrated by the data in Table 1.

The mechanism of an inner-sphere reaction requires that the reactants combine to form a single primary bond system with a ligand bonded simultaneously to both metal centers. Various types of indirect evidence for such a mechanism have been offered, but the most persuasive argument is provided when an actual ligand transfer from one metal center to another is established, as in the following example:

Table 1. Comparison of experimental rate constants for cross-reactions with rate constants calculated from the Marcus expression.[a]

	k_{12} ($M^{-1}\,s^{-1}$)	
Reaction	Observed	Calculated
$\mathrm{IrCl_6^{2-} + W(CN)_8^{4-}}$	6.1×10^7	6.1×10^7
$\mathrm{IrCl_6^{2-} + Fe(CN)_6^{4-}}$	3.8×10^5	7×10^5
$\mathrm{IrCl_6^{2-} + Mo(CN)_8^{4-}}$	1.9×10^6	9×10^5
$\mathrm{Mo(CN)_8^{3-} + W(CN)_8^{4-}}$	5.0×10^6	4.8×10^6
$\mathrm{Mo(CN)_8^{3-} + Fe(CN)_6^{4-}}$	3.0×10^4	2.9×10^4
$\mathrm{Fe(CN)_6^{3-} + W(CN)_8^{4-}}$	4.3×10^4	6.3×10^4
$\mathrm{Ce^{IV} + W(CN)_8^{4-}}$	$>10^8$	4×10^8
$\mathrm{Ce^{IV} + Fe(CN)_6^{4-}}$	1.9×10^6	8×10^6
$\mathrm{Ce^{IV} + Mo(CN)_6^{4-}}$	1.4×10^7	1.3×10^7
$\mathrm{L\text{-}Co[(-)PDTA]^{2-} + Fe(bipy)_3^{3+}}$	8.1×10^4	$\geqslant10^5$
$\mathrm{L\text{-}Fe[(-)PDTA]^{2-} + Co(EDTA)^-}$	1.3×10^1	1.3×10^1
$\mathrm{L\text{-}Fe[(-)PDTA]^{2-} + Co(ox)_3^{3-}}$	2.2×10^2	1.0×10^3
$\mathrm{Cr(EDTA)^{2-} + Fe(EDTA)^-}$	$\geqslant10^6$	10^9
$\mathrm{Cr(EDTA)^{2-} + Co(EDTA)^-}$	$\simeq3\times10^5$	4×10^7
$\mathrm{Fe(EDTA)^{2-} + Mn(CyDTA)^-}$	$\simeq4\times10^5$	6×10^6
$\mathrm{Co(EDTA)^{2-} + Mn(CyDTA)}$	9×10^{-1}	2.1
$\mathrm{Fe(PDTA)^{2-} + Co(CyDTA)^-}$	1.2×10^1	1.8×10^1
$\mathrm{Co(terpy)_2^{2+} + Co(bipy)_3^{3+}}$	6.4×10	3.2×10
$\mathrm{Co(terpy)_2^{2+} + Co(phen)_3^{3+}}$	2.8×10^2	1.1×10^2
$\mathrm{Co(terpy)_2^{2+} + Co(bipy)(H_2O)_4^{3+}}$	6.8×10^2	6.4×10^4
$\mathrm{Co(terpy)_2^{2+} + Co(phen)(H_2O)_4^{3+}}$	1.4×10^3	6.4×10^4
$\mathrm{Co(terpy)_2^{2+} + Co(H_2O)_6^{3+}}$	7.4×10^4	2×10^{10}
$\mathrm{Fe(phen)_3^{2+} + MnO_4^-}$	6×10^3	4×10^3
$\mathrm{Fe(CN)_6^{4-} + MnO_4^-}$	1.3×10^4	5×10^3
$\mathrm{V(H_2O)_6^{2+} + Ru(NH_3)_6^{3+}}$	1.5×10^3	4.2×10^3
$\mathrm{Ru(en)_3^{2+} + Fe(H_2O)_6^{3+}}$	8.4×10^4	4.2×10^5
$\mathrm{Ru(NH_3)_6^{2+} + Fe(H_2O)_6^{3+}}$	3.4×10^5	7.5×10^6
$\mathrm{Fe(H_2O)_6^{2+} + Mn(H_2O)_6^{3+}}$	1.5×10^4	3×10^4

[a] Data from Bennett (1973).

$$\mathrm{Co(NH_3)_5X^{2+} + Cr(H_2O)_6^{2+} \rightarrow [(NH_3)_5Co..X..Cr(H_2O)_5]^{4+} + H_2O \rightarrow Co(NH_3)_5(H_2O)^{2+} + Cr(H_2O)_5X^{2+}} \quad (\mathrm{X{=}halide\ ion}). \tag{62}$$

The rate of such a reaction is found to depend on the specific halide ion, being determined by the relative strengths of the metal–halide bond being broken and that being formed.

The limiting case of the inner-sphere mechanism is provided by an intramolecular electron transfer within a compound containing two metal atoms. An extension of the Marcus theory to this case indicates a strong dependence of the rate constant on the distance between the metal centers, due to the effect of the charge separation on the solvent reorganization energy and also due to the distance dependence of the electronic coupling between the metal atoms. This dependence has been demonstrated experimentally by measuring electron transfer rates for a series of compounds of the type

$$\mathrm{[(NH_3)_5Os(III)...(L)_{\mathit{n}}...Co(NH_3)_5(III)]^{5+}},$$

where L_n is a peptide bridge of varying length (Isied *et al.*, 1985, 1988). Pulse radiolysis in aqueous solution leads to rapid reduction of Os(III) to Os(II), followed by intramolecular oxidation of Os(II) and reduction of Co(III). The rate constant for electron transfer varies over seven orders of magnitude for L=proline as n increases from zero to four.

2.4 Catalysis by Enzymes

Earlier in this section, acid and base catalysis in aqueous solution is discussed; in Sec. 1.3.3 an example of catalysis in the gas phase is given, and in Sec. 3.2.5 the important topic of heterogeneous catalysis on surfaces is treated briefly. The importance of these types of catalysis pales by comparison with enzyme catalysis which is essential to life itself, as it is involved in numerous biochemical reactions. The effects of enzymes on chemical reactions are profound—increases in reaction rates by factors of 10^{10} to 10^{14} over the uncatalyzed reaction are possible. Moreover, this type of catalysis (unlike industrial heterogeneous catalysis) takes place at ambient temperatures and atmospheric pressure in neutral aqueous solutions.

Enzymes are proteins—copolymers of amino acids with the general formula $\mathrm{RCH(NH_2)COOH}$ or $\mathrm{RCH(NH_3^+)COO^-}$ that are fastened together by peptide linkages:

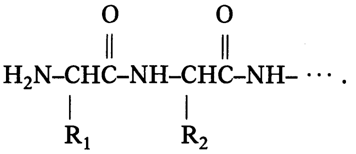

$$\mathrm{H_2N{-}\underset{\displaystyle R_1}{\underset{|}{C}}H\overset{\displaystyle O}{\overset{\|}{C}}{-}NH{-}\underset{\displaystyle R_2}{\underset{|}{C}}H\overset{\displaystyle O}{\overset{\|}{C}}{-}NH{-}} \cdots .$$

The number of amino acid units in these proteins is very large, of the order of hundreds, and the molecular weights are of the order of 10^4–10^5 daltons. Nevertheless, they are specific chemical compounds that can be obtained in pure form, and can even be crystallized in some cases. Only a very small part of the very large enzyme molecule, the "active site," is intimately involved in the catalytic reaction. The rest of the molecule may serve as a skeleton which constrains the active site to the appropriate geometry.

A complete description of the mechanism of action of an enzyme contains the following information (Fersht, 1977):

1. The sequence of amino acids in the protein. Automated methods can now be used for this determination.
2. The detailed structure of the active site and of the enzyme–substrate complex. The first x-ray determinations of protein structures (myoglobin and hemoglobin) were carried out in the 1950s, and this pioneering work is described in the 1962 Nobel Prize lectures (Kendrew, 1963; Perutz, 1963). Since that time the structures of some 500 to 1000 proteins, including many enzymes, have been determined by diffraction methods. Enzyme–substrate complexes have a fleeting existence along the reaction path, but it is sometimes possible to determine structures of less reactive analogs, or of enzymes blocked with inhibitors that occupy the active site. Only a limited number of functional groups can be used by an enzyme in the catalytic process, and it is not surprising, therefore, that the active site regions of different enzymes may be very similar.
3. The chemical mechanism of the enzyme action, and a determination of the rate constants for the individual reaction steps. The basic approach to this problem is outlined below.

4. An understanding of the magnitudes of the rate constants based on thermochemistry and stereochemistry.
5. The specificity of substrates, and the mechanism by which an enzyme can discriminate between substrates of similar structure competing for the active site.

The basic understanding of the rates of enzyme reactions was developed by Michaelis and Menten (1913). Experimentally, it is found that the initial rate of the enzymatic reaction is proportional to the enzyme concentration. At low concentrations of substrate, the rate is also proportional to substrate concentration. However, as the substrate concentration increases, a saturation effect becomes evident, and the rate reaches a maximum value (Fig. 8). This effect can be expressed by the rate equation

$$v=k_{\mathrm{cat}}[E]_0[S]/(K_m+[S]), \tag{63}$$

where v is the rate, k_{cat} is a rate constant, K_m is essentially an equilibrium constant, called the Michaelis constant, $[E]_0$ is the initial concentration of free enzyme, and $[S]$ is that of the substrate. An observed rate constant can be defined from the expression

$$v=k_{\mathrm{obs}}[E]_0,$$

from which

$$k_{\mathrm{obs}}=k_{\mathrm{cat}}[S]/(K_m+[S]). \tag{64}$$

At very high substrate concentrations such that $[S]\gg K_m$, the limit on the rate is

$$v_{\max}=k_{\mathrm{cat}}[E]_0 \quad \text{and} \quad k_{\mathrm{obs}}^{\infty}=k_{\mathrm{cat}}. \tag{65}$$

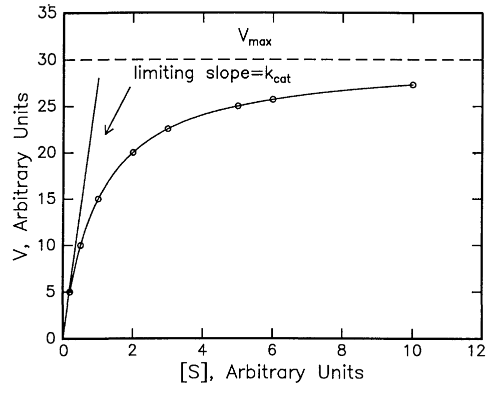

FIG. 8. Reaction rate plotted against substrate concentration for a reaction obeying Michaelis–Menten kinetics.

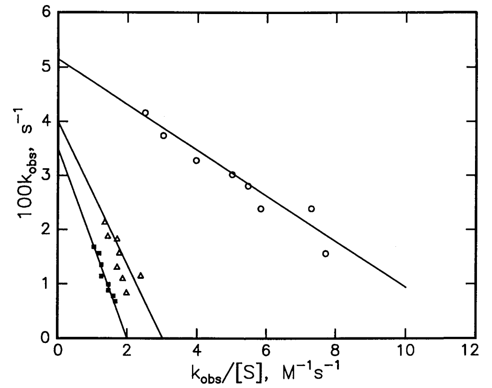

FIG. 9. Data for the α-chymotrypsin–catalyzed hydrolysis of N-acetyl-*L*-tryptophanamide, plotted according to Eq. (67). Circles, *p*H 7.9, $[E]_0=5$ μmol; squares, *p*H 9.6, $[E]_0=5$ μmol; triangles, *p*H 9.6, $[E]_0=20$ μmol. Data from Himoe *et al.* (1967).

The Michaelis constant is given by the substrate concentration at which $v=v_{\max}/2$. At low concentrations of the substrate, where $[S]\ll K_m$, the rate is

$$v=(k_{\mathrm{cat}}/K_m)[E]_0[S],$$

and

$$k_{\mathrm{obs}}^{0}=k_{\mathrm{cat}}[S]/K_m. \tag{66}$$

Rearrangement of Eq. (64) leads to a form linear in $1/[S]$:

$$k_{\mathrm{obs}}=k_{\mathrm{cat}}-K_m k_{\mathrm{obs}}/[S]. \tag{67}$$

A plot of the data in the form of k_{obs} vs $k_{\mathrm{obs}}/[S]$ gives a straight line with an intercept of k_{cat} and a slope of $-K_m$ (Fig. 9). This expression is also written in the form (the Eadie–Hofstee equation)

$$v=v_{\max}-K_m v/[S]. \tag{68}$$

The rate expression of Eq. (63) can be derived from a mechanism

$$\mathrm{E+S\rightarrow ES}, \quad k_1, \tag{69}$$

$$\mathrm{ES\rightarrow E+S}, \quad k_{-1}, \tag{70}$$

$$\mathrm{ES\rightarrow P}, \quad k_{\mathrm{cat}}, \tag{71}$$

where ES is an enzyme–substrate complex and P is the product. Under steady-state conditions (valid when the concentration of ES is much smaller than that of E or S), this mechanism leads to a rate expressed by Eq. (63), with $K_m=(k_{-1}+k_{\mathrm{cat}})/k_1$. Thus, it is seen that the Michaelis constant is a measure of the stability of the enzyme–substrate complex; in fact, if $k_{cat} \ll k_{-1}$, the constant is just the dissociation constant of the complex.

Some of the above general considerations may become more concrete when applied to a specific enzyme, α-chymotrypsin. This compound belongs to a family of enzymes, the serine proteases, that have in common a reactive serine residue, $\text{-NH-CH(CH}_2\text{OH)-C(O)-}$; they catalyze the hydrolysis of peptides and synthetic ester and amide substrates. Chymotrypsin consists of 246 amino acid residues, and their sequence has been determined. The enzyme is produced in the pancreas by the action of another enzyme, trypsin, on the precursor, chymotrypsinogen.

Examples of data for chymotrypsin that fit the Michaelis–Menten mechanism are shown in Fig. 9. A clue to the location of the active site in this enzyme is given by the *p*H dependence of the reaction, which indicates that k_{cat} is sensitive to groups with acid dissociation constants of $\sim 10^{-7}$ M and $\sim 10^{-10}$ M (pK_a's of 7 and 10), whereas K_m is affected by groups with a *p*K of about 9. There are three groups in the enzyme with a pK_a of about 7: a terminal -NH_3^+ group, unlikely to be active, and two histidine residues containing imidazole rings:

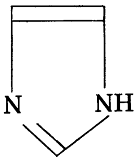

Imidazole is known to have a pK_a of 7.09 in aqueous solution. Thus, at least one of these groups is at the active site. Further reaction studies indicate that this group is histidine-57 (the 57th amino acid residue in the chain). Some substrates, such as *p*-nitrophenyl acetate, form stable acyl–enzyme complexes, and the amino acid sequencing of the complex shows that the acetyl group is linked to serine-195. The acid and alcohol products are not released by the enzyme simultaneously, indicating a slightly more complicated mechanism,

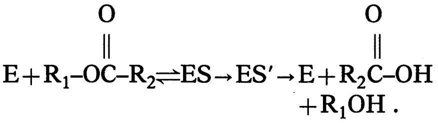

$$E + R_1\text{-}O\overset{\overset{O}{\|}}{C}\text{-}R_2 \rightleftharpoons ES \rightarrow ES' \rightarrow E + R_2\overset{\overset{O}{\|}}{C}\text{-}OH + R_1OH . \qquad (72)$$

Thus, the surprising conclusion is that two amino acid residues separated by 137 other residues are both at the active site. The fact that these two groups are actually close together has been proven by x-ray diffraction studies (Fig. 10). In addition, these structural determinations indicate hydrogen-bonding between the histidine and aspartic acid-102. In fact, some evidence now indicates that the *p*H ~ 7 dependence is due to involvement of this aspartate group in a proton transfer from histidine-57. The effect of higher *p*H mentioned above is now known to be due to the α-amino group of isoleucine-16, which affects the conformation of the enzyme but does not directly influence k_{cat}. This specific example of α-chymotrypsin shows that the combination of structural studies (x-ray diffraction, NMR spectroscopy, infrared spectroscopy) and mechanistic studies can be combined to elucidate the action of an enzyme at the most detailed level.

2.5 Polymerization Reactions

Polymers are very large molecules, made up of recurring structural units, frequently identical. In the discussion of enzyme catalysis in the preceding section, proteins were briefly mentioned as an example of naturally

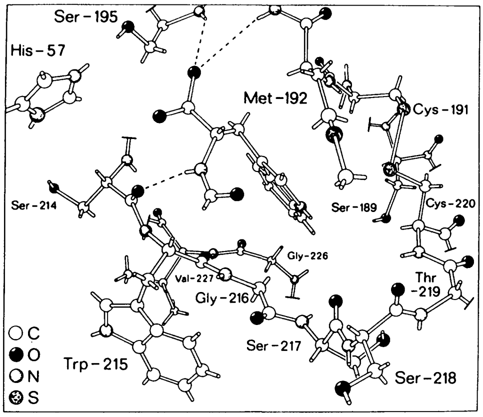

FIG. 10. The active site structure of chymotrypsin. Amino acids are abbreviated as follows: Cys, cysteine; Gly, glycine; His, histidine; Met, methionine; Ser, serine; Thr, threonine; Trp, tryptophan, Val, valine. (From *Enzyme Structure and Mechanism*. By Alan Fersht. © 1977 by W. H. Freeman and Company. Reprinted by permission.)

occurring copolymers, consisting of amino acid residues held together by the amide linkage,

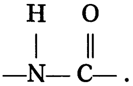

$$\overset{\displaystyle H}{\overset{|}{—N}}—\overset{\displaystyle O}{\overset{\|}{C}}— .$$

Other polymers also occur in nature: starch, cellulose, natural rubber, silk, wool, and DNA (deoxyribonucleic acid) are examples. However, the number of these is overwhelmed by the number of man-made polymers. Only a few of these will be discussed here as examples of general types of polymerization reactions.

Condensation Polymers. One of the oldest synthetic polymers is *Bakelite,* produced by the condensation of phenol, C_6H_5OH, with formaldehyde, H_2CO. The reaction is induced by a basic catalyst that removes a proton from phenol to form the phenoxide ion. Subsequent reactions are:

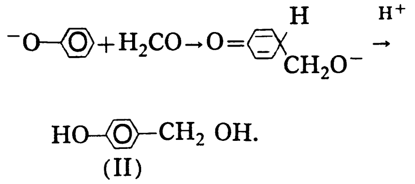

(73)

Further condensation of (II) with another molecule of phenol and loss of a water molecule leads to

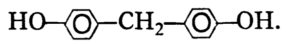

Continuation of this process leads to a cross-linked three-dimensional polymer, with each aromatic ring linked through methylene (CH_2) groups at the 1, 3, and 5 positions of the ring.

Polyester condensations lead to a variety of polymers. For example,

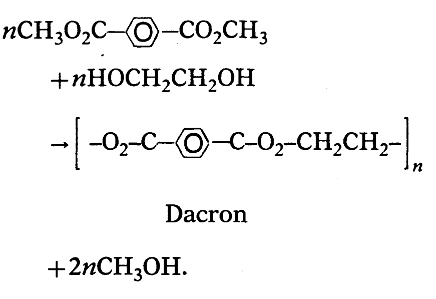

(74)

Polyamides, with linkages like those of proteins, can be made by condensing diamines with dicarboxylic acids:

$$n\mathrm{HO_2C(CH_2)_4CO_2H} + n\mathrm{H_2N(CH_2)_6NH_2}$$

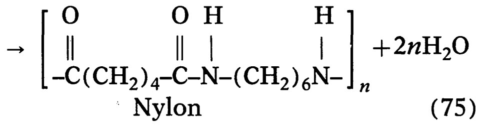

$$\rightarrow \left[-\overset{\displaystyle O}{\overset{\|}{C}}(CH_2)_4-\overset{\displaystyle O}{\overset{\|}{C}}-\overset{\displaystyle H}{\overset{|}{N}}-(CH_2)_6\overset{\displaystyle H}{\overset{|}{N}}- \right]_n + 2n\mathrm{H_2O}$$

Nylon (75)

Addition Polymers. Hydrocarbons containing a double bond (alkenes or olefins) typically form polymers in chain reactions; for example, styrene polymerizes to form polystyrene:

$$n\mathrm{C_6H_5{-}CH{=}CH_2} \rightarrow [-(\mathrm{C_6H_5})\mathrm{CH{-}CH_2}]_n. \quad (76)$$

Free-radical catalysts to initiate these reactions are frequently produced by thermal or photochemical decomposition of organic peroxides, which form alkyl radicals: e.g.,

$$\mathrm{(CH_3)_3C{-}O{-}O{-}C(CH_3)_3} \rightarrow 2\mathrm{(CH_3)_3CO}\bullet \rightarrow 4\mathrm{CH_3}\bullet + 2\mathrm{(CH_3)_2C{=}O}. \quad (77)$$

This chain-initiation step is followed by addition of the radical to the alkene to form a larger radical that can add to yet another alkene molecule; this process is repeated until a chain-terminating step intervenes:

$$\mathrm{R}\bullet + n\mathrm{CH_2{=}CH(C_6H_5)} \rightarrow \mathrm{R[-CH_2{-}CH(C_6H_5)-]}_n. \quad (78)$$

The chain length is large enough so that the product is essentially polystyrene, even if the initiating radical is still part of the polymer molecule. Other alkenes, such as isobutylene (2-methylpropene), undergo acid-catalyzed polymerizations in which a proton adds to the alkene to form a positively charged (cationic) growing chain.

3. REACTIONS INVOLVING THE SOLID PHASE

Chemical reactions that take place entirely in the solid phase are much less common than those taking place in the gas or liquid phase. As in the case of liquid-phase reactions, transport within the condensed phase is often the rate-determining process. From a technological standpoint, the most important class of reactions involving solids is that in which the solid serves as a heterogeneous catalyst for reactions of gases or liquids. Other important reactions of solids, such as corrosion and electrochemical reactions at the surface of metal electrodes, will be discussed in Sec. 4.3.

3.1 Solid–Solid Reactions

In the solid phase, unlike the gas or liquid phase, reactants are restricted from access to each other. Diffusion of one or more reactants is required, through a solid phase that may include both reactant and product species. Mass transfer may take place by bulk diffusion, or may occur along grain boundaries, etc.

Addition reactions are the simplest type of solid–solid reaction. Examples are found in the formation of spinels:

$$\mathrm{ZnO(s)+Fe_2O_3(s)\rightarrow ZnFe_2O_4(s)}, \qquad (79)$$

$$\mathrm{MgO(s)+Al_2O_3(s)\rightarrow MgAl_2O_4(s)}. \qquad (80)$$

In the latter reaction, both Mg^{2+} and Al^{3+} ions diffuse.

In double decomposition reactions, such as

$$\mathrm{AgCl(s)+NaI(s)\rightarrow NaCl(s)+AgI(s)}, \qquad (81)$$

there may be as many as four solid phases involved. In this example, they might be AgCl/NaCl/AgI/NaI, corresponding to diffusion of Ag^+ through NaCl and AgI to react at the AgI/NaI interface, and motion of Na^+ through AgI and NaCl to react at the AgCl/NaCl interface.

3.2 Gas–Solid Reactions

3.2.1 Tarnishing of Metals. Tarnishing is the reaction of a solid metal with a liquid or a gas, typically to form an oxide or a sulfide. An example that again illustrates the importance of material transport is provided by the reaction of silver with molten sulfur at a temperature of 200 °C. An experiment was carried out with the silver surface separated from the liquid sulfur by two separate layers of silver sulfide, Ag_2S, and it was found that the weight of the silver layer decreased as reaction occurred, the weight of the sulfide layer closest to silver remained relatively constant, but the sulfide layer next to sulfur gained weight corresponding to the loss of weight in silver. Thus, the reaction involves diffusion of Ag^+ through the sulfide to react at the silver sulfide–sulfur interface. Electrons flow in the opposite direction to maintain electrical neutrality.

The formation of thick ($\geqslant$100 nm) tarnish films often obeys a linear rate law

$$X=At, \qquad (82)$$

where X is the film thickness, t is the time, and A is a rate constant. The mechanism involves reaction of the gas or liquid with the solid to form a film that does not adhere, so that open surface of the solid is always accessible to the other reactant. Hence, the rate is independent of time and film thickness. Examples are the oxidation of alkali metals and alkaline earth metals. On the other hand, if diffusion through a product forming a coherent solid is rate-limiting, the formation rate follows a parabolic law:

$$X^2=At. \qquad (83)$$

This expression can be derived from the equations for the concentration gradient of metal atoms diffusing to the surface, which is inversely proportional to the film thickness.

The growth of thin films ($<$100 nm) follows several possible rate laws, depending on specific metals and on film thickness. These rate laws reflect the electrostatic effects of space charges in the surface regions, which now represent a significant fraction of the total film thickness. Also, in thin films material transfer occurring at the gas–surface interface may be rate-determining, rather than transport through the film itself. However, the film thickness determines the electric field through which the metal ions move. As the film thickness increases, the potential decreases and the rate of growth slows down. This is the case for the oxidation of aluminum, and indeed, were it not so, aluminum would be virtually useless for practical applications. After an oxide film of about 20 Å is formed (at room temperature), the rate of oxidation decreases to a very low value, and the metal is "passivated" against further reaction.

3.2.2 Chemical Vapor Deposition of Thin Films on Semiconductor Substrates. Chemical vapor deposition (CVD) is a process in which vapor-phase components react to form a solid thin film on some surface. The use of this method for the formation of thin films has become a major process in the production of solid-state electronic devices. It is an extremely versatile technique that permits the formation of solid phases with a wide range of controllable composition at relatively low reaction temperatures. The solids formed, with a high degree of purity and perfection, may be either crystalline or glassy. An indication of the economic significance of CVD is given by the production of silicon for use in solid-state electronics, using the reduction of trichlorosilane ($SiHCl_3$) by hydrogen. The polycrystalline silicon thus produced is grown into single crystals with an impurity content of one part in 10^{11}. About 30 000 metric tons (3×10^7 kg) are produced every year, worldwide (Yablonovitch, 1989).

The chemical reactions taking place in a CVD process can be investigated by determination of the vapor composition *in situ*, using techniques such as mass spectrometry, Raman and absorption spectroscopy, and gas chromatography. In addition, the composition of the solid phase must be determined using techniques such as XPS, Auger spectroscopy, and IR reflectance spectroscopy. Reactors designed for use at both normal and reduced pressures, at ambient and elevated temperatures (1000–1300 °C), are available. At normal pressures, the rates of transport processes and surface chemical reactions are comparable. For this reason, both types of rate-limiting steps must be optimized to ensure uniformity of film characteristics over extended surfaces. The recent innovation of reactors designed for low-pressure CVD has proven to be a significant advance in the semiconductor industry. The advantage is that gas flow is no longer a consideration in the reactor design, since diffusion is no longer the rate-determining step in the process, and the gas-phase composition is more homogeneous. The longer mean free path of the reactants means that limitations on the loading and positioning of the solid substrates are reduced (see CVD).

3.2.3 Vapor Etching of Semiconductors.

The increasing demands on spatial resolution in integrated circuit wafers have made conventional chemical etching with liquids unsuitable. In spite of the name, plasma etching is predominantly a chemical process in which the reactants are produced by an electrical discharge in a low-pressure gas. The reactive species thus produced diffuse to the substrate surface and react to form volatile products. The electrical discharge is typically a microwave rf discharge at a frequency of ~14 MHz and a power level of 0.1 to 1 W/cm^3 of reactor volume.

The chemistry of plasma etching is very complicated and not yet well understood (Jasinski, Meyerson, and Scott, 1987). In the discharge, high-energy electron impact produces charged species by direct and dissociative ionization. Also, lower energy electrons dissociate molecules by processes such as

$$e^- + O_2 \rightarrow e^- + O + O$$

and

$$e^- + CF_4 \rightarrow e^- + CF_3 + F. \tag{84}$$

Electron impact may also lead to electronically excited molecules that dissociate rapidly before they can lose energy in collisions or by radiation.

A specific example is provided by the CF_4–O_2 plasma, which is useful in etching Si, SiO_2, and Si_3N_4. Tetrafluoromethane (CF_4) is itself a very inert gas, which in a discharge produces reactive species such as F, CF, CF_2, and CF_3. The addition of O_2 up to a content of ~30% to the discharge increases the etch rate of silica, but the rate then decreases as more oxygen is added. This is apparently caused by a parallel dependence of the F atom concentration on oxygen pressure, and it is believed that the reaction of oxygen with the CF_3 radical is responsible. The reactor effluent contains CO, CO_2, COF_2, and SiF_4 as stable products. It is important that the gaseous products be relatively stable, since otherwise they may dissociate and redeposit on the substrate. For example, SiH_4 (silane) is volatile, but decomposes in a plasma to form solid deposits of SiH_x. Halocarbons are widely used as sources of halogen atoms and halocarbon radicals because they are relatively inexpensive, easy to handle, noncorrosive, volatile, and relatively nontoxic. (For the deleterious effect of chlorofluorocarbons on the stratospheric ozone layer, cf. Sec. 1.3.3.) Table 2 (from Vossen and Kern, 1978) shows some of the plasma etching reactants and the observed products (see PLASMA ETCHING).

As indicated above, the addition of oxygen to the halocarbon has an important effect on the etching rate. This is in part due to increased concentrations of halogen atoms, and also because the formation of polymers that deposit on the substrate is hindered. Substrate

Table 2. Halocarbon gases for plasma etching.

Etching gas	Etchable materials	Principal fragments and relative intensities from mass spectra					
CF_4	Si,SiO_2,Si_3N_4,Ti, Mo,Ta,W	CF_3	CF_2	F	CF		
		100	12	7	5		
C_2F_6	SiO_2,Si_3N_4	CF_3	C_2F_5	CF_2	CF	C	F
		100	41	10	10	2	1
$CClF_3$	Au,Ti	CF_3	$CClF_2$	CF_2	Cl	CF	
		100	21	10	7	4	
$CBrF_3$	Ti,Pt	CF_3	$CBrF_3$	$CBrF_2$	Br		
		100	15	14	5		
C_3F_8	SiO_2,Si_3N_4	CF_3	CF	C_3F_7	CF_2	C_2F_5	C_2F_4
		100	29	25	9	9	7
CCl_4	Al,Cr_2O_3,Cr	CCl_3	Cl	CCl	CCl_2		
		100	41	41	29		

temperature is another variable that affects the etching rate, as most of the chemical reactions involved have activation energies of 10–50 kJ/mol. For this reason, temperature control is important in establishing reproducible etching conditions.

3.2.4 Adsorption of Gases on Solids. Reactions occurring at the surfaces of solids are extremely important in the production of many industrial chemicals, and in the refining of crude oil into useful fuels and into a variety of feedstocks for other chemical processes. The initial step in any of these heterogeneous catalytic reactions is the adsorption of gases on the solid substrate. The potential chemical reactivity of a surface can be understood if it is imagined that the surface is formed by cleaving or breaking off a fragment of the solid (Fig. 11). In a covalent solid, the surface then contains broken bonds and unsatisfied valencies. (This same principle leads to the excess free energy of a liquid surface, demonstrated by the phenomenon of surface tension.) In a similar way, the surface of an ionic solid has unbalanced electrostatic forces. Chemisorption results when a molecule interacts with the unsatisfied valencies of the solid surface; in effect, this is a chemical reaction since the electronic configurations of the surface and the adsorbed molecule are changed in the process. The process is spontaneous, and hence the free energy change ΔG is negative:

$$\Delta G = \Delta H - T\,\Delta S < 0. \tag{85}$$

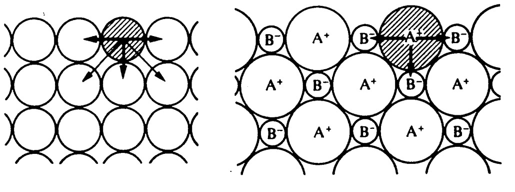

FIG. 11. A schematic representation of the surface energy of a covalent solid (left), and an ionic solid (right). (From Bond, 1987. © 1987 by Oxford University Press. Reproduced by permission of the publishers.)

The adsorbed molecule loses some degrees of freedom when it is confined to the surface, and the entropy change ΔS is negative. Therefore, the enthalpy of adsorption ΔH must also be negative, i.e., the reaction is exothermic. Typical values are -40 to -800 kJ/mol. However, the process is very specific, depending on the metal and even the particular surface plane exposed, and also depending on the adsorbed species. The reverse process of desorption is highly endothermic, and for this reason, chemisorbed layers are stable up to relatively high temperatures. Since the valence forces involved in this adsorption process are short-range forces, only a single layer may be considered to be chemisorbed, although additional layers may accumulate by physical adsorption. In many cases the electronic structure of the adsorbed molecule must be drastically altered to permit reaction with the "broken bonds" of a metal surface. Dissociation may be necessary, as in the case of hydrogen:

$$2M + H_2 \rightarrow 2H\text{-}M. \tag{86}$$

Unsaturated hydrocarbons with a π-electron cloud, such as alkenes or aromatics, can interact by adding on to the "free valencies" of the metal, with subsequent rehybridization of their molecular orbitals. Chemisorption on metal oxide surfaces is less well understood than that on metal surfaces, and differs somewhat between semiconductor and insulator oxides.

Physical adsorption involves forces very similar to those acting between molecules in a liquid or condensed gas: van der Waals forces resulting from dipolar or multipolar interactions, or forces between multipoles induced by fluctuations in electron density. As would be expected, the enthalpies of physisorption are considerably smaller than those associated with chemisorption, and are similar in magnitude to enthalpies of vaporization of liquids (8–20 kJ/mol). The temperatures at which physical adsorption occurs are related to the boiling points of the adsorbates, but aside from this there is not the degree of specificity found in chemisorption. Finally, another distinction from chemisorption is that more than a single layer of adsorbed species can form.

The relation between the amount of adsorbed material and the pressure of the gas phase from which it is adsorbed is known as an adsorption isotherm. In spite of the multitude of known gas–solid interactions, all isotherms appear to be contained within five families (Fig. 12). The simplest of these, the Langmuir isotherm, is based on the following assumptions:

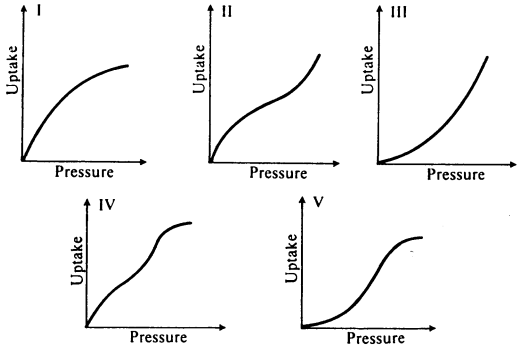

FIG. 12. The five different types of adsorption isotherm. I is typical of a Langmuir isotherm, and II is that for a BET isotherm (see text). (From Bond, 1987. © 1987 by Oxford University Press. Reproduced by permission of the publishers.)

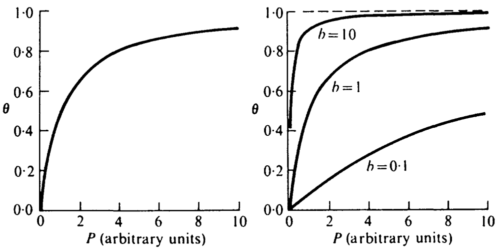

FIG. 13. Langmuir adsorption isotherms, showing the effect of the equilibrium constant for adsorption b. (From Bond, 1987. © 1987 by Oxford University Press. Reproduced by permission of the publishers.)

1. The surface is homogeneous.
2. The surface contains a fixed number of adsorption sites, each of which can be occupied by only a single molecule of adsorbate. (This limits the coverage to a monolayer.)
3. All sites are equivalent, and the adsorption of a single molecule is independent of all surrounding molecules.

These conditions lead to an equilibrium between A, the gas-phase adsorbate, M, the adsorption site, and A–M, the adsorbed species:

$$\mathrm{A(g)+M(s)\rightleftharpoons A\text{–}M(ads)}.$$

The equilibrium constant for this reaction can be expressed in terms of p, the pressure of A at equilibrium, and the fractional surface coverage θ. The ratio [A–M]/[M] is simply $\theta/(1-\theta)$, from which the equilibrium constant is

$$b=\theta/p(1-\theta)$$

and hence,

$$\theta=bp/(1+bp). \qquad (87)$$

The adsorption coefficient b indicates the strength of the interaction (Fig. 13). A convenient way of plotting a Langmuir isotherm is provided by identifying θ with the ratio V/V_m. The volumes V and V_m are those of the adsorbed gas and the equivalent of an adsorbed monolayer, respectively. Then the Langmuir expression can be written in the form (Fig. 14)

$$P/V=(1/bV_m)+(P/V_m). \qquad (88)$$

Another widely used isotherm is the BET (Brunauer, Emmet, and Teller, 1938) isotherm, which is particularly applicable to

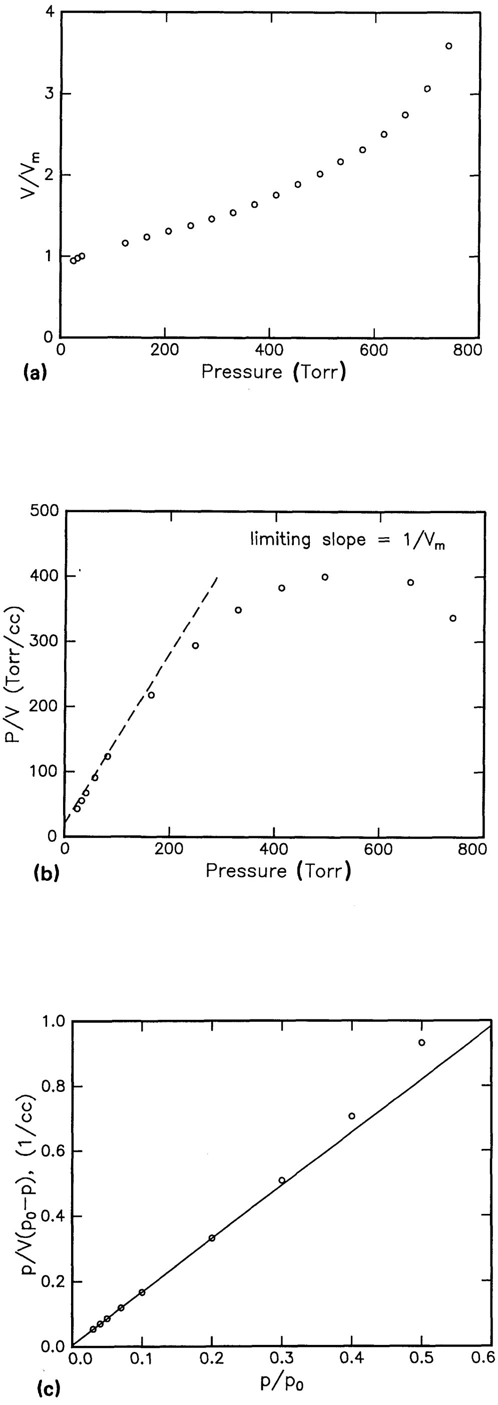

FIG. 14. The adsorption of nitrogen on KCl at 78 K. (a) Adsorption isotherm in the form of $\theta = V/V_m$ vs pressure. (b) Plotted as a Langmuir isotherm; this gives values of 0.77 cm^3 for V_m and 49.4 atm^{-1} for b; (c) plotted as a BET isotherm, yielding 0.58 for V_m and 344 for the constant C. Data from Keenan and Holmes (1949).

physisorption. It assumes neither a monolayer coverage of the surface nor a constant heat of adsorption. Instead, it is assumed that the first layer has an enthalpy of adsorption H_0 while that of subsequent layers is equal to the heat of vaporization of the adsorbate species H_L. The BET isotherm can be derived by considering the equilibrium between each adsorbed layer and the gas phase, with the result

$$\frac{p}{V(p_0-p)} = \frac{1}{V_m C} + \frac{(C-1)p}{V_m C p_0}\,. \tag{89}$$

In this equation, the BET constant is given by $C = \exp(H_0 - H_L)/RT$, where R is the gas constant and T is the absolute temperature. The pressures p and p_0 are the equilibrium pressure and the vapor pressure of the adsorbate at the temperature of the experiment. The BET expression is widely used to obtain the surface areas of catalysts, etc. An inert gas such as Ar or Kr is used as the adsorbate, and measurements of the volume and pressure are made as pressure is changed. A plot of $p/V(p_0-p)$ vs p/p_0 (up to low values of p/p_0) gives a straight line with intercept $1/V_m C$ and slope $(C-1)/CV_m$ (Fig. 14) from which the heat of adsorption for the first layer (from C) and a value of V_m can be derived. With assumptions about the size of adsorbed molecules and their packing on the surface, the volume V_m can be converted to a surface area. In the case of physisorption, where H_0 and H_L are comparable, the BET isotherm gives reasonable results. An expression similar to the BET isotherm can also be derived for a finite number of adsorbed layers. When the number of layers in such an expression is reduced to one, and with the identification of θ with V/V_m, the isotherm becomes identical to the Langmuir adsorption with the equilibrium constant b equal to C/p_0.

The experimental requirements for studies of adsorption, surface chemistry, and heterogeneous catalysis are very demanding. For example, the surfaces to be studied must be prepared under ultra-high vacuum (UHV) conditions. The reason for this can be understood on the basis of the rate at which ambient gaseous molecules collide with a surface. At 298 K, this rate for a gas of mass 28 (N_2 or CO) is 3.8×10^{20} collisions $cm^{-2}\,s^{-1}\,Torr^{-1}$ (1 Torr = 133.3 Pa). A typical metal surface has of the order of 10^{15} sites per square centimeter, and if half the collisions lead to adsorption, the rate at which surface sites are covered is 1.9×10^5 surface layers s^{-1} $Torr^{-1}$. Obviously, if the experiment is to extend over a time period of minutes, the ambient pressure must be of the order of 10^{-8} to 10^{-10} Torr to prevent surface coverage by background gas from blocking the adsorption of the desired species.

The surface scientist has a number of analytical tools at his disposal and seems to delight in designating them with a bewildering collection of acronyms (Gasser, 1985). LEED (low-energy electron diffraction) utilizes the scattering of 20–500 eV electrons from a metal surface to determine the lattice structure, or in some cases, the arrangement of an adsorbed layer. Typically, about 2% of the incident electrons are scattered quasi-elastically, with a penetration depth of the order of one nanometer (i.e., a few atomic layers). In clean metals the lattice spacing is usually within about ±10% of that found for the bulk metal. When a gas is adsorbed on the metal surface, a new electron scattering pattern is observed, which may be due to the metal lattice itself changed by adsorption, or may be caused by scattering from the adsorbate. Several electron emission spectroscopic techniques are used to obtain information about the elemental composition of the adsorbed species and the changes in its electronic configuration upon adsorption. UV photoelectron spectroscopy (UPS), using photons with energies less than 100 eV, provides information about the energy levels of valence electrons. The energy analysis is made by determining the kinetic energy of ejected electrons, given by

$$E_k=h\nu-I-E_{\text{int}}, \tag{90}$$

where the first term is the energy of the UV photon, I is the ionization potential, and E_{int} is the internal (vibrational and rotational) energy of the molecule. According to a theory due to Koopmans, the ionization potential is equal in magnitude to the orbital energy of the electron. Although this is not rigorously correct, it is a useful working approximation. XPS (x-ray photoelectron spectroscopy), also known as ESCA (electron spectroscopy for chemical analysis), is very similar to UPS, except that a monochromatic x-ray source now provides the excitation photons, and the energy is sufficient to eject electrons from core levels of the atom. Auger electron spectroscopy (AES) uses an electron or x-ray source that need not be monochromatic to ionize from core levels. The energy of the primary photoelectron thus ejected is not analyzed. The empty core level with energy E_1 is refilled with an electron from an outer shell, having energy E_2. However, there is still sufficient excitation energy to eject a secondary electron from a level with energy E_3. The energy of this electron is measured, and is given by

$$E_k=E_1-E_2-E_3, \tag{91}$$

so that it is independent of the excitation energy. In principle, Auger electron spectroscopy can also provide information about the chemical bonding of the atom that is ionized, but in practice, spectral resolution limits this application. Instead, Auger spectroscopy is used to determine surface "cleanliness," or the amount of adsorbate on the surface. Other techniques useful in studying chemisorption are EELS (electron energy loss spectroscopy) to determine vibrational energy levels of adsorbed species, and several types of microscopy: field emission microscopy (FEM), scanning tunneling microscopy (STM), etc. Also, EXAFS (extended x-ray absorption fine structure) and related x-ray techniques can be used to obtain information about interatomic distances.

3.2.5 Heterogeneous Catalysis. From an industrial standpoint, the most important chemical reactions that involve solids are those taking place on the surface of catalysts (Bond, 1987; Sinfelt, 1983). Industrial catalysts are produced in many forms: pellets, rings, spheres, tablets, granules, spaghettilike extrudates, etc. The solid phase is not necessarily uniform, as it may consist of a mechanical support, the catalyst itself, and a "promoter" that increases the catalytic activity by a mechanism not always understood. (There is more than a modicum of empiricism in the art of chemical catalysis.) These industrial catalysts may be subjected to extreme conditions of temperature (up to 1600 K) and pressure (350 atm), and yet may have a useful life of several years before they either deteriorate mechanically or are chemically "poisoned" by impurities in the feedstock or products. Catalytic reaction vessels are broadly classified into fixed-bed reactors, in which the catalyst is confined and the reactant gases flow through it, and fluid-bed reactors in which the catalyst consists of fine particles that are sufficiently agitated by flowing gas to behave like a liquid phase.

Some important industrial processes utilizing heterogeneous catalysis are discussed briefly below (see also CATALYSIS).

3.2.5.1 Catalytic Processes Used in the Petroleum and Petrochemical Industry. Many crude oils contain significant amounts (a few percent) of sulfur compounds, such as thiophene, C_4H_4S. These sulfur compounds are undesirable because when burned they produce the air pollutant sulfur dioxide and because the sulfur poisons catalysts used in "cracking" reactions (see below). They can be removed by hydrogenation over catalysts such as Mo–Co sulfides, Ni–Mo sulfides, or Ni sulfide–tungsten, supported on alumina.

Low molecular weight alkanes, such as methane or ethane, can be catalytically oxidized to synthesis gas, a mixture of hydrogen, carbon monoxide, and carbon dioxide. The general reaction is exemplified by that for methane:

$$CH_4+H_2O \rightleftharpoons 3H_2+CO. \qquad (92)$$

About 90% of the world production of hydrogen uses this type of reaction, which is catalyzed by Ni on alumina or magnesia. Depending on the stoichiometry, the water–gas shift reaction

$$CO+H_2O \rightleftharpoons CO_2+H_2 \qquad (93)$$

may also be important. This second step is catalyzed by Fe_3O_4 and Cr_2O_3.

Synthesis gas can be converted to hydrocarbons by a reaction that is essentially the reverse of reaction (92). This Fischer–Tropsch synthesis was used in Germany during World War II to produce liquid hydrocarbon fuels from synthesis gas derived from coal, and it is still used in South Africa. Catalysts for the Fischer–Tropsch synthesis of hydrocarbons are ruthenium and palladium. Other catalysts (Cu–Zn oxides) change the product to oxygen-containing compounds, such as methanol from methane.

A very important class of catalysts is the zeolites, which are aluminosilicates with the general formula $M_v(AlO_2)_x(SiO_2)_y \cdot zH_2O$, where M is a cation. The catalytic properties of these zeolites depend on three factors:

1. their crystal structure forms cavities of a size that depends on the specific zeolite and which can contain only molecules below a certain size;
2. strongly acid hydroxyl groups are exposed on the crystal surface;
3. electrostatic fields of the cations can induce reactivity.

These catalysts are widely used in cracking reactions, which break up large hydrocarbon molecules into smaller ones and also cause some isomerization of normal alkanes to branched-chain alkanes. Thus, the volatility of the hydrocarbons is increased and their antiknock ratings (octane numbers) are enhanced.

As the compression ratio of automobile engines increased, there was a demand for gasoline of higher octane ratings to prevent knocking. This demand could not be met solely by fuels produced in the catalytic cracking process. Needed were catalysts to effect the following general reactions:

1. dehydrogenation of alicyclic molecules such as cyclohexane to aromatics (e.g., benzene);
2. enlargement of cyclopentane rings to cyclohexanes;
3. isomerization of normal (straight-chain) alkanes to branched-chain alkanes;
4. conversion of normal alkanes to alicylics and thence to aromatics;
5. fragmentation (cracking) of alkanes to lower molecular weight alkanes.

This process is called *catalytic reforming*. Typical changes in a hydrocarbon feedstock are:

	Alkanes	Alicyclics	Aromatics	Octane number
Initial %	60	25	15	60
Final %	32	2	66	100

In the 1950s, bifunctional catalysts of ~1% platinum on alumina supports were developed for this purpose. The platinum serves to dehydrogenate alkanes and cycloalkanes to alkenes plus hydrogen. These unsaturated species then migrate to an acidic site on the alumina where a proton transfer reaction converts them to carbonium ions, i.e., molecular cations with the positive charge on carbon. These carbonium ions can undergo isomerization (ring closure or enlargement). After losing a proton, the isomerized species migrates to another Pt site, where it is hydrogenated. More recently (since 1970), bimetallic catalysts of platinum–rhenium or platinum–iridium on alumina have been introduced. In some cases, both of these catalysts are used at different stages of the reforming process. At least part of the beneficial effect of rhenium or indium is due to the inhibition of carbonaceous deposition on the catalyst.

3.2.5.2 Heterogeneous Catalysis in the Heavy Chemical Industry. The synthesis of ammonia by the inorganic fixation of nitrogen

$$N_2+3H_2 \rightarrow 2NH_3 \qquad (94)$$

is an important application of heterogeneous catalysis. The hydrogen for this reaction may be produced from alkanes as described above. Early attempts to find a suitable catalyst were made around the turn of century, and in an "Edisonian" research mode, tests were made of many iron ores reduced to metallic iron. Minor impurities in the ore had significant effects on the catalytic activity. It was eventually realized that alumina and potassium oxide were important constituents. The necessary catalyst is now made by reducing magnetite, Fe_3O_4, together with promoters that may include alumina, magnesia, potassium oxide, calcium oxide, silica, and traces of titanium, zirconium, and vanadium oxides. Large amounts of the ammonia thus produced are destined for use as an agricultural fertilizer, while another major use is in the production of nitric acid. Ammonia is catalytically oxidized to nitric oxide, NO, which is oxidized by air to nitrogen dioxide. This dissolves in water to form nitric acid, HNO_3.

$$3NO_2+H_2O \rightarrow 2HNO_3+NO. \qquad (95)$$

A patent for the catalysis by platinum sponge of the ammonia oxidation was first obtained as long ago as 1838, and the catalyst of choice today remains platinum or a platinum alloy in the form of a finely woven gauze.

4. ELECTROCHEMICAL REACTIONS

Electrochemistry was originally limited to the chemistry of processes taking place at electrodes in aqueous solutions. It is now a wider discipline, including all charge-transfer processes at membranes, which may be biological as well as inorganic (see ELECTROCHEMISTRY).

4.1 Electrochemical Cells

In Sec. 2.3 we described oxidation–reduction, or redox, reactions as those in which an electron is transferred from a reducing agent to an oxidizing agent. If the Gibbs free-energy change ΔG is negative for the reaction, it will occur spontaneously, and it may therefore be possible to obtain useful work from the process. The key to practical implementation is to intercept the electrons in some external circuit and to use the electron flow (or electric current) to do the work. The schematic diagram shown in Fig. 15 depicts a galvanic cell, an electrochemical cell, or simply, a battery. (This specific arrangement is known as the Daniell cell.) Each side of this device comprises a half-cell, with its cell reaction:

Left half-cell:

$$Zn(s) \rightarrow Zn^{2+}(aq)+2e^-, \qquad (96)$$

Right half-cell:

$$2e^-+Cu^{2+}(aq) \rightarrow Cu(s), \qquad (97)$$

Net reaction:

$$Zn(s)+Cu^{2+}(aq) \rightarrow Zn^{2+}(aq)+Cu(s). \qquad (98)$$

Thus, copper ion is reduced to the metal at the cathode (by convention, the right half-cell), while zinc metal is oxidized to the ion at the anode. To prevent the mixing of the copper and zinc sulfate solutions, which would lead to equilibrium and no driving force for the

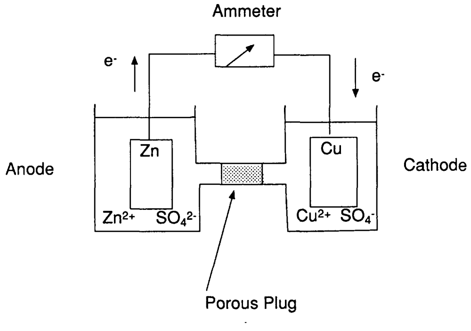

FIG. 15. Schematic diagram of an electrochemical or galvanic cell. This specific cell is known as the Daniell cell.

reaction, a porous barrier that permits electrical contact but not mixing of solutions is interposed between the two half-cells. If the voltage across this cell is measured at 25 °C, it is found to depend on the concentrations of both copper and zinc ions according to the empirical expression

$$\Delta E(\mathrm{V}) = 1.100 - 0.0128 \ln([Zn^{2+}]/[Cu^{2+}]). \qquad (99)$$

However, the measured voltage obviously depends on the combinaton of half-cells, either of which could be combined with some half-cell other than the one used in this example. To establish a reference standard, the potential of the half-cell described by the reaction

$$2\mathrm{H}^+(\mathrm{aq},\ 1m) + 2e^- \rightarrow \mathrm{H}_2(\mathrm{g},\ 1\ \mathrm{atm}) \qquad (100)$$

is *defined* to be zero volts at 25 °C. This voltage scale is entirely arbitrary, but since only relative voltages are important, the absolute voltage of the hydrogen electrode is not important. Relative to this cell, the half-cell potentials (E^0) of the Daniell cell are found to be 0.763 V for the Zn oxidation and 0.337 V for the Cu reduction, when each solution has a concentration of 1 molal (the standard state). The standard half-cell potentials for a large number of reactions have been determined and are tabulated in many reference works. By convention, the cell potential is defined as $\Delta E =$ (cathode potential) − (anode potential), where in each case the reaction is written as a reduction. Thus, in the present example,

$$\mathrm{Zn}^{2+}(\mathrm{aq}) + 2e^- \rightarrow \mathrm{Zn}(\mathrm{s}) \quad E^0 = -0.763\ \mathrm{V}\ (\mathrm{Anode}) \qquad (101)$$

$$\begin{array}{ll} \mathrm{Cu}^{2+}(\mathrm{aq}) + 2e^- \rightarrow \mathrm{Cu}(\mathrm{s}) & E^0 = 0.337\ \mathrm{V}\ (\mathrm{Cathode}) \\ \hline & \Delta E^0 = 1.100\ \mathrm{V}. \end{array} \qquad (102)$$

If the cell voltage is positive, the reaction takes place spontaneously, and electrons flow in an external circuit from the anode to the cathode.

The general expression for the cell potential can be derived from the thermodynamic relation between the Gibbs free energy ΔG and the reversible work done by the cell, dw_{rev}:

$$\Delta G = -dw_{\mathrm{rev}}, \qquad (103)$$

together with

$$\Delta G = \Delta G^0 + RT \ln Q. \qquad (103')$$

In this expression, R is the gas constant (in units of energy/K), T is the absolute temperature, and the quantity Q is given by

$$Q = [C]^c [D]^d / [A]^a [B]^b \qquad (104)$$

for a hypothetical reaction

$$a\mathrm{A} + b\mathrm{B} \rightleftharpoons c\mathrm{C} + d\mathrm{D}. \qquad (105)$$

The quantities in brackets are thermodynamic activities, which become equal to concentrations in infinitely dilute solutions. The reversible work done by the electrochemical cell is related to the cell potential ΔE according to

$$dw_{\mathrm{rev}} = nF\ \Delta E, \qquad (106)$$

where n is the number of electrons transferred in the net cell reaction and F is the charge in coulombs on one mole of electrons, 96 496 [cf. Eq. (124)]. Thus,

$$-nF\ \Delta E = \Delta G^0 + RT \ln Q. \qquad (107)$$

If all reagents are in their standard states, the last term disappears, and $\Delta G^0 = -nF\ \Delta E^0$. Finally,

$$\Delta E = \Delta E^0 - (RT/nF) \ln Q. \qquad (108)$$

This expression, the Nernst equation, is one of the two most basic equations in electrochemistry. At 25 °C, the constant term RT/F has the value 0.025 69 V, thus accounting for Eq. (99). The determination of a cell potential may serve in some cases as a method for determining the equilibrium constant for a reaction.

4.1.1 Batteries. Although many possible half-cell reactions can be combined to give a net reaction with a positive potential, there are obvious restrictions that eliminate most of these from use as practical methods of generating or storing electrical energy (see ENERGY STORAGE). Usually, but not always, the most important requirement is that the cell process be reversible, so that the battery serves as a storage device for electrical energy, and can be easily recharged from an external power source. Mobile electrical energy applications, such as in electric automobiles, place a high priority on the available energy per unit weight. Other factors are economic—the widespread use of platinum electrodes or chemical reactions of rare metals in batteries is hardly practicable.

The lead–sulfuric acid storage battery is very widely used. It has almost complete conversion of chemical to electrical energy and vice versa. The internal resistance is low, enabling it to produce large currents for short periods, and the construction is relatively simple. The chemical reactions of this cell are:

$$PbO_2(s)+HSO_4^-(aq)+3H^+(aq)+2e^- \rightarrow PbSO_4(s)+2H_2O \quad \text{(Cathode)}, \qquad (109)$$

$$Pb(s)+HSO_4^-(aq) \rightarrow PbSO_4(s)+2e^-+H^+(aq) \quad \text{(Anode)}, \qquad (110)$$

$$Pb(s)+PbO_2(s)+2HSO_4^-(aq)+2H^+(aq) \rightarrow 2PbSO_4(s)+2H_2O \quad \text{(Net reaction)}. \qquad (111)$$

The cathode of this cell is made of a lead grid with interstices filled with PbO_2. The anode is a similar grid, with interstices filled with spongy lead to provide a larger surface area. The electrodes are immersed in aqueous sulfuric acid, about 35% H_2SO_4 by weight. The cell potential of this battery, when fully charged, is $\sim$2.0 V, and in typical automotive use, six cells are connected in series to give a total voltage of 12 V. Upon discharge, sulfuric acid is converted to solid lead sulfate and water, decreasing the electrolyte density, which can be monitored as an indication of the state of charge. Typical storage capacities are 90–200 A h. The major disadvantage of the lead storage batttery is its relatively low energy density, about 40 W h/kg.

Another rechargeable battery that has widespread use in small calculators and hand tools depends upon the reactions

$$Cd(s)+2OH^- \rightarrow Cd(OH)_2(s)+2e^- \text{(Anode)}, \qquad (112)$$

$$NiO_2(s)+2H_2O+2e^- \rightarrow Ni(OH)_2(s)+2OH^- \quad \text{(Cathode)}, \qquad (113)$$

$$Cd(s)+NiO_2(s)+2H_2O \rightarrow Cd(OH)_2(s)+Ni(OH)_2(s) \quad \text{(Net reaction)}. \qquad (114)$$

Since the reactants are solids, their concentration does not change with the state of charge, and the cell potential is a constant 1.30 V. The energy density is comparable to that of the lead storage battery, about 30 W h/kg.

The mercury cell is another battery that maintains a constant voltage. The cell reactions, producing a voltage of 1.34 V, are

$$Zn(Hg)+2OH^- \rightarrow ZnO(s)+H_2O+2e^- \quad \text{(Anode)}, \qquad (115)$$

$$HgO(s)+H_2O+2e^- \rightarrow Hg(l)+2OH^- \quad \text{(Cathode)}, \qquad (116)$$

$$Zn(Hg)+HgO(s) \rightarrow ZnO(s)+Hg(l) \quad \text{(Net reaction)}. \qquad (117)$$

The anode consists of a zinc–mercury amalgam. At the cathode, solid mercuric oxide is reduced. The electrolyte is a paste of zinc hydroxide in a strongly alkaline KOH solution. This battery is not rechargeable.

The two batteries just described are "dry cells," which, unlike the lead storage battery, contain no liquid electrolyte. A more ubiquitous dry cell is the LeClanché cell, with a voltage of 1.5 V, provided by the cell reactions

$$Zn(s) \rightarrow Zn^{2+}+2e^- \quad \text{(Anode)}, \qquad (118)$$

$$2MnO_2(s)+2NH_4^++2e^- \rightarrow Mn_2O_3(s)+2NH_3+H_2O \quad \text{(Cathode)}, \qquad (119)$$

$$Zn(s)+2MnO_2(s)+2NH_4^+ \rightarrow Mn_2O_3(s)+Zn^{2+}+2NH_3+H_2O \quad \text{(Net reaction)}. \qquad (120)$$

The anode is a zinc cylinder or cup that serves as a container for the electrolyte, a moist paste of NH_4Cl and $ZnCl_2$ in an inert filler, essentially an aqueous solution in the form of a gel. The cathode is an inert carbon rod in the center of the cylinder, surrounded by a paste of MnO_2 and powdered graphite to provide electrical conductivity. One disadvantage of this cell is that the ammonium ion is a Brönsted acid, and as such it attacks the zinc casing slowly and ultimately may cause leakage.

Recently, interest in electrically powered vehicles has been revived because of environmental concerns: urban air pollution, and the "greenhouse effect" resulting from the combustion of hydrocarbons. As a result, new types of rechargeable batteries are being investigated, with a goal towards achieving higher energy densities than those obtainable with the lead–acid storage battery. For example, the sodium–sulfur battery with a net reaction of $2Na(l)+S(l) \rightarrow Na_2S(s)$ consists of an anode of liquid sodium separated from the cathode of liquid sulfur by a porous β-alumina membrane, acting as a solid electrolyte. This battery generates a potential of 1.8–2.1 V, and has an energy density of 140 W h/kg, considerably better than that of the lead battery. However, it has the significant drawback of requiring an operating temperature of 300–400 °C. Further research in this area will surely lead to new developments.

Another type of battery, not rechargeable and requiring a constant supply of reactants, is the fuel cell. Essentially, this utilizes a combustion (oxidation) reaction as its source of energy, but unlike an electric generator driven by a heat engine, the reaction generates electricity directly. A major advantage of this direct production of electricity over a heat engine-driven generator lies in the relative efficiency. The thermodynamic limit on a heat engine imposed by the second law of thermodynamics leads to a theoretical efficiency of ~40% for engines operating at achievable temperatures, with real-life efficiencies of about 25%. But the fuel cell is *not* a heat engine, and efficiencies of 60%–80% are obtained.

The most highly developed fuel cell, which has been widely used in space vehicles, is the hydrogen–oxygen cell, using the reactions

$$2H_2 \rightarrow 4H^+ + 4e^- \quad \text{(Anode)}, \tag{121}$$

$$4H^+ + O_2 + 4e^- \rightarrow 2H_2O \quad \text{(Cathode)}, \tag{122}$$

$$2H_2 + O_2 \rightarrow 2H_2O \quad \text{(Net reaction)}. \tag{123}$$

The electrodes of this cell include catalytic substances such as finely divided Pt or Ag, and are separated by an ion-exchange membrane such that protons are the current carrier. Fuels other than hydrogen (such as methanol, hydrocarbons, ammonia, and hydrazine) have been used. The oxygen (in terrestrial applications) can simply be obtained from the atmosphere (Bockris and Srinivasan, 1969).

4.2 Electrolysis

The chemical action of the cell in Fig. 15 can be reversed if the ammeter is replaced by an external voltage source with its positive pole connected to the copper electrode. As the external voltage is increased to 1.100 V, the cell current decreases to zero, and a further voltage increase will lead to a reversal in the flow of current. As a result, the direction of the cell reaction will also be reversed and the copper electrode will dissolve, while zinc plates out on the zinc electrode. This is an example of electrolysis, further applications of which are discussed below. In an electrolysis reaction, the relation between the current, the time during which it flows, and the extent of reaction serves to provide a macroscopic experimental determination of the charge on the electron, according to Faraday's laws (1834), which can be written in the form:

$$F = it/n(m/W), \tag{124}$$

where i is the current in amperes, t is the time during which it flows in seconds, m is the weight of material deposited or reacted, and W is its molecular weight. The faraday F is the charge in coulombs on one mole of electrons.

4.2.1 Electroplating. As discussed above in connection with the Cu–Zn or Daniell cell, application of an external voltage greater than the cell potential reverses the spontaneous cell reaction and can lead to oxidation of a metal to its ion at the anode, with migration of the cation through the electrolyte to the cathode, where the ion is reduced and the metal in its elemental form is deposited. A typical example is silver plating, in which the anode is metallic silver (anode reaction, $Ag \rightarrow Ag^+ + e^-$) and the cathode is the object to be plated (cathode reaction, $Ag^+ + e^- \rightarrow Ag$). The electrolyte is a solution of a silver salt, usually silver cyanide, mixed with sodium or potassium cyanide. Similar methods are used for many other metals, from antimony to zinc (Mantell, 1960).

Although given a different name, "electrorefining," the production of pure metals by electrolysis, is another form of electroplating. Copper refining provides a good example. The invention of the dynamo in the latter part of the 19th century created an increasing demand for copper wire, which must be quite pure to have the necessary ductility and electrical conductivity. There is a somewhat symbiotic relationship between the growth in the electric power industry and the electrochemical refining of copper, which uses very large quantities of electric energy. The anodes in this electrorefining process are crude copper, actually 99.0%–99.5% pure copper! A large number of these anodes are interleaved with the cathodes of pure copper (less than 1 mm thick) in an electrolyte of copper sulfate containing other additives. The electrical energy demands are impressive, about 200 kW h/ton of pure copper produced. The raw copper contains, among other impurities, significant amounts of gold and silver which are unoxidized during the electrochemical process and can be regained from the material settling to the bottom of the electrolysis cells. The final stages of purification of these noble metals again rely upon electrorefining.

Electrorefining is also used in the production of nickel, cobalt, lead, antimony, bismuth, indium, mercury, and tin.

"Electrowinning" is a similar process for obtaining pure metals from their ores. The ore is initially leached with a solution that either dissolves or forms a soluble complex with the metal. The resulting solution is purified and modified before it is used as the electrolyte in an electrochemical process that plates out the desired metal. In some applications, the spent electrolyte can be recycled into the leaching step.

4.2.2 Electrolytic Production of Hydrogen and Deuterium. There is a considerable interest in the use of hydrogen as a medium for both storage and transport of energy (Bockris and Nagy, 1974). Over very long distances, the cost per unit of energy is less for the pipeline transport of hydrogen than for electrical transmission by overhead cables. This proposed "hydrogen economy" is usually dependent on the production of hydrogen by the electrolysis of water, using electric power from a nuclear power plant. The advantage of this method over the usual production by chemical reduction of water is that the latter method uses a nonrenewable energy source, such as methane. Aside from these considerations, the only advantage of electrolytic production is in regions where electricity is extremely inexpensive, or in applications where very high purity hydrogen is needed; for example, in the hydrogenation of oils to solid fats, or in the catalytic production of ammonia. The by-product oxygen may also make electrolysis economically feasible.

The electrolyte in this process is usually an alkali hydroxide. Nevertheless, this does not imply that the alkali metal is produced at the cathode, because the potential required to reduce the alkali metal ion is greater than that required to reduce protons or water itself to hydrogen. The relative potentials are 2.714 V for the reduction of Na^+ to Na, and only 0.828 V for the reduction of water to hydrogen and hydroxide ion. The cell reactions are

$$2(2H_2O + 2e^- \rightarrow H_2 + 2OH^-) \quad \text{(Cathode,} \quad -0.828\ \text{V)}, \tag{125}$$

$$2H_2O \rightarrow O_2 + 4e^- + 4H^+ \quad \text{(Anode,} \quad -1.229\ \text{V)}, \tag{126}$$

$$4H^+ + 4OH^- \rightarrow 4H_2O, \tag{127}$$

$$2H_2O \rightarrow 2H_2 + O_2 \quad \text{(Net reaction)}. \tag{128}$$

Thus, the theoretical potential needed to carry out the electrolysis is 0.427 V, but usually an "overvoltage" is required because the electrode processes are not truly reversible. Metal electrodes of steel or nickel are used, separated by a diaphragm of asbestos or perforated metal. It is of some practical interest to note that deuterium, the heavy hydrogen isotope of mass 2, is enriched in the residual water remaining in the electrolysis cell after most of the water has been decomposed. This was noted by Washburn and Urey shortly after the discovery of the deuterium isotope (Urey, Brickwedde, and Murphy, 1932). Soon thereafter, Lewis and MacDonald produced concentrated heavy water (D_2O) by this process. A plant producing heavy water in kilogram amounts operated in Norway before the beginning of World War II. The demand for heavy water as a neutron moderator for nuclear reactors used in the Manhattan Project during World War II led to an investigation of the electrolytic method of isotope separation (Murphy, 1955). A plant in Trail, British Columbia, that produced large amounts of hydrogen for the catalytic synthesis of ammonia was used for this purpose. Actually, the initial enrichment in deuterium to 2%–3% was carried out by means of a chemical exchange, and this was followed by electrolysis to extend the enrichment to 99.8% D. This plant produced about 500 kg of heavy water per month for a period of 2.5 yr, and a total of slightly less than ten tons was produced by the end of World War II. At the present time, electrolysis is generally used only for the final stages of any deuterium enrichment process, because of the high cost of electricity.

4.2.3 Electrolysis in Molten Salts. As pointed out above, many metals cannot be produced from aqueous solutions of their salts, because their reduction potentials are higher (i.e., more negative) than those of hydrogen, i.e., the metals are stronger reducing agents. However, this problem can be overcome by electrolysis in a nonaqueous solvent, or more frequently in a molten salt. Ionic solids are not good electrical conductors because the ions are fixed in a crystal lattice, but in a molten salt they are relatively free to move to the electrode of opposite sign. In fact, the conductivities are considerably higher than those of the same salts in solution.

This property is utilized in the production of metallic sodium and gaseous chlorine from sodium chloride. Molten NaCl at a temperature of about 600 °C is contained in an iron vessel lined with refractory brick. A graphite anode in the center of this cell oxidizes Cl^- to Cl_2, which is collected in a funnel or dome above the anode. Steel cathodes enter through the sides of the vessel, and molten sodium collects in an inverted trough above the molten electrolyte. To prevent the violent recombination of sodium and chlorine, the anode and cathode are separated by an iron screen. The heat required to keep the salt molten is provided by the current flow through the cell.

Aluminum is another metal produced on an even larger scale by electrolysis. This metal is easily oxidized and consequently does not occur in nature as the metal; a major ore is bauxite, hydrated Al_2O_3. This ore is first purified to alumina, with the removal of oxides of iron, silicon, and titanium. The pure alumina is dissolved in a bath of cryolite (Na_3AlF_6), which has a melting point around 1000 °C, considerably lower than that of alumina itself ($\sim$2000 °C). Both the cathode and anode of the electrolysis cell are carbon; however, the anode is gradually consumed by the reaction $C(s)+2O^{2-}\rightarrow CO_2(g)+2e^-$, while Al^{+3} is reduced to the metal at the cathode. The reaction of the anode provides some of the heat required to maintain the high cell temperature. The aluminum produced in this way usually has a purity of only 99.5%–99.8%, and further electrolytic refining is used, with a molten electrolyte consisting of a mixture of fluorides of aluminum, sodium, barium, and calcium in various proportions.

4.3 Corrosion of Metals

The practical aspects of electrochemistry described above have all been applications with desirable economic consequences. Another electrochemical phenomenon, corrosion, has very undesirable economic consequences, leading to losses of billions of dollars per year in the destruction of metal products. The effects of corrosion may be relatively superficial, such as tarnishing, or extremely profound, leading to structural weakening and ultimate failure. Corrosion is the oxidation (in the general sense of the word) of a metal; the oxidizing agent may be a gas (oxygen or air pollutants such as nitrogen and sulfur oxides), water, or an aqueous solution such as seawater. In rare cases, corrosion is actually desirable; as pointed out in Sec. 3.2.1, aluminum forms an oxide film that prevents further reaction. Corrosion is not always an electrochemical process: The example above of aluminum oxide formation, and the formation of other films on metals in contact with a reactive gas or liquid, are more properly described as surface reactions (Sec. 3.2.1).

However, corrosion by water or aqueous solutions definitely is an electrochemical process. For example, in salt water containing oxygen, iron is oxidized according to the cell reactions

$$Fe\rightarrow Fe^{2+}+2e^- \quad \text{(Anode)}, \qquad (129)$$

$$\tfrac{1}{2}O_2+H_2O+2e^-\rightarrow 2OH^- \quad \text{(Cathode)}. \qquad (130)$$

Iron forms a ferrous hydroxide, which reacts further with oxygen to form hydrated ferric oxide. It has been found that the areas where oxide forms tend to act as an anode, while the bare exposed metal is the cathode. Even differences in the metal surface, such as local mechanical stresses, temperature gradients, impurity inclusions, etc., may serve to establish a potential difference leading to electrochemical attack.

The possibility of electrochemical reaction is exacerbated if two dissimilar metals are in electrical contact and are immersed in an electrolyte solution. In fact, the example of the Daniell cell in Sec. 3.1 illustrates this principle. Copper ion is a stronger oxidizing agent than is zinc ion, so the spontaneous reaction will lead to attack on the surface of the zinc electrode. The same principle can be used to protect steel, by connecting it electrically to a more easily oxidized metal, such as magnesium or zinc, which acts as a "sacrificial anode," while the less active steel (i.e., iron) acts as the cathode and is not attacked. This principle is widely used in maritime applications. "Galvanized" steel, made by dipping steel in molten zinc, is protected in the same way. Electroplating may also be used to form the zinc coating.

ACKNOWLEDGMENTS

This research was carried out at Brookhaven National Laboratory under Contract No. DE-AC02-76CH00016 with the U.S. Department of Energy and supported by the Division of Chemical Sciences, Office of Basic Energy Science.

GLOSSARY

***Ab Initio* Method:** Quantum chemical calculations of molecular properties, without using empirical methods.

Activation Energy: An indication of the energy barrier to reaction, which can be determined from the temperature dependence of the rate constant.

Adsorption Isotherm: The relation between the pressure of a gas and the surface coverage of a solid on which it is adsorbed.

Bimolecular Reaction: A reaction involving two molecules interacting during the reaction process.

Brönsted Acid: A compound that readily loses a proton.

Brönsted Base: A compound that readily accepts a proton.

Catalyst: A substance that increases the rate of a chemical reaction, without itself being consumed in the reaction.

Chemical Shift: The change in the energy levels of a nucleus with a magnetic moment in a magnetic field due to the screening effect of orbital electrons. As such, it is sensitive to the chemical environment of the nucleus.

ESCA: Acronym for electron spectroscopy for chemical analysis; see "XPS" below.

EXAFS: Acronym for extended x-ray absorption fine structure. A technique by which the structure of the x-ray absorption spectrum near the absorption edge can be used to determine interatomic distances.

Gibbs Free Energy: A thermodynamic property that indicates the tendency for a chemical reaction to take place.

Isomerization: A reaction that leads to atomic rearrangement within a molecule.

Ligand: An atom or group of atoms that are coordinatively bonded to a central metal atom.

Metathesis Reaction: A chemical reaction involving an atom or radical transfer, of the form: $A + BC \rightarrow AB + C$.

Oxidation Number: An indication of the number of valence electrons in an atom or molecule. In a monatomic ion, it is numerically equal to the ionic charge.

Reaction Rate Constant or Coefficient: A fundamental indicator of chemical reactivity which, combined with concentrations of reactants, allows the rate of reaction to be calculated.

Reaction Mechanism: The combination of elementary reactions that leads to the observed overall reaction.

Redox Reaction: A reaction in which the oxidation numbers of the reactants change.

Unimolecular Reaction: A reaction involving a single, isolated molecule in the rate-determining step.

XPS: Acronym for x-ray photoelectron spectroscopy, also known as ESCA . A method for determining binding energies of electrons (usually core electrons) by measuring the kinetic energy of electrons ejected upon excitation by x-ray irradiation.

List of Works Cited

Bennett, L. E. (1973), *Prog. Inorg. Chem.* **18**, 1–176.

Benson, S. W. (1976), *Thermochemical Kinetics*, 2nd ed. New York: Wiley.

Bernstein, R. B. (1982), *Chemical Dynamics Via Molecular Beam and Laser Techniques*, New York: Oxford.

Berry, R. S., Rice, S. A., Ross, J. (1980), *Physical Chemistry*, New York: Wiley, p. 279.

Bockris, J. O'M., Nagy, Z. (1974), *Electrochemistry for Ecologists*, New York: Plenum.

Bockris, J. O'M., Srinivasan, S. (1969), *Fuel Cells*, New York: McGraw-Hill.

Bond, G. C. (1987), *Heterogeneous Catalysis*, 2nd ed. Oxford: Clarendon.

Brönsted, J. N., Pedersen, K. (1924), *Z. Phys. Chem.* **108**, 185–235.

Brunauer, S., Emmet, P. H., Teller, E. (1938), *J. Am. Chem. Soc.* **60**, 309–319.

Brunschwig, B. S., Creutz, C., Macartney, D. H., Sham, T.-K., Sutin, N. (1982), *Faraday Discuss. Chem. Soc.* **74**, 113–127.

Bull, T. H., Moon, P. B. (1954), *Discuss. Faraday Soc.* **17**, 54–57.

Cicerone, R. J. (1987), *Science* **237**, 35–42.

Fersht, A. (1977), *Enzyme Structure and Mechanism*, San Francisco: Freeman.

Finlayson, B. J., Pitts, J. N., Jr. (1976), *Science* **192**, 111–119.

Gasser, R. P. H. (1985), *An Introduction to Chemisorption and Catalysis by Metals*, Oxford: Clarendon.

Himoe, A., Parks, P. C., Hess, G. P. (1967), *J. Biol. Chem.* **242**, 919–929.

Isied, S. S., Vassilian, A., Magnuson, R. H., Schwarz, H. A. (1985), *J. Am. Chem. Soc.* **107**, 7432–7438.

Isied, S. S., Vassilian, A., Wishart, J. F., Creutz, C., Schwartz, H. A. (1988), *J. Am. Chem. Soc.* **110**, 635–637.

Jasinski, J. M., Meyerson, B. S., Scott, B. A. (1987), *Annu. Rev. Phys. Chem.* **38**, 109–40.

Kendrew, J. C. (1963), *Les Prix Nobel en 1962*, Stockholm: Norstedt, pp. 103–125.

Lovelock, J. E., Maggs, R. J., Wade, R. J. (1973), *Nature* **249**, 194–196.

Mantell, C. L. (1960), *Electrochemical Engineering*, New York: McGraw-Hill.

Marcus, R. A. (1963), *J. Phys. Chem.* **67**, 853–857; *J. Chem. Phys.* **38**, 1858–1862.

Michaelis, L., Menten, M. L. (1913), *Biochem. Z.* **49**, 333–369.

Miller, J. A., Fisk, G. A. (1987), *Chem. Eng. News* **65**, (35), 31 August, 22–46.

Molina, M. J., Rowland, F. S. (1974), *Nature* **249**, 810–812.

Murphy, G. M. (Ed.) (1955), *Production of Heavy Water*, New York: McGraw-Hill.

Okabe, H. (1978), *Photochemistry of Small Molecules*, New York: Wiley.

Perutz, M. F. (1963), *Les Prix Nobel en 1962*, Stockholm: Norstedt, pp. 82–102.

Silverman, J., Dodson, R. W. (1952), *J. Phys. Chem.* **56**, 846–852.

Sinfelt, J. H. (1983), *Bimetallic Catalyst*, New York: Wiley.

Sutin, N. (1983), *Prog. Inorg. Chem.* **30**, 441–498.

Sykes, A. G. (1967), *Adv. Inorg. Chem. Radiochem.* **10**, 153–245.

Taylor, E. H., Datz, S. (1955), *J. Chem. Phys* **23**, 1711–1718.

Vossen, J. L., Kern, W. (Ed.) (1978), *Thin Film Processes*, New York: Academic.

Urey, H. C., Brickwedde, F. G., Murphy, G. M. (1931), *Phys. Rev.* **39**, 164–165.

Wayne, R. P. (1985), *Chemistry of Atmospheres*, New York: Oxford.

Weston, R. E., Jr., Schwarz, H. A. (1972), *Chemical Kinetics*, Englewood Cliffs, NJ: Prentice-Hall.

Yablonovitch, E. (1989), *Science* **246**, 347–351.

Further Reading

Bakhalov, G. T., Turkovskaya, A. V. (trans. Isserlis, G.) (1965), *Corrosion and Protection of Metals*, New York: Pergamon.

Bard, A. J., Faulkner, L. R. (1980), *Electrochemical Methods: Fundamentals and Applications*, New York, Wiley.

Bell, R. P. (1941), *Acid-Base Catalysis*, Oxford: Oxford University Press.

Bell, R. P. (1973), *The Proton in Chemistry*, 2nd ed., Ithaca, NY: Cornell University Press.

Bender, M. L. (1971), *Mechanisms of Homogeneous Catalysis*, New York: Wiley–Interscience.

Boyer, P. D. (Ed.) (1971), *The Enzymes*, Vol. II. New York: Academic.

Chameides, W. L., Davis, D. D. (1982), *Chem. Eng. News* **60**(40), 39–52.

Cheetham, A. K., Day, P. (1987), *Solid State Chemistry: Techniques*, New York: Oxford University Press.

Ember, L. R., Layman, P. L., Lepkowdki, W., Zurer, P. S. (1986). *Chem. Eng. News* **64**(47), 14–64.

Gardiner, W. C., Jr., Olson, D. B. (1980), *Annu. Rev. Phys. Chem.* **31**, 377–399.

Gardiner, W. C., Jr. (Ed).(1984), *Combustion Chemistry*, New York: Springer-Verlag.

Goodman, D. W. (1986), *Annu. Rev. Phys. Chem.* **37**, 425–457.

Graham, A. K. (Ed.)(1971), *Electroplating Engineering Handbook*, 3rd ed., New York: Van Nostrand Reinhold.

Hannay, N. B. (1967), *Solid-State Chemistry*, Englewood Cliffs, NJ: Prentice-Hall.

Heicklen, J. (1976), *Atmospheric Chemistry*, New York: Academic.

Johnston, H. S. (1966), *Gas Phase Reaction Rate Theory*, New York: Ronald.

Levine, R. D., Bernstein, R. B. (1987), *Molecular Reaction Dynamics and Chemical Reactivity*, New York: Oxford University Press.

Lin, M. C. Ertl, G. (1986), *Annu. Rev. Phys. Chem.* **37**, 587–615.

Mantell, C. L. (1983), *Batteries and Energy Systems*, 2nd ed., New York: McGraw-Hill.

Moore, J. W., Pearson, R. G. (1981), *Kinetics and Mechanism*, 3rd ed. New York: Wiley.

Nachtrieb, N. H. (1979), *Annu. Rev. Phys. Chem.* **31**, 131–156.

Newman, J. S. (1973), *Electrochemical Systems*, Englewood Cliffs, NJ: Prentice-Hall.

Newton, M. D., Sutin, N. (1984), *Annu. Rev.*

Phys. Chem. **35**, 437–480.

Robinson, P. J., Holbrook, K. A. (1972), *Unimolecular Reactions*, New York: Wiley–Interscience.

Seymour, R. B., Carraher, C. E., Jr. (1988), *Polymer Chemistry: An Introduction*, New York: Marcel Dekker.

Taube, H. (1970), *Electron Transfer Reactions of Complex Ions in Solution*, New York: Academic.

CHEMILUMINESCENCE

TUAN VO-DINH, *Advanced Monitoring Development Group, Health and Safety Research Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A.*

Introduction 413
1. **Basic Photophysical and Photochemical Principles** 414
1.1 Background 414
1.1.1 Electronic Transitions and Luminophores 414
1.1.2 Chemiluminescence 414
1.2 Chemiluminescent Systems in the Gas Phase 415
1.2.1 Hydrocarbon Reactions With Active Nitrogen 415
1.2.2 Chemiluminescence Reactions With Ozone 416
1.2.3 Other Gas-Phase CL Reactions. 417
1.3 Chemiluminescent Systems in Liquid Solution 417
1.3.1 Luminol 417
1.3.2 Peroxyoxalates 417
1.3.3 Lucigenin 418
1.3.4 Singlet-Oxygen Sensitization Systems 418
1.3.5 Tris(2,2′-bipyridine) ruthenium (II) 419
2. **Instrumentation** 419
2.1 Photometric Detection Systems 419
2.2 CL-Based Detectors For Chromatography and Flow-Injection Systems 421
3. **Applications** 421
3.1 Monitoring Atmospheric Pollutants 422
3.2 Measurement of Trace Metals. 422
3.3 Determination of Alcohols, Ethers, Aldehydes, and Saccharides 423
3.4 Determination of Polycyclic Aromatic Hydrocarbons 423
3.5 Analysis of Proteins, Amino Acids, and Other Biological Systems 423
3.6 Analysis of Drugs and Food Components 424
3.7 Determination of Sulfides 425
3.8 Chemiluminescence Immunoassays 425
3.9 Nucleic Acid Detection, Hybridization, and Sequencing 425
Acknowledgments 427
Acronyms 427
Glossary 427
Works Cited 427
Further Reading 428

INTRODUCTION

Chemiluminescence (CL) refers to emission from excited molecules generated as products of a chemical reaction. If this reaction is derived from biological organisms such as bacteria, fungi, and insects, the luminescence is called bioluminescence. Although CL phenomena were reported more than a half century ago (Albrecht, 1928), only in the last ten years have there been extensive basic studies as well as practical applications. The dynamic field of CL is experiencing considerable growth due to the advent of new instrumentation, sensitive detection methods, and efficient automation technology. The use of CL is also receiving increased interest due to its simplicity, sensitivity, and dynamic range. This article provides an overview of the basic principles of CL, the instrumentation commonly used for detection, and several applications of this unique molecular process in a wide variety of fields ranging from environmental monitoring to protein analysis, food product screening, immunological assays, and DNA sequencing.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. BASIC PHOTOPHYSICAL AND PHOTOCHEMICAL PRINCIPLES

1.1 Background

1.1.1 Electronic Transitions and Luminophores. Luminescence is an emission process that involves transitions between electronic states of an atom or a molecule. The technique deals with optical detection and spectral analysis of light emitted by a substance, often referred to as a luminophore, which undergoes a radiative transition from an excited electronic state to a lower electronic state. Most organic molecules contain an even number of electrons. In the ground state, the electrons fill the various atomic orbitals with the lowest energies in pairs. According to the Pauli exclusion principle, two electrons in one given orbital must have spins in opposite directions and the total spin is equal to 0. For this reason, the ground state has no net electron spin. Such a state is called a singlet state. Excitation of a molecule usually results in the promotion of one electron from the highest occupied orbital to a previously unoccupied orbital having a higher energy. Figure 1 shows a diagram, known as the Jablonskii diagram, illustrating the photophysical processes that ensue from excitation of a molecule into a singlet (spin equal to 0) excited state S_n. At equilibrium, molecules have the greatest probability of having the thermal energy of the lowest vibrational and rotational levels of the ground state S_0 in accordance with the Boltzmann distribution. In condensed media, a molecule in the S_n state deactivates rapidly to lower vibronic levels of S_n within 10^{-13} to 10^{-11} s via vibrational relaxation (VR) processes and acquires the thermal energy distribution of the lowest vibrational levels of S_n. This VR process is faster than electronic transitions. The VR excess energy is released as thermal energy to the surrounding medium. From the S_n state, the molecule deactivates rapidly to the isoenergetic vibrational level of a lower electronic state such as S_{n-1} via a process known as internal conversion (IC). By a succession of VR and IC processes, the molecule deactivates rapidly to the lowest excited singlet state S_1. From S_1 the molecule may

1. further deactivate to the ground state S_0 in a radiationless fashion (i.e., no photon emitted) via IC and VR processes, or
2. emit a photon.

This emission process is known as fluorescence.

From S_1, the molecule may also undergo a transition to some vibrational level of one of the lower triplet states (states with spin equal to 1) via a mechanism known as intersystem crossing (ISC). The molecule undergoes deactivation to the lowest excited triplet state T_1 via VR and IC processes. From T_1, the molecule may return to the ground state S_0 in a radiationless fashion via an ISC process or by emitting a photon. This latter process is known as phosphorescence. Both fluorescence and phosphorescence processes are referred to as luminescence (Harmony, 1972; Vo-Dinh, 1984).

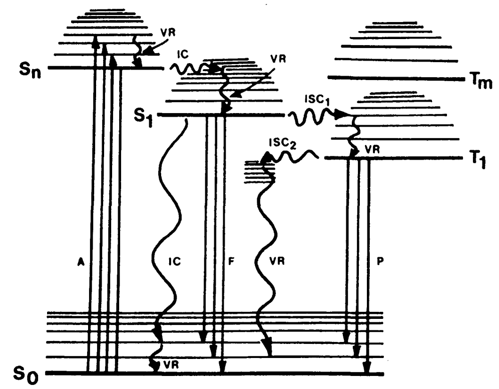

FIG. 1. Jablonskii diagram for molecular photophysical processes. S_0, singlet ground state; S_1, first excited singlet state; S_n, *n*th excited singlet state; T_1, first excited triplet state; T_m, *m*th excited triplet state; *A*, absorption; *IC*, internal conversion; *F*, fluorescence; *VR*, vibrational relaxation; *P*, phosphorescence; *ISC*, intersystem crossing.

The nature of the orbitals involved in an electronic transition is an important factor in determining the luminescent properties of the molecule. Organic compounds that are known to be strongly luminescing are molecules with extensive π-electron systems. It is noteworthy that, in the gas phase, fluorescence may be observed from higher vibrational levels of the excited state. These emission mechanisms, known as single-vibronic-level (SVL) luminescence processes, do not occur in solution.

1.1.2 Chemiluminescence. Chemiluminescence emission is usually a fluorescence process (from the excited singlet state) rather than a phosphorescence process (from an excited triplet state). This is because the triplet state has a relatively long lifetime and is subject to radiationless deactivation by other molecules present such as oxygen. Chemiluminescence in solution generally involves an oxidative reaction. Molecular oxygen and hydrogen peroxide are most often involved.

The production of chemiluminescence usually requires three conditions:

1. adequate chemical energy to populate an excited state of a molecule,
2. an efficient chemical reaction with high yield, and
3. deactivation of the excited state via radiative CL emission.

In some cases the excited molecule does not emit radiation but transfers the energy to another molecule that subsequently emits CL. The efficiency of CL emission may be determined by the number of photons produced per molecule undergoing a chemical reaction, defined as follows:

$$M \xrightarrow{\Omega_E \Omega_C} M^*, \tag{1}$$

$$M^* \xrightarrow{\Omega_L} M + h\nu(\text{CL photon}), \tag{2}$$

$$\Omega_{\text{CL}} = \Omega_E \times \Omega_C \times \Omega_L, \tag{3}$$

where M denotes a molecule in the ground state, and M^*, a molecule in the excited state.

The excitation yield Ω_E, which denotes the proportion of molecules promoted to the excited state, depends on the chemical reaction mechanism. A high CL yield Ω_{CL} depends on a high chemical yield Ω_C and an efficient luminescence yield Ω_L. The values of Ω_{CL} vary from 10^{-8} up to 0.88, the latter for the *in vitro* firefly reaction, which has one of the most efficient CL yields (Seliger and McElroy, 1960).

1.2 Chemiluminescent Systems in the Gas Phase

When a reagent gas is mixed with another gas in a reaction cell, and if the reaction is sufficiently exothermic, one or more of the products may be promoted to an excited electronic state, which can subsequently decay to the ground state with emission of CL. By use of different reagent gases, a wide variety of CL reactions may be induced. Chemiluminescence reagents are not limited to stable molecules. Highly exothermic reactions of atoms, radicals, ions, and electronically excited species can be produced in gas discharges and detected by photometric detectors.

1.2.1 Hydrocarbon Reactions With Active Nitrogen. An active-nitrogen detector has been developed by Sutton *et al.* for monitoring hydrocarbons (Sutton *et al.*, 1979). High-energy electrical discharges (e.g., electrical or microwave discharges) can be used to produce a mixture of nitrogen atoms and highly excited nitrogen molecules $N_2(A\,^3\Sigma_u^+)$. These excited-state nitrogen molecules (active nitrogen) can induce collisional energy transfer to analyte molecules present, providing a pathway to electronic and vibrational excitation of the analytes. For example, active nitrogen reacts with hydrocarbons to produce an excited state of the cyanogen radical (CN*), which emits intense CL emission corresponding to the following transition:

$$B\,^2\Sigma^+ \rightarrow X\,^2\Sigma^+. \tag{4}$$

Since carbon compounds are found to react with active nitrogen to produce excited states of cyanogen radicals, the active-nitrogen detector can be used as a universal monitor for hydrocarbons. Possible reactions capable of producing excited-state cyanogen are

$$CH_2 + N \rightarrow CN^* + 2H, \tag{5}$$

$$CH + N \rightarrow CN^* + H. \tag{6}$$

Figure 2 shows the CL spectrum of CN* obtained when hydrocarbons were added to the active-nitrogen monitor (Sutton *et al.*, 1979). The most intense band of this emission is the violet-blue emission at approximately 385 nm. This system is often used as the universal response to hydrocarbons, and is based on the fact that virtually all carbon compounds react with active nitrogen to produce an excited state of cyanogen radicals. In general, CL emission from saturated hydrocarbons is much less intense than that from the corresponding alkene reactions. Addition of HCl has been found to enhance catalytically the reactions of saturated hydrocarbons. For example, the detection limit of methane was 4 and 200 ng with and without HCI catalysis, respectively (Sutton *et al.*, 1979; Birks, 1989).

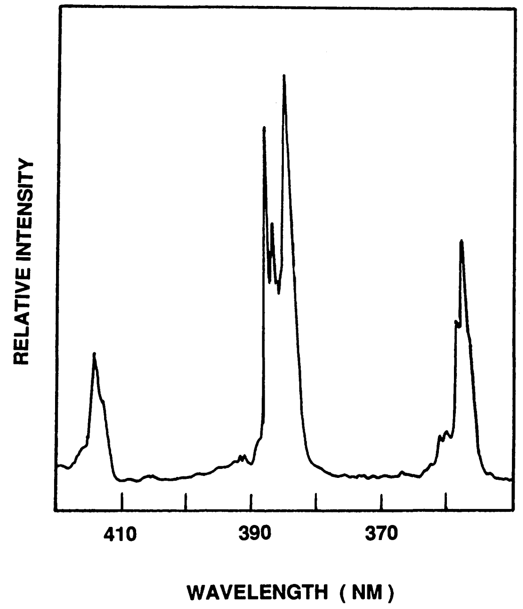

FIG. 2. Chemiluminescence spectrum of CN* for hydrocarbons added to active nitrogen 1979 (from Sutton *et al.*, 1979).

1.2.2 Chemiluminescent Reactions With Ozone.

Ozone (O_3) is one of the most widely used CL reagents since most gas-phase reactions with ozone have been found to be chemiluminescent. The basis of ozone CL reactions is attributed to the weak O–O bond (24.9 kcal/mol), which is broken to form the strong O=O bond (119 kcal/mol) in molecular oxygen and strong C=O bonds (172 kcal/mol) in reactions with many organic compounds (Birks, 1989).

Ozone reacts rapidly with olefins, producing CL emission from various species. Figure 3 shows the CL emission spectrum for the reaction of O_3 (2% in O_2) with ethylene (Finlayson *et al.*, 1974).

Various nitrogen-containing compounds, oxygenated compounds, and reduced sulfur compounds can be detected by use of ozone CL reactions. The basic reaction involved in the detection of nitrogen and oxygenated compounds is

$$NO + O_3 \rightarrow NO_2^* + O_2. \tag{7}$$

Sulfur compounds and their interactions with ozone have also been extensively investigated. Reduced sulfur compounds usually become oxidized to produce electronically excited SO_2^* ($^3B_1$), which exhibits CL emission between 300 and 400 nm (Birks, 1989). Figure 4 shows the CL spectrum of SO_2^* ($^3B_1$) produced by the reaction of O_3 with CH_3SH (Glinski *et al.*, 1981).

The gas-phase reaction between ozone and simple sulfides, e.g., H_2S, CH_3SH, and CH_3SCH_3, has been found to produce strong CL emission under flow-tube conditions (Pitts *et al.*, 1972). The results indicated that SO_2^* was the chemiluminescent species produced by the reaction

$$SO + O_3 \rightarrow SO_2^* + O_2, \tag{8}$$

where SO is produced in a previous step. Glinski *et al.* (1981) reported that, at low pressures, the excited singlet state is quenched to the triplet manifold to yield phosphorescence.

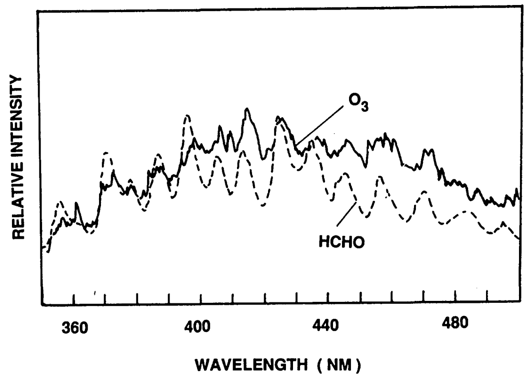

FIG. 3. Chemiluminescence spectra for the reaction of O_3 (2% in O_2) with ethylene (from Finlayson *et al.*, 1974).

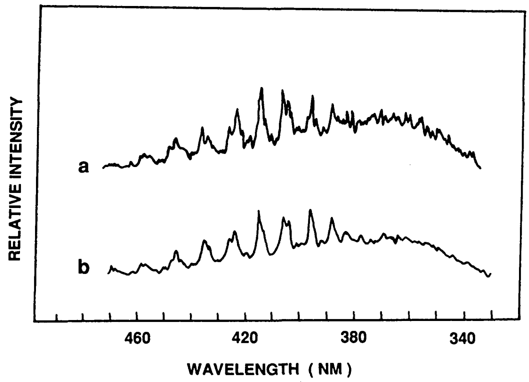

FIG. 4. Chemiluminescence of SO_2 from the reaction of O_3 with CH_3SH: *a*, 20 mTorr O_3; 7 mTorr CH_3SH; *b*, 30 mTorr O_3, 15 mTorr CH_3SH; with 520 mTorr of He in both cases (from Glinski *et al.*, 1981).

1.2.3 Other Gas-Phase CL Reactions. Fluorine molecules, F_2, react with reduced sulfur compounds to produce several chemiluminescent species, providing an important detection method highly selective to organosulfur compounds (Mishalanie and Birks, 1986). When fluorine atoms are used in place of fluorine molecules, the reactions occur at a much faster rate, and the chemical processes are similar to those occurring with active nitrogen. Fluorine atoms can be produced via various methods such as microwave discharges of F_2, CF_4, and SF_6 (Birks, 1989). Other compounds that have been used as gas-phase CL reagents include sodium atoms (Yamada *et al.*, 1982) and chlorine dioxide (Spurlin and Yeung, 1982).

1.3 Chemiluminescent Systems in Liquid Solution

1.3.1 Luminol. Luminol or 5-amino-2,3-dihydrophthalazine-1,4-dione is one of the most investigated CL systems. The oxidation process of luminol has been extensively investigated although the mechanisms have not yet been completely elucidated. In aqueous alkaline solution, luminol is oxidized to produce an excited-state species, viz. 3-aminophthalate, which exhibits CL emission centered at approximately 425 nm (Seliger and Lee, 1972). This reaction is depicted in Fig. 5. Various oxidants (e.g., permanganate, hypochlorite, iodine, and perhaps the most common oxidant, hydrogen peroxide) can be used. When peroxide is used as the oxidant, a catalyst (e.g., peroxidase, hemin, ferricyanide, and transition-metal ions, such as Co^{2+}, Cu^{2+}, Fe^{3+}, etc.) is required. The CL luminol reaction may be summarized as follows:

$$\text{Luminol} + H_2O_2 + \text{Catalyst} + OH^- \rightarrow \text{Product} + \text{Light}. \qquad (9)$$

The above equation indicates that the CL emission depends upon the concentrations of luminol, H_2O_2, and the catalyst. Since the measurement of the CL signal intensity can be used to quantitate any of these species, the luminol CL method can be used to determine the following species:

1. luminol or species labeled with luminol,
2. peroxide or species that can be converted to peroxide, or
3. catalysts (e.g., Cu^{2+}, Co^{2+}, Ni^{2+}, Fe^{3+}, etc.) or species labeled with catalysts.

The wavelength maximum of the luminol CL emission changes with the solvent used. The CL quantum yield has been found to vary also with solvent polarity and, in water, with *p*H, with the maximum CL intensity occurring between *p*H values 10 and 11 (Seliger and Lee, 1972). Although the CL quantum yield of luminol, which is only about 1%, is low in comparison to typical luminescence quantum yields of strong fluorophors, it is one of the more efficient nonenzymatically coupled CL systems. To date, extensive studies have been devoted to investigation of properties of luminol and many of its analogs (Gundermann, 1968; Roswell and White, 1978). The reaction rates of some intermediates of the luminol system have also been investigated by flash photolysis spectroscopy (Meneryi and Lind, 1980). Other luminol derivatives that are commonly used as CL systems include isoluminol and aminobutylethylisoluminol (ABEI). Coupling of these compounds to an analyte is usually performed via the primary amine group. Although the quantum efficiency of luminol is greater than that of isoluminol, substitution at the primary amine group on luminol often decreases the quantum efficiency. On the other hand, substitution at the primary amine group on isoluminol increases the quantum efficiency.

1.3.2 Peroxyoxalates. Some commonly used CL reactions are peroxyoxalate reactions which involve the hydrogen peroxide oxidations of aryl oxalate esters. The quantum efficiencies of these reactions are very high (usually 20%–30%), with the production of a

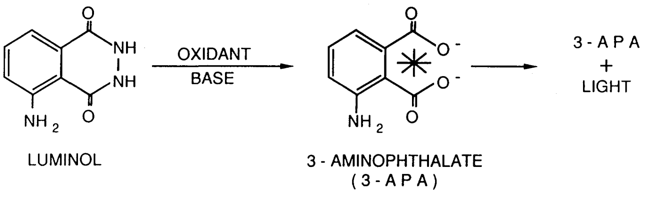

FIG. 5. Chemiluminescence reaction for luminol (asterisk indicates luminescent excited state).

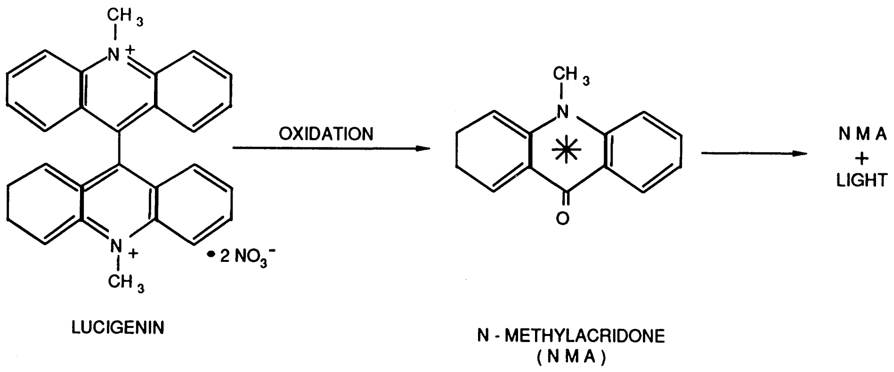

FIG. 6. Chemiluminescence reaction for lucigenin (asterisk indicates luminescent excited state).

high-energy intermediate. This intermediate and a fluorophore (fluorescent product) then form a charge-transfer complex which dissociates to produce an excited-state fluorophore. The fluorophore subsequently emits CL. Peroxyoxalate CL has useful applications in the quantification of hydrogen peroxide or fluorophores such as polycyclic aromatic hydrocarbons. Two of the most commonly used oxalate esters are bis-(2,4,6-trichlorophenyl)oxalate (TCPO) or bis-(2,4-dinitrophenyl)oxalate (DNPO). Peroxyoxalate CL can be used over a wide *p*H range. TCPO can be used from *p*H 5 to 9, with the maximum CL intensity occurring at approximately *p*H 7. Measurements at more acidic *p*H can use DNPO. A weak base such as triethylamine, tris(hydroxymethyl)aminomethane, or imidazole is usually included in the reaction mixture in order to catalyze the reaction and increase CL intensity (Birks, 1989).

1.3.3 Lucigenin. Lucigenin (N,N'-dimethyl-9,9'-diacridinium nitrate) is another commonly used system that undergoes CL reaction in alkaline solution to produce N-methylacridone with emission of CL light. Figure 6 shows a schematic diagram for the lucigenin CL reaction. The CL emission originates from an excited singlet state of N-methylacridone, which exhibits a CL spectrum from 420 to 500 nm with a maximum at about 440 nm (Totter, 1975). The chemical reactions are complicated and probably involve both oxidation and reduction steps, although the overall reaction is an oxidation. Lucigenin can react either with hydrogen peroxide or with a reductant plus oxygen. The presence of catalysts like Co(II), Fe(III), Fe(II), Cu(II), Cr(III), and Ni(II) can considerably increase the CL emission.

Acridinium esters and acridans (Fig. 7) react similarly to lucigenin to also yield N-methylacridone- and CL light.

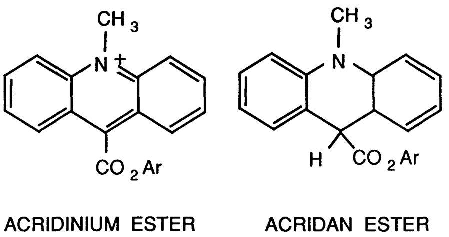

FIG. 7. Structure of acridinium and acridan phenyl esters. Ar, aromatic system.

1.3.4 Singlet-Oxygen Sensitization Systems. Various CL reactions involve singlet oxygen which can be formed by the folowing reactions:

$$S + h\nu \rightarrow {}^1S^* \rightarrow {}^3S^*, \tag{10}$$

$${}^3S^* + {}^3O_2 \rightarrow S + {}^1O_2, \tag{11}$$

where S denotes the ground state of a sensitizer for singlet oxygen; ${}^1S^*$, an excited singlet state of the sensitizer; ${}^3S^*$, an excited triplet state of the sensitizer; ${}^1O_2$, the singlet state of molecular oxygen; and ${}^3O_2$, the triplet state of molecular oxygen.

In the presence of a singlet-oxygen acceptor, A, the following reaction, which involves direct energy transfer from singlet-oxygen dimers to a dye capable of fluorescence, may occur:

$$2\ {}^1O_2 + A \rightarrow 2\ {}^3O_2 + {}^1A^* \rightarrow A + h\nu, \tag{12}$$

where A denotes the state of acceptor A, and ${}^1A^*$, an excited singlet state of A. The formation of singlet-oxygen dimers in Eq. (12) is well studied and many investigators have reported CL by this process. Acceptors have included violanthrene, methylene blue, and chlorophyll (Birks, 1989). Another reaction that can produce CL involves formation of a fluorescent product:

$$^1O_2 + A \rightarrow (AO_2) \rightarrow \text{Product}^* \rightarrow \text{Product} + h\nu. \tag{13}$$

In this reaction the acceptors may form dioxetane intermediates that can decompose to yield CL. Although most dioxetanes decompose in solution to yield predominantly excited triplet states producing little CL, there are acceptors that form dioxetanes exhibiting good excited singlet yields. These include lophine (2,4,5-triphenylimidazole) and some of its derivatives (White and Harding, 1964).

1.3.5 Tris(2,2′-bipyridine)ruthenium(II). Another CL system is noteworthy because the reagent is regenerated and thus can be recycled. This system, ruthenium tris-bipyridine(bpy), involves reaction of $Ru(bpy)_3^{3+}$ and $Ru(bpy)_3^{+}$ to produce an excited state of $Ru(bpy)_3^{2+}$, a stable species which decays to the ground state by emitting an orange emission. $Ru(bpy)_3^{3+}$ and $Ru(bpy)_3^{+}$ can be electrogenerated from $Ru(bpy)_3^{2+}$ by reduction at approximately -1.3 V, and oxidation at approximately $+1.3$ V.

2. INSTRUMENTATION

2.1 Photometric Detection Systems

Chemiluminescence can be very sensitive because it involves an emission technique. The basis of this sensitivity in emission spectrometry is well known. Absorption spectrometry involves the detection of small changes in a large signal (light beam passing through the sample into the detector) that is always present and produces an intrinsic photon noise which determines the detection limit. On the other hand, emission spectrometry is not limited by the high level of the excitation radiation since most of this radiation is eliminated by excitation monochromators or optical filters, and appropriate excitation/detection geometry. In CL, there is no excitation light source, and thus no scattered excitation light problem. The background noise of the sample is the limiting factor.

The instrumentation commonly used to detect CL is relatively simple and is essentially similar to that used for luminescence. Most of the discussion on detection instrumentation in this section also applies to conventional luminescence.

A typical CL detection system may have the following basic components:

(a) a dispersive element, such as an optical filter, a monochromator, or a polychromator; and
(b) an appropriate detector.

No excitation light source is required since the chemical reactions provide the necessary excitation energy. Because the luminescence signals of commonly used CL systems usually have broadband features, the monochromators and related instrumentation for CL do not need to be sophisticated and costly. The CL emission from the sample is collected through appropriate optics and focused onto the entrance slit of a dispersive element. There are two basic classes of luminescence spectrometers: single-channel instruments and multichannel instruments. The first type of instrument generally uses a monochromator coupled to a photomultiplier, whereas the latter uses a polychromator interfaced to a multichannel detector, such as a vidicon, a photodiode array, or a charge-coupled device (CCD).

A wide variety of manufacturers of luminescence spectrometers are available, each offering a variety of models with different performance characteristics and each offering different options. Basic instrument components are also commercially available and easily assembled. An investigator may assemble off-the-shelf components for his or her particular applications.

The most commonly used luminescence spectrometer is a photomultiplier coupled with a double-grating monochromator. The spectral resolution and the sensitivity of the spectrometer are strongly dependent on the characteristics of the monochromator and the photomultiplier tube. Modern holographic gratings generally offer excellent stray-light rejection. The single-photon counting method can be used for detection. Unlike the more conventional analog detection method, the single-photon counting signal output is digital in nature, producing discrete pulses of charge. The digital technique has proven to have several advantages over the analog method, especially for low-level signal detection (Malmstadt *et al.*, 1972; Vo-Dinh and Wild, 1973, 1974).

Photomultipliers are widely used as CL detectors for their high spectral sensitivity, wide operating range, low cost, and relatively simple electronics. Whenever low-intensity signals are observed in CL measurements, it is desirable to improve the sensitivity of detection by recording the entire spectrum simultaneously. Interferometric methods, e.g., Fourier-transform techniques, can enhance the detection sensitivity only if the noise of the measured signal is detector limited.

With monochromators, only one spectral resolution element, or channel, can be monitored at a time. Detectors that permit the recording of the entire spectrum simultaneously, thus providing the multiplex advantage, are known as multichannel detectors. A well-known example of a multichannel detector is the photographic plate. More modern multichannel detectors include the vidicon, the diode array, and the charge-coupled device (Talmi, 1975; Campion and Woodruff, 1987). In multichannel spectrometers, the detector is placed at the focal plane of a polychromator, which is a monochromator with the exit slit removed. As a result, the entire emission, dispersed at all wavelengths within the polychromator, is detected simultaneously. The simultaneous detection of all the dispersed emission using n spectral resolution elements reduces the measurement time by a factor of n in the case of an S/N (signal-to-noise)–limited measurement, or improves the S/N ratio by a factor of n in the case of a time-limited measurement.

There are several types of multichannel detectors commercially available. One device used in the 1970s is the vidicon, which is essentially a television-type device comprised of an array of microscopic photosensitive diode junctions that are grown upon a single silicon crystal wafer. These diode junctions provide individual microelements that are used to detect simultaneously the radiation dispersed by the polychromator. Prior to detection, each diode is charged to a preset reversed-bias potential by a fast-scanning electron beam. The electromagnetic radiation that strikes the diode causes depletion of the charge in each diode. A current proportional to the incident radiation recharges the diode to the preset potential when the electron beam scans again, thus providing information on the intensity of the incident light that has struck the diode. Consequently, an almost instantaneous recording of the spectral intensity at each diode is generated in the computer memory of the vidicon detector. Vidicons are capable of integrating radiation intensity over multiple scanning cycles through their charge-storage capabilities. Sensitivity can be further enhanced by incorporating an image-intensification section in front of the vidicon to yield a silicon-intensified target (SIT) vidicon.

Another multichannel detector is the diode array, which also utilizes photodiodes as detection elements. Recording of the signal is performed with direct on-chip circuitry rather than with a scanning electron beam. Signal amplification is achieved by a microchannel-plate image intensifier. Gated detection down to 5-ns time resolution can be performed with the intensified diode array. Intensified diode arrays are currently the most widely used multichannel detectors.

Recently, other types of devices, such as charge-coupled devices (CCDs), have received increasing use because of their high quantum yields, two-dimensional imaging capabilities, and very low dark currents (see CHARGE-COUPLED DEVICES). Charged-coupled devices have several advantages, including

1. high quantum yields and wide spectral response compared to those for photomultipliers,
2. low dark current when cooled to less than 140 K (permitting integration of the luminescence emission over extended periods of time), and
3. immunity from light shock or intense bursts of radiation.

This immunity is an important feature in CL because of the wide range in reaction rates (over 15 orders of magnitude), resulting in a very wide range of photon fluxes. An important advantage of CCDs is that the read noise is much lower than that of photodiode arrays. With CCDs a process involving on-chip charge combining, often referred to as "binning," can reduce the effect of read noise. The advent of CCD detectors allows the investigation and applications of weakly CL systems that could not be detected otherwise. Jalkian and Denton (1988) have developed CCD instruments for wavelength-dispersive and nondispersive CL measurements. Figure 8 shows a wavelength-nondispersive luminometer employing a (512 ×320)-element CCD detector (Jalkien and Denton, 1988). The wavelength-dispersive instrument was used to record the spectrum of luminol, whereas the nondispersive instrument, which is simple in design and offers improved sensitivity, was mainly used to perform quantitative measurements.

2.2 CL-Based Detectors for Chromatography and Flow-Injection Systems

Chemiluminescent reactions in the gas phase and in solution have been successfully used to develop an extremely wide variety of detectors in gas and liquid chromatographs. In general, CL-based detectors are very sensitive and selective to the specific species that undergo CL reactions. For example, the measurement of NO by its CL reaction with O_3 provides the basis for gas chromatography (GC) detectors. The limit of detection of NO is in the part-per-trillion (ppt) range. The dynamic range of the analytical curve for the CL method is over 6 orders of magnitude. Figure 9 shows a schematic diagram of a CL-based detection system for GC. In general, GC detectors include active-nitrogen detectors, ozone-induced detectors, sulfur-based detectors, fluorine-based detectors, and chlorine-dioxide-based detectors (Birks, 1989).

A variety of CL reaction schemes in solution have been used to develop detectors for liquid chromatography (LC). The CL systems include luminol, lucigenin, peroxyoxalates, Ru(bpy), electrochemiluminescence, and the luciferin/luciferase (bioluminescence) reactions. Another example described in a previous section involves the use of CL reactions of hydrogen peroxide for liquid chromatography detection. For details, the reader is referred to a recent book dealing specifically with chemiluminescence and photochemical reaction detection in chromatography (Birks, 1989).

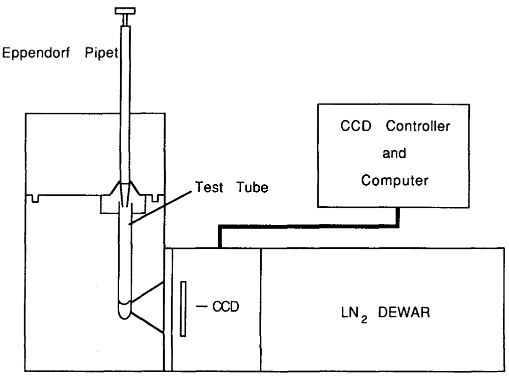

FIG. 8. Wavelength-nondispersive luminometer (from Jalkian and Denton, 1988).

A variety of CL-based detection systems can also be operated as post-column reactions for flow injection as well as chromatographic systems (Birks, 1989) (see CHROMATOGRAPHY). The instrumental device involves a method to deliver and mix the CL reagents with the column effluent and an appropriate detection flow cell that permits monitoring the CL signal. A typical instrumental setup for CL application in flow injection and chromatography is shown in Fig. 10.

Because of the reactions kinetics involved in the CL process, the CL emission intensity is usually a transient signal rather than a constant signal. The CL emission intensity I_{CL} (photons/s) is directly proportional to the rate of chemical reaction dM/dt [(molecules reacting)/s]:

$$I_{CL} = \Omega_{CL}(dM/dt). \qquad (14)$$

Since the rate of the CL reaction varies with time due to consumption of reagents, the CL intensity also varies with time. In order to optimize the sensitivity of detection, one needs to measure the CL signal at its maximum level which occurs after a specific time delay following reagent mixing. In general, the CL signal reaches the maximum level in a second or less after reagent mixing and decays to the background level in a few seconds. The nature of the solvents used in flow-injection analysis (FIA) or LC is an important parameter affecting the CL intensity. Table 1 illustrates the effect of various solvents on the CL intensity of luminol (Birks, 1989).

3. APPLICATIONS

The use of CL is gaining increasing popularity due to its simplicity and sensitivity. The description of various CL applications in this section is not aimed to give an extensive list of all possible applications but rather to provide an overview of selected examples of CL studies in order to illustrate their usefulness in a wide variety of fields.

3.1 Monitoring Atmospheric Pollutants

There is an extremely wide variety of analytical instruments for monitoring gaseous species using CL reactions. These devices are based on various schemes (e.g., $NO+O_3$, CH_2+N, $CH+N$, etc.). Van Zoonen *et al.* have developed a solid-state CL detector for hydrogen peroxide for potential application to acid rain analysis (Van Zoonen *et al.*, 1985). The acidity of rain and cloudwater is affected by the levels of sulfur oxides and nitrogen oxides released into the atmosphere and subsequently dissolved in the water. One of the main reactants in the decomposition cycle of sulfur oxides is believed to be hydrogen peroxide. An important problem related to monitoring hydrogen peroxide is the low concentrations of these species, which are unstable in aqueous samples. The normal hydrogen peroxide concentrations in rain water may vary from less than 1 μg/L (3×10^{-8} *M*) in polluted areas to more than 1 mg/L (3×10^{-5} *M*) in clean areas. The CL monitor used peroxylate CL and consisted of a two-bed reactor packed with solid trichlorophenyloxalate (TCPO) and 3-aminofluoranthene immobilized on controlled-pore glass beads. The detection limit was 1.5×10^{-8} *M* and the dynamic range of the analytical curve was 6 orders of magnitude. The quenching of peroxyoxalate CL was used as a new detection principle in FIA and high-performance liquid chromatography (HPLC) (Van Zoonen *et al.*, 1986). Figure 11 shows a block diagram of the CL-based detector for FIA and HPLC.

3.2 Measurement of Trace Metals

The CL detection method in liquid chromatography has used the luminol reaction for trace metals separated by ion-exchange chromatography (Birks, 1989). Mixtures of Cu^{2+} and Co^{2+} and of Cu^{2+}, Co^{2+}, and Ni^{2+} were separated on an anion-exchange column as their chloride complexes. Manganese(II) has been determined at part-per-billion (ppb) levels by detection of the CL emission produced

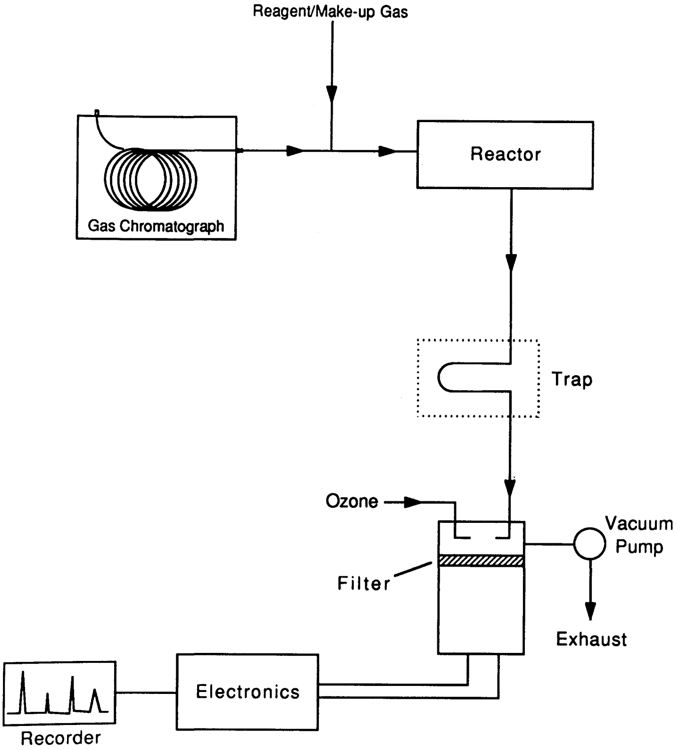

FIG. 9. Schematic diagram of a CL-based detector using the $NO+O_3$ reaction for gas chromatography (from Birks, 1989). A cold trap is used to trap nonvolatile species before they pass into the CL reaction chamber where the effluent is mixed with an excess of O_3. Light produced from the $NO+O_3$ reaction is detected by a photomultiplier tube.

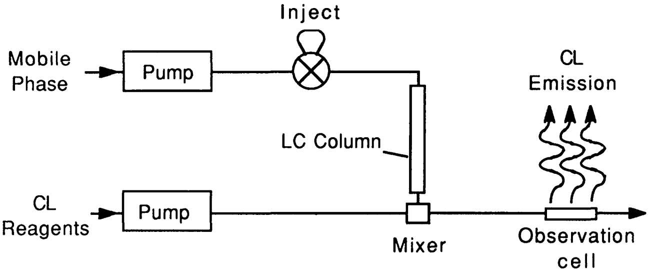

FIG. 10. Experimental setup for CL applications in liquid chromatography (from Birks, 1989).

by oxidation of luminol with hydrogen peroxide (Burguera, 1980).

3.3 Determination of Alcohols, Ethers, Aldehydes, and Saccharides

The approach involving peroxide has been used to detect a variety of aliphatic oxygen-containing compounds, such as alcohols, ethers, aldehydes, and saccharides. This general method is based on the principle that if an analyte of interest can be converted into a specific amount of hydrogen peroxide, then addition of a luminol-catalyst reagent mixture will permit quantitation of the peroxide produced and thus the original analyte. Gendelman and Birks have described a photooxygenation CL system that used anthraquinone-sensitized photooxygenation (Gandelman and Birks, 1982). Various alcohols, ethers, aldehydes, and saccharides were separated on a reverse-phase column using a mobile phase containing anthraquinone-2,6-disulfonate. The column effluent passed through a photochemical reactor where H_2O_2 was produced and subsequently detected via Co^{2+}-catalyzed luminol CL.

Table 1. Solvent effects on luminol chemiluminescence.[a]

	CL intensity[b]		
	Catalyst		
Organic solvent	Co^{2+}	Cu^{2+}	Cr^{3+}
Methanol	0.22	0.95	0.08
Ethanol	0.19	0.93	0.23
Acetone	0.94	0.85	0.80
Acetonitrile	0.09	0.98	1.33

[a] Taken from Birks, 1989.

[b] The CL relative intensities are given for a 10% organic–90% aqueous solvent mixture divided by the intensity for a 100% aqueous solution.

3.4 Determination of Polycyclic Aromatic Hydrocarbons

Chemiluminescence of polycyclic aromatic hydrocarbon (PAH) compounds, emitted at a platinum electrode, can be utilized to determine these compounds in aprotic (i.e., containing no protons) solvents (Cruser and Bard, 1969). This CL method has been applied for detection in liquid chromatography (Hill *et al.*, 1986). Electrogenerated CL of PAHs was observed in micellar solution (Haapakka *et al.*, 1988). When polarized during a cathodic potential pulse in the presence of peroxodisulfate ions in aqueous acetate solution (*p*H =7.0), a micelle-forming surfactant (Brij-35), and a PAH compound, an oxide-covered aluminum electrode produced an intense electrochemiluminescence where the emitting species was the micelle-encapsulated PAH. The limit of detection was $10^{-8}M$ for 9,10-diphenylanthracene.

3.5 Analysis of Proteins, Amino Acids, and Other Biological Systems

Species that do not participate directly in CL reactions can still be detected by their interactions with other species which are catalysts for the reaction. This indirect approach has been used in several variations to detect a wide range of analytes. For example, amino acids have been separated via ion exchange and detected on the basis of their complexation with catalysts such as Co^{2+} or Cu^{2+}. Several amino acids including arginine, glycine, histidine, hydroxyproline, proline, and valine have been analyzed with limits of detection from 1 pmol for histidine to 2 nmol for most of the other amino acids (Koerner and Nieman, 1987). It is noteworthy that this method permits detection of both primary and secondary amino acids without derivatization. This general approach allows detection of substances that complex with a transition-metal ion catalyst. Catecholamines, with Cu^{2+} as the catalyst, can also be analyzed by reverse-phase chromatography and detected at 1 pmol levels by the CL method (Koerner and Nieman, 1987). Thyroglobulin, gamma-globulin, and ovalbumin have been determined via size-exclusion chromatography using CL detection (Hara *et al.,* 1985).

Clinical methods for protein analysis have been developed using changes in either absorption or fluorescence characteristics induced by the formation of complexes with the probe molecule. One of the major advantages of such methods is that chemical modification of the analyte is not required. Woolf and Crayeski have developed a method using CL to quantitate protein-probe complexes (Woolf and Grayeski, 1988). The method was based on chemical excitation of a complex formed by the binding of a fluorescent molecule to the hydrophobic portion of the protein. The fluorescent molecule was excited with an aqueous peroxyoxalate CL reaction. The resulting CL intensity was enhanced by the hydrophobic environment of the protein. This CL method was developed for human albumin and has a detection limit of 21.6 ng/mL.

3.6 Analysis of Drugs and Food Components

Aminoglycoside antibiotics are good Cu^{2+} ligands, and the three components of gentamicin-C in commercial gentamicin sulfate have been determined via ion-pair chromatography using CL detection (Koerner and Nieman, 1987). Many organic compounds such as drugs and food components are slowly oxidized in air and produce an extremely weak CL emission. This phenomenon has been used to investigate the stability of organic compounds (Edo *et al.,* 1985). Chemical change of food components during processing and storage involves the amino-carbonyl reaction. Investigations of the amino-carbonyl reaction have dealt with melanoidines, the main reaction products, and with the effects of amino-carbonyl reactions on the aging of the human body and diseases of adulthood (Edo *et al.,* 1985). Kurosaki *et al.* (1989) have reported a weak CL emission from amino-carbonyl reactions of *L*-lysine with various sugars and aldehydes in aqueous solution. Among the aldehydes and sugars investigated, glycolaldehyde and *D*-arabinose exhibited the most intense CL signal (Table 2). The browning reaction monitored from the UV absorption at 420 nm was used as an index of the amino-carbonyl reaction. The results in Table 2 shows a gradual increase during the course of the reaction and a similar trend was observed for the CL emission intensity. The studies indicated that the CL signal of the amino-carbonyl reaction reached a maximum after approximately 1 h. The CL emission also depended on the *p*H, and a linear relationship between the CL intensity and hydrogen-ion concentration was reported. In nitrogen-purged solutions, no CL was observed, thus supporting the hypothesis that the CL of the amino-carbonyl reaction is produced by an oxidation reaction. The study indicated that CL may be useful in evaluating the stability of

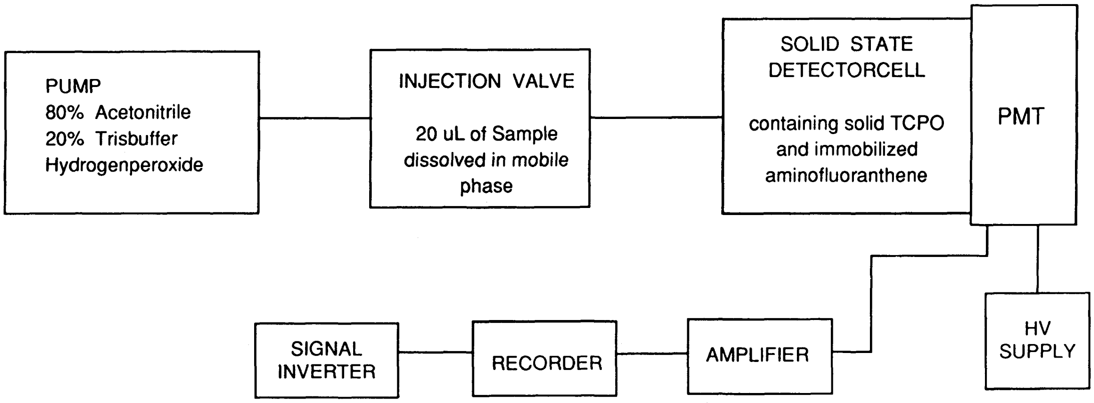

FIG. 11. A CL detector for flow-injection analysis and high-performance liquid chromatography (from Van Zoonen *et al.*, 1986).

Table 2. Chemiluminescence in the amino-carbonyl reactions of sugars and other carbonyl compounds with *L*-lysine.[a] (Taken from Kurosaki *et al.*, 1989.)

	Absorbance at 420 nm (counts/s)	
Sample	Maximum CL	Browning[b]
Hydroxy aldehyde		
Glycoaldehyde	46 370	206.0
Polyhydroxy aldehydes and ketones		
Aldotriose		
Glyceraldehyde	30 450	12.9
Aldopentoses		
D-arabinose	42 130	3.3
D-Ribose	30 530	3.6
D-xylose	37 890	3.0
Aldohexoses		
D-glucose	31 030	0.8
D-galactose	23 450	1.6
Ketohexoses		
D-fructose	27 980	1.5

[a] Aqueous solutions of *L*-lysine and sugars (each 50 nmol/L) were heated at 100 °C.

[b] The browning absorbance at 420 nm was determined after 25 min.

crude drugs extracted from natural resources that contain many amino acid derivatives including proteins and sugars.

3.7 Determination of Sulfides

Several oxidizing systems have been used to determine sulfides by monitoring their CL emission (Klockow and Teckentrup, 1976). These CL systems involved sodium hypochlorite with and without a sensitizer, and hydrogen peroxide with osmium tetroxide, peroxidase, or catalase as catalyst. Sulfides could be detected over a large concentration range from 10^{-4} to 100 ppm.

3.8 Chemiluminescence Immunoassays

Chemiluminescence immunoassays can be performed by labeling either antibodies or antigens with a CL compound of interest. CL immunoassays offer several advantages, including high sensitivity, excellent stability of the label, simplicity of the instrumentation, and flexibility of the measurement. Several types of CL immunoassays can be performed. The simplest method involves direct replacement of the radioactive label, e.g., ^{125}I, by a CL label. Such applications have used dioxetanes (Wynberg *et al.*, 1981), acridium phenyl esters (Weeks *et al.*, 1983), and isoluminol (Schroeder *et al.*, 1978) to label haptens (parts of antigens that induce an immune response) or small peptides. A variety of assays using CL have been developed for the quantification of protein and peptide hormones such as hCG, human chorionic gonadotrophin (Kohen *et al.*, 1984), and *alpha*-fetoprotein (Weeks *et al.*).

3.9 Nucleic Acid Detection, Hybridization, and Sequencing

One of the most widely used techniques in molecular biology deals with nucleic acid hybridization to detect specific DNA sequences immobilized on membranes. In addition, the increasing interest in human genome mapping and sequencing has induced the rapid development of many DNA cloning and sequencing techniques. The usual procedure for detection and quantification of specific DNA or RNA sequences is mainly based on the procedure originally described by Southern (1975). This method, often referred to as Southern blotting, involves transferring electrophoretically separated DNA fragments from a gel to a membrane support followed by hybridization with a labeled probe. The blot is washed and the DNA of interest is detected by autoradiography or other detection techniques. Conventional methods use radioactive labeled probes which give high sensitivity for detection. Separated DNA fragments are most often detected with radioisotopic labels such as ^{125}I, ^{32}P, and ^{35}S. Radiometric techniques generally suffer from several disadvantages: radioisotopes have limited shelf-life (e.g., about 60 days or ^{125}I); instruments for detection and reagents are relatively expensive; and special precautions are required for the shipping, handling, and waste disposal of radioactive materials. Consequently, a number of techniques have been developed for nonradiative probe labeling and detection. Various nonradioactive labels include fluorophors (e.g., fluorescein and tetramethylrhodamine), metal atoms (e.g., colloidal gold), enzymes (e.g., alkaline phosphatase, acid phosphatase, microperoxidase, and horseradish peroxidase), and antigen:antibody or biotin:avidin enzyme-conjugated systems (Wilchek and Bayer, 1988).

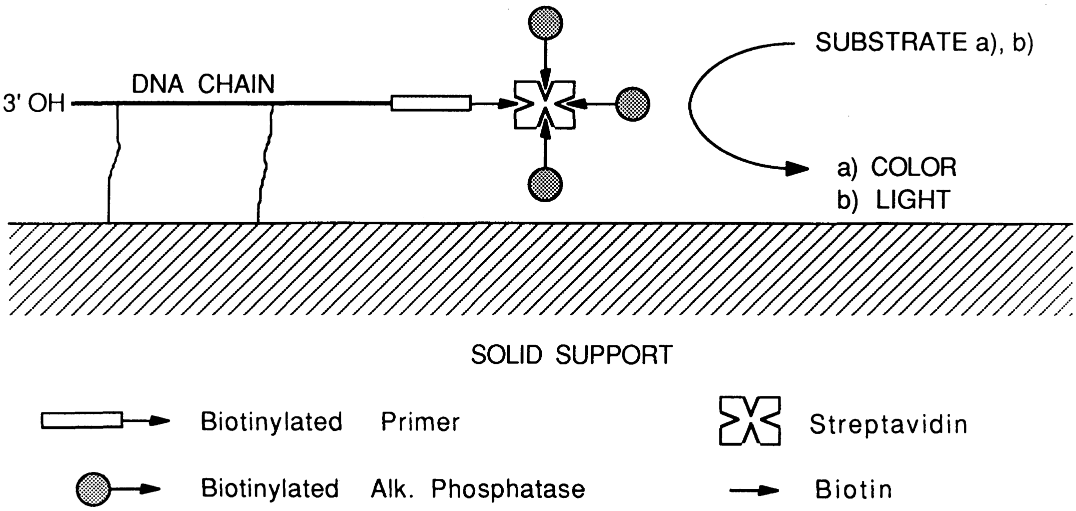

FIG. 12. Chemiluminescence detection of immobilized DNA using an enzyme-catalyzed reaction (from Beck *et al.*, 1989). The curved arrow indicates reaction of the substrate with biotinylated alkaline phosphatase to form a luminescent or colored product.

Because the luminol moiety can be quantitated conveniently at trace levels, there is current interest in the use of luminol or related derivatives as labels in immunoassays (Kricka and Carter, 1982).

Recently, enzyme-based chemiluminescence and bioluminescence assays have been developed for detecting DNA. Enzyme-based labels offer a nonradioactive technique that is very sensitive because of their catalytic amplification process. Alkaline phosphatase (AP) is an enzyme system that has been used recently for CL assays. Alkaline phosphatase exhibits a high degree of heat stability. The crude enzyme, extracted from *E. coli,* shows no loss of activity when heated at 85 °C for 30 min, or at 95 °C for 8 min in the presence of magnesium (Schlesinger and Barret, 1965). Alkaline phosphatase is also stable in low-*p*H environments. These properties of remarkable thermal stability and excellent resistance to environmental conditions makes alkaline phosphatase a commonly used label for DNA detection. Bronstein *et al.* reported recently the development of a direct CL dioxetane-based enzyme substrate which provides improved sensitivity (Bronstein *et al.*, 1990). Dioxetanes are four-membered ring peroxides incorporating a weak oxygen-oxygen bond which easily decomposes to form strong carbonyl products. This chemical reaction is accompanied by a release of up to 100 kcal/mol of energy capable of producing a strong CL emission. The luminescence signal is produced from a direct CL substrate, disodium 3-(4-methoxyspiro [1,2-dioxetane-3-2'-tricyclo$[3.3.1.1.^{3,7}]$decan]-4-yl) phenyl phosphate (AMPPD), which decomposes upon dephosphorylation with alkaline phosphatase. The CL signal obtained from alkaline-phosphatase cleaved AMPPD on a Southern blot may be detected on instant film or x-ray film.

Beck *et al.* (1989) reported a nonradiative detection technique for DNA sequencing based on CL using biotinylated alkaline phosphatase. The CL detection methodology was similar to

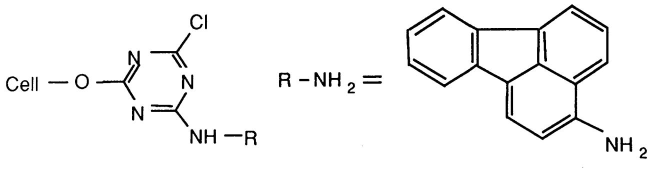

FIG. 13. Example of reactant immobilized on solid substrate: 3-aminofluoranthene immobilized on cellulose for CL detection of hydrogen peroxide (from Gubitz *et al.*, 1985).

colorimetric staining and involved a biotin/streptavidin interaction. A 5′-biotinylated oligonucleotide was used as a primer extension reaction for DNA sequencing or as a probe for hybridization to a target DNA. Biotinylated alkaline phosphatase was bound to the biotinylated 5′-end of the oligonucleotide via a streptavidin bridge and catalyzed the light reaction by cleaving off a phosphatase group from an added CL substrate. The CL emission was detected by a Polaroid or x-ray film. The method, capable of detecting DNA in the hundred-attomol range, has been developed for DNA sequencing. Figure 12 shows a schematic diagram for the CL detection of DNA using an enzyme-catalyzed reaction. The results currently reported by many researchers have indicated that luminescence detection technology using CL systems immobilized on solid substrates or membranes (Fig. 13) has permitted the development of a new generation of simple, rapid, and ultrasensitive detection assays for environmental and biochemical applications (Beck *et al.*, 1989, Gubitz *et al.*, 1985).

ACKNOWLEDGMENTS

This work was sponsored by the Office of Health and Environmental Research, U.S. Department of Energy, under Contract No. DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc. The assistance of D. L. Stokes is gratefully acknowledged.

ACRONYMS

ABEI: aminobutylethylisoluminol
CCD: charge-coupled device
CL: chemiluminescence
DNA: deoxyribonucleic acid
DNPO: bis-(2,4-dinitrophenyl)oxalate
FIA: flow-injection analysis
GC: gas chromatography
HPLC: high-performance liquid chromatography
IC: internal conversion
ISC: intersystem crossing
LC: liquid chromatography
RNA: ribonucleic acid
SIT: silicon-intensified target
SVL: single vibronic level
TCPO: bis-(2,4,6-trichlorophenyl) oxalate
VR: vibrational relaxation

GLOSSARY

Aprotic Solvent: Solvent containing no proton.

Fluorophore: A species that emits fluorescence.

Luminescence: An emission process that involves transitions between electronic states of an atom or a molecule.

Singlet State: Electronic state having a spin equal to 0.

Triplet state: Electronic state having a spin equal to 1.

Lists of Works Cited

Albrecht, H. O. (1928), *Z. Phys. Chem.* **136**, 321.

Beck, S., O'Keeffe, T., Coull, J. M. Koster, H. (1989), *Nucleic Acids Res.* **17**, 5115.

Birks, J. R. (Ed.) (1989), *Chemiluminescence and Photochemical Reaction Detection in Chromatography*, New York: VCH Publishers.

Bronstein, I., Voyta, J. C., Lazzari, K. G., Murphy, O., Edwards, B., Kricka, L. J. (1990), *Biotechniques* **8**, 310.

Burguera, J. L. (1980), *Anal. Proc.* **17**(2), 43–44.

Campion, A., Woodruff, W. H. (1987), *Anal. Chem.* **59**, 1299A.

Cruser, S., Bard, A. (1969), *J. Am. Chem. Soc.* **91**, 267.

Edo, K., Sato, H., Kato, M., Mizugzki, M., Uchiyama, M. (1985), *Chem. Pharm. Bull.* **33**, 3042.

Finlayson, B. J., Pitts, J. N., Atkinson, R. (1974), *J. Am. Chem. Soc.* **96**, 5356.

Gandelman, M. S., Birks, J. W. (1982), *J. Chromatogr.* **242**, 21.

Glinski, R. J., Sedarski, J. A., Dixon, D. A. (1981), *J. Phys. Chem.* **85**, 2440.

Gubitz, G., Zoonen, P. V., Gooijer, G., Velthorst, N. H., Frei, R. W. (1985), *Anal. Chem.* **57**, 2071.

Gundermann, K. D. (1968). *Chemilumineszenz Organischer Verbindungen*, Berlin: Springer-Verlag.

Haapakka, K., Kankare, J., Lipiainen, K. (1988), *Anal. Chim. Acta* **215**, 341.

Hara, T., Toriyama, M., Ebuchi, T. (1985), *Bull. Chem. Soc. Jpn.* **58**, 109.

Harmony, M. D. (1972), *Introduction to Molecular Energies and Spectra*, New York: Holt, Rinehart & Winston.

Hill, E., Humphreys, E., Malcolm-Lawes, J. (1986), *J. Chromatogr.* **370**, 427.

Jalkian, R., Denton, M. B. (1988), *Appl. Spectrosc.* **42**, 1194.

Klockow, D., Teckentrup, J. (1976), *Talanta* **23**, 889.

Koerner, P. J., Jr., Nieman, T. A. (1987), *Mikrochim. Acta* **2**, 79–90.

Kohen, F., Bayer, E. A., Wilchek, M., Barnard, G., Kim, J. B., Collins, W. P., Beheshti, I., Richardson, A. P., McCapra, F. (1984), in: L. J. Kricka, P. E. Stanley, G. H. G. Thorpe, and T. P. Whitehead (eds.), *Analytical Applications of Bioluminescence and Chemiluminescence,* London: Academic, pp. 149–158.

Kricka, L. J. Carter, T. J. N. (Eds.) (1982), *Chemical and Biological Luminescence,* New York: Dekker.

Kurosaki, T., Sato, H., Mizugaki, M. (1989), *J. Biolumin. Chemilumin.* **3**, 13.

Malmstadt, H. V., Franklin, M. L., Horlick, G. (1972), *Anal. Chem.* **44**, 63A–76A.

Meneryi, G., Lind, J. S. (1980), *J. Am. Chem. Soc.* **102**, 5830.

Mishalanie, E. A., Birks, J. W. (1986), *Anal. Chem.* **58**, 918.

Pitts, J. N., Jr., Kummer, W. A., Steer, R. P., Finlayson, B. J. (1972), *Adv. Chem. Ser.* **113**, 246.

Roswell, D. F., White, E. H. (1978), in: M. DeLuca (ed.), *Methods in Enzymology,* Vol. 57. New York: Academic, p. 409.

Schlesinger, M. J., Barret, K. (1965), *J. Biol. Chem.* **240**, 4282.

Schroeder, H. R., Boguslaski, R. C., Carrico, R. J., Buckler, R. T. (1978), in: M. Deluca (Ed.), *Methods in Enzymology,* Vol. 57. New York: Academic, pp. 424–445.

Seliger, H. H., Lee, J. (1972), *Photochem. Photobiol* **15**, 227.

Seliger, H. H., McElroy, W. D. (1960), *Radiat. Res., Suppl.* **2**, 528.

Southern, E. M. (1975), *J. Mol. Biol.* **98**, 503.

Spurlin, S. R., Yeung, E. S. (1982), *Anal. Chem.* **54**, 318.

Sutton, D. G., Westberg, K. R., Melzer, J. R. (1979). *Anal. Chem.* **51**, 1399.

Talmi, Y. (1975), *Anal. Chem.* **47**, 658A.

Totter, J. R. (1975), *Photochem. Photobiol.* **22**, 203.

Van Zoonen, P., Kamminga, D. A., Gooijer, C., Velthorst, N. H., Frei, R. W., Gubitz, G. (1985), *Anal. Chim. Acta.* **174**, 151.

Van Zoonen, P., Kamminga, D. A., Gooijer, C., Velthorst, N. H., Frei, R. W., Gubitz, G. (1986), *Anal. Chem.* **58**, 1245.

Vo-Dinh, T. (1984), *Room Temperature Phosphorimetry for Chemical Analysis,* New York: Wiley.

Vo-Dinh, T., Wild, U. P. (1973), *Appl. Opt.* **12**, 1286.

Vo-Dinh, T., Wild, U. P. (1974), *Appl. Opt.* **13**, 2899.

Weeks, I., Campbell, A. K., Woodhead, J. S. (1983), *Clin. Chem.* **29**, 1480.

White, E. H., Harding, J. J. C. (1964), *J. Am. Chem. Soc.* **86**, 5686.

Wilchek, M., Bayer, E. A. (1988), *Anal. Biochem.* **171**, 1.

Woolf, E. J., Grayeski, M. L. (1988), in: D. Eastwood and L. J. Cline (Eds.), ASTM STP 1009, Philadelphia: American Society for Testing and Material, pp. 67–74.

Wynberg, H., Meijer, E. W., Hummelen, J. C. (1981), in: M. DeLuca and W. D. Elroy, (Eds.), *Bioluminescence and Chemiluminescence,* New York: Academic, pp. 687–689.

Yamada, M., Ishiwada, A., Hobo, T., Suzuki, S., Araki, S. (1982), *J. Chromatogr.* **238**, 347.

Further Reading

Birks, J. W. (Ed.) (1989), *Chemiluminescence and Photochemical Reaction Detection in Chromatography,* New York: VCH Publishers.

Burr, J. G. (Ed.) (1985), *Chemi- and Bioluminescence,* New York: Dekker.

DeLuca, M., Elroy, W. D. (Eds.) (1981), *Bioluminescence and Chemiluminescence,* New York: Academic, pp. 687–689.

Gundermann, K. D. (1968), *Chemilumineszenz Organischer Verbindungen,* Berlin: Springer-Verlag.

Kricka, L. J., Carter, T. J. N. (Eds.) (1982), *Chemical and Biological Luminescence,* New York: Dekker.

Kricka, L. J., Stanley, P. E., Thorpe, G. H. G., Whitehead, T. P. (Eds.) (1984), *Analytical Applications of Bioluminescence and Chemiluminescence,* London: Academic, pp. 149–158.

VanDyke, K. (Ed.) (1982), *Analytical Applications of Bioluminescence and Chemiluminescence,* New York: Academic.

Vo-Dinh, T. (1984), *Room Temperature Phosphorimetry for Chemical Analysis,* New York: Wiley.

30
Chromatography

James M. Miller

30.1 **Introduction** **1057**
30.1.1 Historical Summary 1057
30.2 **Basic Concepts** **1059**
30.2.1 Definition of Chromatography 1059
30.2.2 Classification 1059
30.2.3 Mechanism of Operation 1059
30.2.4 Other Definitions and Symbols 1061
30.2.5 Qualitative and Quantitative Analysis 1064
30.3 **Gas Chromatography (GC)** **1065**
30.3.1 Instrumentation 1065
30.3.1.1 Carrier Gas Supply 1065
30.3.1.2 Sample Introduction 1065
30.3.1.3 Columns 1067
30.3.1.4 Detectors 1068
30.3.2 Modes of Separation 1069
30.3.2.1 Gas–Liquid Chromatography (GLC) 1069
30.3.2.2 Gas–Solid (GSC) 1071
30.4 **Liquid Chromatography** **1072**
30.4.1 Instrumentation 1073
30.4.1.1 Solvent Delivery Systems 1073
30.4.1.2 Injection Valves 1073
30.4.1.3 Columns 1074
30.4.1.4 Detectors 1074
30.4.2 Modes of Separation 1076
30.4.2.1 Liquid–Solid Chromatography (LSC) 1076
30.4.2.2 Bonded-Phase Chromatography (BPC) 1078
30.4.2.3 Reversed Phase Chromatography (RPC) 1078
30.4.2.4 Ion-Pair Chromatography (IPC) 1079
30.4.2.5 Gradient Elution 1080
30.4.2.6 Ion-Exchange Chromatography (IEC) and Ion Chromatography (IC) 1080

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

30.4.2.7 IC Instrumentation 1081
30.4.2.8 Size-Exclusion Chromatography (SEC) 1082
30.4.3 Preparative Liquid Chromatography 1084
30.4.4 Thin-Layer Chromatography 1085
30.5 Supercritical Fluid Chromatography (SFC) 1085
30.5.1 Operating Characteristics 1086
30.5.2 Modes of Separation 1087
30.5.2.1 Isobaric SFC 1087
30.5.2.2 Pressure Programming 1087
30.6 Spectrochemical Detection 1087
30.6.1 GC/MS 1091
30.6.2 LC/MS 1091
30.6.3 Other Hyphenated Methods 1092
30.7 Selected Other Topics 1093
30.7.1 Multidimensional Separations 1093
30.7.1.1 GC/GC 1094
30.7.1.2 LC/GC 1094
30.7.1.3 LC/LC 1094
30.7.1.4 Other Combinations 1095
30.7.2 Chiral Separations 1095
30.7.3 Capillary Electrochromatography (CEC) 1096
Acknowledgments 1098
Glossary 1098
References 1100
Further Reading 1101

30.1
Introduction

Chromatography is the most widely used separation method for chemical analysis. While it is not a spectroscopic method, it is often combined with or used with spectroscopic methods such as ultraviolet/visible (UV/Vis) absorption spectroscopy or mass spectroscopy (MS). A spectrometer can be an integral part of some chromatographs; for example, a UV/Vis spectrometer is most commonly used as a detector for high-performance liquid chromatography (HPLC). Alternatively, chromatographs are coupled to, or used with, spectrometers such as MS, and the combined technique, gas chromatography–mass spectrometry (GC–MS or GC/MS), is an example of so-called hyphenated methods.

This chapter focuses on chromatography, but it includes relevant information about spectroscopic detection systems, including, in some cases, the interface between the chromatograph and the spectrometer. The outline of the remainder of the article is as follows. The first section is devoted to a description of basic concepts that apply to all forms of chromatography. This is followed by sections devoted to the two most common types of chromatography, GC and LC, followed by a shorter section on the third type, supercritical fluid chromatography (SFC). The final sections elaborate on spectroscopic detection, mainly mass spectrometry (MS), and a few special topics.

For each type of chromatography, a description of the general uses of the technique is followed by sections on instrumentation and modes of separation. In Section 30.4.1, the four essential components of a chromatograph, namely, the mobile-phase (MP) delivery system, the sample injector, the separation column (i.e., the stationary phase (SP)), and the detector, are described. In Section 30.4.2, a discussion of separation principles (i.e., the physical/chemical mechanism of resolving a mixture into its individual components) is combined with examples of a few applications that illustrate the usefulness and limitations of the techniques.

30.1.1
Historical Summary

In 1903, the Russian botanist Tswett devised a separation method that is now called *liquid-column chromatography* (*LC*). He separated structurally similar yellow and green chloroplast pigments

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

by placing a leaf extract onto the top of a column of calcium carbonate particles and then washing the column with carbon disulfide (by gravity flow of the liquid). The colored pigments formed visible bands that became separated as they moved down the column. Tswett coined the term *chromatography* (from Greek, meaning "color writing"), although he also realized that the method was applicable to colorless compounds. Significant achievements in biochemistry and the chemistry of complex natural organic compounds were made possible by this method. However, the technique was time consuming and tedious, thereby restricting more widespread use in chemical separations and analysis. Faster and more efficient separations would only come years later, as a result of the theoretical and instrumental advances.

In a classic, Nobel-prize-winning paper, Martin and Synge (1941) advanced the theory of chromatography and predicted that high-speed LC would require very small column particles and high pressure. The technology for doing this was not available for many years, but, in the interim, James and Martin (1952) published the first paper on GC as had been suggested by Martin and Synge 10 years before. The potential of the technique was quickly recognized by other workers, who provided engineering advances that made gas chromatographs practical and useful, and GC became a major separation tool, providing additional advances in the theory of chromatography. It took about 10 more years until it was suggested that the principles of GC should be applied to LC. Giddings, among others, published important papers (Giddings, 1963) that gave rise to high-pressure liquid chromatography (HPLC), resulting in improved or higher performance LC, also known by the label *HPLC*. As a result of advances in instrumentation, HPLC methods have gradually approached the quality of GC methods in terms of speed and resolution. Further information about the history of chromatography can be found in the many publications by Ettre and Zlatkis (1979), Ettre (2002) and Ettre (2003).

Unfortunately, GC has some inherent limitations; it requires samples to be in the vapor phase, which limits its usefulness to compounds that can be volatilized. Current estimates are that only about 20% of known organic compounds can be satisfactorily separated by GC without some chemical modification (such as esterification of organic acids, which produces more volatile analytes). Modern LC, (HPLC), is suitable for analyzing most of the remaining 80% of known organic compounds, as well as many inorganic compounds.

Chromatography, primarily GC and HPLC, is used for rapid, efficient separation and analysis of complex mixtures of organic and inorganic compounds in the pharmaceutical, chemical, biomedical, biochemical, food, energy, cosmetics, and environmental industries, to name a few. With over 160 000 chromatographers around the world, chromatography forms one of the largest fields in the physical sciences. Clearly, it is not possible to do justice to all areas of such a large field in this short chapter. Therefore, in the sections that follow, the emphasis is on the most important theoretical, chemical, and procedural aspects of chromatography, along with an emphasis on spectroscopic detection.

30.2
Basic Concepts

30.2.1
Definition of Chromatography

Chromatography is a separation method in which the components of a sample partition between two phases: one phase is a stationary bed with a large surface area, and the other is a gas or liquid (MP), which percolates through the stationary bed. The sample is carried by the MP through the column. Samples partition (equilibrate) between the MP and the SP based on their solubilities or their relative tendencies to sorb on/in each of the respective phases. For this generalized introduction, we refer to the phases using these general terms and the symbol S for stationary and M for mobile. The components of the sample (called *solutes* or *analytes*) separate from one another based on their *relative* affinities for the two phases. This type of chromatographic process is called *elution*.

The "official" definition of the International Union of Pure and Applied Chemistry (IUPAC) is "Chromatography is a physical method of separation in which the components to be separated are distributed between two phases, one of which is stationary (SP) while the other (the MP) moves in a definite direction. Elution chromatography is a procedure in which the MP is continuously passed through or along the chromatographic bed and the sample is fed into the system as a finite slug" (Ettre, 1993).

30.2.2
Classification

Chromatography is classified according to the physical state of the MP. Thus, if the MP is a *gas*, the technique is called *GC*, and if it is a *liquid*, it is called *LC*. A subclassification can be made according to the state of the SP. If the SP is a solid, the GC technique is called *gas–solid chromatography* (*GSC*), and if it is a liquid, *gas–liquid chromatography* (*GLC*). The names used in liquid chromatography are more diverse and do not usually follow this simple pattern. A complete classification scheme is shown in Figure 30.1.

30.2.3
Mechanism of Operation

The objective in any chromatographic separation is to move individual materials through a column at different velocities so that they exit from the column at different times. In a typical experiment, a mixture of materials is injected into a flowing stream of a carrier medium (the MP) at the top of a chromatographic column, containing the SP. The MP is either a gas, a liquid, or a supercritical fluid. Common SPs are either solids or liquids. Liquids are kept stationary in the column by coating them on the inner wall of the column or by coating them on solid supports that are then packed (like solids) in the column. An alternative way to classify columns is to differentiate between packed columns and nonpacked, or open tubular (OT) columns.

A component in a sample mixture can interact with the mobile and SPs in a number of given ways. Operative partition mechanisms in chromatography are often discussed in terms of the nonpolar or polar nature of the sample, MP, and SP. A material is said to have nonpolar character if the electronic charge is uniformly distributed in the individual

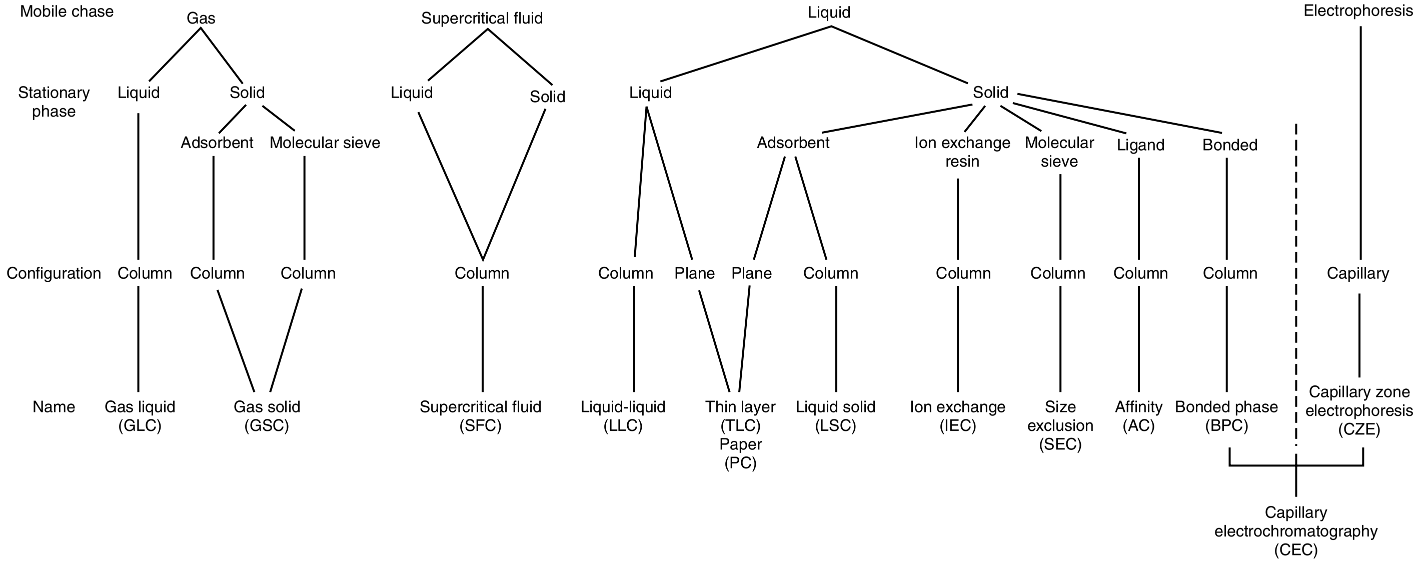
Mobile chase
Stationary phase
Configuration
Name
Gas
Liquid
Solid
Adsorbent
Molecular sieve
Column
Column
Column
Gas liquid (GLC)
Gas solid (GSC)
Supercritical fluid
Liquid
Solid
Column
Supercritical fluid (SFC)
Liquid
Liquid
Solid
Adsorbent
Ion exchange resin
Molecular sieve
Ligand
Bonded
Column
Plane
Plane
Column
Column
Column
Column
Column
Liquid-liquid (LLC)
Thin layer (TLC) Paper (PC)
Liquid solid (LSC)
Ion exchange (IEC)
Size exclusion (SEC)
Affinity (AC)
Bonded phase (BPC)
Electrophoresis
Capillary
Capillary zone electrophoresis (CZE)
Capillary electrochromatography (CEC)

Fig. 30.1 Classification of chromatographic techniques (Miller, 2005); reprinted with permission.

molecules that compose the material. An example of a nonpolar molecule is methane (CH_4), which has a zero dipole moment because of the symmetrical arrangement of the individual dipoles constituted by the carbon–hydrogen bonds. In contrast, a material is said to have polar character if any region of the individual molecules that compose the material possesses an excess or deficiency of electronic charge relative to the rest of the molecule. Many electrically neutral molecules have polar character that can be exploited in chromatographic separations. As an illustration, consider the polar nature of the molecules benzene, butanol, 2-pentanone, nitropropane, and pyridine (Miller, 2005). The π-electron cloud of the benzene molecule can interact with the mobile and SPs through an electron donation mechanism. Butanol can interact via hydrogen bond formation with proton-accepting sites on the liquid or SP. The 2-pentanone molecule can interact through interactions of its dipoles with those of the mobile or SP. The nitro group on nitropropane is an electron-accepting group, which will be retained by SPs that possess electron-donating character. Finally, the unpaired electrons on the pyridine molecule can participate in the formation of hydrogen bonds with mobile or SPs that possess proton-donating character.

A final example demonstrates the concept of relative polarity. Consider a family of aliphatic, straight-chain alcohol molecules, each member of which is classified as polar. As the length of the carbon chain is increased, the polarity of the molecule decreases (relative polarity index: methanol > ethanol > propanol > butanol, etc.). In the examples provided in this chapter, we observe that polar sample molecules have a strong tendency to partition into polar mobile and/or SPs through one or more of the mechanisms described above. Conversely, nonpolar molecules will tend to partition into nonpolar phases. Ionic species are separated using familiar ion-exchange reactions as well as other methods that exploit their ionic nature.

30.2.4
Other Definitions and Symbols

The hypothetical chromatogram shown in Figure 30.2 displays the response of a detector as a function of time, t. The sample (composed of species 1 and 2) is applied to the top of the column as a narrow band. If a species has no interaction with the SP, it elutes from the end of the column when the total volume of MP introduced at the top of the column equals the void volume of the column. The time for elution of unretained species is t_M. The molecules of species 1 and 2 are swept along at the velocity of the MP, and they would all exit from the column at the same time were it not for the differences in their interactions with the SP. The time that each species actually spends in the SP, the adjusted retention time $t_R{}'$, determines when it will exit from the column. In practice, retention is often defined in terms of the retention time t_R:

$$t_R = t'_R + t_M \tag{30.1}$$

With reference to Figure 30.2, species 1 spends less time in the SP and therefore has a shorter retention time than species 2. Thus, the retention time serves to characterize a given analyte in a given chromatographic system. The x-axis in Figure 30.2 can also represent the

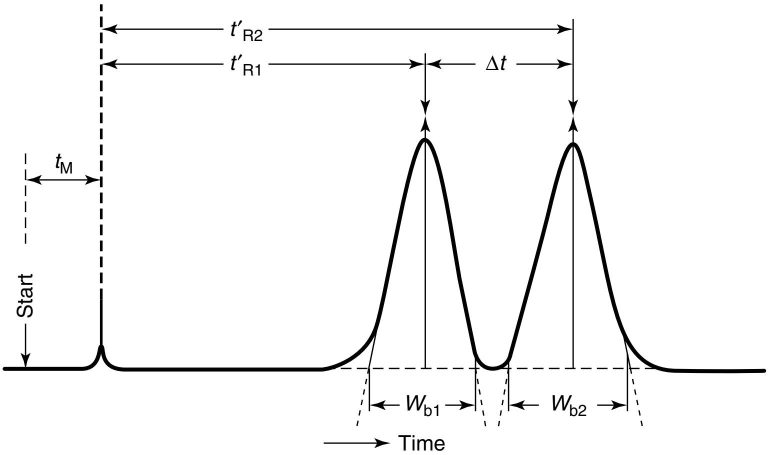

Fig. 30.2 Hypothetical chromatogram showing the separation of two analytes. The adjusted retention time is shown for each analyte as t'_R, and the retention time for a nonretained species as t_M. Chromatographic peak widths at base are labeled w_b. (Yost *et al.*, 1980); reprinted with permission.

volume V of MP that has flowed in a given time t, because, at a constant MP flow rate, F, the following relationship holds

$$V_R = t_R \times F \tag{30.2}$$

Consequently, volume and time are often used interchangeably on chromatograms and chromatographic data.

Another parameter elated to retention time is the retention factor, k:

$$k = \frac{V'_R}{V_M} \tag{30.3}$$

It too can be used for qualitative analysis for a specified chromatographic system.

While the components in the sample are being separated and eluted, the once-narrow bands broaden. This is primarily the result of three effects: (i) the differences in the path that each component travels through the column; (ii) the longitudinal diffusion effects; and (iii) slow equilibration of the components between the mobile and SPs. As a result of this band spreading, the components elute as Gaussian-like peaks like the one shown in Figure 30.3. They are broadened in proportion to the square root of the retention time, so the latter peaks are the broadest. It is this square root relationship that makes it possible to separate peaks even though the peaks are themselves being broadened, because the peaks are being separated from each other on a linear basis.

As shown in Figure 30.3, the peak width is denoted as w_b when measured at the baseline. For a Gaussian peak, that width is four times the standard deviation of the curve, 4σ. To get an estimate of the efficiency of a given separation, one can compare the retention time with the peak width. The parameter that has been defined for that purpose is the plate height, N.

$$N = \left(\frac{V_R}{\sigma}\right)^2 = 16\left(\frac{V_R}{w_b}\right)^2 \tag{30.4}$$

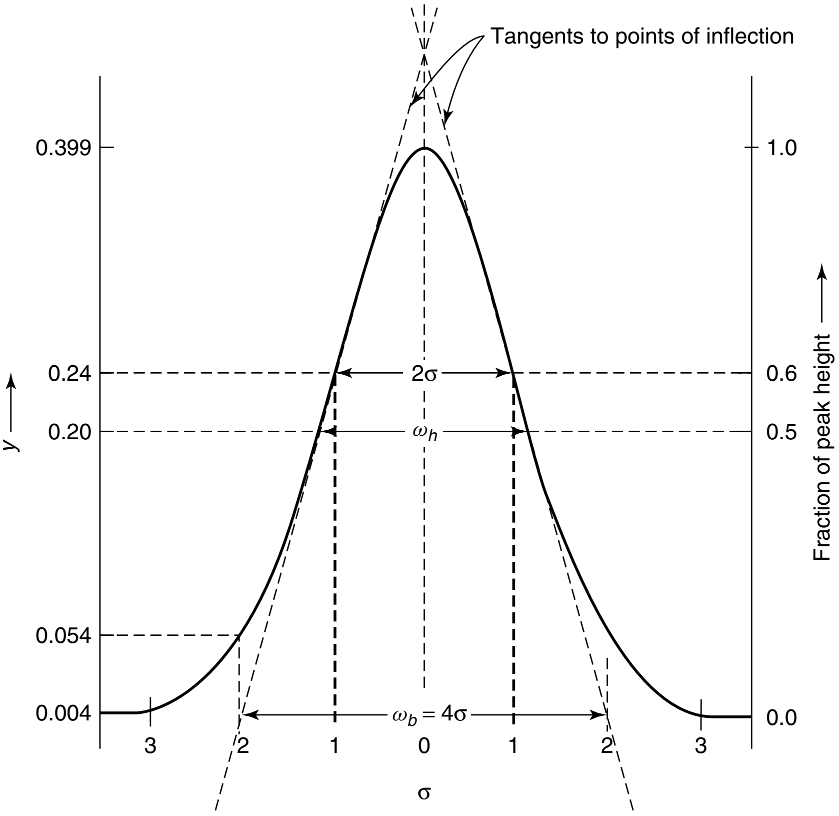

Fig. 30.3 Idealized Gaussian or normal distribution (Miller, 2005); reprinted with permission.

The name plate height came from the concept of theoretical plates that was in use in distillation processes when GC was first applied to petroleum separations. Briefly, the term *theoretical plate* referred to an imaginary unit in the chromatographic column wherein one complete equilibration of an analyte between the stationary and MPs occurred, as was visualized in distillation columns. In the entire column, there would be N of these units, but of course they do not exist in reality. For a given set of chromatographic conditions, N is approximately constant for all the peaks, and it is used in the industry to specify the quality of a column. It is a useful measure of column efficiency because it conveys the ability of the column to provide narrow peaks and good separations, and it can be easily calculated from chromatograms. This measure of column efficiency is frequently normalized to the column length L to give the plate height, H:

$$H = \frac{L}{N} \tag{30.5}$$

Other fundamental parameters are used not only to measure the quality of a separation scheme but also in the design of a separation. The separation factor α is a measure of the ability of a particular separation scheme to separate two components, and it can be defined in terms of the quantities represented in Figure 30.2:

$$\alpha = \frac{k_2}{k_1} = \frac{V_{R2}'}{V_{R1}'} \tag{30.6}$$

If the separation factor α is 1.0, the peaks coincide and there is no separation.

The efficiency of separation depends on the difference in retention times and the sharpness of the peaks, and is commonly described in terms of the resolution (R_s) of a particular method:

$$R_s = \frac{2(t'_{R2} - t'_{R1})}{(W_{b1} + W_{b2})} \tag{30.7}$$

Narrow peaks with very different retention times are said to be separated with high resolution. A well-resolved peak arising from a pure substance is very desirable from both an analytical and a preparative viewpoint. R_s can also be expressed in terms of N, k, and α (Purnell, 1960).

$$R_s = \frac{1}{4}(N)^{\frac{1}{2}}\left[\frac{(\alpha - 1)}{\alpha}\right]\left[\frac{k_2}{(1 + k_2)}\right] \tag{30.8}$$

A fundamental knowledge of the physical and chemical processes occurring on the column is critical to the successful design and performance of a separation. The experimental variables that control N, k, and α (and consequently R_s) are different for GC, LC, and SFC, and are briefly discussed in the sections that follow. Chromatographers experienced in these techniques know how to adjust the experimental conditions to maximize the resolution of a given separation.

30.2.5
Qualitative and Quantitative Analysis

Since most chromatographs are equipped with detectors, it is possible to use the chromatographic data (the chromatogram) for both qualitative and quantitative analysis. The chromatographic retention parameters such as retention time or retention factor do serve to characterize a given solute and can be used for making qualitative identification when compared with standards. Correct peak assignments can be made by comparison with a literature chromatogram or by comparison with a chromatogram from injection of a known standard. However, for positive confirmations of the identities of unknown components in a mixture, spectral measurements are required, most commonly by MS. Chromatographic data alone are not sufficient for identifying a true unknown chemical. Alternatively, identification can be done by collection of effluent fractions (corresponding to individual components) followed by off-line spectroscopic analysis.

Quantitative analysis is usually based on measurements of the heights or areas of the chromatographic peaks and comparison with heights or areas resulting from known standards in accordance with normal analytical practice. Especially in GC, it is common to use the internal standard method or the method of standard addition. When standards are not available, or when peaks have not been identified, it is common practice to calculate an area percentage by dividing the peak area of the desired peak by the sum of the areas of all the peaks. Such a calculation provides only an approximate quantitative analysis because the detection sensitivity of various compounds varies and only true calibration with standards can account for these differences. Also, it is often the case that some compounds may not have been detected by the particular detector being used (a nonuniversal detector) so the true sum of the peak areas cannot be found. Modern chromatographs do have good data acquisition and processing equipment (e.g., computers with integration and calibration software) for fast and accurate quantitation.

30.3
Gas Chromatography (GC)

GC was introduced as an analytical and separation technique in the 1940s and 1950s, in large part because of the work of Martin and Synge, referred to earlier. The first commercial chromatographs became available in 1955. The first ones were made to use packed columns, so all of the early literature uses them. When OT capillary columns were invented, the proponents attempted to show that they were superior, but these columns were patent-protected and not too widely used. Also, they required some modifications of the early instruments, resulting in two types of GC instruments. Today, the instruments designed to accept capillary or OT columns are the more popular ones and they are featured in this discussion.

GC is used for analysis of mixtures containing inorganic or organic compounds that have vapor pressures greater than about 0.1 Torr (10 Pa) at instrument operating temperatures. GC can be used to analyze solids, liquids, and gases. Amounts ranging from 10^{-2} to 10^{-12} g can be introduced and analyzed.

30.3.1
Instrumentation

A schematic of a capillary GC instrument is shown in Figure 30.4. Each of the individual parts is briefly described.

30.3.1.1 Carrier Gas Supply

The most popular carrier gases used as GC MPs are helium, and hydrogen, followed by nitrogen and argon. Where available, helium is preferred over hydrogen because of the latter's explosive characteristic. Carrier gases must be very pure, and gas scrubbing techniques for removal of oxygen, hydrocarbons, and water are used to assure purity at the point of use. Flow rates are accurately controlled and, for those cases where the pressure may vary, modern instruments are equipped with automatic pressure regulation to achieve constant flow rates. Typical flow rates are 20–100 ml min^{-1} for packed columns and 0.1–10 ml min^{-1} for capillary columns.

30.3.1.2 Sample Introduction

Microliter syringes are most commonly used for the introduction of liquid samples and solutions. Automated systems using microsyringes are highly desirable and readily available. For gases, gas-tight milliliter syringes or valves are used. Injector temperatures are regulated so as to be high enough to vaporize the injected sample, but not so high as to decompose it. Injectors usually have a rubber septum over the top of the column, which is pierced by the syringe during sample introduction. Direct introduction of the sample onto the column is possible and is known as *on-column injection.*

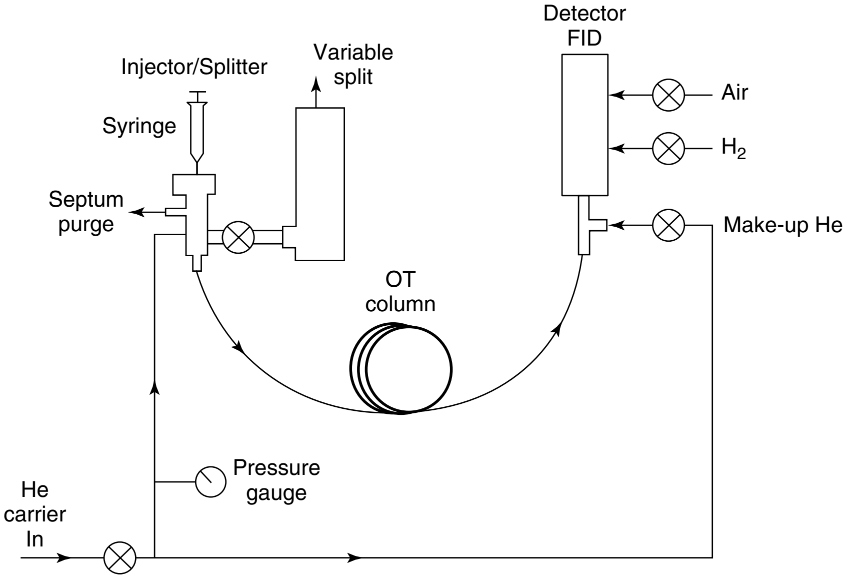

Fig. 30.4 Schematic of a typical gas chromatograph (McNair and Miller, 1998); reprinted with permission.

Because of the difficulty in injecting very small volumes of samples as required for capillary column GC, a sample inlet splitter is usually used. Small samples from split injection maintain separation efficiency and resolution. Sample splitters have a mixing chamber in which the injected sample and carrier gas are thoroughly mixed prior to being split. Typical splitting involves sending 50–99.9% of the mixture to waste through a vent tube, with the remainder being introduced onto the column. It is also possible to make a so-called splitless injection, whereby the entire mixture is directed to the column using the same apparatus, but a more involved procedure is required. A stated earlier, on-column injection is a third alternative.

The introduction of gas samples is most often accomplished with a gas sampling valve. In one valve position, a sample loop of known volume is filled with the sample gas at a known pressure. When the valve is activated, a precise amount of sample is swept out of the sample loop and onto the column by the carrier gas. This is the same type of valve that finds use in HPLC.

Other techniques can be used to introduce samples from unusual matrices. Injection of volatile components present in a relatively nonvolatile liquid can be performed by the headspace (HS) method. In one form of HS, called *purge-and-trap technique*, the liquid is sparged with an inert gas, and the volatile components are trapped by adsorption on a solid support or by condensation in a cold trap below the boiling point of the most volatile component. The trapped species are then volatilized and directed into the carrier gas stream.

Pyrolysis GC is useful for nonvolatile samples such as organic polymers. In this method, the sample is thermally pyrolyzed under controlled conditions, and the decomposition products are then analyzed by GC.

30.3.1.3 Columns

There are two types of GC columns: packed and capillary. Typical packed columns are constructed from 2–4-m lengths of stainless-steel or glass tubing with inside diameters (i.d.) of 2–4 mm. Columns are packed with 100–1000-μm-diameter porous particles. The particles are left untreated for separation by GSC and are coated with a liquid phase for GLC. Typical loading levels of liquid phases on the solid supports are about 0.5–10.0% by weight.

Typical capillary columns are 10–100-m lengths of 0.1–1.0-mm-i.d. metal or glass tubing. These columns have no packing material, and are often referred to as *OT columns*. The inner wall of the tubing has a coating that functions as the SP. GSC capillary columns have a porous layer deposited on the inner wall (porous-layer open tubular, (PLOT)). In GLC, the liquid phase may be directly applied to the inner wall (wall-coated open tubular, (WCOT)) or to a support material coating the inner wall (support-coated open tubular, (SCOT)). The thickness of a typical liquid SP film in a capillary column ranges from 0.2 to 0.5 μm. The tremendous popularity of capillary columns is due to the extremely high resolution separations that can be obtained. SPs with a variety of chemical and physical properties are available. As is seen in Section 30.3.2, the choice of an SP is dictated by the separation problem at hand.

Accurate control of column temperature is essential for reproducible GC. Column ovens that regulate temperatures within 0.2 °C between 25 and 450 °C are adequate for this purpose. Cryogenic equipment is used for temperature regulation down to about −100 °C for those applications requiring lower temperatures.

Temperature Programming (PTGC) Isothermal GC, the maintenance of a constant temperature on the column, can result in poor resolution of early-eluting peaks and significant band spreading of late-eluting peaks. Consequently, it is often desirable to use temperature programming whereby the column temperature is raised from a low initial temperature to a high final temperature during the GC run. Most GC ovens have the capability to provide one or more temperature gradients for this purpose.

A comparison of the effect of temperature programming can be seen by comparing Figure 30.5, an isothermal separation and the corresponding programmed temperature separation of the same sample. In the latter case, the early peaks are much better separated, the late peaks are eluted earlier, the peak shapes are nearly the same and are symmetrical, the total time required is decreased, and six additional compounds are eluted after the C_{15} peak. Good quantitation is obtained using temperature programming gas chromatography (PTGC) and it is also ideal for initial screening of new samples. There are a few disadvantages, the major one being the requirement that the SP have a wide temperature range of stability. In some cases, the SP may show increased bleeding during the run, since its vapor pressure increases with the temperature increase, causing the baseline to drift upward. However, in most cases, the advantages outweigh the disadvantages, and PTGC is very popular.

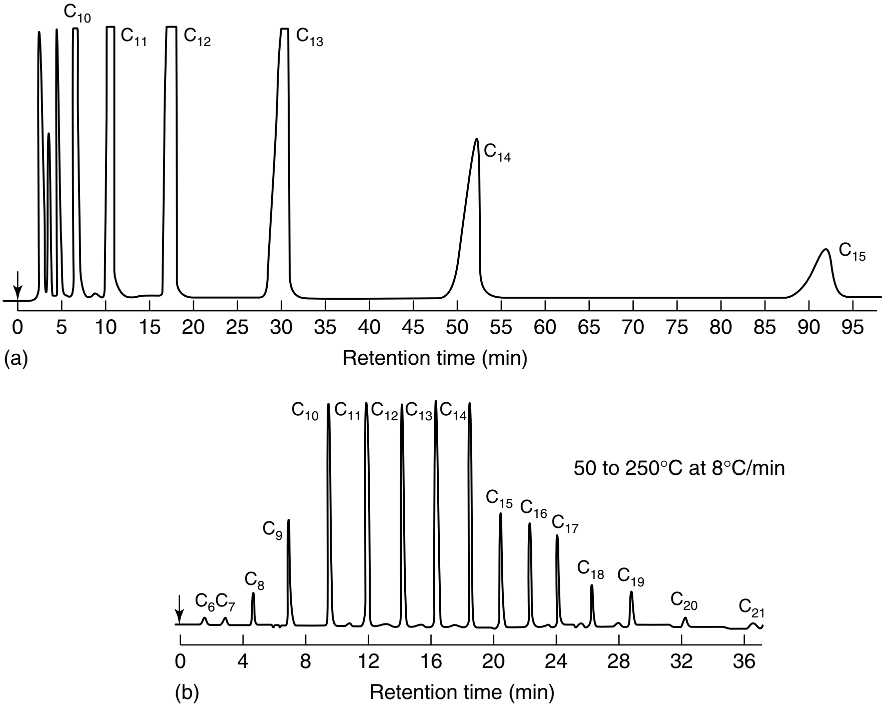

Fig. 30.5 A comparison of (a) isothermal and (b) temperature-programmed separation of *n*-paraffins (McNair and Miller, 1998); reprinted with permission.

30.3.1.4 **Detectors**

Most of the detectors used in GC were invented for this use, and many of them are based on ionization phenomena. The main exception is the thermal conductivity detector (TCD), which was already in use to measure the thermal conductivity of gases and, with minor modifications, became an ideal GC detector in the early days. For optimal use, a carrier gas of high thermal conductivity, like helium or hydrogen, is required; of course, safety precautions must be taken if hydrogen is used. It is a universal, nondestructive detector, but one with only a modest detectivity (about 10^{-8} g). The most popular one at present is the flame ionization detector (FID), which uses an oxy-hydrogen flame to burn the eluents from the GC column, thereby producing ions and electrons to form a current.

Only a few of the detectors are spectroscopic, but MS is a detector that is ideally suited to provide data for both qualitative and quantitative analysis. It is discussed later along with its use for HPLC. Infrared (IR) detection is similarly useful and it is also discussed later. The only other spectroscopic detector that finds use is the flame photometric detector (FPD). It employs a photomultiplier tube in combination with optical filters to monitor light emission generated in a flame similar to that used in the FID. Bandpass filters that transmit 394- and 526-nm radiation make this detector specific for luminescence

Tab. 30.1 Gas chromatography detectors.

Detector	Detection limit (g)	Selectivity
Thermal conductivity (TCD)	10^{-8}	Universal
Flame ionization (FID)	10^{-11}	Organics
Electron capture (ECD)	10^{-13}	Halogens, nitro-groups
Flame photometric (FPD)	10^{-10}–10^{-11}	Sulfur, phosphorus
Mass spectrometric (MS)	10^{-12}	Organics, others
Infrared spectrometric (FTIR)	10^{-8}	Most organics, others
Photoionization (PID)	10^{-10}	Aromatic organics, arsenic, sulfur, phosphorus
Hall electrolytic conductivity (HECD)	10^{-7}	Sulfur, nitrogen, phosphorus, halogens
Nitrogen phosphorus (NPD)	10^{-9}	Sulfur, nitrogen, phosphorus, arsenic
Discharge ionization (DID)	10^{-9}	Universal for liquids and gases
Gas density (GADE)	10^{-5}	High molecular weight
Helium ionization (HID)	10^{-9}	Universal for gases
Ultrasonic	10^{-7}	Universal for gases

from sulfur- and phosphorus-containing species, respectively. The detection limit of the FPD is about 10^{-11} g for phosphorus-containing and 10^{-10} g for sulfur-containing analytes. A few characteristics of GC detectors are given in Table 30.1.

30.3.2
Modes of Separation

30.3.2.1 Gas–Liquid Chromatography (GLC)

In GLC, the important parameters are the nature of the liquid SP and the temperature. The MP is inert and relatively unimportant in determining the mode of separation. Historically, GLC is also referred to as *partition chromatography* since the separation mechanism is based on partition equilibration of the analyte between the liquid and gas phase.

Hundreds of high boiling liquids and some solids have been used as SPs, providing a wide variety of available chemical properties (e.g., polarity, chemical functionality) and physical properties (e.g., particle size, film thickness). In fact, there are too many to enable one to select the very best one for a given analysis. However, with the current emphasis on the very efficient capillary columns, the need for high selectivity has decreased and only a few carefully chosen SPs are necessary to handle most analyses. A family of silicone polymers that differ in the functional groups they contain has become the most popular SPs. The least polar one is a polydimethylsiloxane polymer abbreviated as PDMS. Replacement of the methyl groups with more polar ones like phenyl, cyano, or fluoro results in more polar phases. Another common polar SP is a polyethyleneglycol like Carbowax 20M. Interest in using the class of compounds known as *room temperature ionic liquids* (*RTILs*) is increasing. Often, phases are chemically bonded to the inside wall of a silica capillary tube to make WCOT columns. For a more complete description of the range of available choices and the criteria used in their selection, the reader is referred to works by Grob and Barry (2004) and McNair and Miller (1998) which cover this subject in detail. In this section, some examples are provided to illustrate the capabilities of GLC.

The gas chromatograms in 30.5, which were used to illustrate PTGC, show the separation of a mixture of aliphatic hydrocarbon standards as described earlier. The hydrocarbons are typical of those found in gasoline-range petroleum products. The subscript *i* in the designation C_i refers to the number of carbon atoms found on the straight-chain backbone of the molecule. These chromatograms were obtained on a capillary column (coated with a nonpolar liquid phase) using column temperature programming and an FID. The nonpolar hydrocarbons have a high affinity for the nonpolar SP. A given hydrocarbon will reside in the SP until the temperature is raised high enough to generate a partial vapor pressure sufficient to increase its equilibrium concentration in the carrier gas, relative to its concentration in the SP. Therefore, the separation mechanism is primarily based on the differences in boiling points among the different hydrocarbons. The PTGC analysis begins with the column at a low temperature, below the boiling point of all the hydrocarbons. As the temperature of the column is gradually increased, hydrocarbons with successively higher boiling points are successively eluted. This type of analysis is a standard test method for determining the boiling range distribution of petroleum products, a so-called simulated distillation.

Columns are often developed and optimized for specific applications. For example, a polar SP, consisting of methyl phenyl cyanopropyl silicone bonded to a fused-silica open tubular (FSOT) column, was developed specifically for the trace analysis of tetrachlorodibenzo-p-dioxins. Analysis of these compounds is very important because of their extreme toxicity, particularly 2,3,7,8-tetra-chlorodibenzo-p-dioxin, which is classified as a supertoxic material. The chromatogram shown in Figure 30.6 is for nine isomers of tetra-chlorodibenzo-*p*-dioxin that differ only in the position of the chlorine atoms on the aromatic ring structure. The GC effluent is monitored with mass spectrometric detection, using selective monitoring of ions with mass-to-charge ratios of 320 and 322. These values correspond to the molecular weights of ions formed by ionization of the tetra-chlorodibenzo-*p*-dioxin molecule (i.e., no fragmentation), taking into account the natural relative abundance of the chlorine isotopes, ^{35}Cl and ^{37}Cl. The location of the chlorine atoms influences the relative polarities and boiling points of the individual isomers. Both properties are exploited in the separation shown in Figure 30.6.

The effect of column polarity on retention in GC is further illustrated by the example shown in Figure 30.7. Here, the injected sample, containing 25 different organic solvents, is split and chromatographed simultaneously on two capillary columns. The upper trace is the chromatogram obtained on a nonpolar column, while the lower trace is the chromatogram obtained on a polar column. Detection is achieved by the use of flame ionization detectors for effluent from each column. Notice that polar compounds such as ethanol, isopropanol, *n*-butanol, and 3-methyl-1-butanol elute more quickly from the nonpolar column than they do from the polar column, where their affinity

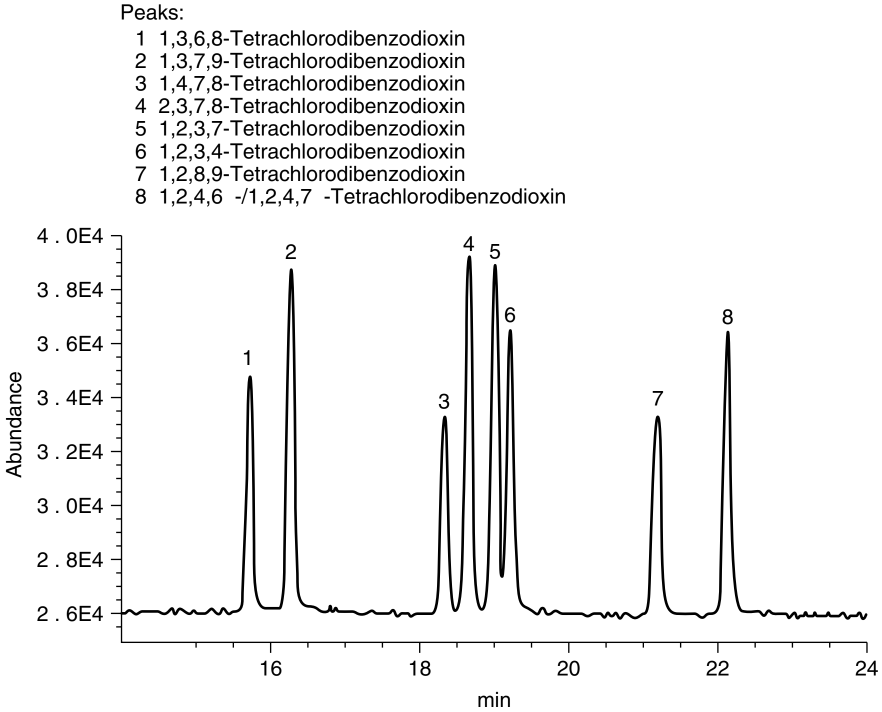

Fig. 30.6 Gas chromatogram of nanogram mixture of dioxins. Column: 35-m-long × 0.32-mm-i.d. fused-silica column, coated with 0.25-µm film of methyl phenyl cyanopropyl silicone. Column temperature: held for 0.5 min at 70 °C, then at 30 °C min^{-1} to 200 °C, then at 5 °C min^{-1} to 270 °C. Carrier gas: helium. Sample introduction: splitless injection. Detector: mass spectrometer, selected ion monitoring at m/z 320 and m/z 322. Peak assignments as noted in legend. (Quadrex Corp., 1990); reprinted with permission.

for the SP results in longer retention times. Also notice that within a series of like compounds (e.g., methyl, ethyl, isopropyl, and butyl acetate), elution is in the order of increasing boiling point on both columns, similar to the effect observed in Figure 30.5.

30.3.2.2 Gas–Solid (GSC)

GSC refers to GC performed on solid SPs. Whereas in GLC every attempt is made to minimize adsorption of solutes on active sites on the support (so that partition is based solely on liquid–liquid interactions), adsorption is the principal separation mechanism in GSC. Typical materials used as SPs include porous polymers, silica gel, diatomaceous earth (a natural form of silica), crushed firebrick, porous carbon, and zeolites (also commonly referred to as *molecular sieves*). One unique GSC separation using a molecular sieve can separate oxygen and nitrogen, as well as the other permanent gases (e.g., hydrogen, nitrogen, oxygen, argon, carbon monoxide, and carbon dioxide) and low-boiling compounds (e.g., methane, ethane, ethylene, and acetylene).

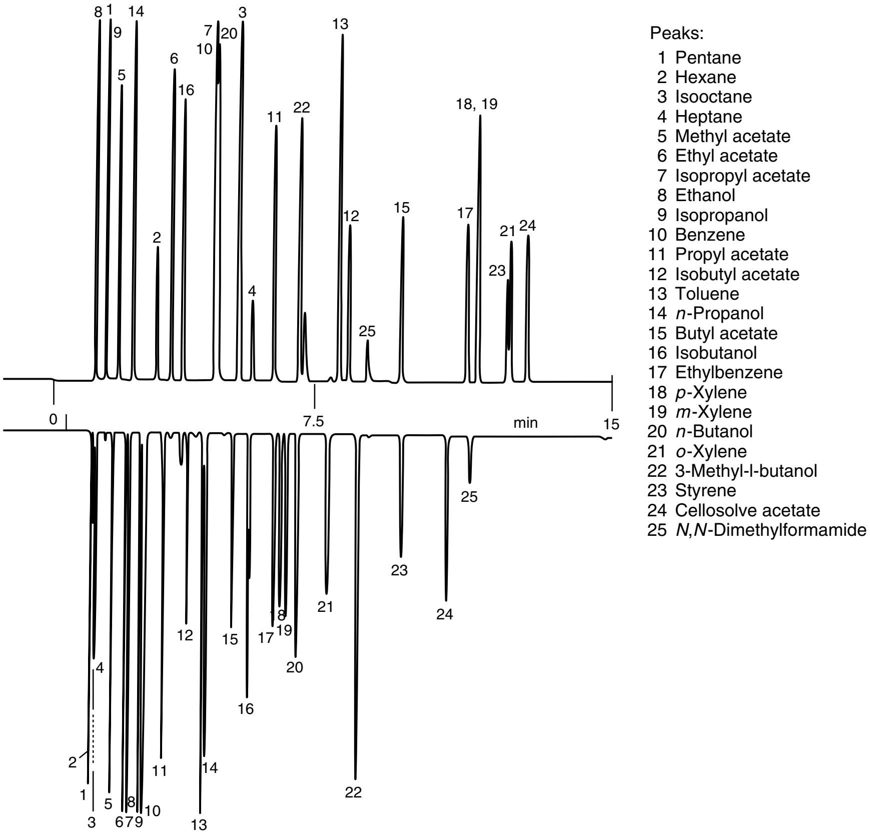

Fig. 30.7 Gas chromatograms of a mixture of organic solvents. Upper chromatogram column: 25-m-long × 0.32-mm-i.d. fused-silica column, coated with 1.0-μm phenyl vinyl methyl silicone film. Lower chromatogram column: 25-m-long × 0.32-mm-i.d. fused-silica column, coated with 0.50-μm polyethylene glycol film. Column temperature: held for 3 min at 35 °C, then at 5 °C min^{-1} to 100 °C. Carrier gas: helium, 1.5 mL min^{-1}. Sample introduction: split injection. Detectors: dual flame ionization (FID). Peak assignments as noted in legend. (HNU/Nordion Systems, 1990); reprinted with permission.

30.4
Liquid Chromatography

Modern LC, also referred to as HPLC, was introduced as an analytical separation technique in the early 1970s, and it has become the most widely used chromatographic technique. HPLC complements GC in that it provides rapid analysis of thermally labile, nonvolatile compounds that would be difficult or impossible to do by GC. Some typical examples of applications of HPLC are separation and quantitation of individual components in complex mixtures, analysis of nominally pure compounds for trace impurities, and isolation of pure compounds for synthetic and/or identification purposes. It can be used to analyze solids or liquids. A typical sample is 10 μl of a 0.1–100-mg/ml solution of the sample in a suitable solvent, although quantities as small as 10^{-6} g can be analyzed.

30.4.1
Instrumentation

A typical HPLC instrument is shown schematically in Figure 30.8. A basic, simple instrument is depicted; additional solvents, mixing valves, precolumns, and filters are often necessary, in addition. The following discussion is limited to the components shown in the figure.

30.4.1.1 Solvent Delivery Systems

Unlike GC where the MP is inert, the HPLC MP actively competes with the SP for analytes. A partitioning takes place between the two phases, so the composition of the MP is very important and may require mixtures of liquids. In one type of solvent delivery system, proportioning valves deliver precisely controlled volumes of solvent from reservoirs to the mixing chamber. A liquid chromatograph may have as many as four reservoirs, enabling solvent mixtures to be used as the MP. The liquid MP is delivered from the mixing chamber to the top of the chromatographic column by a high-pressure pumping system with output capabilities as high as 40 MPa and flow rates in the range of 0.1–10 ml min^{-1}. Gradient elution techniques (*see* Section 30.4.2.5), where the composition of a two-, three-, or four-component MP is changed in a stepwise or continuous manner, are achieved by microprocessor control of the solvent proportioning valves. Gradient elution techniques in LC are used in much the same way that temperature programming is used in GC to enhance resolution. MPs commonly used in LC include water, dilute aqueous pH buffer solutions, acetonitrile, tetrahydrofuran, methanol, chloroform, hexane, and methylene chloride, as well as miscible mixtures composed of one or more of these solvents.

30.4.1.2 Injection Valves

Injector valves are designed to accommodate sample introduction of microliter or milliliter sizes. Sample loops are completely filled with sample while the MP

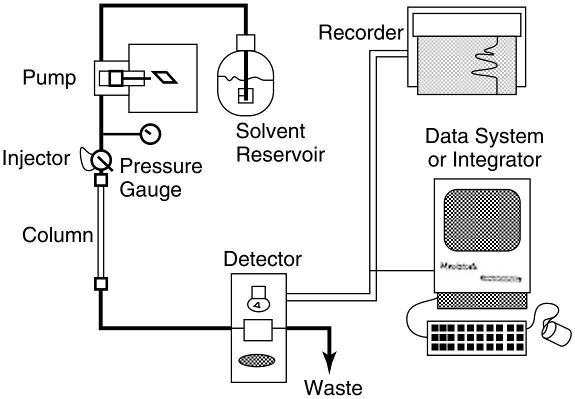

Fig. 30.8 HPLC schematic (Miller and Crowther, 2000); reprinted with permission.

is pumped through a different port on the valve. When the port to the sample loop is opened, the sample is flushed out of the loop by the MP and a precise amount of sample is delivered to the top of the chromatographic column. A typical injection volume is 1–100 µl. Automated systems based on these injection valves are commercially available.

30.4.1.3 **Columns**

The heart of the liquid chromatograph is the high-performance column on which the separation takes place. Typical columns are constructed from 3- to 25-cm lengths of 2–5-mm-i.d. stainless-steel tubing, although larger diameter columns (e.g., 60 cm) are used for preparative work and smaller diameter "microbore" columns (1 mm) for applications requiring advanced detectors (such as MS or Fourier-transform infrared (FTIR) spectroscopy). Columns are packed with small (3–50 µm) particles that constitute the SP. SPs with a variety of chemical and physical properties are available. As seen in Section 30.4.2, the choice of a SP/MP combination is dictated by the separation problem at hand.

Since chromatographic theory indicates that packed column performance will improve as the size of the SP particles is decreased, there has been a trend in column development toward smaller and smaller particles. There is a limit beyond which it is too difficult to pack these tiny particles and it is reached around 2- or 3-µm diameter. The smaller diameters that produce better performance also require higher pressures for fast analysis, and this situation has produced the latest version of HPLC commonly known as (*UHPLC*), *ultrahigh performance liquid chromatography*. UHPLC columns work at the limits of current technology, but they are the most efficient packed columns in use.

Another, alternative, commercial column introduced in 2001 is the monolithic column. Manufactured by a sol–gel process into a single stable rod, the monolithic columns so produced are enclosed in a plastic PEEK sheath. Compared to conventional microparticulate columns, the monolithic columns usually exhibit a lower pressure drop and can be operated with high efficiencies at higher flow rates.

30.4.1.4 **Detectors**

Most HPLC detectors are adapted from existing detection systems, many of which are spectroscopic. By far, the most common is based on the classic UV/Vis absorption spectrophotometer. As is the case with GC, MS is one of the most powerful and useful detectors to be used with HPLC, and it is discussed together with GC/MS. Table 30.2 summarizes some of the characteristics of the most common

Tab. 30.2 Common HPLC Detectors.

Spectroscopic
Ultraviolet absorption (UV/Vis)
Fixed wavelength UV
Variable wavelength UV
Photodiode array (PDA), also called *diode array detector* (*DAD*)
Fluorescence
Mass spectrometer (MS)
Electrical
Electrochemical detector (ECD); amperometric
Conductivity detector
Other
Refractive index (RI)
Evaporative light scattering detector (ELSD)

detectors and a brief discussion of some of them follows.

The refractive index (RI) detector functions as a nearly universal detector, that is, a detector that responds to nearly all analytes. This results from the fact that with the large number of available LC solvents, it is usually possible to design an experiment so that the analytes have refractive indices different from the MP. RI detectors are nondestructive and typical detection limits are about 10^{-6} g for a component in the injected sample.

UV/Vis absorbance detectors are among the most popular LC detectors because most organic compounds absorb 190–700-nm radiation, while many other compounds that do not absorb can be chemically reacted to form absorbing species. UV/Vis detectors are of three types: (i) fixed-wavelength detectors, which operate at only one wavelength (usually 254 or 280 nm, where a large percentage of organic compounds exhibit at least some absorbance); (ii) variable-wavelength detectors, which operate at any selected wavelength between 190 and 700 nm; and (iii) photodiode array (PDA) detectors, which can monitor the entire wavelength region between 190 and 700 nm.

PDA detectors have found increasing use in the past several years because of their utility in qualitative identification of unknowns by their characteristic UV/Vis absorbance spectra. UV/Vis detectors are nondestructive, and for a strongly absorbing compound, the lower limit of detection is about 10^{-9} g. Further details on UV/Vis absorbance can be found elsewhere in the Encyclopedia.

Fluorescence detectors are more specific than UV/Vis detectors because not all compounds that absorb light subsequently fluoresce. The use of a fluorescence detector restricts the choice of a MP to one that does not fluoresce, and minute traces of fluorescent materials commonly present in solvents can cause problems. Fluorescence detectors are nondestructive and have detection limits of about 10^{-10} g. Further details on molecular fluorescence can be found elsewhere in the Encyclopedia.

There are several electrochemical detectors (ECDs). One measures the electrical conductivity of the LC eluent by applying an alternating voltage between two electrodes and measuring the resulting current. Conductivity is related to the current–voltage ratio and represents the contribution of the individual equivalent conductivities of all ions in solution. Another is amperometric in principle and causes a chemical transformation of the analyte. The effluent from the column passes a working electrode (one of three electrodes in the detector cell) set at a specific potential with respect to the reference electrode. If a species in the effluent can be oxidized or reduced at this potential, the transfer of electrons at the electrode–effluent interface serves as the detection signal. The main advantages of the amperometric detector are sensitivity (10^{-12} g) and selectivity (by prudent choice of the potential).

In the light-scattering detector (LSD), also known as the *evaporative light-scattering detector* (*ELSD*), the effluent from the column is nebulized and the resulting fine mist is introduced into a stream of preheated gas. The mist is passed through a high-temperature drift tube where the volatile solvent is vaporized, leaving very fine particles of solute in the carrier gas stream. Laser light scattered by these particles is detected by a PDA. These detectors are meeting the need for a universal detector, which has lower detection limits than the RI detector and is compatible with gradient elution techniques, unlike RI.

Other detectors are based on specific characteristics of the analyte (radioactivity, rotation of plane of polarized light). It should also be noted that the effluent leaving a nondestructive detector (RI, UV/Vis, fluorescence, FTIR) can be collected (in effluent fractions corresponding to separated components) and analyzed off-line by other analytical techniques, such as MS, FTIR spectroscopy, or nuclear magnetic resonance (NMR) spectroscopy. This very common approach for combining powerful separation and identification tools can allow unambiguous identification of unknown components in a complex mixture.

30.4.2
Modes of Separation

As was shown in Figure 30.1, the classification of HPLC methods is not primarily according to the states of the two phases, as was the case with GC, but is more extensive. The exception is liquid–solid chromatography (LSC). Other modes, as shown in Figure 30.9, usually carry the name of the mechanism by which the separation is achieved. Also note in Figure 30.9, that another difference in the classifications, compared with GC, is the identification of those methods performed on plane surfaces, not in columns. The major one of this type is thin-layer chromatography (TLC), which is discussed separately from the column methods in Section 30.4.4.

30.4.2.1 Liquid–Solid Chromatography (LSC)

LSC is also called *adsorption chromatography* since the mechanism of the separation is mainly adsorption. It is the oldest of the various LC modes and was the one used by Tsweet. It combines a polar SP (silica or alumina) with a nonpolar or slightly polar MP (e.g., heptane, hexane, or pentane, possibly mixed with a more polar solvent like methylene chloride, chloroform, or isopropanol). Variable amounts of trace

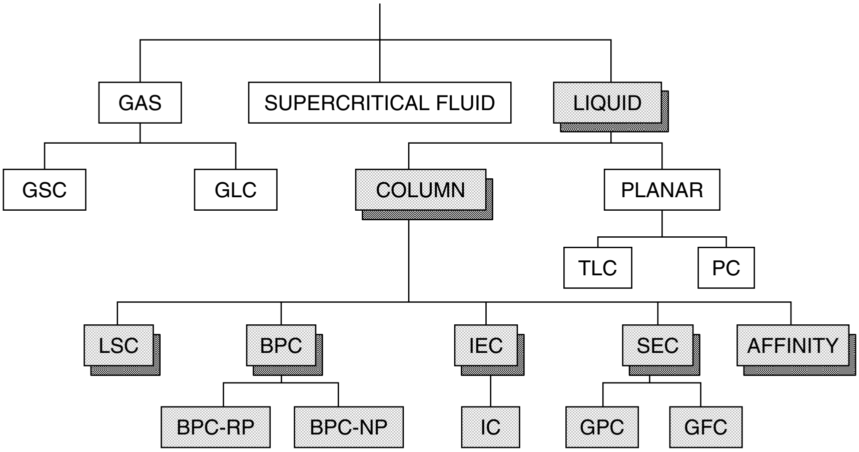

Fig. 30.9 Classification of liquid chromatographic techniques (Miller and Crowther, 2000); reprinted with permission.

water in these commonly used solvents presented a major problem with stability and reproducibility, which led to its loss of popularity. Since this system was originally the major one in use, it is also referred to as *normal phase* (*NP*) HPLC. Ironically, that name is still used even though LSC is currently not the normal mode of operation.

In general, polar molecules are strongly adsorbed on these SPs; therefore, they are the last to elute from the column. The opposite is true for nonpolar molecules. In Figure 30.10, a NP

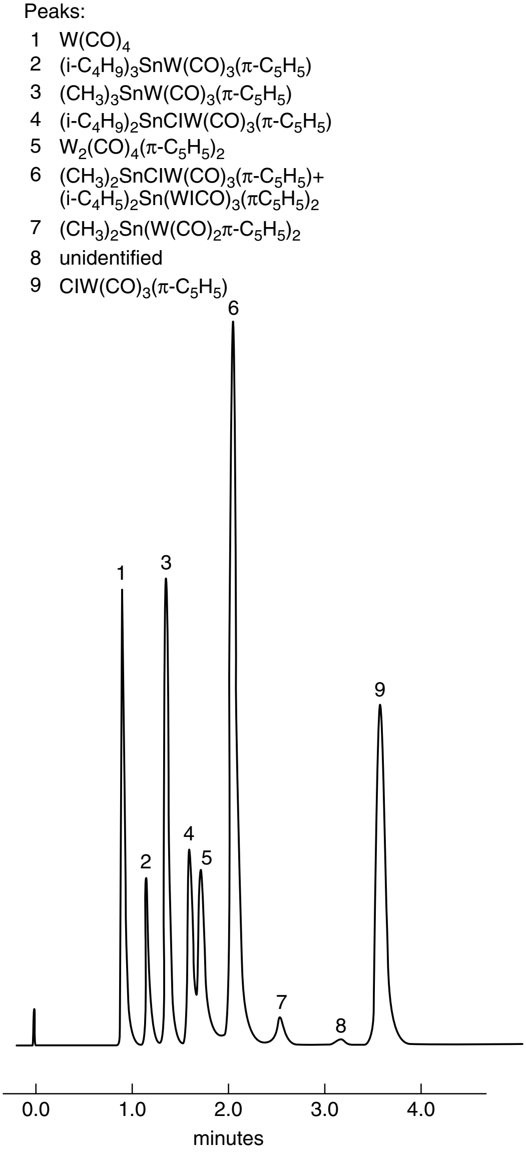

Fig. 30.10 Normal-phase liquid chromatogram of a mixture of tungsten carbonyl and tungsten tin carbonyl complexes. Column: 15-cm long × 4.6-mm i.d., 3-µm particles, cyano. Mobile phase: 10 : 90 mixture of tetrahydrofuran:hexane, 2 mL min^{-1}. Detector: UV absorbance at 220 nm. Peak assignments as noted in legend (Dewaele, 1990); reprinted with permission.

chromatogram of a mixture of tungsten carbonyl complexes and tungsten tin carbonyl complexes is shown (Dewaele, 1990). This chromatogram was obtained on a cyano column with detection based on UV absorbance at 220 nm. While LC is most commonly used for separation of organic materials, this example demonstrates its utility in resolving a complex synthetic mixture of organometallic compounds. While normal-phase chromatography (NPC) is particularly suited to nonpolar or moderately polar compounds that are not water soluble, it is not restricted to these cases.

30.4.2.2 Bonded-Phase Chromatography (BPC)

LSC and LLC have been replaced for the most part by bonded-phase chromatography (BPC) in which a liquid phase is chemically bonded to a solid support to become the SP. The main advantage of BPC compared with LLC is that the SP is stable and not susceptible to physical or chemical removal by action of the MP. The most common SPs are those bonded to siloxanes (see Table 30.3), although organic resins are returning to use after a decline in popularity in the 1970s and 1980s. Siloxane-based packing materials are prepared by silanization of the surface hydroxyl groups on the silica particles with alkylchlorosilane compounds. For example, an octadecyl column would be prepared by reaction of the silica particles with a solution of dimethyloctadecylchlorosilane to produce a chemically modified surface. It will be seen that the functional group bonded to the SP critically affects the types of separations that can be performed; some are for NPC, and others are for phases of opposite polarity, called *reversed phase* (*RP*).

30.4.2.3 Reversed Phase Chromatography (RPC)

The production and use of BPC in which the SP is nonpolar has become very popular. Since the polarity of such columns is the opposite or reverse of those used in LSC, their use has been designated as reversed-phase chromatography (RPC). This definition applies to nonpolar columns, whether or not they are the bonded type. The MP used with them is polar, making RPC complementary to NPC.

Typically in RPC, a hydrocarbon-bonded support, such as methyl-, octyl-, or octadecyl-siloxane (ODS), is used with an MP composed of water, methanol, acetonitrile (ACN), tetrahydrofuran, or a mixture of some or all of these solvents.

Tab. 30.3 Bonded-phase chromatography stationary phases.

Designation	Surface chemical formula
Silica (Si)	Si-OH
Amino (NH_2)	$Si\text{-}O\text{-}Si(OH)_2\text{-}(CH_2)_3\text{-}NH_2$
Cyano (CN)	$Si\text{-}O\text{-}Si(CH_3)_2\text{-}(CH_2)_n\text{-}CN$
Methyl (C_1)	$Si\text{-}O\text{-}Si\text{-}(CH_3)_3$
Phenyl (C_6)	$Si\text{-}O\text{-}Si\text{-}(CH_3)_2\text{-}C_6H_5$
Octyl (C_8)	$Si\text{-}O\text{-}Si\text{-}(CH_3)_2\text{-}(CH_2)_7\text{-}CH_3$
Octadecyl (C_{18})	$Si\text{-}O\text{-}Si\text{-}(CH_3)_2\text{-}(CH_2)_{17}\text{-}CH_3$

In RPC, the more polar materials have greater affinities for the MP and so elute first.

Consider, for example, the RP chromatogram of a mixture of sugars shown in Figure 30.11. This chromatogram was obtained on an amino column with gradient elution using an acetonitrile/water MP and detection based on laser light scattering (Letter, 1991). The individual sugars elute in three groups, with the low molecular weight (MW = 180) fructose and glucose molecules first, the higher molecular weight (MW = 342) sucrose, maltose, and lactose molecules next, and the highest molecular weight (MW = 504) raffinose molecule last. For these sugar molecules, polarity (and order of elution) is a function not only of the molecular weight but also of the relative position of the polar hydroxyl and hydroxymethyl groups on the molecule. The latter accounts for the excellent separation even among molecules with the same molecular weight.

30.4.2.4 **Ion-Pair Chromatography (IPC)**

When ions are present in samples run by RPC, they typically elute very quickly (with little or no retention) because they do not interact appreciably with the nonpolar SP. As a consequence, they are not separated from one another. One way to remedy this situation is to add a counterion to the MP and form ion pairs that are uncharged. Reversed-phase ion-pair chromatography (RPIPC) is performed with conventional RP LC columns and with polystyrene-divinylbenzene resins that have no ion-exchange sites. The MP contains ions that combine with the sample ions to create neutral ion pairs.

For example, the separation of a mixture of carboxylic acids by RP could be done using a MP that has a low pH, so that the acids would be present

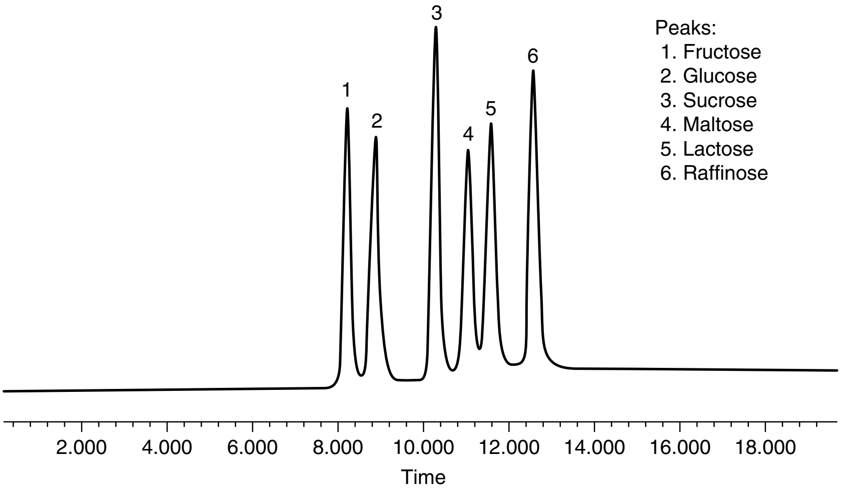

Fig. 30.11 Reversed-phase liquid chromatogram of a mixture of sugars. Column: 25-cm long × 4.6-mm i.d., 10-μm particles, amino. Mobile phase: initially 85 : 15 acetonitrile:water, then linear gradient to 60 : 40 acetonitrile:water over 10 min; 1.5 mL min^{-1}. Detector: laser light scattering. Injection size: 20 μL. Peak assignments as noted in legend (Letter, 1991); reprinted with permission.

as neutral carboxylic acids. However, if it were necessary to run the sample at a pH above that of the pK_a of the acids, the acids would be present as the carboxylate anions. To make them neutral, the addition of a cationic tetraalkylammonium ion would form neutral ion pairs with the carboxylate anions and they would likely be separated by this methodology, RPIPC.

30.4.2.5 **Gradient Elution**

Isocratic separations, where the MP is unchanged throughout the chromatographic process, often give poor resolution of early-eluting peaks, significant band spreading of late-eluting peaks, and unnecessarily long separation times. The solution to these problems is gradient elution, where the MP composition is continuously changed during the separation. It is used to speed the elution of strongly retained compounds or to retard the elution of early-eluting compounds. For optimum resolution and speed of analysis time, it is necessary to begin with a MP just strong enough to dissolve the most soluble species in the sample, and finish with a MP strong enough to remove the most insoluble species from the column. The resulting separation is optimized similar to the way GC separations are optimized by programmed temperature operation.

30.4.2.6 **Ion-Exchange Chromatography (IEC) and Ion Chromatography (IC)**

IC and ion-exchange chromatography (IEC) are used for the separation of ionic compounds, inorganic and organic. IEC is performed on SPs that possess ionic sites on the surface, using aqueous, ion-containing MPs. Columns are packed with small (5–25 μm) particles of polystyrene-divinylbenzene resin or (less commonly) silica whose surfaces are chemically modified for specific applications. For anion analysis, the SP particles have surface cationic sites such as amines ($NR_3{}^+$). For cation analysis, the SP has surface anionic sites such as sulfonates ($SO_3{}^-$).

In classical IEC, separation is based on ion exchange of sample ions (X) and MP ions (Y) with the charged groups (S) of the SP. If anions are to be separated (anion exchange), the SP contains cationic groups and the reaction in the column follows Eq. (30.9).

$$X^- + S^+Y^- \longrightarrow Y^- + S^+X^- \quad \text{(anion exchange)} \tag{30.9}$$

Conversely, to separate cations (cation exchange), the stationary contains anionic groups as shown in Eq. (30.10).

$$X^+ + S^-Y^+ \longrightarrow Y^+ + S^-X^+ \quad \text{(cation exchange)} \tag{30.10}$$

Sample ions are in competition with MP ions for sites on the ion-exchange resin that serves as the SP. Sample ions that do not bind as strongly to the resin as MP ions will elute early in the chromatogram, while sample ions that interact more strongly will be retained and elute later in the chromatogram.

A significant advance in the practice of IEC resulted from the work of Small, Stevens and Bauman (1975) who modified IE methods so that universal detection of ions at very low concentrations became possible. The method, called *IC*, uses a conductivity detector and a second column that permits the reduction of the background conductivity of the MP to reduce the magnitude of the baseline signal. The sample must be in aqueous form for introduction into the chromatograph. A typical sample is 50 μL of a 0.1–100-mg/L solution of the sample in a suitable solvent.

MPs commonly used in IC are dilute (approximately 10^{-3}–10^{-5} molar) aqueous solutions of ionizable materials such as sodium hydroxide, hydrochloric acid, mixtures of sodium carbonate and sodium bicarbonate, and mixtures of organic acids such as oxalic acid, citric acid, and tartaric acid. In RP IC, water-soluble organic solvents such as methanol and acetonitrile can be added to the MP. Quantities as small as 10^{-6} g can be analyzed.

30.4.2.7 **IC Instrumentation**

Measuring the small conductivity due to an eluting ion in the presence of the high background conductivity of the MP is difficult. Suppression of the conductivity of the MP, or eluent, dramatically improves the sensitivity (Small, Stevens and Bauman, 1975). Consider the example where a sodium hydroxide MP is used to separate anions. A low concentration of sample anions exits from the column containing the SP along with a high concentration of sodium and hydroxide ions. The column effluent is then passed through a second cation-exchange column (suppressor column), with hydrogen ions attached to the ion-exchange sites on the resin. Sodium ions displace hydrogen ions from these sites, with two beneficial results. First, water is formed by the reaction

$$\mathrm{Na^+OH^-} + S^-\mathrm{H^+} \longrightarrow \mathrm{H_2O} + S^-\mathrm{Na^+} \qquad (30.11)$$

Therefore, the effluent from the suppressor column has the very low conductivity of deionized water (as opposed to the high conductivity of a sodium hydroxide solution) and provides an excellent background for sensitive conductivity measurements on the anions. Second, the sample anions are paired with hydrogen ions instead of sodium ions as a consequence of the reaction

$$\mathrm{Na^+}X^- + S^-\mathrm{H^+} \longrightarrow \mathrm{H^+}X^- + S^-\mathrm{Na^+} \qquad (30.12)$$

As a result, detector sensitivity to the analyte anion is maximized because the equivalent conductivity of the hydrogen ion is approximately seven times as large as that of the sodium ion. The situation for cation analysis is completely analogous. Here, an anion-exchange suppressor column loaded with hydroxide ions is used to suppress the conductivity of the MP. The development of fiber suppressors and membrane-type suppressors simplified the procedures for this type of detection and provided improvements in resolution and sensitivity compared to the use of column suppressors.

Nonsuppressed conductivity uses the same conductivity detector just described, but without suppression of the MP conductivity. This technique for ion analysis is sometimes referred to as *single-column IC*. Low background conductivity is achieved by use of low concentrations (10^{-3}–10^{-5} molar) of MPs composed of ions that have a high affinity for the SP. UV/Vis absorbance detectors can be used to detect ions that absorb radiation in the 190–700-nm range. Alternatively, light-absorbing species produced by post-column reaction of a nonabsorbing ion with a complexing reagent can be detected. An example of this is absorbance detection at 520 nm of the complexes formed between 4-(2-pyridylazo)-resorcinol (PAR) and transition-metal ions such as Fe(II), Fe(III), Cu(II), Ni(II), Zn(II), Co(II), and Pb(II). Finally, absorbance detection can be accomplished using a UV/Vis-absorbing MP (e.g., phthalate buffers) to detect nonabsorbing ions. Here, eluting ions decrease the measured absorbance and are observed as negative peaks in the chromatogram. Limits of detection for electrochemical and UV/Vis detectors are approximately the same as for LC.

In Figure 30.12, the ion chromatogram of an aqueous mixture of 34 anions, both inorganic and organic, is shown. This chromatogram was obtained with suppressed conductivity using a ternary solvent system of sodium hydroxide.

In Figure 30.13, the separation of alkali-metal cations, as well as of the ammonium ion, is achieved using a cation-exchange resin. This chromatogram was obtained on a surface-sulfonated cation-exchange resin using dilute hydrochloric acid as the MP and eluent-suppressed conductivity detection. The observed retention behavior results from the increasing affinity of the cations for the SP as the ionic radius is increased.

30.4.2.8 **Size-Exclusion Chromatography (SEC)**

Size-exclusion chromatography (SEC) is used to separate molecules on the basis of their physical size. SEC with aqueous MPs is sometimes referred to as *gel filtration chromatography* (*GFC*), while SEC with nonaqueous MPs is sometimes called *gel permeation chromatography* (*GPC*). SPs are porous silica microspheres (e.g., 5-µm particles) or polymer

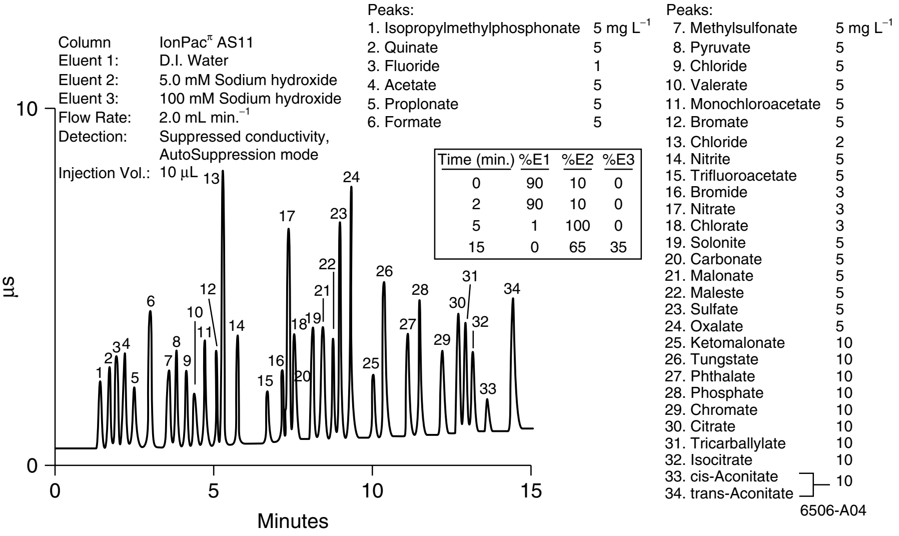

Column	IonPac$^{\pi}$ AS11
Eluent 1:	D.I. Water
Eluent 2:	5.0 mM Sodium hydroxide
Eluent 3:	100 mM Sodium hydroxide
Flow Rate:	2.0 mL min.$^{-1}$
Detection:	Suppressed conductivity, AutoSuppression mode
Injection Vol.:	10 µL

Peaks:

1. Isopropylmethylphosphonate	5 mg L^{-1}
2. Quinate	5
3. Fluoride	1
4. Acetate	5
5. Proplonate	5
6. Formate	5

Time (min.)	%E1	%E2	%E3
0	90	10	0
2	90	10	0
5	1	100	0
15	0	65	35

Peaks:

7. Methylsulfonate	5 mg L^{-1}
8. Pyruvate	5
9. Chloride	5
10. Valerate	5
11. Monochloroacetate	5
12. Bromate	5
13. Chloride	2
14. Nitrite	5
15. Trifluoroacetate	5
16. Bromide	3
17. Nitrate	3
18. Chlorate	3
19. Solonite	5
20. Carbonate	5
21. Malonate	5
22. Maleste	5
23. Sulfate	5
24. Oxalate	5
25. Ketomalonate	10
26. Tungstate	10
27. Phthalate	10
28. Phosphate	10
29. Chromate	10
30. Citrate	10
31. Tricarballylate	10
32. Isocitrate	10
33. cis-Aconitate 34. trans-Aconitate	10

Fig. 30.12 Ion-exchange separation of anions; reprinted with permission of Dionex Corporation, Sunnyvale, CA.

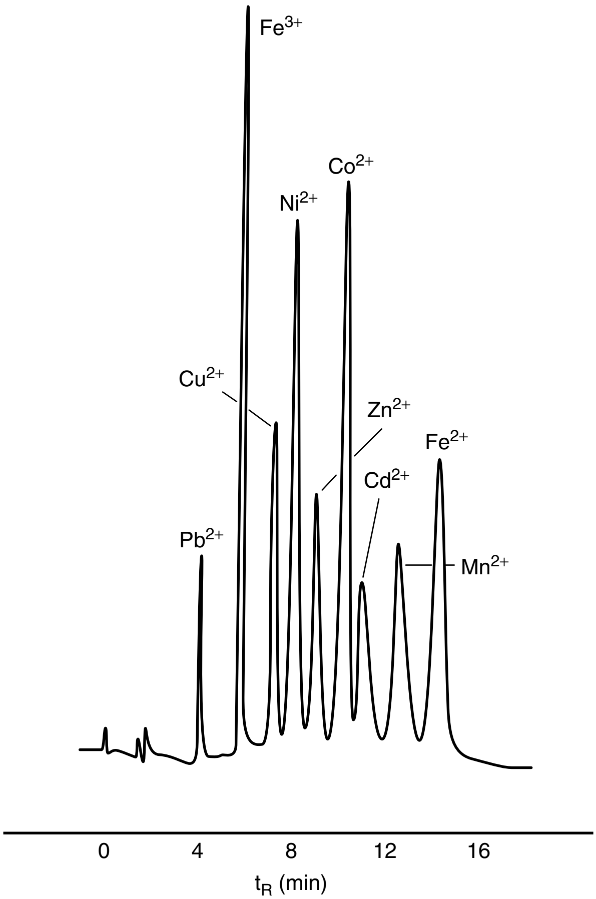

Fig. 30.13 Separation of nine transition-metal cations by ion chromatography; reprinted with permission of Dionex Corporation, Sunnyvale, CA.

beads that have porous microstructures with very well controlled pore size distributions. Typical pore sizes in these particles range from 6 to 100 nm. Molecules that are too large to permeate the pores in the particle travel around the particles and appear first in the chromatogram. Small molecules that can permeate the pores travel a long and tortuous path through the particles; therefore, they have longer retention times. MPs are chosen on the basis of their solvating power for the sample. In order to restrict the retention behavior to size-exclusion effects, it is essential that the components in the sample have a very high solubility in the MP. In the size-exclusion chromatogram shown in Figure 30.14, the components elute in the order of decreasing molecular weight (Yost, Ettre and Conlon, 1980). An important application of SEC is the analysis of molecular weight distributions of polymeric materials.

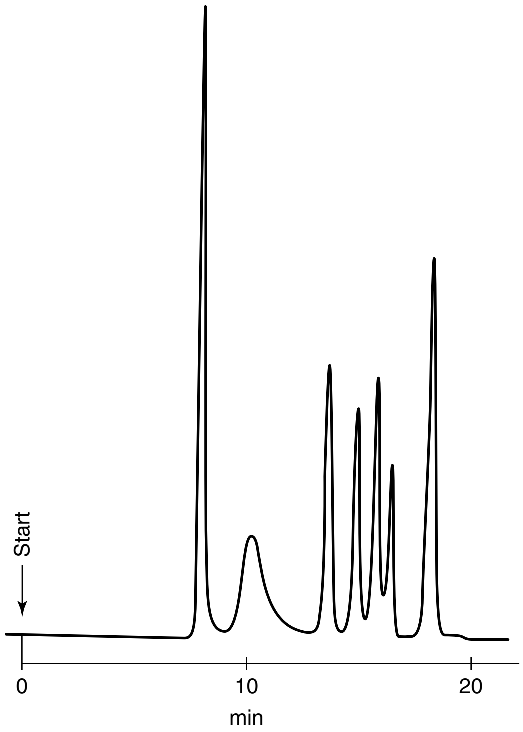

Fig. 30.14 Size-exclusion chromatogram of a mixture of organic molecules and polystyrene molecular weight standards. Two columns in series: 25-cm long × 8-mm i.d., divinylbenzene gel stationary phase. Mobile phase: tetrahydrofuran, 1 mL min^{-1}. Detector: UV absorbance at 260 nm. Peak assignment in order of elution: 1, polystyrene (molecular weight (MW) = 2, 400); 2, polystyrene (MW = 2, 100); 3, dioctyl phthalate (MW = 391); 4, dibutyl phthalate (MW = 278); 5, diethyl phthalate (MW = 222.2); 6, dimethyl phthalate (MW = 194); 7, benzene (MW = 78); (Yost *et al.*, 1980); reprinted with permission.

30.4.3
Preparative Liquid Chromatography

The objective in preparative LC is to isolate significant quantities of chemical compounds in a high-purity form. The amount of material can be very large (grams), as might be required for synthetic applications, or quite small (milligrams) for the purpose of chemical identification and analysis by other instrumental methods. For small-scale preparative LC, a standard liquid chromatograph can be used, with the addition of high-sample-capacity columns (diameters of 0.7–2.5 cm) if necessary. Large-scale industrial preparative liquid chromatographs have large MP reservoirs, large flow-rate capabilities, large-capacity columns (some as high as 60 cm in diameter), and automated fraction collection systems. In preparative LC, the capacity of the column is frequently exceeded, so that resolution is much poorer than would be realized in a normal analytical separation. A common strategy is to load as much sample onto the column as possible without degrading the resolution to the point where it is not possible to isolate a pure fraction. This strategy minimizes the number of repeated injections and fraction collections necessary to obtain the required quantity of purified material. Prep LC continues to develop and some recent papers can be found in the *Journal of Chromatography A* (2003, Vol. 1006. Seidel-Morgenstern, A.).

Preparative separations are especially useful in chiral separations whereby production quantities of chiral drugs can be produced. For more information about chiral separations, see Section 30.7.2.

30.4.4
Thin-Layer Chromatography

Most of the techniques shown in the chromatographic classification schemes in Figures 30.1 and 30.9 are carried out in columns. The exception is the one-labeled planar, consisting of two techniques, paper chromatography (PC) and TLC. PC was the older technique and it has been largely replaced by TLC. In classical TLC, the SP is dispersed on the surface of a planar surface like a square glass plate, and the samples are applied as spots on one edge of the plate using a pipet or an eye-dropper. The plate is then dipped in a solvent reservoir and MP-flow through the SP occurs as a result of capillary forces. The components in the sample are spatially separated on the plate and are detected by charring, chemical staining, UV/Vis reflectance, radiographic, or fluorescence techniques. The physical and chemical mechanisms for separation in TLC are the same as those in modern column LC, but there are some small differences due to the fact that the MP is moving on a dry bed and the flow is not controlled. The most common SP is silica gel, which means that the separation mode is most often NP. However, RP TLC plates are also available, and most of them are available in a small particle size, giving rise to higher performance, so-called high-performance thin-layer chromatography (HPTLC).

Instrumental and automated methods for TLC have been developed. New techniques for applying the sample to the plate have also dramatically improved the resolutions that can be achieved. Instrumental methods for recording chromatograms such as densitometry, UV/Vis absorbance, and fluorescence have reduced the time required for analysis, while providing detection limits as low as 10^{-12} g (using fluorescence detection). Most dramatic are instrumental methods for forced flow of the MP, which can reduce separation times and improve separation efficiencies, because smaller SP particles and larger plate sizes can be used. The most popular of these methods is called *overpressurized layer chromatography* (*OPLC*) the instrumentation for which has been available for about 25 years. Excellent details are given in the chapter on OPLC in Sherma and Fried's, 2003 *Handbook*, a useful reference for all of TLC. Instrumental methods are more popular in Europe than in the United States, where most TLC is still performed manually. The determination of acrylamide in drinking water by HPTLC combined with MS is a typical example of instrumental TLC (Alpmann and Morlock, 2008); the authors show that their results are comparable to those obtained by HPLC/MS.

30.5
Supercritical Fluid Chromatography (SFC)

SFC is not as popular as GC or HPLC. The name implies that the MP is a supercritical fluid and has properties between a gas and a liquid. The classification system shown earlier in Figure 30.1 appropriately shows SFC placed between GC and LC. It was introduced in the early 1960s by Klesper, Corwin and Turner (1962) at a time when HPLC was being developed and some chromatographers used sufficiently high pressures to create supercritical fluids. Advocates of SFC claim that it has the best attributes of GC and HPLC, and that it has the performance characteristics of a *unified* chromatographic system. Carbon dioxide is the most commonly used MP, and therefore SFC more closely resembles normal-phase HPLC.

In an age of concern about the environment, its use is highly desirable and has caused SFC to be labeled as a green process (Anon, 2007). This characteristic is especially important for preparative work.

Numerous separations and analyses of high-molecular-weight, thermally unstable, relatively nonvolatile compounds not suitable for analysis by GC were pursued during the remainder of the decade. However, with the advent of the more convenient method of HPLC in the 1970s, interest in SFC subsided rapidly. There was renewed interest with the introduction of the first commercial SFC instrument in 1982, and since then there have been technological advances in SFC instrumentation that have further popularized the technique. Applications that would be difficult or impossible to do by either GC or HPLC have been shown to be amenable to analysis by SFC (Griffiths, 1988). Examples include the separation of lipids and phospholipids, pharmaceutical products, as well as fuel oil distillates that can be separated into fractions corresponding to particular chemical families (e.g., paraffins, olefins, naphthelenes, and aromatic compounds). SFC can be used to analyze solids and liquids in amounts ranging from 10^{-2} to 10^{-12} g.

30.5.1
Operating Characteristics

A supercritical fluid is a substance that is above its critical temperature and pressure. By far the most common supercritical fluid used as a chromatographic MP is carbon dioxide, which has a critical point of 31 °C and 7.4 MPa. At the critical point, carbon dioxide has a density of 0.448 g/mL. Since carbon dioxide is relatively nonpolar, modifiers such as methanol, ethanol, acetonitrile, propylene carbonate, and water are sometimes added to increase MP polarity and therefore the solvating power for some samples. Other less commonly used MPs in SFC include ammonia, nitrous oxide, and sulfur dioxide.

SFC requires well-regulated control of both temperature and pressure. The high pressures are usually achieved by compression of a gas (e.g., carbon dioxide) with a high-pressure syringe pump. The condensed MP is then heated to supercritical temperatures. Some properties of supercritical fluids, such as solvating power, diffusion, and viscosity, are functions of density (Smith, Wright and Yonker, 1988). Density programming (also called *pressure programming*, since MP density is a function of pressure) is therefore used to improve resolution in SFC separations in an analogous way to that in which temperature programming and gradient elution are used in GC and LC, respectively. Pressure programming is achieved by microprocessor control of the pressure in the fluid delivery system.

SFC instrumentation can be modeled after either GC or LC instrumentation and both types are available commercially; the columns used can be packed or OT. Packed-column instruments resemble HPLC instruments and OT column instruments resemble GCs. In either case, a restrictor is required at or near the exit to keep the pressure high enough to keep the fluid above its critical pressure. The simplest restrictor is a 10-μm-i.d. capillary tube joined to the column outlet by a zero-dead-volume union fitting. Instruments using packed columns have become more popular, and, like HPLC, the most popular detector is UV. SFC is flexible: preparative separations are possible as well as hyphenated SFC/MS instrumentation.

30.5.2
Modes of Separation

30.5.2.1 Isobaric SFC

The simplest form of SFC is performed at constant pressure, isobaric. As in GC and LC, the selection of a MP/SP combination is based on the requirements of the separation. Among the more important factors are sample polarity and solubility of the sample in the supercritical fluid. In Figure 30.15, the supercritical fluid chromatogram of a mixture of hydroxybenzoic acids is shown (Berger and Deye, 1990). This chromatogram was obtained on a NP (cyano) LC column with a carbon-dioxide-based MP. Methanol and citric acid are added to the supercritical carbon dioxide to increase the solubility of the polar hydroxybenzoic acids in the MP. In the absence of these additives, the elution of the more polar, strongly retained components would occur with poorer resolution and only after much longer times.

30.5.2.2 Pressure Programming

Isobaric separations often suffer from the same problems that occur in isothermal GC and isocratic LC. The solution to these problems is pressure programming, where the density of the MP (and therefore the solubility of a given solute in the MP) is continuously increased during the separation by increasing the pressure. In Figure 30.16, three supercritical fluid chromatograms of a solution of the surfactant Triton X-114 are shown (Giorgetti *et al.*, 1989). These chromatograms were obtained on an RP (C18) LC column with a carbon-dioxide-based MP. The designator *n* in the figure denotes the number of monomer units in each polymeric surfactant molecule. Methanol was added to the supercritical carbon dioxide to increase the solubility of the surfactants in the MP. Comparison of Figure 30.16(a) and (c) clearly illustrates the beneficial effect that pressure programming can have on resolution. In addition, comparison of Figure 30.16(a) and (b) demonstrates the improvement in resolution that can be realized by increasing the methanol content in the MP in a linear fashion, much like gradient elution in LC.

30.6
Spectrochemical Detection

In prior sections, several spectroscopic detectors have been described, but one of them is of special importance: MS. The combination of a chromatographic instrument interfaced to MS is often referred to as a *hyphenated method* and designated as GC–MS or GC/MS, HPLC/MS, and so on. Multiple hyphenation, as in GC/MS/MS, is also common and sometimes called *hypernation*. These combinations can be considered to be multidimensional as well, and that topic is covered in Section 30.7.1.

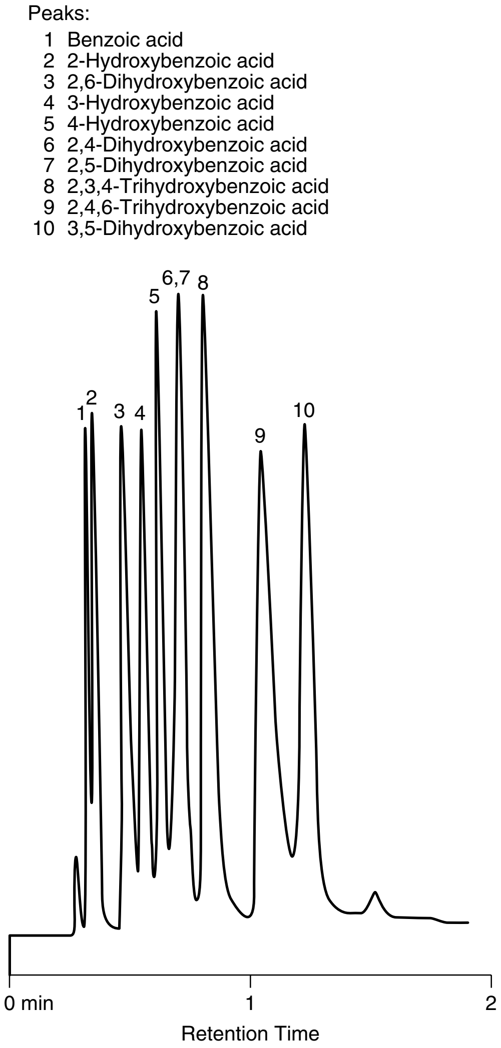

Fig. 30.15 Supercritical fluid chromatogram of a mixture of hydroxybenzoic acids. Column: 10-cm long × 2-mm i.d., normal phase, 7-µm particles, cyano. Column temperature: 60 °C. Mobile phase: carbon dioxide, containing 7.4% added methanol and 0.04% added citric acid. Pressure: 20 MPa. Flow rate 1.0 mL min^{-1}. Sample introduction: 0.2 µL on-column injection. Detector: UV/Vis absorbance at 230 and 430 nm. (Berger and Deye, 1990); reprinted with permission.

An alternative way of describing these combinations is simply to consider the mass spectrometer to be the detector attached to the chromatographic system. Thus, it is sometimes called the *MSD*, or the *mass selective detector*. Mass spectroscopists could also consider the chromatograph to be a special inlet system for MS. No significance is attached to either concept in this section, which is mainly concerned with GC/MS and LC/MS. Later, some discussion is included on the other hyphenated MS methods and other chromatographic combinations such as those with FTIR and NMR spectrometry.

With MS, it is possible to determine the molecular weight and molecular formula of an unknown compound.

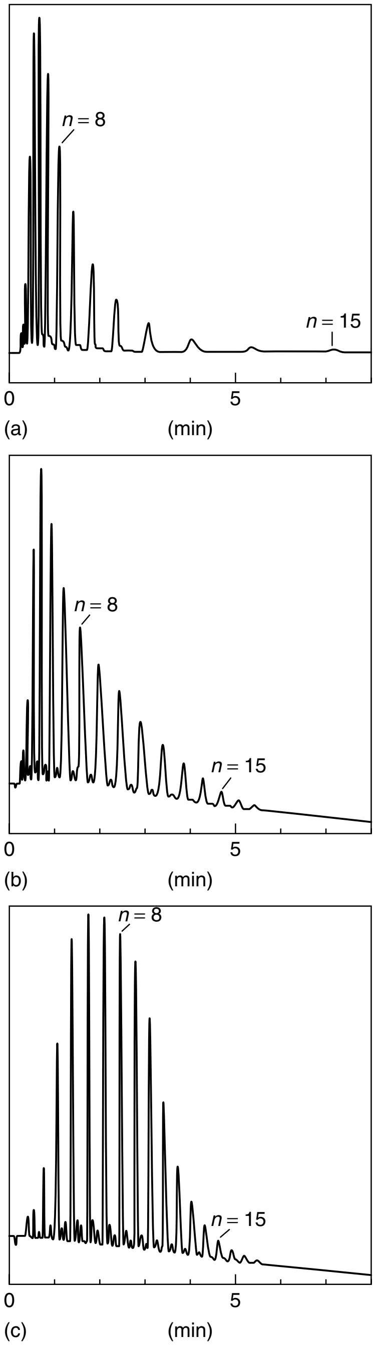

Fig. 30.16 Supercritical fluid chromatograms of the surfactant Triton X-114. Column: 10-cm long × 2-mm i.d. reversed-phase, 3-µm particles, C_{18}. Column temperature: 170 °C. Mobile phase: carbon dioxide with methanol modifier. Flow rate: 2 mL min^{-1}. Sample injection loop. Detector: UV absorbance at 278 nm. (a) Isobaric at 21.0 MPa, constant modifier flow (0.125 mL min^{-1}). (b) Isobaric at 21.0 MPa, modifier gradient (0.025–0.4 mL min^{-1} over 8 min, linear). (c) Pressure program (13.0–37.5 MPa in 8 min), constant modifier flow (0.125 mL min^{-1}) (Giorgetti *et al.*, 1989); reprinted with permission.

In addition, characteristic fragmentation patterns (produced by sample ionization) can be used to deduce molecular structure. Spectrometers in common use with chromatographic systems include quadrupole analyzers, quadrupole ion-trap mass spectrometers, and magnetic/electrostatic dual-focusing instruments. Time-of-flight (TOF) mass spectrometers are becoming increasingly popular. Typical detection limits of about 10^{-12} g can be realized with MS.

Data from a conventional mass spectrometer can be collected in two ways, total ion chromatogram (TIC) and single ion, called *single ion monitoring* (*SIM*). In TIC (also called *scan mode*), all ions detected are displayed as one signal, so the output looks just like a chromatogram from any other chromatographic detector. Any peak in this TIC can be examined to see what ions it contains, and this information is used primarily to identify unknowns as just described. By comparison, the SIM output is set to display only those ions (of a given mass-to-charge ratio) that have been chosen for collection. So, for example, if the mass of a particular analyte or analyte fragment is chosen as the single ion, the only peaks in the chromatogram will be for peaks that contain that ion. If a characteristic fragment ion is chosen, all analytes that fragment to give that ion, and only that ion, will appear in the chromatogram. Up to six ions can be acquired using conventional MS software. SIM is more selective and sensitive than TIC. Consider the difference between the acquisition of data over a range, which covers all possible ions (say 43–500), for a TIC, compared to one that collects data for only six ions, for a SIM. The former collects data for 76 times as many ions (457/6), so much more time is spent on each ion in SIM, increasing its sensitivity. Consequently, SIM is used primarily for quantitative analyses, often at trace levels.

Conventional chromatographic columns are normally operated so that the column exits are at atmospheric pressure, although it is possible to run a GC with a vacuum outlet. By comparison, mass spectrometers must be run under high vacuum. Thus, the process of joining the two involves an interface that can take a chromatographic effluent at atmospheric pressure and reduce it to a vacuum condition (about 10^{-6} torr). As will be seen, this process is different in GC and HPLC, but has been ideally achieved in both cases. It is much easier to couple a GC effluent to an MS than an HPLC effluent, which contains a liquid MP and often a buffer both of which have to be removed or enriched or exploited (solvent-assisted ionization).

Consequently, GC/MS was the first hyphenated technique to be successfully produced (1959) and it has become a relatively simple and cost-effective instrumental method. Bench-top instruments have become smaller and mass production has allowed the costs of instruments to decrease. Both parts of the instrument as well as all data are handled by computer, and these too have improved and become less expensive.

The actual identification of an unknown analyte can be accomplished by matching its mass spectrum from the TIC with reference spectra available in the MS software. The largest MS data base, containing over 130 000 spectra, is available from NIST.

30.6.1
GC/MS

The basic incompatibility between a gas chromatograph and a mass spectrometer is the pressure difference. The chromatograph separates components at column pressures of about 10^5 Pa with carrier gas flow rates of 0.1–100 ml min^{-1}, while the ion source of the mass spectrometer operates at pressures of about 10^{-4} Pa (10^{-6} torr). OT columns up to 0.32 mm i.d. are well suited for use in a GC coupled to an MS because of their low flow rates, typically 1–2 ml min^{-1}. The type of interface, called *direct*, is used with OT (capillary) columns. The end of the OT column is extended from the GC directly into the ion source of the MS. The GC flows are low enough, and the vacuum pumping high enough, that the necessary vacuum required by the mass spectrometer can be easily maintained. In this arrangement, the outlet of the GC column is operated at vacuum. This is not common, but it has been shown that vacuum operation can, in fact, be advantageous (Cramers, Scherpenzeel and Leclercq, 1981).

Further details on mass spectrometers can be found elsewhere in the Encyclopedia as well as in books by Oehme (1999) and McMaster (2008).

Today, GC/MS has become almost as routine as GC itself, and a GC/MS system is an essential part of most GC research labs. For the analysis of volatile samples, there is no better instrument.

30.6.2
LC/MS

MS can also be interfaced with a liquid chromatograph for perhaps the ultimate in identification capabilities. The past few years have seen major changes and improvements in LC/MS. Interfacing was initially more difficult, but a major innovation was the realization that interfaces could also serve as the ionization devices. This field is now in a growth spurt and has been adapted to handle biological samples that were previously considered impossible to be run by MS. Currently, LC/MS plays a major role in biotechnology due to its capability to perform rapid, high-resolution separations and identifications of large biomolecules. While LC/MS itself is not new, books covering the new advances and applications continue to be written (Ardrey, 2003; Niessen, 1999).

The key advances in instrumentation that made LC/MS possible were the development of narrow-bore (1-mm-i.d.) LC columns that can be operated at very low MP flow rates (as low as 0.1 mL min^{-1}), and interfaces for introducing liquids into the ion source of the mass spectrometer. In HPLC/MS, the ion sources also serve as the interfaces; thermospray ionization (TSI), electrospray ionization (ESI), and atmospheric pressure ionization (API) are the most popular. The latter two can be carried out at atmospheric pressure, a method of sampling that seemed impossible only a few years ago.

In adapting a HPLC method to LC/MS operation, several changes in MP composition need to be considered. These include

- increasing the volatility of the buffer;
- changing the pH;
- adding adducts, such as Na.$^+$

The MP needs to be volatile. Reversed-phase methods are usually ideal because the major components, water and methanol or ACN, meet this requirement. However, when buffers are used in RPLC, they also need to be volatile. Ammonium acetate is one example, and trifluoroacetate (TFA) can be used to form ion pairs with basic analytes, thereby making them neutral and decreasing their interference with MS detection.

30.6.3
Other Hyphenated Methods

Other spectroscopic instruments that have been interfaced to chromatographs are FTIR and NMR. Like the mass spectrum, an IR spectrum is a highly characteristic property of an organic compound. The complementary nature of IR and MS data makes the combination of them in series the most powerful method for analysis of unknowns by GC. In the case of GC/FTIR, large samples are desired because they facilitate detection by IR, but the vapor-state analytes are not easily sampled by IR. A light pipe has been designed that can be heated to keep the analytes in the vapor state and also meet the following requirements of a good IR cell for use with GC: small volume, long path length, and high transmission. A typical light pipe is 50 cm × 1 mm in size and has a reflecting gold coating. Its description and details of its use including the data handling requirements have been discussed (Griffiths, de Haseth and Azarraga, 1983). The other interface is a type of matrix isolation in which the analytes are frozen on a rotating band, which moves into the IR beam. A comparison of these two types has been published (Schreider, Demirian and Stickler, 1986).

The liquid from a HPLC is compatible with normal IR sampling and is less of a problem. However, LC MPs may not be transparent in the IR, and water is a particularly difficult solvent to handle. For volatile organic solvents, evaporation is possible, and the remaining nonvolatile analytes can be deposited into KBr and pressed into pellets. Further discussion can be found in reference (McDonald, 1986).

Mixed tandem combinations like GC/FTIR/MS are also possible since IR is nondestructive; it is possible to combine the three instruments into one. The special requirements and some applications have been described (Wilkins, 1983). Other tandem combinations are also possible. A common one is HPLC/UV/MS.

NMR does not have the low detectability limits of the other spectroscopies used in tandem with chromatography, so it has not been used as extensively. Also, the high magnetic field of current NMR instruments requires that the chromatograph be located at a considerable distance from the NMR probe. There are only a few reports of its use with GC, and, for HPLC, special glass flow cells are needed. For proton studies, the MP should ideally have no protons, so for RP work, deuterium oxide is used. Other deuterated solvents are quite expensive. LC/NMR is finding use in the pharmaceutical industry for impurity analysis (Lindon, Nicholson and Wilson, 2000). At least one commercial instrument offers LC/MS/NMR.

Other chromatographic methods that have been coupled with MS include TLC, SFC, capillary electrochromatography CEC) and capillary zone electrophoresis (CZE). SFC is handled much like GC, and CEC much like HPLC. TLC is quite different and requires special techniques. Details can be found in Poole's book (Poole, 2003).

30.7 Selected Other Topics

30.7.1 Multidimensional Separations

When a conventional one-dimensional separation is not adequate to resolve all the components of a sample, it may be possible to devise a multidimensional process that will yield better resolution. The concept is to spread out the separation in two dimensions rather than just one. The basic theory of two-dimensional separations was presented by Giddings, first in a paper (Giddings, 1984) and then in Cortes' book (Giddings, 1990) on the subject. Although the latter covers multiple dimensions, most of the actual applications are only two-dimensional. These two dimensions need to be orthogonal (at right angles) in order to maximize the two-dimensional space created. Indeed, Giddings' definition is based on that premise (Giddings, 1984). Orthogonal methods are those that use different mechanisms or modes of action in the separation. For multidimensional column methods, each column should provide a very different type of separation. For example, in GC/GC, one column could be very polar and the other nonpolar. The combination of LC and GC would be considered two-dimensional by virtue of the large difference in their separating mechanisms. A two-dimensional LC separation might combine RP in one column with SEC in the second, thereby attaining true orthogonality. A second requirement for true orthogonality is that components separated in the first dimension must not be recombined in the second dimension. The number of peaks that can be resolved in a two-dimensional separation is the product of the peak capacities of the two operations, thus showing the advantage of two-dimensional methods. As opposed to planar methods, in which the components are distributed in two-dimensional *space*, two-dimensional column separations give components distributed in two-dimensional *time*. Column separations have the added advantage of on-line detection.

A two-dimensional separation can be run off-line or on-line, the latter being preferred for convenience, ease of automation, and improved data handling. The disadvantage to the on-line mode is that the MPs may be incompatible, especially in LC, as in a combination of RP with NP HPLC; the former MP is usually aqueous and the latter nonaqueous. This problem seldom occurs in GC, where the MP can be helium for either polar or nonpolar column separations, and GC/GC is a very common two-dimensional method.

Optimization of the separation is usually called *tuning*, which can be achieved in several ways, including

- using one column of normal length and one of very short length;
- adjusting column variables such as temperature and flow rate;
- using innovative valving or modulation between the columns;
- using cryogenic trapping; and
- using original software to make optimal use of complex data such as that obtained by MS detection.

Descriptions of the different methods and combinations of multidimensional separations and some applications can be found in the volume edited by Mondello, Lewis and Bartle (2002).

30.7.1.1 GC/GC

Much of the early GC/GC work involved heart cutting, which refers to the process of isolating a poorly separated portion (a cut) of the molecules from the first dimensional process and running it on the second dimension. Heart cutting is not a true two-dimensional process since the whole sample is not run in the second dimension. It is still a valuable procedure and one that provided the first GC examples of improved resolution by the use of two dissimilar columns.

For a true two-dimensional GC separation, a cryotrap or modulator is usually required (Ledford, TerMaat and Billesbach, 2007; Hinshaw, 2004). A fraction from the first column is isolated with the modulator and transferred to the second column. Since only a single detector is used, the separation in the second column has to occur in less time than it takes to elute a single peak from the first column. Usually it is desirable to take three or four samples from each peak from the first column, so the run time for the second column can be only a few seconds at most. Since the second separation must be very fast, the second column is short and is operated at a high temperature and fast flow rate. One of the most impressive applications has been the analysis of cigarette smoke; in one case, over 30 000 peaks were isolated using GC/GC/TOFMS (Dalluge, Beens and Brinkman, 2003). In effect, the use of an MS detector to give a GC/GC/MS system results in a separation that is indeed three dimensional.

The interface/modulator between the two columns can take several forms; the two most common ones are a thermal sweeper and a cryogenic trap. At least two commercial instruments are available, including a retrofit for an existing GC and a complete GC/GC/TOFMS instrument. Papers from the First International Symposium on Comprehensive Multidimensional GC, which include reports of modulation devices and many applications have been published (Brinkman, Th., U.A. and Vreuls, R.J.J. 2003).

30.7.1.2 LC/GC

The combination of LC and GC is usually performed in the order indicated – LC being the first dimension. The interface between the systems must be capable of transferring the relatively large amount of MP from the LC into the GC. The popular transfer techniques are (i) retention gap, also called *on-column*; (ii) concurrent eluent evaporation, employing a loop type interface; and (iii) hot injectors, most often with programmed temperature vaporization (PTV). To divert the large volume of gas generated by vaporizing the LC solvent, types 1 and 2 are usually operated with a solvent vapor exit (SVE) placed between the precolumn and the analytical column. The relatively volatile solvents often used in NP LC can be handled, but the aqueous phases used in RP, which may also contain nonvolatile salts, are more difficult, and as a consequence there has not been much development of RPLC/GC.

30.7.1.3 LC/LC

A variety of two-dimensional LC methods have been described, many of which show similarities to the two-dimensional GC methods, including heart cutting. In designing an LC/LC procedure, however, one has to deal with the fact that the MPs differ between modes. In LC, unlike in GC, the MP, as well as the SP, participates in the process of defining the retention characteristics of the system. Thus, the combination of two LC columns of differing separating characteristics is likely to require the use of two different MPs. Consequently, a truly orthogonal LC/LC system is often limited by the necessity to have compatible MPs. For example, NP followed by RP might present difficulties with immiscibility of the NP fraction injected on the RP column. Some LC combinations that do work are

IEC and RPLC;
SEC and RPLC;
SEC and NPLC; and
combinations of RPLC with electrophoretic methods.

30.7.1.4 Other Combinations

These GC and LC examples of multidimensional chromatography serve as a fitting introduction to many other combinations. The major ones are with SFC and capillary electrophoresis (CE). Since SFC is intermediate in characteristics between GC and LC (*see* Section 30.5), and some consider it to be the ideal unified technique, it is a natural to combine with GC or LC or both. Also, the CE methods (CZE and CEC) are very similar to capillary RPLC and IEC and make ideal combinations with HPLC.

30.7.2
Chiral Separations

Chromatography is even capable of separating optical isomers, a process that is becoming increasingly important, especially in the pharmaceutical industry. For many drugs, only one optical isomer is pharmacologically active, and a total analysis that does not separate and quantitate the enantiomers is unsatisfactory. Enantiomers have identical physical properties and can rarely be separated by conventional chromatographic systems. Thus, new approaches to the separation of chiral compounds have been developed.

This separation challenge has been addressed by various methods, the oldest being the formation of diastereomers. A racemic mixture of optical isomers (designated *R* and *S*) cannot be separated chromatographically unless a chiral element is introduced into the procedure. One way is to react the mixture with a chiral reagent (designated, e.g., R'), to form reaction products that will have two chiral centers: one isomer will generate an RR' reaction product, and the other the SR' combination. These two diastereomers will have different physical properties and can be separated chromatographically with nonchiral phases. This approach was the major one used in GC when the field began to develop over 30 years ago. However, it has some disadvantages, and since not all analytes are amenable to analysis by GC, other approaches to chiral separations have been developed. These can be divided into two groups: those using chiral stationary phases (CSPs), and those using chiral mobile phases (CMPs). The use of CSPs is the most popular method for achieving chiral separations because chiral groups can be bonded to the surface of a stationary support. For GC, CSPs are the only option, since it has not been found possible to operate GC with CMPs. Considerable development work on CSPs has been done for HPLC phases. Pirkle (1986) has discussed the early work on CSPs for HPLC, giving examples and describing the way they work. Much less use has been made of CMPs in chiral HPLC separations, where it is assumed that the analytes form strong associations with a chiral component in the MP and that these complexes interact differently with the achiral SP. Although the mechanisms are not all clear, intermolecular associations that seem to form the basis of chiral separations are hydrogen bonding, metal chelation, ligand exchange, and ion-pair formation.

It should also be noted that SFC is of great utility in chiral separations, a major advantage being its relatively low pressure drop over the analytical column, which allows the use of multiple columns in series to achieve higher efficiencies. These SFC columns do not have to contain identical packing materials; use of different column trains can provide unique selectivities. In addition, SFC is ideal for preparative chiral separations since the MP is volatile, easily removed, and nontoxic. A single reference for more information about chiral separations is the book by Beesley and Scott (1998).

30.7.3
Capillary Electrochromatography (CEC)

Electrophoresis is not a chromatographic method, but a technique often used by chromatographers, and one that has been combined with HPLC to produce a technique called *CEC*. The common use of capillary columns brought electrophoresis and HPLC together. The first reported use of electrophoresis in OT capillary columns was published in 1967, but it was later, in 1979–1981, that a real impetus was given to the field by the use of fused-silica columns similar to those used in capillary GC. Later, when a pseudo-SP was introduced as a secondary phase, the technique took on some characteristics of LC and became known as *electrokinetic chromatography* (*EKC*) or, if the second phase was micellular, *micellar electrokinetic chromatography* (*MEKC*). Although scientists call MEKC a chromatographic process, others do not. Subsequently, however, the use of WCOT columns, and packed capillaries with SPs of the type used in RP HPLC, launched the operational mode now called *CEC*, a bona fide chromatographic method, one which has features of both RPLC and capillary zone electrophoresis, CZE (see Figure 30.1).

The CEC apparatus most closely resembles CZE, an illustration of which is shown in Figure 30.17. A packed column similar to one that could be used in HPLC, having a small (50–250 μm) i.d. is used along with a electrophoretic voltage gradient (up to 35 kV) and a flow electroosmotic flow (EOF) of the MP (or run buffer) caused by the resulting electroosmotic pressure. Since the highest electrophoretic mobilities are obtained with acetonitrile–aqueous buffer mixtures, this mixture has become the most popular; a good buffer concentration is about 1 mM. In the most common usage, the detector (usually a UV detector similar to one that would be used in HLC) is located near the cathode end and samples are introduced into the capillary at the anode end. Ionic analytes are driven by the electrostatic field, and, in addition, ions and molecules are transported by the electroosmotic flow, which performs the same function as the pump needed to move the solvent in HPLC. The SP in the column serves as a partitioning agent just as it does in HPLC, and the combination of these forces causes a separation of the analytes.

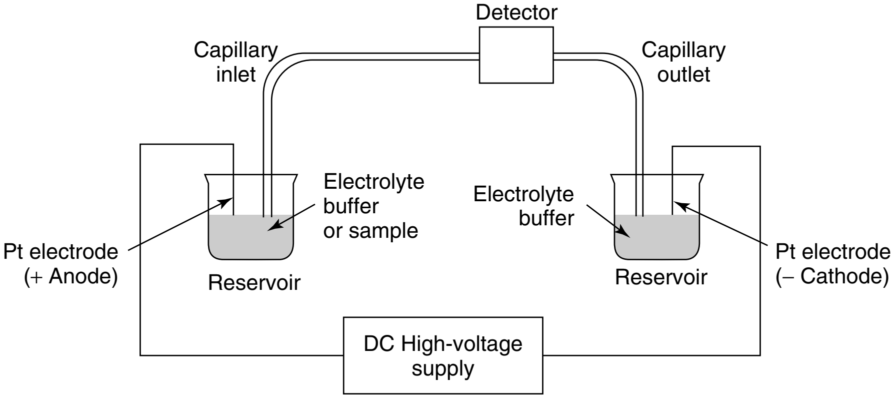

Fig. 30.17 Typical apparatus for capillary electrochromatography (CEC) and capillary electrophoresis. Reproduced with permission from Christian, G. D. *Analytical Chemistry*, 6th edition, John Wiley & Sons. Copyright 2004, John Wiley & Sons.

Tab. 30.4 Advantages and disadvantages of capillary electrochromatography (CEC). Reprinted from reference (McNair and Miller, 1998) with permission from John Wiley & Sons.

Advantages	**Disadvantages**
1. Greater efficiency	1. Poor detectivity
2. Wide range of applications	2. Poor quantitative precision
3. Faster	3. Poor migration time precision
4. No troublesome pumps	4. High-voltage hazard
5. Less expensive for chiral analyses	

The advantages and disadvantages of CEC listed in Table 30.4 are mainly comparisons with HPLC. The main advantage is the higher efficiency that results from the ideal flow pattern obtained in electrophoresis. The combination of HPLC and CZE provides CEC separations based on both effects for ionized materials and thus can effect separations not possible by HPLC alone. And, since the electrophoretic effects are predictable, the combined effects should be predictable if one has separate chromatographic data. Of the disadvantages, the poor detectivity and precision are sufficiently severe to discourage many chromatographers from the use of CEC, although CE is finding use for chiral separations.

The 1981 paper by Jorgenson and Lukacs (1981) is generally considered to mark the beginning of CEC, and the fundamentals have been summarized by Pyell (2001) who reported that efficiencies are about 10 times better for CEC than for HPLC. An entire volume of the *Journal of Chromatography A* is devoted to CEC (Horvath, 2000), and two books are available (Deyl and Svec, 2001; Krull *et al.*, 2000).

Acknowledgments

Portions of this chapter have been adapted by the author from other works that have been published previously by John Wiley. Some of the text has been updated and revised from the original edition written by Michael J. Kelly.

Glossary

Bonded-Phase Chromatography: Liquid chromatography performed with a stationary phase that is chemically bonded to a particulate supporting material such as silica powder. See also normal-phase chromatography, reversed-phase chromatography.

Capacity Factor: An old term that is no longer used. See Retention Factor.

Capillary Column: In gas or supercritical fluid chromatography, the small-diameter, open tubular column that contains the stationary phase. Inner wall of the tubing is coated with a solid or liquid film that functions as the stationary phase. The major type is a wall-coated open tubular (WCOT) column.

Carrier Gas: Mobile phase in gas chromatography.

Chromatogram: The output from a chromatographic separation that shows the response of the detector as a function of time.

Chromatograph: The instrumentation used in chromatography (noun). To separate and analyze components in a mixture by chromatography (verb).

Column: In chromatography, the glass, metal, or plastic tubing that contains the stationary phase.

Detector: In chromatography, the device used for dynamic monitoring of the chemical compounds leaving the column.

Gas Chromatography (GC): Chromatography with a gaseous mobile phase.

Gas–Liquid Chromatography (GLC): Gas chromatography with a liquid stationary phase that is dispersed on or bonded to a particulate packing material or to the inner wall of the column.

Gas–Solid Chromatography (GSC): Gas chromatography with a solid stationary phase.

Gel Filtration Chromatography: Size-exclusion chromatography with an aqueous mobile phase.

Gel Permeation Chromatography: Size-exclusion chromatography with a non-aqueous mobile phase.

Gradient Elution: In liquid chromatography, a technique for improving resolution, which consists of continuously changing the composition of the mobile phase during the separation.

Height Equivalent of a Theoretical Plate: Originally, a quantity used to describe the efficiency of a particular separation scheme, but no longer used. See plate height.

High-Performance Liquid Chromatography (HPLC): Modern liquid chromatography using mobile phases pumped at high pressures, stationary phases consisting of columns containing densely packed, small-diameter particles, and dynamic detection of the components eluting from the column.

Injector: In chromatography, the device used to introduce a sample onto the top of the column.

Ion Chromatography (IC): A group of liquid chromatography techniques used for the separation and analysis of ions.

Ion-Exchange Chromatography (IEC): Ion chromatography with an ion-exchange resin as the stationary phase.

Ion-Pair Chromatography: Ion chromatography with a reversed-phase or polymeric stationary phase, and a mobile phase containing chemical reagents that form neutral ion pairs with the sample ions.

Liquid Chromatography (LC): Chromatography with a liquid mobile phase. See also high-performance liquid chromatography.

Liquid–Solid Chromatography (LSC): Liquid chromatography with a solid stationary phase.

Mobile Phase (MP): In chromatography, the flowing gas, liquid, or supercritical fluid that carries the individual components of a mixture through the stationary phase.

Nonpolar: A property of a material resulting from the distribution of electronic charge in the individual molecules. A material is said to have nonpolar character if the electronic charge on each molecule is uniformly distributed through the molecule. See also polar.

Normal-Phase Chromatography (NPC): Liquid–solid chromatography or bonded-phase chromatography with a polar stationary phase and a nonpolar mobile phase.

Number of Theoretical Plates: A measure of the efficiency of a particular separation scheme, now called *plate number*.

Open Tubular Columns (OT): See capillary column.

Packed Column: A column that contains the stationary phase as a packed powder or resin.

Plate Height (H): A measure of column efficiency equal to the column length (L) divided by the plate number (N).

Plate Number (N): A measure of column efficiency that is based on a comparison of an analyte's retention volume and its peak width.

Polar: A property of a material resulting from the distribution of electronic charge in the individual molecules. A material is said to have polar character if any region of the individual molecules that compose the material possesses an excess or deficiency of electronic charge relative to the rest of the molecule. See also nonpolar.

Resolution (R_S): In chromatography, a quantity which describes the efficiency and degree of separation achieved in the chromatographic separation of two chemical compounds.

Retardation Factor for Column Chromatography (R): A measure of chromatographic retention behavior, equal to the average speed of a given analyte through the column divided by the average mobile phase velocity. Also equal to V_M/V_R.

Retardation Factor for Thin-Layer Chromatography (R_F): A measure of chromatographic retention behavior, equal to the distance traveled on a TLC plate by a given analyte divided by the distance traveled by the solvent (MP) front.

Retention Factor (k): A measure of chromatographic retention behavior equal to V'_R/V_M. Formerly called *capacity factor*, *capacity ratio*, and *partition ratio*, none of which should now be used.

Retention Time (t_R): In chromatography, the length of time a chemical compound spends in the column.

Reversed-Phase Chromatography (RPC): Bonded-phase chromatography using a nonpolar stationary phase and polar mobile phase.

Separation Factor (α): In chromatography, a measure of the ability of a separation scheme to separate two components. Also referred to as the *selectivity*.

Size-Exclusion Chromatography (SEC): Liquid chromatography technique that separates chemical compounds on the basis of their physical size. See also gel filtration chromatography, gel permeation chromatography.

Stationary Phase (SP): In chromatography, the solid or liquid phase that resides in the column. The stationary phase is a powder or resin packed into the column or a film deposited on or bonded to the inner wall of the column.

Supercritical Fluid: A material that is above its critical temperature and pressure.

Supercritical Fluid Chromatography (SFC): Chromatography with a supercritical fluid as the mobile phase.

Temperature Programming (PTGC): In gas chromatography, a technique for improving resolution, which consists of continuously changing the column temperature during the separation.

Thin-Layer Chromatography (TLC): Liquid chromatography with a stationary phase dispersed on a solid planar surface. Mobile-phase flow through the stationary phase occurs as a result of capillary forces.

References

Alpmann, A. and Morlock, G. (2008) *J. Sep. Sci.*, **31**, 71–77.

Anon (2007) *Chem. Eng. News*, **85**(43), 49.

Ardrey, B. (2003) *Liquid Chromatography-Mass Spectrometry: An Introduction*, John Wiley & Sons, Hoboken, NJ.

Beesley, T.E. and Scott, R.P.W. (1998) *Chiral Chromatography*, John Wiley & Sons, Chichester, England.

Berger, T.L. and Deye, J.F. (1990) *J. Chromatogr. Sci.*, **28**, 446.

Th. Brinkman, U.A. and Vreuls, R.J.J. (2003) *J. Chromatogr. A*, **1019**, 1–285.

Cramers, C.A., Scherpenzeel, G.J., and Leclercq, P.A. (1981) *J. Chromatogr.*, **203**, 207.

Dalluge, J., Beens, J., and Brinkman, U.A.Th. (2003) *J. Chromatogr. A*, **1000**, 69–108.

Dewaele, C. (1990) *J. Chromatogr. Sci.*, **28**, 649.

Deyl, Z. and Svec, F. (eds) (2001) *Capillary Electrochemistry*, Elsevier, Amsterdam.

(a) Ettre, L.S. (1993) *Pure Appl. Chem.*, **65**, 819–872; (b) Ettre, L.S. (1993) *LC-GC No. Am.*, **11**, 502.

Ettre, L.S. (2002) *LC-GC N. Am.*, **20**, 128, 452.

Ettre, L.S. (2003) *LC-GC N. Am.*, **21**, 458.

Ettre, L.S. and Zlatkis, A. (eds) (1979) *75 Years of Chromatography – A. Historical Dialog*, Vol. 17, J. Chromatogr., Library, Elsevier, Amsterdam.

Giddings, J.C. (1963) *Anal. Chem.*, **35**, 2215.

Giddings, J.C. (1984) Two-dimensional separations: concept and promise. *Anal. Chem.*, **56**, 1258A–1270A.

Giddings, J.C. (1990) in *Multidimensional Chromatography: Techniques and Applications* (J.H. Cortes), Marcel Dekker, New York, pp. 1–27.

Giorgetti, A., Pericles, N., Widmer, H.M., Anton, K., and Datwyler, P. (1989) *J. Chromatogr. Sci.*, **27**, 318–324.

Griffiths, P.R. (1988) *Anal. Chem.*, **60**, 593A–597A.

Griffiths, P.R., de Haseth, J.A., and Azarraga, L.V. (1983) *Anal. Chem.*, **55**, 1361A.

Grob, R.L. and Barry, E.F. (2004) *Modern Practice of Gas Chromatography*, 4th edn, Wiley Interscience, Hoboken, NJ.

Hinshaw, J.V. (2004) *LC-GC No. Am.*, **22**, 32–40.

HNU/Nordion Systems (1990) *J. Chromatogr. Sci.*, **28**, 389.

Horvath, C. (ed.) (2000) *J. Chromatogr. A*, **887**, 1–513. 38 papers on CEC.

James, A.T. and Martin, A.J.P. (1952) *Biochem. J.*, **50**, 679.

Jorgenson, J.W. and Lukacs, K.D. (1981) *J. Chromatogr.*, **218**, 209–216.

Klesper, E., Corwin, A.H., and Turner, D.A. (1962) *J. Org. Chem.*, **27**, 700.

Krull, I.S., Stevenson, R.L., Mistry, K., and Swartz, M.E. (2000) *Capillary Electrochromatography and Pressurized Flow Capillary Electrochromatography*, HNB Publishing, New York.

Ledford, E.B. Jr, TerMaat, J.R., and Billesbach, C.A. (2007) *Introduction to GC x GC*, Zoex Technical Note KT 039595-1, Zoex Corp., Lincoln, NE; available at: www.zoex.com/html/technote_kt030505-1.html.

Lindon, J.C., Nicholson, J.K., and Wilson, I.D. (2000) *J. Chromatogr. B*, **748**, 233.

Martin, A.J.P. and Synge, R.L.M. (1941) *Biochem. J.*, **35**, 1358.

McDonald, R.S. (1986) *Anal. Chem.*, **58**, 1906.

McMaster, M. (2008) *GC/MS: A Practical User's Guide*, 2nd edn, Wiley-VCH Verlag GmbH, Weinheim, Germany.

McNair, H.M. and Miller, J.M. (1998) *Basic Gas Chromatography*, Wiley Interscience, Hoboken, NJ.

Miller, J.M. (2005) *Chromatography: Concepts and Contrasts*, John Wiley & Sons, Hoboken, NJ.

Miller, J.M. and Crowther, J.B. (eds) (2000) *Analytical Chemistry in a GMP Environment*, Wiley Interscience, Hoboken, NJ.

Mondello, L., Lewis, A.C., and Bartle, B.D. (eds) (2002) *Multidimensional Chromatography*, John Wiley & Sons, Chichester, England.

Niessen, W.M.A. (1999) *Liquid Chromatography-Mass Spectrometry*, 2nd edn, Marcel Dekker, New York.

Oehme, M. (1999) *Practical Introduction to GC-MS Analysis with Quadrupoles*, John Wiley & Sons, New York.

Pirkle, W.H. (1986) in *Chromatography and Separation Chemistry* (ed. S. Ahuja), ACS Symposium Series 297, American Chemical Society, Washington, DC.

Poole, C.F. (2003) *The Essence of Chromatography*, Elsevier, Amsterdam, Netherlands.

Purnell, J.H. (1960) *J. Chem. Soc.*, **1960**, 1268.

Pyell, U. (2001) Fundamentals of capillary electrochromatography, in *Advances in Chromatography* (P.R. Brown and E. Grushka), Vol. 41, Chapter 1, Marcel Dekker, New York.

Quadrex Corp. (1990) *J. Chromatogr. Sci.*, **28**, 216.

Schreider, J.F., Demirian, J.C., and Stickler, J.C. (1986) *J. Chromatogr. Sci.*, **24**, 330.

Seidel-Morgenstern, A. (2003) *J. Chromatogr. A*, **1006**, 1–293.

Sherma, J. and Fried, B. (eds) (2003) *Handbook of Thin-Layer Chromatography*, 3th edn, Marcell Dekker, New York.

Small, H., Stevens, T.S. and Bauman, W.C. (1975) *Anal. Chem.*, **47**, 1801–1809.

Smith, R.D., Wright, B.W., and Yonker, C.R. (1988) *Anal. Chem.*, **60**, 1323A–1336A.

Snyder, L.R. and Kirkland, J.J. (1979) *Introduction to Modern Liquid Chromatography*, John Wiley & Sons, New York.

Wilkins, C.L. (1983) *Science*, **222**, 291.

Yost, R.W., Ettre, L.S., and Conlon, R.D. (1980) *Practical Liquid Chromatography: An Introduction*, The Perkin–Elmer Corporation, Norwalk, CT.

Further Reading

Cunico, R.L., Gooding, K.M., and Wehr, T. (1998) *Basic HPLC and CE of Biomolecules*, Bay Bioanalytical Laboratory Inc., Richmond, CA.

Dong, M.W. (2006) *Modern HPLC for Practicing Scientists*, John Wiley & Sons, Hoboken, NJ.

Eeltink, S., Desmet, G., Vivo-Truyols, G., Rozing, G.P., Schoenmakers, P.J. and Kok, W.Th. (2006) *J. Chromatogr. A*, **1104**, 256–262.

Grob, R.L. and Barry, E.F. (2004) *Modern Practice of Gas Chromatography*, 4th edn, Wiley Interscience, Hoboken, NJ.

Hahn-Deinstrop, E. (2007) *Applied Thin-Layer Chromatography: Best Practice and Avoidance* *of Mistakes*, 2nd edn, Wiley-VCH Verlag GmbH, Weinheim, Germany.

Hunt, B.J. and Holding, S.R. (eds) (1989) in *Size Exclusion Chromatography*, Chapman and Hall, New York.

Josic, D. and Clifton, J.G. (2007) *J. Chromatogr. A*, **1144**, 2–13.

Kromidas, S. (ed.) (2006) *HPLC Made to Measure*, Wiley-VCH Verlag GmbH, Weinheim, Germany.

Letter, W.S. (1991) *J. Chromatogr. Sci.*, **29**, 129.

McCalley, D.V. (2002) *J. Chromatogr. A*, **965**, 51–64.

McMaster, M. (2005) *LC/MS: A Practical User's Guide*, Wiley Interscience, Hoboken, NJ.

McMaster, M. (2006) *HPLC: A Practical User's Guide*, 2nd edn, Wiley-VCH Verlag GmbH, Weinheim, Germany.

McNair, H.M. and Miller, J.M. (1998) *Basic Gas Chromatography*, Wiley Interscience, Hoboken, NJ.

Meyer, V.R. (2004) *Practical High-Performance Liquid Chromatography*, Wiley-VCH Verlag GmbH, Weinheim, Germany.

Miller, J.M. (2005) *Chromatography: Concepts and Contrasts*, John Wiley & Sons, Hoboken, NJ.

Poole, C.F. (2003) *The Essence of Chromatography*, Elsevier, Amsterdam, Netherlands.

Rood, D. (2007) *The Troubleshooting and Maintenance Guide for Gas Chromatographers*, 4th edn, Wiley-VCH Verlag GmbH, Weinheim, Germany.

Schmidt-Traub, H. (ed.) (2005) *Preparative Chromatography: of Fine Chemicals and Pharmaceutical Agents*, Wiley-VCH Verlag GmbH, Weinheim, Germany.

Smith, J.H. and McNair, H.M. (2003) *J. Chromatogr. Sci.*, **41**, 209–214.

Snyder, L.R. and Kirkland, J.J. (1979) *Introduction to Modern Liquid Chromatography*, John Wiley & Sons, New York.

Weiss, J. and Weiss, T. (eds) (translator) (2005) *Handbook of Ion Chromatography*, 3rd edn, Two Volumes, Wiley-VCH Verlag GmbH, Weinheim, Germany.

CIRCUIT ELEMENTS

WILLIAM L. HOFFMAN, *Newark, Ohio, U.S.A.*

Introduction 455
1. **Conductors** 456
1.1 Introduction 456
1.2 Applications 456
1.2.1 Power Transmission 456
1.2.2 Signal Transmission 456
1.2.2.1 Coaxial Cable 456
1.2.2.2 Low-Frequency Conductors ... 456
1.2.2.3 Instrumentation 457
1.2.2.4 Digital Signal Conductors 457
1.3 Superconductors 457
2. **Resistors** 457
2.1 Introduction 457
2.2 Physical Parameters 457
2.3 Applications 458
2.3.1 Fixed Resistors 458
2.3.2 Variable Resistors 458
2.3.3 Chip Resistors 459
3. **Capacitors** 459
3.1 Introduction 459
3.2 Physical Parameters 459
3.3 Applications 459
3.3.1 *RC* Filter Circuits 459
3.3.2 *LC* Filter Circuits 461
3.3.3 Timing Circuits 461
3.3.4 Energy Storage 461
3.3.5 Power Factor Correction 461
4. **Inductors** 462
4.1 Introduction 462
4.2 Physical Parameters 462
4.3 Applications 462
4.3.1 Filter Circuits 462
4.3.2 Low Frequency 462
4.3.3 High Frequency 463
5. **Diodes** 464
5.1 Introduction 464
5.2 Physical Parameters 464
5.3 Applications 464
5.3.1 Rectifier Circuits 464
5.3.2 Overvoltage Protection 465
5.3.3 Zener Diodes 465
5.3.4 Varactor Diodes 467
5.3.5 Photodiodes 467
5.3.6 High-Frequency Diodes 467
5.3.7 Light-Emitting Diodes 468
6. **Transistors** 468
6.1 Introduction 468
6.2 Applications 469
6.2.1 Analog Applications 469
6.2.2 Switching Applications 469
7. **Transformers** 470
7.1 Introduction 470
7.2 Applications 470
7.2.1 Power Distribution 470
7.2.2 High Frequency 470
7.2.3 Circuit Isolation 471
7.2.4 Variable Transformers 471
7.2.5 Constant-Voltage or -Current Transformers 472
8. **Switches** 472
8.1 Introduction 472
8.2 Applications 472
8.2.1 Signal and Control 472
8.2.2 Low to Medium Power 473
8.2.3 Industrial Power 474
8.2.4 Power Distribution 474
9. **Fuses and Circuit Breakers** .. 475
9.1 Introduction 475
9.2 Applications 475
Glossary 476
Further Reading 476

INTRODUCTION

This article dwells on the common, discrete circuit elements, including semiconductors, found in electronic circuitry. Some attention is given electric power elements as well. The emphasis is placed on the basic applications of various circuit elements with a number of circuit diagrams to illustrate these. Integrated circuits are not discussed as such but in some applications their inclusion is a necessary part of a functional relationship, e.g., the operational amplifier.

The arrangement of the article is in nine sections each devoted to a particular circuit element. Each section is composed of a short

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

introduction describing the element followed by a discussion of various applications of the element as it is commonly, and sometimes not so commonly, used.

The writing is in terms that can be understood by a scientist or technician who has some basic knowledge of electricity but is not experienced in electronic or electrical circuits. A list of references for further reading is included at the end of the article which provide details of specialized topics and in some cases include more general tutorial chapters for electronics tyros.

1. CONDUCTORS

1.1 Introduction

The elements that are the best conductors of electricity at ordinary temperatures are the metals. The best conductor is silver, followed closely by copper, gold, and aluminum in that order. The most widely used material for conductors is copper because of its abundance and low cost relative to silver. Aluminum conductors are sometimes used to minimize weight and cost, but aluminum's ability to conduct electricity is only 60% of that of copper.

1.2 Applications

1.2.1 Power Transmission. Long-distance electric power transmission lines often span distances up to several hundred miles. Because the resistance of a conductor increases with length, it is important to minimize the conductor resistance when a large amount of power is to be transmitted over long distances. Otherwise an unduly large amount of energy loss will occur. Decreasing the resistance can be accomplished by using a large-diameter conductor. However, there are practical limits of excessive weight and cost resulting from too large a diameter.

The energy loss in a conductor is proportional to the square of the current. The power transmitted is proportional to the voltage–current product, VI. Consequently, for a given VI product, if the voltage is doubled and the current halved, the power can be transmitted at one-quarter the loss in the transmission line. The result is that long-distance power transmission lines deliver power at voltages up to a practical limit of a few hundred-thousand volts. This requires a current of only a few amperes to transmit power at megawatt rates on reasonably sized conductors.

1.2.2 Signal Transmission. Signal currents are encountered in communications and computer networks, and generally in all electronic systems. The currents are usually quite small. Consequently, conductors for signal currents are usually selected for properties other than low resistance.

Two important properties of signal conductors are characteristic impedance and shielding from noise. Conductors are usually not used for transmitting signals over distances longer than a few kilometers. For longer distances, microwave links or fiber-optic systems provide larger bandwidths and signal-to-noise ratios.

The main consideration for selecting conductors for signals is protecting the information content of the signal from corruption. Just which type of conductor should be used in a particular application depends upon the frequency spectrum of the signal, the electrical noise environment, and the power level of the signal.

1.2.2.1 Coaxial Cable. High-frequency signals, ranging from 1 MHz to 1 GHz, usually require coaxial cables. Such cables consist of a conductor surrounded by insulation and covered by a tubular conducting shield as shown in Fig. 1. The shield prevents energy from being lost through outgoing radiation, and noise pickup from external sources of radiation. The coaxial cable should be terminated in its characteristic impedance to prevent signal energy from being reflected back to the source.

1.2.2.2 Low-Frequency Conductors. For frequencies ranging up to the end of the audio spectrum (about 20 kHz), no special consideration generally needs to be given for the

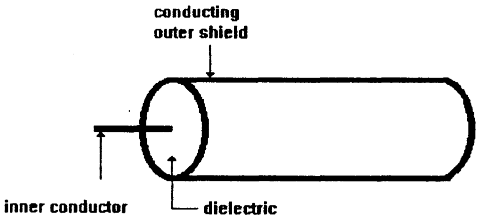

FIG. 1. Typical section of coaxial cable.

choice of conductors. However, when signal power levels are quite low, some shielding or twisted-pair configurations may be required to prevent noise pickup.

1.2.2.3 Instrumentation. Instrumentation and measurement systems often employ frequencies up to 100 kHz applied to balanced circuit transducers. Conductors employed with these circuits can cause serious unbalances and crosstalk errors because of their capacitance coupling to other conductors or system ground. These conductors should be shielded cables in matched pairs or terminated with adjustable capacitors to balance out their capacitance effects.

1.2.2.4 Digital Signal Conductors. For distances up to a few meters, digital signals can be transmitted in parallel mode over multiconductor cables without loss of information. Quite often these cables have their conductors molded into a flat ribbon. At the low impedances and relatively high power levels encountered in digital systems, noise pickup and capacitance coupling between conductors is not often significant as long as the frequency is kept low.

For distances up to a few kilometers, serial mode transmission of digital signals is more appropriate. Where bit transmission rates exceed 4000 bits per second, properly terminated coaxial cable is usually required to prevent loss of information. Bit rates less than 4000 can usually be safely handled by a twisted-pair cable configuration such as encountered in many telephone circuits.

1.3 Superconductors

Certain metals and alloys exhibit zero resistivity when their temperature is lowered to below 20 K. Conductors made from these materials are referred to as superconductors and indeed have zero resistance. Superconducting wires are very expensive compared to copper conductors, and so their use is limited to special applications where cost is a secondary consideration.

One application of superconducting wire is in the design of solenoids to produce very large magnetic fields. Such solenoids are only used in research laboratories and in special medical or industrial processes. See the articles SUPERCONDUCTING DEVICES and SUPERCONDUCTING MAGNETS for more details on superconductor applications.

In 1987, higher temperature (90 K) superconducting materials were discovered. Although these original materials were not suitable for most commercial uses, efforts are under way to develop high-temperature materials for commercial applications.

2. RESISTORS

2.1 Introduction

A resistor is a bilateral circuit element that obeys Ohm's law. Its purpose is either to limit the amount of current in a circuit for given applied voltage, or to produce a desired voltage for a given amount of current.

2.2 Physical Parameters

Resistors are configured in many sizes and shapes and usually employ carbon or metal as the active material. The physical size and materials of a resistor are determined by the amount of power it must dissipate without undue temperature rise. The electrical resistance is determined by the resistivity and geometry of the conducting material used. See the article RESISTORS for more information about the technology of resistor construction.

Resistance values may be indicated on the body of a resistor by alphanumeric coding, but most often by an industry standard color code given in Table 1.

Resistors are usually marked with colored bands. The color of the first band denotes the first significant figure of the resistor value, the color of the second band the second significant figure and the color of the third the power of ten by which the significant figures are multiplied. For example, the colors brown, red, and orange would denote a resistor value of 12 000 Ω. The industry also uses a set of standard values for resistors with a tolerance of 5% given in Table 2.

In addition to those with standard fixed values, resistors whose values are variable are

Table 1. Resistor color codes.

Black	0	Green	5
Brown	1	Blue	6
Red	2	Violet	7
Orange	3	Gray	8
Yellow	4	White	9

Table 2. Standard significant figures for 5% resistors.

10	18	33	56
11	20	36	62
12	22	39	68
13	24	43	75
15	27	47	82
16	30	51	91

also available. These are identified by various names such as potentiometers, photoresistors, thermistors and varistors.

2.3 Applications

2.3.1 Fixed Resistors. The principal application of resistors is in electronic circuits. In addition to their applications of current or voltage limiting, resistors are often used in conjunction with capacitors to produce time delays or to alter the frequency response of an electronic circuit as described in the section on capacitors. The termination of coaxial cables in their characteristic impedances is usually accomplished with an appropriately sized resistor.

2.3.2 Variable Resistors. Potentiometers are the most common form of variable resistors. An example of a potentiometer is shown schematically in Fig. 2. A potentiometer provides a way to divide a resistance of fixed value into two resistances in series whose ratio can be varied. The result is a variable voltage divider that finds wide application in electronic circuitry. For example, potentiometers are employed as volume controls or other adjustable controls in audio, radio, or video systems. Figure 3 is a typical example of a volume control circuit.

The potentiometer can be configured to provide a single variable resistance. This is accomplished by using the adjustable terminal and one fixed terminal as the terminals of a resistor leaving the other fixed terminal unconnected as shown in Fig. 4.

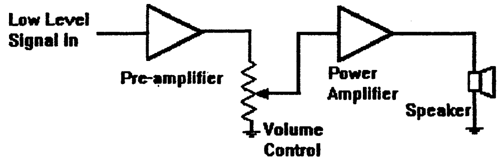

FIG. 3. Typical use of a potentiometer as a volume control.

Photoresistors are made from materials whose electrical resistance varies with the intensity of light incident on them (see PHOTODETECTOR DEVICES). These devices are used to produce voltage or current values that represent light intensity levels. Photoresistors are used to activate light-control circuitry and for photometric applications. Figures 5 and 6 are examples of photoresistor applications.

Thermistors are resistors whose resistance varies with temperature. They usually employ a semiconductor as the active material and therefore have a negative temperature coefficient, i.e., increasing resistance with decreasing temperature. Thermistors are used to produce circuit resistance that varies with temperature. An obvious thermistor application is in electronic thermometer circuits for the measurement and control of temperature.

Their negative temperature coefficient makes thermistors useful in compensating for the positive temperature coefficients of other resistors in a circuit. Figure 7 is an example of thermistor applications for temperature compensating fixed resistors.

Varistors are resistors whose resistances vary with the applied voltage. They are quite nonlinear. Characteristically, varistors have high resistance at low voltages and change rapidly to low resistance as the applied volt-

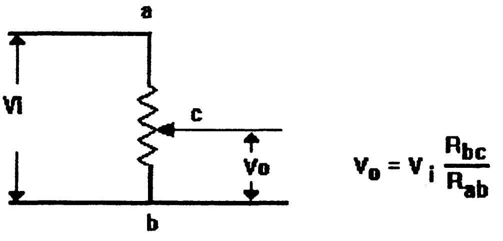

FIG. 2. A potentiometer used as a voltage divider.

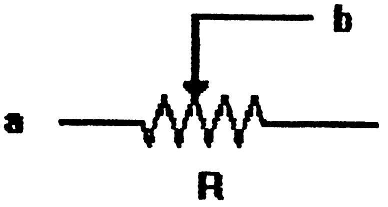

FIG. 4. Potentiometer used as a variable resistor. The resistance between *a* and *b* can be varied from 0 to *R* where *R* is the maximum resistance of the potentiometer.

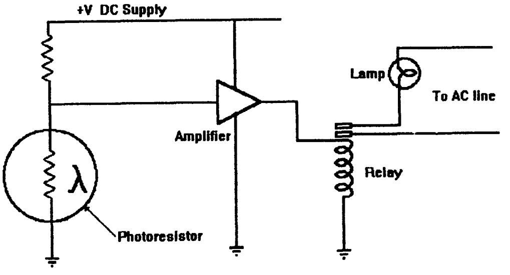

FIG. 5. Photoresistor on-off control. As the incident light on the photoresistor decreases the resistance increases until sufficient voltage is applied to the amplifier to energize the relay which turns the lamp on.

age increases beyond a certain threshold value. A widespread application of varistors is in overvoltage protection circuits for electrical equipment that is sensitive to power-line surges.

2.3.3 Chip Resistors. With the advent of surface mount technology in printed circuit-board electronics, a monolithic or chip resistor has been developed in recent years. A resistor of this type is formed into a rectangular parallelepiped with tinned surfaces for direct soldering to printed circuit-board conductors.

3. CAPACITORS

3.1 Introduction

A capacitor is a device that stores electrical charge. It is a two-terminal bilateral circuit element that obeys the relation $Q=CV$, where Q is the charge stored in the capacitor in coulombs, V is the voltage across its terminals in volts, and C is the capacitance in farads.

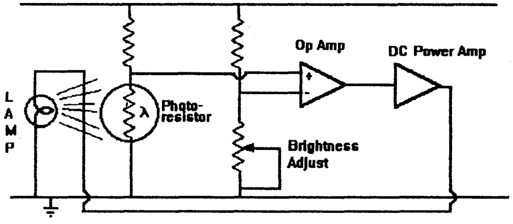

FIG. 6. Lamp intensity control circuit. If the lamp intensity changes, the photoresistance changes in such a way as to cause the amplifier to produce an opposite change in the lamp intensity.

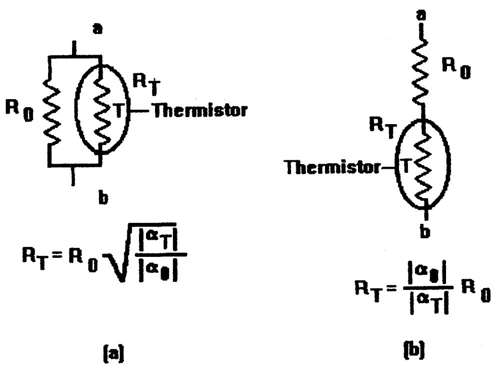
a
R_0
R_T
T
Thermistor
b
$R_T = R_0 \sqrt{\frac{|\alpha_T|}{|\alpha_0|}}$
(a)
a
R_0
R_T
Thermistor
T
b
$R_T = \frac{|\alpha_0|}{|\alpha_T|} R_0$
(b)

FIG. 7. Temperature compensation circuits. (a) A thermistor is used in parallel to compensate for thermal variation in R_0. (b) A thermistor is used in series for compensation. In both circuits the temperature coefficients α_T and α_0 must be of opposite sign.

3.2 Physical Parameters

Capacitors are manufactured in numerous shapes and sizes. Their basic construction consists of two parallel conducting surfaces insulated from each other by some dielectric material. Typical examples of dielectrics are air, paper, oil, and some form of plastic or metal oxide. Capacitance values range from picofarads to farads.

Most capacitors have a fixed value of capacitance. However, a few varieties of variable capacitances exist for use in circuits that require adjustment to a precise value of capacitance for best performance. Generally, the maximum capacitance values of variable capacitors range from a few to a few hundred picofarads.

3.3 Applications

3.3.1 *RC* Filter Circuits. Since the capacitive reactance varies inversely with the frequency, the zero-frequency, or dc, reactance of a capacitor is infinite. Conversely, the capacitive reactance at high frequencies can become very small compared to other elements in a circuit. Because of their frequency dependence, when capacitors are combined with resistors or other circuit elements, a class of circuits can be formed which are frequency selective, i.e., they may transmit or attenuate bands of frequencies. These circuits are generally referred to as filters.

Oscilloscope input circuits employ the infinite dc reactance of a capacitor to block dc

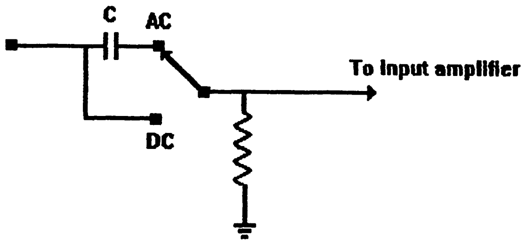

FIG. 8. Oscilloscope input circuit. When the switch is in the ac position a capacitor blocks dc voltage from the oscilloscope amplifier.

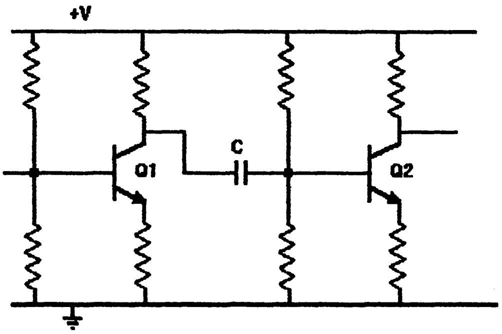

FIG. 9. Amplifier interstage coupling. The capacitor C prevents the dc output of first stage transistor Q_1 from reaching the input of the second stage transistor Q_2.

voltages that would preclude the effective display of small ac signals combined with the dc. Figure 8 is typical of an oscilloscope input circuit.

The dc blocking circuit is also used to interconnect stages of an amplifier to prevent the dc bias voltage of a preceding stage from upsetting the dc bias of a subsequent stage. An example of this circuitry is shown in Fig. 9.

The oscilloscope circuit is an example of an elementary filter circuit that is referred to as a high-pass filter. A high-pass filter generally transmits high frequencies while attenuating lower ones. If the roles of the capacitor and resistor are reversed, as shown in Fig. 10, a low-pass filter results, which passes dc and low frequencies while attenuating higher frequencies.

An altered version of the low-pass filter is shown in Fig. 11. Here, the resistor of Fig. 10 has been replaced by a rectifier diode. In this circuit the capacitor is charged to a dc voltage from current pulses passing through the diode. The capacitor discharges through the load connected across the capacitor. The result is a dc voltage across the load accompanied by a relatively much smaller ac voltage. This circuit is an elementary form of a rectifier and filter power supply circuit used to convert ac voltage into dc. See the section on diodes for more information about power supplies.

An example of a more elaborate filter circuit that attenuates a band of frequencies is the twin-T filter shown in Fig. 12. This is referred to as a band-stop or notch filter because of the shape of the response curve. In theory, zero transmission at the notch frequency is possible if the resistors and capacitors of the circuit are precisely matched. In practice, an attenuation of 60 dB is achievable.

A band-pass filter can be achieved if the twin-T filter is used as a feedback element in

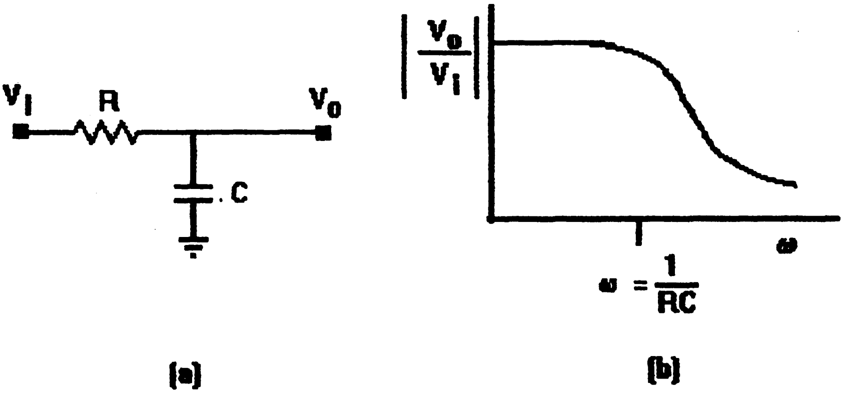

FIG. 10. Low pass filter. (a) A simple *RC* low-pass filter circuit. (b) Response of output voltage V_0 relative to the input as a function of the frequency.

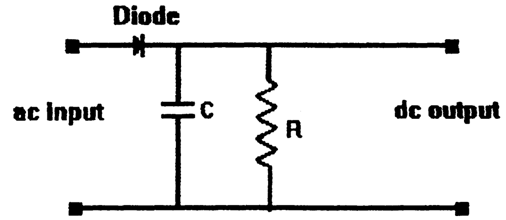

FIG. 11. Rectifier filter. The capacitor C and the diode form a low-pass filter to smooth out variations in the dc output produced by the ac. If the load resistor R is chosen so that the time constant *RC* is at least five times the period of the ac voltage, the dc variation will be approximately ±10% of the dc average. This assumes that no other load is connected in parallel with R.

an op amp circuit as shown in Fig. 13. In this case the high- and low-frequency parts of the response curve will be determined by the input resistor used. The sharpness of the band pass will also be affected by the value of this resistor.

3.3.2 *LC* Filter Circuits. Applications of capacitors in conjunction with inductors are discussed in Sec. 4.3.

3.3.3 Timing Circuits. The rate at which a capacitor charges or discharges through a resistor is determined in part by the product of the resistance and capacitance. This *RC* product is called the time constant of the circuit. Practical time constants can range from less than one microsecond to several minutes, depending on the values of resistance and capacitance chosen for the circuit.

Often the charging of a capacitor in an *RC* circuit is the basis for control applications that require on–off timing or timed intervals between events. One such application for the control of a photo enlarger lamp is described in Fig. 14.

Another important application of *RC* timing is in relaxation oscillator circuits. One such application is described in Fig. 15. Here, the period of the square wave generated by the oscillator is determined by the *RC* time constant which can easily be adjusted by varying *R*.

3.3.4 Energy Storage. As a capacitor is charged it accumulates energy proportional to the square of the charge accumulated. If the capacitor is charged slowly and then discharged quickly, the power obtained from the discharge can be many times the power of the charging source. This power amplification makes the capacitor useful in applications where large bursts of power are needed for short periods with relatively long intervals between bursts. Some typical applications are photo flash units, detonators, spot welders, and plasma pulse generators.

3.3.5 Power Factor Correction. Because most of the load on a power distribution system is large electric motors, the current and voltage on a power line may have a significant phase difference. This phase difference is measured by the power factor which is the cosine of the phase angle. A power factor significantly different from 1 results in excessive energy loss in the power system.

When capacitors are placed across the power line, they effectively tune out the inductive effects of the electric motor load, thereby reducing the phase difference to an acceptable level. These power-factor correction capaci-

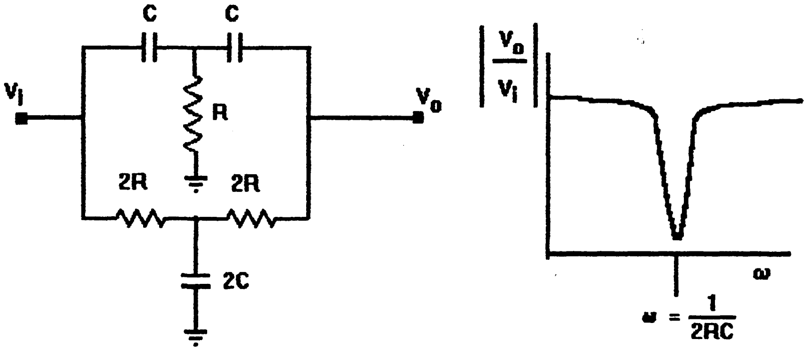

FIG. 12. Twin-T band-stop filter. When the capacitors and the resistors are carefully matched, the depth of the notch in the response curve can be as much as 60 dB.

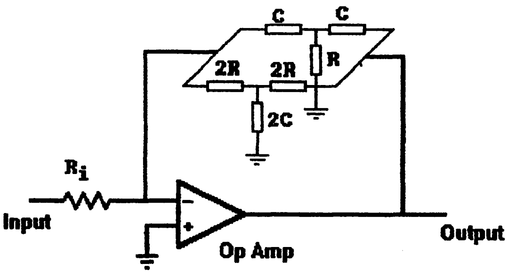

FIG. 13. Twin-T band-pass filter. When the circuit of Fig. 12 is used as the feedback network shown, a narrow band-pass filter is the result.

tors are usually very large compared to capacitors used in electronic circuits, since they must withstand thousands of volts across their terminals as well as severe weather and large variations in temperature.

4. INDUCTORS

4.1 Introduction

An inductor is a bilateral circuit element that obeys the circuit relation $V=L\,dI/dt$, where V is the voltage across the inductor terminals, dI/dt is the rate of change of current, and L is the inductance in henries. There is no voltage across the terminals of an ideal inductor if the current is not changing.

4.2 Physical Parameters

Inductors are usually formed by winding wire into a coil or coils. The inductor may be wound on a form containing ferromagnetic material to increase the value of inductance. Inductance values range from nanohenries to henries. Their physical sizes cover a wide range of volume and weight. For more information about their construction, see the article INDUCTORS.

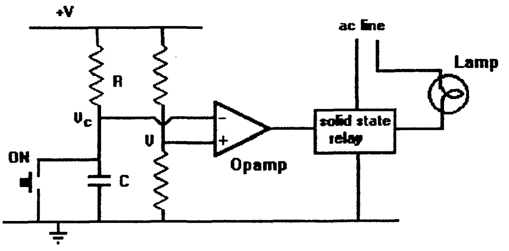

FIG. 14. Lamp timer circuit. When the on button is pushed, C discharges, which lowers V_c below V and turns the lamp on. When the button is released, C begins charging through R. When V_c exceeds V, the lamp turns off. The length of time the lamp is on depends on the time constant RC.

4.3 Applications

4.3.1 Filter Circuits. Since inductive reactance varies with frequency, inductors may be used in conjunction with resistors to form elementary filter circuits. This use of inductors is not common however, since resistor–capacitor filters are usually less expensive and bulky.

4.3.2 Low Frequency. Generally the physical size, weight, and cost of an inductance appropriate for low-frequency applications are very large compared to those characteristics required for high-frequency use. Consequently, circuit designers try to avoid using inductors for low-frequency applications wherever possible. However there are a few low-frequency applications that are commonly found.

An important low-frequency application of inductors is in the crossover networks used in stereo speaker systems. Here, inductors and capacitors form low-pass, band-pass, and high-pass filters that guide the audio signals to their respective speakers. Figure 16 is an example of a crossover network for a speaker system.

Another low-frequency application for inductors is the switching power supply for converting from one dc voltage to another. Here, the energy storage character of the

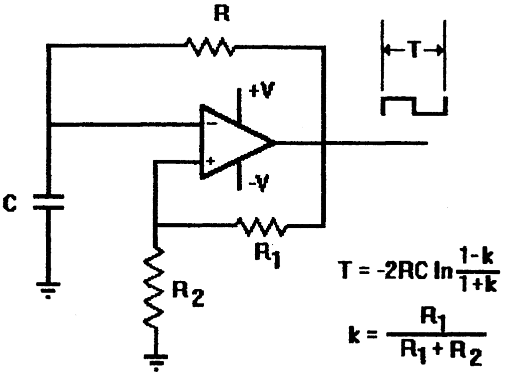

FIG. 15. Relaxation oscillator.The period of oscillation T depends on the RC time constant and the relative values of R_1 and R_2. When the resistors are all equal, the square-wave output is symmetrical.

inductor is used to provide for either a stepped-up or stepped-down voltage output from a primary dc source.

Elementary step-up and step-down circuits are shown in Fig. 17. Both of these circuits follow the same principle of operation. When the switch is closed, current builds up in the inductor. When the switch is opened, the inductor voltage adjusts to try to maintain the current flow and so maintains the capacitor voltage.

Usually the switches in the circuits of Fig. 17 are metal-oxide-semiconductor field-effect transistor (MOSFET) devices designed specifically for fast switching. The addition of sensing and processing circuitry allows for voltage regulation of these switching circuits as shown in Fig. 18. The output voltage of the switching regulator is controlled by controlling the duty cycle of the transistor switch.

Switching regulators offer improved efficiency over analog regulators that use pass transistors as the regulating elements. Switching frequencies may range from 20 kHz to 1 MHz or more.

4.3.3 High Frequency. When a capacitor is combined with an inductor, either in series or in parallel, a resonance phenomenon occurs. Over a narrow band of frequencies the impedances of the capacitor–inductor combinations vary drastically. For a capacitor and inductor in series, the impedance goes nearly to zero at the resonant frequency, while for a parallel combination the impedance becomes very large compared to the off-resonant values. These types of *LC* circuits are called resonant circuits and are most often used at high frequencies to produce very narrow-band filters. Figures 19 and 20 are examples of elementary resonant-circuit band-pass filters.

An elementary tuned amplifier that uses a resonant circuit is shown in Fig. 21. In practice, several of these amplifiers are cascaded and used as intermediate-frequency amplifiers in radio and television circuits. Usually each amplifier in the cascade is tuned to a slightly different frequency to amplify a band of frequencies sufficiently wide to pass the desired information.

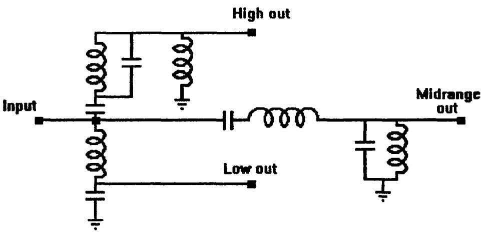

FIG. 16. Speaker system crossover network.

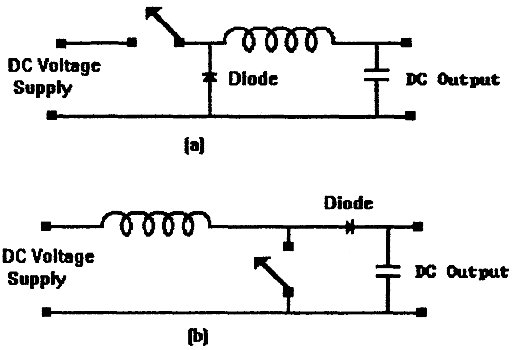

FIG. 17. Elementary switching power supply circuits. (a) This step-down circuit produces a voltage from zero to the supply voltage depending on the fraction of the time the switch is closed. (b) This step-up circuit produces an output voltage equal to or higher than the supply voltage again depending on the fraction of the time the switch is closed.

One inductor application that is used widely in radio-frequency circuits is the *LC* circuit shown in Fig. 22. The circuit is used to prevent high-frequency signals from reaching the power supply leads. The resonant frequency of the inductor–capacitor combination should be well below the effective range of the filter. Inductors used in this application are called radio-frequency chokes.

The electromagnetic interference (EMI) filter of Fig. 23 is an extension of the choke circuit. Here, inductor–capacitor combinations form a low-pass filter circuit, which is designed to prevent electromagnetic radiation

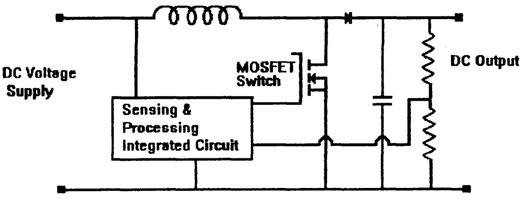

FIG. 18. Switching regulator circuit. The integrated circuit compares a fraction of the output voltage with an internal reference. If the output is too low the integrated circuit increases the "on" time of the MOSFET switch. If the output is too high the regulator decreases the "on" time.

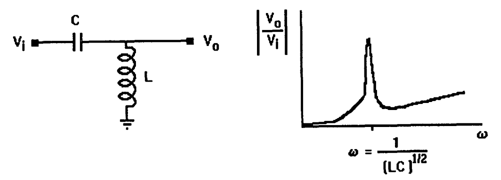

FIG. 19. Series resonant band-pass filter. This circuit has an input impedance of virtually zero at resonance and an output impedance that is very high.

from entering or leaving shielded enclosures containing high-frequency circuits.

Another form of an inductor can be made from a spiral trace etched in a printed circuit board. Such inductors range in value up to about 100 nanohenries and are useful at frequencies of several hundred megahertz. Also surface-mounting inductors for printed circuit-board applications are available. These inductors are contained in monolithic blocks that solder directly to a printed circuit board and range in value up to a few microhenries.

5. DIODES

5.1 Introduction

A diode is formed by a *p–n* semiconductor junction. It is a two-terminal unilateral device, allowing current to pass easily in one direction while virtually blocking current in the opposite direction. A typical diode volt–ampere curve and circuit symbol are shown in Fig. 24. For more details on the physics of diodes see the article SEMICONDUCTORS, ELEMENTAL—ELECTRONIC PROPERTIES.

5.2 Physical Parameters

The amount of current to be conducted, the reverse breakdown voltage, and the mechanical and thermal properties determine the physical size of a diode. Usual sizes range from cylinders 1 mm in diameter, 5 mm long to more than 3 cm in diameter, 7 cm long.

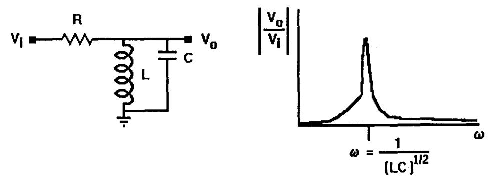

FIG. 20. Parallel resonant band–pass filter. This circuit has a very high input impedance at resonance and an output impedance that depends on the resistance of *R*.

5.3 Applications

5.3.1 Rectifier Circuits. The most universal application for diodes is the conversion of ac voltage into dc. In this application the diodes are called rectifiers. Circuits employing diodes for this purpose are called rectifier circuits.

Figure 25 is an example of an elementary, half-wave rectifier circuit. The output of this circuit is often called raw dc. For it to be useful in most applications the output must be filtered. An example of a capacitor filter is included in Fig. 25.

The half-wave rectifier circuit uses only half of the alternating voltage cycle to derive its output. A more efficient use of the ac source is the full-wave rectifier circuit, two versions of which are shown in Fig. 26. In this case the filtering needs only half as much capacitance as in the half-wave case because the fundamental ac component of the rectifier output is at twice the frequency.

Rectifier circuits usually are the primary source of dc power for electronic circuits. They are also used for less exacting applications such as battery charging and electroplating.

Automatic gain control circuits in radio and television receivers often employ the simple half-wave circuit of Fig. 25. In this application the voltage across the capacitor is a dc representation of the signal strength. This

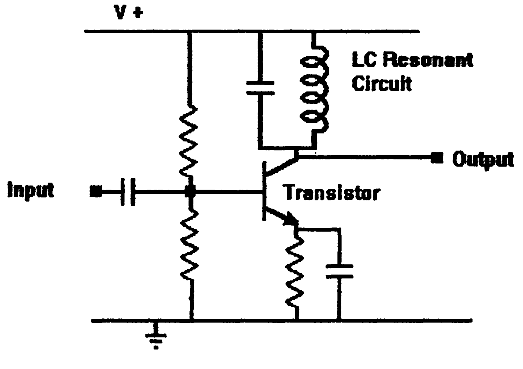

FIG. 21. Tuned amplifier. The action of the transistor supplies a current to the parallel resonant circuit proportional to the input voltage. When the input is at the resonant frequency of the *LC* combination, the output voltage becomes relatively large compared to nonresonant frequencies.

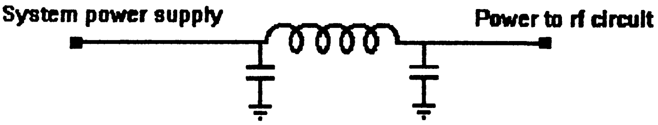

FIG. 22. Radio-frequency filter. High-frequency current is blocked from the power supply by the inductor. The current is shunted off to ground through the lower impedance paths of the capacitors.

voltage is used to control a gain determining element of the receiver to mitigate against fluctuations in the signal strength. Diodes for this application must operate at much higher frequencies than those used for power supply rectifiers; however, the power handling capability required of such diodes is insignificant compared with that of rectifiers. Consequently, these diodes are usually much smaller than rectifiers and have enhanced performance at high frequencies.

An interesting application of the ac-to-dc diode conversion is that of the clamp or dc restorer circuit shown in Fig. 27. This circuit adds a dc value to the ac output that is equal to the peak ac voltage less the forward voltage drop of the diode. The dc voltage added may be positive or negative depending on which direction the diode is installed. The circuit causes the output voltage to be clamped either above or below ground.

The dc restorer circuit suffers from error due to the diode forward voltage. For signals smaller than the forward voltage the circuit will not produce any significant dc restoration. This disadvantage is eliminated with the introduction of an operational amplifier, as shown in Fig. 28.

5.3.2 Overvoltage Protection. Circuits sometimes include diodes as protection devices to prevent over-voltage conditions that might cause damage or degradation of circuit performance. An example of such a circuit is shown in Fig. 29. Here, a relay coil is shown controlled by a transistor. When the transistor cuts off the current, the inductance of the relay coil could cause the collector voltage of the transistor to rise far above the supply voltage and permanently damage the transistor. The diode in the circuit prevents the collector voltage from rising more than a few tenths of a volt above the supply. At this voltage the diode begins to conduct, directing the inductive energy to the power supply without further voltage increase.

In automobile electrical systems, when current is cut off from the starter relay, the inductance of the relay coil can cause voltage spikes in the system that can damage the electronic control modules incorporated in modern vehicles. Diodes are used in a way similar to that of Fig. 29 to prevent damaging voltage spikes.

For logic gates, considerable damage can result from excessive input voltages. Figure 30 is an example of how diodes may be used to prevent excessive voltages from damaging gate inputs.

5.3.3 Zener Diodes. Normally, a diode reverse voltage that is sufficient to produce

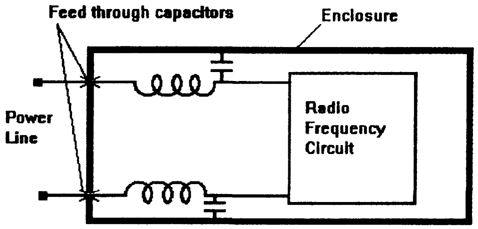

FIG. 23. Electromagnetic interference (EMI) filter. These circuits are identical to that of Fig. 22. They are used here to prevent high-frequency currents from escaping to the power line where they may cause interference in other devices attached to the line.

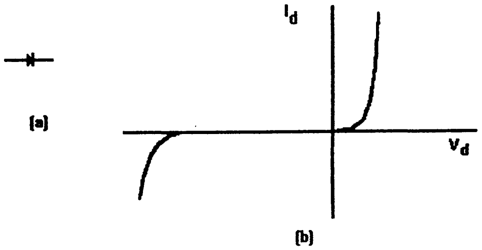

FIG. 24. The diode. (a) Circuit symbol. (b) Typical current–voltage characteristic curve.

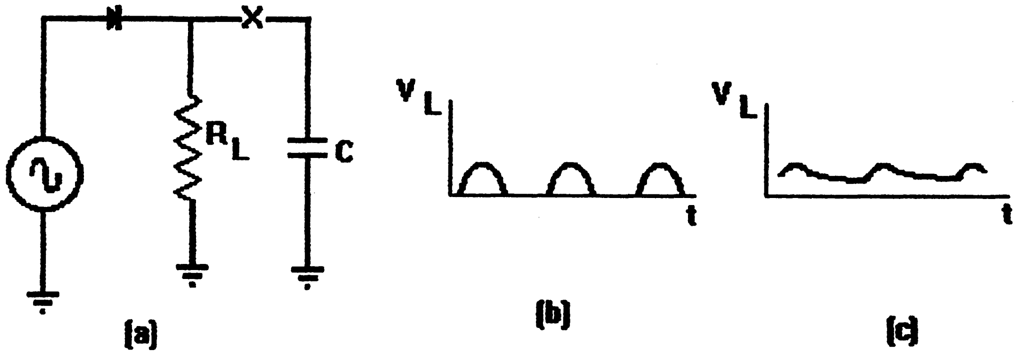

FIG. 25. (a) Half-wave rectifier circuit. (b) Variation in the voltage across the load resistor when *C* is not connected. (c) Load voltage variation when *C* is connected.

breakdown will permanently damage the diode. However, a class of diodes exist that are designed to accept reverse breakdown voltage and then recover to their original state when the voltage is removed. These diodes are called Zener diodes and the breakdown voltage is called the Zener voltage. Zener diodes are manufactured with a range of Zener voltages from 2.4 to 200 V.

One widely used application of Zener diodes is a voltage stabilizer or regulator circuit shown in Fig. 31. Here the output voltage will be at the Zener voltage of the diode as long as the input is above the Zener voltage.

The circuit of Fig. 31 has adequate regulation for some applications, but there will be some variation in output voltage as the load varies. If the current through the Zener diode can be held constant, the voltage will remain fixed except for some reaction to temperature variations. Figure 32 is an example of a regulated power supply that uses a Zener diode as a voltage reference which does not vary as the load varies. Zener diodes used for this purpose are usually temperature compensated units, which have a forward conducting diode incorporated in series with the reverse conducting Zener diode. Modern regulator inte-

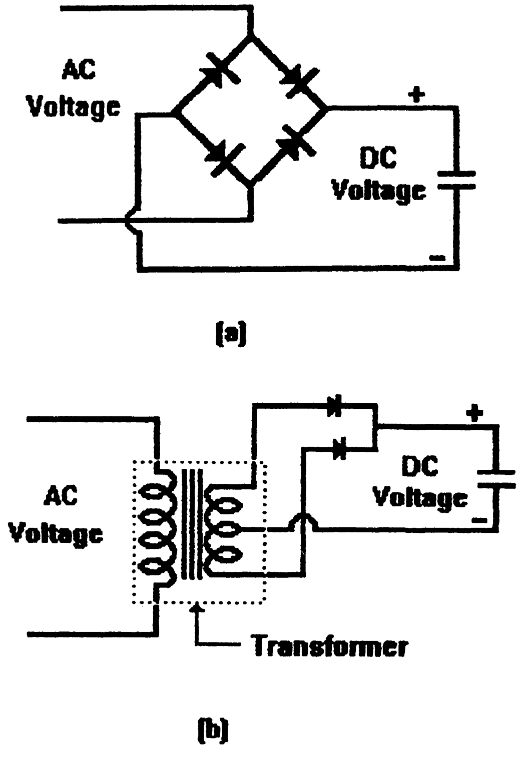

FIG. 26. Full-wave rectifiers. (a) Full-wave bridge circuit. (b) Center-tapped full-wave circuit.

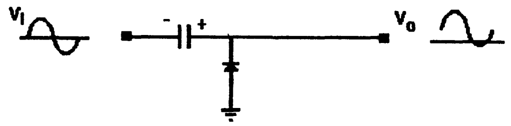

FIG. 27. dc restorer circuit. On the negative half of the input voltage the capacitor charges nearly to the peak value of the input with the polarity indicated. This capacitor voltage adds to the input voltage producing an output that is clamped above ground with the exception of a few tenths of a volt equal to the diode forward voltage.

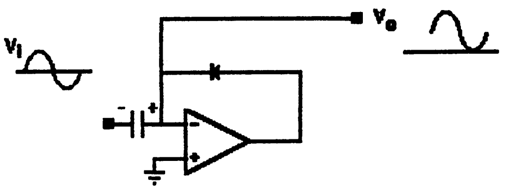

FIG. 28. Modified dc restorer. This circuit operates essentially the same as that of Fig. 27 except that the amplifier reduces the effect of the diode forward voltage to a millivolt or less.

grated circuits usually employ a Zener reference within the circuit package.

Zener diodes are often incorporated in amplifier inputs to prevent overvoltage damage from too large input signals. Figure 33 is an example of this protection application. The Zener voltage must be selected to be less than the minimum damaging voltage, but larger than the normally expected signal voltage.

5.3.4 Varactor Diode. Diodes behave as small capacitors when they are reverse biased. Often this is a detriment to circuit performance. However, since the size of this capacitance is determined by the amount of reverse voltage, specially designed diodes, called varactors, are employed as voltage controlled capacitors.

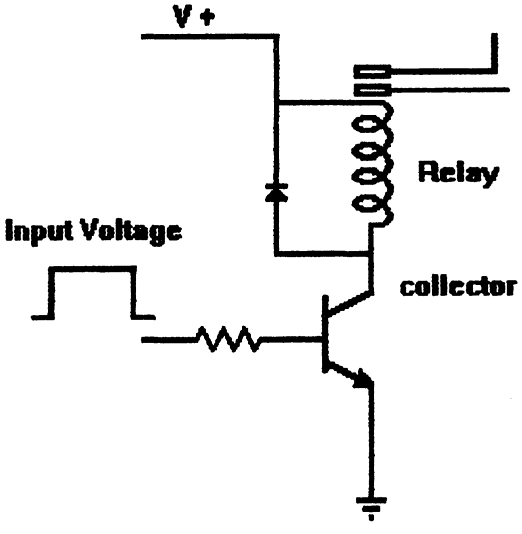

FIG. 29. Inductive spike protector. When the transistor is cut off at the falling edge of the input pulse, the inductance will cause the collector voltage to rise above V^+. The diode will then conduct and hold the voltage rise to less than 1 volt above V^+.

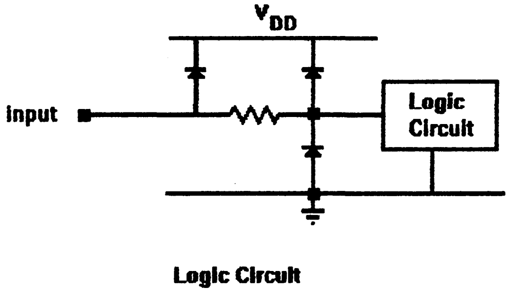

FIG. 30. Overvoltage protection. The upper diodes protect the logic circuit from positive-input overvoltages; the lower diodes protect against negative-input voltages.

Varactor diodes are the principal tuning element in most modern radio and television receivers. Figure 34 is an example of a voltage controlled oscillator (VCO). Such oscillators are used in conjunction with frequency control circuits to generate the tuning of most receivers and transmitters that operate at more than one frequency.

5.3.5 Photodiode. Photodiodes are used as light detectors, especially where nanosecond responses are required. They respond to light incident on a thin layer of intrinsic semiconductor sandwiched between *p*-type and *n*-type material. The light creates electron-hole pairs and consequently a diode current. This configuration is referred to as a *p-i-n* or PIN diode. The sensitivity of these diodes to incident light is usually less than one microampere per microwatt of incident light. See the articles on electronic properties of semiconductors for details concerning *p-i-n* structures. See also PHOTODETECTOR DEVICES.

5.3.6 High-Frequency Diodes. Gunn diodes are used as low-power oscillators to

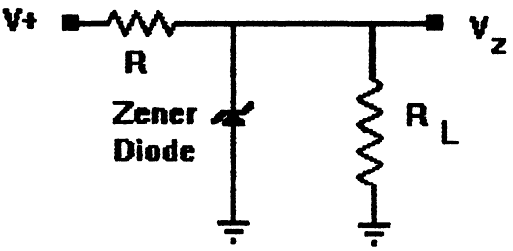

FIG. 31. Simple Zener regulator. The output voltage V_z will be equal to the Zener voltage as long as the input voltage minus the drop across R is greater than the Zener voltage.

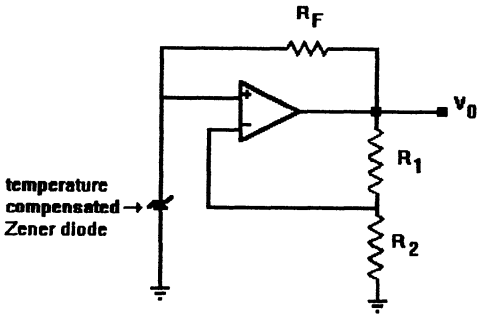

FIG. 32. Improved voltage regulator. The output voltage depends upon the ratio of R_1 and R_2. The resistor R_F is chosen to provide a Zener current that will produce a zero temperature coefficient for the temperature-compensated diode.

generate frequencies in the 5–100-GHz range. They utilize the many-valley scattering phenomenon that occurs in some III-V semiconductor compounds. See the article SEMICONDUCTORS, COMPOUND—ELECTRONIC PROPERTIES for details concerning III-V semiconductors.

Another class of PIN diodes, distinct from the photodiodes discussed above, serve as voltage-variable resistances and switches. Because of their nanosecond response times they are often used in voltage-controlled attenuators at UHF frequencies and as switches in microwave circuits. The amount of resistance that such a diode presents depends upon the amount of forward bias current passing through it. When the bias current is at the saturation point the diode has a low resistance to small high-frequency currents and acts as a closed switch.

Diodes called IMPATT diodes have on and off rise times of less than one picosecond and are used to generate frequencies in the 100-GHz range. The acronym IMPATT is derived from the "Impact Avalanche Transit Time" description of the diode operation. This diode exhibits a negative resistance region in its characteristic volt–ampere curve. When it is coupled to a cavity resonator and with the proper voltage applied, power at microwave frequencies is generated.

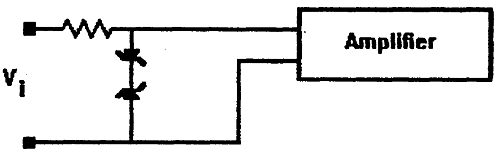

FIG. 33. Amplifier input protection. The "back-to-back" combination of Zener diode voltages is selected to begin conducting when the input voltage V_i becomes excessive for either polarity.

5.3.7 Light-Emitting Diodes. A specially designed diode called a light-emitting diode (LED) uses the recombination of electron-hole pairs in a forward-biased *p-n* junction as a source of light in both the visible and infrared parts of the spectrum. Light-emitting diodes have replaced small indicator lamps in many electronic applications and have formed the basic elements of the popular seven-segment display as well as other alphanumeric indicators.

With the advent of gallium arsenide semiconductors, laser diodes have been under rapid development for use in optical fiber communication circuits. Gallium arsenide lasers produce coherent light in the red and near-infrared regions. See LASERS, SOLID STATE for details concerning these devices.

6. TRANSISTORS

6.1 Introduction

A transistor is a three-terminal device in which the current between two of the terminals is controlled by a current or voltage applied to the third terminal. In some applications an approximately linear relationship exists between the control terminal state and the controlled current. In other cases the transistor may be used as a switch simply to provide on–off control of the current.

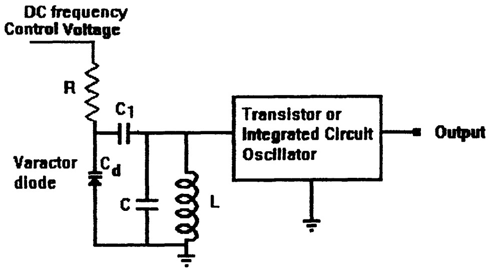
DC frequency Control Voltage
R
C_1
Varactor diode
C_d
C
L
Transistor or Integrated Circuit Oscillator
Output

FIG. 34. Oscillator frequency control. The varactor capacitance is coupled to the resonant *LC* circuit through C_1. The resonant frequency is altered by a change in the control voltage. The resistor R must be chosen large enough to have negligible effect on the resonant circuit.

There are two basic types of transistors, the junction or bipolar transistor and the field-effect transistor (FET). Many different configurations of the FET exist, but all operate by using a small voltage applied to one terminal to control current between the other two terminals, as shown in Fig. 35. The junction transistor exists in two different forms, *npn* and *pnp*. Both use current into one terminal to control the current between the other two, as shown in Fig. 36.

6.2 Applications

6.2.1 Analog Applications. The transistor is often described as a power amplifying device. In linear applications, a low-power signal is applied to the control terminal (gate or base), which causes significantly larger current to flow through the other two terminals of the transistor. When a resistive load is connected in series with this current, a reproduction of the signal voltage is produced across the resistance at a power level several orders of magnitude higher than that of the signal source. The transistor actually converts the power of the dc supply to an amplified reproduction of the power of the signal source and consequently serves as a power amplifier.

6.2.2 Switching Applications. Transistors are very useful switching devices both in low-power, high-speed applications found in digital circuits and in the higher-power cases such as switching power supplies.

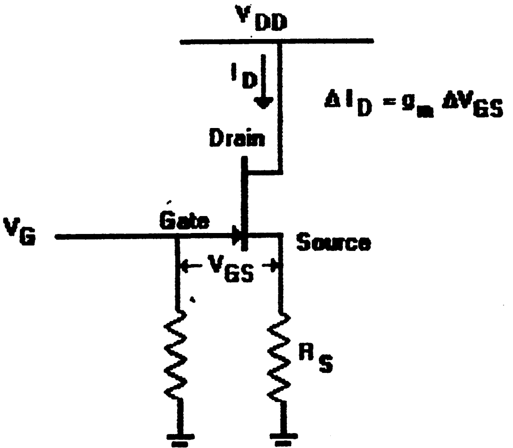

FIG. 35. A JFET transistor. The resistor R_s causes the gate to be negative relative to the source, which is necessary for the linear operation described by the equation for ΔI_D.

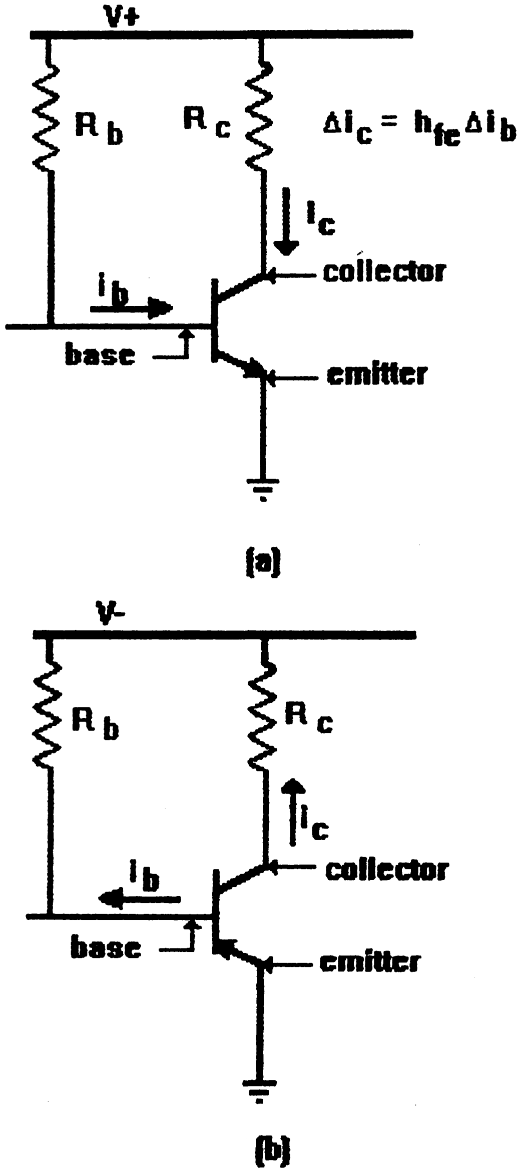

FIG. 36. Junction transistors. The equation for ΔI_C is applicable to both the *npn* (a) and the *pnp* (b) transistors.

The metal-oxide semiconductor FET or MOSFET has been the transistor design of choice among circuit designers for high-current switches. With current ratings up to 60 A, forward or on resistances as low as 0.028 Ω, and holdoff voltages of hundreds of volts, MOSFET transistors have been widely used as current switches in power supplies with ratings as high as 1000 W or more.

In recent years, most electronic designs include few transistors as discrete circuit elements. Instead transistors have become an important component of complex integrated circuits that perform many different, higher level analog and digital circuit functions. In exceptional cases where high voltage or power dissipation preclude the use of integrated circuits, discrete transistor devices still are incorporated in circuit designs.

7. TRANSFORMERS

7.1 Introduction

A transformer consists of two coils of wire that are magnetically coupled. If an ac voltage is applied to one of the coils, called the primary, an ac voltage will appear across the other coil, the secondary. Ideally, the ratio of the voltages will be the same as the ratio of the number of turns in their respective coils.

Usually transformer coils are wound around some sort of ferromagnetic core material to enhance the magnetic coupling between primary and secondary windings. For frequencies ranging up to a few kHz, this material consists of iron in the form of thin laminations bonded together. At higher frequencies, various kinds of ceramics containing ferromagnetic particles are used for core material.

The power input to the primary of a transformer will ideally equal the power absorbed by a load connected to the secondary. Realistically, most transformers are quite efficient with only a small amount of power loss in the transformer itself. Transformers are manufactured in a myriad of shapes and sizes depending upon the frequency and level of the power being transformed. Their weights range from a few grams to thousands of kilograms with corresponding volumes of a cubic centimeter to many cubic meters.

7.2 Applications

The fundamental purpose of a transformer is to change an existing ac voltage to a value more suitable for the application at hand, without incurring a significant amount of power loss. For example, a resistive voltage divider can be used to reduce a voltage for a particular purpose, but the power loss in the resistors usually far exceeds the loss from using a transformer to obtain the same reduction.

7.2.1 Power Distribution. Power generating stations use transformers to produce secondary voltages of several hundred kilovolts, which are suitable for efficient transmission of power over long distances. At the consuming end of the transmission line, transformers reduce the voltage to a more suitable value for the application of the power.

Power supplies for electronic equipment provide another example of the use of transformers to produce voltages more suitable for the applications. Most power for electronic equipment is obtained from the standard 120-V, 60-Hz lines in homes and commercial buildings. Most electronic devices require dc voltages of 30 V or less. The first step in converting the power is to use a transformer to reduce the ac voltage to a more suitable value for rectification to dc.

7.2.2 High Frequency. For frequencies of several kilohertz to hundreds of megahertz, transformers are used with a different point of view from simple voltage changing. Here, the concept of impedance matching is the dominant consideration in order to transfer a maximum amount of signal power.

High-frequency transformers may be designed to operate at a single frequency or to cover a broad frequency range, depending on the application. Single-frequency designs are found in high-frequency communications equipment, especially in the intermediate-frequency amplifiers of television and radio receivers. Usually, one or both of the windings of these transformers is tuned with a capacitor to provide frequency selectivity, i.e., to filter out unwanted frequencies.

The use of tuned transformers has diminished in recent years with the advent of surface acoustic wave (SAW) filters, which provide better selectivity and preclude the need for tuning adjustment or temperature compensation.

Broad-band high-frequency transformers are most commonly used for impedance matching between transmission-line circuits where signal frequencies may vary from 1 to 500 MHz. These transformers consist of a few

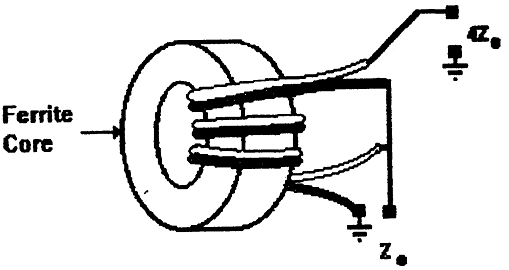

FIG. 37. Impedance matching transformer.

turns of paired wires wound around a torroidal ferrite core as shown in Fig. 37.

7.2.3 Circuit Isolation. Transformers are often used to isolate electrically a circuit connected to the secondary from the circuit connected to the primary. Since the transfer of power between the primary and secondary of a transformer is only through magnetic coupling, there is no direct electrical connection between the two windings. This isolation allows the primary and secondary circuits to have different levels of dc voltage and still be able to transfer ac power between them without disturbing their respective dc levels.

7.2.4 Variable Transformers. A special kind of transformer, called an autotransformer, has its secondary terminals connected directly to the primary. In most cases the secondary is comprised of some fraction of the primary turns that is accessed through a tap on the primary coil, as shown in Fig. 38. The secondary voltage is determined by the ratio of the number of turns between the tap and the common terminal to the total number of turns in the coil.

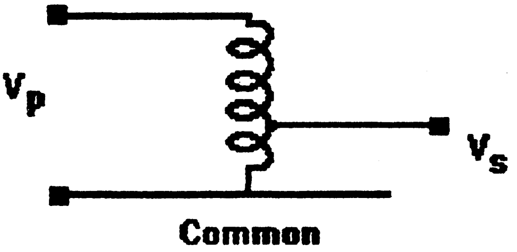

FIG. 38. The autotransformer uses a section of the primary winding as a secondary. There is no electrical isolation between primary and secondary.

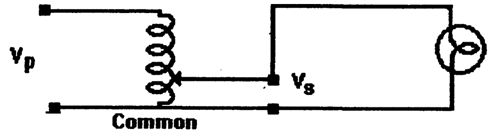

FIG. 39. Variable autotransformer. Moving the tap adjusts the secondary voltage to control the lamp intensity.

Often autotransformers will be wound in a configuration that allows the tap to be movable so the number of turns in the secondary can be varied. Thus an autotransformer can be used to control the brightness of lamps or the temperature of an electric heater as shown in Fig. 39. Autotransformers do not provide circuit isolation between primary and secondary. Consequently, their use should be limited to situations where danger of electrical shock or damage to equipment will not occur because of the lack of isolation.

Another type of variable transformer is the differential or linear transformer. This device consists of a single primary winding and two secondary coils wound on a rod-shaped core, as shown in Fig. 40. The core is movable, which allows the coupling between the primary and the secondaries to be varied. With the core centered, as shown in Fig. 40, the voltage in the two secondaries will be equal. When the core is moved to the right or left the two secondaries will have unequal voltages, the larger voltage appearing in the coil with the larger share of the core.

Differential transformers are useful for detecting small linear motions and are widely used in instrumentation systems for this purpose. Displacement ranging from 0.01 to 50 cm can be measured with accuracies of better than 0.1%.

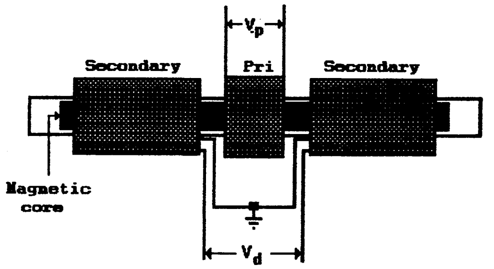

FIG. 40. Differential transformer. The two secondaries are differentially connected so that when the core is centered for symmetrical coupling, V_d is zero.

7.2.5 Constant-Voltage or -Current Transformers. A current-limiting transformer has been specially designed to operate gas-discharge devices, particularly the neon signs seen in nearly every city of the world. This transformer has a high-voltage secondary winding with a high resistance. Once the high voltage ionizes the gas, the current is limited by the high resistance of the secondary to a value that will maintain the discharge without overloading the transformer. Such transformers will withstand indefinitely a short circuit on the secondary without damage from overheating.

A cousin to the current-limiting transformer is the constant-voltage transformer. This device is designed to maintain a nearly constant secondary voltage when the primary voltage varies. Usually, this is accomplished by using a variable-reluctance core material, which is controlled to vary the coupling between primary and secondary to offset variations in the secondary voltage.

8. SWITCHES

8.1 Introduction

A switch is a device used to complete the circuit between an electrical load and its power source. Switches exist in a vast number of forms and sizes for a myriad of applications. In some cases, the load current passes through the switch contacts; in other cases, control of a load current in one switch may be done remotely by use of a second. The latter cases usually employ magnetically operated switches, called relays, or solid-state devices designed specifically for remote operation.

Applications for switches must take into account both electrical and mechanical requirements. The principal electrical requirements are the amount of current the switch contacts can carry without overheating, the amount of current the switch can interrupt for resistive and inductive loads without arc damage, and the amount of voltage the switch can withstand without insulation breakdown or arcing across contacts or terminals.

The mechanical requirements include the physical stress that must be withstood from the mounting and actuation method, the temperature and humidity extremes of the operating environment, and the number of operating cycles expected before failure.

8.2 Applications

The application areas can be divided into categories, namely, signal and control switching, low to medium power switching, heavy industrial switching, and power distribution switching.

8.2.1 Signal and Control. Switches for signal and control are subject to little electrical stress. Their main requirements are reliable operation and low noise generation. They may be either mechanical or solid state devices.

Mechanical switches are by far the most important for interfacing with human operation. Many reliable mechanisms have been designed for switches to be operated by pushing a button, turning a knob or flipping a lever. Some disadvantages exist for mechanical switches when they are used in wide-bandwidth or digital circuits, however.

One of the more important defects of mechanical switching is contact bounce. When their contacts close, almost all mechanical switches will alternate between an open and closed condition for as many as 100 times, for a period that may range up to a millisecond. Most digital circuits are capable of responding to this rapid opening and closing of contacts, which can result in serious errors when mechanical switches are used to interface with computers.

Contact bounce can be eliminated either by software methods, in the case of computer applications, or by electronic "debouncing" circuits, such as that shown in Fig. 41.

Another disadvantage of mechanical switches is their effect on high-frequency or low-level signals. The capacitance of the switch and its attendant wiring may seriously affect the high-frequency performance of a circuit. For low-level signals, the switch and its wiring may introduce noise.

In these applications the mechanical switch should be used to control a more suitable switching device that minimizes the capacitive and noise effects. An example of a combination of mechanical and solid-state switching to minimize noise in a low-level audio circuit is shown in Fig. 42. In a similar manner a mechanical switch can be used to control a relay specially designed to switch radio frequency signals.

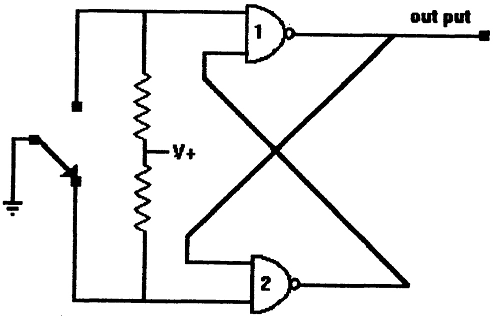

FIG. 41. Debounce circuit. The output is the logical NAND function of the inputs to gate 1. When the switch is in the lower position the output will be zero volts, a logical zero. When the switch is in the upper position the output will be approximately +V, a logical 1. These outputs will occur on initial contact of the switch terminals and remain fixed during contact bounce.

An interesting example of a switch designed specifically for interfacing with microprocessor circuits is the so-called sine–cosine switch shown in Fig. 43. This switch is capable of continuous rotation in either direction. The two rotor poles are identical, except that one pole is set 90 degrees around the shaft from the other. This results in the contacts emitting a different sequence of two bit binary numbers for clockwise rotation than for counterclockwise rotation. The software for the microprocessor must be designed to detect the rotation direction. The result then can be used to generate control signals for frequency, amplitude, position, etc.

8.2.2 Low to Medium Power. Applications for low- to medium-power switching are widespread among automobiles, households, offices, and commercial establishments. Mostly, such applications are for lighting and small electric motor control. In other than automobiles, the switches used usually carry the full load current of the devices that they control. Electrical specifications for such switches rarely exceed 15 A and 230 V. In the case of electric appliances, such as refrigerators, washers, dryers, and air conditioners, relays are usually employed to control the electric motors.

For automobiles, most of the dashboard switches are of the control type. These are used in conjunction with relays to provide high current capacity without requiring high-current conductors routed under the dash.

With the advent of solid-state switching devices, two of the most important in power control applications have been the silicon controlled rectifier (SCR) and the triac. The circuit symbols for these devices are shown in Fig. 44.

The SCR is a unilateral device like the diode with one important exception: The SCR will not conduct unless it is triggered on by a current from gate to cathode. Once triggered the SCR behaves as a diode rectifier, regardless of the presence or absence of gate current. When the anode voltage becomes equal to or negative, with respect to the cathode, the SCR stops conducting and will remain in a nonconducting state until triggered again by gate current.

The gate control property of the SCR allows it to be turned on any time during the portion of the ac cycle when the anode is positive with respect to the cathode. Consequently, the device can be used as a rectifier that will produce a variable amount of average dc current depending on when it is triggered during the positive half of the ac cycle. One application for the SCR is its use as a speed control for dc motors, as shown in Fig. 45.

The triac is similar to the SCR in that it is triggered on by a current through the gate-to-MT2 junction. The device differs from the SCR in that it is bilateral, i.e., it can conduct in either direction. Once the triac is turned on, the gate terminal no longer controls it until the voltage between MT1 and MT2 goes to

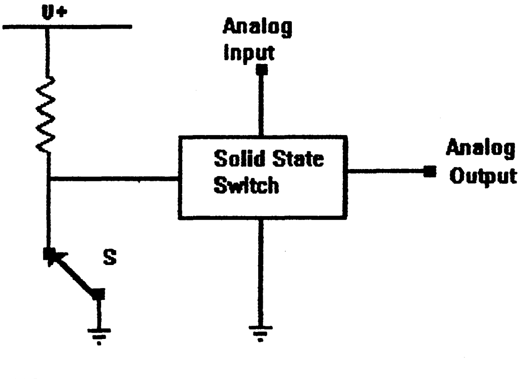

FIG. 42. The solid-state switch is an integrated circuit incorporating a MOSFET transistor in the analog signal path. These switches are turned on and off with digital signals such as that generated by the mechanical switch S.

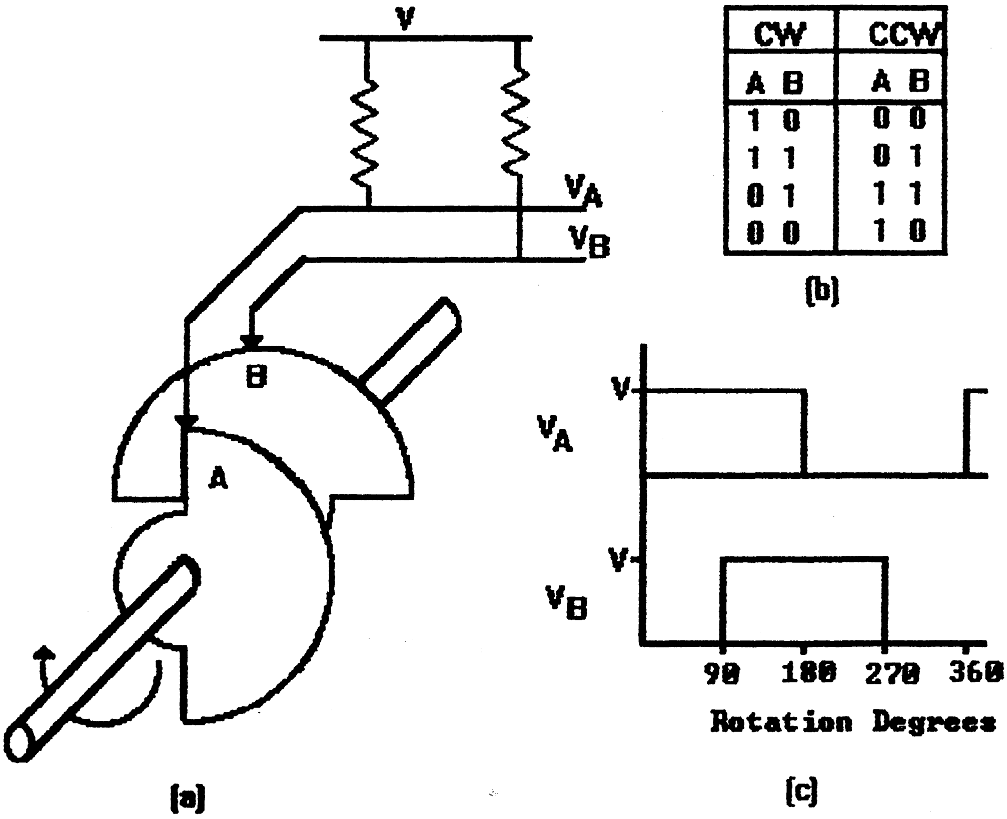

CW		CCW	
A	B	A	B
1	0	0	0
1	1	0	1
0	1	1	1
0	0	1	0

FIG. 43. (a) Electromechanical arrangement of the sine–cosine switch. (b) Switch output logic functions for clockwise and counterclockwise rotation. (c) A and B voltages for clockwise rotation.

zero. The device then goes to a nonconducting state and must be turned on again by gate current.

If the phase of the gate current in the triac is varied relative to the ac line voltage phase, the device can be made to conduct for a variable fraction of the ac cycle. When used in this way, the triac can control the average ac power delivered to resistive loads, such as heaters and incandescent lamps.

Another application of the triac is that of zero-voltage switching, which is used to prolong the life of expensive incandescent lamps, such as those used in traffic signals or flood lighting. The life of such lamps is materially shortened by high current surges when they happen to be turned on at or near the peak of the ac line voltage. The control of these lamps by a triac circuit, shown in Fig. 46, minimizes lamp failure due to current surges at turnon.

8.2.3 Industrial Power. Industrial electrical requirements usually include large electric motors, ranging up to several hundred horsepower, and electric heating and lighting loads, ranging up to hundreds of kilowatts. Switches for the large currents and voltages required of these loads must be specially designed. Most are magnetically operated to provide fast operation to minimize the arcing time on opening and closing. The contacts must be large and of sufficient heat capacity and thermal conductivity to prevent excessive temperature rise. Remote control of these switches is usually done from push buttons or, in some cases, directly from the output of computer ports or integrated logic circuits.

8.2.4 Power Distribution. Switching for power distribution requires the most severe specifications for switch gear. These devices must be capable of handling thousands of amperes at tens of thousands of volts. The contacts are usually immersed in oil or other fluids that will quench arcs and prevent contamination. Such switches are usually oper-

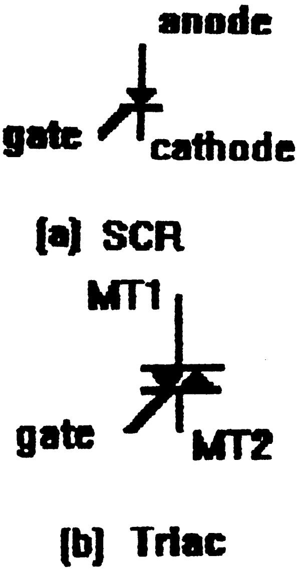

FIG. 44. (a) The silicon controlled rectifier (SCR) will conduct only when triggered by a current passing from gate to cathode. Once triggered it will conduct as an ordinary diode without further control from the gate until the anode voltage becomes negative with respect to the cathode. At this point the SCR resets and will not conduct again until the anode becomes positive and gate current is initiated. (b) The triac operation is similar to the SCR except the triac conducts in both directions. It resets to the off state each time the voltage between MT1 and MT2 passes through zero and must be triggered on again by gate current.

ated remotely from a control panel or computer, located at some central point in the distribution system, many miles from the switching site.

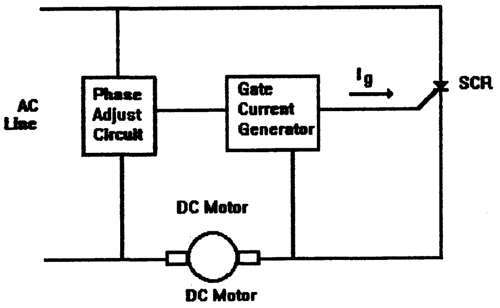

FIG. 45. dc motor speed control. The average current through the SCR is controlled by adjusting the phase of the gate current relative to the phase of the ac line.

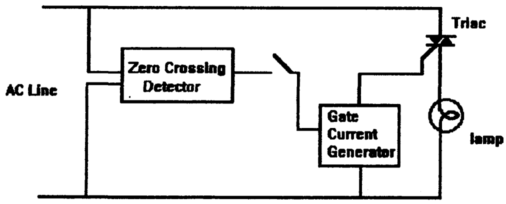

FIG. 46. Zero-voltage switch. When the switch is closed to turn the lamp on the gate-current generator will not trigger the triac until the line voltage is at or near zero.

9. FUSES AND CIRCUIT BREAKERS

9.1 Introduction

A fuse is a device that opens a circuit if the current becomes larger than a preset maximum. Its active element is a low-melting-point alloy that melts to open the circuit when the current becomes excessive. A circuit breaker performs the same task as a fuse but differs in operation in that a circuit breaker can be reset and used again, while a melted fuse must be replaced. In either case, the purpose of the device is to protect wiring or electrical equipment from damage due to excessive current.

Fuses and circuit breakers are rated according to the maximum current they will allow before opening. Current ratings vary from small fractions of an ampere to hundreds of amperes. Their response times to current overloads depend on the severity of the overload and vary inversely with the square of the overload current. Response times may range from a few milliseconds to several minutes.

9.2 Applications

For any application, the response time of the fuse or circuit breaker must be taken into account to ensure that circuit damage will not occur before the protective device can interrupt current flow. For the protection of semiconductor devices, the response times must be very short—shorter than that of most protective devices. Consequently, special fuse types have been developed to handle these short response-time requirements.

Some applications may involve large inrush currents at turnon that can be ten times larger than the normal operating current of a circuit. In these cases, the response time of the fuses or breakers used must be long enough that they will not interrupt the circuit before normal current can be established. Large inrush currents are usually encountered with the turnon of incandescent lamps and electric motors. Sometimes large inrushes occur in magnetic core devices, such as transformers and solenoids, depending upon the magnetization state of the core at turnon.

Another class of circuit protective devices, called thermal overload protectors, serve to interrupt current to a circuit, much the same as fuses and circuit breakers. These devices do not react to overcurrent, however. Instead, they operate from heat generated in their immediate surroundings, much in the same way as thermostats. Thermal overload protectors are often used in electric motors and transformers, where they are embedded in the winding or core space, and cut off current if the temperature rises beyond a safe limit. These thermal protectors usually operate mechanically and reset to the operating state when their temperature is reduced below their operating limit.

GLOSSARY

Band Pass: Range of frequencies transmitted by a circuit.

Characteristic Impedance: Termination impedance for a transmission line necessary to prevent reflection of transmitted energy back to the source.

CMOS: Complimentary metal oxide semiconductor.

dB: decibel.

Duty Cycle: Fraction of time that a switch is on during a switching cycle.

FET: Field effect transistor.

High Pass: Used to describe the behavior of a circuit that transmits all frequencies above a selected value.

III-V Compounds: Binary compound with one element taken from column III of the periodic table of elements and the other from column V.

IMPATT: Impact avalanche transit time, used to describe carrier behavior in a semiconductor structure designed specifically for generating energy in the GHz range.

JFET: Junction field effect transistor.

Low Pass: Used to describe the behavior of a circuit that transmits all frequencies below a certain selected value.

MOSFET: Metal oxide semiconductor field effect transistor.

Op Amp: Operational amplifier.

Pass Transistor: Transistor used as a series element to regulate the output of a dc power supply.

SCR: Silicon controlled rectifier.

Surface Mount: Describes the technique of mounting circuit elements on a printed circuit board by soldering them directly to the component side of the board.

Triac: Semiconductor device used for electronically controlled switching of alternating current at any selected point during the cycle.

UHF: Ultra high frequency. Frequencies ranging from approximately 450 to 950 MHz.

Further Reading

Bar-Lev, A. (1984), *Semiconductor and Electronic Devices*, 2nd ed., Englewood Cliffs, NJ: Prentice Hall.

Boylestad, R., Nashelsky, L. (1982), *Electronic Devices and Circuit Theory*, Englewood Cliffs, NJ: Prentice Hall.

Carr, J. J. (1981), *The Complete Handbook of Oscillators, Amplifiers and Multivibrators*, Blue Ridge Summit, PA: Tab.

Helms, H. L. (1986), *Contemporary Electronics Deskbook*, New York: McGraw-Hill.

Horowitz, P., Hill, W. (1990), *The Art of Electronics*, New York: Cambridge University Press.

Marston, R. M. (1988), *Optoelectronics Circuit Manual*, London: Heinemann.

Mellard, T. (1987), *Automotive Electronic Systems*, London: Heinemann.

Pippenger, D. E., Toboben, E. J. (1986), *Linear and Interface Circuits and Applications*, Texas Instruments Electronics Series, New York: McGraw-Hill.

Sum, K. K. (1984), *Switch Mode Power Conversion*, New York: Marcel Dekker, Inc.

Sze, S. M. (1981), *Physics of Semiconductor Devices*, New York: Wiley.

CIRCUITS, ANALOG

THOMAS H. LEE, *Tewksbury, Massachusetts, U.S.A.*

Introduction 479
1. **Operational Amplifiers** 479
2. **Review of Linear System Theory** 482
3. **Device Physics and Circuit Models** 486
3.1 Bipolar Transistors 486
3.2 MOSFETs 489
3.3 The Charge-Control Model 490
4. **Amplifier Design** 494
4.1 The Three Basic Amplifier Topologies 494
4.2 Biasing 496
4.3 Bandwidth Considerations 497
5. **Data Converters** 498
5.1 Analog-to-Digital Converters .. 499
5.2 Digital-to-Analog Converters... 503
6. **Miscellaneous Applications**... 503
Glossary..................... 505
Works Cited 505
Further Reading 505

INTRODUCTION

Analog circuits and systems process signals that may be continuous both in time and in amplitude. Even systems that are nominally digital may exhibit behavior that can be understood only in an analog context, especially at extremes of speed. Furthermore, despite the growing use of digital systems, there remain many instances in which analog circuits can implement a given function more conveniently (although perhaps with less accuracy) and can provide substantially higher speed operation. Finally, digital systems often require analog-to-digital (A/D) and digital-to-analog (D/A) converters (ADCs and DCAs) to provide interfaces to the "real" world.

This article examines operational amplifiers first (in Sec. 1) because many of the global issues pertaining to analog circuits are conveniently illustrated in that context. Section 2 provides a brief review of the theory of signals and systems and considers in particular the issue of system stability. A cursory treatment of semiconductor devices follows in Sec. 3, with a focus on issues relevant to circuit modeling, while Sec. 4 surveys amplifier design with an emphasis on bandwidth considerations. Section 5 presents a number of techniques for performing A/D and D/A conversion, and Sec. 6 concludes this article by examining a number of miscellaneous applications, such as switch-mode power conversion.

Throughout this article conventional, rather than electron, current flow is assumed, as is the use of Système International units.

1. OPERATIONAL AMPLIFIERS

The term "operational amplifier" derives from the variety of mathematical operations that such devices may realize conveniently, as will be shown shortly. The symbol for an operational amplifier (op-amp) is shown in Fig. 1, as are the reference directions for a number of associated electrical quantities. All potentials are with respect to an implicit ground.

An op-amp produces at its output a voltage that is some multiple of the differential input voltage, that is,

$$v_{\text{out}}=A(v_+-v_-). \qquad (1)$$

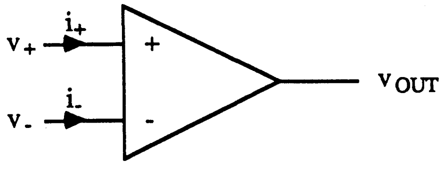

FIG. 1. Operational amplifier.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

An ideal op-amp may be defined as one possessing zero input current ($i_+ = i_- 0$), infinite differential gain (or amplification factor) A, infinite speed (that is, dynamic limitations are ignored so that A is a purely scalar quantity), and infinite output current capability. By assuming these characteristics, one can expediently analyze any ideal op-amp circuit. Corrections for the failure of real op-amps to satisfy all of these conditions can also be applied if necessary.

As an example of this type of analysis, and also to illustrate the origin of the term operational amplifier, consider the collection of circuits shown in Fig. 2. Figure 2(a) is known as an inverting amplifier and may be analyzed as follows. Assume, in addition to ideal op-amp behavior, that the output voltage of the op-amp is finite as a consequence of the negative feedback connection (there is a net inversion of signals propagating around the feedback loop). If the output is finite, the infinite differential gain assumed necessarily implies a zero input differential voltage; the voltages v_+ and v_- must be equal and, in this case, equal to zero. Because the inverting input is always forced to be at zero potential in this case, it is also known as a virtual ground.

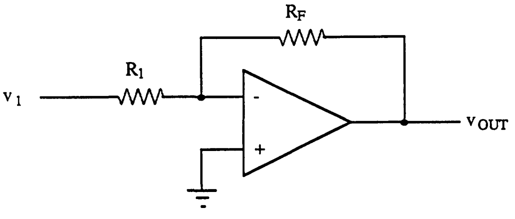

a) Inverting Amplifier

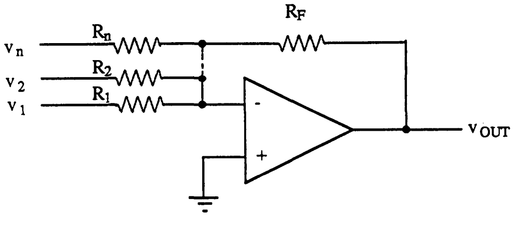

b) Inverting Summer

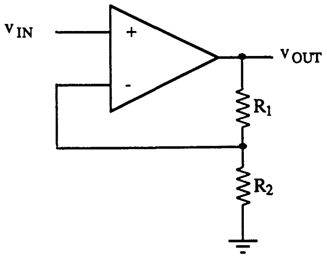

c) Non-Inverting Amplifier

FIG. 2. Examples of op-amp circuits.

The current through the input resistor is thus easily computed because all node voltages are now known. Furthermore, by Kirchhoff's current law this current must equal that through the feedback resistor R_F since we have assumed that no current flows into the input terminals of the op-amp proper. Thus,

$$i_1 = i_F \rightarrow v_1/R_1 = -v_{\text{out}}/R_F, \tag{2}$$

so that

$$v_{\text{out}}/v_1 = -R_F/R_1, \tag{3}$$

that is, this circuit provides a voltage amplification factor (or gain) that is equal to $-R_F/R_1$, with the minus sign indicating that this circuit also provides an inversion.

Figure 2(b) shows how this circuit may be extended to provide an output that is a weighted sum of any number of inputs. The analysis is simplified by observing that the current i_F is simply the sum of the currents through the input resistors, and that the output voltage is simply the product of this total current and the feedback resistor (with a minus sign). Thus,

$$v_{\text{out}} = -R_F\left(\frac{v_1}{R_1} + \frac{v_2}{R_2} + \cdots + \frac{v_n}{R_n}\right). \tag{4}$$

By appropriate choices of the various resistors, this circuit can compute a variety of linear functions of the inputs, including the average of n input signals, as well as the special case of a pure sum.

Figure 2(c) shows how a noninverting amplifier may be realized. Note that negative feedback is still employed, but that the input signal is now applied to the noninverting input of the op-amp. Proceeding as before, observe that the voltage at the inverting terminal is simply

$$v_- = v_{\text{out}}\left(\frac{R_2}{R_1 + R_2}\right). \tag{5}$$

Since we have assumed an infinite differential voltage gain, $v_+ = v_-$ as before. Further-

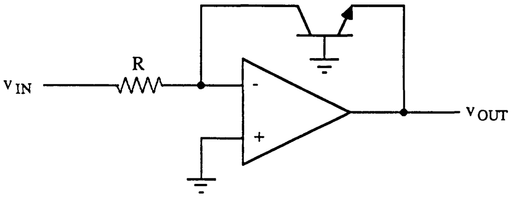

a) Logarithmic Amplifier

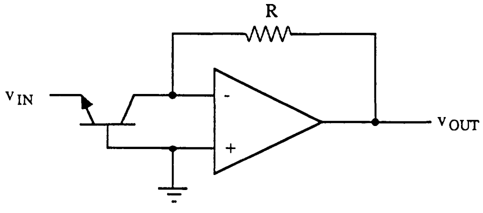

b) Exponentiator

FIG. 3. Nonlinear op-amp circuits.

more, because v_+ is equal to the input voltage, one finds that

$$\frac{v_{\mathrm{out}}}{v_{\mathrm{in}}}=\left(\frac{R_1+R_2}{R_2}\right). \tag{6}$$

With the addition of nonlinear elements (those whose current-voltage characteristics are not expressible in linear form), operational amplifiers can perform a rich variety of functions. As an example, consider the circuits, shown in Fig. 3, that combine transistors with op-amps. Although transistors will be discussed in detail in Sec. 3, assume for present purposes that a transistor obeys the following relationship at a fixed value of collector-base voltage:

$$i_C=I_S\exp(qv_{\mathrm{BE}}/kT), \tag{7}$$

where i_C is the collector current, I_S is a parameter known as the saturation current, q is the magnitude of the electronic charge, k is Boltzmann's constant, and T is the absolute temperature. Although this relation is an approximation, well-made devices conform to it within a few percent over more than seven or eight decades of collector current.

With this exponential $i_C - v_{\mathrm{BE}}$ relationship it is trivial to show that the circuit of Fig. 3(a) is a logarithmic amplifier. By virtue of the negative feedback connection, the inverting terminal is a virtual ground. Therefore the current through the input resistor is simply v_{in}/R. This current is also the collector current, and the base-emitter voltage is logarithmically related to the collector current. The output voltage is $-v_{\mathrm{BE}}$, whose value is readily computed as

$$v_{\mathrm{out}}=-v_{\mathrm{BE}}=-\frac{kT}{q}\ln\left(\frac{v_{\mathrm{in}}}{RI_S}\right). \tag{8}$$

From this last equation we see that as long as the temperature is held constant, the output voltage is proportional to the natural logarithm of the input voltage. It should be noted that there are modifications of this basic circuit that reduce greatly its temperature dependence.

The exponentiator of Fig. 3(b) can be analyzed in an exactly analogous way. The exponentiator can combine with summers and log-takers to permit multiplication entirely in the analog domain.

The computational power of op-amp circuits increases greatly as soon as energy-storage elements such as capacitors (and infrequently, inductors) are used since such elements provide some memory of past history. As an example consider the circuit shown in Fig. 4. This connection realizes an electronic integrator, that is, a circuit whose output is proportional to the integral of the input voltage (or current).

Analysis of the integrator is straightforward. Recall that charge (Q) and voltage are related in the following manner:

$$Q=CV. \tag{9}$$

Furthermore, note that the capacitor current i_C is simply the time integral of the capacitor charge. Finally, note that the negative feedback connection once again guarantees that the current through the input resistor is equal to that through the capacitor and that

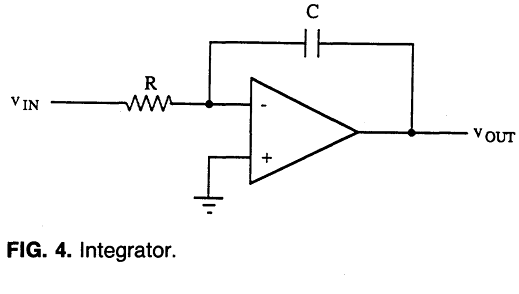

FIG. 4. Integrator.

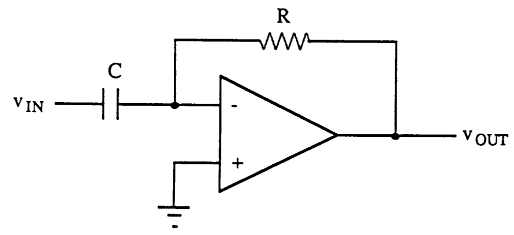

FIG. 5. Differentiator.

the capacitor voltage is the negative of the output voltage. Hence,

$$Q(t)=\int_0^t i_C\,d\tau=\int_0^t \frac{v_{\text{in}}}{R}\,d\tau \qquad (10)$$

$$\rightarrow v_{\text{out}}(t)=\frac{-Q(t)}{C}=-\frac{1}{C}\int_0^t i_C\,d\tau$$

$$=-\frac{1}{RC}\int_0^t v_{\text{in}}\,d\tau. \qquad (11)$$

The foregoing equation implicitly assumes that the capacitor possesses no initial charge. Any initial voltage across the capacitor merely adds to that computed by the integral term.

The differentiator of Fig. 5 exploits the relationship between capacitor current and voltage in a complementary manner by placing the capacitor in the input circuit, rather than in the feedback circuit.

Note that the integrator and differentiator may operate on several input signals simultaneously, so that integrals or derivatives of sums can be computed by these circuits. Additionally, a number of integrators and summers can connect to simulate a variety of differential equations (see Fig. 6 for an example).

One may perform a great many mathematical operations by combining these various building blocks in appropriate ways, and indeed the term "analog" traces its origin to the use of such circuits to construct electronic analogs of physical processes. Even though digital computers have all but completely replaced analog ones, there remain important classes of problems (such as voice and image recognition) which have resisted solution by digital means despite decades of effort. That biological systems (which are essentially analog in character) have successfully solved many of these problems has recently sparked a resurgence of interest in analog computation [e.g., in the form of so-called "neural networks" (*q.v.*)].

2. REVIEW OF LINEAR SYSTEM THEORY

The foregoing review of operational amplifiers glibly ignores the important issue of stability; there is no guarantee, *a priori*, that a feedback system will be stable. We now review the properties of systems with a focus on feedback structures and stability.

This discussion strictly applies only to linear systems, that is, those that obey superposition. Thus doubling the amplitude of excitation doubles the amplitude of response in linear systems. Although virtually every system exhibits nonlinearities if the excitation is made sufficiently large, many systems are approximately linear over some suitably restricted range. The development that follows assumes that excitations are within this linear range.

Recall that the input-output behavior of a linear system is completely characterized by either the time response to an impulse excitation (i.e., to the Dirac delta "function") or by the total frequency response (magnitude and phase over all frequencies) of the system to sinusoidal excitation.

The link between the time-domain and frequency-domain descriptions of a linear system is the Laplace transform. Let us denote by $h(t)$ the impulse response of a system, and by $H(s)$ the complex frequency response (also

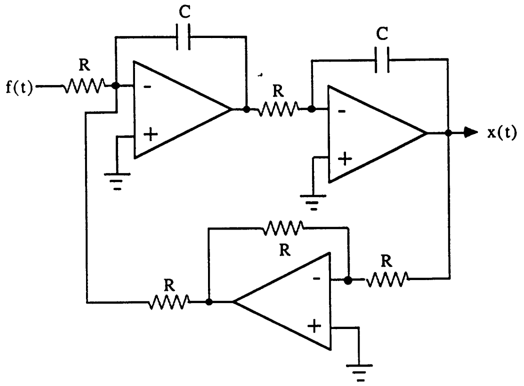

$$\frac{d^2x}{dt^2}=\frac{1}{(RC)^2}[f(t)-x(t)]$$

FIG. 6. Example of an analog computer.

called the system function or transfer function) of the linear system. Then,

$$H(s) = \int h(t)e^{-st}\,dt, \qquad (12)$$

where appropriate limits of integration are chosen depending on the situation. In general, one employs the unilateral Laplace transform if the system is *causal*, that is, if the system produces no output to any excitation (e.g., an impulse) before the excitation is applied. In such cases the limits of integration are 0 and ∞.

The steady-state response of a system to sinusoidal excitation at a frequency of ω rad/s is found by setting $s=j\omega$ in the expression for $H(s)$. The magnitude of $H(j\omega)$ is the magnitude of the gain and the phase of $H(j\omega)$ is the phase angle of the output relative to the input.

The values of complex frequency s that result in infinite values of $H(s)$ are the poles of the system, while those that result in a zero value are known, appropriately enough, as the zeros. Because $H(s)$ is a generalized (complex) gain and therefore is the ratio of Laplace transforms of the output and input, an infinite value of $H(s)$ necessarily implies that no input signal is required to produce a nonzero output signal at a pole frequency. Poles are thus also known as the *natural frequencies* or *eigenvalues* of the system, and their values determine whether a system is stable. Correspondingly, no output results when a system is excited with complex exponentials at the frequencies of its zeros.

Definitions of stability abound, each with its own subtle nuances. This article employs the *bounded-input, bounded-output* (BIBO) definition of stability, which states that a system is stable if every bounded input always produces a bounded output. It may be shown formally that a linear system is stable in the BIBO sense if the real parts of all of the poles of $H(s)$ are negative in sign.

This stability criterion makes intuitive sense since complex exponentials of the form e^{st} are eigenfunctions of linear systems. Thus natural frequencies with negative real parts correspond to components in the output that decay, rather than grow, with time.

In designing systems (particularly feedback systems), simply achieving BIBO stability usually is not sufficient. For example, the response to a particular bounded input may be undesirably large while remaining bounded. Thus one may imagine *degrees* of stability, a notion we shall quantify after examining feedback systems in detail.

Figure 7 is a general block diagram of a feedback system. It comprises a forward path [some functional block whose transfer function is denoted by $a(s)$], a feedback path denoted by $f(s)$, and a summing junction or error comparator. Because most feedback systems employ negative feedback, the requisite minus sign is customarily absorbed by the summing junction, as shown. With negative feedback, the intent is to respond to a perturbation in a manner that tends to reduce the influence of the perturbation. A familiar example is a thermostatically controlled heater, which responds to a drop in temperature by turning on. Hence the response of the system is in a direction that opposes the perturbation.

After a little algebraic manipulation the overall input-output transfer function of our general block diagram is readily found to be (see Fig. 7)

$$H(s) = a(s)/[1+a(s)f(s)]. \qquad (13)$$

With this expression in hand we are now in a position to explore some of the properties of feedback systems. Note that the poles of $H(s)$ are the roots of the denominator polynomial, $1+a(s)f(s)$. The denominator polynomial is also known as the *characteristic equation* or the *desensitivity*. The product $a(s)f(s)$ is the negative of a quantity known as the *loop transmission*. The loop transmission is found by setting all inputs to zero, cutting the loop anywhere, injecting a signal into one end (preserving the direction of signal flow that prevails in normal operation), and observing what emerges from the other end. The ratio of the output to the input in this experiment is defined as the loop transmission. The reason this quantity is important is that it determines

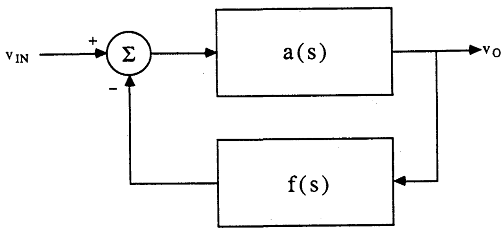

FIG. 7. Feedback system.

the roots of the characteristic equation of $H(s)$ and hence the natural frequencies of $H(s)$.

The loop transmission is an important quantity for reasons other than the control it exerts on the dynamic properties of the system. Let us consider the behavior of $H(s)$ when the magnitude of the loop transmission is large compared to unity. In this case,

$$H(s) \approx 1/f(s). \tag{14}$$

We see that the overall input-output behavior of a feedback system is largely *independent* of the characteristics of the forward path as long as the loop transmission magnitude is large compared to unity. Hence the forward path element may (and typically does) exhibit nonlinearities, temperature dependence, drifts in parameters with time, and a whole host of other undesirable properties without deleteriously affecting the overall system as long as the loop transmission magnitude (or desensitivity) remains large. In many instances, $f(s)$ is realized with linear, passive components whose characteristics are relatively stable with time and temperature so that $H(s)$ also possesses these desirable characteristics.

Let us also consider the other extreme, that is, very small loop transmission magnitudes. In this case the overall system function becomes

$$H(s) \approx a(s). \tag{15}$$

When the loop transmission magnitude is small compared with unity, we see that the system function is approximately equal to the forward path transfer function. Thus any desirable properties of $f(s)$ are *not* conferred upon $H(s)$ in this case.

Recognizing that large loop transmission magnitudes are desirable from the standpoint of greatly suppressing the dependence of overall loop behavior on any indiscretions of $a(s)$, it would appear that one should provide as much amplification in the loop as possible. Unfortunately, stability considerations preclude applying with abandon any such remedy. The reason, of course, is that the loop transmission also controls the natural frequencies and thus affects the dynamic behavior of the loop.

To illustrate this sort of problem consider a three-pole loop transmission of the following form:

$$-a(s)\, f(s) = -a_0\, f_0/(\tau s+1)^3, \tag{16}$$

where f_0 is a purely scalar quantity whose value we shall vary. The characteristic equation associated with this loop transmission is

$$P(s) = 1 + a_0\, f_0/(\tau s+1)^3 \tag{17}$$

and has two roots that acquire positive real parts as $a_0\, f_0$ increases. In fact, the system is unstable in the BIBO sense for any $a_0\, f_0$ equal to or greater than 8. Thus, obtaining large loop transmission magnitudes is at odds with stability considerations.

Even though this conclusion appears strictly to apply only to this particular example, the result is in fact general. The reason is that all physical systems are essentially *low-pass* in nature, at least ultimately so. As a consequence there are always additional poles in the loop transmission that will eventually cause trouble for a large enough magnitude of $a_0\, f_0$. Loop transmission magnitudes cannot be made arbitrarily large as a consequence.

Fortunately, the tradeoff between stability and desensitivity can be made less severe by a number of compensation methods. To understand how these compensators function, let us explore the frequency response of a system that is nearly unstable. Specifically, consider a system in which $a(j\omega)f(j\omega)$ is nearly -1 over some range of frequency ω and that this frequency range is relatively narrow in span. It follows that the magnitude of $H(j\omega)$ will be quite large over a similar frequency range compared with its magnitude elsewhere. We regard this large peaking in the frequency response as a sign of encroaching instability.

Appropriate modification of the magnitude and/or phase behavior of the loop transmission can help the product $a(j\omega)f(j\omega)$ avoid ever getting close to a value of -1. Stated in greater detail, satisfactory operation is obtained as long as one avoids the simultaneous occurrence of a unit magnitude and $-180°$ phase shift of $a(j\omega)f(j\omega)$.

Quantitative measures of stability based on these ideas are the gain margin and phase margin. These are defined as follows:

$$\text{Gain margin} \equiv \frac{1}{|a(j\omega_{-180})f(j\omega_{-180})|}, \tag{18}$$

Phase margin

$$\equiv 180° - \angle a(j\omega_{\text{unity}})f(j\omega_{\text{unity}}), \tag{19}$$

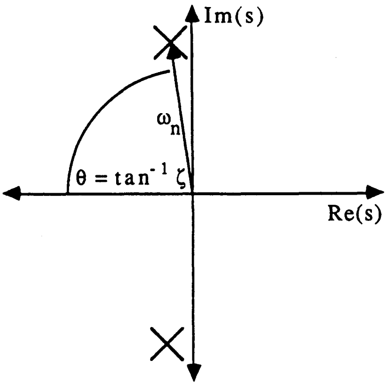

FIG. 8. Complex pole parameters.

where ω_{-180} is the frequency at which $a(j\omega)f(j\omega)$ has a phase shift of $-180°$, and ω_{unity} (also known as the crossover frequency) is the frequency where the loop transmission magnitude is unity.

These quantities have the following interpretations. The gain margin is the factor by which the loop transmission magnitude may be increased before instability results. The phase margin is the amount of additional (negative) phase shift the loop can tolerate at the crossover frequency before becoming unstable. Generally, a gain margin greater than or equal to about three, coupled with a minimum phase margin of approximately 30°–60°, will yield satisfactorily stable performance.

Use of these stability measures is attractive because only the frequency response (magnitude and phase) of the loop transmission needs to be determined. Since one can usually measure the frequency response with relative ease, application of these stability criteria is generally a straightforward exercise in practice.

Unfortunately, no simple general expression relating these measures to parameters such as overshoot (in response to a step excitation) or frequency response peaking exists. However, if the system is dominated by a complex pole pair (as is common), and if $H(s)$ is never zero for any finite values of s, then $H(s)$ may be approximated as

$$H(s) \approx \frac{K}{(s^2/\omega_n^2) + (2\zeta s/\omega_n) + 1}, \qquad (20)$$

where K is some constant factor (the dc gain of the closed-loop system) and ω_n and ζ (known as the damping ratio) are defined as shown in the *s-plane* plot of Fig. 8, where the poles are represented by ×'s.

For this two-pole case the maximum step-response overshoot P_0, measured as the ratio of the maximum value to the final value, is

$$P_0 = 1 + \exp[-\pi\zeta(1-\zeta^2)^{-1/2}], \qquad (21)$$

and the maximum peaking M_p in the frequency response, measured as the ratio of the maximum magnitude of $H(j\omega)$ to its magnitude at low frequencies (e.g., dc), is

$$M_p = (2\zeta)^{-1}(1-\zeta^2)^{-1/2} \quad \text{for } \zeta < 2^{-1/2}. \qquad (22)$$

Note that there is no peaking in the frequency response for values of ζ greater than 0.707 in an all-pole second-order system.

Having presented some measures of stability, we may now discuss several strategies for improving system stability at a given value of $a_0 f_0$. As one might suspect, compensation methods center around either modifying the magnitude of the loop transmission near frequencies where the phase shift of $a(j\omega)f(j\omega)$ is near $-180°$, or modifying the phase characteristics of $a(j\omega)f(j\omega)$ near the crossover frequency, or both.

A lag compensator (Fig. 9) reduces the magnitude of the loop transmission above a certain frequency that may be controlled by design (consider the low- and high-frequency asymptotic gains of this network). Because this type of compensator reduces the crossover frequency, the speed of the system is reduced at the same time the stability is improved. Furthermore, the lag compensator introduces a phase lag (hence its name) over some range of frequencies. Care must be taken in the design of the compensator to guarantee that only a small amount of this negative phase shift remains at the crossover frequency. A general rule of thumb is to place the zero of the lag compensator at least a

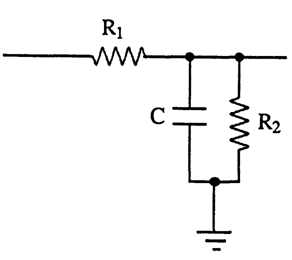

FIG. 9. Lag network.

factor of 10 below the crossover frequency to avoid this problem.

The lead compensator shown in Fig. 10 functions by adding a positive phase shift (a leading phase) to the loop transmission. There also is an attendant increase in magnitude that forces the crossover frequency to increase as well. While a higher crossover frequency increases the speed of the system, care must be taken not to push out crossover too far since many loop transmissions exhibit a rapid increase in negative phase shift as frequencies increase. Hence an excessive increase in crossover frequency may conceivably result in a net degradation in stability rather than the anticipated improvement. A general rule of thumb is to place the zero of the lead compensator at about the anticipated crossover frequency. Some iteration usually is required for a satisfactory design.

While the notion of phase margin has formed the primary focus for this discussion on stability and compensation, it must be noted that there are some pathological cases in which phase margin may not be defined. For example, it is possible for a loop transmission to exhibit multiple crossover frequencies; it is even possible not to have an explicit crossover frequency at all (at least mathematically). In these cases one must resort to the Nyquist stability criterion (Roberge, 1975) as the final arbiter of the question of stability. As with gain and phase margin, the Nyquist criterion requires only frequency response information and thus may operate on experimentally derived data. While gain and phase margin essentially attempt to decide the stability question by noting the behavior of the loop transmission at only two discrete frequencies (and primarily one, the crossover frequency), the Nyquist criterion examines the loop transmission at all frequencies and is able to determine stability in all cases. Because the Nyquist criterion is somewhat cumbersome to apply, the gain and phase margin criteria (primarily the latter) are most widely used. Fortunately, the gain and phase margin ideas presented here are adequate in nearly all cases of practical interest, and one must resort to use of the Nyquist criterion only rarely.

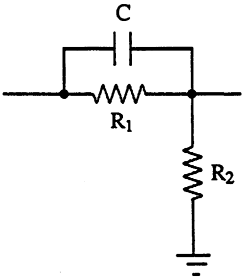

FIG. 10. Lead network.

3. DEVICE PHYSICS AND CIRCUIT MODELS

The two broad classes of active devices presently in most common use are the bipolar transistor and the field-effect transistor (FET). We will briefly review operation of the bipolar transistor and the metal-oxide-semiconductor FET (MOSFET) with an eye toward the development of linearized circuit models of these devices. This linearization is necessary because of the significant mathematical barriers that attend any effort to analyze exactly an aggregation of elements as nonlinear as transistors. Although machine computation may always be employed, useful design insight rarely emerges from such an exercise. Thus simulation is best used as a final, verification tool rather than as a primary design tool. The focus of this section is on the development of models that are more helpful in circuit design as well as in initial analyses.

3.1 Bipolar Transistors

An *npn* bipolar transistor is depicted in simplified form in Fig. 11 along with its schematic symbol. A *pnp* device is constructed similarly, although with materials of the opposite doping type. The schematic symbol for a *pnp* transistor is the same as that for an *npn* device except that the emitter arrow points in the other direction.

In normal operation (known as "forward active" operation) the potential at the base terminal of an *npn* transistor is made more positive than that at the emitter. Additionally, the collector potential is made greater than that at the base. Under these conditions the collector current is related to the base-emitter voltage as follows:

$$i_C = I_S\,[\exp(qV_{BE}/kT) - 1], \qquad (23)$$

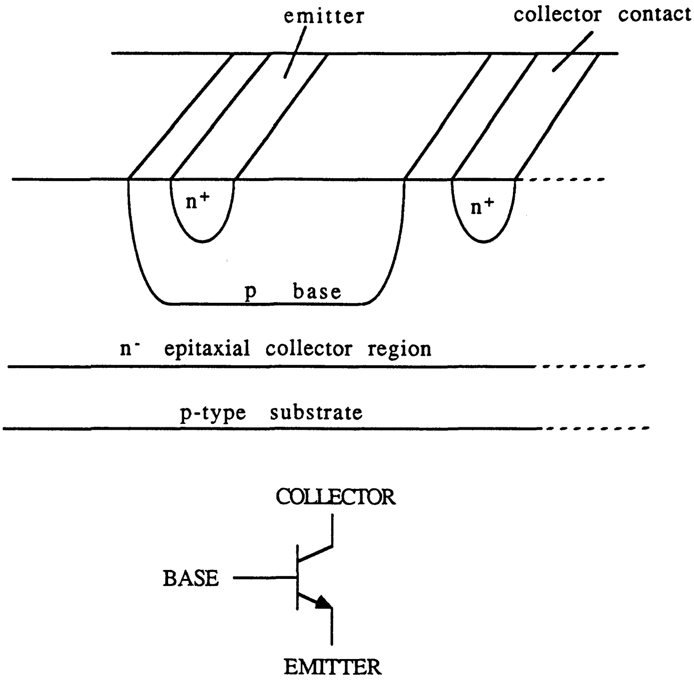

FIG. 11. *npn* transistor.

where I_S is known as the reverse saturation current, q is the magnitude of the electronic charge, k is Boltzmann's constant, and T is the absolute temperature in kelvins. Practical devices conform to this relationship over several decades of collector current if the collector-to-emitter potential is held fixed. Additionally, the exponential term dominates over the -1 term to such an overwhelming extent at all practical currents that the -1 term is often omitted.

The base current is typically a small fraction (ideally zero) of the collector current. The ratio of collector current to base current is denoted by β and is typically on the order of 100, although it can range from a few tens to several thousand depending on the details of device fabrication. The value of β depends greatly on temperature, processing, and operating current levels and is thus by no means constant, even for devices of nominally the same type. For example, it is not uncommon to observe a 2 or 3 to 1 change in β as the temperature changes from −55° to 125 °C (β increases with temperature). Hence care must be taken to avoid designs that are overly sensitive to β.

As noted earlier the highly nonlinear dependence of collector current on base-emitter voltage makes exact analysis of transistor circuits somewhat difficult. Linearized models, valid only over some restricted operating range, are typically used instead. In this case the exponential relationship is approximated by a linear one, centered about some *operating point* value.

Figure 12 illustrates this approach to linearization. The operating point on the i_C-v_{BE} curve is simply some value about which one performs a Taylor-series expansion of the true exponential relationship. Only the linear term in the expansion is preserved. As a consequence all variables may be considered the sum of two components, an operating point value (also known as the bias point, dc point,

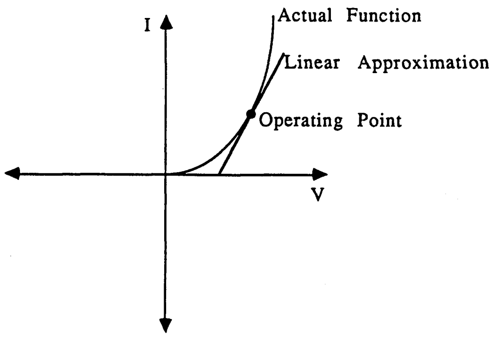

FIG. 12. Linearization.

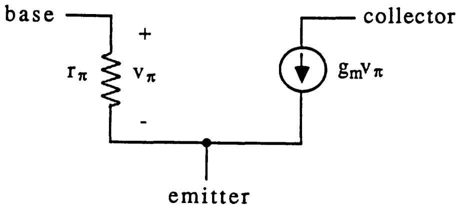

FIG. 13. Hybrid-π model.

or quiescent point) that is assumed constant, and an incremental or small-signal component. The incremental quantities thus represent small perturbations about the operating point and are linearly related to one another.

A combination of upper- and lower-case variable names and subscripts is customarily used to denote the various quantities. A dc point is represented by an upper-case variable with upper-case subscripts, while lower-case variables with lower-case subscripts denote small-signal quantities. The sum of a dc and incremental quantity is known as a *total variable* and is represented by a lower-case variable with an upper-case subscript. Thus as an example

$$i_B = I_B + i_b. \tag{24}$$

With these formalities out of the way it is a straightforward exercise to verify that Fig. 13 is a small-signal model of a bipolar transistor. This model, known as the *hybrid-π* model for historic reasons, employs a dependent current source (also known as a *transconductance*), with a value g_m, to represent the linearized relationship between a change in collector current and a change in base-emitter voltage. The resistance from base to emitter, traditionally called r_π, accounts for the signal-dependent base current that flows. The values of g_m and r_π are given by the following equations:

$$g_m = qI_C/kT, \tag{25}$$

$$r_\pi = \beta/g_m. \tag{26}$$

The quantity kT/q occurs frequently enough that it is often referred to as V_T, the "thermal voltage," and has a value of about 25 mV at (warm) room temperature.

Although the model of Fig. 13 is adequate for the analysis of static circuit behavior in many cases, there are instances where its limitations become particularly significant. As an example, consider the incremental model of a circuit known as a common-emitter amplifier shown in Fig. 14. The voltage gain of this amplifier is easily found by noting that the input voltage is equal to the voltage v_π and that the output voltage is (minus) the product of the transconductance and the output load resistance R_L. Thus,

$$\text{gain} = -g_m R_L. \tag{27}$$

This last expression suggests that the gain may be increased without limit simply by increasing the effective load resistance. However, the gain per stage in practice is typically limited to values of a few thousand because of a phenomenon called *base-width modulation*, also known as the Early effect. Since the collector-base space-charge layer (SCL) has a voltage-dependent width, variations in collector potential cause variations in the effective base width. An increase in collector voltage causes the SCL to extend further into the base, effectively narrowing it. Because β is a critical function of base width, the base current then decreases at the same time the collector current increases. These effects are modeled as additional resistances as shown in Fig. 15.

The base-width modulation factor η is typically in the range of 10^{-5}–10^{-3}. Sometimes a quantity known as the Early voltage (denoted V_A) is specified instead. The ratio of the thermal voltage V_T to the Early voltage is equal to η.

The model as presented so far has at least one important deficiency—it fails to represent any dynamic limitations of the transistor. This oversight is readily corrected with the improved model of Fig. 16. Capacitances are added across the base-emitter and collector-base junctions since each junction has a capacitance associated with the space-charge layer. The junction capacitances are depen-

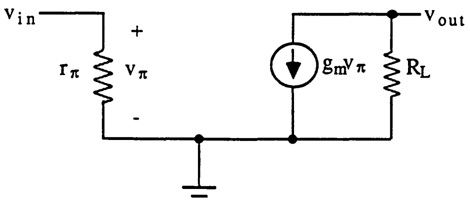

FIG. 14. Incremental model of a common-emitter amplifier.

dent on junction voltage as well as on doping levels and geometries and generally behave as follows:

$$C_j = \frac{C_{jo}}{(1 - V_J/\Psi_o)^m}, \tag{28}$$

where C_{jo} is the capacitance at zero bias, V_J is the dc voltage across the junction, Ψ_o (the *contact potential*) is a doping-level- and temperature-dependent "constant" on the order of a volt, and m is a doping-profile-dependent constant, usually between 0.2 and 0.5. The small-signal junction capacitance associated with the collector-base region is usually denoted by C_μ while that associated with the base-emitter region is usually denoted by C_{je}.

In the case of the base-emitter junction in forward bias (base voltage greater than the emitter voltage in an *npn* device) there is an additional capacitive component due to the minority carriers present in the base region under such a condition. Because a change in base-emitter voltage changes the amount of charge injected (and thus present) in the base region, and since this change requires the flow of base current, there is a *current*-dependent capacitance, known as the *diffusion capacitance*, denoted by C_b. The sum of the two base-emitter capacitances is known as C_π.

With this more elaborate model one may compute quantities such as the bandwidth of an amplifier provided that the various device parameters are known.

As a last comment on the hybrid-π model, it should be noted that even the augmented model is not valid over an infinitely wide frequency range. Owing to implicit assumptions of quasistatic device operation, the model eventually breaks down when the frequencies involved are high enough to invalidate these assumptions. In general, the model is valid as long as the frequencies are at least a factor of 3–5 lower than a value denoted as ω_T (sometimes called the current gain-bandwidth product), defined as

$$\omega_T = g_m/(C_\pi + C_\mu). \tag{29}$$

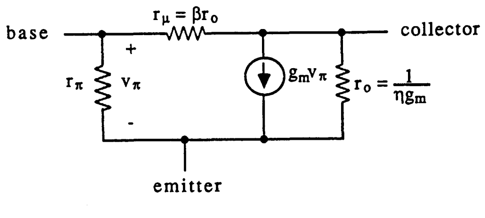

FIG. 15. Augmented hybrid-π model.

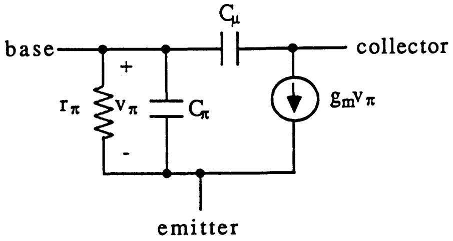

FIG. 16. Dynamic hybrid-π model.

3.2 MOSFETs

A typical *n*-channel MOSFET is shown in Fig. 17, along with two of its commonly used symbols. From the figure it is clear that a MOSFET is actually a four-terminal device. The fourth terminal, the substrate or body, is not always explicitly shown (e.g., as in one of the schematic symbols). If one ignores the substrate, the drain current of the *n*-channel MOSFET behaves in the following manner on a static basis:

$$i_D = \frac{\mu C_{\text{ox}} W/L}{2}(v_{GS} - V_T)^2, \quad v_{GS} - v_{DS} < V_T, \tag{30}$$

$$i_D = \frac{\mu C_{ox} W/L}{2}[2(v_{GS} - V_T)v_{DS} - v_{DS}^2], \quad v_{GS} - v_{DS} > V_T. \tag{31}$$

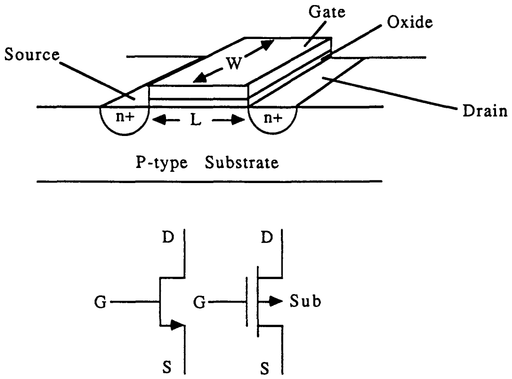

FIG. 17. *N*-channel MOSFET.

The condition $v_{GS} - v_{DS} < V_T$ defines the *saturation* or *pinch-off* region of operation while $v_{GS} - v_{DS} > V_T$ defines the *Ohmic* or *triode* region of operation. The saturation region is the usual operating regime of the MOSFET, where its behavior is similar to that of a bipolar transistor in the forward active region of operation. Two important differences, however, are that the gate current is practically zero independent of the drain current, and that the relationship between drain current and gate-source voltage is square law, not exponential.

An incremental model for the MOSFET in saturation, including capacitances, is shown in Fig. 18. This model also includes an additional resistance between the drain and source to account for the effects of *channel-length modulation,* a phenomenon analogous to the Early effect in bipolar transistors.

Figure 19 shows a simplified incremental model for a MOSFET in the triode region of operation. Note that the device simply acts as a voltage-controlled resistor at frequencies where the capacitances may be ignored.

3.3 The Charge-Control Model

So far we have examined only the small-signal behavior of transistors. We now consider the development of a model that will permit us to analyze the large-signal dynamic behavior of the bipolar junction transistor (BJT) in all regions of operation. This model is the "charge-control" model, and provides a tremendous amount of insight that is especially valuable in the design of switching (e.g., digital) circuits.

We will initially ignore the effects of space-charge-layer capacitances (the large-signal counterparts of C_{je} and C_{μ}). Furthermore, we will focus at first on the forward-active (normal) mode of operation, since the charge-control equations for reverse-active (i.e., roles of collector and emitter interchanged) operation can be written by analogy.

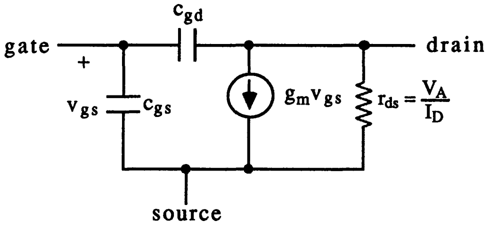

FIG. 18. Incremental model of MOSFET in the saturation region.

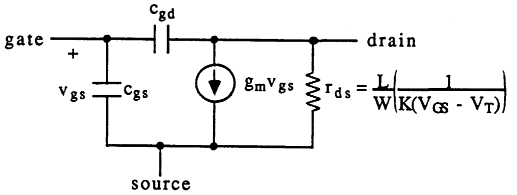
C_{gd}
gate
+
v_{gs}
C_{gs}
$g_m v_{gs}$
$r_{ds} = \frac{L}{W}\left(\frac{1}{K(V_{GS} - V_T)}\right)$
drain
source

FIG. 19. Incremental model of MOSFET in the triode region.

Also, since saturation involves forward bias on both the base-emitter and base-collector junctions, we will write the charge-control equations for the saturation region simply by adding the equations for the forward-active region of operation to those for the reverse-active region. Although the resulting set of equations will appear imposing, we will find that a greatly simplified set of equations will often suffice to describe device behavior in saturation. In all that follows, the subscript "*F*" will denote those quantities pertaining to the forward-active mode of operation, while the subscript "*R*" will denote quantities pertaining to reverse-active operation.

After deriving the basic charge-control equations we will relate charge-control and hybrid-π model parameters to underscore the relationship between small-signal and large-signal behavior. We will then examine several examples of how to use the model to carry out actual calculations. We will conclude by augmenting the model to include SCL capacitances, and by making some observations about the validity and accuracy of the charge-control model.

To begin the development of our model, recall that a "typical" static excess-carrier distribution for an *npn* transistor looks as shown in Fig. 20. The excess-carrier concentration at the base-emitter SCL edge, $n_b'(0)$, is equal to $n_{bo}[\exp(v_{BE}/v_T) - 1]$, where n_{bo} is the equilibrium minority-carrier concentration in the base. Assume that this relationship holds over essentially infinite bandwidths.

If we may neglect recombination, the distribution will be a perfect triangle, and the total amount of excess charge in the base can be computed readily as $qAWn_b'(0)/2$, where W is the base width and A is the area of the emitter junction. Hence the total excess charge

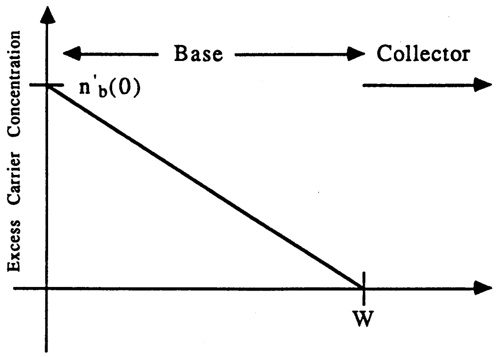

FIG. 20. Minority-carrier distribution in the base region in the foward-active region.

in the base varies exponentially with the applied junction voltage.

Recall that charge transport in the base region is dominated by diffusion so that the slope of the excess-carrier distribution at the base-emitter SCL edge is proportional to the emitter current and the slope of this distribution at the collector-base SCL edge is proportional to the collector current. Since we are neglecting recombination at this point, these two slopes are equal. Because the total excess charge in the base region, q_F, is proportional to the slope of the excess-carrier distribution (via a simple geometrical consideration, that is, that the area of a right triangle is proportional to the slope of the hypotenuse), we may write

$$i_C=q_F/\tau_F, \tag{32}$$

where the constant of proportionality τ_F possesses the dimensions of time.

We need a second charge-control equation to account for the base current. In actual fact, the base current consists of three components. One is proportional to q_F and supports recombination of carriers in the base region. Another component results from the reverse injection of holes from the base into the emitter, and is also proportional to q_F. The third component accounts for currents that support *changes* in q_F. Thus this last current is proportional to dq_F/dt. Consequently, one may express the base current as follows:

$$i_B=q_F/\tau_{BF}+dq_F/dt, \tag{33}$$

where the proportionality constant τ_{BF} also possesses the dimensions of time.

Finally, the emitter-current charge-control equation is readily derived by applying Kirchhoff's current law:

$$-i_E=i_C+i_B. \tag{34}$$

If we now consider operation of the BJT in the reverse-active region, that is, with the roles of emitter and collector reversed, we can write the three relevant charge-control equations by analogy with the forward-active case:

$$i_E=q_R/\tau_R, \tag{35}$$

$$i_B=q_R/\tau_{BR}+dq_R/dt, \tag{36}$$

$$-i_C=q_R(1/\tau_R+1/\tau_{BR})+dq_Rdt. \tag{37}$$

The six equations presented in the foregoing development describe completely the large-signal dynamic behavior of bipolar junction transistors in all regions of operation, neglecting the influence of space-charge-layer capacitances. As we shall see, operation of the BJT even in the saturation region can be analyzed by combining the equations for the forward- and reverse-active modes.

Before proceeding further, recognize that the charge-control model must be compatible with the small-signal model. Let us now turn to the task of relating hybrid-π model parameters to those of the charge-control model. This exercise is not purely pedagogic; transistor data sheets rarely provide charge-control parameters in any useful form, so that one is usually forced to derive them from the small-signal parameters that *are* usually given.

One would hope that the charge-control and small-signal models converge for small-signal dynamic analyses. In fact, we will *require* this convergence and exploit this requirement to relate charge-control parameters to hybrid-π model parameters. For example, consider the forward-active region collector-current and base-current charge-control equations. In the steady state, the ratio i_C/i_B is equal to τ_{BF}/τ_F. The hybrid-π model tells us that this ratio is β_F. Hence, one relationship we have discovered is

$$\tau_{BF}=\beta_F\tau_F. \tag{38}$$

Similarly, for the reverse-active region,

$$\tau_{BR}=\beta_R\tau_R. \tag{39}$$

Thus the base and collector time constants are simply related by the respective β.

Now consider τ_F itself. Perhaps the most expedient way to relate it to hybrid-π parameters is to equate base currents in the limit of infinite frequency (while continuing to neglect SCL capacitance terms). In the charge-control (CC) model, the base current approaches

$$i_b=dq_F/dt, \tag{40}$$

while in the hybrid-π model the base current approaches

$$i_b=C_b\,dv_{be}/dt. \tag{41}$$

Now recall from elementary device physics that we may express the diffusion capacitance C_b as follows:

$$C_b=g_mW^2/2D_n=g_m\tau, \tag{42}$$

where g_m is the transconductance, W is the effective base width, and D_n is the diffusivity of electrons (recall that we are considering *npn* transistors here). Equating the expressions for base current and substituting for C_b, we find that

$$q_F/\tau=g_mv_{be}. \tag{43}$$

Noting that g_mv_{be} is simply the collector current so that

$$i_c=q_F/\tau, \tag{44}$$

we conclude that τ is simply equal to τ_F by comparing this equation with the collector-current CC equation. Thus, we have discovered an important link between CC and small-signal model parameters, namely, that τ_F is equal to C_b/g_m.

In many circuits, the drive levels are high enough to saturate the transistor. In saturation, both transistor junctions are forward biased. Since both the base-emitter and base-collector junctions consequently inject excess carriers into the base region, we need to sum the reverse- and forward-active CC equations to model the combined effects of these two sources of charge. Thus the relevant charge-control equations are

$$i_C=\frac{q_F}{\tau_F}-\frac{dq_R}{dt}-q_R\left(\frac{1}{\tau_R}+\frac{1}{\tau_{BR}}\right), \tag{45}$$

$$i_B=\frac{dq_F}{dt}+\frac{q_F}{\tau_{BF}}+\frac{dq_R}{dt}+\frac{q_R}{\tau_{BR}}, \tag{46}$$

$$-i_E=i_B+i_C. \tag{47}$$

At this point, it is best to consider a picture of the charge distribution in the base region under conditions of saturation (Fig. 21). The form of the charge distribution in the base region in saturation has the general shape of a trapezoid. This shape in turn may be regarded as the result of summing two triangular distributions, one of which corresponds to q_F and the other of which corresponds to q_R. The areas of the triangles are proportional to their respective charge components, but be aware that the constants of proportionality are *not* the same for the two components. Thus one cannot simply solve the problem purely graphically. The primary virtue of this picture is that it allows us to rethink the problem of describing the dynamic behavior of transistors in the saturation region.

Let us call the amount of q_F necessary to put us at the edge of saturation q_{BO}. The additional base charge in saturation will then be some mixture of forward and reverse charge. Denote the additional amount of forward charge as q_{FS} and let the reverse charge be q_R. It is also useful to lump these last two components together in a single quantity called q_S, the saturation charge. Hence the total base charge may be expressed as

$$q_B=q_{BO}+q_{FS}+q_R=q_{BO}+q_S. \tag{48}$$

Now, from Eq. (46), we can develop a relationship between the amount of reverse charge and the amount of excess forward charge in saturation. It can be demonstrated that there is a wide separation in natural frequencies of the saturation charge control

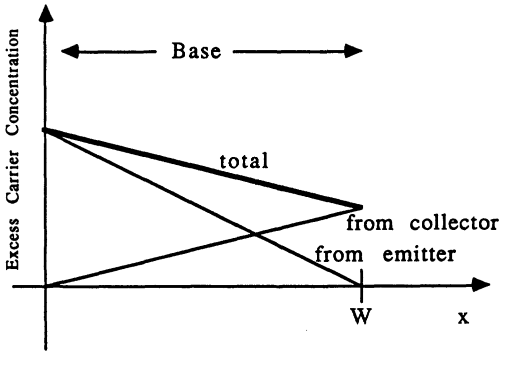

FIG. 21. Minority-carrier distribution in the base region in saturation.

equations and thus we may neglect the d/dt term in Eq. (46) to obtain

$$\frac{q_{FS}}{\tau_F}=q_R\left(\frac{1}{\tau_R}+\frac{1}{\tau_{BR}}\right). \tag{49}$$

This last equation is a key equation, since it tells us that only two of q_R, q_{FS}, or q_S are independent. Thus one can rewrite Eq. (46) in terms of just q_S, for example. If this is done, one obtains the following expression for the base current in saturation:

$$i_B=\frac{dq_S}{dt}+\frac{q_S}{\tau_S}+\frac{q_{BO}}{\tau_{BF}}. \tag{50}$$

Recognize that the last term of this equation is simply the base current necessary to put the transistor at the edge of saturation. Calling this current i_{BO} allows one to write

$$i_B-i_{BO}=\frac{dq_S}{dt}+\frac{q_S}{\tau_S}, \tag{51}$$

where the time constant is

$$\tau_S=\frac{\tau_{BF}(\beta_R+1)+\beta_F\tau_{BR}}{\beta_F+\beta_R+1}. \tag{52}$$

Notice that the dynamics in saturation are first order in nature. Additionally, the driving "force" behind changes in the saturation charge is simply the difference between the applied base current and the amount of base current necessary to put the device at the edge of saturation. Furthermore, these changes take place with a characteristic time constant that is a particular weighted sum of the forward and reverse time constants.

These equations reveal an important insight. Since both junctions are in forward bias in saturation, the base-emitter and base-collector voltages are essentially fixed. However, base charge continues to build (as long as the base current exceeds i_{BO}), and the transistor cannot be returned to a nonconducting state until this charge is removed. Hence, saturated transistors are slow to turn off. This property is particularly troublesome for digital circuits, where rapid transitions between off and on states are always desired.

One way to increase switching speed in such cases is to avoid saturation in the first place. Two methods for accomplishing this end are shown in Fig. 22. Both of these circuits employ diode clamping to prevent significant forward bias of the collector-base junction. In Fig. 22(a) (known as a "Baker clamp"), the upper diode turns on as the transistor approaches saturation. Any additional base current diverts through this diode rather than causing increases in base charge, thus preventing saturation.

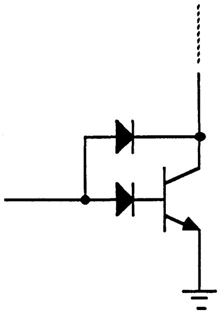

(a) Baker Clamp

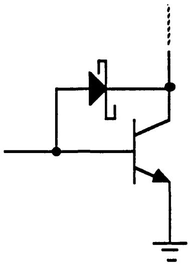

(b) Schottky Clamp

FIG. 22. Antisaturation circuits.

The circuit of Fig. 22(b) employs a single Schottky diode to perform the same function. A Schottky diode may be formed at the interface of a metal with a lightly doped semiconductor, and has a relatively small forward voltage (e.g., 0.3 V) in the conducting state, compared with the 0.6–0.7 V that is typical of standard junction diodes. Hence, the Schottky clamp allows the collector-base junction to be forward biased, but only weakly so. Because of its simplicity and efficacy, Schottky clamping is employed in the fastest families of TTL (transistor-transistor logic) digital circuits (see CIRCUITS, DIGITAL).

It may be shown that SCL capacitances can control switching speed in many cases, just as they can control bandwidth in the small-signal realm. Thus we need to augment our CC model and equations to reflect this reality. With the addition of the capacitances and their associated charges, q_{VE} and q_{VC}, come associated terms in the CC equations. These SCL currents are

$$i_B=\frac{d(q_{VE}+q_{VC})}{dt}, \tag{53}$$

$$i_C=-\frac{dq_{VC}}{dt}, \tag{54}$$

$$i_E=-\frac{dq_{VE}}{dt}. \tag{55}$$

These terms need to be added to the CC equation previously derived, and would seem to increase to unpleasant levels any analyses we would wish to carry out, especially since the charge terms are nonlinearly related to the junction voltages. However, note that not all of the terms in the equations are important in all regions of operation.

In the cutoff region, which we shall arbitrarily define as associated with less than a 0.4-V forward bias on the base-emitter junction, only the SCL terms are important since there is essentially no q_F in such a case.

In the forward-active region (FAR), the base-emitter voltage is roughly constant, so we can often neglect the base-emitter SCL capacitance. Finally, in the saturation region, all of the terminal voltages are constant so that one may neglect all of the SCL capacitances (there are no d/dt's). Thus the additional complexity imposed by the SCL terms is actually quite modest.

Both the hybrid-π and charge-control models assume that the base charge can be represented by a single "lump." However, as frequencies approach f_T $(=\omega_T/2\pi)$ one must take the true distributed nature of the base charge into account. Recall that a consequence of this neglect is the inability of the hybrid-π model to give us accurate answers for frequencies near f_T. A similar limitation applies to the CC model—it simply cannot provide any reliable predictions on time scales comparable to, say, τ_F (for the FAR). Thus while it is true that the CC model allows us to analyze the large-signal dynamic behavior of switching circuits over all regions of operation, CC analysis does not evade the fundamental frequency limitations of the hybrid-π model.

4. AMPLIFIER DESIGN

4.1 The Three Basic Amplifier Topologies

The three fundamental bipolar transistor amplifier topologies are the *common-emitter, common-base,* and *common-collector* (more commonly known as the *emitter follower*), as shown in Fig. 23. Analogous structures realized in MOS form are obtained by recognizing that the base terminal corresponds to the gate, the emitter to the source, and the collector to the drain.

The incremental model for the common-emitter amplifier is repeated in Fig. 24(a) (ignoring capacitances for the moment). As noted earlier, the voltage gain of this amplifier is $-g_mR_L$ (the minus sign indicates that the amplifier inverts the signal). From the figure it is also easy to see that the input impedance is simply r_π while the output resistance is R_L. At a collector bias current of 1 mA, r_π is about 2.5 kΩ at a temperature of 300 K. This value of resistance is considered neither particularly high nor low, and the input impedance of the common-emitter amplifier is thus considered moderate.

The incremental model for the common-base amplifier is shown in Fig. 24(b). It is straightforward to show that the input impedance of this amplifier is

$$z_{\text{in}}=\frac{r_\pi+R_S}{\beta+1}\approx\frac{r_\pi}{\beta}+\frac{R_S}{\beta}=\frac{1}{g_m}+\frac{R_S}{\beta}. \tag{56}$$

At a collector current of 1 mA, the reciprocal of transconductance is approximately 25 Ω while β is generally a fairly large number (e.g., a couple of hundred). As a consequence the input impedance of the common-base amplifier is considered low.

Both the output resistance of this amplifier and the voltage gain magnitude are the same as for the common-emitter amplifier with the important difference that the common-base amplifier does not invert.

The emitter follower's incremental model

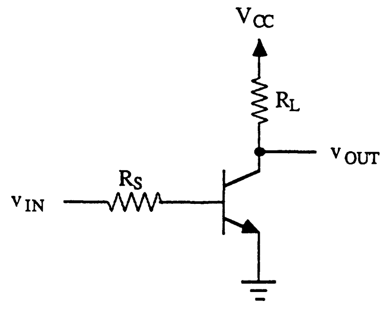

(a) Common-Emitter

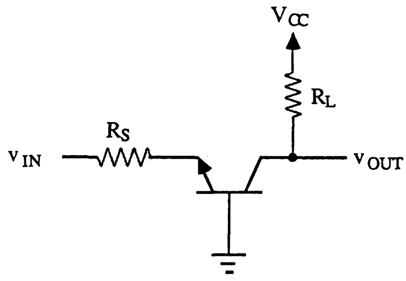

(b) Common-Base

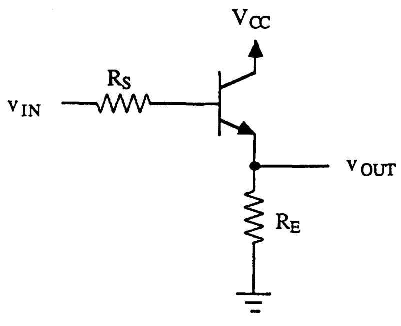

(c) Emitter Follower

FIG. 23. The three basic amplifier topologies.

is shown in Fig. 24(c). The voltage gain is readily found as

$$\text{gain} = \frac{(\beta+1)R_E}{(\beta+1)R_E + R_S + r_\pi} < 1, \tag{57}$$

while the input impedance is

$$z_{\text{in}} = (\beta+1)R_E + r_\pi \tag{58}$$

and the output impedance is

$$z_{\text{out}} = \frac{R_S + r_\pi}{\beta+1} \| R_E \tag{59}$$

where the symbol "$\|$" denotes a parallel resistance combination $[a\|b = ab/(a+b)]$.

Examination of these expressions reveals that the input impedance of the emitter follower can be quite large, because of the multiplicative factor $\beta+1$, while the output impedance can be quite small, because of the appearance of the same factor in the denominator. Thus the emitter follower can drive relatively heavy loads while not appreciably disturbing whatever stage is driving the follower. For this reason the follower is also known as a buffer, and this action reveals the value of the amplifier even though its voltage gain is unity at best.

These three topologies are often combined in a variety of ways to exploit the various advantages of each. As an example consider

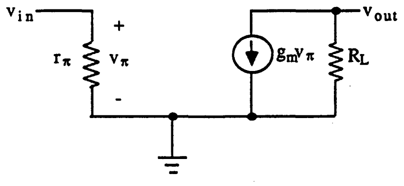

(a) Common-Emitter

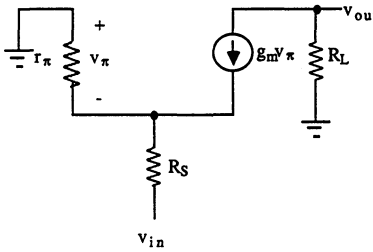

(b) Common-Base

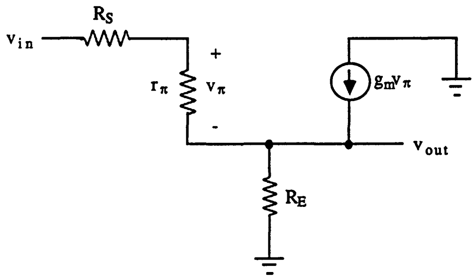

(c) Common-Collector (Emitter Follower)

FIG. 24. Incremental models of amplifier topologies.

the *differential amplifier* of Fig. 25. For perfectly differential (antisymmetric) input signals this amplifier may be regarded as a pair of common-emitter amplifiers. This connection has value because all practical amplifiers exhibit a variety of drifts with time and temperature. Drifts (and noise) that affect both devices in the same way are effectively suppressed by taking the output differentially. With some care thermal drifts referred to the input of approximately 1 μV/°C are routinely achievable, and with extraordinary care 50 nV/°C has been achieved. This performance is a substantial improvement over the basic −2 mV/°C drift that a typical common-emitter amplifier exhibits by itself. For these reasons the differential amplifier forms the core of a great many practical circuits, including the first stage of almost every operational amplifier.

4.2 Biasing

Implicit in much of the discussion thus far is some means to establish an operating point. This function is explicitly provided by bias circuitry, and takes a number of forms as shown in Fig. 26. In Fig. 26(a) bias is established by a voltage divider whose tap drives the base terminal, and by a resistor inserted in the emitter circuit. In practice the quiescent voltage at the emitter is often chosen a minimum of 2 or 3 V to suppress variation in collector current with base-emitter voltage variation. The resistors in the voltage divider at the base are chosen low enough in value to avoid modulation of the base voltage by variations in base current (such as might be caused by β variation), but high enough in value to avoid excessive loading of the previous stage.

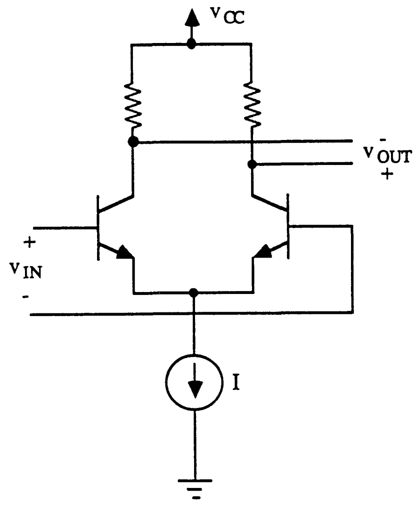

FIG. 25. Differential amplifier.

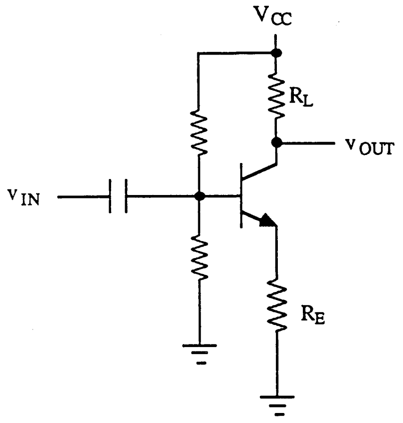

a) Classic Bias Scheme

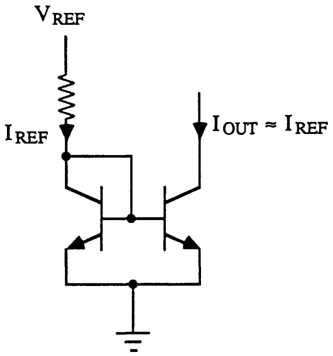

b) Current Mirror

FIG. 26. Bias circuits.

When this type of bias circuit is used, coupling capacitors are generally required in series with the input and output (the differential amplifier is a notable exception to this rule). These capacitors are necessary to prevent the upset of bias voltages that may occur if stages were directly connected together. Coupling capacitors necessarily block dc, and such a bias circuit is thus not suitable for dc amplifiers. Additionally, since the required capacitances are usually many orders of magnitude larger than can be realized in integrated form, alternative means must be employed in integrated circuits (ICs).

The bias circuit shown in Fig. 26(b) is known as a current mirror and exploits the matched characteristics of two transistors on a given integrated circuit. If such matching is assumed then it follows that the two collector currents are equal since the base-emitter voltages are forced to be equal. Such a current mirror may be employed as a current source for a differential amplifier, for example.

4.3 Bandwidth Considerations

The explicit computation (by hand) of the 3-dB bandwidth of an arbitrary network is generally impractical for all but the simplest systems. For example, finding the bandwidth of a network comprising just four independent energy-storage elements may require finding the roots of a fourth-order equation. Considering that the incremental model for a transistor includes two capacitances, we see the difficulty in extracting these roots if we use more than, say, just two transistors in a circuit.

While it is certainly true that machine computation can circumvent this limitation in the analysis of networks, it is less helpful in their design.

The method of open-circuit time constants (OCTs) allows one to estimate the bandwidth of a system almost by inspection, although at the expense of some accuracy. Because this method often yields useful answers with relatively little effort one is often willing to accept the approximate nature of those answers. Most importantly, and unlike typical circuit simulation programs, OCTs identify which elements are responsible for bandwidth limitations. The great value of this property in the design of amplifiers hardly needs expression.

To begin development of this method, consider all-pole transfer functions only. Such system functions may be written as follows:

$$\frac{V_o}{V_i}(s)=\frac{a_o}{(\tau_1 s+1)(\tau_2 s+1)\cdots(\tau_n s+1)}, \tag{60}$$

where the various time constants may or may not be real.

Multiplying out the terms in the denominator leads to a polynomial we shall express as

$$b_n s^n+b_{n-1}s^{n-1}+\cdots+b_1 s+1, \tag{61}$$

where the coefficient b_n is simply the product of all of the time constants and b_1 is the sum of all of the time constants. (In general, the coefficient of the jth term is computed by forming all unique products of the n time constants taken j at a time and summing all $n-j+1$ such products.)

Now, near the -3-dB frequency, the first-order term typically dominates over the higher-order terms so that (perhaps) to a reasonable approximation

$$\frac{V_o}{V_i}(s)\approx\frac{a_o}{b_1 s+1}=\frac{a_o}{(\sum_{i=1}^{n}\tau_i)s+1}. \tag{62}$$

The bandwidth of the original system in radian frequency as estimated by this first-order approximation is then simply the reciprocal of the effective time constant:

$$\omega_h\approx\frac{1}{b_1}=\frac{1}{\sum_{i=1}^{n}\tau_i}. \tag{63}$$

It may be shown that this bandwidth estimate is conservative in the sense that the actual bandwidth will always be at least as high as estimated by this method.

So far, it has only been shown that a first-order estimate of the bandwidth is possible if one is given the sum of the pole time constants. Unfortunately, such information is almost never available in this explicit form.

This limitation can be circumvented by relating the desired time-constant sum to (more or less) easily computed network quantities. Specifically, consider an arbitrary linear network comprising only resistors, sources (dependent or independent), and m capacitors. Compute the effective resistance R_{jo} facing the jth capacitor with all of the other capacitors removed (open-circuited), then form the product $t_{jo}=R_{jo}C_j$ (the subscript "o" refers to the open-circuit condition). Finally, sum all such "open-circuit" time constants. Remarkably, this sum is in fact *precisely* equal to the sum of the *pole* time constants used earlier. Thus at last, we have

$$\omega_{h,\mathrm{est}}=\frac{1}{\sum_{j=1}^{m}R_{jo}C_j}. \tag{64}$$

The method of OCTs is relatively simple to apply because each time-constant calculation involves the computation of a single pure resistance, although one must be wary of the impedance-modifying potential of dependent sources (such as the transconductance of a transistor model). In any event, the amount of computation required is typically substantially (indeed, often even fantastically) less than that needed for an exact solution.

The greatest value of the technique lies in its identification of those elements implicated in bandwidth limitations, that is, those elements whose associated open-circuit time constants dominate the sum. This knowledge can guide modifications to circuit values or even suggest wholesale topological changes. In contrast, typical simulators only provide a numerical value for the bandwidth while conveying little or nothing about what can be done to alter performance in a desired direction.

The origin of this property of OCTs may be regarded intuitively as follows. The reciprocal of each *j*th open-circuit constant is the bandwidth that the circuit would exhibit if that *j*th capacitor were the *only* capacitor in the system. Thus each time constant represents a local bandwidth degradation term. The method of OCTs effectively states that a linear combination of these individual, local limitations yields an estimate of the total bandwidth. The design value of OCTs derives directly from the identification and quantification of the local bandwidth "bottlenecks."

One must be careful not to place too much faith in the ability of OCTs to provide accurate estimates of bandwidth, however. This situation should hardly be surprising in view of the rather brutal truncation to first order of the denominator polynomial. However, there are numerous conditions under which OCT estimates are fairly reasonable.

It should be clear that an OCT bandwidth estimate is, in fact, *exact* for a first-order network since no truncation of terms is then involved. Similarly, OCT estimates will be quite accurate if a network of higher order happens to be dominated by one pole (that is, one pole is much lower in frequency than all of the other poles). There are many systems of practical interest, such as operational amplifiers, that are designed to have a dominant single pole.

Unfortunately, there are so many conditions under which OCTs give terribly pessimistic estimates that some caveat is necessary. For example, complex poles quite commonly arise in the design of wideband multistage amplifiers. The presence of complex poles upsets OCT estimates because OCTs effectively ignore the imaginary part of the pole time constants since complex poles must appear in conjugate pairs. The act of summing as required by the OCT method thus leaves only the real part. However, the true bandwidth of systems *does* depend on both the real and imaginary parts. The larger the imaginary part of the time constant, the greater the actual bandwidth and the greater the error. In fact, it is not uncommon for OCTs to estimate bandwidths that are one-half (or less) that actually observed.

Since it is generally not possible to tell by inspection of a network if complex poles are going to be an issue, one must always keep in mind that the primary value of OCTs is in identifying those portions of a circuit that control the bandwidth, rather than in providing accurate bandwidth estimates. Circuit simulators take care of the latter task.

Although application of open-circuit time constants is reasonably straightforward, there are one or two final issues that deserve consideration. An extremely important idea is that not all capacitors in a network belong in the OCT calculations. We have seen that the method of open-circuit time constants is an extremely valuable tool for designing amplifiers for good dynamic performance mainly because of its ability to identify the problem areas of the circuit. Because of the tremendous insights gained with extremely modest effort we are generally willing to overlook its quantitative limitations such as the often highly conservative nature of the estimated bandwidth. As long as we take care to apply the method only to models that apply to the high-frequency regime, we are assured of reasonable answers.

Concluding, the method of open-circuit time constants is an indispensable guide in the design of amplifiers. With it one can design intelligently and confidently to satisfy a given bandwidth specification. To be sure, the method has its quantitative shortcomings, but the valuable intuition provided is usually more than sufficient compensation.

5. DATA CONVERTERS

As technology moves increasingly toward digital representation of data, one is confronted with the fact that the world is analog in nature. In this section we examine analog-to-digital (A/D) and digital-to-analog (D/A) converters.

5.1 Analog-to-Digital Converters

Without question A/D conversion is more difficult than D/A conversion. The tradeoffs among complexity, resolution, and conversion speed tend to be somewhat severe, certainly more so than in D/A conversion. The great variety of A/D converter topologies is a consequence of this situation.

Since an analog signal spans a continuum of amplitudes, and since a digital representation is necessarily finite, an A/D converter maps a continuous range of signal amplitudes into a discrete digital representation. Given that this action thus irrevocably discards certain information about the amplitude of the original signal, one must determine how many bits are sufficient from a consideration of the requirements at hand. Additionally, we need to develop a criterion that identifies required conversion speeds.

One can derive some rough guidelines regarding resolution if certain simplifying assumptions are made. Referring to Fig. 27, we see that the quantization error has a uniform probability density function (PDF) for a ramp input (Δ is the size of the minimum quantum the converter can resolve). If we assume that this PDF also applies for a sine-wave input, the root mean square (rms) signal-to-noise (S/N) ratio thus becomes $(6N+1.8)$ decibels (dB), where N is the number of bits used in the representation. Hence, an eight-bit A/D converter can provide a nearly 50-dB rms S/N ratio.

There is a theorem due to Nyquist that determines how fast one must sample a wave form. Consider a signal that is band limited (i.e., its spectrum has zero energy beyond some frequency). Nyquist's sampling theorem states that one must sample the signal at more than twice the frequency of the highest-frequency component of the signal. If this sampling is performed with infinite amplitude resolution, the original wave form can in principle be recovered perfectly using an ideal, interpolating ("brickwall") filter. Thus a signal that possesses energy up to, say, 4 MHz (e.g., a broadcast television signal) must be sampled at least at an 8-MHz rate. Furthermore, since ideal filters are unrealizable, one must actually sample at a somewhat higher rate (e.g., 10 MHz) to accommodate this lack.

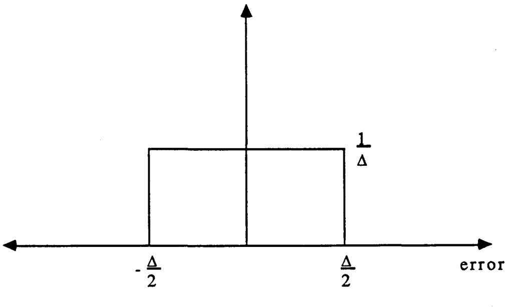

FIG. 27. PDF of error for ramp input.

Although various exotic technologies exist for performing high-speed analog-to-digital conversion (e.g., electrooptical and electron-beam methods), this review concerns circuit techniques only.

In principle, the fastest A/D conversion method is the parallel or "flash" method (Fig. 28). To perform an N-bit conversion, the analog input is presented simultaneously to an array of $2^N - 1$ comparators. The threshold of each comparator is derived from a voltage divider driven by a reference voltage. Thus the output of the comparator whose threshold is set just below the value of the input signal is high, as are the outputs of all the comparators underneath. The resulting "thermometer code" is sent to a digital encoder to provide an N-bit binary output.

This converter is fast because of its parallel nature. The conversion time is simply the sum of the delay of a single comparator and the delay of the encoding logic. Comparators with six-bit resolution and speeds in excess of 400 MHz are available, illustrating the high speeds attainable with this method. The greatest resolution achieved so far is 10–11 bits.

Unfortunately, the exponential dependency of the number of required comparators on the value of N limits the practicality of this approach for N larger than about 10. In addition, the substantial and nonlinear input capacitance associated with such a large number of comparators presents a nontrivial drive problem.

When greater resolution is required than can comfortably be provided by flash converters, a number of techniques can be used, most of which also ameliorate the drive problem.

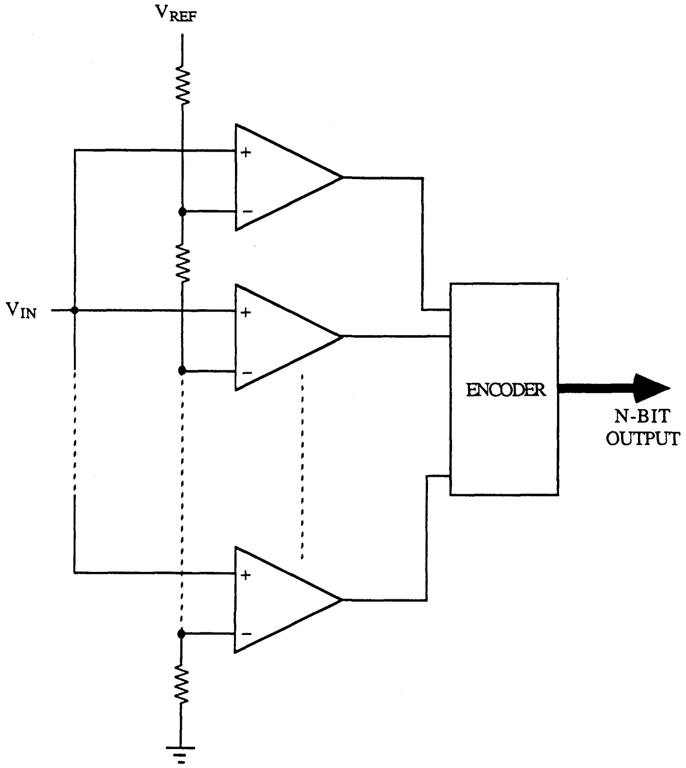

FIG. 28. Flash converter.

Many of these methods employ flash converters as components within the overall converter, and these consequently can digitize at high speeds. One such method, the dual-rank or subranging technique, is illustrated in Fig. 29. Here two flash converters of relatively low resolution combine to realize a system possessing higher resolution than either of the two-component A/D converters alone. The first-rank converter coarsely digitizes the input signal to M bits. This M-bit output feeds an N-bit DAC (digital-to-analog converter) which possesses at least the accuracy desired of the overall converter (N bits). The analog output of the DAC is subsequently subtracted from the input, and this residue (the "subrange") is converted by the second-rank ADC, which possesses $(N-M+1)$-bit resolution. The outputs from both internal ADCs are then combined to yield the final N-bit output.

A practical conversion system employing this type of converter generally requires the use of a sample-and-hold (S/H) stage to guarantee that the input signal remains stable during both steps of the conversion. Although a S/H stage is critical for the dual-rank converter, it also reduces conversion errors in virtually all other converters as well.

If somewhat lower conversion speeds can be tolerated, the successive approximator (S/A) shown in Fig. 30 can be used. As can be seen, the S/A converter comprises a comparator, a DAC, and some specialized digital logic known as a successive-approximation register (SAR).

Upon receipt of a "start convert" command, the SAR makes an initial estimate by setting the most-significant bit (MSB) high and all other outputs low. The comparator output remains high if this guess is less than the digital equivalent of the input. The MSB is then known to be high and remains on; otherwise it is turned off. The other bits are tested in succession in this manner. An end-of-conversion (EOC) signal is asserted upon completion of the procedure, signaling to the external world that the digital output of the SAR is valid. In general, then, we see that at least N clock cycles are required to complete an N-bit conversion using the S/A method.

Clearly, the input must remain stable during the conversion interval or gross errors will

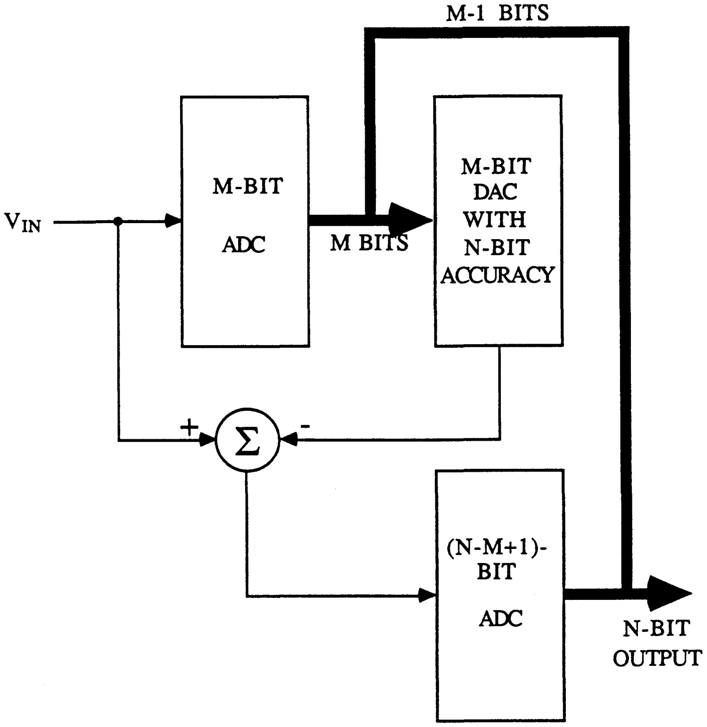
M-1 BITS
M-BIT
ADC
M BITS
M-BIT
DAC
WITH
N-BIT
ACCURACY
V_{IN}
+
Σ
-
(N-M+1)-
BIT
ADC
N-BIT
OUTPUT

FIG. 29. Dual-rank converter.

result. Hence, a S/H stage must precede the converter proper.

In some instances the analog input signal varies quite gradually over a conversion interval. In such cases the tracking or servo A/D converter shown in Fig. 31 may be used. Here,the difference between the analog input and an analog version of the digitized value is quantized to one bit with a comparator. An up/down counter is either decremented or incremented depending on the sign of the comparator's output.

The tracking converter performs well as long as the input changes by less than one least-significant bit per clock pulse, but fails miserably if the input signal changes by a large amount. In the worst case, $2^N - 1$ clock cycles are required to catch up to a full-scale

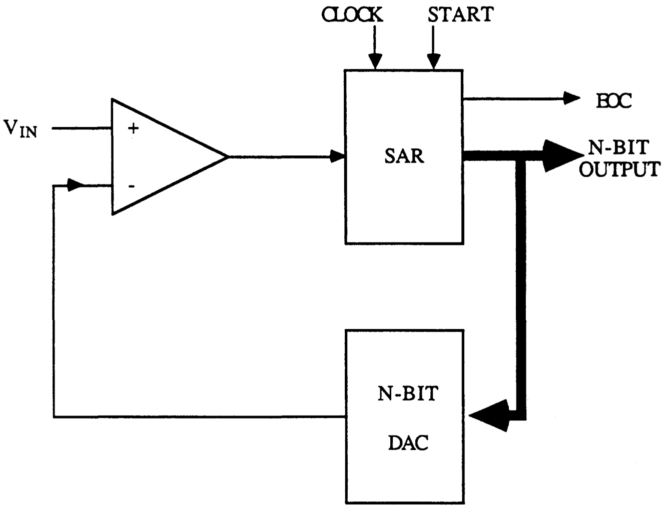
CLOCK
START
EOC
V_{IN}
+
-
SAR
N-BIT
OUTPUT
N-BIT
DAC

FIG. 30. Successive approximator.

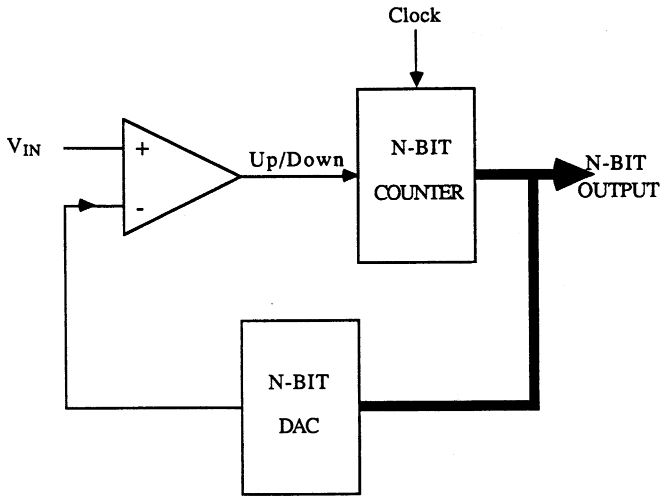

FIG. 31. Tracking converter.

step change in the input signal. Hence, even though this converter provides faster tracking performance than the S/A converter, the worst-case performance is much worse.

When high speed is not critically important, but high resolution is, the dual-slope A/D converter shown in Fig. 32 is widely used. In this method, a unipolar input signal is connected to a previously zeroed integrator for some fixed amount of time. Then a reference of the opposite polarity is connected to the integrator at the same time a counter is enabled. The counter is stopped as soon as the output of the integrator crosses zero. The counter's value is proportional to the input signal.

Because the same components (particularly the resistor and capacitor) are used in both the up and down portions of the integration, they only need to be stable over one conversion cycle; slow drifts in the characteristics of these elements can be tolerated. With the addition of special circuitry to compensate for inevitable system offsets, resolution in excess of 12–16 bits is routinely achievable. It is primarily for these reasons that digital

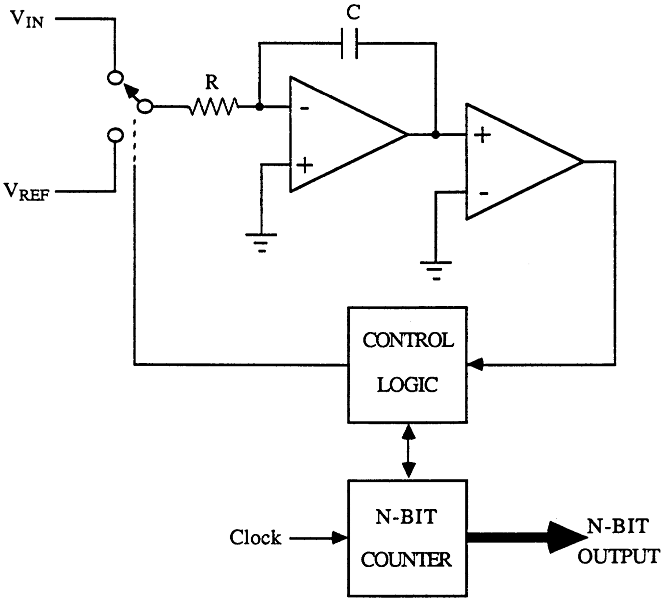

FIG. 32. Dual-slope converter.

voltmeters use this technique almost exclusively, although the conversion speeds tend to be quite slow (e.g., 100 Hz or less).

Another attractive property of this type of converter is that frequencies at integral multiples of $1/T$ are rejected. Thus by a judicious choice of the integration time T, the influence of certain periodic interfering signals (e.g., 60-Hz power line noise) can be greatly reduced.

5.2 Digital-to-Analog Converters

Virtually all modern DACs depend on the controlled summation of binary-weighted current sources. If a voltage output is desired (as is often the case) this current is simply made to flow through a suitable resistance.

Figure 33(a) shows how a transistor may be connected as a current source. By establishing a voltage on the base, the emitter voltage is also roughly fixed. Since the emitter voltage is the voltage across the resistor, the emitter current is fixed. Finally the collector current and emitter current are nearly equal if the transistor possesses a high value of β.

Figure 33(b) shows one popular method of combining current sources in a binary-weighted fashion. Under the control of the digital input, the various current sources are activated or deactivated as appropriate, providing an output current proportional to the digital input. This particular realization is amenable to realization in integrated circuit form because only current sources weighted in integer ratios are required. Thus one may exploit the excellent matching of components that monolithic construction offers.

A substantial side benefit of implementing DACs in a current mode is that the conversion speeds can be quite impressive. Because speed is usually limited in most electronic systems by parasitic capacitances rather than inductances, switching currents rather than voltages results in high-speed operation. As an example, many current-mode DACs offer conversion times as low as a few nanoseconds, while the best voltage-mode DACs require 100 ns or more.

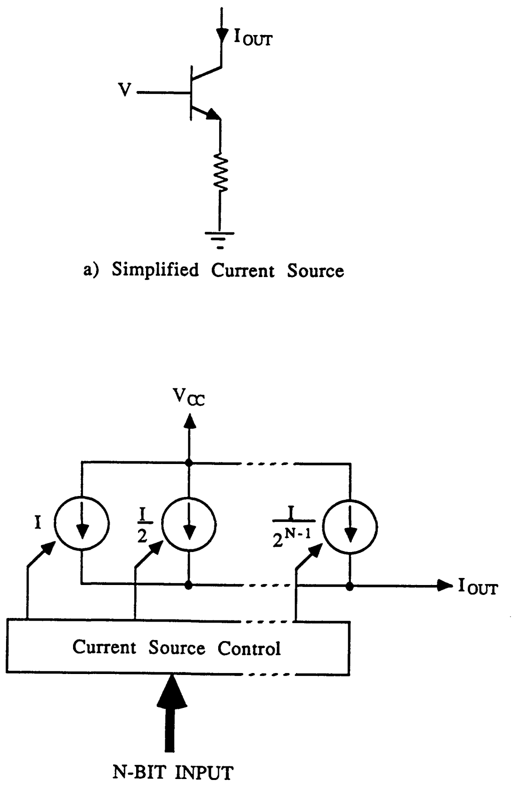

a) Simplified Current Source

b) Simple DAC

FIG. 33. Current-mode DAC.

6. MISCELLANEOUS APPLICATIONS

Because analog circuit design spans such a vast area it is simply not possible to provide anything resembling a representative sampling of applications. In this section we examine just three types of circuits that have not been covered earlier in this article.

Figure 34 shows a *switching power supply* in a general form. Switching supplies are used when it is necessary to convert voltages (or currents) from one value to another with high efficiency. The key to this topology's efficiency is the use of a switch to perform the actual modulation of power from source to load. Since a perfect switch does not dissipate energy when it is open (because the current is zero) or closed (because the voltage across it is zero), modulating the time the switch is on allows power to flow to the load without loss.

Practical realizations of this type of power supply employ transistors as the switching elements. Although transistors are not ideal switches, conversion efficiencies in excess of 90% are routinely achievable in many cases.

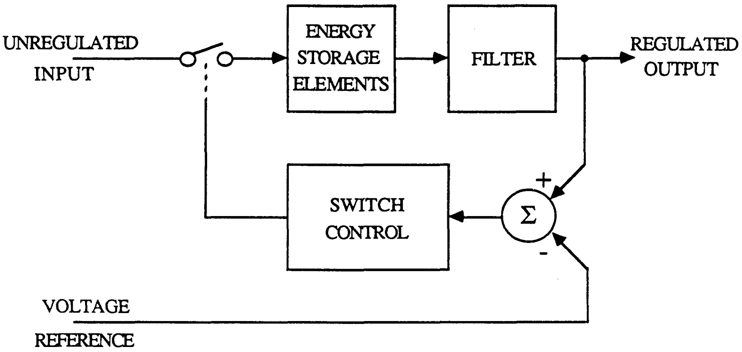

FIG. 34. Switching converter model.

Figure 35 shows a method for exploiting the temperature dependency of bipolar transistors. If one assumes that the two transistors are identical but are operated at a precisely known ratio of currents, then one may compute the difference in base-emitter voltages as follows:

$$\Delta V_{\mathrm{BE}}=\frac{kT}{q}\ln\left(\frac{I_1}{I_2}\right). \qquad (65)$$

Thus this circuit provides an output voltage that is proportional to absolute temperature (PTAT). With care practical circuits of this type readily conform to this law over a wide temperature range [e.g., (−55)–125 °C] to within tenths of a percent.

The last example, shown in Fig. 36, is a *band-gap voltage reference.* Such a circuit is useful for generating a stable, temperature-independent, known voltage for a variety of applications. Its operation exploits the detailed nature of I_S, the saturation current. A quasiempirical expression for the saturation current is

$$I_S=ABT^r\exp\left(-qV_{Go}/kT\right), \qquad (66)$$

where A is the base-emitter junction area, B is some constant, r is a process-dependent constant (usually between 2 and 5), and V_{Go} is the band-gap energy, extrapolated to 0 K, expressed in volts. With this expression one may show that the base-emitter voltage of a transistor operated at constant collector current exhibits a nearly constant temperature coefficient of about −2 mV/°C. When such a negative-temperature-coefficient voltage is combined in the right proportion with a positive-temperature-coefficient voltage (such as that produced by the thermometer circuit of Fig. 35) a stable voltage results. It may be shown that this voltage is nearly the band-gap voltage (roughly 1.2 V for silicon devices).

In Fig. 36, Q_1 and Q_2 are operated at a constant 2:1 ratio of collector currents by action of the op-amp and associated feedback

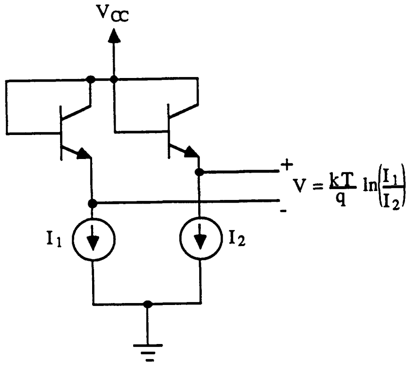

FIG. 35. Thermometer.

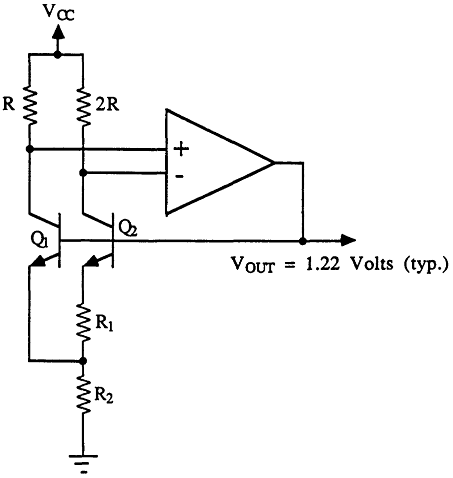

FIG. 36. Band-gap voltage reference.

loop. Hence their base-emitter voltages possess a PTAT component. The current through the resistors in the emitter legs is thus PTAT, and this PTAT voltage drop is added to the base-emitter voltage in just the right proportion so that the two temperature coefficients cancel. The output voltage is typically 1.22 V, and usually varies less than a few millivolts over a range of (−55)–125 °C.

GLOSSARY

BIBO Stability: A system is said to be stable in the bounded-input, bounded-output (BIBO) sense if every bounded input produces a bounded output.

Characteristic Equation: The denominator of the transfer function $H(s)$ whose roots are the poles of $H(s)$.

Crossover Frequency: The frequency at which the negative of the loop transmission [that is, $a(s)f(s)$] of a feedback system has a phase of $-180°$.

Desensitivity: The factor by which any incremental changes in the forward path are suppressed at the output. It is equal to $1 + a(s)f(s)$.

Frequency Response: The magnitude and phase of a transfer function $H(j\omega)$ under steady-state sinusoidal excitation at an angular frequency ω.

Gain Margin: The factor by which the loop transmission may be increased before unstable behavior results.

Incremental Model: One which relates *changes* in excitation to *changes* in response, rather than their absolute values.

Loop Transmission: The gain around a feedback loop. It is equal to $-a(s)f(s)$.

Phase Margin: The amount of negative phase shift that may be added to the loop transmission at the crossover frequency before unstable operation results.

Pole: Any value of complex frequency s that results in an infinite value for a transfer function $H(s)$. These frequencies are also known as natural frequencies or eigenvalues.

Thermal Voltage: The energy kT expressed as a voltage $(=kT/q)$. It has a value of about 25 mV at 300 K.

Zero: A value of complex frequency s for which the transfer function $H(s)$ is zero.

List of Works Cited

Roberge, J. K. (1975), *Operational Amplifiers: Theory and Practice,* New York: Wiley.

Further Reading

Gray, P. R., Meyer, R. G. (1984), *Analysis and Design of Analog Integrated Circuits,* New York: Wiley.

Roberge, J. K. (1975), *Operational Amplifiers: Theory and Practice,* New York: Wiley.

CIRCUITS, DIGITAL

THOMAS H. LEE, *Tewksbury, Massachusetts, U.S.A.*

	Introduction	507
1.	**Boolean Algebra**	507
2.	**Karnaugh Maps**	509
3.	**Combinational Circuits**	510
4.	**Sequential Circuits**	513
5.	**Memory Circuits**	515
6.	**A Survey of Digital Logic Families**	516
6.1	Transistor–Transistor Logic (TTL)	516
6.2	Metal-Oxide-Semiconductor (MOS) Logic	517
6.3	Emitter-Coupled Logic (ECL)	519
6.4	Gallium–Arsenide (GaAs) Logic	520
7.	**Some Representative Applications**	521
8.	**Fundamental Performance Limitations**	522
	Glossary	524
	Works Cited	524
	Further Reading	524

INTRODUCTION

The processing of information by digital circuits and systems has become ubiquitous, primarily because the manipulation of information in digital form offers computational repeatibility and freedom from distortion during transmission and replication, in contrast with the relatively noise-, distortion-, and otherwise error-prone behavior that is typically associated with analog systems. Furthermore, the existence of mathematical formalisms for the design of digital circuits allows algorithmically based automated design of such circuits and systems, while automated design of analog circuits remains largely an elusive goal.

This article provides an overview of digital electronics and begins with a summary of the rules of Boolean algebra in Sec. 1, followed by a survey of a variety of "combinational" logic functions in Sec. 2. Section 3 then examines an important class of circuits that possess memory, while Sec. 4 focuses on the memory elements themselves. Following this macroscopic examination, Sec. 5 presents the electrical (as opposed to logical) characteristics of a number of common logic "families." A survey of representative applications follows in Sec. 6, and Sec. 7 concludes the article with a discussion of some of the fundamental issues associated with digital computation.

1. BOOLEAN ALGEBRA

Digital, or "logic," circuits process information according to rules of mathematical logic developed by Augustus De Morgan (1806–1871) and George Boole (1815–1864). In particular, Boole's *An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities*, published in 1854, defines rules by which logical statements, expressed in an abstract symbolic form, can be manipulated. These rules comprise what is now known as Boolean algebra.

Boole's accomplishment went largely unnoticed for quite some time. In fact, nearly sixty years elapsed before Alfred North Whitehead and Bertrand Russell proved the essential equivalence of mathematics and logic in their *Principia Mathematica* (1910–1913), acknowledging Boole's significant contribution to this end.

Boolean algebra subsequently was adopted as a tool primarily by semanticists. It was not until 1937 that then-MIT graduate student

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Claude Shannon recognized that the discrete nature of electrical relay circuits was particularly suited to analytical treatment by Boolean methods. His ground-breaking paper, "A Symbolic Analysis of Relay and Switching Circuits," introduced symbolic logic into the domain of electronic engineering and thus may be said to mark the birth of digital electronics.

We define *Boolean functions* as those that map n binary input variables into a single binary output. Common practice is to assign values of "1" and "0" (or "true" and "false" or "high" and "low") to these variables. As a consequence of the finite, discrete nature of the mapping, any Boolean function can be completely defined by a table (known as a "truth table") which specifies this mapping. Since each variable can only take on a finite number of values (two), there is only a finite number of possible truth tables, and thus there is only a finite number of distinct Boolean functions of a particular number of variables. Consider Boolean functions of two variables, for example. There are four possible combinations of input variables, and there are therefore 2^4 possible output mappings. Generalizing from this example we see that a Boolean function of m variables possesses $n=2^m$ possible input states and that there are therefore 2^n possible Boolean functions of these n states. We call digital circuits that implement Boolean functions *combinational* circuits.

Some Boolean functions occur so frequently (and are fundamental in some sense) that they have been given mnemonic names. One such function is the AND function, which has the property that its value is 1 only if all of the inputs are 1. We will represent this function in operational form as follows: $Y=AB$, which means "the output Y is 1 only if inputs A *and* B are both 1."

Another common and useful function is the OR function, $Y=A+B$ ("the output Y is 1 if either input A *or* input B is a 1"). It is somewhat unfortunate that the plus sign denotes both the logical OR function as well as common addition, but this usage is standard.

Yet another useful function is the (unary) NOT function, whose output is the logical inverse (or complement) of the (single) input. The NOT function is usually denoted by an overbar ($\bar{A}$).

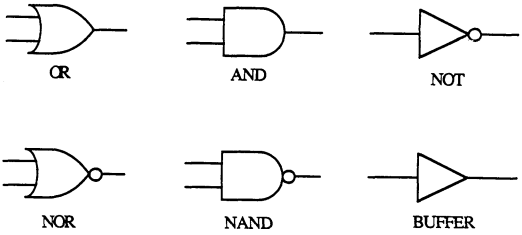

FIG. 1. Basic logic gates.

The circuit symbols for the AND, OR, and NOT operators are shown in the top row of Fig. 1. Note that the addition of a small circle to either input or output is a schematic shorthand notation for inversion—it is equivalent to adding a NOT function at the indicated node, so that the three lower symbols represent NOR (NOT+OR), NAND, and buffer (output=input) functions.

The following list is a collection of axioms, postulates, theorems, and otherwise useful properties of Boolean algebra:

Commutativity:

$$AB=BA, \tag{1a}$$

$$A+B=B+A. \tag{1b}$$

Associativity:

$$ABC=(AB)C=A(BC), \tag{2a}$$

$$A+B+C=(A+B)+C=A+(B+C)\ . \tag{2b}$$

Distributivity:

$$A(B+C)=AB+AC, \tag{3a}$$

$$A+BC=(A+B)(A+C). \tag{3b}$$

(Logical) Sums:

$$A+0=A. \tag{4}$$

$$A+A=A. \tag{5}$$

$$A+\bar{A}=1. \tag{6}$$

$$A+1=1. \tag{7}$$

$$A+AB=A. \tag{8}$$

$$A+\bar{A}B=A+B. \tag{9}$$

(Logical) Products:

$$A0=0. \tag{10}$$

$$AA=A. \tag{11}$$

$$A\bar{A}=0. \tag{12}$$

$$A1=A. \tag{13}$$

$$A(A+B)=A. \tag{14}$$

$$A(\bar{A}+B)=AB. \tag{15}$$

DeMorgan's Theorem:

$$\overline{A+B}=\bar{A}\,\bar{B}. \tag{16a}$$

$$\overline{AB}=\bar{A}+\bar{B}. \tag{16b}$$

With the use of these identities (which clearly are not all independent) it is possible to express *any* Boolean function as either a (logical) sum of (logical) products or as a product of sums. Thus the OR, AND, and NOT functions are sufficient to express any Boolean function. Actually, we can make an even stronger statement since DeMorgan's theorem shows that the AND and OR operators are not independent; one can be derived from the other (with a suitable number of inversions). As a consequence, one can implement any Boolean function using only NOT and either OR or AND functions.This observation is more than simply of purely pedagogic value. The computers of the Apollo spacecraft were implemented entirely with only NOR circuits ("gates"), and this single type of gate was sufficient to get the astronauts to the Moon and back.

2. KARNAUGH MAPS

Although a truth table provides a complete description of a combinational logic function, it does not always present this information in a form that may be amenable to easy synthesis by available building blocks (e.g., AND or OR gates). The Karnaugh ('kar-no) map can aid synthesis by rearranging the truth table in a particular manner that allows one to write down, almost by inspection, a sum-of-products (or product-of-sums) Boolean equation for the logic function. It is then a straightforward (although not necessarily trivial) matter to devise a circuit to implement this desired equation.

The use of a Karnaugh map is best illustrated with a practical example. Let us suppose that we wish to design a circuit that performs arithmetic addition (as opposed to logical addition). Such a module takes three binary arguments (two addends and possibly a "carry in" signal from another like module) and produces two outputs (a sum and carry out). By connecting an appropriate number of such modules together, one can add two binary numbers of arbitrary length.

The truth table for this addition module is as follows:

Carry in	*A*	*B*	Sum	Carry out
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

Now, to construct a Karnaugh map for, say, the carry output, we rearrange the truth table for this function so that the input variables are represented along two axes, and so that only one input variable changes at a time in going from one entry to an adjacent entry:

AB *C* In	00	01	11	10
0	0	0	1	0
1	0	1	1	1

To form a sum-of-products expression from this map, we organize the cells that contain 1's into groups of rectangles that are as large as possible. Since only one input variable changes as we go from one cell to the next, any pair of adjacent cells (either vertically or horizontally) corresponds to a term in which one of the variables is irrelevant (missing). For example, the two cells in the third column of our map correspond to $A=B=1$ and $C=1$ or 0, that is, to the term AB. In general, any rectangle containing 2^n cells corresponds to a term in which n variables are missing from the Karnaugh map. By identifying the largest such rectangles comprising 1's in our map, we can rapidly read off a sum-of-products expression.

In our case we find three such rectangles (here identified with ellipses, as is common), each of which is 1 by 2 in size:

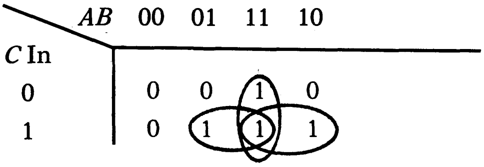

C In \ *AB*	00	01	11	10
0	0	0	1	0
1	0	1	1	1

We see that the cells identified in our map correspond to the terms *AB*, *BC*, and *AC*, where C=Carry In. Since the desired function is the union of all such terms, the corresponding Boolean equation is found by logically summing the terms corresponding to these rectangles. Thus, we may write the following expression for the carry output:

$$\text{Carry Out}=AB+BC+AC. \tag{17}$$

The advantages of using a Karnaugh map to simplify the identification of logical equations is limited in practice to systems of four or fewer input variables since the map becomes a volume (then a hypervolume, and so on) beyond four variables. One generally employs computational design aids for the more complex situations.

There are a few other fine points associated with the use of these maps. When trying to identify the largest rectangles containing 1's, recognize that the boundaries of the map are logically connected. Thus, 1's in the leftmost and rightmost columns (or top and bottom rows) would actually be part of one large rectangle.

Often in constructing these maps, one encounters situations in which there are irrelevant values for certain input combinations. These "don't care" states may be handled as 1's to encourage formation of the largest rectangles. However, one should not add extra logical terms in the sum-of-products expression simply to accommodate the added states; the "don't care" states should be used only to simplify the resulting maps. Thus these states should be marked with X's and treated as 1's at one's discretion.

It should also be clear that the resulting Boolean equation is not necessarily unique; one may choose to group the terms in several different, though logically equivalent, ways.

Finally, although the foregoing discussion has only mentioned the sum-of-products formulation of Boolean functions, one could just as well identify rectangular blocks of contiguous 0's. Such blocks represent the result of OR operations so that one can form a product-of-sums expression that is logically equivalent to the sum-of-products formulation.

3. COMBINATIONAL CIRCUITS

Since Boolean functions expressed in a sum-of-products (or product-of-sums) form explicitly use only AND, OR, and NOT operations, they do not necessarily lead one directly to practical circuit implementations. As can easily be imagined, the design of complex digital systems would be quite cumbersome if only those three primitive functions were available (the decision to use only NOR gates in the Apollo computer was the result of reliability considerations). Fortunately, the design of digital circuits is simplified by the great variety of logic functions available in integrated circuit form, some of which we now survey.

The least sophisticated of these functional blocks are simple combinations of the basic gates. An example (see Fig. 2) is the exclusive-OR (XOR) function. This function produces a high output only if the inputs differ, and thus the XOR can form the core of modules that perform addition.

Multiplexers (Fig. 3) route one of n signal inputs to a single output under the command of a control input. For convenience, many multiplexers offer both an output and its inverse. Aside from the conventional routing

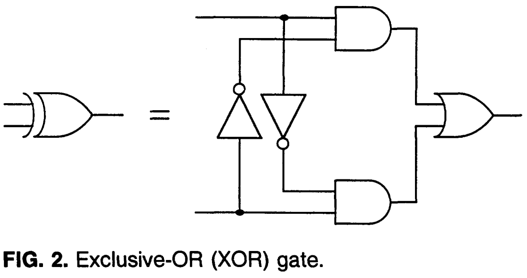

FIG. 2. Exclusive-OR (XOR) gate.

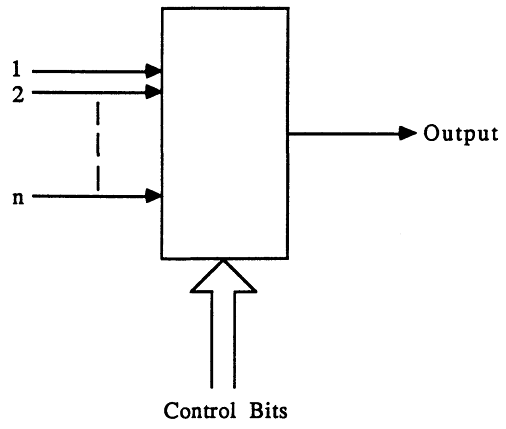

FIG. 3. Multiplexer.

capabilities, an *n*-input multiplexer can synthesize any Boolean function specified by a truth table of *n* entries (even of 2*n* entries, with the addition of, at most, one inverter), if one treats the control signals as the input variables. By connecting the conventional inputs to appropriate high or low states, arbitrary Boolean functions of the control inputs are realizable.

Demultiplexers (Fig. 4), also known as decoders, perform the complementary operation—they take a single input and route it to one of *n* outputs, again under the command of some control bits. Depending on the particular implementation, the unselected outputs may either assume a high-impedance state that approximates an open circuit, or they may be placed in some arbitrary inactive default state (e.g., high). In many cases the input is not an explicit one; the selected output simply assumes its active default state (e.g., low).

Priority encoders (Fig. 5) rank inputs and provide a binary output whose value corresponds to that of the highest-valued input that is asserted. In the example shown in Fig. 5, input number five of the 8-input encoder is the highest-valued input that is asserted (in this particular case, assume that "asserted" means "high"). Thus, the output of the encoder is five in binary form (101).

The priority encoder is particularly valuable in "interrupt-driven" systems in which several entities may vie for a processor's attention. By conveying the priorities associated with these various requests, the encoder allows the processor to order the sequence in which these interrupts will be serviced.

Parity generators/checkers are used in simple forms of error detection. When used as a parity generator, this circuit counts the number of 1's in the input bits, and provides two outputs, either of which may be used. One output, the "even" parity output, is high if there is an even number of 1's in the input, while an odd number of input 1's produces a high state on the "odd" parity output. One of these parity outputs is then sent out along with the original input bits. At the receiving end, an identical parity circuit computes the parity of the incoming data. This computation is compared with the parity bit that was transmitted with the original data. If an odd number of corruptions has occurred (the most likely distortion presumably being a single-bit error), the transmitted parity and that computed at the receiver will disagree, signaling the presence of an error so that one may then take appropriate action.

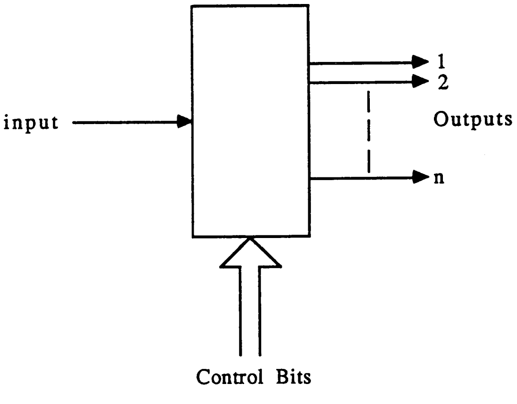

FIG. 4. Demultiplexer.

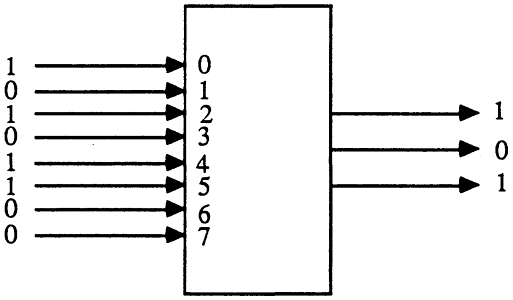

FIG. 5. Priority encoder.

Another important class of combinational circuits consists of those that perform arithmetic operations. *Adders* that typically take two 4-bit arguments and produce their arithmetic sum are available. Furthermore, an arbitrarily large number of these circuits may be combined to handle addends of arbitrary size.

Integrated *multipliers* are also available. These typically accept two 8-bit or 16-bit arguments and generate their arithmetic product.

Magnitude comparators take two arguments and provide outputs either indicating equality or identifying which is greater. As with adders and multipliers, these modules are expandable to permit comparison of inputs of arbitrary size.

Arithmetic logic units (ALUs) can perform a variety of arithmetical and logical operations. Arithmetic functions include addition, subtraction, and magnitude comparison, while logical functions include AND, OR, XOR, and inversion. Because of this smorgasbord of functions, ALUs are found at the heart of virtually all computers (including microprocessors). Finally, ALUs are expandable as well.

Read-only memories (ROMs) and *programmable logic arrays* (PLAs) allow the user to implement custom logic functions. Although ROMs will be discussed shortly in greater detail, it suffices to note that a ROM has m inputs (*address* lines) and n outputs (*data* lines), and that each of the n data outputs can be made an arbitrary Boolean function of the m address inputs. Thus ROM implementations can be superior to those of other combinational circuits, especially when the desired logic function possesses little structure. Furthermore, since n is typically greater than one (e.g., eight), a ROM can simultaneously provide several Boolean functions of a common set of input variables. This attribute is often significant.

It is sometimes necessary to realize a function that involves a greater number of input variables than ROMs can comfortably accommodate. Note that a large number of input variables can be problematic for ROMs since the number of memory cells required grows exponentially with m. In these instances PLAs and their variants (known variously in several commercial versions as PALs, for programmable array logic, or GALs, for generic array logic, etc.) are attractive alternatives. These devices provide custom sum-of-products implementations of one or more Boolean functions of a set of inputs.

In their most general incarnation, PLAs possess the structure shown in Fig. 6. An array of AND gates known as an *AND plane* allows the formation of the individual product terms via programmable interconnections, while a programmable *OR plane* similarly allows the formation of the sum terms. If there are recurrent product terms in the desired sum-of-products expression this structure may afford some economies over a ROM-based implementation because each product term needs to be computed only once.

Programming is accomplished with a special device and typically involves the blowing of microscopic integrated fuses to define the AND and OR connections. Once programmed in this way, such a PLA cannot be "unprogrammed." In other cases, programming relies on tunneling phenomena or hot-electron injection to transport charge onto/into the insulating control electrodes of transistors whose function is analogous to that of fuses in the previous case. Depending on the particular type of device, either tunneling or photoexcitation via ultraviolet light can return the electrons to their original state, allowing reprogramming. Permanent programming can also be performed at the time of manufacture of the integrated circuit. This last option is the most practical one for high-volume production.

As a final comment on this class of circuits, note that there are many PLAs offering a great

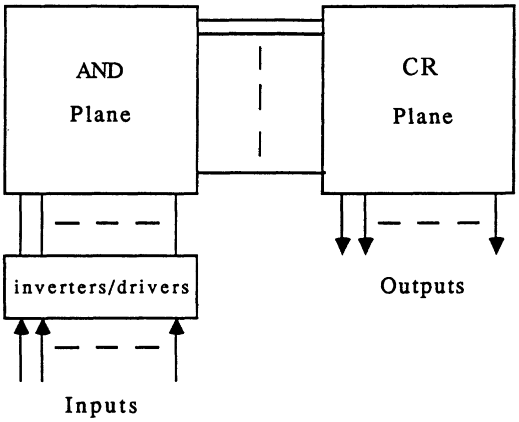

FIG. 6. Internal structure of a PLA.

variety in the number of inputs and outputs as well as in the internal architecture.

Up to now, we have implicitly assumed that logic circuits respond instantaneously. However, clearly all physical systems require some time to respond to changes in the values of input variables, with typical delay times on the order of one to ten nanoseconds for modern computational elements. The logic designer must allow for these delays, even for those associated with interconnecting wiring (approximately one nanosecond per foot of conductor is typical) as well.

Although physical circuits based on the methods described thus far will eventually produce the correct output after the inputs have remained stable for some suitable amount of time, it is quite possible to observe spurious outputs on a transient basis when and if the inputs change. These "glitches" may be harmful by themselves, or they may cause malfunctions in the circuits to which they may be connected. These undesired (and hopefully temporary) outputs are known as *hazards*, and their frequency of occurrence can be eliminated or at least reduced greatly by guaranteeing that all adjacent 1-filled rectangles of the Karnaugh map share terms; that is, that there are overlaps among the rectangles. Unfortunately, this remedial measure is not readily applicable to all forms of combinational logic. In particular, it is difficult to eliminate hazards in purely ROM-based combinational circuits since it is usually impractical to anticipate the many possible intermediate states through which the output may traverse. As will be seen in the next section, however, this problem may be solved quite easily by using an array of flip-flops on the outputs of the ROM. The outputs of the flip-flops are commanded to change only after the outputs of the ROM have settled, thus shielding the external world from hazards.

4. SEQUENTIAL CIRCUITS

The foregoing discussion of combinational circuits has necessarily concerned only functions whose present value is dependent solely on the present values of inputs; there has been no explicit mention of any temporal aspects of computing, except in the context of hazards and related practical issues.

By virtue of their time-invariant nature, combinational circuits lack a property that is crucial to the functioning of computers and virtually all other important systems—that of memory. The ability of a circuit to keep track of past inputs (and decisions) is directly responsible for the great variety of powerful behavior that computational devices (both digital and analog) exhibit.

Of necessity, digital circuits possessing memory cannot be constructed *as* combinational circuits. Interestingly, however, they can be constructed *from* combinational elements. We now consider *sequential* circuits, which can exhibit such behavior.

The crucial difference is the presence of feedback. As an example, consider the block diagram of Fig. 7. A set of signals (state variables) defines the present state of the circuit and is fed back to the input of the combinational circuitry. The combinational logic uses the old state information and the present value of the inputs to compute the next state.

As a practical example consider the circuit shown in Fig. 8 . The lower drawing emphasizes the similarity of this circuit, known as a reset-set (RS) *flip-flop*, to the model of Fig. 7. To analyze this circuit, note that if $R=1$ the output Q will be 0 independent of S or X. By symmetry, if $S=1$, then X will always be zero. As a consequence, Q will equal $\overline{R}$; the output is the inverse of the R input in this case.

If $S=R=0$, then $Q=\overline{X}$, that is, the present output equals the inverse of the state. Under this set of conditions, the output Q can be *either* 0 or 1. The actual value of Q cannot be determined by examining the static solutions to these Boolean constraints.

To deduce the output of this circuit when $S=R=0$, assume initially that $X=1$ and $Q=0$. If S subsequently changes to 1, X will change to 0 as noted earlier, and Q will change to 1. The set of conditions $R=X=0$ and $Q=1$ is

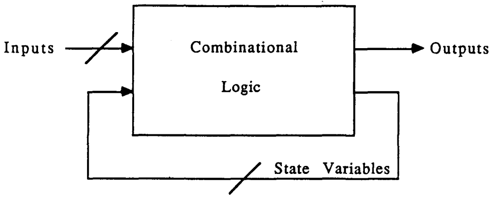

FIG. 7. Asynchronous sequential circuits model.

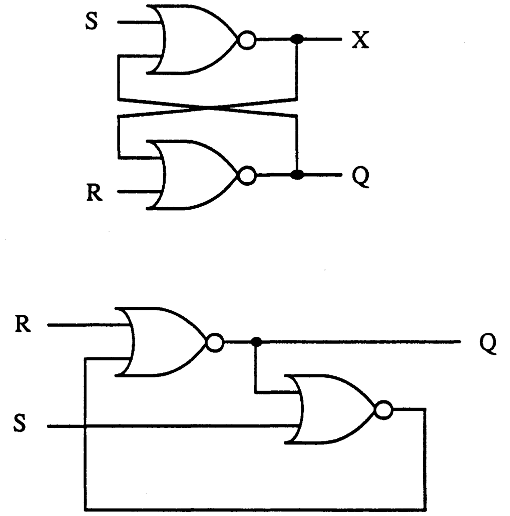

FIG. 8. RS flip-flop.

stable; no further changes will take place, even if S returns to 0.

Now assume that R changes to 1 while $S=X=0$ and $Q=1$ initially. The output Q will change to 0, causing X to change to 1. Again, the output does not change, even if R returns to 0. From these observations we conclude that the circuit remembers which of S or R was most recently in the high state. If S was high more recently, then Q will be 1 (set). If R was high most recently, then Q will be reset to 0. Finally, if $S=R=1$, Q takes on a value of 0 until S changes to 1 and $R=0$.

From the verbosity of the foregoing description, it is clear that there is a need for a more efficient means of describing the behavior of sequential circuits. The state transition diagram of Fig. 9 completely specifies the behavior of the RS flip-flop in a particularly economical way.

Let us define a state as some subset of stable (on a static basis) solutions to the Boolean constraints imposed by the gates and the associated connections, and represent it as a circle on our diagram. The equations inside each state circle define the outputs as Boolean functions of the inputs, while the labels on the arcs (which denote transitions) are the Boolean functions which define the conditions under which transitions take place.

In the RS flip-flop example shown in Fig. 9, there are two states, defined by the value of Q. When $Q=1$, it will remain so until R changes to 1. When $Q=0$, it will remain so until $S\bar{R}=1$.

In addition to the RS flip-flop, there are several other types of flip-flops (shown in Fig. 10). The *D flip-flop* has a single data input, a *clock* input (denoted by the wedge in the symbol), and often two outputs (Q and its inverse, as shown). When the clock input makes a transition (usually from low to high), the output assumes the value of the input at that time, and remains there until the next activating transition of the clock. This type of flip-flop is commonly used to "deglitch" the outputs of ROMs, as discussed earlier.

The *JK flip-flop* has two data inputs and a clock input. If the J and K inputs are inverses of each other, the output after an activating clock transition (e.g., high to low) is equal to the value of J at the time of the transition. If both inputs are low, the output does not change after a clock transition. If both inputs are high, the next output is the inverse of the present one.

A *latch* has a data input and a latch-enable input. As long as the enable signal is low, the output follows the input. When the enable is high, the output stays at whatever state pre-

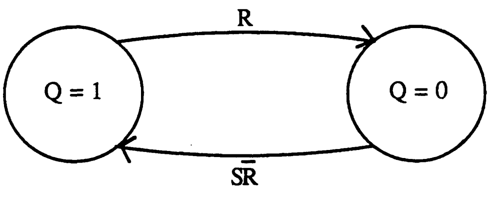

FIG. 9. State diagram for RS flip-flop.

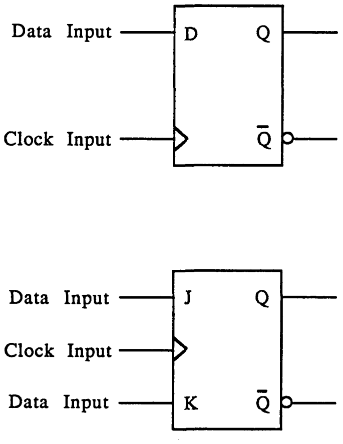

FIG. 10. Other flip-flops.

vailed at the time the enable signal made the transition from low to high.

The particular paradigm for sequential circuits shown in Fig. 7 is sometimes called the primitive or fundamental mode model, and is fraught with difficulties of a subtle nature. Note, for example, that state changes may take place whenever any input signal changes. Hence, state changes occur at unpredictable times, and may even be triggered spuriously by noise (from whatever source) on the input signals.

To reduce the susceptibility of digital systems to such noise-induced errors, and to enforce a discipline that encourages the design of systems that are amenable to tractable analysis, a modified sequential circuits model is almost universally used. Shown in Fig. 11, this model explicitly separates the storage of state information from the combinational logic. State information is stored in some memory (e.g., an array of flip-flops) at intervals controlled by an external clock. Because the state memory updates its contents only at times controlled by the clock, independent of the behavior of the inputs at other times, the likelihood of noise-induced errors is greatly reduced.

The design of such *synchronous* systems is straightforward, and begins with the construction of a state diagram appropriate to the function desired. Combinational logic is then designed to produce the desired outputs given a particular state and inputs, in addition to computing the next state and presenting this information to the input of the state memory. When the clock next updates the state memory, this new information is presented to the combinational logic, and on it goes.

Rounding out a survey of sequential circuits are *counters* and *shift registers*. Some counters can count both up and down, and both decimal and binary moduli are available. Many counters are also presettable so that counting may start from values other than zero.

Shift registers are merely collections of flip-flop chains. Inputs and outputs can be in serial or parallel form, depending on the particular type of shift register. The example in Fig. 12 shows, in simplified form, a serial-in/parallel-out type of shift register. Input data are clocked through a cascade of flip-flops, and the output data are taken as a group of *n* bits every *n* clock cycles. Such a circuit thus converts data in a serial format to data in a parallel format.

5. MEMORY CIRCUITS

The previous section examined a class of digital circuits that exhibit or employ memory. In this section, we survey some of the many types of circuits whose sole function is that of providing memory.

The two broad classes of memory circuits are *random-access memories* (RAMs) and *read-only memories* (ROMs), with various taxa within these two classes. The key distinction is that RAMs typically require power to retain their data while ROMs do not.

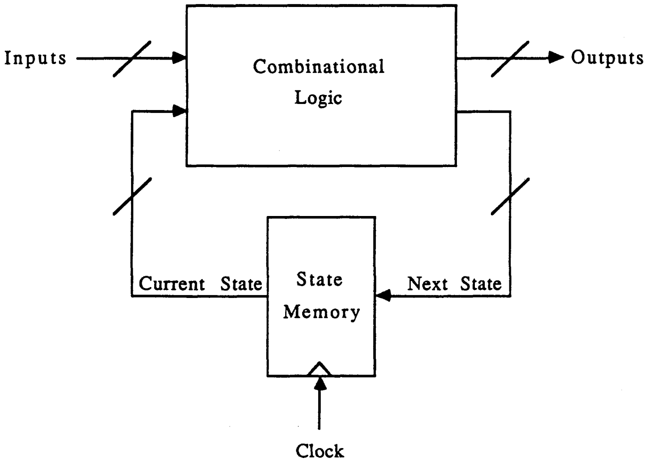

FIG. 11. Synchronous sequential circuits model.

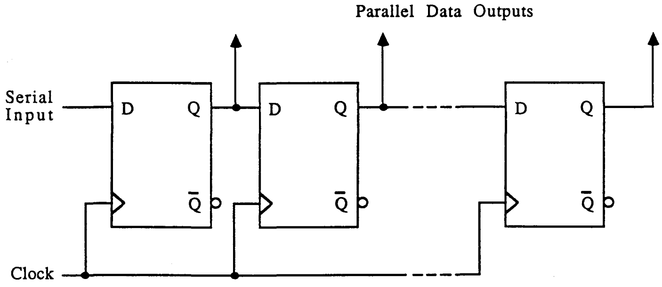

FIG. 12. Shift register.

As the basic static RAM's (SRAM's) memory cell is essentially a flip-flop (each usually comprising four to six transistors), they retain their data as long as power is applied; no further intervention is required. On the other hand, dynamic RAMs (DRAMs) typically represent data as charge stored on an array of capacitors. Because of unavoidable leakage effects, the stored charge cannot persist indefinitely. Thus DRAMs require periodic "refreshing" (e.g., every 2–4 ms) to restore the charge to full value. Despite this need, DRAMs are attractive because their simple internal structure (only one transistor and one capacitor is required per cell) permits a denser packing of memory cells than is possible with SRAMs. However, the need to charge and discharge substantial capacitances makes DRAMs somewhat slower than SRAMs.

Nonvolatile RAMs (NVRAMs or NOVRAMs) may employ magnetic materials to store data and are exceptions to the power requirement rule. These "bubble memory" devices are considerably slower (typically by approximately three orders of magnitude) than their purely electrical counterparts, however. Ferroelectric NVRAMs (and DRAMs) are making a tentative egress from the laboratory, though, and may provide performance that is significantly superior to that of bubble devices.

Read-only memories are found in greater variety than RAMs. *Mask-programmed ROMs* are factory-programmed memories intended for mass production and whose contents cannot be altered once the devices are fabricated, while the cells of *fuse-link PROMs* (programmable ROMs) can be programmed only once since the fusible links that determine the contents of the cells cannot be repaired once they are blown.

Ultraviolet-erasable programmable ROMs (UV-EPROMs) depend on tunneling or hot-electron injection for programming and UV exposure for erasure. The erase process is necessarily not selective since the UV radiation illuminates all cells simultaneously through a transparent window on the device.

Electrically erasable programmable ROMs (EEPROMs) typically exploit tunneling phenomena both to store and to remove charge, permitting completely electrical programming and erasure. Erasure is global in some EEPROMs (known as "flash" EEPROMs), while others allow the erasure of individual cells or groups of cells without disturbing the contents of other cells. Programming/erasure cycle times are steadily decreasing so that EEPROMs are suitable for use as NVRAMs in an increasing number of situations.

6. A SURVEY OF DIGITAL LOGIC FAMILIES

The presentation so far has concerned primarily the global issues associated with digital circuits. To extend the discussion to a more detailed level, we now survey the electrical characteristics and basic underlying circuitry of the dominant logic families.

6.1 Transistor–Transistor Logic (TTL)

The oldest surviving family of digital integrated circuits is TTL, named for the fact that all logic functions are performed entirely with bipolar junction transistors. This family of devices operates from a 5-V power supply, and recognizes any input voltage from 0 to 0.8 V as a valid 0, and any input from 2.4 to 5 V as a valid 1.

Figure 13 is a schematic diagram of the earliest TTL NAND gate. The most distinctive feature of this circuit is the use of a dual-emitter transistor. When a voltage within the valid low range is applied to either emitter, Q_1 turns on, withdrawing charge from the base

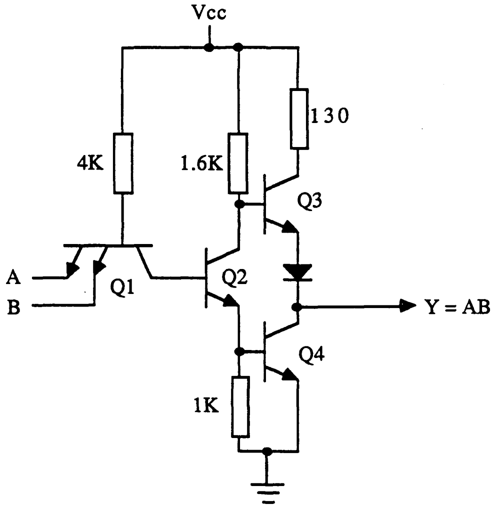

FIG. 13. Classic TTL NAND gate.

of transistor Q_2, causing that transistor to turn off. Transistor Q_2 is known as a phase-splitter since the voltages on its collector and emitter are 180° out of phase. As a consequence, Q_3 turns on while Q_4 turns off, causing the output to go high.

If both inputs are high, the base-collector junction of Q_1 allows current to flow into the base of Q_2 (conventional current is assumed) from the power supply through the 4-kΩ resistor. Since Q_2 is then turned on, Q_3 turns off while Q_4 turns on. The result is a low output. Since we have already deduced that a low at either input causes a high output, we see that this circuit in fact performs the NAND function.

Note that the source that drives the inputs must *sink* current (about 1.6 mA in the original TTL family). Because of the finite current-handling capability of TTL devices (approximately 16 mA), each standard TTL output can drive ten inputs. Beyond this level of loading, the outputs are no longer guaranteed to produce valid logic levels.

Improvements in circuit design and fabrication techniques have spawned a number of TTL subspecies. The use of Schottky clamps (see CIRCUITS, ANALOG) to improve speed significantly without requiring an increase in power is employed in all of the newest TTL devices. Schottky diodes prevent the buildup of excessive base charge when a bipolar transistor is on. Since less excess charge must subsequently be removed, the transistor can turn off much more quickly.

Typical propagation delay times for one type of simple TTL gate are now approximately 3 ns, with power requirements of roughly 4 mW/gate, while another TTL subspecies offers 1–2 ns delays and 10 mW/gate power "dissipation." In both cases, the product of delay and power (conventionally known, in defiance of dimensional considerations, as a speed–power product) is about the same (10–20 pJ/gate). Yet another version of TTL parts requires only 1 mW/gate while switching in about 4 ns, for a 4-pJ product. To the extent that the speed–power product is a figure of merit, smaller numbers are clearly desirable.

As one last comment on this logic family, it should be noted that Schottky (metal-semiconductor) diodes take the place of double-emitter transistors for reasons of speed in modern "TTL" devices.

6.2 Metal-Oxide-Semiconductor (MOS) Logic

Metal-oxide-semiconductor field-effect transistors (MOSFETs) do not require any expenditure of power to maintain their state (i.e., on or off). In principle, then, digital circuits using such devices could drive an infinite number of like gates and would require no power on a static basis.

As with bipolar transistors, MOSFETs are found in two broad classes, termed *n* channel and *p* channel for the polarity of the current-carrying charge. Figure 14 shows the commonly used shorthand symbols for MOSFETs. Where more than one symbol is shown for a particular type of MOSFET, all are equivalent for analyzing digital circuits at this level. Note that all integrated MOSFETs are in reality four-terminal devices. The fourth terminal, the substrate, is shown explicitly in only some of the symbols of Fig. 14, and is almost universally connected to the source terminal in digital circuits and may thus be ignored for the most part. Also, as far as digital circuits are concerned, MOSFETs behave essentially as switches. The on state is characterized by a low (ideally zero) resistance between the drain and source terminals, while the off state is a nearly infinite resistance between these terminals. The state of the device is primarily

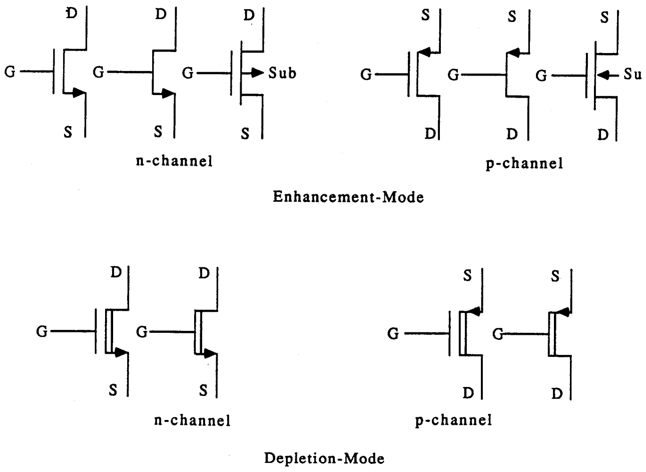

FIG. 14. Shorthand MOSFET symbols.

controlled by a voltage between the gate and source terminals.

In addition to differences in the polarity of the current-carrying charge, MOSFETs are either enhancement-mode or depletion-mode devices (again, see Fig. 14). In enhancement-mode devices, the transistor is constructed so that it is normally off, and thus requires the application of a nonzero gate-to-source voltage to turn it on. This voltage is positive for *n*-channel devices and negative for *p*-channel devices.

In depletion-mode devices, the transistor is normally on and a nonzero gate voltage must be applied to turn it off (negative for *n* channel and positive for *p* channel).

Figure 15 shows a NAND gate realized with *n*-channel devices. To a first approximation, the depletion-mode device may be thought of as a resistor (although a very nonlinear one). Thus the output will be high unless both inputs to the enhancement-mode devices are high.

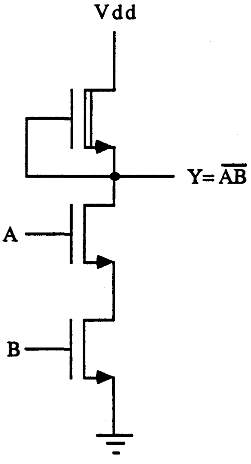

FIG. 15. NMOS NAND gate.

Although a completely analogous circuit can be constructed with *p*-channel devices, *n*-channel realizations are generally favored because electron mobility in silicon is significantly greater than that of holes, leading to superior operating speeds.

A drawback of this enhancement-/depletion-mode combination is that current always flows whenever the output is low; the circuit thus consumes energy in that state. Since the gates of MOSFETs themselves require no current on a static basis, the idling current in the low state represents a significant waste.

This problem can be solved by combining enhancement-mode *n*- and *p*-channel devices to create complementary MOS (CMOS) logic circuits, as shown in Fig. 16. Whenever the inputs are valid logical levels, *n*- and *p*-channel devices are never on simultaneously. In the NAND circuit example shown in Fig. 16, the output is taken to the low state only when both M_1 and M_2 are simultaneously on; that is, only when both inputs are high. When any input is low, either M_3 or M_4 is on, forcing the output to the high state. Since there is thus never a conducting path from the supply to ground (only from supply to output or from ground to output), no energy is consumed on a static basis, with the exception of that due to negligible leakage currents.

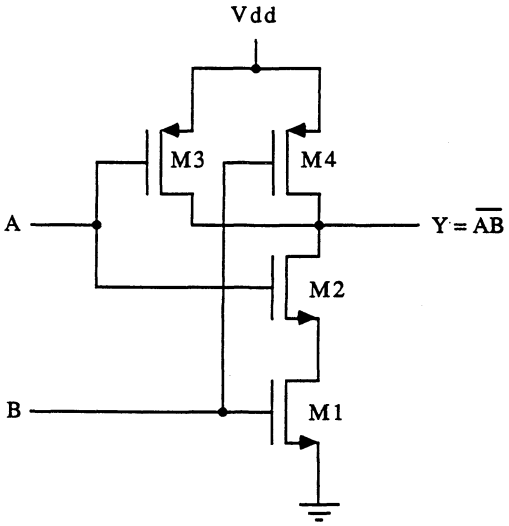

FIG. 16. CMOS NAND gate.

Although CMOS circuits are somewhat less area-efficient than purely *n*-channel ones because of the need to accommodate two transistors of opposite polarity, its vanishingly small static power requirements (typically the order of 10 nW) have led CMOS to dominate MOS logic technologies. Furthermore, the greater power requirements of *n*-channel technologies (or of any technology not using complementary devices) imposes a packing-density limitation because of thermal problems. Hence, CMOS is supplanting all other MOS technologies as densities increase.

To maintain compatibility with established TTL logic circuits, most CMOS devices possess characteristics that are sufficiently similar to TTL in most respects that they may be considered roughly equivalent. Older CMOS devices, however, can function with supply voltages ranging from 3 to 15 V, while newer CMOS devices impose a maximum of 5–7.5 V, compared with the narrow 5% supply voltage tolerance of TTL.

One other attribute of CMOS devices deserves discussion. Since MOS transistors are essentially electronically controlled switches, the outputs of CMOS gates swing within millivolts of the supply and ground, while TTL outputs are typically a maximum of about 3.5 V (out of 5) and a minimum of roughly 0.35 V. Thus CMOS devices can tolerate somewhat more noise than can TTL devices. On the other hand, CMOS devices can generate a great deal more noise than TTL devices because there can be a conducting path from supply to ground through the transistors on a *transient* basis, when perhaps both *n*- and *p*-channel transistors are momentarily on simultaneously. If the impedance of the power supply distribution wires is not sufficiently low, some unintended modulation of the power supply local to the devices can occur with attendant negative consequences. As with all high-speed circuits, this problem is overcome by the liberal use of *bypass capacitors* connected from power supply to ground as close to the device as practical.

6.3 Emitter-Coupled Logic (ECL)

Emitter-coupled logic was developed around the same time as TTL. Recall that early TTL devices were unnecessarily slow because the transistors were allowed to go into saturation, causing excessive buildup of base charge, with attendant reductions in speed.

The philosophy underlying ECL is to avoid saturation and to separate the 0 and 1 states by a smaller voltage than in TTL so that faster switching results for a given maximum rate of change of voltage.

Figure 17 shows an ECL OR/NOR gate in simplified form. Note that the outputs are "open-emitter," meaning that external resistors must be supplied to provide a current path for the output stage. This arrangement allows the outputs of several gates to be wired together, effectively providing an OR function at the common node.

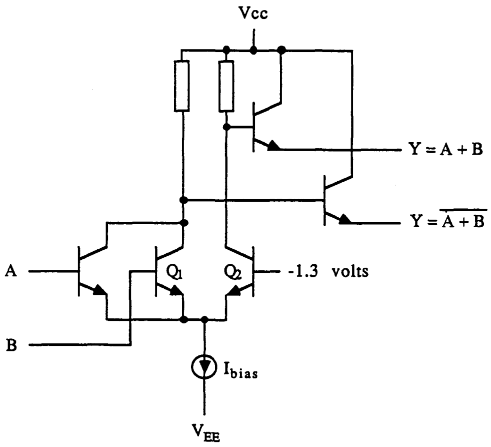

FIG. 17. ECL OR/NOR gate (simplified).

In contrast with TTL and CMOS circuits, the most positive supply in ECL circuits is usually ground, while a negative supply of approximately -5.2 V is normally used, although most ECL circuits will accommodate negative supplies from -4 to -6 V.

The fundamental property on which ECL devices rely is the fact that an emitter-coupled pair of transistors such as Q_1 and Q_2 cannot saturate in the connection shown. Furthermore, base voltage differences of only several kT/q (where k is Boltzmann's constant in joules per kelvin, T is the absolute temperature in kelvins, and q is the magnitude of the electronic charge in coulombs) are required for one of the pair to conduct nearly all of the current I_{bias}. Since kT/q at room temperature is about 25 mV, far smaller voltage differences are required to separate 1 from 0 in ECL circuits than in TTL. With ground and -5.2 V at the supply connections, a valid input 0 is anything lower than approximately -1.475 V, while a valid input 1 is anything greater than about -1.105 V.

By avoiding transistor saturation and by employing small voltage swings, ECL circuits achieve high speeds, with typical propagation delays of 0.5–2 ns for simple gates.

By virtue of the current source I_{bias} (which can be a simple resistor), the current drawn from the supplies is relatively constant, even during switching, since the current merely steers from one transistor of the emitter-coupled pair to the other. Thus not as much power-supply noise generation occurs as with CMOS. However, since the logic swings are smaller, there is less room to accommodate noise in the first place.

Another attractive attribute of ECL circuitry derives from the symmetry of the input stage. Since the collector voltages of the emitter-coupled pair are 180° out of phase, two outputs that are logical inverses of one another are available if both collector outputs are used, as in the example of Fig. 17. Because of the topological symmetry, these two outputs will have essentially no skew in time with respect to one another. This consideration can be particularly important as one pursues extremely high-speed operation.

Unfortunately, the relatively high power requirements (e.g., 10–20 mW/ gate) have limited the density of ECL circuitry because of thermal considerations. Thus only a relatively narrow range of functions is available in ECL form. Nevertheless, ECL offers the highest operational speeds of any silicon-based logic family.

6.4. Gallium–Arsenide (GaAs) Logic

Primarily because electron mobilities in GaAs materials greatly exceed those in silicon, *n*-channel GaAs field-effect transistors (FETs) exhibit substantially faster switching speeds. While this much has been known for quite some time, various impediments primarily of a materials-processing nature have forced a slow development of integrated circuits employing these transistors. For example, the lack of a high-quality insulating oxide has prevented the realization of GaAs MOSFETs. Instead, MESFETs (metal-semiconductor FETs) must be used. Additionally, the tendency of arsenic to evaporate from the crystal during processing has hindered the use of annealing to repair the crystal damage caused by ion-implantation of dopants.

Despite these difficulties, GaAs logic circuits are available and offer speed–power products as small as 15 fJ/gate at 0.1 mW/ gate. At 1 mW/gate, delays of 75 ps are not atypical. Such miserly power consumption means that thermal considerations will not set the bound on attainable density as it does in silicon logic families, although inhomogeneities in processing do limit wafer sizes at present.

Because of the greatly inferior mobility of holes in GaAs, complementary circuits are impractical. Consequently, all logic operations must be carried out with *n*-channel transistors of the type mentioned earlier, the depletion-mode MESFET. An example of a two-input NOR gate is shown in Fig. 18, and is somewhat reminiscent of *n*-channel depletion-mode silicon circuits.

In this circuit, M_1 may be treated as a resistor for purposes of this discussion. As we saw in the silicon case, use of depletion-mode transistors as resistors exacts a penalty in power consumption. It is therefore remarkable that GaAs circuits use as little power as they do. In any event, transistor M_2 is normally off in the absence of any inputs, or when both inputs are low, because of the action of M_1. Transistor M_3, which again may be treated as a resistor, thus pulls the output high.

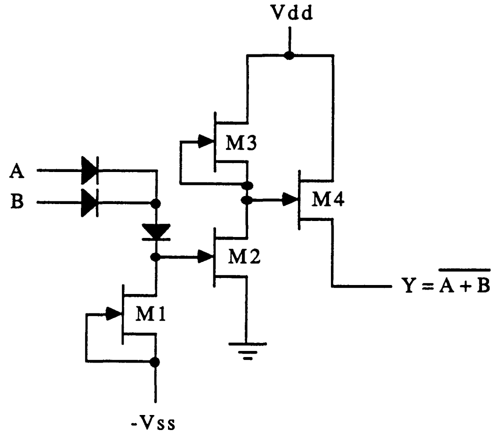

FIG. 18. GaAs NOR gate.

If any input is high, M_2 turns on, pulling the output low. Hence, we see that this circuit does perform the NOR function. Furthermore, any number of arguments can be accommodated simply by adding more diodes to the input stage.

7. SOME REPRESENTATIVE APPLICATIONS

Figure 19 is a collection of some applications of digital circuits. Figure 19(a) shows how a flip-flop may be used as a frequency divider. At each rising edge of the input clock (possessing a frequency "fclk"), the flip-flop output changes state by virtue of the feedback. Thus the output signal is one-half the frequency of the input clock. Many such stages can be chained to yield division factors of any integer power of 2, and these divider chains are integral parts of digital watches, for example. In this application, a master clock frequency of 32 768 Hz ($=2^{15}$ Hz) is often used. A chain of 15 divide-by-two stages thus produces one pulse per second. These pulses in turn drive stages which handle the counting in modulo-12 (or 24) and modulo-60. Finally, subsequent circuitry translates the counter outputs into signals that drive a display.

Figure 19(b) shows how a counter may be used to provide integer divide factors that are not necessarily powers of 2. Most counters have a provision for allowing the preloading of a count value; that is, counting can begin from some arbitrarily set value. Furthermore, many counters provide a separate indication (here, available as signal "Max/Min") of when the output is either a maximum or minimum value. In the circuit of Fig. 19(b), the value of the divide factor is supplied to the preload inputs. At each tick of the input clock the counter decrements toward zero, at which point the Max/Min line goes high. Assuming synchronous preloading, the next tick of the clock results in a return to the preset count value (because the "load" line has been asserted) and the cycle repeats. The Max/Min output is used as the divided output, as shown. Note that the output does not spend equal amounts of time in the high and low states, in contrast with the pure flip-flop divide-by-2^n case.

Figure 19(c) is a sum-of-products implementation of the "carry out" function ($=AB + BC + AC$) of the adder example studied in

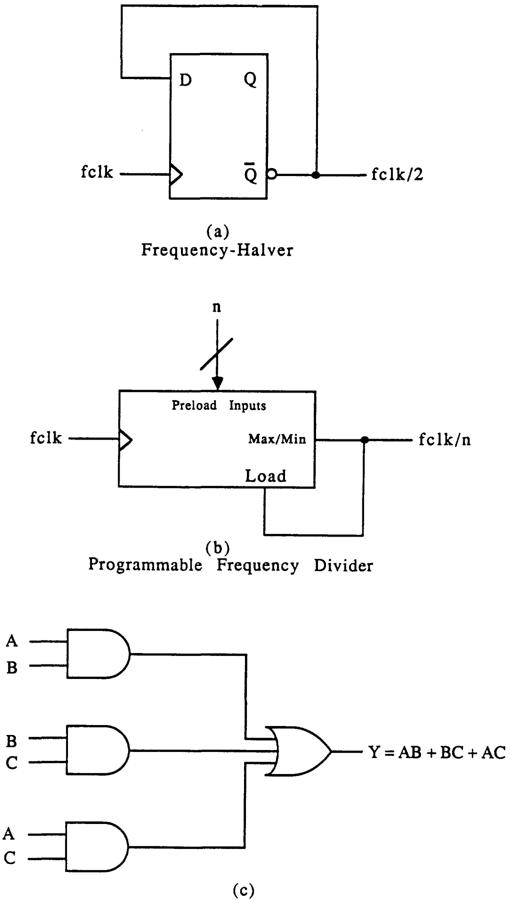

FIG. 19. Miscellaneous applications.

Sec. 3. Recall that, in principle, any Boolean function can be expressed in a sum-of-products form so that only a single level of AND gates and a single level of OR gates is needed. Since signals only propagate through two levels of circuitry, such implementations are capable of exhibiting high-speed performance.

As one last example, Fig. 20 is a simplified block diagram of a microprocessor. Even though all microprocessors are not identical, this diagram is representative of "typical" architectures.

The arithmetic logic unit (ALU) is the heart of the microprocessor, and can perform a wide variety of arithmetic and logical operations as described in Sec. 4. The instruction decoder converts instructions stored in external memory into signals that will initiate appropriate operations in the ALU as well as elsewhere in the system.

The locations of program instructions are identified with a collection of bits known collectively as an address, and the program counter holds the address of the next instruction to be executed. While program execution normally proceeds in a linear sequence, this flow may be interrupted or otherwise modified conditionally on the outcome of operations. The program counter's contents are then altered as necessary.

The accumulator stores the arguments and results of operations, while the condition code register provides information about the status of operations, such as overflow resulting from the addition of overly large numbers.

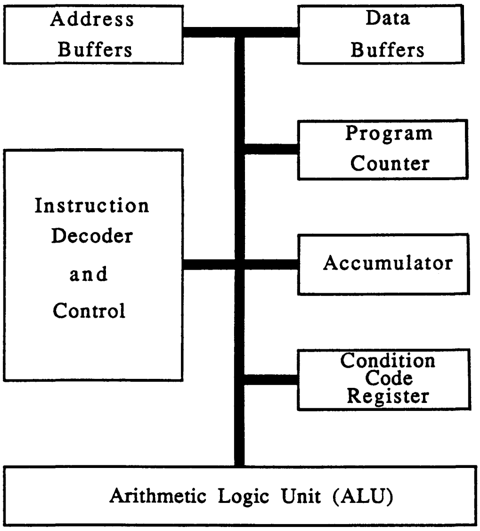

FIG. 20. Microprocessor block diagram.

Finally, address and data buffers provide connections to the external world. Any group of signals possessing some related function collectively forms what is known as a bus, and the data and address buses are the lines of communication with devices such as ROM (which store the basic operating software), RAM (which stores data, including software loaded from an external source such as a disk), and a variety of input/output (I/O) devices. Buffers merely boost signal power from relatively feeble intrachip levels to the values necessary to drive external loads.

8. FUNDAMENTAL PERFORMANCE LIMITATIONS

We have seen that a variety of technological improvements in recent years have yielded large increases in computing speed per unit energy. It is natural to consider what the fundamental limits on such improvements may be.

Central to this end is an acknowledgment that computation is a *physical* process since all systems that manipulate data ultimately must represent information as physical, sensible quantities such as charge or magnetic flux. With this idea in mind, we can establish some rough bounds on computation speed and power consumption in a systematic way. Furthermore, these bounds allow us to evaluate and compare various technologies, both real and proposed.

Let us first examine how much energy is required to perform computation, where we shall define "computation" as a cyclical change in state (i.e., from 0 to 1 to 0). One factor that sets a lower bound on this energy derives from a consideration of thermally induced errors; the energy separating the 0 and 1 states must be chosen sufficiently large to guarantee unambiguous identification of those states, even in the presence of thermal fluctuations.

The probability of a thermal fluctuation of an amount of energy E_0 occurring in a system within a response time τ is given by the Boltzmann relation:

$$P=\exp(E_0/kT), \tag{18}$$

so that the mean time between errors (MTBE) is

$$\mathrm{MTBE} = \tau \exp(E_0/kT), \tag{19}$$

or, solving for the energy fluctuation, we obtain

$$E_0 = kT \ln(\mathrm{MTBE}/\tau). \tag{20}$$

Since we need to separate our 0 and 1 states by an amount that exceeds E_0, we conclude that the "switching energy" (the energy required to switch from high to low and back to high) must be chosen as a large multiple of kT (say, $100kT$) in order to provide a reasonable MTBE from thermal fluctuations. At a temperature of 300 K, $100kT$ is approximately 0.4 aJ, which is four or five orders of magnitude smaller than the energy consumed by a typical GaAs gate.

One reason for the great difference between the minimum switching energy computed on the basis of thermal considerations and the energy typically consumed by physical circuits is the need to choose power supply voltages that are large enough to accommodate manufacturing variations of various types. In the case of MOSFETs, for example, a parameter known as the threshold voltage is dependent on a host of factors such as temperature and doping levels. Because threshold voltage is related to the value of gate-to-source voltage that determines a device's transition from off to on, any variation in this parameter is effectively a noise term that corrupts the interpretation of logic signals. Since the variation of threshold voltage can be controlled in practice only to a certain extent, the logic levels must be chosen with a sufficient separation to guarantee that threshold-variation-induced errors are satisfactorily improbable. Use of higher voltages in turn results in higher power consumption.

A more significant reason for seemingly excessive power consumption is that operational speeds generally increase with power. In FET logic, for example, there are various parasitic (undesirable but irreducible) capacitances that must be charged and discharged to effect changes in state. Hence faster switching can be obtained, but at the expense of increased power consumption. In general, then, we see that various factors of a practical nature can force the use of switching energies that greatly exceed the thermally limited value.

Because the foregoing statement initially appears to apply strictly only to the MOSFET case studied, it is interesting to evaluate superconducting logic circuits with these ideas in mind. Consider a technology in which information is represented as flux stored in a superconducting ring and controlled with some superconducting switching element (e.g., a Josephson junction). Assume that one flux quantum Φ_0 ($= 2\times10^{-15}$ Wb) is stored in the ring. In this case, the switching energy is simply

$$E_{\mathrm{sw}} = LI^2 = \Phi_0^2/L, \tag{21}$$

where L is the inductance of the ring and I is the current circulating in the ring.

Assuming that the inductance L can be any chosen value, we see that the switching energy can be made as small as desired so that the thermally limited value can be approached. However, since the inductance L scales with the size of the loop, large inductances are obtained at the expense of reduced circuit density.

If higher densities are pursued, higher switching energies must be employed. In general, it may be shown that *any* logic device working at its quantum limit requires a higher switching energy as dimensions are made smaller. Furthermore, the quantum-limited response time τ_Q grows at the same time switching energy decreases according to the following relation:

$$\tau_Q \gtrsim h/\pi E_{\mathrm{sw}}, \tag{22}$$

where h is Planck's constant, 6.6×10^{-34} J s/rad. If one substitutes the thermally limited value of $100kT$ for E_{sw}, we obtain the following relationship:

$$T\tau_Q \gtrsim h/100\pi k \approx 1.5\times10^{-13}\ \mathrm{K\,s}. \tag{23}$$

That is, the product of absolute temperature and response time in any system possessing an acceptably low rate of thermally induced errors must always exceed about 10^{-13} K s.

Using this relationship, we find that the quantum-limited response time at liquid-nitrogen temperatures exceeds about 2 fs, while this value is tens of femtoseconds at liquid-helium temperatures. Even though one may employ smaller loops to reduce the inductance and consequently increase speed, one may not reduce delay times below these quantum limits.

Mead and Conway (1980) have compared switching energies of room-temperature MOSFET and superconducting logic as a function of feature size and have found that MOSFET logic actually does better than superconducting logic systems (on a switching energy basis) for feature sizes smaller than about 0.5–1.0 μm. Thus as higher densities are pursued, superconducting logic looks progressively less attractive as far as switching energy is concerned. Alternatively, as lower switching energies are pursued in superconducting logic, both allowable densities and attainable speeds suffer. At 1-μm feature sizes, the switching time of a minimum switching-energy superconducting device at 4 K is approximately 20 ps. Heterojunction high-electron mobility FETs (HEMTs) operated at a temperature of 77 K have achieved speeds of the same order.

When the maximum switching speeds of individual elements are considered, however, superconducting logic is superior, with switching times as fast as a few hundred femtoseconds at liquid-helium temperatures (with the choice of a suitably large switching energy), compared with the tens of picoseconds typically attainable in FET technologies. Although one can trade speed for power in virtually any technology, this tradeoff extends to higher speeds for superconducting logic than for FET logic because the former allows operation nearer the quantum limits of switching energy and speed if density considerations are ignored.

Even though the switching speeds of individual superconducting elements are superior, much of this advantage can erode when larger systems of these elements are constructed because of unavoidable speed-of-light delays (and dispersion) along the interconnecting wires. Although this delay is approximately 3 fs/μm in free space, it may be as large as 5–10 fs/μm in the materials commonly used in integrated circuits. This delay can easily dominate in superconducting systems, not only because a given delay represents a higher fraction of the intrinsic response time, but also because the lower allowed densities force the use of longer interconnecting paths and thus greater delays than for FET technologies.

In general, use of superconducting logic at present may be justifiable only in special circumstances, where its superior raw speed can accomplish what other technologies cannot (even with architectural enhancements such as a high degree of parallelism), and where one is thus willing to tolerate lower circuit density, higher power consumption (including that associated with the cryogenic apparatus), and increased difficulty of fabrication.

GLOSSARY

Bus: A collection of related signals (e.g., as in an address or data bus).

Clock: Usually a periodic signal used to control the timing of events in digital circuits.

Combinational Logic: Logic circuits in which the outputs are functions of only the present values of the inputs.

Flip-Flop: A sequential logic element with two stable states.

Karnaugh Map: A particular arrangement of a truth table that aids in the synthesis of logic circuits.

Random-Access Memory (RAM): Memory in which read and write operations may both be performed, and in which all memory cells are equally accessible. Most RAMs lose their contents when power is removed.

Read-Only Memory (ROM): A memory circuit that is generally programmed (written to) only once or rarely. The stored data are not lost when power is removed.

Sequential Logic: Logic circuits in which the outputs may be functions of present and past values of the inputs, as well as past values of outputs.

State Machine: A synonym for sequential logic.

Synchronous Systems: Logic circuits in which states may change only at intervals controlled by a clock.

Truth Table: The input–output behavior of a combinational circuit expressed in simple tabular form.

List of Works Cited

Mead, C., Conway, L. (1980), *Introduction to VLSI Systems,* Reading, MA: Addison–Wesley.

Further Reading

Brodsky, M. H. (1990), *Sci. Am.,* February, 68–75.

Horowitz, P., Hill, W. (1980), *The Art of Electronics,* Cambridge, UK: Cambridge University, Chaps. 8 and 9.

COAL

WENDELL H. WISER, *Department of Fuels Engineering, University of Utah, Salt Lake City, Utah, U.S.A.*

Introduction 528
1. **Coal Resources** 528
1.1 Coal Resources versus Coal Reserves 528
1.1.1 Parameter Values for Estimating Resources 529
1.1.2 Parameter Values for Determining Reserves 529
1.1.2.1 Degree of Exploration 529
1.1.2.2 Depth of Coal Seam 529
1.1.2.3 Seam Thickness 529
1.1.2.4 Recoverability 529
1.2 World Coal Reserves and Resources 529
2. **Origin and Classification of Coal** 530
2.1 Origin of Coal 530
2.2 Coal Type (Petrography) 531
2.3 Coal Rank 531
3. **Mining of Coal** 532
3.1 Deep Mining 532
3.1.1 Room-and-Pillar Mining 532
3.1.2 Longwall Mining 532
3.2 Surface Mining 533
3.3 Thickness of Overburden in Underground Mining 533
4. **Transportation, Storage, and Handling of Coal** 533
4.1 Transportation 533
4.1.1 Transportation Methods 533
4.1.1.1 Rail—Unit Trains 533
4.1.1.2 Trucking 534
4.1.1.3 Slurry Pipeline 534
4.1.1.4 Water Transport 534
4.2 Storage 535
4.2.1 Spontaneous Ignition and Weathering 535
4.3 Handling 535
5. **Analyses and Composition of Coal** 536
5.1 Sampling 536
5.2 Proximate Analysis 536
5.3 Ultimate Analysis 537
5.4 Physical Structure of Coals 537
5.5 Chemical Structure of Coals 538
5.6 Mineral Matter in Coals 538
6. **Properties of Coals** 538
6.1 Physical Properties 538
6.1.1 Reflectance 538
6.1.2 Porosity and Surface Area 539
6.1.3 Electrical Conductivity 539
6.1.4 Magnetic Susceptibility 540
6.1.5 Specific Heat and Enthalpy 540
6.2 Mechanical Properties 540
6.2.1 Elasticity 540
6.2.2 Strength 540
6.2.3 Hardness and Abrasiveness 540
6.3 Plastic Properties of Coal 541
7. **Coal Preparation** 541
7.1 Size Reduction 542
7.2 Beneficiation 542
7.3 Dewatering and Drying 542
7.4 Automation 542
8. **Coal Carbonization** 543
8.1 Principles 543
8.2 Low-Temperature Carbonization 543
8.2.1 Products and their Uses 543
8.3 High-Temperature Carbonization 544
8.3.1 Metallurgical Coke Manufacture 544
8.3.2 By-Products and their Uses 545
9. **Coal Combustion** 545
9.1 General Principles 545
9.2 Coal-Combustion Systems 546
9.2.1 Combustion on a Grate 546
9.2.2 Atmospheric Fluidized-Bed Combustion 546
9.2.3 Pressurized Fluidized-Bed Combustion 547
9.2.4 Entrained-Flow Combustion 547
10. **Synthetic Fuels and Other Products from Coal** 547
10.1 Gaseous Fuels from Coal 548
10.1.1 Producer Gas 548
10.1.2 Water Gas 548
10.1.3 Low-Heating-Value Gas 548
10.1.4 Intermediate-Heating-Value Gas 549

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

10.1.5 High-Calorie Gas or Substitute Natural Gas 549
10.1.6 Underground Gasification..... 549
10.1.7 Integrated Gasification Combined Cycle.............. 550
10.2 Liquid Fuels from Coal....... 550
10.2.1 Coal Pyrolysis................ 550
10.2.2 Solvent Extraction............ 551
10.2.3 Catalytic Liquefaction 551
10.2.4 Synthesis Gas from Coal...... 552
10.2.5 Indirect Liquefaction 552
10.2.5.1 Liquid Hydrocarbon Fuels ... 552
10.2.5.2 Alcohol (Methanol) 552
10.3 Briquetting of Coal........... 553
10.4 Methane Production from Coal Seams 553
11. Environmental, Health, and Safety Impacts from Coal Utilization 554
11.1 Impacts from Coal Mining.... 554
11.1.1 Explosions.................. 554
11.1.2 Mine Cave-In and Subsidence.................. 554
11.1.3 Surface-Mining Environmental Impact........ 554
11.1.4 Impact upon Health 555
11.2 Impacts from Carbonization .. 555
11.3 Impacts from Combustion 555
11.3.1 Particulates 555
11.3.2 Oxides of Sulfur............. 555
11.3.3 Oxides of Nitrogen 556
11.3.4 Carbon Dioxide 556
11.3.5 Clean Coal Technologies...... 557
11.4 Impacts from Coal Conversion.................. 557
12. Economic Considerations 557
Glossary..................... 558
Works Cited 558
Further Reading............. 558

INTRODUCTION

Coal is, by a very large margin, the largest known fossil-fuel resource occurring in the world, and, as technologies exist at this writing, the largest reserve of controllable energy in the world. In the early part of the twentieth century, coal became the dominant energy resource used in the industrialized world. However, as petroleum and natural gas resources were developed, and systems for their distribution installed, coal yielded its position of dominance to these two resources, primarily on the basis of cleanliness and convenience (fluids as opposed to solid). As reserves of petroleum and natural gas decline, most observers in the field of energy supply predict a return of coal to its position of dominance, particularly in applications where electric energy may not be suitable. This article will introduce the reader to this important energy resource, identifying many of the aspects of coal that must be addressed if it is to move to the position of prominence being predicted for it.

1. COAL RESOURCES

Quantities of coal in the United States are stated to be as small as 225 billion metric tons, and as large as 3180 billion metric tons. It is important to understand the bases for these various estimates. The parameters used in these estimates have not been completely standardized among nations, and sometimes not even among states or provinces within a nation. The system of parameters utilized in this treatise is one quite commonly employed in the literature.

1.1 Coal Resources versus Coal Reserves

In this system four parameters are utilized in the determination or estimation procedure, namely degree of exploration, depth of the coal seam (thickness of overburden), seam thickness, and fraction of coal recoverable. The degree of exploration, in turn, is utilized under four definitions:

- Measured—core holes are drilled not more than 0.8 km apart, thereby establishing the thickness of the seam, the lateral extent of the seam, and the quality of the coal.
- Indicated—core holes are drilled not more than 2.4 km apart.
- Inferred—an occasional reference or control point established, as by core holes more than 2.4 km apart, or outcropping, then extrapolation beyond the control point to distances greater than 2.4 km, based upon geological factors.

- Potential—an outcropping or other visual observation, followed by extrapolation based only upon geologic and geographic patterns.

1.1.1 Parameter Values for Estimating Resources. The term *resources* is a general term comprising many combinations of applications of the factors presented above. *Gross resources,* a very large number often quoted, is the sum of coal quantities estimated by all of the above degrees of exploration, with no restrictions placed upon seam depth or seam thickness, and assuming 100% recoverability. This is the basis for the estimate of 3180 billion metric tons of coal in the United States. Other smaller numbers are often quoted by placing some restriction upon certain of the above factors, most often excluding the "Potential" category under exploration and placing some limitation upon thickness of overburden. These estimates, in general, do not consider whether the coal thus estimated can actually be mined, and are without economic connotation.

1.1.2 Parameter Values for Determining Reserves. The term *reserves* generally means *demonstrated reserve base,* meaning known and recoverable by current mining technology. Although not all countries have adopted the same restrictions for what is deemed to be recoverable, the limitations applied quite generally in the United States, and certain other countries are given here.

1.1.2.1 Degree of Exploration. Estimates derived from both measured and indicated resource categories are considered to be quite reliable for this purpose, and are usually combined in determining coal reserves. Estimates in the inferred and potential resource categories are usually considered to be too speculative and are not generally included in determining reserves. It must be recognized that there are exceptions to these uses in some countries.

1.1.2.2 Depth of Coal Seam. Deep mining of coal at depths greater than about 300 m is not practiced in many countries; hence, for those countries only coal with less than 300 m of overburden is counted in reserve determinations. This is not uniform, however. In countries or regions where coal is deep-mined to depths up to 900 m, the 900-m restriction may be used in determinations of reserves. (See Sec. 3.3.5.)

1.1.2.3 Seam Thickness. The restriction applied relative to this factor strongly depends upon the degree of depletion of the coal reserves of the area. For example, in the United States where reserves are still very large, deep mining of seams less than 70 cm in thickness is considered to be uneconomic; hence, such seams are not included in reserves. By contrast, in certain European countries where good quality coal is largely deleted, seam thicknesses down to 35 cm are included in reserves.

1.1.2.4 Recoverability. The degree of recoverability is dependent upon the mining technique to be used as well as other factors, and may range from less than 50% of the coal in place to near 100%.

1.2 World Coal Reserves and Resources

In Table 1 are presented world coal reserves, as discussed above, and two categories of world coal resources frequently reported

Table 1. World coal reserves and resources (billions of short tons).

Continent or Region	Demonstrated reserve base (1)	Identified resources (2)	Gross resources (3)
Asia, including USSR	752	3000	11 000
North America	476	1900	4 400
Europe	75	300	800
Africa	23	90	250
Australia, New Zealand, New Caledonia	18	70	130
South and Central America	8	30	40
Total	1352	5390	16 620

in the literature. Column (1) presents quantities determined by the following criteria: measured and indicated resource categories, 0–300 m depth, and seam thicknesses as utilized in various countries or regions (usually either 70 or 35 cm). Column (2) includes inferred resource quantities along with measured and indicated resources, and depths 0–900 m. Column (3) represents gross resources as defined in Sec. 1.1.1 above. Because of the different criteria used by the various estimators, the values in column (1) involve reasonable adjustments to some data to arrive at a consistent determination. The values presented are conservative. In further years, as coal reserves are more widely used, coal resources not now included in the demonstrated reserve base will be further explored, moving many of those resources into the demonstrated reserve base and greatly increasing the known and recoverable coal reserves.

2. ORIGIN AND CLASSIFICATION OF COAL

Coal is considered by geologists to be a rock. Hence the scientific study of coal is called *coal petrology*. The description or classification of coal is called *coal petrography*. Coal is a heterogeneous material, so much so that the use of the word *polymeric* in relation to coal is quite inaccurate. Rather, coal is composed principally of microscopically distinguishable entities of organic origin, distinctive both physically and chemically, called *macerals*. Varying amounts of inorganic material are also found distributed through the coal seams.

2.1 Origin of Coal

Organic coal macerals are derived from plants, ranging from large trees to tiny plants, even spores and exudates. Over long periods of time the plant materials accumulated and were preserved in the moist environment of a swamp. As the trees and other plants fell, they were covered by water, thus isolating the decaying plant material from the atmosphere. Organic material preserved in this fashion is called *peat*. It was necessary that the level of the water in the swamp adequately cover the peat, yet not preclude plant growth. Considering the rather delicate balance required, one readily can understand that a coal seam of thickness of 30 m or more would be, and is, an unusual occurrence. In the course of time sediments washed in upon the peat, burying it and sealing it from air oxidation and erosion processes.

During the time of accumulation of the plant material in the peat swamp, inorganic substances entered the swamp, mostly carried in by water, either as particulate mineral grains or as ionic compounds dissolved in the water. These materials were deposited throughout the period of organic accumulation, resulting in mineral grains or combined with organic fractions. In general, the mineral deposits occurred from rather quiescent water movements, resulting in small mineral grains in the coal, although rapid water flows occasionally deposited rocks or large concentrations of inorganic material within the peat accumulation.

Following burial of the organic and inorganic materials under thick layers of sediment, geological activity in the earth subjected the peat to varying degrees of pressure and temperature. The alteration of these accumulated plant substances resulting from the effects of elevated pressure and temperature is termed *metamorphism*. The severity of the metamorphism determined many of the physical and chemical properties of the coal.

The coalification process represents a phenomenon whereby hydrogen and oxygen were removed from the carbon skeletal units of the woody material and any other organic materials present, being released chiefly as water, carbon dioxide, and methane. This release from a solid skeletal structure created tiny micropores throughout the coal, even down to the molecular scale. Hence coal is a highly porous material with large internal surface areas. The transformation of plant substances into coal is generally believed to involve two stages: a biochemical stage in which the woody material is transformed into peat, and a geochemical (metamorphic) stage in which peat is transformed into coal. The nature of the original plant material (e.g., tree trunk, leaves, etc.) passes its imprint through both stages. This imprint can be observed by viewing the coal maceral under a microscope; this viewing is called petrography and defines the coal type. The product from the geochemical stage additionally depends upon the severity of the metamorphism, which establishes coal rank.

2.2 Coal Type (Petrography)

One is interested in understanding, or even predicting, the response of a particular coal in a utilization process. The nature of the plant material as it existed before coalification began may greatly influence how the coal maceral formed from that plant material behaves in the utilization process. The maceral formed from one portion of a plant may have a different visual appearance or other different optical properties than one formed from another portion of the plant material. If a carefully prepared sample from the coal, such as a thin section or a polished surface, is examined under a microscope, it is possible to determine the relative proportions of the various maceral components and other optical features that correlate with chemical behavior. When carefully performed, this procedure can provide reasonably accurate analyses of the coal in terms of organic origin of its component macerals. This procedure is known as a petrographic analysis.

Three principal classes of substances are generally recognized as the major petrographic constituents of coal, namely, vitrinite, liptinite (sometimes called exinite), and fusinite (also called inertinite).

Vitrinite originates from the larger portions of plant material such as tree trunks, large branches, etc. In most coals vitrinite constitutes from about 50% to as much as 90% of the volume of the coal (commonly 75%–80%).

Liptinite originates from leaves, spores, resins, and other chemically similar materials in the plant. It typically possesses a higher hydrogen content than vitrinite and is usually more reactive than vitrinite. Liptinite is typically about 3%–5% of the coal.

Fusinite is relatively inert in many utilization processes, hence the alternate name "inertinite" preferred by some petrologists. Its origin relates more to the history of the woody material than to the chemical nature of the original plant material. The carbon content is considerably higher than observed in vitrinite or liptinite. Fusinite particles appear to have resulted from a forest fire or other extreme phenomenon such as could alter the chemical structure in the direction of carbon enrichment. Fusinite is normally present at lower than 10% concentration in the coal. In liquefaction of coal, it does not participate as a reactant. In the manufacture of metallurgical coke it does not participate in the transition to coke, but remains as an "inclusion" in the coke walls or in larger boundary definitions.

2.3 Coal Rank

As the peat accumulation was covered by sediments of increasing thickness, geothermal processes inside the earth provided heat, which caused the chemical release of oxygen and hydrogen. The heat and pressure generated by geothermal activity were often supplemented by other geologic activity, sometimes generating extremely high pressures, with accompanying increases in temperature. By this means the organic material was converted progressively toward more carbon-rich material. The physical and chemical properties of the resulting organic mass (coal), as well as the chemical composition, were largely determined by the severity of this metamorphic process. Consequently, coals are ranked in relation to the severity of the metamorphism by which they were formed.

Peat is a precursor of coal, not yet having been subjected to the geochemical processes, and hence is not generally included among coal ranks. Four principal categories, with a few subcategories, are used to classify coals according to rank. The principal categories are anthracite, bituminous, sub-bituminous, and lignite. (In some countries the term *brown coal* is used, which encompasses lignite and part of the sub-bituminous rank.)

Two or more properties are usually combined to define coal rank. The property that most consistently correlates with coal rank is elemental carbon content. Two additional properties frequently used (usually one but not both) are volatile-matter content (Sec. 5.1) and heating value.

Table 2 presents typical values of the most often used properties associated with each principal rank. For the higher-rank coals (anthracite, low-volatile bituminous, and medium-volatile bituminous) the percentage of volatile-matter content is usually used in combination with elemental carbon content to define rank. For the lower-rank coals (high-volatile bituminous, sub-bituminous and lignite), heating

Table 2. Typical parameter values for rank definition.

Coal rank	Fixed carbon (dry basis) (wt %)	Volatile matter (dry basis) (wt %)	Heating value (moist basis) (kJ/kg)
Anthracite	86–98	2–14	
Low-volatile bituminous	78–86	14–22	
Medium-volatile bituminous	69–78	22–31	
High-volatile bituminous[a]	52–69		25 500–32 500
Sub-bituminous[a]	50–58		19 000–30 000
Brown coal, including lignite	46–56		< 19 000

[a] An additional parameter distinguishes between high-volatile bituminous and sub-bituminous, i.e., high-volatile bituminous is agglomerating, while sub-bituminous is nonagglomerating.

value is used in combination with elemental carbon content. Anthracite is usually found in steeply pitched seams, indicating strong geologic activity accompanied by great pressures and high temperatures during coalification. By contrast, lignites, and generally sub-bituminous coals, are found in essentially horizontal seams.

3. MINING OF COAL

Coal, after its formation, must remain protected from the atmosphere to prevent slow degradation by a process known as *weathering*, a process of chemical attack by oxygen and moisture. Consequently any coal deposit of value will be covered by soil and/or rock ranging up to several thousand meters thick; this covering is called *overburden.*

3.1 Deep Mining

Coal rank bears a direct relationship to the severity of the metamorphic process. The higher-rank coals (anthracite and bituminous) were formed under conditions of high pressures and elevated temperatures, with overburden layer thick enough and of sufficient nature and strength to contain the organic mass while subjected to great geologic forces. Geological activity subsequent to formation of the coal may have altered the overburden thickness, and the erosive forces of water and wind have often reduced the thickness of the overburden, sometimes exposing the coal seam in an *outcropping.* Following all of these activities, anthracite and bituminous coals are usually found under a layer of rock and soil ranging in thickness from a few tens of meters to several thousand meters. The most practical means of removing these coals from seams situated under thick overburden is to follow the coal-seam contour without removing or disturbing the overburden.

3.1.1 Room-and-Pillar Mining. High-rank coals are usually located under rock overburden. However, when the coal is removed, which in the seam is supporting the overburden, the roof of the mine will collapse unless adequate support is provided. The oldest and still most widely used technique is to leave coal pillars in place to support the roof. Quite typically several parallel entry or development channels are mined into the coal seam and connected by cross channels. This mining pattern leaves behind supporting coal pillars about 18×24 m in cross section. The height of the cut into the coal seam depends upon the mining equipment used, and is typically about 3 m. Thus, of the 3 m thickness of coal potentially minable in this pattern, only about 40% is actually removed during the advance into the coal seam. If the seam is more than 3 m thick, additional coal will be left behind, generally in the roof, but sometimes in the floor, depending upon the composition and condition of the floor. In some instances additional coal is mined when retreating from the mine, by removing part of the pillars or some of the coal left in the mine roof or floor. On average only about 50% of the coal in place is removed in room and pillar mining. The entry channels provide for ventilation of the mine, to remove dust and flammable gases, particularly methane, which is present in all coal seams.

3.1.2 Longwall Mining. Following completion of the above-described retreating operation and withdrawal of the equipment, it is not considered to be practical, consistent with safety requirements, to reenter that section of the coal seam to attempt to recover a portion of the remaining coal. With the objective of increasing the percentage removal of the coal in place, by refraining from leaving coal behind in pillars, a technique known as *longwall mining* is used in some countries. Several parallel entry channels are mined into the coal seam. A pair of these channels are typically 180 m apart and extend as far as 1800 m, or even further, into the coal formation, following the coal-seam contour. A cross channel is then mined in the seam at this depth into the seam, connecting the two parallel channels. A specially designed mining machine then removes coal from the face along the cross channel by moving back and forth, advancing toward the mine mouth on each new cut. As the operation progresses toward the mine mouth, leaving a cavern behind, the roof caves in behind the mining machine. The special design and construction of the longwall mining machine provides protection to the operator, as well as to the complete operation. No coal pillars are thus left unmined, but the thickness of the coal mined is restricted to the height minable in the single pass, typically 3 to 4 m. On average, about 80% of the coal originally in place is removed in this technique.

3.2 Surface Mining

Low-rank coals, particularly brown coal (which includes lignite) and some sub-bituminous coals, are situated in environments that do not allow underground mining. The environments are often swampy, or the overburden is of such nature that it will not support itself above any entry channel of practical width. If such coals are to be mined, the operation involves removal of the overburden to expose the coal for recovery. The operation is variously known as surface mining, strip mining, or sometimes open-pit mining. The economics of surface mining dictate whether such a technique should be applied, governed by the thickness and nature of the overburden to be removed in relation to the coal-seam thickness and the quality and value of the coal. In general, at least 95% of the coal in place can be recovered in surface mining. However, even for very thick seams, overburden thickness greater than about 45 m usually discourages surface mining as a means of recovery.

There are several advantages in surface mining of coal. These include lower mining cost per ton of coal recovered, higher recovery, and greater safety. A major disadvantage is destruction of the surface. However, this disadvantage may be overcome by restoring the surface, which is considered to be practical if the surface is relatively flat.

3.3 Thickness of Overburden in Underground Mining

In general, in consideration of weight of the overburden, deep mining is not conducted at depths greater than 300 m. In a few areas of the world where the nature of the rock constituting the overburden may indicate sufficient strength, such as the Rocky Mountain region of Utah and Colorado, some mining is conducted to 900 m, but is not commonly done. It is for this reason that recoverable coal reserves generally do not include coal seams at depths greater than 300 m.

4. TRANSPORTATION, STORAGE, AND HANDLING OF COAL

4.1 Transportation

The demonstrated reserve base for coal is several hundred times larger than identified and recoverable petroleum reserves. Although coal exhibits serious disadvantages when considered as a substitute for petroleum and natural gas, the developed nations of the world will almost certainly be compelled to make such substitutions. Coal reserves usually are not located near large coal markets, mostly requiring transport of the coal over distances ranging from a few kilometers to 1600 km or more on land, or up to several thousand kilometers on the oceans.

4.1.1 Transportation Methods

4.1.1.1 Rail—Unit Trains Railroads are the principal means of long-distance transport for coal, the result of consideration of both cost and availability for most coal markets. The individual freight car, as part of a typical freight-hauling system, spends about half of its time in railroad yards, 40% in customers' yards or sidings, and only 10% of its time moving. This realization gave rise to the concept of the *unit train.* Since coal movements often involve very large quantities of coal, it is practical for a complete freight train—all of the cars in the train—to carry only coal, all loaded at one origin, all unloaded at one destination each trip, and moving in both directions on a predetermined schedule. This is the general concept of the unit train. Unit-train cars spend virtually no time in railroad yards, time available for the other functions of the cars is doubled at the outset, and productivity is greatly increased. Energy consumption per ton-km is about one-third of the consumption for mixed-train freight. The unit train in the United States typically contains 100 to 125 cars, carrying an average of 90 metric tons per car, and is moved by five or six diesel-powered locomotives of 3000 horsepower each.

An important aspect of coal transport by open-top rail cars is the loss of coal along the route due to wind, and called in the industry *windage*—about a 1% loss of the coal transported. This loss is often reduced by spraying the coal with water or oil, or sometimes by covering with a lid.

4.1.1.2 Trucking. Although a large amount of coal is moved by truck, the larger portion of this movement is with off-highway vehicles, defined as those which are too long, too wide, or too heavy for use on public highways. They are used in short-haul movements, such as from the mine mouth to an electric power plant built nearby. Trucks built for use on public highways may carry coal payloads up to 60 metric tons, while off-highway coal trucks may haul up to 180 metric tons of coal. The smaller trucks are often rear-dump, although the larger highway vehicles may be bottom-dump. The larger off-highway trucks are bottom-dump, hence of the tractor-trailer type (sometimes also called semi-trucks). Trucks for coal transport are essentially all diesel powered.

4.1.1.3 Slurry Pipeline. While it is known to most people that transport of liquids and gases in pipelines is an efficient and proven method of conveyance, it is not so generally known that transport of solids by pipeline enjoys the same advantages of economy, simplicity of operation and protection of the environment. For transport of coal by this means, coal is ground to a fine size, mixed with water to form a slurry, then pumped through a pipeline. The pipeline hydraulics require the coal to pass through 20 Tyler mesh (all particles smaller than 833 μm). Such transport is particularly suited to coals to be used in pulverized coal-combustion systems, which require grinding until about 80% will pass through 200 mesh (74 μm). Quite typically the coal–water slurry is about 50% solids by weight and moves through the pipeline at a velocity of about 1.5 m/s. If the coal is too coarse (greater then 20 mesh) a higher pumping velocity is required. If the coal is too fine, the slurry pumps more easily, but dewatering before combustion becomes more difficult. One seeks for the optimum balance between pumpability and dewatering characteristics. The slurry is usually transported through steel pipelines, and corrosion is controlled by adding corrosion inhibitors. Top-size particle control helps to limit erosion and prevents pipeline plugging. Booster pump stations are typically located from 100 to 160 km apart, depending upon the terrain. Continuous centrifuges are most commonly used for dewatering the slurry.

Although no serious technical problems exist relative to the coal-slurry pipeline, difficult political problems do exist. The coal-slurry pipeline is in direct competition with the railroad for coal transport. In the United States, transportation of coal represents a major source of revenue for the railroads. The railroads almost unanimously refuse to grant a right-of-way for coal-slurry pipelines to cross the railroad. Some form of eminent-domain legislation will be required to resolve this right-of-way problem. At the time of this writing, the railroads and other special-interest groups have successfully blocked such legislation in the United States. Without rights-of-way across railroads, coal transport by slurry pipeline will remain extremely limited.

4.1.1.4 Water Transport. Transportation by water is the least expensive method for shipping coal, therefore, the preferred route where a waterway exists. The vast majority of coal moved on rivers is carried on barges. "Jumbo" barges are as large as 10 m wide and 60 m long, and carry up to 1350 metric tons of coal. The barges are moved in "tows" of as many as 48 in number for waterways where locks are not required. For river transport where locks are involved, lock sizes are the major limitation in determining the number of barges in a tow. Other limiting factors may be the capacity of the towboat, current velocity, and the width and depth of the river channel. The towboat actually pushes the barges and usually its power ranges from 3000 to 10 000 horsepower. The tow speed usually ranges from 5 to 20 km/h. Energy consumption for transportation of coal on barges ranges from 260 to 450 kJ/ton km.

Because of the lower cost of coal transport by waterway, it may be cost-effective to transfer the coal from railroad to barge during a long-distance shipment. For example, western U.S. coal is shipped by unit trains from mine mouth to the western end of the Great Lakes, transferred to vessels, and hauled by waterway to the final destination along the shores of the lakes. Very large self-propelled vessels, hauling up to 40 000 metric tons of coal per trip, are used for large coal shipments on lakes. These vessels are equipped with a self-unloader able to discharge the cargo in about four hours.

Since the large increase in world prices of petroleum in the 1970s, transport of coal across oceans for such uses as electric power generation has increased. Ocean transport is coupled with either rail–land transport or pipeline. Where the land transport system is the pipeline, adequate water removal from the slurry must be achieved before loading into the ship, to ensure that the ship's stability will not be endangered. Furthermore, shipping of excess water is uneconomical. In terms of tonnage, coal is now the third largest commodity of international trade, exceeded only be petroleum and iron ore.

4.2 Storage

The use of coal as a major source of energy, such as at modern electric power generating plants, requires storage of large quantities of coal, both for day-to-day use (called live storage) and to ensure long-term availability (called dead storage). The live-storage area may contain as much as 100 000 metric tons of coal, with an additional 1 million metric tons in dead storage. Storage of large quantities of coal is usually achieved in open stockpiles on the ground. The site is prepared for stockpiling by removing the topsoil and installing a layer of crushed rock or other material as a stabilizing layer. Adequate drainage is also provided. While live storage piles may remain exposed to the atmosphere and weather, dead storage piles are usually covered, either by a roof or by spraying the surface of the pile with a chemical that dries to form a "skin" over the pile. As an alternative, coal is sometimes stored in bins and silos, constructed of concrete or steel. Such storage facilities are more environmentally acceptable, as they control dust as well as runoff from rain water.

4.2.1 Spontaneous Ignition and Weathering. A major danger in coal storage is spontaneous combustion. Coal is a porous material with a large internal surface area, typically greater than 50 m^2/g. Oxygen in the air reacts with the surface of the coal slowly at ambient temperatures. If the rate of liberation of heat by these oxidation reactions exceeds the rate of dissipation of heat from the pile, the temperature increases, resulting in an increase in oxidation rate, until a combustion flame appears. Moisture tends to promote these reactions. The heat of wetting of dry coal by rain increases the temperature of the pile, increasing the rate of oxidation. In order to prevent spontaneous ignition, run-of-mine coal should not be stockpiled. Because air enters the bottom of the pile and diffuses upward, the pile must either allow air circulation freely enough to dissipate the heat of oxidation, or be compact enough to prevent air circulation. In piling the coal, it should be laid down in layers about 0.5 m thick and each layer compacted by suitable equipment. Even when great care in stockpiling is exercised, the heating value of the coal will be reduced by 1% to 2% during the first year of storage. Coking quality is also reduced during storage.

4.3 Handling

From removal of coal from the seam to final utilization in an electric power plant or other use, coal handling is a major part of the total coal utilization system. Most of the equipment is highly specialized and has been developed specifically for the coal industry, e.g., underground mining machines; conveyors; loading, unloading and trans-shipment terminals; equipment for stacking and reclaiming in coal yards; and coal feeders. The trend has been to continuous systems, as well as to larger units. A total handling system approach is now normally used to permit complete integration of the various equipment items and automation to reduce the number of operating personnel to a minimum.

5. ANALYSES AND COMPOSITION OF COAL

Every chemical element occurring in nature can be found in some coal, somewhere, and all coals contain a very large number of the elements. Most of these elements are present only in trace amounts and may not influence the behavior of the coal in its final use. A few trace elements, however, may be harmful in a particular process, such as the ability of very small amounts of certain metals, occurring in the organic portion of the coal as organic metallic structures, to deactivate a catalyst during a conversion process.

Five elements usually account for more than 99% of the atoms in the organic portion of any coal, namely carbon, hydrogen, oxygen, sulfur, and nitrogen. These five elements are responsible for essentially all of the properties, both good and bad, of a coal. Hence any chemical analysis of a coal will usually specify the quantities of these five elements. For special uses of a coal, other elements may exhibit effects, usually adverse effects, but these will not be discussed.

Two types of analysis of coals are widely used in the selection of a coal for a particular use, namely proximate analysis and ultimate analysis. For a detailed discussion of these analyses, as well as other specialized tests, the reader is referred to the supplemental reading lists at the end of this chapter (Elliott, 1981).

5.1 Sampling

When one considers the manner in which the raw materials accumulated prior to the formation of coal (Sec. 2.1), one readily understands that a coal seam is not uniform in composition. If one proceeds a few feet through a coal seam, significant differences in composition are often observed. The analysis should be representative of the total amount of coal in the *lot*. The lot may consist of a few kilograms for laboratory use, a truck load of 20 tons, a unit-train load of 10 000 tons, or a ship load of 100 000 tons. In general, the value of the analysis data may not be greater than the care exercised in obtaining the samples for analysis.

A number of sampling techniques has been developed for various applications. Factors that influence selection of a procedure include the size of the lot to be sampled, the number of increments required to be taken for the lot size, and the top size of the coal particles in the increments, as well as others. It is important that some form of grid system be applied to the taking of increments from the lot to ensure their systematic spacing. If adequate care in sampling is exercised, a 50-g sample may be representative of a 10 000-ton unit train of coal. Coal exposed to the atmosphere will slowly oxidize, and hence change its chemical composition over time. In order to obtain an analysis of coal as it exists in the coal seam, the sample must be stored under an inert atmosphere.

5.2 Proximate Analysis

The proximate analysis reports the content of four quantities, namely moisture, volatile matter, fixed carbon, and ash. The analysis is conducted in three distinct steps, each step precisely defined in terms of temperature and time, as prescribed by the American Society for Testing and Materials (ASTM) under the designation ASTM-D3172. In order for results to be compared for various coals, in relation to the intended use, the various steps and parameters must be precisely followed.

- Moisture—A 1-g sample of coal of minus 250-μm particle size (all particles smaller than 200 μm diameter) is dried at a temperature of 110 °C for 1 hour. The weight loss of the sample is reported as moisture content, expressed as a percentage of the original sample weight.
- Volatile matter—A 1-g sample of coal of minus 250-μm particle size is heated to 950 °C in an inert atmosphere, and held at this temperature for 7 min. The weight loss of the sample during this operation, less the moisture, is reported as volatile-matter content, expressed as a percentage of the original coal sample.
- Fixed carbon—The remaining portion of the sample, with moisture and volatile matter removed as described above, is burned in air until a constant weight is attained. This weight loss is reported as fixed carbon,

expressed as a percentage of the original sample.

- Ash—The weight of the remaining material after combustion to constant weight is reported as ash content, expressed as a percentage of the original sample.

These quantities, however, are not absolute quantities present in the original coal. There does not exist an absolute method for determining the true moisture content of coal by heating, because of the wide variance in temperature at which different forms of moisture are liberated (surface, chemically combined, and water of hydration in clay minerals). Ash is a product of the oxidation process, during which iron pyrite, FeS_2, is converted to oxides of iron with liberation of oxides of sulfur, and other chemical changes. Hence ash is really not a component in the original coal. The fixed-carbon portion should not be confused with elemental carbon content, since a large amount of the carbon atoms in the original coal is released in the volatile matter, and some hydrogen and other elements may remain in the solid residue after completion of the volatile-matter evolution. The quantity of volatile matter evolved is a function of the maximum temperature achieved, the heating rate applied in achieving that temperature, and the time at that temperature.

Volatile-matter content, as determined in the proximate analysis, is very useful in relation to predicting combustion characteristics of a coal, and hence in burner design. In relation to assignment of rank designation, volatile-matter content is combined with percent carbon (elemental carbon) for the higher-rank coals. It also is useful in evaluating coals for coking. Thus, although not an absolute analysis, the proximate analysis is a very useful analysis.

5.3 Ultimate Analysis

The ultimate analysis of a coal reports the percentage by weight of the five major elements: carbon, hydrogen, oxygen nitrogen, and sulfur. The analysis is usually made on the coal after removal of moisture by the technique of the proximate analysis and is reported as being on a *dry basis*. The analysis is made on the coal with the mineral matter in the coal, sometimes reported on this basis, and sometimes reported after calculation to an *ash-free* basis. The usual procedure involves chemically determining the quantities of carbon, hydrogen, nitrogen, and sulfur, then calculating the oxygen content by difference. Hence the oxygen content as reported contains the accumulation of the errors that may exist in the other determinations and may be in error by as much as a few percent. An instrumental determination of oxygen, such as neutron activation, where available, is preferred. A prescribed procedure for the ultimate analysis for coal is given in ASTM-D3176. Inasmuch as the reported results are intended to be absolute quantities of these five elements, any accurate method of measurement is valid.

5.4 Physical Structure of Coals

During the metamorphic phase of the coalification process, some of the carbon, hydrogen, and oxygen present in the woody material was released. The peat mass was solid during this transformation; hence, pores were formed where these materials were released from the solid structure. The resulting coal thus contains pore spaces ranging in size from large macropores greater than 3000 Å diameter down to micropores a few angstroms in diameter. This pore structure gives rise to large internal surface areas, typically ranging from about 30 to 200 m^2/g or higher. The internal volume is also large, the greater portion usually being located in the large macropores, cracks, and fissures for lower-rank coals (less than 75% carbon), while for bituminous coals (75% to 84% carbon) pores smaller than 200 Å in diameter account for the major portion of the pore volume. Consequently, the bulk density of coal is typically of the order of 1.27–1.35 g/cm^3. The pore volume and pore size distribution in coal are important in governing the rate of diffusion of methane through and from a coal seam when the seam is opened. They are also of importance in processes for conversion of coal to liquids or gases.

Surface areas of coals are most often measured by adsorption of a vapor (most commonly N_2 at $-196\,^\circ C$ or CO_2 at $-78\,^\circ C$) into the pores and applying the Brunauer–Emmet–Teller theory to the adsorption isotherms. True densities of coals are usually measured by displacement of helium, methanol, or water. Pore volumes are calculated from true density measurements. Electron microscopy and x-ray diffraction have revealed additional features within the solid matrix in coals (Brunauer, 1945).

5.5 Chemical Structure of Coals

Because of the great heterogeneity of coal, it is not possible to draw a picture that represents all of the structural units in coal. It is not accurate to represent coal as a system of repeating units of the nature of a polymer. It is practical to try to represent types of structural units in coal which seem to be dominant, and the manner in which these units are joined in the solid matrix. Such a representation is helpful in visualizing how coal is attacked chemically in combustion or conversion processes.

Various chemical and instrumental techniques have been utilized in an effort to assemble such a representation. Generally each technique reveals but one small element of the picture. The sum of all of these experiments leads to the following description of the chemical structure of bituminous coal. In the organic portion of the solid matrix, which constitutes the major portion of the coal, from 60% to 75% of the carbon atoms are aromatic carbons, from 15% to 25% are saturated carbons located in rings with aromatic carbons, called hydroaromatic carbons (e.g., the saturated carbons in tetralin), and the remaining 5% to 10% of carbons are aliphatic hydrocarbon or other carbon functional groups. The dominant structures in the solid matrix are clustered-ring-type structures, on average about 2 to 3 rings per unit cluster for bituminous coals, somewhat less for sub-bituminous, even less for lignite, but much higher in anthracite. These ring-type clusters are joined principally by short aliphatic bridges and ether linkages. For a particular ring cluster, the number of covalent chemical bonds joining it to a neighbor in the plane of the cluster is fewer in a second dimension than in the first (perhaps one-fourth) and far fewer in the third dimension (perpendicular to the plane of the cluster), where hydrogen bonding and other van der Waals–type forces seem to predominate. Coal thus has a layered appearance at the unit structure level, but not layered in the same plane over long distances. In bituminous coals the oxygen is mostly phenolic and etheric, while in lignite a rather large number of carboxylic acid groups occur. In addition to the solid matrix, varying amounts of a *mobile phase* (organic material not chemically bound to the solid matrix) seem to be present, but always in amounts much smaller than the solid matrix. For purposes of attempting to visualize how coal must be attacked, particularly in conversion to liquids, the representation of structural groups and connecting bridges in bituminous coal (Fig. 1) has been prepared by the author.

5.6 Mineral Matter in Coals

Mineral matter was deposited along with the accumulation of plant material. It is usually quite uniformly distributed throughout the coal seam, as grains or thin bands. Coal typically contains between 4% and 20% mineral matter on a weight basis, although some coals, from certain portions of a seam, may be as high as 30% mineral matter. The dominant compounds in the mineral matter are usually SiO_2 and Al_2O_3, with some sulfates and pyrite (in most coals), and smaller amounts of carbonates and oxides of alkali and alkaline-earth metals. Trace amounts of radioactive elements may be present. Several of these compounds undergo chemical change during combustion; hence, ash produced in the analysis (by combustion) is quite different from mineral matter in the coal, as to both weight and chemical composition.

6. PROPERTIES OF COALS

6.1 Physical Properties

The physical properties of coal are a consequence of the cumulative effects of the nature of the original plant material, the manner in which the plant material accumulated, the severity of the metamorphism that determined the rank of the coal, and the specific form of plant debris that determined the petrography of the coal. Only a few of the most important physical properties of coals can be identified here, and these only superficially. For a detailed discussion of instruments and techniques, the reader is referred to Elliott (1981).

6.1.1 Reflectance. The determination of reflectance of a coal provides valuable infor-

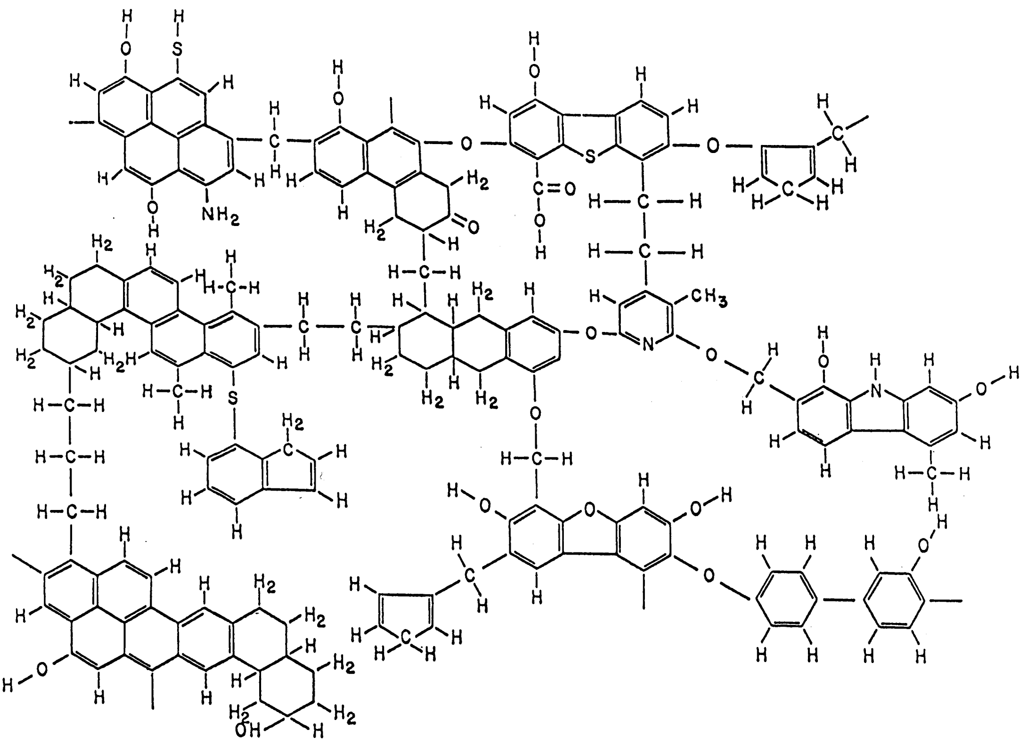

FIG. 1. Schematic representation of structural groups and connecting bridges in bituminous coal. [From Wiser (1978).]

mation concerning the origin of coal and coal structure, as well as predictions of behavior of a coal in a particular process. Various optical constants of a vitrinite, the principal component of most coals, can be derived from reflectance measurements. The optical constants of vitrinites in high-rank coals depend upon the orientation of the bedding planes with respect to the impinging radiation. A continuous increase of reflectance with rank of the coal is ascribed to the growth and ordering of aromatic layer clusters.

6.1.2 Porosity and Surface Area. These features of coals represent physical properties because of their importance in various uses of coal. The diffusion of reactant gases into coal pores during combustion, followed by reaction on the surfaces within these pores, is important to the combustion process. Diffusion of reactant gases and/or liquids, and sometimes catalysts, into the pores of the coal in coal liquefaction and gasification processes is vital to the success of such conversion processes. The formation of adsorbants from coal, such as activated carbon, relies upon the porosity and surface area of the parent coal in order to yield a product satisfying the requirements of high surface area.

6.1.3 Electrical Conductivity. Coals are generally classed as semiconductors, with specific resistance in the range $1\text{–}10^{14}$ Ω cm. The specific resistance decreases with increase in rank, the increase in rank being accompanied by an increase in mean aromatic layer diameter, then falls abruptly for coals of carbon content greater than 90% (the anthracite range). Specific resistance also decreases with increase in temperature, a behavior characteristic of semiconductors. The anisotropy observed for reflectance is also observed with electrical conductivity, the resistivity being much higher when measured perpendicular to the layer planes than when measured parallel to them.

6.1.4 Magnetic Susceptibility. The organic portions of coal are essentially diamagnetic, but traces of paramagnetism are also observed, possibly due to measurable unpaired electrons or free radicals in coal. By comparing diamagnetic susceptibilities of coals with studies on pure aromatic hydrocarbons, deductions have been possible concerning the number of rings in a condensed aromatic cluster in coal, yielding the value of 2–3 rings per cluster for bituminous coals, cited earlier.

6.1.5 Specific Heat and Enthalpy. Although a large number of experimental investigations have been conducted in measuring specific heats on coals, the data obtained are of limited value except for the coal for which the measurements were made. This is due to the heterogeneous nature of coals and the widely different compositions. The problem is aggravated by the occurrence of a plastic state for coals in the bituminous ranks and extending into the sub-bituminous rank. The degree of plasticity also varies widely.

6.2 Mechanical Properties

The mechanical properties of coals are important for coal mining in two ways: ensuring stable and safe mine operation and ensuring that coal is removed from the seam effectively. As greater reliance upon coal as a source of fuel and chemicals becomes necessary, the traditional trial-and-error methods of mine development must be superseded by added emphasis upon better knowledge of the mechanical properties of coals and their associated strata. This will make possible development of innovative methods to achieve higher productivity, greater recovery, lower costs, greater confidence relative to safety, and easier access to deposits in difficult locations.

Most classification of coals is based upon the "nonmechanical" characteristics, namely coal rank, carbon content, volatile matter, reflectance, gross specific energy, and petrographic composition. Attempts to correlate these characteristics with mechanical properties, or to use them to predict mechanical properties, are mostly unfruitful. Hence the application of nonmechanical characteristics of coals to mining needs are of little value. At the same time few data have been obtained by measuring and describing the overall mechanical properties of virgin coal seams. Measurements on coal seams would have to rely largely upon dynamic methods, e.g., sound waves and radar. Interpretation of such data for application to mine operation is difficult and somewhat imprecise at best.

6.2.1 Elasticity. When a hard coal (i.e., bituminous or anthracite) is subjected to a stress in compression, initially it exhibits a nonlinear stress–strain curve associated with closure of cracks, which is followed by an essentially linear elastic range. These observations are made upon coal samples that have been removed from the seam, and the initial nonlinear portion of the stress–strain curve may be the result of the mining operation rather than a property of the coal *in situ*. Coals exhibit anisotropy with respect to compression. The elastic modulus perpendicular to the bedding plane is significantly lower than that parallel to the bedding plane. However, within the bedding plane the moduli parallel to it appear to be independent of direction of applied stress around an axis perpendicular to the plane. Hard coal thus tends toward transverse isotropy.

6.2.2 Strength. Failure of a hard coal under stress is signified by an abrupt fracture and major structural change. The value of the stress that produces such a change is usually quite distinct, and defines the strength of the coal. Little attention has been given to the measurement of tensile strength, or shear strength, in coals. It is difficult to clamp a coal sample in a way to perform tensile and shear tests without altering these properties in the coal sample. On the basis of limited tests, the compression strength of coal is observed to be greater than its tensile strength by an order of magnitude. As a minimum, energy consumed in breakage of a solid must provide for creation of new surface and crack propagation. One such minimum value has been reported to be approximately 10 J/m^2.

6.2.3 Hardness and Abrasiveness. Hardness is a convenient measure of the tendency of the coal to cause abrasive wear, as well as of the deformability of the coal. Two techniques are most widely used for evaluating hardness in coals. In the rebound technique a hammer is dropped a certain distance to impact the surface of the coal, and the rebound of the indenter is a measure of the dynamic hardness of the coal. The height of drop must be controlled to ensure that the coal is not stressed beyond its elastic limit, at which point plastic failure occurs and the share of energy returned in rebound energy is reduced. In the static indentation technique a load is applied to the coal surface and the area of the permanent indentation is measured, the static hardness being expressed as the ratio of the load applied to the contact area of the indentation. No anisotropy has been observed in the hardness of coal samples.

The type, quantity, and distribution of minerals in coal determine its abrasiveness. Typically the hardest of these minerals are quartz (Vickers hardness numbers of 1100–1260) and pyrites (Vickers hardness numbers of 840–1130), compared with a hard steel at 600–700.

6.3 Plastic Properties of Coal

When bituminous coal is heated to temperatures above about 350–375 °C, the coal softens and becomes plastic. Both the rate at which plasticity develops and the degree of plasticity increase as the temperature increases. The appearance of the plastic state is accompanied by extensive evolution of organic molecules. If the coal sample is held at a temperature above 375 °C, the degree of fluidity increases to a maximum value, then decreases until the coal mass again becomes solid. The anthracites do not exhibit plastic behavior. Similarly the brown coals or lignites do not. Some degree of plasticity may be observed in some sub-bituminous coals, although to a much lesser degree than in bituminous coals.

In Sec. 5.5, bituminous coal is represented as consisting of clusters, both aromatic and hydroaromatic, joined by short bridges such as a methylene bridge. At temperatures above 350–375 °C, rupture of these connecting bridges becomes increasingly extensive, the rate increasing as the temperature increases. Utilizing electron spin resonance, an increase in free-radical concentration has been observed during the time when fluidity is increasing in the coal mass. Thermal bond rupture of the connecting bridges in this temperature range proceeds via a free-radical mechanism. During this time of increased concentration of free radicals and increasing fluidity, particle agglomeration occurs, forming a coke mass. The plastic coal mass adheres strongly to the walls of the container, leading to plugging. Both the agglomerating and plugging tendencies are the direct result of the formation and reactions of free radicals.

Various theories have been set forth to explain the existence and behavior of the plastic state. One can visualize the coexistence in the plastic mass of solids, liquids and gases, accompanied by a comparatively high concentration of free radicals. This is a time period of extensive thermally initiated chemical reactions. Particle agglomeration and growth are the consequence of stabilization of the free radicals via condensation reactions. Sticking to vessel walls is also a consequence of the behavior of free radicals. The fluidity is the manifestation, even down to a molecular level, of the coexistence of solids, liquids, and gases. When the chemical reactions are completed, and the volatiles evolved, the physical state becomes solid. Most of the organic material remaining has become a part of the solid mass or coke, the result of the agglomeration processes.

The formation of the plastic state in coal has been erroneously referred to as a melting phenomenon. The phenomenon is a chemical process that bears essentially no relationship to melting. Very serious problems have developed in reactor systems when there was an attempt to accommodate the behavior of the coal on the basis of melting, and real chemical processes were not recognized.

7. COAL PREPARATION

The principal objective of coal preparation is the separation of impurities from coal, most notably mineral matter (often, although incorrectly, called ash) and sulfur. A second objective, of growing importance, is to provide a coal feed of uniform particle size. The vast majority of coal utilization is in combustion. In recent years the urgent need to protect the environment during fuel combustion has dramatically increased the role of chemistry in coal preparation. Special emphasis is directed to the problem of sulfur oxide emissions from combustion of coal. Adverse effects resulting from sulfur emissions, as well as other negative impacts of coal combustion, are discussed in Sec. 11.

7.1 Size Reduction

Quite typically in a coal containing more than 2% sulfur, more than half of the sulfur is in the inorganic portion of the coal. For high-sulfur coals, most of the sulfur may occur as pyrites and sulfates in the inorganic portions. Pyrite and sulfate sulfur may be liberated by crushing. A physical separation technique, such as, for example, a sink–float operation, may then effect separation of the liberated sulfur-bearing particles from the organic portion of the coal. In many coals the total mineral matter content can be greatly reduced by physical separation following size reduction.

An increasing number of coal combustion operations utilize entrained flow or fluidized beds, or some other system where uniform particle size is important. For such processes the coal is crushed and mechanically cleaned, in order to liberate and separate impurities and provide a feed of uniform size. Whereas in the mid-1900s only a few percent of the total bituminous coal production was crushed, today more than 90% is crushed before use.

The dominant machine used for years for coal size reduction is the roll crusher. This machine feeds run-of-mine coal, usually delivering a fairly coarse product (e.g., 20-cm top size). The hammermill is the most commonly used coal pulverizer. Wet rod mills and ball mills are used or proposed for preparing a fine coal slurry for pipeline transport. For preparation of ultrafine coal, attrition mills and jet mills are of special interest. Several good reviews describing these and other techniques may be found in the literature. (For references, see Elliott, 1981, p. 578).

Of greatest theoretical and practical concern in size reduction are energy consumed in relation to change in size distribution, wear on machines, and size separation during transport. A mathematical model is used that attempts to correlate the characteristics of the mill and the operating variables to predict the nature and quality of the product. At this writing, modeling activity is demonstrating the ability to optimize size reduction in mineral processing at greatly reduced cost; potential for application to coal size reduction is obvious.

7.2 Beneficiation

Following size reduction to liberate impurities from the solid coal matrix, actual separation of the impurities from the coal is called *beneficiation.* A large number of processes have been devised and are used for this purpose, each with special features for specific applications. Some of these separation processes operate upon dry feed, while others are wet processes. A few examples, by name only, are jigs, cyclones (both wet and dry), concentrating tables, froth flotation, and high-gradient magnetic separation. A description of these and other beneficiation techniques may be found in the literature. (For references, see Elliott, 1981, pp. 580–585.)

7.3 Dewatering and Drying

The drive toward full mechanization of the coal-mining operation has resulted in an increase in production of coal fines and dust within the mine. Underground dust suppression operations add water to the coal leaving the mine. Furthermore, the widespread use of size-reduction and beneficiation operations, many of them wet processes, leads to a need to remove water from the coal. The increased moisture, if not removed, represents an economic burden in transporting the coal, as well as a reduction in heating value per pound of feed coal due to vaporization of the water during combustion. In addition, excess moisture may result in freezing during transport and increased difficulty in handling.

Examples of water-removal and moisture-reduction equipment, by name only, include screens, cyclones, centrifuges, filters, and thermal dryers. Interested readers may find a description of these and other techniques in the literature. (For references, see Elliott, 1981, pp. 587–594.)

7.4 Automation

In order to improve the competitive position of coal in the world energy mix, the need becomes urgent to increase operational efficiency throughout the full train of preparation processes, from preliminary storage to delivery to the marketplace. Automation represents a key element in the drive to improved efficiency, spurred by recent developments in instrumentation in relation to the various steps in the sequence of events. Mathematical models are used to simulate the various steps, which may be combined into one large model defining a complete beneficiation process. Optimization procedures utilizing computers make possible the preparation of flow sheets for new plants, identifying optimum parameters as dictated by required yields at specified ash and/or sulfur contents. (For references, see Elliott, 1981, pp. 596–597.)

8. COAL CARBONIZATION

The term *carbonization*, when applied to coal, generally means heating coal to a temperature above about 400 °C, in an inert atmosphere, yielding as the principal product a solid richer in carbon than the initial coal. This carbon enrichment is the result of expelling from the solid mass gaseous and liquid compounds that are richer in hydrogen, or oxygen, or perhaps other elements, than was the starting coal. The term *pyrolysis* is often used in connection with the same type of heating operation with coal. Although the distinction between the two terms is not clear-cut, and the uses not universal, in general the term carbonation is applied when the solid, carbon-rich product is of the greatest interest, and pyrolysis is applied when the volatile products to be liberated by heating are of primary interest. Thus the primary product of the operation when viewed as carbonization becomes the by-product when viewed as pyrolysis.

8.1 Principles

The chemical structure of coal is described in Sec. 5.5. When coal is heated, at temperatures above 375 °C the connecting bridges between clusters begin to rupture. The degree of rupture increases with temperature, becoming essentially complete at 500–550 °C. The free radicals thus formed seek stabilization by such means as may be available to them. Principal among the options is the abstraction of hydrogen from a hydroaromatic structural unit in the vicinity. If the unit thus formed exhibits a sufficiently high vapor pressure, it will evolve as a volatile product. A direct correlation between hydroaromatic hydrogen in the coal and the quantity of evolved product has been experimentally observed. When all such options for producing stable molecules of sufficient vapor pressure have been exhausted, the balance of the skeletal structural units of the coal undergo condensation reactions that result in a solid product, even at the elevated temperature, called char or coke. In practice, at all temperatures considerably more than half of the coal by weight remains as char or coke.

8.2 Low-Temperature Carbonization

Low-temperature carbonization is defined as occurring at temperatures between the onset of thermal bond rupture, as described in Sec. 8.1, and about 750 °C. In this temperature range, in general, hydrogen is not released from aromatic carbon atoms to a significant extent. The formation of carbon–carbon bonds occurring in the production of char or coke in this temperature range is limited chiefly to the carbon atoms involved in the rupture of bridges between condensed configurations, in removal of side chains or of functional groups from these condensed configurations, and in other weak bonds involving ring carbons that can be thermally ruptured in this temperature range. The resulting char formed is not very strong because of limited cross-linking, thereby restricting its uses.

8.2.1 Products and their Uses. The quantity and properties of char produced are a function of the coal being carbonized and the temperature of carbonization. In general, low-temperature carbonization processes operate at temperatures between 450 and 700 °C. Char yields range from about 60% to 75% by weight of the coal. In most cases the volatile by-products are about equally divided between liquids and gases.

Low-temperature carbonization has usually been practiced primarily to produce a char or coke. In general these cokes are irregular in structure and in shape, and dull black in appearance. They are less dense and weaker than high-temperature coke, and hence tend to be rather dusty. Because of the moderate temperature of low-temperature carbonization, the volatiles are only partially expelled, as related to the standard volatile-matter test performed at 950 °C. The volatile-matter content of low-temperature coke ranges from 7% to 20% by weight of the coke.

Probably the most valuable characteristic of low-temperature coke is its high reactivity. This property, coupled with the moderate content of volatiles, results in a fuel that ignites easily and burns reliably. These cokes are very good fuels for use in furnaces, space heaters, and fireplaces. While they are good solid fuels for some foundry uses, their lack of strength precludes their use in the blast furnace, or other uses where a substantial overburden must be supported. Because of their high porosity and reactivity, they are also advantageous for use in production of water gas.

A typical low-temperature carbonization operation may yield as liberated volatile products about 80 to 120 standard cubic meters of gases with heating values of the order of 30 000–37 000 kJ/m,^{3} 9.5 to 11 L of light oil, and 50 to 120 L of tar, all per ton of coal carbonized. A large number of retort processes have been developed, most of them prior to 1960. The interested reader will find a good review of these processes in Lowry (1963).

In recent years carbonization of coal has been under development under the name of pyrolysis, with production of liquid fuels and/or chemicals as the principal objective. This potential application, with this objective, is discussed in Sec. 10.2.1.

8.3 High-Temperature Carbonization

The primary objective of high-temperature carbonization is the production of a strong and uniform coke for use in metallurgical operations, such as the blast furnace. To achieve this objective, temperatures in the range 900–1300 ° C are required in the carbonization process. Consideration of the fundamental chemistry involved reveals that a strong coke cannot be produced at lower temperature. In practice coke ovens operate at temperatures up to 1300 °C, the upper limit controlled by safe operating limit of the wall brick (now usually silica brick).

8.3.1 Metallurgical Coke Manufacture. The required physical properties of a metallurgical coke can be illustrated by reference to the blast furnace in reduction of iron ore. Three essential ingredients are fed to the top of the furnace, in sequence, namely coke, iron ore, and limestone. The iron component of the ore, chiefly Fe_2O_3, is to be reduced to metallic iron, by interaction with a reducing agent, principally gaseous carbon monoxide:

$$Fe_2O_3+3CO \rightarrow 2Fe++3CO_2.$$

The CO is produced by oxidation of the coke with a limited supply of air according to the reactions:

$$C+O_2 \rightarrow CO_2,$$

$$C+CO_2 \rightarrow 2CO.$$

The limestone acts as a flux to remove silicates and certain other minerals present in the ore, forming a slag that is molten at furnace temperatures. The air for reaction with the coke is introduced at the bottom of the furnace. The coke must maintain its structural integrity as it moves downward through the furnace. If it breaks and crumbles, the bed porosity will be lost and the efficiency of reduction of the iron ore greatly reduced. For satisfactory operation the coke must be strong to support the overburden above, uniform to provide for the chemistry of ore reduction throughout the cross section of the bed, and porous for rapid reaction with the air.

The strength of the coke relates directly to the degree of cross-linking between the condensed ring structures of the coal, which occurs during the carbonization process. At temperatures above 800 °C, the hydrogen atoms present on aromatic carbons increasingly are thermally released, providing opportunities for increased cross-linking. In the temperature regime of 1200–1300 °C this hydrogen release nears completion and a maximum degree of cross-linking is achieved. A high degree of plasticity, with attendant expansion of the coal mass during formation of plasticity, is always observable during production of a metallurgical coke of the required strength. Anthracite, brown coal or lignite, and in most cases sub-bituminous coal do not exhibit plasticity, and will not produce a strong metallurgical coke. Medium-volatile bituminous coal will yield strong metallurgical coke when carbonized alone. High-volatile bituminous coal can yield a strong coke when blended with an appropriate amount of low-volatile bituminous coal. The horizontal slot-type oven, also known as a by-product coke oven, is the most important and most used means of producing high-quality blast furnace or foundry coke. For a technical description of the horizontal slot-type oven, see Chap. 14 of Elliott (1981).

8.3.2 By-Products and their Uses. A typical yield of volatile products from a horizontal slot-type coke oven is 250 to 350 standard cubic meters of gases, 40 to 60 L of tar and light oil, and 2 to 4 kg of ammonia, all expressed per ton of coal carbonized. Principal components in high-temperature coke-oven gas, expressed on a volume basis, are hydrogen, about 60%, and methane, about 30%. Most of the balance consists of combustible hydrocarbons. This gas can be used as a fuel with heating value of about 20 000 kJ/m^3. Coke-oven gas is frequently separated into products with greater market value than the value in combustion.

Benzene, toluene, and the xylenes are the principal components of the light oil. Separation techniques most often employed are distillation, absorption in liquids, absorption by solids, and compression and cooling of the gas.

The tars contain a very great number of compounds, including phenols, cresols, naphthalenes, and benzene and its homologs. The tar fractions are marketed in such products as tar acids, tar bases, anthracene oil, creosote, and pitches. Separation of the tars into these fractions is usually accomplished by distillation. The pitches constitute the largest fraction, and are used for fuel, for road and roofing tars, and for manufacture of other pitch products. Creosote is highly toxic to wood-destroying fungi and insects. Wood preservation is its most important application.

Ammonia is recovered from coke-oven gas. It is the principal nitrogen-containing ingredient in agricultural fertilizers, which products represent a major market for ammonia.

9. COAL COMBUSTION

For several centuries coal has been an important energy resource. Early in the 19th century it began to replace wood as a fuel, and later in that century became the dominant fuel in industrialized countries where coal resources existed. As petroleum and natural gas were developed, they in turn replaced coal as the dominant fuels in many uses. However, in countries where coal resources occur, coal retains a dominant position in a few uses, most important among them being raising steam for electric power generation. As an example, at this writing 80% of all coal consumption in the United States is in electric power generation, and 56% of electric power in the United States is generated from the combustion of coal. Coal combustion, at this writing, supplies 23% of the total energy consumed in the United States, and is increasing its percentage of the total energy supply.

In recent years, particularly since the close of World War II, a substantial effort in research and development has been directed toward improving the efficiency of coal combustion. Initially these efforts were motivated by economic considerations, but more recently environmental and health considerations would seem to represent the principal motivators. The most recent and sophisticated coal combustion technologies of today include provisions for major reductions in unburned carbon and sulfur emissions, when compared with pre–World War II technologies.

9.1 General Principles

Coal contains widely varying amounts of volatile materials ranging from as low as 4% by weight for anthracite to more than 40% by weight for some bituminous and sub-bituminous coals. Most of these volatiles are combustible. The volatile matter content of a coal greatly influences its combustion characteristics, particularly the ease of ignition and of sustaining combustion, thereby dictating to some degree the burner design for combustion of a coal. As a quantity of coal is heated en route to its ignition temperature, volatiles are liberated from the coal by action of the heat applied. The combustion process thus becomes concerned with three basic phenomena, namely pyrolytic liberation of volatiles, combustion of carbon, and parallel or sequentially prior combustion of volatiles. When one desires to study the fundamentals of the overall combustion of coal, one usually examines these three aspects separately. While the solid matrix that remains as volatiles are liberated is predominantly carbon, usually called char, it is not graphite. The most important reactions involved in coal combustion are those occurring with the char. The char reacts at both external and internal surfaces, at rates determined by both chemical and physical processes. An excellent, detailed treatment of the studies in coal combustion may be found as Chap. 19 in Elliott (1981).

9.2 Coal-Combustion Systems

The principal objectives in the design of coal-combustion equipment are to achieve continuous economical operation, while minimizing the discharge of sulfur oxides, nitrogen oxides, and particulates to the environment, at the same time avoiding slagging and/or fouling problems resulting from the ash. In practice the coal is either suspended on a grate or in the air. Air-suspension combustion systems involve either fluidization of the coal bed with the air or entrainment of the coal in the air stream.

9.2.1 Combustion on a Grate. In the more common grate system, raw coal is fed to the fuel bed from above the fire. As combustion proceeds, the solid material yet unreacted moves downward, progressing through several stages, with the mineral matter (ash) passing through the grate. Primary air is admitted from beneath the grate, moving upward and countercurrent to the stream of solids. A series of zones may be visualized in the bed. Immediately on top of the grate is an ash zone, essentially all combustible material having been removed. Next above the ash zone is an oxidizing zone. It has been experimentally observed that the reaction between the fixed carbon of the solid and oxygen, occurring in the vicinity of the solid, proceeds to CO_2 via the overall reaction

$$C+O_2 \rightarrow CO_2.$$

Next above the oxidation zone occurs a reduction zone in which the following reaction occurs:

$$CO_2+C \rightarrow 2CO.$$

The top layer, containing the raw coal, responds to the high heat evolution from below, and the volatiles in the raw coal are liberated. Secondary air is admitted above the bed providing for the combustion of the volatile material (VM) and the carbon monoxide from the reduction zone:

$$(VM+CO)+O_2 \rightarrow CO_2+H_2O.$$

The volatile matter and carbon monoxide reaching this region above the raw coal layer may represent as much as 60% of the total heating value of the original coal. A high degree of turbulence in this region, involving the secondary air, is required to achieve good combustion efficiency as well as to avoid excessive smoke.

Mineral matter from the coal is a potential source of mechanical difficulty. In a top-feed grate system the ash layer on top of the grate, assisted by the cool primary air stream, insulates the grate from the very high temperatures of the combustion zone above, thus avoiding melting of the cast-iron grate. If temperatures are too high, fusion of the ash occurs (slagging), forming clinkers which may not pass through the grate. Furthermore, the slag may cause fouling of surfaces, or corrosion of key elements of the combustion system.

9.2.2 Atmospheric Fluidized-Bed Combustion. If coal is crushed, and a feed stream to a combustion reactor selected to contain only particles smaller than about 6 mm diameter (the usual size limitation) the upward velocity of the air, admitted at the bottom of the coal bed, can be increased until the coal particles begin to be lifted by the air stream. As the gas velocity increases, the pressure drop ΔP through the bed increases, reaching essentially a constant value. This condition defines the minimum fluidizing velocity. Operation will be conducted at some velocity above this value. The bed into which the crushed coal is fed may consist of an inert material, such as sand or ash, or a chemically reactive material, such as limestone or dolomite. The heat-transfer surface, such as tubes in which water is to be converted to steam, is submerged in the bed. The operating temperature in the bed is generally between 750 and 1000 °C.

The fluidized bed provides several advantages relative to problems historically associated with coal combustion for steam raising.

1. Fluidization of small coal particles with air establishes good contact between coal particle and oxygen, resulting in improved combustion efficiency. Good combustion efficiency has been achieved with excess air quantities as low as 2%.
2. Fluidization of coal in a bed containing limestone achieves removal of sulfur from the combustion product gases according to the reaction

$CaO+SO_2+\frac{1}{2}O_2 \rightarrow CaSO_4.$

3. The fluidized bed provides for maintaining a uniform temperature, characteristically about 870 °C, low enough to limit formation of NO_x. At this temperature, NO_x formation appears to result almost wholly from fuel-bound nitrogen, rather than fixation of air nitrogen.
4. The uniform temperature in the fluidized bed remains below the melting temperature of the mineral matter of the coal, thus avoiding slagging with its attendant problems of corrosion and fouling of surfaces.

9.2.3 Pressurized Fluidized-Bed Combustion. When a fluidized bed is operated for coal combustion at elevated pressures up to 10 atm, the principles and advantages presented in the preceding subsection apply. A slight improvement in reduction of NO_x formation is also observed at the elevated pressure. However, interest in operating with increased pressure in the combustion zone is born of the realization that as combustion pressure increases, the size of the combustor decreases, as also does the area of heat-transfer surface required for maintaining a specified bed temperature. These factors reduce investment cost. Furthermore, if electric power generation is an objective of the combustion operation, the combustion gases at elevated pressures (e.g., 10 atm.) may be expanded in a gas turbine, thus providing a form of *combined cycle* generation system and resulting in an improved power generating efficiency (see Sec. 10.1.7).

9.2.4 Entrained-Flow Combustion. Large industrial boilers often operate at a steam output demand of a few million kilograms of steam per hour. Certain other processes require temperatures in the combustion zone of 1500 °C or higher. Pulverized-coal (entrained flow) firing is readily adaptable to these demands, and has gained prominence in such uses. The coal is crushed to a very fine size, the feed typically being about 80% through 200 mesh, Tyler screen series (-74 μm), and is transported into the combustion chamber entrained in the primary air supplied for combustion. A high degree of turbulence is provided resulting in excellent combustion efficiency. Temperature in the combustion chamber is typically 1600 °C or higher.

Combustion of the char or fixed-carbon portion of the coal particle (the solid remaining after evolution of the volatiles) occurs at the surface of the particle. For particles larger than 65 μm the surface flux of volatiles is large enough to prevent oxygen from reaching the particle surface, and combustion of the volatiles precedes solid burning. The surface flux of volatiles decreases as particle size decreases. For particles between 65 and 15 μm diameter, gas phase and solid burning proceed together. For particles smaller than 15 μm in diameter, the initial burning appears to be completely heterogeneous. In practice, in pulverized-coal-fired furnaces most of the coal fired is smaller than 65 μm, and 20%–30% is smaller than 15 μm. Not more than 2% is larger than 300 μm. These burning patterns for very small coal particles provide for extremely high heat flux, with resulting high temperatures.

10. SYNTHETIC FUELS AND OTHER PRODUCTS FROM COAL

As discussed in Sec. 1.2, known and recoverable reserves of coal in the world are very great. Coal has been used as an energy resource much longer than petroleum and natural gas. But coal has suffered a decline as an energy resource, both with respect to total consumption and as a percentage of consumption. It is generally recognized that coal must play a larger role in supplying the energy needs of the industrialized nations as reserves of petroleum and natural gas decline. However, there are several reasons why substituting coal for either petroleum or natural gas is difficult. Principal among these reasons are these:

1. Coal is dirty to use—when burned it emits both solid and gaseous pollutants to the environment (see Sec. 11) in large quantities.
2. Coal is much less convenient to distribute and to use than are petroleum and natural gas, because it is a solid.
3. Many areas of energy demand in modern society, such as transportation by auto, truck, or airplane, as well as certain uses in the industrial sector, simply cannot use a fuel in solid form.

For these, as well as several other reasons, if coal is to be extensively substituted for petroleum and/or natural gas, it must first be converted to a liquid or a gaseous form.

10.1 Gaseous Fuels from Coal

Coal reacts with steam to produce combustible gases. At temperatures of 950 °C or higher, the reaction proceeds noncatalytically at a satisfactory rate, according to the reaction

$$\text{coal (mostly carbon)} + H_2O \rightarrow CO + H_2. \quad (1)$$

The reaction between carbon and steam is strongly endothermic in the amount of 136 kJ/g mol of carbon determined at 925 °C. Energy in the form of heat must be supplied to raise the temperature of the coal bed to 950 °C and to hold the temperature at that value, as well as to sustain the endothermic reaction. The usual practice is to introduce either air or oxygen to the gasifier in a controlled amount, to provide for the following overall reaction:

$$C + \tfrac{1}{2}O_2 \rightarrow CO. \quad (2)$$

This reaction is exothermic in the amount −110.5 kJ/g mol of carbon reacted, at 25 °C. CO as a fuel gas exhibits a heat of combustion of −283 kJ/g mol, and hydrogen exhibits a heat of combustion of −285.9 kJ/g mol, both determined at 25 °C.

Three additional reactions may occur in a coal gasifier, depending upon reaction conditions. Steam added as a reactant may react with CO produced in one or more reactions occurring in the reactor, according to the reaction

$$CO + H_2O \rightleftarrows CO_2 + H_2. \quad (3)$$

This reaction, known as the water–gas shift reaction, is exothermic in an amount of −33.5 kJ/g mol, determined at 700 °C. The reaction exhibits an equilibrium constant K_p equal to 1 at 827 °C, with K_p increasing as the temperature decreases.

When hydrogen is present in the gasifier, it is possible that some methane may be formed by the following reaction:

$$CO + 3H_2 \rightleftarrows CH_4 + H_2O. \quad (4)$$

This reaction is exothermic in an amount −225.9 kJ/g mol, determined at 700 °C. The equilibrium constant K_p increases as the temperature decreases. In addition, hydrogen may react directly with carbon according to the following reaction:

$$C + 2H_2 \rightleftarrows CH_4. \quad (5)$$

This reaction is exothermic in an amount −91.4 kJ/g mol, determined at 700 °C. The reaction exhibits an equilibrium constant K_p equal to 1 at 546 °C, with K_p increasing as the temperature decreases.

10.1.1 Producer Gas. Perhaps the simplest concept for producing a gaseous fuel from coal, when only a gaseous fuel is desired, is to react coal with a limited quantity of air, according to Eq. (2) above, yielding a gas consisting essentially of CO and the nitrogen that entered with the air. Typically the heating value of such a gas, called producer gas, is about 4460 kJ/m^3 [120 Btu/standard cubic foot (SCF)]. With such a low heating value, this fuel gas is only practical for use at or very near the site where produced.

10.1.2 Water Gas. The objective is to burn a portion of the coal, the carbon proceeding completely to CO_2, to provide heat to sustain the steam–carbon reaction of Eq. (1). The process is cyclic. In the first, or "blow" portion, excess air is blown through the fixed coal bed, the combustion increasing the bed temperature to a value above 950 °C, typically about 1300 °C. The exit gas from the blow is principally CO_2, H_2O, and N_2, with no heating value. The air is discontinued and steam is blown under pressure through the hot coal bed. The reaction proceeds according to the endothermic reaction of Eq. (1) until the bed has cooled to a temperature where the reaction rate is too low. The gas from this portion of the cycle contains principally CO and H_2, with a heating value of about 11 297 kJ/m^3 (300 Btu/SCF) and is collected as the product fuel gas. The cycle is then repeated.

10.1.3. Low-Heating-Value Gas. Air and steam are introduced simultaneously into the coal feed, with the intent that a portion of the carbon in the coal will react with the limited air to yield CO according to Eq. (2), while the balance of the carbon in the coal will react with steam according to Eq. (1) to yield CO and H_2. Some gasifiers (e.g., Lurgi) operate with an exit gas temperature low enough to permit formation of a significant amount of methane according to Eqs. (4) and (5). Other gasifiers (e.g., Koppers-Totzek) operate at high temperatures in order to obtain maximum efficiency in decomposition of the steam, but producing no methane because of the high temperature. The product gas consists of CO, H_2, and perhaps some CH_4, but is diluted by the N_2 that entered in the air. The heating value of the product gas ranges from about 4854 kJ/m^3 (130 Btu/SCF) to about 6695 kJ/m^3 (180 Btu/SCF), depending upon the amount of CH_4 in the product gas.

10.1.4 Intermediate-Heating-Value Gas. In order to produce a fuel gas of higher heating value, by eliminating nitrogen and its diluting effect from the product gas, oxygen is first separated from the nitrogen and other components of the air, usually by distillation of liquid air. The procedure is then similar to that of Sec. 10.1.3. In this case the product gas contains principally CO, H_2, and perhaps some CH_4, depending upon temperatures in the gasifier. Heating values of product gases range from about 10 460 kJ/m^3 (280 Btu/SCF) to about 12 970 kJ/m^3 (350 Btu/SCF), depending upon the amount of CH_4 in the product gas.

10.1.5 High-Calorie Gas or Substitute Natural Gas. The ratio of H_2 to CO in the product gas of Sec. 10.1.4 is usually considerably less than 1. In order to produce CH_4 by the hydrogenation of CO according to Eq. (4) one must first increase the ratio of H_2:CO to at least 3:1. This can be done by the water–gas shift reaction of Eq. (3). If the equilibrium constant K_p of Eq. (3) is to be great enough to achieve a ratio of H_2:CO of 3:1, the temperature must be no greater than about 430 °C in the shift converter. At this low temperature a catalyst (e.g., Fe_3O_4) is required to achieve a satisfactory reaction rate. In practice, this reaction is continued to a H_2:CO ratio of about 3.1:1. This gas, enriched in H_2, is then reacted over a nickel catalyst at a temperature of about 450 °C to yield a product gas of high CH_4 content (usually greater than 95% CH_4, with CO content less than 0.1%) and a heating value greater than 35 150 kJ/m^3 (950 Btu/SCF). This product gas is completely interchangeable with natural gas as far as its behavior in a burner is concerned, hence its name substitute natural gas (SNG).

Many technologies have been developed to achieve a high-calorie gas product from coal. The steps in the process do not always appear in the same order, but the chemistry is essentially that described by Eqs. (1)–(5).

The chemistry described above is initiated by the noncatalyzed reaction of steam with carbon at temperatures of 950 °C or higher. One may conduct the steam–carbon reaction at a considerably lower temperature (e.g., 700 °C) by employing an appropriate catalyst (e.g., potassium). Under such conditions the yield of CH_4 in the gasifier may be great enough to represent the only net product from the process. The basic chemistry occurring is as represented in Eqs. (1)–(5) above. At temperatures above about 800 °C, the reaction between steam and carbon tends to be as shown in Eq. (1). As the temperature is decreased, a different reaction pathway begins to appear, namely

$$2C + 2H_2O \rightarrow CH_4 + CO_2. \qquad (6)$$

At a temperature of 260 °C, the reaction tends to be mostly as indicated by Eq. (6), and is almost thermally neutral.

If the reader is interested in technologies that have been developed to carry out the chemistry indicated in Eqs. (1)–(5), a description of those processes may be found in Elliott (1981), Chap. 24.

10.1.6. Underground Gasification. As discussed in Sec. 1, large quantities of coal have been identified that are not considered to be economically recoverable by mining techniques. In particular, very thin seams less than 35 cm thick, and seams with overburden greater than 900 m, are not considered to be minable. For such coal seams, as well as certain other special types of resources, the chemistry identified in Eqs. (1)–(5) and more especially Eqs. (1) and (2) may be carried out *in situ*.

The technique involves drilling a well into the coal seam through which is injected air, and water or steam if the moisture content in and adjacent to the coal seam is insufficient for the gasification process. A second well is drilled at an appropriate distance from the injection well, from which product gases will be removed. Ignition may occur at the bottom of the injection well (forward burn) or at the bottom of the product recovery well (reverse burn). With air injection, the product gas is similar in composition to a low-heating-value gas, described above, but with a heating value slightly lower, generally ranging from 3350 to 4850 kJ/m^3 (90 to 130 Btu/SCF). If the overburden is tight enough to prevent loss of the injected gases, oxygen rather than air may be injected, thereby producing a gas similar in composition to an intermediate-calorie gas, but again of slightly lower heating value. (For processes, see Elliott, 1981.)

10.1.7 Integrated Gasification Combined Cycle. An attractive application for intermediate-calorie gas is electric power generation. If this fuel gas is burned to produce steam, and the steam expanded in a turbine to drive a generator, the overall efficiency is limited by the temperature and pressure of the turbine operation, about 38% in most fossil-fuel plants. An alternative is to expand the hot gasification-product gases in a turbine, generating some electric power by that operation. This is followed by the ordinary fuel-gas-combustion–steam cycle normally operative in fossil-fuel power plants. This combination of gas turbine and steam turbine power generation results in overall efficiency improvement of several percent, e.g., from 38% to 42–44%, a significant improvement. The concept is known as combined cycle and the system is integrated gasification combined cycle (IGCC).

10.2 Liquid Fuels from Coal

Coals of interest for conversion to liquid fuels generally exhibit an atomic hydrogen: carbon ratio of 0.8 to 0.9, with a few coals approaching a value of 1.0. Liquid fuels derived from coal must have atomic H:C ratios greater than 1.0, in order to be liquids. The atomic H:C ratios of a few fuels with which liquid fuels from coal would be compared are listed in Table 3 with characteristic values of atomic H:C ratios. Referring to Fig. 1, the bridges that join the condensed units in coal begin to rupture thermally at about 350 °C, and the rupture of the bridges is essentially complete at 450 °C. Nearly all direct coal liquefaction processes are carried out near 450 °C, or slightly higher, in order to realize full benefit of this thermal action. If the free radicals thus formed are stabilized immediately, as by addition of a hydrogen atom, a stable molecule is formed, which may be of such size (vapor pressure) as to become part of a liquid fraction, or at least be soluble in the liquids present. Inasmuch as the objective is to form liquid-size molecules, rather than something larger, stabilization of the free radicals with hydrogen is usually the preferred practice.

Hydrogen is an expensive commodity and represents a large expense in the conversion of coal to liquids. The heavy, raw liquid formed in direct coal conversion processes does not represent a large consumption of hydrogen. Starting from a bituminous coal with an atomic H:C ratio of 0.90, the primary, raw liquid product typically has an atomic H:C ratio of about 1.1. However, a considerable amount of hydrocarbon gases may be produced during the conversion process (accompanied by large consumption of hydrogen), the principal component of which is CH_4. The ratio of liquids to hydrocarbon gases usually ranges from about 1 to about 4, depending upon the process approach. The various approaches presented below for conversion of coal primarily to liquids are classified to a degree on the basis of the manner in which hydrogen is provided for stabilization of the thermally produced free radicals.

10.2.1 Coal Pyrolysis. By definition, coal pyrolysis is the process of heating coal in the absence of a reacting atmosphere, and without an externally applied catalyst. A variation of pyrolysis is hydropyrolysis, wherein coal is heated in the presence of molecular hydrogen, but still in the absence of an externally applied catalyst. In a typical pyrolysis process for production of liquids, the coal is heated to a temperature between 450 and 525 °C. As the

Table 3. Characteristic atomic H:C ratios of some liquid fuels.

	Atomic H:C ratio
Bituminous coal (as reference)	0.8–0.9
Naphthenic crude (petroleum)	1.6
Unleaded gasoline	1.6
Diesel fuel	2.1
Fuel oil	1.8

connecting bridges rupture, the free radicals thus formed abstract hydrogen from the hydroaromatic structures in close proximity to the bond rupture, thereby forming molecules that may be evolved as volatiles. For a bituminous coal the volatile product yield, liquids plus gases, is 35%–40% by weight of the dry, mineral-matter-free (dmmf) coal, about equally distributed between liquids and gases, resulting in about 1 barrel (42 U.S. gallons) to 1.25 barrels of liquid per ton of bituminous coal. If heating rates and product removal are very rapid, thus reducing the time for the liquids in the hot zone, the liquid yield may be increased by decreasing thermal cracking of the liquid molecules to form gases.

In hydropyrolysis, heating of the coal is conducted in a hydrogen atmosphere. In the temperature range for liquids production as the desired product (e.g., below 525 °C), hydrogen does not react very readily to stabilize the thermally produced free radicals without the aid of a catalyst, hence volatile yields are not greatly increased in this temperature range compared with ordinary pyrolysis. At higher temperatures, e.g., 600 °C and higher, hydrogen reacts more readily, increasing conversions substantially, with most of the increased yield being gases.

10.2.2 Solvent Extraction. In solvent extraction processes the coal is first slurried in a hydrogen-donor solvent, usually coal-derived. The slurry is pumped into a reactor at a temperature usually between 430 and 450 °C, and hydrogen pressure of 14 MPa (2000 psi) or higher. An external catalyst is not applied with the raw coal. As the free radicals are thermally produced, hydrogen from two sources is available for radical stabilization, namely from the hydroaromatic structures of the coal, and from hydroaromatic structures in the donor solvent. The former usually enters into the reactions before the latter, probably because it is nearer to the bond-rupture sites. Most of the organic portion of the coal is dissolved, and the remaining solids, consisting of the mineral matter and a few percent of unreacted organic matter, are removed while the product is still at the elevated temperature (e.g., 430–450 °C). Solid-removal techniques include flash distillation, critical solvent de-ashing, filtration, hydrocyclone, and centrifuging.

When addition of hydrogen from external sources is limited to about 2% by weight of coal, and the solvent removed, the remaining liquid at the elevated temperature becomes solid at ambient temperature, with a solidification temperature (or fusion point) of about 175 °C. This product is called solvent-refined coal (SRC). Usually the mineral matter content has been reduced to about 0.1% to 0.5% by weight, and the sulfur content to about 0.7% (from a coal of 3% sulfur or higher). The net SRC product represents about 56% of the dmmf coal fed, the balance being recycle solvent and gases.

If addition of hydrogen is increased to 3.5%–4.0% by weight of coal, a product that is liquid at ambient temperature is obtained. The ratio of liquids to HC gases is about 3:1 to 4:1. Again, most of the hydrogen consumed appears in the gases. For a self-sustained process, after providing for energy input and hydrogen production, the net liquid-product yield is about 2.5 barrels per ton of coal. The practice of slurrying the coal in a hydrogen-donor solvent achieves three important objectives:

1. permits pumping the coal into the pressurized region,
2. controls agglomeration of a caking coal, and
3. provides a source of hydrogen without adding a catalyst.

10.2.3 Catalytic Liquefaction. Catalytic liquefaction involves the addition of a catalyst to the system to assist in the conversion of coal to liquids. Usually the coal is slurried in a hydrogen-donor solvent, process-derived, as described for solvent extraction processes. Temperatures and hydrogen pressures are similar to those identified for solvent extraction. The conversion process is initiated by thermal bond rupture in the connecting bridges between clusters in the coal structure. In this system three sources of hydrogen are available for stabilization of radicals and other reactions that may occur, namely hydrogen from the hydroaromatic structures in the coal, hydrogen from the donor solvent, and hydrogen from the gas phase as may be effected by the catalyst. If the catalyst is a solid phase, such as cobalt-molybdenum supported on alumina, contact between the coal units and catalyst is limited until the process of dissolving the coal in the donor solvent is well advanced. Thus the catalytic hydrogenolysis occurs in the later stages of the liquefaction process. Certain catalysts may be dissolved in a solvent, then impregnated into the coal pores prior to coal heat-up. A vapor-phase catalyst may achieve penetration into the coal pores prior to heat-up. In these cases catalytic hydrogenolysis may begin as soon as the appropriate temperature and pressure are achieved.

In catalytic liquefaction the catalyst encounters a severe environment. Not only must it withstand deactivation by sulfur and nitrogen in the coal, but there are also many elements, especially metals in the mineral matter of the coal and organometallic structures, which may "poison" active sites on the catalyst, perhaps making regeneration of the catalyst impractical, or even impossible. Essentially all of the organic portion of the coal is converted to liquids and gases, with ratios of liquids to gases similar to solvent extraction (about 3:1 to 4:1). Solid-removal techniques are the same as identified above for solvent extraction. Gross liquid yields up to four barrels per ton of coal may be obtained, but net yields per ton of coal, for a self-sustained process, are similar to those for solvent extraction, although they may be of somewhat higher quality due to the activity of the catalyst. In this case hydrogen consumption is also higher.

10.2.4 Synthesis Gas from Coal. In Secs. 10.1.3 and 10.1.4, conversion of coal to a gas consisting principally of carbon monoxide and hydrogen is described. When that gaseous product is intended for use as a fuel, it is referred to as intermediate-calorie gas. If the intended use of the gas is to produce other products, such as the high-calorie gas of Sec. 10.1.5, or alcohol, or a host of other chemicals, or other liquid fuels, that same gas, produced by the same technology, is called synthesis gas. The pathway followed in further reaction of the CO and H_2 and the resulting products depends upon the function and selectivity of the catalyst used, and the reaction conditions.

10.2.5 Indirect Liquefaction. Coal is reacted with oxygen and steam by the reactions of Eqs. (1) and (2) of Sec. 10.1. The ratio of H_2 to CO present in the gas thus produced may range from 1:2 to 2:1, depending upon the technology and temperature used to produce the synthesis gas. This gas is then reacted over a catalyst [Eq. (3)] to increase the H_2:CO ratio. The ratio of H_2:CO required depends upon the intended product, which in turn dictates the required temperature: the higher the required H_2:CO ratio, the lower the temperature. The CO may then be hydrogenated to a wide range of products, with the choice of catalyst providing some measure of selectivity toward certain products. Two examples follow as illustrations.

10.2.5.1. Liquid Hydrocarbon Fuels. Two general reactions are observed to occur over an iron catalyst:

$$y\mathrm{CO}+2y\mathrm{H}_2\rightarrow(\text{—CH}_2\text{—})_y+y\mathrm{H}_2\mathrm{O}, \qquad (7)$$

$$2y\mathrm{CO}+y\mathrm{H}_2\rightarrow(\text{—CH}_2\text{—})_y+y\mathrm{CO}_2, \qquad (8)$$

with an exothermic heat of reaction of -165 kJ/mol calculated at 25 °C. The product contains the full range of linear hydrocarbons, from CH_4 all the way to waxes. Oxygenated compounds, e.g., alcohols, ketones, etc., are also produced. An example of typical operating parameters is as follows: catalyst, Fe_3O_4, promoted with potassium; temperature, 315 °C; pressure, 2.1 MPa. About 30% of the product is hydrocarbons in the gasoline range (e.g., C_6 to C_{12}), with a research octane number of about 60. The process is called the Fischer–Tropsch synthesis, after the two German researchers who first produced gasoline from coal by this technique in the 1930s. The general technology produced a significant portion of the liquid hydrocarbon fuels used in Germany during World War II, and produces more than half of the liquid hydrocarbon fuels used in South Africa today.

10.2.5.2 Alcohol (Methanol). When synthesis gas is passed over a catalyst consisting of zinc, chromia, and copper, the principal reaction is

$$\mathrm{CO}+2\mathrm{H}_2\rightarrow\mathrm{CH}_3\mathrm{OH}, \qquad (9)$$

with an exothermic heat of reaction of -158 kJ/mole calculated at 25 °C. Typical operating parameters are catalyst, $ZnO:CuO:Cr_2O_3=6:3:1$; temperature, 300 °C; pressure, 4.9 MPa.

At this writing, methanol is being promoted as a transportation fuel, principally from environmental considerations. While it does exhibit certain advantages, it also presents several disadvantages. Chief among its advantages are that it

1. is clean burning, with low CO emissions, and lower NO_x emissions than gasoline due to lower combustion temperature;
2. exhibits high octane number, permitting high compression ratio up to 16 (compared with 8 to 9 for automobiles burning unleaded gasoline), and hence higher efficiency.

Chief among the disadvantages are that

1. it has low heating value per gallon, about 0.54 that of gasoline;
2. it is toxic to humans;
3. it does not start well at temperatures below about 15 °C.

Alcohol also attacks certain components in the standard automobile fuel system, such as the carburetor and fuel pump. For extended use of alcohol fuels the automobile engine should be built for that purpose.

Methyl tertiary-butyl ether (MTBE) can be produced from methanol, to be used as an octane-improving additive to gasoline. Its use will permit higher compression ratios, hence higher efficiency, than is achievable with unleaded gasoline.

Mobil Oil Company developed a special catalyst with high selectivity for conversion of CH_3OH to gasoline, a molecular sieve known as ZSM-5. The overall reaction is

$$x\mathrm{CH_3OH} \rightarrow (\text{—}\mathrm{CH_2}\text{—})_x + x\mathrm{H_2O}. \qquad (10)$$

Reaction parameters are as follows: temperature, inlet 360 °C and outlet 415 °C; pressure, 2.1 MPa. The hydrocarbons released from the catalyst are chiefly (80%) in the gasoline range. Because of this phenomenon the catalyst is sometimes referred to as a shape-selective catalyst. From a material balance on Eq. (10) one can observe that the conversion of CH_3OH to gasoline-range hydrocarbons is about 40% on a weight basis. The research octane number of that gasoline is about 85. For process descriptions for conversion of coal to liquids, see Elliot (1981), Chap. 28.

10.3 Briquetting of Coal

The fine sizes of coal, and sometimes of coke, may be transformed into larger structures, called briquettes, by application of heat, pressure, and usually also a binder, thereby becoming suitable for use as domestic fuels, particularly in fireplaces and space heaters. Briquettes also find use in certain commercial and industrial operations. The coals most frequently used in manufacture of briquettes are anthracite (noncaking) and low-volatile bituminous (caking). The material most often used as a binder is coal tar pitch. A few other materials, such as petroleum bitumen, are being used increasingly as binders, helping to reduce smoke emissions from briquette combustion. Briquettes made for industrial plants are usually cubes or bricks weighing between 0.5 and 2.0 kg. Briquettes for domestic use usually are shaped as eggs or pillows, with a weight from 16 to 80 g.

In briquetting of noncaking coals, the coke-forming components of the pitch used as binder are of great importance. The strength of the briquette must be derived from the coke-forming components of the binder. Since coal-tar pitches contain larger quantities of coke-forming components than do bitumens, they are often preferred in briquetting anthracite. Briquetting of caking coal without binder has been extensively investigated, but the quality of the briquettes has been found to be inferior to those produced using a binder.

10.4 Methane Production from Coal Seams

As a consequence of the coalification process, and the high porosity of most coals, varying amounts of methane are trapped in the coal pores and confined by the pressure maintained in the coal seam. In coals in the United States, the estimated quantities of methane range from about 2 standard cubic meters (SCM)/metric ton in a western coal to about 20 SCM/metric ton for an eastern coal. The total estimated quantity of methane in U.S. coals (about 11 trillion SCM) is more than twice as great as known and recoverable reserves of natural gas in the United States. Various techniques of recovery of the methane from coal seams are being investigated. However, the methane exists in an adsorbed state on the internal surfaces in the coal, and rates of diffusion through the porous coal structure are slow. Consequently a very large external surface must be created by fracturing in order to realize a practical rate of recovery of methane from coal.

11. ENVIRONMENTAL, HEALTH, AND SAFETY IMPACTS FROM COAL UTILIZATION

As noted in Sec. 3.1, the occurrence of large coal reserves under thick overburden necessitates sending people and machines underground to recover the coal, thus exposing them to health and safety hazards. As described in Sec. 5, the skeletal structure of coal is composed principally of polynuclear aromatic units, many of which form carcinogenic compounds when released from the coal structure. Coals contain varying amounts of sulfur and nitrogen; these elements form compounds during combustion that are detrimental to human health and destructive to the environment if released into the atmosphere. Thus, along the full train of events, from recovery to utilization, coal may produce negative impacts unless properly handled.

11.1 Impacts from Coal Mining

Underground mining operations have always involved grave dangers to the mining personnel, and while the dangers have been greatly reduced during this century, mine disasters still occur and dangers persist. The mining of coal produces the largest number of accidental deaths and injury to workers of any sector of the coal energy system. Underground mining causes about four times as many deaths or injuries as surface mining per million tons of coal mined.

11.1.1 Explosions. The principal cause of fatalities in underground coal mining is gas explosions. Mining operations release methane, somewhat in relation to the fineness of the coal produced during the mining process. If the concentration of methane in the air is allowed to rise above the lower flammability limit of 5.0% by volume, a spark of sufficient energy, such as an electrical spark, may lead to an explosion or a fire. Air is circulated through the mine in sufficient quantities to attempt to ensure that the methane concentration will remain well below the lower explosion limit, and usually less than 1% by volume. Explosion-proof lights and motors are also required.

Reaction between coal and air occurs at surfaces. If in the mining operation large external surface areas are generated by formation of extremely fine coal dust, such as may be produced in high concentrations in longwall mining, the resulting increased reaction rates may also lead to an explosion. The dust concentration is controlled by air circulation through the mine.

11.1.2 Mine Cave-In and Subsidence. In room-and-pillar mining, pillars of coal are left in place to support the roof of the mine. Mining safety regulations restrict the width of any development or entry channel within the mine, unless additional ceiling reinforcement is provided, including floor-to-ceiling support, between the walls of the channel. In spite of these precautions, occasionally an unanticipated instability in the overburden may create pressure upon the supports beyond their capacity thereby causing the supports to fail and the roof to cave in. Such occurrences are rather rare today.

In longwall mining the roof is allowed to cave in behind the mining machine as it progresses into the coal seam, resulting in cavitation in the overburden above the cavern. Depending upon the strength and stability of the overlying rock, the cavitation may extend to the surface, resulting in a depression in the surface above the mine. This phenomenon is known as *subsidence*. After many years the mine roof may fail and settle. It is a fairly common occurrence for the cavitation thus created to extend to the surface, causing subsidence. This phenomenon has occurred in locations where towns had been built above the region that was subsequently mined, causing serious environmental damage, even property destruction. Where resources of good coals are scarce, towns have been relocated to permit underground mining of coal in regions under the prior location of the town.

11.1.3 Surface-Mining Environmental Impact. When coal is recovered by surface mining, plant life on the affected surface is destroyed. Environmental considerations, often even legislation, dictate the restoration of the overburden and revegetation. If these steps are not properly carried out, the result may be erosion of the exposed surface, destruction of wildlife habitats, and disruption of previous land use. Erosion by water may also increase sediment loading in streams. In areas of sloping terrain, terracing may be required. In the United States, surface mining legislation mandates restoration of the overburden to a slope approximating the original slope, as well as providing stipulations concerning revegetation and certain other factors. As the slope of the terrain increases, it becomes more difficult to comply with the terms of the legislation, and in some cases this precludes application of surface mining for recovery of coal.

11.1.4 Impact upon Health. Coal dust is hazardous to the health of the miners, particularly those who work in underground mining at the face of the coal seam being mined. Severe respiratory diseases often result from prolonged exposure to high dust concentration in underground mines. Examples of consequences of such exposure are tuberculosis, chronic bronchitis, and coal workers' pneumoconiosis, as well as others. As legislation has been enacted to provide benefits to victims of these ailments, the various obstructive airway diseases thus inflicted have been brought together under the name "black lung disease."

11.2 Impacts from Carbonization

Carbonization processes include those operations wherein the coal is heated in the absence of air or other reacting atmosphere. These processes are usually conducted at or near atmospheric pressure and are usually not sealed against evolution of volatiles to the environment. Bituminous coals are most commonly used in such processes, which coals evolve up to 40% or more of the weight of the coal at process conditions. Both epidemiologic studies involving human workers and laboratory studies involving animals have confirmed the carcinogenicity of many of these coal-derived compounds. The principal cancer-causing compounds thus identified are polynuclear aromatic hydrocarbons and polycyclic amines. High-temperature coal tars, such as from metallurgical coke ovens, contain more of these compounds than the volatiles usually produced in gas plants (Lloyd, 1971).

11.3 Impacts from Combustion

Coal is utilized predominantly in combustion, with the largest segment being in electric power generation, which represents 80% of all coal consumed in the United States. Inasmuch as coal combustion in power generation and other industrial applications is conducted in a fuel-lean mode, essentially no carbon monoxide or unburned hydrocarbons are emitted to the environment from this source. However, combustion in a fireplace, residential space heater, or home furnace is not subject to regulation to the fuel-lean mode, and may release polynuclear aromatic hydrocarbons or nitrogen-bearing heterocyclic compounds, as well as other pollutants. The principal products from coal combustion viewed as pollutants are unburned carbon, ash particulates, sulfur oxides, and oxides of nitrogen. Very recently some groups have begun to relate CO_2 emissions to potential global warming, and in that context view CO_2 as a pollutant.

11.3.1 Particulates. The very fine dispersion of inorganic particles within the coal matrix, as a consequence of the manner in which the coal bed was formed, coupled with disintegration of fine particles during combustion (e.g., oxidation of pyrites in the mineral matter), gives rise to liberation of extremely fine particles of inorganic matter, called *fly ash*. In addition, some combustion processes do not accomplish complete conversion of the carbon, resulting in tiny carbon particles being emitted. These particles from both sources are small enough to be emitted with the combustion gases. The particles themselves may be harmful to the lungs when inhaled, especially to people suffering from respiratory ailments. In addition, the mineral matter often contains metallic elements in trace amounts, such as arsenic, cobalt, lead, mercury, and several others, which are toxic to the body, and possibly carcinogenic. The particulates also cause defacing of buildings and other structures.

These particulates, both varieties, may be effectively removed in an electrostatic precipitator, the manufacturers stating an effectiveness of at least 99.9% solid removal. This technique is most widely used for solid removal from combustion effluent gases in electric power generation. Other devices finding use include the bag house and cyclone separator.

11.3.2 Oxides of Sulfur. When coal is burned in air, the sulfur in the coal, both organic and pyritic, is converted to sulfur oxides, in a ratio of about 5 parts SO_2 to 1 part SO_3, mole basis. The sulfur content in coals presently used industrially varies from as low as 0.3% by weight to 5.0% by weight, or even higher. While combustion of the very-low-sulfur coal may meet acceptable emission standards without sulfur removal, serious environmental and health effects result from the combustion effluent from high-sulfur coals, unless the sulfur oxides are removed. Coal combustion is the principal source of fuel-generated sulfur oxide emissions to the atmosphere.

Sulfur oxides are very soluble in water, SO_2 forming H_2SO_3, a weak acid, and SO_3, forming H_2SO_4, a strong acid. Many of the harmful effects of sulfur oxides are due to the action of these acids. When air containing sulfur oxides is inhaled, moisture in the respiratory system may lead to acid formation and oxidation reactions in the system. As a consequence, breathing air that contains sulfur oxides may aggravate existing respiratory illness, and prolonged exposure to sulfur oxides may cause respiratory illness.

As rain droplets pass through an atmosphere than contains sulfur oxides, these oxides dissolve in the liquid droplets, forming the acids identified above and producing "acid rain." Sulfur oxides are the largest contributors to the formation of acid rain. Plants in general are sensitive to pollutants in the air, and many plants are adversely affected by acid rain. While a certain amount of damage to crops and forests has been traced to acid rain, the effects upon aquatic life in lakes and rivers is still, at this writing, the subject of strong controversy.

Sulfur oxides lead to large amounts of damage to buildings and other structures, both as pollutants added to rain and by direct attack. The cost of this damage, resulting from defacing of the structure as well as structural damage, is great.

Sulfur oxides may be removed from combustion product gases by chemical reaction with a number of compounds, a popular example being limestone. The "scrubbing" compound may be added to the coal bed in the case of fluidized-bed combustion, one of the advantages of that type of combustion system. Alternatively, the combustion-product gases may be passed through an appropriate scrubbing system, called a *stack gas scrubber*, before being emitted to the atmosphere. At this writing, all new electric-power-generating plants constructed in the United States to burn coal are required to be equipped with stack gas scrubbers for removal of sulfur oxides, regardless of the sulfur content of the coal being fed.

11.3.3 Oxides of Nitrogen. Nitrogen can react with oxygen to form several oxides of nitrogen, designated NO_x. From the point of view of air pollutants resulting from coal combustion, the principal oxides of interest are NO and NO_2. The extent of reaction between nitrogen and oxygen to form NO is a function of temperature, the degree of reaction increasing as temperature increases. Some coal-combustion systems attempt to operate at temperatures low enough to limit NO_x formation. At higher temperatures, molecular nitrogen from the air reacts with atmospheric oxygen to form NO.

Coals in general contain bound nitrogen ranging from about 1.0% to 1.5% by weight of the coal. Research in recent years has demonstrated that the bound nitrogen in the coal reacts more readily with atmospheric oxygen during the combustion process to form NO_x than does atmospheric nitrogen, and it does so at a lower temperature. The stack gas srubbers currently used to remove sulfur oxides from coal combustion product gases do not remove NO_x.

NO_2 is one of the essential ingredients in the formation of photochemical smog, which is the smog of greatest concern and most frequent occurrence. The other two essential ingredients are hydrocarbons and sunlight. The chemistry of smog formation is discussed in the article AIR POLLUTION: COMPONENTS CAUSES, AND CURES.

Photochemical smog causes serious loss of productivity of crop plants. It also attacks rubber, paints, fabrics, and other materials. Coal combustion is a principal contributor of NO_x to the formation of photochemical smog. In cities with smog problems, the automobile is the principal contributor of hydrocarbons to photochemical smog formation.

11.3.4 Carbon Dioxide. Coal combustion is a principal source of addition of CO_2 to the atmosphere. In a direct sense CO_2 would not be considered to be a pollutant, but some groups are expressing concern about a possible global warming effect and a perceived contribution by CO_2 to such a warming trend. Some groups have advocated a dramatic reduction in the use of coal combustion, as well as other fossil fuels, to address this perceived atmospheric temperature phenomenon. At this writing the question of prediction of a global warming is controversial. Different scientific and engineering groups disagree as to whether a global warming trend is occurring, or whether what is now being observed is a part of the cyclic pattern established over the past 100 years. There is further disagreement among these groups as to what contributions, as a percentage of all contributing factors, may be attributable to CO_2 production. Finally, there is strong disagreement among scientists as to the adequacy of the models used in predictions of global warming. Further information concerning the basis for the global warming concern is presented in the article AIR POLLUTION: COMPONENTS, CAUSES, AND CURES.

11.3.5 Clean Coal Technologies. Direct combustion of coal presents several hazards to the environment. It is possible by means of precombustion processing of the coal to reduce some of those environmental hazards substantially, such that pollutant removal from the combustion effluent may not be required. Of particular interest is reduction of sulfur and reduction of the mineral matter, which contains many potential pollutants in addition to a portion of the sulfur. Included under the umbrella of clean coal technologies are several concepts for delivering a cleaner-burning fuel. At this writing, these concepts are either among topics already discussed, or are only research concepts which may or may not prove to be successful.

11.4 Impacts from Coal Conversion

Coal-conversion technologies of greatest interest and importance are conducted in a closed system under pressure. Consequently gaseous or liquid products of environmental concern are, or should be, contained during the conversion train, and can be collected or processed before release to the environment. In the reducing atmosphere within the reactor system, sulfur is converted to H_2S and nitrogen is converted to NH_3. Both of these products are readily removed from the exit gases by well-developed and reliable technologies.

Workers need to avoid contact with coal-derived liquids and vapors in coal-conversion plants. However, because of the nature of the coal-conversion systems, adverse health effects from the closed systems are expected to be less extensive than those from coke ovens.

12. ECONOMIC CONSIDERATIONS

The economic position occupied by coal in relation to other energy resources is subject to many factors. These factors in turn vary widely among the different countries, affected by political considerations, resource availability, existing and projected energy consumption patterns, and others. Only a few general observations can be made.

1. Combustion of oil or natural gas exhibits two major advantages that establish its preference over coal, namely cleanliness in combustion and convenience in delivery to the combustion site (even to the burner), since they are fluids, while coal is a solid. In a free-market environment, coal must demand a lower price per unit heating value than oil or natural gas to offset these factors, or it cannot compete in the marketplace. These factors may be overruled by government subsidy or other non-free-market factors.
2. In the development of synthetic fuels from coal, the projected, required selling price has always been higher than the world market price of petroleum or natural gas delivered to the market location. Synthetic fuels plants are all very expensive to construct, and they are constructed with existing energy costs factored into the cost of materials for constructing the plant. When the high cost of plant construction is amortized, and factored into the selling price of the product, that price at the time of entry of the product into the market will always be higher than the market price of the product with which it must compete. At this writing that market price differential is about 50%–100%.

GLOSSARY

BET (Brunauer–Emmet–Teller): A technique for calculating surface areas from gas adsorption isotherms.

dmmf: Dry, mineral-matter-free.

Metamorphism: Alteration of plant substances resulting from the effects of elevated pressure and temperature.

Mobile Phase: Organic material in coal not chemically bound to the solid matrix.

Overburden: The layer of rock and soil on top of a coal seam.

Petrographic Analysis: Identification of the types and quantities of organic material in coal, as determined by observation under a microscope.

Petrography: A description or classification of coal (from the word roots petros = rock, graphy = description of).

Petrology: The scientific study of coal (from the word roots petros = rock, logos = the science of).

SCM: Standard cubic meter, measured at 0 °C and 1 atm pressure.

SCF: Standard cubic foot, measured at 0 °C and 1 atm pressure.

List of Works Cited

Brunauer, S. (1945), *The adsorption of Gases and Vapors*. Princeton, New Jersey: Princeton University Press.

Elliott, M. A. (Ed.) (1981), *Chemistry of Coal Utilization*, 2nd Suppl. Vol. New York: Wiley.

Lowry, H. H. (Ed.) (1963), *Chemistry of Coal Utilization*, Suppl. Vol. New York: Wiley.

Lloyd, J. W. (1971), *J. Occupat. Med.* **13**, 53–68.

Wiser, W. H. (1978), in: B. R. Cooper (Ed.), *Scientific Problems of Coal Utilization*, DOE Symposium Series No. 46, Oak Ridge, TN: U.S. DOE Technical Information Center.

Further Reading

Elliott, M. A. (Ed.) (1981), *Chemistry of Coal Utilization*, 2nd Suppl. Vol., New York: Wiley.

van Krevelen, D. W. (1961), *Coal: Typology, Chemistry, Physics and Constitution*, Amsterdam: Elsevier.

COHESIVE ENERGY

JAMES R. CHELIKOWSKY, *Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, Minnesota, U.S.A.*

	Introduction: Cohesion in Solids	559
1.	**Binding in Solids: Classification by Type of Chemical Bond**	560
1.1	Closed-Shell Binding: van der Waals Solids	560
1.1.1	Dipole–Dipole Interactions	561
1.1.2	Lennard–Jones Potentials	562
1.2	Hydrogen Bonding	563
1.3	Ionic Bonding	563
1.3.1	Electrostatic Interactions in Ionic Solids	563
1.3.2	Madelung Constant	564
1.4	Covalent Solids	565
1.4.1	Classical Chemical Bond in Solids	565
1.4.2	Example of the H_2 Molecule	566
1.4.3	Solid-State Treatments of the Covalent Bond	567
1.5	Metallic Bonds	568
1.5.1	Delocalized Electrons	569
1.5.2	Wigner–Seitz Method	569
1.6	Mixed Bonding in Solids	569
1.7	Noncrystalline Solids	570
2.	**Total-Energy Calculations of the Cohesive Energy**	572
3.	**High-Pressure Trends in Cohesion**	572
4.	**Chemical Trends in Cohesion and Structure**	574
	Glossary	575
	Works Cited	576
	Further Reading	576

INTRODUCTION: COHESION IN SOLIDS

Cohesion in solids arises when the free energy of the atoms in the solid is lower than that of the vapor or liquid. The cohesive energy of a solid can be defined as the difference in energy per atom in the solid at 0 K at 1 atm versus the solid separated into isolated, neutral atoms in their lowest-energy state. Since the solid is taken to be at absolute zero, entropic considerations do not enter and the free energy corresponds to the enthalpy of the solid. A related term is the energy to separate an ionic crystal into the constituent ions. This energy difference is known as the "lattice energy."

From an experimental point of view, the cohesive energy of a solid can be determined by precise calorimetry measurements. For example, if the heat of sublimation is measured and the heat capacities of the gas and solid are known as functions of temperature, it is possible to determine the enthalpy difference between the gas and the solid at a given temperature. This difference at absolute zero would correspond to the cohesive energy. In some cases, it is possible to estimate the cohesive energy from known experimental data and some simple approximations. For example, the Born–Haber cycle can be used to determine the heat of formation of an ionic solid. If we are interested in reacting metallic sodium and chlorine gas to form solid sodium chloride, then we could *conceptualize* this process taking place in several steps. First, we form a neutral monatomic gas of Na and Cl. We could vaporize the sodium metal and measure the heat of sublimation. For the chlorine gas, we could measure the energy to dissociate the Cl_2 molecule to monatomic chlorine. Second, we consider creating a gas of ionic species Na^+ and Cl^- by ionizing the Na and adding an electron to Cl to form a negative ion. To determine the energy of this step we would measure the ionization energy of Na and the electron affinity of Cl. Third, we consider bringing together the Na^+ and Cl^- species to form a crystal composed of the ionic species. This energy may be determined by summing up all the classical electrostatic

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

interactions in the crystal as discussed in the section on ionic bonding (Sec. 1.3). By considering the difference in energy from forming the crystal of ions and the "preparation energy" from sublimating the Na metal, dissociating the Cl_2 molecule, and forming the (Na^+,Cl^-) ions, we can obtain an accurate measure of the heat of formation (Swalin, 1972).

In general, calorimetry measurements are difficult, as the difference between competing solid states can be quite small. For example, the energy difference between carbon in the diamond and graphite structures can be a few kcal/mol. Nonetheless, for most elemental solids and many binary alloys, the cohesive energy has been measured. In Table 1, the cohesive energy is given for most of the elemental solids.

From a theoretical point of view, binding in solids is fairly well understood. The forces in solids are purely electronic. The interactions can be categorized as electron-nuclear, electron-electron, and nuclear-nuclear. Unlike the situation in nuclear physics where the forces are not well defined between nucleons, the forces in solids are. Specifically, the Hamiltonian for a solid can be written down exactly. Quantum chemists have demonstrated for molecules that the calculated values for binding energies agree well with experiment. One has no reason to believe the situation for solids would be different.

Although the interactions in solids are well understood, it is not an easy task to evaluate the total energy of the solid, even at absolute zero. In principle, two factors complicate the problem. The energy of an isolated atom can be very large. For example, the energy of an isolated, heavy atom such as uranium can be estimated to be on the order of $\sim 10^6$ eV or more, but the cohesive energy of uranium is only about 6 eV/atom. (This is an extreme example to be sure, but it illustrates just how difficult the problem can be.) To obtain a reasonable estimate of the cohesive energy in this case, one must have a method that is accurate to one part in 10^6, or better. The other fact that greatly complicates the evaluation of cohesive energies by theoretical methods is the number of particles involved. Given that a macroscopic solid may contain 10^{23} nuclei and electrons, it is impossible to determine the total energy of the crystal without some approximations.

Several advances have made it possible to predict the cohesive energy of solids by numerical solutions of the quantum-mechanical equations of motion, i.e., the Schrödinger equation. One advance has been of a technological nature: the invention of high-speed computers. Another major advance was to devise one-electron potentials that greatly simplify many-body interactions. Also, several electronic structure methods have been developed that separate the problem of describing strongly bound core electrons from that of weakly bound valence electrons. For further discussion of these developments, see ELECTRONIC STRUCTURE OF SOLIDS. The accuracy of these computations is usually not at the same level as experiment. Nonetheless, it is now possible, for elemental solids and simple compounds, to predict whether a given crystal structure is the most stable one at 0 K and 1 atm.

1. BINDING IN SOLIDS: CLASSIFICATION BY TYPE OF CHEMICAL BOND

It is useful to classify solids by the type of chemical bond that stabilizes the solid (Kittel, 1986). For example, electrostatic forces between ions dominate the properties of solids such as sodium chloride. By attributing the binding in NaCl to electrostatic interaction between Na^+ and Cl^- ions, it is possible to account for many properties such as the cohesive energy, compressibility, elastic properties, vibrational spectrum, and crystal structure. Broadly speaking, there are five types of bonding in solids: van der Waals, hydrogen, ionic, covalent, and metallic.

1.1 Closed-Shell Binding: van der Waals Solids

Inert gases have filled *s* and *p* shells (see ELECTRONIC STRUCTURE OF ATOMS AND MOLECULES). These elements have no orbitals available for forming bonds. Thus, it is not surprising that these elements form weakly bound solids. Crystals made of inert gases have low melting points, sublimation energies, and bulk moduli. They are also electrically inert and transparent. The closed-shell atomic structure of these elements results in crystal structure in which directional forces are not significant.

Table 1. Cohesive energies, melting points, and bulk moduli for elemental solids as tabulated by Kittel (1986), Brewer (1977), and Gschneidner (1964).

Element	Cohesive energy (eV/atom)	Melting point (K)	Bulk modulus (10^{11} N/m^2)	Element	Cohesive energy (eV/atom)	Melting point (K)	Bulk modulus (10^{11} N/m^2)
Actinium	4.25	1324	0.25	Neon	0.002	24.56	0.010
Aluminum	3.39	933.5	0.722	Neptunium	4.37	910	0.68
Americium	2.73	1449		Nickel	4.44	1728	1.86
Antimony	2.75	903.9	0.383	Niobium	7.57	2750	1.702
Argon	0.08	83.81	0.016	Nitrogen	4.92	63.15	0.012
Arsenic	2.96	1089	0.384	Osmium	8.17	3306	4.18
Barium	1.9	1002	0.103	Oxygen	2.6	54.36	
Beryllium	3.32	1562	1.003	Palladium	3.89	1827	1.808
Bismuth	2.18	544.6	0.315	Phosphorus	3.43	317	0.304
Boron	5.81	2365	1.78	Platinum	5.84	2045	2.783
Bromine	1.22	265.9		Plutonium	3.6	913	0.54
Cadmium	1.16	594.3	0.467	Polonium	1.5	527	0.26
Calcium	1.84	1113	0.152	Potassium	0.934	336	0.032
Carbon	7.37		4.43	Praseodymium	3.7	1205	0.306
Cerium	4.32	1072	0.239	Radium	1.66	973	0.132
Cesium	0.804	301.6	0.020	Radon	0.202		
Chlorine	1.4	172.2		Rhenium	8.03	3459	3.72
Chromium	4.1	2133	1.901	Rhodium	5.75	2236	2.704
Cobalt	4.39	1770	1.914	Rubidium	0.852	312.6	0.031
Copper	3.49	1358	1.37	Ruthenium	6.74	2527	3.208
Curium	3.99	1613		Samarium	2.14	1346	0.294
Dysprosium	3.04	1684	0.3.4	Scandium	3.9	1814	0.435
Erbium	3.29	1797	0.411	Selenium	2.46	494	0.091
Europium	1.86	1091	0.147	Silicon	4.63	1687	0.988
Fluorine	0.84	53.48		Silver	2.95	1235	1.007
Gadolinium	4.14	1587	0.383	Sodium	1.113	371	0.068
Gallium	2.81	302.9	0.569	Strontium	1.72	1042	0.116
Germanium	3.85	1211	0.772	Sulfur	2.85	388.4	0.178
Gold	3.81	1338	1.732	Tantalum	8.1	3293	2.00
Hafnium	6.44	2504	1.09	Technetium	6.85	2477	2.97
Holmium	3.14	1745	0.397	Tellurium	2.19	722.7	0.23
Indium	2.52	429.8	0.411	Terbium	4.05	1632	0.399
Iodine	1.11	386.7		Thalium	1.88	577	0.359
Iridium	6.94	2720	3.55	Thorium	6.2	2031	0.543
Iron	4.28	1811	1.683	Thulium	2.42	1820	0.397
Krypton	0.116	115.8	0.018	Tin	3.14	505.1	1.11
Lanthanum	4.47	1194	0.243	Titanium	4.85	1946	1.051
Lead	2.03	600.7	0.43	Tungsten	8.9	3695	3.232
Lithium	1.63	453.7	0.116	Uranium	5.55	1406	0.987
Lutetium	4.43	1938	0.411	Vanadium	5.31	2202	1.619
Magnesium	1.51	922	0.354	Xenon	0.16	161.4	
Manganese	2.92	1520	0.596	Ytterbium	1.6	1098	0.133
Mercury	0.67	234.3	0.382	Yttrium	4.37	1801	0.366
Molybdenum	6.82	2895	2.725	Zinc	1.35	692.7	0.598
Neodymium	3.4	1290	0.327	Zirconium	6.25	2128	0.833

Since pairwise forces dominate the bonding in inert-gas crystals, the structures are close-packed.

1.1.1 Dipole–Dipole Interactions. If we consider interactions present between static, electrically neutral, spherical charge distributions, then no cohesion is possible. However, if we allow dynamical fluctuations of dipoles that arise from the motion of the electrons in the inert-gas atom, then an attractive interaction is possible. Specifically, a dipole present in one atom can induce a dipole moment in a neighboring atom. Suppose at some time an atom A has an instantaneous dipole moment $\mathbf{p}_A$. An atom B will experience an electric field that is proportional to p_A/R^3 at a distance R from A. This field will induce a dipole on B given by $p_B=\alpha E_A \sim \alpha p_A/R^3$, where α is the polarizability of the atom. The energy of this interaction is given by $U_{AB} \sim -p_B p_A \sim -\alpha p_A^2/R^6$. The time average of $\mathbf{p}_A$ vanishes. However, since the dipole enters as the square, the interaction energy does not vanish. More precise treatments of this interaction require the consideration of three or more interacting atoms (Horton, 1968).

1.1.2 Lennard–Jones Potentials. The attractive dipole-dipole interaction is opposed by the Pauli repulsive forces. A fundamental principle of quantum mechanics is that two electrons cannot occupy the same state, i.e., they may not have the same quantum numbers. If atoms with closed shells overlap some electrons must be promoted to unoccupied, higher-energy states. This promotion results in a higher-energy state and an effective repulsion between the atoms. Empirically, this repulsive interaction is found to be very short ranged. One of the more common approximations is to assume a pairwise interaction that scales as $\sim 1/R^{12}$. Another form often used is exponential scaling, i.e., $\sim \exp(-R/\rho)$, where ρ is a parameter that indicates the range of the interaction.

Combining the repulsive and attractive terms results in the Lennard–Jones potential. The potential between two atoms can be written as

$$U(R)=4\epsilon\left[\left(\frac{\sigma}{R}\right)^{12}-\left(\frac{\sigma}{R}\right)^{6}\right], \tag{1}$$

where U is the interaction energy between two atoms, R is the interatomic distance and (ϵ,σ) are parameters. Values for ϵ and σ can be extracted from gas-phase measurements and transferred for solid-state calculations.

At very low temperatures, and for heavy atoms, we may neglect the kinetic energy and use Eq. (1) to determine the cohesive energy of an inert-gas solid. The total energy per atom, E_{total}, is given by

$$E_{\text{total}}=2\epsilon\left[\sum_{m,n\neq m}\left(\frac{\sigma}{R_{nm}}\right)^{12}-\sum_{m,n\neq m}\left(\frac{\sigma}{R_{nm}}\right)^{6}\right], \tag{2}$$

where we have taken the nth atom as the reference atom, R_{nm} is the distance of atom m from the reference atom, and the m sum is over all the atoms in the system, except the reference atom. If we define a dimensionless parameter ρ_{nm} such that $R_{nm}=\rho_{nm}R$ where R is the nearest-neighbor distance, then we may define E_{total} as

$$E_{\text{total}}=2\epsilon\left[\alpha\left(\frac{\sigma}{R}\right)^{12}-\beta\left(\frac{\sigma}{R}\right)^{6}\right], \tag{3}$$

where (α,β) correspond to structurally dependent parameters, e.g., $\alpha=\Sigma_{n,m}(1/\rho_{nm})^{12}$. For the face-centered-cubic (fcc) structure, $\alpha=12.131\,88$ and $\beta=14.453\,92$; the corresponding values for the hexagonal-close-packed (hcp) structure are 12.132 29 and 14.454 89. One can see how small the differences are between hcp and fcc structures. Both structures have 12 nearest neighbors and differ only in the placement of the second neighbors. The short-range nature of the van der Waals forces accounts for $(\alpha,\beta)\approx$ number of nearest neighbors. Thus a body-centered structure with eight nearest neighbors would have values of $(\alpha,\beta)\approx 8$.

By taking the derivative of E_{total} with respect to R, one can find the equilibrium value of the bond length. For the fcc structure, $R_0 \approx 1.09\sigma$. Using this value, the cohesive energy can be found. For fcc crystals, we find the cohesive energy $\approx 8.6\epsilon$. For example, for a dilute argon gas, $\epsilon=0.0104$ eV/atom. Using Eq. (3), we would predict a cohesive energy of $\cong 0.09$ eV/atom; the experimental value is about 0.08 eV/atom. In Table 2, we summarize the Lennard–Jones parameters for the inert gases. Note how the binding energy increases with electron number; this is a consequence of the higher polarizability of the heavier atoms. Heavier inert-gas atoms have outer electrons that are more loosely bound than those in lighter inert gases. Hence, it is easier to induce larger dipoles in the heavier atoms. As a consequence, the binding energy for Xe is nearly an order of magnitude larger than Ne. This feature is unusual. For example, group-IV solids show the reverse

Table 2. Inert-gas parameters for the Lennard–Jones potential. From Kittel (1986) and Horton (1968).

	ϵ (10^{-16} erg)	σ (Å)	Nearest-neighbor distance (Å)	Cohesive energy (eV/atom)	Melting point (K)
Helium	14	2.56	···	···	···
Neon	50	2.74	3.13	0.02	24
Argon	167	3.40	3.76	0.08	84
Krypton	225	3.65	4.01	0.12	117
Xenon	320	3.98	4.35	0.17	161

trend: Diamond exhibits a much stronger binding energy than does lead. This reversal is a consequence of weaker covalent bonds being formed by heavier atoms, as discussed in the section on covalent bonding.

The differences between the experimental and theoretical values *increase* as the mass of the inert gas atom *decreases*. This effect is "quantum mechanical" in origin. The zero-point motion of the atoms constitutes a significant correction for light atoms as the term scales inversely with the atomic mass. In the case of helium, a solid cannot be formed at 1 atm even at absolute zero as the zero-point motion exceeds the attractive Lennard–Jones potential.

Although our discussion has concentrated on inert gases, solids such as benzene, nitrogen, methane, and chlorine are bound via van der Waals interactions. The bonding orbitals of the constituent species in these solids, e.g., N_2, CH_4, etc., are completely filled. Thus, just as in the inert-gas atoms, only weak dipole interactions are present.

Dipole-dipole interactions also occur in polar molecules that have permanent dipole moments. Such molecules can form weakly bound solids similar to molecular solids.

1.2 Hydrogen Bonding

Another example of weak bonding in solids arises from *hydrogen bonding* (Pauling, 1960). Hydrogen bonding was first recognized in 1912. It is present in organic solids and in solids such as ice, boric acid, and hydrogen fluoride, i.e., the solid contains hydrogen and an electronegative element such as oxygen, fluorine, or nitrogen. In such solids, the hydrogen atom serves to bond two negative ions. This is a surprising result as one might expect hydrogen to form a single covalent bond to only one atom. However, given that hydrogen bonds form only with highly electronegative atoms, it is believed that the hydrogen is ionized and that the forces involved are electrostatic in nature. In the most extreme case, such as $[HF_2]^-$, the bare proton is thought to form a "bond" between two F^- ions.

The binding energy of a hydrogen bond is typically on the order of 0.1–0.5 eV/atom. Given this weak bond, the hydrogen bond plays an important role in organic reactions at room temperature. It is also responsible, along with the permanent dipole moments, for the striking physical properties of water and ice. It is for these reasons that Pauling has stated that the hydrogen bond may be the most important bond for understanding the biochemistry of living organisms.

1.3 Ionic Bonding

Solids made of elements from the extremes of the periodic table, e.g., from groups I and VII, form ionic solids. In these solids, electrons are transferred from the cation to the anion. In the language of Pauling, the electropositive atom such as Na yields a valence electron to form a closed-shell positive ion. The electronegative atom such as chlorine gains an electron to form a closed-shell negative ion. The binding energy in these crystals is large, and the structures formed tend to be close-packed.

1.3.1 Electrostatic Interactions in Ionic Solids. Although the bonding in ionic solids is very different from the bonding found in molecular solids, it can be described accurately by *pairwise* forces. The attractive origin of bonding in ionic solids is electrostatic in nature between the positively charged cations and the negatively charged anions. In sodium chloride, Na^+ ions attract Cl^- ions. One can envision forming a sodium chloride crystal as follows. Energy is required to ionize a Na atom to form Na^+ and a free electron. The electron is added to Cl to form Cl^-. Energy is liberated in this latter process; however, the net process is endothermic. The ionization energy of Na is 5.14 eV; the electron affinity of Cl is 3.61 eV. The energy to create an ion pair is 1.53 eV. When these ions are brought together we gain a "lattice energy" contribution from the electrostatic interactions. This term is about 7.9 eV/molecular unit; the cohesive energy is $(7.9-1.5)=6.4$ eV/molecular unit.

1.3.2 Madelung Constant. The lattice energy in an ionic solid can be evaluated by summing over the electrostatic interactions. If we consider a rocksalt structure (a simple-cubic lattice with alternating cation and anion sites), then the summation would proceed as

$$E_S=-(e^2/R)[6-12/\sqrt{2}+8/\sqrt{3}-\cdots], \qquad (4)$$

where E_S is the electrostatic interaction, R is the interatomic distance between the cation and anion, and we have taken the charge on the ions to be unity. The summation in Eq. (4) arises from 6 nearest neighbors at a distance R, 12 second-nearest neighbors at $\sqrt{2}R$, 8 third-nearest neighbors at $\sqrt{3}R$, and so on. The sign on each shell of neighbors differs. The summation in Eq. (4) is conditionally convergent. This poor convergence is a consequence of the long-range nature of the Coulomb forces. The sum in brackets in Eq. (4) is structure dependent and is called the *Madelung constant*, A_M. Consider a simple ionic solid with a basis of two ions with opposite charges. An example of such a solid would include the rocksalt, cesium chloride, and zinc-blende structures. The Madelung constant can be written as

$$A_M=R\sum_n\left(\frac{1}{R_n}-\frac{1}{|\mathbf{R}_n-\boldsymbol{\tau}|}\right), \qquad (5)$$

where R is the nearest-neighbor distance, $\mathbf{R}_n$ is a lattice vector, and $\boldsymbol{\tau}$ is the basis vector that gives the atomic position within the crystalline unit cell. The first term in the sum excludes the divergent term, i.e., the $R_n=0$ term. Equation (5) can be evaluated by a technique developed by Ewald. In Table 3, values for the Madelung constant are given for some common ionic crystals. For simple binary solids, the value of A_M resides between 1.5 and 2.0. For fluorite structures the value is near 5 and for alumina the value is 25.

Table 3. Madelung constant for some binary structures.

Structure	Madelung constant
Rocksalt	1.7476
Cesium chloride	1.7627
Zinc blende	1.6381

Just as for the inert-gas case, at very short distances there is a repulsive term that prevents the crystal from collapsing. A simple form of this interaction is $E_R=B\exp(-R/\rho)$. With this term the total lattice energy (Born and Huang, 1954) is given by

$$E_{\text{lattice}}=B\exp(-R/\rho)-A_Mq^2/R, \qquad (6)$$

where B is a constant that depends on the structure and we assume that the charge on the ions is q. By taking the derivative of Eq. (6) with respect to R and setting it equal to zero to find the nearest distance, one can eliminate B from the equation. The equilibrium lattice energy per molecular unit is given by

$$E^0_{\text{lattice}}=\frac{-A_Mq^2}{R_0}\left(1-\frac{\rho}{R_0}\right), \qquad (7)$$

where R_0 is the equilibrium nearest-neighbor distance. The bulk modulus of the solid can be used to estimate ρ; typically $\rho/R\approx0.1$. Thus, the repulsive term in Eq. (7) is a 10% correction to the lattice energy. Another common form used for the repulsive term is $E_R\approx B/R^n$ where $n\approx6$–9. In Table 4, values for the ionic parameters in Eq. (6) are given along with the experimental lattice energy for the alkali halides.

From Table 3, one might expect the most ionic binary structure to be of the cesium chloride structure as this structure has the highest Madelung constant. However, the structure of a highly ionic crystal such as CsF is not cesium chloride, but rocksalt. Subtle differences in the bonding orbitals may play a more important role in determining which structure is favored (Tosi, 1964).

While this picture of ionic bonding is simple and able to describe the physical properties of alkali halides, it is limited in its applicability. It may be difficult to determine the charge state on the cation and anion. For the alkali halides, it does not seem unreasonable to assume that the charge transfer is complete.

Table 4. Properties of alkali halide crystals [see Eq. (6)]. From Tosi (1964).

Alkali halide	B (10^{-8} erg)	ρ (Å)	R_0 (Å)	Lattice energy (kcal/mol) Theory	Lattice energy (kcal/mol) Experiment
LiF	0.30	0.291	2.01	242.2	242.3
LiCl	0.49	0.330	2.57	192.9	198.9
LiBr	0.59	0.340	2.75	181.0	189.8
LiI	0.60	0.366	3.00	166.1	177.7
NaF	0.64	0.290	2.32	215.2	214.4
NaCl	1.05	0.321	2.82	178.6	182.6
NaBr	1.33	0.328	2.99	169.2	173.6
KF	1.31	0.298	2.67	189.1	189.8
KCl	2.05	0.326	3.15	161.1	165.8
KBr	2.30	0.366	3.30	154.5	158.5
KI	2.85	0.348	3.53	144.5	149.9
RbF	1.78	0.301	2.82	180.4	181.4
RbCl	3.19	0.323	3.29	155.4	159.3
RbBr	3.03	0.338	3.45	148.3	152.6
RbI	3.99	0.348	3.67	139.6	144.9

However, for other seemingly ionic materials, the issue is problematic. For example, in oxides the charge state of oxygen appears to be O^{2-}; however, an isolated oxygen atom cannot bind a second electron. It is unclear what to use for the preparation energy of the oxygen anion in the solid. Quantum-mechanical calculations have shown that the charge transfer is rarely complete for solids other than the alkali halides.

1.4. Covalent Solids

Crystals such as carbon, silicon, and germanium (group IV of the periodic table) form in the diamond structure. They exhibit strong binding. For example, carbon in the diamond structure has a cohesive energy of about 7.3 eV/atom (graphite has a slightly higher cohesive energy). The origin of the bond in these crystals cannot be as simple as that found in the inert gases or ionic solids. The binding in these latter solids results in close-packed structures, which is consistent with *pairwise* interactions. Unlike the close-packed structures, the diamond structure is open and each atom is bonded to four nearest neighbors. The forces here are not pairwise and are angularly dependent. The existence of a stable open structure such as diamond is a strong signature for the existence of quantum-mechanical forces. Classical forces that arise from purely Coulombic interactions would be pairwise in nature and would lead to close-packed structures.

1.4.1. Classical Chemical Bond in Solids. The origin of the traditional covalent bond is from the sharing of electrons between two atoms. Organic chemistry is usually considered the domain for the study of the covalent bond. This is true in some sense; inorganic covalent solids are rare. Consider that of the 100-plus elements that form solids, fewer than a dozen form covalent solids. Nonetheless, the covalent bond is crucial to understanding the behavior of semiconductors and superconductors.

In Fig. 1, the valence charge density is presented for both carbon and silicon in the diamond structure. Carbon is unusual in that the covalent bond between the atoms consists of a "double hump" as opposed to the single maximum in the silicon structure. This difference arises from differences in the core electronic structure of carbon as opposed to silicon, i.e., carbon has no *p* states in the core; silicon does.

If the number of nearest neighbors in a crystal equals the number of valence electrons, then localized covalent bonds may be formed (Pauling, 1960). In the diamond structure the number of nearest neighbors is four. If we were to consider a hypothetical structure for silicon such as simple cubic which has six nearest neighbors, we would expect silicon to be a metal in this structure. The number of valence electrons present would not saturate the bond. Partially filled bonding orbitals allow electrons to be easily propagated through

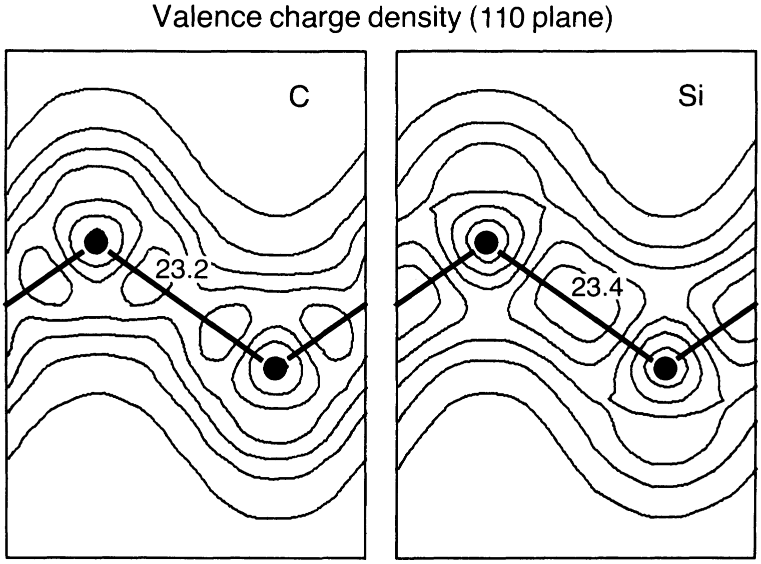

Valence charge density (110 plane)

FIG. 1. Contour plot of the valence charge density of carbon and silicon from Yin and Cohen (1982). The core states have been omitted from the plot. The density is normalized to the number of electrons in the unit cell.

the solid, creating a metallic crystal. Quantum calculations have verified this picture.

1.4.2 Example of the H_2 Molecule. The energetics of the covalent bond can be addressed best from a quantum-mechanical perspective. One of the first cases to be examined was the hydrogen molecule. This is the simplest system that illustrates the covalent bond and includes electron-electron interactions.

The molecule contains four interacting particles: two protons and two electrons. To determine the binding energy of this system, we solve the Schrödinger equation for the electronic wave functions and energy levels. The Hamiltonian for the electronic interactions is easy to write down:

$$\mathscr{H}(\mathbf{r}_1,\mathbf{r}_2)=-\frac{\hbar^2\nabla_1^2}{2m}-\frac{\hbar^2\nabla_2^2}{2m}-\frac{e^2}{r_1}-\frac{e^2}{r_2}-\frac{e^2}{|\mathbf{R}-\mathbf{r}_1|}-\frac{e^2}{|\mathbf{R}-\mathbf{r}_2|}+\frac{e^2}{|\mathbf{r}_1-\mathbf{r}_2|}. \qquad (8)$$

The first two terms correspond to the kinetic energy of the electrons, and the last five terms correspond to the electron-proton and electron-electron interactions. The protons are assumed to be at rest and a distance R apart.

The separation of the nuclear coordinates from the electronic coordinates is a consequence of the Born–Oppenheimer approximation. In this approximation, one assumes the electronic motions are so rapid, as compared to the nuclear motions, that the electrons respond instantaneously to a fixed set of nuclear coordinates. This approximation works better for heavier nuclei, as one would expect.

A quantum mechanical treatment of the hydrogen molecule was performed by Heitler and London (1927). The first step in their treatment is to write an approximate wave function for the molecule. Their wave function is based on assuming the molecular wave function could be written as follows:

$$\Psi(\mathbf{r}_1,\mathbf{r}_2)=a_1\phi_a(\mathbf{r}_1)\phi_b(\mathbf{r}_2)+a_2\phi_b(\mathbf{r}_1)\phi_a(\mathbf{r}_2), \qquad (9)$$

where ϕ_a (ϕ_b) is a hydrogenic wave function centered on the a (b) proton. The first term in Eq. (9) can be interpreted as electron 1 localized on proton a and electron 2 localized on proton b. The reverse situation is also possible so we include this term in our wave function.

The binding energy of the molecule can be found by solving a simple 2×2 determinant for the parameters a_1 and a_2 in Eq. (9). The results for the electronic energies and wave functions are given by

$$E_{\pm}=\frac{H_{11}\pm H_{12}}{1\pm S_{12}}$$

and

$$\Psi_{\pm}=\frac{\phi_a(1)\phi_b(2)\pm\phi_a(2)\phi_b(1)}{\sqrt{2(1\pm S_{12})}}, \tag{10}$$

where $\phi_a(1)=\phi_a(\mathbf{r}_1)$, $\phi_a(2)=\phi_a(\mathbf{r}_2)$, etc., and H_{11}, H_{12}, and S_{12} are defined as

$$H_{11}=\int\int \phi_a(1)\phi_b(2)\mathscr{H}(1,2) \times\phi_a(1)\phi_b(2)d^3r_1d^3r_2,$$

$$H_{12}=\int\int \phi_a(1)\phi_b(2)\mathscr{H}(1,2) \times\phi_a(2)\phi_b(1)d^3r_1d^3r_2,$$

$$S_{12}=\int\int \phi_a(1)\phi_b(2)\phi_a(2)\phi_b(1)d^3r_1d^3r_2.$$

The matrix elements H and S were evaluated using the hydrogen-atom wave function: $\phi(r)=\exp(-r/a_0)/\sqrt{\pi}a_0^{3/2}$, where a_0 is the Bohr radius (i.e., 0.529 Å). H and S are only functions of the interatomic distance and the electron charge and mass. The two signs in Eq. (10) correspond to "bonding" for the + sign and "antibonding" for the − sign. If the molecule is in the antibonding state, the electronic interactions are purely repulsive. The state with attractive electronic interactions corresponds to the bonding state. The corresponding wave function is Ψ_+, i.e., the symmetric combination. By the Pauli principle the wave functions must be antisymmetric upon exchange of the two electrons. Thus, the spin part of the wave function must be antisymmetric, i.e., the electrons must have opposite spin. For Ψ_- the electrons have parallel spin. The difference of the interaction energy between parallel and antiparallel electron spins is called the "exchange interaction."

To evaluate the binding energy of the molecule, we must add to the electronic energy the proton-proton interaction energy:

$$E_{\text{binding}}=\frac{e^2}{R}+\frac{H_{11}+H_{12}}{1+S_{12}}. \tag{11}$$

The binding energy as a function of the interatomic distance is plotted in Fig. 2. The minimum energy occurs at $R_0=0.8$ Å and the difference in energy from the isolated hydrogen atom is about 3.2 eV. The corresponding values from experiment are 0.74 Å and 4.75 eV. We also show in Fig. 2 the energy of the molecule if the electrons were in the antibonding state. The interaction between electrons of parallel spins (paired electrons) is always repulsive compared to the isolated hydrogen atoms.

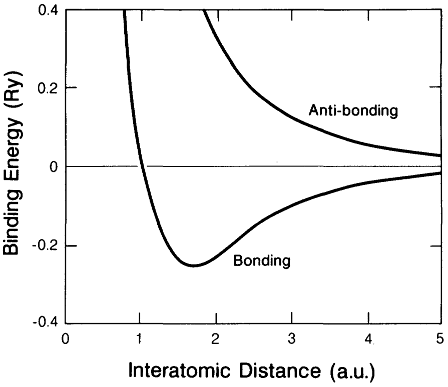

FIG. 2. Energy of the hydrogen molecule as a function of interatomic distance. The most stable curve is the bonding state where the electrons are of opposite spin. The antibonding state occurs if the interacting hydrogen atoms have the same spin. After Heitler and London (1927).

In Fig. 3, the electron distribution functions for the bonding and antibonding states are sketched. The square of the wave function is related to the probability of finding the electron in a given region of space. From Fig. 3, the probability of finding the paired electrons (bonding electrons) between the two atoms is much higher for the bonding than for the antibonding state.

1.4.3 Solid-State Treatments of the Covalent Bond. Although the agreement between experiment and the Heitler–London theory is only fair by contemporary standards (McLean *et al.* 1960), the basic principles of the covalent bond were demonstrated. Electrons of opposite spin can pair up to form a covalent bond between like species of open-shell atoms. For the solid state, the problem is compounded by the number of interacting atoms. However, the essential features are the same.

To handle the large number of electrons and nuclei in solids, an additional approximation is usually incorporated beyond the Born–

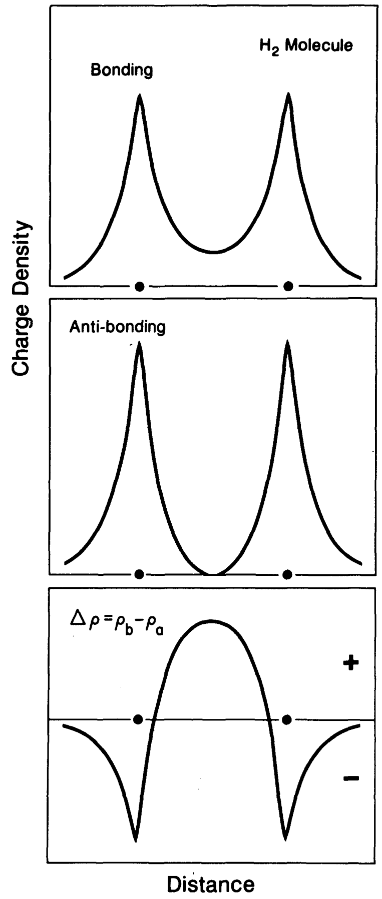

FIG. 3. Schematic charge density for the bonding and antibonding states of the hydrogen molecule. Also shown is the charge density difference between the bonding and antibonding states. The bonding state puts more charge between the atoms.

Oppenheimer approximation. The "one-electron" (or Hartree) approximation treats the interaction of an electron in the solid by averaging over the motion of all the other electrons in the solid. With this approximation, the Schrödinger equation is written as

$$\mathscr{H}\Psi=\left(-\frac{\hbar^2\nabla^2}{2m}+V(\mathbf{r})\right)\Psi(\mathbf{r})=E\Psi(\mathbf{r}). \qquad (12)$$

One of the chief efforts in theoretical investigations of solids concerns the construction of accurate one-electron potentials. Various approximations have been used to address this issue, which is called the "many-body" potential problem. A useful approximation is to assume an effective potential for the exchange energy. Often, this effective potential is based on the electron distribution function, or the electron charge density. Such theories are called "density-functional" theories (Hohenberg and Kohn, 1964) (see ELECTRONIC STRUCTURE OF SOLIDS).

Given Eq. (12), it is often convenient in the solid state to form wave functions from combinations of atomic orbitals matching the symmetry of the neighboring atoms (Madelung, 1981). Crystalline silicon forms in the diamond structure, which has four nearest neighbors arranged in a tetrahedral pattern. The ground-state configuration of silicon is $3s^2 3p^2$. From the two s and two p orbitals, one can form so-called sp^3 hydrids. Such a combination is often very accurate for the covalent bond. For silicon, one might write

$$\begin{aligned}
\Psi_1&=(s+p_x+p_y+p_z)/2,\\
\Psi_2&=(s-p_x-p_y+p_z)/2,\\
\Psi_3&=(s+p_x-p_y-p_z)/2,\\
\Psi_4&=(s-p_x+p_y-p_z)/2,
\end{aligned} \qquad (13)$$

where (s,p_x,p_y,p_z) are atomic wave functions for the valence states in silicon. Other common combinations of wave functions are sp^2 (trigonal planar), d^2sp^3 (octahedral), and d^4sp^3 (dodecahedral). Carbon can exhibit both sp^2 and sp^3 bonding in graphite and diamond structures, respectively. Linus Pauling (1960) was one of the first to recognize the correspondence between orbital symmetry of the bonding orbitals and the resulting structure.

1.5 Metallic Bonds

Metals are fundamentally different from the solids we have discussed. Unlike van der Waals solids, ionic crystals, or covalent crystals, metals have electrons not localized on atoms (or ions), or in a bond. Simple wave-function descriptions based on Eq. (13) are inappropriate for delocalized, or metallic, bonds. The electrons are free to traverse the solid at large. The resistivity of a metal ranges from zero for a superconductor to about 10^{-2} Ω cm. From 10^{-2} to about 10^9 Ω cm, the solid is a semiconductor, and above 10^9 Ω cm the solid is considered an insulator. Another important distinction between a metal and a semiconductor (or insulator) is that as the temperature of the solid is increased, the conductivity of a metal is reduced, whereas for a semiconductor the conductivity increases. The conductivity of the semiconductor increases because as the temperature of a solid increases, the electronic energy increases. This increase in energy can be exhibited by exciting an electron out of a bond and into a delocalized state that can conduct electricity. In the language of band theory, the electron is excited from a completely filled valence band to an empty conduction band. For a metal, the temperature increases the possibility of a lattice vibration (phonon) scattering an electron, thereby inhibiting its flow through the solid, but does not change the accessibility of the empty states.

From a bonding point of view, if the solid has more nearest neighbors than valence electrons, then an elemental crystal is usually a metal. The covalent bonds cannot be saturated by the available number of electrons. (Inert gases are the notable exception, as the constituent atoms possess closed shells).

1.5.1 Delocalized Electrons. The cohesive energy of metals is attributed to a delocalization of the valence electrons of the atoms into wave functions that extend throughout the solid. By "delocalizing" an electron over several atoms the kinetic energy of the electron is reduced and the electronic energy of the solid is lowered. As a very simple example, consider an electron localized in a cube. The lowest-energy state of this electron is given by $E_0 = 3(h^2/8mL^2)$, where L is the length of the cube. If the length of the cube is doubled, E_0 is reduced to $E_0/4$.

1.5.2 Wigner–Seitz Method. Eugene Wigner and his first graduate student, Frederick Seitz, were among the first to account for the cohesive energy of a simple metal. Their paper "On the Constitution of Metallic Sodium" was published in 1933 in *The Physical Review*. This paper presented a simple model for calculating the electronic structure of sodium. A key feature was to view the crystal as being composed of identical cells that surrounded single metallic ions (see ELECTRONIC STRUCTURE OF SOLIDS). The conduction electron in each cell would interact only with the ion within the given cell.

Wigner and Seitz (1933,1934) considered the cell to be spherical and equal in volume to the primitive crystallographic cell. At the cell boundary they assumed the derivative of the wave function must vanish:

$$\left.\frac{\partial \Psi(r)}{\partial r}\right|_{r=R_0} = 0, \tag{14}$$

where R_0 is the size of the Wigner sphere. This approximation is appropriate for close-packed metals such as fcc, hcp, or bcc structures. For bcc sodium, $R_0 = 3.95$ Å.

Wigner and Seitz suggested the reason metals cohered was that their wave functions were probably "smoother" than the wave functions of isolated atoms. In a free atom the wave function of the electron must go to zero at large distances, whereas in the solid it does not. The boundary condition in Eq. (14) reduces the curvature of the wave function at the cell boundary relative to the atomic limit. The reduction in curvature reduces the kinetic energy of the electron relative to the free atom. (Recall that the quantum-mechanical operator for the kinetic energy is the curvature of the wave function. Little curvature or "smooth" wave functions have less kinetic energy than rapidly oscillating wave functions.) The reduced kinetic energy leads to the binding in Na and other metals. The cohesive energy determined by the Wigner–Seitz method is about 1 eV/atom. This number agrees with experiment.

The Wigner–Seitz paper was important as one of the first papers to calculate the cohesive energy of a *real* material, as opposed to an idealized model. Moreover, it spelled out in some detail how to proceed in a logical progression to determine the properties of a metal. In this sense, their paper marked the beginning of the application of electronic structure theory to the quantitative prediction of materials properties.

Cohesion in transition metals and rare-earth metals is more complex than in simple metals such as sodium. In transition metals, the *d* orbitals are sufficiently localized to form "covalent"-like bonds between the atoms.

1.6 Mixed Bonding in Solids

We have concentrated on archetypical examples of bonding as a guide to explaining the cohesive energy of a solid. However, real solids are not perfectly ionic or covalent or metallic. The constituents of all solids are nuclei and electrons which interact via the Coulomb law subject to the Pauli principle. The imposition of terms such as "ionicity" or "covalency" on the chemical bond is useful only so far as it aids us in predicting the properties of solids.

Even in elemental solids, one can have "mixed" bonding. Carbon in the graphite structure shows strong covalent bonding within the graphitic planes, whereas between the planes van der Waals forces are thought to be important. The intraplanar bonds are thought to be among the strongest covalent bonds formed for any solid; the interplanar bonds are among the weakest solid-state bonds.

An important theory for examining semiconducting or insulating solids has been developed by Phillips (1973). He applied a dielectric theory to describe the fractional ionicity of the chemical bond in binary crystals of the form $A^N B^{8-N}$, where A and B are not transition metals, and N is the number of valence electrons on the cation. Ionicity is a measure of the charge transferred from one atom to another upon the formation of a chemical bond. This method has been very useful in predicting chemical trends in solids, including heats of formation. This theory is based on dividing the "average optical gap" in the AB crystal into a covalent and ionic contribution. The average optical gap corresponds to the energy at the maximum photon absorption in the optical spectrum of a solid. In the dielectric theory the ionicity of a crystal is of the form

$$f_i = C^2/(E_h^2 + C^2), \qquad (15)$$

where C is the ionic contribution to the gap and E_h is the covalent contribution. The average gap, E_{gap}, is given by $E_{\text{gap}}^2 = C^2 + E_h^2$. If the optical spectra are measured, this determines E_{gap}. E_h can be found by scaling the covalent contribution with the lattice constant of a perfectly covalent solid such C, Si, Ge, or Sn.

For elemental crystals such as silicon, $C = 0$, and the ionicity of the crystal $f_i = 0$. For a highly ionic crystal such as RbF, $C \gg E_h$ and $f_i \approx 1$. It can be shown that this ionicity scales roughly as $f_i \approx 1 - N_a/4$, where N_a is the number of valence electrons on the free-atom cation. In Table 5 ionicity values for some common semiconducting crystals are given.

Another successful theory for covalent solids treats a solid as a "large molecule." The wave functions for the solid are handled as in Eq. (13). The various matrix elements that enter the theory are parametrized or extracted from simple scaling arguments (Harrison, 1980).

Table 5. Phillips's ionicity parameters for selected tetrahedral semiconductors. From Phillips (1973).

Crystal	E_h (eV)	C (eV)	E_{gap} (eV)	f_i
C	13.5	0	13.5	0
Si	4.77	0	4.77	0
Ge	4.31	0	4.31	0
Sn	3.06	0	3.06	0
BP	7.44	0.68	7.47	0.006
AlSb	3.53	2.07	4.41	0.250
BeSe	5.65	3.36	6.57	0.261
AlP	4.72	3.14	5.67	0.307
GaP	4.73	3.30	5.75	0.327
AlN	8.17	7.30	11.0	0.449
InN	5.93	6.78	8.99	0.578
ZnO	7.33	9.30	11.8	0.616
HgTe	2.92	4.0	5.0	0.65
CuI	3.66	5.50	6.61	0.692
CuBr	4.14	6.90	8.05	0.735
AgI	3.09	5.70	6.48	0.770
HgS	3.76	7.3	8.3	0.79

In intermetallic alloys, the interplay between ionic, metallic, and covalent contributions can be complex. For example, CsAu is an intermetallic compound that is strongly ionic and is a semiconductor. NaTl can be thought to contain elements of covalent, ionic, and metallic binding simultaneously (Anderson, 1963).

In Table 6, we summarize the types of bonding that lead to cohesion in solids.

1.7 Noncrystalline Solids

Another issue that increases the complexity of understanding cohesion comes from noncrystalline solids such as glasses. Normally when solids are formed from the melt, a discontinuous change in the structure occurs. However, when a material forms from a normal liquid state and shows no discontinuous change, it is called a glass. A glass becomes more rigid through a progressive increase in its viscosity. In principle, a liquid cannot support a shear whereas a solid does. In practice, a liquid whose shear viscosity is greater than about 10^{13} P is called a glass. This value defines the boundary between the liquid and glassy state (Kittel, 1976; Jones, 1971; Goldstein and Simha, 1976). Glasses are char-

Table 6. Properties of the chemical bond in the solid state, after Anderson (1963).

Type of chemical bond	Examples	Origin of cohesion	Properties	Occurrence
Molecular	Xe,Ne,N_2	Fluctuating dipoles	Good insulator, weak binding, close-packed structures	Rare gases, halogen dimers
Hydrogen bond	H_2O,$[HF_2]^-$	Bare proton bond	Fairly weak bonds, loose structures. Peculiar dielectric properties	H with electronegative elements such as O or F
Ionic	NaCl,CsCl	Electrostatic attraction between closed-shell ions	Strong bonds, close-packed structures, ionic conduction	Electropositive with electronegative elements
Metallic	Na,Mg,Al	Lowered kinetic energy; delocalized bonds	Moderate to strong bonds, high conductivity, close-packed structures	Most common bond of elemental solids
Covalent	C,Si,Ge	Shared electrons, exchange forces important	Hard, open structures, strong bonds	Uncommon, tetravalent elements

acterized by a lack of crystalline order as evidenced by the absence of sharp x-ray diffraction patterns. Glasses may have one or two diffuse rings in the x-ray diffraction pattern, which is in contrast to the sharp rings of polycrystalline aggregates. Glasses also have the elastic properties of an isotropic solid.

Glasses, or amorphous solids, are often formed by rapid cooling. In Fig. 4, a schematic volume-versus-temperature curve is illustrated for a glassy material. With rapid cooling, the individual atoms have not had time to assume the lowest-energy structure; the resultant structure is often noncrystalline. Although a glass may not be in absolute thermodynamic equilibrium, it may be stable for extended periods of time with respect to the crystalline state.

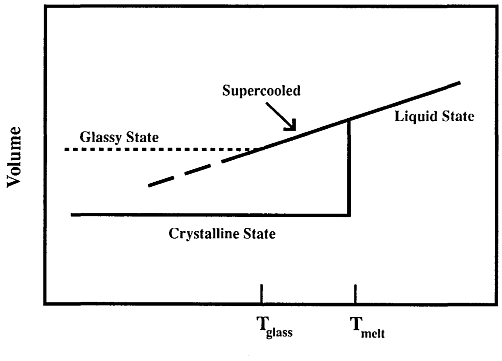

FIG. 4. Schematic volume-vs-temperature relation for a glassy material. The liquid may be supercooled below the melting point T_{melt}. At the glass transition temperature T_{glass}, the supercooled liquid becomes a glass. The details of the glass transition are somewhat dependent on the cooling rate. [Kittel (1976).]

Thus, it may be possible to discuss quasiequilibrium thermodynamic variables, e.g., a glass transition temperature. The microscopic structure of glassy solids is poorly understood and is a contemporary research topic. One suggestion for the structure of a glassy solid is that such a solid is really a disordered assemblage of microcrystallites. Another suggestion is that such a structure is a continuous network with short-range order. The lack of detailed structural information greatly inhibits our understanding of these materials.

One promising approach to remedy this situation is to model interatomic forces, or use quantum mechanics to find the true interatomic forces. These forces are then used within molecular dynamics simulations (Ciccotti *et al.*, 1987) to model the crystallization process. In this approach, the atoms are given initial velocities corresponding to a high temperature. The classical trajectories are computed as a function of time. The kinetic energy of the assemblage is lowered systematically, corresponding to a reduced temperature. Initially, if the temperature is above the melting point, the system will correspond to a liquid. As the temperature is reduced below the melting point, the system will solidify. If this process could be performed slowly, the solid-state form could be crystalline. However, the time scale for computer simulations are often orders of magnitude shorter than the experimental times for crystallization. Hence, the final solid state is invariably amorphous. In addition to the problem of time scales, several issues have limited this technique. One is the construction of accurate interatomic potentials, or quantum-mechanical forces. This is particularly true of covalent solids where the interaction between atoms is not pairwise.

2. TOTAL-ENERGY CALCULATIONS OF THE COHESIVE ENERGY

In 1929, the famous physicist P.A.M. Dirac wrote: "The underlying physical laws necessary for the mathematical theory of ... the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble." This has been the case until recently. While some approximations are necessary, to handle many-body forces such as exchange and correlation, it is now possible, for many solid-state systems, to determine without experimental input the cohesive energy of a crystalline solid.

Two major advances have contributed to the present capabilities. First, fast computing machines now allow complex equations to be solved. A second advance was the construction of simple potentials for handling exchange and correlation. (*Correlation* interactions are any many-body terms other than *exchange*.) These advances allow the total electronic energy of an atom in the solid and the energy of an isolated atom to be computed. These two values known, the cohesive energy is determined from the difference.

The spatial and the energy distribution of electrons in solids must be known to compute the total energy of a solid. These distributions can be determined by a solution of the one-electron Schrödinger equation:

$$\left(\frac{-\hbar^2\nabla^2}{2m}+V_{\text{ion}}(\mathbf{r})+V_{\text{H}}(\mathbf{r})+V_{\text{xc}}(\mathbf{r})\right)\Psi(\mathbf{r})$$
$$=E\Psi(\mathbf{r}). \qquad (16)$$

The first term corresponds to the kinetic energy. $V_{\text{ion}}(\mathbf{r})$ is the bare potential of nuclei. $V_{\text{ion}}(\mathbf{r})$ can also be taken to be an effective potential that includes the potential of the bare nuclei plus the tightly bound core electrons, i.e., a pseudopotential (Cohen and Chelikowsky, 1989). $V_{\text{H}}(\mathbf{r})$ is the electrostatic potential that can be solved from the Poisson equation. $V_{\text{xc}}(\mathbf{r})$ is the exchange-correlation energy, which is taken to be a functional of the charge density. The total energy of the crystal can be determined by the following expression:

$$E_{\text{total}}=\sum_i E_i-\frac{1}{2}\int \rho(\mathbf{r})V_{\text{H}}(\mathbf{r})d^3r+\int \rho(\mathbf{r})$$
$$\times[\epsilon_{\text{xc}}(\mathbf{r})-V_{\text{xc}}(\mathbf{r})]d^3r+E_{\text{ion-ion}}, \qquad (17)$$

where the charge density is given by

$$\rho(\mathbf{r})=\sum_{\text{occupied}} |\Psi(\mathbf{r})|^2,$$

summing over the occupied states. The lead term in Eq. (17) is a sum over occupied energy levels. The second term is to subtract off double-counting terms. The exchange and correlation terms require some care: The potential energy term has been subtracted from the total-energy expressions, and the energy density ϵ_{xc} has been added.

The total-energy expression in Eq. (17) is nontrivial to evaluate as the electrostatic terms diverge individually. If we have an infinite solid, then we have an infinite number of nucleus–nucleus and electron–electron interactions. These terms can be handled by an Ewald summation method.

Table 7 illustrates the accuracy of this method for some elemental solids. The determination of the cohesive energy is the most difficult part of these calculations. The total energy of the solid and atom is determined by independent calculations and the energy difference taken. The local-density approximation is expected to be accurate to about ≈10%. If the error in the atom and the error in the solid are in the same direction, then the cohesive energy can be quite accurate. Typically, the cohesive energy determined in this manner is in error by about 5%–10%. Generally, the cohesive energy determined by this method is greater than the experimental value.

3. HIGH-PRESSURE TRENDS IN COHESION

If we can predict the cohesive energy of a solid as a function of pressure, then we can compare to pressure measurements. Specifically, if we determine the cohesive energy for different crystal structures as a function of volume, we can predict which structure will be most stable as a function of pressure.

Table 7. Calculated cohesive energies and structural properties for selected elemental solids (Yin and Cohen, 1982; Chelikowsky *et al.*, 1986; Chelikowsky and Chou, 1988). In the case of ruthenium, two lattice parameters are required to specify the hexagonal-close-packed (hcp) structure (Kittel, 1986).

Crystal	Lattice parameters (Å)	Cohesive energy (eV/atom)	Bulk modulus (10^{11} N/m^2)
Carbon (diamond)			
Theory	3.60	8.10	4.33
Experiment	3.57	7.35	4.43
Silicon (diamond)			
Theory	5.45	4.84	0.98
Experiment	5.43	4.63	0.99
Germanium (diamond)			
Theory	5.66	4.26	0.73
Experiment	5.65	3.85	0.77
Ruthenium (hcp)			
Theory	2.68(a) 4.16(c)	7.70	3.5
Experiment	2.71(a) 4.28(c)	6.62	3.2
Copper (fcc)			
Theory	3.62	3.83	1.88
Experiment	3.60	3.50	1.42

This area has been a successful use of cohesive energy computations. Many calculations now exist for elemental solids and simple compounds; they are quite accurate when compared to experiment (Yin and Cohen, 1982; Moruzzi *et al.*, 1978). In Fig. 5, the total energy of silicon in various crystal structures is shown as a function of volume. This calculation was one of the first total-energy calculations to reproduce accurately the phases of silicon. A typical pressure-versus-volume relationship is given by the Murnaghan (1944) equation of state:

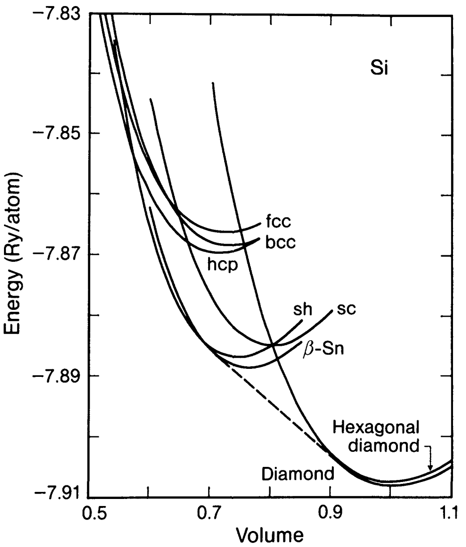

FIG. 5. Calculated equations of states for silicon in various crystal structures. The volume has been normalized to the experimental atomic volume (20.0 Å^3). [Yin and Cohen (1982).]

$$P(V)=\frac{B_0}{B_0'}\left[\left(\frac{V_0}{V}\right)^{B_0'}-1\right], \tag{18}$$

where P is the pressure, B_0 is the equilibrium bulk modulus, B_0' is the pressure derivative of the bulk modulus, and V_0 is the equilibrium volume. Values for V_0, B_0, and B_0' can be determined from fitting the calculated energy versus volume to an equation-of-state function.

In Fig. 5, several polytypes of silicon are considered. The lowest-energy structure is the diamond structure. As the pressure increases, we might expect that closer-packed structures would become stable versus the open diamond structure. This is the case. As the pressure increases, the white-tin or β-Sn structure becomes more stable than the diamond structure. The white-tin structure is a transition structure between diamond and more close-packed structures. The number of nearest neighbors is four, but the two next-nearest neighbors are only 5% more distant. The theoretical transition pressure where diamond transforms to β-Sn is near 10 GPa; this pressure is in good agreement with experiment. At higher pressures, other structures have lower energies than diamond or β-Sn.

Total-energy calculations predicted successfully that hexagonal-close-packed silicon would be stable at higher pressure, i.e., in the range of 40 GPa. Also, high-pressure superconducting forms of silicon were predicted using total-energy methods.

4. CHEMICAL TRENDS IN COHESION AND STRUCTURE

While total-energy, or cohesive-energy, computations are useful for predicting the most stable crystalline phase, these methods are computationally intensive. Empirical methods have been developed that provide a "global" perspective without the computational complexity. For example, the surface energy of an elemental metal, J, can be estimated by $J \approx E_{\text{cohesive}}/V_a^{2/3}$, where E_{cohesive} is the cohesive energy and V_a is the atomic volume. While this expression is only an estimate, it is probably as accurate as any experimental or computational method. Correlations of metallic properties such as surface segregation with these empirical surface energies have been quite successful (de Boer *et al.*, 1988).

Cohesive energies of elemental solids also correlate roughly with the melting points and the bulk moduli. Intuitively, this seems reasonable. High cohesive energies imply strong bonds. Concurrently with strong bonds one expects resistance to compression and thermal disruption. Examples of this are given in Table 1 where the melting points and the bulk moduli for elemental solids are compared to the cohesive energy. As extreme examples, osmium has a cohesive energy of 8.2 eV/atom, a melting point of 3306 K, and a compressibility of about 0.02×10^{-10} m^2/N; for neon the cohesive energy is 0.02 eV/atom, the melting point 25 K, and the compressibility 10.0×10^{-10} m^2/N.

Another example of how such chemical trends can be important involves the search for "superhard" solids. For example, carbon in the diamond structure is much harder than any other material. Corundum is the mineral ranked next below diamond on the hardness scale devised by Mho. Cubic boron nitride, a synthetic material, is at least as hard as corundum and is the second hardest material. Nonetheless, boron nitride is only half as hard as diamond (Bundy, 1974).

Such materials have great technological applications, e.g., making drill bits or protecting electronic devices with superhard films (Bundy, 1974; Sarin, 1987). Often superhard materials are very brittle, and the mechanical properties may be dominated by defects. A contemporary research topic is the use of modern computational methods for cohesive energies to predict the "hardness" of a hypothetical crystal (Sarin, 1987). An example of this effort is given by Liu and Cohen (1989). This work has centered on low-compressibility covalent solids such as diamond and boron nitride. A key relationship used in this study is that the compressibility scales with the nearest-neighbor distance, R, as $R^{3.5}$. Short bond lengths, which correlate with a large cohesive energy, also characterize solids with low compressibilities. Experimentally, there has been considerable interest in growing diamond films (Sarin, 1987; Moustakas, 1989) (see DIAMOND AND DIAMONDLIKE MATERIALS).

A related issue to understanding the chemical forces that lead to cohesion is to construct indices by which to judge whether two solids have similar chemical bonds. This approach is similar to the Phillips ionicity theory discussed in Sec. 1.6. If an index can be found that suggests that the bonding of one solid is "very similar" to another in some quantitative sense, then one would expect the properties of the two solids to be similar. For example, if the atomic volume and electronegativity of two atoms are similar, one expects the respective solids to have similar structures and cohesive properties. Pauling pioneered this effort by the development of electronegativity and atomic radius scales (Pauling, 1960).

Quantitative indices have been developed that accurately predict many structural properties of binary semiconductors and insulators. Many of them are based on a pseudopotential model of a solid. In this model, the valence and core electrons are treated separately. The tightly bound core states are considered to be inert. The pseudopotential reproduces only the valence charge densities of the solid. This type of potential can be used to determine orbitally dependent radii, i.e., radii are different for *s*, *p*, and *d* electrons. Correlations between the cohesive energies and structural energies of simple binary solids with orbital radii have been quite successful. In Fig. 6, we illustrate how the structural energies of binary solids can be systematized

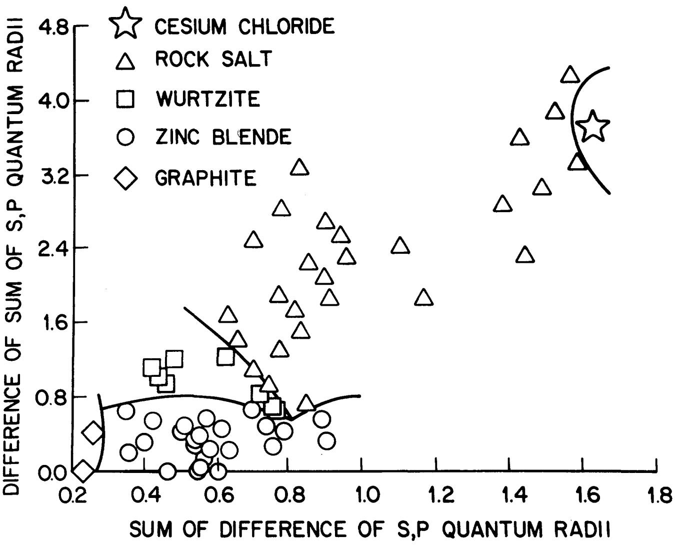

FIG. 6. Quantum structural diagram that systematizes the most stable crystals of simple binary solids. Orbital radii based on pseudopotentials are used as chemical indices to characterize the solid-state bonding. Five different crystal structures are separated on this diagram. The abscissa corresponds to the sum of the difference of the *s,p* radii, i.e., $|r_p - r_s|_A + |r_p - r_s|_B$, and the ordinate corresponds to the difference in the sum of the *s,p* radii, i.e., $|(r_p + r_s)_A - (r_p + r_s)_B|$. $(r_p, r_s)_A$ correspond to the *s*- and *p*-orbital radii for atom *A*. [St. John and Bloch (1974)].

by orbital radii. These types of plots are called *quantum-structural diagrams.* In Fig. 6, if two binary solids share similar combinations of ionic radii, the solids can be expected to have similar structures. One of the most surprising results of this diagram is that the wurtzite and zinc-blende structures can be separated into two domains. These structures differ only with respect to the third nearest neighbors. It would be very difficult to distinguish these two structures with *ab initio* total-energy calculations.

A similar scheme based in part on electronegativities, orbital radii, and the number of valence electrons has been applied to intermetallic binary, ternary, and even some quaternary compounds. It is possible to predict cohesion in hypothetical solids via such schemes. Many of these hypothetical solids have later been found to be stable (Villars and Calvert, 1985).

GLOSSARY

Born–Oppenheimer Approximation: The electronic and nuclear coordinates can be separated in a solution of the Schrödinger equation.

Cohesive Energy: The energy necessary to separate a solid into its constituent atomic species to free atoms at 0 K and 1 atm.

Covalent Bond: A bond formed by the sharing of two electrons with opposite spins.

Dielectric Theory of the Covalent Bond: A theory of the covalent bond based on the optical spectra of tetrahedral semiconductors. It is used to define an ionicity scale.

Exchange Interaction: A spin-dependent Coulomb energy. The spin nature of this Coulomb interaction arises from the Pauli principle.

Hydrogen Bond: A largely ionic bond in which a bare proton stabilizes an interaction between two negatively charged ions.

Ionic Solid: A solid that coheres through the electrostatic interaction of charged ions.

Ionicity: A measure of the charge transferred from one atom to another upon the formation of a chemical bond.

Lattice Energy: The energy necessary to separate an ionic solid into free ions at 0 K and 1 atm.

Lennard–Jones Potential: A pairwise potential commonly used to model the interactions between inert-gas atoms. The form of this interaction between two atoms is $U(R) = 4\epsilon[(\sigma/R)^{12} - (\sigma/R)^6]$, where R is the interatomic distance, ϵ is the binding energy, and σ is a parameter that characterizes the interaction range.

Madelung Constant: A structurally dependent constant determined by the summation of the Coulomb interactions in an ionic crystal.

Metallic Bond: A bond formed by the lowering of the kinetic energy of the conduction electrons.

Pauli Principle: A fundamental law of quantum mechanics: two electrons cannot have the same quantum numbers.

van der Waals Forces: Attractive interaction between closed-shell atoms or molecules. Arises from the interaction of dynamical dipoles.

Wigner–Seitz Method: One of the first methods for quantifying the cohesive energy of simple metals.

List of Works Cited

Anderson, P. W. (1963), *Concepts in Solids*, New York: Addison-Wesley.

Born, M., Huang, K. (1954), *Dynamical Theory of Crystal Lattices*, Oxford, Oxford Univ. Press.

Brewer, L. (1977), Lawrence Berkeley Laboratory Report 3720.

Bundy, F. P. (1974), *Sci. Am.* **231**, 62–70.

Chelikowsky, J. R., Chou, M.Y. (1988) *Phys. Rev. B* **38**, 7966–7971.

Chelikowsky, J. R., Chan, C. T., Louie, S. G. (1986), *Phys. Rev. B* **34**, 6656–6661.

Ciccotti, G., Frenkel, D., Mcdonald, I. R. (1987), *Simulation of Liquids and Solids*, Amsterdam: North Holland.

Cohen, M. L., Chelikowsky, J. R. (1989), *Electronic Structure and Optical Properties of Semiconductors*, Berlin: Springer.

de Boer, F. R., Boom, R., Mattens, W. C. M., Miedema, A. R., Niessen, A. K. (1988), *Cohesion in Metals: Transition Metal Alloys*, Amsterdam: North Holland.

Goldstein, M., Simha, R. (Eds.) (1976), *Ann. NY Acad. Sci.* **279**, 1.

Harrison, W. (1980), *Electronic Structure and the Properties of Solids*, San Francisco: Freeman.

Heitler, W., London, F. (1927), *Z. Phys.* **44**, 455–472.

Hohenberg, P., Kohn, W. (1964), *Phys. Rev.* **136**, 864–871.

Horton, G. K. (1968), *Am. J. Phys.* **36**, 93–119.

Jones, G. O. (1971), *Glass*, 2nd ed., New York: Chapman and Hall.

Kittel, C. (1976), *Introduction to Solid State Physics*, 5th ed., New York: Wiley.

Kittel, C. (1986), *Introduction to Solid State Physics*, 6th ed., New York: Wiley.

Liu, A., Cohen, M. L. (1989) *Science* **245**, 841–842.

Madelung, O. (1981), *Introduction to Solid-State Theory*, Berlin: Springer.

McLean, A. D., Weiss, A., Yoshimine, M. (1960), *Rev. Mod. Phys.* **32**, 211–218.

Moruzzi, V. L., Janak, J. F., Williams A. R. (1978), *Calculated Electronic Properties of Metals*, New York: Pergamon.

Moustakas, T. D. (1989), *Solid State Ionics* **32**, 861–868.

Murnaghan, F. D. (1944), *Proc. Nat. Acad. Sci. USA* **30**, 244–247.

Pauling, L. (1960), *The Nature of the Chemical Bond and the Structure of Molecules and Crystals*, 3rd ed., Ithaca: Cornell Univ. Press.

Phillips, J. C. (1973), *Bonds and Bands in Semiconductors*, New York: Academic.

Sarin, V. K. (1987), *Proceedings of the Third International Conference on the Science of Hard Materials*, Oxford: Elsevier.

Swalin, R. A. (1972), *Thermodynamics of Solids*, New York: Wiley.

St. John, J., Bloch, A. (1974), *Phys. Rev. Lett.* **33**, 1095–1101.

Tosi, M. P. (1964), in: H. Ehrenreich, F. Seitz (Eds.), *Solid State Physics: Advances in Research and Applications*, Vol. 16. New York: Academic.

Villars, P., Calvert, L. D. (1985), *Pearson's Handbook of Crystallographic Data for Intermetallic Phases*, Metals Park, OH: American Society of Metals.

Wigner, E., Seitz, F. (1933), *Phys. Rev.* **43**, 804–810.

Wigner, E., Seitz, F. (1934), *Phys. Rev.* **46**, 509–524.

Yin, M.-T., Cohen, M. L. (1982) *Phys. Rev. B* **26**, 5668–5687.

Further Reading

Anderson, P. W. (1963), *Concepts in Solids*, New York: Addison-Wesley.

Ashcroft, N. W., Mermin, N. D. (1976), *Solid State Physics*, New York: Holt, Rinehart and Winston.

Callaway, J. (1974), *Quantum Theory of the Solid State*, San Diego: Academic.

de Boer, F. R., Boom, R., Mattens, W. C. M., Miedema, A. R., Niessen, A. K. (1988), *Cohesion in Metals: Transition Metal Alloys*, Amsterdam: North Holland.

Harrison, W. (1980), *Electronic Structure and the Properties of Solids*, San Francisco: Freeman.

Kittel, C. (1986), *Introduction to Solid State Physics*, 6th ed., New York: Wiley.

Madelung, O. (1981), *Introduction to Solid-State Theory*, Berlin: Springer.

Moruzzi, V. L., Janak, J. F., Williams, A. R. (1978), *Calculated Electronic Properties of Metals*, New York: Pergamon.

Pauling, L. (1960), *The Nature of the Chemical Bond and the Structure of Molecules and Crystals*, 3rd ed., Ithaca: Cornell Univ. Press.

Pearson, W. B. (1972), *Crystal Chemistry and Physics of Metals and Alloys*, New York: Wiley.

Phillips, J. C. (1973), *Bonds and Bands in Semiconductors*, New York: Academic.

Slater, J. C. (1968), *Quantum Theory of Matter*, 2nd ed., New York: McGraw-Hill.

Villars, P., Calvert, L. D. (1985), *Pearson's Handbook of Crystallographic Data for Intermetallic Phases*, Metals Park, OH: American Society of Metals.

Wells, A. F., (1984), *Structural Inorganic Chemistry*, 5th ed., Oxford: Oxford Univ. Press.

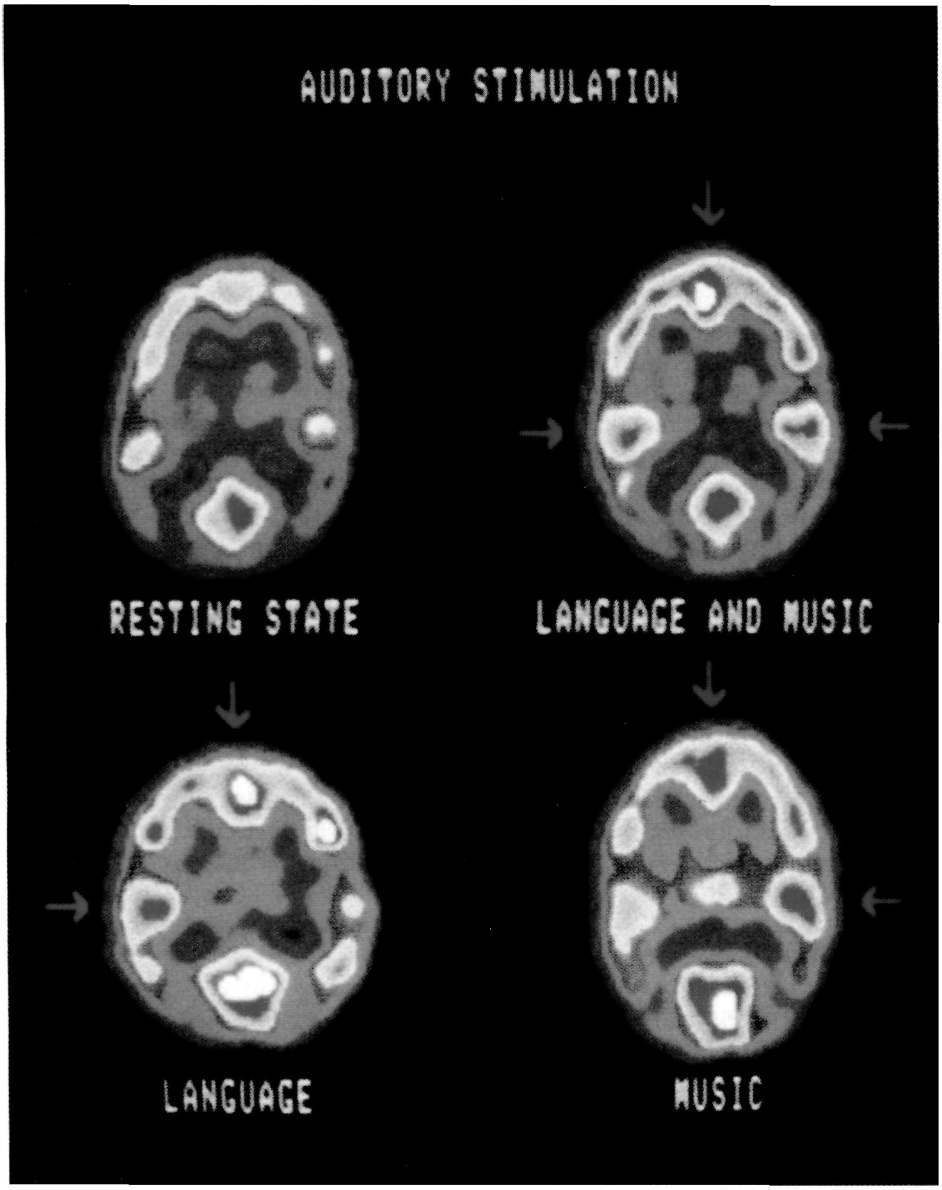
AUDITORY STIMULATION
RESTING STATE
LANGUAGE AND MUSIC
LANGUAGE
MUSIC

Vol. 2 – PLATE 1. PET images of the utilization of deoxyglucose in the brain during different auditory stimuli. (Courtesy of E. Hoffman, University of California, Los Angeles.)

COLLECTIVE PHENOMENA IN SOLIDS

L. J. SHAM, *Department of Physics, University of California, San Diego, La Jolla, California, U.S.A.*

	Introduction	579
1.	**Concepts in Interacting Systems**	581
1.1	Hartree–Fock Approximation	582
1.2	Linear Response Functions	583
1.3	Correlation Energy	584
1.3.1	Total Energy in Jellium	584
1.3.2	The Density-Functional Theory	585
1.4	Quasiparticles	586
2.	**Collective Excitations in the Electron Plasma**	588
2.1	Dielectric Screening and Plasma Oscillations	589
2.2	Landau–Silin Theory of Fermi Liquid	591
2.2.1	Quasiparticle Interaction	591
2.2.2	Local Equilibrium Distribution	592
2.2.3	Specific Heat and Magnetic Susceptibility	593
2.2.4	Landau–Boltzmann Equation	593
2.2.5	Density and Current Response	594
2.2.6	Helicon and Magnetoplasma Modes	594
2.2.7	Spin Waves	596
3.	**Collective Effects Associated with Interband Transitions**	597
3.1	Excitons and the Electron-Hole Liquid	598
3.1.1	The Phase Diagram	599
3.1.2	Plasma Oscillations and Alfvén Waves in EHL	600
3.2	Dynamical Response of the Fermi Sea to Interband Transition	600
3.2.1	Principle of the Mahan or MND Effect	600
3.2.2	Experimental X-Ray Spectra in Metals	603
3.2.3	Optical Spectra in Doped Semiconductor Quantum Wells	603
4.	**Solitons**	604
4.1	The Sine-Gordon and the Nonlinear Schrödinger Equations	606
4.2	Some Applications in Solids	606
5.	**Phase Transition from Collective-Mode Instability**	607
5.1	Itinerant-Electron Ferromagnetism	608
5.2	The Wigner Crystal	608
5.3	Charge- and Spin-Density-Wave States	609
	Acknowledgment	610
	Glossary	610
	Works Cited	610
	Further Reading	612

INTRODUCTION

The simplest model used to understand the properties of a solid is in terms of a collection of independent electrons and harmonic phonons. (See ELECTRONIC STRUCTURE OF SOLIDS; PHONONS IN CRYSTAL LATTICES.) When the interaction between pairs of particles is taken into account, a dichotomy ensues. On the one hand, the particles are turned into "quasiparticles." These quasiparticles have weak interaction between each other so that they can be treated like independent particles although the properties of each quasiparticle may not have the same numerical values as the individual electron or phonon. We use the field-theory term of the particle having been dressed or renormalized. For example, the low-temperature electronic contribution to the specific heat may still be considered as that of a collection of noninteracting fermions, with, however, a heavier mass than the free-electron mass. There is a large class of properties of a solid which may be understood on the

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

basis of these quasiparticles. On the other hand, there are phenomena which involve collective motion of the particles of the system, which are possible only because of the interaction between the particles. The aim of this article is to provide a physical understanding of both aspects of the consequence of the particle interaction. It is not a treatise on many-body techniques for computing the interaction effects, nor is it a complete catalog of all the collective phenomena in solids. It attempts to explain a few key concepts in the many-body theory upon which is built the understanding of the renormalization of the particle and of a number of important collective phenomena.

We begin with the introduction of the concept of exchange and correlation and of conceptual tools needed to describe the collective motion and the single-particle renormalization. An important tool is the linear response function, the change of a property of the interacting system to first order in the external perturbation. A collective mode is deduced from the resonance behavior of a linear response. The ground-state energy of a system contains the effects of interaction. The analysis of the contributions such as exchange and correlation is simplest to illustrate in the homogeneous electron gas, a model of a solid in which the crystal lattice is replaced by a uniform background of positive charges. To cope with the total energies in real solids, the density-function theory is introduced. How a quasiparticle in an interacting electron system is formed is described in terms of the linear response of the system to the addition or removal of an electron.

The collective excitations in a homogeneous electron gas (Sec. 2) provide an excellent introduction to the collective behavior in general and a good model for the simple metals where these modes have been observed. An examination of the linear density response yields the plasma oscillation and the Kohn effect in lattice vibrations. The Landau Fermi-liquid theory is a powerful phenomenological theory which treats the low-lying excited state of an interacting-fermion system. It is used to bring forth the quasiparticle properties and the low-energy collective modes such as the magnetoplasma waves and the spin waves formed in a sea of mobile electrons.

In a solid, the electronic structure consists of a number of energy bands. The transition of an electron from one band to another due to excitation by light is influenced by the interaction with the other electrons in the system. One interaction effect in an insulator or semiconductor may be described as the attraction between the excited electron and the "hole" left behind, forming a hydrogen atom-like object called the exciton. An intense light source can excite a large number of excitons which condense into a two-component plasma, termed the electron-hole droplet (Sec. 3.1). The collective modes in the electron-hole droplet illustrate excitations both in a two-component plasma and in a system of finite spatial extent. In a metal or a doped semiconductor, the conduction electrons react to the presence of the hole left behind by the optical excitation of an electron. Such a reaction leaves remarkable signatures (the Mahan singularities) in the optical spectrum. They are observed in metals and in doped semiconductor heterostructures in which the conduction electrons are confined to a two-dimensional layer (Sec. 3.2).

When nonlinear effects are taken into account, there are collective modes which cannot be described in terms of small disturbances as in the linear response theory. The solitary wave is an important example. Section 4 introduces the soliton solutions of two commonly occurring nonlinear equations, and physical examples in solids.

The ground state of the system upon which the edifice of collective phenomena is built is the "normal" state, which is close to the noninteracting particle state. Under different external conditions, such as temperature or pressure, a collective mode may become unstable, leading to the formation of a different ground state, such as the superconducting, magnetic, or ferroelectric state. For the properties of such states the reader is referred elsewhere in the *Encyclopedia* (see FERROELECTRIC DEVICES; FERROMAGNETISM AND FERRIMAGNETISM; HIGH-TEMPERATURE SUPERCONDUCTIVITY; SUPERCONDUCTIVITY, LOW-TEMPERATURE). The study of instabilities which presage the phase transitions is undertaken in Sec. 5 to introduce charge-density waves, structural transformations, spin-density waves, and the Wigner lattice, which in a two-dimensional system competes with the fractional quantum Hall state.

Each section begins with an overview of the subjects treated. Throughout the article, a few simple concepts have been used to describe a number of important collective modes, spanning one- to three-dimensional solids, as well as to provide an understanding of the idea of quasiparticles. These few important concepts and their illustrations are also meant to facilitate further pursuit in the literature of the collective phenomena not included here.

1. CONCEPTS IN INTERACTING SYSTEMS

For an independent particle, the potential determines its motion. The first approximation in an interacting system is to simulate the influence of the other particles by a potential. This is done first in the Hartree–Fock approximation. The effective potential then has two additional terms. The electrostatic potential due to the charge distribution of the other electrons is easy to understand. The second term is a nonlocal potential due to the exchange interaction of the electron with the others.

What the Hartree–Fock approximation misses is called the correlation effect. Although in principle it is possible to obtain the correlation with progressively more complicated many-electron wave functions, for solids this procedure has not been practical. It is also difficult to extract information on the collective excitations from further refinements of the wave function. A practical alternative for studying correlation effects and collective excitations is to use the idea of linear response. This concept is well known in circuit theory. One might, for instance, subject a circuit to a small ac voltage and measure the resulting change in current, from which some properties of the circuit may be deduced. In general, in responding to a weak external stimulus, a property of the system makes a small change. The ratio of the change to the external perturbation is, to the lowest order approximation, the linear response function, which is characteristic of the system. If the time dependence of the perturbation oscillates with a single frequency, a resonance behavior, i.e., a sharp rise in the response in the neighborhood of a particular frequency, signals the existence of a collective mode, which is self-sustaining without the external stimulus. From the frequency width of the resonance neighborhood, the lifetime of the collective mode may be deduced. A simple example is the plasmon (a quantum of plasma oscillation) as a resonance in the electron-density response to an oscillating electric field in a simple metal.

The total energy of a system is where the interaction effects are most easily delineated. The Hartree–Fock approximation represents a first step in including the interaction effects while maintaining the independent particle form in the ground state. To understand the correlation effect, a fruitful simplication is to remove the crystal lattice effect by replacing the lattice with a uniform positively charged background. In the resulting model, known as jellium (a shorter name for the homogeneous electron gas), the interaction energy is due entirely to the Coulomb interaction between electrons and may be separated into the exchange energy, which is due to the exclusion of electrons with the same spin and momentum, and the remainder called the correlation energy which includes the effect of repulsion between two electrons regardless of whether their spins are parallel or not. Analysis of the interaction energy in terms of the density response leads to the interpretation of the correlation energy partly as screening of the Coulomb potential in the exchange energy and partly as contribution from the zero-point energy of the plasma oscillations.

Quantitative computation of the correlation energy in jellium requires sufficiently draconian measures that extension of the same approach to real solids has only recently achieved acceptable accuracy for a couple of cases. The density-functional theory provides a widely used alternative approach for the real solids. The idea of the theory is to replace the role of the nuclear potential by the ground-state density distribution as the agent to characterize a system. Thus, one would specify a hydrogen atom not by the proton potential but by the electron density distribution of the 1*s* orbital. Then the philosophy of the Thomas–Fermi approximation may be used in approximating a small neighborhood of the system by jellium at the corresponding density. The important difference between the Thomas–Fermi approximation and the local-density approximation within the density-functional theory is that the latter applies the local approximation only to the local exchange and correlation energy and not to the kinetic energy, which provides a marked improvement in including the quantum effects.

Low-lying excited states of an interacting system may be enumerated by analogy with the independent-particle system. The entity corresponding to the independent particle is known as the quasiparticle, which possesses similar properties as the independent particle, such as mass, except that the values may be changed by the interaction effects. The concept of a quasiparticle is established by examining the response of the system to the addition or removal of an electron. The resonance pole in the response function as a function of energy yields the quasiparticle energy and the width of the resonance in energy gives the inverse lifetime of the quasiparticle. The particle in an excited state can decay by exciting other electrons. Phase-space restriction of the fermion system due to the Pauli exclusion principle causes the quasiparticles with energy near the Fermi energy (the smallest energy to add or remove a particle) to maintain their identity for a long time. The effect of all the other particles is lumped into the self-energy term, which is the extension of the Hartree–Fock potential to its most general form including velocity and time dependence (or, equivalently, spatially nonlocal and energy dependence) and the possibility of decay. The renormalized energy includes a term due to the electrostatic potential of all the particles and an exchange and correlation term.

The physical picture of exchange, correlation, the linear response, and collective excitations sketched here will be elaborated upon in the following sections.

1.1 Hartree–Fock Approximation

In adding interaction to a system of independent particles, the first approximation, both historically and conceptually, is to keep the role of the particles and to take into account the interaction effects by allowing each particle to be governed by an effective potential due to all the other particles. Thus, a wave function of a system of N noninteracting electrons is a Slater determinant of one-electron orbitals, which is an $N \times N$ determinant, with each row having the same orbital and each column having the same particle coordinates. The wave function vanishes if two orbitals are the same, thereby satisfying the Pauli exclusion principle. The Hartree–Fock approximation consists in finding the best set of one-electron orbitals in the Slater determinant representing the ground state. According to the variational principle, the best set is the one which yields the lowest total energy for the interacting system.

The resultant orbitals are given by a set of modified one-particle Schrödinger equations:

$$\left[-\frac{\hbar^2}{2m} \nabla^2 + v(\mathbf{r}) + \int dr' \, \frac{e^2}{|\mathbf{r}-\mathbf{r}'|} \, n(\mathbf{r}') \right] \psi(\mathbf{r}) - \int dr' \, \frac{e^2}{|\mathbf{r}-\mathbf{r}|} \, n(\mathbf{r},\mathbf{r}')\psi(\mathbf{r}') = E\psi(\mathbf{r}), \tag{1}$$

where the spin summation is understood. Each electron moves under the combined influence of the external potential $v(\mathbf{r})$; the Hartree potential, which is expressed here as the electrostatic potential due to the electron density distribution $n(\mathbf{r})$ via the Coulomb interaction $e^2/|\mathbf{r}-\mathbf{r}'|$; and the nonlocal Fock exchange potential due to the one-particle density matrix $n(\mathbf{r},\mathbf{r}')$ as a consequence of the Pauli principle. Both the density and the density matrix are given by the N lowest energy orbitals of the above equation. The interactions of an orbital with itself in the Hartree and Fock terms cancels.

The total ground energy can be correspondingly classified as the kinetic energy term, the potential energy term due to the external potential, the Hartree term, and the exchange term. In addition to the ground state, the Hartree–Fock equations yield orbitals which can be used to construct excited states. The excitation energy over the ground state is simply the sum and difference of the single-particle Hartree–Fock energies of the orbitals added to or taken from the ground state.

Whatever in the total energy the Hartree–Fock approximation leaves out is termed the correlation energy. Similarly, if one is to focus on the motion of a single electron, the Hartree–Fock picture provides among the interaction effects the electrostatic interaction due to the other electrons and the exchange interaction with electrons of the same spin orientation due to the Pauli principle. Beyond these terms, the single determinant has no correlation between two electrons due to the Coulomb repulsion. The missing correlation effect can, as we shall see below, be pictured as due to interaction with collective excitations which the motion of the particle under examination excites. Before discussing correlation, we need a formalism for treating the collective excitations, which are also of interest on their own right.

1.2 Linear Response Functions

A powerful conceptual tool to treat the collective phenomena, taking into account the mutual interaction among the constituent particles of a system, is the linear response function. Suppose that the interacting system is subjected to a weak external probe which couples to the property $B(\mathbf{r})$ at the position $\mathbf{r}$ by a "potential" $\delta\phi(\mathbf{r},t)$ at time t. The perturbation to the Hamiltonian is given by

$$\delta H = \int d^3r\, B(\mathbf{r})\delta\phi(\mathbf{r},t). \tag{2}$$

The change of property A to first order in the external perturbation is of the form

$$\delta A(\mathbf{r},t) = \int d^3r' \int dt'\, \chi_{AB}(\mathbf{r},\mathbf{r}';t-t')\delta\phi(\mathbf{r}',t'). \tag{3}$$

The linear response function χ_{AB} is a property entirely of the system. The time dependence follows from the fact that the system itself has no time-dependent forces. Causality makes the response vanish if $t<t'$. The Fourier transform with respect to time of the response function, $\chi_{AB}(\mathbf{r},\mathbf{r}';\omega)$, measures the change of property A corresponding to a perturbation at an angular frequency ω (referred to hereafter as frequency for short).

As a simple example of how the dynamics of the interacting system determines the response function, consider the density response to an electric disturbance in the electron gas, keeping fixed the uniform background of compensating positive charges. Let the perturbing electric potential have a single wave vector $\mathbf{q}$ and a monochromatic frequency ω:

$$\delta\phi(\mathbf{r},t) = \delta\phi(\mathbf{q},\omega)e^{-i\omega t+i\mathbf{q}\cdot\mathbf{r}}. \tag{4}$$

Both the observables A and B are in this case the density distribution $n_{\mathbf{q}}$. In the classical approximation, each electron of charge $-e$ and mass m is set in oscillation with the displacement given by

$$\delta\mathbf{r} = -e\mathbf{E}/m\omega^2, \tag{5}$$

driven by the electric field which is composed of the external perturbation and the induced electric field:

$$\mathbf{E} = -i\mathbf{q}(\delta\phi+\delta\phi_i), \tag{6}$$

where the induced field is due to the change in the electron density:

$$\delta\phi_i = 4\pi\delta n_q/q^2. \tag{7}$$

The change in density is related to the displacement by the volume dilatation:

$$\delta n_q = -nei\mathbf{q}\cdot\delta\mathbf{r}, \tag{8}$$

where n is the uniform density of the unperturbed system. Elimination of the induced potential leads to an expression of the density response which is the ratio of δn to $\delta\phi$:

$$\chi_{nn}(\mathbf{q},\omega) = \frac{ne^2q^2}{m\omega^2}\left(1-\frac{\omega_p^2}{\omega^2}\right)^{-1}, \tag{9}$$

where

$$\omega_p^2 = 4\pi ne^2/m. \tag{10}$$

If the perturbing frequency ω is equal to the plasma frequency ω_p, the density response blows up. In other words, the system is able to sustain a density oscillation of this frequency without external disturbance. This response mode of the jellium is called the plasma oscillation. This example illustrates how microscopic dynamics (in this case a rather crude combination of Newtonian mechanics and electrostatic theory) leads to an expression for a response function and how its behavior yields the interpretation of a collective mode.

If the perturbation is adiabatically turned on or if a weak damping is introduced, a small imaginary number may be added to the frequency ω in the response function. Then the imaginary part of the response function, known as the spectral density, will peak in the neighborhood of the pole of the response function. The peak position gives the collective mode energy and momentum, and the width of the resonance gives the inverse lifetime of the collective mode.

Two concepts related to the density response which arise in the example above are the electric susceptibility function $\tilde{\chi}_{nn}$, which is the density response to the total potential in Eq. (6), and the dielectric function ϵ, which is the ratio of the perturbing potential to the total potential. These two quantities are the extension of the susceptibility and dielectric constant to finite wave vector and frequency. From these definitions and Eq. (6) the two quantities are related to the density response and to each other by the relations

$$\epsilon^{-1}=1+\frac{4\pi}{q^2}\chi_{nn}, \quad \epsilon=1-\frac{4\pi}{q^2}\tilde{\chi}_{nn}. \tag{11}$$

An approximation in which the proper polarization part is replaced by the density response of a noninteracting system is called the random phase approximation. This contains the long-range Coulomb interaction effects.

The spin density response with both A and B replaced by the spin vector σ, known as the magnetic susceptibility function, is a second-rank tensor related to the proper polarization part since the spin-density wave does not create long-range interaction as a charge-density wave does. In an isotropic system, the tensor is diagonal with elements identical to the proper polarization part of the density response.

1.3 Correlation Energy

Interaction plays an important role in the cohesion of a solid (see COHESIVE ENERGY), which is the difference between the total energy of the solid and the total energy of the constituent atoms isolated from one another. In metals, it is found that the correlation energy contribution to cohesion is as important as the Hartree and exchange terms.

An understanding of the essential physics of the correlation effects may be garnered by first neglecting the lattice potential. Thus, the properties of the jellium model (defined in Sec. 1.2) have been studied. Jellium is also a useful model for carriers in a doped semiconductor. (See ELECTRON STATES IN ZERO-, ONE- AND TWO-DIMENSIONAL STRUCTURES.) With the addition of the lattice pseudopotential as a perturbation, the jellium model describes the behavior of simple metals such as the alkali metals and aluminum quite well. (See ELECTRONIC STRUCTURE OF SOLIDS.) Quantitatively, it turned out not to be so straightforward to extend the methods developed for jellium to most of the other solids. Among the various approaches, the density-functional theory is the most widely used and, perhaps, the most successful. Its limitation lies in the inability to describe accurately the strong correlation of d or f electrons in narrow bands. This state of affairs brings us to the present-day research on strong correlation in narrow-band models such as the Hubbard model. (See METAL–INSULATOR TRANSITIONS; HEAVY-FERMION PHENOMENA.) Here we review the concepts arising out of the study of the interaction energy in jellium and the principles of the density-functional theory.

1.3.1 Total Energy in Jellium. The noninteracting part of the total energy is given by the kinetic energy of a single determinant of single-particle states, each of which, in jellium, is a plane wave, with wave vector **k** and spin state σ (± 1 for up and down states), and thus energy

$$\epsilon_{\mathbf{k}}=\hbar^2k^2/2m. \tag{12}$$

The determinant (and, hence, the state of the system) may be specified by the distribution function $n_{\mathbf{k},\sigma}$ which, at zero temperature, assumes the value of unity if the single-particle state $(\mathbf{k},\sigma)$ is occupied and the value zero otherwise. The lowest energy state of n electrons per unit volume has single-particle states occupied up to the wave vector with magnitude k_{F}, the Fermi wave vector. The total energy per particle in the noninteracting approximation is the average kinetic energy per particle:

$$\epsilon_s=\frac{3}{5}\,\epsilon_{k_{\mathrm{F}}}=\frac{2.21}{r_s^2}\ \mathrm{Ry}. \tag{13}$$

The last expression is expressed in terms of the mean electron radius in units of the Bohr radius a_{B} related to the electron density by

$$n=\left[\frac{4\pi}{3}(r_sa_{\mathrm{B}})^3\right]^{-1}. \tag{14}$$

Since the mean distance between two electrons is of the order r_sa_{B}, the Coulomb energy between them is r_s^{-1} Ry. The mean radius r_s is a measure of the ratio of the electron interaction energy to the average kinetic energy.

If the expectation energy is calculated for the single-determinant state, the approximation is the Hartree–Fock approximation (Sec.

1.1). The interaction-energy part is the exchange energy, which, per particle, is

$$\epsilon_x = -\frac{3}{4\pi} k_F = -\frac{0.92}{r_s}. \tag{15}$$

The remainder of the total energy per particle, the correlation energy, is denoted by ϵ_c, so that

$$\epsilon = \epsilon_s + \epsilon_x + \epsilon_c. \tag{16}$$

The exchange and correlation energies can be formulated in terms of the inverse dielectric function defined by Eq. (11) (Pines and Nozières, 1966). If the inverse dielectric function is taken to first order in the Coulomb interaction, the exchange energy is recovered. If the dielectric function is taken in the random phase approximation (see Sec. 1.2), then in addition a correlation-energy term results, which contains the correct logarithmic term in the high-density (or small-r_s) limit:

$$\epsilon_c = [0.062 \ln(r_s) - 0.096] \text{ Ry}. \tag{17}$$

Physically, there are two contributions, one being the static screening of the Coulomb interaction in the exchange due to the excitation of the electron-hole pairs at low energies and the other being the zero-point energies of the plasmon modes. The constant term is corrected by a second-order exchange to the random phase term.

The best expression for the correlation energy is obtained numerically by a Monte Carlo evaluation of the correlated wave function (Ceperley, 1978).

1.3.2 The Density-Functional Theory. Ever since the days of Thomas and Fermi, there had been attempts to formulate quantum mechanics in terms of density in place of the wave function. The advantage over the solution of the many-particle Schrödinger equation would be a more direct relation between the density and the potential and a better way to include the interaction effects. However, the methods remained semiclassical in nature and the quantum effects, such as the density oscillations, were not treated quantitatively. The modern density-functional theory (Hohenberg and Kohn, 1964; Kohn and Sham, 1965) is formally based on the density variable instead of the wave function, but the density is solved by an effective one-particle Schrödinger equation which in principle contains all the interaction effects. This theory avoids the shortcomings of the semiclassical theory and provides a formal framework for approximations for the interaction effects and for the gradient expansion including the Thomas–Fermi approximation as a special case.

Consider a class of systems of electrons. The electrons all have the same mass and interact with the same Coulomb potential. What determines a system in this case uniquely is the potential $v(\mathbf{r})$ due to the nuclei, aside from a trivial additive constant to the potential. In the Schrödinger approach, once $v(\mathbf{r})$ is given, solution of the many-electron Schrödinger equation determines all the properties of the system. In particular, the ground-state density distribution $n(\mathbf{r})$ can be determined. Thus, the density $n(\mathbf{r})$ is said to be a functional of the potential $v(\mathbf{r})$. The fundamental theorem in the density-functional theory due to Hohenberg and Kohn is the converse of the preceding statement. The external potential (to an arbitrary additive constant) and, hence, all the properties of a system are uniquely determined by the ground-state density $n(\mathbf{r})$.

With the help of this theorem, it is possible to write the ground-state energy of a system in the form

$$E = \int d^3r\, n(\mathbf{r}) v(\mathbf{r}) + T_s[n] + \frac{1}{2} \int d^3r \times \int d^3r'\, n(\mathbf{r}) \frac{e^2}{|\mathbf{r}-\mathbf{r}'|} n(\mathbf{r}') + E_{xc}[n]. \tag{18}$$

The first term on the right is the potential energy due to the external potential. The second is the kinetic energy of the noninteracting system with the same density as the interacting system. The notion of a noninteracting system with the same ground-state density as the real system is important to the reduction of the many-electron problem to the solution of a one-particle problem with an effective potential, as we shall presently see. The third term is the electrostatic energy of the electron charge distribution. The fourth term is the exchange-correlation energy. In the limit of constant electron density, these energy terms reduce to the corresponding ones already defined for jellium, with the electrostatic energy terms adding to zero.

A corollary to the Hohenberg–Kohn theorem is the variational principle with respect to the density. The change in the ground-state energy to first order in the density change is zero. This yields a variational equation for solving the density. It only remains to find the energy functional. Two series constructions are possible. One is the perturbation from jellium. The other is a gradient expansion in terms of the density and its derivatives to all orders. If the exchange and correlation energy $E_{xc}[n]$ is neglected, only the noninteracting kinetic energy functional needs to be expanded in Eq. (18). The leading terms without the gradients of the density, corresponding to the energy density of jellium at the local density, reproduce the Thomas–Fermi approximation. The gradient terms also correspond to some attempts to improve the Thomas–Fermi approximation. If the leading term in the gradient expansion of the exchange energy functional, corresponding to the exchange energy density of jellium at the local density, is kept, the Dirac correction to the Thomas–Fermi approximation ensues. In practical terms, computational results of the leading term in the gradient expansion, constituting the addition of exchange and correlation effect to the Thomas–Fermi approximation, provide a good asymptotic approximation, but the gradient corrections are more trouble than they are worth.

A key to the improvement over the Thomas–Fermi-type approximation is to keep the whole single-particle kinetic energy functional $T_s[n]$. For the corresponding noninteracting system, the variational equation in terms of the density variable is just the sum of the functional derivative of $T_s[n]$ and the external potential. The solution for the density can be obtained by solving the one-particle Schrödinger equation. The variational equation for the interacting system, from Eq. (18), has the same form as in the noninteracting case, except that the effective potential now includes the external potential, the Hartree potential (strictly speaking, the electrostatic potential due to the density distribution of all the electrons), and the exchange-correlation potential defined as the functional derivative of $E_{xc}[n]$, i.e.,

$$v_{\text{eff}}(\mathbf{r}) = v(\mathbf{r}) + \int d^3r' \, \frac{e^2}{|\mathbf{r}-\mathbf{r}'|} \, n(\mathbf{r}') + v_{xc}(\mathbf{r}). \tag{19}$$

Thus, solving the many-electron problem for the density is reduced to solving a one-particle Schrödinger equation. What remains is the construction of the exchange-correlation functional. There is a perturbation-theory expression for the functional (Sham, 1985). The most widely used is the local-density approximation, which is the leading term in the gradient expansion, i.e., the exchange-correlation part of the chemical potential of the electron in jellium. (See ELECTRONIC STRUCTURE OF SOLIDS; Schlüter and Sham, 1982; Cohen and Chelikowsky, 1988). An interpolated form of the electron-gas results of Ceperley (1978) is widely used.

The functional can be extended to include the spin density, to superconductors (Oliveira *et al.*, 1988), and to finite magnetic fields by extending the density variable to include the current density (Vignale and Rasolt, 1988).

1.4 Quasiparticles

In this section, we devote our attention to concepts arising out of consideration of one-particle properties. Clearly, an electron propagating in an interacting system has to push other electrons out of the way. Thus, the particle is "dressed." In the Hartree–Fock approximation, Eq. (1), the particle sees, besides the external potential, the electrostatic potential due to the electrons and the nonlocal exchange potential.

To include the correlation effect, the nonlocal and energy independent (or static) exchange term has to be replaced by a more general term, $\Sigma(\mathbf{r},\mathbf{r}',E)$, called the self-energy, which includes the exchange term and is now energy dependent (or dynamic) and complex as well as being nonlocal. An infinite perturbation series can be developed for the self-energy, which shows that the form adopted is the most general form. The energy dependence reflects the retardation effects or the finite energies of the density oscillations and other correlation effects excited by the motion of the electron. The complex self-energy allows for the decay of the electron state under study through exciting other electrons. The effective Schrödinger equation becomes

$$\left[-\frac{\hbar^2}{2m}\nabla^2+v(\mathbf{r})+\int d^3r'\,\frac{e^2}{|\mathbf{r}-\mathbf{r}'|}\,n(\mathbf{r}')\right]$$
$$\times\psi_j(\mathbf{r},E)+\int d^3r'\;\Sigma(\mathbf{r},\mathbf{r}',E)\psi_j(\mathbf{r}',E)$$
$$=W_j(E)\psi_j(\mathbf{r},E). \qquad (20)$$

From the eigenvalues $W_j(E)$ and eigenfunctions $\psi_j(\mathbf{r},E)$, a Green's function can be constructed (see Mahan, 1990):

$$G(\mathbf{r},\mathbf{r}',E)=\sum_j \frac{\psi_j(\mathbf{r},E)\psi_j^\dagger(\mathbf{r}',E)}{E-W_j(E)}\,. \qquad (21)$$

The one-electron Green's function may be regarded as the response function to the addition or removal of an electron to the system. By analogy with the collective-mode concept in relation to the response function, a pole of the one-electron Green's function gives a self-sustaining mode which is identified as a state of the quasiparticle. The quasiparticle energy is given by the zeros of

$$E=\text{Re}\ W_j(E), \qquad (22)$$

where the right-hand side is the real part of $W(E)$. The imaginary part of $W(E)$ at the quasiparticle energy gives the reciprocal lifetime of the quasiparticle.

The quasiparticle state resembles that of an independent electron with its motion much modified by the interaction. If there is no interaction among the electrons, the eigenvalue $W_j(E)$ is real and independent of E (yielding infinite lifetime) and the residue of the pole, when integrated over all space for $\mathbf{r}=\mathbf{r}'$, is unity. The inclusion of interaction reduces the corresponding residue, known as the wave function renormalization. Thus, a quasiparticle is not completely equivalent to an electron including the interaction effects since there is a remaining piece of the Green's function besides the pole part. Nonetheless, for one-particle properties of the system, a simple picture of quasiparticles with one-to-one correspondence between the quasiparticle states and the independent-electron states is useful. The Fermi surface in momentum space which separates the unoccupied quasiparticle states from the occupied ones must enclose the same volume as the Fermi surface of the independent electrons. This is known as the Luttinger theorem. Quasiparticle states close to the Fermi surface are long-lived. Because of the phase-space restriction by the exclusion principle, the reciprocal lifetime of the quasiparticle close to the Fermi surface is proportional to the square of the energy difference from the Fermi energy. This follows from the process of decay of the particle coming primarily from the Auger process of exciting an electron-hole pair, which must remain within roughly the same energy range from the Fermi surface. The final state of the particle contributes no phase-space restriction since its state is determined by energy and momentum conservation from the initial state and the states of the electron-hole pair. The quasiparticle picture is, thus, valid for low-lying excited states.

For jellium, the self-energy can be constructed in the random-phase approximation in terms of the dielectric function. It is of a similar form to the exchange term with the Coulomb potential screened by the energy-dependent, nonlocal dielectric function. The physical picture is that of the Fock exchange term statically screened and with the addition of a correlation hole. At the Fermi level, because of the hindrance of the other electrons and of phonons, the speed of the quasiparticle is lower than that of a free electron at the Fermi level. The effective mass m^* of the quasiparticle at the Fermi surface is defined as the ratio of its momentum to its velocity. Thus, the interaction effects cause the effective mass to be larger than the bare electron mass m. This effect manifests itself in the enhanced density of states and, hence, the low-temperature specific heat. The isotropy of jellium means that the quasiparticle Fermi surface is spherical as is the one for the free electrons. Since both Fermi surfaces enclose the same volume of phase space, the Fermi momentum for the quasiparticles is equal to that of the noninteracting system with the same number of electrons. Without the isotropy of jellium, the shape of the Fermi surface (in general) is changed by the interaction effects.

For real solids, the simplest method to construct the self-energy is to start with the density-functional orbitals (see Godby *et al.*, 1987, and references therein). Calculations of the quasiparticle energies demonstrate that the band structure is not well given by the eigenvalues of the one-particle Schrödinger with the potential in Eq. (19), particularly the band gap. The discrepancy in the band-gap values can be understood in terms of the discontinuous jump of the exchange-correlation potential v_{xc} when one electron is added to the insulating ground state. On the other hand, for metals, the density-functional eigenvalues yield a good approximation for the Fermi surface although they cannot give the density-of-states enhancement.

2. COLLECTIVE EXCITATIONS IN THE ELECTRON PLASMA

In this section, we examine the collective modes due to interaction among the electrons, in which the band structure is of secondary consideration. Thus, the jellium model introduced in the last section serves admirably to explore the physics of a variety of collective oscillations of the electron system, which acts like a low-temperature degenerate plasma. The three-dimensional jellium model is a good approximation to simple metals, which are generally elements on the left-hand side of the periodic table although lead is included, and to the charge carriers in heavily doped semiconductors. Two-dimensional jellium is a good approximation to the confined layers to carriers in semiconductor heterostructures (see HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR). One-dimensional jellium can be used to glean some collective effects and instabilities of the linear conductors (see ELECTRONIC STATES IN ZERO-, ONE-, AND TWO-DIMENSIONAL STRUCTURES).

The collective excitations in the electron plasma are obtained by studying the linear response to a weak electric or magnetic field with a sinusoidal spatial variation (with wave vector $\mathbf{q}$) and a monochromatic time oscillation (with angular frequency ω). Allowing the heavy positively charged background to oscillate yields in addition a model for the long-wavelength lattice vibrations (phonons) in a simple metal. A revisit of the density response including the quantum effects at a finite wave vector and frequency yields the static screening of a charged impurity, the plasma oscillations at high frequencies, screening of the acoustic phonons at low frequencies, and the image of the Fermi surface in the Kohn effect at the wave vector equal to the Fermi diameter.

To treat low-energy charge- and spin-density oscillations under more complicated circumstances, such as in the presence of a strong static magnetic field, we use the Fermi-liquid theory due to Landau and Silin, which is a powerful phenomenological theory treating the low-lying excited states of the interacting system as a number of quasiparticles which have been described in Sec. 1, but now including the residual interaction between quasiparticles which is essential to the existence of the collective modes. Such a phenomenological theory has been justified from first principles starting with bare electrons and Coulomb interaction between electrons. To describe a nonequilibrium state of the interacting system, Landau's Fermi-liquid theory divides the position and momentum phase space of the quasiparticle into cells, with dimensions large enough for each cell to be a thermodynamic system but small compared with the corresponding variations of the external perturbation and, therefore, of the collective mode, and defines a quasiparticle distribution for all the cells. The time evolution of the quasiparticle distribution is governed by the kinetic theory. Knowledge of the quasiparticle distribution in thermal equilibrium leads to thermodynamic properties. Knowledge of the time dependence of the quasiparticle distribution leads to the collective modes. Moreover, the results delineate clearly for each property whether the interaction effects are due to the dressing of the quasiparticle or to the interaction between quasiparticles.

The difference between an uncharged Fermi liquid, such as ^{3}He, and the electron system is the existence in the latter of a long-range electrostatic potential due to the density variation of the quasiparticles. The long-range interaction, besides giving rise to the plasmon in the electron system, tends to shield from experimental observation some of the spin-independent collective modes which exist in the uncharged Fermi liquid. A good example is the zero sound which is due to the collective oscillation corresponding to an isotropic change of the Fermi surface. In the charged Fermi liquid, such a mode is coupled to the compressional part of the lattice vibration.

The effect of the quasiparticle interaction in the zero sound is submerged in the lattice effect.

The presence of a uniform and static magnetic field opens up the possibility of interesting collective modes which are not totally screened by the long-range interaction. There are two classes, those which owe their very existence to the quasiparticle interaction and those which exist because of the long-range electric potential but are modified by the quasiparticle interaction. The former are particularly important in extracting experimental information of the quasiparticle interaction. The cyclotron motion of the electrons is inaccessible by the electromagnetic field except in the surface region (the skin effect). With the magnetic field parallel to the metal surface, arrangement can be made to impart electromagnetic energy in the skin depth to electrons passing through it in their execution of the cyclotron motion. The resultant resonance is known as the Azbel'–Kaner resonance. The dielectric response to electromagnetic waves in a static magnetic field shows transverse plasmons at high frequencies, helicon waves at low frequencies with their polarization vectors winding along the static magnetic field in the same sense as the cyclotron motion, and branches of magnetoplasma waves around the cyclotron frequency or its harmonics. The magnetoplasma modes exist under the condition that they do not create a net density wave or a current, either of which will induce the long-range Coulomb interaction screening out the collective modes. Thus, the Fermi-surface deformation in such a magnetoplasma mode must be neither spherically symmetric (which would produce a density wave) nor a rigid displacement (which would produce a current). Spin waves in metals are truly collective modes in that they cannot exist without quasiparticle interaction. They are unlike the spin waves in a ferromagnet which are due to oscillations of localized spins from their equilibrium orientations. In a metal, the electrons which carry spins are highly mobile. It is their net spin density at each locality which oscillates in time and varies from locality to locality. Like the sound wave in a fluid, the individuality of the particles is lost, but unlike the sound wave, the spin wave in driven by particle interaction and not by collision.

2.1 Dielectric Screening and Plasma Oscillations

The procedure of obtaining the density response discussed in Sec. 1.2 is general and will be followed here. An external electric disturbance induces a change in electron density, which in turn creates an induced electric potential. The total electric potential seen by a charge is the sum of the external potential and the induced potential. The density response expressed as a linear coefficient to the total potential is the proper polarization part $\tilde{\chi}$ in Eq. (11). The only difference is that we forsake the classical treatment of the electron motion to obtain the density response to the total potential in Sec. 1.2, and use an approximation which is also reasonable for finite wave vector and frequency. In the so-called random phase approximation (RPA), the density response to the total potential is equated with the density response of the noninteracting system, which is easily obtained, known as the Lindhard expression:

$$\tilde{\chi}(\mathbf{q},\omega)=\sum_{\mathbf{k},\sigma}\frac{n_{\mathbf{k},\sigma}-n_{\mathbf{k+q},\sigma}}{\omega+i0+\epsilon_{\mathbf{k}}-\epsilon_{\mathbf{k+q}}}, \tag{23}$$

consisting of the excitations of all possible electron-hole pairs, with occupation function $n_{\mathbf{k},\sigma}$ and energy $\epsilon_{\mathbf{k}}$ from Eq. (12). The small imaginary part $i0$ is due to the adiabatic switching on discussed in Sec. 1.2.

Since the induced potential is related to the induced electron density change, it can be expressed in terms of the external potential via $\tilde{\chi}$. The result is that the total potential is given by the external potential divided by the dielectric function which can be expressed in terms of $\tilde{\chi}$ as in Eq. (11). The Lindhard expression for $\tilde{\chi}$ can be evaluated explicitly for the free-electron energy dispersion. In particular, the static dielectric function in three dimensions is

$$\epsilon(q,0)=1+\frac{4\pi e^2}{q^2}N_0\left[\frac{1}{2}+\frac{4k_F^2-q^2}{8k_F}\ln\left|\frac{2k_F+q}{2k_F-q}\right|\right], \tag{24}$$

where N_0 is the density of states at the Fermi level. In the long-wavelength or small-q limit, the polarization part reduces to the density of states and the dielectric screening is of the form

$$\epsilon(q,0)=1+4\pi e^2N_0/q^2. \quad (25)$$

In this Thomas–Fermi approximation, a static external Coulomb potential, e.g., due to a charged impurity, is screened to a short-ranged Yukawa potential.

When the wave vector is equal to Fermi surface diameter, there is a sharp decrease in the dielectric function (24). It has an $x \log x$-type singularity for q near $2k_F$, known as the Kohn singularity. The behavior may be understood by looking at the electron-hole pair contribution to the polarization. The most important ones are those nearly satisfying energy conservation (thus, making the denominator small). As q increases through the diameter of the Fermi surface, the number of such pairs drops sharply. The Kohn singularity is, therefore, an image of the Fermi surface. From this physical explanation, it is clear that the flatter the two pieces of the Fermi surface across one of its diameters are, the stronger the singularity. For two-dimensional or cylindrical Fermi surfaces (Brovman and Kagan, 1974) the Kohn singularity is stronger, being a square-root decrease. In one dimension or parallel plane sections of the Fermi surface, the Kohn singularity is stronger still, having a logarithmic divergence. The significance of the Kohn singularity is that it is reflected in the dispersion of any collective mode which involves the low-frequency density response in the neighborhood of the Fermi diameter. This is observable in the phonon dispersion in metals where the force constants between ions are screened by the electrons. In one dimension, the giant Kohn singularity sets up such a strong charge- or spin-density wave (Overhauser, 1965) that the normal ground state could become unstable. This is an alternative way of viewing the Peierls instability (see Sec. 5).

The Fourier transform to the three-dimensional coordinate space of the screened impurity potential due to the Kohn singularity has a power-law oscillation in space $\sim\cos(2k_Fr)/r^3$, known as the Friedel oscillation. The Friedel oscillation of the electron density in the presence of a charged impurity may be observed by the ensuing change in the hyperfine interaction.

A lattice vibration may be regarded as a collective mode (phonon) in the linear response of the nuclear displacement (see PHONONS IN CRYSTALS). As a simple model for the electron screening of the lattice vibration with a long wavelength, consider the oscillation of the positively charged background. The dielectric response can be analyzed in the same way as for jellium with a rigid positive background. We simply need to add the polarization contribution due to the oscillating positive background to Eq. (11), which in the long-wavelength limit is given by $-\Omega_p^2/\omega^2$, where Ω_p is the ion plasma energy given by Eq. (10) with the electron mass replaced by the ion mass M. Since the ion plasma frequency is much lower than the electron plasma frequency, the electron screening term should be taken at low frequency in determining the zero of the dielectric function which corresponds to the electron-screened ion plasma frequency, then given by

$$\omega^2=\Omega_p^2q^2/4\pi e^2N_0, \quad (26)$$

keeping only the leading term in q using Eq. (25). This is the acoustic branch of the phonon.

In the long-wavelength limit, the electron polarization at a finite frequency is, from the Lindhard expression,

$$\tilde{\chi}(\mathbf{q},\omega)=ne^2q^2/m\omega^2. \quad (27)$$

For small q, the number of electron-hole pairs is limited and this expression can be shown to exhaust the sum rule. Thus, the classical derivation given in Sec. 1.2 and the RPA both actually yield the right plasma frequency for three-dimensional jellium in this limit. In a two-dimensional system, the polarization expression is the same if n denotes the number of electrons per unit area. However, the Fourier transform of the Coulomb potential in two dimensions is $2\pi e^2/q$, yielding from the zero of the dielectric function the 2D plasma frequency for small q given by

$$\omega_p^2=2\pi ne^2q/m. \quad (28)$$

The oscillation of the heavy positively charged background changes the plasma frequency slightly, replacing m by the reduced mass.

In metals, the plasma oscillation is observed by energy loss of injected electrons (Schnatterly, 1979; Raether, 1980). Figure 1 gives an example in aluminum, which also shows the frequency dispersion of the plasmon. The quadratic dispersion can be deduced by keeping the next order term in q in addition to the zero-q limit in Eq. (27). In a

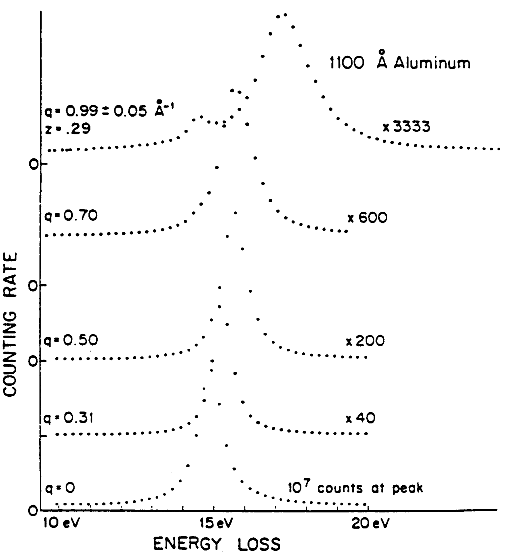

FIG. 1. Measured plasmon spectra of Al from $q=0$ to $1\ \text{Å}^{-1}$. [From Gibbons *et al.* (1976), Fig. 2; reproduced with permission.]

two-dimensional electron gas in a semiconductor heterostructure, the plasma frequency is in the infrared range and can be observed by light absorption (Heitmann, 1986). The problem is that the charge oscillation and, thus, the electric vector of light have to be in the plane of the electron gas and, therefore, the confining interface. The finite wave vector of the plasma oscillation is obtained by laying down metal wire gratings on the interface. Figure 2 shows the observed dispersion of plasma frequency, confirming the square-root wave-vector dependence. Equation (28) has to be modified by replacing the mass with the effective band mass and by making allowance for the dielectric function of the background heterostructure.

2.2 Landau–Silin Theory of Fermi Liquid

In Sec. 1.3, the single-particle properties are represented by quasiparticles which are electrons dressed by interaction with the other electrons. There is a residual interaction between quasiparticles. A phenomenological formulation due to Landau represents the low-energy excited states of the interacting system in terms of a number of quasiparticles and holes excited near the Fermi surface. By including the residual interaction between the quasiparticles as phenomenological parameters, Landau's theory is able to describe all the low-energy collective excitations. Although the Fermi-liquid theory was originally designed for the uncharged ^{3}He liquid, Silin extended it to the electron gas by including the effects of the long-range Coulomb interaction. In the limit of temporally and spatially slowly varying weak external perturbations, the Fermi-liquid theory has been derived from first principles by means of infinite perturbation series. (For reviews, see Pines and Nozières, 1966; Mahan, 1990; Abrikosov, 1972).

2.2.1 Quasiparticle Interaction. A low-energy excited state of the whole many-electron system may be described by the addition of quasiparticles to the ground state and removal of quasiparticles from it, resulting in a change of distribution:

$$\delta n_k(\mathbf{r},t) = n_{\mathbf{k}}(\mathbf{r},t) - \theta(\mu - E_k), \tag{29}$$

where k is short for the wave vector $\mathbf{k}$ and spin σ. It is a matter of convenience to describe the addition and removal of quasiparticles to and from, respectively, the ground state as a difference between two quasiparticle distributions, $n_k(\mathbf{r},t)$ for the excited state and the step function $\theta(\mu - E_k)$ for the equilibrium distribution of unity when the quasiparticle energy E_k is below the chemical potential μ and zero above. It is, however, misleading to think of the ground state as made up exclusively of quasiparticles. A quasiparticle distribution is well defined only if the quasiparticle lifetime

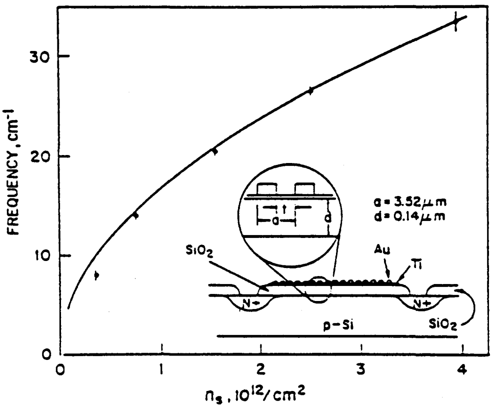

FIG. 2. Measured two-dimensional plasma spectrum of the inversion layer in Si MOSFET, compared with theory (solid curve). The inset shows the structure of the Si MOSFET with the metal gratings which produce the finite wave vectors. [From Allen *et al.* (1977), Fig. 2; reproduced with permission.]

is long, i.e., for energies close to the Fermi energy. For an external disturbance which is slowly varying in space and time (i.e., with wave vector $q \ll k_F$ and frequency $\omega \ll \mu$), we may think of a neighborhood around r with dimension larger than the Fermi wavelength as locally a homogeneous electron gas and, over a time interval around t much shorter than the period of the external disturbance, having a distribution $n_k(\mathbf{r},t)$ of quasiparticles. The excited state has a total energy in excess of the ground-state energy given by

$$\delta E[\delta n_k] = \sum_k \int d^3r\, v_{\text{ext}}(\mathbf{r},t)\delta n_k(\mathbf{r},t) + \frac{1}{2}\sum_{kk'} \int d^3r \int d^3r' \frac{e^2}{|\mathbf{r}-\mathbf{r}'|}\delta n_k(\mathbf{r},t)\delta n_{k'}(\mathbf{r}',t)$$
$$+ \sum_k \int d^3r\, E_k \delta n_k(\mathbf{r},t) + \frac{1}{2}\sum_{kk'} \int d^3r \int d^3r'\, f(k,k')\delta n_k(\mathbf{r},t)\delta n_{k'}(\mathbf{r}',t) + O(\delta n^3). \quad (30)$$

The first two terms are familiar from Sec. 1.1, being the change due to the external potential directly and via the induced electrostatic potential (the Hartree term). In the next two terms, the coefficient to the first-order term in δn_k, E_k, is the quasiparticle energy in addition to the external potential; and the coefficient to the second-order term in δn_k, $f(\mathbf{k},\mathbf{k}')$, is the interaction between quasiparticles, which is a short-range interaction on top of the long-range Hartree term. In jellium, the Coulomb interaction is spin independent but the quasiparticle interaction is made up of a spin-independent and a spin-dependent term:

$$f(\mathbf{k}\sigma,\mathbf{k}'\sigma') = f_s(\mathbf{k},\mathbf{k}') + \sigma\cdot\sigma' f_a(\mathbf{k},\mathbf{k}'). \quad (31)$$

For low-energy quasiparticles close to the Fermi surface, the interaction functions depend only on the relative angle between the wave vectors on the Fermi surface. Each function can be written in the form

$$f_s(\mathbf{k},\mathbf{k}') = N_0^{-1} \sum_{l=0}^{\infty} A_l(2l+1)P_l(\hat{k}\cdot\hat{k}'), \quad (32)$$

in terms of the exact density of states N_0 at the Fermi level and the Legendre function P_l, where $\hat{\mathbf{k}}$ is the unit vector along $\mathbf{k}$. Thus, the set of parameters A_l specifies the spin-symmetric interaction $f_s(\mathbf{k},\mathbf{k}')$, and a similarly defined set, B_l, characterizes the spin-asymmetric interaction $f_a(\mathbf{k},\mathbf{k}')$.

2.2.2 Local Equilibrium Distribution. From the energy difference between the excited state and the ground state, Eq. (30), we can extract the energy of a single quasiparticle in the presence of others by taking the first functional derivative with respect to δn_k. In addition to the Hartree energy,

$$v_{\text{ext}}(r,t) + \sum_{k'} \int d^3r' \frac{e^2}{|\mathbf{r}-\mathbf{r}'|}\delta n_k'(\mathbf{r}',t), \quad (33)$$

which is shared by all the quasiparticles in the same locality and time, the energy of a quasiparticle in state k at $\mathbf{r}$ and t is

$$\tilde{E}_k = E_k + \sum_{k'} f(k,k')\delta n_{k'}. \quad (34)$$

It is sometimes convenient to specify the quasiparticle distribution as the deviation from the local equilibrium distribution,

$$\delta\tilde{n}_k(\mathbf{r},t) = n_k(\mathrm{r},t) - \theta(\mu - \tilde{E}_k), \quad (35)$$

with both the quasiparticle energy and the Fermi energy measured from the local Hartree value, Eq. (33). Comparison of this with the quasiparticle deviation from the global equilibrium, Eq. (29), yields the relation

$$\delta\tilde{n}_k(\mathbf{r},t) = \delta n_k(\mathbf{r},t) + \delta(E_k - \mu)\sum_{k'} f(k,k')\delta n_{k'}. \quad (36)$$

The delta function at the Fermi level arises from the small deviation from the step-function distribution. If we express the quasiparticle deviation in terms of the spherical harmonics about the Fermi surface,

$$\delta n_k = \sum_{lm} \delta(E_k - \mu)\delta n_{lm} Y_{lm}(\hat{k}), \quad (37)$$

then the spin-symmetric coefficients for the local and global deviations are related by

$$\delta\tilde{n}^s_{lm} = (1 + A_l)\delta n^s_{lm}. \quad (38)$$

The spin-asymmetric coefficients are related by

$$\delta\tilde{n}^a_{lm} = (1 + B_l)\delta n^a_{lm}. \quad (39)$$

2.2.3 Specific Heat and Magnetic Susceptibility. As an example of the use of the Fermi-liquid theory, consider the low-temperature thermodynamic properties. The entropy and, hence, the thermal distribution of the quasiparticles are derived in the same way as for the noninteracting Fermions:

$$n_k(T)=1/[1+\exp\{(\tilde{E}_k-\mu)/k_BT\}]. \qquad (40)$$

The difference lies in that this relation through Eq. (34) depends implicitly on the quasiparticle distribution on the right-hand side. At low temperatures, to the lowest order in the quasiparticle deviation, the interaction term to the energy may be neglected and the specific heat is given by the noninteracting expression with the effective mass of the quasiparticle in place of the free-electron mass.

In the presence of a constant and uniform magnetic field H, each quasiparticle with spin up or down, $\sigma=\pm 1$, has an additional Zeeman energy $-\mu_B\sigma H$, where μ_B is the Bohr magneton. Then, the local deviation is

$$\delta\tilde{n}_{k\sigma}=\mu_B\sigma H\delta(E_k-\mu). \qquad (41)$$

The global deviation, which is isotropic around the Fermi surface, is, by Eq. (39), the local one, divided by the factor $1+B_0$. The susceptibility of the interacting system is, therefore, that of the noninteracting Fermions (with the effective mass in the density of states) divided by the same factor.

2.2.4 Landau–Boltzmann Equation. The response of the system to external perturbations at low frequencies and long wavelengths is governed by the kinetic theory of the quasiparticles suitably modified to take into account the interaction between the particles. The rate of increase of the quasiparticle distribution at $(k,\mathbf{r})$ is equal to the scattering rate $I_k[n]$ into the same point of the phase space:

$$\frac{\partial n_k(\mathbf{r},t)}{\partial t}+\frac{\partial n_k(\mathbf{r},t)}{\partial \mathbf{r}}\cdot\frac{\partial \tilde{E}_k(\mathbf{r},t)}{\partial \mathbf{k}}$$
$$-\frac{\partial n_k(\mathbf{r},t)}{\partial \mathbf{k}}\cdot\frac{\partial \tilde{E}_k(\mathbf{r},t)}{\partial \mathbf{r}}=I_k[n]. \qquad (42)$$

The total rate includes the partial rate with respect to time, the partial rate of change with respect to position times the rate of change of position which is equal to the velocity of the wave packet, and the partial rate with respect to the wave vector times the rate of change of the wave vector which is equal to the acceleration and, therefore, the force on the quasiparticle. The quasiparticle energy here is that of Eq. (34) plus the Hartree energy, Eq. (33). To the first order in the deviation from equilibrium distribution, the Landau–Boltzmann equation becomes

$$\frac{\partial}{\partial t}\,\delta n_k(\mathbf{r},t)+\mathbf{v}_k\cdot\frac{\partial\delta\tilde{n}_k(\mathbf{r},t)}{\partial \mathbf{r}}-\mathscr{F}(\mathbf{r},t)\cdot\mathbf{v}_k\delta(\mu-E_k)$$
$$=I_k[\delta\tilde{n}], \qquad (43)$$

where v_k is the quasiparticle velocity and $\mathscr{F}$ is the total force due to the external field and the induced field. The quasiparticle interaction is hidden in the deviation from the local equilibrium $\delta\tilde{n}_k(\mathbf{r},t)$, related to the global deviation $\delta n_k(\mathbf{r},t)$ by Eq. (36). By neglecting this difference, Eq. (43) reduces to the usual Boltzmann equation.

The terms in the linearized Landau–Boltzmann equation may be given the following physical interpretation. The external force which drives the quasiparticle distribution away from equilibrium is coupled to the states in the thin shell around the Fermi surface. The ensuing time rate of change of the distribution gives the deviation from the global equilibrium. The rate of change of the distribution due to the spatial variation must be counted in terms of the deviation from the local equilibrium. Since the scattering events can only reduce the nonequilibrium state to a local equilibrium, the collision term involves only the deviation from local equilibrium. The difference between the global and local distributions is caused by the quasiparticle interaction. Thus, for a particular phenomenon, by examining the relative importance of the terms involving the two distributions, one can determine whether the residual quasiparticle interaction plays a role. For a low-frequency phenomenon such that $\omega \ll qv_F$, the first term on the left of the linearized equation is unimportant compared with the second term. The equation becomes just like the Boltzmann equation with the entire equation involving only the "renormalized" distribution $\delta\tilde{n}_k(r,t)$. Then the residual interaction plays no explicit role and cannot be measured. For a high-frequency phenomenon such that $\omega \gg qv_F$, because the time variation term involves $\delta n_k(r,t)$ and the collision term involves $\delta\tilde{n}_k(r,t)$, in principle the quasiparticle interaction is explicitly present. However, the theory is generally insufficiently accurate to distinguish between relaxation time with and without the residual interaction. It is in practice not possible to deduce the interaction from such high-frequency phenomena. Only from phenomena in which both the temporally and spatially varying terms are comparable can the Landau interaction parameters be deduced.

2.2.5 Density and Current Response. Density and current responses to the longitudinal electric field yield information on the transport properties. The collision term, which by restoring the local equilibrium plays an important role in transport, may be written in terms of relaxation times τ_l in the form

$$I_k[\delta\tilde{n}] = -\sum_{lm} \frac{\delta\tilde{n}_{lm}}{\tau_l}\,\delta(\mu - E_k)Y_{lm}(\hat{k}), \tag{44}$$

in terms of the spherical harmonic components, Eq. (37). In general, the relaxation-time expression is an approximation. In some cases, such as impurity scattering, this is exact. The $l=0$ term is absent, so that the spherical average of the linearized Landau–Boltzmann equation leads to the equation of continuity with the density and current density given by

$$\begin{aligned} \rho(q,\omega) &= -e\sum_k \delta n_k, \\ \mathbf{j}(q,\omega) &= -e\sum_k \mathbf{v}_k\delta\tilde{n}_k. \end{aligned} \tag{45}$$

Note the global and local distributions which go, respectively, into the density and current density expressions.

The density and current responses may be obtained by solving the Landau–Boltzmann equation in the spherical harmonic components. The proper polarization part (the density response to the total field) is given in terms of the spherically symmetric deviation of the quasiparticle distribution, yielding

$$\tilde{\chi}(q,0) = -e^2N_0/(1+A_0). \tag{46}$$

The modification of the Landau parameter is the counterpart of the spin susceptibility. (See Sec. 2.2.3.)

The current response to a uniform static field, which creates an $l=1$ type Fermi surface distortion, leads to the conductivity

$$\sigma(0,\omega) = ne^2\tau^*/(1-i\omega\tau^*), \tag{47}$$

where the relaxation time is given by

$$\tau^* = \tau_1/(1+A_1). \tag{48}$$

This is an example in which the Landau interaction parameter A_1 does enter into a transport coefficient but in which the inaccessibility of the unrenormalized relaxation time τ_1 makes it impossible to determine the Landau parameter.

2.2.6 Helicon and Magnetoplasma Modes. In an uncharged Fermi liquid, the Landau–Boltzmann equation leads to a variety of collective oscillations about the Fermi surface. However, in the electron gas, because of the long-range Coulomb interaction, these are masked by the strong screening due to the high plasma frequency or by the lattice vibrations. For example, zero sound in the Fermi liquid, corresponding to a spherical distortion of the Fermi surface, is excited by a compressional wave. In an electronic solid, this is inextricably bound with the lattice wave. A clean separation of the quasi-interaction effects is not possible (Sham, 1969). In this section we consider the collective excitations as first-order responses to a weak electromagnetic field in the presence of a strong uniform static magnetic field $\mathbf{H}_0$, which opens up new possibilities for collective modes. (See Kaner and Skobov, 1966; Platzman and Wolff, 1973.) The screening current which limits the electromagnetic wave to the skin depth is somewhat curtailed by the cyclotron motion of the electrons in the strong magnetic field, giving rise to windows for the propagation of electromagnetic waves along the magnetic field direction.

To the left-hand side of the Landau–Boltzmann equation, Eq. (43), a term corresponding to the Lorentz force due to the static field is added:

$$-\frac{e}{c}\,\mathbf{v}_k\times\mathbf{H}_0\cdot\frac{\partial}{\partial\mathbf{k}}\,\delta\tilde{n}_k(r,t). \tag{49}$$

The force $\mathscr{F}$ is now the Lorentz force due to the weak electromagnetic field.

Consider the propagation of a right-handed circularly polarized electromagnetic wave along $\mathbf{H}_0$. In the long-wavelength limit, the dispersion relation from the wave equation is

$$\omega^2\epsilon(0,\omega) = q^2c^2. \tag{50}$$

From the Landau–Boltzmann equation, the dielectric response is of the form

$$\epsilon(0,\omega)=1-\frac{\omega_p^2}{\omega(\omega-\omega_c-i/\tau)}, \tag{51}$$

where $\omega=eH_0/m^*c$ is the cyclotron frequency involving the effective mass of the quasiparticle at the Fermi surface.

At low frequency and weak magnetic field, with $\omega \ll \omega_c \ll 1/\tau$, the relaxation-time term dominates the dielectric function and the wave vector is complex,

$$q\equiv q'+i/\delta=(i\omega\tau)^{1/2}\omega_p/c. \tag{52}$$

An applied low-frequency electromagnetic wave is confined to a skin depth δ of the surface of the metal. For normal metals, the plasma frequency is much larger than the cyclotron frequency even for the largest magnetic field strength attainable nowadays. Thus, the uniform cyclotron resonance is screened by the plasma mode. The only possible cyclotron resonance is the Azbel'–Kaner resonance, in which the electron cyclotron radius is much larger than the skin depth and in which the resonance occurs when the applied electromagnetic wave is in synchronization with the electron passing through the skin depth. At high frequencies, $\omega \gg \omega_c, 1/\tau$, the dispersion relation for the transverse plasma waves is

$$\omega^2=\omega_p^2+q^2c^2. \tag{53}$$

For a high magnetic field so that the cyclotron term dominates in the dielectric function, $\omega_c\tau \gg 1$, there is a low-frequency mode:

$$\omega=q^2c^2\omega_c/\omega_p^2. \tag{54}$$

This is known as the helicon mode. The frequency is lower than ω_c. The propagation of the right-handed circularly polarized mode is made possible by the static magnetic field. The experimental arrangement is for the magnetic field to be normal to the metal surface and the applied electromagnetic wave incident normal to the surface. The helicon propagation provides an ac method to measure the Hall coefficient $R=1/nec$ since the coefficient in Eq. (54) is

$$\omega_c/\omega_p^2=RH_0/4\pi, \tag{55}$$

and the sense of the circular polarization is determined by the sign of the carrier charge. The Landau interaction term is unimportant in the helicon since it is at low frequency and high magnetic field.

In contrast to the helicon wave whose frequency is much lower than the cyclotron frequency, there are magnetoplasma modes with frequencies of the same order as the cyclotron frequency. We follow Mermin and Cheng (1968) in obtaining these modes from the Landau-Boltzmann equation, Eq. (43), which is rewritten in the form

$$(L-\omega)\delta n_k=-ie\mathscr{E}_{\text{ext}}\cdot\mathbf{v}_k\delta(\mu-E_k). \tag{56}$$

The operator L includes from Eq. (43) the spatial variation term, the cyclotron motion term, the collision term, and the induced driving-force term. The eigenfrequencies of the operator L are the frequencies of the magnetoplasma modes. We apply perturbation theory for small q in terms of the spherical harmonic components δn_{lm} of the quasiparticle distribution about $\mathbf{H}_0$ as the polar axis. From the expressions in Eq. (45), the induced current involves only the $l=1$ components and the induced density fluctuation only the $l=0$ component. Thus, the $l=0,1$ deformation of the Fermi surface would induce a strong screening current masking the collective modes. The magnetoplasma modes of interest have components with $l>1$. Consider separate cases of different orientations of the wave vector $\mathbf{q}$ relative to the magnetic field $\mathbf{H}_0$.

Case I, $\mathbf{q}\perp\mathbf{H}_0$: The frequency is given by

$$\omega=a_l m\omega_c+\frac{q^2v_{\mathrm{F}}^2}{\omega_c}\sum_{l'>1,m'}\frac{a_l|C_{l'm'}^{lm}|^2a_{l'}}{ma_l-m'a_{l'}}. \tag{57}$$

The Landau interaction parameter in $a_l = 1 + A_l$ splits the degeneracy in the modes at zero wave vector. These magnetoplasma modes are, therefore, strongly influenced by the quasiparticle interaction. The coefficient $C_{l'm'}^{lm}$ is the $3j$ symbol, i.e., the matrix element between the two spherical harmonics of $Y_{1,0}$ if the electric polarization vector is parallel to $\mathbf{H}_0$ and of $Y_{1,\pm1}$ if the electric vector is normal to $\mathbf{H}_0$. The former mode is known as the ordinary wave and the latter the extraordinary wave.

Case II, $\mathbf{q}\|\mathbf{H}_0$: The normal-mode frequencies are given again only for $l>1$ by

$$\omega=a_l m\omega_c+\frac{q^2v_{\mathrm{F}}^2}{\omega_c}\sum_{l'>1}\frac{a_l|C_{l'm}^{lm}|^2a_{l'}}{m(a_l-a_{l'})}, \tag{58}$$

where the coefficient $C_{l'm}^{lm}$ is the matrix element between the two spherical harmonics of $Y_{1,0}$. The contrast with Case I is that the sum here in the q^2 term is over the same m components. Hence, if the interaction parameters A_l are neglected, the q^2 term blows up. For noninteracting particles, the modes form a continuum with the dispersion

$$\omega=\omega_c+\mathbf{q}\cdot\mathbf{v}_k. \tag{59}$$

In contrast to Case I, the discrete modes exist only by virtue of the quasiparticle interaction.

Observations of both types of modes have been used to deduce the Landau parameters A_l for alkali metals.

2.2.7 Spin Waves. The spin waves considered here are the spin counterpart of the magnetoplasma waves in jellium, unlike the spin waves in magnetic materials. We follow the treatment of Platzman and Wolff (1967). A review of the experimental situation is given by Dunifer *et al.* (1974).

In order to include spin flip, the quasiparticle distribution $n_k(\mathbf{r},t)$ is now considered as a 2×2 density matrix. Its trace gives the density distribution considered for the magnetoplasma waves. Trace with the Pauli matrices yields the spin-density distribution $\mathbf{m}_\mathbf{k}(r,t)$. The corresponding Landau–Boltzmann equation is derived from the kinetic theory in the same way as above. Through the asymmetric part of the quasiparticle interaction in Eq. (31), the spin density creates an internal magnetic field:

$$\mathbf{h}=\sum_{\mathbf{k}'} f_a(\mathbf{k},\mathbf{k}')\mathbf{m}_{\mathbf{k}'} \tag{60}$$

which also relates this deviation of the spin distribution from global equilibrium to the deviation from local equilibrium, $\tilde{\mathbf{m}}_\mathbf{k}(r,t)$. To the driving force is now added the precession of the electron spin acted upon by the magnetic field.

If the magnetic field is taken to be in the z direction, the longitudinal component $m_{z,\mathbf{k}}(r,t)$ yields the spin susceptibility as above. The transverse component

$$m_+=m_x+im_y \tag{61}$$

yields information on spin precession. The Landau–Boltzmann equation for m_+ may be solved in terms of the spherical harmonic components of its $\mathbf{k}$ orientation on the Fermi surface, in much the same way as the density fluctuation. The crucial difference is that the spin-density distribution creates no screening current. Thus, the dominant spherical harmonic components are $l=0,1$, which, when summed over $\mathbf{k}$ space, yield, respectively, the magnetization and magnetization current:

$$M_+=\sum_\mathbf{k} m_{+,\mathbf{k}}, \qquad \mathbf{J}_+=\sum_\mathbf{k} v_\mathbf{k} m_{+,\mathbf{k}}. \tag{62}$$

If the $l>1$ components of the spin distribution are neglected, the Landau–Boltzmann equation reduces to a closed set of equations for the magnetization and current:

$$\begin{aligned} &-i(\omega-\omega_s)M_+ + i\mathbf{q}\cdot\mathbf{J}_+ = i2\mu_\mathrm{B}M_0H_+,\\ &-i\left(\omega-\omega_s\frac{b_1}{b_0}\right)\mathbf{J}_+ + i\mathbf{q}\,\frac{1}{3}\,v_\mathrm{F}^2 b_0 b_1 M_+\\ &\qquad + \frac{1}{\tau}\,\mathbf{J}_+ + b_1\omega_c\times\mathbf{J}_+ = 0, \end{aligned} \tag{63}$$

where $b_l = 1 + B_l$ involves the spin part of the quasiparticle interaction, $\omega_s = 2\mu_\mathrm{B}H_0$ is the spin precession frequency in terms of the Bohr magneton μ_B, H_+ is the rf field normal to the static magnetic field, and ω_c is the cyclotron frequency due to the static field, directed along the same direction.

The terms in the two equations may be interpreted as follows. The terms $-i\omega M_+ + iq\cdot\mathbf{J}_+$ equated to zero would give the equation of continuity. However, the magnetization M_+ precesses about $\mathbf{H}_0$ with frequency ω_s and is driven by the rf field $\mathbf{H}_+$. The first equation shows no relaxation of the magnetization M_+ because the scatterings due to quasiparticle interaction (characterized by the relaxation time τ) which appear in the second equation do not change the spin component. The equation could be supplemented by a term M_+/T on the left if spin-flip scatterings with a characteristic time T are included. The first two terms in the second equation have a similar physical meaning, with the precessing spin current driven by the gradient of the magnetization. The term $\mathbf{J}_+/\tau$ is the current relaxation due to non-spin-flip particle scattering. The last term, $\omega_c \times \mathbf{J}_+$, represents the cyclotron motion of the current. The factors of the interaction parameters b_l come from the terms in the Landau–Boltzmann equation which appear as deviation from the local equilibrium.

An approximate solution of these equations yields the transverse spin response:

$$\chi_+(\mathbf{q},\omega+i0)\equiv\frac{M_+}{H_+}=\frac{\chi_+(0,0)\omega_s}{P(\mathbf{q},\omega+i0)}\,. \tag{64}$$

The denominator has the form

$$P(\mathbf{q},\omega+i0)=\omega_s+\alpha q^2-\omega-i/T, \tag{65}$$

where the coefficient of the dispersion is

$$\alpha=b_0b_1\,\frac{v_\mathrm{F}^2}{3}\left[\frac{\cos^2\Delta}{\tilde{\omega}_s+i/\tau}+\frac{(\sin^2\Delta)(\tilde{\omega}_s+i/\tau)}{(\tilde{\omega}_s+i/\tau)^2-b_1^2\omega_c^2}\right], \tag{66}$$

Δ being the angle between $\mathbf{q}$ and $\mathbf{H}_0$, and $\tilde{\omega}_s = \omega_s(1 - b_1/b_0)$.

In the strong-scattering limit, $\tilde{\omega}_s\tau\ll 1$, the denominator assumes the form

$$P(\mathbf{q},\omega+i0)=\omega_s-iD_sq^2-\omega-i/T, \tag{67}$$

with the diffusion coefficient

$$D_s=b_0b_1\,\frac{v_\mathrm{F}^2}{3}\left[\cos^2\Delta+\frac{\sin^2\Delta}{1+(b_1\omega_c\tau)^2}\right]. \tag{68}$$

The diffusion coefficient and the spin-flip relaxation time give the Dysonian line shape (Dunifer *et al.*, 1974) for the conduction electron spin resonance at the frequency ω_s, unrenormalized by the interaction parameters.

In the weak scattering limit $\tilde{\omega}_s\tau\gg 1$, the coefficient α is real and the poles of χ_+ give the spin waves with the dispersion $\omega=\omega_s+\alpha q^2$. These spin waves cannot occur without the presence of the interaction parameters B_l. Figure 3(a) shows the frequency dispersion for three directions of the wave vector $\mathbf{q}$ relative to the static magnetic field, normal, parallel, and at the critical angle Δ_c where the two terms in α cancel each other. Only the $\sin^2\Delta$ term in α depends on the field strength H_0. The reason is that only the component of the magnetization current parallel to the wave vector influences the magnetization M_+ and that the cyclotron motion of the current vector is about the static magnetic field. If $\mathbf{q}$ is parallel to $\mathbf{H}_0$, the cyclotron motion of the current vector has no influence on the magnetization. If $\mathbf{q}$ is perpendicular to $\mathbf{H}_0$, $\mathbf{J}_+\cdot\mathbf{q}$ precesses with frequency $b_1\omega_c$, thus modifying the spin current precession with frequency $\tilde{\omega}_s$. Figure 3(b) shows typical spin-wave signals in relation to the conduction electron spin resonance line. The wave vector $q=n\pi/L$ is determined by the standing waves confined by the thickness L of a thin-film sample.

3. COLLECTIVE EFFECTS ASSOCIATED WITH INTERBAND TRANSITIONS

In a solid, the electronic structure consists of a number of energy bands. In the ground state of the system, a completely filled band is known as a valence band and a partially filled or empty band is known as a conduction band. Optical processes which involve interband transitions may incur collective effects through the particle interaction. These phenomena are clearly outside the scope of the jellium model even though the model has been shown in the last section to be extremely useful in illustrating the collective effects of electrons in a single conductor band.

In an insulator or semiconductor, the conduction bands are all empty. Light can excite the transition of an electron to a conduction band leaving behind a hole in the valence band. The electron-hole pair binds into a hydrogen-atom-like entity, known as an exciton, which survives from a nanosecond to a microsecond before the electron deexcites giving out a photon (see EXCITONS; Rashba and Sturge, 1982). With a sufficiently intense light source, the density of excitons can increase so much that Keldysh predicted that they would form a new metallic phase, known as the electron-hole liquid (see Rice, 1977; Hansel *et al.*, 1977; Jeffries and Keldysh, 1983), with a number of new collective phenomena, which we shall examine in Sec. 3.1. The phase diagram is like that of the gas–liquid transition. The collective modes include the two branches of the plasma oscillations and the analog of the helicon wave in a neutral plasma known as the Alfvén wave which has a different frequency dispersion from the helicon and which can now have either circular polarization about the magnetic field axis.

In a metal or a doped semiconductor, with one or more partially filled conduction bands, the appearance of a valence hole following the interband transition causes the conduction electrons to react (Sec. 3.2). If the valence hole is localized, the reaction of the conduc-

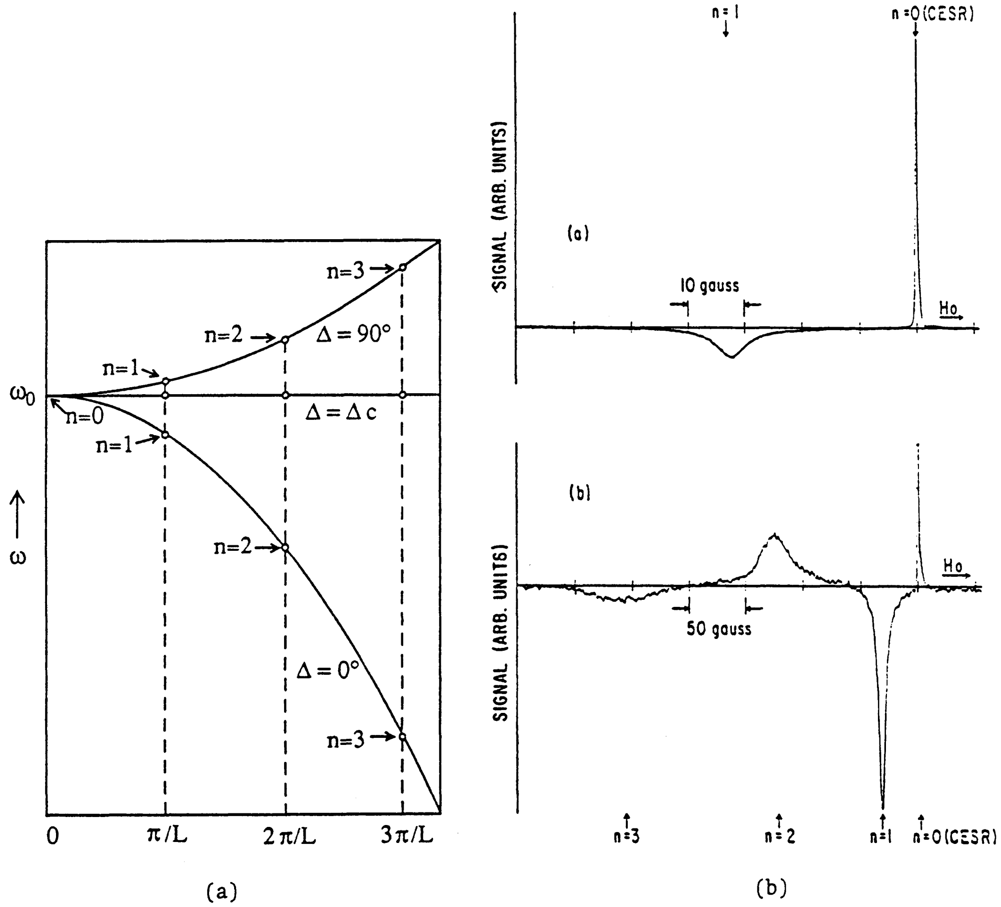

FIG. 3. (a) Schematic spin-wave dispersion for three values of the angle between the wave vector and the magnetic field normal to a metal film of thickness L. (b) Typical spin wave signals for $\Delta=90°$ sweeping the static magnetic field, showing the conduction electron spin resonance line and the spin waves at three wave vectors $q=n\pi/L$, L being the thickness of the Na film. "CESR" stands for conduction electron spin resonance. [From Dunifer *et al.* (1974), Figs. 1 and 6(b); reproduced with permission.]

tion electrons creates a large number of density waves around the Fermi surface, leading to a threshold singularity at the Fermi level in the optical spectrum. In x-ray emission, the hole is deep in the core of an atom and is well localized. Mahan predicted that an edge singularity should be seen. The relevance of the edge singularity in the experimental x-ray emission and absorption spectra and in x-ray photoemission is surveyed below. The strength of the edge singularity in the optical spectrum of a doped semiconductor quantum well varies with samples, seemingly dependent on whether the hole is localized or is mobile with a finite mass. The theory of the finite mass effect of the valence hole on the edge singularity is discussed below.

3.1 Excitons and the Electron-Hole Liquid

In a semiconductor, an electron and a hole created by the absorption of a photon relax, respectively, to the bottom of the conduction band and the top of the valence band. There, they are bound by the mutual Coulomb attraction mediated by the dielectric constant of the medium to form a hydrogen-like entity called the exciton. As Keldysh (Jeffries and Keldysh, 1983) puts it, the exciton resembles more the positronium than the hydrogen atom in that after a time the electron-hole pair recombines emitting a photon. Two excitons have the tendency to form a biexciton molecule. A large number of excitons may form a liquid phase of electrons and holes, called the elec-

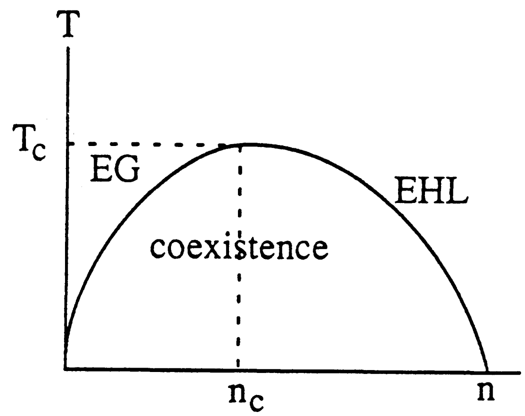

FIG. 4. Schematic phase diagram of the electron-hole liquid (EHL) and exciton gas (EG) in the temperature T and density n plane.

tron-hole liquid (EHL), in which the individual electron and hole are no longer bound to each other. Such a system may be considered in quasiequilibrium because the relaxation times of the carriers to the conduction-band valley and the valence-band peak, of the order of picoseconds, are shorter than the recombination time of the electron-hole pair, which is of the order of nanoseconds in a direct bandgap semiconductor such GaAs and of the order of microseconds in an indirect gap material such as Ge or Si.

3.1.1 The Phase Diagram. Figure 4 shows a schematic phase diagram of equal numbers of electrons and holes as a function of density n and temperature T. At low density and high temperature they form a gas of excitons (EG), and at high density and low temperature they form an electron-hole liquid (EHL). The phase transition between the two phases is first order except, of course, at the critical point (n_c, T_c). Thus, there is a coexistence region, in which the liquid phase is composed of droplets. The calculation of the phase boundary requires a careful comparison of the total energy of the exciton gas with that of the two-component plasma, including the kinetic energy, exchange, and correlation, as well as the dependence on the bulk electronic structure of conduction valleys and the valence band peak. (See Rice, 1977; Vashishta, *et al.*, 1983.) Strain in some instances favors the formation of the liquid phase and has been utilized in Ge to produce very large drops (see Wolfe and Jeffries, 1983).

The quasiequilibrium properties of EG and EHL are observed by the luminescence spectra due to electron-hole recombination after the semiconductor is excited by a light source. (See Westervelt, 1983.) Figure 5 is an excellent example. The top panel shows the spectrum of the exciton gas phase. Since Ge has an indirect band gap, recombination of an electron-hole pair must be accompanied by the emission of a phonon to conserve momentum. In this case, the phonon assistance is rendered either by the longitudinal acoustic branch (the luminescence peak labeled LA) or the transverse optical branch (the peak labeled TO). With the increase of the intensity of the excitation light, at a well-defined threshold, a broad peak marked EHD appears below the free-exciton (LA) peak, marked FE in the lower panel of Fig. 5. The broad peak is due to the recombinations in the electron-hole drop, with the width of the peak being a measure of

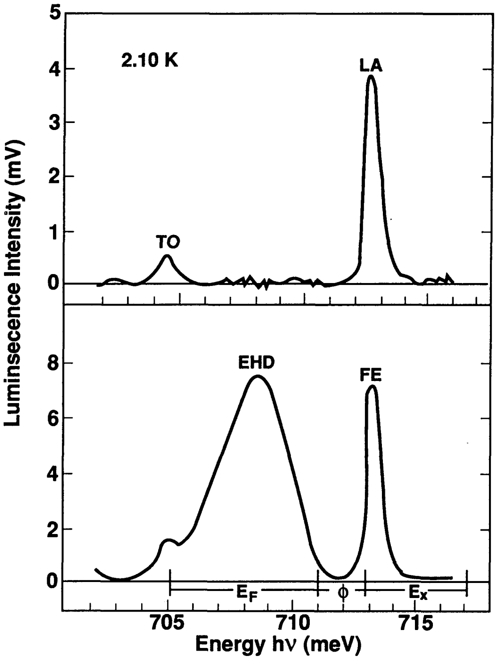

FIG. 5. Upper panel: Luminescence spectrum of the free excitons (FE) in Ge at 2 K. The phonon replicas are marked by LA and TO. Lower panel: Luminescence of the electron-hole drop (EHD) when the excitation light power is twice as high as above. The LA replicas are marked FE and EHD. [Taken from Westervelt, (1977), Figs. 1.3 and 1.4; also published in Jeffries and Keldysh (1983), p. 196, Fig. 6; reproduced with permission.]

the sum of the Fermi energies of the electrons and holes in the liquid.

3.1.2 Plasma Oscillations and Alfvén Waves in EHL. The electron-hole liquid is a two-component neutral plasma, analogous to jellium with a mobile positive background. The plasma oscillation, therefore, has two branches, a high-frequency branch with the long-wave limit frequency $\omega_p^2 = 4\pi ne^2/m$ with m being the reduced mass of the electron-hole pair, and an acoustic branch with frequency dependent linearly on the wave vector. The energy scale is, however, different from that in the simple metals. Since the Fermi energy in an EHL is of the order 10 meV, the collective excitation is in the infrared range. (See Markiewicz and Timusk, 1983.)

The electrons in jellium form a light-mass mobile component, the restriction of whose motion by the magnetic field allows the propagation of helicon waves. The counterpart in EHL, in which the two components of the plasma have comparable masses and are both mobile, is the Alfvén wave. In the wave equation (50), the dielectric function for the positively circularly polarized wave is changed from Eq. (51) to

$$\epsilon(0,\omega) = 1 - \sum_{j=e,h} \frac{\omega_{pj}^2}{\omega(\omega - \omega_{cj} - i/\tau_j)}, \qquad (69)$$

where the index $j=e,h$ denotes the quantity for the electron or hole component. The hole cyclotron frequency is $\omega_{ch} = eH_0/cm_h$, whereas the electron one is $\omega_{ce} = -eH_0/\mathrm{c}m_e$. In the high frequency regime we retrieve the plasma oscillation. In the high-field regime, with $\omega_{cj}\tau_j \gg 1$, the cancellation of the first-order term in ω on the left-hand side of the wave equation (50) leads to the low-frequency mode being linear instead of quadratic in the wave vector q:

$$\omega^2 = q^2 H_0^2/4\pi(m_e + m_h). \qquad (70)$$

The other difference between the helicon wave and the Alfvén wave lies in the sense of the circular polarization, which in the former can be in only one sense dependent on the charge of the active component of the plasma but which can be both right- and left-handed in the neutral EHL.

Other interesting phenomena in the light scattering of the EHD arise out of the finite size of the droplets. See Jeffries and Keldysh (1983) for further details.

3.2 Dynamical Response of the Fermi Sea to Interband Transition

During an interband optical transition, the presence of a hole in a valence band causes the Fermi surface to respond dynamically. The most dramatic manifestation of this is in the power-law singularities in energy at the Fermi level in the emission and absorption spectra and in the asymmetric line shape of the photoemission spectrum. (For reviews see Mahan, 1990; Almbladh and Hedin, 1983.) Here we shall examine the effects which occur in the soft x-ray range of core-level transitions in metals and in the optical range of valence- to conduction-band transitions in semiconductor quantum wells. The basic cause is the sudden perturbation by the appearance or disappearance of the core hole, which provides a charged-impurity-like potential (Hopfield, 1969). That the initial and final states of the system are under the influence of different potentials, and thus differ by more than a single-particle transition, resembles the Franck–Condon effect. The edge singularity arises out of the high density of states of the electron-hole pair density fluctuations at the Fermi level. We first explain the principle of the phenomenon before detailing the experimental observations in the x-ray range in metals and in the optical range in modulation-doped semiconductor quantum wells.

3.2.1 Principle of the Mahan or MND Effect. The effect of the dynamical response of the Fermi sea to the core hole is known as the Mahan or MND effect, in honor of the originator, Mahan (1967), and of Nozières and de Dominicis (1969), who gave a more complete theory, thereby elucidating the subtle competing factors which determine the power-law indices. Here, we shall first give a qualitative description of their theories and then use the boson model to supply a somewhat more quantitative understanding of the phenomenon. In this subsection, the core level is taken to have no dispersion or, equivalently, the core hole is assumed to be localized.

In the state of the system with a core hole, a conduction electron in the Fermi sea and the hole attract each other with a Coulomb interaction screened by the other electrons. The change of the electron state is limited by the available phase space because of the presence of the Fermi sea. In addition, the tendency to excitonic formation is countered by the ability of the core hole to excite electron and (conduction-band) hole pairs in the Fermi sea. The net result is the elimination of the formation of an exciton in the metal above the Fermi level, placing sufficient oscillator strength in the neighborhood of the Fermi level to cause the absorption or emission spectrum to have a power-law singularity, $|E-E_F|^{-\alpha}$. So far, the consideration has not included the fact that all the conduction-band states are changed in the presence of the core potential. When this is taken into account, the overlap matrix element of the many-electron state in the presence of the core hole and the state in its absence is vanishingly small as the number of electrons tends to infinity. This effect is known as the infrared catastrophe (Anderson, 1967). The orthogonality between the two many-electron states tends to lower the oscillator strength for the transition and to reduce the power-law singularity. The existence of a Fermi edge singularity is the delicate net result of these three competing factors (electron-core-hole attraction, production of conduction-band electron-hole pairs, and the infrared catastrophe).

The appearance or disappearance of the core hole between the initial and final states creates a sudden perturbation in time. Its effect is very different from a perturbation which is adiabatically turned on, even if the strength of the sudden perturbation is weak. The response of the system to a small adiabatic perturbation is adequately described by the linear response (Sec. 1.2). To illustrate the difference between the two types of perturbation, consider the example of a harmonic oscillator of frequency ω in the ground state at time $t=0$ subject then to a weak force

$$F_0(1-e^{-t/\tau}), \tag{71}$$

with a time constant τ. What is the probability of finding the oscillator in the first excited state after a long time? In the adiabatic limit $(\tau \gg 1/\omega)$, the probability is $|F_0 x_{10}/\omega^2\tau|^2$, where x_{10} is the matrix element of the displacement. The probability of exciting the oscillator from its ground state goes to zero with the rate of the adiabatic switching on. In the sudden limit $(\tau \ll 1/\omega)$, the probability is $|F_0 x_{10}/\omega|^2$. A sudden perturbation of the same strength as the adiabatic perturbation can make the system leave its ground state.

Now consider a boson model more relevant to the optical transition problem and justify its relevance to the Fermi system after we have seen the results of the boson model. Suppose that the system in state a with energy E_a is uncoupled to the bosons but in state b with energy E_b is coupled linearly to the bosons of frequency ω_q by the Hamiltonian

$$H_1 = \sum_q V_q(b_q + b_q^\dagger), \tag{72}$$

in terms of the linear coefficient V_q and boson creation and destruction operators. If H_r denotes the light excitation with frequency ω, the absorption spectrum $P(\omega)$ for the transition from state a to state b is given by the Fourier transform of the correlation function,

$$P(t) = \langle H_r(t) H_r(0) \rangle_a, \tag{73}$$

the ensemble average being over the bosons with the system in state a. For simplicity, the excitation matrix element is taken to be independent of the boson states; then $P(t)$ is proportional to the correlation function

$$S(t) = \langle e^{iH_b t} e^{-i(H_b + H_1)t} \rangle, \tag{74}$$

where H_b is the boson Hamiltonian and $\hbar = 1$. The matrix element with respect to the state a has been evaluated [a factor of $\exp\{i(E_b - E_a)t\}$ is understood] and only the boson ensemble average remains. This correlation function is easily evaluated in the same way as the derivation of the Debye–Waller factor for harmonic phonons (Maradudin *et al.*, 1963):

$$S(t) = e^{-i\delta E t + g(t) - 2M}, \tag{75}$$

where

$$\delta E = -\sum_q |V_q|^2/\omega_q \tag{76}$$

represents the change in the transition energy to $E_t = E_b - E_a + \delta E$ due to the reaction of the bosons,

$$2M = \sum_q (2n_q + 1)|V_q|^2/\omega_q^2 \tag{77}$$

is the Debye–Waller factor, n_q being the boson distribution, and

$$g(t)=\sum_q [n_q e^{i\omega_q t}+(n_q+1)e^{-i\omega_q t}]|V_q|^2/\omega_q^2 \tag{78}$$

is essentially the boson propagator. Note that if the boson perturbation H_1 were turned on adiabatically the terms $g(t)$ and $2M$ would be absent.

The spectrum $S(\omega)$ has two types of behavior at threshold depending on whether M is finite or not. The Debye–Waller factor is from Eq. (77) expressed in frequency terms at zero temperature as

$$2M=\int_0^\infty d\omega N(\omega)|V(\omega)|^2/\omega^2, \tag{79}$$

where $N(\omega)$ is the boson density of states. The Debye–Waller factor is finite, e.g., for phonons with deformation-potential coupling

$$N(\omega)\propto\omega^2;\quad V(\omega)\propto\omega^{1/2}. \tag{80}$$

If the coupling V is weak, a series expansion of the spectrum yields

$$S(\omega)=\delta(\omega-E_t)e^{-2M}+e^{-2M}\sum_q\left|\frac{V_q}{\omega_q}\right|^2 \times\delta(\omega-E_t-\omega_q)+\cdots. \tag{81}$$

The first term is the no-phonon line and the second term is the one-phonon side band, etc. The reduction in strength by the Debye–Waller factor can be understood as the overlap between the initial and final states reduced by the change of the final state by the presence of the phonons.

The Debye–Waller factor is infinite for the example of phonons with piezoelectric coupling,

$$V(\omega)\propto\omega^{-1/2}, \tag{82}$$

or for the more relevant example of the density waves in a fermion system:

$$N(\omega)\propto\omega;\quad V(\omega)\propto\omega^0. \tag{83}$$

In this case, the no-phonon line disappears and the spectrum has a power-law singularity at the threshold.

This simple boson model has a diverse range of applications: infrared spectrum of electronic transitions in impurities coupled to phonons in insulators (see the review by Duke and Mahan, 1965), Doppler shift in the Mössbauer effect (Maradudin, 1966), and paraelectric resonance (Sander and Shore, 1970). We shall now adapt it to the fermion system in place of the bosons.

We follow the treatment of Schotte and Schotte (1969) as being the most physical approach. The sudden introduction of a core-hole potential creates electron-hole pairs in the vicinity of the Fermi surface or density waves:

$$b_k^\dagger=c_{k_>}^\dagger c_{k_<}. \tag{84}$$

The right-hand side depicts the creation of an electron outside the Fermi surface at $k_>$ and destruction of one inside at $k_<$. The creation of the density wave at $k=(k_>,k_<)$ is approximately a boson operator. The density wave has energy

$$\omega_k=\epsilon_{k_>}-\epsilon_{k_<}. \tag{85}$$

The density of states of the density waves is given by the convolution of the electron-hole pair densities of states, which for small energy difference from the Fermi level yields

$$N(\omega)=\omega N_0^2, \tag{86}$$

where N_0 is the electron density of states at the Fermi level. The linear ω dependence of the boson density of states and the coupling constant V from the Coulomb interaction between the core electron and the density waves means that the Debye–Waller factor diverges. The evaluation of the two exponents in Eq. (75) together gives for large t

$$g(t)-2M=-(\delta/\pi)^2\ln|\omega_F t|, \tag{87}$$

where N_0V has been replaced by the s-wave phase shift and ω_F is the cut-off energy for the density waves of the same order as the Fermi energy.

The correlation function $S(t)$ may be thought of as that of the core hole after photoemission, which after a long time behaves like

$$S(t)\sim e^{-iE_t t}(\omega_F t)^{-(\delta/\pi)^2}. \tag{88}$$

The photoemission spectrum near the threshold E_t is, therefore,

$$S(\omega)=\left(\frac{\omega-E_t}{\omega_F}\right)^{(\delta/\pi)^2-1}. \tag{89}$$

Physically, this is what has happened. Photoemission from the core level leaves behind a core hole which provides a sudden attractive potential creating density waves in the conduction-band Fermi sea. The number of such waves is large $[N(\omega)\sim\omega]$, compared with the low-energy phonons $[N(\omega)\sim\omega^2]$. The sheer number of density waves makes the excitation of the system out of the ground state very probable, so much so that the overlap of the current state with the old ground state is zero. Hence, the "zero-phonon" line disappears. The effect of the sudden perturbation of such a large number of density waves persists a long time, like the grin of the Cheshire cat, causing the power-law divergence of the spectrum at the threshold.

The x-ray absorption (or emission) spectrum involves the transition of a core electron to the conduction band and is given by

$$P(t)=\sum_{kk'}\langle e^{iH_b t}c_k e^{-i(H_b+H_1)t}c^\dagger_{k'}\rangle, \tag{90}$$

where, as in the boson model, H_b is the Hamiltonian for the density waves and H_1 couples the density waves to the core hole. By comparing the commutation relation with that for the density wave b_k, one obtains the boson approximation for the fermion operator:

$$\sum_k c_k\sim\exp\left\{\sum_k (b^\dagger_k-b_k)/\omega_k N_0\right\}. \tag{91}$$

The difference between the absorption spectrum P here and the photoemission spectrum S above is that now after the introduction of the sudden perturbation a conduction electron is added in the same locality as the core hole, as given by the operator in Eq. (91). Since this is just a linear displacement operator of the harmonic oscillator, the correlation function $P(t)$ can be evaluated in a similar way to that for $S(t)$. The threshold behavior is

$$P(\omega)\sim\left(\frac{\omega-E_t}{\omega_F}\right)^{[(\delta/\pi)-1]^2-1}. \tag{92}$$

The presence of the conduction electron changes the phase shift δ by $-\pi$.

3.2.2 Experimental X-ray Spectra in Metals.

Transitions from deep core levels in metals fit the localized core-hole model discussed in the previous section very well. In x-ray photoemission (XPS), we expect the spectrum of a deep core level, which would be a sharp line in the independent-particle model, to become a broadened asymmetric line when the power-law singularity in Eq. (89) is rounded off by phonons and by decay through excitation of the Fermi sea. Such asymmetry is indeed observed in the XPS of Na, Mg, and Al from 1*s*, 2*p* levels (Citrin *et al.*, 1977). Quantitative fit of the power law yields indices in good agreement with theory.

The x-ray absorption and emission spectra, involving electron transitions from the core level to states near the Fermi level, should also have power-law singularities with the same index, different from that of the XPS. This is confirmed by observation in Na, Mg, and Al. (See a comprehensive analysis of the power-law fit to the experimental data by Citrin *et al.*, 1979.) Figure 6 clearly shows that in the emission and absorption spectra of Na, both peak structures near the Fermi level are observed. One exception is Li, where the broadening by phonons dominates over the dynamical response of the Fermi sea (Mahan, 1990).

3.2.3 Optical Spectra in Doped Semiconductor Quantum Wells.

Thin layers of different semiconductors with very sharp interfaces can be grown epitaxially (see HETEROSTRUCTURES AND SUPERLATTICES, SEMI-

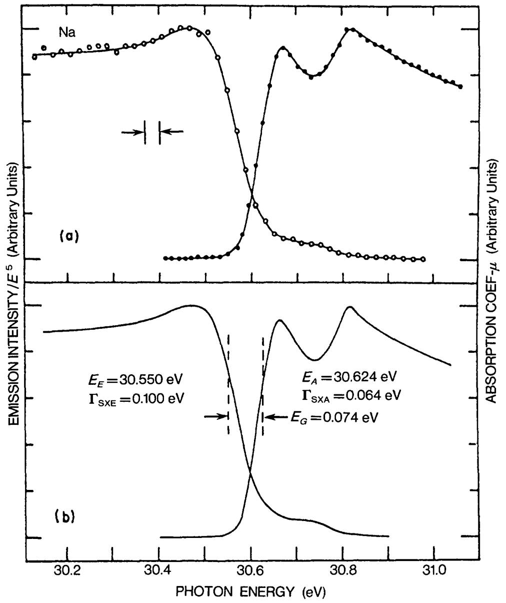

FIG. 6. Soft x-ray emission and absorption spectra with $L_{2,3}$ shell in Na: (a) spectral curves fitted to original data; (b) after correction for instrumental broadening. [From Callcott *et al.* (1978), Fig. 7; reproduced with permission.]

CONDUCTOR). If a semiconductor, such as GaAs, is sandwiched between two layers of another with a larger band gap, such as AlGaAs, electrons released by doping in the AlGaAs layers will be confined in the GaAs layer (see ELECTRON STATES IN ZERO-, ONE-, AND TWO-DIMENSIONAL STRUCTURES). Hence, the structure is called a quantum well. Quantization of the electron motion normal to the interfaces yields a set of subbands. In an undoped quantum well, the optical excitation of an electron from the topmost valence subband to the lowest conduction subband yields an exciton with a binding energy much stronger than the three-dimensional counterpart, such that the exciton line may be seen in the optical spectrum even at room temperature. Doping in the quantum well to produce a Fermi sea of electrons unbinds the excitons, not so much through screening of the attractive Coulomb potential as through the phase-space exclusion by the Fermi sea (Schmitt-Rink *et al.*, 1989). In a doped quantum well, are the Fermi edge singularities seen in the absorption and emission spectra for transitions between the conduction and high lying valence subbands?

Observations vary with samples (Fig. 7). In both InGaAs/InAlAs and GaAs/AlGaAs quantum wells, some samples (Skolnick *et al.*, 1987; Lee *et al.*, 1987) show the edge singularity behavior in both absorption and emission while others (Soorykumar *et al.*, 1987; Livescu *et al.*, 1988; Zhang *et al.*, 1990) show the edge singularity in absorption but not in emission.

The theory of the edge singularity reported above assumes that the valence electron energy has no dispersion. Thus, it has been assumed that when the edge singularities are seen in the quantum wells, the transition is not from the valence subband but from a localized state due either to the alloy fluctuations or to the interface roughness. A finite dispersion rounds off the power-law singularity (Müller-Hartmann *et al.*, 1971). In the last section, the singularity is shown to arise out of the density of states of the electron-hole pair being linear in the energy ω from the Fermi level. If the valence-band dispersion is taken into account, the density-wave energy is modified to be

$$\omega_k=\epsilon_{k_>}-\epsilon_{k_<}+(k_>-k_<)^2/2m_h, \qquad (93)$$

where m_h is the valence-hole mass. Then the density of states is quadratic in ω for $\omega < 4E_F m_e/m_h$ and linear in ω for larger ω. Thus, in the valence correlation function the power law is depressed to zero accompanied by the rise of the zero-phonon line if $\omega < 4E_F m_e/m_h$.

A more careful treatment of the finite valence-hole mass (Uenoyama and Sham, 1990) shows that, while the finite valence-band dispersion rounds off the power-law singularity in both the absorption and emission spectra, it also allows for the Boltzmann distribution of the valence hole before recombination which decreases the intensity of emission as the energy increases towards the Fermi level. This effect is clearly seen in the luminescence spectrum of a GaAs quantum well [Fig. 7(c)]. The roundoff is apparently not sufficient to remove the peak at the Fermi level in the absorption spectrum altogether.

4. SOLITONS

So far in this article, we have relied entirely on the concept of linear response to understand the collective excitations in solids. Consequently, these collective excitations are small oscillations of a property about its equilibrium value. A characteristic of the linear theory is that the collective modes of a particular kind can be superposed on one another without changing their properties. There are nonlinear couplings of these excitations, such as the anharmonic coupling of the phonons and the electron–phonon interaction. If the nonlinear corrections are treated as perturbations, the results are renormalization of the excitation dispersion (energy vs momentum) relations, modification of lifetime, and generation of harmonics or sum and difference frequencies (Nakajima *et al.*, 1980). Perturbation treatment of nonlinearity, however, misses important nonlinear effects, among them the collective excitation known as the soliton. We shall first study the soliton solutions of two important nonlinear equations and then survey their applications in solids.

When nonlinear interaction is taken into account there are collective modes which cannot be described in terms of small disturbances. The solitary wave is an example of a collective mode in which the deviation from

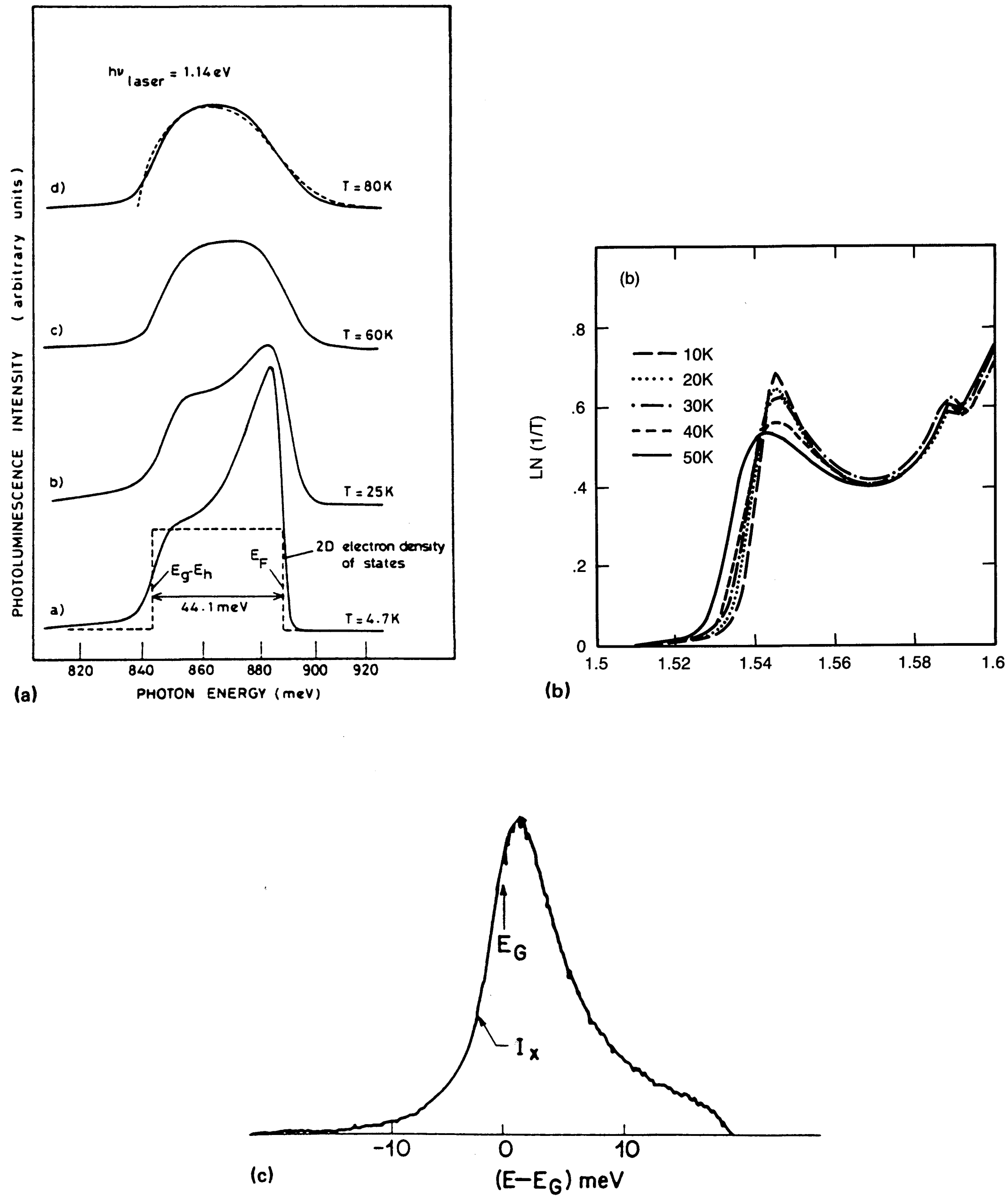

FIG. 7. (a) Emission spectra of an *n*-doped InGaAs/InP quantum well with an electron density of 9.1×10^{11} cm^{-2} at various temperatures. The dashed curves show the spectra calculated without the interaction effects. [From Skolnick *et al.*, 1987, Fig. 1; British Crown copyright/MOD, reproduced with the permission of the Controller of Her Britannic Majesty's Stationery Office.] (b) Absorption spectra at various temperatures of an *n*-doped GaAs/AlGaAs quantum well. [From Livescu *et al.*, 1988, Fig. 5(b); reproduced with permission. ©1988 IEEE.] (c) Emission spectrum of an *n*-doped GaAs/AlGaAs quantum well with a density of 9×10^{11} cm^{-2} at 2 K. [From Sooryakumar *et al.*, 1987, Fig. 2(a); reproduced with permission.]

equilibrium is large and yet which has a long lifetime. A wave packet of linear waves will in time disintegrate if different components of the linear waves move with different speeds. The soliton keeps its shape. In this section we describe the soliton solutions of two ubiquitous nonlinear equations in one dimension: the sine-Gordon and the nonlinear Schrödinger equations. The former solution is more accurately depicted as a kink, i.e., a finite deviation from the equilibrium value in one-half of the one-dimension system and no deviation in the other half of the system. The soliton solution of the nonlinear Schrödinger equation is truly an isolated wave of finite extent which moves through the system without changing its shape. The occurrence (or nonoccurrence in one instance) of the solitons in solids is surveyed. A particularly important role played by the soliton is domain formation near a phase transition, be it a transition from a paramagnetic solid to ferromagnetic, from a commensurate to incommensurate charge-density wave, or from a normal to superconducting metal. Soliton propagation of electromagnetic waves in a dispersive optical fiber may have important potential application in fiber-optic telecommunications.

4.1 The Sine-Gordon and the Nonlinear Schrödinger Equations

There is a large class of nonlinear differential equations with soliton solutions. We shall not discuss the method of solving such nonlinear equations and refer the reader to Eilenberg (1981) and Newell (1985), which are readable by a physicist with a normal mathematical background. We concentrate here on describing the soliton solutions to two nonlinear differential equations which appear in a great number of areas of solid-state physics.

The first is the sine-Gordon equation:

$$\frac{\partial^2\psi}{\partial t^2}-c^2\,\frac{\partial^2\psi}{\partial x^2}+\omega_p^2\,\sin\psi=0. \tag{94}$$

If ψ is small so that the sine term can be replaced by the argument, the solution is a harmonic wave with the frequency wave vector relation given by

$$\omega^2=\omega_p^2+q^2c^2, \tag{95}$$

which we have seen in the case of the transverse plasma oscillations in Sec. 2.2.6. In the linear approximation, a superposition of the waves can make up a wave packet localized in space. However, the components are dispersive so that after a time the width of the packet grows and the wave packet begins to unravel. The nonlinear term in the sine-Gordon equation acts to counter that tendency. A suitably localized wave packet can maintain its shape forever (or for as long as damping, which has not been included in the equation, does not come into play). Such a soliton solution is given by

$$\psi=4\,\arctan[e^{(x-x_0-vt)/\xi}]. \tag{96}$$

The solution changes from 0 to 2π in the neighborhood of x_0+vt within a width of the order ξ, where

$$\xi=\frac{c}{\omega_p}\,\sqrt{1-\frac{v^2}{c^2}}\,. \tag{97}$$

It is a rounded step propagating with speed $v<c$, not otherwise dependent on the parameters ω_p and c of the differential equation. Although the equation is nonlinear, a characteristic of the solitons is that two solitons can pass through each other with no change in shape. A more apt description is the term "kink" solution. Superposed on the soliton can be small oscillations (see Krumhansl in Bishop and Schneider, 1978).

The second nonlinear equation which appears frequently in solids is the nonlinear Schrödinger equation:

$$i\,\frac{\partial\psi}{\partial t}+\frac{\partial^2\psi}{\partial x^2}+|\psi|^2\psi=0. \tag{98}$$

The soliton solution is

$$\psi=\operatorname{sech}(x)e^{it/2}. \tag{99}$$

This is indeed a solitary wave peaked about the origin.

4.2 Some Applications in Solids

The soliton solution of the sine-Gordon equation occurs in a number of areas in solids (Bishop and Schneider, 1978; Davydov, 1985). The most direct application is in the Josephson junction transmission line (see also Superconductivity, Low-Temperature), which is composed of two superconducting metal strips separated by a thin insulating barrier permitting Josephson tunneling of Cooper pairs between the metal strips. The superconducting order parameter is given by ψ.

Close to the second-order phase transition, the Landau theory for the free energy quartic in the order parameter (see PHASE TRANSITIONS) yields a static version of the nonlinear equation with the sine term replaced by the first two terms of its power series. The quartic term is temperature dependent. Above the transition temperature, the free energy has only one minimum, which corresponds to the normal state. Below the transition temperature, the quartic polynomial develops two minima. The ground state is given by one of the two minima of the quartic free energy. A soliton solution is an excitation which connects one order parameter value in one spatial region of the system to the other order parameter value in another region. This applies to all second-order transitions with complex order parameters, including superconductivity, order–disorder, magnetism, and charge-density waves. For a quasi one-dimensional superconductor, the soliton is the counterpart of the vortex structure in three dimensions. In a ferromagnet, it is a domain wall. The soliton solutions are important to the statistical mechanics of critical phenomena around the transition.

In a one-dimensional charge-density-wave state commensurate with the lattice, such as the Peierls distortion (see the next section), the soliton connects two equivalent distortions in different sections of the linear chain and provides an excited state lying in the energy gap of the semiconductor. (See the review by Heeger *et al.*, 1988.) The kink solution in this case does not share the soliton property of two of them being able to pass through each other without changing shape. It has an unusual charge and spin relation. With the soliton, there appears near midgap a single electronic state. The unpaired electron gives a 1/2 spin but the whole complex is neutral (charge zero). If the midgap electronic state is empty, or is occupied by two electrons if their Coulomb repulsion is insufficient to put one of them above the gap, then the soliton is charged but its spin is zero. Such states are seen in the optical spectra of the conducting polymers. In a one-dimensional charge-density-wave state incommensurate with the underlying crystal lattice, e.g., in $NbSe_3$, the excited state of the sliding charge-density wave obeys the sine-Gordon equation. (See the review by Grüner, 1988.) The sliding charge density gets pinned by defects and, consequently, its effects on the transport show up not in the dc conductivity but in the ac and nonlinear conductivity. Tunneling through a pinning center is possible (Bardeen, 1986). Claims of solitons playing a role in the charge-density transport (Rice in Bishop and Schneider, 1978) were disputed by Grüner (1988). Near a commensurate–incommensurate phase transition, the incommensurate phase consists of large commensurate regions separated by domain walls, which have been shown to be solitons of the sine-Gordon equation (Bak in Bishop and Schneider, 1978).

Hasagawa and Tappert (1973) first suggested the possibility of soliton propagation of light in an optical fiber. Provided that the optical medium is dispersive (i.e, the frequency–wave-vector relation is nonlinear), the electromagnetic pulse propagation is governed by the nonlinear Schrödinger equation, in which $x \rightarrow t - x/v$ and $t \rightarrow x$ with the new distance and time variables expressed as dimensionless quantities in units of the critical length and time scales, which are dependent on the medium parameters. Mollenauer *et al.* (1986) have verified experimentally that such soliton propagation in a fiber is possible. They suggest that the inevitable slight loss be compensated by energy boosting with Raman scatterings at periodic intervals of the long fiber. This is potentially of great importance in optical communications.

5. PHASE TRANSITION FROM COLLECTIVE-MODE INSTABILITY

In the foregoing, we have, with the exception of the electron-hole drop, mostly described collective excitations from a "normal" state: resistive as opposed to superconducting, paramagnetic as opposing to magnetic, and high crystal symmetry as opposed to lower symmetry. The excitation of a collective mode increases the free energy of the system in a stable state. If it is found in theory to decrease the free energy of the system, then the assumed ground state is unstable against a finite distortion of the collective mode, resulting in a phase transition. This can come about by varying a system parameter, such as temperature, pressure, density, or magnetic field. The instability presages a phase transition. This view is taken here to introduce a model ferromagnetic transition, the charge- and spin-density-wave states and concomitant structural distortions, and the Wigner lattice, which, in a two-dimensional system, competes with the fractional quantum Hall effect.

The finite change of the collective-mode coordinate serves as the order parameter of the new phase. In solids, such a phase transition is accompanied by a change in symmetry. The change from high to low symmetry is often called broken symmetry. In the new phase, there is often a collective mode associated with the broken symmetry tending to restore the old phase. Notable examples not treated here are the collective mode of the superconducting order parameter and spin waves in antiferromagnets. For more detailed expositions of the consequences of such phase transitions, we refer to Anderson (1984) and Nakajima *et al.* (1980). In this section, we survey the possible unstable collective modes as preludes to phase transitions.

5.1 Itinerant-Electron Ferromagnetism

The simplest example of an instability is the spin-density response to a uniform and static magnetic field in jellium. The spin susceptibility given in Sec. 2.2.3 has a denominator $1+B_0$. The Landau parameter B_0 is negative. As the density is lowered, it reaches the value of -1. Then the paramagnetic ground state of equal populations of up and down spins of electrons becomes unstable against the state which has more electrons in one spin direction than the other. The new state possesses a spontaneous magnetization. From the total-energy point of view, the kinetic energy of the system is raised if, instead of the electrons being distributed equally in two spin states, more electrons are put in one spin state than in the other. On the other hand, this act would increase the exchange and correlation of the electrons, thereby lowering the total energy of the system since the exchange and correlation energy of the system is negative. Since the ratio of the exchange and correlation energy to the kinetic energy is r_s, which is inversely proportional to the cube root of the density (see Sec. 1.3.1), in the small-r_s or high-density limit the state is paramagnetic, and in the large-r_s or low-density limit the state is ferromagnetic. Ceperley and Alder (1980) have computed the density at which the transition of jellium to the ferromagnetic state occurs.

The above is a modern rendering of Bloch's theory of itinerant-electron ferromagnetism. The actual cause of ferromagnetism in real metals is more complicated. In *d*-electron ferromagnets, such as iron, cobalt, and nickel, the *d* orbitals are considered partly to provide localized spins and partly to provide itinerant spins. We refer to the article on MAGNETIC ORDERING IN SOLIDS for a proper discussion. We note in passing that the density-functional theory, in the local spin-density approximation, which describes the local exchange-correlation effect as that of an electron gas with the same local density and local spin density, produces magnetization or its absence for the 3*d* and 4*d* transition-metal series in good agreement with observation.

An example of a transition analogous to the ferromagnetic transition in jellium is provided by a quasi two-dimensional electron system in silicon metal-oxide-semiconductor field-effect transistors (MOSFETs), in which the role of the conduction valleys replaces the role of spins in the ferromagnetic transition. In a silicon MOSFET, a layer of electrons is trapped at the silicon interface with the oxide. If the silicon surface is in the (111) crystalline direction, there are six conduction valleys which contribute to the layer. At high density, the layer is made of equal populations from the six valleys. At low density, the splitting of the valley degeneracy has the same origin as the ferromagnetic transition of the electron gas (Bloss *et al.*, 1979). The exchange-correlation energy is stronger in a two-dimensional system than in a three-dimensional system because of the confinement in one direction. Thus, the valley transition in a quasi-two-dimensional system may be observed at a more accessible density.

5.2 The Wigner Crystal

In jellium at a sufficiently low density, the kinetic energy of the electron is small compared with the interaction energy. The Coulomb repulsion between pairs of electrons keeps the electrons apart from one another so that they form a lattice, known as the Wigner lattice. In the calculation by Ceperley and

Alder (1980) of the total energy for jellium, as the density decreases, the system first transforms into a ferromagnetic state, and, at still lower density, the electron gas is unstable against the Wigner crystal formation. The lattice of positive ions is, however, not favorable for the formation or observation of the Wigner crystal, nor is the density range in simple metals for which jellium is otherwise a successful model for many collective modes.

A more suitable system is the two-dimensional electron gas confined in a semiconductor heterostructure (Sec. 3.2.3) and under a strong magnetic field normal to the layer. The effect of the background semiconductor lattice is negligible. The appropriate electron density may be obtained by the design of the heterostructure system. The kinetic energy may be lowered relative to the interaction energy by increasing the magnetic field. In two dimensions, the electronic structure is composed of sets of Landau levels broadened by the impurities. As the magnetic field is increased, the number of Landau levels occupied by the electrons decreases. Although there are localized states between the Landau levels due to the two-dimensional disorder, they do not contribute to the Hall voltage. Thus, the famous plateaus are observed (von Klitzing *et al.*, 1980) in the Hall measurement around integral occupations of the Landau levels as the magnetic field is varied. (See QUANTUM HALL EFFECT, or Prange and Girvin, 1987.) The Hall plateaus are also observed (Tsui *et al.*, 1982) at fractional occupation of a Landau level. The state of a fractionally occupied Landau level is a new phase of an incompressible fluid with a gap in its excitation spectrum (Laughlin in Prange and Girvin, 1987). The fractional quantum Hall states compete with the two-dimensional Wigner crystal, which has been claimed to be observed at very low density (Etienne *et al.*, 1988).

5.3 Charge- and Spin-Density-Wave States

A collective mode of oscillating charge or spin density at a finite wave vector may become unstable. Then the ground state of the system develops a static charge- or spin-density distribution with a finite wave vector. The Wigner lattice may be regarded as a linear combination of density waves.

The giant one-dimensional Kohn anomaly which occurs in both the charge- and spin-density responses (Sec. 2.1) causes instability in the charge- and spin-density waves. Whether the system adopts the charge- or spin-density-wave state depends on the interaction energy. In the case of the charge-density state, the lattice of charged ions distorts with the electron-density wave because of the enormous Coulomb attraction between the ions and the electrons. In one-dimensional conductors (see the review volume edited by Devreese *et al.*, 1978), the Kohn anomaly thus causes the phonon at the Fermi diameter $2k_F$ at low temperatures to become unstable, creating a charge-density-wave state.

The formation of the charge-density wave may alternatively be viewed via consideration of the total energy of the system. In the simple example of a linear chain, such as polyacetylene, in which each molecule contributes one conduction electron, the half-filled conduction band of an evenly spaced chain is unstable against dimerization of the chain (displacement of every other molecule, known as the Peierls distortion) opening up a gap in the electron band structure at the Fermi level, transforming the ground state to a semiconductor. If u denotes the displacement of every other molecule, the increase in elastic energy of the chain is proportional to u^2 but the decrease of the total electron energy at zero temperature due to the gap formation is proportional to $u^2 \ln u$, as can be seen by summing the electron energies $-\sqrt{(\epsilon^2+\Delta^2)}$ of the occupied states near the gap where ϵ is the energy measured from the Fermi level in the absence of the distortion and Δ is the half-gap, proportional to u. The balance of the two energy terms yields a minimum of the total energy at a finite lattice distortion, u. In this example, the wavelength of the charge-density wave bears a simple rational ratio to the lattice constant and the charge-density wave is known as a commensurate state. It is possible to have the charge-density wavelength incommensurate with the lattice constant. In that case the static charge density may slide over the periodic lattice potential, yielding a new collective mode. This happens in $NbSe_3$ (see Sec. 4.2; Bardeen, 1986; Grüner, 1988).

In two- or three-dimensional systems, parallel sections of the Fermi surface will also give rise to the giant Kohn anomaly and, hence, a charge- or spin-density-wave instability. These states are common in the layer compounds of transition-metal dichalcogenides, e.g., $NbSe_2$ (Wilson *et al.,* 1975), where the quasi two-dimensional Fermi surface often has parallel sections. A famous example of the spin-density wave in a three-dimensional solid is the explanation of antiferromagnetism in Cr (Fawcett, 1988).

Phonon modes in solids driven unstable by the electron-density response sometime precipitate structural phase transitions. The structural distortion in the quasi one-dimensional conductor potassium cyanoplatinate, at low temperatures, and the soft-phonon modes at the corresponding wave vector above the transition temperature have both been observed (Devreese *et al.,* 1978). A number of the quasi two-dimensional transition-metal dichalcogenides show soft phonons above the charge-density-wave transition temperatures. In some of the A15 compounds, such as Nb_3Sn (reviewed by Sham, 1978), the Sn atoms form a body-centered lattice and the Nb atoms form chains running on the cube crystal planes in all three $\langle 100 \rangle$ directions. Charge-density waves along the chains cause the C_{11}–C_{12} shear to become unstable at low temperatures, accompanied by a Γ_{12} optical-mode distortion, creating a cubic-to-tetragonal transformation.

In this article, we have not touched on diplacive transitions which are not caused by the electron response, as occur in ferroelectric crystals (see DIELECTRIC PROPERTIES OF INSULATORS).

ACKNOWLEDGMENTS

My part of the research is supported by NSF Grant No. DMR 88-15068.

GLOSSARY

Fermi Liquid: A system of particles obeying Fermi statistics and having pairwise interactions. Liquid ^{3}He is an example of an uncharged Fermi liquid and an interacting electron system is an example of a charged Fermi liquid.

Functional Derivative: If a quantity E is determined by a function $n(k,r)$ of variables k and r, an infinitesimal change of the function $\delta n(k,r)$ brings about a change in E of the form

$$\delta E = \int dk \int dr\, f(k,r)\delta n(k,r),$$

where $f(k,r)$ is known as the functional derivative, denoted by $\delta E/\delta n$.

Incommensurate Charge-Density-Wave State: A state with a sinusoidal spatial variation of the electron density of a wavelength not bearing a simple rational ratio to the lattice vector.

MOSFET: Metal-oxide-semiconductor field effect transistor. The MOS forms a capacitor which can confine a layer of electrons on the semiconductor side.

List of Works Cited

Abrikosov, A. A. (1972), *Introduction to the Theory of Normal Metals,* Solid State Physics Suppl. **12**, New York: Academic.

Allen, S. J., Tsui, D. C., Logan, R. A. (1977), *Phys. Rev. Lett.* **38**, 980–983.

Almbladh, C.-O., Hedin, L. (1983), in: E. E. Koch (Ed.), *Handbook on Synchrotron Radiation,* Vol. 1. Amsterdam: North-Holland, p. 607–904.

Anderson, P. W. (1967), *Phys. Rev. Lett.* **18**, 1049–1051.

Anderson, P. W. (1984), *Basic Notions of Condensed Matter Physics,* Menlo Park: Benjamin/Cummings.

Bardeen, J. (1986), *Physica B* **134**, 14–18.

Bishop, A. R., Schneider, T. (1978), (Eds.) *Solitons and Condensed Matter Physics,* Berlin: Springer-Verlag.

Bloss, W. L., Sham, L. J., Vinter, B. (1979), *Phys. Rev. Lett.* **43**, 1529–1533.

Brovman, E. G., Kagan, Yu. M. (1974), in: G. K. Horton and A. A. Maradudin (Eds.), *Dynamical Properties of Solids,* Vol. I. Amsterdam: North-Holland, pp. 191–300.

Callcott, T. A., Arakawa, E. T., Ederer, D. L. (1978), *Phys. Rev. B* **18**, 6622–6630.

Ceperley, D. M. (1978), *Phys. Rev. B* **18**, 3126–3138.

Ceperley, D. M., Alder, B. J. (1980), *Phys. Rev. Lett.* **45**, 566–569.

Citrin, P. H., Wertheim, G. K., Baer, Y. (1977), *Phys. Rev. B* **16**, 4256–4282.

Citrin, P. H., Wertheim, G. K., Schlüter, M. (1979), *Phys. Rev. B* **20**, 3067–3114.

Cohen, M. L., Chelikowsky, J. (1988), *Electronic Structures and Optical Properties of Semiconductors,* Berlin: Springer-Verlag.

Davydov, A. S. (1985), *Solitons in Molecular Systems,* Dordrecht: Reidel.

Devreese, J. T., Evrard, R. P., van Doren, V. E. (1978) (Eds.), *Highly Conducting One-Dimensional Solids,* New York: Plenum.

Duke, C. B., Mahan, G. D. (1965), *Phys. Rev.* **139**, A1965–A1982.

Dunifer, G. L., Pinkel, D., Schultz, S. (1974), *Phys. Rev. B* **10**, 3159–3185.

Eilenberger, G. (1981), *Solitons: Mathematical Methods for Physicists,* Berlin: Springer-Verlag.

Etienne, B., Paris, E., Andrei, E., Deville, G., Glattli, D. C., Williams, F. I. B. (1988), in: W. Zawadzki (Ed.), *Proceedings of the 19th International Conference on the Physics of Semiconductors,* Warsaw: Institute of Physics, Polish Academy of Sciences, pp. 139–142.

Fawcett, E. (1988), Rev. Mod. Phys. **60**, 209–283.

Gibbons, P. C., Schnatterly, S. E., Ritsko, J. J., Fields, J. R. (1976), *Phys. Rev. B* **13**, 2451–2460.

Godby, R. W., Schlüter, M., Sham, L. J. (1987), *Phys. Rev. B* **37**, 10,159–10,175.

Grüner, G. (1988), *Rev. Mod. Phys.* **60**, 1129–1181.

Hansel, J. C., Phillips, T. G., Thomas, G. A. (1977), in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics: Advances in Research and Applications,* Vol. 32. New York, Academic, pp. 88–202.

Hasagawa, A., Tappert, F. (1973), *Appl. Phys. Lett.* **23**, 142–144.

Heeger, A. J., Kivelson, S., Schrieffer, J. R., Su, W. -P. (1988), *Rev. Mod. Phys.* **60**, 781–850.

Heitmann, D. (1986), *Surf. Sci.* **170**, 332–345.

Hopfield, J. J. (1969), *Comments Solid State Phys.* **11**, 40–49.

Hohenberg, P., Kohn, W. (1964), *Phys. Rev.* **136**, B864–B871.

Jeffries, C. D., Keldysh, L. V. (1983) (Eds.), *Electron-Hole Droplets in Semiconductors,* Amsterdam: North-Holland.

Kaner, E. A., Skobov, V. G. (1966), *Usp. Fiz. Nauk.* **89**, 367–408. [English translation: *Sov. Phys. Usp.* **9**, 480–503 (1967)].

Kohn, W., Sham, L. J. (1965), *Phys. Rev.* **140**, A1133–A1138.

Lee, J. S., Iwasa, Y., Miura, N. (1987), *Semicond. Sci. Technol.* **2**, 675–678.

Livescu, G., Miller, D. A. B., Chemla, D. S., Ramaswamy, M., Chang, T. Y., Sauer, N., Gossard, A. C., English, J. H. (1988), *IEEE J. Quantum Electron.* **24**, 1677–1689.

Mahan, G. D. (1967), *Phys. Rev.* **153**, 882–889.

Mahan, G. D. (1990), *Many-Particle Physics,* 2nd edition, New York: Plenum Press.

Maradudin, A. A., Montroll, E. W., Weiss, G. H. (1963), *Theory of Lattice Dynamics in the Harmonic Approximation,* New York: Academic Press, p. 283.

Maradudin, A. A. (1966), in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics: Advances in Research and Applications,* Vol. 18. New York: Academic, pp. 273–420.

Markiewicz, R. S., Timusk, T. (1984), in: C. D. Jeffries, L. V. Keldysh (Eds.), *Electron-Hole Droplets in Semiconductors,* Amsterdam: North-Holland.

Mermin, N. D., Cheng, Y. C. (1968), *Phys. Rev. Lett.* **20**, 839–842.

Mollenauer, L. F., Gordon, J. P., Islam, M. N. (1986), *IEEE J. Quantum Electron.* **QE-22**, 157–173.

Müller-Hartmann, E., Ramkrishnan, T. V., Toulouse, G. (1971), *Phys. Rev. B* **3**, 1102–1119.

Nakajima, S., Toyozawa, Y., Abe, R. (1980), *The Physics of Elementary Excitations,* Berlin: Springer-Verlag.

Newell, A. C. (1985), *Solitons in Mathematics and Physics,* Philadelphia: Society for Industrial and Applied Mathematics.

Nozières, P., DeDominicis, C. T. (1969), *Phys. Rev.* **178**, 1097–1107.

Oliveira, L. N., Gross, E. K. U., Kohn, W. (1988), *Phys. Rev. Lett.* **60**, 2430–2434.

Overhauser, A. (1962), *Phys. Rev.* **128**, 1437–1452.

Pines, D., Nozières, P. (1966), *The Theory of Quantum Fluids,* Reading, MA: Benjamin.

Platzman, P. M., Wolff, P. A. (1967), *Phys. Rev. Lett.* **18**, 280–283.

Platzman, P. M., Wolff, P. A. (1973), *Waves and Interactions in Solid State Plasmas,* New York: Academic Press.

Prange, R. E., Girvin, S. M. (1987), (Eds.), *The Quantum Hall Effect,* New York: Springer-Verlag.

Raether, H. (1980), *Excitation of Plasmons and Interband Transitions by Electrons,* Springer Tracts in Modern Physics, Vol. 88, Berlin: Springer.

Rashba, E. I., Sturge, M. D. (1979) (Eds.), *Excitons,* Amsterdam: North-Holland.

Rice, T. M. (1977), in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics: Advances in Research and Applications,* Vol. 32. New York: Academic, pp. 1–87.

Sander, L., Shore, H. B. (1972), *Phys. Rev. B* **3**, 1472–1482.

Schlüter, M., Sham, L. J. (1982), *Physics Today* **35**, 36–43.

Schmitt-Rink, S., Chemla, D. S., Miler, D. A. B. (1989), *Adv. Phys.* **38**, 89–188.

Schnatterly, S. E. (1979), in : H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics; Advances in Research and Applications,* Vol. 34. New York: Academic, pp. 275–358.

Schotte, K. D., Schotte, U. (1969), *Phys. Rev.* **182**, 479–482.

Sham, L. J. (1969), in: R. H. Enns, R. R. Haering (Eds.), *Modern Solid State Physics, II. Phonons and Their Interactions,* London: Gordon and Breach, pp. 143–198.

Sham, L. J. (1978), in: M. Balkanski (Ed.), *Proceedings of the International Conference on Lattice Dynamics, Paris, September 5–9, 1977,* Paris:

Flammarion, pp. 557–561.

Sham, L. J. (1985), *Phys. Rev.* **32**, 3876–3882.

Skolnick, M. S., Rorison, J. M., Nash, K. J., Mowbray, D. J., Tapster, P. R., Bass, S. J., Pitt, A. D. (1987), *Phys. Rev. Lett.* **58**, 2130–2133.

Sooryakumar, R., Pinczuk, A., Gossard, A. C., Chemla, D. S., Sham, L. J. (1987), *Phys. Rev. Lett.* **58**, 1150–1153.

Tsui, D. C., Störmer, H. L., Gossard, A. C. (1982), *Phys. Rev. Lett.* **48**, 1559–1562.

Uenoyama, T., Sham, L. J. (1990), *Phys. Rev. Lett.* **65**, 1048–1051.

Vashishta, P., Kalia, R. K., Singwi, K. S. (1984), in: C. D. Jeffries, L. V. Keldysh (Eds.), *Electron-Hole Droplets in Semiconductors,* Amsterdam: North-Holland.

Vignale, G., Rasolt, M. (1988), *Phys. Rev. B* **37**, 10,685–10,696.

Von Klitzing, K., Dorda, G., Pepper, M. (1980), *Phys. Rev. Lett.* **45**, 494–497.

Westervelt, R. M. (1977), Ph. D. thesis, University of California, Berkeley.

Wilson, J. A., DiSalvo, F. J., Mahajan, S. (1975), *Adv. Phys.* **24**, 117–201.

Wolfe, J. P., Jeffries, C. D. (1984), in: C. D. Jeffries, L. V. Keldysh (Eds.), *Electron-Hole Droplets in Semiconductors,* Amsterdam: North-Holland.

Zhang, Y.-H., Jiang, D.-S., Cingolani, R., Ploog, K. (1990), *Appl. Phys. Lett.* **56**, 2197–2200.

Further Reading

Abrikosov, A. A. (1971), *Introduction to the Theory of Normal Metals,* Solid State Physics Suppl. **12**, New York: Academic Press. An excellent account of the electromagnetic properties in metals.

Davydov, A. S. (1985), *Solitons in Molecular Systems,* Dordrecht: Reidel. Despite its title, it contains a comprehensive account of the soliton phenomena in various solids.

Jeffries, C. D., Keldysh, L. V. (1984), *Electron-Hole Droplets in Semiconductors,* Amsterdam: North-Holland. The latest review of both theories and experiments on the electron-hole liquid by experts.

Mahan, G. D. (1990), *Many-Particle Physics,* 2nd ed., New York: Plenum Press. An excellent treatment of several aspects of the many-body problem in solids, including further details of the topics covered here on the electron gas and on the edge singularity.

Nakajima, S., Toyozawa, Y., Abe, R. (1980), *The Physics of Elementary Excitations,* Berlin: Springer-Verlag. An excellent introductory text on collective excitations.

Parr, R. G., Yang, W. (1989), *Density-Functional Theory of Atoms and Molecules,* New York, Oxford University Press. A comprehensive review with emphasis on applications in quantum chemistry.

Pines, D., Nozières, P. (1966), *The Theory of Quantum Fluids,* Reading, MA: Benjamin. A very readable account of the Landau Fermi-liquid theory.

Riste, T. (1977) (Ed.), *Electron-Phonon Interactions and Phase Transitions,* New York: Plenum. Contains reviews of phase transitions due to soft modes, in particular in the transition metal dichalcogenides.

Trickey, S. B. (1990) (Ed.) *Density-Functional Theory of Many-Fermion Systems,* a special issue of *Adv. Quan. Chem.* **21**. Designed for the 25th anniversary of the theory with perspectives by the originators and reviews by many authorities on the subject.

COLLISION PROCESSES OF ATOMS AND MOLECULES

See MOLECULAR AND ATOMIC COLLISION PROCESSES

COMBUSTION

CHARLES K. WESTBROOK, *Physics Department, Lawrence Livermore National Laboratory, Livermore, California, U.S.A.*

	Introduction	1
1.	**Combustion Fundamentals**	1
1.1	Fuels	2
1.2	Oxidizers	2
1.3	Concept of Stoichiometry	2
2.	**Mixing of Fuel and Oxidizer**	3
2.1	Premixed Systems	3
2.2	Laminar Burning Velocity	3
2.3	Flammability Limits	4
2.4	Detonations	4
2.5	Turbulent versus Laminar Combustion	5
2.6	Diffusion Flames (Nonpremixed)	6
2.7	Spray Combustion and Diffusion Flames	6
3.	**Combustion Chemistry**	7
3.1	Initiation Reactions	8
3.2	Abstraction Reactions	8
3.3	Radical Decomposition	9
3.4	Radical Recombination	9
3.5	Pollutant Chemistry and Emissions	9
4.	**Practical Combustion Systems**	11
4.1	Internal Combustion Engines	11
4.1.1	Engine Knock	12
4.1.2	Diesel Engines	13
4.2	Burners, Boilers, and Furnaces	14
4.3	Novel Combustion Systems	14
4.4	Incineration and Toxic Wastes	14
5.	**Summary**	15
	Glossary	15
	Works Cited	15
	Further Reading	16

INTRODUCTION

Over millennia, combustion has provided society with much of its practical energy production. From the fire that cooked food for cavemen, to the forges that enabled prehistoric man to emerge from the Stone Age, to the gunpowder that transformed relations between societies in the Middle Ages, and to the boilers and furnaces that drove the Industrial Revolution, combustion has been intimately interwoven with the technological factors that have moved civilization. Even today, although alternative sources of energy production are being intensively pursued, a large fraction of practical energy production is the result of combustion of hydrocarbon fuels.

Combustion has also been responsible for many of the environmental problems associated with modern civilization. The killer fogs and smogs of the past century and the problems of modern air pollution and acid rain are products of the use of combustion for practical energy production. Even more recently, the production of carbon dioxide, an apparently benign product of complete combustion, is now being associated with global warming and the so-called greenhouse effect. (See AIR POLLUTION: COMPONENTS, CAUSES, AND CURES.)

Combustion interacts with society in many ways, influencing economics, foreign policy, and other political and sociological problems. The import and export of hydrocarbon combustion fuels has an enormous impact on the balance of payments for many countries, and access to fuels has played a large role in many foreign policy decisions over the past hundred years.

1. COMBUSTION FUNDAMENTALS

Combustion consists of the reaction of one or more chemical species to produce other species, together with the release of heat. As stated by Strehlow (1984), combustion occurs when "...weak molecular bonds are broken and replaced by stronger bonds and the excess

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

bond energy is liberated to the system, usually causing a very large increase in temperature." This heat can then be used to do work in a thermodynamic sense, driving a practical system such as an internal combustion engine or a furnace.

1.1 Fuels

In most familiar systems, combustion takes place between a fuel, often a hydrocarbon species, and an oxidizer, usually air. The physical properties of the hydrocarbon fuel can vary considerably. Many fuels exist as gases under normal conditions. For example, natural gas consists primarily of methane (CH_4) with small amounts of ethane (C_2H_6), propane (C_3H_8), and other trace species, all of which are gases at room temperature and pressure. Automobile gasoline and diesel fuel are complex mixtures of larger hydrocarbons, with components as large as toluene (C_7H_8), octane (C_8H_{18}), and dodecane ($C_{12}H_{26}$), all of which are liquids at room temperature. Familiar fuels such as firewood, paper, gunpowder, and many others are solids under normal conditions. Fuel storage and distribution, as well as the configuration of the combustion system that consumes them, will then depend on the nature of the fuel.

The fuel need not consist of a hydrocarbon species; one simple fuel is a mixture of molecular hydrogen and carbon monoxide. Other fuels can include N-H species such as ammonia (NH_3) or hydrazine (N_2H_4), sulfur-bearing species such as hydrogen sulfide (H_2S) and carbon disulfide (CS_2), and mixed species such as hexyl nitrate ($C_6H_{13}NO_2$) and hydrogen cyanide (HCN).

In the familiar case of toxic waste incineration, the "fuel" can consist of any of a wide variety of chlorinated hydrocarbon species. Some of the most common of these are methyl chloride (CH_3Cl), vinyl chloride (C_2H_3Cl), and trichloroethylene, TCE (C_2HCl_3). Combustion of such species leads to carbon dioxide, water, and hydrogen chloride (or hydrochloric acid), HCl.

1.2 Oxidizers

Similarly, in the most common combustion situations, the fuel reacts with oxygen, either that present in air (which includes 21% O_2) or in the form of pure oxygen. From a historical point of view, combustion has been nearly synonymous with oxidation of these fuels in air, and this is still the basis of most practical combustion. However, when higher combustion temperatures are desirable, it is possible to use pure oxygen as an oxidizer, and other species such as N_2O_4 and ozone (O_3) are sometimes used as oxidizers for specialized applications. The key to all of these systems is that the reaction of the fuel with the oxidizer releases a significant amount of thermal energy.

1.3 Concept of Stoichiometry

The final product temperature and composition of a combustible mixture, and often the overall rate of combustion, depend on the ratio of fuel to oxidizer. A *stoichiometric* mixture is defined as one in which there is exactly the correct amount of oxidizer present to convert all of the fuel to final products. For example, if propane (C_3H_8) is the fuel, five oxygen (O_2) molecules are required to convert each propane molecule to CO_2 and H_2O:

$$C_3H_8+5O_2=3CO_2+4H_2O. \tag{1}$$

Therefore, a stoichiometric mixture of propane and oxygen would consist of $\frac{1}{6}$ propane and $\frac{5}{6}$ oxygen. If the oxidizer in the same example is air instead of oxygen, then the overall reaction is given by

$$C_3H_8+5O_2+18.8N_2=3CO_2+4H_2O+18.8N_2, \tag{2}$$

since the ratio of N_2 to O_2 in air is approximately 3.76, and a stoichiometric mixture of propane and air would include approximately 4% propane [i.e., one molecule of propane in the 24.8 molecules on the reactant side of reaction (2) above]. The same number of chemical bonds are broken and made in reactions (1) and (2), and so the total heat release is equal in the two reactions. Since that heat must be shared by more product molecules in reaction (2), the product temperature in reaction (1) will be considerably higher than that of reaction (2).

A given fuel-oxidizer mixture can include more or less fuel than a stoichiometric amount. The ratio of the fuel concentration actually present to that which would be the stoichiometric amount is termed the *equivalence ratio*, usually indicated by the symbol Φ, and a mixture with more than a stoichiometric amount of fuel (fuel-rich) has $\Phi > 1$, whereas a mixture with less than a stoichiometric amount of fuel (fuel-lean) has an equivalence ratio $\Phi < 1$.

The importance of the equivalence ratio can be illustrated by showing the variation in product temperature for a methane-air mixture in Fig. 1. The product temperature is defined as the final temperature achieved by a mixture of fuel and oxidizer at a particular equivalence ratio when all of the heat released by the reaction is used to heat the reaction products. This condition is also known as the *adiabatic* condition, and the product temperature is known as the *adiabatic flame temperature*, T_{ad}. The maximum value of T_{ad} is achieved for values of the equivalence ratio close to stoichiometric, falling quite rapidly as Φ moves towards lean or rich. Values of T_{ad} for many fuel-oxidizer mixtures have been collected and tabulated by Lewis and von Elbe (1987) and by Gaydon and Wolfhard (1979). Generally speaking, fuel-oxygen mixtures have much higher values for T_{ad} than fuel-air mixtures, but for both fuel-air and fuel-oxidizer mixtures maximum product temperatures are achieved at nearly stoichiometric values of the equivalence ratio. One result of this trend is that the greatest rate of combustion and the highest power production are achieved for stoichiometric conditions. Unfortunately, for combustion in air (which contains nearly 80% molecular nitrogen), production of nitric oxide (NO), a major chemical pollutant caused by combustion, grows very rapidly with increasing temperature, so that the rate of NO production is also greatest for stoichiometric conditions. This results is one of the great conflicts in much of practical (i.e., air-breathing) combustion: the greatest efficiency and power production occurs at $\Phi = 1$, but the greatest rate of pollutant production also occurs at $\Phi = 1$. One must often sacrifice some degree of performance in order to reduce the unwanted pollutant emissions.

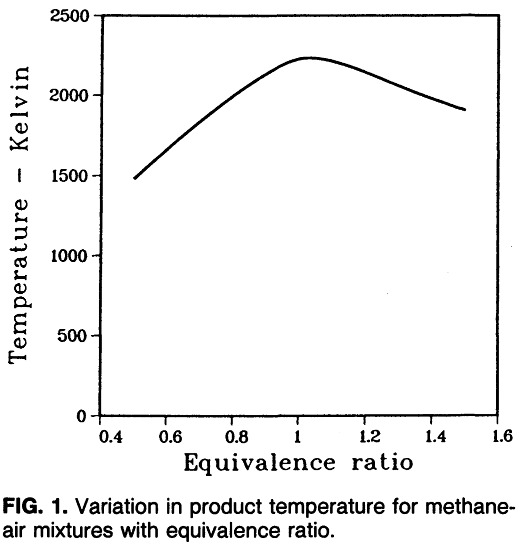

FIG. 1. Variation in product temperature for methane-air mixtures with equivalence ratio.

In addition to the final temperature, the product chemical composition depends strongly on Φ. In particular, on the fuel-rich side of stoichiometric, insufficient O_2 results in incomplete oxidation of the fuel, and products such as carbon monoxide and soot will result.

2. MIXING OF FUEL AND OXIDIZER

In order for combustion to occur, the fuel and oxidizer must interact on a molecular level. There are two general situations that can be used to characterize most combustion environments, based on the mechanism by which the fuel and oxidizer are combined. In the first limit, the fuel and oxidizer are already mixed prior to combustion, whereas in the opposite limit the fuel and oxidizer are initially separated and must be mixed during the combustion process.

2.1 Premixed Systems

In so-called premixed systems, molecular-scale mixing is accomplished prior to combustion. For example, in the familiar case of a laboratory Bunsen burner, the fuel and air are first mixed and stored in a single container, from which the combustible mixture flows into the burner itself. Another perhaps less obvious premixed system is found in the conventional internal combustion engine, where the initially liquid fuel is vaporized and mixed with air in the carburetor immediately prior to being introduced into the engine combustion chamber.

2.2 Laminar Burning Velocity

Premixed combustion represents an instructive limiting condition, and many fundamental combustion properties are defined in terms of premixed processes. For example, if a stoichiometric premixed fuel-air mixture is confined in a linear tube, and the mixture is ignited at one end of the tube, a flame will propagate down the tube. If the gas mixture is initially quiescent, and heat and mass transport processes take place on a molecular level (the laminar limit), then the rate of propagation of this flame is a fundamental combustion characteristic of the fuel, termed the laminar burning velocity, S_u. The laminar burning velocity is found experimentally to depend on the equivalence ratio, on the gas pressure, and on the unburned gas temperature. It also depends on the chemical properties of the fuel, so that S_u varies from one fuel to another. In Fig. 2, taken from Westbrook *et al.* (1983), the laminar burning velocity of an ethylene-air mixture is plotted as a function of equivalence ratio at two different pressures. The shapes of these curves are seen often in combustion problems, having maximum values for mixtures close to stoichiometric and falling for both rich and lean conditions.

Different fuels and oxidizers can be compared by noting their respective laminar burning velocities. Because heat release rates and adiabatic flame temperatures are higher for fuel-oxygen mixtures than for fuel-air mixtures, fuel-oxygen mixtures are found to have larger laminar burning velocities. The laminar burning velocities for a variety of stoichiometric fuel-oxidizer mixtures are shown in Table 1.

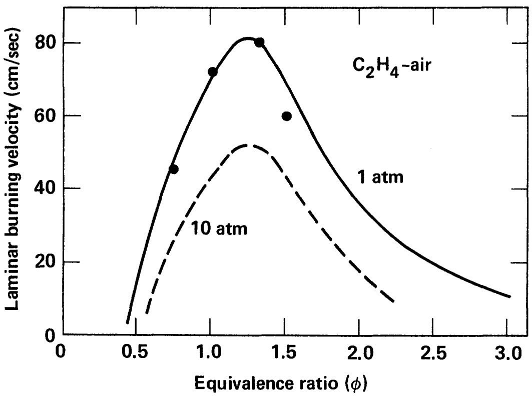

FIG. 2. Variation in laminar burning velocity of ethylene-air mixtures with equivalence ratio at atmospheric pressure and at 10-atm pressure.

Table 1. Laminar burning velocities for different stoichiometric mixtures of fuel and oxidizer.

Fuel-oxidizer mixture	S_u (cm/s)	Fuel-oxidizer mixture	S_u (cm/s)
H_2-air	310	H_2-oxygen	1100
Methane-air	38	Methane-oxygen	450
Propane-air	40	Acetylene-oxygen	1140
Ethylene-air	70	Ammonia-oxygen	115
Methanol-air	44		
Octane-air	40		

2.3 Flammability Limits

In Fig. 2, as the equivalence ratio differs from nearly stoichiometric for either lean or rich conditions, the laminar burning velocity decreases steadily from its maximum value. Extrapolating this curve in either direction suggests that S_u will fall to zero for a critical lean value Φ_L and a critical rich value Φ_R of the equivalence ratio. Experimentally, it is observed that for fuel-oxidizer mixtures outside these values of equivalence ratio, flame propagation is impossible to achieve. These limiting values Φ_L and Φ_R represent the lean and rich flammability limits for premixed flame propagation and are often used as fundamental parameters for combustion of different fuels. Lean and rich flammability limits for a variety of fuel-oxidizer mixtures are given in Table 2. Extensive studies of lean and rich flammability limits have been made by Simon (1951) and by NACA (1959).

2.4 Detonations

In a manner similar to laminar flame propagation, it is found that a detonation wave can propagate down a linear tube filled with a premixed fuel-oxidizer mixture. The detailed properties of detonations are quite complex (e.g., Fickett and Davis, 1979; Nettleton, 1987), but the fundamental feature of a detonation is that combustion is caused by the propagation of a shock wave into the reactive mixture (see EXPLOSIVES).The shock wave compresses and heats the premixed gases to a point where they react rapidly, releasing heat and reinforcing the propagating shock wave. The structure of such a detonation wave is characterized by the existence of cell-like structures that can be recorded as the detonation proceeds down the tube. A typical record of detonation cells is shown in Fig. 3. The sizes of these cells depend

Table 2. Lean and rich flammability limits; values shown are percentages of fuel in the total mixture.

Fuel-oxidizer mixture	Lean limit	Rich limit
Hydrogen-air	4.0	75.0
Methane-air	5.0	15.0
Propane-air	2.1	9.5
Ethylene-air	2.7	36.0
Hydrogen-O_2	4.0	95.0
Methane-O_2	5.1	61.0
Ammonia-air	15.0	28.0
Methanol-air	6.7	36.0
Acetylene-air	2.5	100.0

on the fuel-oxidizer equivalence ratio and on the unreacted mixture temperature and pressure. As in the case of laminar flame propagation, there are lean and rich limits to detonability that represent fundamental combustion properties of a given fuel.

2.5 Turbulent versus Laminar Combustion

The fluid mechanical features of combustion systems depend very strongly on the magnitude of gradients of various quantities, including the temperature, pressure, or fluid velocity field. When these quantities vary slowly throughout the physical domain of interest, then the fluid flow is laminar, and momentum, mass, and heat transport are controlled by molecular properties of the medium. However, when there are significant gradients, particularly in fluid velocity, then fluid mechanical instabilities are observed in which the flow system is found to break up spontaneously into chaotic, statistically random motions. Under such conditions, the fluid system is made up of three-dimensional structures often called turbulent eddies that move and interact among themselves, exchanging heat, mass, and momentum while retaining their identity as eddies for relatively long periods of time. These eddies are much larger than characteristic length scales encountered in laminar systems, and net transport of heat, mass, and momentum is found to be much more efficient and rapid than in laminar environments.

The most reliable quantitative measure that divides laminar from turbulent regimes is the *Reynolds number*, denoted as *Re*. The Reynolds number of the fluid is defined (Williams, 1965) in terms of the fluid velocity V, a characteristic length scale L of the system being studied, the gas density ρ, and the viscosity μ, as

$$Re = LV\rho/\mu = LV/\nu, \tag{3}$$

where ν is the kinematic viscosity, $\nu = \mu/\rho$. The Reynolds number can be interpreted as the

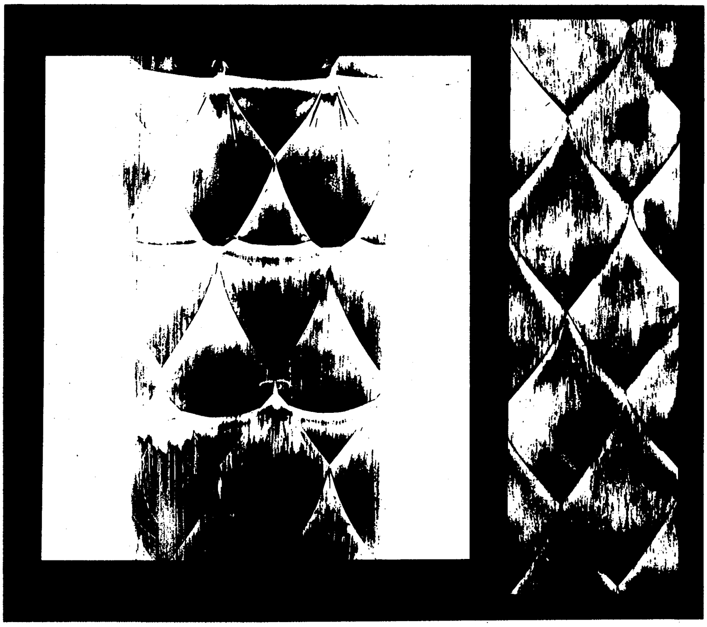

FIG. 3. Detonation cells observed using smoked foils in a detonation tube. Picture provided by Prof. R. A. Strehlow, University of Illinois.

ratio of inertial fluid forces to viscous forces in the fluid, with the inertial forces driving the system unstable and viscous forces trying to restore stability to the flow. When *Re* is small, laminar conditions are observed, and when *Re* is large, turbulent regimes are observed. Between these limits, the flow is found to be intermittently unstable. Values of *Re* in typical laminar flows are generally less than 100, values in fully turbulent flows are often 10^3 or greater, with the intermittent regime characterized by values of *Re* between these values.

In turbulent premixed combustion, the turbulent unburned fuel-oxidizer mixture produces a highly wrinkled flame surface. These wrinkles vary randomly in time and magnify the total amount of surface area between burned and unburned gases, thereby increasing dramatically the overall rate of heat release that can be achieved.

2.6 Diffusion Flames (Nonpremixed)

In contrast with premixed combustion systems, often the fuel and oxidizer are initially separate and then mixed together as they burn. This is actually the case in the great majority of practical combustion systems. For example, in a common gas-burning kitchen stove, jets of pure fuel gas emerge from the burner, where the fuel mixes with the air in the room. If an ignition source is present (e.g., a match or pilot flame), an essentially stationary flame is established, and the location of this flame coincides very closely with the position where fuel and oxidizer exist in stoichiometric proportions.

This type of flame is thus characterized by a high-temperature flame zone in which fuel and oxidizer react rapidly. Fuel approaches the flame from one side and oxidizer from the opposite direction. The physical process by which the reactants approach the flame and mix is diffusion, resulting in the term *diffusion flame* for this type of flame. Diffusion flames can be either laminar or turbulent. In *laminar* diffusion flames, molecular diffusion of fuel and oxidizer is the dominant transport process, and the features of these flames are controlled by molecular quantities. In *turbulent* diffusion flames, the mixing is dominated by the fluid mechanics of the combustion environment.

In addition to gas stoves, some other familiar diffusion flames include furnaces, gas turbines, and large-scale industrial burners. A forest fire and the combustion of a fireplace log are diffusion flames, as are a candle flame and a simple wooden match.

In spite of the ubiquity and enormous economic importance of turbulent diffusion flames, and the fact that they have been studied in one form or another for thousands of years, there is still a great deal of theoretical activity being devoted to these systems, and there is considerable controversy as to their controlling physical and chemical factors. The mathematics of turbulence phenomena is very complex, and these systems are very "stiff" (Gear, 1971)—that is, the characteristic time and space scales vary over many orders of magnitude, from a molecular level to the size of a gas furnace, and from a chemical reaction time scale of a fraction of a microsecond to the time required for a turbulent flame to cross an engine chamber. These problems cannot be resolved with current computational techniques, and a thorough theoretical understanding of turbulent combustion is still not available.

2.7 Spray Combustion and Diffusion Flames

Many hydrocarbon fuels are liquids at room temperature and pressure. Since rapid combustion requires the fuel to mix with oxidizer at a molecular level, it is necessary for such fuels to evaporate at the time of combustion, and the most common means of vaporizing the fuel is to disperse it into the combustion environment in the form of a spray. There are a considerable number of practical combustion systems in which the fuel is initially present in the form of a spray of liquid droplets. For example, in a diesel engine (Heywood, 1988), a liquid fuel spray is injected into the combustion chamber where the fuel spray droplets begin to evaporate. The vaporized gaseous fuel then diffuses away from the droplet surface into the surrounding air. At some distance away from the droplet, the ratio of fuel to oxygen achieves a stoichiometric value, and a flame zone will be established there (see DIESEL ENGINES). Subsequently, this diffusion flame is supplied by fuel evaporating from the surface of the droplet and diffusing out to the flame zone. Of course, in situations where the spray has enough time to vaporize completely and mix with the air prior to combustion, premixed conditions will prevail, but in most practical systems, the sequential processes of evaporation, diffusion, and oxidation all occur during the combustion process.

3. COMBUSTION CHEMISTRY

It is often convenient to write a simple expression for the overall reaction in a combustion system. This was done above as reaction (2) for the combustion of propane in air. However, on a molecular level the actual chemistry of combustion is very much more complex. For example, in the case of methane (CH_4), oxidation begins by removing a single hydrogen atom from the methane molecule. The resulting CH_3 species then can be consumed by any of several possible reactions, leading ultimately to CH_2O, which then produces HCO, CO, and finally CO_2. A complete chemical kinetic description of methane oxidation requires a detailed reaction mechanism with approximately 30 distinct chemical species, including both stable and radical species, and approximately 200 elementary chemical reactions between these species (Warnatz, 1984). For fuels with more carbon and hydrogen atoms, reaction mechanisms can be much larger; for heptane C_7H_{16}, a reasonably complete reaction mechanism must consider 350 chemical species and 2000 elementary chemical reactions (Westbrook *et al.*, 1991).

Combustion chemistry is best understood as a classical chain reaction (Semenov, 1959), in which the chain carriers are radical species. In hydrocarbon systems these radical species include, roughly in order of importance, OH, H, O, HO_2, CH_3, CH_3O_2, and other less significant radicals. In other systems radicals of importance can include NH_2, NH, N, HS, CN, NCO, and many others. Chain-propagation reactions maintain the total number of radical species, such as in the reaction of H_2 with OH:

$$H_2+OH=H_2O+H, \tag{4}$$

which consumes OH and produces H radicals. Termination reactions reduce the size of the so-called radical pool, as in the reaction

$$H+OH=H_2O. \tag{5}$$

In reaction (5) two radical species are consumed and none produced, resulting in a net loss of radicals to the combustion system. Chain termination can also result from nonkinetic processes such as molecular diffusion of radicals away from a flame zone.

Perhaps most important in understanding the chemistry of combustion are the chain-branching reactions, including

$$H+O_2=O+OH, \tag{6}$$

$$CH_4+O=CH_3+OH, \tag{7}$$

and numerous others in which the number of radical species increases. Reaction (6) is of particular significance for hydrocarbon combustion, and many practical combustion phenomena can be understood by the ways they interact with it. Reactions and processes that produce H atoms lead to accelerated rates of combustion because they promote reaction (6), whereas reactions that consume H atoms tend to retard combustion.

For illustration, it is well known that addition of small amounts of halogenated species to flammable mixtures dramatically reduces the rate of combustion and often makes them nonflammable (Westbrook, 1980, 1982). For example, the active chemical in commercial flame extinguishers is often CF_3Br. Chemical kinetic analysis of the influence of such additives show that they inhibit flame propagation through a sequence of reactions such as

$$HBr+H=H_2+Br, \tag{8a}$$

$$H+Br_2=HBr+Br, \tag{8b}$$

$$Br+Br=Br_2, \tag{8c}$$

where bromine is the inhibitor. Most often, HBr is the form in which Br is present as an inhibitor, but even when molecular bromine (Br_2) is used, the above reactions rapidly establish an equilibrium among the Br-containing species so that these reactions proceed in the directions shown here. The net result of these reactions is the conversion of two H atoms into one H_2 molecule, and each H atom that is consumed in this way is therefore not able to react with molecular oxygen in reaction (6) and provide chain branching to the reaction mechanism.

Another illustration is provided by the observation that the laminar burning velocity S_u decreases with increasing pressure. An important part of the analysis of this effect is that a reaction that competes with reaction (6) for H atoms is

$$H+O_2+M=HO_2+M, \tag{8d}$$

where M refers to any so-called third-body species required to conserve energy and momentum in the reaction. Since this reaction depends on the cube of the density and converts three reactants into two products, its rate increases more rapidly with density than the rate of reaction (6). Reaction (8d) is a chain-propagation step, producing one HO_2 radical for each H atom radical consumed. In addition, HO_2 radicals are considerably less reactive than the O and OH products of reaction (6). For these reasons, increasing the rate of reaction (8d) relative to the rate of reaction (6), which happens as the pressure and density increase, results in overall inhibition of the overall rate of combustion.

In some environments, a sequence of several reactions is responsible for chain branching, as in the intermediate-temperature pair of reactions

$$CH_4+HO_2=CH_3+H_2O_2, \tag{9}$$

$$H_2O_2+M=OH+OH+M, \tag{10}$$

which consumes one HO_2 radical and produces three radicals, including one CH_3 and two OH radicals.

In combustion chain reactions, as in nuclear and other chain reactions, a constant level of chain carriers (radical species) results in a steady rate of reaction, whereas a net loss of radicals due to termination will extinguish the reactive system, and a net increase in radical levels due to chain branching can lead to very rapid reaction or explosion.

The reactions in a detailed mechanism fall into several general classifications, consisting of initiation, abstraction, radical decomposition, and radical recombination (Westbrook and Dryer, 1984).

3.1 Initiation Reactions

A fuel molecule can decompose thermally to produce two radical species. Examples of such reactions are

$$C_3H_8=C_2H_5+CH_3, \tag{11}$$

$$C_5H_{12}=C_2H_5+C_3H_7, \tag{12}$$

for propane and pentane. These reactions require strong bonds to be broken and are therefore very slow except under high-temperature conditions. Their primary task in combustion is to provide a small number of radical species, following which other reactions carry out the process of fuel consumption.

3.2 Abstraction Reactions

Most fuel consumption is accomplished in combustion systems through reactions with the radical species listed previously. Radicals remove (abstract) a single H atom from a hydrocarbon fuel species, as in the case of ethane reacting with OH or O:

$$C_2H_6+OH=C_2H_5+H_2O, \tag{13}$$

$$C_2H_6+O=C_2H_5+OH. \tag{14}$$

Abstraction reactions can function as either chain-propagating, as in reaction (13), or chain-branching reactions, as in reaction (14).

There are three types of C–H bonds in hydrocarbon fuels, varying according to the structure of the fuel molecule and termed primary, secondary, and tertiary bonds. In primary C–H bonds, the carbon atom is connected to only one other carbon atom; in secondary C–H bonds, the carbon atom is bonded to two other carbon atoms; and in tertiary C–H bonds, the carbon atom is connected to three other carbon atoms. Primary C–H bonds are strongest and tertiary bonds are the weakest. These can be shown as

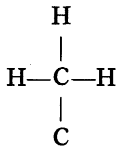

primary
C atom is connected
to only one
other C atom

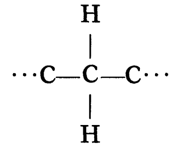

secondary
C atom is connected
to two other
C atoms

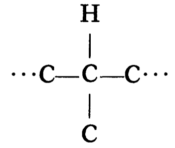

tertiary
C atom is connected
to three other
C atoms

Abstraction of H atoms from primary C–H bonds is slower than abstraction from secondary C–H sites, and both are slower than abstraction from tertiary sites. These distinctions are responsible for variations in ignition and combustion rates among fuels that differ only in their molecular structure. For example, two isomeric forms of butane each consisting of 4 carbon atoms and 10 hydrogen atoms, differing only in the way they are put together, are *n*-butane and iso-butane, shown as

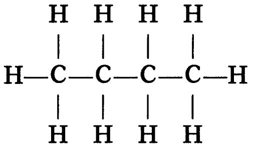

$$\begin{array}{ccccccccccc} & & \mathrm{H} & & \mathrm{H} & & \mathrm{H} & & \mathrm{H} & & \\ & & | & & | & & | & & | & & \\ \mathrm{H} & - & \mathrm{C} & - & \mathrm{C} & - & \mathrm{C} & - & \mathrm{C} & - & \mathrm{H} \\ & & | & & | & & | & & | & & \\ & & \mathrm{H} & & \mathrm{H} & & \mathrm{H} & & \mathrm{H} & & \end{array}$$

n-butane

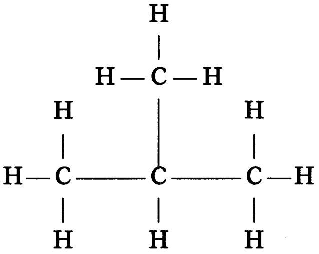

iso-butane

Iso-butane has nine H atoms at primary sites and is therefore more resistant to H-atom abstraction than *n*-butane with only six such primary H-atom sites. As a result, iso-butane ignites more slowly than *n*-butane in combustion environments.

3.3 Radical Decomposition

After an H atom has been abstracted from the fuel, the resulting large radical species will in most cases decompose to produce smaller radical and stable species. The product radical species then decompose further until only small radical and stable species remain. Thus even for large hydrocarbon fuels, this process quickly produces small species with one to three carbon atoms. In this sense, a large fraction of the combustion chemistry of virtually all hydrocarbon fuels eventually is reduced to the reactions of small-fragment hydrocarbon species. Many of these decomposition reactions produce H atoms, including the following in the cases of the ethyl, iso-propyl, and formyl radicals:

$$C_2H_5 = C_2H_4 + H, \qquad (15)$$

$$iC_3H_7 = C_3H_6 + H, \qquad (16)$$

$$HCO = H + CO. \qquad (17)$$

The H atoms that are produced in these decomposition reactions can then react with O_2 via reaction (6) to provide chain branching to the overall reaction system.

3.4 Radical Recombination

Rather than decomposing, radical species can recombine to produce larger, more stable hydrocarbon species. Perhaps the most important of these recombination reactions involves the recombination of two methyl radicals to produce ethane:

$$CH_3 + CH_3 = C_2H_6 \qquad (18)$$

This process reduces the number of radical species by two and provides chain termination, slowing the overall rate of reaction.

3.5 Pollutant Chemistry and Emissions

In addition to desirable products such as heat and radiation, combustion processes also generate unwanted and harmful products. These emissions can include products of incomplete combustion such as carbon monoxide, soot, and hydrocarbon species, as well as combustion by-products such as oxides of nitrogen or sulfur. Finally, there is some concern that an otherwise benign product of complete combustion, carbon dioxide, may be leading to global warming through the greenhouse effect and may ultimately be the most serious combustion problem to be considered.

Incomplete combustion results either when insufficient oxidizer was supplied to the fuel or when the fuel-oxidizer mixture was given inadequate time to complete its reaction. Fuel-rich hydrocarbon combustion produces carbon monoxide, even with ample time to achieve chemical equilibrium, and some CO is produced even under stoichiometric and lean conditions. In nonpremixed systems, if the time for mixing of fuel and oxidizer is not considerably shorter than the residence time in the combustion zone, then a variety of partially oxidized species will result. Coal combustion produces flyash and larger particulates, and many combustion systems produce particulates loosely referred to as soot.

Production and destruction of soot particles in combustion are controlled by both chemical kinetics and transport processes (Harris and Weiner, 1983; Frenklach *et al.*, 1986). Gas-phase kinetic processes lead to production of large aromatic hydrocarbon species that are often termed soot precursors. The chemistry involved in production of soot precursors is a current research topic and apparently involves reactions of both neutral and ionized chemical species, leading to species with molecular weights of nearly 10 000. At this size, these molecules are almost large enough to be considered particles, and they form nuclei upon which small hydrocarbon species begin to collect, eventually growing by many orders of magnitude in mass over time scales of milliseconds.

Soot can be both harmful and necessary. In furnaces and boilers, soot is an essential element in transferring heat of combustion to the metal surfaces that in turn heat the working fluid, usually air or water. Much of the energy released by combustion is absorbed by soot particles in the flame and then radiated from the luminous soot particles to the walls of the combustor. Without soot, the hot product gases are much less efficient in providing this transfer of heat. The soot is intended to be oxidized to CO_2 and water later in the combustion process, downstream from the luminous flame zone. However, soot that is not consumed or is produced unintentionally and then emitted along with the other exhaust gases can be a source of visible smoke, odor, and toxic and carcinogenic chemical species.

Chemical species that are present in the fuel in small amounts often lead to pollutant emissions. The best example of this is provided when liquid hydrocarbon fuels and coal and coal-derived fuels contain quantities of sulfur. Combustion temperatures and the presence of oxygen then convert whatever sulfur is present into oxides of sulfur, particularly SO_2, with small fractions of SO_3, denoted collectively as SO_x. If these species are allowed to escape into the air, they combine with water in the atmosphere and lead to acid rains. To remove SO_x from combustor exhausts, the most common strategy is to provide calcium carbonate or lime with the exhaust and produce $CaSO_4$ that can be collected in the solid phase.

The other common constituent of acid rain consists of oxides of nitrogen or NO_x. Some NO_x is produced from nitrogen that is present in the fuel, as in the case of SO_x above. However, the majority of NO_x production from combustion results from the oxidation of N_2 in the air via the catalytic series of reactions called the Zel'dovich mechanism,

$$N_2 + O = NO + N, \tag{19}$$

$$N + O_2 = NO + O. \tag{20}$$

Since both reactions require breaking strong chemical bonds, especially the triple bond between the two N atoms, these reactions proceed rapidly only at elevated temperatures, and their rates are strongly temperature dependent. Thus most NO_x production in a typical combustion system takes place only in the very highest temperature parts of the system.

Combustors can be modified to reduce the production rate of NO_x by removing or limiting the regions of high temperature. Based on the variability of product temperature with equivalence ratio as shown in Fig. 1, the most common strategy is to carry out the combustion in two or more stages, none of which is stoichiometric with its associated high temperature. In industrial burners and furnaces, often the first combustion zone is fuel-rich ($\Phi > 1.3$), resulting in temperatures considerably lower than the peak value at $\Phi = 1$, together with large fractions of CO and other products of incomplete combustion. Then, downstream from this rich combustion zone, additional air is injected into these gases in what is effectively a fuel-lean combustion zone, completing the oxidation of the fuel and the overall heat release process, but again avoiding high temperature that would promote production of NO_x.

Another strategy used to reduce NO_x emissions is to dilute the reactant gases with inert gases, a process called exhaust-gas recirculation (EGR). The principle involved was noted earlier, where it was observed that product temperatures in fuel-O_2 mixtures are considerably higher than in fuel-air mixtures because the additional N_2 in air dilutes the mixture and must be heated by the reaction along with the chemical products. In exactly the same way, other inert species such as H_2O and CO_2 can also be added to combustible reactants and will result in lower combustion temperatures. These lower product temperatures will then result in lower production rates of NO_x. This process is often very simple to implement, since the combustion exhaust consists of mixtures of N_2, H_2O, and CO_2, all of which are suitable as noncombustible diluents. This strategy is used in virtually all modern automobile engines to lower NO_x production.

Most of these techniques for NO_x control focus on temperature reduction to limit the rate of NO production during combustion. However, it is also possible, under tightly controlled conditions, to consume some of the NO that has already been produced. Several chemical kinetic schemes consume NO and produce N_2 by addition of nitrogenous species such as ammonia (NH_3) (Lyon, 1976), urea, and cyanuric acid (Perry and Siebers, 1986) to the product gases of a combustor. The details of these processes are chemically quite complex and have only recently become reasonably well understood (Miller and Bowman, 1989).

Finally, even the products of chemically complete combustion are now suspected of having serious harmful effects as leading to the greenhouse effect and causing long-term warming of the earth's climate. Worldwide rates of use of hydrocarbon fuels and their resulting production of CO_2 have risen steadily since the Industrial Revolution and are still increasing at present. Some computer simulations of the results of ever-increasing CO_2 emissions suggest a direct connection between such emissions and global warming, although many of the details of CO_2 circulation throughout the earth's land, ocean, and atmospheric systems are still extremely uncertain.

4. PRACTICAL COMBUSTION SYSTEMS

Although at first glance very different from each other, essential elements of most practical combustion systems are often quite similar.

4.1 Internal Combustion Engines

The fuel generally consists of a mixture of liquid hydrocarbons (Heywood, 1988). Both premixed and diffusion combustion operation can be utilized. In carbureted engines, the fuel is vaporized and mixed with air in the carburetor and then drawn into the combustion chamber. This mixture is ignited by means of localized deposition of energy, usually a spark plug. However, other ignition systems can and are used, including plasma and laser discharges. The mixture must be within the flammability limits or ignition and flame propagation are not possible. A premixed turbulent flame then propagates across the combustion chamber. The turbulence is the result of fluid-dynamic instabilities caused by the large Reynolds number of the inlet flow into the combustion chamber. This situation is illustrated in Fig. 4, showing the combustion flame front in a typical engine chamber (Cowart *et al.*, 1990), both from above and in a side view. In the top view, the circle at the left indicates the inlet and exhaust valve, while the circle at the right indicates the location of the spark plug. Heat release from the rapid chemical reactions causes a large pressure increase within

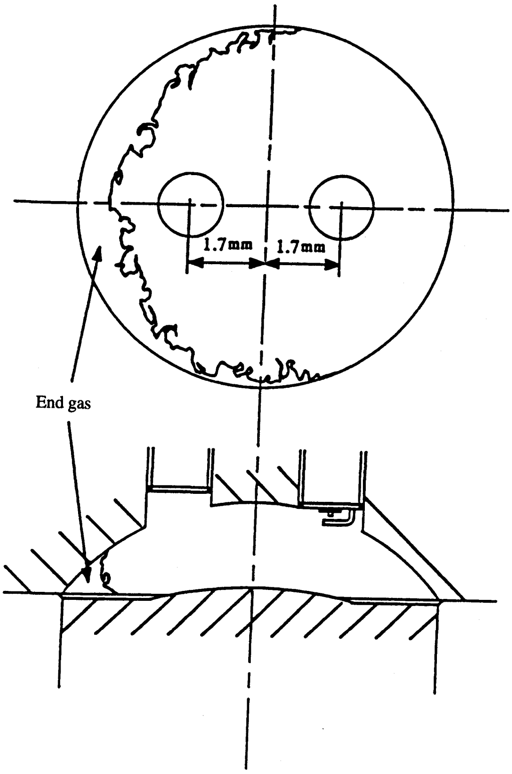

FIG. 4. Side view and top view of an automotive engine chamber, showing the flame propagating away from the spark plug where the spark ignition started the flame. The end gases are also indicated.

the combustion chamber that results in a strong force exerted on the piston, creating torque and driving the motor.

As the flame propagates across the combustion chamber, the Zel'dovich reactions produce NO from the nitrogen in the combustion air. In modern automobile engines, cold exhaust gases, consisting almost entirely of water, carbon dioxide, and nitrogen from previous combustion cycles, are recirculated into the combustion chamber together with the reactant gases, and this dilution, by exhaust gas recirculation (EGR), lowers the combustion temperatures in order to reduce the production rate of NO_x. Carbon monoxide and unburned hydrocarbon species that are present in the exhaust gases pass through a catalytic treatment that completes their oxidation at low temperature, reducing the hydrocarbon emissions without producing further NO_x.

In other cases the fuel is injected directly into the air already present in the combustion chamber. Most often this fuel is normally a liquid and is injected as a spray that then vaporized and mixes with air, creating a mixture in which the local equivalence ratio varies as a function of position in the combustion chamber. If there were sufficient time, eventually all of the liquid fuel would vaporize and the entire combustion chamber would consist of a homogeneous mixture of fuel and air very much like the carbureted engine. However, in the fuel-injected engine, ignition occurs quite early in the cycle, again generally by means of a spark plug. A difficulty in this type of operation is that the spark plug must "see" a flammable mixture when it discharges; otherwise no stable flame will result and an engine misfire will result.

A motivation for this type of engine configuration is its potential for fuel economy and lower NO_x emissions. The quantity of fuel to be injected can be matched accurately to the power needs of the engine, and by confining the fuel to a subvolume of the combustion chamber in which some combustion will take place under rich, and at other places, under lean conditions, the higher temperatures of stoichiometric combustion can be avoided, lowering NO_x production.

Under these conditions all of the flammable mixture burns very quickly, long before all of the remaining fuel spray has evaporated and mixed with air. There is then an extended period during which this remaining fuel burns as a diffusion flame, with fuel and air approaching the relatively stationary flame zone from opposite sides. Otherwise, combustion in this system is very much the same as in the carbureted engine. The heat release drives the piston, NO_x is produced via the Zel'dovich mechanism and controlled with EGR, and unburned hydrocarbons and CO are consumed in a catalyst.

4.1.1 Engine Knock. During the combustion cycle illustrated in Fig. 4, a premixed turbulent flame crosses the combustion chamber. The unreacted gases farthest from the spark plug, called the "end gases," are therefore the last gases to be consumed by the flame. These end gases, shown in Fig. 4, are steadily compressed and heated both by the motion of the piston and by the heat release of the propagating flame. As a result, although they are not in contact with the flame front, these end gases begin to react and, given sufficient time, would spontaneously ignite. Under normal operating conditions, the flame front arrives at the end gases and consumes them, resulting in smooth combustion. However, if these end gases are sufficiently reactive, or if the flame propagation is slow, then the end gases will explode. This explosion, and the pressure wave that results, produce the characteristic ringing phenomenon known as engine knock.

Engine knock can be very destructive if permitted to persist. An example of an engine piston damaged by prolonged knocking is shown in Fig. 5. Therefore, it is desirable to avoid operating conditions that encourage knocking performance. Two of the engine variables that influence knock are the compression ratio and the chemical nature of the automotive fuel. The compression ratio CR is the ratio of the combustion chamber volume at its greatest value [when the piston is at the bottom of its cycle, called *b*ottom *d*ead *c*enter (BDC)] to its smallest value at *t*op *d*ead *c*enter (TDC). The overall efficiency of engine combustion is known to increase steadily with increasing compression ratio, and so there is motivation to operate at as high as possible a value of CR. However, the rate of end-gas ignition also increases rapidly with CR, even-

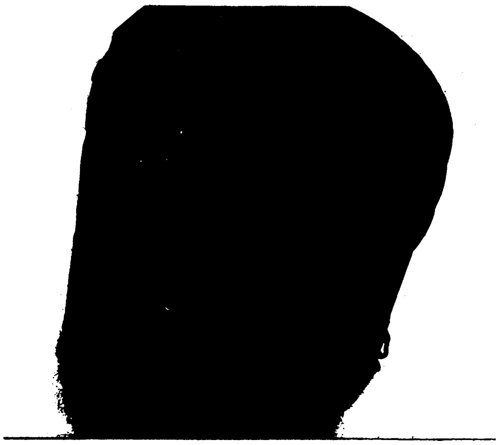

Fig. 5. Photograph of an automobile engine piston which has been severely damaged by engine knock. (Courtesy of P. Jessup, UNOCAL.)

tually providing a limit to the increase in efficiency that could otherwise be achieved. The use of antiknock additives slows the rate of end-gas ignition, permitting the engine to increase its compression ratio further and to achieve a higher practical efficiency.

The chemical nature of the fuel also has a great influence on its tendency to knock (Lovell, 1948). That tendency is measured by the octane number of the fuel, with high values of octane number signifying resistance to knock and low values indicating significant tendency to knock. Straight-chain hydrocarbon species, with large percentages of secondary C–H bonds, have very low octane ratings. For example, the fuel that has, by definition, an octane rating of zero, is the straight-chain species *n*-heptane, indicated schematically as

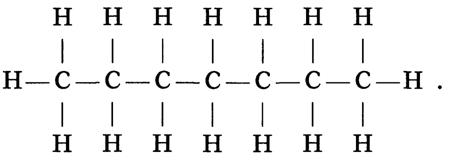

$$\begin{array}{ccccccccccccccccc} & & \mathrm{H} & & \mathrm{H} & & \mathrm{H} & & \mathrm{H} & & \mathrm{H} & & \mathrm{H} & & \mathrm{H} & & \\ & & | & & | & & | & & | & & | & & | & & | & & \\ \mathrm{H} & - & \mathrm{C} & - & \mathrm{C} & - & \mathrm{C} & - & \mathrm{C} & - & \mathrm{C} & - & \mathrm{C} & - & \mathrm{C} & - & \mathrm{H}\,. \\ & & | & & | & & | & & | & & | & & | & & | & & \\ & & \mathrm{H} & & \mathrm{H} & & \mathrm{H} & & \mathrm{H} & & \mathrm{H} & & \mathrm{H} & & \mathrm{H} & & \end{array}$$

In contrast, highly branched fuels with large numbers of primary C–H bonds are extremely knock resistant and have large octane numbers. The fuel that has, again by definition, an octane number of 100, is iso-octane:

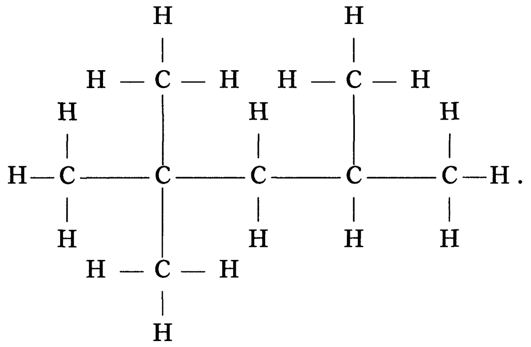

Although the octane numbers and knock tendencies of pure-component fuels have been known for many years and some structural trends related to knocking are recognized, only very recently have these properties been placed on a quantitative and predictive basis (Westbrook *et al.*, 1991). Other factors can influence engine knock, including carbon deposits on engine chamber walls and fluid mechanical details in the flow patterns in the engine, but the above mechanism is the most important factor leading to engine knock in modern spark-ignition engines.

4.1.2 Diesel Engines. From a point of view of combustion principles, the diesel engine is very similar to the spray-injection spark-ignition engine above. One obvious difference is that ignition is accomplished by compression, much like in engine knock, rather than by means of a spark plug. Combustion is believed to take place in a combination of partially premixed, partially diffusion regime. It has been suggested that a diffusion flame exists around each fuel droplet, but it has not been possible to observe the phenomenon in sufficient detail to determine for certain the structure of the combustion zone in a diesel engine.

Diesel engines operate at much higher compression ratios than spark-ignition engines. This is possible because compression ignition or knock is effectively encouraged in the diesel engine. Desirable diesel fuels are denoted with high values of the *cetane* number, indicating good compression ignition tendency. As a result, often fuels with poor or low octane numbers have good or high cetane numbers.

4.2 Burners, Boilers, and Furnaces

A great number of practical combustors consist of stationary systems (as opposed to mobile systems such as automobiles), each of which can produce an enormous amount of thermal energy. These systems heat product gases that then heat a working thermal fluid, usually water, which in turn does the work desired.

All of these systems consist of diffusion flames, although the fuel can be gaseous natural gas, liquid fuel oil, or solid coal particles. Some systems in fact burn mixed fuels such as coal slurries, solid coal particles suspended in a liquid. Most such systems are turbulent, including virtually all of the industrial-scale combustion systems. This is because turbulent systems are able to produce much greater overall rates of heat release per unit volume than any laminar system of comparable combustor volume.

Because the combustion system is large and stationary, it is possible to apply a wider variety of pollution control equipment and technologies to these larger systems than to smaller units such as automobile engines. This includes scrubbing units, precipitators, and other devices to remove SO_x, NO_x, and burned hydrocarbons from exhaust gases.

4.3 Novel Combustion Systems

In addition to the familiar combustors already described, there are many other types of systems that employ the same combustion principles but have different boundary and other operating conditions. In supercritical water oxidation (SCWO), the fuel and oxidizer are dissolved in water at pressures well above the critical point of the water. Under such conditions of approximately 250 atm, water is strictly neither a liquid nor a gas, and such a closed environment has been used to consume toxic waste materials that must not be released into the atmosphere.

Pulse combustion systems have been used for many years, but only recently have the principles of pulse combustors become understood. In such systems, combustion in a chamber forces product gases out through a tailpipe assembly, and a rarefaction wave in the combustion chamber then sucks fresh fuel and air in through flapper valves. These fresh reactants then mix with hot combustion products from the previous combustion event, leading to a spontaneous ignition of the fresh mixture. This then drives a new cycle, and the process repeats itself continuously. Pulse combustors are now used in propulsion systems, in industrial processes, and in home heaters.

4.4 Incineration and Toxic Wastes

There are a wide variety of chemical species that are closely related to hydrocarbon fuels, but in which one or more of the H atoms have been replaced by chlorine atoms. These species, many of which are poisonous or carcinogenic, are used widely in industrial processes. Examples of such species are vinyl chloride, trichloroethylene, and chlorobenzene. When compounds containing these species, such as excess plastic pipe, used transformer fluids, and many others, are no longer needed, there is a need to consume them in an environmentally safe way. A simple method of consumption is combustion, in which the hydrogen and carbon constituents are oxidized to water and carbon dioxide, and the chlorine atoms are converted to HCl. The HCl products can then be dissolved in water and the acid water then concentrated and removed to be treated in a sensible and safe way.

The difficulty with this process of combustion, or incineration, is that all combustion processes are subject to the same problems noted above for simple hydrocarbon combustion. Specifically, incomplete combustion, which in the case of hydrocarbon fuels leads to soot production, in the case of toxic species can produce very harmful toxic wastes such as dioxin and phosgene (mustard gas). Thus the combustion devices intended to consume toxic species are much less tolerant to incomplete combustion than simple hydrocarbon fuels, and the environmental penalties for combustion problems are very severe. The thermochemical product of chlorine atom combustion is definitely HCl, but the necessity of producing only HCl and the opportunities for producing much-less-desirable other products make incineration a problematical science.

5. SUMMARY

Combustion is a part of everyday life, from the production of home heating and home cooking, through the energy required to propel the automobile, the bus, the truck, the train, and the airplane, to the firing of electricity generating plants, industrial boilers, and furnaces. In addition, combustion produces much of our air pollution and other environmental problems.

It is likely that combustion will continue to be an important source of energy and power production for many years, and it is very important to understand the fundamental processes, chemical and physical, that control combustion and its associated processes.

GLOSSARY

Cetane Number: Comparative rating of ignition for fuels under diesel conditions. Higher cetane numbers indicate greater ease of ignition; small cetane numbers indicate a fuel difficult to ignite.

Compression Ratio (CR): Ratio of the combustion chamber volume at its greatest value, at bottom dead center, to its volume at its smallest value, at top dead center.

Detonation: Reactive shock wave.

End Gases: The last fuel-air charge to be consumed in the combustion chamber; the reactants that can explode to produce engine knock.

Equivalence Ratio (Φ): Ratio of fuel concentration actually present to that required to consume exactly the available oxidizer.

Exhaust Gas Recirculation (EGR): Practice of dilution of reactants with reaction products, generally consisting of CO_2, H_2O, and molecular nitrogen, usually employed to reduce product temperatures and slow production of NO_x.

Laminar: Combustion regime in which transport of heat and mass occurs via molecular diffusion and conduction.

NO_x: Nitrogen oxides ($NO+NO_2$).

Octane Number: Relative measure of susceptibility to knocking behavior. Fuels with large octane numbers resist knock; fuels with small octane numbers are likely to knock.

Stoichiometric: Conditions in which the ratio of fuel to oxidizer is exactly correct to produce complete oxidation.

Turbulent: Combustion regime in which transport of heat and mass occurs via unstable large-scale motions of the fluid.

List of Works Cited

Cowart, J. S., Keck, J. C., Heywood, J. B. Westbrook, C. K., Pitz, W. J. (1990), "Engine Knock Predictions Using a Fully-Detailed and a Reduced Chemical Kinetic Mechanism," in: *Proceedings of the 23rd Symposium (International) on Combustion,* Pittsburgh; The Combustion Institute, pp. 1055–1062.

Fickett, W., Davis, W. C. (1979), *Detonation,* Berkeley: University of California Press.

Frenklach, M., Clary, D. W., Gardiner, W. C., Jr., Stein, S. (1986), "Effect of Fuel Structure on Pathways to Soot," in: *Proceedings of the 23rd Symposium (International) on Combustion,* Pittsburgh: The Combustion Institute, pp. 1067–1076.

Gaydon, A. G., Wolfhard, H. G. (1979), *Flames, Their Structure, Radiation, and Temperature,* London: Chapman and Hall.

Gear, C. W. (1971), *Numerical Initial Value Problems in Ordinary Differential Equations,* New York: Prentice-Hall.

Harris, S. J., Weiner, A. M. (1983), "Surface Growth of Soot Particles in Premixed Ethylene/Air Flames," *Combust. Sci. Tech.* **31**, 155–167.

Heywood, J. R., 1988, *Internal Combustion Engine Fundamentals,* New York: McGraw-Hill.

Lewis, B., von Elbe, G. (1987), *Combustion, Flames, and Explosions of Gases,* Orlando: Academic.

Lovell, W. G. (1948), "Knocking Characteristics of Hydrocarbons," *Indus. Eng. Chem.* **40**, 2388–2438.

Lyon, R. K. (1976), *Int. J. Chem. Kinet.* **8**, 315–318.

Miller, J. A., Bowman, C. T. (1989), "Mechanism and Modeling of Nitrogen Chemistry in Combustion," *Prog. Energy Combus. Sci.* **15**, 287–338.

National Advisory Committee for Aeronautics (1959), Report No. 1300, "Basic Considerations in the Combustion of Hydrocarbon Fuels with Air," Cleveland: Lewis Flight Propulsion Laboratory.

Nettleton, M. A. (1987), *Gaseous Detonations, Their Nature, Effects, and Control,* London: Chapman and Hall.

Perry, R. A., Siebers, D. L. (1986), *Nature,* **324**, 657–658.

Semenov, N. N. (1959), *Some Problems in Chemical Kinetics and Reactivity* (trans.), Princeton: Princeton University Press.

Simon, D. M. (1951), "Flame Propagation," *Indus. Eng. Chem.* **43**, 2718–2725.

Strehlow, R. A. (1984), *Combustion Fundamen-*

tals, New York: McGraw-Hill.

Warnatz, J. (1984), "Rate Coefficients in the C/H/O System," in *Combustion Chemistry,* edited by W. C. Gardiner, Jr., New York: Springer-Verlag.

Westbrook, C. K. (1980), "Inhibition of Laminar Methane-Air and Methanol-Air Flames by Hydrogen Bromide," *Combus. Sci. Tech.* **23**, 191–202.

Westbrook, C. K. (1982), "Inhibition of Hydrocarbon Oxidation in Laminar Flames and Detonations by Halogenated Compounds," in: *Proceedings of the 19th Symposium (International) on Combustion,* Pittsburgh: The Combustion Institute, pp. 127–141.

Westbrook, C. K., Dryer, F. L. (1984), "Chemical Kinetic Modeling of Hydrocarbon Combustion," *Prog. Energy Combus. Sci.* **10**, 1–57.

Westbrook, C. K., Dryer, F. L., Schug, K. P. (1983), "Numerical Modeling of Ethylene Oxidation in Laminar Flames," *Combus. Flame* **52**, 299–313.

Westbrook, C. K., Pitz, W. J., Leppard, W. M. (1991), "The Autoignition Chemistry of Paraffinic Fuels and Pro-Knock and Anti-Knock Additives: A Detailed Chemical Kinetic Study," Society of Automotive Engineers Paper No. SAE-912314.

Williams, F. A. (1965), *Combustion Theory,* Reading, Mass.: Addison-Wesley.

Further Reading

Glassman, I. (1987), *Combustion,* Orlando: Academic.

Heywood, J. B. (1988), *Internal Combustion Engine Fundamentals,* New York: McGraw-Hill.

Lewis, B., von Elbe, G. (1987), *Combustion, Flames, and Explosions of Gases,* Orlando: Academic.

Sawyer, R. F. (1981), "The Formation and Destruction of Pollutants in Combustion Processes: Clearing the Air on the Role of Combustion Research," in: *Proceedings of the 18th Symposium (International) on Combustion,* Pittsburgh: The Combustion Institute, pp. 1–22.

Spalding, D. B. (1979), *Combustion and Mass Transfer,* New York: Pergamon.

Strehlow, R. A. (1984), *Combustion Fundamentals,* New York: McGraw-Hill.

Williams, F. A. (1965), *Combustion Theory,* Reading, Mass.: Addison-Wesley.

COMMUNICATION ENGINEERING

See TELECOMMUNICATION

COMPOSITE MATERIALS

CHARLES P. BEETZ, JR., *Advanced Technology Materials, Inc., Danbury, Connecticut, U.S.A.*

Introduction 17
History of Composite Materials 17
Constituent Materials and Properties of Composites 18
1. **Types of Composite Materials** 19
1.1 Fiber-Reinforced Composites 20
1.2 Particulate and Flake-Reinforced Composites 21
1.2.1 Particulate Composites 21
1.2.1.1 Metal Carbide Cermets 22
1.2.1.2 Metal Oxide Cermets 22
1.2.1.3 Dispersion-Strengthened Systems 22
1.2.2 Flake-Reinforced Composites 23
2. **Fibers and Matrix Materials** 23
2.1 Fiber Types 23
2.1.1 Glass Fibers 24
2.1.2 Carbon Fibers 25
2.1.3 Polymer Fibers 27
2.1.4 Inorganic Fibers 28
2.1.4.1 Boron Fibers 29
2.1.4.2 Silicon Carbide Fibers 30
2.1.4.3 Alumina Fibers 30
2.1.5 Whiskers 30
2.1.6 Single-Crystal Fibers 31
2.1.7 Metal Wires 32
2.2 Matrix Materials 32
2.2.1 Polymers 32
2.2.2 Metals 33
2.2.3 Ceramics 33
2.2.4 Carbon 34
3. **Composite Fabrication** 34
3.1 Fabrication Processes 34
3.1.1 Molding Processes 34
3.1.1.1 Contact Molding 34
3.1.1.2 Vacuum-Bag Molding 35
3.1.1.3 Expansion- and Pressure-Bag Molding 35
3.1.1.4 Autoclave Molding 35
3.1.1.5 Injection Molding 35
3.1.1.6 Compression Molding 36
3.1.2 Filament Winding 36
3.1.3 Pultrusion 37
3.1.4 Infiltration and Diffusion Bonding 37
4. **Composite Applications** 37
4.1 Military Aircraft 37
4.2 Aerospace 37
4.3 Commercial Aircraft 38
4.4 Ground Transportation 38
4.5 Recreation and Sporting Equipment 38
5. **Future Outlook for Composite Materials** 39
Glossary 39
Works Cited 40
Further Reading 41

INTRODUCTION

History of Composite Materials

Since the earliest of times it has been recognized that certain combinations of different materials produce a composition whose properties are superior to those of the constituents alone. An ancient example is the use of straw to reinforce mud to produce reinforced bricks for construction purposes. Another important structural composite material was concrete developed and used by the Romans in the 2nd or 3rd century B.C. The Romans developed an extensive knowledge of the chemistry of cementitious limes as recorded in the writings of Vitruvius (Murdok and Brook, 1979). The Roman concrete was a combination of volcanic ash, lime, sand, gravel or crushed rock, and water. This composition, when properly mixed and dried, produced an extremely strong composite material that made possible the construction of truly massive structures. Another advantage of concrete was the ability to cast it into any desired shape to

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

produce the great arches, domes, and vaults characteristic of much of Roman architecture. The great aqueducts, the dome of the Pantheon, and the Colosseum are examples of the large-scale application of man-made composite materials.

The use of concrete declined with the fall of the Roman Empire and did not reappear on a similar scale until the late 19th century, when the addition of steel reinforcing bars to concrete provided a major advancement. The steel reinforcement extended the use of concrete to areas such as beams, columns, and floor slabs subject to bending loads that place the structural element in a state of tension on one side and compression on the other. The tensile load is transferred to the steel reinforcement, which has a much greater tensile strength than the unreinforced concrete.

Another example of an early man-made composite material is paper, invented by the Chinese over 1800 years ago. The early Chinese papers were composed of pulp from the bark of mulberry trees, hemp, and cloth fibers. The pulp and fibers were mixed together in water, precipitated on a cloth screen, and dried to form the paper sheet. More recent examples since the middle of the 20th century include wood laminates such as plywood, directionally solidified eutectics, fiber-reinforced polymers such as fiberglass, and fiber-reinforced concrete, ceramics, and metals.

Constituent Materials and Properties of Composites

Composites encompass an extremely wide variety of constituent materials, as illustrated in the brief historical introduction. This extreme diversity makes it difficult to develop an acceptable definition of a composite material. In this article we will limit the scope of our definition of a composite material to those materials that are man-made by combination of two or more constituent materials to form a third material that has specific properties not possessed by the constituent materials individually. Thus we will not consider, for example, naturally occurring composite materials such as wood. This still leaves an almost unlimited range of combinations of constituent materials that could be considered as composite materials. In spite of this great diversity, the commonly used constituents in composites fall into three basic categories: reinforcements, fillers, and matrix materials. The reinforcing constituents are usually fibers, particulates, or flakes, which provide the principal structural properties of the composite. The matrix material provides the means of encapsulating and binding the reinforcing constituents, giving the composite material its bulk shape and form. Filler constituents are usually added to the matrix material to provide specific engineering, functional, or processing properties.

The properties of the composite material are derived from

1. the basic properties of the constituents,
2. their geometrical arrangement, and
3. the resulting complex interactions between the constituents.

The central idea encompassing all composite materials development is that different constituent materials can be combined to produce a material having specific properties or combinations of properties that are different from those of the constituents. As a simple example, consider a fiber-reinforced composite in which the fiber is unidirectionally oriented in an isotropic matrix material. The tensile strength of the composite measured along the fiber direction will in general be greater than that of the matrix and representative of the fiber, assuming that we use a fiber whose tensile properties are greater than those of the matrix. The matrix transfers the load to the fibers that support the applied load. In structural composites, the fibers form the majority phase, uniformly distributed and separated from one another by a thin layer of the matrix that serves to maintain the fibers in a fixed orientational relationship and to distribute the load uniformly over the fibers. Continuous-fiber–reinforced composite materials are in general anisotropic; properties measured along the fiber direction are different from those measured perpendicular to the fiber direction. An interesting and important aspect of fiber-reinforced composites is the ability to design the composite to make use of the anisotropy and tailor the properties of the composite to carry increased loads along specific directions. The properties of the composite ultimately depend on the constituents and the orientational relationship of the reinforc-

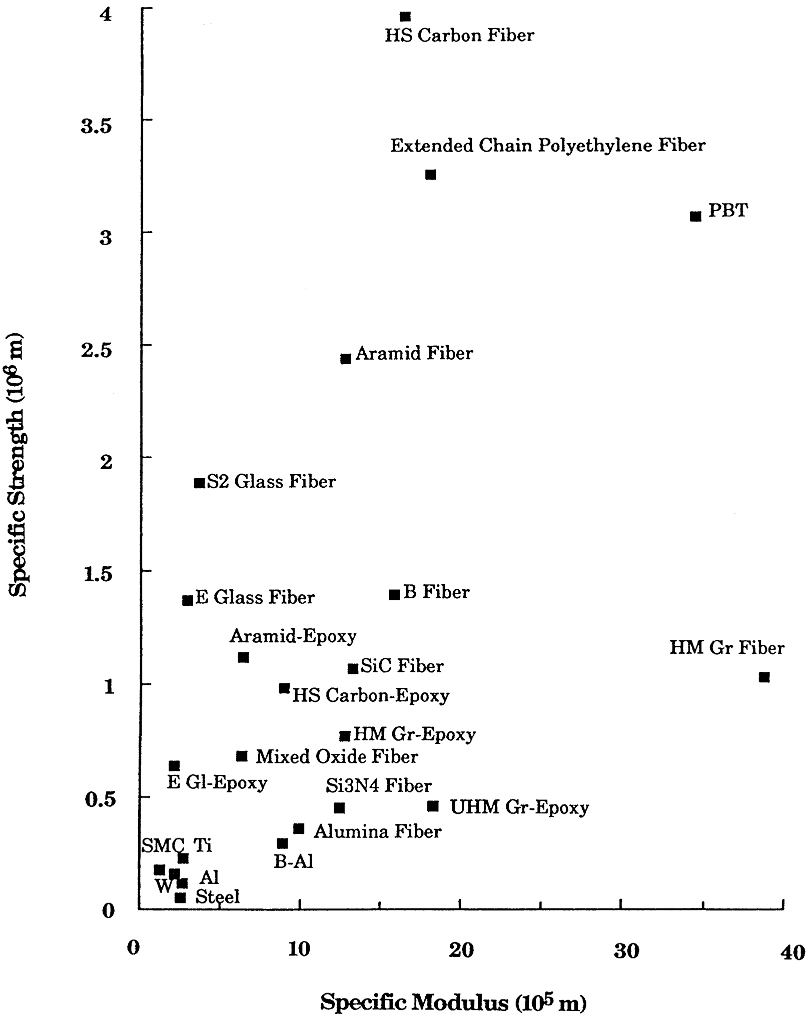

FIG. 1. A comparison of the specific ultimate strength and specific tensile modulus of a number of commonly used structural materials with various uniaxial fiber-reinforced composite materials (abbreviations: HS, high strength; HM, high modulus; UHM, ultrahigh modulus; SMC, sheet molding compound; see text for others.

ing fiber relative to the direction of the applied load. Figure 1 compares the specific strength (ultimate strength divided by density) and specific modulus of elasticity (Young's modulus divided by density) for a variety of commonly used structural and composite materials. Composite materials constructed from light-weight, high-performance fibers and polymeric matrix materials can exhibit specific strength and stiffness 3 to 4 times higher than those of commonly used structural materials such as steel and aluminum.

1. TYPES OF COMPOSITE MATERIALS

Composite materials can be classified (Schwartz, 1984) in several ways, such as by the type of matrix material, by the distribution of the reinforcement, e.g., continuous or discontinuous, or according to the functional purpose such as high-temperature, structural, or conductive composite materials. Composites can be most broadly classified according to the nature of the structural constituent. There are three basic categories of composite materials, as illustrated in Fig. 2. The three

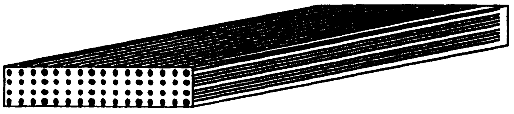

Fiber Reinforced

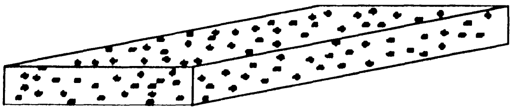

Particulate Reinforced

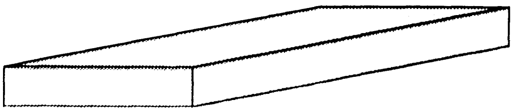

Dispersion Strengthened

FIG. 2. An illustration of the three basic categories of composite materials: fiber-reinforced, particulate-reinforced, and dispersion-strengthened matrices.

categories are (1) fiber-reinforced composites, (2) particulate-reinforced composites, and (3) dispersion-strengthened composites. The effect of the size of the reinforcing agent on the composite strength is illustrated in Fig. 3 for particulate- and fiber-reinforced composites (Broutman and Krock, 1967). In each case very different mechanisms are responsible for strengthening and stiffening of the matrix material. In the case of fiber-reinforced composites, the fiber must have an aspect ratio (length divided by diameter) that is at least ~10 and preferably > 100 to transfer the load from the matrix to the fiber completely. The three categories of composite materials are discussed in the following section and are readily distinguished by the range in size and volume concentration of the reinforcing phase.

1.1 Fiber-Reinforced Composites

In the area of structural composites, fibers are the most important class of reinforcement. This is driven by the necessity to produce ever stronger, stiffer, and more lightweight materials. Fibers provide a unique opportunity to obtain the ultimate mechanical properties of a given type of material. In general, materials when made in fiber form tend to exhibit more nearly intrinsic tensile properties than do bulk forms. The reasons for this will be discussed in Sec. 2. Glass fibers, for example, have much higher tensile strength than bulk glass plates. Similar statements can be made for metal, carbon, ceramic, and polymeric fiber forms. The most weight-efficient method for using the large Young's modulus and tensile strength of fibers is to combine them in a matrix of a lightweight material usually, but not necessarily, having lower tensile strength and modulus. Glass-fiber–reinforced resin is one of the first modern structural composites and has been devel-

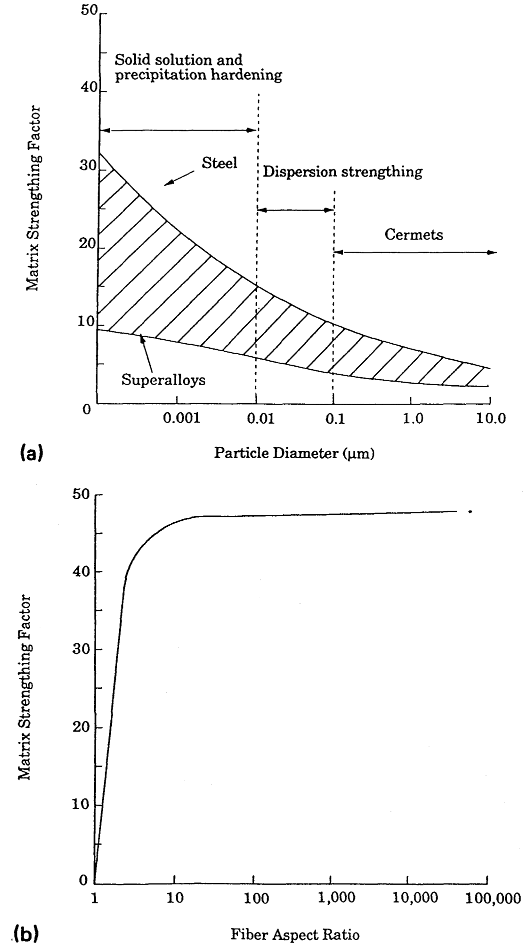

FIG. 3. Effect of the size of (a) particulates and (b) fiber aspect ratio (aspect ratio is defined as the ratio of fiber length to diameter) on the composite strength; after Broutman and Krock (1967).

oping an ever-increasing scope of utilization since the end of World War II. There have been significant developments since the 1960s in the area of carbon fibers and carbon-fiber–reinforced composites, resulting in some of the highest-strength lightweight materials. Carbon fibers have been used to reinforce a variety of polymeric resins and lightweight metals such as aluminum and magnesium. Besides reinforcing lower-modulus matrix materials, there has also been significant progress in using fibers to provide toughness to brittle matrix materials such as ceramics and glasses.

The performance of fiber-reinforced composites is determined by many variables such as

1. fiber orientation, length, shape, and composition,
2. composition of the matrix material,
3. the volume fraction of the reinforcing phase,
4. the interaction between the fiber and the matrix, and
5. the presence of any additional filler materials (Weeton *et al.*, 1987).

The mechanical properties of the composite are highly anisotropic and depend on the orientation of the fiber in the matrix. The strength and modulus of a uniaxially oriented fiber-reinforced composite will be largest along the direction of the fiber. If the applied load is at an angle that is not along the fiber direction, the strength and modulus of the composite will be dramatically reduced. Figure 4 illustrates the directional dependence of the tensile modulus as a function of angle between the applied load and the fiber direction for a carbon-fiber–reinforced epoxy composite. Since in actual applications structural elements are rarely subject to pure uniaxial loads, it is necessary to combine different fiber orientations in a single composite. This is typically done by forming a laminate made up of uniaxial plies oriented at various angles to one another. A typical laminate construction is illustrated in Fig. 5. The ability to tailor the mechanical properties of the composite to accommodate complex loads is a major advantage of composite materials over isotropic materials. The directional selectivity available with composite material design allows for more mass-efficient structures. The stiffness and strength of a uniaxial fiber-reinforced composite are directly proportional to the volume fraction of the fiber in the composite and can be described in a first-order approximation by a simple rule-of-mixtures expression such as

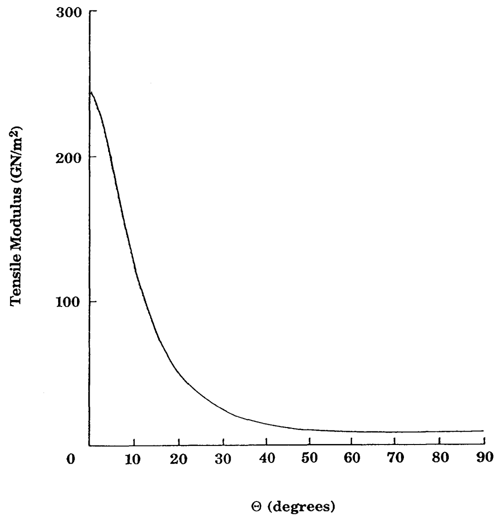

FIG. 4. Variation of the tensile modulus of a uniaxial composite as a function of the angle between the fiber direction and the applied load. After Hull (1981).

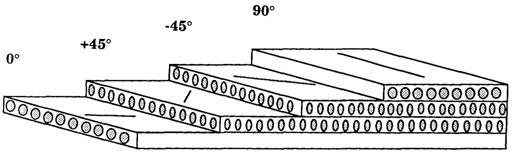

FIG. 5. Example of an angle-plied laminate; the angles indicate the amount of rotation of the fiber orientation in each lamina from the uniaxial direction.

$$E_c = E_f V_f + E_m V_m,$$

where E_c, E_f, and E_m are the moduli of elasticity of the composite, fiber, and matrix, respectively, and V_f and V_m are the volume fractions of the fiber and matrix in the composite. Similar expressions also can be written for the strength, density, Poisson's ratio, etc., of the composite.

1.2 Particulate and Flake-Reinforced Composites

1.2.1 Particulate Composites.

Particulate composites include cermets, metal and ceramic particles in metal, and plastic- and ceramic-matrix materials. Particulate composites differ from fiber-reinforced composites in that the reinforcing constituent does not have a shape anisotropy. The particulate can be viewed in a first approximation as spherical in shape, although in actuality the particulate may be spherical, disk shaped, cubic, or polyhedral. In cases where the particulate is randomly distributed in the matrix and lacks shape anisotropy, the properties of the resulting composite material are isotropic. Since the particulate composite does not have a homogeneous composition, the properties depend on the size, shape, concentration, and composition of the particulate as well as the strength of the interfacial bond with the matrix.

An important class of particulate-reinforced composites intended for use at high temperatures is comprised of cermets, a combination of the words ceramic and metal. Cermets are usually composed of ceramic particles 10%–97% by volume, supported in a metal matrix. Examples are cobalt-cemented tungsten carbide, nickel-cemented titanium carbide, alumina-reinforced chromium, and uranium-plutonium cermets. Cermets are usually produced by powder metallurgical techniques. The bonding between the particulate and the metal matrix is usually due to partial solubility of the two phases in one another. The properties of the cermets are variable over a wide range through variation of the ceramic-to-metal ratio in the starting powder mixture and the processing conditions. There are two commercially important classes of cermets, based on

1. metal carbide and
2. metal oxide particulates.

Many other cermet materials have been produced on a laboratory scale for applications as diverse as selective optical absorbers for solar collector panels.

1.2.1.1 Metal Carbide Cermets. A commonly used metal carbide cermet is cemented tungsten carbide, which finds broad application as a cutting-tool material. The most commonly used binder metal is cobalt. Cobalt-cemented tungsten carbides in general are abrasive and have high compressive strength, hardness, and rigidity. These properties can be varied through changes in the amount of cobalt binder used. Typically the amount of cobalt ranges from 1% to 35% by volume. Increased cobalt content results in increased toughness and ductility and decreased hardness, rigidity, and abrasiveness. Most commercial tungsten carbide cutting tools have small amounts of TiC, NbC, and TaC powders (0%–1%) to improve the high-temperature mechanical properties and the chemical wear resistance of the tool (North, 1987). Typical particle sizes are on the order of 1–10 μm, although new fine-grained tungsten carbide cobalt cermets are being developed that have superior properties. Titanium-carbide-particulate–reinforced nickel or nickel-alloy–matrix cermets have high stiffness and strength at temperatures in excess of 2200 °C.

1.2.1.2 Metal Oxide Cermets. Besides the carbide cermets, metal oxide cermets also find wide commercial application. Oxide cermets find use as high-speed cutting tools, low-friction mechanical seals, nuclear reactor fuel elements, and selective solar absorber coatings. The oxide cermets have the unique property that the metal and the ceramic may be interchanged as either the particulate or the matrix material. A typical metal oxide cermet composition is 72% Cr, 28% Al_2O_3; the properties of the cermet can be controlled by altering the ratio of metal to ceramic. The oxide cermets are typically formed by powder metallurgical manufacturing methods.

1.2.1.3 Dispersion-Strengthened Systems. Dispersion-strengthened composite materials consist of a matrix material that has fine particles of a second phase, typically 0.01–0.1 μm in diameter, dispersed uniformly throughout the matrix. In contrast to the particulate or fiber-reinforced composites, the dispersed particles constitute only a small fraction of the total volume of the composite, usually less than 3%. The strengthening mechanism in dispersion-strengthened systems is very different from that in either fiber- or particulate-reinforced materials. The fine particles reinforce the matrix by restricting the motion of dislocations. An excellent example of a dispersion-strengthened (DS) metal is commercial DS copper, which consists of pure copper that contains fine Al_2O_3 particles (Boyer and Gall, 1985). The effects of varying the Al_2O_3 content on the mechanical and electrical properties of the copper are shown in Fig. 6. The DS copper alloys provide excellent high-temperature stability over pure copper and copper alloys. Other metals in which dispersion strengthening has been successfully employed include

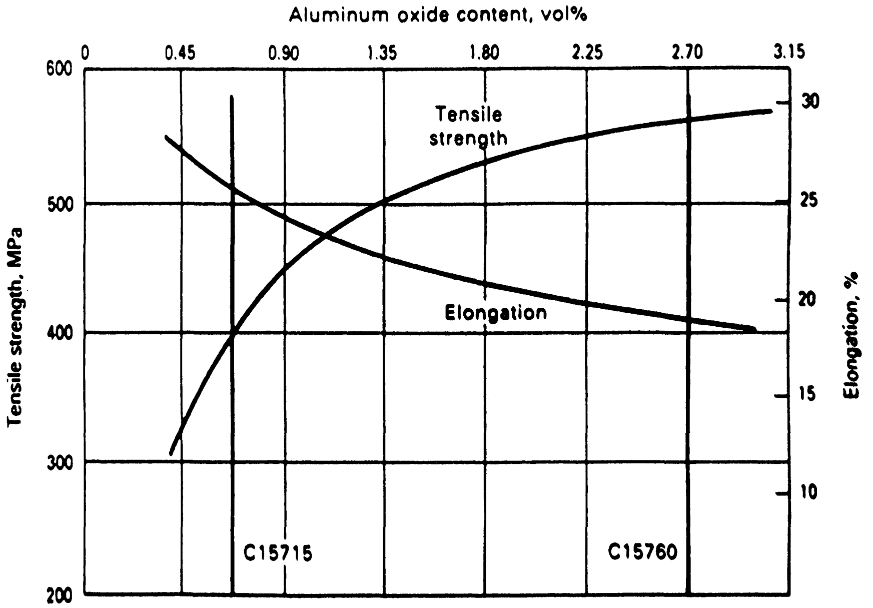

FIG. 6. The effect of aluminum oxide content on the mechanical properties of dispersion-strengthened copper. Reproduced from Boyer and Gall (1985) with permission from ASM International.

sintered aluminum powder alloys, thoria dispersed in tungsten, and thoria dispersed in nickel.

1.2.2 Flake-Reinforced Composites. A special case of a particulate-reinforced composite occurs when the particulate is in the form of a flake. A flake is a particulate that exhibits two-dimensional shape anisotropy, most easily visualized in the form of a plate. In the plane of the flake the shape is isotropic, and perpendicular to the plane of the flake the shape is highly anisotropic, resembling a fiber. The aspect ratio of the flake (diameter divided by thickness), just as in the case of fiber-reinforced materials (Fig. 3), is an important variable that determines how effective the flake is at supporting the load applied to the composite. The two-dimensional shape anisotropy of the flake leads to interesting composite properties provided

1. the aspect ratio is large (~ 100) and
2. the flakes can be incorporated into the composite such that they orient parallel to one another.

An oriented flake-reinforced composite exhibits isotropic mechanical properties in the plane of the flakes, in contrast to fiber-reinforced materials where special laminating or filament-winding procedures must be used to realize in-plane isotropic properties. The two-dimensional shape anisotropy allows the flakes to pack more closely together in the composite, yielding a higher volume fraction of reinforcement and increasing the modulus and strength. The most commonly used flakes are metal, graphite, glass, and mica. Metal flakes are typically used to increase the electrical and thermal conductivity of a polymer matrix material. Mica and glass flakes are used to provide high-temperature insulation properties. Flake-reinforced materials also find use as impermeable barrier materials because of the high packing densities and overlapping of the flakes, preventing fluids and gases from penetrating the composite.

2. FIBERS AND MATRIX MATERIALS

2.1 Fiber Types

A variety of materials can be produced in fiber form and used as reinforcing agents in a composite material. Consequently, fibers come in many sizes, shapes, and structures that usually reflect the manufacturing method. The range of structures commonly found in fibers is illustrated in Fig. 7, ranging from amorphous to single-crystalline to multiphase. Fibers are commonly classified according to material type such as glass, polymer, inorganic, metallic, or carbonaceous. Since many materials can be produced as continuous fibers, this section will describe some commonly used fibers for structural composite materials and recent trends in fibers for high-temperature composite applications.

2.1.1 Glass Fibers. The most frequently used material for manufacturing fibers for structural composite applications is glass. Glass fibers consist primarily of silica (SiO_2) with varying amounts of other metal oxides such as calcium, boron, sodium, aluminum, and iron. There are four major types of glass fiber:

1. A-glass (high alkali content) for good chemical resistance,
2. C-glass (chemical) for extremely good resistance to chemical attack,
3. E-glass (electrical) for high electrical resistivity for insulation purposes (also, because of its strength and resistance to attack by water, it is the most frequently used

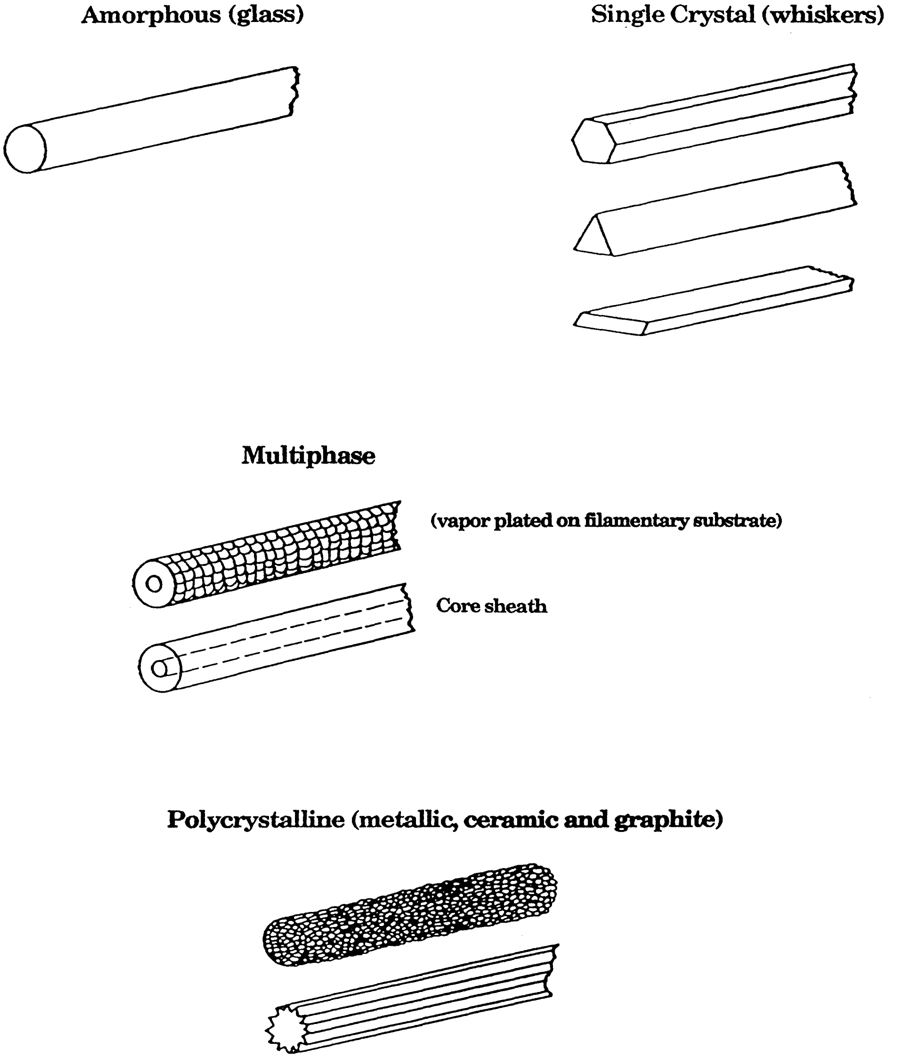
Amorphous (glass)
Single Crystal (whiskers)
Multiphase
(vapor plated on filamentary substrate)
Core sheath
Polycrystalline (metallic, ceramic and graphite)

FIG. 7. Typical structures exhibited by fibrous reinforcements. Reproduced from Weeton *et al.* (1987), with permission from ASM International.

Table 1. Composition of the major types of glass fibers (in percent).

	Glass fiber type			
Constituent	A	C	E	S
Silicon oxide	72.0	64.6	54.3	64.2
Aluminum oxide	0.6	4.1	15.2	24.8
Ferrous oxide	···	···	···	0.21
Calcium oxide	10.0	13.2	17.2	0.01
Magnesium oxide	2.5	3.3	4.7	10.27
Sodium oxide	14.2	7.7	0.6	0.27
Potassium oxide	···	1.7	···	···
Boron oxide	···	4.7	8.0	0.01
Barium oxide	···	0.9	···	0.2
Miscellaneous	0.7	···	···	···

glass fiber for structural purposes), and

4. S-glass for structural applications where higher strength is needed.

S-glass also has higher strength at elevated temperatures. The compositions of the major types of glass fibers are shown in Table 1.

Glass fibers are produced by melting mixtures of the appropriate oxides. The molten glass is extruded under the action of gravity through a heated die containing many small-diameter holes; see Fig. 8. The purpose of the die is to ensure that the fibers have a uniform and well-defined diameter. The extruded glass is drawn away from the die at high speed, resulting in a large number of small-diameter fibers. The fibers are then given a coating of a sizing agent and taken up on a winder. This produces a continuous roving that is used to make yarns, tapes, and various forms of fabrics used for composite material fabrication. Other methods disrupt the flow of molten glass through the heated die using a gas jet to produce short discontinuous staple fibers. Such fibers are used in making fiber mats and nonwoven materials.

Glass fibers produced by spinning processes have an amorphous structure that exhibits no long-range order typically characteristic of crystalline solids. The glass structure consists of a random network of silicon atoms covalently bonded to oxygen atoms. The randomness of this arrangement results in a completely isotropic structure whose properties are independent of direction, i.e., parallel or perpendicular to the fiber axis. The modulus of the glass fiber is primarily determined by the silicate network. Addition of other metal oxides can disrupt or strengthen the silicate network as illustrated in Table 2. Glasses having modulus and strength values comparable to or greater than S-glass have been produced using beryllium oxide additions (Diefendorf, 1985); however, because of toxicity concerns, these glasses have not received wide utilization.

2.1.2 Carbon Fibers.

Over the course of the past 30 years, carbon fibers have gone from being a laboratory curiosity to a major industry. Carbon-fiber–reinforced composites are used in a variety of applications ranging from advanced military fighter aircraft to sporting equipment such as fishing rods, tennis racquets, and golf clubs. Carbon fibers have unsurpassed specific strength and stiffness that make them attractive for applications where mass efficiency is important. Commercially available carbon fibers are produced by controlled pyrolysis, carbonization, and graphitization of one of several precursor fibers:

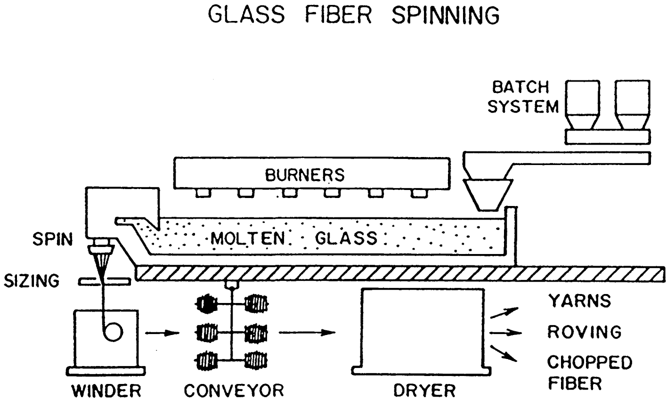

FIG. 8. Schematic diagram of the process used for the production of continuous glass fiber. Reproduced from Diefendorf (1983), with permission.

Table 2. Variation of glass fiber properties with composition (Rauch *et al.*, 1966).

Constituent	E-glass	I	II	III	IV	V
	Composition (%)					
SiO_2	54.3	45.5	45.5	50.7	47.7	42.0
B_2O_3	8.0	19.0	19.0	17.2	14.2	18.2
LiO_2	...	0.5	0.5	0.45	0.3	0.5
MgO	4.7	17.5	10.6	6.30	15.6	12.7
Al_2O_3	15.2	17.5	24.4	25.3	22.2	19.5
BeO	...	...	...	...	...	6.6
Fe_2O_3	...	...	...	...	...	0.5
Tensile Strength (GPa)	3.45	4.05	3.97	4.06	4.36	5.03
Young's Modulus (GPa)	74.5	75.2	62.7	62.1	73.1	82.0
Density (g/cm^3)	2.54	2.33	2.05	2.00	2.20	2.20

1. polyacrylonitrile (PAN),
2. petroleum or coal-tar-pitch–based precursors, or
3. rayon (cellulose).

Of these precursors, PAN is the most widely used for commercial carbon-fiber production. A schematic diagram of the manufacturing process is shown in Fig. 9 for both the PAN- and pitch-based fibers. In the PAN process, a starting fiber of polyacrylonitrile is subjected to a stabilization heat treatment during which the fiber is stretched, orienting the polymer chains parallel to the fiber axis. During the oxidative stabilization process, the linear carbon-nitrogen backbone of the polymer is converted to a ladder polymer of aromatic hexagonal rings. The stabilization process produces a fiber whose glass transition tem-

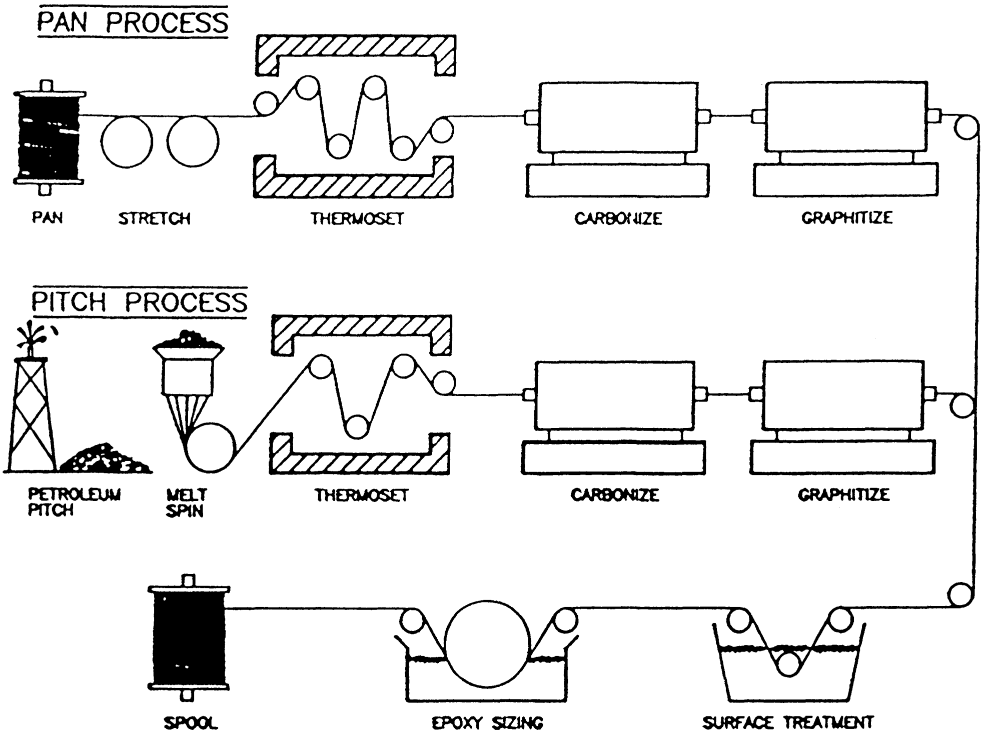
PAN PROCESS
PAN
STRETCH
THERMOSET
CARBONIZE
GRAPHITIZE
PITCH PROCESS
PETROLEUM PITCH
MELT SPIN
THERMOSET
CARBONIZE
GRAPHITIZE
SPOOL
EPOXY SIZING
SURFACE TREATMENT

FIG. 9. Schematic diagram of the processes used for the production of PAN- and pitch-based carbon fibers. Reproduced from Diefendorf (1983), with permission.

perature is well above its decomposition temperature. The stabilized fiber can then maintain the orientation of the ladder polymer structure during the subsequent carbonization process. The carbonization heat treatment drives off hydrogen, oxygen, and nitrogen, leaving behind a partially oriented graphitelike network of carbon atoms. The alignment and crystalline perfection of the carbon network are substantially improved by giving the fibers a high-temperature heat treatment, frequently termed the graphitization heat treatment. Fibers manufactured from PAN precursors in a strict sense do not actually develop a fully three-dimensional graphite lattice until heated to temperatures on the order of 3000 °C.

Pitch-based carbon fibers are produced by conversion of a petroleum or coal-tar-pitch-based precursor fiber into a carbon fiber through a series of heat-treatment processes similar to the PAN process. The pitch-based precursor fiber is produced by first extracting a liquid-crystalline mesophase that is melt-spun into a precursor fiber. The precursor fibers have a high degree of orientation of large platelike aromatic hydrocarbon molecules. This precursor fiber is then subjected to a series of heat treatments analogous to the PAN precursor fiber. The large platelike hydrocarbons become aligned parallel to the fiber axis and develop three-dimensional ordering into a true graphite structure at much lower temperatures than PAN-based fibers.

The Young's modulus of the fiber is highly dependent on the temperature of the graphitization heat treatment. The Young's modulus increases with increasing orientation of the carbon planes and therefore with increasing heat-treatment temperature. Carbon fibers can be produced with a wide range of properties, depending in detail on the process conditions. The higher degree of preferred orientation of the graphite planes parallel to the fiber axis in pitch-based fibers permits pitch-based fibers to develop higher values of Young's modulus at lower values of graphitization heat-treatment temperature.

The tensile strength of the fiber is dependent not only on the degree of orientation of the carbon planes parallel to the fiber axis but also on the number of critical defects present in the fiber volume. Many types of structural defects such as voids, particulate inclusions, misoriented crystallites, and microcracks are present in the fibers as a result of the fiber-manufacturing process. These defects limit the strength of the fiber in much the same way as the weakest link limits the strength of a chain. The tensile strength of the fiber increases, for example, with decreasing fiber diameter as a result, in part, of the reduced number of critical defects present in the smaller fiber volume.

The degree of orientation of the carbon planes parallel to the fiber axis and the concentration of defects present per unit volume can be controlled through variation of the process parameters so that carbon fibers having a variety of properties can be produced. Most carbon fibers fall into one of three different grades:

1. aerospace grade,
2. intermediate modulus, and
3. high modulus.

The Young's modulus and tensile strength of these three grades are given in Table 3.

2.1.3 Polymer Fibers. In carbon fibers, the high degree of preferred orientation of the sp^2-bonded carbon planes parallel to the fiber axis accounts for the extremely large modulus values achieved in these fibers. An analogous situation can be realized using organic polymers, such as paraphenylene terephthalamide (PPT), that can be spun from a nematic liquid-crystalline solution into fibers having a high degree of preferred molecular orientation. In the nematic liquid-crystalline phase, the polymer chains are oriented parallel to one another in local domains. The liquid-crystalline domains are randomly oriented with respect to one another. During spinning, the molecular orientations of the liquid-crystal domains are redirected so that they are parallel to one another, aligning the polymer chains parallel to the fiber axis. The perfection

Table 3. Properties of the three major grades of commercial carbon fibers.

Fiber grade	Modulus range(GPa)[a]	Strain to failure (%)
Aerospace	220–241	1.5–2.0
Intermediate modulus (IM)	275–345	~2.0
High modulus (HM)	379–586	~1.0

[a] Young's modulus.

of the molecular orientations in the fiber can be further improved by a post-spinning heat treatment. The first commercially available fiber based on this process was Kevlar™. The specific strength and modulus of Kevlar 49 fibers are ~5× and ~2× those of steel, respectively. More recently a new grade of aramid fibers has been introduced that has a Young's modulus of 179 GPa and tensile strength of 3.45 GPa at a density of 1.44 g/cm^3 (see Table 4). These fibers have specific properties comparable to those of aerospace-grade carbon fibers. The compressive properties of aramid fibers are somewhat low because of weak intermolecular bonding in the structure that allows it to buckle at low compressive stresses (Diefendorf, 1985). Kevlar-fiber–reinforced composites find use in tensile-dominated loading applications such as rocket motor casings, belts in radial car tires, ropes, and cables. Kevlar-reinforced composites have excellent impact damage resistance and are also used in ballistic protection applications.

Since the discovery of the aramid class of solution-spinnable polymers in the early 1970s, there have been other rigid-rod polymers discovered that also can be solution-spun into highly oriented fibers. These include the aromatic heterocycles such as polyphenylene bisbenzthiazole (PBT) (Helminiak *et al.*, 1975). PBT is a rigid-rod polymer that does not have the kink structure of PPT and therefore produces a fiber that has a substantially larger Young's modulus; see Table 4. Besides solution spinning, other processes have been developed for producing fibers composed of highly oriented rigid-rod molecules such as melt spinning of liquid-crystalline polymers (LCP). This technique has recently been successfully demonstrated on a commercial scale (Vectran™) using thermotropic copolyesters whose properties are comparable to those of the aramid-based fibers (Table 4). The organic fibers offer strength and stiffness intermediate between glass and carbon fibers.

The molecular bonds in linear chain polymers are very strong and when oriented appropriately can produce mechanical properties comparable to those of steel. Calculations of the modulus of fully extended polyethylene, for example, have predicted values ranging from 180 to 285 GPa (Lindenmeyer, 1976). Recently very-high-modulus extended-chain polyethylene fibers have been manufactured using a solution-spinning process. These fibers have a Young's modulus of 172 GPa, tensile strength of 3 GPa, and a density of 0.97 g/cm^3, in good agreement with the early predictions. The key step in the production of this type of fiber was developing a spinning solution that kept the long-chain polyethylene molecule extended, preventing it from folding back on itself. These new fibers are being used in applications such as ropes, cables, ballistic fabrics, and sporting equipment.

2.1.4 Inorganic Fibers. This section describes the major types of ceramic fibers that can be used to reinforce metal, glass, and ceramic composite materials. These fibers are usually intended for use in applications requiring high temperature stability (mechanical and chemical) that cannot be realized using polymeric composites. Many of the ceramic fibers discussed have also found application in polymer-based composites. Although

Table 4. Properties of several high-performance organic fibers.

Fiber type	Modulus[a] (GPa)	Tensile strength (GPa)	Density (g/cm^3)
Kevlar™ 29[b]	58	3.62	1.44
Kevlar 49	124	3.62	1.45
Kevlar 149	179	3.45	1.44
PBT	304	2.7	···
LCP (Vectran™)[c]	62–86	2.5–3.1	1.40
Extended-chain polyethylene (Spectra 1000™)[d]	172	3.09	0.97

---

[a] Young's modulus.
[b] Kevlar is a registered trademark of E. I. duPont de Nemours & Co., Inc.
[c] Vectran is a registered trademark of Hoechst Celanese Corp.
[d] Spectra 1000 is a registered trademark of Allied Signal, Inc.

many different types of ceramic fibers have been produced over the years, only the three most commonly used fibers will be discussed. The most frequently used ceramic fibers are boron, silicon carbide, and alumina. The reader is directed to Bracke *et al.* (1984) and Sowman (1988) for a more complete discussion of past work and future trends in ceramic fibers and composites.

2.1.4.1 Boron Fibers. Boron is an attractive candidate material for producing in fiber form as it has an atomic weight less than carbon, a Young's modulus about twice that of steel, and a density of 2.34 g/cm^3. The preparation of fibers from refractory materials such as boron via direct melt spinning as in the case of glass fibers or by pyrolysis of a precursor fiber as in the case of carbon fibers is usually very difficult. Consequently, commercial boron fibers are produced by chemical vapor deposition (CVD) of a thick boron coating onto a continuous core of tungsten wire. The boron fiber is itself a composite consisting of a small-diameter tungsten wire core coated with a thick layer of polycrystalline boron. Boron fibers are produced as a monofilament instead of a tow of many small fibers. The diameter of the tungsten wire core is typically ~20 μm and the final diameter of the boron filament is variable but is usually ~100–140 μm. A schematic diagram of the CVD process for making boron fibers is shown in Fig. 10. Boron is deposited by thermal decomposition of a boron-containing gas such as boron trichloride on the tungsten wire that is resistively heated by passing a current through the wire. The thickness of the boron coating increases as the wire is drawn through the reactor tube. The boron coating is built up to the desired thickness by passing the wire through many reactors placed in series. The boron fibers produced by this method have a Young's modulus of 400 GPa and tensile strength of 3.6 GPa with a density of 2.57

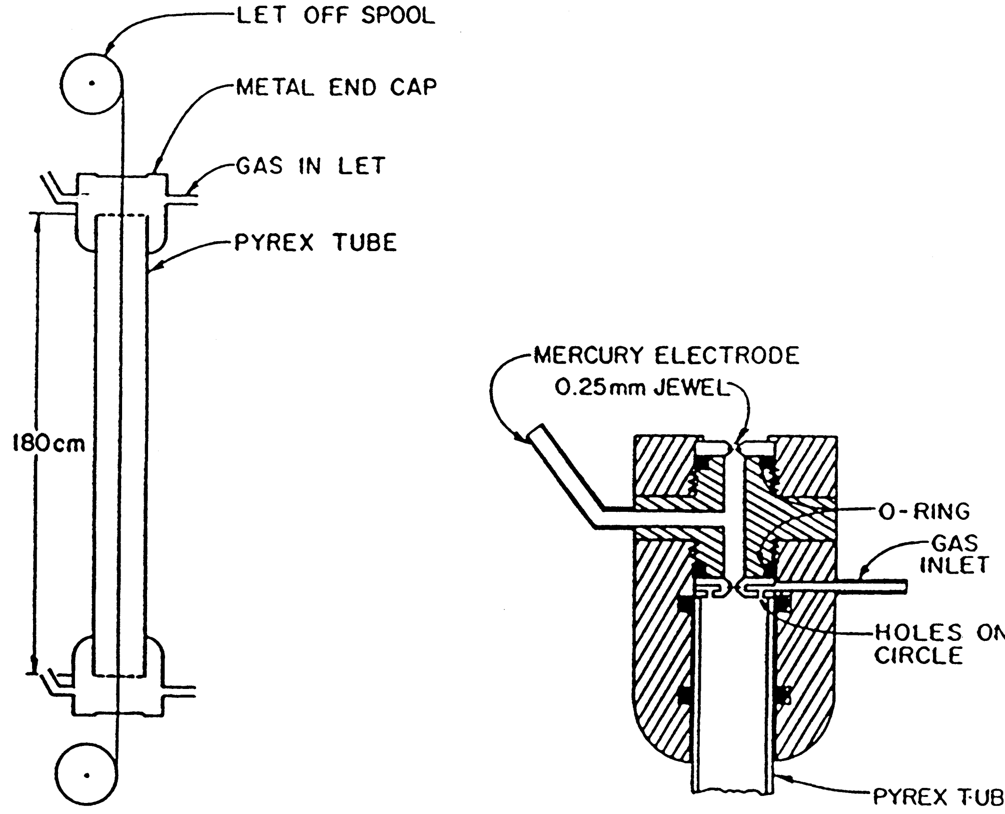

FIG. 10. Apparatus for chemical vapor deposition of a boron fiber onto a tungsten wire. The wire is resistively heated to temperatures high enough to decompose a boron-containing gas such as BCl_3 in hydrogen to deposit elemental boron on the tungsten wire. Reproduced from Diefendorf (1983), with permission.

g/cm^3. One of the largest applications for boron fibers is in reinforcing epoxy matrices for sporting goods such as lightweight racquets and shafts for golf clubs. Boron is also used as a reinforcement in aluminum, magnesium, and titanium metal matrix composites.

2.1.4.2 Silicon Carbide Fibers. Silicon carbide fibers can be produced by CVD on a core carbon monofilament or by pyrolysis of a precursor fiber. Silicon carbide fibers prepared by CVD are most frequently deposited on a carbon monofilament using a process very similar to that for CVD boron fibers. The SiC coating proceeds by reaction of a mixture of chlorinated alkyl silane compounds with hydrogen on the surface of the resistively heated carbon core. The carbon monofilament is $\sim$50 μm in diameter, much larger than the tungsten wire used for boron fibers. The CVD SiC fibers have a Young's modulus of 400 GPa, tensile strength of 3.45 GPa, and a density of 3.3 g/cm^3.

Silicon carbide fibers also can be produced by pyrolysis of Si-containing organometallic polymers such as polycarbosilane. The polymer is first melt-spun into a fiber that undergoes a curing process, which produces a high degree of cross-linking; the cured fiber then is pyrolyzed forming the product SiC fiber. There are many difficulties in the curing process such that easily spinnable high-molecular-weight Si polymers cannot be used because during the curing step they produce structures that result in breakup of the fiber structure. Thus the polymers used contain elements other than Si and C such as nitrogen and oxygen that introduce stable compounds, resulting in a nonstoichiometric silicon carbide fiber. The properties of these fibers are significantly lower than CVD-produced fibers, i.e., Young's modulus $\sim$196 GPa, tensile strength $\sim$2.9 GPa, and density $\sim$2.55 g/cm^3.

2.1.4.3 Alumina Fibers. Fibers of polycrystalline Al_2O_3, Al_2O_3-SiO_2, and Al_2O_3-SiO_2-B_2O_3 mixtures are commercially produced using slurries or sols containing the inorganic phases, various salts, and polymers to control spinnability of the fiber. The solution-spun fibers are given a pyrolysis heat treatment that nucleates the crystallization of the oxide ceramic phase and concurrently removes the polymers or other binders. On further heat treatment, the ceramic particles sinter together, densifying the structure; increased grain growth then follows. The fibers produced in this manner have a very fine grain structure, with fibers produced from sols having the finest microstructures. The properties of fiber FP (Al_2O_3) produced by DuPont, using a slurry-modified sol process, are Young's modulus $\sim$380 GPa, tensile strength $\sim$1.4 GPa, and density 3.9 g/cm^3. Although the tensile strength of these fibers is low, they have excellent compressive properties. The fiber FP also has good stability up to the temperature of 1000 °C and can be processed in aluminum- or magnesium-matrix materials.

2.1.5 Whiskers. It has been known since the early 1950s that materials in whisker form can exhibit superior mechanical properties over bulk forms. Whiskers are small, elongated, single crystals that are almost free of macroscopic defects common to bulk forms of the material. The strength of whiskers increases with decreasing whisker diameter: at diameters less than $\sim$1 μm, whiskers exhibit strengths approaching theoretical estimates; above $\sim$100 μm, whiskers exhibit strengths comparable to bulk forms. The nearly ideal mechanical properties of small whiskers have made them attractive for composite material applications, especially since whiskers of many interesting materials can be grown that have reasonable aspect ratios. Many materials have been grown in whisker form and evaluated in one way or another for composite-material applicability; in this section we will focus on the most promising whisker materials. The reader is directed to Evans (1972) for a review of much of the early work on whiskers.

Primarily for economic reasons, the only commercially available whisker reinforcements are composed of silicon carbide. The SiC whiskers are produced by the controlled pyrolysis of rice hulls that inherently have a high silicon content. SiC whiskers have also been produced by the vapor–liquid–solid (VLS) process in which a solid catalyst particle is placed in a growth reactor and exposed to the reactive gases H_2, CH_4, and SiO. The catalyst particle becomes saturated with Si and C and forms a molten ball that mediates the growth of the whisker through the liquid–solid interface. The crystal grows by precipitation from the supersaturated liquid. The growing crystal is fed from the vapor by the reaction of the constituent molecules at the vapor–liquid interface. β-SiC whiskers having tensile strengths over 8 GPa and a Young's modulus of 581 GPa have been reported (Petrovic *et al.*, 1985). The growth mechanism of the rice-hull–produced SiC whiskers is a variant of the VLS process.

Another whisker material that has recently shown promise as a potentially low-cost reinforcement for polymer-based composites is carbon (Tibbetts *et al.*, 1986). Carbon whiskers can be produced by a vapor growth process that uses a small-diameter iron catalyst particle to grow the whisker. The iron particle is exposed to a mixture of CH_4 in H_2 at elevated temperatures, saturating the iron particle with carbon. The carbon then precipitates from the particle as a hollow graphite tube approximately the diameter of the catalyst particle. The graphite tube produced by this process has the structure of high-quality graphite and is extremely small, ~30 nm in diameter, and difficult to handle. The length of the hollow graphite tube is dependent on the processing conditions, but tubes tens of centimeters in length have been produced. To produce larger-diameter whiskers, a chemical vapor deposition step is carried out that increases the diameter to the desired value, typically ~10 μm. This step represents a departure from the traditional whisker process, and the structures produced are more appropriately called fibers. The properties of the resulting CVD-coated graphite tubes are highly dependent on the final diameter but typically fall in the range of aerospace-grade carbon fibers. The CVD carbon coating is highly imperfect as deposited and accounts for the reduction of the fiber properties from those expected for single-crystal graphite. However, high-temperature heat treatment of the fibers produces properties approaching those of single-crystal graphite. Fibers produced by this method have been shown to have thermal conductivities comparable to single-crystal graphite and should be useful for composite applications requiring this property.

2.1.6 Single-Crystal Fibers. An outstanding deficiency of currently available ceramic fibers is creep at high temperatures. Above ~1000 °C, the structure of polycrystalline Al_2O_3 and SiC fiber changes, causing slippage along grain boundaries, greatly reducing the physical properties of the fiber. Many of the mechanisms responsible for high-temperature creep, such as grain growth and slippage at grain boundaries, can be eliminated through the use of single-crystal fibers. Single-crystal fibers can be produced by many different techniques, such as growth from solutions, melts, and the vapor phases. Production of single-crystal fibers is difficult, and it is likely that only three techniques will be exploited in the future: controlled drawing from the melt, vapor growth, and zone refining.

The most versatile method for producing single-crystal fibers is the recently developed laser-heated pedestal-growth (LHPG) technique (Feigelson, 1988). This is a variation on the float-zone pedestal-growth process. In the LHPG technique, a source rod having the desired composition is heated by a laser focused onto one end of the rod, creating a molten zone from which the fiber is pulled. The process does not employ a crucible. The position of the laser-heated zone is fixed in space while the source rod is translated into the laser beam at a rate necessary to conserve the melt volume. The principal advantage of this process is the ability of the laser to produce extremely sharp temperature gradients (~1000 °C/cm) that account for the high growth rates, from millimeters to tens of millimeters per hour. In addition, the steep temperature gradients, fast growth rates, and high surface-to-volume ratio of the fibers imply that the fibers cool very rapidly. This helps to suppress high-temperature solid-state phase transformations and thus can be important for stabilizing metastable forms of the material for nonstructural applications. Another of the major advantages of this technique is the complete absence of hot furnace elements that can cause contamination of the fiber. The field of single-crystal fiber growth is relatively new and has been primarily focused on fibers for optical applications. There is still need for much more work to be done relating processing conditions to defects in the crystal structure and how these defects influence fiber properties. However, the LHPG technique has demonstrated potential for producing high-quality single-crystal fibers. Czochralski fibers, by comparison, suffer from low pull rates, and therefore, without significant process improvements, it is highly likely that fibers produced by this method will be very expensive.

2.1.7 Metal Wires. Metal wires have also been used to reinforce polymer, metal, and ceramic matrices. Wires are typically made by drawing a rod of the metal through a series of dies of decreasing diameter; typical diameters range from several millimeters to ~10 μm. Wires of steel, tungsten, molybdenum, and niobium and alloy wires such as tungsten-rhenium-hafnium-carbon have been evaluated for the reinforcement of superalloy materials such as iron-, nickel-, and cobalt-based alloys. The reader is guided to Jones (1989) for a review, a compilation of properties, and a comment on the status of wire-reinforced composites.

2.2 Matrix Materials

The purpose of the matrix phase is varied, depending on the intended use of the composite structure. The role of the matrix material in structural composites is to distribute and transfer the applied load to the reinforcement. Even this statement needs to be clarified, as it applies only to those composite systems in which properties of the reinforcement such as tensile strength and modulus are substantially greater than those of the matrix. Examples of such systems are carbon- or glass-fiber–reinforced polymers. Here the strength and modulus of the composite are dominated by the properties of the reinforcement. At the other extreme are systems such as ceramic matrix composites that seek to use the low-density, high-temperature, and environmental stability of bulk ceramic materials while employing a second-phase fiber reinforcement to thwart crack propagation and prevent catastrophic failure modes common to bulk ceramic materials. In both cases, the interfacial reactions between the matrix and the reinforcement are critical to developing the expected properties. Most fibers, for example, are given a special surface treatment to facilitate the interaction of the fiber with the matrix material. In some metal- or ceramic-matrix composites, a thin coating is applied to the fiber to prevent the matrix from aggressively attacking the fiber and degrading its properties. The physical and chemical nature of the fiber surface and how it interacts with the matrix both during the fabrication of the composite and during the actual in-service use determine how effectively the reinforcement is utilized and governs the primary performance of the composite; see Jones (1989), Metcalf (1974), and Kerans *et al.* (1989).

Besides the reinforcement aspects, matrix materials also must provide a whole set of process-oriented properties that permit them to be readily fabricated into a composite. These are properties such as matrix viscosity that permit infiltration of the fiber tows, processing temperatures that are compatible with the thermodynamic stability of the reinforcement, and interaction with and bonding to the reinforcing phase. These comments apply to all four basic classes of matrix materials—polymers, metals, and ceramics including glass and carbon.

2.2.1 Polymers. Polymers represent the most commonly used matrix materials. This is due to their low cost, low density, and acceptable mechanical properties, and the relative ease with which they can be processed. Polymer matrices are readily separated into two types, thermoplastic polymers and thermosetting polymers, depending on their thermorheological behavior.

Thermoplastic polymers have a relatively well-defined melting temperature and exhibit reversible softening or melting behavior. Thermoplastic matrices are processed by traditional polymer processing techniques such as compression molding and extrusion. Examples of traditional thermoplastic polymers are polyethylene, polypropylene, nylon, and polystyrene. These polymers have not been extensively used in structural composite applications, primarily because of creep at elevated temperatures. Thermoplastic matrices are commonly used in noncritical structural applications such as equipment housings and instrument panels in automobiles. Recently, polymers such as polyetherimide, polyphenylene sulfide, and polyetheretherketone have been explored as high-temperature thermoplastic matrices. The excellent temperature resistance and superior impact strength exhibited by these polymers should allow them to be used in higher-temperature structural applications (National Materials Advisory Board, 1987).

Thermosetting polymers do not exhibit the reversible melting behavior of thermoplastics. On initial heating or curing, thermosets develop extensive cross-linking of the polymer chains, and on subsequent heating, the polymer will decompose before melting. The rheological properties of the uncured polymer and the conditions necessary for initiation of the cross-linking reactions can be varied by changing the basic formulation of the thermoset. This is especially important for impregnation of fiber bundles and tows before curing. An interesting property of thermosets that makes them very useful for composite-material fabrication is their ability to be placed in a state of partial cure and held in that condition for long periods of time before being fully cured. This ability to be carried to a stage of partial cure gives the impregnated fiber sheet mechanical properties similar to those of a fabric, facilitating fabrication of complex laminated structures. Typical thermosetting polymers are epoxy, polyester, and phenolic polyimide resins. Thermosetting polymers are widely used in structural composite applications such as military and civilian aircraft, boat hulls, and automotive parts.

2.2.2 Metals. In contrast to polymers, metals offer greater strength and stiffness and higher use-temperature capability. The major purpose for fiber and particulate reinforcement in metals is to improve their high-temperature mechanical properties. The most investigated metals are aluminum, magnesium, titanium, and copper, although there has also been a considerable amount of research on superalloy matrices such as nickel-based, cobalt-based and iron-nickel–based alloys (Islam *et al.*, 1988). These metals, with the exception of copper and the superalloys, are primarily chosen for their low densities. The most commonly used reinforcements are boron, silicon carbide, alumina, and carbon fibers. The fabrication of fiber-reinforced metal requires high temperatures for melting or diffusion bonding of fiber-interleaved metal foils. The interaction of the molten or hot metals with the fiber can significantly degrade fiber properties. Special diffusion-barrier coatings may be required, such as SiC coatings on boron fibers for use in titanium matrices. Similarly, carbon fibers require a protective coating before they can be used in the fabrication of aluminum-matrix composites.

The introduction of a continuous fiber into a metal matrix results in an anisotropic composite whose strength and modulus are greater along the fiber axis than in the transverse direction. Thus continuous-fiber–reinforced metals cannot be readily fabricated using traditional metal-forming machinery. The use of SiC whiskers or particulate reinforcements in aluminum has resulted in metal-matrix composites that are more nearly isotropic. A commercial application of metal-matrix composites is the use of alumina fiber to reinforce and provide superior wear resistance to the upper piston-ring groove region of an aluminum diesel engine piston (U.S. Congress, Office of Technology Assessment, 1988).

2.2.3 Ceramics. Bulk ceramic materials have excellent performance characteristics, such as superior retention of strength and stiffness at high temperature, high hardness, low density, and excellent oxidation resistance. The problems with bulk ceramic materials are their poor tensile strength, poor thermal shock resistance, and poor impact resistance, significantly limiting their scope of applicability. The development of ceramic-matrix composite materials is an attempt to impart toughness to these brittle materials through the use of fibrous and particulate reinforcements to alter the failure mode of the ceramic. The inclusion of fibers and particulates into ceramic matrices can alter the failure mode and toughen the ceramic in one of several ways (Diefendorf, 1985):

1. by increasing the local driving forces necessary to propagate a crack to failure,
2. by increasing locally the amount of mechanical energy per unit area necessary to propagate a crack, and
3. by decreasing the local strain in the matrix by cracking along fiber-matrix interfaces that reduces the stress concentration, preventing catastrophic failure.

In most structural composites, the fibers are needed to provide improved stiffness and strength. In many ceramic matrices, however, the modulus of the matrix is comparable to that of the fiber. This creates a situation in which the strain necessary for transferring the load from the matrix to the fiber is larger than the strain to failure of the matrix; thus addition of reinforcing fibers does not always result in improved properties. The addition of fibers to improve the brittle fracture characteristics of the ceramic may be viewed as the introduction of a homogeneous population of defects that impede crack formation and propagation and may require compromising other properties of the ceramic such as tensile strength. Many ceramic materials have been explored as matrices for composite structures (Schioler and Stiglich, 1986), including glass, silicon carbide, silicon nitride, alumina, mullite, and cements. One system that has demonstrated encouraging results is the use of carbon and silicon carbide fibers to reinforce glass and glass ceramic matrices such as borosilicates and aluminosilicates (Prewo, 1982). The glass and glass ceramic matrices have an important advantage over many other ceramic materials in that their viscosity can be substantially reduced by heating above the softening point so that the matrix can flow and fully impregnate the fiber bundles, producing fully dense composite structures.

2.2.4 Carbon. Almost since the introduction of carbon fibers as reinforcements for composite materials, carbon-fiber–reinforced carbon matrix (carbon/carbon) composites were envisioned (Stoller *et al.*, 1974). Carbon as graphite has excellent high-temperature properties. The high strength and stiffness of graphite are maintained to temperatures well in excess of 2700 °C. Applications such as rocket nozzles and space reentry vehicles, which experience very high service temperatures, have made great use of carbon/carbon composites. A major area of research activity in carbon/carbon composites is environmental protection. Carbon/carbon composites are susceptible to oxidative erosion when heated in oxygen-containing environments and require a protective coating such as silicon carbide.

Carbon/carbon composites are usually fabricated by liquid infiltration of a fiber preform with coal-tar or petroleum pitches. The infiltrated preform is heated to pyrolyze the pitch, leaving behind the carbon matrix. This process must be repeated several times to reduce the porosity in the carbon matrix caused by the release of volatile degradation products during the pyrolysis of the pitch. Other fabrication methods include vapor infiltration in which a carbonaceous gas infiltrates and decomposes on the heated fiber preform. This approach also has problems with porosity due to incomplete filling of the interior of the preform and sealing off the porous interior regions, as the gas phase is most rapidly depleted of the carbon species on the outermost regions of the preform.

Carbon/carbon composites have also received attention for use in biomedical applications such as artificial implant structures for replacing bone. Carbon/carbon composites are attractive for this purpose because of their light weight, excellent mechanical properties, and chemical inertness.

3. COMPOSITE FABRICATION

3.1 Fabrication Processes

There are many methods for fabrication of composite materials. This section describes the more commonly used processes for fiber-reinforced composites in a generic sense and directs the reader who desires a more in-depth description to specific references. The fabrication of a fiber-reinforced composite structure can be broken down into two basic operations:

1. placement and orientation (lay-up) of the various layers of fibers impregnated with resin and
2. consolidation of the laminate to induce bonding between the fibers and the matrix.

The fabrication of laminated structural composites is very hand-labor intensive, although significant progress has been made in automating the hand lay-up procedures.

3.1.1 Molding Processes

3.1.1.1 Contact Molding. The simplest and one of the most widely practiced forms of composite fabrication is contact molding, more commonly referred to as wet lay-up. This molding technique is more suited for fabrication using fibers in the form of fabrics and is not generally used for unidirectional composites because of limitations with selecting fiber orientation. The basic raw materials for this process are a two-part thermosetting resin system and fibers usually in the form of mats or woven fabrics. In this process, the resin is coated onto or into a mold. Fibers in the form of mats or woven fabric sheets are hand placed over the resin-coated mold, and more resin is applied. The resin is worked into the fiber mats, and air bubbles are removed through the use of squeegees or rollers. This process is repeated until the required thickness is achieved. The fiber-and-resin–coated mold is then covered with a release film to provide a smooth surface to the exposed surface. The lay-up is then allowed to cure either at room temperature or by heating.

3.1.1.2 Vacuum-Bag Molding. Vacuum-bag molding is widely used for fabrication of laminates of simple to complex geometries. The technique is an extension of the wet lay-up process, using a simple vacuum technique to remove air bubbles and to apply a modest hydrostatic pressure to the laminate to aid in consolidation and removal of excess resin. The vacuum-bag process is shown schematically in Fig. 11. Fibers in the form of resin-preimpregnated mats, tapes, fabrics, or tows are placed on a mold that has the desired form or shape. The fibers are placed on the mold in the desired orientation, and a laminate having the desired number of layers is built up; a porous release film is then placed over the laminate and covered with a bleeder fabric that takes up excess resin as the laminate structure cures. A nonpermeable bag material such as a polyvinyl alcohol film is placed over the entire structure and sealed around the edges, and a system of vents or passages is provided so that a vacuum can be drawn in the space between the mold and the vacuum bag. The bagged mold is then placed in an oven while the vacuum is maintained to consolidate and cure the laminate. Curing also can be effected by heating the mold. After curing, the vacuum bag, bleeder cloth, and release film are removed, and the laminate can be removed for further processing such as edge trimming, cutting, or further machining. The reader is guided to Schwartz (1984) for a more detailed discussion of the vacuum-bag process.

3.1.1.3 Expansion- and Pressure-Bag Molding. Pressure-bag molding is a variation of the vacuum-bag molding technique in which the fiber lay-up is constructed on a mold in a manner similar to the vacuum-bag technique. The lay-up is covered with a rubber bag that is pressurized to remove voids and excess resin during the curing step. The process is shown schematically in Fig. 12(a). An interesting modification of this process, shown in Fig. 12(b), is the use of silicone rubber inserts in the mold that expand when heated and apply pressure to the lay-up during the curing cycle. The vacuum fitting removes entrapped air and volatiles evolved during curing of the resin. This technique is frequently used in the fabrication of complex structures.

3.1.1.4 Autoclave Molding. The majority of high-performance structural composites for aerospace applications are produced by autoclave molding. Autoclave molding is also very similar to the vacuum-bag process, the difference being that the resin is cured in an autoclave held at elevated temperature and pressure. The vacuum bag allows the removal of air early in the curing cycle while the high pressure permits a higher degree of consolidation than either vacuum- or pressure-bag molding.

3.1.1.5 Injection Molding. Injection molding is widely used for volume production of short-fiber–reinforced articles. The process involves heating a thermoplastic polymer that contains reinforcing materials such as short fibers, flakes, or particulates and injecting the heated charge into a mold. The mold is cooled so that the polymer material rapidly solidifies. The mold is opened and the part is removed. This process is very flexible and has been

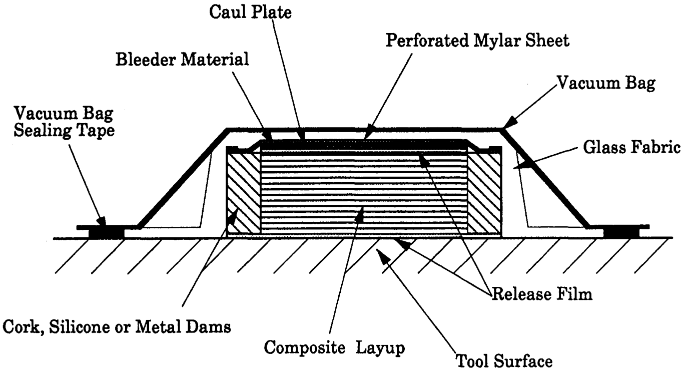

FIG. 11. Schematic diagram of the vacuum-bag molding process. After Schwartz (1984).

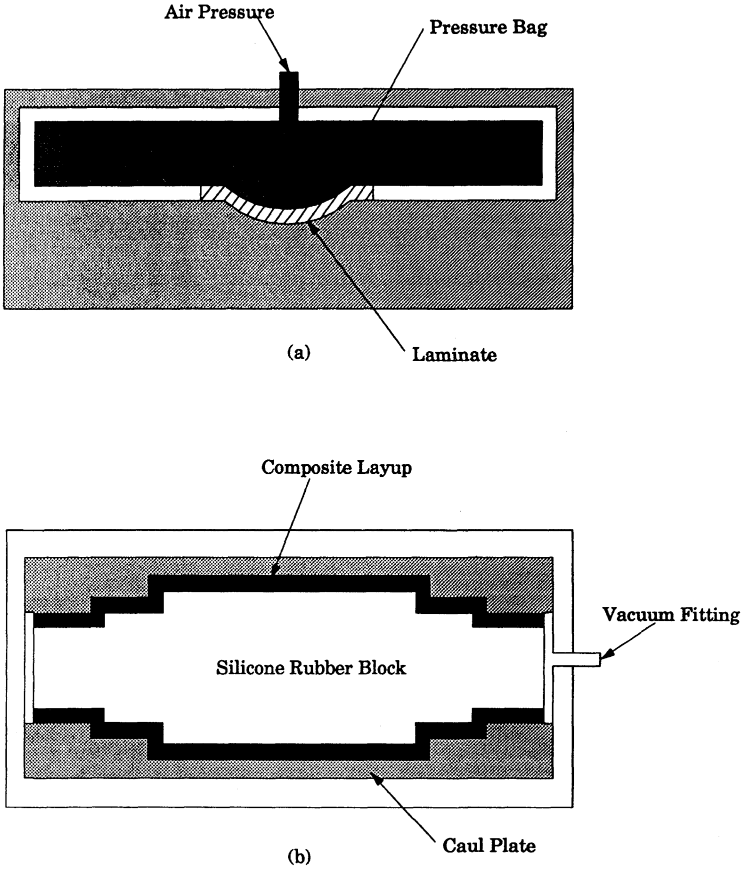

FIG. 12. Schematic diagrams of the (a) pressure-bag and (b) expansion-molding processes. After Schwartz (1984).

adapted to a variety of shapes and sizes. The process cycle time can be very short and therefore attractive for high-volume applications such as glass-fiber–reinforced automotive parts. An interesting modification of the injection-molding process is reaction injection molding, in which a two-component thermosetting polymer such as an epoxy resin is injected into a mold that is prefilled with reinforcing fibers. The two-component resin system is mixed together just before the injection point into the mold. The two components begin to react as they are forced into the mold. On completion of the injection process, the mold is then heated to complete the curing of the resin. This process is very adaptable for complex composite forms that involve the use of several different reinforcements used at selected places in the composite article. The process has been used for limited production of automotive body components and large structures such as boat hulls. The main limitation of the process is the long period necessary to cure the polymer matrix.

3.1.1.6 Compression Molding. This process is primarily used for volume production of fiber-reinforced components such as automotive fascias, fenders, hoods, and trunk lids. The process consists of loading a mold with a charge of a fiber-filled thermosetting resin commonly called sheet molding compound, closing the mold, and forcing the charge to conform to the shape of the mold, followed by heating to effect the cure of the thermoset resin. The cured part is then ejected from the mold and the process is repeated. Typical cycle times are about 30 s to several minutes.

3.1.2 Filament Winding.

The filament-winding fabrication process has received a great amount of attention initially as a method for high-rate production of composite tube structures. The process consists of winding fiber tows or tapes onto a rotating mandrel. Either the winding can be carried out using preimpregnated fibers or the resin can be applied to the fiber by passing it through a resin bath as it is being wound onto the mandrel. In either case, the mandrel is covered with the desired number of plies of fiber and then placed in an oven for curing. The filament winding has been computer-automated to a large extent so that very complex shapes can be produced using multiple-axis filament-winding machines (Schwartz, 1984). The concept of automated filament winding has been generalized to the extent that machines have been produced that can lay down fiber tapes or tow in very complex patterns, removing much of the hand-labor–intensive work and producing composites that have a high degree of accuracy and precision in fiber placement.

3.1.3 Pultrusion. Pultrusion is a technique that was initially developed to produce solid shapes having a constant cross section such as round rod (Schwartz, 1984). In this process, continuous-fiber tows are pulled through a heated die that is the shape of the desired cross section. The resin cures during this heating step and a solid, uniaxially reinforced rod is produced. The major advantage of the pultrusion process is speed.

3.1.4 Infiltration and Diffusion Bonding. These techniques are commonly used for the fabrication of metal-matrix composite structures. In the vacuum infiltration process, molten metal is drawn, through the use of a vacuum, into a mold containing a fiber preform. The fibers are placed in the mold in the desired locations and orientations and the molten metal is allowed to fill the mold. The mold is cooled to solidify the metal after which the composite is removed from the mold. This process is difficult to carry out if selective positioning of the fibers is required so that they are not uniformly distributed in the mold. The process is very dependent on the viscosity of the molten metal and its ability to wet the surface of the fiber to facilitate full impregnation of the fibers to make a fully dense composite.

In the liquid-metal infiltration process the fiber tow is pulled through a crucible containing the molten metal. The infiltrated tow passes through an orifice having the desired diameter, and the metal is cooled quickly on passing through the orifice. The resulting uniaxially reinforced wire is then wound up on a large-diameter spool. The wire can be cut and used at a later time as the raw material in a hot-pressing consolidation process to fabricate a composite structure.

In the diffusion-bonding process, fibers are attached to metal foils by use of an organic binder. The foils are stacked up to form a preform of the laminate having the desired fiber orientations. The preform is placed in a collapsible container and subjected to high temperature and pressure to cause the metal matrix to flow around the fibers and bond the laminae together. The organic binder is pyrolyzed to burn off volatile components during the beginning of the heating cycle. Diffusion-bonding processes have been used to make boron-reinforced aluminum and SiC-reinforced titanium composites.

4. COMPOSITE APPLICATIONS

4.1 Military Aircraft

Historically, military aircraft have provided the driving force behind the research, development, and utilization of composite structures. Composite construction on aircraft is continually increasing from 1% of the structural weight on the F-14 to 26% on the AV-8B (Schwartz, 1984). This trend does not show any sign of changing in the near future, as many of the new military aircraft programs on composite structures are aimed at further mass reduction. For example, the goal of the Air Force Ultra-Lightweight Structures program is to reduce the structural weight of several baseline aircraft by 50%. The extended and successful use of composites in military aircraft is now spilling over to other military applications—for example, the use of ceramic-matrix composites for vehicle armor and of polymer-matrix composites for structural components on ships, submarines, and other weapon systems.

4.2 Aerospace

Carbon-fiber–reinforced polymer and metal-matrix composites have played an important role in providing lightweight, dimensionally stable platforms in space for communications satellites. These applications are primarily performance-driven by the need to reduce the mass of structural components and increase the amount of active payload. The space shuttle orbiter uses many composite components to provide the performance necessary for a reusable reentry vehicle. Programs such as the National Aerospace Plane, a passenger version of the space shuttle for very fast transcontinental flight, are relying on composite materials to provide enabling technologies. The NASA-proposed space station project will require major advances in composite-material fabrication and processing to enable a massive structure to be assembled in orbit.

4.3 Commercial Aircraft

The use of composite materials on commercial aircraft has increased as experience with composites on military aircraft has grown, with both cost and safety mediating the rate of acceptance. Fiberglass-epoxy composite fairings and control-structure applications have increased dramatically and are now being supplanted by more advanced Kevlar-epoxy composites. Graphite-reinforced epoxy composites are also being more extensively used in secondary and primary structural applications such as elevator structures, rudders, and vertical fin structures. The use of carbon/carbon composite brake disks on commercial aircraft has resulted in significant weight savings (~590 kg) and lower wear rates. The use of advanced composites on commercial aircraft is expected to increase, especially as lower-cost materials and processing are developed. The increased penetration of carbon-fiber–reinforced polymer composites into primary load-bearing structures on commercial aircraft depends on the ability to improve the properties of the present composites such as

1. toughness,
2. higher temperature capability,
3. strength and modulus,
4. environmental resistance, and
5. compressive strength.

If these problem areas are solved, the use of carbon-fiber–reinforced polymers will also increase in other applications as well.

4.4 Ground Transportation

One area in which composites potentially could have large impact is automotive applications. At present there is significant use of nonstructural composite materials on automobiles, such as sheet molding compounds reinforced with chopped glass fiber for the front and rear fascias, hoods, trunk lids, and body panels. There have been a few small-volume applications of carbon-fiber–reinforced drive shafts in light-duty vans and trucks. Continuous-filament–wound, glass-fiber–reinforced polyester leaf springs are now used in a variety of passenger cars and light trucks. The composite leaf spring represents a major structural application of composites, having a superior fatigue lifetime over the metal component. Composites have been used in a number of prototype engine parts such as pistons and connecting rods made from both graphite-fiber and silicon-carbide–particulate reinforced aluminum. Prototype engines have been made that are constructed almost entirely from polymer-matrix composites including the engine block. Carbon/carbon composite disk brakes are presently used in racing cars because of their superior performance characteristics and lighter weight. Although many prototype parts have been evaluated, carbon-fiber–reinforced composites have not penetrated the automotive marketplace primarily because of the high costs associated with the raw materials and composite fabrication. The real potential of composites to lower manufacturing and assembly costs through integration of many individual components into a single composite structure has not yet been realized because of the absence of high-rate fabrication processes, and remains a challenge to future scientists and engineers.

4.5 Recreation and Sporting Equipment

In direct contrast to the cost-sensitive automotive applications, sporting and recreational applications of composites are performance-driven. Glass fibers were first used in composite fishing rods to improve performance over bamboo and metal poles. The fiberglass was supplanted by graphite and Kevlar fibers. Tennis racquets that are constructed from graphite-epoxy composite are stiffer than wooden or aluminum racquets and do not dissipate as much energy in deformation of the racquet, allowing more energy to be given to the ball, giving the player a definite performance advantage. Boron-fiber–reinforced epoxy is also experiencing increased interest for tennis racquet applications. Advanced composites are being used in nearly every possible sporting-goods application including fishing rods, snowmobiles, motorcycles, bicycles, skis, track and field equipment, archery bows and arrows, protective helmets, and golf clubs, to mention a few. The application of advanced composites to sports and recreational equipment will continue to grow as new materials are developed, such as the extended-chain polyethylene fibers.

5. FUTURE OUTLOOK FOR COMPOSITE MATERIALS

The use of composite materials is cost- and performance-benefit–driven. Composite materials usually cost more than metals so that direct replacement of metal parts with their composite counterparts is not an effective strategy unless the benefit such as mass reduction, increased lifetime, additional performance, etc., can offset the additional cost of the composite. Substitutional strategies work for performance-driven applications such as aircraft where mass savings have a direct economic impact, such as improved fuel economy or increased payload capability. In cost-driven applications such as passenger automobiles, mass reduction does not result in a sufficient savings to the user in terms of either decreased operating costs or increased payload capability that can offset the increased cost of the composite structure. Thus, the composite must provide an economic benefit in terms of lower overall vehicle construction cost through parts consolidation and lower assembly costs; simple part-substitution strategies do not work, because of the higher cost of the composite part. In order to address high-volume, cost-driven applications, radically new engineering approaches are necessary that depart from traditional engineering practices used for metallic structures. Such new approaches are being explored as high-speed resin-transfer molding of highly integrated structures, high-speed reaction injection molding of large integrated structures, and filament winding of thermoplastic structural components (U.S. Congress, Office of Technology Assessment, 1988).

The development of new low-cost composite fabrication techniques to supplant traditional fabrication practices in cost-driven high-volume applications represents an important goal for domestic industries to achieve in order to remain competitive in the world marketplace.

In addition to the development of new composite-fabrication technology, the development of lower-cost, high-performance constituent materials will also lead to the more widespread utilization of composites. As the use of composite materials has increased, the need for a more basic understanding of the science of composite materials has also increased. Since the early 1980s, our understanding of the complex interactions between the constituents in composites has dramatically improved through the application of more sophisticated analytical techniques. The improved understanding in turn challenges researchers to develop new materials and processes.

GLOSSARY

Anisotropic: Having properties that are different when tested along different directions relative to a coordinate system fixed in the material.

Aspect Ratio: The ratio of the length of a fiber to its diameter.

Autoclave: A pressure vessel that can maintain a prescribed pressure and temperature of a gas for curing polymeric matrix composites.

Carbonaceous: Relating to or composed of carbon.

Cermet: A composite material made by mixing, pressing, and sintering a mixture of metal and ceramic powders.

Die: A tool or mold used to impart shapes to or form impressions on materials such as metals, glasses, polymers, and ceramics. An example is a die used in spinning glass fibers; a molten glass charge is forced through small orifices in the die, which shapes the glass into small continuous cylinders or fibers.

Fabric: Planar sheets of woven or nonwoven materials consisting of intertwined yarns, fibers, or filaments.

Fiber: A term usually used to describe short lengths of a material having very small cross-sectional areas and aspect ratios > 1.

Filament: A fiber having an infinite aspect ratio, such as a continuous fiber.

Laminate: A composite material consisting of multiple layers of a single material or different materials bonded together.

Matrix: The principal phase in which the reinforcing phase is embedded.

Petroleum Pitch: A distillate product from petroleum used in the production of carbon fibers.

Poisson's Ratio: A constant relating the change in sample cross-sectional area to the change in length when a material is elastically deformed.

Polyacrylonitrile (PAN): A polymer of acrylonitrile used in the manufacture of carbon fibers.

Stiffness: A general term used to describe the ratio of the stress applied to a material and the deformation produced. The stiffness of an elastic material subjected to a uniaxial stress is referred to as Young's modulus.

Superalloys: Alloys based on iron, nickel, and cobalt that are an extension of stainless-steel alloys. Superalloys have higher tensile strength than stainless-steel alloys at elevated temperatures and are primarily for use at temperatures above 540 °C.

Tensile Strength: The maximum tensile force per unit area that a sample can sustain.

Thermoplastic: A material that can be repeatedly softened by increasing or hardened by lowering its temperature.

Thermosetting Polymer: A polymeric material that irreversibly converts into an infusible and insoluble solid on curing by heating or chemical means.

Tow: An untwisted loose collection of continuous filaments.

Yarn: Strands of fibers or filaments suitable for weaving.

Young's Modulus of Elasticity: The ratio of the tensile stress to the elastic strain or deformation produced.

List of Works Cited

Boyer, H. E., Gall, T. L. (1985), *Metals Handbook*, Metals Park, OH: American Society for Metals.

Bracke, P., Schurmans, H., Verhoest, J. (1984), *Inorganic Fibers and Composite Materials*, New York: Pergamon.

Broutman, L. J., Krock, R. H. (1967), *Modern Composite Materials*, Reading, MA: Addison-Wesley.

Diefendorf, J. R. (1983), *Fiber Producer*, August, pp. 56–68.

Diefendorf, J. R. (1985), in: W. Watt, B. V. Perov (Eds.), *Strong Fibers*, Vol. 1 of *Handbook of Composites*, A. Kelly, Yu. N. Rabotnov, series eds. Amsterdam: North Holland, p. 46.

Evans, C. C. (1972), *Whiskers*, M & B Monograph ME/8 series, J. G. Cook (Ed.), London: Mills & Boon.

Feigelson, R. S. (1988), *Mater. Sci. Eng.* **B1**, 67–75.

Helminiak, T. E., Arnold, F. E., Brenner, C. L. (1975), *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **16**, 659–671.

Hull, D. (1981), *An Introduction to Composite Materials*, Cambridge: Cambridge University Press, p. 111.

Islam, M. U., Wallace, W., Kandell, A. Y. (1988), *Artificial Composites for High Temperature Applications*, New Jersey: Noyes Data Corporation.

Jones, F. R. (Ed.) (1989), *Interfacial Phenomena in Composite Materials 89*, London: Butterworth.

Kerans, R. J., Hay, R. S., Pagano, N. J., Parthasarathy (1989), *Ceram. Bull.* **68**, 429–442.

Lindenmeyer, P. H. (1976), in: L. E. Murr, C. Stein (Eds.), *Frontiers of Material Science Distinguished Lectures*, New York: Dekker, Chap. 6, p. 219.

Metcalfe, A. G. (Ed.) (1974), *Interfaces in Metal Matrix Composites*, New York: Academic.

Murdok, L. J., Brook, K. M. (1979), *Concrete Materials and Practice*, New York: Wiley.

National Materials Advisory Board (1987), Report of the Committee on Thermoplastic Composites as Structural Components, *The Place for Thermoplastic Composites in Structural Components*, Washington, D.C.: National Academy (NMAB-434).

North, B. (1987), *Int. J. High Technol. Ceram.* **3**, 113–127.

Petrovic, J. J., Milewski, J. V., Rohr, P. L., Gac, F. D. (1985), *J. Mater. Sci.* **20**, 1167–1173.

Prewo, K. M. (1982), *J. Mater. Sci.* **17**, 3549–3557.

Rauch, H. W., Sutton, W. H., McCreight (1966), *Survey of Ceramic Fibers and Fiberous Composite Materials*, Springfield, VA: National Technical Information Service. AFML-TR-66-365.

Schioler, L. J., Stiglich, Jr., J. J. (1986), *Am. Ceram. Soc. Bull.* **65**, 289–292.

Schwartz, M. M. (1984), *Composite Materials Handbook,* New York: McGraw-Hill.

Sowman, H. G. (1988), *Am. Ceram. Soc. Bull.* **67**, 1911–1916.

Stoller, H. M., Butler, B. L., Theis, J. D., Phillips, D. C. (1974), in: J. W. Weeton, E. Scala (Eds.), *TMS-AIME Proceedings of Symposium on Composites: The State of the Art,* Warrendale, PA: The Minerals, Metals & Materials Society, p. 69.

Tibbetts, G. G., Beetz, C. P., Endo, M. (1986), *SAMPE J.* **22**, 30–36.

U.S. Congress, Office of Technology Assessment (1988), *Advanced Materials by Design,* Washington, D.C.: U.S. GPO (OTA-E-351), p. 106.

Weeton, J. W., Peters, D. M., Thomas, K. L. (1987), *Engineer's Guide to Composite Materials,* Metals Park, OH: American Society for Metals.

Further Reading

Jones, F. R. (Ed.) (1989), *Interfacial Phenomena in Composite Materials 89,* London: Butterworth.

Kelly, A. (Ed.) (1990), *The Concise Encyclopedia of Composite Materials,* Cambridge, MA: MIT Press.

Lee, S. M. (Ed.) (1990), *The International Encyclopedia of Composites,* New York: VCH.

Schioler, L. J., Stiglich, J. J., Jr. (1986), *Am. Ceram. Soc. Bull.* **65**, 289–292.

Schwartz, M. M. (1984), *Composite Materials Handbook,* New York: McGraw-Hill.

Weeton, J. W., Peters, D. M., Thomas, K. L. (1987), *Engineer's Guide to Composite Materials,* Metals Park, OH: American Society for Metals.

COMPRESSIBLE FLOWS

J. E. SHEPHERD, *Department of Mechanical Engineering, Rensselaer Polytechnic Institute, Troy, New York, U.S.A.*

	Introduction	43
1.	**Thermodynamics of Compressibility**	44
2.	**Fluid Dynamics—The Euler Equations**	46
3.	**Criteria for Compressibility**	47
4.	**Simple Flows**	47
4.1	Quasi One-Dimensional Flows	47
4.2	Shock Waves	50
5.	**Detonation Waves**	52
6.	**One-Dimensional Unsteady Flows**	53
7.	**Multidimensional Flows**	57
7.1	Steady Supersonic Flows	57
7.2	Aerodynamics	59
7.2.1	Linearized Flow Theory	59
7.2.2	Subsonic Flow	60
7.2.3	Supersonic Flow	61
7.2.4	Transonic Flow	61
7.3	Hypersonic and Hypervelocity Flows	62
7.4	Unsteady Flows	64
8.	**Flows with Phase Change**	65
	Glossary	67
	Works Cited	67
	Further Reading	68

INTRODUCTION

Compressible flow is the study of fluid motion with significant induced density changes. Sufficiently strong forces will cause any real material to flow as a compressible fluid. This covers the range from the nonlinear elastic deformation of solids under high explosive loading to the flow of compressed air from an automobile tire. Compressible flows are significant in a wide range of technical devices and natural phenomena. These include high-speed flight, detonations and explosions, gas turbines, rocket engines, nuclear weapons, blast waves and sonic booms, high-velocity impact, power generation, chemical synthesis, laser fusion and high-pressure research, volcanic eruptions, and shock waves within supernovae.

Compressible flows are essentially nonlinear phenomena that arise from the interaction of thermodynamics and fluid dynamics. This article describes the general theory of compressible flow starting from the basic thermodynamic properties of sound speed, the fundamental derivative, the Grüneisen coefficient, and the conservation relations for a compressible fluid. The special case of adiabatic, inviscid flow governed by the Euler equations is given particular attention. Many practical flows can be analyzed in this approximation, since the effects of viscosity and thermal conduction are often confined to thin layers near walls or at slipstreams. Solutions to the Euler equations are observed to consist of smooth regions of isentropic flow joined by discontinuous jumps in properties through shock waves or contact surfaces.

This article begins with basic considerations about thermodynamics and fluid dynamics. The two basic elements of compressible flow, isentropic motion and shock waves, are then introduced. Steady flow within a converging–diverging or de Laval nozzle is discussed with one-dimensional theory. The complexity of real nozzle flows is illustrated with selected experimental results. Shock waves are treated from both the classical Rankine–Hugoniot discontinuity approach and the molecular viewpoint. Detonations are treated as shock waves followed by chemical reaction zones.

One-dimensional unsteady flow is illustrated with a discussion of the problem of shock-wave formation, the theory of shock tubes, and the propagation of blast waves. Steady supersonic flows in two dimensions are characterized by regions of isentropic flow

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

(uniform motion or expansion waves) interrupted by contact surfaces or shock waves. If the shock waves are curved, a continuous variation of entropy and vortical flow will be produced. Simple flows without vorticity can be analyzed by the method of characteristics in the isentropic regions, supplemented by the shock jump conditions and matching conditions at contact surfaces.

Aerodynamic applications are considered in the context of two-dimensional steady flows over streamlined bodies. Corrections to the lift and drag due to compressibility are derived from linearized equations. Optimal airfoil design for modern aircraft illustrates the issues associated with flow near the sound speed and the interaction between shock waves and boundary layers. Flows at very high velocities are characterized by strong shock waves. Lift and drag on reentry vehicles are determined by the shock configuration and viscous heating at the body surface.

Unsteady compressible flows in two dimensions are illustrated by the problem of Mach reflection. Experimental and numerical results on the interaction of a shock wave with a ramp are discussed. The article ends with some selected results on compressible flow with phase change.

A few comments on the history and origins of compressible flow are appropriate here. Early efforts, starting with Newton, originated in the study of sound generation and propagation. Later the technological problems of steam engines and steam turbines provided motivation for the engineers and scientists of the 1800s. A genuine theory of compressible flow was possible only after the foundations of thermodynamics had been completed in 1850. Ideas from classical fields of mechanics, hydrodynamics, elasticity, and acoustics were ultimately combined with thermodynamic notions to obtain the understanding of today.

The basic foundations of compressible flow were developed in the period 1850–1910. A significant step was the experimental observation of "choking" by Napier (1867) and others and the theoretical explanation using isentropic flow ideas by Reynolds (1885). The problem of nonlinearity and wave steepening was formulated by Riemann in 1860, and the theory of the shock wave as a discontinuity was completed by Rankine (1870) and Hugoniot (1889). Supersonic flow and shock waves were revealed by Mach's experimental observations (1870–1900) and analysis of shock waves produced by high-speed bullets, spark discharges, and gas jets. Berthelot and Vielle discovered (1878–1883) the nature of the detonation process in gases and measured detonation velocities. Chapman (1899) and Jouguet (1904) analyzed the solutions to the shock jump conditions for explosive gases and determined that the minimum-wave-speed solution agreed with the experimental measurements. Vielle (1900) experimentally confirmed Hugoniot's theory for the relation between pressure rise and wave speed for moving shock waves.

Stodola's experiments and analysis of the de Laval nozzle flow (1903) clarified the role of shock waves in steady flows. This was followed by Prandtl's comprehensive experiments and analysis of the nozzle flow problem (1906). Prandtl's analysis (1908) of oblique shock waves and steady expansion waves in two-dimensional flow allowed a complete explanation of the jet phenomena first observed by Mach. These basic ideas were then extensively applied to gas turbines and supersonic flight from 1910 to 1940. Since that time, the technological problems of high-speed aircraft, gas turbines and rockets, and explosives have continued to motivate the further developments of experimental techniques and more sophisticated theory. In the last 20 years, numerical computational methods have played an increasingly important role in both engineering and research.

1. THERMODYNAMICS OF COMPRESSIBILITY

Most relevant to compressible flows is the relation between pressure and volume of a fluid, which can be determined from classical thermodynamics. The relative change in specific volume v for small changes in pressure P and entropy s about an undisturbed state $(\ )_0$ is given by

$$\frac{v-v_0}{v_0} = -K_s(P-P_0) + \Gamma K_s^2(P-P_0)^2 + G\frac{K_sT_0}{v_0}(s-s_0) + \cdots. \qquad (1)$$

The quantity K_s is the *adiabatic compressibility,*

$$K_s = -\frac{1}{v}\left(\frac{\partial v}{\partial P}\right)_s, \tag{2}$$

which is inversely related to the velocity of small-amplitude adiabatic (acoustic) disturbances, the *sound speed* $c=\sqrt{v/K_s}$. Compressibility of low-pressure gases such as ordinary air is on the order of the inverse pressure or about 10^{-5} Pa^{-1} at normal atmospheric pressure. Liquid and solid compressibilities are much smaller, typically 10^{-10} Pa^{-1}.

The *fundamental gasdynamic derivative* Γ is defined (Thompson, 1988) to be

$$\Gamma = \frac{c^4}{2v^3}\left(\frac{\partial^2 v}{\partial P^2}\right)_s, \tag{3}$$

a nondimensional measure of the nonlinear response of the fluid. Values of Γ range from slightly above unity for ordinary gases up to 4–7 for typical organic liquids. Normal fluids have values of $\Gamma > 1$. In some circumstances Γ can be less than 1 and in exceptional cases Γ can be negative. The *Grüneisen coefficient* G is a nondimensional measure of thermal expansion,

$$G = \frac{1}{c_p K_s}\left(\frac{\partial v}{\partial T}\right)_P, \tag{4}$$

which has a value between 0 and 0.67 for gases and around 2 in most solids.

K_s, Γ, and G are thermodynamic properties that are determined by the molecular composition of the fluid and the extent of intermolecular interactions. Representative values of these and related properties are given in Table 1 for selected substances at conditions of $P=1$ bar (10^5 Pa) and $T=20$ °C.

Specific volume may change as a result of either applied pressure or entropy increases that are the consequence of irreversible processes such as external heating or shock waves. Many compressible flow problems of significance to applied physics are adiabatic and consist of smooth regions of *isentropic flow* with zero entropy change interrupted by thin layers or *shocks* in which the entropy irreversibly increases.

It is the variation of the sound speed with thermodynamic state that provides compressible flow with its true nonlinear character. There are two aspects of nonlinearity: the strength of the disturbances, and the degree of nonlinear response of the fluid to these disturbances. Disturbance strength is quantified by the ratio of the amplitude of the pressure variation ΔP to the quantity $1/K_s = \rho c^2$. Material response will be nonlinear when $\Delta P/\rho c^2$ becomes nonnegligible. This will occur for disturbance pressures of >0.2 bar for atmospheric air and pressures of thousands of atmospheres for substances such as water.

Nonlinearity is intimately related to Γ, both as the coefficient of the second term in Eq. (1) and by the relation to the variation (Thompson, 1971) of sound speed with pressure,

$$\left(\frac{\partial c}{\partial P}\right)_s = \frac{v}{c}(\Gamma - 1). \tag{5}$$

Wave propagation in compressible flows is profoundly influenced by nonlinearity. The dependence of propagation velocity on amplitude results in the creation of propagating sharp jumps, *shock waves,* or continually spreading *expansion waves* in the fluid state. The character and degree of nonlinearity are determined by the numerical value of Γ. The sign of Γ determines the curvature of the isentropes in (P,v) coordinates and the type of shock waves that are formed.

Table 1. Properties of some compressible materials.

	ρ (kg/m ^{3})	c (m/s)	K_s (Pa^{-1})	Γ	G
Air	1.205	344	7.01×10^{-6}	1.20	0.40
Water	998.2	1484	4.55×10^{-10}	4.4	0.11
Ethyl Alcohol	790	1159	9.42×10^{-10}	6.4	0.64
Steel (304S)	7896	4569	6.07×10^{-12}	2.98	2.47

2. FLUID DYNAMICS—THE EULER EQUATIONS

Changes in specific volume or mass density $\rho = 1/v$ will produce fluid motion. Conservation of mass applied to a fluid element requires that a local motion inward or outward, indicated by a nonzero divergence in the velocity field $\mathbf{u}$, be produced by any change in mass within that volume:

$$\mathrm{div}\mathbf{u} = -\frac{1}{\rho}\frac{D\rho}{Dt}. \tag{6}$$

The *substantial* or *convective* derivative used in this expression is defined as $D(\cdots)/Dt = \partial(\cdots)/\partial t + \mathbf{u}\cdot\mathrm{grad}(\cdots)$ and represents the time derivative taken in a frame of reference moving with the fluid.

The complete description of the motion is determined by the thermodynamic properties discussed above together with the balance equations for mass, momentum, and energy including the appropriate boundary and initial conditions, and thermodynamic and transport properties (see FLUID DYNAMICS, EQUILIBRIUM). An approximation of wide applicability is that of inviscid, adiabatic flow. Such flows are determined by solutions to the *Euler equations*:

$$\frac{\partial}{\partial t}\rho + \mathrm{div}(\rho\mathbf{u}) = 0, \tag{7}$$

$$\frac{\partial}{\partial t}\rho\mathbf{u} + \mathrm{div}(\rho\mathbf{u}\mathbf{u}) = -\mathrm{grad}P, \tag{8}$$

$$\frac{\partial}{\partial t}\rho\left(e + \frac{|\mathbf{u}|^2}{2}\right) + \mathrm{div}\left[\rho\mathbf{u}\left(e + \frac{P}{\rho} + \frac{|\mathbf{u}|^2}{2}\right)\right] = 0. \tag{9}$$

Pressure plays a special role. As in incompressible flow, the pressure field plays a mechanical role in the balance of momentum as a force field that accompanies material acceleration. An equally important role for pressure in compressible flows is as a thermodynamic variable, which is implicitly coupled to both internal energy $e = e(P,\rho)$ and the mass density through the Euler equations. This dual role of pressure is responsible for many of the special features of compressible flow, including the phenomena of wave propagation and the interconversion of thermal and mechanical energy.

An equation of state for the internal (thermodynamic) energy $e = e(P,\rho)$ must be supplied to complete the Euler-equation model of compressible flow. High-pressure experiments or molecular-dynamics simulations are required for solids and liquids to obtain the necessary thermodynamic information. Ideal models of thermodynamic behavior can often be applied in the case of gases. Most often studied has been the compressible flow of a *perfect gas*, an ideal gas with constant specific heats. Ordinary air and most other gases can be described by this model up to moderate pressures (50 bar) and temperatures (1000 K). Only one parameter is required for this model, the ratio of specific heats $\gamma = c_p/c_v$, which takes on values between 1 and 1.667. The equation of state is

$$e = Pv/(\gamma - 1). \tag{10}$$

Temperature can be determined by the usual ideal-gas law $P = \rho RT$. The various derivatives significant to compressible flow behavior are

$$K_s = 1/\gamma P,$$

$$c = \sqrt{\gamma P v},$$

$$\Gamma = \tfrac{1}{2}(\gamma + 1),$$

$$G = \gamma - 1. \tag{11}$$

Solution of the Euler equations with the perfect-gas equation of state is the subject of classical *gasdynamics*. These equations must be solved numerically except for special cases in which analytic solutions have been obtained for steady flows and self-similar transient flows. Solutions to compressible flow problems have the special feature of the spontaneous generation of discontinuities such as shock waves and contact surfaces.

Fluid composition must be specified in order to analyze chemically reacting flows such as detonations. This requires an equation of state appropriate to a mixture of species and an evolution equation for the composition. Changes in the species concentrations are due to simultaneous material transport and chemical reaction. A reaction-rate model must be developed from molecular considerations of the processes of reactive collisions (combustion chemistry), vibrational–translational energy exchange, and electronic excitation or ionization dynamics. Supersonic and hypersonic vehicle aerodynamics, spacecraft heating during reentry, high-speed flames in gases, and detonation of explosives (gaseous and condensed) (see EXPLOSIVES) offer a rich variety of applications that motivate the study of *compressible reacting flows.*

3. CRITERIA FOR COMPRESSIBILITY

Under what circumstances is a flow considered compressible? Combining our thermodynamic analysis with the Euler equations shows that the nondimensional pressure $\Delta P/\rho c^2$ is of the same magnitude as the *Mach number* $M=u/c$ for steady flows. Therefore, the Mach number must exceed some minimum value for compressibility to be considered an essential feature of a steady flow. A value of $M>0.3$ is often taken as a criterion for ordinary gases. A more important distinction is between *subsonic* flow, $M<1$, and *supersonic* flow, $M>1$, with the case of *sonic* flow, $M=1$, playing a special role.

In the case of unsteady flows, compressibility effects must be considered whenever the acoustic-wave transit time L/c across a characteristic distance L is comparable to or greater than other significant time scales present in the flow. Small-amplitude motion with $M\ll 1$, $\Delta\rho/\rho\ll 1$, and $\Delta P/\rho c^2\ll 1$ can be described by the theory of *acoustics* (see ACOUSTICS, LINEAR). The approximation of acoustics is to neglect the convective derivative terms compared to the unsteady derivatives in the Euler equations and to treat the motion as isentropic. Acoustic flows are described by the linear wave equation, for which a large number of exact solutions can be obtained. Large-amplitude motions are intrinsically nonlinear and require experimental or numerical investigation (see ACOUSTICS, NONLINEAR).

4. SIMPLE FLOWS

Many technical applications of compressible flow require only simple notions about the dynamics of motion in one dimension. Useful information can be obtained by considering only variations in one spatial direction even in the case of apparently multidimensional flow within a nozzle or a tube of variable cross section. Transient flows in ducts of a constant cross section play an important role in the operation of devices such as shock tubes, mufflers, and the phenomenon of water hammer. Shock waves can be analyzed by considering only the one-dimensional structure normal to the wave front, even in cases of curved waves. Detonation waves and transient shock waves (blast waves) resulting from explosions can also be productively examined from a one-dimensional viewpoint.

4.1 Quasi One-Dimensional Flows

A typical application of compressible flow is accelerating fluids from rest to high velocities. This can be accomplished through a de Laval nozzle, first used in steam turbines. It has since been widely applied in the generation of supersonic flow in devices such as wind tunnels and rocket motors. Fluid at high pressure is accelerated through a duct that converges and then diverges as shown in Fig. 1.

Simple explanations of the principal features of these flows are possible by assuming a slowly varying duct cross section and steady-state ($\partial/\partial t=0$), adiabatic flow. This ensures that the flow is almost parallel and that the

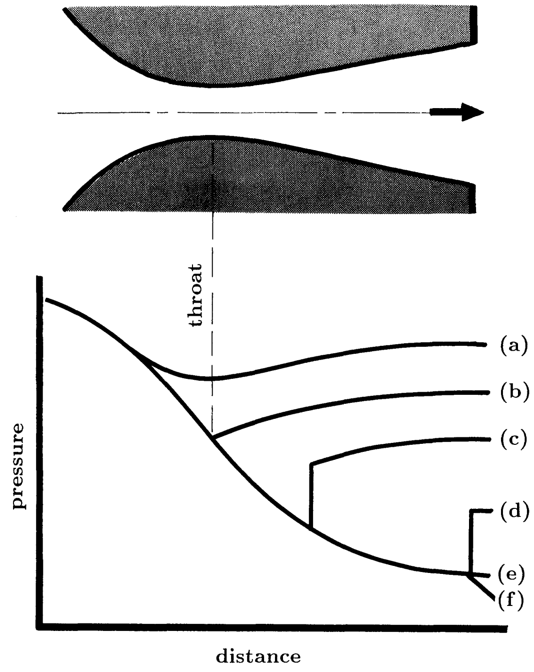

FIG. 1. Flow through a de Laval nozzle and the associated changes in the fluid pressure. Each curve represents a different downstream pressure as discussed in the text.

accelerations perpendicular to the main flow direction are also small. In this *quasi one-dimensional* approximation, the Euler equations can be compactly written as

$$\frac{d}{dz}(\rho uA)=0,$$

$$\frac{dP}{dz}=-\rho u\frac{du}{dz},$$

$$\frac{d}{dz}\left(h+\frac{u^2}{2}\right)=0, \tag{12}$$

where $A(z)$ is the duct cross-sectional area perpendicular to the main flow direction z. In this situation, the total mass flow rate is constant and equal to ρuA. The energy equation is written in terms of the *enthalpy* $h=e+P/\rho$. The combination $h_t=h+u^2/2$ is known as the *total* or *stagnation* enthalpy, which is constant in adiabatic, steady flows.

If the motion is inviscid and no shock waves are present, adiabatic flow can be further simplified as isentropic, and variations in pressure and density can be related by Eq. (1) to obtain the isentropic flow equations

$$\frac{d\rho}{dz}=\frac{1}{c^2}\frac{dP}{dz}, \quad \frac{1}{u}\frac{du}{dz}=\frac{1}{M^2-1}\frac{1}{A}\frac{dA}{dz}. \tag{13}$$

For initially subsonic flow, $M<1$, as in the case of accelerating a fluid from rest, a converging duct $dA/dz<0$ will increase the fluid velocity as the pressure and density decrease. This acceleration is equivalent to the conversion of thermal energy into mechanical energy, as shown by the integrated form of the energy equation $u=\sqrt{2(h_t-h)}$, where the total enthalpy h_t is just the enthalpy of the stationary fluid in the reservoir supplying the duct.

The *stagnation state* is a hypothetical state produced by isentropically bringing a flow to rest (stagnation). If the fluid reservoir supplying the nozzle is large enough, then the pressure P within the reservoir will be equal to the *stagnation pressure* P_t. Stagnation pressure in a steady, isentropic flow is constant. The stagnation pressure in a steady flow can be changed only by external work, energy addition, or dissipative processes such as friction or shock waves.

The Mach number will increase monotonically with the velocity for usual fluids, which have $\Gamma>1$, since

$$\frac{du}{u}=\frac{1}{1+(\Gamma-1)M^2}\frac{dM}{M}. \tag{14}$$

Given sufficient stagnation pressure, continuous contraction in the duct cross section will result in the Mach number approaching 1, a phenomenon known as *choking*. In a constant-area duct, choking can also occur as a result of frictional forces or heat addition.

It is apparent from the differential relation equation (13) for velocity that transition through the *sonic point*, $M=1$, is possible only at an area extremum. This takes the shape of a minimum forming a *throat* in the nozzle for the case of normal fluids, $\Gamma>0$. If the downstream pressure is sufficiently low, as in cases (c)–(f) of Fig. 1, sonic flow will occur at the throat and the flow downstream will be supersonic for some distance. In the supersonic flow downstream of the throat, increasing the duct area results in further velocity and Mach-number increases. Pressure, temperature, and density will decrease with increasing Mach number in a supersonic flow.

For a given nozzle and stagnation pressure, the flow within the nozzle depends crucially on the pressure at the exit of the nozzle. If the flow is subsonic at the nozzle exit, the exit pressure will be approximately equal to the ambient pressure in fluid outside the nozzle. The situation is more complex for sonic or supersonic exit flows and the exit pressure is, in general, not equal to the ambient pressure.

Several possible situations are shown in Fig. 1. First, in case (a), the ambient pressure must be less than the stagnation pressure in order for flow to take place. Second, if the ambient pressure is not too low, then the flow remains subsonic throughout. This is the case for exit pressures above case (b) of Fig. 1. Third, when the ambient pressure is sufficiently low, cases (c)–(f), the flow will be *sonic* at the nozzle throat ($M=1$) and supersonic ($M>1$) for some distance downstream. Sonic flow at the throat will occur when the ratio of stagnation pressure to throat pressure exceeds a critical value. This value is determined by the thermodynamic properties of the fluid and can be estimated to be about 1.9 for air, 5300 for water, 2000 for ethyl alcohol, and 5.21×10^5 for steel. The pressure required at the nozzle exit to first produce this state, case (b), must be determined from the specific nozzle geometry and operating conditions.

Supersonic flow will exist for some distance downstream of the throat for all cases with an exit pressure lower than case (b). For exit pressures between cases (b) and (e), the nozzle flow is termed *overexpanded.* For exit pressures between those of cases (b) and (d), a special situation within the nozzle must exist. A sudden increase in pressure occurs through a nonisentropic *shock wave,* discussed in detail in the next section. The shock wave compresses and decelerates the fluid, resulting in subsonic flow downstream with the pressure increasing with distance up to the exit value. There is a loss of total pressure across the shock wave and an increase of entropy. The shocks are idealized as sharp discontinuities normal to the flow in Fig. 1, but in reality these waves are usually obliquely oriented to the flow (described in Sec. 7.1). An additional feature of real nozzle flow is the interaction between the shock and the boundary layer at the nozzle surface. This interaction can result in separation of the flow from the nozzle surface and the formation of a recirculation or reverse-flow region downstream of the interaction point.

For pressures between cases (d) and (e), shock waves will form in the jet outside the nozzle. The *pressure-matched* situation of case (e) has a smooth acceleration of flow up to nozzle exit where the pressure is exactly equal to ambient. Case (f) is an *underexpanded* flow and the nozzle exit pressure is higher than ambient. Further expansion must take place outside the nozzle. Conservation of energy limits the maximum fluid velocity that can be achieved by expansion to zero pressure to be $u_{\max} = \sqrt{2h_t}$. However, for ordinary fluids, the Mach number will be infinite in this state.

Analytic and tabulated solutions (Brower, 1990) are available for the thermodynamic states and flow variables of quasi one-dimensional isentropic flow of a perfect gas. For other fluids, each case can be treated numerically given sufficient thermodynamic information to compute enthalpies and sound speeds on isentropes. Approximate similarity principles based on Γ are available (Thompson and Sullivan, 1977) as guidelines for liquids and dense gases.

Real fluid flows in nozzles are multidimensional and have complicating factors such as condensation in moist air and a viscous boundary layer on the duct walls. Accurate computations of the flow should include the displacing effect of this layer and the motion transverse to the main (axial) flow direction. Realistic two-dimensional steady-state computations including these effects have been carried out and are in quantitative agreement with experiments. Figure 2 shows a schlieren photograph of the flow (Schnerr, 1989) of moist air through a two-dimensional circular arc (radius 100 mm) nozzle with a throat width of 120 mm. The flow is from left to right. Dark regions in the photograph are expansions; light regions represent compressions.

The fluid in the reservoir was at a pressure of 1 bar and temperature of 295 K, and had a relative humidity of 37.2%. The rapid cooling rate ($> 10^6$ K/s) due to the expansion in the nozzle causes supersaturation and homogeneous nucleation of the water vapor. This occurs downstream of the throat in the supersonic flow region near a Mach number of 1.2 to 1.3, resulting in energy addition to the flow from the heat of condensation. Addition of thermal energy to the supersonic flow causes

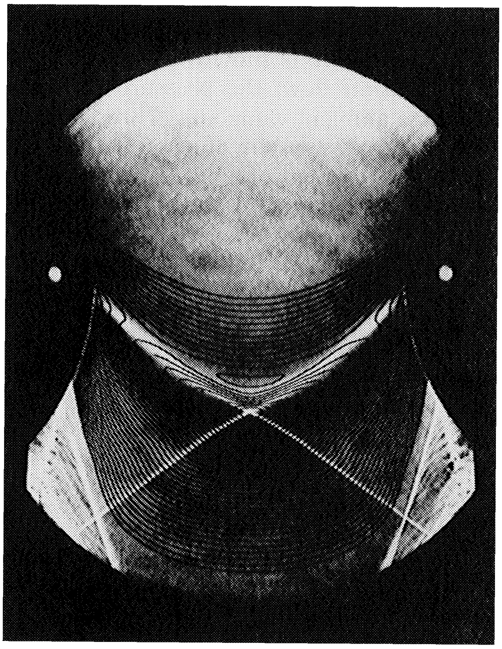

FIG. 2. Experimental visualization of the flow of moist air through a circular arc nozzle with superimposed calculated iso-Mach contours.

the density to increase and oblique shocks to form. These form the bright X-shaped pattern visible in the photograph. Directly downstream of the shocks, a rapid expansion takes place. Unrelated to condensation, but equally important in real nozzles, is the wedge-shaped separated-flow region visible at the downstream edges of the nozzle arcs. Lines of constant Mach number are superimposed on the flow image. These were computed from the numerical solution (Schnerr and Dohrmann, 1990) of the Euler equations supplemented with a condensation model. These contours start with the sonic line, $M=1$, at the left and increase in increments of 0.02 to the right.

The supersonic acceleration process in a nozzle is limited only by the decoupling of translational and internal energy modes that occurs when the density becomes too low for molecular collisions to be effective. When the supersonic expansion occurs in a very low-pressure region (vacuum), the resulting flow is known as an *underexpanded supersonic jet*, illustrated in Fig. 3. This is the multidimensional flow resulting from a converging nozzle with choked flow ($M=1$) at the exit. The sonic exit flow is further accelerated through an expansion fan (discussed subsequently in Sec. 7) after exiting from the nozzle.

The effect of a small but not negligible ambient pressure is shown in Fig. 3. Within the expansion flow, the pressure decreases below the ambient value and is brought back up by a complex system of shock waves. This shock "cell" consists of a barrel-shaped shock surrounding the jet and a nearly planar normal shock, or Mach disk, downstream. The boundary between the jet and ambient fluid is outside of the barrel shock and actually is a region of mixing or a *shear layer*, as indicated by the shaded region. There are large changes of density across this boundary, resulting in the pressure-wave reflections shown in Fig. 3. There are a series of shock cells that can be observed within the jet flow at lower pressure ratios in sonic and supersonic jets, corresponding to cases between (d) and (e) of Fig. 1.

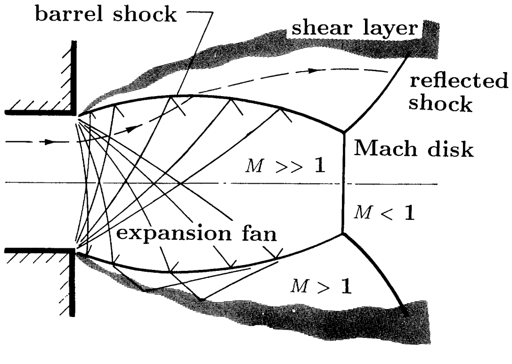

FIG. 3. Features of a free jet expansion from a sonic flow produced by a converging nozzle. The pressure at nozzle exit is much higher than the pressure of the stationary ambient fluid.

Free jet expansions are used to create very low-temperature, supersonic molecular beams essential to modern chemical physics (Levine and Bernstein, 1987). A population inversion can be produced in the molecular vibration–rotation states if the expansion process is sufficiently rapid. This can create the conditions necessary for light amplification, the principle behind the *gasdynamic laser* (Anderson, 1976) (see LASERS, GAS).

4.2 Shock Waves

Steady supersonic flow and transient compressible flows are distinguished by the potential for spontaneous generation of shock waves. Shock waves occur in steady supersonic flows whenever flow deflections are associated with pressure increases. In transient flows, rapid energy addition or boundary motions will produce traveling compressive disturbances that may form shock waves if sufficiently strong. Steady and unsteady shocks are a consequence of the nonlinearity of compressible flow and the underlying *hyperbolic* nature of the Euler equations. In the context of the Euler equations, shock waves appear as jumps or discontinuities in the solutions. Physically, shock waves are found to have a finite thickness, on the order of microns for gases at atmospheric pressure and nanometers for liquids or solids.

All shock waves can be locally analyzed by focusing on a small spatial region (control volume) surrounding the wave and by transforming to a coordinate system moving at the local shock velocity. Conservation of momentum implies that only the component of velocity u normal to the wave experiences change. The tangential component of velocity must be continuous. For simplicity, the tangential component is set equal to zero in this analysis. Including a nonzero tangential velocity leads to the concept of oblique shock waves, discussed in subsequent sections. The analysis carried out here is valid for the normal component of velocity in both cases. Irrespective of the shock orientation or motion, only a single set of equations is required.

Integration of the Euler equations through the thin control volume around the shock yields the *jump conditions*

$$\rho_1 u_1 = \rho_2 u_2, \tag{15}$$

$$P_1 + \rho_1 u_1^2 = P_2 + \rho_2 u_2^2, \tag{16}$$

$$h_1 + \frac{u_1^2}{2} = h_2 + \frac{u_2^2}{2}, \tag{17}$$

in which the subscripts denote states upstream (1) and downstream (2) of the wave and u is the fluid velocity normal to the wave surface in a wave-fixed coordinate system. These equations can be combined into a single expression, the *Rankine–Hugoniot equation*:

$$h_2 - h_1 = (P_2 - P_1)(v_2 + v_1)/2. \tag{18}$$

In the wave-fixed frame, the velocity component normal to the wave is always supersonic, $M_1 = u_1/c_1 > 1$, upstream and subsonic, $M_2 = u_2/c_2 < 1$, downstream. The flow can be supersonic downstream of shock waves oriented obliquely (see Sec. 7.1) to a supersonic flow, since only the normal component of velocity changes across a shock. Flow behind a moving shock wave may also be supersonic relative to a stationary observer. Pressure, density, and temperature always increase across shocks in normal fluids.

Solutions to this equation are known as *shock adiabats* or *Hugoniots*. Shock adiabats can be determined from thermodynamic data $h(P,v)$ by solution of the Rankine–Hugoniot equation. In the case of the perfect-gas model, analytic solutions can be obtained and tabulations are available for many values of the specific-heat ratio γ. The results for $\gamma = \frac{7}{5}$, used for low-temperature air, are shown in Fig. 4. For perfect gases, pressure and temperature increase indefinitely as the shock strength M_1 is increased. However, the density and Mach number reach finite, limiting values ($\rho_2/\rho_1 \rightarrow 6.0$ and $M_2 \rightarrow 0.378$) for large shock strengths. For a general perfect gas,

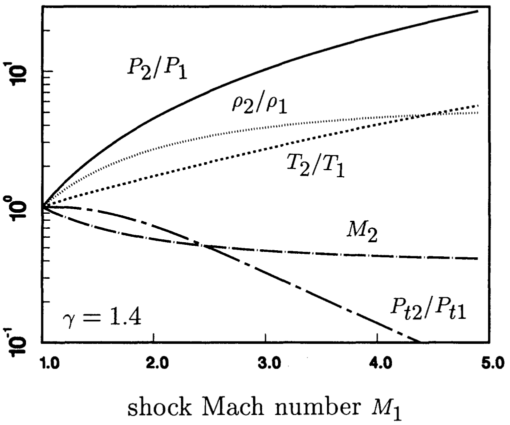

FIG. 4. Solutions to shock jump conditions, perfect-gas model, $\gamma = 1.4$. The Mach numbers M_1 and M_2 are for the velocity component normal to the shock front as measured in a coordinate system attached to the wave.

$$\frac{\rho_2}{\rho_1} \rightarrow \frac{\gamma+1}{\gamma-1} \text{ and } M_2 \rightarrow \sqrt{\frac{\gamma-1}{2\gamma}} \text{ as } M_1 \rightarrow \infty. \tag{19}$$

These limits are consequences of the equipartition of the energy in the incoming flow into internal and kinetic energies in the shocked fluid when the shock strength is large. Stagnation pressure, P_{t2}/P_{t1} in Fig. 4, always decreases across shock waves. This is due to the entropy increase across the shock and represents a source of losses in supersonic flows containing shocks.

For dense fluids and liquids, numerical solutions can be readily obtained using tabulated or empirical equations of state. In the case of solids, the experimental measurements needed for Rankine–Hugoniot computations are often obtained from actual shock-wave propagation measurements. Consider a shock wave propagating with velocity U_s through a material initially at rest. Behind the shock wave, the fluid particles are set in motion with velocity u_p. This moving shock can be considered as equivalent to a stationary wave through a Galilean transformation, yielding the velocities $u_1 = U_s$, $u_2 = U_s - u_p$. Shock data for solids are usually given in form of tabulated (U_s, u_p) pairs. These data (March, 1980) are often found to obey the empirical expression $U_s = c_0 + s u_p$, in which s is an empirical parameter, approximately equal to $\Gamma/2$. Shock data are frequently combined with information about G and K_s to produce an approximate *thermoelastic* or *Mie–Grüneisen* (Zel'dovich and Raizer, 1966) equation of state used for numerical solutions of the Euler equations.

A shock wave is an example of an irreversible natural process and, in addition to the jump conditions, must satisfy the second law of thermodynamics,

$$s_2 > s_1. \tag{20}$$

The amount of entropy generated by the shock depends on the strength of the shock wave. This can be measured by either the shock Mach number U_s/c_1 or any of the nondimensional jumps such as $[P]/\rho_1 c_1^2$ or $[v]/v_1$, where $[P]=P_2-P_1$, $[v]=v_2-v_1$. Shock adiabats are initially close to isentropes, so that weak shock waves can be considered as acoustic disturbances. Shock adiabats progressively deviate from the isentropes as the shock strength is increased. The entropy jump $[s]=s_2-s_1$ is third order in the shock strength,

$$\frac{T_1[s]}{c_1^2}=\frac{\Gamma_1}{6}\left(\frac{[P]}{\rho_1 c_1^2}\right)^3+\cdots, \tag{21}$$

so that the effect of nonlinearity [the second term in Eq. (1)] is more significant than entropy generation in weak shock waves. It is this condition that dictates that shock waves in normal fluids ($\Gamma > 0$) are compressive waves and in unusual fluids ($\Gamma < 0$) are expansive (rarefaction) waves.

A mechanical interpretation of the entropy jump across the shock is possible if we consider the molecular processes involved. Large gradients of temperature and velocity occur within the shock. Molecular transport of energy and momentum across these gradients results in dissipation throughout the shock front region. The dependence of $[s]$ on shock strength given above suggests that the shock layer becomes very thick for weak shocks, $M_1 \rightarrow 1$. Experimental studies of shock structure in gases confirm this, and simple computations demonstrate (Thompson, 1983) that the thickness of strong shocks in gases can even be predicted by continuum estimates of the dissipation rate. The detailed structure of such waves does require the solution of kinetic equations (Bird, 1976) describing the evolution of the distribution function of the molecular velocities. In the case of shocks in complex molecular systems, coupling between vibration–rotational and translation states can result (Vincenti and Kruger, 1986) in unusual wave structures. Low-amplitude waves are *fully dispersed* structures that contain continuous compression waves. Intermediate-amplitude waves are *partially dispersed* structures with both continuous compression and shock jumps. Large-amplitude waves consist of a shock followed by a zone of equilibration ("relaxation") between energy excitation modes.

Compression and entropy generation in strong shock waves will result in a substantial temperature increase across the shock wave. Strong shock waves ($M > 5$) can be produced in gases by use of laboratory shock tubes. The temperature jump across such waves is sufficient (> 1200 K) to initiate molecular dissociation and chemical reactions rapidly (Zel'dovich and Raizer, 1966). Explosion (see EXPLOSIVES) or detonation can occur when strong shock waves are created in combustible mixtures. Ionization and radiative transport of energy are important features of very strong shocks (Zel'dovich and Raizer, 1966), particularly in determining the structure of the shock front.

High explosives or high-velocity impact are used in solids to generate strong shock waves. A variety (Davison and Graham, 1979) of unique electrical, mechanical, and chemical effects can be produced in this fashion. It is possible to simulate conditions (pressures of up to 3.5 Mbar) encountered in geological materials (Jeanloz, 1990) deep within planetary cores. Unique phenomena (Liepmann and Laguna, 1984) occur when superfluid liquid helium II is shocked. The two-fluid nature of helium II gives rise to two shock waves that travel at distinct velocities. One shock wave is a pressure wave with very little temperature rise and the other is a temperature wave at nearly constant pressure.

5. DETONATION WAVES

Detonation waves can be envisioned as shock waves induced and sustained by exothermic chemical reaction, i.e., combustion. The high temperature behind the shock wave initiates the combustion reactions, and the volume expansion resulting from the reaction acts as a piston to drive ahead the shock wave. Gas, liquid, and solid explosives are all capable of detonation and have been extensively studied (Fickett and Davis, 1979; Gibbs and Popalato, 1980). Detonation waves are observed under many conditions to travel with a constant velocity characteristic of the material and the initial conditions. Detonation velocities are 2–3 km/s for gases and up to 8 km/s in solids or liquids. Pressures behind gaseous detonations are 15–30 times the initial pressure. Pressures behind solid or liquid explosives can be as large as 400 kbar.

Detonation velocities can be predicted by supposing that the fluid changes composition within the shock layer and is in chemical equilibrium downstream. Detonation adiabats can be obtained by combining the shock jump conditions with an appropriate multicomponent equation of state and chemical equilibrium relations for the downstream state. The solution for shock velocity is nonunique with the exception of a single distinguished case, the minimum possible wave speed, known as the *Chapman–Jouguet velocity.* The Chapman–Jouguet solution has the special feature that fluid velocity downstream is sonic with respect to the wave and the entropy jump is an extremum. Physically, this can be interpreted as providing a sonic barrier to prevent acoustic disturbances from disrupting the wave from the rear.

Computations of the Chapman–Jouguet (CJ) velocity are readily carried out for gases by numerical methods. Experimental velocities are found to be close (0–5 % below) to the theoretical values computed in this fashion. Detonation products for solid and liquid explosives are in a highly compressed dense-gas state and the thermodynamic computations are less reliable than in gases. Modern statistical mechanical methods based on thermodynamic perturbation theory (Ree, 1984) are moderately successful for detonation velocity prediction in these cases. Of great importance to high explosives is the yielding of the confining container, which can produce a significant influence on the propagation velocity and even result in detonation failure. As the diameter of a cylindrical charge of high explosive is decreased, the detonation velocity decreases until a *critical* or *failure diameter* is reached. Steady detonation is found to be impossible for charges smaller than the failure diameter (Gibbs and Popalato, 1980).

Zel'dovich, von Neumann, and Döring (ZND) independently formulated the theory (Fickett and Davis, 1979) of steady, laminar detonation waves. In this ZND model, the detonation consists of a nonreactive shock wave traveling at the CJ velocity, followed by a reactive boundary layer or reaction zone in which reactants are converted to products. However, both the CJ and ZND models conceal the true unsteady character of detonations. It is observed experimentally and predicted from stability analyses that transverse shock-wave disturbances spontaneously appear on a planar detonation front and that propagating gaseous detonations almost invariably display this instability. The quasiperiodic system of transverse waves forms a cellular pattern with a characteristic wavelength that is indicative of the chemical sensitivity of the mixture. Computations using detailed chemical kinetic models of the combustion process demonstrate (Westbrook and Urtiew, 1982) that the instability wavelength is proportional to the length of the chemical reaction zone computed on the basis of the ZND model. Other types of transient behavior such as initiation and failure are intimately linked (Lee, 1984) to the length scale and dynamics of the reaction zone. The simplest notion is that explosives with larger reaction zones are the least sensitive, have the greatest instability wavelength, and have the largest failure diameters. These ideas are relevant to both gaseous and condensed explosives.

6. ONE-DIMENSIONAL UNSTEADY FLOWS

Analytic solutions to the unsteady Euler equations can be obtained for flows with spatial variation in only one direction. This is particularly true for planar flow since the equations can be written in *characteristic form* and solved exactly for several cases of technical importance. The characteristic form of the Euler equations can be compactly written in the case of nonreactive flow as

$$\frac{du}{dt} \pm \frac{1}{\rho c}\frac{dP}{dt} = \mp \frac{n}{x}\, cu \quad \text{on} \quad \frac{dx}{dt} = u \pm c, \tag{22}$$

$$\frac{ds}{dt} = 0 \quad \text{on} \quad \frac{dx}{dt} = u. \tag{23}$$

The dimension index n takes on a value of 0 for planar flow, 1 for cylindrical flow, and 2 for spherical flows. The time derivatives in Eqs. (22) and (23) are to be interpreted as derivatives obtained by an observer traveling on the lines in space-time defined by the velocities $u \pm c$ and u, known as the *characteristics* $C^{\pm}$ and C^0. Physically, the characteristics associated with velocities $u \pm c$ correspond to the speed of adiabatic disturbances in pressure and velocity *relative to the flow*. Entropy disturbances are associated with fluid motion itself and travel with the particle velocity u. In the present case of adiabatic flow, the entropy of each element remains unchanged with the exception of passage through shock waves. Equation (23) does not apply at the location of the shock, and the jump conditions, Eqs. (15)–(17), must be applied to determine the changes across the shock.

A set of characteristics passes through each point in the flow. The wedge-shaped region determined by characteristics C^+ and C^- coming from the past defines the region of fluid that can influence motion at that point. The projection of the characteristics into the future determines the region that will be influenced by disturbances at that point. For adiabatic, nonreactive flows it is possible to define a quantity F

$$F(P,s_0) = \int_{P_0}^{P} \frac{dP'}{\rho c}, \tag{24}$$

and to simplify further the equations for pressure and velocity to read

$$\frac{d}{dt}(u \pm F) = -\frac{n}{x}cu \quad \text{on} \quad \frac{dx}{dt} = u \pm c. \tag{25}$$

The right-hand side vanishes in planar flows, $n=0$, and the quantities $u \pm F$ are known as the *Riemann invariants*. The Riemann invariants are constant on the corresponding characteristics, except when the characteristics cross shock waves. The value of the invariant must change across the shock in order to be consistent with Eqs. (15)–(17). For the perfect-gas model, the quantity F can be computed in closed form and is equal to $2c/(\gamma-1)$.

As an application of the theory, consider the initial-value problem associated with a discontinuity in pressure separating uniform regions of gas, the so-called *Riemann problem* of compressible flow. This situation occurs in *shock tubes* (Oertel, 1966), used in the laboratory study of shock waves. A closed, gas-filled tube is divided into two sections by a thin diaphragm of metal or plastic. The *driver* section is filled with high-pressure gas. On the other side of the diaphragm is the low-pressure *driven* section. A space-time diagram and the characteristic network associated with this process are shown in Fig. 5.

The diaphragm is ruptured by either overstress or mechanical puncturing. Referring to Fig. 5, a compression wave (shock) propagates into the low-pressure region (1) and an expansion wave, denoted by the "fan" of characteristics C^-, propagates into the high-pressure region (4). A contact surface (dashed line) separates shocked gas (2) from the accelerated driver gas (3). The shock reflects from the end wall to produce a stagnant region (5) in which testing is usually carried out. The reflected-wave interaction (a) with the contact surface produces a secondary shock between regions (5) and (6) that limits the test time to a value τ as shown. The actual test time is significantly influenced by real fluid effects such as the boundary layer behind the shock wave. By varying the pressure and gas composition in driver and driven sections, a wide range of shock strengths can be produced.

For an ideal shock tube, a shock wave is formed immediately on the low-pressure side of the diaphragm. However, if the diaphragm in our example shock tube were to rupture slowly, a shock-free compressive wave could be produced that ultimately would steepen and form a shock wave at some distance from the diaphragm. This process is illustrated in Fig. 6 for a piston (P), analogous to the contact surface in the shock tube, being accelerated into an initially undisturbed region. The particles are accelerated when the first signal arrives along the C^+ characteristic as shown by the trajectory (dashed line) or particle path PP. The slope of the C^+ characteristics decreases as time increases and the piston accelerates. Adjacent characteristics first cross at point (a), producing a shock wave. The point of shock formation can be computed from the piston path and the Riemann invariant on both C^+ and C^-.

The nonlinearity of convection and the thermodynamic dependence of sound speed [Eq. (5)] are both significant to shock formation. Solution of the initial-value problem for

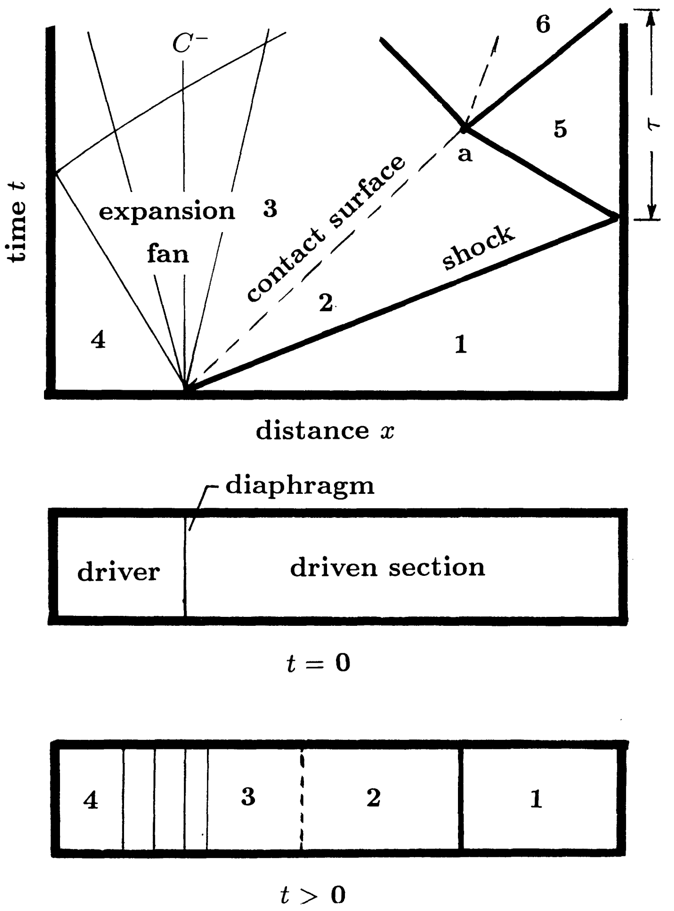

FIG. 5. Space-time (*x-t*) diagram for shock-tube operation.

a simple compressive wave, as shown in Fig. 6, yields the result that the characteristic velocity changes with pressure,

$$\frac{d}{dP}(u+c)=\frac{\Gamma}{\rho c}\,. \tag{26}$$

This leads to the following conclusions about the evolution of pressure disturbances. Compression shocks form when $\Gamma > 0$, shocks never form in the exceptional case $\Gamma = 0$, and expansion shocks form if $\Gamma < 0$.

Another feature of the Riemann problem is the spreading wave or *expansion fan* that propagates into the high-pressure region. This flow is characterized by a self-similar motion dependent only on the coordinate x/c_0t, which results in acceleration of fluid towards the shock wave. The fluid isentropically expands as it accelerates. This unsteady expansion process will result in a maximum particle velocity or *escape speed* of $u_e = 2c_0/(\gamma-1)$ for perfect gases. This is achieved when the low-pressure region is a vacuum. An interface or *contact surface* is created between the fluid originally in the high-pressure region and the shocked fluid. Pressure and velocity are continuous across this interface, but the entropy and composition may change arbitrarily.

In the Riemann (shock-tube) problem, the key elements of many compressible flow problems appear: spontaneous formation of contact surfaces, expansion waves, and shock waves. These features are universal and appear in both one- and higher-dimensional flows. The shock-tube flow is the prototype for many similar technical problems such as *water hammer*, which is simply wave motion in fluid-filled pipes induced by sudden valve operation or other transients. These problems can be analyzed (Moody, 1990) by characteristic methods or other numerical techniques. For small-amplitude waves, a linear analysis suffices and the characteristic relations reduce to $dP = \pm\rho_0c_0du$ for waves moving to the right

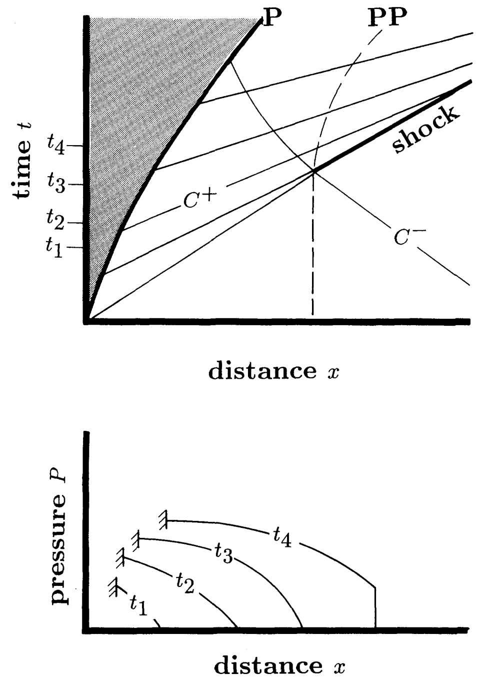

FIG. 6. Compression wave steepening to form a shock, illustrating crossing of characteristics.

$(dx/dt = +c_0)$ and the left $(dx/dt = -c_0)$, respectively. The quantity ρc is referred to as the *acoustic impedance* of the fluid. The ratio of the fluid and solid acoustic impedances determines the pressure produced when a wave interacts with a fluid–solid boundary. At hard boundaries, a large increase in impedance occurs and total reflection results in doubling the pressure. At soft boundaries, a large decrease in impedance occurs and the pressure changes sign. The reflected wave has equal magnitude to the original, i.e., a compressive wave reflects as an expansion and vice versa.

Similarity solutions have played an important role in the development of compressible-flow theory. Perhaps the best-known example is the solution (Taylor, 1950) for the propagation of a decaying spherical shock wave produced by a rapid release of energy within a uniform atmosphere. This solution utilizes the strong shock approximation to the jump conditions, in which the initial pressure in the atmosphere is neglected. If an amount of energy E is suddenly released into the atmosphere of density ρ_0, the radius R_s of the resulting shock wave is proportional to $(E/\rho_0)^{1/5}t^{2/5}$. The pressure P_s immediately behind the shock can be shown to decay inversely with the cube of the distance, $P_s/P_0 \sim EP_0^{-1}R_s^{-3}$. Many variations on this problem can be solved analytically through similarity transformations, including the inverse problem of collapsing or imploding shock waves (Whitham, 1974).

The Taylor solution is an example of the decay of a transient shock or *blast wave* created by an explosion. Blasts from accidental and deliberate explosions have been extensively studied (Kinney and Graham, 1985). Shock pressures from intense explosions follow scaling suggested by the Taylor solution $P_s/P_0 \sim (R_e/R_s)^n$ in which the scaling distance $R_e = (E/P_0)^{1/3}$. In regions close to the energy source, the exponent $n \sim 3$ as given by the Taylor solution. Far from the source, the exponent $n \sim 1$ as given by the acoustic solution. Solutions uniformly valid in the near and far fields can be constructed by the technique of *nonlinearization* (Whitham, 1974). Linear (acoustic) solutions are used for amplitudes but are evaluated on the nonlinear characteristics described by Eq. (25). Shock waves are inserted at the points of characteristic crossing. The shock amplitude is chosen so that the conservation laws, Eqs. (15)–(17), are satisfied.

The dependence of the scaling distance on energy released leads to the concept of *cube-root* scaling for blasts produced by high explosives. Blast pressures from two different charges (mass W) will be equal if the scaled distance $R_sW^{-1/3}$ are the same. Such scaling relations are valuable engineering tools (Kinney and Graham, 1985) for estimating blast waves in air and water. In addition to similarity solution methods, the very useful approximate technique of *shock dynamics* (Whitham, 1974) can be used to solve for shock-wave motion in complex geometries (Schwendeman, 1988).

Because of the importance of one-dimensional compressible flow in technical problems as diverse as plumbing systems and nuclear weapons, numerous methods have been developed for numerical solutions of these flows. A reliable, but cumbersome, method is based directly on the method of characteristics (Ritchmeyer and Morton, 1967) with insertion of shock waves when dictated by characteristic crossing. More flexible but potentially less accurate methods are based on finite-difference approximations to the Euler equations. The accuracy of the technique is determined by the ability to resolve shock waves and contact surfaces. A crude but useful version is to employ *artificial viscosity* to smooth out shock waves. Many more sophisticated techniques such as the methods of random choice (Sod, 1985) and of flux-corrected transport (Oran and Boris, 1987) and the piecewise parabolic method (Collela and Woodward, 1984) have been developed in recent years. Numerical techniques developed for one-dimensional nonreactive flows can be extended to solve reactive, multidimensional flows by the techniques of operator splitting. A solution is built up through a sequence of one-dimensional solutions for each physical process and dimension in space. Very complex problems such as transient supersonic jets, oblique shock reflection, and detonation instability can be treated in this fashion.

7. MULTIDIMENSIONAL FLOWS

Compressible flow in two and three dimensions is of great significance in designing aircraft, space vehicles, propulsion systems, and experimental facilities such as wind tunnels. Realistic design requires including transients, real gas effects, and complex geometrical configurations. Solution of such problems requires a concerted effort with both experimental investigation using scale models and numerical solutions to some approximation such as the Euler equations. However, it is possible to indicate the features of such flows by examining the analytic solutions possible for steady flows of a perfect gas.

7.1 Steady Supersonic Flows

An essential distinction for steady multidimensional flows is the presence or absence of *vorticity*, $\mathbf{\Omega}=\text{curl}\,\mathbf{u}$, in the flow. As in low-speed, incompressible flow, vorticity is a measure of rotational relative motion (rate of strain) within the fluid. Rotational motion can only be created in an initially irrotational fluid by the action of viscous forces at boundaries or by the existence of entropy gradients within the flow. The Euler equations can be used to deduce a transport equation for vorticity, the *Crocco–Vazsonyi equation*:

$$\frac{\partial\mathbf{\Omega}}{\partial t}+\mathbf{\Omega}\times\mathbf{u}=T\,\text{grad}\,s-\text{grad}\left(h+\frac{u^2}{2}\right). \qquad (27)$$

For steady flow, this implies that vorticity exists solely due to gradients in total enthalpy or entropy in the direction transverse to the local flow direction $\mathbf{\Omega}\times\mathbf{u}=T\,\text{grad}\,s-\text{grad}(h+u^2/2)$. Total enthalpy is conserved in steady adiabatic flows (with or without shocks) since the Euler equations directly yield $\text{div}(h+u^2/2)=0$. Therefore any variation in h_t must originate upstream and be a consequence of nonuniform boundary conditions. Transverse entropy variations can spontaneously arise from the presence of shock waves of varying strength. A single curved shock wave or a system of transversely nonuniform waves are the most common origins of such entropy variation. We conclude that general multidimensional flows containing curved shocks must also be rotational flows and cannot be isentropic even if the upstream flow is uniform.

We conclude that steady, adiabatic, supersonic flow can be treated as isentropic in the absence of curved shock waves or viscous effects. Therefore the Euler equations can be written in characteristic form. For simplicity, only two-dimensional flows will be considered here. This is most readily done in the *natural coordinate system* (s,n) in which s is distance measured along the flow direction or *streamline* and n is the distance measured normal to the streamlines; see Fig. 7. Characteristic directions are oriented at angles $\pm\mu$ to the main flow direction (Fig. 7), where μ is the angle that a weak disturbance or *Mach line* will make with a uniform flow,

$$\mu=\sin^{-1}(1/M). \qquad (28)$$

These characteristic directions in space are analogous to the characteristic directions in space-time introduced in the previous discussion on unsteady one-dimensional flow. The similarity of the governing equations results in a very close analogy between one-dimensional unsteady flow and two-dimensional steady flow. Primitive variables for the characteristic equations in natural coordinates are the magnitude of the flow velocity u and its orientation θ with respect to a fixed axis. The equations can be compactly written as

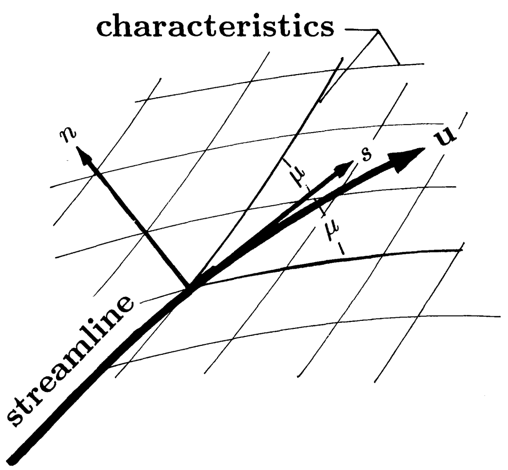

FIG. 7. Natural coordinate system (*s*,*n*) for two-dimensional flow showing Mach lines for supersonic flow.

$$\frac{\partial}{\partial s}(\omega \pm \theta) \mp \tan\mu \frac{\partial}{\partial n}(\omega \pm \theta) = 0, \qquad (29)$$

using the *Prandtl–Meyer function* ω defined by

$$d\omega = \frac{dM}{M}\frac{\sqrt{M^2-1}}{1+(\Gamma-1)M^2}. \qquad (30)$$

The quantities $\omega \pm \theta$ are the Riemann invariants for this theory. The results of quasi one-dimensional flow theory find direct application in steady multidimensional problems. The differential relationships between thermodynamic and flow variables defined in Sec. 4.1 are valid along the *stream tubes* defined by surfaces of constant n. The energy equation (12) can be used to relate the Mach number M to the thermodynamic state in order to evaluate Γ and complete the integration for the Prandtl-Meyer function.

The Mach number appears as a parameter just as in one-dimensional flow. The distinction between subsonic and supersonic flows is of great importance. Characteristic directions and shock waves can exist only when the flow is locally supersonic. Mathematically, this is a result of the *hyperbolic* character of equations in which all characteristics are real. For steady subsonic flows, the equations are *elliptic* and no real characteristics exist.

The characteristic directions, $dn/ds = \pm\tan\mu$, define the cone of influence for a disturbance generated at a point. Only locations inside of this *Mach cone* can be influenced by events at the apex of the cone. These directions also define the characteristics along which the Riemann invariants $\omega \pm \theta$ are constant. The shock jump conditions discussed in Sec. 4.2 imply that the Riemann invariants discontinuously change when the characteristics cross shock waves.

Solutions for simple changes in flow direction can be constructed for initially uniform flows by use of the Riemann invariants originating in the undisturbed flow. Consider a flow moving from left to right in the x direction. Flow deflections $d\theta$ are related to changes in the Prandtl–Meyer function and velocity by

$$d\omega = \frac{du}{u}\sqrt{M^2-1} = \pm d\theta \qquad (31)$$

along the characteristic directions $dy/dx = \pm\tan\mu$. Compressive turns $du<0$ result in velocity decreases associated with stream-tube contraction in supersonic flow. Expansive turns $du>0$ result in velocity increases associated with stream-tube expansion in supersonic flow. For the perfect-gas model, $\omega(M)$ can be computed in closed form. Tabulations are available for many values of γ. A maximum change in ω is associated with the maximum velocity that can be obtained in quasi one-dimensional flow expansion. For a value of $\gamma = 1.4$, $\omega_{\text{max}} = 130.45°$ for acceleration from an initially sonic flow.

A strong compressive turn will result in the crossing of adjacent characteristics of the same family, generating a shock wave just as in the one-dimensional unsteady case. This wave will, in general, be an ***oblique shock***. That is, the wave is not normal to the flow direction. As discussed in Sec. 4.2, only the component of the flow velocity in the direction normal to the shock is changed. An oblique shock at an angle β (Fig. 8) with respect to the incident flow (velocity U_1) is therefore equivalent to a shock at normal incidence to a flow with velocity $u_1 = U_1 \sin\beta$. The flow is deflected through an angle θ which depends on β and u_1.

With this simple geometric transformation, all oblique shocks can be analyzed as equivalent normal waves using the methods described previously. A consequence of the transformation is that the flow deflection is a multivalued function of wave angle and a maximum flow deflection exists at a critical

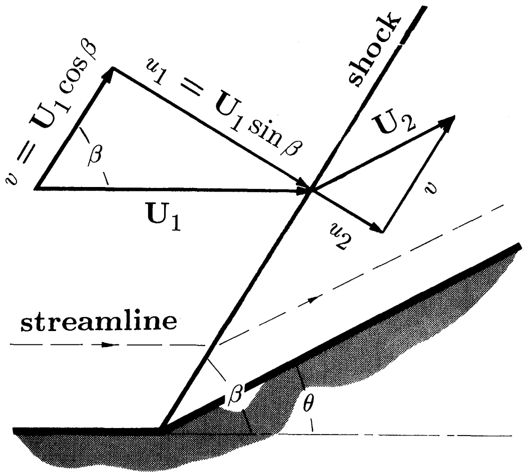

FIG. 8. Oblique-shock velocity diagram, showing wave angle β and resolution into normal and tangential components.

value of the wave angle, which is dependent on the shock strength. For strong oblique shocks in a perfect gas, $\gamma = 1.4$, the maximum flow deflection has the largest value of 45.5° and occurs at $\beta = 68°$. The transition between downstream subsonic and supersonic flow occurs at a critical wave angle β^* that is close to occurrence of maximum deflection. Oblique shock waves with $\pi/2 > \beta > \beta^*$ result in locally subsonic flow downstream of the wave and are known as *strong* waves. Oblique shocks with $\beta^* > \beta > 0$ are locally supersonic downstream and are known as *weak* waves. These properties of oblique shocks play an important role in determining the flow over supersonic projectiles.

7.2 Aerodynamics

Aerodynamics (*q.v.*) is the study of fluid motion over bodies such as wings, fuselages, missiles, and entire airplanes. Flow over streamlined surfaces is characterized by the absence of flow *separation* from the body, and the effects of viscosity are confined to thin *boundary-layer* and wake regions. The flow is usually assumed to consist of two regions. Far from the surfaces and the wake is the *outer* region which is usually treated as inviscid and adiabatic, approximately governed by Euler's equations. Within the wake and near the surfaces is an *inner* region in which the effects of viscosity and heat conduction dominate. Some form of the Navier–Stokes equations must be applied within these regions, and the problem of turbulent flow must usually be addressed.

The outer region is initially computed by supposing that the viscous layer is infinitely thin and that the flow next to the surface must follow the surface contour but *slips* on the surface, i.e., has a nonzero velocity parallel to the surface. The inviscid flow is then used to compute the viscous boundary layer. The displacing effect of the inner boundary layer can then be used to correct the outer flow for viscous effects. In supersonic flow, this procedure is complicated by the presence of shock waves in the outer flow and the interaction of shocks with the boundary layers, which in some cases, can lead to flow separation.

7.2.1 Linearized Flow Theory. Aerodynamic design is a mature subject that uses sophisticated numerical simulations and wind-tunnel testing to develop optimal shapes for lifting surfaces (wings) and entire vehicles. There is also a substantial body of theory on approximate analytical solutions for *thin* or *slender* shapes such as wings and fuselages. One class of theories involves the actual mechanism of lift in subsonic flight, intimately related to the production and maintenance of vorticity in the fluid. Another class of theories attempts to describe the influence of compressibility on the flow, without explicitly considering vorticity.

We are concerned with the second category, which analyzes *irrotational* and isentropic motion. Shock waves can be accommodated by this theory as long as they are weak. Solutions based on potential flow notions are very useful in aeronautics. For wing sections, these theories give very useful estimates of lift, and the drag induced by lift. They also provide the outer boundary conditions necessary for the viscous boundary-layer computations. Linearized analyses usually come to the happy conclusion of an analytical solution that is invaluable in correlating and understanding the experimental data. The irrotational and compressible analyses have particular relevance to near-sonic and supersonic flight. For supersonic and hypersonic flight, the actual mechanism of lift (and drag) is almost exclusively determined by compressibility effects rather than the vorticity mechanism of subsonic flight.

Consider a streamlined body at rest being approached by a flow that is uniform with velocity U_∞ at large distances upstream from the body. If the flow is irrotational, curl$\mathbf{u}=\mathbf{0}$, then we can use a velocity potential Φ to define $\mathbf{u}=\text{grad}\ \Phi$. Linearized flow is the case in which the body has a small aspect ratio (transverse thickness divided by length) and therefore produces only a small perturbation to the incoming flow. Using a velocity potential $\Phi=U_\infty x+\phi$ and linearizing the momentum equation for isentropic flow in a perfect gas, we find that in Cartesian coordinates (x,y,z) the approximate equation for the perturbation potential ϕ can be written

$$(1-M_\infty^2)\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}+\frac{\partial^2\phi}{\partial z^2}=(\gamma+1)M_\infty^2\frac{1}{U_\infty}\frac{\partial\phi}{\partial x}\frac{\partial^2\phi}{\partial x^2} \qquad (32)$$

in which $M_\infty=U_\infty/c_\infty$ is the freestream Mach number. In the case of two-dimensional flows, used in idealized wing theory, the z derivative can be omitted or else the z and y terms can be combined to produce a radial derivative used in the axially symmetric theory for fuselages and missiles. The term on the right-hand side is nonlinear and required only in the *transonic flow* regime, $M_\infty\sim 1$. Solutions for subsonic and supersonic flow are greatly simplified by omitting this term. The main results of the theory, based on solutions to the linearized potential equation, are outlined in the subsequent sections.

7.2.2 Subsonic Flow. Subsonic, compressible flow, $M<1$, is of great technical significance in the design of the modern passenger aircraft, in which $M_\infty\sim 0.78$–0.82 during high-altitude cruise. Acceleration of the flow over the lifting surfaces of the aircraft results in local regions of sonic and supersonic flow, which must be treated as a transonic flow problem, discussed in Sec. 7.2.4. A critical flight Mach number is reached when sonic flow first occurs on the lifting surface. Compressibility effects are manifested before the critical Mach number as an increase or apparent *divergence* in the drag coefficient. This important effect can be understood by simple dimensional analysis of the subsonic potential equation.

The linearized potential equation for subsonic flow can be written as a *Laplace equation*

$$\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y'^2}+\frac{\partial^2\phi}{\partial z'^2}=0, \qquad (33)$$

in which the scaled coordinates are $y'=y\sqrt{1-M_\infty^2}$ and $z'=z\sqrt{1-M_\infty^2}$. The pressure P is conventionally presented in aerodynamic analyses as a nondimensional *pressure coefficient*

$$C_P=\frac{P-P_\infty}{\frac{1}{2}\rho_\infty U_\infty^2}\approx-\frac{2}{U_\infty}\frac{\partial\phi}{\partial x}. \qquad (34)$$

Consider a low-Mach-number ($M_\infty\to 0$) subsonic flow which results in a measured pressure coefficient $C_{P,0}$. If this flow is governed by the subsonic potential equation, the scaling implicit in the spatial transformation used to obtain the Laplace equation implies that for any finite Mach number the pressure coefficient will be given by

$$C_P=C_{P,0}/\sqrt{1-M_\infty^2}. \qquad (35)$$

This is the *Prandtl–Glauert rule* of subsonic flow, clearly indicating the divergence effect as $M_\infty\to 1$. Since lift and drag are simply the net forces, pressure and viscous shear, integrated over the body surface, the drag divergence phenomenon follows. The viscous contribution will also change with Mach number but not directly in a divergent fashion; rather, indirectly through the phenomenon of boundary layer separation. The drag in question is induced drag due to lift since the form or shape drag is known to be zero (d'Alembert's paradox) for potential flows. More sophisticated drag correction rules can be obtained by integrating the correction factor over the surface or using other approximations to the linearized potential equation.

The drag divergence effect described by the linearized potential equation arises purely because of the rise in surface pressures as the local Mach number approaches unity. This is a reliable guide as long as the flow everywhere remains subsonic. For supercritical flows, realistic computations of the compressibility effect must incorporate other significant effects. There is a loss in total pressure if shock waves are formed. The pressure rise on the surface will result in thickening or even separation of the boundary layer due to the adverse pressure gradient. Separation of the boundary layer will result in a large *form* drag associated with the change in the pressure distribution due to the presence of reverse flow region downstream of the flow separation point. It is this effect which is primarily responsible for the severe degradation of performance when airfoils designed for subsonic flight are used near $M=1$.

7.2.3 Supersonic Flow. Supersonic flow, $M>1$, results in a wave equation for the potential function

$$\frac{\partial^2\phi}{\partial y'^2}+\frac{\partial^2\phi}{\partial z'^2}-\frac{\partial^2\phi}{\partial x^2}=0 \tag{36}$$

in which $y'=y\sqrt{M_\infty^2-1}$ and $z'=z\sqrt{M_\infty^2-1}$ are now the scaled coordinates. Solutions to the wave equation are fundamentally different from those of the Laplace equation. However, it is clear from the coordinate transformations in the supersonic case that a divergence in pressure coefficients will also occur as $M_\infty \rightarrow 1$. Consideration of the transonic version of the potential equation is required to resolve this singularity.

Thin wings in supersonic flight can be analyzed via the wave equation, and the results are equivalent to a linearized solution using characteristics. The pressure coefficient of a wing element at a small angle θ to the flow can be computed from Eq. (31),

$$C_P=2\theta/\sqrt{M_\infty^2-1}\,. \tag{37}$$

Airfoils designed for supersonic fight have sharp leading and trailing edges in contrast to the blunt leading edge of subsonic airfoils. A blunt leading edge in supersonic flight is undesirable since it will be proceeded by a detached shock wave with a substantial loss in total pressure. Shock waves emanate from the leading and trailing edge of a thin, sharp-edged airfoil in a supersonic flow. Expansion waves exist between the two shocks, resulting in the *shock-expansion* configuration characteristic of flow field around supersonic bodies. Far from the wing or aircraft, a pressure wave is created that starts and ends with shock waves. The pressure decreases between the shocks, leading to the characteristic *N-wave* pressure profile of a *sonic boom* produced by a supersonic aircraft. The dissipative nature of the shock waves results in an inviscid contribution to the drag.

Flow near the front of a supersonic wedge or cone is of great interest since this is a common configuration, shown in Fig. 9, in vehicles and projectiles. Deflection of the flow by the body results in a shock wave ahead of the body surface. Depending on the angle subtended by the wedge or cone, this shock wave may either be *attached* to the tip of body or *stand off* from it. This behavior is a simple consequence of the maximum turning angle that is possible through an oblique shock, mentioned in the previous discussion. It is possible to obtain similarity solutions for flows over simple wedges and cones when the shock is attached.

The supersonic wedge flow consists of an oblique shock, at a constant angle to the incoming flow, of the appropriate strength to turn the flow parallel to the wedge, and a uniform flow between the shock and the wedge. Supersonic cone flow is similar in that an oblique shock at a constant angle to the incoming flow emanates from the cone tip. However, the flow between the shock and the cone surface is not uniform. The flow does have a similarity character, and the flow field can be obtained from a single differential equation with the similarity variable r/x. This solution reveals that a smooth transition between subsonic and supersonic flow can exist in these conical flows. Such transitions and the associated sonic lines play an important role in the case of a wedge or cone that is blunted or has an included angle that exceeds the maximum turning angle for an oblique shock. In these situations, a subsonic region exists between the shock and body near the tip of the body since the flow behind a normal shock wave is subsonic. Acceleration of the flow over the body can result in a transition to supersonic flow, as shown in Fig. 9. Analysis of such flows is complicated by both the transonic character and the nonlinearity originating in the substantial magnitude of the flow deflections.

7.2.4 Transonic Flow. Transition between subsonic and supersonic flow, the *transonic* flow regime, is of particular significance to modern aerofoil technology. Transonic flow problems are also very challenging to analyze (Cole and Cook, 1986). They are intrinsically nonlinear, i.e., the right-hand side of the po-

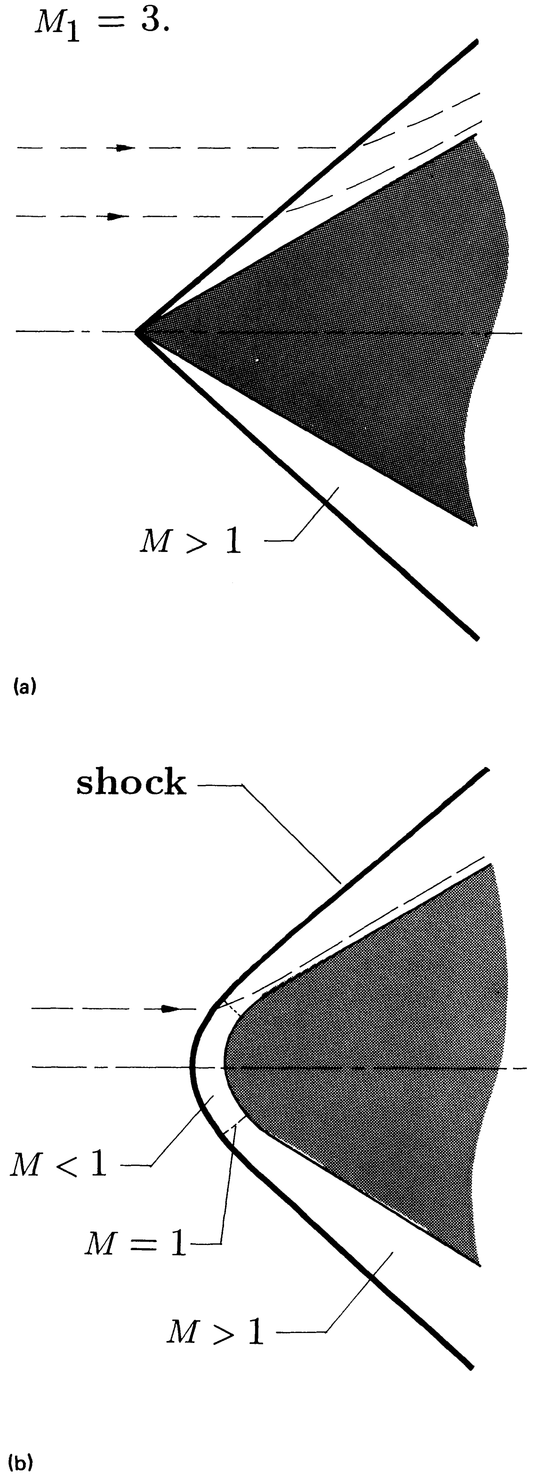

FIG. 9. Flow near the tip of (a) slender and (b) blunt supersonic bodies, illustrating the transition between subsonic and supersonic flow, the distinction between attached and detached shocks, and the streamlines in self-similar flows.

tential equation, Eq. (32), cannot be neglected. The character of the flow and solution methodology changes type, elliptic to hyperbolic, when the sonic line is crossed. Experiments on transonic flows require special care since the cone of influence is quite broad ($\mu \sim 90°$) for near-sonic flow. As a consequence, the configuration of the wind tunnel or testing range is quite significant and substantial corrections for *tunnel interference* are often required.

Supersonic flow will occur over portions of the wings on modern subsonic passenger aircraft. This can result in a shock wave associated with the ultimate deceleration of the supersonic flow. A bubble-shaped region of supersonic flow occurs between the upper surface of the wing, the sonic line, and the shock wave as shown in Fig. 10. By careful choice of the airfoil shape, the critical freestream Mach number at which the shock wave occurs can be increased substantially over conventional designs. With highly optimized surfaces, smooth deceleration from supersonic to subsonic flow, *without a shock wave*, can be obtained. These airfoils, known as *supercritical* designs, can result in substantial drag reduction over more conventional designs and are extensively used on modern aircraft. Design procedures for supercritical airfoils have been developed in some detail (Sobieczky and Seebass, 1984). The design process requires determining an inviscid flow through a solution of the Euler equations, then computing an actual wing profile by correcting for the computed viscous boundary-layer thickness.

7.3 Hypersonic and Hypervelocity Flows

When the Mach number of a vehicle or projectile exceeds a value of $M \sim 5$, the flow is considered to be *hypersonic* (Anderson, 1989), a flow regime encountered in spacecraft launch and reentry. In such flows, the characteristic features are very strong shock waves that dominate the flow field. Lift and drag in hypersonic flows are almost exclusively produced by the high-pressure region generated by the shock on the lower surface of a lifting body. The upper surface is at a very low pressure due to the strong Prandtl–Meyer expansion flow created at the leading edge. Hypersonic flows are relatively simple for vehicles with a sharp nose or leading edge and

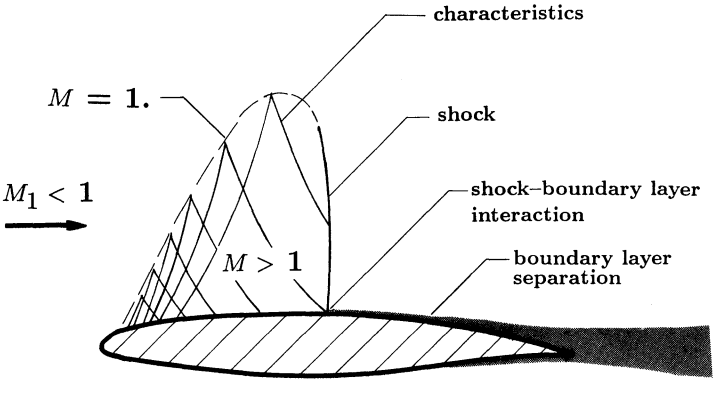

FIG. 10. Principal features of transonic flow over an airfoil.

a small inclination (angle of attack) to the incoming flow. In that case, the region between the shock and vehicle is very narrow since the Mach lines and shock waves are inclined only slightly relative to the flow direction. The region between the shock and the lower surface of the body contains high-enthalpy fluid that has been processed by the oblique shock. Closer to the vehicle can be an *entropy* layer, consisting of fluid processed by the bow shock; and finally, adjacent to the vehicle will be a viscous boundary layer in which friction heating further increases the temperature of fluid. Vehicles with blunt noses (reentry vehicles) will have detached shock and a rather thick associated entropy layer.

Hypersonic flows are distinguished by a large Mach number, although the actual flow velocity may be modest, $u<4$ km/s. This is typical of flows in laboratory shock tunnels with stagnation enthalpies less than 8 MJ/kg. Molecular dissociation and ionization effects are often negligible under these conditions and the gas can be treated as ideal. Hypervelocity flows are distinguished by large velocities, $u \geqslant 8$ km/s, but could have modest Mach numbers. Recall that it is the kinetic energy, $u^2/2$, that is significant in the energy equation, so that these two regimes differ by a factor of 4 in energy. More significant is the strong dependence of chemical reaction rates on temperature. Chemical reaction is much more important in hypervelocity flows, and effects such as dissociation can dominate the observed flow phenomena (Stalker, 1989). In hypersonic flows, the shocks can all be described as strong and the downstream fluid characterized by the limiting density ratio, but hypervelocity flows are better characterized (Stalker, 1989) by a constant-temperature flow downstream of the shocks. It is possible for the two flow regimes to occur simultaneously in the flow over a single body such as a blunt cone.

High-speed flight and atmospheric reentry of space vehicles are the primary applications of hypersonic and hypervelocity research. Atmospheric reentry occurs under both conditions simultaneously, $M=25$ and $u=8$ km/s for a typical near-earth orbit. Behind the shock waves, the temperatures are high enough that dissociation produces atomic nitrogen N and oxygen O, and chemical reaction produces NO. Gas-surface reactions can also be important (Park, 1990) to the survival of reentry vehicles. Scaling methods have been developed so that laboratory hypervelocity experiments (Stalker, 1989) can be used, particularly in problems where near-normal shocks and blunt bodies are important.

A simple means of estimating lift and drag in hypersonic flows is to assume that the fluid molecules inelastically collide with the lower surface of the vehicle and the component of momentum normal to the surface is completely transferred to the vehicle. This is known as the *Newtonian theory* after Newton's early efforts to model low-speed flows. The shock wave and resulting flow next to the body constitute the physical mechanism that produces the inelastic collision effect. If we apply these notions to a surface inclined at an angle θ to the freestream, the local pressure coefficient is found to be

$$C_p=2 \sin^2 \theta. \tag{38}$$

When we apply this concept to a flat plate with velocity U and an angle of attack α in an atmosphere of density ρ_0, the lift L and drag D forces are found to be

$$L=\rho_0 U^2 \sin^2 \alpha \cos \alpha, \tag{39}$$

$$D=\rho_0 U^2 \sin^3 \alpha. \tag{40}$$

These results can be extended to more complex bodies by taking account of the pressure field induced by curved stream lines (centripetal force) and the nonuniformity of the flow between the shock and the body. Another solution technique is indirect and specifies a *waverider* body shape that is a stream surface of a known flow behind a shock wave of specified shape (Anderson, 1989). Yet another solution method is possible for slender bodies, termed *hypersonic small-disturbance theory* (Hayes and Probstein, 1966). Linearizing the equations in the cross-stream direction effectively reduces two-dimensional hypersonic flow problems to an equivalent one-dimensional unsteady flow problem. Numerical and similarity solution methods developed for one-dimensional unsteady flow can then be directly applied to solutions of hypersonic flow.

Hypersonic flows can be created in the laboratory by a variety of steady-flow wind tunnels, shock tubes, and blowdown wind tunnels. All of these systems create in some fashion, either by combustion, shock heating, or electrical heating, a high-enthalpy, high-pressure reservoir of gas. This gas is then expanded through a de Laval nozzle to produce the desired hypersonic flow. Hypervelocity facilities have the highest stagnation enthalpies, up to 30 MJ/kg. These systems are usually transient in operation since the high-enthalpy fluid must be created by a dynamic process such as shock compression. An example of this type of facility is the *free piston compressor* (Stalker, 1990), in which a massive piston is driven by compressed gas into the driver fluid. The compressed driver gas ($P\sim 1$ kbar) then ruptures a diaphragm and drives a shock through the test gas and out the nozzle at the end of the test section, which initiates a quasisteady flow through the nozzle. The model or experiment is conducted in a vacuum tank surrounding the nozzle. Another technique for generating the driver gas is to use a conventional shock-tube arrangement with high-pressure helium or to burn a combustible mixture initially at high pressure. Finally, an alternative method for producing hypersonic flows is the ballistic range. A projectile is accelerated by a gas or powder gun to hypersonic velocities and injected into a test chamber filled with the atmosphere of interest.

7.4 Unsteady Flows

Many flows of interest in applied physics are unsteady flows with significant multidimensional motion. Great advances have been made in understanding these flows through careful experimentation and numerical simulation of the Euler equations. These simulations are based on numerical solution techniques that can either *capture* shock waves as part of the general solution algorithm or attempt to isolate and treat shock waves and contact surfaces separately through special *front tracking* techniques. Common to all methods is the use of the *conservation form* of the equations, generically written as

$$\frac{\partial f}{\partial t}+\operatorname{div}\mathbf{F}=0. \tag{41}$$

The quantities f and $\mathbf{F}$ are the generalized densities and fluxes. For the Euler equations in two-dimensional Cartesian coordinates with velocity components $\mathbf{u}=(u,v)$,

$$f=\begin{pmatrix} \rho \\ \rho u \\ \rho v \\ \rho u\left(e+\frac{u^2}{2}\right) \end{pmatrix},$$

$$\mathbf{F}=\begin{pmatrix} \rho u & \rho v \\ \rho u^2+p & \rho uv \\ \rho vu & \rho v^2+p \\ \rho u\left(h+\frac{u^2+v^2}{2}\right) & \rho v\left(h+\frac{u^2+v^2}{2}\right) \end{pmatrix}. \tag{42}$$

Typical applications are to the problems of unsteady transient jets, blast propagation, and shock diffraction and refraction. As an example, consider the problem of unsteady shock reflection from a wedge. This is a configuration that often occurs when blast waves interact with structures.

Experimental studies (Hornung, 1986) have shown that there are at least two distinct types of reflection, *regular* [Fig. 11(a)] and *Mach* [Fig. 11(b)]. Regular reflection produces a

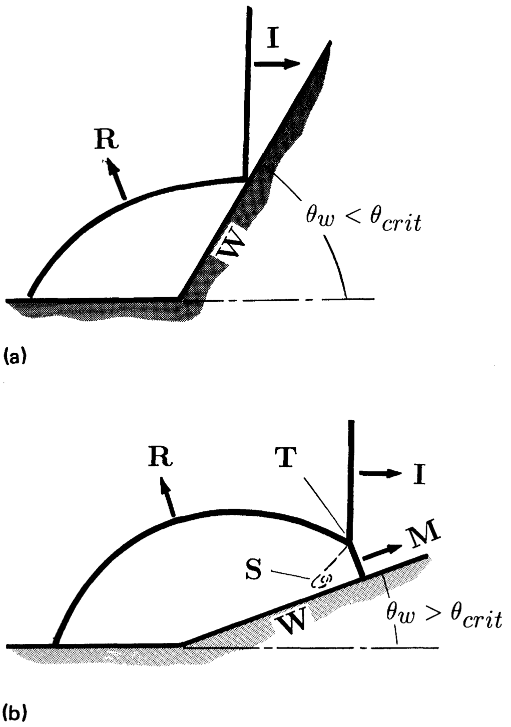

FIG. 11. (a) Regular and (b) Mach reflection of an unsteady shock wave interacting with a wedge.

reflected wave (R) at the point where the incident wave (I) impacts the wedge (W). Mach reflection produces a *Mach stem* (M), a shock wave that is oriented nearly normal to the ramp and intersects the incident (I) and reflected waves (R) in a *triple point* (T), from which a contact surface and associated vorticity [manifested as the shear layer (S)] emerge into the shocked fluid. Mach reflection is observed for wall angle θ_W smaller than a critical value; above that, transition to the regular configuration occurs. The critical wedge angle depends on the shock Mach number. Its determination is apparently complicated (Hornung, 1986) by the displacement effect due to the viscous boundary layer on the ramp surface and the shear layer at the contact surface.

If the viscous effects are ignored, the problem can be viewed as pseudosteady by transforming to a frame attached to the triple point. Oblique shock adiabats (known as *shock polars*) can then be used to construct a simple approximation to the flow field. The critical angle for transition can be determined by applying a causality condition on disturbance originating from the tip of the wedge. Despite the success of the pseudosteady theory and the appealing simplicity of a classification into either Mach or regular reflection, a rich variety of flows are actually observed experimentally. Notable features include secondary and tertiary triple points, vortex structures, and jetlike flows behind the Mach stem. Numerical simulation based on the Euler equations offers insight into the rich variety of these flows. Shown in Fig. 12 is a comparison between observed density contours (isopycnics) in an interferogram image [Fig. 12(a)] and computed isopycnics [Fig. 12(b)] for a Mach 3.51 shock wave in SF_6 incident on a 20° wedge. This is an example of double Mach reflection that clearly shows the "kink" in the reflected wave where the second Mach stem occurs. Other features on the figure include the incidence shock separating undisturbed fluid (0) from shocked fluid (1), a Mach stem between (0) and (3), a reflected wave between regions (2) and (1), and a shear layer between (3) and (2). A vortex is clearly visible behind the curved Mach stem. Letters label the corresponding isopycnics in the experimental and calculated results.

8. FLOWS WITH PHASE CHANGE

A recent development in compressible flows is the exploration of adiabatic phase change induced by rapid compression. While the phenomenon of water condensing from moist air upon isentropic expansion (shown in Fig. 2) is well known, more startling and less widely appreciated is the possibility of liquefying a pure substance by adiabatic compression. Shock-induced liquefaction can occur (Detleff *et al.*, 1979) for fluids with sufficiently large molar heat capacity. Partial and complete liquefaction shocks have been observed in a shock tube when the test (driven) fluid is composed of molecules similar to, or more complex than, perfluoroheptane or octane.

A unique feature of the complete liquefaction shocks is the presence of an extraordinary number of vortex rings in the liquid behind the shock wave. Photographic observations (Thompson *et al.*, 1986) of partial liquefaction shocks demonstrate that the shock wave may be *split* into a precursor shock and a separate condensation wave. These waves

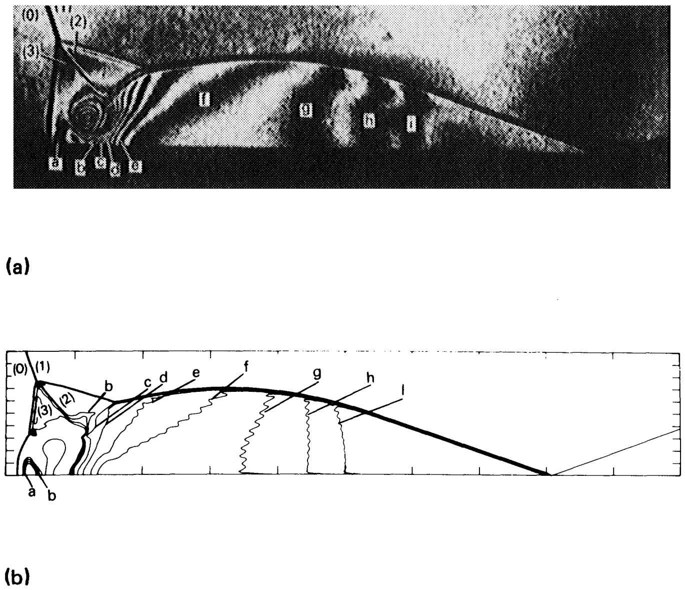

FIG. 12. (a) Experimental and (b) computational results for double Mach reflection of a shock wave in SF_6 (Glaz *et al.*, 1986).

can also be unstable, as manifested by a highly wrinkled wave front. A mild version of this instability is exhibited by the wave shown in Fig. 13. This photograph is an oblique image of a partial liquefaction wave front produced behind a $M=2.63$ shock in isooctane initially at 383 K and 1.1 bar.

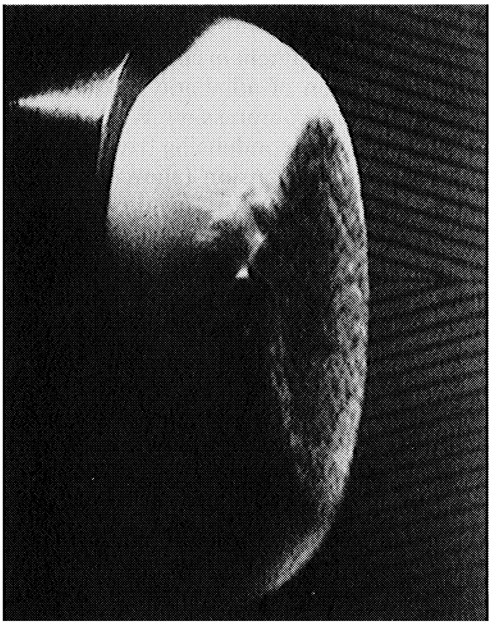

FIG. 13. Diffraction of an unstable liquefaction shock front from a tube (Thompson *et al.*, 1986).

Wave splitting and spontaneous wave-front distortions are two features that may be related to the more general problem (Menikoff and Plohr, 1989) of shock stability. It is known that under certain circumstances, often associated with phase transformations (Duvall and Graham, 1977), shock waves can be unstable to both longitudinal and transverse disturbances. Such instabilities are associated with unstable feedback mechanisms between the mechanics of fluid flow and the thermodynamics of adiabatic compression. Additional instabilities are possible due to the wave structure associated with the finite rate of phase transformation. These instabilities may be similar to those observed in detonations; however, the origin is presently unknown.

Another type of compressible flow with phase change is a two-phase mixture of a liquid with its vapor, typically characterized as a bubbly liquid. Experimental observations (Nakaryakov *et al.*, 1979) of shock waves propagating in such mixtures indicate that many different types of wave structures can exist. The large difference between the vapor and liquid properties combined with the dynamic nature of the phase change results in dispersive waves with complex pressure profiles, including smooth compressions and oscillatory tails, rather than just simple shocks. If the bubbles also contain a combustible mixture, then explosions within the bubbles can result in a self-sustaining wave (Hasegawa and Fujiwara, 1982). Unique wave phenomena (Kutateladze *et al.*, 1987) have also been observed in a pure substance near its thermodynamic critical point. It is possible to obtain very low or even negative values of the fundamental gasdynamic derivative, making possible the demonstration of *rarefaction shock waves*, a possibility hinted at by Eq. (21).

GLOSSARY

Acoustic Impedance: Product of mass density and sound speed ρc, proportionality constant between pressure and velocity changes for acoustic waves.

Adiabatic Compressibility: Proportionality between fractional specific volume changes and pressure changes, $K_s = -(1/v)(\partial v/\partial P)_s$.

Blast Wave: Transient shock wave created by explosion.

Chapman–Jouguet Velocity: Theoretical velocity at which a self-sustaining detonation propagates.

Characteristics: Directions in space-time along which information travels.

Choking: Limitation of mass flow rate due to fluid velocity approaching the sound speed.

Contact Surface: Physical surface in flow separating fluids of different composition, density, or entropy. Pressure and velocity are continuous across contact surfaces.

de Laval Nozzle: Duct containing a converging–diverging section that is used to accelerate the flow to supersonic velocities.

Expansion Wave: Isentropic steady or unsteady flow resulting in a decrease in pressure.

Fundamental Gasdynamic Derivative: Basic measure of the nonlinearity of the fluid, $\Gamma = (c^4/2v^3)(\partial^2 v/\partial P^2)_s$.

Grüneisen Coefficient: Measure of the thermal expansion of a fluid, $G = (1/c_p)K_s(\partial v/\partial T)_P$.

Hugoniot: Relationship between state variables for a shocked fluid, also known as shock adiabat.

Hypersonic Flow: Fluid motion where the velocity greatly exceeds the sound speed, i.e., $M > 5$.

Mach Number: Ratio of fluid velocity to sound speed, $M = u/c$.

Nonlinearization: Process of using nonlinear characteristics together with linear wave equation solutions to obtain approximate solutions to blast wave problems.

Prandtl–Meyer Function: Function used to relate flow deflection angle to velocity changes in steady two-dimensional isentropic flow.

Quasi One-Dimensional: Approximation used to analyze flow in variable-area ducts.

Shock Wave: Adiabatic jump in fluid properties that occurs in a propagating thin layer.

Sonic Flow: Condition in which the Mach number is 1 i.e., the fluid velocity equals the sound speed.

Stagnation State: Thermodynamic state reached when a fluid is isentropically brought to rest.

Substantial Derivative: Derivative following the path of the fluid particles.

Supercritical Airfoil: Airfoil capable of shock-free flight even though the Mach number locally exceeds 1.

Supersonic Flow: Fluid motion with the velocity greater than the sound speed, i.e., $M > 1$.

Thermoelastic Equation: Also known as the Mie–Grüneisen equation, an equation of state used for shock-wave calculations in solid materials.

Transonic Flow: Fluid motion near the speed of sound.

Water Hammer: Shock waves generated by transients in plumbing systems.

Waverider: Shape of vehicle designed by examining streamline solutions for hypersonic flow problems with specified shock profiles.

List of Works Cited

Anderson, J. D., Jr. (1976), *Gasdynamic Lasers: An Introduction*, New York: Academic.

Anderson, J. D., Jr. (1989), *Hypersonic and High-Temperature Gas Dynamics*, New York: McGraw-Hill.

Bird, G. A. (1976), *Molecular Gasdynamics*, Oxford: Clarendon.

Brower, W. B., Jr. (1990), *Theory, Tables and Data for Compressible Flow*, New York: Hemisphere.

Cole, J. D., Cook, L. P. (1986), *Transonic Aerodynamics*, Amsterdam: Elsevier.

Collela, P., Woodward, P. D. (1984), *J. Comp. Phys.* **54**, 174–201.

Davison, L., Graham, R. A. (1979), *Phys. Rep.* **55**, 255–379.

Detleff, G., Thompson, P. A., Meier, G. E. A., Speckmann, H.-D. (1979), *J. Fluid Mech.* **95**, 279–304.

Duvall, G. E., Graham, R. A. (1977), *Rev. Mod. Phys.* **49**, 523–579.

Fickett, W., Davis, W. C. (1979), *Detonation*, Berkeley: University of California Press.

Gibbs, T. R., Popalato, A. (1980), *LASL Explosive Property Data*, Berkeley: University of California Press.

Glaz, H. M., Walter, P. A., Glass, I. I., Hu, T. C. J. (1986), in: J. R. Bowen, J.-C. Leyer, R. I. Solouhkin (Eds.), *Dynamics of Explosions*, Progress in Astronautics and Aeronautics, Vol. 106. New York: AIAA, pp. 359–387.

Hasegawa, T., Fujiwara, T. (1982), *19th Symposium (International) on Combustion*, Pittsburgh: The Combustion Institute, pp. 675–683.

Hayes, W. D., Probstein, R. F. (1966), *Hypersonic Flow Theory*, New York: Academic.

Hornung, H. A. (1986), *Annu. Rev. Fluid Mech.* **28**, 33–58.

Jeanloz, R. (1990), *Annu. Rev. Earth Planetary Sci.* **18**, 357–386.

Kinney, G. F., Graham, K. J. (1985), *Explosive Shocks in Air*, 2nd ed., Berlin: Springer.

Kutateladze, S. S., Nakoryakov, V. E., Borisov, A. A. (1987), *Annu. Rev. Fluid Mech.* **19**, 577–600.

Lee, J. H. S. (1984), *Annu. Rev. Fluid Mech.* **16**, 311–336.

Liepmann, H. W., Laguna, G. A. (1984), *Annu. Rev. Fluid Mech.* **16**, 139–177.

Levine, R. D. Bernstein, R. B. (1987), *Molecular Reaction Dynamics*, Oxford: Oxford University Press, pp. 243–247.

March, S. P. (Ed.) (1980), *LASL Shock Hugoniot Data*, Berkeley: University of California Press.

Menikoff, R., Plohr, B. J. (1989), *Rev. Mod. Phys.* **61**, 75–130.

Moody, F. J. (1990), *Introduction to Unsteady Thermofluid Dynamics*, New York: Wiley.

Nakoryakov, V. E., Pokusaev, B. G., Shreiber, I. R., Pribaturin, N. A., (1988), *Int. J. Multiphase Flow* **14**, 655–677.

Oertel, H. (1966), *Stossrohre*, Berlin: Springer.

Oran, E. S., Boris, J. P. (1987), *Numerical Simulation of Reactive Flow*, Amsterdam: Elsevier.

Park, C. (1990), *Nonequilibrium Hypersonic Aerothermodynamics*, New York: Wiley.

Ritchmeyer, R. D., Morton, K. W. (1967), *Difference Methods for Initial Value Problems* , 2nd ed., New York: Wiley.

Ree, F. H. (1984), *J. Chem. Phys.* **81**, 1251–1263.

Schnerr, G. (1989), *Exp. Fluids* **7**, 145–146.

Schnerr, G., Dohrmann, U. (1990), *AIAA J.* **28**, 1187–1193.

Schwendeman, D. W. (1988), *Proc. R. Soc. London Ser. A* **416**, 179–198.

Sobieczky, H., Seebass, A. R. (1984), *Annu. Rev. Fluid Mech.* **16**, 337–364.

Sod, G. A. (1985), *Numerical Methods in Fluid Dynamics*, London: Cambridge University Press.

Stalker, R. J. (1989), *Annu. Rev. Fluid Mech.* **21**, 37–60.

Stalker, R. J. (1990), in Y. W. Kim (Ed.), *Current Topics in Shock Waves*, New York: AIP.

Taylor, G. I. (1950), *Proc. R. Soc. London Ser. A* **201**, 159–174.

Thompson, P. A. (1971), *Phys. Fluids* **14**, 1843–1849.

Thompson, P. A. (1983), *Phys. Fluids* **26**, 48–49.

Thompson, P. A. (1988), *Compressible-Fluid Dynamics*, Troy, NY: RPI Bookstore.

Thompson, P. A., Sullivan, D. A. (1977), *Phys. Fluids* **20**, 1064–1071.

Thompson, P. A., Carofano, G. C., Kim, Y.-G. (1986), *J. Fluid Mech.* **166**, 57–92.

Vincenti, W. G., Kruger, C. H. (1986), *Introduction to Physical Gasdynamics*, Malabar, FL: Krieger, pp. 281–293.

Westbrook, C. K., Urtiew, P. A. (1982), *19th Symposium (International) on Combustion*, Pittsburgh: The Combustion Institute, pp. 615–623.

Whitham, G. B. (1974), *Linear and Nonlinear Waves*, New York: Wiley.

Zel'dovich, Ya. B., Raizer, Yu. P. (1966), *Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena*, Vols. I and II, New York: Academic.

Further Reading

Anderson, J. D., Jr. (1976), *Gasdynamic Lasers: An Introduction*, New York: Academic.

Anderson, J. D., Jr. (1989), *Hypersonic and High-Temperature Gas Dynamics*, New York: McGraw-Hill.

Anderson, J. D., Jr. (1990), *Modern Compressible Flow*, New York: McGraw-Hill.

Cole, J. D., Cook, L. P. (1986), *Transonic Aerodynamics*, Amsterdam: Elsevier.

Emanuel, G. (1986), *Gasdynamics: Theory and Application*, New York: AIAA.

Fickett, W., Davis, W. C. (1979) *Detonation*, Berkeley: University of California Press.

Liepmann, H. W., Roshko, A. (1957), *Elements of Gasdynamics*, New York: Wiley.

Moody, F. J. (1990), *Introduction to Unsteady Thermofluid Dynamics*, New York: Wiley.

Oertel, H. (1966), *Stossrohre*, Berlin: Springer.

Park, C. (1990), *Nonequilibrium Hypersonic Aerothermodynamics*, New York: Wiley.

Sedov, L. I. (1959), *Similarity and Dimensional Methods in Mechanics*, New York: Academic.

Stanyucovich, K. P. (1960), *Unsteady Motion of Continuous Media*, London: Pergamon.

Sod, G. A. (1985), *Numerical Methods in Fluid Dynamics*, London: Cambridge University Press.

Thompson, P. A. (1988), *Compressible-Fluid Dynamics*, Troy, NY: RPI Bookstore.

Vincenti, W. G., Kruger, C. H. (1986), *Introduction to Physical Gasdynamics*, Malabar, FL: Krieger.

Whitham, G. B. (1974), *Linear and Nonlinear Waves*, New York: Wiley.

Zel'dovich, Ya. B., Raizer, Yu. P. (1966), *Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena*, Vols. I and II, New York: Academic.

COMPUTER-AIDED DESIGN IN ELECTRONICS

ANTUN DOMIC, *Digital Equipment Corporation, Hudson, Massachusetts, U.S.A.*

Introduction 71
1. **History** 72
2. **Computer-Aided Design in Electronic Systems** 73
2.1 Electronic Design of Integrated Circuits 74
2.2 The Design Process 74
3. **High-Level Design** 75
4. **Logic Design** 76
4.1 Circuit Design 76
4.2 Logic and High-Level Synthesis 76
5. **Simulation and Verification** .. 77
6. **Layout Editing and Verification** 79
6.1 Layout Editing 80
6.2 Layout Verification 81
7. **Automatic Layout** 82
7.1 Placement 82
7.2 Routing 83
8. **Higher-Level Design Tools** ... 84
Glossary 85
Works Cited 86
Further Reading 87

INTRODUCTION

The application of computers to the design of electronic systems has steadily increased since its introduction in the late 1950s. During the last two decades, the prices of minicomputers and desktop machines (workstations and personal computers) have been steadily decreasing while their overall performance has been continuously improving. This upward trend in computer performance has been accelerated by rapid advances in microprocessors, which have doubled their execution speeds approximately every 18 months, and faster and larger-capacity semiconductor memories. Increased performance has allowed the development of sophisticated new applications whose functionality, in turn, makes possible the efficient design of more powerful machines. This cycle will continue.

To illustrate the increase in workstation computational power, we cite two sets of numbers: The Motorola 68020 family of CISC microprocessors (complex instruction set computer), widely used in these machines a few years ago, had in 1986 an execution speed of approximately 4 MIPS (million instructions per second). Two machine generations later, by mid-1990, Motorola's 68040 microprocessor had achieved a speed of 20 MIPS. In four years, processing power had grown five times. The progress in chips implementing RISC (reduced instruction set computer) architectures has been even more rapid. The first workstation that implemented SUN Microsystems' SPARC architecture appeared in the second half of 1987 and offered 10 MIPS, while in the first half of 1991 Hewlett-Packard's Precision Architecture machines achieved 70 MIPS. That is a sevenfold increase in approximately four years. Hardware prices have decreased to the point where the difference between the low end of the workstation market and the high-end PCs (personal computers) has disappeared.

The progress in the application of computers to the design of computer systems has been equally impressive. New and better computer-aided design (CAD) tools have been one of the key factors making possible the successful design of extremely complex integrated circuits (ICs or chips). To point out the increase in IC complexity, we offer the following illustration. In 1985, the microprocessor designed by Digital Equipment Corporation for use in its MicroVAX II series machines contained about 125 000 transistors. By 1990, Intel's 80486 microprocessor, used in many personal computers, contained around 1 200 000 transistors. Without important ad-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

vances in computer-aided design, systems of such complexity would be impossible to design.

The field of computer-aided design includes a wide span of applications. We will cover only aspects related to the design of electronic systems, with emphasis on the design of integrated circuits. Given the many areas of interest, we are forced to mention some of them only briefly. In general, CAD refers to specialized software that allows data entry and verification, simulation of the behavior of pieces of a system, automatic synthesis of increasingly detailed descriptions, and a number of data management facilities. Examples include graphical editors to enter logic schematics and chip layouts, programs to verify that layout design guidelines have been respected, systems to simulate the electrical behavior of a set of interconnected transistors, or programs that automatically generate geometries in order to connect or "route" the desired pins among components placed on a printed-circuit board.

As its name indicates, CAD is an aid to design. So it is not a discipline with its own set of problems, even though many CAD problems have models with great theoretical importance. CAD's effectiveness is measured by what it makes possible. Two areas are emphasized: the ability to design very-large-scale integration (VLSI) chips and systems using them, and the contribution to productivity, such as savings in the size of design teams and overall time-to-market schedules. In consequence, developments and topics of interest have changed through time, tracking the leading-edge technologies used in the design of computers and integrated circuits. Nevertheless, one tendency has been consistent. As systems become more complex, the design process moves toward more and more abstract levels of description. This phenomenon has required a corresponding increase in the development and use of automatic synthesis and simulation tools. While advances will continue in most CAD areas, the most visible increases in design productivity will be achieved through synthesis (at logic and layout levels) and innovative uses of simulation and other verification techniques.

1. HISTORY

CAD, as part of the broader field of design automation, started in the mid-1950s. The use of transistors in computer circuits created new problems, and to solve them, designers took advantage of the wider availability of computers. The first CAD paper probably was by Cray and Kisch in 1956 (see Breuer, 1972), describing a system to check and simulate Boolean equations. Another area of early interest was the entry of data to machines that automatically placed wires to connect components on a board, and programs to verify these data. In these early beginnings we immediately recognize topics that have persisted until today: logic simulation, interconnect estimation, and entry and verification of layout-related data.

In the 1960s, the problems of interest were related to board design. The placement and routing of components on a board were important issues, and we can see early work that is still relevant today (Lee, 1961; Hightower, 1969). Simulation of Boolean descriptions (gate-level simulation in today's terminology) and generation of appropriate input values to exercise the logic (test vectors) became an alternative to the building of increasingly complicated hardware prototypes, often called "bread boards," for system verification. The desire to automate steps in the design process—for example, the automatic generation of logic—also received attention quite early (Friedman and Yang, 1968).

Traditionally, the development of CAD tools was confined to internal efforts in computer and semiconductor companies and research carried out at universities. The CAD development groups were closely tied to the design groups that used their tools, and they were considered truly part of the overall system- or chip-development effort. In the 1970s, the appearance of minicomputers that could be adapted to dedicated tasks in a cost-effective way created new opportunities. Independent commercial companies appeared that offered systems, combining hardware and software, for design entry and for automating some design tasks. The first systems were devoted to the editing and routing of boards and, with the growing importance of the IC, to IC layout entry and geometrical design checking. Schematic entry was the next category of systems that attracted these companies.

Important advances in all areas were achieved in the 1970s. A circuit simulator developed at the University of California, SPICE (Nagel, 1975), while having gone through many enhancements, is still in use today. New ways to do IC layout were proposed (Williams, 1978), and significant attempts to shift the design process toward more abstract levels were reported (Johanssen, 1979). The increased transistor density in ICs allowed by the newest metal oxide semiconductor (MOS) technologies brought the introduction of the term VLSI to describe these new chips that could really be considered a fairly complex computer system. The appearance of the book *Introduction to VLSI Systems* by Mead and Conway (Mead and Conway, 1980) significantly turned the perspective in IC design from a pure "device" to a "system" orientation. Clearly, this approach emphasized the importance of automation. In the United States, the funding provided by the Defense Advanced Research Projects Agency (DARPA) to support research through its VLSI projects had a major impact on CAD development through the mid-1980s. Technology advances, such as the emergence of denser MOS gate arrays, shifted the focus even more towards CAD for ICs.

In the early 1980s, the success of the microprocessor provoked a shift from minicomputers to cheaper workstations. In the commercial CAD world, companies that were not able to port software from their proprietary, minicomputer-based hardware to these new machines ended up disappearing from the market, replaced by new players that competed on the basis only of their software, with UNIX as the dominant operating system. As the entry barriers for developing software are much lower than those for hardware, a large number of companies were formed, most of them offering narrowly focused programs (layout verification, simulation, schematic editors, etc.). But as the decade progressed, two phenomena occurred. Internal CAD groups at computer and semiconductor design houses started to rely increasingly on CAD companies for important pieces of their software, and the CAD companies went through a process of consolidation that continues until today (1992). The survivors of this consolidation process, such as Mentor Graphics (Beaverton, OR) and Cadence (San Jose, CA), are large, publicly traded companies with yearly sales on the order of hundreds of millions of dollars. These companies have concentrated their offerings on fairly wide, integrated systems for design instead of isolated point tools.

Today, CAD is advancing along three different lines. Universities continue to play an important role in new research. The University of California at Berkeley, Stanford University in Palo Alto (both in CA), and Carnegie Mellon University in Pittsburgh, PA, have been leading centers in the United States, while the Universities of Edinburgh in Scotland, Grenoble in France, Louvain in Belgium, and Eindhoven in The Netherlands have been important European contributors. In addition, research centers associated with companies, such as IBM's Thomas J. Watson Research Center at Yorktown Heights, NY, and AT&T's Bell Laboratories in Holmdel, NJ, have played a key role. Commercial CAD companies and internal CAD groups have traditionally functioned using different models. Internal CAD groups have focused more on specialized applications that serve their companies's technologies, particularly in areas that today are leading edge but soon will be more generic, while CAD companies have been generally interested in providing systems with the widest possible appeal. Currently, however, most computer and semiconductor companies use a combination of commercial and in-house tools. The evolution of this mix is not clear at this time, but it will determine much of the CAD development in the coming years.

2. COMPUTER-AIDED DESIGN IN ELECTRONIC SYSTEMS

We will consider only certain aspects in the design of electronic systems, in particular, digital computer systems. We will include CAD tools for the design of a system as a logical unit, including component ICs, for the simulation of the system, and for the design of boards. We will exclude the semiconductor process design and manufacturing, thermal analysis, testing and reliability, cabinet design, specialized software, and other topics that, while significant in the design of a computer system, fall outside our description of a "logical unit."

As already indicated, the focus of CAD has changed through time. While most of the efforts in the 1960s were dedicated to board-level design, the phenomenal increase in the complexity of integrated circuits in the 1980s has shifted its focus. In 1990, a state-of-the-art microprocessor contained between one and two million transistors. It must be noted that, while much more complex than other parts, the microprocessor is only one component in a system. Many CAD tools are equally applicable to boards and IC design, even though the emphasis can be different. Examples are simulation tools, or placement and routing systems that use similar algorithms. Our descriptions will mainly address the IC case, although we will mention differences and applicability to systems.

2.1 Electronic Design of Integrated Circuits

There is a large variety of integrated circuits, using different semiconductor technologies and having equally different kinds of circuit functionality. The first important distinction among the types of ICs is between digital and analog chips. As these names indicate, the main difference is whether the signals can assume a continuous range or are restricted to two discrete values (high and low, or 0 and 1) for basic design purposes. A microprocessor is a mostly digital chip, while a filter can be purely analog. A second distinction is made according to the semiconductor technology used. The most common technologies today are CMOS (for complementary metal oxide semiconductor) and bipolar technologies such as ECL (emitter coupled logic). Lastly, a third distinguishing characteristic is the IC's design style. The most important categories of design styles are custom, semicustom, and gate arrays. Special chips such as memories do not fit any of these types, and traditionally their advances have been made possible through progress in semiconductor-processing technology rather than through CAD. Many developments in IC CAD tools, particularly related to synthesis versus entry and simulation, have been motivated by applications to semicustom and gate-array MOS chips.

The distinctions among these design styles are based on the degrees of freedom we assume. In the case of custom design, maximum flexibility is assumed: Essentially each transistor can be individually tailored. While it achieves the best use of a technology, it levies a large design-effort cost. So it is prevalent in high-speed microprocessors and analog parts. In the case of semicustom design, there is extensive use of a predesigned library of functions, referred to as standard cells, with each usually consisting of a few logic gates. Larger units, called macros, are also employed as single entities. This approach causes a number of inefficiencies, but it allows much more automation. Frequently, custom and semicustom approaches are combined. Finally, in the case of gate arrays, the transistors are laid out in a fixed, regular fashion, and a library of basic gates is constructed, very much as in the case of a semicustom chip but usually with lower performance. A key difference is that in a gate array, we "design" only the interconnection layers, bringing an important economic advantage as only a few layers are customized. Technology advances have increased significantly the capacity of the CMOS sea-of-gates types of arrays, due to the use of more interconnect layers (usually three metal layers), resulting in a higher percentage of utilization of the available gates, and reduced transistor sizes, increasing the density in the chip. Today, the largest gate-array chips contain hundreds of thousands of logic gates and are offered by many semiconductor vendors. Gate arrays constitute the largest component of the market for special-purpose chips (application-specific ICs or ASICs).

2.2 The Design Process

We give now a brief overview of the design process, with the intent of relating design steps to the different CAD tools. The steps involved vary in importance among the types of systems and chips, but they apply quite generically to most computer systems. The design process is a sequence of transformations, each of which produces a description that progressively refines the previous one by adding further detail. This process is usually referred to as "top-down" design.

First, certain technology choices are made and translated into design parameters. For example, a system intended for a certain market will have specific requirements for speed, power consumption, and overall capabilities. A suitable technology is selected to meet these requirements, and the system's basic characteristics are established. In the case of a microprocessor, basic problems include establishing an acceptable clock rate, determining the amount of memory that can go inside the chip, calculating the total power dissipation, and other global factors.

The design proceeds in stages, with each adding more detail in terms of the actual system implementation. Usually, a high-level or "behavior" model is constructed and extensively simulated. From here, logic and circuit design is done to transform the high-level model into Boolean functions and their circuit implementations. In the case of chips, the circuits are laid out, that is, the different mask layers required by manufacturing are drawn. On the other hand, for boards, the positions of the components and the wires connecting them are determined. At each step, extensive verification tasks are performed to establish compliance to technology rules and to establish the equivalence of the different descriptions. Now, we will proceed to describe in greater detail these major steps and the CAD tools employed in the process.

3. HIGH-LEVEL DESIGN

The logic design process can be divided into two major phases. First, a high-level description of the system or IC is usually created using a special language referred to as hardware description language (HDL). Depending on its level of detail, this model is called a behavior or an RTL (register transfer level) model. Second, this high-level or functional model is converted, through manual or automated steps, into a logic description. This description includes the precise components being used at a level of logical units such as Boolean gates, ALUs (arithmetic logic units), etc., and, in the case of boards, components such as ICs that are treated only from the point of view of their input/output behavior.

The functional model indicates what the system does, but does not indicate exactly how it does it. The difference between a pure "behavior" and an "RTL" model is the level of detail, and it is not clear cut. A way to indicate the distinction is through an example: We may want to add two numbers, and we write such a statement in a behavior model, while in the RTL version we indicate exactly which arithmetic unit and registers are actually involved in performing the operation. A functional model resembles a software program. It is written in an HDL that includes a simulation environment so that the system's behavior under specific input patterns can be observed. Architectural decisions and their impact on system performance are usually evaluated by use of a behavior model and extensive simulation.

Many specialized languages, such as DPL or DECSIM BDS (see De Micheli *et al.*, 1987) have been used as HDL. Until the early 1980s, most of the HDLs used by industry were developed by computer or semiconductor companies. In some cases, general software programming languages (such as C) have been used directly, even though only a restricted subset of the language's constructs is allowed. The key tradeoff in the design of an HDL and its simulation environment has been between flexibility and execution speed. Implementations based on interpreters allow excellent observability and user interaction but at a serious cost in simulation speed.

An effort to create an HDL that could be widely used was initiated by the U. S. Department of Defense as part of their VHSIC program (very-high-speed integrated circuits). This resulted in the definition of VHDL (VHSIC HDL), now an IEEE standard (Dewey and Gadient, 1986). VHDL has gained increased acceptance as a natural candidate for standardization. Commercial CAD companies had been offering their own languages and simulators for years, contributing to the creation of a large installed base of models that are exchanged among the system or chip users. Cadence's Verilog HDL, for instance, has played a significant role due to its large market share (Cadence Design Systems, San Jose, CA).

The functional description has taken another increasingly important role as input to the logic synthesis tools now being used in IC design. The logic synthesis process consists of the transformation of a functional description into a structural representation, usually at the gate level, including steps to minimize the logic and to meet timing constraints. This has caused reexamination of the HDL, as the exact interpretation of its statements in terms of Boolean equations has become very important.

4. LOGIC DESIGN

The creation of a representation of the design using only units whose functionality corresponds essentially to Boolean equations and storage elements is called logic design. The most usual description is a set of schematic diagrams involving simple logic gates (such as multiple-input AND, OR, NOT, etc.) and registers. In the case of systems, these functions are divided into components. The representation of a design in these terms is also called a structural description. Many systems, particularly before the advent of VLSI, were designed with schematics, not functional descriptions, as their starting point. Only as complexity increased has the use of functional models become the norm.

Specialized editors to enter gate- and transistor-level descriptions are called schematic editors. Output from the schematic editors are lists of components or cells and their interconnections (netlists), including several levels of detail ranging from large library items (such as an ALU) to transistor-level descriptions, and plots. The netlists are used by simulators to investigate the behavior of the circuits. Special software and computers for schematic editing followed a development path similar to layout editors. This will be described in a later section. Today, software systems for editing on workstations and PCs are offered by many companies, including component libraries for the major gate-array vendors.

4.1 Circuit Design

Circuit design consists of determining the exact implementations of the logic functions and storage elements in terms of the primitive elements of the specific technologies. In the case of ICs, these primitive elements are transistors and other basic elements, including their electrical parameters. Each technology has its own specific circuit techniques. For the case of CMOS the reader can consult Glasser and Dobberpuhl (1986). A design comprises a library of items, each implementing a logic function and parameters such as delays, drive, and power consumption; a method for distributing special signals such as clocks; and general usage rules. In the case of a custom design, there is much more flexibility in the creation of circuits. For example, some tools have been developed to assist the circuit design process, such as CMOS transistor sizers (Hill *et al.* 1989), but it still requires a large amount of trial and detailed simulation.

4.2 Logic and High-Level Synthesis

As already mentioned, the use of logic synthesis has become a part of the design process. Even though work in aspects of logic synthesis started decades ago, it was during the 1980s that the extraordinary progress achieved in this area changed the logic design process. Most of the work has been centered on the generation of the combinational logic (excluding the storage elements that convert the design into a sequential type). The process is divided into two steps: a technology-independent step, where Boolean logic is minimized, followed by a technology-mapping phase where the logic is implemented using a set of predefined gates dependent on specific technology choices.

A Boolean function of N variables can be represented by three subsets of the N-dimensional Boolean cube (the set product of $\{0,1\}$ with itself N times): the points where its value is 1, or ON set, the points where its value is 0, or OFF set, and optionally a "don't care" set, that is, those input value combinations for which the function's output value is unimportant. This case occurs typically when we know that a specific combination of input values will never happen. Minimizing the function is closely related to the problem of "covering" the ON set optimally by subsets of the N-dimensional cube that are themselves lower-dimensional cubes (or "faces") with particular properties. Techniques to obtain a minimal realization of a Boolean function using two levels of logic date back to the 1950s (McCluskey, 1956). One subproblem encountered in this process (finding the collection of optimal special subsets) is "*NP* complete." In loose terms, *NP*-complete problems cannot be solved by any algorithm the number of whose steps is bounded by a polynomial function of the problem variables (Garey and Johnson, 1979), making it necessary from a computational point of view to develop reasonable heuristics. When MOS technology started to predominate in VLSI chips, interest in two-level representations was renewed, as it was possible to produce efficient layouts for a structure called PLA (programmable logic array) to implement two-level logic functions. A detailed description of a widely used program, ESPRESSO-II, and problems found in the logic minimization of PLAs is given in Brayton *et al.* (1984).

Logic synthesis for the general case of a multilevel network, including sequential elements, has been done with a variety of techniques. Systems such as IBM's LSS (Darringer *et al.*, 1984) have combined "local" and "global" transformations to modify the logic until the desired constraints are met. Local transformations examine only a small portion of the circuit, an example being the application of equivalences between AND-NOT and NOT-OR combinations (DeMorgan laws; see Dietmeyer, 1978). Global ones take a larger view, such as following input–output signal flow to identify particular relations among variable values, or collapsing a selected portion of the circuit into two levels to apply special minimization. Several objectives are considered, such as area minimization, timing optimization, and testability. General Electric's SOCRATES (de Geus and Cohen, 1985) employed a rule-based approach to the technology-dependent phase following Boolean minimization. A completely algorithmic approach is followed in MISII (Brayton *et al.*, 1987), applied to strictly combinational circuits. The program executes a user-controlled sequence of optimizations that include tradeoffs between area and time. The technology-mapping phase is based on a technique introduced in DAGON (Keutzer, 1987), where the predefined logic gates and the network are decomposed into tree graphs and matched using local analysis.

A number of new areas closely related to logic synthesis are being actively researched. Because of space constraints, we can only mention them briefly: the relation between the testability of a circuit and its minimal Boolean implementation, the problem of encoding a finite-state machine (see, for example, Dietmeyer, 1978), improved timing optimization algorithms, and the possibility of looking at the full sequential circuit and taking advantage of operations that cross the register boundaries. The interested reader is referred to the CAD conferences cited for further reading.

A different type of logic synthesis is usually referred to as high-level synthesis. The intention is to find a structural implementation for a given behavior and set of constraints, but the basic elements from which we are selecting have a higher level of complexity than logic gates—typically registers, ALUs, multiplexers, and buses. The intended output is a data-path structure and the specification of its control.

Two important problems in high-level synthesis are scheduling and allocation. In scheduling, we determine which operations take place at each control step under a set of restrictions. The restrictions can include, for instance, the number of hardware units available, the total delay that operations can take, or the estimated power consumption. Many algorithms have been investigated, but most fall into two categories. These are exhaustive search combined with some heuristics or standard mathematical techniques such as linear programming (Brayton *et al.*, 1985; Hafer and Parker, 1983), and constructive approaches that proceed by building the operations incrementally (Marwedel, 1986). During allocation, we choose the units in the data path and explicitly assign the operations. Constraints are total area, number of units, and bus interconnects, among others. Techniques used for allocation can be separated into constructive and global ones. In the latter case, graph models have been employed (Hafer and Parker, 1983). Given the generality of the problems, rule-based expert systems have also been applied to high-level synthesis (Thomas *et al.*, 1988).

5. SIMULATION AND VERIFICATION

The need to verify that a system description actually performs as intended under sequences of input values is obvious. Extensive simulations are performed at all levels of system description to ensure two key objectives: that the descriptions match the design intention, and that the different descriptions are actually equivalent. The first of these objectives is usually called the validation problem, even though, unless we are able to do exhaustive analysis, we cannot totally validate a design. The second, or verification, is used in a looser sense to describe approaches that are used to compare different descriptions and obtain a reasonable degree of confidence that the design performs as intended. While detailed electrical simulations—for instance, at the transistor level—would give us a larger degree of assurance that a complex chip will operate properly, the time it would take a computer to perform such a simulation makes this approach completely unfeasible. Thus simulations are performed with models that have very different levels of accuracy in order to cover a large chip or system "adequately."

The most general description available is an HDL program. This representation is usually simulated at the functional or RTL level. The simulation parallels closely the execution of a general software program, and the data that we are observing can be quite general (such as an integer, for instance). The advantages of functional simulation include high execution speed and the ability to perform extensive simulations before the design of the actual circuits has progressed sufficiently to be useful.

A more detailed description is obtained at the logic level. In this case, the description consists of logic gates (ANDs, ORs, etc.) and the values we observe are 0 and 1, plus some "undefined" states, producing what is referred to as gate-level simulation. The concept of time can be now introduced in a much more precise way. Relations are established in discrete units among the different clocks, the behavior of the different register elements is established, and delays through each gate type are indicated. In addition, delays due to interconnects can be modeled. These are usually based on extracted wire lengths or total capacitance including parasitic effects such as couplings. Gate-level simulation has been the most prevalent way to establish a design's correctness, at least from a functional perspective that includes simple timing models, even though in most cases no proof of correctness is achieved. Gate-level descriptions are often combined with RTL sections, creating a "mixed-mode" simulation. An example of such a system, used widely by a large computer company, is found in Insinga (1984). For topics closely related to gate-level simulation, such as test-pattern generation and fault simulation, the interested reader is referred to Abramovici *et al.*(1990).

Many techniques have been used to execute gate-level simulations. We mention the two most basic: compiler and event driven (see Saleh and Newton, 1990). In a compiler-driven simulator, the elements in the circuit are ordered according to the time at which they may execute a transition, usually according to their logic level. Code is then generated to evaluate all gates sequentially. In an event-driven simulation, we try to take advantage of the fact that the ratio between the total number of nodes and the nodes that actually change values is normally below 10% so that it is not necessary to simulate all nodes at each time step. On the basis of the inputs that have changed values, we schedule evaluations only for those nodes that may be affected. As a way to accelerate gate-level simulations, special-purpose hardware accelerators have been built by computer companies for their internal use, and some small companies have offered these machines commercially. Two examples can be found in Denneu (1982) and Agrawal *et al.* (1987). Many attempts to use parallel processing (see PARALLEL COMPUTING) have also been described. The possibility of building systems based on field-programmable gate arrays (FPGA) to do simulation (versus the actual implementation of the chip or system) is being actively researched.

In the case of MOS circuits, the use of switch-level simulators has been quite successful as a way to combine a simulation in terms of obtaining digital information (high and low values at each node) with the inclusion of electrical and timing information. The idea is to model the transistors as fixed-value resistors and capacitors (*RC* model), controlled by a switch (the value on the transistor's gate), and wires as ideal conductors. This simplification allows the use of an efficient approximate model to calculate delays (White and Sangiovanni-Vincentelli, 1987). In addition to high and low values, some electrical notions can be introduced. For instance, based on the transistor's width and length, a concept of "strength" can be introduced to model behavior in a feedback loop. A problem has been the fact that in many common MOS circuits, a transistor can apparently act as a bidirectional device. While this actually may not happen during the circuit operation, to determine automatically the direction of signal propagation is difficult. Efforts to improve the actual processing of the transistor interconnection list and to transform it into a set of Boolean equations has produced important speed improvements in simulation times (Bryant *et al.*, 1987).

In addition to the use of the switch-level model for simulation, it has become common to use it for chip-level timing verification of MOS circuits. The problem here is different: to determine paths that do not meet a timing constraint—for example, whether a signal is arriving at the input of a storage element in time with respect to the clock controlling the element. The analysis now is not based on actual input patterns but on looking at worst-case delays using the *RC* model. In addition to the directionality problem, another concern is the appearance of "false paths," that is, paths that will never actually be logically exercised. An early example of such a system was RSIM (Terman, 1983). The applicability of these techniques to the processing of very large transistor networks has continued to make them extremely attractive (Pan *et al.*, 1991). Many attempts have been made to speed up the analysis of these *RC* networks. Very large networks can be obtained by detailed layout extraction to include the effects of actual interconnect delays. A recent technique called asymptotic wave-form estimation (Pillage and Rohrer, 1990) has been successfully applied to networks with hundreds of thousands of elements.

The most detailed simulations are done at the device level. In this case, very accurate models of transistors, transmission lines, resistors, and other basic circuit elements are used. Analysis of such aspects as transients, steady state, or sensitivity can be performed with great accuracy but involve significantly increased simulation times. In the case of ICs, the number of transistors that can be simulated at this level in a reasonable amount of time rarely exceeds a few thousand. Probably the most widely used circuit simulator is SPICE and its many successive versions from the University of California and commercial CAD companies (Nagel, 1975).

The main components in a circuit simulator are as follows: The complex elements, such as transistors, are broken into simpler components, usually two-terminal linear and nonlinear resistors, capacitors, inductances, and controlled sources, using predefined models (for the case of SPICE MOS devices, see Vladimirescu and Liu, 1980). By use of Kirchhoff's current and voltage laws, systems of differential equations can be written and solved by numerical integration methods. The numerical techniques used account for the large amount of computer resources needed by circuit simulators, the difficulties in solving very large circuits, and problems related to convergence and stability of the solutions. For a more detailed discussion we refer the reader to White and Sangiovanni-Vincentelli (1987) and Saleh and Newton (1990). Also, for a deeper level of analysis of the detailed simulation of devices such as transistors as two- and three-dimensional structures, the interested reader can consult Sze (1984).

A large amount of simulation is done just to ensure that small logic changes have not altered the circuit's behavior. Recently, new techniques have been applied to prove formally the equivalence of two Boolean networks, avoiding multiple iterations through a simulation loop. The general problem of deciding if a Boolean expression is a tautology, that is, is true for all possible input values, is *NP* complete (Garey and Johnson, 1979). But special representations of a function that allow fast comparisons using reasonable amounts of memory for important subsets of functions, particularly those generated by logic synthesis tools, have been introduced (Bryant, 1986; Brace *et al.*, 1990). These functions are given canonical representations as special acyclic graphs called binary decision diagrams. This representation has allowed the verification of equivalence inside logic synthesis programs such as MISII. Combined with the ability to abstract a transistor-level netlist into a Boolean representation, the technique offers new possibilities for the functional verification of MOS circuits.

6. LAYOUT EDITING AND VERIFICATION

An important piece of the design process is the representation of the detailed data describing the exact geometries for the actual masks used to manufacture ICs, or the metal areas and holes for a board. Before the appearance of electronic editors, these large amounts of data were manually transferred (or "digitized") from drawings into computers, so that they could be further processed and used in manufacturing. The basic problem in the design of the first electronic editors was the actual hardware, particularly the graphic display systems. These displays were connected to general-purpose computers or were sold as part of a complete system that included the editing software and a modified minicomputer. Plotting hardware followed a similar path. In the 1970s and early 1980s, several companies offered these very successful systems, running on proprietary hardware closely tied to their specific programs. Then in the 1980s, the appearance of the workstation changed this picture radically. Layout editors could operate on these smaller computers that, in addition, were used for a multiplicity of other tasks. A number of companies, unable to port their software quickly to the new computers, simply disappeared from the marketplace and were replaced by new, software-oriented players.

6.1 Layout Editing

The representation of layout data has two aspects. First, we need standard exchange formats in order to transfer data. Second, each editor needs efficient internal data representations in order to manipulate data in an interactive fashion. Some standard exchange formats, such as CIF (Caltech intermediate form) and GDSII (originally from Calma Company), widely used for IC layout data, Gerber (property of Gerber Scientific Instrument), and the more general EDIF (electronic design interchange format) are described in Rubin's book (Rubin, 1987). As the amount of data increased significantly, dealing with flat polygon representations became difficult. The use of replication and hierarchy became common in IC layout, reducing complexity and memory usage. New data structures, able to include relations among polygons, began to be used. An editor widely used in academia, at least in the United States, is MAGIC (Osterhout *et al.*, 1984), where the concept of links among some neighboring polygons (called "corner stitching") was exploited to perform operations beyond the usual polygon editing such as geometrical design rule checking. Other two-dimensional data organizations, such as dividing the available area into quadrants and storing the data in related trees ("quad-trees"; see Rosenberg, 1985), have been used to allow quick searches.

Entering polygons to represent an integrated circuit is an extremely time-consuming and error-prone task. While reliable data are difficult to obtain, the problem being similar to the question "How many lines of code does a programmer write per day?," the figure is generally a few dozen unique transistors per day. Representing the masks at an abstract level could simplify the process and increase productivity significantly. The concept of symbolic layout (Williams, 1978) was introduced in the context of MOS IC design, even though the idea of representing basic objects as symbols is older. Connections are represented by dimensionless wires along centerlines on the correct process layers, while primitive elements are treated as objects with design parameters. For instance, a MOS transistor would be described in terms of its type, channel width, and length. The editing functions are simplified greatly as the designer does not need to enter the actual polygons and is not concerned with the exact spacings but only with the topology of the layout. A program acts on the symbolic description and expands it into the actual polygons with required spacings based on the technology rules. Unfortunately, the density of the layout obtained after this simple expansion is generally very poor. It has been necessary to alter the relative position of transistors and interconnect in order to reduce the total layout area. This step is called compaction. A compactor program creates a mathematical model to represent the many geometrical constraints (such as spacings, avoiding the intersection of some layers, or keeping some elements fixed) among the large number of basic elements (transistors and connection polygons) and uses an optimization technique to minimize the total layout area subject to the derived constraints. Many algorithms for performing this task continue to be proposed, starting from the first, one-dimensional graph models, to flexible probabilistic ones (Boyer, 1988). In spite of the progress achieved, the use of this technique to design large circuits has not been widespread. The best successes reported have been in the use of symbolic layout to design at the cell level (circuits up to a few dozen transistors), particularly because of the simplicity in transporting the cell layouts across similar semiconductor processes.

Another way of describing layout data is by creating a procedural description that, when executed as a program, generates the actual polygons (Johanssen, 1979; Batali and Hartheimer, 1980). The layout is then represented by text. Typical basic elements, such as a transistor or a contact, could be instantiated by calls of the parametrized procedures. While there is some loss of design freedom as we cannot operate directly on the polygons, the layout representation could take much less memory and storage than the full polygon expansion. Another advantage is the possibility of using more abstract geometrical concepts to express relative positions, such as "to_the_right_of." A useful abstraction introduced in the late 1970s was the simplification of the geometrical design rules so that they could all be described as multiples of a basic unit, resulting in the so-called "lambda style rules" (Mead and Conway, 1980). The use of procedural layout languages and simplified design rules has been especially fruitful in the design of special array-type compilers, such as memory or data-path generators. They form a very important part of today's layout systems, including those offered by commercial CAD companies.

6.2 Layout Verification

The drawings describing the mask layers which form the integrated circuit contain millions of polygons. Whether these are generated by hand or by software programs, errors can be introduced making it necessary to perform exhaustive analysis on the data. One looks for three types of problems: compliance to the geometrical design rules, equivalence with the schematic data, and appearance of electrical effects detrimental to the chip performance.

Each semiconductor process has a set of geometrical design rules describing its layers, the minimum dimensions acceptable for each layer, the minimum separation among polygons, and the interactions between the different layers. These interactions allow, for example, the overlapping of a number of layers to form transistors or the prohibition of intersection of certain layers to prevent connections. A geometrical design rule checking (DRC) program verifies the polygon database measuring polygon sizes, distances, and intersections and indicates the places where a rule has been violated. Usually, the violations are superimposed on the layout in such a way that they can be displayed in the layout editor. Some editors, such as MAGIC (Osterhout *et al.*, 1984), allow running a DRC program interactively as the layout is being done.

The second problem that needs to be resolved is to ensure the equivalence of the devices formed by the masks with the database generated from the schematics (or some other data). Again, the process has rules that describe how electrical elements are formed and how the geometrical dimensions are translated into electrical parameters. A typical case in MOS technologies is how a transistor is formed by intersecting some rectangles in specific layers, and how the resulting rectangle's dimensions relate to the transistor's basic parameters, its channel width and length. So, we are considering two related problems: conversion of the geometrical database into devices and interconnections, and proof that the extracted device database is actually equivalent to another of the same basic type. A CAD program performing the first step is called a layout or node extractor, while the checking one is a wirelist comparator. Unfortunately, another set of circuit elements is created by the layout. For instance, rectangles overlapping in different layers, such as the metal layers used for interconnect, will create a capacitive coupling between the corresponding nets. Each technology has a number of effects introducing unintended circuit elements called parasitics. The parasitics can alter the circuit performance or, worse, the actual function that the circuit was designed to implement. The layout extractors usually extract the parasitics of importance to the design, and these are examined according to design restrictions and passed to simulators.

The approaches to rule checking and extraction have been similar. We can operate on the polygon database through a set of basic operations, such as intersections, unions, differences, and expansion, and express the extraction and spacing rules in terms of these operations. Another approach has been to rasterize the polygon data and "slide" a window recognizing specific patterns, taking advantage of the fact that the rules can be localized. In all cases, use of sorting techniques and special representations has been common. Since many rules are independent and apply on a layer-by-layer basis, the checking process lends itself well to the use of multiprocessors. A much more difficult issue has been the checking of three-dimensional effects, such as conductor couplings, where recognition of particular geometry patterns and scaling are used to avoid solving complex differential equations.

7. AUTOMATIC LAYOUT

Efforts to automate the layout tasks (as opposed to simplifying the entering and checking of data) by use of special algorithms started in the early days of CAD. Traditionally, layout tasks have been divided into placement and routing. Placement consists of finding the exact locations that the basic components or cells (small collections of gates that are treated as a single unit) will occupy on a chip or a board. Routing consists of determining the precise paths, including layers, that nets will follow to connect the basic elements' pins. Since many placement and routing problems fall into the class of *NP*-complete problems, it became necessary to develop heuristic algorithms to obtain solutions in reasonable computer times.

7.1 Placement

A key difficulty in placement problems is the huge search space. If we try to place N elements into N slots, there are $N!$ possible configurations. Today, a block in a chip can easily contain 10 000 cells, making it impossible to do anything remotely close to an exhaustive search. Algorithms to place components on a board started to appear in the 1960s. The basic ideas in the early approaches were first to cluster elements according to some specific criteria, such as having a large number of nets in common, and then to perform element exchanges to improve the configuration on the basis of some measure of its "goodness" or cost. The selection of which components to exchange was based on connectivity considerations. The cost function tried to reflect the congestion of the nets in the areas available for routing and some easily computable approximation to the actual net lengths that would result after routing. The simplest approximation to net length used is the half perimeter of a rectangle circumscribing all pins, followed in complexity by some special trees. The expectations were that routes would be short, and so the same tracks could be used for several nets achieving full routing within the available space. Examples of these approaches are described in Breuer's book (Breuer, 1972). One key deficiency in these initial approaches is that they could only drive the placement into a local minimum of the cost function.

A more elaborate model involved considering a placement as a graph, with the components as vertices and the nets as edges, and application of graph-partitioning techniques. In the early 1970s, the "min-cut" approach was introduced (Kernighan and Lin, 1970). As its name indicates, we try to exchange components so that if we "cut" the graph into two, the number of nets crossing between the two sides is minimized. In the process, many iterations are performed, and we are allowed to go through configurations that are not always better than the previous ones, allowing an escape from local minima. There have been multiple refinements of this basic idea. These include the use of elaborate net-weighting schemes to give preferences to some nets over others, the design of special software data structures to simplify the algorithmic complexity of the procedure (Fidduccia and Mattheyses, 1982), and the introduction of fictitious pins in a recursive approach to divide the available area (Dunlop and Kernighan, 1985).

In the early 1980s, a probabilistic approach to solving combinatorial problems was introduced (Kirkpatrick *et al.*, 1983). The overall process was analogous to the melting of a solid at high temperature and then slowly cooling it to form a crystal (annealing to a state of minimum energy), thus the name simulated annealing. The method had a number of important advantages: the cost function could be quite general, the method allowed escaping from local minima, and the process could be stopped at any desired point. More general schemes following the annealing basic principles were called probabilistic hill-climbing algorithms. The algorithms required determining three basic operations: the generation of new configurations, the modification of the algorithm's control parameters, and the configuration acceptance or rejection criteria. Formal proofs of convergence with probability equal to 1 to a global minimum were given under fairly general assumptions on the specific problem, based on Markov chain theory (Romeo and Sangiovanni-Vincentelli, 1985). In the case of simulated annealing, the control parameter is called temperature. Initially, the temperature was decreased by a constant factor in a geometrical fashion (new_temperature = old_temperature×ratio, where $0 < \text{ratio} < 1$), but a number of adaptive schemes have been introduced. The acceptance criterion is based on the comparison of an exponential function of the cost difference between the current and newly generated configuration, exp(−cost_difference/temperature), with a randomly generated value in the range of 0 to 1. Many problem-specific stopping criteria have been described.

Simulated annealing has been applied with great success to the placement of row-style, standard cell blocks in integrated circuits. A large number of refinements, particularly related to techniques to cut down the computer run times of the algorithm and to more general placement problems, have been introduced. Timberwolf, a system originally developed in the University of California at Berkeley and subsequently at Yale University, deserves particular mention due to its industrial acceptance (Sechen, 1988).

In general, the basic measure used in placement algorithms has been a combination of total area after routing and average interconnect net length. Recently, interest has shifted toward "performance-driven" placement, meaning taking into account the effects of the placement on the overall performance of the circuit in a more direct fashion. For instance, nets can have very different net lengths (roughly proportional to delay) and still meet the timing requirements because of their specific path constraints (Benkoski and Strojwas, 1991). While minimal area and short connections are certainly related to these objectives, the emphasis and resolution have changed. Closer ties to the timing verification systems, and possibly changes in the circuit parameters, may be required to meet all constraints automatically.

7.2 Routing

Routing is the process of determining the actual paths assigned to the interconnections among the pins of the placed components. Each technology allows a certain number of conducting layers to carry nets and of "vias" or "contacts" to switch among the layers. Each layer has its own electrical and geometrical characteristics, such as resistivity per square, fixed or variable widths, and many others. The vias also vary widely; in some cases we can jump directly across several layers, in others only between neighboring ones. To illustrate the ranges, we give two examples. First, until recently, CMOS technologies only allowed the use of up to two metal layers for routing, while boards had dozens of layers available. Second, other typical constraints are that all paths must fit inside a fixed area, that paths must never get closer than a given minimum spacing, or that certain layers are assigned a fixed direction (north–south or east–west). A router is a program that assigns paths to the nets so that paths corresponding to different nets never overlap on the same layers.

A common situation is that only two directions are allowed, so that "wires" can turn 90°, resulting in what is referred to as "Manhattan geometry." If we add to this assumption a fixed width and minimum spacing per layer, then each layer can be modeled as a grid. The first algorithm reported to find paths, still of interest today, is the Moore–Lee algorithm (Lee, 1961), based on graph-theory shortest-path algorithms. The basic idea is to expand a net, starting at its pins, to the neighboring grid points successively until all required connections are achieved. Routers based on this technique are called maze routers. While very flexible, their results depend on the order in which the nets are routed, and their implementation can consume large amounts of memory since we need to store a matrix representing the grid points. A new algorithm was proposed on the basis of a technique called line search (Hightower, 1969). Instead of advancing by a minimum unit, we try to determine in one step a whole line segment containing the "current" point, extending the segment until an obstacle or a target is found. Once a segment is constructed, we explore new segments in the direction orthogonal to the previous one, starting from each of its end points. The search is conducted from the two points we wish to connect, and when the two set of segments intersect, the connection has been achieved. While the algorithm does not guarantee that a path is found, or even that a path exists, it is much faster than a Moore–Lee type algorithm and it takes less memory. Many refinements and combinations of these basic approaches have been used through the years, particularly for board routing. Another problem frequently studied was how to "rip out" a connection already made to allow the completion of others and to "reroute" the previous one.

As integrated circuits moved into the VLSI scale in the late 1970s, routing among their basic cells or blocks required dealing with areas that, properly scaled for the interconnect layer dimensions, were much larger than those found in boards. The routing problem was then split into subproblems, each with a different objective. The available area was divided into routing regions, usually rectangular, with special characteristics so that different algorithms could be applied in each type of area. The overall routing process was divided into phases, each adding increased levels of detail. In the case of ICs, the division of the chip into major functional blocks and routing regions is called floor planning.

The most important division introduced was between global and detailed routing, a concept applicable to IC and board problems. Global or "loose" routing is the process of simply distributing nets among the identified routing areas. Once this assignment is complete, the detailed routing phase determines the exact paths inside each area using any appropriate router. For global routing, there are two key considerations: to minimize an estimated net length and to avoid excessive congestion in any area (the allocation of so many nets into the area that detailed routing becomes physically impossible). One common approach has been to model the routing areas as vertices in a graph whose edges connect only neighboring areas. The model is refined by adding capacity to the edges to represent the number of nets that can be allowed in the region, and a length to indicate an expected average distance that a path will traverse inside the region. Once a model like this is set up, many algorithms can be called to assign the nets to the graph arcs.

Some special areas have become typical in IC layout, generating considerable research to solve their specific problems. The most important has been channel routing. In this case, we have a rectangular region with pins located only at its top and bottom edges; the vertical separation can be changed, and two conducting layers are available, one for each orthogonal direction. The problem is to connect all pins using the minimum possible number of horizontal tracks or, equivalently, vertical spacing (Hashimoto and Stevens, 1971). Many algorithms have been developed to solve this problem (see Lengauer, 1990). Another important region is the switch box, differing from a channel essentially by having pins on all sides and nonrectangular shape.

A related problem is the interaction among the different regions through the routing process. We may be forced to enlarge a channel region, altering the overall structure of the initial division: New regions can easily appear. More integrated approaches to resolve routing without resorting to complex area divisions are starting to receive attention. A key issue is the ability to route these large regions in reasonable computer times. The most recent CMOS gate arrays, offering hundreds of thousands of gates and three metal layers for interconnect, have made the problem particularly interesting for industry.

8. HIGHER-LEVEL DESIGN TOOLS

A number of approaches have been developed to move the design process as much as possible toward higher levels of abstraction. To reach this objective successfully, we need to increase the automation of the "lower-level" design tasks. Systems that automatically produced layouts starting from high-level descriptions were referred to as silicon compilers (Johanssen, 1979, Siskind *et al.*, 1982). The name was based on an analogy with a compiler for a software language, the behavior-level or RTL description playing the role of the language and the mask-level layout the equivalent of the object code. A large amount of research was done on the topic, and the approach generated considerable controversy. Ultimately, these early tools failed in actual applications because of their too broad scope. For a general-purpose microprocessor, many problems needed to be resolved simultaneously at a time when none of them, considered individually, had a fully satisfactory solution. For more recent developments, the reader is referred to De Micheli *et al.* (1987), and Gajski (1988).

Some silicon compilers attempted to map a description into a specialized chip architecture, restricting their applicability but taking advantage of optimizations valid only in their more restrictive domains. The most successful area has been in compilers targeted to digital signal processing algorithms. An early effort, a system developed at Edinburgh called FIRST, targeted filter structures using a bit serial architecture implemented in MOS technology (Denyer and Renshaw, 1985). Other ongoing efforts in this area that can be mentioned are the CATHEDRAL system at the Interuniversity Micro-Electronics Center in Belgium (Note *et al.*, 1991) and LAGER at the University of California at Berkeley (Pope *et al.*, 1987).

Other types of specialized systems in existence today are also called "silicon compilers." In general, these tools generate layouts for special kinds of circuits starting from a schematic like description and using libraries of cell layouts. So most are really special module generators that perform, beyond assembling a layout, specific optimization steps and typically generate appropriate simulation models. Standard examples include PLA and memory structure generators and data-path assemblers. These tools are offered by commercial CAD vendors and have been widely used in industry.

GLOSSARY

Bus: A set of bit signals of related type—for instance, the 32-bit signals commonly used as a word length in computer architectures.

Cell: A set of gates that is treated as a single entity. Cells are usually placed in libraries, with multiple descriptions available such as layout, schematic symbols, and simulation models.

Circuit: A collection of interconnected electrical devices. It consists of electrical devices and connected pins (nets or nodes). See Network.

CMOS: Complementary metal oxide semiconductor, an integrated-circuit technology. See MOS and NMOS.

Contact: Vertical connection from one layout layer to another; also referred to as via.

Data Path: A regular structure used in layout, consisting of functional units and busses connecting them.

Electrical Device: Primitive circuit element, such as resistor, capacitor, inductance, or transistor.

Floor Plan: The physical layout of the chip in terms of its main units, such as data paths, memories, and power and ground busses.

Gate Array: An integrated circuit where transistors are arranged in a two-dimensional array, and the design is done by specifying only the metal interconnect layers.

Grid: Uniformly spaced array of points in a coordinate space. Usually displayed in layout editors to guide the editing process and indicate the minimum geometrical resolution available.

Interconnect: A conductor electrically joining two or more pins.

Layout: The physical description of a design, in terms of polygons attached to technology layers, forming interconnect and electrical devices such as transistors.

Logic Gate: A functional unit of combination Boolean logic, such as an inverter (NOT function) or a NAND-2 (logical AND followed by NOT of two input signals).

Mask: Glass usually covered with chrome etched away in particular areas where specific material will be placed on the wafer during the semiconductor manufacturing process.

MOS: Metal oxide semiconductor, an integrated-circuit technology. NMOS and CMOS are the two most common MOS technologies.

Net: A set of electrically connected pins.

Netlist: A list of all nets in a circuit.

Network: A collection of interconnected electrical devices. It consists of electrical devices and connected pins (nets or nodes). See Circuit.

NMOS: *n*-type metal oxide semiconductor, an integrated-circuit technology.

Pin: A connection point in a device.

Router: A software program used to create the interconnections in a network where the devices have physical locations.

Standard Cell: A cell whose layout conforms to some uniform characteristics (such as given height). Usually, they form libraries, and designs are done with choice restricted to only those elements.

Transistor: An active semiconductor device with three (or more) terminals. It is a basic element in integrated-circuit design.

Via: Vertical connection from one layout layer to another; also referred to as a contact.

VLSI: Very-large-scale integration, a term used to describe ICs containing over 100 000 transistors.

Wire: The physical realization of a connection, indicating a detailed conducting path among pins.

List of Works Cited

Abramovici, M., Breuer, M. A., Friedmann, A. D. (1990), *Digital System Testing and Testable Design,* Rockville, MD: Computer Science Press.

Agrawal, P., Dally, W. T., Fisher, W. C., Jagadish, H. V., Krishnakumar, A. S., Tutundjian, R. (1987), "MARS: A Multiprocessor Based Programmable Accelerator," *IEEE Design Test Comput.,* **4** (5), 28–36.

Batali, J., Hartheimer, A. (1980), *The Design Procedure Language Manual,* VLSI Memo No. 80-31, Artificial Intelligence Laboratory, MIT, Cambridge, MA: MIT Press.

Benkowski, J., Strojwas, A. J. (1991), "The Role of Timing Verification in Layout Synthesis," *Proc. 28th Design Automation Conference, 1991,* Los Alamitos, CA: IEEE Computer Society Press, pp. 612–619.

Boyer, D. (1988), "Symbolic Layout Compaction Review," *Proc. 25th Design Automation Conference, 1988,* Los Alamitos, CA: IEEE Computer Society Press, pp. 383–388.

Brace, K. S., Rudell, R. L., Bryant, R. E. (1990), "Efficient Implementation of a BDD Package," *Proc. 27th Design Automation Conference, 1990,* Los Alamitos, CA: IEEE Computer Society Press, pp. 40–45.

Brayton, R. K., Hachtel, G. D., McMullen, C. T., Sangiovanni-Vincentelli, A. L. (1984), *Logic Minimization Algorithms for VLSI Synthesis,* Boston: Kluwer Academic.

Brayton, R. K., Brenner, N., Chen, C., De Micheli, G., McMullen, C. T., Otten, R. (1985), "The Yorktown Silicon Compiler," *Proc. Int. Symposium Circuits and Systems, 1985,* New York: IEEE Publishing Services, pp. 391–394.

Brayton, R. K., Rudell, R., Sangiovanni-Vincentelli, A. L., Wang, A. (1987), "MIS: A Multiple-Level Logic Optimization System," *IEEE Trans. Comput.-Aided Design,* **CAD-6** (6), 1062–1081.

Breuer, M. A. (Ed.) (1972), *Design Automation of Digital Systems,* Vol. 1, Englewood Cliffs, NJ: Prentice-Hall.

Bryant, R. E. (1986), "Graph Based Algorithms for Boolean Function Manipulation," *IEEE Trans. Comput.* **C-35**, (8), pp. 677–691.

Bryant, R., Beatty, D., Brace, K., Cho, K., Sheffler, T. (1987), "COSMOS: A Compiled Simulator for MOS Circuits," *Proc. 24th Design Automation Conference, 1987,* Los Alamitos, CA: IEEE Computer Society Press, pp. 9–16.

Darringer, J. A., Brand, D., Gerbi, J. V., Joyner, W. H., Trevillyan, L. (1984), "LSS: A System for Production Logic Synthesis," *IBM J. Res. Devel.* **28**, (5), 537–545.

de Geus, A., Cohen, W. (1985), "A Rule-Based Expert System for Optimizing Combinational Logic," *IEEE Design Test* **2** (4), 22–32.

De Micheli, G., Sangiovanni-Vincentelli, A., Antognetti, P. (1987), *Design Systems for VLSI Circuits; Logic Synthesis and Silicon Compilation,* Dordrecht, Netherlands: Martinus Nijhoff.

Denneu, M. (1982), "The Yorktown Simulation Engine," *Proc. 19th Design Automation Conference, 1982,* Los Alamitos, CA: IEEE Computer Society Press, pp. 55–59.

Denyer, P., Renshaw, D. (1985), *VLSI Signal Processing: A Bit Serial Approach,* Reading, MA: Addison-Wesley.

Dewey, A., Gadient, R. (1986), "VHDL Motivation," *IEEE Design Test Mag.* **3** (1), 12–16.

Dietmeyer, D. L. (1978), *Logic Design of Switching Functions,* 2nd ed., Boston: Allyn and Bacon.

Dunlop, A. E., Kernighan, B. W. (1985), "A Procedure for Placement of Standard-Cell VLSI Circuits," *IEEE Trans. Comput.-Aided Design,* **CAD-4** (1), 92–98.

Fiduccia, C. M., Mattheyses, R. M. (1982), "A Linear-Time Heuristic for Improving Network Partitions," *Proc. 19th Design Automation Conference, 1982,* Los Alamitos, CA: IEEE Computer Society Press, pp. 175–181.

Friedman, T. D., Yang, S. C. (1968), "Methods Used in an Automated Logic Design Generator (ALERT)," *IEEE Trans. Comput.* **C-18**, 593–613.

Gajski, D. (Ed.) (1988), *Silicon Compilation,* Reading, MA: Addison-Wesley.

Garey, M., Johnson, D. (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness,* San Francisco: Freeman.

Glasser, L. A., Dobberpuhl, D. W. (1985), *The Design and Analysis of VLSI Circuits,* Reading, MA: Addison-Wesley.

Hafer, L. J., Parker, A. C. (1983), "A Formal Method for the Specification, Analysis, and Design of Register-Transfer Level Digital Logic," *IEEE Trans. Comput.-Aided Design* **CAD-2** (1), 4–18.

Hashimoto, A., Stevens, J. (1971), "Wire Routing by Optimizing Channel Assignment within Large Apertures," *Proc. 8th Design Automation Conference, 1971,* Los Alamitos, CA: IEEE Computer Society Press, pp. 155–169.

Hightower, D. W. (1969), "A Solution to Line Routing Problems on the Continuous Plane," *Proc. 6th Design Automation Conference, 1969,* Los Alamitos, CA: IEEE Computer Society Press, pp. 1–24.

Hill, D., Shugard, D., Fishburn, J., Keutzer, K. (1989), *Algorithms and Techniques for VLSI Layout Synthesis,* Boston: Kluwer Academic.

Insinga, A. (1984), "Behavioral Modeling inside a Structural Level Simulator," *Proc. International Conference on CAD, 1984,* Los Alamitos, CA: IEEE Computer Society Press, pp. 42–44.

Johanssen, D. (1979), "Bristle Blocks: A Silicon Compiler," *Proc. 16th Design Automation Conference, 1979,* Los Alamitos, CA: IEEE Computer Society Press, pp. 310–313.

Kernighan, B. W., Lin, S. (1970), "An Efficient Heuristic Procedure for Partitioning Graphs," *Bell Syst. Tech. J.* **49**, 291–308.

Keutzer, K. (1987), "DAGON: Technology Binding and Local Optimization," *Proc. 24th Design Automation Conference, 1987,* Los Alamitos, CA: IEEE Computer Society Press, pp. 341–347.

Kirkpatrick, S., Gelatt, C., Vecchi, M. (1983), "Optimization by Simulated Annealing," *Science* **220**, 671–680.

Lee, C. Y. (1961), "An Algorithm for Path Connections and its Applications," *IRE Trans. Electronic Comput.* **EC-10**, (3), 346–365.

Lengauer, T. (1990), *Combinatorial Algorithms for Integrated Circuit Layout,* New York: Wiley.

Marwedel, P. (1986), "A New Synthesis Algorithm for the MIMOLA System," *Proc. 23th Design Automation Conference, 1986,* Los Alamitos, CA: IEEE Computer Society Press, pp. 271–277.

McCluskey, E. J. (1956), "Minimization of Boolean Functions," *Bell Syst. Tech. J.* **35**, 1417–1444.

Mead, C. A., Conway, L. (1980), *Introduction to VLSI Systems,* Reading, MA: Addison-Wesley.

Nagel, L. W. (1975), *SPICE2: A Computer Program to Simulate Semiconductor Circuits,* ERL Memo 520, 1975, University of California at Berkeley.

Note, S., Geurts, W., Catthoor, F., De Man, H. (1991), "CATHEDRAL-III; Architecture Driven High-Level Synthesis for High-Throughput DSP Applications," *Proc. 28th Design Automation Conference, 1991,* Los Alamitos, CA: IEEE Computer Society Press, pp. 597–602.

Osterhout, J., Hamachi, R., Mayo, R., Scott, W., Taylor, G. (1984), "Magic: A VLSI Layout System," *Proc. 21th Design Automation Conference, 1984,* Los Alamitos, CA: IEEE Computer Society Press, p. 152.

Pan, J-W. Biro, L., Grodstein, J., Grundmann, W., Yen, Y-T. (1991), "Timing Verification of a 1.2 M Device Full-Custom CMOS Design," *Proc. 28th Design Automation Conference, 1991,* Los Alamitos, CA: IEEE Computer Society Press, pp. 551–554.

Pillage, L. T., Rohrer, R. A. (1990), "Asymptotic Waveform Estimation for Timing Analysis," *IEEE Trans. Comput.-Aided Design* **CAD-9**, (4), 352–366.

Pope, S. P., Rabaey, J., Broderson, R. W. (1987), "Automatic Generation of Digital Signal Processing Circuits," in: G. De Micheli, *et al.* (Eds.), *Design Systems for VLSI Circuits; Logic Synthesis and Silicon Compilation,* Dordrecht, Netherlands: Martinus Nijhoff.

Romeo, F., Sangiovanni-Vincentelli, A. (1985), "Probabilistic Hill-Climbing Algorithms: Properties and Applications," in: H. Fuchs (Ed.), *1985 Chapel Hill Conference on VLSI,* Rockville, MD: Computer Science Press, pp. 393–418.

Rosenberg, J. B. (1985), "Geographical Data Structures Compared: A Study of Data Structures Supporting Region Queries," *IEEE Trans. Comput.-Aided Design,* **CAD-4** (1), 53–67.

Rubin, S. M. (1987), *Computer Aids for VLSI Design,* Reading, MA: Addison-Wesley.

Saleh, R. A., Newton, R. A. (1990), *Mixed-Mode Simulation,* Boston: Kluwer Academic.

Sechen, C. (1988), *VLSI Placement and Global Routing using Simulated Annealing,* Boston: Kluwer Academic.

Siskind, J., Southard, J., Crouch, K. (1982), "Macpitts: Generating Custom VLSI Design from Succinct Algorithmic Description," in: P. Penfield (Ed.), *Proc. MIT Conference on Advanced Research on VLSI, 1982,* Dedham, MA: Artech House, pp. 28–39.

Sze, S. M. (Ed.) (1984), *VLSI Technology,* New York: McGraw-Hill.

Terman, C. J. (1983), "RSIM: A Logic-Level Timing Simulator," *Proc. International Conference on Computer Design, 1983,* Los Alamitos, CA: IEEE Computer Society Press, pp. 437–440.

Thomas, D., Dirkes, E. M., Walker, R. A., Rajan, J. V., Nestor, J. A., Blackburn, R. L. (1988), "The System's Architect Workbench," *Proc. 25th Design Automation Conference, 1988,* Los Alamitos, CA: IEEE Computer Society Press, pp. 337–343.

Vladimirescu, A., Liu, S. (1980), "The Simulation of MOS Circuits Using SPICE2," ERL Memo 80/7, 1980, University of California at Berkeley.

White, J. K., Sangiovanni-Vincentelli, A. (1987), *Relaxation Techniques for the Simulation of VLSI Circuits,* Boston: Kluwer Academic.

Williams, J. D. (1978), "STICKS, a Graphical Compiler for High Level LSI Design," *Proc. AFIPS Conference* **47**, Montvale, NJ: AFIPS Press, pp. 289–295.

Further Reading

De Micheli, G., Sangiovanni-Vincentelli, A., Antognetti, P. (1987), *Design Systems for VLSI Circuits; Logic Synthesis and Silicon Compilation,* Dordrecht, Netherlands: Martinus Nijhoff.

Dietmeyer, D. L. (1978), *Logic Design of Switching Functions,* 2nd ed., Boston: Allyn and Bacon.

Gajski, D. (Ed.) (1988), *Silicon Compilation,* Reading, MA: Addison-Wesley.

Glasser, L. A., Dobberpuhl, D. W. (1985), *The Design and Analysis of VLSI Circuits,* Reading, MA: Addison-Wesley.

IEEE Proceedings (1990), Special Issue on the Future of Computer-Aided Design, *IEEE Proc,* **78**, (2), February 1990.

Lengauer, T. (1990), *Combinational Algorithms for Integrated Circuit Layout,* New York: Wiley.

Lund, P. (1986), *Printed Circuit Board Precision Artwork Generation and Manufacturing Methods,* Westlake Village, CA: Bishop Graphic.

Noble, P. J. W. (1989), *Printed Circuit Board Assembly,* Milton Keynes: Open University Press.

Tredennick, Nick (1987), *Microprocessor Logic Design,* Bedford, MA: Digital.

Saleh, R. A., Newton, R. A. (1990), *Mixed-Mode Simulation,* Boston: Kluwer Academic.

Weste, N., Eshraghian, K. (1985), *Principles of CMOS VLSI Design: A Systems Perspective,* Reading, MA: Addison-Wesley.

COMPUTER-AIDED MANUFACTURING (CAM)

Kazuaki Iwata, *Department of Mechanical Engineering for Computer-Controlled Machinery, Osaka University, Osaka, Japan*

	Introduction	89
1.	**History**	90
2.	**Basic Concept and Major Components of CAM**	92
3.	**Process Planning**	92
3.1	Processing Flow of Process Planning in Machining for Machine Parts	92
3.1.1	Understanding Design Information and Drawings, and Recognition of Production Features	92
3.1.2	Determination of Production Processes and Their Sequence	92
3.1.3	Determination of Production Machines	93
3.1.4	Selection and Design of Jigs and Fixtures	93
3.1.5	Estimation of Operation Time	93
3.2	Process Planning for Assembly of Printed Circuit Boards	94
3.3	Computer-Aided Process Planning System in Machining	94
3.3.1	Variant-Type CAPP System	94
3.3.2	Generative-Type CAPP System	94
3.3.3	Expert-Type CAPP System	96
4.	**Operation Planning**	96
4.1	Processing Flow in Operation Planning	96
4.1.1	Determination of the Operation Sequence and Operators within a Process	97
4.1.2	Selection of Tools	97
4.1.3	Determination of Conditions of Operation	97
4.1.4	Tool Path Generation	97
4.1.5	Postprocessing	97
4.2	Computer-Aided Operation Planning Systems	97
5.	**Production Control**	98
6.	**Computer Numerical Control and Direct Numerical Control**	99
6.1	CNC	99
6.2	DNC	99
7.	**Development of CIM**	99
8.	**Summary**	101
	Glossary	101
	Works Cited	102
	Further Reading	102

INTRODUCTION

Manufacturing is defined as a function of transforming raw materials into valuable products that meet the public's needs, by means of facilities, information, and labor power. It is also expressed more concretely as a series of interrelated activities and operations involving the design, material selection, planning, production, quality assurance, management, and marketing of the product by the manufacturing industries (CAM-I, Standard Committee, 1978). Automation in manufacturing has made great strides since the invention of electronic computers. Some typical examples of manufacturing technologies developed so far in conjunction with computers are CAD (computer-aided design), CAM (computer-aided manufacturing), CAD/CAM, FA (factory automation), and CIM (computer integrated manufacturing). CAM, which is the subject of this article, is defined as execution of process and operation planning with the proper input data such as product design specification and production order, and control of production processes with the assistance of computers. The interrelationships between CAD, CAM, FA, and CIM are explained in Sec. 2 (see also Fig.1).

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

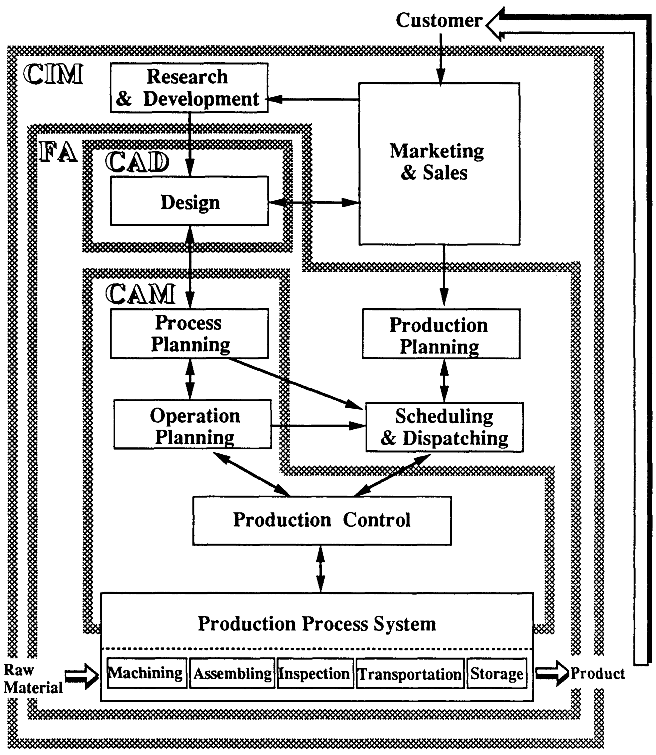
Customer
CIM
Research & Development
FA
CAD
Design
Marketing & Sales
CAM
Process Planning
Production Planning
Operation Planning
Scheduling & Dispatching
Production Control
Production Process System
Machining
Assembling
Inspection
Transportation
Storage
Raw Material
Product

FIG. 1. Role of CAM within framework of CIM.

1. HISTORY

The history of CAM dates back to the early 1950s when research was initiated for the automation of cutting processes by means of NC (numerical control).

NC is a control method in which motions of machines are controlled according to a number of pulses that are generated by a digital circuit depending on input numeric data. Once a program that represents the numeric data is prepared, the operation of the machine can be controlled automatically. NC can be applied to various machines such as machine tools, industrial robots, plotters, etc. The control equipment for NC is generally called the NC controller.

The developments of numerical control have been made mainly in two areas. One is the development of the NC controller that automates production processes, and another is development in the NC program preparation that leads to automation in the area of information processing for manufacturing.

In the first generation of NC controllers, functions such as arithmetic operations, comparisons, and decisions were carried out by a hard-wired logic circuit. However, complicated circuits became necessary with the expansion of the functions. In the early 1970s, CNC (computer numerical control), which used a computer for performing these functions, was developed to improve functionality, expansibility, operability, and maintainability. Nowadays microprocessors are used. DNC (direct numerical control) was also developed in the late 1960s to control and coordinate a number of numerically controlled machines. The technical developments of NC controllers are summarized in Table 1.

Efforts have also been made to add new control features to the NC controller, such as adaptive control and learning control. An adaptive controller identifies the changes in the states of the manufacturing processes and system components, adjusts the control parameters so as to fulfill the control objectives, and then executes the control. Adaptive controllers have already been put into practical use in certain applications. In the next generation of NC controllers, the development of a learning controller is also desired. In addition to the adaptive control function, a learning

Table 1. Development of NC Controller.

Year Item	1976–1979	1980–1985		1986–1991	
		Microprocessor			
Processor	Bit slice	8 bit	16 bit	32 bit	CISC/RISC
Technical trend	Expansion of functions and improvement of human–machine interface through computerization of NC			Improvement speed and accuracy through the use of 32-bit processor Human friendly CNC	
	Expansion of NC functions, cutter compensation, fixed cycle, memory operation, etc.	Improvement of human–machine interface, color CRT, interactive symbolic programming, graphic simulation		High-speed, high-accuracy machining control, easy programming, intelligent control	
Servo system	DC analog servo		AC analog servo	AC digital servo	

controller has the function of accumulating and making use of the data obtained in the preceding operations to improve the control performance.

The automation of information processing for manufacturing originates in the development of a general-purpose NC programming language called APT (automatically programmed tools). It was developed at MIT for automatic tool path generation. Later on, IITRI (Illinois Institute of Technology Research Institute) took over the responsibility of the development program and improved the system. The project resulted in APTIII and APTIV, which have greatly contributed to the automation of NC data generation.

Tens of general-purpose NC programming languages have been developed in the world since then. At the same time, a large number of special-purpose programming languages that can be applied to specific machine tools or products have been developed by many companies. The success of the APT system has stimulated the development of programming languages for many other manufacturing processes, such as those for plastic forming, assembling, and inspection.

The APT system can only generate tool paths of cutting processes. However, there has been a demand for the extension of the programming system functions to cover the automatic selection of tools and the determination of operating conditions. EXAPT (extended subset of APT), developed in the mid-1960s, is the first general-purpose system that was provided with such functions. Many special-purpose NC programming systems with these functions have been also developed.

The next step in the development of CAM information systems was to realize a system that could execute all necessary information processing from obtaining CAD data to generating data for production processes, such as NC data and process instructions. This kind of system is called the CAD/CAM system. The main problem of the system was that of developing CAPP (computer-aided process planning). Three approaches have been proposed for CAPP so far; they are a variant type, a generative type, and an expert type. The studies of these three types of approaches were started in the late 1960s, the mid-1970s, and the early 1980s, respectively.

In the early 1980s, people became interested in producing a variety of products; therefore, the integration of CAPP and production scheduling in addition to the integration of CAD and CAPP were proposed (e.g., Chryssolouris and Chan, 1985; Iwata and Fukuda, 1989).

In 1988, the concept of concurrent engineering or simultaneous engineering was proposed suggesting parallel processing of design and manufacturing information with constant interactions with each other rather than serial processing (Pennell and Winner, 1989).

FA came into practical use in the early 1980s, and from the mid-1980s, CIM (computer integrated manufacturing) has been put into practice. (The first book about CIM is Harrington, 1973.) With the advance of FA and CIM, CAM is regarded as the integrating component of the total manufacturing system.

2. BASIC CONCEPT AND MAJOR COMPONENTS OF CAM

CAM covers the functions of process planning, operation planning, production control, and production process systems as shown in Fig. 1. FA, on the other hand, includes design and production management (production planning, scheduling, and dispatching) in addition to the activities involved in CAM. Further expansion of the concept of FA leads to CIM, which includes R&D (research and development), marketing, and sales. Therefore, CAM can be understood as a subset of FA and CIM.

The information processing relevant to design, process planning, and operation planning is generally called technological information processing. Thus CAM plays a main role in technological information processing.

Process planning is the planning of production processes of parts or products based on the information and drawings generated in the design stage. Production processes include operations (machining, assembling, painting, packaging, etc.) and interprocess operations (transportation, storage, etc.). The results of the process planning are used as input information for operation planning and scheduling/dispatching.

In operation planning, operations comprising the process are identified and various items required for the executions of operations are determined, such as machining conditions and tool paths.

Production control is processing of control and management information for achieving efficient and reliable execution of various operations and interprocess operations. The production process is a series of operations to transform material into valuable products. CNC and DNC are the important elements of systems that execute the production processes.

It should be noted that in some countries the term CAM is used with a wider or narrower meaning than the definition given here. It may include the functions of production planning, scheduling, and dispatching in addition to the functions defined here, or on the contrary, it may exclude production process systems. The term process planning may also be used in a different way. It may involve functions of both the process planning and operation planning.

3. PROCESS PLANNING

3.1 Processing Flow of Process Planning in Machining for Machine Parts

As defined in the previous section, process planning is the process of generating the production process information such as information about production processes, production machines, jigs, fixtures, and so on, based on the data obtained from design information and drawings (geometry, dimension, tolerance, working conditions, etc.). The results of process planning are usually described in the form of a process chart. Figure 2 shows the information processing flow in the process planning of machine part production. Although the figure shows a straightforward process, there could be a number of feedback loops in the process. This processing flow can be applied to process planning for the production of other types of products such as semiconductors, printed circuit boards, apparel, food, etc., though there are specific features and particular considerations in the details of processing according to the type of product.

3.1.1 Understanding Design Information and Drawings, and Recognition of Production Features. The first step of process planning is to understand the information represented in assembly drawings, part drawings, and part lists, which are prepared in the design stage usually using CAD systems. The information is then checked to see whether it provides all the necessary and error-free information for the subsequent steps. At the same time, the features of products or parts are identified from the aspect of production; these are generally called production features. For example, typical forms or surfaces to be machined are recognized to prepare for the subsequent steps.

3.1.2 Determination of Production Processes and Their Sequence. In this process, possible production processes are selected on the basis of design information, drawings,

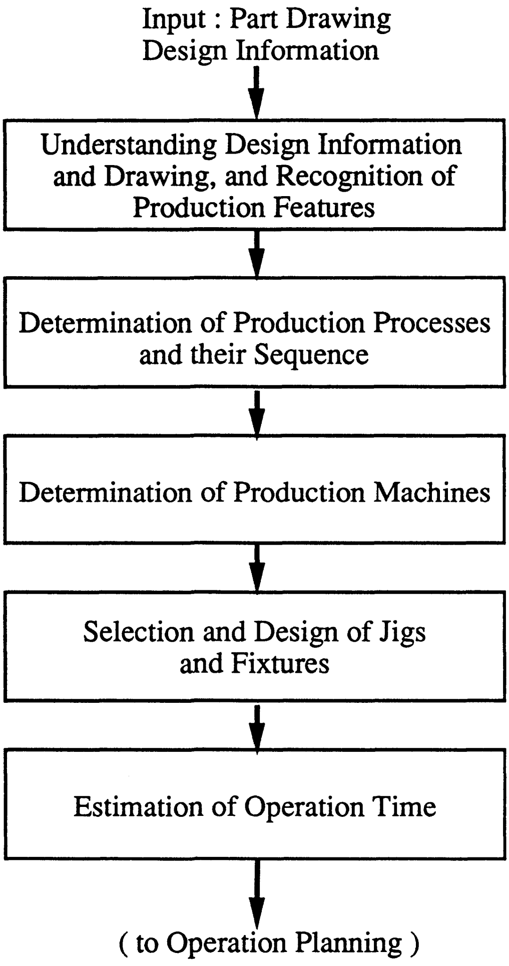
Input : Part Drawing
Design Information
Understanding Design Information and Drawing, and Recognition of Production Features
Determination of Production Processes and their Sequence
Determination of Production Machines
Selection and Design of Jigs and Fixtures
Estimation of Operation Time
(to Operation Planning)

FIG. 2. Information processing flow of process planning of machine parts production.

and manufacturing features. There can be many possible combinations of various production processes such as deformation processing (casting, forging), removal processing (machining, grinding), joining (welding, adhesion), and treatment (surface treatment, heat treatment) for producing the same part with the same geometry. The optimum combination of production processes is determined on the basis of various evaluations in terms of quality, lot size, economic efficiency, and resource savings.

In the case of machining, for example, possible machining processes (such as turning, drilling, end milling, cylindrical grinding, etc.) are selected for each form or surface to be machined in the part. Another example is the selection of a welding method. There can be arc welding, resistance welding, electroslag welding, electron beam welding, laser welding, etc., depending on material and geometries of parts.

If there are multiple possibilities, the optimum processes and their sequences are selected in terms of available facilities, technological level and experience, economic efficiency, production time, etc.

3.1.3 Determination of Production Machines. The production machines (machine tools, forming machines, assembling machines, etc.) are determined corresponding to the production processes selected in the previous step. Since the forms and surfaces of the part are generated by the relative motion of the tool and the part, the possible tool geometry can be identified in this step.

It should be noted that one production machine may be able to perform multiple production processes. Therefore, there could be a number of possible combinations of machines. The optimum combination should be selected considering production time, the number of machines, load balance of the machines, economic efficiency, etc.

3.1.4 Selection and Design of Jigs and Fixtures. Fixtures for fixing workpieces on the pallets or tables of the production machines, and jigs for positioning and clamping the workpieces and for guiding the tools, are selected in this step. This is called auxiliary equipment and is used for improving accuracy, simplifying operations, and shortening the working time.

There are various technological and technical factors to be considered in designing and selecting the auxiliary equipment. These factors are classified into mainly three categories. They are the factors about the production machine such as the position and orientation of the spindle; the factors about the workpiece such as the gravity center and the reference planes for clamping, machining, and assembling; and the factors about the production operations such as vibration and thermal deformation.

3.1.5 Estimation of Operation Time. The operation time is estimated on the basis of the results of the steps described in Secs. 3.1.1–3.1.4. If there are multiple options in the process plan, the shortest operation time is adopted as the estimated time. The estimated time in this step is, however, a rough estimation based on standard times. The precise estimation is carried out after the operation planning.

3.2 Process Planning for Assembly of Printed Circuit Boards

Assembling methods, sequences, and machines are determined in assembly process planning. Few general-purpose process planning systems have been developed for assembling operations that differ from those for machining operations. Process planning systems developed so far are specific to the assembling operations by particular machines. In the following, the process planning for assembly of printed circuit boards, which are widely used in modern industries, is explained.

Figure 3 shows the processing flow of process planning for mounting electronic parts (resistors, condensers, ICs, etc.) on printed circuit boards. Obtaining the data from the CAD system, assembling processes are assigned first. Then collisions between insertion heads of inserters and parts are examined. This process, which is generally called collision check, is important, because the proper insertion sequence should be selected so as to avoid these collisions. Next the assembling processes are determined. Finally, the optimum operation sequence is determined, if there are multiple possibilities in determining insertion sequences. The total mounting time is usually adopted as a criterion for optimization (Japan Society for Precision Engineering, 1991).

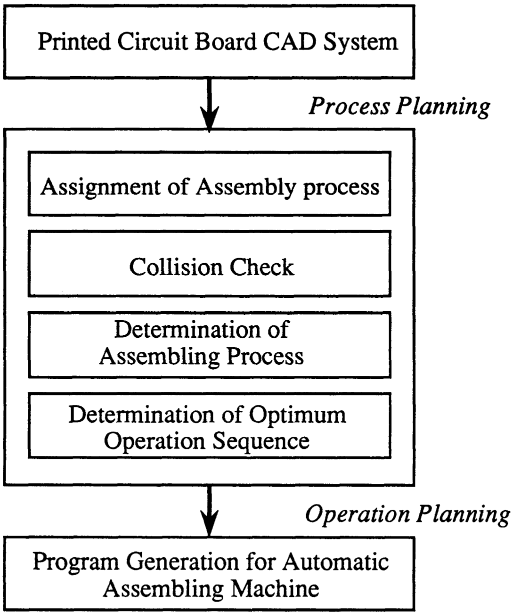
Printed Circuit Board CAD System
Process Planning
Assignment of Assembly process
Collision Check
Determination of Assembling Process
Determination of Optimum Operation Sequence
Operation Planning
Program Generation for Automatic Assembling Machine

FIG. 3. Information processing flow of process planning for mounting electric parts on printed circuit boards.

3.3 Computer-Aided Process Planning System in Machining

Automation of process planning has been studied mainly for machining. The efforts in developing the CAPP system have contributed to the standardization of manufacturing processes and facilities, and for the rationalization of organizations. However, technologies have not matured to the point that current CAPP systems provide sufficient capabilities for CAD/CAM integration. Thus, the improvement of CAPP systems is an important subject for future factories.

The CAPP systems developed so far are categorized into the following three types:

1. variant type,
2. generative type, and
3. expert type.

3.3.1 Variant-Type CAPP System. The variant-type system uses a decision table for planning the production processes as shown in Fig. 4. The decision table represents the logic of the process planning that has been acquired from past experience. Many systems utilize the concept of GT (group technology) in which parts are grouped into families based on similarities in their production. The code representing the parts family is called the GT code. G2030 and H1819 in Fig. 4 are examples of GT codes. The decision table can be formed so as to represent the sequences of production processes corresponding to the specific GT codes. The systems of this type were the first of the three types to be developed. AUTOPROS (Nissen, 1969), MIPLAN (TNO, 1981), CAPP (CAM-I, 1976), and AUTAP (Eversheim and Fuchs, 1980) are some well-known systems of this type.

3.3.2 Generative-Type CAPP System. In the generative-type CAPP system, the process planning is performed on the basis of the logic that is obtained from the analysis of the part

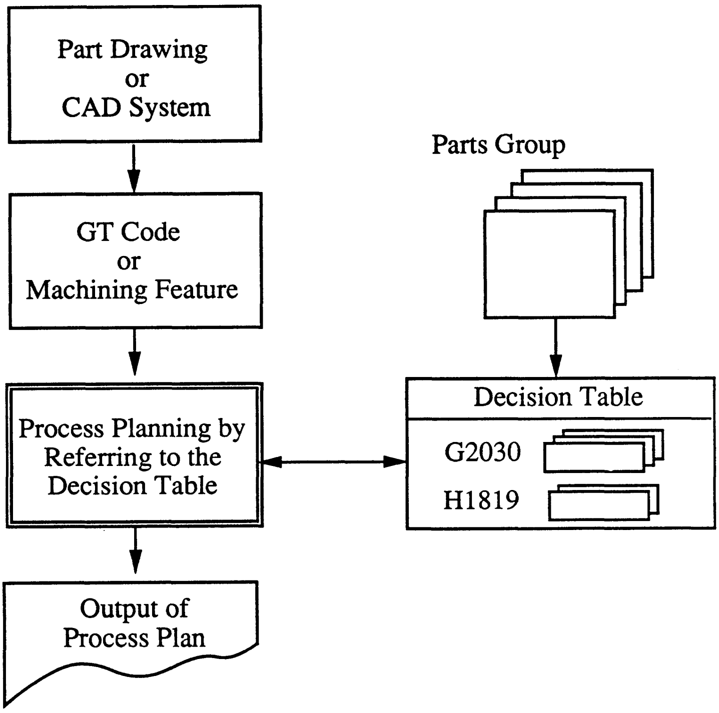
Part Drawing
or
CAD System
GT Code
or
Machining Feature
Process Planning by
Referring to the
Decision Table
Output of
Process Plan
Parts Group
Decision Table
G2030
H1819

FIG. 4. Basic concept of variant-type process planning.

and the machine tool, in terms of part geometry, required accuracy, form-generating functions of the machine tool, and other related manufacturing technologies. Figure 5 illustrates the basic concept of the generative-type system. Although systems of this type have been studied for many years, further studies are required to develop practical systems, because the logical relations between the production features of the part and the production processes are not clarified.

Some typical research works about systems of this type are CAPSY (Spur *et al.*, 1978), CIMS/PRO (Iwata *et al.*, 1977), and CPPP (Mann, 1977).

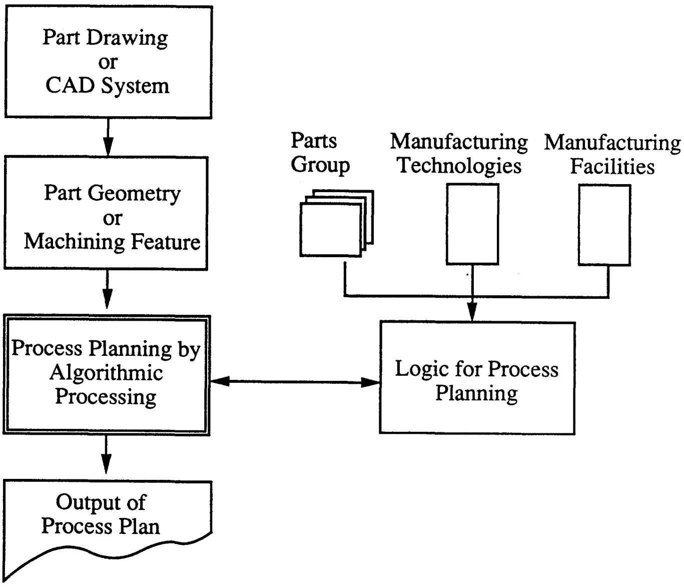
Part Drawing
or
CAD System
Part Geometry
or
Machining Feature
Process Planning by
Algorithmic
Processing
Output of
Process Plan
Parts
Group
Manufacturing
Technologies
Manufacturing
Facilities
Logic for Process
Planning

FIG. 5. Basic concept of generative-type process planning.

3.3.3 Expert-Type CAPP System. In recent years, as it became clear that the knowledge and know-how of the process planner played an important role in the practice of the process planning, many expert systems were developed. These systems consist of the inference engine, the knowledge base (containing the know-how and knowledge of the experts), the database (containing general data), the system that acquires the knowledge from the experts, and the interface between the CAD system and the users, as shown in Fig. 6. Many systems of this type have been developed, such as GARI (Descotte and Latombe, 1985), TOM (Matsushima *et al.*, 1982), CIMS/PRO (Iwata and Sugimura, 1984), XPLAN (Van't Erve and Kals, 1986), and KAPPS (Iwata and Fukuda, 1987). Current research work is mainly focusing on a feature-based approach. Process planning is associated with production features identified in the geometric models of parts that are generated in the design stage using CAD systems.

Until 1980, research and development work about CAPP was carried out in the field of removal processing, especially for machining. In the 1980s, however, research was extended to cover assembling and deformation processing (casting, forging, sheet-metal forming). Since 1990, it has been recognized that research on automation of process planning for all manufacturing processes is needed.

4. OPERATION PLANNING

4.1 Processing Flow in Operation Planning

The task of operation planning is to determine the detailed information necessary for the execution of production processes (such as machining and assembling), on the basis of the output of the process planning. The information processing flow for the operation planning of machine parts includes

1. determination of the operation sequence and operators within a process,
2. selection of tools,
3. determination of conditions of the operations,
4. determination of the tool path, and
5. postprocessing.

Operation planning of products or parts with simple geometries does not require much experience. However, for products with complicated geometries even the determination of

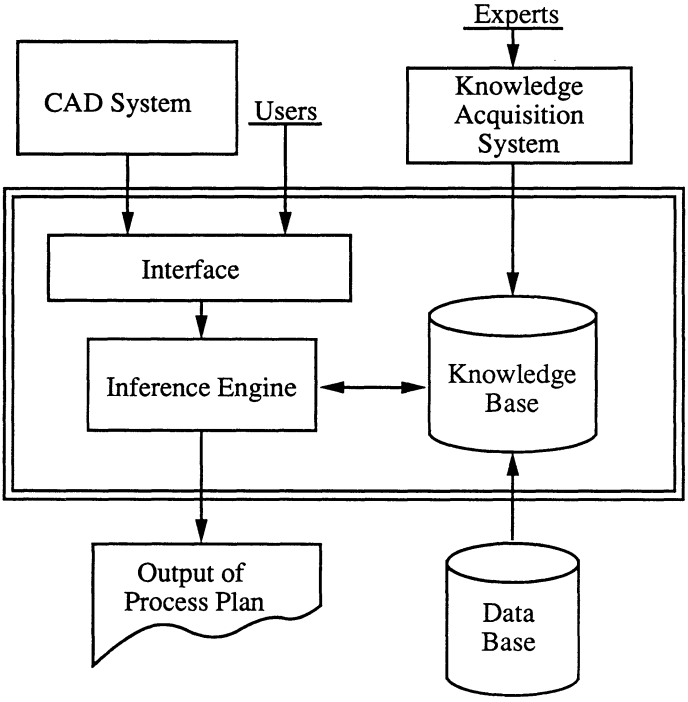

FIG. 6. Basic structure of expert-type process planning system

the tool path is a difficult task and requires a lot of work, time, and cost. A lot of effort has, therefore, been made on the automation of operation planning.

4.1.1 Determination of the Operation Sequence and Operators within a Process. Each production process determined by process planning is analyzed to identify what operations should be executed in which sequence and what actions of the operators are required within the process. Let us consider the operation of external cutting in turning, for example. The required operations are chucking the workpiece, clamping by the tail stock, air cutting, cutting (a number of times, if necessary), measuring, detaching the workpiece, cleaning the machine, disposing of chips, etc. In operation planning, such operations must be properly selected and their sequence must be determined. The series of operations for one production process is usually called the work cycle.

4.1.2 Selection of Tools. In this step, proper tools in terms of material, geometries, and dimensions are selected on the basis of the selected machines (type, size, rigidity, power, structural configuration, etc.), workpieces (material, hardness, rigidity dimension, and accuracy before and after machining, etc.), and types of operations.

4.1.3 Determination of Conditions of Operation. The proper conditions of the operations are determined in this step on the basis of the selected machine, tools, fixtures, and the role of the operators if operators are necessary. The conditions have to be determined so as to produce a product or part that meets the specifications and also to achieve efficient, economic, reliable, and safe operations.

Many studies have been carried out about the optimization of the condition of operations. Note that there are multiple operations to be considered in optimization, and different companies have different objective functions and constraints.

4.1.4 Tool Path Generation. In this step, proper tool path is generated according to the geometry of the part, the movement of the machine tool or assembling machine, and the geometry of the tools. The path must be determined to avoid collision between the tool and tool holder, on the one hand, and the workpiece and the jigs, on the other, and also to achieve minimum operation time.

In the past, the problem of tool path generation was dealt with independently whether the part geometry was $2\frac{1}{2}$ dimensional or 3 dimensional. However, efforts have been made to unify these problems in recent years. The term $2\frac{1}{2}$-dimensional is a special term in the field of manufacturing. It is defined as the dimensionality that is composed of a contouring motion in a 2-dimensional plane and a linear motion along the line perpendicular to the plane. Examples of $2\frac{1}{2}$-dimensional machining are drilling operations, face milling operations, and turning operations.

4.1.5 Postprocessing. In the previous steps, the tool path data were obtained in the form of cutter location data (CL data), which has a general format independent of the configuration of the machine tool. Postprocessing is a process to convert CL data into NC data adapted for the specific machine tool. The postprocessor should be designed so that NC data generated by it should not contain waste motions, should be adjusted to the timing of the mechanisms such as an automatic tool changer, and should be accompanied by proper documents that can be used for various checks by the operator.

The results of operation planning as described above are available through various media such as paper, tapes, and floppy disks. Nowadays they are often transmitted directly from the computer to the machine.

4.2 Computer-Aided Operation Planning Systems

Most computer-aided operation planning systems have been developed as special-purpose systems in companies; thus very few general-purpose systems have been developed.

The most famous general-purpose systems are APT (automatically programmed tools) and its subsets that were developed for the automation of tool path generation. EXAPT (extended subset of APT) is also well known as a general-purpose system that covers the determination of operation sequence, the selection of tools, and the determination of conditions of operations in addition to the tool path generation (e.g., Budde and Weisswellar, 1973).

The EXAPT processor reads a part program written in EXAPT language, executes technological calculations, and outputs the tool path data in the form of CL data, which are then processed by the post processor and converted to the NC data adapted to the specific machine. It can also issue work lists and work orders to the operator.

In the APT and EXAPT systems, NC language is used to input part programs into the system. Other types of input methods have also been developed such as a menu-driven interactive system and a direct interfacing to CAD system.

5. PRODUCTION CONTROL

As described previously, the objective of production control is to realize effective and safe operations of production processes such as machining, assembling, and inspection, and interprocess operations such as transportation and storage. As seen in Fig. 1, the data utilized for production control are classified into the following categories:

1. technological data generated by process planning and operation planning and management data prepared by production planning and scheduling, and
2. the data gathered in the production process (shop floor).

Figure 7 shows the software organization of the production control system which generates the control information using various relevant data. The system has five major functions: data management, operation control, operation monitoring, performance tracking, and status display.

The basic subfunctions and typical application-oriented subfunctions of each function are represented in the figure. The extent of the functions that are covered by a system is dependent on the level and organization of the system. Taking for example FMS (flexible manufacturing system: computer-controlled automated manufacturing system for mid-volume, mid-variety parts production, in which NC machine tools or NC manufacturing cells are combined organically by automatic transportation equipment), the operation control function is, so far, limited to the control of machining operations. There are very few systems that have the functions of assembling control and washing control. Recent systems, however, tend to be equipped with all these functions.

The operation monitoring function is an important function in order to realize automated operation of the production processes.

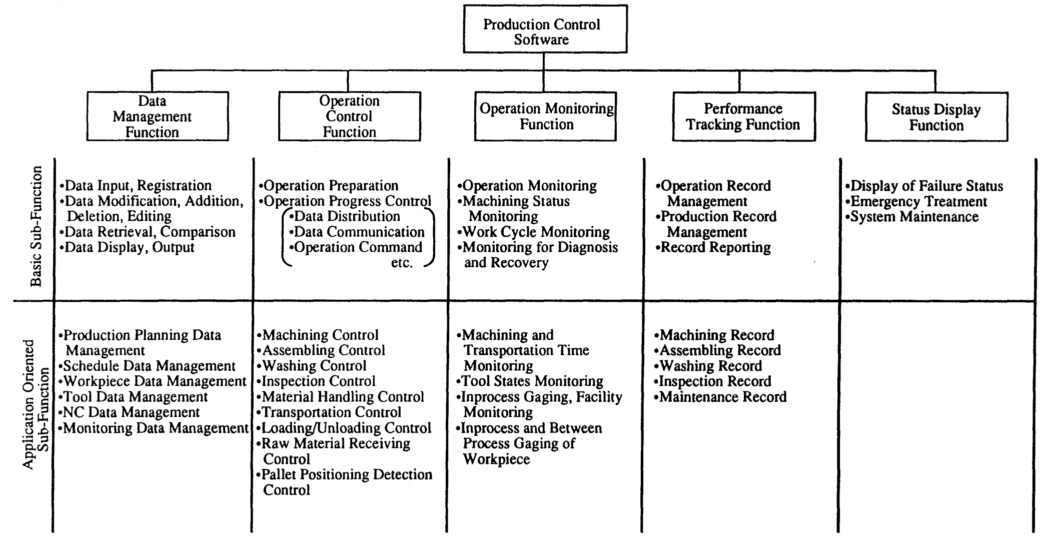
Production Control Software
Data Management Function
Operation Control Function
Operation Monitoring Function
Performance Tracking Function
Status Display Function
Basic Sub-Function
Application Oriented Sub-Function
•Data Input, Registration
•Data Modification, Addition, Deletion, Editing
•Data Retrieval, Comparison
•Data Display, Output
•Operation Preparation
•Operation Progress Control
•Data Distribution
•Data Communication
•Operation Command
etc.
•Operation Monitoring
•Machining Status Monitoring
•Work Cycle Monitoring
•Monitoring for Diagnosis and Recovery
•Operation Record Management
•Production Record Management
•Record Reporting
•Display of Failure Status
•Emergency Treatment
•System Maintenance
•Production Planning Data Management
•Schedule Data Management
•Workpiece Data Management
•Tool Data Management
•NC Data Management
•Monitoring Data Management
•Machining Control
•Assembling Control
•Washing Control
•Inspection Control
•Material Handling Control
•Transportation Control
•Loading/Unloading Control
•Raw Material Receiving Control
•Pallet Positioning Detection Control
•Machining and Transportation Time Monitoring
•Tool States Monitoring
•Inprocess Gaging, Facility Monitoring
•Inprocess and Between Process Gaging of Workpiece
•Machining Record
•Assembling Record
•Washing Record
•Inspection Record
•Maintenance Record

FIG. 7. Software organization for production control.

This function includes monitoring of machining and assembling operations, monitoring of work cycles and monitoring for failure diagnosis, and recovery actions.

It is also very important to pay attention to the human interface that provides the means of communication between the system component and personnel, especially when the system advances in automation.

6. COMPUTER NUMERICAL CONTROL AND DIRECT NUMERICAL CONTROL

Machines and equipment that comprise the production process system are commanded by production control. Computer numerical control (CNC) and direct numerical control (DNC) play a major role in controlling these machines and equipment.

6.1 CNC

The typical architecture of a CNC system is shown in Fig. 8. In a CNC system, most of the NC functions such as input/output, arithmetic and control, and servo control are performed by microprocessors instead of hardwired circuits. As seen in Fig. 8, various processing units are contained, such as a communication control processor, a graphic processor, a peripheral equipment control processor for robot and sequence control, and a servo processor. It also has a PTR/FD interface and main spindle drive interface.

Input devices are dependent on the input method of information. Industrial robots, for example, often use a teaching pendant instead of a paper-tape reader by which an operator can instruct the controller in motions of robots. It was a common practice to use paper tapes or floppy disks containing NC programs for inputting the control information to the CNC controller, but recently, an interactive programming function has become available on CNC controllers. With this function, an operator inputs necessary data under the guidance of messages and menus displayed on a CRT in an operational panel. The CNC controller then generates NC programs automatically.

6.2 DNC

DNC systems have been developed for the purpose of controlling and coordinating a number of machine tools and robots on a shop floor by a central DNC computer. The DNC computer manages part programs and distributes them to the machines. The management functions of the part programs include the functions of input/output, storage, editing, copy, deletion, modification, renewal, retrieval, etc. While the DNC computer transmits the part programs to the machines, it monitors the operating status of the machines and coordinates their operations. At the same time, it can perform other functions such as production management and part programming. A DNC system is effective in improving the availability of machines because of the timely distribution of NC data. It is also effective in improving the reliability of distributing part programs, because it can transmit them directly to the machine without use of NC tapes that have been commonly used with NC machine tools.

To make the most of the DNC system advantage, it is important to use effective DNC computer software for production management, such as software for scheduling, controlling machine loading and material supply, and report generation (percentage of machine utilization, production performance, etc.).

7. DEVELOPMENT OF CIM

The application of computers in the manufacturing field has made rapid progress since the 1960s. Until the mid-1980s the systems were developed separately in each area of manufacturing such as design, production planning, production management, and shop floor control. In the late 1980s, however, there appeared a strong movement to the integration of the systems of different areas, which is called CIM (computer integrated manufacturing). The trend from CAD and CAM to CIM is illustrated in Fig. 9.

The several levels of CIM can be considered in the course of development as seen in Fig. 9. In the early 1990s, most CIM systems correspond to level 3, which integrates the activities from the acceptance of customer's

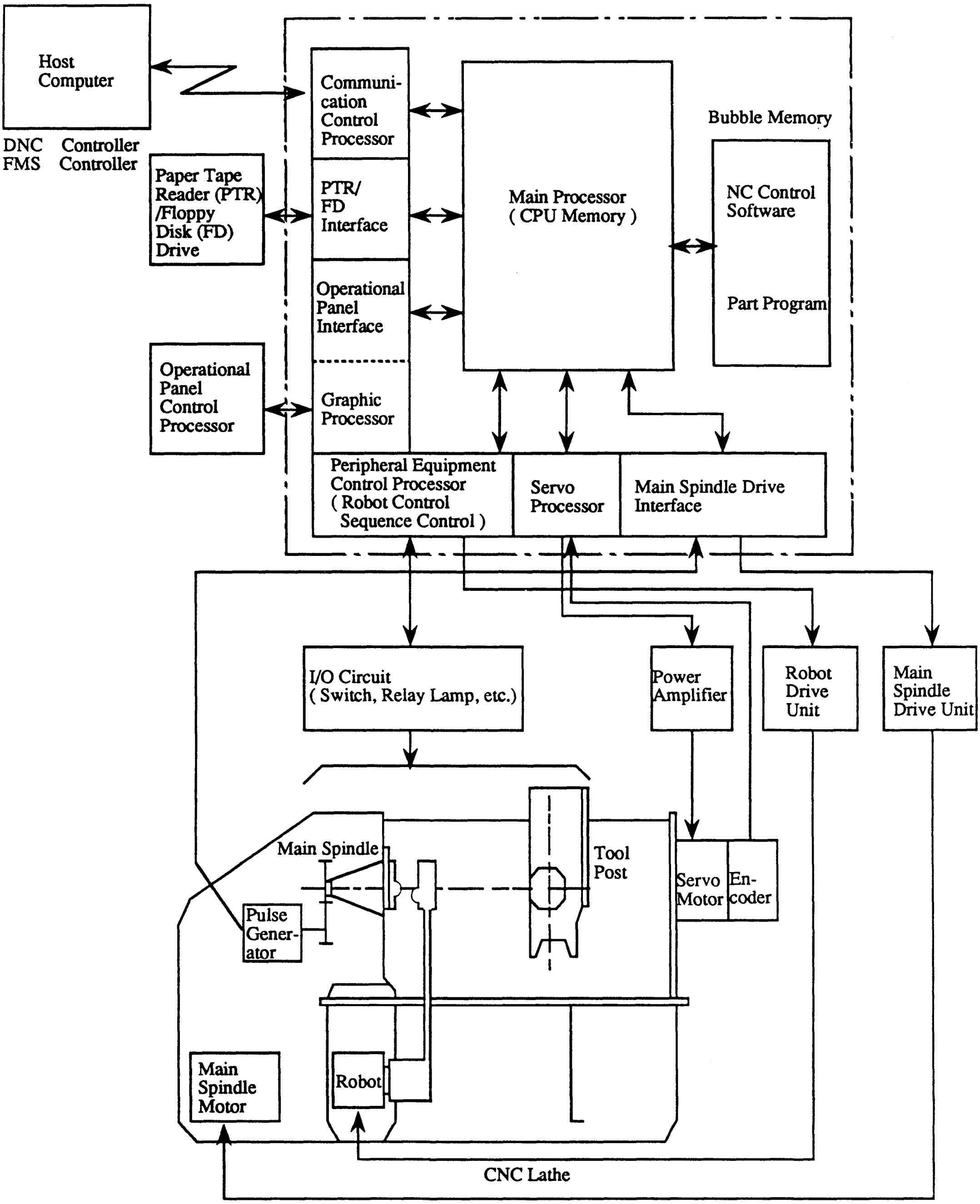
Host Computer
DNC Controller
FMS Controller
Paper Tape Reader (PTR) /Floppy Disk (FD) Drive
Operational Panel Control Processor
Communication Control Processor
PTR/ FD Interface
Operational Panel Interface
Graphic Processor
Main Processor (CPU Memory)
Bubble Memory
NC Control Software
Part Program
Peripheral Equipment Control Processor (Robot Control Sequence Control)
Servo Processor
Main Spindle Drive Interface
I/O Circuit (Switch, Relay Lamp, etc.)
Power Amplifier
Robot Drive Unit
Main Spindle Drive Unit
Main Spindle
Pulse Generator
Tool Post
Servo Motor
Encoder
Main Spindle Motor
Robot
CNC Lathe

FIG. 8. Basic structure of CNC system (Courtesy of Okuma Iron Co., Japan).

needs and orders to shipping and distribution of products. The system aims to achieve the integrated processing of manufacturing tasks by realizing a coherent flow of information and material. As the number of companies and researchers engaged in CIM increases, however, the concept of CIM tends to diversify and to become ambiguous. The organization of level-3 CIM is represented in Fig. 1.

The base technologies that support CIM are

1. communication network,
2. database, and
3. interface.

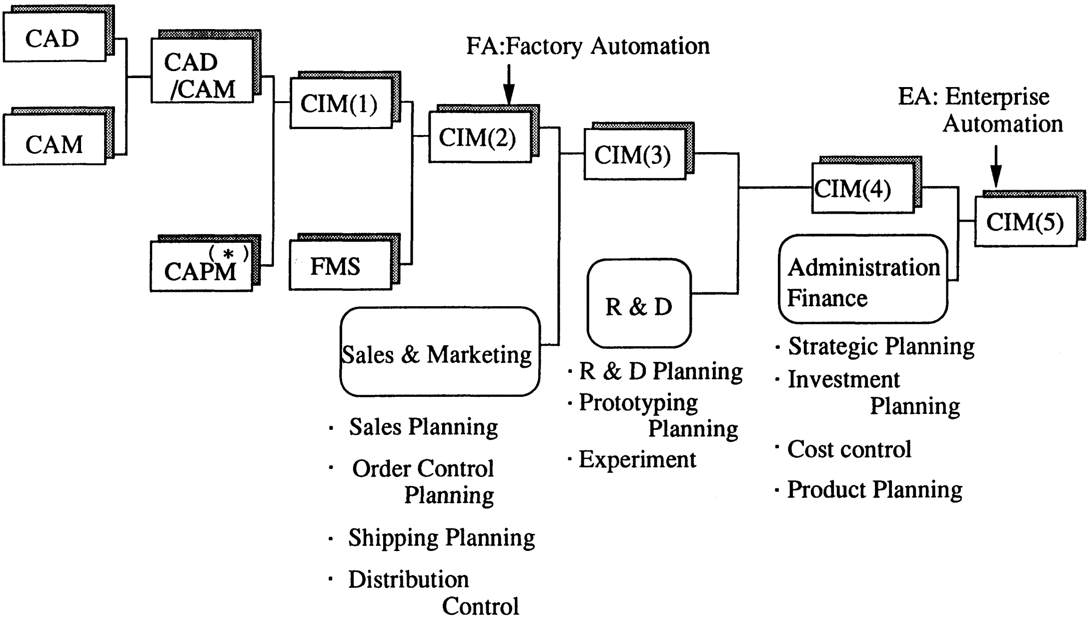
CAD
CAM
CAD /CAM
(*) CAPM
CIM(1)
FMS
FA:Factory Automation
CIM(2)
CIM(3)
Sales & Marketing
· Sales Planning
· Order Control Planning
· Shipping Planning
· Distribution Control
R & D
· R & D Planning
· Prototyping Planning
· Experiment
CIM(4)
Administration Finance
· Strategic Planning
· Investment Planning
· Cost control
· Product Planning
EA: Enterprise Automation
CIM(5)

(*) CAPM: Computer-Aided Production Management

FIG. 9. Concept of CIM on five levels.

Besides CIM,

1. concurrent engineering or simultaneous engineering in which CAD and CAM are integrated and
2. integration of CAM and scheduling

are the noteworthy topics in relation to the CAM integration.

8. SUMMARY

CAM involves two aspects of computer application in manufacturing. One is the computer control of machines to automate the production processes, which originated in the development of NC machines. The other is the computer processing of production data, which involves computer-aided process planning, computer-aided operation planning, and production control.

Nowadays CAM is widely realized in most areas of manufacturing, and the level of automation is becoming more and more sophisticated. Along with this tendency, expert systems that can accumulate and utilize human knowledge and experiences relevant to CAM were recognized as a practical approach and have been rapidly developed in recent years. Automated operation of the factory has also made progress resulting in the automated operation of the 66 h from Friday evening to Monday morning. In addition to the development of each CAM technology, the integration of various manufacturing activities such as design, management, marketing, etc., has attracted strong attention. The trend is now moving from FA to CIM. CIM is, and will be, a key technology of manufacturing in the 1990s and will develop further, moving toward "intelligent CIM."

GLOSSARY

CAD: Computer-aided design.

CAM: Computer-aided manufacturing.

CAPM: Computer-aided production management.

CAPP: Computer-aided process planning.

CIM: Computer integrated manufacturing.

CNC: Computer numerical control.

DNC: Direct numerical control.

FA: Factory automation.

FMS: Flexible manufacturing system.

GT: Group technology.

IC: Integrated circuit.

Manufacturing: A series of interrelated activities and operations involving the design, materials selection, planning, production, quality assurance, management, and marketing of parts or products.

NC: Numerical control.

Operation: A defined action, i.e., the act of obtaining a result from one or more elements.

Operation Planning: Planning of operations based on the plan prepared by the process planning.

Process: A systematic sequence of operations to produce a specified result.

Process Planning: Planning of processes for the production of a part or product on the basis of the information and drawings generated in the design stage.

Production: Fabrication and assembly of a part or product.

Production Control: The function of directing or regulating the ordinary movement of parts or products through the entire production process from the requisitioning of raw materials to the delivery of the finished products.

List of Works Cited

Budde, W., Weissweiler, H. (1973), in: *Proceedings of the 14th International Machine Tool Design and Research Conference, Manchester, U. K.*, London: Macmillan, Paper 404, pp. 1–13.

CAM-I (1976), *Automated Process Planning System (CAPP), Systems Manual*, Version 2, Release 1. Arlington, TX: Cam-I, Inc.

CAM-I Standards Committee (1978), *Glossary of Computer Aided Manufacturing Terms*, R-78-SC-01. Arlington, TX: Cam-I, Inc.

Chryssolouris, G., Chan, S. (1985), *Ann. CIRP* **34**, 413–417.

Descotte, Y., Latombe, J. C. (1985), *Artif. Intelig.* **27**, 183–217.

Eversheim, W., Fuchs, H. (1980) *Advanced Manufacturing Technology*, Amsterdam: North-Holland, pp. 303–313.

Iwata, K., Kakino, Y., Ohba, F. (1977), in: *CIRP International Seminar on Manufacturing Systems, 6th, Cranfield, England*, Paris: Collége Internationale de Recherches pour la Production, pp. 25–40.

Iwata, K., Sugimura, N. (1984), in: *CIRP International Seminar on Manufacturing Systems, 16th Tokyo, Japan*, Paris: Collége Internationale de Recherches pour la Production, pp. 83–92.

Iwata, K., Fukuda, Y. (1987), in: *CIRP International Seminar on Manufacturing Systems, 19th, University Park, Pennsylvania*, Paris: Collége Internationale de Recherches pour la Produciton, pp. 287–294.

Iwata, K., Fukuda, Y. (1989), in: *CIRP International Workshop on Computer Aided Process Planning, Hannover, Germany*, Paris: Collége Internationale de Recherches pour la Production, pp. 73–83.

Japan Society for Precision Engineering (Ed.) (1991), *Manufacturing Software System*,Tokyo: Ohm, Inc. (in Japanese).

Mann, W. S. (1977), *A Computer Process Planning Technology*, R76-942625-14, East Hartford, CT: United Technology Co.

Matsushima, K., Okada, N., Sata, T. (1982), *Ann. CIRP* **31**, 329–332.

Nissen, K. (1969), in: *Proceedings of the CIRP International Conference on Applications of Computers in Manufacturing*, Paris: Collége Internationale de Recherches pour la Production, pp. 47–52.

Pennell, J. P., Winner, R. I. (1989), in: *Globecom '89. IEEE Global Telecommunications Conference & Exhibition. Conference Record*, New York: IEEE, Vol. 1, pp. 647–655.

Spur, G., Anger, H.-M., Kunzendorf, W., Stuckmann, G. (1978), in: Davies, B. J. (Ed.), *Proceedings of the Nineteenth International Machine Tool Design and Research Conference held in Manchester, 13th–15th September, 1978*, London: Macmillan.

TNO (1981), *Introduction to MIPLAN*, Appeldoorn, The Netherlands: Organization for Industrial Research, Inc.

Van't Erve, A. H., Kals, H. J. J. (1986), *Ann. CIRP* **35**, 325–329.

Further Reading

Ham, I., Lu, C.-Y. (1988), *Ann. CIRP* **37**, 591–601. (An overview of the current status and future trend in an area of CAPP.)

Japan Society for Precision Engineering (Ed.) (1991), *Manufacturing Software System*, Tokyo: Ohm Inc. (in Japanese). (A detailed text on computer aided manufacturing software.)

Rembold, U., Blume, C., Dillmann, R. (1985), *Computer-Integrated Manufacturing Technology and Systems*, New York and Basel; Marcel Dekker. (A detailed text on computer automation in manufacturing.)

Zozaya-Gorostiza, C., Hendrickson, C., Rehak, D. R. (1989), *Knowledge-Based Process Planning for Construction and Manufacturing*, San Diego; Academic. (An overview of expert-type CAPP systems and their research project.)

COMPUTER DATABASES

H. BEHRENS AND J. W. TEPEL *Fachinformationszentrum Karlsruhe, Eggenstein-Leopoldshafen, Germany*

	Introduction	103
1.	**Data Models and File Design**	104
1.1	Data Storage	104
1.2	Data Models	105
1.2.1	Relational Model	105
1.2.2	Network Data Model	106
1.2.3	Hierarchical Model	106
1.3	File Design	107
2.	**Database Management Systems and Their Components**	108
2.1	Dedicated Databases	108
2.2	Structure of Database Management Systems	108
2.3	DBMSs for Personal Computers	109
2.4	DBMSs for Mainframes	110
2.5	Choosing a DBMS	110
3.	**Information Retrieval and Hypertext**	110
3.1	Information Retrieval Systems (IRS)	110
3.2	Hypertext	111
3.3	Information Retrieval Models	111
4.	**Bibliographic Databases**	112
5.	**Factual Databases**	114
6.	**Numeric Databases**	115
7.	**On-line Services**	121
8.	**Other Data Distribution Channels**	123
	Glossary	123
	List of Works Cited	125
	Further Reading	125
	Registered Trademarks	126

INTRODUCTION

Data of every kind (citations, facts or numeric data) have been collected and handled since the beginning of the scientific era. Until a few decades ago, this handling was carried out by pencil, by ink, or in printed form, but all on paper. Thus we can speak of the paper age in science. In physics, data were published in journal articles, in monographs, in textbooks, and in data compilations. In addition, citations were already regularly compiled and published in abstract journals in order to facilitate the access of physicists to literature. Good examples were *Physikalische Berichte,* founded as *Fortschritte der Physik* in 1845, and *Physics Abstracts,* founded as *Science Abstracts* in 1898.

As far as numerical data are concerned, major efforts have been made to compile and evaluate these in handbooks. One of the best examples is Landolt–Börnstein, *Numerical Data and Functional Relationships,* of which the first volume was published in 1883. However, even in this paper age the physics community was confronted (as also holds true for other science communities) with an exponential growth of the literature and data, a fact that was, however, not realized by the community because the absolute numbers were not so large as they are today.

After the second World War, at the end of the 1940s and the beginning of the 1950s, the development of computers began (see COMPUTERS). In the middle of the 1950s we had the first mainframes, in the middle of the 1960s the first minicomputers, and in the middle of the 1970s the first microcomputers. In parallel, new supercomputers became available from the middle of the 1970s onwards. This computer development had two effects. On the one hand, it speeded up the increase in the amount of data and information generated, by some kind of feedback effect (e.g., because of the automatization of the measurement process), and, on the other hand, the capacity to handle and store numeric as well as alphanumeric information also increased. For the first time large amounts of data and information could be handed easily and inexpen-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

sively. Thus the electronic revolution offered the possibility to overcome at least partly the problems connected with the data and information explosion.

In the electronic age we now have data in electronic form or, as it is normally said, in machine-readable form. Usually a collection of electronically stored items is called a database. These electronically stored items can be citations, facts, numerical data, or even the full text of an article. Accordingly, one distinguishes between ***bibliographic, factual, numerical***, and *full-text* databases. The bibliographic databases contain only references (titles, authors, abstracts, sources, etc.) together with other items for the corresponding subject analysis (classification, descriptors, etc.) The factual databases contain other structured but mainly non-numeric data. The numerical databases are defined by the fact that their content is dominated by structured numeric data. The full-text databases contain complete and unshortened texts from articles or other written papers. Next-generation databases will be bigger and more complex, containing a mixture of the above types of data, and will be used extensively for decision support purposes (Silberschatz *et al.*, 1990).

Let us now come back to the computers themselves: What are their advantages? They are ideal for handling data insofar as they are able to store, manipulate, display, and retrieve large amounts of data quickly. Even printed versions can be produced with very little human intervention, so that the transfer of data from the electronic form into the old-fashioned paper form is greatly facilitated. Another advantage of the computer technology development is that the costs of storage and access are constantly decreasing, that we can more easily replace old data by new data, i.e., it becomes easier to update our data collections, and that we can search more specifically for data in a database than was possible in the printed works via the indexes.

As was to be expected, the introduction of computers has also led to unwanted and less favorable developments. In the paper age each scientist and engineer interacted directly with each data item. He had to select the data and make all the necessary manipulations himself. Therefore the contact of each scientist and engineer with his data was very close. The introduction of computers had the consequence of a certain decoupling between humans and the data, which in some instances has led to inconsistencies, ambiguities and a loss of overall quality. On the other hand, whenever computational cross-checks can be made, typical human errors have been all but eliminated.

1. DATA MODELS AND FILE DESIGN

1.1 Data Storage

A computer is an excellent instrument for storing information. Even the simple $5\frac{1}{4}$-in. magnetic floppy disks, which have been around for many years, manage to store a respectable 360 kB (B=byte) of data. This represents about 140 typewritten pages. The $3\frac{1}{2}$-in. high-density disks popular with IBM-compatible personal computers store 1.46 MB, or as many as 580 pages. Hard disks with storage capacities of up to 300 MB are readily available for microcomputers. Optical disks or CD-ROMs (compact-disks read-only memory) have typical capacities of 550 MB.

With the advent of such cheap mass storage devices, storage of data is no longer a problem. Also, the user does not have to worry about the low-level physical organization of the data on disk, since this is taken care of by the operating systems. Care has to be taken, however, to preserve any relationships between the data items. A telephone directory is not created on a computer by storing at random names and telephone numbers. At the very least, one has to know which number "belongs" to which person. In addition, of what use is any data collection if the data needed cannot be retrieved? It is this aspect of computer-based data collections or databases that has become of increasing importance with the advent of the CD-ROM, needing special retrieval and indexing techniques in order to guarantee acceptable retrieval times.

Small data collections can easily be handled by modern text editors or word-processing systems. Most of these software packages are extremely flexible, they provide for rapid searching of text passages, they provide a nicely formatted output, and they have many other useful features. Managing, for example, a collection of 50 or so cooking recipes does not warrant the expense of a database management system. Any relationships among the data items can be recognized by just "looking" at the data, i.e., reading the recipe.

1.2 Data Models

More extensive data collections have to be structured in some way and cannot be retrieved by just "browsing" through them. The starting point in designing a database is an informal model of the real world, the *entity–relationship* (ER) model. This starts off by listing all relevant "objects" or *entities* and the *relationships* among them. An entity is a thing that exists and is distinguishable, such as a PERSON or an AUTOMOBILE. An entity has *attributes*, e.g., a PERSON has a name, age, address, etc., each of which in turn has a certain value from a *domain* of values for that attribute in the whole set of entities. The domain of the attribute AGE belonging to the entity PERSON is the set of integer numbers in the range 0 to 120—assuming that no person is older than 120 years. An attribute or a set of attributes that uniquely defines each entity in a set of entities is called a *key* for that entity set. A relationship among entity sets is simply an ordered list of entity sets. The list may consist of only two members, e.g., PERSONS and AUTOMOBILES with the relationship "has a" or "belongs to."

Entities and their relationships are best shown in the form of a diagram, the ER diagram (Fig. 1). Entity sets are represented by rectangles, their attributes by circles, and the relationships by diamond-shaped diagrams. Note that relationships may be of the kind one to one (1:1), many to one (*n*:1) or many to many (*m*:*n*). In our case we have a *n*:1 relationship between AUTOMOBILES and PERSONS when we assume that a person may own more than one car. The *n*:1 relationship is shown by the arrow in the diagram. More precisely, the arrow stands for "determines uniquely." A *m*:*n* relationship would exist if an AUTOMOBILE could in addition have more than one owner. More details on ER diagrams may be found in Ullman (1982).

Closely connected to the ER model are "data models," three of which, viz., the *relational*, *hierarchical*, and *network* models, have been used in practically all commercially available database systems (DBSs). These data models can be transcribed more directly into the physical structures and organization needed to make a DBS work on a computer.

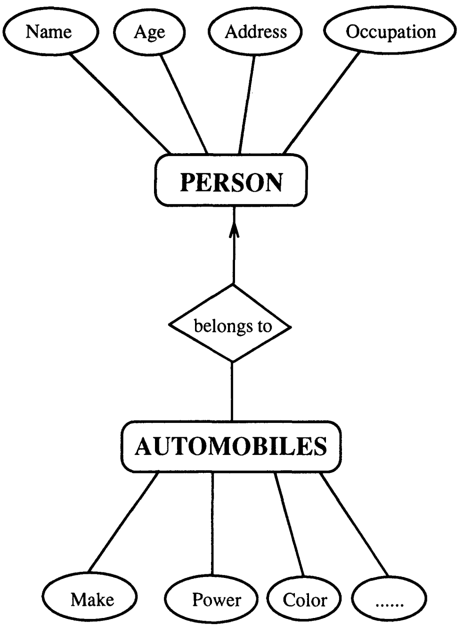

FIG. 1. Entity–relationship model of PERSONS and AUTOMOBILES. Objects (rectangles) together with their attributes (ellipses) are linked via undirected graphs (lines). The diamond-shaped relationship "belongs to" between AUTOMOBILES and PERSONS is of the "many-one" kind (*n*:1). This is expressed by the arrow into PERSON.

1.2.1 Relational Model.

A relation is any subset of the Cartesian product $D_1 \times D_2 \times \cdots \times D_k$ of a list of domains $D_1, D_2, \cdots, D_k$. The members $(v_1, v_2, \cdots, v_k)$ of a relation, where v_i is a value from domain D_i, are called *tuples*. An example illustrates this more clearly. Let us assume that the entity PERSONS has the attributes NAME, AGE, and TELEPHONE. Let us assume that NAME consists of only two values, Smith and Jones, i.e., the domain $D_1 =$ (Smith, Jones). Simplifying in a similar way, the domains for AGE and TELEPHONE might be $D_2 = (10,24,60)$ and $D_3 = (456\,789,\ 234\,567)$. The Cartesian product of the domains D_1 to D_3 or $D_1 \times D_2 \times D_3$ is given by the twelve 3-tuples (v_1, v_2, v_3):

(Smith, 10, 456 789), (Jones, 10, 456 789),
(Smith, 10, 234 567), (Jones, 10, 234 567),
(Smith, 24, 456 789), (Jones, 24, 456 789),
(Smith, 24, 234 567), (Jones, 24, 234 567),
(Smith, 60, 456 789), (Jones, 60, 456 789),
(Smith, 60, 234 567), (Jones, 60, 234 567).

Any subset of this set of 3-tuples including the set that might correspond to the real world,

(Smith, 24, 456 789),
(Jones, 60, 234 567)

is a relation of degree or arity 3. The relation scheme is written PERSONS (NAME, AGE, TELEPHONE). It helps to view relations simply as tables of values of given attributes, where the order of the attributes is unimportant.

The relational model may be used to construct a DBS for the entities PERSONS and AUTOMOBILES as follows. A relation scheme describing AUTOMOBILES might by written as AUTOMOBILES (AKEY, TYPE, COLOR, SERIAL__NO). Here AKEY serves to uniquely identify an automobile. It might consist of the letter *A* followed by a running number, e.g., A001, A002, ..., A376. Such a quantity plays the role of a "key" for accessing any particular automobile. To the PERSONS relation scheme we would add a unique person key PKEY consisting of, say, P001, P002, etc. Then the relationship between PERSONS and AUTOMOBILES is described by HAS__A__CAR (PKEY, AKEY), i.e., by tuples of the kind (P001, A002), (P001, A009), (P002, A001), etc., where we have assumed that person P001 owns two cars, A002 and A009. The three relations PERSONS(···), HAS__A__CAR(···), and AUTOMOBILES(···) suffice to describe the ER diagram in Fig. 1.

Note that by introducing a relation HAS__A__CAR(···) to describe the relationship between PERSONS and AUTOMOBILES, we can in effect handle *m*:*n* combinations: A person need not have a single car, nor is it necessary for a car to have one owner only. Restricting ourselves to *n*:1 relationships between automobiles and owners makes the relation HAS__A__CAR unnecessary: Augmenting the scheme AUTOMOBILES(···) with PKEY clearly defines the ownership of a car.

Relational database systems (RDBSs) have many advantages, and for this reason commercial systems for microcomputers are almost exclusively of this type. Data can be easily changed or updated. Deleting a car from the DBS would involve deleting a record from AUTOMOBILES without affecting the PERSONS file. Furthermore, the theory of RDBSs is well advanced and other data models can be reduced to the relational model at the cost of data storage (see below).

1.2.2 Network Data Model. This is the most general form in which data and their relationships can occur. Many real-life situations are of the network type. In order to avoid confusion with the relational model, it is best to talk here about logical records instead of tuples and logical record formats containing fields instead of relational schemes (Ullman, 1982). Instead of *n*:1 relationships one uses links connecting different record types. Figure 2 shows network data at the record level. The corresponding ER diagram (not shown) consists simply of the entities PARENTS, CHILDREN, and TOYS joined by undirected links. Parent records (P1,P2,...), children records (C1,C2,C3), and toy records are linked by pointers (not shown in the figure). Toy T4 is shared between children C2 and C3, whereas T5 is also owned by parent P4. Data are found by navigating along the links and not by performing relational algebra. In order to simplify implementation of networks on a computer, *m*:*n* relations are replaced by *m*:1 and 1:*n* relations by introducing intermediate record types containing the respective link information.

1.2.3 Hierarchical Model. The hierarchical model can be considered to be a special case of a network. Bidirectional links are replaced by parent-children 1:*n* links. The data structure looks like an inverted tree (Fig. 3), where the lower nodes and leaves can be reached only via the root. A linear physical record representing the hierarchical structure might look as shown in Fig. 4. The child and toy data are contained in a repeating group of fields called a periodic group (PG). In this case three such groups occur, i.e., the periodic

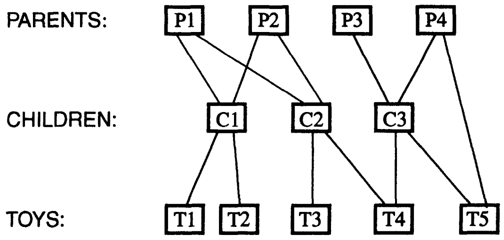

FIG. 2. Parent–child–toy diagram as an example of a network. The record types PARENTS, CHILDREN, and TOYS are linked by pointers. Note that toy T5 is owned by parent P4 and child C3. Complex relationships can be represented in a network diagram.

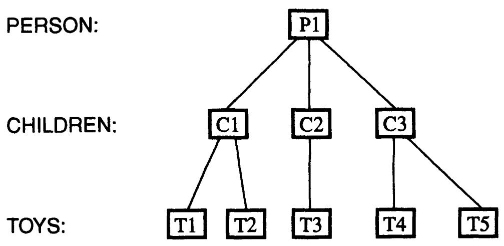
PERSON:
P1
CHILDREN:
C1
C2
C3
TOYS:
T1
T2
T3
T4
T5

FIG. 3. Person–child–toy diagram as hierarchical structure.

group multiplier Pm is 3, since there is one instance of the group for each child. PG1 consists of data belonging to child C1 together with its toys T1 and T2. The toy multiplier Tm is needed to tell the computer the number of toy names that follow. Note that updating a record of this type is much more complex than in the relational case. Deletion of a node in Fig. 3 (say C1) eliminates the access paths (physical links) to the lower levels of the tree. Removing a child node from the tree also eliminates its toys. This can be considered to be an advantage or disadvantage depending on whether or not one argues that a toy is an entity even when not linked to its child owner.

Both the network and hierarchical data models can be transformed to the relational case. Linkages are replaced by relations. The hierarchical data in Fig. 4 can be represented by the relational schemes MOTHER(MKEY, ...), CHILDREN(CKEY, MKEY, ...), and TOYS(CKEY, ...). Here the three levels of hierarchy are linked by the unique keys CKEY and MKEY, identifying tuples in the CHILDREN and MOTHER relations, respectively.

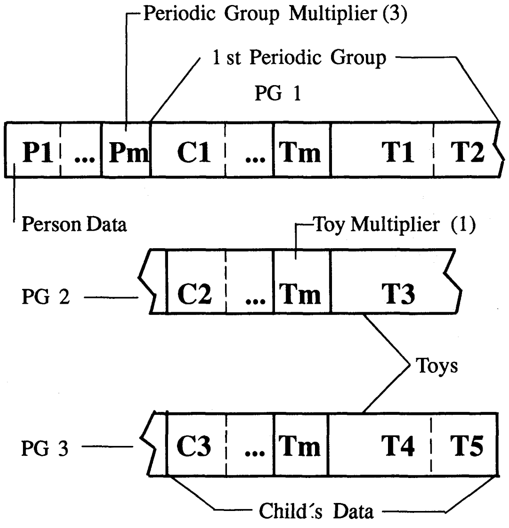
Periodic Group Multiplier (3)
1 st Periodic Group
PG 1
P1
...
Pm
C1
...
Tm
T1
T2
Person Data
Toy Multiplier (1)
PG 2
C2
...
Tm
T3
Toys
PG 3
C3
...
Tm
T4
T5
Child's Data

FIG. 4. Data of Fig. 3 arranged as a linear record (ADABAS).

1.3 File Design

Designing a database is an iterative process consisting of a number of steps shown in Fig. 5 (see Lockemann and Schmidt, 1987). The first step involves an analysis of information requirements in the broadest sense. What data are to be stored, who has access, and what information has to be provided by the DBS are some of the topics that have to be clarified. The *conceptual design* involves specification of objects and their relationships as in the entity–relationship model. At the *logical design* level, the data model most suited for the problem at hand is chosen, and the record and field formats for each of the files comprising the database are specified. The type and length of fields needed and the questions that have to be answered will determine to a large degree the database management system (DBMS) that can be used. The nature of a DBMS is explained in Sec. 2.2.

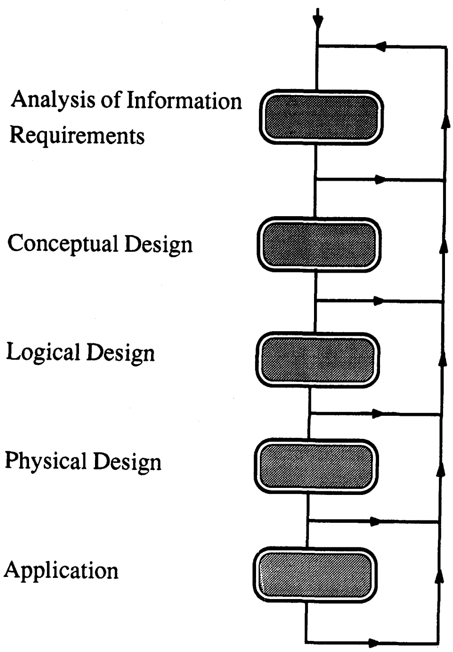
Analysis of Information Requirements
Conceptual Design
Logical Design
Physical Design
Application

FIG. 5. Steps in designing a database.

The *physical design* is taken care of by the manufacturer of the DBMS—unless the whole system is to be developed from scratch, something that was only done for the older physics databases, where dedicated systems written in FORTRAN or PL/1 were used. Usually the data are organized sequentially in a direct access file with pointers to the organizational units of the data file (documents, fields) stored in one or more index files. Access to the data is usually by one or more keys constructed by concatenating the contents of some data fields. Lookup of these keys in the index files may be by use of a hashing algorithm, by an index sequential access method (ISAM) or by use of Bayer trees (BTREE/ISAM techniques) (see Bayer and McCreight, 1972); these terms are defined in the Glossary. Physical data organization is described in detail in the book by Ullman (1982).

The *applications* are the real test for the DBS. At this stage the gap between planning and reality becomes evident, and, depending on how well the database has been modeled to fulfill the overall requirements, some modifications will be necessary at the higher design stages of Fig. 5.

2. DATABASE MANAGEMENT SYSTEMS AND THEIR COMPONENTS

2.1 Dedicated Databases

Even today, most DBSs in use in physics are dedicated systems developed for some specific purpose and are usually written in a high-level language such as FORTRAN or PL/1 (see COMPUTER PROGRAMMING LANGUAGES). Many databases originated at a time when commercial DBMSs were either nonexistent or too expensive, and it is only natural that physicists used the programming language that they knew best and that they had been using for their numerical computations, FORTRAN. Systems written in PL/1 having more sophisticated input/output (I/O) routines and indexed access to the data emerged in the 1970s. Modern programming languages such as C are little used up to now in the physics community.

Dedicated DBSs can be very efficient as regards performance and the need for data storage. Their main drawback is their inability to adapt to changing user requirements. A new file design means major changes to the software. Updates are commonly done by generating the files anew. Developing a dedicated database or retrieval system is worthwhile only in very few isolated instances (Tepel and Müller, 1990).

2.2 Structure of Database Management Systems

A database management system (DBMS) is simply a database system (DBS) with additional management facilities. Some of these features are listed below (see Ross, 1978).

1. *Database security.* Provision has to be made to control user access to the data. This is achieved by introducing different external or user views of the data (in effect a filter allowing users to see only certain data fields), by disallowing certain commands that modify the data, by password protection, and by encoding of the data.
2. *Database integrity.* Routines have to be provided to allow backups to be made of the DBS. Larger DBMSs have transaction logging, checkpointing, and capabilities for restart after system failures.
3. *Data definition language (DDL).* The DDL provides a means of describing the data structures and maps them onto the storage media. As an example, Fig. 6 shows the ADABAS™ DDL module for Fig. 4. Low-

```
01,NA,20,A,DE          **Person's Name
01,AG,2,U,DE           **Age
01,AD,60,A             **Address
01,TL,12,U             **Telephone No
01,CH,PE               **Periodic Group
      02,CN,20,A,DE    **Child's Name
      02,CA,2,U        **Age
      02,TY,30,A,MU,DE **Toys
```

A = Alphanumeric Field DE = Descriptor Field
U = Numeric Field MU = Multiple Values

FIG. 6. ADABAS™ data definition module for the data of Figs. 3 and 4. Each entry consists of a level identifier (01 or 02), the field name, field length and type, and additional information such as whether the field is to be indexed (DE) or not.

level routines for disk access and file management are usually taken from the operating system.

4. *Data dictionary.* The data dictionary or database system catalog provides the user or database administrator with an overview of the DBS. It lists the files and data structures (tables and attributes) present in the system. Data definition modules appear in the data dictionary, where they are kept for future reference and modifications.
5. *Data acquisition.* Data input routines in batch mode and via masks depending on the data structures as described by the DDL are usually provided.
6. *Data manipulation language (DML).* The DML is the command language of the DBS and allows operations on the database such as the creation or deletion of whole files or groups of files and the start and termination of user sessions as well as manipulations on the data itself, such as deletion, insertion, and updating of individual logical fields or records. The query language used for retrieving data is often considered to be a subset of the DML. Database producers often call their DMLs programming languages of the fourth generation, since they provide direct database support with commands such as FIND, ADD, and DELETE. In addition, most commercial DBMSs have interfaces to high-level programming languages such as PL/1 or C, and to Assembler. Well-known query languages are SQL, which is a de facto standard for relational database systems, and NATURAL™, which is used together with the DBMS ADABAS™ on mainframes.
7. *Report generator.* Most DBMSs include a report-generating facility that presents retrieved data as properly formatted reports. Personal computer (PC) systems often provide advanced features such as graphics capabilities and support for a variety of printers including systems using PostScript™.
8. *Support for multiple access.* An important feature of a DBS is the centralized storage of data. In addition to reducing data redundancy, this consolidates updating procedures, since only the master copy of the file has to be maintained. It is preferable to use a single database under a single control rather than having scattered collections of computer files maintained by individual groups or users. Multiple and simultaneous access to such a centralized system is important, and the DBMS software has to preserve the integrity of the data and prevent deadlock situations.

2.3 DBMSs for Personal Computers

The three types of DBMSs for use on PCs are the flat-file, relational, and network or hierarchical types. Of these, the relational database management systems (RDBMSs) are by far the most common.

Flat-file systems consist of linear arrays of data and permit no logical file couplings. Each database consists of a single file only. Records are usually subdivided into fields of fixed length, a certain number of which may be indexed.

Data acquisition and report generation are supported. These systems are inexpensive, are often available as public-domain or user-supported software, and are ideal for keeping simple lists of data. Through their menu-driven user interfaces they are very user-friendly.

A few professional network-type DBMSs exist for special applications on personal computers. An overview may be found in Seymour (1988).

Relational DBMSs for PCs are very common. Very widespread is dBASE™ by Ashton–Tate. Although more a DBS than a DBMS, it has a powerful programming language and is supported by a large number of software firms. Procedures can be rapidly developed and tested, since dBASE runs as an interpreter. These procedures can then be compiled with dBASE compilers, such as CLIPPER™, which produce executable files. dBASE relies heavily on the PC operating systems for file management.

An example of a full relational DBMS for PCs is ORACLE™. ORACLE was originally developed for minicomputers, but nowadays the whole spectrum of computer hardware is supported, from microcomputer and workstations to large mainframes. It uses SQL as query language (Date, 1987) and has interfaces to several high-level programming languages, in particular to C. Data management is reminiscent of mainframe systems, with password protection, clustering of files in definite partitions (areas on the hard disk), rollback and commit statements (*rollback* makes the changes to data at the end of a session undone, *commit* confirms the changes), and provision for a systems supervisor—the DBA or database administrator.

2.4 DBMSs for Mainframes

Most of the older types of DBMSs for mainframes are of the network or hierarchical type, e.g., Information Management System (IMS) of IBM. In recent years, relational DBSs have become more important. The advantages of such RDBMSs are easier maintenance of individual files and flexibility with respect to user queries. The data structure and linkages in a hierarchical system determine what type of questions can or cannot be answered, whereas the linear file design of relational systems allows future applications, which were not foreseen at the time of the file design. Possible drawbacks of RDBMSs, viz., the need to compute joins between many different relations with a corresponding loss in performance and the need to construct displays from relations stored on different physical regions of the hard storage devices, are partly offset in commercial systems such as ADABAS by allowing some types of repeating fields (multiple fields) and repeating groups (periodic groups) in each record of a file. In this way, retrieving a record for display purposes usually places all data at one's disposal, while still allowing indexing on individual values of multiple fields, such as author names in a bibliographic document.

The DBS used in the Messenger retrieval system of STN-International, FIDO, is a hierarchical system allowing linkage of terms on several different levels (word, sentence, paragraph, and linkage proximity). However, the requirements of a DBS for information retrieval systems used mainly for bibliographic and full-text databases (see below) are somewhat different from those of other DBSs.

2.5 Choosing a DBMS

Choosing a DBMS for a proposed application is not an easy task and needs a good understanding of the underlying principles of electronic data processing and data modeling. Unfortunately, often a choice has to be made before the scope of the application or the abilities of the DBMS software are known completely. Changing subsequently from one DBMS system to another can be very costly.

There is probably no single system that satisfies all the user requirements. A DBMS that works well with some type of data, e.g., records of fixed length subdivided into fields, performs very poorly or is unable to handle data of varying length. A number of criteria are helpful in determining which system is best suited for the particular purpose:

1. Structure of the data—free text, free text with some structure (e.g., a book), documents from on-line retrieval sessions, data in fixed format.
2. Data types—text, numeric, graphics, or a combination of these.
3. Lifetime of data—static or changing data (frequent updates).
4. Size of the database.
5. Searchability of the data—free text search, full text search, search by means of indexed fields or inverted lists.
6. Data relationships (1:1, n:1, n:m) and models.
7. User interfaces—what provision is made in the software package for user interaction with the data.

3. INFORMATION RETRIEVAL AND HYPERTEXT

3.1 Information Retrieval Systems (IRS)

Systems in use with bibliographic databases fall into the category IRS rather than DBMS. Of primary importance are the user interface and the retrieval possibilities for the documents. Less emphasis is placed on the DBM part of IRSs. The data are more static than with classical DBS, and much effort is spent in building index files for efficient data retrieval. The index parts of such files are normally larger than the actual data. Updates to older data are rarely done; files change because new documents are added at regular intervals. In addition, IRSs usually support structured thesauri containing hierarchical and associative relationships between terms (keywords), which makes document retrieval very effective. These terms are usually called descriptors and are classified into superordinate or broader terms, subordinate or narrower terms, and associative or related terms.

3.2 Hypertext

A comparatively recent development, hypertext combines the virtues of full-text DBSs with user-defined access paths to the data (Marchionini and Shneiderman, 1988; Franklin, 1989). In hypertext the underlying DB structure permits the user to traverse complex networks of information quickly. The user can select his own topics of interest, link these with related concepts, and navigate on-screen through the full text by choosing at each stage among a set of alternatives. The ability to browse through the text using a filter of selected topics is characteristic of hypertext systems.

3.3 Information Retrieval Models

In order to retrieve documents from an IRS the user formulates a query in the query language of the particular system. The goal of such a query is to maximize the *recall* of documents, defined as the ratio (number of relevant items retrieved)/(number of relevant items in the DB), while keeping the *precision*, defined as (number of relevant items retrieved)/(total number of retrieved items), as large as possible. Broad query formulations tend to have a high recall but a low precision, whereas narrow formulations have high precision but low recall figures.

IRSs used by large on-line hosts employ the so-called Boolean retrieval model. A search query is formulated and all documents are retrieved that exactly match this query. The query may contain terms connected by the Boolean operators AND, OR, and NOT, as in the following example from the database PHYS (comments have been added within the delimiters /*....*/):

```
/* user entries are expected at the system prompt ⇒  */

⇒ s database                          /* search for the term database  */
L1      527  DATABASE                 /* found 527 documents  */
     /* query results are labeled consecutively by L1,L2,...  */
⇒s information retrieval              /* find information retrieval  */
     37241  INFORMATION
      1039  RETRIEVAL
L2      387  INFORMATION RETRIEVAL    /* 387 documents found  */
/* search for documents which contain both the terms DATABASE and
   INFORMATION RETRIEVAL */
⇒s      L1 AND L2
L3      31 L1 AND L2
⇒ s  L3 NOT optical                   /* of the 31 documents found
                                         previously show only
                                         those that do not contain
                                         the word OPTICAL  */
   125284 OPTICAL                     /* 125284 occurrences
                                         of OPTICAL in DB  */
L4      29 L3 NOT OPTICAL             /* final result of combined
                                         queries  */
```

Here documents are retrieved that contain the word "database" and "information retrieval," but not the term "optical." One of the documents found is displayed in Fig. 7.

Apart from the Boolean model, where an exact match between terms in the query and in the documents is required, other retrieval models such as the probabilistic model and the vector-space model relax this condition (see Salton, 1989). Here a measure of similarity is introduced between the terms in the query formulation and in the documents. Documents may be ranked according to the degree of overlap with the query. Recall is improved by these methods, and users can decide by themselves at which point in the

```
L4   ANSWER 2 OF 29
AN   10(14):66440  PHYS
TI   Data structures and organisation: Special problems
     in scientific applications.
AU   Read, B.J. (Rutherford Appleton Lab., Chilton
     (UK))
SO   Comput. Phys. Commun. (Dec 1989) v. 57(1-3) p.
     164-168
     International Conference on Computing in High Energy
     Physics, Oxford
     (UK), 10-14 Apr 1989
     ISSN 0010-4655; CODEN CPHCB
CY   NETHERLANDS
DT   Journal; Conference
TC   Computing
LA   English
AB   In this paper we discuss and offer answers to the
     following questions: What, really, are the benefits
     of databases in physics. Are scientific databases
     essentially different from conventional ones. What
     are the drawbacks of a commercial database manage-
     ment system for use with scientific data. Do they
     outweigh the advantages. Do database systems have
     adequate graphics facilities, or is a separate gra-
     phics package necessary. SQL as a standard language
     has deficiencies, but what are they for scientific
     data in particular. Indeed, is the relational model
     appropriate anyway. Or, should we turn to objecto-
     riented databases. (orig.)
CC   *0270
CT   *COMPUTER CALCULATIONS; COMPUTER CODES; COMPUTER
     GRAPHICS; PROGRAMMING LANGUAGES; *INFORMATION SYSTEMS;
     INFORMATION RETRIEVAL; DOCUMENTATION
```

FIG. 7. Example of a PHYS documentation unit.

ranked list they want to place their cutoff of useful documents. Little practical experience is available with these more advanced retrieval models.

4. BIBLIOGRAPHIC DATABASES

The databases, producers, and hosts mentioned in Secs. 4–7 are listed and described in the Cuadra directory of databases (Cuadra, 1990), which is updated regularly. For this reason, we have omitted the administrative details, such as addresses, from our overview.

As already mentioned in the Introduction, bibliographic databases are defined by the fact that they only contain references (titles, authors, sources, etc.) to published literature. Of course, there exist a very large number of small, mostly private files of this type. It is, however, not the intention to describe these in more detail, but to restrict ourselves to the large bibliographic databases covering whole fields of science, such as physics, chemistry, materials science, etc. They are produced professionally by large producers like the Institution of Electrical Engineers, Fachinformationszentrum Karlsruhe, Chemical Abstracts Service, etc., and are offered on line via so-called hosts.

Retrieving simply bibliographic citations is not enough for an on-line researcher. More is involved. Thus, in order to provide the prospective user with the details of the content of an original document and in order to make a specific and selective retrieval possible, each bibliographic item (documentation unit) also contains terms derived by what is known as subject analysis (see Fig. 8). This subject analysis, often also called indexing, is usually carried out manually by the document analysts. In some, very few, modern systems, parts of this analysis process (the indexing and adding other information) are nowadays also handled automatically (or semiautomatically) by application of suitable computer systems (automatic indexing and retrieval).

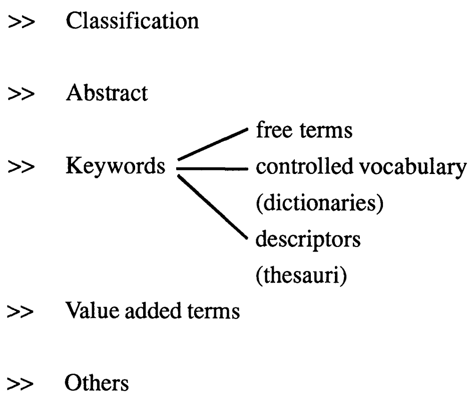

FIG. 8. The subject analysis of a document consists of adding the terms shown.

A normal documentation unit, which is the smallest unit of a bibliographic database and which contains all necessary information about an original document, consists of two parts, the technical bibliographical or subject cataloging part and the subject analysis part.

Documentation unit:
Technical bibliographical part

Field	Prefix
Accession number	AN
Authors, editors	AU
Title	TI
Source	SO
Journal title	JT
Corporate source or affiliation	CS
Publication year	PY
Publication date	PD
Meeting date	MD
Number of report	NR
Country of publication	CY
Language	LA
Document type	DT
Patent country	PC
Patent kind code	PK
Patent number	PN
Update code	UP

Documentation unit: Subject analysis part

Field	Prefix
Abstract	AB
Classification code	CC
Controlled terms (descriptors)	CT
Treatment code	TC
Value-added (e.g., element) terms	ET

A typical documentation unit making use of this scheme and taken from the database PHYS is shown in Fig. 7. Other information items, so-called value-added terms assigned in the process of subject analysis, are data flagging, chemical indexing with or without chemical structures, numerical indexing, astronomical objects, etc.

In physics, normally classifications are applied that are strongly related to the international ICSTI (International Council of Scientific and Technical Information) classification. It consists of a hierarchical scheme four levels deep of which the first two levels consist of 10 and 67 classes, respectively (see ICSTI, 1991).

For indexing free terms, controlled terms and descriptors from a thesaurus might be used. A thesaurus contains controlled terms only with additional hierarchical relationships in the form of so-called broader or narrower terms (see, e.g., Salton, 1983) and with so-called forbidden (synonyms) and other related terms. Typical examples in physics are the thesauri of the database INSPEC and of the database PHYS, which are described below.

Of the discipline-oriented bibliographic databases, those of major importance for the field of physics are the databases PHYS and INSPEC. Both have a long tradition because they already existed before the computer age in printed form as abstract journals.

PHYS originates from the *Physikalische Berichte*, which was founded in 1845 in Berlin under the title *Fortschritte der Physik.* In 1920 this abstract journal was renamed *Physikalische Berichte*; today its title is *Physics Briefs/Physikalische Berichte.*

In 1979 the abstract journal was computerized so that a database called PHYS today exists in parallel to the printed version with citations back to the publication year 1979. The producer since that time is the Fachinformationszentrum Karlsruhe or, briefly, FIZ Karlsruhe in Germany. Actualization takes place biweekly. In addition to the normal bibliographic terms such as title, authors with affiliations, sources, etc., this database contains document analysis terms such as abstracts, international physics classification codes, and descriptors from a thesaurus. For the field of astronomy the names of the astronomical objects are added as searchable key words. Searching for special substance designations is supported by a mixture of manual and automatic adding of substance names. The database PHYS covers the whole of physics, which includes elementary particles and fields, nuclear physics, atomic and molecular physics, classical areas of phenomenology, fluids, plasmas, discharges, condensed matter, cross-disciplinary physics, geophysics, astronomy, and astrophysics. As of September 1990, it contained about 1.4 million citations with an increase per year of about 125 000 entries.

INSPEC goes back to the abstract journal *Science Abstracts*, which was founded in 1898 in England. Very soon this abstract journal was divided into two parts, one for physics and one for electrical engineering, the physics part being named *Physics Abstracts*. At the end of the 1960s the production process for this abstract journal and other abstract journals was mechanized by the producer, the Institution of Electrical Engineers, with the consequence that a database was created in parallel to the printed version.

In the case of INSPEC, subject analysis is done by assigning classification codes (the classification used is similar to the ICSTI international classification) by indexing with a free and controlled vocabulary, and by chemical (for the designation of substances) and numerical indexing. The database covers the fields of electronics, computers, computing, electrical engineering, physics, and information technology. The physics part corresponds to the printed version *Physics Abstracts*. The total number of citations as of September 1990 was about 3.7 million (50%–60% physics). The updating is carried out biweekly, which results in an increase per year of about 270 000 citations.

In addition to those physics-specific databases that cover the field of physics completely, there are others that either are of a multidisciplinary character or focus on other fields like energy, chemistry, materials science, etc. A database of the first type is PASCAL, which covers fields such as chemistry, biology, engineering sciences, geoscience, medicine, and also physics (about 13.5%, without nuclear and elementary physics). The citations stored go back to the publication year 1973. PASCAL contains presently about 8 million citations with an increase per year of about 450 000. It is produced by l'Institut de l'Information Scientifique et Technique du Centre National de la Recherche Scientifique in Nancy, France. Titles are bilingual, French and original language; the abstracts are in French; and the descriptors are trilingual, French, English, and Spanish.

Other multidisciplinary databases are ENERGY [produced by an International Energy Agency (IEA) international cooperation], which covers energy research and technology, but contains also a large part of physics-relevant citations, and NTIS, which is devoted to reports on research and development sponsored by U.S. and non-U.S. governments. In this same context, the Science Citation Index (SCI or SCISEARCH), which contains references from approximately 3800 journals to worldwide literature across many disciplines, should also be mentioned. Discipline-oriented databases focusing on fields other than physics, but containing partly also physics-relevant citations, are CA (chemistry) produced by Chemical Abstracts Service in the U.S., INIS (nuclear research and technology) produced by an International Atomic Energy Agency (IAEA) international cooperation, METADEX (metallurgy and material sciences) produced by the Institute of Metals in London and by the American Society for Metals (ASM) in the U.S., MATH (mathematics) produced by FIZ Karlsruhe in Germany, and MathSci (mathematics) produced by the American Mathematical Society (for corresponding addresses see Cuadra, 1990).

What are significant quality criteria for bibliographic databases? First, the degree of in-depth subject cataloging and subject analysis, which varies from bibliographic database to bibliographic database, should be mentioned. Then, currency, completeness, and consistency are very important. The last points are self-evident since for researchers, for instance, timeliness is of major relevance because otherwise they will lose citations to the most recent publications. The same is true for completeness, because researchers must be sure not to miss citations to important papers in a well-defined area.

5. FACTUAL DATABASES

By definition, factual databases contain other structured but mainly non-numeric data. The factual databases are sometimes also alluded to as referral and textual-numeric databases. Conversely, factual and numeric databases together are sometimes called factual databases. The majority of their data consists of text data. This type of database refers to organizations, projects, experts, patents, products, conferences, etc. Typical examples are research databases, product databases, and conference databases.

Research databases contain information on research institutes and research projects (research in progress). The usual content of the file segment INSTITUTE is

- Institute (company) name and address
- Telecommunication numbers
- Classification code
- Director of the institution
- Staff (names)
- Description of the main research activities
- Facilities of the institute
- Accession numbers of projects carried out at the institute
- Update code

and of the file segment PROJECTS is

- Title of project
- Classification code
- Description of the project
- Performing organization
- Cooperating organizations
- Name of project leader
- Staff (names)
- Supporting organization
- Duration begin, date
- Duration end, date
- Number of contract
- Funding
- Update code

Typical examples relevant for physics are in Germany the databases VADEMECUM (information on institutes), INFOR (information on institutes and projects), and FORKAT (projects funded by the German Federal Ministry for Research and Technology); in Britain the database Current Research in Britain (descriptions of projects conducted at British universities); and in France the database Telelab (information on governmental and university laboratories).

Another prototypical example consists of the conference databases that refer to future meetings and conferences. The components of the record are the following:

- Accession number
- Meeting title
- Meeting organizer
- Meeting location
- Meeting date
- Meeting secretariat
- Classification code
- Status date

A representative example is the database CONF produced by the FIZ Karlsruhe in Germany. A record of this database is shown in Fig. 9. Perhaps it should be mentioned that many of the factual databases are based on questionnaire actions for the acquisition of the necessary information needed for the input. The examples shown before show this very clearly. It is here our intention not to give a complete collection of the existing factual databases (such a complete survey can be found in Cuadra, 1990), but only to present some representative examples and to explain typical features.

6. NUMERIC DATABASES

A database is called numeric when it consists mainly of structured numeric data. Before considering the databases in more detail let us look at the kind of data stored in such a database. In physics we are dealing with numeric values of observables. The data we have to consider usually has the following structure:

1. Measured value
2. Uncertainty
3. Units (SI)
4. Data type
5. Range of validity
6. Method of measurement
7. Combination with other quantities
8. Bibliographic citations

Generally speaking a *datum* can also be any parameter deduced with the help of a model or a theory. It should, however, be noted, that very often some of the fields listed above are empty.

The automatic techniques that are nowadays employed in experiments either in laboratories or outside, for instance in satellite experiments, often generate large amounts of data, sometimes more than can be handled afterwards. These raw or primary data are recorded continually or at specific intervals.

```
L2   ANSWER 12 OF 32
AN   PAS:48318  CONF        SN  K99805312
MT   International ANS Topical Fast Reactor Safety Meeting.
MO   American Nuclear Society (ANS, La Grange Park, IL (USA); Euro-
     pean Nuclear Society, Petit-Lancy (Switzerland); Internatio-
     nal Atomic Energy Agency (IAEA), Vienna (Austria). Internatio-
     nal Working Group on Fast Reactors
ML   Snowbird, UT (USA)
MD   19900812 - 19900816
NTE  Contact: ANS/Idaho Section  International Fast Reactor Safety
     Meeting  attn.: R.W. King  P.O. Box 2528  Idaho Falls, ID 83402
     USA
     Deadlines: for summaries (1500 words): 1 Jan 1990; for complete
     papers with abstracts: 4 May 1990
     Abstracts: (summaries) will not be published
CC   *10P Nuclear reactor technology
      13L Safety engineering
      10O Nuclear power plants
UT   Safety design criteria; public perception/implication of the
     Chernobyl accident; safety experiments/fast reactor testing and
     operational safety experience; designing for passive safety;
     instrumentation and control; safety codes and calculational
     techniques; consequences modeling of accidents/severe accidents/
     containment design and analysis; seismic and external event design
     and accomodation
STA  Entered          Revised
STD  19880519         19891016
```

AN	Accession Number	MT	Meeting Title
MO	Meeting Organizer	ML	Meeting Location
MD	Meeting Date	NTE	Notes
CC	Classification Codes	UT	Uncontrolled Terms
STA	Status	STAD	Status Date

FIG. 9. Retrieval example from CONF.

Later on they are analyzed and interpreted according to the physical quantity involved. Very often this methodology leads to databases containing large volumes of raw data. We shall, however, not consider these in the following because their usage is mostly restricted to the very few specialists carrying out the particular experiments. We shall also not discuss further how to build private numeric files and databases (this has been sketched in the sections before) or go into further details concerning the numerous small home-built data files in physics. As before, we shall only elucidate the publicly available larger databases that are produced in a professional and organized manner.

In physics we are mostly dealing with time- and location-independent data. Therefore, in the following we shall restrict ourselves to the handling and storage of this type of data. In certain subject areas, including astronomy, astrophysics, and geophysics, time series are important. However, that would be a topic of its own.

More often than not we have more than a single experimental result available for a physical quantity. Sometimes these measurements even contradict one another. It is therefore very important to know what must be done in order to store these data in machine-readable form. It is now customary to distinguish between the following forms of data handling and presentation:

- *Tabular lists.* These are simple systematic lists of numerical information.
- *Data compilations.* These are systematic collections of numerical data with some

reorganization for an optimal presentation.

- *Evaluations.* These are collections of numerical data, but with just one recommended best value for one measured quantity. This recommended best value is derived by a critical consideration of available information.

It is obvious that the last form, "evaluations," is of major importance when we are considering storage of data in machine-readable form in a field where we have very often more than one value for one physical quantity.

The structure of numerical databases under consideration can be described in four parts as follows:

1. Bibliographic part—title, authors, bibliographic citation.
2. Information part—for example, substance name, molecular formula, material description, experimental conditions.
3. Data part—for example, melting points, lists of spectrum peaks.
4. Computational part—For example, general program packages, including statistical packages, interpolation routines; specialized program packages, including calculation of thermodynamic properties, spectrum estimation.

This outline shows that a numerical database in physics will contain, aside from the dominant numerical information, also other alphanumeric information (for example, the source where the data came from in the bibliographic part). Very important, sometimes even essential, are also the specialized computer packages integrated in a numerical database. Their purpose is normally to offer a possibility for calculating data not stored on the basis of basic data or of some kind of knowledge base.

The structure of the numeric part is described by matrices (data tables) with m rows and n columns (m and n have any numerical value) where the elements A_{mn} of these matrices contain the numerical information (sometimes also alphanumeric comments). Further details are best illustrated by examples. For this we have selected three: one from the field of crystallography, one from the field of nuclear physics, and one from the field of spectroscopy.

The first database we would like to consider is the Inorganic Crystal Structure Database (ICSD). It contains crystallographic data (crystal structure data) from inorganic compounds:

- reference (authors, journal, year of publication, volume and page)
- chemical name (and phase designation)
- mineral name (and origin)
- chemical formula
- Chemical Abstracts registry number (if available)
- density
- unit-cell dimensions
- space group
- atomic coordinates, site occupation, oxidation state, and temperature factors for each element of the structure
- reliability index (R value)
- remarks concerning conditions of measurement or specialties of the structure not described by the above parameters

ICSD is produced by FIZ Karlsruhe and Gmelin-Institut in cooperation with the University of Bonn (for more details see IUCR, 1987). A typical record is shown in Fig. 10. ICSD contains about 30 000 structures that have been compiled from the open literature. Before the data were entered into ICSD, they were carefully checked both manually and automatically by application of suitable test programs. There are also two specialized program packages integrated into ICSD, one for the calculation of the different angles between the various atoms in the crystal and one for the calculation of distances between specified atoms. Both components can be accessed with the retrieval software without any problems.

As stated above, our next database example has been taken from the field of nuclear physics. In this area the database ENSDF (Evaluated Nuclear Structure Data File), where the printed version is the *Nuclear Data Sheets,* plays a major role. It contains the nuclear structure and radioactive decay characteristics of all isotopes known so far, i.e., the existing energies of the energy levels, the spins, the parities, and the half-lives and all decay energies of the related α, β, and γ transitions. It consists of a data part and bibliographic part, with the data part containing only evaluated data. ENSDF is produced and kept up to date by an international coop-

```
L1    ANSWER 1 OF 4
AN    34308  ICSD
CN    Calcium carbonate
        ***Aragonite***
LSF   Ca C O3
MF    C Ca O3
TI    Refinement of the crystal structure of the aragonite phase of Ca C
      O3
AU    Dickens, B; Bowen, J S
SO    Journal of Research of the National Bureau of Standards, Section
      A.
      Physics and Chemistry. (1971) Vol. 75 p. 27-32; CODEN=JNBAAR
CLP   A=4.9598(5) B=7.9641(9) C=5.7379(6)    unit: Angstrom
      ALPHA=90. BETA=90. GAMMA=90.    unit: Degrees
      NFU  4
      VOL  226.65  Angstrom**3
DEN   2.947(2)  g/cm**3
CSG   PMCN; 62
RVAL  0.04
ST    MINR Compound with mineral name
      TEM  Temperature in Kelvin at which the determination was made (if
      not room temperature)
      TEM 298
LAU   MMM
PRS   OP20
CCLS  MMM; D2H
CSYS  ORT
CSYM  CEN; NPOL
FTYP  ABX3

ATP   Atomic Parameters
      | At | Nr| Ox |   Wy | X          | Y           | Z           SOF |
      |----|---|----|------|------------|-------------|-----------------|
      | Ca | 1 | +2 |   4c | 0.2        | 0.58507(5)  |  0.25974(6)     |
      | C  | 1 | +4 |   4c | 0.2        | 0.2386(2)   | 0.4148(3)       |
      | O  | 1 | -2 |   4c | 0.25       | 0.0770(2)   | 0.4043(2)       |
      | O  | 2 | -2 |   8d | 0.4737(2)  | 0.3196(1)   | 0.4131(2)       |

      TF  Temperature Factors      unit: Angstrom **2
      |Atom  | B(1,1)   | B(2,2)   | B(3,3)   | B(1,2)    | B(1,3)    | B(2,3)   |
      |------|----------|----------|----------|-----------|-----------|----------|
      | Ca 1 | 0.664(9) | 0.599(9) | 0.601(9) | 0.0       | 0.0       | -0.01(1) |
      | C  1 | 0.75(5)  | 0.85(5)  | 0.46(5)  | 0.0       | 0.0       | 0.07(5)  |
      | O  1 | 1.50(5)  | 0.54(4)  | 1.04(4)  | 0.0       | 0.0       | 0.03(4)  |
      | O  2 | 0.71(3)  | 1.03(3)  | 1.02(3)  |-0.32(3)   | -0.02(3)  | -0.09(3) |

      TFLG 52 At least one temperature factor is implausible but agrees
      with the value given in the paper
```

FIG. 10. Data set from ICSD.

eration coordinated by the IAEA. In detail its contents can be characterized by the following figures.

1. Bibliographic part: contains about 120 000 bibliographic citations (from 1910 to 1991).
2. Data part: contains about 920 000 card image-based records. Data sets number about 10 700 of which 24% are adopted levels data sets, 27% are radioactivity data sets, and 49% are nuclear reaction data sets.

The structure of the database is shown in Figs. 11 and 12. An example of a typical data set is presented in Fig. 13.

The next database we would like to consider in this context is the spectroscopic C13 NMR/IR (carbon-13 nuclear magnetic resonance/infrared) database, which is produced jointly by Chemical Concepts and BASF (a large chemical company in Ludwigshafen, Germany) (the on-line version is offered by STN International).

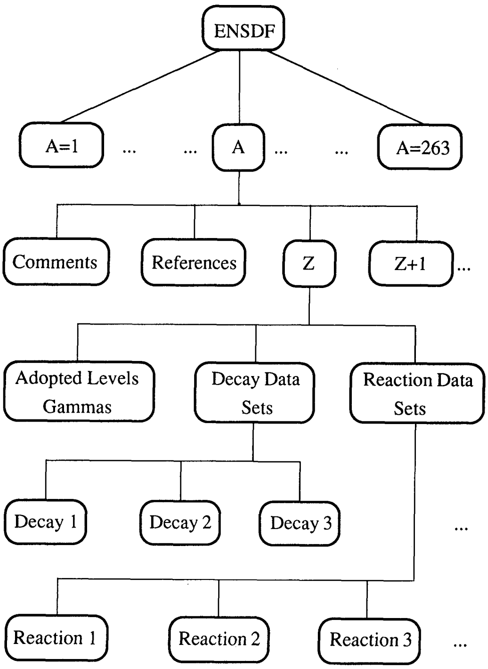
ENSDF
A=1
A
A=263
Comments
References
Z
Z+1
Adopted Levels Gammas
Decay Data Sets
Reaction Data Sets
Decay 1
Decay 2
Decay 3
Reaction 1
Reaction 2
Reaction 3

FIG. 11. File organization in ENSDF.

Typical contents	
^{13}C NMR part	IR part
Accession number	Accession number
Molecular formula	Molecular formula
Chemical name	Chemical name
Chemical Abstracts registry number	Chemical Abstracts registry number
Chemical structure (connection tables)	Chemical structure (connection tables)
Molecular weight	Molecular weight
Solvent	Source (sample)
Standard	Sample form
Instrument code	Melting point
Temperature	Peak lists
Conversion shift	Number of peaks
Origin number	Normalization factor
Chemical shift	
Relaxation times	
Bibliographic reference	

A sample for a ^{13}C NMR record is presented in Fig. 14 (see also Luksch, 1990). It should be noted that the database also has a graphical component that allows graphic representation of chemical structures and corresponding spectra. The database contains 102 000 ^{13}C NMR spectra (85 000 chemical structures) and 16 000 IR (15 000 structures). The data are taken from journal articles, spectral catalogs, and unpublished sources. In addition, specialized software packages are connected with the database. These are packages for spectrum estimation, spectrum interpretation, and similarity search. Of these, the most important and interesting one is that for spectrum estimation. It allows the calculation of unknown ^{13}C NMR spectra (chemical shifts) from a known chemical structure. This is based on the correlation between spectral features and the environment of an individual atom in a molecule described by the HOSE and HORD codes (hierarchically ordered spherical description of environment, hierarchically ordered ring description; see Bremser, 1978). A special library with spectral information linked to these HOSE and HORD codes is then built up which allows the derivation of spectra from topological matrices (see Bremser, 1978 and

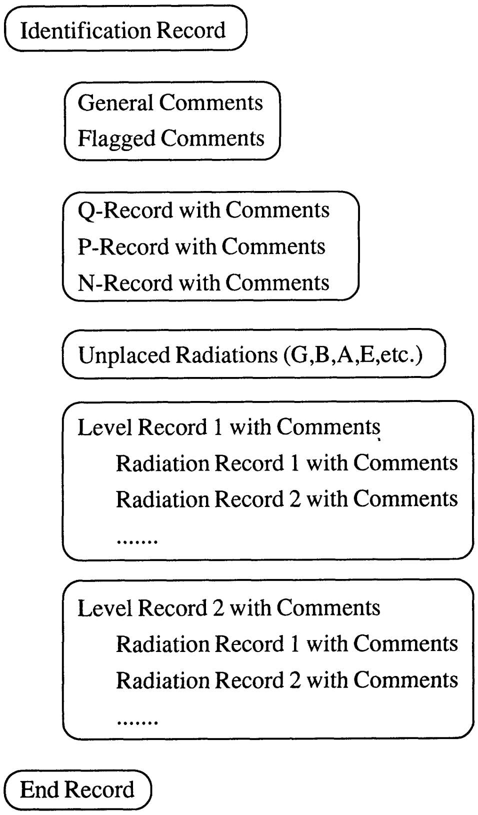
Identification Record
General Comments
Flagged Comments
Q-Record with Comments
P-Record with Comments
N-Record with Comments
Unplaced Radiations (G,B,A,E,etc.)
Level Record 1 with Comments
Radiation Record 1 with Comments
Radiation Record 2 with Comments
....
Level Record 2 with Comments
Radiation Record 1 with Comments
Radiation Record 2 with Comments
....
End Record

FIG. 12. Structure of a single ENSDF data set. All records are 80-byte card images. The letters Q, P, N, G, B, A, and E, respectively, stand for quantities important in nuclear physics, viz., *Q* values, parent record, normalization record of emitted radiation, γ, β, and α emissions, and energy data.

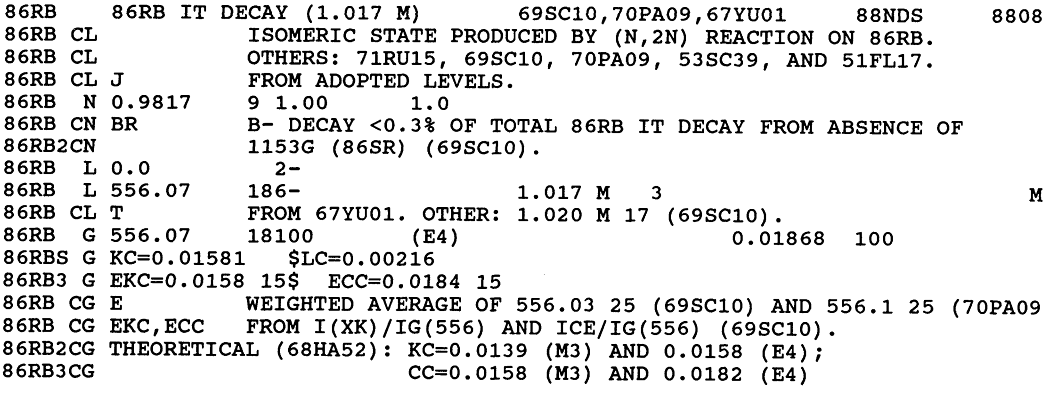

```
86RB     86RB IT DECAY (1.017 M)      69SC10,70PA09,67YU01      88NDS     8808
86RB CL             ISOMERIC STATE PRODUCED BY (N,2N) REACTION ON 86RB.
86RB CL             OTHERS: 71RU15, 69SC10, 70PA09, 53SC39, AND 51FL17.
86RB CL J           FROM ADOPTED LEVELS.
86RB  N 0.9817      9 1.00      1.0
86RB CN BR          B- DECAY <0.3% OF TOTAL 86RB IT DECAY FROM ABSENCE OF
86RB2CN             1153G (86SR) (69SC10).
86RB  L 0.0           2-
86RB  L 556.07      186-                1.017 M   3                      M
86RB CL T           FROM 67YU01. OTHER: 1.020 M 17 (69SC10).
86RB  G 556.07      18100       (E4)                  0.01868  100
86RBS G KC=0.01581   $LC=0.00216
86RB3 G EKC=0.0158 15$   ECC=0.0184 15
86RB CG E           WEIGHTED AVERAGE OF 556.03 25 (69SC10) AND 556.1 25 (70PA09
86RB CG EKC,ECC     FROM I(XK)/IG(556) AND ICE/IG(556) (69SC10).
86RB2CG THEORETICAL (68HA52): KC=0.0139 (M3) AND 0.0158 (E4);
86RB3CG                       CC=0.0158 (M3) AND 0.0182 (E4)
```

FIG. 13. Example of an ENSDF data set.

1987; Zupan, 1986). Why is this so important? As we said before, 102 000 spectra are in the database. We have, however, many more compounds existing in nature (about 12 million). Thus in case a spectrum must be interpreted that is not in the database, the corresponding chemical shifts for a postulated chemical structure can be estimated. In a retrieval session it is also very easy to manipulate the structures of the chemical compounds under investigation in order to achieve the greatest possible correspondence between calculated and measured spectra.

Since each of the spectroscopic methods [^{13}C NMR, IR, mass spectroscopy (MS)] has its own specific analysis potential for an unknown chemical structure, it is of great advantage to apply several spectroscopic techniques in parallel (multidimensional spectroscopy) to one unknown compound. The future of spectroscopic databases is therefore devoted to multidimensional spectroscopic databases in which the different sublibraries of specific spectroscopic data such as ^{13}C NMR data, IR data, etc., are linked to one central structure library with connections to the CAS, Beilstein, and Gmelin databases via corresponding registry numbers (see Fig. 15). A typical example is the SpecInfo database originally developed by BASF in Germany and now offered by Chemical Concepts.

These multidimensional databases can be combined with suitable intelligent software to build a multidimensional interpretation system, which is shown schematically in Fig. 16. The flow of information can be described in the following way. The MS, NMR, and IR spectra for the unknown substance are transferred to the system. On the basis of the MS, NMR, and IR subdatabases, chemical structure sets are found by applying similarity searches and statistical analysis of the resulting "hit" list. A structure generator then assembles the substructures and derives structures. MS, NMR, and IR spectra can afterwards be predicted and then fed back with suitable ranking into the interpretative system. Finally, the solution space of the structure gen-

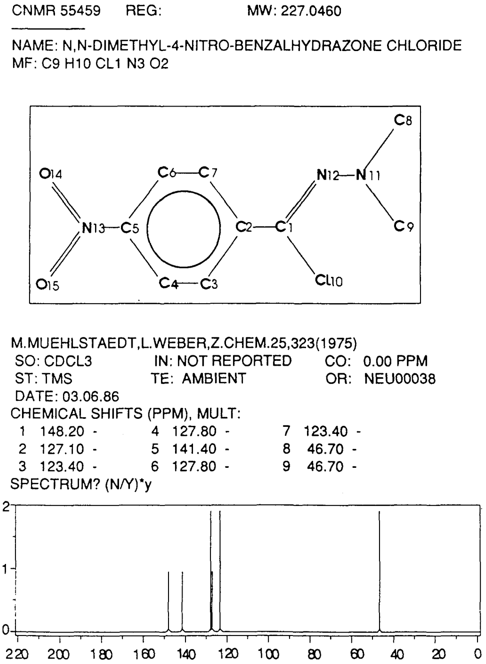

```
CNMR 55459   REG:              MW: 227.0460

NAME: N,N-DIMETHYL-4-NITRO-BENZALHYDRAZONE CHLORIDE
MF: C9 H10 CL1 N3 O2

M.MUEHLSTAEDT,L.WEBER,Z.CHEM.25,323(1975)
 SO: CDCL3          IN: NOT REPORTED     CO:  0.00 PPM
 ST: TMS            TE: AMBIENT          OR:  NEU00038
 DATE: 03.06.86
CHEMICAL SHIFTS (PPM), MULT:
 1  148.20 -        4  127.80 -          7  123.40 -
 2  127.10 -        5  141.40 -          8   46.70 -
 3  123.40 -        6  127.80 -          9   46.70 -
SPECTRUM? (N/Y)*y
```

FIG. 14. An example of a retrieved ^{13}C NMR data set.

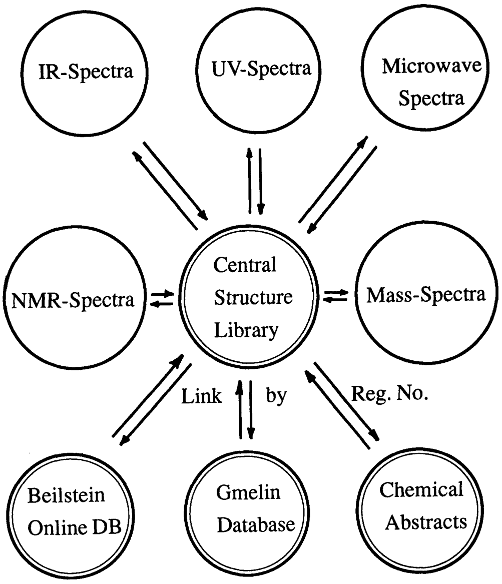
IR-Spectra
UV-Spectra
Microwave Spectra
NMR-Spectra
Central Structure Library
Mass-Spectra
Link
by
Reg. No.
Beilstein Online DB
Gmelin Database
Chemical Abstracts

FIG. 15. Spectral data pools with a central chemical structure library.

erator is restricted in a kind of convergence process such that a small number of possible candidates for the structure remain (see Bremser, 1978; Zupan, 1986).

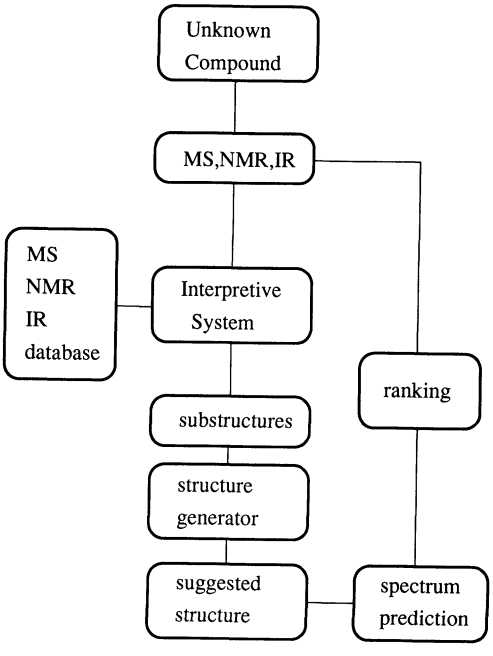
Unknown Compound
MS,NMR,IR
MS
NMR
IR
database
Interpretive System
substructures
structure generator
suggested structure
ranking
spectrum prediction

FIG. 16. Organization of a multidimensional spectral database system.

Contrary to the situation we have for the bibliographic databases, the coverage of the whole field of physics by existing numeric databases is very uneven. For example, the fields of crystallography (see Table 1), of nuclear physics (Table 2), of thermophysics and thermodynamics, and of spectroscopy are well covered. In the wide field of solid-state physics, on the other hand, there are hardly any databases. Those that exist contain mostly materials data and strictly speaking do not belong directly to the field of solid-state physics.

Concerning quality criteria there are specific differences between bibliographic and numeric databases. Reliability of data is most important in numeric databases, whereas this point is not so critical in bibliographic databases. In addition, it is important to store data and metadata (information necessary to define data), to refer to the sources of the data, to give information about type of apparatus, experimental conditions etc., and to make available special comments whenever necessary. In many cases the storage of the uncertainties belonging to the data is also of importance. In addition, continuous data maintenance and updating are often essential.

7. ON-LINE SERVICES

Usually, the producers of databases do not make their databases available to the public themselves, but pass these on to database vendors or on-line service centers, the so-

Table 1. The most important crystallographic databases (for more details see IUCR, 1987).

Atomic coordinates

1. Cambridge Structural Database
   (organic compounds)
2. Inorganic Crystal Structure Data Database
   (inorganic compounds)
3. Metals Data File
   (metals and alloys)
4. Protein Data Bank
   (biological macromolecules)

Unit-cell parameters
  NIST Crystal Data Identification File
  (Organic and inorganic compounds)
Powder patterns
  Powder Diffraction File
  (organic and inorganic compounds)

Table 2. The most important nuclear physics databases (for more details see Behrens and Rittberger, 1985).

1. Nuclear structure and decay data databases:
   ENSDF (Evaluated Nuclear Structure Data File)
2. Nuclear reactions (non-neutron) databases:
   None
3. Neutron reaction databases:
   a. Experimental data:
      EXFOR (EXchange FORmat database for experimental nuclear reaction data)
   b. Evaluated data:
      BROND 90 (USSR Evaluated Neutron Data Library)
      ENDF/B-VI (Evaluated Nuclear Data File, Brookhaven, USA)
      ENDL-84 (Evaluated Nuclear Data Library, Lawrence Livermore Labs.)
      IRDF 90 (International Reactor Dosimetry File, IAEA)
      JEF-2 (Joint Evaluated File, Nuclear Energy Agency, NEA)
      JENDL-23 (Japanese Evaluated Nuclear Data Library)

called *hosts*. These hosts have suitable mainframes at their disposal where the databases are loaded. By offering also a more or less refined retrieval software they make the databases available to the public. The on-line access for the normal user will be via the usual telecommunication networks. As the word "on-line" implies, this search technique is an interactive process, where user and computer are conversing in real time. For the user the following type of equipment is required (see COMPUTER HARDWARE). First, a terminal or, even better, a microcomputer with ASCII standard and with some suitable communication software is needed. To link the PC (or the terminal), which should have an RS232 port and an asynchronous communication card, to the phone, a *modem*, which stands for modulator/demodulator, is also needed. The modem is a device that converts the digital signals coming out of the PC into the analog language of the telephone system, and vice versa. An important point in this context, however, is the number of characters to be transmitted per second over the line. Usually, the modem speed is measured in baud (bits per second). A 300-baud modem can handle 30 characters per second (approximately five average-length words) and a 2400-baud modem, 240 characters per second (approximately 40 words). Thus it is advantageous to have at least 2400-baud facilities, but today even faster lines and modems are available.

In the first part of the communication link, the connection between the user's PC and the host is made via the ordinary telephone line. Because of cost considerations, the second and major part of the path makes use of special telecommunication networks. These are accessed by making a telephone call to the nearest node; from this the call is automatically switched through the telecommunication networks (packet-switching systems) to the host. At present, one sometimes finds more complicated schemes, viz., a mainframe at the center with telecommunication links to both the host and the user. Of course, the customer has to sign an agreement with the host in order to receive a password that permits access to the databases.

Today we have about 5000 databases publicly offered on line and about 600 hosts (Cuadra, 1990). Such a host may offer from very few to hundreds of databases for all fields of science. For physics the most important public hosts are STN International, DIALOG, BRS, ESA-IRS, Questel, and Data-Star (addresses and more detailed description can be found in Cuadra, 1990). Each of these hosts uses its own retrieval software. This has the consequence that the corresponding search languages are different. In case of STN International, for instance, the name of this language is Messenger. In many cases, however, the necessary commands are similar to a large degree. It is clear that the difference between the various command languages complicates the human–host interaction since a user must learn the language peculiar to each service. In the interaction process that takes place between a human user and the host, there are also other frustrating although often trivial problems. One of these is the log-in procedure to a host computer, which is a trivial but very often cumbersome procedure. It could require typing in a certain number of consecutive digits with restrictions for timing. One typing error normally has the consequence of an invalid call and of the obligation to repeat it. Therefore a large number of on-line searching aid software packages are nowadays available. When installed on the PC they can, for instance, be used for retrieval. Their purpose is to improve the user-friendliness of the host access. These on-line searching aid systems can be classified as (Efthimiadis, 1990)

- gateway systems,
- front ends, or
- intermediary systems.

Gateway software allows auto-log-on (automatic dialing, automatic selection of a communication network, automatic connection to the host computer, automatic transmission of the user's log-in password) and possibly uploading and downloading. Front-end software includes the above features, but goes beyond by simplifying steps of the searching process itself (presearch formulation, editing, uploading of search formulation, downloading, and post-processing). Intermediary systems are often real expert systems and have the aim of replacing human intermediary searchers in some way (query reformulation, query expansion, natural language interface). For more details, see, e.g., Efthimiadis (1990).

From the above it follows that on-line searching aids also simplify the problem of accessing more than one host. In addition, file transfers such as downloading (capturing of data sent by the host computers) and uploading (sending data from the user's computer to the on-line host) are easily performed by making use of online searching aids. This is important because the users very often want to manipulate (post-process) the data they retrieved, which can be done after downloading by application of various packages. Examples for gateway systems are PC Net/Link and iNet; for front-end systems ProSearch, Sci-Mate, STN-Express, STN Personal File System, Molkick, etc.; and for intermediary systems CONIT, NDA, etc.

8. OTHER DATA DISTRIBUTION CHANNELS

An on-line connection between user and host computer is but one of many data distribution channels. Nowadays, quite a few alternative methods of storing and transferring data from one location to another are being used. The possibilities include tape drives, Bernoulli cartridge systems, removable hard disks, write-once optical disks (WORMs), erasable optical disks, floppy disks, and even VHS video tapes.

Tape drives are by far the most popular medium. Data are transferred between mainframes on nine-track, half-inch, open-reel tapes with capacities between 100 and 300 MB depending on reel size and recording density (e.g., 6250 bits per inch). Capacities for systems for use with mini- and microcomputers range from 40 MB to 5 GB. These include cartridges of various makes, such as TK50 and TK70 (VAX computers), DC 6000 QIC, DC 2000 minicartridges, Teac's D/CAS cassettes, and 4-mm digital audio tape (DAT). An overview of the different systems available can be found in Stone (1990).

Because of the predominance of MS-DOS™ and its offspring in the world of personal computers, exchange of smaller amounts of data on standard $5\frac{1}{4}$- or $3\frac{1}{2}$-in. floppy disks is very common. The use of optical disks with their larger storage capacities is increasing (CD-ROMs). WORMs and erasable optical disks are still quite expensive at present, and lack of standardization hampers their use for data exchange.

GLOSSARY

Bibliographic Database (also called Reference Database): A computer-readable file containing literature citations or references to published articles and documents.

Bit: A single linear digit having one of two states, 0 or 1.

BTREE (or Bayer Tree): A way of organizing structured data in a computer in order to minimize the time needed to access it.

Byte: Group of eight bits, used as the smallest computer storage and processing unit. In addition, it is the basis of our character representations.

CAS: Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210.

CD-ROM: Compact-disk read-only memory. Read-only optical disk for the storage of digitized data.

DBMS or Database Management System: A software system with management and administrative capabilities for maintaining and controlling access to large amounts of structured data on a computer. A DBMS is a DBS with management facilities.

DBS or Database System: A software system supporting structured data on a computer, usually arranged in one or more files.

DDL: Data definition language, used to describe the logical file structures in DBS.

DDM: Data definition module, containing the data definitions for a file in a DBS in original text or compiled form.

Descriptor: A keyword contained in a thesaurus.

Direct Access File: A collection of data in a computer file to which access can be gained directly, i.e., without scanning the data from the start of the file.

DML: Data manipulation language. Command language of a DBMS. Sometimes there is no clear distinction between DDLs and DMLs.

Documentation Unit: The smallest unit of a bibliographic database containing all the data of a literature citation together with the terms of the subject analysis.

ER or Entity–Relationship Model: Conceptual model of the real world used in a DBS containing objects (entities) and their relationships.

Entity: Any object that exists on its own and is distinguishable (e.g., a person, an automobile).

Factual Database: A database containing factual information, mainly of non-numerical nature.

Field: The smallest subdivision of a record and usually contains numeric or alphanumeric information for a single topic.

File: A named organized collection of related data elements stored in a computer system.

File Design: Conceptual and logical design of a DBS.

Front-End Software: Software that is loaded on the user PC in order to improve the user-friendliness of the host access by allowing automatic log-on, prepared search formulations, editing, etc.

Full-Text Database: A database containing the entire text of a document.

Gateway Software: PC software allowing automatic log-in to a host computer.

Hardware: Any piece of physical equipment associated with computer processing or communication networks.

Hashing Algorithm: A function that computes, from a character string serving as key for a file, an integer in the range 0 to some maximum value. This "hash value" is used for directly gaining access to the required data in the file.

Hierarchical Model: Data model structured like an inverted tree. Each item may have one or more descendants, but only one parent.

HORD Codes: Hierarchically ordered ring descriptions. They describe rings and their environments in chemical molecules.

HOSE Codes: Hierarchically ordered spherical description of environments of atoms in chemical molecules.

IMS: IBM's Information Management System.

Index: An ordered list of the contents of one or more fields or parts thereof, each element or key pointing to or identifying the location of the corresponding record.

Indexing (of Documents): The assignment of keywords (descriptors) to bibliographic documents. The term is sometimes also used as a synonym for subject analysis.

Information or Data Retrieval: Transfer of selected information from long-term storage to computer memory for subsequent processing, e.g., for display on a terminal or printer, or downloading to some other device.

Inverted File: A file that contains the addresses or locations or records in which the values for given keys exist. It is arranged for each key in order of record number.

ISAM: Index sequential access method. A method for gaining access to data in a file containing records sorted by key values. A dictionary and a telephone book are everyday examples of index sequential lists. Access to these is by first locating the correct page, i.e., by looking at the entries on the top left or right corners or each page, and then sequentially reading the individual entries on the page.

Link: Connection between two data items or records. A link may be physical, i.e., one item may contain a pointer (or address) to the other, or logical, where two items are connected by fields having identical contents.

Modem: An acronym for modulator/demodulator, the name for a hardware device that interfaces between a telephone line (analog signals) and a terminal or PC (digital pulses).

Network Model: Data model in which the elements have complex (*m*:*n*) relationships. Each item may be connected to a number of parents and descendents.

Numerical Database: A database that predominantly contains structured numerical data such as physical properties of substances.

On-Line: Interactive process where the user and the host computer are conversing in real time by a constant connection via a computer terminal or a PC.

On-Line Searching Aids: Software systems providing support for on-line searching.

Packet-Switching Network: A communication network which, in order to transmit a message from one location to another, breaks the message into small units called packets containing the destination address. Through multiple use of a single transmission line the costs are shared and are much lower than with normal telephone lines.

Pointer: A data element containing the address of the object being pointed to. In the programming language C, pointers are 2- or 4-byte quantities that may be used in simple arithmetic operations (addition, subtraction).

Record: A linear logical unit containing data associated with one entity or object of a database.

Relational Model: Data model in which the elements can be ordered linearly like the rows of a table.

RS232: A specification defined by the Electronic Industries Association (EIA) detailing the physical configuration of the serial interface connector between data terminal equipment and data communication equipment.

Subject Analysis: The assignment of terms such as classification codes, keywords (descriptors), and other "value-added" codes to bibliographic citations.

Thesaurus: A controlled vocabulary of keywords to be used for indexing of documents. Entries are ordered hierarchically in narrower (NT) and broader (BT) terms. An entry may be classified as forbidden or point to related (RT) or preferred terms.

WORM: Write once, read many times optical disk for storing digitized information.

List of Works Cited

Bayer, R., McCreight, E. M. (1972), "Organization and Maintenance of Large Ordered Indices," *Acta Informatica* **1**:3, 173–189.

Behrens, H., Rittberger, W. (1985), "Nuclear and Particle Physics," in: D. F. Shaw, (Ed.), *Information Sources in Physics*, London: Butterworths.

Bremser, W. (1978), *Anal. Chim. Acta* **103**, 355–365.

Bremser, W. (1987), *ESN–European Spectroscopy News* **75**, 10–27.

Cuadra (1990), *Directory of Online Databases*, Vol. 11, No. 3, New York: Cuadra/Elsevier.

Date, C. J. (1987), *A Guide to the SQL Standard*, Wokingham: Addison-Wesley.

Efthimiadis, E. N. (1990), *J. Documentation* **46**, 218–262.

Franklin, C. (1989), "Hypertext Defined and Applied", *ONLINE* **13**, No. 3, 37–49.

ICSTI (1991), *International Classification Scheme for Physics*, ICSTI 1991.

IUCR (1987), *Crystallographic Databases*, Bonn: Data Commission of the International Union of Crystallography.

Lockemann, P. C., Schmidt, J. W. (Eds.) (1987), *Datenbank-Handbuch*, Heidelberg: Springer-Verlag.

Luksch, P. (1990), "Spectroscopic Online Databases—A Tool for Structure Elucidation," International Symposium on the Future of Scientific, Technological and Industrial Information Services, IAEA-SM-317, p. 22.

Marchionini, G., Shneiderman, G. (1988), "Finding Facts versus Browsing Knowledge in Hypertext Systems," *IEEE Computer*, **21**(1), 70–80.

Ross, R. G. (1978), *Data Base Systems*, New York: AMACOM.

Salton, G. (1983), *Introduction to Modern Information Retrieval*, New York: McGraw-Hill.

Salton, G. (1989), *Automatic Text Processing*, Reading, MA: Addison-Wesley.

Seymour, J. (1988), "Project 3: Programmable Databases", *PC Magazine* **7**, No. 9, 93–272.

Silberschatz, A. Stonebraker, M, Ullman, J. D. (Eds.) (1990), *Database Systems: Achievements and Opportunities*, Lagunita Report of NSF Invitational Workshop on the Future of DBS Research, Palo Alto, CA.

Stone, M. D. (1990), "QIC 150 Tape Drives: The Rising Star for Backup," *PC Magazine*, **9**, No. 17, 269–338.

Tepel, J. W., Müller, H.-W. (1990), "GAMCAT—A Personal Computer Database on Alpha Particles and Gamma Rays from Radioactive Decay," *Nucl. Instrum. Methods Phys. Res.* **A286**, 443–446.

Ullman, J. D. (1982), *Principles of Database Systems*, Rockville, MD: Computer Science Press.

Zupan, J. (Ed.) (1986), *Computer-Supported Spectroscopic Databases*, Chichester: Ellis Horwood Limited; New York: Wiley.

Further Reading

Date, C. J. (1986), *An Introduction to Database Systems*, Wokingham: Addison-Wesley.

Hall, J. L. (1986), *Online Bibliographic Databases*, London: ASLIB.

Jaschek, C. (1989), *Data in Astronomy*, Cambridge, England: Cambridge University Press.

Nadrachal, J. (Ed.) (1984), *Data Bases and Data Structures in Physics*, Proceedings of the Fifth Summer School on Computational Physics, *Computer Phys. Commun.* **33**, 1–297.

Perry, J. T., Lateer, J. G. (1989), *Understanding ORACLE*, San Francisco: Sybex.

Rossmassler, St. A., Watson, D. G. (1980), *Data Handling for Science and Technology*, Amsterdam: North Holland.

Rumble, J. R., Hampel, V. E. (Eds.) (1984), *Database Management in Science and Technology*, Amsterdam: North Holland.

Salton, G., McGill, M. J. (1983), *Introduction to Modern Information Retrieval*, New York: McGraw-Hill.

Salton, G. (1989), *Automatic Text Processing*, Reading, MA: Addison-Wesley.

Schwarz, St., Watson, D. G. Alvfeldt, O. (1983), *Nonbibliographic Data Banks in Science and Technology*, Paris: ICSU Press.

Shaw, D. F. (1985), *Information Sources in Physics*, London: Butterworths.

Shneiderman, B. Kearsley, G. (1989), *Hypertext Hands-On! An Introduction to a New Way of Organizing and Accessing Information*, Reading, MA: Addison-Wesley.

Registered Trademarks

ADABAS and NATURAL are products of the Software AG, Darmstadt, Germany.

dBASE is produced by Ashton-Tate, Torrance, California.

CLIPPER is produced by the Nantucket Corporation, Los Angeles.

MS-DOS is the trademark of Microsoft Corporation.

ORACLE is produced by the Oracle Corporation, Belmont, California.

PostScript is the trade mark of Adobe Systems, Inc.

COMPUTER GRAPHICS

Bruce Naylor and Arun N. Netravali, *AT&T Bell Labs, Murray Hill, New Jersey, U.S.A.*

Introduction 127
1. **Geometric Modeling** 128
1.1 Affine Transformations 128
1.2 Polytopes 129
1.3 Nonlinear Sets 131
1.3.1 Parametric Curves and Surfaces 132
1.4 Constructive Methods 134
1.5 Discrete-Space Methods 135
2. **Rendering** 135
2.1 Visibility Determination 136
2.1.1 Z-Buffer Algorithm 136
2.1.2 Scan-Line Algorithms 137
2.1.3 Visibility Priority 137
2.1.4 Ray Tracing 137
2.2 Illumination 139
2.3 Antialiasing 139
2.4 Surface Color Variations 140
3. **Animation** 140
4. **Hardware** 140
5. **Applications** 141
Glossary 142
Further Reading 143

INTRODUCTION

Computer graphics is the subfield of the computing sciences concerned with the computational creation of visuals, that is, of visually perceived and interpreted information with a predominantly spatial character. This is in contrast to the display of language-based information using symbols (e.g., letters). Visuals may take the form of something as simple as a graph plot of empirically acquired data, in which the geometric setting is used to convey nongeometric information, or it may be as complex as synthesizing realistic images that are not easily distinguishable from photographs, in which the visual component is the essence.

Regardless of the objective, computer graphics uses analytic geometry embedded in various data structures, such as arrays, lists, and graphs, as the basis for representing the basic geometry. This geometry is augmented by a variety of attributes, most notably color, which are appropriate for the intended use. Collectively this forms the *geometric model.* Operations on this model, such as creating a picture of it, are provided by a variety of algorithms.

Since visual perception is a dominant component for much of human experience, and is the sense that receives information at the greatest rate, computer graphics has many application domains. To date, these include visualization of data in business, science, medicine, engineering, architecture, mathematics, and the visual arts, and the simulation and control of processes in manufacturing. Its reliance upon computational representations of geometry, a subject addressed by the subfield of geometric modeling, is shared with other computational fields such as robotics, computer vision, image processing, and computational geometry, as well as with various forms of scientific geometric computation.

Computer graphics began in the early 1960s, most notably with a system called Sketch Pad developed by Ivan Sutherland (1989 Turing Award). Sketch Pad was an interactive design program for electronic circuits that used for its display a variant of an oscilloscope: a CRT in which the electron beam draws the image in a manner analogous to a person drawing with pen and paper. Such a device has been called variously a stroke or vector or random-scan display, and has with rare exceptions been monochromatic.

A flowering of basic research occurred in the field at the University of Utah's Computer Science Department, headed by Ivan Sutherland, during the very late 1960s and throughout much of the 1970s. This work focused on synthesizing more realistic images using TV-based CRTs, called raster-scan displays, that could easily paint a full-color image on the entire screen. Similar work was taking place

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

in industry with the creation of flight simulators: mock cockpits used to train aircraft pilots in which the images in the cockpit windows were generated by computer graphics, thus creating a "virtual reality." Raster-scan-based graphics has since become the dominant display method, made possible by greatly reduced costs of memory and much greater processor performance. It is now available on almost all computer systems, including personal computers. Accompanying this technological advance has been a flowering of the field in research, academic, and commercial arenas.

1. GEOMETRIC MODELING

Creating visuals is primarily a spatial activity, and so one must have a computational representation of space. This space is usually taken to be a Euclidean space, and it is almost always two dimensional (2D) or three dimensional (3D). Euclidean space is modeled algebraically either as an affine space, in which parallel lines never meet, or as a projective space, which adds to an affine space the "points at infinity" at which parallel lines meet (and so has the elliptic parallel property). Real affine spaces use the standard Cartesian coordinates to denote points: (x,y) if in 2D, or (x,y,z) if in 3D, where the coordinates are real numbers. Projective spaces add one more coordinate so that the affine point (x,y) corresponds to the projective points (xw,yw,w) for all $w \neq 0$. If $w=0$, then it is a point at infinity in the (x,y) direction. Points are typically represented computationally as one-dimensional (1D) arrays of reals, and a real number is usually represented by "floating point," where a number is represented in a manner comparable to scientific notation.

Geometric objects, such as lines, circles, or polygons, are interpreted as sets of points. These points are usually defined as those satisfying some equations and inequalities. Thus, for instance, a line in 2D is the set of points satisfying the equation $Ax+By+C=0$. A line segment is the subset of a line between two points on the line, say (x_1,y_1) and (x_2,y_2), which determine a set of inequalities, viz., $x_1<x<x_2$ and $y_1<y<y_2$. These expressions can then be used to test if an arbitrary point lies on the line segment by seeing if it satisfies the line equation and the inequalities. It is also possible to generate any number of points known to be on the line segment using these expressions. The data structure for a line segment can be just its two end-points, since they determine the line and bound the segment as well. Analogously, a circle can be represented by its center and radius.

1.1 Affine Transformations

A fundamental class of operations on sets of points consists of those that map each point onto another point to form a one-to-one correspondence between points, while preserving the neighborhoods of the points; that is, the relative closeness of points is preserved. As a consequence, they preserve the topology or "connectedness" of the set, and such transformations are called *homeomorphisms*. The simplest and most important member of this class are the *affine transformations*. These include translation, scaling, and rotation. If one excludes translation from this class, one is left with the class of *linear transformations*, also called linear operators, which has been studied thoroughly in the context of linear algebra on vector spaces. One can represent a 2D point (x,y) as a 2D vector $[x\ y]$ of a 2D vector space, and coordinate axes can be represented by the basis vectors of that space: the x axis by [1 0] and the y axis by [0 1]. Similarly for any dimensionality. Linear transformations are equivalent to a change of the vector basis: each basis vector is mapped to any arbitrary vector. The length of each new basis vector provides scaling along each axis. A rotation is achieved if the new basis vectors remain mutually orthogonal; otherwise, shearing is introduced. Any linear transformation on a d-dimensional space can be represented by a $d \times d$ matrix, whose computational representation is usually a 2D array of reals.

However, a change of vector basis cannot provide translation. While vector addition can do so, the use of only linear transformations can be retained if instead one uses projective space, so that a 2D Euclidean point will be represented by a 3D projective point (xw, yw,w). Most often $w=1$, so that one can think of the original 2D space as simply the $w=1$ plane in the 3D projective space (see Fig. 1). This solves the translation problem by mapping the 2D origin/zero-vector [0 0] to the new w basis vector [0 0 1]. A translation of $[Tx\ Ty]$

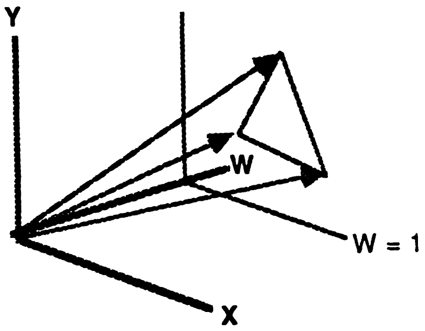

FIG. 1. Homogeneous coordinates.

is achieved by mapping the w basis vector to [Tx Ty 1], while the standard 2D linear transformations affect only the x and y basis vectors. In general, then, affine transformations of a d-space are represented by a restricted class of linear transformations of a $(d+1)$-space.

Projective space is also useful for central projections, e.g., the perspective projection in 3D, as well as for representation of surfaces as rational functions, and is generally preferable to affine space. The coordinates of a projective space are called *homogeneous coordinates*, because a point in d-space is mapped to a line through the origin in $(d+1)$-space. Similarly, lines are mapped to planes that contain the origin; that is, a line $Ax+By+C=0$ is represented by a vector [A B C], which can be interpreted as the normal of a plane containing the origin. As a result, the zero vector is a solution to the new equation representing each original point or line: thus, the term "homogeneous."

This mapping of a point or line to a set of one greater dimension applies to spaces of any dimension. The dimension-independent versions of point, lines, and planes are called *flats* or *affine subspaces* or *linear varieties*. Therefore, homogeneous coordinates map k-dimensional flats, or simply k-flats, to $(k+1)$-flats that contain the origin. A k-flat can be thought of as a k-dimensional space, say determined by the first k coordinates of a d-space, to which an arbitrary d-dimensional affine transformation has been applied. They are also defined as the intersection of hyperplanes, where a hyperplane is a $(d-1)$-flat defined by the equation $a_1x_1+a_2x_2+\cdots+a_dx_d=a_{d+1}$, i.e., a linear equation in d variables. In 2D a line is a hyperplane, and in 3D a plane is a hyperplane. The intersection of n hyperplanes is in general a $(d\text{-}n)$-flat, $0\leqslant n\leqslant d$.

1.2 Polytopes

Points, line segments, polygons, and polyhedra are the most commonly used geometric sets, and they are respectively the 0D (zero-dimensional), 1D, 2D, and 3D instances of the dimension-independent class known as *polytopes*. These are objects with "flat faces" (Fig. 2). One method of defining polytopes is by a recursive definition in terms of their boundary. Observe that a line segment, which is a 1D polytope, has for its boundary two points, which are 0D polytopes. Similarly, the boundary of the 2D polygon is comprised of a set of the 1D line segments, and the 3D polyhedron is bounded by a set of 2D polygons. In general, a d-polytope is bounded by a set of $(d-1)$-polytopes, which are referred to as the $(d-1)$-faces of the polytope. If one unwinds this recursive definition, it becomes evident that a d-polytope has a boundary comprised of k-dimensional faces for all k, $0\leqslant k<d$.

One can construct a data structure representing a polytope based on this property. Given the 0-faces (vertices), each 1-face (edge) is represented by its two vertices. Each 2-face (polygon) is then a list of 1-faces, and in general each k-face is represented by a list of $(k-1)$-faces. These lists of k-faces are further organized into connected sets, called manifolds. This method of representing each polytope by its boundary is called a *boundary representation*, and there is a one-to-one correspondence (an isomorphism) between connected components of the boundary and elements of the data structure (Fig. 3). This data structure is a graph in the form of a hierarchy

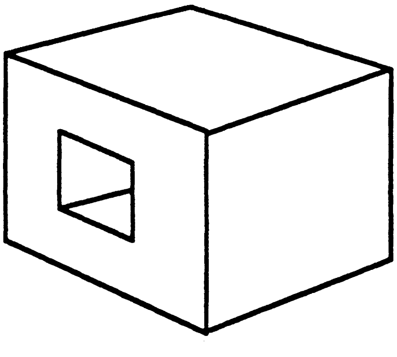

FIG. 2. A polyhedron.

of lists: nodes at one level correspond to all faces of a particular dimension. Arcs in the graph are of two types: between a k-face node and the nodes of its bounding $(k-1)$-manifolds, and between elements of the same manifold. The only purely geometric information is the coordinate values of the vertices; all other information is topological, i.e., information indicating which faces are connected to which other faces. Thus, a boundary representation, also denoted as *b-rep,* is often characterized as being a topology-based representation.

A crucial property of polytopes is that all of the faces are flat; that is, each k-face is contained in a k-flat. This restriction to sets defined in terms of linear equations (hyperplanes) greatly simplifies computations, as the intersection of two flats is another lower-dimensional flat. The fact that any k-face lies in the intersection of $d-k$ hyperplanes can be used as a basis for a representation of polytopes.

First note that a hyperplane h partitions space into two half-spaces, denoted as h^+ and h^-. The intersection of any set of half-spaces is a convex polytope (or the empty set), where a convex set is one for which the intersection of it with a line is a single line segment (or single point). All faces of a convex polytope lie in the hyperplanes that form the half-spaces. However, since a half-space is a convex set, and the intersection of convex sets is always another convex set, this approach can only represent convex polytopes.

The approach can, nonetheless, be extended to represent any polytope, not just convex ones, by allowing both half-spaces of a hyperplane to contain parts of the polytope. This is achieved by recursively subdividing space by the hyperplanes that contain the faces of the polytope. A single local operation is needed: partition a region r by a hyperplane h that intersects r. Usually, h contains a face that lies in r. This binary partitioning of r creates two subregions $r^+=h^+\cap r$ and $r^-=h^-\cap r$, where the subregions are open sets, i.e., they do not include h (*open sets* are sets that have had their boundary points removed). This process can be represented by a three-node graph in which the nodes correspond to regions r, r^+, and r^-, and arcs are between r and its two "child" regions (Fig. 4). The right arc can be interpreted as corresponding to the operation $h^+\cap x$ for any set $x\in r$, and the left arc to $h^-\cap x$. Now the two subregions can be treated just like r; that is, they can be partitioned by hyperplanes containing faces lying in them, and then their children can be partitioned, and so on recursively. The result is a hierarchy of regions, and the coresponding graph is a binary tree in which each node of the tree corresponds to one region, and arcs indicate the relation between parent and child regions of creating the later from the former by partitioning. Such a geometric structure is called a *binary space partitioning tree* or, alternatively, a *partitioning tree* or *bsp tree.*

Generally, the initial region is all of d-space, and a consequence of inducing a partitioning tree structure on d-space is that a disjoint convex decomposition/partitioning is formed by the unpartitioned leaf regions, called *cells.* Convexity arises from the fact that

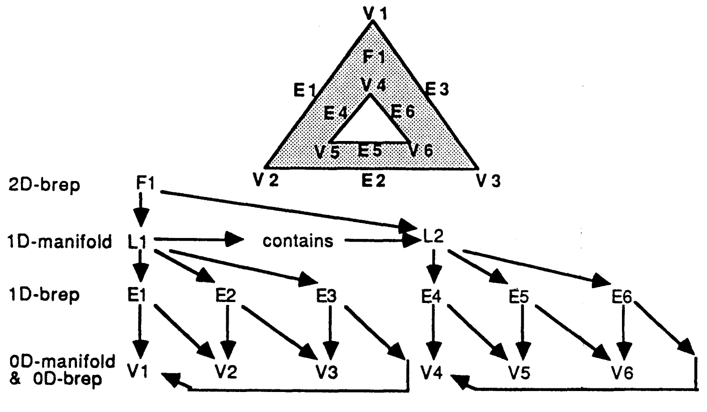

FIG. 3. Boundary representation of a polygon.

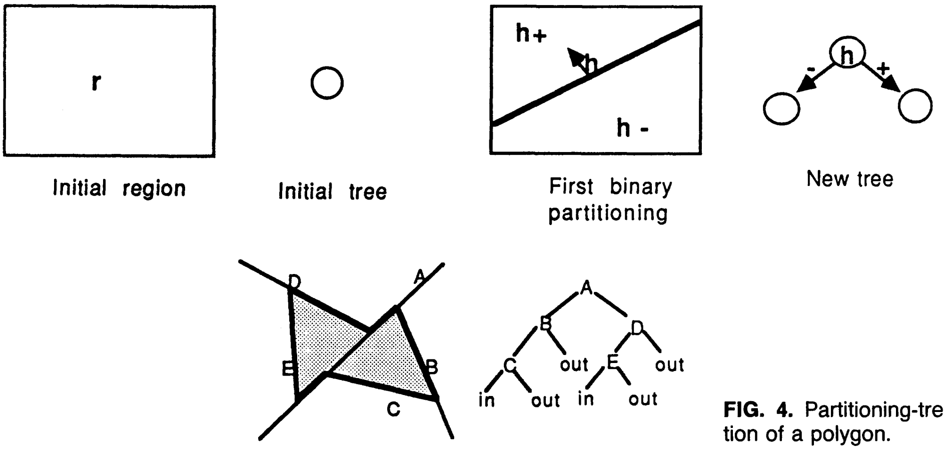

FIG. 4. Partitioning-tree representation of a polygon.

each region *r* of the partitioning tree is created by the intersection of half-spaces corresponding to the arcs in the path of *r* starting at the root of the trees. Since the partitioning tree has been constructed so that all polytope faces lie in the partitioning hyperplanes, none of the boundary lies in any cell, and so each cell with be either entirely in the polytopes's interior or in its exterior. The interior cells then form a convex decomposition of the polytope (and the out-cells form a convex decomposition of the exterior as well). The polytope is defined to be the union of the in-cells, which are open sets, followed by a closure operations that adds boundary points and fills in the cracks between in-cells.

The development of polytopes has implicitly assumed that the flatness of faces is an intrinsic property. However, it is easy to imagine warping the faces of a polytope into curved faces, In fact, systems have been developed that use nonflat polyhedral faces (called *algebraic polytopes,* if the faces lie in polynomials). So flatness need not be considered an intrinsic property of polytopes. However, the idea of a geometric set whose boundary is a set of "smooth" faces, i.e., faces without "interior edges," is intrinsic to the concept of polytopes; that is, polytopes have *piecewise* smooth boundaries. Thus the intrinsic properties of polytopes are that they have smooth faces, and they have a finite number of faces. The former is satisfied through the use of continuous functions, developed in classical mathematics, whose gradient (tangent function) is continuous. The latter is a combinatorial property, which together with the computational setting is an issue addressed by computer science; in particular, it is concerned with the development of representations and efficient algorithms providing operations on those representations. The boundary representation and the partitioning tree are examples of this.

1.3 Nonlinear Sets

While using only linear expressions to define a geometric set simplifies considerably all computational aspects, restricting faces to being flat is too limiting, since many desirable surfaces are not piecewise flat. Defining surfaces that are curved requires using nonlinear functions, such as polynomials of degree greater than 1, or transcendental functions such as sine or cosine or exponential functions, or any easily computable function whose gradient is continuous. Unfortunately, nonlinear functions can be considerably more complicated than linear functions. Because of this, it is very common to use the nonlinear functions to define the model, and then to construct a piecewise linear approximation of the nonlinear sets so that the difficult computations are performed using linear polytopes.

There are two fundamental forms for defining sets: the *characteristic function* that determines whether a point is a member of a set, and the *generating function* that enumerates the points that are members of the set. For sets of integers, the computable sets, or *recursive sets,* are those that have a computable characteristic function, and the semicomputable sets, or *recursively enumerable sets,* are those with a computable generating function. For polynomials, implicit representations provide characteristic functions and parametric functions provide generating functions. The implicit form defines the sets in d-space as $\{\mathbf{x} \mid \mathbf{f}(\mathbf{x})=0\}$, where $\mathbf{x}$ is a d-dimensional vector and $\mathbf{f}$ is a scalar-valued function. For example, $x^2+y^2-r^2=0$ is the implicit form defining a circle of radius r centered at the origin. The parametric form uses a set of parameters to generate the set $\{\mathbf{x} \mid \mathbf{g}(\mathbf{t})=\mathbf{x}\}$, where $\mathbf{x}$ is a d-dimensional vector and $\mathbf{t}$ is a vector of dimension $<d$. The dimension of the set is that of t. A parametric form of the circle above using parameter t is $x=r \cos t$ and $y=r \sin t$, for $0 \leqslant t < 2\pi$.

Any set whose characteristic function is given by a polynomial is referred to as an *algebraic set*, and the 1D and 2D algebraic sets are known, respectively, as algebraic curves and algebraic surfaces. The study of algebraic sets is the principal topic in the branch of mathematics known as *algebraic geometry*. While algebraic geometry considers polynomials of any degree, the lower-degree algebraic sets are the most widely used. For example, the second-degree polynomials in 2D define the conics: circle, ellipse, hyperbola, and parabola. In 3D they are the quadric surfaces: sphere, ellipsoid, hyperboloid, paraboloid, and cone. For sets defined by higher-degree polynomials, comparable classifications are not always known; and, unlike first- or second-degree sets, they may intersect themselves, creating singularities, and they may not admit a simple parametrization. In general, the greater the degree of the polynomial, the greater the computation that is involved and also the greater the numerical error due to the floating-point representation of reals. Thus, in practice, only polynomials of relatively low degree are computationally viable if the polynomial is an *irreducible* polynomial. A polynomial is said to be reducible if it is the product of two or more lower-degree polynomials, and irreducible if it is not. Thus $x^2 \equiv x \times x$ is reducible, whereas x^2+y^2 is irreducible. Irreducible polynomials are an important class in that they provide the primitive algebraic sets from which all other algebraic sets can be constructed, and they are referred to as *algebraic varieties*. Planes and spheres are examples of algebraic varieties.

1.3.1 Parametric Curves and Surfaces. Defining a complicated surface using only a single polynomial requires using one of high degree, which as noted above is computationally unattractive. An effective alternative is to define a set piecewise, where each piece is defined by a low-degree polynomial. The pieces then connect to each other in a patchwork manner to form a continuous set. Since each piece is a subset of an algebraic set, it is continuous for all derivatives. But at the places where the pieces join together, the continuity varies according to need. Tangent continuity C^1 is required to ensure that at each join the set does not form a corner, whereas curvature continuity C^2 is used to induce greater smoothness; only rarely are higher degrees of continuity needed.

An important instance of constructing piecewise polynomial curves with smooth joins in the class of curves known as *splines*, which are usually piecewise cubic curves with C^2 continuity at the joins. Piecewise definitions are most often formulated using the parametric rather than the implicit form because the desired piece of the polynomial can be specified by restricting the parametric domain. For example, a set in 3D is defined by the coordinate functions $x=X(t)$, $y=Y(t)$, and $z=Z(t)$, where each coordinate function is a polynomial, and a piece can be defined by limiting the coordinates of t to be in the interval [0 1]. If t is a 1D vector, i.e., a scalar, then the set is a space curve. If t is a 2D vector, a *biparametric* surface is defined. The biparametric surfaces, often denoted as $P(u,v)$, can in fact be interpreted as being a set of parametric curves: holding one parameter constant determines a curve in the other parameter (Fig. 5). This property has given rise to calling these surfaces *cross-product surfaces or tensor product surfaces.*

If one wants to design curves and surfaces interactively, that is, to use a person's visual esthetics to select the shape, the method of shape specification must be simple and intuitive. Specifying the coefficients of polynomials directly does not provide this. Consequently, geometric methods, rather than algebraic ones, have been developed for the interactive design process. A classical approach is *interpolation:* given a set of user-defined points, compute a piecewise polynomial curve or surface that goes through the points and provides the desired degree of continuity at the joins. In fact, splines were developed for exactly this purpose. Interpolation is sufficient if one does not want to adjust

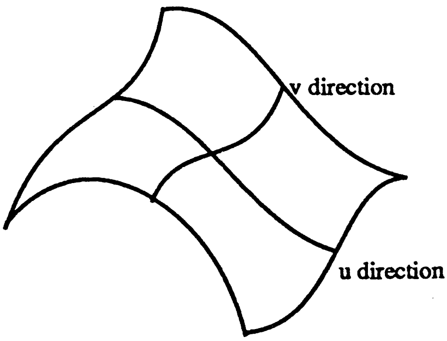

FIG. 5. A parametric surface patch.

the points subsequently, for doing this may cause the shape to change everywhere. With interpolation, any change can have a global effect, whereas the intent is usually to make a local modification, that is, to move only one point and the part of the curve around that point. To circumvent this problem, methods have been developed based on the idea of approximating a set of points rather than interpolating them. The points that are approximated are called *control points* since the designer uses them to control the shape; but the set only lies near the points, not through the control points (in some cases a few of the control points are indeed on the set).

The most well known of these are the Bezier curves and surfaces. A Bezier curve, which is defined by a single polynomial of degree n, is determined by $n+1$ control points, numbered P_0, P_1, ..., P_n (see Fig. 6). To obtain local control, a curve is constructed piecewise from a set of Bezier curves. Thus, only a subset of each Bezier curve is used. The end points of that curve segment are given by the first and last control points, P_0 and P_n. These two control points lie on the curve itself and are the only control points that do so. The tangent line at P_0 is the line through P_0 and P_1, and similarly the tangent line at P_n is the line through P_{n-1} and P_n. An analogous relation holds for higher derivatives at the end points (e.g., the curvature at P_0 is determined by P_1 and P_2), and so there is an intuitive geometric relation between control points and the curve.

The control points taken by themselves as a list of points can be linearly interpolated to form what is called the "control polygon" (a bit of a misnomer, since the first and last control points are not considered "connected," and so it is really a piecewise linear curve). The importance of the control polygon is that the Bezier curve is an approximation of it, and this is why Bezier curves are considered an approximating method. The algebraic analogue to the geometric formulation of Bezier curves is the approximation of functions using the Bernstein basis. Any polynomial of degree n can be represented exactly by the weighted sum of $n+1$ Bernstein basis functions $B_{i,n}(t)$, $0 \leqslant i \leqslant n$, where each is a polynomial of degree n. In particular, $B_{i,n}(t) = C(n,i)t^i(1-t)^{n-i}$ where $C(n,i)$ are binomial coefficients. They have the property that they sum to 1, i.e., $\Sigma_i B_{i,n}(t) = 1$. The control points are used to determine the weights of each basis function; that is, P_0 is the weight for $B_{0,n}$, P_1 for $B_{1,n}$, and so on, so that the curve is $P(t) = \Sigma_i B_{i,n}(t) P_i$. While the algebraic form is important for analysis and computation, it is the geometric characteristics of manipulating control points that has made the Bezier form appealing for geometric design.

When one is constructing a curve by joining several Bezier curve segments, maintaining continuity at the joins requires that the control points of the two segments at a join be geometrically related. For example, tangent continuity between curve segments A and B requires that the control points $A_n = B_0$ and that the tangent line at the join contains control points A_{n-1} and B_1. Constraining control points to achieve continuity is a consequence of the fact that the definition of each segment is independent of every other; and so, only constraining the control points can induce a relation between segments.

An alternative that performs this correlation intrinsically is the *B-spline* method. It retains from Bezier curves the idea that the curve is a piecewise polynomial and that each segment is the weighted sum of $n+1$ basis functions with the weights being the control

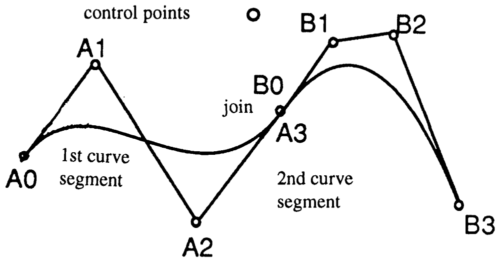

FIG. 6. Cubic Bezier curve with two segments.

points (the unweighted basis functions sum to 1 as well). The difference lies in the fact that a single control point is used as a weight in $n+1$ consecutive segments, and this then provides a means of ensuring continuity at the joins up to C^{n-1}. Thus, a cubic *B*-spline curve is piecewise cubic with C^2 continuity at the joins. Control point P_i is used as a weight in curve segments $S_i, S_{i+1}, \ldots, S_{i+n}$, but it is the weight of a different basis function in each of the segments. Since P_{i+1} is used in segments $S_{i+1}, S_{i+2}, \ldots, S_{i+n+1}$, two consecutive control points contribute to n common segments, which result in the desired automatic correlation between segments to ensure continuity.

Unfortunately, parametric piecewise polynomial curves in 2D of any degree cannot exactly define a conic, which is an important class of curves. However, if the curves are defined in projective space, they indeed can define conics. This is because the perspective projection of a 3D, degree 2, parametric space curve can be exactly a conic. Thus, it is common to use projective space for defining parametric curves, and such curves are called *rational curves,* because the perspective division of $X(t)$ and $Y(t)$ by $Z(t)$ results in a ratio of two polynomials. Analogous reasoning yields rational surfaces.

1.4 Constructive Methods

The methods just described define a shape directly, primarily in terms of its boundary. These direct methods can be supplemented by constructive ones in which two or more of the direct methods are used in conjunction to construct a new shape. The most common such method is the use of Boolean set operations: union, intersection, and both symmetric and asymmetric difference, which correspond to the Boolean functions of "or," "and," "exclusive-or," and "subtraction" (A&~B). Figure 7 illustrates their use in forming a new polygon from set operations on two other polygons. The result can then be used in subsequent set operations, resulting in a hierarchy of shapes. This method of shape creation is called, when in 3D, *constructive solid geometry.*

Another constructive method, called *instancing,* creates multiple copies of a shape, or "instances," by applying a unique affine transformation to the original for each copy. Inherent to this concept is the property that any subsequent change to the original will cause all instances to change; that is, the instances are always defined to be the current value of the original transformed by the current value of the transformation.

A third basic constructive method utilizes the idea of a shape moving in time. For example, if a line segment is moved some distance along a particular direction, all of the points "visited" by the line segment form a quadrilateral. If a line segment is rotated 360° about one of its points, the locus of points swept out is a disk. If instead the center of rotation is not on the segment but is on the line containing the segment, a disk with a hole in it is formed. There are all methods of creating *swept objects.* The technique of moving a shape along some path, i.e., along a curve, is called *translational sweeping,* while rotating a shape about some axis is called *rotational sweeping* (also called, in 3D, generation of surfaces and solids of revolution).

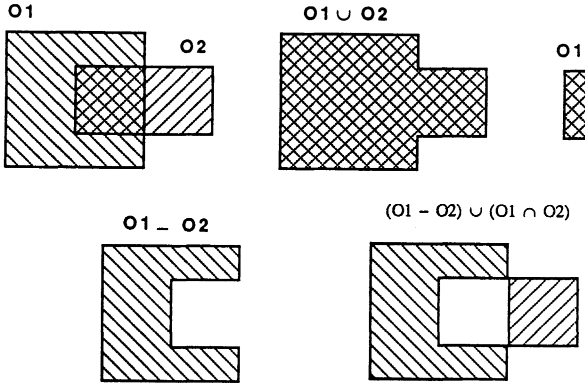

FIG. 7. Constructing sets with set operations.

1.5 Discrete-Space Methods

The methods discussed so far have all made explicit use of the idea that space is continuous. An alternative is to treat space as discrete in the form of a regular lattice of points created by restricting the coordinates to the integers. At each lattice point, an arbitrary set of attributes, like color, can be defined, and a *d*-dimensional space can then be represented computationally by a *d*-dimensional array of attributes. In 3D, each point is referred to as a *voxel* (derived from "volume element"). This representation is typically generated by discretely sampling the analog output of a sensing device, such as a TV camera, which detects some form of electromagnetic energy. Medical devices used for imaging the human body are another example of such devices.

While the array representation is very simple, it can require substantial amounts of memory when compared to typical continuous space representations. An approach to reducing this is to find connected regions of elements that have the same attributes, called *homogeneous regions,* and to represent each such region as one element in the data structure. A standard approach to this is to subdivide space into homogeneous regions using recursive subdivision, as was done by the partitioning tree discussed above. For discrete domains a commonly used method is called the *quadtree* in 2D and the *octree* in 3D. A quadtree is formed by recursively partitioning a square region into four "quadrants" (Fig. 8), and an octree is formed by recursively partitioning a cubic region into eight "octants." During the recursive subdivision, the partitioning terminates whenever a region is produced that is homogeneous; otherwise the region is partitioned. The limit to the subdivision occurs whenever a lattice point is reached, as this is a homogeneous region by definition.

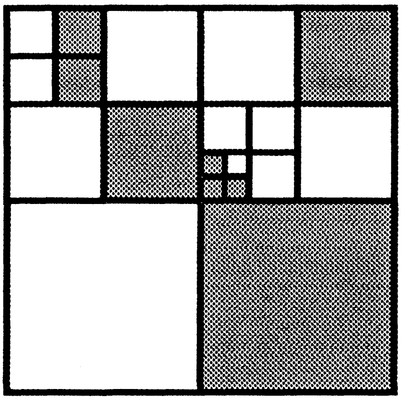

FIG. 8. Quadtree representation.

2. RENDERING

Given a geometric model representing a 3D scene, defined using any of the above techniques, the most important and computationally intensive operation is to synthesize a 2D image of the scene for display with video, film, or print media. This is called *rendering.* The goal is to create an image that will resemble, to some degree, what a person or camera would see if the scene were in some sense real. If the scene is a realizable 3D world, then rendering may try to achieve photorealism; that is, the image generated is indistinguishable from a photograph of a real scene. However, an alternative objective is simply to communicate information of some sort without being concerned with photorealism. This information may have an intrinsic 3D geometric form, such as weather data, or the 3D setting may be used simply as a means of communication, as in the display of statistical data points. But even when photorealism is not an objective, the appropriate rendering algorithms are essentially simplified versions of those needed for photorealism, since the human visual system has evolved for the express purpose of perceiving a 3D physical world.

Rendering algorithms simulate, to varying degrees of accuracy, the physics of the propagation of light through a homogeneous media and the interaction of light with matter, such as reflection, transmission, diffraction, and scattering. However, to do this accurately, one might think that all such propagation and interaction with objects in the scene would need to be simulated at the level of photons. It would, however, be computationally intractable to perform such a simulation. Fortunately this is not necessary since the resolution of the human visual system is limited, and so every detail provided by such a simulation could not be perceived. Therefore, the real challenge in rendering is to develop computationally attractive algorithms that produce realistic-appearing replicas of as large a class of phenomena as possible. Over the last two decades, many such algorithms have been developed that can indeed effectively simulate at reasonable costs a variety of optical effects; among these are reflections, refraction, transparency, shadows, and surface texture.

Synthesizing an image using a geometric model of a 3D scene composed of a variety of objects and sources of illumination requires, in addition, a virtual camera described by its viewing position, orientation, and field of view, as well as an image plane that models the display surface onto which the scene will be projected. Rendering can then be understood as being performed in two distinct steps: *visibility calculations,* which compute the propagation of light through homogeneous media, and *shading,* which performs the computation needed to simulate interaction of light with matter. (While in a number of algorithms these steps may not be very distinct, such a distinction will simplify the presentation.)

For visibility calculations, we consider the light that reaches the viewing position along rays emanating from the object's surfaces. However, objects closer to the viewing position will occlude objects farther away, and this occlusion must be computed. If we then consider rays emanating from the viewer instead of from the surfaces ("antioptic" rays), then any such ray may intersect any number of surfaces, but it will intersect the image plane on which we will create the picture at only one point. From which of these surfaces will the light reach the image plane point? Only the light from the closest surface will, and so this surface point is considered visible. If for all rays emanating from the viewing position the visible surface point is determined, then we will have solved what is called the visible surface or hidden surface problem.

The second component, called *shading* or *local illumination,* computes how incident light is reflected, refracted, and absorbed within a small area on a surface, on the basis of its geometric and optical properties. Models have been developed, based on physics, to cover the spectrum from purely diffuse matte surfaces to purely specular mirrorlike surfaces. Diffuse surfaces are characterized by radiation heat-transfer models (blackbody radiation), and specular surfaces by geometric optics (including refraction and Fresnel effects). Modeling the range between these two extremes is achieved by combining models for the extremal cases.

Since the geometric model is defined in continuous space whereas the image is defined in discrete space as a set of pixels, care must be taken when converting between these two different fundamental forms of space. For example, an inclining line will be represented as a set of pixels that, when examined closely, will resemble a staircase instead of a smooth line. This effect is called *aliasing.* It arises from attempting to represent rates of change (frequencies) that are greater than the spacing of the pixel lattice. Visible artifacts due to aliasing can be reduced by higher sampling rates combined with low-pass filtering; this is called *antialiasing.*

2.1 Visibility Determination

Visibility can be computed in a number of ways. For example, one can directly simulate the propagation of light by constructing a ray from the viewpoint through each image point, determine the objects it intersects, and choose the object that has the smallest distance along the ray. This approach is called *ray tracing.*

An alternative is to transform the model to the coordinate system shown in Fig. 9 in which the viewing position is at the origin and the x and y coordinates correspond to the horizontal and vertical directions of the display, respectively. In addition, the image plane is orthogonal to the z axis and at unit distance from the origin. As a consequence, the z coordinate corresponds to depth. One can then apply a perspective projection to all the points in the model by mapping each point (x,y,z) to $(x/z, y/z, -1/z)$. This has the effect of mapping a ray through the image plane point P to one that is parallel to the z axis but still contains P, and also preserves the depth ordering along the ray. As a consequence of this mapping, the visible surface at image point (x,y) is the surface that has the smallest z coordinate.

2.1.1 Z-Buffer Algorithm. This algorithm is the most commonly used visible surface algorithm. It operates on any set of 3D points in the above post-perspective space whose xy coordinates lie on the pixel lattice. All points with the same xy coordinates project onto the same pixel. Since the visible surface is composed of those points that are closest, we

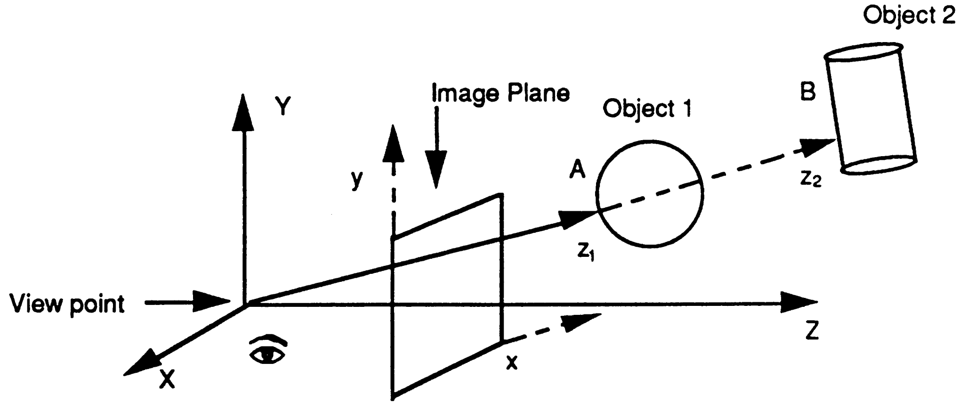

FIG. 9. Visibility: A ray intersects object 1 at distance z_1 and object 2 at z_1, $z_1 < z_2$. Only light from object 1 reaches the image plane.

simply need to sort the points into buckets, one for each pixel denoted as *pixel*[*x*,*y*], while keeping only the point with smallest depth (z value). This can be accomplished by maintaining both a color buffer, *pixel*[*x*,*y*].*color*, and a z buffer, *pixel*[*x*,*y*].*z*, which are set initially to the background color and maximum depth, respectively. Each surface is quantized, one at a time, into a set of 3D points. For a point (x,y,z), if $z < pixel[x,y].z$, then this point is closer than any previous point with those xy coordinates, and its color and z value replace the contents of *pixel*[*x*,*y*]. Otherwise, the point is known to be occluded and so is discarded. After all points have been processed, the contents of the color buffer is the desired image.

2.1.2 Scan-Line Algorithms. Here the image is treated as being composed of a set of horizontal *scan lines*, and it is synthesized one scan line at a time, going from top to bottom. For each scan line, the plane consisting of the viewpoint and the scan line is extended to determine the objects it intersects. If the objects are described by polygons, then such intersections are easy to calculate and this results in a collection of horizontal line segments, called *spans*, one span for each intersected polygon. Visibility calculation can then be carried out on these spans, thereby reducing a 3D problem to a set of 2D problems. These spans are kept in an "active polygon" list sorted left to right by their left end points. Between two consecutive end points, only one polygon can be visible. The algorithm sweeps left to right, computing at each span end point the closest span, and then drawing that span. When a scan line is completed, the next scan line in the top-to-bottom order is processed. This requires updating the active list of polygons by deleting old polygons that do not intersect the new scan line and adding new polygons that do. Polygons remaining in the list have the end points of their spans updated to reflect the change in position between two consecutive scan lines.

2.1.3 Visibility Priority. Another method is based upon visibility priority. If a ray from the view intersects two surfaces A and B such that A is closer to the viewer than B, then A is said to *occlude* B. If it is also the case that A occludes B for every ray that intersects both of them, then A is said to have visibility priority over A. If this fact is known, then it is possible to draw B entirely first, and then draw A "on top of" B; as a consequence the part of A that occludes B will overwrite it. This has been dubbed the *painter's algorithm*, alluding to the occasional practice of painters to overpaint a background with a foreground. If a visibility priority ordering of all surfaces can be generated, then drawing the surfaces "back to front"—that is, lowest priority to highest priority—will produce an image with correct visibility. If one uses the partitioning-tree representation of polyhedra, as discussed in Sec. 1.2, one can generate such a priority ordering by simply traversing the tree in an order determined by the current viewing position.

2.1.4 Ray Tracing. All three of the methods just described are not suitable for computing interobject reflections and transmittance. To accomplish this, ray tracing can be employed (Fig. 10). Here for each pixel on the image plane, a ray is propagated from the viewpoint through the pixel into the object space. Among all the objects that this ray intersects, the object that is the closest to the viewpoint is chosen. From this point of intersection, a second ray is traced toward the light sources. If this secondary ray encounters an object between the surface and the light source, then the intersection point is considered to be in shadow. Rays are also propagated in the directions of reflection and refraction. Such rays may intersect other objects and so generate other reflection and refraction rays, and so on recursively until the ray goes out into empty space or hits a light. This process creates a tree of reflection, refraction, and shadow rays forming a geometric network by which light flows back to the viewer. The intensity and chromaticity of the light is determined by the local illumination models at each surface in the network.

Ray tracing is capable of generating photorealistic images that contain most of the optical effects, except those that can only be modeled using the wave nature of light. There are two principal difficulties, however. As the number of objects in the scene increases, ray-surface intersection calculation increases, particularly if there is a large number of secondary rays. To give an indication of the magnitude of the problem, consider an image (1000×1000) containing a million pixels generated from a scene containing a thousand objects. If there are, say, three secondary rays per primary ray (a very modest estimate), and if each object is represented by a cubic polynomial, then intersection calculations could require computing the real solutions of cubics 4×10^9 times, which is quite formidable.

The second problem for ray tracing arises from the infinitely thin nature of the rays that result in point sampling of the scene. This generates aliasing artifacts. Several techniques have been proposed to overcome these difficulties. Ray-surface intersection can be simplified if surfaces that do not lie near a ray can be excluded. A hierarchy of bounding boxes that enclose objects has been used for this purpose. If a ray does not intersect a box, which is easy to check since boxes are simple, then it does not intersect any object in that box. However, if it intersects the box, then the ray is tested for intersection with each of the smaller boxes contained within this box. This process is continued until the smallest box is reached, at which point the ray is tested for intersection with all the objects in that box.

Another similar technique is to divide the object space into disjoint cells using, for example, a partitioning tree or an octree. A list of all surfaces that intersect a cell is stored with that cell. A ray can then be inserted into the tree to quickly find the cells that it intersects, and the cells can be visited in a near-to-far priority order. Ray-surface intersections are then limited to those surfaces that lie in the cells traversed by the ray, and can be performed in the near-to-far ordering. The first intersection is then the visible surface, and so the rest of the surfaces in the remaining cells need not be tested. Such algorithms

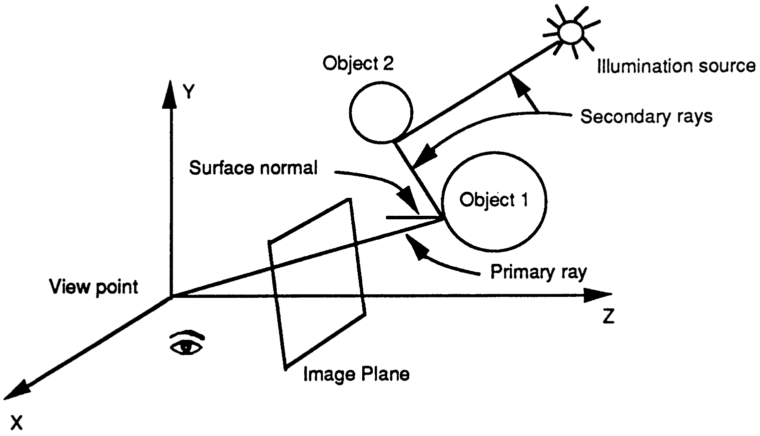

FIG. 10. Primary and secondary rays in ray tracing.

can improve the efficiency of ray-surface intersection calculations by orders of magnitude. But despite these improvements, ray intersection continues to be the principal computational burden for ray-tracing algorithms.

2.2 Illumination

Having determined the point on a surface in the object space that is visible, one must determine the color of light traveling in any particular direction, for example, from a visible surface to the viewer. This requires modeling the interaction between light incident upon a surface with the surface itself. As an example, for reflection, the intensity I in any direction can be modeled as a sum of three components:

$$I=I_d+I_s+I_a,$$

where I_d is diffuse reflection, I_s is specular reflection, and I_a is reflection due to ambient light. The diffuse component is due to light being reflected by a rough surface uniformly in all directions and is treated as being direction independent. Specular reflection corresponds to reflections along the direction of reflection, determined by the angle of incidence and the surface normal. For mirrors this is the only illumination term, but for less perfect reflectors this component is what models highlights. For matte surfaces, the specular component is 0 and reflections are modeled soley by the diffuse term.

I_a is determined by light reflected off of other nearby surfaces. This can be treated simply as being a constant level of direction-independent illumination or it can be computed exactly. *Radiosity,* originally developed for studying radiation heat transfer, can be employed for this. The radiosity of a surface is the total light emanating from it. It consists of all the reflected light and, for light-emitting surfaces (lights), all emitted light. The objects must be subdivided into surface elements that are small enough so that the radiosity term can be treated as being constant over the element. The unknown radiosities are then computed from a system of simultaneous linear equations, one for every pair of these surface elements, expressing the quantity of light transferred between the two elements.

2.3 Antialiasing

Since objects are modeled as having value-continuous, but not tangent-continuous, surfaces, they may have arbitrary sharpness; i.e., they can have edges. Sharpness represents a high rate of change as one moves, say, horizontally across an image, and this rate of change may be thought of in terms of frequency. More exactly, one can in principle perform a Fourier transform on a spatial function to obtain its representation in terms of frequencies. Thus edges correspond to spatial regions with high frequencies since the function changes very rapidly. (If the function is value discontinuous, then there is no bound on the highest frequency.) The displays, being physical systems, are capable of showing only images whose bandwidth is limited. It is, therefore, necessary to carry out low-pass filtering on the signal before it is displayed. If filtering is not done, artifacts such as jagged edges (spatial aliasing) and backward spinning of wagon wheels (temporal aliasing) are clearly seen. However, too much filtering reduces the sharpness of the images, and certain filter shapes can cause other visible artifacts (e.g., ringing). Thus the filter design represents a trade-off between a number of subjective factors such as sharpness and visibility of aliasing artifacts.

Prefiltering a signal by an amount such that the bandwidth of the resulting signal makes the sampling rate a Nyquist rate is a preferred mode of overcoming aliasing. In a number of situations, this can be easily done. However, in ray tracing such prefiltering may require computing a large number of samples by tracing many rays. It is typical, for example, to compute 8 or 16 ray-traced samples for each pixel and then filter them. Another method is to filter the objects directly in 3D space to make them smoother. Unfortunately, such filtering, if properly done, must depend on the direction and distance of the viewpoint and the image plane relative to the object. Moreover, only simple surfaces such as spheres and polygons can at this time be handled, and special effects such as shadows and transparency are quite difficult. A preferred approach is to do adaptive sampling of the image plane. In areas of high detail (or uncertainty) the density of the rays is increased and the location of the rays is randomized. The former method allows more local bandwidth where needed, and the latter makes aliasing noise more random and makes it contain higher frequencies, thus reducing its visibility.

2.4 Surface Color Variations

Visibility calculation and shading described in the last section can, in many cases, create images that look artificial because their surfaces look too smooth. Metallic objects may look fine, but natural objects such as fruits and human skin will not. Realism can be increased in many cases by adding detail, i.e., texture, on the surface of the objects. This is usually done either by adding a prespecified two-dimensional pattern of colors on top of a smooth surface or by making the surface rough. The former can be considered as a mapping of a 2D "image" onto a 3D surface with associated warping of the image, whereas the latter can be modeled as randomization of one or many of the surface attributes, such as the reflection property or surface normal. Since the addition of texture patterns on the object's surface can be thought of as a mapping from texture space onto the object space, it can be done in one of two ways:

1. compositing the mapping from texture space to object space with the mapping from object space to the image plane used in visibility calculation; or
2. utilizing an inverse mapping from the image plane to texture space.

A different approach to varying the color across a surface is to interpolate the color. As mentioned earlier, nonlinear surfaces can be approximated by polygons in order to reduce the complexity of the algorithms dealing with such surfaces. However, the illumination will be discontinuous at polygon edges even though the nonlinear surface would have been continuous in this region. If, however, the color is known at the vertices, then these values can be easily interpolated across the polygon. This guarantees that two polygons that share an edge will have the same color along that edge, and so the image will remain value-continuous at any edge not lying on the silhouette. This method is called *color interpolation* or *Gouraud shading.* A similar idea is to interpolate surface normals defined at the vertices across the polygon and compute the illumination using these normals at each interpolation point. This method is called *normal interpolation* or *Phong shading,* and it allows polygonal models to display highlights.

3. ANIMATION

Computers have been used in creating image sequences showing animation. Techniques of the previous section can be used to create a set of *key frames* in an image sequence. Key frames are those frames where a discontinuity in motion or content occurs during the animation sequence. Computer algorithms are then used to create the "in-between" frames. Smooth trajectories of motion are traditionally used to compute these in-between frames.

In traditional commercial animation, the motion is selected directly by the animator without any explicit reliance upon physics (other than what is contained in the animator's mind). In contrast, *physically based modeling* incorporates the physics of moving bodies explicitly using kinematics or dynamics to determine differential equations for modeling motion. Here the objects, environment, motion equations, and constraints are modeled directly and can be used to generate a significant quantity of the motion in an animation. Fairly sophisticated motion models, such as for the motion used in walking and running by human beings, have been developed.

4. HARDWARE

Computer graphics has from its outset required computing hardware specific to the task of generating images. This arises both from the basic requirement of displaying an image (for example, on a TV screen), and from the performance requirements needed to synthesize the image in an acceptable length of time. For real-time applications, such as simulators for aircraft training, acceptable means 60 images per second, while interactive applications require at least 1–6 images per second. Given the significant performance requirements, hardware designers have turned to creating special-purpose VLSI (very large scale integration) ICs and to designing parallel computer architectures.

A digital image is represented by a 2D array of color values, called *pixels* or *pels.* Each pixel value encodes the color to be drawn at the corresponding point of the display surface. It may be only a single bit if a bilevel black and white display is used; or it may be as many as 24 bits if full color is desired, so that the three primary colors of red, green, and blue (RGB) each have 8 bits, for a total of 24 bits per pixel. Alternatively, one can add one level of indirection and let the value at each pixel be an index into a 24-bit wide *color table*. This allows, for example, a system with only 8 bits per pixel to select, at any one time, any 2^8 colors from a palette of 2^{24} possible colors. This can be used to introduce color into a grey-scale picture (called pseudocolor), as is common with satellite images.

For display on a standard TV monitor (NTSC), a typical resolution is 640 pixels horizontally by 480 pixels vertically for a total of $640 \times 480 = 307{,}200$ pixels displayed at 30 Hz. Higher-resolution monitors are widely used, and are commonly $1280 \times 1024 = 1\,310\,720$ pixels displayed at 60 Hz. Thus, a full-color high-resolution display requires almost 4 Mb of memory accessed at a rate of around 250 Mb/s. Current memory technologies cannot easily provide this video refresh rate in addition to the image-drawing access rate, so special memory chips called VRAMs (video random-access memory) have been designed that provide an additional access mechanism specifically for video refresh. Many of these used in parallel (e.g., 128) can provide the required bandwidth over today's bus technology. The specific block of memory formed from these VRAMs that is used to hold the pixels being drawn on the CRT is called the *frame buffer*, since it holds at any one time a single frame/image of a sequence of frames/images.

Rendering the geometric model at high rates has always relied on specially designed graphics systems. Performing the basic operations of transforming, clipping, and drawing visible polygons, in that order, has led to a number of pipelined architectures, using a variety of hardware implementations. One such design used as many as 16 copies of a graphics VLSI processor to implement the transformation and clipping operations, while an alternative implementation used nine programmable DSPs (digital signal processors) to provide the same functions. Similarly, the drawing of visible polygons (possibly with texture mapping and individual shading calculations for each pixel) has been performed by specially designed VLSI chips (e.g., 20) or programmable DSPs (up to 64). The image computation is partitioned among the k drawing processors so that each is responsible for one out of every k pixels. To obtain a uniform load balancing among the drawing processors, the pixels assigned to a single processor are distributed uniformly throughout the image.

5. APPLICATIONS

Since geometry and images are fundamental components of our experience, it is not surprising that computer graphics has been applied widely. The most common and simplest use is the depiction of statistical data using graphs, bar charts, pie charts, etc., and the creation of illustrations to accompany text, such as the figures in many papers and books. This kind of applying computer graphics carries on activities in the computational domain that were mature long before computers. Since by and large they represent a transference of manual tasks, their computational complexity is not great, and the principal software issues are those dealing with the design of good user interfaces.

Engineering has from the outset been a significant consumer of computer graphics. Electrical engineering uses graphics for designing printed circuit boards on which the IC chips are mounted, for designing the logic diagrams specifying the function of a circuit, and for creating the physical layout of the ICs themselves. All of these, however, are treated as multiple layers of 2D sheets, even though printed circuits and chips are 3D physical objects.

Mechanical, aerospace, civil, chemical, and industrial engineering are concerned with the creation of physical objects that are inherently three-dimensional. To exploit computer graphics fully, interactive design capabilities are essential, as well as nongraphics computations that determine physical properties such as mass, effects of stress, and heat-transfer properties. In certain cases, the designed object can be manufactured directly from the computational representation using numerically controlled milling machines, or, if the target material is plastic, "drawn" directly using lasers.

Closely allied to the needs of these engineering disciplines are those of architecture. In this case, the esthetic dimension is of greater importance than physical properties, and much more of the design is decided by visual criteria as opposed to functional requirements, such as heating and power distribution. Since architecture is concerned with designing spaces for people to be in, having the ability to "walk through" a computer model of a building is considered to be an important capability, a goal yet to be satisfactorily attained. Ideally, the user would be completely enveloped in a *virtual reality* by wearing a head-mounted display. Such a display has separate images for each eye, forming a stereo pair, and everything in the person's field of vision is computer generated. The position of the head is tracked as the person moves, and the images correspond to what would be seen if the environment were physical, thus giving a feel of authenticity to the environment.

Physical scientists are increasingly using computer graphics both to visualize data and to observe simulations of physical processes, which has been dubbed *scientific visualization.* This can provide considerably more insight into the nature of a physical process than just the equations that describe state transitions. It can also provide access to processes whose space-time scale is significantly different from ours. Thus astronomers can see a process entailing millions of light-years develop in a few minutes of computer animation, and similarly for processes at the atomic level involving only a microsecond or less.

Another important area of data visualization is medical imaging. Various sensing devices, such as CAT scan or MRI, produce a set of slices where each slice is a 2D array of values correlated to the kind of tissue at each point in the sample. These slices, when stacked on top of each other, form a 3D array of voxels (volume elements) showing the part of the body that was scanned. Medical diagnosis can then be greatly enhanced if it is possible not only to see a static view of this data, but to manipulate it interactively as well. Providing such capabilities is called *volume rendering.*

Since computer graphics generates pictures, one should expect that it could be of great use to the visual arts. This is indeed the case, and has been since its very beginnings. Initially, much of this work was 2D in character, emulating painting, and in fact many computer paint programs have been written. More recent efforts have included 3D sculptures seen through stereographic viewers. Another visual art form, film, has its computational correlate, namely computer animation, which has been used for both artistic and commercial purposes. Much use has been made in the making of commercials for TV and for traditional graphics, such as the introductions to TV shows (referred to as "flying logos"). More ambitiously, several full-length movies have employed computer graphics to generate an appreciable percentage of the footage, while still more have utilized it for special effects.

GLOSSARY

Affine Subspace: Any subspace defined by the intersection of hyperplanes; *flat* is an alternative name.

Affine Transformations: Transformations that preserve flatness (linearity): translation, rotation, scaling, and shearing.

Algebraic Set: Any set of points that satisfies a polynomial (i.e., its solution set).

Antialiasing: Various filtering operations to remove the artifacts due to discretizing continuous functions, such as those that define curves and surfaces. A discrete image of a line without any filtering appears "jagged" or "staircased."

Binary Space Partitioning Tree: A computational representation of polytopes that partitions space into disjoint sets of convex cells. Each cell is in the interior or the exterior of the polytope.

Boundary Representation: A computational representation of polytopes in which there is a one-to-one correspondence between faces of the polytope and elements of the representation.

Constructive Solid Geometry: Defining 3D solid objects using set operations, affine transformations, and instancing (copying).

Diffuse Reflection: Reflection from a rough or matte surface that is perceived as uniform in all directions within the hemisphere above the surface.

Frame Buffer: A special 2D array of memory containing the image that is displayed on a CRT monitor 30–60 times a second.

Geometric Model: The computational representation of geometric objects and any of their attributes such as color.

Homogeneous Coordinates: Coordinates used in projective space; a normal 2D point (x,y) has in projective space the coordinates (wx,wy,w) for any $w \neq 0$.

Hyperplane: Dimension-independent entity defined by a single linear equation. Examples are a plane in 3D and a line in 2D. Always one dimension less than the embedding space.

Illumination: Modeling the interaction of light with surfaces. Local illumination is concerned with how the light that is incident upon a very small area is reflected, refracted, and absorbed. Global illumination is concerned with the interaction of light among all of the surfaces.

k Flat: A set in d-dimensional space of dimension k, $0 \leqslant k < d$, that can be defined by the intersection of $d-k$ hyperplanes or as all the weighted (affine) combinations of $k+1$ points. For example, a plane is a 2-flat and a line is a 1-flat regardless of the dimensionality of the space.

Nonlinear Sets: Sets defined by nonlinear equations.

Nyquist Rate: The minimum rate at which a continuous function can be sampled and then subsequently reconstructed without introducing errors (aliasing). That rate is twice the highest frequency present in the function.

Parametric Set: A set of points defined by a one-to-one mapping of a parameter t to a point p, where t is a vector whose dimension determines the dimension of the set. The parametric representation of a line through the points p_0 and p_1 is $p=tp_1+(1-t)p_0$.

Pixels: The points of color out of which an image is composed when represented in discrete space. The pixels lie on a uniformly spaced 2D array.

Polynomial: An expression that is the weighted sum of a set of monomials. A monomial is a constant times any product of coordinate variables where each coordinate is raised to some nonnegative integer power. Example: $x^2y^3+3x+2y^2-5$ ("missing" coordinates in a monomial have been raised to the zeroth power).

Polytope: Any set of points with an interior and an exterior, and whose boundary lies entirely in hyperplanes. The boundary is composed of a set of faces. A polyhedron is a 3D polytope, and a polygon is a 2D polytope.

Ray Tracing: Rendering a scene by tracing rays throughout the scene, analogous to geometric optics. The rays originate from lights and viewing positions, and are reflected, refracted, and absorbed by surfaces.

Rendering: Modeling the propagation of light and its interaction with surfaces to generate an image of a geometric model.

Set Operations: Forming the union, intersection, or difference of two sets.

Specular Reflection: Reflection of light from perfectly smooth surfaces; the light is concentrated along one direction, as with a mirror.

Splines: Curves that are comprised of pieces, each defined by a polynomial, in which the pieces fit together to form a smooth curve (tangent or curvature continuous).

Texture Mapping: Defining the color of a surface by a function that varies the color across the surface. A common example is mapping an image onto a surface lying in 3D.

Virtual Reality: Computer-generated sensory input correlated to human motion so that the subjective experience is interpreted as having a realistic quality.

Visibility: Determining for each member of a set of rays that surface the ray intersects first, i.e., what is the visible surface along these rays.

Volume Rendering: Generating images of 3D data that was produced by sampling space using electromagnetic radiation. Common example is medical data generated by scanning a body.

Further Reading

Farin, G. (1988), *Curves and Surfaces for Computer Aided Geometric Design,* San Diego: Academic.

Foley, J., van Dam, A., Feiner, S., Hughes, J. (1990), *Computer Graphics Principles and Practices,* Reading, MA: Addison-Wesley.

Glassner, A. (Ed.) (1989), *An Introduction to Ray Tracing,* San Diego: Academic.

Hearn, D., Baker, P. (1986), *Computer Graphics,* Englewood Cliffs, NJ: Prentice-Hall.

Mantyla, M. (1988), *An Introduction to Solid Modeling,* New York: W. H. Freeman.

COMPUTER HARDWARE

KENNETH H. SLATER, *Semiconductor Engineering Group, Digital Equipment Corporation, Hudson, Massachusetts, U.S.A.*

Introduction 145
1. **General Aspects** 146
1.1 System-Wide Characteristics and Metrics 146
1.1.1 Data Representations 146
1.1.1.1 Data Representation Standards 147
1.1.2 Word Width or Word Length 147
1.1.3 Addressing Characteristics 148
1.1.4 Performance Metrics 149
2. **Digital Computer Components** 150
2.1 The Central Processor 150
2.1.1 Operation of a Basic CPU 150
2.1.2 RISC vs CISC 151
2.1.3 Pipelining and Super Scalar (Multiple Issue) 152
2.1.4 Floating-Point Processors 152
2.1.5 Vector Processors 152
2.1.6 Multiprocessors 153
2.1.7 Interrupts 153
2.2 Main Memory 153
2.2.1 Main Memory Organization and Interleaving 154
2.2.2 Parity and Error-Correcting Code (ECC) 154
2.2.3 Memory Mapping 155
2.2.4 Virtual Memory 155
2.2.5 Caches 156
2.3 Busses 156
2.3.1 Standardized Busses 156
2.4 Input/Output 157
3. **Peripheral Devices** 157
3.1 Persistent Mass Storage 157
3.1.1 Magnetic Disks—Floppy and Hard 157
3.1.2 Optical Disks 158
3.1.3 Magnetic Tape 158
3.2 Communications and Ports 158
3.2.1 Serial and Parallel Ports 158
3.2.2 Modems 158
3.2.3 Networks 159
3.3 User Interface Devices 159
3.3.1 Hard-Copy Devices—Printers and Plotters 159
3.3.2 Soft-Copy Devices—Terminals and Displays 159
3.3.3 Input Devices 159
3.3.4 Optical Scanners 160
4. **Summary** 160
Acronyms 160
Glossary 160
Works Cited 161
Further Reading 161

INTRODUCTION

Computers are now ubiquitous in industrialized society. While the cost, performance, and capacity of computer hardware span several orders of magnitude, the vast majority of machines are stored-program digital computers built of common subsystems. Our focus will be to describe and discuss hardware aspects of these machines, including common contemporary peripheral devices.

Computers have helped fuel the worldwide electronics industry, providing a ready market for semiconductors, especially dynamic memories, which are the key to driving advanced semiconductor technology. Advances in semiconductors have, in turn, helped create the worldwide consumer electronics market. With the advent of the personal computer, computer hardware has become a mass market item, and consumer electronics technology such as compact disks and digital audio tapes are now found in both industrial and personal computers. Over all, the demand for computer hardware has helped create and shape the worldwide electronics economy as we know it today.

Preceded by many mechanical devices and electronic analog computers, the first generation of electronic digital computers was built

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

with vacuum tube technology in the 1950s. The advent of the discrete transistor ushered in the second generation of cheaper, smaller, and more reliable machines in the 1960s. The third generation in the 1970s introduced minicomputers, made practical by small-scale integrated circuits. With the advent of large-scale integrated circuits in the 1980s, the personal computer and workstation emerged as the leading products of the fourth generation. The fourth generation also saw the widespread use of high-speed networks and the addition of optical storage devices in addition to traditional magnetic storage media. The fifth generation of machines, still in their infancy with no one architecture having emerged as dominant, will likely take advantage of the low cost and high density of very large-scale integrated circuits and rely on multiple-processor techniques. Different and promising approaches include massively parallel machines, neural networks, systolic arrays, and data-driven machines (see NEURAL NETWORKS; PARALLEL COMPUTING).

1. GENERAL ASPECTS

Today's general-purpose digital computers, be they common personal computers or large supercomputers, have many hardware similarities (Fig. 1). They all incorporate one or more central processing units (CPUs) which execute the instruction stream and process data. The CPU retrieves data and instructions from a main memory, and writes modified data back to it. The ability to interpret items in memory as either instructions or data is a key attribute of a stored-program digital computer.

Often the CPU is much faster than the main memory, and a memory cache is employed to match CPU and memory speeds better. Data travel between main memory and the CPU over a pathway called a bus, to which input/output (I/O) adapters (also called peripheral controllers or adapters) are also attached. These controllers connect the CPU and memory system to other devices such as user interfaces, data storage, or communications devices.

Contemporary computer hardware relies on fast and dense electronic integrated circuits for processing logic and main memory. Rigid magnetic or optical disks are used for persistent mass storage, and flexible magnetic media, be they floppy disk* or magnetic tape, for archival storage and data transport. The primary human interface is usually a cathode-ray-tube display, with an electromechanical keyboard and pointing device (such as a mouse or tablet) for human interaction. Despite this common technology, the performance, capacity, and cost of these machines vary widely. The number, size, speed, and configuration of these hardware elements determine the overall characteristics and cost of the computer system.

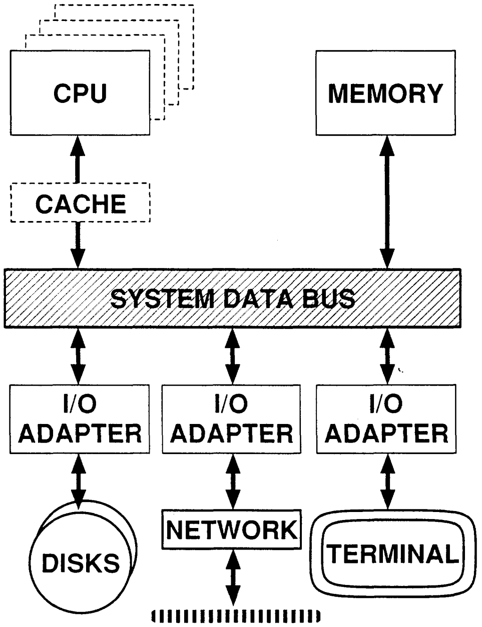

FIG. 1. The major functional blocks of a typical stored-program digital computer. Optional components are shown in dashed lines. Not shown is the possibility of separate memory and I/O busses, used to improve performance by not burdening the CPU memory bus with lower speed I/O traffic.

1.1 System-Wide Characteristics and Metrics

1.1.1 Data Representations. Computer hardware directly interprets only a small number of data representations; more complex data representations are created by software building upon these basic entities.

---

*Some inexpensive or older personal computers use floppy disks for persistent magnetic storage instead of a hard disk.

The fundamental unit of information in a digital computer is the bit (short for Binary digIT). A bit can have one of two values, 0 or 1, and easily maps to mechanisms in the underlying technologies, such as the presence or absence of electrical charge or the polarity of a magnetic field.

In order to represent data that has more than two values, bits must be aggregated into sets that can represent computer instructions or data. There has been widespread convergence towards representing data in multiples of 8 bits, called a byte, which has one of 256 (2^8) values. Common uses for the byte include

1. representing one of 256 possible characters,
2. representing an unsigned integer in the interval of 0 to 255,
3. representing a two's-complement signed integer in the interval of -128 to $+127$, and
4. representing a simple computer instruction.

Sets of bytes are used to represent larger integers, floating-point data, memory addresses, or complex computer instructions. Common sizes for integer data are 2, 4, and 8 bytes, and floating-point data are typically 4, 8, or 16 bytes in length. While these types are nearly universal, many machines implement additional hardware data types, such as linked lists, character strings, or decimal integers.

1.1.1.1 Data Representation Standards. Character data first became widely standardized with adoption of the American Standard Code for Information Interchange (ASCII), which was based upon a 7-bit character code (ANSI, 1977). While ASCII is still in widespread use, more recent standards issued by the International Standards Organization are beginning to supersede it. ISO Standard 8859 offers 1, 2, or 4 byte-per-character representations (in ISO terminology, an 8-bit grouping of bits is an Octet) which permits simultaneous use of many character sets, a growing requirement in international markets and in desktop publishing. For example, ISO Latin-1 is a one-octet character set in which characters 32 through 126 map directly to the popular 7-bit ASCII set; other standardized single-octet character sets include Arabic, Greek, Cyrillic, and Hebrew. Many Asian character sets, with their rich store of characters, are standardized as two-octet sets, including Katakana, Kanji, Hanzi, and Korean.

Integer data have not been formally standardized, but a *de facto* standard does exist. The vast majority of current computing hardware uses integers sized in multiples of 8-bit bytes (8, 16, 32, 64, and 128 bits being most common) for both signed and unsigned data.

Two's-complement notation for signed integer data has become standard in the industry, being selected over one's-complement and sign-magnitude notations used in some earlier machines. It has the advantage of being able to perform arbitrary signed operations on the same hardware that handles unsigned addition, and it has a unique representation of zero (the other representations have both a positive and negative zero). Given an N-bit word, two's-complement notation can represent numbers in the range of -2^{N-1} to $+2^{N-1}-1$ (for example, an 8-bit byte can represent -128 to $+127$). To convert a number to two's-complement notation, convert the absolute value of the number to an unsigned binary integer. If the original number was negative, then invert (logically complement) each bit and add 1 to the result.

Floating-point data format and floating-point data operations have been standardized through the Institute of Electrical and Electronics Engineers (IEEE) standard 754, commonly called IEEE Floating Point (IEEE, 1985). While many proprietary formats are still in widespread use, largely for historical reasons, the IEEE format is steadily becoming accepted and is supported by most newly designed computer architectures. The standard defines

1. single (32-bit) and double (64-bit) precision basic formats, and single ($\geqslant$ 43-bit) and double ($\geqslant$ 79-bit) precision extended formats (Fig. 2);
2. add, subtract, multiply, divide, square root, remainder, and compare operations;
3. conversions between
    a. integer and floating point formats,
    b. different floating point formats, and
    c. basic floating point formats and decimal strings; and
4. floating-point exceptions and their handling.

1.1.2 Word Width or Word Length. The word width (word length is used interchange-

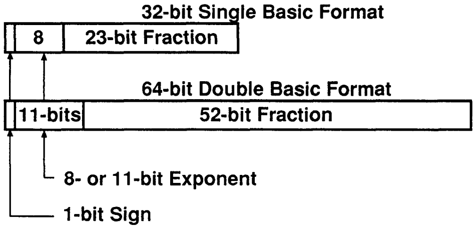

FIG. 2. The sign, exponent, and fraction fields of the IEEE-754 single and double precision basic floating-point data formats. The standard defines the most significant bits of each field to be on the left.

ably) indicates how many bits of data can be processed at once; the wider the word, the higher the performance of a machine compared to a similar machine with a smaller word width. Also, larger word widths provide added precision or range of values in computations directly (the same precision can be achieved by using multiple smaller words and appropriate software; the penalty is reduced speed). Typical word widths in current machines are 8, 16, 32, or 64 bits. Word widths may vary between different parts of the machine; currently popular personal computers based on the 386SX microprocessor have a 32-bit word width in the central processor, a 16-bit word width to memory, and 16- or 8-bit word widths to peripheral devices. Reducing the word width reduces required hardware, reducing hardware costs, but also reduces overall system performance.

1.1.3 Addressing Characteristics. Addressing granularity, or the smallest addressable unit, describes the smallest unit of memory that may be uniquely addressed. Current practice is to address memory at the byte level, which greatly facilitates character processing. This usually does not preclude addressing and accessing larger units, including and up to the word width of the machine.

There are two different conventions for addressing bytes within multiple-byte words (which includes integer, floating-point, and address data): Does the address indicate the most significant bits of the word or the least? Because of a famous paper by Cohen (1981) these two conventions have come to be known as "Big Endian" and "Little Endian." While on an isolated computer either convention works just as well, there are problems to overcome when exchanging programs or data between machines that use different conventions. Both conventions are in widespread use on contemporary machines.

Data alignment is an important, and sometimes restrictive, characteristic of how data may be accessed in memory. Memory subsystems are designed with a natural word width, usually in multiples of 2, 4, 8, or 16 bytes. Access to full words that are aligned to occupy a single word in memory is naturally fast and efficient. Unaligned memory access to full words requires access to two aligned memory locations, selection of the proper bits from each aligned word, and assembly of the bits into the requested unaligned word, as illustrated in Fig. 3. Needless to say, support of unaligned access to memory requires additional hardware and/or software and almost always incurs a performance penalty.

Some computers, particularly the newer reduced instruction set computer (RISC) designs, do not provide any hardware support for unaligned memory access, and if it is supported at all it is usually done with low-level software. Other machines, including several popular complex instruction set computer (CISC) models, provide transparent hardware support for unaligned access, although in most cases aligned data has a noticeable performance advantage.

Total address space is another key characteristic of a machine and is the number of unique addresses that are provided by mem-

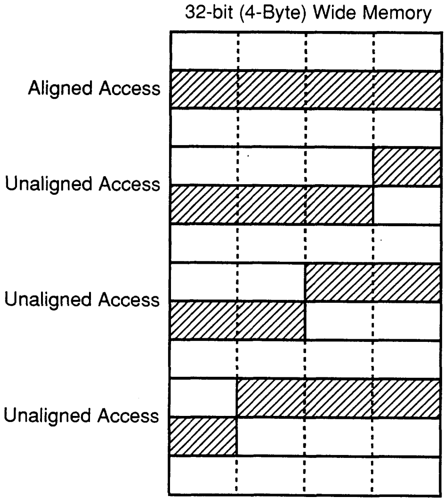

FIG. 3. Aligned and unaligned memory access. Depiction of an aligned memory access of a 32-bit word from a 32-bit byte addressable memory. Also depicted are the three possible cases of unaligned access.

ory. This has direct impact on the data and instruction capacity of the machine. Machines that utilize a 16- to 32-bit address space are very common today, with 48- and 64-bit address machines now arriving on the market. The following table indicates the relationship of the number of bits in the machine address and the number of uniquely addressable items in memory:

Bits in Address	Uniquely Addressable Memory Items
16	65 536 (2^{16})
32	4 294 967 296 (2^{32})
48	281 474 976 710 656 (2^{48})
64	18 446 744 073 709 551 616 (2^{64})

1.1.4 Performance Metrics. "How fast is that computer?" This is a complicated question to answer; even if theoretical or measured performance characteristics of the major subsystems are known, at best they will provide an upper bound of the machine's overall performance. The actual performance of the system is a function of the application to be run, its data, the compilers and operating system being employed, the hardware configuration, and operating environment overhead factors. Benchmark programs have had a particularly difficult time keeping up with the rapid increases in main-memory size and cache sizes; many benchmark programs now fit completely within the machine's high-speed caches and give unrealistically good results (an actual application would likely not be able to fit entirely within the fast cache memory, and would perform significantly slower, because of cache misses which cause reference to the slower main memory). Despite these difficulties, computer users and vendors still like to sum up their system performance in a simple numerical rating.

Millions of instructions per second (MIPS) attempts to quantify the performance of a computer by how many instructions it can execute per second. Computations performed per instruction is not well defined and varies widely over different hardware. Instruction mix and frequency varies widely over different workloads. With these ambiguities, the MIPS number is often ridiculed as the "meaningless indicator of performance." Nonetheless, the MIPS rating of a computer is still often quoted and serves at least as a first-order approximation to relative performance.

Millions of floating-point operations per second (MFLOPS) is used to quantify overall floating-point performance (as MIPS quantifies integer and logical performance). MFLOPS numbers are also widely quoted. As all current computers have comparable sets of floating-point instructions, the use of MFLOPS as a performance indicator between different machines is less problematic than the use of MIPS is for integer and logical performance.

In the personal computer area, clock speed in megahertz is often used as a performance indicator. This has worked well to date as an overall indication of performance between different models of PCs because most PCs are based on a common family of processors and have common hardware configurations; in general the faster the clock runs, the faster the processor will execute instructions and process data, assuming the overall system is reasonably well balanced to operate at that clock speed. Between processor families, however, the number is largely meaningless.

Recently, the Systems Performance Evaluation Cooperative (SPEC) was formed by hardware vendors to promote the use of standardized benchmarks of representative problems of realistic size. A suite of benchmarks which stress different performance aspects and applications of a computer system are provided, with the results summarized by a single figure of merit, the SPECmark.

The SPEC suite consists of four integer-intensive benchmark programs written in C and six floating-point-intensive benchmarks written in FORTRAN. SPECmark numbers are reported for each individual benchmark program, and are combined via a geometric mean to give integer, floating-point, or overall computer system performance. The overall performance of a Digital VAX-11/780™ using VMS™ compilers is defined to be 1.0 SPECmark; in late 1991 the fastest RISC workstations are reporting SPECmarks of about 70, current model VAX processors are about 40, and an Intel™ 50-MHz 486™ is near 18 SPECmarks. See Weicker (1990) for further information on benchmarks and performance measurement.

2. DIGITAL COMPUTER COMPONENTS

2.1 The Central Processor

The central processing unit (CPU) executes a program stored in main memory, manipulating data to achieve a desired calculation. A minimal CPU will have an instruction unit that retrieves instructions from the memory and controls the elements of the CPU to execute the instruction; high-speed temporary storage called registers or accumulators to store intermediate calculations; and an arithmetic/logic unit (ALU) which actually performs the logical or arithmetic operations on the data. The program counter is a register that holds the address of the next instruction in main memory to be executed; a change in program flow (a branch or jump) can be accomplished by modifying its value. A program status word (PSW) register is usually implemented to record specific state information, such as whether the last add instruction resulted in carry or if an interrupt should be generated on a divide by zero operation.

2.1.1 Operation of a Basic CPU. The following example illustrates the operation of a basic CPU, as shown in Fig. 4. A RISC-like machine is used for this example (see Sec. 2.1.2 for a discussion of RISC and CISC machines). The example illustrates the operation of the two instructions:

1. load the contents of a word of memory into a CPU register, and
2. add that register together with a small constant integer and store the results into another register.

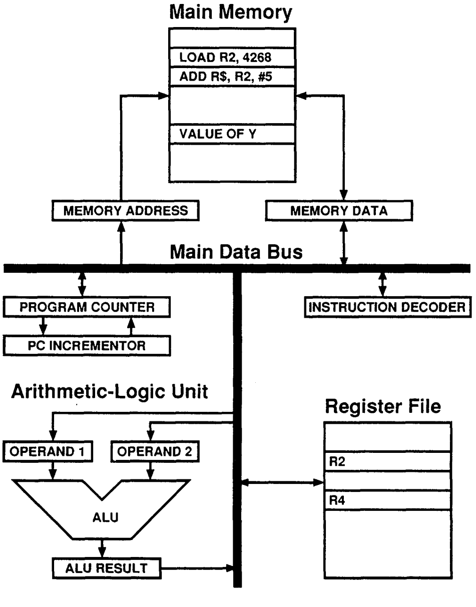

FIG. 4. The CPU used in the example of instruction execution in Sec. 2.1.1.

In a high-level programming language, the programmer might express this operation as $X = Y + 5$. After compilation to machine instructions, this statement is comprised of two instructions:

1. Load the value of variable *Y*, which is stored in main memory at a known address, into a CPU register, and
2. add the value of *Y*, now in a register, with the integer constant 5 contained in the add instruction itself, and place the result in another CPU register (which holds the value of *X* in our example).

In summary, we are assuming the following locations of integer data:

Y in main memory at address 4268, temporarily stored in CPU register *R*2.
X stored in CPU register *R*4.
5 a constant, embedded in the add instruction as an immediate operand.

The first instruction is LOAD *R*2, 4268. This instructs the CPU to load the value stored at main memory location 4268 into CPU register *R*2. Our example starts with the program counter register in the CPU containing the address of the example LOAD instruction and proceeds as follows:

1. The CPU uses the value of the program counter to fetch (read) the next instruction from main memory. It sends the value of the program counter to the memory address register, and the main memory returns the instruction to the memory data register. The memory address register and memory data registers are invisible to the programmer, as are several other registers we will encounter in this example, being part of the internal workings of the CPU.
2. The program counter is incremented by the length of an instruction, so that it now holds the address of the next instruction to execute. Typical of many machines, the program counter is incremented with a dedicated piece of hardware to make this

frequent operation efficient.

3. The fetched instruction is moved from the memory data register to the instruction decoder. The instruction is a "load from memory to register" instruction. The decoding circuits separate out two important fields of information within the instruction: The memory address to read from (4268) and the destination register ($R2$).
4. The decoded memory address is used to read the value of the data from memory, by moving it from the instruction decoder to the memory address register. The returned data from memory is temporarily held in a memory data register awaiting storage.
5. The data returned from memory is moved to the specified destination, CPU register $R2$ in the register file (an array of programmer-visible registers).

The second instructions is ADD $R4$, $R2$, #5. This instructs the CPU to add the contents of CPU register $R2$ to the integer constant held in a field of the instruction (in this case with a value of 5) and to place the result in CPU register $R4$.

1. The CPU uses the value of the program counter to fetch (read) the next instruction from main memory.
2. The program counter is incremented by the length of an instruction, so that it now holds the address of the next instruction to execute.
3. The fetched instruction is decoded inside the CPU. It is an "add constant to register" instruction. The decoding circuits separate out three important fields of information within the instruction: The value of the constant (5); the source register ($R2$); and the destination register ($R4$).
4. The value of the constant is sent to the first operand holding register of the arithmetic logic unit (ALU).
5. The value in the source register is sent to the second operand holding register of the ALU.
6. An ADD operation is done by the ALU, and the result is temporarily held in the ALU result register.
7. The value in the ALU result register is stored in the destination CPU register ($R4$).

As the above simple example illustrates, even the simplest instructions require several steps, the movement of data within the machine, and the control of various registers and logic units.

2.1.2 RISC vs CISC. Until the middle of the fourth generation of computers, the richness and complexity of the instruction set understood by the CPU continued to grow. Machines with literally thousands of instructions were designed and are in widespread use. These were later dubbed CISC after RISC appeared in the mid-1980s. RISC computers gain a significant speed advantage over CISC machines by keeping both their CPU and instruction set simple and uniform. They usually employ large numbers of registers to improve performance further. The availability of high-level language optimizing compilers allows programmers to be productive while shielding them from the simple yet often difficult to program underlying machine instructions. The compilers could effectively optimize the small set of fast RISC instructions, while they often could not effectively optimize or even use many of the large set of CISC instructions. RISC machines are now broadly accepted and are the basis for most new designs.

CISC machines are characterized by

1. variable-length instructions formats,
2. a large number of often specialized instructions,
3. a small number of general-purpose registers ($\leqslant 16$),
4. performing arithmetic and logic operations directly to memory, and
5. instructions that have multiple results, such as adding two numbers and incrementing an address of one of the operands in the same instruction (autoincrement)

RISC machines are characterized by

1. fixed-length instructions formats, often only of one basic size,
2. a small number of very general instructions,
3. a large number of general-purpose registers ($\geqslant 32$),
4. arithmetic and logic operations performed only on registers; operands must be loaded and stored from memory with separate instructions, and
5. instructions having only one result with no other side effects (change of machine state).

2.1.3 Pipelining and Super Scalar (Multiple Issue). Pipelining is a technique to improve CPU performance by processing more than one instruction at a time. A typical instruction has several distinct phases of execution: fetch the instruction; decode the instruction; fetch the operands; process the operands; and store the result. In many CPUs, these various phases are performed by distinct pieces of hardware that sit idle during part of the instruction execution cycle. A pipelined CPU will process multiple instructions simultaneously, by overlapping the end of one instruction with the beginning of another. A continuous flow of instructions moving through the CPU "pipeline" overlapped by phase keeps more of the CPU hardware productive, improving performance (Fig. 5). The simplicity of RISC instructions allows pipelines to be built generally with less hardware and to perform more effectively than comparable CISC implementations.

A super scalar, or multiple-issue, machine has several replicated elements, and can process and complete multiple instructions simultaneously. While the pipelined machine processes instructions sequentially with partial overlap, the multiple-issue machine processes instructions in parallel with full overlap (Fig. 5).

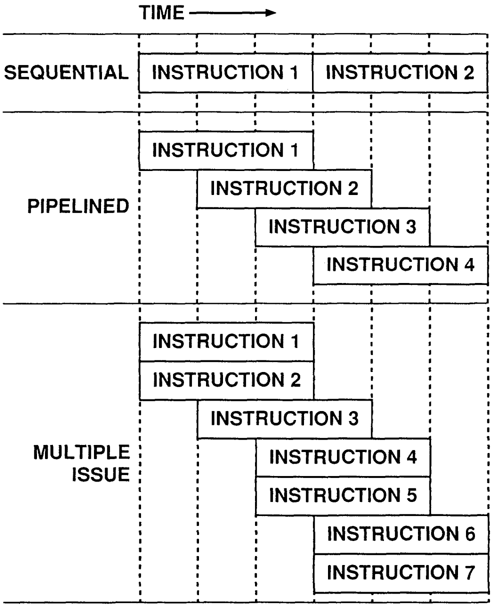

FIG. 5. Instruction execution sequences in sequential, pipelined, and multiple-issue machines. Only one of two possible instructions is being issued in the multiple-issue machine in the second time period, presumably because of a dependence that prevents instruction 4 from being issued with instruction 3.

The standard sequential processor executes only one instruction at a time. In a pipelined machine, instruction execution is overlapped to improve performance, taking advantage of the inherent multiple phases to complete most instructions. Super scalar machines have sufficient additional hardware to process multiple instructions in parallel through multiple pipelines. In both pipelined and super scalar machines, interinstruction data or control dependencies usually limit the theoretical maximum performance, as some instructions must wait for results of previous instruction to become available.

2.1.4 Floating-Point Processors. Floating-point processors are hardware especially designed for floating-point computations; they perform floating operations which the main CPU's arithmetic and logic unit cannot do (it is designed for integer and logical operations). A floating-point processor typically has its own set of registers and control unit, and the main CPU "hands off" entire floating-point instructions to the floating-point processor for rapid execution. Many floating-point units are designed to let the main CPU continue to process other instructions while the floating-point unit completes the floating-point operation independently. Current practice is for most computers to include floating-point processors; low-end personal computers are about the only common machines in which a floating-point processor is optional. When it is not present, floating-point operations can be performed in software by the integer unit at a considerable performance penalty.

2.1.5 Vector Processors. Vector processors are special hardware that can process vectors of data faster than a general-purpose CPU. A vector processor has its own vector register set, with each register holding from 64 to 256 vector elements. A single vector-machine instruction will operate on all elements of a vector or pair of vectors, completing the operation on all vector elements much faster than a scalar processor could complete the operation iteratively. Typical operations would include arithmetic or logical operations on a vector by a constant or between two vectors. Vector processors are usually implemented as pipelined scalar machines which are optimized for vector operations.

2.1.6 Multiprocessors. A multiprocessor is a computer system with multiple CPUs that share a common main memory. Each CPU executes its own set of instructions and operates on data that is possibly shared with other processors. With a properly designed program, a multiprocessor can execute portions of a computation in parallel, achieving a significant performance improvement over a single processor. Because of simultaneous memory access limitations, most multiprocessors in use today have between 2 and 64 CPUs. Programs that wish to exploit the speedup offered by a multiprocessor currently must contain special instructions to control the splitting (and later merging) of parallel instruction streams. Software technology to automate this task fully is still maturing and a programmer is usually required to obtain satisfactory results. The extra development effort required for creating parallel programs, and the lack of parallel programming standards, has limited the availability of programs that take advantage of parallel processors. Multiprocessors are most easily exploited by multitasking operating systems; they can assign each task to a specific processor and effectively use them without special programming effort by their users.

Multiple processors with local main memory avoid the bottleneck of shared memory, but the new bottleneck to performance becomes interprocessor communication. This is the approach taken by massively parallel machines, discussion of which is beyond the scope of this article (see PARALLEL COMPUTING).

2.1.7 Interrupts. Most every processor supports some form of interrupt mechanism. When this mechanism is enabled, the receipt of an interrupt causes the current instruction stream to be suspended and an interrupt service routine program to be executed. The interrupted program can be resumed transparently when the service routine completes. Interrupt hardware usually supports a number of distinct interrupt types, and a priority scheme so that the most important interrupts can preempt ones of lesser importance. Individual interrupts can usually be enabled, which allow them to interrupt the processor, or disabled, in which case they do not effect the processor.

Interrupts allow computer systems to respond to events as they occur without the overhead and possible delay of explicitly checking to see if an event of interest occurred (polling). They are particularly important for computers involved in real-time control of equipment, for efficient interfacing with computer peripherals, for time-keeping and scheduling functions, and for handling unexpected conditions and errors. On pipelined machines, excessive interrupts can adversely affect performance, as the switch in the instruction stream caused by the interrupt causes a delay while the new instruction stream is fetched.

2.2 Main Memory

Computer memory is arranged in a memory hierarchy (Fig. 6) in order to strike a balance between cost, capacity, and performance. The fastest, most expensive memory in the machine is the CPU registers, which have a small total data capacity. Successively larger, less expensive, and slower memory is arranged in order of decreasing speed (and cost) and increasing capacity. A well balanced memory hierarchy is key to good overall system cost and performance. This section will discuss main memory and main memory caches.

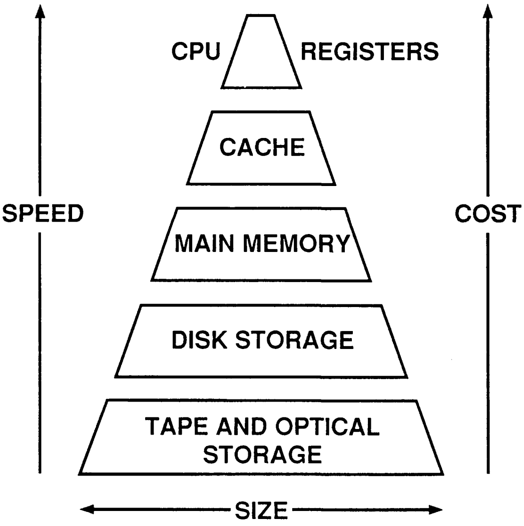

FIG. 6. The relationship between cost, size, and speed in the hierarchy of computer memory systems.

2.2.1 Main Memory Organization and Interleaving. Main memory is random-access: when presented with a memory address, it will directly read or write a given word (or smaller unit if the addressing scheme is so organized). The word width of the memory is usually the word width of the CPU, but designs in which the memory word is twice the size of the CPU to increase performance or half the size to reduce cost are not uncommon. In both cases additional hardware makes the size mismatch invisible to the computer user. Physical memory sizes of computers varies widely: personal computers will have from a few hundred kilobytes up to 16 megabytes; workstations typically have from 4 to a couple of hundred megabytes; larger machines may have hundreds of megabytes or several gigabytes. Memory access times also vary widely, but are generally in the range of 1 to 100 ns. Semiconductor memory density has quadrupled every two to three years throughout the 1980s permitting ever larger main memories and caches, and this trend continues unabated.

Memory systems require a fixed amount of time, called the cycle time, to perform a read or write operation and become ready to perform the next operation. There are usually three or more distinct time periods within the memory cycle:

1. accept an address and operation code from the CPU;
2. accept data (write cycle) or return data (ready cycle); and
3. perform internal operations to complete the cycle, which ensures that the data are properly stored, and to prepare for a new memory cycle.

Interleaving is a technique, similar to pipelining in the CPU, which offsets the cycles of two or more banks (sections) of memory so that the CPU does not have to wait for the entire memory cycle before it can issue a new memory operation.

For example, imagine we had a memory system with two banks or sections, with one section handling even-numbered addresses and the other handling odd-numbered addresses. Further imagine that the third phase of the cycle takes half of the memory cycle time. These memory banks are configured such that the memory cycle of second bank always starts halfway through the cycle time of the first (conversely, the first always starts halfway through the cycle time of the second). With such a scheme, if the processor alternately accesses even and odd memory addresses, it can complete memory transactions at twice the basic memory cycle time. The portion of the memory cycle that does internal operations completes while the processor works with the alternate bank on new memory operation in the early phases of its cycle.

Memory refresh is required for data retention in dynamic semiconductor memories, which are most commonly used in main memories. These devices store information as the presence or absence of electric charge in memory cells of 1 to 4 metal–oxide–semiconductor (MOS) transistors. When a charge is present, it will leak away over time and requires periodic replacement. The memory refresh process periodically reads all memory cells and rewrites them, thereby recharging all previously charged cells. Memory refresh cycles typically need to occur every few milliseconds, while normal memory operations are measured in nanoseconds, and so the overhead is not too large.

Some memory systems provide for protection of sections of memory. Memory protection schemes are commonly used on multiprocessing systems, to prevent independent programs from inadvertently or maliciously reading or altering another program's memory.

2.2.2 Parity and Error-Correcting Code (ECC). To protect against hardware errors in the memory, parity or an error-correcting code (ECC) is used to detect and correct memory errors. In the parity scheme, an additional error-detection bit, the parity bit, is stored with every byte by the memory subsystem. This bit is set by the memory hardware such that the total number of "1"s that are stored is odd (odd parity). Should a read operation result in an even number of ones, a parity error has been detected. This scheme detects single or triple bit errors, but cannot correct errors or detect double bit errors.

ECC memory (error correction code) stores several additional bits with each word of data both to detect memory errors and, in many cases, to correct them. A common scheme is to store 7 ECC bits with each 32-bit word. Such a scheme uses a Hamming code (Hamming,

1950) to create the 7 ECC check bits, which duplicate enough information from the stored 32-bit data to allow all single-bit errors to be corrected and all double-bit errors to be detected (Siewiorek and Swarz, 1982).

2.2.3 Memory Mapping. When the address space available to programs is smaller than the available physical memory (typical of 16-bit machines), techniques are employed to allow access to the larger physical memory within the confines of the smaller program address space.

One popular technique is memory mapping. In this scheme the upper portion of the logical address is used to select one of several mapping registers. The contents of the selected mapping register are used as the upper portion of the physical address while the remaining low bits of the logical address form the low bits of the physical address. In Fig. 7 this is illustrated for a system with a 16-bit logical address space and a 28-bit physical address space. The upper four bits of the logical address are used to select 1 of 16 mapping registers which each provide the upper 16 bits of a physical address. They are concatenated with the lower 12 bits of the logical address to produce a 28-bit physical address. By manipulating the contents of the mapping registers, all of a large physical memory can be accessed.

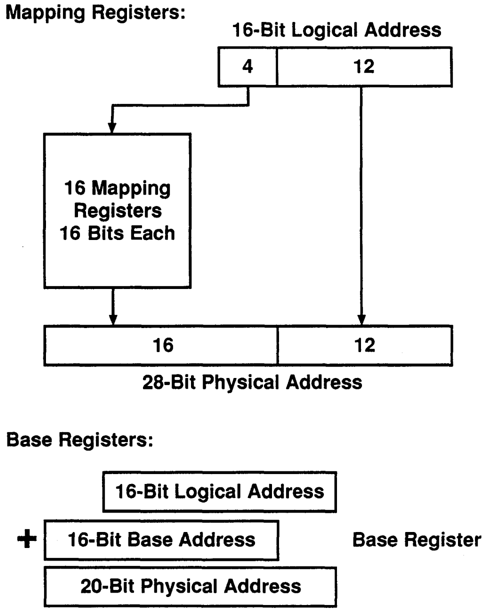

FIG. 7. Typical configurations of mapping registers and base registers.

Another technique is the use of base registers. These registers contain high-order memory addresses, and they are offset and added to the logical address automatically by the hardware. Often the choice of base register is based upon the context of the reference, be it for an instruction, to a stack, or for data. Figure 7 illustrates the scheme used by the 8086 family of processors used in PCs; the 16-bit base register is offset from the program address by 4 bits to produce a 20-bit physical address. There are many variations on these two techniques; see Siewiorek *et al.* (1982) for further examples.

2.2.4 Virtual Memory. Virtual memory is a technique that permits a physical memory to appear to be very much larger than it actually is, giving the illusion of a large, linear address space, which makes programs easier to write and more independent of the physical resources of a particular computer. The illusion of virtual memory is accomplished with additional logic in the memory subsystem, a large backing store which is usually a disk, and supporting software (Hennessy, 1990). Both virtual and physical memory is divided into blocks called pages. Each virtual page has an entry in a section of memory called the page table, which keeps track of the location of a page, whether it is currently resident in physical memory or paged out onto backing store. When a referenced page is resident, the translation of the virtual page address to a physical page is handled entirely through hardware via the page table, with only a very slight delay. However, when a referenced page is not resident, the supporting software is signaled that a page fault has occurred by the memory hardware. This supporting software will then suspend the requesting program, make the requested page resident (which may involve making room by moving another page out to backing store), and then resume the requesting program. The program is not aware that a page fault occurred, except for the fact that some additional time elapsed while the physical memory was being rearranged to satisfy the virtual address reference.

In order to improve virtual memory performance, a cache of recently translated virtual addresses is kept in a translation buffer. When a virtual address is found in the translation buffer, the computer hardware can accomplish the memory access without reference to the page tables, further improving performance.

Virtual memory systems depend on an observed characteristic of most all programs known as locality—that when a reference to memory is made, there is a very high probability that subsequent references will be to a location near the previous ones. Consequently, the probability that a subsequent reference is to a page already in memory is high, and overall system performance will be good. Thrashing is a problem that occurs when a program accesses virtual memory in such a way it is constantly page faulting, usually because it does not exhibit locality. It therefore has a very high page fault rate and consequently poor performance.

2.2.5 Caches. Caches are fast memories located between the main memory and the central processor and used to buffer the CPU from the relatively slow main memory. They maintain a copy of selected items in main memory, and rely on the fact that most programs make repeated references to clustered areas of memory (see discussion of locality in Sec. 2.2.4). Caches are designed to have a high probability of a cache hit (finding the required information in the cache), so that the faster cache can supply the information to the processor without incurring the speed penalty of a main memory reference. When the information is not in the cache, it is fetched from main memory and supplied to both the processor and the cache. Many caches maintain information in blocks larger than one word, and the cache will fill the entire block on the first cache miss, betting that other references will be made to that block.

Multiple-level cache systems are used with large main memories and fast processors, each cache successively smaller and faster. A system with a multiple-level cache would have several trapezoids in Fig. 6 in place of the single one shown.

Cache memory usually uses static semiconductor memories (static RAM). These devices, while significantly less dense then the dynamic RAMs used in main memory, are usually an order of magnitude faster and do not require refresh cycles (each memory cell is active and can maintain its state while the part is powered up).

While reads dominate cache access, information must also be handled on write. A write-through cache writes both the cache and the main memory, slowing the processor. A write-back cache writes only to cache initially, and the main memory is not updated until the cache block is replaced by other information. Other design issues include the lookup (mapping) scheme, block size, total size, and block replacement scheme (Hennessy 1990). Cache coherency is an important issue in multiple processors where several processors may be writing and reading simultaneously. If each processor in a multiple processor has its own cache, there can exist several copies of the same data. Should one processor modify a piece of data, it must ensure that all other copies of the data in other caches or in main memory are either updated or marked as invalid. Various schemes to maintain cache data coherency have been invented, with various performance characteristics and hardware complexities; see Dubois (1990) for approaches to this issue.

2.3 Busses

Busses (see BUS SYSTEMS AND COMPUTER INTERFACES) connect various subsystems together in the computer, moving addresses and data over a well-defined interface. Many computers have just one major bus that connects the CPU, memory, and input/output (I/O) adapters together (Figs. 1 and 4). Faster machines usually employ separate busses for the slower I/O, connecting the I/O busses to the faster CPU memory bus through an adapter or directly to memory.

2.3.1 Standardized Busses. Standardized busses allow peripheral devices to be designed once and used on a variety of machines, and for machines to be expanded after they have been designed and manufactured. The bus of the IBM PC-AT, probably the most used bus in the history of computing, is a *de facto* standard. A few of the most common formally standardized busses include the VME bus (IEEE 1014) and the SCSI bus (ANSI X3.131).

2.4 Input/Output

Input/output, or I/O, is the general term used to describe the movement of data between main memory and any peripheral device, such as disks, tapes, networks, terminals, or printers. Peripheral devices connect to the computer's I/O bus through an adapter. Adapters convert the specialized interfaces of the peripheral to a series of registers that can be read to or written from the I/O bus. Most adapters have some form of buffering, a small local memory that temporarily stores I/O data and helps to overcome processor and peripheral speed differences.

Programmed I/O is the simplest technique, using the central processor to read and write data, peripheral control, and peripheral status to or from the adapter registers and move the data to or from memory. Direct memory access (DMA) logic in the controller allows the CPU to set up a data transfer by writing control information to the adapter registers, with the actual data movement to or from memory being performed by the adapter, freeing the CPU for other work. Some of the most powerful computers employ separate I/O processors (sometimes called channels) that run specialized I/O programs to provide high-speed I/O within minimum impact on the CPU.

3. PERIPHERAL DEVICES

There is a wide variety of peripheral devices for use with computers. We will discuss the main features of the most common ones in this section.

3.1 Persistent Mass Storage

The vast majority of mass storage is electromechanical, in contrast to the purely electronic main memories discussed previously. While access to information is much slower, the data capacity of these devices is much larger, and the cost per bit of information stored is much lower. The stored information is said to be persistent, i.e., the information is retained when the device is powered off.

Fully electronic persistent storage is available, but currently with very low capacity and high cost, such as flash memories used in some portable computers. Vigorous development continues on these technologies and there are predictions that they will have a significant presence later in the decade.

3.1.1 Magnetic Disks—Floppy and Hard. A magnetic disk stores information on a disk-shaped platter coated with a thin magnetic film (see MAGNETIC STORAGE OF INFORMATION). The disk rotates at constant velocity under a read/write transducer called a head. The head is moved over one of many tracks on the surface of the disk, each track occupying a specific radial distance from the center of the disk. The head can detect previous magnetic patterns stored on the disk in a read operation, or write new magnetic patterns to store information. Information is stored in sectors, which is a subsection of a track. A sector will store from a hundred to a few thousand bytes of data, along with overhead bytes that synchronize the read/write circuits, identify the sector, and store error detection and correcting codes.

Disk striping is a technique used to improve performance in situations where files are large and there are multiple disks available. A file is distributed over several disks (members of a stripe set) such that when a section of a file fills an entire track, the next part of the file is found on the next disk in the stripe set. The goal is to eliminate the delay that moving the disk head from track to track causes, by switching the operation to another disk that has prepositioned its heads to the proper track. Switching the data stream among disks is done at electronic speeds, while moving heads is done at mechanical, inertia-limited speeds. Consequently, a stripe set of several disks with good controlling software can significantly outperform single drives in many situations.

A floppy disk is so named because the magnetic disk is made of a thin flexible plastic disk packaged in a flexible or rigid covering. Information may be stored on one or both sides of the disk. Floppy disks come in 8-, $5\frac{1}{4}$-, and $3\frac{1}{2}$-in. diameters with data capacities up to 2.8 megabytes. Floppy disks may be removed from the disk drive, and consequently a single drive may service a library of disks, or the disk may be used to transport data from one computer to another.

A hard disk is made of rigid material, and usually has several disks stacked on a common hub. Heads are ganged together and moved in unison by a single positioner, with each disk surface being serviced by a single head. 14-, 9-, and 8-in. diameter disks are produced usually for use with large computers, while $5\frac{1}{4}$- and $3\frac{1}{2}$-in. units are popular in workstations and PCs. The advent of laptop computers has kept sizes shrinking, and 2-in. and smaller models are now on the market. Disk capacity for $5\frac{1}{4}$-in. and larger units is now above 1 gigabyte and increasing; $3\frac{1}{2}$-in. hard disks can currently store about half that. Removable hard disks are available but declining in popularity; most hard disks today are built for permanent installation with the disks sealed inside the drive.

The key performance criteria for disks are transfer rate and average access time. Transfer rate, in megabytes per second, indicates how fast information may be moved to or from the disk once it is located. Average access time describes how long it takes to access any piece of information on the disk. Access time depends on seek time (how fast the heads can position to a given track) and rotational latency (how long it takes to rotate the desired information under the head).

3.1.2 Optical Disks. Mechanically, optical disks are very similar to magnetic disks, employing a rotating disk and positionable heads. However, the disk surface is designed to store data on the basis of altering the reflectivity or polarization of light, and storing information involves changing these characteristics, usually by heating a small area with a laser pulse (Hamacher, 1984) (see OPTICAL STORAGE). All current optical disk drives have removable media.

Initial optical disks were read-only devices (most drives could not write information on them). CD-ROM is the most popular format, and stores 600 megabytes of data on a $4\frac{3}{4}$-in. diameter compact disk identical in size and technology to compact audio disks. In fact, many newer CD-ROM drives for computers include headphone jacks and volume controls, and will also play audio CDs.

Write-once read-many (WORM) optical disks are available, useful mainly for archival storage. 14-, 8-, and 5-in. units are available with capacities from 600 megabytes to over a gigabyte.

Finally, read-many/write-many devices are becoming available. Capacities and sizes are similar to the CD-ROM and WORM devices.

3.1.3 Magnetic Tape. Magnetic tape, lacking random access, is the slowest mass storage technology currently in use. However, its low cost per bit and large capacity make it a popular archive and transport mechanism. While nine-track reel-to-reel tapes are still in use, a variety of less expensive and more convenient cartridge-based devices are now more popular. The quarter inch cartridge (QIC) is widely used in workstations, and 8-mm video tapes and 4-mm digital audio tapes (DAT) are now being used, often with stack loaders that can automatically change up to 16 tapes. Cartridge capacities of 5 gigabytes are available.

Tapes store data in a thin magnetic film layered on flexible tape. Nine-track and QIC tapes store data linearly along the length of the tape, using one head per track. 8- and 4-mm units use helical scan technology, similar to that used in video cassette recorders, storing bits on diagonal stripes. Helical scan technology is more complex, but permits much higher bit densities.

The useful capacity of all types of magnetic tapes is improved by data compression technology, which uses hardware and/or software to encode the stored data in such a way that it occupies less space on the storage medium. Extremely effective data compression schemes have been devised and it is not unusual to be able to double the raw capacity of a tape for many common data storage situations. The same technology is finding application in portable computer disk drives, where the user is willing to accept the lower performance caused by the compression/decompression process in favor of significantly increased storage capacity.

3.2 Communications and Ports

3.2.1 Serial and Parallel Ports. A port is a defined interface on the computer that can read and/or write digital data to an external device. A parallel port transfers all the bits of a word at once, or in parallel. A serial port transfers the bits one at a time in sequence, at rates that typically range from 300 to 38,000 bits per second (MaNamara, 1988). Most terminals and modems attach to serial ports, while printers may use either.

3.2.2 Modems. A modem (MOdulator/DEModulator) is used to transfer serial data over an audio channel such as a telephone line or radio link. Common modem speeds are 1200, 2400, and 9600 bits per second. Many newer modems have data compression and error correction features. A typical modem attaches to a serial port and to a modular telephone jack, and can dial the telephone under computer control. As all-digital telephone services, such as ISDN, become widely available, the need for modems will be eliminated.

3.2.3 Networks. Networks allow a group of computers to share programs and data easily, and usually provide shared facilities such as file and printer servers. See NETWORKS, COMPUTER for additional information.

3.3 User Interface Devices

3.3.1 Hard-Copy Devices—Printers and Plotters. Hard copy is a term used to refer to printed output. While printers have traditionally focused on character-based documents and plotters drew charts and graphs, the distinction is blurring as newer hard-copy devices, including laser, ink-jet, and 24-pin printers can do both. Laser printers use much of the same technology as a plain-paper office copier; the optical input used by the office copies is replaced by a computer-controlled laser to form the image (see XEROGRAPHY). Resolutions of 300 dots per inch (DPI) on $8\frac{1}{2}$ by 11 or 11 by 17 in. sheets are available, and color laser printers are now available. Ink-jet printers squirt ink drops out of a nozzle to form dots on the paper. Ink-jet printers are quiet and can print at 150 DPI in both single color and multiple colors. Impact printers, now predominantly using thin wires in the print head to strike an inked ribbon and then the paper, also can produces 150-DPI text and graphics. Multiple-color ribbons offer the ability to produce limited multiple-color output slowly.

3.3.2 Soft-Copy Devices—Terminals and Displays. Computer terminals and displays (see DISPLAY TECHNOLOGY) are used to produce soft-copy output on a screen. They may be character only or mixed characters and graphics, in color or monochrome. The basic character display today displays 80 columns and 24 or 25 rows of characters. Each character may be displayed with a small number of special characteristics, such as an underline, high intensity, or blinking. Color displays usually allow each character to have selected foreground and background colors.

While character displays model the display space in terms of character cells only, a graphics display models the display as a two-dimensional array of picture elements (pixels) which can display both characters and arbitrary graphics. Typical PC displays have 640 by 480 pixel displays, while workstations commonly have 1024 by 768 or 1280 by 1024 pixel displays.

Monochrome displays use one bit per pixel to indicate whether a pixel is on or off. Grayscale displays use multiple bits per pixel (usually 2, 4, or 8) to represent the brightness of a particular pixel. Color displays also use multiple bits per pixel (up to 24). Frequently, color displays use the value of a pixel to index into a color map, which then selects the displayed color from a very large palette of potential colors. A typical color display uses 8 bits per pixel, so that 256 different colors can be displayed simultaneously. The color map translates 8-bit pixel values to 24-bit color values, allowing the selection of 256 colors from a palette of 2^{24}.

Most terminals use cathode-ray tube (CRT) technology, with diagonal measures of 10 to 19 in. The demand for laptop computers has driven the development of flat-panel display technology. Liquid crystal displays (LCDs) are in widespread use in laptops, with early units offering 640 by 480 notebook-sized grayscale displays. 16-color units are available at higher cost. With strong worldwide demand for portable high-resolution color LCD displays for both computer and television use, the technology is expected to evolve rapidly to high resolutions and lower cost.

3.3.3 Input Devices. Input devices have remained remarkably stable in an otherwise rapidly changing field. Keyboards still dominate, and the mouse has established itself as the predominant pointing device. A mouse is a palm-sized device that may be rolled across a work surface and transmits relative movement to the computer. Software or hardware then moves a cursor on a display to mimic the motion of the mouse. From one to three buttons on the mouse allow a user to signal some action to software, usually manipulating a displayed object that the cursor is positioned on. A trackball is a partially exposed sphere that a user may roll in place; the relative motion is transmitted to the computer as in the case of a mouse. An advantage of a trackball is that it does not require desk space for motion. Tablets have a rectangular surface over which a puck with buttons is moved (not unlike a mouse); a tablet transmits absolute coordinates to the computer, based upon the position of the puck on the surface of the tablet.

Recent work in handwriting recognition software foreshadows the introduction of so-called "pen-based computers" in the near future, where the keyboard and mouse would be replaced by a stylus which "draws" on a display panel.

3.3.4 Optical Scanners. Optical scanners convert printed or handwritten copy into bit-mapped images, so that they may be manipulated, displayed, or printed by the computer. Hand-held and single-sheet scanners are low in cost and typically produce 300 to 600 dots per inch monochrome or gray-scale images. Color scanners are available at greater cost and at various resolutions. Software is available to convert scanned-in character images to computer character codes, permitting further processing by most existing programs.

4. SUMMARY

Computers continue to evolve, providing ever higher performance and larger capacities at lower costs. Despite this constant evolution, the majority of successful machines over the first four generations have had a common gross structure and similar subsystems, which we have reviewed in this article.

The advent of the personal computer has moved computing into the mass market, and consumer electronics technology such as compact disks and digital audio tape are being applied to solve computing problems. Other mass market technology, such as cameras that use floppy disks as "film," cellular radio networks, handwriting recognition software, and voice recognition systems (Kamel, 1990) all indicate a healthy and vigorous future for computers with even broader application.

ACRONYMS

ALU: Arithmetic logic unit.

ASCII: American standard code for information interchange, a character encoding standard.

CD-ROM: Compact disk—read-only memory.

CISC: Complex instruction set computer.

CPU: Central processing unit.

CRT: Cathode-ray tube.

DAT: Digital audio tape.

DMA: Direct memory access.

ECC: Error correction code.

ISDN: Integrated services digital network, a packet-switching digital telephone service.

MFLOPS: Millions of floating-point operations per second. A popular but nonstandardized figure of merit for floating-point performance.

MIPS: Millions of instructions per second, a popular but nonstandardized figure of merit for integer performance.

PC: (1) Personal computer, (2) program counter.

QIC: Quarter-inch (tape) cartridge.

RAM: Random access memory. While ROM is also random access memory, the term has come to be used almost exclusively to refer to read-write memory based upon charge storage in semiconductor chips.

RISC: Reduced instruction set computer.

ROM: Read-only memory, written once with a fixed set of instructions and/or data can be repeatedly read thereafter.

WORM: Write once read many.

GLOSSARY

Bit: The smallest unit of information represented by a digital computer, which can have one of two values: 0 or 1. From Binary digIT.

Byte: A set of eight bits, commonly used to represent a character. On many computers, the smallest addressable entity.

Core or **Core Memory:** Random-access read-write memory based upon information storage in magnetic cores. Some operating systems refer to any type of main memory as core for historical reasons.

Pixel: A single dot or picture element on a graphic display.

SPECmark: An overall figure of merit of computer performance, based on benchmarks distributed by the Systems Performance Evaluation Cooperative (SPEC).

Word: An addressable collection of bits. In current practice, words are usually a multiple of 1, 2, 4, or 8 bytes (8, 16, 32, or 64 bits, respectively).

List of Works Cited

ANSI (1977), *X3.4-1977 Code for Information Interchange*, New York: American National Standards Institute.

Cohen, D. (1981), *Computer* **14**, 48–54.

Dubois, M., Thakkar, S. (Eds.) (1990), Special Issue on Cache Architectures in Tightly Coupled Multiprocessors, *Computer* **23** (6).

Hammacher, V. C., Vranesci, Z. G., Zaky, S. G. (1984), *Computer Organization*, 2nd ed., New York: McGraw-Hill.

Hamming, R. W. (1950), *Bell Syst. Tech. J.* **29**, 147–160.

Hennessy, J. L., Patterson, D. A. (1990), *Computer Architecture: A Quantitative Approach*, San Mateo, CA: Morgan Kaufmann.

IEEE (1985), *SIGPLAN Notices* **22**, 9–25.

Kamel, R. (Ed.) (1990), Special Issue on Voice in Computing, *Computer* **23** (8).

McNamara, J. E. (1988), *Technical Aspects of Data Communication*, 3rd ed., Bedford, MA: Digital Press.

Siewiorek, D. P., Swarz, R. S. (1982), *The Theory and Practice of Reliable System Design*, Bedford, MA: Digital Press.

Siewiorek, D. P., Bell, C. G., Newell, A. (1982), *Computer Structures: Principles and Examples*, New York: McGraw-Hill.

Weicker, R. P. (1990), *Computer* **23**, 65–75.

Further Reading

Duncan, R. (1990), *Computer* **23**, 5–16. Describes the multitude of parallel computer architectures and classifies them.

Gajski, D. D., *et al.* (Eds.) (1987), *Computer Architecture*, Washington: IEEE Computer Society Press. A tutorial on computer architecture presented as a collection of important papers from 1977 through 1985. Includes sections on microprocessors, RISC machines, and special purpose architectures.

Hennessy, J. L., Patterson, D. A. (1990), *Computer Architecture: A Quantitative Approach*, San Mateo, CA: Morgan Kaufmann. An excellent, easily read textbook that covers all aspects of computer architecture and hardware design, with an emphasis on performance and cost implications. Includes analysis of contemporary RISC machines.

Hennessy, J. L., Juppi, N. P. (1991), *Computer* **24**, 18–29. An article tracing the impact that advances in technology have had on computer architecture and suggesting what the future holds.

Horning, R. J., *et al.* (1991), *IEEE Spectrum* April 1991, 58–68. A good paper on the choices and challenges facing contemporary RISC workstation architects and designers.

Krick, R. F., Dollas, A. (1991), *Computer* **24**, 5–15. Presents a history of advances in instruction sequencing technology and discusses current strategies.

Stone, H. S., Cocke, J. (1991), *Computer* **24**, 30–38. Discusses the impact that high speeds will have on computer design, and the need to consider optical interconnects, greater parallelism, and use of locality to increase performance.

Weicker, R. P. (1990), *Computer* **23**, 65–75. A good overview of popular benchmark programs and the pitfalls of comparative benchmarking.

COMPUTER PROGRAMMING LANGUAGES

HERBERT J. BERNSTEIN, *Bernstein+Sons, Bellport, New York, U.S.A.*

- **Introduction** 163
- 1. **History** 164
  - 1.1 Plug-Board Programming, Punch Cards, and Paper Tape 165
  - 1.2 Stored Programs and Machine Language Programming 165
  - 1.3 Assemblers 166
  - 1.4 Compilers 166
  - 1.5 Interpreters 167
  - 1.6 Software Engineering 168
  - 1.7 Supercomputing 168
  - 1.8 Parallel Computing 169
  - 1.9 4GLs, Packages, and Nonprogramming 169
- 2. **Programming Language Concepts and Styles** 170
  - 2.1 Imperative Programming 171
  - 2.2 Object-Oriented Programming 172
  - 2.3 Functional Programming 173
  - 2.4 Logic Programming 173
  - 2.5 Concurrent Programming 174
- 3. **Languages for Major Scientific Calculations** 174
  - 3.1 FORTRAN 175
  - 3.2 C 178
  - 3.3 ALGOL and PASCAL 181
  - 3.4 Libraries and Language Extensions 183
- 4. **Languages for the Laboratory Environment** 183
  - 4.1 BASIC 184
  - 4.2 FORTH 185
- 5. **Languages for Graphics** 185
- 6. **Languages for Data Retrieval** 186
- 7. **Summary** 186
- **Acknowledgments** 186
- **Glossary** 186
- **Works Cited** 187
- **Further Reading** 188

INTRODUCTION

Computer programming languages are used by people to specify the actions to be taken by computers in solving problems. As the needs of different types of users to solve problems with computers vary, so do the computer programming languages that have been developed. In this article we concentrate on the aspects of computer programming languages of interest to physicists, both for major scientific calculations and for control of experiments. Thus we will give only brief attention to languages and approaches to languages that, while of major importance in other fields, such as computer science, are not critical to working physicists. The so-called imperative or statement-oriented programming languages, which specify sequences of actions to be taken on objects (such as adding a list of numbers, and then dividing by the count of things in the list to get an average), are the major tools of modern physics.

In the space available we cannot fully describe the rich structure of even this small subset of computer programming languages. At best we can give some of the essential terminology, and examples of the most heavily used features. Languages used by people to communicate with people are learned by using them. The same is true of languages used by people to communicate with computers.

Computers have their own internal languages, private jargons of strings of numbers that are difficult for most people to learn and rarely used as programming languages. A single computer may have more than one such language, and different types of computers, even though made by a single manufacturer, may have different languages. The lists of numbers direct the machine to perform certain simple actions. Because the choice of numbers may vary from machine to machine, the number 103344 might be an "instruction"

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

(an "order" or "command") to add two numbers stored in locations 33 and 44 on one machine, and an instruction to print the character in location 3310 on another. Such lists of numbers are almost as cryptic as the codes and ciphers used to hide secrets from spies. These "machine languages" require a programmer to specify in full detail what is to be done, and slight conceptual changes can require large portions of a program to be rewritten.

In order to save the effort of writing a new and different program for each machine to solve the same problems, to make computers accessible to more people, and to increase the chance of writing instructions for computers correctly, many programming languages for computers have been created, languages that express the actions to be performed in terms more comprehensible to people, and in ways that allow the same programs to be used on many different types of computers.

There is a large range of problems in physics for which computer languages are used. For problems expressed as formulas and equations, there are languages, such as FORTRAN, that accept formulas in ordinary scientific notation. For problems of process control and control of scientific experiments, languages that allow a finer control over a machine and facilitate management of real-time processes running in parallel, such as C and assembly languages, are used. For many problems it is important to be able to make rapid changes in a program, and the efficiency of execution is not as critical. In such cases, interpretive languages, such as APL, BASIC, and FORTH, that allow "on-the-fly" changes are used.

Some languages are close in style to the operation of computers, consisting of sequences of actions to be taken on objects. Such a style simplifies efficient translation of a problem specification into the internal language of a computer, but may force a strained interpretation of the original problem, increasing the chances of programming errors, and increasing program maintenance costs. To reflect more naturally the demands of theproblems to be solved rather than the demands of computers, languages using other styles of programming have been developed: object-oriented, functional, and logic programming. These offer promise of increased productivity for appropriate problems, but for the classic problems of theoretical and experimental physics, FORTRAN and its statement-oriented, imperative relatives predominate.

We will begin by considering the history of computer programming languages and the differing programming styles that have been developed. Then we will look in some detail at the languages for major scientific calculations, exemplified by FORTRAN, C, and PASCAL, and the additional languages used in control of experiments, exemplified by BASIC and FORTH. For each we will give some of the vocabulary of the languages: the "operators," which are effectively the verbs of the language, and the "data structures" and "data types" that define the nouns and their modifiers.

1. HISTORY

The history of computer programming languages is intimately related to the history of computers themselves (see COMPUTERS; COMPUTER HARDWARE). Many of the restrictions and inconveniences in early languages were reflections of the limitations of the computers of the time. Because our focus here is on languages, rather than computers, we will omit many critical contributions to the design of computers and discuss only some examples of the changes in hardware that had a significant impact on the development of languages.

The first consideration of programming was for Charles Babbage's proposed Analytical Engine by Augusta Ada, Countess of Lovelace, in 1842 (Morrison and Morrison, 1961, pp. 245–295; Hollingdale and Toothill, 1965). The programming was in a direct machine language, with a manual translation from normal mathematical notation. Babbage's machine had a memory, a processor for arithmetic and control, and input-output capabilities. At this early date, the fundamental concepts of operations (actions to be performed), variables (information that may change during the calculation), and constants (numbers, such as e or π, that would not change) were established.

1.1 Plug-Board Programming, Punch Cards, and Paper Tape

In Babbage's time, cardboard cards with holes punched into them were used for the control of looms, and he chose such punch cards as a mechanism for specification of both operations and variables. In 1890, Herman Hollerith invented a series of electromechanical punch-card devices that could do census calculations under the control of a wired-in program of simple actions, such as adding, subtracting, or printing certain columns of numbers. The numbers were stored as holes punched into cardboard cards; hence the name "punch cards." "Programming" such machines consisted of throwing switches and inserting jumper wires into plug boards (like an old-fashioned telephone switchboard). Programs were saved by simply keeping the entire plug board with its wires, and putting in a fresh one for the next program. These electromechanical machines were put to increasingly complex and general tasks, culminating in the design and construction of the Automatic Sequence Controlled Calculator, the Mark I, by Howard Aiken from 1937 to 1944 (Slater, 1987, pp. 81–90). In addition to its internal plug board wiring, it accepted a general program of instructions on paper tape. The operations of the machine were composed from addition, subtraction, multiplication, division, and retrieval of previously computed variables. Data came from three more paper tapes and two stacks of punch cards. The values of 60 constants were set from 1400 rotary switches. Programming the machine was a complex process of preparing tapes, cards, and switch settings. The program on the control paper tape consisted of lines of 24 holes each, broken into three fields, two for operands and one to specify the operation (Bashe *et al.*, 1986).

1.2 Stored Programs and Machine Language Programming

A great simplification of programming came with the creation of digital electronic stored-program computers because a program was stored in memory as numbers. The concept was pioneered during work on Eckert and Mauchly's Electronic Numeric Integrator and Calculator (ENIAC); see Rosen (1969). ENIAC was programmed by wiring, but Wilkes' EDSAC in 1949 fully implemented stored-program computer design. Early internal memories were small, and would have greatly limited program size, so that programs still existed primarily as holes in cards and paper tape external to computers. However, having the ability to store large portions of a program internally made it possible to reuse the same sequence of instructions rapidly without having to reload them for each execution from cards or tape, thereby having efficient "loops" and "subroutines." Commonly used modules of code, e.g., routines to do calculations on floating-point numbers (i.e., real numbers in which the decimal point might be at different places) or complex numbers on machines that did not have built-in hardware for such calculations, were programmed in the native language of the machine, "machine language," and saved for use with many programs. Machine languages settled down to a variety of simple numeric representations of operations of the form

Operation Operand-Address Operand-Address...

where the number of operand addresses (the places from which data was to be fetched, or to which results were to be stored) was determined by the design of the machine and the particular operation being done. Sometimes all the operands were explicitly listed with the instruction, and sometimes one or more operands were implicit in the operation chosen. On some machines the address from which the next instruction would be fetched had to be explicitly given as part of the instruction. In others, that was needed only to create conditionally executed sections of code—loops and subroutines.

If the computer used decimal numbers internally, programming could be a simple transliteration of operations into appropriate decimal digits, and laying out of locations of instructions and operands in memory. This was a complex extension of the problem of seating many people for a large dinner party.

Hardware economies could be found by building computers based on binary numbers (base 2) rather than decimal (base 10), so, for such machines, instructions had to be created in that numbering system. Octal (base 8) notation was sometimes convenient for creating such instructions.

1.3 Assemblers

Translation of names for operations and names of the things that need to be fetched and stored is itself a calculation, one for which the computer itself is well suited. Thus, in the 1950s machine-language programming was replaced for most purposes by assembly-language programming (e.g., the work of Rochester and McClelland; see Bashe *et al.*, 1986, pp. 323 ff.). An assembler translates names for operations, such as "ADD" and "SUB," into the appropriate numbers, and keeps track of which memory locations have already been used for instructions, variables, and constants, so that the programmer might refer to them by name instead of by number. Though each machine had its own assembly language, at least the pieces of code for that machine saved in symbolic form did not have to be rewritten when used with other code that would force changes in addresses.

For example, let us invent a decimal machine that has the following instructions:

01ssdd:

MOV ss [source to] dd [destination]

02ssdd:

ADD ss [source to] dd [destination]

03vvll:

IF vv [variable to test is] NOT-ZERO [then goto location] ll

04ssdd:

READ [a number from source device] ss [and store in] dd

05ssdd:

PRINT [the number in location] ss [on device] dd

060000:

END

and assume that the source of input data is device 5 and the printer is device 6. Then to read in 10 numbers, add them, and print the result, we might write

```
TEN:       DATA      10
NEGONE:    DATA      -1
ZERO:      DATA      0
INPUT:     VARIABLE
ICOUNT:    VARIABLE
ISUM:      VARIABLE
START:     MOV            TEN,ICOUNT
           MOV            ZERO,ISUM
LOOP:      READ           05,INPUT
           ADD            INPUT,ISUM
           ADD            NEGONE,
                           ICOUNT
           IF             ICOUNT NOT-
                           ZERO, LOOP
           PRINT          06,SUM
           END
```

rather than

```
      10
      -1
      0
      0
      0
      0
      010004
      010205
      040503
      020305
      020104
      030408
      050506
      060000
```

This was a significant improvement in making it easier to write programs. However, programmers still had to write out every instruction in full detail. To save time in writing similar sequences, assemblers were expanded to become macro-assemblers, in that sequences of instructions could be given names.

1.4 Compilers

Assemblers were and still are useful tools, but even with macro facilities, they are still strongly linked to particular machines, not to the problems of primary interest to users of the machines. Translators for languages that do not directly represent machine language, but that require compilation of the meaning of an entire program or major segments of a program into appropriate instructions as large units, are called "compilers." In 1952, Grace Hopper created a more user-oriented computer language called A-O for UNIVAC computers. By 1955, a group at IBM led by John

Backus had designed a language called FORTRAN (for IBM Mathematical Formula Translating System), which would take statements written in ordinary algebraic mathematical notation and translate them into a program in machine language. This language, and its successors, became the workhorses of scientific calculations in the United States, first on the IBM 704 and its descendants, and then on other computers.

In FORTRAN, the example of adding ten numbers, above, might be written as:

```
        ISUM=0
        DO 100 ICOUNT = 1,10
        READ (5,1000) INPUT
100     ISUM=ISUM+INPUT
        WRITE (6,1000) ISUM
1000    FORMAT (I6)
        STOP
        END
```

For business applications, compilers for a "Common Business Oriented Language" (COBOL) based on earlier work by Grace Hopper were developed in 1959, and for such applications the language is still in heavy use. The style is closer to English than the mathematical formulas of FORTRAN, as in

```
ADD MYVALUE TO MYSUM
```

The great success of FORTRAN and COBOL and the increasing power of computers led to attempts to create better languages. Starting in the late 1950s, Peter Naur and others worked on "Algorithmic Language" (ALGOL), intended to provide code more structured and easier to read and to maintain than FORTRAN. Partially because it was oriented toward character sets not well supported on the card punches in the United States, and because FORTRAN was adequately meeting the scientific programming needs of the time on the computers available in the United States, ALGOL did not achieve great popularity in the United States, but did attract a wide following in Europe in the 1960s. By creating PL/I in 1964, IBM attempted to correct some of the limitations of FORTRAN and combine it with all the business-oriented features of COBOL. PL/I is still in use, but has not attracted a large following. ALGOL evolved into two different languages. ALGOL-68 addressed the need for more general data structures, but was hard to read. In 1971 Niklaus Wirth proposed an alternative language, called PASCAL, that proved effective for that purpose, and was simple and clear. Use of ALGOL-68 faded out.

The special needs for computing in the military environment were addressed by Jules Schwartz's JOVIAL ("Jules' Own Version of the International Algebraic Language"), but real-time programming for control computers was done in a mixture of assembler and FORTRAN. This produced a body of code that was hard to debug and maintain as computers changed, and in the mid-1970s development of a language specially suited to programming computers operating in real time and embedded in military hardware was begun. Attempts were made to address all aspects of the cycle of creation and maintenance of software by providing tools needed for effective software engineering in the language, and to learn the lessons of the languages developed up until then. The result, developed primarily by J. Ichbiah and colleagues in 1979, was called ADA, in honor of the Countess of Lovelace; see Booch (1983). We will return to this language in considering software engineering below.

A compiled language called C was designed in 1972 by D. Ritchie (Kernighan and Ritchie, 1978). C provides the user-oriented economy of expression of FORTRAN and PASCAL with the intimate access to hardware features of assemblers, and has replaced a major portion of the use of assemblers for time-critical real-time programming in operating systems and device control code.

1.5 Interpreters

A compiler takes the user source code as a whole and translates it all to machine language before executing any of it. This can imply a long delay between changes in a program and seeing the results, and also causes difficulty in highly dynamic decisions about the use of memory. For applications requiring a more immediate response to code changes or highly dynamic uses of data structures, interpretive languages have been developed. The computer is effectively programmed to simulate a machine closer to the user needs. The first such interpreter, called the Speedcoding Interpreter, was developed by John Backus and others at IBM in 1953. Interpreters have been used for many significant languages since. Examples are APL ("A Programming Language"), developed by Kenneth Iverson in 1962, BASIC, developed by Kemeny and Kurtz in 1965, LISP, developed by John McCarthy in 1962, and FORTH, developed by Charles Moore in 1971.

The languages implemented by use of interpreters often also have compilers to allow for more efficient execution once the programmer is past the debugging stage. Many languages that are more appropriate to compilation because of complex context, such as FORTRAN and C, also have had interpreters created to aid in debugging programs.

1.6 Software Engineering

As an increasing investment was made in programs for use on computers, it became clear that the costs of software development would exceed by far the costs of the computers on which the software would run, and that the major source of failures of computer-based systems would be unreliable software, rather than unreliable hardware. Languages were proposed that more closely reflected the functions to be performed, rather than the internal workings of computers. John McCarthy developed LISP (for List Processing) from 1956 to 1962 which, though it could be used in the step-by-step style of statement-oriented (or imperative) languages, introduced the ability to solve problems simply by evaluating functions made up of yet more functions, perhaps defined recursively. LISP became the basic tool of artificial intelligence research and computerized symbolic algebra. PROLOG, in 1972, introduced an extremely high level of functional programming, permitting general statements of problems to be effectively their code. The development of ALGOL and PASCAL had provided a basis for programming in a hierarchical structure or routines using hierarchical structures of data. These approaches were the beginnings of a serious attempt to convert the art of programming into a process of engineering—creation of a quality product in a predictable time that would perform according to its specifications, and that could be maintained in an operational state over considerable periods of time despite changes in the hardware being used. It was also hoped that an engineering approach would allow the rapid and reliable development of new software products and new versions of existing products in response to new problem specifications.

The most serious effort in this direction, of creating an environment for "software engineering," has been in the ongoing development of ADA, where, because of the large expenditures on software in the military, and the high risks from failures of that software, an engineering approach is desirable. The degree of success of that particular effort is not known as of this writing. It is clear that a highly disciplined and organized approach to creating programs, in which each part of the software is self-documenting and small enough to be thoroughly understood and tested, with the cleanest possible interfaces to other pieces, greatly increases the chances of producing correctly functioning programs in a reasonable length of time.

One of the most significant changes in approach to programming adopted by software engineering has been in the area of "object-oriented" languages. Beginning with work on SMALLTALK from 1971 to 1980, languages supporting this concept allow the creation of objects that respond to messages from other objects by taking appropriate actions, shifting the emphasis in software design from complex sequences of actions to more modularized actions within the mechanisms of objects. Such an approach is particularly appealing when one works with systems requiring many interacting parallel threads of execution. Languages for object-oriented programming provide the ability to define new "abstract data types," in which the user may specify not only the representation of data within memory, but the operations that may be performed on that data, and hide the internals of the representation from processes using the type, so that unneeded dependences on the specifics of a representation will not proliferate.

For more on the history of software engineering, see Ghezzi and Jazayeri (1987).

1.7 Supercomputing

In the major mathematical calculations of physics, the largest and fastest computers of the moment, so-called supercomputers, are pressed into service. The programming languages used are adapted to the special capabilities of those machines. In 1963, Control Data introduced the 6600, a computer capable of 1 million floating-point operations per second and the supercomputer of its day. The machine had a complex and subtle interaction among its operations because some of them happened in parallel, making it difficult to get maximum execution speeds from ordinary FORTRAN compilers. It was common to code critical sections of programs in assembler for speed. Since then, supercomputer hardware has become more complex. The CDC 7600, the Cray-1, CDC Cyber 200 series and NEC SX series, Fujitsu VP series, and Hitachi S-800 series pushed the limits of available hardware speeds for properly organized programs and data. The present generation of machines combines the processing of entire vectors of data in single instructions with the availability of more than one processor to handle a single problem. Preprocessors to FORTRAN, highly optimizing compilers, extensions to FORTRAN, and use of C allow close to optimal execution without having to resort to assembler code, but the programs often have to be extensively redesigned when moved from one machine to another in order to realize maximum speeds.

See Perrott and Zarea-Aliabadi (1986) for more detail on supercomputer languages.

1.8 Parallel Computing

It appears that we may be reaching the limits of speed achievable by a small number of powerful processors, and that the supercomputers of the future will have much greater parallelism. Languages designed for sequential execution of instructions on one processor may not be appropriate for this new generation of parallel computers. Suppose we have a machine with 1000 processors that are able to work in parallel, and we try to convert an existing sequential code to run on our new system. We must ensure that the result of the parallel execution is the same as that of the sequential execution of the program, so we introduce code to synchronize the use of data that must be updated to ensure that it is updated in the desired ordering. If, as a result of this synchronization, we force, say, 1% of the executed instructions to execute on just one processor in total control of the machine, then we can never gain a speed improvement of even 100 times, much less the 1000 times we would like to see from 1000 processors (see PARALLEL COMPUTING).

Many paradigms need to be considered in the design of a language for parallel processing: shared memory versus message passing, tightly coupled versus loosely coupled processors, homogeneous versus heterogeneous processors, many small processors versus a few powerful processors versus many powerful processors. When a large number of processors attempt to share a common memory, mechanisms must be provided to synchronize their use efficiently, or access to a single location may force the processors to "serialize" execution, i.e., to run one after the other, rather than in parallel. When many processors use private memories instead, then careful estimation of the network traffic needed to keep feeding each one of them data is needed. The 1980s have seen an explosion of new languages and extensions of existing languages to handle various types of parallel systems. Of particular interest is the language LINDA (see Ahuja *et al.*, 1986) with data structures based on the concept of Tuple Space. Tuples are ordered sets of values, usually given as a parenthesized list. In a sequential language we would expect to have a fixed address telling us where the tuple is to be found. In the Tuple Space approach, we find the tuple partly by the contents of its fields and partly by its structure, so that the necessary addressing information is carried by the tuple.

The basic operations are "out," which adds a new tuple to Tuple Space, "read," which reads a tuple but leaves it in Tuple Space, and "in," which reads a tuple and removes it from Tuple Space. Trying to "read" or "in" a tuple that is not yet in Tuple Space causes the requester to wait until someone else puts the desired tuple there. This provides the essential interlock to allow effectively simultaneous access to the same data. See Bal *et al.* (1989) for more on the subject of languages for distributed computing.

As with existing supercomputers, there is a certain attractiveness to programming in a variant of FORTRAN, and then having the parallel conversion done by a preprocessor or the language translator. An example is Alliant FX/Series FORTRAN (see Alliant, 1988).

1.9 4GLs, Packages, and Nonprogramming

In the early days of computing, there was a feeling of moving from one generation of computers to another, and from one family of languages to its descendants. With the faster pace of development today, it is hard to distinguish any clear boundaries marking the end of one generation and the start of another. However, the term "4GL," for "Fourth-Generation Language," has come to mean packages and systems of powerful tools for programming that allow someone who is not particularly familiar with the intricacies of programming to solve problems with computers. This includes powerful database query languages, spreadsheet systems, graphics tool boxes, and systems integrated with text processing for the production of complex compound documents. In many cases the need to program is avoided by the use of menus or graphic user interfaces allowing the nonprogrammer to piece together the solution to his problem out of simple known steps and objects. It is to be hoped that the fourth-generation packages will be followed by a "Fifth Generation" of even more powerful systems functioning as expert systems able to assist the user actively in finding the solution to his problems.

With the spread of individual workstations and the decentralization of the management of information, the 1980s have seen the rapidly increasing use of more powerful tools by a wider community of users. Solutions to problems can be pieced together with much greater speed than ever before, but with that speed also comes the risk that the solution, being composed of unexamined black boxes, may not be what was initially intended. There may no longer be a rigorous trace of what was done to achieve that solution, nor even of what it actually is. With powerful tools comes the risk of large mistakes.

For more background on fourth- and fifth-generation languages, see Chorafas (1986).

2. PROGRAMMING LANGUAGE CONCEPTS AND STYLES

The wide variety of uses to which computers may be put has led to the creation of more languages than any one person could learn or use in a lifetime. Indeed, it is common to extend languages, effectively to invent whole new vocabularies, when faced with even modest new problems. Good programmers write extensive libraries of "functions" and "subroutines" that they try to use and further extend for additional applications.

In order to deal with this veritable Tower of Babel, we will present the general concepts of programming languages, with supporting examples from some of the more popular languages used in physics. In any particular application on any particular machine, the specific instructions for use of the dialects of languages used there must be consulted.

Digital stored-program computers provide the means to manipulate numbers and other data by sequences of operations. The data and the memory locations within which they are stored are "objects" on which a computer performs a sequence of "actions." Computer programming languages are tools to perform actions, explicitly or implicitly, on objects. They are designed either to work directly with representations of concepts that derive from the structure and function of digital stored-program computers or to work with representations of concepts that derive from representations of other concepts. In that case, the language translator must translate those other representations into terms appropriate to computers. The objects that must be manipulated are the equivalent of nouns and pronouns in human languages. These objects are called "variables," "constants," "parameters," "arrays," "strings," "pointers," "arguments," etc. in programming languages. There are the actions that may be taken either simply as actions by themselves or on some of these objects, the equivalent of intransitive and transitive verbs in human languages. These actions are called "operations," "functions," "subroutines," "procedures," "handlers," "drivers," etc., in programming languages. The basic elements of a language are enriched by additional descriptive elements that describe various properties of objects and actions, and provide a context that may change the meaning of the objects and the actions. Think of these as descriptive sentences involving adjectives and adverbs. Thus, if we were writing a procedure for cooking a chicken, we might begin with descriptive sentences setting the environment for our procedure:

1. The oven should be at 350 degrees.
2. The chicken should be at room temperature.

Then we might list the actions to be taken on the objects involved:

1. Clean and season the chicken.

2. Put the chicken into the pan.
3. Put the chicken and pan into the oven until 2 hours have gone by.
4. Baste the chicken every 15 minutes.
5. Remove the chicken from the oven.
6. Let the chicken set for 15 minutes.

In this example, we can think of the "chicken," the "pan," and the "oven" all as objects, and "clean," "put . . . into," "baste," "remove," and "let . . . set" as the actions in a programming language.

In programming languages we usually use the name "statement" for sentences of the language. As with human languages, the actions (verbs) and objects (nouns and pronouns) may be in a variety of relative positions, and some of these objects and actions may be understood, rather than explicitly included, in the statement. The three most common orderings of actions and objects are these:

1. Put an action before the objects it affects (prefix notation).
2. Put an action after the objects it affects (postfix notation).
3. Put an action between the objects it affects (infix notation).

For example, if we wished to add the numbers 5 and 6 to get 11, this would be expressed as "+ 5 6" in prefix notation, "5 6 +" in postfix notation, and "5 + 6" in infix notation.

When our emphasis is on the sequence of actions, on doing one thing after another, we consider a statement-oriented, or "imperative" programming, approach. When our emphasis is more on the objects, and we consider the actions as an internal mechanism of the objects that communicate by passing messages among themselves, we are concerned with "object-oriented" programming. Instead of giving the detailed steps involved in chicken cooking, we might define a "chicken cooker" as an object, which might contain a raw chicken or a cooked chicken. We would still have to define somehow the internal mechanisms of the chicken cooker, but we could hide the choice of gas stove versus microwave, basting, etc., from the user of the chicken cooker. If we divorce ourselves entirely from considerations of the order in which actions are to be performed, but consider only the most pure function evaluations, then we are concerned with "functional" programming; and if we refuse to consider actions at all, but only the relationships among objects, we are concerned with "logic" programming. In this last case, we might simply define the relationships between characteristics of a raw chicken and a properly cooked chicken, and leave it to the computer to find a way to transform raw chicken into cooked chicken.

2.1 Imperative Programming

FORTRAN, BASIC, ALGOL, PASCAL, C, and MODULA-2 are examples of languages that support statement-oriented programming, following the approach of our chicken-cooking example above. FORTRAN established the accepted programming style for these languages. BASIC is a simplified, FORTRAN-like language often run interpretively to allow rapid changes and experimentation with a program. ALGOL was a competitor to FORTRAN that allowed a more hierarchical structuring of a program. PASCAL is a descendant of ALGOL with a rich control over data structures and types. C is a more flexible language with less rigid enforcement of data structures and types. MODULA-2 is a descendant of PASCAL that provides support for the concepts of modules and information hiding (see below).

A program in such an imperative programming language will consist of some number of related modules containing some declarations that establish the environment in which the work is to be done (primarily specifying the data structures of the objects to be manipulated) and then a well-defined sequence of steps to perform on those objects. The most powerful high-level action to be taken is the execution of an assignment statement, i.e., a statement that puts new information into an object, usually based on the evaluation of arithmetic expressions, possibly containing references to functions of previously computed data. This statement is a generalization of the simple computer order to "store." It looks like an algebraic equation, but has a different meaning.

Consider the FORTRAN statement

```
X=X**2
```

As an algebraic equation, it would have two solutions, 0 and 1. As a statement in FORTRAN, it is an order to take what is currently stored in the variable X, multiply it by itself, and store it back into X. If X started out as 0.5, it would now become 0.25. It need not be 0 or 1 or any other particular value. Thus, in order to determine the state of an object, what value is stored in a variable or whether any value has been stored at all, we must follow the sequence of actions taken from the beginning and see what has been done to the object. This makes debugging of statement-oriented language programs difficult. Special programs, called debuggers, are used to provide traces of all changes to an object or even of all steps taken.

In order to improve the readability and maintainability of imperative programming language programs, it is desirable to keep small the context of objects that any actions may affect. This is done by breaking up the program into procedures, modules, packages, etc.

A consequence of breaking up a program into small pieces in an imperative programming language is the possibility of "side effects." A procedure may have internal data structures, objects only it knows about, or access to less-than-obviously accessible global data structures. By changes in the state of these hidden objects, different invocations of the same object may produce totally different results. Consider the analogy of an employer dealing with an employee who is having problems outside of work, perhaps with bills that are piling up. On a day when his bank is threatening to foreclose on his house, the employee may simply ignore what his boss tells him and sulk. On a day when he has caught up on his payments, he may perform his job quickly and efficiently. In a global sense, the behavior of the employee is deterministic, but from the point of view of the employer, the employee seems to respond differently at different times to the same inputs. Such variability can be highly desirable when, for example, a procedure for printing pages puts a heading on at the top of each page because it is keeping an internal counter of lines, but in general side effects make it difficult to write reliable software.

2.2 Object-Oriented Programming

In object-oriented programming, we attempt to solve problems by working with objects that can send messages to one another. Those messages may then cause the objects to take some appropriate actions, but, essentially, the actions are hidden, internal mechanisms within the objects. This is a realistic model of the behavior of systems such as trucking companies. Goods delivered to them cause them to dispatch trucks to appropriate destinations. Thus we might define objects called "warehouse" and "restaurant A," "restaurant B," and "restaurant C." The warehouse might contain a variety of inventories of goods, say, "chickens" and "ducks," and allow the operations "arrival" (i.e., store) and "dispatch." The arrival of a truckload of ducks might result in the dispatch of three smaller truckloads of ducks, one to each of the restaurants. The mechanism of computing the quantities of ducks to dispatch to each restaurant need not concern us; it is an internal mechanism of the warehouse, perhaps relating to earlier messages (e.g., standing orders from the restaurants). In a similar way we might define a chicken-cooking machine, which accepts the message that a raw chicken has arrived, and which dispatches a cooked chicken by some unspecified internal means. Other than timing, we would not be concerned that the chicken cooker used a gas oven or a microwave, just that the arrival of a raw chicken resulted in the production of a cooked chicken.

An early example of a computer system operating in a similar manner was the CDC 6600 of 1963, a complex of 11 processors communicating via messages left in mail boxes. This approach was used to provide a highly efficient operating system, but was not made available for user programs. SIMULA-67 introduced the concept of "class," which allowed a data structure to be associated with the code that would manipulate that data structure. MODULA-2, SMALLTALK, and C++ provided similar modularization of "objects."

The important distinction from the norms of simple imperative languages is that in, say, FORTRAN, one could organize data and create routines to manipulate those data. The entire mechanism of how the data would be manipulated normally remained visible, so that a programmer could and often did build assumptions about the internals of that manipulation into the programs that used it. Then when any change was made at the lower level, say to improve the efficiency of the routines handling the data structure, or to add new features to the data structure, every piece of code that used those routines had to be examined, and perhaps changed. In an object-oriented language, by associating the manipulations of the data with the data structures themselves, the internal mechanisms can be hidden from the higher-level routines, and the effects of changes within the lower level routines can be kept purely internal.

However, there can still be side effects in an object-oriented language. Indeed, because of the hidden nature of actions within objects, side effects may become the dominant problem-solving mode. Also, even though there is a considerable relaxation of concern about strict sequential execution of actions on objects by allowing many messages to be outstanding at one time, there is still a strong preference for solving problems by performing actions in some sequence on objects.

2.3 Functional Programming

In functional programming (or "applicative" programming) problems are solved by pure function evaluations, without any possibility of side effects. The concepts began with work of Church on the "Lambda Calculus" in the 1930s and found a viable programming language expression in McCarthy's LISP. Iverson's APL, while not a purely applicative language (nor is LISP for that matter), allows and encourages programming in that style. ML is a more recent example of a functional programming language.

In such a language, the procedures that express functions themselves are valid data, which will be evaluated in some manner consistent with their definition, but not in a strictly determined order. Thus if we were doing a problem in surveying, and were using some lengths and the sine, cosine, and tangent of an angle to compute the boundaries of a field, we would simply write the equations for each of the boundaries in terms of the known quantities, and leave it to the computer to decide on the actual order in which to compute the sine versus the cosine versus the tangent. This functional style can be combined with concepts from other programming styles to achieve a richer language. For example, the abstract data types of object-oriented programming, with their hidden internal mechanisms for performing required actions, can be viewed as valid data types in a functional programming language and the actions recast simply as operators on data to be used in functions without side effects.

Languages for symbolic manipulation of equations are often based on one of the functional programming languages, such as LISP, since they are oriented toward evaluation of formulas.

Because the order of execution is not of concern in functional programming, this programming style may have significant application in efficient use of parallel processors.

For more background on functional programming, see Hudak (1989).

2.4 Logic Programming

We may move still one step further in trying to focus on the application rather than on the operation of computer hardware and consider only the relationships among objects of interest rather than even the implicit actions of functions. This style is called logic programming.

PROLOG was developed by A. Colmerauer in 1972. It deals with the specification of problems on a high level, as statements about objects and the relationships among them, and then allows the user to ask questions about those relationships, making the logical inferences from what is known to what has been asked. This leads to the term "logic programming." PROLOG is an appropriate language for work in artificial intelligence. However, it may be argued that any application of a computer to any problem is an exercise in artificial intelligence, and that the normal human specification of a problem in terms of objects and their relationships could be addressed directly by PROLOG without further programming.

In PROLOG, if we wish to state as a fact that "the child is father to the man," we would write

```
father (child,man).
```

In general we state a fact by putting the relationship first, before a parenthesized list of objects associated by that relationship. Questions about relationships are posed by placing "?-" before the hypothetical relationship. If it can be inferred from the known facts, PROLOG answers "yes"; otherwise "no."

If an object is named starting with a capital letter, then it is treated as a variable that may stand for known or unknown objects. In

PROLOG, when the object for which the variable stands is known it is called "instantiated." When asked a question involving an uninstantiated variable, PROLOG tries to find a match to facts that will allow it to instantiate the variable. The answer may be accepted or rejected by the user. If rejected, the attempts to match continue.

PROLOG uses a comma between facts to indicate "and," and ":-" to show that a fact depends on other facts (making a "rule"), so that we could say "Sally loves men who love dancing and dining" as

```
loves(sally,Men):- loves(Men,dancing),
  loves(Men,dining).
```

PROLOG can work with numbers (integers or real) as well as symbolic identifiers. It has the arithmetic operators + (add), − (subtract), * (multiply), / (divide), and mod (take the remainder), and the relational operators > (greater than), < (less than), >= (greater than or equal to), =< (less than or equal to), and =\= (not equal to). For more detail on programming in PROLOG, see Clocksin and Mellish (1987).

2.5 Concurrent Programming

With all concerns about evaluation as well as order of evaluation taken out of the hands of the programmer, a logic programming language has the maximum freedom to adapt the solution of a problem to parallel and distributed execution. By adopting the view that the system is in open-ended continuous operation, reacting to an environment, rather than transforming input data into output data, and by adopting the attitude that partial problem solutions are interesting, logic languages such as PROLOG have been changed into "concurrent logic" programming languages, one of which is CONCURRENT PROLOG. The approach blends naturally with object-oriented programming style in which objects send messages to other objects, and perhaps delegate them (passing them on) to yet other objects to handle when they are not understood.

These languages are not a particular improvement over imperative languages such as FORTRAN for straight numerical algorithms, since they may execute orders of magnitude more slowly, and not even produce the expected result. They are better suited to solving the challenging problems of making computers able to make humanlike decisions in reasonable times, to reason, and to display artificial intelligence. The role of such languages would then not be in the execution of a numeric algorithm but in helping with the design of the algorithm, or the design of an experiment on the basis of an expert system.

For more on this subject, see Shapiro (1989).

3. LANGUAGES FOR MAJOR SCIENTIFIC CALCULATIONS

When a problem pushes computer resources to their limits, languages that will make optimal use of those resources are needed. This is most easily achieved with languages whose structures correspond to the structures of computers, in terms of sequences of actions to be taken on appropriate data. The first and most enduring of these languages is FORTRAN. It is an evolving language with many dialects and variants, but (with minor modifications) routines written to do numeric calculations in 1959 could still be used in 1989, and most likely will still be useful in 2019. For some things, however, even with the best optimization, FORTRAN may not allow a machine to achieve maximum performance. In earlier days, an assembler would have been used to allow instructions to be written that would extract that last bit of efficiency. Today, this would more commonly be done in a FORTRAN-like language such as C.

We will also look at ALGOL and its descendant, PASCAL, which avoids some of the clumsiness of FORTRAN and which was a popular competitor to FORTRAN for a while. Many important algorithms were published in ALGOL, and an ability to read the language is important. PASCAL is used extensively to handle complex data structures.

FORTRAN is evolving and picking up features from PASCAL and other languages. It is not clear whether the resulting language will be able to retain the performance capabilities of the simpler, existing language.

If one looks carefully at FORTRAN, C, and PASCAL, it is obvious that almost anything one might wish to program could be programmed in any of the three, and with almost equal ease. Indeed, aside from the "syntactic noise" of different statement delimiters, programs in the three languages can be hard to distinguish, and the same well-written codes can be, and often are, transliterated from one to another of the languages to good effect. The significant differences among these languages lie in two areas: ability to help detect errors in writing a program, and what we may call "depth of context."

In any language it is possible to write a program that is expressed correctly in that language, but that does not do what we had intended. PASCAL provides very strong cross-checks to make such errors less common than in FORTRAN or C. However, it is precisely those cross-checks that often make it difficult to use PASCAL for applications that must efficiently push the limits of a computer, as in device drivers for experiment control, which would be more likely to be written in FORTRAN and C.

C, however, shares common ground with PASCAL in allowing a depth of context lacking in FORTRAN. In FORTRAN, the environment of definitions within which we work is usually on one level. Everything about the definition of a name in FORTRAN as a particular type of data structure is visible within the routine manipulating it, because in most dialects of FORTRAN it is not possible to nest complete routines, with their definitions, within other routines. Such nesting can be done for both procedure and data structure definitions in PASCAL and for data structure definitions in C. This allows some of the definitions being used to be brought much closer to the point where they will be used, allowing a disciplined programmer to keep the context that someone reading the code needs to understand to a page or two, but also allows an undisciplined programmer to create a complex hierarchy of contexts, within the depths of which one can get lost.

The difference between approaches is similar to differences between English and German. In normal English, sentences work on a small number of levels of structure; while in German, a single sentence may go on for a paragraph, with complex nested subordinate clauses, and verbs gathered far away from the rest of the sentence. FORTRAN encourages a style familiar to English speakers, while C and PASCAL encourage a more Germanic style of coding. It must be emphasized, however, that such stylistic biases need have little effect on most programs, since the common practice now is to write most programs as sets of independently compiled routines, each of which is so small as not to be likely to make use of many levels of internal structure, providing a depth of context in the chain of external calls, rather than in internal blocks. This approach can be followed in any of these languages, and produces code that is easy to move from one language to another.

3.1 FORTRAN

The family of FORTRAN languages has a simple, imperative style. These languages work with data as integer, real, double-precision real, complex, or logical, and arrays of up to three or more subscripts of such data types. They form expressions by applying the arithmetic operators + (addition), − (subtraction or negation), * (multiplication), / (division), or ** (exponentiation) to numeric constants, variables, array elements (elements of vectors or matrices), and function calls (values returned by other pieces of code that evaluate the value of a function of some given arguments), and assign the values of those expressions to variables or array elements.

Though many variations are allowed on different systems, the following rather restrictive approach is valid on almost all systems. This rigid format is relaxed in many modern dialects, to make typing of programs from terminals simpler and to make it easier to read programs.

A program consists of statements.

Each statement is formatted as a sequence of fixed-length lines (originally intended as the images of 80-column punch cards), with four fixed fields:

Columns 1–5 may contain a number (a nonzero integer in some machine-dependent range, but rarely restricted to less than 32767) to be used as a statement label or may indicate a comment by a "C" in column 1.

Column 6 may contain a continuation mark (not blank, and not the digit zero), indicating that the line on which it appears contains a continuation of the statement that was on the prior line.

Columns 7–71 contain the body of the statement itself.

Columns 72–80 may contain an optional comment (often used for a line sequence number).

The only valid characters that may be used are blanks, the uppercase letters, the digits 0 through 9, the special symbols “=”, “+”, “−”, “*”, “/”, “(”, “)”, “,”, “.”, “$”, and, on all except the oldest systems, the single quote “ ' ”. Most newer systems now allow the use of lowercase letters and additional special symbols, but a program written in the limited “FORTRAN character set” is portable from machine to machine.

Blanks and leading zeros are not significant in statement numbers, and, with minor exceptions, blanks are not significant within a statement and may be introduced or removed as needed to make a program more readable.

A program consists of its main-line code (the “main program”) and whatever subroutines and functions it may need to perform its functions. Each such routine consists of a list of statements, the first group of which (the declarations and other nonexecutable statements) establish the environment within which the following executable statements will work.

The main program is distinguished from the other routines by not having a special header, or, in some dialects, starting with a “PROGRAM” statement. Thus, the general structure of a FORTRAN program would be

PROGRAM name
declarations and other nonexecutable statements
executable statements
END
SUBROUTINE name (argument,argument,...,argument)
declarations and other nonexecutable statements
executable statements
END
FUNCTION name (argument,argument,...,argument)
declarations and other nonexecutable statements
executable statements
END
⋮

continuing with as many subroutines and functions as may be needed in any order.

Logically, these routines may be intended to fit into a multilevel hierarchical structure, but visually there are only two levels: the main program and the other routines. This simple structure may cause visual confusion in large systems with many routines.

Variables in FORTRAN are named with one to six letters and/or digits, but the name of a variable must start with a letter. (More than six characters are allowed in a variable name in most modern dialects.) If the first letter of the name of a variable is “I”, “J”, “K”, “L”, “M”, or “N”, it is assumed to be an integer unless declared otherwise. It is assumed to be (single precision) real for the other letters. Use of variables of other types requires an explicit declaration. The implicit typing of the most commonly used data by the first letter of its name is a distinctive feature of FORTRAN, significantly reducing the need for declarations of variables, but also removing an opportunity for the compiler to check the declared use of a variable against the ways in which it is used by a program.

Arrays are indicated by a name formed according to the rules for variable names followed by a parenthesized list of up to three (or in some dialects more) integer indices. The set of indices picks which element of an array is of interest, as with XX(1) or XX(3) to select the first or third elements of one-dimensional array XX, or ZZ(3,4) to pick the fourth element of the third row of a two-dimensional array. Functions are referenced in the same way as arrays, except that the parenthesized list contains arguments of the function. The arguments are used in some internal calculation of the resulting “return value” of the function and, unlike array indices, need not be integers. The objects upon which the language acts may be constants, variables, array elements, the values returned by function calls, and, via special input-output statements, values found in files of data stored on disks, tapes, or other external media.

The constants and variables are of type INTEGER (whole numbers), type REAL (numbers with decimal points, usually floating-point numbers in scientific notation), type DOUBLE PRECISION (real numbers of higher precision, often double the precision, than those of type REAL), type COMPLEX, and type LOGICAL (having values “.TRUE.” or “.FALSE.”)

There is also a way to handle characters. Early FORTRAN compilers supported character strings in a very limited way, given as “nnHaaaaaaaaaa”, where the “nn” was a count of the characters “aaaaaaaaaa” that followed. Newer compilers, since the mid-1970s, have supported a type CHARACTER, quantities that may have values that are character strings of a known length, given as "'aaaaaaa'", i.e., strings delimited by single quotation marks.

In FORTRAN, files of data are an important construct, but the language does not handle them like other data types. Considering all types except files, the actions that may be taken are the following:

Assignment. A statement of the form destination = expression, where the destination is a variable or array element, first computes the value of the expression and then stores it into the location implied by the variable or array element. If the expression is of a different type than the destination, a conversion of type is done before the value is stored. Thus, the statement

```
I=X+2
```

is a directive to fetch the contents of the variable "X", add the constant "2", and store the result converted by truncation to an integer into the variable "I". If "X" had contained the value 3.6, then, following the execution of this assignment statement, "I" would contain the value 5.

Arithmetic Operators. "+" for addition, "−" for unary negation or subtraction, "*" for multiplication, "/" for division, and "**" for exponentiation. The values being combined may be numbers of type integer, real, double precision real, or complex. In the strictest versions of FORTRAN, the objects being added (or subtracted or multiplied, divided or exponentiated) must be of the same type (or at least compatible types as with complex multiplied by real or real taken to an integer power), but more general mixed-mode addition is allowed in newer variants.

Relational Operators. For integers, reals, and double precision reals, under the same compatible mode restrictions as subtraction, numbers may be compared to produce a type logical result as follows:

".GT.": greater than
".GE.": greater than or equal to
".LT.": less than
".LE.": less than or equal to
".EQ.": exactly equal to
".NE.": not exactly equal to

Caution is needed in comparing type REAL values, since the reduced internal precision of calculations on a computer may make tests for exact equality fail unexpectedly.

Logical Operators. Expressions of type logical may be combined using the operators ".NOT.", ".AND.", and ".OR.". Some dialects define additional logical operators, but these three are sufficient to allow the most general logical inferences by forming expressions [e.g., .NOT.(A .AND. B) is the "nor" of A and B].

Expressions are formed by taking sequences of constants, variables, array elements, and function references separated by operators and grouped by balanced pairs of parentheses. Evaluation works from the innermost level of function calls and parenthesized groups out, subject to a hierarchy of operators. First all the arguments of function references within the expression are evaluated, and then the functions are evaluated on those arguments. Then parenthesized expressions are evaluated from the innermost level out. Operators are applied in the following order of precedence (highest on the list first):

"**"
"*" and "/"
"+" and "−"
".GT.", ".GE.", ".LT.", ".LE.", ".EQ.", and ".NE."
".NOT."
".AND."
".OR."

so that X/Y/Z+D−E+F is the same as (X/(Y*Z))+(D+F)−E, with the compiler having the right to reorder the evaluation to an algebraically equivalent expression and to evaluate common subexpressions fewer times than they actually appear.

In addition to the assignment statement, FORTRAN includes the following executable statements that control the sequence of execution:

The GOTO statement. Transfers control to the indicated statement, as in "GOTO nnnn", where "nnnn" is a statement label in the same routine, which will be the next statement to be executed. There are two additional GOTO statements, the "computed" GOTO and the "assigned" GOTO, which allow the choice of destination to be decided within the program while running, instead of being chosen beforehand.

The arithmetic IF statement. Transfers control conditionally, as in "IF (expression) nneg, nzero,npos", where the numeric value of "expression" is computed, and then control passes to the statement with label "nneg" if the value of the expression is negative, the statement with label "nzero" if the result is zero, and to the statement with label "npos" if the value of the expression is nonzero positive.

The logical IF statement. Executes a statement conditionally, as in "IF (expression) statement", where the logical expression is evaluated, and the statement is executed if the result is .TRUE.

The block IF statement. Similar to the logical IF statement, except, of the form:

```
IF (expression) THEN
   statements to execute if the expression
     is .TRUE.
ELSE
   statements to execute if the statement
     is .FALSE.
ENDIF
```

allowing most GOTOs to be avoided.

The DO statement. Executes a block of code repetitively, incrementing a control variable on each pass. Originally of the form

```
          DO nnnn ivar = ilow,ihi,istep
             statements
nnnn last statement of the loop
```

where the variable "ivar" would start at the value "ilow", and be incremented by the value "istep" (by 1 if "istep" was not specified) until it reached or exceeded the value "ihi". The statements between the DO statement and the statement with statement label "nnnn" would be executed for each value assigned to "ivar" between "ilow" and "ihi". In earlier dialects the statements would be executed at least once (for ivar equal to ilow), even if ihi was less than ilow. This is not done in more modern dialects.

In some dialects this is now written as

```
DO ivar = ilow,ihi,istep
   statements
ENDDO
```

avoiding the need to use a statement label.

The CONTINUE statement. A do-nothing statement convenient for attaching statement labels for the end of DO-loops.

The CALL statement. Transfers control to a subroutine, as in "CALL name(arg1,arg2, arg3,...)" or just "CALL name".

The RETURN statement. Returns from a subroutine to the statement just after the CALL.

The STOP statement. Terminates execution of the program.

The PAUSE statement. Temporarily suspends execution of the program. This statement is not appropriate in a batch environment.

Files are manipulated by a variety of statements, such as "OPEN", "CLOSE", "READ", "WRITE", "REWIND", and "BACKSPACE". Files may include complex formatting of data to and from external representations, either using separate "FORMAT" statements to hold the descriptions of the various fields or including appropriate character strings in READ and WRITE statements. The environment within which the executable statements are executed is established by the nonexecutable statements:

The COMMON statement. Declares that the variables or arrays listed should be made available to other routines listing the same COMMON statement. Otherwise variables and arrays are local to each routine. There may be more than one common area, distinguished by using labels conforming to the rules for variable names. The names for the same variables in different routines need not be the same. The shared use is determined by their physical position within the list of variables, and their types.

The EQUIVALENCE statement. Declares that particular variables or arrays will share the same storage locations. Originally intended to economize on memory when it was a scarce resource, it has come to be used as a way to view the same bit patterns as being of different types.

The DIMENSION and TYPE statements. These statements declare the types of both variables and arrays, and the size of arrays (by listing the maximum values of each of the indices) or, in the case of the DIMENSION statement, just the size of arrays.

FORTRAN includes a large library of built-in functions, which the programmer may replace or add to with his own routines.

For more on FORTRAN, see ANSI Committee No. X3J3 (1989).

3.2 C

C is a language that even more closely reflects the capabilities of the machine being used than does FORTRAN, allowing bit manipulations and use of internal registers so that an assembler need not be used. C normally produces well-optimized code. The syntax avoids much of the card-oriented clumsiness of FORTRAN, using some of the style of ALGOL-like languages (see below).

A C program consists of statements terminated by semicolons, made into nested blocks by "{" and "}" (the equivalent of "begin" and "end" in ALGOL, described below). The objects that may be manipulated in C are built from three precisions of integer, which may also be unsigned for use as bit patterns, from two precisions of real, and from data of type "character." The types of all variables must be declared. Arrays are declared by using brackets with the range of the index (the quantity that selects the array element desired), e.g., "int xx[10];", which has essentially the same meaning as the FORTRAN "DIMENSION xx(10)". Users may define types. Complex data structures may be created by listing the field names with their types in braces in the "struct" declaration. A "union" allows each field to be mapped to the same locations, acting somewhat like an "EQUIVALENCE" statement in FORTRAN. Bit fields can be given field names in structures.

In FORTRAN, most access to data was explicitly by variable name, but a limited ability to point to particular data items in a larger set of data items was implicitly provided by array indices. C allows for the explicit declaration of pointers to objects (including pointers to pointers, etc.) and extraction of the machine address of objects, and types of objects may be explicitly ignored. This flexibility helps in allowing C programs to be used in place of assembler code, but means that a valid C compiler may provide little protection against programming errors.

C names variables as strings of letters and digits, beginning with a letter, but, unlike FORTRAN, accepts both upper- and lower-case letters, and (in most dialects) treats them as distinct. Only the first eight characters of a C variable name are significant in some dialects.

C describes the objects it manipulates in terms of "storage class" and type. Storage class declarations allow the programmer to direct that a variable will claim storage only during the invocation of the block that uses it ("automatic"), or that the variable will have a life extending past that of any particular invocation of the block using it ("static"), or that the variable has a life external to the entire program ("external"). Variables may also be declared as being best stored in fast machine registers, if possible ("register").

The types are

1. short int, int, long int. The representation of three integer types, intended to provide access to the integer sizes natural to the machine.
2. unsigned int. The representation of an integer without a sign bit, obeying the arithmetic rules of "2's complement arithmetic" commonly used for manipulating machine addresses.
3. float. The representation of a real number. The format of type float constants is similar to that used in FORTRAN, except that all such constants are double precision.
4. double. The representation of a possibly higher precision real number than provided by type float.
5. char. The representation of characters appropriate to the machine. A character constant is given as a single character in single quote marks or by quoting special sequences starting with a backslash:

\n	newline (or line-feed)
\t	horizontal tab
\b	backspace
\r	carriage return
\f	form feed
\\	backslash
\'	single quote
\"	double quote
\nnn	the bit pattern given by the 1–3 octal digits nnn.

Character strings are given as "array of char" strings in double quote marks. A zero character ('\0') is appended to string constants to help programs scan for the end of a string.

C also provides an enumerated data type, in the style of PASCAL (see below).

C has FORTRAN- and ALGOL-like expression evaluation, with the addition of a combinedassignment and increment or decrement (a concept from ALGOL-68). The operator "=" is used for assignment. It is just another operator in C, not a special statement as in FORTRAN or PASCAL. The form "xx += 1" is the same as "xx = xx + 1", and "xx −= 1" is the same as "xx = xx −1". Other appropriate operators may be combined with assignment in the same way.

Another special feature of C is the use of "++xx" to mean increment xx before using it, and "xx++" to mean return the value of xx, and then increment it.

C has an even more complex order of precedence for the actions of its operators than does FORTRAN. Parentheses are recommended.

The operators available in C are

1. Unary operators. Indirection ("*") to retrieve the value pointed to by the object, address of ("&") to obtain a pointer to the object, arithmetic negation ("−"), logical negation ("!"), ones' complement ("~"), increment ("++"), decrement ("−−"), type cast (specified by a type-name in parentheses), size in bytes ("sizeof").
2. Arithmetic operators. The operator "*" is used for multiplication as in FORTRAN; the operator "/" is used for division as in FORTRAN; the operator "%" is used to get the remainder; the operator "+" indicates addition; the operator "−" indicates subtraction.
3. Shift operators. Left shift with zero fill ("<<"), right shift (">>"). The right shift has zero fill for unsigned expressions, but may be an arithmetic right shift for signed expressions.
4. Relational operators. Less than ("<"), greater than (">"), less than or equal to ("<="), greater than or equal to (">="), equal to ("=="), not equal to ("!=").
5. Bit-by-bit operations. And ("&"), exclusive or ("^"), inclusive or ("|").
6. Logical operations. And ("&&"), or ("||").
7. Conditionals. The question mark and colon are used to form an if-then-else operation, as in

```
e1 ? e2 : e3
```

   which returns e2 if e1 is nonzero (i.e., true), and returns e3 otherwise.
8. Assignment operations. Simple assignment ("="), other operations (+, −, *, /, %, >>, <<, &, ^ , |) combined with assignment in the form "op"= [e.g., x1+=x2 is equivalent to x1 = (x1) + x2].
9. Comma operation. The comma causes left-to-right evaluation of its arguments and returns the rightmost value.

Operators are applied in the following order of precedence (highest on the list first):

unary operations: * & − ! ~ ++ −− type-cast sizeof

multiplicative operations: * / %

additive operations: + −

shift operations: << >>

ordered relational operations: < > <= >=

equality relational operations: == !=

bit-by-bit and: &

bit-by-bit exclusive or: ^

bit-by-bit inclusive or: |
logical and: &&

logical or: ||

conditional: ... ? ... : ...

assignment = += −= *= /= %= >>= <<= &= ^= |=

comma

Control of the sequence of execution is provided by the following statements allowing compound statements, grouped by braces "{" and "}" with declarations private to the block of statements given after the opening brace, and statements similar to the FORTRAN "GOTO", "LOGICAL IF", and "DO", in this case "goto", "if", and "while", "do...while", and "for" statements.

Additional statements that improve the clarity of C code over FORTRAN in some cases are the following:

1. The "switch" statement. This statement executes one of several statements selected by the value of an expression. It is of the form

```
switch(expression){
  case c1 : s1;
  case c2 : s2;
       ...
  case cn : sn;
  default : sd
};
```

where c1, c2, ..., cn are integer constant expressions. The expression is evaluated and control transferred to the statement listed after the matching constant expression, or to the default if no match is made.

2. The "break" statement. This statement terminates the innermost repeated execution statement or switch statement, avoiding the need for some goto statements.
3. The "continue" statement. This statement ends the current pass of the innermost repeated execution statement, but allows subsequent passes, if any.
4. The "return" statement. This statement is used to return a value from a function and continue execution in the statement where the call was made.

C includes a preprocessor that translates lines beginning with the symbol "#" to allow replacement of identifiers by strings ("#define"), inclusion of lines from other source files ("#include"), and conditional compilation of lines ("#if", "#ifdef", "#ifndef").

Because of its intimate connection with major operating systems development, there is a very extensive library of functions for use with C programs for file manipulations and other useful activities. The libraries used in support of other languages often are written in C.

For more on the C language, see Kernighan and Ritchie (1978).

3.3 ALGOL and PASCAL

ALGOL is a FORTRAN-like language with a less card-oriented, more readable syntax. The major difference from FORTRAN is structuring into nested blocks, allowing a hierarchy of procedure declarations, and a more rigid requirement that the types of all variables be specified, rather than allowing the default integer and real declarations of FORTRAN. The current descendant of ALGOL in heaviest use is PASCAL.

Let us first look briefly at ALGOL syntax. A program is composed of statements terminated by semicolons, organized into at least one BEGIN-END block:

```
BEGIN
   statements declaring the types of variables
   executable statements and procedures
   comments
END
```

Procedures follow the same structure as the outer level after a statement of the form

```
PROCEDURE name;
```

immediately followed by declarations of variables to be used only in that procedure.

On any level the types of the objects to be used are declared by statements beginning "REAL," "INTEGER," or "BOOLEAN," possibly modified by the word "ARRAY" to indicate an array. In FORTRAN each array index implicitly starts with 1. In C indexing starts with 0. In ALGOL, a lower and an upper value for each index are given to allow indexing to start from other values.

The actions that can be taken on these objects are similar to those allowed in FORTRAN, but slightly different symbols are used, depending on the characters available on the particular computer. The FORTRAN "**" for exponentiation can be replaced by an up-arrow (often printing as "^"), and the relational operators by their usual mathematical symbols; the assignment operator is indicated by ":=" to distinguish it from the test for equality. Expression evaluation is similar in algebraic style.

Labels are symbolic, rather than numeric as in FORTRAN, and delimited by a ":", so that no special field needs to be reserved for them. There are the equivalent of the FORTRAN logical IF, the block IF, the GOTO, and the computed GOTO in ALGOL.

Only minimal file handling was specified in the language and varied from system to system until later languages, such as PASCAL.

PASCAL (see Jensen and Wirth, 1974; Wilson and Addyman, 1982) is sometimes referred to as "FORTRAN with semicolons," mixing some of the style of FORTRAN with that of ALGOL. Programs are written as sequences of statements in an imperative style, and formulas and equations translate into PASCAL similarly to the way it is done in FORTRAN and ALGOL. However, PASCAL provides a more natural set of data structures and control structures than was available in FORTRAN at the time PASCAL was introduced, which allowed the writing of cleaner, more comprehensible code, especially when dealing with problems that involved more than simple iteration of the evaluation of mathematical formulas.

The objects that can be used in PASCAL are constants or variables, and variables must have a "type," i.e., the attributes of the values and the legitimate range of values that may be assumed by a variable must be specified. The language starts by defining "simple" types, then provides "structure" types that may be built out of simple types, and "pointer" types that facilitate linking of data together. The four simple types are as follows:

integer-type. A subset of the whole numbers (..., −3, −2, −1, 0, 1, 2, ...).

real-type. A subset of the real numbers (such as 3.141 59, −6 483 221., 0.000 03, etc.).

char-type. The characters that can be handled on that machine, including the upper- and lowercase letters, and the decimal digits, with alphabetic and numeric ordering preserved.

Boolean-type. The logical values "true" and "false."

Except for the real-type, the simple types can all be ordered by integer values, and are called "ordinal-types."

The user can create new types. One can simply enumerate the values that may be taken on by a new type in their appropriate ordering ("enumerated-types"), or pick a range of values out of an existing type ("subrange-types"), or make more complex structures out of existing types. Structure-types may be packed for storage efficiency, or unpacked.

The possible "structure-types" are as follows:

array-type. Vectors ("tuples") of existing types specified by an index set. Because the elements of an array-type may themselves be of array type, matrices may be represented, with the usual comma notation to separate the indices.

record-type. A structure consisting of groups of named fields, where (unlike array elements, each of which must be of the same type) fields may be of different types. This ability to mix types allows one to, for example, create strings of characters with a field of integer-type giving the length and a field or array of char-type holding the characters. Unlike arrays, however, the value of a field identifier cannot be computed. It is just a name. Different record structures ("variants") may be used in a given record-type, determined by the value of a tag field used to select the variant to be used.

set-type. A set composed of elements all of the same type (the "base-type"), which must be an ordinal type. A PASCAL implementation imposes a limit on the cardinality of the base type, usually somewhere between 48 and 256.

file-type. A structure consisting of a sequence of values of a given component type. Such structures provide PASCAL with an input-output mechanism.

In order to facilitate working with lists that may grow and shrink, pointer types are defined:

pointer-type: Either "nil" (i.e., not pointing to anything) or having a value that identifies (points to) data of a type (the "domain-type") associated with the pointer.

This rich structure of types allows the user to build up complex data structures as the fundamental objects of the language. Objects and actions are used to create expressions. PASCAL does not have the same finely detailed precedence of actions seen in FORTRAN, and requires more use of parentheses to avoid ambiguities. The objects and expressions are used in invocations of functions (that return values) and procedures (that do not return values) and in assignment statements. Control over the sequence of execution of statements is provided by a variety of control structures, similar to those discussed for FORTRAN and C. The language allows portions of code and appropriate data structure to be packaged as procedures and functions, and for them to recursively call themselves.

A PASCAL program to add ten numbers and print the result might be written as:

```
program addnum (input,output);
var myinput,count,sum:integer;
begin
  sum := 0;
  for count := 1 to 10 do
  begin
    readln (myinput);
    sum := sum + myinput;
  end;
   writeln (sum)
end.
```

As with our other examples of this problem, we have ignored the niceties of what the program could (or should) do if there are not ten numbers to read, or if there is an error in reading.

For more on PASCAL, see Wilson and Addyman (1982) and Feuer and Gehani (1982).

3.4 Libraries and Language Extensions

Each application of computers and computer languages tends to develop a specialized set of language extensions and libraries appropriate to problems commonly encountered. There are extensive and well-tested libraries of mathematical routines for solving linear equations, differential equations, etc. By imposing a style of programming, of use of data structures, and file formats, these libraries extend the base languages into jargons appropriate to their applications. To pick one example among many, consider the "Basic Linear Algebra Subprograms" (BLAS) by Dongarra *et al.* (1988). This package provides routines for the manipulation of matrices and vectors, using a clear consistent style. The first letter of a routine specifies the precision ("S" for single precision real, "D" for double precision, etc.). The next two letters classify the kind of matrices being handled ("GE" for general, "HE" for Hermitian, "TR" for triangular, etc.), and the next two the operation ("MV" for matrix-vector product, etc.). Storage conventions are specified for triangular matrices.

By adhering to such conventions, the user not only gains the ability to use the routines in the library involved, but also the ability more efficiently to produce new programs that achieve clarity and maintainability by following known conventions; i.e., by speaking a known language.

4. LANGUAGES FOR THE LABORATORY ENVIRONMENT

Computers have become an integral part of the laboratory environment. They control the operation of instruments, move the data from instruments to storage, analyze the data, compare it to models, display it, and archive it. There is a wide range of specialized languages, packages, and systems for this purpose. In some cases the computer aspects are so well integrated with a set of instruments that the user need not be concerned with programming. The system functions as a black box. In other cases, detailed programming is required. In the latter cases, the choice of an appropriate language may be forced by considerations of what languages, if any, may be available for the computers involved. However, when the choice is unconstrained by such questions, typical choices are FORTRAN, PASCAL, C, Assembler, BASIC, and FORTH.

What drives such choices are questions of time—either of the computer or of the person using it. Compared to the demands of some modern experiment designs, computers are slow and can have difficulty moving all the data involved without falling seriously behind. In such cases, highly efficient code, or at least code of highly predictable timing, may be required. Languages oriented toward the automatic management of complex data structures can place demands on the computer for activities such as garbage collection, for which there simply may not be time. In the most extreme cases, use of C or an assembler that can give the maximum control over use of machine resources may be required. Since FORTRAN has a direct and static mapping of its data structures to machine implementation, and has benefitted from decades of effort at optimization of the code it produces, it can often provide almost the same efficiency. PASCAL has been used as a cleaner, simpler language for such applications, but in time-critical PASCAL work has not been as popular as mixtures of FORTRAN with C and assembler.

If, however, the primary consideration is time of the person, then interpretive languages such as BASIC and FORTH become attractive. The speed lost in interpretation with BASIC can be recovered by taking the final version and compiling it. FORTH manages to achieve good speeds by providing what is called threaded code (see Sec. 4.2) and by allowing embedded assembler code use.

The creation of device control code in FORTRAN or PASCAL is difficult because the languages themselves do not provide intimate access to hardware, nor is the timing of code execution easily predicted. They also do not deal directly with asynchronous parallel processes. They do, however, provide a comfortable environment for the creation of data structures, management of external files, and numeric data analysis. Thus a common structure of laboratory code is an outer layer of control logic in FORTRAN or PASCAL, using time-critical services provided by C or assembler subroutines.

Efficient communication among the pieces is essential. For large streams of data, the structure most commonly used is the "circular buffer" or "circular queue." In this structure, two pointers are used, each advancing from the bottom of the structure to the top, and then returning to the bottom to scan the structure again. One of these pointers designates the next location into which data are to be stored, and the other the next location from which data are to be taken. As the pointers chase each other, the gap between them may grow and shrink to deal with the vagaries of timing of the arrival of data and its processing, thereby compensating for the unknown timing of the higher-level language code.

When timing is less critical or ease of modification is the main issue, an interpretive language such as BASIC or FORTH may be used for laboratory work.

4.1 BASIC

There are many dialects of BASIC. The language was developed by Kemeny and Kurtz as a simplified FORTRAN and ALGOL derivative appropriate for students. It is most often used as an interpretive language on time-sharing systems and small computers. Programs consist of sequences of numbered lines. The variables used may be restricted to single-letter names, but it is more common to allow at least a letter and a digit or, in most recent versions, full strings of letters and digits starting with a letter. The types of variables may be either type real or type string, and most current versions allow additional types. The types are usually designated by an extra symbol appended to the name of the variable:

real single precision	!	(or nothing as the default)
integer	%	
double-precision real	#	
string	$	
long integer	&	(on some machines)

Arrays are indicated by a variable name followed by subscript indices in parentheses, and are declared by a "DIM" statement.

Expression evaluation is FORTRAN-like, using the following operations:

1. Arithmetic operations. Addition ("+"), subtraction ("−"), multiplication ("*"), division ("/"), exponentiation ("^").
2. Relational operations. Less than ("<"), less than or equal to ("<="), greater than (">"), greater than or equal to ("=>"), equal to ("="), not equal to ("<>").
3. Logical operations. "AND", "OR", "NOT".

The flow of program control is keyed to the numbers put on each line:

The "GOTO" statement transfers control to the designated line number.

The "GOSUB" statement transfers control to the designated line number. That line and following lines are executed as a subroutine to be executed until a "RETURN" statement. The "RETURN" statement transfers control back to the place immediately after the "GOSUB".

The "FOR" statement causes repeated execution of the following statements in statement number order until a matching "NEXT" statement.

The "IF" statement, of the form "IF expression THEN statement" or "IF expression THEN statement ELSE statement," evaluates the logical expression and executes the statement after the "THEN" if the expression is true.

The assignment statement is as in FORTRAN, except some versions require the word "LET" before the statement.

Files are manipulated by INPUT and PRINT statements.

For device control, many versions of BASIC add an "INP" statement to read device registers and an "OUT" statement to store new values into device registers.

In BASIC the example of adding 10 numbers might read

```
100  SUM = 0
110  FOR COUNT%= 1 TO 10
120     INPUT MYINP
130     SUM = SUM + MYINP
140  NEXT COUNT%
150  PRINT SUM
160  END
```

There are compiled, batch versions of BASIC, some of which avoid the need for statement numbers on every line and which allow the use of symbolic statement labels.

For more on BASIC, see Nolan (1974).

4.2 FORTH

FORTH is a simple stack-oriented, postfix notation language, with a highly structured programming style oriented toward use of small computers. A typical FORTH program is built up from function definitions that are used to create higher-level function definitions. The highest level created is invoked as the solution to the problem. Most implementations of the language make use of a programming technique called "threaded code," which is particularly effective in achieving a compromise between the ease of modification of an interpretive language, and the efficiency of execution of a compiled language.

In threaded code, we create a virtual machine for our application as a set of subroutine calls, and use lists of those calls as the instructions in the virtual machine. Suppose we start with a machine that has only integer arithmetic. Let us define a routine to do floating-point addition, and store it at, say, location 100, and let us define a routine to print a floating-point number and store it at location 200. To keep our machine simple, let us address all the objects we care about by a pointer, called a "stack pointer," which acts like a spring-loaded pile of plates at a buffet. We can only see the top item, the item to which we currently point. We can remove that item ("pop" it off the stack) by changing the pointer by one data object, and place a new item on top ("push" it onto the stack) by changing the pointer in the other direction. Then we can keep our new language simple, by having the addition routine add the two items at the top of the stack, removing both and replacing them by the sum, and having the print routine print what is on top of the stack. A routine to add and print then would just be the sequence 100, 200. This is a compact representation of a complex action. This routine can itself be stored and used as an extension to the language.

For more on FORTH, see Brodie (1981).

5. LANGUAGES FOR GRAPHICS

Computers are a useful tool for capturing, drawing, and modifying images. Their use in such applications increases as raw computing power and storage capacity increase, allowing manipulation of increasingly complex images.

Languages for computer graphics are implemented as extensions to more general languages. The objects to be manipulated are images and the pieces of images, such as lines and circles, as well as the devices that draw and display images, such as pens and graphics screens. A language for graphics must deal both with what is being drawn and the way in which it is being drawn.

The most elementary view is that of a pen being dragged along a piece of paper, as in the language LOGO (Abelson, 1982). The paradigm in LOGO is that of a turtle pointed in a direction, carrying a pen, which may be up or down. The pen is placed down by the command "PENDOWN" and raised by the command "PENUP". A stroke is made in the current direction or its reverse by the commands "FORWARD nnn" and "BACK nnn", where nnn is the length of the stroke. The direction of strokes is changed by "RIGHT ddd" and "LEFT ddd", which specify a rotation in degrees.

Similar minimal graphics can and have been added as libraries or language extensions of most procedural languages, providing at least a vector-drawing capability. However, management of images as objects is simplified when issues of scaling and presentation location are separated from the object definitions themselves by the establishment of windows and viewports. Three-dimensional graphics in particular benefits from separation of the definition of the object from the mechanism of viewing the object, so that orientation, hidden-line removal, ray tracing, and other considerations in looking at the object do not require redefinition of the object. There are many contending approaches to presentation of images, including GKS and PHIGS, with rapid changes being driven by rapid advances in the ability of computers to present increasingly detailed images in real time.

For more information, see COMPUTER GRAPHICS.

6. LANGUAGES FOR DATA RETRIEVAL

Computers manage information. They provide the tools to scan large amounts of information and extract what is needed. Data retrieval works with objects that are identified by their contents, rather than by their physical location. The contents may include explicit keys to help in the search or may stand by themselves.

The paradigm now most used for database languages is one in which data are considered as a set of tables, the columns of which identify attributes of data, and the rows of which are the data items of interest. Some column, or set of columns, serves as the key to identify uniquely the data items to be retrieved.

The operators used are union (combining the rows of two tables with matching columns to form one with the combined information from both), intersection (combining to retain only the common information from both), difference (combining to obtain information in the first table that is not in the second), Cartesian product (combining tables with different columns to make all possible combinations), select (extracting selected rows from a table), project (extracting selected columns from a table), inner join (combining tables by matching entries in a common column), outer join (like inner join, but including rows with no match), and divide (extracting particular rows and columns on the basis of a match in other columns).

A popular language for data retrieval is SQL (Structured Query Language). The typical command in this language has the form

SELECT	the names of columns
FROM	the name of a table
WHERE	the criteria to be satisfied in selecting
ORDER BY	the order of presentation of the columns

The criteria may include relational, arithmetic, and pattern-matching operators.

There are some difficulties in applying this paradigm to data that have search criteria based on considerations other than positions of values within tables. For example, a search on the shape of particular images, or on the symmetries of lattices involving experimental error, may strain this paradigm, and may require the explicit creation of inverted indices organized on the basis of the anticipated search.

For more information, see COMPUTER DATA BASES.

7. SUMMARY

In order to save the effort of writing a new and different program for each machine to solve the same problems, to make computers accessible to more people, and to increase the chance of writing instructions for computers correctly, many programming languages for computers have been created, languages that express the actions to be performed in terms more comprehensible to people, and in ways that allow the same programs to be used on many different types of computers. Some languages are close in style to the operation of computers, consisting of sequences of actions to be taken on objects. Such a style leads to efficient translation of a problem specification into the internal language of a computer, but may force a strained interpretation of the original problem, increasing the chances of programming errors, and increasing program maintenance costs. To reflect more naturally the demands of the problems to be solved, rather than the demands of computers, languages using other styles of programming have been developed: object-oriented, functional, and logic programming. These offer promise of increased productivity for appropriate problems, but for the classic problems of theoretical and experimental physics, FORTRAN and its statement-oriented, imperative relatives predominate.

ACKNOWLEDGMENTS

We wish to thank Frances C. Bernstein, Daniel J. Bernstein, and Lawrence C. Andrews for their helpful comments and suggestions.

GLOSSARY

Assembler: A translator for a symbolic representation of a computer machine language.

Compiler: A translator for a computer programming language that converts an entire program, or major segments of it, as a unit to machine language.

Constant: Datum that does not change.

Function: A self-contained portion of code that returns a single value computed from the parameters given in invoking the function.

Functional Programming: Programming in terms of function evaluations.

Imperative Programming: Programming in terms of a sequence of actions to be taken.

Infix Notation: Programming notation in which an operator is given between the objects upon which it acts.

Instruction: An elementary action that a computer may take, or that a computer programming language defines.

Interpreter: A translator for a computer programming language that converts small portions of a program to machine language immediately.

Logic Programming: Programming in terms of the relationships among objects.

Main Program: The outermost layer of code, which starts and controls the execution of code to solve a problem. It may invoke other routines to solve the problem, in which case control would eventually normally return to the main program, or it may pass on control to some other main program, without the expectation of the return of control.

Object-Oriented Programming: Programming in terms of objects that take actions on the basis of messages passed to them.

Operator: An elementary action on objects defined in a computer language. If there is only one object, the operator is called "unary"; if there are two, "binary," etc.

Order: An elementary action a computer may take; an instruction.

Prefix Notation: Programming notation in which an operator is given before the objects upon which it acts.

Procedure: A routine. See below.

Postfix Notation: Programming notation in which an operator is given after the objects upon which it acts.

Routine: A self-contained portion of code used by other code. It may be a main program, a function returning a value, or a subroutine performing an arbitrary calculation. It may invoke other routines.

Side Effect: A nonobvious change to the state of data structures resulting from the invocation of a function, subroutine, or other procedure.

Statement: The specification of an action to be taken (an executable statement) or the environment (a declaration) in which actions are to be taken in a programming language. It is usually built upon objects and actions, and used in turn to build routines.

Subprogram: A subroutine. See below.

Subroutine: A self-contained portion of code that performs a calculation that may depend on parameters given in invoking the function.

Variable: An identifier for data that may change in the course of execution of a program.

List of Works Cited

Abelson, H. (1982), *Byte* **7**(8), 88–115.

Ahuja, S., Carriero, N., Gelernter, D. (1986), *Computer* **19**, 26–34.

Alliant (1988), *FX/Fortran Programmer's Handbook*, Littleton: Alliant Computer Systems Corp.

ANSI Committee No. X3J3 (1989), *American National Standard for Information Systems Programming Language Fortran S8 (X3.9–198x), version 112*, Washington, DC: Computer and Business Equipment Manufacturers Association (CBEMA).

Bal, H. E., Steiner, J. G., Tanenbaum, A. S. (1989), *ACM Comput. Surveys* **21**, 261–322.

Bashe, C. J., Johnson, L. R., Palmer, J. H., Pugh, E. W. (1986), *IBM's Early Computers*, Cambridge, MA: MIT Press.

Booch, G. (1983), *Software Engineering with Ada*, Menlo Park: Benjamin/Cummings.

Brodie, L. (1981), *Starting FORTH*, Englewood Cliffs: Prentice-Hall.

Chorafas, D. N. (1986), *Fourth and Fifth Generation Programming Languages*, Vols. 1 and 2. New York: McGraw-Hill.

Clocksin, W. F., Mellish, C. S. (1987), *Programming in Prolog*, 3rd, revised and extended ed. New York: Springer-Verlag.

Dongarra, J. J., Du Croz, J., Hammarling, S., Hanson, R. J. (1988), *ACM Trans. Math. Software* **14**, 1–17.

Feuer, A. R., Gehani, N. H. (1982), *ACM Comput. Surveys* **14**, 73–92.

Ghezzi, C., Jazayeri, M. (1987), *Programming Language Concepts*, 2nd ed. New York: Wiley.

Hollingdale, S. H., Tootill, G. C. (1965), *Electronic Computers*, Harmondsworth: Penguin.

Hudak, P. (1989), *ACM Comput. Surveys* **21**, 359–411.

Jensen, K., Wirth, N. (1974), *Pascal User Manual and Report*, 2nd ed. New York: Springer-Verlag.

Kernighan, B. W., Ritchie, D. M. (1978), *The C*

Programming Language, Englewood Cliffs: Prentice-Hall.

Morrison, P., Morrison, E. (Eds.) (1961), *Charles Babbage and His Calculating Engines*, New York: Dover.

Nolan, R. L. (1974), *Introduction to Computing through the BASIC Language*, New York: Holt, Rinehart and Winston.

Perrott, R. H., Zarea-Aliabadi (1986), *ACM Comput. Surveys* **18**, 5–22.

Rosen, S. (1969), *ACM Comput. Surveys* **1**, 7–36.

Shapiro, E. (1989), *ACM Comput. Surveys* **21**, 412–510.

Slater, R. (1987), *Portraits in Silicon*, Cambridge: The MIT Press.

Tenenbaum, A. M., Augenstein, M. J. (1981), *Data Structures Using Pascal*, Englewood Cliffs: Prentice-Hall.

Wilson, I. R., Addyman, A. M. (1982), *A Practical Introduction to Pascal*, 2nd ed. (including the Pascal Standard). New York: Springer-Verlag.

Further Reading

Dijkstra, E. W. (1976), *A Discipline of Programming*, Englewood Cliffs: Prentice-Hall.

Knuth, D. E. (1973), *The Art of Computer Programming*, 2nd ed., Vol. 1: *Fundamental Algorithms*, Reading, MA: Addison-Wesley.

Knuth, D. E. (1981), *The Art of Computer Programming*, 2nd ed., Vol. 2: *Seminumerical Algorithms*, Reading, MA: Addison-Wesley.

Knuth, D. E. (1973), *The Art of Computer Programming*, Vol. 3: *Sorting and Searching*, Reading, MA: Addison-Wesley.

March, S. T. (ed.) (1989), *ACM Computing Surveys: Special Issue on Programming Language Paradigms*, New York: Association for Computing Machinery.

Saib, S. H., Fritz, R. E. (eds.) (1983), *The ADA Programming Language: A Tutorial*, New York: IEEE.

COMPUTERS

JOHN W. CARR, III, *Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.*

Introduction 190
1. **Paradigms for Computer Usage** 191
2. **History of Computer Development** 192
2.1 General Structure of Computer Usage 193
2.1.1 Operating System and Language Interactions 194
2.2 Changes in System Technology Simplifying Usage 194
2.2.1 Changes in Software Technology 195
2.2.2 Changes in Programming Procedures 196
3. **The Classical Sequential Computer Paradigm** 198
3.1 The Control and Data Units 198
3.1.1 Data Unit 198
3.1.2 Control Unit 198
3.1.3 Initial Conditions 199
3.1.4 Continuing Operation 199
3.2 A Corresponding Structural Model 200
3.3 Methods of Speeding Up Performance 201
3.3.1 Use of Static Memory 201
3.3.2 Combined Control and Arithmetic–Logic Unit 201
4. **Actual Implementation of Hardware Configuration** 201
4.1 A Typical PC Architecture 202
4.2 Standard Microprocessor Families 202
4.2.1 Changes Within a Family 202
4.3 Hardware and Software Buses 203
4.3.1 The Central Hardware "Motherboard" 203
4.3.2 Standard Buses 204
4.3.3 Operating Systems as Software Buses and Interactive Interfaces 204
4.3.3.1 Combined Hardware and Software Extensions 204
4.3.3.2 Graphic User Interfaces 204
4.3.4 Asynchronous Hardware–Software Interaction 205
4.4 From Integrated Circuit to Area Network 205
4.4.1 The Integrated-Circuit Chip as a Similar System 205
4.4.2 Design of Integrated Circuits 206
4.4.3 A Higher Level of Area Networks 206
4.5 The Philosophy of Interchangeable Parts 206
5. **The Logical Equivalence of Software and Hardware** 208
6. **Modern Technology** 208
6.1 Applications Programs 208
6.2 RISC vs CISC 209
6.3 Artificial Intelligence 210
6.3.1 Division of Artificial Intelligence into Problem Solving and Learning 210
6.3.2 Neural Nets 211
6.4 Creation and Standardization of Languages 211
6.5 Creation and Standardization of Operating Systems 215
7. **Sequential vs Parallel Processing** 216
7.1 Classes of Parallel Equipment 216
7.1.1 MISD and MIMD Systems 217
7.1.2 Cost of New Large Systems 218
7.2 Relative Serial and Parallel Efficiencies 218
7.2.1 Methods of Improving Performance 218
7.2.2 Examples of Problem Solution Implementations 218
7.3 Areas in Which Parallel Computers Have Succeeded 219
8. **Special-Purpose Equipment Structures** 219

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

8.1 Simplified Microprocessor Controllers and Coprocessors . 219
8.1.1 Real-Time Applications 219
8.1.2 Future Control Devices 219
8.2 Digital Signal Processors...... 220
9. Computer Economics in 1991 220
9.1 Overall Economics of Computers 220
9.2 Standards Organization....... 221
9.2.1 International Economic Conflicts 221
9.2.2 Hardware Becomes a Commodity Product 221
9.3 Fundamental Changes in Scientific Economics.......... 222
Glossary..................... 223
Acronyms 225
Further Reading............. 225

INTRODUCTION

The description of the structure and continuing growth of a new, rapidly growing science, technology, and industry such as computer usage is difficult within an encyclopedia's pages. These three elements interact within an area that is oriented to

1. intellectual effort,
2. information handling over wide geographic areas,
3. computation, and
4. equipment design and sale.

The reader, like the users of such machines, faces a rush of technological and economic details that continuously changes. A static description here will surely be altered continuously in the future. This article will, however, attempt to emphasize those elements which will remain invariant or change slowly as time progresses.

Classically, electronic computers may be classified as analog, using continuous voltages to represent (usually numerically encoded) data, and digital, in which analog pulses are used to represent bits (binary digits). Each bit can be assigned alternative values of 0 or 1 (absence or presence of a pulse). A string of bits can be used to represent many different informational elements by encoding such elements into so-called data structures. The basic element in such a data structure is usually an 8-bit byte or a 32-bit word.

These can be used to represent integers and floating-point approximations to real numbers. In addition, strings of bytes or words can be used to represent alphabetic and other symbolic characters, as well as pictures. With further interconnection into sets of strings, one can represent sentences, equations, and the internal graphical structure of pictures as well. Video images can be represented by digital numerical pixels, or picture elements. Much of computer science involves theory and practice of manipulation and transformation of such elements.

In this discussion, we will concentrate on electronic stored-program digital computers, that class which has gained dominance in modern society. Through success in a competitive market place of vendors, programmers, and users, along with strong support from governmental organizations, digital computers as described herein are at present dominant in their range of applications. Several possible replacement technologies will be mentioned only as a possible prediction of future directions of both research, development, and production.

Digital computers are complex switching devices, often generically labeled hardware, operating upon

1. data, furnished from outside the computers, by either humans or other machines, and
2. programs, labeled software, created and furnished, up until now, outside the computers, almost completely by human programmers.

Data input into computers must be collected by hand or by machine, and communicated on paper or by electronics. Data collection and assemblage must be considered as outside the range of this discussion. Data output is on paper, on cathode-ray tube or alternative electronic presentation devices, or by audio sounds or video, television-like graphical images.

Programming is now, in most cases, a labor-intensive human-performed operation. The process is marked by possible errors in the finished product. In addition, one needs continuing so-called maintenance of the created sequences of instructions over time, since the program itself may attempt to solve a real-world changing problem.

Programming in the most common case consists of the generation of a series of control operations. These, intermixed with data transformation operations, direct the hardware through a complex path. The end result of the process is transformation of the input data to produce output. The output data are required to satisfy some specified goal—to yield the solution of the problem.

In this survey a top-down approach will be presented, which emphasizes the fundamental structure of the elaborate computational system that has been developed.

1. PARADIGMS FOR COMPUTER USAGE

Because of the interposition of elaborate programs between the user and the computer itself, many users interact using paradigms in which the computer is disguised as a special purpose device, or machine. Among the external interfaces that have been set up by which users may interact with computers are the following:

1. The classical file cabinet, with files which can be viewed by "popping them up" on a cathode-ray tube (CRT) screen. Various "windows" operating systems allow the user to point at or name the wanted files and act upon them, opening and closing windows into different sets of text and data.
2. The classical matrix-like spread sheet, in which the user can name or point at entries, describe them in terms of equations, enter parameters, and operate on rows, columns, individual elements, or the entire matrix, with the computer automatically updating the entries.
3. The classical drafting procedure, in which the CRT face acts like the drawing board, and the user operates on various lines and surfaces with various pseudodrawing tools, including simulated pens, paint brushes, erasers, etc., actually consisting of electronically interacting artifacts. Such devices can be used to animate synthetic pictures in time.
4. The classical video screen, with operations to merge, delete, overwrite, recolor, etc., frames from external sources.
5. The classical notebook, with pen-like devices being used to point to and write onto an interactive representation, which changes and operates the computer inside a notebook-sized container.
6. The classical library card catalog, where the user can search for the desired information, which is presented on the CRT screen. Such information systems (with or without the actual card image) are called databases. As yet the card paradigm is incomplete, in that more often special languages are used to search out the information.
7. The classical typewriter, extended to the word-processor and desktop print shop, where copy is edited on the CRT face and then output using various sized type fonts intermixed with drawings and photograph, on laser printers.
8. The classical teletype system, extended into electronic mail or e-mail, such that users with keyboards can mail text or other information, including fax-like material, to other computer users on a worldwide net.
9. The classical bulletin board in which users, using e-mail, can post documents of their own choosing in a formal or informal exchange of information throughout a group.
10. Cooperative exchange of equipment, in which one user is able, by means of network interconnection of computers, to use programs and equipment located elsewhere geographically or administratively.

Still generally unavailable are widespread voice-mail integration of the telephone network paradigm into these systems, or video conferencing. However, as technology provides wider communications bandwidths, these appear to be coming. So-called video-tex systems, mainly in France and elsewhere overseas, have provided rudimentary picture-information network systems, like libraries or information desks. It has thus become apparent that almost any human information behavior, network, or system is a possible paradigmatic model for computer application.

Many of these paradigms have made it easy, as time went on, for nonusers of computers to learn to solve a restricted but important class of their own problems. However, it is still very difficult, in many cases, to move spread sheets onto video screens, images onto e-mail, etc., pictures and graphs into reports. This interconnection of the paradigmatic systems remains a partially achieved task.

Each time a paradigm becomes useful and important, it may need to be connected to previous such structures, and the sets of these interfaces grow markedly, often geometrically. Behind these user–interface paradigms, of course, flows a history that deals with actual equipment and the complicated programs that run upon it. As these change, the implementation of the paradigms for users must also, in many cases, be reworked.

2. HISTORY OF COMPUTER DEVELOPMENT

Historically, computer equipment arose from the vacuum tube technology of World War II radar. At that time fabrication was both expensive and unreliable. Assembly of individual discrete parts into large one-of-a-kind systems required detailed handwork (such as wiring) by human beings, which raised costs and produced many fabrication errors. These became known as hardware "bugs"; the term was carried over into software or programming.

The first electronic digital computer to be publicly operational was the Electronic Numerical Integrator and Calculator (ENIAC) (shown in Fig. 1), at the University of Pennsylvania in 1946. Other early systems were soon operational in laboratories of U.S. security agencies, the Atomic Energy Commission, and the National Bureau of Standards. Research projects were started in many British, Continental, and Soviet governmental laboratories, and in other universities in the U. S., Canada, Great Britain, and elsewhere. By the mid-1950s commercial equipments, pioneered by such firms as Univac, IBM, NCR, and others in the U.S., and Ferranti, ICL, Bull, Siemens, and others on the Continent, began to make their appearance in extensions of problem solutions previously attempted with classical punched-card equipment.

Costs of these electronic systems began to decrease and reliability to improve in the 1960s. The use of discrete transistors as the switching devices still required them to be interconnected by hand labor at high cost. Only after the labor-intensive circuitry interconnection methods for such devices were replaced by printed circuits and circuit boards produced by machine operations did the cost of digital computer hardware begin to fall.

Mass production of computers came only after 1970 with the invention and development of the integrated circuit (IC). This combined automated production of transistors and wiring on the same silicon base (or substrate). Using photolithographic-like technology, firms such as Texas Instruments, Intel, Motorola, and their Japanese competitors NEC, Toshiba, and others began to produce small versions of centralized computers.

These so-called microprocessors were produced using a different kind of high technology assembly line, in which solid-state physics and chemistry replaced mechanics as the underlying science guiding design and fabrication. This mass production of the central kernel of the digital computer had been paralleled by mass reproduction of software and data input, using, first, punched cards, then magnetic tape and the portable "floppy diskette" of magnetic material. The remaining economic bottleneck to truly efficient mass production and distribution remained in the design and fabrication start-up costs for both equipment and programs.

Design costs for integrated circuits began to drop with the development of computer-aided design (CAD) tools for integrated circuits. As quantities of an individual IC unit, or chip, began to increase, fabrication costs came down. Analogous design techniques for programs, called computer-aided software engineering (CASE), have not yet proved as successful. With both techniques, beginning about 1980, a mass market for both dealer and direct-mail sale of desktop computers began.

Continuing improvements were made in the distributed software. Programs are now sold using mass-published and mass-marketed copies, at relatively low costs with respect to the hardware, on small, inexpensive floppy diskettes. These fabrication and distribution advances now make individual computer interaction a potential for every technologically sophisticated worker in the modern

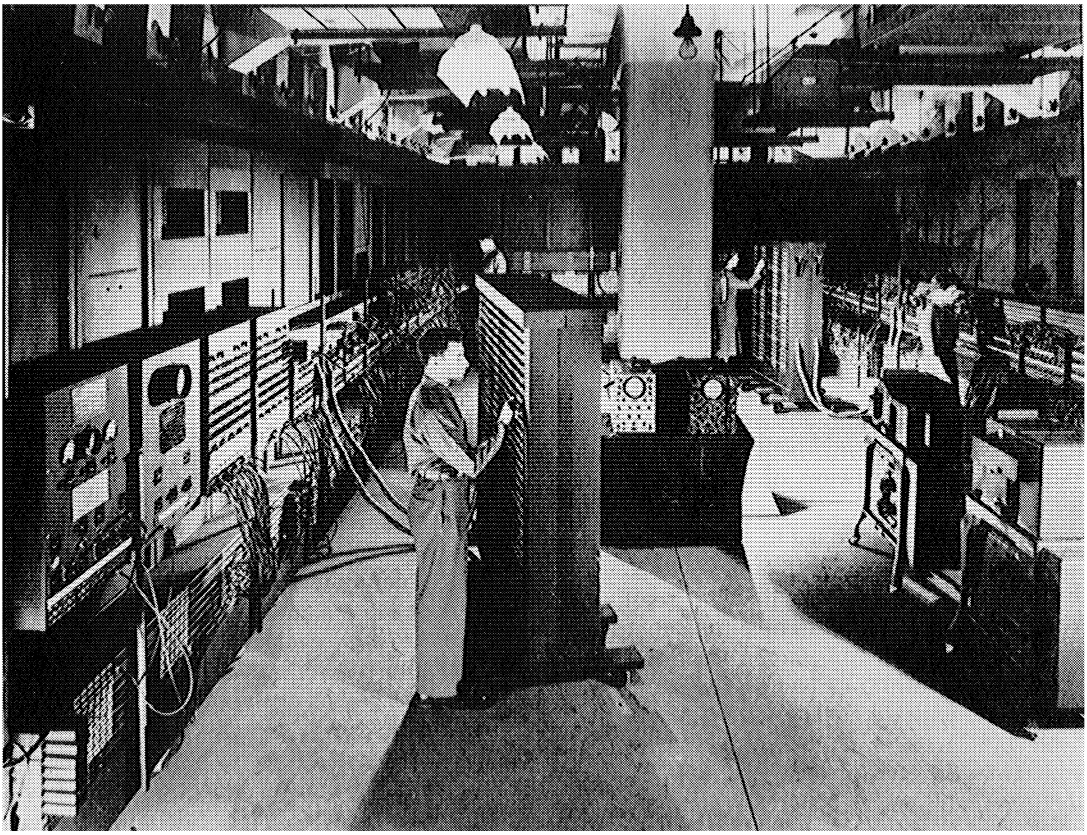

FIG. 1. The electronic numerical integrator and calculator (ENIAC), first designed and completed in 1946 at the Moore School of Electrical Engineering of the University of Pennsylvania, and used at the Ballistic Research Laboratories of the Aberdeen Proving Ground in the post-World War II era. (Courtesy of Moore School of Electrical Engineering, University of Pennsylvania.)

world, and for many for whom their use will be routinely integrated into their jobs.

2.1 General Structure of Computer Usage

Users should, it appears, view solution of problems of any sort in light of the entire computational system which has become economically available. Such problems may be computational or data processing, real-time control, or database searching. A useful model is to view the process of computation from the top down, or outside in. One starts with the problem and works inward towards program and computer, rather than starting with the detailed technology of the system and working outward.

Of primary importance is the problem, or class of problems, to be solved. Some type of applications program, a solution procedure, must be selected by the user. Such a program is the highest level performer in the hierarchy of program, operating system, and computer. The program (or programs) may have already been in operation, available for purchase, or may need to be written from the beginning. The program must interact with at least two other systems elements:

1. an operating system, or control program that performs the housekeeping for the applications program, on
2. a specific computer, or integrated system of computers, that accepts the language of the applications program, as integrated with the operating system program.

2.1.1 Operating System and Language Interactions. The operating system serves as a buffer between the applications program and the actual machine hardware, the computer itself. It manages the specific idiosyncracies of the input–output elements of the computer, and of the secondary mass storage devices such as magnetic disks. The applications program communicates with the operating system in a language which that system can interpret. In general, the applications program does not have to communicate directly with the hardware. For this reason, an operating system has the potential of offering a unique interface to many applications programs. Yet at the same time it may be tailored to control the hardware of many different computer versions.

Similarly, the language in which an applications program is written may be passed through a compiler–translator which outputs machine code for a particular machine. Different versions of this compiler–translator may produce resulting machine codes that are compatible with different machine hardware. Both operating systems and compiler–translators offer the possibility for the introduction of standardization, in which the programming languages, operating systems languages, or actual computer hardware is defined in a standard fashion. This allows ease of use, design, and manufacture in the interaction between these various elements.

The potential user, having decided on an applications program to solve the proposed problem, must decide on an operating system to accept the applications program, and a computer itself to accept the operating system. In many cases, the latter two are already decided upon as a result of previous analyses and decisions. This may restrict the available applications programs for solution purposes.

In some cases, if no solutions program exists as yet, or none can be obtained or modified to suit the need, someone—user or applications programmer—must write a program to interact with the operating system and at the same time solve the problem. In today's complicated computational systems, a professional programmer will most often be needed here, although for certain problems, so-called higher-level languages or graphical interfaces may allow problem-oriented users to create programs.

2.2 Changes in System Technology Simplifying Usage

In the early 1970s, computers were mainly used through the good offices of professional programmers, who wrote programs under the guidance of noncomputer specialists, fed in data, and monitored and produced the results. Three technological advances changed this situation: time sharing, integrated circuits and cheaper memory systems, and interactive graphical computing.

1. "Time sharing" was developed as a method by which many terminals (CRTs with circuitry) were connected to a large central main-frame computer, and shared its machine cycles among them. These terminals were originally mechanical teletypes, later electronic displays with keyboards. The central computer responded to individual users at these terminals so that in appearance, at least, each user seemed to have control of the central machine. The latter was actually sharing its operation sequences among the many terminals. This change allowed (particularly important) interactive usage of central machines from keyboards at remote sites. Such interactive time sharing included both computing and programming from the terminals. This sped up a previously much slower cycle based on punched cards, not keyboard strokes, as input. Time sharing also required design and development of networks connected between terminals and computers. Terminals themselves soon began to contain microprocessors, and thus became smaller local computers. The original time-sharing networks became computer-to-computer networks, expanding to many-computer interconnections called local area networks (LANs).
2. The development of the integrated circuit described above, along with the parallel development of computer-aided design, began to produce small digital computers. Such personal computers (PCs) based on these microprocessors became available for a few thousand dollars. Speeds continued to increase with decrease in size of the integrated microcircuits. Costs and size of the resulting computers decreased. The combined functions of the resulting computers also increased markedly. Low-cost, small-

size computers (see Fig. 2) began to appear not only in homes but also in small and large businesses. These—combined into networks—began to replace or extend the original far more costly main frames, as the physically large and more expensive original designs were called. By 1991, microprocessor chips with half a million to a million transistors were being produced in large quantities. These were sold at lower and lower costs, with former large-size circuitry and interconnections moving onto the chips themselves. At the same time, the cost of computer memory, mainly dynamic random access memory (DRAM), was coming down because of decreased transistor and wiring sizes on the IC chips. The result was the production and sale of large numbers of these cheaper microcomputers, which could be used in the same way as the time-sharing terminal, but without the need for the central main-frame computer. The sharp distinction between large main-frame computer and microcomputer began to blur with the expansion of personal computer performance and the development of so-called "work stations" with power equivalent to that of previous large computers. It is possible as of 1992 to obtain a computer that can comfortably be held on one's lap, yet performs at levels comparable to those of desktop computers (see Fig. 3).

3. Interactive graphical computing was bolstered by addition of electronic screen-pointing devices for the cathode-ray tube terminal. The "mouse," first developed by Xerox, was commercialized successfully by the firm Apple Computer. Its availability allowed the Macintosh computer to remain a firm competitor of IBM's ubiquitous PC as both began to enter the offices of major businesses internationally. Only after so-called windowing systems began to appear on the work stations and PC systems (about 1990) did they begin to compete in effective nonprogrammer user interaction. Hardware was thus combined with elaborate software, to produce what became known as graphical user interfaces (GUIs) (Sec. 4.3.3.2). These began to replace previous programmer-oriented, rather than user-oriented, interactive interfaces. Newer systems thus place the emphasis on making the interaction between user and computer less complex and more "user friendly." The paradigms for users mentioned above have sped up the so-called learning curve on the part of novice and nonprofessional programmer users. For the first time the computer user rather than professional programmer is more nearly in control.

2.2.1 Changes in Software Technology. Technology in software design and "construction" has lagged behind that of hardware. This is probably because the problems involved in such creation are more difficult and have often been ignored or passed on to the pro-

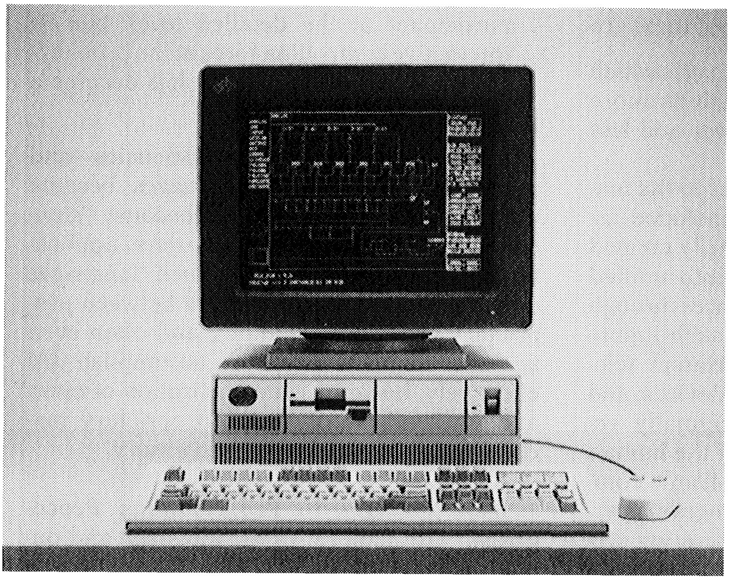

FIG. 2. A personal computer, rated about 100 times as fast as the original microprocessor-based computers, containing up to 8 megabytes of primary storage, a hard disk drive with up to 200 megabytes of secondary storage, and a powerful graphical display and high-resolution laser printer. (Courtesy of IBM Corporation.)

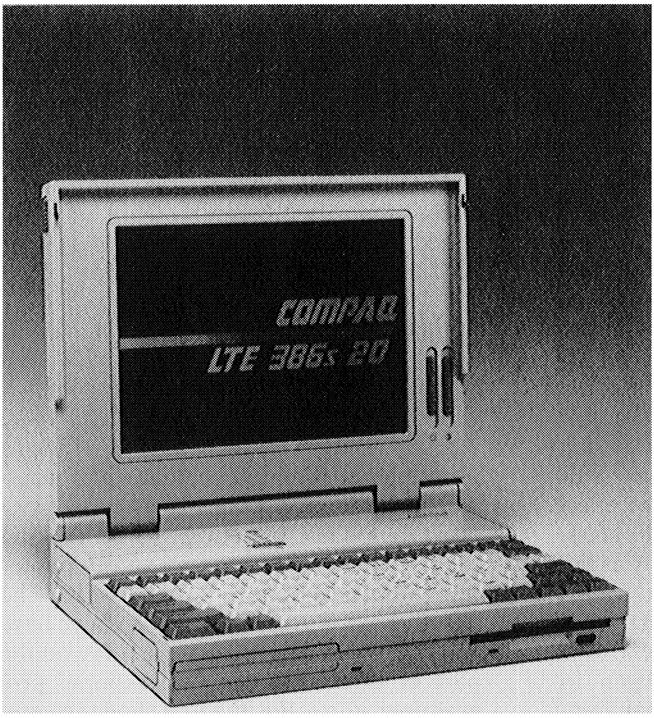

FIG. 3. A laptop computer, weighing about 6 lb, and containing up to 8 megabytes of storage and a $3\frac{1}{2}$-in. floppy diskette drive. Such computers perform at speeds and with programs comparable to desktop computers, which are also microprocessor oriented. (Courtesy of Compaq Computer Corporation.)

gramming creators by the hardware designers. Until now, there has been little computer-aided design for the generation of software. Here again handcrafting of program sequences has most often been the rule. The classical procedure is diagrammed in Fig. 4.

Many software tools, however, are now in use to aid applications programmers in creating and running their problem solution program. Among the more important of these are

1. applications programming languages which made the sequence of instructions more nearly that of the problem solver and less that of the machine itself;
2. translators from such languages to the machine language itself, which included assemblers that assemble previously created and tested sequences (code) into unified programs, and automatic interfaces through the operating system (OS) to such input–output devices as terminals, printers, telephone lines, and mass storage devices; and
3. operating systems, which originally replaced the behavior patterns of the human operators of the original main-frame computers by automatic control programs performing equivalently. Later operating systems changed many decision-making procedures from the human programmer to the system itself at run time. They also provided standard interfaces that applications programmers could use to guarantee correct performance. Thus the OS began to perform more and more "housekeeping" tasks for the applications program and programmer. The computer operator more and more became an onlooker rather than participant at the detailed level, but an interactive controlling force at the problem-solving level, made at what has become a problem-solving console.

These changes led to new efficiencies—and constraints. As certain languages became widely used, they began to become "standards" for a growing group of users. Applications programs written in such languages could be passed back and forth between proprietary hardware machines, and often even across machine families. To accomplish this effectively, however, standardization of operating systems beyond such proprietary machine families also became necessary.

2.2.2 Changes in Programming Procedures. In addition to the language and operational "utilities" described above, certain

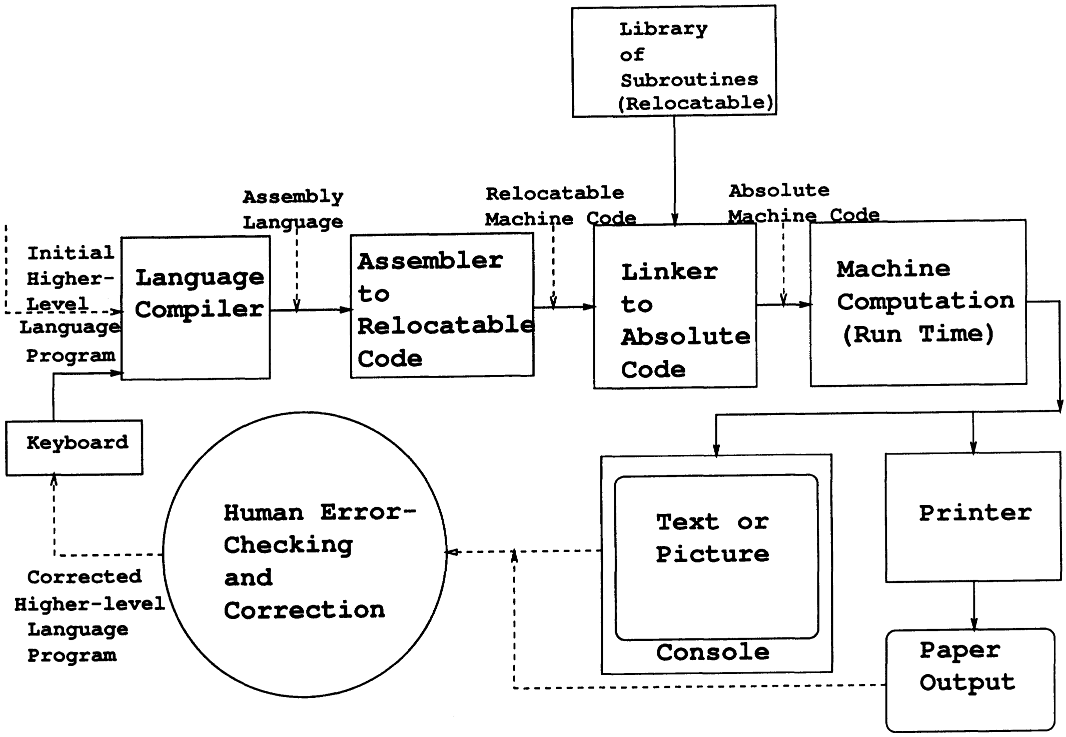
Library of Subroutines (Relocatable)
Initial Higher-Level Language Program
Assembly Language
Relocatable Machine Code
Absolute Machine Code
Language Compiler
Assembler to Relocatable Code
Linker to Absolute Code
Machine Computation (Run Time)
Keyboard
Human Error-Checking and Correction
Text or Picture
Console
Printer
Paper Output
Corrected Higher-level Language Program

FIG. 4. A diagram of the classical programming procedure, originally with teletype or punched cards as input and only printed paper as output. Now interactive keyboards and cathode-ray tube displays speed up the feedback in correcting errors and changing performance of programs.

fundamental strategies for producing computer programs more nearly automatically began to arise. Some of these were highly complex, such as the tactics of artificial intelligence (*q.v.*) and expert systems. These proposed to copy the behavior of human mental processes to aid in solving real-life problems. In certain areas, such as governmental intelligence and certain medical diagnostics, claims were made that such programs produced results beyond the behavior of human experts.

The concept of "objects," as developed with object-oriented programming, created classes of programs and data structures that passed along historical traits by a process of "inheritance," and abstracted out many programming details so that programs might be used over a broad range of analogous applications. This type of programming, combined with special graphical user interfaces, appeared to produce an interactive design process that benefitted from rapid interaction between programmer and programming system.

More immediately applicable computer-aided design procedures for programs have been collected under the title of computer-aided software engineering (CASE). Here the new programming processes of the graphical user interfaces, operating systems, and interactive programming are combined to aid in the generation of alternative versions of certain standard applications programs. CASE has most often been applied to replace the generation of simple, but nevertheless expensive, programs in commercial and data processing applications, previously produced by hand coding. The results are intended to be user-friendly programs. The generated programs themselves interact with nonprogrammer users to create specialized data handlers and transformers.

The resulting programming process is now a mixture of both labor-intensive human analysis and design, combined with machine aids that attempt to guarantee ease of maintenance of the resulting code. Some systems keep a complete record of the historical changes in large programs, combined with histories of certain tests and their results. As software generation has become a more nearly engineering process (as opposed to art form), the introduction of engineering design and test procedures, similar to those of integrated circuit hardware design, increases.

Software costs continue to grow as a major economic factor in the use of computers. Until adaptive program generation techniques begin to produce correct programs automatically, human programming will remain a major constraint on computer utilization.

3. THE CLASSICAL SEQUENTIAL COMPUTER PARADIGM

This section describes the original stored-program computer and gives both abstract and concrete descriptions of how it has been extended with memory and input–output, becoming the central element in local area networks, client server systems, etc. Later, how it has been replicated into parallel computer systems will be discussed.

The simplest model for a digital computer, now only an approximation to the smallest portions of very elaborate systems, is that of the stored-program digital computer. Often credited to Burks, Goldstine, and von Neumann—who published first—it was first proposed by the creators of the ENIAC, Eckert and Mauchly, at the University of Pennsylvania in the mid-1940s. It is useful to compare these basic ideas with the structure of the most popular present design, the IBM Personal Computer.

Parts of such present-day popular microcomputers copy the separated-memory model of the "Harvard machine" of the 1940s, with separate program storage [now in read-only memory (ROM) in the IBM-PC basic input-output system (BIOS)] and separate data storage (in the PC, always in DRAM).

Most of the systems today, however, store other program instructions and data intermixed in the same storage medium, usually DRAM, following the model of the Pennsylvania designers and von Neumann. In the discussion that follows, the classical model is slightly modified to adapt to changes in structure and perception since it was originally proposed. The structure of such a computer can be viewed, as with most such artifacts, by means of both a behavioral and an architectural model.

3.1 The Control and Data Units

The behavioral model provides an idealized, abstract description of how a single digital computer behaves, described in terms of two units, each formally so-called finite-state machines—"FSMs." The behavioral units are pictured in Fig. 5.

Two separate units, control and data, act in a time-interleaved fashion in this model. Each unit has inputs, outputs, and internal states, drawn in each case from a finite set of alternatives. Passage of information between the two units, as described in Fig. 5, is performed in a "ping-pong" fashion, starting with original initial address and data states inserted from outside.

First the control unit uses its initial starting address to generate or retrieve an instruction, which becomes the control unit output. This becomes the data unit input, and is then performed by the data unit, transforming the internal data in the data unit. At the same time, an output is sent back from the data unit to the control unit. This output becomes the control unit input. The control unit uses this input to choose a new address, and the cycle is repeated, beginning at the first word of this paragraph.

3.1.1 Data Unit. In more detail, the data unit holds the data state (consisting of many bits representing numerical and linguistic or structural information). It possesses arithmetic and information-transforming operations which change the data states. The data state is transformed at each step or cycle of machine operation, as a function of the input data operation coming from the control unit. Its outputs are test results sent back to the control unit as the result of test operation inputs.

3.1.2 Control Unit. Outputs from the control unit are operations sent to the data unit, which perform specific transformations on the data states, or test them (in present microcomputers, usually as "flags") for properties such as sign, overflow, etc., generating outputs which are sent back to the control unit, usually of "true" or "false" nature.

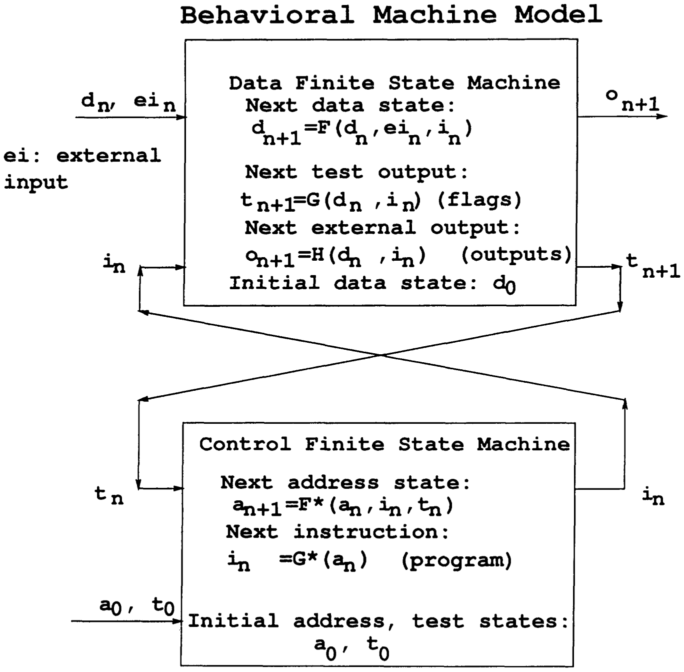
Behavioral Machine Model
d_n, ei_n
ei: external input
Data Finite State Machine
Next data state:
$d_{n+1}=F(d_n, ei_n, i_n)$
Next test output:
$t_{n+1}=G(d_n, i_n)$ (flags)
Next external output:
$o_{n+1}=H(d_n, i_n)$ (outputs)
Initial data state: d_0
o_{n+1}
i_n
t_{n+1}
Control Finite State Machine
Next address state:
$a_{n+1}=F^*(a_n, i_n, t_n)$
Next instruction:
$i_n = G^*(a_n)$ (program)
Initial address, test states:
a_0, t_0
t_n
a_0, t_0
i_n

FIG. 5. A behavioral diagram of the classical sequential computer, described in the text. A data unit transfers data upon input of instructions from the control unit, which changes its address as the result of test information from the data unit, sending instructions on as a function of the address. The latter is the program.

The control unit has its own internal state, usually called the address state, which changes at each cycle of the joint system. Ordinarily, the succeeding address is obtained in integer succession, with addresses represented as integers. However, results of test operations on data change addresses out of the regular integer succession sequence. It is the latter result which allows "transfer of control" and "loop control" between instructions in the control unit.

Some operations sent to the data unit are more complicated, such as test operations which also transform data states (such transformations are often termed "side effects" for operations or programs that are not logically clean).

3.1.3 Initial Conditions. The user of such a system (which may be a human, or another program such as the operating system) must furnish

1. the initial address state (which is where the program is to be started) in the control unit,
2. the initial data state (which generally consists of all of the initial numbers, integers, and other parameters which specify the exact problem to be solved),
3. the next-address function, which takes data outputs and the previous address, and decides on the new or next address. This function, in the mathematical sense, must be supplied as a part of the computer program from outside this system, and
4. the next-data function, which takes control outputs (operations to be performed) and the present data state, and generates a new data state. This function is hardware, wired into the system and generally not changeable from outside.

3.1.4 Continuing Operation. Once these four elements have been set (either at hardware design time or program and data entry time), starting the computer initiates the cycle of successive (address, data) changes until a "halt" instruction is reached, which leaves the control unit in an address with no output instruction. In actual systems, there generally are no halt instructions. Once one program terminates, it supplies the initial address and data to another program (which might be some other program controlling it, for example).

3.2 A Corresponding Structural Model

Corresponding to the behavioral description given above, which only describes elements being manipulated as abstract, coagulated units, one can produce a structural model, which is usually what the actual hardware designer works from. One such model, with four units interconnected in a more elaborate fashion, is given in Fig. 6. Parts of these units may belong to one or the other of the control and data units above. Engineering and hardware realities dictate a more elaborate representation.

The four units of the present usual hardware configuration are usually connected by one (or sometimes more than one) common electronic highway or "bus." Some systems have separate instruction and data buses; others have special buses to match different speeds of information transfers of the units that follow:

1. Memory or storage unit: In "stored-program machines" the address-instruction function (or program) in the control unit is stored as instructions in positions given by machine addresses (usually implemented in the integrated-circuit DRAM). In addition, data states are stored in addressable positions, usually as short (8-, 16-, or 32-bit) sequences, as are the instructions.
2. Control unit: In this, the control or address state is represented as a changeable stored integer, along with the hardware next-address function. This latter function changes addresses depending upon test results of flags, or individual bits developed as the result of test instructions.
3. Arithmetic–logical unit (ALU): In this unit, data states, with changes occurring in words in the memory, are transformed according to the definition of the arithmetic or other data operations, wired in and defined by the machine's designers. Key data information is generally stored in accumulators or registers (analogous to the accumulators on the original desk calculators) and transformed by operations on one or more of these, with results fed back to the same or a different accumulator.
4. Input unit: This unit brings in new data from outside sources such as keyboards, telephone lines through modems, analog-to-digital converters bringing in audio and video information, and floppy diskettes and hard disks with programs and data. These are stored in registers or memory to change the computer data state or instruction program.
5. Output unit: This unit passes information, including data and control commands (instructions), to effector output devices such as printers, cathode-ray tube and video displays, digital-to-analog converters, telephone lines, and the floppy diskettes and hard disks.

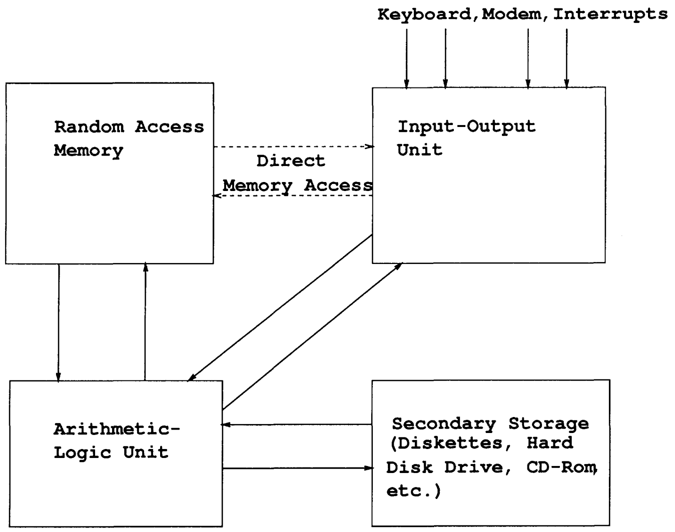
Keyboard, Modem, Interrupts
Random Access Memory
Input-Output Unit
Direct Memory Access
Arithmetic-Logic Unit
Secondary Storage (Diskettes, Hard Disk Drive, CD-Rom, etc.)

FIG. 6. The classical four-unit structural model of a sequential computer, with arithmetic-logic unit, storage unit, and input and output units to communicate with the outside world.

One can also consider the floppy diskettes and hard disks, along with compact optical disks and magnetic tape units, as representing a unit:

1(a) Secondary storage unit or memory, in which a mechanical device backs up the initial electronic random access memory (RAM).

3.3 Methods of Speeding Up Performance

Hardware designers, in efforts to speed up performance often slowed by mechanical or lower-speed electronic information transfers, use buffering techniques and parallel memory transfer as expediters. Most common is the so-called "cache" memory, a smaller, more expensive high-speed memory that holds most-often or most-recently used instructions and data. In many cases, sophisticated compilers and assemblers allow these artifacts to be "transparent" to the user-programmer at some level, although quite complicated algorithms may be necessary to achieve best performance.

3.3.1 Use of Static Memory. At the present time, the memory or storage units are most often DRAM. These are mass-produced integrated circuits whose time of information access has continually decreased during the past ten years. At the same time, the number of memory units (bytes or words) per chip has moved upwards the megabyte range. With increase in the density of the storage transistors on the chip, costs have come down. These units are called "dynamic" because they do not hold their information without constant refreshing—reading and writing back the storage contents. Alternative, higher speed but higher priced, storage elements are the static RAM, which do not need this refreshing process. Other memories (as used in the PC BIOS) are used with permanent program or data storage, ROM. These contain information (data or instructions) stored at fabrication time, or later changed only by substitution of new chips.

3.3.2 Combined Control and Arithmetic-Logic Unit. The control unit and ALU have from the time of the first microprocessors been combined on a single integrated circuit chip, or silicon element. Transistors are used to perform the switching functions necessary to change addresses and data. More recently, as more transistors became available on a single chip, further input and output units of increasing complexity have been added to the central structures. All of these units contain smaller portions of their own memory, usually fabricated as high-speed local static RAM.

The overall speed of both memory and the other units is dependent on the speed of action of individual transistors. This is usually described in terms of a "clock" or high-speed counter, which synchronizes storage and shifting of pulses between different elements on the chip. Present commercial microprocessor clock rates range from 20 to 100 MHz, or millions of cycles per second. This means that the smallest time duration for local information transfer might be in the neighborhood of 10^{-8} s. (Actual arithmetic operations, which require more complex pulse manipulation, would be longer in duration.)

4. ACTUAL IMPLEMENTATION OF HARDWARE CONFIGURATION

The model for actual hardware design is not of the single sequential computer, but, as on the integrated circuit (IC) chip, of a set of parallel subsystems made up of transistors mounted on a common substrate. On the chip, the silicon substrate is the physical common element that allows electronic interaction by transistor switching. The parallel subsystems interconnected on this substrate can interact both synchronously and asynchronously.

A synchronous system is locked into a common clock, which coordinates the timing of the many subsystems in the overall systems. Common actions—moving or complementing pulses—are performed at (more or less) the same time when a clock pulse, propagated all over, triggers local transistor switching.

Most simple ICs are controlled synchronously; the upper physical size of a synchronous system becomes dependent on the maximum delay in clock pulse propagation from one point of the chip to another (called "clock skew"). An asynchronous system is one in which a local action may have to wait on the completion of a more remote operation elsewhere in the system. Such systems contain inherent delays, obviously, as well as possible blocking conditions (two subsystems waiting on each other to complete each one's present task, for example). There exist some asynchronous circuits at the IC level (the classical "flip-flop" circuit is an example), but most asynchronous interaction comes at higher levels, between chips, computers, independent memory units, etc.

4.1 A Typical PC Architecture

A typical computer of the PC variety will have a synchronous, clock-controlled microprocessor as its central processing unit (CPU). The CPU, however, will change its control asynchronously upon being interrupted by such input–output units as keyboard, hard-disk controller, and telephone modulator–demodulator (modem). In addition, if other so-called coprocessors are present (such as a high-speed floating-point arithmetic unit), the microprocessor will asynchronously pass information back and forth at arbitrary, irregular intervals dependent only on the size of different parameters and corresponding calculations. Sometimes this "coprocessing" technique moves information so as to by-pass the CPU, such as in the direct memory access (DMA) process which captures a bus (hardware highway) for a time, in order to move information back and forth between storage and input–output.

4.2 Standard Microprocessor Families

Large-scale main frames and minicomputers more often have had structures which were later copied in microprocessors, and so it is useful to look at microprocessors as the most prominent examples (at least in manufactured quantity) of single processors. In practice, microprocessors have over time come to be fabricated in large quantities. In this process, the number of different hardware architectures has come to be narrowed down to families, usually created by one manufacturer, with upward compatibility of programs. Original systems, such as the Intel 8080, were later expanded into more complicated 8086, and then 80186, 80286, 80386, and 80486 (Fig. 7) systems, over a period of about ten years. A comparable family developed by Motorola started with a very simple 6800 microprocessor and expanded upward through the 68000 series. There are other families for less standard purposes (e.g., digital signal processing, by Texas Instruments), but quantity occurrence in the marketplace has often been restricted to a few families.

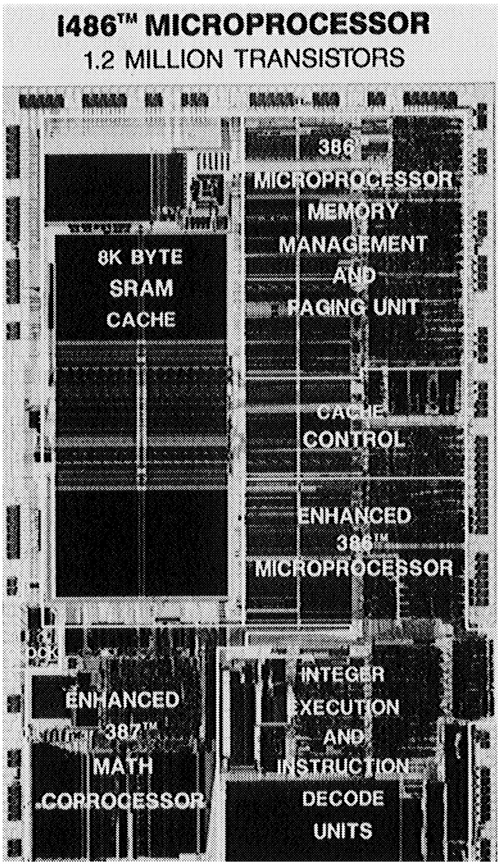
i486™ MICROPROCESSOR
1.2 MILLION TRANSISTORS
386 MICROPROCESSOR MEMORY MANAGEMENT AND PAGING UNIT
8K BYTE SRAM CACHE
CACHE CONTROL
ENHANCED 386™ MICROPROCESSOR
ENHANCED 387™ MATH COPROCESSOR
INTEGER EXECUTION AND INSTRUCTION DECODE UNITS

FIG. 7. An Intel 80486 integrated-circuit microprocessor, showing the detailed wiring and transistors that make up this state-of-the-art complex instruction-set computer (CISC). (Courtesy of Intel Corporation.)

4.2.1 Changes Within a Family. Successive numbered versions of such a processor family usually consist of chips in which added architectural features are adjoined at each stage. In addition, basic clock rates increase as transistor rectangle size decreases with improved fabrication technology. "Upward compatibility" of programs attempts to guarantee that programs written for previous versions in the family will perform on later versions, thus allowing older programs to be used with newer processors. Through the use of interfacing software such as operating systems and compilers, this is often accomplished.

Among the architectural features which vary are those which improve speed of computation and decrease cost of fabrication. The central limiting factor in such a process is often the bus or joint information communication path on the chip (see BUS SYSTEMS AND COMPUTER INTERFACES). Such a bus appears abstractly not only in hardware on a chip, but at a higher circuit-board level, and abstractly in the software operating system as well.

4.3 Hardware and Software Buses

Abstractly, the common highway or bus of the microprocessor chip, from which subsystems are connected asynchronously, has a parallel in the common interconnection structure of the software operating systems. Just as subcircuits are designed independently and then connected on a chip, subprograms (often called subroutines) are designed independently with standard behaviors, and then interconnected to form a working program. Such programs can be "utility" programs, used to perform housekeeping tasks for the overall computer system, or parts or wholes of application programs created to solve specific problems.

The standard hardware bus will have a set of wires, divided into subsets transmitting data, instructions, and control information back and forth between the asynchronous (and sometimes synchronized) subunits. The addition of a new electronic subunit consists of attaching its inputs and outputs to the appropriate data, instruction, and control wires of the bus, so that those common bus signals can be manipulated by the new subunit. Thus, a special memory, input, or output system can be mounted on a printed circuit board and plugged into the bus to expand the capability of the original system. Obviously, along with the new circuitry, in most cases appropriate changes and additions to the central microprocessor program instructions will also be necessary in order to synchronize with the control signals to the newly added unit.

4.3.1 The Central Hardware "Motherboard". Those parts of the computer hardware which are fixed at fabrication time (e.g., of the standard PC) are mounted on what is called a motherboard. This contains wired connections between sockets into which the IC chips are to be inserted. These standard chip sockets are usually mounted on a printed circuit board, along with a series of open sockets connected to the bus. Such sockets on board or bus may be filled at manufacturing time, or later by the vendor or actual user. The standard chip sockets allow the possible potential substitution, e.g., of a later microprocessor in a family series for an earlier slower version. This is a substitution at the chip level. The bus sockets, on the other hand, allow the interconnection of extra boards, fitted with appropriate pins, into the bus. Such extra boards, often from a different manufacturer than that of the original motherboard or of the entire system itself, allow an extension of the memory, control, or operating capability of the original motherboard configuration.

Thus, the flexibility of the motherboard with its sockets along the standard bus allows many different designers and fabricators to add new or different subsystems to the original system, provided they follow standard socket and bus specifications.

These generally have been set up with standard signal, timing, and electrical characteristics, by specific manufacturers or groups of manufacturers, or by IEEE or other standards committees. Such buses, unless accepted widely, tend to again become proprietary in nature, locking in users for a period of time to a restricted set of sources of hardware modifications and extensions. The hardware bus is thus the standard interface by which disparate subsystems can interact.

With these ideas of flexible socketing, memory and microprocessors can be extended or replaced, at least within the electrical and physical limits of the bus system itself. Many of the newer paradigms for computer users, including audio, video, and telephone interaction, have been achieved by development of special-purpose boards able to be plugged into available sockets on a bus. Sequential and parallel data transmission, and digital-to-analog converters (as well as the inverse a-to-d devices), were first added to original personal computer configurations by special boards. Powerful graphics systems, themselves often containing microprocessors, can be added into standard buses, provided accompanying programs are made available. Later, many such functions have been moved down to the central microprocessor ICs themselves, since these have been growing continuously in available functionality.

4.3.2 Standard Buses. Present standard buses in the PC world include ISA (for Industry Standard Architecture; used in the original PC), EISA, an extension, and the Micro Standard Architecture of IBM's PS-2 newer configuration. Newer standards generally have a wider, faster bus to fit the later-developed microprocessors and memory that are to be combined on the motherboard configuration.

Other standard buses used in larger, faster work stations (which have grown out of proprietary buses), such as Multibus, VME bus, etc., were originally designed around a particular microprocessor family, but later in some cases were able to be used with other equipment. The larger minicomputers and mainframe computers also have their standard buses, to allow "third party" vendors to design and furnish extensions to the original architecture.

4.3.3 Operating Systems as Software Buses and Interactive Interfaces. A similar commercial-industrial method of furnishing software extensions to computer control systems uses the operating system as a "software bus," which allows the "socketing" of new programs and data structures into the overall computer software structure. Corresponding to the wires of the hardware bus are parameters in the programs which are to be "plugged in" to the operating system's data storage. Standard transfer of control procedures correspond to the control wires of the hardware bus. Since CPUs most often contain only one processor at a time, classical operating systems have asynchronous-like features only when multiprogramming (performance of two programs on the same single processor) is being carried out.

4.3.3.1 Combined Hardware and Software Extensions. Standard operating system definitions allow separate hardware producers or vendors to specify extensions to the software. The results are intended to operate correctly, if they correspond to the system constraints. Among the best known of operating systems was the original OS-360 for the IBM-360 series, which allowed many applications to be successfully integrated into this IBM main-frame series of computers. For the PC systems, Microsoft's disk operating system became such a standard that all application programs for PCs, for many years, have used the MS-DOS rules for interaction with that operating system. More recently, the Windows program for the same computer family has belatedly introduced a GUI standard. The power of the Apple Macintosh computers was in the existence of a combined GUI and operating system many years earlier, using the original "mouse" screen-pointing concept pioneered by Xerox, and now available on Windows.

Broader standards for windowing systems, e.g., X-Windows, are appearing as GUIs for faster work-station systems made by Sun, Hewlett-Packard, and others. These systems, like many recent computers now being marketed, have as their common extended software "bus-like" structure the UNIX system. This operating system, originally developed at Bell Laboratories, is becoming, more and more, the OS of choice for hardware manufacturers, applications program vendors, and users. Its combination with a successful common standard GUI may become the future overall standard for programs generated by different originators.

UNIX was originally designed for word processing and text manipulation at Bell Laboratories, but its use has been markedly extended over about 20 years. It contains the software equivalent of many of the hardware techniques implanted on chips or motherboards, including software coprocessors, splitting of processes into parallel tasks, and complete control of files, following the file-cabinet paradigm. It works with many interacting terminals dealing with a central system, and allows mutual interactive operation on common files, under appropriate controls. Only recently have memory prices decreased enough so that UNIX systems appeared economic on PC platforms.

4.3.3.2 Graphic User Interfaces. GUIs like the Macintosh clipboard–file-cabinet paradigm extend user interaction to graphical presentations which allow the mouse pointer to manipulate files by pointing to "icons" or "buttons" on the screen (Fig. 8), rather than requiring lengthy key-stroke sequences as with UNIX or the PC-oriented DOS. Microsoft's windows 3.0 extends the Macintosh behavior by directly linking various programs through

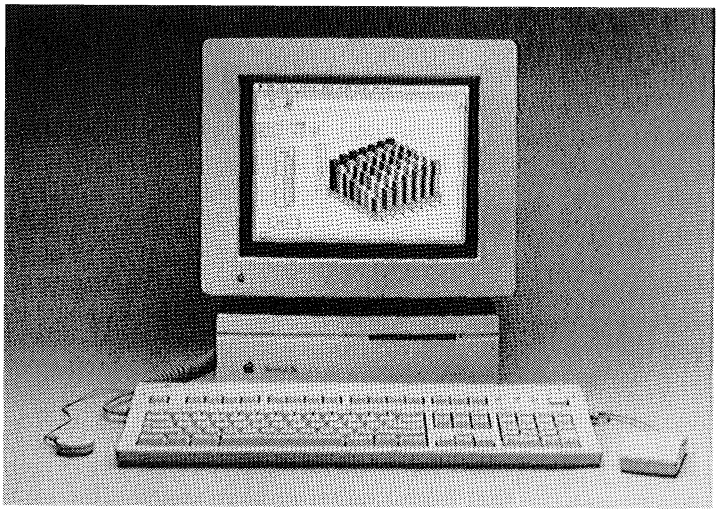

FIG. 8. A graphical user interface, in use with a "mouse" to point to icons, "buttons," and specific windows, text, and pictures on the screen. This point-and-click process bypasses much of earlier character-oriented interfaces on microprocessor operating systems. (Courtesy of Apple Corporation.)

dynamic data exchange, which will cause a chart from a spread sheet, embedded in a word-processing program, to change as individual cells in the spread sheet are changed. This is an example of computational paradigms being interconnected.

4.3.4 Asynchronous Hardware–Software Interaction. The asynchronous behavior of hardware systems obviously causes a need for appropriate response by programs. Two methods of having CPUs such as microprocessors (and much larger machines) respond to the real-time (often mechanical) behavior of the outside system are as follows:

1. "Polling": The CPU synchronously, at regular intervals, looks at electronic control signals from other units, and responds by changing its internal software control to respond to that unit with highest priority to be serviced. This of course places a burden on the CPU, which must spend a portion of its time in checking on its environment. (This is indeed analogous to a "poll" of a population to learn each individual's situation.)
2. "Interrupt processing": The CPU only responds when interrupted by an external unit (analogous to a telephone ring), and then transfers its internal control to software which responds to the particular situation. This system is much faster and much more efficient, but also requires more complex software and computer operation capability on the part of the ALU and peripherals.

This programming behavior is, of course, very similar to the behavior of interconnecting hardware devices interacting in a "master-slave" relationship or requiring a "handshake" procedure to guarantee that both subsystems have the same common control information. Many of the tactics of circuits have found their way into programs, and conversely.

4.4 From Integrated Circuit to Area Network

4.4.1 The Integrated-Circuit Chip as a Similar System. Because integrated-circuit chips need efficiency both in space and speed, the structure of a chip follows many of the characteristics of the larger systems, particularly since entire microcomputers are now being moved onto single chips. Individual chip configurations are, in general, synchronous with respect to the fundamental clock cycle frequency, which times basic transfers between transistor–capacitor storage elements.

Larger subsystems on many chips act relatively independently and asynchronously, waiting on completion of actions by other subsystems before starting on a local task. Such interactions often contain feedback that corresponds to rudimentary hand shaking. This is the more elaborate interaction between higher-level subsystems to guarantee that mutual tasks can be continued because local subgoals have been reached.

Communication on a two-dimensional planar chip is of course constrained by that planarity, somewhat relaxed by the availability of multiple levels of wire crossing.Thus, not every subsystem can easily connect directly to every other, and the concept of bus is forced upon the designer, as a standard pathway with standard interfaces that allow independent subdesign.

4.4.2 Design of Integrated Circuits.

IC chips were originally designed by "hand," with drawings made of the many levels of silicon and metallic compounds that were to be superimposed. The entry of programming technology, particularly in the case of graphics, began a continuing development of "silicon compilers" (see COMPUTER-AIDED DESIGN IN ELECTRONICS). These latter were oriented towards linguistic or GUI descriptions of the circuits, which were to be translated into the final sequences of rectangles to be deposited on the silicon substrate. The most sophisticated of such programming systems were aimed at creating application-specific integrated circuits (ASICs). Among the many applications that have been moved from software onto hardware circuit boards and finally onto ASIC chips are graphics, telephone modems, spelling checkers, games, disk controllers, and many other key "bottleneck" processors where lower cost or higher speed is absolutely necessary to make such behavior economically feasible. These similarities in structure, creation, and economics of various levels of computer software and hardware emphasize the abstract (although certainly not concrete) similarities of these varied forms of information storage and manipulation.

4.4.3 A Higher Level of Area Networks.

Once low-cost computers began to appear on many desks, it became apparent that their interconnection had the potential for mutual information processing and control among many users. Microcomputers and work stations are limited to the number of common interactive users that they can accept; they cannot carry out such interactions and at the same time perform requisite calculations. From this need and opportunity came the LAN concept (Fig. 9), which allows machines to call on storage, programs, and even program performance from and by other machines. The development of such commercial standards as that of Ethernet and corresponding packet-switching networks first produced LANs, and then began their interconnection into wide-area networks.

At present, LANs are usually configured out of a collection of PCs or work stations, acting asynchronously except for bus message communication, which is synchronized to a standard cyclical system. This means that LAN designs are constrained, like the structure of chips, as well as the structure of PC interrupt processing, to those design tactics that guarantee the absence of deadlocks and other fatal interactions. One can thus note that from smaller to large systems, the same rules of synchronous and asynchronous interaction apply:

1. individual transistor circuits;
2. interactive subsystems on a chip;
3. chip sets in a cooperating system, such as on a PC motherboard;
4. interactive multiprocessing systems;
5. common-bus parallel machines;
6. many-processor locally connected massively parallel machines; and
7. LANs.

It is this potentially recursive, replicative nature of the hardware and software artifacts that have been created that allows individual humans the possibility of understanding their behavior. At some near future stage, however, it may be only through embedded computer descriptions within themselves that humans will be able to maintain enough knowledge to make satisfactory use of such systems.

4.5 The Philosophy of Interchangeable Parts

The sometimes chaotic growth of all these various combined hardware and software systems illustrates a growing organization of their structure. Under the constraints of technology and the economic-sociological environment of the marketplace, a process classically

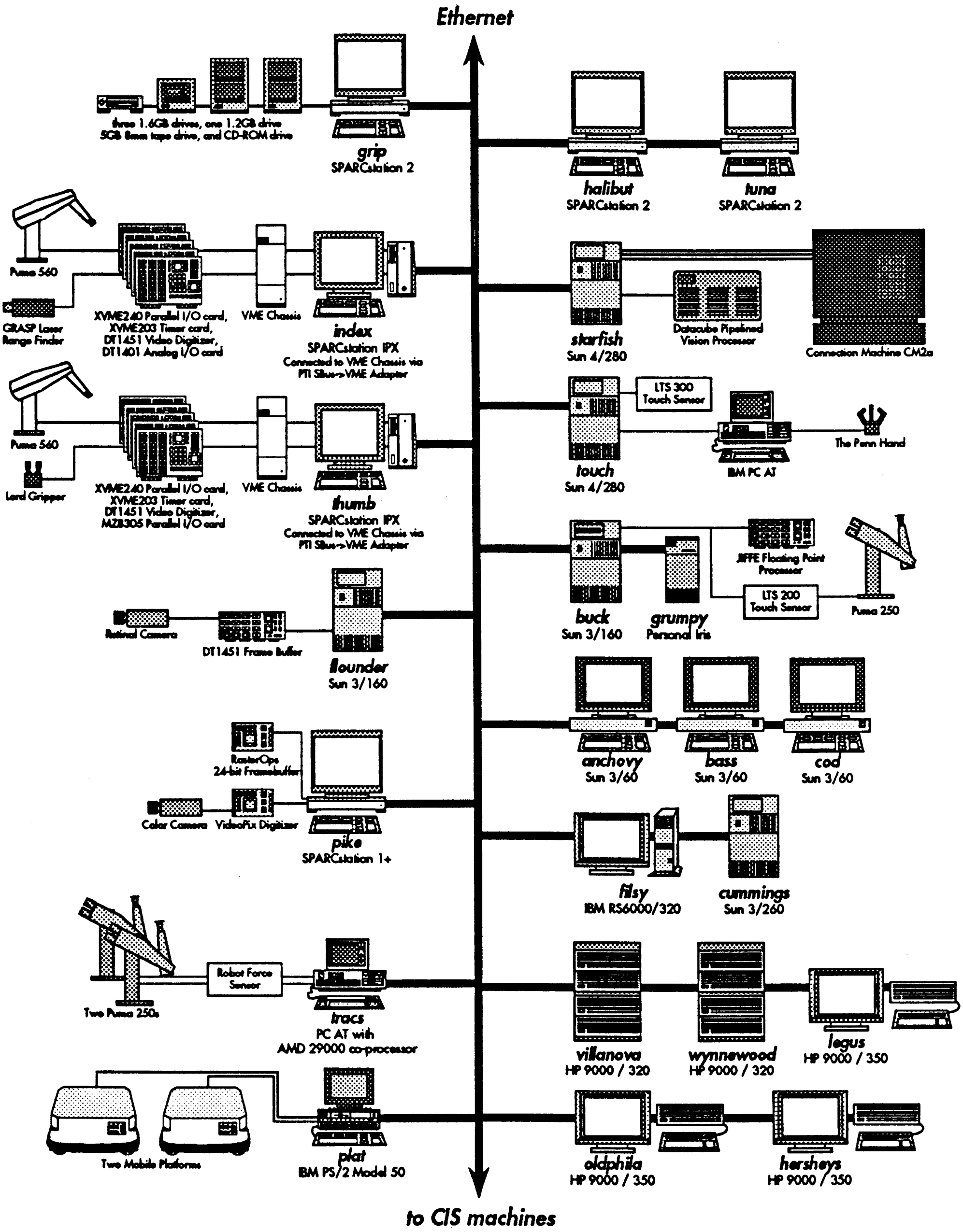
Ethernet
three 1.6GB drives, one 1.2GB drive
5GB 8mm tape drive, and CD-ROM drive
grip
SPARCstation 2
halibut
SPARCstation 2
tuna
SPARCstation 2
Puma 560
GRASP Laser
Range Finder
XVME240 Parallel I/O card,
XVME203 Timer card,
DT1451 Video Digitizer,
DT1401 Analog I/O card
VME Chassis
index
SPARCstation IPX
Connected to VME Chassis via
PTI SBus->VME Adapter
starfish
Sun 4/280
Datacube Pipelined
Vision Processor
Connection Machine CM2a
Puma 560
Lord Gripper
XVME240 Parallel I/O card,
XVME203 Timer card,
DT1451 Video Digitizer,
MZ8305 Parallel I/O card
VME Chassis
thumb
SPARCstation IPX
Connected to VME Chassis via
PTI SBus->VME Adapter
LTS 300
Touch Sensor
touch
Sun 4/280
IBM PC AT
The Penn Hand
buck
Sun 3/160
grumpy
Personal Iris
JIFFE Floating Point
Processor
LTS 200
Touch Sensor
Puma 250
Retinal Camera
DT1451 Frame Buffer
flounder
Sun 3/160
anchovy
Sun 3/60
bass
Sun 3/60
cod
Sun 3/60
RasterOps
24-bit Framebuffer
Color Camera
VideoPix Digitizer
pike
SPARCstation 1+
filsy
IBM RS6000/320
cummings
Sun 3/260
Two Puma 250s
Robot Force
Sensor
tracs
PC AT with
AMD 29000 co-processor
villanova
HP 9000 / 320
wynnewood
HP 9000 / 320
legus
HP 9000 / 350
Two Mobile Platforms
plat
IBM PS/2 Model 50
oldphila
HP 9000 / 350
hersheys
HP 9000 / 350
to CIS machines

FIG. 9. Diagram of a local-area network (LAN), showing the asynchronous interaction between semi-independent processors which nevertheless may call on each other for data, program, or actual execution of program and data in combination at a remote site. (Courtesy of GRASP Laboratory, Dept. of Computer Science, University of Pennsylvania.)

akin to the "interchangeable parts" systems of 19th century New England has been developing.Standards are beginning to guarantee that individual subpieces of the overall system are replaceable, sometimes by software, sometimes by hardware. As long as the interfaces are well-defined as to interconnection syntax and behavioral semantics, individual designer groups can work on improving subsystems without interfering with the remainder of the system.

The evolution of computer systems thus can be viewed as a continued design process, making use of improvements and changes in technology as increased physical information is discovered, obeying flexible but locally fixed laws of interconnection.

5. THE LOGICAL EQUIVALENCE OF SOFTWARE AND HARDWARE

Through the systems optimization, or at least improvement, process that goes on alternatively with new hardware and software design, runs a common thread. It may be stated as a somewhat constrained principle: Computer software and computer circuitry are logically equivalent; at any time an example of one may be replaced by an equivalently behaving example of the other.

This is certainly true in that any circuit can be simulated by a program. Such simulators have become increasingly important in the CAD process for circuits, where simulators are used to test integrated circuit behavior. The hardware design language VHDL (a very high-level hardware description language) is an example of a language for creating such programs which simulate hardware. In the other direction, only programs that, intuitively speaking, do not modify themselves can be replaced by circuitry, since at present circuitry is not able to modify itself.

With the coming of low-cost, high-speed ASICs, these units are designed rapidly through CAD procedures. Portions of operating systems and application programs are beginning to be replaced by individual chips or subportions of larger systems on a chip. This process seems to indicate that the flow of innovation in digital computers will move from application and utility programs, able to be easily modified and improved. Such programs, relatively expensive to create, use, and maintain, will be transformed to chips, chip sets, and portions of larger microprocessor systems. These hardware systems will run much faster than the original programs. Primitive examples, mentioned elsewhere, include many cases where microprocessor programs were later reimplemented as peripheral processors in hardware.

Replication of design and fabrication costs is a chief factor contributing to overall systems cost. Such movement from software to hardware will probably prove most effective in applications in parallel to produce hardware parallel processes. The chief immediate application of such replication in this fashion, of course, is in the development of parallel processors consisting of many interconnected microprocessors (or microprocessor-like subsystems) performing the same or different operations at the same time. Such hardware will be described further below.

6. MODERN TECHNOLOGY

6.1 Applications Programs

The structure of applications programs has evolved towards a concept of "user friendly." A definition of this term might initially be that the program does not require from the user any specific knowledge of programming skills or operating system or computer structure. In addition, the program, through its GUI, should guide the user through the computational process by asking questions. For example, it can request responses by simple user actions (motions of the pointer-mouse, single "hot-key" keyboard responses). These replace the elaborate and arcane sequences of keystrokes originally required in many of the first applications programs, tailored to the programming technology of that time. At its most straightforward, the interactive program might offer to the user an elaborate version of "20 questions" in which the program is the questioner, working to discover the true desire of the user.

Such programs may be centered around a single (or more than one) paradigm or conceptual programming structure. The GUI, which likens the terminal screen to a notebook or blackboard, is one such presentation. The spread sheet, a matrix-like representation in which the user enters numbers and/or letter sequences to be processed by standard actions (preprogrammed or under user-control) is another. Standard word-processing programs offer a visual structure not unlike the pages of a manuscript. The more a user can call on previous experience, the easier use of such a system can be.

Many applications programs (like some operating system interfaces) offer menus, either in printed characters or in pictures on a graphical user interface. These constrain and simplify the choices that a user has when interacting. The system may offer the ability to browse through the menus, or additionally, through the text, pictures, or data that are being processed—or the program itself. The system should offer complete information and preventive editing so that incorrect or out-of-range data does not cause the program to "crash" (end up with insufficient information to recover automatically or manually, or to continue).

Applications programs for more complicated problems may require even more complicated interfaces. These should produce graphical results posed at high speed for the user to assimilate and respond to. Computer user and system thus become a pair of joint co-processors working interactively to solve a specific problem or class of problems. This process can of course be extended to many computers and many machines, provided an understanding of the needed interactions can be placed into the hardware and software.

6.2 RISC vs CISC

The advent of CAD for computers themselves, in the form of silicon compilers and layout programs, has produced basic changes. These are changing the output of the silicon chip assembly lines in such fabrication firms as Intel, Motorola, Texas Instruments, AMD, and National Semiconductors in the United States, and Siemens, NEC, Toshiba, and others overseas.

Because of the increased ease of designing and fabricating logical structures, a new overall systems approach including both hardware and software is now being undertaken. Most indicative of this process is the development of new reduced instruction-set computer (RISC) devices. System architects analyze the entire system structure consisting of software (program, compiler, operating system) and hardware (central microprocessor and peripheral chip set). The resulting optimization, originally carried out by researchers at Stanford University, the University of California at Berkeley, and IBM, among others, produced a different design from that of the classical main-frame computer and microprocessor—a class usually called complex instruction-(CISC) (for example, the Intel 80486 shown in Fig. 7). (It is the instructions themselves that are reduced or complex, not the set.)

In such RISC machines, much of the behavior of a complex instruction in a classical CISC machine is generated by the RISC machine's compiler at compile time, rather than by the hardware designer far earlier at the instruction design time itself. The combination of an optimized compilation sequence of simplified instructions is claimed (and often verified in actual hardware designs) to be much faster than a possible shorter sequence of the more complicated complex instructions. The inclusion of pipelined instruction operation and storage caches yields an advantage, in many cases, to simplified RISC instructions as compared to their CISC complex counterparts. Since the cost of memory has decreased with integrated circuit costs, a larger program instruction memory is now available, and thus brings down overall RISC cost.

One major efficiency of RISC designs (which may be possible to introduce into more classical designs such as the Intel 80x86—x = null, 1,2,...)—is the introduction of parallel performance inside individual operations themselves, so that additions and multiplications, e.g., can be performed simultaneously on different pairs of operands. As the cost of memory decreases, compilers can use more storage for instructions, translating and unwinding loops so that the product (memory access/instruction) times (number of instructions) continues to decrease.

Although RISC designs have been used in some main-frame and/or minicomputers, their most publicized appearance has been in microprocessors, which have become more powerful as their speeds encroach upon those of past minicomputers. RISC designs are competing markedly with the more classical Intel and Motorola CISC chips, both for performance and in the market place as parts of so-called work stations. The Intel 80x86 family, e.g., has served as the central processor in the family of IBM PCs and their clones (duplicate copies). Partially because of its speeds, this chip sequence in time has come to constitute an estimated 80% of the microprocessor sales market for personal computers, as the clock frequency, directly related to the overall performance rate, has increased to about 50 MHz (early 1992).

On the other hand, comparisons of performance of a recent RISC-based Hewlett-Packard work station, using the manufacturer's own RISC chip, generate claims of speeds "up to ten times faster" than the most recent, highest-speed Intel CISC microprocessor, the

80486. Machines based on this and similar RISC designs, by HP, MIPS, Sun, and IBM, and even by Intel and Motorola, have produced work stations fundamentally different from the PC architecture of the 80*x*86 family, and much faster. Part of the difference is masked by differing compiler technologies, and by different operating systems, as previously discussed.

Even so, the coming of these new architectures seems to presage further overall system optimizations that may change the software–hardware combination more markedly than before.

The future of mass-produced and -marketed machines (almost everything, in the near future, but the largest computers) will therefore depend on conditions of production and marketing among the large international companies that now increasingly dominate integrated hardware–software systems, as well as the larger number of entrepreneurial smaller organizations that appear to effect distribution.

Just as important, perhaps, will be subsequent software–hardware optimizations that may produce further marked efficiencies, not predictable before the integrated-circuit design and fabrication revolution. These may produce classes of designs as yet untried and perhaps not yet thought of.

6.3 Artificial Intelligence

Beginning in the mid-1950s, researchers began using the term "artificial intelligence" (*q.v.*) to describe potential computers that would behave like intelligent beings—e.g., humans. Competition for research funds has moved emphasis back and forth in this area between mechanical translation, game playing, so-called planning, and later "expert systems." Portions of work in image-processing analysis appeared to fall into this area.

6.3.1 Division of Artificial Intelligence into Problem Solving and Learning. Early efforts seemed to divide the area of study into applications to problem solving on the one hand, and efforts to create learning systems on the other. One of the key methods, now appearing in some commercial and governmental applications, was the attempt to copy a human expert's capabilities into a corresponding computer program running on a machine.

Applications have included copying of such human activities as geologic exploration techniques, testing and tasting of soup, recognition and discovery of terrorists and their activities, medical diagnosis, and a host of other less exotic human behaviors. Special programming procedures, built around a list processing language (LISP) or PROLOG (a logic programming language), bypassed some of the difficulties of the more classical machine-oriented and numerically oriented languages. This was in an attempt to move computer behavior much further into the logical area. So-called knowledge engineers produced "protocols" of medical experts' diagnosis investigations, in order to translate them into computer programs that would produce a similar behavior.

Most of these systems, independent of language, were based on a (condition, action) paradigm, in which the presence of a particular condition in the case at hand (whether medical, geological, or whatever) caused the computer to produce a corresponding information-processing action, whether inside its memory, or output to the observer for possible outside action. The priority of which conditions to give immediate action to, whether to perform them in parallel, etc., in some ways copied the structure of certain computers themselves.

The computer-expert system and the external user were once again a co-processing system, each responding to the other's input by changing state and yielding output. The design of a controlled interactive sequence of (condition, action) pairs thus became a first approximation at generating an intelligent computer program, whose instructions were made up of these individual responses.

Although the knowledge engineer might be considered to be learning from the expert whose protocols he recorded, the need for systems that modify themselves to produce better performance is still there. Such "adaptive systems" have been developed to aid various control systems for aircraft and factories to improve their performance, but these in general have depended upon classical statistical techniques—after all, a type of learning to predict behavior.

A new engineering-oriented theory, the "fuzzy sets" of Zadeh, attempts to introduce enough vagueness into the measurement of the external world that the computer can produce results that might be considered qualitative rather than quantitative. Results from such systems appear to remain experimental.

6.3.2 Neural Nets. The most sophisticated proposed technique for fine-grained intellectual behavior is based on an attempt to copy brain behavior at some detailed level. Both programs and hardware have been designed to copy individual neuron action, with self-adaptation through feedback to the artificial "synapse-like" input elements in individual units. Actual neural net machines, with variable control, have been designed and constructed using special ICs (see NEURAL NETWORKS). Compiler–assemblers are to interconnect substructures both initially, and perhaps later dynamically. Neural net programs so far have been successful in certain classes of simple tasks, but their competitiveness in practice still remains to be seen. They do offer a potential common method of applying learning techniques which otherwise appear markedly different in different problems to be attacked.

The most impressive examples of human-like behavior have perhaps been with chess-playing machines, which in certain cases have been able to defeat very competent human players. However, neither learning nor complex logical expert systems are deeply involved with such machines. Instead, they make use of the brute power of massive integrated circuit techniques to search widely and deeply within the tree of possible moves available to the machine and its opponent. A simple but powerful intellectual heuristic technique is able to limit this search by ignoring certain branches in which it would be of no value to continue the search. For the moment, the expert behavior of human chess masters does not appear to have been captured.

6.4 Creation and Standardization of Languages

A major factor in the development and usage of computers has been the creation of a mass of computer languages (see COMPUTER PROGRAMMING LANGUAGES) to specify the behavior of a specific machine or class of machines.

Two basic techniques of responding to computer languages, which are usually input through a programmer's interface, involve the level of the system which responds to the language:

1. Compilation-translation: This term describes the replacement of external language strings, understandable by the human programmer, by equivalently behaving computer language strings immediately interpretable as programmed instructions by the target computer. Such compilers are one-time performers, which do their translation job, and then retire from the system and leave the computer to interact with its own language translated strings.
2. Interpretation: This term describes a process in which a program (interpreter) stored permanently inside the computer continuously intercepts, interprets, and acts upon the external language strings, dynamically, responding element by element to the external language.

Obviously interpreters are more powerful than compilers, since they include the compiler capability, and are always resident, and thus contain the potential of changing the linguistic program strings with time—adaptively. Correspondingly, compiler-translators, because they look at the external language only once, yield a much faster overall system performance.

The classical example of a language that is compiled is FORTRAN, created by IBM programmers in the early 1960s. This is a formula translator that translates numerical formulas, appropriately modified at the programming input stage, into the languages of various target machines. FORTRAN has proved both effective and long-lasting; it is still being used as a standard language for certain classes of numerical analysis problems even with massively parallel computer systems, in a modified form to include control of the synchrony and asynchrony of individual computers performing concurrently.

An alternative example of an interpretive language is that of BASIC, first created at Dartmouth University at about the same time. BASIC proved particularly useful for interactive computing, which arose at the same time, because the need for recall of a compiler was obviated. It was adopted, because of its relative simplicity in usage (first designed for students), as the original programming language for the IBM PC, embedded in its ROM as a fundamental artifact purchased with the machine. Because of its interpretative slowness, however, BASIC does not compete in large problems, and does not appear as a widespread language of use, except in certain special cases.

Languages evolve as a function of classes of problems being solved, and classes of users or programmers. LISP and PROLOG, oriented towards functional and logical descriptions of problem solution techniques, came out of mathematical studies almost completely independent of computers. On the other hand, the language C, developed at Bell Laboratories, was conceived as a human-oriented programming language nevertheless allowing close interaction with actual machine instructions. Many important non-numerical applications, such as operating systems and utility programs for newer machines, are written in C. A sample is shown in Fig. 10.

An important step forward in computer language structure was the development of command languages (in classical computer command form) that were recursive in the formal mathematical recursive-function behavior that allows functions to call on themselves. From LISP, which had this characteristic, came the algorithmic language (ALGOL) languages, which for the first time had well-studied plans for such recursive behavioral implementation.

A family of computers, first created by Burroughs, and later adapted by microprocessor designers, used a so-called "push-down stack" to store data and addresses automatically on a "last-in, first-out" basis. Without these continually developed hardware capabilities, programs that call on other programs and each other recursively would be much more costly to implement. The extension of this initial structure to more elaborate register "context switching" is purported to produce much of the power of the new RISC architectures.

University users, in attempting to codify the fundamental structure of the more pragmatic commercial languages and simplify the ALGOL complexity, invented other languages such as PASCAL and MODULA, with which many students wrote their first programs.

The U.S. Government, a large user of programs and programmers, demanded standardization of its data processing procedures and produced (with other outside organizations) a common business oriented language (COBOL). The Department of Defense went further to produce an extremely complex language, ADA (named after an English woman regarded as the first computer programmer). This has elaborate certification procedures to guarantee that the results of translation, compilation, and performance would be the same for any machine on which it was to be run. Complex language such as COBOL and ADA combine behavior of both translators and interpreters. They perform certain transformations initially, and then the system interacts dynamically from time to time at run time.

Similar languages, such as VHDL for integrated circuit and chip description, were formally codified and have become part of the design process for microprocessors, ASICs, and other assorted hardware boards and systems.

In fact, one can view this expansion of languages as describing a second level of paradigm, in which the computer is viewed as an instrument to which a programmer communicates, talking about how to solve a restricted class of problems, but in many cases using a very complicated set of nouns (data structures) and verbs (commands).

While higher-level languages have offered tremendous power to the programmer in developing more elaborate data structures and transformations, they have not solved the problem of guaranteeing correctness of the program to produce a specific result. This is the goal of the logic languages (of which PROLOG is a somewhat modified offshoot) which attempt to describe the goals desired and from them produce commands to achieve those goals. Most such languages have achieved only research performance, although Japanese commercial plans indicate use of PROLOG.

The world of computer languages is Darwinian-like in its evolutionary behavior. The power and number of users, as well as investment in previous programs in a language, may guarantee its survival. IBM's imprimatur produced FORTRAN's longevity, but failed to preserve a more elaborate in-house language, PL-1, in the external world. LISP, a powerful and classical artificial intelligence language, failed to produce a much-wanted hardware interpretive commercial machine to run LISP

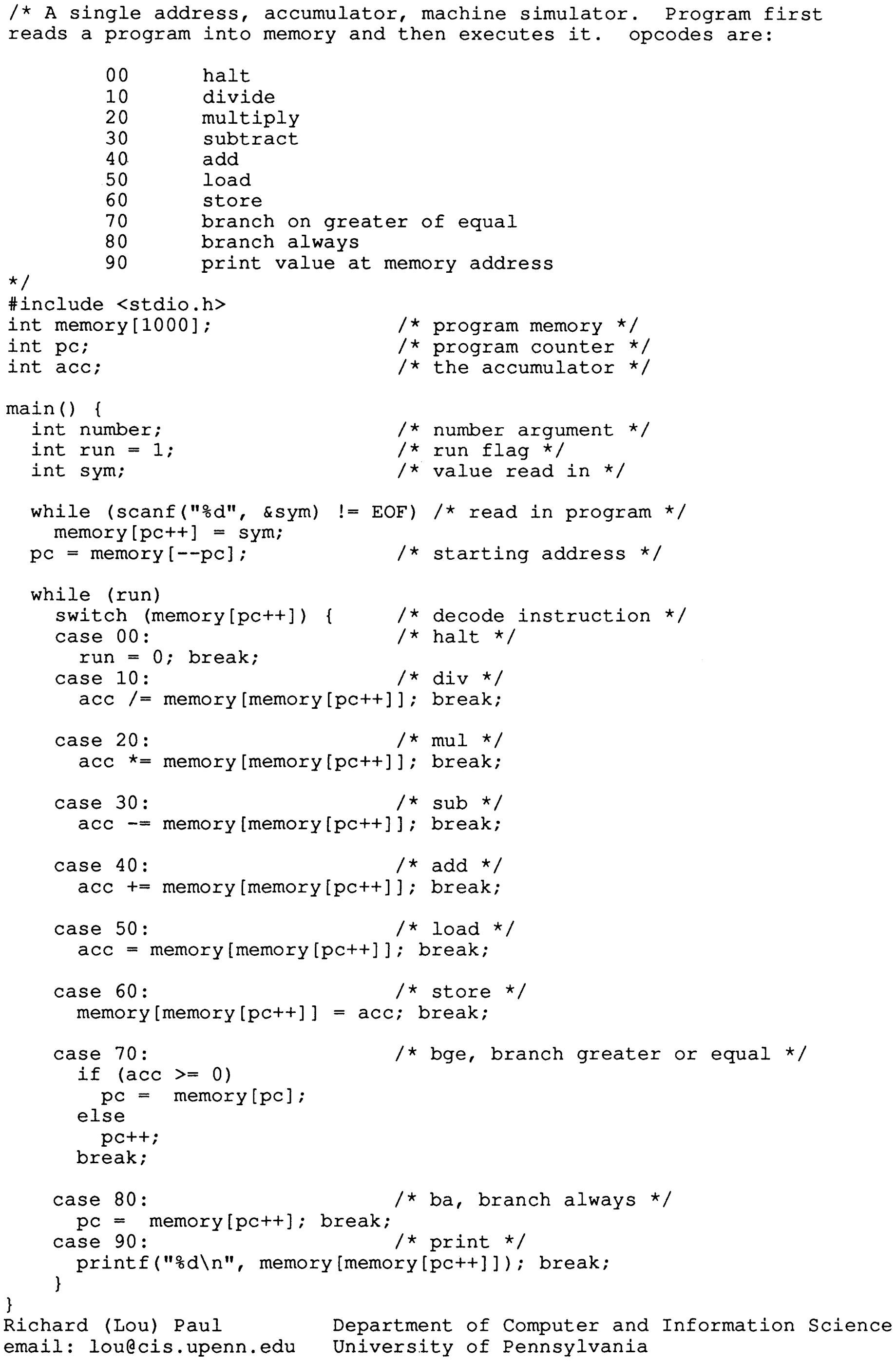

```
/* A single address, accumulator, machine simulator.  Program first
reads a program into memory and then executes it.  opcodes are:

        00      halt
        10      divide
        20      multiply
        30      subtract
        40      add
        50      load
        60      store
        70      branch on greater of equal
        80      branch always
        90      print value at memory address
*/
#include <stdio.h>
int memory[1000];               /* program memory */
int pc;                         /* program counter */
int acc;                        /* the accumulator */

main() {
  int number;                   /* number argument */
  int run = 1;                  /* run flag */
  int sym;                      /* value read in */

  while (scanf("%d", &sym) != EOF) /* read in program */
    memory[pc++] = sym;
  pc = memory[--pc];            /* starting address */

  while (run)
    switch (memory[pc++]) {     /* decode instruction */
    case 00:                    /* halt */
      run = 0; break;
    case 10:                    /* div */
      acc /= memory[memory[pc++]]; break;

    case 20:                    /* mul */
      acc *= memory[memory[pc++]]; break;

    case 30:                    /* sub */
      acc -= memory[memory[pc++]]; break;

    case 40:                    /* add */
      acc += memory[memory[pc++]]; break;

    case 50:                    /* load */
      acc = memory[memory[pc++]]; break;

    case 60:                    /* store */
      memory[memory[pc++]] = acc; break;

    case 70:                    /* bge, branch greater or equal */
      if (acc >= 0)
        pc =  memory[pc];
      else
        pc++;
      break;

    case 80:                    /* ba, branch always */
      pc =  memory[pc++]; break;
    case 90:                    /* print */
      printf("%d\n", memory[memory[pc++]]); break;
    }
}
Richard (Lou) Paul         Department of Computer and Information Science
email: lou@cis.upenn.edu   University of Pennsylvania
```

FIG. 10. A program written in C, one of the most prevalent languages at present in use by programmers developing applications programs for commercial marketing. (Courtesy of Professor Richard Paul, University of Pennsylvania.)

programs directly. (This would have speeded up productivity many times using this language. Nevertheless, users apparently could not support the costs of the equipment involved.) A very powerful language, APL, once supported by IBM and popular in Canada, the land of its creator, has almost disappeared despite its power in certain classes of vector computations.

The disappearance of such once-popular languages, like APL and PL-1, and the rise of the use of C, demonstrate how the constraints of technology, programmer acceptance, compiler complexity, and user demand on the part of government and industry produce survival of the "fittest" languages only in terms of very complex, far from only technical, measures of effectiveness.

The rise of C as a programming language, similar to, but less sophisticated theoretically and technically than, university-oriented PASCAL, shows the technical and market forces at work in language adoption and usage. The language C is closer to machine language than almost any other higher-level language, and contains operators that approximate many machine operations in the standard CISC microprocessors. Its close association with UNIX at Bell Laboratories and University of California, Berkeley, also gave a wide impetus to its usage. Moreover, it is a programmer's language, not forced upon its users by educational curricula.

Specialized languages include those for such user paradigms as the database search process. Designers here created so-called "fourth generation languages," such as SQL, successful in this area, but not extended more widely. The interest of the Japanese in interacting with potential European users of Japanese computers has boosted the potential of PROLOG, originally created and nurtured in France and England. Japanese "fifth generation computers," supported in development by their Ministry of International Trade, may use this logic language based on nonimperative structures similar to declarative sentences. On the other hand, the most powerful new Japanese supercomputers may competitively use FORTRAN, because of the nature of available programs.

A powerful, potentially useful class of languages are those first based on SMALLTALK, another product of the Xerox Park Research Center (which also developed the "mouse" and portions of the IC design systems). SMALLTALK is labeled an object-oriented language, in which the objects are generalized data structures and operations on each class of object are associated with that class. Object classes have a hierarchical hereditary structure in which descendants have the same characteristics as their predecessors, but with potentially more detailed structure and corresponding operations.

Object-oriented languages apparently owe their initial concepts, among others, to mathematical category theory, in which structures are essentially viewed in terms of pointing arrows describing abstract interactions. Despite what might be considered an arcane beginning, pragmatic object-oriented languages such as SMALLTALK and C++, a C derivative extension, are beginning to be used with claims of markedly increased programmer productivity. The abstract view of elements, as opposed to the details of the more classical languages, offers an alternative top-down view.

In the university research community, so-called "typed languages," again corresponding to mathematical structures in abstract algebra, allow powerful but correct manipulation of complex data structures. Such languages have not yet gained a foothold in any commercial or production sense.

Certain small groups of users, such as astronomers, have adapted special languages that prove easy to use and exchange (in this case FORTH, a "growing machine" language, which provides semantic clarity and abbreviational simplicity by sequences of Polish suffix operators, using interpreters).

Throughout this Babel-like set of language developments runs the pragmatic thread of appealing jointly to a body of users on one hand and to a specific computer architecture on the other. Nevertheless, up until very recently, the desire for control of the result by the user population has generally kept resulting programs, in almost any language, highly labor intensive in generation.

Clashes between esthetic formal theory and pragmatic user acceptance generally have found the latter winning out. The coming of the GUI, with its amplification of user performance, has begun to produce acceptance of automatic programming generation, triggered from such graphical interactive interfaces, by nonprogrammer users who wish to produce programs.

The clash here becomes one between the classical programmer, who desired freedom to create at will, on the one hand, and the more casual problem solver, who is willing to accept constraints on what can be communicated in return for more speed and power in producing results. Cheaper computation would appear to increase the potential for more constraints and less creative freedom. Among such constraint systems now are syntax-directed compilers, which only allow correct language strings to be typed. Also there are strong restrictions, in many newer systems, on what commands are allowed, and on use of replication features that manipulate previously written code, just as hardware designers use previously designed subsystems.

In general, this process of automatic generation of computer programs, still unsuccessful, illustrates the paradox that relates the two philosophical extremes: freedom and constraint. Machine languages offer almost absolute freedom to the programmer (with the sometimes crude constraints of the actual hardware). Only very crude syntax is required. As languages become higher level, stricter syntax is required, but semantics becomes more widely defined. The user is constrained as to what program primitives can be used. But these can be used, hopefully error-free, freely, and at very low cost. Just how the Darwinian process of language evolution will proceed in this situation is still to be seen.

6.5 Creation and Standardization of Operating Systems

It has been noted above that the operating system is abstractly the equivalent of the hardware bus inside various computer families, such as that of the PC. Its construction allows addition of subroutines and applications programs into slot-like positions, in a standard process, so that the set of programmed operations can be extended in a growing fashion.

The origin of the operating system, perhaps first developed on the Digital Equipment Corporation's model PDP-8, was the result of applying a classical expert system technique. One observes the sequence of interaction protocols of an operator seated at the console. These sequences, translated into a program, allowed replacement of the expert, in this case the operator, by a computer program that performed the same actions much more quickly and effectively.

The development of the operating system OS-360, for IBM's family of similarly named machines, powerfully demonstrated the potential of the operating system as the integrating factor for a family of computers. Its use continued over many users, and down through tens of years of time. The OS-360 system, however, did not have the interactive power that would allow humans to program and solve problems easily on line.

The development of the operating system Multics by the government-sponsored Project MAC at Bell Labs and MIT first proved the large-scale potential of interactive computing. Honeywell and General Electric used this design to produce the first commercially successful time-sharing operating system. The specialization of Multics to UNIX by workers at Bell Laboratories produced a system originally emphasizing processing of strings using terminals connected to a minicomputer. Over almost 20 years the efforts of the telephone company, AT&T, as well as other firms and universities, such as University of California, Berkeley, produced adaptations that showed success.

UNIX is now expanded to a broader range of equipment, including larger machines as well. Despite its requirements for large amounts of storage, improvement in hardware performance has moved it closer to becoming a possible standard for even personal computers. Since many work stations and larger computers now also use UNIX, this will allow passage of programs and data back and forth electronically between computers from different vendors. Users of one computer can perform portions of their programs on more powerful computers in the same network, with data from elsewhere in a network file system. Such capabilities require interactive network utility programs tailored to fit the cooperating operating systems at both ends.

The development of similar small operating systems on microprocessors, often restricted UNIX-like versions, culminated with Microsoft's disk operating system (DOS) on the IBM PC. This became a de facto standard for those operating systems running on that common de facto standard personal computer, the IBM PC.

The embedding of operating systems in graphical interfaces was first developed by Xerox and commercialized by Apple with the Macintosh. The Macintosh used the paradigm of the desk top, the clipboard, and the file cabinet. The transporting of this technology to the PCs by Microsoft recently produced a powerful, popular GUI, Microsoft's Windows 3.0, as an extension or replacement of that vendor's ubiquitous MS-DOS.

Other technological attempts to improve operating systems include placement of operating systems on ROMs or IC chips. The BIOS on the PC is one such example. Special operating systems have also been developed for real-time systems controlling man-made artifacts, such as aircraft, rockets, automobiles, and factory systems.

At a certain level, the operating system may be viewed as replacing the hardware computer itself as a machine paradigm, with its own operations, data structures, and interfaces. Portions of it may obviously move into the hardware itself. Probably more of the applications programs operations will become codified within the operating system, in a classical process of movement from human information and behavior into automated instructions in the machine.

7. SEQUENTIAL VS PARALLEL PROCESSING

With the coming of low-cost microprocessors, fabricated using VLSI technology, construction of massively parallel-processing systems became possible (see PARALLEL COMPUTING). These consist of large numbers of simpler subprocessors. Originally, parallel systems contained few subprocessors and were very expensive. With new technology, designs have grown in size from tens to thousands of subprocessors, connected in arrays that attempt to pass information and control from element to element in as efficient a fashion as possible.

Such systems require elaborate extensions of previous sequential-programming technology, its algorithms, routines, and languages. An entirely new set of control instructions, many borrowed from parallel hardware design procedures, had to be added to classical languages such as FORTRAN. Alternatively, new languages had to be developed. Investigations into which classical sequential algorithms could be easily and efficiently extended into parallel computation have been necessary. For those problems where this extension technique would not work, entirely new parallel algorithms were needed to make use of the new experimental parallel programming structures.

7.1 Classes of Parallel Equipment

In general, an individual subprocessor in an array is some type of sequential processor. Since each individual subprocessor in an array has both program and data, it is straightforward to classify parallel processing systems according to program (instruction) and data. If all of the basic subprocessors scan the identical program stream (and the same instruction) at once, the system is termed single instruction (SI). Otherwise, if different subprocessor control units scan and interpret different instructions, the term multiple instruction (MI) is used. If all of the data units act upon the same data element, the system is called a single-data (SD) parallel system. Otherwise, with different data being processed in different subprocessors, the label is multiple data (MD).

Thus, to a first approximation, such systems may be classified into four obvious categories: SISD, SIMD, MISD, and MIMD.

SISD machines are classical von Neumann–Eckert–Mauchly digital computers such as the first stored-program computers of the early 1950s. Recently, designers have come up with possible extensions to produce very long instruction word (VLIW) instruction processing. Such processing is now making its way into RISC microprocessor, as well as into single high-performance processor, design. VLIW instructions perform different operations, such as multiplication and addition (included in one instruction), in parallel on local data sets which are made independent of one another through pipelining, vectorizing, or other tactics.

SIMD computers most often operate on vectors of many individual data elements, with the same instruction in parallel. Such architectures very early became important in large numerical calculations able to be expressed in vector-like form. Searches of many databases in parallel, for the same information, also falls into this category. Such problems replace sequential many-input calculations, repeated in time, by the same calculation sequence performed by duplicate processors over space. With these, there are local problems whenever programs produce different conditional transfers for different data inputs. This obviously no longer permits easy use of the same instruction simultaneously in every subprocessor.

7.1.1 MISD and MIMD Systems. MISD systems involve solving many problems with one data stream. A computer with multiple searches on a single data-base stream is an example of such a processor type. Another technique is the use of pipelining, which takes sequential instructions and performs them concurrently on successive iterates in a sequential data stream. The most sophisticated of such systems might well be termed MIMD systems, as described below.

Most important in the future will probably be MIMD systems, which can support multiple instruction streams by pipelining and parallel processing. In the latter case, the interconnection structure by which information, both instructions and data, is passed between processors becomes important, and can be used for further classification. MIMD systems can, e.g., use a bus, or common information channel, to pass information at lower cost but lower overall speed. Most lower-cost recent commercial systems, such as the Sequent and Encore computers, originally used standard microprocessors as the processors connected with common memory, along a bus.

More complicated MIMD systems, such as *n*Cube, and the Hypercube and later versions from Intel, connect each memory-equipped processor to its *n* nearest neighbors, to form an abstract logical hypercube (see Fig. 11). Messages are sent between the node subprocessors in the hypercube to synchronize the concurrent computations.

Hierarchical direct connections include such geometrical forms as trees, pyramids, and prisms. Crossbar machines (in which every processor is connected to every other), are overly expensive, so far, except for small systems. Marked changes in cost of such communication between subprocessors would be necessary before complete crossbar systems became commercially viable.

A most important characteristic of a parallel processing system is its granularity, defined in terms of the size of unit of work allocated to individual processors. Until recently, the most powerful systems were termed "coarse-grained," such as Cray, IBM, and previous ETA systems, with four to ten very high-speed processors, each individually constructed, generally not using VLSI fabrication. Such systems have computing power ranging up to thousands of megaflops. ("Flops" stands for "floating-point operations per sec-

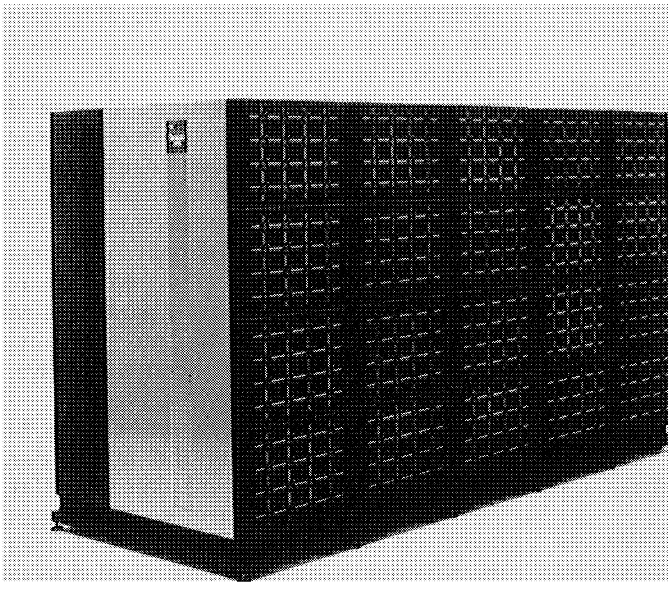

FIG. 11. A multiple-microprocessor parallel machine with potential for more than 1000 individual separate processing units connected to "nearest neighbors" in a cube-like "hypercube" array. (Courtesy of Intel Corporation.)

ond;" mega and giga are the standard prefixes for 10^6 and 10^9.)

The relative cost performance of such coarse-grained processors, compared with alternatives using many-times replicated VLSI microprocessors, appears to be decreasing through multifabrication economies of scale. These "supercomputers," therefore, may tend to be replaced by the massively parallel newer designs.

7.1.2 Cost of New Large Systems. Current 1991 prices for Cray's smallest to largest computers range from $300,000 to $30 million, according to published reports. The largest of these, the Y-MP C-90, is announced as having a central processing unit that can perform at the rate of one gigaflops. For the top price unit, there will be 16 such processors, each program-compatible with previous systems.

New examples of massively parallel designs have been announced recently. An Intel announcement of a new massively parallel processor (Paragon XP/S) with up to 4000 internal processors, gives a range in cost from $2 to $55 million. This system uses standard Intel 860XP RISC-like processors in each node, and is claimed to generate 5 to 300 gigaflops. Such parallel systems with so many internal subprocessors would tend to fall into the medium-grained category, based on the necessary detailed work each subprocessor might do. An Alliant Computer announcement gives price ranges of $2.5 million for a 100-processor system to $16 million for an 800-processor system.

Among other medium-grained commercial systems would be those using tens of Intel 80*x*86 or competitive RISC processors connected along a bus (Sequent, Encore, and other less powerful commercial versions). So-called fine-grained systems were built around simpler and cheaper basic units. The Connection Machine, built in hypercube interconnection with a special chip as subprocessor, is an SIMD machine which replicates the same computation in every one of its thousands of elements.

7.2 Relative Serial and Parallel Efficiencies

Simulation and direct experimentation on existing equipments is being used to test classes of problems according to a criterion of "efficiency": the ratio of actual performance of processors, interacting in an inefficient manner due to delays in waiting for the results of other computations, to that performance possible with perfect undelayed response from all other processors at the same time. [A sequential machine, which does not wait for outside results, has an efficiency of 1, for one processor. An n-element parallel processor with perfect efficiency of 1 would have an efficiency ratio of $n \times (n\text{-machine parallel eff.}) / (\text{sequential eff.})$, with a maximum value of n.]

Certain types of near-linear partial differential equations are solved on key MIMD machines with a very high efficiency, near 1, and efficiency ratio close to n. Nonlinear partial differential equations, on the other hand, which require more elaborate communication (therefore larger communication delays) when solved on parallel processors, have a much lower efficiency ratio.

7.2.1 Methods of Improving Performance. Methods of improving performance of such complex solution procedures in parallel include the attempted use of Monte Carlo (probabilistic) analyses, easily replicated in parallel. Recent conference reports on high-performance computing include examples of this and other new tactics. These develop parallelized solution processes, alternatives to sequential classical procedures that are not easily extended.

In any case, of course, even with low efficiency of usage of parallel architectures, any marked improvement means that solutions to otherwise impossible problems may be obtainable in a finite time. Most of the improvements in productivity in analysis and design of large physical and engineering systems will probably depend on improved usage of such massive architectural computer complexes. Two recent descriptions of experience at Sandia Laboratories and IBM Research make the claim that massively parallel MIMD architectures are becoming more commercially competitive and application effective.

7.2.2 Examples of Problem Solution Implementations. Among the applications whose efficiencies have been studied for MIMD shared and distributed memory architectures is the use of "farm parallelism," with many workers doing the same task, applied to the modeling of gas-phase reactive systems. The problem is broken into several elementary chemical reactions, each investigated separately and concurrently. Many classical events (so-called trajectories) are calculated concurrently. Experimental results on an IBM-3090 multiprocessor (6 subprocessors) and an *n*Cube 2 with 128 nodes were carried out for 4096 trajectories. Efficiencies of over 90% were obtained with several alternative grains of parallelism.

7.3 Areas in Which Parallel Computers Have Succeeded

Other numerical calculations recently described, where parallel computers improved performance, include Navier–Stokes equations (IBM-3090-VF, 6 processors); simulation of laminar–turbulent transition (Cray 2); structure of semiconductor strained superlattices (IBM-3090VF); 3D shallow-water models (Alliant FX/4, 4 processors); and Poisson problem using domain decomposition (IBM-3090VF, Alliant FX/80). Again, future effectiveness of such architectures will depend upon growth in effectiveness of the overall complex of programming languages, compilers, and operating systems, as applied to such parallel performance.

8. SPECIAL-PURPOSE EQUIPMENT STRUCTURES

Because of the increasing ease of design using silicon (or alternative technologies, of which gallium arsenide is the only remote future contender, at the present time), many special-purpose devices for replacement or extension of standard classical machines have been tried; some have been successful. Almost all retained the stored program.

Among past entries not able to be successful commercially were special-purpose database machines. Among those which have proved successful is the graphics image creation machine originally developed at Stanford University, and mapped into silicon by Silicon Graphics. This forms a powerful chip set that has enabled computer graphics presentations on work stations to be markedly sped up at a large decrease in price over previous techniques using the central processing unit.

8.1 Simplified Microprocessor Controllers and Coprocessors

Special simplified controllers have been developed for controlling portions of a microcomputer system. For example, a separate microprocessor, of simplified design, often controls the microcomputer keyboard or the hard-disk drive. Even more simplified special-purpose computers, often without stored programs, control interrupt processing, real-time processing, or analog-to-digital and digital-to-analog conversion outside of the central microprocessor.

The original microprocessors had too few transistors to carry out floating-point arithmetic operations. Therefore, arithmetic coprocessors, on a separate chip, accepted operational commands and data from the central microprocessor. This chip performed the actual multiplications and divisions for scientific and engineering applications programs that required high-speed arithmetic. Successful co-processors of this type were originally off-chip (such as the 80*x*87 floating-point processor for the Intel 80*x*86 series). With growing transistor count on chips, many of these original special-purpose co-processors have since been incorporated into the newer designs which allow for far more functions on the same chip.

8.1.1 Real-Time Applications. Special-purpose real-time microprocessor controllers, such as the Motorola 6802, have become integral parts of automotive systems, acting to control antilock brake systems and to monitor engine behavior. Other classes of specialized digital computers, mass produced in silicon, have found mass applications. Large areas of technology, previously based on analog electronic voltages, are now being manipulated with digital devices. For example, the telephone industry has spawned the digital signal processor, especially made to handle digitized audio signals. High-definition television (HDTV), which has competing analog and digital proposed systems, will probably end up almost completely digital in character, with digital computers as the central controlling and processing elements.

8.1.2 Future Control Devices. Such devices are example of mass-produced, low-cost, stored-program computers, interacting through sensors inputting signals from the outside world. Their output, through effectors, can be used to change and control the outside world. The ability to change their internal behavior, either from outside, or by means of their own program, through tactics of control theory, artificial intelligence, or other heuristic techniques, offers the capability of small, local devices, independently responding, where needed. Their applications to experiment in many areas are still ahead.

8.2 Digital Signal Processors

With its capabilities, an important section of digital computer usage, not ordinarily known to the standard user, is that of the digital signal processor (DSP). These specially designed digital computers are particularly oriented to handle streams of signals in such devices as cellular phones, private branch exchanges, central telephone exchanges, industrial controllers, telephone answering machines, motor controls, multimedia applications, and such internal digital computer mechanisms as hard-disk drives.

The DSPs themselves are now being produced by use of customized design procedures that allow inclusion of new and specialized logical functions on a central computer chip. An example is the Texas Instruments TMS320 series of chips, which allows customized inte gration of analog-to-digital and digital-to-analog converters on the central integrated circuit. Such input–output devices translate from classical audio and other signal levels into digital samples, and conversely.

Such combinations are now being fabricated in low-power CMOS technology, with 32-bit floating point arithmetic units included on the same chip. Unit prices are quoted at "$1/Mflops in 1992," with 27-Mflops units currently being shipped.

The integration of a standard central unit (here the DSP itself, oriented to handling streams of numbers very efficiently), with peripheral elements designed in at the silicon level, produces a chip specific to the application at hand. This is an example of the fundamental logical equivalence of software and hardware and of the continuing movement of functions, originally implemented as programs, into the design of the central silicon chips themselves.

Almost any electronic consumer product, from automotive brake and audio systems to modems and telefacsimile devices, will probably use something like a DSP as its central control element. Most users of standard commercial computers and networks will not directly interface with DSPs. Nevertheless, their existence offers the capability of extension of real-time networks into a broad range of data collection and control procedures, far beyond the standard artifacts most often thought of as computers. As applications arise, other analogous designs, still with stored programs, may extend the range of classical computer systems even further.

9. COMPUTER ECONOMICS IN 1991

9.1 Overall Economics of Computers

Early usage of digital computers entailed high cost of equipment and low productivity due to individual human programming. Financial support was originally economically centered on governmental or large corporation funding.

Moreover, within the United States and Western Europe, construction and distribution of the machines themselves was controlled by means of proprietary systems. These were specially designed so as to prevent easy passage of computer programs from one computer manufacturer's machine to that of another manufacturer. A commercial or scientific user, therefore, was locked into its manufacturer's equipment and programs. Most scientists or engineers involved with large calculations, until recently, at least, may have experienced this situation.

A similar proprietary situation began with the software, although it too generated a conflict: proprietary restrictions vs ease of circulation. On one side, computer programs are easily reproducible using computers themselves (like books with photocopying). On the other, their purchase has become surrounded with complex licensing procedures, copyrights, and other attempts by the vendors to restrict unauthorized reproduction. Just how exactly innovation will be encouraged or discouraged by such processes is still to be worked out, since proprietary families of programs obviously become tied into operating systems, for example.

9.2 Standards Organization

To allow crossover of programs from proprietary hardware or software family to other families, a standardization process began to be developed for hardware, languages, and operating systems. Such standardization mitigates against the restrictions of proprietary hardware, operating systems, and applications programs, all of which aid individual manufacturers in protecting their market share. There is thus a conflict between the forces of commercial hardware and software producers, and the users. The former want to control certain parts of the overall system, to optimize market share and profit. The users want some optimal method of producing the resulting computation at low overall cost in time and money. This clash provides the key continuing dialectic for computers today. Much of the ongoing changes, in particular the effort at standardization, are attempts at resolving this conflict.

9.2.1 International Economic Conflicts. Among the hardware devices that came to be fabricated in quantity, one became an international *cause cèlèbre* rivalling that over imported automobiles. DRAMs became a cheap but powerful method of storing programs and data during processing. Without them, the PC would have been impossible. In the United States, only the largest manufacturers such as IBM and AT&T were able to compete with their own in-house versions of these devices.

Overseas, this part of the computer technology market became economically dominated by Japanese integrated circuit manufacturers, often accused of "dumping" products below cost onto the American market. This alleged oligopoly and its future effect on production of U.S. computers (mirrored by alleged U.S. control of microprocessor production) became a prime "political football" in U.S.–Japanese economic relations. This economic conflict emphasizes the rapid rise of computers to a most important level in the international economy. The DRAM's marketing as a commodity product demonstrates the importance of the standardization process for hardware, here built around the concept of interchangeable electronic parts.

With standardization, the sociological structure of machine usage began to form around standardized technological artifacts: standard hardware, standard languages, standard operating systems, and standard graphical user interfaces. Behind these standards, the hardware and software being distributed began to be commodity products, sold on price and effectiveness, not the proprietary products sold on the historical necessity of the user not losing the investment in past hardware and software. Standardization began when U.S. and European governmental organizations, such as the Department of Defense (DOD), and industrial user or vendor consortia began to negotiate common standards. Among classically important ones were the ALGOL standards, the International Standards Organization agreements on the open systems interface, and the standards on the FORTRAN, COBOL, and ADA languages, the latter two required for U.S. governmental program procurement. Such standards increased the probability that hardware and software from different origins would nevertheless interact with each other successfully.

9.2.2 Hardware Becomes a Commodity Product. As a result of this effort, at the end of the 1980s, the economics of computer manufacture, marketing, and usage began to change markedly. Before, the use of proprietary hardware (the IBM series of IBM-360 and subsequent mainframes) or of proprietary operating systems (such as the VAX-VMS system of Digital Equipment Corporation) locked users into a continuing sequence of same-vendor equipments.

During the last years, much of the computer industry has been forced out of this pattern. A lead story in a September 1991 Wall Street Journal, entitled "Computers Become A Kind of Commodity, To Dismay of Makers," indicates the extent of this change. This report, as well as somewhat more technical reports in trade journals, lists the changes:

1. Customers, except for the buyers of the larger machines, are buying computers as if they were commodities, attempting to buy as cheaply as possible.
2. Profit margins have dropped; sales channels have been transformed away from retailers towards mail order distribution and mass purchasing. The result of decreased margins has forced virtually every computer manufacturer to try to "reinvent itself." Some manufacturers (e.g., Wang)

no longer sell their own computers, only software for another manufacturer. New alliances, such as that between once archrivals Apple and IBM, are springing up.

3. With the loss of proprietary computer system monopolies and the coming of "open systems," the sale of "clones" (or near-exact duplicates), such as for the IBM PC, brings new competition. New quasimonopolies, for the moment, have arisen in software (such as the Microsoft DOS PC operating system and its Windows derivative). In hardware, Intel, the maker of the system 8086 microprocessors and derivatives, may also be momentarily considered a quasimonopoly, despite the beginning fabrication of duplicates by competitors.
4. The coming of UNIX as a standard operating system offers a new level of open system and commodity product structure. UNIX has now become the standard operating system of computers at the work station level above the IBM PC and its clones. For much of the past, implementation of full-scale UNIX on PC-level machines was impossible because of RAM and hard-disk storage limitations, but recent pricing decreases offer storage capabilities that make UNIX competitive. However, the complexities of the system, without appropriate graphical interfaces such as the open-system X-Windows, do not make it a system for the casual user.

The result predicted is that the number of vertically integrated hardware vendors will decrease markedly, replaced by a smaller number of fabricators in the U.S. and overseas. Consulting and integration companies will become more important. Software companies, with proprietary programs, may replicate the past domination by proprietary hardware systems—until their products, too, become commodity products.

All this, of course, does not take into consideration very much what foreign (most likely Asian) competitors in both areas may do. It also assumes that computers will merge "gracefully" into all areas of modern industrial society, which is far from likely. Nevertheless, such critiques seem to indicate that, whereas the technological future of computers might be considered predictable, their place in the economy may shift drastically due to such technological changes themselves.

New paradigms for computer users that will be designed to hide the actual software–hardware structure will continue to be invented. The beginning capability for computer statistical, neural net, or other "learning" will begin to grow if small practical commercial attempts prove successful. The size of computers, decreasing from main frame to desktop to laptop to notebook size and below, makes new applications suddenly feasible. The interactive one-person, portable computer, connected by cellular phone networks to e-mail and giant database servers, appears almost within the present technological-economic range.

There thus appears to be a dichotomy in future usage: On the one hand, smaller and smaller physical computers will become more tailored to individual users, with these individual software–hardware systems adapting themselves to the needs and idiosyncracies of the individual human with whom they are in synergy. On the other hand, such devices (and others) will also serve as end organs of input–output to giant client-server network systems. These will provide knowledge, computational, and retrieval services for varied forms of information, with many alternative methods of presentation. Interaction with other humans on this network will be rapid and powerful. Problems of misuse of such systems are among the most important not yet posed.

9.3 Fundamental Changes in Scientific Economics

One major effect of the computer revolution on scientists and engineers has been the disappearance of many "wet laboratories." These classical laboratories, where experiments were carried out on the actual physical, chemical, or mechanical devices, have in many cases been replaced by digital computer computations to predict comparable behavior. Wherever the theory is trusted, stored-program computers, with programs to solve problems based on that theory, should often be able to produce predictable results far less expensively.

On the frontiers of a new field, experiments obviously are necessary before any theory can be created or verified. Nevertheless, many highly sophisticated experimental laboratories, such as wind tunnels, are no longer being duplicated in quantity. Predictive behavior of supersonic planes, in most cases, are the results of elaborate computer calculations. The software component of such classical organizations as Bell Laboratories has become an increasingly important financial factor. Originally, the experimental laboratories were often most expensive; they are still present, but economically may no longer dominate. After that the hardware costs of the computer equipment, so often first supported by the government in national laboratories, became a marked burden. Finally, in time sequence, the software development for such computer hardware has become increasingly expensive. The process of prediction and verification of experiment is thus intricately entwined, both technologically and economically, with use, design, and development of both hardware and software. In certain cases, the computer has become the experiment.

GLOSSARY

Adaptive System: A system which changes its behavior as a function of its input.

Analog: Method of representing information by continuously varying physical quantities.

Arithmetic–Logic Unit (ALU): A portion of a computer which performs arithmetic calculations and also makes logical decisions based on the arithmetic results.

Application-Specific Integrated Circuit (ASIC): Usually an IC designed especially for some specific problem solution, as opposed to general-purpose microprocessor ICs.

Basic Input–Output System (BIOS): That part of the operating system concerned with handling the details of input and output procedures; usually stored in ROM.

Bus: A (usually parallel) set of wires connecting many units together in a common pathway for data or instruction passage.

Cache: A small, high-speed storage unit between RAM and the CPU or microprocessor's register storage, that requires far less time to access, and thus allows the computer to run much faster.

Computer-Aided Software Engineering (CASE): An area of study and development of ways to produce effective programs semiautomatically or automatically.

Central Processing Unit (CPU): That part of a computer which interprets the program and carries out instructions.

Chip: A technical term for the usually rectangular piece of silicon substrate onto which integrated circuits are deposited and fabricated as a microelectronic device.

Complex Instruction-Set Computer (CISC): One in which individual operations carry out elaborate data processing. Such designs are alleged by RISC protagonists to be much less efficient, slower, and more costly.

Clone: A "duplicate copy," usually marketed, designed so as to duplicate the performance of the original design. For computers, most recent clones are copies of the IBM PC.

Coarse-Grained: A parallel processing system with each subprocessor working on large subsets of a programming procedure at the same time.

Compiler-Translator: A service program that translates from a higher-level language to the basic machine code language.

Computer Operator: A technician who ordinarily controls the sequence of operations of a computer, originally manually.

Converter: Electronic circuit to transform analog signal representation to digital, and conversely.

Digital: Method of representation of information by variables that can take on only a limited set of discrete values.

Digital Signal Processor: A special type of microprocessor designed specifically to manipulate streams of digital representations, usually of external analog signals.

Dynamic Random-Access Memory (DRAM): Storage device which must continually be refreshed to retain information.

Expert Systems: Programs which attempt to copy the effective behavior of human experts.

Fine-Grained Processing System: A parallel processing system with each subprocessor working on a very small portion of a programming procedure at the same time.

Finite-State Machine: A formal representation of mechanistic behavior describing the overall flow of performance in terms of inputs, outputs, internal states, and functions relating their local behavior.

Flag: Single binary digit indicating the existence of a particular situation—usually the recent occurrence of a particular event, such as an addition operation with overflow.

Floating Point: Numerical encoding in terms of powers of 2 or 10 with fractional coefficients, to allow computation without the overflow obtainable with fixed-length fixed-point numerical representation.

Fuzzy Set: Nontechnically, an imprecisely defined set which allows imprecise but hopefully useful reasoning to be carried out on its behavior.

Graphical User Interface (GUI): A pictorial presentation, usually on a cathode-ray tube, of the user's range of choices for the next step, allowing much easier programming and operation of a computer.

Handshake: A two-way passage of signals between two concurrent portions of an interacting system, that allows synchronization of their behavior.

Hardware: Generic term for any or all computer electronic equipment.

Hot Key: A key on the computer or terminal keyboard that is programmed to cause a sequence of operations to occur that would otherwise require a more elaborate sequence of key strokes.

Icon: A pictorial representation of a file or program, on a GUI.

Interpreter: A continuously resident service program that translates higher-level language on line at run time, performing actions based on the semantics of the linguistic elements.

Interrupt: An external signal from a peripheral to the CPU or microprocessor stating that it needs some type of service.

Local Area Network (LAN): An interconnection of computers usually within a contiguous geographical area.

Machine Code: The simplified code of the actual machine hardware, which translates at run time into actual electronic switching actions.

Macintosh: A very successful personal computer with a file and clipboard paradigm for its CRT presentation, using a mouse.

Main Frame: Classical term for a physically very large expensive CPU with all its peripherals.

Menu: A listing, as in a restaurant, for options—programs, data, etc.—that a programmer may choose to have the computer perform or operate on.

Microprocessor: A CPU fabricated, usually from silicon, with very small transistors and a very small chip substrate.

Multiple Instruction, Multiple Data (MIMD): Description of parallel processing systems that allow many individual processors to work simultaneously, in one program, on different data with different instructions.

Motherboard: A large printed circuit board, constituting the basic physical structure onto which a computer's electronic elements would be mounted.

Mouse: Small hand-held device used to move a cursor around the screen.

MS-DOS: A small operating system originally applied to the IBM PC as PC-DOS; often called simply DOS. (From Microsoft Disk Operating Systems.)

Neural Net: A detailed interconnection of processors with the aim of producing behavior comparable to that of human brain neurons or neuron groups.

Object Oriented Programming (OOP): The use of object-oriented languages to manipulate classes of data structures with common properties, including the partial ordering of inheritance.

Operating System: Control program that performs housekeeping for applications programs, including file storage and retrieval and input–output, also offering simplified language for control.

Parallel Processor: A digital computer made up of more than one subprocessor, which performs more than one instruction at the same time.

PC: Personal computer—a computer, first available about 1978, which usually interacted with only one person.

Pipeline: A processing structure taking sequential information, such as numerical elements in a vector, and processing more than one at a time, to speed up arithmetic or data manipulation.

Polling: Cyclic consultation of peripherals by a CPU or microprocessor.

Protocol: In nondiplomatic usage, the detailed recording of experiments, human interactions, or computer behavior, for example.

Pushdown Stack: A hardware or software structure allowing "last-in, first-out" behavior of storage into it, usually broadened to include even more elaborate storage procedures.

Random-Access Memory (RAM): An electronic read-and-write memory in which every stored element can be consulted equally easily.

Read-Only Memory (ROM): An electronically unchangeable memory that can be consulted at very high speed.

Reduced Instruction-Set Computer (RISC): A computer based on a set of simplified instructions; claimed because of its instructions' simplicity to perform far more effectively than previous designs.

Run Time: That interval when the computer is actually performing a program, as opposed to compile time, when the compiler is operating, or other preliminary intervals not associated with actual performance of the solution program.

Sequential Processor: A digital computer which performs operations only in sequence, not several at the same time.

Silicon Compiler: A compiler–translator whose target language is a linguistic description of an actual chip-fabrication plot.

Software: Generic term for any and all types of programs.

Spread Sheet: A computer program that presents a diagram ruled horizontally and vertically into which numbers, equations, and other information may be entered, to be operated upon by very general operations (such as matrix-vector multiplication, for example). The term "spread sheet" comes from the original paper equivalent.

Static Random-Access Memory (SRAM): Random-access memory that retains its contents without the need of repeated refreshing.

Stored Program: Class of computer where the problem solution procedure is stored in a changeable memory.

Time Sharing: A computer usage method by which many users at terminals share the performance of a powerful central processor.

UNIX: An operating system originally developed by Bell Laboratories about 1970.

User-Friendly: An adjective describing hardware or software which is easy for a nontechnically oriented user to interact with.

Utility Program: A program to aid in understanding and manipulating the contents of storage and behavior of the machine, often after nonpredicted behavior.

Very High-Level Description Language (VHDL): A language for describing computer architecture along with more detailed circuitry description down to integrated circuits.

Video-Tex: An interactive computer-driven public system for interrogating data bases such as telephone books, airline schedules, etc.

Very Long Instruction Word (VLIW): Describing a type of processor that allows concurrent behavior in a single instruction, and therefore speeds up each instruction.

Windows: Rectangular CRT areas in which different data or instruction sets can be presented, so that the user can deal with more than one group of information on the screen at the same time.

ACRONYMS

ALU: Arithmetic-Logical Unit
ASCII: American Standard Code for Information Exchange
ASIC: Application-Specific Integrated Circuit
BIOS: Basic Input–Output System
CAD: Computer-Aided Design
CASE: Computer-Aided Software Engineering
CISC: Complex Instruction-Set Computer
CPU: Central Processing Unit
CRT: Cathode-Ray Tube
DMA: Direct Memory Access
DOS: Disk Operating System
DRAM: Dynamic Random-Access Memory
DSP: Digital Signal Processor
GUI: Graphical User Interface
LAN: Local Area Network
MIMD: Multiple Instruction, Multiple Data
MISD: Multiple Instruction, Single Data
PC: Personal Computer
RAM: Random-access memory
RISC: Reduced Instruction-Set Computer
ROM: Read-Only Memory
SIMD: Single Instruction, Multiple Data
SISD: Single Instruction, Single Data
SQL: Standard Query Language
SRAM: Static Random-Access memory
VHDL: Very High-Level Development Language
VLIW: Very Long Instruction Word

Further Reading

Bach, M. J. (1986), *The Design of the UNIX Operating System*, Englewood Cliffs: Prentice Hall.

Barkakati, N. (1991), *X Window System Programming*, Carmel, IN: SAMS.

Birnes, W. J., ed. (1989), *McGraw-Hill Personal Computer Programming Encyclopedia: Languages and Operating Systems*, New York: McGraw-Hill.

Brinch-Hansen, P. (1977), *The Architecture of Concurrent Programs*, Englewood Cliffs: Prentice Hall.

Camposano, R., Wolf, W. (1991), *High-Level VLSI Synthesis*, New York: Kluwer Academic.

Curry, D. (1991), *Using C on the UNIX System (A Nutshell Handbook)*, Sebastopol, CA: O'Reilly and Associates.

DeCegama, A. L. (1989), *The Technology of Parallel Processing*, Vol. 1, *Parallel Processing Architectures and VLSI Hardware*, Englewood Cliffs: Prentice Hall.

Denker, J. ed. (1986), *Neural Networks for Computing*, New York: American Institute of Physics.

Durand, M., El Dabaghi, F., eds. (1991), *High Performance Computing II, Proceedings of the Second Symposium on High Performance Computing, Montpellier 1991*, Amsterdam: North Holland.

Goldberg, A. (1984), *SMALLTALK-80, The Interactive Programming Environment*, Reading MA: Addison-Wesley.

Huber, J. P., Rosneck, M. W. (1991), *Successful ASIC Design the First Time Through*, New York: Van Nostrand Reinhold.

Hwang, K., Briggs, F. A. (1984), *Computer Architecture and Parallel Processing*, New York: McGraw-Hill.

Iannucci, R. A. (1990), *Parallel Machines: Parallel Machine Languages*, Boston: Kluwer Academic.

Katevenis, M. G. H. (1985), *Reduced Instruction Set Computer Architecture for VLSI*, Cambridge, MA: MIT Press.

Kochan, S. G., Wood, P. H. (1985), *UNIX Shell Programming*, Indianapolis: Hayden.

Lipsett, R., Schaefer, C., Ussery, C. (1989), *VHDL: Hardware Description and Design*, Boston: Kluwer Academic.

Mead, C., Conway, L. (1980), *Introduction to VLSI Systems*, Reading, MA: Addison-Wesley.

Patterson, D. A., Hennessy, J. L. (1990), *Computer Architecture: A Quantitative Approach*, San Mateo, CA: Morgan Kaufman.

Peek, J. (1991), *MH & xmh*, Sebastopol, CA: O'Reilly and Associates.

Reiss, L. (1987), *Introduction to Local Area Networks with Microcomputer Experiments*, Englewood Cliffs: Prentice Hall.

Rubin, S. M. (1987), *Computer Aids for VLSI Design*, Reading MA: Addison-Wesley.

Stevens, (1989), *Unix Networking Programming*, Englewood Cliffs: Prentice Hall.

Tabak, D. (1990), *RISC Systems*, New York: Wiley.

Veronis, A. (1991), *Survey of Advanced Microprocessors*, New York: Van Nostrand Reinhold.

Wechsler, H., ed. (1990), *Neural Networks for Perceptrons*, Vol. 2. New York: Academic.

Wiener, R., Pinson, L. J. (1988), *An Introduction to Object-Oriented Programming and C++*, Reading, MA: Addison-Wesley.

CONFORMATIONAL ANALYSIS

S. SCOTT ZIMMERMAN, *Department of Chemistry, Brigham Young University, Provo, Utah, U.S.A.*

Introduction 229
1. **Overview of Conformational Analysis** 230
1.1 The Goals of Conformational Analysis 230
1.2 Dihedral Angles 230
1.2.1 The Dihedral Angle in Ethane and Butane 231
1.2.2 The Dihedral Angles in Peptides 232
2. **Experimental Techniques** 234
2.1 X-Ray Diffraction 234
2.2 Nuclear Magnetic Resonance 235
2.3 Ultraviolet Absorption Spectroscopy 236
2.4 Infrared Spectroscopy 237
2.5 Circular Dichroism and Optical Rotatory Dispersion 237
2.6 Other Experimental Techniques 238
3. **Theoretical Techniques** 238
3.1 Empirical Methods 239
3.2 Quantum-Mechanical Approaches 239
3.3 Applying Theoretical Techniques 240
Glossary 240
Works Cited 241
Further Reading 241

INTRODUCTION

Conformational analysis is the study of the three-dimensional structure of molecules. It specifically deals with spatial arrangements that result from rotations about single bonds between the atoms in molecules (Carey, 1987). Conformational analysis is a major field of physics, chemistry, and biology because the conformation of a molecule has a significant effect on its physical and chemical properties and on its biological activity (Orville-Thomas, 1974).

The following are the objectives of this article:

Provide an overview of conformational analysis, including a complete definition of conformational analysis, a discussion of its importance, and a description of how scientists determine the conformation of a molecule.

Describe some of the major experimental approaches to conformational analysis, including x-ray diffraction, nuclear magnetic resonance (NMR), ultraviolet spectroscopy (UV), infrared spectroscopy (IR), circular dichroism (CD), and optical rotatory dispersion (ORD). This article will not describe all the physical techniques employed for conformational analysis. For a more complete description of experimental approaches, consult the list of Further Reading at the end of this article.

Describe the major theoretical (computational) approaches to conformational analysis, including *ab initio*, semiempirical, and empirical methods, which are part of the more general procedures of conformational analysis called molecular modeling, molecular mechanics, and molecular dynamics.

The example systems described in this article are restricted to organic (carbon-containing) compounds: simple hydrocarbons, peptides, carbon-based polymers, and biological polymers, with emphasis on biological applications. This article does not treat inorganic compounds, which often do not have rotatable single bonds and therefore occur in only one conformation. The reader should understand, however, that conformational analysis of inorganic compounds is also an important field of chemistry and physics. Fortunately, the principles and techniques discussed in this article also apply to inorganic compounds.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. OVERVIEW OF CONFORMATIONAL ANALYSIS

Organic and biological molecules are made up of atoms joined together with covalent bonds. Most carbon-carbon, carbon-nitrogen, and carbon-oxygen single bonds can readily rotate at room temperature, whereas double and triple bonds have severely restricted rotations. A rotation of more than about 15° in a double or triple bond is extremely rare. Therefore, conformational analysis deals primarily with the different shapes of a molecule as a result of rotations about single bonds or about partially double bonds.

1.1 The Goals of Conformational Analysis

The major goals of conformational analysis can be expressed in terms of the following questions (see Cantor and Schimmel, 1980a; Zimmerman, 1985):

What is the three-dimensional shape of the molecule? The immediate aim of experimental and theoretical conformational analysis is to know the relative locations in space of all the atoms in the molecule being studied.

How does the shape of the molecule change in response to environmental changes? For example, what are the effects of temperature, pressure, and solvent? How does the conformation in the solid state differ from the conformation of the molecule in solution? What are the effects of small ions and other solutes on the conformation of the molecule in solution?

What is the relationship between the conformation of the molecule and its physical, chemical, and biological properties? This question addresses the important structure-function relationship in organic and biological molecules. The question is also important in materials science and polymer science, since the development of new materials depends on an understanding of the conformational properties of the molecules involved.

Can theoretical methods be used to predict the conformation of a molecule? One of the major goals of the modern field of molecular modeling, for example, is to use computational methods to predict the conformational properties of organic and biological molecules.

Given the conformational characteristics of a molecule with known biological activity, can other molecules be designed and synthesized having similar physical and biological properties? This question is the basis for drug design in modern pharmaceutical companies. If a certain molecule, for example, has desirable properties but also has some undesirable side effects, a related molecule can possibly be designed to enhance the desirable effects and diminish the undesirable effects.

1.2 Dihedral Angles

The conformation of a molecule is described completely by the *X*, *Y*, and *Z* Cartesian coordinates of all the atoms in the molecule. Since the conformation is independent of the location of the molecule in space, the number of independent variables is $3N-3$, where N is the number of atoms.

Often the molecule is treated as having fixed bond lengths and bond angles, and therefore the conformation can be described as a set of angles of rotation (called *dihedral angles* or *torsional angles*) about the single bonds. Because of bond stretching and bending, the bond lengths and angles between atoms are in fact not fixed but fluctuate about equilibrium locations. The concept and use of dihedral angles is still valid because the equilibrium locations are essentially constant.

Any dihedral angle θ is defined by the position of four atoms: two central atoms (see atoms 2 and 3 in Fig. 1) connected by the covalent bond whose torsional angle is being described, and two flanking atoms (see atoms 1 and 4 in Fig. 1), each attached to one of the

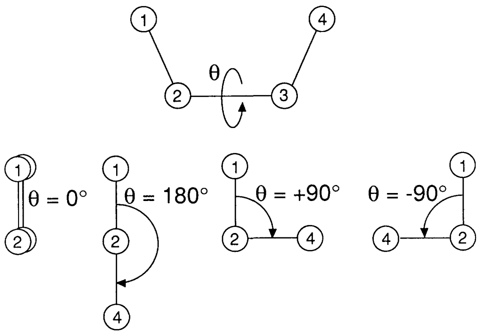

FIG. 1. Definition of the dihedral (torsional) angles θ. When the 1–2 and 3–4 bonds are eclipsed, $\theta=0°$; when those bonds are *anti* or *trans*, $\theta=180°$. Clockwise rotation of the 3–4 bond relative to the 1–2 bond gives positive values of θ; counterclockwise rotation gives negative values of θ.

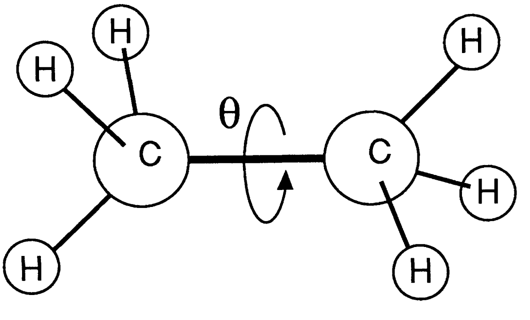

FIG. 2. Ball-and-stick model of ethane.

central atoms. If the four atoms are in the same plane with the two flanking atoms on the *same* side of the bond under consideration, then (by definition) the dihedral angle θ is zero (that is, $\theta=0°$). When atoms 1, 2, and 3 are kept stationary and atom 4 is moved by a clockwise rotation about the bond between atoms 2 and 3, the value of θ increases from 0° to 180°. When the bond is rotated counterclockwise, the value of θ decreases from 0° to $-180°$. When all four atoms are in the same plane but on *opposite* sides of the bond under consideration, the dihedral angle θ equals $\pm 180°$ (IUPAC-IUB, 1970). See Fig. 1.

1.2.1 The Dihedral Angle in Ethane and Butane. One of simplest organic compounds that can exist in different conformations is ethane, CH_3–CH_3. Figure 2 shows a ball-and-stick representation of the structure of ethane. Since the three hydrogen atoms on each carbon are equivalent, the four atoms used to specify the dihedral angle θ (for rotations about the C–C bond) include the two carbons and any one of the three hydrogens on each of the two carbons. The shape of the molecule with a specific θ value can be represented in several ways. The three most common are wedge-and-dash drawings, sawhorse drawings, and Newman projections. Figure 3 shows ethane in all three of these representations with $\theta=60°$, called the *staggered* conformation. Figure 4 shows ethane in the three representations with $\theta=0°$, called the *eclipsed* conformation.

Although all values of θ from $-180°$ to $+180°$ for rotations about the C–C bond in

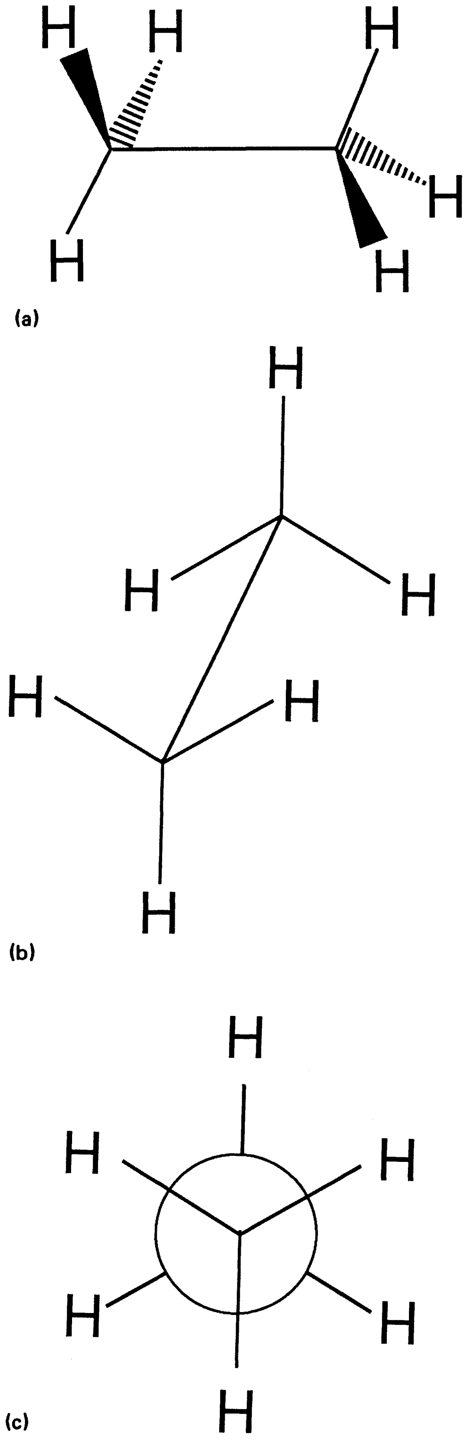

FIG. 3. Models of ethane with $\theta=60°$. (a) Wedge-and-dash model. (b) Sawhorse model. (c) Newman projection.

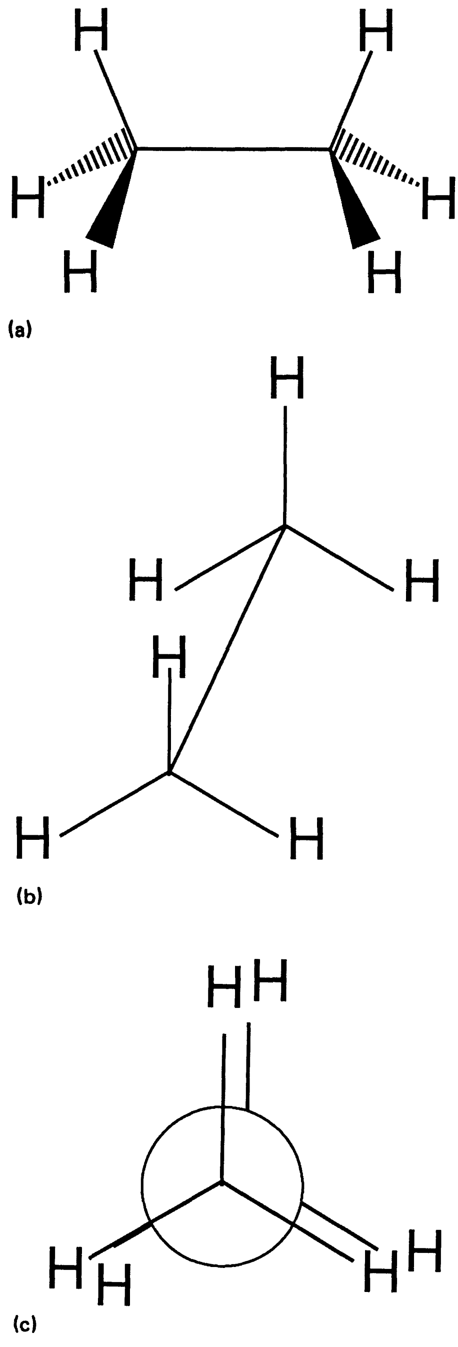

FIG. 4. Models of ethane with $\theta=0°$. (a) Wedge-and-dash model. (b) Sawhorse model. (c) Newman projection.

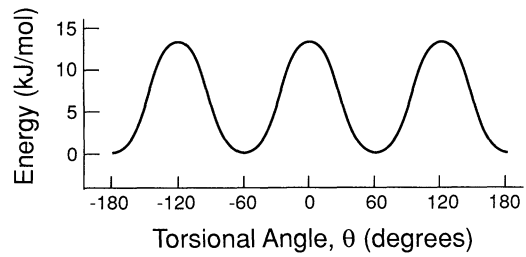

FIG. 5. The potential (conformational) energy of ethane as a function of the torsional angle θ.

ethane are possible, all values of θ are not equally probable, and therefore rotation about the C–C bond is not completely free. The diagram in Fig. 5, derived from experimental and theoretical conformational analysis, shows that the potential energy is a minimum at $\theta=+60°$, $-60°$, and 180°. These three values of θ are, in fact, the same conformation, which is the most stable (lowest-energy) form of ethane. The potential energy is a maximum (13 kJ) at $\theta=0°$ and $\pm 120°$. These values of θ also yield the same conformation, which is the least stable form of ethane.

The conformational properties of organic compounds are further illustrated by examining the rotations about the central C–C bond of butane, CH_3–CH_2–CH_2–CH_3. This molecule can exist in two different stable conformations, one called *gauche* (with $\theta=\pm 60°$) and one called *anti* (with $\theta=180°$). Although both the gauche and anti conformations are stable—that is, they both occur in minimum-energy conformations—anti is the *global* (overall) minimum, having a potential energy of 3.3 kJ lower than the gauche. The potential-energy diagram of butane is shown in Fig. 6.

1.2.2 The Dihedral Angles in Peptides.

Conformational analysis of peptides and proteins is a major field of biophysical chemistry

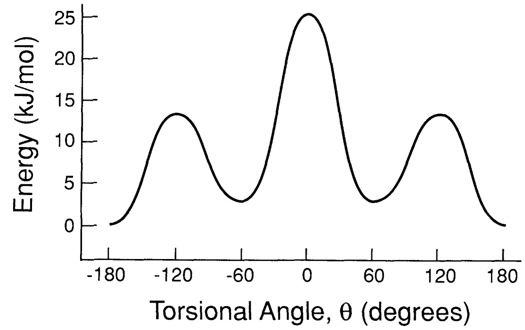

FIG. 6. The potential (conformational) energy of the central C–C bond of butane as a function of the torsional angle θ.

because the shape of a protein is the major factor in determining its biological function.

Peptides and proteins are composed of amino acids joined together with peptide bonds. Naturally occurring amino acids have the general structure shown in Fig. 7(a). Two amino acids joined together by a peptide bond are generally called dipeptides [Fig. 7(b)]; three amino acids joined together with two peptide bonds are generally called tripeptides; and so forth. Long, unbranched chains of amino acids joined together with peptide bonds are called polypeptides. If the polypeptide is a naturally occurring biological molecule, it is called a protein.

The dihedral angles that dictate the conformation of peptides and proteins are ϕ, ψ, ω, and χ (see Fig. 8). Angle ϕ is the dihedral angle described by the four atoms C′–N–C$^\alpha$–C′ that is, rotation about the N–C$^\alpha$ bond. (C$^\alpha$ is the α carbon, adjacent to the carboxyl carbon in an amino acid. C′ is the carbonyl carbon in a peptide unit.) Angle ψ is described by the atoms N–C$^\alpha$–C′–N, that is, rotation about the C$^\alpha$–C′ bond. Angle ω is described by the atoms C$^\alpha$–C′–N–C$^\alpha$, that is, rotation about the C′–N or peptide bond. The peptide bond has partial double-bond character because of the overlap of the unpaired electrons on the nitrogen with the π-electrons of the C═O double bond. Since the peptide bond has partial double-bond character, ω is usually found to have the value 180° (planar trans) or, on rarer occasions, the value 0° (planar cis). Angle χ is a rotation about a side-chain bond (see Fig. 8).

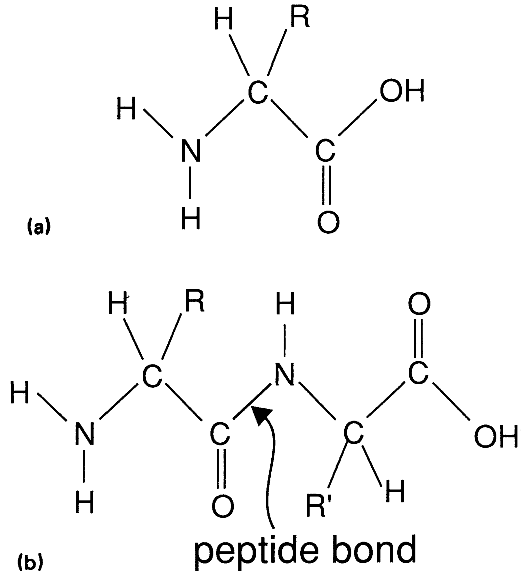

FIG. 7. (a) The structure of an amino acid (shown here in its un-ionized form). (b) The structure of a "dipeptide." R represents an organic substituent that defines a specific amino acid. For example, R is –CH$_3$ in the amino acid alanine.

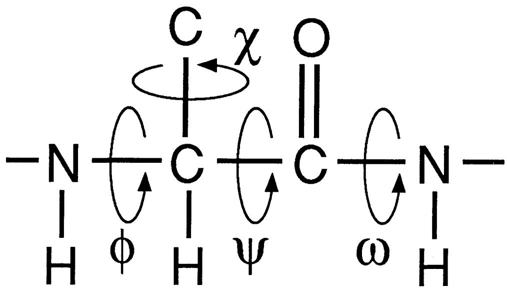

FIG. 8. The Greek letter designations of the angles of rotation in peptides and proteins.

A theoretical analysis of the molecule N-acetyl-N′-methylalanineamide (Fig. 9), a model of an amino acid unit within a polypeptide or protein, yields the contour plot (Fig. 10) of the conformational potential energy as a function of the dihedral angles ϕ and ψ, with ω and χ being fixed at 180° and 60°, respectively (see Zimmerman *et al.*, 1977). (A conformational-energy contour plot is a three-dimensional potential-energy curve analogous to the two-dimensional curves shown in Figs. 5 and 6.) The diagram shows lines of constant energy at increments of 1 kcal mol^{-1} (4.186 kJ mol^{-1}) above the global minimum at about $\phi = -80°$, $\psi = 80°$ (marked with a +). For simplicity, lines for conformational energies greater than 9 kcal mol^{-1} (40 kJ mol^{-1}) above the global minimum have been omitted from the plot. This large blank space indicates disallowed conformations in which the ϕ-ψ values cause

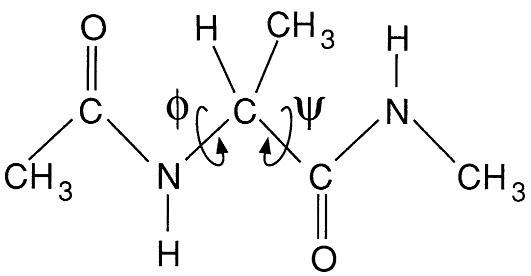

FIG. 9. The structure of N-acetyl-N′-methylalanineamide, a model of a single alanine residue in a peptide or protein.

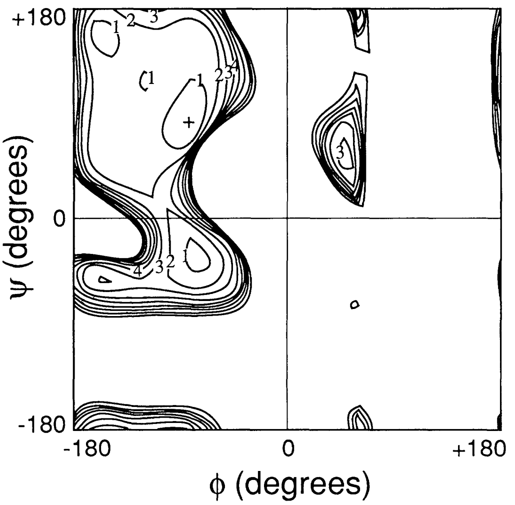

FIG. 10. The conformational energy contour map of the alanine residue (*N*-acetyl-*N'*-methylalanineamide). The + marks the global minimum. Lines of constant energy are drawn for 1–9 kcal mol^{-1} above the global minimum.

overlaps between one or more pairs of atoms in the peptide molecule.

Figure 11 shows the conformation of the global minima of N-acetyl-N'-methylalanineamide. This conformation ($\phi = -80°$, $\psi = 80°$) is called C_7^{ax} because the hydrogen bond forms a seven-member ring (or cyclic) structure and the alanine side chain is approximately in the same plane as (that is, axial to) the seven-membered ring.

The above discussion of butane, ethane, and peptides demonstrates the nature of conformational analysis and lays the foundation for the discussions of experimental and theoretical techniques in conformational analysis.

2. EXPERIMENTAL TECHNIQUES

Chemists and physicists have developed a wide range of experimental techniques for analyzing the conformation of organic and biological molecules. Each technique has its own special set of advantages and disadvantages, depending on the type and amount of information that the technique provides, the purity and size of sample required to perform the experiment, the ease of varying environmental conditions, the effort required to prepare samples and obtain results, and so forth (Cantor and Schimmel, 1980b; see MOLECULAR SPECTROSCOPY). In many cases, two or more techniques are required to yield the desired conformational data.

The following is a brief description of some of the most popular experimental techniques in conformational analysis.

2.1 X-Ray Diffraction

X-ray diffraction is a powerful technique for determining the precise location of the atoms in a molecule (see X-RAY DIFFRACTION). The technique is based upon three principles:

1. Electrons scatter x rays such that the amplitude of the scattered wave is proportional to the electron density of the sample;
2. the scattered waves recombine constructively or destructively; and
3. the way in which the scattered waves combine (interfere) depends upon the distances between the reflecting planes and the wavelength of the electromagnetic radiation.

For constructive interference to occur (that is, for the reflected waves to be reinforced), the equation

$$2d \sin\theta = n\lambda \qquad (1)$$

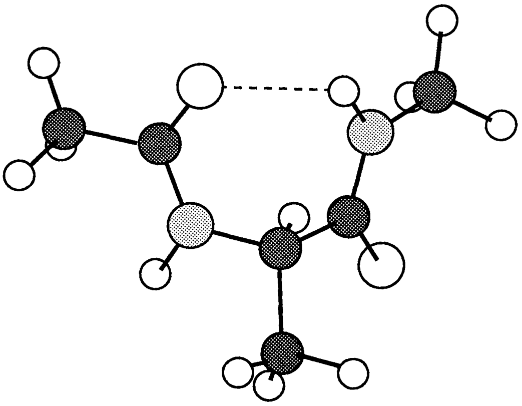

FIG. 11. The conformation of the alanine residue (N-acetyl-N'-methylalanineamide) in its global minimum. The dotted line represents a hydrogen bond. This conformation is called C_7 because the hydrogen bond forms a cyclic group of seven atoms. The large white atoms are oxygen; the small white atoms are hydrogen; the light gray atoms are nitrogen; and the dark gray atoms are carbon. The backbone atoms of alanine begin with the first nitrogen from the left and proceed right to the C^{α} atom and then to the C' (carbonyl) atom.

must be satisfied, where *d* is the perpendicular spacing between the lattice planes in the crystal, θ is the angle between the incident (or the reflected) beam and the plane of the reflecting particles, *n* is an integer, and λ is the wavelength. This is called the Bragg law.

Since bond lengths in molecules are $\sim 1.5 \times 10^{-10}$ m (~ 15 nm), the wavelengths of the radiation must also be in the range of ~ 15 nm. That is, the diffracting electromagnetic radiation must be x rays, which have wavelengths of 10^{-12}–10^{-7} m.

To determine the locations of the atoms within a molecule using x-ray diffraction requires these general steps:

1. Purifying and crystallizing the sample.
2. Measuring the diffraction angles and intensities using an x-ray diffractometer.
3. Calculating electron-density maps based upon the data gathered from the diffractometer and analyzed using Fourier transforms (see FOURIER AND OTHER MATHEMATICAL TRANSFORMS).
4. Relating the location of the atoms within the molecules to the electron-density maps.

The major advantage of x-ray diffraction is that the precise location of all the atoms (except hydrogen) within the sample molecule can be obtained. The conformation of the molecule in the crystalline state is therefore known unequivocally. The major disadvantages of the technique are that

1. producing a crystal of sufficient purity and size can be difficult or even impossible, especially for large molecules;
2. gathering and analyzing data can be extremely time consuming (although modern diffractometers and computers have removed much of the tedium from the process); and
3. determining the dynamics of conformational change in response to environmental changes is usually impossible.

An x-ray structure represents the single most stable conformation of the molecule based upon intermolecular and intramolecular interactions. Molecules in the gas phase or in solution, on the other hand, can often exist in (and interconvert among) many different conformations. Therefore, the x-ray structure may have little relationship to the physical and chemical properties of the molecule in solution.

2.2 Nuclear Magnetic Resonance

Nuclear magnetic resonance (NMR) is one of the most popular tools for analyzing the conformation of organic and biological molecules (see NUCLEAR MAGNETIC RESONANCE; MAGNETIC RESONANCES, DETERMINATION OF STRUCTURES BY). This popularity results from three factors (Freifelder, 1982):

1. the theory of NMR is sufficiently developed to determine—at least in principle—the arrangement of all the atoms in a molecule from the NMR spectrum;
2. hydrogen atoms—whose electron densities are too low for observation by x-ray diffraction—can be located using NMR; and
3. different types of nuclei—such as ${}^{1}_{1}$H, ${}^{13}_{6}$C, ${}^{15}_{7}$N,—can be examined separately.

Moreover, NMR is effective in analyzing conformations in solution and in measuring conformational changes in response to environmental changes. For species with molecular weights < 500, NMR can usually yield the locations of all the atoms in the molecule. For larger species, where determining the precise locations of all the atoms may be impossible, NMR can still yield much valuable information about the conformation, including hydrogen-bonding schemes, interatomic distances between certain atoms, and binding by other species.

The principles of NMR can be understood from the quantum-mechanical description of atomic nuclei (see ELECTRONIC STRUCTURE OF ATOMS). Consider, for example, the nucleus of a hydrogen atom. Its proton possesses two possible "spin" quantum numbers (angular momentum eigenvalues): $+1/2$ and $-1/2$. Each of these spin states gives rise to a magnetic field whose strength and direction is expressed as the magnetic moment μ. When a proton is placed in an external magnetic field of strength H_0, the magnetic moment of the proton can assume one of two angles relative to the direction of H_0: θ (aligned with) or $180-\theta$ (aligned against the external magnetic field). The difference in potential energy between these two magnetic states is $2\mu H_0 \sin\theta$.

When a proton absorbs this amount of energy—which lies within the radio-frequency range of the electromagnetic spectrum—the proton undergoes a transition between the two magnetic states. The proton is said to "flip" from one orientation to the other.

What has been stated about protons (the nuclei of hydrogen atoms ^1_1H) is also true of other atomic nuclei, most notably $^{13}_6C$, $^{15}_7N$, $^{19}_9F$, and $^{31}_{15}P$. All have a quantum-mechanical spin number I of $\frac{1}{2}$, which gives rise to two nuclear energy levels in the presence of an external magnetic field.

If all nuclei exposed to the same applied magnetic field underwent a transition at the same frequency, NMR would have little value in determining molecular conformations. In reality, however, when a molecule is placed in an applied magnetic field, the individual nuclei within the molecule do not see the same magnetic field. The effective magnetic field, H_{eff}, seen by each nucleus differs from the applied magnetic field H_0, because each nucleus in a molecule is surrounded by electrons that also generate a magnetic field of their own, thus "shielding" the nucleus to some degree from the applied magnetic field. Moreover, the effective magnetic field seen by nucleus is influenced by the magnetic field of nearby nuclei. Thus, the observed resonance frequency of a particular nucleus depends on its environment. This is why NMR spectroscopy is valuable in conformational analysis; each nucleus of the same type has a resonance frequency that depends on the conformation of the molecule.

The major advantages of NMR are that the technique can

1. yield data from ensembles of conformation (although they are sometimes difficult to understand);
2. provide information about conformational dynamics;
3. yield spectra of samples in the solid, liquid, or gas states, including aqueous and nonaqueous solutions; and
4. provide a considerable amount of conformational information, including precise locations of atoms in small molecules.

The major disadvantages are that

1. the signals from water can sometimes interfere with the sample signal from proton (^1H) NMR (although taking the spectrum in D_2O or using high-resolution Fourier-transform NMR instruments can usually overcome this problem);
2. the spectra from dilute or impure samples are difficult to obtain; and
3. the spectra of large, complex molecules are difficult (and sometimes impossible) to interpret.

2.3 Ultraviolet Absorption Spectroscopy

Ultraviolet (UV) absorption spectroscopy has long been used as an analytical technique for identifying compounds and determining concentrations, but it also has value in analyzing conformational properties of compounds. For example, the UV absorption spectrum of a polypeptide can be used to demonstrate the presence of α-helices, β-pleated sheets, or nonregular ("random coil") structure, and to analyze helix-to-coil transitions.

UV spectroscopy cannot match x-ray crystallography or NMR spectroscopy for the amount of detailed information obtained, but UV spectroscopy has fewer experimental constraints. The major advantage of UV spectroscopy over most other popular techniques is that spectra can be obtained from only partially purified samples, from dilute solutions, and from small samples, with relatively inexpensive instruments. Moreover, changes in conformation as a function of change in temperature, pressure, or other environmental conditions are easily monitored with UV spectroscopy.

UV spectroscopy is based upon transitions between different electronic energy levels of the electrons upon absorbing UV light (electromagnetic radiation with a wavelength below 400 nm). For example, the amide unit in peptides and other molecules undergoes the so-called n-π^* transition at ~220 nm, the π_1-π^* transition at ~190 nm, and the π_2-π^* transition at ~165 nm. The exact values of these transitions depend upon the specific type of molecule and on its conformation. In the synthetic polypeptide poly-L-lysine, for example, the location of the maximum in the absorption spectrum is at 191 nm for the α-helix, 193 nm for the random coil, and 196 nm for a β-pleated sheet. Although these changes are small, they can be measured accurately. Moreover, the differences in the amount of absorption are more dramatic: the three types of conformation have molar absorptivities of 4300, 7100, and 7800, respectively (Gratzer, 1967).

UV spectroscopy, therefore, is a powerful technique in conformational analysis when the spectra of different conformations have been well characterized and the goal of the analysis is to determine the general conformation of the sample or to study its conformational changes.

2.4 Infrared Spectroscopy

Infrared spectroscopy (IR) is best known as a tool for determining structures of organic compounds, but it can also be used as a tool in conformational analysis. For example, O–H and H–H groups produce different IR absorption bands depending on whether or not they are involved in a hydrogen bond (see MOLECULAR SPECTROSCOPY).

IR spectroscopy is based upon the absorption of electromagnetic radiation by transitions in the vibrational levels (due to bond stretching and bending) in a molecule. The energies of these transitions are generally in the range of infrared radiation, which has a wavelength range of 10^{-6}–10^{-4} m (1–100 μm). IR spectra usually cover the range 2.5–16 μm, which corresponds to the wave-number range of 400–625 cm^{-1}. (Traditionally, IR absorption is plotted as a function of wave numbers rather than wavelengths.)

One of the major disadvantages of traditional IR spectroscopy is that the absorption band for H_2O (and in some cases even D_2O) occurs in the same wave-number range as the main absorption bands of many organic and biological molecules, greatly diminishing the value of IR spectroscopy in analyzing conformations in aqueous solutions. With the advent of Fourier-transform infrared spectroscopy (FT-IR), the technique has greatly increased in its usefulness as a method of conformational analysis.

2.5 Circular Dichroism and Optical Rotatory Dispersion

One of the most useful techniques for analyzing the conformations and the conformational changes of asymmetric (or optically active) compounds, particularly biopolymers, is circular dichroism (CD) and the related technique, optical rotatory dispersion (ORD). In concept, CD and ORD provide the same information, but in practice, CD has become the method of choice because its spectra show greater sensitivity to conformation.

When plane-polarized light passes through a sample of a chiral compound, the light is changed in two ways:

1. The maximal amplitude is no longer confined to a plane but rather traces out an ellipse. The resulting ellipticity of the light is defined as the arc tangent of the ratio of the minor axis to the major axis of the ellipse. CD is a measure of the ellipticity caused by the chiral compound.
2. The orientation of the ellipse may be rotated. In other words, the plane of polarized light is rotated. Optical rotatory dispersion is a plot of the amount of rotation as a function of wavelength.

Consider an alternative way of looking at ellipticity. Plane-polarized light can be resolved into two circularly polarized components of equal amplitude. The sum of the two components at all points results in the original plane-polarized wave. In an optically active medium, the left circularly polarized light (A_L) and the right circularly polarized light (A_R) are absorbed differently. This is called circular dichroism. When the two opposite circularly polarized waves are combined, the result is an ellipse, as shown in Fig. 12. Thus, CD is equivalent to ellipticity. A CD spectrum is a plot of the ellipticity θ (or specific ellipticity, $[\theta]$) as a function of wavelength.

The major advantages of CD and ORD are that the spectra are sensitive to conformational changes and that the environmental conditions are easily modified. Thus, conformational changes in biopolymers are easily monitored using ORD and CD. For example, Fig. 13 shows the CD spectra of poly-L-lysine (a model polypeptide) in α-helix, β-pleated sheet, and random coil conformations (Greenfield and Fasman, 1969). In the figure, $[\theta]$ is

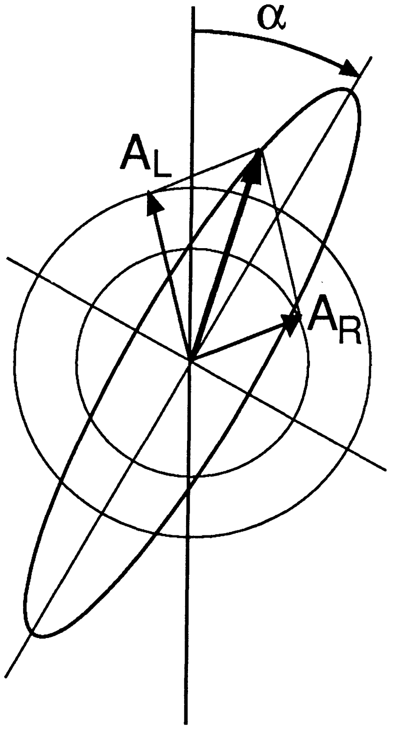

FIG. 12. Circular dichroism and optical rotation. The circular dichroism (or ellipticity) is proportional to the difference between the amplitude of the left circularly polarized wave A_L and of the right circularly polarized wave A_R. The optical rotation is given by α.

the specific ellipticity, and the wavelength λ is taken in the ultraviolet region where the peptide unit absorbs light. Notice the huge difference in the spectra for the three different types of conformation. CD can likewise be used to analyze the general conformation of proteins, carbohydrates, and nucleic acids, as well as other kinds of molecules.

2.6 Other Experimental Techniques

The five experimental techniques discussed above are the most popular methods for analyzing the conformation of molecules, but they are by no means the only methods. Other popular experimental techniques include fluorescence spectroscopy, ultracentrifugation (and other hydrodynamic techniques), light scattering, Raman spectroscopy, Mössbauer spectroscopy, electron paramagnetic resonance spectroscopy, and so forth. For a general description of these techniques, see Cantor and Schimmel (1980b) and Freifelder (1982).

3. THEORETICAL TECHNIQUES

Theoretical (computational) techniques are valuable in the conformational analysis of organic and biological molecules because the results of a theoretical analysis can

1. provide a model or framework to aid in choosing proper experiments;
2. help explain experimental results; and
3. provide information about conformational properties of molecules when experiments are difficult or impossible to perform.

The major objectives of theoretical analysis are

1. to calculate the conformational properties of a molecule given its covalent structure;

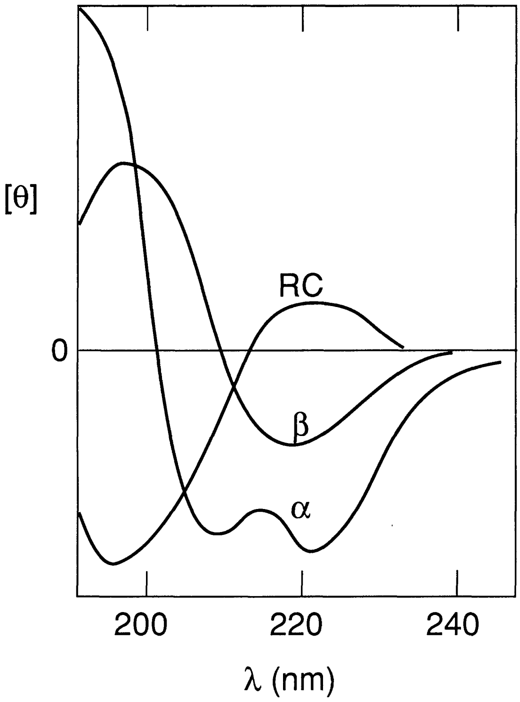

FIG. 13. The circular dichroism spectrum of poly-L-lysine in three different conformations: α-helix, β-pleated sheet, and random coil (RC). The conformational changes are induced by changing the pH. From Greenfield and Fasman (1969).

2. to determine the effects of the environmental conditions on the conformation; and
3. to relate the calculated conformation of a molecule to its physical properties and its biological activity.

3.1 Empirical Methods

Empirical methods of conformational analysis involve empirical equations (that is, equations based on experiment or observation without regard to fundamental theory) that relate the positions of the atoms to the conformational energy, and involve parameters used in the empirical equations. In a so-called molecular mechanics system, computer software employs the equations and parameters to calculate the conformational energy, minimize the total conformational energy for locating local and global minima, and display on a computer monitor a graphic representation of the conformations.

A commonly used empirical equation (Momany *et al.*, 1975; Feller *et al.*, 1989) is the so-called Lennard-Jones "6-12" potential:

$$U_{\mathrm{NB}}=F\epsilon\left(\frac{r_g}{r_{ij}}\right)^{12}-2.0\epsilon\left(\frac{r_g}{r_{ij}}\right)^{6}, \qquad (2)$$

where r_g is the equilibrium distance between the atoms i and j, ϵ is the energy at the equilibrium distance, and r_{ij} is the distance between atoms i and j. The positive (repulsive) term contains r^{12}, and the negative (attractive) term contains r^6. The attractive term is the London dispersion force (London, 1937). The F is an adjustment factor taken as 1.0 for 1–5 interactions (those between atomic pairs separated by at least four bonds) and as 0.5 for 1–4 interactions (those between atomic pairs separated by three bonds). The subscript NB stands for "nonbonded," meaning the energy of interaction between atoms that are not bonded to each other. The values of r_g and ϵ depend on the particular types of atoms involved in the interaction.

Along with the Lennard-Jones 6-12 potential, conformational energy calculations also require an equation to take into account the electrostatic interactions. For example, the most common equation for determining electrostatic energy between two atoms is

$$U_{\mathrm{el}}=kq_iq_j/Dr_{ij}, \qquad (3)$$

where k is a constant to ensure that the energy has the correct units, q_i and q_j are the partial charges on atoms i and j, D is the dielectric constant, and r_{ij} is the distance between atoms i and j.

For certain types of bonds (particularly double bonds), the nonbonded and electrostatic energies do not completely describe the restricted rotation observed for the bond. Therefore, torsional energy is also used in some procedures to determine the conformational energy:

$$U_{\mathrm{tor}}=\frac{U_0}{2}(1\pm\cos n\theta), \qquad (4)$$

where U_0 is the maximum torsional energy, θ is the torsional angle, and n is the symmetry of the barrier.

The total conformational energy of a molecule in a particular conformation is the sum of nonbonded energies U_{NB} and electrostatic energies U_{el} for all i,j pairwise interactions and of all the torsional energies U_{tor}. This procedure was used to calculate the total conformational energy at each ϕ, ψ point of the conformational energy contour map shown in Fig. 10. It can also be used to analyze the conformational properties of larger peptides and other biological and organic compounds. Because the calculations require large amounts of computer time, the procedure does not permit the complete conformational analysis of proteins, nucleic acids, carbohydrates, or other macromolecules.

In many approaches to molecular mechanics, the bond lengths and bond angles are also considered variable. Appropriate potential-energy functions include not only the interactions between the nonbonded atoms, but also between the bonded atoms (Burkert and Allinger, 1982).

3.2 Quantum-Mechanical Approaches

Pullman (1976) has argued that empirical energy functions "suffer from two obvious drawbacks. In the first place, whatever the practical justification for the partitioning of the total potential energy into a series of components, the procedure involves necessarily an element of arbitrariness and, possibly, incompleteness. In fact, most authors take into consideration different selected components only and frequently different sets of them in different publications. Secondly, the fundamental formulas and parameters used to define the various components are far from being well established and differ, often appreciably, from one author to another."

The conclusion that Pullman and others draw from these observations is that only quantum-mechanical methods, not empirical methods, have the ability to account for the subtle quantum-mechanical effects that may be part of the conformational properties of a molecule.

These arguments certainly have some merit. Quantum mechanics has indeed contributed to the field of theoretical conformational analysis (see QUANTUM MECHANICS). But in actual practice, quantum-mechanical methods also involve a certain degree of arbitrariness, require broad approximations, and involve huge amounts of computer time (Zimmerman, 1985). Both quantum-mechanical and empirical methods are valuable in addressing the problems of molecular conformations, since each approach brings to the problem its own strengths and weaknesses.

The most popular semiempirical molecular-orbital methods used for conformational analysis include the extended Hückel (EHT) method (Hoffman, 1963), the complete neglect of differential overlap (CNDO/2) methods (Pople and Segal, 1966), and the all-valence electron molecular-orbital (perturbative configuration interaction using localized orbitals) method (Diner *et al.*, 1969) (see ELECTRONIC STRUCTURE OF ATOMS AND MOLECULES).

EHT is a molecular-orbital treatment in which the interactions between electrons in different orbitals are ignored. CNDO/2 assumes the atomic orbitals are spherical when electron repulsions are evaluated. Directionality of p orbitals is included, however, through the so-called one-electron resonance integrals. CNDO calculations are aimed primarily at reproducing electronic properties, such as dipole moment (Clark, 1985).

Perturbative configuration interaction using localized orbitals goes beyond CNDO/2 by including a refined molecular-orbital all-valence electron procedure that incorporates an appreciable fraction of the correlation energy (Maigret *et al.*, 1971).

Ab initio methods (see QUANTUM MECHANICS) are also used in theoretical conformational energy analysis (e.g., see Shipman and Christofferson, 1973; Peters and Peters, 1981).

3.3 Applying Theoretical Techniques

Whether using empirical, semiempirical (molecular-orbital), or *ab initio* theoretical methods, the theoretician utilizes the following general steps in determining the conformational properties of a molecule:

1. Selecting a set of initial trial conformations, either randomly (e.g., using a Monte Carlo method; see Li and Scheraga, 1987) or systematically (e.g., by stepping through all of conformational space).
2. Calculating the total conformational energy of the molecule in each of the trial conformations. If the theoretician selects and calculates enough trial conformations, the results will yield the most favored (lowest-energy) conformation or will provide an ensemble of conformational states on which to base a statistical-mechanical analysis.
3. Minimizing the total conformational energy for each of the trial conformations to yield local and global minima. If this method yields a large number of minima, the set of minima may constitute the partition function for a statistical-mechanical analysis of the conformational properties. If minimization yields a small number of minima, inspection of those minima may provide all the necessary information about the conformational properties of the molecule.

Applying these techniques to peptide molecules, as an example, has yielded important information about the structure-function relationship in biological systems (see Zimmerman, 1985).

GLOSSARY

α-Helix: A polypeptide or protein conformation in which the backbone chain forms a right-handed screw with 3.6 amino acid residues per turn and a pitch (rise per turn) of 54 nm.

β-Pleated Sheet: A polypeptide or protein conformation in which the backbone chains are extended and hydrogen bonded to adjacent chains.

Conformation: The overall three-dimensional shape of a molecule.

Contour Plot: A plot of the conformational energy of a molecule as a function of two variable dihedral angles. Each contour line represents conformations of equal energy.

Dihedral Angle: The rotational angle of a bond, defined such that the angle is 0° when the two bonds flanking the torsional bond are eclipsed when looking along the bond and 180° when the two flanking bonds are *anti* (or *trans*) to each other.

Global Minimum: The conformation of the lowest energy of a molecule.

Local Minimum: A conformation found at a minimum-energy point (not necessarily the lowest, or global, minimum).

Molecular Modeling: A technique that employs computer graphics and calculations to study molecular conformations and dynamics.

Molecular Mechanics: A method for calculating the conformations of molecules based upon empirical energy functions.

Random Coil: A statistical ensemble of polymer conformations in which the backbone of the molecule has no order.

Torsional Angle: See dihedral angle.

List of Works Cited

Burkert, U., Allinger, N. L. (1982), *Molecular Mechanics,* Washington, DC: American Chemical Society.

Cantor, C. R., Schimmel, P. R. (1980a), *Biophysical Chemistry. Part I. The Conformation of Biological Macromolecules,* San Francisco: Freeman.

Cantor, C. R., Schimmel, P. R. (1980b), *Biophysical Chemistry. Part II. Techniques for the Study of Biological Structure and Function,* San Francisco: Freeman.

Carey, F. A. (1987), *Organic Chemistry,* New York: McGraw-Hill.

Clark, T. (1985), *A Handbook of Computational Chemistry,* New York: Wiley.

Diner, S., Malrieu, J. P., Jordan, F., Gilbert, M. (1969), *Theor. Chim. Acta* **15**, 100–110.

Feller, D. C., Delmoe, E. F., Zimmerman, S. S. (1989), *Comput. Chem.* **13**, 337–341.

Freifelder, D. (1982), *Physical Biochemistry,* 2nd ed. San Francisco, Freeman.

Gratzer, W. B. (1967), in G. D. Fasman (Ed.), *Poly-α-amino Acids,* New York: Marcel Dekker, pp. 177–238.

Greenfield, N., Fasman, G. D. (1969), *Biochemistry* **8**, 4108–4116.

Hoffman, R. (1963), *J. Chem. Phys.* **39**, 1397–1412.

International Union of Pure and Applied Chemistry–International Union of Biology Commission on Biochemical Nomenclature (1970), *J. Mol. Biol.* **52**, 1–17.

Li, Z., Scheraga, H. A. (1987), *Proc. Natl. Acad. Sci. U.S.A.* **84**, 6611–6615.

London, F. (1937), *Trans. Faraday Soc.* **33**, 8–26.

Maigret, B., Perahia, D., Pullman, B. (1971), *Biopolymers* **10**, 1649–1660.

Momany, F. A., McGuire, R. F., Burgess, A. W., Scheraga, H. A. (1975), *J. Phys. Chem.* **79**, 2361–2381.

Orville-Thomas, W. J. (1974), in: W. J. Orville-Thomas (Ed.), *Internal Rotation in Molecules,* New York: Wiley.

Peters, D., Peters, J. (1981), *Int. J. Quantum Chem.* **19**, 1121–1132.

Pople, J. A., Segal, G. A. (1966), *J. Chem. Phys.* **44**, 3289–3296.

Pullman, B. (1976), in: B. Pullman (Ed.), *Quantum Mechanics of Molecular Conformations,* New York: Wiley.

Shipman, L. L., Christofferson, R. E. (1973), *J. Am. Chem. Soc.* **95**, 1408–1416.

Zimmerman, S. S. (1985), in: V. J. Hruby (Ed.), *The Peptides: Analysis, Synthesis, Biology,* Vol. 7, *Conformation in Biology and Drug Design,* Orlando, FL: Academic.

Zimmerman, S. S., Pottle, M. S., Némethy, G., Scheraga, H. A. (1977), *Macromolecules* **10**, 1–9.

Further Reading

Derome, A. E. (1987), *Modern NMR Techniques for Chemistry Research,* New York: Pergamon.

Glass, R. S. (1988), *Conformational Analysis of Medium Sized Heterocycles,* New York: VCH.

Glusker, J. P., Trueblood, K. N. (1985), *Crystal Structure Analysis: A Primer,* Oxford: Oxford University.

Marchand, A. P. Takeuchi, Y. (Eds.) (1986), *Applications of NMR Spectroscopy to Problems in Stereochemistry and Conformational Analysis,* Deerfield Beach, FL: VCH.

Mason, S. F. (1982), *Molecular Optical Activity and the Chiral Discriminations,* New York: Cambridge University.

Rasmussen, K. (1985), *Potential Energy Functions in Conformational Analysis,* New York: Springer-Verlag.

Stout, G. H. Jensen, L. H. (1989), *X-ray Structure Determination,* New York: Wiley.

3

Fundamental Constants

J. L. Flowers and B. W. Petley

3.1 Introduction 35
3.2 Conceptual Origins and Importance of the Constants 39
3.2.1 Conceptual Origins of the Constants 39
3.2.2 Importance of the Constants 40
3.2.3 QED and the Josephson Effect Determination of $2e/h$ 42
3.2.4 QED and the Quantum Hall Effect Determination of h/e^2 46
3.2.5 The Quantum Metrology Triangle 48
3.3 Obtaining "Best Values" of the Constants 49
3.3.1 Relationships among the Constants 49
3.3.2 The 2006 CODATA Adjustment 50
3.3.2.1 Uncertainty Assignment and Discrepant Data 55
3.3.2.2 Recommended Values of the Constants 55
3.3.2.3 Comparison of the 1986 and 2006 Adjustments 56
3.4 Josephson and Quantum Hall Effect Representations of the Volt and Ohm 57
3.4.1 The SI and Representations of the Volt and Ohm 58
3.4.2 The Conventional Values K_{J-90} and R_{K-90} 59
3.5 Implications of Recent Advances in the Fundamental Constants Field 59
3.6 Possible Role of the Fundamental Constants in Replacing the Kilogram 60
3.6.1 Problems with Present Kilogram 60
3.6.2 An Alternative Based on Fundamental Constants 61
3.6.2.1 The Measurement of h 61
3.6.2.2 Methods Involving the Avogadro Constant, N_A 62
3.6.3 Practical Mass Metrology Making Use of a New Definition of the Kilogram 65
3.7 Numerology and Time Variations of the Constants 66
3.8 Future Trends and Advances 68
Glossary 69
References 71
Further Reading 73
Useful Web Sites 73

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

3.1
Introduction

This chapter discusses the fundamental physical constants of nature such as the speed of light in vacuum c, the mass of the electron m_e, the Newtonian gravitational constant G, the elementary charge e, the Planck constant h, the Avogadro constant N_A, Faraday constant F, the Rydberg constant (for infinite mass) R_∞, the fine-structure constant α, the Josephson constant K_J, and the von Klitzing constant R_K. Such constants are nature's true invariants in contrast to, say, the freezing point of gallium, or the time required for the Earth to revolve once about its axis, which can be changed–for example, the former by impurities and the latter by seasonal effects. It does not necessarily mean that there might not exist a deeper set of constants in terms of which the above fundamental constants may ultimately be expressed, as indeed some of them, such as the above Faraday, fine-structure constant, the Rydberg constant, and others may be already be expressed in terms of e, h, μ_0, and c. Thus, they have a unique place in science and technology and have even acquired a certain mystique. Determination of their values spurred a considerable amount of activity that peaked between about 1970 and 1985. There are many important combinations of the fundamental constants that are widely used in science and technology. The 2006 CODATA (Committee on Data for Science and Technology) evaluation of the ''best'' values for the fundamental constants contains some 300 recommended values for various combinations of the constants. It is convenient to discuss progress in this field of endeavor with respect to the recommended values referred to the chapter by Barry Taylor in the earlier edition of the *Encyclopedia of Applied Physics* on this topic that we update here.

In all of our fundamental theories, there appear a certain few parameters that characterize the fundamental particles and the interactions that we find in nature. These are termed the *fundamental physical constants*.The precision determination of the numerical values of these constants, in terms of International System of Units (SI) units when they are dimensioned constants, has long been and still remains one of the principal objectives of experimental science. This is not because there is any intrinsic virtue in accumulating lists of numbers with ever more decimal places, but because the fundamental constants are the quantitative links between our most basic theories and the physical reality of our universe that we wish them to describe. Our theories must stand

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

or fall according to their ability to make quantitative predictions that agree with experimental observations to the maximum achievable accuracy. The progress of our understanding of the physical world is therefore very much interwoven with the advance of the art of precision measurement and its application to the determination of the fundamental physical constants.

This chapter provides an illustration of the *raison d'etre* for striving for increased measurement accuracy that Maxwell and many other scientists have long recognized:

> *... the opinion seems to have got*
> *abroad that in a few years all of*
> *the great physical constants will*
> *have been approximately measured*
> *and the only occupation which*
> *will be left to men of science will*
> *be to carry on these measurements*
> *to another place of decimals....*
> *But the history of science shows*
> *that even during that phase*
> *of her progress in which she*
> *devotes herself to the improving*
> *the accuracy of quantities with*
> *which she has long been familiar,*
> *she is preparing materials for the*
> *subjugation of new regions which*
> *would have remained unknown*
> *if she had been content with*
> *the rough methods of her early*
> *pioneers.* (Maxwell, 1871)

Advances in accuracy continue to be made. Over the last 80 years, measurement accuracy has advanced about 10-fold every 15 years, a process that may well continue until natural limits are reached. Not only has this topic had its share of Nobel Prize winners but, in recent years, there also have been notable advances in both theory and experiment relevant to the fundamental constants with applications moving out into the market place. New technologies have made possible new levels of precision: Computers have made possible more complex calculations. The advent of increasingly stable lasers has led to new levels of accuracy being achieved in many different experiments, for example, optical spectroscopy in the visible region with Hertz-level, that is fractional 10^{-16} accuracy (Coddington, Swan, and Newbury, 2008; Ludlow *et al.*, 2008), and measurements of the Rydberg constant with fractional 10^{-12} accuracy level (Fischer *et al.*, 2004; Hänsch *et al.*, 2005). The discovery of the Josephson effects (JEs) in superconductors and the quantum Hall effect (QHE) in semiconductors have provided new ways to determine several important fundamental constants. Ion traps and atom traps of single particles have increased measurement times. There has been steady progress in precision measurements of the properties of fundamental particles and simple atomic systems.

Current theoretical and experimental efforts in this field span almost the whole of physics, from elementary particle theory to experimental solid-state physics. The intercomparison of measurements at the highest accuracy via very different methods in different parts of science permits detailed examination to be made of the coherence of science as a single entity. In an increasing number of cases, the experimental accuracy has reached an artificial limit set by our ability to realize the appropriate combination of the practical units of measurement, the SI. A consequence of this has been the increasing application of methods of measuring fundamental constants either to maintain more stable SI units, or to incorporate them into the definitions of the unit(s) concerned–as happened in 1983 when the value of the speed of light was adopted as the present definition of the meter. This is approaching the point where most of the SI base units are definable in terms of fixed values for fundamental constants.

There are also interrelationships between the measurable combination of fundamental constants and related quantities, for example, the Rydberg and fine-structure constants, $2e/h$ via the JE, realizations of the absolute ohm and ampere, the quantized Hall resistance, the electron and muon anomalous magnetic moments, the proton-to-electron mass ratio, the proton gyromagnetic ratio, Faraday, and Avogadro constants. Consequently it is appropriate, indeed even necessary, for the solid-state theorist to be aware of the work of the electrical metrologist; the atomic mass spectroscopist of the work of the precision electrochemist; and the quantum electrodynamics (QEDs) theorist of the work of the experimental semiconductor physicist. The fundamental constants of nature and closely related precision measurements are truly the common meeting ground for many of the disciplines of science and technology.

It is clear from the earlier part of this introduction, that the field of fundamental constants too broad for any one does more than touch upon a few key selected topics. However, these should provide the reader with a reasonable picture of the field, especially the very important interplay between theory and experiment that is the key to scientific progress. As elsewhere in science, much of the beauty of the subject lies in the detail, including how the measurements were made in the laboratory or the detailed calculations of the theoretician were brought to the meeting ground between theory and experiment at the fractional part in a billion (1×10^{-9}) level of accuracy to $<1 \times 10^{-12}$ and beyond. Both the measurement and the calculation may well have taken 10–15 years but the work is distilled into a four-page article!

The earlier reviews of the recommended values of the fundamental constants proceed at appropriate intervals of about once every 10 years. The date attributed to each review refers to the latest date for which published data was included. Thus the 2006 evaluation had a closure date of 31 December 2006 and was published in 2008. The values in this review replaced those in the 2002 review. It is important for the coherence of science that as far as possible scientists use the same values for the fundamental constants. Birge pioneered the subject and published the first major review of the values in 1929 (Birge, 1929) and most subjects were happy to adopt his recommended values. Since then the reviews have usually been mainly made by two or three people. However, since the 1970s, reflecting that the subject has become increasingly international, the reviewers have been supported by an international committee known as the *CODATA Task Group on Fundamental Constants.*Recently, the CODATA Task Group reviews have taken place every four years. From time to time others have been motivated to perform their own reviews, for example, the Taylor, Parker, and Langenberg review in 1969 (Taylor, Parker, and Langenberg, 1969) and the 1975 review by Tuninskii and Kholin (1975) and Tuninskii (1986). These reviews have been welcomed and, in several cases, have helped clarify discrepancies and have thereby served to progress the subject.

The main body of the chapter (Section 3.2) begins with a brief discussion of the "origins" of the fundamental constants, that is, how they might be usefully categorized; and a lengthier discussion of "the challenge and romance of the next decimal place,'' their importance and why their determination to ever greater levels of accuracy can have a profound effect on physics as a coherent whole by allowing any small differences between constants measured by very different measurements to be thoroughly investigated.

Besides the earlier example of the Millikan oil-drop determination of the elementary charge, a later illustration, which is featured here, comes from work in the late 1960s through the determination of the quotient $2e/h$ via the JE in superconductors (a topic in which two of the present authors participated). The derivation of a value of the fine-structure constant (α) from it, when combined with other constants, enabled the use of the resulting value of α to test the QED theory of the ground-state hyperfine splitting of the hydrogen atom to well beyond the sixth decimal place. A similar example that is still under active investigation is also presented in which a value of α obtained from the QHE is used to test the QED theory of the electron magnetic moment anomaly a_e.

In Section 3.3, we discuss how sets of the ''best values'' constituting the recommended values of the fundamental constants are derived, particularly concentrating on the 2006 CODATA least-squares adjustment of the constants, which is the most recent internationally accepted evaluation currently available. In this section, we include some of these recommended values for a comprehensive group of constants, which are taken from the 2006 CODATA set.

The realization of the definitions of the SI units is very time consuming. Modern science and technology relies on making measurements involving time and frequency and electrical quantities, and there is a universal need throughout the world to maintain a continuous representation of the SI units. In Section 3.4, we discuss the importance of the fundamental constants to technology by reviewing the 1 January 1990 introduction of new practical representations of the volt and ohm as defined in the SI. These representations, based on the JEs and quantized Hall effects, and the consequent internationally adopted values of the Josephson and von Klitzing constants K_J and R_K, are now being used worldwide in electrical metrology to derive and maintain representations of the volt and ohm to fractional accuracies of 10^{-9} or better.

In Section 3.5, we discuss the principal advances made in the fundamental constants field since the 1986 CODATA adjustment was completed and discuss their implications for the least-squares adjustment of 2006. In Section 3.6, we further consider the importance of the fundamental constants to modern technology, and illustrate these by exploring the likely future role of the constants in replacing the definition of the kilogram, the only SI base unit that is still defined in terms of a prototype material artifact, and to similarly replace the definitions of other SI units. In Section 3.7, we briefly discuss the question of why the constants have the particular values that they do, and their possible variation with time. Finally in Section 3.8, which concludes the main body of the chapter, we look ahead in an attempt to see where the fundamental constants field might be heading over the next decade.

It is useful at this point to mention briefly some aspects of the history of the measurements of the fundamental constants. Modern measurements of the fundamental constants really began at the start of the twentieth century modern physics, with the advent of quantum theory, relativity, and atomic and nuclear physics. The two notable exceptions were c and

G, which were first measured in the eighteenth century.

Extraordinary advances in determining the values of the constants were made immediately after 1945 as a result of the significant advances made in electronics, microwaves, and other technologies during World War II. A second period starting in about 1970 has also been especially productive because of the wide introduction of highly sophisticated digital electronic instruments, lasers, computers, and cryogenic techniques. To a considerable extent, the accuracy with which the fundamental constants have been known at any given epoch has depended upon the level of sophistication of the measuring instruments and apparatus available at the time. However, it is not only improvements in measurement technology that have led to more accurate values of the constants. The discovery of new phenomena, such as the JEs and the quantized Hall resistance effects, that provide new and better methods for determining certain constants have also led to significant advances in the field.

Measurements of the constants and the frontiers of measurement science or metrology, especially the development of new reference standards for the units of measurement, have always been closely linked, but in recent years the distinctions between them have all but disappeared. Perhaps the most striking example of this is the determination of c in the early 1980s based on extremely difficult measurements of the frequencies and wavelengths of several different laser radiations. As a direct consequence of this work, the definition of the SI unit of length, the meter, which, since 1960s was based on a particular optical transition in the ^{86}Kr atom, was replaced in 1983 with one that was based on an adopted value for the speed of light (Petley, 1983; Giacomo, 1984; Hudson, 1984).

As discussed in Section 3.6, if current efforts to determine h or N_A with a relative uncertainty of about 1×10^{-8} are successful, it is very probable that this scenario will be repeated with regard to the kilogram in the second or third decade of this century. There may be consequent changes to our definitions of the electrical units, while a fixed value for the Boltzmann constant may well be incorporated into a revised definition of the kelvin. This will be an important topic for the next meeting of the General Conference on Weights and Measures (CGPM) scheduled for 2011. Any changes to the definitions of the SI units are framed so as to ensure continuity in the sizes of the units concerned while ensuring that they anticipate or keep pace with the improvements in measurement accuracy enabled by, and demanded by, future progress in science and technology.

3.2
Conceptual Origins and Importance of the Constants

3.2.1
Conceptual Origins of the Constants

Three distinct ''sources'' of the fundamental constants may be identified. The first is the physical theories that have been developed to describe nature's basic laws, for example, Maxwell's theory of electromagnetism, Einstein's theories of relativity, quantum mechanics, and QED, and their application as appropriate to atomic systems. c, G, h, α, and R_∞, fall within this category. These particular constants (see Section 3.1) are of particular importance in science because c is the velocity of all electromagnetic radiation or photons (in vacuum) and the limiting velocity of all energy and matter; G determines the strength of the gravitational force between bodies; h is the fundamental constant of quantum mechanics and determines the energy of a quantum of radiation of a given frequency; R_∞ sets the scale of the energy levels of atoms; and α determines the strength of the electromagnetic force that governs how electrically charged elementary particles and electromagnetic radiation interact.

The second source is the fundamental particles such as the electron, proton, and neutron. The elementary charge e, the rest mass of the electron m_e, the rest masses of the proton and neutron m_p and m_n, the magnetic moment of the electron in units of the Bohr magneton, namely μ_e/μ_B, where $\mu_B = eh/4\pi m_e$, and the magnetic moments of the proton and neutron in units of the nuclear magneton, μ_p/μ_N and μ_n/μ_N, respectively, where $\mu_N = eh/2m_p$, respectively. These are all examples of quantities that characterize a basic property of the elementary building blocks of nature.

Conversion factors are the third source. Although not true fundamental constants like those in the first two categories, these quantities nevertheless play a critical role in the measurement of fundamental constants in terms of the SI units. Examples include the local value of the gravitational acceleration g, which relates mass to force; the ratio of the old X-ray unit of length or kilo-x-unit to the nanometer; and also the ratio of various so-called as-maintained electrical units in a particular national standard's laboratory to their SI definitions, for example, Ω_{NPL} where Ω_{NPL} is the National Physical Laboratory (NPL) as-maintained ohm. Prior to 1 January 1990 the laboratories local ohm standards were maintained as the mean resistance of a particular group of wire-wound standard resistors that were kept in an oil bath at 25°C.

It should be noted that $\alpha = e^2/(4\pi\varepsilon_0\hbar c)$ and $R_\infty = m_e c\alpha^2/(2h)$ are examples of constants that are actually combinations of other quantities, but are considered fundamental constants in their own right since the combination always appears in theoretical equations in the same way. (Here the magnetic constant, $\mu_0 = 4\pi \times 10^{-7}\ \mathrm{N\,A^{-2}}$, exactly, is the permeability of vacuum, and the electric constant $\varepsilon_0 = c^2/\mu_0$.). Many other similar examples are found in Table 3.1 of this chapter.

3.2.2
Importance of the Constants

There are at least four reasons why the fundamental physical constants play a critical role in science and technology and hence should be known as accurately as possible.

1. Accurate values of the constants are required for the critical comparison of theory with experiment, and such comparisons that advances in our understanding of the physical world. A closely related outcome is that by comparing the numerical values of the same fundamental constants obtained from experiments in the different fields of physics, the self-consistency of the basic theories of physics themselves can be tested with increasingly high accuracy (see Section 3.1).
2. The determination of the values of the fundamental constants more accurately fosters the development of new state-of-the-art measurement methods that may have a wider application. Determining the next decimal place is never trivial and usually requires an entirely new measurement technology that, as mentioned elsewhere, can take

Tab. 3.1 Recommended values of some selected physical constants.

Quantity	**Symbol**	**Value**	**Units**	**Relative uncertainty**
Speed of light in vacuum	C	299 792 458	$\mathrm{m\ s^{-1}}$	Exact
Permeability of vacuum	μ_0	$4\pi \times 10^{-7}$	$\mathrm{N\ A^{-2}}$	(Exact)
Permittivity of vacuum, $1/\mu_0 c^2$	ε_0	$8.854187817\ldots \times 10^{-12}$	$\mathrm{F\ m^{-1}}$	(Exact)
Newtonian constant of gravitation	G	$6.67428(67) \times 10^{-11}$	$\mathrm{m^3\ kg^{-1} s^{-2}}$	1.0×10^{-4}
Planck constant	H	$6.62606896(33) \times 10^{-34}$	J s	5.0×10^{-8}
$h/2\pi$	$\bar{H}$	$1.054571628(53) \times 10^{-34}$	J s	5.0×10^{-8}
Elementary charge	E	$1.602176487(40) \times 10^{-19}$	C	2.5×10^{-8}
magnetic flux quantum, $h/2e$	Φ_0	$2.067833667(52) \times 10^{-15}$	Wb	2.5×10^{-8}
Conductance quantum,	G_0	$7.7480917004(53) \times 10^{-5}$	S	6.8×10^{-8}
Electron mass	m_e	$9.10938215(45) \times 10^{-31}$	kg	5.0×10^{-8}
Proton mass	m_p	$1.672621637(83) \times 10^{-27}$	kg	5.0×10^{-8}
Proton-electron mass	m_p/m_e	1836.15267247(80)		4.3×10^{-10}
Fine-structure constant, $e^2/4\pi_o\hbar\ c$	α	$7.2973525376(50) \times 10^{-3}$		6.8×10^{-10}
Inverse fine-structure constant,	α^{-1}	137.035999679(94)		6.8×10^{-10}
Rydberg constant	R_∞	10 973 731.568527(73)	$\mathrm{m^{-1}}$	6.6×10^{-12}
Avogadro constant	N_A, L	$6.02214179(30) \times 10^{23}$	$\mathrm{mol^{-1}}$	5.0×10^{-8}
Faraday constant	F	96 485.3399(24)	$\mathrm{C\ mol^{-1}}$	2.5×10^{-8}
Molar gas constant	R	8.314472(15)	$\mathrm{J\ mol^{-1}\ K^{-1}}$	1.7×10^{-6}
Boltzmann constant	K	$1.3806504(24) \times 10^{-23}$	$\mathrm{J\ K^{-1}}$	1.7×10^{-6}
Stefan–Boltzmann constant $(\pi^2/60)k^4/\hbar^3c^2$	σ	$5.670400(40) \times 10^{-8}$	$\mathrm{W\ m^{-2}\ K^{-4}}$	7.0×10^{-8}
Non-SI units accepted for use with the SI				
Electron volt	eV	$1.602176487(40) \times 10^{-19}$	J	2.5×10^{-8}
(Unified) atomic mass unit	U	$1.660538782(83) \times 10^{-27}$	kg	5.0×10^{-8}

The digits in parentheses represent the one-standard-deviation uncertainty in the last digit(s) of the given value.

Taken from the 2006 CODATA least-squares adjustment by Mohr, Taylor, and Newell (2008).

10–40 years to bring from conception to fruition. An example is the determination of the Avogadro constant N_A, which necessitated the development of techniques to measure the lattice spacing of pure, single crystals of silicon directly in terms of meter with a relative uncertainty of 1/10th of a part per million, all uncertainties are one-standard-deviation estimates.) The result has been the extension of our ability to make accurate measurements of length in SI units to the picometer or 10^{-12}-m range and the demise of the x-unit used for measurements of X-ray and γ-ray wavelengths. Work toward an improved Avogadro constant continues using isotopically pure silicon-28.

3. Measurements of the fundamental constants also provide the key to the development of a reproducible and invariant system of measurement units, which is also a major global goal of measurement science or metrology. This marks the culmination of a long path from the use of properties of seeds and the human body, and later via properties of water and properties of the Earth, to the units used at present based on fundamental constants. Notable early examples of the latter were the involvement of the Faraday constant to maintain the International ampere at the end of the nineteenth century and the more recent involvement of the speed of light in the definition of meter resulting from the improved measurement accuracy of high frequency radiation and other technology associated with the development of radar, which were later allied with the development of highly monochromatic and stable lasers through to the visible region and beyond, and the later satellite Global Positioning System (GPS).

From the time of Maxwell, Johnstone-Stoney, Planck, Dirac, and countless others, the development of SI units that are really constant in space-time has progressed to become a reality today. The future role of artifacts such as the platinum–iridium cylinder kept in a vault at the International Bureau of Weights and Measures (BIPM), that at present defines the SI kilogram, will be to act as a flywheel to maintain sufficiently constant representations of a mass unit defined in terms of a fundamental constant. The development of ever more accurate atomic clocks for terrestrial time in applications have led to astrophysical investigations well beyond the Solar System such as monitoring the stability of the millisecond pulsars or the comparison of different types of atomic clock enabling investigations pertaining to the invariance of dimensionless constants such as α and m_p/m_e over the age of the universe.

4. Finally, values of the fundamental constants are required for computations and measurements throughout science and technology–for example, the calculation of the excited states of atoms and molecules that play an important role in air pollution monitoring, nuclear energy, and nuclear fusion.

3.2.3

QED and the Josephson Effect Determination of $2e/h$

As an example of the significant role improved measurements of fundamental constants can play in increasing our understanding of physical theory, consider the accurate determination of the quotient $2e/h$ using the JE in weakly coupled superconductors by a method that was first reported in the late 1960s (Parker, Taylor, and Langenberg, 1967). This also provides an illustration of the way in which limitations in the accuracy of the realization of the electrical units in various countries impacted in a dramatic way on our knowledge of the fundamental physical constants.

When a Josephson tunnel junction–for example, two thin films of lead separated by a 1-nm-thick thermally grown oxide layer–is irradiated with microwave radiation of frequency f, its current versus voltage curve exhibits current steps at highly precise quantized Josephson voltages U_J (Barone and Paterno, 1982; Gallop, 1991). The voltage of the nth step, $U_J(n)$, where n is an integer, is related to f by

$$U_J(n) = nf/K_J \tag{3.1}$$

where K_J is known as the Josephson constant $(= 2e/h)$. Note that it is the inverse of the superconducting magnetic flux quantum $\Phi_0(= h/2e)$.

The theory of the JEs predicts that

$$K_J = 2e/h \tag{3.2}$$

and, to date, the experimentally observed universality of Equation 3.1 is entirely consistent with the prediction. The exactness of this equation when subject to changes in the operating conditions (frequency, temperature, material, different apparatus, etc.) has been verified with fractional uncertainties ranging from parts in 10^9 to 10^{16}. It follows from Equation 3.1 that a measurement of f and U_J for the nth step will yield a value of $2e/h$. The result reported in 1969 (Parker *et al.*, 1969) using 8–12 GHz microwaves was

$$K_J = 2e/h = 4.835\,976(12) \times 10^{14}\,\mathrm{Hz/V_{NIST-69}} \tag{3.3}$$

where $\mathrm{V_{NIST-69}}$ is the unit of voltage maintained at the US National Institute of Standards and Technology (NIST) in 1969[1]. We mention in passing that symbols for units are in upright Roman type, and in an italic font for quantities.

Their value was rapidly confirmed by the 36-GHz measurements of Petley and Morris (1970) at the NPL after due allowance was made for the different voltage units maintained in the two countries. It was swiftly confirmed by other national standards laboratories, notably in Australia and Germany. At that time the maintained volt was based on the mean electromotive force (emf) of a group of electrochemical standard cells. These had the disadvantage that their emf depends on temperature and other factors. The JEs enabled provided a known constant voltage and the existence of this enabled the technology of standard cell manufacture to be improved rapidly. In 1972, the Consultative Committee on Electricity (CCE) recommended a value for K_{J-72} for volt maintenance purposes. This value has since been revised as K_{J-90} and provides an internationally agreed representation of the SI volt.

---

[1] A note on uncertainties. In Equation 3.3, the number in parentheses represents the one-standard-deviation uncertainty in the last digits of the quoted value. It should also be noted that the uncertainty is a quasi-statistical combination of essentially two types of uncertainty, essentially one that represents the fluctuations of the measured value during the measurement and the second from quantities, which may be in error but have not varied during the measurement. Currently, these are characterized by the terms Type A and Type B, respectively, and have built on the earlier concept of random and systematic uncertainties (fluctuations in the measured value may well not be truly random and part of the art of this type of measurement lies in identifying as many as possible of these causes (e.g., someone using a hammer, switching off equipment, or thermostats, etc.). International agreement has been made on how to evaluate such uncertainties, see the document known as the

GUM : Guide to the expression of the Uncertainty of Measurement JCGM (1995). ISO/IEC Guide 98:1995 Guide to the expression of uncertainty in measurement (GUM). Geneva, Switzerland, International Organization for Standardisation (International Organization for Standardisation) (see also, e.g., *http://www.bipm.org/en/committees/jc/jcgm/wg1_bibliography.html* and *http://physics.nist.gov/Pubs/guidelines/TN1297/tn1297s.pdf*. Interestingly, uncertainty estimates are made by theoreticians as well as experimenters and it has also become essential to give an uncertainty for the numerical results of theoretical calculations.

The difference between K_{J-72} when expressed in SI units (hertz per volt) and the previous best value, namely, that resulting from the 1963 least-squares adjustment of the fundamental constants (Cohen and DuMond, 1965), was (35 ± 10) parts per million. As discussed later, the cause of this rather severe discrepancy was the use in the 1963 adjustment of a value of α derived from the early 1950s measurement of the fine-structure splitting of the deuterium atom (energy difference between the $2^2\mathrm{P}_{3/2}$ and $2^2\mathrm{P}_{1/2}$ atomic energy levels), which subsequently turned out to be incorrect.

Because QED, the relativistic quantum field theory that describes the interaction of charged particles and electromagnetic radiation or photons, is one of the most precise and important theories of modern physics (Kinoshita, 1990a), and involves knowledge of the fine-structure constant, it is essential to see how well the topic can withstand stringent experimental tests. The Josephson $2e/h$ determination is therefore significant for QED mainly because a reliable indirect value of QED-independent value of α can be derived from it, and this value can in turn be used to compare QED and experiment critically and unambiguously. This is in marked contrast to the situation that existed prior to 1967 when no such value was available and tests of QED were mainly checks of internal consistency.

The equation relating α to $2e/h$ may be written as

$$\alpha = \left[\frac{4R_\infty(\Omega_{\mathrm{LAB}}/\Omega)\gamma_p'(lo)_{\mathrm{LAB}}}{c(\mu_p'/\mu_B)(2e/h)_{\mathrm{LAB}}}\right]^{1/2} \tag{3.4}$$

where, as before, $\Omega_{\mathrm{LAB}}/\Omega$ is the ratio of the laboratory as-maintained ohm to the SI ohm; μ_p'/μ_{B} is the magnetic moment of the proton in units of the Bohr magneton (throughout, the prime means that the protons are not free but are in a spherical sample of pure water at 25°C); and $\gamma_p'(\mathrm{lo})_{\mathrm{LAB}}$ is the gyromagnetic ratio of the proton obtained by the so-called low (magnetic) field method and is measured in terms of the laboratory as-maintained unit of current, $A_{\mathrm{LAB}} = V_{\mathrm{LAB}}/\Omega_{\mathrm{LAB}}$. (Note that the more general subscript LAB (for laboratory) may be replaced by the subscript NIST, say, for measurements made at the US national standards' laboratory); and that $\gamma_p' = \omega_p'/B$, where ω_p' is the proton nuclear magnetic resonance or spin-flip angular frequency in the magnetic field (magnetic flux density) B. In the low-field method, B is established via a precision solenoid of known dimensions carrying a current known in terms of A_{LAB}.) The value of α derived from Equation 3.3 in 1969 (Taylor, Parker, and Langenberg, 1969) using the best data available was

$$\alpha^{-1} = 137.036\,08(26) \tag{3.5}$$

Included among the quantities that require an accurate value of α for comparing theory and experiment are the g-factors of the electron and muon; the energy levels in hydrogen-like atoms, especially the $n = 2$ Lamb shift ($2^2\mathrm{S}_{1/2} - 2^2\mathrm{P}_{1/2}$ splitting); and the ground-state hyperfine splitting in hydrogen, muonium (an electron bound to a positive muon or μ^+e^- atom), and positronium (a bound electron–positron pair or e^+e^- atom).

Of particular interest in the late 1960s was the hydrogen hyperfine splitting ν_{Hhfs}. This transition frequency, which corresponds to essentially the energy difference between a hydrogen atom in which the electron and proton spins are aligned and one in which they are in opposite directions, has been determined experimentally with the extraordinary accuracy of 1 part in 10^{12} using the hydrogen maser. The theoretical QED equation for ν_{Hhfs} is

$$\begin{aligned}\nu_{\mathrm{Hhfs}}(\mathrm{theor}) = (16/3)\, cR_\infty \left(\mu_p/\mu_B\right) \\ \times \left(1 + m_e/m_p\right)^{-3} \\ \times \alpha^2 \left(1 + \mathrm{h.o.c.}\right) \qquad (3.6)\end{aligned}$$

where h.o.c. are higher order corrections.

In contrast to the 1×10^{-12} relative uncertainty of $\nu_{\mathrm{Hhfs}}(\mathrm{expt})$, Equation 3.6 is limited to an accuracy of a few parts in a million because of the difficulty in calculating some of the higher order corrections. (Even in the late 1960s, c, R_∞, μ_p/μ_B, and m_e/m_p were sufficiently well known for their total contribution to the fractional uncertainty of $\nu_{Hhfs}(\mathrm{theor})$ to be considerably less than 10^{-6}.)

The most uncertain higher order correction to Equation 3.6 was $\delta_N^{(2)}$, the proton polarizability contribution, which arises from the various excited states or internal structure of the proton. In the late 1960s, theoretical calculations predicted $\delta_N^{(2)} = (0 \pm 5) \times 10^{-6}$. This was in conflict with what was implied by the value of α accepted at that time, which, as noted earlier, was derived from a measurement of the fine-structure splitting in deuterium. The deuterium value of α was used to calculate $\nu_{Hhfs}(\mathrm{theor})$ from Equation 3.6, when this result was compared to $\nu_{Hhfs}(\mathrm{expt})$ as obtained from the hydrogen maser and, when their difference was assumed to arise solely from the existence of a polarizability term, it was found that $\delta_N^{(2)} = (43 \pm 9) \times 10^{-6}$. This meant that the probability for $\delta_N^{(2)}$ to be as small as predicted by direct calculation was only 1 in 20 000, an obvious inconsistency. In contrast, when the value of α derived from the JE measurement of $2e/h$, Equation 3.5, was used in place of the deuterium value, it was found that $\delta_N^{(2)} = (2.5 \pm 9) \times 10^{-6}$, in keeping with the theoretical calculations.

Hence the JE value of α removed a discrepancy that during the 1960s was termed one of the major unsolved problems of QED. This example further illustrates the overall unity of science–a low-temperature, solid-state physics experiment has provided information about the excited states of the proton–as well as how precise measurements of the fundamental constants can illuminate apparent inconsistencies in our physical description of nature. It also illustrates why one of the most accurately measured quantities, the hydrogen maser frequency, does not contribute to our knowledge of fundamental constants other than the polarizability of the proton.

The other aspect of Equation 3.4 lies in the two methods of measuring the gyromagnetic ratio of the proton that are characterized by the terms $\gamma_p'\left(\mathrm{lo}\right)_{\mathrm{LAB}}$ and $\gamma_p'\left(\mathrm{hi}\right)_{\mathrm{LAB}}$, respectively.

In the low-field determination, the spin precession frequency is measured in a weak magnetic field, such as is produced by passing a current $i(\text{lo})_{\text{LAB}}$ through a precisely wound helical solenoid and in the high-field method the flux density is "measured" by weighing the force mg on a precisely wound rectangular that is suspended between the poles of an electromagnet, carrying a current $i(\text{hi})_{\text{LAB}}$, where m is the mass required to balance the magnetic force when the local gravitational acceleration is g. Taking both currents as having measured in terms of the same laboratory units one may combine the resulting two equations in two ways, one to eliminate the currents to obtain γ_p' and the other to eliminate γ_p' and hence obtain a realization of the SI definition of the ampere, also remembering that

$$\begin{aligned}&\gamma_p' B(\text{lo}) = \gamma_p'(\text{lo})_{\text{LAB}} B(\text{lo})_{\text{LAB}} \quad \text{and}\\ &\gamma_p'(\text{hi}) B(\text{hi})_{\text{LAB}} = \gamma_p B(\text{hi}) \qquad (3.7)\end{aligned}$$

Much of the difficulty of either method came from the need to measure the coil dimensions very accurately, yet the dimensions of γ_p' are coulombs per kilogram. This led Kibble (1975) to wonder whether there might be a way of adapting and combining the two methods into a realization of both quantities (Flowers and Petley, 2001). This led him to propose the method of realization of the electrical watt, which is the method behind the present best measure of determining h and the proposed new definition of the kilogram that is discussed in Section 3.5.

3.2.4
QED and the Quantum Hall Effect Determination of h/e^2

Another example of a critical test of QED, is one that is still under active experimental and theoretical investigation at present, this involves the QHE and the magnetic moment anomaly of the electron $a_e = g_e/2 - 1 = \mu_e/\mu_B - 1$, where g_e is the g-factor of the electron.

The current best experimental value for $a_e = (g_e-2)/2$, is that by Gabrielse *et al.* ([2006, 2007]) at Harvard using a single electron stored in a Penning-type ion trap. This work achieved a formidable fractional accuracy of 6.6 parts in 10^{10} and this is about six times smaller than an earlier Penning trap measurement at the University of Washington by Van Dyck, Schwinberg, and Dehmelt (1987).

$$a_e = 1.159\,652\,85(76) \times 10^{-3} \qquad (3.8)$$

The QED theoretical expression for a_e may be written as (Kinoshita and Lindquist, 1990b)

$$\begin{aligned} a_e(theor) &= A_1^{(2)}(\alpha/\pi) + A_1^{(4)}(\alpha/\pi)^2\\ &\quad + A_1^{(6)}(\alpha/\pi)^3 + A_1^{(8)}(\alpha/\pi)^4\\ &\quad + \ldots + \delta a_e A_1^{(2)} = 1/2\\ A_1^{(4)} &= -0.328478965579\ldots\\ A_1^{(6)} &= 1.181241456\\ A_1^{(8)} &= -1.7283(35) \qquad (3.9)\end{aligned}$$

The total uncertainty in a_e(theor) assuming α is exactly known is now $2.4 \times 10^{-10}\ a_e$, with the dominant uncertainty being that ascribed to the tenth order term, $A_1^{(10)} = 0.0(3.7)$. A test of QED, by comparing a_e(theor) with a_e(expt), to the full accuracies of a_e(theor) and a_e(expt) therefore requires a measured value of α having an uncertainty at or significantly less than $2.4 \times 10^{-10}\ a_e$. Such a value is not yet available, but one close to this has been obtained from the Harvard (Gabrielse *et al.*, 2006) and Washington (Van Dyck, Schwinberg, and Dehmelt, 1987) values for the fine-structure constant that are

$$\alpha^{-1} = 137.035999711(96) \quad \text{and}$$
$$\alpha^{-1} = 137.03599883(50) \qquad (3.10)$$

respectively.

The fractional uncertainty contribution from the QED theory to the conversion from a_e to α is estimated to be about 2.4×10^{-10}–a triumph of achievement for both theory and experiment.

The theory and measurement of the muon anomaly at Brookhaven, (Brown *et al.*, 2001) has similarly led to a measurement of a value of α^{-1}, denoted α_μ^{-1} of

$$\alpha_\mu^{-1} = 137.03567(26) \qquad (3.11)$$

which provides valuable confirmation of our understanding of the muon as an elementary particle. The hyperfine splitting in muonium has also been measured, and this can lead to a value for α^{-1}, of 137.036 0017(80).

Other tests of QED have been made by comparing the spin precession and cyclotron frequencies of hydrogen-like heavier atoms (highly ionized atoms with one electron remaining). Measurements of the g-factor ratio of hydrogen-like carbon and oxygen have been made at GSI Darmstadt and the University of Mainz (Werth *et al.*, 2002) by comparing the ratios of the spin precession and cyclotron frequencies in a Penning trap an experimental value of

$$\frac{g_g - (^{12}\mathrm{C}^{5+})}{g_e - (^{12}\mathrm{O}^{7+})} = 1.000497273218(41) \qquad (3.12)$$

and similarly a double Penning trap measurement of the ratio of the spin precession to the cyclotron frequency of hydrogen-like $^{16}\mathrm{O}^{7+}$ (that is oxygen from which all but one of its electrons have been removed) gave

$$\frac{f_s(^{16}\mathrm{O}^{7+})}{f_c(^{16}\mathrm{O}^{7+})} = 1.000497273218(41) \qquad (3.13)$$

Both of these are also in good agreement with some formidable QED calculations.

The QHE is a behavior characteristic of a completely quantized two-dimensional electron gas (Prange and Girvin, 1990; Yoshioka, 2002). Such a gas may be realized in a high-mobility semiconductor device such as a silicon metal-oxide-semiconductor field-effect transistor (MOSFET), or a device made from gallium–aluminum–arsenide as a heterojunction of standard Hall bar geometry in an applied magnetic flux density of the order of 10 T and cooled to about 1 K. Hence, in common with the JE, the QHE is a low-temperature, solid-state physics phenomenon.

It is found that for a fixed current I (typically 10–50 μA) through the device, there are regions in the curve of Hall voltage U_H, versus gate voltage for a MOSFET, or of U_H versus B, for a heterojunction, where U_H remains constant as either the gate voltage or B is varied. These regions of constant U_H are termed *quantum Hall plateaus.*Inthe limit of zero dissipation (zero voltage drop) in the direction of current flow, the Hall voltage-to-current quotient $U_H(\mathrm{i})/I$ or Hall resistance $R_H(\mathrm{i})$ of the ith plateau, where i is an integer (we consider only the integral QHE), is quantized and given by

$$R_H(i) = U_H(i)/I = R_K/i \qquad (3.14)$$

where R_K is the von Klitzing constant (named in honor of Klaus von Klitzing, who discovered the QHE in 1980 and for which he was awarded a Nobel Prize in 1985).

The theory of his effect predicts that

$$R_{\mathrm{K}} = \frac{h}{e^2} = \frac{\mu_0 c}{2\alpha} \tag{3.15}$$

Measurements show consistency with Equation 3.9 within a few parts in 10^{10}.

Since in the present SI, $\mu_0 = 4\pi \times 10^{-7}$ N A^{-2} exactly, and $c = 299792458\,\mathrm{m\,s}^{-1}$ exactly as a result of the 1983 redefinition of the meter in terms of the speed of light, it is apparent from (Equation 3.10) that a measurement of R_K in SI units (i.e., ohms) with a given uncertainty will yield a value of the fine-structure constant α with the same uncertainty.

In practice, R_K is measured in terms of Ω_{LAB} (indicated as $(R_K)_{\mathrm{LAB}}$) so that the conversion factor $\Omega_{\mathrm{LAB}}/\Omega$ must be determined in a separate experiment notably by using an apparatus known as a *calculable cross-capacitor* in which the unknown resistance of a reference resistor is compared with the known impedance of the capacitor. Nevertheless, the QHE has already yielded a value of α with an uncertainty of 2.4 parts in 10^8 (Shields, Dziuba, and Layer, 1989). When used to compare a_e(theor) with a_e(expt), it gives a difference of a_e(expt)$-a_e$(theor) $= a_e = (4.1 \pm 2.5) \times 10^{-8}$. Since the 4.1×10^{-8} difference is only 1.7 times the 2.5×10^{-8} standard deviation of the difference, the difference between theory and experiment a_e is within statistically acceptable limits. Since QED actually predicts the value of the entire electron g-factor, then a consequence of the near equality of $g_e(\mathrm{expt}) = 2(1 + a_e(\mathrm{expt}))$ and $g_e(\mathrm{theor}) = 2(1 + a_e(\mathrm{theor}))$, is that this agreement confirms the validity of the renormalization processes inherent in QED theory to the hitherto unprecedented level of four parts in 10^{11}.

3.2.5
The Quantum Metrology Triangle

The JE provides a measurement of $2e/h$ in terms of voltage, V, and similarly the quantized Hall resistance a measurement of h/e^2 in terms of resistance. There is an interesting third measurement that can be obtained from single electron tunneling (SET), which provides a measurement of the elementary charge e in terms of a current, I, and a driving frequency ν giving three relations with two unknowns.

The measurement of e by SET has been implemented on the charging of an isolated capacitor by the transfer of single electrons by the application of an alternating voltage at close to the maximum that a stochastic frequency can be counted; if this is 10 GHz, then one has a charging current of about 1.6 nA on a 10^{-16} F capacitor–a rather small current to measure with a fractional accuracy of about 10^{-6} or better. However, Mark Keller and his colleagues have achieved this accuracy at NIST in the United States (Keller, Zimmerman, and Eichenberger, 2007). This has excited considerable interest throughout the world and work is presently in progress in at least four national institutes: in France, Germany, United Kingdom, and United States. To date, this leg of the triangle has been verified to about a fractional part in 10^6 and is hoped to improve this accuracy by a further two orders of magnitude.

The closure provided by combining the three types of measurement has become known as closing the quantum metrology triangle. That is, if sufficient accuracy can be achieved with the SET work, the three measurements should provide an internally consistent set and thereby provide a valuable check of the underlying theory.

The 2006 evaluation showed that the JE $2e/h$ leg could be verified by internal consistency involving the moving coil watt determinations to a fractional uncertainty of about $\varepsilon_J = (-77 \pm 80) \times 10^{-9}$. There is some doubt over the reliability of this value due to the small amount of data and it is suggested that the agreement may be partly fortuitous. Similarly, the QHE leg could be verified to an uncertainty $\varepsilon_K = (20 \pm 18) \times 10^{-9}$ by both external and internal consistency. The SET method at present limits the accuracy and, to date, the triangle has been shown to provide closure with a fractional uncertainty at the 10^{-6} level (Gallop, 2005; Keller, 2008).

3.3
Obtaining "Best Values" of the Constants

As science and technology advances, so too our ability to measure particular combinations of the fundamental constants improve in accuracy. The particular combinations of the constants that can be best measured also change.

3.3.1
Relationships among the Constants

Equations 3.4 and 3.7 together with Equations 3.8 and 3.10 can provide values of α. However, because, in practice, a value of $\text{unit}_{\text{LAB}}/\text{unit}_{\text{SI}}$ is required to obtain α from Equation 3.10 as well as from Equation 3.4, the two values will not be completely independent. Moreover, $2e/h$ (in SI units) can be obtained from the relation

$$\frac{2e}{h} = \left[\frac{16R_\infty\left(m_p/m_e\right)N_\text{A}}{\mu_0 c^2 M_p \alpha}\right] \tag{3.16}$$

where M_p is the molar mass (atomic weight in units of kilograms per mole) of the proton and $\gamma_p'(\text{lo})$ from the equation

$$\gamma_p'\left(\text{lo}\right)_{\text{LAB}} = \left(\frac{F_{\text{LAB}}K_{A,\text{LAB}}^2\left(\mu_p'/\mu_B\right)\left(m_p/m_e\right)}{M_p}\right) \tag{3.17}$$

where F_{LAB} (Faraday constant) is measured in terms of the as-maintained ampere A_{LAB} in a particular laboratory, and

$$\begin{aligned} K_{A,\text{LAB}} &= A_{\text{LAB}}/\text{A} = (V_{\text{LAB}}/\Omega_{\text{LAB}})/(\text{V}/\Omega) \\ &= (V_{\text{LAB}}/\text{V})/(\Omega_{\text{LAB}}/\Omega) \\ &= K_{V-\text{LAB}}/K_{\Omega-\text{LAB}} \end{aligned} \tag{3.18}$$

Further, the Faraday constant is related to the Avogadro constant by $F = N_A e$, and $K_{A,\text{LAB}}$ may be obtained from direct measurements using a so-called ampere force balance as well as from the equation

$$K_{A,\text{LAB}}^2 = \frac{\gamma_p'\left(\text{lo}\right)_{\text{LAB}}}{\gamma_p'\left(\text{hi}\right)_{\text{LAB}}} \tag{3.19}$$

where $\gamma_p'\left(\text{hi}\right)_{\text{LAB}}$ is the gyromagnetic ratio of the proton as obtained by the so-called high (magnetic) field method. (The magnetic flux density B is determined by measuring the force it exerts on a conductor of known dimensions carrying a current known in terms of A_{LAB}.)

It is apparent that complex relationships can exist among groups of fundamental constants and conversion factors, and that a particular constant may be obtained either by direct measurement or indirectly by appropriately combining other directly measured constants. If the direct and indirect values have similar accuracy, both must be considered in order to arrive at a best value for the quantity. However, each of the various routes that can be followed to a particular constant, both direct and indirect, provides a somewhat different numerical value. The best way to treat this is by using the mathematical technique of least-squares, a technique that is backed by the rigor of statistics.

Least-Squares Adjustments of the Constants

The least-squares method furnishes a well-defined procedure for calculating best ‘‘compromise’’ values of the constants from all of the available data. It takes into account all possible routes to a particular constant and yields a single value for each constant by weighting the different routes according to their relative uncertainties. The weights themselves are obtained from the *a priori* uncertainties assigned to the individual direct measurements that constitute the original set of data.

The least-squares study of the constants was pioneered in the late 1920s by Birge (1929) and continued by others, notably DuMond and Cohen (1953) and Cohen and DuMond (1965). More recent reviews were those of Taylor, Parker, and Langenberg (1969), based on their JEs determination of $2e/h$; and of Cohen and Taylor ([1973, 1987]); and more recently by Mohr and Taylor ([1999, 2005], and Mohr, Taylor, and Newell (2008). Since 1973, reviews have usually been undertaken under the auspices of CODATA. The 2006 adjustment is the most recent one currently available, and the recommended set of constants resulting from it, are those officially adopted by CODATA for international use. Subsequent evaluations are scheduled to take place at four-year intervals and the values will be posted on the NIST Web Site at *http://physics.nist.gov/cuu/Constants/*.

3.3.2
The 2006 CODATA Adjustment

An important guiding principle for the wide acceptance of the evaluations for scientists is

1. to be clear how they are performed;
2. what data is included; and
3. the weight given to a particular input datum.

The standard statistical method gives weights to the input of the data according to the inverse square of its associated uncertainty. Common sense, backed by the history of the subject, supports the reality that there may be some unknown science that is as likely to be affecting the most accurate value as those of lesser accuracy. Indeed, history suggests that it would have been more realistic to double the uncertainty associated with the output values. This option has so far been rejected on the grounds that it might slow the pace of the subject because it could mask the recognition that there were unanticipated problems or unknown science associated with particular measurements.

An inherent consequence of applying the least-squares method is that a determination with an uncertainty that is three or more times greater than that of the smallest, has less than one-ninth of the impact on the weighted mean value than does the most accurate result. This puts very strong pressure on the ability of scientists to estimate their uncertainties. Therefore, it is usual to check the effect on the evaluated values of omitting the result with the smallest uncertainty. Similarly, one can test for the effect of all data pertaining to the QHE (say) or the JE and thereby verify the extent to which these phenomena do represent a measurement of the particular combination of the fundamental constants

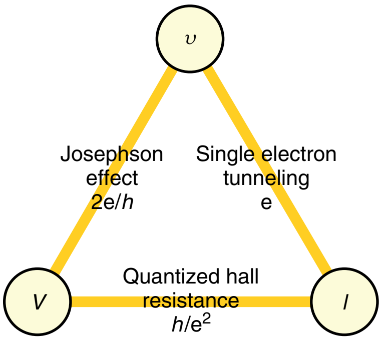

Fig. 3.1 The quantum metrology triangle links the Josephson effect, the quantum Hall effect and single electron tunneling, the sides of the triangle, to the electrical units and frequency, represented by the vertices, by using Ohm's law and tests the self-consistency of the expressions that are assumed in Equations 3.2 and 3.10.

concerned. The preliminary assessment of the input data is the most critical part of an evaluation because the computing thereafter is largely automatic–so the prudent approach to the rejection of "rogue" data suggested by the computer adage "garbage in: garbage out" can only be followed by subjective decisions by the reviewers backed by advice from the CODATA Task Group and other expert scientists.

If sufficient data on a particular topic are available, one may use a statistical quantity termed the *Birge ratio*,which amounts to comparing the value of χ^2 with the number of degrees of freedom for the group and expanding the assigned uncertainty to make the two values equal. Thus, in the 2006 evaluation, the determinations of G were analyzed in this way and the output uncertainty expanded by a factor of 4.26, and this is illustrated in Figure 3.2.

The present leading experimental values for G are also given in Table 3.2. It will be noted that the uncertainty in the CODATA recommended value was conservatively increased above that obtainable by statistical weighting of the results because it was judged that there were significant discrepancies among the published measurements.

Unfortunately, one of the major problems for the measurements of G result from the small size of the gravitational force between objects in the presence of the 10^9 times greater gravitational attraction of the Earth, or spurious electromagnetic fields. This places formidable requirement for the suspension, which must have a small and stable torsional constant combined with adequate strength and very good shielding from outside effects (including sidereal heating of the laboratory, or gravitational gradients). Consequently, the achievable experimental accuracy has not improved as rapidly as that of other constants and has only improved about a 100-fold over two centuries. Thus, the measurements of Cavendish (1798) led to $G = 6.75(5) \times 10^{-11}\ \mathrm{m^2\ s^{-2}\ kg^{-1}}$ and this may be compared with the CODATA 2006 recommended values of $6.67428(67) \times 10^{-11}\ \mathrm{m^2\ s^{-2}\ kg^{-1}}$.

In earlier evaluations, the data entering a least-squares adjustment of the constants were divided into two groups:

1. the more precise data or auxiliary constants and
2. the stochastic input data.

The auxiliary constants used in the 1986 adjustment, none of which had an assigned fractional uncertainty greater than 2×10^{-8}, included μ_0, c, R_∞, M_p, m_p/m_e, g_e, μ_e/μ_p, and the g-factor of the muon:

$$g_\mu = 2(1 + a_\mu) = 2\mu_\mu/(eh/4\pi m_\mu) \quad (3.20)$$

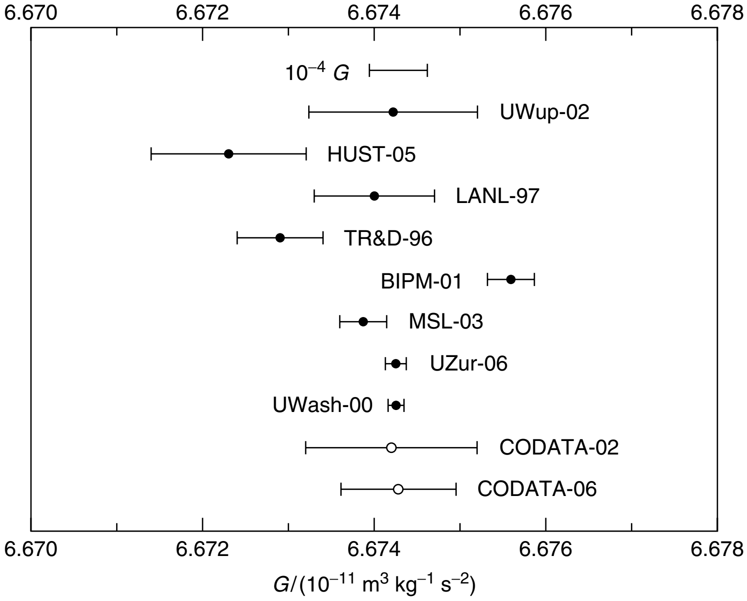

Fig. 3.2 Recent values of the gravitational constant. (Taken from CODATA 2006, see Mohr, Taylor, and Newell (2008) for details and designations.)

where a_μ is the magnetic moment anomaly of the muon and μ_μ is its magnetic moment.

Also included in the auxiliary constant category in 1986 were the representations of the volt, ohm, and ampere, expressed in terms of the common set of electrical units known as the *BIPM volt, ampere, ohm*, and the 1972 internationally agreed value for the Josephson constant $K_{J\text{-}72}$. In 1990, at the suggestion of the CCE, the latter was revised to the present internationally agreed value of K_J and assigned the designation $K_{J\text{-}90}$, together with a value for the von Klitzing constant $R_{K\text{-}90}$. The best values for K_J and R_K are obtained from the evaluations.

In the 1986 evaluation, the stochastic data were grouped into sets providing information on five constants. Since then, recognizing the greatly improved sophistication of calculations provided by modern computers that have progressively replaced the slide-rules and hand-cranked mechanical calculators of the Birge era, the distinction between auxiliary constants and stochastic input data has been dropped and all the relevant data has been included in the evaluation. This advance has now allowed any correlations between the input data to be taken into account.

Since the Rydberg constant, for example, is now measured to 12 decimal places, then, for the 2006 evaluation, the computer had to be able to perform a least-squares multivariate analysis comprising some 79 unknowns from some accepted 135 input data (the latter from an initial set of 150 input data), with the resulting large number of associated matrices, evaluated to a large number of decimal places. Birge could have scarcely imagined what scientific and technological progress would

Tab. 3.2 Data contributing to the CODATA 2006 value of the Newtonian gravitational constant G.

Source (see Mohr, Taylor, and Newell (2008), and references therein)	**CODATA 2006 Identification**	**Method**	**Value 10^{11} Gm^{-3} kg^{-1} s^{-2}**	**Relative standard uncertainty u_r**
2002 CODATA Adjustment	CODATA-02	–	6.6742(10)	1.5×10^{-4}
Karagioz and Izmailov (1996)	TR&D-96	Fiber torsion balance, dynamic mode	6.672 9(5)	7.5×10^{-5}
Bagley and Luther (1997)	LANL-97	Fiber torsion balance, dynamic mode	6.674 0(7)	1.0×10^{-4}
Gundlach and Merkowitz (2000), Gundlach and Merkowitz (2002)	UWash-00	Fiber torsion balance, dynamic compensation	6.674 255(92)	1.4×10^{-5}
Quinn *et al.* (2001)	BIPM-01	Strip torsion balance, compensation mode, static deflection	6.675 59(27)	4.0×10^{-5}
Kleinevos (2002), Kleinvos *et al.* (2002)	UWup-02	Suspended body, displacement	6.674 22(98)	1.5×10^{-4}
Armstrong and Fitzgerald (2003)	MSL-03	Strip torsion balance, compensation mode	6.673 87(27)	4.0×10^{-5}
Hu *et al.* (2005)	HUST-05	Fiber torsion balance, dynamic mode	6.672 3(9)	1.3×10^{-4}
Schlamminger *et al.* (2006)	UZur-06	Stationary body, weight change	6.674 25(12)	1.9×10^{-5}
2006 CODATA Adjustment	CODATA-06		6.674 28(67)	1.0×10^{-4}

Taken from Table 27 of CODATA 2006, see Mohr, Taylor, and Newell (2008) for details.

accomplish within 70 years of his seminal work–but he would surely have approved.

Although the output values have associated uncertainties, the least-squares analysis results on correlations between the various output values and, for some purposes, the user may need to use the associated variance and covariance matrices that are also available from the NIST web site. Thus, the recommended uncertainty of the quotient e/h may differ from root sum of the squares of the uncertainties assigned to e or h.

An equation that relates a stochastic input datum to the M adjustable constants (with the aid of the auxiliary constants if necessary) is known as an *observational equation*.Thus, if there are n_1 items of stochastic input data, there are n_1 observational equations. However, the observational equations for data items of the same type are of the same form. The weight w_i of each stochastic datum and hence of each observational equation is related to the datum's *a priori* assigned uncertainty s_i by $w_i = 1/s_i^2$.

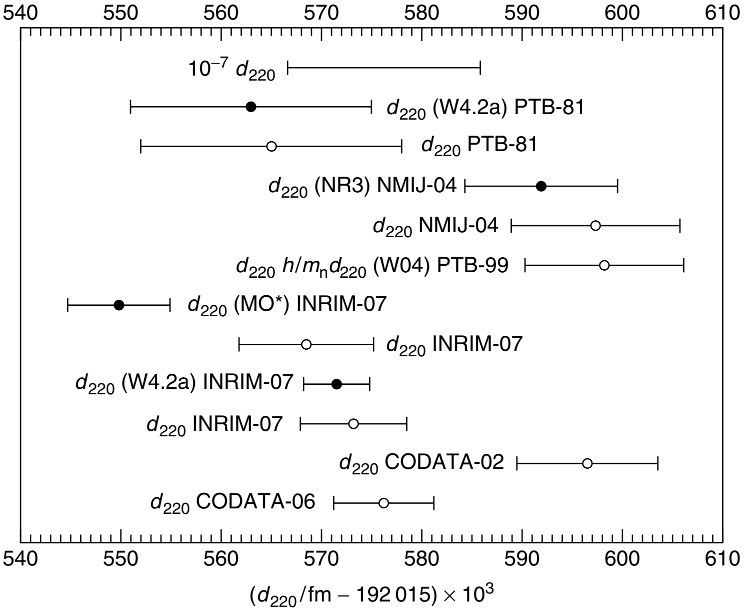

Fig. 3.3 Values of d_{220} inferred from various measurements. (Taken from CODATA 2006, see Mohr, Taylor, and Newell (2008) for details and designations.)

Consequently, if one value of a datum for a particular quantity has half the uncertainty of another, it carries four times as much weight in the adjustment. Taking into account the weight assigned to each observational equation, and with the aid of a present-day computer, the least-squares adjusted values of the M unknowns may be readily obtained. "Best values" in the least-squares sense for these M adjustable constants, with their variances and covariances, are thus the immediate output of the adjustment. The optimal values in the least-squares sense for most constants of interest not directly subject to adjustment may then be calculated from the M adjustable constants. In contrast to the auxiliary constants of the 1986 adjustment, for which no uncertainty exceeded 2×10^8, the fractional uncertainties assigned to the 138 items of stochastic input data initially considered in the 2006 adjustment were broadly in the range 10^{-12}–10^{-6}.

Because the d_{220} data were in such disagreement, as seen in Figure 3.3, a separate working group has been set up to coordinate the work in several countries in order to produce recommended value(s) for d_{220} and V_m(Si). The work in progress includes determination(s) of N_A resulting from measurements made with a sample of isotopically pure silicon produced in Russia (Becker, 2003; Becker *et al.*, 2006).

Because new results that can influence a least-squares adjustment of the constants are reported continually, it is always difficult to choose an optimal time at which to carry out a new adjustment and to revise the recommended values of the constants. In the case of the 2006 adjustment, all data available up to 1 January 2006 were considered for inclusion.

3.3.2.1 Uncertainty Assignment and Discrepant Data

Because the weight of each stochastic datum included in a least-squares adjustment of the constants is taken as $w_i = 1/s_i^2$, critical analysis of the input data and deciding what *a priori* uncertainty s_i should be assigned to each measurement is the main problem in successfully carrying out an adjustment. Modern data acquisition techniques mean that instead of making tens of observations one has thousands, but without sophisticated analysis and intelligent investigation of possible variables (e.g., by taking advantage of the sophistication provided by factorial replicate and other statistical analysis techniques), the apparently improved reduction in the uncertainty estimate may prove highly illusory. The uncertainty problem is made especially difficult because in most experiments sufficient data are taken to reduce the Type A uncertainties (formerly termed *random*effects) to negligible amounts, and the dominant uncertainty assigned to the experiment is primarily from estimates of the Type B uncertainties associated with the effects that remain substantially constant during the period of the measurement. Because such effects frequently arise from a phenomenon that the experimenter is not so familiar with, the quantitative estimation of the uncertainties is subjective and is necessarily based on "educated" guesses.

As shown above with the measurements of G, discussed above, another problem area is that of deciding how to handle "discrepant" data. These comprise measurements that differ from each other by statistically significant amounts in comparison with their assigned uncertainties s_i. Such data cannot be included in an adjustment uncritically because the inconsistencies imply either incorrect uncertainty assignments or the presence of unknown measurement errors. When confronted with such a situation, the reviewer can in general either include the inconsistent data, but only after expanding (increasing) their s_i so that they are no longer discrepant, or decide, on as sound a theoretical and experimental basis as possible, which of the inconsistent data are least reliable, and discard them. These two approaches clearly imply that different reviewers might treat the same data differently, thereby obtaining somewhat different sets of best values.

Motivated by the all-too-frequent occurrence of large changes in the recommended values of many constants from one adjustment to the next, new algorithms for handling discrepant data in least-squares adjustments of the constants were developed during the years between the 1973 and 1986 adjustments. These added increased objectivity to expanding *a priori* uncertainties and discarding inconsistent data were used extensively by Cohen and Taylor in carrying out the 1986 adjustment and, with their aid, the initial group of 38 items of stochastic input data was reduced to 22 items by deleting those that were either highly inconsistent with the remaining data or had assigned uncertainties so large that they carried negligible weight.

3.3.2.2 Recommended Values of the Constants

Table 3.3 gives the 2006 CODATA recommended values and the associated one-standard-deviation uncertainties ($\sim$65% probability) of a number of fundamental physical constants. As noted in Section 3.3, the direct output of a least-squares adjustment provides the best values (in the least-squares sense) for the adjustable constants, or unknowns, and their variances and covariances. For the 1986 adjustment, the five unknowns chosen for evaluation were α^{-1}, K_V,

Tab. 3.3 Comparison of the 1973, 1986, and 2006 CODATA recommended values and the assigned uncertainties of a selected group of constants implied by the data currently available.

Quantity	1973 fractional uncertainty σ_{73}	1986 fractional uncertainty σ_{86}	2006 fractional uncertainty σ_{2006}	Ratio $\sigma_{86}/\sigma_{2006}$	Fractional change δ of 2006 value from 1986 value	Change δ/σ_{86}
α^{-1}	8×10^{-7}	4.5×10^{-8}	6.8×10^{-10}	66.2	7.4×10^{-8}	1.73
E	2.9×10^{-6}	3.0×10^{-7}	2.5×10^{-8}	12	-5.3×10^{-7}	−1.75
H	5.4×10^{-6}	6.0×10^{-7}	5.0×10^{-8}	12	-9.9×10^{-7}	−1.65
m_e	5.1×10^{-6}	5.9×10^{-7}	5.0×10^{-8}	11.8	-8.3×10^{-7}	−1.40
N_A	5.1×10^{-6}	5.9×10^{-7}	5.0×10^{-8}	11.8	8.4×10^{-7}	1.43
m_p/m_e	3.8×10^{-7}	2.0×10^{-8}	4.3×10^{-10}	46.5	-1.6×10^{-8}	−0.79
F	2.8×10^{-6}	3.0×10^{-7}	2.5×10^{-8}	12.0	3.2×10^{-7}	1.07
$2e/h$, K_J	2.6×10^{-7}	3.0×10^{-7}	2.5×10^{-8}	12.0	4.6×10^{-7}	2.0
R	31×10^{-7}	8.4×10^{-6}	1.7×10^{-6}	4.9	-4.5×10^{-6}	−0.54
K	32×10^{-6}	8.5×10^{-6}	1.7×10^{-6}	5.0	-5.5×10^{-6}	−0.65
σ	125×10^{-6}	34×10^{-6}	7.0×10^{-6}	4.9	1.9×10^{-5}	−0.57
G	6.15×10^{-4}	3.4×10^{-4}	1.0×10^{-4}	3.4	2.5×10^{-4}	1.98
R_∞	7.6×10^{-8}	1.3×10^{-9}	6.6×10^{-12}	197	3.2×10^{-9}	0.70

K_Ω, d_{220}(Si), and μ_μ/μ_p. The number of "unknowns" was increased dramatically in the 2006 adjustment, However, the uncertainties of many of the output values were still largely determined by a comparatively few different types of measurements.

3.3.2.3 Comparison of the 1986 and 2006 Adjustments

The 1986 CODATA adjustment represented a major advance over its 1973 counterpart; the uncertainties of many recommended values were reduced by roughly an order of magnitude, and further progress has been made through advances made throughout the precision measurement-fundamental constants field between 1986 and 2006. This is apparent from column 4 of Table 3.3, which compares the recommended values and uncertainties of a selected group of constants as obtained from the two adjustments and from Figure 3.4 a plot of the changes in uncertainty tabulated in columns 1–3. It is also clear from the last column of Table 3.3 that unexpectedly large changes have occurred in the 1973 recommended values of several of these constants (there are changes that are large relative to the uncertainty assigned to their 1973 value). Many of these changes are a direct consequence of the 7.8×10^{-6} fractional decrease in the quantity K_V from 1973 to 1986, and the very high correlation between K_V and the calculated values of e, h, m_e, N_A, and F. Since $2e/h = K_{J-72}/K_V$, the 1986 value of K_V also implies that the value of $2e/h$ that was adopted by the CCE in 1972, (which was believed to be consistent with the SI value, and which most national standards laboratories adopted to define and maintain their national representation of the volt), was actually 7.59 μV smaller than the correctly realized SI value. As discussed

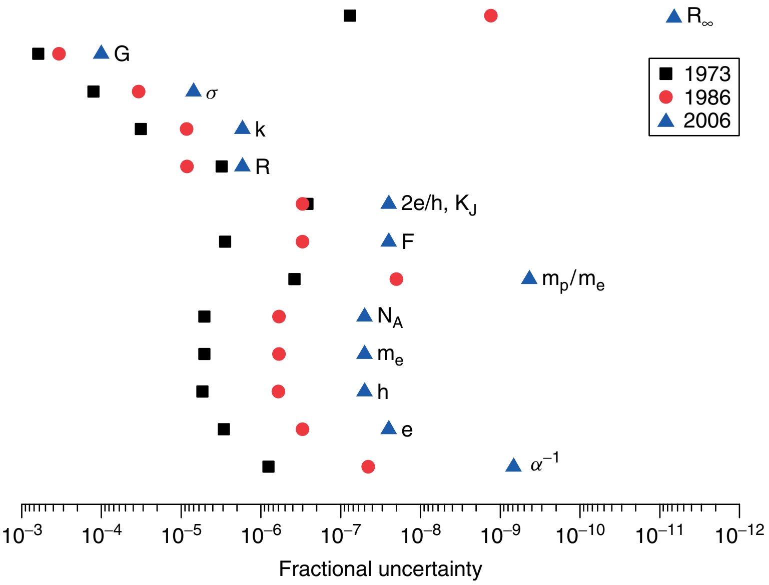

Fig. 3.4 Improvements in the fractional uncertainty of some selected constants from 1973 to 2006.

in Section 3.4, this unsatisfactory situation was rectified adopting the revised value for $2e/h$ with effect from 1 January 1990.

The large change in K_V between 1973 and 1986 would have been avoided if two determinations of F that seemed to be discrepant with the remaining data had not been deleted in the 1973 adjustment. In retrospect, the disagreement was comparatively mild. Fortunately, there are fewer similar disagreements in the 2006 adjustment; the measurements that were deleted were so discrepant, or of such low weight that if retained they would have a negligible effect on the adjusted values. Thus, it is unlikely that any reasonable alternative evaluation of the data considered in the 2006 least-squares adjustment could lead to significant changes in the 2006 recommended values. It is interesting that the spectacular improvement in the experimental and theoretical aspects relating to the accuracy of the Rydberg constant, largely as a result of the development of the ability to make direct frequency multiplication measurements through to the visible region and beyond (de Beauvoir *et al.*, 2000; Hänsch *et al.*, 2005), was not significantly changed when expressed in terms of the fractional uncertainty assigned to the 1986 value. None of the 1986 recommended values in Table 3.3 has been changed by more than two standard deviations.

3.4
Josephson and Quantum Hall Effect Representations of the Volt and Ohm

As pointed out in Section 3.2.2, the fundamental constants have a key role to play in the establishment of reproducible and invariant units of measurement. Exemplifying this role was the worldwide introduction on 1 January 1990 of new practical units of voltage and resistance based on standardized values for the

Josephson constant and quantum Hall constant, in the sections that follow. We consider first some aspects of the SI.

3.4.1
The SI and Representations of the Volt and Ohm

The SI is the system of units used throughout the world for expressing the results of physical and chemical measurements. The three SI base mechanical units are the meter (m), kilogram (kg), and second (s) (Taylor, 1991b). The derived SI unit of force, the newton (N), is then defined in terms of these three base mechanical units by $1N = 1\,\mathrm{m\,kg\,s^{-2}}$. The SI base unit of current, the ampere (A), is defined as that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross section, and placed 1 m apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} N per meter of length. This definition leads to $\mu_0 = 4\pi \times 10^{-7}\ \mathrm{N\,A^{-2}}$ exactly.

The SI derived unit of electric potential difference, the volt (V), and the SI derived unit of electric resistance, the ohm (Ω), are then defined in terms of the three base mechanical units and the base electric unit, the ampere. Formally, $1V = 1\,\mathrm{m\,kg\,s^{-3}\,A^{-1}}$ and $1\ \Omega = 1\ \mathrm{m^2\,kg\,s^{-3}\,A^{-2}}$.

In practice, the volt may be realized using a force balance in which an electrostatic or electromagnetic force is compared with a mechanical force; and the ohm may be realized by comparing the impedance of a calculable capacitor or inductor with that of a resistor. However, because both experiments are exceedingly difficult and time consuming to carry out and are of limited accuracy, it has become necessary to establish practical, or as-maintained, units of voltage and resistance, that is, representations of the volt and ohm that provide better long-term repeatability and constancy than the direct realizations of the SI units themselves.

Historically, practical representations of the volt were based on the mean emf of large groups of electrochemical standard cells. However, in 1972 (Terrien, 1973; Taylor, 1987), the CCE suggested that the national standards laboratories base their national representations of the volt on the JE in order to avoid the variation with time and severe temperature dependence of the emf of such cells. In particular, the CCE proposed, and the Conference International des Poids et Mesures (CIPM) subsequently approved the proposal, that the national standards laboratories adopt $K_{J\text{-}72} = 483\ 594\ \mathrm{GHz\ V^{-1}}$, exactly, as the conventional value of the Josephson constant for use in realizing and maintaining an accurate and stable national representation of the volt by means of the JE (Taylor and Witt, 1989). However, while most national laboratories accepted this value, three did not. Moreover, as pointed out in Section 3.3.3, the 1986 CODATA adjustment showed that $K_{J\text{-}72}$ was nearly 8×10^{-6} smaller than the correct SI value, implying that the national representations of the volt of those countries that used it should be changed accordingly.

During the late 1970s and early 1980s, the CCE also became increasingly concerned that, because most national standards laboratories based their representation of the ohm on the mean resistance of a particular group of precision wire-wound standard resistors, and because these artifact standards age, the various national representations of the ohm differed significantly from each other, and from the ohm of the SI definition, and some groups of resistors were aging excessively. Although, in principle, a calculable cross-capacitor can now be used to realize the ohm with an uncertainty of $<10^{-7}$, it is a difficult experiment to perform routinely. Only one laboratory in the world has had such an apparatus in continuous operation since the method was first developed in the early 1960s. Consequently, electrical metrologists throughout the world enthusiastically welcomed von Klitzing's 1980 discovery of the QHE since it promised to provide a method for basing a representation of the ohm on fundamental constants in much the same manner as the JE, discovered 19 years earlier in 1961, provided a method for basing a representation of the volt on fundamental constants.

3.4.2
The Conventional Values K_{J-90} and R_{K-90}

To address the problems associated with national representations of the volt and ohm discussed earlier, the CCE established the CCE Working Group on the JE and another working group on the QHE. Their proposals were subsequently approved by the CIPM and by the 18th CGPM (Giacomo, 1987; Quinn, 1989) and these proposals went into effect worldwide from 1 January 1990. The values are

$$\begin{aligned} K_{J-90} &= 483\,587.9\text{GHz/Vexactly} \\ R_{K-90} &= 25\,812.807\Omega \quad \text{exactly} \end{aligned} \tag{3.21}$$

The respective CCE Working Groups estimated that $K_{J\text{-}90}$ was consistent with the SI value of K_J to within a conservatively assigned one-standard-deviation uncertainty of 0.2 GHz V^{-1} or relative uncertainty of 4×10^{-7}, and that $R_{K\text{-}90}$ is consistent with the SI value of R_K to within a conservatively assigned one-standard-deviation uncertainty of 0.005 Ω or a relative uncertainty of 2×10^{-7}. This implies that an ideal representation of the volt based on the JE and $K_{J\text{-}90}$ should be consistent with the SI volt to within 0.4 μV, and that an ideal representation of the ohm based on the QHE and $R_{K\text{-}90}$ should be consistent with the SI ohm to within 0.2 μΩ. Reassuringly, the 2006 evaluation improved on this uncertainty and suggested that the maintained units were actually 0.019(25) μV lower and 0.0212 μΩ higher, respectively, and that the corresponding maintained ampere A_{90} differed from that of the SI definition by −0.003(25) μA. Hopefully, these small differences will be confirmed by measurements currently in progress.

Thus, measurements of the fundamental constants of nature have led to a very practical and significant improvement of the international uniformity of electrical measurements and their consistency with the SI. This will no doubt be of major benefit to science, commerce, and industry throughout the world for many years to come; not least because all measurements are a comparison of one quantity relative to another like quantity, and one is never otherwise sure whether the problem lies with a measurement or the reference quantity.

3.5
Implications of Recent Advances in the Fundamental Constants Field

As clearly discussed in Section 3.4, many advances have been made in the fundamental constants field since the completion of the 1986 CODATA least-squares adjustment. The new results lead to values of the constants that in general have uncertainties five to seven times smaller than the uncertainties assigned to the 1986 values. However, the changes in the values themselves are less than twice the 1986 one-standard-deviation uncertainties and thus are not highly significant. Although much new data has become available since 1986, three new results dominate the analysis: a value of h obtained from a realization of the watt; a value of α^{-1} obtained from a_e; and a value of the molar gas constant R obtained from a new measurement of the speed of sound in argon.

Surprisingly, the measurement of $K_W = W_{90}/W$ is actually a measurement of the Planck constant. This follows from the fact that $W_{90} = V_{90}^2/\Omega_{90}$, $W = V/\Omega$, so that $K_{\mathrm{W}} = K_{\mathrm{V}}^2/K_{\Omega}$. Also $K_V = K_{J\text{-}90}/K_J$, $K_{\Omega} = R_K/R_{K-90}$, $K_J = 2e/h$, and $R_K = h/e^2$. These relations lead to $h = 4/K_J^2 R_K$ and hence

$$h = \frac{4K_W}{K_{J-90}^2 R_{K-90}} \tag{3.22}$$

Since $K_{J\text{-}90}$ and $R_{K\text{-}90}$ are defined constants (with no uncertainty), a measurement of K_W (i.e., a measurement of W_{90} in watts or realization of the SI watt) with a given relative uncertainty yields a value of h with the same relative uncertainty.

Since 1986, the value of the molar gas constant R from which the Boltzmann constant k, Stefan–Boltzmann constant σ, and other thermophysical constants were derived has been replaced by a new value of R with a relative uncertainty of 1.7×10^{-6}, nearly five times smaller than the 8.4×10^{-6} relative uncertainty of the 1986 value.

3.6
Possible Role of the Fundamental Constants in Replacing the Kilogram

As indicated in Section 3.4, this section further demonstrates the important role that the fundamental constants can play in the practical metrology.

3.6.1
Problems with Present Kilogram

The unit of mass in the SI, the kilogram, is the only SI base unit still defined in terms of a material artifact; it is equal to the mass of the international prototype of the kilogram, a 90% platinum, 10% iridium (by weight) right circular cylinder about 39 mm in height and diameter. Established by the CGPM in 1889, it is rarely used for fear of damage. In fact, it has only been used on three occasions to compare against copies of the prototype since it was made.

The problems with the present material artifact kilogram may be summarized as follows (Quinn, 1991; Taylor, 1991):

1. it can be damaged, or even destroyed;
2. it cannot be used routinely for fear of wear;
3. it is available in only one laboratory;
4. its value may have changed in an unknown manner, perhaps by as much as 75 µg in the last 100 years (7.5 parts in 10^8); and
5. its mass is not well defined because, despite its being enclosed by three concentric bell jars, it accumulates foreign material and the latter are difficult to remove reproducibly and an empirical, "lavage" process has been specified.

Using the best available balances and cleaning the masses under the proper experimental conditions, Pt–Ir kilogram standards can be intercompared with an uncertainty of about 0.001 µg (or a fractional amount of 1×10^{-9}) or even less. However, it is generally agreed that a new definition of the kilogram based on a constant of nature, such as the mass of a stable elementary particle or atom, that could be realized, in practice, with a relative uncertainty of the order of 1 × 10^{-8} at the 1-kg level is all that is required to replace the current definition. The reason is that the uncertainties associated with the cleaning and aging of the international prototype, the air buoyancy correction required when masses of different densities are compared, and the scaling of masses up and down from the 1-kg level limit practical mass measurements to a relative uncertainty of a few parts in 10^8 at best. There are also losses of accuracy in moving to other in-house mass balances and to other national standards laboratories and to the ultimate users.

3.6.2
An Alternative Based on Fundamental Constants

There are at present three alternative approaches that might be developed to have the required precision.

3.6.2.1 The Measurement of *h*

The most promising method for redefining the kilogram, to date, arose from the result of the measurement of $\gamma_p'(\text{hi})$ by Kibble and Hunt, as mentioned in Section 3.2.3. The method proposed by Kibble (1975) effectively combined aspects of the measurements of $\gamma_p'(\text{hi})$ and $\gamma_p'(\text{lo})$ into one measurement process that is performed on the same coil of n turns and horizontal width w in the same magnetic flux density B in the presence of a vertical flux gradient $\partial\Phi/\partial z$. The two-aspect measurement process involves

1. measuring the force mg, required to balance the coil when a current I was passed through it and
2. measuring the induced emf, E, when the coil was moved vertically in a flux rate of change of $\partial\Phi/\partial t\,(= u\partial\Phi/\partial z)$, at constant velocity, u.

This resulted in two equations from which $\partial\Phi/\partial z$ could be eliminated to give the electrical power VI in absolute mechanical units. If the voltage is measured in terms of a Josephson voltage $V_J = hf_J$ and the current $i = \left(f_J'n\right)e/2$ is measured by combining the JE ($v = hf/2e$) and QHE ($R = h/ne^2$), (where n is a small integer, and f and f' are small frequencies). Thus, since $(h/2e)^2 \big/ (e^2/2h)) = h/8$, the Planck constant h can be measured directly in terms of its SI units, that is, in Joule-second. The experimental determination requires knowledge of the mass m, the vertical coil velocity u (which is measured by optical fringe-counting techniques), and the gravitational acceleration g local to the apparatus (again involving optical fringe-counting of a cube corner reflector catapulted vertically in a commercially available vacuum apparatus). The remaining parameters involve frequency measurements that can readily be measured with the required accuracy. To date, there have been two groups making determinations, Williams *et al.* at NIST, and Kibble *et al.* at NPL. The latest results from the two groups are

Source	**Value for *h***
Steiner *et al.* (NIST) (Steiner *et al.*, 2007)	6.626 068 91(24) × 10^{-34} J.s
Robinson and Kibble (NPL) (Robinson and Kibble, 2007)	6.626 070 95(44) × 10^{-34} J.s

These two-watt balance measurements differ at present by 308(48) nW/W, with the

Robinson and Kibble value being closer to that obtained from the Avogadro constant as shown in Figure 3.5.

It seems likely at present that this route will prove the most capable of achieving the desired 10^{-8} level of accuracy. That being so, the definition of the kilogram might be defined via the Einstein mass–energy equation ($E = mc^2$) in terms of the mass equivalence $h\nu_{def}/c^2$ (where ν_{def} is the frequency defined, e.g., in the definition below) and read, for example, Mills *et al.* (2005):

> *the kilogram, unit of mass*
> *is the mass of a body whose*
> *equivalent energy is equal to*
> *that of a number of photons*
> *whose frequencies sum to exactly*
> *$[(299\,792\,458)^2/662\ 606\ 93] \times$*
> *10^{41} hertz.*

It has to be admitted that this definition is likely to be difficult for many unsophisticated users to comprehend, despite representing the amount of mass converted into energy in a very powerful hydrogen bomb. Also, $h\nu_{def}$ is largely independent of h, rather depending on $\alpha^4 m_e$ instead (Flowers and Petley, 2005). The latter is a consequence of there being no single fundamental constant whose dimensions have solely the dimension of time, aside from the Planck time $t_P\left(=\left(\hbar G/c^5\right)^{1/2}\right)$, unless one considers, for example, such quantities as the Compton frequencies of particles, or the Rydberg constant as being more fundamental than those of the conventional basic set comprising h, e, c, G, k, and m_e. Much will depend on the degree of agreement achieved between the various methods, together with the combined views of users throughout the world, by the time that the decision is discussed by the CGPM at its next meeting in 2011.

3.6.2.2 Methods Involving the Avogadro Constant, N_A

An alternative approach to eliminating the artifact kilogram as the definition of the SI unit of mass is to replace it with a definition based on the mass of an atom, for example, carbon-12 (^{12}C). A possible definition might read as follows: the kilogram is the unit of mass; it is equal to the mass of 5.018×10^{25} free ^{12}C atoms at rest. To avoid introducing a discontinuity in the SI mass unit, the actual number $n(^{12}\mathrm{C})$ given in the definition would, to the fullest possible extent, be taken equal to the number of free ^{12}C atoms at rest whose total mass equals the mass of the international kilogram prototype. Thus we have

$$n\left(^{12}\mathrm{C}\right) = \frac{1kg}{m\left(^{12}\mathrm{C}\right)} = 1kg\frac{N_{\mathrm{A}}}{M\left(^{12}C\right)} \qquad (3.23)$$

where $m(^{12}\mathrm{C})$ is the mass in kg of a free ^{12}C atom at rest; N_A is the Avogadro constant in units of mol^{-1}; and also $M(^{12}\mathrm{C})$ is the molar mass of ^{12}C in units of kilograms per mole, and from the definition of the base unit mole in the SI is exactly 0.012 kg mol^{-1}. It follows from Equation 3.15 that once the new definition is implemented (i.e., once an exact value for n(^{12}C) is adopted), N_A and the atomic mass unit $u = m_u = m(^{12}\mathrm{C})/12$ will become exactly defined constants.

Mass–ratio measurements of single charged particles and ions cooled and trapped in a Penning trap have reached the point where the molar masses of most particles or atoms of interest, such as the electron, proton, or ^{1}H, are now known or soon will be known with a relative uncertainty of $(1–3) \times 10^{-9}$. Similarly, the 20×10^{-9} relative uncertainty of the reported value for the proton–electron mass ratio, $m_p/m_e = M_p/M_e$, should soon

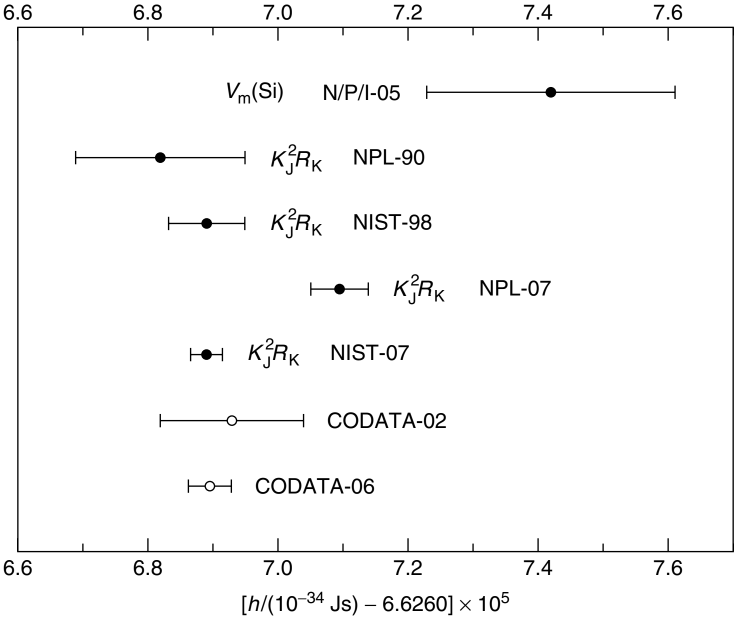

Fig. 3.5 Recent values of the Plank constant based on watt balance ($K_J^2 R_K$) and Avogadro V_m(Si) methods. (Taken from CODATA 2006, see Mohr, Taylor, and Newell (2008) for details).

be reduced by a factor of 10 or so. These small uncertainties imply that if the mass m of any elementary particle or atom can be obtained in kilograms with a relative uncertainty of 1×10^{-8}, then $m(^{12}\mathrm{C})$ and hence $n(^{12}\mathrm{C})$ will be known with essentially the same uncertainty. Further, Equation 3.16 implies that determining N_A with a relative uncertainty of 1×10^{-8} will lead to a value of $n(^{12}\mathrm{C})$ with the same uncertainty. Consequently, one needs to only consider the determination of N_A and, since the electron is the particle whose mass can be obtained most accurately at present, the determination of m_e.

Determining N_A and m_e As discussed earlier, values for constants may be obtained by direct measurement or indirectly from appropriate combinations of other directly measured quantities. Hence, as explained in Section 3.3, N_A may be determined

1. directly by the X-ray crystal density (XRCD) method in which one measures the molar mass, density, and unit-cell volume of a material such as silicon that can be prepared in the form of extremely pure and nearly perfect single crystals. For silicon, a cubic crystal with eight atoms per unit cell of edge length $a = d_{220}\sqrt{8}$ and volume $v = a^3$, the basic relation is $N_A = 8M(Si)/\rho(Si) \times \left(d_{220}\sqrt{8}\right)^3$.
2. N_A may also be obtained indirectly by a number of different routes but, since $m_e = M_e/N_A = M_p/[N_A(m_p/m_e)]$, any indirect measurement of N_A may be viewed as an indirect measurement of m_e.

Since m_e cannot yet be measured directly (one cannot yet "weigh" electrons), only indirect measurements (other than via N_A) need be considered. On the basis of the equations $R_\infty = m_e c\alpha^2/2h$, $\alpha = \mu_0 ce^2/2h$, and relations previously discussed, one can show that

$$
\begin{aligned}
m_e &= \frac{8R_\infty \alpha^{-2} K_W}{cR_{K-90} K_{J-90}^2} \\
&= \frac{16R_\infty \alpha^{-1} K_V^2}{\mu_0 c^2 K_{J-90}^2} \\
&= \frac{4\mu_0 R_\infty \alpha^{-3} K_A^2}{R_{K-90}^2 K_{J-90}^2}
\end{aligned}
\tag{3.24}
$$

where K_W, K_V, and K_A have the same meaning as in Section 3.5, namely, $K_W = W_{90}/W$, $K_V = V_{90}/V$, and $K_A = A_{90}/A$.

Now, in the SI, μ_0, c, R_{K-90}, and K_{J-90} are all exact constants; R_∞ is known with a relative uncertainty that is currently limited by the practical realization of the meter at visible wavelengths or about 6.6×10^{-12} and α^{-1} as obtained from a_e has a relative uncertainty of 6.9×10^{-10} and it should eventually be possible to reduce this to $(1 - 2) \times 10^{-10}$. Consequently, realization of the watt, volt, or ampere (i.e., determination of K_W, K_V, or K_A) with a relative uncertainty of about 1×10^{-8} could lead to a new definition of the SI unit of mass.

One can also show that

$$
\begin{aligned}
m_e &= \frac{2\left(\mu_p'/\mu_B\right)}{C_{K-90} K_{J-90} \gamma_p'(hi)_{90}} \\
&= \frac{2M_p \alpha^{-2} K_w}{R_{K-90} K_{J-90} (m_p/m_e) F_{90}}
\end{aligned}
\tag{3.25}
$$

where again the various quantities entering these relations have the same meaning as in the previous sections. The current relative uncertainty of μ_p/μ_B is 11×10^{-9} and can likely be significantly reduced if need be. The current relative uncertainty of m_p/m_e is 20×10^{-9}, and, as noted in Section 3.6.2, this will eventually be reduced by a factor of 10 or more. Consequently, a determination of either $\gamma_p(\mathrm{hi})_{90}$ or F_{90} with an uncertainty of about 10^{-8} can also lead to a value of m_e, and thus $m(^{12}\mathrm{C})$ and $n(^{12}\mathrm{C})$, having essentially the same uncertainty and therefore possibly lead to a new definition of the kilogram.

It should be recognized that Equations 3.17 and 3.18 are potentially useful as a basis for replacing the kilogram only because it is possible to calibrate laboratory voltage standards in terms of V_{90} with a relative uncertainty of about 1×10^{-9} or better using a Josephson-array voltage standard, and laboratory resistance standards in terms of Ω_{90} with a similar uncertainty using the QHE and all-cryogenic techniques.

The data pertaining measurements of the Avogadro constant provided an important input to the evaluations. This is a modernized version of the method from the 1930s era of Bragg and others, whereby the Avogadro constant is determined from a measurement of the lattice spacing d_{220}, and molar density $M(\mathrm{Si})$ of a pure silicon crystal.

There is considerable interest in this work for technological reasons because it is one of the two types of project providing the basis for decisions of whether and how to redefine the definition of the kilogram. This is a truly international project and technological input and measurements of various aspects are being made in USA, Japan, Belgium, Italy, Germany, Russia, and Australia.

A quantity d_{220} was introduced as an unknown in the 1986 and later adjustments. The two quantities d_{220} and $V_m(\mathrm{Si}) = M(\mathrm{Si})/(\mathrm{Si})$ were treated as separate items of stochastic input data, because the two available experimental values of d_{220} were in gross disagreement. If they had not been discrepant, then their weighted mean could have been used in conjunction with the single available value of $V_m(\mathrm{Si})$ to obtain a single value of N_A through the relation $N_A = 8V_m\,(\mathrm{Si}) \,/\, \left(d_{220}\sqrt{8}\right)^3$. This result could then have been included in the adjustment directly and there would have been no need to introduce d_{220} as an unknown. For the 2006 evaluation, four measurements of d_{220} were considered. One measurement differed markedly from the other three, but was included because it was supported by the value deduced from measurements of the Compton wavelength of the neutron, expressed as a measurement of $h/(m_n d_{220})$, where m_n is the neutron mass.

Work on resolving the present discrepancies continues. Silicon is not a single isotope element, and the relative isotopic abundances vary from one sample to another, or even throughout the volume of a particular sample. Consequently, it is very difficult to produce silicon crystals that are completely homogeneous. Moreover, there can be small faults and possibly impurities affecting both the density and lattice parameters. The vacuum zone refining method of purification produces gradients in the silicon isotopic abundances and, with this in mind, considerable efforts are being made to produce single isotope samples of very pure silicon and results are expected shortly.

3.6.3
Practical Mass Metrology Making Use of a New Definition of the Kilogram

Suppose that one of the above approaches eventually yields a value of m_e or h with the requisite relative uncertainty of about 1×10^{-8} and the kilogram is redefined in terms of either $n(^{12}\mathrm{C})$ or $h\nu_{\mathrm{def}}/c^2$, how would the newly defined kilogram be used in practice?

First, it should be recognized that either present-day platinum–iridium national kilogram standards (particularly those also made along with the prototype), or the prototype kilogram itself, would not be discarded. Rather, they would serve to act as ‘‘flywheels,’’ that is, as a means of preserving the unit of mass between realizations based on the new definition. Their periodic calibration would be performed by carrying out, for example, a watt-realization experiment, or an Avogadro-constant determination, in terms of the mass standard to be calibrated, taking as its value that is currently assigned to it. Any deviation of the measured value of h, or m_e and thus h, or $n(^{12}\mathrm{C})$ from their adopted value would/could be attributed to a shift in the mass standard (always assuming that there was not a really major problem with the data being input to the ‘‘best values’’ available by 2011). A new value would then be assigned to the standard, thereby recalibrating it.

For many years, only a few laboratories in the world would need to have the apparatus necessary to carry out such experiments, because the results obtained could be readily transferred between laboratories with sufficient accuracy by intercomparing mass standards. If more accurate values of relevant fundamental constants (e.g., R_∞, α^{-1}) became available, the adopted value of $n(^{12}\mathrm{C})$ or $h\nu_{\mathrm{def}}/c^2$ would not be changed. Rather, one would take into account any change in $n(^{12}\mathrm{C})$, or $h\nu_{\mathrm{def}}/c^2$, implied by the improved values of the constants by assigning new values to mass standards. Of course, before any new kilogram definition is adopted, the international prototype would be measured periodically (i.e., monitored) in terms of the proposed new definition to minimize introducing a discontinuity in the SI unit of mass.

A second major advantage of the change would be the elimination of a common, correlated, uncertainty from the values of many of the dimensioned fundamental constants, with the choice in terms of h having the added advantage of improved accuracy by removing the uncertainty in m_p/m_e from many constants.

The new kilogram definition would be realized for a kilogram that was in vacuum, but the transfer to a kilogram that is in air can already be made with the requisite accuracy, see Robinson and Kibble (2007), Robinson (1990), and Davidson, Brown, and Berry (2004). It remains to be seen whether future primary kilogram standards will be sustained in air or vacuum, or both.

3.7
Numerology and Time Variations of the Constants

Any chapter on the fundamental physical constants would not be complete without at least a brief discussion of why the constants have the values they do, and whether they are in fact really constant. For example, why does the ratio of the rest mass of the proton to that of the electron happen to be 1836.15267247(80)? Why does α^{-1} happen to equal 137.035999679(94)? Why not other, perhaps simpler, values?[2)] Three rather different points of view have evolved about the constants, especially their dimensionless combinations:

1. The numbers are not at all arbitrary, but can be calculated from some as yet unknown basic theory (or theories) in much the same way that the g-factors of the electron and muon can be calculated from QED, or even in the way that the ratio of the circumference of a circle to its diameter need not be experimentally determined, but can be calculated to arbitrary accuracy from mathematics. For example, in the late 1960s, an expression was derived for α and m_p/m_e involving only simple integers and, based on the volumes of

---

[2)] Viewed in the round, there are probably, to some extent, relationships between the digits in any group of numbers, with the frequencies of particular integers of the first and successive digits depending on the empirical Benford's Law. This was proposed in 1938 by the late Dr Frank Benford, and extended a discovery reported in 1881, by the astronomer/mathematician Simon Newcomb. If absolute certainty is defined as 1 and absolute impossibility as 0, then the probability of any number "d" from 1 to 9 being the first significant digit is given by

$$P\{d\} = \frac{\ln\left(1+\frac{1}{d}\right)}{\ln(10)} = \log_{10}\left(1+\frac{1}{d}\right) \tag{3.26}$$

and for the nth decimal place it is

$$P_n\{d\} = \frac{1}{\ln(10)} \sum_{k=10^{n-1}}^{10^n-1} \ln\left(1+\frac{1}{10k+d}\right) = \sum_{k=10^{n-1}}^{10^n-1} \log_{10}\left(1+\frac{1}{10k+d}\right) \tag{3.27}$$

(For numbers to base B replace the number 10 by B) Benford found that this formula also predicts the frequencies of numbers found in many categories of statistics. Although after a few decimal places all integers are equally probable, it is interesting that this law also applies to the fundamental constants (see, e.g., *http://www.mathpages.com/home/kmath302/kmath302.htm* and Torres *et al.*, 2007) and their determination. Thus, for the first digit is 1 for about 30% of the constants and 7, 8, or 9, for about 5% of the constants, rather than an equal spread of 10% for the digits 1–9 that one might have otherwise expected.

certain bounded spaces associated with the invariance group of a relativistic quantum wave equation, and giving a value for them which were in excellent agreement with the recommended value at the time (Wyler, 1969), although not in accordance with present-day values. However, the physical basis for his derivation was not at all clear, and it remains safe to say that there does not yet exist a physically meaningful ''derivation'' of an accurate value for a fundamental constant, or dimensionless ratio or combination of constants. When one considers that evaluation of the tenth order QED term of the calculation of the electron $g_e - 2$ requires the evaluation of some 12 672 Feynman diagrams, one can imagine that this task remains a formidable challenge for the future. It would, of course, be nice if some development led to the problems associated with renormalization in QED going the way of the epicycles of pre-Keplerian astronomy.

2. The second viewpoint is not aimed at explaining the values of the physical constants, but simply notes that if they were terribly different from their currently observed values, we would not be here to measure them–life as we know it on earth, and perhaps the existence of the universe itself, depends on complex physical processes that require the constants to have their observed values. This is the idea behind the so-called weak anthropic principle (or WAP) (Barrow and Tipler, 1988). The strong anthropic principle (or SAP) goes even further by speculating that the fundamental constants and laws of physics must be such that life, and the universe as we know it, can exist–which represents a considerable extension of the ''Cogito ergo sum,'' of Descartes of nearly six centuries ago.
3. An extreme statement of the third point of view is that the fundamental constants change with time as the universe evolves and therefore the currently observed values are not particularly significant. This idea had its origins in 1937 when Dirac noted that the ratio of the electric to gravitational forces between the electron and proton is of the order 10^{40}, very nearly equal to the age of the universe expressed in an appropriate atomic time unit, for example, the time required for light to travel a distance equal to the classical electron radius. He speculated that either the equality is accidental or it indicates a causal relationship among electromagnetism, gravitation, and cosmology. If such a relationship is in fact true, then since the universe is continually aging, the equality would remain unchanged only if one or more of the fundamental constants entering the two ratios changed–for example, the Newtonian gravitational constant G, or the elementary charge e. Grand unification theories (GUTs) imply that the fundamental constants were very different during the first second of the ''ig bang'' when there was a unification among the strong, weak, and electromagnetic forces, thereby providing a basis for some of Dirac's ideas on the topic.

These speculations have stimulated a considerable amount of theoretical and experimental work over the last 50 years, the latter quite often involving astronomical or geophysical observations. As a result, stringent limits (albeit frequently model dependent) have been placed on the possible time dependence of a number of physical constants. For example, it has been claimed that h changes less than 2 parts in 10^{12} per year; that G changes less than 3 parts in 10^{11} per year (these are essentially the variations of α and $\alpha_G = Gm_p^2 \big/ \hbar c$, respectively).

We have already entered the situation where the accuracy of laboratory determinations gives measures of the constancy of certain of the constants. This particularly applies to measures of the stabilities of atomic clocks that rely on energy transitions in different atoms.

The frequencies of different types of atomic clock using transitions in different atoms can also be intercompared and also provide information about the stability of α over laboratory timescales. Thus, by comparing transitions, frequencies in Al^+ and Hg^+ single-ion optical atomic clocks Rosenband *et al.* at NIST (Rosenband *et al.*, 2008) measured

$$\nu_{\text{Al}}^{+}/\nu_{\text{Hg}}^{+} = 1.052871833148990438(55) \tag{3.28}$$

and, by comparing measurements taken at various times over a period of about a year, they deduced that α changes by less than $(-1.6 \pm 2.3) \times 10^{-17}$ per year.

One can only infer the constancy of dimensionless quantities because one cannot usually assume that the reference quantity remains constant. In the above example, other dimensionless quantities are also involved in the theoretical expressions for the transition frequencies.

In the near future, the changes in the values of the fundamental constants caused by mankind's present limitations in understanding and implementing the science behind measuring them, will dominate any changes that might be caused by the inconstancy of Nature's laws. The underlying assumptions behind some of the above can be questioned. For example, there has been at least one example of a cosmological "measurement" of the constancy of h where a null result was guaranteed through measuring a quantity by a method that predetermined a null result.

3.8
Future Trends and Advances

The fundamental-constants field is ever evolving; new experiments are undertaken with each major new advance in measurement science. Indeed, roughly speaking, our knowledge of the values of many of the fundamental constants improved by an order of magnitude each decade during the last half of the twentieth century. However, the question naturally arises as to whether this sort of improvement can continue indefinitely. Will phenomena and measuring techniques yet to be discovered lead to continued reductions in uncertainty of this magnitude into the twenty-first century?

We have almost arrived at the era predicted by Maxwell (1870) when he forecast, at a time when the atomic nature of our universe was increasingly emerging, that

> *If we wish to obtain standards*
> *of mass of length, mass or time*
> *that shall be absolutely permanent,*
> *we must seek them not in the*
> *dimensions, or motion, or the*
> *mass of our planet, but on the*
> *wavelength, the period of vibration*
> *and mass of those imperishable*
> *and unalterable and perfectly*
> *similar molecules.*

The projected change to the international measurement system (the basis for the SI units) means that theoretical and experimental science will at last be performed in terms of the same sets of quantities with an exactness that our forbears could only dream of. An important consequence will be that it will be possible for experiments relating to the measurement of dimensioned fundamental constants to revert from the national standard's laboratories to the universities and elsewhere.

We are discussing changing the definition of the SI unit of mass at a time when the discovery of the Higgs boson seems imminent. Why is the Higgs so important? Because it is thought to pervade the universal vacuum; not, as with the old ether, to provide a material substrate for the propagation of electromagnetic waves, but rather to interact with particles and confer mass upon them. The Higgs ministrations are usually hidden away in the vacuum, but if enough energy is brought to bear in a tiny volume of space, at the point where two energetic particles collide, then the Higgs can be turned into an actual particle whose existence can be detected. Theoretical calculations made using the standard model of particle physics combined with previous experiments serve to limit the possible range of masses for the Higgs particle. Right now that mass is thought to lie between 114 and about 190 GeV. However, it is difficult to imagine that an increased understanding of the nature of mass will immediately impact on how we choose to define the mass unit of our very practical SI unit system, but in another 50 years metrology may be heading along a very different direction.

The American Physicist Richtmyer (1932) echoing Maxwell of 50 years earlier, again asked and answered the question: "*Why should one wish to make measurements with ever increasing precision?* He too responded: *Because the whole history of physics proves that a new discovery is quite likely to be found lurking in the next decimal place.*" The last 50 years have amply demonstrated the wisdom of their response. A reply that will undoubtedly continue to be demonstrated in the future.

Some day we must reach the ultimate limits. However, we can be sure that scientists will continue their efforts to advance our knowledge of the fundamental constants unabated as long as they believe there is even a slim chance of success.

Glossary

a_e: Electron magnetic moment anomaly; $a_e = g_e/2 - 1 = \mu_e/\mu_B - 1$, where g_e is the g-factor of the electron, μ_e is the magnetic moment of the electron, and μ_B is the Bohr magneton: $\mu_B = eh/2\pi m_e$, where e is the elementary charge, h is the Planck constant, and m_e is the electron mass.

α: Fine-structure constant, approximately 1/137, is the dimensionless coupling constant or the measure of the strength of the electromagnetic force that governs how electrically charged elementary particles (e.g., electron, muon) and light (photons) interact.

α^{-1} Inverse fine-structure constant; $\alpha^{-1} \approx 137$.

BIPM: Bureau International des Poids et Mesures (International Bureau of Weights and Measures), located in Sèvres, France, a suburb of Paris. BIPM is the international standards laboratory.

CCE: Comité Consultatif d'Électricité (Consultative Committee for Electricity), one of the eight advisory committees to the CIPM.

CGPM: Conférence Général des Poids et Mesures (General Conference on Weights and Measures), the highest international metrological authority established by the Convention du Mètre (Treaty of the Meter) signed in 1875; it is a diplomatic organization.

CIPM: Comité International des Poids and Mesures (International Committee for Weights and Measures), one of the organizations established by the Convention du Mètre (Treaty of the Meter) signed in 1875. The CIPM supervises the BIPM and is under the authority of the CGPM.

CODATA: Committee on Data for Science and Technology, an interdisciplinary committee of the International Council of Scientific Unions. It seeks to improve the compilation, critical evaluation, storage, and retrieval of data of importance to science and technology.

JE: Abbreviation for Josephson effect, a solid-state, low-temperature physics phenomenon characteristic of weakly coupled superconductors cooled below their transition temperatures. An example is two thin films of superconducting lead separated by a thermally grown oxide layer about 1-nm thick.

K_J: Josephson constant: characteristic of the Josephson effect and assumed equal to 2 e/h, where e is the elementary charge and h is the Planck constant.

K_J-90: Conventional value of the Josephson constant adopted by the CIPM for international use in basing a representation of the volt on the Josephson effect starting 1 January 1990; $K_{J-90} = 483\,597.9\,\mathrm{GHz\,V^{-1}}$ exactly.

Muonium: Hydrogen-like atom composed of a positive muon and an electron (μ^+e^-atom).

NIST: National Institute of Standards and Technology, the national standards laboratory of the United States (called the National Bureau of Standards or NBS prior to 23 August 1988).

NPL: National Physical Laboratory, the national standards laboratory of the United Kingdom.

ppm: Symbol often used to denote an uncertainty of a fractional part in 10^6. Being an English language, specific symbol it is not part of the internationally supported set of quantities in the International Standards Organization Standard: ISO 31 and, because it is not used in a consistent way, is not an internationally recommended symbol, and that consequently, where it is used by authors, its meaning must always be defined in the text.

QED: Quantum electrodynamics, the relativistic quantum field theory of the interaction of charged particles and electromagnetic radiation (photons). QED is an amalgamation of Maxwell's theory of electromagnetism, quantum mechanics, and Einstein's special theory of relativity.

QHE: Abbreviation for quantum Hall effect, a solid-state, low-temperature physics phenomenon characteristic of a two-dimensional electron gas. The latter may be realized in a high-mobility semiconductor device such as a silicon metal-oxide-semiconductor field-effect transistor (MOSFET) or GaAs-$\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ heterojunction cooled to about 1 K. The applied magnetic field (magnetic flux density) B must be of order 10 T.

R_K: von Klitzing constant characteristic of the quantum Hall effect and assumed equal to h/e^2, where h is the Planck constant and e is the elementary charge.

R_K-90: Conventional value of the von Klitzing constant adopted by the CIPM for international use in basing a representation of the ohm on the quantum Hall effect starting 1 January 1990; $R_{K-90} = 25812.807\Omega$ exactly.

SI: Système International d'Unités (International System of Units), the modernized metric system with the seven base units meter (m), kilogram (kg), second (s), ampere (A), Kelvin (K), mole (mol), and candela (cd); and numerous derived units such as the Hertz (Hz), Joule (J), Newton (N), watt (W), volt (V), and ohm (Ω).

References

Armstrong, T.R. and Fitzgerald, M.P. (**2003**) *New Measurements of G using the Measurement Standards Laboratory Torsion Balance*, Phys. Rev. Lett., **91**, 201101.

Bagley, C.H. and Luther, G.G. (**1997**) *Preliminary results of a determination of the Newtonian constant of gravitation: a test of the Kuroda hypothesis*, Phys. Rev. Lett., **78** (16), 3047–3051.

Barone, G. and Paterno, A. (**1982**) *Physics and Applications of the Josephson Effect*, John Wiley & Sons, Inc., New York.

Barrow, J.D. and Tipler, F.J. (**1988**) *The Anthropic Cosmological Principle*, Oxford University Press.

de Beauvoir, B., Schwob, C. *et al.* (**2000**) Metrology of the hydrogen and deuterium atoms: determination of the Rydberg constant and Lamb shifts. *Eur. Phys. J. D*, **12**, 61–93.

Becker, P. (**2003**) Tracing the definition of the kilogram to the Avogadro constant using a silicon single crystal. *Metrologia*, **40** (6), 366.

Becker, P., Schiel, D. *et al.* (**2006**) Large-scale production of highly enriched ^{28}Si for the precise determination of the Avogadro constant. *Meas. Sci. Technol.*, **17** (7), 1854.

Birge, R.T. (**1929**) Probable values of the general physical constants (as of January 1, 1929). *Phys. Rev. Suppl.*, **1** (1), 1–73.

Brown, H.N., Bunce, G. *et al.* (**2001**) Precise measurement of the positive muon anomalous magnetic moment. *Phys. Rev. Lett.*, **86** (11), 2227–2231.

Cavendish, H. (**1798**) Experiments to determine the density of the earth. *Philos. Trans. R. Soc. Lond. (1776–1886)*, **88**, 469–526.

Coddington, I., Swann, W.C., and Newbury, N.R. (**2008**) Coherent multiheterodyne spectroscopy using stabilized optical frequency combs. *Phys. Rev. Lett.*, **100** (1), 013902.

Cohen, E.R. and DuMond, J.W.M. (**1965**) Our knowledge of the fundamental constants of physics and chemistry in 1965. *Rev. Mod. Phys.*, **37** (4), 537–594.

Cohen, E.R. and Taylor, B.N. (**1973**) The 1973 least-squares adjustment of the fundamental constants. *J. Phys. Chem. Ref. Data*, **2** (4), 663–734.

Cohen, E.R. and Taylor, B.N. (**1987**) The 1986 adjustment of the fundamental physical constants. *Rev. Mod. Phys.*, **59** (4), 1121–1148.

Davidson, S., Brown, S., and Berry, J. (**2004**) A Report on the Potential Reduction in Uncertainty from Traceable Comparisons of Platinum– Iridium and Stainless Steel Kilogram Mass Standards in Vacuum, NPL. CMAM 88.

DuMond, J.W.M. and Cohen, E.R. (**1953**) Least-squares adjustment of the atomic constants, 1952. *Rev. Mod. Phys.*, **52** (3), 691–708.

Fischer, M., Kolachevsky, N. *et al.* (**2004**) New limits on the drift of fundamental constants from laboratory measurements. *Phys. Rev. Lett.*, **92** (23), 230802–230804.

Flowers, J.L. and Petley, B.W. (**2001**) Progress in our knowledge of the fundamental constants of physics. *Rep. Prog. Phys.*, **64** (10), 1191–1246.

Flowers, J.L. and Petley, B.W. (**2005**) The kilogram redefinition-an interim solution. *Metrologia*, **42** (5), L31.

Gabrielse, G., Hanneke, D. *et al.* (**2006**) New determination of the fine structure constant from the electron g value and QED. *Phys. Rev. Lett.*, **97** (3), 030802.

Gabrielse, G., Hanneke, D. *et al.* (**2007**) Erratum: new determination of the fine structure constant from the electron g value and QED [Phys. Rev. Lett. 97, 030802 (2006)]. *Phys. Rev. Lett.*, **99** (3), 039902.

Gallop, J. (**2005**) The quantum electrical triangle. *Philos. Trans. R. Soc. A*, **363**, 2221–2247.

Gallop, J.C. (**1991**) *SQUIDs, the Josephson Effects and Superconducting Electronics*, CRC Press.

Giacomo, P. (**1984**) News from the BIPM. *Metrologia*, **20** (1), 25–30.

Giacomo, P. (**1987**) News from the BIPM. *Metrologia*, **24** (1), 45–51.

Gundlach, J.H. Merkowitz, S.M. (**2000**) Measurement of newton's constant using a torsion balance with angular acceleration feedback. *Phys. Rev. Lett.*, **85** (14), 2869–2872.

Hänsch, T., Alnis, J. *et al.* (**2005**) Precision spectroscopy of hydrogen and femtosecond laser frequency combs. *Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci.*, **363** (1834), 2155–2163.

Hu, Z.-K., Guo, J.-Q., and Luo, J. (**2005**) Correction of source mass effects in the HUST-99 measurement of G.. *Phys. Rev. D*, **71** (12), 127505.

Hudson, R.P. (**1984**) Documents concerning the new definition of the metre. *Metrologia*, **19** (4), 163–177.

JCGM (**1995**) ISO/IEC Guide 98:1995. International Organization for Standardisation Guide to the Expression of Uncertainty in Measurement (GUM), Geneva, Switzerland.

Karagioz, O.V. and Izmailov, V.P. (**1996**) Measurement of the gravitational constant with a torison balance. *Izmer. Tekh.*, **39** (10), 3–9.

Keller, M.W. (**2008**) Current status of the quantum metrology triangle. *Metrologia*, **45** (1), 102–109.

Keller, M.W., Zimmerman, N.M., and Eichenberger, A.L. (**2007**) Uncertainty budget for the NIST electron counting capacitance standard, ECCS-1. *Metrologia*, **44** (6), 505.

Kibble, B.P. (**1975**) A measurement of the gyromagnetic ratio of the proton by the strong field method, in *Atomic Masses and Fundamental Constants*, vol. **5** (eds J.H. Sanders and A.H. Wapstra), Plenum Press, London and New York, pp. 545–551.

Kinoshita, T. (ed.) (**1990a**) *Quantum Electrodynamics*, Advanced Series on Directions in High Energy Physics, vol. **7**, World Scientific.

Kinoshita, T. and Lindquist, W.B. (**1990b**) Eighth-order magnetic moment of the electron. V. Diagrams containing no vacuum-polarization loop. *Phys. Rev. D*, **42** (2), 636.

Kleinevoß, U. (**2002**) Bestimmung der Newtonschen Gravitationskonstanten G, *University of Wuppertal*. Ph.D. Thesis.

Ludlow, A.D., Zelevinsky, T. *et al.* (**2008**) Sr lattice clock at 1 × 10-16 fractional uncertainty by remote optical evaluation with a Ca clock. *Science*, **319** (5871), 1805–1808.

Maxwell, J.C. (**1870**) Address to the mathematical and physical sections of the British Association, liverpool, Reproduced in *The Scientific Papers of James Clerk Maxwell*, vol. **2** (ed. W.D. Niven), Cambridge University Press, Cambridge, 1980, p. 245 [From the British Association Report, Vol. XL].

Maxwell, J.C. (**1871**) Inaugural address as Cavendish Professor, Reproduced in *The Scientific Papers of James Clerk Maxwell*, vol. **2** (ed. W.D. Niven), Cambridge University Press, Cambridge, 1980, p. 244.

Mills, I.M., Mohr, P.J. *et al.* (**2005**) Redefinition of the kilogram: a decision whose time has come. *Metrologia*, **42** (2), 71–80.

Mohr, P.J. and Taylor, B.N. (**1999**) CODATA recommended values of the fundamental physical constants: 1998. *J. Phys. Chem. Ref. Data*, **28** (6), 1713–1852.

Mohr, P.J. and Taylor, B.N. (**2005**) CODATA recommended values of the fundamental physical constants: 2002. *Rev. Mod. Phys.*, **77** (1), 1–107.

Mohr, P.J., Taylor, B.N., and Newell, D.B. (**2008**) CODATA recommended values of the fundamental physical constants: 2006. *Rev. Mod. Phys.*, **80** (2), 633–698.

Parker, D.H., Taylor, B.N., and Langenberg, D.N. (**1967**) Measurement of $2e/h$ using the ac Josephson effect and its implications for quantum electrodynamics. *Phys. Rev. Lett.*, **18** (8), 287–291.

Parker, W.H., Langenberg, D.N., Experiment, I. *et al.* (**1969**) Determination of e/h, using macroscopic quantum phase coherence in superconductors. *Phys. Rev.*, **177** (2), 639–664.

Petley, B.W. (**1983**) New definition of the metre. *Nature*, **303** (5916), 373–376.

Petley, B.W. and Morris, K. (**1970**) A measurement of $2e/h$ by the ac Josephson effect. *Metrologia*, **6** (2), 46.

Prange, R.E. and Girvin, S.M. (eds) (**1990**) *The Quantum Hall Effect*, Springer-Verlag.

Quinn, T.J. (**1989**) News from the BIPM. *Metrologia*, **26** (1), 69–74.

Quinn, T.J. (**1991**) The kilogram: the present state of our knowledge. *IEEE Trans. Instrum. Meas.*, **40** (2), 81–85.

Quinn, T.J., Speake, C.C., Richman, S.J. *et al.* (**2001**) A new determination of G using two methods. *Phys. Rev. Lett.*, **87**, 111101.

Richtmyer, F.K. (**1932**) The romance of the next decimal place. *Science*, **75**, 1–5.

Robertson, B. (**1971**) Wyler's Expression for the Fine-Structure Constant α. *Phys. Rev. Lett.*, **27** (22), 1545.

Robinson, I.A. (**1990**) Comparing in-air and in-vacuum mass standards without buoyancy corrections via in-vacuum weighing. *Metrologia*, **27** (3), 159.

Robinson, I.A. and Kibble, B.P. (**2007**) An initial measurement of Planck's constant using the NPL Mark II watt balance. *Metrologia*, **44** (6), 427–440.

Rosenband, T., Hume, D.B. *et al.* (**2008**) Frequency ratio of Al+ and Hg+ single-ion optical clocks; metrology at the 17th decimal place. *Science*, **319** (5871), 1808–1812.

Schlamminger, S., Holzschuh, E., Kundig, W. *et al.* (**2006**) Measurement of Newton's gravitational constant. *Physical Review D*, **74** (8), 082001–25.

Shields, J.Q., Dziuba, R.F., and Layer, H.P. (**1989**) New realization of the ohm and farad using the NBS calculable capacitor. *Instrum. Meas., IEEE Trans.*, **38** (2), 249–251.

Steiner, R.L., Williams, E.R. *et al.* (**2007**) Uncertainty improvements of the NIST electronic kilogram. *IEEE Trans. Instrum. Meas.*, **56** (2), 592–596.

Taylor, B.N. (**1987**) History of the present value of 2e/h commonly used for defining national units of voltage and possible changes in national units of voltage and resistance. *IEEE Trans. Instrum. Meas.*, **IM-36**, 659–664.

Taylor, B.N. (**1991**) The possible role of the fundamental constants in replacing the kilogram. *IEEE Trans. Instrum. Meas.*, **40** (2), 86–91.

Taylor, B.N., Parker, W.H., and Langenberg, D.N. (**1969**) *The Fundamental Constants and Quantum Electrodynamics*, Academic Press, New York and London. Reprint of *Reviews of Modern Physics*, 41 (3), 375 (1969).

Taylor, B.N. and Witt, T.J. (**1989**) New international electrical reference standards based on the Josephson and quantum Hall effect. *Metrologia*, **26**, 47–62.

Terrien, J. (**1973**) News from the Bureau International des Poids et Mesures. *Metrologia*, **9** (1), 40.

Tuninskii, V.S. (**1986**) Adjusting correlated values of fundamental constants. *Meas. Tech.*, **29** (8), 702–706.

Tuninskii, V.S. and Kholin, S.V. (**1975**) Concerning changes in the methods for adjusting the physical constants. *Metrologia*, **8**, 3 (in Russian).

Van Dyck, R.S.Jr , Schwinberg, P.B., and Dehmelt, H.G. (**1987**) New high precision comparison of electron and positron g-factors. *Phys. Rev. Lett.*, **59** (1), 26–29.

Werth, G., Beier, T. *et al.* (**2002**) A new value for the mass of the electron from an experiment on the g factor in 12C5+ and 16O7+. *Can. J. Phys.*, **80** (11), 1241–1247.

Wyler, A. (**1969**) *C. R. Acad. Sci. Ser. A*, **269**, 743.

Yoshioka, D. (**2002**) *The Quantum Hall Effect*, Springer.

Further Reading

Flowers, J. (**2004**) The route to atomic and quantum standards. *Science*, **306** (5700), 1324–1330.

Mills, I.M., Mohr, P.J. *et al.* (**2006**) Redefinition of the kilogram, ampere, kelvin and mole: a proposed approach to implementing CIPM recommendation 1 (CI-2005). *Metrologia*, **43** (3), 227–246.

Mohr, P.J. and Taylor, B.N. (**2000**) CODATA recommended values of the fundamental physical constants: 1998. *Rev. Mod. Phys.*, **72** (2), 351–495.

Petley, B.W. (**1985**) *The Fundamental Physical Constants and the Frontier of Measurement*, Adam Hilger, Bristol.

Quinn, T. and Burnett, K. (**2005**) The fundamental constants of physics, precision measurements and the base units of the SI. *Philos. Trans. R. Soc. A*, **363** (1834), 2097–2327.

This special issue contains many chapters on the status of aspects of fundamental constants and measurement.

Uzan, J.-P. (**2003**) The fundamental constants and their variation: observational status and theoretical motivations. *Rev. Mod. Phys.*, **75**, 403.

Useful Web Sites

SI brochure *http://www.bipm.org/en/si/si_brochure/*

NIST fundamental constants *http://physics.nist.gov/cuu/Constants/*

BIPM web site *http://www.bipm.org/*

COPPER

Mary W. Covington,* *Consultant*

Nathan L. Church and William H. Dresher, *International Copper Association, New York, U.S.A.*

Bernard R. Cooper, *Dept. of Physics, West Virginia University, Morgantown, West Virginia, U.S.A.*

Introduction 267
1. **Basic Electronic and Physical Behavior** 267
2. **Characterization and Properties** 270
2.1 Fatigue 270
2.2 Magnetic Properties 270
2.3 Formability 270
2.3.1 Strengthening 270
2.4 Corrosion and Oxidation 271
2.5 Joining 272
3. **Resources and Raw Materials** 272
4. **Production** 273
5. **Copper and Alloy Terminology and Standards** 274
6. **Applications** 275
7. **Summary** 276
Works Cited 276
Further Reading 276

INTRODUCTION

Copper and its alloys are used when the application requires high electrical and thermal conductivity, formability, ease of joining, corrosion resistance, and moderate to high strength. In addition to its traditional uses in building construction, electrical and electronic applications, industrial machinery, transportation equipment, and consumer products, new uses for copper are constantly being found. Some of these are in solar heating, electric vehicles, superconductors (both high- and low-temperature types), fire safety systems, and even pharmaceuticals.

We first discuss the basic electronic and physical behavior of copper. All of the valuable functional properties of copper find their foundation in this electronic behavior. As an example, we describe the features of the electronic structure that lead to the high electrical conductivity of copper.

The useful properties of copper in practice are determined not only by the ideal behavior based on the fundamental electronic and physical behavior, but also by the modifications of that behavior brought about by processing and the presence of impurities. In Sec. 2, we therefore discuss the impurity content characteristic of high-conductivity copper and then summarize some of the key physical and mechanical properties of copper. After doing this, we discuss the practical aspects affecting the utilization of copper: resources and raw materials, production, alloy terminology, and standards. In Sec. 6, we describe the utilization of copper in a variety of applications before providing a brief summary in the final section.

1. BASIC ELECTRONIC AND PHYSICAL BEHAVIOR

Copper (atomic number $Z=29$) is in subgroup 1*B* of the periodic table of the elements. The copper atom contains 2, 8, and 18 electrons in its inner shells and 1 electron in its outer shell. The cuprous ion (Cu^+) is formed by loss of the first electron and the cupric ion (Cu^{2+}) is formed by loss of a second electron. The atomic weight is 63.54. There are two stable isotopes, ^{68}Cu and ^{65}Cu. The specific gravity is 8.89 and the melting point is 1083 °C.

---

*Deceased.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

The most important attribute of the electronic behavior of copper is that it behaves like a monovalent metal such as the alkali metals, but with behavior modified by the existence of the filled *d*-shell states at an energy about 2 eV below the unfilled *s*-like states that provide the monovalent behavior. The two most conspicuous electronic-behavior-based properties of copper, its high electrical conductivity and its distinctively reddish "copper" color, follow as consequences of these features of the electronic structure.

To envision the relationship of electrical conductivity, optical behavior, and other properties to the electronic structure of copper we have to consider the first Brillouin zone for the face-centered-cubic (fcc) structure, as shown in Fig. 1, and the nature of the Fermi surface encompassing the filled electronic states within that zone (Callaway, 1964). For a perfect three-dimensional lattice, there are three primitive translations, $\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3$, which define the periodicity of the lattice. The requirement that the potential acting on an electron moving through a solid be identical at any two points $\mathbf{r}$ and $\mathbf{r}'$ connected by any linear combination of multiples of these primitive translation vectors leads to the requirement that the electronic wave functions obey Bloch's theorem and are labeled by a crystal-momentum quantum number $\mathbf{k}$, which is specified in terms of primitive reciprocal lattice vectors $\mathbf{b}_1,\mathbf{b}_2,\mathbf{b}_3$. Thus,

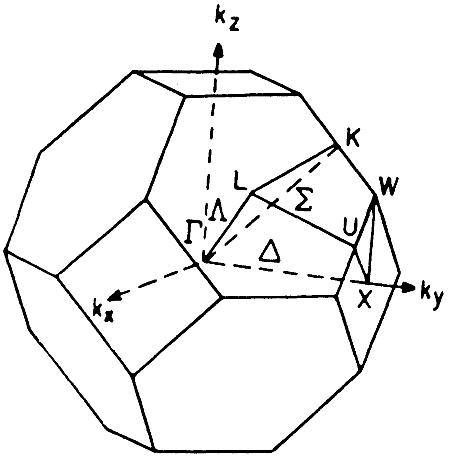

FIG. 1. The first Brillouin zone for the face-centered-cubic structure with points and lines of high symmetry labeled in conventional notation.

$$\mathbf{k}=2\pi\left(\frac{k_1\mathbf{b}_1}{N_1}+\frac{k_2\mathbf{b}_2}{N_2}+\frac{k_3\mathbf{b}_3}{N_3}\right), \qquad (1)$$

where k_1, k_2, k_3 and N_1, N_2, N_3 are integers with $k_1,k_2,k_3<N_1,N_2,N_3$, and

$$\mathbf{b}_1=\frac{\mathbf{a}_2\times\mathbf{a}_3}{\mathbf{a}_1\cdot(\mathbf{a}_2\times\mathbf{a}_3)}, \qquad (2)$$

with $\mathbf{b}_2$ and $\mathbf{b}_3$ obtained by cyclic permutation of the indices of the $\mathbf{a}$'s.

The *reciprocal-lattice vectors* $\mathbf{K}_m$, which are linear combinations of integral multiples of $\mathbf{b}_1$, $\mathbf{b}_2$, and $\mathbf{b}_3$, define a periodic lattice in reciprocal (or *lattice-momentum*, $\mathbf{k}$) space; The part of $\mathbf{k}$ space closer to the origin than to any nonzero $\mathbf{K}_m$ is called the *first Brillouin zone*, and is constructed by setting up planes that perpendicularly bisect the lines connecting the origin to all $\mathbf{K}_m$, and then taking the volume about the origin enclosed by these intersecting planes. Figure 1 shows the first Brillouin zone for the face-centered-cubic structure with conventional labeling of points and lines of high symmetry. Since the product $N_1N_2N_3$ is a very large number essentially corresponding to the number of copper atoms in a macroscopic sample ($\sim 10^{24}$), and k_1, k_2, k_3 are integers ranging from zero up to N_1, N_2, N_3, the $\mathbf{k}$ form a very finely spaced semicontinuum that in practice can be treated as a continuous variation of lattice momentum across the first Brillouin zone.

At each value of lattice momentum $\mathbf{k}$ in the first Brillouin zone there are a number of discrete electron energy states (E_m); and as $\mathbf{k}$ varies across the Brillouin zone, the set of curves given by the variation of the E_m with $\mathbf{k}$ gives the electronic energy band structure (Callaway, 1964). Because of the Fermi statistics obeyed by electrons, the first Brillouin zone can hold two electrons per real-space crystal lattice site per band. At absolute zero temperature the band states will be filled with electrons up to an energy level, the Fermi energy (E_F), such that all the electrons are accounted for. For a metal there are one or more partially filled bands. Thus there is a surface in reciprocal space separating filled from empty states. This surface is the *Fermi surface*. Since E_F is typically several electron volts and 1 eV is the energy equivalent of a temperature of 11 606 K, there is only a very slight blurring of the Fermi surface at finite temperatures. The properties of a metal are dominated either by the behavior of the electrons at or near the Fermi surface or by transitions between filled electron states below the Fermi energy and unfilled states above the Fermi energy. For example, the electrical conductivity is determined by the situation that, under the influence of a dc electric field, the only electrons that can respond are those at the Fermi surface, since they can be displaced into adjacent empty states with different lattice momenta. On the other hand, an optical-frequency electric field can cause transitions between filled states below the Fermi energy and unfilled states above E_F, and either or both states can be well separated in energy from E_F.

The simplest example of a Fermi surface is the spherical surface that occurs for a noninteracting electron gas, and the Fermi surface given by band theory for the monovalent alkali metals is quite close to spherical. For copper, as shown in Fig. 2, the Fermi surface is also basically spherical, but with distinctive *necks* contacting the hexagonal < 111 > faces of the first Brillouin zone (Segall, 1962). This behavior has its origin in the outer electron configuration of atomic copper and the way in which these outer electrons interact in the solid to form the energy bands. There is one valence electron state per copper atom, corresponding to the atomic 4*s* state; and it is these states that basically form the partially filled band, the filling of which determines the Fermi energy and Fermi surface. If this were the only effect determining the Fermi surface, it would be very similar to those of the alkali metals, quite close to spherical. However, the filled 3*d* states affect the electronic structure and consequent properties of copper in a very important way.

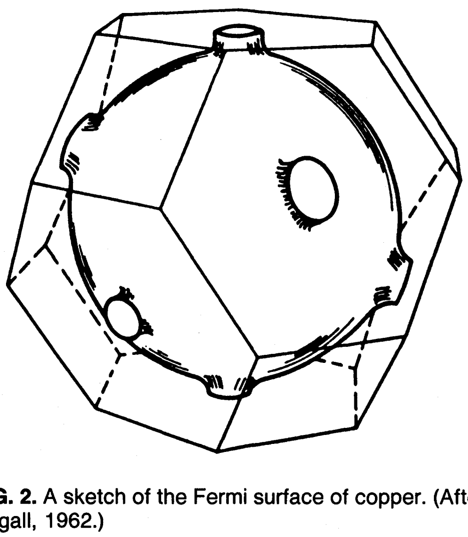

FIG. 2. A sketch of the Fermi surface of copper. (After Segall, 1962.)

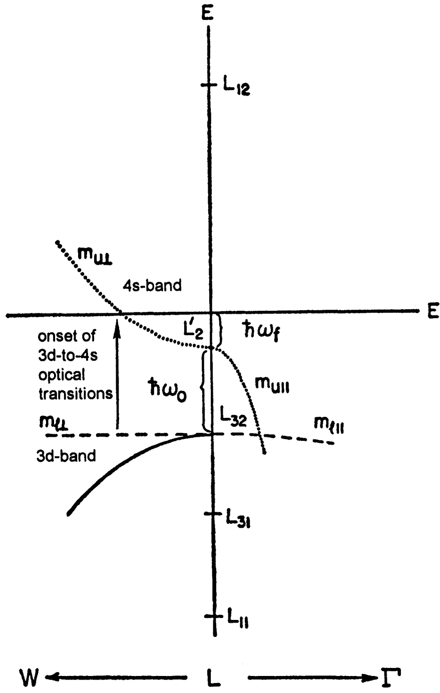

FIG. 3. The band structure near *L* in the first Brillouin zone for the noble metals. (After Cooper *et al.*, 1965.)

The 10 3*d* electrons per copper atom form filled bands in the solid and do not give a Fermi surface; however, these 3*d* states interact with the 4*s* states causing changed behavior (Segall, 1962) of the 4*s* band as shown in Fig. 3. The 4*s* band is "bent" upward by this interaction, so that the Fermi surface intercepts the hexagonal faces of the Brillouin zone with a ring-shaped intersection centered around the < 111 > direction in **k**-space. Thus it is the 4*s*-3*d* interaction (mixing) that causes the *necks* in the copper Fermi surface shown in Fig. 2.

The distinctive reddish color of copper corresponds to the sharp onset in frequency of strong interband optical transitions between filled states at the upper part of the 3*d* band (which has a relatively flat variation of energy with **k**) and empty conduction-band 4*s*-like states (Ehrenreich and Philipp, 1962; Cooper *et al.*, 1965). As can be seen from Fig. 3, these transitions become allowed only at an energy given by the energy difference between the $3d$ and $4s$ bands at the **k**-space locations where the $4s$ band cuts the Fermi energy E_F; and it is the associated sharp onset of strong interband optical transitions (indicated by the arrow in Fig. 3) that dominates the optical behavior of copper.

The high electrical conductivity of copper also arises from the modification of its free-electron-like (alkali-metal–like) monovalent behavior caused by the filled $3d$ states. The filled d shells cause (see Ziman, 1960, Chap. IX) the compressibility of copper to be much smaller than that of the alkali metals. Since the electrical resisitivity of copper at room temperature and above is dominated by scattering of electrons by thermal vibrations, the substantial decrease in amplitude of these vibrations, corresponding to the decreased compressibility, gives rise to the increased electrical conductivity.

2. CHARACTERIZATION AND PROPERTIES

The useful properties of copper in practice are determined not only by the ideal behavior based on the fundamental electronic and physical behavior described in Sec. 2, but in the modification of that behavior brought about by processing and by the presence of impurities. For example, we can consider such effects on the electrical conductivity. High-purity copper's electrical conductivity is slightly over 100% IACS (as based on a volume electrical resistivity of 0.0168 $\mu\Omega$ m). This value refers to the International Annealed Copper Standard (IACS) established in 1913 and is based on a copper standard defined as 100%, being given by 100 times the reciprocal ratio to the volume resistivity of annealed copper at 20 °C as measured at that time (0.017 241 $\mu\Omega$ m). Subsequent improvements in refining have raised the actual value of copper to slightly over 100% IACS. Electrical conductivity is temperature- and processing-sensitive; for copper it is 800% IACS at -240 °C and 38% IACS at 425 °C. Although copper does not become superconductive, its conductivity decreases markedly as temperature approaches absolute zero. Cold working of annealed copper to about 90% reduction in area results in a slight drop in electrical conductivity of 2–3% IACS. Conductivity is also affected by impurities or alloying elements.

Impurity content characteristics of high-conductivity copper are given in Table 1, while Table 2 summarizes some of the key physical and mechanical properties of such high-purity copper. Below we briefly summarize a number of the useful and interesting characteristic properties of copper.

2.1 Fatigue

Fatigue strength, usually specified at a life of 10^8 cycles, is enhanced by fine grain size and the absence of nonmetallic inclusions. The techniques used to strengthen coppers (see Sec. 2.3.1) are also effective in increasing resistance to fatigue.

2.2 Magnetic Properties

Copper and most of its alloys are nonmagnetic. Copper is a diamagnetic material and has a mass susceptibility (χ) of -0.086×10^{-6} cgs units per gram at room temperature. This is equivalent to a gram-atomic susceptibility of -5.1×10^{-6}. Susceptibility varies with increasing temperature; for example, it is -0.085 at 300 °C, -0.082 at 600 °C, and -0.079 at 900 °C.

2.3 Formability

Copper and its alloys are readily hot worked by all of the usual processes, and may be fabricated into end-use forms by rolling, bending, drawing, stretching, and coining. All operations that deform the metal result in work hardening to various extents depending on chemical composition and previous annealing or fabricating history.

Casting to shape, using sand or permanent molds, is a widely used process for copper and many of its alloys, including many specifically designed for this purpose.

2.3.1 Strengthening. In addition to cold working, copper alloys can be metallurgically strengthened by solid solution hardening, phase transformations, precipitation hardening (including spinodal decomposition), and grain size refinement.

Table 1. Chemical composition of some coppers.

Composition	C10100 Oxygen-free Electronic[a]	C10200 Oxygen-free[a]	C11000 Electrolytic Tough Pitch[b]
Copper (including silver) (min %)	99.99	99.95	99.90
Trace element (ppm max)			
Antimony	4	—	—
Arsenic	5	—	—
Bismuth	1	—	—
Cadmium	1	—	—
Iron	10	—	—
Manganese	0.5	—	—
Mercury	1	—	—
Nickel	10	—	—
Oxygen	5	10	Variable
Phosphorus	3	—	—
Selenium	3	—	—
Silver	25	—	—
Sulfur	15	—	—
Tellurium	2	—	—
Tin	2	—	—
Zinc	1	—	—

[a] ASTM B 170 Standard Specification for oxygen-free electrolytic copper-refinery shapes.
[b] Impurity elements vary according to process.

Solid solution hardening is achieved by short-range order, chemical interaction, lattice misfit, electrical and valence interaction, or a combination of these mechanisms.

Precipitation hardening is used primarily for chromium- and beryllium-containing coppers. The latter are the strongest generally available copper alloys, exceeding 1100 MPa tensile strength compared to 700–965 MPa reached by other forms of treatment.

Grain size refinement hardens copper alloys while retaining a high degree of ductility and toughness. A low-temperature anneal following cold working produces grain refinement in wrought material such as strip or wire.

2.4 Corrosion and Oxidation

Copper corrodes at low rates in unpolluted air, water, and deaerated nonoxidizing acids. It is not resistant to oxidizing acids, aerated nonoxidizing acids, oxidizing heavy-metal salts, sulfur, hydrogen sulfide, ammonia, and some sulfur and ammonia compounds, the latter of which forms highly soluble complexes with copper.

Copper and its alloys resist corrosion because of their low driving force toward the oxidized state and because their corrosion products form adherent protective films in some media. Under most outdoor conditions, the film eventually forms a green patina by reacting with moisture and oxides of sulfur in the air to form an artificial mineral, brochantite [$Cu_2SO_4 \cdot 3Cu(OH)_3$]. The actual color of the patina, ranging from blue-green to yellow- or brown-green will vary with the atmospheric environment to which the copper is exposed. The presence of sulfides in the atmosphere tends to blacken the patina and the presence of acids tends to dissolve it. Hence, the status of the patina on copper roofs, etc., is a good indicator of the degree of air pollution present.

Pure copper oxidizes in air at temperatures of approximately 250 °C and a film of Cu_2O is formed. Up to 100 °C, the oxide film grows linearly with the logarithm of time. From 100 to 1025 °C, the scale consists of an inner layer of Cu_2O and an outer layer of CuO. Oxidation behavior of alloys depends on alloying elements.

Hydrogen embrittlement at high temperatures occurs when hydrogen diffuses into copper and reacts with oxygen to form water vapor under high pressure. Subsequent defor-

Table 2. Physical and mechanical properties of oxygen-free copper C10100.

Property	English units	Metric units
Melting point	1981 °F	1083 °C
Electrical resistivity (annealed)[a]	10.3 Ω (circ mil/ft)@68 °F	1.71 $\mu\Omega$ cm@20 °C
Electrical conductivity (annealed)[b]	101% IACS@68 °F	0.586 MS cm@20 °C
Thermal conductivity	226 Btu/ft^2 ft h °F@68 °F	0.934 cal/cm^2 cm s °C@20 °C
Coefficient of thermal expansion	9.4×10^{-6}/°F (68–212 °F)	17.0×10^{-6}/°C (20–100 °C)
	9.6×10^{-6}/°F (68–392 °F)	17.3×10^{-6}/°C (20–200 °C)
	9.8×10^{-6}/°F (68–572 °F)	17.7×10^{-6}/°C (20–300 °C)
Specific heat (thermal capacity)	0.092 Btu/lb °F@68 °F	0.092 cal/g °C@20 °C
Density	0.323 lb/in.^{3}@68 °F	8.94 g/cm^3@20 °C
Specific gravity	8.94	8.94
Modulus of elasticity (tension)	17 000 ksi	117 000 MPa
Modulus of rigidity	6,400 ksi	44 000 MPa
Tensile strength	42 ksi	290 MPa
Yield strength, 0.1% offset	36 ksi	248 MPa
Fatigue strength[c,d]	17 ksi	117 MPa
Elongation, % in 2 in.	14	14

[a] Maximum resistivity of Certified Grade.

[b] Minimum conductivity of Certified Grade (volume basis). Mechanical properties are for flat products, one-half hard temper.

[c] Rod, drawn to 35% reduction in area.

[d] At 100 million cycles.

mation results in cracking. Embrittlement is most common in oxygen-containing copper, typically C11000 (numbering system designates chemical composition, see Sec. 6.1), electrolytic tough pitch copper commonly used for applications where the highest electrical conductivity is not a requirement. Embrittlement is avoided by the use of oxygen-free or deoxidized grades.

2.5 Joining

Copper is easily soldered and brazed and can be welded. Mild fluxes or cleaning in a diluted acid bath is used to remove tarnish films. Commercial solders of lead plus tin in the range of 35–60% achieves a superior metallurgical bond, although such solders are prohibited in many areas for use in potable-water plumbing systems because, under certain conditions, the lead can leach into water. For plumbing use, antimony-tin, silver-tin, and copper-antimony solders are available. Copper is also easily brazed for higher-temperature applications. The welding processes most used for copper are resistance welding, gas-metal-arc (GMAW), and gas-tungsten-arc (GTAW). Electron-beam welding can be used for precision welds.

The brass alloys can be soldered, brazed, or welded. The other copper alloys are generally brazed or welded.

3. RESOURCES AND RAW MATERIALS

For at least 6000 years, copper has been one of the most important materials in the development of civilization and culture. The earliest copper objects were made by beating and hammering native copper, a form of metallic copper found mixed with copper-bearing ore. The earliest known copper object to have been melted and cast, a mace head dated at 5000 BC, was found at Can Hasan in Anatolia (now Turkey). Archaeological evidence shows that copper ore was first smelted between 4000 and 3500 BC, thus marking the beginning of the science of metallurgy. Current theory is that the smelting process evolved from observations of the reduction of copper minerals in the glaze used in pottery. Thus, the first smelting furnaces were undoubtedly pottery kilns. In 1965 archaeologists excavated the earliest furnace designed specifically for smelting copper at Timna in the southern Sinai in Israel. The furnace was dated at 3200 BC. Charcoal had been the reducing agent and iron oxide the fluxing agent.

By 3000 BC, silver, tin, and lead were being used, and copper began to be alloyed with these other metals. Alloying with tin produced bronze, a hard alloy used for tools, plows, weapons, armor, and decorative objects. Brass, an alloy of copper and zinc, has been in use for almost 2000 years. The first intentional use of brass was by the Romans, who used it for coins, ornaments, and decorative objects.

There was a thriving trade in bronze by 1500 BC, which gave this era the name of the Bronze Age. In the ancient world, copper was mined in Cyprus and other Middle Eastern sites; however, since there are no known tin deposits at the bronze-making centers of this region, the tin had to have been transported to the region. Evidence exists that it could have come from as far away as Cornwall, where it was transported from Penzance by the Phoenicians—the great metal traders of that era. Since the tin had to have come from a distant site, it is clear that the Bronze Age could not have started until a reliable network of trading contacts had been established.

By 1100 BC, iron had replaced bronze in most functional uses. The Iron Age came about both because of iron's wider availability and because swords made from iron were proven to be superior to those of bronze when the Hittites conquered Assyria and captured Babylon in 1300 BC.

The modern history of copper begins with the discovery of electricity, and electrical applications continue to be one of its principal uses. Copper ranks third in metal consumption after steel and aluminum. Copper is mined in about 58 countries, with the major producing countries accounting for nearly 95% of world production. The major producing countries, in order of production in 1990, were Chile, United States, Soviet Union, Canada, Zambia, Poland, China, Zaire, Australia, Peru, and Mexico. The leading copper refining countries in 1990 were United States, Soviet Union, Chile, Japan, Germany, China, Canada, Zambia, Belgium, and Poland. World mine production was 9 100 500 metric tons in 1990. Of this amount, the United States produced 1 565 100 metric tons. Despite copper's long history, about 75% of all copper ever mined has been consumed since World War II. Copper ore reserves in the United States and the world are considered more than adequate to meet cumulative demand in the foreseeable future.

4. PRODUCTION

In present-day mining operations, large quantities of low-grade ore are crushed and loaded into trucks or conveyor belts for delivery to the ore treatment plant. In the United States, the ore averages about 0.6% copper, the richer values having long since been mined out. The ore is treated by grinding and froth flotation to produce a beneficiated product containing about 25% copper. These concentrates are reduced to the metallic state by smelting to produce copper matte containing up to about 60% copper. The matte is converted by selective oxidation to blister copper containing 98.5% or more copper. This intermediate product is refined electrolytically into copper cathodes containing 99.9% or more copper.

In recent years, more and more mines are leaching waste dumps and oxide ores for recovery of copper. Sulfuric acid, recovered from smelter off-gases as part of air-pollution control, is used as the leachant. Acid liquors containing copper are purified by solvent extraction, and the copper is recovered, in the form of cathodes, by electrowinning.

Cathodes from both the electrolytic refining process and the electrowinning process are, for the most part, melted, cast, and rolled into wire rod—a precursor for drawing into wire. The remainder is melted and cast into cakes or billets for conversion into semifabricated forms by manufacturers of wrought copper and copper alloys. Approximately 75% of all newly mined copper and 50% of all copper (including scrap) is made into wire and cable by wire mills. Most of the remainder is made into sheet, strip, plate, rod, bar, pipe, tube, wire, forgings, and extrusions by brass mills. A small amount is used in castings by foundries or manufactured into powder. Certain mill products, such as wire and most tubular products, are used without further fabrication. Most flat-rolled products, bar, rod, mechanical wire, and forgings go through multiple forming, finishing, machining, and assembly operations before their ultimate use in a manufactured product.

Free-cutting brass rod, which exhibits outstanding machinability and good corrosion resistance, and brass strip, which has high strength, corrosion resistance, excellent formability, and good electrical properties, constitute more than three-fourths of the total tonnage of copper alloys shipped from U.S. brass mills. Other alloy types of major commercial significance include tin bronzes (phosphor bronzes), which are noted for their excellent cold-forming behavior and strength; tin brasses, known for outstanding corrosion resistance; copper nickels, which are strong and particularly resistant to seawater; nickel silvers, which combine a silvery appearance with good formability and corrosion resistance; beryllium coppers, which provide outstanding strength when hardened; and aluminum bronzes, which have high strength along with good resistance to oxidation, chemical attack, and mechanical abrasion.

Foundries use prealloyed ingot, scrap, and virgin metal as raw materials. Their chief products are shaped castings for many different industrial and consumer goods, the most important of which are plumbing products and industrial valves. Centrifugal- and continuous-cast products find major application as bearings, cylinders, and other symmetrical components. Powder plants produce powder and flake for further fabrication into powder metallurgy parts, chiefly small sintered bronze bushings for bearings.

In 1990, the United States consumed 3 064 000 metric tons of copper mill and foundry products. The total supply of copper and its alloys (wire mill, brass mill, foundry, and powder products) to U.S. domestic users totaled 6643 million pounds (metal content including alloying ingredients). Of this, the amounts to the five major end-use markets were (in millions of pounds) building construction, 2627; electrical and electronic, 1723; industrial machinery and equipment, 923; transportation equipment, 746; and consumer and general products, 624.

A large proportion of copper supply is provided by scrap. This recycled material is responsible for about 45% of the total copper supply in the United States.

In 1990, the United States was self-sufficient in copper, exporting 400 000 metric tons of concentrates and refined copper more than it imported, and exporting 4400 metric tons of scrap more than the 95 600 metric tons of copper contained in imported semifabricated products it imported. Imports of copper metal, in the form of primary refinery shapes or mill products, were about 7% of U.S. total supply.

5. COPPER AND ALLOY TERMINOLOGY AND STANDARDS

To ensure orderly identification of specific wrought and cast copper chemical compositions, a six-digit Unified Numbering System is used in the United States and Canada. This system has been adopted by North American specifying bodies such as the American Society for Testing and Materials (ASTM), Society for Automotive Engineers (SAE), and various federal organizations in specifying standards for copper and copper alloy products.

Besides the ASTM standards used in the United States and Canada, other national specifications are BS in the United Kingdom and DIN in Germany.

In the designation system, numbers from C10000 through C79999 denote wrought alloys. Cast alloys are numbered from C80000 through C99999. Within these two categories, the compositions are grouped into the following families of copper and copper alloys:

Coppers (C10000–C15999 and C80000–C81299): Metal that has a designated minimum copper content of 99.3% or higher.

High-Copper Alloys (C16000–C19999 and C81300–C82999): For the wrought products, these are alloys with designated copper contents less than 99.3% but more than 96% that do not fall into any other copper alloy group. The cast high-copper alloys have designated copper contents in excess of 94%, to which silver may be added for special properties.

Brasses: These alloys contain zinc as the principal alloying element with or without other designated alloying elements such as iron, aluminum, nickel, and silicon. The wrought alloys comprise three main families of brasses: copper-zinc alloys (C20500–C28599); copper-zinc-lead alloys (leaded brasses) (C31200–C38599); and copper-zinc-tin alloys (tin brasses) (C40400–C49999).

The cast alloys comprise four main families of brasses: copper-tin-zinc alloys (red, semi-red, and yellow brasses) (C83300–C85899); "manganese bronze" alloys (high-strength yellow brasses)(C86100–C86399); leaded "manganese bronze" alloys (leaded high-strength yellow brasses)(C86400–C86899); and copper-zinc-silicon alloys (silicon brasses and bronzes) (C87300–C87999).

Bronzes: Broadly speaking, bronzes are copper alloys in which the major alloying element is not zinc or nickel. Originally "bronze" described alloys with tin as the only or principal alloying element. Today, the term is generally used not by itself but with a modifying adjective. For wrought alloys, there are four main families of bronzes: copper-tin-phosphorus alloys (phosphor bronzes)(C50100–C52499); copper-tin-lead phosphorus alloys (leaded phosphor bronzes) (C53200–C54899); copper-aluminum alloys (aluminum bronzes) (C60600–C64499); and copper-silicon alloys (silicon bronzes) (C64700–C66199).

The cast alloys have four main families of bronzes: copper-tin alloys (tin bronzes) (C90200–C91799); copper-tin-lead alloys (leaded and high-leaded tin bronzes) (C92200–C94599); copper-tin-nickel alloys (nickel-tin bronzes) (C94700—C94999); and copper-aluminum alloys (aluminum bronzes) (C95200–C95900).

The family of alloys known as "manganese bronzes," in which zinc is the major alloying element, is included in the brasses, above.

Copper-Nickels (C70100–C71999 and C96200–C96999): These are alloys with nickel as the principal alloying element, with or without other designated alloying elements.

Copper-Nickel-Zinc Alloys (C73000–C79999 and C97300–C97999): Known commonly as "nickel silvers," these are alloys that contain nickel and zinc as the principal and secondary alloying elements, with or without other designated elements. They contain no silver.

Leaded Coppers (C98200–C98999): These comprise a series of cast alloys of copper with 20% or more lead, with usually a small amount of silver present, but without tin or zinc.

Special Alloys (C99300–C99999): Alloys whose chemical compositions do not fall into any of the above categories.

6. APPLICATIONS

Copper and its alloys are widely used commercially because of their excellent electrical and thermal conductivity, resistance to corrosion, and ease of fabrication, with good strength and fatigue resistance. Pure copper is used for electrical applications such as wire and cable, bus bars, electrical contacts, and any part required to conduct electricity. Copper and certain brasses, bronzes, and copper nickels are used for automobile radiators, heat exchangers, home heating systems, solar-energy panels and other applications requiring high thermal conductivity. Copper, brass, some bronzes, and copper nickel are used for corrosion-resistant pipe, valves, and fittings in potable-water, process-water, and other aqueous fluid systems. The copper-nickel and bronze alloys are most noted for their application in marine systems. Copper and the brass and bronze alloys are widely used in architecture, for both functional and decorative purposes, in such applications as roofing and guttering, locks and hinges, railings, and window and door mullons.

Copper chemicals, such as copper sulfate, carbonate, oxychloride, and oxides, are used in a wide range of applications: as an algicide in water treatment; as a micronutrient in chicken and cattle feed and on copper-deficient agricultural soil; as a fungicide and bactericide on crops grown in tropical areas; as a wood preservative; in antifouling marine paint; and as oxidation catalysts in petroleum product and organic chemical manufacture.

Copper and copper alloys are used in a number of newly developed technologies. For example, copper is an important material in the use of semiconductors and microcircuitry, where it is used in lead frames and in printed circuit boards. It is used as the stabilizer in superconducting cables. Here the superconducting alloy filaments are embedded in a matrix of pure copper. In this role, the copper acts as the electricity conductor when the filaments are not in their superconducting state and acts as a heat transfer agent to cool the filaments to the temperature at which they are superconducting. In the conventional, low-temperature system, the superconducting filaments are niobium-titanium or niobium-tin alloy. In the new high-temperature system, the superconducting filaments are ceramics that may themselves contain copper oxide.

Copper has also made possible the use of a number of new technologies, and its use is growing as a result. For example, in telecommunications the FAX machine would not be possible without copper wire, and increasing use of FAX for information transmission has greatly increased the use of copper telecommunications cable. Likewise, more and more automation and instrumentation is being used in automobiles. This has been made possible by copper wire and, as a result, has significantly increased the use of copper in an average automobile, now standing at approximately 45 lb per vehicle.

New uses are continually being developed for copper and its alloys and compounds. Among these are applications that take advantage of its bacteriostatic properties such as for clean surfaces in hospitals and public buildings; as a biofouling- resistant surface on ships and other marine structures; in fire safety sprinkler systems and gas distribution systems; and for pharmaceuticals used in diseases where tissue damage is involved.

7. SUMMARY

Copper has been one of the most important materials in the development of civilization, and even today is a significant factor in the development and application of new technologies. The modern growth of the industry is linked to the widespread use of electricity and the continual development of new ways to use electricity. Copper's high electrical and thermal conductivity, formability, and corrosion resistance account for its major applications today.

List of Works Cited

Callaway, J. (1964), *Energy Band Theory,* New York: Academic.

Cooper, B. R., Ehrenreich, H., Philipp, H. R. (1965), *Phys. Rev.* **138**, A494–A507.

Ehrenreich, H., Philipp, H. R. (1962), *Phys. Rev.* **128**, 1622–1629.

Segall, B. (1962), *Phys. Rev.* **125**, 109–122.

Ziman, J. M. (1960), *Electrons and Holes: The Theory of Transport Phenomena in Solids,* Oxford: Oxford University Press.

Further Reading

AMAX Copper, Inc. (1974), *OFHC Copper: A Survey of Properties & Applications,* New York: AMAX Copper, Inc.

American Society for Metals (1979), *Metals Handbook,* Vol. 2: *Properties and Selection: Nonferrous Alloys and Pure Metals,* Materials Park, OH: ASM.

American Society for Testing and Materials (1991), *Anual Book of ASTM Standards,* Vol. 02.01: *Copper and Copper Alloys,* Philadelphia: ASTM.

Butts, A. (1954), *Copper, The Science and Technology of the Metal, Its Alloys and Compounds,* New York: Reinhold.

Copper Development Association Inc. (1975), *Low-Temperature Properties of Copper and Copper Alloys,* Application Data Sheet No.104/5, New York: Copper Development Association Inc.

Copper Development Association Inc. (1990), *Standard Designations for Copper and Copper Alloys,* New York: Copper Development Association Inc.

International Copper Association, Ltd. (1973), *Cryogenic Properties of Copper* (wall chart), New York: International Copper Association.

Nonferrous Metal Works of the World, 5th ed. (1989), New York: Metal Bulletin Inc. (Lists world copper and brass suppliers by country.)

Touloukian, Y. S. (1967), *Thermophysical Properties of High Temperature Solid Materials,* New York: Macmillan.

U. S. Bureau of Mines (1985), Copper chapter in *Mineral Facts and Problems,* Bull. No. 576, Washington, DC: U.S. GPO.

U. S. Bureau of Mines, *Minerals Yearbook,* Published annually, Washington, DC: U.S. GPO.

Ziman, T. M. (Ed.) (1969), *The Physics of Metals,* Vol. 1: *Electrons,* Cambridge, UK: Cambridge University Press.

COSMOLOGY

See ASTROPHYSICS
GALAXIES AND COSMOLOGY

CRITICAL PHENOMENA

CYRIL DOMB, *Department of Physics, Bar-Ilan University, Ramat-Gan, Israel*

Introduction ... 277

1. **The Classical Period** ... 279

1.1 Liquid–Gas Critical Point ... 279

1.2 Magnet–Fluid Analogy: Curie Point of a Ferromagnet ... 281

1.3 Microscopic Critical Behavior of Fluids ... 282

1.4 Critical Behavior of Binary Alloys ... 283

1.5 Universality of Second-Order Transitions (Landau) ... 284

2. **Ising Model: Exact Solutions** . 285

2.1 Ising Model ... 285

2.2 The Lattice-Gas Model ... 287

2.3 Transfer Matrix ... 287

2.4 Onsager's Solution for the Rectangular Lattice ... 288

2.5 Other Exact Results ... 290

3. **Three-Dimensional Models** ... 290

3.1 The Challenge of Onsager's Solution ... 290

3.2 More General Models ... 290

3.3 Series Expansions ... 291

3.4 Critical Exponents ... 293

3.5 Spherical and Gaussian Models ... 294

3.6 Systems with Weak Long-Range Forces ... 295

3.7 Experimental Results ... 295

3.8 Conclusion ... 298

4. **Scaling and Universality** ... 298

4.1 Reconciliation ... 298

4.2 Equation of State in the Critical Region: Scaling ... 299

4.3 Universality ... 299

4.4 Crossover Behavior ... 300

4.5 Kadanoff's Treatment ... 302

5. **Renormalization Group** ... 302

5.1 Background Ideas ... 302

5.2 RG Derivation of Critical Behavior ... 304

5.3 Real-Space Renormalization .. 305

5.4 Momentum-Space Renormalization ... 305

5.5 Critical Exponents: Corrections to Scaling ... 307

5.6 Conclusion ... 307

6. **Miscellaneous Additional Topics** ... 307

Glossary ... 308

Works Cited ... 309

Further Reading ... 309

Review Articles ... 309

Books ... 309

INTRODUCTION

If the thermodynamic properties of matter in equilibrium are surveyed, it will be found that they fall into two groups: those that are smooth and continuous, and those that have sharp discontinuities. As examples of the first group we may cite the properties of ideal or nearly ideal gases (energy, entropy, specific heat, and equation of state), of ideal or nearly ideal solids, of ideal or nearly ideal mixtures of gases or solids, paramagnetism and diamagnetism, and properties of electrons and phonons in normal metals. The second group is usually associated with phase transitions of various types: liquid–vapor equilibrium and the critical point, the melting of solids, phase separation in liquid or solid mixtures or solutions, order–disorder transitions in alloys, ferromagnetism, antiferromagnetism, superconductivity, and λ-point anomalies (e.g., liquid helium).

Standard statistical mechanics (see STATISTICAL MECHANICS, CLASSICAL; STATISTICAL MECHANICS, QUANTUM) can deal with the first group with relative ease. The macroscopic body under consideration can be decomposed into microscopic constituents (molecules, harmonic oscillators, spins, and electrons) that can to a good approximation be treated as noninteracting. Of course there must be interactions present to ensure that thermal equi-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

librium is attained, but in describing the behavior of the macroscopic bodies these interactions can be ignored. The energy levels of the macroscopic body can be easily related to the energy levels of the microscopic constituents, and with the aid of some elementary combinatorial analysis, a partition function can be constructed from which the equilibrium behavior of the macroscopic bodies can be derived. Elementary mathematical considerations show that the thermodynamic properties will be continuous.

In general, the discontinuities described in the properties of the second group above originate in *interactions* between the microscopic constituents, and the problem of how to take such interactions into account in calculating a partition function will be central to our discussions. Much of the mathematics involved in a proper treatment of interactions is specialized and sophisticated. The major aim of the present article will be to convey the results and significance of the calculations, which, we believe, can be understood by the general physicist even if the detailed mathematical techniques are beyond him or her.

When the interactions between the constituents are weak, PERTURBATION METHODS (*q.v.*) can be used, but as long as only a finite number of terms are considered, the continuity of the thermodynamic properties will not be destroyed. Discontinuous behavior can be introduced only by taking the perturbation series to infinity so that the mathematical character of the partition function is changed. In fact, the problem to be tackled in dealing with phase transitions is the "strong interaction" problem in which the interactions can no longer be treated as a small perturbation, but play a dominant role in the calculations and in the resulting physical properties.

We will be largely concerned with behavior near critical points of fluids, magnets, and solutions, and with various types of λ-point transition. The past two decades have seen great progress in the theoretical understanding of how this behavior depends on the nature of the intermolecular interactions, e.g., their symmetry and range. We shall endeavor to describe this progress and bring the reader up to the frontiers of current research.

Our treatment will follow the historical development of the subject, and we shall divide this history into four periods:

1869–1944	The classical period
1944–1965	The Onsager revolution
1965–1971	Reconciliation of Onsager with van der Waals
1971–	Renormalization group

In the first period the term classical is used in the sense of "well established" and not with any relationship to quantum theory. The subject starts with the introduction of the term "critical point" by Thomas Andrews in his Bakerian Lecture of 1869 to the Royal Society. Andrews had used the term before in 1863, but the exposition of the true nature of the critical point of a fluid took place in the Bakerian Lecture in which he discussed in detail the experimental behavior of the isotherms of carbon dioxide. A few years later van der Waals provided a theoretical explanation of Andrews's results. The classical period is characterized by the interplay of experiment and theory, and the introduction of new concepts that seemed adequate to provide a qualitative and quantitative description of the phenomena under consideration. Developments that took place in different areas of physics seemed to have many features in common and led eventually to the idea of a unified description of all phenomena involving critical points or λ points.

The second period was initiated by a mathematical *tour de force*, the exact solution by Onsager of a realistic two-dimensional model, giving results that were in basic disagreement with the classical calculations. This period was characterized by further exact calculations of the Onsager kind and also by "down-to-earth" numerical calculations that made extensive use of computers; in addition, new experimental techniques were exploited to furnish more accurate data than had been available previously. As a result the classical theories were completely discredited in relation to quantitative predictions of critical behavior. However, no coherent new pattern was formulated to put in its place.

This led Uhlenbeck to suggest at a conference on "Phenomena in the Neighbourhood of Critical Points" at the National Bureau of Standards in April 1965 that the most important task facing workers in the field was to reconcile Onsager with van der Waals. Indeed, shortly after this, a wider framework emerged into which the new theoretical and experimental results seemed to fit. During the period 1965–1971 this framework was explored in more detail, and the pattern of behavior was summarized most succinctly in the hypothesis of universality. However, the pattern was obtained empirically — there was no theoretical explanation.

The latest phase supplied the theoretical understanding that has made the field "respectable." It was initiated in 1971 by Kenneth Wilson, who suggested a brilliant application of the renormalization group, a concept taken from field theory, to behavior near a critical point. The burst of research activity stimulated by Wilson's ideas has opened up many new avenues of exploration; Wilson's work received rapid recognition and he received a number of important awards culminating in the Nobel prize for physics in 1982. The recognition accorded Wilson is also a recognition of the field of critical phenomena and of the work of all those who helped build the structural foundation on which the renormalization group could base itself.

1. THE CLASSICAL PERIOD

1.1 Liquid–Gas Critical Point

The relation between gases and liquids was the subject of much attention during the first half of the 19th century. It was realized quite early that the liquid phase ceased to exist above a certain temperature, but the general assumption was that it disappeared into the gaseous phase. Thus Faraday talked of the "disliquefying point," while Mendelev used the term "absolute boiling temperature" for the point at which the latent heat of evaporation becomes zero.

In addition to careful and accurate experiments on carbon dioxide described in the Bakerian Lecture of 1869, Thomas Andrews introduced a new concept of symmetry between the liquid and gaseous phases; the two phases merged at the critical point into one fluid phase, "but if any one should ask whether it is now in the gaseous or liquid state, the question does not, I believe, admit a positive reply." He emphasized this feature in the title of his lecture "On the Continuity of the Gaseous and Liquid States of Matter," and pointed out how it was possible, by suitable choice of path, to pass from the liquid to the gaseous phase without any discontinuity. Figure 1 represents the isotherms of carbon dioxide taken from Andrews's Bakerian Lecture (1869). The difference of behavior below and above the critical temperature is manifest. Andrews estimated the critical temperature at 30.92 °C.

Only four years elapsed before van der Waals used the newly developing ideas on the kinetic theory of gases to give a plausible theoretical explanation of Andrews's experimental data. Van der Waals assumed that a gas is made up of molecules with a hard core and a long-range mutual attraction. The range of the attractive forces was assumed to be long compared to the mean free path, and they give rise to a negative "internal pressure." To calculate its value van der Waals made use of the virial theorem that Clausius had introduced in 1870 only a few years before, which states that

$$\left\langle \sum_i \tfrac{1}{2} m v_i^2 \right\rangle = -\tfrac{1}{2} \left\langle \sum_i (x_i X_i + y_i Y_i + z_i Z_i) \right\rangle , \qquad (1)$$

where v_i is the velocity, (x_i, y_i, z_i) the coordinates, and (X_i, Y_i, Z_i) the force on the ith molecule. He found that

$$P_{\text{internal}} = 1/v^2, \qquad (2)$$

where $v = V/N$ is the volume per molecule. For the hard core he made the simplest assumption that the available volume is reduced from v to $v-b$. Hence the equation he put forward was

$$P = P_{\text{internal}} + RT/(v-b), \qquad (3)$$

$$(P + a/v^2)(v-b) = RT. \qquad (4)$$

Van der Waals was a doctoral student, and the above ideas were put forward in his thesis in Dutch. Their importance was quickly recognized by Maxwell, who reviewed the thesis in *Nature* in 1874, and in a lecture to the Chemical Society in 1875 he wrote as follows:

"The molecular theory of the continuity of the liquid and gaseous states forms the subject of an exceedingly ingenious thesis by Mr. Johannes Diderick van der Waals, a graduate of Leyden. There are certain points in which I think he has fallen into mathematical errors, and his final result is certainly not a complete

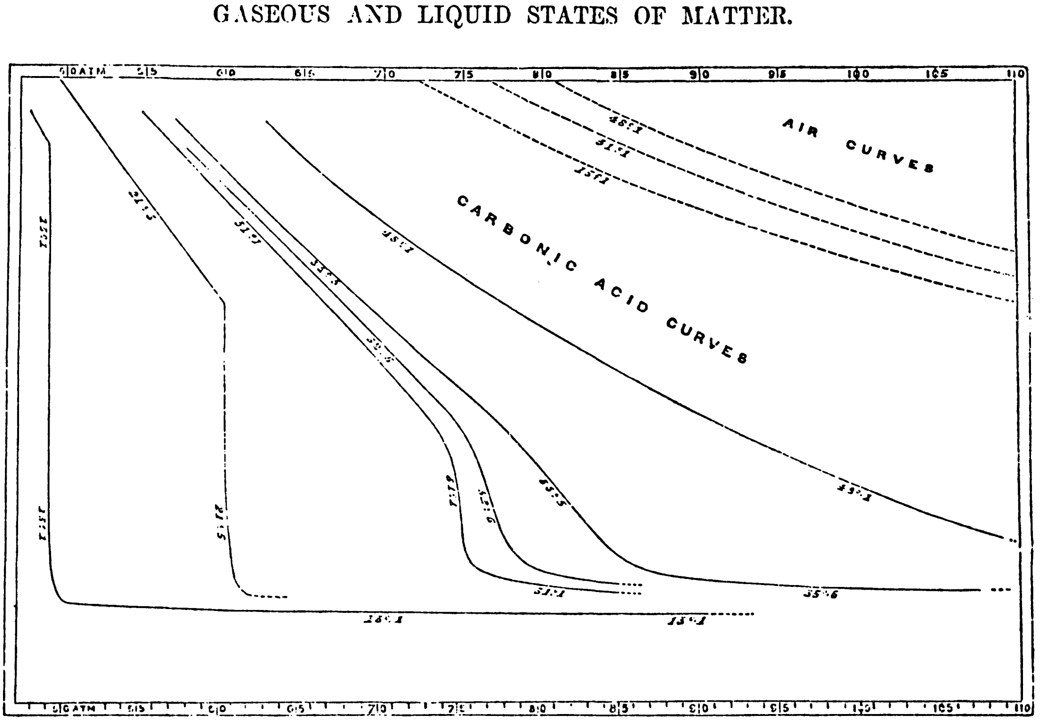

FIG. 1. Isotherms of carbon dioxide taken from Andrews's Bakerian Lecture to the Royal Society in 1869. Andrews's estimate for the critical temperature was 30.92 °C.

expression for the interaction of real molecules, but his attack on this difficult question is so able and so brave, that it cannot fail to give a notable impulse to molecular science. It has certainly directed the attention of more than one inquirer to the study of the Low-Dutch language in which it is written."

The form of isotherms given by van der Waals's equation (4) is shown in Fig. 2. In order to compare them with the experimental results of Andrews, it was necessary to decide where to draw the horizontal line that represents the liquid and gas in equilibrium. Here Maxwell put forward his "equal area" construction, which is illustrated in Fig. 3.

Boltzmann too was impressed by van der Waals's results, and although he said that "unfortunately, van der Waals had to abandon mathematical rigour at a certain point in order to carry out his calculation," he was nevertheless impressed by the practical value of his theory.

It is significant that both Maxwell and Boltzmann were reasonably happy with the attractive a/v^2 term; their main anxiety was the repulsive term.

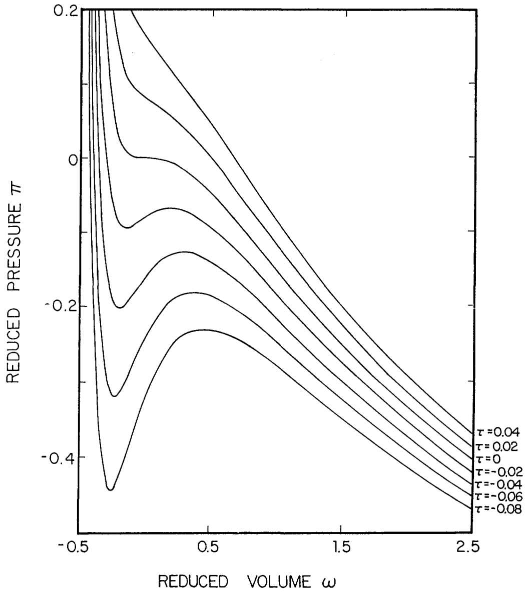

FIG. 2. Isotherms corresponding to van der Waals's equation (4). Reduced pressure $\pi(=P/P_c)$ is plotted against reduced volume $\omega(=V/V_c)$.

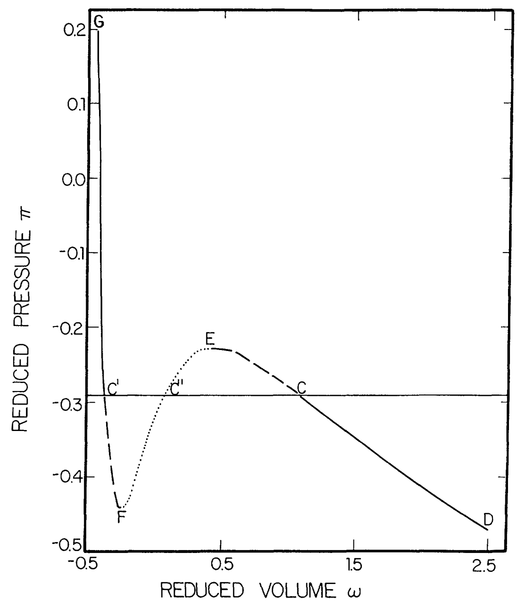

FIG. 3. Maxwell's "equal area" construction to determine the horizontal part of the isotherm in van der Waals's equation. Area $C'FC''$ = Area $C''EC$. The construction was proposed in a lecture to the Chemical Society in 1875.

The new concept of "internal pressure" that van der Waals introduced was to bear fruit 30 years later in a completely different area of physics.

1.2 Magnet–Fluid Analogy: Curie Point of a Ferromagnet

The fact that a magnet loses its magnetic power at high temperatures was noted by Gilbert in his famous treatise "De Magnete" in 1600. A more detailed quantitative investigation started in the 19th century involving Faraday in the 1830s, Barrett in 1874, Bauer in 1880, and Hopkinson in 1889, who was the first to introduce the term "critical temperature" for "the temperature at which the magnetism disappears." But the definitive paper for magnets comparable to that of Andrews for fluids was written by Curie in 1895 before he started his more famous investigations of radioactivity.

One of the most interesting new ideas put forward in this paper is the analogy between magnets and fluids. Taking pressure P as the analog of magnetic field and density ρ as the analog of magnetization, Curie points out the close similarity between the P-ρ and M-H isotherms (Fig. 4). The paramagnetic state at high temperatures corresponds to the gaseous phase and the ferromagnetic state at low temperatures to the liquid phase. Curie says that this analogy could be used to suggest new and useful experiments, and he poses the question whether there exists a precisely defined critical point with associated critical constants for a ferromagnet analogous to a fluid.

It was this analogy that led Pierre Weiss in 1907 to postulate his *molecular field* hypothesis, in which he assumed that the mutual interactions between the molecules could be replaced by a uniform field nM proportional to the magnetization in the same direction. He states, "One may give to nM the name *Internal Field* to mark the analogy with the internal pressure of van der Waals."

For the magnetic susceptibility of a gas, Curie had discovered experimentally the inverse temperature dependence

$$\chi = C/T, \tag{5}$$

and Langevin had applied the techniques of statistical mechanics to explain this result theoretically (see PARAMAGNETISM). For an ideal gas of molecules each having magnetic moment μ_0, Langevin derived the magnetic equation of state (analogous to $PV=RT$ for an ideal gas)

$$M = L(\mu_0 H/kT), \tag{6}$$

where $L(x)$ is the Langevin function $\coth x - x^{-1}$.

For a ferromagnet Weiss put forward the simple modification

$$M = L[\mu_0(H + H_{\text{int}})/kT] \quad (H_{\text{int}} = nM), \tag{7}$$

and this led to far reaching conclusions. There is indeed a sharply defined critical point analogous to that for a fluid; Weiss and Kamerlingh Onnes later termed this the *Curie point* in memory of Curie, who had been killed in an accident in 1906. Below the Curie temperature there is a nonzero spontaneous magnetization; in the paramagnetic state above the Curie temperature T_c, Curie's relation (5) for the susceptibility is modified to

$$\chi = C/(T - T_c), \tag{8}$$

and this is usually called the Curie–Weiss law.

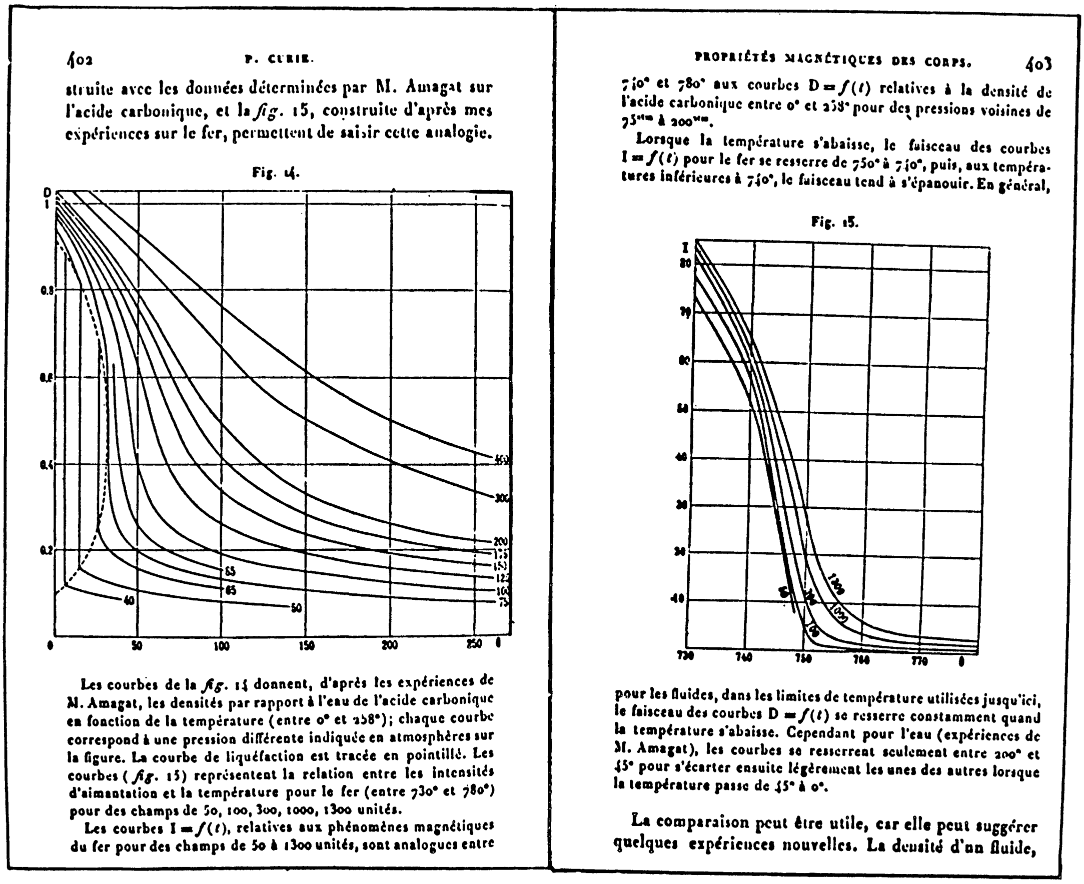

402 P. CURIE.

struite avec les données déterminées par M. Amagat sur l'acide carbonique, et la *fig.* 15, construite d'après mes expériences sur le fer, permettent de saisir cette analogie.

Fig. 14.

Les courbes de la *fig.* 14 donnent, d'après les expériences de M. Amagat, les densités par rapport à l'eau de l'acide carbonique en fonction de la température (entre 0° et 258°); chaque courbe correspond à une pression différente indiquée en atmosphères sur la figure. La courbe de liquéfaction est tracée en pointillé. Les courbes (*fig.* 15) représentent la relation entre les intensités d'aimantation et la température pour le fer (entre 730° et 780°) pour des champs de 50, 100, 300, 1000, 1300 unités.

Les courbes $I = f(t)$, relatives aux phénomènes magnétiques du fer pour des champs de 50 à 1300 unités, sont analogues entre

PROPRIÉTÉS MAGNÉTIQUES DES CORPS. 403

740° et 780° aux courbes $D = f(t)$ relatives à la densité de l'acide carbonique entre 0° et 258° pour des pressions voisines de 75^{atm} à 200^{atm}.

Lorsque la température s'abaisse, le faisceau des courbes $I = f(t)$ pour le fer se resserre de 750° à 740°, puis, aux températures inférieures à 740°, le faisceau tend à s'épanouir. En général,

Fig. 15.

pour les fluides, dans les limites de température utilisées jusqu'ici, le faisceau des courbes $D = f(t)$ se resserre constamment quand la température s'abaisse. Cependant pour l'eau (expériences de M. Amagat), les courbes se resserrent seulement entre 200° et 45° pour s'écarter ensuite légèrement les unes des autres lorsque la température passe de 45° à 0°.

La comparaison peut être utile, car elle peut suggérer quelques expériences nouvelles. La densité d'un fluide,

FIG. 4. The analogy between fluids and magnets noted by Curie in 1895. On the left-hand side are Amagat's density–temperature curves for CO_2, at different pressures, and on the right-hand side are Curie's own measurements of the magnetization–temperature curves for iron at different magnetic fields.

The magnet–fluid analogy that proved so fruitful seems to have been forgotten for more than 30 years until it was rediscovered in the lattice-gas model by Cernuschi and Eyring in 1939. It was fruitful again in subsequent developments after 1944.

1.3 Microscopic Critical Behavior of Fluids

The experiments of Andrews were concerned with the thermodynamic properties of bulk fluids. An experimental phenomenon that excited theoretical interest at the beginning of the 20th century was critical opalescence, in which a colorless transparent fluid suddenly becomes opaque in a narrow region of temperatures near T_c; the onset of opacity is accompanied by dramatic changes of color. (These experimental results had been known much earlier, an initial observation having been made by Andrews.) It was clear that a considerable increase in light scattering takes place near the critical point, and this was explained by Smoluchowski and Einstein as arising from large fluctuations in density as follows.

In 1910 Einstein showed how Boltzmann's formula relating the probability of any state of a system to its entropy (see STATISTICAL MECHANICS, CLASSICAL) could be used to calculate fluctuations of thermodynamic quantities. For the density he derived the formula

$$\langle \Delta\rho^2 \rangle = (\rho^2 kT/V) K_T, \tag{9}$$

where K_T is the isothermal compressibility,

$$K_T = \frac{1}{\rho}\left(\frac{\partial \rho}{\partial P}\right)_T = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T. \tag{10}$$

Assuming that a change in density is accompanied by a change of refractive index according to the Lorentz–Lorenz law

$$(\epsilon-1)/(\epsilon-2)=A\rho \tag{11}$$

and that scattering takes place randomly, Einstein came to the conclusion that the scattering of light of wavelength λ should be proportional to K_T/λ^4. Since van der Waals's equation leads to an infinite value of K_T at T_c, this seemed to provide a satisfactory explanation of critical opalescence.

It was remarkably perceptive of Ornstein and Zernike (1914, 1916) to note an inconsistency in the treatment arising from the assumption that the fluctuations in all elements of volume are independent of one another. In fact, they pointed out that there must be a correlation between different elements that increases indefinitely in range as the critical point is approached.

To treat this correlation, Ornstein and Zernike introduced a basic new function, afterwards called the "pair distribution function," which has played a central role in the theory of liquids ever since. Let $v(d\mathbf{r})$ be a random variable representing the number of molecules in a volume $d\mathbf{r}$ centered at $\mathbf{r}$. Since $d\mathbf{r}$ is very small, the probability of occupation by more than one particle being of order $d\mathbf{r}^2$ can be neglected. Hence we can write

$$\langle v(d\mathbf{r})\rangle=n_1(\mathbf{r})d\mathbf{r}, \tag{12}$$

where $n_1(\mathbf{r})$ is the density, which Ornstein and Zernike took to be a constant ρ. Similarly for the correlation between particles at points $\mathbf{r}_1$ and $\mathbf{r}_2$,

$$\langle v(d\mathbf{r}_1)v(d\mathbf{r}_2)\rangle=n_2(\mathbf{r}_1,\mathbf{r}_2)d\mathbf{r}_1d\mathbf{r}_2, \tag{13}$$

and for a homogeneous isotropic fluid $n_2(\mathbf{r}_1,\mathbf{r}_2)$ is of the form $n_2(r)$ $(r=|\mathbf{r}_1-\mathbf{r}_2|)$.

Ornstein and Zernike introduced the function

$$g(r)=n_2(r)/\rho^2, \tag{14}$$

which approaches 1 at large distances where the correlation becomes negligible. They used the fluctuation relation (9) to derive a fundamental identity between K_T and $g(r)$. But they also obtained an integral equation for $g(r)$ that they were able to solve explicitly.

The basic physical idea behind Ornstein and Zernike's treatment was to differentiate between the direct influence of molecular interactions, which should be short ranged and was represented by a function $f(r)$, and the correlation between densities, represented by $g(r)$ above, which should become long ranged as the critical temperature is approached. The integral equation relates $g(r)$ and $f(r)$.

Their calculations suggested that the correlations fall off asymptotically as

$$\frac{1}{r}\exp(-\kappa r), \tag{15}$$

where the value of κ can be determined from van der Waals's equation [$\kappa \sim (T-T_c)^{1/2}$]. Their detailed conclusions about light scattering differed from those of Einstein; near the critical temperature the λ^{-4} Rayleigh dependence on wavelength ceases to be valid, and there is a "whitening" of the scattered light. By temperature T_c the wavelength dependence has become of the form λ^{-2}.

1.4 Critical Behavior of Binary Alloys

The early years of the 20th century saw the development of x-ray diffraction as a powerful tool in the investigation of crystal structures, and compounds such as NaCl were found to have a regular ordered structure. However, the ionic bond between Na and Cl is so strong that no significant disordering occurs when the temperature is raised; the crystal melts before it disorders.

In 1919 Tammann suggested that similar ordering might occur in metallic alloys, and this was demonstrated experimentally a few years later (Bain, 1923; Johansson and Linde, 1925) by the existence of superlattice lines in the x-ray diffraction pattern of copper–gold alloys. However, for such systems there were soon indications that significant disordering takes place with an increase in temperature and that this is accompanied by an anomalous specific heat.

The mathematical description of the disordering process is usually associated with the names of Bragg and Williams. In their classic paper in 1934 they introduced a parameter S to characterize the *degree of order* and used Boltzmann's principle to calculate its behavior as a function of temperature. They found a pattern closely analogous to the Weiss theory of ferromagnetism, with S falling rapidly to zero at a critical temperature T_c and remaining zero for $T > T_c$. In fact, for a binary alloy with equal concentrations of constituents the equation derived for S was

$$S=\tanh(ST_c/T), \tag{16}$$

which is of the same form as Eq. (7) with $H=0$ and $\tanh x$ replacing the Langevin function $L(x)$.

In 1935 Bragg and Williams wrote a second paper elaborating their ideas. Because of the analogy with ferromagnetism they called T_c the *Curie temperature* of the alloy; S was now termed *long-distance order* to differentiate it from another parameter that had been introduced by Bethe to characterize the *short-range order* that persists above T_c. But they also apologized in this paper for having ignored the work of other investigators, notably Gorsky, Borelius, and Dehlinger, who had been thinking along similar lines.

The important new concept of long-range order is clearly described in the second paper, and a significant distinction is drawn between long-range order and long-range forces. It is stated clearly that short-range forces can give rise to long-range order, a conclusion exactly parallel to that of Ornstein and Zernicke for fluids described above, of which Bragg and Williams were apparently unaware.

1.5 Universality of Second-Order Transitions (Landau)

It was L. D. Landau (1937) who first attempted to provide a unified description for all second-order transitions. In addition to the phenomena described above, experimental evidence was accumulating about specific heat anomalies that were termed λ-point transitions in liquid helium, ammonium chloride, and a number of other substances, and the superconducting transition was also in this category. In 1933 Ehrenfest had introduced a thermodynamic classification of higher-order transitions, but there were difficulties associated with his treatment, and we shall avoid describing λ-point transitions as second-order phase transitions although this terminology is common.

Landau generalized the idea introduced in the theory of alloys to all systems manifesting λ-point transitions. He suggested that for every such system one must identify an *order parameter* analogous to the long-range order in an alloy, which would be zero on the high-temperature side of the transition and nonzero on the low-temperature side. He emphasized the important role that symmetry plays in phase transitions, and suggested that the important features of behavior in the vicinity of a λ point could be determined by expanding the free energy in a power series as a function of the order parameter η. In the ferromagnetic transition, the order parameter is the spontaneous magnetization, and in the fluid, the density difference between liquid and gas; Landau argued that for reasons of symmetry the form of the expansion of the Gibbs function would be

$$\phi(P,T,\eta)=\phi_0(P,T)+A(P,T)\eta^2+B(P,T)\eta^4 +\cdots, \tag{17}$$

and the Curie temperature would correspond to

$$A(P,T)=0, \quad B(p,T)>0. \tag{18}$$

Above the Curie temperature $A>0$ and the solution corresponding to a minimum of ϕ is $\eta=0$; this is the phase of higher symmetry. Below the Curie temperature the minimum of ϕ corresponds to a nonzero value of η given by

$$\eta^2=-A/2B, \tag{19}$$

and this is the phase of lower symmetry.

Landau's theory enabled one to understand why all the systems discussed previously had the same essential pattern of critical behavior; even though equations of state such as Eqs. (4) and (7) look very different, they both conform to Eq. (18). Critical exponents are the same for all λ-point transitions, and in later terminology the transitions would be described as universal.

The behavior of typical thermodynamic quantities for a ferromagnet near T_c were as follows ($t = T/T_c - 1$):

$$\begin{array}{lll} \text{spontaneous magnetization} & M_0 \sim (-t)^{1/2} & (t<0), \\ \text{initial susceptibility} & \chi_0 \sim t^{-1} & (t<0), \\ \text{critical isotherm} & H \sim M^3 & (t=0), \\ \text{derivative of susceptibility} & \dfrac{d\chi_0}{dH} \sim t^{-4} & (t>0). \end{array} \qquad (20)$$

The specific heat tended to one value C_- on approaching the transition from a lower temperature, and to another C_+ from above. The transcription to appropriate thermodynamic variables can be readily made for other systems.

Landau developed an alternative approach to that of Ornstein and Zernike to deal with critical fluctuations. His method was to seek a Taylor expansion of the excess free energy arising from variations in density (for a magnetic system it would arise from variations in magnetization, and in other systems from variations in the appropriate order parameter). We can divide the total Helmholtz free energy ΔF into contributions from different spatial regions,

$$\Delta F = \int \Delta a(\mathbf{r}) dV, \qquad (21)$$

where $a(T,\rho)$ is the local free energy expressed in terms of the density. $\Delta a(\mathbf{r})$ depends on $\delta\rho(\mathbf{r})$ and combinations of the derivatives of ρ, which are spherically symmetrical. In an expansion of $a(\mathbf{r})$ the first derivatives of ρ can enter only as $(\operatorname{grad}\rho)^2$ and the second derivatives as $\nabla^2\rho$. Terms in $\nabla^2\rho$ can be integrated to give a surface contribution that can be ignored and to give a bulk contribution of the form $(\operatorname{grad}\rho)^2$. We are thus led in the first order to an expansion of the form

$$\Delta a(\mathbf{r}) = \tfrac{1}{2}c\delta\rho^2 + \tfrac{1}{2}b(\nabla\rho)^2 + \cdots, \qquad (22)$$

where $c = (1/\rho)(\partial P/\partial\rho)_T$ is zero at T_c, and b must be greater than 0 for $\Delta a(\mathbf{r})$ to correspond to a minimum at T_c.

If $\delta\rho$ is broken down into Fourier components, then ΔF is a sum of contributions from independent normal modes. The absence of interaction between the modes is due to the truncation of expansion (22) at the square term. Assigning an average energy kT to each mode, Landau was led to the universal formula (15) for the pair correlation function.

The approximation of stopping with the gradient terms in the expansion in Eq. (22) is reasonable for long-wavelength fluctuations which are responsible for critical behavior. However, since c is zero at T_c it is reasonable to assume that a higher-order term $d(\delta\rho)^4$ is needed to get a valid picture in the critical region. This was the suggestion of Landau and Ginzburg for the order parameter of the superconducting transition. Its adoption complicates the above treatment very considerably and has far-reaching consequences, as we shall see in Sec. 5.4.

2. ISING MODEL: EXACT SOLUTIONS

2.1 Ising Model

In 1925 W. Lenz, who had been looking for a simple model that might serve to explain ferromagnetism, put forward the following suggestion to his graduate student E. Ising. Assume that each atom possesses a spin and hence a magnetic moment μ_0, which can orient either parallel or antiparallel to an external magnetic field H. There is an interaction between nearest-neighbor spins in the lattice, parallel spins having an interaction $-J$ and antiparallel spins $+J$ ($J>0$). Lenz hoped that the model might manifest a nonzero spontaneous magnetization, i.e., that as $H\rightarrow 0$ the ratio

$$(N_2 - N_1)/(N_2 + N_1) \qquad (23)$$

might tend to a nonzero value. Here N_1 and N_2 are the number of spins antiparallel and parallel to the field H.

Ising was able to solve the problem in one dimension only, where the statistical problem is quite simple, and the solution did not possess a spontaneous magnetization at any temperature above zero. He concluded wrongly that the model would not give rise to a spontaneous magnetization in higher dimensions. In 1936 Peierls was the first to demonstrate convincingly that for a two-dimensional Ising model, at sufficiently low but nonzero temperatures, the ratio (23) does not tend to a nonzero value as $H \to 0$.

The Hamiltonian for the Ising model can be conveniently written in the form

$$\mathscr{H} = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j - \mu_0 H \sum_i \sigma_i, \tag{24}$$

where the sum i is taken over all N spins of the lattice and $\langle i, j \rangle$ over all of the nearest-neighbor pairs of the lattice, counting each pair once only. Each σ_i takes on the values ± 1 according as the spin i is pointing in the direction 1 or 2. Ising considered only a model of spin $\frac{1}{2}$; the natural generalization to spin s is

$$\mathscr{H} = -\frac{J}{s^2} \sum_{\langle i,j \rangle} s_{zi} s_{zj} - \frac{\mu_0 H}{s} \sum_i s_{zi}, \tag{25}$$

where s_{zi} are standard quantum spins taking on the values $s, (s-1), \ldots, -(s-1), -s$. The normalization is chosen to keep the maximum interaction between nearest-neighbor spins, and the maximum interaction with the external field, independent of s. The term Ising model without qualification usually refers to the model of spin $\frac{1}{2}$; it is sometimes called the simple Ising model for greater clarity.

Let us now consider the partition function (see STATISTICAL MECHANICS, QUANTUM) for the Ising model. Any configuration of spins characterized by the numbers N_1 and N_2 of spins pointing in directions 1 and 2, and N_{11}, N_{12}, and N_{22} different types of nearest-neighbor pairs, will have energy

$$\mu_0 H(N_1 - N_2) + J(N_{12} - N_{11} - N_{22}). \tag{26}$$

It is easy to establish the relations

$$2N_{11} + N_{22} = qN_1, \quad 4N_{12} + N_{22} = qN_2, \tag{27}$$

by enumerating all q bonds that originate in each atom of the lattice (q being the coordination number of the lattice).

For convenience, to avoid surface effects, we adopt the usual solid-state physics device of cyclic boundary conditions, so that all lattice sites are equivalent. Also we know that

$$N_1 + N_2 = N; \tag{28}$$

hence Eq. (26) can be written

$$-(\mu_0 H + \tfrac{1}{2} qJ)N + 2\mu_0 H N_1 + 2JN_{12}. \tag{29}$$

If $J > 0$ the lowest-energy state corresponds to $N_1 = 0$, $N_{12} = 0$, i.e., all the spins are aligned in direction 2. Of the five numbers introduced only two are free, and these are conveniently taken to be N_1 and N_{12}.

The major difficulty in the calculation of the partition function is the combinatorial problem of determining how many states there are on a lattice with a given (N_1, N_{12}). If we denote this by $g(N; N_1, N_{12})$, the partition function can be written as

$$\begin{aligned} Z_N^I &= \exp(-\beta \mathscr{H}) \\ &= y^{-N/2} z^{-qN/4} \sum_{N_1, N_{12}} g(N; N_1, N_{12}) y^{N_1} z^{N_{12}} \end{aligned}$$

$$[y = \exp(-2\beta\mu_0 H), \quad z = \exp(-2\beta J),$$

$$\beta = 1/kT]. \tag{30}$$

We now consider the problem of an alloy consisting of N_A atoms of type A and N_B atoms of type B on the same lattice ($N_A + N_B = N$). Since the alloy has a fixed composition, N_A and N_B remain fixed, in contrast to N_1 and N_2 above. But since N_A and N_B represent particular values of N_1 and N_2, relations (27) are valid for N_A, N_B, N_{AA}, N_{AB}, and N_{BB}. To calculate the energy of a given configuration it is reasonable to use a model taking account only of nearest-neighbor interactions, since interatomic forces decrease rapidly with distance. Let us denote the interactions of AA, AB, and BB pairs of neighbors by $-\epsilon_{AA}$, $-\epsilon_{AB}$, and $-\epsilon_{BB}$, respectively. Consider a configuration in which there are N_{AA}, N_{AB}, and N_{BB} such pairs. The energy of the configuration is

$$\begin{aligned} w &= -N_{AA}\epsilon_{AA} - N_{AB}\epsilon_{AB} - N_{BB}\epsilon_{BB} \\ &= -\tfrac{1}{2} q(N_A \epsilon_{AA} + N_B \epsilon_{BB}) - N_{AB}\epsilon \end{aligned}$$

$$[\epsilon = \epsilon_{AB} - \tfrac{1}{2}(\epsilon_{AA} + \epsilon_{BB})]. \tag{31}$$

If $\epsilon < 0$ the lowest-energy state corresponds to the minimum of N_{AB}, i.e., to two separated phases of A and B atoms. This gives rise to the analog of an Ising ferromagnet in the realm of alloys, a *solid solution*.

The partition function for the solid solution is, from Eq. (30),

$$Z_N^s = z_A^{(1/2)qN_A} z_B^{(1/2)qN_B} \sum_{N_{AB}} g(N;N_A,N_{AB}) z^{N_{AB}}$$

$$[z_A = \exp(\beta\epsilon_{AA}),\ z_B = \exp(\beta\epsilon_{BB}),\ z = \exp(\beta\epsilon)]. \quad (32)$$

On comparing Eqs. (32) and (30) we see that, except for the ground-state terms, the partition function for the alloy is obtained by picking out the coefficient of y^{N_A} in the partition function for the magnet. In statistical-mechanical terminology (see STATISTICAL MECHANICS, QUANTUM) the partition function for the magnet corresponds to the *grand partition function* for the alloy. Writing

$$\Lambda_N(y,z) = \sum_{N_1,N_{12}} g(N;N_1,N_{12}) y^{N_1} z^{N_{12}}, \quad (33)$$

the behavior of both the ferromagnet and the alloy can be calculated from a knowledge of $\Lambda_N(y,z)$.

In Sec. 1.4 we referred to order–disorder transitions in binary alloys that form superlattices at low temperatures. This corresponds to $\epsilon > 0$ in Eq. (31) which gives rise to a lowest-energy state with N_{AB} a maximum. The magnetic analog with $J < 0$ is an Ising model of an antiferromagnet in which the lowest-energy state corresponds to alternating up and down spins. Both of these problems are likewise determined by $\Lambda_N(y,z)$ defined in Eq. (33) but for a different range of values of the variable z.

The introduction of a specific model (25) enables us to mention an important alternative method of deriving the classical result. If we focus attention on the field acting on a given spin as a result of interactions with its neighbors, this field will fluctuate as the orientation of the neighbors changes. The exact solution must take these fluctuations into account exactly. But we can approximate by replacing the fluctuating field by its mean value, and from this a self-consistent equation can be derived for the magnetization. This is equivalent to the classical result, and is usually called the *mean-field approximation.* We would expect from the statistical considerations that it would be a reasonable approximation when the number of interacting neighbors is sufficiently large.

2.2 The Lattice-Gas Model

In 1939 Cernuschi and Eyring realized that the Ising model could readily be adapted to provide a simple model of liquid–gas equilibrium. If, in the solid-solution model of the previous section, type A atoms are replaced by the molecules of the fluid, and type B atoms by holes, i.e., unoccupied cells, the possibility of varying the number of holes allows for changes in free volume, one of the essential features of a liquid. Any two neighboring occupied cells have an attractive interaction; the fact that an occupied cell cannot accept another molecule provides a crude representation of a hard-core repulsion, and the model thus represents the two basic features of intermolecular forces. The concentration of holes is not fixed but allowed to vary, and this is taken into account by means of the grand partition function, as indicated in the previous section. The interactions ϵ_{AB} and ϵ_{BB} are zero, ϵ_{AA} represents the intermolecular attraction, and ϵ defined in Eq. (31) is equal to $-\frac{1}{2}\epsilon_{AA}$.

For this model the magnet–fluid analogy discussed in Sec. 1.2 is formulated precisely, and the thermodynamic quantities associated with fluids, such as pressure and density, can be given exact magnetic analogs. All of the classical results discussed in Sec. 1 can be related to a simple function of two variables $\Lambda_N(y,z)$ defined by Eq. (33), and we can understand the similarities in experimental behavior of the various systems. However, when we come to examine the critical behavior of the function $\Lambda_N(y,z)$ we shall find that it differs very significantly from the Landau pattern described in Sec. 1.5.

2.3 Transfer Matrix

For a one-dimensional assembly of systems each having t possible energy levels, a general treatment is possible that reduces the evaluation of the partition function to the determination of the largest eigenvalue of a matrix. Let $U(x_i,x_{i+1})$ denote the energy between any two systems x_i and x_{i+1}, each x_i taking t possible values. For simplicity we shall deal with a ring of systems in which the Nth is connected cyclically to the first. The partition function for N systems is then given by

$$Z_N = \sum_{x_i} \exp\{-\beta[U(x_1,x_2)+U(x_1,x_2)+\cdots$$
$$+U(x_N,x_1)]\}$$
$$= \sum_{i,j,\ldots,k=1}^{t} V_{ij}V_{jk}\cdots V_{kl}V_{li}, \qquad (34)$$

where $V_{ij}=\exp(-\beta U_{ij})$ corresponds to the interaction between neighboring systems in states i and j. The elements V_{ij} thus form a $t\times t$ matrix V and the continued product (34) will readily be identified as the trace of V^N. Thus if $\lambda_1,\lambda_2,\ldots,\lambda_t$ are the eigenvalues of V, then

$$Z_N=\lambda_1^N+\lambda_2^N+\cdots+\lambda_t^N. \qquad (35)$$

For sufficiently large N only the largest eigenvalue, λ_1, is significant, and we may identify Z_N with λ_1^N. It should be noted that if the interaction U_{ij} is symmetric in i and j, the matrix V is symmetric. Also the ith diagonal element of V^N is proportional to the probability that a particular system is in state i.

The above method of treating one-dimensional assemblies was suggested independently in 1941 by Kramers and Wannier, by Lassettre and Howe, and by Montroll. It constituted the first preliminary step towards modern theories of critical behavior, focusing attention on the exact calculation of the partition function. The matrix V is usually called the *transfer matrix.*

For the one-dimensional simple Ising model $t=2$, and the eigenvalues in Eq. (35) can readily be calculated:

$$\left.\begin{matrix}\lambda_1\\ \lambda_2\end{matrix}\right\}=\tfrac{1}{2}(1+y)\pm[(1-y)^2+4yz^2]^{1/2}. \qquad (36)$$

This is the solution found by Ising in 1925. For large N we can ignore λ_2, and since λ_1 is a smooth analytic function of (y,z) there are no singularities. There is no spontaneous magnetism for $T>0$; the only manifestation of singular behavior occurs at $T=0$ where the susceptibility takes on the value $\beta\mu_0^2\exp(\beta J)$, which becomes infinite.

The method outlined above can be applied to a finite strip of width l, and the number of different states t to be taken into account for the transfer matrix is then 2^l. Similarly, for a three-dimensional block of dimensions $l\times m$ the transfer matrix is of order 2^{lm}. As long as the transfer matrix is finite and the largest eigenvalue is nondegenerate, it is easy to show that it involves no singularities. To achieve singular behavior the matrices must be infinite; hence for the Ising model only systems infinite in at least two dimensions can manifest singular critical behavior. Such dimensional considerations were completely absent in the classical Landau theory (Sec. 1.5).

2.4 Onsager's Solution for the Rectangular Lattice

In 1936 Peierls had demonstrated rigorously that the Ising model on the quadratic lattice in two dimensions does indeed possess a spontaneous magnetization. In 1944 Onsager published an exact solution for this model in zero magnetic field, which exercised a revolutionary impact on the whole subject, since its properties differed essentially from those of the classical theories. This will be seen from the following brief summary:

1. *The specific heat* manifests a striking singularity at the Curie point. It is logarithmically infinite on both sides of the Curie temperature, and is shown in Fig. 5 for the symmetric net. For temperatures near to T_c the following approximation may be used:

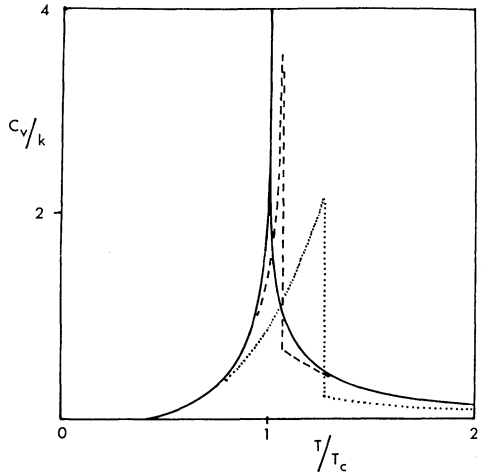

FIG. 5. Onsager's classic calculation (1944) of the specific heat of the spin-$\frac{1}{2}$ Ising model on a simple quadratic lattice. The solid curve represents the true specific heat, which is logarithmically infinite. The dotted curve represents a closed-form approximation due to Bethe, and the dashed curve an improved closed-form approximation due to Kramers and Wannier.

$$C_v/NK \cong (2/\pi)(\ln \cot \pi/8)^2 \times (K_1 - 1 - \pi/4)$$

$$\cong 0.4945 \ln |T - T_c|$$

$$[K_1 = \ln(2^{1/2}|K - K_c|),\ K = \beta J]. \quad (37)$$

It should be noted that the specific-heat curve differed substantially from any curves that had been observed experimentally, particularly in regard to the magnitude of the "tail" above the Curie temperature. This can be seen more clearly from the critical value of the entropy, which shows that the entropy change above T_c ($=0.387$) is greater than that below T_c ($=0.306$).

For an asymmetric net with interactions J, J' in the principal directions, the logarithmic nature of the specific-heat anomaly is unchanged. As one of the interactions J' becomes small the Curie temperature moves to zero, but this movement is very slow; if the interaction J' is J/v, then the reduction in T_c is of order $\ln v$ for large v. The specific-heat curve for an asymmetric net in which $J' = J/100$ was calculated by Onsager, and is shown in Fig. 6 for comparison with the symmetric net, and with the specific heat of a one-dimensional chain ($J' = 0$). The abscissa is $2kT/(J+J')$ so that the areas under all the curves are the same. It will be seen that the large asymmetry reduces the Curie temperature by a factor of 2 only. Also there is an essential difference below the Curie temperature between a two-dimensional net with one small interaction and a one-dimensional chain.

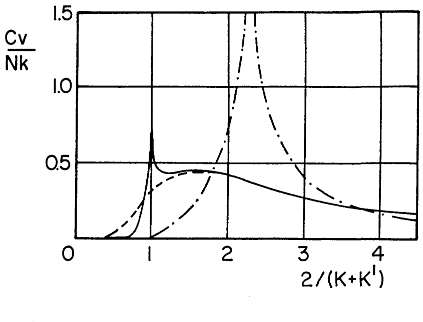

FIG. 6. Onsager's calculation of the specific heat of an asymmetric rectangular lattice with a ratio of interactions 100:1. The logarithmic infinity is retained in contrast to the specific heat of the linear chain (dashed line). The central curve (–·–) corresponds to the symmetric quadratic lattice.

2. *Other thermodynamic quantities.* The internal energy E has a vertical tangent on both sides of the Curie point. Near the Curie point it is of the form

$$A_0 + B_0(T - T_c)\ln|T - T_c|. \quad (38)$$

The free energy F involves an integral of the internal energy and therefore has infinite curvature at the Curie point. All thermodynamic quantities have a singularity at T_c, so that *no Taylor expansion of the type envisaged by Landau* (Sec. 1.5) *is possible.*

3. *The specific heat of a finite crystal.* The specific heat of a finite crystal of n rows is finite at all temperatures, since the partition function is an analytic function of the temperature, but a maximum occurs near the transition point of the infinite crystal. Onsager showed that for large n the maximum specific heat is given approximately by

$$C_v/Nk = 0.4945 \ln n + 0.1879. \quad (39)$$

4. *Correlations.* In Sec. 1.3 we described the classical theory of correlations between the molecules of a fluid as formulated by Ornstein and Zernike. By analogy, for a magnetic system we introduce the correlation between spins. For the simple Ising model we define

$$\langle \sigma_0, \sigma_r \rangle = (\Sigma \sigma_0 \sigma_r \exp(-\beta \mathscr{H}))/Z_N^I, \quad (40)$$

which plays a role analogous to $g(r)$ in Eq. (14). The calculation of correlations is more sophisticated than that of thermodynamic quantities, and a number of important exact results were derived by Kaufman and Onsager in 1949. Of particular significance was their conclusion that at T_c the correlations decay with distance as $r^{-1/4}$. In his original paper Onsager had calculated the mean range over which correlation persists for $T > T_c$ as $\sim (T - T_c)^{-1}$. For the classical theory [Eq. (15)] the corresponding behavior was r^{-1} and $(T - T_c)^{-1/2}$.

5. *Long-range order and spontaneous magnetization.* It can be readily shown that for the simple Ising model the long-range order and spontaneous magnetization are effectively identical. Onsager in 1947 and Yang in 1951 concluded independently that its critical behavior is $\sim (T_c - T)^{1/8}$, whereas classical theory gave $(T_c - T)^{1/2}$ [Eq. (20)].

2.5 Other Exact Results

For some time after 1944 hopes ran high that a solution of the three-dimensional Ising model might soon be forthcoming. But as a variety of different methods were advanced to rederive the Onsager solution, none of which could be extended to three dimensions, the realization spread that the three-dimensional model is a problem of exceptional difficulty.

However, solutions were derived for a variety of other two-dimensional lattices, e.g., the triangular and honeycomb. The analytic character of the solution was individual to each lattice. But the critical behavior followed the same pattern as Onsager's solution and indicated that properties like the coordination number of the lattice have little relevance to critical behavior.

In 1952 Yang and Lee proved that the function $\Lambda_N(y,z)$ in Eq. (33) is analytic for all values of y except for $y=1$ $(H=0)$ and $T<T_c$. This important result established the pattern of singularities and showed rigorously that the lattice-gas model is capable of accounting for the pattern of discontinuities in the isotherms representing liquid–gas equilibrium that were described in Sec. 1.

3. THREE-DIMENSIONAL MODELS

3.1 The Challenge of Onsager's Solution

Onsager's solution demonstrated clearly that the classical treatment was incorrect for the two-dimensional Ising model, and there was no valid reason for thinking that it would be correct for a three-dimensional model. But the thermodynamic functions arising from Onsager's solution differed radically from experimental results. The spontaneous magnetization, which comes to zero with an eighth power law, was much steeper than anything that had been observed in magnetic materials. The specific heat has a large tail that did not accord with measurements of order–disorder and λ-point transitions, and the logarithmic singularity symmetric on both sides of T_c seemed strangely unreal. Was the disagreement with experiment due to the dimensionality of the model, the neglect of forces other than nearest neighbor, or perhaps the inadequacy of the Ising interaction?

It was therefore of great importance to look for alternative methods that could provide reliable information on three-dimensional models and on alternative forms of interaction. The most useful source of such information in the initial period following Onsager was the development of series expansions at high and low temperatures. Further important information was extracted from Onsager's solution, and other exact solutions were discovered corresponding to unrealistic interactions but nevertheless providing insight into the relationship between critical behavior and intermolecular forces. The cluster integral theory of fluids, which had been one of the triumphant theoretical developments of the late 1930s, had seemingly come up against a barrier in its attempt to explain critical behavior. Onsager's work aroused interest in the far more primitive lattice-gas model, which is isomorphic with the Ising model. And finally the exact results for the two-dimensional Ising model and the new theoretical conjectures that began to emerge for other models provided a challenge to experimentalists to look for systems in nature that might be expected to conform to the theoretical models, and to provide accurate measurements of their critical behavior.

3.2 More General Models

In 1928 Heisenberg suggested that the origin of the large interactions giving rise to ferromagnetism lay in the quantum-mechanical exchange forces, which can be represented by a vector coupling of spins,

$$J\mathbf{s}_i\cdot\mathbf{s}_i=J(s_{xi}s_{xj}+s_{yi}s_{yj}+s_{zi}s_{zi}). \qquad (41)$$

Here s_x,s_y,s_z are quantum-mechanical noncommuting operators representing the components of spin. Because of the noncommutation of the spin operators the quantum-mechanical Heisenberg model presents special difficulties in calculation. If we allow s to become infinite, we derive the *classical Heisenberg model* with an interaction

$$J\sigma_i\cdot\sigma_j, \qquad (42)$$

where $\sigma_i\sigma_j$ are unit vectors that take on all possible orientations.

The Heisenberg model has spherical spin symmetry. A generalized Ising model of spins s has been defined in Eq. (25) by the interaction of

$$Js_{zi}s_{zj}, \tag{43}$$

where each s_z can take on $2s+1$ values. The Ising model has anisotropic spin interaction.

It is possible to combine the above interactions into a more general Heisenberg–Ising interaction

$$J_1s_{xi}s_{xj}+J_2s_{yi}s_{yj}+J_3s_{zi}s_{zj}. \tag{44}$$

The particular case $J_3=0$, $J_1=J_2$ is called the x-y model; it has two-dimensional circular spin symmetry. When $s\to\infty$ it leads to a classical Heisenberg model with two-dimensional spins.

We shall see in Sec. 6 that the spin dimension plays an important role in modern theory and is usually denoted by n. Confining attention to isotropic spin interactions, the Ising model corresponds to $n=1$, the x-y model to $n=2$, and the Heisenberg model to $n=3$.

3.3 Series Expansions

During the two decades that followed the publication of Onsager's solution, the critical properties of model systems had to be established on an individual basis — each model and each property entailed a separate calculation. The most useful tool turned out to be the generation of lengthy perturbation series expansions at high and low temperatures.

The idea of generating lengthy series expansions and correlating the coefficients with critical behavior was first advanced by Domb (1969). Onsager's solution applies only in zero field, and the first target was to explore the properties of the model in a nonzero field.

The method was developed with the help of Rushbrooke and Wakefield at Oxford and Sykes, Fisher, and other members of the Theoretical Physics Department at King's College, London, and applied to a number of problems. A few general remarks are necessary to clarify the basis of the method. In normal statistical work it is hazardous to attempt to extrapolate asymptotic behavior from a finite number of terms. However, in the case of a ferromagnet one could anticipate on physical grounds a pattern of expected critical behavior, and use the coefficients of the series to obtain the best fit to parameters such as critical temperature and critical exponents in the assumed asymptotic form for the coefficients. Moreover, the methods of analysis and fitting could be tested on Onsager's exact solution to see how well they worked and to get an idea of the error to be expected. The method proved capable of furnishing important reliable information, and as numerical techniques improved, its use became widespread.

The actual generation of terms required a close familiarity with graph theory, and the programming of sophisticated graphical enumeration problems on computers. There are basically three different types of series expansion.

1. *High-temperature expansions* are derived by expanding the partition function in powers of β:

$$\langle \exp(-\beta\mathscr{H})\rangle = 1-\beta\langle\mathscr{H}\rangle+\frac{\beta^2}{2!}\langle\mathscr{H}\rangle-\cdots$$
$$+\frac{(-\beta)^n}{n!}\langle\mathscr{H}^r\rangle-\ldots, \tag{45}$$

where $\langle\ \rangle$ denotes the classical average or quantum-mechanical trace.

Those familiar with statistics will recognize Eq. (45) as the analog of a *moment-generating function* and will doubtless remember that it is more convenient to take logarithms and evaluate the *cumulant generating function*

$$\ln\langle e^{-\beta\mathscr{H}}\rangle = -\beta\langle\mathscr{H}\rangle+\frac{\beta^2}{2!}(\langle\mathscr{H}^2\rangle-\langle\mathscr{H}\rangle^2)$$
$$-\frac{\beta^3}{3}(\langle\mathscr{H}^3\rangle-3\langle\mathscr{H}\rangle\langle\mathscr{H}^2\rangle$$
$$+2\langle\mathscr{H}\rangle^3)+\ldots\,. \tag{46}$$

Such expansions were developed for all of the models described in the previous section (Sec. 3.2). For the Ising model of spin $\frac{1}{2}$ simplifications arise, and the expansions could be taken to considerable length. The terms of the expansion for the free energy in zero field, and the magnetic susceptibility and its field derivatives, are usually positive in sign. These expansions were the most useful from which to obtain information about critical behavior.

2. *Density expansions* use the variable y [$=\exp(-2\beta\mu_0H)$, Eq. (30)] or its equivalent, and since one expects asymptotically that

$$\Lambda_N(y,z) \sim [\Lambda(y,z)]^N, \tag{47}$$

they give rise to an expansion of the form

$$\ln \Lambda(y,z) = yg_1(z) + y^2 g_2(z) + \cdots + y^r g_r(z) + \cdots . \tag{48}$$

For a lattice-gas model these expansions are equivalent to the well-known Mayer cluster integral expansions that were developed in the late 1930s.

In principle, such expansions could be determined for all the models of the previous section, but calculations were practicable only for the Ising model. The expansions start from a state of complete alignment of the spins in a large magnetic field, and terms of Eq. (48) correspond to groups of 1,2,3,... overturned spins.

3. *Low-temperature expansions* have proved practicable only for the Ising model. They use z [$=\exp(-\beta J)$, Eq. (30)] as expansion variable, the coefficients being functions of y. They are simply a rearrangement of the density expansions (48).

As a typical example of a high-temperature series we reproduce the expansion of the initial magnetic susceptibility χ_0 for the simple Ising model on the fcc lattice:

$$\begin{aligned}\chi_0/\beta\mu^2 &= 1 + 12w + 132w^2 + 1404w^3 \\ &\quad + 14\,652w^4 + 151\,116w^5 \\ &\quad + 1\,546\,322w^6 + 15\,734\,460w^7 \\ &\quad + 59\,425\,580w^8 + \ldots \\ &\quad (w = \tanh \beta J).\end{aligned} \tag{49}$$

Before discussing the analysis of the terms we make a few observations on power series. Consider a function $f(z)$ defined by

$$f(z) = \sum_{n=0}^{\infty} a_n z^n. \tag{50}$$

Then, if

$$\lim_{n\to\infty} |a_n|^{1/n} \tag{51}$$

exists and is equal to $1/z_c$, the series converges for $|z| < z_c$. We can then write

$$|a_n| \sim \phi(n)/z_c^n, \tag{52}$$

where

$$\lim [\phi(n)]^{1/n} = 1. \tag{53}$$

There is always a singularity on the circle $|z| = z_c$. If all the a_n are consistent in sign, then the dominant singularity lies on the positive real axis. Replacing z by $-z$, we see that if the a_n alternate regularly the dominant singularity lies on the negative real axis. More irregular alternations indicate dominant singularities in the complex plane. If a_n is real, these must occur in complex pairs $(1/z_c) \exp(\pm i\sigma)$. For a single pair we should expect

$$a_n \sim [\phi(n)/z_c^n] \cos n\sigma.$$

If σ is a simple fraction, this gives rise to cyclic behavior; otherwise it is more random.

If the a_n are all known *exactly*, we can (in principle) continue the function analytically across the whole plane. Asymptotic values of a_n determine the behavior near to the dominant singularity. A dominant singularity on the positive real axis corresponds to a positive temperature and can reasonably be identified with the Curie point. Hence we see that series of terms consistent in sign are particularly useful since a numerical analysis of the a_n provides direct information about critical behavior.

In view of Eq. (53) a simple assumption for the asymptotic behavior of $\phi(n)$ is

$$\phi(n) \sim An^g, \tag{54}$$

and to test whether this assumption is in accord with the numerical data, it is convenient to plot a_n/a_{n-1} for series (49) as a function of $1/n$. If assumption (54) is valid, then

$$a_n/a_{n-1} \sim (1/w_c)(1+g/n), \tag{55}$$

and hence a straight line should be obtained whose intersection with $1/n = 0$ determines w_c and whose slope determines g. Such *ratio plots* are reproduced in Fig. 7 for the susceptibility and specific heat, and it will be seen that they approach an asymptotic linear value quite rapidly. Both lead to the same value of $1/w_c$, as one might expect on physical grounds.

In contrast to high-temperature series, low-temperature series for the simple Ising model are inconsistent in sign and irregular in behavior for many common lattices. For example, the spontaneous magnetization m_0 for the fcc lattice can be put in the form

$$m_0/\mu_0 = 1 - 2u^6 - u^{11}(24 - 26u + 48u^4 + 252u^5$$

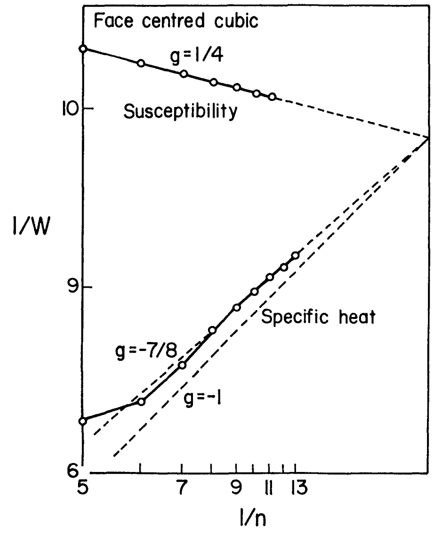

FIG. 7. An example of the use of series expansions to determine critical behavior. The susceptibility and specific heat are expressed as high-temperature expansions $\Sigma_n a_n w^n$ where $w = \tanh J/kT$. The ratios a_n/a_{n-1} of successive terms in the series expansion are plotted as a function of $1/n$, and from this an estimate can be obtained of their asymptotic behavior for large n. The ratios approach linearity, the extrapolated lines meeting the axis $1/n=0$ at $w=w_c$, the critical value, and the slopes give the critical exponents. If $a_n \sim n^{g-1}$, the corresponding thermodynamic quantity has a singularity of the form $(1-T_c/T)^{-g}$. *From this plot the estimates $\gamma \cong 5/4$, $\alpha \cong 1/8$ were derived (based on work by M. F. Sykes and his collaborators in* 1967 *at King's College London).*

$$-720u^6+438u^7+192u^8+984u^9 +1008u^{10}-12\,924u^{11} +19\,536u^{12}+\cdots)\quad (u=z^2), \qquad (56)$$

and it is difficult to detect any pattern in the coefficients.

A breakthrough in dealing with series of this kind occurred in 1961 when Baker introduced the Padé approximant into the field. This piece of mathematics, which had lain dormant since the end of the 19th century, enables *all* the singularities of a function defined by a power series to be studied; the unphysical singularities can be ignored, and attention can be focused on the singularity of physical interest.

Density series were of importance in attempting to find the m-H behavior at T_c. A power series in y is obtained from Eq. (48), the coefficients being evaluated at the critical value, z_c, of z estimated from high-temperature expansions.

3.4 Critical Exponents

Assumption (54) corresponds to a branch-point singularity at T_c. We therefore define a critical exponent (or index) θ of a thermodynamic quantity Ψ by the relation

$$\theta = \lim_{t\to 0} \frac{\ln \Psi}{\ln t}, \quad \Psi \sim A t^{\theta}. \qquad (57)$$

This definition ignores logarithmic factors. In a more refined analysis one might need to modify Eq. (57) to

$$\Psi \sim A t^{\theta} (\ln t)^{\phi}. \qquad (58)$$

Fortunately, in the practical application to critical phenomena in two and three dimensions, Eq. (57) is usually sufficient, although in four and more dimensions Eq. (58) is often needed. Also, when $\theta=0$ it is usual to specify whether the behavior is discontinuous (e.g., classical specific heat), or logarithmic (e.g., specific heat in the Onsager solution).

The notation for magnetic critical exponents developed empirically and the standardization provided by Fisher in 1965 at the Washington Conference on "Phenomena in the Neighbourhood of Critical Points" received wide acceptance. Table 1 defines the commonly used exponents and lists their classical values.

The Ising model served as a *trailblazer* in unravelling the general pattern of behavior of critical exponents. For the simple two-dimensional model α, α', β, ν, and η were known from the exact calculations quoted in Sec. 2. The remaining exponents were estimated from series expansions (e.g., $\gamma = 7/4$, $\delta = 15$) and were later shown to be exact.

In general, high-temperature series expansions relating to the initial susceptibility χ_0 were the most useful, and as the exponent γ was evaluated for different lattices and different models, a number of general conjectures were advanced as follows:

1. Critical exponents for a given model depend on dimension and not on lattice structure.
2. Critical exponents for the Ising model are independent of the spin s.
3. High-temperature critical exponents are

Table 1. Magnetic critical exponents ($t=T/T_c-1$, $t'=-t$).

Exponent	Quantity	Definition	Curve	Classical Value
$T>T_c$				
α	Specific heat	$C_m\sim t^{-\alpha}$	$m=0$	0 (discontinuous)
γ	Initial susceptibility	$\chi_0\sim t^{-\gamma}$	$m=0$	1
ν	Coherence length $\sim\delta$	$\kappa\sim t^{\nu}$	$m=0$	$\frac{1}{2}$
2Δ	Derivatives of susceptibility	$d^2\chi_0/dH^2\sim t^{-\gamma-2\Delta}$	$m=0$	3
$T<T_c$				
α'	Specific heat	$C\sim t'^{-\alpha'}$	$m=0$	0 (discontinuous)
β	Spontaneous magnetization	$m_0\sim t'^{\beta}$	$H\rightarrow 0+$	$\frac{1}{2}$
γ'	Initial susceptibility	$\chi_0\sim t'^{-\gamma'}$	$m=m_0$	1
ν'	Coherence length	$\kappa\sim t'^{\nu'}$	$m=m_0$	$\frac{1}{2}$
Δ'	Derivatives of susceptibility	$d\chi_0/dH\sim t'^{-\gamma'-\Delta'}$	$m=m_0$	$\frac{3}{2}$
$T=T_c$				
δ	Field vs magnetization	$H\sim m^{\delta}$		3
η	Pair correlation function	$\Gamma(\Gamma,0)\sim r^{d-2+\eta}$	$m=0$	0

ordered in a one-to-one manner with the symmetry of the magnetization in the ground state.

These conjectures represented important steps towards the hypothesis of universality (Sec. 4.3). The significance of the last conjecture can be illustrated by reference to the anisotropic Heisenberg model (44) with the following interaction:

$$J\sigma_{zi}\sigma_{zj}+J'(\sigma_{xi}\sigma_{xj}+\sigma_{yi}\sigma_{xy}). \qquad (59)$$

If $J'=0$, the model reduces to the simple Ising model in three dimensions for which γ had been estimated at 1.25, whereas if J' grows to J, the model becomes classical Heisenberg with corresponding estimate 1.38. What happens when $0<J'<J$? Conjecture 3 suggests that the change takes place discontinuously at $J'=J$, and that $\gamma\cong 1.25$ when $0<J'<J$. Likewise if $J=0$ the model is classical x-y with $\gamma\cong 1.32$; conjecture 3 suggests that if $J'>J$, γ retains the value 1.32.

Critical exponents for the three-dimensional Ising model lay between those for the two-dimensional model and those of classical theory. Thus, as the dimensionality increased, classical values were approached. However, increase in spin dimension generated a move away from the classical results. In keeping with this it was shown rigorously that the two-dimensional classical Heisenberg model has no ordered state except at $T=0$ and is in this respect similar to the *one-dimensional* Ising model.

3.5 Spherical and Gaussian Models

In 1952 Berlin and Kac discussed two new models for which they were able to calculate an exact solution in three dimensions in a nonzero field. The Gaussian model consists of independent spins with a nearest-neighbor Ising interaction, the magnitude of each spin being governed by a probability distribution

$$(2\pi)^{-1/2}\exp(-\sigma^2/2)d\sigma. \qquad (60)$$

The spherical model allows the spins to take any value subject to the overall restriction

$$\sigma_1^2+\sigma_1^2+\cdots+\sigma_N^2=N. \qquad (61)$$

Neither model is realized by any known physical system, and the solutions exhibit certain marked unphysical features.

For many years the models were regarded as of mathematical interest only. The discovery by Stanley in 1968 that the spherical model can be regarded as the limit of the classical vector spin model in which the *spin* dimension becomes infinite secured an important place for the model in the general pattern of critical behavior. The Gaussian model found its place as a first step in Wilson's application of the renormalization group (Sec. 5.4).

Like the Heisenberg model, the spherical model does not manifest critical behavior in two dimensions. In three dimensions Kac and

Berlin found that $\gamma=2$, and because of Stanley's result, the variation of γ as a function of spin dimension could be understood, starting with 1.25 for $n=1$, 1.32 for $n=2$, and 1.38 for $n=3$, and reaching 2 for $n=\infty$.

An important contribution of the spherical model to our understanding of critical behavior arose from the calculation by Joyce in 1966 of the properties of the model when the dimension $d \geqslant 4$. Joyce found that with the exception of logarithmic factors, all critical exponents attain their classical values when $d=4$ and retain them for higher dimensions. Joyce also calculated the properties of the model for long-range forces of the form $J(r) \sim 1/r^{d+\sigma}$ and found that for a sufficiently small σ, exponents attain classical values (even when $d=1$). In such a case a large number of neighbors effectively contribute to the interaction, so that the mean-field approximation discussed at the end of Sec. 2.1 gives the correct answer.

We shall see that these results are of great help in constructing an overall pattern of critical behavior.

3.6 Systems with Weak Long-Range Forces

In 1959 Kac introduced a one-dimensional continuum model with long-range forces that seemed to offer the possibility of exact treatment. The molecules interact with a hard-core repulsion of length b, and an exponential attraction $\lambda e^{-\lambda r}$. For this model he showed that the partition function can be derived as the maximum eigenvalue of a linear integral equation with a positive-definite Hilbert–Schmidt kernel.

The calculations were pursued by Kac, Uhlenbeck, and Hemmer in 1963. They first showed that for any finite value of λ the partition function is continuous, and there is no phase transition. However, in the limit of $\lambda \to 0$ there is a phase transition, and it can be evaluated in closed form. It gives rise, in fact, to the van der Waals equation [Eq. (4)] with the Maxwell construction automatically included. At the end of their paper, Kac, Uhlenbeck, and Hemmer discussed the three-dimensional model and advanced arguments to show that in this case also the attractive and repulsive parts of the potential make independent contributions to the free energy. In three dimensions the hard-sphere partition function is not known in closed form, but Monte Carlo estimates are available.

Calculations of a similar kind for the one-dimensional Ising model led to a Weiss type of solution [Eq. (7)]. In fact, it seemed that a mean-field solution will always be obtained for a model with a weak long-range attractive force of the form

$$J(r)=\lim_{\lambda\to 0} \lambda^{d} e^{-\lambda r}. \tag{62}$$

Classical theory was thus partially reinstated—it is a valid theory for long-range forces, but in practice intermolecular forces in nature are usually short ranged.

The work of Joyce on the spherical model indicated that a mean-field solution should result for models with forces of the form $J(r) \sim 1/r^{d+\sigma}$ for a sufficiently small σ.

3.7 Experimental Results

The experimental investigation of the behavior of a substance near a critical point is a matter of considerable difficulty. Very accurate temperature measurements are needed since small changes of temperature are accompanied by large differences in equilibrium values. Substances must be very pure; a small concentration of impurity can exercise a significant influence. As T_c is approached the time needed to reach equilibrium becomes large, so that a long time scale is needed for the performance of experiments. For fluids gravity must be taken into account. The theoretical values of several of the thermodynamic properties investigated become infinite at T_c. But such an infinity applies only to an unbounded sample; in practice the finite size of the sample becomes important since it provides a bound on the values that can be obtained. In solids this involves a consideration of grain boundaries.

The first suggestion that experimental results for real fluids deviated from the van der Waals picture came from Verschaeffelt at the end of the 19th century. Unfortunately his work was ignored by his contemporaries. In 1945 Guggenheim undertook a critical analysis of experimental data on the coexistence curves of a number of gases. According to van der Waals theory these gases should obey a law of corresponding states, i.e., if reduced units T/T_c and ρ/ρ_c are used (where ρ_c and T_c are the density and temperature, respectively, at the critical point), the coexistence curves for the different gases should fall on a single universal curve. Guggenheim found that the data provided good support for such a law of corresponding states (Fig. 8), but the form of the curve was cubic,

$$\Delta\rho \sim (-t)^{1/3},$$

rather than quadratic as required by van der Waals theory.

In 1954 Habgood and Schneider published precise measurements of the isotherms of xenon near the critical point. They found that the critical isotherm is much flatter than the cubic curve predicted by van der Waals (Fig. 9), and they suggested that the third and fourth derivatives of P with respect to ρ are zero at the critical point.

The theoretical developments described in Sec. 3.4 stimulated a new wave of precise experimental measurements, making use of new techniques and new magnetic materials to attain a much greater accuracy than had been available previously. We shall quote a few representative examples. In 1957 Fairbank, Buckingham, and Kellers published the result of a new investigation into the nature of the λ point of liquid ^{4}He. Improved low-temperature techniques enabled them to get very close to the λ point, and domains and

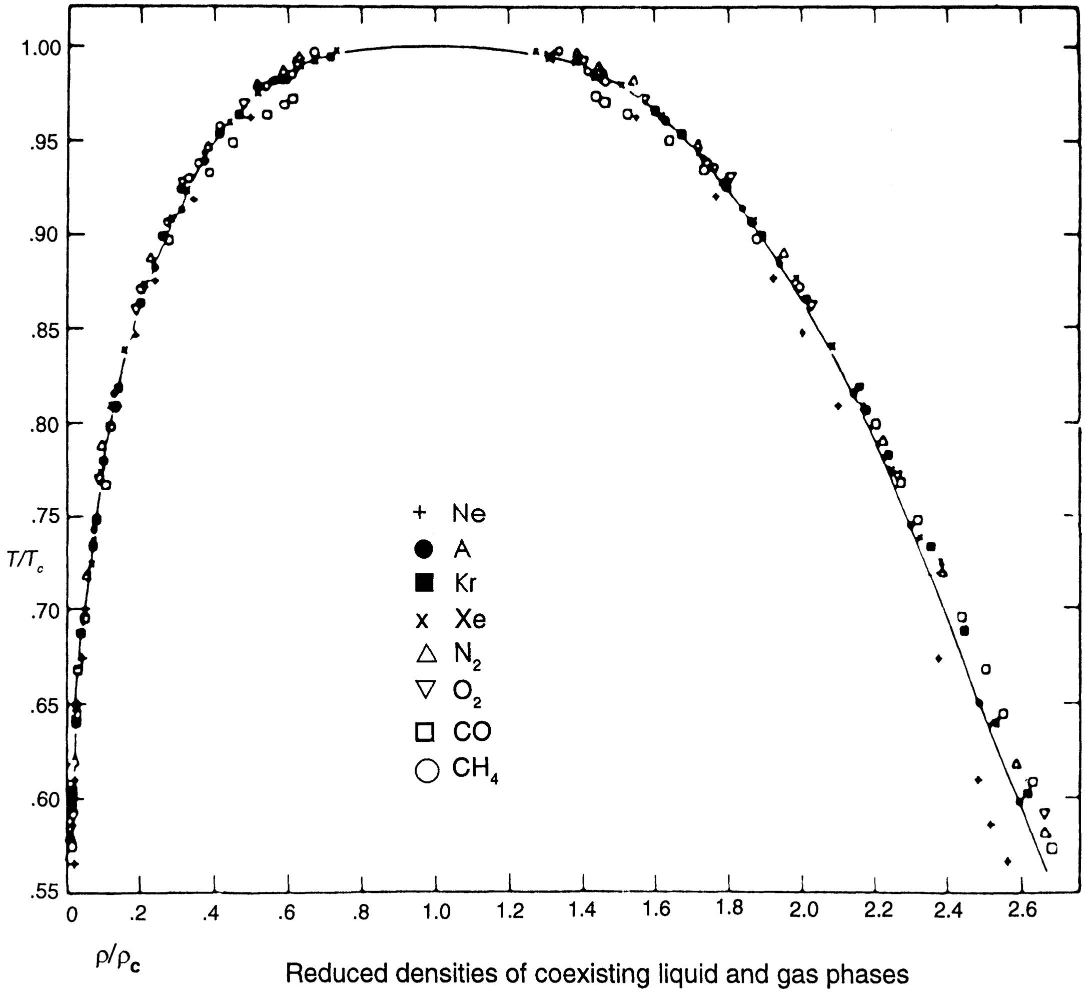

FIG. 8. Reduced densities of coexisting liquid and gas phases (T/T_c vs ρ/ρ_c). The data fit a law of corresponding states, but the resulting curve is cubic rather than quadratic as given by van der Waals's equation (after Guggenheim, 1945).

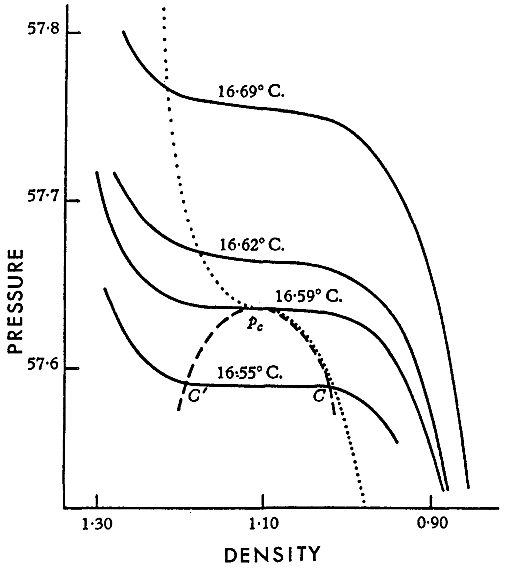

FIG. 9. Isotherms of xenon near the critical point. The dashed line marks the region of coexistent phases. The dotted line is the critical isotherm according to van der Waals's equation to be contrasted with the measured 16.59 °C isotherm (after Habgood and Schneider, 1954).

grain boundaries that limit the growth of specific heats of crystals are absent in liquid helium. They concluded that the specific heat is logarithmically infinite on both sides of the λ point (Fig. 10) and is in many respects similar to the Onsager specific heat depicted in Fig. 5. Of course they did not suggest that there was any basic reason for the similarity between a three-dimensional quantum fluid and a two-dimensional Ising ferromagnet, but at least they demonstrated that logarithmically infinite specific heats are a practical reality.

Conventional ferromagnets like iron, nickel, and cobalt have high Curie temperatures, and accurate measurements of critical behavior are difficult. However, new ferromagnets were discovered, such as europium sulfide ($T_c \sim 16.50K$), which are much more amenable to accurate measurements, and nuclear magnetic resonance can be used to measure the fraction of spins oriented in a particular direction at a given temperature and hence to deduce the spontaneous magnetization. Figure 11 is an example of such a measurement by Heller and Benedek in 1965 from which they deduced that the cube of the spontaneous magnetization is very nearly linear.

It is difficult to find a magnet that conforms satisfactorily to the Ising interaction. However, we have seen in Sec. 2.1 that an order–disorder transition in a binary alloy can also provide a realization of an Ising model. β-brass (CuZn) was chosen by Als-Nielsen and Dietrich for neutron-scattering experiments to test the theoretical predictions for the three-dimensional Ising model, and they estimated for the critical exponents $\beta = 0.30 \pm 0.005$ and $\gamma = 1.25 \pm 0.02$. These were in good agreement with the predictions of series expansions.

The new experimental results seemed to fit well to the theoretical calculations, even though the theoretical models were simple and crude,

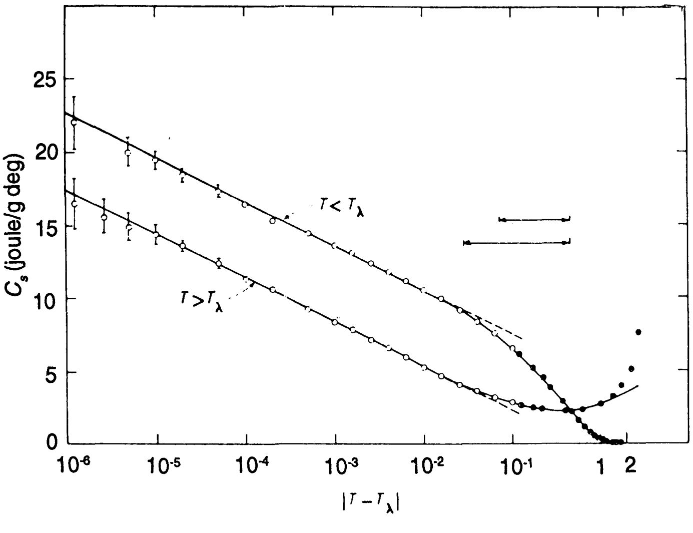

FIG. 10. Specific heat of liquid ^{4}He near the λ point. Accurate measurements by Fairbank, Buckingham, and Kellers in 1957 indicated a logarithmic curve of the type derived by Onsager for the Ising model.

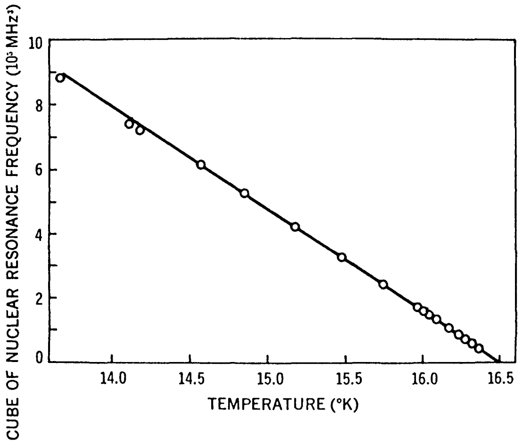

FIG. 11. Temperature dependence of the cube of the nuclear frequency of ^{153}Eu in EuS between 13.60 and 16.33 K. In this range the law $m_0 \sim Dt'^{1/3}$ is quite accurately obeyed (after Heller and Benedek, 1965).

and did not adequately map the true physical interactions. It seemed as if critical behavior was insensitive to the details of the interaction mechanism.

3.8 Conclusion

In this section we have described the chain of theoretical and experimental work set in motion by Onsager's solution. Agreement between theory and experiment was gratifying, but there was no coherent framework to replace the classical theory. In the next section we shall describe how such a framework was constructed.

4. SCALING AND UNIVERSALITY

4.1 Reconciliation

The first international conference devoted specifically to critical phenomenon took place in April 1965 in Washington at the National Bureau of Standards. It could reasonably be termed the founding conference of critical phenomena, since for the first time all the different strands of the subject were woven into a coherent fabric, and the major obstacles to progress were clearly delineated.

The keynote address was given by G. E. Uhlenbeck, and it is particularly fascinating now to quote the actual words of one of his concluding paragraphs:

> "If there is such an universal, but non-classical behaviour, then there must be a universal explanation which means that it should be largely independent of the nature of the forces. The only corner where this can come from is I think the fact that the forces are not long range. The Onsager solution gives I think a strong hint. It may well be so that away from the critical points the classical theories give a good enough description but that they fail close to the critical point where the substance remembers so to say Onsager. I think that to show something like this is the central theoretical problem. One can call it the reconciliation of Onsager with van der Waals."

Uhlenbeck was looking for a new type of universality for short-range forces that would replace the universality of the classical theories, which are valid only for very-long-range forces.

The first requirement in effecting the reconciliation was a coherent description of the "nonclassical universality" analogous to the van der Waals description of classical behavior. A preliminary attempt in this direction was the "droplet model" introduced by Essam and Fisher in 1963 and elaborated by Fisher a few years later. The model attempted to find an empirical approximation to $g_r(z)$ in Eq. (48) by analogy with droplet formation in a condensing gas. The resulting partition function depended on only two parameters and led to a relation between critical exponents,

$$\alpha'+2\beta+\gamma'=2, \qquad (63)$$

which was satisfied by the exact exponents for the two-dimensional Ising model. Rushbrooke noted that from thermodynamics alone one could derive an inequality

$$\alpha'+2\beta+\gamma'\leqslant 2. \qquad (64)$$

Rushbrooke's observation triggered off thermodynamic investigations that led to other inequalities of a similar kind. However, the heuristic droplet model had a number of unphysical features, and could not be regarded as providing a representation of the equation of state over the total range of the variables.

A satisfactory form of a nonclassical equation of state valid in the critical region emerged within a few months of the above conference.

It was suggested independently by three different groups: Widom (1965) in the U.S., who searched for a generalization of van der Waals's equation that could accommodate nonclassical exponents; Domb and Hunter (1965) in the U.K., as a result of an analysis of the behavior of series expansions of higher derivatives with respect to magnetic field at the critical point; and Patashinskii and Pokrovskii (1966) in the U.S.S.R., who considered the behavior of multiple correlations near the critical point.

Widom used fluid language in developing his ideas. Because of the exact symmetry $H \to -H$, $m \to -m$ in magnetic systems we shall find it convenient to reproduce the results in magnetic language. The transition to fluids is achieved by replacing H by $p-p_c$ and m by $\rho - \rho_c$.

4.2 Equation of State in the Critical Region: Scaling

The critical magnetic equation of state was derived from high-temperature series expansions of the form Eq. (45) and involved the plausible assumption (which was afterwards checked numerically) that the critical behavior of the rth field derivative of the magnetic susceptibility is of the form

$$\frac{d^r \chi_0}{dH^r} \sim t^{-\gamma - r\Delta}, \qquad (65)$$

where Δ was termed the gap exponent. This led to an equation of state of the form

$$m = t^{\beta} U(Ht^{-\Delta}). \qquad (66)$$

However, since one wishes to use the equation for positive and negative values of t, the form (66) is not very convenient, and Griffiths reformulated the equation so as to involve t only linearly,

$$H = m^{\delta} h(tm^{-1/\beta}). \qquad (67)$$

Equation (67) is correct only for positive H and m. By symmetry H and m can both be reversed without making any physical difference to the system. Therefore the general equation, valid for all m,H is

$$H = m|m|^{\delta-1} h(t|m|^{-1/\beta}). \qquad (68)$$

It is convenient to use Eq. (67) with the symmetry assumption latently understood.

There are two characteristic features of Eq. (67). First, all critical exponents depend on only two parameters, β and δ, and this is the basis for relations like Eq. (63) that are indeed satisfied exactly. The critical exponents of other thermodynamic quantities can be readily derived, and are reproduced in Table 2.

Second, critical data satisfy a *scaling relation*, i.e., if $Hm^{-\delta}$ is plotted against $tm^{-1/\beta}$, the two-dimensional data will fall on a single curve $h(x)$. As a result of the Yang–Lee theorem quoted in Sec. 2.5, $h(x)$ is analytic and the variable x ranges from $-x_0$, the zero of $h(x)$ (which corresponds to a spontaneous magnetization), to $x=\infty$. For a mean-field theory $h(x)$ is linear; as the dimension decreases and the spin dimension increases, $h(x)$ becomes more curved. $h(x)$ is termed the *scaling function.*

Patashinskii and Pokrovskii related susceptibility derivatives to correlation functions and were therefore able to relate the correlation exponent ν to Δ:

$$2\Delta = \gamma + d\nu. \qquad (69)$$

Such a relation was usually termed a *length scaling* relation as opposed to *thermodynamic scaling* relations of type (63).

These scaling relations were satisfied by the known exact exponents for the two-dimensional Ising model and for the spherical model (Table 2). For the three-dimensional Ising model, numerical estimates supported thermodynamic scaling, but there were small but persistent discrepancies in the length scaling relations, which will be discussed subsequently (Sec. 5.5). They were subject to experimental test for a large number of systems of fluids and magnets, and the results were very encouraging.

Similarly, the hypothesis of a scaling function $h(x)$ was tested for fluids and magnets, and a typical result using data for ^{4}He is reproduced in Fig. 12. It will be seen that the data fit well to a single curve.

4.3 Universality

Equation (67) says nothing about possible values of β, δ, and $h(x)$. We have already mentioned empirical information accumulating about regularities in the pattern of critical exponents. These were collected together in the *hypotheses of universality* put forward by

Table 2. Scaling relations between critical exponents.

Thermodynamic quantity	Exponent	Simple Ising model in two-dimensions	Mean field	Spherical model in three dimensions	Spherical model—long-range forces[a]
Spontaneous magnetization	β	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
Magnetization vs field at T_c	δ	15	3	5	$(d+\sigma)/(d-\sigma)$
High-temperature susceptibility	$\gamma=\beta(\delta-1)$	$\frac{7}{4}$	1	2	$\sigma/(d-\sigma)$
High-temperature gap	$\Delta=\beta\delta$	$\frac{15}{8}$	$\frac{3}{2}$	$\frac{5}{2}$	$(d+\sigma)/2(d-\sigma)$
High-temperature specific heat	$2-\beta(\delta+1)$	0	0	-1	$(d-2\sigma)/(d-\sigma)$
Low-temperature susceptibility	$\gamma'=\gamma$	$\frac{7}{4}$	1	…	…
Low-temperature gap	$\Delta'=\Delta$	$\frac{15}{8}$	$\frac{3}{2}$	…	…
Low-temperature specific heat	$\alpha'=\alpha$	0	0	…	…
		Length scaling			
High-temperature coherence length	$\nu=\beta(\delta+1)/d$	1	$\frac{1}{2}$	1	$1/(d-\sigma)$
Low-temperature coherence length	$\nu'=\nu$	1	$\frac{1}{2}$	…	…
Spin pair correlation at T_c	$\eta=d(\delta-1)/\delta+1$	$\frac{1}{4}$	0	0	$2-\sigma$

[a] Force $J(r)\sim r^{-(d+\sigma)}$ in d dimensions $(1<d/\sigma<2,\ 0<\sigma<\min\{2,d\})$.

Kadanoff in 1970 at a summer school on critical phenomena; an analogous *smoothness postulate* was advanced independently by Griffiths. This nonclassical universality is more sophisticated than had been envisaged by Uhlenbeck; different classes are defined by the space dimension d and the spin dimension n, and within a given class critical behavior is universal. If a third parameter σ is introduced to take account of the range of the intermolecular forces, both classical and nonclassical behavior can be taken into account in a wider pattern of universality. Once d, n, and σ have been specified, the exponents δ, β, and the function $h(x)$ are determined, and the behavior is smooth and universal within the class. Changes in critical exponents with accompanying discontinuities occur in the *crossover* between one universality class and another (e.g., in the transition from a two-dimensional to a three-dimensional system by the introduction of a new interaction), and we shall now discuss this crossover behavior in more detail.

4.4 Crossover Behavior

It was noted by Hankey and Stanley in 1972 that scaling laws are closely related to homogeneous functions. A homogeneous function of degree p in two variables u,v is defined by the relation

$$g(\lambda u,\lambda v)=\lambda^p g(u,v) \qquad (70)$$

for all λ. But if we use Eq. (70) with the value $\lambda=1/u$, we deduce that

$$g(u,v)=u^p g(1,v/u)=u^p G(v/u). \qquad (71)$$

Hence a homogeneous function of two variables scales to a function of a single variable.

To take account of the types of exponent arising in relations like (66) or (67), Hankey and Stanley introduced the concept of a *generalized homogeneous function* that satisfies the relation

$$g(\lambda^a u,\lambda^b v)=\lambda g(u,v). \qquad (72)$$

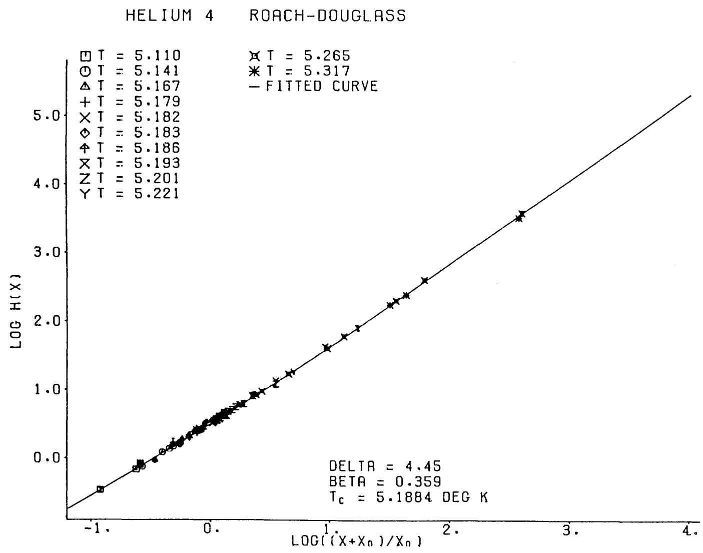

FIG. 12. Scaling relation for the fluid ^{4}He. Experimental data of Roach (1968) fitted by Vicentini-Missoni (1971).

A standard homogeneous function (70) corresponds to $a=b$. Taking λ to be $u^{-1/a}$ we find as a generalization of Eq. (71)

$$g(u,v)=u^{-1/a}g(1,v/u^{b/a})=u^{-1/a}G(v/u^{b/a}), \tag{73}$$

and this is of the same form as Eqs. (66) or (67). From (67), for example, we have

$$a=1/\delta, \quad b=-1/\beta\delta. \tag{74}$$

We have seen in Sec. 3.4 that critical exponents for a given model depend on dimension and not on lattice structure. If we consider a Hamiltonian of the form

$$\mathscr{H}=-J\sum_{\mathbf{i},\mathbf{j}}\sigma_{\mathbf{i}}\sigma_{\mathbf{j}}-gJ\sum_{\lambda,\mu}\sigma_{\lambda}\sigma_{\mu}=\mathscr{H}_0+g\mathscr{H}_1 \tag{75}$$

for a simple-cubic lattice, where $\mathbf{i},\mathbf{j}$ refer to nearest neighbors in horizontal planes, and λ,μ to nearest neighbors in the vertical direction, then for $g=0$ the behavior will be two-dimensional in character, but for $g\neq 0$ it will be three-dimensional. How does the transition or *crossover* depend on the parameter g when g is small? We shall investigate this example in detail as typical of the crossover from one universality class to another.

For the Hamiltonian (75), by analogy with Eq. (66), we postulate for the singular part of the free energy a generalized homogeneous function

$$f(g,t)=t^{2-\alpha_2}V(gt^{-\phi}), \tag{76}$$

where the subscript 2 is to remind us that we are dealing with the two-dimensional critical exponent, and ϕ is the new *crossover* exponent analogous to the gap exponent Δ in Eq. (66).

In presence of a magnetic field H, Eq. (76) would generalize to

$$f(g,H,t)=t^{2-\alpha_2}V(gt^{-\phi},Ht^{-\Delta_2}). \tag{77}$$

By differentiating Eq. (77) we conclude that thermodynamic quantities corresponding to the derivatives of the free energy with respect to H satisfy scaling relations similar to Eq. (76); for example, for the initial susceptibility we can write

$$\chi_0(g,t)\cong t^{-\gamma_2}Z(gt^{-\phi}). \tag{78}$$

However, there is an important difference between the scaling relations with respect to g and those with respect to H of Eqs. (66) and (67). For the latter the transition disappears when $H\neq 0$, whereas for the former a new transition point $T_c(g)$ appears with new critical exponents. For this to be possible the scaling function $Z(x)$ in Eq. (78) must have a singularity at some point x_c, and the singular behavior must be such as to enable the susceptibility when $g\neq 0$ to correspond to the three-dimensional Ising model:

$$\chi_0(g,t)\cong A(g)\dot{t}^{-\gamma_3} \quad [\dot{t}=T/T_c(g)-1]. \tag{79}$$

From this we first deduce that the new critical point is given by $gt^{-\phi}=x_c$, or

$$t_g = T_c(g)/T_c(0) - 1 = g/x_c^{1/\phi}. \quad (80)$$

To divide the crossover region more precisely into areas of temperature where the behavior is two-dimensional in character and those where it is three-dimensional, for a given g, we require knowledge of the scaling function $Z(x)$. But the general pattern of these areas can readily be outlined. Let us fix arbitrarily a percentage deviation p from two-dimensional behavior that we regard as significant, and let x_A be the value of x for which $Z(x_A)$ deviates from $Z(0)$ by p%. Then for $0 < x < x_A$ we can reasonably consider the behavior as two dimensional, i.e.,

$$t > (g/x_A)^{1/\phi}. \quad (81)$$

Let us now determine x_B at which the deviation of $Z(x)$ from its singular behavior is less than p%. For $x_B < x < x_c$ we can reasonably consider the behavior as three dimensional, i.e., for

$$(g/x_B)^{1/\phi} > t > (g/x_c)^{1/\phi}. \quad (82)$$

4.5 Kadanoff's Treatment

The progress described at the beginning of this section depended on an empirical and pragmatic approach. In a key paper published in 1966 Kadanoff tried to provide a theoretical basis for understanding the scaling properties that had been discovered in critical behavior.

Kadanoff noted that near T_c the coherence length ξ becomes very large, and hence it is possible to find a length L large compared with the lattice spacing but small compared with ξ. He then considered replacing the interaction of individual spins by the interaction of blocks of L^d spins. One might perhaps expect that in each block the spins would be nearly all up or nearly all down, and the original Ising model of spin σ and interaction J could then be replaced by a new model with block spin $\tilde{\sigma}$ and interaction $\tilde{J}$. If the new block spin model was effectively the same as the original model, the free energies would be related by

$$f(h,t) = L^{-d} f(\tilde{h},\tilde{t}) \quad (h = \beta H). \quad (83)$$

But how would $\tilde{h}$ and $\tilde{t}$ be related to h and t? Kadanoff suggested that

$$\tilde{h} = L^x h, \quad \tilde{t} = L^y t. \quad (84)$$

It is then easy to show that all the exponents (including ν) can be expressed in terms of x and y, that relations (63) and (69) are satisfied, and that a scaling-type equation of state (66) results.

Relation (69) is satisfied by the Ising model in two dimensions, but, as we have noted, numerical calculations for the three-dimensional model showed a small but persistent discrepancy. Also, there are several features of Kadanoff's argument that do not stand up to critical examination. However, his ideas stimulated the next major theoretical advance that completed the reconciliation sought by Uhlenbeck.

5. RENORMALIZATION GROUP

5.1 Background Ideas

The renormalization group (RG) was developed in the 1950s in connection with field theory (see PARTICLES, ELEMENTARY; QUANTUM ELECTRODYNAMICS), but it had no apparent practical consequences and had not been taken very seriously. Wilson saw that the theory was capable of providing an understanding of universality and a framework for the detailed calculation of critical behavior.

The RG does not produce exact solutions of the Onsager type, and its application involves quite drastic approximations, but it endeavors to retain all the essential physical characteristics of the problem near T_c, while rejecting those features that are not significant. No reliable method has been found of assessing directly the magnitude of the error at any stage of approximation of an RG treatment, and comparison with the results of exact solutions and series expansion estimates has played, and continues to play, a key role in the successful application of the RG.

Before discussing the RG approach we shall outline (following Wilson) the nature of the problem to be tackled in trying to account for critical behavior. We shall then give a qualitative account of how the RG deals with it.

If we consider the behavior of a fluid such as water, far from the critical point there are microscopic density fluctuations on an atomic scale (wavelengths ~1 Å). If the temperature and pressure are increased towards the critical point, fluctuations become important at longer wavelengths. Sufficiently close to T_c and P_c there are fluctuations on the scale of 1000–10 000 Å that scatter ordinary light and give rise to critical opalescence, making the water look milky. However, the microscopic fluctuations have not decreased in magnitude. Close to the critical point there are fluctuations at all wavelengths from 1 Å up to the correlation length, which tends to infinity at the critical point. The essence of the problem is how to deal with an assembly with so many scales of length.

The aim of the RG approach is to reduce in a systematic way the enormous number of degrees of freedom of the system, eliminating those that are irrelevant and do not contribute to critical behavior, and focusing attention on those which are relevant. A transformation is set up in the parameter space of the Hamiltonian,

$$\mathscr{H}' = \mathrm{R}\{\mathscr{H}\}, \tag{85}$$

which preserves the dimensionality and symmetries of the system and reduces the correlation length by a factor $b > 1$ and the number of degrees of freedom from N to $N' = N/b^d$. The transformation is chosen so that the partition function is preserved:

$$Z_{N'}(\mathscr{H}') = Z_N(\mathscr{H}). \tag{86}$$

A very wide choice of transformations is possible that can be characterized in *real space* or in *momentum space*. In real space the transformation usually follows the Kadanoff idea of introducing block-spin variables or eliminating some of the spins by partial summation (decimation transformation). However, the edge of the block is not long compared with the lattice spacing, but contains one or two lattice spacings; also the Kadanoff picture must be amplified, and the relation between the original spin and the block spin must be described precisely. In momentum space the aim is usually to eliminate high-momentum variables.

The transformation is iterated:

$$\mathscr{H}' = \mathrm{R}[\mathscr{H}],\ \mathscr{H}'' = \mathrm{R}[\mathscr{H}'],\dots . \tag{87}$$

A brief reference to the behavior of standard elementary iterative processes of the form

$$x_{n+1} = f(x_n) \tag{88}$$

will help to clarify the next step. It can easily be shown in Eq. (88) that the x_n tend to a limit x^* (fixed point), which is a solution of the equation

$$x = f(x). \tag{89}$$

If there are several solutions of this equation some may be *stable* (any deviation leads back to the fixed point) and others *unstable* with the opposite property. It can be shown easily that the fixed point is stable if $|f'(x^*)| < 1$ and unstable if $|f'(x^*)| > 1$. If $f'(x^*) = 1$ the fixed point is *marginal*, and we must proceed to higher derivation of $f(x)$ at x^*.

For iterative processes in s variables we can write Eq. (88) in the more general form

$$x_{(n+1)i} = f_i(x_{nj}) \quad (i,j = 1,2,\dots,s), \tag{90}$$

and fixed points x_i^* will be determined by solutions of the s equations

$$x_i = f_i(x_j). \tag{91}$$

The behavior at the fixed point is now governed by the eigenvalues of the matrix $|\partial f_i/\partial x_j|$ at x_i^*. If they are all less than 1, the fixed point is stable; if they are all greater than 1, it is unstable. But there are now other possibilities of partial stability; if t are > 1 and $s - t$ are < 1, we can describe the fixed point as t-unstable. In the theory of critical phenomena we are concerned with t-unstable fixed points with $t = 1$, 2, and perhaps 3.

The characteristic of a stable fixed point is that for a certain range of variables the solution is independent of the starting point. Each stable fixed point $x_i(r)^*$ has a *range of attraction*, and for any starting value within this range the limit will be $x_i(r)^*$. It is thus assumed in Eq. (87) that convergence takes place in the parameter space of the Hamiltonians to a fixed-point Hamiltonian, satisfying

$$\mathscr{H}^* = \mathrm{R}[\mathscr{H}^*], \tag{92}$$

and that critical properties are determined by behavior near the fixed point. Each fixed point has its own range of attraction, and all Hamiltonians within this range will have the same pattern of critical behavior (universality). All such Hamiltonians lie on the *surface of criticality* in the parameter space of Hamiltonians. But different fixed points in the parameter space give rise to different universality classes.

5.2 RG Derivation of Critical Behavior

Consider the partition function

$$Z_N = \langle \exp(-\beta\mathscr{H}) \rangle, \tag{93}$$

which is characterized by a number of coupling constants $K^{(1)}, K^{(2)}, K^{(3)}, \ldots$ that depend on $T, H, \ldots$ and can be conveniently written as a vector $\mathbf{K}$. Suppose that the correlation length associated with this Hamiltonian is $\xi(K)$. We now apply a block-spin or decimation transformation R_b to obtain a new set of parameters $\mathbf{K}'$ satisfying

$$\mathbf{K}' = \mathbf{f}(\mathbf{K}), \tag{94}$$

where the function will usually be nonlinear. Since we have reduced the unit of length by b we have the relation

$$\xi(\mathbf{K}') = b^{-1}\xi(\mathbf{K}), \tag{95}$$

and because of Eq. (86) we have for the partition function per particle $Z = (Z_N)^{1/N}$

$$[Z(\mathbf{K}')]^{N'} = [Z(\mathbf{K})]^{N} \tag{96}$$

and hence for the free energy per particle

$$\psi(K') = b^d\psi(\mathbf{K}). \tag{97}$$

Applying the transformation successively as in Eq. (87) in the form

$$\mathbf{K}_{n+1} = \mathbf{f}(\mathbf{K}_n), \tag{98}$$

we find that

$$\xi(\mathbf{K}_{n+1}) = b^{-1}\xi(\mathbf{K}_n) = \cdots b^{-n}\xi(\mathbf{K}_0). \tag{99}$$

Suppose now that $\mathbf{K}_n$ approaches a fixed point $\mathbf{K}^*$ satisfying

$$\mathbf{K}^* = f(\mathbf{K}^*). \tag{100}$$

It then follows from Eq. (99) that if we start at $\mathbf{K}_0$, then

$$\xi(\mathbf{K}_0) = b^n\xi(\mathbf{K}_{n+1}), \tag{101}$$

so that if $\xi(K^*)$ is not zero the correlation of the original system must be infinite, i.e., $\mathbf{K}_0$ must correspond to the critical point.

To obtain critical behavior near $\mathbf{K}^*$ we write

$$\mathbf{K}_n = \mathbf{K}^* + \mathbf{k}_n, \tag{102}$$

where $\mathbf{k}_n$ is small. Applying this to Eq. (98) we can develop an expansion and retain only linear terms, obtaining

$$\mathbf{K}_{n+1} = \mathsf{L}\mathbf{k}_n, \tag{103}$$

where relations (103) are linear so that L is a matrix. Let $\mathbf{l}_i$ be the eigenvectors and Λ_i the eigenvalues, satisfying

$$\mathsf{L}\mathbf{l}_i = \Lambda_i\mathbf{l}_i. \tag{104}$$

Expand $\mathbf{k}_{n_0}$ in the form

$$\mathbf{k}_{n_0} = \sum_i u_i^{(0)}\mathbf{l}_i \tag{105}$$

where n_0 is sufficiently large for Eq. (103) to be valid. We then have

$$\mathbf{k}_{n+n_0} = \sum_i \Lambda_i^n u_i^{(0)}\mathbf{l}_i. \tag{106}$$

Any eigenvalue Λ_i whose modulus exceeds 1 will grow in importance (*relevant*); one whose modulus is less than 1 will be *irrelevant*; an eigenvalue for which $|\Lambda_i| = 1$ is called *marginal*. The $u_i^{(0)}$ are called *scaling fields* (and we can drop the superscript); for example, u_1 is the temperature t, u_2 the magnetic field H, etc.

From Eq. (99) we have the relation for the correlation function

$$\xi(u_1, u_2, \ldots) = b^n\xi(\Lambda_1^n u_1, \Lambda_2^n u_2, \ldots) \tag{107}$$

while the corresponding relation for the free energy follows from Eq. (97),

$$\psi(u_1, u_2, \ldots) = b^{-n}\psi(\Lambda_1^n u_1, \Lambda_2^n u_2, \ldots). \tag{108}$$

When $H=0$ and the only relevant field is t we find from Eq. (107)

$$\xi(t, 0, \ldots) = b^n\xi(\Lambda_1^n t) \cong b^n(\Lambda_1^n t)^{-\nu}. \tag{109}$$

Hence we find that

$$b\Lambda_1^{-\nu} = 1 \quad \text{or} \quad \nu = \ln b / \ln \Lambda_1. \tag{110}$$

For the free energy we define scaling indices y_T and y_H by

$$\Lambda_1 = b^{dy_T}, \quad \Lambda_2 = b^{dy_H}, \tag{111}$$

and we then have, for all λ,

$$\psi(t, H) = b^{-dn}\psi(b^{ndy_T}t, b^{ndy_H}H), \tag{112}$$

i.e.,

$$\psi(\lambda^{y_T}t, \lambda^{y_H}H) = \lambda\psi(t, H) \tag{113}$$

where $\lambda = b^{dn}$.

The free energy is a *generalized homogeneous function* (Sec. 4.4). Take λ to satisfy

$$\lambda^{y_T}t = 1, \quad \lambda = t^{-1/y_T}, \tag{114}$$

and we have

$$\psi(t,H)=t^{2-\alpha}K(Ht^{-\Delta}), \tag{115}$$

where

$$2-\alpha=y_T^{-1}, \quad \Delta=y_H/y_T, \quad K(u)=\psi(1,u), \tag{116}$$

i.e., we have derived a relation analogous to Eq. (66). Finally we have the additional relation, from Eqs. (110) and (111),

$$\nu^{-1}=dy_T. \tag{117}$$

Hence *all* exponents are determined by y_T and y_H in agreement with Kadanoff, and in support of length scaling.

The extension to crossover behavior (Sec. 4.4) is immediate. One needs to use a scaling field g and to introduce a scaling index y_g satisfying a relation corresponding to Eq. (111), and this will lead to a free energy of the form (77).

5.3 Real-Space Renormalization

To give a specific illustration of the above ideas for real-space renormalization, we take an example used by Wilson in relation to the simple Ising model on the two-dimensional square lattice of lattice spacing a. The block spin relates to a 2×2 cell of the lattice that contains 9 spins. The value of the block spin is determined by a simple majority rule; if 5 or more of the 9 spins point upwards, the block spin has value $+1$, otherwise it has value -1. It is a straightforward matter with the help of a computer to transform any configuration of the original lattice to one of the transformed lattice. The 2^9 possible configurations of a cell in the original lattice are reduced to 2 in the transformed lattice, the value of b above is 3, and the new lattice has spacing $3a$. Speaking roughly, we can say that the transformation averages out fluctuations in spin direction whose scale is smaller than the block size $3a$. It is as if one looks at the lattice through an out-of-focus lens that blurs the smaller features but does not affect the larger ones.

When the transformation is applied for a second time, fluctuations in spin direction whose scale is smaller than 3^2a will have been averaged out. After n applications, fluctuations whose scale is smaller than 3na will have been averaged out. Eventually fluctuations at all scales up to the correlation length are averaged out.

Block spins do not have the same couplings as the spins in the original configuration lattice. In the original formulation, there is one nearest-neighbor interaction parameter. The first stage of the block spin transformation introduces a finite but substantial number of more distant neighbors and multispin couplings. If the second state is carried out exactly, a new round of coupling parameters will be introduced; if this is continued indefinitely an infinite number of coupling parameters will be required. An infinity of parameters is needed to furnish an exact solution. Wilson's crucial conjecture was that all but a finite and relatively small number of coupling parameters can be ignored without impairing the physical description of behavior near the critical point. He later tested this conjecture by performing a calculation for the simple Ising model on the simple quadratic lattice using a decimation transformation (developed by Kadanoff) and retaining 217 parameters! His conclusion is as follows:

"The dominant interaction at the fixed point is the nearest-neighbor coupling. For accurate calculations one must include many more; but for a qualitative picture the nearest-neighbor constant K plus perhaps the next-nearest-neighbor coupling L should be enough. The remaining couplings are at least a factor of 5 smaller than L.... Thus the old idea of Kadanoff that there would be effective nearest-neighbor Ising models for block spins is very close to the truth."

5.4 Momentum-Space Renormalization

For the use of the RG in momentum space it proved convenient to employ a continuous-spin adaptation of the Ising model. This enabled contact to be made with the Gaussian model (Sec. 3.5) which gives valid results for critical exponents when $d>4$; on replacing the lattice structure by a continuum (as for a fluid) the theory could be merged with the Landau–Ginzburg development (Sec. 1.5) and with the field theories with which the RG originated.

The partition function for the simple Ising model (24) can be written in the alternative form

$$Z^I=\int_{-\infty}^{\infty} d\sigma_1 2\delta(\sigma_1^2-1)$$

$$\times \exp\left(K \sum_{\mathbf{l},\mathbf{m}} \sigma_{\mathbf{l}}\sigma_{\mathbf{m}} + \beta\mu_0 H \sum_{\mathbf{l}} \sigma_{\mathbf{l}}\right) \quad (118)$$

with $\sigma_{\mathbf{l}}$ a continuous variable. Wilson replaced the function $2\delta(\sigma_{\mathbf{l}}^2-1)$ by the continuous function

$$w(\sigma_{\mathbf{l}}) = \exp(-\tfrac{1}{2}c\sigma_{\mathbf{l}}^2 - u\sigma_{\mathbf{l}}^4), \quad (119)$$

where u is a positive number. When $c=-4u$, $w(\sigma_{\mathbf{l}})$ has the form

$$w(\sigma_{\mathbf{l}}) = \exp[-u(\sigma_{\mathbf{l}}^2-1)^2+u], \quad (120)$$

which reproduces the simple Ising model when $u \to \infty$. For finite positive u, $w(\sigma_{\mathbf{l}})$ then has maxima at $\sigma_{\mathbf{l}} = \pm 1$, and if u is large, the probability function is sharply peaked at $\sigma_{\mathbf{l}} = \pm 1$. When $u=0$ Eq. (119) the Gaussian model is retrieved for positive c.

Following the idea of universality classes enunciated in Sec. 4.3 one might expect that any weighting function of the form (120) would belong to the Ising universality class. Wilson made the more daring suggestion that as long as $u>0$ *any* weighting function of the form (119) will correspond to the Ising universality class, i.e., that the only value of the parameters (c,u) giving rise to a change in behavior is $u=0$.

The transition to a continuum-space model can be achieved on replacing the $-\sigma_{\mathbf{l}}\sigma_{\mathbf{m}}$ in Eq. (118) by

$$\tfrac{1}{2}(\sigma_{\mathbf{l}}-\sigma_{\mathbf{m}})^2. \quad (121)$$

Since the lattice spacing is small and not expected to exercise a significant influence on critical behavior, the discrete σ is replaced by the continuous function $\sigma(x,y,z)$ and becomes

$$\tfrac{1}{2}(\nabla\sigma)^2. \quad (122)$$

To operate in momentum space it is convenient to introduce the Fourier-transformed spin variables

$$\hat{\sigma}_{\mathbf{q}} = \sum_{\mathbf{l}} \exp(i\mathbf{q}\cdot\mathbf{l})\sigma_{\mathbf{l}}. \quad (123)$$

For a lattice structure of spacing a, the momenta $\mathbf{q}$ are essentially restricted by $|q| \leqslant \pi/a$ (neglecting the shape of the Brillouin zone which, as one might expect, proves irrelevant). This cutoff is characteristic of the physics of condensed matter, and contrasts with quantum field theory where there is normally no natural cutoff and the momenta must be taken to infinity; the latter is therefore represented by zero lattice spacing. The RG transformation R of Eq. (85) is defined by averaging out all the $\hat{\sigma}_{\mathbf{q}}$ variables satisfying

$$b^{-1} < |q|a/\pi \leqslant 1. \quad (124)$$

This corresponds as before to a spatial rescaling factor b.

The first practical application by Wilson of the RG was in momentum space taking $b=2$. Severe approximations were required in evaluating the integrals to put the results into a form from which iterative recursion relations could be derived. They were solved numerically and gave fair values for the critical exponents.

In a subsequent paper by Wilson and Fisher the key suggestion was advanced that the appropriate expansion parameter for the critical exponent is $\epsilon = 4-d$ treated as a continuous parameter, and expansions to the first order in ϵ were derived for a number of models. Wilson then generalized the treatment to the n-vector classical Heisenberg model, using for the exponent in Eq. (119)

$$-\tfrac{1}{2} c_0 \sum_{i=1}^{n} \sigma_i^2 - u_0\left(\sum_{i=1}^{n} \sigma_i^2\right)^2. \quad (125)$$

For this model he used Feynman graph methods of field theory (see QUANTUM ELECTRODYNAMICS) to derive the following expansions:

$$\gamma = 1 + \frac{n+2}{2(n+8)}\,\epsilon + \frac{(n+2)(n^2+22n+52)}{4(n+8)^3}\,\epsilon^2 + O(\epsilon^3), \quad (126)$$

$$\eta = \frac{n+2}{2(n+8)^2}\,\epsilon^2 + \frac{n+2}{2(n+8)^2}\left(\frac{6(3n+14)}{(n+8)^2} - \frac{1}{4}\right)\epsilon^3 + O(\epsilon^4), \quad (127)$$

$$\phi = 1 + \frac{n\epsilon}{2(n+8)} + \frac{\epsilon^2(n^3+24n^2+68n)}{4(n+8)^3}\,\epsilon^2 + O(\epsilon^3). \quad (128)$$

Even though ϵ is assumed to be a small parameter, substituting $\epsilon=1$ gave results that compared surprisingly well with the series expansion estimates.

This seminal work initiated a new diagrammatic approach to the calculation of critical exponents, and powerful field-theoretical techniques were used to extend the ϵ expansion to higher orders, and to improve the accuracy of the critical exponent estimates.

5.5 Critical Exponents: Corrections to Scaling

Using the above techniques Nickel added several terms to the ϵ expansion, but the series was found to be divergent. However, Nickel and Lipatov showed how to calculate the asymptotic behavior of the general term in the expansion, and this enabled a Borel transformation to be used to convert the series into a convergent one. From this Le Guillou and Zinn-Justin in 1977 made the accurate estimates of the exponents of the simple Ising model with $\epsilon = 1$ as follows: $\gamma = 1.240$, $\beta = 0.325$, $\nu = 0.630$. Series expansions had favored $\gamma = 1.25$, $\beta = 0.3125$, $\nu = 0.64$. The differences, although relatively small, were beyond the expected error bounds. Of more fundamental importance was the failure of the second set of exponents to satisfy the length scaling (or hyperscaling) relation (69).

The key to the resolution of this discrepancy lay in the form of the correction terms to scaling. It is latently assumed in the ratio analysis, Eq. (55), that there is no *confluent singularity*, i.e., that the critical behavior of the thermodynamic function under consideration is of the form

$$t^{-h}G(t), \tag{129}$$

where $G(t)$ is analytic. However, Wegener showed in 1972 that the RG leads to the conclusion that there is a correction term

$$t^{-h-\Delta_1}G_1(t), \tag{130}$$

where Δ_1 need not be an integer. Moreover, it was possible to derive an ϵ expansion for Δ_1, and for $\epsilon = 1$ this gave $\Delta_1 \simeq \frac{1}{2}$.

In 1980 the discrepancy was resolved in a paper by Nickel presented at the International Conference on Statistical Mechanics in Edmonton, Alberta. Nickel added a significant number of new terms to the series expansion for the susceptibility of the bcc lattice and showed that with a correction term of the form (130) it was possible to achieve agreement between the series expansion and RG estimates with acceptable error bounds.

It is now generally assumed that the most accurate estimates of critical exponents in three dimensions are provided by the ϵ expansion. The results shown in Table 3 are taken from the final 1980 paper of Le Guillou and Zinn-Justin.

5.6 Conclusion

It is very gratifying that such good agreement should finally have been achieved between the two most accurate methods of estimating critical behavior. As regards the contributions to progress in the RG, real-space renormalization has proved useful for illustrating the general principles but has not proved very accurate in three dimensions. Momentum-space renormalization has surprisingly provided the most accurate estimates of three-dimensional critical behavior to date.

6. MISCELLANEOUS ADDITIONAL TOPICS

Although we have described in detail the mainstream developments in critical phenomena, a number of related topics have been passed over for lack of space. We summarize them briefly in this section.

Exact solutions have played a role of special importance in critical phenomena. We have seen that the modern era was initiated by Onsager's exact solution, which revealed the defects of the classical theories and served as a test for new techniques of calculation. For about 20 years after Onsager's work there were no new solutions for realistic interactions that differed in a substantial way from

Table 3. ϵ-expansion estimates of critical exponents, from Le Guillou and Zinn-Justin (1980).

n	γ	ν	η	β	α	Δ_1
0	1.1615 ± 0.0020	0.588 ± 0.0015	0.027 ± 0.004	0.302 ± 0.0015	0.236 ± 0.0045	0.470 ± 0.025
1	1.241 ± 0.0020	0.630 ± 0.0015	0.031 ± 0.004	0.325 ± 0.0015	0.110 ± 0.0045	0.498 ± 0.020
2	1.316 ± 0.0025	0.669 ± 0.0020	0.033 ± 0.004	0.3455 ± 0.0020	-0.007 ± 0.006	0.522 ± 0.018
3	1.386 ± 0.0040	0.705 ± 0.0030	0.033 ± 0.004	0.03645 ± 0.0025	-0.115 ± 0.009	0.550 ± 0.016

that of Onsager. Then in 1967 Lieb produced a variety of new solutions for two-dimensional ferroelectric models. The critical behavior followed a completely different pattern from the Ising model, and it seemed that new ground was being broken.

Lieb was followed by Baxter, whose solution of another ferroelectric model, the eight-vertex model, gave rise to exponents that vary continuously with the strength of the interaction. This was a major challenge to the idea of universality, since the model is equivalent to an Ising model with a two- and four-spin interaction. The issue was resolved by an analysis by Kadanoff and Wegener in 1971, who indicated that the model possesses a special symmetry that gives rise to continuously varying exponents; this particular feature would not be expected to arise in normal physical systems.

Baxter subsequently derived a number of other new solutions and as a by-product extended very significantly exact series expansions for two-dimensional models which he was unable to solve exactly. (For more details of this exact work, see Baxter, 1982).

An interesting point that led to fruitful ideas is the behavior of two-dimensional models when $n=2$. Peculiarities in this model were noted some years ago, and their investigation by Kosterlitz and Thouless in 1973 has suggested a new type of critical phenomenon related to *topological order*. Experimental work has already confirmed some of the theoretical predictions.

Other problems in condensed-matter physics that have been profoundly influenced by progress in critical phenomena are the shape and size of polymer chains in solution (see, e.g., de Gennes, 1979), and percolation processes (see, e.g., Stauffer 1985). The model of a self-avoiding walk on a lattice (see, e.g., Domb, 1969) had for many years been regarded as a useful representation of a polymer chain with a realistic hard-core "excluded volume." It had been noted empirically that the properties of such self-avoiding walks paralleled those of the Ising model. The precise connection was revealed by de Gennes in 1972, who noted that the self-avoiding-walk configurations correspond to a ferromagnetic model with spin dimension n when n is allowed to take the value zero (see Table 3).

A theoretical model that burst into prominence after being ignored for nearly two decades is the Potts model. This was first suggested in 1952 as a generalization of the Ising model having q orientations but only two different energies of interaction; the critical point could be located exactly. At the Kyoto Conference on Statistical Mechanics in 1968, Kasteleyn and Fortuin showed that percolation processes associated with random mixtures that have applications to dilute magnets, amorphous semiconductors, and a number of other problems in condensed-matter physics correspond to the Potts model with $q=1$! A number of physical systems have been identified for which the Potts model is a reasonable representation, and the difference in symmetry between the n-vector ferromagnetic model and q-component Potts model gives rise to important differences in critical behavior. The literature on percolation and the Potts model has escalated sharply in the past few years.

GLOSSARY

Coherence Length: The range over which the behavior of atoms is significantly correlated.

Critical Exponents: Parameters that characterize the singular behavior of thermodynamic functions near a critical point.

Crossover: A change from one specific pattern of critical behavior to another.

Curie Temperature (Point): The temperature at which a ferromagnet loses its intrinsic magnetization.

Gaussian Model: A model of ferromagnetism in which the values of the atomic spins follow a Gaussian probability distribution.

Heisenberg Model: A model of ferromagnetism introduced by Heisenberg in 1928 with a quantum interaction between atomic spins.

Ising Model: A model of ferromagnetism suggested by Lenz to his student Ising in 1925 in which each spin has two values, up or down.

Lambda-Point Transition: A phase transition with no *latent* heat but with a *specific heat anomaly* in the shape of the Greek letter λ.

Lattice-Gas Model: A lattice model of a fluid in which each lattice cell is either occupied or unoccupied by a molecule.

Long-Range Order: Correlation of molecules or spins that extends to infinite range.

Mean-Field Solution: An approximation in the theory of critical behavior in which a fluctuating quantity is replaced by its average value.

Opalescence: Changes in color and appearance arising from light scattering.

Order Parameter: A key thermodynamic quantity identified by Landau that characterizes the critical behavior of a system.

Partition Function: A function constructed from the *microscopic* properties of molecules of a system from which *macroscopic* behavior can be derived.

Relevant (or Irrelevant) Fields: Interactions that do (or do not) significantly affect critical behavior.

Scaling Relations: Relations arising from a simplified equation of state in which two-dimensional data scale onto a single curve.

Universality: Systems of widely different origins whose critical behavior follows the same standard pattern.

List of Works Cited

Andrews, T. (1869), *Philos. Trans. R. Soc. London* **159**, 575–591.

Bain, E. C. (1923), *Trans. AIME* **68**, 625–641.

Baxter, R. J. (1982), *Exactly Solved Models in Statistical Mechanics*, New York: Academic.

Bragg, W. L., Williams, E. J. (1934), *Proc. R. Soc. London Ser. A* **145**, 699–730.

Bragg, W. L., Williams, E. J. (1935), *Proc. R. Soc. London Ser. A* **151**, 540–566.

Curie, P. (1895), *Ann. Phys. Chim.* **5**, 289–405.

DeGennes, P. G. (1979), *Scaling Concepts in Polymer Physics*, Ithaca, NY: Cornell University Press.

Domb, C. (1969), *Adv. Chem. Phys.* **15**, 229–259.

Domb, C., Hunter, D. L. (1965), *Proc. Phys. Soc. London* **86**, 1147–1151.

Guggenheim, E. A. (1945), *J. Chem. Phys.* **13**, 253–261.

Habgood, H. W., Schneider, W. G. (1954), *Can. J. Chem.* **32**, 98–112.

Heller, P., Benedek, G. (1965), *Phys. Rev. Lett.* **14**, 71–74.

Johansson, C. H., Linde, J. O. (1925), *Ann. Phys. (Leipzig)* **78**, 439–460.

Landau, L. D. (1937), *Phys. Z. Soviet Union* **11**, 26, 545. [English Translation in: D. ter Haar (Ed.) (1965), *Collected Papers of L. D. Landau*, London: Pergamon, 193–216].

Onsager, L. (1944), *Phys. Rev.* **65**, 117–149.

Ornstein, L. S., Zernike, F. (1914), *Proc. Akad. Sci. Amsterdam* **17**, 792–806.

Ornstein, L. S., Zernike, F. (1916), *Proc. Akad. Sci. Amsterdam* **18**, 1520–1527.

Patashinskii, A. Z., Pokrovskii, V. L. (1966), *Sov. Phys. JETP* **23**, 292–297.

Roach, P. R. (1968), *Phys. Rev.* **170**, 213–223.

Tammann, G. (1919), *Z. Anorg. Chem.* **107**, 1–239.

Vicentini-Missoni, M. (1971), in C. Domb and M. S. Green (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 2, New York: Academic, Chap. 2.

Weiss, P. (1907), *J. Phys. (Paris)* **6**, 661–690.

Widom, B. (1965), *J. Chem. Phys.* **43**, 3898–3905.

For detailed references of other papers mentioned in the text see

Domb, C. (1985), *Contemp. Phys.* **26**, 49–72.

Further Reading

Review Articles

De Jongh, L. J., Miedema, A. R. (1974) *Adv. Phys.* **23**, 1–269.

Domb, C. (1960) *Adv. Phys.* **9**, 149–361.

Fisher, M. E. (1967), *Rep. Prog. Phys.* **30**, 615–730.

Fisher, M. E. (1974), *Rev. Mod. Phys.* **45**, 597–616.

Heller, P. (1967), *Rep. Prog. Phys.* **30**, 731–826.

Wilson, K. G., Kogut, J. (1974), *Phys. Rep.* **12**C, 75–200.

Books

Baxter, R. J. (1982), *Exactly Solved Models in Statistical Mechanics*, New York: Academic.

De Gennes, P. G. (1979), *Scaling Concepts in Polymer Physics*, Ithaca, NY: Cornell University Press.

Domb, C. (1992), *The Critical Point*, London: Taylor and Francis, in press.

Domb, C., Green, M. S. (Eds.) (1972–6), *Phase Transitions and Critical Phenomena*, Vols. 1–6, New York: Academic.

Domb, C., Lebowitz, J. L. (Eds.) (1983–9), *Phase Transitions and Critical Phenomena*, Vols. 7–13, New York: Academic.

Ma, S. K. (1976), *Modern Theory of Critical Phenomena*, Reading, MA: Benjamin.

Stanley, H. E. (1971), *Phase Transitions and Critical Phenomena*, London: Oxford University Press.

Stauffer, D. (1985), *Introduction to Percolation Theory*, London: Taylor and Francis.

CRYOGENICS

Hans M. Bozler and Christopher M. Gould, *Department of Physics, University of Southern California, Los Angeles, California, U.S.A.*

Introduction 311
1. **What Happens at Low Temperatures?** 312
2. **Production of Low Temperatures** 313
2.1 Thermal Insulation 314
2.2 Cryogenic Refrigeration Using Fluids 315
2.2.1 Stirling Cryocooler 316
2.2.2 Joule-Brayton Recuperative Refrigerator 316
2.2.3 Joule-Thomson Refrigerator 316
2.2.4 Evaporative Coolers—1 K, 0.3 K 319
2.2.5 Dilution Refrigeration 319
2.3 Magnetic Refrigerators 320
2.3.1 Introduction to Spin Entropy and Adiabatic Demagnetization 320
2.3.2 Demagnetization of Paramagnetic Materials 321
2.3.3 Nuclear Demagnetization 321
3. **Measurement of Low Temperatures** 323
3.1 Temperature Scales 323
3.1.1 The Absolute Temperature Scale 323
3.1.2 Other Temperature Scales 324
3.1.3 International Temperature Scale of 1990 (ITS-90) 324
3.2 Classes of Thermometers 325
3.3 Practical Thermometers 326
3.3.1 Constant-Volume Gas Thermometer 326
3.3.2 Thermocouples 326
3.3.3 Cryogenic-Liquid Vapor Pressure 326
3.3.4 Resistance Thermometers 326
3.3.5 Diode-Voltage Thermometers 327
3.3.6 Capacitance Thermometers 327
3.3.7 Ferromagnetic and Antiferromagnetic Frequency Shifts 327
3.3.8 Electron Paramagnetic Susceptibility 327
3.3.9 Nuclear Susceptibility 327
3.3.10 ^{3}He Melting Pressure 328
3.3.11 Noise Thermometry 328
3.3.12 Nuclear Orientation 328
3.3.13 Quasiparticle Damping Thermometry 328
3.3.14 Nuclear Quadrupole Thermometry (NQT) 328
3.3.15 Second-Law Thermometry 329
4. **How Low Can We Go?** 329
4.1 Negative Temperatures 329
4.2 Present Ultralow-Temperature Limits 330
Acknowledgments 332
Glossary 332
Works Cited 333
Further Reading 333

INTRODUCTION

Cryogenics is the field of science and engineering that deals with the production of low temperatures. This subject includes the techniques needed to obtain low temperatures, the methods to measure low temperatures, and the study of the physics of matter at these temperatures. A quest for obtaining low temperatures followed the proposal of an absolute zero temperature by Guillaume Amontons in 1702. Progress accelerated after the middle 1800s when a series of experiments succeeded in liquefying the materials that were gases at room temperature. These were liquefied in the descending order of their boiling points—oxygen, nitrogen, hydrogen, and finally helium. In this article, we will briefly review the sorts of physical changes that take place in matter at cryogenic temperatures. Then we will discuss the techniques used to obtain these temperatures. After this we will introduce the very important subject of measuring temperature. Since the range of

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

temperature varies over a number of orders of magnitude, we will frequently refer to logarithmic graphs as shown in Fig. 1. Here, each factor of 10 in absolute temperature is one division on the scale.

1. WHAT HAPPENS AT LOW TEMPERATURES?

The understanding of absolute temperature, first postulated by Guillaume Amontons, was incorporated by William Thomson (Lord Kelvin), James P. Joule, Rudolf Clausius, and several other thermodynamicists into the laws of thermodynamics during the mid-1800s (Mendelssohn, 1977). The absolute temperature is zero when all thermal energy has been extracted from matter. Thus when a material is at zero temperature, only the quantum-mechanical zero-point motions remain.

As a material is cooled from room temperature many changes take place. Liquids freeze, and solids become stiffer and more brittle. Pure metals become less resistive, and many metals become superconductors. All of the gases that exist at room temperature eventually liquefy, and most of these also freeze into solids. At ordinary pressures, helium (the common isotope ^{4}He and the rare isotope ^{3}He) remains a liquid at absolute zero in temperature.

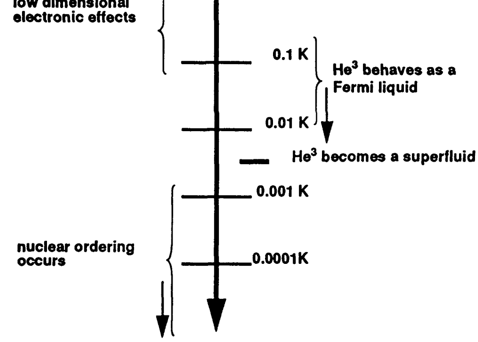
1000 K
room temperature
boiling point of water
freezing point of water
highest T_c superconductor
100 K
O_2,N_2 liquefy
H_2 liquefies
10 K
many metals become superconducting
He liquefies
He^4 becomes a superfluid
1 K
low dimensional electronic effects
0.1 K
He^3 behaves as a Fermi liquid
0.01 K
He^3 becomes a superfluid
0.001 K
nuclear ordering occurs
0.0001K

FIG. 1. Logarithmic temperature scale indicating significant events as the temperature is lowered from room temperature. The temperatures are on the kelvin scale.

The process of reducing temperature lowers the energy in materials. This is the energy that causes the random motions. For this reason, as the temperature is reduced, interactions between molecules, atoms, and even nuclei of neighboring atoms progressively become more important. We can substitute an energy scale, with ϵ being a typical single-atom or -molecule energy, for the temperature scale in Fig. 1 by using

$$\begin{aligned}\epsilon &= k_B T \\ &= (1.38\times 10^{-23}\ \mathrm{J/K})T \\ &= (8.63\times 10^{-5}\ \mathrm{eV/K})T. \qquad (1)\end{aligned}$$

As the temperature (or energy) is reduced below the interaction energy, important physical changes are likely to occur. These energies are small and become tiny when the temperatures are reduced. For this reason, cryogenics can be thought of as the physics of small energy.

A second way of thinking about low temperatures is to consider the entropy that is contained within materials (Kittel and Kroemer, 1980; Zemansky and Dittman, 1981). Entropy is a measure of disorder or randomness. The classical definition of entropy is written in terms of the number of ways that one can microscopically rearrange a system consistent with whatever physical constraints exist (e.g., temperature, energy, pressure or volume, magnetic field or magnetization). In usual systems, this number of rearrangements is a very large number, but the entropy is the logarithm of this number. If we use S for the entropy and Γ for the number of combinations we have

$$S = k_B \log \Gamma. \qquad (2)$$

When the thermal energy, as shown in Eq. (1), is smaller than interactions, the system goes into its ground (lowest-energy) state. The system becomes highly organized. Thus when a system is highly organized, the number of possible arrangements of particles Γ is small and S approaches zero. This corresponds to the low-temperature state of a system.

The universe has a background temperature of about 3 K due to the residue of radiation from the "big bang." Current measurements show this temperature to be 2.735 K (Mather, 1990). For this reason, there are no naturally occurring regions anywhere in the universe that have a temperature below this 2.7 K. Thus, with cryogenics, we are capable of producing unique states of matter (below 2.7 K) that are not duplicated in the natural world. By contrast, in other frontiers of technology (high energy, high pressure, high temperature, high magnetic fields, low pressure or high vacuum, etc.) manmade efforts are inevitably exceeded by nature.

At very low temperatures (below 1 K) the usual vibrations (phonons) in most materials become insignificant. Properties of materials are dominated by the interactions of electrons and magnetic spins, and eventually as the temperature is reduced, the magnetism of the nuclei becomes important. (Zemansky and Dittman, 1981; White, 1979; Lounasmaa, 1974).

2. PRODUCTION OF LOW TEMPERATURES

In 1877 L. Cailletet expanded high-pressure oxygen, which produced a fine mist of liquid (Mendelssohn, 1977). Simultaneously, R. Pictet had produced a similar result in another laboratory. After this, the race was on to liquefy all of the remaining gases. Sir James Dewar was the first to liquefy hydrogen in 1898; but, in the race to liquefy helium (the last natural gas), he was beaten out by Kamerlingh Onnes, who succeeded in 1908. Soon after this time, temperatures near 1 K were achieved.

As a by-product of the highly competitive effort to be the first to liquefy gases, several inventors worked to improve cryogenic refrigerators. The first patent for an expansion engine was given to Solvay in 1887. Hampson in England and Linde in Germany each filed a patent in 1895 for a process using a compressor, a counterflow heat exchanger, and a Joule–Thomson expansion valve to produce large quantities of liquid oxygen and nitrogen. Shortly after this, Claude filed a patent in France for a refrigerator that placed an expansion engine before the Joule–Thomson expansion valve in the Hampson–Linde refrigerator. An expansion engine greatly improves the refrigerator by letting the gas do work on a moving piston and thus lose heat energy. This staged approach to refrigeration is characteristic of low-temperature refrigerators.

Magnetic cooling was first proposed independently by William Francis Giauque and Peter Debye in 1926 (Mendelssohn, 1977). Giauque succeeded in reaching 0.25 K by 1933.

Currently, the production of low temperatures follows the techniques developed by the pioneers that we have mentioned above. Heat energy is extracted from materials with various methods of refrigeration depending on which methods efficiently remove entropy in the temperature range where the refrigerator is operating. Figure 2 shows, on a logarithmic temperature scale, the approximate temperature ranges in which these various techniques function best. As the temperature is reduced, the preferred method of cooling changes. For this reason, devices that produce or operate at low temperatures (cryostats) frequently employ a series of cooling techniques. This cascade approach is used in state-of-the-art cry-

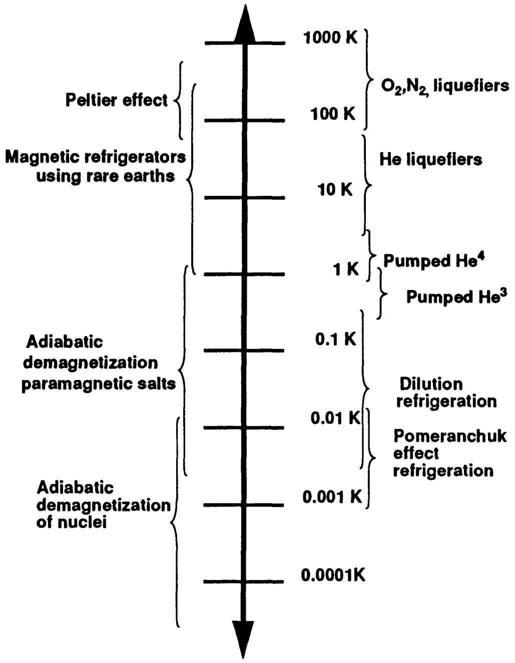
1000 K
100 K
10 K
1 K
0.1 K
0.01 K
0.001 K
0.0001K
O_2,N_2 liquefiers
He liquefiers
Pumped He^4
Pumped He^3
Dilution refrigeration
Pomeranchuk effect refrigeration
Peltier effect
Magnetic refrigerators using rare earths
Adiabatic demagnetization paramagnetic salts
Adiabatic demagnetization of nuclei

FIG. 2. Logarithmic temperature scale indicating the ranges of temperature in which various refrigeration techniques work most efficiently. Many of these techniques are discussed in the text, with the exception of Peltier-effect refrigerators and the Pomeranchuk effect.

ostats for obtaining microkelvin temperatures. These cryostats start with liquid helium at 4.2 K (a result of a cascade of mechanical coolers), then use an evaporation refrigerator to obtain 1 K; this is followed by a dilution refrigerator that can get to less than 10 mK. Next a nuclear cooling stage gets the temperature to well below 1 mK. Sometimes, a second nuclear cooling stage is added that can obtain a final (nuclear) temperature in the nanokelvin range.

2.1 Thermal Insulation

The production of low temperatures requires that we remove heat energy. Our ability to lower the temperature is ultimately limited by the amount of heat that leaks back in. For this reason, any low-temperature system needs to be surrounded by thermal insulation. This insulation can be in the form of materials that have low thermal conductivity, or a vacuum. Heat flows in three primary ways. First, there is diffusive conduction, which is usually high in pure metals and crystals and low in glasses and composite materials such as ceramics, plastics, and fibers (White, 1979; Powell and Childs, 1972). A second mode of heat flow is convection, in which heat is carried by a gas or liquid. A temperature inversion (hot region below the cold region) generally induces convection so that a fluid moves from a low-density hot region to a high-density cold region. Most gases are actually poor thermal conductors. However, in open regions, they can start a convective flow that transports much more heat than conduction. For this reason, many forms of insulation such as Styrofoam rely on trapping air in small cavities. The third common mode of heat flow is blackbody radiation. The heat flow due to radiation depends on the emissivity of the material and is proportional to T^4 (Stefan's law) (Kittel and Kroemer, 1980).

Containers for cryogenic fluids and equipment are referred to as Dewar flasks—named for James Dewar. Dewars rely on vacuum and reflective surfaces to minimize heat flow. In Fig. 3, we illustrate two types of Dewars. The first is the original vacuum flask. Here the walls are made from glass or stainless steel. Glass Dewars have a shiny metal film added to the walls to reduce radiation. (These Dewars are the same as the household Thermos™ bottles for holding coffee or cold drinks.) The second Dewar is an improved version that uses superinsulation in addition to the vacuum. Superinsulation is made of many layers of Mylar (polyester) film coated with a thin film of aluminum. The superinsulation then acts as a very efficient radiation shield since, going from higher temperature, each layer radiates less then the one outside. Also, residual gas in the vacuum space is trapped between the layers of the superinsulation. In modern Dewars, the neck, in which there is a rapid change in the temperature, is frequently made from an epoxy and fiberglass composite. For efficient storage of cryogenic liquids, the neck is usually kept small since much of the heat leaking into the Dewar is conducted down the neck. Dewars can be made in many shapes ranging from the simple cylindrical shapes shown here to large tank cars.

Many low-temperature devices use "heat switches," which are capable of turning on and off heat flow (White, 1978; Lounasmaa, 1974). Heat switches are used when there is a need to have parts of the cryogenic apparatus in close thermal contact at some times, and well isolated at other times. For example, a

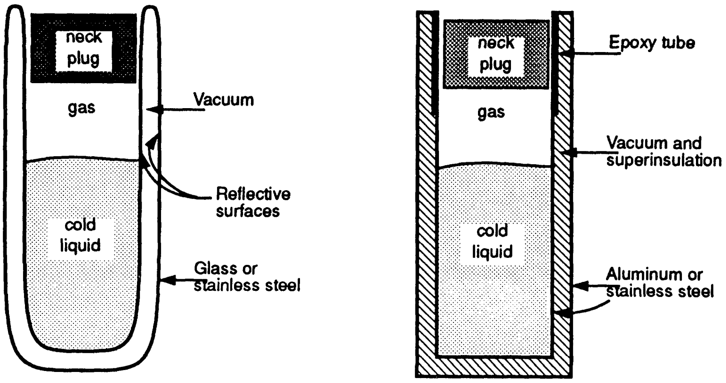

FIG. 3. Typical arrangements of Dewar flasks. The one on the left is close to the original design and uses a vacuum and highly reflective surfaces to insulate the liquid from heat coming through the walls. The walls are made from a poor heat conductor so that the top region, or neck, conducts very little heat. The Dewar on the right uses a more modern construction with the neck constructed of an epoxy tube (usually with fiberglass as a filler) and the walls filled with superinsulation.

heat switch can be used to allow cooling and later isolation of a material, the heat capacity of which is being measured. Such a switch can be a mechanical switch, or a system using a gas that can be later evacuated to exchange heat. At temperatures below 1 K, heat switches are most frequently made from pure superconducting metals such as Pb, Sn, In, and Al. (See the inset in Fig. 14.) When these metals are pure, their heat conductivity is very high in the "normal" state but very low in the superconducting state. A superconducting heat switch is made by surrounding the superconductor with a variable-field magnet. When the magnetic field is raised above the superconductor's critical field, the conductor becomes normal and switches from being a poor conductor of heat to a very good conductor of heat.

2.2 Cryogenic Refrigeration Using Fluids

All large-scale cryogenic refrigerators, and most small-scale refrigerators, use heat engines that have either a liquid, a gas, or a mixture of a liquid and gas as the refrigerant. These engines use the expansion of a compressed fluid as the cooling process. A few heat engines use magnetic or thermoelectric cooling. Graham Walker in his book on cryogenic heat engines (Walker, 1983) has coined the term "cryocoolers" to describe heat engines intended to produce cryogenic temperatures. In this section, we will briefly discuss three processes that can be used singly or in combinations to produce low temperatures. In practice, there are many variations of these processes named after different inventors. The three processes we discuss here illustrate the essential features that exist in all of these refrigerators.

In general, heat engines are devices that use heat to produce mechanical work. Examples of these would be steam turbines or gasoline and diesel engines. In Fig. 4 we have a schematic drawing of two heat engines, one running forward using high-temperature heat to produce work, and one running in reverse using work to remove heat. A forward-running heat engine is characterized by its efficiency ϵ in converting the high-temperature heat Q_h (that's what you pay for) into work W_{out}:

$$\epsilon = W_{\mathrm{out}}/Q_h. \tag{3}$$

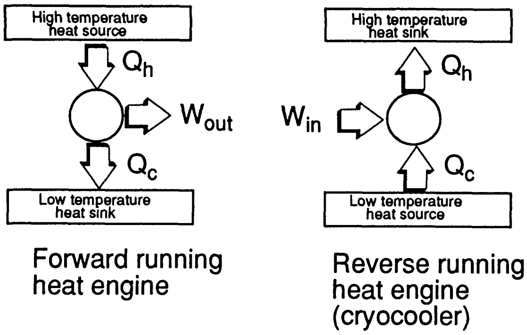
High temperature heat source
Q_h
W_{out}
Q_c
Low temperature heat sink
High temperature heat sink
Q_h
W_{in}
Q_c
Low temperature heat source

Forward running heat engine

Reverse running heat engine (cryocooler)

FIG. 4. Schematic diagrams for continuously operating heat engines. The forward-running heat engine uses heat Q_h to produce work W_{out} and rejects heat Q_c. The reverse-running heat engine uses work W_{in} to remove heat Q_c from a low temperature and adds heat Q_h to the high-temperature heat sink.

The first law of thermodynamics (energy conservation) requires that $W_{\mathrm{out}} = Q_h - Q_c$. The change in entropy due to heat flow is given by $\Delta S = Q/T$, which means that proportionately more entropy is carried away by the heat carried to the low-temperature reservoir. The second law of thermodynamics requires that the net entropy of the heat engine cannot decrease. Thus there is a maximum efficiency which is equal to $1 - T_c/T_h$ and can never exceed 1.

A cryocooler is a reverse-running heat engine, which can be characterized by the ratio of the extracted heat Q_c to the input work W_{in}. This ratio is called the coefficient of performance C,

$$C = Q_c/W_{\mathrm{in}} \tag{4}$$

Here, the second law of thermodynamics shows that the maximum value for C is $T_c/(T_h - T_c)$, which can exceed the value 1 (which occurs in common refrigeration devices such as room air conditions), while the first law requires that $W_{\mathrm{in}} = Q_h - Q_c$. (Note that the directions of heat flow and work in the cryocooler are reversed.)

An ideal heat engine follows a Carnot cycle. The Carnot cycle has four parts, which are illustrated in a *TS* plane in Fig. 5. This figure shows the heat engine going through a series of constant-temperature ($1 \rightarrow 2, 3 \rightarrow 4$) processes and constant-entropy processes ($2 \rightarrow 3, 4 \rightarrow 1$). These correspond to combinations of heating or cooling and compressing or expanding the fluid. In the Carnot cycle, the thermodynamic system plotted is usually a

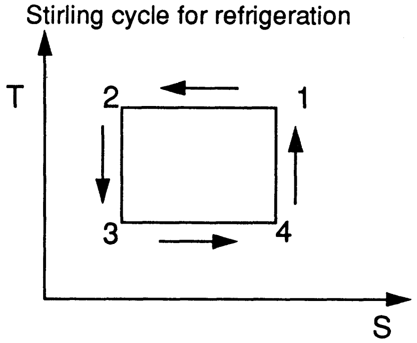

FIG. 5. The Stirling cycle as represented in a *TS* plane. The arrows indicate the direction taken when a Stirling engine is operating as a refrigerator.

single substance—often a gas. However, obtaining the heat switching necessary to have constant-entropy processes in a single substance is impractical.

2.2.1 Stirling Cryocooler. A Stirling heat engine uses a process approximating a Carnot cycle. In contrast to the Carnot engine, the Stirling engine uses two "working substances" that exchange heat between themselves so that the composite system follows a Carnot path. Figure 5 represents the ideal path for the refrigerator, but a real Stirling refrigerator relies on mechanical devices that only approximate the path. Also, friction and thermal gradients reduce the performance. Thus, the real coefficient of performance is less than $T_c/(T_h-T_c)$. Frequently, this cycle is shown in a pressure–volume (*PV*) plane; however, this presentation can be deceptive since a Stirling engine relies on a reversible heat source to produce adiabatic (constant *S*) changes in the temperature.

A still idealized model of a Stirling refrigerator using a gas is shown in Fig. 6. Here step 1→2 is accomplished by compressing the gas on the right (hot) side and simultaneously removing heat Q_h through a heat exchanger (not shown) to the high-temperature reservoir. The next step, 2→3, involves moving both pistons simultaneously with the gas delivering heat to the regenerator. Step 3→4 is the isothermal expansion of the left (cold) cylinder. Now, heat from the low-temperature side Q_c is absorbed.

In a real Stirling engine, the reversible heat source is called a regenerator and is typically made from an assembly of finely divided metal in the form of strips, wires, or tubes. This provides a large heat capacity, efficient heat exchange, and low heat conduction between the hot and cold ends. An actual Stirling engine only approximates the steps described in Fig. 6, since the mechanical linkages to produce these exact piston motions would be unnecessarily complex. On the other hand, smooth oscillatory motion of the pistons does not deteriorate the efficiency much. Stirling engines have a significant advantage in that they do not require valves, which makes them highly reliable. They are among the most efficient cryocoolers, and modern versions of these are used in both large-scale and small lightweight systems. Most modern military aircraft have Stirling cryocoolers on board to cool their infrared detection systems.

2.2.2 Joule–Brayton Recuperative Refrigerator. A Joule–Brayton recuperative cryocooler is very similar to a Stirling engine except in the manner in which the gas is precooled. The Joule–Brayton cryocooler uses a compressor, a counterflow (recuperative) heat exchanger, and an expander. A schematic diagram for this type of refrigerator is shown in Fig. 7. First the gas is compressed, which heats it up. The high-temperature heat exchanger brings the gas back to the temperature of the surroundings. The expander is a reversed compressor. Compressed gas is allowed to expand while doing work on the piston. The walls of the expander are poor heat conductors so that the process is adiabatic and nearly reversible. Thus the work done by the expanding gas on the piston in the expander removes internal energy from the gas, which results in a temperature drop in gas. After passing through the cold region, the returning gas flows through a counterflow heat exchanger which precools the incoming gas.

2.2.3 Joule–Thomson Refrigerator. The third refrigeration technique, the Joule–Thomson expansion, differs from the expander explained above because in this case the gas does no external work. Instead of the gas doing work on a piston, it does work on itself. This is the procedure that is most commonly used in air conditioning, home refrigerators, and freezers, where the working fluid is Freon 12 (a chlorofluorocarbon) or ammonia. A Joule–Thomson refrigerator uses a compressor to feed gas through a narrow

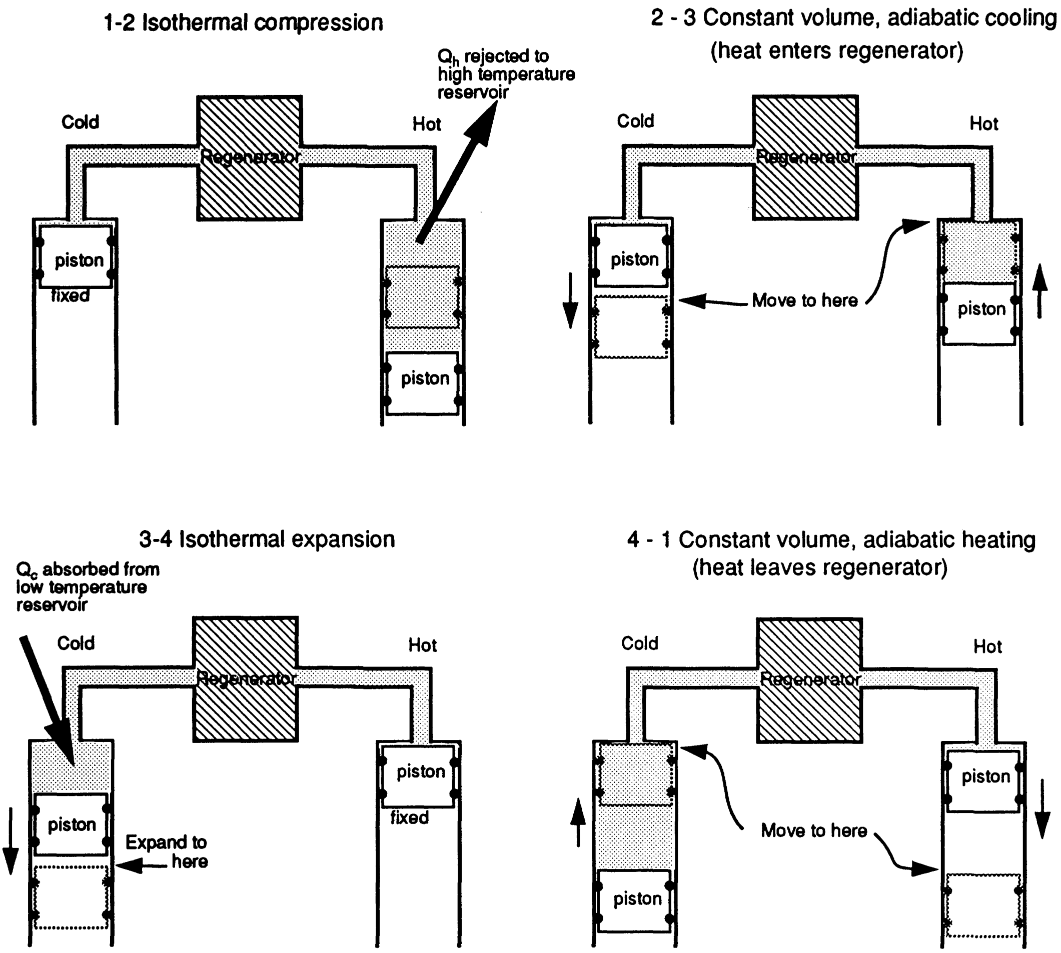
1-2 Isothermal compression
Q$_h$ rejected to high temperature reservoir
Cold
Hot
Regenerator
piston
fixed
piston
2 - 3 Constant volume, adiabatic cooling (heat enters regenerator)
Cold
Hot
Regenerator
piston
Move to here
piston
3-4 Isothermal expansion
Q$_c$ absorbed from low temperature reservoir
Cold
Hot
Regenerator
piston
Expand to here
piston
fixed
4 - 1 Constant volume, adiabatic heating (heat leaves regenerator)
Cold
Hot
Regenerator
piston
Move to here
piston

FIG. 6. An idealized model for a Stirling engine. The pistons are connected to external mechanical devices to achieve the correct sequencing of the motion.

restriction. This narrow restriction can be made with a piece of fine capillary tubing or a fine porous material. In many cryogenic refrigerators, the restriction can be adjusted to fine tune the performance of the refrigerator. Thus this restriction is often called a Joule–Thomson valve. We can easily show that a thermodynamic potential called the enthalpy is the same before and after the gas passes through the restriction. The enthalpy is defined by $H = U + PV$, where H is the enthalpy and U is the internal energy. In general, a constant-enthalpy expansion can cause either cooling or warming depending on the sign of the Joule–Thomson coefficient $\mu \equiv (\partial T/\partial P)_H$. In general, gases have a negative μ at high temperatures and warm when expanded, but have an inversion point where μ goes through zero and becomes positive at lower temperatures. For gases at temperatures below the Joule–Thomson inversion point, the result is a reduction in temperature. The maximum inversion temperatures, which also depend on pressure, for N_2 and O_2 are about 621 and 764 K—above room temperature—while the inversion temperatures for H_2 and He are about 202 and 40 K.

Although Joule–Thomson expanders can reduce the temperature of a gas at any starting temperature below the inversion point, they are efficient refrigerators and liquefiers only if the gas going into the Joule–Thomson valve can be precooled to significantly lower temperatures. As a rule of thumb, starting the

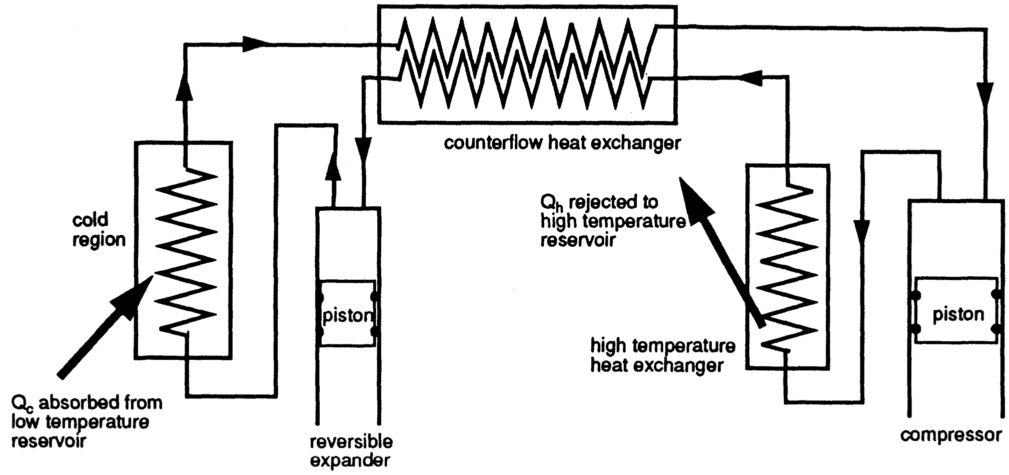
counterflow heat exchanger
cold region
Q$_c$ absorbed from low temperature reservoir
piston
reversible expander
Q$_h$ rejected to high temperature reservoir
high temperature heat exchanger
piston
compressor

FIG. 7. A schematic diagram of Joule–Brayton recuperative cryocooler. Heat is removed from the gas by allowing it to do work on the piston in the reversible expander.

expansion near the critical temperature of the liquid greatly increases the efficiency of the refrigerator.

A Hampson–Linde cycle (named for its inventors) uses a Joule–Thomson expansion and a recuperative heat exchanger. Figure 8 shows a schematic diagram of this cycle, and the path taken by the gas in a temperature–entropy diagram is shown in Fig. 9. The numbers in the *ST* diagram refer to the points in this Joule–Thomson refrigerator.

1→2 is a compression from the low-pressure side of the system at P_1 to the high-pressure side at P_2. This compression raises the temperature of the gas above the ambient (room) temperature, and as the gas is fed through the high-temperature heat exchanger, the temperature is returned to the ambient temperature, 2→3. Then the gas passes through the counterflow heat exchanger, where it is cooled by the returning gas, to point 4. The Joule–Thomson expansion, 4→5, follows a line of constant enthalpy and ends in the two-phase region. Thus part of the gas liquefies and can be removed with new make-up gas coming in at the intake of the compressor, or alternatively, the system can be used as a refrigerator using the latent heat of the liquid

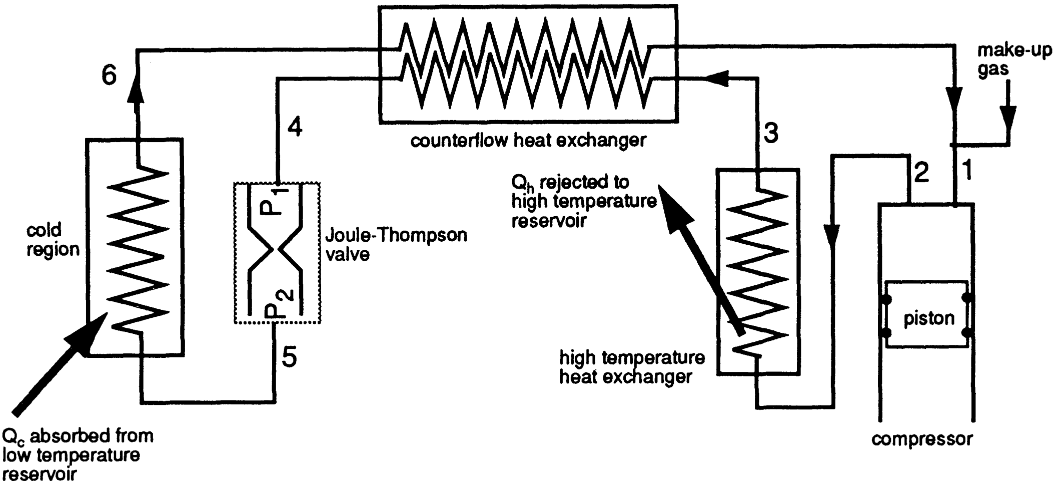
6
4
counterflow heat exchanger
3
make-up gas
2
1
cold region
P_1
P_2
Joule-Thompson valve
5
Q$_h$ rejected to high temperature reservoir
high temperature heat exchanger
piston
compressor
Q$_c$ absorbed from low temperature reservoir

FIG. 8. Joule–Thompson expansion refrigerator. Here, the compressed gas is first returned to close to room temperature and then further cooled by returning gas in the counterflow heat exchanger. The Joule–Thompson expansion results in a liquid–gas mixture if the inlet temperature to the Joule–Thompson valve is sufficiently low.

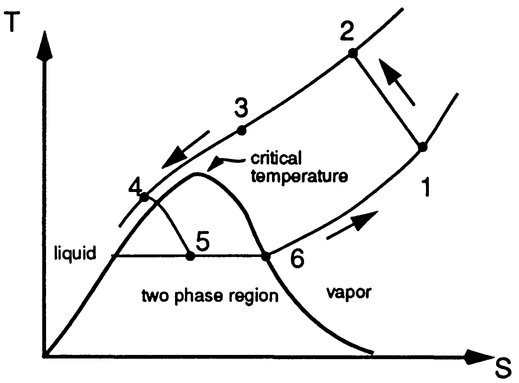

FIG. 9. A temperature-vs-entropy diagram for a typical Joule–Thompson refrigerator. The numbered points in the diagram correspond to points in Fig. 8.

in the two-phase region, 5→6. If the precool temperature (corresponding to point 4) is not sufficiently close to the two-phase region so that the Joule–Thomson expansion does not reach the two-phase region, this system still operates as a refrigerator, but at a greatly reduced efficiency.

By itself, a Joule–Thomson refrigerator can be used for the liquefaction of air and other gases with the exception of hydrogen and helium. However, this process is usually coupled with one or more other expansion engines. For example, Collins developed a helium liquefier during the late 1940s that used two Joule–Brayton type expanders followed by a Joule–Thomson expansion stage. This three-stage liquefier is now common to low-temperature laboratories throughout the world.

2.2.4 Evaporative Coolers—1 K, 0.3 K. Apparatus that requires temperatures below 4 K usually use liquid helium (^{4}He, which boils at 4.2 K at atmospheric pressure) as a starting point. In most low-temperature experiments that use liquid helium, the liquid is produced elsewhere and then transferred via an insulated line to a Dewar. The process of producing lower temperatures involves adding stages to the refrigeration process. As a first stage in cooling below 4.2 K, a simple evaporative cooler can be used.

Figure 10 illustrates a simple pumped-^{4}He refrigerator. The main Dewar holds a supply of ^{4}He at atmospheric pressure. Inside the Dewar, there is a leak-tight container ("can") whose high vacuum acts as insulator. Inside the vacuum can, a container holding liquid ^{4}He is connected to a vacuum pump, usually located outside the Dewar. A small amount of liquid ^{4}He is allowed to trickle into the "1-K pot" through a tiny capillary. The pump reduces the pressure in the 1-K pot causing the liquid ^{4}He to boil, removing the latent heat of vaporization and cooling the remaining liquid. As the temperature is reduced, the ^{4}He vapor pressure is reduced, as shown on the left-hand side of Fig. 10. The lower vapor pressure in turn lowers the amount of ^{4}He that can be pumped away so that eventually a steady-state temperature is obtained that depends on the vapor pressure and the size of the vacuum pump. With a good vacuum pump and fast pumping lines, temperatures near 1 K (more frequently nearer 1.2 K) can be obtained this way.

^{3}He is a rare isotope of helium containing one less neutron than the common isotope ^{4}He. Isotopically pure ^{3}He has significantly different properties from ^{4}He at low temperatures. We have included a sketch of the ^{3}He vapor-pressure curve in Fig. 10. Since ^{3}He has a higher vapor pressure, we can reach lower temperatures (about 0.3 K) by substituting ^{3}He for ^{4}He in the cryostat. Of course, in that case we are very careful to recover the ^{3}He from the back of the vacuum pump since it is very expensive.

2.2.5 Dilution Refrigeration. Most cryostats that reach temperatures below 0.5 K use a dilution refrigerator as the next cooling stage below the 1-K pot. A dilution refrigerator uses a mixture of ^{4}He and ^{3}He. This mixture separates into two phases below 0.86 K. At low temperatures, one phase (the "concentrated" phase) is nearly pure ^{3}He and the other phase is mostly ^{4}He with about 6.4% ^{3}He (the "dilute" phase). When concentrated ^{3}He is diluted, latent heat is absorbed in a way very much analogous to the vaporization of a liquid (such as the ^{4}He discussed above). A dilution refrigerator can operate either by having ^{3}He circulate with ^{4}He remaining stationary or with ^{4}He circulating. Dilution refrigerators with circulating ^{3}He are the most common form of these refrigerators.

The essential parts of this type of refrigerator are shown in Fig. 11. The entire low-temperature portion of the refrigerator is

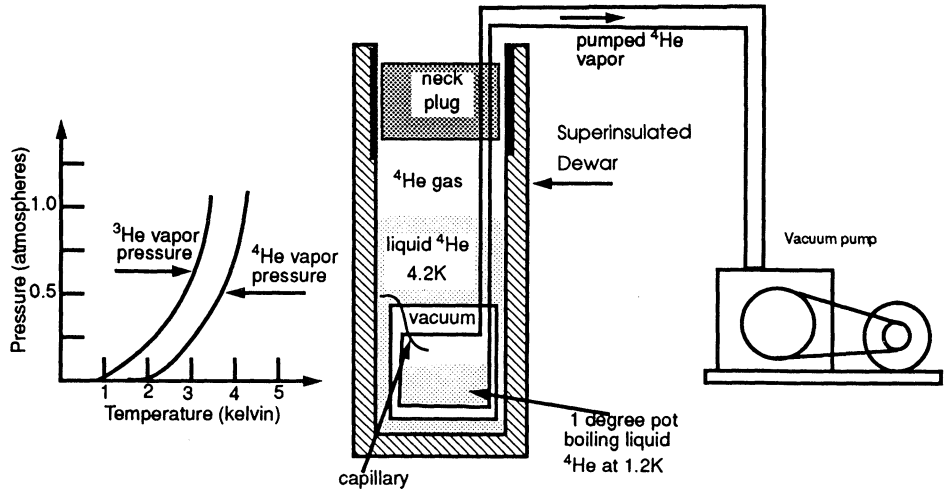

FIG. 10. A schematic diagram showing a Dewar with a 1-K refrigerator inside. The vacuum pump removes helium gas from the 1-K pot, causing cooling. The 1-K pot is refilled by a small capillary connected to the helium in the main Dewar. The inset on the left shows approximate values for the vapor pressures of ^{4}He and ^{3}He. When ^{3}He is used for refrigeration, a second ^{3}He pot is placed below the 1-K pot and the ^{3}He vapor is returned to the ^{3}He pot after passing through a series of heat exchangers.

surrounded by a vacuum can that is immersed in liquid ^{4}He in the same configuration as is shown in the ^{4}He refrigerator. With the dilution refrigerator, the cooling takes place in the mixing chamber, which has both concentrated and dilute phases in coexistence. The ^{3}He is moved through the refrigerator by pumping on the still, which is held near 0.8 K. In that temperature range, the vapor pressure of the ^{3}He is much greater than that of the remaining ^{4}He *even though the concentration of ^{3}He is small.* For this reason, the pump removes from the still almost pure ^{3}He, which is recovered and then returned. The continuous removal of ^{3}He from the still causes a depletion of ^{3}He in the mixing chamber, which in turn causes the concentrated ^{3}He to dilute continuously into the ^{4}He. This process is similar to evaporation, in which the concentrated ^{3}He is the liquid phase and ^{3}He in the dilute phase is analogous to a vapor. (Here the "vapor" phase lies below the "liquid" phase since liquid ^{4}He is denser than ^{3}He.) The temperature of the mixing chamber depends on how much heat it is removing, the size of the pumping system, and the efficiency of the heat exchangers that cool the returning ^{3}He. Dilution refrigerators with large pumping systems and very efficient heat exchangers are capable of reaching temperatures below 0.005 K (5 mK) at the mixing chamber, and some have even reached 2 mK.

Since the ^{3}He in both the concentrated and dilute phases is a Fermi liquid, a closer analogy can be drawn with the Peltier effect (Ebner and Edwards, 1971) in which current through a thermocouple junction cools (or heats). However, Peltier refrigerators do not function well much below room temperature because of the Joule (Ohmic) heating due to the electrical resistance of the wires.

2.3 Magnetic Refrigerators

2.3.1 Introduction to Spin Entropy and Adiabatic Demagnetization. Magnetic refrigerators take advantage of the heat energy associated with the magnetization of materials. These refrigerators are most easily understood in terms of entropy, with the heat required to raise the entropy being given by $\Delta Q = T\Delta S$. In the simplest case, a magnetic material can be thought of as a collection of small magnetic moments. With no magnetic field, the orientation of the spins is random. In a magnetic field, the individual moments have an energy given by

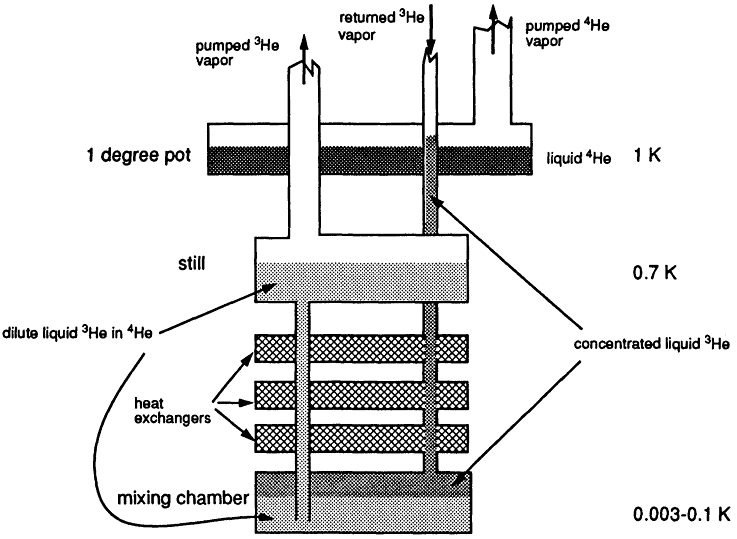

FIG. 11. A schematic diagram of a dilution refrigerator. The parts shown are placed inside a vacuum can that is inside a Dewar similar to that in Fig. 10. The still is connected to an external vacuum pump that removes mostly ^{3}He from the dilute liquid ^{3}He mixture. The 1-K pot acts as a precooler to condense the returning ^{3}He vapor. The heat exchangers further cool the returning ^{3}He. The mixing chamber removes the latent heat of mixing, which causes the mixing chamber to cool. Typical values for the temperatures of the various stages are indicated on the right.

$$\epsilon = -\mathbf{m}\cdot\mathbf{B}, \tag{5}$$

where **m** is the magnetic moment and **B** is the magnetic field. At high temperatures, where the thermal energy $k_B T$ is greater than ϵ, the orientation of the spins is almost random. In this situation, pictured in Fig. 12(a), the number of possible rearrangements of the spins having the same total magnetization is large. Thus according to Eq. (2), the entropy at high temperatures is large. At lower temperatures where $\epsilon \cong k_B T$, the individual moments are partially aligned, as shown in Fig. 12(b), and the entropy is reduced. At very low temperatures, the spins are almost completely aligned by the field, as shown in Fig. 12(c). Here there remain very few rearrangements of the spins consistent with the same magnetization and so the entropy is near zero. Figure 13 shows a sketch of the entropy versus temperature at two magnetic fields. For isolated spins, the entropy reduction is a function only of the ratio B/T.

2.3.2 Demagnetization of Paramagnetic Materials. Figure 13 shows a typical demagnetization cycle. In step 1→2 the material is magnetized in a large magnetic field. An external refrigerator connected to the sample through a "heat switch" removes the entropy. The magnetic material is then thermally isolated and the field reduced in step 2→3. Ideally, there is no heat flow, so that this step is adiabatic and we see that the temperature is reduced. Heat flowing into the material eventually warms it up, as indicated by step 3→4. Finally the magnetic field can be increased to complete the cycle, making a magnetic heat engine; however, most practical magnetic refrigerators act as "one-shot" systems only using steps 1→2 and 2→3.

Many materials can be used to make a magnetic refrigerator. The choice of material depends on the size of the magnetic moment and the strength of the residual interactions between the magnetic moments. If the magnetic moment is small, the entropy reduction during the magnetization is small, and the refrigerator will not be effective. Interactions between magnetic moments cause magnetic ordering (either ferromagnetic or antiferromagnetic), which reduces the entropy and makes the entropy independent of the magnetic field. Thus, very little magnetic cooling can occur at temperatures below which magnetic ordering occurs.

At high temperatures Gd is frequently used for magnetic refrigeration. Below 1 K a number of rare-earth salts are used. The most common of these in low-temperature laboratories is CMN (cerium magnesium nitrate). This salt orders ferromagnetically near 2 mK (0.002 K), and this sets the minimum temperature that it can achieve.

2.3.3 Nuclear Demagnetization. Nuclear demagnetization follows the same process as that for paramagnetic materials except that the magnetic moments that are used are the nuclear moments. While these moments are much smaller, the interactions between the

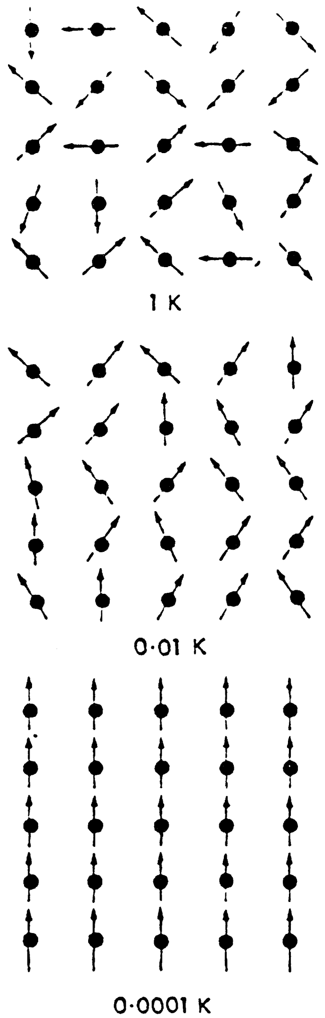

FIG. 12. Arrangements of magnetic spins at a high, intermediate, and low temperatures. The high-temperature arrangement is nearly random with a small total magnetization. There are many arrangements of spins with small magnetization causing a high entropy. At the intermediate temperature, the magnetic energy is approximately the same as $k_B T$. At low temperatures, the spins are aligned with a higher magnetization, but fewer possibilities. Thus the entropy is low. (Mendelssohn, 1977, p. 168.)

moments are also small. A nuclear demagnetization apparatus typically uses a dilution refrigerator to precool a sample such as copper in a high magnetic field (8 T or more) to temperatures near 10 mK. Sometimes, two stages of nuclear demagnetization are used with the first stage providing the precooling for the second. The entropy reduction may be less than 10% of the total entropy. Nevertheless, these refrigerators are capable of reaching extremely low temperatures because these systems, which have small interactions, do not order (ferromagnetically or antiferromagnetically) until the temperature is much smaller than in the case of electron paramagnetic materials. Nuclear demagnetization cryostats hold all of the current low-temperature records, as we will discuss in the last section.

A diagram of a nuclear demagnetization cryostat is shown in Fig. 14. The materials that can be used for nuclear demagnetization include In, Cu, Ag, and even Tl and Pt. Cu is used in the greatest number of nuclear demagnetization cryostats by far.

There is a group of metal alloys with rare earths that have a hybrid form of magnetic moment through which the atomic polarization enhances the field at the nucleus, which gives a larger effective moment. This is called Van Vleck enhancement of the nuclear moment, and these alloys can be very effective for nuclear demagnetization (Andres, 1978). The most effective of these Van Vleck alloys is $PrNi_5$, which can be used to reach temperatures in the range of 0.5 mK.

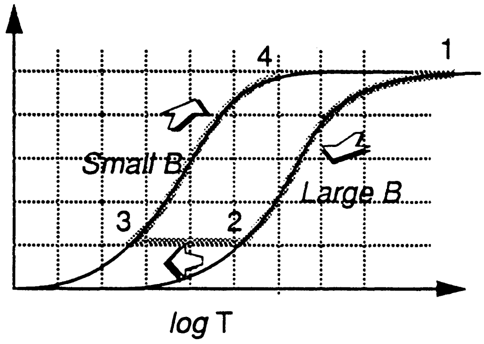

FIG. 13 The path taken in entropy-vs-temperature space for a magnetic refrigerator. The magnetic material is precooled in large magnetic fields (1 → 2), then demagnetized while it is thermally isolated (2 → 3). Heat is removed from the surroundings at low field (3 → 4), and finally the field is raised which increases the temperature (4 → 1).

3. MEASUREMENT OF LOW TEMPERATURES

3.1 Temperature Scales

3.1.1 The Absolute Temperature Scale. The definition of the true thermodynamic temperature is subtle and annoyingly unsatisfying. The intuitive sense of temperature, being what registers on a thermometer when it has come into thermal equilibrium with the material to be measured, is, however satisfying, insufficient to determine uniquely an absolute universal scale. A correct definition of temperature is that it is the quantity that connects entropy changes and heat exchanges in reversible processes:

$$dS = \text{đ}Q/T. \tag{6}$$

Equivalently, $1/T$ is the integrating factor that converts differential heat exchanges into a perfect differential, so that thereby entropy is a state variable.

Obviously, every practical thermometer does not immediately refer to this definition, but each does have a connection to it through whatever chain of calibrations establishes its accuracy.

The definition implicitly defines $T=0$ at the absolute zero of temperature, so that all that is required for a complete scale is an overall scale factor, set by defining one special temperature. That special place is chosen by in-

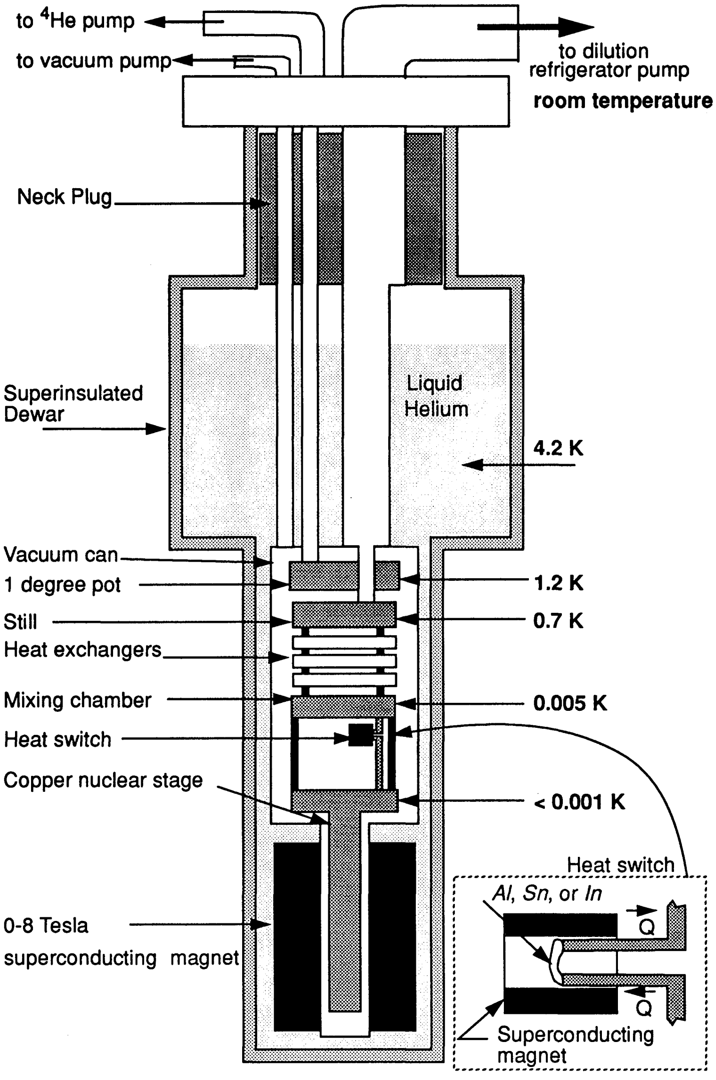
to ^{4}He pump
to vacuum pump
to dilution refrigerator pump
room temperature
Neck Plug
Superinsulated Dewar
Liquid Helium
4.2 K
Vacuum can
1 degree pot
Still
Heat exchangers
Mixing chamber
Heat switch
Copper nuclear stage
1.2 K
0.7 K
0.005 K
< 0.001 K
0-8 Tesla superconducting magnet
Heat switch
Al, Sn, or In
Q
Q
Superconducting magnet

FIG. 14. Nuclear demagnetization. This cryostat combines the features of several refrigeration processes. The main components are indicated on the left. Three pumping lines are connected to external pumps and gas recovery systems as shown. The bold printed temperatures indicated along the right-hand side of the figure are typical operating temperatures of the various stages in the cryostat. Liquid ^{4}He is delivered to the main Dewar. A 1-K pot maintains 1.2 K. The dilution refrigerator provides precooling for the nuclear demagnetization stage, which is connected by a heat switch. When the demagnetization stage is isolated from the dilution refrigerator, the main magnet field is lowered, lowering the last stage to submillidegree temperatures. The inset shows a heat switch using a superconducting bridge between copper rods.

ternational convention to be the triple point of water (at which ice, water, and water vapor are simultaneously in equilibrium, $P=4.579$ Torr$\approx$6.1 mbar) and defined to be $T=273.16$ K, where K stands for the international unit of temperature, the kelvin (not degree kelvin). The Kelvin scale is used principally in scientific work, particularly at low temperatures.

3.1.2 Other Temperature Scales. There are three other temperature scales in common usage today. Each was originally defined by choosing two special points to define two temperatures. Each is thus linearly related to the Kelvin temperature scale. Closely connected to the Kelvin temperature scale is the Celsius scale, with temperature changes numerically equal in each. Today the definition is $T(^\circ\mathrm{C})=T(\mathrm{K})-273.15$. The original definition of this scale used the equilibrium temperature of melting ice at 1 atm as 0 °C, and the normal boiling point of water (where the pressure is 1 atm) as 100 °C. This exact 100 °C interval was preserved (to the then existing level of accuracy of experiments) by the specific choice of Kelvin scale factor given by choosing the triple point of water temperature given above. Accounting for the negative slope of the water–ice coexistence curve, the temperature of the normal melting point is about 0.003 °C. The Celsius scale, having zero closer to ordinary human conditions than the Kelvin scale, is the most commonly used temperature scale in the world, and is used in most scientific work near room temperature.

An earlier scale is the Fahrenheit scale, originally defined to be 0 °F at the lowest temperature then available (ca. 1700, with a salt–ice mixture) and 12 °F (carrying over a proposal by Isaac Newton) at normal human body temperature, later multiplied by eight to become 96 °F. (Nonlinearities in early thermometers led this number to differ from the now standard value 98.6 °F.) This scale is currently defined in terms of the Celsius scale as $T(^\circ\mathrm{F})=\frac{9}{5}T(^\circ\mathrm{C})+32$. It is the one most commonly used in the United States, but is (very) slowly being displaced in favor of the Celsius scale.

Finally, analogous to the relationship between the Kelvin and Celsius scales, in which the size of a unit temperature change is the same but with a zero offset, is the relationship between the Fahrenheit and Rankine scales. The Rankine scale is an absolute Fahrenheit scale and defined equivalently by either $T(^\circ\mathrm{R})=T(^\circ\mathrm{F})+459.67$, or $T(^\circ\mathrm{R})=\frac{9}{5}T(\mathrm{K})$.

3.1.3 International Temperature Scale of 1990 (ITS-90). The current international temperature scale, ITS-90, was established at the 17th session of the Consultative Committee on Thermometry (CCT) in September 1989 (Preston–Thomas, 1990). The CCT is one of eight Consultative Committees of the International Committee of Weights and Measures, which is, in turn, a committee of the General Conference of Weights and Measures, the authority, by international agreement, for the International System of Units (SI). The CCT is composed of representatives of national standards laboratories throughout the world and of other laboratories concerned with the accurate definition of temperature.

ITS-90 establishes temperatures down to 0.65 K. It also defines overlapping ranges and the functions to be used in calibrating the various thermometers used in each range. (See Fig. 15.) From 0.65 to 5.0 K, temperature is defined as a specified function of the vapor pressure of each of the helium isotopes in separate subranges. From 3.0 to 24.5561 K and from 13.8033 to 1234.93 K the scale is defined in terms of, respectively, a constant-volume gas thermometer using each of the helium isotopes and a platinum resistance thermometer. Up to room temperature each

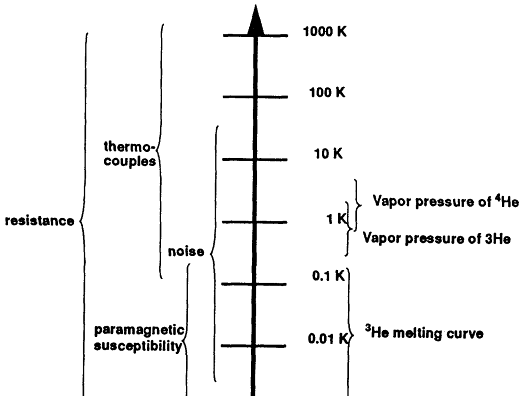

FIG. 15. A logarithmic temperature scale indicating the various types of thermometry that are used.

Table 1. Summary of classes of thermometers for use at cryogenic temperatures.

Class	Theoretical foundation	Examples	Usability	Range	Accuracy	Sample-to-sample reproducibility
Primary	Rigorous, absolute	Ideal gas law, 2nd law, nuclear orientation	Generally poor	Limited	High	Excellent
Secondary	Strong	NMR, noise	Poor to good	Limited	Needs calibration point	Good
Tertiary	Weak	Resistance, diode voltage, thermocouples	Good	High	Needs calibration curve	Poor
Transfer standards	None	Vapor pressure of ^{3}He and ^{4}He, solidification pressure of ^{3}He	Poor	Limited	High	Excellent
Transfer standards	None	Triple points, superconducting transitions	Poor	None	High	Excellent

thermometer is calibrated at the triple points of hydrogen (with equilibrium ortho–para ratio), neon, oxygen, argon, mercury, and water at 13.8033, 24.5561, 54.3584, 83.8058, 234.3156, and 273.16 K, respectively.

ITS-90 superseded the previous temperature scale, known as ITPS-67(75), by extending its lower bound below the triple point of H_2 and changing temperatures below the triple point of H_2O by a mean amount of 10 mK. It also provides better agreement with thermodynamic values than the 1976 Provisional 0.5 K to 30 K Temperature Scale (EPT-76).

3.2 Classes of Thermometers

All thermometers operate by measuring some physical property of a material that changes with changing temperature. Every thermometer can be classified on the basis of its ability to define the absolute temperature. The classification scheme given here is consistent with general use, although there is no universally recognized definition. Examples of each class of thermometer will be given in the following section. These classes are summarized in Table I.

A primary thermometer is one that is capable of defining the temperature by itself, without reference to any other device. It must therefore be grounded in a rigorous thermodynamic principle. For example, a gas thermometer measuring the pressure P of a fixed amount of gas and comparing it to the pressure when in equilibrium with water at its triple point, P_{tr}, would define temperature as

$$T = 273.16 \lim_{P_{\text{tr}} \to 0} \left(\frac{P}{P_{\text{tr}}} \right), \tag{7}$$

where the limiting process successively reduces the amount of gas in the volume to eliminate the effect of intermolecular interactions. The reliance on the ideal gas law and the experimental production of an ideal gas (through the limiting process) makes this a good primary thermometer, although in practice it is rather limited. In general, the reliance on thermodynamic principles renders primary thermometers independent of the working substance, so that reproducibility from sample to sample (and, more importantly, from laboratory to laboratory) is excellent.

A secondary thermometer is one in which an absolute temperature can, in principle, be measured (at least in some approximation) but for which technical difficulties prevent absolute calibration. For example, the nuclear susceptibility in platinum (far from any ordering temperature) to a high accuracy obeys a Curie law with a known coefficient. The difficulty, however, is that an absolute calibration would require an accurate knowledge of the coupling of the platinum nuclei to the thermometer's electronics. Usually the process of determining this coupling is foregone in favor of calibrating at some known point. In principle, a single calibration point is sufficient to define a secondary thermometer's entire absolute scale.

All other thermometers without simple connection to fundamental thermodynamic principles, which includes almost all thermometers in use, are classified as tertiary. Certainly the physics of their operation is well known, but difficulties in a first-principles calculation prevent their being classified as secondary. For example, the voltage developed by a thermocouple arises from the temperature-dependent chemical potential of the two involved materials, but this is exceedingly difficult to evaluate even for simple metals, let alone the metal alloys usually used in practical thermocouples.

Finally, we mention the role of transfer standards and fixed points. These provide an excellent mechanism to transfer temperature scales between laboratories. These are exclusively associated with phase transitions (fixed points simply being the intersection of at least two phase transitions) and are thus not calculable from first principles without a known equation of state, which we do not have. These transitions can be used either directly, as in the case of ^{3}He melting-curve thermometers in the low-millikelvin regime, or indirectly, to calibrate other, potentially more convenient, thermometers such as by using the pressure-dependent superfluid transition temperature in ^{3}He, $T_c(P)$, which establishes a reliable scale from 0.9 to 2.5 mK (Greywall, 1986).

3.3 Practical Thermometers

A good thermometer should have high sensitivity, short internal thermalization time constants (so that the thermometer itself will have a well-defined temperature), good thermal contact with the system, and negligible total heat capacity so as not to load thermally the system it is measuring (which may be achieved simply by making the thermometer physically small), and should introduce negligible heat to the system through its operation. It should furthermore be insensitive to all environmental influences other than temperature. A good thermometer for one measurement may be a poor one under different conditions. We summarize here some of the thermometers used in cryogenic measurements. The ordering of these thermometers generally follows their useful temperature range, in decreasing order.

3.3.1 Constant-Volume Gas Thermometer. As mentioned above, the pressure of a fixed amount of an approximately ideal gas can be used, although corrections for nonideality are important. This thermometer can be used until the pressure becomes so small that thermomolecular corrections become significant. This limits its practical range to that above several kelvins. For precision metrology there is also a significant problem associated with adsorption of the gas on its container's walls. Even for the noble gases, the binding energy for initial layers to surfaces is 100 K or more, which can significantly affect the inferred molar volume if the surface-to-volume ratio is not negligible.

3.3.2 Thermocouples. Thermocouples have a long history of importance, being used in the definition of the Kelvin scale until the advent of the current scale, ITS-90. They have the advantage of simplicity, but the disadvantage of relatively small voltage signals which vanish as $T \rightarrow 0$. Below 10 K they are generally not useful, although special Kondo-enhanced thermocouples (see HEAVY-FERMION PHENOMENA) have been demonstrated to work in a limited range around 10 mK.

3.3.3 Cryogenic-Liquid Vapor Pressure. The vapor pressure of cryogenic liquids can be used as a thermometer. The ITS-90 is, in fact, *defined* in terms of the helium liquid vapor pressures from 0.65 to 5.0 K. This technique has a range limited from above by the liquids's critical point and from below by the rapidly vanishing vapor pressure, decreasing as $\exp(-\Delta/k_BT)$, where Δ is a constant (at low temperatures) proportional to the latent heat of vaporization. Even using ^{3}He, with its small latent heat of vaporization, the lowest measurable temperatures with this technique are only 0.3 K.

3.3.4 Resistance Thermometers. One of the simplest tertiary thermometers to use is a resistance device. Any material with a temperature-dependent resistance can be used. Typical materials are carbon (including ordinary carbon composition resistors), germanium, platinum, and numerous alloys. All resistors exhibit some degree of magnetoresistance so that calibrations are distorted by modest (few teslas) magnetic fields, but most resistors retain sensitivity to thermal changes, which may be sufficient in a given application. Off-the-shelf carbon resistors are convenient, but vary in their characteristics from batch to batch and may show significant irreproducibility during the first few thermal cyclings.

The low-temperature limit is set by different effects for different materials. For carbon resistors, the resistance of which climbs with decreasing temperature, the low-temperature limit is set by the maximum resistance measurable without significant interference from spurious effects, such as leakage resistance. A typical maximum resistance is 1 MΩ. For commercial thermometers the low-temperature limit is around 1 K, although uncalibrated carbon resistors routinely used in ultralow-temperature laboratories can be used down to a few millikelvins. Carbon resistors generally also have a problem with internal thermalization in the millikelvin regime because they are poor thermal conductors.

Platinum resistors decrease in value with decreasing temperature. Their practical low-temperature limit is about 10 K, which arises from their vanishing temperature coefficient at low temperatures and the power dissipation from the temperature-independent part of the resistance.

3.3.5 Diode-Voltage Thermometers. Closely connected to resistance thermometers are diode thermometers. The difference between diodes and resistors is only in the diode's nonlinear current–voltage relationship. For a fixed current, as is used conventionally, the two are identical. Driven with 10 μA forward current, silicon diodes yield a sensitivity of 2–3 mV/K from room temperature down to about 30 K when the carriers begin to freeze out and the sensitivity increases to 50–100 mV/K. GaAs diodes have a similar sensitivity at high temperatures, but this sensitivity decreases below about 20 K. The low-temperature limit is about 1 K, set by the rather large power dissipation of 10 μW.

3.3.6 Capacitance Thermometers. Capacitance thermometers use the temperature dependence of the dielectric susceptibility of ferroelectric (and other ordered) crystals. They have the advantage of insensitivity to magnetic fields and are used principally to regulate temperatures as magnetic fields are applied. Capacitance thermometers utilizing $SrTiO_3$ exhibit a nonmonotonic temperature dependence, peaking around 65 K. They also suffer from irreproducibility when thermally cycled.

3.3.7 Ferromagnetic and Antiferromagnetic Frequency Shifts. The resonance frequency of nuclei in solids depend sensitively upon the internal magnetic fields. Therefore there have been a number of proposals to measure temperature via nuclear frequency shifts in magnetically ordered materials. By choosing the appropriate ferromagnet or antiferromagnet, essentially any temperature interval can be covered. The disadvantages of this technique are the sensitivity to sample preparation and the narrow temperature interval over which this technique has high sensitivity.

3.3.8 Electron Paramagnetic Susceptibility. The magnetic susceptibility of electrons in paramagnetic salts provides a convenient thermometer at low temperatures. For example, by choice of a material with a suitable magnetic ordering temperature (and by use of a large measuring field), temperature resolution of 1 nK near the λ point of ^{4}He can be achieved (Chui and Lipa, 1984; Steinberg and Ahlers, 1983; Marek, 1987). Below 1 K a convenient material is cerous magnesium nitrate (CMN), which has a large susceptibility, but a relatively low ordering temperature of 1.8 mK. In the low-millikelvin regime it is much more common to dilute the magnetic Ce ions with nonmagnetic La in a 1:19 ratio. This material is conventionally named LCMN. The temperature dependence of the susceptibility of LCMN is accurately given by a Curie–Weiss form, $\chi \propto 1/(T-\Delta)$, down to 0.3 mK, although this needs to be checked for each sample, given that the randomness of the La-for-Ce substitution means that there must exist CMN regions which are ordered.

3.3.9 Nuclear Susceptibility. Analogous to the magnetic susceptibility of electrons, the nuclear susceptibility also makes a convenient thermometer that is very common in ultralow-temperature laboratories. Because of the much smaller magnetic moments of nuclei relative to electrons, the susceptibility is smaller, pushing down the upper temperature limit of usability to a few tens of millikelvin. Also because of the reduced moments, the nuclear interactions are smaller than electronic ones, suppressing the ordering temperatures and pushing down the low-temperature limit of usability. As long as the temperature is significantly above 1 μK, well above the nuclear ordering temperatures, the susceptibility is accurately described by a Curie form, $\chi \propto 1/T$. For purposes of measuring temperatures in the low-millikelvin regime, the most commonly used material is platinum, which has a spin-lattice relaxation time of about (30 mK)/T s. As usually implemented, pulsed NMR spectrometers show pulse-to-pulse amplitude reproducibility of only about 1%, requiring considerable averaging for high accuracy.

3.3.10 ^{3}He Melting Pressure. An interesting thermometer used in several ultralow-temperature laboratories uses the melting pressure of ^{3}He directly as a thermometer. All that is needed to measure temperature is an accurate calibration of a pressure gauge. Zero offsets can be eliminated by adoption of the A transition of superfluid ^{3}He as the reference point. The melting curve is roughly independent of magnetic field for small fields ($B \leqslant 0.5$ T). This thermometer suffers from a rather large thermal load, but this is not significant for experiments studying superfluid ^{3}He or using it in an auxiliary role. This thermometer loses sensitivity at the low end below about 1 mK, where the solid orders magnetically. It is also insensitive at temperatures approaching the minimum in the melting curve around 0.3 K.

3.3.11 Noise Thermometry. Thermal noise (known as Johnson noise or Nyquist noise) is a limiting factor in resistance thermometry. However, this noise can itself be used as a thermometer directly. The maximum available noise power per unit bandwidth of any device is k_BT. In the case of a pure resistor of value R, at temperature T, thermal fluctuations of local charge density produce an effective voltage source in series with the resistor with mean square amplitude in a frequency interval Δf of $V^2 = 4k_BTR\Delta f$. The fundamental difficulty at low temperatures is efficient coupling of the noise into a detector that itself has a noise temperature small compared to that being measured, for which several solutions have been attempted. With commercial rf superconducting quantum-interference devices (SQUIDs), detector noise temperatures of order 1 mK have been demonstrated in noise thermometers. In principle, this thermometer could be a primary thermometer, but difficulties in precisely specifying the system transfer function over the entire bandwidth for which noise contributes power has limited devices to date to secondary status, requiring one calibration point to fix the temperature.

3.3.12 Nuclear Orientation. Nuclear orientation thermometry is a primary thermometer in that all of its parameters can be specified by studies of radioactive decay modes. The fundamental idea is that γ rays associated with a chain of radioactive decays are not emitted isotropically, but rather have a calculable probability of being emitted at any specified angle with respect to an axis easily defined by an applied magnetic field. The radiation pattern is isotropic at high temperatures and is maximally anisotropic at $T=0$. By measuring the anisotropy one can deduce the temperature. This thermometer suffers from self-heating associated with the nuclear decay by-products. Reducing the activity of the source to reduce the heating increases the time required to count decays. Furthermore, the accuracy of this device is not high, limited by instrumental difficulties to about 1%. Although there is no shortage of nuclei and hosts to consider, it appears unlikely that nuclear orientation will be a useful primary thermometer below several millikelvin.

3.3.13 Quasiparticle Damping Thermometry. A unique thermometer developed at the University of Lancaster, United Kingdom, measures the damping of a wire vibrating in liquid ^{3}He. At the lowest temperatures, the damping arises from individual quasiparticles being scattered from the wire. Since the number of quasiparticles as a function of temperature is known, this can be a good secondary thermometer. At temperatures above about 500 μK the damping becomes dominated by ordinary hydrodynamic viscosity, but this device is still useful as a tertiary thermometer. Since the ^{3}He itself is the thermometer, this avoids problems of thermal contact between ^{3}He and some other thermometer. This device has been used to measure the lowest temperatures in ^{3}He and has been used in dilute mixtures with some attendant controversy.

3.3.14 Nuclear Quadrupole Thermometry (NQT). A speculative technique being investigated by several laboratories uses a small Zeeman splitting (see ZEEMAN AND STARK

EFFECTS) of nuclear quadrupole levels to establish a primary thermometer. Nuclear resonance measures transitions between different quadrupole levels, with slightly shifted frequencies arising from the Zeeman fine structure. The ratio of the spectral weights of the different transitions depends upon relative populations of different energy levels E_i, which are determined by a Boltzmann factor $\exp(-E_i/k_BT)$. The possibility of such a thermometer has been demonstrated using a variety of nuclei at temperatures from 10 mK to 1 K.

3.3.15 Second-Law Thermometry. Finally, at the coldest temperatures yet reached, none of the above thermometers is reliable. Fortunately, the second law of thermodynamics can be used directly. The basic idea is to measure some thermometric property (such as magnetization) from which we can deduce the entropy while injecting a known amount of heat $\text{đ}Q=TdS$. The temperature is then immediately deducible.

4. HOW LOW CAN WE GO?

The second and third laws of thermodynamics state that the entropy of any system at zero temperature is zero, but the efficiency of removing entropy also approaches zero at zero temperatures. For this reason, $T=0$ stands as one of Nature's ultimate limits, much like the speed of light in a different context.

4.1 Negative Temperatures

In any discussion of low-temperature limits, it is natural to consider the possibility of exceeding the ultimate limit. While it is impossible to achieve negative temperatures solely by exchange of heat between reservoirs, it is possible to produce negative temperatures by nonthermal means.

In a discussion of *low*-temperature limits, it may appear odd to discuss negative temperatures because, simply put, negative temperatures are hotter than positive temperatures. However, the theoretical considerations involved in producing negative temperatures are identical to those in producing the current ultralow-temperature limits.

The range of systems in which negative temperatures can be achieved is quite limited. Consider an ensemble of two-state systems with energy levels $E_i = \pm\Delta/2$, as shown in Fig. 16, such as a spin-$\frac{1}{2}$ system in a magnetic field where $\Delta=\mu B$. In thermal equilibrium, the probability of level i being occupied is proportional to the Boltzmann factor $\exp(-E_i/k_BT)$. At ordinary (positive) temperatures this is consistent with the usual perception that lower-energy levels are more populated than higher-energy levels, as shown in Fig. 16(a).

If, however, by some means, such as magnetic resonance or rapidly reversing the field direction, the upper level can be made more populated, then the effective temperature describing this situation would be negative, as shown in Fig. 16(b). In fact, exactly this procedure of rapidly (quasiadiabatically) reversing the field direction has demonstrated negative temperatures in silver nuclei and their ferromagnetic tendency in the *minus* 10-nK temperature range.

What was required for this description to work? The obvious requirement was that the spectrum of energy levels needed to be bounded from above. Otherwise, at negative temperatures, the levels at infinitely high energy would be infinitely occupied. This is why a classical gas with translational kinetic energy (unbounded above) cannot exhibit negative temperatures. Since energy levels are

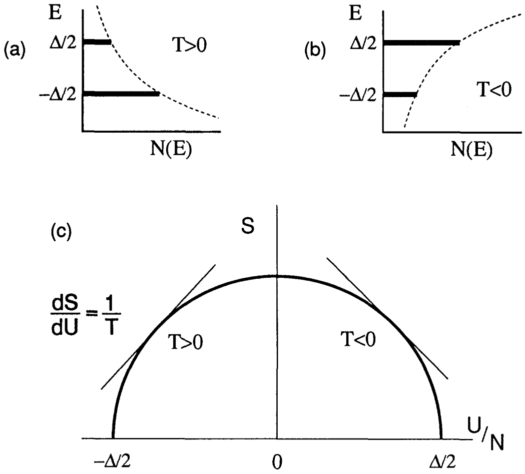

FIG. 16. Negative temperatures. The relative occupation probability of different energy levels in thermal equilibrium defines a temperature which can be (a) positive or (b) negative. In (c) the same problem is viewed from the perspective of the entropy, suggesting an alternative scale of temperature.

always bounded below, there is no converse argument against positive temperatures.

A more subtle requirement is that the system itself must be in thermal equilibrium. That is, there is a hierarchy of characteristic times that must be properly ordered. The characteristic time for the energy in the system to distribute itself thermally must be short compared to the time scale on which temperature is being defined by the experimental measurement, which in turn must be short compared to the time scale on which external influences change the system's temperature. This is analogous to the requirements given earlier for a thermometer to measure accurately the temperature of its environment.

The requirement of fast internal equilibration is somewhat obscured in the two-state example because any ratio of occupation probabilities uniquely defines some temperature, $N(E_1)/N(E_2)=\exp[(E_2-E_1)/k_BT]$. In an N-state problem, however, there are $N-1$ independent ratios of occupation probabilities, all of which must be derivable from one unique temperature if that quantity is to make sense. Given an arbitrary nonthermal disturbance of population levels, these $N-1$ ratios will not yield a unique temperature during an initial time interval before energy can be exchanged within the system.

An entirely equivalent description of negative temperatures is given in Fig. 16(c), where the entropy of the ensemble is shown as a function of its total internal energy. The derivative of this function is, by definition, the inverse of the temperature. The evolution of the temperature as a function of increasing energy can be read off easily. When the energy is minimum, the slope dS/dU is infinite, so that $T=0$. As the energy increases, the (positive) slope decreases until it reaches 0, at which point $T=+\infty$. Further increases in the energy cause the slope to decrease so that the temperature continuously evolves from $-\infty$ to -0, at which point the energy is a maximum.

This leads naturally to the frequent suggestion for an alternative scale of temperatures given by the negative slope $-dS/dU=-1/T$. This scale, called the "minus beta" scale ($-\beta=-1/k_BT$) or names equally unpoetic, satisfies the intuitive sense that it increases with increasing energy, and has the further advantages of correctly ordering negative and positive temperatures, and providing a ready answer to the question of why absolute zero is unattainable: you cannot reach infinity.

4.2 Present Ultralow-Temperature Limits

Finally, we summarize the current state of the art in achieving low temperatures. A logarithmic temperature scale showing the present limits of low-temperature technology is shown in Fig. 17. This is a dynamic field, and by the time this article appears in print some of the records listed here will be out of date. Nonetheless, here is a snapshot of the lowest records known to us as of August 1991.

The reason that there is more than one record is that for different systems there are different tradeoffs. For example, when trying to reach the lowest temperatures in a nuclear system, it is advantageous to leave the electrons at an elevated temperature to act as a larger buffer against heat leaks from the outside. The thermometers used at the end of the temperature scale here are also working at their limits, so that some revision in the reported temperatures is to be expected.

The lowest temperature achievable in a continuous mode is approximately 2 mK (Frossati, 1978). Short of mechanical failure or conscious decision to stop, this temperature could be maintained indefinitely. All of the remaining records listed use one or more stages of nuclear refrigeration, and so are intrinsically limited in duration, though, as mentioned earlier, this duration can be on the order of weeks at temperatures only fractionally above the low limit.

The lowest temperature at which a metal is known to become superconducting is 325 μK for rhodium (Buchal *et al.*, 1983) as discovered in the laboratory of F. Pobell, now at the University of Bayreuth in Germany.

In trying to establish a low temperature, both a low specific heat and low thermal boundary resistance are desirable. Dilute solutions of ^{3}He in ^{4}He do not have either of these advantages. The lowest measured temperature in these dilute mixtures is about 220 μK (Owers-Bradley *et al.*, 1983), although this measurement was limited by thermal contact to the thermometer so that the actual temperature may well be considerably lower.

Moving to ^{3}He lowers the boundary resistance, but increases the problem of high spe-

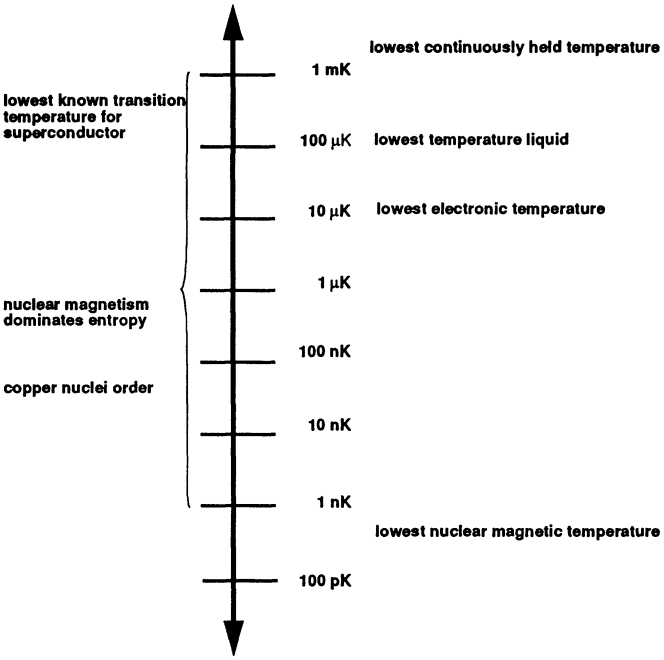
lowest continuously held temperature
1 mK
lowest known transition temperature for superconductor
100 μK
lowest temperature liquid
10 μK
lowest electronic temperature
1 μK
nuclear magnetism dominates entropy
100 nK
copper nuclei order
10 nK
1 nK
lowest nuclear magnetic temperature
100 pK

FIG. 17. A logarithmic temperature scale showing the approximate locations of the lowest temperatures that have been achieved to date and some of the new physical phenomena observed in the submillidegree temperature regime.

cific heat. In superfluid ^{3}He-A, with a specific heat varying as T^3, the current record is about 140 μK, while in superfluid ^{3}He-B, with specific heat decreasing faster as $\exp(-\Delta/k_BT)$, the record is just under 100 μK (Carney *et al.*, 1989). Both of these records were established at the University of Lancaster in the United Kingdom, and were measured via the quasiparticle damping of a vibrating wire immersed in ^{3}He-B. Much earlier measurements in Lancaster using a platinum NMR thermometer showed that ^{3}He-B could be cooled to about 110 μK.

Solid ^{3}He, with its much lower heat capacity for T less than 1 mK, has been cooled to about 40 μK (Takano *et al.*, 1985; Mamiya *et al.*, 1990).

Moving to metals considerably reduces the specific heat, and consequently gives the lowest achievable temperatures. Both of the groups in Lancaster and in Bayreuth have demonstrated temperatures in copper of 12 μK as measured by a platinum NMR thermometer in good thermal (i.e., electrical) contact with the copper (Gloos *et al.*, 1988).

Finally, we consider the temperature of the nuclear system in a metal. Here we gain the advantage of measuring the temperature of the actual refrigerant and not that of a material cooled by it. At the very lowest temperatures, the time constants for nuclear thermalization with the electronic system in the same block of metal can be made quite long, on the order of hours. Since the internal thermalization time within the nuclear system is much shorter, typically milliseconds, it makes good sense to define separate temperatures for the nuclear and electronic systems, even though they are physically inseparable (cf. the discussion above, regarding negative temperatures).

In copper, the nuclear ordering temperature is 58 nK. Since most of the nuclear entropy is removed in the vicinity of the transition, its cooling power rapidly diminishes, limiting the lowest temperature in copper nuclei to 30 nK, established at the Helsinki University of Technology in Finland (Huiku *et al.*, 1986). (At the same time, it should be noted that the electronic system in the same sample is maintained at about 50 μK, i.e., about a thousand times hotter.)

Silver, with much weaker nuclear interactions than copper, allows a much lower temperature to be reached. Using silver, the Helsinki group established the current nuclear low-temperature record of 560 pK (Hakonen *et al.*, 1991), 9 orders of magnitude below the bottom end of the International Temperature Scale of 1990.

ACKNOWLEDGMENTS

We wish to thank W. J. Gully, G. R. Pickett, and P. J. Hakonen for useful comments on this manuscript. Much of the research by the authors has been supported by the National Science Foundation Low Temperature Program.

GLOSSARY

Antiferromagnetism: A state of magnetic order in which alternate individual spins have opposite orientations, resulting in a reduced total magnetization.

Blackbody Radiation: Electromagnetic radiation coming from a heat source.

Carnot Cycle: A thermodynamically reversible closed cycle consisting of two adiabatic (see entropy) and two isothermal processes.

Compressor: A machine that uses mechanical energy to do work on a gas by reducing its volume.

Convective Flow: A spontaneously circulating flow caused by a gravitational instability associated with a hot, low-density fluid placed below a cold, higher-density fluid.

Counterflow Heat Exchanger: A device that allows heat to flow between two fluids that are moving in opposite directions.

Dewar Flask: Originally, a double-walled glass container with a vacuum between the walls and reflecting surfaces that provides thermal insulation. Currently applies to any well thermally insulated container. This term is often shortened to Dewar.

Dilution Refrigerator: A refrigerator that obtains cooling by diluting ^{3}He into a mixture of ^{3}He and ^{4}He.

Enthalpy: A thermodynamic potential obtained by subtracting the mechanical term ($-pV$) from the internal energy, i.e., $H=U+pV$.

Entropy: The thermodynamic variable that measures the quality of thermal energy. From a statistical-mechanics view, entropy measures the degree of disorder in a system.

Evaporative Cooler: A simple refrigerator that cools by evaporating a liquid.

Expansion Engine: The reverse of a compressor. Here, a high-pressure gas does work on a moving piston, thereby removing thermal energy.

Ferromagnetism: A state of magnetic order in which the individual spins have the same direction, resulting in a high total magnetization.

Heat Switch: A device that can interrupt heat flow. These are often made from superconducting metals, where the thermal conductivity is high in the normal state and low in the superconducting state. Other heat switches have been made from movable mechanical contacts and gas-filled containers.

Joule–Thomson Coefficient: The rate of cooling or heating caused by adiabatic (Joule–Thomson) expansion of a gas.

Joule–Thomson Expansion: An adiabatic expansion of gas through a restriction such as a small tube, a porous plug, or an adjustable valve (Joule–Thomson valve). This is a constant-enthalpy process.

Joule–Thomson Inversion Point: A temperature point in a phase diagram where the Joule–Thomson coefficient changes sign. These points depend on pressure and volume, and thus form a line in a phase diagram.

Latent Heat: The heat of vaporization or melting. This corresponds to the change in enthalpy in going from one phase to another.

Magnetic Ordering: An increase in the organization of magnetic spins as temperature is lowered. The magnetic order can be either ferromagnetic, ferrimagnetic, or antiferromagnetic and is accompanied by a reduction in the entropy.

Magnetic Refrigerator: A refrigerator that uses a magnetic material as the working substance and a magnetic field to provide the external work.

Nuclear Demagnetization: A refrigeration technique that uses the magnetic moments of nuclei at the magnetic material. (See Magnetic Refrigerator.)

Paramagnetic Material: A material that has no magnetic order.

Phonons: Lattice vibrations in a solid. Formally, phonons form a basis set to describe lattice deformations and are thus pseudoparticles.

Recuperative Refrigerator: A refrigerator that regains cooling by the use of a counterflow heat exchanger to precool returning gas.

Regenerator: Formally, an ideal heat source and sink. In heat engines, a regenerator temporarily stores heat during part of a cycle. An efficient regenerator can maintain a thermal gradient so that heat can enter and leave at a series of different temperatures.

Stirling Engine: A heat engine that follows the Stirling cycle, which is equivalent to a Carnot cycle using a gas plus a regenerator. Often, any heat engine that uses a regenerator is referred to as a Stirling engine.

Superinsulation: A highly effective cryogenic insulation made from multiple layers of aluminized polyester film. It operates in a vacuum. Frequently, the layers are spaced by a web of plastic fiber. This insulation works as an effective radiation shield because of the layers of reflective aluminum. The polyester and other plastics are poor thermal conductors. Also, the multiple layers trap any residual gas atoms or molecules.

Working Substance: The part of a heat engine that undergoes mechanical (pressure, volume, magnetization, etc.) and thermal (temperature, entropy) changes during a cycle. Typically, this is a gas, liquid, or magnetic material.

Zero-Point Motion: The minimum motion of atoms and molecules about their equilibrium position that is required by the wave nature of these particles. The energy associated with this motion cannot be removed even at $T=0$, and thus is not part of the thermal energy.

List of Works Cited

Andres, K. (1978), *Cryogenics* **18**, 473–477.

Buchal, Ch., Pobell, F., Mueller, R. M., Kubota, M., Owers-Bradley, J. R. (1983), *Phys. Rev. Lett.* **50**, 64–67.

Carney, J. P., Guenault, A. M., Pickett, G. R., Spencer, G. F. (1989) *Phys. Rev. Lett.* **62**, 3042–3045.

Chui, T. C. P., Lipa, J. A. (1984) in: U. Eckern, A. Schmid, W. Weber, H. Wühl (Eds.) *Proceedings of 17th International Conference on Low Temperature Physics* (LT17) (North-Holland, 1984), p. 931–932.

Ebner, C., Edwards, D. O. (1971), *Phys. Rep.* **2C**, 77–154.

Frossati, G. (1978), *J. Phys. (Paris), Colloq.* **39**, C6-1578–1589.

Gloos, K., Smeibidl, P., Kennedy, C., Singsaas, A., Sekowski, P., Mueller, R. M., Pobell, F. (1988), *J. Low Temp. Phys.* **73**, 101–136.

Greywall, D. S. (1986), *Phys. Rev. B* **33**, 7520–7538.

Hakonen, P. J., S. Yin, and K. K. Nummila (1991), *Europhys. Lett.* **15**, 677–682.

Huiku, M. T., Jyrkkio, T. A., Kyynäräinen, J. M., Loponen, M. T., Lounasmaa, O. V., Oja, A. S., (1986), *J. Low Temp. Phys.* **62**, 433–487.

Kittel, C., Kroemer, H. (1980), *Thermal Physics*, 2nd ed., New York: Freeman.

Mamiya, T., *et al.* (1990), *Physica B* **165 & 166**, 837–838.

Marek, D. (1987) in *Proceedings of the 18th International Conference on Low Temperature Physics* (LT18), *Jpn. J. Appl. Phys.* **26**, Suppl. 3, 1683–1684.

Mather, J. C., *et al.* (1990), *Astrophys. J. Lett.* **354**, L37–L40.

Mendelssohn, K. (1977), *The Quest for Absolute Zero*, 2nd ed. New York: Wiley.

Owers-Bradley, J. R., Chocholacs, H., Mueller,R. M., Buchal, Ch., Kubota, M., Pobell, F. (1983), *Phys. Rev. Lett.* **51**, 2120–2123.

Powell, L. R., Childs, G. E. (1972), in: B. H. Billigs *et al.* (Eds.), *American Institute of Physics Handbook*, New York: McGraw-Hill, p. **4**-142.

Preston-Thomas, H. (1990), "The International Temperature Scale of 1990," *Metrologia* **27**, 3–10.

Steinberg, V., Ahlers, G. (1983), *J. Low Temp. Phys.* **53**, 255–283.

Takano, Y., *et al.*, (1985) *Phys. Rev. Lett.* **55**, 1490–1493.

Walker, Graham (1983), *Cryocoolers*, New York: Plenum.

White, Guy K. (1979), *Experimental Techniques in Low Temperature Physics*, 3rd ed., New York: Oxford University Press.

Zemansky, M. W., Dittman, R. H. (1981), *Heat and Thermodynamics*, 6th ed., New York: McGraw-Hill.

Further Reading

Hudson, R. P. (1972), *Principles and Application of Magnetic Cooling*, Amsterdam: North-Holland.

Mendelssohn, K. (1977), *The Quest for Absolute Zero*, 2nd ed., New York: Wiley. An interesting history of the efforts to reach ever lower temperatures and to understand physics that is revealed at low temperatures.

Richardson, Robert C., Smith, Eric (1988), *Experimental Techniques in Condensed Matter Physics at Low Temperatures*, Reading, MA: Addison-Wesley.

Temperature, Its Measurement and Control in Science and Industry, Pittsburgh: Instrument Society of America. An episodic overview of the state of the art in articles focusing on specific issues by the workers themselves. Volume 5 was published in 1982.

Tilley, D. R., Tilley, J. (1990), *Superfluidity and Superconductivity*, 3rd ed., Bristol: Adam Hilger.

Wilkes, J. and Betts, D. D. (1987), *An Introduction to Liquid Helium* 2nd. ed., Oxford: Clarendon.

CRYSTAL GROWTH

Erich Schönherr, *Max-Planck-Institut für Festkörperforschung, Stuttgart, Germany*

Introduction 336
1. **Physical and Technical Principles** 337
1.1 Driving Force, Undercooling, and Supersaturation 337
1.2 Formation of an Initial Crystal 337
1.3 Crystal Growth 338
1.3.1 Mass Transfer of Crystal Species and their Incorporation 338
1.3.2 Mass Transport through Chemical Reactions 340
1.3.3 Heat Transfer into the Crystal 340
1.3.4 Growth Rates from Theoretical Models 341
1.4 Principles of Crystal-Growth Procedures 341
1.4.1 Crystal Growth in Monocomponent Systems 341
1.4.2 Crystal Growth in Multicomponent Systems 342
1.5 Selection of Crystal-Growth Method 342
2. **Basic Components of Crystal-Growth Equipment** 342
2.1 Containers 342
2.2 Heating Apparatus 342
2.2.1 Resistance Heater 342
2.2.2 Induction Heater 343
2.2.3 Other Heat Sources 344
2.3 Temperature Control 344
2.4 Growth-Rate Control 344
2.5 Environment 345
3. **Techniques of Crystal Growth** 345
3.1 Techniques of Crystal Growth from the Vapor Phase 345
3.1.1 Pizzarello Method 345
3.1.2 Temperature Gradient Technique 347
3.1.3 Temperature Difference Technique 347
3.2 Techniques for Crystal Growth from the Liquid Phase 347
3.2.1 Czochralski Technique 347
3.2.2 Bridgman Technique 348
3.2.3 Temperature Gradient Technique 349
3.2.4 Liquid-Zone Techniques 349
3.2.4.1 Liquid-Zone Techniques with Crucible 349
3.2.4.2 Floating-Zone Methods 349
3.2.4.3 Verneuil Technique 350
3.3 Techniques of Crystal Growth from a Solution 350
3.3.1 Solvent Evaporation 350
3.3.2 Temperature Difference Techniques 350
3.3.2.1 Normal Solution Conditions 350
3.3.2.2 Hydrothermal Solution Conditions 351
3.3.3 Techniques with Gel Media 351
3.3.4 Slow-Cooling Techniques 352
3.4 Techniques of Crystal Growth from the Solid Phase (Recrystallization) 353
4. **Techniques of Thin-Film Growth** 353
4.1 Liquid-Phase Epitaxy 353
4.2 Chemical Vapor Epitaxy 354
4.3 Molecular-Beam Epitaxy 354
5. **Characterization of Crystals** 355
5.1 Morphology 355
5.2 Segregation of Impurities 356
5.2.1 Segregation Coefficients 356
5.2.2 Impurity Distribution through Zone Refining 356
5.2.3 Precipitation 357
5.3 Structural Defects 357
5.3.1 Point Defects 357
5.3.2 Line Defects 357
5.3.3 Two-Dimensional Defects 357
6. **Examples of Applications** 358
6.1 Czochralski Techniques 358
6.1.1 Edge-Defined Film-Fed Growth 358

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

6.1.2 Dendritic Silicon Web Growth 358
6.2 Float Zoning of Silicon 358
6.3 High-Pressure Technique of Diamond Growth............. 359
Glossary.................... 360
Works Cited................. 360
Further Reading............ 362
Series and Journals 362

INTRODUCTION

Single crystals consist of a highly ordered atomic assembly. They can be produced in highest purity or with defined impurity (dopant) concentrations to have various important properties of scientific and technological interest. For these reasons crystals are produced in two scales: large crystals (e.g., 1.5-m-long silicon crystals) grown at industrial sites for technical applications and smaller crystals at laboratories for research (e.g., 10^{-3} m for structure analysis and 10^{-2} m for optical investigations).

Extensive study of crystal growth from solutions started in the late 19th century. Historical reviews are given by van Hook (1961), Buckley (1951), and Bohm (1985). The essential crystal-growth techniques which are used at the present time were developed between 1900 and 1960. In the last 30 years these methods have been enormously improved by the realization of control of heat, mass flow, and crystal form through electronic instrumentation. In recent years, experimental investigations of crystal-growth systems in space together with extensive numerical investigations of heat and mass flow have given sufficiently quantitative analyses of some crystal-growth systems that their state of art is now transforming to a state of science. The invention dates of the most important techniques are listed in Table 1.

In this article, the fundamentals of various growth methods of bulk crystals are emphasized with the goal of giving a general insight into the various procedures and their functions. Hence the common properties of different techniques are identified to improve understanding of particular procedures. In

Table 1. Dates of most important inventions of crystal growth methods.

Method	Year	Reference	Comments
Verneuil	1902	Verneuil (1902)	Flame fusion of ruby
Hydrothermal technique	1905	Spezia (1905)	Hydrothermal synthesis of quartz
	1943	Nacken (1950)	Start of industrial quartz production
Solution	1910	Krüger and Finke (1910)	Two tanks technique
Bridgman	1914	Tammann (1914)	Metal crystals in vertical tubes
	1924	Obreimov and Schubnikov (1924)	Metal crystals in sealed ampoules with cone
	1925	Bridgman (1925)	Metal crystals in sealed ampoules with capillary end
	1936	Stockbarger (1936)	LiF in graphite crucibles, controlled environment
Czochralski	1915	Nacken (1915)	Salol crystals from melt through top seeding
	1918	Czochralski (1918)	Pulling of crystalline wires from melt
	1926	Kyropoulos (1926)	Alkali halides from melt through top seeding
Pyrolysis	1922	Koref (1922)	W deposition on hot wire through WCl_6 decomposition
Chemical transport	1923	van Arkel (1923)	W transport through $WCl_6 = W + 3Cl_2$
Zone melting	1952	Pfann (1952)	Horizontal zone refining of Ge
	1953	Keck and Golay (1953)	Float zoning of Si
Pizzarello	1954	Pizzarello (1954)	PbS crystals through sublimation
Diamond synthesis	1955	Bundy, Hall, Strong, Wentorf (1955)	Diamond synthesis from Ni solvent under high pressure

addition, the advantages, disadvantages, and economic aspects of each method are pointed out. The basic physical processes of crystal growth, namely, mass transfer, heat transfer, incorporation of crystal species, and impurity segregation are undoubtedly rather complex and are therefore presented only by very simple quantitative relations. Crystallization is not considered that which is commercially applied to the production of sugar, salt, and fertilizer as reviewed by Mullin (1985). Also, electrocrystallization is ignored which is described, e.g., by Bockris and Razumney (1967).

Section 1 deals with physical and technical principles of crystal growth to the extent that they are needed to comprehend at least qualitatively the crystal-growth procedures, which consist of a controlled change of a mobile phase into a solid. The conduction of the crystal species through the fluid phase to the crystal interface by a suitable temperature field appears as the most important principle of crystal-growth processing in a monocomponent system. In a multicomponent system, change of the composition is an additional way of processing.

In Sec. 2 the basic components which are common to most crystal-growth systems are outlined. The components are essentially heat sources, fluid containers, and particular crystal environments.

The growth methods are described in Sec. 3 and are classified by the kind of mobile phase: crystal growth from the vapor phase in Sec. 3.1, from the melt in Sec. 3.2, from the solution in Sec. 3.3, and from the solid phase in Sec. 3.4. The particular techniques of each category are essentially characterized by technical aspects such as support, attachment, and geometry of fluid and crystal. Additional categories are obtained from physicochemical aspects such as the change of a temperature distribution or composition of the fluid phase (solution).

Since thin films are grown by the same physicochemical principles as those of bulk crystals, layer deposition techniques are briefly discussed in Sec. 4.

Some essential points of crystal characterization such as growth form, impurity segregation, and point, line, and two-dimensional defects are discussed in Sec. 5.

Section 6 closes the article and gives some examples of sophisticated applications of crystal-growth methods.

1. PHYSICAL AND TECHNICAL PRINCIPLES

1.1 Driving Force, Undercooling, and Supersaturation

From the view of thermodynamics, crystal growth consists of the change of a fluid phase into a solid phase which represents the crystal. The transition requires a driving force which is given by the difference in chemical potential $\delta\mu = \mu_f - \mu_c < 0$ between the species in the fluid (f) and crystalline (c) phase. Measurable thermodynamic quantities such as temperature T, vapor pressure p, and concentration C are preferred for the description of practical crystal growth. In particular,

1. the undercooling $T_m - T_c = T_m \delta\mu / H_m$ is used for a melt monosystem where T_m is the melting point and H_m the heat of fusion;
2. the supersaturation $p_f - p_c$ with $\ln(p_f/p_c) = -\delta\mu/kT$ for a vapor monosystem; and
3. the supersaturation $C_f - C_c$ with $\ln(C_f/C_c) = -\delta\mu/kT$ for a solution.

1.2 Formation of an Initial Crystal

The formation of an initial crystal is for all self-seeding techniques an important procedure. The creation of only one nucleus can be achieved by a sufficiently slow undercooling or supersaturation of the fluid in a cone or capillary. When the particular seeding area is idealized by a small circular cavity, the formation of a single seed can be explained by the thermodynamic approaches described by Sigsbee (1969) and Walton (1969). For an arbitrary cylindrical cluster of height h, the Gibbs energy of formation G_i results in

$$G_i = ivG_v + iv(\sigma_{12} + \sigma_{13} - \sigma_{23})/h + 2(\pi v h)^{1/2} \sigma_E\, i^{1/2}, \qquad (1)$$

where i is the number of cluster members; v the solid volume per monomer; G_v (<0) the difference in Gibbs energy between solid and fluid (driving force); σ the surface Gibbs interfacial energy among cluster (1), fluid (2), and wall (3); and σ_E the edge Gibbs energy. The energy G_i assumes a maximum value G_i^* at

$$i^* = -\pi h \sigma_E^2 v^{-1} [G_v + (\sigma_{12} + \sigma_{13} - \sigma_{23})/h]^{-2}.$$

The size i^* is denoted as critical since clusters smaller than i^* are unstable while those greater than i^* form growing nuclei. An increased possibility of nucleation exists for the cavity since the edge free energy σ_E is reduced through the reaction with the wall. The preferred nucleation is schematically illustrated in Fig. 1. Curve (1) represents the derivative of the interfacial energy of G_i as a function of the size i for a cluster on a plane crucible wall. Curve (2) shows the derivative for that cluster in a cavity of matching radius. For the actual cavity where the radius is in accordance with the cluster size i_c^* a nucleus is formed as soon as the driving force reaches vG_{vc}. The formation of other nuclei is unlikely since clusters of improbable sizes larger than i^* are required.

1.3 Crystal Growth

Crystal growth occurs through the transport of crystal species from the fluid phase to the crystal interface followed by incorporation of these species into the crystal. The incorporation is accompanied by chemical reactions at the surface if the fluid species differ from the structure units of the crystal, e.g., in the cases of chemical transport and associative and dissociative sublimation. In the case of a multicomponent system additional chemical reactions can happen in the fluid. When the reactions in the fluid can be ignored the three processes

1. mass transfer,
2. chemical reactions at the interface, and
3. incorporation

are sequentially coupled so that one of these determines the growth rate.

1.3.1 Mass Transfer of Crystal Species and their Incorporation. A quantitative description of the motion of fluid species is very complex since four conservation laws have to be considered—namely, conservation of momentum (Navier-Stokes), chemical species (transport), mass (continuity), and energy (heat flow). General treatments are presented by Bird *et al.* (1960), Hershey (1973), and Peube (1977). In general, mass transport controls the growth rate through volume diffusion for most multi-component systems. In comparison to a purely monocomponent fluid, here diffusive fluxes, caused by thermal (Soret effect) or concentration gradients, are superposed to the convective flux of the species as discussed by Wilcox (1971), Faktor and Garrett (1974), and Rosenberger (1979). A one-dimensional steady-state approximation for the two-component nonviscous fluid leads to the velocity

$$V_{cf}=(D/\delta)\ln[(C_t-C_{cf})/(C_t-C_f)] \qquad (2)$$

of the crystal species in the fluid, where D is the binary diffusion coefficient and C_t the total concentration, assumed to be spatially constant. The concentrations C_{cf} and C_f are those of the crystal species at the interface and at the

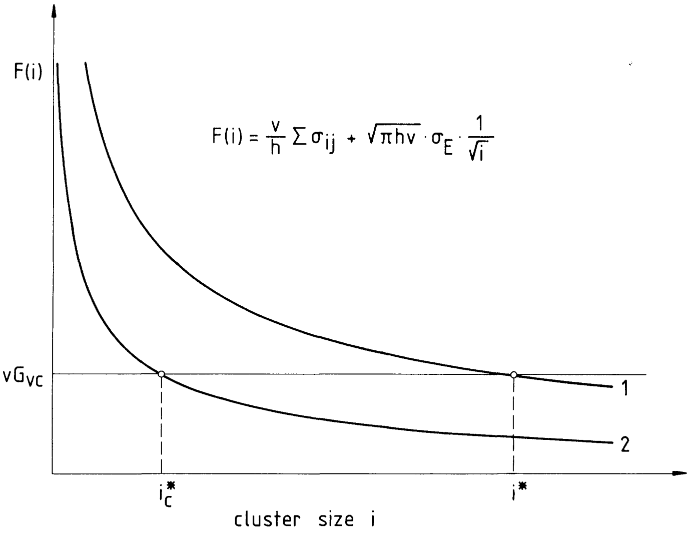

FIG. 1. Formation of an initial crystal. (See text.)

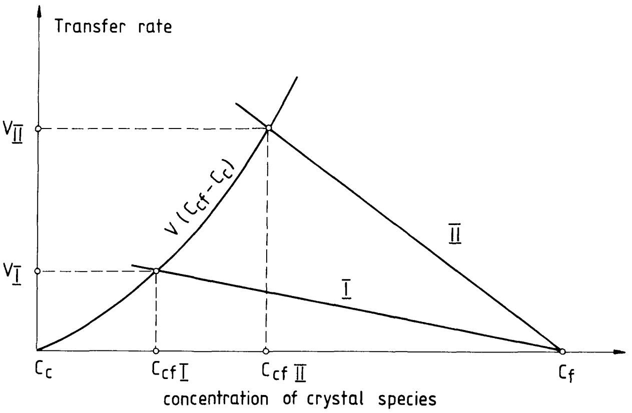

FIG. 2. Stationary crystal growth at the rates V_I and V_{II} for two different transport conditions. The mass transfer rates I and II equal the incorporation rate at $C_{cf\,I}$ and $C_{cf\,II}$, respectively. In both cases the saturation concentration difference $C_f - C_c$ is the same.

effective distance δ (see Sec. 5.2.1) from the interface, respectively. The concentration distribution, and thus the thickness δ, are strongly affected by convections resulting, e.g., from stirring, buoyancy, or tension gradients of wall-free fluid surfaces (Marangoni convection); therefore, the same total driving force $C_f - C_c$, where C_c is the saturation concentration of the crystal, can lead to different growth rates as illustrated in Fig. 2. Curve $V(C_{cf} - C_c)$ represents the surface reactions which are driven by the supersaturation $C_{cf} - C_c$. Curve I decribes the mass transfer for a convectively undisturbed flux. The intersection of the two curves determines the growth rate V_I. By pumping the fluid at a high speed towards the crystal, the effective diffusion zone δ becomes small resulting in curve II and thus in an increased growth rate V_{II}.

The growth is denoted as interfacial kinetics controlled if $C_{cf} - C_c \gg C_f - C_{cf}$ and mass transfer controlled if $C_f - C_{cf} \gg C_{cf} - C_c$. The growth rate V_c and fluid mass velocity V_{cf}, normal to the interface, are linked by the conservation relation

$$\rho_c V_c = \rho_{cf} V_{cf}\,, \qquad (3)$$

where ρ_c and ρ_{cf} are the densities of crystal and fluid, respectively, as shown by Wilcox (1971).

The interfacial processes are mainly affected by the surface structure, i.e., by the number of available bonds per approaching crystal-growth unit. The actual structure is between atomically rough and atomically smooth as schematically shown in Fig. 3. In the first case nearly every surface site has sufficient free bonds to bind an incoming growth unit, so that the growth occurs continuously normal to that interface. Rough interfaces achieve the highest incorporation rates and may appear close to the melting point or at high supersaturation as described by Jackson (1975) and Bennema and Gilmer (1973).

In most cases smooth plane as-grown surfaces are observed which run parallel to (low-

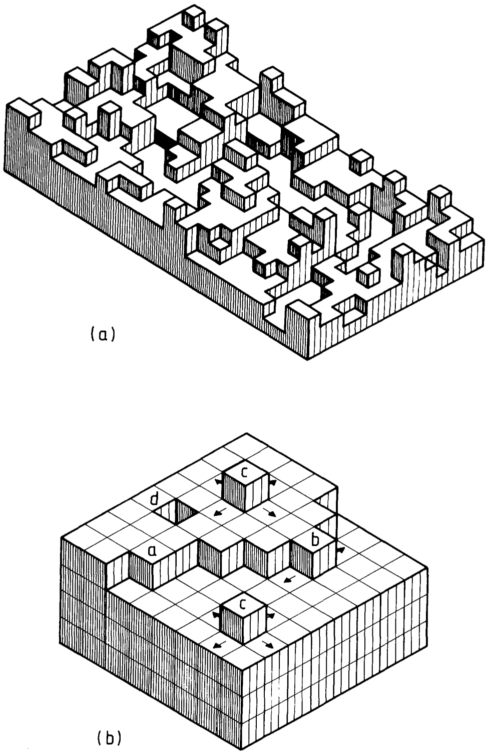

FIG. 3. Crystal surface models. (a) Rough surface. (b) Kossel-Stranski surface (Kossel, 1927; Stranski, 1928) characterized by some particularities: An atom is in a kink position (*Halbkristallage*) at *a*, at a ledge at *b*, and adsorbed at *c*. A hole (*Lochkeim*) is formed at *d*.

index) crystallographic faces (singular faces) of low surface energy. They grow layer by layer through the capture of growth units in kinks along monomolecular steps (ledges). The energy of kink formation is very low, about a tenth of the heat of evaporation, so that the growth rate is mainly determined by the creation of steps. Sources of ledges are screw dislocations (BCF theory) as discussed by Burton, Cabrera, and Frank (1951), Strickland-Constable (1968), Mutaftschiev (1980), and Chernov (1984) and two-dimensional seeds as described by Voronkov (1983). Screw dislocations continuously provide the surface with self-perpetuating spiral steps. In contrast to the free energy of a two-dimensional seed, the free energy of the first ledge winding around a screw dislocation is so low that the growth along the spiral step occurs at nearly zero driving force without observable critical behavior. In cases of growth from vapors and solutions, the kinks are mainly fed from adsorbed species so that the ledge spreading is strongly affected by the average surface diffusion length and desorption probability of the adsorbed species. At low supersaturations the distance between steps is much greater than the diffusion path and the growth rate becomes proportional to the square of the driving force. At high supersaturations the step density of the spirals becomes high and the growth rate linearly proportional to the driving force. When, in addition, heterogeneous reactions are involved the surface processes become very complex as described by Jones and Shaw (1975).

Surface nucleation models, which explain the growth of perfect crystals, lead to exponential, linear, or more complex relations between growth rate and driving force as, e.g., discussed by Brice (1973).

1.3.2 Mass Transport through Chemical Reactions. Chemical vapor transport is extensively discussed by Schäfer (1964) and Faktor and Garrett (1974). Transport of a solid in the vapor phase requires the presence of a volatile product of the solid. In the simplest case the volatile product is formed by the reaction of the solid with a gas (transporting agent), e.g., $M(s)+X_2(g)=MX_2(g)$, where M is a divalent metal, X a halogen, and MX_2 the volatile reaction product. When thermal diffusion and any kind of sublimation is excluded, the one-dimensional stationary flux J_i of the vapor species i is expressed by

$$J_i=\frac{D}{RTd}\,\frac{p^*(d)p_i(0)-p_i(d)p^*(0)}{p^*(d)-p^*(0)}\times\ln\frac{p^*(d)}{p^*(0)}, \tag{4}$$

where T is the average temperature between crystal (at the location 0) and source (at d). The average diffusion coefficient is given by

$$D=D_0(T/T_0)^n(p_0/p) \tag{5}$$

with $T_0=273$ K, $p_0=1$ Pa, $p=\Sigma p_i$ (total pressure), and $1.5 \leqslant n \leqslant 1.8$. The standard diffusion coefficient D_0 varies between 1.2×10^{-4} and 0.01×10^{-4} m^2/s. The mass velocity of the total vapor is

$$V_t=(D/d)\ln[p^*(d)/p^*(0)].$$

The stoichiometric average pressure, $p^*=\Sigma a_j p_j$, follows from the condition of insolubility ($J_j=0$) of the transporting agent j in the solid. The coefficients a_j are the stoichiometric ones of the gaseous transport agent compounds j. Transport of species i occurs from hot to cold places for

$$p^*(d)p_i(0)>p_i(d)p^*(0)$$

and from cold to hot for

$$p^*(d)p_i(0)<p_i(d)p^*(0).$$

In the example $M(s)+X_2(g)=MX_2(g)$ one has $D=D_1=D_2$, $p_1=p_{X2}$, $p_2=p_{MX2}$, and $p^*=p_{X2}$.

1.3.3 Heat Transfer into the Crystal. The balance of heat at the crystal interface leads to a quantitative description of crystal growth. In the case of growth in one dimension x the crystal-growth rate becomes

$$V_c=[k_c(dT_c/dx)-k_f(dT_f/dx)]/(\rho_c H_c), \tag{6}$$

where k_c and k_f are the heat conductivities, dT_c/dx and dT_f/dx the temperature gradients of crystal (c) and fluid (f), ρ_c the crystal density, and H_c the heat of phase transition. When A is the cross section of crystal and fluid, $Ak_f(dT_f/dx)$ is the heat transfer rate from the fluid into the crystal, and $Ak_c(dT_c/dx)$ the heat removal rate from the crystal. Most of the crystal-growth procedures consist of controlling these rates. A complete heat-flow description of a crystal fluid system leads to the Stefan problem as described, e.g., by Parker (1970) and Crank (1988).

1.3.4 Growth Rates from Theoretical Models. Maximum crystal-growth rates for a monocomponent system can be estimated from basic physical relations.

1. For the melt, from the Wilson-Frenkel relation

$$V = a\alpha \exp\left(-\frac{H_0}{RT_c}\right)\exp\left(-\frac{H_m}{RT_m}\right) \times\left[1-\exp\left(-\frac{H_m}{R}\frac{T_m-T_c}{T_m-T_c}\right)\right], \tag{7}$$

where a is the interatomic distance (10^{-10} m), α the lattice vibration frequency (10^{13} Hz), H_0 the activation enthalpy of self-diffusion in the melt as discussed by Jackson (1975), H_m the heat of fusion, R the gas constant, T_m the melting point temperature, and T_c the temperature of the crystal.

2. For the vapor, from the Hertz-Knudsen relation

$$V = \frac{M_f}{\rho_c(2\pi M_f RT_c)^{1/2}}(p_f-p_c), \tag{8}$$

if every atom is incorporated which hits the (rough) interface. In this expression, M_f is the molar mass of the vapor, ρ_c the density of the crystal, p_f the vapor pressure in front of the crystal interface, and p_c the saturation pressure of the crystal.

3. For the vapor in a cylindrical ampoule with interfacial equilibrium, from the Hagen-Poiseuille relation

$$V = \frac{M_f a_d^2}{16\rho_c \eta LRT}(p_f^2-p_c^2), \tag{9}$$

where η is the dynamic viscosity, a_d the ampoule diameter, p_f the vapor pressure of the source, L the distance between crystal and source, and T the average temperature of crystal and source.

4. For the vapor with an additional inert component A, from

$$V = \frac{M_{fc}D}{\rho_c LRT}\ln\left(\frac{p_t-p_c}{p_t-p_{fc}}\right), \tag{10}$$

where M_{fc} is molar mass of the crystal species of the vapor phase, D the binary diffusion coefficient defined in Eq. (5), $p_t=p+p_A$ the constant total pressure, p_{fc} the partial vapor pressure of the crystal vapor species at the distance L in front of the crystal, and T an average temperature.

Growth rates described by the relations (7)–(10) are shown in Fig. 4.

1.4 Principles of Crystal-Growth Procedures

1.4.1 Crystal Growth in Monocomponent Systems. In monocomponent systems, crystal growth is generally forced by adjusting a temperature distribution across the interface in such a way that heat is continuously flowing from the hotter fluid into the colder crystal. The growth rate is essentially determined by the rate of heat flow according to Eq. (6). Continuous growth in one direction toward the fluid is maintained by two principal techniques:

1. A temporary constant temperature distribution is slowly shifted toward the fluid by moving the crystal or in a few cases by the moving of the heater.
2. The container and heater are kept in a fixed

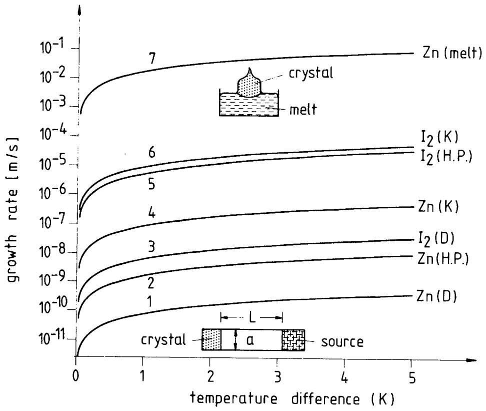

FIG. 4. Various growth rates from theoretical models for different monocomponent systems as a function of undercooling. *Curve 7:* Growth of Zn from the melt according to Eq. (7), H_0 = 21.31 kJ/Mol (from Ramachandrarao, 1978). *Curves 1–6:* Crystal growth rates of Zn and I_2 from vapor phase as a function of temperature difference between source and crystal. The parameters are T_c = 410 °C, $D_0=10^{-5}$ m²/s for Zn, and T_c=40 °C, $D_0=5\times10^{-6}$ m²/s for I_2. The common parameters are L=0.15 m, $a=2\times10^{-2}$ m, η = 2×10^{-4} P, and P_A = 133 Pa. Curves 1 and 3 represent Zn and I_2 crystal growth by diffusion through argon according to Eq. (10). Curves 2 and 5 represent crystal growth of Zn and I_2 by a pure viscous flow according to Eq. (9). Curves 4 and 6 represent Zn and I_2 incorporation according to Eq. (8).

position while the temperature of the heater is slowly lowered.

In a few cases crystals are isothermally grown from supercooled melts by extremely slow cooling rates as described by Scheffen-Lauenroth *et al.* (1981).

The shape or the temperature distribution of the fluid near the interface affects the stability of the interface, particularly if the fluid contains impurities or dopants. The possibility of parasitic nucleation is reduced by a strong normal temperature gradient through superheating the fluid in front of the interface.

1.4.2 Crystal Growth in Multicomponent Systems. In principle, the methods for crystal growth from a monocomponent system (melt) can be transferred to a multicomponent system. In that case the drive rates have to be reduced by a factor between 10^4 and 10^5 since the growth rate is essentially determined by the mass transfer according to Eq. (2) or by surface reaction kinetics.

In addition to the selection of the temperature, the choice of fluid composition is possible. Hence crystal growth can be carried out under isothermal conditions. Four principal isothermal procedures may be employed:

1. slow cooling (or heating) of the fluid-crystal system;
2. slow enrichment of the fluid with crystal species, e.g., by evaporation of the solvent or by injection of an additional fluid;
3. continuous supply of a supersaturated nutrient into the isothermal environment of the crystal (temperature difference technique); and
4. decomposition or heterogeneous chemical reaction in an isothermal environment.

1.5 Selection of Crystal-Growth Method

The choice of crystal-growth method depends on

1. external aspects such as desired size, shape, purity, doping, and perfection of the crystal;
2. economics;
3. available equipment; and
4. thermodynamic behavior.

The thermodynamic properties such as structural change below the melting point, peritectic reaction, decomposition, sublimation, reaction of the melt with the environment, and solubility in a suitable solvent exclude or indicate the applicability of a particular fluid and therefore determine the growth method.

In particular, the structural change limits heating to below the transition temperature so that solvents, reactive vapors, or eventually sublimation have to be employed. If an incongruent melting compound A_mB_n, e.g., forms solid A and a liquid at the peritectic temperature T_p, then crystals of A_mB_n can be grown below T_p from a solution which is obtained from a relevant mixture of B and A_mB_n. Materials which decompose into one or more (e.g., into two in the case of CdS) volatile components can be grown from the melt in closed containers or by encapsulation in a suitable liquid. Crystal growth through sublimation is possible at reasonable rates for vapor pressures higher than 100 Pa.

2. BASIC COMPONENTS OF CRYSTAL-GROWTH EQUIPMENT

2.1 Containers

In most crystal-growth methods containers are used to hold the fluid. Commercially available containers are often made of fused silica, sintered alumina, graphite, tungsten carbide, platinum, and iridium. It is necessary that the container material does not contaminate the fluid and does not react with the ambient atmosphere. In addition, when the crystal reacts with the crucible wall, the contact between crystal and wall must be prevented. For many substances these conditions of compatibility are hard to fulfill.

The principal designs of containers can be classified as regular crucibles, boats, tubular crucibles, and ampoules as shown schematically in Figs. 5(a), 5(b), 5(c), and 5(d), respectively.

2.2 Heating Apparatus

2.2.1 Resistance Heater. Resistance furnaces are employed for the heating of fluids; therefore, most heaters consist of tubes which closely surround the crucibles or ampoules. Ceramic tubes wound with NiCr or AlFe alloy wires and SiC resistor tubes are commonly

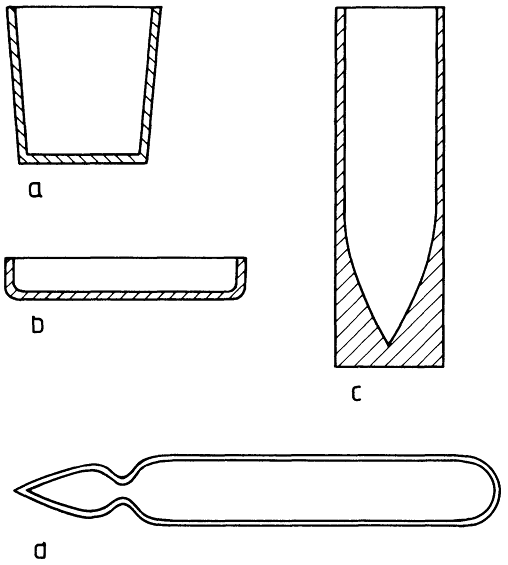

FIG. 5. Principal container designs for crystal growth. (a) Regular crucible, (b) boat, (c) tubular crucible, and (d) ampoule.

used in air. They are enclosed in an insulating housing to reduce the heat loss. Lengthwise-sliced cylinders of graphite or high-melting metals surrounded with metallic heat shields are preferred in oxygen-free equipment. Distinct temperature distributions are obtained through separate heating elements along the furnace axis. Chamber furnaces are mainly employed for growing crystals from high-temperature solutions through slow cooling. The available temperature of a resistance furnace depends on the corrosiveness of the ambient atmosphere. Temperatures up to 1800 °C can be reached in air with rhodium or $MoSi_2$ heaters. In vacuum and noble gases, temperatures up to 3000 °C are obtainable with tungsten, while in nitrogen, noble gases, and vacuum, graphite can be heated up to 2600 °C.

Advantages: Except those with noble-metal wiring, resistance furnaces have a low initial cost. Their temperature can be easily controlled. *Disadvantages:* In general, a desirable high temperature gradient cannot be achieved.

2.2.2 Induction Heater. An induction heater consists of a water-cooled copper coil which is supplied with an alternating voltage at a frequency between 100 kHz and 13.6 MHz. A body within the coil is heated by an induced current if the electrical conductivity of the body is sufficiently high and if the distance between coil and body is sufficiently small. Thus conductive fluids can be directly heated without using a container. Metal and graphite crucibles as well as any appropriately shaped conducting part (susceptor) can be inductively heated, while the surrounding wall can be kept cool with water [Fig. 6(a)]. Moreover, direct heating of a corrosive fluid is possible in a cold-wall boat [Fig. 6(b)], obtained by indenting a copper tube, or in a cold-wall crucible [Fig. 6(c)] described, e.g., by Aleksandrov *et al.* (1978). In both cases the melt is held by a shell which is formed between melt and crucible.

Advantages: Inductive heaters offer the possibility of directly heating the fluid and susceptors at higher temperatures than those which can be reached with resistance heaters. *Disadvantages:* High-frequency generators have a high initial cost, high consumption of cooling water, and difficulty in making an efficient power transfer.

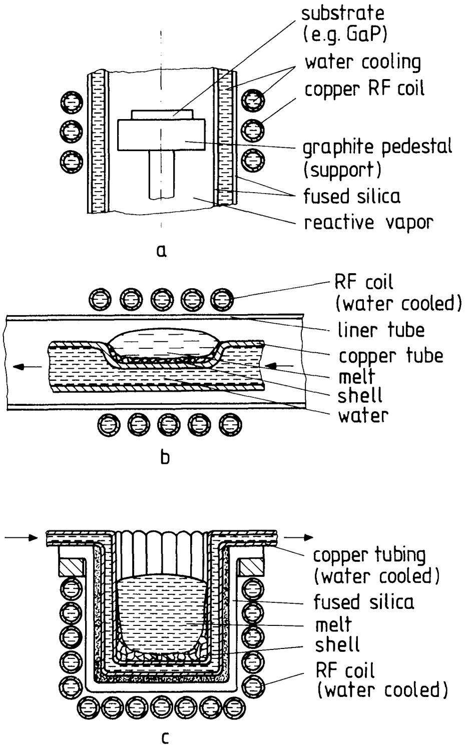

FIG. 6. Cold-wall heating designs. (a) Substrate heating enclosed in a cold-wall tube. (b) Cold-wall boat. (c) Cold-wall crucible.

2.2.3 Other Heat Sources.

The temperature of flames is difficult to control. They are used for the Verneuil method only.

Very high temperatures at an optimal efficiency can be obtained by electron bombardment; its application is preferred for floating-zone crystallization of conductive materials with a high melting point and low vapor pressure, as reviewed by Shah and Wills (1975). In the simplest assembly the fluid acts as an anode which is surrounded by a negative tungsten glowing filament. The method requires a vapor pressure less than 10^{-2} Pa of the environment and material. Insulating materials can be bombarded using a water- cooled slitted-ring anode.

Pure optical radiation is suitable for formation of molten zones in nonmetallic rods. The highest temperatures can be obtained by irradiation with a CO_2 laser beam. The light from the filament of a halogen lamp or arc of a xenon-tungsten lamp can be transferred by imaging with elliptical reflectors.

Heat pipes consist of a closed double-wall cylinder. The space between the walls is partly filled with a liquid. Heat is steadily transferred by boiling the fluid at some points and condensing it at others by placing the tube in a resistance furnace. Heat pipes provide an extremely spatially constant temperature distribution. Detailed descriptions are given by Chisholm (1971) and Winter and Barsch (1971).

Fluid baths are preferred for heating or cooling of tanks used for crystal growth from low-temperature solutions. The baths are heated with immersed resistance elements.

Other heating methods such as gas discharge arcs and plasmas are of minor importance. They have been reviewed by Shah and Wills (1975).

2.3 Temperature Control

To grow high-quality crystals and films it is very important to provide fluid and crystal with a temperature distribution according to a prescribed function of time. This is achieved with commercially available closed-loop control systems (personal computer). In general, the input function of the control system is a voltage which results from the difference between prescribed (set point) and measured (actual) temperatures. Temperature sensors are usually thermocouples. Platinum resistors and thermistors are frequently used at low temperatures ($<200\,^\circ C$) and pyrometers at very high temperatures ($>1500\,^\circ C$). The output function of the controller is a voltage which drives the power of the heat source, e.g., the current of the resistance element through a thyristor or the power of a high-frequency generator through the gate of the oscillation tube.

2.4 Growth-Rate Control

Growth-rate control is required for optimal crystal perfection and dopant distribution. With respect to process parameters (input functions), which can be varied during processing, the growth systems can be classified into two groups: (1) systems which are driven with one input function and (2) those with several input functions.

1. When only one input function is needed, the control of growth rate is reduced to control of that function (open-loop system). For optimal operation the functional relation between input and crystal shape have to be known. In all cases the input rates have to stay below a critical amount to prevent degeneration of the crystal interface. Bridgman, liquid-zone, and Pizzarello techniques are examples where the input is the displacement velocity between the temperature distribution and the fluid. These systems are approximately linear when the crystal grows at a constant diameter and can be described by the governing equation

   $$\beta d(t) = a(t) + \tau \frac{da(t)}{dt}, \tag{11}$$

   where $\beta \approx 1$ is the transition factor describing the temperature change through the crystal, $d(t)$ the input function (ampoule displacement), $a(t)$ the response (crystal length), and τ the time constant. The relaxation time is a function of kinetics, mass, and heat transfer. Figure 7 shows solutions of Eq. (11) for a ramp input.

   Other examples of linear behavior are solvent evaporation, where the input function is the evaporation rate, and solution cooling (with a linear relation between saturation and temperature) where the input is the cooling rate. Both systems are approximately linear when the rate of crys-

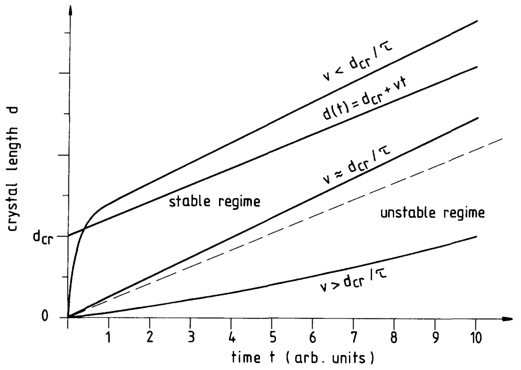

FIG. 7. Crystal lengths in response to different drive rates v as functions of time, according to Eq. (11), using the ramp function $d(t)=d_{cr}+vt$ as an input function. The drive length d_{cr} determines the critical undercooling of seed formation. Crystal growth degenerates for too high drive rates $v>d_{cr}/\tau$.

tallized mass is considered as the response. The growth rates are therefore monotonic and not constant for constant input rates.

2. For complex systems with more than one input possibility, closed-loop systems (or the skill of the crystal grower) have to be applied, where the shape of the crystal is measured and compared with a prescribed one. The Czochralski method is an example where the change of the melt temperature and pulling rate are two independent input functions. Two types of sensors, namely, a video camera for measuring the crystal diameter [described by Draper (1975)] and a balance for weighing the crucible [described by Bardsley *et al.* (1977) and Mecseki *et al.* (1987)], are preferably employed. The signals of the sensors are used to control the heating rate at a prescribed pulling rate.

2.5 Environment

The environment around the crystal and fluid may have different purposes:

1. to prevent contamination of the crystal, e.g., by oxidation,
2. to protect the laboratory from poisoning,
3. to reduce evaporation of the fluid, and
4. to stabilize decomposing compounds.

In most cases the crystal-fluid assembly is therefore enclosed in a modified housing which can be supplied with vacuum, static, or flowing high-purity inert gas. From the kind of heating design two types of housings can be classified: hot wall and cold wall. The hot-wall housings are heated by a resistance furnace from the outside. They are generally simple in design, i.e., consist, e.g., of sealed silica ampoules. Cold-wall housings are either heated inductively from the outside or enclose (in a more or less sophisticated manner) the heating elements.

The purest available environment is a vacuum. It can be applied to techniques of crystal growth from the melt if the vapor pressure of the melt is sufficiently low.

Slight evaporation of the fluid from open crucibles is reduced through heavy ambient inert gases (N_2, Ar, and Kr) at pressures of about 10^4 Pa. Strong evaporation of fluids at pressures $> 10^4$ Pa can be suppressed by covering the fluid with an inert liquid and applying a sufficiently high ambient gas pressure. The liquid must be of a lower density than that of the fluid and must also adhere to the crucible and crystal.

The decomposition of a fluid A_mB_n into a liquid phase and vapor component B [e.g., $2Li_3N(l)=6Li(l)+N_2(g)$] is prevented through an excess vapor of B (e.g., N_2 atmosphere).

3. TECHNIQUES OF CRYSTAL GROWTH

3.1 Techniques of Crystal Growth from the Vapor Phase

3.1.1 Pizzarello Method. A fused silica ampoule is supported with source material either together with a transport agent or a high vacuum ($<10^{-2}$ Pa), then sealed off and hung in a vertical two-zone furnace. The growth method, reviewed by Schönherr (1980), is illustrated in Fig. 8. The upper end of the ampoule is distinctly shaped with a pointed tip and a neck [Fig. 9(a)] or capillary [Fig. 9(b)] for seeding. Starting with the upper seeding end in the middle of the upper heating zone (Fig. 8), the ampoule is slowly ($\approx 2\times 10^{-8}$ m/s) pulled upwards so that the upper end becomes cooler than the source. After a crystal is formed its growth velocity matches the drive rate after an initial high speed as shown in Fig. 7.

The mounting of a seed is possible in similar designs as illustrated in Fig. 9(c).

According to Eq. (10), small amounts of an impurity vapor reduce the mass transfer rate

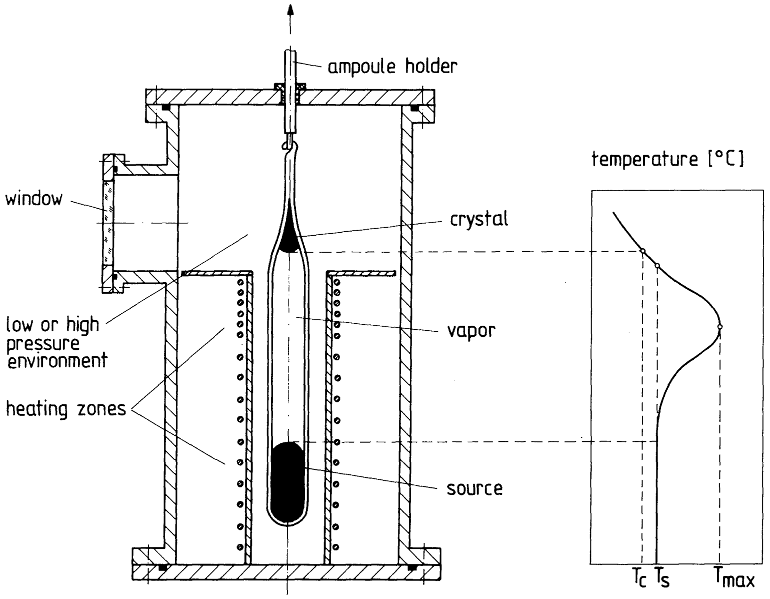
ampoule holder
window
crystal
low or high pressure environment
vapor
heating zones
source
temperature [°C]
T_c T_s T_{max}

FIG. 8. Pizzarello method. (See text.)

dramatically. Foreign vapors are removed by using "almost-closed" crucibles with a small hole ($\approx 0.1 \times 10^{-3}$ m) [Fig. 10(a)], or with self-sealing ends [Fig. 10(b) and 10(c)] which are enclosed in an evacuated tube as described by Faktor and Garrett (1974). In particular, partial pressures of materials which dissociate into two vapor components [e.g., $2CdS(s) = 2Cd(g) + S_2(g)$] can be adjusted to the stoichiometric ratio through almost closed ampoules.

Advantages: The Pizzarello technique is achieved with low-cost equipment. The easy processing only requires a constant temperature distribution and a constant drive rate. High crystalline perfection and high chemical purity usually result. *Disadvantage:* The crystals grow at low rates of about 10^{-8} m/s.

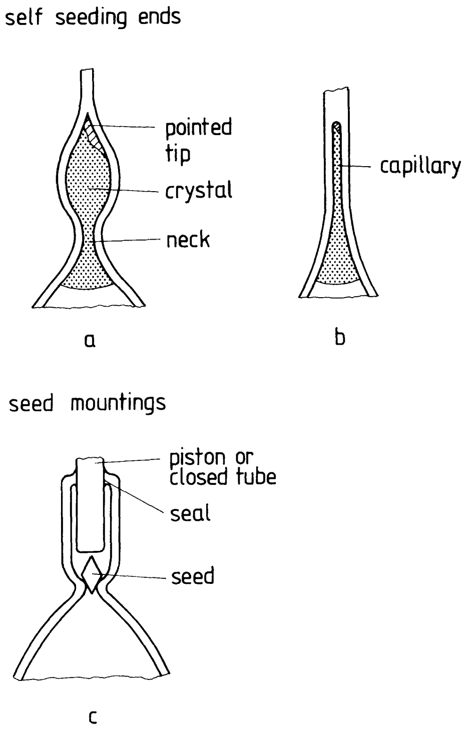
self seeding ends
pointed tip
crystal
neck
capillary
a
b
seed mountings
piston or closed tube
seal
seed
c

FIG. 9. Designs of ampoule ends for seeding. (a) Seeding tip with constriction and (b) capillary for seed selection. (c) Seed support in a closed ampoule.

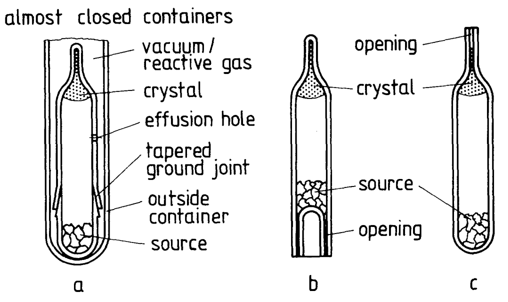
almost closed containers
vacuum / reactive gas
crystal
effusion hole
tapered ground joint
outside container
source
a
opening
crystal
source
opening
b
c

FIG. 10. Designs of almost closed containers. (a) Two-section ampoule with effusion hole. (b) Ampoule with self-sealing source end. (c) Ampoule with self-sealing seeding end.

3.1.2 Temperature Gradient Technique. This technique is used for growing large crystals from the vapor phase at low temperatures ($\approx 120\,^{\circ}C$), e.g., HgI_2 crystals as described by Kaldis *et al.* (1987). The principle of the method is shown in Fig. 11. A seed crystal is mounted on a coolable pedestal at temperature T_c in the center of a cylindrical, evacuated, sealed glass container. The source material initially deposited at the upper part of the container wall is held at a temperature T_s higher than T_c of the pedestal. A heating ring at the bottom at a temperature T_r higher than T_s of the source prevents parasitic nucleation on the container bottom and pedestal wall. Crystal growth is forced by lowering the bottom temperature T_r while the pedestal T_c and source temperature T_s are kept constant.

Advantages: The low-cost equipment offers the possibility of growing crystals with high crystalline perfection and high chemical purity. A seed crystal may be easily supported. Although low linear growth rates of 10^{-8} m/s result, a large total crystalline mass is condensed because of radial growth. *Disadvantage:* The processing is rather complicated since a variation of source and crystal temperature are required to control the crystal growth.

3.1.3 Temperature Difference Technique. For synthesis of new compounds through chemical vapor transport, an ampoule is filled with source material together with an appropriate reaction agent and sealed at sufficiently low pressure. Then the nutrient is distributed at one end of the capsule. After that procedure the ampoule is placed in a horizontal two-zone furnace which heats the ampoule ends to different constant temperatures. The furnace is tilted slightly to increase the fluid velocity by gravitational convection. Numerous small crystals are formed opposite the source since no single distinctly supersaturated small area can be achieved.

Advantage: The technique needs very low-cost equipment only. *Disadvantage:* A single crystal cannot be achieved.

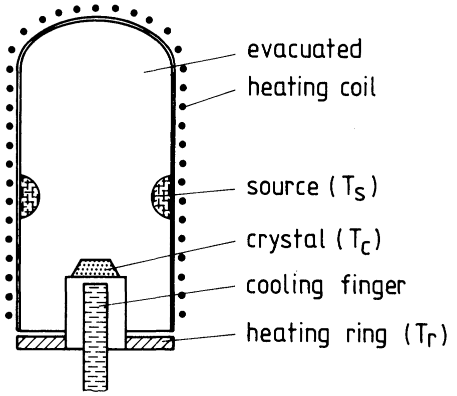

FIG. 11. Design of a closed vessel for crystal growth from vapor phase. (See text.)

3.2 Techniques for Crystal Growth from the Liquid Phase

3.2.1 Czochralski Technique. The modern Czochralski technique (Fig. 12) results from the ideas of Nacken (1915), Czochralski (1918), and Kyropoulos (1926). The primary steps are

1. the melting of the source material in a crucible,
2. the preheating of a seed by holding it close above the melt,
3. touching the melt with the seed,
4. slight backmelting of the seed to achieve coherent growth,
5. pulling of the seed with decreasing diameter (necking) to reduce or eliminate structural faults (e.g., grain boundaries and dislocations) of the seed, and
6. pulling with increasing diameter until the desired radius is reached. The crystal shape is controlled by variation of the melt temperature and pulling rate. The crystal and crucible are rotated in opposite directions at 5–40 rpm to achieve a radially symmetric temperature distribution.

Quickly evaporating materials can be crystallized by the liquid-encapsulation technique where the melt is covered with a suitable liquid and the vessel filled with an inert gas of sufficiently high pressure. The Czochralski technique is also employed for growing crystals from solution, as described by Elwell and Scheel (1975).

Advantages: Fast crystal-growth rates of between 10^{-6} and 10^{-5} m/s are attainable. Very large crystals (e.g., silicon 2.2 m in length and 0.16 m in diameter) can be grown. Crystals of highest perfection (e.g., dislocation-free silicon) can be obtained. *Disadvantages:* Medium- to high-cost equipment is required.

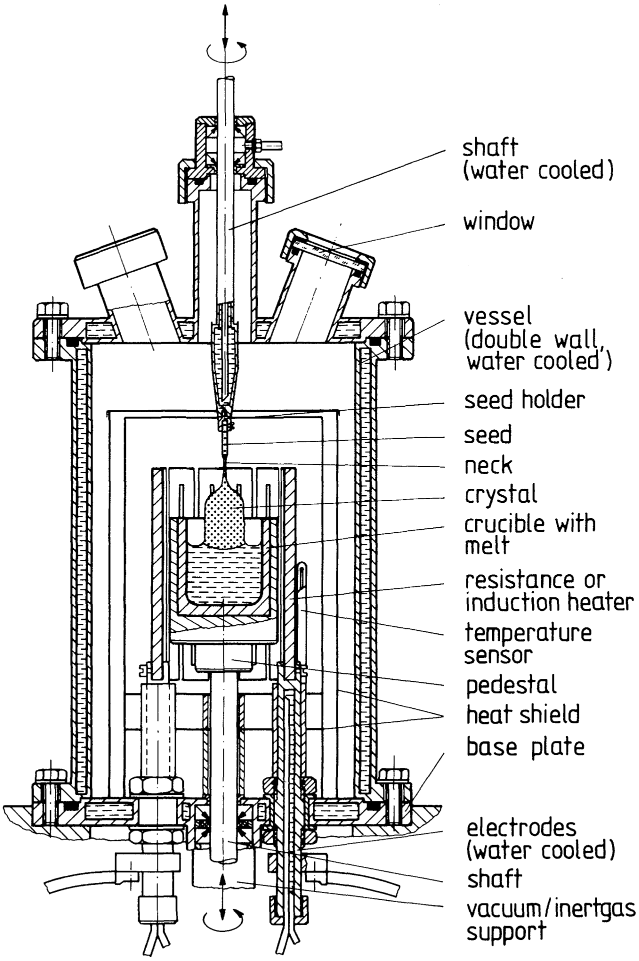
shaft
(water cooled)
window
vessel
(double wall,
water cooled)
seed holder
seed
neck
crystal
crucible with
melt
resistance or
induction heater
temperature
sensor
pedestal
heat shield
base plate
electrodes
(water cooled)
shaft
vacuum/inertgas
support

FIG. 12. Czochralski puller.

The crystal growth has to be continuously controlled since the heat-flow conditions change with crystal length.

3.2.2 Bridgman Technique. In the Bridgman method, crystallization of a melt is carried out in a boat [Fig. 5(b)], vertical tubular crucible [Fig. 5(c)], or ampoule [Fig. 5(d)]. Tubular containers are designed with a seeding end at the bottom similar to that shown in Figs. 9(a) and 9(b), while boats can have a seed installed at one end [Fig. 13(a)]. Figure 14 displays the principle of the vertical technique. The container is initially hung in the upper hot zone of a furnace for complete melting of the source material. Then the crucible is moved down into the cooler furnace zone, where an initial nucleus is formed in the seeding end. The growth of the nucleus is achieved by a sufficiently low drive rate. Alternatively, the heater can be moved while the container is held in a fixed position. Quickly evaporating materials can be processed in closed hot-wall containers (Fig. 14). Decomposition of a compound which forms a volatile constituent B can be prevented by an excess of B in a closed container at a sufficiently high wall temperature. Crystals of lower density than that of their melt (e.g., Ge) are processed in horizontal boats with a trapezoidal cross section where the upper side is greater than the bottom side.

Advantages: In general, low-cost equipment is sufficient. Medium growth rates between 10^{-7} and 10^{-6} m/s are reached. The processing is very simple, since it employs a constant temperature distribution and constant drive rate. The method offers the possibility of

1. crystal growth from solution, e.g., growth of peritectically melting compounds,

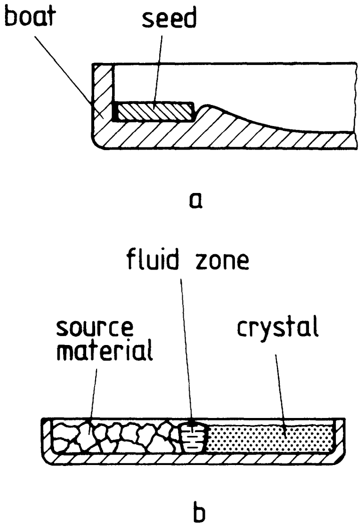

FIG. 13. Use of a boat for zone melting. (a) Seed mounting. (b) Crystal-growth method.

2. crystal growth of volatile or decomposing materials, and
3. doping with volatile elements.

Disadvantage: Dislocations may be formed if the thermal expansion of the container differs from that of the crystal.

3.2.3 Temperature Gradient Technique. The temperature gradient technique is, in principle, a modified vertical Bridgman method, where the crucible is always maintained at an initially fixed position in a particular furnace. The temperature of the heater locally decreases from the upper crucible part to the seeding end. Nucleation and crystal growth is forced by slow cooling of the furnace while the slope of the temperature distribution stays approximately constant.

Advantage: No movable mechanical parts are required. *Disadvantages:* The disadvantages are the same as those of the Bridgman techniques.

3.2.4 Liquid-Zone Techniques

3.2.4.1 Liquid-Zone Techniques with Crucible. The liquid-zone techniques are very similar to those of Bridgman with one difference: Only a small portion [Fig. 13(b)] of the source material is melted by heating with a short resistance oven or a short induction coil around the container. Crystal growth is achieved by driving the fluid zone from the seeding end to the opposite end by moving the container or heater. The method is widely applied for material refining through the use of a multi-heating-cooling zone assembly.

Advantage: A nearly constant (macroscopic) doping distribution may be obtained as shown in Fig. 22 (Sec. 5.2.2). *Disadvantages:* The disadvantages are the same as those of the Bridgman techniques.

3.2.4.2 Floating-Zone Methods. In the floating-zone techniques the melt hangs on a vertical feed rod (or crystal) and is supported by the vertical crystal (or feed rod) below without use of a container. The zone is achieved by heat sources, modified for particular materials as described in Sec. 2.2.3. The liquid is held in position by the surface tension σ which compensates for the weight. The resulting maximum crystal radius is given by

$$r_{\max}=0.92\sigma^{1/2}(\rho_f g)^{-1/2}, \tag{12}$$

where ρ_f is the density of the melt and g is the acceleration due to the gravity, as described by Brice (1965). The method is mainly applied to industrial production of silicon crystals of low oxygen content (Sec. 6.2).

Advantages: The method provides the fluid and crystal with the purest environment. The highest growth velocities (e.g., 10^{-4} m/s for silicon) can be achieved. *Disadvantage:* A

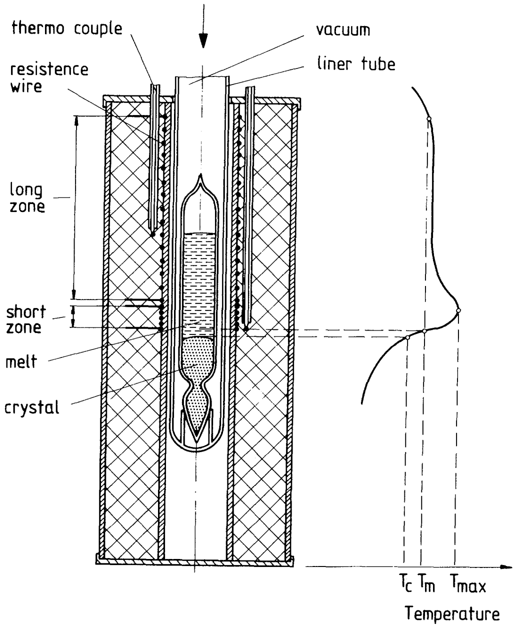

FIG. 14. Vertical Bridgman apparatus.

complex growth control is required by setting power, movement rate, and zone length. A suitable temperature distribution has to be applied for sufficient structural perfection.

3.2.4.3 Verneuil Technique. The Verneuil technique, described, e.g., by Falkenberg (1978), consists in passing powdered source material through a flame downwards to the top of a crystal on which a thin molten feeding zone is formed by the heat of the burner (Fig. 15). Crystal and flame are surrounded with an insulating ceramic tube to reduce heat loss. Oxygen and powder [(60–125)$\times10^{-6}$ m in diameter] are conducted through the center tube and hydrogen is passed through the outer tube. The powder chamber is divided by a fine-mesh screen into an upper large part which contains the powder source and a lower tapered part through which oxygen is conducted. The powder is mixed with the oxygen flow through the screen by tapping the chamber at a frequency of about 45 taps/min. Crystal growth from the thin molten zone is achieved by lowering the crystal while preserving the distance to the burner. Higher temperatures than those of a oxygen-hydrogen burner are obtained with an induction-coupled plasma as reviewed by Brice (1973) and Laudise (1970).

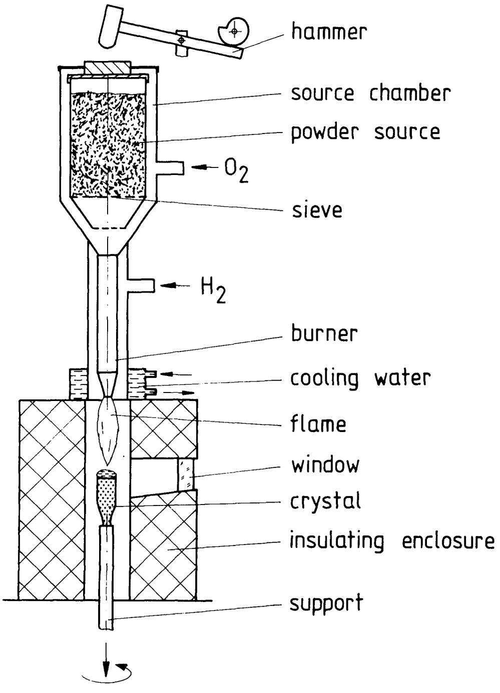

FIG. 15. Verneuil technique.

Advantages: Only low-cost equipment is needed for crystal growth of high melting oxides at temperatures below 2800 °C. *Disadvantages:* The heat control of the burner is difficult. The growth control is complicated, requiring variation of powder supply and crystal movement.

3.3 Techniques of Crystal Growth from a Solution

3.3.1 Solvent Evaporation. In general, a solute crystallizes from a solution at the container wall by slow evaporation of the solvent. A single crystal can be grown if one of the deposited crystals is selected and suspended in that (saturated) solution. The solvent evaporation is controlled by effusion through a hole in the closing top plate. The method requires a solvent whose vapor pressure is much greater than that of the solute. Therefore, the procedure is mainly restricted to low-temperature solvents, e.g., water. The technique is improved by temperature control, stirring, and feedback of the solution. Feedback is achieved by placing a sieve plate together with solute material between the top plate and the solution surface. The evaporating solvent condenses at the cooled top plate, flows back, and is saturated by solute material on the sieve.

Advantages: The method can be carried out with equipment of lowest cost. Large crystals can be grown by very simple processing. *Disadvantages:* The reachable growth rates are low (10^{-9}–10^{-7} m/s) and the optimum growth rate is difficult to control.

3.3.2 Temperature Difference Techniques

3.3.2.1 Normal Solution Conditions. In the temperature difference techniques a saturated aqueous solution, together with a great excess of solute material (nutrient), is maintained in a tank at the temperature T_s. The solution is continuously pumped through pipes to a second tank and back to the source tank. To grow a crystal the temperature T_c of the second tank is held lower than that of the source tank. The principle of the method is illustrated in Fig. 16. Here, in addition, the saturated solution first passes an overheated tank at the temperature T_u on the way to the

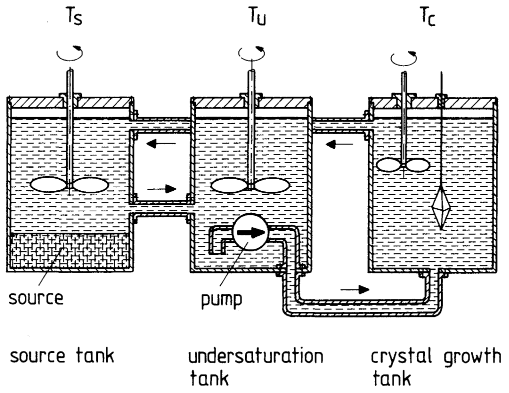

FIG. 16. Three-tank technique for growth of large ADP ($NH_4H_2PO_4$) crystals from aqueous solution, after Walker and Kohman (1948).

growth tank. Thereby residual floating nuclei of the source solution are dissolved.

Advantages: Isothermal crystal growth can be achieved at a constant supersaturation $C(T_s)$–$C(T_c)$. The growth rate is easily controlled by adjusting the temperatures. *Disadvantages:* The attainable growth rates are low (10^{-9}–10^{-7} m/s).

3.3.2.2 Hydrothermal Solution Conditions. Aqueous solutions under hydrothermal conditions (i.e., at temperatures higher than 100 °C and pressures greater than 10^5 Pa) are industrially applied to the growth of β-quartz crystals. The method requires a high-pressure autoclave as, e.g., described by Laudise (1970) and depicted in Fig. 17. The autoclave, of special steel (up to 1 m in diameter and 4 m in length), contains in the lower part a grit of natural quartz as source material (nutrient) and in the upper part suspended single-crystalline quartz seed plates. Nutrient and crystals are separated by a baffle plate with a small opening (5%) to reduce convective instabilities. For crystal growth the nutrient is held at 400 °C and the crystals at 380 °C (temperature difference technique). An 80% (at room temperature) filling with water results in a pressure of about 1.4×10^8 Pa. The solubility of quartz is increased by adding NaOH. A growth rate of 10^{-8} m/s is obtained with a 1.0 M NaOH solution. A review of hydrothermal synthesis of other materials is given by Rabenau (1985).

Advantages: Low-cost equipment is sufficient for synthesis of small crystals of new materials. *Disadvantages:* The costs for safe processing increase enormously with increasing autoclave volume.

3.3.3 Techniques with Gel Media. Most kinds of fine-crystalline precipitation in solutions at room temperature can be directed to single-crystal growth by precipitation in a gel as described by Henisch (1988) and Arora (1981). The possible chemical processes are, e.g., the formation of a reaction product, reduction of a compound, decomplexing of complex compounds, or outsalting. Water glass (sodium metasilicate solution), agar (carbohydrate polymer), and gelatin (protein) are the most commonly used gels which sufficiently reduce the nucleation frequency.

Figure 18 shows two principles of reaction-product precipitation in a silica hydrogel. The gel is prepared from Na_2SiO_3, H_2O, and a suitable acid at a *p*H value between 6 and 8. In Fig. 18(a) the gel contains the acid (here 2 M $C_4H_6O_6$) as one reactant. The other reactant (here 1 M $CaCl_2$) is dissolved in water which is placed above the set gel. Crystals of the solid reaction product (here $CaC_4H_4O_6$) are formed in the gel. In the U-tube method [Fig. 18(b)], the gel contains a minimal amount (here 0.2

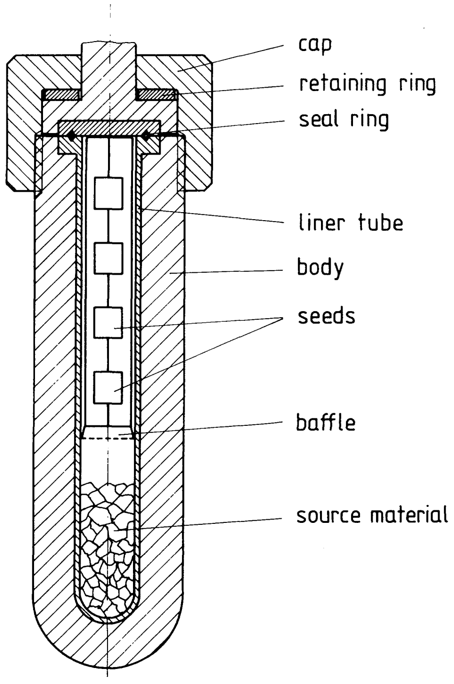

FIG. 17. Autoclave for quartz growth under hydrothermal conditions.

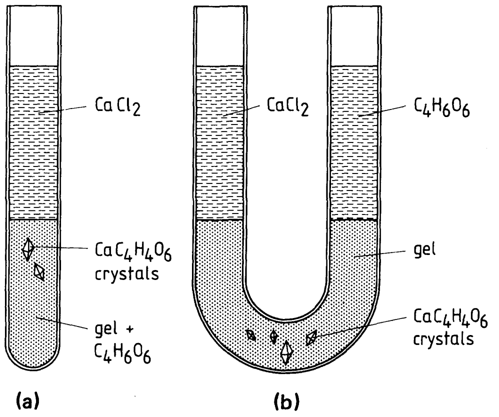

FIG. 18. Crystal growth in gel.

M $C_4H_6O_6$) of acid, which is sufficient forsetting. The reactants (here 1 M $CaCl_2$ and 2 M $C_4H_6O_6$) are placed on opposite sides of the set gel.

Advantages: Crystal growth in gels is one of the cheapest methods. *Disadvantages:* The attainable growth rates are very low (10^{-9}–10^{-8} m/s) and crystals above 10^{-2} m are hard to obtain.

3.3.4 Slow-Cooling Techniques. Growth of crystals by slow cooling is widely applied with aqueous, metal, and fluxed melt solutions. Tanks of glass enclosed in fluid baths and equipped with stirrers, and rotatable seed holders are used for aqueous solutions. Molten salt solvents with fluoride constituents are mainly employed for growth of oxide crystals. Detailed descriptions are given by Elwell and Scheel (1975) and Wanklyn (1975). High-temperature solutions are contained in crucibles of noble metals (platinium, iridium, etc.). They are sufficiently closed to release the pressure without noticeable solvent evaporation. The growth procedure is composed of several steps and illustrated in Fig. 19, where $LiAlSiO_4$ is used as a typical example.

1. The procedure is started by slowly heating (e.g., at 50 °C /h) the source materials in a chamber furnace to the soaking temperature T_{soak}. This temperature is preset 20–50 °C above the saturation temperature T_i.
2. The soaking temperature is held between 2 and 24 h for complete dissolution.
3. Then follows a fast cooling (e.g., at 100 °C /h) to the initial cooling temperature T_i, and

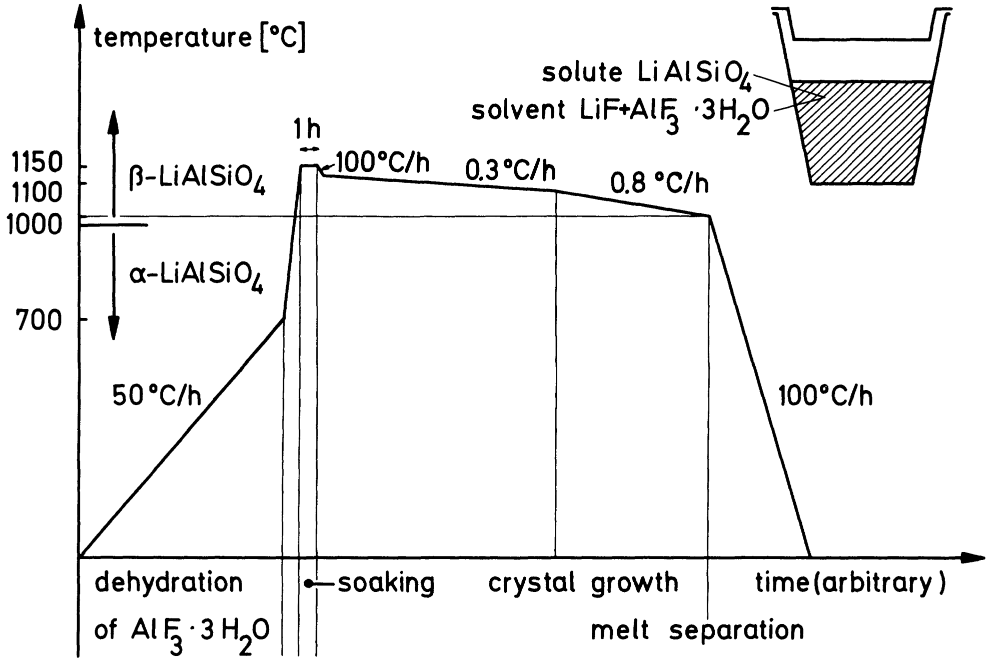

FIG. 19. Growth of eucryptite from fluoride flux as an example of the high-temperature solution technique (see text).

4. a slow cooling (e.g., at 0.3 °C/h) over several days. The slow cooling achieves a low number of nuclei and their undisturbed growth.
5. The slow cooling is continued at an increased rate (e.g., at 0.8 °C/h) since the normal growth rates decrease with the cube root (for linear solubility) of the cooling rate.
6. The crystal growth is terminated by decanting the solution from the crystals.
7. A slow cooling (e.g., at 20 °C) to room temperature prevents mechanical damage of the crystals.

A sophisticated example (growth of $Y_3Fe_5O_{12}$ crystals) with a preset seed is described by Tolksdorf and Welz (1978).

Advantages: The method is easily carried out and generally requires low-cost equipment. *Disadvantages:* The properties of fluid, conditions of surface reactions, and incorporation of the crystal species change with decreasing temperature.

3.4 Techniques of Crystal Growth from the Solid Phase (Recrystallization)

It is fundamentally difficult to grow a single crystal from a solid phase since the mobility of the crystal-forming species is strongly restricted. Precipitation of a dissolved phase or devitrification of glass leads to very small crystals in general. Only materials of high dislocation mobility, like metals, can in some cases be transformed by recrystallization into single crystals. The process consists of the release of Helmholtz free energy through migration of grain boundaries, i.e., through the growth of large grains by the consumption of adjacent small ones. The procedure is achieved in two steps: the sample is (a) loaded with high free energy by straining and (b) either annealed in a homogeneous temperature field or moved through a temperature distribution in the direction of increasing temperature (in contrast to the Bridgman technique). The methods are described in detail by Laudise (1970) and Burgers (1963).

Advantage: The method generally requires low-cost equipment. *Disadvantages:* The method is restricted to materials of high dislocation mobility. Large single crystals are hard to obtain.

4. TECHNIQUES OF THIN-FILM GROWTH

Thin ($<100\times10^{-6}$ m) crystalline layers of various materials are of great importance for modern electronic devices. Crystalline films are grown by deposition or precipitation of the film species on a particular plate, called the substrate. The procedure is called (homo)epitaxy if the film consists of the same material as the substrate, heteroepitaxy if not (see HETEROSTRUCTURES AND SUPERLATTICES, METALLIC; HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR; THIN FILMS, NONELECTRONIC PROPERTIES). The control of layer thickness requires sufficiently low growth velocity, which is obtained by using a low-density fluid. Therefore layers are grown from a solution, i.e., by the technique of liquid-phase epitaxy, and from various kinds of vapor phases for which the techniques of chemical vapor epitaxy [see CVD (CHEMICAL VAPOR DEPOSITION)] and molecular-beam epitaxy (*q.v.*) are most widely used. Methods with low growth rates, low processing temperatures, and fast exchangeable fluids are applied to produce technologically important multilayers which consist of a periodic sequence of very thin ($<0.1\times10^{-6}$ m), variably doped semiconducting films.

4.1 Liquid-Phase Epitaxy

Liquid-phase epitaxy is mainly applied to layer growth of III-V compounds (e.g., GaAs) using the metal of the compound as a solvent. The methods are characterized by the procedure of attachment between substrate and solution:

1. Tipping: The solution and substrate are separately placed in a modified (graphite) boat. The solution is tipped over the substrate by tilting the boat.
2. Dipping: Similar to the Czochralski technique, the substrate attached to a rod is immersed into the solution.
3. Sliding: The substrate is held in a tray of a horizontal (graphite) plate. A second plate on top of the support plate contains a sufficiently large bore filled with solution and is slid over the substrate. Multilayers can be grown using a support plate with several solution chambers.
4. Centrifuging (especially designed for grow-

ing multilayers) as described by Bauser (1987): A rotating vertical graphite cylinder holds the solution at low revolution rates in two opposite inner tanks around a rotation axis, and at high revolution rates in two opposite outer storage channels. The substrates are mounted in radially connected channels between inner and outer storage chambers. The solution comes in contact with the substrate when the initially low revolution rate is increased.

In each method the layer growth is achieved either

1. by cooling, i.e., a saturated solution is brought in contact with the substrate and is cooled by a linear or sudden temperature change, or
2. by isothermal growth, i.e., a supersaturated solution is brought into contact with the substrate on which the crystal then grows at constant temperature by the consumption of the supersaturation.

4.2 Chemical Vapor Epitaxy

In most cases epitaxial layer growth through a heterogeneous chemical reaction is achieved in an open transport system (reactor), i.e., in a fused silica tube or vessel which contains the substrate and eventually source materials. The reactor is fed by appropriate vapor species (precursor molecules) in an H_2 or He carrier gas through a pipework which controls the flow rate. The various reactor designs are directed to different purposes—e.g., a laminar gas flow over the substrate, small exchangeable gas volume for multilayers, and high mixing of the vapors for homogeneous growth. The simplest type consists of a horizontal chamber with an obliquely mounted substrate or a vertical tube with the substrate attached to a horizontal pedestal. Two different principal reactor types result from the means of heating: namely, the cold-wall reactor where only the substrate support is (inductively) heated [Fig. 6(a)], and the hot-wall reactor where the substrate is heated by a resistance furnace outside of the reactor.

Three different thermochemical processes —pyrolysis, pyrolytic synthesis, and chemical transport—are employed for epitaxial deposition.

1. Pyrolysis (thermal decomposition) is preferred for deposition of elements and is achieved in a cold-wall reactor. An example is silicon epitaxy from a $SiHCl_3$-H_2 feed which decomposes at 1100 °C, i.e.,

$$2SiHCl_3(g) \rightarrow 2Si(s) + H_2(g) + 3Cl_2(g).$$

2. Pyrolytic synthesis consists of the layer formation from a reaction of at least two introduced vapor compounds. An example is the MOVPE (metalorganic vapor-phase epitaxy) of III-V compounds from a metal alkyl and phosphine in a cold-wall reactor, e.g., indium phosphide according to the reaction

$$(C_2H_5)_3In(g) + PH_3(g) \rightarrow InP(s) + 3C_2H_6(g)$$

at about 550 °C.

3. Chemical transport is a reversible process. It additionally requires a heated source chamber. Hot-wall reactors are preferred. An example is the chloride procedure according to the (simplified) reaction

$$2GaP(s) + 2HCl(g) \\ = 2GaCl(g) + P_2(g) + H_2(g)$$

at about 850 °C. The deposition rate is adjusted by the H_2 flow rate, chlorine/hydrogen concentration ratio, and temperature difference between source and substrate.

4.3 Molecular-Beam Epitaxy

A molecular-beam apparatus consists of a very sophisticated assembly (see MOLECULAR-BEAM EPITAXY). It achieves precisely controlled epitaxial layer growth by incident fluxes [molecular beams, 10^7–10^8 molecules/(cm^2 s)] of atoms and molecules in an ultrahigh vacuum ($<10^{-8}$ Pa). Compound layers, e.g., $Ga_{1-X}Al_XAs$, are synthesized by the appropriate fluxes of Ga, Al, and As_2 species. Each species is evaporated from a particular Knudsen effusion cell. Monolayers of monatomic height are obtained by closing and opening the cells with specific shutters. Several apparatus are integrated such as

1. a mass spectrometer to control the vapor phase,
2. an electron gun to analyze the surface structure by angle-resolved photoelectron emission spectroscopy and reflection high-energy electron diffraction, and

3. an ion gun to determine chemical composition by secondary-ion mass spectroscopy.

5. CHARACTERIZATION OF CRYSTALS

5.1 Morphology

Depending on the growth procedure, a single crystal can display rounded or flat as-grown faces. Primarily, the morphology of growing crystals, i.e., the macroscopic shape, is determined by the density of free bonds at the surface or alternatively by the number of periodic bond chains (PBC) in different directions parallel to the interface within a distinct slice. According to the PBC theory, which is discussed by Hartman (1973), three types of interfaces can be classified—namely, kinked, stepped, and flat faces. Kinked (*K*) faces contain no PBC and are therefore rough, hence fast growing and unlikely to appear. Stepped (*S*) faces contain one type of PBC and provide the surface with parallel growth steps. Flat (*F*) faces contain at least two types of PBC, need step sources, grow slowly, and are therefore most likely to appear. In general, *F* faces have the highest lateral atom density, lowest Miller indices, and lowest surface energy. Reaction of free bonds with impurities or formation of surface bonds from neighboring free bonds (reconstruction) can lead to other faces as the PBC analysis predicts.

The formation of rounded faces depends on the temperature, the form of the fluid isotherms, and fluid composition. Strong rounding is not observed in crystal growth far below the melting point. In the case of crystal growth from solution or reactive vapor (e.g., chemical transport), the final morphology is strongly polyhedral: composed of *F* faces and eventually *S* faces. When the crystal is dissolved the morphology becomes different from that of growth in general. While the growth morphology consists of slowly growing faces, the solution morphology is preferably composed of fast-growing faces.

In the case of growth from a nonisothermal melt or vapor the crystal form approaches the surrounding isotherms. Faceting is pronounced by the presence of impurities in the fluid. Figure 20 illustrates an example. The faceting and rounding of the crystal interface are represented by interface contours. The sample, an iodine crystal, was grown by the Pizzarello technique (Fig. 8) from iodine vapor which contained a small amount of krypton. The initial (F_i) and final (F_f) interfaces were obtained by holding the ampoule at fixed positions x_i and x_f, respectively. The sequential growth contours from F_i to F_f result from a sudden ampoule displacement from x_i to x_f. Only rounded faces are observed for backsublimation where the interface moved from F_f to F_i after the reverse displacement from x_f to x_i.

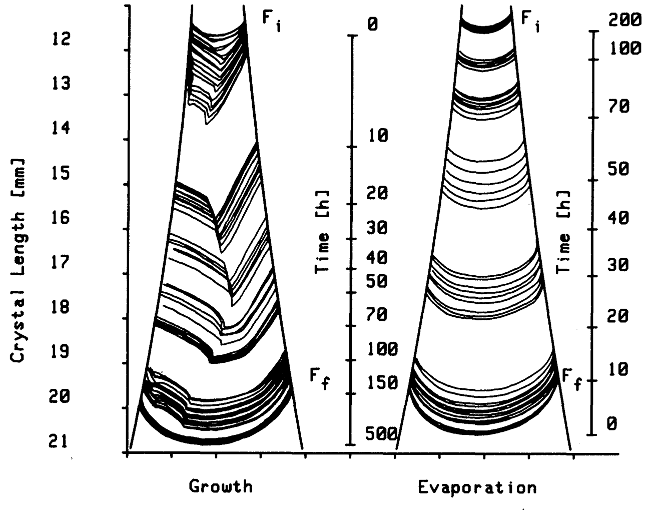

FIG. 20. Vapor-solid interface contours of iodine in the presence of krypton (400 Pa) at different times: growth at about 62 °C through sudden ampoule displacement; backsublimation (evaporation) after sudden reverse displacement.

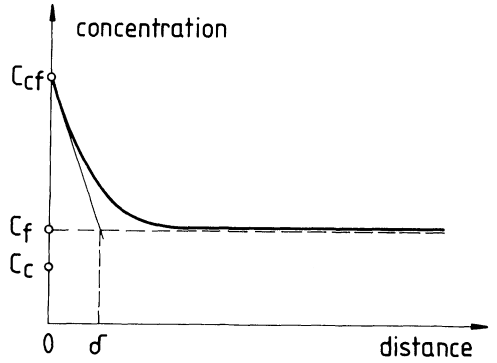

FIG. 21. Impurity distribution in front of a crystal growing from a melt. C_c is the concentration of the solute in the crystal, C_{cf} at the interface, and C_f in the melt at a large distance from the interface. δ is defined as the width of the boundary where transport of melt is achieved by diffusion.

5.2 Segregation of Impurities

5.2.1 Segregation Coefficients. In general, impurities dissolve in the coexisting melt and solid phases at different concentrations. The ratio $k_{\text{eq}}=C_{ce}/C_{fe}$ of the crystal impurity concentration (C_{ce}) to that (C_{fe}) of the melt is called the equilibrium segregation (or distribution) coefficient. The coefficient is obtained from the appropriate binary phase diagram. If $k_{\text{eq}}<1$ the impurity becomes enriched in front of the growing crystal interface as illustrated in Fig. 21 and discussed, e.g., by Pfann (1952, 1966), Carruthers (1975), and Brice (1965, 1973, and 1986). The impurity enrichment leads to surface concentration C_{cf}. The local (exponential) change of the impurity concentration in the fluid is approximated by a linear change which leads to the thickness δ of the so-called solute boundary layer. Two additional segregation coefficients are defined: the interfacial $k^*=C_c/C_{cf}$ and the effective $k_{\text{eff}}=C_c/C_f$ where C_c is the impurity concentration in the crystal near the interface and C_f the concentration in the bulk fluid. Under steady-state conditions the two coefficients are related by the crystal-growth rate V_c and diffusion coefficient D through

$$k_{\text{eff}}=k^*/[k^*+(1-k^*)\exp(-V_c\,\delta/D)]. \qquad (13)$$

The concentration distribution of the crystal is thus severely affected by fluctuations of the growth rate V_c and boundary-layer thickness δ. Fluctuations of both parameters are caused by fluctuations in the fluid which may arise from stirring or gravity-driven convection. If free-fluid surfaces are exposed to thermal gradients additional, irregular convection (Marangoni convection) appears driven by the difference in surface tension. The fluctuating impurity incorporation leads to changes of impurity densities perpendicular to the growth fronts, i.e., striations parallel to the interface, which can be revealed by a suitable etchant.

5.2.2 Impurity Distribution through Zone Refining. The difference in impurity concentrations in coexisting fluid and solid phases is used for purification through sublimation, distillation, and zone refining. For zone refining the principles of floating-zone techniques used for crystal growth are applied. A high efficiency is maintained by designs with several heating zones. Techniques are described by Pfann (1966) and Parr (1960). When a single melt zone of a width L is passed through a solid rod the concentration profile of the ingot assumes the form

$$C_c(x)=C_i[1-(1-k_{\text{eff}})\exp(-k_{\text{eff}}x/L)] \qquad (14)$$

[after Pfann (1966)], where C_i is the initial mean concentration of the ingot (molten zone) and k_{eff} an effective segregation coefficient. In the Bridgman and Czochralski processes the solute concentration of the crystal of a total length L_t is given by

$$C_c(x)=k_{\text{eff}}C_i(1-x/L_t)^{k_{\text{eff}}-1} \qquad (15)$$

for complete solidification of the melt and a constant crystal cross section. Concentration

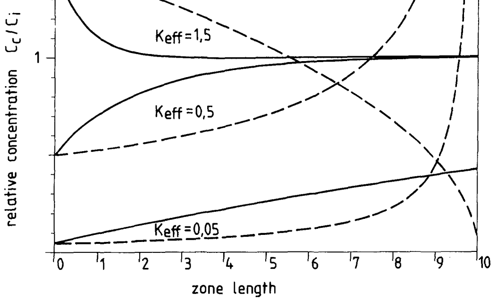

FIG. 22. Impurity distribution in a crystal for different segregation coefficients. Solid line for zoning. Dashed line for Bridgman procedure.

profiles according to Eqs. (14) and (15) are shown for different segregation coefficients in Fig. 22.

5.2.3 Precipitation. When the solubility of a solute decreases with decreasing temperature the solute may be precipitated in the crystal as small particles. Transparent crystals become opaque by this process. Preferably precipitation occurs along dislocations and grain boundaries.

5.3 Structural Defects

There are three classes of structural defects—

1. point defects,
2. line defects, and
3. two-dimensional defects—

that are or may be formed during crystal growth. Their formation is discussed by Vere (1987) and Chernov (1984). (See also IMPERFECTIONS IN CRYSTALS.)

5.3.1 Point Defects. The requirement of a minimum value of the Helmholtz free energy leads to some disorder of the structure above $T > 0$ K. There are two types of disorder:

1. vacant lattice sites, the so-called vacancies or Schottky defects, and
2. interstitials (Frenkel defects) which occupy sites other than those of the perfect structure.

Other point defects are formed by impurities which can occupy vacancies or interstitials. Point defects mainly influence the electrical and optical properties.

5.3.2 Line Defects. Line defects are called dislocations. They cannot terminate within the crystal, and so they either run between two surfaces, form a closed loop, or form a branch of another dislocation. A dislocation is characterized by the direction of its axis and Burgers vector. The vector results from comparison of the perfect structure with the actual structure around the dislocation. The defect is called a step dislocation when the Burgers vector runs perpendicular to the dislocation axis and a screw dislocation when the vector runs parallel to the axis. The screw dislocations serve as growth centers.

In Czochralski or floating-zone techniques, dislocations usually originate from the seed. If the dislocations are sufficiently strongly bonded by the lattice (e.g., silicon) they can be eliminated by growing a neck after seeding. During the crystal growth additional dislocations can be formed by shear strains which may result from

1. radial temperature gradients in the crystal,
2. precipitated impurities,
3. trapped impurity particles, or
4. sticking of the crystal to the container wall.

The formation of new dislocations is enhanced by multiplication of already existent dislocations as described by Hull (1975) and Friedel (1964).

5.3.3 Two-Dimensional Defects. Essentially there are three types of two-dimensional defects: (a) grain boundaries, (b) twins, and (c) stacking faults.

(a) Low-angle grain boundaries are mainly formed by multiplication of dislocations due to a straining stress field. Therefore, in the simplest case, low-angle boundaries consist of an array with parallel dislocations. For example, one family of step dislocations forms a pure tilt grain boundary where the dislocation axis is parallel to the tilt axis, while a family of screw dislocations forms a pure twist grain boundary where the dislocation axis is perpendicular to the boundary plane. The structure of high-angle boundaries is complex. They arise from intergrowth of differently oriented grains or from noncoherent nucleation on the growing surface (see GRAIN BOUNDARIES).

(b) Twin boundaries form a mirror plane where the structure is reflected. These particular faults originate either from a noncoherent nucleation or from mechanical deformation. Twinning is frequently observed for III-V and II-VI compounds which crystallize in the sphalerite structure. Here a twin is formed by a 180 °C rotated nucleus on a {111} plane. In contrast to grain boundaries, these twins are unmovable during heat treatment. Multiple twinning is caused by a too high supersaturation. Particular multitwin structures lead to dendritic growth as described, e.g., by Lindberg and Faust

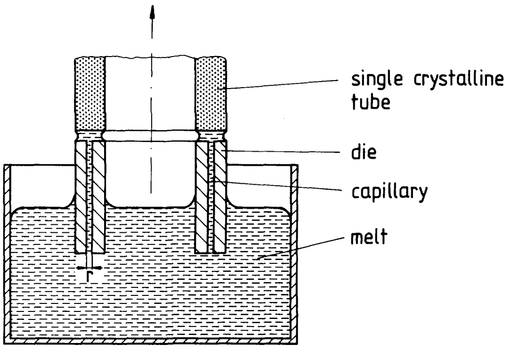

FIG. 23. Growth of a tube by the edge-defined film-fed method.

(1962).

(c) Stacking faults are mainly formed by precipitation of vacancies or interstitials so that either an atomic layer (intrinsic) is left out or an additional one (extrinsic) is inserted. Stacking faults are often observed on {111} planes of the sphalerite and diamond structures. Stacking faults extend within the crystal over a limited area, in contrast to twins, grain boundaries, and dislocations. More or less periodically different stacking (polytypism) of crystallographic planes occurs in some layered structures. The particular stacking faults are caused by dislocation movements. The faults are described by Trigunayat and Verma (1976) and Farkas-Jahnke (1983).

6. EXAMPLES OF APPLICATIONS

6.1 Czochralski Techniques

6.1.1 Edge-Defined Film-Fed Growth. Particular shapes of cross sections can be made through the Czochralski technique by placing a wettable die with pores between the melt and the crystal interface as illustrated in Fig. 23 and described by Ciszek (1981) and Wald (1981). The crystal is supplied with melt through a thin layer on top of the die. The layer is fed by the fluid which creeps through the pores if the die thickness is less than $2\sigma_f \cos(\phi)/\rho_f gr$, where σ_f is the melt surface tension, ϕ the wetting angle, ρ_f the melt density, g the acceleration due to the gravity, and r the radius of the capillary bores. Preferably, the method is applied to the growth of thin sheets (e.g., Si with a graphite die) and tubes (e.g., Al_2O_3 with a Mo die).

6.1.2 Dendritic Silicon Web Growth. Dendritic Si webs, candidates for solar cells, are grown by the Czochralski technique as indicated in Fig. 24. The sheets contain at least two coplanar twin planes parallel to the wide (111) sides; in addition, they are bounded by two [112] dendrites and grown in the [112] direction where the particular structure of two neighboring twin boundaries acts as a reentrant corner for the Si incorporation. The growth of the webs is started with a thin dendritic [112] seed (10^{-3} m diameter) from a melt of asymmetric temperature distribution. A review is given by Seidensticker (1982).

6.2 Float Zoning of Silicon

Figure 25 shows the principle of (a) normal and (b) pedestal zoning of silicon with inductive heating at a frequency between 3 and 5 MHz as reported by Keller and Mühlbauer (1981). A solid wedge-shaped coil (pancake coil with inner diameter of 35×10^{-3} m) with bores for water cooling (10^3-A current) superposes on the surface tension an electrodynamical force. In this way, a five times larger diameter (10^{-1} m) than that given in Eq. (12) can be achieved. The feed rod and crystal rod are oppositely rotated and displaced with respect to the fixed coil. The procedure of seeding and necking for disloca-

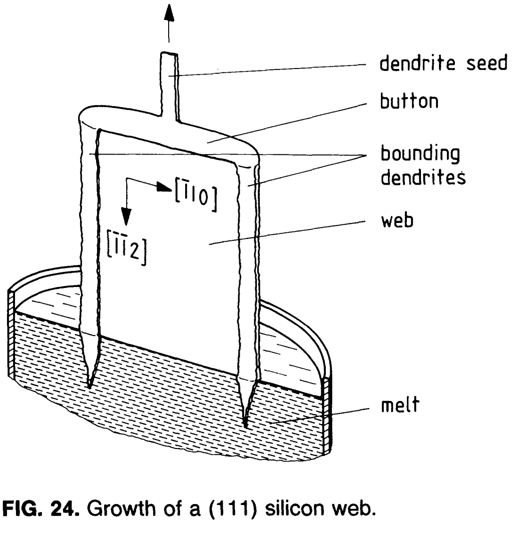

FIG. 24. Growth of a (111) silicon web.

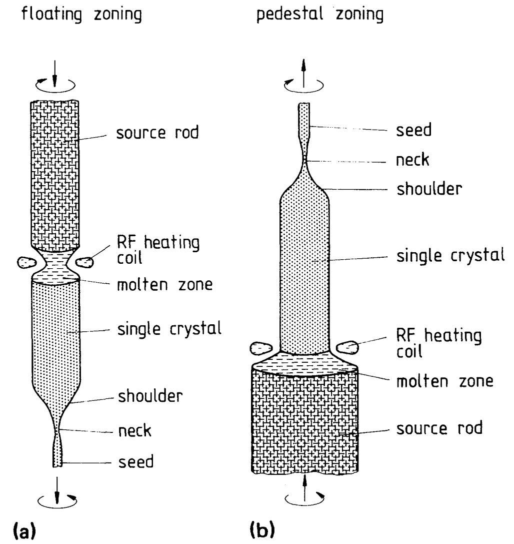
floating zoning
pedestal zoning
source rod
RF heating coil
molten zone
single crystal
shoulder
neck
seed
seed
neck
shoulder
single crystal
RF heating coil
molten zone
source rod
(a)
(b)

FIG. 25. Silicon crystal growth by float zone methods.

tion-free growth is, in principle, equal to the Czochralski method.

6.3 High-Pressure Technique of Diamond Growth

Diamonds of one carat (5×10^{-3} m in diameter) can be grown within a week from graphite dissolved in nickel at an external pressure of 6 GPa and a temperature of 1400 °C. The principle of this highly sophisticated method has been reviewed by Bundy *et al.* (1973), Wentorf (1974), and Nassau and Nassau (1979), and is schematically illustrated in Fig. 26. (See also DIAMONDS AND DIAMONDLIKE CARBON; HIGH-PRESSURE EFFECTS IN SOLIDS; HIGH-PRESSURE TECHNIQUES.) Two sections of a graphite heating tube are filled with a mixture of nickel and graphite powder. For better pressure control a diamond grit (synthetic or natural) is preferred instead of graphite powder. The two mixture compartments are lined with pyrophyllite $[Al_2Si_4O_{10}(OH)_2]$. The particular material is insulating and plastic at the employed high pressures and temperatures. A small seed plate of diamond is placed at the upper and lower pyrophyllite covering. The graphite tube is closed with nickel plates for current transfer and placed in a special belt press. The pestles and inner ring are made of cobalt-bonded tungsten carbide (carballoy: about 90% WC+10% Co). Both are surrounded by several belts of special steel. The free space between the tapered pistons and spheroid belts as well as that between the inner belt and graphite tube are insulated by pyrophyllite. By applying an electric voltage between the pistons the graphite tube is resistively heated slightly hotter at the middle than at the ends. Therefore the diamond grit dissolves in the middle and precipitates epitaxially at the seeds mounted at the tube ends (temperature difference technique).

Small diamond particles of about 0.25×10^{-3} m in diameter are crystallized within a short time (a few minutes) in a pyrophyllite

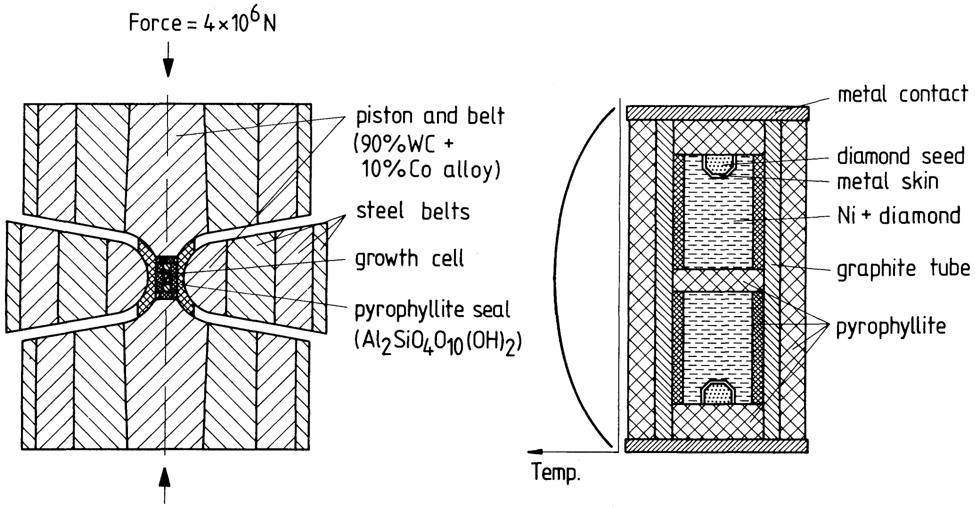
Force = 4×10^6 N
piston and belt (90%WC + 10%Co alloy)
steel belts
growth cell
pyrophyllite seal $(Al_2SiO_4O_{10}(OH)_2)$
metal contact
diamond seed
metal skin
Ni + diamond
graphite tube
pyrophyllite
Temp.

FIG. 26. Diamond growth. (See text.)

tube from a nickel graphite solution as shown in Fig. 26. The graphite is completely converted into diamond since the solubility of graphite is higher than that of diamond.

Advantages: Diamond grit can be fabricated at reasonable costs, and the grains are more suitable for tools than crushed natural diamond since they are free from cracks. *Disadvantage:* The fabrication of large crystals of high quality is more expensive than the cost of natural diamonds.

GLOSSARY

Associative Sublimation: Atoms of a solid join together to form molecules before entering the vapor phase.

Chemical Vapor Transport: Reversible transfer of a solid as a volatile compound through reaction with a gas.

Dendrite: Needlelike, often with branches, fast-growing crystal.

Devitrification: Transformation of glass into crystalline state.

Dissociative Sublimation: Decomposition of molecules of a solid when entering the vapor phase.

Dopant: A small defined quantity of substance added to the crystal.

Cluster: A group of like atoms or molecules forming an unstable condensate.

Nucleus: The smallest group of like atoms or molecules forming a stable condensate.

Peritectic Reaction: Formation of a solid phase β and liquid phase l from a solid phase α by heating of α, i.e., $\alpha \rightarrow \beta + l$.

Peritectic Temperature: The temperature at which the three phases α, β, and l coexist.

Segregation: Change of solute concentration during crystallization.

Stefan Problem: Problem of solving the heat-flow equation with a moving boundary.

Striations: Narrow bands of nearly constant solute concentration in crystals.

Substrate: A thin plate (in most cases a single crystal) onto which a thin solid layer is deposited from a vapor or liquid phase.

Supersaturation: State of vapor (or solute concentration) at pressures (concentrations) greater than that in equilibrium with the crystal.

Undercooling: Supercooling, state of melt at temperature below the melting point.

List of Works Cited

Aleksandrov, V. I., Osiko, V. V., Prokhorov, A. M., Tatarintsev, V. M. (1978), in: E. Kaldis (Ed.), *Current Topics in Materials Science*, Vol. 1. Amsterdam: North-Holland, pp. 421–480.

van Arkel, A. E. (1923), *Physica* **3**, 76–87.

Arora, S. K. (1981), *Prog. Cryst. Growth Charact.* **4**, 345–378.

Bardsley, W., Hurle, D. T. J., Joyce, G. C., Wilson, G. C., (1977), *J. Cryst. Growth* **40**, 21–28.

Bauser, E. (1987), in: R. F. C. Farrow, S. S. P. Parkin, P. J. Dobson, J. H. Neave, A. S. Arrott (Eds.), *Thin Film Growth Techniques for Low-Dimensional Structures*, New York: Plenum, pp. 171–194.

Bennema, P., Gilmer, G. H. (1973), in: P. Hartman (Ed.), *Crystal Growth: An Introduction*, Vol. 1, Amsterdam: North-Holland, pp. 263–327.

Bird, R. B., Stewart, W. E., Lightfood, E. N. (1960), *Transport Phenomena*, New York: Wiley.

Bockris, J. O'M., Razumney, G. A. (1967), *Fundamental Aspects of Electrocrystallization*, New York: Plenum.

Bohm, J. (1985), *Acta Phys. Hungar.* **57**, 161–178.Brice, J. C. (1965), in: E. P. Wohlfarth (Ed.), *Selected Topics in Solid State Physics*, Vol. 5, *The Growth of Crystals from the Melt*, Amsterdam: North-Holland.

Brice, J. C. (1973), in: E. P. Wohlfarth (Ed.), *Selected Topics in Solid State Physics*, Vol. 12, *The Growth of Crystals from Liquids*, Amsterdam: North-Holland.

Brice, J. C. (1986), *Crystal Growth Processes*, Glasgow: Blackie (Wiley).

Bridgman, P. W. (1925), *Proc. Amer. Acad. Arts Sci. (Boston)*, **60**, 303–383.

Buckley, H. E. (1951), *Crystal Growth*, New York: Wiley.

Bundy, F. P., Hall, H. T., Strong, H. M., Wentorf, Jun, R. H. (1955), *Nature (London)* **176**, 51–55.

Bundy, F. P., Strong, H. M., Wentorf, Jr., R. H. (1973), in: P. L. Walker, Jr., P. A. Thrower (Eds.), *Chemistry and Physics of Carbon*, Vol. 10. New York: Marcel Dekker, pp. 213–263.

Burgers, W. G. (1963), in: J. J. Gilman (Ed.), *The Art and Science of Growing Crystals*, New York: Wiley, pp. 416–450.

Burton, W. K., Cabrera, N., Frank, F. C. (1951), *Philos. Trans. R. Soc.* London, A **143**, 299–358.

Carruthers, J. R. (1975), in: N. B. Hannay (Ed.), *Treatise on Solid State Chemistry*, Vol. 5. New York: Plenum, pp. 325–406.

Chernov, A. A. (1984) in: B. K. Vainshtein, A. A. Chernov, I. A. Shuvalow (1984) (Eds.), *Modern Crystallography III. Crystal Growth*, Berlin: Springer, pp. 1–297.

Chisholm, D. (1971), in: G. Cook (Ed.), *M & B Technical Library, TL/ME/2, The Heat Pipe,* London: Mills & Boon.

Ciszek, T. F. (1981), in: J. Grabmaier (Ed.), *Crystals Growth, Properties, and Applications,* Vol. 5, *Silicon,* Berlin: Springer-Verlag, pp. 109–146.

Crank, J. (1988), *Free and Moving Boundary Problems,* Oxford: Clarendon.

Czochralski, J. (1918), Z. *Phys. Chem.* **92**, 219–221.

Draper, R. C. J. (1975), in: B. R. Pamplin (Ed.), *Crystal Growth,* Oxford: Pergamon, pp. 497–520.

Elwell, D., Scheel, H. J. (1975), *Crystal Growth from High-Temperature Solutions,* London: Academic.

Faktor, M. M., Garrett, I. (1974), *Growth of Crystals from the Vapour,* London: Chapman and Hall.

Falkenberg, R. (1978), in: C. H. L. Goodman (Ed.), *Crystal Growth Theory and Techniques,* Vol. 2. New York: Plenum, pp. 109–184.

Farkas-Jahnke, M. (1983), *Prog. Cryst. Growth Charact.* **7**, 163–210.

Friedel, J. (1964), *Dislocations,* Oxford: Pergamon.

Hartman, P. (1973), in: P. Hartman (Ed.), *Crystal Growth: An Introduction,* Vol. 1. Amsterdam: North-Holland, pp. 367–402.

Henisch, H. K. (1988), *Crystals in Gels and Liesegang Rings,* Cambridge: Cambridge University Press.

Hershey, D. (1973), *Transport Analysis,* New York: Plenum.

van Hook, A. (1961), *Crystallization, Theory and Praxis,* New York: Reinhold.

Hull, D. (1975), *Introduction to Dislocations,* Oxford: Pergamon.

Jackson, K. A. (1975), in: R. Ueda, J. B. Mullin (Eds.), *Crystal Growth and Characterization,* Amsterdam: North-Holland, pp. 21–32.

Jones, M. E., Shaw, D. W. (1975), in: N. B. Hannay (Ed.), *Treatise on Solid State Chemistry,* Vol. 5. New York: Plenum, pp. 283–323.

Kaldis, E., Cadoret, R., Schönherr, E. (1987), in: H. U. Walter (Ed.), *Fluid Sciences and Material Science in Space,* Berlin: Springer-Verlag, pp. 355–404.

Keck, P. H., Golay, M. J. E. (1953), *Phys. Rev.* **89**, 1297–1227.

Keller, W., Mühlbauer, A. (1981), in: W. R. Wilcox (Ed.), *Preparation and Properties of Solid State Materials,* Vol. 5, *Floating-Zone Silicon,* New York: Marcel Dekker.

Koref, F. (1922), *Z. Elektrochem.* **28**, 511–517.

Kossel, W. (1927), *Nachr. Ges. Wiss. Göttingen, Math.-Phys. Kl.,* pp. 135–143.

Krüger, F., Finke, W. (1910), Patentschrift Nr. 22 82 46, Klasse 12c, Gruppe 2, Berlin: Reichsdruckerei.

Kyropoulos, S. (1926), *Z. Anorg. Allg. Chemie* **154**, 308–313.

Laudise, R. A. (1970), *The Growth of Single Crystals,* Englewood: Prentice-Hall.

Lindberg, O., Faust J. W., Jr., (1962) in: R. K. Willardson, H. L. Goering (Eds.), *Compound Semiconductors,* Vol. 1. New York: Reinhold, pp. 294–310.

Mecseki, A., Mezö, B., Voszka, A., Galántai, A., Szabó, B. (1987), *Cryst. Res. Technol.* **22**, 677–683; **22**, 685–693.

Mullin, J. W. (1985), "Crystallisation," in: *Encyclopedia of Chemical Technology,* Vol. 7. New York: Wiley, pp. 243–285.

Mutaftschiev, B. (1980) in: F. R. N. Nabarro (Ed.), *Dislocations in Solids,* Vol. 5. Amsterdam: North-Holland, pp. 57–126.

Nacken, R. (1915), *Neues Jahrb. Mineral. Geol. (Abt. B.)* **2**, 133–164.

Nacken, R. (1950), *Chem. Ztg.* **74**, 745–749.

Nassau, K., Nassau, J. (1979), *J. Cryst. Growth* **46**, 157–172.

Obreimow, I., Schubnikow, L. (1924), *Z. Phys.* **25**, 31–36.

Parker, R. L. (1970), in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics,* Vol. 25. New York, Academic, pp. 152–299.

Parr, N. L. (1960), *Zone Refining and Allied Techniques,* London: G. Newnes.

Peube, J.-L. (1977), in: R. Balian, J.-L. Peube (Eds.), *Fluid Dynamics,* London: Gordon and Breach.

Pfann, W. G. (1952), *Trans. AIME* **194**, 747–753.

Pfann, W. G. (1966), *Zone Melting,* New York: Wiley.

Pizzarello, F. (1954), *J. Appl. Phys.* **25**, 804–805.

Rabenau, A. (1985), *Angew. Chem. Int. Ed. Engl.* **24**, 1026–1041.

Ramachandrarao, P. (1978), *Z. Metallk.* **69**, 656–660.

Rosenberger, F. (1979), *Fundamentals of Crystal Growth I,* Berlin: Springer-Verlag.

Sauveur (1912), *Proc. Int. Assoc. Test. Mater.* **2** (11), No.II.6.

Schäfer, H. (1964), *Chemical Transport Reactions,* New York: Academic.

Schönherr, E. (1980), in: H. C. Freyhardt (Ed.), *Crystals Growth, Properties, and Application,* Vol. 2. Berlin: Springer-Verlag, pp. 51–118.

Seidensticker, R. S. (1982), in: H. C. Freyhardt (Ed.), *Crystals Growth, Properties, and Application,* Vol. 8, *Silicon, Chemical Etching,* Berlin: Springer-Verlag, pp. 145–172.

Shah, J. S., Wills, H. H. (1975), in: B. R. Pamplin (Ed.), *The Science of Solid State,* Vol. 6, *Crystal Growth,* Oxford: Pergamon, pp. 326–396.

Sigsbee, R. A. (1969), in: A. C. Zettlemoyer (Ed.), *Nucleation,* New York: Marcel Dekker, Chap. 4.

Scheffen-Lauenroth, Th., Klapper, H., Becker, R. A. (1981), *J. Cryst. Growth* **55**, 557–570.

Spezia, G. (1905), *Atti R. Accad. Sci. Torino* **40**, 254–262.

Stockbarger, D. C. (1936), *Rev. Sci. Instrum.* **7**, 133–136.

Stranski, I. N. (1928), *Z. Phys. Chem.* **136**, 259–278.

Strickland-Constable, R. F. (1968), *Kinetics and Mechanism of Crystallization*, London: Academic.

Tammann, G. (1914), *Lehrbuch der Metallographie*, Leipzig: Leopold Voss.

Tolksdorf, W., Welz, F. (1978), in: C. J. M. Rooijmans (Ed.), *Crystals Growth, Properties, and Applications*, Vol. 1. Berlin: Springer-Verlag, pp. 1–52.

Trigunayat, G. C., Verma, A. R. (1976) in: F. Lévy (Ed.), *Crystallography and Crystal Chemistry of Materials with Layered Structures*, Vol. 2. Dordrecht: D. Reidel, pp. 269–340.

Vere, A. W. (1987), *Crystal Growth Principles and Progress*, New York: Plenum.

Verneuil, M. A. (1902), *C. R. Acad. Sci. (Paris)* **135**, 791–794.

Voronkov, V. V. (1983), in: A. A. Chernov, H. Müller-Krumbhaar (Eds.), *Crystals Growth, Properties, and Applications*, Vol. 9. Berlin: Springer-Verlag, pp. 75–111.

Wald, F. V. (1981), in: J. Grabmaier (Ed.), *Crystals Growth, Properties, and Applications*, Vol. 5, *Silicon*, Berlin: Springer-Verlag, pp. 147–198.

Walker, A. C., Kohman, G. T. (1948), *Trans. Am. Inst. Electr. Eng.* **67**, 580–585.

Walton, A. G. (1969), in: A. C. Zettlemoyer (Ed.), *Nucleation*, New York: Marcel Dekker, Chap. 5.

Wanklyn, B. M. (1975), in: B. R. Pamplin (Ed.), *Crystal Growth*, Oxford: Pergamon, pp. 217–288.

Wentorf, Jr., R. H. (1974), in: R. H. Wentorf, Jr. (Ed), *Advances in High-Pressure Research*, Vol. 4. London: Academic, pp. 249–261.

Wilcox, W. L. (1971), in: R. A. Lefever (Ed.), *Preparation and Properties of Solid State Materials*, Vol. 1. New York: Marcel Dekker, Chap. 2.

Winter, E. R. F., Barsch, W. O. (1971), *Adv. Heat Transfer* **7**, 219–320.

Further Reading

Brice, J. C. (1973), in: E. P. Wohlfarth (Ed.), *Selected Topics in Solid State Physics*, Vol. 12, *The Growth of Crystals from Liquids*, Amsterdam: North-Holland.

Brice, J. C. (1986), *Crystal Growth Processes*, Glasgow: Blackie (Wiley).

Buckley, H. E. (1951), *Crystal Growth*, New York: Wiley.

Chalmers, B. (1964), *Principles of Solidification*, New York: Wiley.

Elwell, D., Scheel, H. J. (1975), *Crystal Growth from High-Temperature Solutions*, London: Academic.

Faktor, M. M., Garrett, I. (1974), *Growth of Crystals from the Vapour*, London: Chapman and Hall.

Gilman, J. J. (Ed.) (1963), *The Art and Science of Growing Crystals*, New York: Wiley.

Goodman, C. H. L. (Ed.) (1974), *Crystal Growth Theory and Techniques*, Vol. 1. New York: Plenum.

Goodman, C. H. L. (Ed.) (1978), *Crystal Growth Theory and Techniques*, Vol. 2. New York: Plenum.

Henisch, H. K. (1988), *Crystals in Gels and Liesegang Rings*, Cambridge: Cambridge University Press.

van Hook, A. (1961), *Crystallization, Theory and Praxis*, New York: Reinhold.

Laudise, R. A. (1970), *The Growth of Single Crystals*, Englewood: Prentice-Hall.

Lawson, W. D., Nielsen, S. (1958), *Preparation of Single Crystals*, London: Butterworths.

Mullin, J. W. (1972), *Crystallisation*, London: Butterworths.

Mullin, J. W. (1985), in: "Crystallization," in: *Encyclopedia of Chemical Technology*, Vol. 7. New York: Wiley.

Nassau, K. (1980), *Gems Made by Man*, Radnor: Chilton.

Pamplin B. R. (Ed.) (1975), *Crystal Growth*, Oxford: Pergamon.

Parr, N. L. (1960), *Zone Refining and Allied Techniques*, London: G. Newnes.

Pfann, W. G. (1966), *Zone Melting*, New York: Wiley.

Rosenberger, F. (1979), *Fundamentals of Crystal Growth I*, Berlin: Springer-Verlag.

Strickland-Constable, R. F. (1968), *Kinetics and Mechanism of Crystallization*, London: Academic.

Tarján, I., Mátrai, M. (1972), *Laboratory Manual on Crystal Growth*, Budapest: Akadémiai Kiadó.

Vainshtein, B. K., Chernov, A. A., Shuvalow, I. A. (Eds.), (1984), *Modern Crystallography III. Crystal Growth*, Berlin: Springer.

Vere, A. W. (1987), *Crystal Growth Principles and Progress*, New York: Plenum.

Verma, A. R. (1953), *Crystal Growth and Dislocations*, London: Butterworths.

Wilke, K. Th. (1988), *Kristallzüchtung*, Frankfurt/M: Verlag Harri Deutsch.

Series and Journals

Crystals Growth, Properties, and Applications, Berlin: Springer-Verlag (1978).

Crystal Res. Tech. (1981).

Current Topics in Materials Science, Amsterdam: North-Holland (1978).

Growth of Crystals (Translated from Russian), New York: Consultants Bureau (1959).

J. Cryst. Growth (1967).

North-Holland Series in Crystal Growth, Amsterdam: North-Holland (1973).

Preparation and Properties of Solid State Materials, New York: Marcel Dekker (1981).

Prog. Cryst. Growth Charact. (1977).

CRYSTALLINE STATE

A. I. Goldman, *Ames Laboratory of the United States Department of Energy and Department of Physics, Iowa State University, Ames, Iowa, U.S.A.*

Introduction: Crystalline and Noncrystalline Solids 365
1. **Crystal Structures** 366
1.1 Two-Dimensional Crystal Lattices 366
1.1.1 Bravais Lattices and Translational Symmetry 366
1.1.2 Point Groups and Space Groups 367
1.2 Crystallography in Three Dimensions 370
1.3 Describing Planes and Directions 371
1.4 Examples of Crystal Structures 372
1.4.1 Close-Packed Structures 372
1.4.2 The Diamond, Cubic Zinc-Blende, and Hexagonal Wurtzite Structures 373
1.4.3 The Sodium Chloride and Cesium Chloride Structures ... 374
1.4.4 The Calcium Fluoride Structure 374
1.5 More Complicated Crystal Structures 375
1.6 Surface Structures 375
1.7 Incommensurate Crystals and Quasicrystals 375
2. **Binding Forces and Cohesive Energy** 377
2.1 The Van der Waals–London Interaction 377
2.2 Ionic Bonding 378
2.3 Covalent Bonding 378
2.4 Metallic Bonding 379
3. **Defects in Crystals** 379
3.1 Vacancies, Interstitials, and Impurities 379
3.2 Chemical Disorder in Alloy Systems 380
3.3 Stacking Faults and Line Defects 380
4. **Crystallization** 382
4.1 Nucleation 382
4.2 Growth 382
4.3 The Recrystallization of Glasses 383
5. **Summary** 383
Works Cited 383
Further Reading 383

INTRODUCTION: CRYSTALLINE AND NONCRYSTALLINE SOLIDS

Historically, we have categorized solids as either crystalline or glassy (often referred to as amorphous solids) as illustrated in Fig. 1. A simple operational definition of crystallinity, which will be expanded upon in detail below, is the state of matter characterized by long-range positional order of its constituent atoms. That is, there is a set of well-defined rules that may be used to specify the positions of all atoms in the solid. In general, the task of describing the various crystalline structures is simplified by the presence of periodic translational order. Once the local arrangement of atoms in one region of the crystal is described, this basic structural unit may be translated through some set of well-defined vectors to construct the entire solid. For glassy materials, on the other hand, no such set of rules or relations exists beyond an average local structure enforced by the forces that bind atoms together. This means that although local short-range positional order may exist in glassy solids, there is no corresponding long-range positional order. A third class of structures which, though well-ordered, are not periodic crystals includes incommensurate structures and quasicrystals. These are solids which possess long-range positional order which is aperiodic. Although our primary purpose here

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

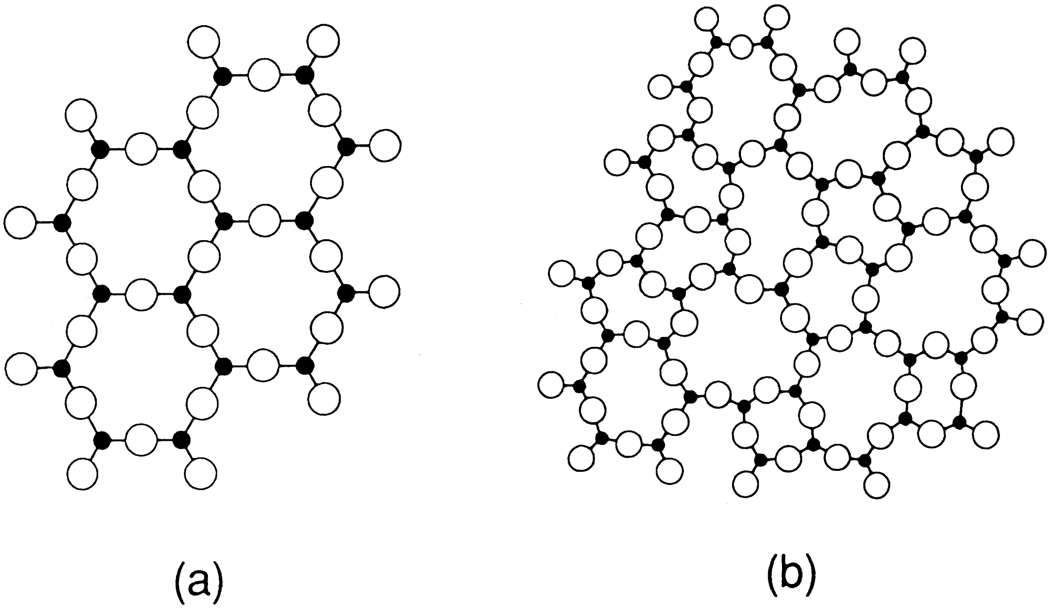

FIG. 1. Crystalline (a) and glassy (b) networks of atoms. (After Kittel, 1986.)

is the description of periodic crystalline solids, the latter two classes of structures will also be discussed where appropriate.

1. CRYSTAL STRUCTURES

This section deals with the definition and details of periodic crystal structures. In particular, we consider what restrictions periodicity places on the symmetries of crystal structures and the placement of atoms within these structures. We first develop the language of crystallography by considering two-dimensional crystal lattices and then move on to three-dimensional structures.

1.1 Two-Dimensional Crystal Lattices

1.1.1 Bravais Lattices and Translational Symmetry. Consider the two-dimensional array of points shown in Fig. 2. These points are said to belong to a *Bravais lattice* if the arrangement and orientation of lattice points, as viewed from one particular point, is identical to that viewed from all other points in the lattice. All points in Fig. 2 satisfy this condition. The hexagonal net of Fig. 3, however, clearly does not satisfy this requirement since points *X* and *Y* are not equivalent. A more formal definition of a two-dimensional Bra-

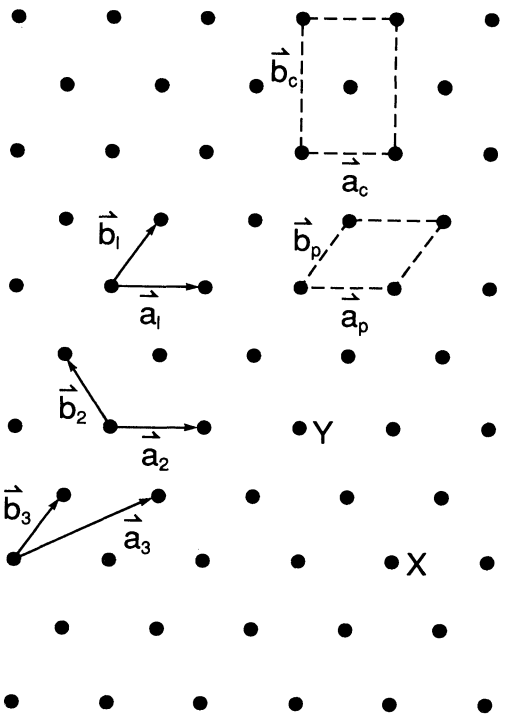

FIG. 2. The centered rectangular Bravais lattice. Primitive translation vectors and primitive and conventional unit cells are illustrated.

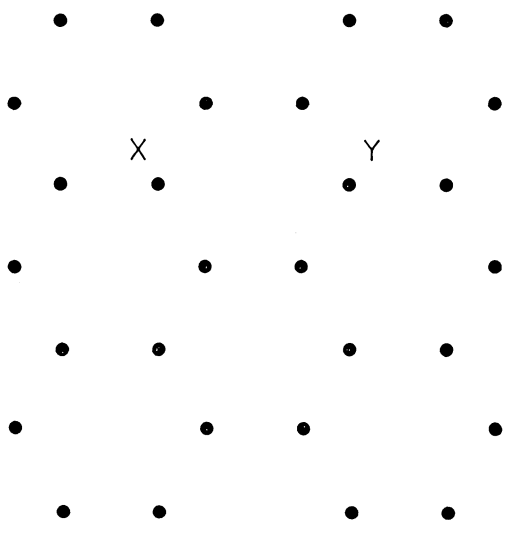

FIG. 3. Two-dimensional hexagonal net. Points X and Y are not equivalent.

vais lattice is that all points may be reached by a translation of the form

$$\mathbf{R}=n_1\mathbf{a}+n_2\mathbf{b},$$

where n_1 and n_2 are integers, and **a** and **b** are referred to as the *primitive translation vectors*. In three dimensions this definition must be modified to include translations out of the plane along a third primitive translation vector, **c**. A general translation in three dimensions is then described by the vector

$$\mathbf{R}=n_1\mathbf{a}+n_2\mathbf{b}+n_3\mathbf{c}.$$

For the two-dimensional Bravais lattice in Fig. 2, we can see that there are many possible choices for **a** and **b**. In terms of the three sets of primitive vectors illustrated in Fig. 2, a translation from point X to point Y can be represented by $-2\mathbf{a}_1+2\mathbf{b}_1$ or $-2\mathbf{b}_2$ or $-2\mathbf{a}_3+4\mathbf{b}_3$.

In the same way that we chose primitive translation for a Bravais lattice, we can define a *primitive unit cell* as the minimum-volume cell that will completely tile the plane. Note that the particular choice of this cell is also nonunique. One example of a primitive unit cell for the Bravais lattice of Fig. 2 is illustrated by the parallelogram spanned by vectors $\mathbf{a}_p$ ($\mathbf{a}_1$) and $\mathbf{b}_p$ ($\mathbf{b}_1$). If we now take this parallelogram and translate it through all linear combinations of $n_1\mathbf{a}_p+n_2\mathbf{b}_p$, for integer n_1 and n_2, we can tile the entire plane without overlap or gaps. It is often more useful, however, to define a second type of unit cell, termed a *conventional unit cell*, as depicted by the rectangle in Fig. 2. The reason for this is that the rectangular unit cell more clearly illustrates the underlying rectangular symmetry of the lattice. Note that while the primitive unit cell contains one lattice point, the conventional unit cell in this figure contains two lattice points. The four lattice points at the corners are shared by four unit cells while the lattice point at the center is completely within the domain of a single cell. Again, if we translate the rectangle by integer combinations of $\mathbf{a}_c$ and $\mathbf{b}_c$, we can tile the entire plane without overlap or gaps.

Before going on to describe the symmetries of lattices, we complete the picture by pointing out that the crystal structure is then fully determined by associating a set of atoms with each lattice point or unit cell. Clearly, if we decorate each unit cell with atoms in an identical manner, the same translation vectors used to move between points on a Bravais lattice can be used to move between identical atoms in all unit cells. In other words, the atomic arrangement is also periodic.

1.1.2 Point Groups and Space Groups.

Although there are an unlimited number of possible lattices in two dimensions, there are only five distinct two dimensional Bravais lattices, as shown in Fig. 4. The most general Bravais lattice is the oblique lattice since it places no restrictions on **a**, **b**, or the angle (γ) between them.

The symmetries of a crystal are defined by the kinds of operations that leave the structure invariant. In addition to the translations described above, there are also *point-group operations* to consider. The point group is the

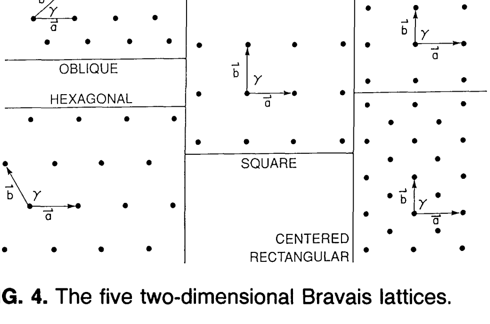

FIG. 4. The five two-dimensional Bravais lattices.

set of all operations which leave one point fixed. In two dimensions these operations include rotations about a fixed point and reflection across a line. The group of such actions which leave the lattice invariant (map the lattice points back onto themselves) defines the point symmetry of the lattice. The oblique lattice, for instance, may be left invariant only under rotations of π (180°) or 2π (360°). If we rotate the lattice points about one fixed point by these values, the rotated lattice is again coincident with the unrotated lattice. By further restricting the possible values of **a**, **b**, and γ, we obtain the special lattice types in Fig. 4. For instance, by restricting $a=b$ and $\gamma=90°$, we see (Fig. 5) that the square Bravais lattice is invariant under rotations of $\pi/4$, as well as reflections across lines x, y, x', and y'. Similarly, one can place constraints on **a**, **b**, and γ that lead to hexagonal Bravais lattices which are invariant under rotations of $2\pi/3$ and $2\pi/6$.

It should be pointed out here that the requirement of periodic translational order restricts the allowed rotational symmetries of two- and three-dimensional lattices to rotations by 2π, $2\pi/2$, $2\pi/3$, $2\pi/4$, and $2\pi/6$. In other words, in addition to a simple rotation by 2π (360°), periodic crystals may have twofold, threefold, fourfold, and sixfold rotation axes, but not five-fold or seven-fold or higher rotational symmetry. The reason for this is illustrated in Fig. 6 for the case of fivefold rotational symmetry. Consider two nearest-neighbor lattice points **A** and **B** separated by some minimum distance d. If we rotate **B** by an angle of $2\pi/4$ about **A** to get **B**′, and rotate **A** about **B** by the same angle to get **A**′, then both **A**′ and **B**′ are also lattice points of a square lattice with a minimum separation distance of d. If we perform the same set of operations on **A** and **B**, now specifying the rotation angle to be $2\pi/5$, the new lattice points **A**′ and **B**′ are now separated by less than d, contradicting the original assumption that d is the minimum distance. The conclusion from all of this is that periodic lattices may not have fivefold rotation axes.

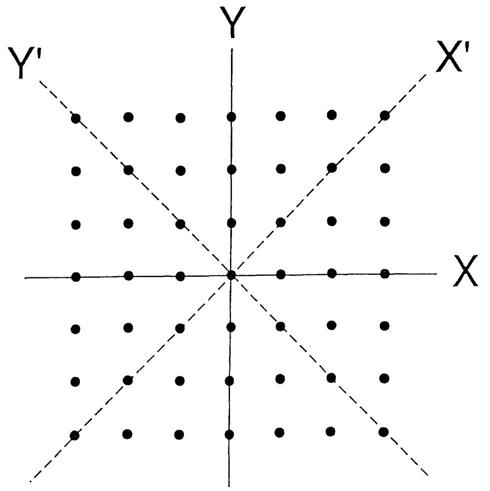

FIG. 5. Square Bravais lattice showing reflection symmetry across the lines x, y, x', and y'.

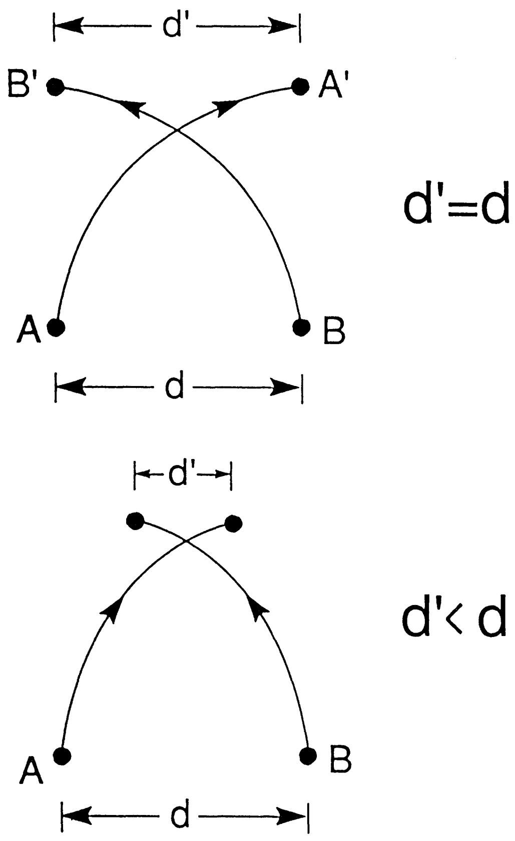

FIG. 6. An illustration of why fivefold rotation axes are not permitted for periodic structures.

The combined actions of all point-group operations (rotations and reflections) result in a total of ten two-dimensional point groups which are illustrated in Fig. 7. The oblong, triangular, square, and hexagonal filled symbols represent twofold, threefold, fourfold, and sixfold rotation axes, respectively. The heavy lines denote *mirror lines*. The presence

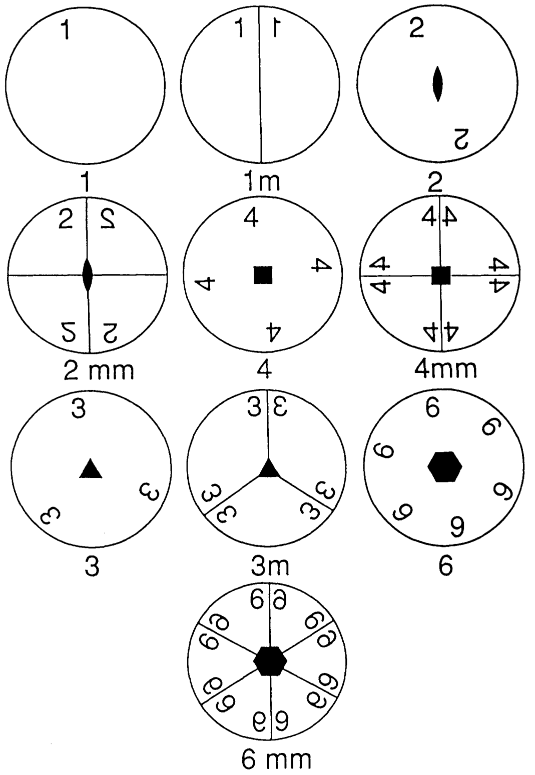

FIG. 7. The ten two-dimensional point group operations. The oblong, triangular, square, and hexagonal symbols denote twofold, threefold, fourfold, and sixfold rotation axes. The heavy lines represent mirror lines. (After the *International Tables for X-ray Crystallography.*)

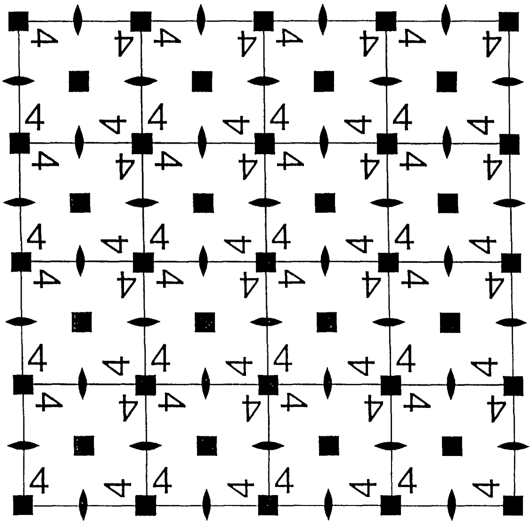

FIG. 8. The point group 4 translated through the vectors of a square lattice.

of a mirror lines indicates that reflection across that line leaves the pattern invariant. It is interesting to note that the combination of certain point-symmetry elements lead to others. For example, for the group 2*mm* in Fig. 7, the combination of a twofold rotation axis and a mirror line results in a second mirror line perpendicular to the first.

New symmetries appear when we consider the effect of compounding point-group and translational symmetries. For example, the point group 4 translated through the vectors of a square lattice results in a set of twofold axes midway between the lattice points and fourfold axes at the center of each unit cell (Fig. 8). Another operation, termed a *glide,* is composed of a reflection across the glide line and a translation of one-half the repeat distance along that line (see Fig. 12). Taken together, the combinations of point-group operations with translations, and glides, yield 17 two-dimensional *space groups.* In other words, there are 17 ways of arranging a set of identical objects periodically in two dimensions. All two-dimensional periodic structures belong to one of these space groups.

The three two-dimensional space groups of fourfold (square) symmetry are shown in Fig. 9 along with the diagram of spatially distributed symmetry operations found in the *International Tables for X-ray Crystallography.*

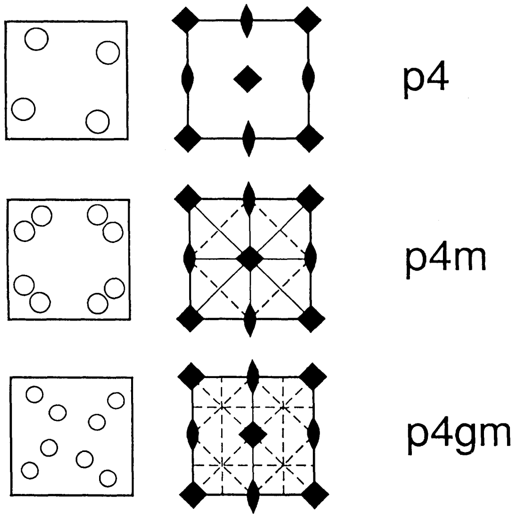

FIG. 9. The three two-dimensional space groups derived from the square space lattice and point-group operations. (After *The International Tables for X-ray Crystallography.*)

The left-hand figure illustrates an important point concerning the demands that various symmetry operations place upon the distribution of atoms in the unit cell. The space group label $p4$ indicates that the group is primitive (one lattice point per unit cell) and has fourfold rotational symmetry. If we place an atom at some *general position* (x,y) within the unit cell (where x and y are some fraction of the repeat distance a along the two orthogonal axes), three other atoms must be placed at $(-x,-y)$, $(y,-x)$, and $(-y,x)$ to satisfy the requirement of fourfold rotational symmetry. Likewise, the other space groups associated with the square lattice, which include additional mirror and glide symmetries, pose their own requirements for the symmetric placement of atoms. Note that there are also *special positions* such as (0,0), which only require one atom per unit cell.

1.2 Crystallography in Three Dimensions

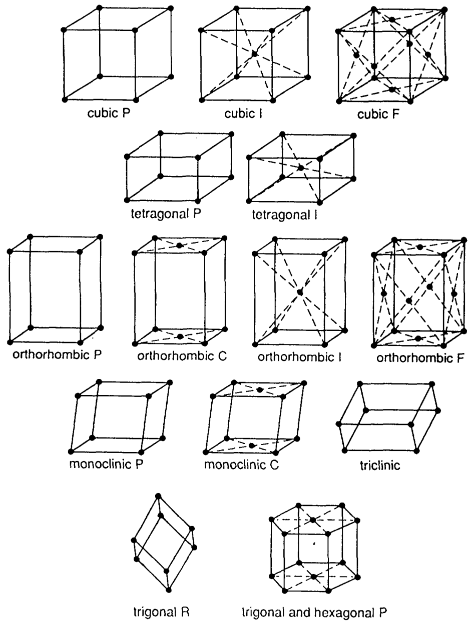

FIG. 10. Conventional unit cells of the 14 three-dimensional Bravais lattices. (After Kittel, 1986.)

TABLE 1. The seven crystal systems in three dimensions.

System	Conventional cell edge lengths	Restrictions on angles
Cubic	$a=b=c$	$\alpha=\beta=\gamma=90°$
Tetragonal	$a=b\neq c$	$\alpha=\beta=\gamma=90°$
Orthorhombic	$a\neq b\neq c$	$\alpha=\beta=\gamma=90°$
Hexagonal	$a=b\neq c$	$\alpha=\beta=90°$ $\gamma=120°$
Trigonal	$a=b=c$	$\alpha=\beta=\gamma\neq 90°$ but less than 120°
Monoclinic	$a\neq b\neq c$	$\alpha=\gamma=90°$ $\beta\neq 90°$
Triclinic	$a\neq b\neq c$	$\alpha\neq\beta\neq\gamma$

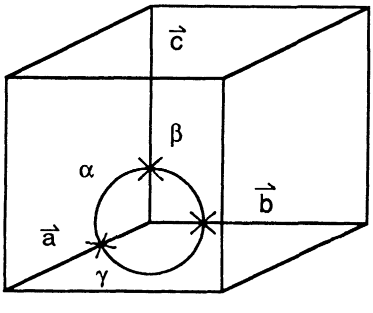

All of the language that we developed in the last two sections carries over to the description of three-dimensional periodic crystalline lattices. In three dimensions there are 7 crystal systems (summarized in Table 1) and 14 Bravais lattices with conventional unit cells as shown in Fig. 10. Each Bravais lattice may also be described by a minimum-volume primitive unit cell as described in Sec. 1.1.1. For instance, the primitive unit cell of the face-centered-cubic (fcc) structure is a 60° rhombohedron defined by the vectors connecting a corner of the conventional unit cell with the lattice points at the centers of the three faces. The primitive unit cell of the fcc (cubic F-type) lattice contains one lattice point, while the conventional unit cell contains four (each of the corners is shared by eight unit cells, while each of the six faces connects two unit cells). As before, the conventional unit cell is preferred since it more closely reflects the symmetry of the underlying lattice.

Including the 10 two-dimensional point groups, which remain valid operations, there are 32 three-dimensional point groups which include new symmetry operations such as mirror planes perpendicular to the axis of rotation, rotary-inversion axes (where a rotation about an axis of $2\pi/n$ is accompanied by a reflection across a plane perpendicular to the line), and inversion through a point [where points $(x,y,z) \rightarrow (-x,-y,-z)$].

There are two additional operations which are required to enumerate the 230 three-dimensional space groups completely. First, a *screw axis* is defined by a rotation accompanied by a translation along the rotation axis by a vector which is some rational fraction of a Bravais lattice vector. Figure 11, for instance, shows examples of right-handed and left-handed screw tetrads. In this case, equivalent sites are related by a 90° rotation and a translation by $c/4$, where c is the repeat distance along the screw axis. One can similarly define screw diads, triads, and hexads. The second operation defines a *glide plane*, and consists of a translation through a vector which is not in the Bravais lattice and a reflection across a plane containing that vector. A glide plane consisting of a translation by $c/2$ is illustrated in Fig. 12.

For further details on space groups and point groups in crystallography, we refer the reader to the first volume of *The International Tables of X-ray Crystallography.*

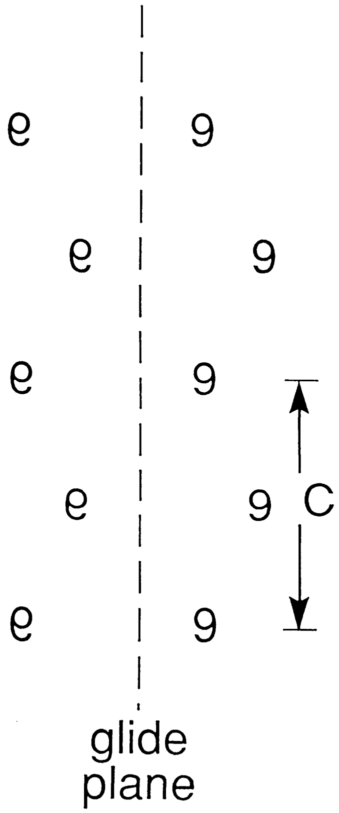

FIG. 12. A glide plane. The arrangement to the right of the plane is obtained by translating the assembly on the left by $c/2$ and reflecting across the plane.

1.3 Describing Planes and Directions

Quite often we are concerned with the arrangement of atoms in a particular plane of a three-dimensional crystal or the direction of the normal to that plane. Lattice planes in the cubic system are shown in Fig. 13. The plane itself is specified by a set of three *Miller indices* (h,k,l) which are defined as a set of integers which are inversely proportional to the intercepts of the plane along the crystal axes. For example, the (2,1,1) plane in Fig. 13 intercepts the a, b, and c axes at points $a/2$, b, and c. Similarly, the $(-1,1,1)$ or $(\bar{1},1,1)$ planes lies along the body diagonal of the cube and has intercepts of $-a$, b, and c. We may also group symmetry-related planes in a crystal into a family denoted by $\{h,k,l\}$. For instance, in the cubic system the planes (1,0,0), (0,1,0), (0,0,1), $(\bar{1},0,0)$, $(0,\bar{1},0)$, and $(0,0,\bar{1})$ all belong to the {1,0,0} family of planes. In contrast, for a crystal with tetragonal symmetry the (0,0,1) and $(0,0,\bar{1})$ planes do not belong to the {1,0,0} family. The direction of the plane is defined by the normal to the lattice plane of interest, and denoted by the same set of Miller indices enclosed in square brackets, $[h,k,l]$. The [2,1,1] direction in Fig. 13 is therefore normal to the (2,1,1) plane.

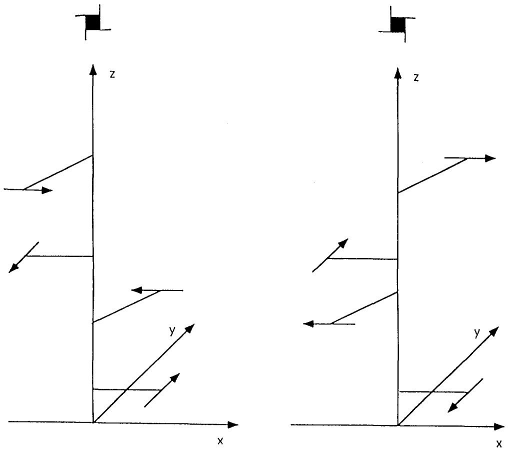

FIG. 11. Right-handed and left-handed screw tetrads. The arrangements in planes displaced by $c/4$ are related by a 90° rotation about the line.

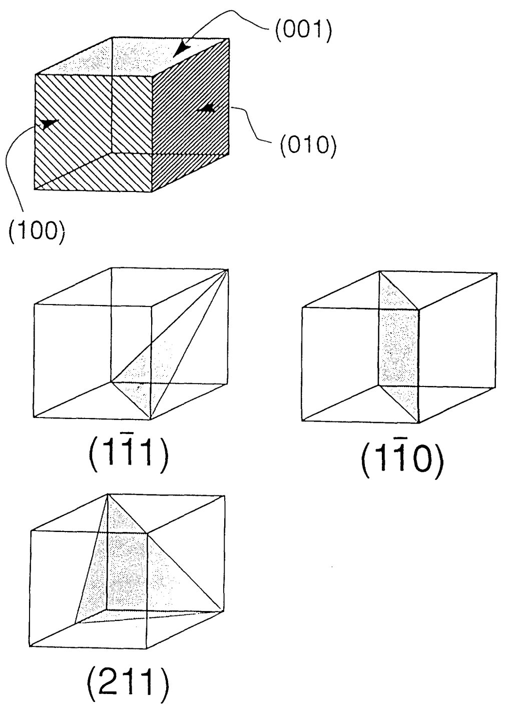

FIG. 13. Lattice planes in a cubic crystal. Directions are specified by the outward normals to these planes.

1.4 Examples of Crystal Structures

1.4.1 Close-Packed Structures. With few exceptions, metals crystallize into highly symmetric, densely packed structures of the cubic and hexagonal lattices shown in Fig. 14. As described below in Sec. 2.4, this arises from the nondirectionality of metallic bonding.

The *face-centered-cubic* (fcc) close-packed structure [Fig. 14(a)] may be described in terms of a cubic conventional unit cell with four atoms at positions (0,0,0), $(0,\frac{1}{2},\frac{1}{2})$, $(\frac{1}{2},0,\frac{1}{2})$, and $(\frac{1}{2},\frac{1}{2},0)$. Each atom is in contact with 12 nearest neighbors along the face diagonals of the unit cell at a center-to-center separation of $a/\sqrt{2}$, where a is the edge length of the unit cell. The density of the fcc structure may be calculated as follows: The volume of the cubic unit cell is a^3, and the total volume occupied by the four atoms per unit cell is $4\times 4\pi R^3/3$. If we take the radius of each atom as half the nearest- neighbor distance the packing fraction of the fcc close-packed structure is 74%.

The *hexagonal-close-packed* (hcp) structure itself is not a Bravais lattice, but may be viewed as two interpenetrating hexagonal lattices displaced from each other by $\mathbf{a}/3$, $2\mathbf{b}/3$, and $\mathbf{c}/2$. The atomic positions of the two atoms per hexagonal unit cell are then (0,0,0) and $(\frac{1}{3},\frac{2}{3},\frac{1}{2})$. Each atom is again coordinated by 12 nearest neighbors (six in the plane and two sets of three from the planes above and below). For the ideal close packing of hard spheres, the ratio of the c-axis lattice constant to the in-plane (a axis) lattice constants is $c/a = \sqrt{8/3}$.

There is a close similarity between the fcc close packing and hexagonal close-packing schemes. This is most clearly seen if we consider the packing of atoms in the planes perpendicular to the [111] directions of the fcc lattice or the [001] axis of the hcp structure. As illustrated in Fig. 15, the closest packing of atoms (equal sized spheres) in a plane is

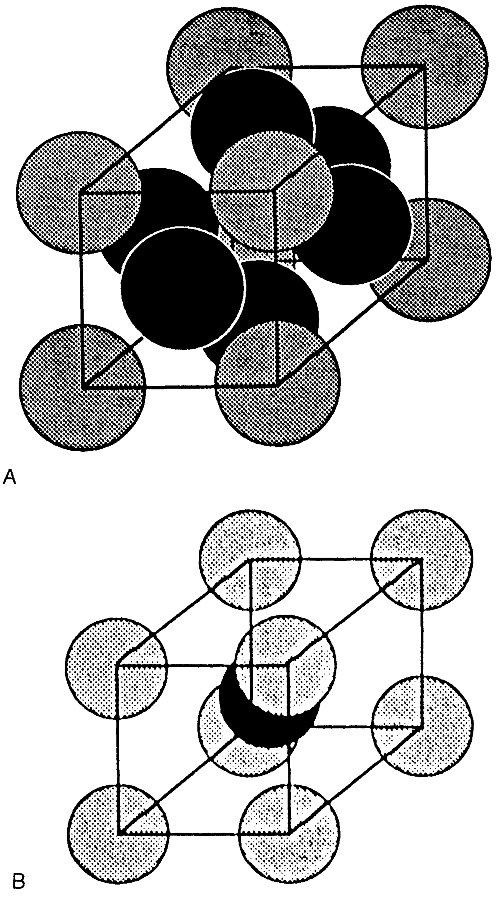

FIG. 14. Packing of equal-size spheres in the (a) fcc and (b) bcc structures. The light and dark shaded spheres represent atoms at the corners and faces of the fcc lattice, respectively, and the corners and body center of the bcc lattice.

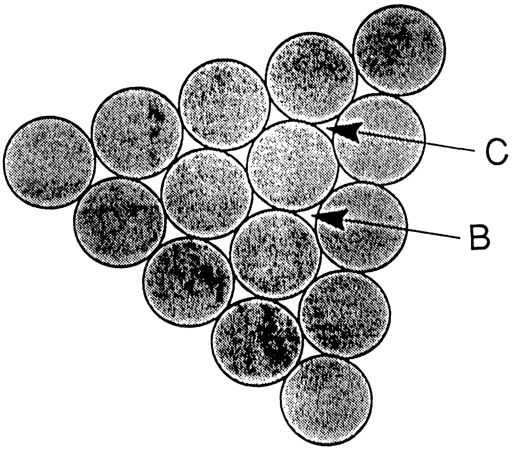

FIG. 15. Atomic packing in the (111) or (001) planes of the fcc or hcp structures. Interstices *B* and *C* denote sites available for atoms in the next layer.

accomplished by sixfold coordination of each atom. The pattern of atoms is precisely analogous to the two-dimensional hexagonal Bravais lattice that has threefold and sixfold symmetry. The voids between the atoms in this plane define the set of positions available for atoms in the next plane. Note that once one of the voids labeled *B* or *C* is chosen as the site for an atom in the next layer, all atoms in that layer sit in the *B* sites. If we continue up one more layer, there is a choice. The next layer of atoms may be situated either in the interstices labeled *C*, or directly over the atoms in layer *A*. If the packing proceeds in a sequence *ABCABCABCABC*,..., the fcc close-packed structure results, with the normals to these planes defining the [111] direction in the cubic lattice. On the other hand, if the stacking sequence is *ABABABABABAB*,..., the layers correspond to the planes perpendicular to the [001] direction of the hcp structure. A third stacking sequence, observed in some rare-earth metals, is *ABACABACABACABAC*,..., corresponding to the double hexagonal close-packed arrangement (dhcp).

The *body-centered-cubic* (bcc) close-packing arrangement [Fig. 14(b)] consists of two atoms per cubic unit cell at (0,0,0) and $(\frac{1}{2},\frac{1}{2},\frac{1}{2})$. Each atom is coordinated by eight nearest neighbors along the body-diagonal directions at a separation of $\sqrt{3}a/2$, and six atoms along the cube edge directions at a separation of a. The packing fraction of this structure is somewhat lower than the fcc structure (67%). The closest-packed planes are the (110) layers where the atoms are arranged in a square rather than the triangular network of the fcc (111) layers.

1.4.2 The Diamond, Cubic Zinc-blende, and Hexagonal Wurtzite Structures. The *diamond* structure is commonly found for covalently bonded semiconductors such as silicon and germanium. The structure (Fig. 16) is based upon a fcc space lattice with a two atom basis in the unit cell at (0,0,0) and $(\frac{1}{4},\frac{1}{4},\frac{1}{4})$. The unit cell, therefore, contains eight atoms. The resulting structure may be viewed as two equal-sized, interpenetrating fcc sublattices with a relative displacement of one-quarter of the distance along the body diagonal. Each atom is tetrahedrally coordinated by four nearest neighbors at a separation of $\sqrt{3}a/4$, and the packing fraction of these structures is quite low (34%). The *zinc-blende* structure (Fig. 17) of many other structures with covalent or partially covalent bonding is a simple modification of the diamond structure. In ZnS, e.g., the Zn atoms are situated on one of the fcc sublattices while the S atoms sit on the other. Each atom is tetrahedrally coordinated by atoms of the other species.

The relationship between the hexagonal *wurtzite* and zinc-blende structures is analogous to that between the fcc and hcp structures described in Sec. 1.4.1. Viewing along the [111] direction of the zinc-blende struc-

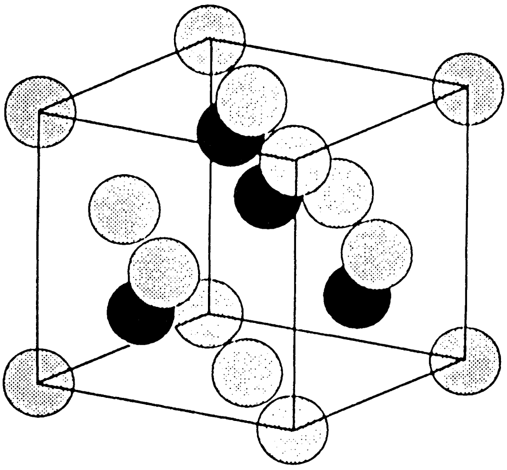

FIG. 16. Atomic packing in the diamond lattice. The light and dark shaded spheres represent atoms on the two fcc sublattices which are displaced along a body diagonal by $(\frac{1}{4},\frac{1}{4},\frac{1}{4})$.

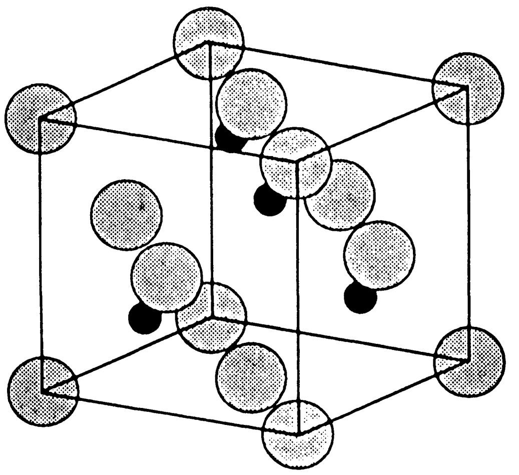

FIG. 17. Packing of different size atoms in the cubic ZnS lattice. The large spheres represent Zn while the small spheres represent the S atoms. Note the similarity to the diamond structure.

ture, we see planes of atoms of the same species stacked in the *ABCABCABC...* sequence. These same planes, stacked in the *ABABABAB...* sequence along the [001] direction, are the basal planes of the wurtzite structure. The space lattice of the wurtzite compounds is hexagonal with four atoms per unit cell. One species is situated at the position (0,0,0) and $(\frac{1}{3},\frac{2}{3},\frac{1}{2})$, while the other is found at $(0,0,u)$ and $(\frac{1}{3},\frac{2}{3},\frac{1}{2}+u)$. The offset u depends upon the compound but is close to $\frac{3}{8}$.

1.4.3 The Sodium Chloride and Cesium Chloride Structures. Many ionic solids crystallize into the NaCl or CsCl structures (Fig. 18). In the *sodium chloride* arrangement, each ion is coordinated by six oppositely charged ions in an octohedron. The space lattice is face-centered cubic with the halogen at positions (0,0,0) and the alkali atom at $(\frac{1}{2},\frac{1}{2},\frac{1}{2})$. The *cesium chloride* structure is simple cubic with the alkali ion at (0,0,0) and the halogen at $(\frac{1}{2},\frac{1}{2},\frac{1}{2})$. Each ion is coordinated by a cube of eight ions of opposite charge.

1.4.4 The Calcium Fluoride Structure. Many compounds of the form AB_2 crystallize in the CaF_2 structure (Fig. 19). The space lattice is fcc with a two-point basis and eight atoms per unit cell. The structure may be viewed as three interpenetrating fcc sublattices. Atoms of type A, and half of the B atoms, are found in the ZnS arrangement described

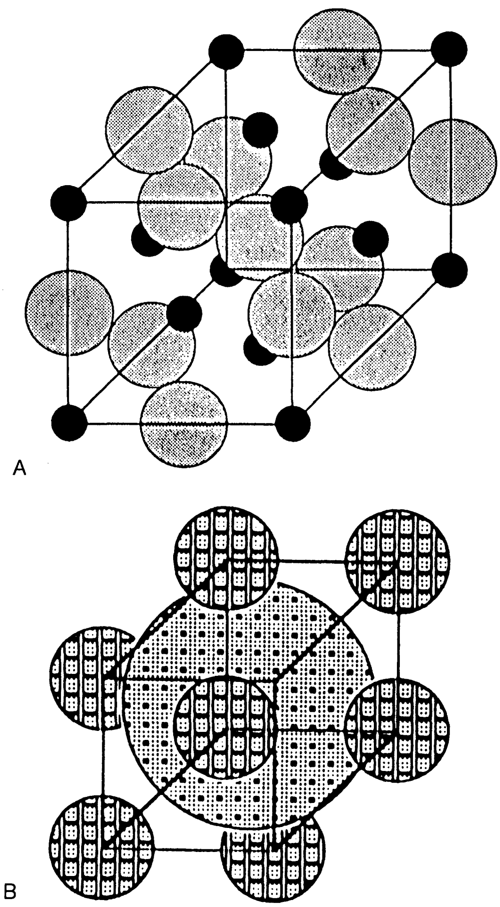

FIG. 18. The (a) sodium chloride and (b) cesium chloride structures.

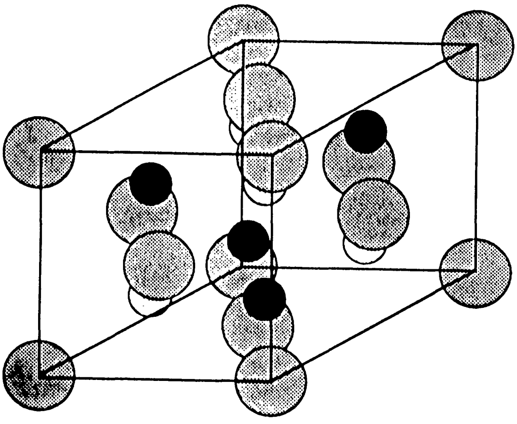

FIG. 19. The CaF_2 structure. Ca atoms are found at the corners of the cubic cell. Half of the F atoms (lightly shaded) are found at the *S* sites of the ZnS arrangement,and the remaining F (darkly shaded) are located on the third fcc sublattice as described in the text.

in Sec. 1.4.2. The remaining B atoms are found at positions $(-\frac{1}{4},-\frac{1}{4},-\frac{1}{4})$ plus face-centered translations. Each A atom is coordinated by a cube of eight B atoms, and each B atom is found within a tetrahedron of A atoms.

1.5 More Complicated Crystal Structures

All of the crystalline structures described in the last few sections are relatively simple, involving only a few atoms per unit cell. It is important to realize that this is not the general case. Unit cells with tens, hundreds, or even thousands of atoms are not uncommon and a description of the local environment of a particular component may be quite difficult. As an example, consider the structure of Al_6Li_3Cu illustrated in Fig. 20. The space lattice is bcc ($a = 13.9$ Å) with 160 atoms in the unit cell. The structure may be more simply described as a bcc packing of clusters of atoms centered at the corners and body center positions. Each cluster contains 105 atoms and has the shape of a truncated icosahedron.

1.6 Surface Structures

The surface of a crystal contains only a very small fraction of the total number of atoms in the solid, but is of importance since it represents the interface between the bulk crystal and its external environment. The structure of the surface layers may be essentially unchanged from that of the bulk crystal except for some modification in layer spacing due to the absence of atomic layers above the surface. On the other hand, there may be significant *reconstruction* of the surface, so that the top layer is out of registry with lower lying layers.

The structure of the surface layer may be cast in terms of one of the 17 space groups detailed in Sec. 1.1.2. Figure 21 shows three possible configurations, or *nets*, for the fcc (111) or hcp (001) surfaces. Here we imagine that the surface atoms are placed at the intersections of the net. The $p(1\times1)$ net is an unreconstructed surface. The $p(2\times2)$ net represents a reconstructed surface with primitive translation vectors twice as long as those of the unreconstructed $p(1\times1)$ net. The $(\sqrt{3}\times\sqrt{3})R30°$ net is also a reconstruction with translation vectors $\sqrt{3}$ times longer than the $p(1\times1)$ net, and rotated 30° with respect to the in-plane translation vectors of the bulk.

1.7 Incommensurate Crystals and Quasicrystals

While the crystal structures of the vast majority of elements, compounds, and alloys we encounter may be classified in terms of one of the 230 three-dimensional space groups, these do not exhaust all of the possibilities.

Incommensurate solids are materials which do not have periodic translational order but exhibit only the crystallography allowed rotational symmetries. As the term implies, incommensurability is derived from a periodic crystalline structure through the modulation of some property of that solid. The modulation itself often results from interactions which have a periodicity that is different from that of the basic lattice. In metals, e.g., instabilities can arise from interactions between the conduction electrons resulting in a modulation of the density of electrons in the solid (a charge density wave). The charged ion cores are also driven to readjust their position through their interaction with the electron gas. Therefore, as illustrated schematically in Fig. 22, the positions of like atoms in neighboring unit

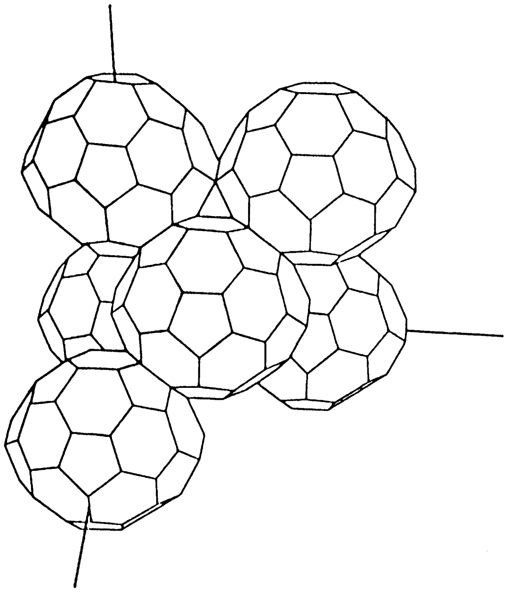

FIG. 20. The cubic phase of Al_6Li_3Cu may be viewed as a bcc packing of clusters of 105 atoms in the shape of a truncated icosahedron. Only the overall shape of the clusters is shown.

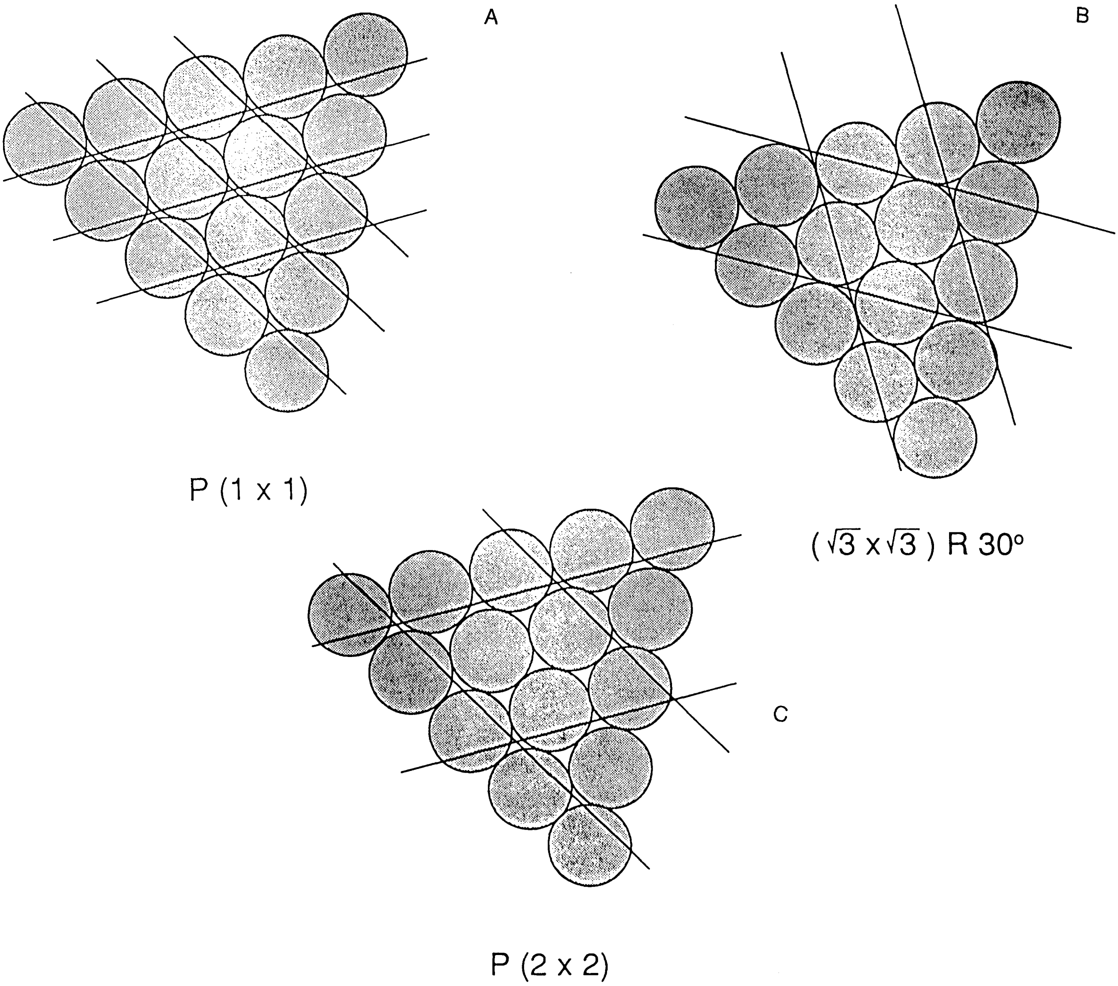

FIG. 21. Two-dimensional nets on the surface of the (111) planes of an fcc structure or (001) planes of an hcp structure.

cells are no longer connected by a simple translation vector, but by a translation plus a displacement from the average position. This displacement is also periodic, but may not have a period that is some rational multiple of the underlying lattice. In general, the incommensurability of the structure varies as a function of temperature and may exhibit "lock-in" transformations to a commensurate structure at particular temperatures.

Quasicrystalline solids not only lack periodic translational order, but exhibit rotational symmetries (e.g., five-fold rotation axes) which are crystallographically forbidden. The first quasiperiodic solid was discovered in 1984 in an alloy of aluminum and manganese with icosahedral symmetry (see Fig. 23). Since the icosahedral group admits rotations of $\pi/5$, the structure is inconsistent with any of the 230 periodic space groups discussed in Sec. 2.2. It is important to note, however, that the lack of periodic translational order does not imply that these materials are disordered glassy solids. For example, Fig. 24 shows a one-dimensional quasiperiodic Fibonacci sequence of atoms separated by two distances, $L(=\tau)$ and $S(=1)$. The sequence of atomic positions, R_n, from some origin may be defined by the rule

$$R_n=\frac{n}{1+1/\tau^2}-\frac{1}{\tau}\{n\tau\},$$

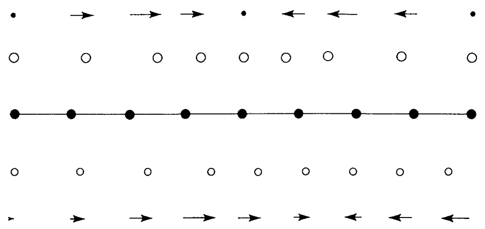

FIG. 22. One-dimensional modulated structures. The filled circles represent the atomic positions in the absence of a modulation, while the open circles denote the modulated structure. Top: The period of the modulation is a rational multiple of the underlying lattice. Bottom: The modulation is incommensurate with the underlying lattice.

where $\tau=\frac{1}{2}(\sqrt{5}+1)$ is known as the golden mean, n is an integer, and the curly braces denote the fractional part of the enclosed expression. The very existence of a rule specifying the atomic positions means that the structure is well-ordered, albeit aperiodic. Since our example here is one dimensional, the idea of forbidden rotational symmetries does not enter the discussion. Therefore, the Fibonacci sequence is perhaps more correctly described as an incommensurate structure. In two dimensions, however, one can construct a quasiperiodic tiling of the plane which is closely related to the one-dimensional Fibonacci sequence, using two rhombic "unit cells." The *Penrose tiling* shown in Fig. 25, or rather its three-dimensional cousin, is one of the candidate structures for the icosahedral alloys.

2. BINDING FORCES AND COHESIVE ENERGY

Having described and discussed the various classifications of crystalline solids, we now turn to a description of the interactions which bind atoms together into these structures. Binding forces act to lower the free energy of the solid with respect to an assembly of isolated atoms. The *cohesive energy* (*q.v.*) of a solid is, in fact, defined as the difference in total energy between the solid and its isolated constituents.

2.1 The Van der Waals–London Interaction

The weakest of the crystal binding forces is found in the inert-gas solids such as krypton, neon, xenon, and argon. These atoms have closed-shell configurations and are highly stable. When two atoms closely approach each other, however, an attractive interaction can result from induced dipole moments. This attractive interaction is a strongly decreasing function of interatomic separation (R), falling off as R^{-7}. There is also, of course, an even shorter-range repulsive interaction which results from electron overlap at very small separations. Taken together, the two terms combine to yield a total energy for the solid which is usually written as

$$U(r)=2\epsilon\left[A_{12}\left(\frac{\sigma}{R}\right)^{12}-A_6\left(\frac{\sigma}{R}\right)^{6}\right],$$

and is known as the Lennard-Jones 6-12 potential. Here σ and ϵ are parameters derived from the gaseous state data, and the constants A_{12} and A_6 depend upon the crystal structure. It should be pointed out that the $1/R^{12}$ dependence of the repulsive energy is introduced for calculational convenience and because it seem to work well for the description of binding energies in the inert-gas crystals. In fact, we only require that the repulsive interaction be much shorter-ranged than the attractive forces, and other forms for this term have been used.

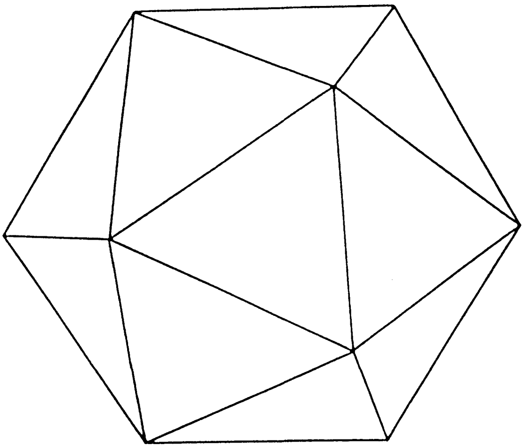

FIG. 23. An icosahedron is characterized by 12 vertices (fivefold axes), 20 triangular faces (threefold axes), and 30 edges (twofold axes).

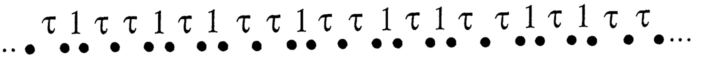

FIG. 24. A one-dimensional quasiperiodic Fibonnaci sequence. Atoms are separated by either a long (τ) or short (1) distance. The ratio of the number of L to S separations approaches τ as the length of the sequence increases.

All of the inert-gas solids crystallize into the fcc structure, and are characterized by binding energies on the order of 0.01 eV per atom. Since the binding is weak, solidification takes place only at low temperatures (< 10 K) and zero-point fluctuations play an important role in the stability of these structures. Since there are no free electrons in these solids, they are extremely good insulators.

2.2 Ionic Bonding

For ionic crystals, the dominant contribution to the binding energy is the interionic Coulomb attraction. NaCl is an example of a purely ionic solid; a Na 3s electron is transferred to fill the 3p shell of Cl, resulting in a positively charged Na^+ and negatively charged Cl^-. The stability of the structure is owed to the fact that the total energy decrease due to the interionic electrostatic attraction is greater than the energy required to transfer electrons between isolated neutral atoms. The electrostatic interaction is much stronger than the dipole interaction, and is longer ranged (the electrostatic force decreases like $1/R^2$). The total cohesive energy, including a term for the repulsive interaction between ion cores at small separation, can be written as

$$U(R)=C/R^n-\alpha q^2/R.$$

Here, C is a constant that can be determined from the known equilibrium separation of the ions and q is the net charge of the ions. The *Madelung constant* α essentially sums the contributions to the electrostatic potential between an ion and all of its neighbors (both near and distant).

In contrast to the weakly bound inert-gas crystals discussed above, ionic crystals are usually hard and have high melting points. Since the excess charge is highly localized around the negative ion, there are no free electrons to carry charge. Ionic crystals, in the absence of defects such as vacancies, are therefore poor conductors.

Alkali halide ionic solids crystallize into either the sodium chloride or cesium chloride cubic lattice (see Sec. 1.4.3). Although most crystallize into the NaCl structure, it is interesting that the electrostatic contribution to the cohesive energy is actually a bit larger (by about 1%) for the CsCl structure. The repulsive interaction between ions, however, increases with increasing coordination number and slightly favors the NaCl arrangement. The particular choice of structure is largely governed by the relative sizes of the ions.

2.3 Covalent Bonding

Covalent bonding in crystals is characterized by the sharing of electrons between constituent atoms. For instance, the 2p orbital of carbon, the 3p orbital of silicon, and the 4p orbital of germanium are all four electrons short of being filled. The energy of an assembly of carbon atoms, e.g., can be lowered by overlap between the unfilled orbitals, thereby sharing the valence electrons. Because of this sharing, the charge distribution is not completely localized around each ion as is found

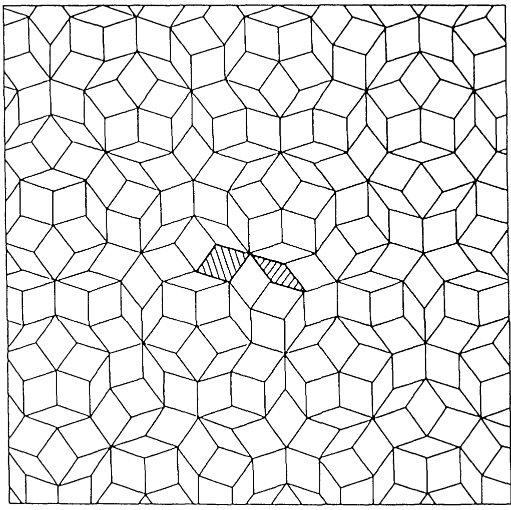

FIG. 25. A two-dimensional Penrose tiling of "fat" and "skinny" rhombuses. Note the presence of local regions with clear fivefold rotational symmetry. The three-dimensional analog of this tiling is one model for the icosahedral phase alloys.

for the ionic crystals. In particular, substantial electron density is found in the interstitial region between atoms. Covalent bonds are highly directional, forming tetrahedrally coordinated structures of the diamond lattice type (see Sec. 1.4.2), and very strong, making them hard and difficult to deform. Since the electron distribution is at least partially delocalized in covalent bonding, covalent crystals are not as good insulators as the ionic or inert-gas crystals. All semiconductors (i.e., Ge, Si, GaAs, etc.) are fully, or nearly, covalent.

For some other compounds, the cohesive forces may best be described as mixed ionic and covalent bonding. The bonding in systems such as ZnO, ZnTe, BeO, ZnSe, CdS, and CdTe, e.g., is characterized as $\approx$60%–70% ionic. The first three crystallize into the hexagonal wurtzite structure and the latter three are found in the cubic zinc-blende structure (see Sec. 1.4.2). At the other end, the bonds between atoms in semiconductors such as InAs, GaAs, and GaSb are described as mostly covalent with some ionic character.

2.4 Metallic Bonding

Metallic bonds differ from the covalent bonds of semiconductors in that the valence electrons are shared among all atoms in the crystal and are relatively free to wander through the crystal lattice. This accounts for the good conductivity of metals. The mechanisms which give rise to metallic bonding are complex and must account for the interaction between the ion cores and the electron gas, as well as the interaction between the electrons themselves. Since the electrons are not assigned to any particular ion core, or pair of ions, metallic bonds are nondirectional. Therefore, metals tend to crystallize into the dense, highly symmetric arrangements—the close-packed structures described in Sec. 1.4.1.

3. DEFECTS IN CRYSTALS

The Gibbs free energy of a system at some temperature, T, may be written as $G=H-TS$. At $T=0$ K an ideal crystal, free of defects, is the state of minimum free energy. Defects increase both the enthalpy H and the entropy S of a system. Consequently, the free energy can be a minimum for a crystal with defects at finite temperature. Even for cases in which G is not a minimum, however, defects in crystals may be stabilized for kinetic reasons since their removal usually requires the diffusion of atoms.

All real crystals contain defects. One must keep in mind, though, that in many cases the presence of defects is not undesirable. For instance, pure semiconducting crystals of Si require the intentional addition of certain impurities to become useful solid-state devices. The presence of dislocations at the surface of a crystallite generally enhances the rate of growth at the solid–liquid interface.

3.1 Vacancies, Interstitials, and Impurities

The simplest defect is a *vacancy*. As illustrated in Fig. 26, a vacancy corresponds to an empty site in the structure which is normally occupied in the ideal crystal. Another simple point defect corresponds to an atom situated at an *interstitial position,* not normally occupied in the ideal crystal.

The concentration of isolated defects in thermal equilibrium may be calculated if we know the energy E required for their formation. For N atoms in the crystal, the number n of defects of some particular type is given by

$$n=Ne^{-E/k_BT},$$

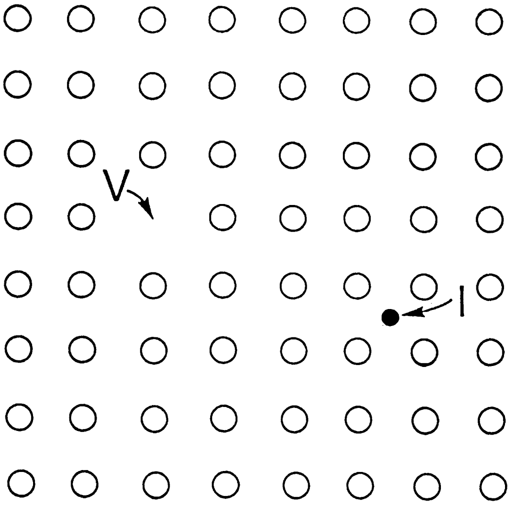

FIG. 26. A schematic illustration of simple defects in a monatomic Bravais lattice. The absence of a circle denotes a vacancy, while the filled circle represents an interstitial atom.

where k_B is the Boltzmann constant and T is the temperature. Similar calculations for clusters of defects are more complicated since the presence of one defect may influence the energy required to form another nearby. There may be other constraints placed on the number and type of simple defects in crystals. For instance, the relative number of positive-, or negative-ion vacancies of interstitials in ionic crystals is regulated by the requirement of charge neutrality.

Chemical impurities can play a crucial role in modifying the physical properties of materials, even in minute concentrations. The electrical properties of semiconductors, e.g., depend largely upon the number of electrons that can be excited across the narrow energy gap separating the valence band from the conduction band. Electrons can be excited across this gap thermally (so that the conductivity of semiconductor increases with temperature), or by an applied electric field. The intentional addition of certain chemical impurities can introduce energy levels within the gap close to either the valence or conduction band, thereby reducing the temperature or applied potential required to populate the conduction band, or alternatively, depopulate the valence band. Chemical impurities may be either *substitutional*, sitting at a site normally occupied by an atom of the pure crystal (e.g., an arsenic impurity at a germanium site in a germanium crystal), or interstitial.

3.2 Chemical Disorder in Alloy Systems

In *ordered alloy* systems, specific sites are assigned to the constituent elements (Fig. 27). For instance, at low temperature, the equiatomic alloy of CuZn (brass) is chemically ordered and may be described as simple cubic with the copper atoms situated at the corners of the cube and Zn atoms at the body center (the CsCl structure). As the temperature of this system is raised above some critical temperature, T_C, there is an *order–disorder transformation* (see ORDER–DISORDER TRANSITIONS) which results in a mixing of the Cu and Zn sites. That is, both the corner and body center sites are occupied by copper or zinc with equal probabilities. The structure is then more correctly described as body-centered cubic. If this alloy is cooled rapidly from some temperature above T_C, the high-temperature phase may be "frozen in," resulting in a metastable disordered phase with some degree of *site-exchange disorder*.

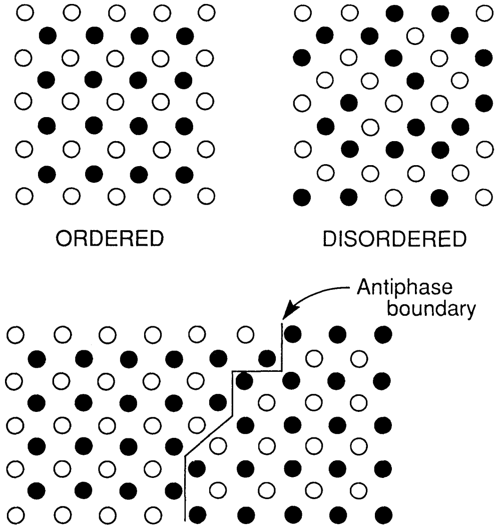

FIG. 27. Ordered (left) and disordered (right) alloys of A (open circles) and B (closed circles) atoms. The full line in the bottom panel represent an antiphase domain boundary between two ordered domains.

Antiphase domains may also occur in chemically ordered alloys. The sites of A and B atoms within any given domain are well-defined, but the A sites in one domain become the B sites in the adjacent domain (Fig. 27).

3.3 Stacking Faults and Line Defects

In contrast to the situation for point defects, extended defects in crystals are metastable since the energy required for their formation is not fully compensated by the increased entropic contribution to the free energy. Therefore, most extended defects may be removed by appropriate heat treatment or *annealing*.

A common defect found in fcc and hcp metals is the *stacking fault*. As described in Sec. 1.4.1, both structures may be viewed as a stacking of close-packed atomic planes. The layer sequence *ABCABCABC...* corresponds to the stacking of planes along the [111] direction of the fcc lattice, while the sequence *ABABABAB...* corresponds to layers stacked along the [001] direction of the hcp lattice. There are, in fact, an infinite number of such possible sequences which correspond to the intermixture of local fcc or hcp order. For instance, the sequence *ABCABABCABCABC...* corresponds to an fcc crystal with a stacking

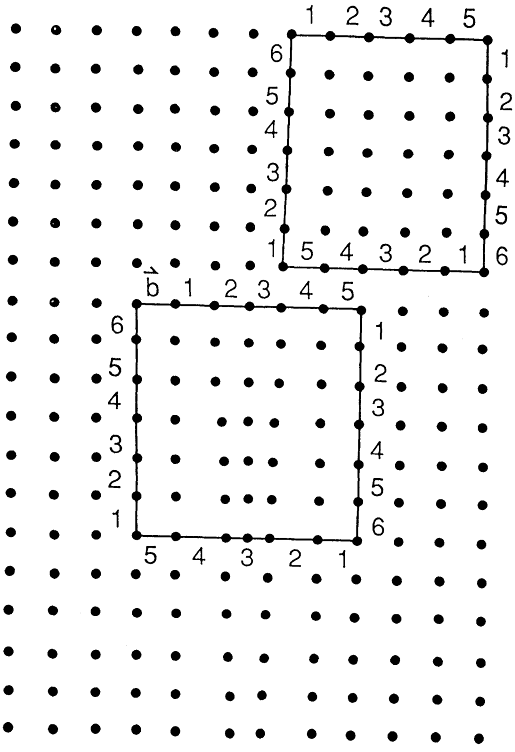

FIG. 28. An edge dislocation. The solid line at the upper right shows a Burgers loop around a defect-free region of the crystal. The solid line in the middle describes the same set of translations around the edge dislocation. The Burgers vector of the dislocation is indicated by the arrow which closes the loop.

fault (a missing *C* layer in the sixth position), which produces a local region of hexagonal order.

These sequences may be used to describe another type of planar defect commonly seen in crystals, termed a *twin*. Consider the sequence of layers *ABCABCACBACBA*..., where the ordering of the layers reverses after the seventh layer. The sequences on either side of this *twin boundary* are mirror images.

Line defects, also termed *dislocations*, have a profound effect upon the mechanical properties of crystals. An *edge dislocation*, depicted in Fig. 28, may be constructed by removing a plane of atoms from one-half of the crystal, leaving the other half unmolested. Alternatively, one may view an edge dislocation as arising from the insertion of a half-plane of atoms. The *dislocation line* runs parallel to the end of the half-plane, out of the plane of Fig. 28. Note that in order to maintain long-range order throughout the entire crystal, the two halves of the crystal must be rejoined so that, near the dislocation line, atoms are crowded together on one side and pushed apart on the other. The presence of edge dislocations greatly reduces the crystal's resistance to shear deformation since the mobility of this defect is quite high. Dislocations may be pinned, however, by impurities.

If we cut partway through a crystal, and then shear the two halves of the cut as shown in Fig. 29, the resulting defect is termed a *screw dislocation*, defined by the dashed line in the figure. At the surface of the crystal, a screw dislocation produces a ledge. Sites immediately adjacent to the ledge on the surface are particularly favorable for the addition of new atoms. Therefore, screw dislocations perform the important function of providing nucleation sites for the next layer of surface atoms during the growth of a crystal from the melt (see Sec. 4). The surface of the crystal grows in a spiral fashion around the dislocation line.

Other extended defects may be formed from combinations of edge and screw dislocations, and may be described by a *Burgers vector*. The Burgers vector is defined by constructing a closed path around the region of the dislocation. In a perfect crystal, a closed loop may be accomplished by a series of translations through the Bravais lattice vectors as shown in Fig. 28. If a dislocation is present within this region, the same sequence of translations will not yield a closed loop. The vector joining the end of the sequence to

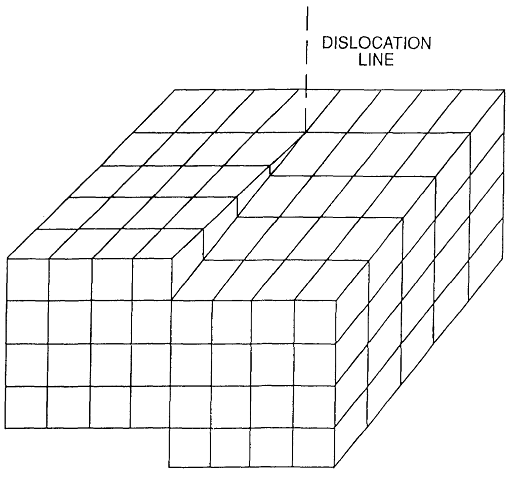

FIG. 29. A screw dislocation. The dislocation line is indicated by the dashed line.

the beginning is the Burgers vector of the dislocation. The Burgers vector is perpendicular (parallel) to the dislocation line characterizing an edge (screw) dislocation.

Low-angle *grain boundaries* are extended defects that separate crystallites, and may be constructed from arrays of edge dislocations. The relative orientation of the two crystallites depends upon the Burgers vector and spacing of the array of dislocations. The presence of grain boundaries often enhances reactions in solids since they provide easy pathways for the diffusion of atoms.

4. CRYSTALLIZATION

In this section, we are primarily concerned with the growth of crystals via a first-order transformation from a liquid phase (see PHASE TRANSITIONS). Below the melting point, the free energy of the well-ordered crystalline phase is lower than that of the liquid, favoring crystallization.

In the next sections, we consider only the case of transformations between a single-phase liquid and a single-phase solid at the same composition. More complex transformations, such as the solidification of eutectic alloys, will not be addressed here.

4.1 Nucleation

There are two distinct stages in the crystallization process: nucleation and growth. For the case of *homogeneous nucleation*, seeds of the new crystalline solid are dispersed uniformly throughout the volume of the melt. The formation of these nuclei is a dynamic process where microscopic regions of crystalline order form and then dissolve through thermal fluctuations. The growth of atomic clusters of the crystalline phase is favored by its lower bulk free energy, but inhibited by the unfavorable positive free energy associated with the surface of the cluster. Since the surface-to-volume ratio decreases with increasing cluster size, clusters larger than some critical radius are stable and can grow; they become the nuclei of the solid phase.

The nucleation process can be enhanced by the presence of interfaces or foreign particles in the liquid. *Heterogeneous nucleation* occurs when the seeds of the solid phase form about these defects, offsetting the positive surface energy of the cluster with the decrease in free energy associated with removing the defect. This process is analogous to "seeding" a cloud with foreign particles to nucleate rain drops.

4.2 Growth

Once nuclei are formed, their growth is controlled by a complex array of processes at the interface between the nuclei and liquid. These include the success of atoms in finding the proper sites on the surface of the crystallite; the probability that atoms will "stick" to this site; and the rate at which excess energy, released in the form of heat once the atom joins the cluster, is removed from the region of the interface.

For some period of time after nucleation, the crystallites grow independently at a rate controlled by the factors mentioned above. As the size of the crystallites increases, neighboring crystallites of different orientations will come in contact. At low enough temperatures (e.g., relatively fast cooling rates), polycrystalline material results. At a high temperature, however, the larger grains can grow at the expense of smaller grains yielding a finite number of single crystals. Certain techniques can be used to promote the growth of large single crystals.

The morphology of grain growth in a liquid medium depends largely upon the rate of cooling. For strong departures from equilibrium conditions, such as large undercooling, grains can be highly irregular or *dendritic*, consisting of tree-like projections into the melt (Fig. 30). These dendrites have a preferred orientation, with their axes along the direction of fastest growth (e.g., the [100] direction in fcc metals).

Under conditions closer to equilibrium, the grain often develops into a well-defined polyhedron closely related to the symmetry of the underlying crystal lattice. *Faceted interfaces* occur along directions of slow growth (Fig.

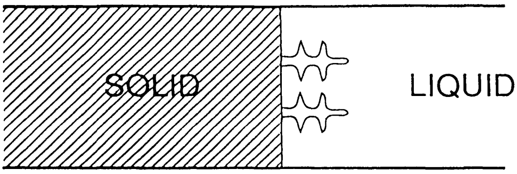

FIG. 30. Dendritic growth at a solid–liquid interface.

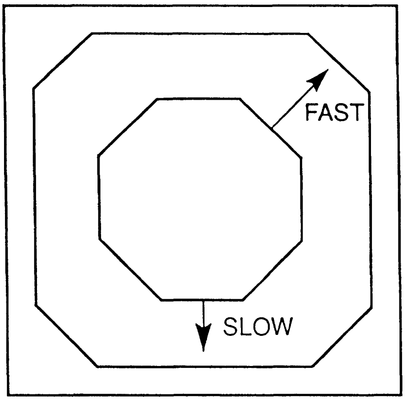

FIG. 31. The development of facets during the equilibrium growth of crystals. Note that the faces perpendicular to the slow growth directions dominate the shape of the crystal at later stages of growth.

31) which generally correspond to directions normal to the most dense planes of the structure. In an fcc metal, e.g., faceting is most likely to occur along the [111] directions.

4.3 The Recrystallization of Glasses

If the rate of undercooling is extremely large, the liquid can solidify into a glassy structure. The recrystallization of glasses often proceeds in a manner analogous to the nucleation and growth process described above. At temperatures far below the melting point, the free energy of the crystalline phase is significantly lower than that of the glass. Crystallization is inhibited, however, by slow kinetics (atoms have to move some distance to effect the transformation). As we heat the system above the glass transition temperature T_g, seeds of the crystalline phase nucleate and grow. Since the difference in free energy between the glass and crystal is so large, the rate of homogeneous nucleation in the glass phase is much larger than that found for liquids; the critical size for nuclei is much smaller. Recrystallized glasses, therefore, are generally much finer-grained than crystals grown from the melt.

5. SUMMARY

In this article, some of the basic features and language of crystalline solids including crystal structure, binding forces, defects, and growth have been introduced. All of these topics have been the subject of intense investigation over the past century which has led to the development of important new technologies such as semiconductor devices. Given the brevity of this article, many interesting and important topics could not be covered in any detail. The reader is referred to the list below for a more comprehensive treatment of the subject.

List of Works Cited

Henry, N. F. M., Lonsdale, K. (1952), *International Tables for X-ray Crystallography*, Birmingham, England: Kynoch Press, Vol. 1.

Kittel, C. (1986), *Introduction to Solid State Physics*, New York: Wiley.

Further Reading

Ashcroft, N. W., Mermin, N. D. (1976), *Solid State Physics*, New York: Holt, Rinehart and Winston.

Wyckoff, R. W. G. (1974), *Crystal Structures*, New York: Interscience.

Jaric, M. J. (Ed.) (1988), *Introduction to Quasicrystals*, New York: Academic.

Verhoeven, J. D. (1975), *Fundamentals of Physical Metallurgy*, New York: Wiley.

CRYSTALLOGRAPHY

W. B. PEARSON AND C. CHIEH, *Department of Chemistry, University of Waterloo, Ontario, Canada*

	Introduction	385
1.	**The Crystalline State**	386
1.1	Unit Cells and Crystal Systems	387
1.2	Lattices	387
1.3	Point-Group Symmetry	388
1.4	Space Groups	389
1.5	Lattice Planes and Crystal Faces	390
1.6	The Reciprocal Lattice	391
2.	**Crystal Structure Determination**	392
2.1	Radiations for Crystallography	392
2.1.1	X-Ray Emission Spectra	392
2.1.2	X-Ray Atomic-Scattering Factors	393
2.1.3	Neutron Atomic-Scattering Factors	393
2.2	X-Ray Crystal Structure Analysis with Single Crystals	394
2.2.1	The Laue Methods	396
2.2.2	Cylindrical-Film Methods	396
2.2.3	Precession Methods	396
2.2.4	Diffractometers	397
2.3	Macromolecule Crystallography	397
2.4	X-Ray Analysis of Polycrystalline Samples	398
2.5	Other Diffraction Methods	398
2.5.1	Neutron Diffraction	398
2.5.2	Electron Diffraction	399
3.	**The Results of Crystal Structure Determinations**	399
3.1	Metals	399
3.2	Insulators/Semiconductors	400
3.3	Molecular Crystals	401
3.4	Assessment and Collation of Crystal Structure Data	403
4.	**Crystallography in Industry and Medicine**	405
5.	**Epilogue**	406
	Glossary	407
	Works Cited	407
	Further Reading	408

INTRODUCTION

Crystallography is the science of crystals and crystalline materials, which include most terrestrial and industrial solids. The science includes the study of crystal structure on an atomic scale and also the shapes, forms, defects, growth, transitions, and properties of crystals. Chemical analysis to determine the elemental composition of a substance and the determination of its crystal structure are the important experiments of the solid state. Principally, x-ray- or neutron-diffraction experiments are used to determine crystal structures, that is to say, the crystal symmetry, the relative positions, and the elemental nature and arrangement of the atoms in the structure. X-ray diffraction by a crystal was first demonstrated in 1912, following the discovery of x rays by Röntgen in 1896. Nevertheless, using earlier mathematical developments and measurements of the angles between the faces of crystals, all of the rudiments of three-dimensional crystal symmetry had been worked out by this time and were at hand to be applied to the new atomic-scale crystallography. Prominent among these early developments were observations on crystal form by René Haüy (1784), Wollaston's (1809) construction of the optical goniometer, Hessel's (1830) demonstration of the 32 symmetry classes, Bravais's (1848) demonstration of the 14 lattices, Sohncke's (1867) discovery of new symmetry elements, and finally, the derivation of the 230 space groups independently by Schoenflies (1891) and Fedorov (1891).

Thus, when in 1912 von Laue, Friedrich, and Knipping demonstrated the diffraction of x rays by crystals, the door was opened to the possibility of gaining exact information on the

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

arrangement of atoms within a crystal. This discovery was rapidly followed by the Braggs' (1913) interpretations of their experiments on rocksalt and other crystals; however, the wavelengths of characteristic x radiations still had to be determined. So began a science of the most fundamental significance. The history of crystallography is complicated and enlightening, and a short description thereof can be found in *Fifty Years of X-Ray Diffraction* by Ewald (1962). Over the years the subject became so important that in 1947–48 the International Union of Crystallography (I.U.Cr.) was created to be concerned with all viable aspects of crystallography through its Commissions, its triennial General Assembly and convention, and its publishing activities.

Much of the realm of crystallography is covered in specific articles elsewhere in the *Encyclopedia,* and some of these articles are cross-referenced in the text. Therefore this article introduces crystallography to the general reader, providing an overview, yet without sacrificing physical and mathematical precision to simplicity. With very many known crystal structures at hand, we can only illustrate this article with information on a few simple types, upon the basic structural principles of which many more complex types are formed.

Finally, it is observed that the rate at which the current technological and medical revolutions are proceeding results in no small degree from the information provided by crystallographic studies.

1. THE CRYSTALLINE STATE

A crystal is characterized by a long-range order. That is to say, the atomic arrangement is periodic in a three-dimensional space throughout the crystal (see CRYSTALLINE STATE). Other solid material may be amorphous, so that there is only short-range order that extends merely for a few shells of atoms in the surroundings of any given atom (see AMORPHOUS MATERIALS, STRUCTURE OF).

Any solid that one views is likely to be made up of crystals and/or amorphous material, and of crystals of more than one type and chemical composition. A compound may exist in the form of gas, liquid, or several forms of solids. Each of these has been conventionally called a *phase,* which is a homogeneous portion separated by a boundary from other portions in a multicomponent system. Thus, the term *crystal* here is the solid of a *single phase,* and the term *crystal structure* refers to the geometry, symmetry, and the arrangement of atoms or molecules of the phase. The solid of a single phase generally is composed of many independent crystals, variously oriented with respect to their symmetry axes, the whole being joined solidly together. Such a state is referred to as *polycrystalline* (see GRAIN BOUNDARIES). In general, solids may also contain crystals of several compounds with quite different chemical compositions and crystal structures or several types of crystals of the same compound. When an element or chemical compound, e.g., SiO_2, exists in several different structural arrangements, this is known as *polymorphism* (see PHASE EQUILIBRIA and PHASE TRANSITIONS). A phase is stable at some temperatures and pressures, and a metastable phase stays indefinitely at temperatures and pressures in which it is energetically unfavorable. Phase transitions take place by changing either the temperature or the pressure or both (see HIGH-PRESSURE EFFECTS IN SOLIDS). For example, diamond, the metastable phase at room temperature and pressure, must be synthesized at high temperature and pressure.

When one sets out to determine the atomic arrangement of a crystal, one not only seeks a sample of the solid that is pure (i.e., that does not also contain crystals of other compounds), but one that is a single crystal, wherein one crystal orientation extends throughout the solid. This is a practical requirement (see CRYSTAL GROWTH), as the diffraction patterns from polycrystalline samples confuse the data analysis since crystal orientation is random, thus limiting the useful information obtainable.

A crystalline arrangement of atoms is described by the laws of symmetry, which include the translation operations. Theoretical consideration of crystal symmetry verified 230 *space groups* (see Sec. 1.4), to one of which the crystal structure must belong. Although the number of space groups is limited, the number of different crystalline arrangements is essentially unbounded. This is because these laws permit continuously variable atom positions to be assumed, although they are fixed in any particular crystal structure.

Furthermore, the positions can be occupied by any combinations of about 100 natural and manmade elements.

A periodic crystalline arrangement gives rise to well-developed crystal faces of single crystals that are independently grown. These faces display the symmetry elements inherent in the crystal type. Sometimes such crystals exhibit a phenomenon known as *twinning,* where, for example, two different orientations are related by a *mirror plane,* whereon they unite coherently.

A crystal may contain organized coherent defects such as twinning, or a given crystal structure may be composed, e.g., of layers of two simple structural types that join coherently and repeat in regular sequence (see IMPERFECTIONS IN CRYSTALS). It may also exhibit irregular or random defects; indeed a polycrystalline material exhibits such defects along the boundaries of the crystallites of different orientations. In other crystals where the atoms occupy ordered sites, the ordering may "get out of step," e.g., be displaced in one region of the crystal compared to another, either on a random or regular basis. Regular ordering sequences or regular atom-displacement sequences may cause a *superstructure,* where a new unit cell that recognizes the regular ordering or displacements relative to the basic unit cell (see Sec. 1.1) may require multiplication of the edges by 2, 3, etc., along the axes **a**, **b**, or **c** concerned. Thus the ordered tetragonal Al_3Ti is an F_{112} superstructure (multiplication $2\times\mathbf{c}$) of the cubic copper structure, *F* (see Sec. 1.2 and HETEROSTRUCTURES AND SUPERLATTICES, METALLIC; HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR).

Liquid crystals are ordered fluids. The ordering in one, two, and three directions gives rise to phases called *nematic, smectic,* and *cholesteric* liquid crystals, but they are not considered solid states (see LIQUID CRYSTALS). On the other hand, the recently discovered *quasicrystals* are solids, but their symmetry and geometry do not follow those observed for crystals. Nevertheless, their existence was confirmed by diffractions, which proclaimed the presence of noncrystallographic fivefold symmetry (see Sec. 1.3) axes in these solids (see QUASICRYSTALS).

1.1 Unit Cells and Crystal Systems

Since a single crystal is an aggregate of millions of atoms periodically arranged in space, systematic methods of description have been devised. The first thereof is the *unit cell,* which is (generally) the smallest parallelepiped that stacked repetitively in three dimensions generates the structure of the crystal, regardless of the atomic content within the parallelepiped. This unit cell is defined by three noncoplanar vectors of magnitudes a, b, and c and three angles between them, α, β, and γ (Fig. 1). Such unit cells must belong to one of seven *crystal classes* that are defined purely by symmetry, although they also have metrical constraints as indicated in Table 1. Thus even though a unit cell may have $a=b=c$ and $\alpha=\beta=\lambda=90°$, if the atomic arrangement does not exhibit at least the minimum symmetry of a cubic system (*point group* 23, see Sec. 1.3), the unit cell must belong to a class lower than cubic. Further remarks on the choice of unit cells are made in the next section.

1.2 Lattices

The concept of a *lattice* is essential to the study of crystal structure. A lattice is an infinite periodic array of points at the termini of the vectors $\mathbf{L}_{uvw}=u\mathbf{a}+v\mathbf{b}+w\mathbf{c}$, where **a**,**b**,**c** are the three noncoplanar vectors defining a primitive unit cell and u,v,w are integers. Thus the integers $[uvw]$ are used to define a lattice direction (note the use of [ ] to define directions, whereas ( ) are used to define lattice planes (see Sec. 1.5)).

There are seven primitive lattices (one lattice point per cell) corresponding to the seven crystal classes. However, because of

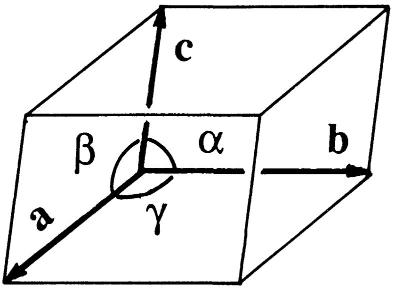

FIG. 1. Quantities a, b, c, α, β, and γ used in defining a unit cell. A right-handed coordinate system is always used.

Table 1. Axial and angular relationships, characteristic symmetry, and the 14 Bravais lattice types of the seven crystal systems.

System	Axial and angular relationships	Min. symmetry and (point group)[a]	Lattice type and symbol[b]
Triclinic	$a \neq b \neq c$; $\alpha \neq \beta \neq \gamma \neq 90°$	$1(\bar{1})$	P
Monoclinic	$a \neq b \neq c$; $\alpha = \gamma = 90° \neq \beta$	$2(2/m)$	P,C[c]
Orthorhombic	$a \neq b \neq c$; $\alpha = \beta = \gamma = 90°$	$222\ (mmm)$	P,I,F,C[d]
Tetragonal	$a = b \neq c$; $\alpha = \beta = \gamma = 90°$	$4(4/mmm)$	P,I
Cubic	$a = b = c$; $\alpha = \beta = \gamma = 90°$	23[e] $(m3m)$	P,I,F
Trigonal	$a = b = c$; $\alpha = \beta = \gamma$	$3(\bar{3})$	R
Hexagonal	$a = b \neq c$; $\alpha = \beta = 90°$, $\gamma = 120°$	$6(6/mmm)$	P

---

[a] The point group is the maximum symmetry.

[b] Lattice symbols: *P*, primitive; *C*, *ab*-face centered; *I*, body centered; *F*, all-face centered; *R*, rhombohedral.

[c] Lattice type *C* is for *b*-axis unique or *A* for *c*-axis unique.

[d] Depending on the choice of axes, lattice type *A* (*bc*-face centered) or *B* (*ca*-face centered) may also be assigned.

[e] The twofold and threefold axes for cubic are along the face and body diagonals, respectively. The angle of intersection is 35.26°.

symmetries permitted in some of these lattices, it is more convenient for structural description to choose the face-centered lattices *A*, *B*, or *C*, the body-centered lattice *I* (all with two lattice points per cell), and the all-face-centered lattice *F* (four lattice points per cell), rather than the primitive lattices. This was already recognized by Bravais (1848), who saw that there were 14 such lattices; they now bear his name (Table 1 and Fig. 2). Although some crystal structures such as that of copper do have atoms only at lattice points, so that it is correct to refer to either the *Cu lattice* or the *Cu structure*, in general this is not so, and it is incorrect to use *crystal lattice* as a synonym for *crystal structure.*

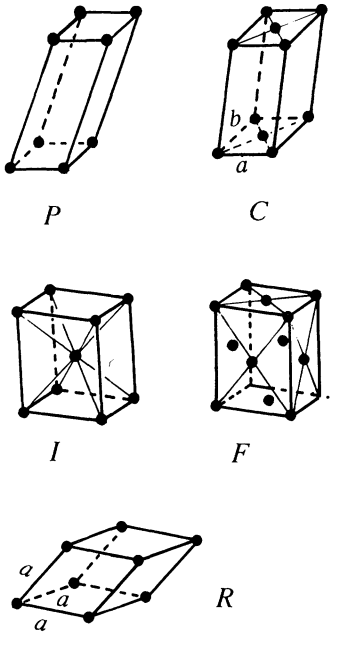

FIG. 2. Primitive *P*, *ab*-face-centered *C*, body-centered *I*, all-face-centered *F*, and rhombohedral *R* (trigonal) lattices (see Table 1).

1.3 Point-Group Symmetry

The manifestation of crystal symmetry is present both in the exterior forms and in the atomic arrangement within the crystals. This is how all of the elements of the crystal symmetry could be determined before any of the internal arrangements of atoms therein were discovered. Thus early morphological and mathematical studies led to the correct conclusion that there are 32 symmetry types that are called *point groups.* The morphology of crystals led to the recognition of *n*-fold rotation axes (where $n = 1, 2, 3, 4$, or 6, but not 5), mirror planes *m*, and a possible center of inversion *i* or $\bar{1}$. Combinations of these symmetry elements, which leave at least one point fixed, give rise to the 32 point groups. It is worth noting that some texts include *axes of rotary inversion* $\bar{n}$ or *improper rotation axes* as symmetry elements, but they result from combining rotation axes with a center of inversion or with a mirror plane.

Notations for describing point groups make use of some characteristic symmetry operation in the group, but they do not give all symmetry elements. Sometimes symmetry elements n or m in three directions are involved. When a mirror plane perpendicular to the n-fold axis is a major feature, it is represented by n/m. A twofold axis, if present, is often perpendicular to another major n-fold axis, and mirror planes can contain the major axis; these symmetry elements are placed side by side. For example, the symmetry point group of a cube with equal parts cut off from the corners is 432, representing the fourfold, threefold, and twofold rotation axes as major features. The pyramidal arrangement of H and N atoms in NH_3 gives rise to a threefold axis and mirror planes, its point group being $3m$. The planar benzene molecule C_6H_6 has a mirror perpendicular to the sixfold axis and two other types of mirror planes; its symmetry belongs to the point group $6/mmm$. Whereas these symbols of the Hermann–Mauguin notation are popular with crystallographers, the Schoenflies symbols are used more often by spectroscopists. The Schoenflies notations of the above point groups are $D_{6h}=6/mmm$, $C_{3v}=3m$, and $O=432$. Schoenflies symbols stress the nature of the group, being dihedral (D), cyclic (C), or octahedral (O). Although point groups were of interest to mineralogists at some time in the past, they do not play a key function in crystal structure descriptions today; however, they are still significant in crystallography as they provide an easy classification of crystals. A regular tetrahedron belongs to the point group 23 (Schoenflies symbol T), where the twofold axes go through the midpoints of opposite edges and the threefold axes go through the vertices and the geometric center. The relationship between a regular tetrahedron and a cube illustrates that 23 is a point group of the cubic crystal system.

1.4 Space Groups

Crystallographic lattices are generated from a point by translations, which are thus a symmetry operation. Hence it is apparent that a complete description of crystal symmetries must include translations as well as those present in point groups. Again the study is executed by the techniques of group theory. Combination of crystallographic point-group symmetry operations with translations gives new symmetry elements such as *screw axes* (2_1, 3_1, 4_1, and 6_1) and *glide planes* (axial glides a,b,c, diagonal glide n, and diamond glide d). The theory showed that there are 230, and only 230, distinct combinations of symmetry elements in three-dimensional space, which are known as the 230 *space groups.* These are each described in the Hermann–Mauguin notation to show essential symmetry elements, e.g., $P6_3/mmc$. The *International Tables for Crystallography* (Hahn, 1983) of the International Union of Crystallography, the *vade mecum* of crystallographers, give essential data for all space groups such as the possible sets of symmetry-equivalent positions therein, their multiplicity (number of equivalent positions), Wyckoff notation (see the Glossary), and point symmetry, and their coordinates. The space group to which a crystal belongs is an essential piece of information, without which the description cannot be completed from the published set of crystal data.

A space-group symbol consists of two parts: a letter depicting the lattice type and some essential symmetry elements. The triclinic system has two primitive space groups, $P1$ and $P\bar{1}$. Crystal structures belonging to the latter have a center of inversion, i.e., if there is an atom located at a point x,y,z, there is also one at $-x,-y,-z$; thus only half of the atomic coordinates in the unit cell needs to be given, as the other half is implicit or equivalent by symmetry. By way of illustration, the 13 space groups of the monoclinic system are listed in Table 2 with a remark about the symmetry features. Space groups with the same symmetry feature, but of different lattice type, are given on the same line. Table 2 also illustrates the relationship between space-group and point-group symmetry. The translation component of the former is absent in the latter.

Table 2. The 13 space groups of the monoclinic system.

Space group	Symmetry feature	Point group
$P2$, $C2$	Twofold axis	2
$P2_1$	2_1 screw axis	2
Pm, Cm	$m \perp$ the unique axis[a]	m
Pc, Cc	c glide	m
$P2/m$, $C2/m$	$m \perp 2$	$2/m$
$P2_1/m$	$m \perp 2_1$	$2/m$
$P2/c$, $C2/c$	c glide $\perp 2$	$2/m$
$P2_1/c$	c glide $\perp 2_1$	$2/m$

---

[a] $\perp$ means perpendicular to.

Therefore screw axes in space groups become simple n-fold rotation axes in point groups. Similarly, glide planes of space groups become mirror planes of point groups.

A symmetry element or operation will transform a set of coordinates x,y,z to a new location. For example, a twofold axis, if parallel to the b axis, gives $-x,y,-z$, and a 2_1 screw gives $-x,\frac{1}{2}+y,-z$. A c glide gives a new location at $x,-y,\frac{1}{2}+z$. Symmetry operations in each space group can be specified by a set of coordinates, which generate all atomic positions in a unit cell. For the space group of common occurrence, $P\,2_1/c$, the set of general coordinates (b axis unique) is: x,y,z; $-x,\frac{1}{2}+y,\frac{1}{2}-z$; $-x,-y,-z$; $x,\frac{1}{2}-y,\frac{1}{2}+z$. A diagram depicting these points is given in Fig. 3. However, if x, y, or z have values such as 0 or $\frac{1}{2}$, the set of four points degenerates into a set of two points only. These are called *special positions*. The special values give rise to several distinct sets, each with its own number of positions or multiplicity, and site symmetry. The sets are denoted by Wyckoff notations. A complete list of all Wyckoff sites for the space group $P\,2_1/c$ mentioned above is given in Table 3. Obviously, the origin for the coordinate system given in this example is at the center of inversion. The choice of origin and setting complicates matters for the comparison and classification of structure types.

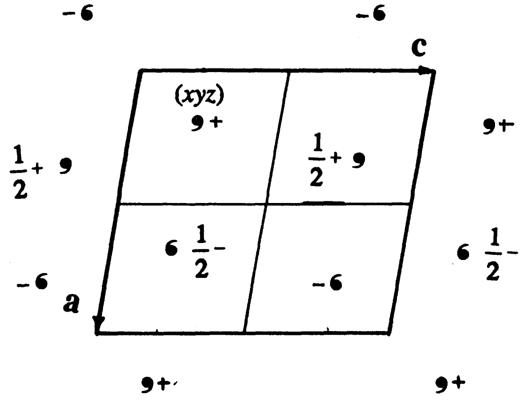

FIG. 3. Symmetry-equivalent images of a pattern represented by a comma at x,y,z (fractional coordinates with respect to a,b,c) in the space group $P2_1/c$. The + signs beside the comma indicate a point below the page. The coordinates $\frac{1}{2}\pm y$ of the points are represented by $\frac{1}{2}\pm$. Depending on the value for y, these points can be above or below the page. Inverted commas are due to mirror planes or centers of inversion.

1.5 Lattice Planes and Crystal Faces

For use in crystal diffraction and for morphological studies, it is necessary to consider a set of parallel planes so constructed that every lattice point lies on some member of it. These are called lattice planes, rational planes, or simply *planes*. When such planes cut an axis of the unit cell, they always divide it into an integer number of equal parts. If we start with some lattice point on a plane as the origin and proceed from it along the axes to the first member of the same set of planes, the equation for the first member plane can be expressed in terms of the intercepts. If the intercepts along the three axes are a/h, b/k, and c/l, respectively, the (h,k,l) are referred to as the Miller indices of the set of planes. The commas may be omitted if the integers or h, k, and l lead to no ambiguity. For example, referred to a lattice point as the origin, the first member of planes (321) intercepts the three axes at $a/3$, $b/2$, and c, respectively; see Fig. 4. The set of planes (100) is parallel to the bc plane. It should be noted, however, that the distance d_{100} between (100) planes is not a, except in orthorhombic, tetragonal, and cubic systems, which have orthogonal axes. The interplanar distance d_{hkl} for the general triclinic system is rather complicated, but for structures with orthogonal axes it is easily recognized.

Miller indices may be used in many ways: to designate a set of lattice planes, a particular member of the set, or a crystal face parallel to the set. The plane (100) intersects at unit length on the axis [100], which represents a lattice point and a direction (see Sec. 1.2). The direction [100] is not perpendicular to the plane (100) except in crystal systems with orthogonal axes, but [100]* referred to a direction in reciprocal space is perpendicular to (100).

The Miller indices are sometimes used to represent a *form* $\{hkl\}$ which represents a set of symmetry-equivalent planes. For example, $\{100\}$ includes (100), $(\bar{1}00)$, (010), $(0\bar{1}0)$, (001), and $(00\bar{1})$, which generate a cube for a cubic crystal. For a triclinic crystal, $\{100\}$ represents only (100) and its symmetry-equivalent plane $(\bar{1}00)$; $\{010\}$ represents (010) and $(0\bar{1}0)$, and $\{001\}$ represents (001) and $(00\bar{1})$. The faces or planes represented by the form $\{hkl\}$ depend on the crystal system.

Table 3. Multiplicity, Wyckoff notation, site symmetry, and coordinates of symmetry-equivalent sites in the space group $P\,2_1/c$.

Multiplicity	Wyckoff notation	Site symmetry	Coordinates of equivalent points
4	*e*	1	$x,y,z;\ \bar{x},\bar{y},\bar{z};\ \bar{x},\frac{1}{2}+y,\frac{1}{2}-z;\ x,\frac{1}{2}-y,\frac{1}{2}+z$
2	*d*	$\bar{1}$	$\frac{1}{2},0,\frac{1}{2};\ \frac{1}{2},\frac{1}{2},0$
2	*c*	$\bar{1}$	$0,0,\frac{1}{2};\ 0,\frac{1}{2},0$
2	*b*	$\bar{1}$	$\frac{1}{2},0,0;\ \frac{1}{2},\frac{1}{2},\frac{1}{2}$
2	*a*	$\bar{1}$	$0,0,0;\ 0,\frac{1}{2},\frac{1}{2}$

All ($hk0$) planes are parallel to **c** or [001], and all ($hk0$) faces have one direction [001] in common. If the crystal is elongated in [001], these faces make it appear as a prism. Thus these are called *zonal planes* or *zonal faces*. The *zones* formed by these faces are convenient in the description of morphology and diffraction patterns. Diffracted beams from zonal planes form a cone with the apex at the crystal.

1.6 The Reciprocal Lattice

A crystal is an identity in a three-dimensional space, and its repeating lattice is called a *direct lattice*, for which a *reciprocal lattice* can be defined either for the convenience of diffraction studies or for other applications. A reciprocal-lattice point is defined to be lying on the normal of a crystal plane (hkl) at a distance K/d_{hkl}, K being a constant. The three reciprocal-lattice vectors **a***, **b***, and **c*** are normals to the planes (100), (010), and (001) with magnitudes of K/d_{100}, K/d_{010}, and K/d_{001}, respectively. Note that since $2/d_{100} = 1/d_{200}$, the reciprocal point corresponding to (200) is twice the distance from the origin as that of (100) or **a***. By the same token, other points in the reciprocal space can be deduced, generating the reciprocal lattice using the basic set of vectors **a***, **b***, and **c***.

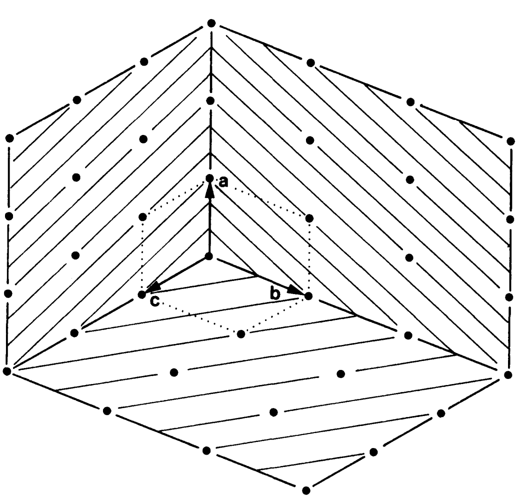

FIG. 4. A set of planes with Miller indices (321) showing the intercepts on axes *a*, *b*, and *c*. Every lattice point lies on some member of the set. The diagram includes three unit cells (one outlined by dashed lines) in each of the three directions.

When applied to x-ray diffraction studies, the constant K can be taken as λ (wavelength of the x rays used); in this case **a*** ($=\lambda/d_{100}$), **b***, and **c*** are unitless pure vectors. The convenience for such a choice has a direct connection with the Bragg equation written in the form $2\sin\theta=\lambda/d_{hkl}$ (see Sec. 2.2). Alternatively, K may be taken as 1, and the reciprocal-lattice space has a unit of reciprocal length. Using vector notation, the relationships between reciprocal and direct lattice vectors are

$$\mathbf{a}^*=\frac{\mathbf{b}\times\mathbf{c}}{V},\quad \mathbf{b}^*=\frac{\mathbf{c}\times\mathbf{a}}{V},\quad \mathbf{c}^*=\frac{\mathbf{a}\times\mathbf{b}}{V},$$

where V is the volume of the unit cell in direct space, and it has the expression

$$V=abc(1+2\cos\alpha\cos\beta\cos\gamma-\cos^2\alpha-\cos^2\beta-\cos^2\gamma)^{1/2}.$$

The reciprocal-lattice angles α^*, β^*, and γ^* bear the following relationships:

$$\cos\alpha^*=\frac{\cos\beta\cos\gamma-\cos\alpha}{\sin\beta\sin\gamma},$$

$$\cos\beta^* = \frac{\cos\gamma\,\cos\alpha - \cos\beta}{\sin\gamma\,\sin\alpha},$$

$$\cos\gamma^* = \frac{\cos\alpha\,\cos\beta - \cos\gamma}{\sin\alpha\,\sin\beta}.$$

Of course, the above relationships are simple for orthogonal systems, in which $\alpha = \beta = \gamma = 90°$.

2. CRYSTAL STRUCTURE DETERMINATION

Although determination of chemical composition and crystal structure are the first two steps in the characterization of a substance, they are but a component in the investigation of its chemical, physical, and other properties. Crystal structure determination provides a three-dimensional map of the atoms in the solid. Thus it gives information on the crystal symmetries involved in the atomic arrangement, on the dimensions and crystal class of the repeat volume or unit cell, and on what atoms occupy which crystal sites fully, partially, and in ordered or random fashion, as well as giving the values of the atomic coordinates. Thus information is obtained about the surroundings of each atom in the structure and the distances and angles between it and its neighboring atoms.

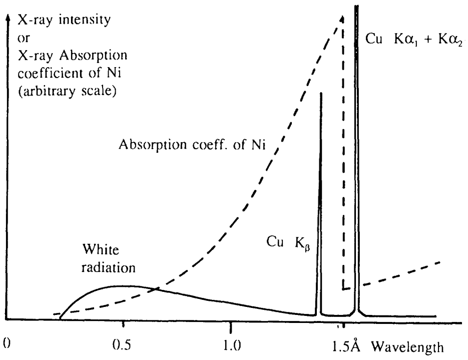

FIG. 5. A typical x-ray spectrum generated by bombarding a copper target with high-energy electrons; also shown are the absorption coefficients of nickel. The sharp drop in absorption coefficients at wavelength between *Kβ* and *Kα* is called the absorption edge.

2.1 Radiations for Crystallography

X-ray, neutron, and electron beams are the major radiations used for crystal structure determination. Each radiation has a unique characteristic and diffraction mechanism.

2.1.1 X-Ray Emission Spectra. Figure 5 shows a typical low-wavelength x-ray intensity spectrum, here generated by 50-kV electrons from a copper target. It is characterized by relatively low-intensity *white* or continuous radiation above a certain wavelength that is made use of in Laue photographs (see Sec. 2.2.1) and by the relatively intense $K\beta$ and $K\alpha_1$ and $K\alpha_2$ peaks at specific wavelengths, with intensities $K\alpha_1 > K\alpha_2 \gg K\beta$. Wavelength separation between $K\alpha_1$ and $K\alpha_2$ is small, and both are used for single-crystal and powder-diffraction photographs and diffractometer data. To simplify the data, the $K\beta$ radiation can be removed by placing a thin nickel foil in the path of the entrant beam, since it strongly absorbs $K\beta$ but not $K\alpha_1$ and $K\alpha_2$. The sharp difference in absorption coefficients within a short range of wavelength is called the ***absorption edge.*** The absorption edge of nickel falls between $K\beta$ and $K\alpha$ of copper x rays, i.e., nickel strongly absorbs $K\beta$ but absorbs little $K\alpha$ (Fig. 5). If the $K\alpha_1$ and $K\alpha_2$ frequencies have to be separated, it requires very careful setting of a *bent-crystal* monochromator, but generally this refinement is considered unnecessary. The principles outlined here are generally applied appropriately to x radiation from molybdenum or other metal targets. Of course, each target metal has characteristic radiations of some unique wavelengths with high intensities and corresponding filter material. The intensity of the x radiation is a function of the accelerating potential and intensity of the electron-generating beam. The potential and intensity are limited by the probability of "burning through" the water-cooled metal target. Stronger x-ray beams can be generated by a rotating-target generator, and very much more intense x-ray sources have now become available at specific *synchrotron radiation* (SR) locations. These are pulsed (intermittent) x-ray beams of continuous wavelength generated when electrons are accelerated to an energy of the order of 10^9 eV or 1 GeV (see SYNCHROTRON RADIATION).

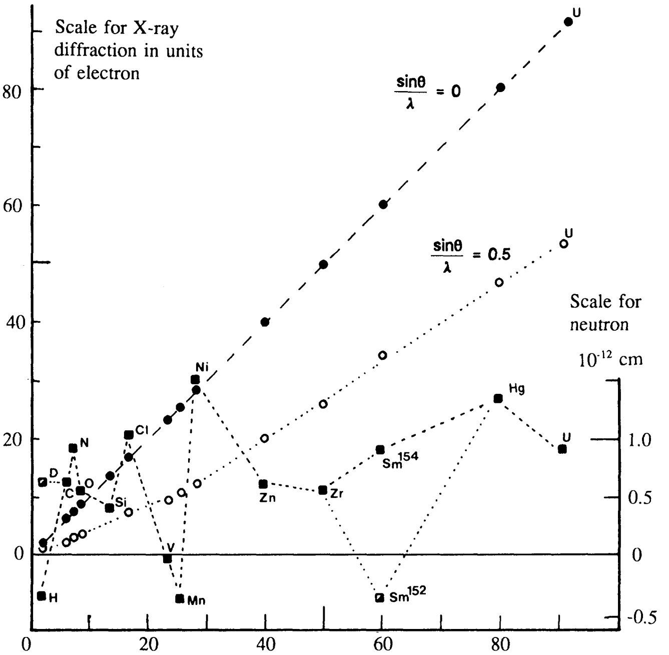

FIG. 6. Comparison of atomic-scattering factors for x rays and for neutrons. Two sets of values are given for x rays, filled circles for $\sin\theta/\lambda = 0$ and open circles for $\sin\theta/\lambda = 0.5$. Corresponding neutron data of the most abundant isotope are represented by filled squares. Isotopes of H and Sm are given for contrasting their difference. Units for the factors are electrons for x rays, and 10^{-14} m for neutrons.

2.1.2 X-Ray Atomic-Scattering Factors. X rays are scattered by the shells of electrons surrounding the atomic nuclei; the amplitudes of x rays scattered by an element, defined as the *atomic-scattering factors*, are positive and increase monotonically with the number of electrons or atomic number (Fig. 6). This has several important consequences in structural analysis.

1. *Location of atoms.* Since for hydrogen, particularly, the scattering factor is very small, very little electron density will appear in Fourier syntheses (see Sec. 2.2) at the locations of the hydrogen atoms, especially if the substance contains other atoms of much greater scattering power. This is why locations for hydrogen atoms cannot be determined with confidence.
2. *Assignments of elements to crystal sites.* When the atoms in a structure have scattering powers that differ little from the average value, it is difficult to assign elements to specific crystal sites. Furthermore, superstructure reflections that might be expected due to atomic ordering will be very weak and probably not detectable. In special cases this situation may be improved by judicious choice of x radiation and making use of the anomalous scattering factor, which is pronounced in the vicinity of an element's absorption edge.
3. *Phase-angle determination.* If a heavy atom is introduced in organic compounds that contain elements H, C, N, and O, its dominant atomic-scattering factor can be used to assist phase-angle determination (see Sec. 2.2).

Finally, since x rays are scattered by the electrons surrounding an atomic nucleus, interpretation of x-ray-diffraction experiments gives information not on the locations of the atomic nuclei but on the electron densities surrounding the nuclear sites.

2.1.3 Neutron Atomic-Scattering Factors. Neutrons, discovered in 1935, are particles emitted from the nuclei of atoms in nuclear reactions such as the fission of heavy atoms in reactors. The wave properties of neutrons make them suitable for diffraction studies. First, whereas x rays are scattered by electrons, neutrons are scattered by the atomic nuclei, whose radii are about $\frac{1}{100\,000}$ of those of atoms. Furthermore, the combined effect of nuclear scattering and magnetic scattering due to intrinsic spin and magnetic moment of the neutrons and of atomic nuclei gives rise to a phase-angle shift of the diffracted neutrons.

An out-of-phase (angle) shift results in a negative neutron-scattering factor. Therefore neutron atomic-scattering factors can have positive or negative values, or indeed have a value approximating zero (Fig. 6). They may even differ appreciably between atomic isotopes, as between hydrogen and deuterium. Indeed, with respect to other elements, the hydrogen and deuterium neutron-scattering factors are relatively much larger than their x-ray-scattering factors, which generally makes neutron-diffraction experiments more effective than x-ray experiments for locating the positions of these atoms in structures. Second, the particular values of neutron-scattering factors can be used with judicious advantage in structure determinations. For example, in the case of an ordered Ni-V alloy, neutron-diffraction experiments can well be combined with x-ray-diffraction experiments to locate the sites of the atoms, since the neutron atomic-scattering factor for vanadium is approximately 0. Finally, since neutrons have intrinsic spin and magnetic moment, diffraction experiments can be used to gain information on atomic moments and their orientations for magnetic atoms in structures by separating from the observed scattering that which is normally expected, thus revealing the *magnetic* scattering, which then may be interpreted.

2.2 X-Ray Crystal Structure Analysis with Single Crystals

W. L. Bragg's interpretation of the x-ray-diffraction patterns from NaCl, KCl, KBr, KI, and ZnS laid the foundation for x-ray structure analysis. A simple illustration of the principle of x-ray diffraction is the Bragg equation: $2d_{hkl}\sin\theta=\lambda$, where d_{hkl} is the interplanar spacing for (hkl), θ the angle of inclination of the x-ray beam to the crystal plane, and λ the wavelength of radiation used (Fig. 7). Thus a constructive beam results when both the wavelength and orientation of the crystal with respect to the beam are proper. Note that the diffracted x-ray beam or *reflection* [of plane (hkl)] deviates by an angle of 2θ from the incident beam (see X-RAY DIFFRACTION).

Bragg's law gives the condition for a reflection to develop, while the intensity I_{hkl} of the diffracted x-ray beam is related to the arrangement of atoms in the crystal. In short, I_{hkl} is proportional to the square of a quantity known as the *structure factor* F_{hkl}, which can be expressed as

$$F_{hkl}=\sum_{j=1}^{j=n} f_j \exp[2\pi i(hx_j+ky_j+lz_j)],$$

where $i=\sqrt{(-1)}$; x_j,y_j,z_j are the coordinates of the atom j expressed as fractions of the unit-cell lengths; f_j is the atomic scattering factor; and the summation is taken over all atoms (number n) in the unit cell. Generally, the higher the atomic number of the atom, the larger the f_j value, which decreases as θ increases. When a reflection develops, the path difference of x rays scattered from neighboring planes (hkl) is one wavelength, or a complete cycle of 2π (Fig. 7). The atom j with atomic scattering factor f_j located away from the origin at x_j,y_j,z_j will have a phase-angle difference of $2\pi(hx_j+ky_j+lz_j)$ relative to that from the origin. X-ray diffraction is the superposition of waves from all atoms, and thus the sum is the structure factor, which when squared is proportional to the intensity of the reflection. Thus structure determination involves the investigation of diffraction patterns, measurements of the angles 2θ and intensities of the reflections, and comparisons of the structure factors calculated for the proposed structural arrangement with those derived from experiment. However, to derive the structure from x-ray measurements is not

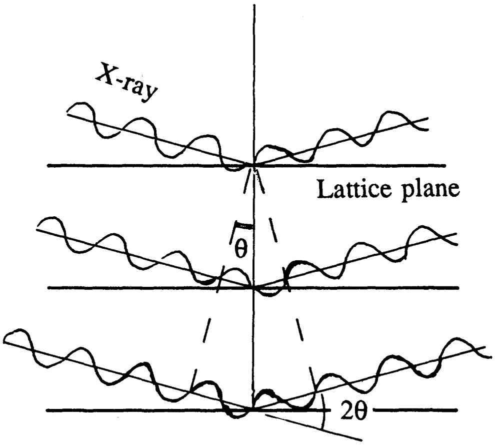

FIG. 7. A simple illustration of the principle of x-ray diffraction, showing the path-length difference of one wavelength for reflected x rays from adjacent lattice planes.

a trivial matter. If the structure factors are known, the electron density $\rho(x,y,z)$, the peaks of which correspond to the positions of atoms, can be calculated from the Fourier series (see FOURIER AND OTHER MATHEMATICAL TRANSFORMS)

$$\rho(x,y,z)=\frac{1}{V}\sum_{-h}^{h}\sum_{-k}^{k}\sum_{-l}^{l}F_{hkl}\exp[2\pi i(hx+ky+lz)],$$

where x,y,z is a point in the unit cell. Unfortunately, the diffraction experiments only provide the magnitudes $|F_{hkl}|$, but not the phase angles α_{hkl}, of F_{hkl} $[=|F_{hkl}|\exp(2\pi i\alpha_{hkl})]$. Thus, the diffraction pattern is analogous to only half of a holographic image. Therefore, solving a crystal structure is essentially a matter of solving the *phase-angle problem* or what crystallographers call the *phase problem*. Various methods have been devised for solving this problem, and the Patterson method invented in 1934 dominated until the 1980s (Glusker *et al.*, 1987). The Patterson function employs the square of the structure factors in a Fourier synthesis,

$$P(xyz)=\frac{1}{V^2}\sum_{-h}^{h}\sum_{-k}^{k}\sum_{-l}^{l}|F_{hkl}|^2\cos[2\pi(hx+ky+lz)],$$

where $P(xyz)$ is the Patterson density at the coordinates x,y,z, and its peak positions correspond to the interatomic vectors transferred to a common origin. Analysis of the Patterson map often reveals relative pairs of atoms. Since heavy atoms, such as chlorine or iron, say, in an organic substance, dominate the x-ray scattering, vectors between them in the crystal are commanding features of the Patterson map, from which locations of heavy atoms can easily be determined. Once the locations of the heavy atoms are found, structure factor calculations using them give phase angles that approximate those using all atoms. Thus borrowing the phase angles derived from the heavy atoms, but using the measured amplitudes for the Fourier synthesis, reveals the locations of the light atoms N, O, C, and H of most organic compounds. This is known as the *heavy-atom method* of Robertson (1953) for solving organic and protein crystal structures (see Sec. 2.3). However, the situation for metallic alloys is more difficult, since generally atomic-scattering factors of the different atoms may differ little from the average value [see Sec. 2.1.2, point **2**].

Hauptman and Karle (1953) applied statistics to the intensities of reflections, and their ideas led to wide success in solving the phase-angle problem during the 1970s. These are called *direct methods*, because the phase angles are derived directly from a mathematical treatment of the intensities themselves.

When phase angles for many reflections are derived from the locations of some atoms or by direct methods, these reflections can be used in the Fourier synthesis to reveal more atomic positions. Successive rounds of calculations generally lead to a satisfactory structure, and when all atoms are located with the help of chemical knowledge, better values of the atomic coordinates can be obtained by carrying out least-squares refinements. These minimize the factor R, which is the normalized difference between calculated and observed intensities of all reflections for a given structural arrangement:

$$R=\left(\frac{\Sigma^{hkl}(|F_{hkl}|_o-|F_{hkl}|)^2}{\Sigma^{hkl}|F_{hkl}|^2}\right)^{1/2}.$$

Early structure determinations from film methods gave R values of 10% or more, but modern data collected using diffractometers give R values between 9% and 2%.

Failure to obtain, after successive refinements, what would be considered a satisfactorily low R value should indicate possibly

1. a wrong choice of space group,
2. significant errors in phase-angle determination,
3. the incorrect positioning of some atoms in the structure,
4. incorrect assignment of elements to crystal sites or unrecognized random distribution of elements among the sites, or
5. any combination of points **1**–**4**.

A final check of the correctness of the structure can be obtained by calculation of a *difference Fourier synthesis*, $\Delta\rho$, which gives a three-dimensional map. Here the calculated electron-density map is subtracted from the observed one. Significant maxima or minima of electron density in the difference map indicate the possible presence of unplaced atoms, misplaced atoms, or incorrect assignment of atoms to crystal sites.

So far we have ignored two major factors that influence the observed intensities of x-ray reflections, compared to those calculated for a proposed structure:

1. *Thermal vibrations of the atoms.* This is taken into account by introducing into calculations a term known as the *temperature factor.*
2. *Absorption of x rays in the crystal studied.* To minimize the effects of x-ray absorption relative to crystal shape, ideally very small spherical single crystals should be examined.

When this is not possible, especially when the substance contains elements of high atomic number, estimates of the effects of absorption on the observed intensities of reflections relative to crystal shape, orientation, and dimensions must be made and taken into account.

Although modern crystal-structure analysis is now based on computer-controlled diffractometers and does not require much use of photographic techniques, many structures have been determined previously using various methods, so we give a general outline of these here.

2.2.1 The Laue Methods. The Laue methods (Amorós *et al.*, 1975) place a flat photographic film in a plane perpendicular to a beam of x rays of continuous wavelength to record the diffraction pattern from a stationary single crystal (Fig. 8). The film can be placed on either the same or the opposite side as the x-ray source, giving back-reflection or transmission Laue photographs, respectively. The Laue method is now used only to study crystal symmetry, perfection, or alignment for other experiments, although it is the easiest experiment to perform. Reflections from some zones appear on the film, and spots of each zone form ellipses on transmission and hyperbolas on back-reflection Laue patterns.

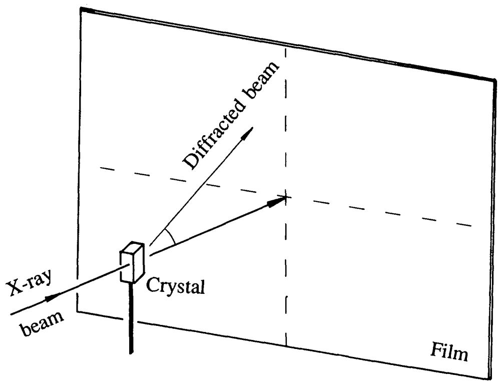

FIG. 8. A simple sketch showing the major arrangements in the Laue methods of x-ray diffraction.

2.2.2 Cylindrical-Film Methods. A flat film in the Laue method covers only a limited area for recording the reflections. More data can be collected when the film forms a cylinder coaxial with the rotation axis of the crystal. The axis is normal to the x-ray beam, giving rotation or oscillation photographs. Furthermore, if the film is allowed to translate while the crystal rotates, and if only one zone of reflection is allowed to be recorded at a time, a distorted reciprocal-lattice plane, e.g., (*hk*2), is recorded. This technique is called the Weissenberg method, which allows reflections from each set of planes (*hkl*) to be separated and their intensities measured. In the simple rotation methods, many reflections coincide in a single spot. Before 1970, the majority of intensity data for crystals was collected using the Weissenberg method, and the instrument, the Weissenberg camera, could also take cylindrical-film Laue, oscillation, and rotation photographs.

2.2.3 Precession Methods. In the precession method (Buerger, 1964), a simple direction, e.g., [100], of the crystal is aligned in the direction of the x-ray beam and is then tilted with respect to and caused to precess around it. This gives reflections of undistorted reciprocal-lattice planes for selected zones. Techniques used in aligning the crystal and selecting zones are useful in the study of crystal orientation as well as symmetry investigations. Careful study of such photographs is required to discover superstructures, especially if reflections due to the superstructure are weak. Precession cameras can also be used to take Laue photographs.

In summary, x-ray photographs reveal crystal symmetries and make possible the determination of cell dimensions, and generally of the space group. Estimation of the intensities of reflections from films was the major technique for crystal structure determination until the 1970s, when the use of automatic diffractometers that record intensities with scintillation detectors became common.

2.2.4 Diffractometers. A diffractometer is an instrument that rotates the crystal about three independent directions using three concentric circles. The angles on the circles are called ϕ, χ, and ω. The detector scans through the angle 2θ to measure the x-ray intensity. Thus the instruments are called four-circle diffractometers. The required orientations of the crystal for it to reflect the x-ray beam from (*hkl*) planes towards the scintillation detector used to measure the intensity are calculated from the unit-cell dimensions and orientation already determined. Computers are essential not only for these calculations, but also for controlling the angle driver and the detector. By the early 1980s, computer programs for solving the crystal structure had been integrated with those for diffractometer control to create complete operating packages.

2.3 Macromolecule Crystallography

Structure determination of proteins is a young science started in the 1930s, but its progress was slow at first, rapid growth only taking place since the 1970s. Compounds with large molecules include enzymes, redox proteins, carrier and storage proteins, hormones, antibodies, and viruses (see PROTEINS). Proteins are thousand-atom, chainlike, very large molecules, the building blocks of which are amino acids. The long chains of proteins are always folded into a globular shape (Blundell and Johnson, 1976).

Compared to small-molecule structural studies, the most difficult task in macromolecule crystallography has been the growth of suitable crystals and the preparation of heavy-atom derivatives. Nowadays, conditions for growing some protein crystals are well understood. Suitable methods have been found for growing an increasing number of proteins and enzymes. Growing crystals with heavy atoms inserted made the solution of the phase-angle problem easier.

Once the crystals are available, data collection, phase-angle determination, calculation of electron-density maps, interpretation of the maps, and least-squares refinements can follow, as in normal crystallography. However, because of the large molecular size, the unit cells of their crystals are very large, and many more reflections have to be collected, often using numerous crystals, since prolonged exposure to x rays may be destructive (see RADIATION DAMAGE OF CRYSTALS), and the calculation and interpretation of data are truly tedious. Many years of work used to be required for such a structure analysis; since faster data collection facilities and high-speed computing facilities have become available, this is no longer the case. When initially a small number of intensities is used, a low-resolution blurry image of the molecule may appear. When more reflections are added as time goes on, better resolution may reveal the backbone of the molecule and eventually even atomic sites. Thus the solution of protein and similar structures proceeds, as indeed do all structural solutions, by successive iterations whereby in the case of proteins, the resolution is successively increased, say from a scale of 10 to 5 to 2 Å, when individual major atom positions may appear in electron-density maps or sections. At 1 Å resolution, all atomic positions are accurately determined. Enzyme projects include the study of interactions between the protein and the *substrate* or between the protein and the *inhibitors*. These results lead to the understanding of biochemical reactions at the molecular level that are of great consequence in medicine and drug design. Furthermore, computer displays are used to model interactions between enzyme molecules, and these facilities enable biochemists and medical scientists to "see" biochemical reaction mechanisms and to facilitate the design of certain drugs.

Many countries have sites for macromolecular data collections, e.g., Cornell High Energy Synchrotron Source (CHESS), Stanford Synchrotron Radiation Laboratory (SSRL), National Synchrotron Light Source (NSLS) in the U.S.A.; Laboratoire d'Utilisation de Rayonnement Electromagnetique (LURE) in France; European Molecular Biology Laboratory (EMBL) in Germany; and Daresbury Laboratory in the United Kingdom.

Diffraction studies of synthetic and natural polymers usually give information about crystallinity and texture of the sample; low-angle scattering experiments are often employed. However, high-angle scattering experiments for polymeric molecules have lately been made possible by use of synchrotron radiations from which a suitable wavelength can be selected from a continuous spectrum of high intensity.

2.4 X-Ray Analysis of Polycrystalline Samples

Before 1960, many simple and even complex structures, particularly of metals and alloys, were solved by cylindrical-film powder methods, because single crystals could not be obtained. However, since crystal orientation is random in polycrystalline materials, reflections due to equivalent planes appear as arcs on the cylindrical film, and much information is lost compared to that obtainable from Weissenberg photographs of single crystals. More recently, computer-aided data analysis has increased the effectiveness of structure determination by powder methods when single crystals are unavailable.

Generally, powder-diffraction techniques are most useful for accurate determination of unit-cell dimensions and therefore for measurement of thermal expansion coefficients, since room-temperature films can be used to record reflections from samples cooled to as low as liquid-helium temperatures, or, enclosed in thin silica tubes, heated to high temperatures. However, diffractometer techniques are now used also for these purposes. Since most natural or industrial materials are polycrystalline, powder-diffraction methods are widely used for their identification, particularly in the metal and mineral realms. To this end, the Joint Powder Diffraction Commission (JPDC) maintains both printed and computerized indexes of powder-diffraction patterns of all known substances for their identification. Access to these can be had by various means but principally by the matching interplanar (*d*) spacings of the three most intense lines in the powder-diffraction pattern.

2.5 Other Diffraction Methods

X rays are electromagnetic waves, whereas neutrons and electrons are usually considered particles; however, a beam of each has the dual property that it can be treated either as a wave motion or as a beam of particles. In diffraction experiments, the former property is of interest. When just emitted from the nucleus of an atom during a nuclear reaction or fission, neutrons have an energy of 0.5 to 5 MeV, but are quickly slowed down when they collide with light atoms such as hydrogen in water, and become *thermal neutrons* of energy 0.025 eV at 293 K, with a de Broglie wavelength of $\lambda = h/mv = 18$ nm (1.8 Å). On the other hand, only high-energy electrons give a desirable wavelength for diffraction experiments.

2.5.1 Neutron Diffraction. The main thermal-neutron sources are nuclear reactors. Elaborate equipment is required to protect researchers and detectors from the harmful radiation; otherwise, the geometric requirements are similar to those for x-ray-diffraction experiments (see NEUTRON DIFFRACTION).

Since neutrons are scattered by the nuclei, rather than by the electrons, neutron-diffraction experiments give the locations of the nuclei in structures. Although neutron diffractions "see" time-averaged crystal structures, the inelastically scattered neutrons can also be measured by comparing the energies of neutrons before and after they interact with atoms in solids. Inelastic scattering experiments make possible the study of dynamic phenomena in the solid state. Thermal neutrons have velocities that are comparable to the thermal waves in solids; hence neutron-scattering experiments allow the observation of the Doppler shift of the scattered neutrons. This property makes possible the investigation of the dynamics (vibration spectrum) of atomic systems, by separating the inelastic and elastic scattering data (Bacon, 1975).

Compared to x-ray diffraction, neutron-diffraction experiments offer three distinct advantages:

1. Since neutron atomic-scattering factors can be either positive, negative, or approximately 0, judicious design of a neutron-diffraction experiment or comparison of the results of x-ray- and neutron-diffraction experiments can give specific information on the location of elemental species on crystal sites.
2. Since the atomic-scattering factor of hydrogen (and more so of its isotope deuterium) by neutrons is relatively much larger than that by x rays (Fig. 6), neutron-diffraction experiments are greatly favored over x-ray experiments for determining the location of hydrogen atoms in structures.
3. Since the neutron diffraction consists of "normal" and magnetic interactions, the two components of the scattering observed can be separated. Thus magnetic properties

associated with certain atoms and their specific sites in crystal structures can be studied (see MAGNETIC ORDERING IN SOLIDS).

2.5.2 Electron Diffraction. Electrons in a modern electron microscope operating at a potential of 50 000 to 10^6 V can penetrate thin materials of 0.1 μm thickness with a magnification power of 10^7 or more. For diffraction purposes, the effective de Broglie wavelengths range from 0.01 to 0.1 Å, so that the Bragg angles are about 1°. Nevertheless, diffraction patterns of very small crystals can be recorded photographically and used for structural investigation. However, because of the sensitivity of orientation, a small misalignment caused by crystal defects is observable in electron-diffraction patterns. The diffracted electron beam gives the image of the area irradiated by electrons; thus it is used for the study of crystal perfection so that the properties of the material can be correlated to sample treatment (see ELECTRON DIFFRACTION).

Because of the charge on the electron, an electron beam only penetrates slightly into the surface of a solid and can only pass through a very thin film of material. Therefore, the electron-diffraction pattern is essentially only two dimensional, although use of special techniques, such as slightly changing the orientation of the surface of the thin film or solid relative to the beam, may improve this situation when structure determination is necessary. Nevertheless, electron diffraction can well be used to identify small crystals settled in a thin-film metal matrix, and the shallow penetration is used with advantage to determine the arrangement of atoms on the surface of a crystal, a technique known as low-energy electron diffraction (LEED), although precise interpretation of the results of such experiments is subject to difficulties.

3. THE RESULTS OF CRYSTAL STRUCTURE DETERMINATIONS

Here we are able to provide only brief versions of atomic arrangements in the thousands of different crystal structures that have been determined. Broadly speaking, these fall into two classes:

1. structures in which single atoms can be considered bonded together, including metals and insulators/semiconductors, where the atoms either are covalently bonded together, or are charged ions;
2. molecular crystals in which groups of atoms forming molecules are covalently bonded together so that the peripheral atoms of the molecules exhibit filled valency subshells to those of neighboring molecules. Thus cohesion between the molecules in the crystals is very weak (van der Waals forces), and such crystals are generally soft, cleave easily if the geometry is appropriate, and have relatively low melting points.

3.1 Metals

The main characteristic of the crystal structures of metals and alloys is that the atoms generally, though not invariably, have high coordination numbers (CN) (see METALS AND ALLOYS, STRUCTURE OF). The "ideal" metallic arrangement is found in the close-packed face-centered cubic structure of copper in which each atom is surrounded by 12 neighbors in a cubo-octahedron (Fig. 9) or in the close-packed hexagonal structure of magnesium, where the surrounding is a twinned cubo-octahedron. This is because 12 is the maximum number of equal-sized spheres that can pack around a central sphere, all spheres being mutually in contact. This close-packed cubic arrangement is the basis of industrially used metals and alloys (see Sec. 4). When some atoms in an alloy have much greater diameters than others, coordination numbers up to 24, or even greater, may be found.

Another common metallic structure is the body-centered cubic type of wolfram (W), or the primitive β-CuZn type (two different components in ratio 1:1) in which each atom is considered to be surrounded by 8 atoms in a cube or 14 in a dodecahedron, depending on the relative sizes of the atoms (Fig. 10). Such alloys are of less use for industrial construction, but the basic arrangement may be of interest for magnetic and other properties. The close-packed and body-centered models are the basis of many more complex superstructural alloy structures.

Four convex coordination polyhedra that have triangulated surfaces are important in the structures of alloys. These are the CN12 icosahedron, the CN14 two-capped hexagonal

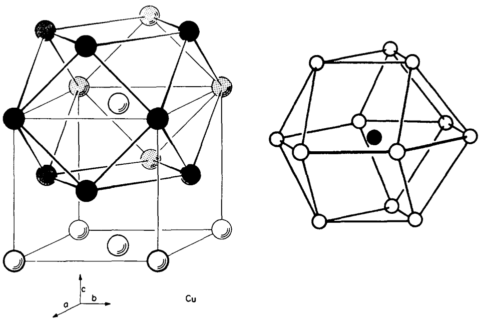

FIG. 9. The cubo-octahedron formed in all-face-centered cubic (fcc) crystals of copper (Cu), and the twinned cubo-octahedron in hexagonal-closest-packed (hcp) magnesium, which results when any one of the triangles is turned about its center by 60°.

antiprism, the CN15 three-capped truncated trigonal prism, and the CN16 four-capped truncated tetrahedron (Fig. 11). Also the trigonal prism, that may be capped by one, two, or three atoms, occurs frequently in the structures of alloys (Fig. 12).

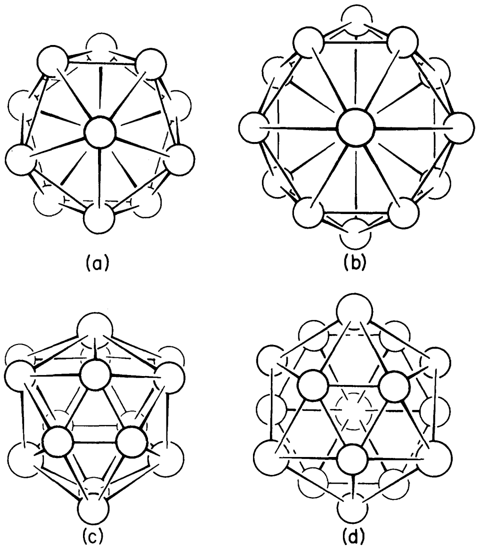

FIG. 11. Pictorial diagrams of CN12 (a) icosahedron, and (b)–(d) CN14, 15, and 16 coordination polyhedra of Frank and Kasper, central atom omitted (Frank and Kasper, 1958).

3.2 Insulators/Semiconductors

The tetrahedron and the octahedron are the most important CN polyhedra of the covalent and ionic insulators/semiconductors (see INSULATORS, STRUCTURE OF). Here the basic covalent structures are the cubic diamond and sphalerite (ZnS), or hexagonal wurtzite (ZnS) types of semiconductors, where each atom is surrounded by a tetrahedron of unlike neighbors so that adamantine arrays of covalent bonds run throughout the crystal in three dimensions (Fig. 13; see DIAMOND AND DIAMONDLIKE CARBON). Many more complex variants of these basic structural types are known. One of the basic ionic types is the rocksalt structure (NaCl), wherein each atom is surrounded by an octahedron of unlike ions

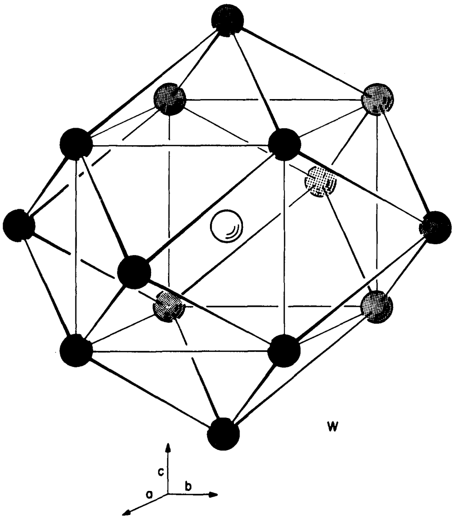

FIG. 10. The dodecahedron, CN14, in the body-centered-cubic (bcc) structure of tungsten (or wolfram W). The 14 neighbors are divided into sets of 8 (corners of cube) and 6 (body centers of neighboring cells).

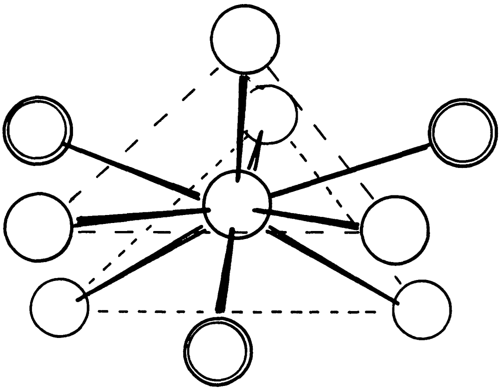

FIG. 12. A tricapped trigonal prism configuration.

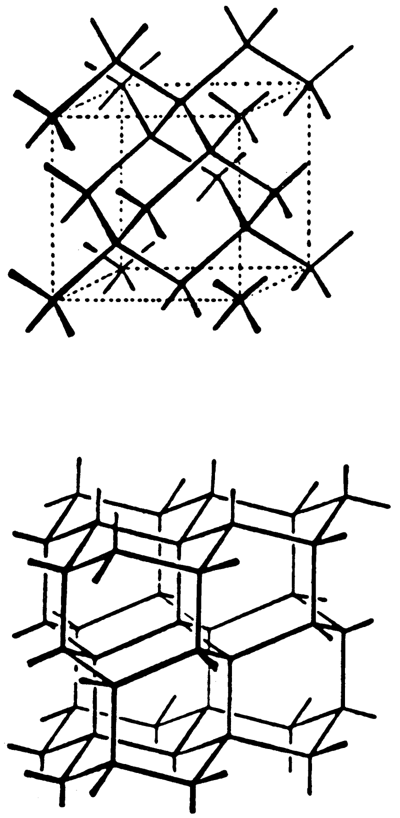

FIG. 13. The tetrahedral bonded framework of diamond or ZnS in the forms of cubic sphalerite (top) or hexagonal wurtzite type (bottom) (Wells, 1962).

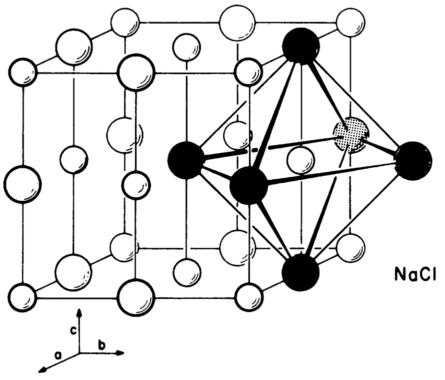

FIG. 14. The rocksalt structure as one of the basic structure types, with one octahedral coordination polyhedron outlined.

(Fig. 14). This arrangement satisfies the central forces of the predominantly ionic bonds. Many complex variants of this basic structure are also known, particularly in oxides, as are very many structures of inorganic compounds and minerals that embrace both tetrahedral and octahedral surroundings, such as spinels (Fig. 15), silicates, sulfosalts, etc. Another basic ionic structure of importance is the CsCl type (identical to the metallic β-CuZn mentioned above) in which each ion is surrounded by a cube of unlike ions. Finally, we draw attention to the frequent occurrence of trigonal prismatic coordination (Fig. 12) also in the structures of insulators/semiconductors.

3.3 Molecular Crystals

These mainly embrace organic materials (substances composed largely of C, H, O, and/or N), but they are also formed by covalent and ionic bonded inorganic materials. Thus the common elemental structures of Se and As form "infinite" molecules that run throughout the crystal in one and two dimensions, respectively (Figs. 16 and 17); the alloy GaTe forms an "infinite" two-dimensional molecule (Fig. 18). Many inorganic substances contain polyatomic anions or polyatomic cations in which a group of atoms is covalently bonded together to form a molecule that has an overall ionic charge, the charges being neutralized in the crystal by other ions of opposite charge: e.g., Na_2TaF_8 (Fig. 19) and $Mo[S_2CN(C_2H_5)_2]I_3$ (Fig. 20).

As well, metallic or insulator structures may contain uncentered clusters of atoms, the

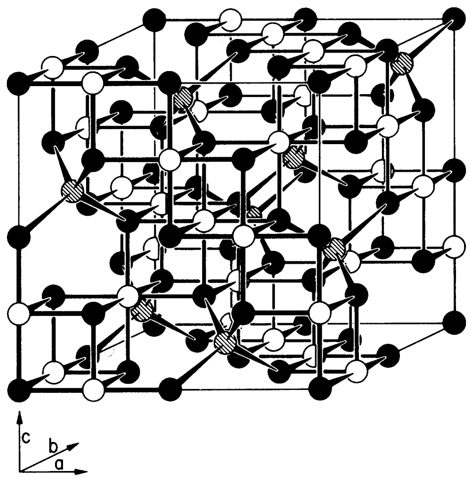

FIG. 15. The spinel crystal structure type.

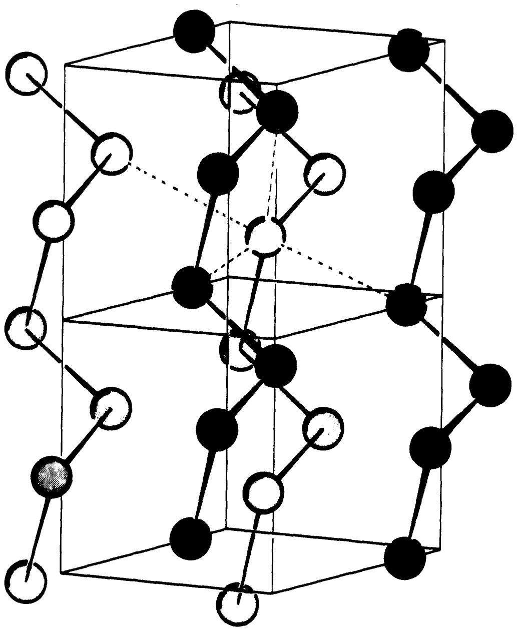

FIG. 16. The structure of Se and Te forming infinite chains in the crystal.

residual valency requirements of which are satisfied by forming occasional covalent bonds to neighboring clusters or intermediate atoms, or by the presence of charged ions in the spaces between the clusters—e.g., boron icosahedral clusters (Fig. 21) or Mo_3Se_4, CN14 (cube+octahedron clusters).

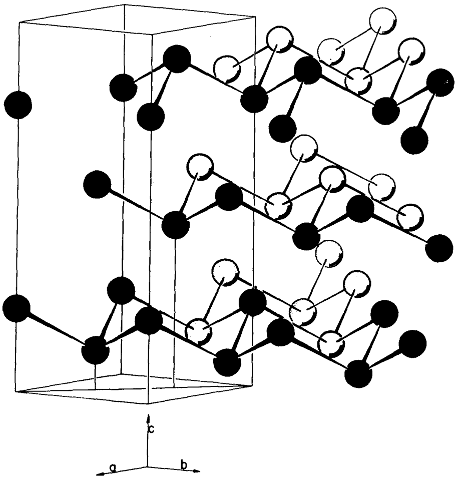

FIG. 17. The structure of sheets of the element arsenic, As.

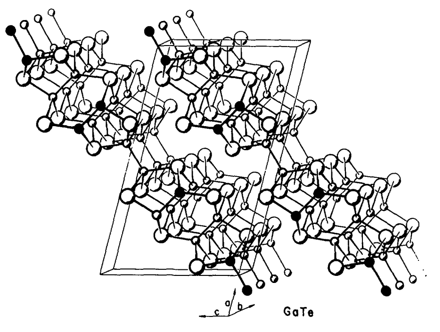

FIG. 18. The infinite-sheet structure of GaTe, with Ga and Te atoms represented by small and large spheres, respectively.

The crystallography of organic molecules is clearly a function of the most efficient packing together of the various molecular shapes as discussed by Kitaigorodski (1963) (see e.g., Fig. 22). However, ionic charges, polarity, and hydrogen bonding are some of the features that can greatly influence molecular packing and affect the cohesion between molecules. Atoms in some organic molecules called *zwitterions* are the sites of distinct positive or negative charges. Such molecules pack together, as far as shape permits, so that negative charges of one molecule attract positive charges of others. The separation of charges is never complete in a bond involving two different elements, and these bonds are said to be *polar*. Molecules usually pack in such a way that their polar branches are attracted to each other, whereas the nonpolar segments also tend to pack together (see Fig. 23). The —OH and —NH groups in a molecule tend to seek negative sites of polar bonds, and *hydrogen bonds* are formed via the positive

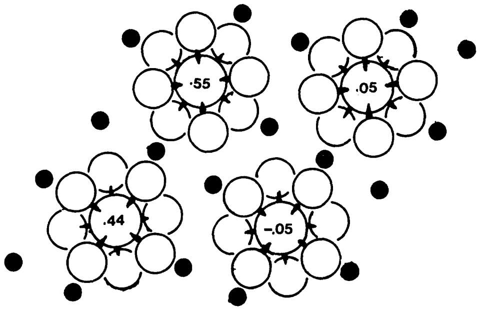

FIG. 19. The arrangement of Na^+ and $TaF_8{}^{2-}$ ions in Na_2TaF_8. Eight F atoms form a tetragonal antiprism. Fractional values in the centers of circles are the *y* coordinates along the *b* axis.

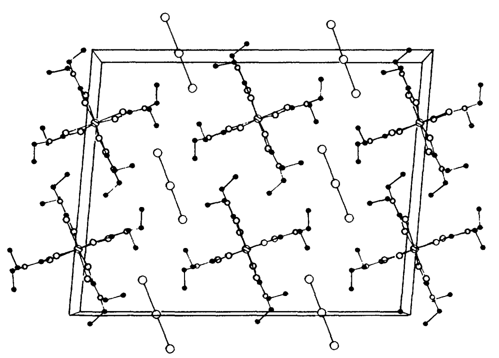

FIG. 20. Large cations $\mathrm{Mo[S_2CN(C_2H_5)_2]^+}$ and linear $\mathrm{I_3^-}$ anions in the crystal of $\mathrm{Mo[S_2CN(C_2H_5)_2]I_3}$.

nature of the H atoms in them. The electronegative atoms O and N in carboxyl —C(=O)OH, hydroxyl —OH, amine —NH and other groups or ions attract these hydrogen atoms. Since molecules may contain many of these groups, their packing together is arranged so as to permit the maximum number of intermolecular hydrogen bonds to form, thus considerably increasing the intermolecular cohesion (Fig. 24).

Finally, there are the structures of very large organic molecules such as proteins and viruses that can be (extracted and) grown in crystalline form for structural analysis. The crystals of these compounds fasten the molecules in rigid positions for x-ray studies. However, the features of their molecular structures are of prime interest. In Fig. 25, a picture of the structure of deoxyribonucleic acid (DNA) is shown.

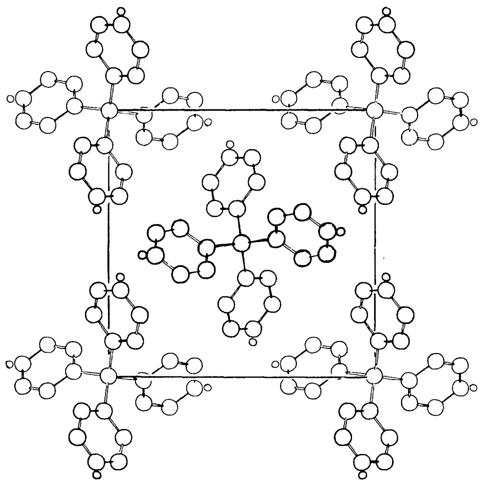

FIG. 22. The packing of tetraphenylmethane in a tetragonal system, a structure common to all tetraphenyl derivatives of the carbon group (C, Si, Ge, and Pb). The molecule in the center is located halfway up along **c**.

It should also be noted that the structures of inorganic polyionic molecules containing water of crystallization (see Fig. 24) can be greatly influenced by hydrogen-bond formation.

3.4 Assessment and Collation of Crystal Structure Data

With the number of determined crystal structures exceeding a hundred thousand, there is an obvious problem of recovering data when required, and perhaps more signifi-

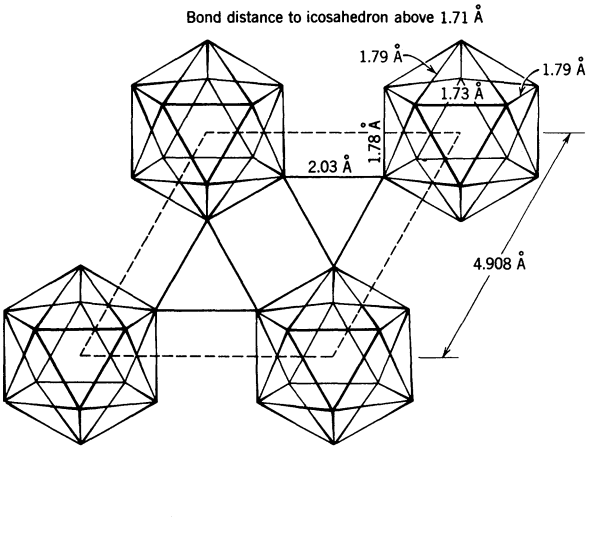

FIG. 21. Predominant icosahedral clusters of boron are the feature of the α-B structure (Decker and Kasper, 1959).

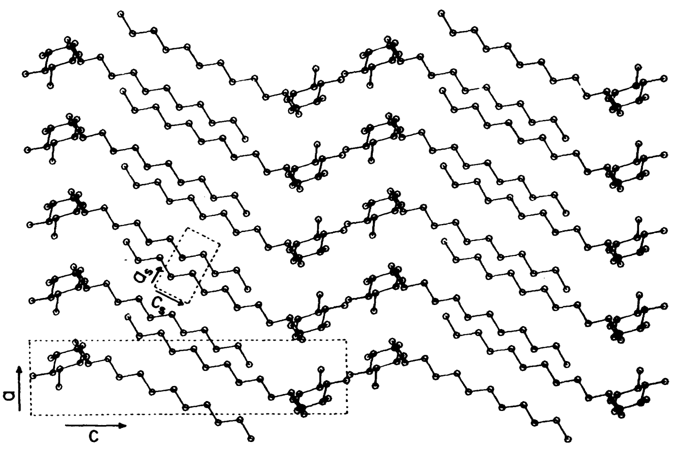

FIG. 23. Molecular packing of 1-decyl α-D-glucopyranoside, $C_{16}H_{32}O_6$, projected along *b*. The polar heads attract each other, leaving the nonpolar chains or tails staying together by themselves. The hydrocarbon subcell also appears in the projection (Moews and Knox, 1976).

cantly of usefully comparing and collating the data at hand. The first problem is now well served by computer data banks that permit retrieval of data under many headings. Of the many such private and public data banks, we shall mention only a few: Cambridge Structural Database (U.K.), organic; National Research Council Canada Metals Data Files (Canada); Gmelin/FIZ Inorganic Crystal Structure Database (Germany); Brookhaven Protein Databank (U.S.A.); and European Molecular Biology Laboratory Database (France).

The second problem of assessment and collation of data by the human mind is much more difficult, since even a mass of computer printouts is, by virtue of its very volume, likely to be intractable. However, in the metal (particularly) and inorganic realms, the concept of a *crystal structure type* developed over the years has proved useful. Although such a type, defined by the crystal's space group and the partially or fully occupied Wyckoff sites therein, may be unique to a single substance, generally it is adopted by several substances and occasionally by a hundred or even several hundred substances. Thus with this concept, the same structure type of many substances can be considered a single unit of information, and can be related to other structure types, and to the observed physical and other properties of its adherents. The advantage of this classification is that it groups together the structures of substances, regardless of their relative unit-cell sizes, relative ratios of the

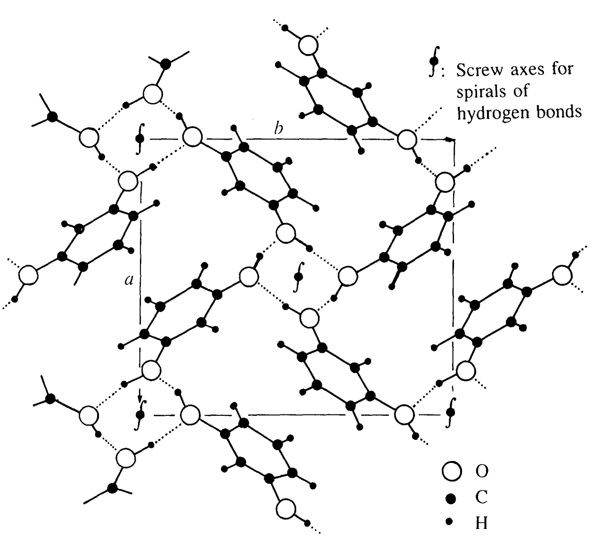

FIG. 24. A projection of the crystal structure of α-resorcinol, $C_6H_4(OH)_2$, showing two hydrogen bonds for each –OH group. The spirals of hydrogen bonds follow the crystallographic screw axes 2_1 which is perpendicular to the paper. The hydrogen positions have been determined by neutron diffraction.

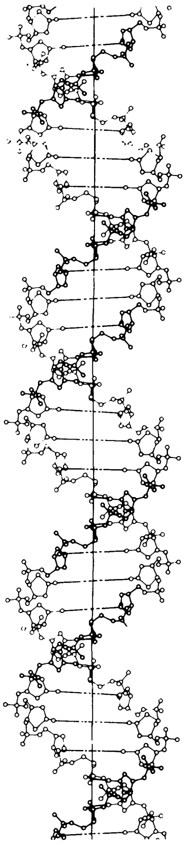

(*b*)

FIG. 25. The structure of DNA, where the hydrogen bonds between the pairs of bases are responsible for keeping the two helices together (Roller-Masser, 1973, where credit is given to M. H. F. Wilkins).

lengths of the unit-cell edges, unit-cell angles, and relative values of variable atomic coordinates x,y,z. Its disadvantage is that it may asssemble together structures in which the coordination of the atoms differs. Generally this consequence is not serious compared to its simplicity and advantages, and it can be taken into account by noting the combinations of structural parameters at which changes of coordination occur.

The advantages of being able to interrelate many different structure types include

1. comparison of structures with similar atomic coordinations that may be significant in relation to particular physical or other properties, and
2. showing that a particular complex structure with a large unit cell is but a superstructure of a much simpler type, the superstructure being caused, e.g., by the ordering requirements of the atoms in the complex structure, or by atomic displacements from regular crystal sites caused by the particular chemical bonding needs of the atoms in the complex structure.

This concept has been basic to many famous books that have been written over the years; some of these are given in the Further Reading list.

4. CRYSTALLOGRAPHY IN INDUSTRY AND MEDICINE

Crystallographic knowledge is basic to the industrial use of materials, which is now known as *materials science.* For example, close-packed cubic metals and alloys are used for construction and mechanical devices because they are isotropic and because the close packing of the equilateral triangular networks of atoms to which the structure gives rise readily permits six active slip systems that facilitate plastic deformation (forming) of a polycrystalline material. Another mechanism for plastic deformation is mechanical twinning, in which the atomic planes slip through a definite fraction of the crystallographic repeat distance, in such a way as to produce a new crystal in twinned orientation relative to the original. Hardening generally results from precipitation of very small crystals of other symmetry in the cubic matrix as a result of suitable adjustment of composition by addition of minor quantities of other elements and annealing at elevated temperatures, followed by rapid cooling (quenching). Information on such precipitates derives from crystallographic studies, as does information on "textures" (crystal orientation) following forming operations. Metals are also important as electrical conductors, a property that arises from the unsaturated metallic bonding or partly filled overlapping energy bands, but which is influenced by interatomic distance, high purity, and perfect (not disordered) occupation of crystallographic sites in intermetallic compounds. Useful magnetic properties of alloys depend on composition ("magnetic" metallic elements present) and particularly on the crystallographic properties of atomic arrangement, coordination, and interatomic distances.

Superconductivity does not have its immediate origins in crystallographic arrangement, but it is interesting to note that all ordinary superconductors ($T_c < 23$ K) with high transition temperatures have one of some half-dozen structure types. Among the new superconductors with much higher transition temperatures, not only is composition important but also structural arrangement and the degree of order or disorder of the components on the crystal sites.

The important electrical properties of inorganic materials (insulators/semiconductors) and their practical use in devices depends not only on composition but also on crystallographic knowledge and crystal-growth techniques for the preparation of highly pure crystals and the creation of graded junctions, etc. Superionic conductivity is directly dependent on crystallographic knowledge, since the structures of such conductors must have linear channels of diameter sufficient for single conductor ions to move through unimpeded on application of an electric field. Magnetic properties of these materials (such as oxides) are, or were, exceedingly important to the recording industry, and structural information has played its part in the development of industrial magnets.

Crystallographic information has been a component in the pharmacological development of "drugs" that have so greatly influenced the modern practice of medicine. However, the greatest triumph of crystallography for modern medicine was Watson and Crick's (1953) determination of the structure of DNA, which resulted also from the concomitant work of many other able crystallographers. This, followed by the determination of the structures of proteins, has lead to the unravelling of the genetic code, the possible recognition of the causes of bodily malfunction that result from defects therein, etc. In addition, diffraction experiments have been used in examining the function of cells and cell membranes.

5. EPILOGUE

This article would not be complete without positive evidence of the importance attached to crystallography by the scientific community, and what better way is there than by listing the commanding number of Nobel Laureates who have been recognized for their achievements related to crystallography. Despite the inexhaustiveness, such a list marks the major milestones since the inauguration of the prize.

Physics:

1901	Wilhelm Röntgen, discovery of x rays
1914	Max von Laue, discovery of diffraction of x rays by crystals
1915	William Bragg and Lawrence Bragg, analysis of crystal structure by means of x rays
1929	Prince Louis de Broglie, discovery of the wave nature of electrons
1935	James Chadwick, discovery of the neutron
1937	Clinton Davisson and George Paget Thomson, discovery of electron diffraction by crystals.

Chemistry:

1936	Peter Debye, studies of dipole moments and the diffraction of x rays and electrons in gases
1946	James Sumner, discovery of enzyme crystallization
1954	Linus Pauling, study of the nature of the chemical bond
1962	John C. Kendrew and Max F. Perutz, determination of the structure of hemoproteins
1964	Dorothy M. C. Hodgkin, determination of the structure of vitamin B_{12}
1976	William N. Lipscomb, revelation of the structure of boranes
1980	Paul Berg, Walter Gilbert, and Frederick Sanger, studies of structure and biochemical function of DNA
1982	Aaron Klug, determination of the structures of viruses
1985	Herbert Aaron Hauptman and Jerome Karle, development of the procedures for the solution of the phase-angle problem in x-ray diffraction
1988	Johann Deisenhofer, Robert Huber, and Hartmut Michel, structure determination of a large (more than 10 000 atoms) bacterial protein complex responsible for photosynthesis.

Physiology and medicine:

1962 Francis H. C. Crick, James D. Watson, and Maurice H. F. Wilkins, elucidation of the structure of DNA.

GLOSSARY

Absorption Edge: The sharp difference in absorption coefficients of x rays within a short range of wavelength. The absorption edges are unique for each chemical element.

Anomalous Scattering: The change of atomic scattering amplitude due to a phase-angle shift for scattered x rays from a chemical element in a crystal, particularly noticeable when the element has an absorption edge near the wavelength of the radiation used for the diffraction study.

Crystal Structure: The arrangement of ions, atoms, and molecules in a crystal. The symmetry and geometry exhibited by their arrangement are part of the structural properties.

Crystal Symmetry: The balanced proportions present in cubes, hexagonal or rectangular prisms, and other polyhedra; see **Point Group**.

Diffraction: The interference phenomenon of waves when they travel through a grating in their path. In x-ray diffraction, the crystal is the grating.

Lattice: An infinite periodic array of points at the termini of vectors $\mathbf{L}_{uvw} = u\mathbf{a} + v\mathbf{b} + w\mathbf{c}$, where **a,b,c** are three noncoplanar vectors.

Miller Indices: Three integers enclosed in parentheses (h,k,l) used to represent a set of (parallel) crystal planes. When enclosed in brackets $[h,k,l]$, they represent a direction normal to the plane; and $\{h,k,l\}$ is a form representing a set of symmetry-equivalent planes.

Organic Compounds: Compounds whose molecules are made up mainly of carbon, hydrogen, oxygen, and/or nitrogen atoms. However, in terms of crystal structure classification, these compounds may also contain other elements such as metals, even though they are then called organometallic compounds by chemists.

Phase: A crystalline, liquid, or gaseous state of matter, as used in Gibbs's phase rule or phase transitions.

Phase Angle: A quantity used in the analysis of interferences of waves, for which sinusoidal models are used. However, most crystallographers simply call it a phase. This term is used here to avoid confusion between it and the Gibbs phase.

Point Group: A group consisting of symmetry elements: n-fold rotation axes, mirror planes m, and/or center of inversion i. The property that makes a crystal indistinguishable from its images after being turned (along an axis) by an angle of $2\pi/n$ is called an n-fold rotation axis. When two halves of a crystal are mirror images of each other, the plane separating the halves is a mirror plane m, and a point from which two opposite directions are equivalent is called a center of inversion i.

Space Group: A group consisting of symmetry elements: those of point groups (see above), screw axes, glide planes, and lattice translations.

Symmetry Element: A process that brings the crystal or object back to itself. For example, a threefold rotation is a rotation of the crystal by 120° ($2\pi/3$). For crystal structure description, symmetry elements includes n-fold rotations, mirror planes m, center of inversion i, glide planes, screw axes, and lattice translations.

Symmetry Group: A point group or space group consisting of symmetry elements. A group consists of a set of elements with a defined group pairing operation, and it satisfies four basic axioms: closure, identity, inverse, and associative; see **Point Group** and **Space Group**.

Unit Cell: The smallest parallelepiped which when stacked repetitively in three dimensions generates the entire crystal structure.

Wyckoff Notation: A letter representing a set of symmetry-equivalent positions in a space group. A general point in a unit cell of a crystal structure belonging to any space group has coordinates x,y,z. However, if one or more of the coordinates has a special value, the number of symmetry-equivalent positions is reduced. Most space groups have several sets of such points, and they are represented by letters a,b,c, etc., which are called Wyckoff notations.

List of Works Cited

Amorós, J. L., Buerger, M. J., de Amorós, M. C. (1975), *The Laue Method*, New York: Academic.

Bacon, G. E. (1975), *Neutron Diffraction*, 3rd ed., Oxford: Clarendon.

Blundell, T. L., Johnson, L. N. (1976), *Protein Crystallography*, New York: Academic.

Buerger, M. J. (1964), *The Precession Method*, New York: Wiley.

Bunn, C. W. (1958), *Chemical Crystallography*, Oxford: Oxford University Press.

Decker, B. F., Kasper, J. S. (1959), *Acta Crystallogr.* **12**, 503–506.

Evans, R. C. (1939), *An Introduction to Crystal Chemistry*, Cambridge, England: Cambridge University Press.

Ewald, P. P. (Ed.) (1962), *Fifty Years of X-Ray Diffraction*, Utrecht: For I.U.Cr., N.A.V. Oosthoek's Uitgeversmaatschappij.

Frank, F. C., Kasper, J. S. (1958), *Acta Crystallogr.* **11**, 184–190.

Glusker, J. P., Patterson, B. K., Rossi, M. (1987), *Patterson and Pattersons*, New York: Oxford Science Publications and International Union of Crystallography.

Kitaigorodski, A. I. (1961), *Organic Chemical Crystallography*, New York: Consultants Bureau.

Hahn, T. (Ed.) (1983), *International Tables for Crystallography*, Vol. A. Boston: International Union of Crystallography and D. Reidel.

Hauptman, H. A., Karle, J. (1953), *Solution to the Phase Problem: I. The Centrosymmetric Crystal*, Ann Arbor: American Crystallographic Association.

Moews, P. C., Knox, J. R. (1976), *J. Am. Chem. Soc.* **98**, 6628–6633.

Reller-Massar, Ann (1973), *Discovering the Basis of Life: An Introduction to Molecular Biology*, New York: McGraw-Hill.

Robertson, J. M. (1953), *Organic Crystals and Molecules*, Ithaca: Cornell University Press.

Watson, J. D., Crick, F. H. C. (1953), *Nature (London)* **171**, 737–738.

Wells, A. F. (1962), *Structural Inorganic Chemistry*, 3rd ed., Oxford: Oxford University Press.

Further Reading

Barrett, C. S., Massalski, T. B. (1966), *Structure of Metals*, New York: McGraw-Hill.

Brice, J. C. (1986), *Crystal Growth Process*, Glasgow: Blackie.

International Tables for X-Ray Crystallography, 3rd ed., Vols. I (1969), II (1972), and III (1968), Birmingham, England: Kynoch.

Jaric, M. V. (Ed.) (1988), *Introduction to Quasicrystals*, Boston: Academic.

Kelker, H., Hatz, R. (1980), *Handbook of Liquid Crystals*, Weinheim: Verlag Chemie.

Pauling, L. (1960), *The Nature of the Chemical Bond*, 3rd ed., Ithaca: Cornell University Press.

Pearson, W. B. (1972), *The Crystal Chemistry and Physics of Metals and Alloys*, New York: Wiley-Interscience.

Villars, P., Calvert, L. D. (1985), *Handbook of Crystallographic Data for Intermetallic Phases*, Metals Park, OH: American Society for Metals.

CVD (CHEMICAL VAPOR DEPOSITION)

BERNARD S. MEYERSON, *IBM Research Division, IBM T. J. Watson Research Center, Yorktown Heights, New York, U.S.A.*

Introduction 409
1. **Classification by Excitation** .. 409
1.1 Thermal CVD 410
1.2 Energetically Enhanced CVD Processes 410
2. **Classification by Reactant Transport** 410
2.1 Atmospheric-Pressure CVD.... 411
2.2 Subatmospheric-Pressure CVD 412
3. **Chemical Vapor Deposition** .. 413
3.1 Atmospheric-Pressure CVD.... 413
3.2 Reduced-Pressure CVD (RPCVD) 417
3.3 Low-Pressure CVD (LPCVD) .. 417
3.4 CVD in Molecular Flow....... 417
3.5 Plasma-Enhanced CVD 418
3.6 Photolytic CVD.............. 419
4. **Models—Equilibrium Thermodynamics Versus Kinetics** 419
4.1 Equilibrium Thermodynamics. 419
4.2 Kinetics-Based Reactor Modeling 421
5. **Applications** 422
5.1 Passive Circuit Elements and Structures.................... 422
5.2 Active Circuit Elements and Device Structures............. 422
5.3 Fundamental Materials Studies 423
6. **Summary**.................... 424
Glossary..................... 424
Works Cited 425

INTRODUCTION

Chemical vapor deposition (CVD) is the process wherein a gaseous source species is decomposed to form a film upon a substrate. This process requires some form of energetic input, whereby the source species is dissociated in the course of film growth. This is distinct from a condensation process, where a vapor is condensed to form a solid, with the source species undergoing only a phase transition. In its most common form, CVD relies upon elevated temperatures to pyrolyze the gaseous precursor for film growth. However, in instances where high growth temperatures are incompatible with the substrate materials, or where metastable materials are to be grown, a wide variety of energetic inputs have been employed to facilitate film growth at reduced temperatures. Examples are plasma excitation (dc, rf, and microwave), photoexcitation (lamp and laser driven), and remote thermal excitation (direct heating of the gaseous source).

CVD is at present the method of choice for the deposition of both mechanical and electronic coatings in the industrial setting. This reflects the numerous desirable characteristics associated with the growth of a material from a gaseous source, key features being excellent film conformality, film uniformity, high crystalline perfection, tight compositional control, high process throughput, and relatively low cost. The intent of this article is to provide an overview of the state of the art in CVD, and to serve as a resource to those who require information on CVD methods for film deposition within a particular material system.

1. CLASSIFICATION BY EXCITATION

CVD processes, independent of the film to be grown, are commonly named employing two distinct strategies. One grouping is done in terms of the excitation source employed, thus the nomenclature plasma-enhanced CVD, photolytic CVD, photothermal CVD, etc. A second grouping is performed in terms of the characteristics of gaseous transport within the

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

reaction vessel. First to be considered will be the grouping of CVD processes by excitation source.

1.1 Thermal CVD

The earliest forms of CVD relied upon the simple pyrolysis of a gaseous source to form film, and this remains the most commonly employed method to date. A wide variety of heating methods are employed, those being the use of an isothermal furnace, lamp-heated or rf-heated susceptors (energy-absorbing substrate holders), photothermal methods using lamps or laser radiation, or the Ohmic heating of a substrate itself. Variants include the heating of a gaseous source so as to prepyrolyze it, forming highly reactive radicals such that films may be grown on substrates held at temperatures lower than would otherwise have been required.

1.2 Energetically Enhanced CVD Processes

The second broad category encompasses those methods where electronic excitation is employed to activate film growth chemistry directly. The primary reason for the development of this methodology was the desire to go to ever lower processing temperatures, thus achieving the growth of nonequilibrium layers, and allowing the use of a wider variety of substrate materials. Two subcategories exist, those being techniques that are based upon electronic versus optical excitation.

When electronic excitation is employed, rf, dc, or microwave energy is imparted to a gaseous precursor. This can be a direct process, where the discharge is within the gaseous source itself, or may be remote, exciting another gaseous species which then equilibrates with the growth source above the substrate to be coated. The excited precursor molecules then decompose to form the layers desired. The use of remote excitation is particularly desirable when employing an unstable growth precursor, or depositing upon a substrate sensitive to the high-energy emissions (e.g., UV radiation) nominally present in the region of plasma excitation.

In the instance of photoexcitation, a light source is employed to impart energy to the gaseous precursor. Direct excitation requires that the growth precursor have an adequately strong absorbance at wavelengths readily accessible to conventional light sources. This coincidence is frequently absent, and many photolytic methods must therefore rely upon the use of a chemically inert sensitizing agent. As an example, minute quantities of mercury vapor may be entrained in a gaseous source, which can then be excited employing a high-power mercury light source. Photolytic CVD must be distinguished from simple photothermal methods, where a substrate may be locally heated with a focused IR source or the like. This is simply a thermally driven growth process, modified with the intent to localize film growth to a desired area.

For completeness, it must be noted that there will be as many energetically enhanced growth methods as there are ways to couple other than thermal energy into a gaseous source. An electron beam may also be employed to impart energy to an adsorbed species such that it decomposes to form film, but it and all other excitation methods share the common trait of decoupling the temperature/energy of the gaseous precursor from that of the substrate to be deposited upon.

2. CLASSIFICATION BY REACTANT TRANSPORT

Without regard to the method of excitation employed, a second commonality among the numerous CVD methods known is the mechanism of gaseous transport within the region of film growth. Reactant transport through a CVD apparatus is a key determinant in how one must approach the application and analysis of a given CVD method. The characteristic of gaseous flow through a reactor—viscous flow at the high-pressure limit, molecular flow in the low-pressure limit—determines such critical reactor parameters as reactant residence time, the homogeneity of reactant distribution, reactor memory of past gaseous content, and the rate at which the source species comes to chemical and thermal equilibrium with its environment. For historical reasons, reactant transport considerations have been included implicitly in the categorization of CVD processes by a nomenclature based around commonly employed operating pressures. This scheme is reasonable as gaseous transport and pressure are coupled for a given geometry. As such, the commonly employed terminology for classifying CVD processes is adhered to below. A cautionary note is that the true determinant of how gaseous species are transported within a given reactor geometry is the ratio between the mean free path (λ) and some typical system dimension (D). For example, where $\lambda/D \gg 1$, molecular flow, transport involving only molecule–surface interactions, dominates. It is possible in principle to achieve this at any given pressure, 1 atm for example, so long as D is appropriately small in value. This would be the case for the vapor deposition of a metallic catalyst on the interior of a pore in a catalyst pellet. This is an extreme case, but it is worth noting that this is a limitation of classifying CVD by any one system parameter.

2.1 Atmospheric-Pressure CVD

Atmospheric-pressure CVD (APCVD) processes are the simplest example of growth from a gaseous source. A gaseous source is flowed through a reactor configured as shown in Fig. 1, where a highly dilute gaseous source is made to flow at high velocity through the cooled quartz vessel. Cooling of the vessel is employed to prevent the desorption of contaminants from the vessel walls, to prevent the homogeneous nucleation of particulates due to gas-phase pyrolysis of the precursor, and to suppress the growth of film on the reactor vessel itself (to minimize particle generation). Gas velocity is maintained at a level adequate to prevent the back-diffusion of contaminating species into the reactor at the necked- down end of the apparatus. The simplicity of this apparatus belies the complexities of its chemical and hydrodynamic behavior.

APCVD processes can display complex flow geometries. Flow in an APCVD reactor, shown in Fig. 1, is a strong function of the boundary conditions of the reactor (e.g., wall temperature, gas velocity, entrance and/or exit geometries, susceptor position and temperatures, etc.). Employing the injection of smoke particles to trace flow contours in a horizontal APCVD reactor, early workers (Eversteijn *et al.*, 1970) demonstrated the existence of roll cells, where two symmetric vortices are formed along the reactor's flow axis. As seen schematically in view A of Fig. 1, roll cells spiral such that the source gas rises above the heated substrate, and returns to the bottom of the reactor cross section along the cooled walls. This greatly complicates film growth because of the associated variations of reactant thermal history. Contaminating species, dopant as an example, evaporated from the leading edge of a substrate and carried off in the spiral gas flow, may reincorporate at a point one full rotation of the cell further along the reactor axis, modifying the chemical composition of the film from that which was intended. An additional issue is that of heterogeneous (surface) versus homogeneous (gas-phase) reactor chemistry. In that conditions of viscous flow exist, the fluid stream (growth source) experiences drag due to nearby surfaces, and a slowly flowing layer of gas is found in the region ($\leqslant$1 cm) immediately above the heated susceptor. As the gaseous source in this slowly flowing region comes to thermal equilibrium with the hot susceptor and substrate, significant gas-phase pyrolysis of the source species may occur, forming a more or less reactive sequence of by-products. This leads to spatial variations of gas composition within the reactor, and complicates the attainment of layers of uniform composition. Thus, the proper design of reactor hydrodynamics is a key issue if uniform, reproducible film growth is to be achieved. To this end, extensive modeling of reactor hydrodynamics has been performed.

Gaseous transport within a reactor operating at or near atmospheric pressure has been studied by a number of workers (Coltrin *et al.*, 1984; Jensen *et al.*, 1989). Complex behaviors such as those in Fig. 1 are reproduced in

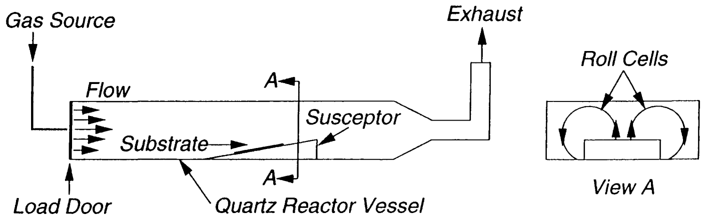

FIG. 1. Atmospheric-pressure CVD system. The tilted susceptor serves to enhance film uniformity. Shown in view "A" are two roll cells, induced by the heating of the gaseous source above the susceptor.

models, but such behavior cannot be generalized in that every reactor geometry produces a unique flow structure. The only generalization made is that of the mathematical formalism employed in the modeling of such complex systems. Hitchman and Jensen (Hitchman *et al.*, 1991) have recently reviewed the state of the art in the area of reactor modeling for a wide variety of reactor configurations.

2.2 Subatmospheric-Pressure CVD

The reduction of operating pressures has been a major factor in the advancement of CVD from a means to grow substrate-like blanket films to a method for the preparation of highly complex layered structures. A liability of many atmospheric-pressure growth processes is, as noted above, the complex hydrodynamic behavior of the system. Reactant residence times in such reactors can be long, in that recirculating flows act as storage nodes for reactants. This leads to an intermixing of layers deposited sequentially, where more abrupt compositional transitions would have been desired. In the instance where a substrate contains volatile species, such as the instance of dopants in silicon, these may be evolved into the gas phase, and reincorporate into film being deposited downstream. To reduce or eliminate such effects, a majority of CVD methods in use today operate in the subatmospheric-pressure regime.

For a fixed flow, an order of magnitude reduction in operating pressure results in an equal reduction of gas residence times within the reactor, facilitating the growth of films with controlled depth-varying compositions. Further, operating at reduced pressures has other fundamental advantages. Low-pressure operation eliminates deleterious flow effects such as the roll cells seen in APCVD. Also, by reduction of gas residence times, undesirable homogeneous chemistry is suppressed, eliminating one mechanism of reactant depletion and/or particle nucleation. When pressures reach the range of 1 to 5 Torr, difficulties encountered with vortices and other "dead spaces" seen in APCVD are absent. The resulting enhancements in the CVD process are remarkable. As an example, one may pack wafers at ever increasing density when gas-phase reactant transport is both rapid and uniform, as is the case at reduced operating pressures. Using an example from silicon technology, consider a process known generically as low-pressure chemical vapor deposition (LPCVD).

LPCVD is a hot-wall CVD growth method, where wafers are packed at high density in a quartz furnace tube. It is common to have 50 or more 125-mm silicon wafers packed coaxially at a spacing of less than 1 cm apart. Shown in Fig. 2, this high packing density makes such systems very efficient. In spite of the high packing density, film thickness across a wafer is uniform to 5% or better by this method for numerous growth processes. Unlike atmospheric pressure CVD, modeling of hydrodynamics in this system is relatively intractable, in that flow in this pressure regime is in transition from laminar, viscous flow to molecular transport (e.g., flow dominated by gas/surface interactions).

A very recent subset of subatmospheric-pressure CVD processes are those operating under conditions where gaseous transport is by molecular flow. In such instances, the assumptions of viscous transport break down, as molecule–molecule collisions become infrequent as compared to molecule–surface interactions. The onset of this condition cannot be specified by a pressure, but rather by the ratio of the molecular mean free path (λ)

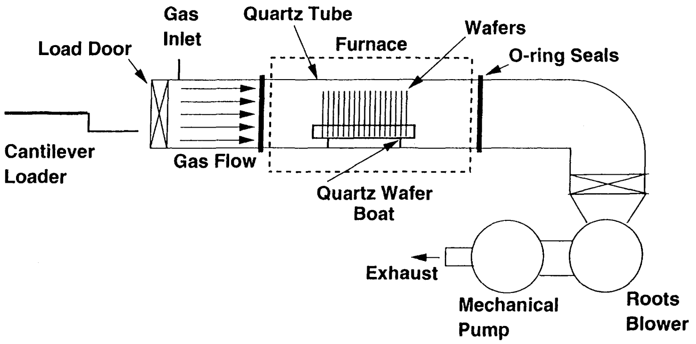

FIG. 2. A schematic of low-pressure CVD as applied to batch processing in silicon technology. High-density wafer packing is possible because of uniform gas distribution at low pressure, although other wafer carrier geometries may be employed.

to a typical system dimension (D), where $\lambda > D$. Using an interwafer dimension of 1 cm, such as in the LPCVD geometry, molecular flow will dominate for $P < 10^{-2}$ Torr. A growth method employing this approach in silicon technology is UHV/CVD (Meyerson, 1986). Similarly, molecular beam epitaxy (MBE) apparatus are being adapted to this approach, where beams of gaseous molecular growth precursors are made to impinge upon heated substrates to form film. This is in contrast to the conventional mode of MBE operation, where a beam of atoms from an evaporative source is employed (see MOLECULAR BEAM EPITAXY). CVD operated at this extreme of pressure creates significant opportunities, in that layer thickness is controllable to better than atomic monolayer dimensions, and applications exploiting this new capability will be discussed further on.

3. CHEMICAL VAPOR DEPOSITION

What has been elaborated thus far is a framework within which various methods of CVD will be discussed. However, it is not within the scope of this article to discuss at great detail any one method of CVD. Rather, this entry is to serve as a resource for workers requiring access to information about a specific growth technique or materials system. A single growth technique, plasma-enhanced CVD (PECVD), may be employed to prepare films of widely varying chemical and structural makeup. Similarly, a single materials system may be deposited employing widely varying deposition processes. Consequently, presented in Tables 1 and 2 are overviews of CVD. Table 1 is broadly organized by CVD growth technique, listing the ranges of operating parameters, and common applications for each method. The exceptionally wide range of operating conditions listed comes about from the large scope of applications for which each method is employed. Application-specific information is found in Table 2, covering silicon-technology-based processes, metal CVD, and compound semiconductor growth by CVD. These entries are organized by material system. It is common for one CVD growth technique to appear under several material systems, it being adapted to the differing requirements by modification of growth chemistry. Noted earlier was the proliferation of acronyms for various deposition processes. Such acronyms frequently attempt to give the reader a hint at the approach to film growth that is being employed. However, a tabulation by either material system or growth method acronym cannot reveal the strategy underlying the development of a given CVD process. For completeness, what follows is a series of entry-level summaries of representative CVD processes in current use.

3.1 Atmospheric-Pressure CVD

Atmospheric-pressure CVD (APCVD) is commonly employed for commercial, high-temperature silicon epitaxy, as well as to prepare materials in many compound semiconductor systems. Producing layers of extraordinary perfection ($\leqslant 10$ defects/cm^2), this method is appealing for its simplicity. High temperatures are employed for substrate preparation, desorbing contaminants, and stabilizing the substrate surface prior to deposition. Examples of this surface preparation chemistry for silicon materials would be the desorption of silicon monoxide at 1000 °C into a hydrogen ambient, while GaAs would be stabilized by a more moderate bake at 800 °C in an arsine-rich hydrogen carrier gas. Growth precursor, diluted in an inert ambient, is then flowed through the cold-walled reactor and pyrolyzed on a heated substrate. High deposition rates are readily obtained, of order several thousands of angstroms per minute. APCVD is primarily employed for the growth of thick, compositionally uniform layers, although work at the research level is under way to lower overall processing temperatures with the goal of producing highly controlled abrupt compositional transitions in film alloy and dopant content. The primary approach to the reduction of the thermal budget for APCVD is to employ rapid thermal processing methods, where samples may be cleaned at high temperatures, followed by rapid temperature reduction to the desired growth temperature. In parallel with this, efforts are under way to operate such reactors in the methodology of UHV, controlling chamber contamination levels by using load-locked introduction chambers and high-purity gaseous sources. Although total pressure remains atmospheric, the partial pressure of contaminating species in such reactors can be reduced to $\simeq 10^{-6}$

Table 1. Operating parameters of commonly employed chemical methods.

Growth technique	Deposition Pressure (Torr)	Deposition Temperature (°C)	Excitation	Film growth applications	Figure	Reference
Atmospheric-pressure CVD (APCVD)	760	150–1200	Thermal	Silicon epitaxy, silicon carbide, III-V and II–VI MOCVD, atomic-layer epitaxy of III–V compounds, tungsten silicide	1	a
Reduced-pressure CVD (RPCVD)	10–80	150–1050	Thermal	Silicon epitaxy, silicon dioxide, silicon nitride, III–V MOCVD	3	b
Low–pressure CVD (LPCVD)	0.2–2.0	150–650	Thermal	Polycrystalline silicon, amorphous silicon, silicon dioxide, silicon nitride, tungsten, copper, aluminum	2	c
CVD in molecular flow	$\leqslant 10^{-3}$	400–800	Thermal	Low-temperature silicon epitaxy, gas-source MBE of silicon and III–V alloys, atomic-layer epitaxy of III–V compounds	4	d
Plasma–enhanced CVD (PECVD)	0.001–0.5	25–800	rf, microwave, ECR	Amorphous silicon, silicon dioxide, silicon nitride, amorphous carbon, diamond, silicon carbide, low-temperature silicon epitaxy	5	e
Photolytic CVD (Photo-CVD)	80	150–600	Photolytic	Silicon dioxide, silicon nitride, low-temperature silicon epitaxy, metal (Cu, Au, W,..,)	6	f

---

[a] Hammond, 1988; Mayer and Shea, 1964; Eversteijn, 1971; Bloem and Classen, 1980; Beers and Bloem, 1982; Fujino *et al.*, 1990; Giunta *et al.*, 1990; Dobkin *et al.*, 1990.

[b] Ogirima *et al.*, 1977; Reif, 1982; Hammond, 1988.

[c] Hitchman, 1990; Brekel and Bollen, 1981; Meyerson and Olbricht, 1984; Gorowitz *et al.*, 1987; Rojas *et al.*, 1990; Beach *et. al.*, 1990.

[d] Meyerson, 1986; Meyerson *et al.*, 1987; Meyerson *et al.*, 1988b; Donahue and Reif, 1985; Yew *et al.*, 1987.

[e] Donahue and Reif, 1985; Gorowitz *et al.*, 1987; Kamins and Chiang, 1982; Emesh *et al.*, 1989; Wong and Saraswat, 1988; Kim *et al.*, 1987; Batey and Tierney, 1986; Lucovsky *et al.*, 1986; Yew *et al.*, 1987; Manabe and Mitsuyu, 1989; Pai *et al.*, 1990; Oehr and Suhr, 1988.

[f] Yamazaki *et al.*, 1990; Yamazaki *et al.*, 1984; Suzuki *et al.*, 1989; Jansinski *et al.*, 1987; Markwalder *et al.*, 1989.

Table 2. Materials commonly prepared by chemical vapor deposition.

Material deposited	Growth technique	Deposition pressure (Torr)	Deposition temperature (°C)	Precursor	Excitation	Comments	Reference
Silicon (cr)	Atmospheric-pressure CVD (APCVD)	760	800–1100	$SiCl_{4-x}H_x$ ($x \leqslant 4$), B_2H_6, PH_3, AsH_3	Thermal	Commercial silicon epitaxy	a
Silicon (cr)	Reduced-pressure CVD (RPCVD)	10–80	800–1150	$SiCl_{4-x}H_x$ ($x \leqslant 4$), B_2H_6, PH_3, AsH_3	Thermal	Commercial silicon epitaxy	b
Silicon (cr)	Plasma-enhanced CVD (PECVD)	0.001–0.1	750–800	SiH_4, Si_2H_6, B_2H_6, AsH_3	Plasma preclean, thermal growth	Low-temperature epitaxy	c
Silicon (cr) and silicon:germanium (cr)	Rapid thermal CVD (RTCVD)/(LRP)	1.0–760.0	600–1200	$SiCl_{4-x}H_x$ ($x \leqslant 4$), B_2H_6, AsH_3, GeH_4	Thermal	Low-temperature epitaxy 600–750 °C, high temperature preclean ≃1000 °C	d
Silicon (cr) and silicon:germanium (cr)	Ultrahigh vacuum/ CVD (UHV/CVD)	10^{-3}	425–800	SiH_4, PH_3, B_2H_6, GeH_4	Thermal	Low-temperature epitaxy	e
Silicon (cr) and silicon:germanium (cr)	Photolytic CVD (Photo-CVD)	80	250–600	SiH_4, $Si_2H_6PH_3$, B_2H_6, GeH_4	Photolytic	Low-temperature epitaxy	f
Si (polycrystalline)	Low-pressure CVD (LPCVD)	0.3–1.0	500– 700	SiH_4, Si_2H_6, B_2H_6, PH_3, AsH_3	Thermal	Commercial process for poly-Si	g
Silicon (amorphous)	PECVD	10^{-3}–10^{-1}	25–425	SiH_4, Si_2H_6, SiF_4, PH_3, B_2H_6, AsH_3	Plasma (rf, dc, microwave)	Typical process for thin-film transistor fabrication	h
Silicon dioxide / silicon nitride	LPCVD	0.5–2.0	400–700	$Si(C_2OH_5)_4$, SiH_4, O_2, NO_2, NH_3	Thermal	Commercial CVD oxide/nitride growth processes	i
Silicon dioxide/ silicon nitride	PECVD	0.1–1.0	80–400	$Si(C_2OH_5)_4$, SiH_4, O_2, NO_2, NH_3	Plasma (rf, dc, microwave)	Low-temperature nitride/oxide CVD processes	j
Silicon dioxide/ silicon nitride	LPCVD	0.5–2.0	400–700	$Si(C_2OH_5)_4$, SiH_4, O_2, NO_2 NH_3	Thermal	Commercial CVD oxide/nitride growth processes	i
GaAs, GaAlAs, InP, ZnSe, CdS, CdTe	Metal-organic CVD (MOCVD/OMCVD)	10–760	400–800	$Ga(CH_3)_3$, $In(CH_3)_3$, $Al(CH_3)_3$, $Ga(C_2H_5)_3$, $In(C_2H_5)_3$, $Al(C_2H_5)_3$, PH_3, AsH_3, $Zn(CH_3)_2$...	Thermal	Commercial compound semiconductor device and substrate fabrication	k

(Continued).

Table 2. Materials commonly prepared by chemical vapor deposition. (Continued.)

Material deposited	Growth technique	Deposition pressure (Torr)	Deposition temperature (°C)	Precursor	Excitation	Comments	Reference
GaAs, GaAlAs, InP, ZnSe, CdS, CdTe	Atomic-layer epitaxy (ALE)	10^{-6}–760	200–700	$Ga(CH_3)_3$, $In(CH_3)_3$, $Al(CH_3)_3$, $Ga(C_2H_5)_3$, $In(C_2H_5)_3$, $Al(C_2H_5)_3$, PH_3, AsH_3, $GaCl_3$	Thermal	Self-terminating monolayer by monolayer growth	l
Aluminum, copper, gold	Metal-organic CVD (MOCVD)	1–760	150–500	$Al(C_4H_9)_3$, $AlH_3\text{-}N\text{-}(CH_3)_3$, $Cu(C_6H_7O_2)_2$, $Cu(C_5F_6HO_2)_2$, $Au(C_5F_6HO_2)(CH_3)_2$, $Cu(C_5H_5)P(C_2H_5)_3$	Thermal or photolytic	Low-temperature metal deposition	m
Tungsten, tungsten silicide	LPCVD, APCVD	1–760	200–400	WF_6, $W(CO)_6$, SiH_4, Si_2H_6	Thermal	Commercial deposition of tungsten for microelectronic applications.	n

---

[a] Hammond, 1988; Mayer and Shea, 1964; Eversteijn, 1971; Bloem and Classen, 1980; Beers and Bloem, 1982.
[b] Ogirima *et al.*, 1977; Hammond, 1988.
[c] Yew *et al.*, 1987; Donahue and Reif, 1985.
[d] Gibbons *et al.*, 1985; Garone *et al.*, 1990.
[e] Meyerson *et al.*, 1988a; Meyerson, 1986; Greve and Racanelli, 1990.
[f] Yamazaki *et al.*, 1984; Yamazaki *et al.*, 1990.
[g] Hitchman, 1990; Brekel and Bollen, 1981; Meyerson and Olbricht, 1984.
[h] Kanoh *et al.*, 1990; Hamaya *et al.*, 1990; Hishikawa *et al.*, 1991; Kushner, 1988.
[i] Rojas *et al.*, 1990; Hitchman, 1990; Matsumura, 1989; Kanoh *et al.*, 1990.
[j] Kamins and Chiang, 1982; Emesh *et al.*, 1989; Kiermasz and Beekman, 1990; Pai *et al.*, 1990; Suzuki *et al.*, 1989; Manabe and Mitsuyu, 1989.
[k] Chen *et al.*, 1989; Bose *et al.*, 1990; Bhat *et al.*, 1990; Mullins *et al.*, 1990; Kuech *et al.*, 1984.
[l] Watanabe *et al.*, 1990; Ohta *et al.*, 1990; Hayafuji *et al.*,; Chiu *et al.*, 1990; Jin *et al.*, 1990; Bedair, 1990.
[m] Saxena and Luehrs, 1990; Beach *et al.*, 1990; Markwalder *et al.*, 1989; Oehr and Suhr, 1988; Kodas *et al.*, 1988; Baum and Jones, 1985.
[n] Gorowitz *et al.*, 1987; Kobayishi *et al.*, 1991; Dobkin *et al.*, 1990.

Torr, with a remarkable improvement (Sedgewick *et al.*, 1991) in film quality at a given growth temperature.

3.2 Reduced-Pressure CVD (RPCVD)

Reduced-pressure CVD (RPCVD) is a term reserved primarily for the subatmospheric-pressure deposition of epitaxial silicon, as shown schematically in Fig. 3. RPCVD is identical to APCVD, other than in the utilization of moderate vacuum (10–80 Torr) to sharpen dopant transition widths in films deposited upon doped substrates. The use of high temperatures and resultant high growth rates are common to both techniques. In the MOCVD (metal-organic CVD) adaptation of this method, it has been shown that the reduction of operating pressures allows the fabrication of layered III–V superlattices, as well as lasers, which have overall quality approaching that found by earlier beam techniques such as MBE.

3.3 Low-Pressure CVD (LPCVD)

Low-pressure CVD (LPCVD) is one of the most versatile film growth methods. Widely used throughout silicon technology, this method operates at significantly lower pressure ($0.2 \leqslant P \leqslant 2.0$ Torr) and temperature ranges ($150 \leqslant T \leqslant 650$ °C) than APCVD or RPCVD. LPCVD is most commonly utilized to deposit noncrystalline layers such as silicon dioxide, silicon nitride, metals, polycrystalline silicon, etc. This method is favored for these applications by virtue of the coatings being both uniform and conformal when deposited. The level of vacuum employed here is moderate, being of order 1.0 Torr, requiring in general a combination Roots blower and mechanical pump to handle the gas loads ($\simeq 100$ std. $\mathrm{cm}^3/\mathrm{min}$) at these pressures.

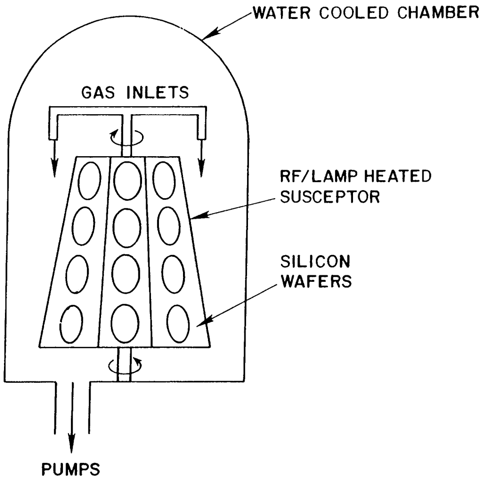

FIG. 3. A typical reduced-pressure CVD reactor as employed in silicon epitaxy or GaAs growth. The rotation of the growth susceptor, gas inlet, and in some cases the individual wafers themselves aids in attaining uniform film thickness and chemical composition.

3.4 CVD in Molecular Flow

There are now several CVD growth methods which operate in this low-pressure regime which accords them numerous desirable characteristics, and such processes constitute the most recent entry into the field of CVD. Gas transport for these processes is by molecular flow. The elimination of molecular collisions in the gas phase leads to reduced reactant depletion by homogeneous chemistry, and, where gas–surface interactions are inefficient, can result in films of extraordinary thickness and compositional uniformity. The appropriate view of a reaction chamber operating in this regime is that of a bath, where reactant distribution is completely isotropic. Transition widths between sequentially grown layers, no longer limited by gas-phase memory effects, become abrupt on the atomic scale, making feasible the large-area production of quantum-scale devices. This method has been employed in the III–V material system during chemical beam epitaxy, where a beam of metal-organic precursor molecules is made to impinge upon a heated substrate in a UHV chamber. UHV/CVD, a batch growth process in the silicon system, also employs reactant transport in this flow regime, as seen in Fig. 4.

Atomic-layer epitaxy (ALE) may also be performed under these flow conditions, where a beam is used to create a self-terminating monolayer of material A on a surface, at which time material B is deposited over A to form another self-limiting monolayer of material.

Self-limiting growth processes are most frequently encountered in polar systems, such as GaAs. Under commonly employed growth conditions, an upper bound of one complete monolayer of arsenic may be adsorbed on a free gallium surface. Upon completion of the monolayer, arsenic no longer accumulates, as

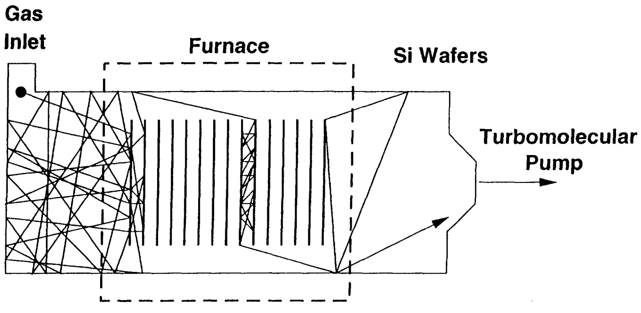

FIG. 4. Molecular transport in a batch-process reactor such as UHV/CVD. In that gaseous transport is to a good estimate similar to that of a particle in a box, concepts such as "upstream" and "downstream" are not applicable.

arsenic clusters are not stable at the temperatures employed in ALE. The gaseous source may then be exchanged for one bearing gallium, again allowing time for saturation coverage to be reached. In a similar way, other III–V species may be introduced with precision, so long as saturation coverage of the growth surface is achieved before the next gaseous source is introduced. This serves to produce artificially structured, highly ordered materials, and is applied most often to III-V systems. In such apparatus, the preferred pumping method is the use of a turbomolecular pump–mechanical pump combination. In processing at such low pressures, the choice of pumping strategy is critical for the following reason.

In the millitorr pressure regime, concepts such as upstream and downstream in a reactor fail. Any contaminant reaching the reaction vessel is uniformly distributed. There is no longer a laminar flow of gas, as in APCVD, which serves to prevent the backstreaming of contaminants. Thus, the maintenance of an appropriately clean reactor environment is determined solely by the nature of the vacuum and gas-handling equipment employed. Turbomolecular pumps, ideally suited to this application, are opaque to the backstreaming of contaminants such as hydrocarbon pump oil, and are unreactive, as pumping is purely by mechanical rather than chemical means as in an ion pump. The use of cryopumping with nominally toxic and explosive growth precursors is hazardous in that condensing out such species accumulates quantities of order many grams over time, which are then released rapidly during regeneration cycles. Thus, great care must be taken when working with CVD processes in this operating regime.

3.5 Plasma-Enhanced CVD

Plasma-enhanced CVD (PECVD) (see Fig. 5) is widely employed where a decoupling of substrate temperature from reactant temperature or energy is required. So that device dimensions may be reduced, semiconductor technologies require that the thermal budget of any given device fabrication scheme be minimized, and thus the trend toward the use of excitation schemes that allow a reduction of substrate temperatures. Plasma excitation for film growth has found application in virtually all aspects of semiconductor growth, insulator growth, metal deposition, and hardcoating deposition. Films are not generally as conformal or uniform as in thermally driven CVD processes, but engineering of discharge geometries has brought these issues to ade-

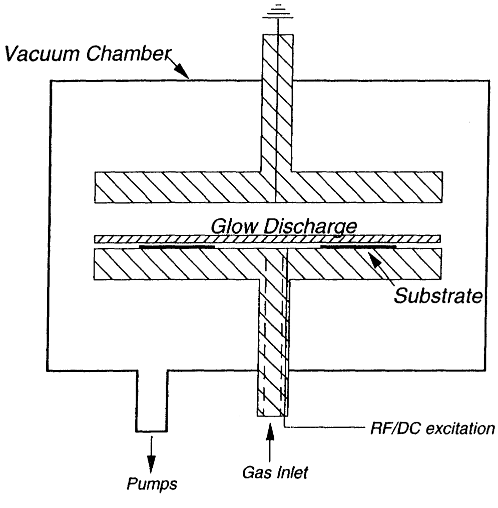

FIG. 5. Parallel-plate rf/dc plasma-enhanced CVD apparatus. Note that although wafers are shown on the powered electrode, this is not universally the case. Further, a dc bias voltage is frequently applied in addition to the rf excitation signal.

quate levels of perfection. A liability of plasma-enhanced methods is that the energies employed are adequate both to fragment the growth precursor, and to sputter impurities from chamber surfaces. Various means have been developed to minimize this effect, one being the confinement of a plasma via electron cyclotron resonance methods, or the formation of a plasma in a clean, remote chamber, from which a beam of excited species is extracted. In spite of some difficulties, PECVD is prevalent throughout the research environment, and in numerous commercial applications as well.

3.6 Photolytic CVD

Photolytic CVD (see Fig. 6) is an example of an alternative strategy for decoupling precursor temperature or energy from the temperature of the substrate. In contrast to PECVD,this method enables one to localize film growth, a trait utilized in most of the applications of this method. Direct photolysis of a source species, such as a metal-organic precursor, is frequently employed in industry to perform local repairs of circuit wiring or lithography masks. Although not well suited to large-area film growth, the prospect of directly writing a circuit by this method maintains a significant level of interest in this area. There are several methods of photo-CVD which employ large- area, high-fluence light sources to perform blanket deposition on wafer surfaces, but the control of film uniformity, and difficulties in eliminating undesirable film growth on chamber optical ports, have slowed the widespread implementation of this method.

4. MODELS—EQUILIBRIUM THERMODYNAMICS VERSUS KINETICS

Modeling of CVD is an active field, where general, accurate, predictive models are noticeably absent. The inability to generalize modeling of CVD comes about through the enormous complexity of the chemistries employed, coupled with the lack of fundamental data required to complete any given model. Moreover, before one even attempts to model a given CVD process, a basic choice must be made as to the approach to be taken. In the limiting case of a closed system, at low growth rates and high temperatures, it is frequently assumed that such a system is in a near-equilibrium state and may be characterized using thermodynamic equilibrium considerations. In the opposed limit of low-temperature, high-flow systems, it is often assumed that reaction kinetics are dominant (rate limiting). Both strategies have been employed in characterizing the CVD process. Hess *et al.* (1985) have recently reviewed this issue.

4.1 Equilibrium Thermodynamics

Equilibrium thermodynamics is frequently employed in the analysis, at least qualitatively, of a CVD process. It is attractive for its simplicity, as details of the exact growth mechanism drop out of the calculation. Rather, one must identify the number n_i of each atomic and molecular species present, as well as the chemical potential μ_i of each species. The Gibbs free energy of the system at a single operating point with well-defined temperature and pressure is simply

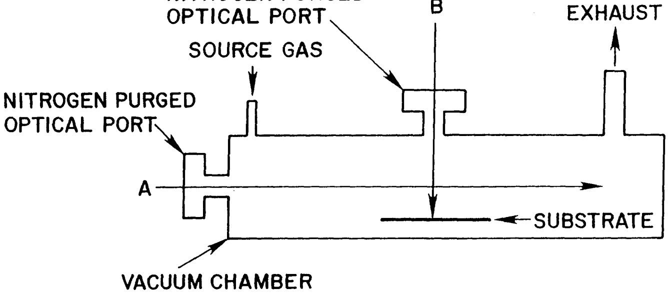

FIG. 6. Photo-CVD reactor shown in two configurations. Optical input via "A" is employed to produce gas-phase reactive intermediates for film growth over large areas. Input "B" is employed for local photothermal or photolytic CVD, allowing direct writing of features upon the substrate.

$$G = \sum_{i=1}^{i=N} n_i \mu_i .$$

Minimization of G for a given set of operating conditions, constrained by the initial values for species present, allows the calculation of the final distribution of species after equilibration with a new environment. The ability to compute the equilibrium distribution of chemical species in the reactor environment allows one to manipulate film growth in the following manner.

In the instance of silicon epitaxy, such a calculation has been performed by Bloem and Giling (1985), for the case of deposition in the silicon:hydrogen:chloride system. Figure 7 displays the calculated distribution of gaseous species present in equilibrium with solid Si over a span of temperatures, holding the Cl/H ratio fixed. From this data, one can compute the total solubility of Si in this gaseous mix when in contact with a silicon surface, shown in Fig. 8. The solubility of Si in the gas phase is limited by the atom count of silicon that may be bound in the computed distribution of hydride and halide species. As an example, for a fixed input of chlorine, conditions favoring the formation of hexachlorodisilane (Si_2Cl_6), at a resultant Si:Cl ratio of 1:3, permit the gaseous source to support a higher silicon fraction than conditions favoring the formation of silicon tetrachloride ($SiCl_4$), at a Si:Cl ratio of 1:4. Thus, using temperature to alter the distribution of silicon-bearing halide species, one may control film growth under near-equilibrium conditions. For example, when the Si-H-Cl chemical system is at equilibrium at 1300 °C, and then flowed into a reactor region held at 1100 °C, the Si content of the gaseous mix must decrease (see Fig. 8) to remain in thermodynamic equilibrium. The system therefore returns to equilibrium by reducing total silicon content by the deposition of Si onto any preexisting Si surfaces. The degree of Si supersaturation that is induced by the rapid temperature drop determines the resulting film growth rate. If this gas mixture were now allowed to equilibrate at 1100 °C, a subsequent reduction in temperature from 1100 to 800 °C would result in the etching of a silicon substrate placed in this flow. This is seen in Fig. 8, where the solubility of Si rises again at 800 °C. This ability to create controllably a supersaturation of Si in the gas phase has been widely employed in conducting selective Si epitaxy. Surfaces other than silicon, such as silicon dioxide, require a finite, non-

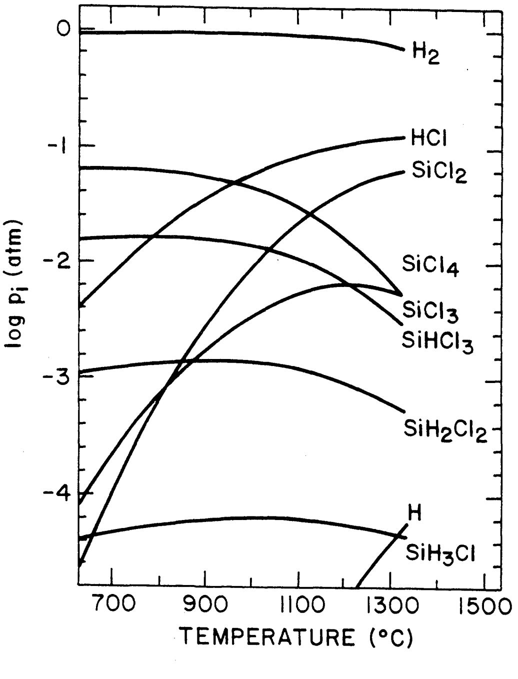

FIG. 7. Calculated equilibrium distribution of gaseous species during APCVD of epitaxial silicon using silicon chloride source species. Data after Bloem and Giling (1985).

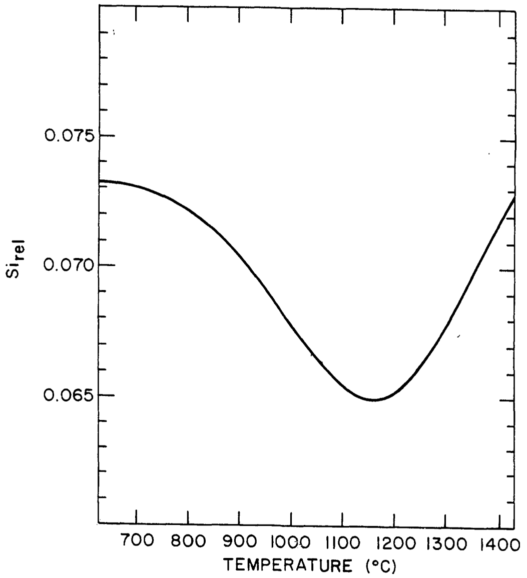

FIG. 8. Relative Si solubility in the gas phase as a function of temperature (see the text). After Bloem and Giling (1985).

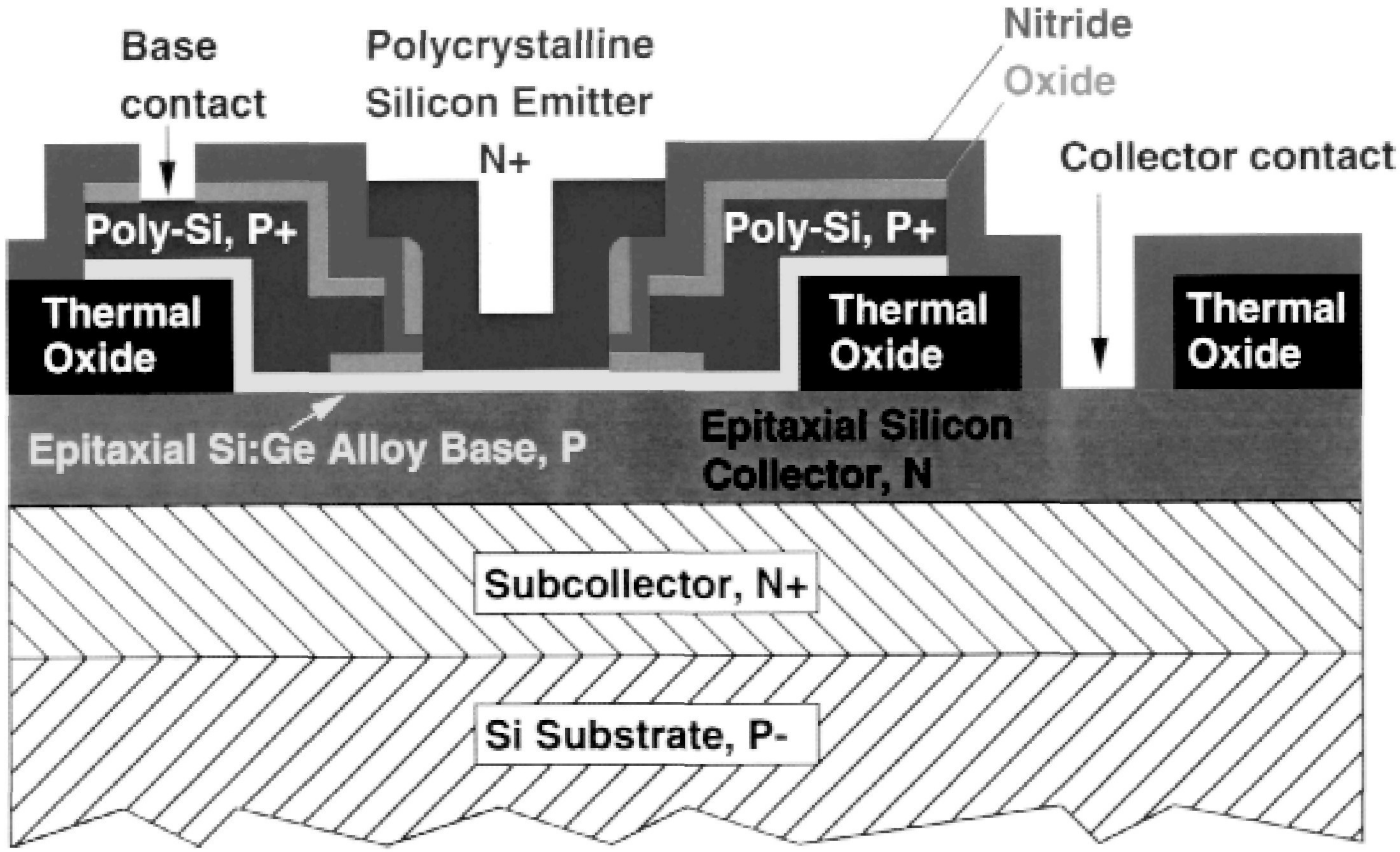
Base
contact
Polycrystalline
Silicon Emitter
N+
Nitride
Oxide
Collector contact
Poly-Si, P+
Poly-Si, P+
Thermal
Oxide
Thermal
Oxide
Thermal
Oxide
Epitaxial Si:Ge Alloy Base, P
Epitaxial Silicon
Collector, N
Subcollector, N+
Si Substrate, P-

Vol. 4 – PLATE 1. A cross-sectional view of a high-performance epitaxial base heterojunction bipolar transistor. All features other than those in black and white are formed using various CVD processes.

zero supersaturation of silicon such that film nucleation and growth proceeds. By contrast, silicon growth on silicon has no minimum requirement on supersaturation for growth to proceed. It is therefore possible to induce selective growth of silicon on silicon for small supersaturations, while silicon dioxide present in the same environment remains free of film. The trends predicted by equilibrium thermodynamic considerations as given in this example are indeed observed, but quantitative agreement between experiment and theory is rare. Further, a cautionary note is due here.

Another system where one may attempt to apply considerations as above would be the growth of III–V materials—e.g., GaAs. One of the difficulties in the growth of GaAs by MOCVD is the unintended incorporation of carbon into GaAs from the Ga precursor molecule, $Ga(CH_3)_3$. If one replaces trimethylgallium with triethylgallium, $Ga(C_2H_5)_3$, the ratio of carbon to gallium increases, and, to first order, one would expect an increase in film carbon content. The opposite occurs, as the use of triethylgallium greatly reduces the formation of surface-active methyl radicals, CH_3, and thus eliminates what was found to be an efficient source for carbon in GaAs. Using only equilibrium thermodynamic considerations to model this chemistry, even qualitatively, would predict the wrong result. This points to the power of having available a full chemical description of a reactive system, so that one may explicitly look at the dynamics of growth chemistry, rather than the equilibrated state, which may not in fact be attained in the time that a mixture is present within the reactor environment.

4.2 Kinetics-Based Reactor Modeling

In contrast to equilibrium-based models, models based upon the detailed chemistry of growth are by definition quantitative, and as perfect as one's choice of chemical model and one's knowledge of all relevant reaction rates. For precisely this reason, models based upon kinetic considerations are for the most part qualitative, and are not suited to general, predictive use. This seeming contradiction results from our lack of knowledge at present as to the exact reaction pathways and the relevant reaction rates for most CVD systems of interest.

The methodology of kinetic modeling requires that one set forth a chemical pathway by which a film is to be deposited. This must include all those reactions postulated to be relevant to film growth, both heterogeneous and homogeneous. The form such a chemical system would take for homogeneous chemistry during Si epitaxy from silane (SiH_4) could be

$$SiH_4 \rightarrow SiH_2 + H_2, \tag{1}$$

$$SiH_4 + SiH_2 \rightarrow Si_2H_6, \tag{2}$$

$$SiH_2 + Si_xH_{2(x+1)} \rightarrow \rightarrow \rightarrow Si_{x+1}H_{2(x+2)} (x \geqslant 2) \tag{3}$$

$$\cdot$$
$$\cdot$$
$$\cdot$$

The heterogeneous component would be described in detail as well, with adsorption (4), heterogeneous decomposition (5), and product evolution (6) represented as follows, where $*$ denotes an active adsite:

$$SiH_4(\text{gas}) + * \rightarrow SiH_4(\text{ads.}), \tag{4}$$

$$SiH_4(\text{ads.}) \rightarrow SiH_x(\text{ads.}) + (4-x)H(\text{ads.}), \tag{5}$$

$$SiH_x(\text{ads}) + (4-x)H(\text{ads}) \rightarrow Si(\text{cr}) + 2H_2(\text{gas}) + * \, (x \leqslant 3). \tag{6}$$

The six reactions enumerated here are a small fraction of the several hundred that need to be and have indeed been considered as part of such a model. It is found, via sensitivity analysis, that many reactions are superfluous to the growth process, and can be eliminated. However, this process relies upon one's making "reasonable" guesses as to the rate constants, forward and reverse, for each of the reactions enumerated above. Thus, though chemical kinetic modeling is potentially an exact science, it suffers from the lack of an adequate database for chemistries of general interest. In that the silane system is among the most intensely studied, Coltrin *et al.* (1984) and Jensen *et al.* (1989) have carried out the full finite-element treatment of silane-based CVD, solving a coupled hydrodynamic and chemical code from the system of reactions as above.

A more common application of kinetic modeling to CVD has been the study of rate-limiting phenomena during film growth. In this simple application of kinetic modeling, one prepares a skeletal series of growth-related chemical reactions, similar to the framework above, but assuming foreknowledge of the active growth species (e.g., the parent molecule, a homogeneously generated radical species, etc). Using what is assumed to be an accurate chemical model, one then derives the functional dependence of film growth rate upon a series of system parameters, such as reactant partial pressure, temperature, and total pressure. This approach will highlight those reactions that are rate-limiting bottlenecks in the film growth process, and has the longest history of application (Bloem and Claasen, 1980) in the analysis of CVD. Upon completion of a model, validation is by experimental verification of the projected dependence of growth rate upon process parameters. However, it must be understood that this does not prove the model to be correct, in that there is no assurance that the model assumed is unique. In spite of this considerable flaw, such early forms of kinetic modeling were frequently useful as qualitative guides as to the optimization of silicon-hydride- and -halide-based CVD.

5. APPLICATIONS

The applications of CVD are as varied as all those applications in which thin-film deposition is required. The current view of where CVD is the method of choice for film deposition is for applications that are mature, where large-area or high-throughput thin-film fabrication is required for commercial products. This is in contrast to beam techniques, notably MBE, where the emphasis is on research and early-stage development of materials and structures. Though this broad assessment has been true in the past, numerous recent applications of the CVD method have shown that the earlier designation of CVD as more of an industrial than a research tool is changing rapidly.

5.1 Passive Circuit Elements and Structures

In technology, CVD is employed to the exclusion of other growth methods in those instances where conformal coatings are required. An extreme example of the conformal nature of CVD growth is shown in Fig. 9. Shown is an advanced trench capacitor, the heart of a standard computer memory chip, and among the most common electronic components in the semiconductor industry. The aspect ratio of the trench shown is 16:1 depth to width. Fabrication consists of the use of anisotropic etching to form a periodic array of trenches in a silicon substrate. The walls of the trench are first thermally oxidized, then further insulated by the deposition of a thin layer of CVD silicon nitride. The trench is then filled conformally by the LPCVD growth of polycrystalline silicon, resulting in the trench capacitor structure shown here. The utilization of the inherent conformality of the CVD process is pervasive. Other applications are the CVD of hard coatings on tool bits employed in high-speed machining, scratch-resistant or antireflection coatings on complex optics, and the use of chemical vapor infiltration (filling of voids in woven carbon fibers in similar fashion to trench filling) to produce lightweight carbon composite parts such as disk-brake rotors.

5.2 Active Circuit Elements and Device Structures

CVD is most widely employed for the conformal growth of passive (e.g., dielectric) device and/or structure components as above. However, CVD is making rapid advances in the deposition of critical electrically active layers. MOCVD is now making inroads into the fabrication of III–V devices, and is of increasing importance at present as III–V (GaAs/GaAlAs) devices and integrated circuits are being commercialized. A more common application of MOCVD is for the fabrication of optically active devices, primarily LEDs and lasers, again in the commercial environment.

CVD, as applied in silicon device technology, has recently come into its own as a method by which active device regions themselves are formed. In existing silicon technology, CVD has long been employed as a means to prepare highly uniform substrate-like material upon which active device regions were defined by either diffusion or ion implantation. For the growth of substrate-like blanket epitaxial layers, straightforward high-temperature epitaxial growth methods were employed—e.g., APCVD and RPCVD. In the past five years, there has been a virtual revo-

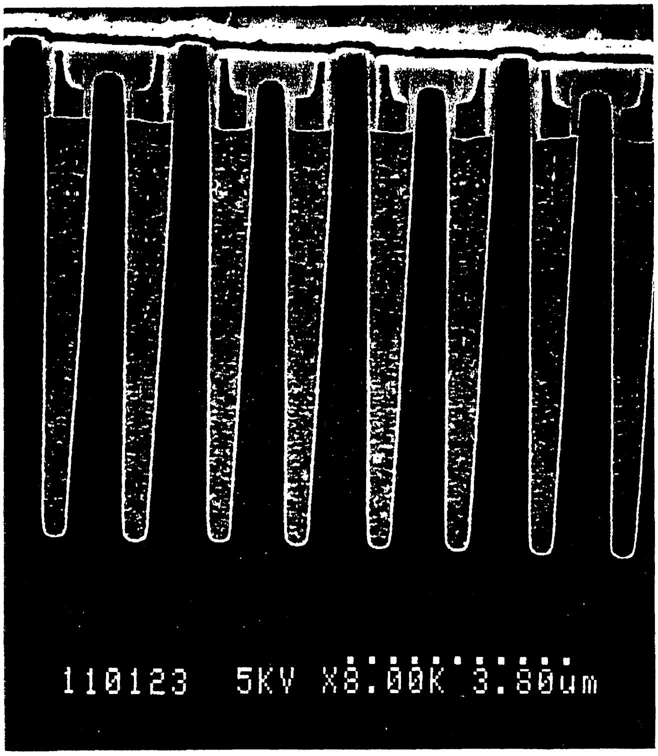

FIG. 9. The conformal nature of the CVD process was employed to fill the high-aspect-ratio features of this deep trench capacitor, the storage node in a modern dynamic random-access memory (DRAM) cell. These trenches are lined with a CVD silicon oxide/silicon nitride dual dielectric, and then filled by the CVD of polycrystalline silicon. The center-to-center spacing between neighboring pairs of trenches is 2.8 μm.

lution in silicon-based devices as a consequence of the application of CVD to this field.

With the advent of several methods of low-temperature silicon epitaxy as enumerated in Tables 1 and 2, new device designs have been implemented, employing *in situ* dopant incorporation, and utilizing as-deposited layers as the active regions of various bipolar and FET structures. The freedom of device design, and the precise control made possible, have rejuvenated the field of Si-based device design. Plate 1 is an example of a Si:Ge-based heterojunction bipolar transistor, and most noteworthy is that all colored regions of the plate represent layers prepared by CVD. Virtually all active and passive device components above the substrate are prepared by CVD, primarily because of the precision possible and the low thermal budget associated with those processes. By working at temperatures below those associated with dopant diffusion, $T \leqslant 700$ °C, the integrity of deposited dopant and alloy profiles is maintained throughout device fabrication.

As a result of advances in device design and materials deposition techniques, presumed limitations on the performance of Si-based devices have been overcome through the application of many of the band-structure tailoring methods applied earlier in the III–V system. One example of this trend was the recent fabrication, by UHV/CVD, of high-performance Si:Ge heterojunction bipolar transistors, having a unity-gain cutoff frequency of 75 GHz at room temperature (Patton *et al.*, 1990), and 94 GHz at 77 K (Crabbe *et al.*, 1990). In these devices, the nominally implanted base region of a bipolar transistor was replaced by a boron-doped, tailored Ge profile in Si. The Ge profile was tailored so as to produce a conduction band grading to accelerate electrons across the base region of the device, resulting in a doubling of previously observed Si device speed, and exceeding the projected limits of Si-based device technology.

5.3 Fundamental Materials Studies

At the forefront of the applications of CVD are those applications where film preparation is being performed (or attempted) for the creation of artificially structured materials.

The intent here is to generate, via atomic layering, materials having unique electronic and optical properties. In the least demanding (but hardly trivial) implementation, quantum-effect devices are fabricated, resonant tunneling diodes being an example (Chang *et al.*, 1983). Dimensional control in such structures must be on the order of a few atomic lengths along one dimension, film thickness in general. This field, heretofore dominated by beam methods (MBE), is now being broadened by the application of CVD methods, showing equal or better dimensional and compositional control in some instances. In the III–V system, atomic-layer epitaxy is employed to prepare structures with atomic-layer control along the growth axis, allowing the fabrication of artificially structured materials. In the case of silicon, recent work reporting the first electron-based double-barrier resonant tunneling diodes fabricated in the silicon system was CVD based, and room-temperature operation of these quantum-effect devices was demonstrated (Ismail *et al.*, 1991).

In the most aggressive application of materials fabrication, one attempts to alter the fundamental band structure of a given material system by appropriate interleaving of constituent atomic layers. This is going a step beyond quantum electronics, in that the operation of a given quantum-effect-based device is governed by the local electronic structure of the host material. Thus, one attempts to create a unique function in known device structures by alteration of the fundamental nature of the material system employed. An example is ongoing work where strain-symmetrized Si and Ge layers are interleaved in an effort to produce "zone folding," the artificial alignment of the valence-band maxima with the conduction-band minima, such that the material becomes a direct-gap semiconductor. The modification of this group-IV alloy system to produce structures with efficient optical activity would open the way for Si-based integrated optoelectronics, in the manner of the far more successful III–V-based systems. In that the outcome of this effort is far from clear, new growth techniques, CVD among them, are being brought to bear to facilitate the preparation of such demanding structures.

6. SUMMARY

What has been discussed in this article is an overview of a field that is evolving at an extremely rapid pace. Although CVD has long been employed in a variety of technical applications, it is now being widely employed within the scientific community as well. Recent advances in the field have allowed atomic-length scale control of materials properties with both abruptness and dynamic range heretofore unavailable, resulting in materials, structures, and devices having unique physical and electronic structure. There are individual articles in the *Encyclopedia* which focus on specific CVD methods discussed in broad terms above, and it is recommended that this text serve as an entry point to the field of chemical vapor deposition.

GLOSSARY

Adsite: Adsorption site at which an impinging species may bind to a surface.

Band Structure: The fundamental electronic structure of a given material, determining the transport properties of carriers in that material.

HBT: Heterojunction bipolar transistor, a bipolar transistor utilizing dissimilar semiconductor materials at its emitter-base or base-collector junctions.

Heterogeneous Chemistry: Reactions occurring at a gas/surface interface.

Homogeneous Chemistry: Reactions occurring in a single reactant phase, that being the gas phase as used in this article.

LED: Light-emitting diode.

Photolysis: Decomposition induced by the electronic energy imparted to a species by incident light.

Radical: A molecular fragment not fully coordinated, which is therefore highly reactive.

RTD: Resonant tunneling diode.

Susceptor: Energy-absorbing substrate holder employed inside a growth chamber to couple to an external energy source.

UHV: Ultrahigh vacuum, generally defined as pressure less than or equal to 10^{-9} Torr.

List of Works Cited

Batey, J., Tierney, E. (1986), *J. Appl. Phys.* **60**, 3136–3146.

Baum, T. H., Jones, C. R. (1985), *Appl. Phys. Lett.* **47**, 538–540.

Beach, D. B., LeGoues, F. K., Hu, T.J. (1990), *Chem. Mater.* **2**, 216–219.

Bedair, S. M., (1990), *Acta Polytech. Scand. Chem. Technol. Metall. Ser.* **Ch195**, 17–37.

Beers, A. M., Bloem, J. (1982), *Appl. Phys. Lett.* **41**, 153.

Bhat, R., Kapon, E., Werner, J., Hwang, D. M., Stoffel, N. G., Koza, M. A. (1990), *Appl. Phys. Lett.* **56**, 863–865.

Bloem, J., Claasen, W. A. (1980), *J. Cryst. Growth* **49**, 435–444.

Bloem, J., Giling, L. J. (1985), *VLSI Electron. Microstruct. Sci.* **12**, 89–139.

Bose, S. S., Szafranek, I., Kim, M. H., Stillman, G. E. (1990), *J. Appl. Phys.* **56**, 752–754.

Brekel, C. H., Bollen, L. J. (1981), *J. Cryst. Growth* **54**, 310–322.

Chang, L. L., Esaki, L., Tsu, R. (1983), *Appl. Phys. Lett.* **43**, 588.

Chen, W. K., Chen, J. C., Anthony, L., Liu, P. L. (1989), *Appl. Phys. Lett.* **55**, 987–989.

Chiu, T. H., Cunningham, J. E., Robertson., Jr., A., Malm, D. L. (1990), *J. Cryst. Growth* **105**, 155–161.

Coltrin, M. E., Kee, R. J., Miller, J. A. (1984), *J. Electrochem. Soc.* **131**, 425–434.

Crabbe, E. F., Patton, G. L., Stork, J., Meyerson, B. S., Sun, J. (1990), *1990 IEDM Tech. Dig.* **17**.

Dobkin, D., Bartholomew, L., McDaniel, G., De-Dontney, J. (1990), *J. Electrochem. Soc.* **137**, 1623–1626.

Donahue, T. J., Reif, R. (1985), *J. Appl. Phys.* **57**, 2757.

Emesh, I. T., Dasti, G., Mercier, J. S., Leung, P. (1989), *J. Electrochem. Soc.* **136**, 3404–3408.

Eversteijn, C. (1971), *Phillips Res. Repts.* **26**, R757.

Eversteijn, F. C., Severin, P. J., Brekel, C. H., Peek, H. L. (1970), *J. Electrochem. Soc.* **117**, 925.

Fujino, K., Nishimoto, Y., Tokumasu N., Maeda, K. (1990), *J. Electrochem. Soc.* **137**, 2883–2887.

Garone, P. M., Sturm, J. C., Schwartz, P. V., Schwarz, S. A., Wilkens, B. J. (1990), *Appl. Phys. Lett.* **56**, 1275–1277.

Gibbons, J., Gronet, C., Williams, K. (1985), *Appl. Phys. Lett.* **47**, 721–723.

Giunta, C. J., Chapple-Sokol, J. D., Gordon, R. G. (1990), *J. Electrochem. Soc.* **137**, 3237–3253.

Gorowitz, B., Wilson, R. H., Gorczyca, T. B. (1987), *Solid State Technol.* **30**, 97–103.

Greve, D. W., Racanelli, M. (1990), *J. Vac. Sci. Technol. B* **8**, 511–515.

Hamaya, H., Murai, H., Kobayashi, K. (1990), *J. Appl. Phys.* **67**, 1356–1360.

Hammond, M. L. (1988), *Solid State Technol.* **31**, 159–164.

Hayafuji, N., Miyashita, M., Kumabe, H., Murotani, T. (1990), *J. Cryst. Growth* **106**, 421–425.

Hess, D. W., Jensen, K. F., Anderson, T. J. (1985), *Rev. Chem. Eng.* **3**, 99–186.

Hishikawa, Y., Tsuge, S., Nakamura, N., Tsuda, S., Nakano, S., Kuwano, Y. (1991), *J. Appl. Phys.* **69**, 508–510.

Hitchman, M. L. (1990), *Vacuum* **41**, 880–884.

Hitchman, M. L., Jensen, K. F. (1991) *Chemical Vapor Deposition*, New York: Academic.

Ismail, K., Meyerson, B. S., Wang, P. J. (1991), *Appl. Phys. Lett.* **59**, 973–975.

Jasinski, J. M., Meyerson, B. S., Nguyen, T. N. (1987), *J. Appl. Phys.* **61**, 431–433.

Jensen, K. F., Fotiadis, D. I., Mountziaris, T. J. (1989), *Proc. Mater. Res. Soc.* **145**, 107.

Jin, Y., Kobayashi, R., Fujii, K., Hasegawa, F. (1990), *Jpn. J. Appl. Phys.* **29**, L1350–1352.

Kamins, T. I., Chiang, K. L. (1982), *J. Electro-Chem. Soc.* **129**, 2331–2335.

Kanoh, H., Sugiura, O., Breddels, P. A., Matsumura, M. (1990), *IEEE Electron. Dev. Lett.* **11**, 258–260.

Kiermasz, A., Beekman, K. (1990), *Semiconduc. Int.* **13**, 108–111.

Kim, S. S., Tsu, D. V., Lucovski, G. (1987), *J. Vac. Sci. Technol. A* **6**, 1740–1744.

Kobayashi, N., Goto, H., Suzuki, M. (1991), *J. Appl. Phys.* **69**, 1013–1019.

Kodas, T. T., Baum, T. H., Comita, P. B. (1988), *J. Cryst. Growth* **87**, 378–382.

Kuech, T. F., Meyerson, B. S., Veuhoff, E. (1984), *Appl. Phys. Lett.* **44**, 986–988.

Kushner, M. J. (1988), *J. Appl. Phys.* **63**, 2532–2551.

Lucovsky, G., Richard, P. D., Tsu, D. V., Lin, S. Y., Markunas, R. J. (1986), *J. Vac. Sci. Technol. A* **4**, 681–688.

Manabe, Y., Mitsuyu, T. (1989), *J. Appl. Phys.* **66**, 2475–2480.

Markwalder, B., Widmer, M., Braichotte, D. van-den-Bergh, H. (1989), *J. Appl. Phys.* **65**, 2470–2474.

Matsumura, H. (1989), *J. Appl. Phys.* **66**, 3612–3617.

Mayer, E., Shea, D. E. (1991), *J. Electrochem. Soc.* **111**, 550.

Meyerson, B. S. (1986), *Appl. Phys. Lett.* **48**, 797.

Meyerson, B., Uram, K., LeGoues, F. (1988), *Appl. Phys. Lett.* **53**, 2555–2557.

Meyerson, B. S., Olbricht, W. L. (1984), *J. Electrochem. Soc.* **131**, 2361.

Meyerson, B. S., LeGoues, F. K., Nguyen, T. N., Harame, D. L. (1987), *Appl. Phys. Lett.* **50**, 113.

Meyerson, B. S., Uram, K. J., LeGoues, F. K. (1988), *Appl. Phys. Lett.* **53**, 2555.

Mullins, J. T., Clifton, P. A., Brown, P. D., Brinkman, A. W., Woods, J. (1990), *J. Cryst. Growth* **101**, 100–104.

Oehr, C., Suhr, H. (1988), *Appl. Phys. A* **45**, 151–154.

Ogirima, M., Saida, H., Suzuki, M., Maki, M. (1977), *J. Electrochem. Soc.* **124**, 903.

Ohta, S. I., Kobayashi, S., Kaneko, F., Kashiro, K. I. (1990), *J. Cryst. Growth* **106**, 166–174.

Pai, C. S., Chang, C. P., Baiocchi, F. A., Swiderski, J. (1990), *J. Appl. Phys.* **68**, 2442–2449.

Patton, G. L., Comfort, J. H., Meyerson, B. S., Crabbe, E. F., Scilla, G. J., DeFresart, E., Stork, J. M., Sun, J. Y., Harame, D. L., Burghartz, J. (1990), *Electron. Dev. Lett.* **11**, 171.

Reif, R. (1982), *J. Electrochem. Soc.* **129**, 1122–1128.

Rojas, S., Modelli, A., Wu, A., Borghesi, A., Pivac, B. (I990), *J. Vac. Sci. Technol. B* **8**, 1177–1184.

Saxena, A. N., Luehrs, F. (1990), *J. Electron. Mater.* **19**, 271–276.

Sedgewick, T. O., Agnello, P. D., Ngoc, D. N., Kuan, T. S., Scilla, G. (1991), *Appl. Phys. Lett.* **58**, 1896–1898.

Suzuki, N., Yoshikawa, T., Masu, K., Tsubouchi, K., Mikoshiba, N. (1989), *Jpn. J. Appl. Phys. 2 Lett.* **28**, L2316–2319.

Watanabe, A., Kamijoh, T., Hata, M., Isu, T., Katamaya, Y. (1990), *Vacuum* **41**, 965–967.

Wong, M., Saraswat, K. C. (1988), *IEEE Electron. Dev. Lett.* **9**, 582–584.

Yamazaki, T., Ito, T., Ishikawa, H. (1984), *1984 VLSI Symposium, Digest of Technical Papers*, Tokyo: Japan Society of Applied Physics, pp. 56–57.

Yamazaki, T., Watanabe, S., Ito, T. (1990), *J. Electrochem. Soc.* **137**, 313–318.

Yew, T. R., Comfort, J. H., Garverick, L. M., Burger, W. R., Reif, R. (1987), *Mater. Res. Soc. Symp. Proc.* **75**, 705–712.

CYBERNETICS

See INFORMATION SCIENCE

13
Cyclotrons

George H. Mackenzie, Paul W. Schmor, and H.R. Schneider

13.1 Introduction 405
13.2 Types of Cyclotron 406
13.2.1 Classic Cyclotron 406
13.2.2 Synchrocyclotron or Frequency-Modulated Cyclotron 410
13.2.3 The Sector-Focused, Isochronous, or AVF Cyclotron 411
13.2.4 The Microtron (Electron Cyclotron) 417
13.3 Design Considerations and Subsystems for an AVF Cyclotron 417
13.3.1 Beam Physics 417
13.3.1.1 Resonances 417
13.3.1.2 Electric Forces 418
13.3.1.3 Charge-Changing Phenomena 419
13.3.1.4 Computer Codes 420
13.3.2 Magnet 420
13.3.3 Accelerating Systems 421
13.3.4 Ion Sources for Cyclotrons 423
13.3.5 Extraction 426
13.3.6 Vacuum 427
13.3.7 Beam Instrumentation 428
13.3.8 Computer Control and Monitoring 429
13.4 Cyclotrons as Injectors or Post-accelerators 429
13.5 Applications of Cyclotrons 431
13.5.1 Interaction of Particle Beams with Matter 431
13.5.2 Basic Research in Nuclear Science* 434
13.5.3 Basic Research in Other Physical Science 435
13.5.4 Isotope Production 436
13.5.5 Particle Beams in Medicine 438
13.5.6 Elemental Analysis of Materials 439
13.5.7 Transmission Radiography 440
13.5.8 Wear and Corrosion 441
13.5.9 Materials Modification and Processing 441
13.6 Facility Design and Shielding 442

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim ISBN: 978-3-527-40691-3

Appendix 13.A 445
Description of a Typical Isotope-Production Cyclotron 445
Acknowledgment 447
Glossary 447
References 448
Further Reading 448

13.1
Introduction

The cyclotron is the first member of a class of cyclic particle accelerators that accelerate charged particles to high energies by repeated passage of the beam through one or more gaps across which a radio frequency (rf) voltage is maintained. It was conceived in 1929 by Lawrence, and the version built by Lawrence and Livingston at the University of California, Berkeley, campus in 1931 (Lawrence and Livingston, 1932) was the first machine to accelerate protons to an energy greater than 1 MeV. In a cyclotron, repeated traversal of the accelerating gap is achieved by the use of a DC magnetic field to deflect the particles in circular arcs whose radius increases as their energy increases. Operation of the Lawrence or classic cyclotron is based on the fact that if relativistic effects can be ignored, that is, when particle velocities are much less than the speed of light, the angular frequency of particle rotation is independent of energy and is determined only by the particle charge-to-mass ratio and the magnetic induction. So long as both are constant, the particle orbit frequency is constant. Acceleration is then possible if the oscillation frequency of the gap voltage is equal to the orbit frequency or is in some cases a harmonic of the orbit frequency.

At the time of its invention, the cyclotron significantly extended the range of particle energies available in the laboratory, and for many years this classic cyclotron, accelerating light atomic and molecular ions, was the mainstay of a number of nuclear research facilities. The restriction, however, to energies for which relativistic effects are negligible meant, for example, that proton energies were limited to about 15 MeV.

Over the years a number of improvements have been made to the original concept. Development of the synchrocyclotron in 1946 to overcome the nonrelativistic energy limitation of the classic cyclotron followed the enunciation of the principle of phase stability by Veksler (1944) and independently by McMillan (1945). This principle showed that stable bunches of particles could be accelerated to high energies by appropriate reduction of the frequency of the accelerating voltage in accordance with the reduction in orbit frequency due to the relativistic mass increase with increasing particle energy. Later in the 1950s an idea originally proposed by Thomas (1938), to employ a periodic variation of the magnetic field strength to provide a transverse focusing force on the beam, was revived. Such an azimuthally varying

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

field (AVF) eliminated the need for a small negative radial gradient in the magnetic field, for focusing in the classic cyclotron. This then permitted the magnetic field to increase with the radius to maintain a constant orbit frequency (isochronism), and continuous wave (CW) operation, to the relativistic energies achievable with synchrocyclotrons, but with much greater beam intensities.

Rickey and Smythe (1962) successfully accelerated H^- ions, that is, hydrogen atoms with an extra electron, loosely bound, and extracted a beam of protons by stripping the associated electrons in a thin foil. This simple technique allowed a single cyclotron to deliver several extracted beams simultaneously with independently variable energy and inspired the construction of machines designed specifically for H^- acceleration.

The higher magnetic inductions achievable with superconducting magnets found an application in cyclotron design with the development of the superconducting heavy-ion (HI) cyclotron in the 1970s.

Today there are about 150 cyclotrons operating around the world, and this number is increasing at the rate of about five per year. The use of cyclotrons in nuclear research is less widespread than before; however, other fields of research and application have been developed. Many cyclotron laboratories offer customers an ion beam service. The beams are being used, for example, to label items with radioactive elements, to detect contaminants present in trace amounts, and to modify the properties of materials. Cyclotron beams can generate high fluxes of secondary particles, such as neutrons; several hospitals operate cyclotron-based neutron therapy facilities, and other cyclotrons are used in an industrial setting for neutron radiography. The majority of new cyclotrons, however, are intended to produce isotopes for incorporation into radiopharmaceuticals. These, and other, applications are described in Section 13.5.

13.2
Types of Cyclotron

13.2.1
Classic Cyclotron

A charged particle of mass m and charge qe, and moving with a velocity $\mathbf{v}$ in a magnetic field of induction $\mathbf{B}$, is subject to a force $qe\mathbf{v} \times \mathbf{B}$. For the special case of a uniform magnetic field with $\mathbf{v}$ perpendicular to $\mathbf{B}$, this results in a circular particle orbit in the plane containing $\mathbf{v}$ and normal to $\mathbf{B}$. Now a particle with an orbit radius r experiences a centripetal acceleration $\mathbf{v}^2/r$, so by Newton's second law of motion

$$qe\mathbf{v}\mathbf{B} = \frac{m\mathbf{v}^2}{r} \tag{1}$$

From this equation, it follows that the particle orbit radius is directly proportional to the particle momentum mv, and also, as Lawrence recognized, that the particle orbit frequency f_{ion} or angular velocity $\omega = 2\pi f$ given by

$$w_{\text{ion}} = \frac{qe\mathbf{B}}{m} \tag{2}$$

is constant, independent of particle energy, so long as relativistic effects are negligible and m may be considered constant.

These equations are the basis of the classic cyclotron illustrated in Figure 13.1. An essentially uniform magnetic field is generated between the circular poles of an electromagnet, and a hollow cylindrical

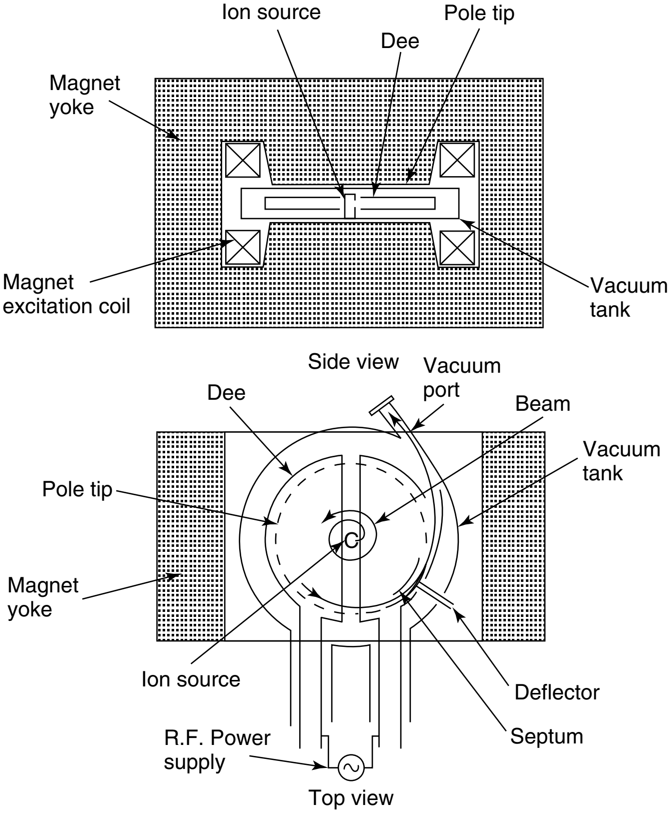
Ion source
Pole tip
Dee
Magnet yoke
Magnet excitation coil
Vacuum tank
Side view
Vacuum port
Dee
Beam
Vacuum tank
Pole tip
C
Magnet yoke
Ion source
Deflector
Septum
R.F. Power supply
Top view

Fig. 13.1 A classic cyclotron, shown in section to reveal the major components.

electrode, split along a diameter, is installed in this gap. In plan views, the two sections resemble the letter *D* and are therefore referred to as *dees*. By maintaining a high-frequency alternating voltage between the dees, an alternating electric field is established in the gap between them. During half the dee oscillation period, the electric field direction is appropriate for extraction of ions from the plasma created by an arc discharge in an ion source located between the dees near the center of the magnet. If the oscillation period of the applied dee voltage is chosen in accordance with Eq. (2), then particles extracted when the electric field is at or near its maximum travel in a circular path, shielded from the electric field while inside the dee, to return to the dee gap half a period later when the field is again at a maximum and in the direction of the particle motion. The beam, therefore, consists of a series of bunches of particles separated by the rf period. At each successive gap crossing the particle, energy is increased with a concomitant increase in orbit radius until the maximum energy and orbit radius, as determined by the magnet pole size, are reached. At this point, in early cyclotrons, the particle beam hit an internal target and initiated nuclear reactions. It is preferable, however, to extract the beam from the cyclotron and lead it to a separate experimental area. This reduces the activation of cyclotron components and permits a much wider range of applications.

The beam is extracted by means of an electrostatic deflector. This consists of a pair of curved electrodes, one a thin grounded metal plate, called a *septum*, which is interposed between the penultimate and final turns of the accelerated beam, and the other a matching electrode located at a slightly larger radius and carrying a high DC voltage. The transverse electric field between the two electrodes provides sufficient radial deflection to enable the beam to be clear of the magnetic field in less than one turn.

Ions that collide with gas molecules while being accelerated may be lost from the beam. To reduce this loss, the volume between the magnet poles, including the dees and other components, is enclosed in a chamber and evacuated.

To produce useful beam intensities in a cyclotron, it is essential that particles having orbits that deviate from the ideal trajectory should also be accelerated and be extractable. An elementary discussion of the beam stability required to achieve this may ignore the effects of acceleration and consider only the motion of particles of selected energies about their corresponding closed or equilibrium orbits. For the classic cyclotron in Figure 13.1, these are circles in the midplane, concentric with the magnet center. The magnetic field in this case is azimuthally uniform but may vary slowly with radius, resulting in curved field lines and nonzero radial field components $B_r(r, z)$ off the midplane, as illustrated in Figure 13.2. To obtain the conditions for focusing in the axial and radial directions, we begin by describing the local midplane field variation by an index $k(r)$ such that

$$B_z(r) = B_z(r_0)\left(\frac{r}{r_0}\right)^{k(r_0)} \tag{3}$$

Here B_z is the midplane magnetic induction, r_0 the equilibrium orbit radius, r a nearby radius, z the axial direction perpendicular to the orbit plane, and

$$k = \frac{\left(\frac{\Delta \mathbf{B}}{\mathbf{B}}\right)}{\left(\frac{\Delta r}{r}\right)} = \left(\frac{r}{B_z}\right)\left(\frac{dB_z}{dr}\right) \tag{4}$$

A particle displaced axially experiences a radial field component $B_r(r, z)$ (Figure 13.2), and an axial force F_z. By retaining only the linear term in the Taylor expansion for B_r about the midplane, $z = 0$, and using the fact that curl $\mathbf{B} = \mathbf{0}$ in the magnet gap, we find that

$$\begin{aligned} B_r(z) &= z\left(\frac{\partial B_r}{\partial z}\right)_{z=0} + \cdots \\ &= z\left(\frac{\partial B_z}{\partial r}\right)_{z=0} + \cdots \end{aligned} \tag{5}$$

and the axial force is

$$F_z = qev_\theta B_r = m\ddot{z} = qev_\theta z\left(\frac{k}{r}\right)B_z + \cdots \tag{6}$$

Particles with small displacements from the magnet midplane experience a linear restoring force, provided k, or more specifically the gradient, $\partial B_z/\partial r$, is negative. Using Eqs (1) and (6), the equation of motion may be written as

$$\ddot{z} - w^2 k z = 0 \tag{7}$$

which, when k is negative, describes simple harmonic motion about the midplane with frequency $\omega_z = \omega_z = \sqrt{-\omega^2 k}$. The oscillation frequency is more commonly expressed in terms of the ion rotation frequency,

$$Q_z = \frac{\omega_z}{\omega_{\text{ion}}} = \sqrt{-k} \tag{8}$$

and Q_z termed the axial betatron frequency, since the theory was first developed

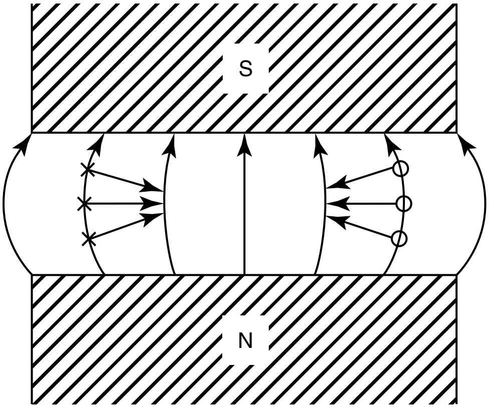

Fig. 13.2 The vertical focusing effect of a radially decreasing magnetic field. The figure shows lines of magnetic induction and the forces experienced by protons traveling in clockwise orbits when seen from above.

for the betatron accelerator. Should k be positive, Eq. (7) describes a steadily increasing axial displacement and particles would eventually collide with the dee or other structures.

A somewhat more complex derivation shows that particles displaced radially from their equilibrium orbits are focused provided that $k > -1$ and that they execute linear betatron oscillations with $Q_r = \omega_r/\omega_{\text{ion}} = \sqrt{1+k}$. (The origin of the restoring force may be comprehended by noting that the centripetal force required for circular orbits decreases as $1/r$ (Eq. (1)). Should the magnetic field decrease more slowly than this, then a particle at a radius larger than its equilibrium orbit will be bent more and focused back toward the equilibrium orbit.) Values of $k < -1$ occur in the edge field of the magnet, the orbit unwinding in a strongly defocusing (radially) field.

For both radial and axial motions to be stable, the midplane field must decrease with radius such that

$$-1 < \left(\frac{r}{B_z}\right)\left(\frac{dB_z}{dr}\right) < 0 \tag{9}$$

A radially decreasing field means that the resonance condition, that is, the equality between the orbit frequency, as in Eq. (2), and a fixed dee frequency, cannot be satisfied everywhere. If $\omega_{\text{rf}} = \omega_{\text{ion}}$ at the cyclotron center, then the revolution time of particles with higher energy will exceed $2\pi/\omega_{\text{rf}}$ and the particles arrive at the dee gap at progressively later times with respect to the peak of the accelerating voltage. If we write the energy gained each turn in the two-dee cyclotron of Figure 13.1 as

$$E_G = 2qe2\widehat{V}\cos\phi \tag{10}$$

where V is the peak voltage on one dee and ϕ is the voltage phase when the particle crosses the dee gap center, then progressively later arrival times correspond to a progressive increase in ϕ until, eventually, the accumulated phase slip exceeds $\pi/2$, at which point acceleration ceases and particles begin to decelerate back toward the machine center where they are lost. The lack of isochronism and its consequent phase slip is exacerbated by the relativistic increase in mass, $m = \gamma m_0 = (1 + T/E_0)m_0$, which causes the rotation time to increase as the kinetic energy T increases. To postpone the inevitable deceleration to the highest possible energy, classic cyclotrons operated with the highest sustainable dee voltage. However, the practical limit corresponded to 100 or 200 turns and $\gamma_{\text{final}} \approx 1.02$.

The U-150 machines built in the Soviet Union are examples of late-period classic cyclotrons. Protons were accelerated to 12 MeV, deuterons to 20 MeV, and α particles (^{4}He) to 40 MeV in a 330-ton magnet with a pole face diameter of 1.5 m. The dee voltage was 85 kV at 10 MHz and the electrostatic deflector held 70 kV. Beam currents up to 250 μA (1.5×10^{15} proton s^{-1}) were obtained.

13.2.2
Synchrocyclotron or Frequency-Modulated Cyclotron

Two methods have been devised to deal with the problem of phase slip and thus permit acceleration to high energies. Frequency modulation, described subsequently, was implemented first; an alternative method, due to Thomas, will be discussed in Section 13.2.3.

In the frequency-modulated (FM) cyclotron, or synchrocyclotron, the frequency of the accelerating voltage is lowered progressively so that it matches the rotation frequency of a hypothetical reference ion, termed the *synchronous particle*, as it is accelerated. A cluster of bunches captured at about the same time can thus be accelerated to energies far beyond the synchronism limit of the classic cyclotron. Following extraction of the cluster, the rf oscillator is returned to its starting frequency, another batch of ions captured, and the process repeated.

If it were essential that the ion orbit frequency and the accelerating frequency match exactly, then the phase acceptance and capture time interval would be small and the beam current low; it would also be difficult to program the rf. Fortunately Veksler (1944) and McMillan (1945) showed independently that exact matching is not necessary because, within certain limits, ions whose energy or phase deviate somewhat from the synchronous ion will oscillate, in energy and phase, about that particle and gain energy at the same average rate. This behavior, termed the *principle of phase stability*, makes synchrocyclotrons and also synchrotrons and linear accelerators practical machines.

The principle was first demonstrated experimentally by Richardson *et al.* (1946). The first synchrocyclotron to surpass the classic cyclotron limit operated at Berkeley, also in 1946, and accelerated deuterons and α particles to 95 MeV amu^{-1}. About 20 synchrocyclotrons have been built. The repetition rates for frequency modulation range from 50 to 4000 Hz, while ions are captured into phase-stable orbits typically for 0.01–0.1 ms; consequently the duty factor and beam intensity are about 1% of that from a cw cyclotron. The long acceleration time means that ions make many turns, and dee voltages are low, typically 10–50 kV.

The highest proton energies, 1000 MeV, are achieved in the synchrocyclotron at the St. Petersburg Nuclear Physics Institute, Russia. The machine at the European Centre for Nuclear Research (CERN), Geneva, Switzerland, has accelerated 7 μA of protons to 600 MeV and ~0.2 particle μA of carbon, oxygen, and neon to ~80 MeV amu^{-1}. The 3000-ton magnet has a 5-m pole-tip diameter and a field falling from 1.94 to 1.81 T. A single dee is used with a voltage 20 kV. A frequency swing between 30.4 and 16.6 MHz for protons and, for example, 7.6–6.6 MHz for ^{20}Ne^{5+} is provided by a rotating capacitor at 360 Hz.

Synchrocyclotrons pioneered research in meson physics, studies of nucleon–nucleon forces, and particle therapy for cancer and other diseases. These roles are being taken over by AVF machines, although there are proposals to build new synchrocyclotrons for therapy since currents of a few nanoamperes are adequate, and the simplicity of construction and operation may give some competitive advantage.

13.2.3
The Sector-Focused, Isochronous, or AVF Cyclotron

To maintain an ion rotation time that is independent of energy (isochronism), the midplane magnetic induction B_z of the classic cyclotron should increase with radius as

$$\begin{aligned} B_{\text{isoch}}(r) &= \gamma(r) B_0 \\ &= B_0 \left(1-\beta^2\right)^{\frac{-1}{2}} \\ &= B_0 \left[1-\left(\frac{r\omega}{c}\right)^2\right]^{\frac{-1}{2}} \end{aligned} \tag{11}$$

where γ and β are the usual relativistic quantities. Such a field is axially defocusing. To overcome the defocusing force and provide adequate axial focusing, Thomas (1938) proposed that an azimuthal variation in B_z be introduced by installing three or more wedge-shaped iron sectors (often called hills) on opposing pole faces of the magnet, as in Figure 13.3. The particle path has a smaller radius of curvature in the hill and a larger one in the intervening valley. Therefore, the closed, or equilibrium, orbit now has a radial modulation, termed *scalloping*, about its average value. Particles have a radial velocity component v_r at the sector edges. Those particles displaced from the midplane experience an azimuthal component of induction $\mathbf{B}\theta$, resulting in an axial force $qev_r\mathbf{B}\theta(r,\theta,z)$ that is focusing at both edges of a sector.

Additional focusing can be realized by having the sectors follow a spiral rather than a radial locus. The curved hill edges then produce a radial component of $\mathbf{B}$ off the midplane that results in an axial force $qev\theta B_r(r,\theta,z)$. This force focuses at the convex hill edges and defocuses at the concave edges. Following the well-known principle of classical optics for a series of focusing and defocusing lenses of equal strength (the strong-focusing principle), the net effect is focusing. In fact, the focusing is even stronger because, as detailed calculations for cyclotron beam optics show, the focusing edge is stronger than the defocusing one.

A Fourier series can be used to describe the midplane induction and the orbit scalloping:

$$\begin{aligned} B_z(r,\theta) = \overline{B}_z(r) \Bigg[1 + \sum_p \big(H_p(r)\cos p\theta \\ + G_p(r)\sin p\theta \big) \Bigg] \end{aligned} \tag{12}$$

$$\begin{aligned} r = \overline{r} \Bigg[1 + \frac{1}{p^2} \sum_p \big(H_p \cos p\theta \\ + G_p \sin p\theta \big) \Bigg] \end{aligned} \tag{13}$$

The magnitude of the azimuthal variation of midplane induction is termed the *flutter* $F(r)$, where

$$\begin{aligned} F &= \frac{1}{2} \sum_p \left(H_p^2 + G_p^2 \right) \\ &\approx \left(B_{\text{hill}} - \overline{B} \right) \frac{\left(\overline{B} - B_{\text{valley}} \right)}{\overline{B}^2} \end{aligned} \tag{14}$$

The betatron frequencies cannot be given in a closed form for the AVF cyclotron; however, several authors, for example, Hagedoorn and Verster (1962), have obtained

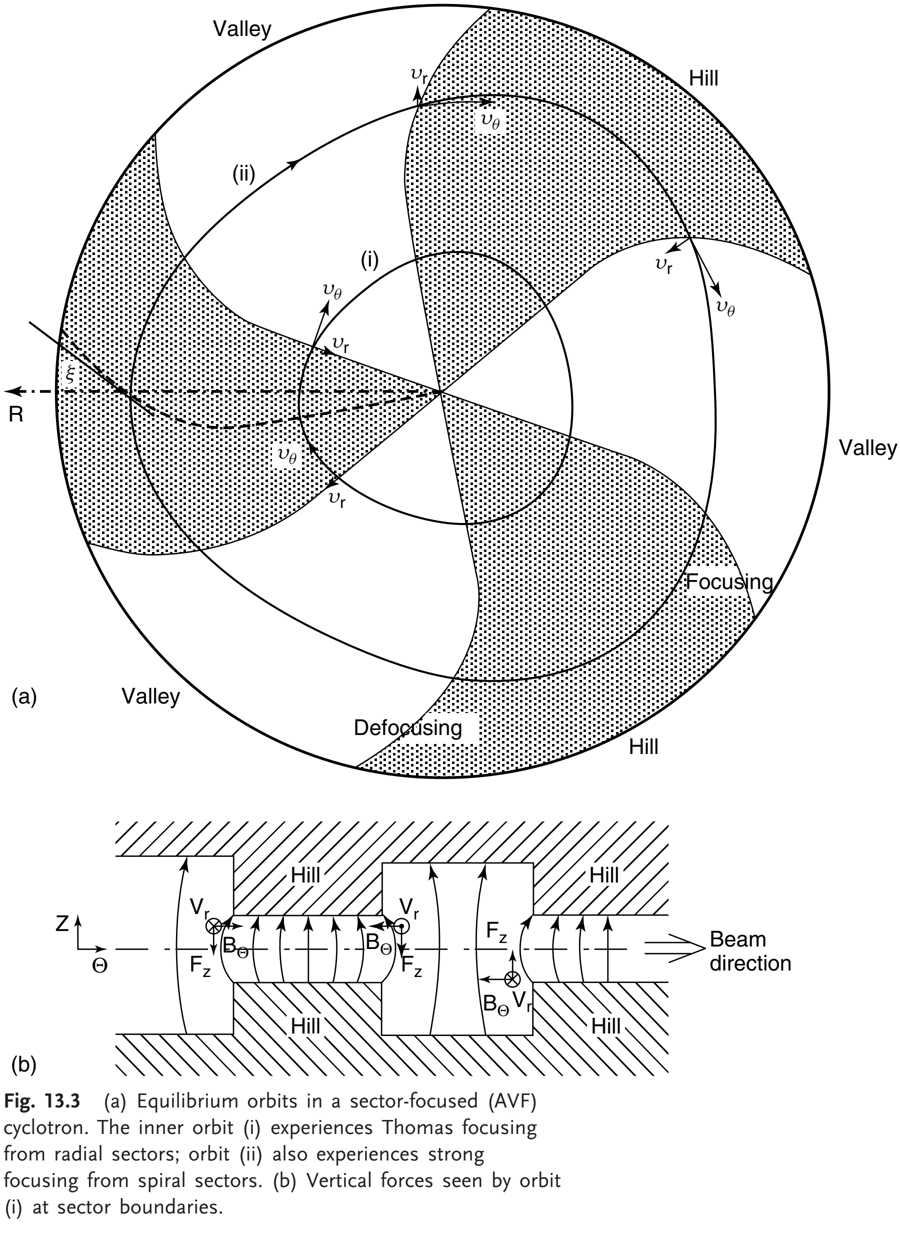
Valley
Hill
υ_r
υ_θ
(ii)
(i)
υ_θ
υ_r
υ_r
υ_θ
ξ
R
υ_θ
υ_r
Valley
Focusing
(a)
Valley
Defocusing
Hill
Hill
Hill
V_r
B_Θ
F_z
V_r
B_Θ
F_z
F_z
B_Θ
V_r
Z
Θ
Beam direction
Hill
Hill
(b)

Fig. 13.3 (a) Equilibrium orbits in a sector-focused (AVF) cyclotron. The inner orbit (i) experiences Thomas focusing from radial sectors; orbit (ii) also experiences strong focusing from spiral sectors. (b) Vertical forces seen by orbit (i) at sector boundaries.

expressions for Q_z and Q_r in the form of infinite series, the leading terms of which are useful approximations, such as,

$$Q_z^2 = -\overline{k} + \left[\frac{N^2}{(N^2-1)}\right] F\left(1 + 2\tan^2\xi\right) + \cdots \tag{15a}$$

$$Q_r^2 = 1 + \overline{k} + \cdots \tag{15b}$$

where $k(r)$ is now the average field index, N is the number of sectors, and $\xi(r)$ is the spiral angle (Figure 13.3(a)).

For perfectly isochronous machines k may be replaced by $\gamma^2 - 1$. From Eq. (15a) it can then be seen that the axial focusing strength will tend to decrease with increasing particle energy. To counteract this, especially in high-energy cyclotrons, designers increase the spiral angle with radius, up to a practical limit of ~70°.

Isochronous cyclotrons can both accelerate a wide variety of ion species and vary the energy of the beam extracted by altering the dee frequency and making the appropriate adjustments to field strength and shape. The relation between the kinetic energy T, the field, and radius may be obtained by writing Eq. (1) in its relativistic form, $\overline{Br} = \beta\gamma m_0 c/qe$, and recalling that $m_0c^2\gamma = E_0 + T$, where E_0 is the rest energy of the ion. The result

$$\overline{Br} = \frac{\left(T^2 + 2TE_0\right)^{\frac{1}{2}}}{qec} \tag{16}$$

is sometimes more conveniently expressed in terms of energy per nucleon T/A, where A is the ion mass in atomic mass units. This relation is plotted in Figure 13.4 for several charge/mass ratios. The diagram can be made more useful for defining the operating range of a cyclotron by adding a scale corresponding to fr, where f is the ion orbit frequency. The radio frequency range required can be reduced by operating, for some T/A, at a harmonic of the rotation frequency, that is, $f_{\mathrm{rf}} = hf_{\mathrm{ion}}$. Two isochronous cyclotrons of the type discussed earlier, which have accelerated many different ion species are the 88-in. AVF cyclotron at the Lawrence Berkeley Laboratory and the Variable Energy Cyclotron at the Atomic Energy Research Establishment, Harwell, UK; both were built in the early 1960s.

The maximum energy attainable by an ion in a given magnet may be set either by its ability to bend the trajectory into a closed orbit (Eq. (16)) or by its ability to provide axial focusing (Eq. (15a)). The bending power is sometimes characterized by K_b, where, in the nonrelativistic approximation, $T/A \approx K_b(q/A)^2$. Saturated iron gives fields in the region of 2 T; however, superconducting excitation coils can contribute an extra 4 T and $K_b \approx 1000$ T m can be achieved in compact, single-pole machines with extraction radius 1 m. The iron sectors in these superconducting cyclotrons are saturated at all working fields; the field modulation is constant and therefore the flutter is inversely proportional to $\overline{B}^2$. This behavior led Blosser and Johnson (1974) to introduce a second magnet parameter K_f to characterize the focusing limit and, for superconducting cyclotrons, the maximum energy per nucleon is the lesser of $K_b(q/A)^2$ and $K_f(q/A)$. The situation is different for cyclotrons using unsaturated iron; for these the flutter is almost independent of excitation. Superconducting machines to accelerate heavy ions began operating at the National Superconducting Cyclotron Laboratory (NSCL-MSU), East Lansing, Michigan, and at Chalk River National Laboratories, Canada, in the 1980s.

Before this, during the 1960s, the concept of the separated-sector cyclotron (SSC) had been developed, first at the

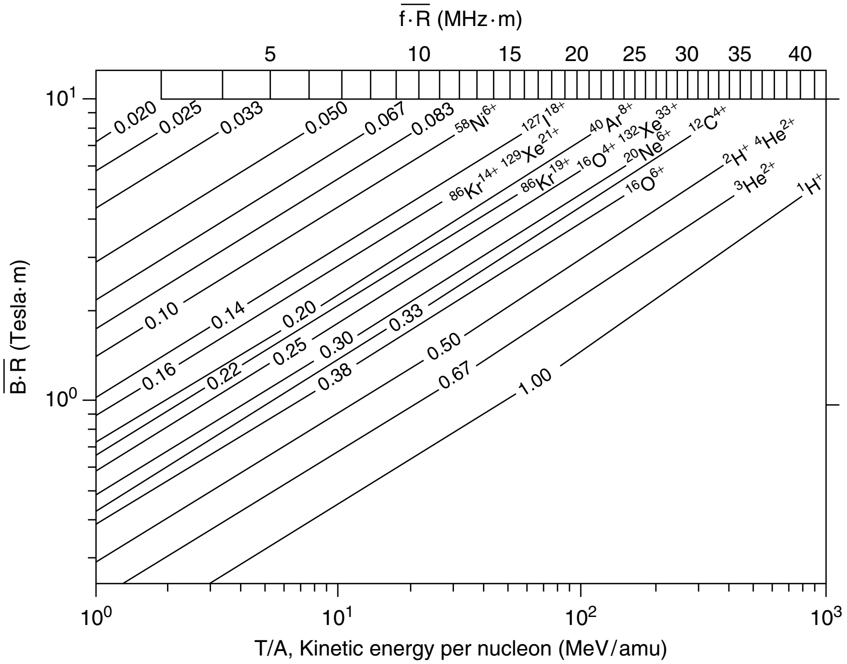

Fig. 13.4 A chart useful for initial design showing the relationship between kinetic energy, magnetic field, extraction radius, and orbital frequency for ions with different q/A. A laboratory with an operating machine may redraw such a figure using axes of field and dee frequency/harmonic.

Paul Scherrer Institute (PSI), Villigen, Switzerland (Willax, 1963), and then at the Indiana University Cyclotron Facility (IUCF), US. In these machines individual sector magnets are arranged in a pinwheel manner (Figure 13.5) and powered in series. The absence of iron in the valleys provides easier access and space for other equipment, especially for efficient, high-Q, rf cavities operating at high voltage that, in turn, lead to more efficient beam extraction. The valley field is small and the flutter F approaches unity. The lower average field, however, means a larger, more costly, cyclotron, and the absence of a conventional center region means that ions must be pre-accelerated to an energy corresponding to the first turn.

Sector-focused cyclotrons are examples of fixed-field alternating-gradient (FFAG) accelerators. This term, however, is usually reserved for machines with magnets of SSC style and with a field shaped such that the betatron frequencies are independent of radius (energy). This would allow ions to be accelerated to high energies, GeV amu^{-1}, while avoiding resonances in betatron motion, but since the field is no longer isochronous, the accelerating frequency must be modulated as in a synchrocyclotron. Several proton FFAG machines of the latter type have been proposed but only models, accelerating electrons, have been built.

More than 100 AVF machines have been built since the first isochronous

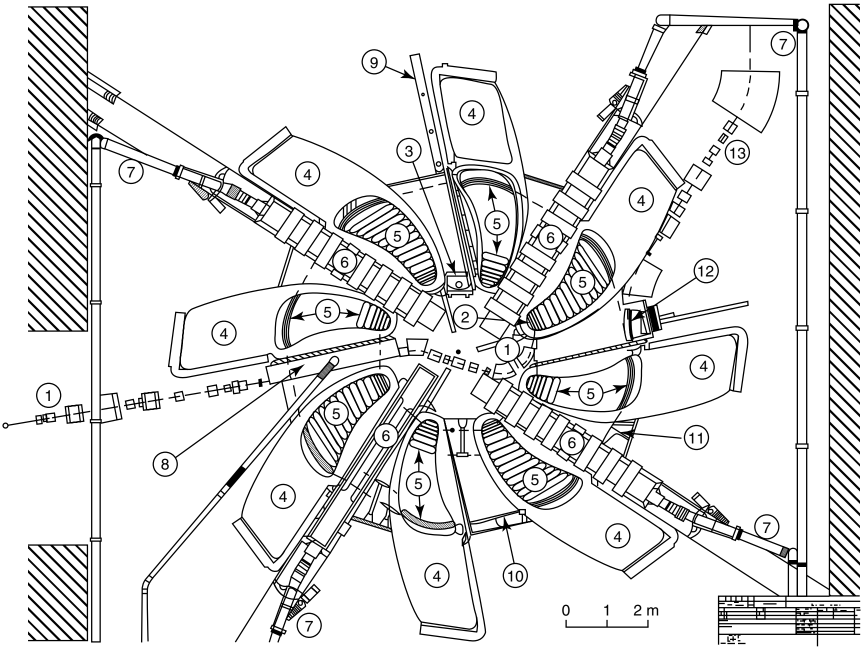

Fig. 13.5 The PSI separated-sector cyclotron. This machine accelerates protons from 72 to 590 MeV. (1) Beam line from 72 MeV injector cyclotron, (2) magnetic inflector channel, (3) electrostatic inflector, (4) one of eight sector magnets, (5) set of correcting (trim and harmonic) coils wound on a magnet pole, (6) one of four 50-MHz accelerating cavities, (7) coaxial rf power transmission line, (8) 150 MHz "flat-top" cavity, (9) beam diagnostic probe (radial drive), (10) electrostatic deflector, (11) magnet to focus the deflected beam, (12) septum magnet, and (13) extracted beam line (590 MeV). (Courtesy of the Paul Scherrer Institute, Villigen, Switzerland).

cyclotron, equipped with radial sectors and accelerating protons to 12 MeV, operated at Delft, The Netherlands, in 1958. They range from small machines accelerating protons to a few mega electronvolts to machines several meters in diameter, some of which accelerate intense beams of protons to hundreds of mega electronvolts, while others are capable of accelerating virtually any element in the periodic table to energies of $\sim$100 MeV per nucleon. The distribution of K values for machines currently operating is given in Figure 13.6. Virtually all new cyclotrons are of the AVF type, and some older machines have been converted; for example, the synchrocyclotrons at the Joint Institute of Research, Dubna, Russia, and the University of Uppsala, Sweden, now have sectors added to their magnet poles to improve performance. Many new cyclotrons accelerate negatively charged hydrogen ions, H^-, in order to exploit the simple and versatile stripping process for extraction. They are intended for nontraditional fields of research and commercial applications. One such machine is illustrated in Figure 13.7, and the parameters of another are given in the appendix.

Perhaps the most novel cyclotron under construction is a separated-orbit cyclotron, called *TRITRON*, at the Technical

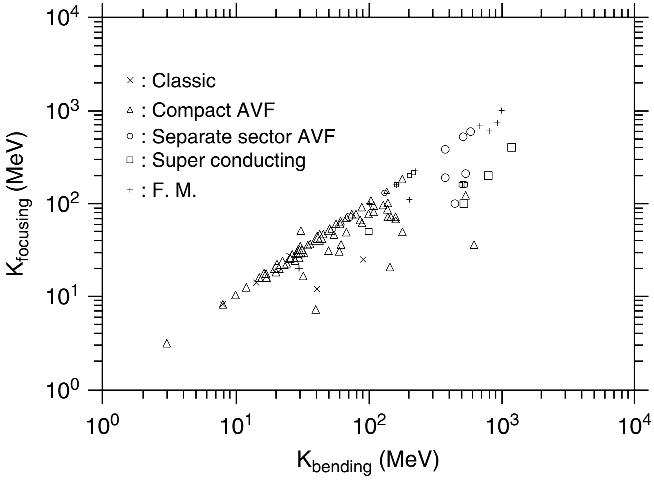

Fig. 13.6 Parameters describing the maximum energy per nucleon attainable in 143 cyclotrons of different types. The symbol size is proportional to the number of machines with the same parameters; however, cyclotrons manufactured commercially are underrepresented. $(T/A)_{max}$ in MeV amu^{-1} is the lesser of $K_b(q/A)^2$ and $K_f(q/A)$. (From the compendium attached to the proceedings of the 11th International Cyclotron Conference).

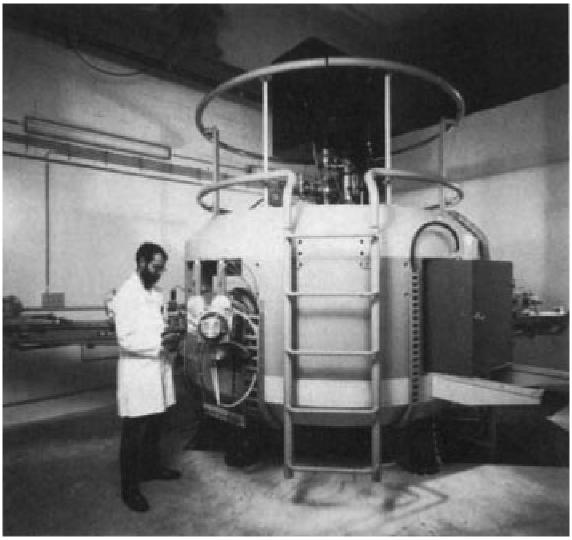

Fig. 13.7 A modern compact cyclotron accelerating H^- ions to 30 MeV. (Courtesy of I.B.A., Belgium).

University, Munich, Germany. It employs 12 superconducting sector magnets and 6 superconducting rf cavities, each capable of 0.5 MV. Ions are injected from a potential-drop accelerator and their energy increased fivefold in only 20 turns. Orbits are separated by ∼4 cm; conductors can be placed around each turn in each sector and the magnetic field and gradient, and hence steering and focusing, adjusted at each step.

13.2.4
The Microtron (Electron Cyclotron)

Microtrons are constant-field, constant-rf, cw electron accelerators. The low electron mass ($m_0c_2 = 0.511$ MeV) means that the particles are relativistic at very low energies and that the orbit frequency is (28 GHz T^{-1})$\mathbf{B}/\gamma$. Electrons would rapidly lose synchronism in a conventional cyclotron; however, Veksler showed in 1944 that a suitable energy gain E_G could be found for which the phase slip for each turn corresponded to an integral number of rf cycles and, that the principle of phase stability applied. The orbits are not concentric but arranged to have a common tangent at which point an accelerating cavity is placed. This cavity may even be a small linear accelerator. The relative energy gain per turn is high, orbits are well separated, and extraction is straightforward. The highest-energy microtron is the 850-MeV stage at Mainz, Germany, which is intended for nuclear research. Most microtrons operate at tens of mega electronvolts and are used as injectors for other electron accelerators or as compact sources for radiation therapy. Currents are of the order of 100 μA. Although occasionally termed electron cyclotrons, microtrons are usually considered to be a separate field and the reader is referred to the general references at the end of the chapter.

13.3
Design Considerations and Subsystems for an AVF Cyclotron

The design of a cyclotron is an iterative process during which many parameters, which interact with one another, are brought to an overall optimum. Some design considerations are introduced in the following text, but the most complete sources of information are the published proceedings of conferences on accelerators and related topics referred to for further reading.

13.3.1
Beam Physics

13.3.1.1 Resonances

Ions accelerated in an AVF machine undergo many transverse oscillations that can exhibit the same phenomena, for example, amplitude growth at resonance or coupled motion, seen in other periodic systems. For example, the inclusion of higher-order terms in the field expansion (Eq. (5)) transforms Eqs (6) and (7) from a description of free to one of forced oscillations. For a field harmonic p (Eq. (12)), the general resonance condition is

$$mQ_r + nQ_z = p, \quad m, n, p \text{ integers} \tag{17}$$

Energy may, in principle, be transferred between transverse and longitudinal motion; however, while the phase in an AVF cyclotron may deviate from an average value it does not perform periodic oscillations. This is not the case for synchrocyclotrons, and a term involving the frequency of phase oscillations would be included in Eq. (17).

Resonances in betatron motion influence the choice of Q_z, the number of magnet sectors and accelerating gaps, and the tolerances of manufacture and assembly.

In an isochronous cyclotron, $Q_r \approx \gamma$ and so may start near 1 and increase with energy; also Q_z frequently rises from a low value near the center of the machine where the field modulation F is small and changes again on approaching the focusing limit or the fringe field where k deviates from the isochronous requirement. In superconducting cyclotrons, different magnet excitations have a marked effect on F and Q_z. Consequently, one or more resonances may be crossed. The Mathieu–Hill differential equation describes particle motion with periodic focusing; analysis shows that the phase shift of a betatron oscillation is $2\pi Q_{r,z}$ per N per sector and that motion becomes unstable when this phase shift is π. Consequently, stable radial motion is not possible in a two-sector cyclotron; at least three sectors are required at low energy. For each N there is a corresponding limiting value for Q_r. Should $N = 3$, for example, then $Q_r \leqslant \frac{3}{2}$, which, in the case of protons in an isochronous cyclotron, corresponds to 469 MeV. In practice, the limit is somewhat lower.

Low-order – m (or n) $= \pm 1$ or ± 2 – resonances may be crossed provided the pth amplitude or gradient is small; small coils are often placed in the resonance region and their currents adjusted empirically to cancel any magnetic field errors remaining after assembly. Higher-order resonances, $m = \pm 3, \pm 4, \pm 4, \cdots$, correspond to nonlinear terms and may be traversed, provided the betatron amplitudes are small (Hagedoorn and Verster, 1962).

Cyclotrons intended for nuclear physics research may be required to preserve polarization of the beam during acceleration. The angular momenta or spins of atomic nuclei are usually randomly distributed. Polarized ion sources select and populate one spin orientation. These ions are injected with spins aligned with the magnetic field and precess about this axis at the Larmor frequency. Those particles whose motion causes them to experience, in the rest frame, a quasi-constant transverse field component will begin to precess about the resultant, nonvertical axis. Conditions for depolarization exist whenever

$$\gamma G + mQ_r + nQ_z = p \qquad (18)$$

where G is 1.79 and -0.143 for protons and deuterons, respectively, and -3.79 and -1.86 for H^- ions and D^- ions. Driving terms arising from field imperfections can, again, be compensated by coils. Depolarization has not been observed in positive-ion cyclotrons, but the large negative G values for H^- and D^- make resonances with the sector symmetry more likely, and H^- polarization changes of 7% have been observed at the six-sector TRIUMF cyclotron, Vancouver, Canada, and complete depolarization at the University of Manitoba, Canada, four-sector machine.

13.3.1.2 Electric Forces

The forces exerted on an ion by a magnetic field are proportional to both charge and velocity, while those from an electric field depend on the charge alone. The two are equivalent when $E = (300\,\mathrm{MV\,m^{-1}\,T^{-1}})\beta\mathbf{B}$. Electric fields, therefore, are relatively more important at low energies.

The electric field at an accelerating gap is illustrated in Figure 13.8. An axially focusing kick is experienced by particles entering the gap, a defocusing kick on

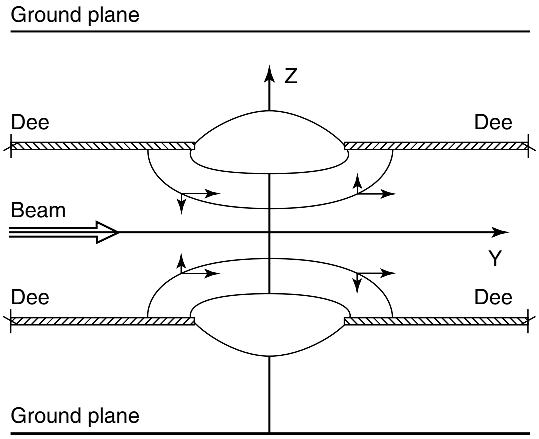

Fig. 13.8 Electric field lines and their components in the acceleration gap.

leaving it. At low energies the transit time may be a substantial fraction of the rf period, leading to a net focusing effect for particles in the gap when the voltage is falling (positive phases). Flutter vanishes at the magnet center, and electric focusing can so dominate the first few turns that the defocusing of ions with negative phases can lead to their removal from the accelerated beam. Another defocusing force arises from the mutual repulsion of the ions. This space-charge force, which is proportional to the charge density, is the major factor limiting the beam current at low energies. At any energy the distribution of accelerating impulses around a turn may have a *p*th harmonic satisfying Eq. (17); gap-crossing resonances may also be driven by the beat frequency arising from the combination of dee and sector symmetries.

The forces from the accelerating structure depend on the ion phase; so do the space-charge forces, usually.

13.3.1.3 **Charge-Changing Phenomena**

An ion that picks up an electron, or has one stripped from it, will alter its charge/mass ratio, drop out of synchronism, slip in phase, and eventually be decelerated and lost. Either may follow from a collision with a residual gas molecule however, and except in highly ionized heavy ions, stripping dominates once acceleration has begun. The cross section per atom $\sigma(T)$ for stripping H^- to H^0 in air is usefully approximated by $7 \times 10^{-19}/\beta^2\,\mathrm{cm}^2$ and the fraction lost by $2.36 \times 10^4\ P(E_0/E_G)R_e$ where P (bar) is the operating pressure and R_e (m) is the extraction radius for a cyclotron with $\gamma \approx 1$. The cross section for stripping in hydrogen is eight times smaller. A vacuum of 10^{-10} bar seems adequate for compact HI or H^- cyclotrons, and 7×10^{-11} bar for larger machines.

At high energies, the Lorentz electric field $\mathbf{E} = \gamma \mathbf{v} \times \mathbf{B}$ lowers the potential barrier of weakly bound negative ions sufficiently to allow quantum-mechanical tunneling to occur and an electron to escape. Measurements and theory are in good agreement, the lifetime for an H^- ion under such conditions being

$$\tau = \left(\frac{a}{\mathbf{E}}\right) \exp\left(\frac{b}{\mathbf{E}}\right) \tag{19}$$

where $a = 2.47 \times 10^{-6}\,\mathrm{V\,s\,m^{-1}}$ and $b = 4.49 \times 10^9\,\mathrm{V\,m^{-1}}$. Figure 13.9 gives, for H^- ions of several energies, the fraction

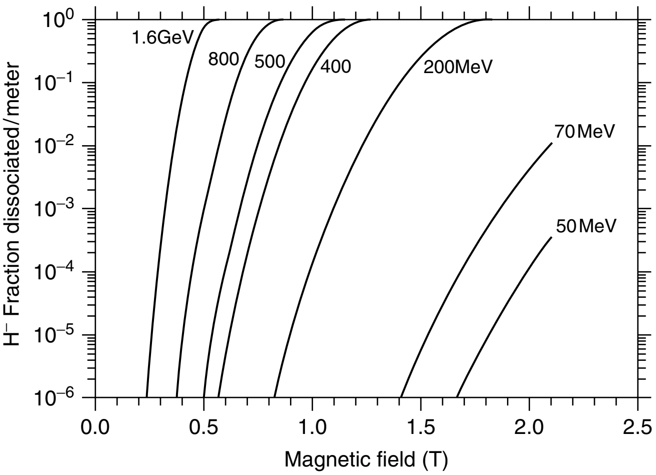

Fig. 13.9 Dissociation of H^- ions as a function of field strength and energy (path length = 1 m).

dissociated per meter path length as a function of the magnetic field strength. The steepness of Eq. (19) means that this electromagnetic stripping is localized to the hills and the highest energies of a given cyclotron.

13.3.1.4 **Computer Codes**

Analytic methods are complemented by the extensive use of computer codes. These are especially important in calculating trajectories of ions in electric and magnetic fields and thence the betatron frequencies and deviation from isochronism (Gordon, 1984). The fields themselves, when they cannot be measured, may be calculated using two- or three-dimensional finite-element codes.

13.3.2
Magnet

The maximum magnetic rigidity ($\overline{Br}$) required can be obtained by use of Eq. (16) or Figure 13.4. The peak hill field is usually at or near magnetic saturation. The field increase required for isochronism and the required flutter F then determine the maximum central field value B_0. This is generally used in order to minimize magnet size and cost. Exceptions are the higher-energy H^- cyclotrons where ion dissociation (Figure 13.9) restricts the hill field. The gap field of a magnet falls rapidly, approaching a pole edge; consequently the maximum isochronous orbit radius is smaller than the pole radius by at least half the hill gap width.

The yoke serves as a low-reluctance flux-return path. Early cyclotrons employed the H-frame design illustrated in Figure 13.1. Beam ports, dee stems, and other services were inserted through the open sides, and modifications were easily made. Superconducting machines and modern compact cyclotrons have a more efficient flux-return structure that surrounds the pole and has midplane and vertical penetrations for extracted beam and other services. This provides, also, some shielding from induced radioactivity (Figure 13.7 and Appendix). Access to the interior of the vacuum vessel is obtained by raising the upper yoke, pole, and vessel lid.

Both the poles and yoke are fabricated either from low-carbon mild steel in the form of forgings or rolled plate, or from soft iron castings. The magnetomotive force is provided by coils fitted around the poles, generally wound from hollow, water-cooled, copper conductor. Power consumption ranges from 30 kW for a minicyclotron to almost 1 MW for separated-sector machines with large K_b.

Two- or three-dimensional finite-element codes can be used to determine flux distributions in the pole and yoke, the ampere-turns required for excitation, and to some degree, the average flux distribution in the gap. With their aid, magnets can be designed that generate fields within 1% of a specified design objective. This is generally not precise enough for cyclotron operation, and an iterative process of field mapping followed by shimming, that is, adding or removing small amounts of iron at various locations in or near the gap, must be included as part of magnet fabrication and commissioning. This procedure reduces deviations from the ideal field to a few milliteslas, which may be adequate for low-energy cyclotrons. For others with tighter tolerances, auxiliary coils are included for empirical adjustments of the isochronous field, or correction of field harmonics, based on measurements of beam properties. These are termed *trim coils* and are wound on the pole face or vacuum vessel (e.g., Figure 13.5). Similar coils of larger power are used to produce the changes in $\mathbf{B}(r)$ required for isochronous operation with different γ_{final}.

The superconducting cyclotron employs superconducting windings in a magnet with iron yoke and poles at room temperature. The high current density, typically 3000 A cm^{-2}, achievable in superconductors makes magnets with large (6 T) gap inductions possible. This results in a considerable economy when designing a cyclotron to accelerate, for example, partially ionized heavy ions with low q/A and high magnetic rigidity. All coils built to date use copper-stabilized multifilament NbTi superconductor, and, because of the large stored energies involved, all are cryostatically stable. A high copper-to-superconductor ratio, typically 20 : 1, is used to make a winding with many internal channels, allowing good thermal contact between the liquid helium bath coolant and the conductor. The iron of the hills and pole is completely saturated and can be treated as uniformly magnetized volumes for design purposes. The coils contribute directly to the gap induction and are generally divided into a pair of coils close to the midplane and a more distant pair; each pair is energized independently to provide some control over the radial variation of $\overline{B}(r)$. Design currents range from 700 to 2300 A, and the stored power from 18 to 60 MJ.

At the Chalk River superconducting cyclotron laboratory, an array of 13 cylindrical steel rods with axes perpendicular to the midplane penetrate the hills. Axial adjustment of a rod produces a local change in the hill fields to provide both radial variation of the field and harmonic adjustment.

13.3.3
Accelerating Systems

Isochronous cyclotrons do not necessarily require a large dee voltage to reach high energies, and in some cases the second dee of Figure 13.1 is replaced by a narrow grounded electrode termed a *dummy dee*. A large energy gain per turn E_G, however, relaxes magnetic field tolerances, opens the center region geometry, assists in conventional extraction, and reduces stripping losses. The dees of an AVF cyclotron are often triangular or spiral shaped, designed to fit within the valleys of the magnet poles. The increased number of gaps increases E_G. The smaller hill gap improves magnet efficiency.

The dees are made part of an rf resonant circuit, usually as the capacitive termination of a shorted quarter-wave line resonator. This generates higher dee voltages and also facilitates impedance matching to the power source. The resonator can also provide the mechanical support for the dees, obviating the need for failure-prone insulator supports. The shorting plate is usually movable for coarse frequency tuning adjustment and to accommodate different operating modes. The high thermal and electrical conductivity of copper makes it the preferred material. The center conductor of the resonator is usually water cooled; other surfaces are also often water cooled to minimize dimensional changes due to thermal expansion and to maintain a stable resonant frequency.

The accelerating fields for many SSCs are developed in more efficient resonant cavities installed in several of the spaces between magnet sectors (Figure 13.5).

Radiofrequency power is coupled to the dees or cavities by either terminating the drive line from the rf source with a plate that is located in a high rf electric field region near a dee or by terminating the drive line with a loop near a magnetic field maximum, for example, the short. Plate or loop dimensions and orientation are adjusted to match source and load impedances.

Cyclotrons tend to operate between 5 and 75 MHz and rf power requirements range from about 15 kW–1 MW. To generate these, a variety of gridded tubes manufactured for use in shortwave broadcasting or telemetry are available. They are used in either of two basic circuits. The earliest and simplest circuit is the self-excited oscillator, illustrated schematically in a grounded-grid configuration in Figure 13.10. It uses positive feedback to

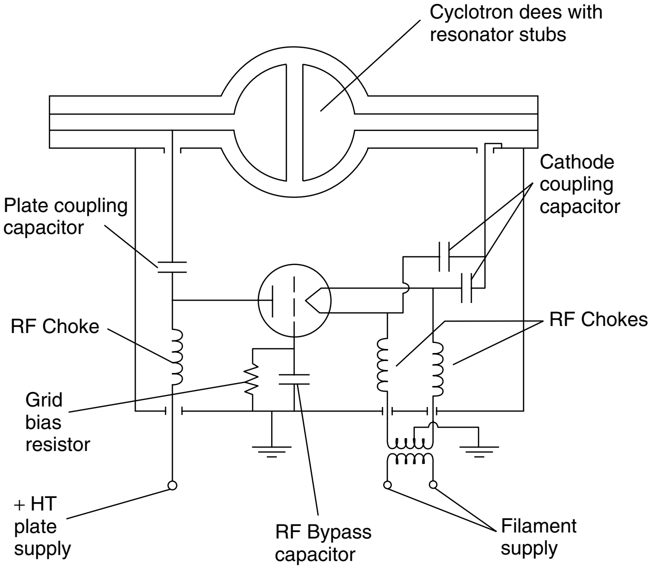

Fig. 13.10 A simplified grounded-grid self-excited oscillator circuit for cyclotron dee excitation.

the power-tube cathode, and the dee and line resonator then determine the resonant frequency. The second type of circuit is the master oscillator–power amplifier. In this case, the feedback line is replaced by a signal obtained from a drive chain consisting of a crystal oscillator and several stages of amplification. In either case, an automatic motor-driven capacitor is usually installed near the dee to maintain the desired resonant frequency by fine tuning.

For a synchrocyclotron, rf power must be supplied as the frequency sweeps over its operating range. A self-excited oscillator system is often used with the transmission-line short replaced with a variable capacitor, either a motor-driven rotating capacitor or a vibrating vane, and the high-power rf is swept in frequency in accordance with the varying resonator capacitance.

13.3.4
Ion Sources for Cyclotrons

Cyclotron ion sources must operate continuously, unlike those used with pulsed machines. Until the 1960s, most ion sources were variants of a type developed at the Oak Ridge National Laboratory, Tennessee, based on the Penning ionization gauge (PIG). In these sources (Figure 13.11), an arc discharge is maintained inside a hollow chimney (anode), aligned parallel to the magnetic field, and passing through the midplane near the cyclotron center. A cathode is placed at one end of the chimney and biased by a few hundred volts with respect to the anode. Electrons drawn from the cathode are constrained by the magnetic field to pass through the arc chamber toward an anticathode placed at the other end of the chimney that reflects them back toward the midplane. The cathode may be a filament or a block heated indirectly by electron bombardment. Each electron can oscillate through the arc chamber many times before collision, thus increasing the probability of ionization. A current of 0.1–10 A at 100–1000 V maintains the discharge. The flow of gas (or vapor) to be ionized is carefully regulated through a small line to the arc chamber. The electric field of the dee penetrates a small slot in the chimney to extract ions from the plasma. The system is otherwise closed and a relatively small gas flow maintains an adequate pressure. These sources provide both light and heavy positive ions and negative hydrogen and deuterium ions. Higher electron energies produce multiply charged ions; however, the cathode lifetime decreases because of ion bombardment. The chimney diameter must be small so as not to interrupt any of the early turns; the relatively small source volume limits the maximum input arc power and consequently the maximum charge state.

An ion source can produce several ion species, with different q/A, at the same time. The requirement that the rf be equal to, or a harmonic of, the orbit frequency means that only certain species will be accelerated. The selection of species can be refined further by placing posts between early turns.

High-charge states, more current, or polarized beams can be provided by larger sources placed outside the cyclotron. These inject their ions either through a valley (Figure 13.5) or down the magnet axis. Axial injection was first demonstrated at Birmingham University, UK, in 1964. A hole is bored along the axis of the magnet yoke and ions, with tens of keV amu^{-1}, led through an evacuated tube to a point just before the magnet midplane (Figure 13.12). They are bent into the plane of acceleration by means of either an

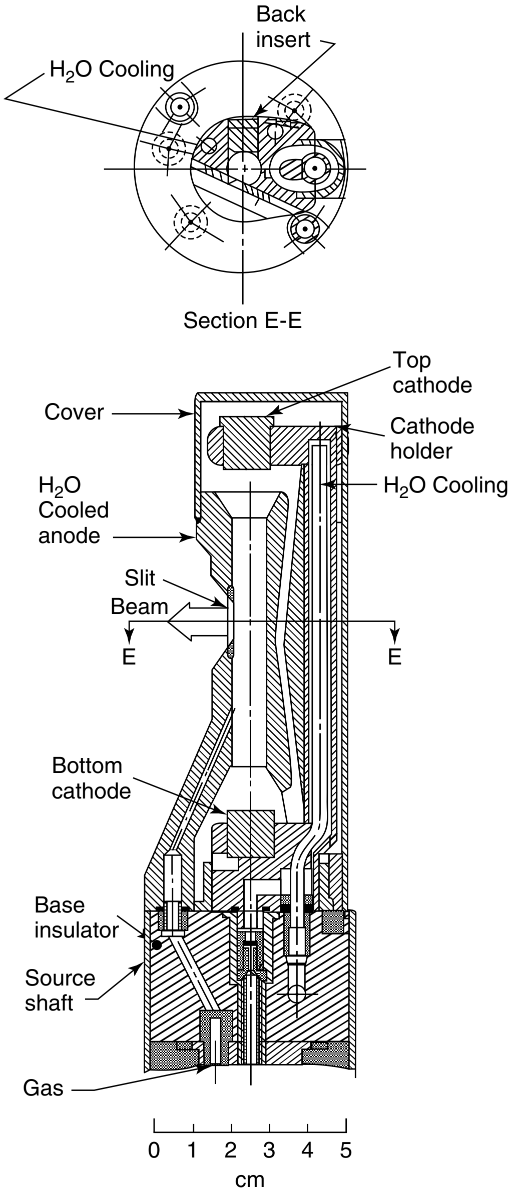
Back insert
H_2O Cooling
Section E-E
Top cathode
Cover
Cathode holder
H_2O Cooled anode
H_2O Cooling
Slit
Beam
E
E
Bottom cathode
Base insulator
Source shaft
Gas
0
1
2
3
4
5
cm

Fig. 13.11 Schematic cross section of the internal PIG source used until 1985 for high-charge-state ions in the 88-in. cyclotron. (Courtesy D. J. Clark, Lawrence Berkeley Laboratory).

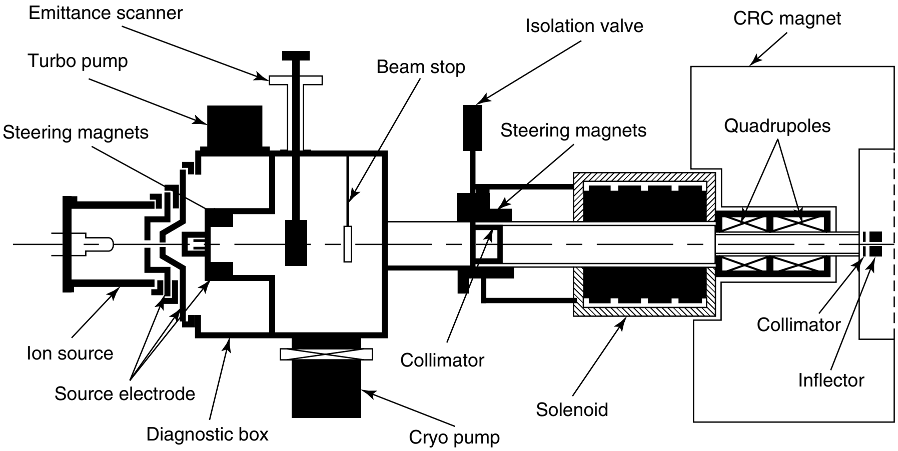

Fig. 13.12 Axial injection from a cusp source providing H^- ions into a 1-MeV test-bed cyclotron (CRC) at TRIUMF. The plane of acceleration is vertical in this cyclotron.

electrostatic mirror (a grounded grid placed above a biased electrode, both at an angle of about 45°) or, preferably, by an inflector consisting of a pair of curved electrodes carrying voltages of opposite sign. Some electrostatic or magnetic focusing elements may be required in the tube. Detailed field calculations are necessary to ensure a smooth match between cyclotron and injection system optics.

Electron cyclotron resonance ion sources (ECRIS) were developed at the Centre d'Etudes Nucléaires, Grenoble, France, in the 1970s and are used by most new HI cyclotrons. Microwaves at a fixed frequency, somewhere between 2 and 20 GHz, are fed into a cavity located within a solenoid. The solenoid is designed to provide at least two surfaces where field and rf satisfy the resonance condition (Eq. (2)) for electrons and where power can be transferred from microwaves to electron motion. In the absence of collisions, the electrons can reach energies of several mega electronvolts. When gas or vapor is introduced, these high-energy electrons collide with atoms or molecules yielding ions and more electrons for the microwaves to accelerate. Multipole (sextupolar or octupolar) magnetic fields confine the plasma radially, which increases the ion lifetime and hence the probability of multiple collisions and consequently the maximum charge state. Several charge states are extracted and the desired species is selected by a bending magnet and slits before injection. The source contains neither filament nor cathode, nor is an arc required to obtain a plasma. Elements that do not easily vaporize may be ionized by placing solid pieces in the resonance zone.

Plasma can be confined within a cylindrical chamber by rows of permanent dipole magnets; a row with north poles facing the plasma alternates with a row with south poles. The shape of the magnetic field flux lines led to the term cusp or multicusp ion source. The chamber is the anode and a heated filament is the cathode. The source chamber may be divided into two regions by a small transverse magnetic field allowing independent optimization of two-step processes. An arc of 25 A at 100 V provides very intense continuous beams of negative hydrogen and deuterium, particularly suited for isotope-producing cyclotrons. The source shown in Figure 13.12 delivers 7 mA of H^-. The Paul Scherrer Institute has developed a proton version capable of 25 mA.

Nuclearly polarized ions of ^1H, ^2H, ^3H, and ^3He are primarily produced by the three different types of source discussed subsequently, namely, atomic (ground state), Lamb shift (metastable), and optically pumped.

An atomic source uses an rf discharge to dissociate $^1H_2$ or $^2H_2$ molecules. The atomic beam is passed through a tapered magnetic sextupole that deflects out of the beam one-half of the hyperfine states that have a particular electron spin direction (Stern–Gerlach effect). The remainder then pass through an rf field that induces transitions from states containing unwanted nuclear spin alignments to the just depopulated hyperfine states. A pass through a second tapered sextupole removes those atoms that have undergone rf transitions. The result is a beam of nuclearly polarized atoms that may now be ionized to $^1H^+$ or $^2H^+$ in helium or to negative ions formed by charge exchange with alkali atoms. Proton currents are about 500 μA, H^- currents are of the order of 10 μA, and polarization is usually about 90%.

Lamb-shift-type polarized-ion sources begin by directing a 550-eV proton beam (or equivalent velocity ^2H, ^3H, or ^3He beam) through a cesium-vapor charge-exchange target. About one-third of the atoms emerging from the cesium are in the $2S_{1/2}$ metastable state. This atomic beam then passes through a magnetic solenoid containing an rf electric field. The magnetic field and rf frequency are chosen such that metastable states with an unwanted nuclear spin become degenerate with the $2P_{1/2}$ state and quickly decay to the ground state. Finally, the atomic beam is passed through an ionizer (iodine for positive ions, cesium for negative) that selectively ionizes the metastable atoms rather than the ground-state atoms. Polarizations of about 80–85% are obtained at currents of 1–2 μA.

An optically pumped polarized-ion source is used at TRIUMF for experiments requiring more intense proton beams. Circularly polarized light from dye lasers tuned to the rubidium D_1 line polarizes the outer electrons of the atoms in a rubidium-vapor charge-exchange target immersed in an axial magnetic field. Protons at ~3 keV have a large probability of picking up a polarized electron in this target to produce an electronically polarized beam of hydrogen atoms. The atomic beam leaves the first solenoid to enter a second solenoid whose magnetic field is in the opposite direction and for which the field gradients on axis have been designed to induce a Sona type transition (i.e., only those atoms with the hyperfine states having opposed electron and nuclear spins change their nuclear spin direction). This results in a nuclearly polarized beam of hydrogen atoms that may then be ionized to positive or negative ions. Polarization of 78% has been achieved to date with 10 μA of H^-; positive-ion currents should be 10 times larger.

13.3.5
Extraction

The two principal methods for extracting the beam from a cyclotron are electrostatic and magnetic deflection and extraction by stripping. In the latter process, the beam passes through a charge-changing medium, usually a very thin foil, located so that particles with altered q/A leave the machine between magnet yokes. A simple change in the radial position of the foil changes the energy extracted; several narrow foils, or ones partially inserted into the beam, allow the simultaneous extraction of several beams with different energies. This method is now chiefly used for negative ions of hydrogen but, in the past, has been used for molecular ions $H_2{}^+$, HD^+, and so on.

The stripped electrons dissipate their energy in the foil; passage of ions both heats the foil and causes radiation damage. Refractory materials, such as carbon or aluminum oxide, are commonly used. Amorphous graphite foils, 200 μg cm^{-2}, extract up to 200 μA from a Cyclotron Corporation CP-42 cyclotron, operated by Nordion International, Canada, at energies between 20 and 40 MeV. Fifteen foils, each lasting about 10 mA-h, are mounted in a cartridge.

Deflecting systems are used for more highly ionized ions or when the charge state must be preserved. The electrostatic deflector of Figure 13.1 may be followed by one or more magnetic channels, which produce a local field reduction to augment the deflection. These channels may be coils powered to oppose the main field or iron bars placed on either side of the deflected beam to divert magnetic flux. The fringe field from such channels often perturbs the orbits before extraction, and additional coils are incorporated to cancel such effects.

Any beam incident on the septum of an electrostatic deflector causes local heating and makes both the septum and neighboring components radioactive. Secondary particles affect voltage holding. Unfortunately the turn-to-turn separation arising from acceleration falls rapidly with energy, as $(B_{0\beta\gamma}{}^{3})^{-1}$, and turns normally overlap at extraction radius. The turn width depends on both radial emittance and the energy spread associated with the phase band being accelerated. Some machines eliminate extraction losses by eliminating overlap between turns at the deflector entrance. This may be achieved by a combination of

1. slits near the machine center to restrict emittance and phase width at the cost of intensity,
2. a large energy gain per turn, and/or
3. the excitation of a coherent radial betatron oscillation.

The latter alternately adds to and subtracts from the turn spacing resulting from acceleration.

The relative alignment of beam and deflectors is quite critical and is usually adjusted empirically. Once set up, this system demands a better machine stability than that required just for acceleration to the final energy or that required for stripping extraction.

The magnet fringe field is a powerful lens, and optics calculations should be performed to ensure that the beam matches smoothly into the external beam transport system.

13.3.6
Vacuum

The tightest constraints for vacuum in cyclotrons arise from either the inability of electrodes to support high voltage or unwanted charge-changing collisions between ions and residual gas molecules. Acceptable operating pressure ranges from 10^{-8} bar for a small proton machine to 10^{-11} bar for H^- or high-charge-state heavy ions. The initial outgassing rate for clean stainless steel or nickel-plated mild steel is about 10^{-8} bar L s^{-1} m^{-2}, and consists chiefly of water vapor. Over several days the evolution of water declines to match the steady-state evolution of hydrogen of $10^{-10}\,\mathrm{bar\,L\,s^{-1}\,m^{-2}}$. Provided that little organic material is incorporated, there are no leaks, and there is no internal ion source, these gases usually constitute the continuing gas load.

The speeds of pumps commonly used lie between 10^3 and $10^5\,\mathrm{L\,s^{-1}}$; to achieve this pumps should be close to, or preferably inside, the cyclotron and be unaffected by magnetic field. Cryopanels cooled to 20 K with 80 K baffles are preferred to the older oil diffusion pumps. Additional hydrogen pumping may be provided by charcoal or zeolite at 20 K or by condensate cryopumps. Initial pump-down is provided by mechanical pumps located in regions free from magnetic field. The magnetic field may also affect the calibration of ion gauges or the operation of mass spectrometers used to analyze the residual gas. Pump-down time can be shortened by venting with dry nitrogen and warming the vacuum tank while roughing. New equipment should be cleaned with solvent or detergent and baked before installation to remove volatile organics and adsorbed water from the surface.

13.3.7
Beam Instrumentation

The beam parameters measured, in a rough order of frequency of measurement, are as follows:

1. both the amount of charge delivered and that lost;
2. the transverse position of the bunch center, its energy, and its phase with respect to rf;
3. the beam size and distribution as a function of these coordinates;
4. polarization; and
5. emittance and correlations between parameters.

In general, these quantities may be measured directly in a beam line, whereas in a cyclotron they must sometimes be inferred by observing the dependence of a related parameter on a machine variable. Most cyclotrons have at least one movable probe giving beam intensity, phase, and density as functions of radius. However, a cyclotron whose operating tolerances are comfortably within the precision of manufacture may monitor only current and the position and size of the beam spot near the production target. A research facility will have the most extensive array of instrumentation.

The sensing instrument may intercept all or part of the ion beam or detect the electromagnetic field associated with a moving bunch of ions. Scintillation screens glow when hit by a beam with density $\sim 1\,\mathrm{nA\,cm^{-2}}$. This glow, when observed through a window by a TV camera, gives an inexpensive display in a control room of beam position, size, and transverse correlations. The system is, however, nonlinear and subject to radiation damage. The beam can be stopped in an electrically insulated conducting target and the incident charge taken via a vacuum feed-through to a measuring instrument that may be a meter, amplifier, oscilloscope, and so on. Information on transverse distributions may be obtained from a subdivided target or from a wire target swept through the beam. At higher energies, the thickness to stop the beam may be inconveniently large, and secondary particles, for example, electrons or X rays, escaping from the surface of a thin target may be detected. Electrons in a cyclotron, including those stripped from H^- ions, spiral along magnetic field lines until collected. In most environments, an incident current >1 nA may be measured. Ion chambers or electron multipliers that boost incident charge have been used for lower currents. Plates or loops placed near the beam detect the electromagnetic field to give information on intensity, position, and phase without perturbing the ions. Various signal-processing techniques enhance the signal-to-noise ratio, but, in general, they are less sensitive than intercepting methods. At IUCF, a monitor that capacitively senses the beam provides a 0.3-mm resolution at 50-nA current.

The cyclotron operating conditions for an ion species sparsely produced by an ion source may be found by using a more intense species with very similar q/A (analog beam).

It is important to be able to bring signals generated by beam users for their own purposes into the central control room. They are concerned that beam properties match their requirements and may be able to generate information beyond the resources of the cyclotron crew.

13.3.8
Computer Control and Monitoring

This is a rapidly changing field. Simple cyclotrons are operated using commercial process controllers. Control systems in research laboratories, however, have to service many more set points and need more frequent revisions to incorporate new equipment or modes of operation; online data-reduction algorithms are also common. These systems, therefore, should be compatible with off-line software development. The performance of some devices should be logged regularly and automatically for fault diagnosis and failure prediction. Closed-loop regulation of subsystems frequently includes microcomputers that exchange data with the main control system; however, closed-loop operation of the entire cyclotron is still rare.

13.4
Cyclotrons as Injectors or Post-accelerators

Several laboratories generate relativistic beams using two or more accelerator stages. This is because a single-stage cyclotron that provides a large energy increase and has a peak field chosen to satisfy the requirements of isochronism and turn spacing at extraction and to avoid saturation will also have a low central field B_0. Relatively few turns take place in the center, and magnets and accelerating structures can be used more economically by a two-stage process that separates the optimization of injection and initial acceleration from the extraction of a high-energy beam. The beam transport line between two machines provides an opportunity to trim off unwanted particles or to increase the charge state of partially ionized atoms by stripping.

The momentum of the beam from an energetic pre-accelerator is too high for axial injection into a cyclotron, and injection schemes tend to mirror the extraction schemes of Section 13.3.5. Fully stripped ions may be directed roughly radially along the valley of an SSC (Figures 13.5 and 13.13) and an electrostatic inflector used to steer them onto the accelerated equilibrium orbit. The radius gain per turn should be such that the next turn clears this inflector. Partially stripped ions may be directed across the face of a solid-pole magnet onto a foil placed at the common tangent of their path and the orbit of stripped particles of the same energy.

The rf of a second cyclotron should be the same as, or a low multiple of, the

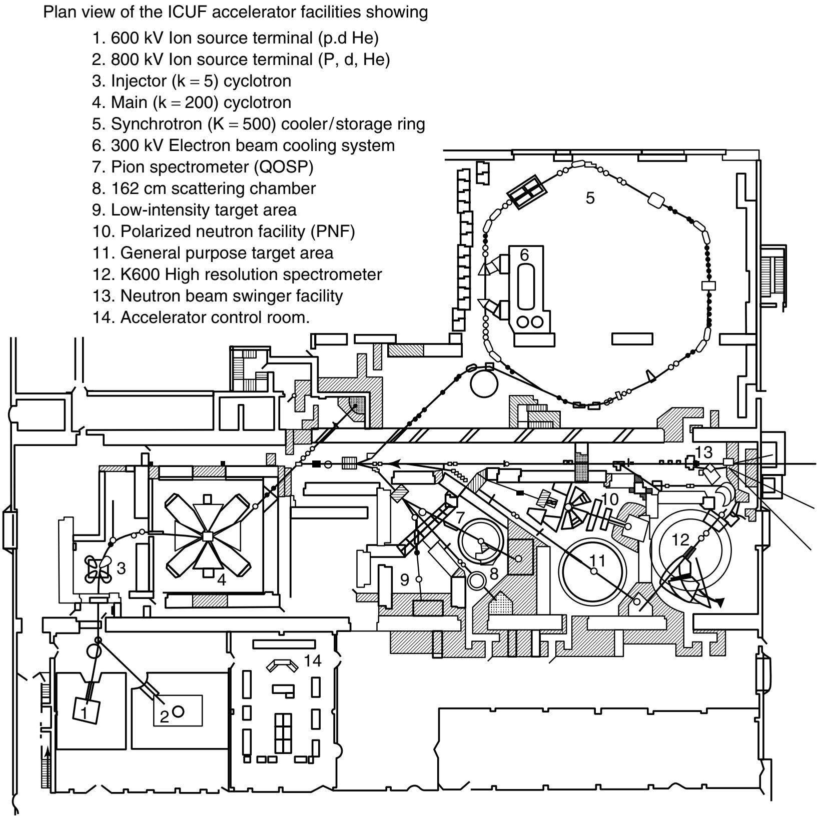
Plan view of the ICUF accelerator facilities showing
1. 600 kV Ion source terminal (p.d He)
2. 800 kV Ion source terminal (P, d, He)
3. Injector (k = 5) cyclotron
4. Main (k = 200) cyclotron
5. Synchrotron (K = 500) cooler/storage ring
6. 300 kV Electron beam cooling system
7. Pion spectrometer (QOSP)
8. 162 cm scattering chamber
9. Low-intensity target area
10. Polarized neutron facility (PNF)
11. General purpose target area
12. K600 High resolution spectrometer
13. Neutron beam swinger facility
14. Accelerator control room.
5
6
13
10
7
12
11
3
4
9
8
14
1
2

Fig. 13.13 Plan view of accelerator facilities at the Indiana University Cyclotron Facility. The figure shows four stages of acceleration from the ion-source terminals, which are potential-drop accelerators, through two cyclotron stages to a cooler/synchrotron. (Courtesy of D. Friesel, Indiana University).

frequency of the injector cyclotron and the phase width of the injected beam much less than the half-period of the new machine. An rf "bunching" cavity is often installed between machines. This buncher improves beam capture by decelerating the beam during one half-cycle and accelerating it during the other to produce a "focus in time" at the first accelerating gap of the next machine.

While cyclotrons are natural partners, with 11 pairs in operation worldwide, other machines have also been used as pre- or post-accelerators. Four laboratories use potential-drop machines as injectors; they have the advantage, for negative ions, of an external ion source and the opportunity to strip off electrons in the terminal, thus increasing q/A. Two laboratories use potential-drop post-accelerators to provide simple and precise energy control over a wide range. A 70-MeV amu^{-1} variable-frequency linear accelerator provides heavy ions for the $K_b = 540$ SSC operated by the Institute of Physics and Chemistry (RIKEN), Saitama, Japan.

Cyclotron beams are also injected into storage rings where they may be accumulated over many turns and ''cooled,'' that is, have betatron and synchrotron oscillation amplitudes reduced either by applying a series of small transverse or longitudinal kicks to the beam or by transferring oscillation energy to an electron beam that shares part of the ring circumference. The experimenter's target is also placed within the ring. The recirculation of the beam through the target increases the interaction rate while the cooling system removes the ''heating,'' or beam degradation, caused by the target so as to maintain the beam size and narrow energy spread. Most rings include an rf cavity so that, after accumulation, the magnet and rf may be ramped, as in a synchrotron, and the beam accelerated to higher energy. Five cooler/synchrotrons have been built, or are under construction, the first (Figure 13.13) being at Indiana University with $K_b = 560$. The first cooling of a beam from a cyclotron occurred in 1984 at the 8-MeV amu^{-1} test accumulation ring TARN I at the Institute of Nuclear Study, Tokyo.

It is proposed, at TRIUMF, to produce 100 μA of 30-GeV protons using the existing cw cyclotron as an injector for a chain of rapid-cycling synchrotrons. Since synchrotrons, like synchrocyclotrons, capture particles over a relatively short injection period, an interjacent storage ring would be necessary in order to utilize most of the cyclotron beams. H^- ions would be extracted from the cyclotron at ~450 MeV and injected into the ring by stripping to protons. The protons would be accumulated, at 450 MeV, over ~20 ms (~10^4 turns) and extracted in one turn for transfer to a synchrotron, whereupon accumulation would start again. The synchrotron would have a 50-Hz repetition frequency.

In general a post-accelerator demands better beam quality and stability of operation from an injector than most other applications.

13.5
Applications of Cyclotrons

The development of early cyclotrons was driven by the needs of basic research in physics and chemistry; however, workers in other fields were soon experimenting with the use of isotopes as tracers and particle beams in medicine. Construction of cyclotron-based facilities intended primarily for use in fields other than nuclear physics began in the 1950s. A number of them were intended specifically for isotope production (e.g., Amersham International (UK)), some for general science and engineering (e.g., Harwell), and others for neutron therapy (e.g., Hammersmith Hospital, London (UK)). The characteristics of cyclotron beams overlap those of potential-drop and linear accelerators, and the machines share some common applications.

In the past, cyclotrons have been designed and sold by large engineering companies, for example, Philips (The Netherlands), A.E.G. (Germany), Thomson–CSF (France), and Japan Steel Works. They have been joined, in the past 15 years, by smaller specialized firms such as Scanditronix (Sweden), IBA (Belgium), RDS–Siemens (US), and Ebco Technologies (Canada).

13.5.1
Interaction of Particle Beams with Matter

Most cyclotron applications depend on the way their beams interact with matter–these processes are summarized subsequently.

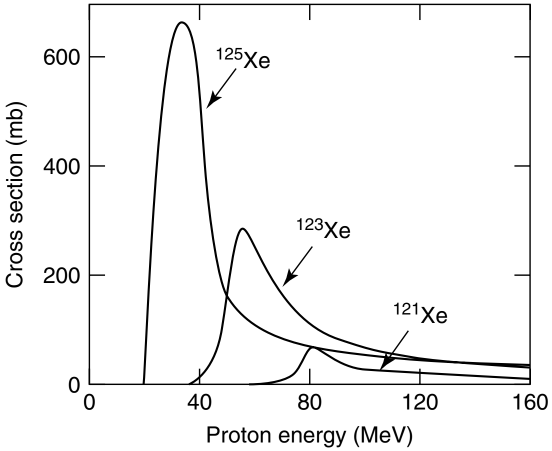

Fig. 13.14 Production of xenon isotopes as a function of bombarding energy; reaction ^{127}I(p, xn)Xe. (After Syme *et al.*, 1978).

Charged particles with energies up to a few mega electronvolts interact chiefly with the Coulomb field of electrons and nuclei. Sufficient momentum may be imparted to atoms to lead to lattice dislocations. Electrons may be detached from atoms (ionization) or excited to more energetic states with subsequent emission of characteristic X rays. Molecules may be dissociated (radiolysis), and the resulting ions or radicals may be very reactive chemically. Detached electrons or displaced atoms may be sufficiently energetic to cause further ionization or dislocation.

At energies above the Coulomb potential barrier (4 MeV for protons on oxygen), charged particles may enter the nucleus and increase nucleon orbital energy with subsequent emission of characteristic γ rays, or initiate a reaction transforming the atom to a different isotope or element. The deuteron binding energy is only 2.2 MeV; at energies above this the nuclear Coulomb field may polarize, and then dissociate, an incoming deuteron. The component neutron may enter a nucleus, since neutrons are not inhibited by the Coulomb barrier, and initiate reactions.

Particles with tens of mega electronvolts may eject several nucleons or light nuclear fragments from a nucleus. Figure 13.14 shows the probability of multiparticle ejection and the creation of different isotopes as the bombarding energy increases. At higher energies, copious numbers of energetic secondary particles are emitted and cascade reactions ensue. Meson production begins around their rest energy, 106 MeV for muons and 140 MeV for pions.

Atomic interactions are many orders of magnitude more likely than nuclear interactions and ions lose their energy in many small steps. The energy deposited along the track, averaged over many samples, is shown in Figure 13.15. Ions, in contrast with neutrons, photons, and electrons, have a specific range, at the end of which there is a peak, the "Bragg peak," in differential energy loss. There is a statistical variation in range caused by absorption reactions and angular scatter. Very low energy π^- are captured in a nucleus that disintegrates into short-range fragments and neutrons enhancing the Bragg peak. Figure 13.16 shows the energy

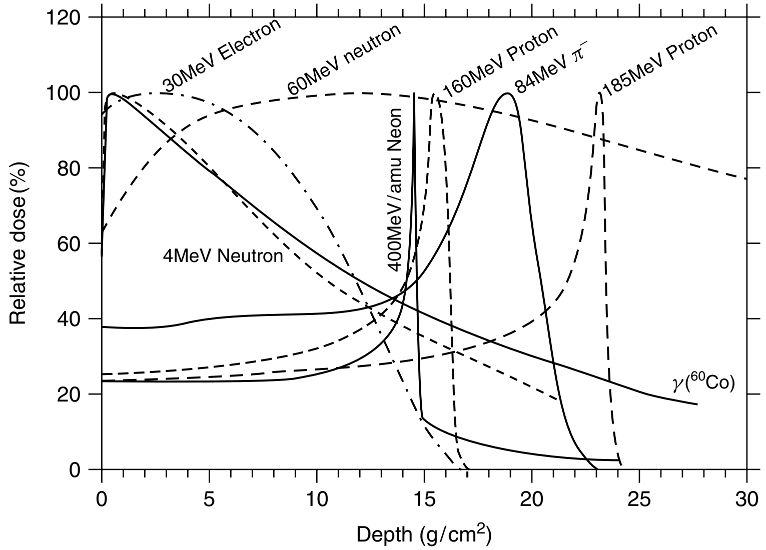

Fig. 13.15 Dose (related to local ionization) as a function of penetration depth in water or tissue for several particle beams and γ rays (1.17 and 1.33 MeV) from ^{60}Co.

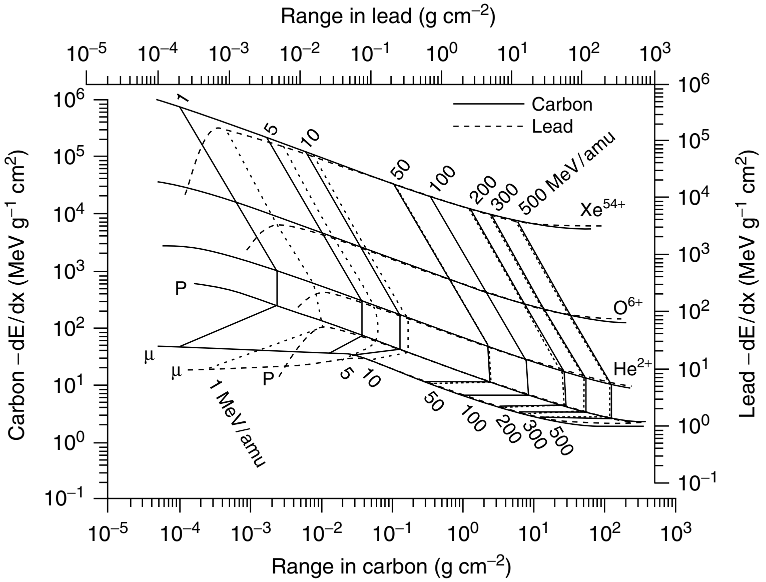

Fig. 13.16 Range and stopping power of muons, protons, α particles, oxygen, and xenon ions in carbon and lead (vertical lines join particles with the same energy per amu and are not intended for interpolation).

dependence of the differential energy loss, or stopping power, and the mean range for several particles in carbon and lead.

13.5.2
Basic Research in Nuclear Science*

The first generation of AVF cyclotrons, with $K \sim 100$ and accelerating light ions ($A < 7$), contributed beams that were more intense and higher in energy than were available from potential-drop machines. The energy resolution could be made comparable by the use of slits to reduce phase-space volume and by magnetic analysis of the extracted beam coupled with magnetic analysis of reaction products. The beam bunch length could be made small enough that reaction products could be identified from their differences in flight time from target to detector. On the other hand, the quasi DC macro duty factor, when compared with that of a pulsed linear accelerator, gave better rejection of uncorrelated events for those experiments that depend on the coincident detection of several correlated reaction products. Some examples of fields investigated using these cyclotrons are quantitative descriptions of nuclear potentials, identification of nuclear states excited at tens of mega electronvolts, and reactions with several emergent nucleons.

Nuclear structure and reaction mechanism studies continue at the higher-energy SSCs and meson factories where, for example, the nucleon–nucleon potential could be explored using spin-polarized beams and the charge distribution within nuclei obtained by comparing π^+ and π^- scattering. The new, cyclotron-injected, coolersynchrotrons offer nuclear physicists superior energy resolution ($\Delta E/E \sim 10^{-5}$) at high energies, from tens of mega electronvolts per amu to 2 GeV. The high flux from meson factories facilitated particle physics experiments, in particular the search for those rare decay modes of mesons that may violate the presently accepted model of strong, weak, and electromagnetic forces or hint at conservation laws not yet included. At bombarding energies of hundreds of mega electronvolts, many different isotopes are produced, many of which are short lived. An electromagnetic analysis system that can isolate individual species online is used by hundreds of experimenters at the CERN synchrocyclotron.

Beams from HI cyclotrons may have less energy per nucleon than from meson factories, but their higher total energy makes them suitable for study of the bulk properties of condensed nuclear matter in a background relatively free from the products of competing, for example, nucleon–nucleon, reactions. Heavy projectiles can generate excited nuclear states with high rotational energy or extreme modes of vibration. A large energy input creates compound nuclei with high thermal energy; subsequent evaporation of nuclear matter produces transuranic elements and nuclei far from the line of stable elements on the A–Z diagram.

Astrophysicists and nuclear chemists are interested in reactions initiated by unstable isotopes and have proposed facilities to produce, collect, ionize, and accelerate radioactive particles. The first such facility in operation is at the Université de Louvain, Louvain-la-Neuve, Belgium, where ^{13}N, with a 10-minute half-life, is generated using 30-MeV protons from one cyclotron, ionized in a radiation-hard ECRIS placed next to the production target, and then accelerated in the adjacent $K = 120$ variable-energy cyclotron CYCLONE.

---

*Chapters 6, 9, and 10 give a general treatment of this subject.

13.5.3
Basic Research in Other Physical Science

Neutrons have long been used in condensed matter research. Accelerator-based neutron sources now complement the traditional reactors and are more easily licensed at present. Storage-ring/synchrotron sources produce pulses of neutrons with a higher instantaneous flux and offer good resolution through time-of-flight techniques. The integrated thermal-neutron flux from the beam dump of a cw-cyclotron meson factory approaches that from research reactors ($2 \times 10^{14}\,\mathrm{cm}^{-2\cdot}\,\mathrm{s}^{-1}$); the cold-neutron flux is superior. An ambitious project under construction at PSI will direct a 1.4-mA beam from their 600-MeV proton cyclotron vertically into a lead target surrounded by horizontal neutron guide tubes. Liquid D_2 or H_2 moderators are placed close to this almost point source. The neutron flux at the experiments, per mA of protons, is $10^{10}n$ $(\mathrm{cm}^2\ \mathrm{s\ nm})^{-1}$ at 0.1 nm to $10^{9}n$ $(\mathrm{cm}^2\,\mathrm{s\,nm})^{-1}$ near 1-nm wavelength.

The muon, μ^- or μ^+, has a mass 206 times that of an electron, or one-ninth that of a proton. It experiences electromagnetic and weak forces but not the strong (nuclear) force. The μ^- mimics the response of a heavy electron to such forces, but the μ^+ acts more like a light proton, often forming the ‘‘element’’ muonium (μ^+e^-) that has almost the same size and binding energy as the chemically reactive hydrogen atom but only one-ninth the mass and three times the thermal velocity. The proton radius is 10^{-13} cm, whereas the μ^+ is pointlike. The muon half-life, 2.2 μs, vastly exceeds the thermalization time (of order 1 ns) when stopping in matter and is long with respect to many chemical and physical processes. Their range may be estimated from Figure 13.16.

The muon beams originate from pion decay and are spin polarized along, or opposite to, their direction of motion. In a local magnetic field, $\mathbf{B}$, a muon will precess at a Larmor frequency of $(135.5\,\mathrm{MHz\ T}^{-1})\mathbf{B}$. The precession of the triplet state of muonium at low fields is $(13.9\,\mathrm{GHz\,T}^{-1})\mathbf{B}$. Muons themselves decay into a neutrino, an antineutrino, and an electron or positron; the latter are emitted preferentially along the muon spin axis with energies up to 52.8 MeV and are easily detected. In a typical muon spin-rotation experiment, a sample is placed in a homogeneous magnetic field and the time interval between a μ^+ entering the sample and its decay positron entering a detector nearby is recorded for millions of events. In the example shown in Figure 13.17, a fast precessional signal is seen to be superimposed on the slower muon decay. The precession frequency depends on the microscopic magnetic environment in which the muon finds itself, for example, a certain crystal lattice site or chemical species. The amplitude depends on the number of muons finding such sites, which may be related to their availability. The term μSR also includes longitudinal spin relaxation experiments carried out in zero field or with a bias field (anti) parallel to the spin direction and muon spin resonance where transitions between states are detected by the absorption of microwave power. Data acquisition may be triggered by a step change in the sample environment, for example, illumination from a pulsed laser. Negative muons are more likely to depolarize during capture by atoms and to undergo nuclear capture. When not captured, they probe magnetism at sites different from μ^+SR.

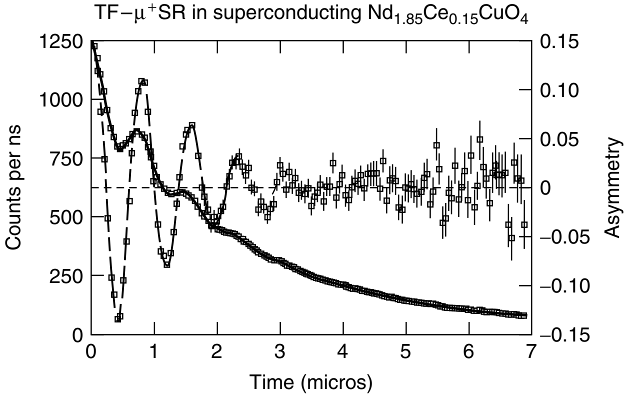

Fig. 13.17 Solid line: Positron counting rate from μ^+ stopped in a superconducting compound immersed in a 0.01 T transverse magnetic field. Dashed line: The same data corrected for background and muon decay. (Courtesy of J. Brewer, University of British Columbia).

Currently μSR is used as a probe of structure (e.g., measurement of magnetic penetration depth in superconductors, or the study of defect states in semiconductors) and dynamics (e.g., quantum diffusion in solids, or following μ^{+-} substituted radicals in chemical reactions).

Muons can bind certain nuclei sufficiently closely, for example, 700 fm for $^2H\mu^{-3}H$, that nuclear fusion may occur. The μ^- is released to participate in about 150 similar events before decay. Muon-catalyzed fusion may be a distance source of energy, and the process is being studied in detail at several laboratories.

The present meson-factory beam lines deliver μ^+ fluxes from 10^5 to 10^7 cm$^{-2\cdot}$ s^{-1}. Upgraded facilities at PSI will increase this severalfold, while the proposed kaon factories would provide a further increase.

The local structure of materials is also examined using heavy ions, for example, at the Hahn-Meitner Institute, Berlin, Germany. Hyperfine structure depends on both nuclear and atomic parameters and the latter are influenced by the solid-state environment. Nuclear probes, which may be constituent or impurity ions, are implanted and the hyperfine levels measured by nuclear techniques such as the Mössbauer effect or magnetic resonance.

13.5.4
Isotope Production

Radioactivity is a well-established field, discussed in Chapter 11, and radioisotopes have been used for decades, primarily as tracers and for nuclear medicine. Most commercial products incorporate isotopes from reactors. Neutron-deficient isotopes, however, require accelerators, and the accelerator of choice is a high-current cyclotron with variable-energy beam. The international trade in cyclotron products amounts to $100 million annually, mostly of radiopharmaceuticals, and is increasing. Some targets are used as irradiated – for example, in wear studies of machine parts – but usually the radioisotope has to be separated from the original target material and other reaction products and synthesized into a chemical compound. For example, ^{18}F-labelled 2-fluoro-2-deoxy-D-glucose is used to study glucose metabolism of the brain; ^{13}N- and ^{15}N-labelled compounds are used to study the fast and slow uptake of nitrates by plants. The investment in hot cells and radiochemical equipment can be compared to the cost of a cyclotron. The target, bombarding particle, and energy (Figure 13.14) are usually chosen to optimize the desired cross section and to reduce contamination from competing reactions. Table 13.1 lists several commercial isotopes and their production reactions.

Short-lived isotopes give a stronger signal with less environmental dose; the use of cyclotron-produced ^{123}I in thyroid studies reduced patient dose 50-fold compared with the longer-lived, reactor-produced ^{125}I or ^{131}I. Production, however, must be sited near a user or near excellent transportation facilities. The isotopes ^{11}C, ^{13}N, ^{15}O, ^{18}F, used in positron emission tomography, have half-lives of minutes. Fortunately they can be produced by low-energy (<20-MeV proton) cyclotrons that can be installed in a hospital. Production facilities at TRIUMF are 2.5 km from the university hospital; however, radiopharmaceuticals are transferred within 2 minutes by means of an underground pneumatic tube (rabbit).

In addition to research laboratories and hospitals with radioisotope production facilities, there are also several companies that make, process, and distribute isotopes and isotope-labeled compounds. These include Amersham International (UK), DuPont-Merck and Mallinckrodt

Tab. 13.1 Examples of popular cyclotron-produced medical isotopes.

Isotope	Production reaction	Beam	Use[a]
Thallium-201	$^{203}\mathrm{Tl}(p, 3n)^{201}\mathrm{Pb}$ $^{201}\mathrm{Pb} \longrightarrow {}^{201}\mathrm{Tl}$	29-MeV protons	Heart studies (SPECT)
Gallium-67	$^{68}\mathrm{Zn}(p, 2n)^{67}\ \mathrm{Ga}$	26-MeV protons	Soft tumor imaging (SPECT)
Iodine-123	$^{124}\mathrm{Xe}(p, 2n)^{123}\mathrm{Cs}$	30-MeV protons	Thyroid imaging and screening of acute stroke patients
	$^{123}\mathrm{Cs} \longrightarrow {}^{123}\mathrm{Xe} \longrightarrow {}^{123}\mathrm{I}$ and $^{124}\mathrm{Xe}(p, pn)$ $^{123}\mathrm{Xe} \longrightarrow {}^{123}\mathrm{I}$		
Xenon-127	$^{133}\mathrm{Cs}(p$, spallation)	High-energy, e.g. 500-MeV, protons	Lung agent
Carbon-11	$^{16}\mathrm{N}(p, \alpha)^{11}\mathrm{C}$	12–14-MeV protons	Brain function (PET)
Fluorine-18	$^{18}\mathrm{O}(p, n)^{18}\mathrm{F}$	~11-MeV protons	Brain function (PET)
Rubidium-82	$^{85}\mathrm{Rb}(p, 4n)^{82}\mathrm{Sr}$	70-MeV protons	Heart function (PET) and brain metastases
	$^{87}\mathrm{Rb}(p, 6n)^{82}\mathrm{Sr}$	85-MeV protons	
	$^{82}Sr \xrightarrow[25d]{} {}^{82}Rb \xrightarrow[1.2m]{\beta+} {}^{82}Kr$		

[a] SPECT: single photon emission computed tomography. PET: positron emission tomography.

(US), Nihon Medi-physics (Japan), and Nordion International (Canada). Most use proton beams below 50 MeV and have two or more cyclotrons to guarantee production. Isotopes from spallation reactions, for example, ^{82}Sr and ^{127}Xe, require much higher energies; however, they are sometimes made parasitically, for example, in targets placed just before the beam dump of a line with primary targets, devoted to other work, upstream.

A modern low- or medium-energy cyclotron can produce 15-kW beam power, which is more than most production targets can withstand. A multifoil H^- or D^- cyclotron can simultaneously share this beam between several targets at the same or different energies.

13.5.5
Particle Beams in Medicine

Cancerous tumors and other structures inaccessible to conventional surgery have been eradicated by using penetrating radiation to ionize and destroy malignant cells. Most treatments employ X rays or γ-rays. Some tens of thousands of patients, however, have been treated with neutron or charged-particle beams that, although less tractable, can offer improved localization of dose and interact differently with tissue. Proton and α-particle treatments were developed, starting in the mid-1950s, at the Berkeley, Uppsala, and Harvard synchrocyclotrons and at the ITEP (Moscow) synchrotron. Neutron studies, begun at Berkeley, were revived at the Hammersmith cyclotron in the 1960s. HI and π^- therapy began in the 1970s.

To reduce the dose to intervening tissue, a patient is irradiated from two or more sides with the beam energy adjusted to place the Bragg peak (Figure 13.15), where applicable, at the treatment site. The energy may be altered and the beam steered transversely to distribute the dose over an irregular volume. The choice of particle is influenced by

1. radiobiological effect (HI, π^-, and neutrons have high linear energy transfer (LET) at the treatment site),
2. the ratio of end-of-range to transit dose (favors π^-, protons and HI),
3. the absence of dose beyond a sharp range end (favors protons and α), and
4. dose delineation (primary beams are smaller than secondary).

A therapeutic dose is 10–50 Gy, delivered in several fractions. Moderate energies are required –250-MeV protons reach the deepest tumors – and since 1 nA cm^{-2} delivers 1 Gy min^{-1}, synchrocyclotrons or synchrotons are adequate for primary beams. Secondary beams, π^- or neutrons, require cw cyclotrons or linacs.

There are 20 neutron treatment centers worldwide, out of which 13 are with cyclotrons. Attention has been focused on neutron radiobiology, especially at hypoxic sites. A typical installation would accelerate 100 μA of deuterium to 50 MeV, and bombard a beryllium target to obtain 20- to 40-MeV neutrons that are then collimated into a fixed beam for treatment. A gantry-mounted, 50-MeV deuteron, superconducting cyclotron, however, designed by H. Blosser and installed at the Grace Harper Hospital, Detroit, is able to rotate completely around the axis of a patient (Figure 13.18).

About a dozen laboratories operate adjunct medical facilities performing radiosurgery by means of charged-particle beams. Charged-particle treatments have been especially advantageous where dose localization is important, for example, in the pelvic region, for tumors of the pituitary (adjacent to the spinal chord and

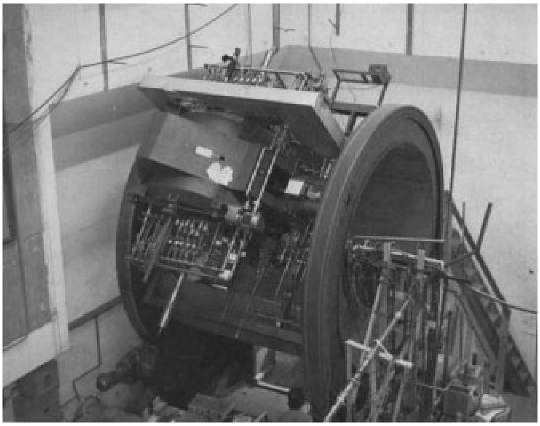

Fig. 13.18 A superconducting, 50-MeV deuteron, cyclotron being mounted on a cylindrical rotating gantry at the Harper Hospital (Detroit). Neutrons are directed toward a patient located near the cylinder axis. The patient position is adjustable and irradiation can take place from all sides. (Courtesy of H. G. Blosser, NSCL-Michigan State University).

brain stem), in choroidal melanoma (adjacent to the optic nerve), and in destruction of arteriovenous malformations in the brain. During the last 6 years, 1000 patients with ocular melanoma have been treated using 70-MeV protons from the PSI injector cyclotron. The success rate, 95%, is the same as for the traditional treatment, which involves removal of the eye, but proton treatment also preserves vision in most cases. Recent improvements in medical imaging (computerized axial tomography (CAT), magnetic resonance imaging (MRI), etc.) make it possible to exploit the 1-mm dose delineation possible with proton beams. Ten new charged-particle treatment facilities are under construction, half with dedicated medical accelerators, six associated with proton cyclotrons.

13.5.6
Elemental Analysis of Materials

Many physical and chemical methods are used to determine the elemental constituents of a material. Those associated with accelerators include

1. the identification of atoms from their characteristic X rays emitted under bombardment (particle induced X-ray emission (PIXE)),
2. the identification of nuclei from their decay products following irradiation,
3. the identification of nuclei from elastic scattering cross sections and angular distributions, and
4. mass analysis following ionization and acceleration of a small sample, a technique termed *accelerator mass spectrometry* (AMS).

These techniques are frequently nondestructive and may be considered when only a small sample (milligrams) is available, when isotopes, and not just the elements, must be identified, or when the spatial distribution within a matrix is required. Absolute resolutions are of the order 1 ppm or less, that is, subnanogram quantities; AMS isotopic ratios, for example, $^{14}C/^{12}C$, have a resolution of $\sim 10^{-14}$. They are primarily the domain of low-energy potential-drop machines; however, cyclotrons have advantages in certain cases and, worldwide, more than 10 cyclotron laboratories offer service to external users and some derive a substantial fraction of their income in this way.

The proton microprobe uses a collimated incident beam focused into a spot of $\sim$10 μm diameter, and X-ray analysis takes place as the beam slowly scans over the sample. The cross section for *K*-shell X-ray production is greater at cyclotron energies, but so is the background; *L*-shell X rays are also produced, however, which may resolve ambiguous cases. A cyclotron beam penetrates further into a sample; however, the spatial resolution is somewhat worse than the standard potential-drop laboratory arrangement. Cyclotron laboratories offering PIXE services include the University of Hamburg and the University of California, Davis Campus. Applications include analysis of thousands of specimens of airborne pollutants per day and the distribution of aluminum in biological tissue.

Proton-induced radioactivity is suited to the detection of trace amounts of light and medium-mass nuclei, whereas neutron activation offers better resolution for heavy elements. Some medium-energy cyclotron laboratories offer both techniques and also irradiation by other ions. Resolutions range from 1 ppb to 10 ppm, depending on the element of interest and the matrix. Commercial applications include the determination of oxygen in metals and impurities or dopants in semiconductors. Many laboratories engage in this work; the more experienced include the Radiation Studies Centre (CERI), Orleans, France; University of California, Davis (US), and RIKEN (Japan).

Cyclotrons have been used as mass spectrometers since, at any one time, only ions whose q/A lies within a narrow band can be accelerated and extracted. However, the measurement of a ratio, say $^{14}C/^{12}C$, for dating purposes would require changes in frequency or field, and potential-drop machines are preferred by AMS facilities that process many samples. The mass resolution of a conventional cyclotron, $\Delta m/m \sim 10^{-3}$, can be enhanced by further analysis of the extracted beam or by designing a machine with very tight synchronism tolerances. The Institute of Nuclear Research, Shanghai, is constructing such a compact and inexpensive cyclotron that can distinguish between ^{14}C, ^{13}CH, and $^{12}CH_2$.

The resolution of isobars, nuclei with the same mass, is effected by measuring another physical property, for example, stopping power. Such measurements are easier at the higher energy available from cyclotrons.

13.5.7
Transmission Radiography

The absorption and scatter coefficient for X rays increases smoothly with atomic number, whereas the coefficient for neutrons is scattered over 5 orders of magnitude and varies considerably from element to element, even from isotope to isotope. This selectivity enables neutron radiography to reveal the distribution of those elements with high cross section, for example, hydrogenous or nitrogenous explosives encased in materials of high *Z*. The isotopic sensitivity makes it possible to observe the distribution of enriched uranium in fuel rods. In other materials neutrons have high penetrability, and large castings can be inspected for voids and aircraft structures for stress cracks and corrosion. The object is backed by material to convert neutrons to electromagnetic radiation, and radiographs are made with X-ray film or pictures recorded in real time using a TV camera. The neutron flux from an easy-to-operate 50-μA, 20-MeV compact cyclotron approaches that from a 5-MW reactor. Japan Steel Works and Sumitomo Heavy Industries have both installed cyclotron-based radiography systems. Oxford Instruments (UK) manufactures a transportable, superconducting, 200-μA, 17-MeV H^- cyclotron. The machine consumes only 20 W, weighs 2 tons, and may be mounted on a "cherry picker" device. The flux of thermal neutrons is estimated to be $2 \times 10^{11}\ n\,\mathrm{cm}^{-2\cdot}\,\mathrm{s}^{-1}$.

Proton radiography is used less extensively; it requires energies above 100 MeV for a useful range. Marginal-range radiography depends on the rapid change in differential energy loss near the end of the Bragg peak. The proton energy is adjusted to match the thickness of the object and resolutions of 0.1% for objects a few millimeters thick and 1% for 50 mm are obtained, which is almost 10 times better than for X rays. In scatter radiography, a broad, parallel beam illuminates the object. Protons passing through the object near the edge will be scattered into the unperturbed stream, enhancing the delineation of the edge. This subject has been developed extensively at the Harwell 160-MeV synchrocyclotron.

These techniques leave little residual activity and are thus nondestructive. Since neither protons nor neutrons are common products of radioactive decay, the methods discussed may be used, with appropriate detectors, on radioactive objects.

13.5.8
Wear and Corrosion

The rate of wear caused by operation of equipment can be measured by irradiating part of the surface that is subject to wear, and then reassembling and operating the equipment while monitoring the radioactivity. In some cases, abraded material may be carried by lubricants or other fluids past a radiation detector to give a dynamic measure of wear as a function of machine operating conditions. In cases where this is not possible, for example, locomotive wheels, the activity remaining is measured periodically while logging the amount of use. The amount of activity required is small; the irradiated area may be large or as small as 1 mm and limited to 0.1 mm. Thickness changes may be measured to 1 nm. Different radionuclides may be created on different components of the same machine and relative wear observed in a single experiment. The Kernforschungzentrum Karlsruhe (Germany) cyclotron has treated 500 parts per year for many years and for different customers. A recent development in this field is the implantation of radioactive ^{7}Be and ^{22}Na carried out at NSCL-MSU.

13.5.9
Materials Modification and Processing

The dislocations, ionization, and reaction products from ion bombardment can alter the physical properties, for example, conductivity, of a target and also lead to chemical change. Normal chemical incompatibilities may be overcome and unusual alloys and compounds may be formed at low temperature and in the solid state. The bombarding energy may be chosen to concentrate the changes at the surface or in a layer buried beneath. Chemical change may be minimized by using one of the target elements as the ion beam; conversely, physical dislocations can often be removed by annealing. The $K \approx 100$ cyclotrons at the Oak Ridge National Laboratory, Harwell, KfK Karlsruhe, and the Joint Institute of Research, Dubna, have exploited such processes for decades. The newer, high-K machines, in particular, the chain of cyclotrons at GANIL, Caen, France, extend the range of ion species and energies, and consequently the depth of dose and LET, available. Irradiations may take place under vacuum, but the majority of them are carried out in air, the beam emerging through a thin metal window at the end of an evacuated beam pipe.

Light- and heavy-ion beams are used to render stainless steel surfaces more resistant to corrosion, friction, and wear. Heavy-ion and α particle beams are used to simulate the damage caused by neutrons in fission or fusion reactors. The neutrons cause dislocations in structural materials, and helium from (n, α) reactions accumulates in the voids. Ion bombardment produces similar effects, but at much higher rates, and also allows mechanical tests to be performed during irradiation.

Ionizing particles may cause intermittent malfunctions in electronic circuits; heavy exposure may lead to failure. Ion beams are used to measure the radiation resistance of military or space-bound electronics in the laboratory. The French agency MATRA ESPACE simulates space radiation effects by using 5–50 MeV per amu carbon to xenon ions from the GANIL cyclotron to obtain penetration depths from 5 to 1300 μm and LET between 0.2 and 68 MeV $(\mathrm{mg\,cm^{-2}})^{-1}$.

Micropore filters are used to remove contaminants from fluids (e.g., the water used in printed circuit manufacture), to separate biological cells, and as semipermeable membranes to preserve the freshness of fruit and vegetables. Superior filters (Figure 13.19) are manufactured at Dubna and also by the company Biopore using beams from GANIL. Rollers feed polycarbonate film, 10–50 μm thick, through $10^{12}\ \mathrm{s^{-1}}$ beams of 5–10 MeV $\mathrm{amu^{-1}}$ xenon or krypton and then into etching and washing solutions. These preferentially dissolve material around the ion path to produce cylindrical holes with a high degree of uniformity and a diameter determined by choice of ion species, energy, and etching process. Diameters range from 0.05 to hundreds of microns. Filters may have 10^9 holes $\mathrm{cm^{-2}}$, whereas membranes used for cell counting have a single hole.

Embryonic commercial processes include the generation of color centers in crystals, and the manufacture of field-emission surfaces and nonreflective glass and plastic. Multi-MeV beams can bury oxide or nitride layers deep in a semiconductor, thus permitting the development of three-dimensional integrated circuits, optical circuits in semiconductor sandwiches, and the micromachining (via etching) of mechanical devices in silicon.

13.6
Facility Design and Shielding

Cyclotron laboratories employ conventional construction methods for the most part, but have the additional requirements of shielding personnel and experimental equipment from primary and secondary particles and of monitoring and minimizing radioactivity in the workplace, at points of egress, and in air and water exhausts.

For beam energies above 10 MeV $\mathrm{amu^{-1}}$ a shielding wall thick enough to attenuate the neutrons produced to a safe level will also attenuate γ rays and stop charged particles. The number of fast neutrons rises rapidly with beam energy (Figure 13.20). They are slowed, initially, by nuclear elastic or inelastic scattering in the shield and some nuclear reactions may occur. They reach thermal equilibrium at some distance from the point of origin and are

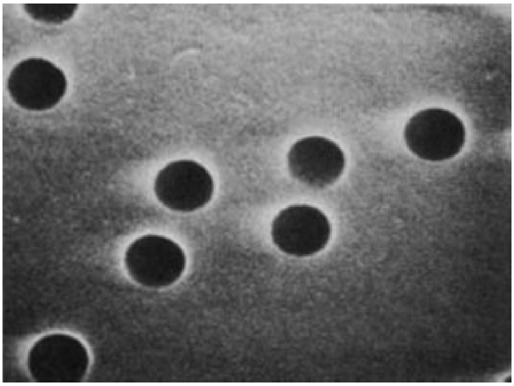

Fig. 13.19 Holes with a 0.2-μm diameter produced in polycarbonate film by heavy-ion bombardment followed by etching. (Courtesy of C. Bieth (GANIL) and Biopore S.A., France).

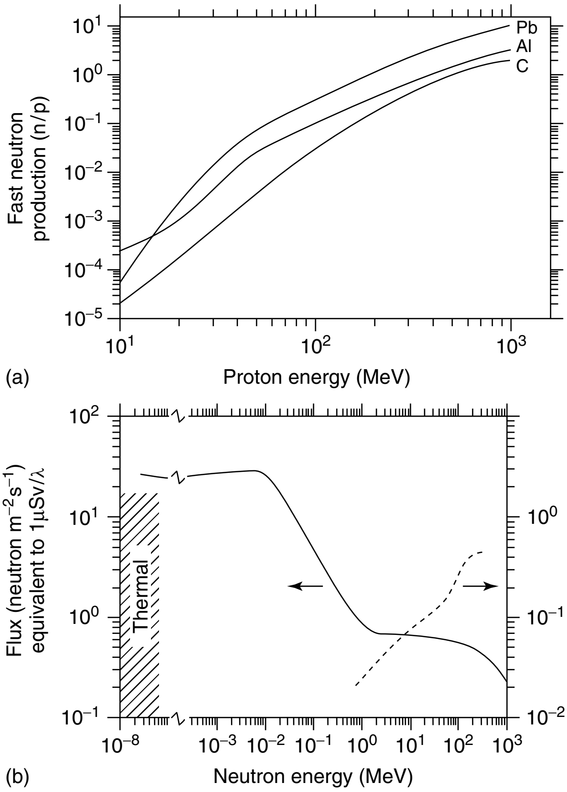

Fig. 13.20 (a) Yield of fast neutrons produced by protons stopping in thick targets. (b) (Solid line) Neutron flux giving unit dose and (dashed line) the thickness of concrete (density 2.4 g cm^{-3}) required to halve the neutron flux as a function of neutron energy.

ultimately captured in a nucleus, usually with the release of a γ ray that may necessitate an additional layer of shielding.

The maximum permissible dose for radiation workers is currently 0.05 Sv/year. This may be transformed to neutron flux and an approximate estimate of shielding thickness made with the aid of Figure 13.20(b). While hydrogenous materials have high neutron-scattering cross sections, materials with high atomic number are better attenuators of γ rays. Concrete is the usual low-cost compromise. Shielding calculations can be quite complicated since both particle production and absorption are multibranched energy-dependent processes, and Monte Carlo simulations are often used.

In practice the greatest dose to humans comes from induced radioactivity, from activities such as processing targets, and from cyclotron or beam line maintenance. Some beam loss is inevitable; good design minimizes the amount and concentrates its deposition in absorbers such as carbon, which minimize residual activity, and where the loss can be monitored by cyclotron operators. Particles likely to be lost, that is, those with large betatron amplitude or extreme phase, may be trimmed at low energy. The absorbers may be removed quickly and safely during maintenance or themselves shielded by a portable lead curtain. Lead of 10 cm thickness attenuates 10-MeV γ rays 100-fold and stops β and α particles. Impurities in construction materials can cause significant residual radiation problems; two important examples are sodium in concrete (which may be a small percentage unless controlled) yielding ^{24}Na and trace amounts of cobalt in steel producing ^{60}Co. The vault and high-intensity beam lines should be sealed and volatile radiation products and airborne radioactivity pulled through a monitored exhaust. The cooling water for components in areas of high loss should recirculate through ion exchangers to remove products of radiolysis. Personnel exposure should be considered at the equipment design stage. For example, screws may be superseded by quick-release clamps, quick disconnects used for electrical and fluid services, and remote removal and handling features incorporated.

Most organic materials and electronic equipment begin to deteriorate at an accumulated dose of 10^4 Sv. Some are more resistant, for example, ethylene/propylene O rings (10^5–10^6 Sv), fiberglass, and certain epoxies (5×10^6 Sv). At higher levels inorganic insulators and metal vacuum seals are used.

Ancillary equipment, such as production targets or particle detector arrangements, is reconfigured quite frequently and may not be commissioned to the same level of reliability as the cyclotron. Some targets may be more radioactive than the cyclotron itself. These facilities are best placed outside the cyclotron vault in their own shielded rooms that can provide low-background areas for experiments requiring them, easier access for maintenance of both cyclotron and equipment, and, when magnets can switch the beam between two or more such areas, a more productive use of cyclotron time. The switching magnets and any initial collimation or common beam-focusing elements are best placed in the cyclotron vault. Site plans for all types of facilities may be found in the proceedings of conferences; Figures 13.13 and 13.21 are examples. While poured-in-place concrete is cheaper and services may be conveniently attached, movable blocks are commonly used to subdivide areas and for some outer walls

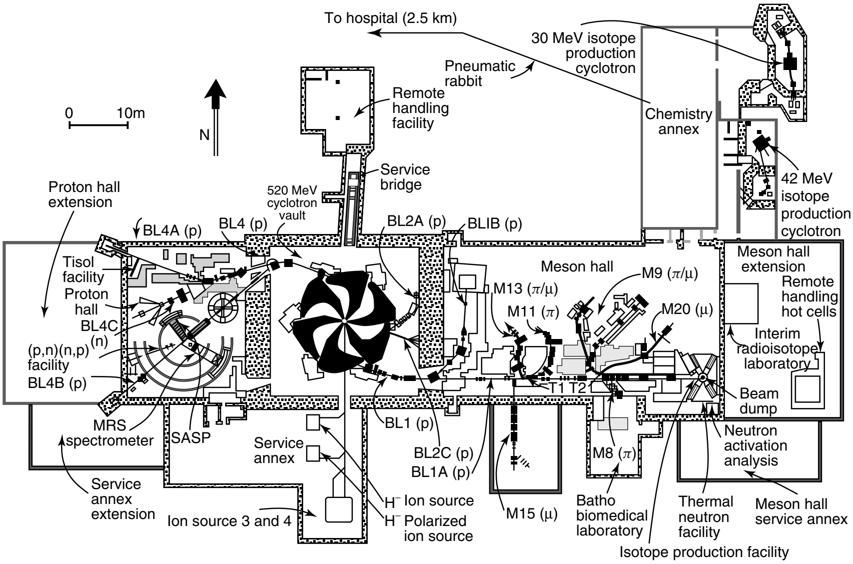
To hospital (2.5 km)
30 MeV isotope production cyclotron
Pneumatic rabbit
Remote handling facility
0 10m
N
Chemistry annex
Service bridge
Proton hall extension
520 MeV cyclotron vault
42 MeV isotope production cyclotron
BL4A (p)
BL4 (p)
BL2A (p)
BLIB (p)
Meson hall extension
Tisol facility
Meson hall
M9 (π/μ)
Remote handling hot cells
Proton hall
M13 (π/μ)
M11 (π)
M20 (μ)
BL4C
(n)
Interim radioisotope laboratory
(p,n)(n,p) facility
BL4B (p)
T1 T2
Beam dump
MRS spectrometer
SASP
BL1 (p)
Neutron activation analysis
Service annex
BL2C (p)
BL1A (p)
M8 (π)
Service annex extension
H⁻ Ion source
Batho biomedical laboratory
Thermal neutron facility
Meson hall service annex
Ion source 3 and 4
H⁻ Polarized ion source
M15 (μ)
Isotope production facility

Fig. 13.21 TRIUMF site plan showing the 520, 42, and 30 MeV H^- cyclotrons. The meson production beam line, BL1 A, services secondary particle channels M8 (π^- therapy), M20, M15 and M13 (μSR), M9 (particle physics), and M11 and M13 (π^- nuclear physics). Spallation isotope production and neutron applications take place at the beam dump. The dump window limits proton currents to 200 μA. Beams of energy 65–100 MeV are used in BL2 C for isotope production. BL4 A, 4B, and 1B accept beams from 180 to 520 MeV with maximum currents of 10, 0.2, and 0.01 μA, respectively. Polarized H_2 and $^2H_2$ targets in 4 A are used for nucleon–nucleon physics; a liquid $^2H_2$ target produces polarized neutrons for BL4 C. The magnetic spectrometers MRS and SASP are used for nuclear physics studies. Large equipment is assembled and checked out in the E and W extensions; the whole area is serviced by a pair of 50-ton cranes. Power supplies, data acquisition and control equipment, control rooms, and offices are located at several levels in the service annexes and the north side of the meson hall. Hot cells for isotope separation and manufacture of radioisotopes are located in the chemistry annex. Speckle indicates concrete poured in place; the unshaded shielding is movable concrete blocks.

in order to facilitate future rearrangement and expansion.

Appendix 13.A Description of a Typical Isotope-Production Cyclotron

The TR30 500-μA, 30-MeV H^- cyclotron (Figure 13.A.1), built by Ebco Technologies, Vancouver, Canada, is located at the TRIUMF site (Figure 13.21), and operated by Nordion for isotope production.

The compact design has four radial sectors. The magnet is approximately square in shape, 2.3 m from side to side and 1.26 m high, and weighs approximately 46 tons. It is split at the midplane, allowing four hydraulic jacks located in the magnet supports to elevate the upper part for access to the cyclotron interior. Two 37 000 A-turn coils mounted on the upper and lower poles provide the magnet excitation. No trim coils are used. Tolerances for isochronism, vertical focusing, and avoidance of resonances were met by a

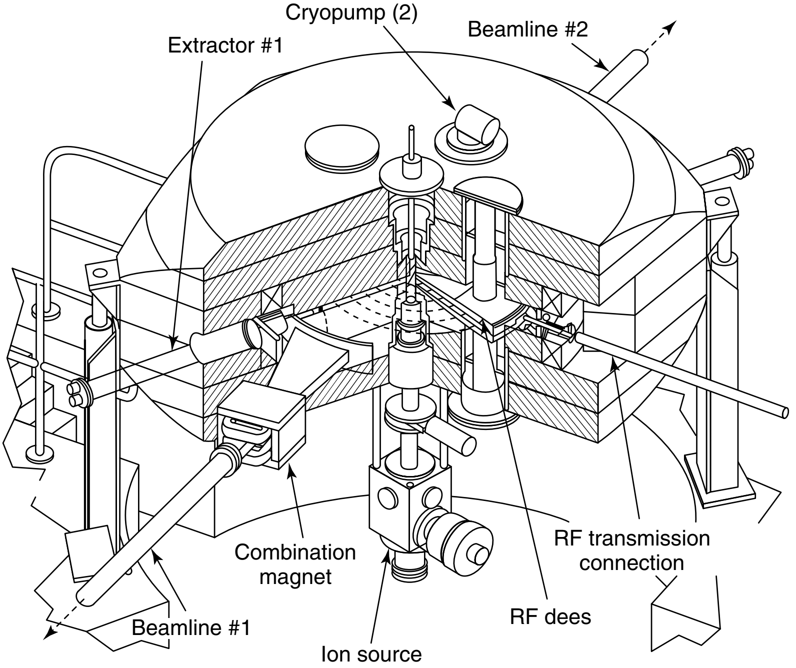

Fig. 13.A.1 Major components of a 30-MeV H$^-$ cyclotron intended for isotope production. (Courtesy of Ebco Technologies, Canada).

slight modulation of hill gap and by shims attached to the hill sides; the latter received their final machining following magnetic measurements.

The cyclotron is mounted over a pit to allow installation of the external H$^-$ ion source below and thus minimize cyclotron-vault head-room requirements. The H$^-$ beam is injected vertically upward toward an electrostatic spiral inflector that bends it into the median plane. Two 45° dees located in opposite valleys operate at 73 MHz, the fourth harmonic of the orbit frequency, and provide acceleration at four gap crossings per orbit. The dee voltage is 50 kV. Radiofrequency power is delivered to the dees through a capacitive coupling to a 50-Ω transmission line that passes through a port in the vacuum-tank wall. For ease of maintenance, the entire 35-kW rf amplifier system is located outside the cyclotron vault.

Four large holes through the yoke in the dee valleys accommodate the coaxial dee resonator stubs. To maintain fourfold magnetic symmetry, there are four additional holes in the unoccupied valleys. Two of these are used as vacuum pump ports in which two 8-in. cryopumps are installed. The vacuum enclosure is defined by the nickel-plated upper and lower pole surfaces and a cylindrical aluminum wall that is sealed to the poles by a double O-ring gasket. The operating vacuum is 7×10^{-10} bar.

Thin graphite stripping foils mounted on two extraction probes traveling in opposite hill gaps deliver two independent,

Tab. 13.A.1 Principal parameters for the TR30 cyclotron.

Magnet	Average induction	1.20 T
	Hill induction	1.90 T
	Valley induction	0.55 T
	Hill gap	4 cm
	Valley gap	18 cm
	Pole radius	76 cm
RF	Frequency	73 MHz
	Harmonic	4
	Dee voltage	50 kV
	Power (for 500 μA)	32 kW
Ion source	Type	H^- cusp
	Output current	5 mA
	Biased at	25 kV
Injection	Energy	25 keV
	Method	Axial
	Inflector field	20 kV cm $^{-1}$
Extraction	Energy	15–30 MeV
	Method	Stripping
	Number of extracted beams	2

variable-energy external beams. The basic cyclotron parameters are given in Table 13.A.1.

Acknowledgment

The authors gratefully acknowledge the assistance of those colleagues who provided information for this chapter and helped with proofreading.

Glossary

amu: Atomic mass unit, $\frac{1}{12}$ of the mass of the carbon isotope ^{12}C.

AVF Cyclotron: An isochronous cyclotron with an aziumthally varying field to provide focusing and thus permit cw operation at relativistic energies.

Betatron Motion/Frequency: Oscillations in the transverse (axial (z), radial (r)) planes executed by the particles constituting a focused beam. The frequency of such oscillations expressed in terms of the particle rotation frequency ($Q_{z,r} = f_{z,r}/f_{\text{rot}}$).

Bunch: A collection of particles accelerated within one rf cycle.

cw Cyclotron: One with a constant accelerating frequency, and hence a continuous pulse train and macro duty factor of 1.

Dee: A hollow electrode, semicircular in early cyclotrons, carrying the rf accelerating voltage. Part of the beam path lies within the dee, acceleration occurring at entrance and exit.

Emittance/Acceptance: A quantitative description of the quality of a beam. The transverse emittance is a combination of the beam size and divergence; the longitudinal emittance, a combination of momentum spread and bunch length. The acceptance of an accelerator corresponds to the emittance of the beam that would just fill it.

Equilibrium (closed) Orbit: A conceptual unaccelerated particle trajectory closing smoothly on itself after one turn. Betatron oscillations can be imagined to take place about this orbit.

FM Cyclotron or Synchrocyclotron or Phasotron: A cyclic particle accelerator in which the accelerating frequency is varied during acceleration to match the particle rotation frequency. The beam is produced in a series of pulses with macro duty factor (0.1–10)% at repetition rates from 0.02 to 2 kHz.

Isochronism: Equality between particle rotation frequency and accelerating frequency.

K-Factors: Parameters associated with a cyclotron magnet. The maximum energy of an ion, charge qe, mass A, which can be contained in a closed orbit is $((T/A) = K_b(q/A)^2$ MeV amu^{-1}; the maximum energy which can be focused axially is $(T/A) = K_f(q/A)$ MeV amu^{-1}.

Macro Duty Factor: The ratio of pulse length to repetition time.

Meson Factory: An accelerator laboratory where intense beams of particles are produced whose energy is well above the threshold for meson production.

Particle Microamp: Cyclotron beam intensities are usually expressed as an electric current, for example, 1 μA being 6.24×10^{12} electronic charges/s. For an ion of charge qe the intensity may be expressed in particle microamperes where 1 pμA is 6.24×10^{12} particles s^{-1}, the electrical current being q times larger.

Pulse (Train): A sequence of bunches.

Stripping: Removal of some, or all, of the atomic electrons from an accelerated ion.

Synchrotron Oscillations: Oscillations in energy and phase about those of an ideal, synchronous particle.

References

Blosser, H.G. and Johnson, D.A. (**1974**) *Nucl. Instrum. Methods*, **121**, 301–306.
Gordon, M.M. (**1984**) *Particle Accl.*, **16**, 39–62.
Hagedoorn, H.L. and Verster, N.F. (**1962**) *Nucl. Instrum. Methods*, **18, 19**, 201–228.
Lawrence, E.O. and Livingston, M.S. (**1932**) *Phys. Rev.*, **40**, 19–35.
McMillan, E.M. (**1945**) *Phys. Rev.*, **68**, 143–144.
Richardson, J.R., MacKenzie, K.R., Lofgren, E.J., and Wright, B.T. (**1946**) *Phys. Rev.*, **69**, 669–670.
Rickey, M.E. and Smythe, R. (**1962**) *Nucl. Instrum. Methods*, **18, 19**, 66–69.
Syme, D.B., Wood, E., Blair, I.M., Kew, S., Perry, M., and Cooper, P. (**1978**) *Int. J. Appl. Radiat. Isot.*, **29**, 29–38.
Thomas, L.H. (**1938**) *Phys. Rev.*, **54**, 580–590.
(a) Veksler, V. (**1944**) *C. R. Acad. Sci. USSR*, **43, 44**, 329–331, 365–368; (b) (**1945**) *J. Physics (USSR)*, **9**, 153–158.
Willax, H.A. (**1963**) *Proceedings of the International Conference on Sector Focused Cyclotrons, 1963*, (eds F.T. Howard and N. Vogt-Nilsen), CERN, No. 63-19, pp. 386–397.

Further Reading

Humphries, S. Jr. (**1986**) *Principles of Charged Particle Acceleration*, Wiley, New York.
Kapitza, S.P. and Melekhin, V.N. (**1978**) *The Microtron*, Harwood Academic, New York.
Livingston, M.S. and Blewett, J.P. (**1962**) *Particle Accelerators*, McGraw-Hill, New York.
Scharf, W. (**1986**) *Particle Accelerators and their Uses*, Harwood Academic, New York.

The best sources for detailed information are the proceedings of international accelerator conferences. The most recent in each series are the following:

Bennett, F. and Kopta, J. (eds) (**1989**) *Proceedings of IEEE Particle Accelerator Conference*, IEEE Publishing Services, New York.
Duggan, J.L. and Morgan, I.L. (eds) (**1991**) Proceedings of the 11th International

Conference on the Application of Accelerators in Research and Industry, *Nucl. Instrum. Methods*, **B56/57**.

Marin, P. and Mandrillon, P. (eds) (**1990**) *Proceedings of the Second European Accelerator Conference*, Editions Frontières, Gif-sur-Yvette.

Martin, B. and Ziegler, K. (eds) (**1991**) *Proceedings of the Twelfth International Conference on Cyclotrons and their Applications*, World Scientific, Singapore.

DATA-ACQUISITION SYSTEMS

WALTER N. GEKELMAN, *Department of Physics, University of California, Los Angeles, California, U.S.A.*

Introduction 463
1. **Definition and Critical Components of a Data-Acquisition System** 464
2. **The Evolution of Data-Acquisition Systems** 465
2.1 NIM 465
2.2 CAMAC 465
2.2.1 CAMAC Bus 466
2.3 GPIB and VME 467
2.4 Microprocessor Control 468
2.4.1 CAMAC Processor Connections 468
2.4.2 CAMAC LAN Connection 469
2.5 FASTBUS 469
2.6 Speed 469
3. **Triggers** 470
3.1 Specialized Instrumentation 470
4. **Digitizers** 471
4.1 Resolution 471
4.2 Sampling Rate 471
4.3 Memory 473
5. **How a Data-Acquisition System Works** 473
5.1 Software 473
5.1.1 Interrupt Handlers 474
5.2 Sample Program 474
5.3 Processor–Crate Communication 475
6. **Speed and Response of Data-Acquisition Systems** 476
6.1 Transfer Rate to Host Computer 476
6.2 Real-Time Analysis 476
6.3 Choice of Operating System 478
7. **An Example of a Data-Acquisition System** 478
8. **Data Management and Visualization** 480
9. **Future Trends** 481
9.1 Communication 481
9.2 Computer Farms 481
9.3 Parallel Computers 481
9.4 Inexpensive, Simple Systems 481
10. **Summary** 482
Acknowledgments 482
Glossary 482
Works Cited 482
Further Reading 483

INTRODUCTION

For the most part, the world we live in is complex, and most events stem from situations which have many degrees of freedom. One approach we use to deal with this is to average over quantities; for example, thermodynamics is used to attempt to describe a system in terms of a few variables such as temperature, pressure, and flow velocities. Even in systems such as these, variables of interest may vary rapidly in space and time. If these variables are recorded in any detail, a data set for a given case can easily contain tens of millions of numbers. As an example, consider an experiment in plasma physics done by the author (Gekelman *et al.*, 1987) in which a fully three-dimensional magnetic field changes its topology rapidly (within 100 ms). The vector field is measured at 4000 locations and sampled at thousands of time steps at each position. Magnetic-field data sets in excess of 10^9 numbers are commonplace. If the experiment is fully characterized by measurement of the electric field, $\mathbf{E}(\mathbf{r},t)$, and particle distribution functions, $f_\alpha(\mathbf{r},\mathbf{v},t)$, gigabyte (GB) data sets are collected for each set of initial conditions. This enormous influx of data is becoming the rule in experimental physics as elaborate detectors capable of detailed measurement are designed. This trend in measurement is most welcome since with it comes the opportunity for detailed comparison with theory, but it is no accident that the theories and

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

experimental techniques have evolved in parallel with data-acquisition technology.

The list of environments which rely heavily on data-acquisition systems is nearly endless, but the following are a few examples: plasma physics and fusion, high-energy physics supercolliders, wind tunnels, laser physics and optics, process monitoring and control in manufacturing, etc. Many devices, such as commercial aircraft and automobiles, have *intelligence* built into them. This is nothing other than one or more sensors attached to a microprocessor which uses the data collected to feed back to the machine operation and report malfunctions. As these data-acquisition systems become cheaper, more and more ordinary devices will contain them. One day we may have "smart" toasters.

The data-acquisition systems (DAS's) that have evolved to meet these demands can range from relatively simple turnkey commercial packages to incredibly complex ones. The former are used in small laboratory experiments, and the latter are used at facilities such as CERN, where high-energy physics detectors with several hundred thousand channels deliver 500 MB of raw data per second (Rüden, 1989). This article will introduce the components within a DAS and attempt to explain how they interact with the experiment and with each other, and finally how data might be stored and reduced. There are, of course, a variety of approaches to this which depend on the speed required, the number of data channels, and on-line processing requirements. Several of these will be reviewed. Examples of DAS's of varying complexity will be presented, and in doing so we will discuss several important elements, such as digitizers, pulse counters, clocks, and control devices. Finally, we will speculate on what will happen in this arena in the near future.

1. DEFINITION AND CRITICAL COMPONENTS OF A DATA-ACQUISITION SYSTEM

A data-acquisition system is anything which accepts a signal from a detector and stores it so that it may be analyzed at a later time. Its basic elements are shown in Fig. 1. Here, the signal source is shown to occur in an experiment, but this definition is too narrow. Anything worth understanding or diagnosing serves as the experiment. This could be the chemicals pouring from a smokestack, the temperature and gas mixture in an automobile engine, or the vital signs of a patient on an operating table. The detector is a device which responds to something within the experiment (fluid flow, temperature, presence of an electric charge, etc.), and usually presents a voltage at its output which is related to that thing. Sometimes the detector output is easily interpreted. The output voltage of a thermocouple is directly proportional to the temperature of the thermocouple. In other cases, the detector output bears no simple relationship to the measured quantity and has to be analyzed on a computer. This is the case for Thomson laser scattering diagnostics in fusion experiments, where the final results are the plasma electron temperature and density. Since detectors vary enormously from case to case, little more will be said about them here. A detector may produce an output signal which is a time-dependent voltage, charge, or current or optical signal. In plasma physics research, for example, sensors typically generate output voltage wave forms. Data acquisition consists of digitizing and recording these rapidly varying analog signals, with a voltage-sensitive

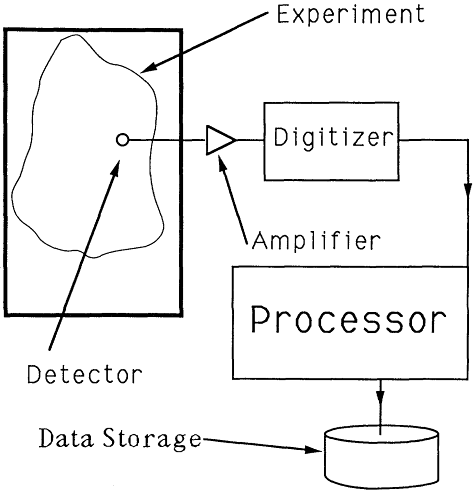

FIG. 1. The basic elements in data acquisition. A detector generates a signal which is usually conditioned by an amplifier and then input to a digitizer. The analog signal is converted to a binary number and then sent via a processor to storage media such as a computer disk or tape.

device, after a start trigger, or before a stop trigger is issued. The gain of any signal-conditioning amplifier is a dimensionless number since the sensor output and digitizer input are both voltages.

In high-energy physics experiments, photomultipliers or wire detectors are used to record the passage of particles. These events produce pulses of charge at the detector output, although the signal generated by each varies and must be separately conditioned. In this case analog-to-digital converters with charge-sensitive input amplifiers are used to encode the signal. The charge present at the amplifier input is integrated over a short period of time ($50\text{ ns} \leqslant t_{\text{gate}} \leqslant 2\ \mu\text{s}$) determined by an input gate pulse. The output of the digitizer is proportional to the number of picocoulombs (pC) of input charge. This is very different from a digitized wave form which may represent the time history of a magnetic field. The gain of charge-sensitive amplifiers is expressed in units of V/pC; that is, the peak output voltage is proportional to the charge received. A third type of conditioning amplifier, the transimpedance amplifier, takes current as the input signal and produces an output voltage proportional to it. The gain of this device has units of V/A or resistance. These amplifiers may have to operate in an electrically noisy environment, and elaborate shielding, or the transformation of the electrical to an optical signal for the purpose of transmission, may be necessary at this stage.

The next operation is signal storage. Sometimes the signal is processed in real time before storage to save storage space. There are a number of ways to accomplish this, and some will be discussed in the next section. In this article, we will concentrate on digital data storage. This involves the transformation of the analog voltage to a time series of digital numbers sampled at various positions on the input wave form. Although, as will be discussed later, information is always lost during sampling, the digitized signal is perfect for use by a computer. In most cases, digitizers have their own local memory and the data must be subsequently transferred to a computer where it may be analyzed mathematically, plotted, or stored on a disk or tape after acquisition.

2. THE EVOLUTION OF DATA-ACQUISITION SYSTEMS

2.1 NIM

Until the 1960s, commercial products such as digitizers did not exist, and the only way to record nonrepetitive, rapidly varying data was by taking pictures from an oscilloscope screen. The budget of a research laboratory may have included tens of thousands of dollars per year for fast Polaroid film. Data analysis involved laboriously measuring the amplitude of the recorded signals from the scope traces with a ruler and entering them by hand into computers when necessary. This meant that experimental researchers generally eschewed the use of computers, and data were sparse. Most experimental results were easily plotted as line drawings. Needless to say, detailed comparison between experiment and theory was not possible.

There was, however, a need for uniformity in the measurement equipment utilized in large high-energy physics experiments. A power supply for devices such as counters and amplifiers with a fixed instrument input voltage and specifications for signal input and output levels was agreed upon by the Nuclear Instrumentation Methods (NIM) committee in the United States. The power supply was cratelike in appearance and allowed for the easy insertion and removal of a variety of modules. The NIM crates had no provision for data readout and any device control had to be done via inputs on the devices themselves.

2.2 CAMAC

The biggest breakthrough in data acquisition occurred with the advent of large-scale integrated (LSI) circuit chips and the introduction of microcomputers which relied on them. This new technology allowed for the control and readout of devices in crates and an interface standard for CAMAC (computer automated measurement and control) was developed in 1969 by the European Standards on Nuclear Electronics Committee and then with further collaboration from the NIM committee in the U.S. (Horelick and Larsen, 1976; U.S. NIM Committee, 1981). The CAMAC standards were updated several times to allow for multiple crate systems that could be connected in parallel and series, and to allow for

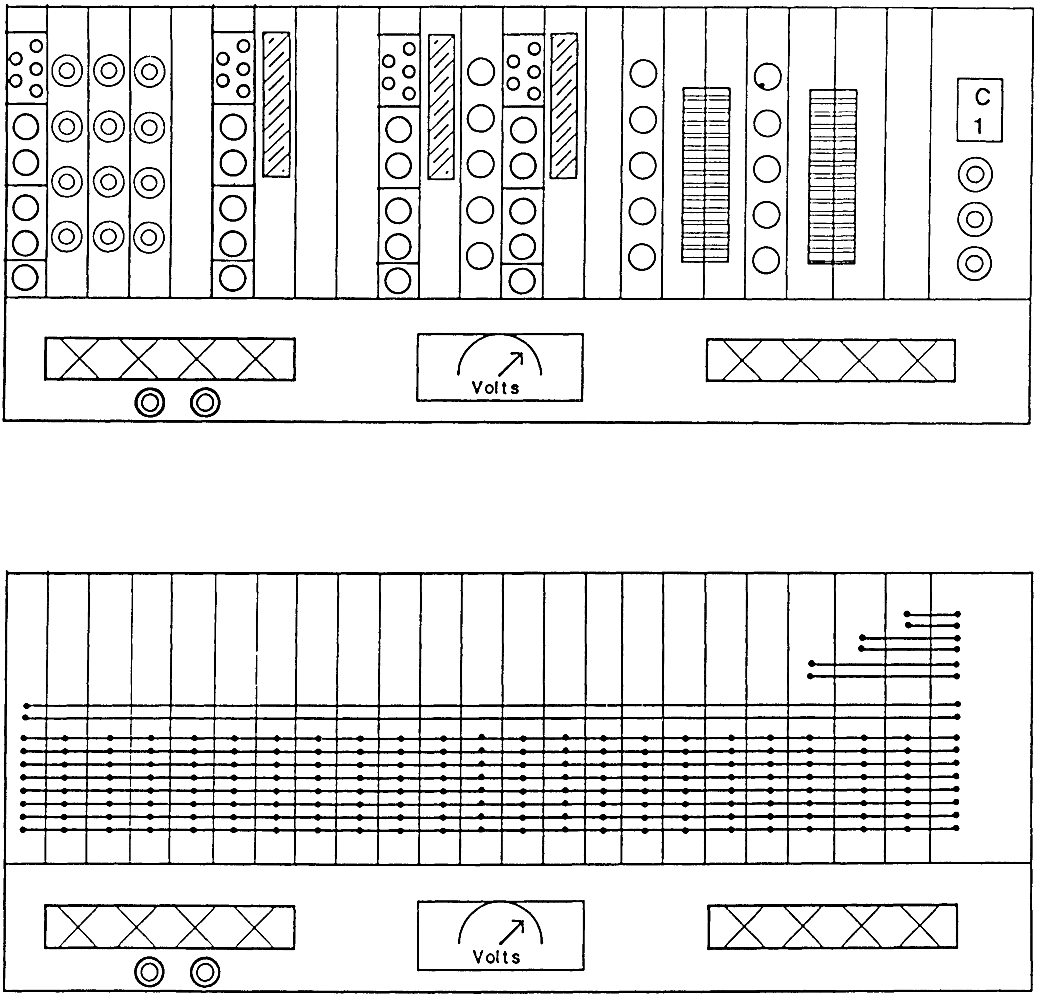

FIG. 2. (a) The front view of a CAMAC crate. There are a number (usually 25) of slots into which various devices such as analog-to-digital converters, memory modules, or clocks (see Table 1) may be placed. The leftmost slots are reserved for the crate controller module(s). (b) A view of the bus arrangement for a CAMAC crate. The top shows four of the point-to-point wires. These go from the controller to each slot and are used for individual addressing. One of the wires is the *N* line, used for station addressing, while the second is the *L* line, used for interrupts. The bussed wires, shown on the bottom, connect all the modules and are used for data transmission.

distributed processor intelligence. These techniques will be discussed in this article. The latest standard was issued by the Institute of Electrical and Electronic Engineers (IEEE, 1982).

CAMAC crates differ from NIM crates in that the former have a built-in computer bus called a CAMAC dataway. The crate has as many as 25 slots, each a standard size, with power and signal contacts at its rear specified by the IEEE convention. The slots are made to house data-acquisition or experimental control devices [Fig. 2(a)]. The variety of devices that can be inserted into CAMAC crates is large and ever growing. An example, which is not exhaustive, is provided in Table 1.

2.2.1 CAMAC Bus. The first slot, or the first and second slots, in a CAMAC crate is usually occupied by a *crate controller,* a microprocessor enabling it to communicate with external computers or other crates. On the back of the slot is an 86-pin connector to the dataway [Fig. 2(b)]. The dataway has two point-to-point wires for each station within it, as well as 24 read and 24 write lines which are shared by all the modules. These lines permit 24-bit parallel data transfers. The point-to-point lines are used to send commands to an individual station or to allow that station to send a signal, such as "now I have data, look at me," a LAM. Other bus lines concern themselves with clock signals and status information about data transfers. The clock signals, or strobes, are generated by the crate controller, which is also the source for all commands to the devices in the crate as well as a funnel through which the device's data move to an external computer. All the CAMAC devices are controlled by issuing instructions to the crate controller, which in turn reissues them to the entire crate or an individual device within it. The commands have been given standard names but always result in words with specific bit patterns being sent along the dataway.

2.3 GPIB and VME

At some point, the crate must interact with a computer to get instructions and send data for storage. The simplest interface is an RS-232 serial port. For this interface, one must buy an RS-232 module for the crate and plug the output from this module into the com-

Table 1. A sample list of devices used in CAMAC and FASTBUS crates.

Device	Purpose	Properties
Analog-to-digital converters	Analog-to-digital signal	16 bit 30 kHz to 8 bit 1 GHz
Programmable amplifier	Computerized signal gain	125-MHz gain, offset, impedance
Digital voltmeter	Digitize many (64) input channels	Slow, 1 mV to 10 V autoranging
Programmable clock	Dynamically control digitization rates	Multiple outputs, 5 ns to 1 ms/tick
Memory modules	Storage of data from digitizers	May be partitioned up to 1 Mbyte each
Stepping motor controllers	Digital control of motors	To move detectors in experiments
Digital-to-analog converters	To operate relays, analog device control	Multiple output channels, 16 bits in
Programmable trigger generator	Generates reliable recording and stop triggers	Variety of triggers, multiple outputs can be combined
Multichannel input gates	Determines electrical state of contacts	32 or more channels; 0=closed contact
Spincoder	Measures shaft rotation	Can feed back to a stepping motor
Gated integrators	Integrates charge	Multichannels (32) for phototube output
Multiplexer	Combines up to 16 input channels	600-ns switch time sequence by computer
Contact closure	Closes physical relays	12 contacts controlled by computer command
Histogrammer	Creates histogram from memory data	10^6 inputs @ 4096 counts/input

puter. The computer then regards the crate as a terminal. The transfer rate is extremely slow (maximum of 19.2 kilobits or 800 24-bit words/s). The next step up is the general purpose interface bus (GPIB). This was first developed by Hewlett Packard in 1965 (and called the HP-IB), and, because of its popularity, it was accepted as an IEEE standard interface. One may now purchase GPIB interface modules for CAMAC crates, available from several vendors. Although quite slow, they are a widely used interface to IBM personal computers and their clones, as well as a number of instruments such as digital oscilloscopes, spectrum analyzers, etc. A faster way for crate/processor communication is to connect the crates directly to the bus of the computer, as shown in Fig. 3(a). There are a variety of standard busses (see BUS SYSTEMS AND COMPUTER INTERFACING), e.g., the Unibus, Qbus, and BI bus for DEC computers, and the VME bus for 32-bit workstations such as SUN computers. Several companies market CAMAC interfaces for these busses. The interface consists of a board set (one for the crate and one for the computer) and software programs called drivers that permit data-acquisition software to allow command and data transfers to occur.

2.4 Microprocessor Control

The introduction of cheap, powerful microprocessors, namely the PDP-11 series by the Digital Equipment Corporation (DEC), enabled a computer to be installed directly in the crate, as shown in Fig. 3(d). This made it possible to deal with situations requiring rapid response, if data had to be preprocessed or moved out of the digitizer's local memory in time to handle the next event. A number of startup companies bought processor chips from DEC and packaged them to fit in the crates along with a crate controller. These intelligent crate controllers had serial ports so that terminals and printers could be connected. This arrangement became an immediate success and is still used in many laboratories. However, new technologies drive the insatiable need for more (in this case, channels of data), and it became necessary to configure systems containing many crates.

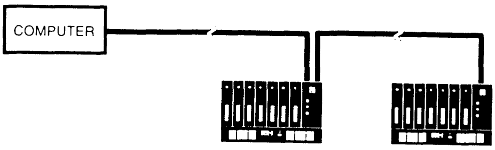

a. One or more crates on a computer bus.

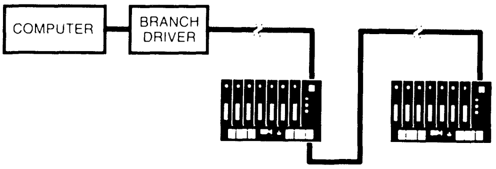

b. Standard CAMAC Parallel Branch Highway.

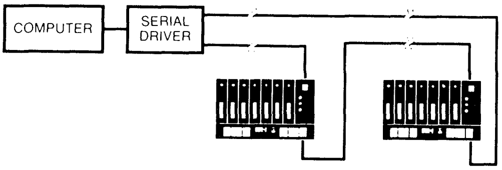

c. Standard CAMAC Serial Highway.

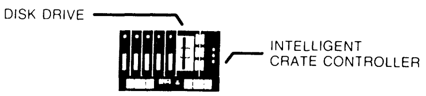

d. Stand-alone CAMAC crate with microcomputer.

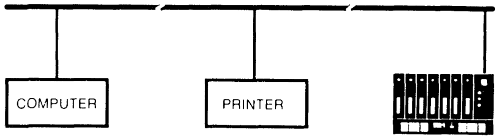

e. LAN-interfaced CAMAC crate.

FIG. 3. Several approaches to CAMAC computer interfacing.

2.4.1 CAMAC Processor Connections. A number of CAMAC crates may be connected in series (serial highways) or in parallel (parallel branch highways). A CAMAC serial branch highway is shown in Fig. 3(c). A serial driver board is installed in the host computer and then connected to the first of up to 64 crates. Each crate has a controller in it which examines the command from the host computer. If it is to one of its modules the command is carried out, if not the command is passed on to the next crate. One of the lines from the host is a clock line to set the timing for all of this. The clock rate can be as high as five MHz, and block transfer rates as high as 2.8×10^5 24-bit words/s have been measured (Cleary, 1985).

It is always faster to interact with many devices in a parallel mode and so parallel branch highways for CAMAC came into being. To handle this the crate interconnections are done by use of a 66-pair ribbon cable as shown in Fig. 3(b), and a special driver board in the host computer. Only seven crates may be connected to the interface (as opposed to 66 in the serial mode), and there is a 50-m length limitation of the cabling, but the data-transfer rate is an order of magnitude higher in this mode.

2.4.2 CAMAC LAN Connection. It is possible to hook a CAMAC crate directly to a local-area network or LAN, with the appropriate crate controller [Fig. 3(e)]. This has the advantage of having the crate accessible by more than one computer on the network. The LAN can be fiber-optics based which makes it immune from electromagnetic noise. This is a very important consideration in many environments. A LAN based on Ethernet hardware can run at up to 10 megabits/s and crate controllers for this are available now. The next technology for LANs is FDDI (fiber distributed data interface) and will be about ten times faster. It is expected that crate controllers for this will be available by mid-1992.

2.5 FASTBUS

The high data rates from some modern experiments have strained serial and parallel CAMAC highways. A response from the industry has been the introduction of FASTBUS which is an IEEE standard (standard No. 960-1986) which allows for the fastest, highest-density data acquisition and transfer today. In addition, researchers in the past few years have taken advantage of the introduction of powerful and inexpensive 32-bit workstations to the computing scene. Since the CAMAC dataway is 24 bits wide, it is incompatible with these processors; 32-bit transfers would take more than one CAMAC cycle. The industry responded with a new type of crate for FASTBUS. Like CAMAC it is a modular data bus but it is 32 bits wide; the crate has additional lines for timing and control status. The FASTBUS crates can be interconnected in a variety of ways which will be outlined in a section to follow. The FASTBUS backplane achieves a bandwidth of 40 MB/s as compared to CAMAC which saturates at about 5 MB/s. For example, a single crate might have analog-to-digital converters with 5 MB of memory to be transferred after the analog-to-digital converters (ADCs) are filled and generate a LAM. At a maximum transfer rate of 1×10^6 24-bit words/s (Cleary, 1985) it would take at least 5 s to move the data to a computer. The actual number depends upon the clock rate of the computer and the software employed, and this is a best-case scenario. One must also consider that a large experiment may have more than one crate delivering data to the host computer. FASTBUS can transfer 32-bit words at about 8 times the above speed. This difference can be critical in experiments with a high repetition rate. There is, however, a cost tradeoff; FASTBUS crates are more expensive than those for CAMAC, and there are far fewer modules available for them at this time. Research groups wishing to use them often are faced with designing their own FASTBUS ADCs.

2.6 Speed

The revolution in 32-bit microcomputers has led to the standardization of several computer bus architectures. Some computer manufacturers have opened up the details of their architectures so that others could build peripheral devices to connect to their machines, thus enhancing the popularity of their product. Two such busses are Multibus and VME. The latter has been made popular by the success of SUN Microsystems, which makes the largest selling workstations in the United States. Crates containing both these busses in their backplanes have been manufactured. The VME crate supports 32-bit transfers just as FASTBUS does. The difference here is that now the crate is an integral part of the host computer. No intermediary crate controller is required. Since VME has been widely adopted there are a large number of cheap memory and processor cards and other modules. It is possible, therefore, to buy your own processor and build a powerful computer directly into the crate. The raw speed of the VME bus is quite high, up to 100 MB/s. This system can stand alone for real-time data processing. A VME interface to a disk can allow for data storage. Unfortunately, there are no fast ADCs which are commercially available for VME busses. Those who require them have had to build their own, which is quite a task. Just as in the case of CAMAC the VME crates can have local processing power, and, in addition, a direct memory access (DMA) interface to the memory of a larger computer which can do further data processing and graphics. We note here that data can be transferred between the memories of two autonomous devices in one of two ways. The first, and much slower, method involves one device placing a single number in a memory register which is mapped into a single memory register of the second device. A flag is raised by device 1 advising the second that the number is there; device 2 copies the number into its memory and sends acknowledgment to the first. The second and much faster way is a DMA data exchange. Device 1 requests the use of the device-2 bus and some of its memory. When the request is granted (usually after device 2 is finished with its latest chore) device 1 fills the allocated memory of device 2 with a stream of data. In this method the handshaking that goes with the transfer of every number is eliminated.

Multiple crates can be connected using a network such as Ethernet (which runs at 10 Mb/s). One such system employing multiple VME crates with 96 8-bit, 100-MHz flash ADCs per crate has been built for a nuclear experiment at CERN, in Geneva (D'Isep *et al.*, 1989). The network can be fiber-optics based and, therefore, immune to electrical noise. In addition, networks even faster than Ethernet, such as FDDI, have developed in parallel with these systems. Interfaces to VME for FDDI already exist. Experimental networks with Gb/s rates are under development, and although there are as yet no standards for them, one company (ULTRANET) has a commercially available product. The trend in the high end of data-acquisition systems is toward systems with networked distributed processing.

Just as important as the leading-edge technology is the availability of very inexpensive and easy-to-use data-acquisition systems for personal computers (PCs). Some inexpensive modules can be installed directly on the backplane of MS-DOS machines and Apple computers. The latest generation of NeXT computers, as well as a new product from Silicon Graphics, have both voice and data digitizer hardware as well as the software needed to use them easily. Menu-driven software packages for DOS computers, that allow the user immediately to acquire data and plot it, are available from a number of third-party vendors. There also exist interfaces and CAMAC crate controllers for PCs, along with the appropriate software. These systems cannot support high-speed data transfers, and are limited by the speed of the PC, but they are perfect for small experiments or student teaching laboratories. In the following sections we present, in greater detail, the elements of some of these systems and give specific examples.

3. TRIGGERS

A simple trigger is an electrical pulse which is issued to stop or start data acquisition at a critical moment during an experiment. Large-scale experiments may involve many levels of triggering. This is the case in high-energy physics or fusion physics where thousands of detectors may be involved in a single experiment. These detectors may generate a signal pattern which, if of interest, may be stored or used to trigger other groups of detectors. In high-energy physics a great number of uninteresting events occur, and these data must be rejected. Multilevel triggers with successively complex discrimination logic are employed in this task. This is interwoven with the need for accurate timing and triggering at each stage. The results from one set of detectors could be compared to a data mask; if a valid event has occurred a secondary trigger is generated and delivered to other parts of the data-acquisition system which are working in parallel. Each trigger level provides a filter for the next level of acquisition.

3.1 Specialized Instrumentation

Generally, these decisions have to be made on microsecond time scales. A great deal of specialized equipment has been developed for this (LeCroy, 1990). Fast-pulse instrumentation often uses discriminators which take input pulses with varying wave forms and produce output pulses of standard amplitude and duration. Discriminators must deal with the varying duration and rise time of the input pulses as well as bursts of pulses which can exceed their own input bandwidth. Logic units are also employed in triggering decisions. These generate an output pulse whenever a number of input pulses occur within a preselected time interval. Logic circuits may produce an output which is either reinforced or inhibited by associating the input pulses to them with preprogrammed logic diagrams. Time intervals between events are measured by time-to-digital converters, which generate binary output. The temporal resolution is a function of the number of digitization bits available and the total time interval involved. Accuracy of tens of picoseconds, which is necessary in time-of-flight measurement, is possible. Internal buffering in time-to-digital converters allows for counting to be resumed before previous data are read out.

4. DIGITIZERS

A great variety of modules are available for CAMAC-based data-acquisition systems. A sample of these is presented in Table 1. The elements listed in the table can be roughly grouped into two categories. The first comprises modules used for timing and control of the experiment. The second, and the key component in data acquisition, includes the digitizers or analog-to-digital converters. The signal to be recorded from the experiment and later processed is assumed to be a current or voltage wave form. Since all digitizers require an input voltage (usually in the range of 0.0–0.5 V) the signal must be properly conditioned before applying it to the digitizer. This conditioning process can be as simple as placing the signal across a resistive voltage divider, if it is too large, or may involve the use of differential amplifiers with variable gain if there is an impedance mismatch between the source and digitizer. Signal conditioning is an enormous topic and is best left to a reference (Horowitz and Hill, 1989).

4.1 Resolution

There are three main features to any digitizer which warrant discussion here. They are the number of bits of resolution, the resident memory, and the digitization rate. In general the more of each feature that a unit possesses, the better it will perform, and the more expensive it will be. Digitizers work by mapping the value of an analog signal (i.e., a voltage) measured at a particular time into an n-bit binary number between 0 and 2^n-1. The number of times per second that a signal is measured and translated into binary is called the digitization rate. Since the measured signal is assigned a discrete value there will always be a quantization error associated with the process. In general, this error is one part in 2^n for an n-bit converter. For example, consider an 8-bit converter which accepts a signal of $0 \leqslant V \leqslant 0.5$ V. The number is stored in the converter as an integer between 0 and 255. One bit, which in this case is roughly 2 mV, is about 0.4% of the maximum signal value. If the signal is close to the maximum that the digitizer can handle, the stored signal will closely resemble the input signal as shown in Fig. 4(a). However, if the signal (or a portion of it) is small, it will appear as a staircase of discrete steps and a considerable error is introduced [Fig. 4(b)]. If the signal exceeds the input range of the converter, it will be assigned the maximum value and information will be lost as shown in Fig. 4(c). There are two ways to avoid large digitization error. The first is to ensure that the input signal is close to the full operating range of the digitizer. This may be accomplished with the use of a preamplifier or, in the case of a signal which may change slowly in time, a computer-controlled programmable preamplifier. The second way to attack the problem is to use analog-to-digital converters of 12 and 16 bits. A 16-bit converter has only a 0.000 15% error. For low digitization rate devices (< 50 kHz) these converters are inexpensive, but their price escalates with frequency. At the time of this writing, there are no digitizers with greater than 8-bit accuracy with sampling rates above 200 MHz.

4.2 Sampling Rate

The second criteria for digitizers is sampling rate. Many inexpensive PC-based data-acquisition systems come with digitizers that work at speeds below 100 kHz (10^5 samples/s). This speed is more than adequate for digitizing voice signals (the human ear generally cannot hear above 20 kHz) and sampling signals of many everyday phenomena. A car

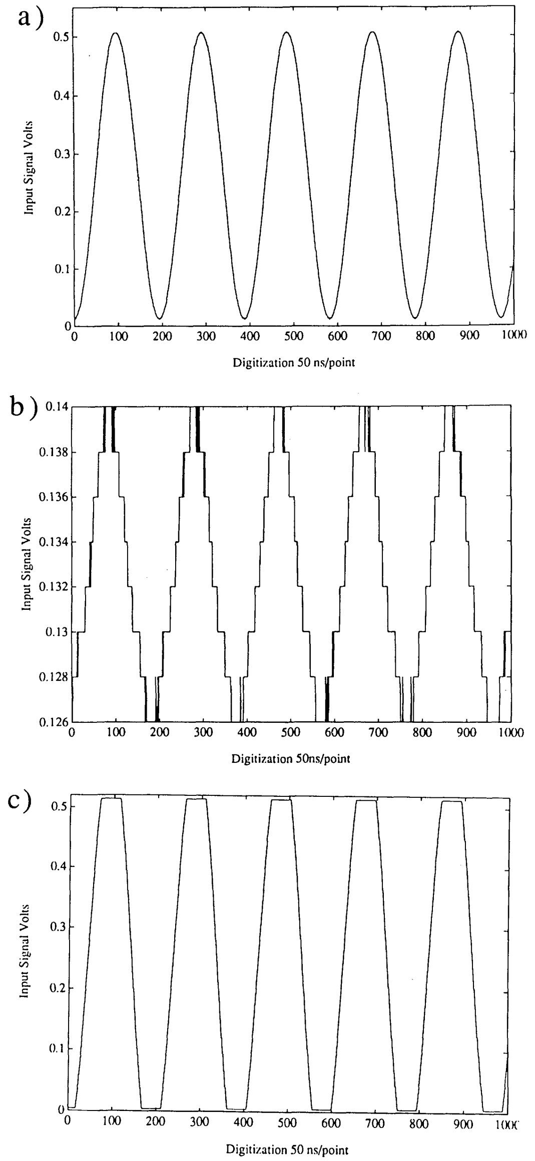

FIG. 4. (a) A digitized 100-kHz sine wave. The digitizer has 8-bit accuracy and can accept a signal of $0 \leqslant V \leqslant 0.5$ V. Since the digitizer is used over its full working range, signal reproduction is excellent. The 1000 samples correspond to a time interval of 50 μs. (b) The same digitizer is used on a signal of very small amplitude. The signal is less than 3% of the 0.5-V working range so that bit noise is apparent. (c) A signal in excess of the 0.5-V limit on the digitizer results in clipping.

moving at 60 mph will move 2.7 mm in one sample time. However, there are a great number of other phenomena such as the oscillations of electrical circuits, or the change of parameters in a plasma, that occur on much briefer time scales. The fastest digitizers today work at about 1 GHz (10^9 samples/s) and are about 100 times too slow for some of these applications. One must, therefore, consider what happens if there are elements present in a wave form which are above the maximum frequency that the digitizer can handle. Figure 5(a) shows an example of a 110-kHz signal digitized at a rate of 1 MHz (1 μs/sample). There are 11 samples on each cycle of the wave form. The theoretical limit for a particular sampling rate is called the Nyquist limit (Bendat and Piersol, 1971). At the Nyquist limit there are only two samples on every cycle, and the wave form, as shown in Fig. 5(b), is not discernable to the unaided eye. Finally, if one introduces a wave form of higher frequency than the digitizer can handle, a curious phenomenon occurs. As shown in Fig. 5(c), a lower-frequency component, which is nonexistent in the original data, appears. The occurrence of this spurious signal is explained in Fig. 6. In the upper graph the original high-frequency signal is shown with dots on it. The dots correspond to the sampling times of the slow digitizer. Suppose it is known that the input signal is a pure sine wave. If the sampled signal points (the dots) are used to reconstruct a sine wave, as shown in Fig. 6(b), a lower-frequency, incorrect wave form results. It is therefore important to exclude all signals of bandwidth greater than that of the digitizer. This can be accomplished with the use of a low-pass filter. Failure to filter the input signal properly can result in a corrupted stored signal, as seen in Fig. 7. The first part of the figure shows the power spectrum of the correctly digitized 10-kHz signal of the type shown in Fig. 5(a). The power spectrum is a measure of how the input signal power varies as a function of frequency, and is found by Fourier transforming the original time-series signal and multiplying by its complex conjugate. In Fig. 7(a) there is a clear peak at 10 kHz corresponding to the input signal. (There is always a peak at zero frequency which is suppressed in these plots.) When an improper digitization procedure is used, such as 1-MHz sampling of a 20-MHz signal [as in Fig. 5(c)], a signal not present in the original wave form appears, as shown in Fig. 7(b). No signal should have been detected at $f < 0.5$ MHz, but the undersampling gives a nonexistent peak at lower frequency.

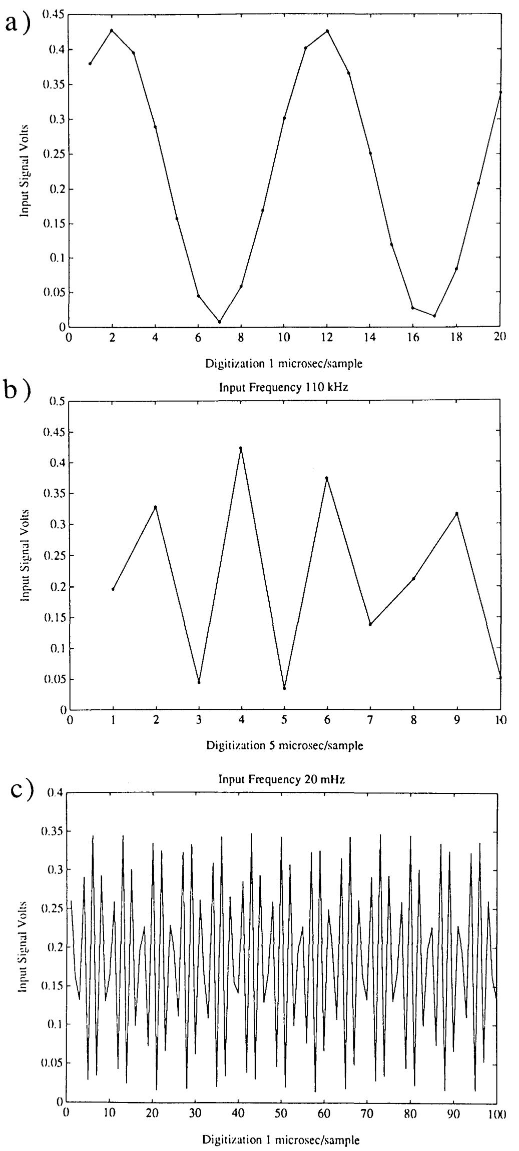

FIG. 5. (a) A 110-kHz signal digitized at 1 μs/sample. The dots on the curve are the digitized data and are connected by straight lines for clarity. The sine wave is choppy [unlike that of Fig. 4(a)] but easily recognizable. (b) The same signal digitized at 5 μs/sample, just above the Nyquist limit. This is truly a limit for the sine wave in (a) is not recognizable. The record shown is 50 μs in length. (c) The effect of undersampling. A 20-mHz signal is digitized at 1-μs intervals; a factor of 20 too slow. The time interval shown is 100 ms but a wave form of five oscillations is not visible.

4.3 Memory

Every analog-to-digital converter must be coupled to a memory. The digitizer performs a conversion of the input signal into a binary number which must then be read into a memory before the next conversion is attempted. In many cases this memory is closely coupled to the converter and is not the same as the memory of the computer which controls the process. When the digitizer's memory is full, its contents must be transferred to the host computer. The cost of an analog-to-digital converter depends upon how much resident memory it has, and how fast that memory must be addressed. Clearly, a digitizer which does one conversion per nanosecond must be matched to a memory fast enough to accept numbers at that rate.

5. HOW A DATA-ACQUISITION SYSTEM WORKS

As outlined above, there are many hardware variations within a DAS, but the process of acquiring data shares many common procedures. As an example of these common elements, we describe the features a software package should have for controlling the interaction between a digitizer (ADC) and a CAMAC-based DAS. When the DAS control software is initiated, the user decides which devices in the crates will participate in the experiment. The software generates a table containing which crate and slot the device occupies. If an experiment is to be repeated, these tables may be read from a file which defines the experiment.

5.1 Software

The software then issues an instruction to the crate controller, which in turn initiates the ADC. Before starting, the ADC must receive a LAM (look at me) enable instruction from the crate controller. In some crate controllers, an instruction to enable an interrupt of the processor when its data acquisition is complete must be issued also. Next, the crate receives a clear signal from the processor controlling it, all the device registers are cleared, and digitization begins. The ADC continually reads and converts data from its input signal. The digitized data are transferred into its local memory. When the memory is filled, the digitizer overwrites the start of memory and continues until a stop trigger is received from the control software. The trigger forces the ADC

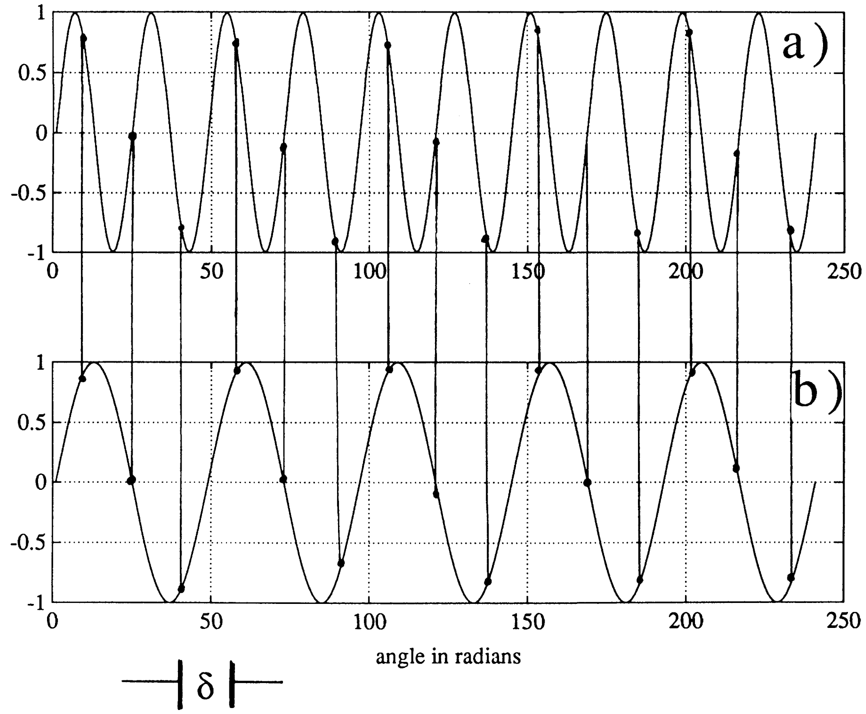

FIG. 6. Effect of undersampling a high-frequency wave form. A sine wave shown in (a) is sampled at an interval δ. Each sample time is denoted by a dot. If it is known that a sinusoidal wave form was input to the digitizer, an incorrectly reconstructed wave form, shown in (b), may result. Undersampling a high-frequency signal can therefore result in low-frequency *ghost* modes not present in the original one.

immediately to stop acquiring data and raise a LAM. When the crate processor detects a LAM, it immediately stops what it was doing and jumps to a location in its memory where it expects to find the starting address of an executable program. Such a program is called an interrupt handler.

5.1.1 Interrupt Handlers. These programs were formerly laboriously written in assembler language by data-acquisition programmers. If there was a programming error and the computer jumped to the wrong address, or there was nothing at the right address, the computer simply crashed. Now, however, companies which sell crate controllers also sell higher language subroutines (usually written in FORTRAN, PASCAL, or C) which activate handlers they have written. The handler is usually quite simple. For example, it might check to see that one of the devices authorized to participate in the experiment generated the interrupt. If an illegal device interrupted, that device could be ignored or the software could branch to an error-handling routine. If a legal device interrupted, the handler disables the LAM and branches off to another program which may be a user-written routine to read data. The interrupt handlers provide an interface between the software control package and internal crate operations.

There is another way to service the device without generating an interrupt. All the above steps are performed, but the device is not authorized to issue a processor interrupt. It does, however, issue a LAM. The software control package must enter a loop in which it looks at all the devices that could raise a LAM and waits for one.

5.2 Sample Program

A generic routine (in FORTRAN) might look like this:

```
       time = secnds(0.0)                      ! set clock to zero
       do i = 1,no_crates                      ! loop over crates
         if(status(crate(i),crate_stat).eq.0)  ! check for problem
           call error_message()                ! something wrong, print
         endif                                 ! an error message
10     if(crate_stat. eq. LAM_mask)then        ! we got a LAM
       {
                 .......do things like read data and turn off the LAM
                 }
       endif
       else                                    ! if we didn't get a LAM
         if (secnds(time).lt.timeout)then      ! continue to check
         goto 10
       else                                    ! we waited long enough
         goto 100                              ! leave the loop
       endif
       enddo                                   ! finish looping over valid modules
100    continue                                ! jump out of search for LAMs
```

Here, the crate controller has a word called crate_stat which it passes to the microprocessor. If the word is blank, there is a problem, and an error message is printed. Further action could be taken here as well. If the word returned from the crate controller equals the bit pattern (the correct sequence of zeros and ones) in the word LAM_mask, the *i*th device has raised a LAM and is then serviced. Finally, a timing routine must be included so that, if nothing happens for a while (timeout seconds), the computer can go on to do other things. Otherwise, it will be caught in an infinite loop!

5.3 Processor–Crate Communication

How do the crate controller (which is a microprocessor) and the master computer actually communicate with each other? This is done by sharing memory locations. Furthermore, the dataway lines previously discussed are directly mapped into these locations. This mapping is done by the hardware in the crate controller. Before the data-acquisition control software is run, the memory locations in the host computer permanently allocated for sharing with the crate controller must be assigned. For some small computers directly housed in the crate, this assignment can be done by setting dip switches in the back of the computer. For larger systems, a software *driver* must be loaded into memory when the computer is booted or turned on. This driver is a short program which permanently resides in memory with the computer's operating system and provides all these memory addresses. When the crate controller interrupts the computer, the driver takes over, finds the correct servicing program, and calls it. When you want to do something such as clear a LAM, the driver actually issues the instruction, which might be something like:

```
MOVB #10., @#164040.
```

This moves a byte, or 16-bit word, which contains the decimal number 10 (or 0 000 000 000 001 010), into memory location 164040, shared by both the controller and host computer. The crate controller then puts this number out onto the dataway, which has the effect of putting positive pulses on two of the signal lines, which in turn forces the hardware to turn off the LAM. Presently, the programmer who writes control software no longer needs to contend with these low-level instructions. Software that handles these processes may be purchased from the company that sells the crate controller.

Once the data is stored in the ADCs, the next step is transferring it to the host computer. In early crate controllers, this process was slow, since data could only be transferred word by word. A CAMAC command initiated the transfer and also passed the address of the data array in the host computer. Each time a word was transferred, the controller also set standard CAMAC bits (the Q and X bit) indicating that all went well. The user could check these bits with the main data-acquisition program. Later models of crate controllers allow for block transfers of data. Blocks of up to 32K (32 767) 24-bit words can be transferred all at once. This procedure saves time as there is considerable overhead in getting the host computer and crate controller to start the transfer (a memory region must be allocated and other parameters initialized). Actual transfer rates are hardware specific, but for a MicroVax (μVAX) running the VAXELN operating system (which is a Digital Equipment Corp. real-time operating system), the first word transfer takes about 700 μs and each additional word 2.3 μs (Antone *et al.*, 1989).

6. SPEED AND RESPONSE OF DATA-ACQUISITION SYSTEMS

6.1 Transfer Rate to Host Computer

The response required of a DAS is entirely dictated by the experimental situation. The relevant parameters are the repetition rate of the experiment (if indeed it is repetitive; a nuclear explosion is not!), the number of data channels and digitizers, and the amount of memory each digitizer must eventually transfer to another computer. The TFTR (Tokomak fusion test reactor) at Princeton is an experiment that lasts for several seconds and has a repetition rate of about 8 min. It can produce 24 MB of data in each shot (Randerson *et al.*, 1986), or about 2.4 GB per day. The experiment utilizes 280 CAMAC crates which contain over 5000 modules (Rauch, 1986). Because there is a fairly long time between shots in the TFTR experiment, the main problem is getting all the data out of the digitizers into storage. Real-time processing is not a great issue. The crates are connected to computers through eight separate serial highways. The CAMAC services a number of Gould SEL computers which gather the data. Data are then transferred to several VAX computers where it is written to disk. Some of it is immediately plotted to give the scientists an idea of what happened on the shot. After a day or two, the data are transferred to magnetic tape where it is archived. One obstacle the TFTR DAS overcame was that no fast digitizer has enough memory to store several seconds worth of data. The problem was solved by designing the system to generate preprogrammed timing pulses to start and stop digitizing. Fortunately, there are modules for CAMAC that can do this.

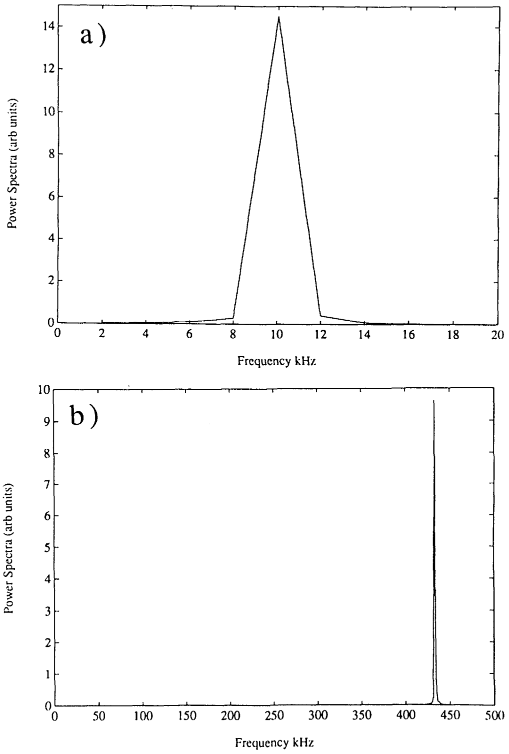

FIG. 7. (a) The power spectrum of a properly digitized 10-kHz signal. There is a peak at the appropriate frequency. The zero-frequency peak has been suppressed. (b) The power spectrum of an improperly digitized wave form. Here a 20-MHz signal is digitized at a 1-MHz sampling rate. Because of the Nyquist limit on the 1-MHz sampling the spectrum extends from 0 to 0.5 MHz. A peak at 440 kHz not present in the original signal is evident.

6.2 Real-Time Analysis

There are other cases where collection of all the data from the modules and storage for subsequent processing is not possible. This occurs in the LEP electron-positron collider at CERN in Geneva (Rüden, 1989). The detector there has several hundred thousand channels and produces 500 MB of data per second. It must be processed in real time and compressed to about 100 kB per second. Each channel, however, may or may not generate

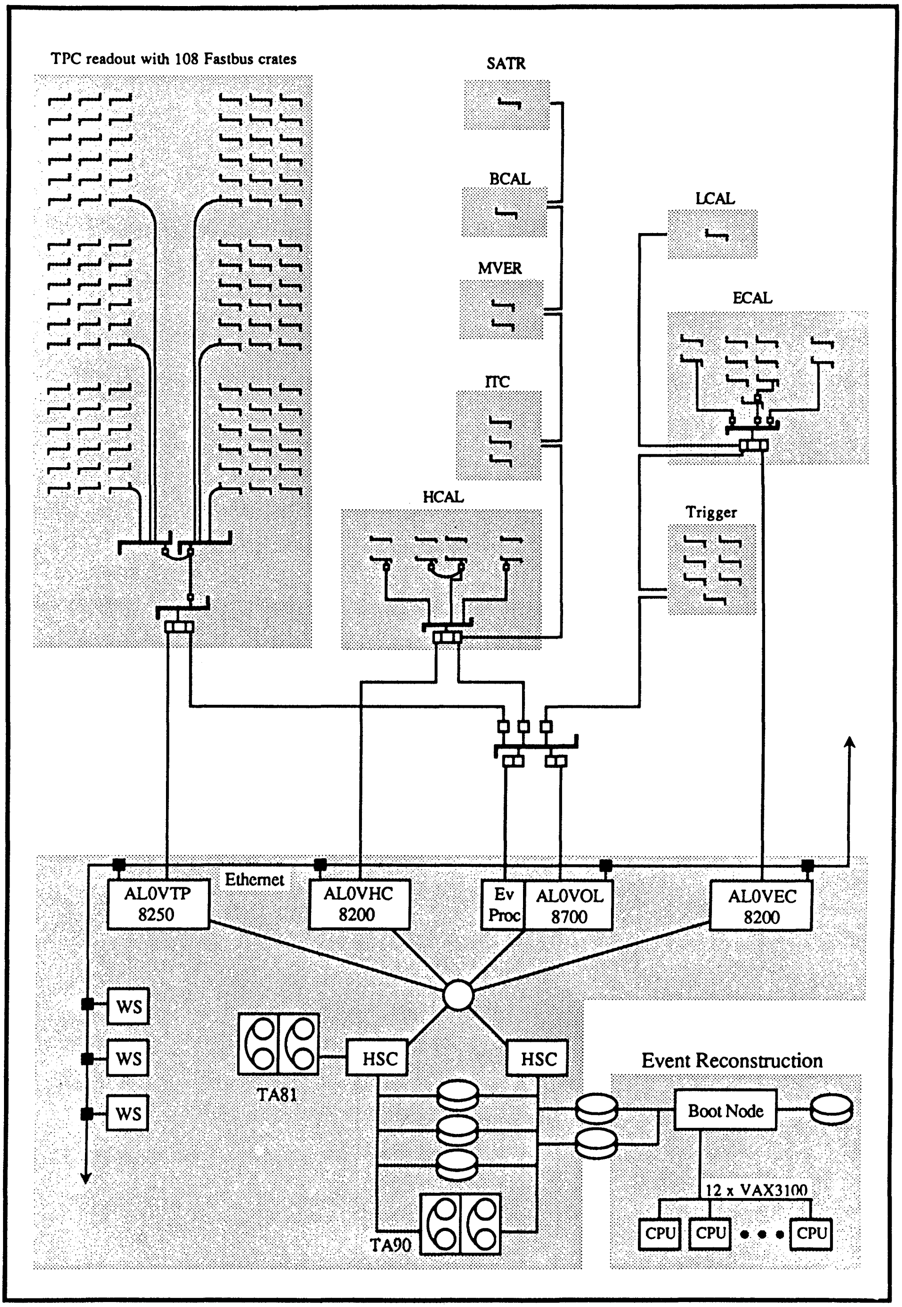
TPC readout with 108 Fastbus crates
SATR
BCAL
MVER
ITC
HCAL
LCAL
ECAL
Trigger
Ethernet
ALOVTP 8250
ALOVHC 8200
Ev Proc
ALOVOL 8700
ALOVEC 8200
WS
WS
WS
TA81
HSC
HSC
Event Reconstruction
Boot Node
12 x VAX3100
CPU
CPU
CPU
TA90

FIG. 8. A large data-acquisition system used in a high-energy physics experiment at CERN. The symbols denote FASTBUS crates each with its own processor. Various abbreviations are as follows: TPC, time projection chamber; SATR, BCAL, HCAL, LCAL, various calorimeters; ITC, inner track chamber; MVER, minivertex detector; ALOV**, VAX computers; HSC, high-speed channel, WS, workstation; TA90, high-speed magnetic tape drive. The system works in parallel to discard an enormous amount of uninteresting data from the 500 MB/s which are continuously acquired.

data depending upon whether a particle passed its detector. In fact, a great many of the channels are empty at any given time. A schematic of the setup is shown in Fig. 8. In this schematic, each shaded area on the top of the page refers to data acquisition associated with a different detector (such as TPC, time projection chamber detector) or monitoring of the machine operation itself. The lower shaded area encloses the group of computers used to acquire, store, and analyze the data. The larger VAX computers are named AL0V** and are networked to each other, the tape drives, and the workstations. Data reduction is accomplished in several stages by use of distributed processing. In the first stage of processing, analog hardware compares signal levels against a threshold and rejects signals that are too small. Meaningful data, properly scaled by the 32-bit FASTBUS crate computers, is next transferred from the crates to a series of DEC KA800 computers through a network. These computers, which are not shown in this diagram, are placed directly on the bus of a larger DEC computer (a VAX 8700 denoted AL0VOL) and rapidly process the event and make a final decision as to whether it is worth saving. If the answer is yes, it is transferred to the memory of the VAX 8700. Final processing of the data is done on a cluster of MicroVAX workstations (Jost, 1989).

6.3 Choice of Operating System

Another issue regarding performance is the choice of computer operating systems. These can be broadly classified into two types, real-time and multitasking. A real-time operating system is stripped down for action. It is geared to serving one process only, the data-acquisition program. These operating systems were first available on single-user microcomputers and occupied very little memory. Real-time operating systems are designed to handle device interrupts quickly. When an interrupt occurs, they can stop the current process and initiate the interrupt handler within several microseconds. These small microprocessors are still extremely popular in DAS's. A multiuser operating system, such as UNIX or VMS, has a large built-in overhead since these systems are continually switching between many processes. A response to an interrupt can take several milliseconds to service, so they cannot work in strictly real time. It is possible to optimize performance by assigning a high priority to the DAS program and carefully designing the software (Appalaraju and Cleary, 1989; Atac *et al.*, 1989). It is also possible to buy an operating system which is tailored for real-time applications, such as VAXELN.

7. AN EXAMPLE OF A DATA-ACQUISITION SYSTEM

Let us consider a specific example that has some of the requirements cited above, i.e., the need for speed and distributed processing, but is not overly complex. The experiment involves the study of magnetic fluctuations in a plasma (a partially ionized gas) embedded in a magnetic field (Gekelman and Stenzel, 1984). When an electrical current is passed through an initially quiescent plasma, large fluctuations in the magnetic field are observed. The strength of these fluctuations is a function of space and time. The experimental goal is to find the source of these fluctuations. They are thought to be due to the random generation of waves in the plasma, but which waves are responsible? Since the waves are chaotic, it is necessary to perform correlation measurements at a number of frequencies to determine the cross spectral function (CSF) of the magnetic-field fluctuations (Bendat and Piersol, 1971). The data-acquisition procedure involves digitizing and storing three signals (B_x, B_y, B_z) that measure the three components of the fluctuating vector magnetic field. The signals from each of two probes are measured at 1024 time steps for an ensemble of 80 shots at 300 spatial positions on several planes. There is insufficient disk space to store all this data, so the CSF is evaluated at each position after the ensemble of data is acquired. The DAS used is shown in Fig. 9. The ADC's reside in CAMAC crates. The crates have a controller as well as an LSI 11/23 computer in them (Gekelman and Xu, 1986). The LSI runs a real-time operating system (RT-11) which rapidly responds to LAMs in the crates and services them. All of the software controlling the devices in the crates and reading data from them resides in the LSI. The LSI in turn is networked to a VAX computer with Ethernet hardware. The VAX has an array processor, which is a very fast calculating engine

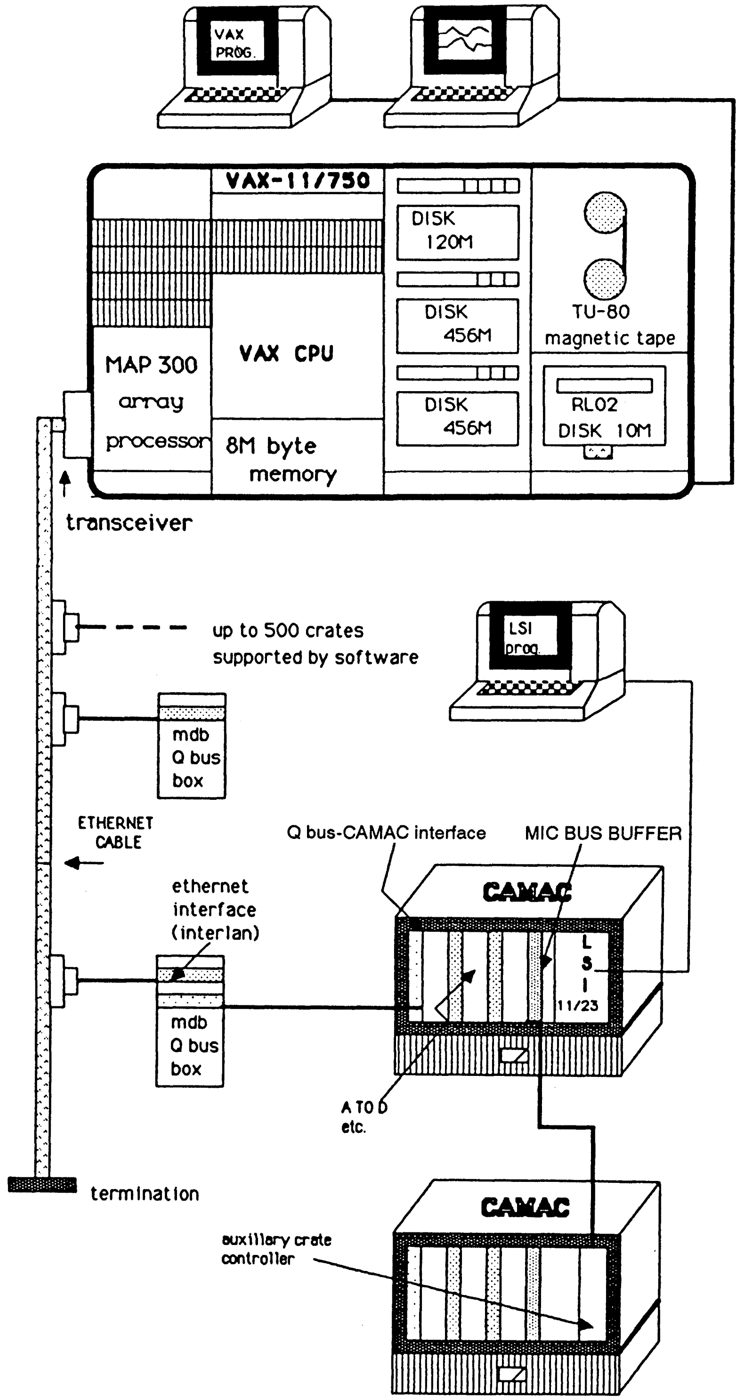
VAX PROG.
VAX-11/750
DISK 120M
DISK 456M
DISK 456M
TU-80 magnetic tape
RL02 DISK 10M
MAP 300 array processor
VAX CPU
8M byte memory
transceiver
up to 500 crates supported by software
LSI prog.
mdb Q bus box
ETHERNET CABLE
Q bus-CAMAC interface
MIC BUS BUFFER
ethernet interface (interlan)
CAMAC
L S I 11/23
mdb Q bus box
A TO D etc.
termination
CAMAC
auxillary crate controller

FIG. 9. A small data-acquisition system in a university physics laboratory. A series of CAMAC crates are interfaced via a microcomputer and network to a minicomputer. The array processor in the minicomputer allows for real-time data analysis such as digital filtering. This combines elements found in Figs. 3(a), 3(d), and 3(e).

capable of directly accessing its memory through a DMA interface. The main data-acquisition program runs on the VAX. The system is configured to make the LSI microcomputer appear as a disk to the VAX. Whenever a command is sent to the LSI, a write statement to a certain device number is issued which activates a device driver setting up a dialog through the network to the LSI. Whenever the LSI sends data to the VAX, the reverse happens. The pseudodisks the two computers read from and write to are divided into logical areas. One area is reserved for instructions and, when data appear, those data are interpreted as a command. Another much larger area is reserved for the actual data. The system can throughput about 250 kB of data per second between the CAMAC crate and the VAX. As soon as data appear in the VAX, a DMA transfer is initiated to the array processor. After 80 shots, the array processor takes the data and calculates four components of the cross spectral tensor. This calculation takes a few seconds, and then the results are sent back to the VAX for storage. In this mode of operation, only 4K real (32-bit) numbers need to be stored for each probe position instead of 480K.

Technology is changing so rapidly that, as far as computing is concerned, systems can become obsolete a year after purchase. In the four years since the above system was designed, extremely powerful workstations have become available, some of them with vector processors. They make attached array processors obsolete. Fast crate controllers networked to these workstations allow for a direct interface to the host computer, making the separate LSI crate computers unnecessary. However, in some other applications, such as the system at CERN, distributed computing is still necessary. This progress is welcome since programming is simpler because it can generally be done entirely in a higher-level language.

8. DATA MANAGEMENT AND VISUALIZATION

The growth of powerful data-acquisition systems has enabled vast amounts of data to be collected in relatively short periods of time. Experimentalists are then faced with trying to analyze enormous data sets. In a situation with many investigators, there may be many separate channels of data, with each scientist chiefly interested in his or her own data. The requirement here is to organize and archive the data to make it simple to retrieve for subsequent analysis. This may involve data compression and migration of data to secondary computers or tapes as it ages. One way to deal with large data sets is to use commercially available database-management software. This has been done on a fusion mirror machine experiment (Stewart *et al.*, 1986), the TEXT Tokomak experiment (Wiley *et al.*, 1989), the Doublet experiment at General Atomics (McHarg, 1989), and others. The key to understanding complex multidimensional data sets is the ability to visualize them. Until recently, this was a difficult process. Data were generally displayed as a series of contour maps or surface plots for separate planes in whatever space (position, momentum, etc.) was of interest. Trying to understand volumetric data involved comparing many of these plots; information was inevitably lost or obscured in the procedure. Recently, a number of vendors have come out with extremely powerful three-dimensional graphics workstations. It is possible to transfer data to one of these and make a full color rendering of volume data as a set of nested isosurfaces of differing transparency, or as voxels, which are color pixel representations. By using appropriate display filters, such as changing the way colors represent key features in the data, many subtle features can be brought out. In addition, the data may be magnified, translated, or rotated in real time. Interfaces to video equipment and computer-controlled slide- and print-making equipment are also available and becoming reasonably priced. Finally, there have been some startling advances in software such that expertise in graphics programming is not required. There are several new commercially available software packages which are menu driven and do all of the above as well as instantly creating vector fields, streamlines, and cut slices through data sets with any number of dimensions. Presently, they run only on powerful graphics computers which are rather costly. However, it is anticipated that the price of these workstations will plummet, and someday they will be as affordable as personal computers. This will herald a new age of discovery of subtle and complex phenomena which now are buried in huge data sets.

9. FUTURE TRENDS

9.1 Communication

From the tone of the preceding text, it is not hard to predict the evolution of DAS's for the next five years. The emphasis on the high end will be speed. This will entail the use of multiprocessors, as mentioned previously, but the old way of connecting them via a bus may change radically. The basic limitation in signal transfer is the speed of light which will not allow an arbitrary number of information packets, each addressed to a different processor, to share one data path effectively. One possible solution is the CSI, coherent scalable interface (Gustavson, 1989), which is basically a point-to-point datapath between processors. It will be designed for GB/s processor-to-processor communication. The CSI is now in a standardization process phase before the IEEE.

9.2 Computer Farms

There is a great deal of experimentation with high-speed DAS's at the large accelerator laboratories. A team at the Fermi National Accelerator Laboratory (Atac *et al.*, 1989) has designed and built a system which links devices from both FASTBUS and VME crates to a farm of computers using a very fast crossbar switch manufactured by Texas Instruments. The switch can make eight simultaneous connections out of a possible 28, each connection having a 20-MB/s bandwidth for a total speed of 160 MB/s. Each connection can be between a crate and a group of parallel processors which receive the data and preprocess it before sending it to storage. As soon as a group of processors have data to work with, the switch changes position and connects more data channels to other free processors. Another approach was used in a nuclear physics experiment in Heidelberg (Männer and Stucky, 1990) in which a multiple 32-bit bus system is used instead of a switch to connect FASTBUS crates to an array of 30 processors. In this experiment, data flows in at a rate of 10^6 events per second, with up to 10^3 parameters per event. The processor farm filters this down to 10^5 numbers per second which actually must be stored, a data reduction of 4 orders of magnitude. Efforts such as the two described here are large because they involve the construction of the hardware and writing of the software to support it. They are, however, the precursors of what may be commercially available systems utilizing very rapid communication technologies between data channels and multiple processors.

9.3 Parallel Processors

One of the recent advances in high-end computation is the introduction of highly parallel processors (see PARALLEL COMPUTING). These computers contain from 64 to 16 384 individual processors. There are various schemes for interprocessor communication and use of memory, but the basic idea is that, if a problem can be broken up into many parallel segments which run independently, significant increase in speed may be attained using processors running in parallel. Although several companies manufacture these systems, they are still largely in the experimental stage since compilers which can generate efficient code for parallel CPUs have yet to be developed. In the future, one could consider the development of a DAS in which one CPU is dedicated to each input data channel. Interprocessor communication would make it possible to compare data from neighboring probes instantly to determine what should be stored, or to do mathematical operations such as the instantaneous evaluation of the curl of vector fields. This will be some time in coming, however, since these multiprocessor machines do not easily interface to any of the standard data-acquisition gear such as FASTBUS or CAMAC. Thus far, the only parallelizing is in the type of system at CERN described earlier, where many FASTBUS crates, each with its own processor, are brought to bear on the problem.

9.4 Inexpensive, Simple Systems

On the other end of the DAS spectrum, we see a trend in development of easy-to-use, off-the-shelf DAS's for small laboratories. A number of digital instruments, such as oscilloscopes and wave-form generators, have GPIB outputs and can be controlled by software running on a PC. Setting them up to acquire, store, and plot data on digital X-Y recorders can take several hours. The most expensive piece of equipment is usually the measurement instrument, not the DAS. As time goes on, these systems will continue to drop in price and become as easy to use as digital voltmeters.

10. SUMMARY

Fifteen years ago, data acquisition was in its infancy. Data were taken in an analog fashion using Polaroid photographs of oscilloscope traces or read off of plots on *X-Y* recorders. No detailed experiment could be performed on a complex multiparameter system. The introduction of data-acquisition standards and microcomputers has radically changed this. Experiments now routinely collect and process billions of numbers from a large variety of detectors. As the price of computers and high-speed instrumentation tumbles, and the ease of implementing them improves, this technology will become available to any research group that needs it.

ACKNOWLEDGMENTS

Writing this made me recall the many hours I spent writing data-acquisition software with Tim Mc Fadden and Lingyu Xu. I have enjoyed working with them and am indebted for all they have taught me. I would like to thank Dr. R.T. Cleary for contributing Fig. 3 and Dr. Wolfgang von Rüden for Fig. 8. I am also indebted to Dr. James Maggs and Marilyn Pagnusat for proofreading.

GLOSSARY

Byte: 8 bits, where a bit is a zero or a one.

CAMAC: Computer automated measurement and control; standardized (IEEE 583) modular data handling system, 1-MHz 24-bit wide data path.

CERN: Center for European Nuclear Research.

CSI: Coherent scalable interface.

DAS: Data-acquisition system.

DMA: Direct memory access; a device which can take over a computer bus and access its memory.

Ethernet: A local-area network with hardware which can sense and correct for collisions between data from different computers; limited to 10 megabits per second.

FASTBUS: Fastest high-density data-acquisition standard available; 32-bit, 40 MB/s (IEEE standard 960-1986).

FDDI: Fiber distributed data interface; fiber-optics–based high-speed network with a token ring topology.

Gigabyte: One billion bytes or 8 billion bits.

GPIB: General purpose instrument bus; most popular instrument bus today; serial transfers limit it to 400 kB/s.

LAM: Look at me; signal generated by device wishing to be serviced by a processor.

LAN: Local-area network; a data communications system for moderately high-speed traffic, usually limited to a building or group of buildings.

LSI: Large-scale integrated circuit.

Megabyte: One million bytes or 8 million bits.

MIPS: Millions of instructions per second.

VOXELS: Representation of volumetric data as colored spots.

List of Works Cited

Antone, I. D., Mandrioli, G., Matteuzzi, P., Sanzani, G., Bloise, C., Grillo, A. F., Marini, A., Ronga, F., Baldini, A., Mancarella, G., Palamara, O., Surdo, A., Lamanna, A., Petrera, S. (1989), *IEEE Trans. Nucl. Sci.* **36**, 1602–1607.

Appalaraju, R., Cleary, R. T. (1989), *IEEE Trans. Nucl. Sci.* **36**, 606–610.

Atac, R., Cook, A., Deppe, J., Fischler, M., Gaines, I., Husby, D., Nash, T., Pham, T., Zmuda, T., Eichten, E., Hockney, G., Mackenzie, P., Thacker, H. B., Toussaint, D. (1989), *IEEE Trans. Nucl. Sci.* **36**, 726–730.

Bendat, J. S., Piersol, A. G. (1971), *Random Data: Analysis and Measurement Procedures*, New York: Interscience.

Cleary, R. T. (1985), *Enhanced Serial Highway System*, Kinetic Systems Corporation Technical Note TN-102, Lockport, IL: Kinetic Systems.

D'Isep, D., Maggiora, A., Minetti, B., Maron, G., Morando, M., Balestra, F., Masoni, A. (1989), *IEEE Trans. Nucl. Sci.* **36**, 713–717.

Gekelman, W., Urrutia, J. M., Stenzel, R. L. (1987), *Magnetotail Physics*, Baltimore, MD: Johns Hopkins University, pp. 261–274.

Gekelman, W., Xu, L. (1986), *Rev. Sci. Instrum.* **57**, 1851–1855.

Gekelman, W., Stenzel, R. L. (1984), *J. Geophys. Res.* **89**, 2715–2733.

Gustavson, D. B. (1989), *IEEE Trans. Nucl. Sci.* **36**, 811–812.

Horelick, D., Larsen, R. S. (1976), *IEEE Spectrum* **13**, 50–55.

Horowitz, P., Hill, W. (1989), *The Art of Electronics*, Cambridge, UK: Cambridge University Press.

Jost, B. (1989), *IEEE Trans. Nucl. Sci.* **36**, 1580–1584 (1989).

LeCroy (1990), *Instrumentation Research Catalog*, Chestnut Ridge, NY, LeCroy Corporation.

Männer, R., Stucky, O. (1990), *Comput. Phys.* **4**, 267–274.

McHarg, B. B. (1989), Proceedings of 13th Symposium on Fusion Engineering, Knoxville, Tennessee (unpublished).

IEEE (1982), *Modular Instrumentation and Digital Interface System (CAMAC)*, ANSI/IEEE Std. 583, New York: IEEE.

Randerson, L., Chu, J., Ludescher, C., Malsbury, J., Stark, W. (1986), *Rev. Sci. Instrum.* **57**, 1892–1894.

Rauch, W. A. (1986), *Rev. Sci. Instrum.* **57**, 1898–1900.

Rüden, W. (1989), *IEEE Trans. Nucl. Sci.* **36**, 1444–1448.

Stewart, M. E., Carter, M. R., Casper, T. A., Meyer, W. H., Perkins, D. E., Whitney, D. M. (1986), *Rev. Sci. Instrum.* **57**, 1929–1931.

U. S. NIM Committee (1981), *Definitions of CAMAC Terms*, U. S. Department of Energy, Office of Energy Research, Washington, DC 20545.

Wiley, J. C., Miner, Jr., W. H., Bravenec, R. V., Patterson, D. M., McCool, S. C. (1989), *Rev. Sci. Instrum.* **60**, 1286–1296.

Further Reading

Carr, J. J. (1978), *Digital Interfacing with an Analog World*, Blue Ridge Summit, PA: Tab Books.

Malcolme-Lawes, D. J. (1988), *Microcomputers and Laboratory Instrumentation*, 2nd ed. New York: Plenum.

DATA ANALYSIS

See STATISTICAL DATA AND ERROR ANALYSIS

DEFORMATION OF MATERIALS

See MATERIALS DEFORMATION

DENRITES

See WHISKERS AND DENDRITES

21
Calorimeters in Particle Physics

Christian W. Fabjan

21.1 **Introduction** **651**
21.1.1 The Task of Particle Detectors 651
21.1.2 Energy Measurement with Calorimeters 652
21.2 **Calorimetry Methods in Nuclear Physics** **653**
21.2.1 Total Absorption and Range of Particles 653
21.2.2 Principal Detection Techniques: Scintillators and Solid-State Detectors 655
21.3 **Calorimetry Methods in Particle Physics: Photons, Leptons, and Hadrons** **656**
21.3.1 Absorption Mechanisms 656
21.3.2 Physics of the Electromagnetic Cascade 657
21.3.3 Physics of the Hadronic Cascade 659
21.4 **Performance of Calorimeters** **662**
21.4.1 Energy Resolution in Homogeneous Electromagnetic Calorimeters 663
21.4.2 Instrumental Contributions to the Electromagnetic Energy Resolution 664
21.4.3 Time, Space, and Angular Resolution in EMCs 665
21.4.4 Energy Resolution of Hadron Calorimeters 665
21.5 **Calorimeter Systems** **667**
21.5.1 Calorimeter Construction Principles 667
21.5.2 Instrumentation Techniques 667
21.5.3 Calorimeter Facilities 670
21.6 **Calorimeters in Astroparticle Physics** **672**
Glossary 675
References 675
Further Reading 676

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

21.1
Introduction

21.1.1
The Task of Particle Detectors

In many areas of experimental physics (solid-state, nuclear, and particle physics), kinematical properties of particles or quanta of radiation need to be determined. The momentum **p** and the energy E, or equivalently the particle's velocity **v** and its mass m [$\mathbf{p} = m\beta\gamma$, $E = m\gamma c^2$, $\boldsymbol{\beta} = \mathbf{v}/c$, and $\gamma = (1 - \beta^2)^{-1/2}$], or momentum **p** and velocity **v** (Fernow, 1986) have to be measured.

A host of different particle detectors has been developed, frequently with very specialized functions. The momentum of charged particles is determined in magnetic spectrometers: particles with momentum p (GeV c^{-1}) in a field B (Tesla) describe a circular trajectory with radius r (meter), such that $p = 0.2998Br$. Position detectors trace the trajectories. The methods used for the velocity measurement depend on the β-range to be covered. The flight time of particles traversing a known distance (time-of-flight (TOF) method) is clocked with sensors sensitive to the arrival time; accuracies of better than 100 picoseconds (10^{-10} seconds) are achieved quite routinely. This method is applicable for particle velocities up to $\beta \leq 0.98$. For even more relativistic particles, the velocity dependence of the electromagnetic interaction of the charged particle in the medium is exploited. If the particle velocity v is larger than the velocity of light c/n in the medium, n being the index of refraction, the medium radiates photons at a characteristic angle $\theta =$ arccos $(1/\beta n)$, called *Cherenkov radiation*. Measuring the angle of the emitted photons provides useful velocity determination up to $\beta \leq 0.9999$ and hence particle identification up to ≈100 GeV c^{-1}. Momentum and velocity measurements have found numerous and diverse applications despite two major shortcomings: (i) the method is limited only to charged particles and (ii) the relative accuracy scales as $\Delta p/p \sim p$; huge engineering efforts are required to employ very large (superconducting) magnets and position detectors comprising typically 10^6 or more position sensors, to achieve adequate momentum resolution at the 100 GeV c^{-1} scale.

For the measurement of neutral and very energetic particles, a further technique is required – calorimetry.

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

21.1.2
Energy Measurement with Calorimeters

The measurement of the particle's energy E relies on "total absorption" in the detector occurring within a few tens of nanoseconds through a cascade of interactions producing increasingly lower-energy particles. The kinetic energy of the particle is initially transformed into excitation or ionization energy of the absorber atoms. After approximately 10 microseconds, all energy appears in thermal form, hence the name *calorimeter*. Calorimeters are instrumented blocks of mostly dense absorbers with sensors embedded to detect these atomic excitations, which reflect the strength of the particle cascade and provide a measure related to the energy of the absorbed particle.

Calorimeters are widely used in experimental particle physics for several reasons:

1. They are sensitive to both charged (e.g., protons, pions, kaons, electrons) and neutral (e.g., neutrons, neutral pions, photons) particles.
2. The absorption process is statistical, so that in well-designed instruments the relative accuracy of energy measurements improves with the energy E, in principle, as $\sigma(E)/E \sim E^{-1/2}$.
3. The total depth D necessary to fully absorb (or "contain") the particle increases slowly with energy as $D \sim \ln E$, an important practical advantage.
4. The fundamental constituents of matter (called *quarks*) manifest themselves as tightly collimated groups of charged and neutral particles, whose total energy can only be measured with calorimeters.

In nuclear physics, the energy of particles (protons, nuclear fragments, electrons) or γ-quanta is typically in the mega electronvolts range; relatively small volumes of absorber (a few tens of cubic centimeters) are sufficient (Section 21.2).

In particle or high-energy physics (Sections 21.3–21.5), energies range from 100 MeV (0.1 GeV) to several 10^6 MeV (1 TeV). Total absorption proceeds, depending on the type of particle, via high-energy electromagnetic, strong, or even weak interactions.

Astroparticle physics (Section 21.6) has spawned the most original and diverse developments covering the widest spectrum of particles – known and unknown – as well as the largest energy range, from milli electronvolts to exa electronvolts (10^{19} eV).

The layout of a generic particle physics experiment is shown in Figure 21.1. The momentum of charged particles is derived from the track curvature in a magnetic field using position detectors, which barely affect the particle momentum. Calorimeters measure the energy of particles, charged or neutral. Photons, electrons, and positrons require relatively little mass to be absorbed. Hadrons, interacting predominantly through the strong interaction, require considerably more material for total absorption. The destructive nature of the calorimetric energy measurement imposes the spatial sequence of the measurement tasks and hence fixes the layout of experiments.

21.2
Calorimetry Methods in Nuclear Physics

21.2.1
Total Absorption and Range of Particles

In nuclear physics, the energy measurement of charged (protons, αs, nuclei, fission fragments, electrons, positrons) and neutral particles (neutral pions and, in particular, neutrons) is frequently required. Precision energy measurements of photons emitted from excited nuclear states are also important. The energies range typically from 100 keV to 100 MeV. In this energy range, the absorption of charged particles proceeds essentially through the electromagnetic (em) interaction.

The "mean energy loss per unit path length x" or "stopping power" is given to a good approximation by the Bethe–Bloch equation (Yao *et al.*, 2006):

$$\left[\frac{dE}{dx}\right]_{\text{inc}} = \frac{CZ_{\text{med}}\rho_{\text{med}}}{A_{\text{med}}}\left[\frac{Z_{\text{inc}}}{\beta}\right]^2 \times\left[\text{In}\left(\frac{2m_e\gamma^2\beta^2c^2}{I}\right) - \beta^2 + \text{correction term}\right] \quad (1)$$

where $C = 0.307\,\text{MeV}\,\text{cm}^2\ \text{g}^{-1}$, $Z_{\text{med, inc}}$ is the charge of the absorbing medium and incident particle, respectively, m_{e} the electron mass, and I the effective ionization potential of the absorber. Owing to the $1/\beta^2$ dependence on the particle velocity, most of the energy loss occurs near the end of the path. The energy loss, when expressed in units of length × density (grams per square centimeter), that is, the number of atomic electrons or scattering centers encountered, is approximately independent of the material (Figure 21.2).

The depth at which half the initial particles are stopped or "ranged out" is called the *mean range* $R(E)$. It is related to the energy loss by

$$R(E) = \int_E^0 \frac{-dE}{\left(\frac{dE}{dx}\right)} \quad (2)$$

Typical values of the mean energy loss and the range of various particles (Yao *et al.*, 2006) are given in Figure 21.2 and can be used to estimate the required absorber depth to range out charged particles in this energy domain. As an example, a pion with a momentum of $100\,\text{MeV}\,c^{-1}$ has a range of approximately $10\,\text{g}\,\text{cm}^{-2}$ in lead, that is, 0.9 cm: a lead layer of ≈1 cm will absorb or range out such a particle.

Neutrons with mega electronvolt energies are detected through elastic scattering (via the strong force) off the detector nuclei. The scattering process transfers a fraction of the neutron kinetic energy to the target nucleus, usually hydrogen, deuterium, or helium, resulting in a recoil nucleus. The recoil nuclei, in turn, lose their energy through ionization of the detector medium. This process is repeated until, essentially, all of the kinetic energy of the neutron has been transferred to the detector nuclei.

Low-energy quanta in the electronvolt to kilo electronvolt range will ionize the atoms of the detector medium and produce detectable electrons (atomic photoeffect). At higher energies, pair production, the conversion into electron–positron pairs in the em field of the nucleus,

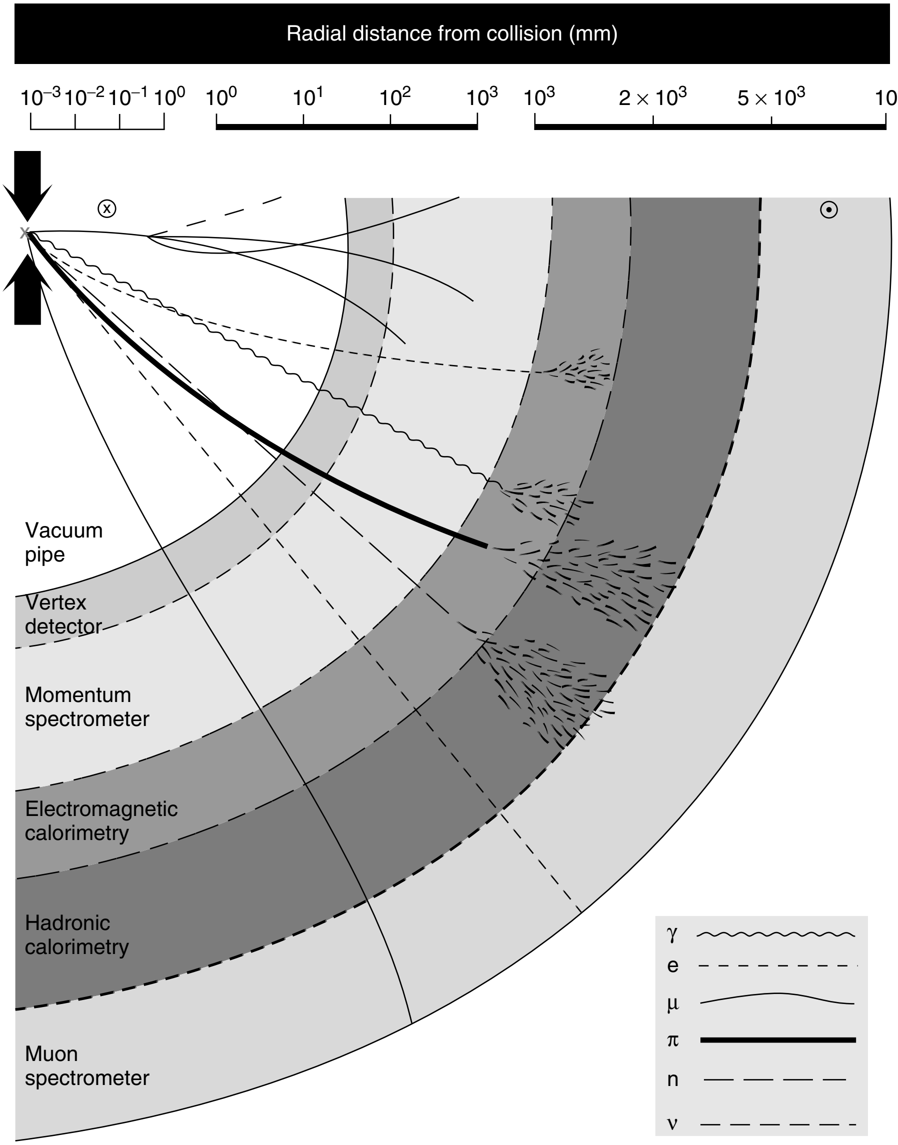
Radial distance from collision (mm)
10^{-3} 10^{-2} 10^{-1} 10^{0}
10^{0} 10^{1} 10^{2} 10^{3}
10^{3} 2×10^{3} 5×10^{3} 10^{4}
Vacuum pipe
Vertex detector
Momentum spectrometer
Electromagnetic calorimetry
Hadronic calorimetry
Muon spectrometer
γ
e
μ
π
n
ν

Fig. 21.1 Cross section through a conceptual configuration of a collider detector: like "Russian Dolls", a sequence of detectors enclose the collision point. In each detector layer, the particle is subjected to a specific measurement, shedding, layer by layer, its physics information. Note the three different logarithmic scales of the radial dimensions.

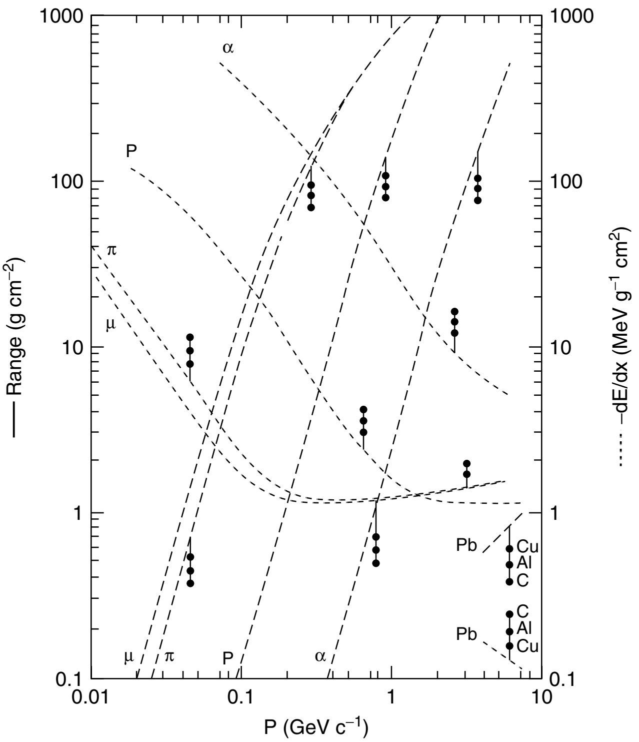

Fig. 21.2 Mean range (left ordinate, long-dashed lines) and energy loss (right ordinate, dashed lines) due to ionization for the indicated particles in Pb, with scaling to Cu, Al, and C indicated, using the Bethe–Bloch equation, including corrections (Yao *et al.*, 2006).

dominates. These processes are described in Section 21.3.

21.2.2
Principal Detection Techniques: Scintillators and Solid-State Detectors

Having described the absorption processes relevant for particles in the mega electron-volt region, we address the two principal techniques, scintillators and semiconductor detectors, used for measuring particles in this energy range.

In scintillators, the molecules are excited by the interaction of the charged particles, converting a small fraction into visible photon emission. Time constants are typically 10^{-9}–10^{-6} seconds. Both inorganic (crystals) and organic scintillators are used. The inorganic scintillators are usually grown with an admixture of dopants acting as recombination centers for ionization charge with subsequent excitation, followed by light emission. Certain inorganic scintillation crystals are compounds of heavy elements and hence have a large absorption coefficient and an excellent energy resolution for γ-rays (see Section 21.4). More recently, the scintillating properties of noble liquids, in particular liquid xenon, have been investigated. These liquids combine high γ-absorption cross sections with a very fast time response (Doke, 2005).

Only a relatively small fraction of the absorbed energy, typically a few percent, is converted into visible light that can be detected with photon detectors (e.g., photomultipliers). Nevertheless, the light output is quite sufficient; in "good" scintillators, there may be as many as 10^4 photons produced for 1 MeV of absorbed energy. This method is suitable because the average fraction of energy converted into light is constant and the light output is proportional, apart from saturation effects (Section 21.4) to the energy deposited.

The second category of detectors, also pioneered by nuclear physicists, uses the measurement of ionization in semiconductors (predominantly silicon and germanium). In these materials, the energy ε required to create the basic measurable excitations, electron–hole pairs, is approximately $\varepsilon \approx 3\,\text{eV}$, compared with typically 100 eV per photon needed in a scintillator. Therefore, for a given energy deposit, the statistical signal fluctuations σ are smaller and the energy resolution can be better. Experimentally, one finds $\sigma^2(E) = (FE/\varepsilon)\varepsilon^2$, and a "Fano" factor $F \approx 0.1$ for Si and Ge; the limit of energy resolution is therefore $\sigma \sim 500\,\text{eV}$ for 1-MeV energy deposit (Fabjan and Gianotti, 2003).

Semiconductor detectors work as ionization chambers in which the charge created by the energy deposit is collected under the influence of an externally applied electric field. Registration of this ionization charge is, however, only possible if it is not masked by free charges, generated thermally or due to the presence of impurities. Adequate suppression is possible through purification of the material or implementing a diode structure in the detector material, operated in reversed bias mode. For Ge, impurity levels as low as 10^9 atoms cm^{-3} have been achieved. Such ultrapure crystals are fabricated in sizes of hundreds of cubic centimeters, making them the detector of choice for γ-ray spectroscopy in the energy range of 10 keV to a few mega electronvolts. Silicon cannot yet be purified to this level, limiting the useful depth of Si detectors to a few millimeters. Such detectors are still well suited for the energy measurement of X-rays and heavy charged particles, such as protons, αs, and fission fragments.

21.3
Calorimetry Methods in Particle Physics: Photons, Leptons, and Hadrons

21.3.1
Absorption Mechanisms

In the "high-energy" domain ($E \geq 100$ MeV) of calorimetry, the incident particle is sufficiently energetic to initiate particle production. This provides a further mechanism for absorption in addition to those previously discussed (ionization, excitation, etc.). With increasing energy of the incident particle, the energy of the secondary particles may be sufficient for subsequent, that is, tertiary production, and so forth. The image of an "avalanche" or shower of particles appropriately describes this process of energy degradation. The particle multiplication continues until the energy of the last generation drops below the energy threshold for further production; at this point, the shower has reached the maximum number of particles – the shower maximum – after which further degradation continues via the low-energy processes described in Section 21.3.2.

There are two distinct types of particle showers. The first type is initiated by energetic electrons (positrons) and photons and propagates through the electromagnetic interaction with the atoms of the absorber. The second type of cascade is produced by particles strongly interacting with the absorber, that is, hadrons, such as pions, kaons, and protons. In both types of cascade, the relevant interaction cross sections reach values that are rather energy independent for particle energies $E \geq 1\,\mathrm{GeV}$. Therefore, the cascade dimensions are described by a mean free path λ between collisions with a specific value of λ for a given material, in general, different for the electromagnetic and hadronic cascades. The geometrical dimensions of a shower of given energy in different absorber materials will be similar, when expressed in the appropriate physics units of the mean free path (scaling in λ).

The principle of the calorimetric energy measurement of energetic particles may be formulated as follows: particle cascading (through the electromagnetic or strong interactions) degrades the energy over rather small geometrical distances. The charged component in the cascade is measured through its ionizing or excitation effect on the medium, which provides a signal approximately proportional to the energy of the incident particle (Fabjan and Gianotti, 2003).

21.3.2
Physics of the Electromagnetic Cascade

In spite of the apparently complex phenomenology of shower development in a material, electrons and photons interact with matter via a few well-understood QED processes, and the main shower features can be parametrized with simple empirical functions.

The average energy lost by electrons and the photon interaction cross section is shown in Figure 21.3 for a frequently used absorber material, lead, as a function of energy. Two main regimes can be identified. For energies larger than $\sim 10\,\mathrm{MeV}$, the main source of electron energy loss is bremsstrahlung. In this energy range, photon interactions produce mainly electron–positron pairs. Above $1\,\mathrm{GeV}$, both processes become roughly energy independent. At low energies, electrons lose their energy mainly through collisions with the atoms and molecules of the material, giving rise to ionization and thermal excitation; photons lose their energy through Compton scattering and the photoelectric effect.

Electrons and photons of sufficiently high energy ($\geq 1\,\mathrm{GeV}$) incident on a block of material produce secondary photons by bremsstrahlung, or secondary electrons and positrons by pair production. These secondary particles in turn produce other particles by the same mechanisms, thus giving rise to a cascade (shower) of particles with progressively degraded energies. The number of particles in the shower increases until the energy of the electron component falls below a critical energy ε, where energy is mainly dissipated by ionization and excitation and not in the generation of other particles.

The main features of electromagnetic showers (e.g., their longitudinal and lateral sizes) can be described in terms of one parameter, the radiation length X_0, which depends on the characteristics of the material (Yao *et al.*, 2006)

$$X_0\left(\mathrm{g/cm^2}\right) \approx \frac{716\,\mathrm{g\,cm^{-2}}A}{Z(Z+1)ln\left(\frac{287}{\sqrt{Z}}\right)} \tag{3}$$

where Z and A are the atomic number and weight of the material, respectively. The radiation length X_0 represents the average distance x that an electron needs to travel

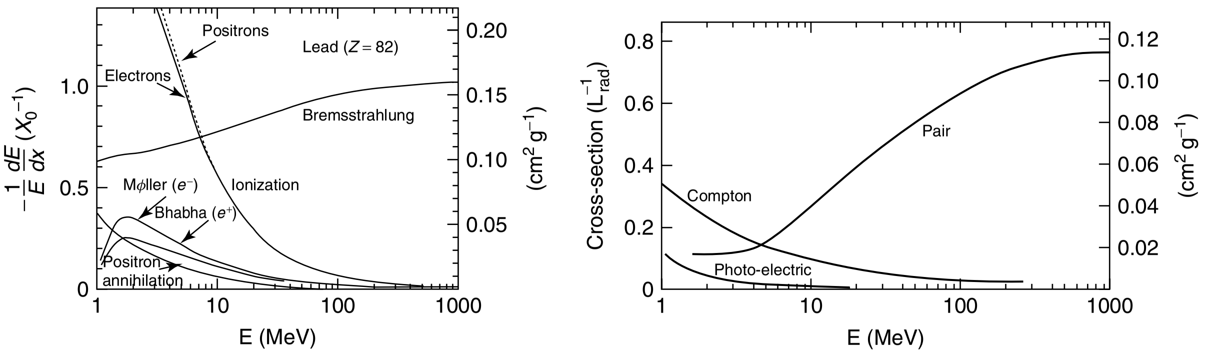

Fig. 21.3 (a) Fractional energy lost in lead by electron and positrons as a function of energy. (b) Photon interaction cross section in lead as a function of energy (Yao *et al.*, 2006).

in a material to reduce its energy to $1/e$ of its original energy E_0

$$\langle E(x)\rangle = E_0 \exp\left(-\frac{x}{X_0}\right) \tag{4}$$

Similarly, a photon beam of initial intensity I_0 traversing a block of material is absorbed mainly through pair production. After traveling a distance $x = 9/7X_0$ its intensity is reduced to $1/e$ of the original intensity

$$\langle I(x)\rangle = I_0 \exp\left(-\frac{7}{9}\frac{x}{X_0}\right) \tag{5}$$

These relations show that the physical scale over which a shower develops is similar for incident electrons and photons, and is independent of the material type if expressed in terms of X_0. Therefore, electromagnetic showers can be described in a universal way by using simple functions of the radiation length.

The shower maximum, that is, the depth at which the largest number of secondary particles is produced, is approximately located at $t_{\max} \approx \ln(E/\varepsilon) + t_0$, where $t_{\max}$ is measured in radiation lengths, E is the incident particle energy, and $t_0 = -0.5\,(+0.5)$ for electrons (photons). This formula shows the logarithmic dependence of the shower length, and therefore of the detector thickness needed to absorb a shower, on the incident particle energy. Longitudinal shower profiles for different energies of the incident particles are shown in Figure 21.4(a) and for electrons at 6 GeV in different material (Figure 21.4b). The calorimeter thickness containing 95% of the shower energy is approximately given by $t_{95\%} \approx t_{\max} + 0.08\,Z + 9.6$, where $t_{\max}$ and $t_{95\%}$ are measured in radiation lengths. As an example, in a calorimeter with thickness $\approx 25X_0$, the longitudinal shower leakage beyond the end of the active detector is less than 1% up to incident electron energies of $\sim$300 GeV. Therefore, even at the particle energies expected at the CERN Large Hadron Collider (LHC), of order about tera electronvolt, electromagnetic calorimeters (EMCs) are very compact devices: the ATLAS lead–liquid argon calorimeter (ATLAS Collaboration, 1996) and the CMS crystal calorimeter (CMS Collaboration, 1997) have thicknesses of $\approx$45 and $\approx$23 cm, respectively (the radiation lengths are $\approx$1.8 and $\approx$0.9 cm, respectively).

21.3.3
Physics of the Hadronic Cascade

Understanding the physics of the hadronic cascade is the basis for the construction of powerful, widely used instruments reaching performance limits imposed by the physics of the hadronic cascade.

By analogy with electromagnetic showers, the energy degradation of hadrons proceeds through an increasing number of mostly strong interactions with the calorimeter material.

In analogy to em cascades, a characteristic scaling length describes the longitudinal shower shape. This is the interaction length λ, defined as

$$\lambda = \frac{A}{N_{\mathrm{Av}}}\sigma_i\left(\mathrm{g\ cm^{-2}}\right) \approx 35A^{\frac{1}{3}}\ \mathrm{g\ cm^{-2}} \tag{6}$$

A being the mass number of the material, N_{Av} the Avogadro number, and σ_i the inelastic cross section. It describes the absorption of N hadrons in matter through the equation

$$\frac{dN}{dx} = \frac{-N}{\lambda} \tag{7}$$

For incident hadrons with sufficient energy to produce secondary hadrons ($E \geq 1$ GeV), a hadron shower is initiated. Typical

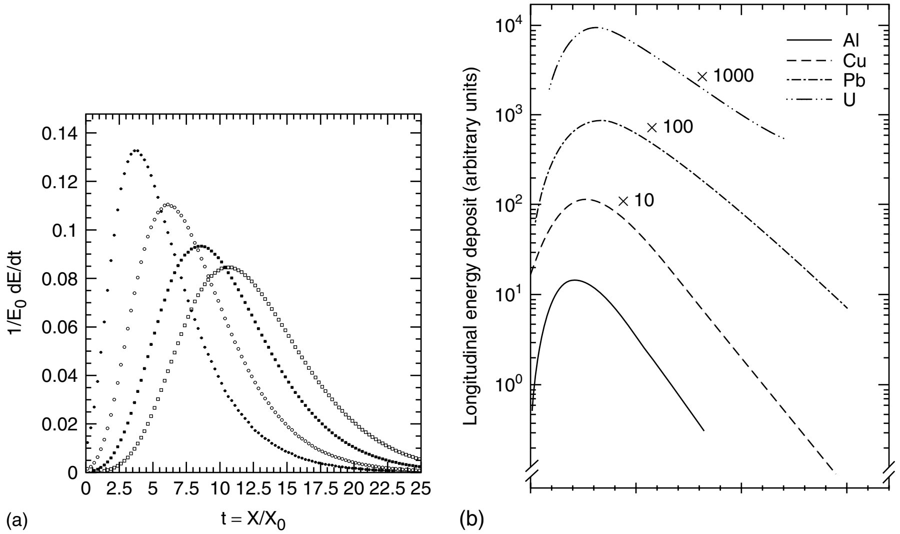

Fig. 21.4 (a) Simulated longitudinal shower profiles in $PbWO_4$ as a function of material thickness (expressed in radiation lengths), for incident electrons of energy (from left to right) 1, 10, 100, and 1000 GeV. (b) Longitudinal shower development of 6 GeV c^{-1} electrons in four very different materials, showing the approximate scaling in units of radiation lengths X_0.

longitudinal shower profiles in different absorbers are shown in Figure 21.5 and some useful features can be parametrized in approximate form as follows:

1. shower maximum $S_{\max}[\lambda] \approx 0.2 \ln E\,(\text{GeV}) + 0.7$;
2. longitudinal depth for 95% containment $L(0.95)\,[\lambda] \approx S_{\max} + 2.5E^{0.13}$ (GeV); and
3. transverse radius for 95% containment $R[\lambda] \approx 1[\lambda]$.

The hadronic interaction produces two classes of effects. Energetic secondary hadrons are produced with a mean free path (interaction length) λ between interactions; their momenta typically add up to a fair fraction of the primary hadron momentum, that is, at the giga electronvolt scale. Second, in the hadronic collisions with the absorber nuclei, a significant part of the hadron energy is consumed in nuclear processes such as excitation, nucleon evaporation, spallation, and so on, producing particles with characteristic nuclear energies at the mega electronvolt scale.

The fast, energetic hadronic component contains protons, neutrons, and charged and neutral pions. Owing to the charge independence of hadronic interactions, in each high-energy collision, on average, one-third of the pions produced will be π^0's, $F_{\pi 0} = 1/3$. This is the ''odd man'' in this group of energetic hadrons because these neutral pions will decay to two photons, $\pi^0 \rightarrow \gamma\gamma$, before having a chance to reinteract hadronically. These energetic photons will induce an electromagnetic cascade, proceeding along the laws of electromagnetic interactions (Section 21.3.2). This physics process acts like a ''one-way diode'', transferring energy

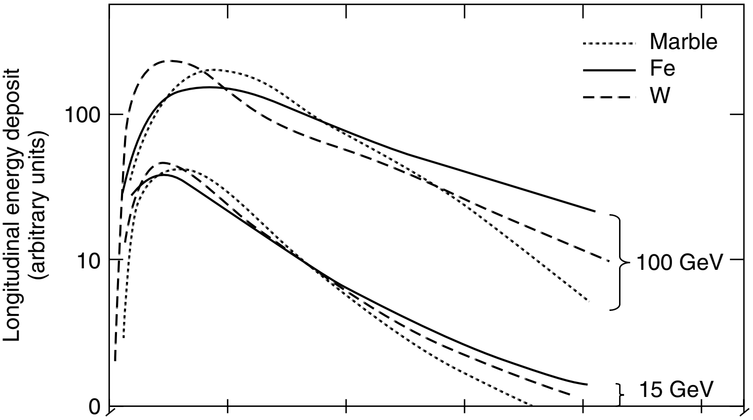

Fig. 21.5 Longitudinal shower development induced hadrons in different materials, showing approximate scaling in interaction length λ. The shower distributions are measured with respect to the face of the calorimeter (Fabjan and Ludlam, 1982; reproduced with permission of Annual Review of Nuclear and Particle Science).

from the hadronic part to the electromagnetic component, which will not contribute further to hadronic processes.

As the number of shower generations increases with increasing incident energy, so will the fraction of the electromagnetic cascade: there is an "identity change" of a predominantly hadron-rich cascade at low incident energy to a cascade developing into a mostly electromagnetic cascade at very high energies.

Second, contributions from nuclear interactions produce photons and neutrons in the few mega electronvolts range. The total energy carried by photons from nuclear reactions is substantial: only a fraction, however, will be recorded in practical instruments, as most of these photons are emitted with a considerable time delay ($\geq$1 microsecond). These nuclear effects, represented by delayed photons and soft neutrons, produce signals that are poorly or not at all detected in practical instruments. One part of the cascade particles is truly "invisible" or "missing", namely energy carried by escaping muons, neutrinos, or backscattered (albedo) particles. Let η_e be the efficiency for observing a signal E^e_meas (measurable energy) from an electromagnetic shower, that is, $E^e_\mathrm{meas} = \eta_e E$ (em); let η_h be the corresponding efficiency for purely hadronic energy to provide visible energy in an instrument. Therefore, for a pion-induced shower the visible energy E^π_meas is

$$E^\pi_\mathrm{meas} = \eta_\mathrm{e} F^0_\pi E + \eta_\mathrm{h} F_\mathrm{h} E \qquad (8)$$

where E is the incident pion energy. In Figure 21.6, the dependence of the various components as a function of the energy of the incident hadron is shown.

The ratio of observable signals induced by electromagnetic and hadronic showers, usually denoted by "e/π", is therefore

$$\frac{E^\pi_\mathrm{meas}}{E^e_\mathrm{meas}} = \left(\frac{e}{\pi}\right)^{-1} = 1 - \left(1 - \frac{\eta_h}{\eta_e}\right) F_h \qquad (9)$$

In general $\eta_h \neq \eta_\mathrm{e}$; therefore, the average response of a hadron calorimeter (HC) as

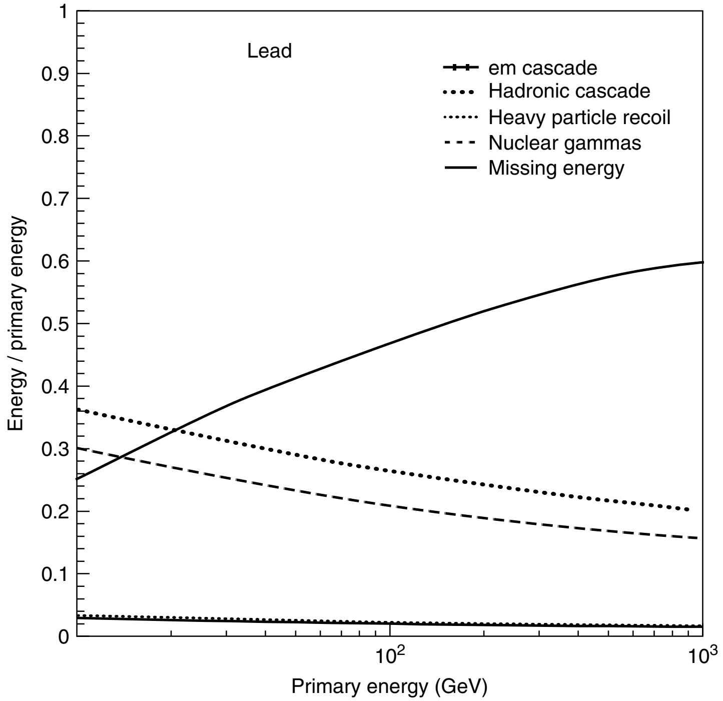

Fig. 21.6 Characteristic components of proton-initiated cascades in lead. With increasing primary energy, the em ($\pi^{0)}$) component increases. The component due to nuclear gammas is, in general, time-delayed and is detected with very low efficiency in practical instruments (Ferrari, 2001, *private communication*).

a function of energy will not be linear because F_{h} decreases with incident energy. More subtly, for $\eta_{\mathrm{h}} \neq \eta_{\mathrm{e}}$, we have to expect that event-by-event fluctuations in the F_{h} and F_{π}^{0} components will have an impact on the energy resolution of such instruments. The relative response "e/π" turns out to be the most important yardstick for gauging the performance of HCs (Wigmans, 1991).

21.4
Performance of Calorimeters

Suitably instrumented calorimeters measure not only the energy but also the time, direction, and impact point of the incident particle. The quality of the measurement (resolution) is driven by fluctuations in the measurement process, both intrinsic, due to the physics of particle absorption, but also due to a range of instrumental contributions.

With increasing energy, the particle content of the shower increases and therefore intrinsic fluctuations will scale as $1/\sqrt{(\text{number of particles})} \sim 1/\sqrt{E}$. Instrumental contributions may scale differently with energy. The conventional parametrization for the energy resolution is $\sigma(E)/E = a/\sqrt{E} \oplus b/E \oplus c$, a denoting the stochastic, b the "noise term", and c a constant term. Measurement of the shower

Tab. 21.1 Characteristics of commonly used crystals for electromagnetic calorimeters.

Crystal	Density ($g\,cm^{-3}$)	Radiation length (cm)	Measured energy resolution (σ at 1 MeV)	Reference
CsI	4.53	1.85	2%	Moses (2002); Lecoq *et al.* (2006)
BGO	7.13	1.12	3.1%	Moses (2002); Lecoq *et al.* (2006)
$PbW0_4$	8.28	0.89	na	Moses (2002); Lecoq *et al.* (2006)

Tab. 21.2 Characteristics of Noble Liquids used in electromagnetic calorimeters.

Liquid	Density ($g\,cm^{-3}$) at boiling point	Radiation length (cm)	Measured energy resolution (σ at 1 MeV)	Reference
Argon	1.40	14	1.2%	Doke (2005)
Krypton	2.41	4.7		Doke (2005)
Xenon	2.94	2.8	1.6%	Doke (2005)

shape provides important information on the particle impact, its direction, and time of arrival. Furthermore, the shower shape encodes information about the particle type.

21.4.1
Energy Resolution in Homogeneous Electromagnetic Calorimeters

Homogeneous calorimeters are instruments in which the absorber serves also as the detecting medium. These fully sensitive instruments provide the most complete information about the shower cascade and hence the potentially best measurement performance. This type of instrument is particularly suitable for precision EMCs, for which crystals (Table 21.1) and liquefied noble gases (Table 21.2) are the most frequently used materials. At low (~1 MeV) energies, the achieved energy resolutions are at the 1–2% level. At these low energies, the measured energy resolution for crystals is quite consistent with counting statistics, based on the number of registered photons. For noble liquids, however, the measured low-energy resolution is a large factor (5–10) worse compared with the theoretical one based on the number of created ion pairs. Recombination and incomplete collection of the generated ion pairs are thought to be the origin of this discrepancy.

For particle physics applications at high energy – above 1 GeV – naïve $1/\sqrt{E}$ scaling from the 1 MeV energy domain is not observed. Such scaling would give resolutions at the 0.03% level at 1 GeV, far better than the measured resolutions, observed to be at the 1% level. This fact is generally attributed to the nonlinear response of the detector medium to the very low energy component (below 1 MeV) of the electromagnetic cascade, frequently parametrized with a cutoff value below which the detector medium effectively becomes insensitive. Neither detailed simulations with state-of-the-art Monte Carlo codes nor the low-energy results support this explanation. Plausibly, several effects conspire: a certain fraction ($\leq 0.5\%$) of the shower energy escapes through the front, resulting in $\sigma(E)/E \sim 0.3\%/\sqrt{E}$. In practical crystal calorimeters, inhomogeneous light production and transport also contribute at a comparable level as do counting statistics and electronics noise. Above 1 GeV, this $1/\sqrt{E}$ scaling is observed to hold to a very good approximation over a very wide energy range from one to several hundred giga electronvolts (CMS Collaboration, 1997; Lecoq *et al.*, 2006).

21.4.2
Instrumental Contributions to the Electromagnetic Energy Resolution

Sampling fluctuations arise in calorimeters that are not uniformly (homogeneously) sensitive, that is, in instruments having their signal sensors interspersed in the absorber (see discussion in Section 21.5.2). Such constructions are very common for EMCs, and almost exclusively used HCs. The shower profile is sampled, leading to incomplete signal information with increased signal fluctuations. The magnitude can be estimated by considering the variations in the number of charged particles N_{ch}, which cross the active layers and which produce the detectable signal. This number is proportional to

$$N_{\mathrm{ch}} \propto \frac{E_0}{\Delta E} \tag{10}$$

where ΔE is the energy loss in one ''unit cell'' (absorber plate + active layer). Assuming statistically independent crossings of the active layers, the ''sampling'' contribution to the energy resolution due to the fluctuation of N_{ch} are

$$\begin{aligned}\frac{\sigma}{E} \div \frac{1}{\sqrt{N_{\mathrm{ch}}}} \div \sqrt{\frac{\Delta E}{E_0(\mathrm{GeV})}} \\ = c\sqrt{\frac{\Delta E(\mathrm{MeV})}{E_0(\mathrm{GeV})}},\ c = 0.032\end{aligned} \tag{11}$$

In reality, the particles are not statistically independent and cross the gap on average under an angle different from normal, increasing ΔE. A coefficient $c \approx 0.05$ is consistent with measurements in many different sampling EMCs (Fabjan and Gianotti, 2003; Wigmans, 2000).

In practical sampling EMCs, the energy resolution is in the range of $5{-}20\%/\sqrt{E(\mathrm{GeV})}$. In well-designed instruments, it is below 1% for 100-GeV photons or electrons, consistent with the detection requirements for the physics discovery at the CERN LHC (ATLAS Collaboration, 1996).

There are several other instrumental contributions to the energy resolution that can be kept at a level below the intrinsic resolution of homogeneous calorimeters, well below the sampling contribution. Incomplete shower containment (transverse or longitudinal shower leakage) affects the performance, as do mechanical- or signal-collection nonuniformities. These instrumental contributions imply a lower limit to the resolution, usually parametrized with a constant term

c. At low energies electronic readout noise may contribute to the resolution, decreasing with energy as $1/E$. Achieving 1% energy resolution implies control of all these instrumental effects at a level well below: a major engineering challenge.

21.4.3
Time, Space, and Angular Resolution in EMCs

While the energy measurement is the principal task of calorimeters, such instruments also provide additional powerful information about the detected photons or electrons (positrons).

Pulse shape analysis of the detected signal can provide timing information on the arrival time of the particle. The performance depends – as always – on the signal/noise ratio and therefore scales – at least in principle – with $1/\sqrt{E}$. In modern instruments, the timing resolution is typically at a level of a few nanoseconds/$\sqrt{E}$ (giga electronvolt).

Space and angular resolution are achieved by appropriate subdivision of the detector in transverse and longitudinal directions. The scale for the subdivision is the radiation length X_0. Longitudinally cells typically have dimensions of a few X_0, while transversely the subdivisions are at the scale of 1 X_0. Such granularity in the readout allows a spatial resolution of the impact point of the particle at the level of $\sim 0.5\ X_0/\sqrt{E}$ (GeV) and an angular measurement of $\sim$50 mrad/$\sqrt{E}$ (GeV) (ATLAS Collaboration, 1996).

21.4.4
Energy Resolution of Hadron Calorimeters

The general analysis presented for EMCs is conceptually also applicable to HCs, albeit with significant quantitative differences due to the physics of the much more complex hadronic cascade.

Inevitably, hadronic cascades imply nuclear interactions with their correlated production of low-energy neutrons and photons, only partially detectable (invisible energy) (Section 21.3). On average, for particles with the same incident energy, the signal response of a calorimeter to hadrons (e.g., pions, π) will be lower than that to electrons, that is, $e/\pi > 1$. For hadrons, the visible energy will fluctuate, event by event, between two extremes: from fully electromagnetic, yielding the same signal as an electron to fully hadronic with a maximum of invisible energy, as shown conceptually in Figure 21.7.

This simple analysis already provides the following qualitative conclusions for HCs with $e/\pi \neq 1$:

- fluctuations in $F_{\pi 0}$ are a major contribution to the energy resolution;
- the average value $< F_{\pi 0} >$ is energy dependent and therefore calorimeters have a response to hadrons, which is nonlinear with energy;
- the above-mentioned fluctuations are non-Gaussian and therefore the energy resolution scales weaker than $1/\sqrt{E}$.

Empirically one finds that such instruments show typically an $e/\pi \approx 1.4$ and an energy resolution of $\sigma/E \geq 0.4/\sqrt{E}$ (GeV).

Understanding these effects suggests how to design instruments in which the response to this nuclear component is boosted or ‘‘compensated for’’, thereby achieving, on average, an equal response to electrons and pions, $e/\pi = 1$. Such instruments are referred to as *compensated calorimeters*.

Remarkably, it is possible to ‘‘tune’’ the e/π response of a calorimeter in the quest to achieve $e/\pi = 1$, and thus optimize the performance. This ‘‘tuning’’

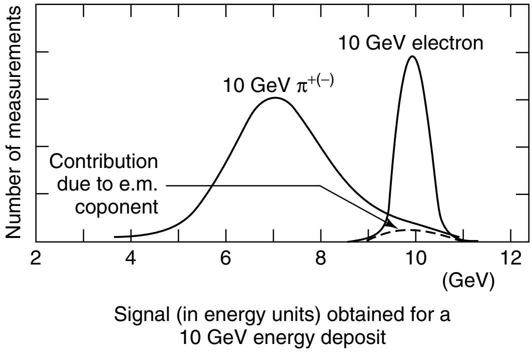

Fig. 21.7 Response function of calorimeters for electrons and hadrons. The curves are for an EMC with $\sigma/E = 0.1/\sqrt{E}$ and for a HC with $\sigma/E = 0.5$ and $e/\pi = 1.4$. A fraction of the hadron-induced cascade propagates almost completely electromagnetically, producing non-Gaussian tails and broadening of the response function.

is based on the fact that practical HCs are almost always built as sampling devices; the energy sampled in the active layers, E_{samp}, is typically a small fraction, a few percent or less, of the total incident energy.

The energetic hadrons lose relatively little energy ($\leq$10%) through ionization before being degraded to such low energies that nuclear processes dominate. Therefore, the response of the calorimeter will be strongly influenced by the response to neutrons (n/mip) and photons (γ/mip) relative to the signal produced by the fast (mip) hadrons (in both the absorber and the readout materials).

As an example, it is possible to influence the response to the abundant neutrons by choosing a readout material rich in hydrogen. The $n - p$ elastic cross section is large and, on average, half of the neutron kinetic energy is transferred to the recoil proton produced in the active material, contributing directly to the calorimeter signal: a 1-MeV proton has a range of ~20 μm in a scintillator and is not sampled unlike a fast hadron (mip). This difference in response between high-Z absorbers and hydrogen-containing readout materials is exploited in the "tuning".

The mip signal will be inversely proportional to the thickness of the absorber plates: increasing the thickness of the absorber plates will decrease E_{samp}, whereas the signal from proton recoils will not be affected and hence neutrons will contribute more significantly to e/π. Changing the sampling fraction changes e/π (Wigmans, 1991).

Having tuned the response to $e/\pi \sim 1$ results also in much improved energy resolution, as conceptually suggested in Figure 21.7, and qualitatively borne out by shower and signal simulations, confirmed by measurements (Fabjan and Gianotti, 2003). One observes a significant reduction in the fluctuations and an intrinsic hadronic energy resolution of $\sigma/E \sim 0.2\sqrt{E(\text{GeV})}$. In principle, tuning of e/π can be applied to all sampling calorimeters, opening the way to better performance. Note that this tuning can only be achieved in sampling calorimeters: a homogeneous calorimeter, for example, a BGO crystal calorimeter would show the typical response of an uncompensated device with a resolution of $\sim 0.4/\sqrt{E}$ (GeV).

Sampling fluctuations contribute, of course, to the energy resolution for which a similar empirical parametrization can be given as for EMCs:

$$\frac{\sigma_{\text{samp}}}{E} = c \cdot \left(\frac{\Delta E(\text{MeV})}{E(\text{GeV})}\right)^{1/2} \tag{12}$$

where ΔE is the energy lost in one sampling cell, with c (hadron) ≈ 0.09 for HCs (Fabjan and Gianotti, 2003).

For high-performance HCs, sampling fluctuations cannot usually be neglected.

We can summarize the principles of modern, optimized hadron calorimetry as follows:

- the key performance parameter is $e/\pi = 1$, which guarantees linearity, $E^{-1/2}$ scaling of the energy resolution, and best resolution;
- by proper choice of type and thickness of active and passive materials the response can be tuned to obtain $e/\pi \approx 1$;
- the intrinsic resolution in practical HCs can be as good as $(\sigma/E) \cdot \sqrt{E} \leq 0.2$;
- sampling fluctuations contribute at the level of $\sigma/E \approx 0.09(\Delta E\,(\mathrm{MeV})/E\,(\mathrm{GeV}))^{1/2}$.

21.5
Calorimeter Systems

The attractive detection properties of calorimeters – energy measurement of charged and neutral particles, improved performance with increasing energy – have made them essential components in modern nuclear and particle physics experiments. In the following we discuss the most important construction principles, frequently used instrumentation techniques, and calorimeter facilities at the LHC.

21.5.1
Calorimeter Construction Principles

A host of calorimeter techniques has been developed to optimize one or several of the following performance parameters:

1. linear response as a function of energy, covering a range from 100 MeV to >1000 GeV;
2. low instrumental noise of the sensors to avoid a degradation of the energy resolution;
3. high degree of readout segmentation is required for position resolution and discrimination between electrons (photons) and hadrons based on the differences in shower shape;
4. insensitivity to magnetic fields, temperature variations, or ionizing radiation on the sensors.

Given the very different properties of electromagnetic and hadronic showers, different optimization criteria apply to EMCs and HCs.

EMCs may be homogeneous or of the sampling type; HCs are essentially always sampling instruments. The calorimeters must also measure the topology (position, direction) of complex particle collisions requiring calorimeters of high lateral and longitudinal subdivision. Many experiments require precision at the 1% level over the full area of calorimeter facilities.

21.5.2
Instrumentation Techniques

All detectors, sensitive to charged particles in the shower cascade, can and are being used to obtain a recordable signal. Two categories are predominant: scintillators (crystals, plastic scintillators, liquefied noble gases) and charge-collecting detectors (liquefied noble gases, semiconductor detectors). A modern example of a homogeneous EMC, using crystals, is the CMS calorimeter, actually being installed at the CERN LHC, Figure 21.8. It uses approximately 76 000 $PbW0_4$ crystals (Table 21.1), surrounding the collision point almost ‘‘hermetically’’. This device has been developed to discover the putative Higgs particle via the decay channel $H \rightarrow \gamma\gamma$, requiring

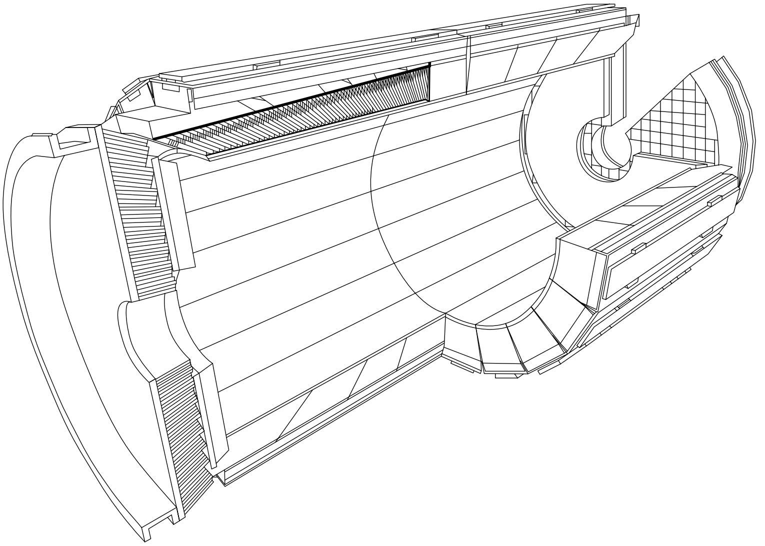

Fig. 21.8 Cutaway view of the CMS crystal EMC. The crystal points toward the collision point and is tapered for crackless coverage of the detection volume. The detector, consisting of 76 000 crystals, has a length of almost 8 m and a diameter of 3.6 m.

a precision of better than 1% at 100 GeV. $PbW0_4$ is the most recent example of a crystal being developed for particle physics by an international, 10-year long R&D effort with the aim to achieve the needed transparency and radiation hardness.

An example of a modern version of plastic scintillator readout is the ATLAS HC, installed at the LHC, Figure 21.9. It is doubly remarkable: the absorber plates, made from steel, are oriented longitudinally, along the direction of the incident hadron. This geometry works as well as the conventional, transverse orientation, because hadrons, after having traversed the preceding EMC, have developed a sufficiently broad shower to be insensitive to the precise orientation of the absorber plates. This geometry was chosen for ease of construction. Scintillation ''tiles'' are nested in slots; the light is transmitted to and absorbed in wavelength-shifting (WLS) fibers, running closely along two edges of the tile. The WLS fibers reemit the absorbed scintillation light at the larger wavelength, to which the fibers are transparent; a group of fibers from tiles covering a certain volume, a ''cell'', are routed to and read by a photomultiplier. The necessary subdivision of the calorimeter is naturally achieved.

Another novel use of scintillators is in the form of scintillating fibers, one to a few millimeters in diameter. Geometries with fibers perpendicular or along the direction of the incident particle have been used successfully (Fabjan and Gianotti, 2003).

Plastic scintillators are chosen for their potential to tune the e/π response (Section 21.4.4) and their relative facility in calorimeter construction. Systematic effects connected with light collection and material ageing need, however, to be very carefully handled, often requiring complex in situ calibration methods.

The second important instrumentation family is based on charge collection from

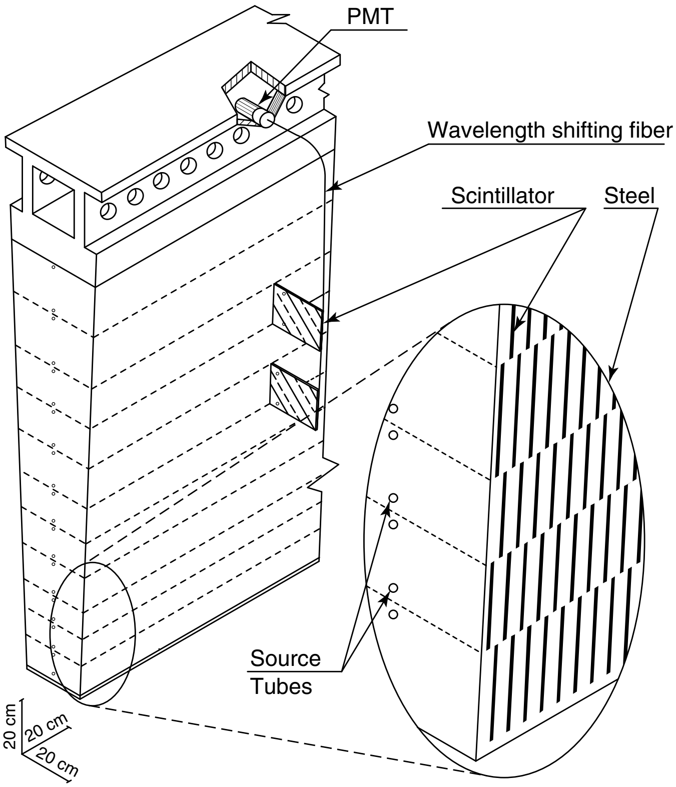

Fig. 21.9 View of one module of the central ATLAS hadronic calorimeter. Thirty-six such modules complete the cylindrical detector. Each of the longitudinally oriented scintillating tiles is read with two wavelength-shifting fibers.

liquid noble gases (Table 21.2). One recent example of a quasi-homogeneous liquid krypton EMC is shown in Figure 21.10. This instrument was built for a fixed target experiment at the CERN SPS to study CP violation, measuring the reaction $K^o_{s,l} \to \pi^0\pi^0$ and $K^o_{s,l} \to \pi + \pi -$. For this measurement, an EMC was needed with an energy resolution of $\sim 5\%/\sqrt{E}$ (GeV) and a position resolution of ~ 1 mm for 25 GeV photons (Fanti *et al.*, 2007). Liquid krypton was chosen as a compromise between the commonly used LAr, which has a quite large radiation length making a homogeneous detector quite bulky and the "ideal" noble liquid, LXe, which is prohibitively expensive for the volume needed. The use of LXe for precision photon and electron calorimetry is extremely attractive, the cost (and the necessary purity) however being a formidable obstacle. The largest LXe device to date is being prepared for a high-sensitivity search for the forbidden decay mode $\mu \to e\gamma$, in which a 53-MeV photon would be emitted. The MEG Collaboration has constructed a 900-l homogeneous LXe EMC with remarkable performance on energy resolution of $\sigma/E \leq 1\%$ at 100 MeV (a world record)

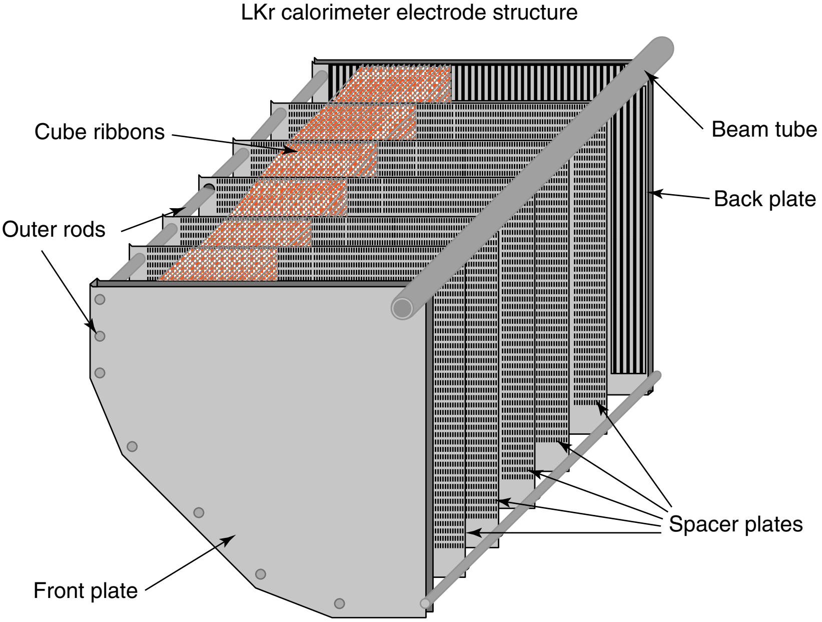

Fig. 21.10 Schematic view of a section of the NA48 liquid krypton electromagnetic calorimeter.

and a time resolution of ≈ 50 ps mm is reported (Sawada, 2007). It should be noted that only the scintillation light is recorded.

Liquid argon has been used for more than three decades as a readout medium: it is cheap, can be easily purchased at sufficient purity, and can be conveniently refrigerated with LN_2. The most advanced incarnation is the ATLAS EMC (Figure 21.11). It is a sampling calorimeter with again an unusual absorber structure in the form of "accordions" made of folded Pb-plates, covered with a stainless-steel skin. This geometry was chosen to achieve the following LHC requirements:

- uniform sensitivity to <1%;
- relatively fast readout;
- good transverse and longitudinal segmentation.

The geometry has allowed the construction of a cylindrical EMC without cracks or insensitive regions. Noble liquids are chosen for the advantages associated with charge collection and its superior control of systematic effects. Above 50 GeV, the ATLAS sampling EMC compares very well in performance with the crystal calorimeter of CMS.

21.5.3
Calorimeter Facilities

The physics and detector requirements for the LHC program have stimulated a major, worldwide R&D effort to meet the long list of conditions:

- *Fast response*: At the LHC design intensity, on average, 30 collisions will take place every 25 nanosecond. Most of these collisions will reflect "conventional" physics with relatively low-energy particles. With fast calorimeter response – 50–100 nanoseconds – the integration over

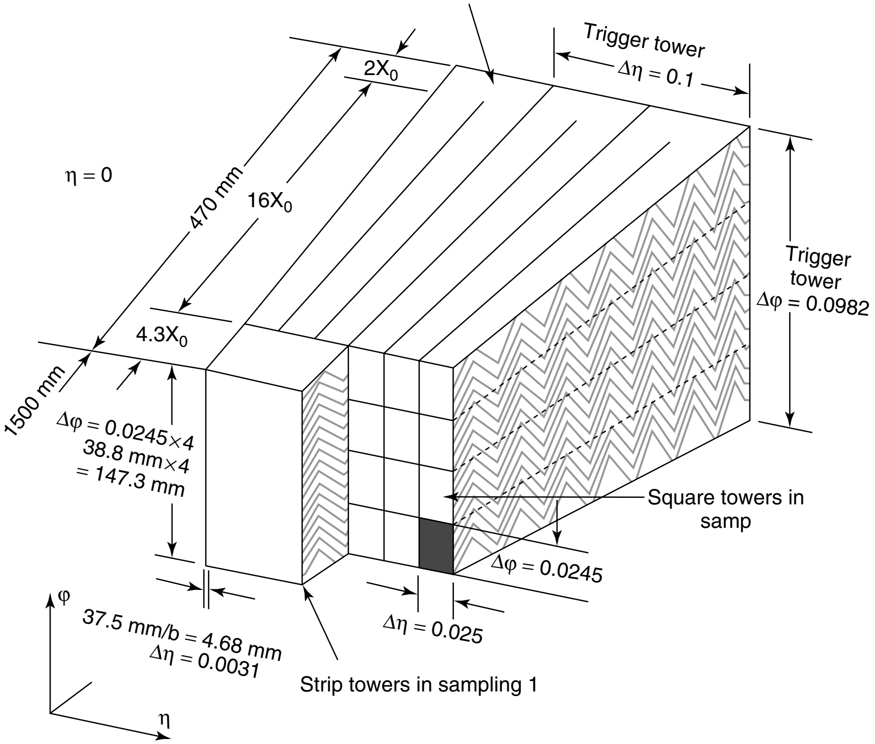
Trigger tower
Δη = 0.1
2X₀
η = 0
470 mm
16X₀
Trigger
tower
Δφ = 0.0982
4.3X₀
1500 mm
Δφ = 0.0245×4
38.8 mm×4
= 147.3 mm
Square towers in
samp
Δφ = 0.0245
φ
37.5 mm/b = 4.68 mm
Δη = 0.0031
Δη = 0.025
Strip towers in sampling 1
η

Fig. 21.11 Schematic view of the geometry and segmentation of the ATLAS electromagnetic calorimeter.

$\leq$100 events produces a "low-energy noise" in which the "new" physics signals are still well detectable.

- *Radiation hardness*: This is one of the severest technical requirements. At small angles to the collision axis radiation levels of $\sim 10^{17}$ neutrons cm^{-2} and $\sim 10^7$ Gy will be accumulated after 10 years of operation.
- *Construction methods that minimize nonsensitive areas*: Neutrinos, a frequent component of "new" physics, are recognized through an apparent "missing" energy. This method works, provided cracks do not simulate this signal.
- *Excellent electromagnetic energy resolution:* The "Holy Grail" of LHC physics is the search for the Higgs particle, which has an important decay channel of $H \rightarrow \gamma\gamma$. A resolution of 1% at 100 GeV is required, combined with an angular measurement of ~5 mrad for 100-GeV photons.
- *Large dynamic energy angle* for electrons, photons, and hadrons from a few giga electronvolts to few tera electronvolts.

In Figure 21.12, the calorimeter facility developed for the ATLAS experiment is shown. The very stringent requirements, combined with more mundane considerations, such as cost, have led to different choices for the EMC and HC; the radiation levels in the small-angle regions are so high that the very robust LAr technique was also chosen for the HC.

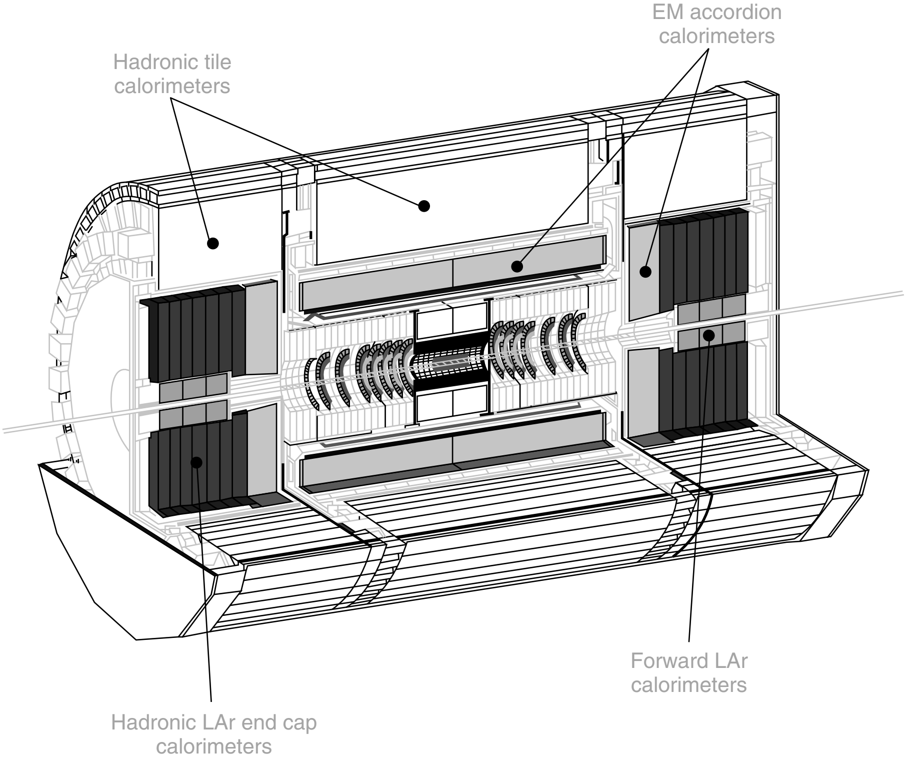
EM accordion calorimeters
Hadronic tile calorimeters
Forward LAr calorimeters
Hadronic LAr end cap calorimeters

Fig. 21.12 Partial cutaway view of the ATLAS calorimeter system (and the inner charged-particle tracker). The diameter is 8.5 m and the total length 13 m. The outer cylinder and the external end caps measure the hadrons; the inner (light gray) cylinders and end caps are the "Accordion" LAr EMCs. In principle, all particles are absorbed in this calorimeter system with the exception of neutrinos and muons, the latter being measured in the large muon spectrometer surrounding the calorimeters (not shown).

21.6
Calorimeters in Astroparticle Physics

This relatively new field addresses questions in cosmology and particle physics, which cannot be addressed with accelerator-based experimentation. At the energy frontier of 10^{20} eV are ultrahigh-energy cosmic rays, impinging on the Earth at a rate of one every few seconds. Another high-energy "cosmic" phenomenon is the study of sources of very high (hundreds of mega electronvolts to few tera electronvolts) photons. Not surprisingly, these studies are mostly based on extensions of the calorimetric principle. Given the rarity of these events, the calorimeter has to be gigantic. It turns out that the Earth's atmosphere is a most precious gift; at sea level, it represents approximately 28 radiations and 17 collisions length: an ideal, homogeneously sensitive calorimeter.

Ingenious approaches have been developed to instrument this homogeneous absorber. Relevant performance parameters are as follows (Hoffman, 1999; Pretzl, 2005):

- the energy threshold for particle detection;
- the systematic error on the energy scale;
- the energy resolution;
- their potential to discriminate between different particles.

One approach explores the Cherenkov light induced in the atmosphere by the relativistic component of a very high energy shower. This technique is tailored to explore the energy window between space-based detectors, rate limited to $E \leq 10\,\text{GeV}$, and ground-based systems, limited by signal to noise to $E \geq 300\,\text{GeV}$. The lateral distribution of the "Cherenkov shower" is narrowly concentrated around the shower axis and can be observed in a dark night sky. The Cherenkov light is collected with large mirrors and focused onto an array of light detectors, providing the energy, direction, and discrimination between photon- and hadron-induced showers. This technique allowed the recent observation of γ point sources (Hoffmann, 1999).

In another strategy, the fluorescence excited by the passage of the shower particle in the atmosphere provides the measure of the particle's energy. An array of telescopes is used to focus the light onto arrays of photo detectors. One observes a very large volume of the atmosphere, close to 10 000 km^2 sr in a fine grid of $1^0 * 1^0$. This technique measures the longitudinal shower profile and the shower maximum, sensitive to the composition of the cosmic rays. With its very high sensitivity, it has recorded some of the highest energy cosmic rays, to date.

The oldest of the three techniques has recently witnessed a remarkable renaissance. It uses ground-based detectors to record the tails of particle cascades initiated by sufficiently energetic primaries. The development of this technique (Cronin *et al.*, 1993) aims at the following:

- increased sensitivity towards higher energies;
- better understanding of the absolute energy scale;
- identification of the primary composition.

The most ambitious project is the Auger observatory, covering an area of 3000 km^2 with a variety of detectors: ground-based devices to measure the ratio between electromagnetic showers and muons, combined with detectors measuring the atmospheric fluorescence, providing complementary energy and particle type information. A spectacular proof of the power of this method has been the recent observation that ultrahigh-energy (close to 10^{20} eV) cosmic rays are emitted from extragalactic sources (Auger Collaboration, 2007).

Neutrino astronomy is another recent frontier requiring the most massive detectors: the water of the oceans and the Antarctic icecap has been instrumented to provide prototypes on the way to instrumenting one cubic kilometer (Learned, 2000).

While high-energy astrophysical applications use known calorimetric techniques in innovative ways, the detection of very low energy deposits in the milli electronvolt to electronvolt range requires totally new detection principles, opening the exploration of other fundamental physics questions, such as the following:

1. improved lower limits on the rest mass of neutrinos;
2. the search for weakly interacting massive particles (WIMPs), which may be responsible for the component of dark (invisible and not yet accounted for) matter in the Universe;

3. the search for magnetic monopoles, which may be relics of the Big Bang; and
4. the search for relic neutrinos from the Big Bang, which may have energies of $\sim$1 MeV.

The basic detection principle, registration of the interaction of the particle with the detector during the absorption process, is again calorimetry. However, the energies involved are below the atomic excitation levels, and different detection processes have to be invoked.

These calorimeters are characterized by operation at cryogenic temperatures, in the few millikelvins to $\sim$1 kelvin range. They exploit either some specific properties of the superconducting phase of matter or the reduction in the thermal noise of the detecting medium. Given the very low thresholds of the basic detection process, the number of elementary excitations can be considerably higher, and these detectors therefore hold the promise of much improved energy resolution.

Operation at cryogenic temperatures opens a number of detection channels:

1. Cooper pairs in superconductors have binding energies in the 10^{-6} to 10^{-3} eV range and may be broken into quasiparticles by phonon (lattice vibration) absorption.
2. The specific heat c for dielectric crystals decreases as $\sim \exp(-1/kT)$, resulting in a measurable, relatively fast temperature rise for very low energy absorption.
3. Energy absorption may induce a change in magnetization, latent heat release, or quasiparticle multiplication.

A powerful application of the detection in superconductors is being developed for dispersive spectroscopy in the wavelength region from 0.5 nm to 2 μm (Peacock, 1999). Arrays of superconducting tunneling junctions serve the dual function of absorber and detector. The number of excess carries N produced by the photon absorption as a function of the wavelength λ is

$N(\lambda) \approx 7 * 10^5/(\lambda\,(\mathrm{nm})\,\Delta\,(T/T_c)\,(\mathrm{meV}))$, Δ being the band gap.

Typical numbers of N are 10^6 at $\lambda \approx 1$ nm and the spectroscopic resolution is at a few percent level.

"Bolometric" detectors are an example of an application of concept 2, the true calorimeters, in which the temperature

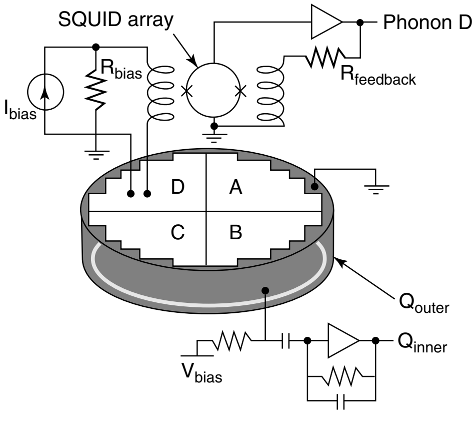

Fig. 21.13 Schematic diagram of a massive (few hundred grams) Ge detector for dark matter searches. Simultaneous measurements of phonons with transition-edge sensors and of the ionization charge provide the required background rejection (Cabrera, 2000).

rise due to the interaction of particles in an insulating crystal maintained at very low temperatures may be measured with a resistive thermometer. The energy resolution (neglecting noise in the readout electronics) is evaluated to be $\sigma(E) \sim T^{5/2}M^{1/2}$ (M being the mass of the bolometer); X-rays have been measured with 0.1% energy resolution (at 6 keV), superior to that obtained with conventional semiconductor detectors. An example is shown in Figure 21.13: using Ge allows simultaneous measurement of temperature and ionization providing powerful (and needed) discrimination between the putative WIMP and conventional background (Booth, 1996). Detectors with a mass between 100 and 1000 kg are being considered for sensitive dark-matter searches (Enns, 2005).

Glossary

Calorimetry: In its most general definition, a technique to measure the energy of a reaction or a process. Sometimes, the energy measurement is related to a temperature measurement, hence the name *calorimetry*. In particle physics, it is the name given to techniques that measure the energy of particles or quanta.

γ-Quanta: Energetic quanta of electromagnetic radiation. They may be emitted from the de-excitation of nuclei or created in very energetic interactions of elementary particles with matter.

Hadrons: Generic name for those elementary particles that are carriers of the ''strong interaction'' (e.g., protons, neutrons, and pions).

Leptons: Generic name for those elementary particles that do not participate in the strong interaction (e.g., electrons, muons, and their corresponding neutrinos).

Nucleon: Generic name for neutrons and protons, the constituents of the atomic nucleus.

Particle Detectors: Instruments to detect or measure properties of particles or radiation quanta. Here, particles designate the microscopic constituents of matter or elementary particles created in energetic collisions of matter. Typically, the trajectory, velocity, energy, or mass of such particles is detected.

Particle Physics: Branch of physics devoted to the study of elementary particles and forces between them. Frequently, very energetic accelerators are used to create or study such particles; this field is therefore often also called *high-energy physics*.

Scintillator: A group of materials that are widely used for particle detection; the material is excited by the passage of charged particles and emits photons (in the UV to visible range) in the de-excitation process. It has important applications outside particle physics.

Solid-State Detectors: Common class of particle detectors. Most frequently used are detectors based on very pure silicon, germanium, or gallium arsenide. The passage of a charged particle creates free charge carriers (''electron–hole'' pairs), which may be collected on suitably arranged electrodes. They are used for the precision measurement of the energy loss or the position of charged particles.

References

(ATLAS Collaboration) (**1996**) CERN/LHCC/96-41.

(Auger Collaboration) (**2007**) *Science*, **318**, 938.

Booth, N.E., *et al.* (**1996**) *Annu. Rev. Nucl. Part. Sci.*, **46**, 471.

Cabrera, B., *et al.* (**2000**) *Physica*, **B 280**, 509.

(CMS Collaboration) (**1997**) CERN/LHCC/97-33.

Cronin, J.W., *et al.* (**1993**) *Annu. Rev. Nucl. Part. Sci.*, **43**, 883.

Doke, T. (**2005**) *Nucl. Instrum. Methods*, **B 234**, 203.

Enns, Ch. (**2005**) *Cryogenic Particle Detection*, Springer.

Fabjan, C.W. and Ludlam, T. (**1982**) *Annu. Rev. Nucl. Part. Sci.*, **32**, 335.

Fabjan, C.W. and Gianotti, F. (**2003**) *Rev. Mod. Phys.*, **75**, 1243.

Fanti, V., *et al.* (**2007**) *Nucl. Instrum. Methods*, **A574**, 433.

Fernow, R.C. (**1986**) *Introduction to Experimental Particle Physics*, Cambridge University Press, Cambridge.

Ferrari A. (**2001**), private communication.

Hoffman, C.M., *et al.* (**1999**) *Rev. Mod. Phys.*, **71**, 897.

Learned, J.G. and Mannheim, K. (**2000**) *Annu. Rev. Nucl. Part. Sci.*, **50**, 679.

Lecoq, P., *et al.* (**2006**) *Inorganic Scintillators for Detector Systems*, Springer.

Moses, W.W. (**2002**) *Nucl. Instrum. Methods*, **A487**, 123.

PDG (Particle Data Group), (**2006**), W.M. Yao *et al.*, *Journal od Physics* **G33**, 1

Peacock, A. (**1999**) *Physica*, **B268**, 595.

Pretzl, K. (**2005**) *J. Phys.*, **G31**, 133.

Sawada, R. (**2007**) *Nucl. Instrum. Methods*, **A581**, 522.

Wigmans, R. (**1991**) *Annu. Rev. Nucl. Part. Sci.*, **41**, 133.

Wigmans, R. (**2000**) *Calorimetry*, Oxford, Clarendon Press.

Further Reading

Froidevaux, D. and Sphicas, P. (**2006**) *Annu. Rev. Nucl. Part. Sci.*, **56**, 378.

Grupen, C. (**1996**) *Particle Detectors*, Cambridge University Press.

Radeka, V. (**1988**) *Annu. Rev. Nucl. Part. Sci.*, **38**, 217.

DETECTORS, SCINTILLATION

ERIC B. NORMAN, *Nuclear Science Division, Lawrence Berkeley Laboratory, Berkeley, California, U.S.A.*

Introduction 503
1. **Physics of Scintillation Detectors** 503
2. **Manufacture of Scintillation Detectors** 505
3. **Applications of Scintillation Detectors** 506
4. **Conclusions** 511
Works Cited 511
Further Reading 512

INTRODUCTION

In many areas of basic scientific research, as well as in a variety of applied fields, one wishes to detect the passage of radiation through matter. The fact that some materials emit light when struck by radiation has been known for many years, and this principle is the basis for the scintillation detector. For example, in 1911, Ernest Rutherford observed scintillation light produced by the interactions of α particles with a zinc sulfide screen in his pioneering scattering experiments which established the existence of the atomic nucleus. Since then, a large number of different scintillating materials have been developed for specific applications in the fields of physics, chemistry, and biology. This article describes the scintillation process and some of the uses of scintillation detectors in the laboratory and in industry.

Before describing the scintillation mechanism, it is useful to review briefly the processes by which radiation is absorbed by matter. For x rays and γ rays (high-energy photons) three processes are important: the photoelectric effect, Compton scattering, and pair production. In the photoelectric effect, the full energy of the photon is absorbed by a bound atomic electron. This ejects the electron from the atom with a kinetic energy equal to the photon energy minus the electron's original binding energy. In Compton scattering, the photon also interacts with an atomic electron, but only a fraction of its energy is transferred to the electron. The remaining energy is carried off by a secondary photon in an amount that depends on the angle of its emission relative to the direction of the primary photon. For photons with energies greater than twice the rest-mass energy of the electron ($2m_ec^2=1022$ keV) it is possible to convert into an electron-proton pair. The kinetic energy of the e^+e^- pair is then equal to the original photon energy minus 1022 keV.

When β particles (electrons) traverse matter, they produce atomic and molecular excitations and ionization through their collisions with atomic electrons. Heavy charged particles, such as protons, α particles (^{4}He nuclei), and nuclei of heavier atoms, also make inelastic collisions with atomic electrons that lead to atomic and molecular excitations as well as ionization. For neutrons, however, the situation is somewhat different. Since neutrons are electrically neutral, the primary way in which they lose energy in matter is through their nuclear interactions with atomic nuclei. They can be scattered off the nucleus or absorbed by it. In hydrogenous materials such as paraffin or polyethylene, neutrons scattering off hydrogen nuclei produce recoiling protons which then interact as described above. Some scintillation materials are doped with ^{6}Li, which has a large cross section for the $^6\mathrm{Li}(n,^3\mathrm{H})^4\mathrm{He}$ reaction. The recoiling triton and α particle then interact as described above.

1. PHYSICS OF SCINTILLATION DETECTORS

The mechanism by which energy deposited by radiation is converted into light within a scintillator is a fairly involved process and is

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

somewhat different in inorganic and organic scintillators. In both types, however, the amount of light produced is proportional to the amount of energy deposited in the scintillator. Inorganic scintillators are made of crystals of materials such as NaI or CsI that are usually activated with an impurity such as Tl (thallium). It has been found that with an impurity concentration on the order of 1 part in 10^3 by number, there is a very efficient transfer of energy from the major constituent to the impurity. In NaI(Tl), for example, the result is that the capture of electrons and/or holes by the thallium impurity leads to excited states of the Tl^+ ion that decay by emitting the light which is observed from this scintillator. Through this process, approximately 12% of the incident energy is converted into Tl^+ light. Because the impurity concentration is so low, this type of scintillator is very transparent to its own light.

Liquid, plastic, and some kinds of crystal scintillators are composed of organic compounds known as aromatic hydrocarbons. These materials consist of molecules containing a benzene ring structure. Within these rings, not all of the electrons that bind the atoms together are strongly localized to individual atoms. These so-called valence electrons occupy the π molecular orbitals which extend above and below the plane of the molecule. Absorption of radiation by these electrons can produce excited states which decay to the ground state or other excited states by emitting light. If the decay proceeds to an excited state, then the photon that is emitted has too little energy to excite another molecule from its ground state to the appropriate excited state. Thus, such transitions produce light to which the scintillator is quite transparent.

Over the years, many different types of scintillators have been developed to meet a variety of radiation-detection needs. The physical properties of a number of different scintillation materials are listed in Table 1. Physical properties such as density, scintillation efficiency (i.e., a measure of the fraction of deposited energy that is converted into light), the wavelength at which the scintillation light appears, and the scintillation decay constant (i.e., the time interval over which light is emitted) vary widely among the different scintillators. These differences are used to decide which scintillator is used for a particular application. High-density materials containing elements with large atomic numbers have relatively high efficiencies for stopping γ rays. Thus, compounds such as NaI(Tl), BaF_2, or bismuth germanate (BGO) are used in such applications. Plastic scintillators have very short decay constants and thus are used in applications where there are needs for fast timing. Scintillators such as ^{6}LiI(Eu) contain nuclei that have high cross sections for neutron-induced reactions and thus are useful in neutron studies.

Table 1. Physical properties of various scintillator materials.

Material	Physical form	Wavelength of maximum emission (nm)	Decay constant (ns)	Density (g/cm^3)	Scintillation efficiency[a] (%)
NaI(Tl)	Crystal	415	230	3.67	100
CaF_2(Eu)	Crystal	435	940	3.18	50
CsI(Na)	Crystal	420	630	4.51	85
CsI(Tl)	Crystal	565	1000	4.51	45
BaF_2	Crystal	325	630	4.88	20
$Bi_4Ge_3O_{12}$	Crystal	480	300	7.13	12
^{6}LiI(Eu)	Crystal	480	1400	4.08	35
Anthracene	Crystal	447	30	1.25	43
NE102A	Plastic	423	2.4	1.032	28
NE115	Plastic	395	320	1.032	10
Pilot U	Plastic	425	1.36	1.032	29
NE213	Liquid	425	3.7	0.874	34

---

[a] Relative to NaI(Tl).

The process by which a complete scintillation detector system converts the energy deposited by the incident radiation into a usable signal involves several steps. Such a system is shown schematically in Fig. 1. As described above, the incident radiation ionizes and excites the scintillation material. Within the scintillator, a fraction of this energy is converted into light. This light travels through the scintillator and is then collected on a photosensitive surface. The resulting photoelectrons are multiplied and collected, thus producing a usable signal. Usually these functions are accomplished by coupling a scintillation material to a light pipe which guides the light to a photomultiplier tube. When photons strike the photocathode, electrons are ejected via the photoelectric effect. These electrons are then accelerated and focused onto a series of dynodes, where through collisions the number of electrons is greatly multiplied. By the time the signal reaches the collection anode, it may have been amplified by more than a factor of 10^6. This signal is usually then sent to an amplifier where further shaping and amplification of the pulse is done before it reaches the data-acquisition system. The size of this final signal provides a measure of the amount of energy deposited by the radiation in the scintillator.

An important characteristic of any type of energy-measuring device is its resolution. This is usually expressed as ΔE=the full width at half maximum (FWHM) of the peak observed from a source of monoenergetic particles. In the scintillation-detector system described above, the measure of the energy of the incident radiation is given by the amount of charge that comes out of the photomultiplier. However, as discussed by many authors (e.g., Rolfs and Rodney, 1988), at each step in the scintillation-detection process there are losses which contribute to the overall detector resolution. As discussed above, in NaI(Tl) detectors the efficiency, ϵ_1, of converting the incident energy into light is about 12%. The mean energy w of these photons is about 3 eV. Of these photons, only a fraction ϵ_2, of order 0.4, actually reaches the photocathode. Only another fraction ϵ_3, of order 0.2, causes the emission of a photoelectron. The fractional energy resolution $\Delta E/E$ of such detectors is limited by the statistical fluctuations in each of these processes and is given by the expression

$$\Delta E/E = 2.35(w/E\epsilon_1\epsilon_2\epsilon_3)^{1/2}.$$

This leads to an expected value of $\Delta E/E$ of approximately 4% for a 1.33-MeV γ ray. As will be shown below, in practice one usually observes slightly worse resolution than this because of imperfections in the detector crystal and/or electronics.

2. MANUFACTURE OF SCINTILLATION DETECTORS

Inorganic scintillators such as NaI(Tl) and CsI(Tl) can be produced as single-crystal ingots up to approximately 36 in. in diameter. To produce crystals of this size, grains or small chunks of the desired chemical are put into a large crucible. The crucible is heated to high enough temperature to melt the material. The crucible is then allowed to cool at a rate which is faster on the bottom than at the top.

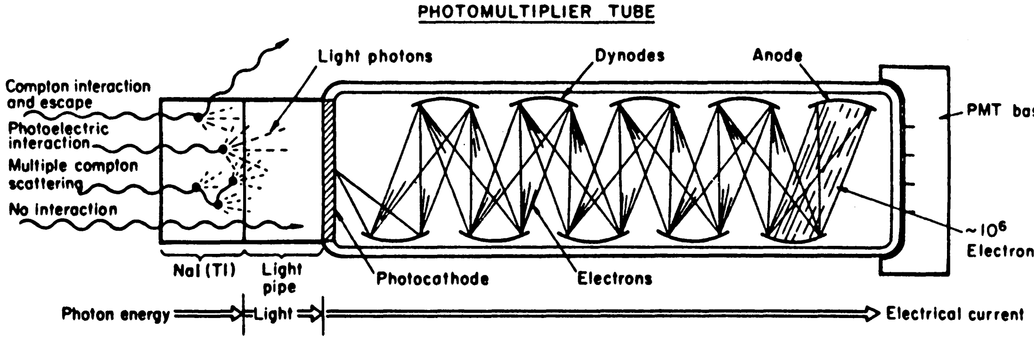

FIG. 1. Schematic view of a scintillation detector system. The interactions of photons in a NaI(Tl) scintillator are shown along with the transmission of the resulting optical photons through a light pipe to a photomultiplier where light is converted into an electrical signal. (Courtesy of S. E. Derenzo.)

Thus crystal growth begins at the bottom and proceeds upwards. In this way large single crystals can be grown. An example of this is shown in Fig. 2, where a large crystal of CsI has just been removed from the crucible in which it was grown. For more information on the production of crystals, see CRYSTAL GROWTH.

Plastic scintillators are usually manufactured in the form of cast sheets. However, these materials are also available in rods, fibers, or beads. They are produced by mixing together a suitable fluorescent organic compound with a polymer and allowing the mixture to solidify. Commonly used mixtures include *p*-terphenyl dissolved in either polystyrene or polyvinyltoluene. A typical plastic scintillator paddle with a light pipe attached is shown in Fig. 3.

Following the machining or forming process, the surfaces of a solid scintillator are polished. A highly reflective coating is usually applied to these surfaces to return the light to the crystal and thus aid in the transport of the scintillation light to the photomultiplier. Crystalline scintillators are normally packaged inside a metal canister so as to protect the scintillator from mechanical shocks. In the case of NaI such protection is also necessary because of the hygroscopic (i.e., moisture sensitive) nature of the material. Liquid scintillators are also often packaged in a metal container. In both cases, the can has a transparent window on one side on which the light pipe and then photomultiplier tube are mounted.

3. APPLICATIONS OF SCINTILLATION DETECTORS

In the areas of nuclear and particle physics research, scintillation detectors are used in a wide variety of experiments. In γ-ray-counting experiments, NaI(Tl) detectors are used in applications where high detection efficiency, good energy resolution, and reasonably good timing properties are needed. An example of low-energy γ rays observed in a small NaI(Tl) detector is shown in Fig. 4. A radioactive source of ^{60}Co, which emits a pair of coinci-

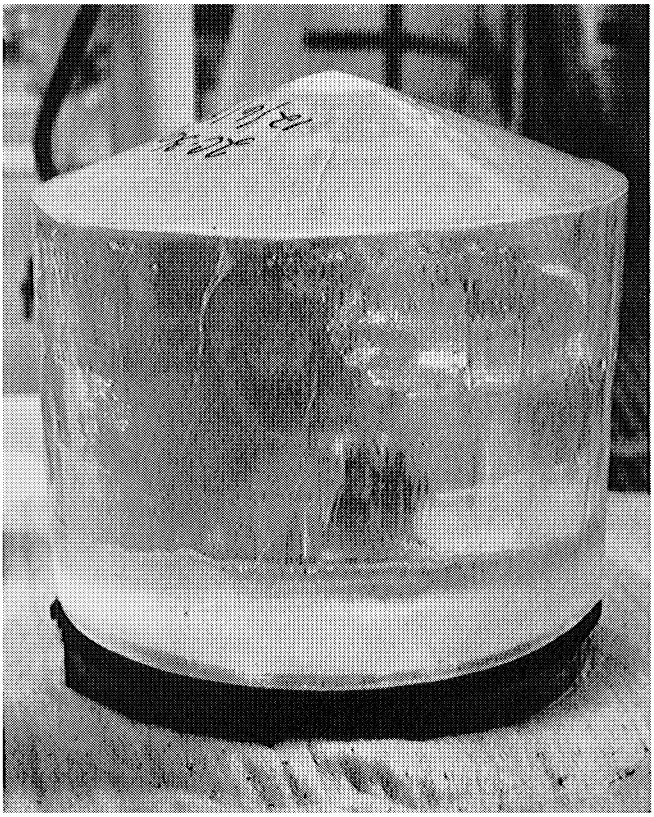

FIG. 2. An ingot of CsI as it emerged from the crucible in which it was grown. (Courtesy of Bicron Corp.)

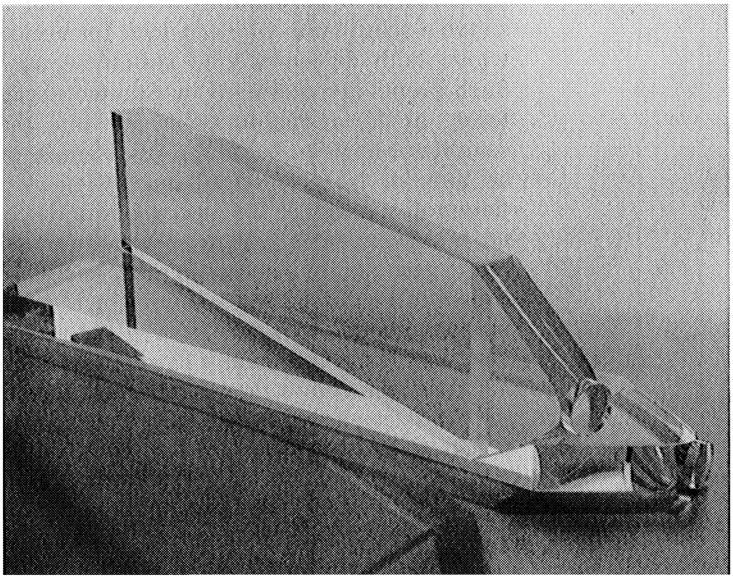

FIG. 3. A plastic scintillator paddle with a light pipe attached to it. (Courtesy of Bicron Corp.)

dent γ rays with energies of 1173 and 1332 keV, was counted with a 7.6-cm-diam × 7.6-cm-long NaI(Tl) detector. The full-energy (photo)peaks of these lines are clearly seen along with a well-defined "sum peak" which is produced when the full energies of both γ rays are absorbed at the same time by the detector. The counts in the region below the photopeaks and in the region between the photopeaks and the sum peak are due to events in which one or both of the γ rays underwent Compton scattering in the crystal or in its surroundings. The energy resolution for the photopeaks is just over 5% FWHM. In contrast, solid-state germanium detectors (described in the article DETECTORS, SEMICONDUCTOR) typically achieve resolutions of approximately 0.15% at this energy.

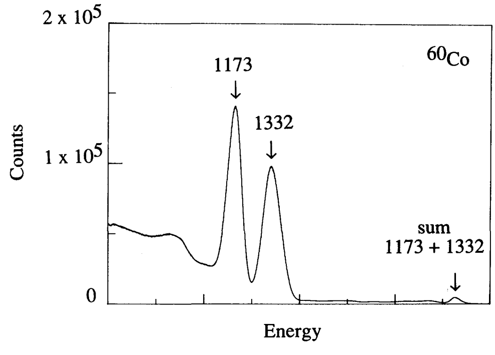

FIG. 4. γ-ray spectrum of ^{60}Co observed in a 7.6-cm-diam × 7.6-cm-long NaI(Tl) detector.

Another use of NaI(Tl) detectors is in nuclear reaction studies. In radiative capture reactions, the projectile fuses with the target nucleus and the resulting excitation energy is carried off by relatively high-energy ($\gtrsim 10$ MeV) γ rays. While solid-state germanium detectors would provide better energy resolution, their detection efficiency for such high-energy γ rays is too low for them to be useful. Therefore, in this type of research, NaI(Tl) detectors are often used. An example of a spectrum observed in the capture of protons by a target of ^{30}Si is shown in Fig. 5. The NaI(Tl) crystal used in this experiment is 25 cm in diameter and 28 cm thick. The proton bombarding energy is 25.5 MeV. The small peak near 32-MeV γ-ray energy corresponds to the case of leaving the resulting ^{31}P nucleus in its ground state; other peaks at lower γ-ray energy correspond to leaving the ^{31}P in one of its excited states.

Another application of scintillation detectors is in the area of Compton suppression. In a typical germanium detector, if a 1-MeV γ ray enters the detector, only about 20% of the recorded events will be in the full-energy peak (photopeak) and 80% will be in a continuous spectrum from zero up to the full γ-ray energy as a result of Compton scattering. These Comp-

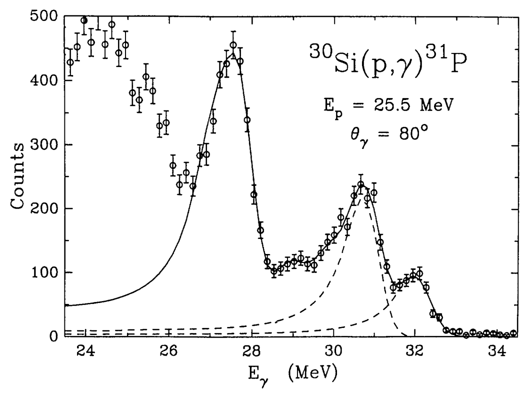

FIG. 5. γ-ray spectrum observed in a 25-cm-diam × 28-cm-long NaI(Tl) detector from the ^{30}Si(p,γ)^{31}P reaction at a proton bombarding energy of 25.5 MeV. (Courtesy of G. Feldman.)

ton events provide no useful information as to the original energy of the incident γ ray. Furthermore, they produce a background which can make it difficult to observe the often small full-energy peaks of interest. By surrounding a germanium detector with a suitable scintillator, one can look for events where both detectors fired simultaneously. Such events are produced by Compton scattering in the germanium detector into the scintillator and can be vetoed. An example of an annular NaI(Tl) scintillator acting as a Compton suppressor for a germanium detector is shown in Fig. 6. γ-ray spectra observed in the germanium detector in this system are shown in Fig. 7. A ^{54}Mn source, which emits monoenergetic γ rays at 835 keV, was counted with a 110-cm^3 germanium detector. The upper spectrum shows the response of the germanium detector without regard to what happened at the same time in the surrounding NaI(Tl) detector. One sees the 835-keV full-energy peak along with a Compton continuum. The lower spectrum is what is observed in the germanium detector when the NaI(Tl) did not observe an event in coincidence with the germanium detector. One can clearly see a reduction in the Compton continuum by a factor as large as 10 in some parts of the

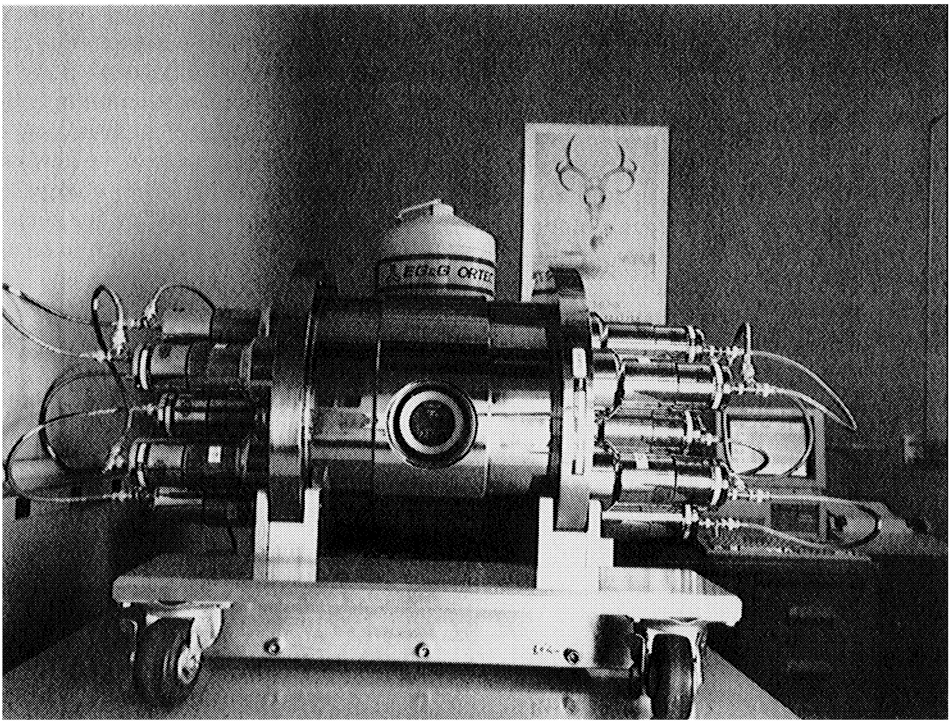

FIG. 6. A 30-cm-diameter × 30-cm-long annular NaI(Tl) detector system surrounding a 110-cm^3 germanium detector. The NaI(Tl) crystal is optically divided into two halves, each viewed by seven photomultiplier tubes. In this system, the NaI(Tl) detector is used to veto events in which a γ ray undergoes Compton scattering in the germanium detector and the scattered photon goes into the NaI(Tl) detector.

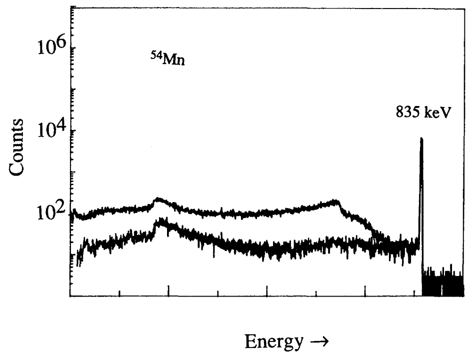

FIG. 7. γ-ray spectra observed in a 110-cm^3 germanium detector placed inside the NaI(Tl) annulus shown in Fig. 6. Data with (lower spectrum) and without (upper spectrum) the NaI(Tl) anticoincidence requirement are shown to illustrate the reduction of the continuum produced by Compton scattering.

spectrum. In a number of multielement germanium detector arrays, such as the High Energy Resolution Array (HERA) at Lawrence Berkeley Laboratory, TESSA at Daresbury Laboratory, and the soon-to-be-built GAMMASPHERE and EUROGAM arrays, BGO anti-Compton shields are used instead of NaI(Tl). The reason for this is that the large atomic number of Bi and the high density of BGO mean that a much smaller-sized scintillator is needed to provide the same amount of Compton suppression. This permits the germanium detectors to be packed together closely, and thus provides high efficiency for detecting the full-energy γ ray in a germanium detector.

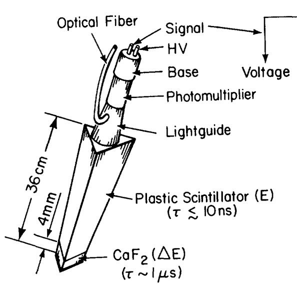

FIG. 8. Schematic view of a phoswich detector. Two scintillators with different decay constants are stacked one behind the other and are both viewed by a single photomultiplier tube. The resulting light signal then has two components which can be separated with appropriate electronics. (Baden *et al.*, 1982.)

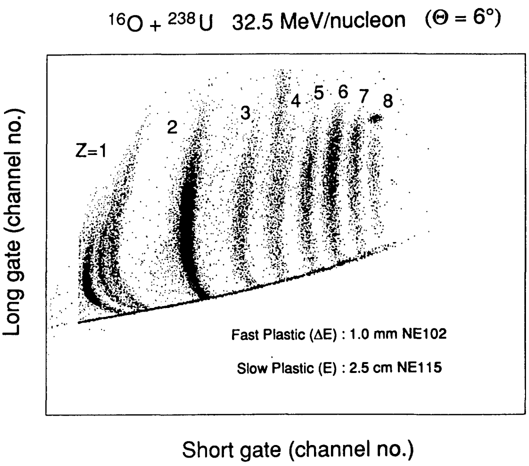

FIG. 9. An example of data obtained from an array of plastic phoswich detectors. By performing a fast integration of the photomultiplier signal, one obtains a measure of the light output from the thin front scintillator. A longer integration provides a measure of the energy deposited in the thicker and slower back scintillator. (Courtesy of Y. D. Chan.)

As can be seen in Table 1, the time dependence of the light output of scintillators varies widely. This variation has been put to practical use in what is called a phoswich detector. This consists of two different scintillators stacked one behind the other, both of which are viewed by one photomultiplier tube. An example of this type of detector is shown in Fig. 8. This detector consists of a thin crystal of CaF_2 mounted on the end of a long plastic scintillator. A charged particle which enters the CaF_2 and comes to rest in the plastic scintillator will produce scintillation light in both materials. Because of the difference in the decay time of the light output in these two materials, the photomultiplier output signal will have two components. There will be a fast component associated with the plastic scintillator and a slower one associated with the CaF_2. By appropriately integrating these signals, one can extract the individual signals from these two scintillators and obtain a measurement of both the energy lost in the front scintillator, ΔE, and the total energy of the charged particle, E. An example of the data resulting from a phoswich detector system is shown in Fig. 9. In this case the products of the fragmentation of an ^{16}O beam by a ^{238}U target were observed by an array of phoswich detectors consisting of two different types of plastic scintillator—one with a short decay time and one with a long decay time. From the "short" and "long" gates set on the photomultiplier output signals, one obtains

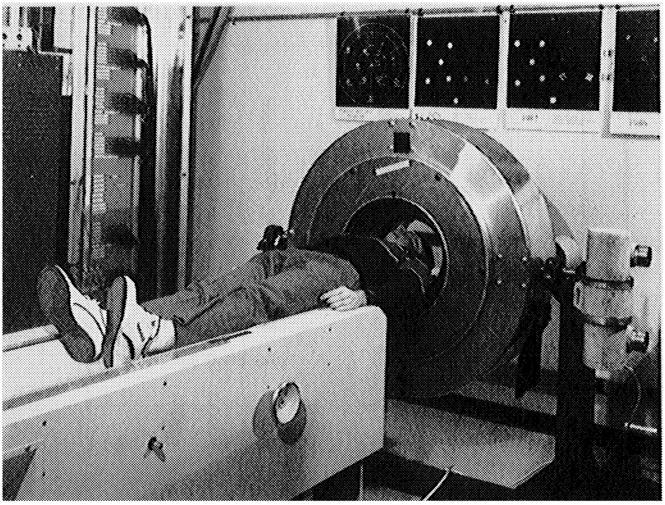

FIG. 10. A patient being examined with a positron-emission tomography system that consists of 280 NaI(Tl) crystals. (Courtesy of S. E. Derenzo.)

the ΔE and E parameters, respectively. The fact that, for a given kinetic energy, nuclei with different atomic numbers (Z) lose differing amounts of energy in the thin front scintillator gives rise to the various bands seen in this two-dimensional plot.

Scintillation detectors have also played a major role in astrophysical studies performed in space. For example, NaI(Tl) detectors on the Solar Maximum Mission detected nuclear γ rays at 847 and 1238 keV produced by the radioactive decay of ^{56}Co (half-life=77 days) coming from Supernova 1987A. This observation helped to confirm that the evolution of massive stars produces a core of iron which collapses and initiates the supernova explosion. This process had been suggested long ago, but the observation of these characteristic γ rays is one of the strongest pieces of experimental evidence that this type of nucleosynthesis actually occurs in supernovae.

In the field of medical physics, scintillation detectors have played major roles in a number of areas. Whole-body counters, which are used to determine the identities and quantities of radioisotopes contained in a person, often consist of large NaI(Tl) crystals. When housed in a low-background environment, such systems are capable of detecting very small (~nanocurie) amounts of γ-ray emitting radioisotopes.

Another area in which scintillators are now being widely used is positron-emission tomography, or PET scanning. In this procedure, a positron-emitting radioisotope, such as ^{18}F, is attached to a compound such as fluorodeoxyglucose and is then introduced into the subject's body. It is concentrated in an organ of

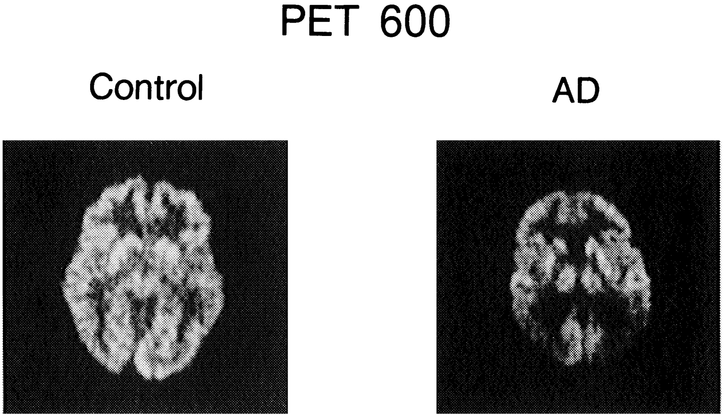

FIG. 11. An example of positron-emission tomography data taken with a PET system consisting of 600 scintillation detectors. Activity of the human brain is indicated by the white areas in the tomograph. A control patient is shown along with a patient suffering from Alzheimer's disease. (Valk *et al.*, 1990.)

interest where the positrons annihilate with electrons. The annihilation process produces two γ rays, each of 511 keV, which are emitted back-to-back. That is, they travel off in opposite directions from the point of annihilation. These annihilation photons are detected in coincidence in two diametrically opposed crystals. By surrounding the subject with a large number of detectors, and by finding the intersection point of all of the lines, one can determine the size and activity of the positron-emitting activity. By observing the time variations in such activity one can observe various body functions in progress. A subject being counted with a PET system of 280 bismuth germanate crystals is shown in Fig. 10. PET data showing activity of the human brain, taken with a system containing 600 individual crystals, are shown in Fig. 11.

In the area of geological exploration, scintillation detectors also play a major role. Measurement of natural radioactivity to locate deposits of uranium and thorium in the earth's crust is an activity in which scintillation detectors are widely used. Low-flying aircraft carrying large NaI(Tl) detectors make surveys of γ-ray activity searching for the characteristic γ rays emitted in the decay chains of ^{232}Th (2.6 MeV) and ^{238}U (1.76 MeV). Borehole measurements are performed in several different ways. Sometimes a scintillation detector is simply lowered down a hole and measurements are made of the local radioactivity. In other applications, sources of either neutrons or γ rays are also placed underground and the scintillation detector is used to measure the amount of scattered or induced activity.

In the area of nuclear safeguards, large plastic scintillators are often used in vehicle or personnel monitoring to look for unauthorized movements of radioactive materials. In such an installation, a vehicle or person drives or walks past the scintillation detector and the characteristic γ rays of ^{235}U or ^{239}Pu are sought for. Walk-through systems are now in place which can detect 0.02 g of ^{239}Pu or 1 g of ^{235}U.

In a number of industrial situations, one is interested in measuring the wearing away of material from an object as a result of its being used. This is the subject of tribology, and scintillation detectors are often used in this field. For example, in the automobile industry, engine wear can be determined by implanting a γ-ray-emitting radioisotope in the part of the engine of interest. A NaI(Tl) or BGO detector is then placed near this part and the γ-ray counting rate is monitored as a function of the engine use. As the part is worn away, one observes a decrease in the γ-ray activity. By taking an identically implanted part and measuring its γ-ray counting rate as one quantitatively removes material from the surface, one can calibrate the technique. Through such tests, it has been found that this method is sensitive to engine part wear as small as 0.00005 in.

Compton scattering of γ rays from an object can be used to determine the thickness and/or distribution of material within the object. This technique is often used in the area of quality control. For example, artillery shells coming off a production line have been irradiated with 662-keV γ rays from a ^{137}Cs source. By measuring the number of scattered γ rays and by comparing the result to that obtained from a standard, one can reject shells that contain voids or other imperfections. This method offers several advantages over more conventional schemes in that it is nondestructive and can be done very quickly.

4. CONCLUSIONS

The conversion of the energy deposited by radiation into light is the basis of the scintillation detector. Many different types of scintillators have been developed to meet a variety of needs in the field of radiation detection. Today scintillation detectors are used extensively in physics, chemistry, medicine, astrophysics, and in a number of industrial applications. With the ongoing development of new and improved scintillating materials, scintillation detectors will continue to play major roles in these fields for the foreseeable future.

List of Works Cited

Baden, A., Gutbrod, H. H., Lohner, H., Maier, M. R., Poskanzer, A. M., Renner, T., Riedesel, H., Ritter, H. G., Spieler, H., Warwick, A., Weik, F., Wieman, H. (1982), *Nucl. Instrum. Methods* **203**, 198–211.

Rolfs, C. E., Rodney, W. S. (1988), *Cauldrons in the Cosmos*, Chicago: Univ. of Chicago Press.

Valk, P. E., Jagust, W. J., Derenzo, S. E., Huesman, R. H., Geyer, A. B., Budinger, T. F. (1990), *Radiology* **176**, 783–790.

Further Reading

Birks, J. B. (1964), *The Theory and Practice of Scintillation Counting,* Oxford: Pergamon.

Brooks, F. D. (1979), *Nucl. Instrum. Methods* **162**, 477–505.

Heath, R. L., Hofstadter, R., and Hughes, E. B. (1979), *Nucl. Instrum. Methods* **162**, 431–476.

Knoll, G. F. (1989), *Radiation Detection and Measurement,* 2nd ed., New York: Wiley.

Leo, W. R. (1987), *Techniques for Nuclear and Particle Physics Experiments,* Berlin: Springer-Verlag.

Neiler, J. H., Bell, P. R. (1965), in: K. Siegbahn (Ed.), *α, β, γ Spectroscopy,* Vol. 1, Amsterdam: North-Holland, p. 245.

22
Silicon Semiconductor Detectors for Particle Imaging

Erik H.M. Heijne

22.1 **Introduction 679**
22.2 **Basics of Silicon Detectors 681**
22.2.1 Device Structure 681
22.2.2 Detector Segmentation and Edges 685
22.2.3 Energy Deposition and Signal Generation in Silicon Detectors 686
22.2.4 Manufacturing of Semiconductor Detectors 689
22.2.5 Degradation of Detectors 691
22.3 **Various Sensor Designs and the Associated Readout Electronics 692**
22.3.1 Custom Design of Chips 693
22.3.2 Imaging Matrix with Linear or 2D Sensor Geometry 695
22.3.3 Factors Affecting Signal/Noise 697
22.4 **Typical Silicon Detector Arrays in a Particle Physics Experiment 698**
22.4.1 Tracking Detectors for Momentum Measurement 699
22.4.2 High-Precision Detectors for Interaction Vertexing 700
22.4.3 Semiconductor Detectors for Particle Energy Measurement 701
22.5 **Other Applications of Semiconductor Detectors 702**
22.5.1 Satellite-Based Experiments 703
22.5.2 Detectors at Synchrotron Light Sources 704
22.5.3 Radiation Imaging Detectors in Industrial or Research Equipment 706
Glossary 706
References 708

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

22.1
Introduction

Radiation, light quanta, as well as energetic particles interact with semiconductor materials by liberating free charge carriers, negative electrons, as well as positive holes. In very pure silicon or germanium crystals, such charge carriers can move with a high mobility and a lifetime of milliseconds. They can move in an electrical field over distances of several millimeters on a nanosecond timescale, and induce electrical signals in the external contacts. Miniaturized electronic circuits then perform the charge signal processing; in this way, it has become possible to study various properties of the incoming radiation.

In the early years, 1950–1970, semiconductor detectors were primarily used as a ‘‘solid-state’’ ionization chamber. They offer the advantages of enhanced stopping power and compactness, compared with magnetic spectrometers or improved precision in the energy measurement compared with other solid-state detectors such as sodium iodide scintillating crystals. This chapter describes the development of semiconductor detector systems starting from about 1980, after a number of fundamental improvements were introduced in detector manufacturing and in systems architecture (Heijne, 2008). In particular, mastering of custom design of analog and mixed-mode integrated signal processing chips in commercial CMOS (Complementary Metal Oxide Semiconductor) technologies proved to be an essential boundary condition for the introduction of large-scale arrays. The properties of the semiconductor materials and the device technology are sufficiently complicated to justify an intensive effort in R&D, with thousands of technically oriented articles already dedicated to the study of these detectors. In exploiting these devices, a multitude of scientific and practical results have been obtained, often linked in a unique way to the innovation in these instruments.

The new applications of semiconductor detectors can be characterized as particle or photon radiation imaging. The detectors now consist of a regular structure of segmented sensor cells, usually diodes, each connected to a dedicated amplifier on an electronics readout chip. The amplifier circuit is matched to the characteristics of the sensor element. The measured positions of the radiation quanta are combined into a 2D or a 3D reconstruction of some object, process, or interaction of elementary particles. These imaging detectors use parallel processing of the signals from simultaneously incident ionizing particles, sometimes even at rates in excess of $10^{10}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. Besides the hybrid detector

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

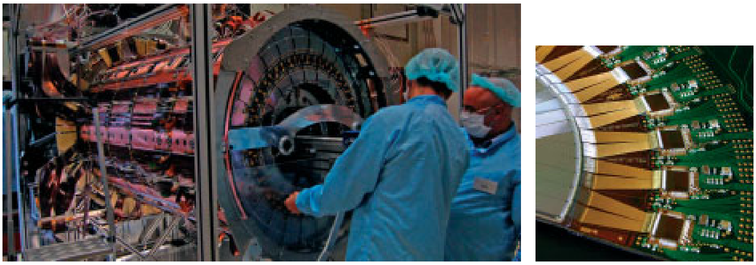
(a) (b)

Fig. 22.1 (a) People are shown assembling the Atlas silicon tracking detector SCT in a clean room at CERN. This detector system consists of a central barrel with several cylindrical layers and perpendicular forward discs on both sides. Fully assembled, it has been placed in the center of the Atlas experiment at the Large Hadron Collider (LHC). The evacuated beam pipe passes along the axis of the cilinder. The energetic proton beams collide around the midpoint of the barrel. Along this longitudinal axis the interactions take place, and particles originating from the vertex point fly in all directions. These will generally cross several planes of sensors and allow at least five space points to be determined. More details are discussed in Section 22.4. (Photo CERN.) (b) Part of a sector of the Atlas forward silicon detector. This module has been assembled at the NIKHEF particle physics laboratory, in Amsterdam, using boards tested in several labs around the world, and equipped with Si detectors that were manufactured by industrial suppliers, readout chips custom designed in a radhard, 0.25 µm CMOS technology, and components for powering, alignment and cooling. (Photo NIKHEF.)

systems, in which the sensors and the readout electronics are still distinct components, monolithic detectors in which all functions are integrated in the same piece of silicon are also being developed. In the hybrid approach, one can optimize functions in the separate parts, at the cost of complexity in interconnections, whereas the advantages of full integration come with loss of flexibility in the electronics design and manufacturing. However, the fully integrated devices profit from the rapid advances in silicon imagers for consumer applications.

In outer space and in elementary particle physics experiments, several large arrays have been constructed with a silicon sensitive area from approximately a few meters squared to over 200 m^2, consisting of thousands of modules with fairly large Si sensor chips and Si CMOS electronic readout chips. Such a large system is illustrated in Figure 22.1(a); Figure 22.1(b) provides a close-up of a detector module.

The turning point in elementary particle physics for abandoning photographic methods in favor of fully electronic experiments was around 1985. The ultimate, very large bubble chambers have been decommissioned and replaced by instruments that feature combinations of silicon vertex tracking detectors, gas-filled time-projection chambers, and heavy calorimeters that measure total particle energy. Large improvements in rate capability and in spatial precision have been achieved with this new equipment. A major driving force for the development and building of the large silicon detector arrays has been the need for improved precision in track reconstruction in physics experiments and the identification of short-lived charm and beauty particles. Applications in particle physics are discussed in some detail in Section 22.4.

A few other applications are mentioned in Section 22.5, such as space-based research and materials analysis with synchrotron light sources. Section 22.2 describes basic aspects of the semiconductor devices and technology and sensor designs are enumerated in Section 22.3.

The textbook *Semiconductor Detectors* by Bertolini and Coche (1968) describes principles and the developments until about 1965. A general, comprehensive book on detectors is *Radiation Detection and Measurement* by Knoll (2000). Recent books on silicon detectors have been written by Lutz (1999) and by Spieler (2005).

22.2
Basics of Silicon Detectors

The physics of semiconductor devices is described, for example, in the famous textbooks of Grove (1967) or Sze *et al.* (2007). Device properties are created by local modifications of electrical charge carrier concentrations at micrometer and recently even at nanometer scale, in an initially homogeneous piece of semiconductor. The extreme purity achieved in manufacturing silicon crystals, which can be less than one electrically active impurity atom on 10^{14}, has allowed reliable and reproducible silicon devices to be used for countless applications.

22.2.1
Device Structure

The p–n junction diode, treated by Sze (1981) in Chapter 3 of his book (Chapter 2 in the third edition by Sze and Ng, 2007), is the basic building block for a silicon detector. A schematic drawing is shown in Figure 22.2. The rectifying property of the diode is needed to limit the current in the sensing element, which at room temperature would flow when an electrical field is applied between ohmic contacts on Si. For many years, the metal-semiconductor Schottky diode gave the best achievable detector performance, but the recent imaging

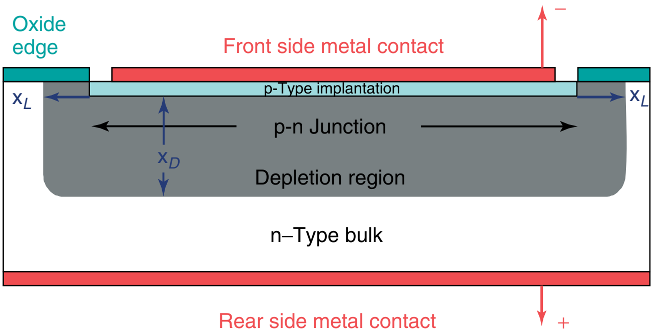

Fig. 22.2 Schematic drawing of a silicon p–n junction diode. The rectifying p–n junction can be obtained by controlled introduction of a p-type impurity such as ion implantation of boron into the bulk n-type silicon. The p–n diode junction area is defined by the opening that is etched in the surface oxide layer. The metal contacts usually are made with aluminum. The diode is reverse biased by applying a potential between the + and − contacts. For high resistivity n-type silicon, this is usually a voltage between 20 and 500 V. A depletion region extends a distance x_D into the n-type bulk as well as some lateral distance x_L under the surrounding oxide edge.

detector systems employ semiconductor–semiconductor p–n junctions obtained by local implantation of doping ions with a keV beam. In both approaches, a depletion layer is formed in which practically no mobile charge carriers exist, so that no current flows unless charges are created by incoming radiation. The concentration of free charge carriers in a semiconductor depends on the bandgap, the temperature, doping with impurity atoms, and crystal defects. A rectifying junction is not needed for semiconductor materials with a large bandgap, such as diamond, as long as the contacts are noninjecting.

Some unusual aspects of the diode structure for semiconductor nuclear particle detectors have to be briefly mentioned here. First of all, whereas most other devices use a surface layer of a few micrometers at most, a *thick sensitive volume* is needed for many of the detector applications and, ideally, the full wafer thickness of several hundred micrometers is exploited. The reasons for doing so may be to slow down and stop the incoming ionizing particle so that its total energy can be determined; a high efficiency may be needed for conversion of X-ray or γ-ray photons; for traversing, swift particles a sufficiently large signal charge is needed in comparison with noise fluctuations. For traversing particles, the signal is proportional to the energy deposition along the path length in the sensitive thickness of the device. Characteristics for energy loss are described later in this section. A second, related aspect is that the *rear-side contact* of the thick device wafer needs special attention, contrary to the usual situation for electronics chips, where the wafer acts simply as a mechanical support and grounded shield. For optimal functioning, only a *small reverse current* in the reverse-biased diode can be tolerated, to detect the randomly arriving, nonrepetitive signal charge above the noise background. The undesirable reverse diode current is increased by crystal or surface impurities that act as generation-recombination centers for free carriers, by injection from contacts, and by ambient factors such as temperature or light. Under optimal conditions, the generation current I_G is the main component, proportional to the sensitive volume: thickness w times area A, and is determined by the intrinsic equilibrium carrier concentration n_i and inversely proportional to the minority carrier lifetime τ:

$$I_{\mathrm{G}} = \left(\frac{n_i}{2\tau}\right) qwA \tag{1}$$

The electron unit charge is q ($q = 1.602 \times 10^{-19}$ C). In pure Si, $n_i = 10^{10}\,\mathrm{cm}^{-3}$ and with the years, the measured values tend to become a little lower as better silicon purity is achieved. By introducing n-type or p-type doping impurity atoms, one can increase the majority carrier concentration n or p over a large range, even close to $10^{20}\,\mathrm{cm}^{-3}$. The product of n and p remains a constant, equal to the value of n_i at the given temperature. The value of the intrinsic carrier concentration is a function of the semiconductor bandgap E_g, the absolute temperature T, and the effective densities of state in the valence band N_V and the conduction band N_C (see Table 22.1).

$$n_i{}^2 = np = N_C N_V \exp\left(\frac{-E_g}{kT}\right) \tag{2}$$

The bandgap E_g in silicon is 1.12 eV at room temperature. When the temperature is increased, the concentrations of electrons n and of holes p increase both and eventually become equal, at which temperature the material is again intrinsic and highly conductive. As the total number of charge carriers in a piece of semiconductor is

Tab. 22.1 Properties of silicon.

Property	**Value**
Atomic number	14
Atomic weight	28.09
Stable isotopes, natural abundance	28, 29, and 30 (92.21%, 4.70%, and 3.09%)
Density (300 K) g cm^{-2}	2.328
Atoms cm^{-3}	5.0×10^{22}
Crystal structure	Diamond type, space group Fd3 m
Lattice constant (300 K) cm	5.43×10^{-8}
Melting point °C	1410
Linear thermal expansion coefficient $°\text{C}^{-1}$	2.6×10^{-6}
Specific heat (25 °C) J $\text{kg}^{-1}\,°\text{C}^{-1}$	760
Thermal conductivity (25 °C) J $\text{m}^{-1}\,°\text{C}^{-1}$	148
(77 K)	2000
Dielectric constant	11.9
Permittivity Fcm^{-1}	1.054×10^{-12}
Bandgap (300 K) eV	1.12
(0 K) eV	1.16
Intrinsic carrier density (300 K) cm^{-3}	1×10^{10}
Intrinsic resistivity (300 K) Ωcm	3.2×10^{5}
Electron diffusion coefficient (300 K) $\text{cm}^2\ \text{s}^{-1}$	35
Hole diffusion coefficient (300 K) $\text{cm}^2\ \text{s}^{-1}$	12
Electron mobility (300 K) $\text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$	1500
Electron mobility (77 K) $\text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$	30 000 (approximate)
Hole mobility (300 K) $\text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$	600
Hole mobility (77 K) $\text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$	20 000 (approximate)
Carrier saturation velocity (300 K) cm s^{-1}	1×10^{7}
Breakdown field V cm^{-1}	$\sim 5 \times 10^{5}$
Electron affinity V	4.05
Work function V	4.8
Effective density of states N_C in conduction band (300 K) cm^{-3}	3.2×10^{19}
idem N_V in valence band (300 K) cm^{-3}	1.8×10^{19}
Effective mass m^*/m_0	$m^*_\ell = 0.98\ \ m^*_t = 0.19$
Electrons	$m^*_{lh} = 0.16\ \ m^*_{hh} = 0.49$
Holes	
Raman phonon energy eV	0.063
Vapor pressure (930 °C) torr	10^{-8}
Energy required per electron–hole pair	
(300 K) eV	3.62
(77 K) eV	3.76
dE/dx m.i.p. MeV $\text{g}^{-1}\,\text{cm}^2$	1.66
Nuclear collision length g cm^{-2}	69
Si cm	29.64
Radiation length g cm^{-2}	21.8
Si cm	9.36

$n + p$, the minimum number of carriers is obtained when n and p are nearly equal, resulting in a low conductivity σ or reciprocally a high resistivity ρ. The resistivity (unit ohm centimeter – Ωcm) is also a function of the respective carrier mobilities μ_e and μ_h (unit centimeters squared per volt second – $\text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$)

$$\rho = \frac{1}{\sigma} = \frac{1}{q(\mu_e n + \mu_h p)} \tag{3}$$

and for detectors, the values used are often in excess of 1 kΩcm, corresponding to $10^{11}\ \text{cm}^{-3}$ electrons in n-type Si. With such a low carrier concentration one can achieve low reverse current, and also obtain a thick depletion region x_D as shown in Figure 22.2. When the diode is heavily doped at the rectifying contact, the extension of the depletion layer is toward the lowly doped side of the diode, because of the equilibrium between the amounts of charge at either side of the p–n junction, as described by the Poisson equation (see Sze *et al.*, 2007). For a p^+n junction[1] one obtains

$$\begin{aligned} x_D &= \sqrt{\frac{2\varepsilon}{qn}(V_0 + V_B)} \\ &= \sqrt{2\varepsilon\mu_e\rho(V_0 + V_B)} \end{aligned} \tag{4}$$

where ε is the permittivity (for Si $1.054\ 10^{-12}\ \text{F cm}^{-1}$), V_B is the externally applied diode reverse bias voltage, and V_0 is the built-in voltage of the unbiased junction. A typically applied voltage is in the range 20–500 V, which can achieve several 100 μm of depletion layer thickness in kilo ohm centimeters Si. An essential feature is the *lateral extension* x_L of the depletion region under the surface oxide. The distance x_L might be equal to x_D, but can also be different, depending on the fixed charge that is present at the interface or in the oxide itself. As discussed later, the lateral extension allows the complete volume to be depleted in detectors with segmented diode contacts.

Such a diode detector presents, proportional to the surface area A of the device, a lower electrical capacitance C_D, the thicker the depletion region x_D.

$$C_D = \frac{\varepsilon A}{x_D} = A\sqrt{\frac{\varepsilon}{2\mu_e\rho(V_0 + V_B)}} \tag{5}$$

The capacitance of a detector is an important factor in the noise performance for small signals, which is discussed in Section 22.3.

A different situation arises in transistors in deep submicron CMOS technology, where the miniaturization implies very shallow junctions. These require a concentration approaching $10^{20}\ \text{cm}^{-3}$, which brings a large capacitance per unit of area. Yet, because of the nanovolume of such devices, the actual values in a device remain small and only ~1000 unit charges q may determine the function of a transistor. Fluctuations on such small numbers cause a significant spread in characteristics from one device to another. In heavily doped Si, the conductivity is not far from that caused by the $\sim 10^{23}$ free electrons cm^{-3} in metallic materials. The minority carrier concentration then becomes lower than 10 to $100\ \text{cm}^{-3}$ because under steady-state conditions the product *np* remains constant as mentioned above. In fact, this can be used to prevent minority carrier injection from an ohmic metal contact by creating such a heavily doped semiconductor layer between the metal and the lowly doped bulk part of the device.

---

[1] Adding + to the carrier concentration indicates which is the more highly doped side, it does not indicate the sign of the carriers.

Table 22.1 provides values for a number of properties of pure crystalline silicon. Some of these have a determining impact on detector characteristics, as mentioned in the text.

22.2.2
Detector Segmentation and Edges

An array of adjacent diodes, each with a separate contact, on a single piece of high-resistivity silicon presents several interesting features for use in physics experiments or radiation imaging applications. The positions of several simultaneously incident particles can be determined simply with reference to the positions of the responding cells in the array. The segmentation, moreover, leads to miniaturization of the elements, and this reduces their capacitance and noise, and also their leakage currents, even if the current per centimeter squared is high. With the use of high resistivity and a reverse bias voltage that achieves full depletion of the wafer thickness, one can avoid insensitive regions in the silicon volume, between the segments, except around the outer edges of the sensor chip. Currently, special techniques are being developed to increase the sensitive volume even quite close to the edges. This is important for elimination of dead areas in applications with multisensor arrays and when detection is needed of interaction products in proximity to a high-intensity beam, which itself should not pass through the detector.

By around 1960, the first segmented surface barrier silicon detectors had been made. However, the more recent development of custom-integrated circuits for signal processing and readout was needed to allow a large-scale use of such devices. It should be obvious that the characteristics and possibilities of the signal-processing chips basically determine the possibilities for different geometries and properties of the sensor matrix. The segments have to be connected to the signal amplifiers and further information processing circuits. Most often the rectifying side of the sensor is segmented, because the electrical field there is the strongest, but it is also possible to segment the side with the ohmic contact, when some precautions are taken. Finally, ‘‘double-sided’’ detectors can be made, in which both sides are segmented.

The sensor-manufacturing process imposes boundary conditions, such as needs for edge protection, maximum wafer size, and minimum achievable dark current. Also the irradiation of the sensors imposes requirements for the sensor architecture. Various designs of segmented sensors and the associated readout circuits are described and illustrated in Sections 22.3 and 22.4.

Segmentation of silicon detectors relies to a large extent on the unique properties of the silicon dioxide surface layer, which is created during very clean oxidation of a silicon wafer. Understanding the physics of the metal-oxide-semiconductor (MOS) system is useful for the design of segmented devices. The interface between the oxide and the silicon acts as one plate of the capacitor structure, and on the other side is a metal or polysilicon plate on top of the oxide layer. The classical book *MOS Physics and Technology* by Nicollian and Brews (1982) treats this in detail. The schematic, simplified drawing (Figure 22.3) for the case of n-type Si shows the accumulation, flat-band, depletion, and inversion states, depending on the external voltage applied to the top capacitor plate. When the swing of the external voltage is applied in a pulsed mode, the behavior depends on the pulse frequency, and above a megahertz the buildup of the inversion

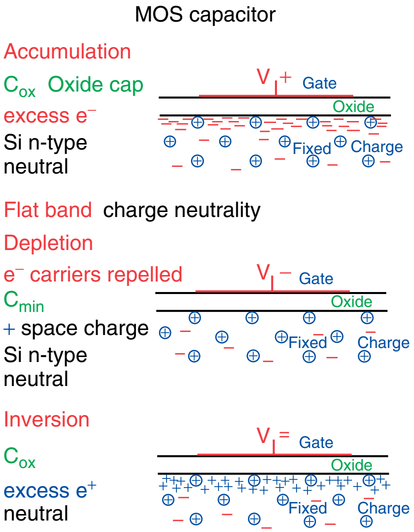

Fig. 22.3 Schematic drawing of the MOS capacitor structure on n-type Si with capacitance C_{ox}. in four different biasing conditions. From top to bottom are sketched the charge and carrier distributions at the oxide–silicon interface layer, for accumulation, flat band (no sketch), depletion, and inversion. Fixed unit charges are encircled, the mobile charges are not. The capacitor plate "gate" on top of the oxide can be a doped, conducting polysilicon or an aluminum metallic contact. The bulk volume, far from the interface, is electrically neutral n-type Si, with mobile carriers (electrons) compensating the positive fixed charges.

state cannot be achieved if the cycle time is shorter than the thermal generation time constant for the minority carriers that have to populate this inversion layer. Then the depletion state with a low capacitance value is maintained throughout the cycle.

The usual acceptor-type surface defects at the interface between the oxide and the silicon attract negative charges, which always provoke a thin depletion layer in the n-type Si. Therefore, an oxide on n-type Si naturally leads to a surface region with high resistance, while on p-type Si it provokes an accumulation layer that can conduct current between segments. To avoid such short-circuiting channels between diodes in p-type Si, special treatments or layouts must be introduced.

In the surface barrier technology, the separation between evaporated metallic gold or aluminum contacts depends on the quality of the native oxide layer of only some 3–4 nm, which is formed as soon as an Si surface is exposed to air. It was found that better stability can be obtained by oxidation in boiling water. The resistance of such a native oxide can be in mega ohms for 0.1 mm. Much better controlled and higher resistance, giga ohms instead of mega ohms, is achieved with modern oxidation technology.

In most segmentation schemes, the electrodes on one plane are all connected to the same potential, provided via the input contacts at the readout chip, so that there is no lateral electrical field between the contacts in the matrix. All field lines then end in a regularly distributed way on the segmented diode contacts.

22.2.3
Energy Deposition and Signal Generation in Silicon Detectors

Swift ionizing particles lose energy while traversing material. Various mechanisms are at work and an introductory description can be found for example, in *'Passage of particles through matter'* by Bichsel, Groom and Klein (2006). At low momentum, the energy loss in Si is some kilo electronvolts per micrometer, and when the momentum of the particle is much higher, ~0.5 GeV/c, the mean specific energy loss goes to a

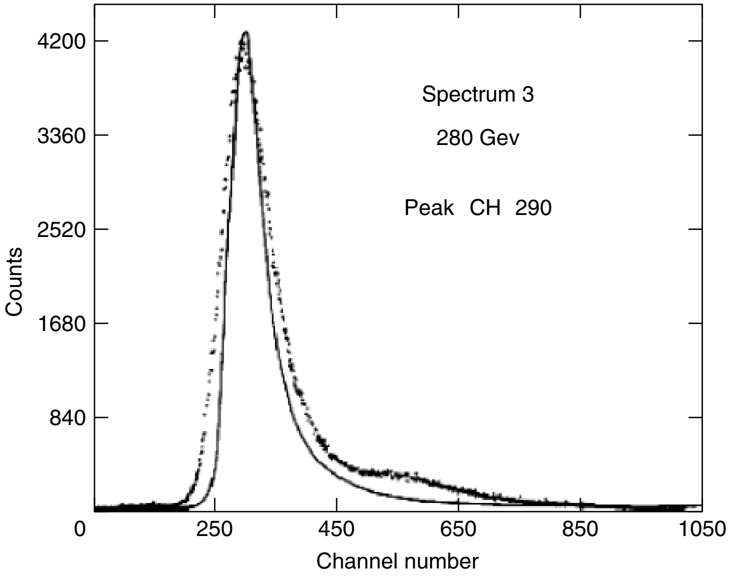

Fig. 22.4 Measured and calculated Landau energy loss distribution for 280 GeV muons traversing a 300 μm thick Si detector in air, from Heijne (1983). The most probable energy deposit is at the peak in channel 290, to which the calculation is normalized. One observes some accompanying secondary particles that give rise to an increased energy deposit probabilty at exactly twice the Landau peak.

minimum of 0.38 keV μm^{-1}. From that momentum range onward the particles are known as *minimum ionizing particles* (m.i.p. or mips). One has to distinguish the energy loss of the particle itself and the energy that is deposited in the absorbing medium. Some of the lost energy escapes from the medium in the form of energetic secondary radiation. Also, the actual energy deposit in a thin silicon layer varies for each particle and statistically it follows the so-called Landau energy distribution, shown in Figure 22.4. In practice, one must consider the "most probable energy deposition" for a particle that corresponds to the value at the peak of the Landau energy loss distribution. Occasionally, large fluctuations may occur, which are caused by close encounters with atomic electrons. Such electrons may receive a large energy transfer, and escape from the sensor volume as "δ-ray" electrons. The specific energy deposited in a thin silicon layer decreases with the layer thickness as shown in Figure 22.5. While for a thick silicon layer of several millimeters the peak of the distribution amounts to 0.36k eV μm^{-1}, the value becomes as low as 0.23 keV μm^{-1} for a layer of $\sim$50 μm thick. The actual value of the energy deposition also depends on the immediate surrounding of the sensitive sensor material.

In a semiconductor, on average about one-third of the deposited energy is spent on excitation of electrons out from the valence band into the conduction band, where carriers can move through the crystal. The remaining positive hole also becomes a mobile charge carrier. The average energy loss per generated e–h

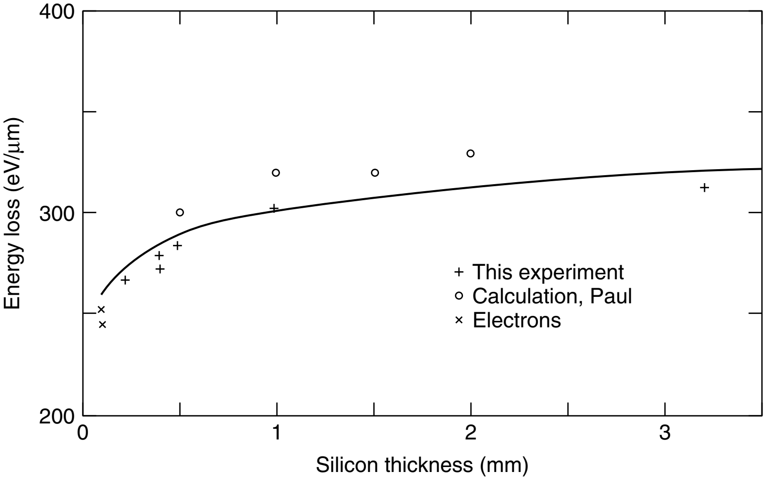

Fig. 22.5 Most probable energy deposition in silicon detectors of different thickness from 0.1 to 3.5 mm, that is, the values in electronvolts per micrometer for the peak of the Landau distributions. The curve is calculated for 200 GeV muons. The + are measured muon data, determined as shown in Figure 22.4, the × are electron data and the o are earlier calculations by Paul. Detailed explanations can be found in Bichsel, Groom and Klein (2006), and Heijne (1983).

pair is 3.62 eV. Of this, only 1.12 eV corresponds to the bandgap energy in Si, and the energy difference is dissipated as phonons and eventually as heat. In a detector of 300 μm thickness, an mip that traverses perpendicularly loses ~80 keV and generates ~22 000 electron–hole pairs. These free electrons (as well as the holes) move under the influence of an electrical field E, with a typical drift speed v, and a mobility μ_e depending on the temperature T:

$$v = \mu_e(T)E \tag{6}$$

The holes that are left in the valence band are mobile as well with mobility $\mu_h(T)$. Values for the mobilities in high purity Si have been measured by Jacoboni *et al.* (1977) and are 1500 V cm^{-2} s^{-1} for e$^-$ and 600 for h$^+$ at room temperature. There is a strong temperature dependence of the carrier mobilities in Si so that the transit time around room temperature varies ~1% for a 0.1° C temperature variation. The mobilities in Si increase with decreasing temperature and reach quite a high value at liquid N_2 temperature ~77 K. Also the thermal conductivity is an inverse function of temperature, and at 77 K, Si conducts actually better than Cu, although it cannot compete in thermal conductivity with diamond or modern diamond-like carbon materials.

The moving charge carriers represent a signal current in the diode and image charges are induced on the diode contacts during the movement as described by Ramo's theorem. The situation for a multielectrode geometry with arbitrary potentials was described by Gatti, Padovini, and Radeka (1982). The actual measured signal may be reduced by "ballistic deficit" if the integration time of the amplifier is shorter than the duration of the signal current, given by the charge transit time in the detector. In practice, for the usual Si detector thickness of 300 μm the hole transit time is <20 nanoseconds for typical bias voltages, and full charge integration can be achieved. In this situation, the integral on nearby electrodes is zero, because the initially induced charge is compensated when the moving carriers reach the collecting electrode(s).

22.2.4
Manufacturing of Semiconductor Detectors

Most of the large radiation imaging detectors currently in use have been made with relatively cheap, commercially manufactured, standard polished, thin silicon semiconductor wafers, sliced from industrial single-crystal ingots. Because of the immense effort in research and development of crystal manufacturing, the silicon industry has achieved extreme electrical and chemical purity and crystal homogeneity, even for a diameter of 300 mm. In some specific applications, different semiconductor materials may present advantages. However, with the exception of hyperpure germanium crystals, all other semiconductor materials have not yet been obtained with similar purity. Also, compound semiconductors such as CdTe often are not monocrystalline, but present a mosaic of subcrystals with different orientations.

The Si crystals are grown in ever-increasing dimensions, over 100-cm long and up to 45 cm diameter, most often along <100> direction, although high resistivity, >kΩcm, is traditionally grown along <111>. High-resistivity n-type or p-type wafers are sliced from float-zone purified crystals and they are available in diameters 2 or 3 in. and 100 mm, 125, 150 mm. The uniformity of resistivity is better in neutron-transmutation-doped (NTD), n-type silicon than in float-zone n-type silicon wafers, where circular striations of approximately millimeter periodicity are observed with tens of percents of resistivity variation. For recent, deep submicron CMOS processing with 90-, 65-, or 45-nm feature size one uses now mostly 300-mm diameter wafers, with resistivity approximately in ohm centimeters. The progress in silicon material is described in the series of four-yearly conference proceedings *'Semiconductor Silicon 19xx'*, ed. R. Huff *et al.* (Huff, Kriegler and Takeishi, 1981).

Standardization in materials and in processing equipment for the CMOS industry has led to much improved device quality and it has lowered production cost. Recent device-manufacturing technologies and parameters are treated in *Silicon VLSI Technology* by Plummer, Deal and Griffin (2000). This industrial development also has facilitated the manufacturing of silicon particle detectors, which started in 1980 with the introduction of "planar" processing by Kemmer (1980) and P. Burger as described in Heijne (2008).

In comparison with advanced CMOS technology, the processing of silicon particle detectors is relatively straightforward. However, special precautions have to be taken because of the unusual materials aspects such as the high resistivity and the need for a long minority carrier lifetime in the bulk silicon. Also the applied reverse bias to the diodes, of several hundred volts, impose a special approach, but which can be similar to that for power devices, as described in *Semiconductor Power Devices* by Ghandhi (1977) and Baliga (1996). The main breakthrough by Kemmer was the introduction of improved cleanliness during the first passivating oxidation step at ~1200 °C. This became possible by advances in MOS technology, but which had not yet been transferred to the detector-manufacturing environment. However, additional long exposures to high temperature, for example, the usual

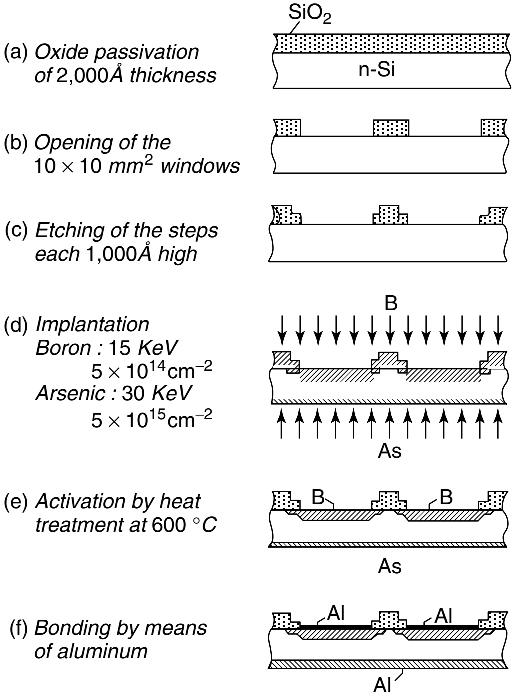
SiO$_2$
(a) Oxide passivation of 2,000Å thickness
n-Si
(b) Opening of the 10 × 10 mm^2 windows
(c) Etching of the steps each 1,000Å high
(d) Implantation
Boron : 15 KeV
5×10^{14}cm^{-2}
Arsenic : 30 KeV
5×10^{15}cm^{-2}
B
As
(e) Activation by heat treatment at 600 °C
B
B
As
(f) Bonding by means of aluminum
Al
Al
Al

Fig. 22.6 Schematic process flow for 1-cm^2 diodes following the description by Josef Kemmer in US Patent 4,442,592 (Kemmer 1980b).

1100 °C anneal step, still have to be avoided as these generally lead to the introduction and diffusion into the bulk volume of metallic impurity defects, which shorten the minority carrier lifetime and lead to an increase of dark reverse bias current.

Kemmer described schematically the basic manufacturing steps for silicon detectors, as shown in Figure 22.6. Some critical details have not been included, such as the profiling of the diode edges and the parameters of the cleaning steps. Commercial manufacturers, understandably, have been reluctant to publish such manufacturing details.

After thorough cleaning of the silicon wafers, the first processing step is the growth of a passivating oxide layer that completely covers the surface. Oxidation takes place in a quartz tube in a high-temperature furnace. The Si wafers are exposed to a homogeneous flow of oxygen or water vapor. At, for example, 1200 °C, a typical SiO$_2$ thickness of 25 nm is grown in 10 minutes. Often, HCl or, preferably, a Cl-containing compound is added in the gas flow, such as trichloroethylene, to prevent diffusion of metal atoms from the outside heating coils through the hot quartz tube into the Si wafers.

The definition of the diode pattern on the wafer is obtained by illuminating a light-sensitive photoresist, which after development presents a barrier layer of polymer material on the nonilluminated places and open windows where the light came through the mask (positive resist). The creation of diodes needs the implantation of doping atoms with an ion-implantation machine, which is, in fact, an ion accelerator with a high current beam that is swept over the wafer area. By chosing a low ion energy, for example, 15 keV, the penetration of the ions is only some tens of nanometers and a thin implanted thickness with high concentration of doping atoms is obtained. To be electrically active, the implanted ions have to be located on lattice positions and for this rearrangement, one needs an activation anneal at elevated temperature, usually above 1000 °C. Care has to be taken that during this step, the radiation damage from the implantation is repaired, and also that no diffusion of undesirable atoms such as Na or Fe takes place. Sometimes, special ''gettering'' structures are provided between the devices or on the back side of the wafer, which can catch such undesirable atoms.

After processing, the individual sensors on the wafer have to be separated. Traditionally, this is done with a wafer-saw that cuts an approximately 60-μm kerf in the ''dicing street'' between the chips. Recently, singulation has also been done with laser cutting or with plasma ion etching. These methods promise to bring less crystal damage along the edge of the cut, so that the insensitive edge width can be reduced.

The precision assembly of the sensors and the electrical, mechanical, and thermal contacts require careful design and construction. The position-sensitive devices aim at measurements with 2–10 μm accuracy. Manipulation and assembly are executed under dust-free conditions and at controlled temperature, in medium-quality clean rooms, as was already illustrated in Figure 22.1.

22.2.5
Degradation of Detectors

During operation, detectors may degrade for various reasons, such as a prolongated heat load that causes metal diffusion into the pure Si material, or humidity that short-circuits isolation layers in critical places of the device. Besides such accidental degradation, the changes induced by particles incident on the sensor material are a fundamental and unavoidable degradation mechanism. Radiation quanta generate free electrons by ionization of the semiconductor material, but at the same time a fraction of the energy deposition leads to dislocations of atoms from their original crystal lattice positions. The ejected atom is mobile as an interstitial; also, the created lattice vacancy can move around for a while, until it combines into a stable defect. The movement of atoms, vacancies, and complex defects is strongly dependent on the temperature and the concentrations of impurity atoms. Whereas in hyperpure Si, the concentration of electrically active impurities initially can be as low as $10^{12}\,\mathrm{cm}^{-3}$, there are usually chemical impurities such as C or O at a concentration of order $10^{15}\,\mathrm{cm}^{-3}$ and these play an important role in the stabilization of the radiation-induced defects. Many studies have been devoted to radiation effects in semiconductors, and Si, in particular (see Series of Conferences on *Defects in Semiconductors* (Oshiyama *et al.*, 2005)). The effects of radiation damage in high-resistivity Si did not receive a lot of attention until recent decades. Some defects that at first are stable at a certain temperature may gradually convert into a different type of defect complex, for example by combining with a diffused oxygen atom. Such a process, called *annealing*, is usually thermally activated, and often leads to reduction of the reverse bias current. When, on the contrary, the new defects lead to a continuing increase of dark current, it is referred to as *reverse annealing*.

Unfortunately, the defect complexes in Si that result from the irradiations, such as vacancy–vacancy, oxygen–vacancy, phosporus–vacancy, and so on, present a variety of electrically active energy levels in the semiconductor bandgap. Because the functioning of semiconductor detectors depends on a small concentration of electrical charges, typically $<10^{13}\,\mathrm{cm}^{-3}$, these devices are quite sensitive to changes in surface charge and to radiation-induced defects. It is found that active defects are introduced at a rate of $\sim 0.1\,\mathrm{cm}^{-3}$, per incident particle per centimeter squared. Therefore, it is not surprising that a fluence (integrated flux) of $\sim 10^{13}\,\mathrm{cm}^2$ particles leads to significant changes in the operation characteristics of a Si detector. It is found that many radiation-induced defects exhibit acceptor-type behavior, and compensate the donor characteristics of the traditional, n-type starting material for Si detectors. It was also found that an operating temperature well below $-10\,^\circ\mathrm{C}$ can prevent formation of the most agressive type of defect, which is believed to be a carbon-related complex. An excess of O in the starting Si seems also to lower the introduction rate of this agressive defect. Watts (1997) gave an overview of work on defects in the particle physics community.

Recently, studies have started on the use of silicon crystals that have been grown in a different way, such as epitaxially grown layers, or p-type instead of n-type material. The first results show that different types of stable defects are being obtained due to different background conditions, and that in some cases operation at $+15\,^\circ\mathrm{C}$ would be preferable above cold operation, owing to beneficial annealing.

22.3
Various Sensor Designs and the Associated Readout Electronics

The architecture of the readout electronics really determines the different sensor designs. Initially, a single amplifier with discrete components was used for each detecting unit. For position-sensitive devices it was possible to limit the number of readout channels to two, or to a maximum of four, by incorporating sophisticated, resistive charge division in the sensor construction, as reviewed by Laegsgaard (1979). In many nuclear physics measurements, such detectors with charge division are adequate, but they cannot handle a situation in which several particles are incident at the same moment, at different positions. Yet, already in 1970, it was possible to exploit detectors with tens to hundreds of segments using voluminous electronics, as in the BoL project in Amsterdam (Koerts *et al.*, 1971).

Using custom-designed integrated circuits, with 64, 128, 256 or even tens of thousands of amplifiers per chip, it has become much easier to apply parallel signal processing for a multitude of microscopic sensor elements. This alternative approach opened up applications in multiparticle experiments in particle physics, light detection, and space, with thousands of cells per sensor. Although there are no formal standards, one may distinguish sensor layouts following a classification of the cell dimensions, as shown in Table 22.2, with 1 mm as the distinguishing criterion for the naming convention.

A basic distinction exists between a hybrid detector system and a monolithic

Tab. 22.2 Segmented silicon detectors – naming convention based on cell dimensions.

Detector type	Coordinate	Long edge	Short edge	Associated chip
Pad detector	2D	≥1 mm	≥1 mm	16–64 channels
Microstrip detector	1D	≥1 mm	<1 mm	128 channels
Hybrid Pixel detector	2D	<1 mm	<1 mm	2048–65 536 channels
Monolithic pixel chip	2D	≤30 μm	≤30 μm	≥250 000 cells

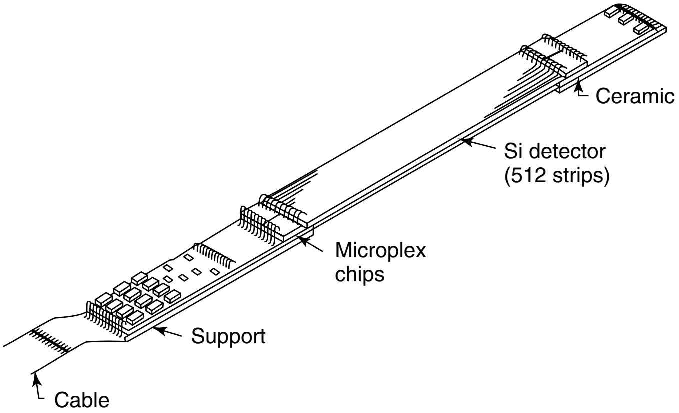

Fig. 22.7 Sketch of one of the first ladder assembly designs for a silicon detector module. This was made around 1987 for the silicon tracker in the MARKII experiment at SLAC. The module contains a single silicon microstrip detector with 512 strips. On both sides of the sensor, alternate strips are connected by wire bonding to the inputs on the two 128-channel Microplex readout chips, which are glued on a ceramic support that also carries the supply and filtering components, and the connector for the cable to the off-detector electronics. Reproduced from LaBarga *et al.* (1991).

detector system.[2] The first three rows in Table 22.2 refer to hybrid detector assemblies in which the sensor cells, manufactured on a separate silicon chip, are each separately connected to the first stages of the amplifier chains, located on another chip. A typical assembly is illustrated in Figure 22.7, which represents a so-called ladder, a module for a detector array in the form of a barrel. Pioneering work on module assembly and readout chip development was undertaken in a collaboration at SLAC, during the mid-1980s (LaBarga *et al.*, 1991).

22.3.1
Custom Design of Chips

Segmented silicon sensors such as microstrip detectors or hybrid pixel detectors need a large number of parallel amplifiers, and parallel-to-serial conversion of the signals in the chip circuit, so as to allow more economical sequential readout by the data acquisition system. At Stanford, Hyams, Parker, and Walker initiated the design of an NMOS chip, called *Microplex* (Walker *et al.*, 1984), which, with 128 amplifiers, was the first integrated design for silicon microstrip detector readout. Their design used a switched capacitor feedback amplifier and this evolved later, thanks to Kleinfelder *et al.* (1988) and Spieler (1994) and coworkers at the Lawrence Berkeley Laboratory, into a family of CMOS chips. These are used even now at Fermilab in the CDF and D0 Tevatron collider experiments. In Germany, a team at the Werner Heisenberg Laboratory, Max Planck Gesellschaft, Munich and the Microelectronics Institute in Duisburg, with Lutz and Buttler, together with Manfredi *et al.* at the University of Pavia, developed a 128-channel CMOS circuit with JFET front-end transistors (Lutz *et al.*, 1988). Later this chip was called *CAMEX* and was used in the electron collider experiment Aleph at CERN, from 1989 to 2001.

---

[2] Monolithic designates the situation that all elements are built on a single "stone" of silicon. A segmented sensor is also monolithic, but here we mean that the sensor as well as the amplifiers and other readout electronics are built on one chip.

An additional feature necessary with high-rate interactions in particle physics experiments is an on-chip temporary data-storage memory, from which selected data that belong to a precise timeframe with a positive trigger can be retrieved later. For the Large Hadron Collider (LHC) experiments, these memories usually operate at 40-MHz frequency and provide storage for the signal charge quantity that is measured every 25 nanoseconds in each of the sensor elements. The memory typically has a depth of 180 cells, which allows all data to be kept until ∼4 microseconds later a first level trigger has determined the data set that may be of interest.

Many of the front-end circuits use an amplifier with switching feedback and double-correlated sampling. A different line of devices, such as the 16-channel *AMPLEX* chip, was initiated by Jarron, as reported by Beuville *et al.* (1990). This family uses the more classical continuous feedback, which can be implemented in different ways. For example, in the first *AMPLEX* a long transistor is used as the feedback element, as shown in Figure 22.8 (Heijne and Jarron, 1988). This approach allows direct connection between sensor and amplifier input, with compensation of the effects of the sensor leakage current. A current up to 450 nA per segment can thus be compensated. The sensor dark current at the input modulates the effective resistance of the feedback so that the operational level

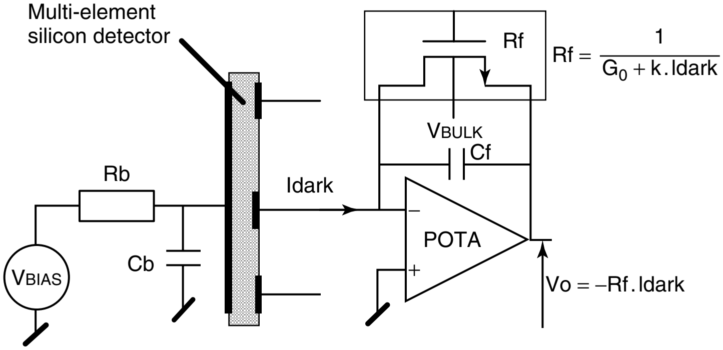
Multi-element
silicon detector
Rb
Cb
VBIAS
Idark
Rf
VBULK
Cf
POTA
–
+
$R_f = \frac{1}{G_0 + k.\mathrm{Idark}}$
$V_o = -R_f.\mathrm{Idark}$

Fig. 22.8 Schematic diagram of the connection of segmented diode elements of a silicon microstrip detector to an operational transconductance amplifier (OTA) with a feedback element presenting resistance Rf and capacitance Cf. This type of feedback restores the operating point also if there is a large diode leakage current Idark in the sensor element at the input of the amplifier (Heijne and Jarron, 1988).

at the output can be maintained at the cost of a slight reduction of amplification. The chip was produced in 1987 in the 3-um CMOS technology of MIETEC in Belgium.

A variety of readout chips with a linear array of amplifiers has been designed since about 1990 for use with silicon microstrip detectors. The pitch of the diode elements on the sensor is not always identical to the pitch of the amplifier input pads on the readout chip. For reliable wire-bonding connections, it would be best to make pitch adaptation on the border of the sensor, using metal pads on a thick oxide in combination with a separate depletion region to make the border independent of the sensitive cells. In some systems, pitch adapters on a separate substrate such as quartz have been introduced, but then the number of connections would need to be doubled.

22.3.2
Imaging Matrix with Linear or 2D Sensor Geometry

Microstrip detectors provide a 1D projection of the particle positions. For 2D position measurement, a second detector is needed, with the strips orthogonal, or at a different angle. When several particles are simultaneously incident on such a sensor sandwich, ambiguities arise and a third sensor plane with another strip orientation is needed to resolve these. The use of a double-sided microstrip detector reduces the number of planes, but does not resolve ambiguities.

A true 2D device with a matrix of sensitive cells makes it possible to assign unambiguous positions to individual particles, even if there are many coincident on the same device. The first experiment that practically adopted this approach at full scale was NA11. Damerell *et al.* (1986) adapted standard charge coupled devices (CCDs) for use with mips. They showed that with true 2D detectors one can obtain a clean pattern recognition. The fully serial readout of the CCD makes it quite a slow detector, and increased parallelism is required for high-rate experiments. The integrated nature and the small pixel capacitance of the CCD result in low electronic noise, and a thin sensitive layer of $\sim 20\,\mu m$ of silicon delivers a sufficiently large signal.

While in a CCD the signal charge is transported from pixel to pixel, in a clocked parallel–serial operation, all over the matrix toward a single output node, a different charge transport mechanism can be implemented by the use of potential gradients in a homogeneous silicon volume. In the early 1970s, this approach had already been adopted for light-sensitive devices. It has been introduced for position-sensitive particle detection by Gatti and Rehak (1984). A lateral drift potential is superimposed on the voltage that depletes the silicon bulk volume. Signal charges are drifted by the potential gradient as well-defined packets toward the readout nodes, which form a linear matrix along one or two borders of the silicon chip, as in a microstrip detector. A drawing of a practical drift detector for the Alice experiment is shown in Figure 22.9, reproduced from a description by Nouais *et al.* (2003).

While the drift detector geometry has a linear, one-dimensional matrix of readout pads similar to that of a microstrip detector, it presents true 2D information on particle impact positions. The second coordinate can be derived from the drift time, if the timing of the impact is measured independently, for example, from a common pulse pickup at the detector backside or from the known bunch crossing time in a collider experiment. An important advantage of the drift detector is the very low capacitance of

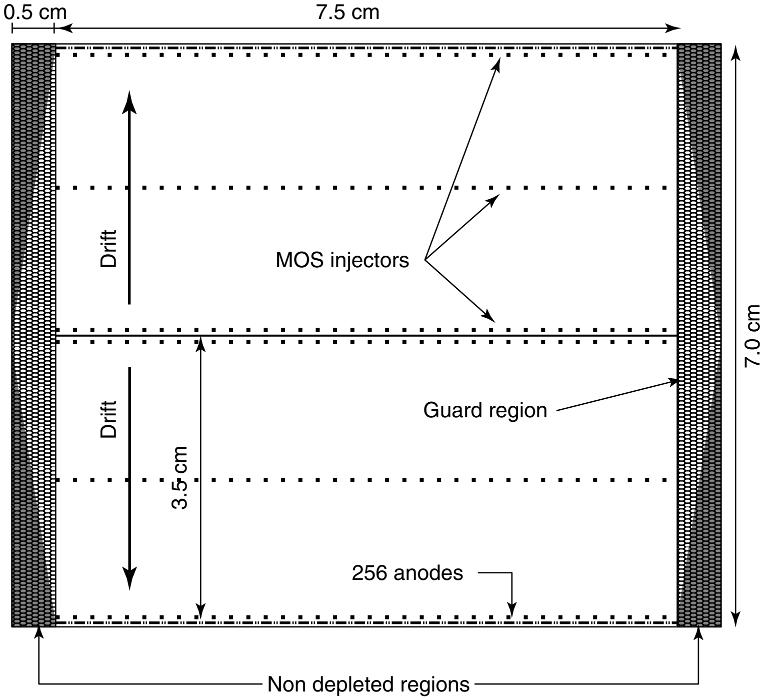

Fig. 22.9 Schematic drawing of a silicon drift detector for the Alice experiment (Nouais *et al.*, 2003), with 256 anode readout elements along the top and the bottom. The drift potential is vertical from the center, both to the top and to the bottom anodes. The drift potential is distributed along the sides, and a typical value is 2000 V over the 3.5-cm drift region. Several rows of MOS injection capacitors are indicated. These can be used to monitor the conditions for the drift velocity, which is strongly dependent on the silicon doping, radiation defects, and the ambient temperature.

the readout node, in comparison with a microstrip detector of similar dimensions. However, the sensitivity of the drift velocity on the uniformity of the devices and on the ambient temperature presents practical limitations that require continuous calibration procedures. Calibration is performed by charge injection via built-in capacitors.

For high-rate experiments, fully parallel hybrid pixel detectors have been developed, which consist of a silicon sensor chip and a matching chip with readout circuits, as illustrated in Figure 22.10. Each sensor cell is individually connected to a readout cell, using a microscopic solder connection, called *bump-bond*. The first such hybrid pixel detectors had ~1000 pixels of dimensions 75 μm × 500 μm, but the most recent one has a matrix of 256 × 256 pixels of 55 μm × 55 μm. Besides the need to register the positions of incident particles, one now aims also to incorporate a variety of additional functions in the pixel circuitry, such as local encoding of the timing and the signal amplitude. For example, a 100-MHz clock distribution has been implemented in the Timepix assembly (Llopart *et al.*, 2007). The schematic circuit diagram of the pixel electronics in this Timepix chip is reproduced in Figure 22.11. This drawing shows the complexity that can be achieved

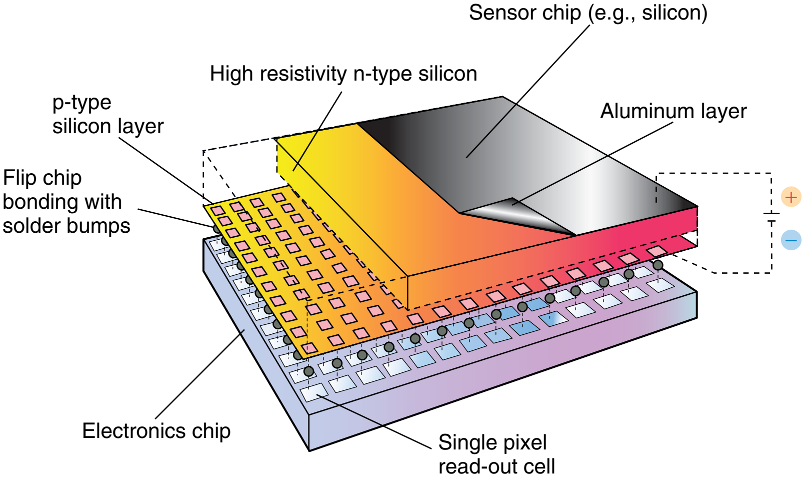

Fig. 22.10 Schematic drawing of a "flip-chip" hybrid silicon pixel detector showing a partially transparent sensor matrix chip and a matching CMOS electronics readout chip underneath. Solder bumps are indicated, which connect the individual sensor cells to the corresponding readout elements in a fully parallel fashion. The sensor is supposed to be a matrix of p–n junctions with a common ohmic rear-side contact covered with an aluminum metallization.

on a small, 55 μm × 55 μm area with recent submicron CMOS technologies.

22.3.3
Factors Affecting Signal/Noise

The signal amplitude in nuclear physics applications is mostly determined by the energy of the incoming particle, often several mega electron volts, provided that the particle comes to a complete stop inside the detector. In particle physics experiments, on the contrary, the ionizing particles generally traverse the silicon detector while undergoing only a fractional, small energy loss, typically of order 30 keV per 100 μm of Si thickness. The signal charge is practically independent of the particle energy, and only determined by the pathlength in the detector, and the thickness of the sensor chip. In order for such an mip particle to be detected with 100% efficiency, the noise of the signal-processing chain has to be sufficiently low. For readout electronics of particle detectors, the noise distribution is usually characterized by the standard deviation of the variance, expressed as the number of equivalent noise, given in units of electron charges (ENC) at the input of the chain, given in units of electron charges. For reliable discrimination of signals above a threshold from the noise distribution, it is desirable that the signal exceeds the noise baseline value by at least six standard deviations. For reliable detection of a swift particle in a segmented detector, one also has to take into account the signal charge division, when this is shared between adjacent segments.

The noise in the signal-processing chain has various sources, and optimization can be achieved in the sensor as well as in the front-end amplifier. The sensor capacitance, the signal shaping time, the power dissipation, and the dark current are prime objectives in the noise reduction. In the pixel detector, one can reduce

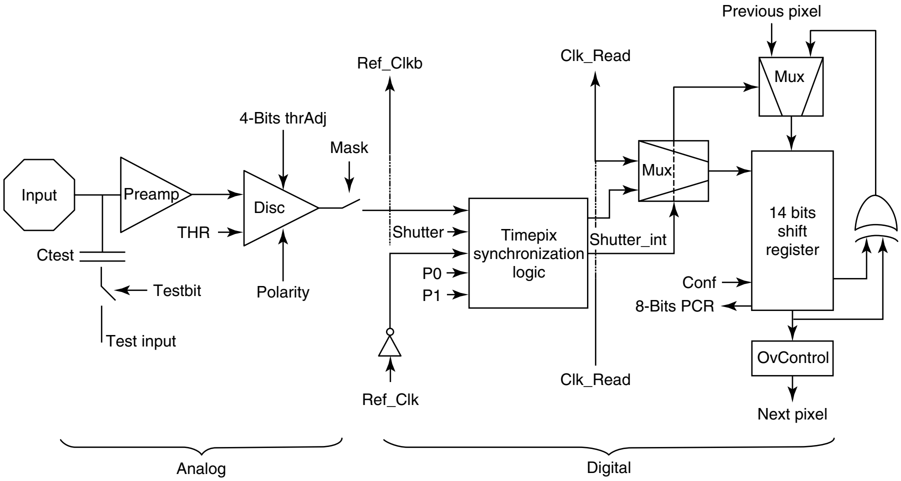
Previous pixel
Ref_Clkb
Clk_Read
Mux
4-Bits thrAdj
Mask
Mux
Input
Preamp
Disc
THR
Shutter
Timepix synchronization logic
Shutter_int
14 bits shift register
Ctest
P0
P1
Testbit
Polarity
Conf
8-Bits PCR
Test input
OvControl
Ref_Clk
Clk_Read
Next pixel
Analog
Digital

Fig. 22.11 Schematic diagram of the pixel electronics in the Timepix chip (Llopart *et al.*, 2007), which is situated in a 0.003 mm^2 area, and reproduced 256 × 256 times on this 0.25-μm technology CMOS chip.

both capacitance and dark current for a single pixel, while the signal amplitude remains practically identical. Therefore, pixel detectors offer a very robust approach to detection with practically 100% efficiency for ionizing particles.

22.4
Typical Silicon Detector Arrays in a Particle Physics Experiment

During the recent decades, silicon detector arrays have become a standard feature of particle physics experiments, around fixed targets as well as in the collider experiments. Immediately after the first publications of test results in 1980, the first fixed target experiments to embrace this new technology were WA82 (in the Omega spectrometer), WA75 and NA11 at CERN, and the photoproduction experiment E691 at Fermilab. Around 1982, the CERN proton–antiproton collider experiments had not yet started using these devices, but in the final stage, 1986–1988, UA2 installed two layers of Si pad detectors to recognize leptons and to reject signals from photon conversions in the inner volume. These arrays had ∼1.1 m^2 area with 6000 sensing elements. The MARKII experiment at SLAC and the Fermilab collider experiments CDF and D0 then adopted, from the beginning, silicon microstrip detectors for their inner trackers, all with tens of thousands of amplifier channels. By the end of the 1980s, most Large Electron Positron collider (LEP) experiments at the CERN were preparing inner silicon detector barrel arrays, equipped with CMOS readout chips, as mentioned in the previous section. Also the BaBar experiment at SLAC and experiments at several other smaller accelerators incorporated silicon detector arrays. Now with the new generation of experiments at the LHC at CERN, the collaborations have installed silicon tracker arrays that are unprecedented in size and complexity. A few examples are mentioned, but already hundreds of pages have been published to provide details of the design, construction, system monitoring, and expected performance for all these silicon detector systems.

The design of the silicon tracker arrays is usually based on the physics goals for particle tracking near the interaction vertex. The precise measurement of the coordinates of the primary interaction point in particle physics experiments and those of possible secondary vertices became more important with the discovery of short-lived particles with a decay length of a few millimeters. A secondary vertex indicates the presence of a charm, beauty, or τ particle among the interaction products. Such a particle is a messenger for a potentially interesting, special reaction. A second goal of the silicon imaging detectors is the improvement on the momentum measurement of the outgoing particles. The precision needed to achieve a given momentum resolution is a function of the radius at which the detector plane is installed, and somewhat relaxed precision can be tolerated toward larger radii. This has led in several experiments to a multitude of different systems that enclose one another in the way of onion-shells. Practical experience will show if this is the optimal approach. Maintenance, replacement and extension operations, and cost may eventually lead to modified designs. It may be mentioned that the use of silicon tracking devices has led to order-of-magnitude improved precision compared to earlier employed gaseous detectors or scintillators, and this has allowed the overall volume of an experiment at a given energy to shrink significantly. The higher cost of the silicon-based systems in the center can be offset by the smaller volume of the magnet and other parts of the detector, including the outer muon chambers.

In recent studies about requirements for future electron collider experiments, the need for improved precision in the particle total energy measurement has become an issue. Some designs of a ‘‘silicon-based’’ calorimeter with tracking capability are being evaluated. In addition, the identification of original leptons from the primary or secondary interaction vertex may be improved by using radiation imaging arrays.

22.4.1
Tracking Detectors for Momentum Measurement

Silicon tracker detectors consist of several layers of silicon position-sensitive devices with associated readout chips. The layout for a typical example is shown in Figure 22.12, with the description of the device in the D0 experiment at Fermilab by Hanagaki (2004). The outer layers in this tracker barrel were supplemented at a later stage with layers 0A-B and 1A-B in the inner volume, very close to the beam pipe. Now being longer and carrying more sensors and readout chips, the modules in such a barrel geometry still continue to follow the original design that was developed for MARKII, as shown in Figure 22.7.

A photograph of part of the real silicon tracker for the CERN CMS collider experiment (Compact Muon Solenoid) is reproduced in Figure 22.13 before complete cabling and encapsulation. Here one can still see the silicon sensors and readout chips, arranged in a way very much similar to the schematic in Figure 22.12. Once all connections for power, signals and cooling have been installed, and thermal insulation is in place, not much of the devices remain visible. The CMS silicon tracker has been described in numerous publications, for example, by Pooth (2006),

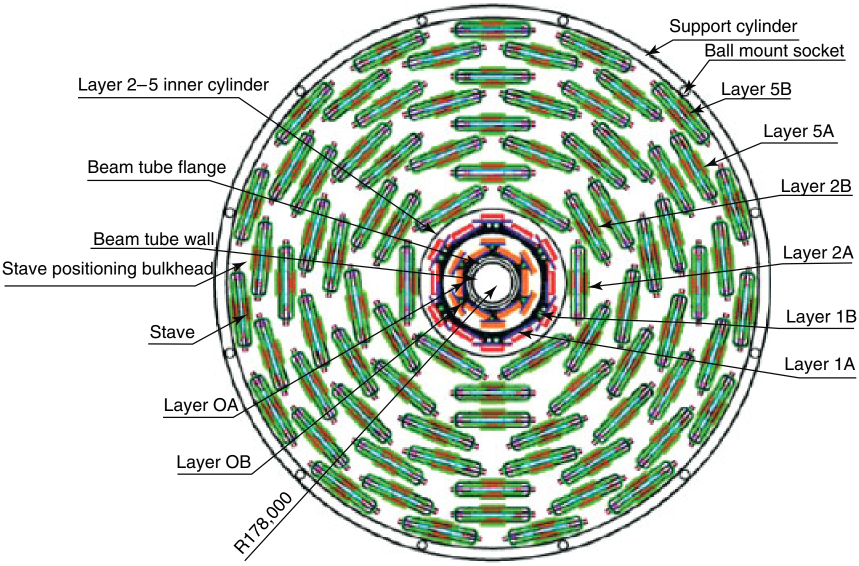

Fig. 22.12 Schematic drawing of the six silicon microstrip detector layers in the Fermilab D0 experiment. The tracker is divided in two groups, with the L0-1 layers in the center and the earlier installed layers 2–5 at radii between 53 and 164 mm. (from Hanagaki 2004).

and is the largest silicon array to date. The CMS tracker array has 10 cylindrical, overlapping layers in the barrel section with radius 20–120 cm, and three inner disks plus nine end-cap disks on either side of the barrel. The overall length is 5.6 m. It covers an active sensor area of 200 m^2, divided over more than 15 000 microstrip sensor units, with diode strip pitch between 100 and 200 μm, depending on the distance of the sensor from the interaction point. All sensors are read out using the 128-channel APV chip, of which ~80 000 have been installed next to the sensors. A critical aspect is the cooling of the instrument, which dissipates several tens of kilowatts, and which has to be maintained below −10 °C to preserve the quality of the high-resistivity silicon under irradiation.

The design of the CMS microstrip tracker allows 8–14 measurement points on each outgoing track, with position precision of the order 10–20 μm. This is complemented by about three points closer to the beampipe, measured in the pixel detector situated within the tracker barrel.

22.4.2
High-Precision Detectors for Interaction Vertexing

For high-precision vertex measurements, one now often uses pixel detector planes, which offer direct 2D space points with precision better than 10 μm. The first use of a pixelized matrix detector was the CCD in the NA11 experiment at CERN, already mentioned (Damerell *et al.*, 1986). Because of the shifting of the charge signals and serial readout via one or a few output nodes, the application of CCD is limited to some specific types of experiments. Some can tolerate low rates, or, as in NA11, the beam

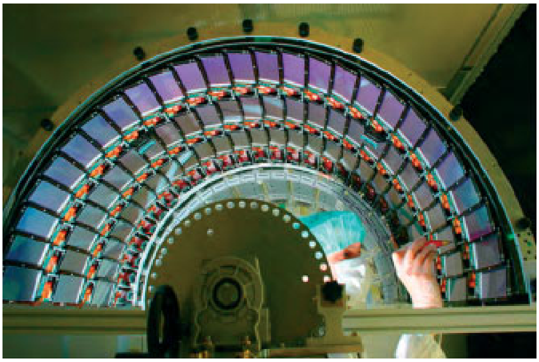

Fig. 22.13 Photograph of the Tracker-Inner-Barrel inside the CMS silicon microstrip detector at CERN, during construction in a clean room. Photo CERN, also published by Pooth (2006).

can be deflected after a positively triggered interaction so that the detectors are not flooded with additional, confusing signals.

In high-rate experiments one can use the hybrid pixel detectors described in Section 22.3. The first application was in the high-intensity environment of the CERN WA97 lead-ion beam, impinging on a lead target (Antinori *et al.*, 1995). A reconstruction of an event is shown in Figure 22.14. The resolving power compared to projective microstrip detectors as well as the low noise are clearly illustrated.

The pixel detector technology is a sophisticated integration of the sensor and the electronics readout function, and it is attractive to think of a full monolithic device, wherein all functions are implemented together. CMOS sensors for visible light now achieve this integration, and have become competitive with CCD in many applications. However, in nuclear and particle physics experiments the devices have to fulfill the need for single-quantum processing, where each particle is recorded with its time of incidence and other characteristics. Continued research is ongoing, in order to build such monolithic detectors for particle physics.

In the LHC collider experiments Alice, Atlas, and CMS at CERN, large arrays of hybrid pixel detectors have been constructed, with sensitive areas of 0.24, 1.7, and 1.2 m^2, respectively. A part of the Atlas pixel detector is shown during assembly in the clean room in Figure 22.15.

A comprehensive description of work on pixel detectors has been published as *Pixel Detectors, From fundamentals to applications* by Rossi *et al.* (2006).

22.4.3

Semiconductor Detectors for Particle Energy Measurement

The energy loss and the conversion into free electric charges in silicon is very reproducible and well understood. If an energetic particle can be made to stop in the detector volume, a constant fraction of its energy is converted into free electron–hole pairs, on average 3.62 eV per pair. For particles of much higher energy, so-called calorimeters have been built for the measurement of their total energy, and these may be more than a meter thick to stop the electromagnetic or the hadronic shower that develops when the particle comes to rest. A tedious calibration process is needed for the current calorimeter devices that employ alternating sheets of passive absorber (Fe, W) and active detector material such as light-emitting scintillator

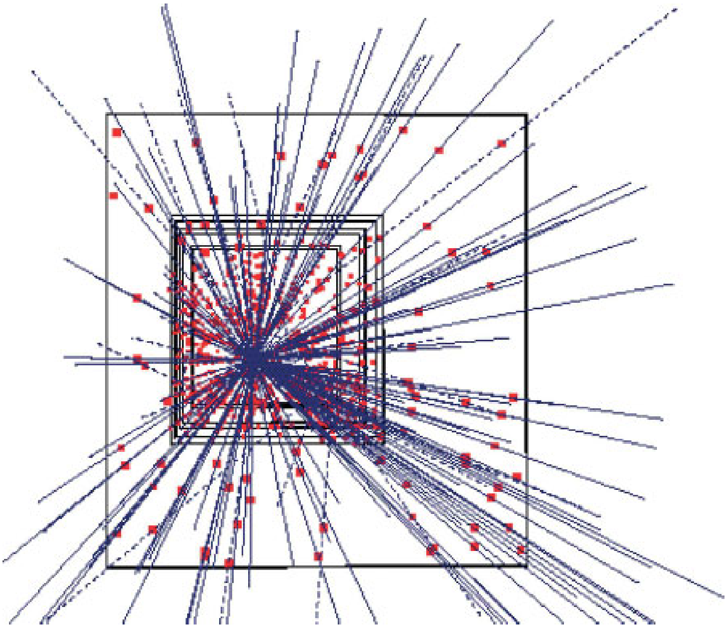

Fig. 22.14 Perspective view of the reconstruction of 153 tracks in a partial angular cone of a lead–lead interaction in the fixed target experiment WA97 (Antinori *et al.*, 1995), using seven planes of silicon hybrid pixel detectors. The registered hits in the planes are indicated by small squares, with smaller squares for the planes closer to the interaction vertex. Most tracks originate at this vertex, and are shown fully drawn. A few tracks cross only a subset of the planes and do not pass through this origin. These tracks are shown by a hatched line. In the analysis, it was found that all hits correspond to a physical particle track, and no uncorrelated noise hits have been observed. This is expected owing to the large noise margin in such a pixel detector. The magnetic field of the Omega spectrometer had been switched off when this interaction took place.

crystals or liquid argon, in which the ionizing secondaries generate mobile free electrons and slow-moving ions.

Simulations show that improved performance can be obtained if the sensor were made out of silicon semiconductor detectors. By about 1985, some prototypes had been tested, but the technology was practically abandoned in view of the high cost of silicon detectors. Since 2006, the development has been taken up again, and in the framework of the EUDET detector R&D for a future international linear collider (ILC), new prototypes are being evaluated (EUDET-Calice Collaboration). Besides the improved energy measurement, the increased calorimeter segmentation is an attractive feature; in addition, such a highly segmented device acts as a continuation of the inner tracker.

22.5
Other Applications of Semiconductor Detectors

The extensive efforts on silicon detector development in particle physics have paved the way for applications in other fields. A few space-borne projects are mentioned here, which have been initiated by particle physicists. In other applications,

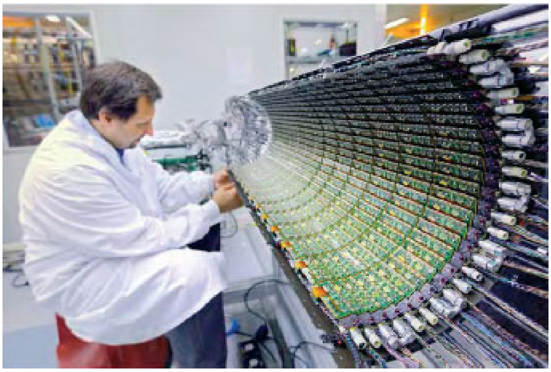

Fig. 22.15 The Atlas silicon pixel detector during assembly in the clean room. A half-shell of the barrel part can be seen. Ladders are supported on a lightweight but rigid carbon-fiber structure. On each ladder a number of through-connected modules can be distinguished. The white tubes are the connections for the cooling fluid (Photo CERN).

for example, protein diffraction with synchrotron light, or materials analysis, scientists from particle physics have been involved as well.

22.5.1
Satellite-Based Experiments

For a long time, silicon detectors have been used on-board satellites for intensity and energy measurement and identification of cosmic rays. There are numerous possibilities for physics exploration that are offered by the natural accelerators that exist far away in the universe. These generate particles with much higher energies than possible on earth, although the fluxes are very small when reaching the solar system. The particle imaging capabilities of multilayer silicon tracker arrays allow a much improved directional detection of these incoming particles, so that these may be correlated with a cosmic source. With this low flux, the sensitivity has to be increased by enlarging the sensitive area. In recent years, balloon experiments such as PAMELA (described by Straulino *et al.*, 2004), have used silicon detector arrays for the measurement and identification of cosmic particles. One of the PAMELA detector modules is shown in Figure 22.16.

The basic ladder structure is also maintained here, as it was adopted originally for the MARKII particle physics experiment at SLAC.

Between 1995 and 2008, two much larger experiments have been built for use in space, in which silicon detector arrays form the main component of the instrumentation. The Fermi Gamma Ray

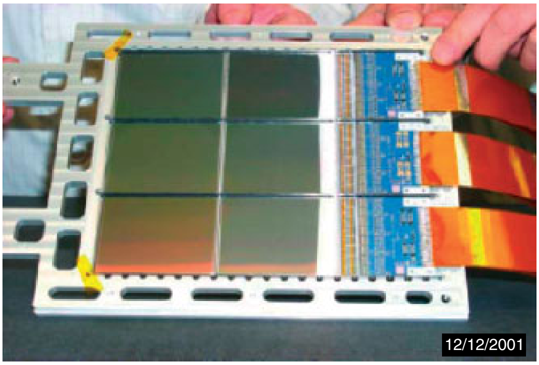

Fig. 22.16 Photograph of a silicon detector plane for the PAMELA experiment. Three ladders, each with two double-sided detectors, are mounted together on an aluminum frame. The diode strips are wire-bonded from sensor to sensor at the left and then bonded to the readout chips at the right. Flexible foils carry the amplified signals further (Straulino *et al.*, 2004).

Space Telescope (*www-glast.stanford.edu/*), formerly called *GLAST*, has a large area tracker (LAT) that consists of 87 m^2 of Si micostrip detectors, 20 000 sensors, and their associated readout chips. The instrument has been described by Sadrozinski (1998), and more recently by Tajima (2007). The total number of sensor readout channels is ~10 million, but the overall power budget is only ~250 W, which is provided by the solar panels. The GLAST/Fermi telescope was succesfully launched on June 13, 2008.

The Alpha Magnetic Spectrometer project is a two-step experiment, described, for example, by Viertel and Capell (1998), with AMS01 (3200 kg) and AMS02 (6000 kg). AMS02 is to be launched by the NASA shuttle and then to be operated at the International Space Station. The website (*http://ams-02project.jsc.nasa.gov/*) gives numerous details. The experiment searches for new physics phenomena in cosmic radiation. Inside a 0.8-T superconducting magnet there is an eight-layer ionizing particle tracker built from double-sided Si microstrip detectors. Some practical aspects of the Si tracker and the construction have been described in publications by Haas (2004) and by Ionica (2003). In this instrument, the total silicon sensor area is 7.2 m^2 and it contains 2286 Si detectors, placed on 192 ladders with a total of 196 608 sensing channels. For the readout, the VA-hdr chip (*http://www.ideas.no/*) is used with 32 channels per chip.

To give an impression of such a space-based instrument, the schematic drawing of AMS02 is shown in Figure 22.17. The distinctive feature is the superconducting, liquid He cooled magnet, that should allow charge determination and precise momentum analysis of the cosmic particles as they can be detected while still outside the earth's atmosphere. The instrument should be transported to the International Space Station by one of the space shuttle flights.

22.5.2
Detectors at Synchrotron Light Sources

Imaging of synchrotron light and X-ray diffraction operate often with photon signals that are similar to that of mips in silicon, in the range 5–100 keV. For the lower end of this range, the absorption of Si for X-rays is quite satisfactory, and the silicon devices that have been developed for particle physics experiments can serve well for different imaging applications at synchrotrons.

Medipix single photon processing detectors have been tested for low-intensity imaging and prove very efficient even for single quanta, as described by Tlustos (2005). A second imaging detector for high photon intensity, with storage capability for up to eight successive timeslots has been developed by Gruner and his team, as published in Rossi *et al.* (1999).

''Pilatus'' is the first large-area silicon array detector for protein crystallography with direct photon conversion, designed at the Swiss Light Source in the Paul Scherrer Institute in Villigen and now produced by Dectris AG (*http://www.dectris.com/*). The design is modular and an array can be composed of a certain number of basic pixel detector modules that each carries 16 readout chips on a Si detector with a matrix of pixels. The cell dimension in the most recent version is 172 μm ×172 μm for a 60 × 97 matrix on a readout chip of approximately 10 × 17 mm^2. The modular structure is illustrated in Figure 22.18.

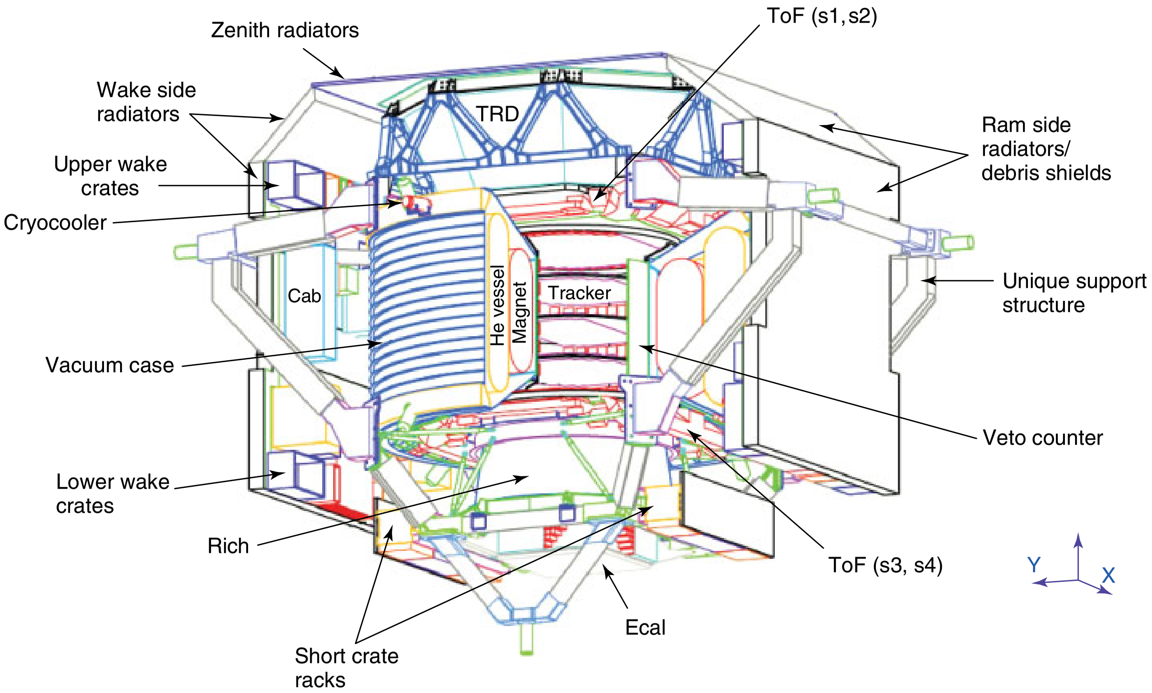
Zenith radiators
ToF (s1, s2)
Wake side radiators
TRD
Ram side radiators/ debris shields
Upper wake crates
Cryocooler
Cab
He vessel
Magnet
Tracker
Unique support structure
Vacuum case
Veto counter
Lower wake crates
Rich
ToF (s3, s4)
Y
X
Ecal
Short crate racks

Fig. 22.17 Schematic drawing of the AMS02 detector with a cutaway that allows a view on the inner parts. The tracker is indicated in the center with a few sheets, surrounded by a veto Counter and the magnet in its liquid He Vessel. Additional particle identification instruments are a time of flight (ToF) and a transition radiation detector (TRD) at the top (entrance side) and a ring imaging Cerenkov hodoscope (RICH), electromagnetic calorimeter Ecal and another ToF at the bottom (exit side). Reproduced from Haas (2004).

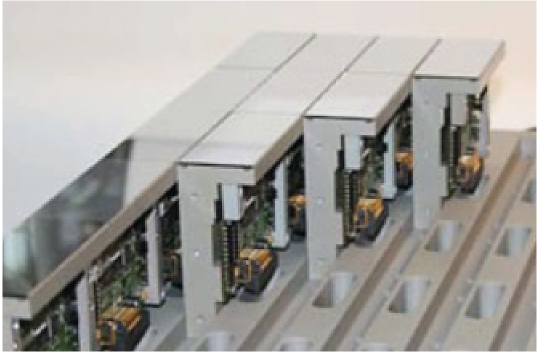

Fig. 22.18 Assembly of a number of "Pilatus" basic modules, which shows the structure that allows a large array to be constructed. A basic module carries a single sensor chip with 2 × 8 readout chips attached to it. The sensor faces the beam with the homogeneous rear side, while the segmented side is toward the readout chips. The dimensions of the sensor on top of the module on the rightmost rail are 84.5 mm × 33.5 mm (from *http://www.dectris.com/*).

22.5.3
Radiation Imaging Detectors in Industrial or Research Equipment

CCD have been in use for quite some time in research equipment and in medical and industrial X-ray imagers. Often the sensor is a scintillator with high-Z atomic number, in which the incident X-rays generate scintillation light. This light is transported and focused onto and detected by the CCD. It would be advantageous to replace such an indirect imaging by a sensor system that transforms the radiation directly into electrical signals. Silicon can be used for X-rays in the range 4–20 keV, and these quanta produce signals that are quite similar to those from mips. On the basis of the experience and advances in particle physics instrumentation, it is then even possible to implement X-ray imaging systems with single-quantum processing. Imaging with individual treatment of each photon still has to be studied in more detail to appreciate all the advantages. An early, striking example is the Medipix range of detectors, which is a spin-off from the Si pixel detector development for the Alice experiment at the LHC. This hybrid pixel assembly has a monolithic silicon sensor matrix of 256 × 256 cells of 55 μm × 55 μm, which is mated by solder bump bonding with a separate readout matrix in 0.25-μm CMOS technology. A sophisticated signal processing for individual, high-rate signals is incorporated in each pixel, such as was illustrated in Figure 22.11. In the most recent version Medipix3, now using 0.13-μm CMOS, one aims to acquire spectroscopic energy information on the individual photons while building up the X-ray image, as described by Ballabriga *et al.*, 2007. The previously mentioned Pilatus imager system is based also on this approach. For efficient absorption of energetic X-ray photons in the range 30–150 keV or even higher, it is needed to use different semiconductor materials such as CdTe.

A variety of applications can be thought of, including the use in visible light imaging with single-quantum photon detection. In this case, one uses a photocathode that produces a primary electron, which is then amplified in a gas-filled structure or in a vacuum tube with a microchannel plate or a kilovolt-accelerating field.

Glossary

Bias Voltage: The usual silicon particle sensor has the structure of a rectifying diode and needs a reverse bias voltage that depletes the sensor volume of free charge carriers.

CCD (Charge Coupled Device): Matrix of imaging cells consisting of capacitors with transparent polysilicon contact plates, so that light can enter the silicon and generate free charge carriers, proportional to the intensity.

Double-Sided Microstrip Detector: Semiconductor detector with a microstrip diode matrix on the rectifying front side, as well as on the ohmic rear side. Often the microstrip patterns make an angle of 90°. It may be useful to have a much smaller angle, such as 7°, which reduces the problem of ambiguities if several particles cross the detector matrix at the same time.

Drift Detector: Position-sensitive detector in which small readout electrodes are placed at the end of one or more carrier drift paths. These drift paths are defined by an electrode structure on the sensor surfaces, and on which a potential gradient is applied, typically $\sim 1\ \mathrm{V}\ \mu\mathrm{m}^{-1}$.

Efficiency: Detection efficiency is the percentage of particles or photon quanta that is detected in the instrument, as a fraction of the number of quanta that are incident on the sensitive detector area. This is equivalent to the probability that the signal charge generated by a particle in the detector volume can be distinguished above the noise charge.

Ion-Implanted Detector: Rectifying diode obtained by creating a p–n junction on a high-resistivity silicon slice, using an ion-implantation accelerator. Boron is often the species that is used to create the p-regions, and phosporous or arsenic is used to create the n+ ohmic rear contact of the detector diode.

Ladder: Linear assembly of two or more sensors on some substrate, often also including the signal readout circuits.

Leakage Current: Current that flows in a reverse-biased diode when no incident radiation is present.

Microstrip Detector: One-dimensional matrix of contiguous sensitive cells (strips) with long dimension above 1000 μm and short dimension (pitch) below 1000 μm.

Minimum Ionizing Particle (m.i.p. or mip): Swift particle with momentum at or above the value for which the specific energy loss reaches a minimum. The material surroundings of the detector modulate the actual value of the energy that is deposited in the sensitive volume. In a ''thin'' (<20-μ m) Si layer in vacuum the average specific energy deposit is $\sim 230\ \mathrm{eV}\ \mu\mathrm{m}^{-1}$, corresponding to an average generation of 63 electron–hole pairs. In an ''embedded, thick'' layer the values can be 360 eV and 99 e–h pairs on average.

Neutron-Transmutation-Doped (NTD) silicon: Under neutron irradiation, usually close to a nuclear reactor, silicon atoms can be converted into phosphorous atoms after neutron capture. This results in a very homogeneous n-type doping profile if the crystal is rotated in the neutron flux. The radiation defects have to be annealed and the doping atoms have to be electrically activated by a high-temperature treatment, usually above 1000 °C.

Noise Charge: The noise in the detector system is expressed as the equivalent noise charge (ENC) at the input of the amplifier in unit electron charges. The noise signal distribution is Gaussian and the width can be expressed by the rms. σ value or by the full width at half maximum (FWHM). The relation between these is $\mathrm{FWHM} = 2.35\sigma$.

Pad Detector: One-dimensional or two-dimensional matrix of contiguous sensitive cells with all lateral dimensions above 1000 μm.

Pixel Detector: Two-dimensional matrix of contiguous sensitive cells with dimensions below 1000 μm.

Surface Barrier Detector: Rectifying diode obtained by metal deposition on a high-resistivity silicon slice, which is used as particle detector.

Type Inversion: During prolonged irradiation of n-type silicon, the progressive introduction of acceptor-type radiation defects leads to effective compensation of the original type and the material exhibits properties of p-type silicon.

References

Antinori, F., Barberis, D., Beker, H., Beusch, W., Campbell, M., Cantatore, E., Catanesi, M.G., Chesi, E., Darbo, G., Di Bari, D., Di Liberto, S., Elia, D., Heijne, E.H.M., Helstrup, H., Jacholkowski, A., Klempt, W., Knudson, K., Kralik, I., Lassalle, J.C., Lenti, V., Lopez, L., Luptak, M., Martinengo, P., Meddi, F., Middelkamp, P., Morando, M., Pellegrini, F., Quercigh, E., Rossi, L., Safarik, K., Segato, G., and Simone, S. (**1995**) Development of silicon pixels for strangeness detection in nucleus-nucleus collisions, in *AIP Conference Proceedings* (ed. J. Rafelski), Conference of Strangeness in Matter, Vol. **340**, AIP, Tucson.

Baliga, B.J. (**1996**) *Power Semiconductor Devices*, PWS Publishing, Boston. isbn 0-534-94098-6.

Ballabriga, R., Campbell, M., Heijne, E.H.M., Llopart, X., and Tlustos, L. (**2007**) The Medipix3 prototype, a pixel readout chip working in single photon counting mode with improved spectrometric performance. *IEEE Trans. Nucl. Sci.*, **NS-54**, 1824.

Bertolini, G. and Coche, A. (**1968**) *Semiconductor Detectors*, North-Holland Publishing, Amsterdam.

Beuville, E., Borer, K., Chesi, E., Heijne, E.H.M., Jarron, P., Lisowski, B., and Singh, S. (**1990**) AMPLEX, a low- noise, low-power analog CMOS signal processor for multi-element silicon particle detectors. *Nucl. Instr. Meth.*, **A288**, 157. (5th Munich Symposium 1989).

Bichsel, H., Groom, D.E., and Klein, S.R. (**2006**) Passage of particles through matter, section 27, particle physics. *J. Phys. G*, **33**, 1, *http://pdg.lbl.gov*

Damerell, C.J.S., English, R.L., Gillman, A.R., Lintern, A.L., Wickens, F.J., and Watts, S.J. (**1986**) Use of charge-coupled devices as high precision detectors. *IEEE Trans. Nucl. Sci.*, **NS-33**, 51. (1985 Nuclear Science Symposium, San Francisco).

Gatti, E., Padovini, G., and Radeka, V. (**1982**) Signal evaluation in multielectrode radiation detectors by means of a time dependent weighting vector. *Nucl. Instr. Meth.*, **193**, 651.

Gatti, E. and Rehak, P. (**1984**) Semiconductor drift chamber - an application of a novel charge transport scheme. *Nucl. Instr. Meth.*, **225**, 608.

Ghandhi, S.K. (**1977**) *Semiconductor Power Devices*, John Wiley & Sons, Inc., New York. isbn 0-471-02999-8.

Grove, A.S. (**1967**) *Physics and Technology of Semiconductor Devices*, John Wiley & Sons, Inc., New York.

Haas, D. (**2004**) The silicon tracker of AMS02. *Nucl. Instr. Meth.*, **A530**, 173. Firenze 6th 2003.

Hanagaki, K. (**2004**) Design of an upgraded Dø silicon microstrip tracker for Run IIb at the tevatron. *Nucl. Instr. Meth.*, **A530**, 12. Firenze 6th 2003.

Heijne, Erik H. M. (**2008**) 1980 a revolution in silicon detectors, from energy spectrometer to radiation imager: some technical and historical details. *Nucl. Instr. Meth.*, **A591**(1), 6.

Heijne, Erik.M.H. (**1983**) Muon Flux Measurement with Silicon Detectors in the CERN Neutrino Beams, Yellow Report *CERN* 83-06 21, July 1983.

Heijne, Erik.H.M. and Jarron, P. A low noise CMOS integrated signal processor for multi-element particle detectors. *Proceedings of the ESSCIRC'88 Conference*, UMIST, Manchester, p. 66.

Huff, H.R., Kriegler, R.J., and Takeishi, Y. (eds) (**1981**) *Semiconductor Silicon*, Vol. **81-5**, The Electrochemical Society.

Ionica, M. (**2003**) Experience in the assembly of the large area silicon tracker for the AMS experiment. *Nucl. Instr. Meth.*, **A513**, 222. PSD-6 Leicester.

Jacoboni, C., Canali, C., Ottaviani, G., and Alberigi Quaranta, A. (**1977**) A review of some charge transport properties of silicon. *Solid-St. Electron.*, **20**, 77.

Kemmer, J. (**1980**) Fabrication of low noise silicon radiation detectors by the planar process. *Nucl. Instr. Meth.*, **169**, 499.

Kemmer, J. (**1980b**) Passivated semiconductor PN junction of high electric strength and process for the production thereof. US Patent 4,442,592, filed Jan. 14 1981 (DE 31 jan 1980).

Kleinfelder, S.A., Carithers, W.C. Jr, Ely, R.P. Jr, Haber, C., Kirsten, F., and Spieler, H.G. (**1988**) A flexible 128 channel silicon strip detector instrumentation integrated circuit with sparse data readout. *IEEE Trans. Nucl. Sci.*, **NS-35**, 171.

Knoll, G.F. (**2000**) *Radiation Detection and Measurement*, 3rd edn, John Wiley & Sons, Inc., New York. isbn 0-471-07338-5.

Koerts, L.A.Ch., Mulder, K., Oberski, J.E.J., and van Dantzig, R. (**1971**) The 'BOL' nuclear research project. *Nucl. Instr. Meth.*, **92**, 157.

LaBarga, L., Adolphsen, C., Gratta, G., Litke, A., Turala, M., Zaccardelli, C., Breakstone, A., Parker, S., Barnett, B., Dauncey, P., Drewer, D., Matthews, J., Jacobsen, R., and Lüth, V. (**1991**) The Mark II silicon strip vertex detector and performance of a silicon detector telescope in the Mark II detector at the SLC. *IEEE Trans. Nucl. Sci.*, **NS-38**, 25.

Laegsgaard, E. (**1979**) Position-sensitive semiconductor detectors. *Nucl. Instr. Meth.*, **162**, 93.

Llopart, X., Ballabriga, R., Campbell, M., Tlustos, L., and Wong, W. (**2007**) Timepix, a 65k programmable pixel readout chip for arrival time, energy and/or photon counting measurements. *Nucl.Instr. Meth.*, **A581**, 485.

Lutz, G. (**1999**) *Semiconductor Radiation Detectors, Device Physics*, Springer. isbn 3-540-64859-3.

Lutz, G., Maloberti, F., Manfredi, P.F., Re, V., Speziali, V., Buttler, W., and Vogt, H. (**1988**) On the design of a JFET-CMOS front-end for low noise data acquisition from microstrip detectors. *Nucl. Instr. Meth.*, **A264**, 391.

Nicollian, E.H. and Brews, J.R. (**1982**) *MOS physics and technology*, Wiley, New York.

Nouais, D., Beolè, S., Bondila, M., Bonvicini, V., Cerello, P., Crescio, E., Giubellino, P., Idzik, M., Kolozhvari, A., Kouchpil, S., Lopez Torres, E., Martinez, M.I., Mazza, G., Piano, S., Piemonte, C., Rashevsky, A., Riccati, L., Rivetti, A., Tosello, F., Trzaska, W.H., Vacchi, A., and Wheadon, R. (**2003**) The ALICE Silicon Drift Detector system. *Nucl. Instr. Meth.*, **A501**, 119. Vertex 2001, Brunnen.

Oshiyama, A., Maeda, K., Itoh, K.M., and Katayama-Yoshida, H. (**2005**) *Proceedings of 23rd International Conference "Defects in Semiconductors"*, Elsevier, Awaji Island; (**2006**) *Physica B*, 376–377.

Plummer, J.D., Deal, M.D., and Griffin, P.B. (**2000**) *Silicon VLSI Technology: Fundamentals, Practice and Modeling*, Prentice Hall. isbn 0-13-085037-3.

Pooth, O. (**2006**) The CMS silicon tracker. *Nucl. Instr. Meth.*, **A569**, 21. Vertex2005.

Report on EUDET-Calice Collaboration arxiv.org/pdf/0707.1245.

Rossi, L., Fischer, P., Rohe, T., and Wermes, N. (**2006**) *Pixel Detectors, From Fundamentals to Applications*, Springer Verlag. isbn-13 978-3-540-28332-4.

Rossi, G., Renzi, M., Eikenberry, E.F., Tate, M.W., Bilderback, D., Fontes, E., Wixred, R., Barna, S., and Gruner, S.M. (**1999**) Test of a prototype pixel array detector for microsecond time-resolved X-ray diffraction. *J. Synchrotron Radiat.*, **6**, 1096.

Sadrozinski, H.F.-W. (**1998**) The GLAST instrument, a gamma-ray large area space telescope. *Nucl. Instr. Meth.*, **A419**, 300. Wire Chamber 98 Vienna.

Spieler, H. (**1994**) Analog front-end electronics for the SDC silicon tracker. *Nucl. Instr. Meth.*, **A342**, 205. Hiroshima STD Symposium.

Spieler, H. (**2005**) *Semiconductor Detector Systems*, Oxford Science Publications, Oxford. isbn-13: 978-90-19-852784-8.

Straulino, S., Adriani, O., Bonechi, L., Bongi, M., Castellini, G., D'Alessandro, R., Gabbanini, A., Grandi, M., Papini, P., Ricciarini, S.B., Spillantini, P., Taccetti, F., Tesi, M., and Vannuccini, E. (**2004**) The PAMELA silicon tracker. *Nucl. Instr. Meth.*, **A530**, 168. Firenze 6th 2003.

(a) Sze, S.M. (**1981**) *Physics of Semiconductor Devices*, 2nd edn Wiley-Interscience, New York. isbn 0-471-09837-X;

(b) Sze, S.M. and Ng, K.K. (**2007**) *Physics of Semiconductor Devices*, 3rd edn, Wiley-Interscience, Hoboken. isbn 978-0-471-14323-9.

Tajima, H. (**2006**) for the GLAST Tracker Team, GLAST Tracker. *Nucl. Instr. Meth.*, **A569**, 140. Vertex 2005.

Tlustos, L. (**2005**) Performance and Limitations of High Granularity Single Photon Processing X-ray Imaging Detectors, PhD Thesis, dir H. Aiginger Wien TU, CERN-Thesis-2005-032 [*http://cdsweb.cern.ch/record/846447/files/thesis-2005-032.pdf*]

Viertel, G.M. and Capell, M. (**1998**) The ALPHA magnetic spectrometer. *Nucl. Instr. Meth.*, **A419**, 295. Wire Chamber 98 Vienna.

Walker, J.T., Parker, S., Hyams, B.D., and Shapiro, S.L. (**1984**) Development of high density readout for silicon strip detectors. *Nucl. Instr. Meth.*, **226**, 200. (3rd Munich Symposium 1983).

Watts, S.J. (**1997**) Overview of radiation damage in silicon detectors–models and defect engineering. *Nucl. Instr. Meth.*, **A386**, 149. Workshop Vertex '96.

DIALYSIS

ROBERT H. BARTH, M. D., *Department of Veterans' Affairs Medical Center, Brooklyn, New York, U.S.A. and State University of New York, Health Science Center at Brooklyn, Brooklyn, New York, U.S.A.*

	Introduction: The Concept of Dialysis	533
	Generalities	533
	Applications	534
	Ion-Exchange Dialysis Methods	534
	Hemodialysis as a Paradigm	535
1.	**Functions of the Kidney: The Goals of Dialysis Therapy**	535
2.	**History**	537
3.	**The Kinetics of Dialysis**	538
3.1	Dialyzer Mass Transfer	538
3.2	Dialysance	539
3.3	Ultrafiltration	540
3.3.1	Convective Clearance	541
3.3.2	Backfiltration	542
4.	**Components of the Dialysis System**	542
4.1	The Dialyzer	542
4.1.1	Dialyzer Geometry	543
4.1.1.1	Coil Dialyzers	543
4.1.1.2	The Parallel Plate	544
4.1.1.3	Hollow-Fiber Dialyzers	545
4.1.2	Membranes	545
4.1.2.1	Classification	545
4.1.2.2	Transport Characteristics	547
4.1.2.3	Biocompatibility	547
4.2	Blood	548
4.2.1	Protein	548
4.2.2	Blood Cells	549
4.3	Dialysis Fluid	550
4.3.1	Composition	550
4.3.2	Water Preparation	551
4.4	Dialysis Monitoring and Delivery Systems	551
4.4.1	Functions	551
4.4.2	The Blood Circuit	551
4.4.3	The Dialysate Circuit	552
4.4.4	Ultrafiltration Systems	552
5.	**The Future**	554
	Glossary	554
	Works Cited	555
	Further Reading	555

INTRODUCTION

Generalities

Dialysis is a process whereby solutes of differing molecular weights are separated according to their varying abilities to diffuse through a membrane. The driving force is the chemical gradient across the membrane for each solute, with no convective exchange or movement of solvent through the membrane. Diffusion is spontaneous and irreversible, leading to an increase in the entropy of the system, and requires no external source of thermal or chemical energy. Dialysis is used in the laboratory, in medicine, and in industry as a separation or concentration process.

The word was coined and the process described during the 1850s by the Scottish chemist Thomas Graham. Graham recognized that certain substances, which he termed "colloids"—probably largely proteins—diffused much more slowly across a parchment membrane than other smaller molecular weight solutes that he called "crystalloids." Using a simple device with a "dialytic septum" of albumin-coated parchment paper, he was able to separate the components of urine, recovering a crystalloid in the diffusate that proved to be primarily urea.

There is little standardization of terminology. The solution on the feed side, with retained macromolecules, may be termed retentate, while the product solution containing diffused solutes is called permeate or diffusate. The term dialysate has been used by different authors for solution on both sides of the membrane. In hemodialysis, however, which is the most widespread application of the process, the feed solution is blood and the

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

solution containing diffused solutes is always called dialysate.

The direct descendant of Graham's method is *equilibrium dialysis,* a laboratory procedure used primarily for purification of proteins, illustrated in Fig. 1. In this process, an aqueous solution containing both small molecules and macromolecules is placed in a collodion or cellophane bag, and the sealed bag placed in a large reservoir of buffer solution. The small molecules diffuse out of the bag along a concentration gradient, while the macromolecules are retained by the bag membrane. If the volume of buffer is large with respect to that of the bag—with, for example, continuous flux of fresh buffer—a relatively complete separation can be effected.

Applications

Industrial applications of dialysis include the recovery of caustic from hemicellulose in the production of rayon, salt removal in pharmaceuticals manufacture, and desalination of water, but the most extensive commercial use is for the medical treatment of end-stage renal disease. In patients with this condition, dialysis substitutes for the failed excretory function of the kidney, removing from the circulation the small [<1000 Da; 1 Da (dalton) $=1.007\,94$ u] and middle (~ 1000–5000 Da) molecular-weight metabolic end products that cause toxic symptoms if allowed to accumulate, while leaving behind proteins and other macromolecules. In 1988, over 110 000 patients received regular dialysis therapy in the United States alone, at a total cost of more than five billion dollars (United States Renal Data System, 1990).

Ion-Exchange Dialysis Methods

The retention of large solutes is ordinarily based on steric effects. The use of ion-selective membranes, which allow preferential passage of positively charged cations or negatively charged anions, has allowed the development of several specialized techniques. In *ion-exchange dialysis,* shown in Fig. 2(a), an anion-exchange membrane allows passage of anions, but excludes transfer of all cations except hydrogen ions. Thus if a solution of an acid and its metal salt is dialyzed against water, the anion diffuses along a concentration gradient and is accompanied by hydrogen ions to preserve electroneutrality. The net effect is separation of the acid from the metal salt, and the process is used for precisely this purpose in the recovery of metal ions in the electroplating industry.

Donnan dialysis, shown in Fig. 2(b), uses a cation-exchange membrane to concentrate a positive ion such as the cuprous ion shown in the figure. Hydrogen ions diffuse down a concentration gradient from the highly concentrated acid solution on the right side of the membrane. Since the membrane prevents diffusion of negative ions, the only way to maintain electroneutrality is to transfer the positive cuprous ion out of the left compartment against a concentration gradient, with resultant increase in concentration of the valuable metal ion by as much as one-hundredfold.

In *electrodialysis,* dialysis takes place in the presence of an electric field. The driving force for solute movement is the electrical attraction of anions to the positive anode and cations to the negative cathode. An array of alternating ion-selective membranes produces concentration and dilution of the ions in alternate cells of the dialyzer [Fig. 2(c)], and the process may be used to remove salts effectively from brackish water. Electrodialy-

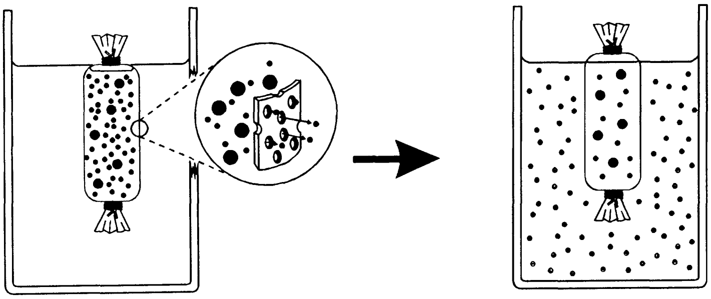

FIG. 1. Equilibrium dialysis. In this laboratory procedure, small and large molecules are separated by their differing ability to diffuse across a membrane.

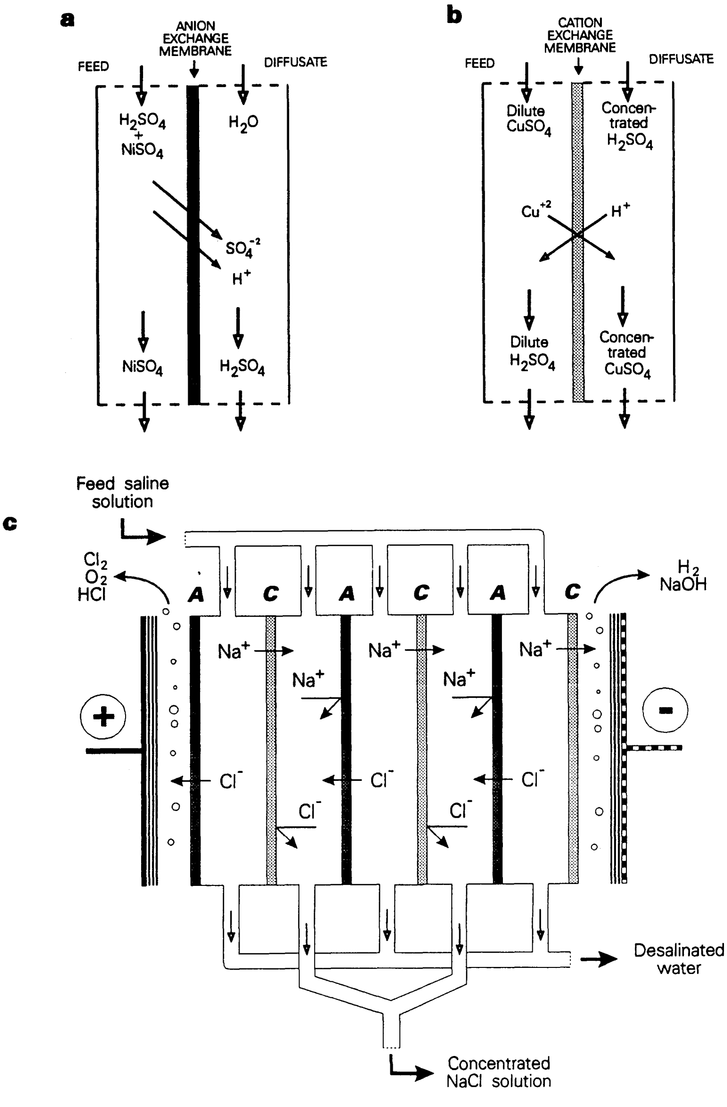
a
ANION EXCHANGE MEMBRANE
FEED
DIFFUSATE
H2SO4 + NiSO4
H2O
SO4-2
H+
NiSO4
H2SO4
b
CATION EXCHANGE MEMBRANE
FEED
DIFFUSATE
Dilute CuSO4
Concentrated H2SO4
Cu+2
H+
Dilute H2SO4
Concentrated CuSO4
c
Feed saline solution
Cl2
O2
HCl
A
C
Na+
Cl-
H2
NaOH
+
-
Desalinated water
Concentrated NaCl solution

FIG. 2. Ion-exchange dialysis techniques. (a) Ion-exchange dialysis. (b) Donnan dialysis. (c) Electrodialysis; ***A***=anion exchange membranes, ***C***=cation exchange membranes. See text for description.

sis is widely used in the preparation of potable water from saline sources.

Hemodialysis as a Paradigm

Therapeutic hemodialysis represents the most highly developed and by far the commonest and most economically important use of the procedure, and has been subjected to the most extensive mathematical analysis. Because of this, and since much of that analysis may be generalized to any dialysis procedure, the remainder of this article will use hemodialysis as a paradigm. After a discussion of the functions of the kidney that hemodialysis must replace, the mathematics of mass transfer and apparatus and techniques will be examined in depth.

1. FUNCTIONS OF THE KIDNEY: THE GOALS OF DIALYSIS THERAPY

Therapeutic dialysis is a "renal replacement therapy," intended to substitute for the normal functions of the kidney that are lost in end-stage renal failure. The dialyzer is thus often referred to as an "artificial kidney," although its simulation of kidney function is incomplete and relatively inefficient.

The physiological role of the kidney includes regulation of body water and electrolytes, excretion of metabolic wastes, and regulation of metabolism through the production of a number of hormones. The first two of these are separation processes, although not the functions of a passive filter. The anatomical unit that is responsible for the formation of urine and adjustment of its solute concentrations is the *nephron,* shown schematically in Fig. 3. Each human kidney contains $(0.8-1.2)\times 10^6$ of these structures, which consist of a tuft of blood capillaries, the *glomerulus,* in intimate contact with a long duct, the *tubule,* which is lined with epithelial cells. Ultrafiltration takes place across the glomerular basement membrane, which divides the capillary lumen from that of the tubule. This membrane is quite permeable to water and to solutes of molecular weight less than about 10,000 Da, but provides an effective barrier to passage of macromolecules over 50,000 Da, especially those that carry a net negative charge at the physiologic *p*H of 7.40.

The rate of glomerular ultrafiltration is determined by the hydrostatic pressure difference between the glomerular capillary and tubular lumina, $P_{GC}-P_T$, and by the osmotic gradient produced by the retained macromolecules, $\pi_{GC}-\pi_T$, where π is the colloid osmotic pressure. This relationship can be expressed as

$$\mathrm{SNGFR}=K_f[(P_{GC}-P_T)-(\pi_{GC}-\pi_T)], \qquad (1)$$

where SNGFR is the ultrafiltration rate of a single nephron and K_f is the glomerular ultrafiltration coefficient, in turn the product of membrane hydraulic permeability and surface area. Using data obtained from rat glomeruli, Dworkin and Brenner (1985) have calculated K_f to be 42.1 nL/(s·cm^2·mm Hg) (1 mm Hg = 133 Pa), a value that is one to two orders of magnitude greater than that of other capillaries. Glomerular filtration thus proceeds rapidly despite a relatively low hydro-

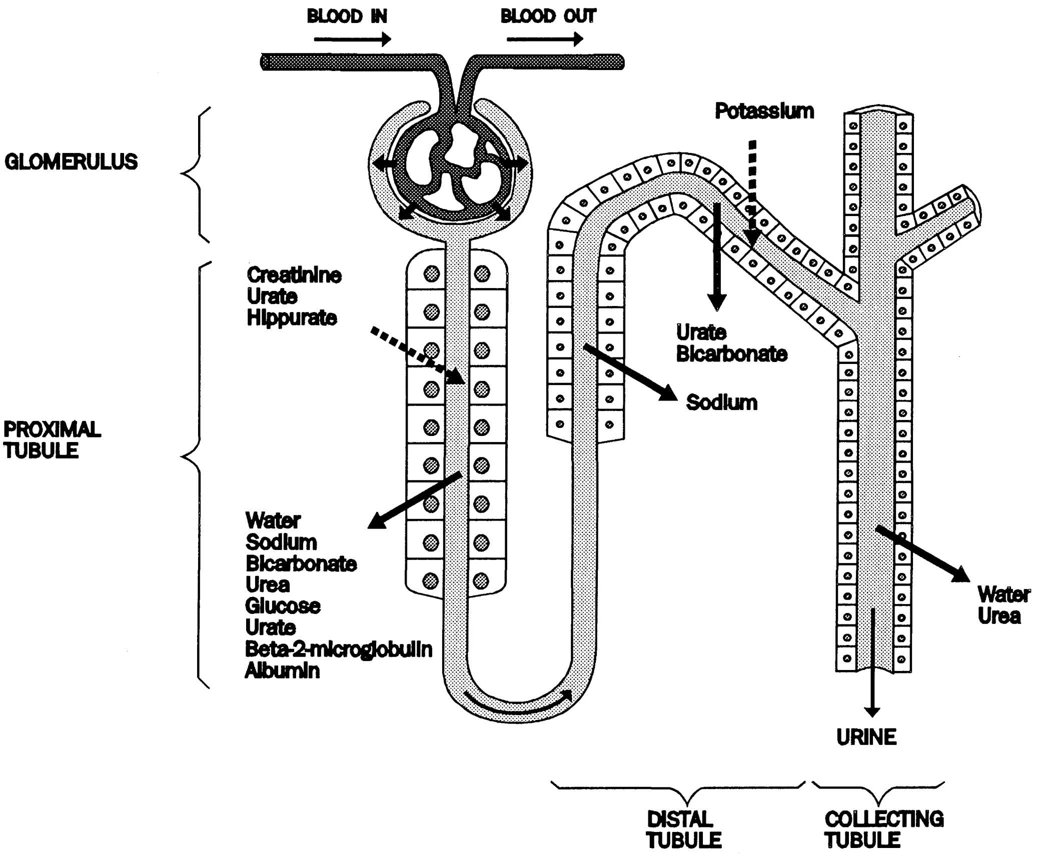
BLOOD IN
BLOOD OUT
GLOMERULUS
PROXIMAL TUBULE
Creatinine
Urate
Hippurate
Water
Sodium
Bicarbonate
Urea
Glucose
Urate
Beta-2-microglobulin
Albumin
Potassium
Urate
Bicarbonate
Sodium
Water
Urea
URINE
DISTAL TUBULE
COLLECTING TUBULE

FIG. 3. The human nephron. Arrows indicate sites of secretion into the tubule (broken arrows) or reabsorption into the blood (solid arrows) of water and solutes.

static driving force in the range of 5–10 mm Hg, reaching levels of 120–130 mL/min, or about 180 L/d.

The ultrafiltrate formed by the glomerulus is practically free of protein, but contains small-molecular-weight solutes in the same concentrations found in plasma. As this solution travels down the tubule, about 99% of the filtered water is reabsorbed into the circulation. Some filtered solutes, such as inulin, a 5200-Da polysaccharide used in the measurement of renal function, remain entirely within the tubule and become progressively more concentrated as water is removed. Other solutes, including sodium, potassium, phosphate, bicarbonate, small proteins, and essential nutrients such as glucose and amino acids, are themselves reabsorbed, so that less than 1% of the solute filtered by the glomerulus is ultimately excreted in the urine (that is, the *fractional solute excretion* is less than 0.01). Finally, potentially toxic organic acids and bases such as hippurate and creatinine are secreted into the tubular fluid, so that their fractional excretion may exceed 1.0. Sites along the nephron where typical solutes are reabsorbed or secreted are indicated in Fig. 3.

By varying the rates of reabsorption and secretion, the tubule adjusts solute excretion to the current needs of the organism. At extremely low levels of salt intake, for example, or in states of dehydration, when there is need to minimize sodium loss, the urine sodium concentration may be virtually unmeasurable. During heavy salt ingestion, however, as much as 4 or 5 g of sodium per day may appear in the urine.

The kidney's ability to excrete a solute x may be represented by the *clearance* of x (C_x), defined as the mass removal rate divided by the plasma concentration. Clearance may also be conceptualized as the volume of plasma required to supply the amount of x appearing in the urine in a given time, usually expressed in mL/min,

$$C_x = U_x \dot{V} / P_x, \tag{2}$$

where U_x and P_x are the concentrations of x in the urine and plasma and $\dot{V}$ is the urine flow rate. Since the rate of excretion of x is equal to $U_x \dot{V}$, and rate of filtration of x is equal to the product of P_x and the glomerular filtration rate (GFR), the fractional excretion of x (FE_x) is

$$\mathrm{FE}_x = U_x \dot{V} / (P_x \times \mathrm{GFR}) \tag{3}$$

and is equal to the ratio of C_x to GFR. Thus, if x is freely filtered at the glomerulus and is neither secreted nor reabsorbed, then its rate of filtration is equal to its rate of excretion, $\mathrm{FE}_x = 1.0$, and C_x is equal to GFR. In clinical practice, the clearance of *creatinine,* a product of muscle metabolism, is generally used as an estimate of GFR, despite the fact that some creatinine is secreted into the tubule and $\mathrm{FE}_{\mathrm{creatinine}}$ is slightly greater than 1.0.

2. HISTORY

In 1913, a little more than fifty years after Graham's first description of dialysis, *in vivo* dialysis was first performed at Johns Hopkins University, and the term "artificial kidney" entered the language. Abel, Rowntree, and Turner passed the blood of nephrectomized dogs through collodion tubes and demonstrated the removal of nonprotein nitrogen. In 1924 Georg Haas, in Germany, performed the first human dialysis, which lasted 15 min and demonstrated primarily that the procedure could be done without harm to the patient.

Two developments of the next decade had profound influence on the nascent hemodialysis. In 1928 the first practical anticoagulant—heparin—became available, making prevention of the coagulation of blood passing through the dialysis apparatus much more manageable. Secondly, cellophane membranes were developed and manufactured commercially, largely for sausage casings. This regenerated cellulose product was far more practical than the fragile collodion for the construction of dialyzers, and in slightly modified forms remains today the mainstay of hemodialyzer manufacture.

Clinical dialysis became a reality in the small town of Kampen, during the Second World War German occupation of the Netherlands. Using a large rotating drum wound with 2.5-in. cellophane sausage tubing and immersed in a 70-L tank of dialysate, Willem Kolff began in 1943 to dialyze patients suffering from uremia, and in 1945 succeeded in keeping a woman alive until her recovery from acute renal failure. After the war Kolff donated machines to clinics in Europe and the United States, and hemodialysis began slowly to grow. By the late 1950's it was an established therapy for acute renal failure, but long-term dialysis was not possible because of the extreme difficulty of repeated cannulation of a patient's arteries and veins.

The breakthrough came in 1960 with the development of the Quinton-Scribner shunt—a Teflon, and later Silastic, cannula that was implanted in an artery and vein in the arm and afforded permanent vascular access. This device was still fraught with problems of infection and coagulation of blood within its tubing, but it eliminated the need for repeated surgical cutdowns, and made chronic hemodialysis a possibility for the first time. Six years later Brescia, Cimino, and their colleagues in the Bronx surgically created an arteriovenous fistula in the wrist, and the resultant high flow and pressure in the draining cephalic vein caused it to dilate and its wall to become thickened. This dilated vein could then be accessed by venipuncture with a large needle, providing sufficient blood flow for effective dialysis. Routine intermittent access to the circulation became a reality, and the modern era of chronic dialysis was established.

Over the scant three decades of its existence as a practical treatment for kidney failure, dialysis has been a remarkable success. In the United States alone, well over 100 000 people receive regular dialysis treatment. The 1-yr mortality rate in industrialized countries is high, ranging from 7% to 23%, but many patients with renal failure, once doomed to a rapid demise from their disease, now continue to live independent and active lives for as much as 20 yr or more.

3. THE KINETICS OF DIALYSIS

Therapeutic dialysis replaces the filtration functions of the kidney—specifically, the establishment and maintenance of physiological solute concentrations and normal body water content. The basic elements, as illustrated in Fig. 4, are simple, with solution (blood) flowing on one side of a semipermeable membrane, and dialysate (an osmotically balanced electrolyte solution) flowing in a countercurrent direction on the other. Solutes diffuse down concentration gradients, and water removal is effected by ultrafiltration, driven by the hydrostatic pressure difference between solution (blood) and dialysate. A mathematical analysis of the process can be divided into several parts: mass transfer characteristics of the artificial kidney, or dialyzer; ultrafiltration; and the kinetics of solute removal from the body. The last of these will not be treated in this article.

3.1 Dialyzer Mass Transfer

Solute transfer across dialysis membranes occurs by diffusion, and is governed by Fick's law

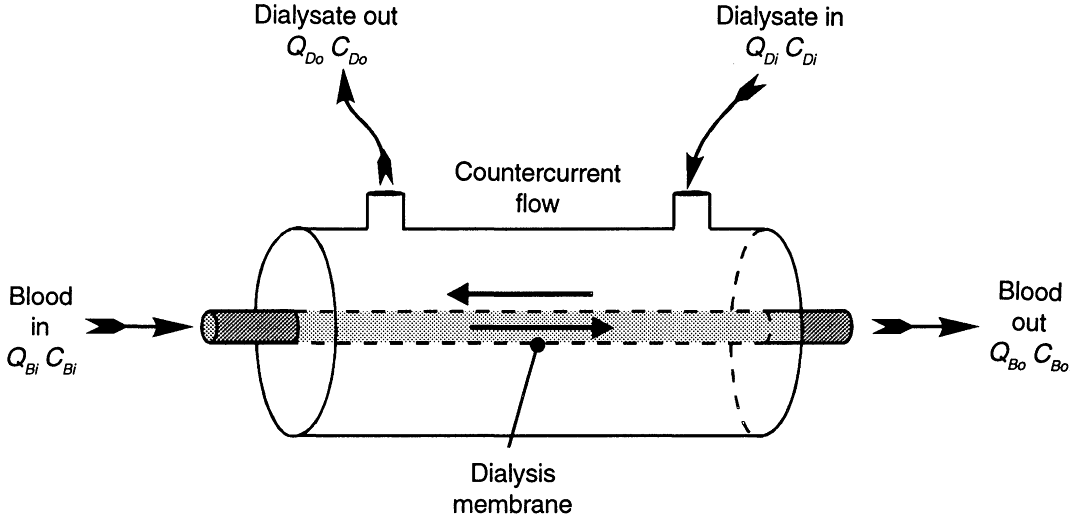

FIG. 4. Schematic representation of an artificial kidney. Blood and dialysate flow in opposite, or countercurrent, direction. Q=flow, C=concentration, Bi=blood inlet, Bo=blood outlet, Di=dialysate inlet, Do=dialysate outlet.

$$J_x = -D_x A \frac{dc}{dx} \qquad (4)$$

or in integrated form

$$J_x = -(D_x A/L)(c_B - c_D), \qquad (5)$$

which states that flux of solute x—J_x (mol/cm^2/s)—proceeds over the distance L at a rate proportional to the concentration difference from solution (blood) to dialysate, $c_B - c_D$, the surface area of the diffusing front, A, and the diffusivity, D_x, which is a constant property of the solute-solvent system at a given temperature. In a dialyzer D_x is constant for a given temperature and solute, and L is determined by membrane thickness and dialyzer geometry and is also constant, so that for a specified solute D_x and L may be combined in the overall mass transfer coefficient, k_0:

$$J_x = -k_0 A (c_B - c_D)_{\text{avg}}. \qquad (6)$$

k_0 is a complex coefficient with components that depend on transport in solution (blood), membrane, and dialysate. Each of these phases has an associated mass transfer coefficient and reciprocal mass transfer resistance, R, related by

$$1/k_0 = R_0 = R_B + R_M + R_D, \qquad (7)$$

$$1/k_0 = 1/k_B + 1/P_M + 1/k_D, \qquad (8)$$

where the subscripts B, M, and D refer to blood, membrane, and dialysate, and P_M is membrane permeability. In some early hemodialyzers, thick and poorly distributed blood and dialysate boundary layers contributed as much as 75–80% of the total resistance to solute transfer. With modern dialyzer design techniques, this figure has fallen to less than 25% (Colton and Lowne, 1981).

Strictly speaking, k_0 is a local property of a point on the dialysis membrane, since k_B and k_D may vary as concentrations change with position along the dialyzer. In practice another value, K_0, is used, defined as the length-averaged value of k_0, and considered constant throughout the dialyzer. The product of K_0 and the membrane surface area, K_0A, is a fixed property of a dialyzer for a given solute (assuming aqueous solution and a fixed temperature, usually 37 °C), and expresses the dialyzer's intrinsic ability to filter that solute.

3.2 Dialysance

Potential dialyzer performance may be expressed by K_0, but in practice it is necessary to quantify the actual amount of filtration of specific solutes taking place under clinical conditions, that is, at known blood and dialysate flow rates. Diffusive dialysance D, the rate constant used to characterize dialyzer solute removal, was defined by Wolff (1951) as "the minute rate of net exchange of a substance between blood and bath fluid per unit blood-bath concentration gradient," or

$$D = \frac{\text{Rate of solute transfer}}{\text{Unit concentration driving force}}. \qquad (9)$$

The concentration driving force appears in Eq. (6) as $(c_B - c_D)_{\text{avg}}$. The correct average concentration difference is the logarithmic mean of the inlet and outlet concentrations:

$$(c_B - c_D)_{lm} = \frac{(c_B - c_D)_1 - (c_B - c_D)_2}{\ln[(c_B - c_D)_1/(c_B - c_D)_2]}, \qquad (10)$$

where 1 and 2 are the two ends of the dialyzer. Nevertheless, because of the inconvenience of measuring inlet and outlet concentrations in both dialysate and blood, in the clinical setting the driving force is approximated by the maximum concentration difference, $c_{Bi} - c_{Di}$, using the terminology of Fig. 4.

The mass transfer rate N, in moles per unit time, is given by the difference between incoming and outgoing solute mass (either in blood or in dialysate):

$$N = Q_{Bi} c_{Bi} - Q_{Bo} c_{Bo} = Q_{Do} c_{Do} - Q_{Di} c_{Di}. \qquad (11)$$

Since in the absence of ultrafiltration $Q_{Bi} = Q_{Bo}$ and $Q_{Di} = Q_{Do}$, dialysance may be written as

$$D_B = \frac{Q_B(c_{Bi} - c_{Bo})}{c_{Bi} - c_{Di}} = \frac{Q_D(c_{Do} - c_{Di})}{c_{Bi} - c_{Di}}. \qquad (12)$$

In modern hemodialyzers, blood and dialysate flows are always countercurrent, and for solutes like urea, creatinine, and phosphate, concentrations in the entering dialysate are 0. In this case, the denominator of Eq. (12) simplifies further to c_{Bi}, and the equation becomes the equivalent of Eq. (2), the definition of clearance for the natural kidney. By this analogy, the rate constant for transfer of solutes not present in entering dialysate is known as dialyzer clearance (K):

$$K = \frac{Q_B(c_{Bi} - c_{Bo})}{c_{Bi}} = \frac{Q_D(c_{Do} - c_{Di})}{c_{Bi}}. \qquad (13)$$

Both dialysance and clearance may be calculated using either blood- or dialysate-side flow rates and concentrations, although in clinical practice the former is more common.

Dialysance is dependent on momentary conditions, that is, blood and dialysate flow rates, while K_0A is a fixed property of a dialyzer for each solute examined. The relationship between K_0A and dialysance is thus an important and useful one, necessary for prediction of dialyzer performance under changing conditions of flow. Equating J as defined in Eqs. (6) and (11), assuming countercurrent blood-dialysate flow and no net ultrafiltration, and substituting with Eqs. (10) and (12), yields

$$D = \frac{e^{-z} - 1}{e^{-z}/Q_B - 1/Q_D}, \qquad (14)$$

where

$$z = \frac{K_0A(1 - Q_B/Q_D)}{Q_B}.$$

Equation (14) is plotted in Fig. 5 for a range of K_0A that varies from 50—a typical value for a large-molecular-weight solute such as vitamin B_{12}—to 900 mL/min, which is about the highest value of K_0A for the small molecule urea obtainable with modern high-performance dialyzers. Examination of this graph—which is calculated for a fixed dialysate flow of 800 mL/min—reveals the boundary conditions for dialysance. When K_0A/Q_B is large, as for small solutes at low blood flow rates, then $D \approx Q_B$. Conversely, when K_0A/Q_B is small, either because the solute is large and poorly diffusible or when blood flows are very high, then the limiting value is $D \approx K_0A$. Clinically attainable blood flows range from 200 to 600 mL/min; in this range small-molecular-weight solute dialysance is flow dependent with most dialyzers. For molecules larger than about 500 Da, increasing blood flow beyond 200 mL/min has little effect on dialysance in conventional cellulose dialyzers. Thus, to increase removal from the circulation of high-molecular-weight solutes, it is necessary to enlarge the surface area of the dialyzer or lengthen the treatment. The situation may, however, be different when newer, more highly permeable membranes, which have higher K_0A's for larger molecules, are used.

3.3 Ultrafiltration

Besides removal of waste solutes, therapeutic dialysis also has the task of removing excess water accumulated by ingestion and by

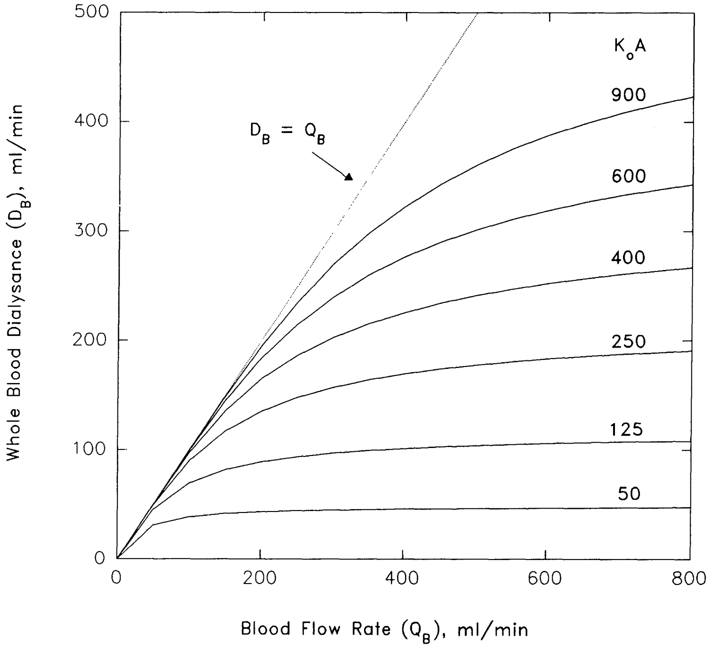

FIG. 5. Dependence of dialysance on blood flow rate, plotted for different values of K_0A.

metabolic formation, often as much as 1–2 L/d. This is accomplished by ultrafiltration across the dialysis membrane, a process described by

$$J_F = \frac{Q_F}{A} = \frac{L_p(\Delta P - \Delta\pi)}{\mu}, \tag{16}$$

where J_F is the water flux (mL/cm^2/min), Q_F is the ultrafiltrate flow rate, A the membrane surface area, L_p the membrane hydraulic permeability coefficient, and μ the filtrate viscosity. The driving force is the sum of the hydrostatic and osmotic pressure gradients. In hemodialysis the transmembrane hydrostatic pressure (TMP) from blood to dialysate varies from 100 to 500 mm Hg, and almost invariably markedly exceeds the osmotic pressure generated by large molecules on the blood side. In clinical practice, an empirically defined ultrafiltration coefficient, K_{UF} [mL/(mm Hg)h], is used to describe dialyzer performance:

$$K_{UF} = \frac{\text{Water flux (mL/h)}}{\text{Transmembrane pressure (mm Hg)}} = Q_F/(P_B - P_D). \tag{17}$$

Table 1 includes approximate ultrafiltration coefficients for a number of membrane materials, with kidney glomerular basement membrane added for comparison.

3.3.1 Convective Clearance. Frictional interaction between water and solute molecules—so-called solvent drag—results in transport of solute across the dialysis membrane with the ultrafiltrate. The restraining action of the membrane to this convective solute movement is expressed as the sieving coefficient S of a membrane for a given solute:

$$S = c_F/c_B. \tag{18}$$

When $S=1$, there is no resistance to solute movement, and the concentration in the filtrate will be unchanged from that of blood. With cuprophane membranes, there is a rapid falloff in permeability with increasing molecular weight, but more "open" membranes such as sulfonated polyacrylonitrile (AN69) have sieving coefficients close to unity for molecules as large as 1500 Da (Table 2) (Green *et al.*, 1976). Since diffusive clearance of these larger molecules is low, convection plays a relatively important part in the transfer of these substances across open membranes.

When ultrafiltration is occurring,

$$Q_{Bo} = Q_{Bi} - Q_F \tag{19}$$

Table 1. Hemodialysis hollow-fiber membrane materials, with glomerular basement membrane for comparison.

Membrane	Synonyms	Membrane thickness (μm)	K_{UF}[a]	Complement activation (%)[b]
Cellulosic				
Cuprammonium cellulose	Cuprophane	8–16	3–5	100
Saponified cellulose ester	SCE	25	3–4	100
Modified Cellulose				
Cellulose acetate	CA	40	4–10	48–67
Diethylaminoethyl cellulose	Hemophan	8	4	60
Cellulose triacetate	CT	85	20	
Synthetic hydrophilic				
Polycarbonate-polyether	PC-PE, Gambrane	15	5	40
Poly(ethylene-co-vinyl alcohol)	EVAL	32	8	
Synthetic hydrophobic				
Polyacrylonitrile	PAN	55	570	
Polymethylmethacrylate	PMMA	25	12	
Polysulfone	PS	40	40	24
Sulfonated polyacrylonitrile	AN69	45	25	5
Kidney glomerular basement membrane		0.2–0.3	1500	0

[a] Ultrafiltration coefficient, mL/(mm Hg) h m^2.

[b] Relative to cuprophan. [Data from Henderson and Chenoweth (1987).]

Table 2. Sieving coefficients for cuprophane and AN69 membranes.

Solute	Molecular weight (Da)	Sieving coefficient Cuprophane	AN69
Urea	60	1.000	1.000
Creatinine	149	0.891	0.985
Sucrose	342	0.791	0.982
Vitamin B_{12}	1355	0.629	0.940
Inulin	5200	0.309	0.776
Egg albumin	44000	0.002	0.044

Data from Green *et al.* (1976).

and

$$Q_{Do}=Q_{Di}+Q_F, \tag{20}$$

and Eq. (11) is thus no longer an accurate expression of solute transfer. Substitution of Eqs. (19) and (20) yields a new expression for dialysance:

$$D=\frac{Q_{Bi}(c_{Bi}-c_{Bo})}{c_{Bi}-c_{Di}}+\frac{Q_F\, c_{Bo}}{c_{Bi}-c_{Di}} \tag{21}$$

$$=\frac{Q_{Di}(c_{Do}-c_{Di})}{c_{Bi}-c_{Di}}+\frac{Q_F\, c_{Do}}{c_{Bi}-c_{Di}}, \tag{22}$$

which now consists of a diffusive term and a term containing Q_F, which represents the convective portion of dialysance. The equation for clearance can be similarly corrected, yielding the convective clearance term $Q_F\, c_{Bo}/c_{Bi}$, which cannot exceed the value of Q_F. Ultrafiltration rates in hemodialysis are rarely greater than 30 mL/min, a value that would range between 10% and 20% of the diffusive clearance for urea, but might account for the majority of removal of solutes with molecular weights over 1500.

3.3.2 Backfiltration. When no net ultrafiltration is occurring ($Q_{Do}=Q_{Di}$), the total net TMP is 0; nevertheless, at single points along the membrane this may not be the case. In fact, dialysate-side pressure is maximal at the blood outlet end of the dialyzer, where blood-side pressure is at a minimum. In the example shown in Fig. 6, net driving pressure on the blood side of the membrane varies from 42 mm Hg at the blood inlet to −16 mm Hg at the outlet. As a result, there is movement of ultrafiltrate out of blood at one end of the dialyzer, and an equivalent movement back from dialysate to blood at the other end, with total net ultrafiltration being 0.

With cellulose membranes, long the mainstay of hemodialysis, ultrafiltration coefficients range from 2 to about 4 mL/(mm Hg)h. Thus the maximum backfiltration rate at any point on the membrane at the pressures in Fig. 6 would be about 1.1 mL/min, probably not a clinically significant effect, especially since negative pressure applied to the dialysate side to produce net ultrafiltration further reduces the potential for backfiltration. With newer membranes, however, K_{UF} may be as high as 80–90 mL/(mm Hg)h, leading to a potential rate of back filtration as high as 24 mL/min at the blood outlet end if net Q_F is 0. At high backfiltration rates, contaminants in the dialysate, which is usually not sterilized, might be convectively transported into the blood by the back-filtrate. Endotoxin, a polymeric product of gram-negative bacteria, or its fragments, is a particularly feared contaminant, since exposure to even trace amounts could cause serious anaphylactic reactions.

4. COMPONENTS OF THE DIALYSIS SYSTEM

The dialysis system consists of a number of discrete components, each with characteristics that influence the process and that may vary considerably. The actual solute exchange takes place within one of several kinds of "artificial kidney," or dialyzer, among which membrane materials, geometry, and flow patterns may differ. Blood has chemical and rheological properties that strongly affect diffusion. The solutes contained in the dialysate determine concentration driving forces for diffusion of numerous small molecules, and the choice of dialysate buffer used to adjust *p*H may have profound effects on the patient. Finally, there is a variety of apparatus used to deliver blood and dialysate to the dialyzer and to control and monitor temperature, pressure, flow rates, and dialysate composition.

4.1 The Dialyzer

Early dialysis was performed with multiple-use devices that generally consisted of a large-surface-area blood-filled cellophane compartment suspended in an open dialysate bath. The ensuing three decades have seen the

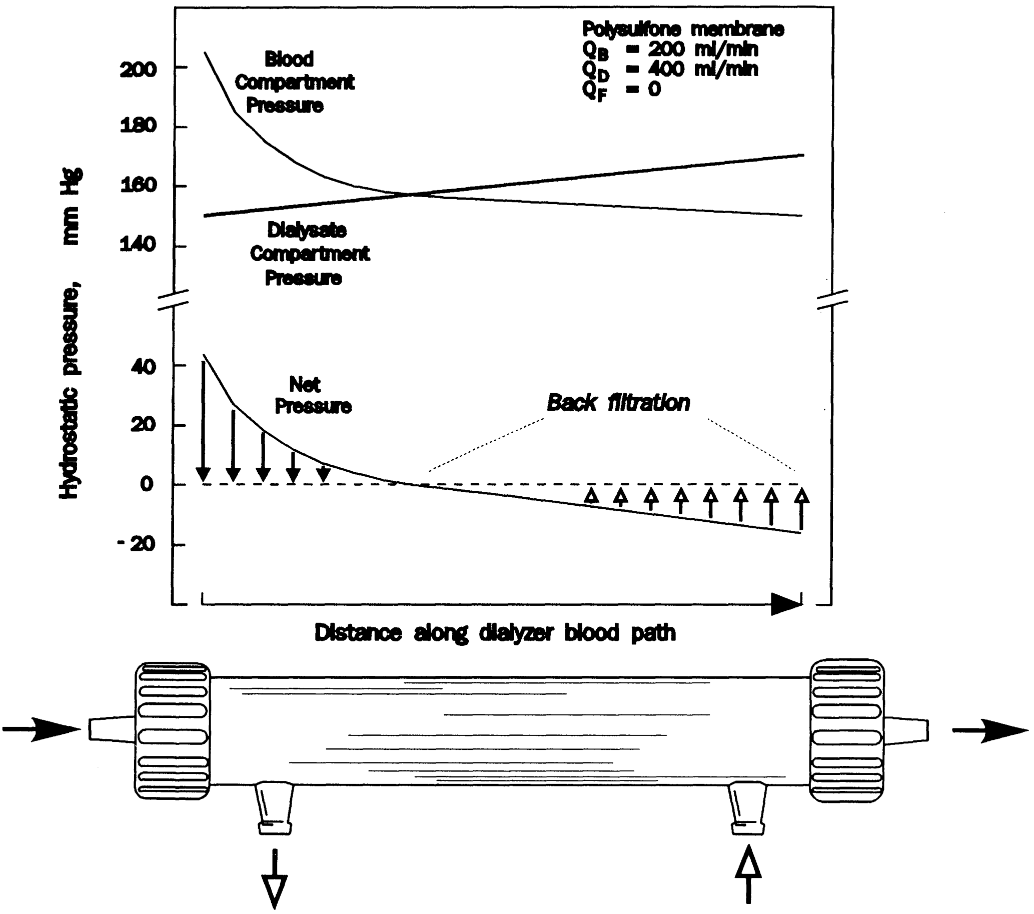

FIG. 6. Backfiltration. With no net ultrafiltration, blood-side hydrostatic pressure exceeds that on the dialysate side over about the first 40% of the blood path, and there is movement of ultrafiltrate out of blood. Over the last 60% of the blood path, the pressure relationships reverse and an equivalent amount of fluid moves from dialysate to blood. Filled arrows, blood compartment; open arrows, dialysate compartment. [Data from Streicher and Schneider (1985).]

development of disposable hemodialyzers, in which all transport processes take place within a self-contained small-volume unit. Cellulosic membranes remain the mainstay of therapeutic dialysis, but a number of new synthetic membrane materials have appeared that have distinctly different properties with respect to mass transfer and to biocompatibility.

4.1.1 Dialyzer Geometry. Hemodialyzers have been produced in several different dialyzer designs, with varying geometrical relationships between blood, membrane, and dialysate. The coil dialyzer dominated the 1950s and early 1960s, when it was succeeded by the parallel plate design. During the 1980s hollow-fiber dialyzers almost completely replaced the earlier designs in clinical use.

4.1.1.1 Coil Dialyzers. The first clinically viable dialyzer constructed consisted of a long cellophane sausage tubing wrapped around a cylindrical drum support. The tube was perfused with the patient's blood, and the entire apparatus was rotated in a tank of dialysis fluid. From this design descended the "twin coil" artificial kidney, introduced in 1956, which was small, efficient (capable of a $K_{D\,\text{urea}}$ of 140 mL/min), inexpensive, and disposable. The dialyzer was placed on end in a canister in an open 100-L tank, and dialysate was pumped crosswise through the supporting mesh. The twin coil was widely used throughout the 1950s and 1960s, and smaller, improved coils were used in many hospitals well into the 1980s, but by 1985 only 1–2% of dialyses in Europe, where the most complete data are available, were performed with this type of dialyzer (Hoenich *et al.*, 1989).

4.1.1.2 The Parallel Plate. The coil dialyzer has now been almost completely replaced by two other designs: parallel-plate and hollow-fiber dialyzers. Parallel-plate devices have evolved from the two-layer 40×14×3-in. Kiil dialyzer of the early 1960s, which was reassembled and fitted with new cellophane membranes before each dialysis. Current versions are small, sterile, disposable devices like the one shown in Fig. 7(a). Blood flows in flat chambers between sheets of membrane with dialysate moving in a countercurrent direction, as illustrated in Fig. 7(b). Modern flat-plate design innovations include precision membrane assembly that allows as many as 17 blood and 34 dialysate layers, and the formation of wavelike microchannels [Fig. 7(c)] in the blood and dialysate chambers to promote turbulent flow. The turbulence en-

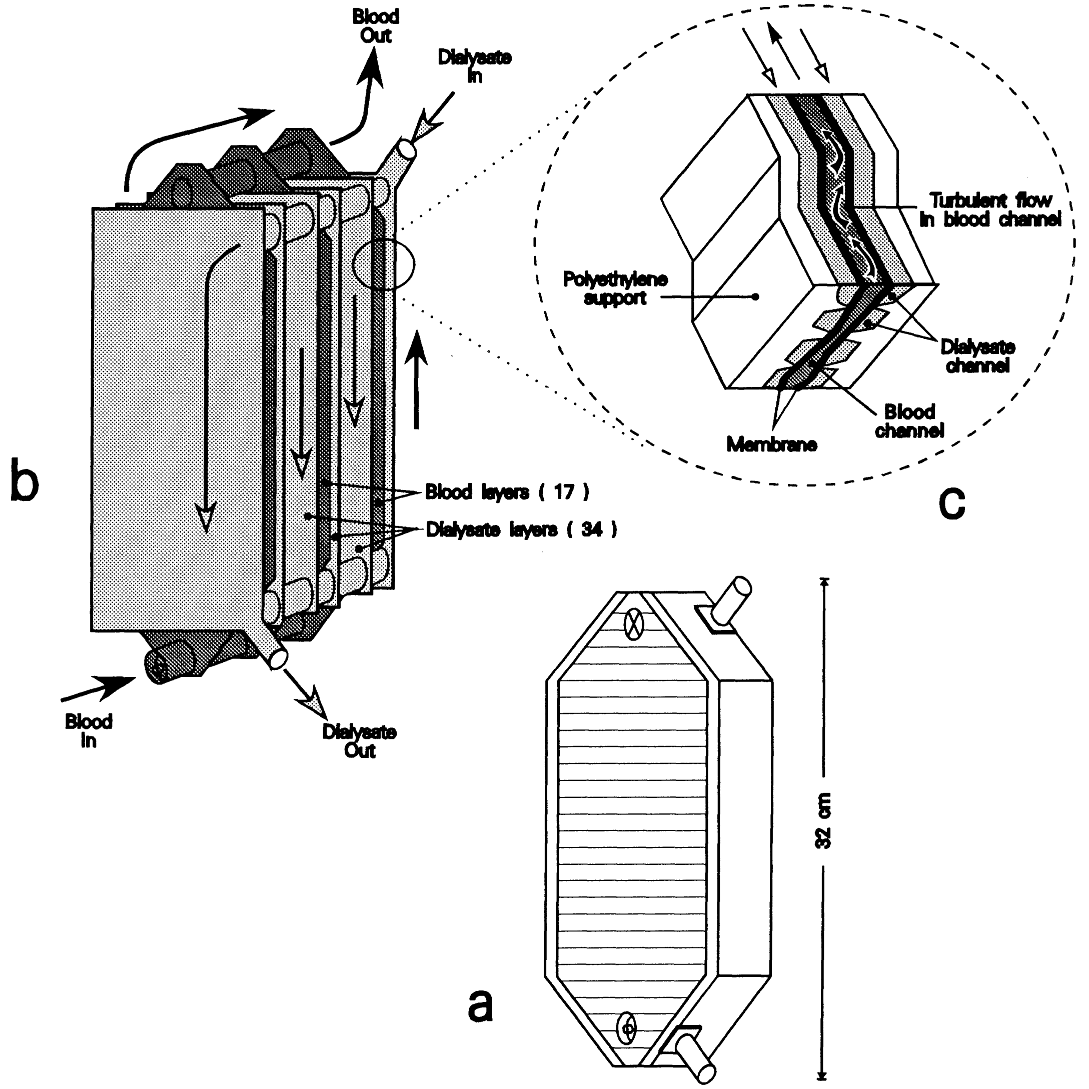
Blood Out
Dialysate In
Turbulent flow in blood channel
Polyethylene support
Dialysate channel
Blood channel
Membrane
b
Blood layers (17)
Dialysate layers (34)
c
Blood In
Dialysate Out
32 cm
a

FIG. 7. Modern parallel-plate hemodialyzer. (a) "Flat-plate" dialyzers, used in about 20% of hemodialyses worldwide. (b) The blood and dialysate paths, flowing countercurrent in parallel flat chambers on opposite sides of a membrane sheet. (c) An enlarged view showing wavelike microchannels within the blood chamber that create turbulent flow and thus better mixing, flow distribution, and transport. Open arrows: dialysate path. Solid arrows: blood path.

sures active mixing, impeding the formation of large stagnant boundary layers, and thereby minimizes mass transfer resistance within the blood and dialysate phases. Parallel plates are probably used in about 20% of dialyses worldwide.

4.1.1.3 Hollow-Fiber Dialyzers. The most widely used design today is the capillary or hollow-fiber dialyzer (Fig. 8), first introduced by Cordis-Dow in the early 1970s. A typical hollow-fiber design consists of a bundle of 6000 to 20 000 capillary tubes, each with an internal diameter of approximately 200 μm and a wall thickness of 8–15 μm for cellulose and up to 40–50 μm for synthetic hydrophobic membranes. The fiber bundle is supported at each end by a "tube sheet" of potting compound, usually polyurethane [Fig. 8(c)]. Hollow-fiber dialyzers have several marked advantages over parallel-plate designs. The ratio of membrane surface area to device volume is greater, leading to a lighter, more compact dialyzer. The fibers are self-supporting, so that effective surface area is not lost through contact with an external rigid framework. Finally, the fibers are less compliant than the parallel-plate blood chambers, so that there is less expansion of the blood compartment during dialysis treatment and consequent decreased pooling of blood outside the body. This is important in minimizing the tendency toward falls in blood pressure in the patient undergoing treatment.

A disadvantage of the hollow-fiber design is its greater vulnerability to plugging of the fibers by particulate matter, making these dialyzers more susceptible to the effects of clot formation in the blood compartment. Severe anaphylactic reactions occasionally occur at the onset of dialysis, which are thought to be largely caused by allergy to residual ethylene oxide gas contained in the dialyzer (Daugirdas and Ing, 1988). This gas, used as a sterilant, is absorbed and slowly released by the polyurethane potting compound, which is not used in parallel-plate devices (Lee *et al.*, 1985); reactions occur over 10 times more commonly with hollow-fiber devices (Villarroel and Ciarkowski, 1985). Nevertheless, hollow-fiber dialyzers now account for up to 80% of dialysis treatments.

4.1.2 Membranes. Dialysis is, after all, a membrane separation process, and the characteristics of the membrane itself strongly affect the nature and efficiency of the dialysis treatment. A number of membranes are in use today, which differ in their structures, their transport properties, and their tendency to interact with blood components.

4.1.2.1 Classification. Dialysis membranes may be conveniently categorized in three groups: cellulose, modified cellulose, and synthetics, with the synthetic membranes further classified as hydrophilic or hydrophobic (Lysaght, 1988). All are natural or synthetic polymers, and are listed along with several properties in Table 1.

Cellulose Membranes. For the first 20 yr of chronic dialysis, there was really only one membrane material—regenerated cellulose. Reconstituted by the cuprammonium process or by saponification of heat-spun cellulose acetate, cellulose is prepared as flat-sheet or hollow-fiber membranes, and still is used in the large majority of dialysis treatments, principally because of its low cost. The structure of cellulose confers particular properties on a hemodialysis membrane. The density of hydroxyl groups on the cellobiose dimer (see Fig. 9) leads to a high polarity, high degree of intramolecular hydrogen bonding, and large affinity for water. The membranes swell markedly when wet, and form a *hydrogel,* a homogeneous single phase containing both water and polymer. Finally, the free hydroxyl groups readily participate in chemical reactions, and probably account for the high interactivity of cellulose membranes with blood cells and proteins.

Modified Cellulose Membranes. To reduce the chemical reactivity of cellulose, blocking groups may be used to replace some or all of the hydroxyl hydrogen atoms (Fig. 9). The most common membrane material of this type is cellulose acetate, in which 60–80% of the available hydrogen has been replaced by acetate radicals. The resultant membrane is less bioactive (see below) and somewhat less hydrophilic, forming a more "open" hydrogel with higher water permeability. The newer cellulose triacetate, with all hydroxyl groups bound to acetate, is so altered that its behavior is more similar to the synthetic hydrophobic materials than to native cellulose. Hemophan is cellulose with about 5% of hydroxyl hydrogen replaced by the diethylaminoethyl moiety, and has properties similar to cellulose acetate.

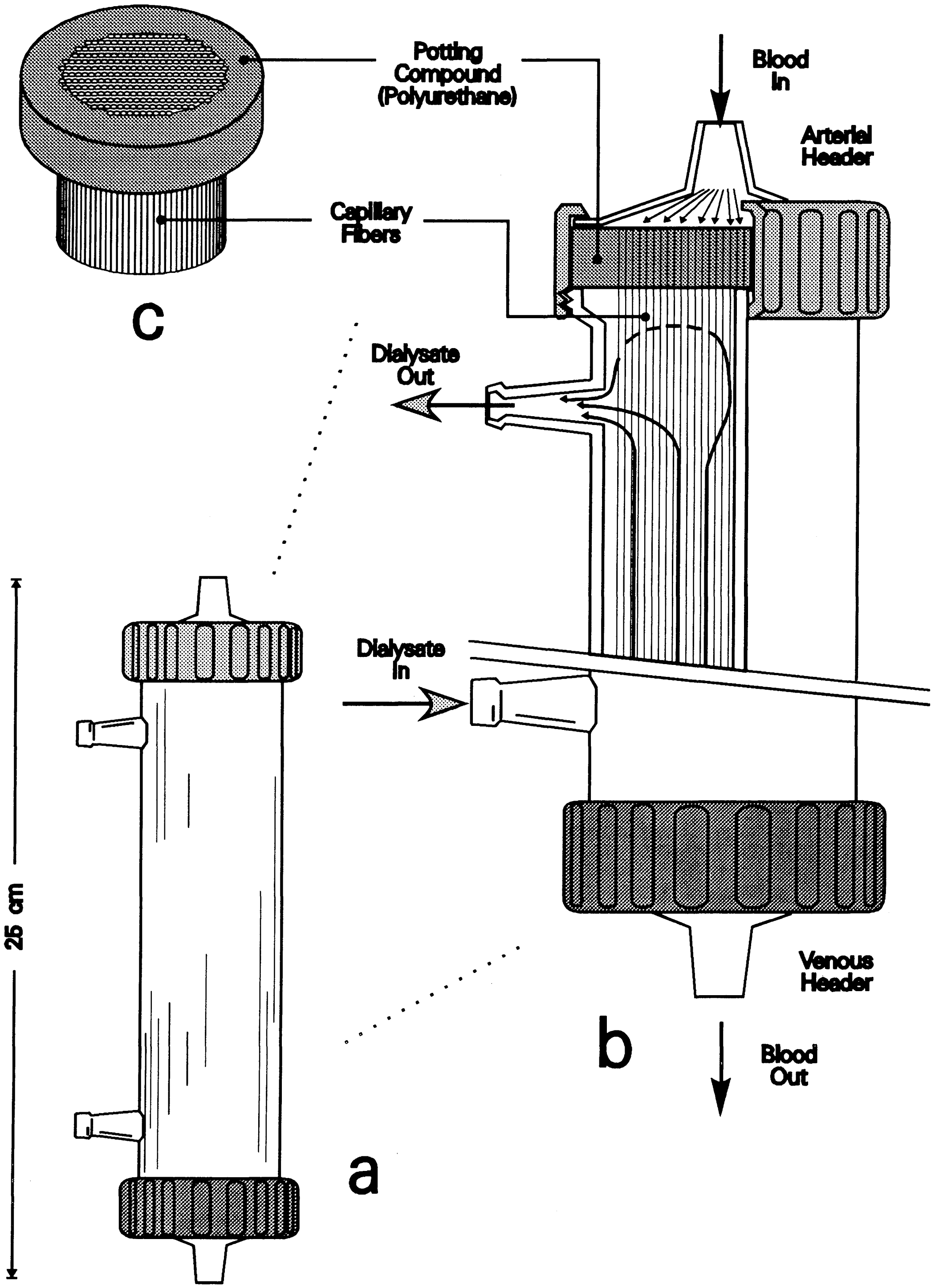
Potting Compound (Polyurethane)
Blood In
Arterial Header
Capillary Fibers
c
Dialysate Out
Dialysate In
25 cm
Venous Header
b
Blood Out
a

FIG. 8. Hollow-fiber hemodialyzer. In a typical hollow-fiber dialyzer (a) blood flows through small (~200-μm internal diameter) fibers that are bathed on the outside by dialysate (b). The fiber ends are supported by a disk of a potting compound such as polyurethane (c).

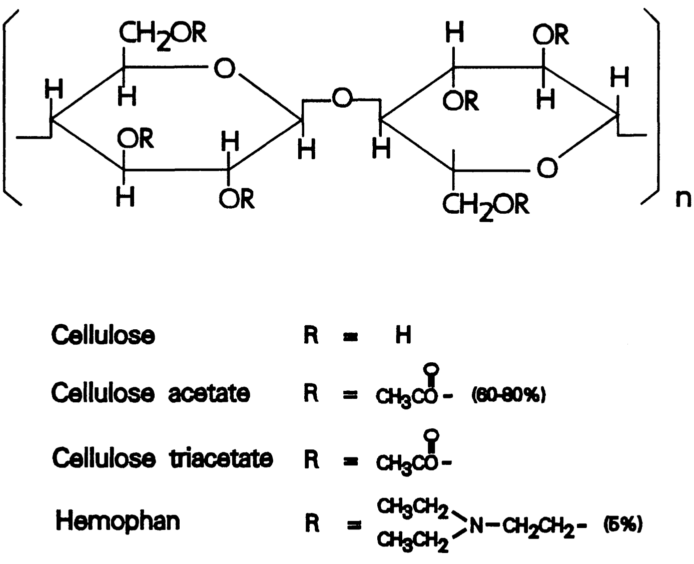
CH2OR
H
OR
O
n
Cellulose R = H
Cellulose acetate R = CH3CO- (60-80%)
Cellulose triacetate R = CH3CO-
Hemophan R = CH3CH2 CH3CH2 >N-CH2CH2- (5%)

FIG. 9. Structure of cellulose and its derivatives.

Hydrophilic Synthetic Polymers. Polycarbonate-polyether (PC-PE) and poly-ethylvinyl alcohol (EVAL) are synthetic membranes with chemical and physical similarities to cellulose. PC-PE has many polar carbonyl groups, and EVAL has unsubstituted hydroxyls; thus both membranes are quite hydrophilic, and form hydrogels. Unlike cellulose, however, PC-PE does not swell when wet.

Hydrophobic Synthetic Polymers. These are nonpolar, resonance-stabilized polymers with low energy of interaction with water, listed in Table 1. Membranes are produced from these materials by a phase inversion process, in which the polymer is dissolved in a solvent and pumped through a spinning jet. As the polymer solution emerges from the jet, it is immersed in a nonsolvent precipitant (often water) that replaces the solvent and converts the solution into a two-phase system, the solid phase forming the rigid membrane structure and the liquid, polymer-poor phase forming pores. Membranes formed in this way have chemical and physical properties quite different from cuprophane and other hydrophilic materials. Their overall structure is that of an anisotropic foam. The inner, blood-contacting surface of a hollow fiber made from these materials is a thin skin, usually about 1 μm in thickness. This relatively dense skin determines the transport characteristics of the membrane. The surrounding open foam-like layer has very little flow resistance and functions mainly as mechanical support. The precise morphology of the membrane may be varied by changes in the polymer concentration, solvent, precipitant, or temperature of the phase inversion.

4.1.2.2 Transport Characteristics. The structural differences among these membranes lead to differences in transport capabilities. In general, clearances of small–molecular-weight solutes such as urea are similar for similar conditions of blood and dialysate flow rate, although individual dialyzer design and varying membrane thickness account for about a 40% difference between the highest and lowest urea clearance in the group CA > PS > PMMA > PAN > Cuprophane (Petersen *et al.*, 1987). Examination of Table 1, however, reveals a marked difference between hydrophilic and hydrophobic membranes in water permeability: K_{UF} for the hydrophobic synthetics is as much as 10 times higher than that of cuprophane. As a result, convective clearance is also higher with the hydrophobic group, and larger–molecular-weight solute removal is considerably more efficient. This marked advantage for control of uremic toxicity is partially offset by the problem of potentially excessive water loss, which necessitates the use of special volumetric ultrafiltration systems, by the increased risk of backfiltration (see Sec. 3.3.2), and by the generally much higher cost of the newer synthetic membranes.

4.1.2.3 Biocompatibility. Besides transport characteristics, there is another important property that determines the suitability of a membrane material for hemodialysis, and that is its propensity to interact, usually in deleterious ways, with components of the body. The term "biocompatibility" has come to be used in this context, defined in negative terms as the *absence* of (1) thrombogenic, toxic, allergic or inflammatory reaction, (2) destruction of formed elements of blood, (3) changes in plasma proteins and enzymes, (4) immunological reaction, (5) carcinogenic effect, and (6) deterioration of adjacent tissues (Klinkmann *et al.*, 1989). It is perhaps more accurate to speak of the occurrence of any of these reactions as "bioincompatibility."

Coagulation. Bioincompatibility of hemodialysis membranes occurs in several ways, primarily through interaction with the blood coagulation system, with the circulating inflammatory mediator proteins of the complement system, and with white blood cells. As occurs with almost any artificial surface, exposure of dialysis membranes to the circulation leads to rapid adsorption of a layer of plasma proteins and adhesion of platelets, the small formed elements in blood that initiate the formation of a "hemostatic plug." This is frequently followed by activation of the blood coagulation system and formation of a thrombus, necessitating the use of anticoagulants, usually heparin, during the dialysis treatment. Membranes differ in thrombogenicity, and although data are scarce, cellulose seems the most reactive, with a decrease in clotting tendency when derivatized to cellulose acetate or hemophan. The least reactivity is shown by the hydrophobic synthetic membranes.

Complement activation. The complement system is a group of more than 20 plasma proteins, normally inactive, which in response to appropriate stimuli may be enzymatically cleaved to yield a number of bioactive proteins and peptides that mediate host-defense and acute inflammatory reactions. Dialysis membranes may serve as the stimuli, and activation of complement to varying degrees is a well-established phenomenon in hemodialysis and a common measure of bioincompatibility. The effects of activated products include agglutination of white blood cells in the lung, leading to transient neutropenia and contributing to the hypoxemia associated with the first hour of dialysis. New, nonreprocessed cellulose membranes, with their exposed hydroxyl groups, are by far the strongest activators of complement, while the hydrophobic synthetics are the most biocompatible. Relative activities of various membranes are shown in Table 1.

Synthetic hydrophobic membranes seem to represent a step forward in hemodialysis technology, with improved transport properties and enhanced biocompatibility with respect to cellulosic membranes. There is also evidence that cellulose may induce the release into the circulation of β-2-microglobulin, a ubiquitous 11 800-Da cell-surface protein that accumulates in dialysis patients and may be deposited in bones and joints, leading to a painful and crippling arthropathy, while the synthetic membranes clear the molecule from the blood with varying effectiveness. Nevertheless, technical difficulties associated with high water permeability, and especially high cost, have greatly hindered acceptance of the synthetics, and in the United States cellulosic membranes are still used for 70–80% of hemodialysis treatments.

4.2 Blood

Much of the preceding analysis of diffusion phenomena in hemodialysis treats blood as a simple aqueous solution. The reality is a good deal more complex. Transport phenomena in blood are influenced by solute interactions with electrolytes, with suspended proteins, and with cellular elements—especially red blood cells—as well as by the particular rheological properties of blood.

4.2.1 Protein. The protein content of blood affects transport in several ways. First, small solutes may bind noncovalently to protein macromolecules, thereby becoming markedly less diffusible. In the extreme case, a small molecule may be almost completely unaffected by dialysis. The trivalent aluminum ion, for example, with a molecular weight of only 27, not only is not removed from the circulation by dialysis, but can actually be transported into the blood from contaminated water against an apparent concentration gradient. In fact, about 80% of blood-borne aluminum is bound to protein, and is therefore not available for diffusion, remaining entrapped in the intravascular space. A similar phenomenon occurs with some drugs, such as digoxin and glutethimide, which are highly protein-bound and not significantly removed by dialysis.

Hemodialysis membranes may themselves interact with plasma proteins, both specifically, as outlined in Sec. 4.1.2.3, and nonspecifically. The high hydraulic fluxes that are often attained with high permeability membranes carry along macromolecules, which accumulate in the immediate area of the membrane. This "concentration polarization," shown in Fig. 10, can lead to formation of a protein gel that may interfere with transport. In fact, perfusion of a polysulfone hollow fiber dialyzer with plasma may lead to a 60% reduction in K_{UF} compared to values obtained with saline perfusion (Bosch *et al.*, 1985).

4.2.2 Blood Cells. Blood is not, of course, a homogeneous liquid, but a suspension in which the dispersed phase—the blood cells—consists of liquid droplets each enclosed in a solid elastic envelope. Its resultant non-Newtonian rheological properties are complex, but probably not of major consequence to dialysis transport phenomena.

The distribution of solutes in the plasma and cell compartments, and the rate of transfer between them, however, have significant impact on the efficiency of dialysis. More urea is removed from aqueous solutions than from equimolar solutions in whole blood, partly because of the decreased volume fraction of free water in plasma and red cells (92% and 72%), and partly because of increased resistance to diffusion caused by plasma protein in the boundary layer. Equilibrium distribution coefficients ($c_{\text{red cells}}/c_{\text{plasma}}$) for various solutes are shown in Table 3. The average concentration in whole blood for most of these solutes, given by

$$\bar{c}=\phi c_{\text{red cells}}+(1-\phi)c_{\text{plasma}}, \tag{23}$$

where ϕ is the hematocrit, or volume fraction of cells, will be less than the concentration measured in plasma. Thus for most solutes the quantity passing through the dialyzer decreases as the hematocrit increases, and solute removal per unit time will correspondingly decrease.

Urea crosses cell membranes with great facility. During the 20 s or so that it takes for an erythrocyte to pass through the dialyzer, the urea concentration in its immediate environment falls by as much as 70–80%. Despite this drastic change in the cell's milieu, the

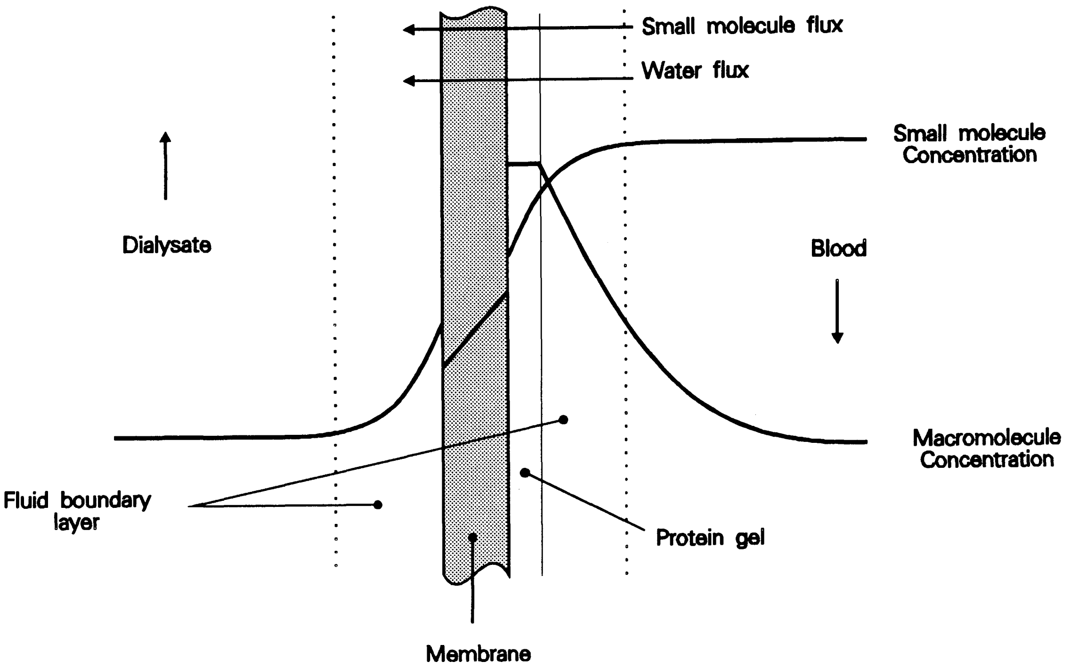

FIG. 10. Concentration polarization. Shown are concentration profiles for small molecules and macromolecules (protein) in the vicinity of a hemodialysis membrane. Water flux from blood to dialysate carries protein molecules to the membrane surface, where a gel may be formed that can interfere with transport of smaller solutes and water. Increasing blood flow rate increases shear forces which tend to limit gel formation.

Table 3. Equilibrium distribution coefficients, $c_{\text{red cells}}/c_{\text{plasma}}$.[a]

Urea	0.86
Creatinine	0.73
Uric acid	0.54
Amino acid nitrogen	1.54
Glucose	0.95

[a] Colton (1981).

ratio between intracellular and extracellular urea is virtually the same at the dialyzer blood inlet and outlet. Other substances, however, such as potassium and phosphate, diffuse more slowly across biological membranes, are retained within the red blood cell, and are relatively less available for removal by dialysis. Late equilibration between the intracellular and extracellular pools of these solutes after the dialysis treatment is finished contributes to a rebound rise in plasma concentration.

4.3 Dialysis Fluid

The driving force for transport of small molecules in dialysis is the transmembrane concentration gradient, which may be manipulated by altering the composition of the dialysis fluid.

4.3.1 Composition. Table 4 lists the composition of typical dialysate solution, with comparative values for serum. The presence of glucose in dialysate may be something of an anachronism. Its originally intended functions—stabilization of serum osmolality and osmotic extraction of water—have been rendered obsolete by high-sodium dialysate and by pressure- and volume-mediated ultrafiltration.

The need for an alkaline buffer to correct the accumulation of hydrogen ions in the patient with renal failure led to the early use of bicarbonate in dialysis solutions. Preparation of dialysate from concentrates, however, made the use of bicarbonate problematic. High concentrations of HCO_3^- are unstable in solution, as the high partial pressure of carbon dioxide leads to loss of the gas into the surrounding atmosphere and fall in HCO_3^- concentration. Further, the presence of calcium and magnesium ions in the dialysate causes precipitation of carbonates with these cations. During the 1960s, the problem was resolved by the use of acetate as the buffer anion. Once in the circulation, acetate is converted quickly and quantitatively to bicarbonate by normal metabolic pathways, and presents no problems of solubility or stability in dialysate solution.

As dialysis became more rapid and efficient, however, and transfer rates of acetate exceeded the rate of metabolic conversion, other problems arose. Acetate in the concentrations achieved in blood produced significant clinical toxicity, including hypotension, headaches, and vomiting. Patients differ in their sensitivity, but by now it is clear that there is a common and marked increase in

Table 4. Composition of serum and dialysate.

	Serum	Dialysate
Electrolytes (mEq/L):		
Sodium	140.0	140.0–145.0
Potassium	4.0	2.0
Chloride	100–110	95–109
Calcium (ionized)	1.8	2.5–3.5
Magnesium	1.6	1.0
Glucose (mg/dL):	120	0–200
Buffer (mEq/L):		
Acetate	0.2	35–40[a] 4[b]
Bicarbonate	25	0[a] 35–40[b]

[a] Acetate dialysate.
[b] Bicarbonate dialysate.

symptoms and hemodynamic instability associated with acetate use. The result has been a return to bicarbonate, but with more sophisticated dialysis delivery systems that allow mixing of two separate concentrates, one containing the buffer and the other the divalent cations Mg^{2+} and Ca^{2+}.

4.3.2 Water Preparation. In most modern dialysis systems the dialysate is prepared continuously by a proportioning system from concentrate and line water. During three 4-h hemodialysis treatments with 500-mL/min dialysate flow, a patient is exposed to 360 L of municipal water. This is roughly 25 times the average exposure by ingestion, and the transport of trace elements and toxins into the blood is far more direct. Moreover, dialysis patients are unable to clear toxins continuously by urinary excretion. Water that is suitable for drinking, therefore, requires further preparation for use in hemodialysis.

Most hemodialysis units use a serial water purification system with several components. After initial screening of particulate matter by a sediment filter, a water softener utilizes a cationic resin to exchange sodium ions for divalent cations, principally magnesium and calcium. An activated carbon filter is used to adsorb chlorine, chloramines, and dissolved organic material. Chlorine and chloramines are used extensively for bactericidal treatment of municipal water supplies, and in high dialysate concentrations can cause serious toxicity. A second sediment filter removes the small carbon particles released from the carbon filter.

For most water supplies, reverse osmosis (RO) is the method of choice for removal of solutes. Within the RO module, feedwater is in contact with a spirally wound or hollow-fiber semipermeable (often cellulose acetate) membrane, and a hydrostatic pressure is applied that exceeds the osmotic pressure generated by the dissolved solutes. As a result, water flows against the osmotic gradient, and a purified product is produced, even small ionic solutes having been removed by molecular sieving. Effectiveness of the device is monitored by measuring resistivity of feedwater and product water and calculating the percentage rejection of solutes, which ranges from 90% to 98% of monovalent and 95% to 99% of divalent ions. RO also effectively removes organic contaminants and bacteria, as well as particle matter; the combination of RO and pretreatment with water softener and carbon filter virtually always produces water pure enough for safe hemodialysis.

4.4 Dialysis Monitoring and Delivery Systems

Hemodialysis has evolved in 30 yr from a difficult and dangerous procedure, requiring constant supervision by a physician, to a routine treatment carried out with a minimum of technical intervention, even practicable in a patient's home. The dialysis apparatus has become more self-regulating, with the latest equipment capable of on-line automatic blood pressure recording and minute-to-minute variability in dialysate sodium concentration.

4.4.1 Functions. The delivery system, commonly referred to as the "dialysis machine," has four principal functions. First, it maintains flow through the blood chamber of the dialyzer, via the extracorporeal blood circuit. Second, it continuously prepares dialysate from line water and concentrate and keeps it flowing through the dialyzer. Third, it regulates ultrafiltration, usually by adjusting volume or hydrostatic pressure in the dialysate circuit. Fourth, it assures patient safety by monitoring and regulating temperature, pressures, and dialysate concentration, and detecting emergency situations like dialysis membrane ruptures or the presence of air in the blood returning to the circulation. A dialysis machine of the conventional "single-pass" type—in which dialysate flows in an uninterrupted path from line water to drain—is shown schematically in Fig. 11.

4.4.2 The Blood Circuit. In modern dialysis, blood is both withdrawn from and returned to the venous circulation; hence there is no pressure gradient to drive blood flow. The driving force for extracorporeal flow is therefore provided by a peristaltic roller pump in the blood circuit. This in turn necessitates a system of air traps and alarms to prevent air, which may be drawn into the system by the negative pressure created by the pump, from being returned to the patient's circulation.

An intermittent or continuous infusion of heparin is necessary to prevent coagulation of blood in the dialyzer or lines. A continuous pressure monitor on the returning ("venous")

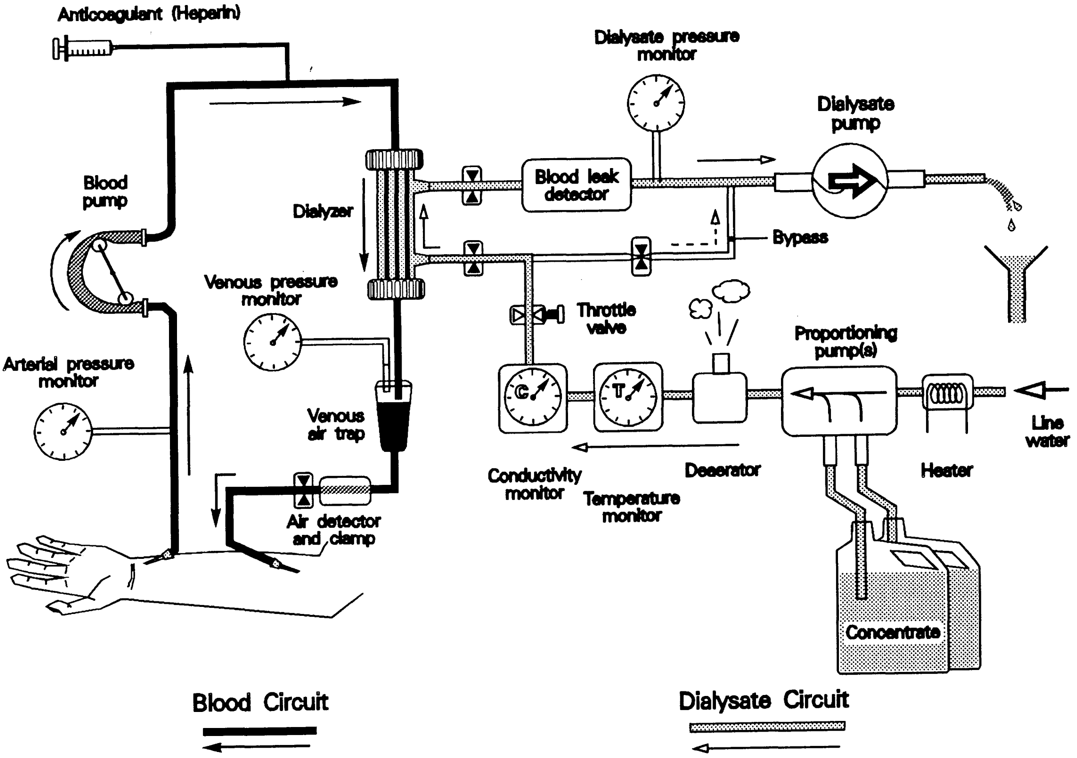
Anticoagulant (Heparin)
Dialysate pressure monitor
Dialysate pump
Blood pump
Blood leak detector
Dialyzer
Bypass
Venous pressure monitor
Throttle valve
Arterial pressure monitor
Proportioning pump(s)
Venous air trap
C
T
Line water
Conductivity monitor
Deaerator
Heater
Temperature monitor
Air detector and clamp
Concentrate
Blood Circuit
Dialysate Circuit

FIG. 11. Single-pass dialysis apparatus. This is still the most widely used configuration for dialysis delivery systems. Ultrafiltration is regulated by change in the tension on the throttle valve, thus varying relative negative pressure in the dialysate circuit.

blood line stops the blood pump if pressure exceeds or falls below preset limits. The dialyzer and all of the tubing that contacts blood are disposable. For economic reasons dialyzers are often reprocessed for repeated use by the same patient; the tubing is always discarded after one use.

4.4.3 The Dialysate Circuit. The dialysate is continuously prepared by a proportioning system from heated line water and bottled concentrate, usually mixing the two in a ratio of about 35 to 1. The solution is deaerated by subjection to negative pressure to prevent later spontaneous degassing, bubble formation, and "vapor lock" of blood channels in the dialyzer. The final composition is monitored by measuring the electrical conductivity of the solution; when the conductivity varies by more than 10% from the target value, an alarm condition results, and bypass valves are tripped that isolate the dialyzer from the dialysate path. A similar alarm system monitors temperature, and an optical sensor stops the blood pump if blood is detected in dialysate exiting from the dialyzer.

4.4.4 Ultrafiltration Systems. Control of the amount of fluid removed during a dialysis treatment requires special apparatus to regulate the hydrostatic pressure across the membrane. The earliest ultrafiltration control system, shown in Fig. 12(a), used with large open tanks of dialysate, consisted of a screw clamp (3) on the venous blood line. Increasing blood outflow resistance with the clamp raised the hydrostatic pressure in the dialyzer blood chamber (shown in figure as + signs) and increased the rate of ultrafiltration. The development of continuously proportioned single-pass delivery systems allowed regulation of the pressure on the dialysate side, and the system shown in Fig. 12(b) is now the most common method of ultrafiltration control. Raising inflow resistance with the throttle valve (4) causes the dialysate pump (5) to develop increasing negative pressure within

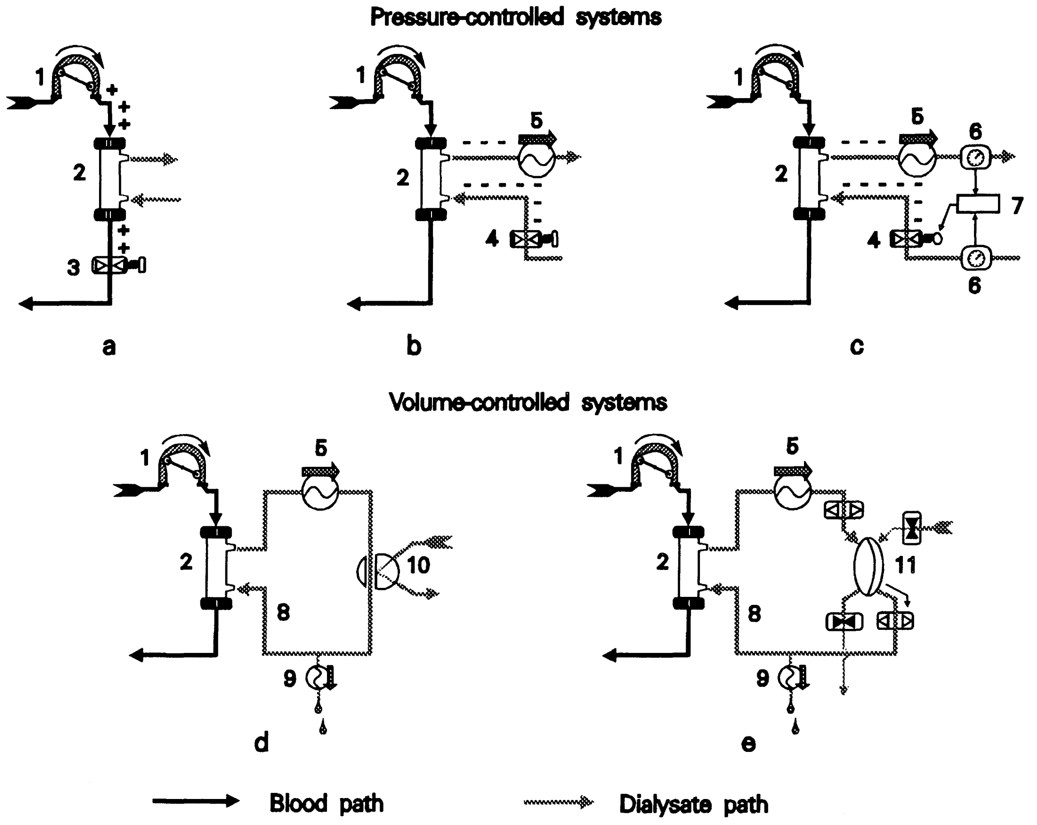

FIG. 12. Systems for ultrafiltration control. (1) blood pump; (2) dialyzer; (3) screw clamp on venous blood line; (4) dialysate line throttle valve; (5) dialysate pump; (6) flow sensor; (7) digital analyzer; (8) closed dialysate circuit; (9) ultrafiltration pump; (10) three-way valve for draining and refilling of dialysate circuit; (11) balancing chamber and valves. See text for explanation.

the dialysate circuit, which in turn increases water flux from blood to dialysate.

This system is vulnerable to a number of inaccuracies. Pressure measurement is not precise and varies with arm movements, fiber clotting, etc., and accurate clinical prediction of ultrafiltration requires exact knowledge of the dialyzer K_{UF} and blood oncotic pressure. A refinement of the system that seeks to overcome these problems [Fig. 12(c)] uses flow sensors (6) on the incoming and outgoing dialysate lines to monitor ultrafiltration more precisely, and to exert feedback control via a digital analyzer (7) that adjusts the throttle valve (4) to maintain the desired rate of fluid removal.

Pressure-based systems serve adequately for dialyzers with cellulose membranes, but use of the newer synthetics, with ultrafiltration coefficients greater than 10 mL/(mm Hg)h, has introduced another problem. The minimum ultrafiltration rate achievable with pressure-controlled systems is that rate generated by the minimum transmembrane hydrostatic pressure. This minimum pressure is the mean blood-chamber pressure, since the system does not generate positive pressure on the dialysate side. Flow resistances in the dialyzer and return lines will almost always produce a pressure of greater than 100 mm Hg; thus, for a dialyzer with a K_{UF} of 30 mL/(mm Hg)h (see Table 1), the minimum ultrafiltration rate would be 3000 mL/h, and could easily reach 5000–6000 mL/h. Removal of 10–20 L of fluid from a patient's circulation over a 3-h period could obviously lead to serious complications, and so in order to use high-UF membranes safely, new ultrafiltration control systems have been devised.

These "volumetric" systems [Fig. 12(d) and 12(e)] regulate ultrafiltration by removing fluid from a closed, rigid dialysate circuit (8) at a measured rate via an ultrafiltration pump (9). Since the dialysate loop is noncompliant, ultrafiltrate crosses the dialysis membrane at a rate that exactly matches the rate of removal by the ultrafiltration pump. Several variations exist, one of which [Fig. 12(d)] uses a closed loop that is drained and refilled at 2-min intervals via a three-way valve (10), and another [Fig. 12(e)] that uses balancing chambers separated by a diaphragm (11) to ensure the equality of dialysate inflow and outflow volume, with exiting dialysate displacing an exactly equal quantity of fresh dialysate into the circuit. In practice, two parallel chambers operating 180° out of phase are used to maintain more constant flow.

5. THE FUTURE

Dialysis is still in many ways a "half-way technology," which prolongs life but does not completely restore patients to a healthy, physiologic state. The definition of adequate dialysis remains elusive, despite our recent enchantment with blood urea. There is much yet to understand about the interaction of the dialysis apparatus with the immune system, and the influence of end-stage renal disease on immunity itself. Microminiaturization of dialysis components for the development of portable or wearable artificial kidneys, and the development of new, efficient membranes with transport profiles that more closely mimic the human kidney are priorities for technical development.

Despite its shortcomings, however, regular dialysis treatment has progressed in 30 yr from a relatively primitive, controversial, and short-term therapy to a highly technological and largely automated procedure backed by a complex and sophisticated body of pathophysiological knowledge, which keeps hundreds of thousands of patients alive for as much as 20 yr and more. If progress continues at that rate during the next few decades, concepts which now seem remote fantasy, such as implantable bioartificial kidneys, may well be realized.

GLOSSARY

Anion: A negatively charged ion.

Biocompatibility: Tendency of a foreign substance not to react with components of the body.

Cation: A positively charged ion.

Clearance: Solute removal capacity of a kidney or dialyzer, defined by total solute removal per unit time divided by plasma concentration.

Collodion: Precipitated nitrocellulose; Used as a dialysis membrane.

Colloid: Large–molecular-weight solutes that diffuse slowly or not at all across a semipermeable membrane.

Complement: A system of blood proteins that participate in host-defense and inflammatory reactions.

Convective Transport: Transfer of solute across a membrane by ultrafiltered solvent; transfer by "solvent drag."

Crystalloid: Small–molecular-weight solutes that diffuse readily across a semipermeable membrane.

Dalton: The unit of molecular weight; the mass of the nucleus of the hydrogen atom; equals 1.00794 u.

Dialysance: Solute removal capacity of a hemodialyzer, defined by total solute removal per unit time divided by plasma-dialysate concentration gradient.

Dialysate: In hemodialysis, the electrolyte solution on the diffusate side of the membrane.

Diffusate: The solution on the side of the dialysis membrane containing diffused crystalloids.

Fractional Excretion: The ratio of rate of urinary excretion to rate of glomerular filtration of a solute.

GFR: Glomerular filtration rate, the rate of ultrafiltration of plasma water at the glomerulus.

Glomerulus: The capillary tuft in the kidney where filtration takes place.

Hemodialysis: Dialysis of blood across an artificial membrane for treatment of kidney failure.

Middle Molecules: Molecules with molecular weight 1500–5000 Da, felt to be responsible for toxicity in kidney failure.

Nephron: The structural unit of renal function, consisting of a glomerulus and tubule.

Permeate: Diffusate.

Retentate: The solution on the side of the dialysis membrane containing retained colloid.

Reverse Osmosis: Transfer of water across a semipermeable membrane against an osmotic gradient, driven by hydrostatic pressure.

Uremia: The complex of symptoms of kidney failure.

List of Works Cited

Bosch, T., Schmidt, B., Samtleben, W., Gurland, H. J. (1985), in: E. Streicher and G. Seyffart (Eds.), *Highly Permeable Membranes* (*Contributions to Nephrology, Vol. 45*) Basel: Karger, pp. 14–22.

Colton, C. K., Lowrie, E. G. (1981), in: B. M. Brenner and F. C. Rector (Eds.), *The Kidney*, Philadelphia: Saunders, Chap. 47.

Daugirdas, J. T., Ing, T. S. (1988), *Kidney Inter.* **33** (Suppl. 24), S37–S43.

Dworkin, L. D., Brenner, B. M. (1985), in: D. W. Seldin and G. Geibisch (Eds.), *The Kidney: Physiology and Pathophysiology*, New York: Raven, Chap. 19.

Green, D. M., Antwiler, G. D., Moncrief, J. W., Decherd, J. F., Popovich, R. P. (1976), *Trans. Am. Soc. Artif. Internal Organs* **22**, 627–636.

Henderson, L. W., Chenoweth, D. (1987), *Blood Purif.* **5**, 100–111.

Hoenich, N. A., Woffindin, C., Ward, M. K. (1989); in: J. Maher (Ed.), *Replacement of Renal Function by Dialysis*, Dordrecht: Kluwer, Chap. 5.

Klinkmann, H., Falkenhagen, D., Stefoni, S., Bonomini, V. (1989), in: V. Bonomini, M. P. Scolari, S. Stefoni, and A. Vangelista (Eds.), *Biotechnology in Renal Replacement Therapy* (*Contributions to Nephrology, Vol. 70*), Basel: Karger, pp. 213–226.

Lee, F. F, Durning, C. J., Leonard, E. F. (1985), *Trans. Am. Soc. Artif. Internal Organs* **31**, 526–533.

Lysaght, M. J. (1988), in: G. D'Amico and G. Colasanti (Eds.), *Nephrology and Dialysis Updated* (*Contributions to Nephrology, Vol. 61*), Basel: Karger, pp. 1–17.

Petersen, J., Hyver, S., Yeh, I. (1987), *Trans. Am. Soc. Artif. Internal Organs* **33**, 265–268.

Streicher, E., Schneider, H. (1985), in: E. Streicher and G. Seyffart (Eds.), *Highly Permeable Membranes (Contributions to Nephrology, Vol. 45)*, Basel: Karger, pp. 1–13.

United States Renal Data System (1990), *1990 Annual Data Report*, Bethesda: National Institutes of Health.

Villarroel, F., Ciarkowski, A. A. (1985), *Artif. Organs* **9**, 231–238.

Wolff, A. V., Remp, D. G., Kiley, J. E., Currie, G. D. (1951), *J. Clin. Invest.* **30**, 1062–1070.

Further Reading

Brenner, B. M., Rector, F. C., Jr. (Eds.) (1991), *The Kidney*, Philadelphia: Harcourt, Brace, Jovanovich. (More than 2400 pages of extremely authoritative information on kidney physiology, kidney diseases, and their management.)

Daugirdas, J. T., Ing, T. S. (Eds.) (1988), *Handbook of Dialysis*, Boston: Little, Brown. (A concise, readable, introductory-level text on medical management that is also quite complete.)

Maher, J. (Ed.) (1989), *Replacement of Renal Function by Dialysis*, Dordrecht: Kluwer. (The most exhaustive and authoritative text on dialysis.)

Tuwiner, S. B., Miller, L. P., Brown, W. E. (1962), *Diffusion and Membrane Technology*, New York: Reinhold. (An older but classic work on industrial dialysis applications.)

United States Renal Data System (1990, 1991), *1990 and 1991 Annual Data Reports*, Bethesda: National Institutes of Health. (A statistical gold mine on the epidemiology and economics of dialysis and transplantation in the United States.)

Wankat, P. C. (1990), *Rate-Controlled Separations*, London: Elsevier. (An engineering approach to dialysis and related processes.)

DIAMAGNETISM

LEOPOLDO M. FALICOV, *Department of Physics, University of California, Berkeley, California, U.S.A., and Materials and Chemical Sciences Division, Lawrence Berkeley Laboratory, Berkeley, California, U.S.A.*

Introduction 557
1. **Thermodynamic Arguments** 558
2. **Diamagnetism of Atoms, Molecules, and Molecular Solids** 558
2.1 Hamiltonian of Electrons in a Magnetic Field 558
2.2 Langevin Theory for Closed-Shell Systems 559
3. **Orbital Diamagnetism of Metallic Electrons** 559
3.1 Quantization of Itinerant Electrons in a Magnetic Field 561
3.2 Landau Diamagnetism 562
3.3 The de Haas–van Alphen Effect 563
4. **Diamagnetism in Superconductors** 564
4.1 The Meissner Effect 565
4.2 The London Theory 565
4.3 The Ginzburg–Landau Theory 566
4.3.1 General Formulation 566
4.3.2 Penetration Depth and Coherence Length 567
4.3.3 Flux Quantization 568
4.3.4 Phase-Current Relationship; The Josephson Effect 568
4.3.5 Magnetic-Field Dependence of the Order Parameter 569
4.3.6 Quantum Interference Phenomena 569
5. **Conclusion** 569
Glossary 569
Works Cited 570
Further Reading 571

INTRODUCTION

The word diamagnetic was introduced—originally as a noun, later as an adjective—by Michael Faraday in 1846. In the *Philosophical Transactions* (Faraday, 1846) one reads, "By a diamagnetic I mean a body through which lines of magnetic force are passing, and which by their action does not assume the usual magnetic state of iron or loadstone." It was constructed from the Greek prefix $\delta\iota\alpha$ (which means by, through, across) and the word *magnetic*—from the Latin word *magneticus*—which described, at that time, the special forces inherent in naturally occurring iron, iron oxides, and other substances.

The operational definition of diamagnetism, quoted in most electromagnetism and materials sciences textbooks for a century and a half, is still valid. A substance shaped in the form of a bar or needle and suspended freely between the poles of magnet is called paramagnetic or magnetic if it orients itself longitudinally with the magnetic lines of force, i.e., pointing towards the poles; it is called diamagnetic if it orients itself equatorially, i.e., perpendicular to the lines of force. Already in 1848 Mrs. Somerville wrote "Substances affected after the manner of bismuth (when suspended between the poles of an electromagnet) are said to be diamagnetic." To this day the semimetal bismuth is taken as the prototype of a diamagnetic substance.

Although the concept of magnetism of materials predated quantum mechanics by some 80 years, magnetism is a purely quantum-mechanical effect. Strictly classical systems in equilibrium cannot display a magnetic moment, even in the presence of a magnetic field. Understanding diamagnetism, therefore, requires a complete quantum-mechanical treatment.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. THERMODYNAMIC ARGUMENTS

In the presence of an applied external magnetic field **B**, the free energy $F(T,\mathbf{B})$ of a given system at a given temperature T can be expanded in a Taylor series of the form

$$F(T,\mathbf{B})=F_0(T)-\mathbf{M}_0(T)\cdot\mathbf{B}$$
$$-(1/2\mu_0)\sum_{ij}B_i\chi_{ij}(T)B_j. \tag{1}$$

In Eq. (1) the vector $\mathbf{M}_0(T)$ is called the temperature-dependent intrinsic magnetization, and the second-rank quantity $\chi(T)$ is the magnetic susceptibility tensor. The magnetization $\mathbf{M}(T)$ is defined by

$$\mathbf{M}(T)=\mathbf{M}_0(T)+(1/\mu_0)\chi\cdot\mathbf{B}. \tag{2}$$

Materials for which $\mathbf{M}_0(T)$ is nonzero at a particular T are called *ferromagnets* or *ferrimagnets* at that temperature, irrespective of the properties of $\chi(T)$. Such is the case, for instance, of the metals Fe, Co, Ni, and Gd and the compounds CrO_2, EuS, and Fe_3O_4, *inter alia*, at low enough temperatures.

If, on the other hand, $\mathbf{M}_0(T)=0$, then the magnetic properties are dominated by $\chi(T)$. In particular, for isotropic systems where χ is a diagonal tensor, $\chi_{ij}(T) = \chi(T)\delta_{ij}$, if $\chi(T)$ is positive, then the substance is said to be *paramagnetic*, whereas if $\chi(T)$ is negative, then it is labeled *diamagnetic*. The definitions can be extended to particular directions of anisotropic systems.

Examples of paramagnets are Fe, Co, Ni, and Gd at high temperatures, most rare-earth compounds, and gaseous molecular oxygen. Examples of diamagnets include the noble gases He, Ne, Ar, Kr, and Xe, most molecular gases, liquids, and solids, and the semimetals As, Sb, Bi, and graphite.

It should be pointed out that the free energy of a diamagnet increases in the presence of a magnetic field, and therefore the system evolves so as to avoid high-field regions of space; there is a net repulsive force, caused by the magnetic field, acting on the diamagnet. Conversely the free energy of paramagnet decreases in the presence of a magnetic field, with an attendant attractive force.

2. DIAMAGNETISM OF ATOMS, MOLECULES, AND MOLECULAR SOLIDS

Lenz's law in electromagnetism (see ELECTRODYNAMICS, CLASSICAL) establishes that when the magnetic flux through an electrical circuit is changed, an induced current appears so as to oppose the flux change. In classical systems that current rapidly decays through dissipation, and a new state of equilibrium is achieved with no current flowing through the circuit, and with the new value of the magnetic flux permanently established. In quantized systems with a discrete, nondegenerate energy spectrum, such as a closed-shell atom or molecule, the induced current persists for as long as the applied field is present, and the net magnetic field—the sum of the applied field and the internal one generated by the electronic current in the atom—is always smaller than the applied field.

In addition, closed-shell systems, because of their nondegenerate ground state, possess no intrinsic magnetic moment and are therefore diamagnetic. (It should be remembered that intrinsic magnetic moments are caused by nonvanishing orbital or spin angular momenta which are able to reorient themselves in the presence of a magnetic field; that capability of reorientation requires degeneracy of the ground-state manifold.)

2.1 Hamiltonian of Electrons in a Magnetic Field

A well-established result, valid in both classical and quantum mechanics, states that a system of electrons in a magnetic field **B**, defined by a vector potential **A**, such that

$$\nabla\times\mathbf{A}=\mathbf{B}, \tag{3}$$

is described by the following Hamiltonian:

$$H=\frac{1}{2m}\sum_i[\mathbf{p}_i+e\mathbf{A}(\mathbf{r}_i)]^2+V-\sum_i e\phi(\mathbf{r}_i)$$
$$+\frac{e}{m}\sum_i\sigma_i\cdot\nabla\times\mathbf{A}(\mathbf{r}_i). \tag{4}$$

In Eq. (4) m is the mass of the electron, e is the (positive) charge of the proton, V is the internal electrostatic energy of the electronic system, ϕ is the electrostatic (scalar) potential, and $\mathbf{r}_i$, $\mathbf{p}_i$, and $(e/m)\,\sigma_i$ are the position, the momentum, and the spin magnetic moment of the *i*th electron, respectively.

For a single *atom* or *ion* in a uniform magnetic field of magnitude B oriented in the z direction, it is convenient to choose the so-called symmetric gauge

$$\mathbf{A}=\tfrac{1}{2}\mathbf{B}\times\mathbf{r}, \tag{5}$$

where the origin of the gauge is located at the nucleus. Under such conditions the Hamiltonian (4) reduces to

$$H=\sum_i\left[\frac{1}{2m}p_i^2+e\phi(\mathbf{r}_i)+\frac{eB}{2m}(L_{zi}+2\sigma_{zi})+\frac{e^2B^2}{8m}(x_i^2+y_i^2)\right]+V. \tag{6}$$

The terms in the summation correspond, for each electron, to the kinetic energy, the outside potential energy, the coupling of the magnetic moments arising from the orbital (L_{zi}) and the spin (σ_{zi}) magnetic moments of the electrons to the magnetic field, and the weak quadratic term, proportional to B^2.

2.2 Langevin Theory for Closed-Shell Systems

Diamagnetism arises in systems where the contribution from the third term in Eq. (6)—linear in B—vanishes, i.e., for atoms and molecules with a closed-shell structure. The magnetic-field dependence of the energy originates then exclusively from the term quadratic in B. For a system of N *atoms* or *ions* per unit volume, each consisting of n electrons which have an average square radial distance from the nucleus equal to $\langle r^2\rangle$, the summation

$$\sum_i (x_i^2+y_i^2)$$

in Eq. (6) can be replaced by $\frac{2}{3}Nn\langle r^2\rangle$. This substitution results in an excess magnetic energy per unit volume of

$$(e^2B^2/12m)\ nN\langle r^2\rangle,$$

and, consequently, in a diamagnetic susceptibility (Langevin, 1905)

$$\chi=-(\mu_0e^2/6m)\ nN\langle r^2\rangle. \tag{7}$$

If the system (solid, liquid, or gas) consists of N_α atoms or ions of the α type per unit volume, and all these species are closed electronic shells, then the diamagnetic susceptibility of the system is given by

$$\chi=-\frac{\mu_0e^2}{6m}\sum_\alpha n_\alpha N_\alpha\langle r_\alpha^2\rangle, \tag{8}$$

where n_α and $\langle r_\alpha^2\rangle$ are, respectively, the number of electrons on the α atom or ion and the mean value of r^2 for the electrons from their nucleus, for that particular species.

It should be noted that the diamagnetism of closed-shell systems is essentially a very weak perturbation of the nondegenerate ground state caused by the presence of the applied magnetic field. It is therefore a purely quantum-mechanical phenomenon—since it requires a discrete, nondegenerate lowest energy state in the spectrum, a feature that can only be achieved quantum mechanically—and an effect that depends extremely weakly on the temperature. This virtual temperature independence is, once again, caused by the same features of the spectrum: Only at inaccessible temperatures are the excited states appreciably populated, the only way in which the various $\langle r_\alpha^2\rangle$ in Eq. (8) can be affected by the temperature.

3. ORBITAL DIAMAGNETISM OF METALLIC ELECTRONS

Metallic electrons are, by their very nature, itinerant. Their wave functions extend essentially over the entire crystal and their energy spectrum is in the form of bands separated, in some cases, by forbidden energy gaps. The single-electron wave functions (see ELECTRONIC STRUCTURE OF SOLIDS) are in the form of Bloch states, i.e., a product of

1. a plane wave of vector **k**,
2. a periodic function with the full periodicity of the crystal lattice, and
3. a spinor.

Each Bloch function is thus labeled by a **k** vector, a band index ν (a positive integer), and a spin index σ (↑ or ↓). The vector **k** is restricted to a finite region in reciprocal space (**k** space), a polyhedron known as the Brillouin zone. The energy ϵ of each state is given by the band-structure function $\epsilon_\nu(\mathbf{k},\sigma)$, which is continuous in **k** space. These features arise from the solution of the Schrödinger equation in a periodic potential, and from the properties of periodic functions and Fourier series in three dimensions.

The occupation of the one-electron energy bands follows Fermi–Dirac statistics (see STATISTICAL MECHANICS, CLASSICAL; STATISTICAL MECHANICS, QUANTUM), with no more than one electron occupying each state. At very low temperatures, $T \to 0$, the Fermi–Dirac distribution is a step function, with all electronic states with energy $\epsilon < \epsilon_F$ being occupied by one electron, and all states with $\epsilon > \epsilon_F$ being empty. The energy ϵ_F is called the Fermi level and, in a metal, falls in the middle of one or more bands, i.e., not in the middle of a forbidden energy gap. The equation

$$\epsilon_v(\mathbf{k},\sigma) = \epsilon_F \tag{9}$$

defines, for a metal, a surface in **k** space called the Fermi surface (see ELECTRONIC STRUCTURE OF SOLIDS). This surface consists of one or more sheets (one or more for each band v which crosses the Fermi level), is defined within the Brillouin zone, and has all the symmetry and periodic properties of the Brillouin zone and **k** space.

Metallic electrons do not form closed electronic shells, and therefore do not fall in the category discussed in the previous section. The fact that the Fermi level falls in the middle of a continuous energy band guarantees that the ground state of the many-electron system is not part of the discrete spectrum, i.e., the ground state is the lowest of a set of continuum states. The response of metallic electrons to an applied magnetic field is, in general, very complex, and contains several contributions of various characteristics which must be treated separately:

1. If the bands of different spin are not identical and are occupied differently then the system develops a spontaneous local magnetization which, depending on the way it distributes itself, results in magnetic ordering (see MAGNETIC ORDERING IN SOLIDS); these ordered states include ferromagnets, ferrimagnets, antiferromagnets, spiral spin arrangements, etc. (see FERROMAGNETISM AND FERRIMAGNETISM).
2. If the bands of different spin are identical—and identically occupied—in the absence of an applied magnetic field, then the electron spin and its associated magnetic moment are responsible for a rearrangement of the band occupation when a magnetic field is applied; the imbalance in the spin distribution results in a temperature-independent positive susceptibility known as Pauli paramagnetism (see PARAMAGNETISM). For simple metals of columns I, II, III, and IV of the periodic table that contain only fully occupied and/or fully empty *d* and *f* atomic shells, a commonly used approximation is the so-called *free-electron* model. This model replaces the periodic potential of the lattice by a constant, i.e., the periodic oscillations of the potential are completely smoothed out. As a consequence the periodic factor in the Bloch state reduces to a plane wave of vector $\mathbf{G}_v$; the set of $\mathbf{G}_v$'s is infinite but discrete and constitutes the so-called the reciprocal lattice. The energy spectrum for the Bloch states in the free-electron model is given by

$$\epsilon_v(\mathbf{k},\sigma) = \frac{\hbar^2}{2m} (\mathbf{k}+\mathbf{G}_v)^2. \tag{10}$$

In this model the Pauli contribution to the (paramagnetic) susceptibility per unit volume is equal to

$$\chi_{\text{Pauli}} = \mu_0 \frac{\mu^2 m}{\pi^2 \hbar^2} (3\pi^2 Nn)^{1/3}. \tag{11}$$

In Eq. (11) μ is the Bohr magneton, N is the number of metal atoms per unit volume, and n is the number of metallic (conduction) electrons per atom.
3. There is, in addition, an orbital diamagnetic contribution to the susceptibility, known as Landau diamagnetism, discussed below. For the free-electron model this contribution is

$$\chi_{\text{Landau}} = -\mu_0 \frac{\mu^2 m}{3\pi^2 \hbar^2} (3\pi^2 Nn)^{1/3}, \tag{12}$$

i.e., $\chi_{\text{Landau}} = -\frac{1}{3}\chi_{\text{Pauli}}$, which yields a total *paramagnetic* susceptibility

$$\chi_{\text{free electrons}} = \tfrac{2}{3}\chi_{\text{Pauli}} \tag{13}$$

for the free-electron model. The free-electron model, however, has only a very restricted range of applicability. For some substances—the semimetals As, Sb, Bi, and graphite being the most notable examples—real band-structure effects, caused by the periodic potential, produce a large enhancement of the Landau diamagnetism, by factors of up to several hundred. The diamagnetic contribution thus overwhelms the paramagnetic Pauli term; the net result is a negative χ and a diamagnetic solid.

4. The orbital contribution of metallic electrons also includes a diamagnetic term which oscillates with the magnetic field, and is periodic in $1/B$. This oscillatory contribution is called the de Haas–van Alphen effect. Its periods and amplitudes give very precise information on the geometrical and differential properties of the Fermi surface.

3.1 Quantization of Itinerant Electrons in a Magnetic Field

To study diamagnetic effects in metals the effect of the electron-electron interaction can be neglected, and the electrons may be treated as independent particles. Independent electrons moving in a crystal lattice in a magnetic field are described by Eq. (4), with the electron–electron interaction made to vanish, $V=0$, and with the electrostatic term $[-e\phi(\mathbf{r}_i)]$ representing the periodic potential acting on electron i and caused by the lattice of nuclei and ionic cores. The solution of the Schrödinger equation under these conditions is still a very complex and difficult process. The difficulty arises from conflicting periodicities imposed separately by the lattice and the magnetic field. The problem is simple only in the free-electron case, $\phi=0$. The solution was obtained by Landau in 1930, the early days of quantum mechanics. Landau chose the field $\mathbf{B}$ along the z axis and an asymmetric gauge:

$$\mathbf{A}=[0,Bx,0]. \tag{14}$$

The resulting spectrum is

$$\epsilon(\nu,k_z)=(\nu+\tfrac{1}{2})\hbar\omega_c+[\hbar^2/2m]k_z^2, \tag{15}$$

where ν is a nonnegative integer, k_z is the wave vector for motion along the direction of the magnetic field, and ω_c, the so-called cyclotron frequency, is given by

$$\omega_c=eB/m. \tag{16}$$

This spectrum for three-dimensional electrons, which depends only on two quantum numbers (a discrete ν and a continuous k_z), is highly degenerate—each energy is shared by a very large number of states. There is a "hidden" quantum number, k_y, necessary to define the wave function completely. The energy ϵ does not depend on k_y, and all wave functions with the same ν and k_z—no matter what the value of k_y—have the same energy. For each fixed value of the set (ν,k_z), there are

$$L_xL_yeB/2\pi\hbar \tag{17}$$

different k_y states, where L_xL_y is the macroscopic area of the crystal—assumed to be prismatic—perpendicular to the direction of the magnetic field.

Each integer ν defines for the perpendicular motion in the (x,y) plane and for fixed k_z a discrete energy level, called a Landau level.

From this spectrum, Eqs. (15)–(17), its degeneracy, and its dependence on the magnetic field intensity B, all thermodynamic quantities can be calculated for the free-electron model. Some of these quantities are discussed explicitly in the next two subsections.

Extension of Eqs. (15)–(17) to the case of a nonvanishing periodic potential was proposed by Onsager (1952), who based his method on a semiclassical approach. This approach consists of determining electron trajectories and applying to them the Bohr–Sommerfeld "old" quantum-mechanical quantization rules, i.e.,

$$\int_{\text{closed orbit}} \mathbf{p}\cdot d\mathbf{r}=(\nu+\gamma)2\pi\hbar, \tag{18}$$

where ν is a nonnegative integer, γ is a (positive) phase correction which takes the value $\frac{1}{2}$ in the free-electron case, the position $\mathbf{r}$ follows the closed semiclassical trajectories of the electron in the (x,y) plane, $\mathbf{p}$ is the *canonical* momentum

$$\mathbf{p}=\hbar\mathbf{k}-e\mathbf{A}, \tag{19}$$

$\hbar\mathbf{k}$ is interpreted as the *kinetic* momentum, and $-e\mathbf{A}$ is the contribution of the field to the canonical momentum.

The semiclassical equations of motion of electrons in a periodic potential and a magnetic field are given by

$$k_{\text{parallel}}=\text{const}, \tag{20}$$

where k_{parallel} is the component of $\mathbf{k}$ parallel to $\mathbf{B}$,

$$\epsilon_\nu(\mathbf{k},\sigma)=\text{const}, \tag{21}$$

and

$$\hbar\dot{\mathbf{k}} = -e\dot{\mathbf{r}}\times\mathbf{B}, \tag{22}$$

which integrated and with proper choice of the origin yields

$$\hbar\mathbf{k} = -e\mathbf{r}\times\mathbf{B} \tag{23}$$

for the motion perpendicular to **B**.

Insertion of Eqs. (23) and (19) into Eq. (18), use of Stokes's theorem, and application of Eq. (3) yields Onsager's famous result for the quantization of itinerant electron orbits in a uniform magnetic field (Onsager, 1952),

$$\int_{\text{closed orbit}} \mathbf{p}\cdot d\mathbf{r} = e\Phi = (\nu+\gamma)2\pi\hbar, \tag{24}$$

for the motion of the electrons perpendicular to **B**. In Eq. (24) Φ is the magnetic flux encircled by the electron in its trajectory in the plane perpendicular to **B**. In other words the only trajectories that are (quantum-mechanically) allowed are those that, in the plane perpendicular to **B**, encircle a magnetic flux Φ which takes the discrete values

$$\Phi_\nu = (\nu+\gamma)\,\frac{2\pi\hbar}{e}. \tag{25}$$

Because the magnetic flux is quantized, and the value of the magnetic field B is externally determined, the areas of the allowed orbits in real space, A_ν, must take well-defined, discrete values. In addition, Eqs. (20), (21), and (23) state that the orbits in real space in the plane perpendicular to **B** are proportional to—and rotated by an angle of $\pi/2$ from—the corresponding trajectories in **k** space. The cross sections (20) at a constant k_{parallel} of the surface of constant energy in **k** space, Eq. (21), are therefore proportional to the areas A_ν of the allowed orbits in real space. The constant of proportionality between areas in **k** space and real space, as given by Eq. (23), is proportional to B^2, i.e., equal to $(eB/\hbar)^2$. Therefore, the allowed cross-sectional areas S_ν of the surfaces of constant energy in **k** space are given by

$$S_\nu = (\nu+\gamma)\,\frac{2\pi e}{\hbar}\,B. \tag{26}$$

Because the cross-sectional areas in **k** space S_ν perpendicular to **B** are quantized—take only the discrete values given by Eq. (26)—the energies for the motion perpendicular to **B** are also quantized. Two successive levels, which differ by $\Delta\nu = \pm 1$, are separated by an energy $\Delta\epsilon$, which defines, for that particular value of the energy and that particular value of k_z, a cyclotron frequency ω_c^* and an effective cyclotron mass m_c^*:

$$\Delta\epsilon = \hbar\omega_c^* = \hbar eB/m_c^*. \tag{27}$$

When the Onsager rules (26) are applied to the free-electron model, the results (15)–(17) are obtained if $\gamma = \frac{1}{2}$. For this particular case $\omega_c^* = \omega_c$, as given by Eq. (16), and $m_c^* = m$, i.e., the cyclotron mass is equal to the electron mass.

It should be mentioned that transition between two adjacent Landau levels, with conservation of k_{parallel}, can be induced, for electrons near the Fermi level, by microwave or far-infrared radiation; when the frequency ω of that radiation is in the vicinity of ω_c, a resonance—the so-called cyclotron resonance or diamagnetic resonance—is observed.

3.2 Landau Diamagnetism

For noninteracting electrons, the total energy of the system—the metal—is obtained by adding the individual energies of each electron. Since electrons satisfy Fermi–Dirac statistics (see STATISTICAL MECHANICS, QUANTUM), in the limit $T\rightarrow 0$ that total energy is obtained by summing one-electron energies for all states with energy below the Fermi level ϵ_F.

The term $\frac{1}{2}$ in the factor $\nu+\frac{1}{2}$ in Eq. (15)—or equivalently the term γ in the factor $\nu+\gamma$ in Eqs. (25) and (26)—, which is ordinarily called the zero-point energy, guarantees that although upon the application of a magnetic field some states increase their energy and others decrease it, the energy sum over states of contiguous occupation up to the Fermi level always shows an *increase* by an amount proportional to B^2, thus yielding a diamagnetic contribution to the susceptibility. This is the Landau diamagnetic contribution to the susceptibility of metallic electrons.

The summation is, in general, difficult to perform. It can be done numerically—and laboriously—for some specific cases, where the number of carriers is small, i.e., for the semimetals. For the free-electron model the summation can be accomplished in closed form. The result is that quoted in Eq. (12).

3.3 The de Haas–van Alphen Effect

The de Haas–van Alphen effect is an oscillatory variation in the magnetic susceptibility of metals as a function of the magnitude of the *static* magnetic field intensity B. It is an effect observed in strong magnetic fields, for pure specimens, and at low temperatures. These three conditions arise from the necessity to have the Landau levels widely separated, not blurred by electron collision effects with impurities, and populated with a sharp discontinuity at the Fermi level, i.e., a discontinuity in occupation of essentially 1 between the last "occupied" and the first "empty" levels. These conditions may be summarized by two inequalities:

$$k_B T \ll \hbar\omega_c^*, \quad \omega_c^*\tau \gg 1, \tag{28}$$

where ω_c^* is the cyclotron frequency defined by Eq. (27), k_B is Boltzmann's constant, and τ, the relaxation time, is the median time between electron collisions (see THERMODYNAMICS, EQUILIBRIUM; TRANSPORT PHENOMENA).

The *loci* of allowed quantum states—in **k** space—for itinerant electrons in a periodic lattice and a magnetic field are "tubes," i.e., cylindrical shells of well-determined, discrete cross sections, given by Eq. (26). The cylinders, not necessarily of circular cross section, are inside each other in order of either increasing or decreasing energy. For the free-electron model the tubes are concentric circular cylinders of axis parallel to **B**. These highly degenerate tubes, when the conditions (28) are satisfied, are populated by electrons if the energy of their states is lower than ϵ_F, i.e., states are occupied for that part of those tubes that lies inside the Fermi surface (9). Only *some* parts of *some* tubes are populated by electrons. The rest—the states for which $\epsilon > \epsilon_F$—are allowed but empty states. This distribution is illustrated in Fig. 1.

As the magnitude B of the field increases the tubes get more widely separated and contain more states—as indicated by Eqs. (26) and (17). The tubes thus get repopulated at each increasing value of the field. For each orientation (θ,ϕ) of the magnetic field **B** there are discrete values of the field intensity B_ν for which a given tube gets completely depopulated, i.e., it "leaves" the Fermi surface and lies completely outside. These values of the field B_ν satisfy the condition

$$S_\nu = (\nu+\gamma)\,\frac{2\pi e}{\hbar}\,B_\nu = S_{\text{Fermi surface}}(\theta,\phi), \tag{29}$$

where $S_{\text{Fermi surface}}(\theta,\phi)$ is the *extremal* cross-sectional area of the Fermi surface in the direction perpendicular to **B**. This phenomenon of allowed-energy tubes "peeling off" the Fermi surface results in a contribution to the free energy, the magnetization, and the susceptibility of the sample which is periodic in $1/B$, with period

$$\Delta\left(\frac{1}{B}\right) = \frac{2\pi e}{\hbar S_{\text{Fermi surface}}(\theta,\phi)}, \tag{30}$$

i.e., a $1/B$ period inversely proportional to the extremal cross-sectional area of the Fermi surface in the direction (θ,ϕ) perpendicular to **B**.

If the Fermi surface is complex and exhibits several sheets and/or several extremal cross-sectional areas for a given sheet, the magnetization becomes a very complicated oscillatory function. The analysis of the various periods in $1/B$ permits the determination of the corresponding cross-sectional areas.

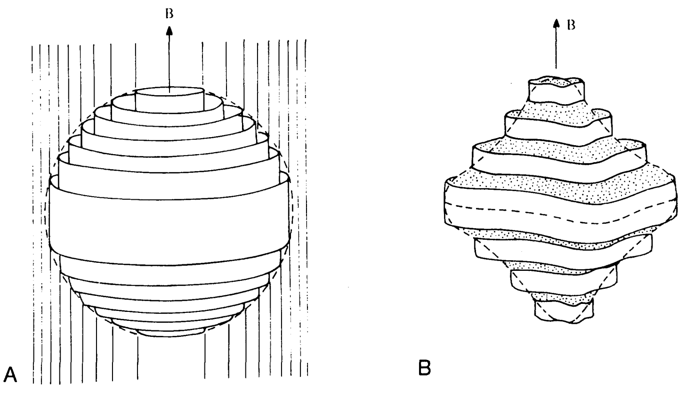

FIG. 1. Schematic representation of the quantization of metallic electrons in a magnetic field. The depicted tubes are the *loci* of allowed states in **k** space. Only the occupied portions of the tubes are drawn. (a) The free-electron model. (b) An arbitrary potential. [From Falicov (1973)].

Studies of the variation of the oscillations with orientation (θ,ϕ) of the magnetic field lead to accurate determination of the geometry and topology of the Fermi surface of the metal under consideration.

The amplitude of the oscillations and its variation with temperature yield information about the separation between the levels—the cyclotron mass m_c^* at the Fermi level for a particular orientation (θ,ϕ), an intrinsic property of the metal—as well as the relaxation time τ, a specific property of the sample under study. The de Haas–van Alphen effect provides the most useful tool to investigate the electronic properties of metals. It yields extremely accurate information on the structural and dynamical properties of itinerant-electron states in solids.

4. DIAMAGNETISM IN SUPERCONDUCTORS

Superconductors are perfect diamagnets for a given range of intensities of applied magnetic fields. They are also perfect conductors, i.e., they have zero resistivity. Although the zero resistivity is a necessary condition for perfect diamagnetism—since it is required for the induced currents in Lenz's law not to decay—by itself it only implies that the magnetic field in the bulk of a superconductor cannot be changed, whatever its value. The expulsion of the field from the bulk of a superconductor is an additional property, known as the Meissner effect (Meissner and Ochsenfeld, 1933). It is illustrated in Fig. 2.

Superconductivity (see SUPERCONDUCTIVITY, LOW TEMPERATURE) is a phase, a state of matter (in the sense that ice and steam are phases of water and diamond and graphite are phases of pure carbon) observed *only* in some solids, mostly metals. The superconducting state has several characteristic properties:

1. When it exists for a given substance, it exists only at temperatures below a so-called transition temperature T_c, and in general down to the absolute zero of the temperature scale 0 K. Historically, substances with the highest known values of T_c have been first pure metals (Hg with $T_c = 4.2$ K, Pb with $T_c = 7.2$ K, and Nb with $T_c = 9.5$ K), later transition-metal compounds (NbN with $T_c = 16.0$ K, V_3Si with $T_c = 17.1$ K, Nb_3Sn with $T_c = 18.05$ K, and Nb_3Ge with $T_c = 23.2$ K), and, since 1986, the so-called *high-T_c* materials, all four- or five-element compounds which contain copper and oxygen (La–Ba–Cu–O with T_c of approximately 30 K, rare-earth–Ba–Cu–O with T_c of approximately 92 K, Bi–Sr–Ca–Cu–O with T_c of approximately 105 K, and—the known record at this time—Tl–Ca–Ba–Cu–O with $T_c = 125$ K) (see SUPERCONDUCTIVITY, LOW-TEMPERATURE; HIGH-TEMPERATURE SUPERCONDUCTIVITY).
2. It exhibits zero dc resistivity, i.e., infinite conductivity for zero-frequency measurements (an effect discovered in mercury by Kamerlingh Onnes in 1911),

$$\rho(\omega=0;T<T_c)=0. \tag{31}$$

3. It exhibits, for weak magnetic fields, perfect diamagnetism, i.e., its magnetic susceptibility is given by

$$\chi=-1 \tag{32}$$

which means that magnetic flux lines are completely expelled from the superconductor and that there is a force pushing superconductors away from magnetic fields. This is the Meissner effect discovered by Meissner and Ochsenfeld (1933).

4. In addition to the effect of high tempera-

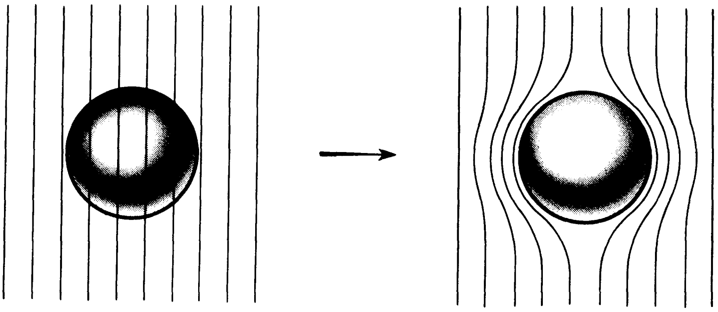

FIG. 2. Schematic representation of the Meissner effect. When a superconducting sphere is cooled in a constant applied magnetic field, below the transition temperature T_c the lines of induction **B** are ejected from the sphere. [From Kittel (1986)].

tures, superconductivity can be destroyed (with a return to the normal metallic state) by either a large enough electric current $I > I_c$ or a large enough applied magnetic field $B > H_c$. The quantities I_c and H_c are called the critical current and the critical magnetic field, respectively.

5. Superconductivity is a macroscopic quantum phenomenon with amplitudes and phases associated with an order parameter ψ. Therefore, interference and diffraction effects can be achieved, in particular the Josephson effect (Josephson, 1962). These effects can be fruitfully employed in processing, storing, and retrieving information, i.e., in computer technology (see SUPERCONDUCTING DEVICES).

4.1 The Meissner Effect

Meissner and Ochsenfeld (1933) discovered that, when a superconductor is cooled below T_c in the presence of a magnetic field, at the transition the lines of induction **B** are completely expelled from the bulk of the sample. In other words, inside the superconductor Eq. (32) is obtained or, equivalently,

$$\mathbf{B} = 0. \tag{33}$$

This result holds up to a given value of the applied field B. In the so-called type-I superconductors there is only one critical field, H_c. For $B > H_c$ superconductivity is destroyed, the flux fully penetrates the sample, and the metal returns to its normal state.

In type-II superconductors there is no field penetration for $B < H_{c1}$ and the superconductor acts as a perfect diamagnet. For intermediate field strengths $H_{c1} < B < H_{c2}$, the magnetic flux lines partially penetrate the superconductor but do not destroy the superconducting state. Since, in general, H_{c2} is very large, this feature of type-II superconductors makes them useful as materials for magnets (see SUPERCONDUCTING MAGNETS). Finally for $B > H_{c2}$ superconductivity is again destroyed with a return of the substance to its normal state. This behavior is illustrated in Fig. 3.

4.2 The London Theory

The London theory (London and London, 1935) is completely classical. It incorporates into Maxwell's equations, phenomenologically, the electrodynamic characteristics of superconductors. The zero-resistivity property (31) requires that, since

$$\rho \mathbf{j} = \mathbf{E},$$

the electric field **E** should vanish everywhere in a superconductor. The Meissner effect, Eqs. (32) and (33), implies vanishing magnetic field **B** in the bulk. These two conditions led London, in 1950, to *postulate* that in a superconductor the current density **j**, instead of being proportional to **E**, as is the case in other media, is proportional to the vector potential **A**,

$$\mathbf{j} = -[1/\mu_0\lambda_L^2]\mathbf{A}, \tag{34}$$

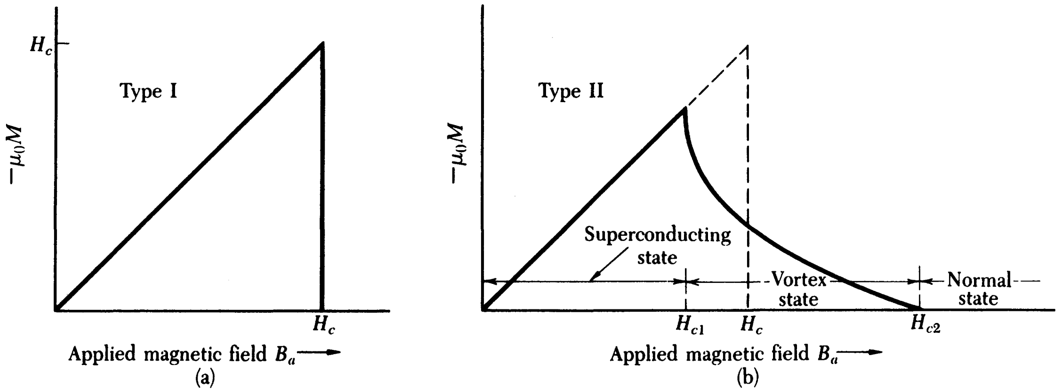

(a) (b)

FIG. 3. (a) Magnetization vs applied magnetic field for a type I superconductor exhibiting the Meissner effect (perfect diamagnetism) below H_c. (b) Magnetization vs applied magnetic field for a type II superconductor. [From Kittel (1986)].

where the constant of proportionality defines the length λ_L, the *London penetration depth.* The London equation (34), by taking the curl on both sides, may also be expressed as

$$\nabla\times\mathbf{j}=[1/\mu_0\lambda_L^2]\mathbf{B}. \tag{35}$$

Since the vector potential **A** is *not*, in general, a physically measurable quantity, and the current density **j** *always is*, Eq. (34) in effect selects a particular gauge, the so-called London gauge. Equation (34) must be consistent with Maxwell's equations. Therefore, it implies that in the absence of charges $\nabla\cdot\mathbf{A}=0$, and that the normal components of **A** are zero on any external surface through which no external current is fed. Additional conditions on the gauge, compatible with the physical boundary conditions, should be imposed on problems with multiply connected geometries.

It is easily seen that the London equation implies the Meissner effect, and that λ_L is the characteristic length describing the penetration and exponential decay of the magnetic field **B** into a superconductor. From Maxwell's equation,

$$\nabla\times\mathbf{B}=\mu_0\mathbf{j}, \tag{36}$$

when the curl is taken on both sides and use is made of Eq. (35) and of Maxwell's other equation, $\nabla\times\mathbf{B}=0$, one obtains

$$\nabla^2\mathbf{B}=\mathbf{B}/\lambda_L^2. \tag{37}$$

Equation (37) does not allow a constant-**B** solution, unless $\mathbf{B}=0$. In addition Eq. (36) guarantees that $\mathbf{j}=0$ wherever **B** vanishes. In a superconductor extending over the half-space $x \geqslant 0$, if the field at the boundary $x=0$ is parallel to the surface and of magnitude B_0, a solution of Eq. (37) yields

$$B(x)=B_0\exp(-x/\lambda_L), \tag{38}$$

i.e., the field decays exponentially into the superconductor, a perfect diamagnet, with characteristic length given by the *London penetration depth.*

4.3 The Ginzburg–Landau Theory

4.3.1 General Formulation. In 1950 Ginzburg and Landau (GL) proposed a macroscopic, phenomenological theory of superconductivity, which was independent of the microscopic aspects of the phenomenon. The theory is quantum-mechanical, in the sense that it includes coherent, macroscopic quantum effects. It is a pioneering theory which, independent of the mechanisms responsible for superconductivity, is still valid today. It contains such diverse phenomena as magnetic-field penetration depths, coherence lengths, magnetic-field flux quantization, magnetic-field dependence of the superconducting order parameter, and the Josephson effect. It can be applied to all superconductors, as well as to superfluid ^{3}He, and has become the prototype theory to study a whole class of phenomena related to spatial dependence in second-order phase transitions.

The GL theory introduces a complex order parameter ψ which is allowed to vary in space. Originally GL interpreted ψ as an amplitude, and $|\psi|^2$ as the density of the "superconducting" electrons (they envisioned a superconductor as two interpenetrating electron fluids: the nondissipative, nonresistive superconducting electron fluid, and the dissipative, resistive "normal" electron fluid). In 1959, however, Gor'kov proved that for temperatures below and close to T_c, equations identical to those of GL could be obtained from the Bardeen–Cooper–Schrieffer (BCS) microscopic theory of superconductivity (1957), and that the GL parameters ψ could be interpreted (except for a trivial constant of proportionality) as what in the BCS theory is known as the energy-gap parameter.

The starting point of the GL theory is the introduction of a magnetic Helmholtz free energy F_{SH} for the superconductor, derived from plausibility arguments:

$$F_{SH}=\int_{\text{superconductor}} d^3r\biggl[F_{N0}+\Delta F(|\psi|^2) + \left(\frac{1}{2m}\right)\left|-i\hbar\nabla\psi-e^*\mathbf{A}\psi\right|^2 + \left(\frac{1}{2\mu_0}\right)B^2(\mathbf{r})\biggr]. \tag{39}$$

Here F_{N0} is the free-energy density of the normal state in the absence of a magnetic field; ΔF is the difference of free-energy densities between the superconducting and the normal states (also in the absence of a field) and is a function of $|\psi|^2$. The third term is the gauge-invariant "superconducting kinetic energy," and the last term is the magnetic-field energy in the superconductor. The vector potential is $\mathbf{A}$, $\mathbf{B}$ is the magnetic field, and e^* is an effective charge, known now to be $-2e$, twice the charge of the electron (the charge of a "Cooper pair" in the BCS theory). All terms in Eq. (39) are functions of the position $\mathbf{r}$, and change with the magnitude and direction of the magnetic field.

Because the proper variables of the magnetic Helmholtz free energy are the temperature T and the magnetization $\mathbf{M}$, where

$$\mathbf{M} = (1/\mu_0) \int_{\text{all space}} d^3r[\mathbf{B}(\mathbf{r}) - \mathbf{B}_a],$$

$$\mathbf{B}_a = \text{applied magnetic field}, \tag{40}$$

F_{SH} is not continuous at the critical fields. The function which is continuous at H_{c1} and H_{c2}, and whose proper variables are T and $\mathbf{B}_a$, is the Gibbs free energy G_{SH}, given by

$$G_{\text{SH}} = F_{\text{SH}} - \mathbf{M}\cdot\mathbf{B}_a. \tag{41}$$

Substitution of Eqs. (39) and (40) into Eq. (41) yields

$$\begin{aligned} G_{\text{SH}} = G_{\text{NH}} + \int_{\text{superconductor}} & d^3r[\Delta F(|\psi|^2) \\ & + (1/2m)|-i\hbar\nabla\psi - e^*\mathbf{A}\psi|^2] \\ & + \frac{1}{2\mu_0} \int \int_{\text{all space}} d^3r[\mathbf{B}(\mathbf{r}) - \mathbf{B}_a]^2, \end{aligned} \tag{42}$$

where

$$G_{\text{NH}} = \int_{\text{superconductor}} d^3r \left[F_{N0} + \frac{B_a^2}{2\mu_0} \right].$$

It should be noted that the last term in Eq. (42) is to be integrated over *the whole space* (both in the superconductor and outside). Minimization of G with respect to the four functions ψ and $\mathbf{A}$ (or equivalently ψ and $\mathbf{B}$) yields the GL equations:

$$\nabla^2\mathbf{A} = \frac{ie^*\hbar}{m}(\psi^*\nabla\psi - \psi\nabla\psi^*) + \frac{\mu_0 e^{*2}}{m}|\psi|^2\mathbf{A}, \tag{43}$$

$$\nabla\times\mathbf{A} = \mathbf{B}_a \text{ on surface}, \tag{44}$$

$$\frac{\partial\Delta F}{\partial\psi^*} + \frac{1}{2m}(-i\hbar\nabla - e^*\mathbf{A})^2\psi = 0, \tag{45}$$

$$(i\hbar\nabla\psi + e^*\mathbf{A}\psi)_{\text{perpendicular}} = 0, \tag{46}$$

where the London gauge

$$\nabla\cdot\mathbf{A} = 0$$

has been chosen.

4.3.2 Penetration Depth and Coherence Length. In singly connected samples with no penetration of the magnetic flux into the bulk superconductor, the phase of ψ can be chosen so that ψ is real throughout the sample. In particular for a one-dimensional, singly connected problem, with quantities varying along the x axis, and with magnetic field and vector potential given by

$$\mathbf{B} = [0,0,B(x)],$$

$$\mathbf{A} = [0,A(x),0],$$

Eqs. (43)-(46) become

$$\frac{d^2A}{dx^2} = (\mu_0 e^{*2}/m)\psi^2 A, \tag{47}$$

$$\frac{dA}{dx} = H_0 \text{ on surface}, \tag{48}$$

$$\frac{\partial\Delta F}{\partial\psi} + \frac{e^{*2}}{m}A^2\psi = \frac{\hbar^2}{m}\frac{d^2\psi}{dx^2}, \tag{49}$$

$$\frac{d\psi}{dx} = 0 \text{ on surface}. \tag{50}$$

For the free-energy difference ΔF, the original GL derivation used a power-series expansion in $|\psi|^2$, and neglected all terms higher than the second. That expansion is still commonly used, and is known to be valid for superconductors at temperatures close to T_c:

$$\Delta F = \frac{H_{cb}^2}{2\mu_0}\left[-2\left(\frac{\psi}{\psi_T}\right)^2 + \left(\frac{\psi}{\psi_T}\right)^4\right], \tag{51}$$

where H_{cb} is the thermodynamic bulk critical field, and ψ_T is the equilibrium value of ψ in the bulk, at temperature T, in the absence of a magnetic field.

The problem of the superconducting half-space discussed above—in Eq. (38)—in connection with the London equation can be solved in the GL theory. Integration of Eqs. (47)-(51), under the assumption of small changes in ψ near the surface, yields for $x > 0$

$$\begin{aligned} \frac{\psi(x,B_a)}{\psi_T} = 1 - \frac{\kappa_0}{(2-\kappa_0^2)\sqrt{8}}\left(\frac{B_a}{H_{cb}}\right)^2 & [e^{-\sqrt{2}\kappa_0 x/\lambda_{\text{L}}} \\ & - \tfrac{1}{2}\kappa_0 e^{-2x/\lambda_{\text{L}}}] \end{aligned} \tag{52}$$

and

$$B(x) \approx B_0 \exp(-x/\lambda_{\mathrm{L}}). \tag{53}$$

Equation (53) reproduces, approximately, Eq. (38). In the GL equations λ_{L}, the London penetration depth, is given by

$$\lambda_{\mathrm{L}}^2 = m/\mu_0 e^{*2}\psi_T^2. \tag{54}$$

In Eq. (52) κ_0 is a dimensionless constant:

$$\kappa_0 = (\sqrt{2}e^*/\hbar)\lambda_{\mathrm{L}}^2 H_{cb}. \tag{55}$$

Two remarks are necessary at this point. First, there are two length scales in the problem: the decay length for magnetic fields, λ_{L}, and the decay length $\lambda_{\mathrm{L}}\sqrt{2}\kappa_0$ for the order parameter ψ, given by the first exponent in Eq. (52). Second, Gor'kov (1959) has shown that

$$\kappa_0 \approx 0.96\lambda_{\mathrm{L}}\xi_0^{-1}, \tag{56}$$

where ξ_0 is the coherence length introduced by Pippard (1953) in connection with the general electromagnetic properties of superconductors (see SUPERCONDUCTIVITY, LOW TEMPERATURE). Values of κ_0 are small (<0.707) for the soft, type-I superconductors (0.01 for Al; 0.3 for Pb), whereas it takes large values (>0.707) for the hard, type-II superconductors [~ 8 for V; extremely large for the new, high-T_c materials (see HIGH-TEMPERATURE SUPERCONDUCTIVITY)].

A type-I superconductor excludes a magnetic field from its bulk completely. If the magnetic field B is increased there is a value H_c for which the superconductivity is suddenly destroyed, the system returns to the normal state, and the magnetic field penetrates the specimen completely. A type-II superconductor excludes the field completely up to a value H_{c1}. Above H_{c1} the field is partially excluded, although the specimen remains superconducting and exhibits zero resistivity. At a higher field, H_{c2}, the flux penetrates completely, superconductivity is destroyed, and the specimen returns to its normal state.

4.3.3 Flux Quantization. In many applications (thin specimens, weak magnetic fields, etc.), the order parameter ψ can be considered to have a constant magnitude $n^{1/2}$, although its phase $\theta(\mathbf{r})$ can vary appreciably in space,

$$\psi = n^{1/2}e^{i\theta(\mathbf{r})}. \tag{57}$$

From standard quantum-mechanical arguments the electrical supercurrent is given in this case by the usual formula

$$\begin{aligned}\mathbf{j} &= \frac{1}{2m}[\psi^*(-i\hbar\nabla - e^*\mathbf{A})\psi \\ &\quad + \psi(i\hbar\nabla - e^*\mathbf{A})\psi^*] \\ &= \frac{ne^*}{m}(\hbar\nabla\theta - e^*\mathbf{A}).\end{aligned} \tag{58}$$

Deep inside any superconductor the magnetic field $\mathbf{B}$ and the electric current $\mathbf{j}$ are zero, and therefore from Eq. (58) one obtains

$$e^*\mathbf{A} = \hbar\nabla\theta. \tag{59}$$

In a multiply connected sample one can find a closed path C which encircles a nonsuperconducting region where there may be a magnetic field. Line integration of Eq. (59) over that path, use of Stokes's theorem, and knowledge that ψ must be single valued yields

$$\begin{aligned}\int_{\text{closed } C} \mathbf{A}\cdot d\mathbf{s} &= \int_{\text{area } C} \nabla\times\mathbf{A}\cdot d\boldsymbol{\sigma} \\ &= \int_{\text{area } C} \mathbf{B}\cdot d\boldsymbol{\sigma} = \Phi \\ &= \frac{\hbar}{e^*}\int_{\text{closed } C} \nabla\theta\cdot d\mathbf{s} \\ &= \frac{\hbar}{e^*}\times 2\pi\nu,\end{aligned} \tag{60}$$

where Φ is the magnetic-field flux, and ν is an arbitrary integer. In other words Eq. (60), taking $|e^*| = 2e$, can be written

$$\Phi = \nu\Phi_0 = \nu\cdot 2.0678\times 10^{-15}\ \mathrm{T\ m^2}, \tag{61}$$

i.e., if a closed path without currents can be established deep inside a multiply connected superconductor, then the magnetic-field flux encircled by that path is quantized in units of Φ_0. The unit of flux, Φ_0, is called a *fluxoid.*

4.3.4 Phase-Current Relationship; The Josephson Effect. From the GL equations it can be easily seen that the order parameter ψ has an indeterminate *arbitrary, constant* phase. In a given superconductor (called 1) its phase θ_1 is completely arbitrary. If, however, there is nearby a second superconductor (called 2), which is *weakly* connected to the first one, although both phases, θ_1 and θ_2, are indeterminate by the *same* additive constant, the *phase difference* between the two,

$$\delta = \theta_2 - \theta_1,$$

is an observable meaningful quantity. As can be seen from Eq. (58), a variation of θ over space is responsible for the existence of a supercurrent. Similarly a phase difference between two weakly coupled, spatially close superconductors produces a current flow between them given by

$$J = J_0 \sin\delta, \qquad (62)$$

where J_0, a constant, describes the maximum possible current which may flow between the two specimens. Equation (62) is Josephson's dc equation relating current and phase difference (Josephson, 1962). It is implicit in the GL equations and applies to any system with a macroscopic, quantum-mechanical, complex order parameter. It can be interpreted in terms of the standard quantum-mechanical uncertainty relation between particle number and wave-function phase.

4.3.5 Magnetic-Field Dependence of the Order Parameter. Detailed solutions of Eqs. (47)–(51) for thin films clearly exhibit a field dependence of the *amplitude* of the order parameter $|\psi|$ on the applied magnetic field strength B_a. As the field is increased the value of $|\psi|$ decreases, and there is a value H_f for which it goes (either continuously or discontinuously) to zero and the film becomes normal. It is found that H_f—the film critical field—depends on H_{cb}, the film thickness d, and the London penetration depth, and that the $|\psi|$ transition to zero at H_f is discontinuous if

$$d/\lambda_L > \sqrt{5}.$$

These results, and many others obtained from the solution of the GL equations for a variety of geometries and situations, have been confirmed by superconducting tunneling experiments.

4.3.6 Quantum Interference Phenomena. Given the following facts:

1. the order parameter ψ is complex;
2. ψ must be single valued;
3. the magnetic field **B** couples to it in a gauge-invariant form and therefore is directly related to the phase θ of ψ; and
4. the GL equations are nonlinear;

it is possible to obtain a large number of interference and diffraction effects which can be fruitfully used in designing interesting electronic devices (see SUPERCONDUCTING DEVICES). It can be said that, in understanding the origins of the diamagnetism in superconductors, scientists and engineers have promoted quantum mechanics to the macroscopic, everyday-use level.

5. CONCLUSION

All diamagnetic effects are quantum-mechanical. The persistence of the currents induced by Lenz's law is caused exclusively by the phase coherence of the quantum-mechanical wave functions, either at the microscopic level—as evidenced in the diamagnetism of atoms, molecules, and metallic electrons—or in their macroscopic manifestations—as in superconductors. One of the principal effects is the gauge-invariant way in which quantum-mechanical phases couple to magnetic fields. This coupling produces, in addition to steady currents that result in a screening of the magnetic field, i.e., normal diamagnetism, a wealth of refraction, interference, and diffraction effects which are directly observable experimentally. The de Haas–van Alphen effect, discussed in Sec. 3, and the Josephson effect and flux quantization in superconductors, presented briefly in Sec. 4, are all manifestations of the relationship between electrical current and quantum-mechanical phase, coherent coupling of magnetic fields to the phases of quantum-mechanical wave functions or order parameters, and the relationship between electrical currents and magnetic fields as given by Maxwell's equations.

GLOSSARY

Antiferromagnet: A substance in which local magnetic moments are arranged in alternating orientations, resulting in no net macroscopic moment.

Bardeen, Cooper, and Schrieffer Theory: See BCS theory.

BCS Theory: A microscopic theory of superconductivity formulated by Bardeen, Cooper, and Schrieffer in 1957.

Bohr–Sommerfeld Quantization Rules: Rules by which the energy of bound electronic states were quantized in the "old" quantum mechanics.

Brillouin Zone: The polyhedron in reciprocal space over which the wave vector **k** is allowed to vary.

Coherence Length: The distance over which the coherence of the quantum-mechanical phase of the superconducting order parameter is preserved.

Cyclotron Resonance: Resonance absorption of electromagnetic radiation obtained when the electromagnetic frequency is equal to the cyclotron frequency of electrons moving in a magnetic field; it is also sometimes called diamagnetic resonance.

de Haas–van Alphen Effect: Oscillatory contribution to the magnetic susceptibility, periodic in $1/B$, caused by the interrelationship of the quantized energy levels and the Fermi surface.

Diamagnet: A substance or body that repels or expels magnetic flux.

Diamagnetic Resonance: See cyclotron resonance.

Fermi–Dirac Statistics: The thermal-equilibrium statistics that applies to particles which satisfy the Pauli exclusion principle, electrons among them.

Fermi Level: The value of the energy that separates occupied from empty states in the limit $T \rightarrow 0$ of Fermi–Dirac statistics.

Fermi Surface: The surface of constant energy in **k** space corresponding to the Fermi level.

Ferrimagnet: A substance in which the arrangement of local magnetic moments exhibits both a uniform value—a macroscopic moment—and a spatially oscillatory component.

Ferromagnet: A substance in which the microscopic magnetic moments are arranged all parallel to one another, with a consequent macroscopic moment.

Fluxoid: The quantum of magnetic flux in multiply connected superconductors (see flux quantization).

Flux Quantization: The property of multiply connected superconductors that restricts the values of the magnetic-field flux inside the sample to integral multiples of a fixed quantity, called the fluxoid.

Ginzburg–Landau Theory (Ginzburg–Landau Equations): A phenomenological quantum theory that describes the electromagnetic properties of superconductors and which allows for spatial variations and coherence effects.

Josephson Effect: Coherent tunneling of electrons between two superconductors separated by a thin nonsuperconducting—normally insulating—layer.

London Theory (London Equation): A macroscopic classical theory which describes the electromagnetic properties of superconductors.

Meissner Effect: The complete exclusion of magnetic flux (perfect diamagnetism) from the body of a superconductor at low enough values of the applied magnetic field.

Paramagnet: A substance that exhibits neither a permanent macroscopic magnetic moment nor long-range order of local moments, and which attracts magnetic flux.

Penetration Depth: The distance over which an outside magnetic field decays into the bulk of a superconductor.

Reciprocal Lattice: The discrete, infinite set of points in reciprocal space which defines the vectors used in the Fourier series for a given periodic structure in three dimensions.

Reciprocal Space (k Space): A mathematical space, with the dimension of inverse length, over which wave vectors and periodicities of Fourier series are defined.

List of Works Cited

Bardeen, J., Cooper, L. N., Schrieffer, J. R. (1957), *Phys. Rev.* **108**, 1175–1204.

Falicov, L. M. (1973), in: *Electrons in Crystalline Solids*, Vienna: International Atomic Energy Agency, pp. 207–279.

Faraday, M. (1846), *Philos. Trans. R. Soc. London* **2**, 2149.

Ginzburg, V. L., Landau, L. D. (1950), *Zh. Eksp. Teor. Fiz.* **20**, 1064–1068.

Gor'kov, L. P. (1959), *Zh. Eksp. Teor. Fiz.* **36**, 1918–1923.

Kittel, C. (1986), *Introduction to Solid State Physics*, 6th ed. New York: Wiley, Chaps. 9, 12, and 14.

Josephson, B. D. (1962), *Phys. Lett.* **1**, 251–253.

Kamerlingh Onnes, H. (1911), *Proc. K. Ned. Akad. Wet.* **14**, 113–114, 818–820.

Landau, L. D. (1930), *Z. Phys.* **64**, 629–637.

Langevin, P. (1905), *J. Phys. (Paris)* **4**, 678–693.

London, F., London, H. (1935), *Proc. R. Soc. London Ser. A* **149**, 71–88.

London, F. (1950), *Superfluids*, New York: Wiley.

Meissner, W., Ochsenfeld, R. (1933), *Naturwissenschaften* **21**, 787–788.

Onsager, L. (1952), *Philos. Mag.* **43**, 1006–1008.

Pippard, A. B. (1953), *Proc. R. Soc. London Ser. A* **216**, 547–568.

Somerville, M. F. G. (1848), *On the Connection of the Physical Sciences*, 7th ed. London: Buckley, Chap. xxiii, p. 369.

Further Reading

General

Kittel, C. (1986), *Introduction to Solid State Physics*, 6th ed. New York: Wiley, Chaps. 9, 12, and 14.

Van Vleck, J. H. (1932), *The Theory of Electric and Magnetic Susceptibilities*, London: Oxford University (reprinted in 1952).

White, R. M. (1970), *Quantum Theory of Magnetism*, New York: McGraw-Hill, Chap. 3.

Diamagnetism of Metals

Cracknell, A. P., Wong, K. C. (1973), *The Fermi Surface*, Oxford: Clarendon.

Falicov, L. M. (1973), in: *Electrons in Crystalline Solids*, Vienna: International Atomic Energy Agency, pp. 207–279.

Pippard, A. B. (1960), *Rep. Prog. Phys.* **23**, 176–266.

Pippard, A. B. (1969), in: Ziman, J. M. (Ed.), *The Physics of Metals. 1 Electrons*, Cambridge: Cambridge University Press, pp. 113–174.

Shoenberg, D. (1957), in: Gorter, C. J. (Ed.), *Progress in Low Temperature Physics*, Vol. 2, New York: Interscience, pp. 226–245.

Wilson, A. H. (1954), *The Theory of Metals*, 2nd ed. London: Cambridge University Press, pp. 150ff.

Ziman, J. M. (1970), *Electrons in Metals: A Short Guide to the Fermi Surface*, London: Taylor and Francis.

Diamagnetism of Superconductors

Chandrasekhar, B. S. (1969), in: Parks, R. D. (Ed.), *Superconductivity*, Vol. 1. New York: Dekker, pp. 1–49.

Tinkham, M. (1975), *Introduction to Superconductivity*, New York: McGraw-Hill.

Werthamer, N. R. (1969), in: Parks, R. D. (Ed.), *Superconductivity*, Vol. 1. New York: Dekker, pp. 321–370.

DIAMOND AND DIAMONDLIKE CARBON

MICHAEL W. GEIS, *Lincoln Laboratory, Lexington, Massachusetts, U.S.A. (Introduction and Secs. 1–3)*

MICHAEL A. TAMOR, *Research Staff, Ford Motor Company, Dearborn, Michigan, U.S.A. (Secs. 4–8)*

	Introduction	1
1.	**Diamond**	2
1.1	Natural Diamond	2
1.2	Synthetic Diamond	3
1.3	Diamond Typing	5
2.	**Diamond Properties**	5
2.1	Mechanical Properties	5
2.2	Surface Properties	6
2.3	Thermal Properties	6
2.4	Optical Properties	7
2.5	Electrical Properties	9
3.	**Use of Diamond**	10
4.	**Diamondlike Carbon—General Considerations**	11
4.1	History and Nomenclature	11
5.	**Hydrogenated DLC (*a*-C:H)**	12
5.1	Formation of *a*-C:H	12
5.2	Properties of *a*-C:H	13
5.2.1	Composition—Hydrogen Content	13
5.2.2	Optical Properties	13
5.2.3	Mechancial Properties	14
5.2.4	Chemical and Thermal Stability	14
5.2.5	Electronic Properties	14
5.3	Structure of *a*-C:H	14
5.3.1	Random Covalent Networks	15
5.4	The *a*-C:H Deposition Process	17
6.	**Modified *M*-C:H**	17
7.	**Hydrogen-Free DLC (*i*-C)**	18
7.1	Formation of *i*-C	18
7.2	Properties of *i*-C	19
7.2.1	Optical Properties	19
7.2.2	Mechanical and Chemical Properties	19
7.2.3	Electronic Properties	19
7.3	Structure of *i*-C	19
7.3.1	Structural Models	21
7.3.2	The *i*-C Deposition Process	21
8.	**Applications of Diamondlike Carbon**	21
	Acknowledgment	22
	Glossary	22
	Works Cited	23
	Further Reading	25

INTRODUCTION

This article discusses the properties of carbon in a range of phases from graphite to diamond, where carbon is generally classified as either diamond or diamondlike. Diamond is a specific cubic crystalline phase of nearly pure carbon. Diamondlike applies to a whole class of carbon phases having some of the properties of both graphite and diamond. The diamondlike phases, which are predominantly amorphous, often contain other elements, primarily hydrogen. All these phases are metastable and will transform into graphite at high temperatures.

Diamondlike material, unlike diamond, is not found naturally. It was first synthesized in the 1950s, and so its development is comparatively recent. Since diamondlike refers to a class of materials with a variety of properties, characterization and deposition methods are emphasized here.

Although diamond and diamondlike carbon are both hard inert materials, the crystalline form of diamond and the amorphous structure of diamondlike carbon give them inherently different properties. Diamond is difficult to deposit as a smooth film because its polycrystalline properties result in a rough-faceted surface. However, thick self-support-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ing diamond substrates can be obtained with a number of growth techniques or by cutting substrates from larger natural diamonds. Diamondlike carbon, on the other hand, is easily deposited as a smooth film. Under some conditions the surface of the diamondlike material is even smoother than the surface substrate on which it was deposited. However, the inherent stress of the diamondlike material makes it very difficult to deposit films thicker than 1 μm, and thick self-supporting substrates have not been made to date.

1. DIAMOND

In spite of diamond's metastability, it does occur naturally and has been used as jewelry since ancient times. Large diamonds, found in alluvial deposits since 800 B.C., were worn by rulers in the early days of India as a sign of rank and wealth. Most historically relevant diamonds, like the Hope and the Idol's Eye, originally came from India. Some of these diamonds have quite a bloody past.

Approximately 100 tons of diamond are produced annually, of which 80% are synthetic (Dismukes *et al.*, 1991; Bush *et al.*, 1991). Gem-quality diamonds are obtained by mining and represent ~5% of the total production. A 1-carat gem-quality diamond varies in price from $600 to $6000 depending upon its quality (Rio Grand, 1992). The price of diamond grit has been continually decreasing, being $12 to $15 a carat in 1969 and $1 to $2 a carat in 1987 (U.S. Department of the Interior, 1989). In industry, diamond is used primarily for cutting and grinding. However, it also has several niche markets. Diamond is used for knives, heat spreaders for temperature-sensitive electronics, and windows. Furthermore, with the recent development of low-pressure diamond growth techniques, new applications in electronics and optical coatings are becoming important.

After discussing both the natural sources and the synthetic production of diamond, this article emphasizes its unique material properties, some of which are reported in Table 1. Diamond's thermal conductivity, hardness, optical transmission, and electrical properties are considered in comparison to other materials. Finally, the present and possible future applications of diamond are discussed.

1.1 Natural Diamond

The primary natural source of diamonds is the core of ancient volcanoes. The material of these cores is called kimberlite, after the town of Kimberley, South Africa. The diamonds

Table 1. Physical properties of diamond.[a]

Property	Description
Lattice constant	0.356725 nm (silicon is 0.54307 nm)
Density	3.515 kg m^{-3}
Coefficient of expansion	8×10^{-7} (fused quartz is 2.5×10^{-7})
Thermal conductivity	1.5×10^{4} W m^{-1} K^{-1} (copper is 4.2×10^{3} W m^{-1} K^{-1})
Heat capacity	6.2 kJ kg^{-1} K^{-1} (copper is 4.2 kJ kg^{-1} K^{-1})
Thermal stability	Reacts with O_2 at >600 °C Reacts to form graphite in vacuum at >1500 °C
Young's modulus	1.054×10^{12} N m^{-2} (drawn tungsten is 3.5×10^{11} N m^{-2})
Dielectric constant	5.7 (silicon is 11.9)
Band gap	5.5 eV (silicon is 1.1 eV)
Resistivity	$>10^{18}$ Ω m, type IIa, 3.4×10^{-5} Ω m, type IIb[b] (copper is 2×10^{-8} Ω m)
Hole mass[c]	Heavy-hole band, $m^*=2.1\ m_0$ Light-hole band, $m^*=0.7\ m_0$ Split-off band, $m^*=1.06\ m_0$
Hole mobility	0.16 m^2 V^{-1} s^{-1} (silicon is 0.05 m^2 V^{-1} s^{-1})
Electron mobility	0.2 m^2 V^{-1} s^{-1} (silicon is 0.15 m^2 V^{-1} s^{-1})

---

[a] All properties are specified at 25 °C unless otherwise stated.

[b] Boron concentration is 2×10^{27} m^{-3}.

[c] m_0 is electron mass.

were not formed in the volcanoes but were brought up to the surface from regions nearer the earth's center by the volcanic action. The cause and character of these ancient volcanoes are unknown, and no such active volcanoes now exist. Many believe the kimberlite did not originate in the cores as hot lava, since the temperatures common to lava (1000–1200 °C) would be expected either to transform diamond into graphite or to burn it into carbon dioxide. Some diamonds, dated by radioactive isotopes, appear to have originated before the earth. While there is some controversy over the interpretation of the isotope dating, most agree that diamonds originate from material dating at least to the earth's formation (Dodson, 1989).

Natural diamonds are usually mined either from the volcano cores or from the sand of streambeds containing material from the cores. In India and Brazil diamonds are found almost exclusively in water-deposited sediments, while in South Africa and Siberia a larger percentage of diamonds is obtained from the volcano cores. Since less than one part diamond by volume is recovered for every 4×10^7 parts mined material, techniques have been developed to extract the diamond from the ore (Burton, 1978a). After the ore is crushed, its components are separated by density. Diamond, having a density of 3.6 $\mathrm{kg\,m^{-3}}$, is denser than most of the other materials. After this step of processing, about one part in 4×10^5 is diamond. The diamond component is then further concentrated using some of diamond's other unique properties. For example, diamonds are not wetted by water and will adhere to grease or wax. On passage of concentrated gravel in water over a moving belt coated with grease, the diamonds adhere to the grease while the other components are washed away. Since diamond is an insulator whereas most of the other components of the ore are conductive, it can also be separated electrostatically. Diamonds, unlike most of the other ore components, fluorescence when exposed to rays and can thus be further concentrated using x-ray fluorescence as an indicator. The last operation, separating diamond from the refined ore, is always accomplished by hand sorting.

1.2 Synthetic Diamond

Synthetic diamond is formed by two techniques, one using a high-temperature (> 1000 °C), high-pressure ($> 10^5$ atm) liquid and the other using a gas at lower temperatures (< 1000 °C) and low pressures (~ 0.1 atm). The high-pressure process takes advantage of the phase diagram of carbon, shown in Fig. 1. At 1 atm, graphite is the most stable form of carbon, but as the pressure is increased, the diamond phase becomes more stable. Since the transformation of graphite into diamond is an extremely slow process, high temperatures and a catalyst (nickel, iron, or cobalt) are usually required. The carbon–catalyst mixture is heated to liquefy the catalyst, at which time it dissolves the graphite and deposits the carbon as diamond. By use of this technique, diamonds several millimeters in diameter have been formed (Wentorf, 1965). At lower pressures, where graphite is the stable form, the same metals will dissolve diamond and deposit the carbon as graphite.

At sufficiently high pressures and temperatures, which can be obtained by chemical explosives, catalysts are not needed. Diamond formed commercially with the explosive technique is composed of very fine crystals and is used primarily for abrasive (Du Pont, 1965).

The second technique uses a nonequilibrium process to form diamond. The diamond is grown in a mixture of gases containing carbon, hydrogen, and oxygen (Zhu *et al.*, 1991). The surface on which the diamond is to be deposited is heated to between 600 and 1000 °C, causing the carbon-containing gas (usually CH_4) to decompose. The carbon deposits both as graphite and as diamond, but under most conditions the graphite is deposited much faster than the diamond, and after a few minutes only the graphite is evident. However, graphite is known to be etched by H and O atoms and OH radicals, while diamond is comparatively inert. By forming these atoms and radicals either in a plasma or on a very hot surface (> 2000 °C) near the region where the carbon is deposited, any deposited graphite will be etched away, while diamond will continue to be deposited. Figure 2 shows an example of a diamond growth system with a heated substrate for diamond deposition and a hot filament to form the graphite-etching chemical species. The reactive chemical species in some diamond growth systems

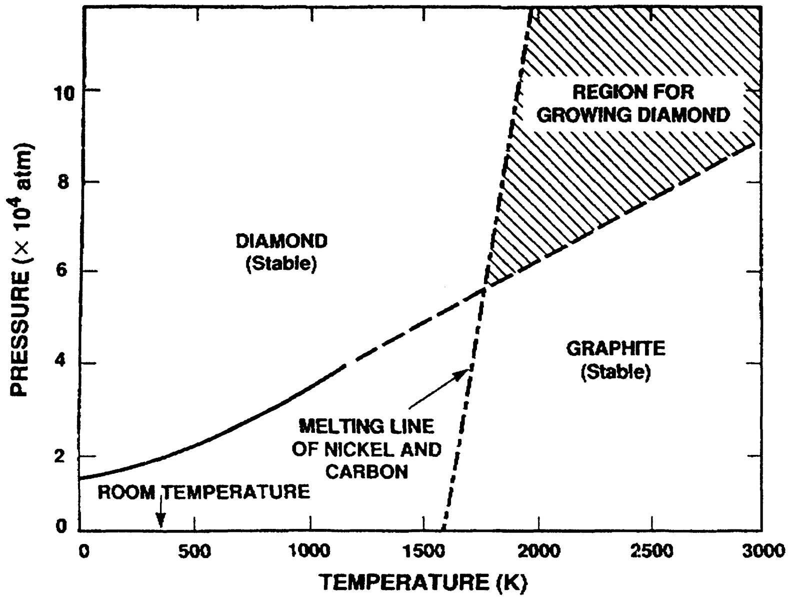

FIG. 1. Pressure–temperature phase diagram of diamond. At high pressures, diamond is the lower-energy phase of carbon. The requirement that a metal catalyst, like Ni, be in the liquid phase to transform graphite into diamond necessitates the growth of diamond above 1700 K.

are formed using a plasma. Some of the highest diamond growth rates (0.1 to 1 mm h^{-1}) are obtained employing chemical reactions between oxygen and acetylene to form the reactive species.

The low-pressure technique deposits diamond over a range of gas compositions. If the concentration of carbon in the gas phase is too small, then the diamond is etched. If the concentration of carbon is too large, then graphite is deposited. Only over a comparatively narrow range of elemental compositions is diamond deposited (Bachmann *et al.*, 1991). Figure 3 shows the experimentally determined atomic concentrations of oxygen, hydrogen, and carbon where diamond is formed. The chemical composition necessary for diamond deposition changes slowly with substrate temperature or gas pressure.

The two diamond growth techniques result in different forms of diamond. The low-pressure approach produces diamond films several micrometers thick and many centimeters across, while the high-pressure technique produces diamonds in the shape of blocks or spheres, usually not much larger than a millimeter across. The low-pressure technique is used to coat windows and cutting tools and

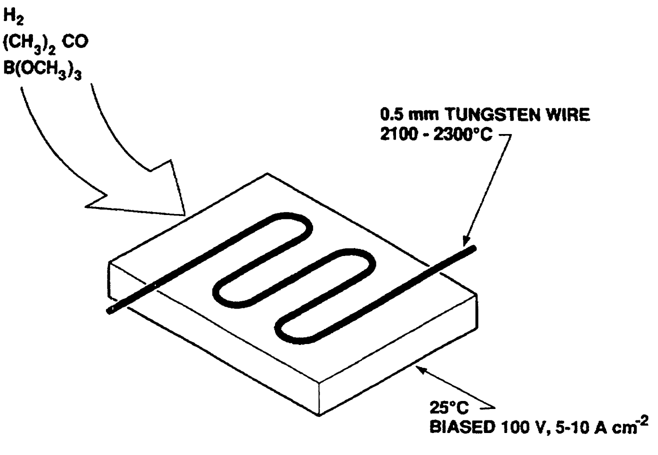

FIG. 2. Schematic diagram of a low-pressure diamond growth system. Although $(CH_3)_2CO$ (acetone) is usually used to grow diamond, nearly any carbon-containing gas will deposit diamond. Boron doping is accomplished with the addition of $B(OCH_3)$ or BH_3. The hot wire can be composed of tungsten (as shown here), tantalum, or rhenium.

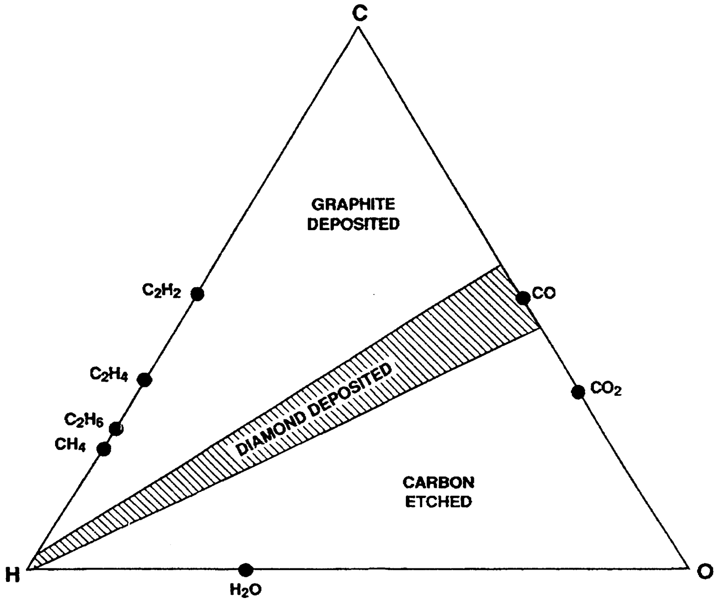

FIG. 3. Chemical atomic concentrations of hydrogen, oxygen, and carbon are shown that will etch diamond and graphite, deposit diamond, and deposit graphite.

has electrical applications. Diamonds produced with the high-pressure process are used primarily for cutting and grinding tools, where the more spherical shape is desired.

1.3 Diamond Typing

Synthetic and natural diamonds can be classified as to their impurity concentration and crystalline perfection. The two most common impurities are nitrogen and boron. Most other elements have extremely low solubility in diamond and therefore are not of general concern. Type I diamond contains high concentrations of nitrogen. This nitrogen either can take only substitutional sites in the carbon lattice or can precipitate as platelets, forming either type Ib or type Ia diamond, respectively. Type IIa diamond contains no measurable nitrogen, as determined by optical absorption. If in addition to containing little to no nitrogen the diamond also contains boron, then it is classified as type IIb. Type IIb diamond is a *p*-type semiconductor.

Natural diamonds are primarily type Ia, with some type IIa and very few type IIb. High-pressure-produced synthetic diamond is almost exclusively type Ib. If metals like Al or Ti that getter nitrogen from the carbon are used with the high-pressure growth technique, then type IIa and type IIb diamonds can be formed.

2. DIAMOND PROPERTIES

2.1 Mechanical Properties

Diamond has several mechanical properties that allow for its use in cutting and machining applications. The theoretical strength α of a brittle material is proportional to $(E\varphi/\lambda)^{1/2}$, where E is the modulus of elasticity, φ is the surface energy, and λ is the distance between atoms. In diamond, $E=1.05\times10^{12}$ N m^{-2}, five times that of steel and ten times that of aluminum alloy; $\varphi=5.3$ J m^{-2}, eight times that of Al; and $\lambda=0.154$ nm, almost half that of iron and Al (Orowan, 1949). Diamond, having among the highest moduli of elasticity and surface energies along with the smallest atomic spacing, is predicted to have a very high fracture strength in tension (1.9×10^{11} N m^{-2}), even along the easiest direction of fracture, normal to the axis with Miller indices $\langle 111\rangle$. The actual strength of diamond is difficult to measure because of the small sample sizes and the inconsistency of samples. However, as for most materials, the theoretical strength is orders of magnitude higher than the measured value under tension of $\sim3\times10^9$ N m^{-2} (Howes, 1965), which is slightly higher than the value for steel wire.

The hardness of many materials scales with their atomic density, and diamond with the highest atomic density is the hardest material at room temperature and pressure, as shown in Fig. 4 (Angus, 1991). Materials with either the diamond or zinc-blende crystal structure are usually harder than materials with the same atomic density but different crystal structures. The hardness also depends on the type and crystal orientation of diamond. Type I with nitrogen impurities is softer than type II. The reason for this is not known, since impurities like nitrogen will harden most other materials.

2.2 Surface Properties

The surface properties of diamond are the result of the almost unique surface chemistry of carbon. Most other materials react with the oxygen and water in the environment. Aluminum and iron, for example, form oxides or hydroxide films. Even very stable compounds, like Si_3N_4, react to form a surface coating of SiO_2. The surface of diamond, however, is extremely stable and unreactive with air, provided the carbon bonds at the surface are terminated with one of a variety of elements, such as hydrogen, fluorine, oxygen, or nitrogen. Just as polyethylene and Teflon are stable and unreactive at room temperature, so the chemical terminated surface of diamond is stable.

This unreactive surface is in part responsible for the extremely low coefficient of friction (0.05) and the low abrasion and wear of diamond. This coefficient, which is less than that for Teflon, actually increases with the addition of lubricating oils. However, since the unreactive property exists only as long as the carbon bonds are terminated, the coefficient can increase to >1 if the chemically bound elements are removed. For example, when diamond is heated to >800 °C in vacuum or when the surface is repeatedly rubbed in vacuum, the elements bound to the surface are removed and the coefficient of friction increases substantially. Often the friction becomes so high as to cause surface damage from abrasion and wear. Both the coefficient of friction, or the wear, and the surface chemistry are dependent upon the crystal orientation and the direction of motion. The surface with Miller indices (111) has the lowest coefficient, 0.05, while the (100) surface has a coefficient of 0.05 in the ⟨110⟩ direction and 0.1 to 0.15 in the ⟨100⟩ direction (Seal, 1953).

2.3 Thermal Properties

Diamond has three unique thermal properties: the highest volumetric heat capacity at high temperatures, one of the lowest heat capacities at lower temperatures, and the highest room-temperature thermal conductivity. At high temperatures near and above the Debye temperature (>1570 °C), the thermal energy of each mole of atoms is $\sim 3RT$, where R is the Boltzmann constant and T is the temperature. Diamond, with the highest atomic density, has the highest volumetric heat capacity. At temperatures <80 °C, well below the Debye temperature, most of the atoms are

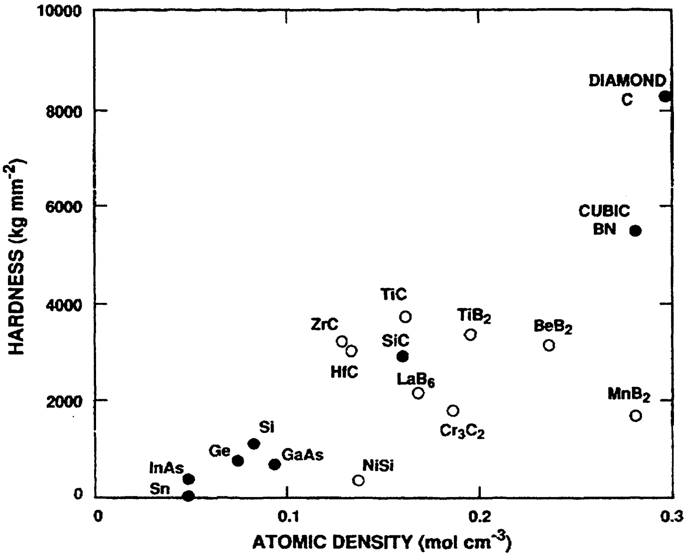

FIG. 4. Material hardness as a function of atomic density. The filled dots represent elements in the diamond crystal structure or compounds in the zinc-blende crystal structure.

in their ground vibrational state and the thermal energy of the lattice is very small. Diamond, having one of the highest Debye temperatures, has one of the lowest low-temperature heat capacities, as shown in Fig. 5.

Diamond's thermal conductivity, shown in Fig. 6, is the result of heat conduction by vibrational energy in the form of phonons. The thermal conductivity is limited by the density of phonons and the density of phonon-scattering centers. The phonon-scattering centers depend upon the purity and crystal quality of the diamond. Type Ia diamond with nitrogen precipitates has from one-fourth to one-tenth the thermal conductivity of the more pure form, type IIa. However, if the nitrogen is not precipitated and is still in substitutional sites, the case for type Ib diamond, then the thermal conductivity is nearly equal to that of type IIa diamond. Most commercial synthetic diamond used for thermal conduction is type Ib, while the natural diamond used for the same property is usually type IIa (Berman, 1979). Low-pressure–produced synthetic diamond, although it has high elemental purity, does not usually have the crystal quality of natural diamond. Therefore, the thermal conductivity of low-pressure synthetic diamond is inferior to that of type IIa natural diamond (Tzeng *et al.*, 1991).

Natural carbon consists of ~99% ^{12}C and ~1% ^{13}C. By making the diamond isotopically pure ^{12}C and by eliminating nitrogen, the density of phonon-scattering centers can be reduced. Such diamonds have ~50% higher room-temperature thermal conductivity than similar diamonds with natural isotopic concentrations. For type IIa diamond, at very low temperatures (<10 K) the mean free path between phonon-scattering events becomes large, and the effective thermal conductivity depends upon the size of the diamond. At these temperatures, diamonds with smooth surfaces have higher thermal conductivity than diamonds with rough surfaces because of specular phonon reflection at the surface (Berman, 1979).

2.4 Optical Properties

Diamond is unique in that it has almost no optical absorption from the extremely low frequencies, approaching dc, to light composed of photons with energies just below the band gap of 5.5 eV (i.e., light with wavelength >220 nm). With the exception of diamond, all of the commonly used window materials have substantial infrared absorption, so that no one material can be used to cover the entire spectrum from 0.2 to >1000 μm. Diamond does have a weak absorption near 5 μm, which is the absorption of a photon by the lattice to form two or more phonons (Edwards and Philipp, 1985). Figure 7 shows the transmission through 1-mm-thick windows of type IIa diamond, LiF, CsI, and fused quartz. LiF is commonly used as an ultraviolet window material, while CsI is a commonly used infrared window material. The ~70% transmission of light through diamond is the result

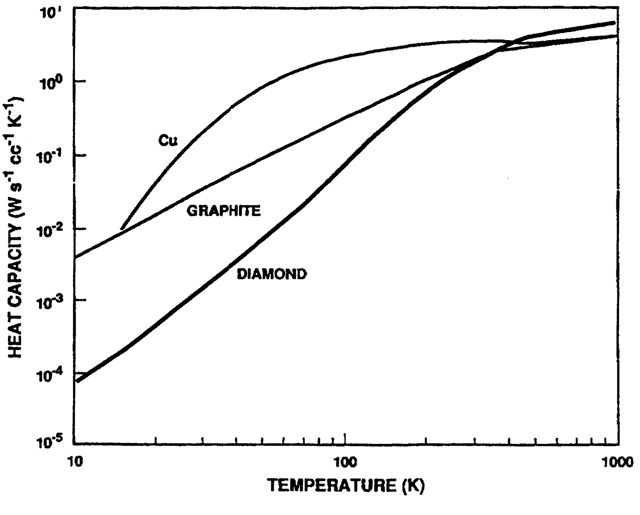

FIG. 5. Volumetric heat capacity as a function of temperature.

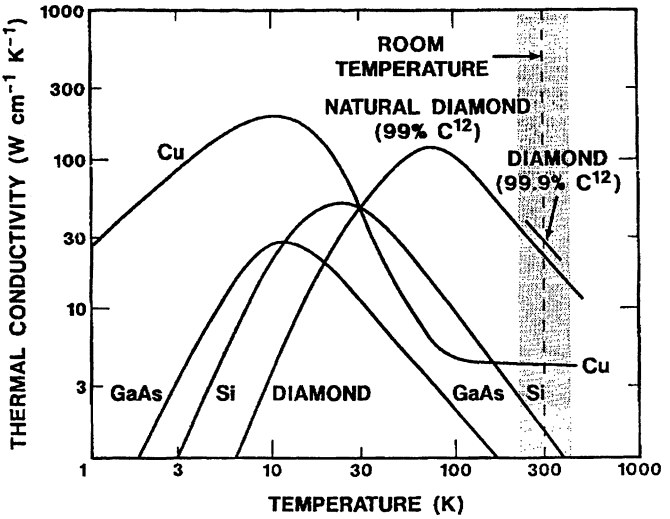

FIG. 6. Thermal conductivity of several materials as a function of temperature. The gray band indicates the temperature range where many semiconductors operate, −25 to 125 °C.

of optical reflection of the light as it passes in and out of the diamond window. Antireflection coating will increase the transmission over a limited spectral range.

Although diamond has the widest range of optical transmission, it has several limitations. Only the highest-purity diamond, type IIa, has this optical property. The other types, containing nitrogen or boron, have additional absorption in both the ultraviolet and infrared electromagnetic spectra. Type IIa diamond often contains internal strains that compromise its optical properties by randomizing the polarization of light passing through it. Types Ia and Ib are reported to have generally less internal strain than type IIa, which may give them superior optical quality over a limited spectral range.

In addition to these transmission properties, color centers can be formed in diamond through radiation damage with > 1-MeV electrons. These color centers are the result of defects that depend upon the impurities in the diamond and heat treatment after radiation. Carbon vacancies in the lattice are believed to be the primary optical defects formed in diamond immediately after radiation. If a radiation-damaged sample is heated above 450 °C, the vacancies diffuse in the crystal, forming new defects. For the case of type Ib diamond, this new defect is believed to be a vacancy associated with a substitutional nitrogen atom. Several other color-center defects have been found and characterized, most of which either are photoluminescent or have

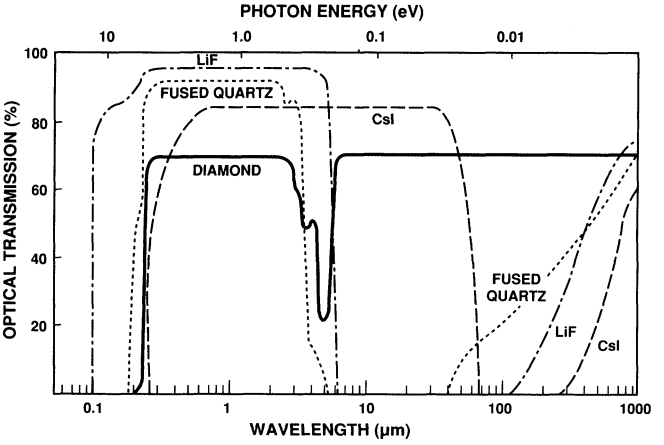

FIG. 7. Optical transmission of diamond, fused quartz, LiF, and CsI as a function of wavelength.

optical absorption in the spectral range from 0.3 to 1 μm.

Color centers produced by radiation damage from electrons or neutrons are often used to change the color of diamonds for cosmetic purposes (Burton, 1978b). Increasing radiation damage usually turns the stones pale green, bottle green, and opaque black. Subsequent heat treatment from 500 to 900 °C for several hours changes type Ia to yellow or amber, type IIa to brown, and type Ib to red or purple.

2.5 Electrical Properties

Diamond is a wide-band-gap (5.5 eV) semiconductor with several unique electrical properties. Two advantageous properties of diamond over the commonly used semiconductors are its high thermal conductivity, as discussed earlier, and its high breakdown electric field. The breakdown electric field of diamonds has not been accurately determined but is between 4×10^6 and 2×10^7 V cm^{-1} (Bazhenov *et al.*, 1985), which is comparable to the breakdown field of $\sim 10^7$ V cm^{-1} for SiO_2. For comparison, the breakdown electric field of silicon and GaAs is between 2×10^5 and 4×10^5 V cm^{-1}.

Semiconducting diamond can be doped either *p* or *n* type to conduct current either by holes in the valence band or by electrons in the conduction band, respectively. Boron is the only known *p*-type dopant for diamond. Other impurities like aluminum or indium, which are used to dope silicon *p*-type, are too large to fit in the diamond lattice and therefore have solubilities too low to be of value as dopants. Even boron is not a very efficient dopant, because it requires ~ 0.36 eV to ionize the boron atom, forming a hole in the valence band. This large ionization energy results in less than 1% of the boron atoms contributing holes to the valence band at room temperature. In addition, most diamond also contains other compensating impurities, like nitrogen, which trap holes, further reducing the hole concentration. If as little as 1% of the boron concentration is replaced with substitutional nitrogen, then the resistance of the diamond can increase by as much as an order of magnitude (Collins and Lightowlers, 1979). When the diamond is heated, more of the boron atoms become ionized, the detrimental effect of the compensating doping is reduced, and boron-doped diamond becomes more conductive. For electrical use, compensating dopants are undesirable, but it is difficult to obtain diamond with < 1% of the boron atoms compensated by impurities or crystal defects.

Phosphorus is the only known *n*-type dopant of diamond (Okano *et al.*, 1990). However, the phosphorus atom is too large to fit in the diamond lattice. This makes it difficult to obtain sufficient phosphorus concentrations for the diamond to be electrically interesting. Carbon-ion implantation has also produced *n*-type diamond (Prins, 1982). Since too much damage results in the formation and graphite and too little damage leaves the diamond insulating, it is difficult to control the doping by this method. Further, if the diamond is heated above 600 °C, much of the damage is annealed and the *n*-type character is degraded. Neither phosphorus doping nor radiation damage produces *n*-type diamond with the degree of resistivity control that can be obtained with silicon and GaAs.

In spite of the difficulty of producing *p*- and *n*-type diamond, both holes and electrons have high mobility in diamond of ~ 2000 cm^2 V^{-1} s^{-1}. This high mobility combined with the large band gap, high breakdown electric field, and high thermal conductivity of diamond make it a desirable semiconductor for high-power, high-frequency, and high-temperature transistors. Diamond transistors have demonstrated power gain at temperatures > 500 °C, and diodes have been operated at 700 °C.

The (111) surface of diamond has a negative electron affinity (NEA) (Himpsel *et al.*, 1979); that is, the energy level of the conduction band is above the minimum energy of electrons in vacuum, as shown in Fig. 8 (Geis *et al.*, 1991b). This NEA property is stable in the presence of oxygen and water. The property is very dependent upon the atoms chemically bound to the diamond surface. Hydrogen and oxygen are believed necessary to maintain the NEA surface of diamond. When the diamond is heated in vacuum to > 800 °C, most of the chemically bounded atoms desorb from the diamond surface and the NEA property is lost. It will be recovered when the diamond surface is again reacted with hydrogen and oxygen. By using the NEA property, cold cathodes have been demonstrated that

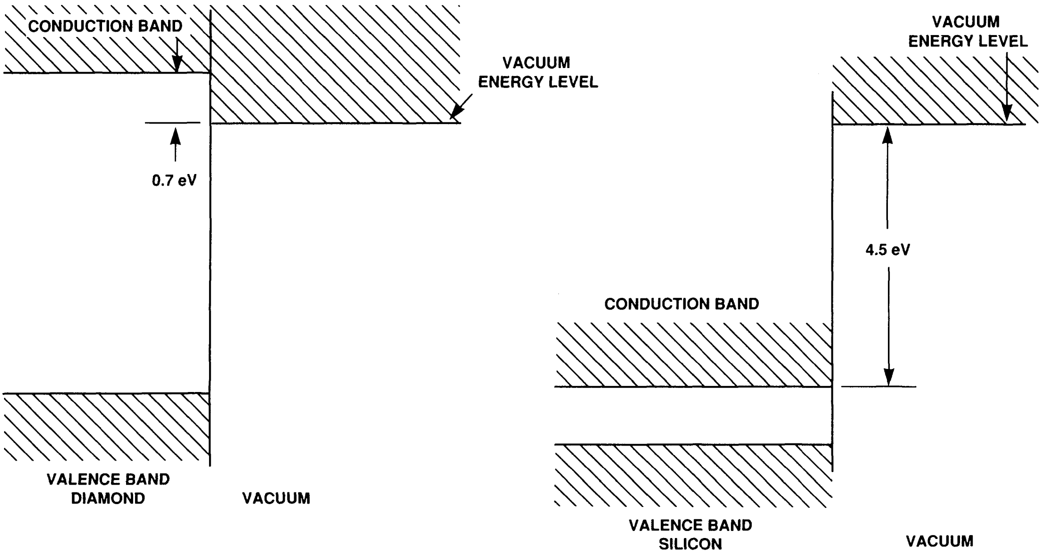

FIG. 8. Energy levels of the conduction and valence bands of diamond and silicon illustrated with reference to vacuum.

are not affected by either oxygen or water (Geis *et al.*, 1991a).

3. USE OF DIAMOND

Diamond, being a comparatively expensive material, generally finds use only when no other cheaper materials will perform as well (Bush, 1979). Diamond's unique hardness, strength, and high thermal conductivity have resulted in its use in cutting and grinding. In this application, diamond usually has an order of magnitude longer lifetime than its competitors, silicon and tungsten carbides. The higher grinding and cutting speeds, longer tool life, and reduced maintenance give diamond a cost advantage. However, ferrous metals like Fe, Ni, and Co, which are used to catalyze the transformation of graphite into diamond at high pressures, will catalyze the transformation of diamond back into graphite at atmospheric pressure. When diamond is used to cut one of these metals, the diamond is quickly eroded away unless it is actively cooled with cutting fluids. Because of this, other materials, like cubic boron nitride, tungsten carbide, and silicon carbide, are often used for machining these ferrous materials.

Diamond's wear resistance and low coefficient of friction have resulted in its use as a bearing in hostile environments, like oil drilling. Because of its ability to maintain a sharp cutting edge with minimal tissue damage, it has also found application in knives for eye and brain surgery. Diamonds are often used as wire drawing dies where high tolerances of wire diameter are desired (Caveney, 1979).

In addition, diamond is used to conduct heat away from temperature-sensitive devices that are required to dissipate high power densities. Semiconductor laser diodes, Gunn oscillators, and IMPATT diodes generate thermal power densities up to 10^6 W cm^{-2}. The performance of these devices degrades quickly with increasing temperature. For most geometries a decrease in thermal resistance of a factor of 2 can be obtained by mounting the device on a diamond substrate rather than a copper substrate.

Diamond is used for windows in spectrometers where a large transmission spectrum is desired. The wide transmission spectrum of diamond combined with its high modulus of elasticity allows diamond anvils to produce extremely high pressure, up to 500 kbar, while still being able to characterize optically the pressurized sample material between the anvils.

Diamond's unique properties could make further applications in lasers and electronics possible. Lasers have been fabricated using the color centers produced in diamond by radiation damage. The high thermal conductivity and heat capacity of diamond might result in high-power visible lasers. Diamond transistors are expected to outperform all other semiconductors in high-frequency high-power generation, if the technical barriers to obtaining large-area, inexpensive device quality diamond substrates can be overcome (Geis, 1991). The NEA property of diamond may result in the replacement of hot cathodes in vacuum electronic applications.

4. DIAMONDLIKE CARBON—GENERAL CONSIDERATIONS

Diamondlike carbon is not a single material, but a broad class of amorphous carbon-based materials that exhibit properties, most importantly high hardness, suggestive of diamond. Carbon can form both a stable threefold-coordinated (π-bonded) crystal structure, graphite, and a metastable fourfold-coordinated (singly bonded) crystalline structure, diamond. In combination with hydrogen, carbon may form polymeric structures containing single or double bonding. By controlled combination of diamond, graphite, and polymeric bonding in an intimately mixed amorphous structure, a wide range of material properties can be obtained. This range of properties can be extended by incorporation of a small fraction of a second element (e.g., Si, W, Ti). Unlike diamond, which can be produced in thin films by chemical vapor deposition (CVD) only at quite high temperature, films of diamondlike carbon can be produced at or even below room temperature. Although these are inferior to diamond in many respects, the ability to deposit a very smooth film of diamondlike carbon on virtually any solid substrate opens a much wider range of application.

4.1 History and Nomenclature

The term "diamondlike carbon" (DLC) was first used by Aisenberg and Chabot (1971) in reference to films deposited from a carbon ion beam. In contrast to other amorphous carbons, this new form of carbon was found to be very hard, electrically insulating, and wear resistant, i.e., diamondlike. Early in the development of techniques for chemical vapor deposition of crystalline diamond, actual diamond films were often conservatively called diamondlike pending confirmation of their crystalline nature. Once the ability to deposit diamond from hydrocarbon vapor was widely accepted, the usage became limited to hard materials composed largely or entirely of carbon but containing no long-range order, diamond or graphitic. Thus, in current usage the term "diamondlike" is descriptive of property only (hard, insulating etc.) and implies the absence of crystalline diamond order. By this definition all diamondlike carbon is amorphous, but not all amorphous carbon is diamondlike. Amorphous carbons (*a*-C) formed by pyrolysis of hydrocarbon resin (glassy carbon), and by sputtering or evaporation of graphite, contain considerable graphite order, are conductive, and are much softer than DLC.

Diamondlike carbon may be produced by a variety of methods and exhibits a wide range of properties. A complex nomenclature has evolved with terms based on the means of preparation (e.g., *i*-C or MSIB *a*-C for mass-selected ion-beam-deposited amorphous carbon) or some inferred structural characteristic of the film (e.g., *ta*-C, meaning tetrahedral amorphous carbon, or even the self-contradictory term "amorphous diamond"). The various forms of DLC may be divided into two broad categories: those formed of carbon only (*i*-C), and those consisting of both hydrogen and carbon, *a*-C:H. A modified form of *a*-C:H contains a small fraction of a third element, *M*, usually a metal, and is designated *M*-C:H. In all cases the film properties are almost entirely controlled by the kinetic energy with which (hydro)carbon species impact the growing surface. *a*-C:H can range from polymeric to graphitic; only *a*-C:H formed near the optimal energy is referred to as hydrogenated DLC. The label *i*-C distinguishes hydrogen-free amorphous carbon, formed from ions or atoms near the optimal energy, from softer, more conductive films formed at lower energy (*a*-C). Table 2 compares several key physical characteristics of the various classes of amorphous and crystalline carbon materials.

A unifying concept for all forms of amorphous carbon is the "hydrocarbon phase diagram" first proposed by Angus and Jansen (1988) (Fig. 9). This diagram maps the degree of fourfold or sp^3 carbon coordination as a function of hydrogen content. The solid line is derived from the theory of random covalent

Table 2. Key physical characteristics of several types of carbon materials.

	Density (g/cm^3)	Optical gap (eV)	Hardness (GPa)	sp^3 (%)	H(%)
Graphite	2.27	0.0	...	0	0
Glassy carbon	1.3–1.6	0.0	2–3	~0	~0
Evaporated C	1.9–2.0	0.4–0.7	2–5	<5	<5
Sputtered C (on Si @ 77 K)	2.2–2.6	0.6–0.9	10	5–10	<5
i-C	2.9–3.4	2.5	100	>85	<5
Hard *a*-C:H	1.6–2.0	1.2–1.6	10–30	40–50	30–50
Soft *a*-C:H	0.9–1.2	1.6–4.0	<5	50–80	50–60
Polyethylene	0.9	6.0	0.01	100	66
Diamond	3.52	5.5	100	100	0

networks (RCN), which is outlined later. Along this curve lie materials with barely sufficient constraint (defined by the degree of carbon–carbon bonding) to be rigid. Materials immediately to the right of the curve may be polymers, while those far to the right must be fluid. Those to the left are rigid solids. Graphite and amorphous carbons with threefold (sp^2) coordination lie at the origin, while diamond and "amorphous diamond" (analogous to amorphous silicon) lie at the top left. Measured sp^3 fractions for hydrogen-free amorphous carbons range from as low 10% to nearly 100%. The dashed curve represents a modification of the random networks that admits large graphitic clusters and is in good agreement with measured sp^3 fractions and compositions of *a*-C:H films.

5. HYDROGENATED DLC (*a*-C:H)

Hard hydrogenated carbon films deposited from a hydrocarbon plasma were first reported by Holland and Ojha (1976). Of all types of diamondlike carbon, *a*-C:H is the easiest and probably cheapest to produce. Properties such as hardness, optical gap, hydrogen content, and density correlate in nearly exactly the same manner for a variety of deposition techniques. This consistency has promoted a good understanding of the underlying atomic structure and its connection to the macroscopic properties of hydrogenated diamondlike carbon.

5.1 Formation of *a*-C:H

Hydrogenated DLC (*a*-C:H) films may be produced by virtually any means that can direct a flux of hydrocarbon ions with kinetic energy in the range 100–400 eV onto a substrate. The substrate temperature may be varied from slightly above to well below room temperature with only a small effect on film properties. Holland and Ojha first reported the CVD of *a*-C:H from hydrocarbon plasma in a parallel-plate reactor in which the substrate was fixed to an electrode to which rf

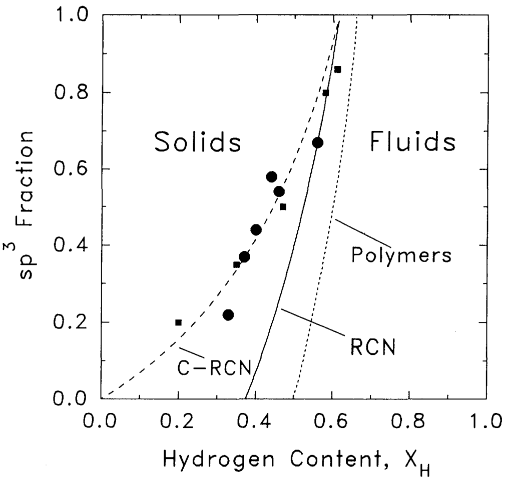

FIG. 9. The hydrocarbon "phase diagram" proposed by Angus and Jansen (1988). The solid curve represents optimally constrained random covalent networks without clustering and the dashed curve is for random networks with large graphitic clusters. The dotted curve represents simple long-chain polymers ranging from polyacetylene to polyethylene. Experimental data are obtained by NMR spectroscopy by Angus and Jansen (squares) and Tamor *et al.* (1991) (circles).

power was applied through a series capacitor. In this configuration, the powered electrode acquires a negative bias determined by the reactor geometry, the applied power, and the type and pressure of the gas. Because virtually any sputtering reactor can be adapted for *a*-C:H deposition, this self-biased CVD technique is widely used. Hydrocarbon plasma may be generated by dc, rf, or microwave discharge, and the substrate may be biased independently, or an accelerated ion source, such as the Kauffmann type, may be used.

5.2 Properties of *a*-C:H

The properties of *a*-C:H films are largely determined by the energy with which hydrocarbon ions arrive at the growing surface. With increasing ion energy *a*-C:H evolves from transparent, hydrogen-rich, polymerlike material to a brittle, nearly opaque, conducting material resembling carbon glass. The most diamondlike properties, maximum hardness and density, are obtained at intermediate energy. Several key properties of *a*-C:H deposited by rf plasma CVD from several precursors are compared as functions of rms self-bias voltage, V_b, in Fig. 10. The average ion energy measured in such systems is typically one-third to one-half the rms bias voltage.

Gas pressure, choice of precursor, and substrate temperature play secondary roles in determining film properties. With increasing gas pressure, collisions with the background of neutral species reduce the effective bombardment energy. At very low energy, precursors may not decompose entirely and some may incorporate nearly intact in the film. Films deposited from benzene may contain pendant benzene groups, while those formed from butane may be highly polymeric. At higher energy densities, film properties are similar for all precursors, but heavier hydrocarbons generally provide proportionately higher deposition rates. Rates up to several microns per hour can be achieved with butane or benzene vapor. The effect of increased substrate temperature during deposition is approximately the same as that of higher ion energy.

5.2.1 Composition—Hydrogen Content. The atomic fraction of hydrogen, X_H, incorporated into *a*-C:H is a key indicator of other film properties. Hydrogen content is controlled mainly by the ion bombardment energy, but is somewhat dependent on precursor. At very low bias X_H can be as high as 0.6, approaching the upper limit of 0.66 permitted for a polymer. The limit on X_H is influenced by the proportion of hydrogen in the source gas; a solid with $X_H > 0.5$ ($H/C > 1$) cannot be formed from benzene or acetylene ($H/C = 1$). For bias above roughly 200 V, where X_H is always less than 0.5, the choice of hydrocarbon has much less influence on film properties. Hydrogen content decreases with increasing bias to minimum of approximately 0.2 for bias of 1000 V.

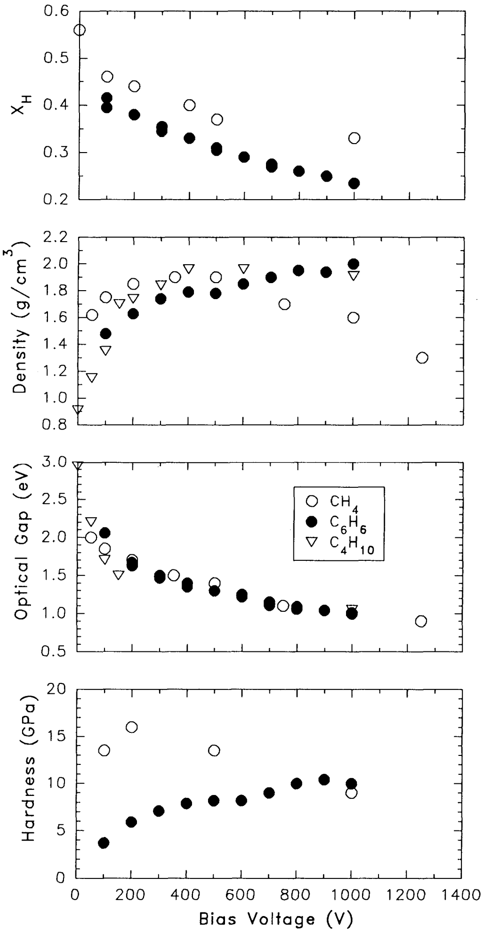

FIG. 10. Key properties of *a*-C:H deposited from hydrocarbon plasma as functions of bias voltage, V_b. Films deposited from three hydrocarbon precursors are compared: methane (○), benzene (●), and butane (▽).

5.2.2 Optical Properties. A critical parameter of a prospective optical coating material is its optical gap, E_g, which is determined from the photon energy ($h\nu$) dependence of the optical absorption (α) near the onset of absorption as

$$(\alpha h\nu)^{1/2}=h\nu-E_g. \qquad (1)$$

This form for the optical absorption edge is typical of amorphous materials where disorder has relieved the requirement of momentum conservation in the absorption process. A gap of 3 eV is required for optical transparency. Figure 10 shows that hard *a*-C:H does not meet this requirement; thin films (less than roughly 0.2 μm) are red-brown in transmission, while thicker films are opaque. Fortunately, this connection between ion energy and optical gap is not rigid. Controlled bombardment by high-energy noble-gas ions (e.g., 1000-eV Ar ions) during growth can increase the optical gap to meet this transparency requirement without sacrifice of other key film properties (Aisenberg and Kimock, 1988). *a*-C:H is quite transparent in the near-infrared range with the exception of a strong absorption band near 3.4 μm due to C-H bond-stretching vibrational modes. Thus *a*-C:H films that are quite opaque in the visible range are well suited for protection of optics for the 4–6-μm atmospheric infrared transmission band.

5.2.3 Mechanical Properties. The density of *a*-C:H correlates well with the hydrocarbon ion energy. Densities as low as 0.9 g/cm^3 for films formed at low energy are consistent with the other polymerlike characteristics (softness, transparency). Such films are similar to another form of hydrocarbon film, called plasma-polymer, which form when little or no potential appears between the plasma and substrate. The film density reaches a maximum of 2.0 g/cm^3 for bias in the 200- to 400-V range, at which the hardness can be as high as 30 GPa, comparable to the hardest ceramics, but far lower than that of diamond (105 GPa). Far less hard, dense, and transparent than diamond, hydrogenated DLC is not very diamondlike in comparison to actual diamond but is certainly the hardest of hydrocarbon materials. With further increase in bombardment energy, the density decreases to only 1.7 g/cm^3, typical of glassy carbon.

Hydrogenated DLC is very slippery and wear resistant even in dry sliding. For steel sliding on *a*-C:H in humid air (55%) coefficients of friction (COF) as low as 0.1 are readily obtained, while much lower values are occasionally observed. For comparison, the friction coefficient for steel on Teflon is approximately 0.05. The COF of *a*-C:H is sensitive to humidity and is reduced considerably in a very dry atmosphere or vacuum (Dimigen *et al.*, 1987). Although quantitative comparison of various measurements is very difficult, wear rates are generally found to be very low. For similar sliding conditions, wear of DLC is always much less than that of other very hard materials such as SiC, Si_3N_4, and TiN.

5.2.4 Chemical and Thermal Stability. Hydrogenated DLC films are extremely resistant to chemical attack by both strong acids and strong bases. However, they are stable only to temperatures between 300 and 400 °C. Although not as hard as films deposited at room temperature, films deposited at elevated temperature retain that hardness and their low friction characteristics to higher temperature (Grill, 1991). Hydrogen devolves from *a*-C:H at higher temperature as the material converts to graphite in several steps (Dischler *et al.*, 1983). After annealing to 600 °C, almost all hydrogen is removed and DLC films become conducting, opaque, and much softer, similar to carbon glass. Turbostatic graphite is formed by annealing to temperatures near 1000 °C.

5.2.5 Electronic Properties. Electronically, *a*-C:H would be best described as a composite of very small conducting islands captured in an insulating matrix. Studies of the temperature dependence of the conductivity indicate that conduction is via thermally activated hopping of charge carriers between localized states. This localization is presumed to occur on graphitic clusters (see later). Addition of dopants including B, P, and various metals increases the conductivity of *a*-C:H. However, this increase is usually accompanied by a decrease in optical gap. Thus it is not certain whether *n*- or *p*-type doping is actually achieved.

5.3 Structure of *a*-C:H

Raman spectroscopy, a probe of the optically active vibrational modes of a material, is one of the most convenient and powerful tools for characterization of both diamond and DLC. Raman spectra of *a*-C:H produced at low and high voltage are compared to the spectrum of a microcrystalline graphite powder in Fig. 11. It is readily apparent that the major features of the *a*-C:H Raman spectrum are derived from those of graphite and that this similarity increases with bias voltage. This indicates that graphitic carbon is present in form of small clusters. (These clusters are two dimensional, with no layer structure as in crystalline graphite.) The correlation between decreasing optical gap and increasingly graphitelike Raman spectrum suggests that the visible optical properties of *a*-C:H are determined by the graphitic component. Robertson and O'Reilly (1987) have made this correlation explicit, showing that the optical gap of a graphitic island or cluster varies as

$$E_g(N) = 6.0/N^{1/2} \text{ eV}, \tag{2}$$

where N is the number of rings in the cluster. The decrease in optical gap corresponds to an increase in cluster size from only a few rings at low voltage to over 100 at high voltage.

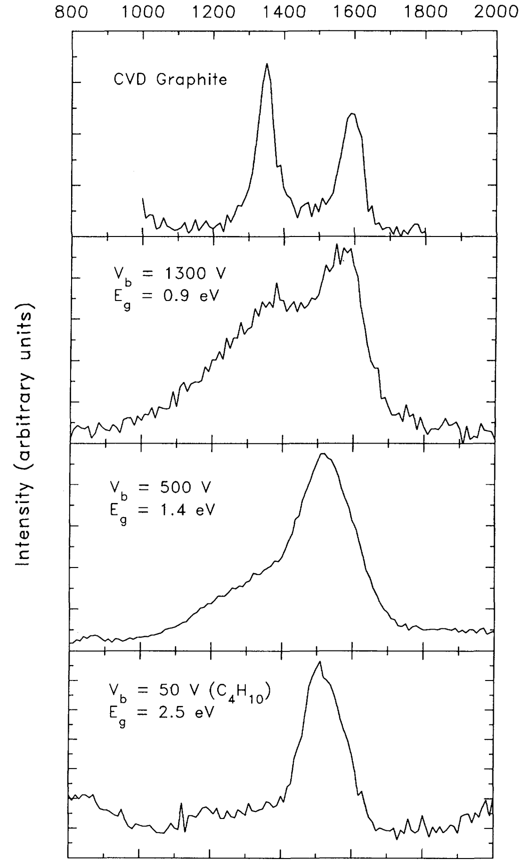

FIG. 11. Raman spectra of a series of *a*-C:H films compared to the spectrum of microcrystalline graphite.

While Raman spectroscopy does suggest the presence of small graphitic structures, it is much less sensitive to the more transparent polymeric or "amorphous diamond" connective structures that determine the bulk properties of DLC. Thus Raman spectroscopy cannot confirm or disprove the presence of a tetrahedrally coordinated component in either *a*-C:H or *i*-C. In contrast, nuclear magnetic resonance (NMR) spectroscopy is a local probe of carbon coordination equally sensitive to all carbon atoms. The strong correlation of increased fourfold carbon coordination with hydrogen content is shown in Fig. 9. A more detailed breakdown of carbon coordinations can permit a qualitative and even semiquantitative connection between local bonding and bulk properties. NMR spectroscopy has been used to sort carbon bonding environments in *a*-C:H into four fractions: threefold with protonation (aromatic or olefinic), threefold without protonation (graphitic), fourfold with protonation (aliphatic or methine), and fourfold without protonation (quaternary) (Tamor *et al.*, 1991). Figure 12 shows the four carbon bonding fractions as functions of bias voltage for a set of films deposited from CH_4. (The fractions of each are labeled f_3^H, f_3^C, f_4^H, f_4^C, respectively, and are normalized such that $f_3^H + f_3^C + f_4^H + f_4^C = 1$.) The overall trend from polymerlike material at low bias to more graphitic at high bias is clear. The fraction of quarternary carbon is maximized at approximately 25% in an intermediate voltage range (200–400 V) and correlates well with film hardness, density, and compressive stress.

5.3.1 Random Covalent Networks.

The connection between local coordination and macroscopic properties is made through the concept of random covalent networks (RCN). Two types of RCN treatments have been applied to *a*-C:H:

1. the theory of optimally constrained networks described by Thorpe (1983), which defines the minimal degree of network connectivity required to achieve macroscopic rigidity, and

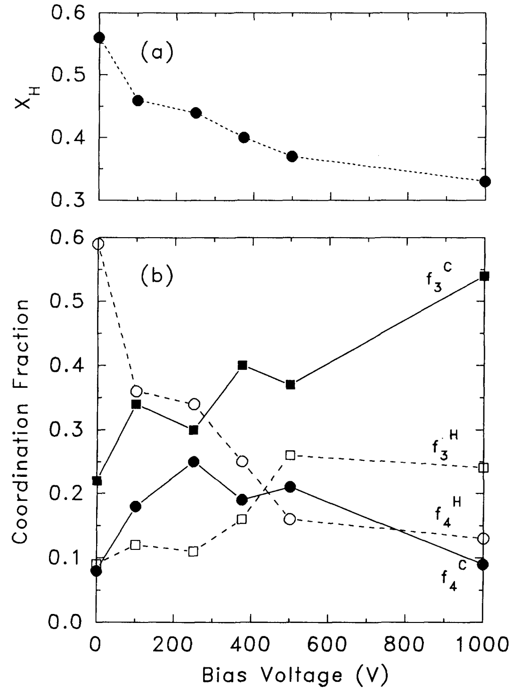

FIG. 12. (a) Hydrogen content of *a*-C:H films estimated from the coordination fractions determined by NMR spectroscopy. (b) The four carbon coordination fractions as functions of bias voltage. The solid symbols are for unprotonated carbon with threefold (f_3^C) and fourfold (f_4^C) coordination as in graphite and diamond, respectively. The open symbols are for protonated carbon with threefold (f_3^H) and fourfold (f_4^H) coordination.

2. a treatment of overconstrained networks in which the degree of overconstraint is connected to macroscopic elastic properties by a simple empirical relationship (He and Thorpe, 1985).

The concept of an optimally constrained network is based on the fundamental statement that a network cannot be rigid unless the number of translational oscillators of nonzero frequency is equal to the number of degrees of freedom in the network, $3N-3$, where N is the (very large) number of nodes in the network. In the absence of clustering, each atom is a node. A straightforward derivation shows that this condition is met when the mean atomic coordination, N_{av} (the average number of bonds to other nodes which are in turn part of the network, i.e., not dangling atoms or clusters), is exactly 2.4. For a random network consisting of carbon and hydrogen only, the overall degree of fourfold carbon coordination must vary with hydrogen content as

$$f_4=f_4^C+f_4^H=(8X_H-3)/5(1-X_H). \qquad (3)$$

This relationship defines the solid curve labeled RCN in the hydrocarbon "phase diagram" shown in Fig. 9. The degree to which overconstraint translates into hardness is determined by the amount of clustering in the network; N_{av} may be inflated by locally overconstrained clusters that cannot contribute to the macroscopic elastic properties (hard peanuts do not make peanut butter any harder). The graphitic clustering inferred from Raman and optical measurements can be accommodated by grouping all threefold-coordinated carbon into larger multiatom nodes and renormalizing to the newly reduced number of nodes. The curve for optimal constraint with large graphitic clusters is given by

$$f_4=5X_H/8(1-X_H), \qquad (4)$$

and is shown on the "phase diagram" as the dashed line labeled C-RCN. The modified form of the RCN model coincides well with the NMR measurements. This indicates that *a*-C:H consists of graphitic clusters, which define most optical properties, locked in a slightly overconstrained network that defines its mechanical properties.

Although formally rigid, optimally constrained networks cannot be macroscopically hard. It must be confirmed that the small degree of overconstraint is sufficient to account for the observed hardness of DLC. He and Thorpe (1985) treated an overconstrained RCN in a simulation in which bonds were randomly deleted from a diamond lattice and the resulting elastic properties calculated as a function of the average constraint in the remaining network. They found an empirical relationship between N_{av} and elastic modulus,

$$E=E_0[(N_{av}-2.4)/(N_0-2.4)]^{1.5}, \qquad (5)$$

where $N_0=4$ and E_0 is the modulus of the original lattice. Because hardness is approximately linear with modulus, this can be used to estimate hardness, H, by replacing E_0 with H_0 for diamond (105 GPa). The hardness of *a*-C:H using N_{av} from the NMR breakdown of carbon coordinations (assuming that graphitic carbon contributes little to hardness) is compared to the measured hardness in Fig. 13. The quantitative agreement confirms that

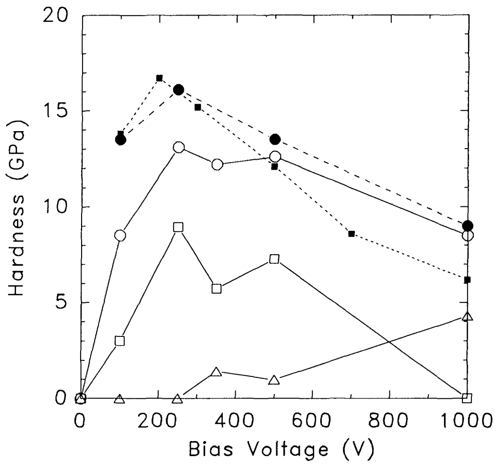

FIG. 13. Measured hardness of *a*-C:H as a function of bias voltage (● and ■) compared to estimates based on NMR measurements of carbon coordination using three assumptions: (1) that all atoms contribute to hardness (○), (2) that graphitically bonded carbon is clustered and so does not contribute to hardness (□), and (3) that quaternary carbon is clustered and so does not contribute (△). These results suggest that while graphitically bonded carbon may contribute somewhat to hardness, dispersed quaternary carbon is essential to sustain the observed hardness of *a*-C:H.

the carbon–carbon bond is so strong that even a small degree of overconstraint is sufficient to account for the hardness of hydrogenated diamondlike carbon. The contribution of quaternary carbon to network hardness is so critical that the opposite assumption of diamond clusters, amorphous or crystalline, implies zero hardness.

5.4 The *a*-C:H Deposition Process

The formation of *a*-C:H involves a complex interaction of chemical and kinetic processes. The mix of ions, neutral species, and radicals in the plasma is dependent upon the source gas, its pressure, and the power applied to the plasma (Koidl *et al.*, 1989). The substrate is bombarded by both energetic ions and diffusing neutral species. Under these conditions, film growth can proceed through a combination of surface and subsurface processes.

On the surface, the bombardment energy promotes formation of a strongly bonded carbon network by driving off hydrogen and loosely bonded carbon. Figure 9 suggests that it may be extremely difficult to produce *a*-C:H that deviates significantly from the C-RCN phase boundary. The strong tendency to hydrogenate is confirmed by studies which show that the hydrogen content of sputtered carbon films is always the same, approximately 35%, unless care is taken to remove residual hydrogen (mainly water vapor) from the reactor (Bhushan *et al.*, 1992). Simultaneous bombardment with a noble gas has been used in attempts to optimize preferential sputtering such that only diamond, the most strongly bonded form of carbon, is accumulated. Simultaneous bombardment has been found to increase the density and transparency of *a*-C:H, probably by reducing the size of the graphitic clusters. Scattered crystals tentatively identified as diamond are sometimes found embedded in the *a*-C:H matrix (Weissmantel, 1982).

Hydrocarbon ions with kinetic energy in excess of 50 eV penetrate several atomic layers into the film. Several calculations have shown that for ion-beam deposition in the absence of hydrogen (*i*-C deposition), this penetration can be tuned to favor a substantial fraction of quaternary bonding (Lifshitz *et al.*, 1989; Kaukonen and Niemenen, 1992). A similar effect may also occur in *a*-C:H, but is impeded by the hydrogen, which ties up dangling bonds, thus impeding formation of a dense quaternary network. The C–H bond is so strong that hydrogen cannot be driven from the film without such violent ion bombardment that the carbon network reverts to graphite. Thus the inability to produce highly overconstrained and harder DLC from hydrocarbon vapor is a consequence of this compromise in ion energy between what would be optimal in the absence of hydrogen (30–80 eV), and what is required to remove a large fraction of the hydrogen.

6. MODIFIED *M*-C:H

Although *a*-C:H is very hard, slippery, and easy to apply to a variety of substrates, it suffers two drawbacks: first, the high compressive stress limits the film thickness to roughly 2 μm, and second, the coefficient of friction is quite high under conditions of high humidity. Dimigen and Hübsch (1983/1984) addressed these problems by incorporating a variety of metals (W, Ta, Ru, Fe, and Ti) into *a*-C:H films by simultaneous sputtering. They found that a inclusion of few percent of metal reduced stress and improved adhesion so that much thicker films could be deposited. The friction coefficients of metal-containing *M*-C:H films were much less dependent on humidity. Iron-containing films exhibited reduced friction (0.04) while Ru-containing films exhibited increased friction (0.7). Incorporation of tungsten was found to increase wear resistance. Other workers have incorporated up to 30% Si into *a*-C:H films (not counting H) and found the friction reduced to 0.05, again independent of humidity, with wear rate comparable to hard *a*-C:H (Oguri and Arai, 1990).

Any element that forms a strong bond to carbon is suitable for incorporation into *M*-C:H. At small *M* content the films remain amorphous, and the additive can be considered a part of the covalent network structure of *a*-C:H. However, as the *M*:C ratio approaches 1, small carbide crystals appear within the amorphous matrix. Annealing also promotes the formation of carbide crystals. Although, strictly speaking, the resulting composite materials do not fit the definition of diamondlike carbon, they have great potential for smooth, slippery, and wear-resistance coatings.

7. HYDROGEN-FREE DLC (*i*-C)

Hydrogen-free DLC represents a subset of the possible carbon structures quite different from those found in *a*-C:H. In the "phase diagram" of Fig. 9, *i*-C lies high on the left axis. As it contains no hydrogen, *i*-C cannot include a polymeric component and must be composed of some combination of networks of sp^3 and sp^2 carbon and possibly graphitic clusters. The first and second dominate the mechanical properties while the last determines the optical properties. The degree of fourfold coordination is controlled by the energy with which carbon atoms or ions arrive at the substrate.

7.1 Formation of *i*-C

The most diamondlike of hydrogen-free DLC is deposited from some variation of a mass-selected or filtered carbon ion beam. Hardness and density are maximized for ion energies in the range 30–80 eV (McKenzie *et al.*, 1991). Just as for *a*-C:H, the substrate may be at or below room temperature. In fact, the substrate must be actively cooled during high-rate deposition. The major challenge in production of *i*-C lies in finding a source of carbon prolific enough to make possible the coating of large areas at reasonable rates. Conventional C-ion sources depend upon ion sputtering of graphite or decomposition of CO gas and ion extraction followed by mass filtering to produce a pure carbon-ion beam. An alternative approach that embodies the essential elements of an ion-beam system employs a cathodic arc discharge on a graphite electrode as an ion source and uses an external bias of the substrate relative to the electrode for the accelerating potential. Magnetic mass filtering must be added to remove the multiatom clusters produced in high-rate erosion of graphite.

Thermal evaporation and sputtering of carbon cannot be used to form *i*-C because the atomic kinetic energy is too low. Under marginal conditions, the ability of the substrate to dissipate the kinetic energy of incoming carbon atoms, through low initial temperature and high thermal conductivity, appears to play an important role in determining the film properties. Carbon films sputtered onto diamond at 77 K exhibit densities as high as 2.8 g/cm^3 and up to 60% sp^3 coordination (Cuomo 1991). Pulsed laser evaporation can impart energies of a few eV and so is an intermediate case between sputtering and ion-beam deposition. Carbon films laser-deposited on low-temperature substrates (77 K) show a systematic increase in density and estimated sp^3 content with increasing thermal conductivity of the substrate. At low laser fluence (of order 10^8 W/cm^2 in a 10-ns pulse), the carbon kinetic energy increases as the laser wavelength decreases, and the sp^3 content and density of the resulting carbon films are found to increase accordingly. For uv-laser (248 nm) evaporation, the ion kinetic energy can exceed 20 eV, approaching the optimal value determined for ion-beam deposition. Indeed, carbon films uv-laser–deposited on high-quality diamond substrates at room temperature have a density 3.2 g/cm^3 and an estimated sp^3 fraction exceeding 85% (Pappas, 1992).

Carbon films produced by laser ablation of carbon at very high laser fluence (10^{12} W/cm^2) appear to be quite different from those formed by evaporation at lower fluence. Above a critical energy density, an extremely hot carbon vapor plume rich in small clusters (C_3,

C_4, etc.) is produced (Collins *et al.*, 1991). Kinetic energies up to several keV are estimated for these small carbon clusters. This carbon stream may be similar to that produced by the cathodic arc. Films deposited from the most energetic portion of the plume are highly insulating and exhibit both density (up to 3.1 g/cm^3) and hardness (typically 40 GPa, but up to 77 GPa) approaching that of diamond, all suggestive of a very high sp^3 content. Microscopic examination of these films shows a nodular structure. The developers of this material have named it "amorphic diamond" and conclude that it is composed of small (100 nm) nodules of quaternary carbon (partially or entirely amorphous) captured in a more graphitic, but nevertheless very strong, matrix. This composite structure permits very high hardness to coexist with relatively low compressive stress. Such a material embodies the mechanical properties expected of "amorphous diamond" but cannot exhibit the corresponding optical properties because of the binding matrix.

7.2 Properties of *i*-C

7.2.1 Optical Properties. In analogy to amorphous silicon, the optical gap of an entirely sp^3 amorphous carbon should be comparable to that of crystalline diamond (5.4 eV). The largest optical gap for *i*-C reported to date is only 2.5 eV. This indicates that even the densest, most diamondlike of carbon films retain some residual nonquaternary carbon. Optical gaps as large as 1.8 eV are reported for uv-laser–evaporated carbon films, while carbon films sputtered onto cold substrates exhibit gaps in the range from 0.6 to 1.0 eV. Because it contains no hydrogen, *i*-C exhibits no C–H stretch absorption band and so is more transparent in the infrared than is *a*-C:H.

7.2.2 Mechanical and Chemical Properties. Without hydrogen to interrupt the connectivity of the carbon network and so reduce the degree of overconstraint, hydrogen-free amorphous carbons are potentially much harder than *a*-C:H even with significant sp^2 bonding. The empirical hardness relationship [Eq. (5)] suggests a minimum of 22 GPa for an all-sp^2 network without clustering. Cold-sputtered carbon films exhibit hardness in this range. For films deposited from a cathodic arc without mass filtering, hardness in the 20- to 40-GPa range is obtained. With mass filtering, hardness and elastic modulus approaching that of diamond are obtained.

The density of *i*-C is maximized with its hardness. Densities as high as 3.1 g/cm^3 have been determined both directly by the sink-float method and indirectly from the electron energy-loss (EELS) spectrum. This is significantly greater than that of graphite (2.2 g/cm^3) and approaches that of diamond (3.4 g/cm^3). Unfortunately, high hardness in *i*-C is accompanied by extremely high compressive stress, which limits film thickness to several hundred nanometers at most. Laser-ablated carbon films are similar in density and hardness, but do not suffer this thickness limitation. Like *a*-C:H, *i*-C exhibits very low wear in sliding friction with steel. The coefficient of friction is usually slightly higher.

7.2.3 Electronic Properties. Like *a*-C:H, *i*-C can be described as a composite of small conducting, presumably graphitic, islands in an insulating matrix. If physically realizable, "amorphous diamond" with absolutely no sp^2 component should be a large-gap amorphous semiconductor. Indeed, the rectifying characteristic of metal–*i*-C junctions are consistent with that of a wide-gap semiconductor (McKenzie *et al.*, 1991).

7.3 Structure of *i*-C

Electron energy-loss spectroscopy (EELS) has proven to be a very powerful analytic tool in the study of thin films of *i*-C. The low-energy EELS spectrum of an *i*-C film deposited from a cathodic arc with mass filtering is compared to the EELS spectra of diamond, graphite, and *a*-C in Fig. 14 (Berger *et al.*, 1988). Two major features are immediately apparent in the EELS spectrum of *i*-C:

1. no features associated with π-π^* transitions near 6 eV are evident, and
2. the $\pi+\sigma$ plasma energy, which is a measure of the volume density of valence electrons, corresponds to a mass density of approximately 3.1 g/cm^3.

Figure 15 compares the higher-energy core-level K(1*s*)-edge EELS spectrum of diamond to those of amorphous carbon films deposited on diamond at room temperature by 248-nm uv-laser evaporation. Although a small feature

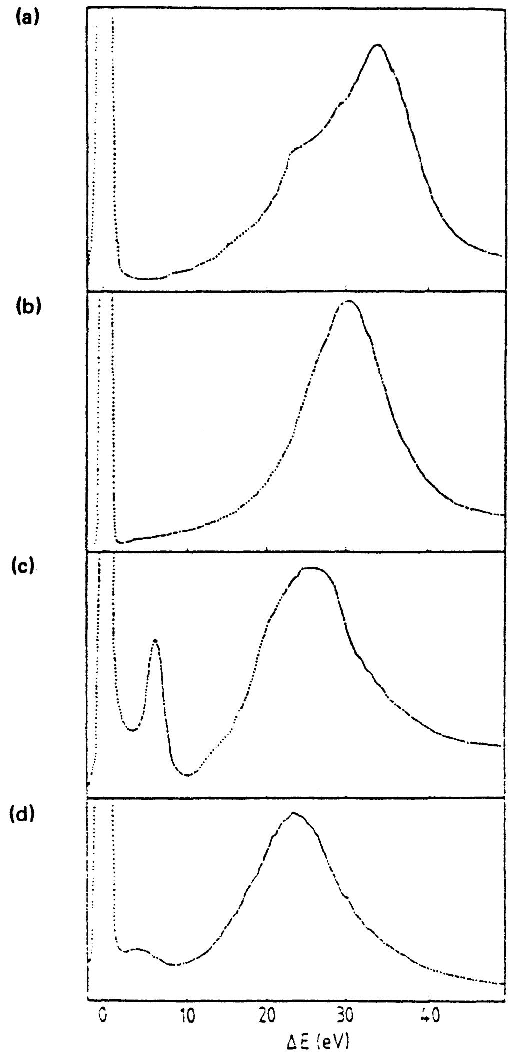

FIG. 14. Low-energy electron energy-loss spectra of carbon materials: (a) diamond, (b) *i*-C, (c) graphite, (d) *a*-C.

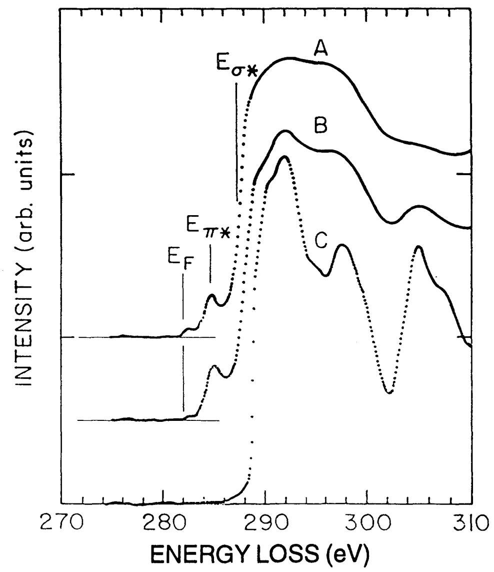

FIG. 15. Core-level [*K*(1*s*)-edge] EELS spectra of carbon films uv-laser evaporated onto cold diamond substrates compared to that of crystalline diamond. Curves A and B are for carbon films deposited on cold (77 K) diamond substrates by uv-laser evaporation. Curve C is for single-crystal diamond. The loss feature associated with residual sp^2-bonded carbon is labeled $E_{\pi*}$. *The edge associated with covalent bonding, both quaternary and graphitic, is labeled* $E_{\sigma*}$. The fraction of integrated intensity below this threshold is used to estimate an sp^3 fraction of 85%. The similarities in the details of the loss features are strongly suggestive of an "amorphous diamond" structure.

due to residual sp^2 bonding is still evident, the similarity of the spectra at higher energy is consistent with an "amorphous diamond" structure.

The radial distribution function (RDF) of *i*-C has been measured by both electron and neutron diffraction (Gaskell *et al.*, 1991). The two most important parameters to be obtained from the RDF are the average distance to the first-nearest neighbors, r_1, which is a measure of bond length, and the average coordination, n_1. For diamond $r_1 = 1.54$ Å and $n_1 = 4$, while for graphite $r_1 = 1.42$ Å and $n_1 = 3$. Amorphous carbon sputtered onto a cold (77 K) silicon substrate has $n_1 = 3.3$, indicative of a small (approximately 10%) fraction of quaternary carbon. However, *i*-C was found to have $r_1 = 1.53$ Å and $n_1 = 4$, essentially the

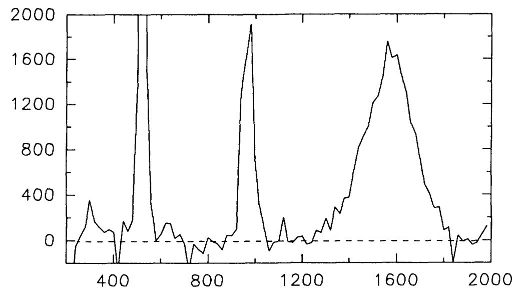

FIG. 16. Raman spectrum of *i*-C. The strong features at 500 and 900 cm^{-1} are the first- and second-order spectra of the Si substrate.

same as diamond. By this measure, *i*-C actually is "amorphous diamond."

Figure 16 shows the Raman spectrum of a film of *i*-C on silicon. The high level of noise in the spectrum is a consequence of the low absorbance of this transparent and very thin (0.1 μm) film. The only clear feature is a broad, symmetric peak at 1600 cm^{-1}. This is evidence for a residual sp^2 component that determines the optical gap. Note that the shape of the peak strongly resembles that for the polymeric *a*-C:H film with a similar 2.5-eV optical gap (see Fig. 11), suggesting graphitic clusters of comparable size. However, the peak is shifted nearly 200 cm^{-1} higher in *i*-C, presumably because of the huge compressive stress in this material. The Raman spectrum of the "amorphous diamond" component is too weak to be detectable in this film, but is expected to resemble that of amorphous silicon with appropriate scaling for the stiffer C–C bond and smaller atomic mass.

7.3.1 Structural Models. It is attractive to describe the densest *i*-C as an amorphous all-quaternary network similar to other amorphous semiconductors. Indeed, the radial distribution function of *i*-C is a very close match to that of amorphous germanium while the EELS spectrum closely resembles that of diamond. Because the stiffness of carbon bonding against bond-angle deformation relative to bond-length distortions is so much greater than in Si and Ge, the stability of such a network was initially somewhat doubtful. However, the large compressive stress in these films may play a key role in stabilizing the highly strained network. Homogeneous "amorphous diamond" may in fact be strongly metastable so long as the compressive stress can be maintained by the film–substrate interface. Phase separation into highly quaternary nodules in a more graphitic matrix may be the mechanism by which laser-ablated carbon maintains a high average sp^3 content without accumulation of unsustainable interfacial stress.

7.3.2 The *i*-C Deposition Process. It appears that the substantial arrival energy of carbon ions correlates with a large steady-state fraction of quaternary carbon. This is counterintuitive in that the excess kinetic energy would be expected to promote annealing to graphite rather than maintaining a highly unstable network structure. Two mechanisms are likely to be in effect. First is "subplantation," a shallow implantation process in which incoming carbon ions continuously rearrange the atomic structure to a depth of a few angstroms with a bias toward disrupting graphitic configurations (which requires only 25 eV) without disturbing quaternary structures (requiring 80 eV). Deposition with too little or too much kinetic energy loses this selectivity. Although a finite steady-state quaternary fraction can be maintained, an entirely quaternary structure should be unattainable by such a mechanism. The second mechanism is stress stabilization. The very large compressive stress that appears in this shallow implantation process may actually shift the relative energies of various bonding configurations in favor of fourfold coordination. As shown in the graphite–diamond phase diagram (Fig. 1), even at room temperature, compressive stress in excess of only 2 GPa is sufficient to stabilize diamond; *i*-C forms with stress up to 8 GPa. It is unclear which mechanism is really driving *i*-C deposition, or whether they can even be separated. In one view, the details of carbon-bonding energetics may be irrelevant and the "optimal" kinetic energy simply that which maximizes compressive stress, while the opposite view holds that stress is an accidental by-product of the kinetic process that maximizes fourfold coordination.

The same interaction between kinetics and compressive stress applies for carbon deposition at suboptimal energy, as in laser evaporation. At lower kinetic energy, insufficient stress is generated to stabilize a largely quaternary network, or rather, the kinetic energy is insufficient to promote a large quaternary fraction. The role of low temperature and high thermal conductivity is to "freeze in" whatever quaternary configurations do form by minimizing and dissipating excess energy, which would otherwise promote annealing to more stable graphitic structures. Immediate covering of the reactive surface also assists in capturing the quaternary component. High instantaneous deposition rates inherent to laser evaporation and ablation assist in this quenching process.

8. APPLICATIONS OF DIAMONDLIKE CARBON

A major impediment to many applications is the very large compressive stresses that appear in all diamondlike amorphous carbon films. For the hardest *a*-C:H the compressive stress approaches 4 GPa, while for the hardest *i*-C the stress is as high as 8 GPa. These enormous compressive stresses limit the thickness of the film to the strength of the interface or the substrate itself. For *a*-C:H, 2 μm is an approximate limit, while for *i*-C, films as thin as 1000 Å are apt to de-adhere. It is possible that the high degree of quaternary coordination required for hardness is actually stabilized by the compressive stress, and so is inseparable from the desirable diamondlike qualities.

Hydrogenated DLC is already in use in several products. Because it is very smooth, thin, flexible, and insulating, *a*-C:H is an ideal protective coating for thin-film magnetic recording media. For hard disks it can virtually eliminate adhesion to the head during stop–start events and so prevent catastrophic failures. For metal-film recording tape, it provides a chemical barrier against oxidation and reduces both tape and tape-head wear. It is used to reduce friction and wear on tape-recording heads and tape-transport guides. *a*-C:H is an ideal protective and antireflection coating for infrared optical objectives made of germanium. With proper choice of thickness and index of refraction, a DLC coating can reduce reflection almost 99% while providing protection against corrosion and particle abrasion. DLC has also been used to stiffen metal-foil high-fidelity tweeter diaphragms and so raise their resonant frequency far beyond the range of human hearing. A form of *a*-C:H formed by a dual–ion-beam process, which increases its transparency and slightly reduces compressive stress while maintaining hardness, is used to form protective coatings on sunglasses and windows for bar-code scanners.

Potential applications include protective coating of glass and plastics, low-friction coatings for sliding and rolling bearings, chemical protection barriers for metals, plastics, and glass, tools (most notably surgical blades), and biocompatible implants. Any material that is inert, transparent, hard as ceramic, and slippery as Teflon and can be deposited on virtually any substrate will certainly find uncountable applications.

ACKNOWLEDGMENT

This work was sponsored in part by the Strategic Defence Initiative Organization/Innovative Science and Technology through the Office of Naval Research.

GLOSSARY

Catalyst: Substance that increases the rate of a chemical reaction. The catalyst is not consumed by the chemical reaction.

Coefficient of Friction: Force required to cause two surfaces to slide on each other. It is unitless, being equal to the force required to cause sliding divided by the force pushing the surfaces together.

Cubic Crystals: Form of crystal in which the minimum repeating unit of atoms has the same symmetry as a cube. Gold, sodium chloride (table salt), silicon, and diamond have cubic crystal forms.

Fluorescence: Emission of light from matter as a result of its being irradiated with energy, usually from ultraviolet light, x rays, or electrons. Dayglow paint is an example of a fluorescent material.

Metastability: Stable condition of an element or compound that is not in its minimum energy state. Diamond is metastable with respect to graphite since graphite is the lowest energy form of carbon.

Miller Indices: Method of defining the planes and axes of crystals. Parentheses ( ) indicate crystal planes, and angle brackets ⟨ ⟩ indicate an axis perpendicular to the planes. For a cubic crystal, consider a cube aligned with the axes of the crystal: the (100) is a set of crystal planes parallel to the square faces of the cube, the (110) is a set of planes perpendicular to the diagonals across any one of the faces of the cube, and the (111) is the set of planes perpendicular to any one of the diagonals connecting any two opposite corners of the cube.

Modulus of Elasticity: Force required to deform a solid elastically. It has units of pressure.

Nonequilibrium: Condition of a chemical system that cannot be characterized by one temperature.

Radical: Atom or molecule in a very chemically reactive state as a result of not having all of its chemical bonding capability satisfied. A hydrogen atom (H) and methane with one hydrogen atom missing (CH_3) are examples of radicals. They react with themselves in a few milliseconds or less, forming H_2 and CH_3CH_3, respectively, However, some radicals can exist in this metastable state for hours or days.

Semiconductors: Solid or liquid whose electrical conductivity is between that of an insulator and a metal. Its conductivity can be changed by several orders of magnitude by incorporating impurities (dopants) at concentrations of a few parts per million.

Semiconductor, *n* Type: Semiconductor in which the electrical current is carried primarily by negatively charged electrons. These mobile electrons originate from positively charged impurities (dopants) or crystal defects in the semiconductor. Phosphorus is an *n*-type dopant of silicon and germanium.

Semiconductor, *p* Type: Semiconductor in which the electrical current is carried primarily by positively charged holes. These holes originate from negatively charged impurities (dopants) or crystal defects in the semiconductor that remove electrons from the conduction band. This forms voids (holes) with an effective positive charge. Boron is a *p*-type dopant of diamond, silicon, and germanium.

Surface Energy: Energy required to increase the surface of a solid or liquid. It has units of energy per area. The surface energy is often very dependent upon impurities that can coat the surface. For example, soap is known to lower substantially the surface energy of water.

Vacancy: Defect in a crystal lattice caused by the removal of an atom.

List of Works Cited

Aisenberg, S., Chabot, R. (1971), *J. Appl. Phys.* **42**, 2953–2958.

Aisenberg, S., Kimock, F. M. (1988), *Mater. Sci. Forum* **52-53**, 1–14.

Angus, J. C. (1991), Case Western Reserve University, private communication.

Angus, J. C., Jansen, F. (1988), *J. Vac. Sci. Technol. A* **6**, 1778–1782.

Bachmann, P. K., Leers, D., Lydtin, H. (1991), *Diamond Related Mater.* **1**, 1–11.

Bazhenov, V. K., Vilkulin, I. M., Gontar, A. G. (1985), *Sov. Phys. Semicond.* **19**, 829–840.

Berger, S. D., McKenzie, D. R., Martin, P. J. (1988), *Philos. Mag. Lett.* **57**, 285–290.

Berman, R. (1979), in: Field, J. E. (Ed.), *Properties of Diamond*, New York: Academic, Chap. 1.

Bhushan, B., Kellock, A. J., Cho, N.-H., Ager, J. W., III (1992), *J. Mater. Res.* **7**, 404–410.

Burton, E. (1978a), *Diamonds*, Radnor, PA: Chilton, Chap. 4.

Burton, E. (1978b), *Diamonds*, Radnor, PA: Chilton, Chap. 19.

Bush, D. M. (1979), in: Field, J. E. (Ed.), *Properties of Diamond*, New York: Academic, Chap. 19.

Bush, J. V., Dismukes, J. P. Nallichern, N. V., Walton, K. R. (1991), in: Tzeng, Y., Yoshikawea, M., Murakawa, M., Feldman, A. (Eds.), *Applications of Diamond Films and Related Materials*, New York: Elsevier, pp. 623–633.

Caveney, R. J. (1979), in: Field, J. E. (Ed.), *Properties of Diamond*, New York: Academic, Chap. 20.

Collins, A. T., Lightowlers, E. C. (1979), in: Field, J. E. (Ed.), *Properties of Diamond*, New York: Academic, Chap. 3.

Collins, C. B., Davanloo, F., Jander, D. R., Lee, T. J., Park, H., You, J. H. (1991), *J. Appl. Phys.* **69**, 7862–7870.

Cuomo, J. J., Doyle, J. P., Bruley, J., Liu, J. C. (1991), *Appl. Phys. Lett.* **58**, 466–469.

Dimigen, H., Hübsch, H. (1983/1984), *Philips Tech. Rev.* **41**, 186–197.

Dimigen, H., Hübsch, H., Memming, R. (1987), *Appl. Phys. Lett.* **50**, 1056–1058.

Dischler, B., Bubenzer, A., Koidl, P. (1983), *Solid State Commun.* **48**, 105–108.

Dismukes, J. P., Bush, J. V. Nallichern, N. V., Walton, K. R. (1991), in: Tzeng, Y., Yoshikawea, M., Murakawa, M., Feldman, A. (Eds.), *Applications of Diamond Films and Related Materials*, New York: Elsevier, pp. 635–644.

Dodson, M. H. (1989), *Nature* **337**, 207–XXX.

Du Pont de Nemours, I. E., and Company (1965), Netherlands Patent Release No. 6506395, 22 November.

Edwards, D. F., Philipp, H. R. (1985), in: Palk, E. D. (Ed.), *Handbook of Optical Constants of Solids*, New York: Academic, pp. 665–667.

Gaskell, P. H., Saeed, A., Chieux, P., McKenzie, D. R. (1991), *Phys. Rev. Lett.* **67**, 1286–1289.

Geis, M. W. (1991), *Proc. IEEE* **79**, 669–675.

Geis, M. W., Efremow, N. N., Woodhouse, J. D., McAleese, J. D., Marchywka, M., Socker, D. G., Hochedez, J. F. (1991a), *IEEE Electron Devices Lett.* **12**, 456–459.

Geis, M. W., Gregory, J. A., Pate, B. B. (1991b), IEEE *Trans. Electron Devices* **38**, 619–626.

Grill, A., Patel, V., Meyerson, B. (1991), *J. Mater. Res.* **5**, 2531–2537.

He, H., Thorpe, M. (1985), *Phys. Rev. Lett.* **54**, 2107–2110.

Himpsel, F. J., Knapp, J. A., VanVechten, J. A., Eastman, D. E. (1979), *Phys. Rev. B* **20**, 624–627.

Holland, L., Ojha, S. M. (1976), *Thin Solid Films*, **38**, L17–L19.

Howes, V. R. (1965), in: Berman, R. (Ed.), *Physical Properties of Diamond*, Oxford: Clarendon, pp. 174–183.

Kaukonen, H.-P. Nieminen, R. M. (1992), *Phys. Rev. Lett.* **68**, 620–623.

Koidl, P., Wild, Ch., Dischler, B., Wagner, J., Ramsteiner, J. (1989), *Mater. Sci. Forum* **52-53**, 41–70.

Lifshitz, Y., Kasi, S. R., Rabalais, J. W. (1989), *Phys. Rev. Lett.* **62**, 1290–1293.

McKenzie, D. R. *et al.* (1991), *Diamond Relat. Mater.* **1**, 51–59.

Oguri, K., Arai, T. (1990), *J. Mater. Res.* **5**, 2567–2571.

Okano, K., Kiyota, H., Iwasaki, T., Nakamura, Y., Akiba, Y., Kurosu, T. Iida, M., Nakamura, T. (1990), *Appl. Phys. A* **51**, 344–346.

Orowan, E. (1949), *Rep. Prog. Phys.* **12**, 185–232.

Pappas, D. L., Saenger, K. L., Bruley, J., Krakow, W., Cuomo, J. J., Gu, T., Collins, R. W. (1992), *J. Appl. Phys.* **71**, 5675–5684.

Prins, J. F. (1982), *Appl. Phys. Lett.* **41**, 950–952.

Rio Grand (1992), Albuquerque, NM, a source of pricing of gem-quality diamonds.

Robertson, J., O'Reilly, E. P. (1987), *Phys. Rev. B* **35**, 2946–2957.

Seal, M. P. (1953), *Proc. R. Soc. London Ser. A* **248**, 279–393.

Tamor, M. A., Vassell, W. C., Carduner, K. R. (1991), *Appl. Phys. Lett.* **58**, 592–594.

Thorpe, M. F. (1983), *J. Non-Cryst. Solids* **57**, 355–370.

Tzeng, Y., Yoshikawea, M., Murakawa, M., Feldman, A. (Eds.) (1991), *Applications of Diamond Films and Related Materials*, New York: Elsevier, Chap. 6.

U. S. Department of the Interior (1989), Bureau of Mines, *Minerals Yearbooks, 1950–1989,* Washington, DC: U.S. GPO.

Weissmantel, C. (1982), *Thin Solid Films* **92**, 55–63.

Wentorf, R. H. (1965), *Adv. Chem. Phys.* **9**, 356–404.

Zhu, W., Stoner, B. R., Williams, B. E., Glass, J. T. (1991), *Proc. IEEE* **79**, 621–646.

Further Reading

Angus, J. C., Hayman, C. C. (1988), *Science* **248**, 913–921. A general review of both diamond and diamondlike materials.

Burton, E. (1978), *Diamonds,* Radnor, PA: Chilton.

Field, J. E. (Ed.) (1979), *Properties of Diamond,* New York: Academic.

Pouch, J. J., Alterovitz, S. A. (1990), *Properties and Characterization of Amorphous Carbon Films,* Aedermannsdorf: TransTech. An extensive collection of short articles on deposition, characterization, and applications of diamondlike carbon.

Proc. IEEE (1991), **79**, special issue on "Large Bandgap Electronic Materials and Components."

Robertson, J. (1992), *Surf. Coatings Technol.* **50**, 185–203. A detailed scientific review of the structure and properties of diamondlike materials.

Tzeng, Y., Yoshikawea, M., Murakawa, M., Feldman, A. (Eds.) (1991). *Applications of Diamond Films and Related Materials,* New York: Elsevier.

DIELECTRIC PROPERTIES OF INSULATORS

JAMES F. SCOTT, *Department of Physics, University of Colorado at Boulder, Boulder, Colorado, U.S.A.*

Introduction 1
1. Polar Dielectrics 2
1.1 Local Fields and Polarizability Theory 2
2. Ferroelectrics 3
2.1 Hysteresis 3
2.2 Switching Kinetics 5
2.3 Mean-Field Theory of Phase Transitions 5
2.4 Order–Disorder versus Displacive Systems 8
2.5 Ultraweak Ferroelectrics 9
2.6 Domain Structures 9
3. Ferroelastics 10
4. Electret Effects 10
5. Electrostriction 11
6. Electrocaloric Effect 11
7. Piezoelectricity 11
8. Numerical Values 12
8.1 Spontaneous Polarization 12
8.2 Curie Constant 12
8.2.1 Quantum Cryogenic Effects 12
8.2.2 Calculating Curie Constants 12
8.2.3 Numerical Self-Consistency of Mean-Field Theory 14
8.3 Dielectric Constants 14
9. Preparation of Dielectric Thin Films 15
9.1 Sputtering 15
9.2 Sol-Gel 15
9.3 Slip-Casting, Thermal Evaporation, and Electron-Beam Evaporation 15
9.4 MOCVD 15
9.5 MBE 15
9.6 Continuous Methods 15
9.7 Ferroelectric Liquid Crystals 16
9.8 New Applications of High-Dielectric Films 16
9.8.1 DRAMs 16
9.8.2 Ferroelectric Memories 16
9.8.3 Optoelectronics 16
9.8.4 Pyroelectrics 17
10. Nonpolar Dielectrics and Dielectric Breakdown 17
10.1 Intrinsic Breakdown in Insulators 17
10.2 dc Thermal Breakdown 17
10.3 Impulse Thermal Breakdown 18
10.4 Avalanche Breakdown 18
10.5 Mathematical Models of Intrinsic Breakdown 18
10.5.1 Polar Materials 18
10.5.2 Nonpolar Materials 18
10.5.3 Collective Breakdown 18
10.5.4 Glasses/Amorphous Dielectrics 19
10.6 Models of Thermal Breakdown 19
10.7 Models of Avalanche Breakdown 19
10.8 Breakdown in Specific Materials 19
Glossary 20
Works Cited 21
Further Reading 21

INTRODUCTION

The analyses of electrical and magnetic properties of insulating solids are two nominally analogous problems in solid state physics that have traditionally borrowed much from each other in both mathematical approach and even terminology: The terms "ferroelectric" and "paraelectric" are, for example, misnomers that found their way into standard usage to describe electrical polarization states in crystals analogous to ferromagnetic and paramagnetic spin states in magnetic materials. There are two important differences, however, between descriptions of the dielectric and magnetic properties of insulators: First, as emphasized by Feynman in 1965, magnetism is an intrinsically quantum mechanical

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

property of solids, and attempts to describe it semiclassically (e.g., in terms of an array of classical spin vectors) are fundamentally unsound; by comparison, except at cryogenic temperatures, the electrical behavior of polar insulators is intrinsically classical, with polarizations representable by localized vectors. It is the aim of the section that follows below to describe the microscopic "local field" averaging procedure by which these vectors are estimated.

The second difference between magnetic and electrical properties of insulators is that the magnetic spins are well-defined spatially localized quantities, even in the high-temperature disordered paramagnetic state in which the expectation value of the magnetization is zero: $\langle M\rangle = 0$. By comparison, in a few polar insulators, a ferroelectric state arises from ordering (e.g., rotation) of rigid dipoles, such as an N–H– –Cl hydrogen bond or a covalent ion (NH_4^+, SO_4^{-2}); but, in general, ferroelectric materials possess no permanent dipoles, ordered or disordered, in their paraelectric phases. In this sense ferroelectrics and antiferroelectrics are far more complicated that their ferromagnetic and antiferromagnetic analogues. For magnets, the spins exist at all temperatures and the spin magnitude S is independent of T, although the thermal averages $\langle S_z\rangle$ can be quite complicated to evaluate; by comparison, for dielectric crystals the local dipoles $P_i(T)$, whose spatial averages yields a macroscopic polarization P, depend in a nontrivial way on T and often vanish individually at elevated temperatures.

1. POLAR DIELECTRICS

1.1 Local Fields and Polarizability Theory

The development of theories for the dielectric properties of polar materials is far more detailed than for nonpolar materials. This is because in nonpolar materials the interaction constant between the electric field and the solid is generally unknown; even the physical mechanism of interaction may be imprecisely known. By comparison, the interaction of polar solids and applied electric fields is assumed to be dominated by simple Coulombic fields, so that some sort of multipole expansion will describe the result. The problems lie in approximating the medium as continuous and defining dielectric constants (and fields) over length scales that are large compared with nearest-neighbor atomic separations, yet small compared to all other characteristic lengths in the system (e.g., sample size).

Historically, there have been two fundamental problems of great interest: First, is the self-alignment of dipoles in a polar crystal (which we call "ferroelectricity") due to Coulomb forces alone, or are short-range interaction among the ions required? And second, how can one express macroscopic quantities, such as the dielectric constant E, in terms of microscopic atomic parameters, such as the polarizability α of a specific atom or molecule (Clausius–Mosotti relationship)?

It was once assumed that dipoles interacting through only Coulomb fields (no short-range interactions) can provide a model of ferroelectricity. Fröhlich's (1937) model assumes an isotropic medium and calculates certain average values over a sphere of radius large compared with a lattice constant a_0, and small compared with the crystal size r (usually assumed infinite). Outside this sphere, a continuous-medium approximation is made, with a dielectric constant ϵ. Inside the sphere is a vacuum, with a (dimensionless) cavity field G. When the sphere is filled with a material of uniform polarization P, the electric field inside the field changes from E_S to E_S+R, where R is called the reaction field. One finds

$$R=[2(G-1)/(2\epsilon+1)]P(4\pi/3). \tag{1}$$

In the limiting case of one polar molecule within each sphere, one obtains $\epsilon-1=[\epsilon/(2\epsilon+1)](4\pi N\mu^2/kT)$, where N is the number of molecules, each of dipole moment μ, per unit volume. This gives $\epsilon\rightarrow\infty$ as $T\rightarrow 0$ but no phase transition. Thus long-range reorientable dipoles in an isotropic medium cannot spontaneously order to produce ferroelectricity at finite T.

If we assume that the local dipoles do order for an unspecified reason (such as short-range forces), then a numerical estimate for the resulting polarization is 10^{-18} esu in a unit cell of order $4\times4\times4$ Å^3, which produces a Curie constant of 1.4×10^3 K. This is a typical Curie constant in a hydrogen-bonded system such as Rochelle salt or tri-glycine sulfate (TGS) and thus formed the basis of ferroelectric models in which re-orientable permanent dipoles played the main role. The first model of that type was provided by Kobeko and Kurchatov (1930).

This internal-field model of Fröhlich leads algebraically to the expression

$$(\epsilon-1)/(\epsilon+2)=4\pi N\alpha/3, \tag{2}$$

which is known as the Clausius–Mosotti relationship. It can be rewritten as

$$\epsilon=[1+8\pi N\alpha/3]/[1-4\pi N\alpha/3], \tag{3}$$

which requires that the dielectric constant diverge at $N\alpha=3/4\pi$. That is referred to as the "$4\pi/3$ catastrophe" and implies that a finite polarization will arise in zero applied electric field, i.e., ferroelectricity will spontaneously arise.

Onsager (1936) extended Fröhlich's model to modify the internal fields. He replaced the depolarization factor 4π discussed above with a general value β that could vary from 4π to $4\pi\times10^{-2}$. It is largest in systems such as tris-sarcosine calcium chloride (TSCC), in which rigid dipoles truly rotate. This eliminates the unphysical result for polar liquids, such as water, for which Fröhlich's theory predicted ferroelectricity at room temperature. Onsager modified the internal field so that the effective field for permanent dipoles can be quite different from that for induced dipoles. The initial application of Onsager's modified theory was made by Slater (1950) for barium titanate. He assumed that α is temperature dependent because it involves ionic displacements. In $BaTiO_3$, he assumed that all the polarization comes from the Ti ion and one of the oxygen ions (designed O_I). If we write $F_{int}=E+[(4\pi/3)+s]P$ in place of Fröhlich's $E+(4\pi/3)P$, we find for barium titanate that the coefficient s is indeed zero at the cell origin $[(0,0,0)$ or $(\frac{1}{2},\frac{1}{2},\frac{1}{2})]$, but is very different at other ion sites: $s(0,\frac{1}{2},\frac{1}{2})=4.3$; $s(\frac{1}{2},\frac{1}{2},0)=-8.7$; $s(\frac{1}{2},0,0)=-15$; *and* $s(0,0,\frac{1}{2})=+30$. (It is useful to distinguish between displacement vector $D=E+4\pi P$ and the internal field $F_{int}=E+\beta P$, where typically $\beta=4\pi/3$. Here, the dielectric constant $\epsilon=dD/dE=1+4\pi\,dP/dE$ is defined for a substance, averaged over a volume large compared with that of the unit cell, whereas the polarizability of any one atom or ion is $\alpha=\mu/F^{int}$.) The criterion for the onset of ferroelectricity in this case then becomes

$$\left(\frac{4\pi}{3}\right)\frac{\Sigma(\alpha_{elec}+\alpha_{ion})}{V}=1, \tag{4}$$

where the sum is over all ions in unit cell of volume V. Slater determined for barium titanate that $(4\pi/3)\Sigma(\alpha_{ion}/V)=0.39$. Thus, in real ferroelectrics, the local fields at different ionic sites may produce an enhancement of the effects originally described by Fröhlich.

2. FERROELECTRICS

In many textbooks, a ferroelectric is defined as a solid that exhibits hysteresis behavior in its $P(E)$ polarization response to applied electric fields. Such a definition is unfortunate because it includes electrets (in which the hysteresis arises entirely from defects and impurities) and also crystals in which $P(E)$ is both lossy and nonlinear but which possess no collective phenomena, ordering temperatures, pyroelectricity, domains, or reversible lattice polarization. A more physical definition of a ferroelectric is that as a polar crystal whose dipoles are intrinsic to the lattice structure and can be reversed by application of a sufficiently large electric field. With this definition, it is necessary but not sufficient for materials to be both piezoelectric and pyroelectric in order for them to be ferroelectric. It is not necessary for ferroelectrics to have a phase transition to a paraelectric phase; many ferroelectrics remain ferroelectric up to their melting point or decomposition temperatures. An operative definition of a ferroelectric is a crystal belonging to a pyroelectric point symmetry class (C_n, C_{nv}, or C_{1h}) that can be switched electrically (coercive field E_c less than breakdown field E_B). Some authors choose to delete the switching requirement and hence to include highly conducting polar metals and semiconductors (e.g., SnTe) as ferroelectric.

Antiferroelectrics can be defined in an analogous way as crystals having pairs of oppositely directed dipoles in their lattice structure, i.e., consisting of two sublattices, each of which is ferroelectrically ordered, but in opposite directions. It is possible to confirm such structures via x-ray crystallography. However, there is no low-field electrical response that yields an unambiguous "signature" of antiferroelectricity. When large electric fields are applied to antiferroelectrics, two *P*(*E*) hysteresis loops are often observed, as shown in Fig. 1(b).

2.1 Hysteresis

It is helpful to discuss the hysteresis *P*(*E*) relationship in Fig. 1 in terms of the three parameters that determine it: E_c, the coercive field, is the *nominal* field required to switch the polarization; P_S, the saturation value of polarization, is extrapolated from asymptotically large applied field *E*; and P_r, the remnant polarization, is the polarization remaining after the applied field is switched off. In reality, E_c is not a constant. It has strong dependence on thickness *d*, increasing from ca. kV/cm in bulk to ~MV/cm at $d \cong 50$ nm. It also depends on the duration of the applied field; as discussed in Sec. 10, a dielectric with a breakdown voltage of 50 V applied as a 30-μs pulse may not tolerate 20 V for 1 s or stand off 5 V applied dc for a month. And E_c depends upon repetitive cycling (fatigue), usually decreasing with increasing number of cycles *N*.

The spontaneous saturation polarization P_S does not vary much with thickness, contrary to textbooks that predict its disappearance, due to depolarization effects at the surfaces, in very thin films. Here, P_S may even increase as $d \rightarrow 40$–50 nm, for reasons discussed in the "finite-size effects" section below. P_r does decrease markedly with fatigue. This behavior is probably due to the decoration of electrode interfaces (and grain boundaries, in ceramic specimens) with accumulated charged defects. Such defects may be intrinsic. For example, pairs of oxygen vacancies and lead vacancies in PZT ($PbZr_{1-x}Ti_xO_3$) form strong dipoles that compensate the lattice polarization and often preclude switching.

An obvious question to ask about the hysteresis curve in Fig. 1 is why the hysteresis loop is not square. Why does P_r not equal P_S? The finite slope of *P* versus *E* at E_C is easier to explain; it arises from a distribution of local coercive fields, due to domain wall structures and to defects. But why is P_S so much larger than P_r, often $\approx 2P_r$? In a pure single crystal, P_r should be only slightly less than P_S. Since a single-domain, polarized state is the ground state of the system below T_c, the effect of a finite temperature *T* is to put a small fraction of dipoles in the antiparallel configuration: % anti$=e^{-\mathbf{E}\cdot\mathbf{P}/kT}/(e^{+\mathbf{E}\cdot\mathbf{P}/kT}+e^{-\mathbf{E}\cdot\mathbf{P}/kT})$. However, under normal circumstances this fraction is small; $EP/kT \approx (10^6 V/m)\ (10^{-2} C/m^2)/(0.03\ eV) \approx (3\times 10ex^5\ J/m^3)/eV \approx 2\times 10^{19}\epsilon_0 = 2\times 10^8$.

There is also a linear response causing $P_S > P_r$, but this effect is also too small to explain observed data with $P_S \gg P_r$. Even if 100% of the dipoles are lined up along $\hat{z}$ in the absence of an applied field, in an applied field the dipoles will increase in length through charge separation of + and −. This linear response $\Delta P = \epsilon_0 \epsilon E$ is fairly small (even though ϵ increases during switching due to domain wall contribution).

So, in hysteresis curves where $P_S \gg 1.1P_r$, one may assume that a certain amount of backswitching is occurring where domains flip back after an insufficiently long voltage pulse is applied. The primary cause of backswitching is defect dipoles (such as Pb-O vacancy pairs in PZT). These are very slow to reorient and thus cause $P_r \ll P_S$ unless very low

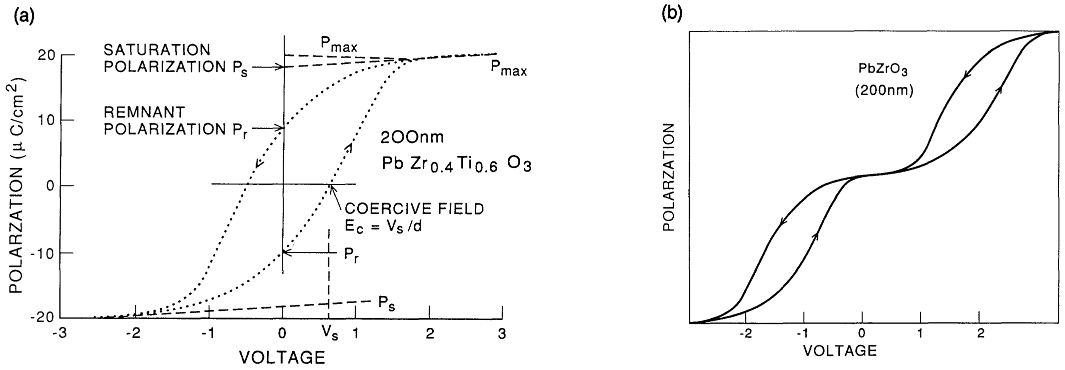

FIG. 1. Hystersis behavior *P*(*E*): (a) in a ferroelectric; (b) in an antiferroelectric.

frequencies are used to measure hysteresis. Backswitching will be greater in ceramics than in single crystals and greater in fatigued ceramics than in virgin specimens.

Backswitching is the reason why both switching and nonswitching current transients ($+E_C$ applied to $-P_r$ and $+P_r$ states, respectively) look qualitatively similar in PZT ceramics [Fig. 2(a)]. If there were no backswitching, we would expect the nonswitching current transient to look like that in Fig. 2(b), which is a voltage spike followed by a monotonic decay.

2.2 Switching Kinetics

Some elementary texts suggest that when a large external field is applied to a ferroelectric crystal, the dipoles reorient simultaneously and almost instantly. This, in fact, would occur at extremely high fields; but at normal fields in the kV/cm–MV/cm range, another mechanism is energetically easier: Domains nucleate, primarily through inhomogeneous nucleation processes at electrode interfaces and defect sites (grain boundaries in ceramics). These domains fluctuate in size until a critical radius is achieved. Then rapid growth occurs as one-dimensional needlelike domains, parallel and antiparallel to the applied field. Finally, the domain walls spread out sideways until the specimen is ~100% reversed from its initial polarization. The last step is usually the rate-limiting step, although in some ferroelectrics (GASH, guanidinium aluminum sulfate hexahydrate) the nucleation step is, and in others (TGS, triglycine sulfate), each step requires comparable time.

When sideways domain growth is the rate-limiting step, Stadler has shown, the switching time varies as $E^{-3/2}$ at high fields. This is in contrast to the activation field dependence $t_S = t_0 e^{-\alpha/E}$ observed generally at low fields.

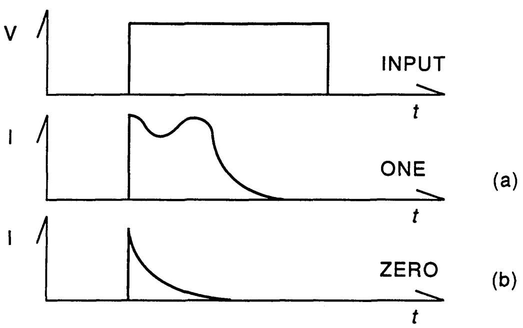

FIG. 2. Displacment current transient $i(t)$ in (a) switching ferroelectric, compared with (b) the nonswitching response.

Ishibashi (1986) has developed a theory of switching based upon earlier models of crystal growth due to Kolmogorov and Avrami. His model, summarized in Eq. (5) below, involves a characteristic switching time t_S, the dominant dimensionality d of domain growth, and the spontaneous polarization P. This three-parameter theory predicts the shape of the displacement current transient $i(t)$ quite well in KNO_3 and PZT at various temperatures and field (see Fig. 2),

$$
\begin{aligned}
i(t) &= \frac{2A\,\Delta P(t)}{\partial t} \\
&= (2P_s A n t_0^{-1})\left(\frac{t}{t_0}\right)^{n-1} \exp\left[-\left(\frac{t}{t_0}\right)^{n}\right], \qquad (5)
\end{aligned}
$$

where A is the electrode area; n, an exponent characterizing domain growth; and t_0, a characteristic growth time depending upon temperature and applied field.

It is found experimentally that there are significant effects in ferroelectrics that arise from finite sizes for the individual grains or, in the case of films, from the film thickness. The most significant effect is that the coercive field increases by about 10^3 as films decrease in thickness from 100 μm to a few tens of nm. Many physical parameters, including coercive field, vary as well with grain size r, typically as $r^{1/2}$ or $r^{-1/2}$.

2.3 Mean-Field Theory of Phase Transitions

In 1937, Landau developed a simple theory that qualitatively describes many of the phenomena exhibited by ferroelectrics near their Curie temperatures. The theory assumes that all dipoles i and j interact equally with each other, independent of separation distance V_{ij}. That is equivalent to an infinite-range potential and hence to replacing the sum of two-particle interactions by an effective field due to $n-1$ particles acting on the remaining particle. This kind of "mean-field" or "molecular-field" theory was first used by Van der Waals to describe the gas–liquid phase transition and by Weiss to describe ferromagnetic–paramagnetic ordering transitions. For fluids and magnets such a mean-field theory gives results only qualitatively in agreement with experiment, because the actual microscopic interactions are extremely short range. However, in ferroelectrics (and, incidentally, in superconductors) the interactions are very long range, being fundamentally Coulombic—which falls off very slowly, as $1/r$—so that mean-field theory gives good quantitative agreement with ferroelectrics experiments.

Landau's assumption is that the crystal's free energy $G(P,T)$ can be expanded in a power series in polarization P near the Curie temperature T_C:

$$G(P,T)=\tfrac{1}{2}A(T-T_C)P^2+\tfrac{1}{4}BP^4+\tfrac{1}{6}CP^6, \qquad (6)$$

where the linear temperature dependence of the quadratic term arises directly from the assumption of infinite-range forces (mean field) and the coefficients B and C are assumed to have no important temperature dependences.

If B and C are both positive, a continuous or "second-order" phase transition occurs at T_C. That is, the expectation value $\langle P(T)\rangle$ varies continuously to zero at T_C and discontinuities are found only in second order, i.e., $\partial P/\partial T$ has a jump at the Curie temperature. If B is negative but C is positive, a discontinuous phase transition is observed at $T_0=T_C$. There is a finite jump in $\langle P(T)\rangle$ and T_C and two (bistable) solutions to minimizing $G(P,T_C)$ exist, one with $\langle P\rangle=0$ and one nonzero. Other quantities may also be discontinuous at such first-order transitions, including spontaneous strain, so that crystals undergoing first-order transitions often crack at the transition temperature.

For first-order transitions, the are three temperatures of importance in addition to the actual transition temperature T_0. These may be understood in general by reference to Fig. 3, which shows the shape of the free energy $G(T,P)$; $G(T,X)$ is equivalent to $G(T,P)$, where X is the displacement of ions that produces the polarization P, and P is proportional to X. Below T_0, only the ferroelectric (FE) phase can exist. Between T_0 and $T_C=T_0+3B^2/16AC$ ($=T_0+7.7$ K in BaTiO$_3$) the nonpolar phase is metastable and the FE phase is stable. (T_C is the extrapolated T at which the FE transition would become continuous, i.e., the Curie temperature extrapolated from the low-T phase.) Between T_C and T_1, the nonpolar phase is stable and the polar phase is metastable and can be induced by applied field;

$$T_1=T_0+4B^2/AC(=T_0+10 \text{ K in BaTiO}_3) \qquad (7)$$

(*n.b.*: T_1-T_0 is the width of the thermal hysteresis). Between T_1 and T_2, the ferroelectric phase can be induced *only* by an applied field

$$T_2=T_0+9B^2/20AC(=T_0+18\ ^\circ\text{C in BaTiO}_3). \qquad (8)$$

In this discussion T_0 is the actual transition temperature at which there are three equal minima in the free energy $G(X)$; T_1 is the temperature at which $G(X)$ develops an inflection point. See Fig. 3.

The coefficient B may be a function of other parameters, such as hydrostatic pressure, uniaxial stress S, or electric field E, along the polar axis (i.e., the field E_j "conjugate" to the spontaneous polarization P_j), and B may change sign as a function of these parameters. The point in phase space where $B\ (E,T)$ [or $B(S,T)$] changes sign is called a "*tri*critical

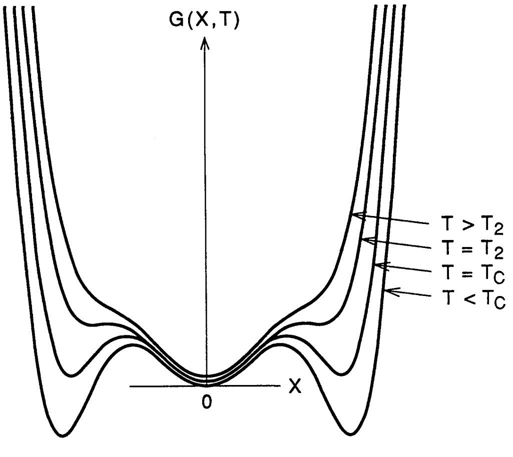

FIG. 3. Free energies $G(T,P)$ versus temperature from Landau theory. Above T_c but below T_1, the free-energy minimum at $P=0$ is deeper than those on the sides ($P\neq 0$); in this range the nonpolar phase is stable but the polar phase is metastable and can remain for long times. Above T_1, but below T_2, the minima in the free energy at positive and negative P grow smaller as T increases, until at T_2 only inflection points in $G(p)$ remain. A stable ferroelectric phase can be induced by an applied electric field between T_c and T_2. A metastable ferroelectric phase can remain even after the field is turned off between T_c and T_1. Ferroelectricity cannot be induced by any applied field above T_2. The text gives numerical values for T_2–T_c and T_1–T_c for BaTiO$_3$.

point" because it connects *three* lines of continuous phase boundaries, one of which lies in the $E=0$ plane, and two which extend into $\pm E$ space, as shown in Fig. 4.

A different kind of critical point can occur when the field is applied perpendicular to the polar axis. Such a case is illustrated in Fig. 5 for $NaNO_2$, which has ferroelectric phase separated from paraelectric phase by a spirally modulated (incommensurate) phase. As the applied field is increased, the intermediate phase disappears and the ferroelectric and paraelectric phases approach each other tangentially. The point where they come together is called a Lifshitz point.

For conjugate fields E_j, there may be quite different phase diagrams, such as that in Fig. 6, which illustrates a triple point (coexistence of three phases), a tricritical point (discussed above) where the transition changes from continuous to discontinuous, and even a critical end point, like that in gas–liquid systems. Note that a critical end point is possible only where the applied field is the conjugate field, because in this case it is not a symmetry-breaking field. That is, there is no need for a phase boundary between ferroelectric and paraelectric states in the presence of a large electric field along the polar direction. [Similarly, there is no need for a phase boundary between ferromagnetic (FM) and paramagnetic (PM) states in finite magnetic fields, as shown in Fig. 7. FI in this figure designates a six-sublattice ferrimagnetic state.]

The simple free energy given above describes many real ferroelectrics extremely well. The coefficients *A*, *B*, and *C* are greatly over-determined from available experimental data. Usually *A* is determined from measurement of dielectric constants $\epsilon(T)=(\partial^2G/\partial P^2)^{-1}=4\pi\mathscr{C}/(T-T_C)$, where $\mathscr{C}$ is called the Curie constant and should not be confused with *C* in the free-energy expansion. *B* is determined from graphing ϵ^{-1} vs P^2 $[\epsilon^{-1}=A(T-T_C)+3BP^2+5\delta P^4)]$ and *C* is determined from P_S just below T_0, where $P_S^2=-B/C$.

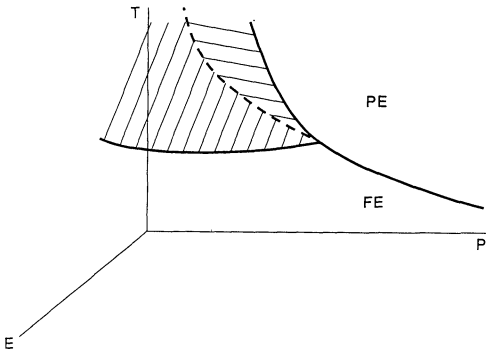

FIG. 4. Free energy surfaces $G(T,E)$ near a tricritical point.

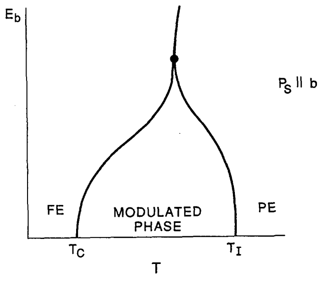

FIG. 5. Phase diagram near a Lifshitz point.

But many additional experimental measurements give a self-consistency check. For example, the width of thermal hysteresis is given by $\Delta T=T_1-T_0=B^2/4AC$. The specific heat jump is given by $\frac{1}{2}AP_S^2(T_0)+\frac{1}{4}P_s^4(T_0)\partial B/\partial T$, and so forth. For the best studied ferroelectrics *A*, *B*, and *C* are each known to better than 5% accuracy; e.g., in TGS, $A=3.92\times10^{-3}\ K^{-1}$; $B=8.0\times10^{-10}$ cgs; $C=5.04\times10^{-18}$ cgs. In $BaTiO_3$, $A=7.4\times10^{-5}\ K^{-1}$; *B* is negative and depends linearly on *T* as $B=18\times10^{-15}(T-T_2)$, $T_2=175\ ^\circ C$; $C=33.4\times10^{-23}$ cgs.

Unfortunately, the free energy discussed above is too simple to describe some complicated ferroelectrics. For some crystals the order parameter that describes the lattice instability is not polarization *P* but some other variable, such as strain e_j, with polarization coupled nonlinearly to e_j. In these systems, termed "improper" ferroelectrics, the spontaneous polarization of the low-symmetry state

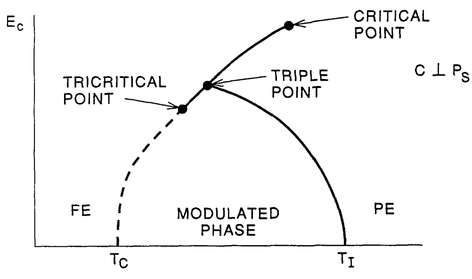

FIG. 6. Phase diagram illustrating triple point, tricritical point, and critical end point. Figures 5 and 6 schematically approximate the real situation in $NaNO_2$.

arises as a kind of "accident" or secondary effect at the phase transition. The complexities and higher-order terms required in the free energies of real ferroelectrics (for which the order parameter is often multidimensional) have generated whole books (e.g., Toledano and Toledano, 1987). A particularly interesting category is that of modulated structures (whose periodic repeat distances are *incommensurate* with the lattice spacing A_0 of the normal phase); these require gradient terms in the free energies G.

2.4 Order–Disorder versus Displacive Systems

In a purely order–disorder system, ions sit in a double-well potential $V(r)$ which is totally independent of temperature T. As temperature rises, the expectation value of average ionic position $\langle r \rangle$ changes dramatically because of Bose population factors, moving from a position often centered on a bond (e.g., O–H–O) to one off center (O–H– –O) with a net dipole. Such an order–disorder model of ferroelectricity is highly attractive by virtue of the fact that the two sides of the double well may be labeled according to fictitious "pseudospin" states $+\frac{1}{2}$ and $-\frac{1}{2}$, and of the extremely complex spin formalism borrowed wholecloth from the analogous problems in ferromagnetism. Unfortunately, there seem to be no real ferroelectrics which closely approximate this spin theory.

In a displacive system, on the other hand, the shape of the double well $V(r,t)$ is explicitly dependent upon T, approximating that of the free energy discussed above. In this case, the ions also move from a centered value ($r=0$, $P=0$) to an asymmetric position, but the change in thermal population of energy levels is negligible; rather, it is the levels themselves that vary in energy with T. [This contrast is analogous to that of raising the energy of a system of molecules isothermally by doing work on it, e.g., compressing it (which changes the energy level spacing without modifying the population of each level), compared with simply heating it, by which the populations of each energy state change but the energy levels remain fixed.] Many real dielectrics approximate a perfectly displacive system. That is, most of the change in $\langle r_j \rangle$ comes from changes in the shape of $V(r,T)$, not from the changes in populations of energy levels within a T-independent potential V.

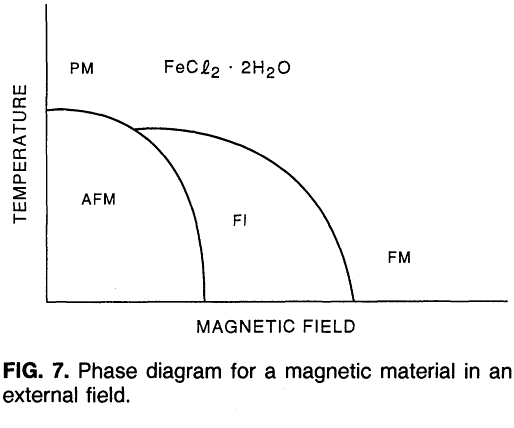

FIG. 7. Phase diagram for a magnetic material in an external field.

Of course, there are real crystals with dynamics intermediate between order–disorder and displacive, in which changes in $V(T)$ account for about half the change in $\langle r_j \rangle$. In some systems, one sublattice undergoes an order–disorder transition (e.g., O–H– –O in KH_2PO_4) which drives a displacive instability in a second sublattice (K^+ or P^+ in KH_2PO_4).

It was once considered that all hydrogen-bonded systems were order–disorder in character. That is untrue: asymmetric hydrogen bonds, such as those in triglycine sulfate, Rochelle salt, or tris-sarcosine calcium chloride (e.g., N–H– –Cl) do not generally produce order–disorder dynamics and are more accurately described as displacive.

An old rule of thumb is that the entropy change is large at an order–disorder transition ($\Delta S \sim 1$ cal/g·K) and small at a displacive transition ($\Delta S \sim 0.1$ cal/g·K), but the experimental values span a continuum without an obvious bimodal distribution. So there is reason to believe that most systems are partially displacive in their kinetics and partially disordered. Precise x-ray data show disorder phenomena very near T_C in even the most displacive of systems (e.g., $PbTiO_3$).

Displacive structural phase transitions are necessarily accompanied by a "soft" (i.e., unstable) optical phonon whose energy approaches zero as $T \rightarrow T_C$. As first outlined by Cochran (1960, 1961) for ferroelectrics, this instability may be described as a subtle cancellation between the long-range and short-range contributions to the frequency of a long-wavelength ($k=0$) transverse optical (TO) phonon:

$$\omega_{TO}^2(T) = R_0' - 4\pi(Ze)^2/V(T), \qquad (9)$$

where $V(T)$ is the unit cell volume, and R_0' is a short-range steric force constant not due to electric charge. [By comparison, for longitudinal (LO) modes the two contributions have the same sign, and so complete softening of $\omega_{LO} \rightarrow 0$ is impossible.] For order–disorder transitions, a soft optical phonon is unexpected. However, damping of one mode often diverges as $T \rightarrow T_C$, so that for $T \rightarrow T_C$ only a diffusional response is observed. In both cases, the peak of the imaginary part of the dielectric constant shifts to $\omega = 0$, but the details differ.

2.5 Ultraweak Ferroelectrics

There are three broad categories of ferroelectrics, classified according to the magnitude of their Curie constants $\mathscr{C}$. In the first category, $\mathscr{C} \approx 10^5$ K and $\mathscr{C}/T_C \approx 10^3$. It is exemplified by $KNbO_3$, $BaTiO_3$, $LiNbO_3$, and other oxides having no permanent dipoles in their paraelectric phases. The second category has $\mathscr{C} \approx 10^3$ K and $\mathscr{C}/T_C \approx 10$. In this group the early 1930 model of Kobeko and Kurchatov is a reasonable approximation; it views ferroelectricity as arising from a rigid rotationlike orientation of permanent dipoles (giving $\mathscr{C} = 3T_C$, as discussed in a section which follows). The third category has $\mathscr{C} < 10^2$ K and $\mathscr{C}/T_C < 1$. It is exemplified by TSCC, $(CH_3NHCH_2COOH)_3.CaCl_2$. In this system, Kozlov *et al.* (1983) were able to show that it is a displacive ferroelectric satisfying Cochran's equation [Eq. (9)], but the instability is in a short-range force constant R_0' in the shell-model description with negligibly small Coulombic terms. As a result, it has both soft transverse TO and longitudinal LO modes, although the TO necessarily reaches zero frequency as T approaches T_0 before the LO does.

2.6 Domain Structures

Magnetic domains are separated by domain walls that are thick ($\geqslant 10$ nm) and permit a continuous rotation either out of the plane of magnetization (Bloch walls) or, in thin films, in-plane (Néel walls). In contrast, ferroelectric domains are usually separated by very thin ($\leqslant 1$ nm) walls of straight edges (Fig. 8). The angles between these walls are restricted by the symmetry of the ferroelectric phase and may include needlelike, zig-zag, and polyhedral arrays (Fig. 9). Under some circumstances thought to be nonequilibrium, the ferroelectric domain boundaries may be quite wavy or "rough," as in lead germanate. These rough domains approximate spin-1/2 Ising model simulations.

As typical examples of differing domain patterns, one might cite TGS (tri-glycine sulfate) as a ferroelectric with only 180° domains (antiparallel polarizations) and barium titanate with both 90° (polarizations at right angles) and 180° domains. In the latter material the 90° domains are not truly at right angles, but lie at $\pm 36'$ from 90°, which permits their observation in crossed polarizers (and also causes considerable internal strain). These 90° domains form angles of 45° with each other. That eliminates (or at least minimizes) the creation of $\nabla \cdot \mathbf{P}$ nonzero space charge; and they form an angle of either 22.5° or 67.5° with the electrodes. In general, domain walls form in a way that permits continuity of strain (lattice deformation) in addition to continuity of the component of electric polarization **P** normal to the domain walls. This requires in piezoelectric Rochelle Salt a chevron structure having 2° differences in the local optical axis, which permits visual observation.

In very thin magnetic films with Néel walls, the critical switching field H_c varies with thickness d as $d^{-4/3}$ (Néel's law). A similar thickness dependence is found for electrical

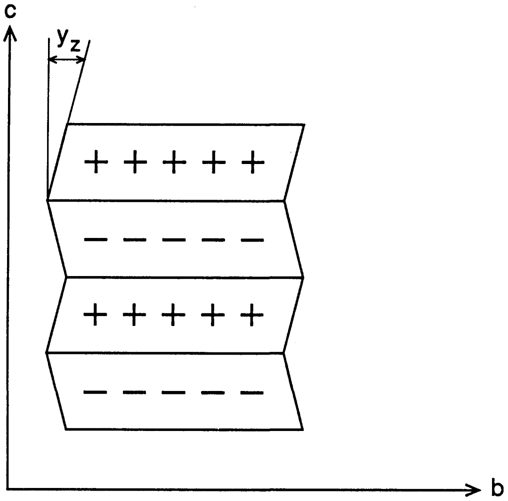

FIG. 8. Straight-line domain wall structures in Rochelle salt. From Kanzig (1957).

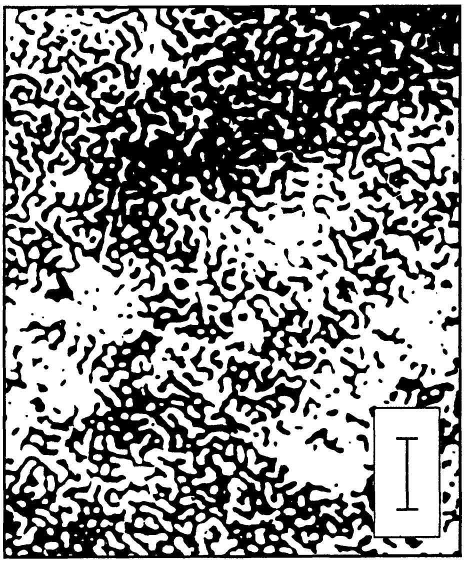

FIG. 9. More complex domain wall configurations in ferroelectrics. From Shur *et al.* (1989).

coercive field E_c versus thickness in dielectric films thinner than 100 nm, but there is no direct evidence that the domain walls differ in geometry for thin and thick ferroelectric films in a manner analogous to Bloch versus Néel walls in ferromagnets (Fig. 10); here the Bloch walls, which have spins rotating out of the surface of the material, are stable for thicker films, whereas the Néel walls, in which spins always lie in the plane of the film, are stable for thinner films.

3. FERROELASTICS

A ferroelastic crystal is one exhibiting hysteresis behavior in its stress–strain relationship, analogous to that between P and E in a ferroelectric. The symmetry classes that can exhibit ferroelasticity have been enumerated by Aizu in 1970. Not all ferroelastics are ferroelectric, or even polar: for example, the C_{2h} phase of LaP_5O_{14} is ferroelastic, and the $D_{2h} \leftrightarrow C_{2h}$ transition is termed a ferroelastic transition (because it is from a ferroelastic state to a nonferroelastic state). Ferroelastic states are bistable at zero applied stress. Not all ferroelectric states are ferroelastic. Ferroelectricity and ferroelastic states can coexist in complex ways; for example, $Ba_2NaNb_5O_{15}$ is both ferroelastic and ferroelectric between 543 and 580 K, and the ferroelastic domains are distinct from the ferroelectric domains, so that at least four physically distinct elastic-electric domains or states coexist.

4. ELECTRET EFFECTS

An electret is a material in which free charge, usually termed "space charge" (electrons or mobile ions), may be electrically driven to the surfaces by application of an electric field. Electrets exhibit many of the macroscopic characteristics usually identified with ferroelectrics—including hysteresis in their $P(E)$ relationship—but they are not ferroelectric: They possess no microscopic dipoles and exhibit no collective phenomena such as an ordering temperature.

The traditional electret is carnauba wax. An example of more current interest is polyvinylidene fluoride (PVF_2). Both materials are able to store charge at their surfaces corresponding to kV of potential differences between the front and back surfaces and to retain it for long times (weeks). Depending upon the charging techniques (whether by proximity or direct contact) and the nature of the electrode interface, the charge state that results can be the same as that of the contact electrode ("homocharge") or opposite ("heterocharge").

Because the polarization in an electret involves ion transport over microscopic distances, it is slow to reverse electrically compared to ferroelectrics (ms vs ns). However, in many real ferroelectrics, electretlike space charge effects may also be present, in addition to the lattice polarization. This greatly com-

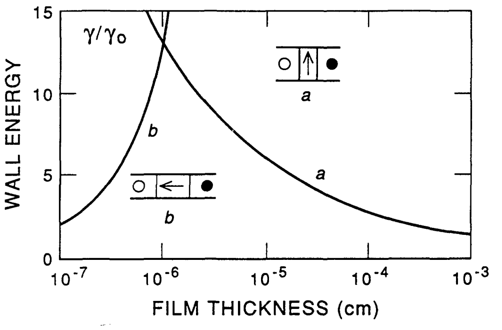

FIG. 10. Néel [curve (a)] and Bloch [curve (b)] wall energies in a ferromagnet versus film thickness. From Chikazumi (1985).

plicates the experimental evaluation of quantities such as remnant polarization P_r, spontaneous polarization P_S, coercive field E_C, and switching time t_S. One characteristic symptom of electret effects is the non-Ohmic current-voltage relationship (Langmuir–Child's law)—a quadratic $I=aV^2$ dependence is observed for space-charge limited current flow in an insulator. Electret effects may be created in insulators exposed to ionizing radiation.

5. ELECTROSTRICTION

Electrostriction is a phenomenon described by a quadratic relationship between strain and electric field. It is present in all solids, including glasses. By comparison, piezoelectricity is a linear relationship between strain and electric field. It is present in only 20 of the 32 point-group symmetries (crystal classes). Of these 20, only 10 have a unique polar axis. These 10 are pyroelectric, with a net polarization that changes with T. Pyroelectrics which can have their polarization reversed (switched) by application of an external electric field are called ferroelectrics. This distinction between pyroelectrics and ferroelectrics is thus purely practical and empirical; there are no symmetry differences between the two.

6. ELECTROCALORIC EFFECT

Various discussions of the electrocaloric effect exist in the literature, which seem to disagree qualitatively with each other. Fatuzzo and Merz (1967) state that the electrocaloric effect is the change in temperature for a ferroelectric crystal whose polarization changes rapidly, and that it occurs (for large fields) only *above* T_C, where $\Delta T=\beta TP^2\times(2C_p\rho)^{-1}$, and application of an electric field changes polarization P from zero to finite values. Below T_C they state that (large) electric fields change $+P$ to $-P$ which produces no change in ΔT. The maximum effect, they state, in $BaTiO_3$ is $\Delta T\approx 2$ K for $T\approx T_C+1$ K.

By comparison, Mitsui *et al.* (1976) claim that the electrocaloric effect can occur only below T_C, since $(\delta T)_{S,x}=(T/C)(\partial Pi/\partial T)_{x,E}$, where $P_i=0$ above T_C and $=P_S$ below T_C, since

$$(\delta T)_{S,x}=\left(-\frac{T}{C}\right)\left(\frac{\partial P_i}{\partial T}\right)_{x,E}(\delta E_i)_{\partial,x}. \tag{10}$$

Since $\partial P_i/\partial T<0$, heat $T\,dS$ must be taken out of the crystal in isothermal conditions if E is $\|P_S$.

Jona and Shirane (1962) disagree with Mitsui *et al.* and point out that ΔT in KDP is 1 K at $E=10$ kV/cm and T_C+1 K.

In any case, the electrocaloric effect arises because of the difference between isothermal and adiabatic dielectric constants. The correct interpretation is as follows.

In ferroelectrics with second-order phase transitions, applied fields cannot shift T_C or induce a polarization above T_C. Hence, in such systems (e.g., TGS), the electrocaloric effect vanishes above T_C, since $\partial P_i/\partial T=0$. Below T_C, $\Delta T=-\beta TP^2/\rho C_p$ is nonzero, provided the applied field does not induce switching from $-P$ to $+P$.

In ferroelectrics with first-order transitions, applied fields above T_0 (but less than T_2) cause large increases in P, from zero to values nearly those of P_S below T_0. In these systems (e.g., $BaTiO_3$, where $T_2=T_C+18$ K) the electrocaloric effect is largest *above* T_0 (1 K per 10 kV/cm) but below T_C and T_2.

7. PIEZOELECTRICITY

All ferroelectrics are necessarily pyroelectric (a change in polarization is generated by a temperature change ΔT) as well as piezoelectric. Piezoelectricity is defined as a change in electric polarization P with change in applied stress σ. The piezoelectric tensor is third-rank, since it couples a vector polarization with a second-rank stress tensor. The inverse effect, of an applied electric field producing a stress (and strain), is termed "electrostriction." Electrostriction necessarily contributes to the free-energy coefficient B in the Landau theory free energy discussed in Sec. 2.3. For example, in KNO_3, $B=-30$ (cgs) when electrostriction is included, whereas $B=-4$ (cgs) when electrostriction is ignored. Thus, the effect of electrostriction in KNO_3 is to make its phase transition far more discontinuous. Note that a quadratic electrostrictive term will be nonzero in all solids (including glasses), independent of their symmetry.

8. NUMERICAL VALUES

8.1 Spontaneous Polarization

The maximum value of P_S that might be possible in a solid is that in which a single electrical charge e is displaced by a bond length L in each unit cell. Assuming $L=1$ Å, this gives a dipole of 10^{-29} C m per unit cell volume. In crystals with small unit cells ($V \approx 3\times3\times3$ Å) such as ABO_3 perovskites, this produces a polarization of approximately $P_S = ed/V = (10^{-29})/(27\times10^{-30}) = 0.4\ \mathrm{C/m^2}$, in rather close agreement with the maximum values observed experimentally in perovskites such as $PbTiO_3$.

P_S can of course be much lower. In smectic-C^* ferroelectric liquid crystals and other materials with large and complex unit cells, the value of P_S is greatly reduced by the large cell volume V, which appears in the denominator in the equation above. In other systems the charge displacement d is very small, because the ferroelectric phase transition involves an eigenvector (often a rigid or nearly rigid rotation) that very slightly displaces net charge. Tris-sarcosine calcium chloride (TSCC) is probably the best example of such a weak ferroelectric; in it the rotation of hydrogen-bonded sarcosine molecules

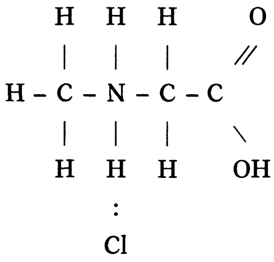

$$\begin{array}{ccccccccc} & & H & & H & & H & & O \\ & & | & & | & & | & & // \\ H & - & C & - & N & - & C & - & C \\ & & | & & | & & | & & \backslash \\ & & H & & H & & H & & OH \\ & & & & : & & & & \\ & & & & Cl & & & & \end{array}$$

yields $P_S = 2.7 \times 10^{-3}\ \mathrm{C/m^2}$, more than a thousand times weaker than in $KNbO_3$ or $PbTiO_3$.

A listing of P_S values for 20 well-studied ferroelectrics is given in Table 1, together with other pertinent parameters. In Table 1 the acronyms TGS, TGSe, and TGFB stand for triglycine sulfate, triglycine selenate, and triglycine fluoroberyllate, respectively. TSCC and MASD are tris-sarcosine calcium chloride and methyl ammonium sulfate dodecahydrate (alum).

8.2 Curie Constant

8.2.1 Quantum Cryogenic Effects. In analogy with ferromagnets, in which the magnetic susceptibility often obeys a Curie–Weiss law of form

$$\chi(T) = \chi_0 + [\mathscr{C}/(T-T_c)], \tag{11}$$

where $\mathscr{C}$ is called the Curie constant and T_c the Curie temperature (which may differ from the actual phase transition temperature T_0), dielectric solids often exhibit the same functional form for their dielectric constant,

$$\epsilon(T) = \epsilon_0 + [\mathscr{C}/(T-T_c)]. \tag{12}$$

If $T_c > 0$ the material is usually ferroelectric, with a spontaneous polarization over some finite temperature range. In some materials, such as $KTaO_3$ and $SrTiO_3$, this Curie–Weiss law is satisfied over a large temperature range but becomes invalid near T_c, which is near absolute zero, and no phase transition actually occurs. That is, a dielectric crystal may satisfy the equation above over a wide temperature range, with a positive value of Curie temperature T_c, but have no actual phase transition.

In the specific cases of pure $SrTiO_3$ or $KTaO_3$, the reason for this anomalous behavior is due to explicitly quantum effects. Equation (12) above can be derived microscopically from the anharmonic theory of lattice dynamics; see for example, Kozlov *et al.* (1983). In such a rigorous treatment a population factor $\bar{n}(T) = [e^{\hbar\omega/kT} - 1]^{-1}$ appears, which in the high-temperature approximation becomes $kT/\hbar\omega$. This term leads to the linear dependence of $\epsilon^{-1}(T)$ upon T in Eq. (12) above. However, it is impossible to make this approximation near $T=0$, and Eq. (12) fails. Such quantum ferroelectric phenomena have been studied carefully in crystals of the $SrTiO_3$ and $KTaO_3$ family, doped with substitutional ions that vary T_c, and in TSCC doped with bromine.

8.2.2 Calculating Curie Constants. In the Weiss theory of magnetism the Curie constant $\mathscr{C}$ in Eq. (11) describing the temperature dependence of the magnetic susceptibility in the paramagnetic phase is not an independent parameter but is given by

$$\mathscr{C} = Ng^2\mu_B^2 S(S+1)/3k_B, \tag{13}$$

Table 1. Experimental properties of some well-studied ferroelectrics.

Substance	Curie constant C (K)	Spontaneous polarization P_S (C/m^2)	Entropy change ΔS (J/mol·K)	Curie temperature T_C (K)	Lorentz factor $\beta = 4\pi T_C/C$	Cell size (g/mol)	Density (g/cc)	Lorentz factor $\beta' = 2T_c\Delta S/P_s^2$	$C\Delta S$ (J/mol)	$2\pi P^2$ (J/mol)	Lorentz factor ratio β'/β
BaTiO$_3$	1.73×10^8	0.26	0.60	393	0.029	233	6.1	0.020	1.04×10^5	1.46×10^5	0.70
KnbO$_3$	2.42×10^5	0.30	0.28	435	0.023	180	4.59	0.007	6.78×10^4	1.99×10^5	0.33
KNO$_3$	5.62×10^3	0.081	5.4	401	0.90	118	2.11	1.3	3.03×10^4	2.07×10^4	1.5
Cd$_2$Nb$_2$O$_7$	1.0×10^5	0.055	0.43	185	0.023	522	6.24	0.070	4.3×10^4	1.43×10^4	3.0
SbSI	2.33×10^5	0.25	0.87	295	0.016	177	5.19	0.027	2.03×10^5	1.20×10^5	1.7
BiSI	1.9×10^5	0.07	0.38	88	0.020	264	6.60	0.038	7.22×10^4	1.11×10^4	1.9
SbSBr	1.21×10^5	0.075	0.17	93	0.010	130	4.86	0.023	2.04×10^4	8.50×10^3	2.3
KH$_2$PO$_4$	3.1×10^3	0.05	4.01	123	0.50	136	2.34	0.75	1.24×10^4	8.21×10^3	1.5
KD$_2$PO$_4$	5.45×10^3	0.062	2.89	213	0.49	138	2.35	0.61	9.97×10^3	1.28×10^4	1.2
TGS	3.2×10^3	0.028	2.31	321	1.26	323	1.69	1.1	7.39×10^3	8.5×10^3	0.87
TGSe	4.05×10^3	0.030	2.38	295	0.91	380	1.85	0.84	9.64×10^3	1.04×10^4	0.93
TGFB	2.24×10^3	0.034	1.17	343	1.92	294	1.63	0.43	2.62×10^3	1.18×10^4	0.22
TSCC	5.8×10^1	0.0027	1.83	128	27.7	411	1.53	26.6	1.06×10^2	1.11×10^2	0.96
(NH$_4$)$_2$SO$_4$	2.9×10^1	0.0062	17.6	224	97.0	142	1.55	249	5.1×10^2	1.99×10^2	2.6
NH$_4$HSO$_4$	2.5×10^2	0.0080	7.8	270	13.6	115	1.73	110	1.95×10^3	2.40×10^2	8.1
NaH$_3$(SeO$_3$)$_2$	1.6×10^3	0.030	4.1	194	1.52	280	3.20	2.2	6.6×10^3	4.45×10^3	1.5
SC(NH$_2$)$_2$	3.7×10^3	0.035	0.167	169	0.57	76	1.41	0.95	6.2×10^2	3.73×10^3	0.17
Rochelle Salt	2.2×10^3	0.0025	0.034	296	1.69	282	1.83	2.23	7.48×10^4	5.44×10^4	1.4
NaNO$_2$	5.0×10^3	0.08	9.14	296	0.74	69	2.16	2.9	4.57×10^4	1.16×10^4	4.0
MASD	5.0×10^2	0.010	9.70	174	4.51	466	1.59	12.8	4.85×10^3	1.66×10^3	2.8

where N is the number of spins per unit volume of magnitude S; g, the gyromagnetic ratio; μ_B, the Bohr magneton; and k_B, Boltzmann's constant. In contrast to this situation, the Curie constant $\mathscr{C}$ appearing in Eq. (12) in the dielectric susceptibility of ferroelectric crystals in their paraelectric phases above T_c is more often regarded as an empirical parameter without a precise microscopic origin. From 1930 through 1955, however, a considerable effort was made to calculate $\mathscr{C}$ for real crystals by physical models. The first of such models was that of Kobeko and Kurchatov (1930), who used a model of rotating dipoles in which the polarizability α of each dipole p_i is given by $\alpha=p_i^2/3k_BT$. The internal field is then given by

$$E_r=E_a+(\beta/4\pi)P, \tag{14}$$

where E_a is the applied field, P is the polarization, and β is called the Lorentz factor. For an array of point dipoles in a simple cubic (or random) lattice, $\beta=4\pi/3$ and $\mathscr{C}=3T_c$ exactly.

A few ferroelectrics empirically satisfy these predictions. Methylammonium aluminum sulfate dodecahydrate alum (MASD) has $\beta=4.51 \approx 4\pi/3$ and $\mathscr{C}=2.9T_c$, in excellent agreement with predictions. TSCC and Rochelle salt are also rather close, suggesting that a rigid-rotation model of localized dipoles is a fair description of their phase-transition kinetics; however, many other materials disagree with these constraints by many orders of magnitude (see Table 1).

Typically, the ABO_3 oxides have Curie constants $\mathscr{C} \approx 10^5\,K \approx 10^3T_c$. These "strong" ferroelectrics do not seem amenable to physical models involving permanent dipoles. Rather, their behavior involves ionic displacements that are created below T_c. Thus, they may often be characterized primarily as "displacive" rather than "order–disorder" systems. Such simple labels may be helpful in generic labeling, but they are misleading in specific cases. Whereas ABO_3 perovskites are probably displacive in their dynamics as a group, $BaTiO_3$ is more accurately described as order–disorder. And among hydrogen-bonded systems (often generically viewed as order–disorder), TSCC is quite displacive, with underdamped "soft" optical photons in both paraelectric and ferroelectric phases—despite its numerical agreement for $\mathscr{C}$ and β with Kobeko–Kurchatov theory mentioned above. Most real crystals possess some displacive characteristics and some order–disorder properties in their behavior near transitions.

8.2.3 Numerical Self-Consistency of Mean-Field Theory. The Landau theory described in Sec. 2.3 has a few empirical parameters. These parameters are overdetermined for most dielectric materials by the available experiments; thus, a check on the self-consistency of the theory is usually possible. One such internal check is that the entropy change at T_c (or T_0) times the Curie constant $\mathscr{C}$ should equal 2π times the square of the spontaneous polarization:

$$\mathscr{C}\Delta s=2\pi P_S^2. \tag{15}$$

In practice, as shown in Table 1, this relationship is satisfied only within factors of 2 or 3. [We compare in Table 1 the Lorentz factor $\beta=4\pi T_c/\mathscr{C}$ with another factor $\beta'=2T_c\Delta S/P_S^2$ that should equal β of Eq. (14) if Eq. (15) is true.] The reason is that the coefficients (e.g., γ and δ) appearing in the free-energy expression of Eq. (15) are assumed in the simplest theory to be temperature independent. In reality, they all vary continuously with temperature (although they are not singular at T_c), so that Eq. (15) above is only an approximation, always good to an order of magnitude, but never exact.

8.3 Dielectric Constants

As discussed in Sec. 8.2.1, many dielectric materials exhibit a Curie-law dependence for their dielectric constants:

$$\epsilon(T)=4\pi\mathscr{C}/(T-T_C)+\epsilon_0. \tag{16}$$

This Curie-law dependence can be derived from a simple model of the dielectric function for the material. Noting that the frequencies of longitudinal and transverse waves in any material must respectively lie at the zeroes and poles of the dielectric constant (required by Maxwell's equations), and that the dielectric function is usually modeled as

$$\epsilon(\omega)=\epsilon_\infty+(\epsilon_0-\epsilon_\infty)/({\omega_{TO}}^2-\omega^2), \tag{17}$$

one obtains the relationship

$$\epsilon_0(T)/\epsilon_\infty=\omega_{LO}^2/\omega_{TO}^2(T), \tag{18}$$

where LO and TO designate longitudinal optic and transverse optic phonon frequencies. In this equation, the transverse optic phonon frequency is often found to be strongly temperature dependent, but the longitudinal mode usually nearly temperature independent. The equation is referred to as the Lyddane–Sachs–Teller relationship (Lyddane *et al.*, 1941), and the temperature-dependent transverse optical phonon branch termed the "soft mode." In Cochran's original soft-mode theory this temperature dependence is explicit, since

$$\omega^2(T)=R_0-4\pi Z^2e^2/3V(T) \tag{19}$$

for the transverse mode, where $V(T)$ is the unit cell volume, permits a subtle cancellation of long-range Coulombic forces (Ze is the effective charge) and short-range forces R_0; whereas for longitudinal modes,

$$\omega_{\mathrm{LO}}^2(T)=R_0+8\pi Z^2e^2/3\epsilon_\infty V(T) \tag{20}$$

is very weakly temperature dependent unless R_0 is extremely small (as in TSCC and one or two other ferroelectrics). Note that the longitudinal phonon frequency above is always nonzero at all temperatures, but the transverse mode can vanish by accidental cancellation of two terms.

A fuller derivation of related equations for cases of many optical modes and phonon damping has been given by Kurosawa (1961) and by Cowley (1963).

9. PREPARATION OF DIELECTRIC THIN FILMS

9.1 Sputtering

Dielectric thin films are often prepared by rf or magnetron sputtering, using either metal or ceramic targets. ABO_3 perovskites deposited in this way are of good stoichiometry and flatness over a full 4- or 5-in. substrate wafer. They can be prepared pinhole-free in thickness from 40 to 1000 nm.

9.2 Sol-Gel

Solution-gelation wet-chemistry techniques using solutions such as barium or lead acetate and zirconium and titanium propoxides can be used to prepare ABO_3 dielectric perovskites. Usually mixed in a common solvent and then hydrolyzed, these solutions gel to yield fine-grained ferroelectric ceramics (40–100 nm) of exceptional quality ($P_S \approx 50\%$ of single-crystal values) when annealed at ca. 850 K. They can be spin-cast at 1500–2000 rpm to give films of superior stoichiometry and uniform density in comparison with those obtained by sputtering.

9.3 Slip-Casting, Thermal Evaporation, and Electron-Beam Evaporation

Other techniques of dielectric thin-film deposition include slip-casting (melting of powder on a metal substrate), thermal evaporation, and electron-beam co-evaporation. Slip-casting and thermal evaporation have both been used for high-quality KNO_3 films; electron-beam coevaporation of BaF_2 and MgF_2 yields high-quality $BaMgF_4$ ferroelectric films.

9.4 MOCVD

Metal-oxide chemical vapor deposition [see CVD (CHEMICAL VAPOR DEPOSITION)] is superior to both sputtering (*q.v.*) and sol-gel spin-on techniques for commercial processing of high-quality dielectric films. Sputtering suffers from target aging (for ceramic targets) and poor defect concentrations. Sol-gel spin-on techniques give poor step coverage for microelectronics applications. A deposition technique superior to both for commercial device processing is MOCVD. In lead-based dielectrics the Pb is often added as tetraethyl lead. More recently liquid-phase room-temperature CVD using sol-gel as the injected material has permitted replacement of this highly toxic tetraethyl lead. uv is often used to break the organic bonds.

9.5 MBE

The highest quality dielectric films can be prepared by molecular beam epitaxy (*q.v.*). This technique is expensive and requires high vacuum. It was developed for semiconductor technology but has recently been applied to dielectrics, including ferroelectric $BaMgF_4$ thin films.

9.6 Continuous Methods

For organic ferroelectrics, continuous sheet production is possible. TGS and PVF_2 (polyvinylidene fluoride) are prepared as continuous sheet and then poled (poling is the process of orienting domains, initially random, by applying an external electric field) at very high fields (MV/cm for PVF_2). Applications include large-area pyroelectric detectors.

9.7 Ferroelectric Liquid Crystals

Certain classes of liquid crystals (the smectic-C* chiral structures) are truly ferroelectric. They may be prepared in oriented thin films in which proprietary surface treatments unwind the cholesteric structures and permit high-speed switching devices to be fabricated. The production of such organic ferroelectric films is in general complex.

9.8 New Applications of High Dielectric Films

The primary recent applications of high dielectric thin films are in

1. capacitors for DRAMs (dynamic random access memories);
2. ferroelectric switching devices (including RAMs);
3. pyroelectric detectors; and
4. integrating capacitors in CCDs (charge coupled devices).

9.8.1 DRAMs. The development of high-density (e.g. 64-Mbit) submicron memory devices has generated new interest in very high-dielectric films to minimize the trenching that would otherwise be required to achieve the requisite capacitance. By replacing SiO_2 with a ferroelectric such as $PbTiO_3$ or a high-dielectric paraelectric such as $Ba_xSr_{1-x}TiO_3$, one can obtain an increase of ×100 in dielectric constant and hence capacitance. At present, several large Japanese corporations favor $Ba_xSr_{1-x}TiO_3$ films for this application, whereas several US corporations are developing $PbZr_{1-x}TiO_3$ films for the same DRAM applications. Note that for this use the film never switches polarization (all the applied voltage pulses have the same polarity), so fatigue (decrease in P_S with operation) is not a problem. However, other aging processes, such as dc breakdown, remain as important failure mechanisms.

9.8.2 Ferroelectric Memories. Thin ferroelectric films may also be used in memory applications involving polarization switching. Three kinds of ferroelectric memories exist in prototype form at this time, but none is in commercial use yet.

The first is a switched capacitor array in which the ferroelectric films (usually 2×2 μm^2 in cross-sectional area and 0.25 μm thick) store the "1" or "0" in the binary Boolean algebra of the computer as $+P_r$ or $-P_r$ polarization states. The memory is read (destructively) by switching the device and comparing the current transient $i(t)$ for the cell in question ($+P_r$ or $-P_r$) with the $i(t)$ from a reference cell that is always in the $-P_r$ state. For this reason it is important to understand theoretically the detailed kinetics of $i(t)$ switching currents. At present, prototype 16-kbit ferroelectric RAM memories have been fabricated by several companies, having 5-V standard silicon VLSI logic levels (or 3 V for GaAs) and very high speeds (10 ns read/write). Their extreme radiation hardness makes them attractive for aerospace applications as electrically erasable programmable read-only memory (EEPROM) or magnetic bubble memory replacement. $PbZr_{1-x}Ti_xO_3$ is favored at present as the cell material.

The second type of ferroelectric memory is the true ferroelectric FET (field-effect transistor). In this device the ferroelectric film is deposited in the gate region of the FET. Depending upon its polarization ($+P_r$ or $-P_r$), the source-to-drain current is altered. This current level can be read without switching (nondestructive read-out, NDRO), a distinct advantage. At present NDRO FETs using this design are being made at Westinghouse with $BaMgF_4$ as the ferroelectric film; the original ferroelectric FETS were made in Osaka in the 1970s with $PbZr_{1-x}Ti_xO_3$.

9.8.3 Optoelectronics. The third generic type of ferroelectric memory of interest is the optically read array. The polarization state of a ferroelectric cell can be read by means of a laser, using beam deflection, phase shift, or other mechanisms. In 1976 Westinghouse developed a 1-kb bismuth titanate memory ($Bi_4Ti_3O_{12}$) that could be read by beam deflection.

The primary disadvantage is that the device must be thick (ca. 100 μm) enough to produce significant beam deflection. This in turn necessitates large operating voltages (>100 V), so that the resulting device cannot be integrated into standard silicon integrated circuits, which operate at 5 V.

9.8.4 Pyroelectrics. Pyroelectric detectors can be fabricated from either bulk or films. They can achieve subnanosecond response times. The pyroelectric coefficient is proportional to the temperature derivative of the spontaneous polarization,

$$p=k\frac{dP}{dT}. \tag{21}$$

This coefficient is large near phase transition temperatures, but in general pyroelectric detectors must operate over a wide temperature range, so that divergences in dP/dT are undesirable.

At present triglycine sulfate (TGS) is the most commonly used pyroelectric. Its figure of merit for this application is not the largest known, but it can be grown cheaply from aqueous solutions in large-area specimens. Other cheaper pyroelectric systems include poled PVF_2 (polyvinylidene fluoride), which has relatively poor figures of merit but can be fabricated in continuous sheet. It is used for inexpensive detectors, such as "intruder alarms." Sol-gel or MOD (metal-organic deposition) KTN ($KTa_{1-x}Nb_xO_3$) is also a good pyroelectric.

10. NONPOLAR DIELECTRICS AND DIELECTRIC BREAKDOWN

Although most of this article is devoted to the dielectric properties of polar insulators, including ferroelectrics, considerable interest exists in the dielectric response of nonpolar materials, including both dielectric liquids (and glasses) and crystals.

As pointed out by O'Dwyer (1964) in his classic text on the subject, all materials conduct electricity to some degree, and all suffer breakdown in sufficiently high fields. However, only "intrinsic breakdown" in dielectric solids is qualitatively unique as a failure mechanism; by contrast, "thermal breakdown" in insulators is not qualitatively different from melting in metals at high currents, and "avalanche breakdown" in solids differs only quantitatively from spark discharge in gases.

In general, these three breakdown mechanisms are important in both polar and nonpolar insulators. They are particularly well understood in alkali halides. We note parenthetically that dielectric breakdown in liquids is much less well understood.

Commercial applications of insulating layers to prevent electrical shorting are especially important at present in the microelectronics industry. For devices such as random-access memories the dielectric materials used are primarily silicon nitride (Si_3N_4), silicon dioxide (SiO_2), and tantalum oxide (Ta_2O_5).

10.1 Intrinsic Breakdown in Insulators

Intrinsic breakdown is a purely electronic process in which the rise in lattice (phonon) temperature is negligible (this, in turn, requires that the ratio of electrical to thermal conductivity is very low).

The principal characteristics of intrinsic breakdown in insulators are as follows.

1. It occurs at low temperatures (i.e., 300 K or below).
2. The breakdown field strength E_B is independent of specimen size or shape and independent of electrode material and shape.
3. The breakdown occurs quickly (μs or less).

Intrinsic breakdown is observed in alkali halides, mica, glass, and most polymers. It was first analyzed by Zener in 1934 on the basis of quantum mechanical tunneling of electrons from the valence band to the conduction band.

10.2 dc Thermal Breakdown

The principal characteristics of thermal breakdown are these:

1. It occurs at high temperatures.
2. The breakdown field E_B depends upon specimen size and shape and on the geometry and thermal properties of the electrodes.
3. Breakdown takes time (ms or longer).
4. ac values of E_B are lower than dc and decrease with increasing frequency.

In this situation the critical electric field E_B (min) is given by

$$-\nabla\cdot(K\nabla T)=\sigma E_B^2(\text{min}), \tag{22}$$

where K is the thermal conductivity tensor.

10.3 Impulse Thermal Breakdown

This is the opposite extreme of dc thermal breakdown. It occurs when the field is applied very quickly, as in a switched capacitor (e.g., ferroelectric memory). Its distinguishing characteristics are the following.

1. It occurs at high temperatures.
2. Breakdown field E_B is fairly independent of sample size and shape.
3. E_B varies greatly with length of time field is applied (i.e., duty cycle in a memory), always decreasing for long times of applied field.

This mechanism is equivalent to what in recent literature has been termed "time-dependent dc breakdown" in ferroelectric memories. In this case the critical field $E_B(\mathrm{min})$ is given by

$$C_v \frac{dT}{dt} = \sigma E_B^2(\mathrm{min}). \tag{23}$$

This process is favored in materials with high electrical conductivity and low thermal conductivity.

10.4 Avalanche Breakdown

Avalanche breakdown is observed in very thin materials, especially mica. It occurs

1. at low temperatures,
2. in thin slabs of dielectrics with very low electrical conductivity and very high E_B,
3. if the applied voltage is ramped up slowly, the current response is very noisy. If a high voltage is applied suddenly, the time required for breakdown has a large statistical variation.

This process physically involves a single electron leaving the cathode and exciting another conduction electron by collision ionization. Avalanche breakdown is a kind of intermediate process between intrinsic breakdown and thermal breakdown inasmuch as it is electronically initiated but has subsequent thermal steps.

10.5 Mathematical Models of Intrinsic Breakdown

10.5.1 Polar Materials. Theories describing intrinsic breakdown in solids are based either upon the single-electron approximation, in which the electron delivers only a small percentage of its total kinetic energy in each collision [the "high-energy" criterion of Fröhlich (1937)], or upon collective excitation models. With the former the result for E_B is

$$E_B = \frac{\pi^2}{2}\frac{e^3}{aIM}\left(m\hbar \ln\frac{\gamma}{\omega_t}\right)^{1/2}\left(1+\frac{2}{e^{\hbar\omega_t/kt}-1}\right)^{1/2}, \tag{24}$$

where

$$\gamma = \frac{\sqrt{2}\pi}{a\omega_t}\left(\frac{I\epsilon_\infty}{m\epsilon_0}\right)^{1/2};$$

ω_t is the transverse optical phonon frequency of long wavelength; ϵ_∞ and ϵ_0 are the high- and low-frequency dielectric constants; a is the nearest neighbor distance; I is the ionization energy; M is the reduced mass of the ions (assuming a two-atom cell, as in alkali halides; and e and m are the charge and (effective) mass of the electron.

10.5.2 Nonpolar Materials. Dielectric breakdown theories for nonpolar bodies were developed on the basis of the deformable-atom hypothesis of Bloch. They involve the solving of the Boltzmann transport equation with an applied field and collisions with both conduction electrons and phonons. Ionization of valence electrons and diffusion are key elements of the model. The specific models consist of neglecting certain of these terms in the general Boltzmann equation, which in its most general form is intractable.

10.5.3 Collective Breakdown. All theories based on the assumption of low conduction-electron densities lead essentially to avalanche breakdown, to be discussed below. The high-density theories describe true intrinsic breakdown. For them a collective electron model is most appropriate.

For polar crystals the result is

$$E_B = f\left(\frac{T_0}{\theta}\right)\left(\frac{me\omega_t}{\hbar}\right)\left(\frac{\epsilon_0}{\epsilon_\infty}\right)^{1/2}(\epsilon_\infty^{-1}-\epsilon_0^{1}), \tag{25}$$

where T_0/θ is a reduced temperature.

Application to non-polar crystals generally involves an unknown interaction constant, which, however, can be approximated by the electron mobility. The low-temperature result is

$$E_B = 0.14\hbar N^{1/2}[m\mu(T)]^{-1}, \tag{26}$$

where N is the number of atoms per unit volume.

10.5.4 Glasses/Amorphous Dielectrics. Amorphous dielectrics are defined as those with a large number of electron states within the bandgap. The breakdown field in this case is given by

$$E_{\mathrm{B}}=\left(\frac{\hbar\omega_t}{2.3D\Delta V}\right)^{1/2} e^{\Delta V/2KT_0}, \tag{27}$$

where ΔV is the binding energy of shallow donors; T_0, the lattice temperature (assumed slightly lower than T_e, the electron temperature); and D, a constant given by

$$D=e^2\tau_c\tau_s\gamma/m\hbar\omega n(t_0), \tag{28}$$

where τ_c and τ_s are the relaxation times for conduction electrons and shallow donors, and n is the conduction-electron population density.

10.6 Models of Thermal Breakdown

Steady-state thermal breakdown occurs when the general equation

$$C_v\frac{dT}{dt}-\nabla\cdot(K\nabla T)=\sigma E_{\mathrm{B}}^2 \tag{29}$$

can be solved by neglecting the first term in Eq. (29) above. It is necessary to consider the realistic case where the electrodes are not perfect thermal insulators; otherwise breakdown occurs at any finite applied voltage.

In the one-dimensional case, the solutions of Eq. (29) yield a maximum breakdown voltage V_{B} (not field E_{B}) given by

$$V_{\mathrm{B}}^2\cong\gamma\int_{T_0}^{\infty}\frac{K}{\sigma}\,dT, \tag{30}$$

which for the case of exponential electrical conduction,

$$\sigma=\sigma_0e^{-\phi/kT}, \tag{31}$$

yields

$$V_{\mathrm{B}}^2\cong\left(\frac{\gamma\kappa KT_0^2}{\sigma_0\phi}\right)^{1/2} e^{\phi/2KT_0}. \tag{32}$$

In the impulse approximation appropriate to fast switching, the result of neglecting the $\nabla\cdot(K\nabla T)$ term in Eq. (27) is

$$E_{\mathrm{B}}=\left(\frac{3C_vKT_0^2}{\sigma\phi t_c}\right)^{1/2} e^{\phi/2KT_0}, \tag{33}$$

where t_c is the rise time of the applied field (assumed linear).

10.7 Models of Avalanche Breakdown

Avalanche can arise by field emission of electrons from the valence band to the conduction band (Zener, 1934) or it can arise by collision ionization of electrons in the valence band. In the former case Zener obtained a result related to Fowler–Nordheim emission:

$$P=(eEa/\hbar)\exp(-\pi^2maI^2/h^2eE) \tag{34}$$

for the probability of valence-conduction band transition per unit time. This results in

$$E_c\sim10^9\ \mathrm{V/m} \tag{35}$$

over a wide range of rise times [E_c is proportional to $\ln^{-1}(At_c^2)$, where t_c is the rise time; this is a rather weak dependence on rise time].

In the latter case (collision ionization), avalanche results at 10^8 V/m in 1-μs rise time and produces 10^{12} electrons (40 generations). Typically this occurs at a lower threshold E_{B} than does Zener avalanche.

10.8 Breakdown in Specific Materials

Alkali halides typically break down at 600 MV/m, with a maximum factor of about 2.0 between their dc breakdown fields (lower) and 10^6-s^{-1} fields (higher). The breakdown is anisotropic; that is, E_{B} is lower along certain directions, e.g., [110]. The breakdown mechanism is intrinsic (or avalanche) below 300 K, thermal above 300 K.

Soda-lime glass and silica glass (fused quartz) exhibit intrinsic or avalanche mechanisms for their breakdown. [This is true of crystalline quartz, also, up to 300 K.] For the glasses, the breakdown is intrinsic up to 300 K and avalanche above ambient temperatures.

Mica has intrinsic (or avalanche) failure up to 400 K and thermal breakdown at higher temperatures; $E_{\mathrm{B}}\approx1.2\times10^2$ V/m.

Most polymers (polymethyl methacrylate, polyethylene, polystyrene, polyisobutylene) are avalanche up to 230 K, and some up to 400 K. The breakdown field E_{B} is typically proportional to the square root of Young's modulus for the polymer.

GLOSSARY

Antiferroelectric: A crystal structure having sublattices polarized such that the net polarization is zero.

Coercive Field: The minimum electric field required to switch the electric polarization in a ferroelectric crystal.

Electret: A substance (not necessarily ferroelectric) in which an electric polarization can be induced and switched by application of an electric field that causes charged-ion transport to the surface(s). Electrets do not produce polarization via switching of lattice polarization (unlike ferroelectrics), but have occasionally been used for similar applications. They include carnauba wax, Teflon, and other noncrystalline materials.

Electrocaloric Effect: Heating of a ferroelectric via application of an electric field.

Electrostriction: The change in linear dimension for a substance under application of an external electric field.

Ferroelastic: A crystal having a hysteretic relationship between uniaxial stress and strain.

Ferroelectric: A crystal having a hysteretic relationship between lattice polarization and applied electric field.

Hysteresis: Nonlinear relationship between electric polarization P (and displacement vector D) and applied electric field E such that two stable states of ("remnant") polarization exist at zero applied field and the polarization $P(E)$ at applied fields depends upon the initial value.

Paraelectric: A crystal having no electric polarization in zero applied electric field.

Pyroelectric: A crystal of point group symmetry C_n, C_{nv}, or C_{1h}, in which a change in temperature produces a change in electric polarization.

PZT: A mixed compound of adjustable percentages of lead zirconate and lead titanate widely used in ferroelectric devices in ceramic form.

Remnant Polarization: The value of electric polarization in zero applied field for a ferroelectric.

Spontaneous Polarization: The value of electric polarization in a ferroelectric that is linearly extrapolated from $P(E)$ values near saturation, at large electric fields E; P_S is usually slightly greater than the remnant polarization and represents the maximum achievable polarization in thermal equilibrium if fields within the specimen were uniform.

List of Works Cited

Aizu, K. (1970), *Phys. Rev.* **B7**, 754–757.

Chikazumi, S. (1985), *Physics of Magnetism*, New York: Wiley.

Cochran, W. (1960), *Adv. Phys.* **9**, 387–396 (1960); (1961) **10**, 401–412.

Cowley, R. A. (1963), in R. F. Wallis (Ed.), *Lattice Dynamics*, New York: Pergamon.

Fatuzzo, E., Merz, W. J. (1967), *Ferroelectricity*, Amsterdam: North-Holland.

Feynman, R. P. (1965), *Lectures on Physics*, Vol. 3, Reading, MA: Addison-Wesley.

Ishibashi, Y. (1986), *Jpn J. Appl. Phys.* **24** (Suppl. 2), 126–131.

Jona, F., Shirane, G. (1962), *Ferroelectric Crystals*, New York: Macmillan.

Kobeko, P., Kurchatov, I. (1930), *Z. Phys.* **66**, 192–197.

Kozlov, G. V., Volkov, A. A., Scott, J. F., Feldkamp, F. E., Petzelt, J. (1983), *Phys. Rev.* **B28**, 255–263.

Kurosawa, T. (1961), *J. Phys. Soc. Jpn* **16**, 1298–1303.

Landau, L. D. (1937), *Phys. Z. Sowjetunion* **11**, 26.

Lyddane, R. H., Sachs, R. G., Teller, E. (1941), *Phys. Rev.* **59**, 673–676.

Mitsui, T., Tatsuzaki, I., Nakamura, E. (1976) *An Introduction to the Physics of Ferroelectricity*, London: Gordon & Breach.

O'Dwyer, J. J. (1964), *The Theory of Dielectric Breakdown of Solids*, Oxford: Clarendon Press.

Onsager, L. (1936), *J. Am. Chem. Soc.* **58**, 1486–1492.

Shur, V. Ya., Gruverman, A. L., Letuchev, V. V., Rumyantsev, E. L., Subbotin, A. L. (1989), *Ferroelectrics* **98**, 29–49.

Slater, J. C. (1950), *Phys. Rev.* **78**, 748–753.

Toledano, J.-C., Toledano, P. (1987), *Landau Theory of Phase Transitions*, Singapore: World-Scientific.

Zener, C. (1934), *Proc. R. Soc. A* **145**, 523.

Further Reading

Bruce, A. D., Cowley, R. A. (1981), *Structural Phase Transitions*, London: Taylor and Francis, Ltd. Modern microscopic theory of ferroelectrics, including renormalization group theory.

Burfoot, J. C. (1967), *Ferroelectrics*, London: Van Nostrand. Standard Engineering reference.

Fatuzzo, E., Merz, W. J. (1967), *Ferroelectricity*, Amsterdam: North-Holland. The standard textbook on ferroelectricity over the past two decades.

Jona, F., Shirane, G. (1962), *Ferroelectric Crystals*, New York: Macmillan. A good text, particularly in its discussion of early work, including that of Fröhlich, Onsager, and Slater.

Kanzig, W. (1957), *Ferroelectrics and Antiferroelectrics*, New York: Academic. Standard text prior to 1960.

Lines, M. E., Glass, A. M. (1977), *Principles and Applications of Ferroelectrics and Related Materials*, Oxford: Clarendon. A comprehensive text including theory and experiment, as well as device applications of ferroelectrics and pyroelectrics.

Mitsui, T., Tatsuzaki, I., Nakamura E. (1976), *An Introduction to the Physics of Ferroelectricity*, London: Gordon & Breach. A more up-to-date text on ferroelectricity, emphasizing statistical mechanics and microscopic models, including soft-phonon modes.

O'Dwyer, J. J. (1964), *The Theory of Dielectric Breakdown of Solids*, Oxford: Clarendon. The standard text on dielectric and thermal breakdown.

Scott, J. F., Araujo, C. A. (1989), *Ferroelectric Memories*, *Science* **246**, 1400–1405. An overview of modern applications of ferroelectric thin films.

Stanley, H. E. (1976) *Introduction to Phase Transitions and Critical Phenomena*, Oxford: Clarendon. A good text on ferroelectric and ferromagnetic phase transitions, and their relationship to liquid–gas systems.

DIESEL ENGINES

CHARLES D. WOOD, *Southwest Research Institute, San Antonio, Texas, U. S. A.*

Introduction 1
1. **Principles of Operation** 1
1.1 Operating Cycle 1
1.2 Features of Combustion 2
1.3 Turbocharging 4
2. **Applications** 5
3. **Design Considerations** 8
3.1 Fuel Injection Systems 8
3.2 Exhaust Emission Control 10
3.3 New Materials 13
Glossary 14
Further Reading 11

INTRODUCTION

The diesel engine, invented in Germany towards the end of the 19th century, is now manufactured and used throughout the world. Its major advantage over other types of power plants is higher efficiency in converting fuel energy to output power. High efficiency in the diesel engine results from a combustion process characterized by spontaneous ignition of fuel injected into the combustion chamber near the end of the compression stroke, nonhomogeneous fuel-air mixtures containing more oxygen than required for complete combustion, an unthrottled flow of air into the engine, and a high compression ratio.

In comparison to the spark-ignition engine, the diesel engine, while significantly more efficient, has a heavier structure, produces more noise, and is more costly. Diesel engine combustion, because of its nonhomogeneous fuel-air mixtures, also produces soot in addition to the gaseous products of combustion.

Since the rate of fuel combustion is controlled by fuel injection rate and the rate of mixing between fuel and air, the diesel engine can be highly turbocharged without significant effects on the rate of combustion. By this means, output power can be significantly increased above that obtained using combustion air at ambient pressure.

These characteristics have led the diesel engine to be used in a wide range of applications. Diesel engines are built in sizes from about 0.5 L/cylinder up to about 1300 L/cylinder. Applications include stationary engines, on-highway vehicle engines, military and civil off-highway engines, locomotive engines, and large marine engines.

Increasingly severe environmental regulations are causing changes in diesel engine design to reduce the quantity of undesirable exhaust emissions. Many of these changes involve the fuel injection system. However, the complete combustion process is under close scrutiny despite difficulties in analysis due to the complexity of the mixing process between fuel and air as well as the subsequent combustion. Since the major attribute of the diesel engine is high efficiency, emission controls must not cause significant depreciation of efficiency; otherwise the major advantage of the diesel engine is lost. A range of potential solutions is available that seem likely to satisfy both environmental and economic goals.

1. PRINCIPLES OF OPERATION

1.1 Operating Cycle

The diesel engine cycle is made up of the following events: First, air is compressed by a piston in a cylinder through a compression ratio of between 14 and 22 (depending on engine size and design). Fuel is introduced into the air by means of a high-pressure fuel injection system when the piston is near the end of its compression stroke, and the fuel ignites spontaneously because of the high air temperature. Heat is added to the air by the exothermic reactions between the fuel and the oxygen contained in the air. Next, the hot, high-temperature products of combustion are expanded to produce work, and the products

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

of combustion are then exhausted to the atmosphere. Finally, the cylinder is filled with a fresh charge of air in preparation for the next cycle.

Diesel engines are built using either two or four piston strokes to accomplish this cycle. In the four-stroke cycle the first (upward) stroke is one of compression. Fuel injection, ignition, and combustion take place with the piston near the top of its stroke. Expansion occurs during the second (downward) stroke. Exhaust valve(s) in the cylinder head are then opened, and the third (upward) stroke displaces the exhaust products out of the engine. Exhaust valves are then closed, intake valves are opened, and the fourth (downward) stroke draws a fresh air charge into the cylinder.

In the two-stroke cycle, compression, injection, ignition, and combustion and expansion occur as in the four-stroke cycle. Near the end of the expansion stroke, the piston uncovers exhaust ports in the cylinder wall (or alternatively, exhaust valves in the cylinder head are opened), and exhaust products are pushed from the cylinder by the residual cylinder pressure. With further downward movement, the piston uncovers another set of (inlet) ports, through which fresh air is pushed by an external air compressor. Additional exhaust products are displaced from the cylinder by the incoming air. After the piston reverses direction and all ports (or valves) are closed, the cylinder contains fresh air as well as some 15%–30% of residual combustion products that escape the scavenging process.

The diesel cycle, as visualized by its inventor Rudolph Diesel (1858–1913), employed a combustion event in which the cylinder pressure was held constant by control of the rate of fuel injection and hence the rate at which heat was added to the air. In practice, combustion rates are tailored for optimum efficiency and minimum exhaust emissions, rather than to meet the constant-pressure criterion.

1.2 Features of Combustion

The principal characteristics of the diesel engine are dictated by its unique ignition and combustion process. The fuel is introduced into the combustion chamber under high pressure in the form of jets issuing through several orifices in the tip of the injector nozzle (see Fig. 1). Each conical fuel jet immediately

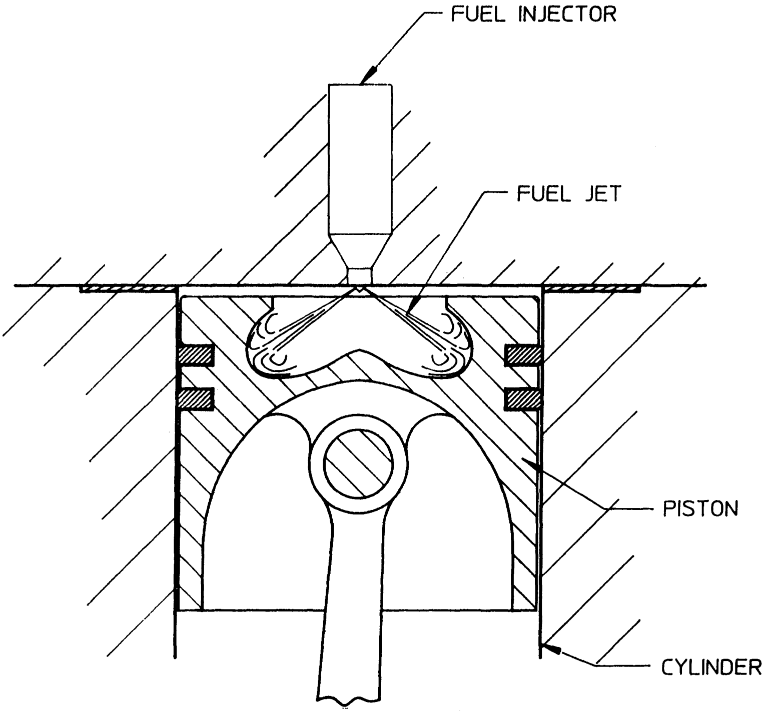

FIG. 1. Fuel injection in a direct-injection diesel engine combustion chamber.

begins to entrain air from the combustion chamber and, as a result, the fuel-air ratio within the jet decreases along the length of the jet as well as from the jet center line towards the outer surface of the jet. Ignition occurs spontaneously at locations within the jet where the fuel-air ratio and temperature combine to minimize ignition delay. Because the fuel-air ratio within the jet varies over a wide range, suitable conditions for ignition are always present and ignition is reliable irrespective of the total quantity of fuel injected, provided air temperatures are sufficiently high. The initial rate of combustion depends strongly on the length of the ignition delay period. A longer ignition delay gives time for more fuel and air to become mixed before ignition. This quantity of mixture burns quickly and gives rise to the objectionable "knocking" sound often associated with diesel engines. Ignition delay increases as the temperature of the compressed air decreases. Therefore reductions in compression ratio, ambient air temperature, or the temperature of the engine coolant increase the ignition delay.

The subsequent combustion rate is controlled by the rate of injection of the remaining fuel and the rate of entrainment of air into the fuel jet. High air entrainment rates increase combustion rates and therefore produce faster rates of pressure rise in the cylinder. Gas temperature is also increased, which in turn increases the quantity of nitrogen oxides in the exhaust. On the other hand, low air entrainment rates produce regions in the combustion chamber where the fuel-air ratio is too high for complete combustion. Partial combustion in these regions produces soot and hydrocarbons in the exhaust gas. Diesel engine combustion-system research is therefore mainly directed to striking an acceptable balance in fuel-air mixing rate.

In engines designed for high-speed service, the mixing rates between fuel and air may be too low to allow combustion to be completed quickly. Such engines often use the indirect injection (IDI) system (see Fig. 2), in which the fuel is injected into a smaller chamber in the cylinder head (precombustion chamber) which is connected to the cylinder volume by one or more orifices. During the compression stroke, air forced into the precombustion chamber is highly turbulent and combustion is therefore rapid. Unburned and partially

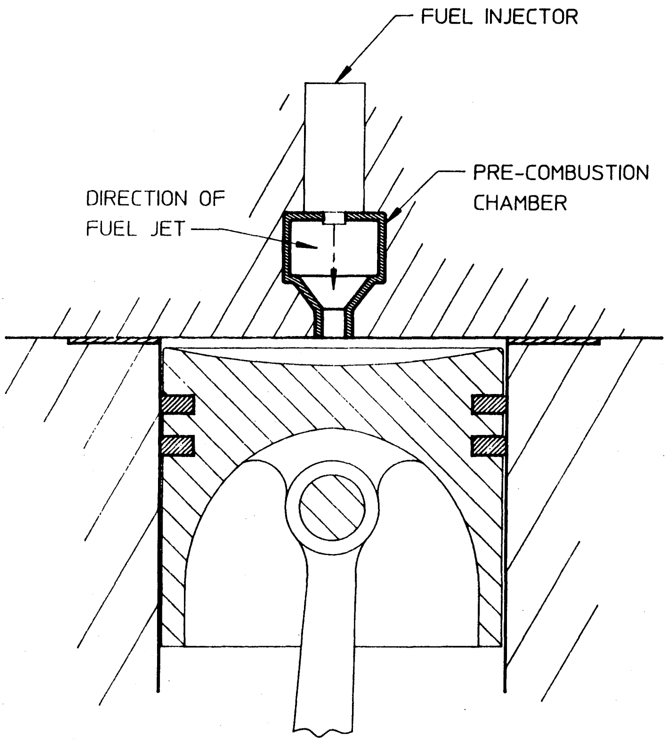

FIG. 2. An indirect injection system with a single orifice between precombustion chamber and cylinder.

burned fuel is ejected from the precombustion chamber by the high-pressure combustion products, and combustion is quickly completed in the cylinder volume as the jet(s) from the precombustion chamber mix with the air in the cylinder volume. Engines with injection directly into the cylinder volume (direct injection, or DI engines, as in Fig. 1) are suitable for lower rotational speeds and have the advantage of higher efficiency than IDI engines because of less turbulence and therefore lower heat losses.

As a result of its combustion process, the diesel engine has certain characteristics that are significantly different from the spark-ignition engine:

1. The diesel engine requires a higher compression ratio in order to produce (spontaneous) compression ignition of the fuel. Maximum cylinder pressures are therefore higher than in the spark-ignition engine.
2. In addition to the higher compression ratio, a fuel having a short ignition delay is required in order to reduce the ignition delay period.
3. The torque produced by the diesel engine tends to increase linearly with the overall fuel-air ratio in the cylinder. At maximum torque, the overall fuel-air ratio is limited to about 60% of the stoichiometric fuel-air ratio in order to prevent excessive exhaust smoke. (The spark ignition engine is constrained to operate with near-stoichiometric mixtures, since flame speeds become too low with large amounts of excess air.)
4. The torque of a four-stroke cycle engine (diesel or spark ignition) is

   $$T=\left(\frac{1}{4\pi}\right)F\rho Q_f V_d \eta_v \eta_i \eta_m, \qquad (1)$$

   where F is the mass ratio of fuel to air in the combustion chamber, ρ is the density of the air inducted into the engine cylinder, Q_f is the fuel heating value, V_d is the displacement of the engine cylinders, and η_v, η_i, and η_m are the volumetric efficiency (actual engine mass air rate divided by theoretical mass air rate), indicated efficiency (net power produced by compression and expansion strokes divided by rate of fuel energy supplied), and mechanical efficiency (output power divided by net power produced by the compression and expansion strokes), respectively. In the diesel engine, the torque of the engine is controlled by varying the fuel-air ratio (F) by changing the quantity of fuel injected. In the spark-ignition engine, fuel-air ratio is constant (approximately), but inlet density (ρ) is changed by means of the throttle.

The use of lower fuel-air ratios and high compression ratios has important consequences on indicated efficiency, as shown in Fig. 3. At full power, current diesel engines can produce brake efficiencies approaching 50%, primarily as a result of these two factors.

1.3 Turbocharging

The relationship for engine power is

$$P=2\pi TN \qquad (2)$$

or

$$P=\tfrac{1}{2}F\rho Q_f V_d N \eta_v \eta_i \eta_m \qquad (3)$$

for the four-stroke cycle, where N is the rotational speed of the engine. The lower fuel-air ratios used in the diesel engine reduce the power produced for a given displacement volume, under the assumption of equal inlet density and engine speed. This disadvantage can be largely offset by turbocharging, which increases inlet density above that of the ambient atmosphere. A turbocharger consists of a centrifugal compressor and a turbine, usually of the radial-inflow type, connected to a common shaft (see Fig. 4). Exhaust gases from the engine cylinders drive the turbine, while ambient air is compressed by the compressor and supplied to the cylinders, thus increasing the inlet density and raising the engine power. A heat exchanger (called either intercooler or aftercooler) is used to cool the compressed air to further increase inlet density. The intercooler uses either engine coolant or ambient air as the heat sink.

The diesel combustion process is particularly well suited for turbocharging since increases in inlet pressure and temperature reduce ignition delay. The amount of pressure boost that can be used is limited only by the loads imposed on the engine structure by the increased cylinder pressures, the increased heat transfer rates to the combustion chamber surfaces, and turbocharger efficiency. This tolerance to high inlet pressure allows diesel engines to be designed with very high mean

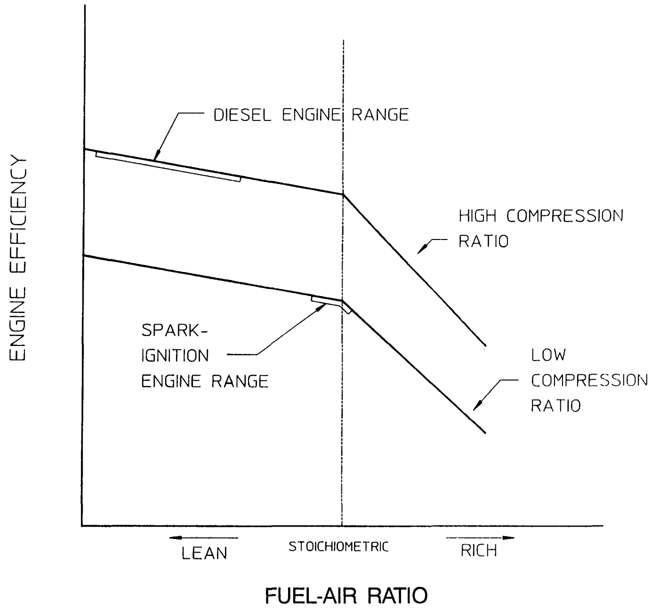

FIG. 3. The effect of compression ratio and fuel-air ratio on indicated efficiency, showing the operating ranges of diesel and spark-ignition engines.

effective pressures (MEP), where, for four-stroke engines,

$$\mathrm{MEP}=\frac{P}{[V_d(N/2)]}. \qquad (4)$$

Turbocharging and intercooling also increase engine efficiency. While output power increases with turbocharging and intercooling, engine friction remains about constant, and so the mechanical efficiency increases. This is reflected in increased fuel economy.

2. APPLICATIONS

The high efficiency of the diesel engine makes it attractive wherever superior fuel economy is of primary concern. Also of importance in some applications (i.e., military and marine applications) is the relatively low vapor pressure of diesel fuel, which lessens fire and explosion hazards.

Weighed against the high efficiency of the diesel are several factors deriving from its thermodynamic cycle. The diesel engine tends to be heavier than the spark-ignition engine because of its higher compression ratio, which requires a stiffer, therefore heavier, structure to resist the higher pressure loads. Engine starting is more difficult in cold climates because ignition delay increases as temperature decreases. Power output for a given engine volume is relatively low unless the diesel is highly turbocharged. Combustion noise tends to be higher than in the spark-ignition engine, particularly under conditions where the ignition delay is long (idling, low ambient temperature). The heavier structure, the precision-machined fuel injection system, and the need for a high-strength forged steel crankshaft increase the cost of the diesel relative to the spark-ignition engine.

As a result, the diesel engine has found its primary applications where the required power levels are high and of long duration (i.e., where fuel costs are a major consideration), where size and weight are of less importance (e.g., in stationary applications or in large vehicles), and in applications not requiring frequent starts or very low noise levels. In such applications, reliability and long engine life are important, and diesel engine designers have responded by providing engines capable of thousands of hours of operation at full power with minimum maintenance needs.

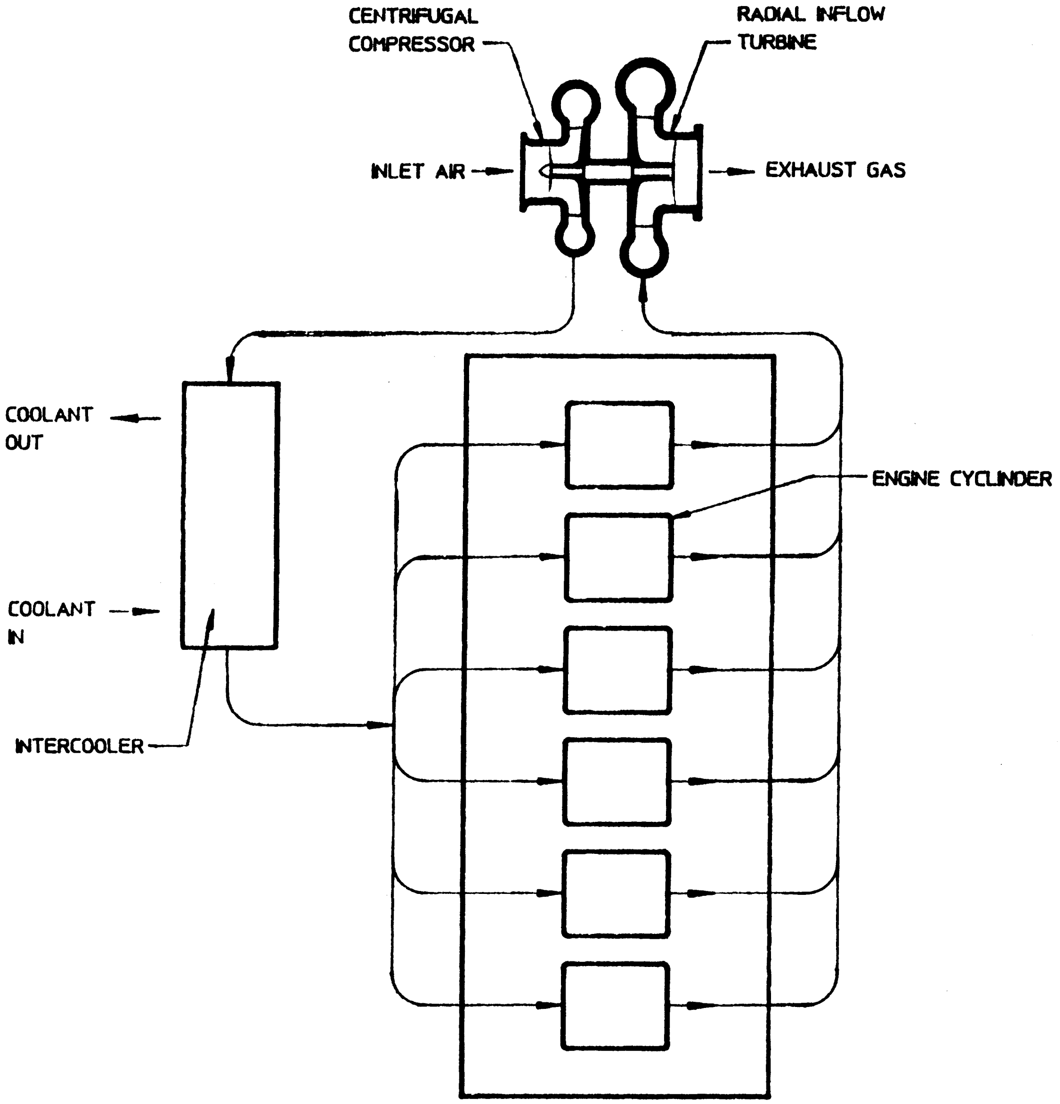

FIG. 4. Turbocharger and intercooler on a six-cylinder diesel engine.

In the United States, where fuel prices are still relatively low, the diesel engine has not supplanted spark-ignition engines for use in automobiles because its advantage in fuel economy is not sufficient to counteract the disadvantages of more difficult starting, and higher weight, volume, cost, and noise. Greater use of automobile diesels is found wherever fuel prices are high. These engines usually employ the indirect-injection (IDI) combustion chamber to allow higher engine speeds, may incorporate a turbocharger with or without an intercooler, and use an electrically heated element (glow plug) in the combustion chamber to assist in starting the engine in cold weather.

In heavy-duty trucks, buses, farm machinery, earth-moving equipment, and military vehicles the diesel engine is dominant. These engines are of the direct-injection design for maximum fuel economy. Turbocharging and intercooling are widely used in such engines. Engine durability is an important design criterion. The diesel engine also enjoys widespread use for electrical power generation, main propulsion as well as auxiliary power

Table 1. Characteristics of typical diesel engines in various applications.

Application	Cylinder diameter (mm)	Piston stroke (mm)	Power per cylinder[a] (kW)	Mean effective pressure[a] (bar)	Speed[a] (rev/s)	Combustion chamber[b]	Strokes per cycle	Air induction system[c]	Compression ratio	Specific engine weight[a] (kg/kW)	Brake efficiency[a,d] (%)
Automobile	90	90	13	6.2	73	IDI	4	NA	22	6.3	30
Heavy-duty truck	130	150	47	12.7	37	DI	4	T,I	16	3.8	41
Smaller stationary	159	159	52	9.3	35	DI	4	T	14	5.2	37
Locomotive	230	250	177	10.5	16	DI	2	T,I	14	6	41
Large stationary	380	460	490	18.7	10	DI	4	T,I	12	13.2	41
Large marine	840	2400	3000	14.9	1.5	DI	2	T,I	10	45	48

[a] At maximum power.
[b] IDI, indirect injection; DI, direction injection.
[c] NA, naturally aspirated; T, turbocharged; I, intercooled.
[d] Based on lower heating value of diesel fuel.

for boats and ships, and railway power for diesel-electric locomotives.

The salient characteristics of typical diesel engines used in various applications are shown in Table 1. As seen in this table, diesel engines are built in an extraordinary range of sizes and powers. Cylinder diameters range over an order of magnitude, and displacement volumes range from less than 0.5 L/cylinder to over 1300 L/cylinder. On the other hand, mean effective pressure is independent of engine size or speed and rather depends on the amount of turbocharging used.

Average piston speed (the product of twice the crankshaft rotational speed and the stroke length) varies between about 7 and 13 m/s, with smaller values used to prolong engine life by reducing wear on cylinder liners, pistons, and piston rings. Given the relatively small range of piston speeds, rotational speed decreases quickly as stroke increases.

The optimum compression ratio is affected by several factors. As the designed rotational speed decreases, longer ignition delays are acceptable and therefore compression ratio can be decreased. Engines that are housed in heated enclosures (e.g., large stationary and marine engines) can accept a lower compression ratio without starting problems because low inlet air temperatures are avoided. Larger engine cylinders lose a relatively smaller amount of heat from the air during the compression stroke, and a lower compression ratio will produce an acceptable ignition delay. Since at the same internal gas pressure engine structural stresses increase as cylinder size increases, it is desirable to decrease compression ratio in the larger engines. This trend is seen in Table 1.

In engines with the same MEP and piston speed, power per unit piston area is constant. At the same time, the weight tends to increase as the cube of cylinder diameter. The specific engine weight (engine weight divided by engine power) therefore tends to be proportional to cylinder diameter. Although the engines of Table 1 have somewhat different MEPs and piston speeds, the trend of increasing specific weight with cylinder diameter is evident.

Brake efficiency of these example engines varies from about 30% to 48%. The lowest value is for an automobile engine with the relatively inefficient IDI combustion system. This engine also has the highest piston speed (which increases friction losses) and these two factors contribute to its relatively low efficiency. The low piston speed, low rotational speed that gives a long time period for combustion, and careful attention to design details that are economically feasible in an expensive engine all contribute to the high efficiency of the large marine engine.

3. DESIGN CONSIDERATIONS

3.1 Fuel Injection Systems

The component most critical to satisfactory diesel combustion is the fuel injection system. This system must inject precise quantities of fuel to each of several cylinders once each engine cycle. Fuel injection pressure may be as high as 2000 bar. The duration of the injection event ranges from 2° to 30° of crankshaft rotation, depending on engine torque demand. This corresponds to injection time intervals from 0.1 to 2 ms at an engine speed of 40 revolutions per second. Injection timing (the crankshaft position where injection starts) is critical and is controlled to within ±1 crankshaft degree.

A common method used to meet these requirements is a "pump-line-nozzle" fuel injection system (see Fig. 5). A camshaft drives several pump pistons (one for each engine cylinder) within their respective cylinders. Each piston delivers high-pressure fuel through a tube to an injection nozzle inserted through the cylinder head of the engine. One fuel injection event occurs each cycle in each engine cylinder. Injection timing is established by the phasing of camshaft rotation relation to crankshaft rotation. Upon delivery of fuel to the injection nozzle, fuel pressure within the nozzle lifts the "needle," permitting fuel to exit from the nozzle through multiple orifices into the combustion chamber. When fuel delivery ceases, a spring returns the needle to its seat.

The quantity of fuel delivered to the nozzle is controlled by rotating the pump piston.

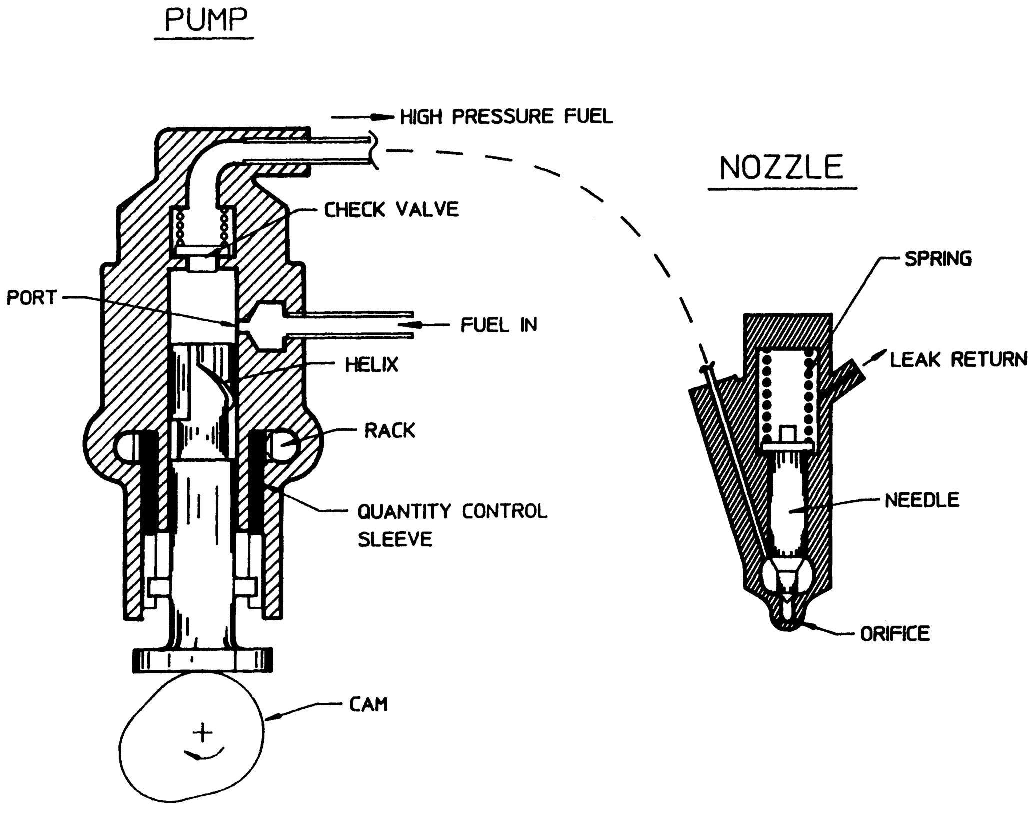

FIG. 5. Pump-line-nozzle fuel injection system for one engine cylinder.

With the piston positioned as shown in Fig. 5, the pump cylinder is filled with fuel from a low-pressure supply pump through a port in the cylinder wall. Fuel delivery to the nozzle commences when the top of the piston covers the port. Fuel delivery ceases when a helix machined in the piston uncovers the port and fuel from above the piston flows out of the port through a slot in the pump piston. By rotation of the piston, the vertical piston position where fuel delivery ceases can be made to occur either sooner or later, thus decreasing or increasing the quantity of fuel delivered.

In smaller engines, such as those used in automobiles and trucks, a multi-cylinder pump is often used in which each pump piston is connected to a nozzle in each engine cylinder. Larger engines may have a separate single-piston pump, connecting tube, and nozzle for each engine cylinder. Alternatively, both large and small engines may combine a single pump piston and injector into a single unit (a "unit" injector) installed in each engine cylinder head that eliminates the connecting tubing and is driven by an overhead camshaft.

With the advent of strict limits on exhaust emissions from diesel engines, the necessity for better control of fuel injection has resulted in the development of ways to control various fuel injection parameters electronically. A few of these systems will be described. Figure 6 shows a pump piston and cylinder containing a movable timing control sleeve. The principles of operation are the same as those described for the pump-line-nozzle system of Fig. 5, except that the port in the cylinder wall is contained in the movable timing control sleeve. Moving the sleeve upward delays the start of injection; injection start becomes earlier when the sleeve is moved downward. The actual start of injection is measured by a needle lift sensor attached to the nozzle body, and a closed-loop control system ensures that the actual start of injection matches the required timing. A computer is used to generate the desired timing schedule on the basis of inputs from other engine sensors (engine temperature, engine speed, delivered fuel quantity). Fuel quantity control is obtained by rotation of the pump piston, as previously described.

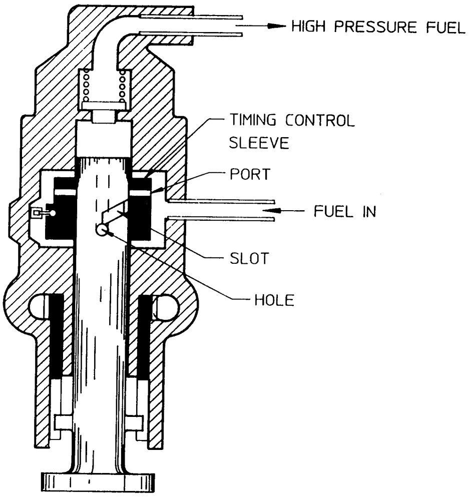

FIG. 6. Injection pump with timing control sleeve.

In another approach, the opening and closing of the port is accomplished by a fast-acting high-pressure solenoid valve rather than by the pump piston (see Fig. 7). The solenoid valve is closed to commence injection. Injection ceases when the solenoid valve is opened and the fuel above the pump piston is returned to the fuel tank. By this arrangement, injection quantity is controlled by the length of time that the solenoid valve is closed, and start of injection is controlled by the time of closing of the solenoid valve. The operation of the solenoid valve is controlled by a computer which obtains various input signals (engine speed, operator's power demand from the accelerator pedal position, engine temperature, and inlet air pressure and temperature).

The need for a camshaft to drive the pump piston can be eliminated with a common-rail fuel injection system. One example is shown in Fig. 8. Here, a high-pressure feed pump (not shown) continuously supplies fuel to all the injectors. Each injector contains an electronically controlled three-way two-position valve. When high-pressure fuel is directed to the space above the injector needle, the needle is held down on its seat and no injection occurs. When the three-way valve connects the space above the needle to the fuel return line, the fuel pressure at the lower end of the needle raises the needle and fuel injection starts. This arrangement permits not only control of fuel delivery quantity and time of injection start, but also the rate of needle lift (by sizing the control orifice diameter) and hence the injection rate during the early part of the injection event. Closed-loop control is

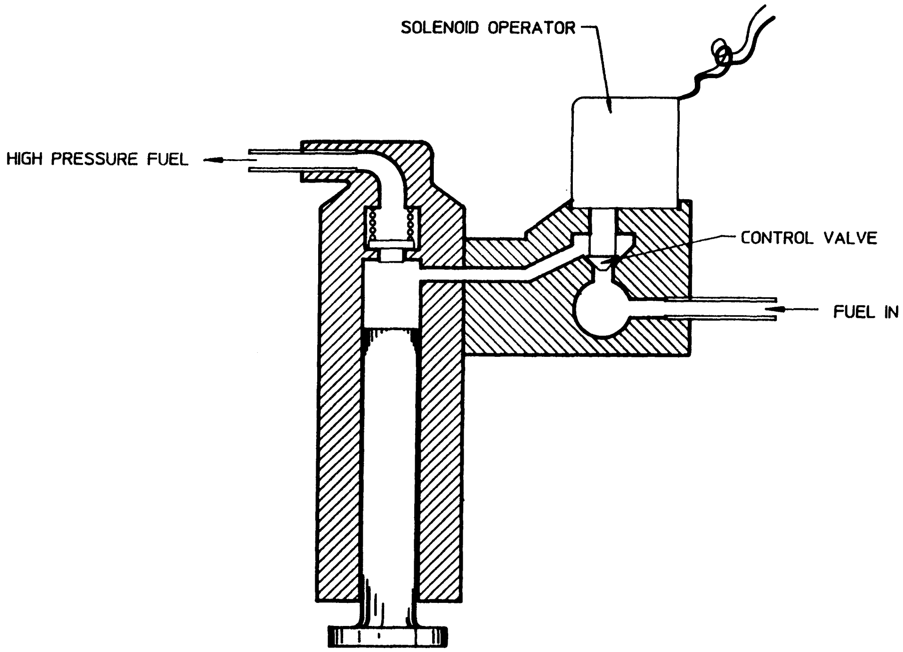
SOLENOID OPERATOR
HIGH PRESSURE FUEL
CONTROL VALVE
FUEL IN

FIG. 7. Fuel injection pump with a solenoid-operated valve for fuel quantity and injection timing control.

effected, as usual, by various sensor inputs to a computer control unit.

3.2 Exhaust Emission Control

Diesel exhaust emissions are now a major concern of diesel engine manufacturers, particularly for engines used in highway vehicles. For these engines, legislative groups have set emission standards that have become increasingly stringent as environmental concerns increase. Emissions standards of similar severity will eventually be applied to nonvehicular engines.

Emittants controlled by laws are nitrogen oxides, carbon monoxide, hydrocarbons, and particulate matter. Nitrogen oxides are formed

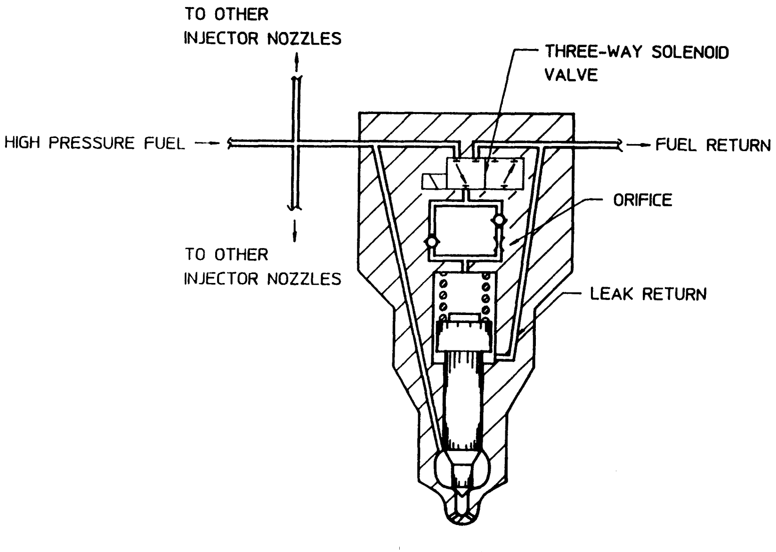
TO OTHER INJECTOR NOZZLES
THREE-WAY SOLENOID VALVE
HIGH PRESSURE FUEL
FUEL RETURN
ORIFICE
TO OTHER INJECTOR NOZZLES
LEAK RETURN

FIG. 8. Common-rail fuel injection system with three-way solenoid valve.

in the combustion chamber at locations where gas temperatures are high. In the burning fuel jet, regions where the fuel-air mixture is nearly stoichiometric produce the highest temperatures and thus are the source of most of the nitrogen oxides. Carbon monoxide is produced during the fuel burning process as the fuel carbon oxidizes, first to carbon monoxide and then to carbon dioxide. Since the $CO \rightarrow CO_2$ reaction rate is relatively fast and the diesel engine combustion chamber contains excess oxygen, carbon monoxide emissions are rarely a problem for diesel engines. Hydrocarbon emissions result from unburned or partially burned fuel, and the sources are the fuel-rich core of the fuel jet, the walls of the combustion chamber that may be wetted by the fuel jet, and crevices in the combustion chamber (e.g., the small space between piston and cylinder above the top piston ring) where the gases are cooled by the adjacent metal and combustion is therefore suppressed. Exhaust particulates consist of agglomerates of tiny carbon particles (soot) as well as various hydrocarbons species that adsorb on the carbon. Particulate emissions are associated with fuel-rich combustion. Lubricating oil that enters the combustion chamber past the piston rings or leaks into the exhaust stream past the intake or exhaust valve guides is also an important source of particulate emissions. An injection event that ends with low-velocity injection of the last fuel injected is certain to produce both soot and hydrocarbon emissions.

While diesel exhaust emissions are strongly affected by engine design details, the overall fuel-air ratio is also an important factor (see Fig. 9). The rapid increase of particulate emissions and exhaust smoke as fuel-air ratio increases limits the diesel engine to fuel-air ratios less than about 60% of the stoichiometric fuel-air ratio. Nitrogen oxides are less sensitive to fuel-air ratio because, regardless of fuel injection quantity and the overall fuel-air ratio, portions of the fuel jet burn at near-stoichiometric levels of fuel-air ratio producing high local temperatures. Carbon monoxide is not formed in large amounts until the overall fuel-air ratio exceeds stoichiometric. Hydrocarbon emissions are too engine-specific to include in Fig. 9.

In diesel engine development, control of nitrogen oxides and particulate emissions is the major concern. The problem is difficult because the most effective control techniques to reduce nitrogen oxides tend to increase particulates. Further, the undesirable emissions are generated during a combustion event which is dominated by a complex turbulent mixing process that is difficult to control. The development of catalysts to reduce nitrogen oxides in the exhaust has not yet been achieved

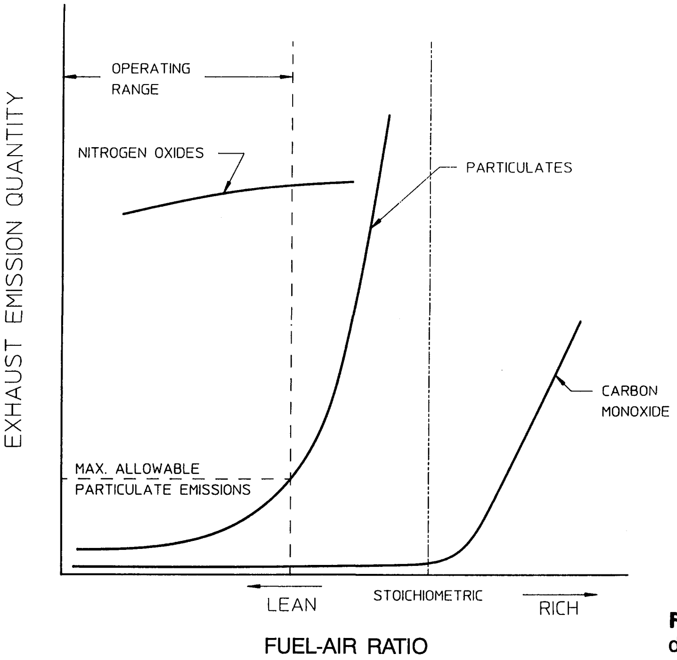

FIG. 9. Effect of overall fuel-air ratio on diesel engine exhaust emissions.

because of the large amounts of excess oxygen in the exhaust.

Despite these difficulties, the exhaust emissions of diesel engines have been greatly reduced from their precontrol levels. Particulate emissions are now about 30% of those before legislative requirements, and nitrogen oxides are about 20% of the precontrol levels. These improvements have been made by alterations to the combustion process through detail changes in design. Some of these are

1. increases in fuel injection pressure to increase the mixing rate of the fuel jet to reduce particulate emissions, combined with a later injection event to reduce emissions of NO_x;
2. reduction in crevice volume, reduction in the volume of unused air spaces in the combustion chamber, and modification of turbocharger air delivery characteristics to increase the amount of air available for combustion, thus allowing a lower overall fuel-air ratio for the same power;
3. design of inlet ports to optimize the swirling motion of the air in the combustion chamber, for improved mixing;
4. improvements in fuel injection systems to eliminate low fuel-jet velocity at the end of the injection event, and to provide better control of injection timing;
5. reduction in oil consumption; and
6. careful matching of the shape of the combustion chamber to the number of fuel jets, the cylinder air motion, and the penetration distance of the fuel jets into the combustion air.

With such modifications, the diesel engine has been able to meet current legislative requirements. As emission regulations become more severe, additional control methods must be applied. Some of these are currently under development.

1. Filters can be used to remove a large fraction of the particulate matter from the exhaust. These filters must be periodically cleaned to prevent excessive exhaust gas pressure losses, and the filtering-cleaning cycle is short because of the relatively large quantity of exhaust soot as well as the limits to filter size imposed by vehicle installation. One method of cleaning is to burn off the collected particulate matter periodically. A burner or electrical heater can be used to ignite the collected soot. Alternatively or in combination, the filter medium (made from porous ceramic) can be doped with a catalytic material to reduce the ignition temperature or a fuel additive containing a soluble catalyst can be used.
2. Cooled exhaust gas can be recirculated and mixed with the inducted air charge. The addition of this inert gas reduces combustion temperature and therefore reduces the quantity of nitrogen oxides, but the decrease in oxygen content of the combustion gas tends to increase particulate emissions.
3. Further modifications to the fuel injection process are possible: higher pressures and more precise control of injection rate, perhaps with these parameters controlled by continuous measurement of exhaust emission levels.

The development of diesel engines with very low emissions can be greatly assisted by a better theoretical understanding of the combustion process: the formation and development of the fuel jet, its interaction with the air motion in the chamber and with the walls of the chamber, the fuel-air mixing process, and the concurrent combustion process.

Exhaust emission control is a serious challenge to the future of the diesel engine. This technical challenge has strong economic overtones; the cost of effective emission control techniques must not prevent the diesel engine from competing with other power sources.

3.3 New Materials

There is much current interest (but as yet no full-scale production) in ceramic materials (e.g., silicon nitride, zirconia) for the walls of diesel engine combustion chambers. Since these materials have a lower thermal conductivity than metals conventionally used in engines, their use can substantially reduce the heat transfer from the combustion gases to the engine coolant. In this way the conventional cooling system can be reduced in size or, in the limit, eliminated. Since the net heat loss from the gases to the coolant is reduced, exhaust gas temperature is increased which provides extra energy in the exhaust stream. This additional energy can potentially be used to drive not only the turbocharger turbine but an additional power turbine geared to the engine crankshaft and by this means can increase engine power output without an increase in fuel consumption. The use of ceramics for this purpose is not without significant problems. Among these are high manufacturing cost, tendency to brittle fracture, problems in matching the thermal expansion of ceramics to that of contacting metal parts, and the selection of lubricating oil properties.

Ceramic materials have begun to be used in limited quantities for engine components subjected to high contact pressure (such as cam followers) or high-temperature corrosion (such as exhaust valves, piston crowns, and precombustion chambers).

GLOSSARY

Average Piston Speed: The product of twice the rotational speed of the crankshaft and the length of piston stroke.

Brake Efficiency: The dimensionless ratio of the power produced at the engine output shaft to the rate of fuel energy supplied to the engine.

Common-Rail Fuel Injection System: A fuel injection system in which a continuous supply of high-pressure fuel is provided to the injector nozzle, and fuel quantity and timing control is effected by the nozzle rather than by the pump.

Compression Ratio: The ratio of the maximum cylinder volume to the minimum cylinder volume.

Cylinder Displacement: The volume swept by the piston in one stroke.

Fuel-Air Ratio: The mass ratio of fuel to air. In diesel engine terminology, the unmodified term refers to conditions at a point within the combustion chamber. See Overall Fuel-Air Ratio.

Fuel Heating Value: The heat energy produced by the complete combustion of a unit mass of fuel.

Fuel Injection Pressure: The pressure of the fuel just upstream of the nozzle orifices.

Ignition Delay: The time between the start of fuel injection and fuel ignition.

Indicated Efficiency: The dimensionless ratio of the net power produced by the compression and expansion strokes of the engine to the rate of fuel energy supplied to the engine. The product of indicated efficiency and mechanical efficiency is brake efficiency.

Mechanical Efficiency: The ratio of the power produced by the engine output shaft to the net power produced by the compression and expansion strokes. Mechanical efficiency is less than 1 because of friction losses (mechanical friction as well as losses due to gas viscosity).

Naturally Aspirated: An engine without a turbocharger or other means for increasing inlet air pressure.

Overall Fuel-Air Ratio: The mass ratio of the total injected fuel quantity to the total air quantity trapped in the cylinder after the intake valve(s) or port(s) is closed.

Stoichiometric Fuel-Air Ratio: The fuel-air ratio that, under conditions of equilibrium, produces combustion products containing no unburned fuel or free oxygen.

Volumetric Efficiency: The mass of air inducted into the engine in one cycle divided by the mass of air that could be contained in the cylinder at the pressure and temperature of the air in the intake manifold.

Further Reading

Heywood, John B. (1988), *Internal Combustion Engine Fundamentals*, Cambridge, MA: MIT. Text on engine combustion system theory.

Lilly, L. R. C. (Ed.) (1984), *Diesel Engine Reference Book*, London: Butterworth. Thorough discussions of diesel engine design practices.

Taylor, Charles Fayette (1985), *The Internal-Combustion Engine in Theory and Practice, Volumes 1 and 2*, Cambridge, MA: MIT. Text on engine theory and design.

DIFFUSION AND IONIC CONDUCTION IN LIQUIDS

DEREK G. LEAIST, *Department of Chemistry, University of Western Ontario, London, Ontario, Canada N6A 5B7*

Introduction 1
1. **Diffusion in Isothermal Liquids** 2
1.1 Historical Background 2
1.2 Chemical Interdiffusion in Binary Liquids 3
1.2.1 Fick's Laws 3
1.2.2 Mobilities and Friction Coefficients 3
1.2.3 Nonelectrolytes 4
1.2.4 Electrolytes 4
1.2.5 Interdiffusion with Fast Chemical Reactions 5
1.3 Multicomponent Interdiffusion 5
1.4 Intradiffusion in Chemically Uniform Liquids 6
1.5 Self-Diffusion in Pure Liquids 6
1.6 Brownian Motion and the Random Walk 6
1.7 Experimental Methods 7
1.7.1 Optical Interferometry 7
1.7.2 Diaphragm Cells 7
1.7.3 Taylor Dispersion 8
1.7.4 Capillary Tubes 8
1.7.5 Nuclear Magnetic Resonance 8
2. **Ionic Conduction in Liquids** 8
2.1 Historical Background 8
2.2 Molar Conductivities 8
2.3 Ionic Mobilities and Transference Numbers 9
2.4 Ion Transport Equations 10
2.5 Experimental Methods 10
2.5.1 Conductivity Measurements 10
2.5.2 Hittorf Experiment 11
2.5.3 Moving-Boundary Experiment 11
3. **Thermal Diffusion** 11
3.1 Historical Background 11
3.2 Diffusion and Heat Flow in Binary Liquids 12
3.3 Soret Steady State 12
3.4 Heat of Transport 12
3.5 Experimental Methods 12
Glossary 13
Works Cited 13
Further Reading 14

INTRODUCTION

Diffusion, the transport of matter by random thermal motions, is one of nature's fundamental processes. If a few drops of ink are injected into the lower end of an unstirred glass of water, more ink diffuses upward than downward, simply because more randomly moving ink molecules are near the bottom of the glass than near the top. The ink gradually spreads from regions of higher to those of lower concentration. Eventually a uniform distribution is reached.

Despite instantaneous velocities of hundreds of meters per second, diffusing molecules make little headway in liquids because they travel only a fraction of a molecular diameter before being deflected by collisions with their neighbors. In fact, it takes about *10 years* for a molecule to diffuse through 1 m of unstirred liquid! For cold or viscous liquids, or macromolecules, diffusion can be orders of magnitude slower.

The slowness of diffusion has practical consequences for the dissolution and precipitation of crystals, gas absorption, mixing, and other mass transfer operations. Speeding up these processes requires mechanical convection, such as stirring, to reduce the distances over which molecules must diffuse. Even with efficient stirring, the rates of many chemical reactions are ultimately limited by how fast diffusion can bring the reactants together.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Measured rates of diffusion provide information about the sizes and shapes of molecules, molecular aggregation, intermolecular forces, and the structure of solutions.

Thermal motion is not the only way to transport matter on the molecular level. If a voltage gradient is applied to an electrolyte (such as molten sodium chloride or sodium chloride dissolved in water), the positive and negative ions will begin to drift in opposite directions. Ion-solvent and ion-ion interactions can be investigated by measuring electrolytic conductivities. Ionic conduction is essential for the operation of batteries and fuel cells, the production of important industrial chemicals, electroplating, corrosion, the transmission of nerve impulses, and many other processes.

Surprisingly, the flow of heat through a mixture can also transport matter. If one region of a convection-free mixture is cooled and another heated, partial demixing occurs. In nonisothermal salt solutions, for example, the dissolved ions usually migrate with the flow of heat to the cooler parts of the solution, while the solvent migrates to the warmer parts. The relative diffusion of substances caused by a temperature gradient is called thermal diffusion. The reciprocal effect—the flow of heat caused by concentration gradients—is also observed.

Diffusion in a concentration gradient and heat conduction in a temperature gradient can be considered "first-order" processes. Thermal diffusion is a more subtle "cross-effect." Understanding thermal diffusion therefore provides a sensitive test of microscopic theories of liquids. In practice, thermal diffusion effects are usually too small to be important, though useful separation of mixtures can be achieved in some cases.

Section 1 of this article covers the principles of diffusion in isothermal liquids. After a summary of the major advances that have been made in diffusion studies, the different kinds of diffusion are discussed: chemical interdiffusion, self-diffusion, tracer diffusion, and Brownian motion. The section concludes with a brief description of experimental techniques for measuring diffusion coefficients.

Section 2 is devoted to ionic conduction in liquids. The discussion begins with historical developments, then moves to the relation between the electrical conductivity and the concentrations and mobilities of the charge-carrying ions. Techniques are outlined for measuring ionic mobilities, from which the drift speeds of ions can be deduced.

Thermal diffusion is the subject of the concluding section (Sec. 3). An empirical approach is taken, using coupled equations for the simultaneous flow of heat and diffusion of matter through mixtures. The Soret and Dufour effects, the thermal diffusion coefficient, Soret coefficient, and heat of transport are introduced.

1. DIFFUSION IN ISOTHERMAL LIQUIDS

1.1 Historical Background

Graham (1850) appears to have been the first to understand that the flux of a diffusing substance is directly proportional to its concentration gradient. A few years later Fick grasped the analogies between diffusion, the flow of heat in a temperature gradient (Fourier's law), and the flow of electric current in a voltage gradient (Ohm's law). He formulated the empirical laws of diffusion (Fick, 1855).

Progress was also being made along theoretical lines. Gibbs (1876) pointed out that gradients in chemical potential, rather than concentration gradients, are the fundamental driving forces for diffusion. This was one of the first steps toward the extension of classical thermodynamics to nonequilibrium systems. Onsager (1931) and others would later develop the thermodynamics of irreversible processes, providing a unified treatment of transport processes.

In the early years of this century the molecular-kinetic theory developed by Einstein (1905) accounted for the chaotic diffusion behavior of colloidal particles observed under a microscope. Visibly linking the molecular and macroscopic worlds, this crucial result helped establish the validity of molecular theory. A foundation had been laid for statistical mechanical treatments of liquid-phase diffusion.

The 1950s were exciting years for diffusion research (Tyrrell and Harris, 1984): accurate gradient techniques for measuring diffusion were developed; the difficult problem of measuring coupled diffusion in multicomponent mixtures was finally solved; radioactive isotopes became widely available for tracer diffusion studies; and molecular-dynamics simulations using digital computers opened a new window on diffusion. In recent years flow techniques and spectroscopic methods based on laser light scattering and nuclear magnetic resonance have become popular sources of diffusion data.

Largely because the forces between molecules in liquids are still not known in detail, a rigorous theory of liquid-phase diffusion is still under development.

1.2 Chemical Interdiffusion in Binary Liquids

Fick studied the interdiffusion of miscible liquids containing the same chemical substances, but at different concentrations. An example would be a solution containing 1.0 mol L^{-1} of heptane dissolved in benzene diffusing into a 0.5-mol-L^{-1} solution of heptane in benzene. Interdiffusion is by far the most common form of diffusion. For two-component systems, it is also called binary mutual diffusion or simply binary diffusion.

1.2.1 Fick's Laws. Suppose that substances *A* and *B* interdiffuse at constant temperature, without convection. Fick's law gives two diffusion equations:

$$J_A=-D_A\frac{\partial C_A}{\partial z}, \tag{1}$$

$$J_B=-D_B\frac{\partial C_B}{\partial z}. \tag{2}$$

C_A and C_B are the concentrations of *A* and *B* in moles per unit volume. J_A and J_B are the fluxes of the substances in moles per unit area per unit time in the *z* direction. The fluxes are measured in the volume-fixed (laboratory) frame of reference defined by

$$J_A V_A+J_B V_B=0. \tag{3}$$

V_A and V_B are the partial molar volumes of *A* and *B*.

Two substances are diffusing, yet only one flux is independent in view of Eq. (3). Moreover, it is easily shown that diffusion coefficients D_A and D_B are equal (Hartley and Crank, 1949):

$$D_A=D_B\,(=D). \tag{4}$$

D is the binary interdiffusion coefficient.

The differential equation for the binary interdiffusion of a substance is obtained by applying the principle of conservation of mass to the flux equation:

$$\frac{\partial C_A}{\partial t}=-\frac{\partial J_A}{\partial z}=\frac{\partial}{\partial z}\left(D\frac{\partial C_A}{\partial z}\right). \tag{5}$$

If *D* can be considered constant, which is a good approximation over small concentration intervals, Eq. (5) simplifies to

$$\frac{\partial C_A}{\partial t}=D\frac{\partial^2 C_A}{\partial z^2}. \tag{6}$$

Equation (6) has been solved for many initial and boundary conditions. Solutions in two and three dimensions are also available (Crank, 1975).

D is found to depend on the nature of the diffusing substances and their concentrations, as well as the temperature and pressure (Jost, 1960). For many liquids *D* is on the order of 10^{-9} m^2 s^{-1} (about one thousand times smaller than typical values for gases).

How fast, on average, is a diffusing substance moving? The flux of a substance equals its concentration times its diffusion velocity v_A,

$$J_A=C_A v_A, \tag{7}$$

so v_A is given by

$$v_A=-\left(\frac{D}{C_A}\right)\frac{\partial C_A}{\partial z}. \tag{8}$$

For a substance at concentration 1000 mol m^{-3} in a concentration gradient of 1000 mol m^{-4}, assuming the value 1×10^{-9} m^2 s^{-1} for *D* leads to a diffusion speed of 1 nm s^{-1}.

For substances of ill-defined molar mass, it is convenient to work with mass fluxes and mass concentrations (e.g., kg m^{-3}) rather than the corresponding molar quantities.

Diffusion fluxes and velocities relative to the center of mass or to the solvent are often used in theoretical treatments of diffusion. Procedures for transforming from the volume-fixed frame used here to other coordinate systems have been developed (Fitts, 1962).

1.2.2 Mobilities and Friction Coefficients. There are more fundamental ways to describe binary interdiffusion. According to the laws of nonequilibrium thermodynamics (de Groot and Mazur, 1962), the flux of a substance is proportional to the gradient in its chemical potential,

$$J_A = -L_{AA}\frac{\partial \mu_A}{\partial z}. \qquad (9)$$

L_{AA}, the Onsager transport coefficient of A, is related to the more familiar binary interdiffusion coefficient as

$$D = L_{AA}\frac{\partial \mu_A}{\partial C_A}. \qquad (10)$$

Both $\partial\mu_A/\partial C_A$ and D drop to zero for mixtures on the brink of phase separation at critical solution points (Haase and Siry, 1968).

When substances interdiffuse, the Gibbs free energy lost during the mixing process is dissipated against viscous frictional forces. For binary interdiffusion, the chemical potential gradient $-\partial\mu_A/\partial z$ can be regarded as a virtual force that propels A through the solution.

Pursuing the mechanical analogy further, the mobility per mole of A may be defined as the ratio of its diffusion speed to the driving force,

$$U_{Am} = -v_A\Big/\left(\frac{\partial \mu_A}{\partial z}\right) = L_{AA}/C_A. \qquad (11)$$

The reciprocal of U_{Am} is called the molar friction coefficient of A. The mobility of a single molecule of A is simply the molar mobility times the Avogadro constant,

$$U_A = N_A U_{Am}. \qquad (12)$$

The mobility is a kinetic quantity that depends on factors such as the size and shape of the diffusing molecules, the composition, and the viscosity of the surrounding medium.

For a better understanding of the meaning of the interdiffusion coefficient, Eqs. (10) and (11) may be combined to give

$$D = U_{Am}\frac{\partial \mu_A}{\partial \ln C_A} \qquad (13)$$

to show that D is the product of kinetic and thermodynamic factors. The latter is often overlooked.

1.2.3 Nonelectrolytes. The chemical potential of a nonelectrolyte relative to a suitable reference value can be expressed as

$$\mu_A = \mu_A^r + RT\ln(y_A C_A), \qquad (14)$$

and therefore

$$D = RTU_{Am}\left(1 + C_A\frac{\partial \ln y_A}{\partial C_A}\right), \qquad (15)$$

where R is the gas constant and T the temperature. The activity coefficient y_A is included to allow for the departures from ideal thermodynamic behavior shown by real solutions. For an infinitely dilute solution, however, the activity-coefficient term vanishes, leading to the limiting relation

$$D^0 = RTU_{Am}^0 \quad (C_A = 0). \qquad (16)$$

Stokes (1903) showed that the mobility of a spherical particle in a continuum fluid is given by

$$U = 1/6\pi\eta r, \qquad (17)$$

where r and η are the particle radius and the viscosity of the fluid. For molecules that are approximately spherical, Eqs. (16) and (17) can be used to estimate limiting interdiffusion coefficients:

$$D^0 = RT/6\pi N_A \eta r = kT/6\pi\eta r; \qquad (18)$$

k is the Boltzmann constant. The corresponding results for nonspherical molecules have been developed.

Equation (18), which is called the Stokes–Einstein equation, suggests that D^0 is proportional to T and inversely proportional to both the radius of the diffusing molecule and the viscosity of the surrounding solvent. These correlations hold reasonably well for molecules that are much larger than the solvent molecules. Since experimental values of the viscosity are invariably used, treatments of diffusion based on Eq.(18) are semiempirical.

1.2.4 Electrolytes. Fick's equations also describe the interdiffusion of a salt dissolved in a solvent, such as aqueous sodium chloride. Although there are two kinds of diffusing ions (the cations and anions), they must diffuse at the same rate to keep the solution electrically neutral. Only one flux is therefore independent, and the binary diffusion equations apply (Robinson and Stokes, 1959).

Because of long-range Coulombic interactions, the values of D for dilute salt solutions are sensitive to the concentration.

For a dissolved salt that dissociates into ν_+ cations and ν_- anions, the limiting interdiffusion coefficient and the ionic diffusion coefficients are related by Nernst's equation,

$$D^0=(\nu_++\nu_-)D^0_+D^0_-/(\nu_-D^0_++\nu_+D^0_-). \quad (19)$$

1.2.5 Interdiffusion with Fast Chemical Reactions. Fick's equations also apply to the diffusion of chemically reacting substances provided the reaction rates are rapid enough to maintain chemical equilibrium at each point along the diffusion path. Diffusion is usually slow enough to satisfy this requirement.

A good example is the diffusion of a partially dissociated electrolyte, $MX \leftrightarrow M^+ + X^-$ (Robinson and Stokes, 1959). There are three dissolved species: molecules MX and ions M^+ and X^-. But since electroneutrality and local chemical equilibrium provide two constraints, only one flux is independent. Fick's laws apply, with J_A and C_A representing the flux and concentration of total electrolyte (dissociated plus undissociated forms).

The diffusion of stepwise self-aggregating solutes ($A+A \leftrightarrow A_2$, $A+A_2 \leftrightarrow A_3$,...) has also been treated. In this case J_A and C_A are the flux and concentration of the total A molecules. It turns out that D is not the simple concentration-weighted average of the monomers, dimers, trimers, etc.

1.3 Multicomponent Interdiffusion

There are significant differences between interdiffusion in multicomponent solutions and the binary interdiffusion of the same components (Tyrrell and Harris, 1984). Not only do the rates of diffusion change, but a new feature appears: coupled diffusion, whereby the gradient in concentration of a substance leads to a flux of another substance. Diffusion can occur from regions of *lower to higher concentration*, opposite to the expected direction.

Diffusion tends to be strongly coupled for multicomponent mixtures that show large departures from ideal thermodynamic behavior. Suppose substance i lowers the chemical potential of substance k. Substance k, seeking regions of lower chemical potential, will diffuse "up" a gradient in the concentration of i. In associating mixtures, diffusing substances can bind and cotransport large amounts of other substances.

Mixed electrolytes provide some of the most striking examples of coupled diffusion. When an electrolyte diffuses, an electric field is generated which speeds up the less mobile ion and slows down the more mobile ion in order to prevent charge separation. The electric field produced by the diffusion of an electrolyte will in turn drive coupled flows of the ions belonging to the other electrolytes in the mixture. In extreme cases, one mole of a diffusing electrolyte has been found to cotransport hundreds of moles of other electrolytes.

If a solution contains $N+1$ independent chemical substances, there are N independent diffusional fluxes. The straightforward generalization of Fick's equations,

$$J_i=-\sum_{k=1}^{N} D_{ik}\frac{\partial C_k}{\partial z}, \quad (20)$$

$$\frac{\partial C_i}{\partial t}=\sum_{k=1}^{N} D_{ik}\frac{\partial^2 C_k}{\partial z^2} \quad (i=1,2,\ldots,N), \quad (21)$$

may be used to define multicomponent interdiffusion coefficients. Coefficient D_{ik} gives the flux of substance i caused by the gradient in the concentration of substance k. D_{ik} coefficients have been measured for a number of three-component systems and a handful of four-component systems.

In general, the matrix $\mathbf{D}$ of coefficients D_{ik} is not symmetric. It can be shown that the eigenvalues of $\mathbf{D}$ are positive, real numbers (Cullinan, Jr., 1965), which rules out damped oscillatory solutions of Eq. (21).

Using Onsager's approach, multicomponent diffusion fluxes may be expressed in terms of the gradients in the chemical potentials of the diffusing substances:

$$J_i=-\sum_{k=1}^{N} L_{ik}\frac{\partial \mu_k}{\partial z}. \quad (22)$$

Using the principle of microscopic reversibility, Onsager proved that the matrix of L_{ik} coefficients is symmetric: $L_{ik}=L_{ki}$ (Miller, 1960). It can also be shown that $L_{ii} \geqslant 0$ and $L_{ii}L_{kk} \geqslant L^2_{ik}$.

The two sets of multicomponent transport coefficients are related by

$$D_{ik}=\sum_{m=1}^{N} L_{im}\frac{\partial \mu_m}{\partial C_k}. \quad (23)$$

1.4 Intradiffusion in Chemically Uniform Liquids

Diffusion in liquids of uniform chemical composition is called "intradiffusion" (Tyrrell and Harris, 1984). It can be closely approximated by tagging diffusing species with isotopes. To measure the intradiffusion of heptane in heptane-benzene mixtures, for example, some of the heptane molecules could be tagged with ^{14}C, keeping the concentrations of benzene and total heptane (labeled + unlabeled) constant throughout the mixture. Intradiffusion can also be studied, without isotopic substitution, by orienting the magnetic moments of nuclei. "Tracer" diffusion is the intradiffusion of minute amounts of foreign species in otherwise uniform solutions.

The intradiffusion coefficient of species i, D_i^*, is defined by using Fick's equation to relate the flux and concentration gradient of the labeled species:

$$J_i^* = -D_i^* \frac{\partial C_i^*}{\partial z} . \tag{24}$$

Intradiffusion differs fundamentally from chemical interdiffusion: the former involves the interchange of labeled and unlabeled species under conditions of uniform chemical composition, while the latter involves the interchange of different chemical substances in composition gradients. Figure 1 illustrates the different coefficients for the heptane-benzene system.

There is no general relation between the interdiffusion and intradiffusion coefficients. For the special case of an infinitely dilute binary solution of nonelectrolyte i, however, the intradiffusion coefficient D_i^* and the interdiffusion coefficient D are identical.

Labeled and unlabeled species form thermodynamically ideal solutions. For this reason activity coefficients are not required to interpret intradiffusion. The intradiffusion coefficient of substance i and its mobility are therefore simply related:

$$D_i^* = RTU_{im}^* . \tag{25}$$

Except at infinite dilution, the mobilities derived from intradiffusion and interdiffusion experiments are different.

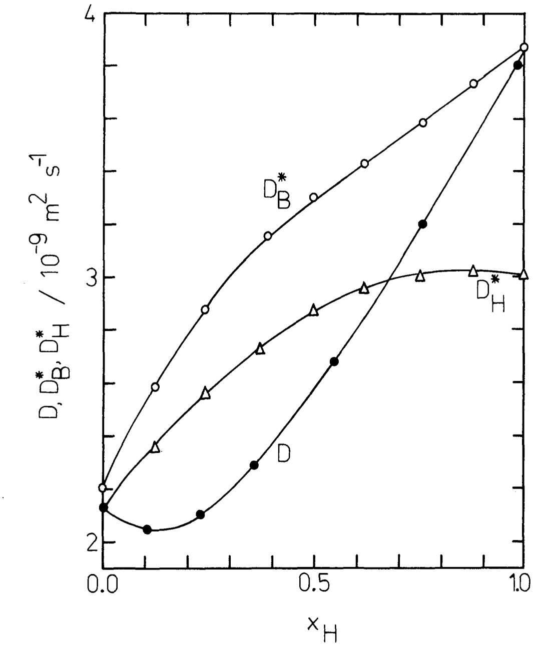

FIG. 1. Interdiffusion coefficient (D) and intradiffusion coefficients (D_H^* and D_B^*) of heptane (H)-benzene (B) mixtures at 25 °C (Harris *et al.*, 1970).

1.5 Self-Diffusion in Pure Liquids

Self-diffusion refers to intradiffusion in a liquid containing two distinguishable species having the same chemical composition. To measure the self-diffusion coefficient of liquid heptane, for example, some of the molecules in a sample of pure heptane could be tagged with ^{14}C.

1.6 Brownian Motion and the Random Walk

Analyzing the motion of a randomly jumping molecule leads to an extremely useful model of diffusion. The simplest case is the one-dimensional diffusion of a molecule at infinite dilution in a uniform solvent. To represent its chaotic thermal motion, the molecule is assumed to take v small steps of length λ per second. Steps are taken at random in the positive or negative directions.

Suppose the molecule is located at position $z=0$ at time $t=0$. Where is it at a later time t? Although its trajectory cannot be predicted, the probability $P(z,t)dz$ of finding the molecule between z and $z+dz$ can be derived for $t \gg v^{-1}$ (Einstein, 1905):

$$P(z,t)dz=(2\pi\lambda^2\nu t)^{-1/2}\exp(-z^2/2\nu\lambda^2 t)dz. \tag{26}$$

When this result is compared with the solution of Fick's equation for diffusion from a point source (Crank, 1975),

$$P(z,t)dz=(4\pi D^0 t)^{-1/2}\exp(-z^2/4D^0 t)dz, \tag{27}$$

a remarkably simple relation emerges between the diffusion coefficient, jump size, and jump frequency:

$$D^0=\nu\lambda^2/2. \tag{28}$$

Figure 2 shows probability distributions calculated with the value 1×10^{-9} m^2 s^{-1} for D^0. As time passes the chances of finding the molecule near the origin drop sharply; information on its position is lost.

The probability distributions shown in Fig. 1 are normal, with variance $2D^0t$. The same kind of distribution also arises in statistical sciences, underscoring the deep connection between diffusion and probability. The root mean square displacement of the molecule from the origin is

$$z_{\text{rms}}=(2D^0t)^{1/2}. \tag{29}$$

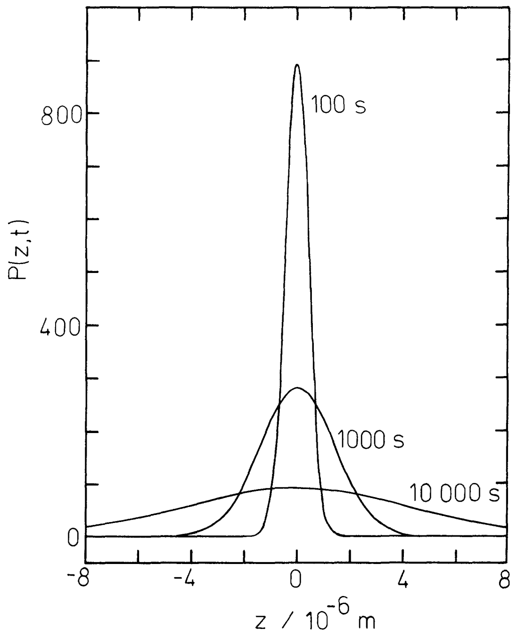

FIG. 2. Probability distributions at 100, 1000, and 10 000 s for a molecule with diffusion coefficient 1×10^{-9} m^2 s^{-1}.

Equation (29) is used to estimate the typical distances that molecules diffuse in one dimension in a given time interval. In two or three dimensions, the displacements are $(4D^0t)^{1/2}$ and $(6D^0t)^{1/2}$.

The diffusion of colloidal particles suspended in a liquid can be observed under a microscope. Bombarded on all sides by solvent molecules, such particles follow a truly chaotic path. This behavior is called Brownian motion.

1.7 Experimental Methods

Diffusion measurements pose two practical difficulties: following small changes in composition, and avoiding convection. Over the years a number of ingeneous diffusion experiments accurate to 1% or better have been developed (Dunlop *et al.*, 1989; Nagashima *et al.*, 1991; Tyrrell and Harris, 1984).

1.7.1 Optical Interferometry.

A sharp boundary between two solutions can be formed by layering the less dense solution over the denser one. When monochromatic light is directed horizontally through the solutions, the light passing near the boundary is bent downward by the sharp gradient in the refractive index. With suitable optics, the refracted light can be made to interfere with undeflected light passing at greater distances from the boundary. Interdiffusion coefficients are calculated from the time-dependent pattern of interference fringes. Two- and three-component liquids have been studied by interferometry. The accuracy of the results (0.1–0.2% of D) is unsurpassed.

1.7.2 Diaphragm Cells.

These simple but effective devices, the workhorses of diffusion experiments, have been used to measure interdiffusion in two-, three-, and four-component solutions, as well as intradiffusion. Diffusion occurs between two solution compartments through stagnant liquid held in the pores of a solution-soaked sintered-glass or sintered-metal frit. After a timed interval (usually several days), the solutions are retrieved and analyzed to determine the changes in concentrations of the diffusing species.

The rate of diffusion depends on the pore area and the effective thickness of the frit. Neither of these quantities is known accurately, so diaphragm cells are calibrated with solutions of known diffusivity, such as aqueous potassium chloride.

1.7.3 Taylor Dispersion. Flow techniques based on Taylor dispersion are a more convenient source of reliable diffusion data. The idea is to inject a narrow pulse of liquid into a laminar carrier stream of different composition confined within a long capillary tube. The injected sample spreads out into a Gaussian distribution as it flows along the tube. A detector, usually a differential refractometer, monitors the broadened distribution of the pulse as it flows out of the tube. Diffusion coefficients are calculated from the eluted pulse shapes.

Interdiffusion in two-, three-, and four-component solutions and intradiffusion have been measured by Taylor dispersion. The collection of data is easily automated using standard liquid-chromatography equipment.

1.7.4 Capillary Tubes. A short length of fine-bore capillary tubing, sealed at one end, is filled with isotopically labeled solution and then lowered into a slowly stirred bath of unlabeled solution. The labeled molecules begin to diffuse out of the tube into the surrounding solution. After a few hours or days, the tube is removed, and the amount of remaining labeled substance is determined, usually by counting of radioactive disintegrations.

1.7.5 Nuclear Magnetic Resonance. The solution is placed in a magnetic field gradient, and then, by use of radio-frequency pulses, the equilibrium distribution of nuclear magnetic moments is disturbed. From the measured decay of the magnetization, which depends in part on how fast the species move into regions of different magnetic field strength, intradiffusion coefficients are calculated.

Nuclear magnetic resonance (NMR) techniques are powerful because the intradiffusion of several different species within the same sample can be measured simultaneously, without isotopic substitution. Measurements can be made over a wide range of intradiffusion coefficient values, down to 10^{-14} m^2 s^{-1}.

2. IONIC CONDUCTION IN LIQUIDS

2.1 Historical Background

When the first electrolyte conductivities were measured, using direct currents, composition changes near the electrodes led to apparent departures from Ohm's law. Kohlrausch overcame this problem by using audio-frequency alternating currents. Electrolytes were found to obey Ohm's law after all (Robinson and Stokes, 1959).

Hittorf had already worked out how to measure the fraction of the total conductivity contributed by each ion. His results together with conductivity data meant that accurate mobilities of individual ions could be calculated. This was a breakthrough, a rare opportunity to determine unambiguous single-ion properties for electrochemical systems (Spiro, 1989).

Conductivity data for partially dissociated electrolytes helped Arrhenius develop the theory of electrolytic dissociation. He proposed that the ions and undissociated molecules of weak electrolytes exist in equilibrium. There was skepticism: Why, at room temperature, should molecules split into ions? It soon became clear that the solvent stabilizes dissolved ions by screening the Coulombic forces and surrounding the ions with a shell of tightly bound solvent molecules.

In the 1930s the general equations for ion transport were developed. Also, Coulombic interactions were shown to account for the strong variations in ionic mobilities at high dilutions (Onsager and Fuoss, 1932). A satisfactory theory for concentrated electrolytes is still elusive, and it is not yet possible to predict accurate ionic mobilities from first principles.

2.2 Molar Conductivities

Ohm's law states that the electric current passing through a conductor is proportional to the applied electric potential difference $\Delta\phi$,

$$I = -\Delta\phi/R\,, \qquad (30)$$

where R is the resistance. For a uniform conductor, the conductance $1/R$ is proportional to the cross-sectional area A and inversely proportional to the length l.

Dividing the current by the area gives the current density i:

$$i = I/A = -(l/AR)(\Delta\phi/l). \qquad (31)$$

Equation (30) may be rewritten as

$$i = \kappa E. \qquad (32)$$

E is the electric field strength $-\partial\phi/\partial z$. κ, an abbreviation for l/AR, is the conductivity. κ depends on the temperature, pressure, and the composition of the conductor, but not its dimensions.

The charge carriers in electrolyte solutions are ions, not electrons. By varying the concentration of ions, the conductivity can be changed over many orders of magnitude. It proves convenient therefore to define the molar conductivity Λ by dividing the conductivity by the concentration of the dissolved electrolyte:

$$\Lambda = \kappa/C. \qquad (33)$$

The molar conductivity of a partially ionized "weak" electrolyte [e.g., acetic acid in water: $CH_3COOH(aq) \leftrightarrow H^+(aq) + CHCOO^-(aq)$] is approximately proportional to its extent of dissociation.

For an electrolyte solution containing positive ions M^{z_M} at concentration c_M (in moles per unit volume) and negative ions N^{z_N} at concentration c_N, the molar conductivity of the electrolyte and the molar conductivities of the ions, λ_M and λ_N, are related by

$$\kappa = C\Lambda_m = c_M\lambda_M + c_N\lambda_N. \qquad (34)$$

Λ_m, λ_M, and λ_N have units S m^2 mol^{-1}.

If the ions moved independently, molar conductivities would be independent of concentration. But ions do, of course, interact, and so molar conductivities vary with concentration, even for very dilute solutions. The interactions are ion-specific. For example, the conductivity of Na^+ ions in 1.0-mol-L^{-1} aqueous NaCl solution differs from that for 1.0-mol-L^{-1} NaI solution. Only at infinite dilution do ions conduct current independently.

2.3 Ionic Mobilities and Transference Numbers

When an electric field is applied to an electrolyte solution, the ions accelerate almost instantaneously to their terminal velocities. The electric force acting on each ion is then balanced by the frictional force offered by the solvent. The mass of an ion plays no role in determining its conductivity. [This simple picture does, however, break down at high field strengths ($\gg$1 kV m^{-1}) or high frequencies ($\gg$10 kHz) (Harned and Owen, 1950).]

The electric mobility u_k of ion k is defined as the ion's terminal drift speed v_k divided by the electric field,

$$u_k = v_k/E. \qquad (35)$$

Taking the product of the drift speed and the concentration gives the flux of the ion, in moles per unit area per unit time:

$$j_k = c_k v_k. \qquad (36)$$

Each mole of ion k carries electric charge z_kF. z_k is called the charge number of ion k (e.g., -2 for $SO_4{}^{2-}$). F, the Faraday constant, is the charge per mole of protons: 96 485 C/mol.

Ion k contributes $c_k v_k F z_k$ to the electric current density. Summing over the different ions gives

$$i = \sum_k c_k z_k F u_k E = \sum_k c_k \lambda_k E \qquad (37)$$

for the total current density. The molar conductivity of an ion is therefore proportional to its mobility:

$$\lambda_k = F|z_k|u_k. \qquad (38)$$

(The equivalent ionic conductivity, $\lambda_{ke} = \lambda_k/|z_k|$, is occasionally used.)

The molar conductivities of many ions are on the order of 0.01 S m^2 mol^{-1}. For a 100-V-m^{-1} electric field, the corresponding ionic

Table 1. Molar ionic conductivities and electric mobilities of ions at infinite dilution in water at 25 °C (Robinson and Stokes, 1959).

Ion	$\lambda_i^0/10^{-4}$ S m^2 mol^{-1}	$u_i^0/10^{-8}$ m^2 s^{-1} V^{-1}
H^+	349.8	36.25
Li^+	38.6	4.00
Na^+	50.1	5.19
K^+	73.5	7.62
Ca^{2+}	119.0	6.17
La^{3+}	209.1	7.22
OH^-	198.3	20.55
F^-	55.4	5.74
Cl^-	76.4	7.91
Br^-	78.1	8.10
SO_4^{2-}	160.0	8.29
NO_3^-	71.5	7.41

speeds are about 10^{-4} m s^{-1}. Table 1 lists the conductivities and mobilities of selected ions in water. The mobilities of H^+ and OH^- are abnormally high because of proton jumps between the water molecules.

The transference number (or transport number) T_k of ion k is defined as the fraction of the total electric current carried by that ion,

$$T_k = c_k\lambda_k \Big/ \sum_i c_i\lambda_i. \tag{39}$$

If the conductivity and the transference numbers of an electrolyte solution have been measured, the ionic conductivities (and therefore the mobilities) can be calculated. For a salt $M_{\nu_M}N_{\nu_N}$ that dissociates to give ν_M cations M^{z_M} and ν_N anions N^{z_N}, the relations are

$$\lambda_M = T_M\Lambda/\nu_M, \tag{40}$$

$$\lambda_N = T_N\Lambda/\nu_N. \tag{41}$$

At very low concentrations (< 1 mol m^{-3}), the ions are far enough apart that the forces between them are essentially Coulombic. Coulombic interactions cause ionic conductivities to decrease linearly with the square root of the concentration (Onsager and Fuoss, 1932).

Although accurate ionic conductivities cannot be predicted, Stokes's law provides a useful qualitative guide for dilute solutions. The electrical force z_kFE per mole of ion k is assumed to be balanced by the frictional force calculated from Stokes's law for one mole of spherical particles of radius r_k:

$$z_kFE = N_A 6\pi\eta r_k v_k, \tag{42}$$

where η is the viscosity of the solvent. Combining Eqs. (35), (38), and (42) gives

$$\lambda_k = z_k^2F^2/6\pi N_A\eta r_k. \tag{43}$$

This result holds reasonably well for large ions, but only poorly for small ions. Ionic radii calculated from experimental conductivities using Stokes's law tend to exceed crystallographic radii, especially for ions with high charge densities (e.g., Li^+). Dissolved ions are effectively larger because they carry tightly bound solvent molecules held in place by ion-dipole forces.

Equation (43) suggests an inverse relation between ionic conductivities and the viscosity of the solvent. Since liquids become less viscous when heated, ionic conductivities *increase* with the temperature (unlike the electronic conductivities of metals).

2.4 Ion Transport Equations

Ionic conduction in electric fields and the diffusion of ions in concentration gradients often occur simultaneously. This more general case of ion migration is described by Onsager's equations (Onsager and Fuoss, 1932)

$$j_k = -\sum_q l_{kq}\frac{\partial\tilde{\mu}_q}{\partial z}, \tag{44}$$

which relate the ionic fluxes to the gradients in the electrochemical potentials of the ions. The matrix of ionic transport coefficients l_{kq} is symmetric.

The electrochemical potential of ion k is the sum of its chemical potential and its electric potential energy, $z_kF\phi$:

$$\tilde{\mu}_k = \mu_k^r + RT\ln(y_kc_k) + z_kF\phi. \tag{45}$$

For a single electrolyte dissolved in a solvent, the coefficients l_{kq} can be evaluated from measured values of D, Λ, and the transference numbers.

At infinite dilution, the ionic activity coefficients y_k are unity and the off-diagonal coefficients l_{kq} are negligible. In this limit Onsager's equations reduce to the Nernst-Planck equations (Newman, 1973):

$$j_k = -D_k^0\frac{\partial c_k}{\partial z} + \frac{z_kFD_k^0c_k}{RT}E, \tag{46}$$

where D_k^0 is the limiting diffusion coefficient of ion k,

$$D_k^0 = RT\lambda_k^0/z_k^2F^2 = RTu_k^0/|z_k|F. \tag{47}$$

Electrolytes show large departures from ideal behavior, even at very low concentrations. Quantitative applications of the Nernst–Planck equations are therefore limited.

2.5 Experimental Methods

2.5.1 Conductivity Measurements. The electrolyte solution is usually placed between two inert platinum electrodes (Robinson and Stokes, 1959). To minimize impedances due to electron transfer reactions, the electrode surfaces may be coated electrolytically with finely divided catalytic platinum.

Unless special cells are used, the length and cross-sectional area of the conducting solution between the electrodes are not accurately known. To determine l/A, so that resistances can be converted to conductivities, cells are calibrated with solutions of known conductivity, such as aqueous potassium chloride.

An ac Wheatstone bridge is suitable for measuring electrolyte conductivities. The bridge is energized with a 1–5-kHz sine-wave current from an audio-frequency oscillator. The ac impedance of the electrolyte consists of a resistance and a capacitance. A variable resistor and a variable capacitor are adjusted to balance the bridge.

Electrolyte conductivities can also be measured inductively, without electrodes contacting the solution, by using a loop of solution to link the cores of two transformers (Robinson and Stokes, 1959). This technique is useful in applications where metal electrodes would be corroded or fouled by adsorption.

2.5.2 Hittorf Experiment. Hittorf's procedure for measuring transference numbers relies on the changes in concentration near electrodes when direct current is passed through an electrolyte solution (Robinson and Stokes, 1959; Spiro, 1989). Figure 3 gives a sketch of a Hittorf cell. The composition of the electrolyte is initially uniform throughout the cell.

For purposes of illustration, suppose the cations are four times more mobile than the anions, and 5 F of charge are passed through the cell. During the experiment, 5 F of cations plate out onto the negative electrode, while 5 F of anions discharge at the positive electrode. Across any plane in the solution, 4 mol of cations are transferred for every mole of less mobile anions.

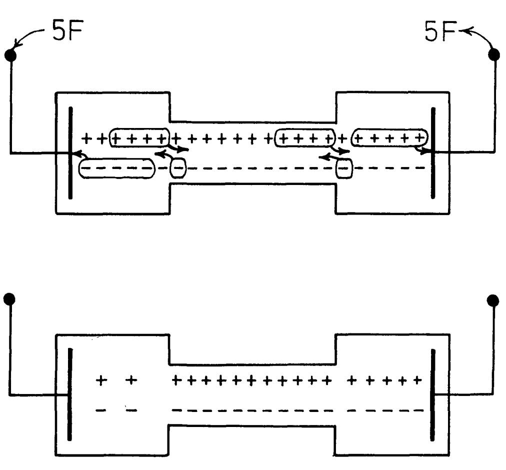

FIG. 3. Sketch of a Hittorf cell before and after the passage of 5 F of direct current. The cations are assumed to be 4 times more mobile than the anions.

After the current is turned off, the solution near the positive electrode is found to contain 4 mol less electrolyte, and the solution near the negative electrode contains 1 mol less electrolyte. The measured changes in concentration give $T_+ = \frac{4}{5}$ and $T_- = \frac{1}{5}$.

2.5.3 Moving-Boundary Experiment. In a moving-boundary cell, the solution is layered over a denser solution of a different electrolyte with an ion in common (e.g., aqueous HCl over aqueous $CdCl_2$). When a direct current is passed in series through the solutions, the boundary between them moves. Surprisingly, the boundary will remain sharp if certain concentrations are used.

By measuring the speed of the boundary, the transference numbers of the upper ion can be determined (Robinson and Stokes, 1959; Spiro, 1989). Suppose the boundary sweeps out volume V during the passage of a total of Q coulombs of charge. The fraction of the total charge carried by the upper ion (1) is $FVCz_1$, and hence $T_1 = FVCz_1/Q$.

3. THERMAL DIFFUSION

3.1 Historical Background

Ludwig was the first to notice diffusion caused by a temperature gradient. He filled a glass tube with a salt solution, heated one end of the tube, and chilled the other. Salt crystals grew in the cooler end, indicating migration of salt to lower temperatures. Soret verified that small but measurable concentration gradients do indeed develop when a temperature gradient is imposed on liquid mixtures. This behavior is called the Soret effect (Tyrrell, 1961; Agar, 1963).

The development of nonequilibrium thermodynamics rekindled interest in thermal diffusion; so did a practical development—the separation of mixtures in thermogravitational columns (Tyrrell, 1961). From the 1950s onward, experimental techniques for measuring the Soret effect were honed to higher accuracy. Reliable data have been obtained for binary electrolyte and nonelectrolyte solutions (Nagashima *et al.*, 1991).

The Dufour effect, the production of a temperature difference by a concentration gradient, is the reciprocal of the Soret effect. The Dufour effect had been measured for gases, but liquids, with their relatively large heat capacities, were not expected to show measurable Dufour effects. Recently, however, the Dufour effect has been measured for liquids (Rowley and Horne, 1980). This advance holds promise for studies of multicomponent liquids.

3.2 Diffusion and Heat Flow in Binary Liquids

When a temperature gradient is applied to a convection-free mixture, heat begins to flow, and the components begin to demix partially. There are many ways to describe such behavior (de Groot and Mazur, 1962; Fitts, 1962). The usual approach is to define first the "reduced" heat flux

$$J'_q = J_q - \sum_{i=1}^{2} J_i H_i, \tag{48}$$

by subtracting from the total heat flux J_q the heat flux arising from the diffusion of matter (Katchalsky and Curran, 1965). H_i is the partial molar enthalpy of substance i.

For thermal diffusion in binary mixtures (one independent diffusional flux of matter), irreversible thermodynamics suggests the coupled equations

$$J_A = -L_{AA}\frac{\partial \mu_A}{\partial z} - \left(\frac{L_{Aq}}{T}\right)\frac{\partial T}{\partial z}, \tag{49}$$

$$J'_q = -L_{qA}\frac{\partial \mu_A}{\partial z} - \left(\frac{L_{qq}}{T}\right)\frac{\partial T}{\partial z} \tag{50}$$

for diffusion, the reduced flow of heat, and their interactions. L_{AA} is the Onsager coefficient for the isothermal interdiffusion of A (Sec. 1.2.2). L_{qq}/T is the thermal conductivity of the solution.

The magnitudes of the Soret and Dufour effects depend on the cross-coefficients L_{Aq} and L_{qA}, respectively. Onsager's reciprocal relation theory holds that $L_{Aq} = L_{qA}$.

The flux of substance A caused by a temperature gradient is roughly proportional to the concentration of that substance. To define a practical transport coefficient that is less sensitive to composition, it is convenient to replace L_{Aq}/T with $C_A D_T$. D_T is called the thermal diffusion coefficient. The flux of substance A can then be expressed as

$$J_A = -D\frac{\partial C_A}{\partial z} - D_T C_A \frac{\partial T}{\partial z}. \tag{51}$$

3.3 Soret Steady State

If convection can be avoided, the concentration gradients induced by thermal diffusion slowly grow in magnitude until they are balanced by interdiffusion in the opposite direction. Heat continues to flow once the steady state is reached, but the diffusional flux of each component is zero.

When the condition $J_A = 0$ is applied to Eq. (51), the result,

$$-\left(\frac{\partial \ln C_A}{\partial T}\right)_{ss} = \frac{D_T}{D} = \sigma, \tag{52}$$

shows that the fractional change in the concentration of A at steady state equals the ratio of the thermal diffusion coefficient to the binary interdiffusion coefficient. The ratio D_T/D is known as the Soret coefficient, σ. (The related quantity $\alpha = T\sigma$ is also used.)

Soret coefficient values lie in the range 0.001 to 0.01 for many liquid mixtures. A 10-K temperature difference typically produces a 1–10% change in concentration. σ does, however, rise to infinity at critical solution points (Haase and Siry, 1968).

3.4 Heat of Transport

To understand thermal diffusion, it is helpful to introduce the heat of transport Q^*, the reduced heat flow per mole of diffusing substance under isothermal conditions:

$$Q^*_A = (J'_q/J_A)_{T\ \mathrm{const}} = L_{qA}/L_{AA}. \tag{53}$$

The heat of transport may be calculated from measured Soret coefficients as

$$Q^*_A = \sigma T C_A \frac{\partial \mu_A}{\partial C_A}. \tag{54}$$

When A diffuses out of a volume element of solution, Q^*_A is the heat that must be absorbed per mole of A to maintain isothermal conditions (Tyrrell, 1961).

3.5 Experimental Methods

Thermal diffusion is usually studied by placing a short column of uniform solution between two metal plates maintained at different temperatures, and then following the subsequent changes in composition induced by the temperature gradient (Tyrrell, 1961; Nagashima, *et al.* 1991). Optical methods can provide accurate results. For electrolytes, thermal diffusion can also be studied by measuring the changes in conductance or potential along the column.

Even small amounts of convection can produce significant errors. One way to minimize this problem is to use a diaphragm cell with a warm upper compartment and a cool lower compartment. Convection can be eliminated by confining diffusion to sufficiently small pores (<15 μm).

Flow cells have also been used to measure thermal diffusion. The liquid mixture is pumped under laminar conditions through a thin (<0.001 m) duct sandwiched between two horizontal metal plates held at different temperatures. A horizontal knife edge bisects the stream into lower and upper portions as it emerges from between the plates. The Soret coefficient is evaluated from the difference in composition between the two streams.

Thermogravitational columns can also be used to measure Soret coefficients, but the analysis is more difficult. A thin layer of solution is sandwiched between two *vertical* surfaces, one of which is cooled and the other heated. Convection causes the denser solution near the cold surface to sink, while the solution near the warm surface rises. Substances that migrate to the warm plate are lofted to the top of the column, while the components that migrate to the cold plate are enriched at the bottom of the cell.

GLOSSARY

Brownian Motion: Random thermal motion of molecules or suspended particles in fluids.

Chemical Potential: Partial molar Gibbs free energy (J mol^{-1}).

Colloidal Particle: Small particle ($\lesssim 0.5$ μm) suspended in a fluid.

Conductance: Reciprocal of the electrical resistance of a conductor (S).

Conductivity: Conductance of a uniform conductor of unit length and unit cross-sectional area (S m^{-1}).

Dufour Effect: Flow of heat caused by concentration gradients.

Electric Mobility: Drift speed of an ion per unit applied electric field strength (m^2 s^{-1} V^{-1}).

Electrochemical Potential: Sum of the chemical and electrical potential energies per mole of an ionic species (J mol^{-1}).

Electrolyte Solution: Liquid or solid that contains mobile ions.

Faraday: Electrical charge per mole of protons (C mol^{-1}).

Flux: Quantity of substance, heat, etc., passing through unit cross-sectional area per unit time (amount m^{-2} s^{-1}).

Interdiffusion: Diffusion between solutions of different chemical composition.

Intradiffusion: Diffusion of labeled (by isotopic substitution or nuclear magnetization) species in systems of uniform chemical composition.

Mobility: Terminal speed of a particle or molecule developed per unit of applied force (s kg^{-1}).

Self-Diffusion: Intradiffusion in systems containing only one form of chemical species.

Soret Effect: Diffusion of matter caused by a temperature gradient.

Thermal Diffusion: See Soret Effect.

Thermogravitational Cell: Device containing a nonisothermal mixture in which both thermally driven diffusion and convection occur.

Tracer Diffusion: Diffusion of a minute amount of species in an otherwise uniform solution.

Transference Number: Fraction of the total electrical current carried by an ion.

List of Works Cited

Agar, J. N. (1963), in: P. Delahay and C. W. Tobias (Eds.), *Advances in Electrochemistry and Electrochemical Engineering,* New York: Interscience.

Crank, J. (1975), *The Mathematics of Diffusion,* 2nd ed., Oxford: Clarendon.

Cullinan, H. T., Jr. (1965), *Ind. Eng. Chem. Fundam.* **4**, 133–139.

De Groot, S. R., Mazur, P. (1962), *Non-equilibrium Thermodynamics*, Amsterdam: North Holland.

Dunlop, P. J., Harris, K. R., Young, D. J. (1989), in: *Physical Methods of Chemistry*, New York: Wiley-Interscience.

Einstein, A. (1905), *Ann. Phys. (Leipzig)* **17**, 549-561.

Fick, A. E. (1855), *Ann. Phys. (Leipzig)* **94**, 59–79.

Fitts, D. D. (1962), *Nonequilibrium Thermodynamics*, New York: McGraw-Hill.

Gibbs, J. W. (1876), *Trans. Conn. Acad. Sci.* **3**, 343–578.

Graham, T. (1850), *Phil. Trans. R. Soc. London* **140**, 1–17.

Haase, R., Siry, M. (1968), *Z. Phys. Chem.* **57**, 56–73.

Harned, H. S., Owen, B. B. (1950), *The Physical Chemistry of Electrolytic Solutions*, New York: Reinhold.

Harris, K. R., Pua, C. K. N., Dunlop, P. J. (1970), *J. Phys. Chem.* **74**, 3518–3529.

Hartley, G. S., Crank, J. (1949), *Trans. Faraday Soc.* **45**, 801–818.

Jost, W. (1960), *Diffusion in Solids, Liquids and Gases*, 2nd ed., New York: Academic Press.

Katchalsky, A., Curran, P. F. (1965), *Nonequilibrium Thermodynamics*, Cambridge, Mass.: Harvard University Press.

Miller, D. G. (1960), *Chem. Rev.* **60**, 15–37.

Nagashima, A., Sengers, J. V., Wakeham, W. A. (Eds.) (1991), *IUPAC Volume on Transport Property Measurement*, London: Blackwell Scientific.

Newman, J. S. (1973), *Electrochemical Systems*, Englewood Cliffs: Prentice-Hall.

Onsager, L. (1931), *Phys. Rev.* **37**, 405–426; **38**, 2265–2279.

Onsager, L., Fuoss, R. M. (1932), *J. Phys. Chem.* **36**, 2689–2778.

Robinson, R. A., Stokes, R. H. (1959), *Electrolyte Solutions*, New York: Academic Press.

Rowley, R. L., Horne, F. H. (1980), *J. Chem. Phys.* **72**, 131–139.

Spiro, M. (1989), in: J. T. Stock, M. V. Orna (Eds.), *Electrochemistry Past and Present*, ACS Symposium Series, Vol. 390, Washington, D.C.: American Chemical Society.

Stokes, G. G. (1903), *Mathematical and Physical Papers*, Vol. 3, London: Cambridge University Press.

Tyrrell, H. J. V. (1961), *Diffusion and Heat Flow in Liquids*, London: Butterworths.

Tyrrell, H. J. V., Harris, K. R. (1984), *Diffusion in Liquids*, London: Butterworths.

Further Reading

Cussler, E. L. (1976), *Multicomponent Diffusion*, Amsterdam: Elsevier.

Cussler, E. L. (1984), *Diffusion*, Cambridge: Cambridge University Press.

Denbigh, K. G. (1951), *The Thermodynamics of the Steady State*, London: Methuen.

Fuoss, R. M., Accascina, F. (1959), *Electrolytic Conductance*, New York: Interscience.

Hamer, W. J. (Ed.), (1959), *The Structure of Electrolytic Solutions*, New York: Wiley.

Hirschfelder, J. O., Curtiss, C. F., Bird, R. B. (1964), *Molecular Theory of Gases and Liquids*, New York: Wiley.

Steele, W. A. (1969), in: H. J. M. Hanley (Ed.), *Transport Processes in Fluids*, New York: Dekker.

DIFFUSION IN BULK SOLIDS

See SOLIDS, DIFFUSION IN

DIFFUSION IN THIN FILMS

D. Gupta, *IBM Research Division, T. J. Watson Research Center, Yorktown Heights, New York, U.S.A.*

Introduction 1
1. **General Aspects of Diffusion** . 2
2. **Diffusion in Polycrystalline Thin Films** 3
2.1 Diffusion Kinetics 3
2.2 Hierarchy of Diffusion and Diffusion Data in Polycrystalline Thin Films 5
2.3 Principles of Modification of Diffusion 5
2.3.1 Modification of Film Microstructure 6
2.3.2 Additions of Solutes 6
2.3.3 Polycrystalline Thin-Film Barriers 7
2.3.4 Sandwich Structures 7
3. **Diffusion in Amorphous Thin Films** 8
3.1 General Aspects of Diffusion in Amorphous Phases 8
3.2 Diffusion in Insulating Amorphous Thin Films 8
3.3 Diffusion in Conducting Amorphous Thin Films 9
4. **Summary** 10
Acknowledgment 10
Glossary 10
Works Cited 11
Further Reading 12

INTRODUCTION

During the last three decades or so, thin films have become essential to advanced technology in a variety of ways. Their use is well known to the scientists and engineers involved in the microelectronic industry. In other industries as well, thin films are used extensively as coatings to control corrosion, wear, friction, hardness, etc. Their applications are very cost effective since specific performance objectives can be met with the use of small amounts of material on relatively cheap substrates. From the microelectronic applications point of view, thin films can be divided into three functional groups.

Electrically conducting film packages are used in Si chip personalization (interconnection) schemes or in subsequent packaging technology (Fried *et al.* 1982; Blodgett and Barbour, 1982). Thin films in these packages invariably have polycrystalline structures and may consist of pure metals, alloys, or intermetallic compounds. To preserve the electrical and chemical characteristics of the thin-film package, thin-film diffusion barriers may be interposed, consisting of either refractory metals or intermetallic compounds. In some recent applications, amorphous thin films have also been considered (Kattelus and Nicolet, 1988). The degradation of the thin-film package takes place by interdiffusion between films or the formation of discontinuities by diffusion-controlled processes such as electromigration, thermomigration, creep and fatigue, etc. (d'Heurle, 1989). Interdiffusion between films may also promote formation of phases with an attendant loss of adhesion.

Dielectric films are also needed in microelectronic thin-film packages between lines used for electrical power distribution and signal propagation. The dielectrics used may be either inorganic compounds, principally the oxides and nitrides of Si (SiO_2, Si_3N_4), or polymeric films such as polyimides that generally have lower dielectric constants and hence are beneficial to prevent noise in dense circuits operating at high speeds (Small and Pearson, 1990; Feigl, 1986). Since dielectric films are always in contact with the conducting lines, control of metallic diffusion and solubility in these films is important to prevent signal shorting and degradation of the dielectric properties.

Finally, semiconducting films are used in Ohmic contacts and Schottky barriers on the

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Si chip itself during the device fabrication process (Katz *et al.* 1990; Wittmer *et al.*, 1986). These films are commonly epitaxial in character. Here, only high-temperature dopant diffusion is relevant, and may be neglected once the transistors have been formed in the Si wafers. In the thin-film applications other than microelectronics, diffusion still plays an important role in developing adhesion and other desirable properties in contact with the substrate. Once again the degradation of the structure involves the formation of undesirable phases, and the development of stresses is controlled by diffusion.

Since diffusion in thin films is highly structure sensitive, the most convenient classification is the one based on the structure itself. As will be seen, such a classification adequately describes the thermal aspects of fabrication and reliability of the thin-film package. In this article, we will describe a hierarchy of diffusion rates due to the various possible defects in polycrystalline thin films and the principles of their modification by alterations of structure or chemistry. Some examples of metallic epitaxial and textured thin films are included to illustrate the role of special boundaries. A brief discussion of diffusion in amorphous conducting and insulating films is provided because they are gaining importance in VLSI (very large scale integration of microelectronic devices) applications. Inclusion of a discussion on amorphous films is appropriate as it is another variation of microstructure of thin films. Comprehensive tabulated lists of diffusion data for the commonly used thin films are provided using the functional classification discussed above. Our discussions will assume largely invariant chemistry during diffusion in thin films with the exception of the presence of small additions of solutes. Consequently, the thin-film systems involving large chemical composition changes such as the ones encountered in heterostructures have not been included. Also, as mentioned earlier, semiconducting films have not been included because of their epitaxial character and freedom from fast diffusion paths such as grain boundaries and dislocations. In view of the limited scope of this article, it has not been possible to cover atomistic aspects of diffusion in thin films, but the interested reader is referred to several publications on this subject (Gupta and Ho, 1988; Nowick and Burton, 1975; Murch and Nowick, 1984).

1. GENERAL ASPECTS OF DIFFUSION

Under the condition of an instantaneous thin-film source of the diffusing species into a semi-infinite solid, the diffusion profile that follows from Fick's second law has a Gaussian mathematical form given by (Shewmon, 1989)

$$c(x,t) = (M/2\sqrt{\pi D t})\exp(-x^2/4Dt), \qquad (1)$$

where M is the total amount of the diffusant per unit area, D is the diffusion coefficient, t is time for diffusion, and $c(x,t)$ is the concentration of diffusant at distance x and time t such that

$$\left(\frac{\partial c}{\partial x}\right)_{(0,t)} = 0.$$

The diffusion profiles are therefore typically linear in plots of ln c versus the penetration distance squared. If diffusion occurs from an infinite source, the profiles are best described by a complementary error function, and the procedure for evaluation of the diffusion coefficient is discussed by Shewmon (1989).

From Eq. (1), the diffusion length $\bar{x}$, which may be defined as that for which $c/M = 1/e = 0.37$,

$$\bar{x}^2 = 4Dt, \qquad (2)$$

is commonly used as a measure of the extent of diffusion in solids.

The temperature dependence of the diffusion coefficient is generally expressed by the Arrhenius relationship:

$$D = D^0 \exp(-Q/kT), \qquad (3)$$

where D^0 is the preexponential factor, Q is the activation energy for diffusion, k is Boltzmann's constant, and T is the temperature in kelvins. The diffusion parameters D^0 and Q implicitly contain a large amount of information on the thermodynamic and kinetic properties of the defects responsible for diffusion. For example, the magnitude of enthalpy and the preexponential term provide clues as to whether vacancies, divacancies, or interstitial atoms are responsible for diffusion. Discussions on the thermodynamic and atomic theories of the diffusion parameters may be found in several references (Nowick and Burton, 1975; Murch and Nowick, 1984). We will use the subscripts a, b, and l on D, D^0, and Q

to denote amorphous, grain-boundary, and bulk lattice diffusion, respectively.

2. DIFFUSION IN POLYCRYSTALLINE THIN FILMS

2.1 Diffusion Kinetics

In polycrystalline thin films, diffusion in the lattice and in grain boundaries generally occurs simultaneously and complicates the kinetics. The coupling of these processes takes place at the grain boundaries, which have an effective thickness δ of the order of the atomic spacing ($\sim 5\times 10^{-8}$ cm). The temperature dependence of the lattice diffusion D_l and that of the grain-boundary diffusion (D_b), together with their respective dimensions—L, the grain size, and δ, the grain-boundary width—determine the overall diffusion profile. It will be seen later on in this section that diffusion processes in general are related to the absolute melting temperature of the host (T_m). Hence, the temperature dependence of diffusion profiles may be described using homologous temperatures (T/T_m) in various materials. We also note here that the problem of diffusion in polycrystalline thin films is identical to the measurement of grain-boundary diffusion in polycrystalline specimens in general. Hence, in the following discussion they have been used in an interchangeable manner.

The interplay of the two atomic diffusion processes in thin films is best described by the three kinetic regimes A, B, and C due to Harrison (1961) that are shown schematically in Fig. 1. Figure 1(a) illustrates A-type kinetics that involves extensive lattice diffusion together with grain-boundary diffusion and mixing across the grains. This regime is typical of diffusion at temperatures above $0.5T_m$ in thin films. Since the diffusion activity is large in both the regions, the diffusion front becomes nearly planar. The diffusion kinetics may resemble that of lattice diffusion, i.e., the diffusion front would advance as a Gaussian with respect to the annealing time t. The measurements would show enhanced apparent diffusivities (D_{app}) with lower activation energies that can be described by Hart's equation (1957):

$$D_{\mathrm{app}}=D_l+fD_b, \qquad (4)$$

where f is the fraction of atoms at the grain boundaries with respect to the grains so that $f=4\delta/L$.

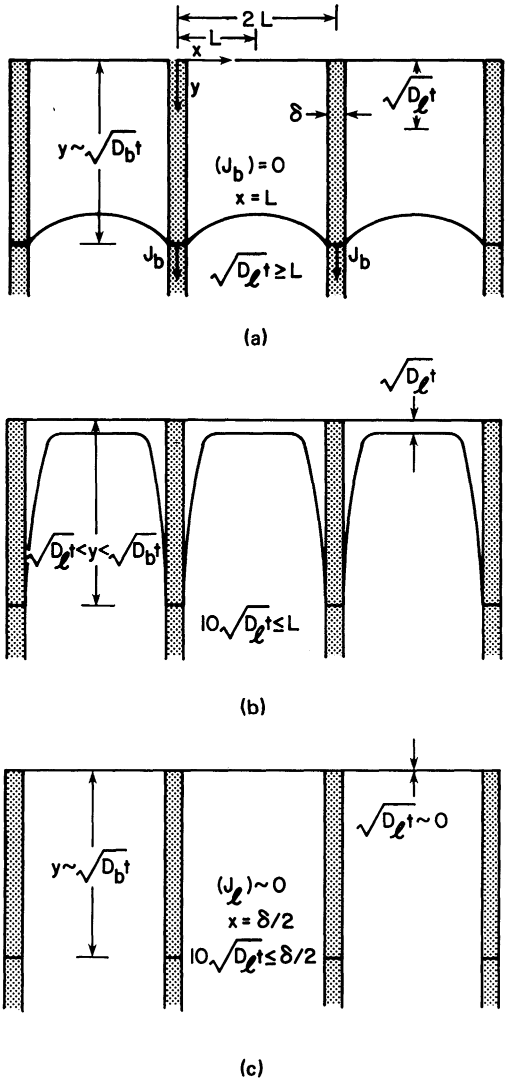

FIG. 1. Schematic representation of (a) type-A, (b) type-B, and (c) type-C kinetics for grain-boundary diffusion after Harrison (1961). The vertical lines indicate the individual grain boundaries of width δ that are perpendicular to the sample surface. The curved lines are isoconcentration contours. The diffusion source coincides for each kinetic regime with the horizontal line.

In the temperature range of $0.3T_m<T<0.5T_m$, B-type kinetics are operative as shown in Fig. 1(b). Here, each boundary is assumed to be isolated, and fast diffusion along the grain boundaries is accompanied by leakage of the diffusant into the grains where lattice diffusion in the *y* direction away from the grain boundaries [Fig. 1(b)] approaches zero.

In the past, almost all experiments on grain-boundary diffusion in metals, alloys, and oxides have been carried out in this regime. It will not be possible to discuss them in detail in this publication. The most commonly used solution is the asymptotic solution due to Whipple (1954) and Suzuoka (1961) and is expressed as

$$P=\delta D_b=0.661\left(\frac{\partial \ln c}{\partial x^{6/5}}\right)^{-5/3}\left(\frac{4D_l}{t}\right)^{1/2}. \quad (5)$$

For metals the grain-boundary width δ is assumed to have a constant value of $\delta=0.5$ nm. Therefore D_b can be determined by plotting the logarithm of the mean tracer concentration c in a thin section parallel to the surface versus $x^{6/5}$. This plot generally contains a straight-line segment in the region where grain-boundary diffusion is dominant and atoms have diffused via grain boundaries into the adjoining grains; for example, see Fig. 2(a) for ^{67}Cu diffusion in polycrystalline W thin films in the range 800–898 °C (Vieregge and Gupta, 1991). To evaluate *P*, a necessary requirement is the knowledge of the bulk diffusion coefficient D_l. In addition, there is a critical parameter $\beta=\delta D_b/2D_l\sqrt{D_l t}$ that must be maintained larger than 30 for Eq. (5) to be valid; otherwise an erroneous activation energy will result (LeClaire, 1963).

Finally, below $0.3T_m$, lattice diffusion is almost frozen out so that only grain-boundary diffusion prevails. This regime is characterized by *C*-type kinetics, shown in Fig. 1(c). This is the only type of kinetic regime offering a straightforward access to D_b. It requires

$$2\sqrt{D_l t}\ll\delta, \quad (6)$$

implying that the diffusing atoms move only inside grain boundaries and lattice diffusion is absent. Analytically, the system can be described in a manner similar to volume diffusion by replacing *D* with D_b in Eq. (1). Therefore, a plot of the logarithm of *c* vs x^2 should yield a straight line the reciprocal slope of which gives D_b:

$$D_b=-\frac{1}{4t}\left(\frac{\partial x^2}{\partial \ln c}\right). \quad (7)$$

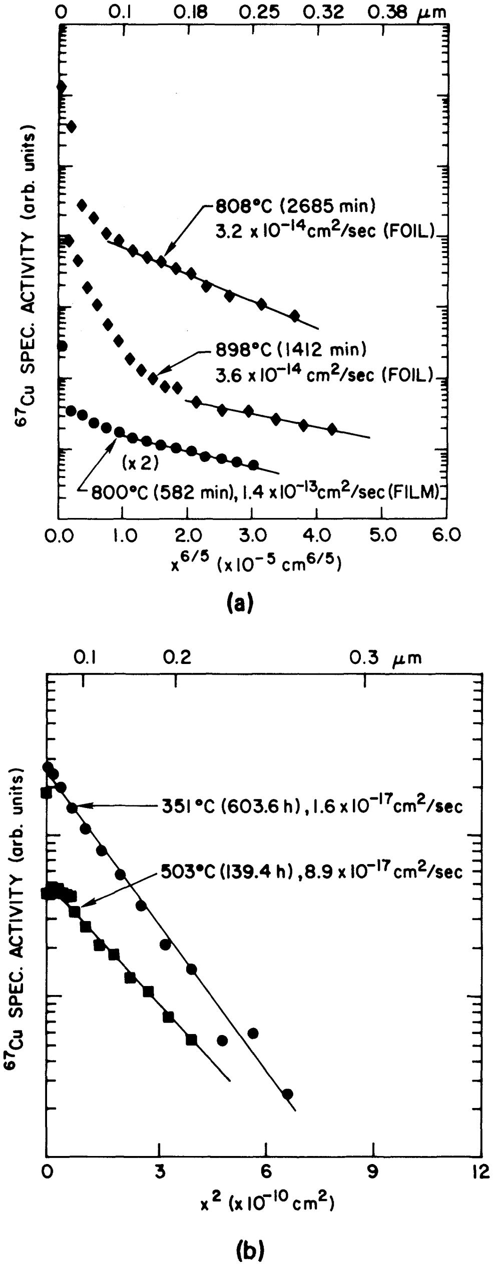

FIG. 2. (a) Penetration profiles for ^{67}Cu diffusion in W films formed by chemical vapor deposition and in sintered W foils. For clarification the profile at 800 °C is multiplied by a factor of 2. The slopes of the linear parts of the profiles are used to calculate *P* in Eq. (5) for type-*B* grain-boundary diffusion regime [after Vieregge and Gupta (1991)]. (b) Penetration profiles of ^{67}Cu diffusion in W films formed by chemical vapor deposition and in W foils at 351 and 503 °C plotted in the penetration distance squared (x^2). The linear plots are strong evidences for the existence of type-*C* kinetics of grain-boundary diffusion in W at temperatures below $0.2T_m$ [after Vieregge and Gupta (1991)].

Because of the constraint imposed by Eq. (6) on the lattice diffusion that must essentially vanish, efforts to study type-*C* kinetics have been largely unsuccessful in the past. Only in recent years has it been possible to investigate type-*C* kinetics unequivocally in fine-grained refractory metals and oxides at temperatures below $0.3T_m$ (Atkinson and Taylor, 1986) by the use of microsectioning techniques. The *C*-type diffusion kinetics is generally encountered only in thin films that have very fine grains. We have recently shown that it is possible to measure grain-boundary diffusion of Cu in the *C* regime in W films formed by the chemical vapor deposition technique at temperatures below $0.3T_m$ (Vieregge and Gupta, 1991). In Fig. 2(b), ^{67}Cu diffusion profiles in polycrystalline W films at 351 and 503 °C are plotted according to Eq. (7). The profiles clearly are linear in the x^2 plots indicating that *C*-type kinetics is, indeed, operative.

At this point, we wish to note that in the case of solute diffusion, the tendency for solute segregation at the grain boundaries has to be taken into account as well. This results in a replacement of the term δ with $s\delta$ in Eqs. (5) and (6), where s is the solute to grain-boundary segregation factor defined by the partitioning of the solute concentration between the lattice, c_l, and the grain boundary, c_b, so that $c_b = sc_l$. The factor s has an inverse Arrhenius temperature dependence compared to the diffusion processes, and it increases as temperature is lowered.

2.2 Hierarchy of Diffusion and Diffusion Data in Polycrystalline Thin Films

In polycrystalline thin films or bulk samples, there are three kinds of defects present that contribute to diffusion: vacancies, dislocations, and grain boundaries. There is actually a hierarchy of diffusion rates along these paths that was first proposed by Gjostein (1973). The lattice diffusion rates due to atom-vacancy exchanges are the slowest, dissociated dislocations have intermediate rates, and high-angle grain boundaries (where the misfit between the adjoining lattices is large) have the fastest diffusion rates. The activation energies for the three kinds of defects can be described, respectively, by $(34, 25, \text{and } 17)T_m$ in calories/mole and the corresponding values for the preexponential terms are of the order of 0.5, 3, and 0.3 cm^2/s. Although the uncertainty is ±20% in the activation energy and an order of magnitude in the preexponential factor, these rules provide reasonable estimates of diffusion rates in systems where no measurements are available. In practice, these parameters lead to measurable diffusion contributions from structural defects even at temperatures below $0.5T_m$ where the normal lattice diffusion is considered to be negligible. In fact, grain boundaries become the dominant diffusion paths in thin films at the temperature of their deposition and subsequent use.

In Table 1, we have compiled diffusion data in thin films of commonly used metals and alloys obtained by various techniques in many laboratories. As mentioned above, in most cases these data are similar to grain-boundary diffusion coefficients measured in the bulk samples. Hence, some bulk materials have also been included. It is not possible to compile all the diffusion data in thin films available in the literature since it is voluminous. For specific systems, the reader is referred to *Defect and Diffusion Forum*, where the results of many diffusion investigations are abstracted on a continuing basis.

Arrhenius plots of some selected systems from Table 1 are shown in Fig. 3. The diffusion coefficients among the various systems are seen to vary by many orders of magnitude. Hence, refractory metals films such as oxygen stabilized β-Ta, prepared by cathodic sputtering (Schwartz and Feit, 1977), and CVD-W may be considered for diffusion barriers against Cu migration from conducting metallizations such as Al-Cu and Cu themselves.

2.3 Principles of Modification of Diffusion

In thin-film technology, diffusion is often detrimental to devices during their fabrication and operation. For example, hillock growth and electromigration failures in thin films have continued to be serious problems in microelectronics (d'Heurle, 1989; Kwok and Ho, 1988). However, ingenious solutions have been found by

1. altering the microstructure of films,
2. addition of solutes,
3. interposing diffusion barriers, and
4. formation of sandwich structures.

These are briefly discussed below.

Table 1. Grain-boundary diffusion parameters in conducting polycrystalline thin films and some bulk specimens.[a]

Matrix solute (wt %)	Diffusant	D_b^0 (cm^2 s^{-1})	Q_b (eV)	Remarks (Poly. films/bulk)
Ag	^{110}Ag	2.6×10^{-2}	0.80	Poly. films
Al	Cu	0.9	1.0	Poly. films
Al	Cu	3.6×10^{-3}	0.87	Poly. films
Al–4% Cu	^{67}Cu	27	1.1	Bulk[b]
Au	^{195}Au	6.2×10^{-3}	0.88	Bulk
Au	Cr	1.0×10^{-3}	1.09	(111) preferred orientation
Au–0.2Co	Cu	4.0	0.56	Poly. films
Au–1.2Ta	^{195}Au	10	1.26	Bulk
Au	Ag	9.6×10^{-5}	1.2	(111) epi. films
Au	Ag	1.0×10^{-6}	1.1	(100) epi. films
Au	Ag	1.0×10^{-5}	0.63	Poly. films
Au	^{195}Au	3.8×10^{-3}	1.16	(100) epi. film
Au	^{195}Au	1.8×10^{-2}	1.0	Poly. films
Au–1.2Ta	^{195}Au	2.0×10^{-2}	1.2	Low angle grain-boundary bulk
Au–0.2Co	Co	1.3×10^{-2}	1.20	Poly. films electrodeposited
Cr	Pt	1.0×10^{-2}	1.69	Poly. films
Cu	Al	0.1	0.94	Poly. films
Cu	Ag	3.2×10^{-6}	0.75	Poly. films
Ni	^{63}Ni	4.4	1.77	10°⟨112⟩ bulk
Ni	^{63}Ni	1.4×10^{-2}	1.08	45°⟨100⟩ bulk
Ni–0.5Co	^{195}Au	0.3×10^{-2}	1.6	Poly. films
Pb	^{203}Pb	0.12	0.46	Poly. films
Pb	^{119}Sn	1.1×10^{-3}	0.46	Foils
Pb	Sn	1.4	0.62	Poly. films
β-Ta	^{67}Cu	9.0×10^{-4}	2.3	Poly. film[b]
$TiAl_3$(0.5Cu)	^{67}Cu	2.7×10^{-3}	1.46	Poly. films[b]
W	^{67}Cu	1.1×10^{-13}	0.46	Poly. films[c]

[a] Unless otherwise mentioned data are from an earlier compilation (Gupta and Ho, 1980).
[b] Vieregge *et al.* (1992a).
[c] Vieregge and Gupta (1991).

2.3.1 Modification of Film Microstructure. It is seen in Table 1 that (100) and (111) epitaxial Au films show a significantly higher activation energy for diffusion compared to the polycrystalline films. This is attributed to the presence of special grain-boundary structure consisting of dissociated dislocations, which is commonly observed in Au. Similarly, films having (111) preferred orientation also have a tendency to show higher activation energy for diffusion compared to the randomly oriented films, due to the preponderance of special boundaries. However, it is almost impossible to produce epitaxy on the substrates commonly employed in devices and other applications. Even films having preferred orientation can only be fabricated in very special circumstances. Therefore, modification of film structure to reduce diffusion has rightly received little attention. Only in the electromigration in submicron-width lines has some success been achieved by producing a bamboo-shaped grain structure. The enhancements in electromigration lifetimes have been possible largely through the elimination of the flux divergence at the triple points of grain boundaries, which are less probable in the bamboo structure (Kwok and Ho, 1988).

2.3.2 Additions of Solutes. Additions of small amounts of solutes to polycrystalline thin films have sometimes resulted in remarkable improvements in performance. For ex-

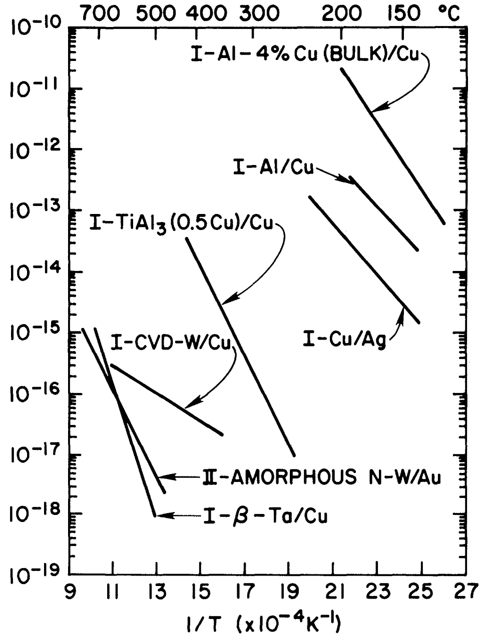

FIG. 3. Arrhenius plots of Au and Cu diffusion in a variety of conducting polycrystalline or amorphous thin films. The large differences in the diffusion coefficients are due to the variable melting temperatures, chemistry, and microstructure of the thin films that may be used as guidance for materials selections for applications such as diffusion barriers. I and II refer to Tables 1 and 2, respectively.

ample, in the Al thin-film interconnects used in VLSI, significant reduction in electromigration damage has been accomplished through additions of 4 wt. % Cu (Ames *et al.* 1970; Rosenberg, 1972). The beneficial effect is attributed to reduction of grain-boundary diffusion through segregation of Cu and formation of an Al-Cu intermetallic compound at the host Al grain boundaries. However, such improvements cannot always be expected as binary alloy systems show all the possible trends, namely enhancement, reduction, and no change of grain-boundary diffusion. In this regard, an empirical correlation between the tendency toward solute segregation and hence grain-boundary diffusion reduction and the shape of the binary phase diagram has been shown to exist (Gupta, 1978). Generally the phase diagrams showing sharp lowering of the liquidus temperature with increasing solute concentration are also prone to solute segregation at the grain boundaries. The data listed in Table 1 seem to substantiate this rule.

2.3.3 Polycrystalline Thin-Film Barriers. The primary purpose of polycrystalline pure metal, alloy, or intermetallic compound films intended for use as diffusion barriers is to block atomic transport in all the possible diffusion paths, namely by vacancies, grain boundaries, and dislocations. Since these processes scale with the melting temperature, the chosen metallization should have a sufficiently high melting point so that even the fastest diffusion process, namely grain-boundary diffusion, is negligible at the device fabrication or performance temperatures. Intuitively, these metals turn out to be transition refractory metals such as Cr, Ni, Ta, Ti, Hf, etc., which also fulfill the above considerations. There are, however, additional requirements for the barrier metals; notably, good adhesion with the adjoining films without excessive stress buildup and, of course, without significant loss of physical properties such as electrical and thermal conductivity. Generally a good compromise can be reached. Pure nickel films are commonly employed in electronic packaging because of inherently low stresses compared to refractory metal thin films. A multilayer low-stress Ti-Cu-Ni-Au thin-film package has been discussed by Morabito *et al.* (1975). Diffusion in pure metal thin-film barriers can be further reduced by incorporating small amounts of solutes to give what have been termed "stuffed barriers" (Kattelus and Nicolet, 1988). As discussed above, in the stuffed barriers the solute tends to segregate at the grain boundaries, thereby blocking the active paths for grain-boundary diffusion.

2.3.4 Sandwich Structures. Howard *et al.* (1978) have investigated Cr, Hf, and Ti intermetallic-compound barriers in Al first-level interconnects to prevent Ohmic and Schottky contact degradation. The Cr, Hf, or Ti metal films of ~1000 Å thickness are deposited as a sandwich with an Al film in between. When heated to 400 °C, CrAl, $HfAl_3$, or $TiAl_3$ compounds are formed depending on the choice of the metal used. These intermetallic-compound barriers are superior to the Cr elemental barrier, and also form structurally coherent electrical contacts without generating excessive stresses. In particular, $HfAl_3$ appears to be an ideal barrier whose effectiveness is not thickness dependent but rather in the fact that diffusion approaches zero. The concept of transition-metal (Cr, Ti, Ta, Hf) intermetallic sandwich structures has been extended to the Al thin-film electromigration problem as well. In the case of 2-μm-wide Al–4% Cu conductors, the Al–intermetallic compound–Al sandwich structures showed improvements of 100-fold in the electromigration lifetimes compared to the Al-Cu system. Again, the $HfAl_3$ and to a lesser extent $TaAl_3$ is the best barrier to prevent electromigration void growth. Diffusion of the radiotracer ^{67}Cu through $TiAl_3$ containing 0.5 wt. % Cu has been recently measured (Vieregge *et al.*, 1992b) and the data are listed in Table 1 and displayed in Fig. 3. The Cu diffusion rates between Al-Cu and $TiAl_3$ (0.5Cu) are seen to differ by a factor of $\sim 10^4$ which directly shows the manner in which transition-metal aluminides are likely to reduce diffusion and improve electromigration lifetimes.

3. DIFFUSION IN AMORPHOUS THIN FILMS

3.1 General Aspects of Diffusion in Amorphous Phases

Amorphous thin films are beginning to receive wider attention for applications in modern technologies because of their unusual properties. Amorphous films, despite their metastable state, have many useful properties below the crystallization temperature. Ionic diffusion in oxide glasses has been studied extensively in the past (Frischat, 1975), and diffusion in organic glasses has also begun to be studied in recent years (Frisch and Stern, 1983). Long-range diffusion in metallic amorphous solids was earlier considered incompatible with the noncrystalline structure due to the onset of crystallization. Measurable amounts of long-range diffusion in amorphous PdSi were first reported by Gupta *et al.* (1975) up to the glass transition temperature without crystallization. Diffusion in metallic amorphous systems has since been investigated extensively in recent years (Mehrer and Dönner, 1989; Kronmüller and Frank, 1989). The thermal energy needed to transform an amorphous structure into a periodic one turns out to be a lot larger than that required for long-range atomic motion of many species.

Because of the absence of periodicity in amorphous phases, it is not possible to have diffusion from defects such as vacancies, grain boundaries, and dislocations in amorphous structure. For diffusion of heavy atoms, the only possible "defect" is the free volume that is generally smaller than vacancy or the atomic volume. Lighter atomic species such as H, however, can diffuse as interstitials. The thermal fluctuation of free volume and the cooperative atomic motion are now considered responsible for diffusion in a disordered structure. Full understanding of the diffusion mechanism, however, is still lacking. In view of the absence of extended defects such as grain boundaries and a single mode of diffusion, amorphous structures are very attractive. Diffusion kinetics in most amorphous phases turn out to lie between those for single crystals and grain boundaries. Hence, attractively small diffusion coefficients can be obtained in some amorphous systems comparable to those in refractory metals for use as diffusion barriers. This may possibly minimize complications due to either high internal stresses or the development of undesirable phases. Amorphous thin films are attractive for VLSI applications as they can be used as dielectrics, Ohmic contacts, and Schottky barriers. Diffusion in these two groups of amorphous films is discussed below.

3.2 Diffusion in Insulating Amorphous Thin Films

The use of glassy or amorphous SiO_2 dielectric films in the microelectronic industry is well known. In recent years, as microelectronic devices have become denser and faster, the need to reduce the dielectric constants in the insulating films has become urgent (Small and Pearson, 1990). Hence other alternatives, both inorganic and organic, are being considered. For example, Si_3N_4, PMDA-ODA (pyromellitic dianhydride-oxydianiline), and BPDA-PDA (biphenyldianhydride-paraphenyldianiline) are being investigated, which also have amorphous structures. Polyimide films, in addition to having smaller dielectric constants, also offer the advantages of ease of fabrication and planarization by spinning and curing. Since the dielectric films always have buried conductors, there is a possibility of metallic dissolution and diffusion. Both processes can be detrimental to the device performance since dissolution would decrease the conductivity and diffusion would impair the dielectric constant leading to shorts if the

Table 2. Diffusion data in amorphous thin films.

Matrix	Diffusant	D_a^0 ($\mathrm{cm^2\,s^{-1}}$)	Q_a (eV)	Remarks (film/foil)
		Insulating amorphous films		
4% PSG	^{67}Cu	4.0×10^{-11}	0.50	Films[a]
PMDA-ODA	^{67}Cu	4.8×10^{-11}	0.42	Films[a]
BPDA-PDA	^{67}Cu	3.4×10^{-10}	0.58	Films[a]
		Conducting amorphous films		
N-W	^{195}Au	5.8×10^{-5}	2.02	18–38 at. % N films[b]
Pd-Si	^{195}Au	—	1.90	20 at. % Si films[c]
Pd-Si	^{110}Ag	2×10^{-6}	1.3	19 at. % Si unrelaxed foils[d]
$PdCu_6Si_{16.5}$	Au	1.5×10^{-4}	1.67	Relaxed films[e]
Gd-Co	^{57}Co	10^{-7}	1.2	16 at. % Gd films[c]

[a] Vieregge and Gupta (1992b).
[b] Dönner *et al.* (1991).
[c] Gupta *et al.* (1982).
[d] Gupta *et al.* (1975).
[e] Chen *et al.* (1978).

distance of line separation is comparable to the diffusion length.

It has been possible to measure diffusion of ^{67}Cu and ^{110}Ag radiotracers in 4 wt % phosphosilicate glass (PSG), PMDA-ODA, and BPDA-PDA dielectric films (Faupel *et al.*, 1989; Faupel *et al.*, 1990; Vieregge and Gupta, 1992b). The measured diffusion coefficients are listed in Table 2. Sizable diffusion of both Ag and Cu is seen to take place. Although the Cu solubility in the dielectric films has not yet been measured, the Cu concentration in the diffusion profiles does indicate that the polyimide films may dissolve an order of magnitude more Cu in comparison to the PSG films. Furthermore, diffusion in the BPDA-PDA and PSG films is comparable, while in the PMDA-ODA films it is substantially larger.

3.3 Diffusion in Conducting Amorphous Films

As mentioned above, amorphous structures are free from extended defects such as grain boundaries and dislocations and hence are attractive as diffusion barriers in Ohmic contacts as well as Schottky barriers. The current status of this class of barriers in semiconductor contacts has been recently reviewed by Kattelus and Nicolet (1988). Amorphous films are deposited by any one of the familiar thin-film techniques, such as electron-gun coevaporation, sputtering, and CVD where vapor quenching is possible. In cases where vapor quenching is not possible, amorphization by ion beam or laser glazing may provide an alternative approach. Recent investigations of solid-state amorphization by gentle annealing of interlaced ultrathin transition-metal films with amorphous Si layers (Herd *et al.*, 1983; Anderson *et al.*, 1983) have added another dimension to this technology. The list of the amorphous films used on semiconductor contacts contains both metal-metal and metal-metalloid systems. The metal-metal systems are typified by the refractory-metal–based alloys, namely $Nb_{53}Ni_{47}$, $W_{35}Zr_{65}$, $Ir_{58}Ta_{42}$, etc. The typical metal-metalloid systems include $N_{21}Ni_{30}W_{49}$, $N_{23}W_{77}$, $Si_{10}W_{90}$, $Si_{10}W_{80}$, etc. These barriers are generally used in thicknesses of ~1000 Å on Si or GaAs substrates against migration of Al and Au conductors.

Because the use of amorphous conducting films in VLSI is rather recent, diffusion measurements of the conductor-metal species are sparse. Most diffusion investigations are in the melt-spun ribbon specimens that have been reviewed by Mehrer and Dönner (1989) and Kronmüller and Frank (1989). Direct diffusion measurements of ^{95}Au radiotracer in amorphous N-W thin films have been carried out by Dörner *et al.* (1991) to elucidate its possible use as a diffusion barrier. The data are listed in Table 2 along with the earlier results of Au and Co radiotracer diffusion in amorphous PdSi and GdCo films by Gupta *et al.* (1975). The Au diffusion data in the N-W thin films are also plotted in Fig. 3. It is seen that they compare favorably with Cu diffusion in β-Ta (Vieregge *et al.*, 1992a), which is considered to be the best diffusion barrier.

4. SUMMARY

In this article diffusion in polycrystalline, epitaxial, and amorphous thin films has been reviewed in conducting and insulating systems. Diffusion in polycrystalline thin films is controlled by their microstructure, and involves fast atomic transport along grain boundaries, and to a lesser extent along dislocations, whereas lattice diffusion due to vacancies within the grains is vanishingly small. Based on the understanding gained of the diffusion processes, it is now possible to provide guidelines for modification and reduction of diffusion in thin films. These involve solute additions that tend to segregate and precipitate at the grain boundaries. In some cases, structure refinement by producing epitaxial and textured films may be possible. An alternative approach is the interposition of thin-film diffusion barriers for which the rules for material selection are discussed. Diffusion data in a large number of polycrystalline metallic systems are compiled in Table 1. Diffusion in several amorphous conducting and dielectric thin-film systems is examined for the first time and the available diffusion data are listed in Table 2. Despite the progress in production of amorphous thin films by many novel techniques, their applications remain largely unexplored with respect to use as diffusion barriers and contacts of various sorts. Direct diffusion measurements in this class of material are much needed to aid in their evaluation for such applications.

ACKNOWLEDGMENT

The author is grateful to Dr. B. S. Berry for valuable suggestions on the manuscript before publication.

GLOSSARY

Activation Energy: Thermal energy needed to accomplish a kinetic process.

Amorphization: A process of producing amorphous structure.

Arrhenius Relationship: A linear semilog plot of a kinetic parameter versus reciprocal of the absolute temperature.

Bamboo Structure: One-dimensional structure where grain boundaries are arranged perpendicular to its length.

Diffusion Barrier: A thin film of metal or an alloy interposed between two conducting films to prevent interdiffusion.

Diffusion Profile: A plot of concentration of the diffusing material versus depth in the specimen.

Electromigration: Diffusion under an electrical field.

Epitaxial Film: A film deposited on a single-crystalline substrate with its orientation related to that of the substrate.

Extended Defects: Defects in solids other than point defects, such as dislocations and grain boundaries.

Intermetallic Compound: A phase formed by two or more metals in specified proportions.

Laser Glazing: Heating of a thin film by a laser with the objective of melting at the surface only to produce an amorphous structure.

Melt Spinning: Quenching of a liquid on a spinning drum to produce amorphous ribbons.

Microelectronic Devices: Microscopic electronic devices consisting of wires, transistors, and resistors.

Ohmic Contact: An electrical contact on a transistor in a Si chip that follows Ohm's law.

Personalization of Si Chip: A scheme of wiring transistors in a Si chip to produce microelectronic devices.

Preexponential Factor: A temperature-independent factor in the Arrhenius relationship.

Preferred Orientation: Grain structure where a certain crystallographic axis is aligned preferentially.

Refractory Metals: Metals having very high melting temperature, in excess of 2500 K.

Sandwich Structure: A scheme in which a film is formed between two films to obtain a specific property.

Schottky Barrier: A *p-n* junction formed with a semiconducting film to prevent electrical and thermal transients.

Solute Segregation: Segregation of impurities at surface of a specimen or at grain boundaries.

Stuffed Barriers: Diffusion barriers containing impurities that segregate at the grain boundaries and prevent diffusion.

Thermomigration: Diffusion under a temperature gradient.

Thin-Film Package: A scheme in which several films are formed to achieve specific objectives.

Vapor Quenching: A method of vapor deposition of amorphous films onto cold substrates.

List of Works Cited

Ames, I., d'Heurle, F. M., Horstman, R. (1970), *IBM J. Res. Develop.* **14**, 461.

Anderson, W. T., Jr., Christou, A., Davey, J. E., in: R. Ludeke and K. Rose (Eds.), *Interfaces and Contacts*, MRS Sym. Proc. 18. North-Holland, New York (1983) p. 57.

Atkinson, A., Taylor, R. I. (1986), *J. Phys. Chem. Solids* **47**, 315.

Blodgett, A. J., Barbour, D. R. (1982), *IBM J. Res. Develop.* **26**, 30.

Chen, H. S., Kimerling, L. C., Poate, J. M., Brown, W. L. (1978), *Appl. Phys. Lett.* **32**, 461.

d'Heurle, F. M. (1989), *Int. Mater. Rev.* **34**, 53.

Dönner, W., Mehrer, H., Pokela, P. J., Kolawa, E., Nocolet, M. A. (1991), *Mat. Sci. Eng.* **B10**, 165.

Faupel, F., Gupta, D., Silverman, B. D., Ho, P. S. (1989), *Appl. Phys. Lett.* **55**, 357.

Faupel, F., Gupta, D., Silverman, B. D., Ho, P. S. (1990), *Proceedings of the Europe Conference on Advanced Material Processes, Aachen,* Oberursel: Deutsche Gesellschaft für Materialkunde, p. 889.

Feigl, F. J. (1986), *Phys. Today* **39**, 47.

Fried, L. J., Havas, J., Lechaton, J. S., Logan, J. S., Paal, G., Totta, P. A., (1982), IBM *J. Res. Develop.* **26**, 362.

Frisch, H. L., Stern, S. A. (1983), *CRC Rev. Solid State Mater. Sci.* **11**, 123.

Frischat, G. H. (1975), *Ionic Diffusion Oxide Glasses,* Aedermannsdorf, Switzerland: Trans. Tech. Pub., p. 27–109.

Gjostein, N. A. (1973), *Diffusion,* Metals Park, OH: American Society of Metals, p. 241.

Gupta, D. (1978), *Electro. Chem. Soc., Symp. Proc.* **78-2**, 498.

Gupta, D., Ho, P. S. (1980), *Thin Solid Films* **72**, 399.

Gupta, D., Ho, P. S. (Eds.) (1988), *Diffusion Phenomena in Thin Films and Microelectronic Materials,* Park Ridge, N.J.: Noyes, p. 1.

Gupta, D., Tu, K. N., Asai, K. W. (1975), *Phys. Rev. Lett.* **35**, 796.

Gupta, D., Tu, K. N., Asai, K. W. (1982), *Thin Solid Films* **90**, 131.

Harrison, L. G. (1961), *Trans. Faraday Soc.* **57**, 1191.

Hart, E. W. (1957), *Acta Met.* **5**, 597.

Herd, S., Tu, K. N., Ahn, K. Y. (1983), *Appl. Phys. Lett.* **42**, 597.

Howard, J. K., White, F., Ho, P. S. (1978), *J. Appl. Phys.* **49**, 4083.

Kattelus, H. P., Nicolet, M.-A. (1988), in: D. Gupta and P. S. Ho (Eds.), *Diffusion Phenomena in Thin Films and Microelectronic Materials,* Park Ridge, N.J.: Noyes, p. 432.

Katz, A., Murarka, S. P., Applebaum, A. (1990), *Mat. Res. Soc. Symp. Proc.* **181**, 3–621.

Kronmüller, H., Frank, W. (1989), *Radiat. Effects Defects Solids* **108**, 106.

Kwok, T., Ho, P. S., in: D. Gupta and P. S. Ho (Eds.), *Diffusion Phenomena in Thin Films and Microelectronic Materials,* Park Ridge, N.J.: Noyes, p. 369.

LeClaire, A. D. (1963), *Br. J. Appl. Phys.* **14**, 351.

Mehrer, H., Dönner, W. (1989), *Defect and Diffusion Forum* **66-69**, 189.

Morabito, J. M., Thomas III, J. H., Lesh, N. G. (1975), *IEEE Trans. Parts, Hybrids, Packaging* **PHP-11**, 253.

Murch, G. E., Nowick, A. S. (Eds.) (1984), *Diffusion in Crystalline Solids,* New York: Academic, p. 1–375.

Nowick, A. S., Burton, J. J. (Eds.) (1975), *Diffusion in Solids: Recent Developments,* New York: Academic, pp. 115–444.

Rosenberg, R. (1972), *J. Vac. Sci. Technol.* **9**, 263.

Schwartz, N., Feit, E. D. (1977), *J. Electrochem. Soc.* **124**, 123.

Shewmon, P. (1989), *Diffusion in Solids, 2nd ed.,* Warrendale, PA: Minerals, Metals and Materials Society, pp. 9–241.

Small, M., Pearson, D. J. (1990), *IBM J. Res. Develop.* **34**, 858.

Suzuoka, T. (1961), *Trans. Jpn. Inst. Met.* **2**, 25.

Vieregge, K., Gupta, D. (1991), in: E. S. Crowson and E. S. Chen (Eds.), *Tungsten and Tungsten Alloys: Recent Advances,* Warrendale, PA: The Minerals, Metals and Materials Society, p. 231.

Vieregge, K., Gupta, D., Spit, F. H. M. (1992a), to be published.

Vieregge, K., Gupta, D. (1992b), to be published.

Whipple, R. T. P. (1954), *Philos. Mag.* **45**, 1225.

Wittmer, M., Stimmell, J., Strathman, M. (Eds.) (1986), *Mat. Res. Soc., Symp. Proc.* **71**, 3–525.

Further Reading

Gafadura, H., *Diffusion and Defect Forum,* Liechtenstein: Vaduz (Journal).

Gupta, D., Ho, P. S. (Eds.) (1988), *Diffusion Phenomena in Thin Films and Microelectronic Materials,* Park Ridge, N.J.: Noyes.

Kaur, I., Gust, W., Kozma, L. (1989), *Handbook of Grain and Interface Boundary Diffusion Data,* Vols. 1 and 2, Stuttgart: Ziegler.

Poate, J. M., Tu, K. N., Mayer, J. W. (Eds.) (1978), *Thin Films: Interdiffusion and Reaction,* New York: Wiley-Interscience.

Shewmon, P. (1989), *Diffusion in Solids,* Warrendale, PA: Minerals, Metals and Materials Society.

DIODES, TUNNEL

See TUNNEL DIODES

DISPERSE SYSTEMS

Makoto Takeo, *Department of Physics, Portland State University, Portland, Oregon, U.S.A.*

Introduction 1
1. **Typical Disperse Systems** 2
1.1 Classification of Two-Phase Disperse Systems 2
1.2 Importance of the Surface for Small Particles 2
2. **Determination of Particle Sizes** 2
2.1 Microscopy, Chromatography, and Sedimentation 2
2.2 Light Scattering 4
2.3 Frequency Broadening and Photon Correlation in Light Scattering 5
3. **Emulsions** 6
3.1 Coarse Emulsions and Microemulsions 6
3.2 Stability of Coarse Emulsions and Microemulsions 6
3.3 Factors Determining Types of Emulsions 7
3.4 Electrical Conductivity of an Emulsion 8
4. **Foams** 8
4.1 Bubble Foams and Concentrated Foams 8
4.2 Stability of Foams 8
4.3 Antifoamers and Foam Inhibitors 9
5. **Colloidal Sols and Suspensions** 9
5.1 Preparation of Colloidal Sols and Suspensions 9
5.2 Colloidal Stability, Flocculation, and Ionic Charges 9
5.3 Zeta Potential 10
6. **Gels** 10
6.1 Structure of Gels 10
6.2 Applications of Gels 10
7. **Aerosols** 11
7.1 Nature and Characteristics of Aerosols 11
7.2 Preparation of Aerosols 11
Glossary 11
Works Cited 12
Further Reading 13

INTRODUCTION

Disperse systems are quite common in nature and in various industries. Examples are aerosols, cement, dyestuffs, emulsions, foams, ink, paint, alloys, paste, paper, plastics, rubber, soil, and dust. Examples of colloidal processes include grinding, fuel carburetion, ion exchange, lubrication, food processing, and pharmaceutical and cosmetic preparation.

Simple disperse systems are two-phase systems, composed of a dispersed phase (small particles or the like), and a continuous dispersion medium. The particles may also be called colloids. The International Union of Pure and Applied Chemistry defines a colloid as any material with at least one of its three dimensions in the range of a nanometer (nm) to a micrometer (μm).

Before the turn of this century, the study of disperse systems did not appear to be more than a descriptive subject. The turning point was the demonstration, made by Sientopf and Zsigmondy in 1903, that particles of certain gold sols were definitely revealed in the ultramicroscope.

In 1923, Svedberg invented an ultracentrifuge for finding particle sizes and molecular weights. Langmuir's concept of adsorption, Einstein's clarification of diffusion of particles in medium, and the ideas of dispersion forces and electric double layers, etc., all contributed to the quantitative understanding of colloids.

The recent development of laser and computer technologies has promoted dynamic methods of investigation for these systems,

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

covering a wide particle-size range, dynamic behavior, and the commercial development of various user-friendly devices for monitoring sizes in industry, as well as improved resolution.

In this article, discussion starts with examples of disperse systems. In addition to classical methods of size determination, a new, powerful approach of dynamic light scattering (photon correlation spectroscopy) is discussed. Subsequent sections deal with various important physical properties of emulsions, foams, sols, gels, and aerosols. We note that aerosols are discussed in more detail in a separate article (see AEROSOLS). In the present article, we have omitted detailed experimental procedures and some important topics, such as biological and environmental suspensions.

1. TYPICAL DISPERSE SYSTEMS

1.1 Classification of Two-Phase Disperse Systems

The main two-phase colloidal dispersions are listed in Table 1.

1.2 Importance of the Surface for Small Particles

A characteristic feature of colloidal dispersions is the large ratio of surface area to volume for the particles involved. As a result, the surface energy is relatively large compared to that of the bulk, making the two-phase disperse system thermodynamically unstable with respect to separation into two bulk phases.

If the dispersion medium wets the particle, the colloid is called lyophilic (solvent loving) and behaves like a solute in solvent; otherwise the colloid is called lyophobic (solvent fearing). Lyophobic colloids may form an unstable disperse system. For instance, oil and water can be vigorously mixed to form a nontransparent, heterogeneous mass; but, on standing, the mixture will separate into two clear, homogeneous layers.

There are surface-active substances that can change a lyophilic to a lyophobic interface or vice versa. Such molecules are called surfactants. An example is a soap. Soap molecules have two active parts: the hydrophilic (water-loving) part, which is an ionic group with a strong electrostatic attraction to water molecules, and another part (hydrocarbon), which is hydrophobic (water fearing) but is lyophilic and wets oil. Thus, if soap molecules cover the surface of an oil drop in water, the hydrophobic colloidal system becomes a hydrophilic colloidal system. If colloidal particles are solvated, then the solvent, immobilized at the surface, must be treated as a part of the particle.

2. DETERMINATION OF PARTICLE SIZES

2.1 Microscopy, Chromatography, and Sedimentation

The particle size is often defined by the radius. However, there are other ways to represent the particle size, such as molecular weight and particle mass. We must note that particles are in general nonspherical, that the definition of size is sometimes conventional,

Table 1. Examples of two-phase systems

Dispersed phase/ medium	Name	Examples
Liquid/gas	Liquid aerosol	Fog, liquid spray, mist
Solid/gas	Solid aerosol	Smoke, dust
Gas/liquid	Foam	Soap foam
Liquid/liquid	Emulsion	Milk, mayonnaise
Solid/liquid	Sol, suspension	Au sol, paste, gel
Gas/solid	Solid foam	Expanded polystyrene
Liquid/solid	Solid emulsion	Opal, pearl
Solid/solid	Solid suspension	Alloy, pigmented plastics

and that the word "molecule" is used here loosely, implying sometimes an aggregate of molecules. A few examples are given in the following.

Optical and electron microscopes are useful for determining geometrical sizes and shapes of particles and the size distribution. Because of diffraction, however, optical microscopes cannot be used for particles smaller than about 1 μm. Thus an electron microscope (transmission or scanning) is important for particle imaging, provided particles have a small vapor pressure and can be put in vacuum without changing their characteristics. In some cases, a colloidal suspension may be freeze-fractured and a surface replica is then used for observation.

Lyophilic colloids form a stable solution, and so one of the methods available for the size determination is chromatography (Yau *et al.*, 1979). When a solution is injected into a stream of liquid flowing through a column packed with small solid spheres, the particles separate according to size, with the largest particles eluting first. The size is calibrated with a colloidal solute of known molecular weight, or measured optically. Thus, some colloids are characterized by molecular weight, which can be easily converted to their geometrical sizes if the density is known. Chromatography provides a rapid method for fractionating polydisperse systems and is capable of size separation of particles between 20 nm and 1 μm (see CHROMATOGRAPHY).

Sedimentation is another useful method for determining molecular weights, but the densities, ρ_1 and ρ_2, of the dispersion medium and colloidal particles, respectively, must differ. In a steady state, the force **F** of buoyancy acting on the particle under gravity or in a centrifugal field is given by

$$\mathbf{F}=V(\rho_2-\rho_1)\mathbf{a}. \qquad (1)$$

This force is balanced by the frictional force against the particle moving in the medium if particles are electrically neutral. Here V is the volume and **a** is the acceleration of the particle. If the experiment is made under the force of gravity, **a** is the gravitational acceleration. When a centrifuge is used with an angular velocity ω, the magnitude a of **a** is ω^2R for a particle at a distance R from the axis of rotation.

The frictional force is proportional to the velocity **v**. The proportionality constant is usually assumed to obey Stokes's law,

$$f=6\pi\eta r_0, \qquad (2)$$

where η is the viscosity of the medium and r_0 is the radius of the particle. If Eq. (2) is assumed for nonspherical particles, then r_0 is called an equivalent radius.

Equation (2) must be modified for aerosols. When a small particle moves in a gas, its collision with gas molecules is intermittent and discontinuous, and the value of the frictional proportionality constant f is reduced (Vold and Vold, 1983). Historically, this effect was first noted in the rate of fall of the colloidal drops in the famous Millikan experiment on the electronic charge.

Because of this friction, a particle under force **F** almost instantaneously reaches a terminal speed v. Since the mass of the particle $m = \rho_2V$, we can then write

$$m(\rho_2-\rho_1)a=\rho_2fv. \qquad (3)$$

This equation still holds for aerosols if the corrected f is used.

The speed v in Eq. (3) may be measured by using a laser Doppler velocimeter or anemometer to estimate the mass m. Alternatively, the speed may be measured under an ultramicroscope with a timer.

Gravity sedimentation is not feasible for particles below a few micrometers, because of disturbances of Brownian motion and thermal gradient. It is easy to overcome this difficulty by using a centrifuge, since the acceleration can be easily increased by a factor of 10^5. From Eq. (3), since $a = \omega^2R$ and $v = dR/dt$, the time for the particle to travel radially from R_0 to R is given by

$$t=\frac{f\ \ln(R/R_0)}{m(1-\rho_1/\rho_2)\omega^2}. \qquad (4)$$

The symbol ln stands for the natural logarithm. The equation may be used to obtain the value of m, provided the effect of diffusion is small.

Experimentally, a suspension is injected at R_0 into a centrifuge while it is spinning with the medium. Assuming that the dispersed particles instantly attain the terminal velocity, the time t is then measured by optically detecting them at $R > R_0$. Thus the rotor is usually made transparent.

In some cases, the frictional coefficient f is not known. Then the method of sedimentation equilibrium may be useful. With a centrifuge, before equilibrium sets in, the particles move outward and build up a radially outward gradient of the particle concentration. The gradient, on the other hand, causes inward diffusion of particles. When equilibrium sets in between them, the particle concentration is constant in time and is given by

$$\ln\frac{c_2}{c_1}=\frac{m}{2k_BT}\left(1-\frac{\rho_1}{\rho_2}\right)\omega^2(R_2^2-R_1^2), \tag{5}$$

where c_i is the concentration at R_i $(i=1,2)$. The symbols k_B and T are respectively Boltzmann's constant and the absolute temperature. The concentration can be measured by attenuation of transmitted light. For this purpose, a transparent, hollow, plastic disk is conveniently used as a rotor.

Recently, sedimentation field-flow fractionation techniques have been developed for high-resolution studies (Caldwell, 1984). They combine the technique of flow separation used in hydrodynamic chromatography with the technique of size separation in centrifugation.

2.2 Light Scattering

There are many other methods applicable for size observation (Provder, 1987; Barth, 1984; Ross and Morrison, 1988). One of them employs scattering of light by colloidal particles and is useful for emulsions, aerosols, air pollution, and microbiological research, sensing the onset of precipitation or nucleation, etc. If particles do not absorb incident light, but have refractive indices distinct from the medium, they scatter incident light into various directions and accordingly reduce the transmitted intensity. The time-averaged attenuation obeys the exponential law

$$I(x)=I_0\exp(-\tau x), \tag{6}$$

where the x axis is in the direction of the incident light and I_0 is the intensity of light at $x=0$. In the exponent, τ is called the turbidity or attenuation.

An exact expression for τ can be derived from a general theory of light scattering. It is not simple, but its limiting cases of Rayleigh scattering and diffraction are easy to handle. The limiting cases can be classified in terms of the particle size r_0 and the wavelength λ of the light as

$r_0<\lambda/20$, Rayleigh scattering;

$r_0>2\lambda$, diffraction theory.

Rayleigh scattering has been widely applied to liquid and solid aerosols in the atmosphere. Brownian motion of the particles causes the particle concentration to fluctuate locally in time, generating local fluctuations in the index of refraction of the solution. On taking the statistical average, the turbidity is given by (Hiemenz, 1986)

$$\tau=\frac{32\pi^3n_0}{3N_A\lambda_0^4}\left(\frac{dn}{dc}\right)^2\frac{1}{(1/cm)+2B'} \tag{7}$$

where c, m, and B are, respectively, the concentration, the mass of solute particle, and the second virial coefficient of the osmotic pressure of the solution. N_A is Avogadro's number, n_0 is the refractive index of the solvent, n is that of the solution, and λ_0 is the wavelength *in vacuo* ($\lambda_0 = n\lambda$). Equation (7) gives a particle mass irrespective of the shape for small particles. For a polydisperse systems, Eq. (7) provides a weight-averaged mass.

If the particle size increases and is larger than $\lambda/20$, scattering into a certain direction from different points of a single particle is out of phase, causing destructive interference and reduced intensity, but phase differences are small for small scattering angles, causing a greater intensity of scattered light at low angles. The overall effect gives rise to maxima and minima in the scattered intensity depending on the scattering angle, with greatest intensity in the forward direction. This behavior depends on the particle size, and the scattering for sizing of spherical particles can be treated with help of the rigorous Mie theory, which was developed in 1809 (Garcia-Rubio, 1987). With the advent of high-speed computers, the treatment has advanced rapidly. Later Kerker (1969) extended the theory to include cylinders and stratified spheres. These theories must be used whenever the particle size is large.

However, if the particle size is larger than about 2λ, the light scattered with large angles becomes very weak because of destructive interference and only the forward scattering is pronounced. This type of scattering is customarily called diffraction.

For this diffraction regime ($r_0 > 2\lambda$), a method has been described by Weiner (1984) to cover observations of spherical particles larger than 1 μm. Since the diffraction pattern spreads more for smaller particles but is narrower in the forward direction for larger particles, the intensity distribution in the diffraction pattern can be used to measure particle sizes in either monodisperse or polydisperse systems.

2.3 Frequency Broadening and Photon Correlation in Light Scattering

Another technique, which has recently attracted widespread attention, is to use temporal fluctuations of scattered intensity by particles in Brownian motion directly (Pecora, 1985). It has also been applied to measurements of interfacial tension (Wielebinski and Findenegg, 1988).

If particles are spherical, the electric field of the light scattered by them behaves, at the detector, as

$$E(t)=E_0 \sum_j \exp\left(i \int_0^t \mathbf{q}\cdot\mathbf{v}_j(t)dt - i\omega t\right), \quad (8)$$

where $i = (-1)^{1/2}$ and $\mathbf{v}_j(t)$ is the velocity of the jth particle at time t. The vector $\mathbf{q}$ is the scattering vector, $\mathbf{q} = \mathbf{k}_0 - \mathbf{k}$, where $\mathbf{k}_0$ and $\mathbf{k}$ are wave vectors of incident and scattered light. In Eq. (8), $\mathbf{v}_j$ is constantly changing in time for all j, because particles are undergoing thermal motion and colliding both with molecules of the medium and with other particles. At a collision, $\mathbf{v}_j$ abruptly changes in its direction and/or magnitude, and so $\mathbf{q}\cdot\mathbf{v}_j$ does also. The phase of the oscillation of the jth term in Eq. (8) accordingly shifts abruptly.

The phase shift at a single collision is sometimes large compared with unity. However, it is usually small, but adds up to unity or even more after many collisions. When a large phase shift occurs or the phase shift accumulates to about unity, the incident long wave train is divided into two incoherent shorter wave trains (Ch'en and Takeo, 1957). If this collisional effect occurs continuously, the scattered light is composed of many wave trains of random length. The shortening of the wave trains results in frequency broadening in the frequency domain, and the unpredictable redistribution of incident photons among the divided wave trains results in a temporally fluctuating scattered intensity.

There is a contribution to the intensity fluctuation from the interference of light scattered by Brownian particles. But it is small if they are randomly distributed or the solution is dilute.

The intensity of light $I(t)$ at the detector is the squared absolute value of Eq. (8). Its fluctuation is handled statistically in terms of the autocorrelation function,

$$G(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} I(t)I(t+\tau)dt. \quad (9)$$

For a monodisperse system, we have

$$G(\tau) = e^{-2Dq^2\tau}, \quad (10)$$

where the normalization is so chosen that the coefficient is unity. D is the diffusion coefficient, given, for a sphere, by

$$D = k_B T/6\pi\eta r_0, \quad (11)$$

and q is the magnitude of the scattering vector $\mathbf{q}$, given by

$$q = \frac{4\pi}{\lambda} \sin\frac{\theta}{2}, \quad (12)$$

where θ is the scattering angle. Since molecules of the medium also scatter the incident light with no correlation, the autocorrelation function contains an additional constant term. Interactions among particles affect the fluctuation of the scattered intensity, so that the interactions can be observed by this method (Pusey and Tough, 1985).

For a polydisperse system,

$$G(\tau) = \int_0^\infty F(\Gamma)e^{-\Gamma\tau}d\Gamma, \quad \Gamma = 2Dq^2, \quad (13)$$

where $F(\Gamma)$ is the normalized distribution function of the particle size parameter Γ. Thus, the autocorrelation function is a Laplace transform of the distribution function.

Experimentally, the intensity fluctuation is measured by counting the number of photons arriving at the detector per unit time. Then Eq. (9) is used to calculate the autocorrelation function on a computer. The result is compared with either Eq. (10) or Eq. (13). We note that, if Eq. (13) is expanded into a power series in τ, the coefficients are the moments of the size distribution function. However, since particles that strongly scatter light contribute more to the observed intensity, the distribution function obtained from the experimental $G(\tau)$ is intensity-weighted. Mie theory may be used to correct the effect, since it gives the scattered intensity as a function of the particle size.

Instead of intensity fluctuation, collisional frequency broadening may be observed by using a spectrometer or an interferometer. In the observed power spectrum, the full width at the half height of the peak intensity is $2Dq^2$ for monodisperse systems of spherical particles. Then Eq. (11) gives r_0.

We note that the power spectrum is the Fourier transform of the autocorrelation function of the fluctuating electric field of the scattered light, instead of the fluctuating light intensity.

If a particle is anisotropic, the particle rotation due to collisions is important, since it causes reorientations of the polarizability tensor and depolarization of the scattered light. The effect is also an intensity fluctuation in the time domain and a frequency broadening in the frequency domain (Pecora, 1985).

Microscopic observations of Brownian motion of the particles reveal an ever-increasing irregularity with reduction of the time interval of observation. A Brownian pathway obtained by joining, with straight lines, successive positions of a particle marked at a given time interval forms a random zig-zagged pattern. Although the scale of the time interval changes, the general feature of the pattern remains the same. This suggests a fractal behavior (see FRACTAL GEOMETRY). In fact, the concept of a fractal dimension has proved to be useful in studying colloids (Zrinyi *et al.*, 1988).

3. EMULSIONS

3.1 Coarse Emulsions and Microemulsions

An emulsion is a disperse system in which the phases are immiscible or partially miscible liquids. The dispersed particles are usually between 0.1 and 10 μm in diameter. If they are between 0.01 and 0.1 μm, the systems are called microemulsions. There is a visual distinction: microemulsions are clear and coarse emulsions are cloudy, because of difference in light scattering.

The practical application of emulsions is very active and important in dairy products, pharmaceutical preparations, cosmetics, the petroleum industry, etc. In many cases, oil will be the dispersed phase and water the continuous phase, designated O/W, or water may be dispersed in oil (W/O). An example of an O/W emulsion is milk, and for a W/O emulsion is butter. Compatibility with either oil or water upon dilution is an easy way of identifying which phase is continuous. Another test for emulsion type is the electrical conductivity—O/W emulsions almost always have high conductivity compared to W/O emulsions.

3.2 Stability of Coarse Emulsions and Microemulsions

The most important physical property of an emulsion is its stability. The state of emulsions can change in various ways; for example, an emulsion may change its type from W/O to O/W or vice versa (inversion). It may also separate into its two constituent phases by coalescence (breaking). Because a density difference may exist between the two phases, the dispersed phase may rise or sink within the medium due to gravity or centrifugation (creaming or settling). The rates of such changes are influenced by various parameters, such as temperature, size distribution of droplets, and the phase volume ratio. Additional factors are described below.

Collisions among droplets, often leading to coalescence, result in emulsion breaking. Consequently, a reduction in collision frequency is effective in promoting stability. This reduction is accomplished by repulsion if droplets are electrically charged with the same sign or by attenuated motion if the viscosity of the medium is high. Droplets covered by mechanically strong and elastic films do not coalesce during collisions.

The free energy (Helmholtz) for the process of coalescence of droplets at constant volume, temperature, and composition is

$$\Delta F=\gamma(A_{\text{final}}-A_{\text{initial}}), \tag{14}$$

where γ is the surface tension and A is the total surface area. Since A in general decreases with coalescence, ΔF is negative, in favor of spontaneous coalescence. For good stability, ΔF must be positive upon coalescence, requiring additional positive terms in Eq. (14).

If an emulsion is prepared by homogenizing two pure liquid components, phase separation (breaking) will usually be rapid. To prepare reasonably stable emulsions, a third component—an emulsifying agent (emulsifier)—is needed in the process of homogenizing. The emulsifier must itself have a negative free energy of adsorption, so that it spontaneously arrives at the interface of droplets. Then if droplets coalesce, A decreases and some of the adsorbed emulsifier is desorbed or removed from the interface at the expense of positive energy, leading to a positive change in ΔF.

Assessment of the stability of an emulsion against coalescence involves droplet counting by means of, say, a microscope or a Coulter counter. In the latter, a dilute solution is made to flow through a small orifice in which electrodes are placed. An electric pulse occurs each time a particle passes between the electrodes. The counting of the pulses gives the particle number, provided there is no "coincidence" (Vold and Vold, 1983). The pulse height depends on the particle size, and if a pulse height discriminator is added, the apparatus can yield a size distribution. However, this technique cannot be used for particles smaller than 0.7 μm.

Although more important in sols, we must note here another mechanism for reducing the particle number density. It is the process called flocculation, aggregation, or coagulation, in which small particles clump together, like a bunch of grapes, due to interaction forces among particles. This phenomenon has been treated, for quantitative discussion of colloidal stability, by the DLVO theory, developed by Deryagin and Landau and by Verwey and Overbeek (and named by their initials), and by improved theories (Tadros and Vincent, 1983). They calculated the energy changes that take place when particles approach one another under mutual interactions between them, and successfully applied the model calculation to the study of stability. Flocculation has been experimentally observed by photon correlation spectroscopy with help of the Mie theory.

Suppose we add a fourth component to an emulsified system. The fourth component is called a coemulsifier, generally an alcohol of short chain length, such as hexanol. Then a transparent and stable microemulsion forms spontaneously in the region where the ratio of components satisfies a certain condition that has been experimentally determined. The coemulsifier, by combining with the emulsifier, achieves a large energy of desorption and a small interfacial tension, which may result in spontaneous formation of smaller droplets. It is also known that certain hydrophobic emulsifiers (nonionic surfactants) may require no coemulsifier for producing microemulsions (Friberg and Venable, 1983). On the other hand, coarse emulsions usually do not form spontaneously. Thus, whether microemulsions should be regarded as true emulsions or as swollen aggregates of surfactant ions (swollen micelles) is a matter of controversy (Hiemenz, 1986). Nevertheless, this technology has opened new applications in countless areas, such as oil recovery from porous rocks and floor waxes.

3.3 Factors Determining Types of Emulsions

In coarse emulsions, what makes W/O or O/W emulsions when two liquids are emulsified? Clearly, the favorable configuration must be that in which the largest amount of emulsifier is accommodated at the interface. In other words (Bancroft's rule), the phase in which the emulsifying agent is more soluble tends to be the continuous dispersion medium. For instance, a water-soluble substance such as sodium oleate emulsifies oil in water. The rule, however, may break down for higher concentrations of the phase that is expected by the rule to form droplets. Emulsifiers are usually soaps or nonionic surfactants.

In addition to surfactants, effective resistance to coalescence of emulsion droplets is conferred by certain finely divided solids. If the solid is preferentially wetted by one phase, then more particles can be found at the interface if that phase forms a continuous medium surrounding drops of the other phase. The moist surfaces of inorganic salts are predominantly hydrophilic and so tend to promote O/W emulsions, while the surfaces of carbon black and resins are predominantly lipophilic and promote W/O emulsions.

It is desirable, in practice, to be able not only to predict the type of an emulsion but also to suggest which surfactant gives maximum stability in emulsification. An attempt at prediction was initiated in 1949 by using Bancroft's rule. This rule assigns a semiempirical scale, called HLB (hydrophile–lipophile balance) numbers, to surfactants. A value of 1 was assigned arbitrarily to oleic acid (predominantly lipophilic) and a value of 20 to sodium oleate (predominantly hydrophilic). The scale (HLB) is constructed for their mixture by

$$\mathrm{HLB}=1\ W_{\text{oleic acid}}+20\ W_{\text{sodium oleate}}, \qquad (15)$$

where W is the weight fraction. This mixture is used as a reference for assigning the HLB number to other emulsifiers by comparing the abilities in forming a W/O or O/W emulsion out of a 50/50 mixture of white oil and water. The HLB numbers have been tabulated (Shinoda and Kunieda, 1983). Those that form W/O emulsions have a low HLB number (3.5–6), and those that form O/W emulsions have a high number (8–18). A good detergent has an HLB number between 13 and 15 and a solubilizer between 15 and 18. The HLB number of a molecule can also be calculated by referring to hydrophilic and lipophilic groups that the molecule carries (Vold and Vold, 1983).

Determination of HLB numbers is usually tedious. But there are various empirical formulas that may be available for nonionic surfactants (Shinoda and Friberg, 1986). One of them is

$$\mathrm{HLB}=20\ W_h/(W_h+W_l), \qquad (16)$$

where W_h and W_l are molecular weights of the hydrophilic and lipophilic moieties of the nonionic surfactant. The HLB number of a nonionic surfactant depends on temperature, although that of an ionic surfactant alone does not change appreciably with temperature around 25 °C.

The HLB concept is fairly satisfactory, but its major disadvantage is that it makes no allowance for temperature, in particular, for ionic surfactants. In fact, an emulsifier can, if nonionic, change at a certain temperature from an O/W to a W/O stabilizer or vice versa. The temperature is called the phase inversion temperature (PIT). The HLB number and the PIT are correlated (Shinoda and Friberg, 1986).

In some practical cases, processes for emulsion instability are desirable. Examples are the creaming, breaking, and inversion of milk to obtain butter and the breaking of W/O oil-field emulsions. Techniques to accelerate breaking include centrifugal separation, freezing, distillation, and filtration. The addition of O/W-promoting emulsifiers may be used to break W/O emulsions, and vice versa.

3.4 Electrical Conductivity of an Emulsion

The electrical conductivity of an emulsion can be used to follow changes in emulsion composition and emulsion type, including inversion. The electrical conductivity of a two-phase system, in which the conductivity of the continuous medium is σ_1 and that of the dispersed phase is σ_2, is given by

$$\sigma=\sigma_1\frac{2\sigma_1+\sigma_2-2(\sigma_1-\sigma_2)V}{2\sigma_1+\sigma_2+(\sigma_1-\sigma_2)V}, \qquad (17)$$

where V is the volume fraction of the dispersed medium. This equation was used to identify experimentally the emulsion inversion as the volume fraction was increased (Manegold, 1952)—an exception to the HLB rule.

4. FOAMS

4.1 Bubble Foams and Concentrated Foams

A foam is a coarse dispersion of gas in liquid (or solid). There are two types: bubble foams and concentrated foams. The former consist of nearly spherical bubbles separated by a bulk liquid, while the latter are mostly polyhedral gas cells partitioned by thin liquid films (exceptions include foam rubber, where the cells are spherical).

4.2 Stability of Foams

Pure liquids do not form stable foams. Thus, certain solutes (foaming agents) are necessary to achieve stability by retarding foam rupture. Good emulsifiers are usually also good foaming agents.

When a large number of spherical bubbles are produced in a medium, a foam is eventually produced, as in a glass of beer. Liquid between the bubbles is drained by gravity, as they rise, and their shapes become polyhedral when in contact through a continuous thin liquid film. In the polyhedron, some portions of the film are flat, where two bubbles meet, and others are curved, where more than two bubbles meet (plateau borders). The pressure of the liquid inside the film is higher at the flat portion than at the curved portion, due to the surface tension. Thus, liquid in the flat portion of the film tends to flow to the curved portion. The flow drags some of the surface layer along with it, so that it locally creates a surface with less foaming agent and hence higher surface tension (Gibbs's effect). The diffusion of a foaming agent is slow if the concentration (gradient) is not high, and then this higher tension promotes a counter flow. Namely, the higher tension exerts a pull on the adjoining areas of lower tension, causing the surface to flow toward the region of higher tension, and the viscous drag of the moving surface carries liquid in the film with it (Marangoni effect). Of course, these effects occur almost simultaneously as complements to each other and thus affect foam stability by opposing the thinning process. However, the film may be ultimately so thinned that its thickness is of molecular-level size, and molecules in the film lose the properties of bulk liquid. It may eventually rupture, or attain an equilibrium thickness, depending on the amount and type of foaming agents and electrolytes. If electric double layers are present at both of the surfaces of the film, the DLVO theory (see Sec. 3.2) predicts the stability of the film. If evaporation occurs, the film is disturbed, affecting the foam stability.

4.3 Antifoamers and Foam Inhibitors

The prevention of foaming and the destruction of existing foams are often matters of practical importance. Antifoamers act against various factors that promote foaming, impairing the ability to foam. For instance, certain substances of low molecular weight may render the surface-active solute more soluble and hence less surface active. Foam inhibitors, on the other hand, totally inhibit the formation of foam. They are always insoluble materials that can spread on the surface of the solution.

A very effective form of antifoam is an emulsion of hydrophobized silica particles in silicone oil, dispersed in water.

5. COLLOIDAL SOLS AND SUSPENSIONS

5.1 Preparation of Colloidal Sols and Suspensions

There are many practically important sols and colloidal suspensions, such as silver halides, oxides and hydrous oxides, clays, carbon black, rubber latex, polystyrene, and paste. In general, the preparation of these systems for any purpose can proceed either by comminution of macroscopic material or by aggregation of smaller molecular units. The mechanical force required to break up crystalline solids into particles of colloidal size is enormous and not practical, unless the solid has glide planes. If rolling mills, ball mills, crushers, ultrasonic irradiation, etc., are not effective, chemical reactions as well as nucleation by vapor condensation may be used to create atomic or molecular aggregation, such as smoke, fume, etc., similar to preparations of aerosols (see Sec. 7.2). The term *sol* is used to distinguish colloidal suspensions from macroscopic suspensions.

5.2 Colloidal Stability, Flocculation, and Ionic Charges

An important cause of colloidal instability is aggregation and flocculation, which is due to dispersion forces, ionic interactions, the overlap of electric double layers, etc. Flocculation has a kinetic analogy with polymerization, which involves, however, covalent bonding. These tendencies are overridden if particles are electrically charged to repel each other to a larger separation. Lyophobic sols are stabilized mainly by electric-double-layer interactions. In hydrophilic sols, if the aqueous medium releases ions from the particles, as in silica, and so confers ionic charges on the surface, then an electric double layer extends into the solution, stabilizing the suspension. A quantitative discussion of these aspects has been successfully developed in terms of the DLVO theory (see Sec. 3.2) (Vold and Vold, 1983).

Usually nonelectrolytes have no effect on sols and suspensions. But when an electrolyte is added, coagulation may be the result. For example, when rivers reach salt water, deltas tend to be formed. For each electrolyte there is a threshold concentration that depends strongly on the charge number of the colloid and has to be exceeded for noticeable coagulation. The DLVO theory provides the theoretical rationale. Here, particular ions are effective, especially those of highest valency and of opposite charge from the colloidal particles. Among monovalent ions, the hydrogen ion and those of high molecular weight, such as alkaloids and dyes, are especially effective (McBain, 1950).

5.3 Zeta Potential

The physical properties of colloidal dispersions in electrolytes depend on the ionic charge on the particles. The amount of electric charge can be best determined by means of measurement of the zeta potential by electrophoresis (Hunter, 1981). When a charged particle is made to move in the electrolyte medium, the ion atmosphere around it is disturbed. Remote ions are left behind, and ions adjacent to its surface are dragged along with it. The concept of a shear plane has arisen, which is a dividing surface, in the mathematical sense, and separates the kinetic unit (colloidal particle plus those ions that move with it) from the medium. The potential due to the ionic charge at this surface is called the zeta potential. The zeta potential by itself is useful to predict the resistance to flocculation by electrolytes.

In solution electrophoresis, let us denote by v the speed of a charged particle moving in an external electric field E. Then the zeta potential is given by

$$\zeta=\eta v/\epsilon DEf\,, \tag{18}$$

where η and ϵ are respectively the viscosity and the dielectric constant of the bulk solution. The parameter f varies between 0.65 and 1.5, depending on the particle size and the ionic strength. Several commercial instruments are available for measurements of the zeta potential (Hiemenz, 1986).

6. GELS

6.1 Structure of Gels

Colloidal particles in sols are very often solvated to a thickness of about one molecule. Sometimes much greater amounts of solvent can be entrapped within particle aggregates or flocculent precipitates. In solutions of long threadlike molecules the polymer chains may cross-link and become mechanically entangled to such an extent that a continuous network is formed. If all of the solvent is trapped within this network, the system as a whole takes on a solid appearance and is called a gel. The solvent within the structure may be immobilized. Jellies are considered as one type of gel and have similar structures. In these, the units of the aggregates are primarily spherical or linear, but these are sufficiently linked so that the whole liquid is enmeshed into a loose framework. Thus, in gels or jellies both the solid and the fluid components are in a highly dispersed state. The entrapped water or ions behave in nearly the same way as in bulk water (Alexander and Johnson, 1949).

In order to form a three-dimensional structure, at least some particles must be joined to three or more others. Heat-reversible gels have junction points that are destroyed on heating and reformed on cooling. Cooling a sol produces eventual gelation at a temperature that seems to be reproducible. Once gelation has occurred, the gel can be maintained at higher temperatures up to some limit. This hysteresis permits holding the systems as either sol or gel within ranges of temperature characteristic of the material and depending also on its concentration and on the presence of other substances.

In some gels, the junction points are permanent. Such gels will come to an equilibrium volume when immersed in a given solvent. Of course, the junction points can move around in thermal motion. If the gels are placed in a suitable solvent, the amount of liquid absorbed can be very large. This is swelling. The theory of swelling can be developed by calculating the chemical potential of the solvent in the gel (Vold and Vold, 1983).

6.2 Applications of Gels

There are wide applications of gel states. It is necessary to maintain agricultural soil in a gel state in order to achieve good aeration and drainage. Treatment with coagulants is common practice. In oil-well drilling, removing the cuttings and sealing the bore wall are made effective if a drilling mud (a bentonite clay suspension) is converted to a gel by a peptizing agent, such as a polyphosphate. A paint should be designed to brush easily and yet not flow on a vertical surface; this is achieved by making the paint thixotropic.

Some gels behave like ultrafilter membranes of very large pore sizes of 100 Å or as large as 8000 Å (cellulose membranes). One application is gel permeation chromatography, which is of great value in biochemical and medical research.

7. AEROSOLS

See AEROSOLS for more details.

7.1 Nature and Characteristics of Aerosols

Liquid or solid colloidal dispersion in a gas, such as dust, fumes, smoke, fog, mist, and clouds, has enormous practical importance. The behavior is quite different from sols, suspensions, and emulsions. The reasons are that the gaseous medium is not as dense as a solvent so that the dispersed particles in aerosols collide easily with each other, and that the stability of the system cannot be readily modified by adding any agent to the gaseous medium. Thus aerosols are always unstable and tend to aggregate and precipitate, unless particles are protected by certain films against aggregation. If particles are volatile, they may disappear by evaporation, depending on temperature.

7.2 Preparation of Aerosols

Dust, fog, clouds, etc. have natural sources. However, the chief methods of preparing aerosols may utilize nucleation or dispersion processes.

Highly supersaturated vapor tends to condense to form large clusters, and if these are larger than a critical size, they grow spontaneously and form droplets. Classical nucleation theory treats the condensation and successfully predicts the nucleation rate (Zettlemoyer, 1969). If ions are present, stable condensation occurs around them at a lower supersaturation value than it would without them. If a liquid droplet forms by condensation on a solid surface with a certain contact angle, which determines the critical size, it grows also at a lower supersaturation value.

Dispersion processes include fuel injection, spray painting, and air humidifiers. Aerosols can also be made by using a nebulizer on solutions of the material in a volatile solvent such as a Freon or a hydrosol. Mechanical disintegration of liquid jets and high-voltage disruption of liquid droplets at the tip of a capillary are also useful.

GLOSSARY

Aggregate: A group of particles joined at their surfaces, with a specific surface area significantly less than the sum of the areas of their constituents.

Coagulation: Joining of primary particles at their faces, making the resulting total surfaces much smaller than the total surface area of their constituents.

Coalescence: The merging of small drops or bubbles into a single large drop or bubble.

Colloidal Particle: Usually a submicrometer particle, but defined academically as any material with at least one of its three dimensions between 10^{-9} and 10^{-6} m.

Dispersion: A finely divided solid, liquid, or gas in a medium.

Dispersion Forces: A long-range, usually attractive force between molecules or particles.

Electric Double Layer: Two parallel sheets of oppositely charged layers, formed by an ionic surface (layer) in contact with an electrolyte. Oppositely charged ions in the electrolyte are attracted to the surface, forming a diffuse layer of opposite charge, very close to the surface.

Electrophoresis: Electrokinetic effect resulting from migration of ions and charged particles relative to the medium, under the influence of an electric field.

Emulsion: A dispersion of an immiscible liquid in another.

Floc: A loose structure formed of primary particles, aggregates.

Flocculate: To create flocs by adding a chemical agent or by changing physical conditions.

Foam: A coarse dispersion of gas in a liquid.

Gel: A sol with strong viscoelastic properties.

Hydrophile: A hydrogen-bonding moiety.

Inversion: The process of reversing the continuous and discontinuous phases of an emulsion.

Lamella: An extended sheet of liquid.

Lipophile: A non-hydrogen-bonding moiety.

Lyophilic: Having an affinity for the medium.

Lyophobic: Lacking affinity for the medium.

Micelle: An organized aggregate of surfactant ions in which the lipophilic hydrocarbon chains are oriented towards the interior of the aggregate, leaving the hydrophilic groups in contact with the aqueous medium, or with reversed orientation of surfactant as well as medium.

Microemulsion: A thermodynamically stable, transparent solution of droplets of sizes between 0.01 and 0.1 μm.

Monodisperse: Dispersed with particles of identical size.

Number Density: Number of particles per unit volume.

Particle Texture: Particle surface characteristics and microstructure.

Peptize: To deflocculate, to disperse, to change a sol to a gel.

Plateau Border: The thin vein of liquid at the intersection of three liquid lamellae.

Polydisperse: Dispersed with particles of various sizes.

Sol: A transparent colloidal suspension.

Solvate: To immobilize solvent by mechanical entrapment at a particle surface or within particle aggregates.

Surfactant: A surface-active agent.

Suspension: Finely divided solid in a liquid medium.

Thixotropy: The time dependence of reforming an internal structure after its dispersion by shear.

Turbidity: The attenuation of transmitted light by scattering.

Ultramicroscope: An instrument to detect light scattered by ultramicroscopic objects when they are lit from the side and against a dark background.

Wet: A liquid wets a solid when it has a contact angle of 0° or less than 90°.

List of Works Cited

Alexander, A. C., Johnson, P. (1949), *Colloid Science*, Vol. II, London: Oxford (Clarendon), pp. 602–604.

Barth, H. G. (Ed.) (1984), *Modern Methods of Particle Size Analysis*, New York: Wiley.

Caldwell, K. D. (1984), in: H. G. Barth (Ed.), *Modern Methods of Particle Size Analysis*, New York: Wiley, pp. 211–250.

Ch'en, S. Y., Takeo, M. (1957), *Rev. Mod. Phys.* **29**, 20–73.

Friberg, S. E., Venable, R. L. (1983), in: P. Becher (Ed.), *Encyclopedia of Emulsion Technology*, Vol. 1, *Basic Theory*, New York: Marcel Dekker, p. 288.

Garcia-Rubio, L. H. (1987), in: T. Provder (Ed.), *Particle Size Distribution*, Washington, DC: American Chemical Society, pp. 162–163.

Hiemenz, P. C. (1986), *Principles of Colloid and Surface Chemistry*, 2nd ed. New York: Marcel Dekker.

Hunter, R. J. (1981), *Zeta Potential in Colloid Science*, New York: Academic.

Kerker, M. (1969), *The Scattering of Light and Other Electromagnetic Radiation*, New York: Academic, pp. 432ff.

Manegold, E. (1952), *Emulsionen*, Heidelberg: Strassenbau, Chemie und Technik, pp. 27–31.

McBain, J. W. (1950), *Colloid Science*, Boston, MA: Heath, pp. 174ff.

Pecora, R. (Ed.) (1985), *Dynamic Light Scattering*, New York: Plenum.

Provder, T. (Ed.) (1987), *Particle Size Distribution*, Washington, DC: American Chemical Society.

Pusey, P. N., Tough, R. J. A. (1985), in: R. Pecora (Ed.), *Dynamic Light Scattering*, New York: Plenum, pp. 85–179.

Ross, S., Morrison, I. D. (1988), *Colloidal Systems and Interfaces*, New York: Wiley.

Shinoda, K., Friberg, S. (1986), *Emulsions and Solubilitization*, New York: Wiley-Interscience.

Shinoda, K., Kunieda, H. (1983), in: P. Becher (Ed.), *Encyclopedia of Emulsion Technology*, Vol. 1, *Basic Theory*, New York: Marcel Dekker, pp. 337–365.

Tadros, T. F., Vincent B., (1983), in: P. Becher (Ed.), *Encyclopedia of Emulsion Technology*, Vol. 1, *Basic Theory*, New York: Marcel Dekker, pp. 152–217.

Vold, R. D., Vold, M. J. (1983), *Colloid and Interface Chemistry*, Reading, MA: Addison-Wesley.

Weiner, B. B. (1984), in: H. G. Barth (Ed.), *Modern Methods in Particle Size Analysis*, New York: Wiley, pp. 135–172.

Wielebinski, D., Findenegg, G. H. (1988), in: K. Hummel, J. Schurz (Eds.), *Progress in Colloid and Polymer Science*, Vol. 77. New York: Springer-Verlag, pp. 100ff.

Yau, W. W., Kirkland, J. J. , Bly, D. D. (1979), *Modern Size Exclusion Liquid Chromatography,* New York: Wiley.

Zettlemoyer, A. C. (1969), *Nucleation,* New York: Dekker.

Zrinyi, M., Kabai-Faix, M., Horkay, F. (1988), in: K. Hummel, J. Schurz (Eds.), *Progress in Colloid and Polymer Science,* Vol. 77. New York: Springer-Verlag, pp. 165ff.

Further Reading

McBain, J. W. (1950), *Colloid Science,* Boston, MA: Heath.

Pecora, R. (1985), *Dynamic Light Scattering: Application of Photon Correlation Spectroscopy,* New York: Plenum.

Provder, T. (1987), *Particle Size Distribution: Assessment and Characterization,* Washington, DC: American Chemical Society.

Ross, S., Morrison, I. D. (1988), *Colloidal Systems and Interfaces,* New York: Wiley.

Vold, R. D., Vold, M. J. (1983), *Colloid and Interface Chemistry,* Reading, MA: Addison-Wesley.

EARTH, INTERIOR STRUCTURE OF THE

SHUN-ICHIRO KARATO, *Department of Geology and Geophysics, University of Minnesota, Minneapolis, Minnesota, U.S.A.*

EIJI OHTANI, *Institute of Mineralogy, Petrology and Economic Geology, Tohoku University, Sendai, Japan*

Introduction 1
1. **Earth's Gross Structure** 2
1.1 Geophysical and Geochemical Observations 2
1.1.1 Seismology and the Earth's Interior 2
1.1.2 Solar Abundance and the Bulk Composition of Silicate Earth 5
1.2 The Composition of the Earth 6
1.2.1 The Crust 6
1.2.2 The Mantle 6
1.2.2.1 The Upper Mantle 6
1.2.2.2 The Transition Zone 7
1.2.2.3 The Lower Mantle 9
1.2.2.4 The D'' Layer and the Core–Mantle Boundary 9
1.2.3 The Core 9
2. **Earth's Thermal Regimes** 10
2.1 Introduction 10
2.2 Heat Flow and Near-Surface Temperature Profile 11
2.3 Temperature Profiles of the Earth's Deep Interior 11
2.4 Thermal History 13
3. **Melting of Silicates and Magmatism** 14
3.1 Melting and Chemical Stratification 14
3.2 Magmatism and Melting of the Earth's Interior 14
3.2.1 Mafic Magmas 14
3.2.2 Silicic and Intermediate Magmas 14
3.2.3 Ultramafic Magmas and Melting in the Deep Mantle 14
4. **Earth's Rheological Structure and Dynamics of the Earth's Interior** 15
4.1 Time-Dependent Deformation of the Earth 15
4.2 Earth's Rheological Stratification 17
4.3 Dynamics of the Earth's Interior 18
4.3.1 Plate Tectonics 18
4.3.2 Seismic Tomography and Deep-Mantle Dynamics 18
4.3.3 Geochemical Observations on Mantle Heterogeneities 19
5. **Summary** 20
Glossary 20
Works Cited 20
Further Reading 20

INTRODUCTION

The Earth's deep interior is largely inaccessible (the deepest hole at Kola peninsula in Russia is ~12 km, only ~0.2% of the Earth's radius), and consequently most of our understanding is based on indirect inferences. These include the average chemical composition of the solar system, the chemical composition of rocks near the Earth's surface, geophysical observations of the Earth's density and of seismic wave-velocity distribution, and laboratory studies of the state and properties of materials at high pressures and temperatures. Recent progress has also been made to increase the resolution in inversion of the Earth's structure through seismic tomography and developments in high-pressure and high-temperature experimental studies.

Section 1 of this article discusses the gross Earth structure, i.e., the major constituent materials of the crust, mantle, and core. Section 2 discusses the Earth's temperature distribution, which is determined by the heat

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

generation and heat transfer processes. The temperature distribution determines the Earth's rheological stratification and has a significant effect on dynamics (see Sec. 4). Section 3 summarizes experimental observations on the melting of silicates. Finally, Sec. 4 gives a brief summary of rheological properties of rocks and the dynamics of the Earth's interior.

1. EARTH'S GROSS STRUCTURE

1.1 Geophysical and Geochemical Observations

1.1.1 Seismology and the Earth's Interior (see also SEISMOLOGY). Table 1 summarizes some of the important parameters of the Earth. Note that the Earth's average density, 5.52×10^3 kg/m^3, is significantly higher than the densities of rocks found on the Earth's surface [$(2.7\text{–}3.4)\times10^3$ kg/m^3]. Another observation is that the ratio $C/MR^2=0.331$ (C, moment of inertia; M, mass; R, radius) is significantly smaller than the theoretical value for a homogeneous body (0.40). These observations show that the density of the Earth must increase significantly with depth. This density increase is due to two factors: the existence of a dense iron-rich core and the increase in density due to hydrostatic compression within the Earth.

The most detailed information on the variation of physical properties or composition of the Earth with depth comes from seismological observations of elastic-wave propagation in the Earth. Various types of elastic waves are used to probe the Earth, including body waves, surface waves, and the free oscillation of the Earth.

The vibrations of the Earth due to natural earthquakes or artificial explosions are recorded on seismograms. Seismic waves recorded at a long distance from the focal point of an earthquake pass through deeper portions of the Earth's interior than waves recorded at a nearby station. Similarly, the surface waves and the free oscillation of the Earth with a longer period reflect the properties of deeper portions of the Earth than those with a shorter period. Thus using the seismograms at different stations, or the surface-wave or free-oscillation records with different periods, one can estimate the depth variation of elastic properties and density. The information used includes the travel times of body waves, the frequency dependence of surface-wave velocities, and the frequencies of various modes of free oscillation.

In an isotropic elastic body, there are two different modes of elastic waves: compressional and shear waves. The velocities of these waves (V_P, velocity of compressional wave; V_S, velocity of shear wave) are related to the physical properties by

$$V_P=\sqrt{(K+\tfrac{4}{3}\mu)/\rho} \qquad (1)$$

and

$$V_S=\sqrt{\mu/\rho}, \qquad (2)$$

respectively, where K is incompressibility, μ is rigidity, and ρ is density. These relations are exact only for an isotropic material, but will be sufficient for the purpose of the present section. However, when the fine details of seismic wave propagation are used to discuss the dynamics of the Earth's interior, the consideration of anisotropy becomes important (see Sec. 4.3.2).

An Earth model (depth variation of seismic wave velocities and densities) based on seismological data is shown in Fig. 1. This model (called PREM, for preliminary reference earth model; see Anderson, 1989) incorporates such details as the frequency dependence of seismic-wave velocities and attenuation but neglects the lateral variation and the azimuthal anisotropy (variation of seismic-wave velocities with the orientation of wave propagation; see Sec. 4.3.2). The gross Earth structure discussed in Sec. 1 is mainly based on the laterally averaged model (except for the crust).

Such an Earth model indicates that the Earth's interior is divided into three layers; the crust, the mantle, and the core. Once an Earth model is obtained, one can infer the internal constitution of the Earth by comparing the model with laboratory data on density

Table 1. Earth parameters.

Total mass, M	5.97×10^{24} kg
Mean radius, R	6371 km
Mean density, ρ	5.52×10^3 kg/m^3
Moment of inertia, C[a]	8.04×10^{37} kg m^2
C/MR^2	0.331

[a] $C=(8\pi/3)\int_0^R\rho(r)r^4dr$

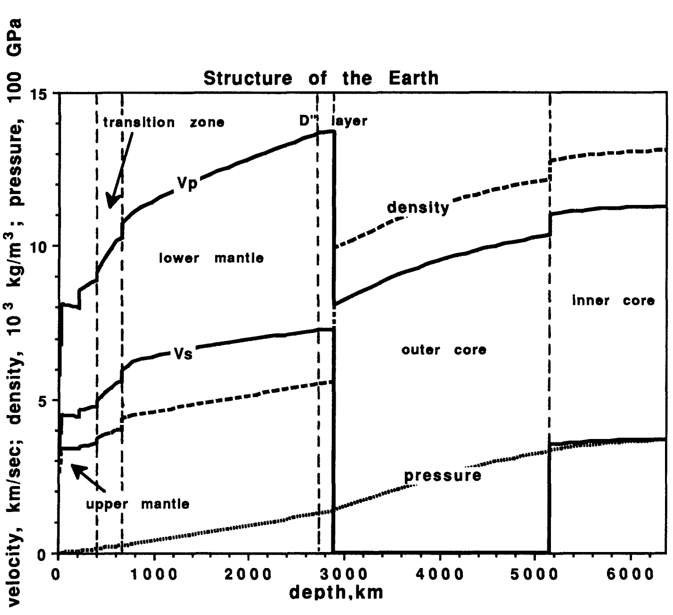

FIG. 1. The depth variation of the average seismic-wave velocities, densities, and pressure according to the PREM model.

and elastic properties with the help of physics of materials under high pressures and temperatures. Such studies indicate that these three layers have quite different chemical composition: The crust is mainly composed of silica-rich minerals such as quartz (SiO_2) and feldspars [$CaAl_2Si_2O_8$, $(K,Na)AlSi_3O_8$], while the mantle is composed of magnesium- and iron-rich minerals such as olivine [$(Mg,Fe)_2SiO_4$] and pyroxenes [$(Mg,Fe,Ca)SiO_3$] and their high-pressure polymorphs. The core is composed mainly of iron and some other alloying elements. The details of compositions of each layer will be discussed in Sec. 1.2.

The Earth structure thus inferred may be compared with a theoretical prediction of depth variation of density and of seismic-wave velocities for an ideal situation, that is, the depth variation of these properties in an adiabatically compressed homogeneous body (a situation corresponding to a vigorously convecting layer without phase transformations). The density change in a homogeneous body due to adiabatic compression is given by

$$\frac{d\rho(r)}{dr}=\left(\frac{\partial\rho}{\partial P}\right)_S\frac{dP(r)}{dr},\tag{3}$$

where r is the distance from the center of the Earth, $(\partial\rho/\partial P)_S$ indicates the pressure dependence of density at adiabatic conditions, and P is the pressure. Now, as we shall see (Secs. 2 and 4), the temperature in the Earth is high enough to make rocks soft, and therefore no significant nonhydrostatic stresses can be maintained over geological time scales. Thus

Table 2. Density and elastic properties of typical minerals (at room temperature and room pressure).

Mineral	Density, ρ (10^3 kg/m^3)	Incompressibility, K (GPa)	Rigidity, μ (GPa)
Quartz, SiO_2	2.65	38	44
Feldspar (anorthite), $CaAl_2Si_2O_8$	2.76	90	40
Olivine, Mg_2SiO_4	3.22	129	82
Modified spinel, Mg_2SiO_4	3.47	174	114
Spinel, Mg_2SiO_4	3.55	184	119
Pyroxene (enstatite), $MgSiO_3$	3.20	104	77
Garnet (majorite), $MgSiO_3$	3.52	175	90
Perovskite, $MgSiO_3$	4.10	246	184
Magnesiowüstite, MgO	3.58	163	131
Iron, Fe	7.87	172	83

the hydrostatic equilibrium is a very good approximation. Hence, we get

$$\frac{dP(r)}{dr} = -g(r)\rho(r) = -\frac{GM(r)\rho(r)}{r^2}, \tag{4}$$

where $g(r)$ is the gravitational acceleration at r, G is the gravitation constant, and $M(r)$ is the mass of a sphere within the radius r, which is given by

$$M(r) = 4\pi \int_0^r \xi^2 \rho(\xi) d\xi. \tag{5}$$

The change in density with pressure is given by the incompressibility $K(r)$ which is related to the observed seismic-wave velocities through

$$\begin{aligned} K(r) &= \rho(r)\left(\frac{\partial P}{\partial \rho}\right)_S \\ &= \rho(r)\left(V_P^2(r) - \frac{4}{3} V_S^2(r)\right) \\ &= \rho(r)\Phi(r), \end{aligned} \tag{6}$$

where $\Phi(r)$ is the seismic parameter defined by $\Phi = V_P^2 - \frac{4}{3}V_S^2$. Combining Eqs. (3)–(6) one gets the Adams–Williamson equation that describes the depth variation of density due to adiabatic compression of a homogeneous body,

$$\frac{d\rho(r)}{dr} = -\frac{\rho(r)g(r)}{\Phi(r)} = -\frac{4\pi G\rho(r)}{\Phi(r)r^2} \int_0^r \xi^2 \rho(\xi) d\xi. \tag{7}$$

The pressure in the Earth can also be calculated from the estimated densities using Eq. (4).

A useful application of the Adams–Williamson equation is to examine how the real Earth structure is different from the idealized state of adiabatic compression of a homogeneous body. This can be done by calculating the inhomogeneity parameter, defined as

$$\eta_B(r) = -\frac{d\rho(r)}{dr} \frac{\Phi(r)}{\rho(r)g(r)}, \tag{8}$$

from the inferred density and seismic-wave velocity distributions. For a homogeneous body compressed adiabatically, $\eta_B = 1$. Here $\eta_B > 1$ implies that the density increases with depth more than in an adiabatically compressed homogeneous body. A typical example includes the density increase due to phase transformations. In contrast, $\eta_B < 1$ implies that density does not increase with depth as much as in an adiabatically compressed homogeneous body. A typical example is the case where the temperature gradient is higher than the adiabatic gradient. Changes in chemical composition with depth can also cause the deviation of η_B from 1. Figure 2 shows the inhomogeneity parameter in the Earth. It is seen that the core and the lower mantle are very close to an adiabatically compressed homogeneous body, but significant deviations are observed in the shallow upper mantle

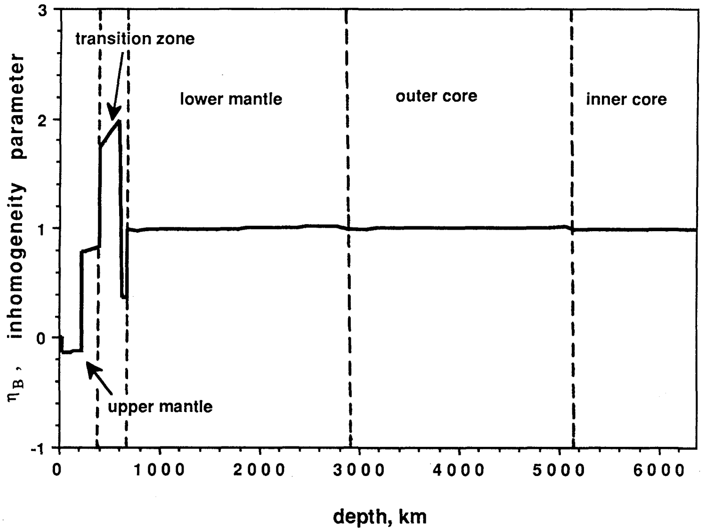

FIG. 2. The depth variation of the inhomogeneity parameter. The deviation from homogeneity ($\eta_B = 1$) is significant in the upper mantle and in the transition zone.

(down to ~200 km) and in the transition zone (400–670 km). We shall discuss possible implications of these observations in the following sections.

A layered Earth model such as shown in Fig. 1 is an idealization. Deviations from a simplified layered model are sometimes important in the discussion of the Earth's structure and dynamics. First, significant lateral variations occur in both the crust and the mantle, reflecting the lateral variation of temperature and/or composition. Second, the structure of the Earth is sometimes anisotropic (anisotropy in the upper mantle is taken into account in a simplified way in the PREM). Studies of fine structure of the Earth (lateral heterogeneity and anisotropy) have made significant progress during the past few years as a result of the development of digital broad-band seismic networks and powerful computers. The results and their implications for the Earth's dynamics are discussed in Sec. 4.3.

1.1.2 Solar Abundance and the Bulk Composition of Silicate Earth. The Earth's internal constitution may also be studied by analysis of the chemical composition of rocks (geochemistry). Unfortunately, we only have access to rocks presently at the surface of the Earth. Rocks are there as a result of complicated processes of differentiation and tectonics and do not represent the average composition of the Earth. To get some insight into the composition of the deep interior of the Earth, we need some understanding of the source materials from which the Earth was formed and the differentiation processes that have occurred throughout its history.

The Earth was formed in the solar nebula at ~4.5 billion years ago. The Earth's present chemical structure is a result of differentiation processes that have occurred in the solar nebula and in the Earth's interior. Therefore, a starting point of understanding the chemical composition and the differentiation processes in the Earth (and in the solar nebula) is the chemical composition of the solar system ("solar abundance"). There are two ways of estimating solar abundance. The first way is through analysis of the chemistry of the outer layer of the sun where convection is vigorous and therefore the chemistry of the outer layer is believed to be representative of the average chemistry of the sun and of the solar system-.The second way is examination of the chemical composition of the "primitive" (meaning undifferentiated) meteorites such as (type I) carbonaceous chondrites. These two estimates agree well, except for some volatile elements which are depleted from chondrites. The average chemical composition of the nonvolatile components of the solar system is given by the "solar abundance" in Fig. 3.

Basaltic magma is generated by partial melting of the upper mantle (see Secs. 1.2.1 and 3.2.1). Ringwood (1979) considered that upper mantle rocks or peridotites (composed mainly of olivine and pyroxenes) sampled as xenoliths by some volcanoes represent a residue of partial melting that generates basaltic magma. Therefore, the composition of the source material from which both basaltic magma and peridotite xenoliths are formed is considered to be some average of the two. This hypothetical source material is called "pyrolite" (see Fig. 3).

The chemical composition of upper-mantle rocks shows a large diversity due to variations in the degree of differentiation by partial melting. Rocks that show evidence of minimum differentiation may be representative of the primitive mantle. The composition of these primitive rocks is similar to that of pyrolite and has the following characteristics (Fig. 3):

1. Most of the refractory elements (except for elements with strong affinity to Fe, called siderophile elements) have abundances similar to the "solar abundance," although the Mg/Si ratio of the mantle (about 1.27 in atomic ratio) is slightly higher than that of the solar abundance (about 1.05);
2. Fe and siderophile elements are depleted; and
3. volatile elements are depleted.

Among the above compositional characteristics of the primitive upper mantle, 2 and 3 are readily explained by differentiation due to the formation of the Earth's core and by differentiation in the solar nebula, respectively. However, the origin of the high Mg/Si ratio [item (1)] has been a controversial matter. Three alternative explanations are possible:

1. The chemical composition of the Earth is different from the "solar abundance," possibly because of the differentiation processes in the solar nebula,

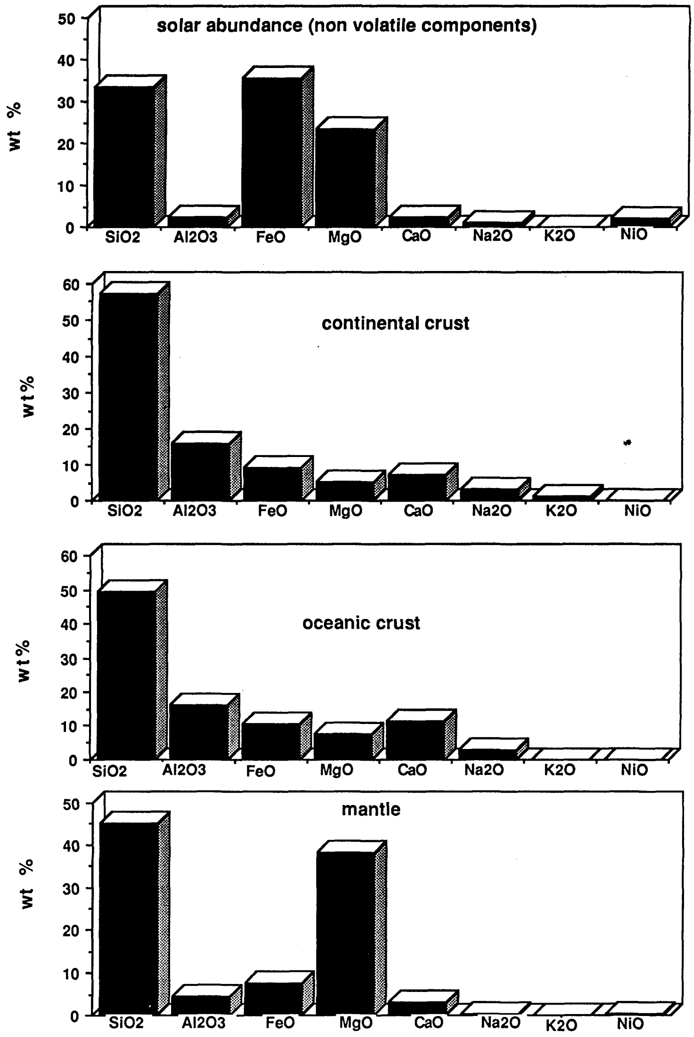

FIG. 3. The chemical composition of the solar system, the mantle (the pyrolite model), and the oceanic and continental crust.

2. the lower mantle (and/or the transition zone) is more enriched with Si than the upper mantle, and
3. a significant amount of silicon is dissolved in the core.

1.2 The Composition of the Earth

(See also VOLCANICS; PLATE TECTONICS; GEOLOGY)

1.2.1 The Crust. The crust is the near-surface layer with variable thickness (~5–50 km). The crust in the oceanic region is considerably thinner than the crust in the continental region and is remarkably homogeneous with ~6-km thickness. In contrast, the thickness of continental crust is highly heterogeneous. It is 10–20 km in the Basin and Range province in the U.S.A., but it is ~50 km in Tibet.

The oceanic crust is made of sediments (major chemical composition being $CaCO_3$ and SiO_2) and an underlying basaltic layer {basalt is a fine-grained rock composed mostly of calcium-rich feldspar ($CaAl_2Si_2O_8$) and pyroxenes [$(Mg,Fe,Ca)SiO_3$]}. The basaltic layer is created at mid-oceanic ridges by the partial melting of upper mantle rocks (see Sec. 3.2.1). The oceanic crust subducts together with the underlying uppermost mantle (together they constitute the oceanic lithosphere) into the deeper mantle at ocean trenches.

The continental crust is more heterogeneous than the oceanic crust, reflecting its complicated history of metamorphism and deformation. A simplified model divides it into two layers: the upper and the lower continental crust. The upper continental crust is mainly composed of silica-rich rocks such as granitoids, the major constituent minerals being quartz (SiO_2), potassium/sodium-rich feldspars [$(K,Na)AlSi_3O_8$], biotite [$K(Mg,Fe)_3(AlSi_3O_{10})(OH)_2$] and hornblende [$Ca_2(Mg,Fe,Al)_5(Si,Al)_8O_{22}(OH)_2$]. The lower crust is composed of more mafic (meaning MgO- and FeO-rich) rocks such as basalt, gabbro, or amphibolite, the major constituent minerals being calcium-rich feldspar, pyroxenes, hornblende, and amphibole [$(Mg,Fe,Ca,Na)_{2\text{-}3}(Mg,Fe,Al)_5(Si,Al)_8O_{22}(OH)_2$].

1.2.2 The Mantle. The crust-mantle boundary is referred to as the Mohorovičić (or Moho) discontinuity. The velocity of compressional waves changes from 6–7 to 8 km/s at this discontinuity. The presence of clear reflection at this discontinuity indicates that this is a sharp boundary: The width must be significantly smaller than the wavelength of seismic waves ($\sim$5 km).

The mantle may be divided into three layers: the upper mantle ($<$400 km), the transition zone (400–670 km), and the lower mantle (670–2900 km).

1.2.2.1 The Upper Mantle. The layer below the Moho discontinuity down to $\sim$400 km is referred to as the upper mantle. The comparison of inferred velocities with laboratory measurements shows that the major constituent minerals are olivine, pyroxenes, and garnets [$(Mg,Fe,Ca)_3Al_2Si_3O_{12}$]. Olivine is the most abundant mineral in the uppermost mantle. It has highly anisotropic elastic properties and the observed anisotropy of seismic waves in the upper mantle (see Sec. 4.3.2) can be explained by the anisotropic alignment of olivine. However, the relative amounts of these minerals may change with depth.

The inhomogeneity parameter defined by Eq. (8) has low values in the upper mantle, indicating that the temperature gradient is much higher than the adiabatic gradient. This is consistent with the temperature gradients estimated from other geophysical and petrological observations (see Sec. 2.1).

The high temperature gradient causes a mechanical stratification of the upper mantle (see also Sec. 4.2). The top cold part is the lithosphere which is characterized by high velocity and low attenuation of seismic waves, low electrical conductivity, and high creep strength (viscosity). A layer beneath the lithosphere is characterized by relatively low velocity and high attenuation of seismic waves, high electric conductivity, and low creep strength, all of which are indicative of high temperatures (temperatures close to the melting temperature). This layer is called the asthenosphere. The lithosphere/asthenosphere boundary is well defined by a sharp change in physical properties, but the bottom of the asthenosphere is not well defined. The depth of the lithosphere/asthenosphere boundary varies from place to place. In the oceanic mantle, the lithosphere thickness increases with age: the lithosphere is very thin near the ocean ridges where new oceanic lithosphere is formed, but its thickness increases up to 50–100 km at the age of $\sim$100 million years.

An important issue about the asthenosphere is whether it represents a layer of partial melting or not. It was often postulated that the geophysical anomalies (low velocity and high attenuation of seismic waves, high electrical conductivity, low viscosity) in the asthenosphere are due mainly to the presence of a small amount of melt. But recent experimental studies on elasticity (and anelasticity), electrical conductivity, and rheology of upper-mantle rocks show that high temperature (and/or the presence of hydrogen) is sufficient to cause these anomalies. Partial melting may not be required throughout the asthenosphere but may be a localized phenomenon beneath volcanoes rather than a global phenomenon.

1.2.2.2 The Transition Zone. Laboratory studies under high pressures and temperatures show that silicate minerals undergo several series of phase transformations at the pressure and temperature conditions corresponding to those in the transition zone (400–670 km). Most of these phase transformations are of the first order and are associated with the increase in density and in seismic-wave velocities (see Table 2). The observed high gradients or discontinuities in density and in seismic-wave velocities (and the resultant high inhomogeneity parameter) in the transition zone is mostly attributed to these phase transformations.

Olivine [$(Mg,Fe)_2SiO_4$] with $Fe/(Fe+Mg) \sim 0.1$ transforms to the modified spinel structure at $\sim$400 km and then to the spinel structure at ~500 km and finally to the perovskite [(Mg,Fe)SiO$_3$]+magnesiowüstite [(Mg,Fe)O] at ~670 km (see Fig. 4). In these cases, the packing of atoms changes significantly in a narrow pressure (or depth) interval (see Fig. 4). The sharp increases in density and in seismic-wave velocities at ~400 and ~670 km are considered to be due mainly to the transitions from olivine to modified spinel and from spinel to perovskite + magnesiowüstite, respectively (the modified-spinel to spinel transformation does not change the density or velocity very much; see Table 2). In the case of the spinel to perovskite + magnesiowustite transformation, the number of oxygen ions surrounding the silicon ions changes with the transformation: Silicon ions are surrounded by four oxygen ions in spinel, while they are surrounded by six oxygen ions in perovskite.

The other example is the transformation of pyroxene [(Mg,Fe)SiO$_3$] into the garnet structure. In this case, the mutual solubility of pyroxenes and alumino-silicate garnet [(Mg,Fe)$_3$Al$_2$Si$_3$O$_{12}$] increases with pressure. Thus the transformation is gradual and this phase transformation is considered to be responsible for the observed gradual increase in density and velocities in the depth range of 400–670 km. In a silica-rich garnet, some lattice sites which are normally occupied with aluminum ions are occupied with silicon and magnesium ions, and therefore those silicon ions are surrounded by six oxygen ions. Garnets which have silicons surrounded by six oxygens are called majorite.

The transformation pressure P_{tr} varies with temperature through the Clausius–Clapeyron relation, viz;

$$\frac{dP_{\mathrm{tr}}}{dT}=\frac{\Delta S}{\Delta V}, \tag{9}$$

where $\Delta S = S_{\mathrm{hpp}} - S_{\mathrm{lpp}}$ (S_{hpp}, entropy of the high-pressure phase; S_{lpp}, entropy of the low-pressure phase) and $\Delta V = V_{\mathrm{hpp}} - V_{\mathrm{lpp}}$ (V_{hpp}, volume of high-pressure phase; V_{lpp}, volume of low-pressure phase). In all cases, ΔV is negative, but the sign of ΔS depends on specific transformations. In the case of the olivine to modified spinel transformation, ΔS is negative, that is, the modified spinel has smaller entropy than olivine, and hence dP_{tr}/dT is positive. In contrast, the phase transformation from spinel to perovskite + magnesiowüstite has negative dP_{tr}/dT, because the

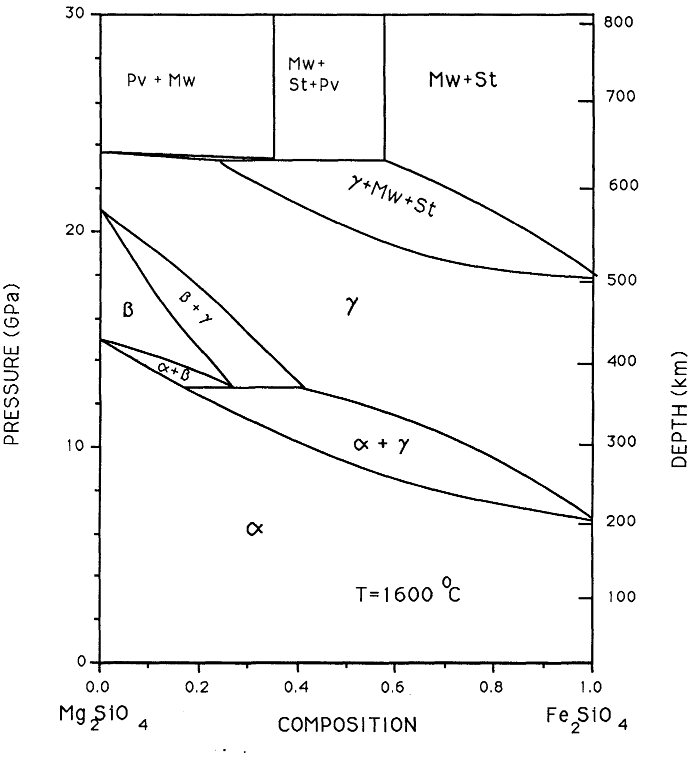

FIG. 4. The phase diagram of (Mg,Fe)$_2$SiO$_4$. α, olivine; β, modified spinel; γ, spinel; Pv, Perovskite; Mw, magnesiowüstite; St, stishovite. For a typical chemical composition appropriate to the Earth's interior [Fe/(Mg+Fe)~0.1], (Mg,Fe)$_2$SiO$_4$ transforms from olivine (α) to modified spinel (β) at ~400 km, to spinel (γ) at ~500 km, and finally to perovskite (Pv) +magnesiowüstite (Mw) at ~670 km at T~1600° C. T~1600° C is the temperature estimated at 670 km on the basis of the comparison of the phase diagram with the seismological observations (see Sec. 2).

perovskite (+ magnesiowüstite) has higher entropy than spinel. Since the transformation pressure (depth) depends on temperature, by measuring the exact depth of phase transformations from seismological observations, one can estimate the temperature at these depths (see Sec. 2.3).

Although the major cause for high velocity gradients and velocity discontinuities in the transition zone is now known to be phase transformations in silicates, it is not well established whether the observed density and velocity changes in this region are solely due to phase transformations or partly due to some chemical stratifications as well. Seismological and petrological observations are also consistent with chemical stratification involving enrichments in pyroxenes and/or garnet in the transition zone.

Major seismic discontinuities in the mantle occur at ~400 and 670 km on the global scale. However, these depths may vary laterally because of heterogeneity in temperature and/or chemistry (or mineralogy). In particular, large variation in these depths is expected in and around subducting slabs where a large lateral variation in temperature and/or chemistry (mineralogy) occurs. High-resolution seismological studies are under way to detect these fine details, which contain important information as to lateral variation in temperature and/or chemistry (or mineralogy) and hence have important clues as to the dynamics of the Earth's interior.

1.2.2.3 The Lower Mantle. The lower mantle is mostly composed of (Mg,Fe)SiO_3 perovskite and magnesiowüstite. The amount of magnesiowüstite is not well constrained and ranges from 0% to 20% depending on the models. Thus the (Mg,Fe)SiO_3 perovskite is the most abundant mineral in the Earth (it constitutes ~40%–50% of the Earth). The velocity and density variations in the lower mantle are smooth, and are compatible with a homogeneous composition without major phase transformations except near the topmost lower mantle (~670–1000 km)where relatively high gradients of these properties are observed. The phase transformation of majorite garnet into perovskite structure, or formation of some aluminous high-pressure minerals, may be responsible for the anomalous density and velocity gradient of the topmost part of the lower mantle.

1.2.2.4 The D'' Layer and the Core–Mantle Boundary. The very bottom of the Earth's mantle (100–200 km thick), which is called the D'' layer, has unusual physical properties: the velocity gradient is low and lateral variation is very large (Lay, 1989).

Two hypotheses have been proposed to explain this layer. One is that the D'' layer corresponds to the bottom thermal boundary layer of a convecting mantle where the temperature gradient is high. The other alternative is that the D'' layer is chemically different from the rest of the mantle. A high temperature gradient is also suggested by the comparison of estimated temperatures in the mantle with the melting temperatures of the core materials (see Secs. 2.3 and 3.3). Both high temperature gradient and large chemical heterogeneities may cause the unusual properties of the D'' layer.

The core–mantle boundary is not smooth but has some topography with amplitude ~0.5 to ~5 km. Such topography may be a result of upwelling and downgoing convection current in the Earth's mantle and may also reflect the density variations due to chemical heterogeneities. The topography at the core–mantle boundary enhances the mechanical interaction of the core and mantle.

1.2.3 The Core. The core is divided into two parts: the outer core and the inner core. The densities of these two layers are much higher than those of silicates, and the major constituent of the Earth's core is believed to be metallic iron (Fig. 5). This conclusion is supported by the comparison of chemical compositions of the mantle (and the crust) with the solar abundance (Fig. 2). The inspection of Fig. 2 shows that the mantle (and the crust) is significantly depleted in iron compared to the solar abundance, suggesting an iron-rich layer in the deep portion of the Earth.

The absence of shear-wave propagation in the outer core indicates that it is largely in a molten state. This metallic liquid outer core may generate the geomagnetic field by convection. The observed density of the outer core indicates that the outer core is less dense than molten iron. Some light elements are needed to explain the observed density. Candidates include sulfur, oxygen, hydrogen, silicon, and carbon. High-pressure experiments show that all of these elements can be dissolved into iron at high pressures and high

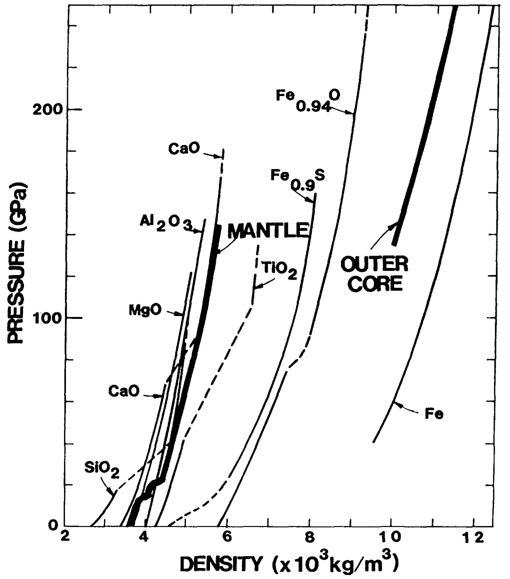

FIG. 5. Density–pressure relationships for various materials. Thin solid lines are those determined by laboratory experiments by shock waves. Thick solid lines are those inferred from seismological observations. The comparison shows that the mantle is composed of silicates or oxides but the core is composed mainly of iron.

temperatures. No consensus has been reached as to which of these elements are dominant. An important effect of light-element alloying is that it significantly reduces the melting temperature.

Shear waves pass through the inner core, and therefore it must be largely in a solid state. The density of the inner core is close to that of pure iron (+ nickel). Thus the Earth's core appears to be chemically layered: a relatively pure, solid inner core, and a liquid outer core containing a significant amount of light elements. An interesting hypothesis is that the inner core may have resulted from the solidification of the outer core as a result of secular cooling. On the basis of the seismologically determined density jump at the outer–inner core boundary and the assumption that the inner core has grown throughout the Earth's history, one can estimate the rate of energy release upon the solidification of the inner core. It is shown that the gravitational energy released by the growth of the inner core can provide sufficient energy to generate the geomagnetic field.

2. EARTH'S THERMAL REGIMES

2.1 Introduction

The temperature profile $T(z)$ of the Earth's interior is determined by the heat generation and heat transfer processes through the energy conservation equation, viz;

$$\frac{\partial T}{\partial t}=\left(\frac{k}{\rho C_p}\right)\nabla^2 T+\frac{H}{\rho C_p}-\mathbf{v}\cdot\nabla T, \tag{10}$$

where k is thermal conductivity, C_p the specific heat, H the heat generation per unit volume, and $\mathbf{v}$ the velocity. The first term in the rhs of Eq. (10) indicates the heat loss due to thermal conduction, the second term the heat generation, and the third term the energy transfer due to convection. The relative importance of each term depends on specific geological settings. For example, the heat production due to radioactive elements is important in the continental crust, but is less important in the oceanic crust and in the mantle. The relative importance of convective and conductive heat transfer depends on the rheology (viscosity) of Earth material (see Sec. 4). Convection becomes important when the Rayleigh number exceeds a critical value ($\sim 10^3$). The Rayleigh number for a fluid layer heated from below is given by

$$Ra=\alpha\rho\Delta T g L^3/\kappa\eta, \tag{11}$$

where α is the thermal expansion, ΔT the temperature difference between the top and bottom of the layer, L the thickness of the layer, κ the thermal diffusivity ($\kappa=k/\rho C_p$), and η the viscosity. Among these parameters, the most critical is the viscosity, which changes significantly with temperature (and pressure) (see Sec. 4). The parameters in the equation are presented in Table 3 which shows that the Rayleigh number of the mantle (below the asthenosphere) is 10^5–10^7 and therefore vigorous convection should occur in most of the mantle (and the core).

When convective heat transfer is not important, the temperature profile is determined by heat generation within a layer and by the boundary conditions [T and the heat flux $J=k(dT/dz)$ at the boundary]. The amount of heat flux varies from place to place and ranges from (20 to 300)$\times 10^{-3}$ W m^{-2}, and the total heat loss from the Earth is estimated to be $\sim 4\times 10^{13}$ W. Since the thermal conductivities

Table 3. Physical constants and parameters relevant to mass and heat transfer in the Earth.

Thermal expansion α^a	3×10^{-5}/K
Acceleration of gravity g^b	10 m^2/s
Density ρ^c	$(3–5)\times10^3$ kg/m^3
Thermal diffusivity κ	10^{-6} m^2/s
Thickness of the layer L^d	400–2900 km
Viscosity η^e	
(asthenosphere and below)	10^{20}–10^{22} Pa s
(lithosphere)	10^{24} Pa s or higher
Temperature difference ΔT^f	1500–4000 K

[a] Thermal expansion in minerals depends on temperature and pressure. This value is for olivine at high T ($T\sim1000$–1500 K) and at 0.1 MPa.

[b] The acceleration of gravity is nearly constant throughout the Earth's mantle.

[c] See the Earth model in Fig. 1.

[d] For the upper mantle, $L=400$ km; for the whole mantle, $L=2900$ km.

[e] The viscosity of mantle materials is a strong function of temperature (and pressure) and varies significantly with depth (see Sec. 4; Fig. 9).

[f] See Sec. 2.

of rocks are generally small (0.1–0.5 W/m K), steep temperature gradients (10–500 K/km) will be formed in a layer where only conductive heat transfer occurs. In contrast, when convective heat transfer becomes important, the temperature profile within a convection cell becomes close to adiabatic ($dT/dz=\alpha gT/C_p$), leading to a much smaller gradient (0.3–0.5 K/km for the upper mantle). One expects therefore that the temperature gradient near the surface (i.e., the lithosphere) is much higher than in the deep mantle.

The temperature profile also varies laterally because the amount of convective heat transport and of heat generation varies from place to place. A large amount of heat is carried to the surface near ocean ridges by an upwelling convective current, but the downgoing current related to subducting slabs at ocean trenches will cool the mantle. The lateral variation in temperatures therefore reflects the convective pattern and provides useful constraints as to the dynamics of the deep interior of the Earth (see Sec. 4.3).

The Earth's temperature profile has also changed through geological time. Current understanding of temporal variation of the Earth's temperature is discussed in Sec. 2.4.

2.2 Heat Flow and Near-Surface Temperature Profile

The processes that determine the temperature profiles are quite different between the oceanic and the continental lithosphere. The oceanic lithosphere contains only a small amount of radiogenic elements and the temperature profiles are mainly determined by the heat flux carried by the underlying convective current. This heat flux is large near ocean ridges and decreases with the distance from ridges. Thus the temperature of the oceanic lithosphere decreases with the distance from ocean ridges. The temperatures near ocean ridges exceed the solidus of upper mantle rocks and partial melting occurs to create a basaltic layer.

In the continental lithosphere, the heat transported by the underlying convective current is relatively small [$(30–50)\times10^{-3}$ W m^{-2}] compared to the oceanic lithosphere. However, the contribution from radiogenic heat (H) in the continental crust is significant and compensates the effects of relatively small heat flux. Thus the temperature profiles of the continental lithosphere are similar to those of old oceanic lithosphere, and the heat flux from the old oceanic lithosphere is similar to that from the continental lithosphere. In total, the heat flux from the Earth amounts to $\sim4\times10^{13}$ W.

The temperature gradient measured in this way ranges from ~10 K/km in the stable continent to more than 500 K/km near volcanic regions (such as ocean ridges). When extrapolated to deeper portions, temperatures would exceed the melting temperatures of most of rocks at the depth of 100–200 km. However, seismological observations indicate that most of the Earth's mantle is solid. The implication is that the temperature gradients at the deeper portions must be significantly smaller than the near-surface gradients.

2.3 Temperature Profiles of the Earth's Deep Interior

Melting temperatures of Earth materials (see Sec. 3) provide important constraints on the temperature profiles. First, the temperatures in the inner core and the mantle must be lower than the melting temperatures of the relevant materials. Second, the temperatures in the outer core must be higher than the melting temperature of iron (plus some light elements).

Melting temperatures of iron (and its alloys) have been determined up to pressures corresponding to the core pressures, using shock waves as well as static compression experiments. The melting temperature of iron increases with pressure and reaches ~7000 K at the pressure of the outer–inner core boundary. If the core were made of pure iron, this temperature would be the temperature at the outer–inner core boundary. However, the outer core contains a significant amount of secondary components such as sulfur, oxygen, and hydrogen (see Sec. 1.2.3) which, in most cases, reduce the melting temperature. For example, if the light element is sulfur, the melting temperature for the eutectic composition will be ~ 4500 K at the outer–inner core boundary and ~3500 K at the core–mantle boundary (Jeanloz, 1990).

In addition to these broad constraints, more detailed estimates of temperature profiles can be made using petrological studies of mantle xenoliths (fragments of rocks carried up by volcanoes) and from the phase diagrams determined by high-pressure experiments (see Sec. 1.2.2.2). Also, some physical properties such as electrical conductivity or seismic-wave velocities (see Sec. 1.1.1, Fig. 2) can be used to estimate the temperature profiles in the deep interior of the Earth. The temperatures in the upper mantle and the transition zone have been estimated using these techniques (Fig. 6) (Jeanloz and Morris, 1986). The following points can be noted.

1. The temperature gradient changes from steep in the upper portion of the mantle to gradual in the deeper portion. The temperature gradient in the deeper mantle (400–670 km and below) thus estimated is close to the adiabatic gradient (0.3–0.5 K/km).
2. The temperature in the upper mantle be-

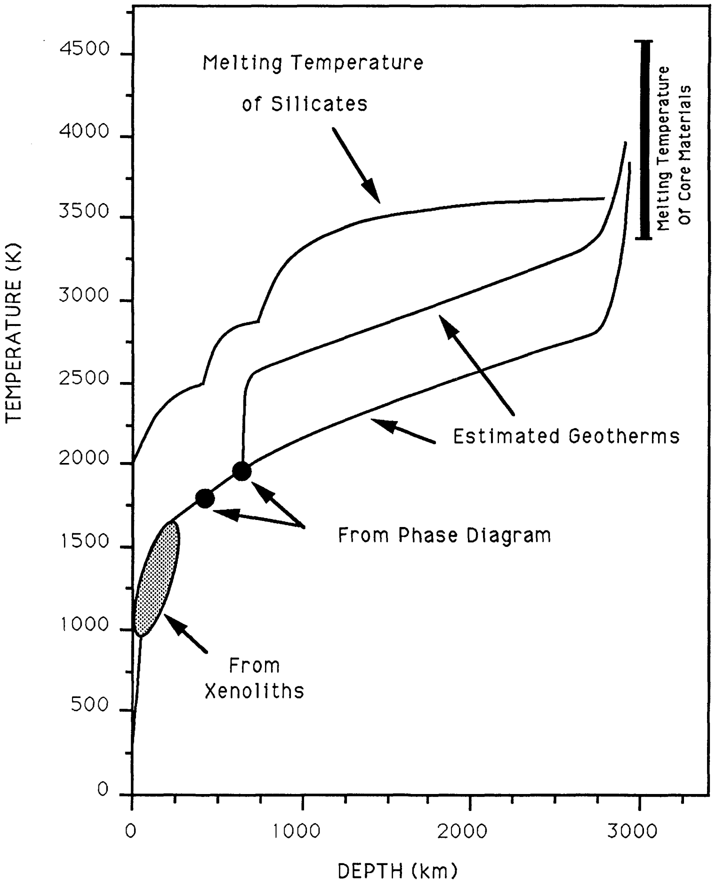

FIG. 6. Average temperature profiles for the Earth's interior. Shown together are the depth variation of melting temperatures of silicates and the melting temperature of iron at the core–mantle boundary determined by laboratory experiments. The two curves in the lower mantle show two possible temperature profiles, one (colder one) corresponding to whole-mantle convection, the other (hotter one) corresponding to layered convection with a thermal boundary layer at ~670 km.

low young ocean is higher than the temperature in a stable continent, but they are indistinguishable when the age of the ocean exceeds ~100 millions years.

When the temperature at ~670 km (~2000 K) is extrapolated to the core–mantle boundary assuming an adiabatic gradient (as suggested by the seismic-wave velocities in the greatest part of the lower mantle), one gets the temperature of ~2500–3000 K, which is lower than the melting temperature of the core materials (4000±500 K). The implication is that there must be some layers in the Earth below 670 km where the temperature gradient is significantly higher than the adiabatic gradient. Possibilities include thermal boundary layers near 670 km and/or at the bottom of the mantle (D'' layer).

The uncertainties in estimated temperatures become larger as we go deeper in the Earth. The uncertainties in the estimated temperatures in the upper mantle and transition zone are ~ ±100 K. The major uncertainty in the temperatures in the lower mantle concerns the presence or absence of thermal boundary layers in the transition zone and this yields ~ ±500-K uncertainties in the estimated temperatures. The uncertainties in the temperatures in the core come from the uncertainties in the melting temperature of core materials and are more than ±1000 K.

2.4 Thermal History

The previous discussions are concerned mainly with the Earth's present temperature profiles. The Earth's temperatures, however, have changed with time since the rates both of heat generation and of heat loss change with time. The major energy sources for the Earth's dynamics are gravitational energy and the heat produced by radiogenic elements. A significant amount of gravitational energy was released during the early stage of evolution of the Earth (during accretion and core formation) and probably throughout geological history as a result of growth of the inner core. The amount of gravitational energy released during accretion is proportional to R^5 (R is the radius of the planet), and therefore the bigger the planet, the larger this effect. The net temperature increase due to gravitational energy depends also on the rate of energy release and the efficiency of heat loss. When a planet is formed quickly, the heat is largely retained in the planet and a significant temperature increase will occur. Also, when a planet has a thick atmosphere, it will reduce the efficiency of energy loss and increase the temperature. Current theoretical studies on the formation processes of Earth suggest that a significant melting leading to the formation of a magma ocean may have occurred during the accretion and the core-formation processes.

The heat production due to radiogenic elements is determined by their concentration and the half-life times. Important radiogenic elements include ^{235}U, ^{238}U, ^{232}Th, ^{40}K and possibly ^{26}Al. ^{26}Al has a short half-life (0.74 million years) and it may have been important in the early stage of the solar nebula. All of other elements have half-lives on the order of a billion years and their activity has been decaying slowly.

The time variation of the Earth's average temperature T is determined by the balance of heat generation and heat loss as

$$\rho C_p V \frac{dT}{dt} = H^+(t) - H^-(t), \qquad (12)$$

where ρ is the density, C_p the specific heat, V the volume of the Earth, and H^+ and H^- the rates of heat generation and heat loss, respectively. The present-day energy loss rate can be estimated from the heat-flow measurements (see Sec. 2.2) to be $\sim 4\times10^{13}$ W. The heat generation due to radioactivity at present can be estimated if the bulk composition of the Earth is assumed to be the same as a primitive chondrite (see Sec. 1.1.2), which yields $\sim 2\times10^{13}$ W. The imbalance between the two indicates that the Earth has been cooling through time. Both the release of gravitational energy and the heat production due to radiogenic elements were more significant in the early stage than they are today. Thus the temperature in the past should have been higher than it is now. However, the rate of temperature change is also controlled by the energy loss mechanisms. The most important energy loss mechanism is the convective heat transfer through the mantle (see Sec. 4). The efficiency of heat loss due to this mechanism depends on temperature because the viscosity of rocks depends strongly on temperature (see Sec. 4.2): the higher (lower) the temperature, the higher (lower) the rate of cooling.

In this way, the temperature-dependent viscosity of rocks tends to stabilize the temperatures in the Earth. Thus a rather moderate temperature drop of ~100–300 K is estimated from the Archean (<2.5 billion years) to the present.

3. MELTING OF SILICATES AND MAGMATISM

(See also VOLCANICS)

3.1 Melting and Chemical Stratification

Melting is a very efficient way to create chemical stratification of the Earth. The major chemical boundaries of the Earth, namely the crust–mantle boundary, the mantle–core boundary, and the outer-inner core boundary, are considered to have been formed through some processes involving melting. Melting can create chemical stratification in two different ways. First, when partial melting occurs, melt may migrate up or down depending on its density relative to the host solid rock. Second, when total melting occurs, fractional crystallization in a melt will create a chemical stratification. The former process is believed to be responsible for the major volcanic activity near the surface, including that at mid-ocean ridges. The latter is believed to occur at the inner and outer core boundary and in magma chambers (or the hypothetical magma ocean).

When partial melting (or fractional crystallization from melt) occurs, selective enrichment or depletion of some elements occurs in the melt (or the solid phases). The nature of element partitioning between melt and solid phases depends strongly on the crystal structure of the solid phases and the degree of partial melting (or crystallization). Thus the measurements of abundance patterns of various elements in rocks provides important information as to the degree of melting (or crystallization) and the crystal structure of solid phases involved (and hence the depth of melting).

The melting temperatures of most Earth materials increase with pressure (Fig. 6). However, the rate of increase in melting temperature within the stability field of a given phase becomes small at high pressures because of the large increase in density of melt relative to solid (melts are more compressible than solids). Secondary components sometimes have important effects on melting. An example is the effect of water on the melting of silicates: Water dramatically reduces the melting temperature and changes the chemical composition of the melt.

3.2 Magmatism and Melting of the Earth's Interior

3.2.1 Mafic Magmas. Magmas enriched in (Mg,Fe)O are called mafic magmas, and the typical one is basaltic magma. The most important igneous rock formed by the solidification of this type of magma is mid-ocean ridge basalt (MORB), and its production rate is approximately ~4.5 km^3/year. MORB comprises a major part of the oceanic crust that descends into the deep interior of the Earth at a subduction zone.

Some ocean islands such as Hawaii are composed of a somewhat different igneous rock called ocean island basalt (OIB), which shows enrichment in incompatible elements (elements that are selectively concentrated into the liquid phase upon partial melting). These rocks are thought to have originated in relatively undepleted regions (or mantle enriched in incompatible elements). In most cases, the volcanoes which supply these basalts remain at the same position for a geologically long period of time, while the surface plates move a significant distance. Thus these volcanoes are often referred to as "hot spots." The production rate of this type of magma is ~0.1 km^3/year.

3.2.2 Silicic and Intermediate Magmas. Magmas enriched with silica (e.g., granite) are called silicic, and magmas with a composition between silicic and mafic are called intermediate. Andesite is a typical igneous rock formed by an intermediate magma. These magmas are formed in the island arcs or the active continental margins, and they are generated by partial melting of the crust or the mantle in the presence of water. Silicic and intermediate magmatism is responsible for the growth of continents and island arcs. The production rate is ~1 km^3/year.

3.2.3 Ultramafic Magmas and Melting in the Deep Mantle. Komatiite is a unique igneous rock formed by the solidification of magmas which are extremely enriched in

(Mg,Fe)O (ultramafic magma). They occur mainly in the Archean period (more than ~2.5 billion years ago). Magmas with this composition indicate a high degree of partial melting and therefore high temperatures in the Archean period.

Theoretical studies of thermal history (see Sec. 2.4) strongly suggest extensive melting in the early stage of the Earth. Large-scale melting in the early stage of Earth's history could have generated a thick layer of ultramafic magmas (magma ocean). The consequences of such extensive melting can be inferred from experimental studies on melting at high pressures, and the results can be compared with geochemical and geophysical observations to shed some light on the degree of chemical differentiation in the Earth.

Melting experiments under ultrahigh pressure indicate that the chemical composition of the upper mantle is similar to the composition of partial melt that is formed at high pressures (13–15 GPa) and that the solidus phase (a solid phase that crystallizes first from the melt upon cooling) changes from olivine at low pressure to majorite at ~15 GPa, and finally to perovskite at ~25 GPa. Partial melting or crystallization involving majorite or perovskite results in quite different (trace) element partitioning than those involving upper mantle minerals (such as olivine or pyroxenes). Therefore some stratification, in both trace- and major-element chemistry, might have been formed in the mantle if large-scale melting had occurred in the early stage of Earth's history.

Another important observation in ultrahigh-pressure melting is that the density differences between melt and solid become smaller at high pressures. This eventually leads to density crossover: Melts will become denser than solids. If this density crossover occurs in the Earth, the transport of magmas from deep interior of the Earth would become difficult.

The consequences of large-scale melting (such as magma ocean) for chemical stratification depend on the processes of melt crystal separation as well as on the density contrasts and the partitioning of elements between melts and crystals. It has often been assumed that large-scale melting inevitably results in chemical stratification, but recent theoretical studies suggest that very limited chemical stratification might result because of a stirring effect of vigorous convection in the magma ocean.

4. EARTH'S RHEOLOGICAL STRUCTURE AND DYNAMICS OF THE EARTH'S INTERIOR

(See also PLATE TECTONICS)

4.1 Time-Dependent Deformation of the Earth

The Earth's dynamics is governed by the balance of the driving force determined by the density heterogeneity and the resistance force determined by rock rheology. The Earth's rheological properties depend on time scales and on the stress magnitude. At short time scales and for small deviatoric stresses, Earth materials behave like an elastic solid (with minor energy dissipation). When the time scale and/or the stress magnitude are large, nonelastic deformation will result, leading to permanent deformation that causes tectonic activities such as mountain building and earthquakes. The transition conditions from elastic to nonelastic therefore involve two criteria, the time-scale criterion and the stress criterion (Fig. 7).

To understand this, we must first understand the two contrasting mechanisms of nonelastic deformation. The first one is brittle deformation, that is, fracture. Fracture occurs mostly along a fault, and therefore the fracture strength of rocks is determined by the resistance for sliding along a fault. The strength σ of a rock in the brittle regime is given by

$$\sigma=\sigma_0+\mu_f\sigma_n, \tag{13}$$

where σ_0 is the cohesive strength, μ_f the friction coefficient, and σ_n the normal stress on the fault. Laboratory studies show that both σ_0 and μ_f are relatively independent of rock type, temperature, pressure, and strain rate. However, the strength in a brittle regime increases with pressure, since σ_n is proportional to pressure.

The second mechanism is ductile deformation. Ductile deformation occurs, in most cases, through the motion of crystalline defects in minerals. The motion of these defects is thermally activated, and therefore ductile deformation is important only at high temperatures. The easiness of the motion of defects depends on the strength of chemical bonds and crystal structure, and therefore the ductile strength depends on the rock type. To a first approximation, the dependence on rock

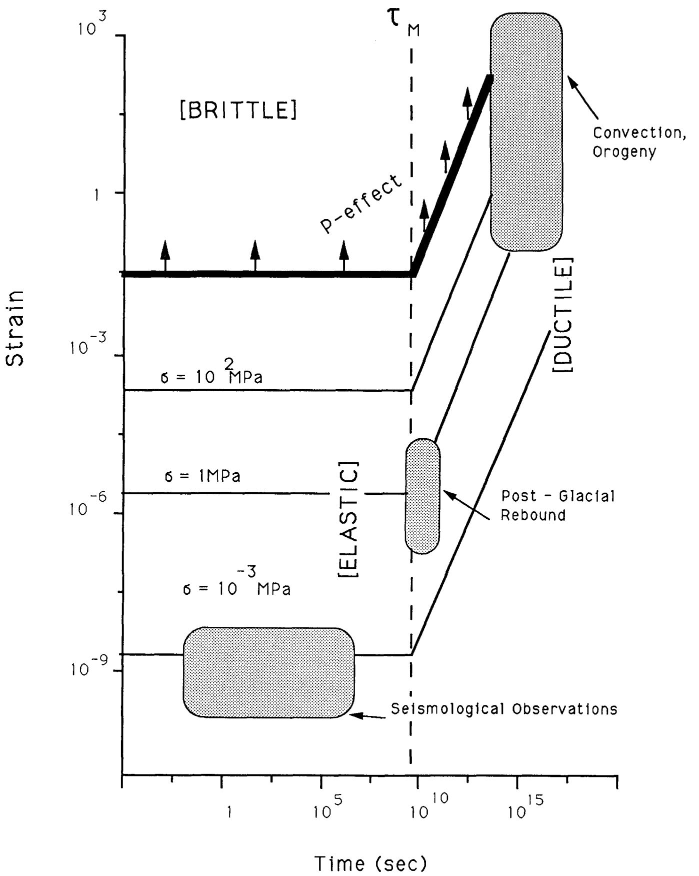

FIG. 7. Mechanical behavior of rocks as a function of time scales and strain (stress) magnitude. Constant stress contours are also shown for both elastic and ductile regimes. When the stress magnitude exceeds a critical value, a material will deform in the brittle manner (fracture), and this critical stress increases with pressure ("*P*-effect"); the critical stress shown in this figure corresponds to the stress in the continental lower crust and the uppermost mantle. Ductile deformation is important compared to elastic deformation when the Maxwell time $\tau_M=\eta/\mu$ (η, viscosity; μ, rigidity) is smaller than the time scale of the relevant process. The Maxwell time depends on temperature and pressure and the value shown in this figure is for the typical asthenosphere. Shown together are the strain magnitudes and time scales of typical deformation processes in the Earth.

type is through the melting temperature, and the flow law is given by

$$\dot{\epsilon}=A\sigma^n \exp[-\beta T_m(P)/RT], \tag{14}$$

where $\dot{\epsilon}$ is the strain rate, σ the differential stress, A a constant preexponential factor, n the stress exponent ($n=1$–5), $T_m(P)$ the melting temperature which depends on pressure, and β a nondimensional constant. The effective viscosity η is defined by

$$\eta=\sigma/\dot{\epsilon}. \tag{15}$$

For Newtonian rheology (i.e., $n=1$), the effective viscosity is independent of stress, whereas for non-Newtonian rheology ($n>1$), the effective viscosity decreases with the stress magnitude. Ductile deformation is important compared to elastic deformation when the Maxwell time

$$\tau_M=\eta/\mu \tag{16}$$

is smaller than the time scale of deformation, where μ is the rigidity. Thus ductile deformation is important in slow processes and/or at high temperatures (see Fig. 7).

Ductile deformation results in changes in rocks' microstructures reflecting deformation conditions. They include the formation of preferential alignment (preferred orientation) of minerals and the creation of crystal dislocations. The preferential alignment of minerals can be observed as an anisotropy in seismic-wave velocities as well as in thin sections of rocks under a microscope. The density of crystal dislocations and some other microstructures can be observed by the analysis of rock specimens and provide an estimate for the magnitude of differential stress in the Earth.

Brittle and ductile deformation are largely independent, and the lower of the strengths associated with these processes will determine the real strength. Thus the mechanical properties of rocks can be summarized as in Fig. 7. At low stresses and short time scale (or low temperatures), rock behaves as an elastic solid. At high stress and short time scales (or low temperatures), rocks deform in a brittle manner. Finally, at long time scales (or high temperatures), rocks deform in a ductile fashion.

4.2 Earth's Rheological Stratification

The Earth's rheological stratification can be predicted on the basis of temperature profiles (Fig. 6) and the constitution and the mechanical properties of rocks [Eq. (14)]. The temperature is low in the near-surface regions of the Earth, and therefore rocks are elastic or brittle depending on the stress levels. The strength of rocks in the near-surface layer increases with depth as a result of increases in pressure. As one goes to deeper portions, temperatures will increase, and this enhances ductile deformation. The depth of onset of ductile deformation depends on the rock type and the temperature profile. In the continental lithosphere, the lower crust will be a ductile layer with strength significantly smaller than that of the upper crust and the uppermost mantle. In the oceanic region, the crust is mostly a brittle layer and ductile deformation occurs only in the upper mantle. The ductile strength of the upper mantle first decreases with depth because of the rapid increase in temperature in near surface regions (Fig. 6). However, the temperature gradient becomes small at the deeper portions, and then the pressure effect (through the pressure dependence of melting temperature) becomes more important, which increases the ductile strength with depth. Thus one expects a strength minimum at depths where the temperature gradient changes from a near-surface high gradient to a low gradient at the deeper portions (this depth depends on the temperature profile and is ~50–200 km in the oceanic mantle; see Fig. 6). At even deeper portions where a series of phase transformations occur in major constituent minerals, rheological properties will also change. Recent laboratory studies suggest that the garnet and spinel-rich transition zone (400–670 km) presumably has a larger viscosity than the olivine-rich upper mantle or the perovskite-rich lower mantle. Thus a rheological stratification of the Earth is a natural consequence of temperature and pressure distribution and of the phase transformations in the Earth.

The Earth's rheological properties can be inferred from some geological and geophysical observations. Attenuation of seismic waves provides information as to the Earth's non-elastic properties at relatively short time scales ($1\text{–}10^3$ s). The attenuation of seismic energy is measured by a quality factor Q defined as

$$Q^{-1}=\Delta E/2\pi E, \tag{17}$$

where ΔE is the energy dissipated per unit cycle, and E is the energy stored per unit cycle. Figure 8 shows the depth variation of Q in the mantle which is characterized by a high-Q (low-loss) lithosphere (0–100 km), low-Q (high-loss) asthenosphere (100–200 km), and a moderate-Q lower mantle.

Time-dependent deformation after the last deglaciation provides the most detailed information as to Earth's rheological stratification. The melting of polar ice sheets ~10 000 years ago changed the distribution of surface load,

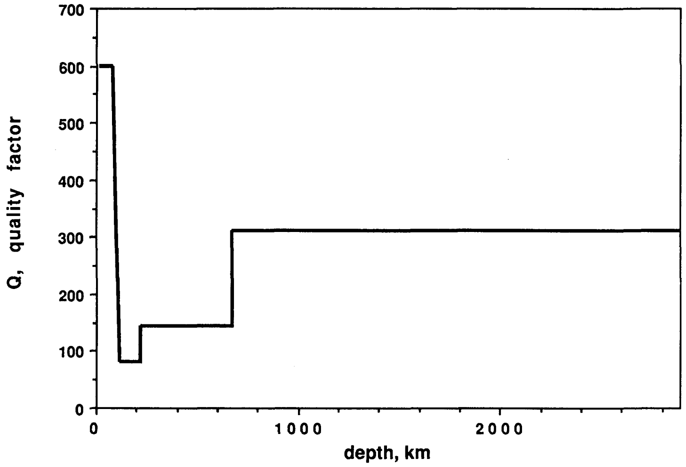

FIG. 8. Depth variation of seismic wave attenuation (Q, quality factor).

which induced plastic flow in the Earth. The resultant deformation is still going on today as observed by the change in sea levels at the coast or by the change in the spin rate of the Earth (as a result of change in the moment of inertia due to internal mass redistribution). Theoretical analyses of such deformation show that the near-surface layer of the Earth behaves like a thin elastic shell (the lithosphere), whereas the deeper portions behave like a viscous fluid.

Another observation related to mantle rheology is the geoid anomalies (anomalies of the height of equipotential surface of the Earth's gravity field). Such anomalies reflect the anomalies of mass distribution in the Earth. Mass anomalies with a long wavelength are supported by viscous flow and therefore provide a clue as to the rheological properties of the Earth's interior. Theoretical analysis of the geoid anomalies associated with subduction of the oceanic lithosphere suggests a significant increase in viscosity in the deep part of the mantle.

Most of these studies assume Newtonian rheology [linear relationship between strain rate and stress, i.e., $n=1$ in Eq. (14)]. The viscosity of the mantle thus inferred is shown in Fig. 9. The average viscosity of the Earth's mantle is well constrained from these studies ($\sim 10^{21}$ Pa s), but the depth variation is poorly resolved. There are some indications that the asthenosphere has lower viscosity and the transition zone and the lower mantle have higher viscosity than the average.

4.3 Dynamics of the Earth's Interior

4.3.1 Plate Tectonics (see also PLATE TECTONICS). The estimation of the Rayleigh number defined by Eq. (11) for the Earth's mantle shows that it significantly exceeds the critical number ($\sim 10^3$), for both the upper (or the lower) and the whole mantle (see Sec. 2.1). Thus the Earth's mantle must convect vigorously. The relative motion of the near-surface layer (lithospheric plates) is the manifestation of mantle convection. Several geophysical or geological observations indicate that Earth's surface is divided into several plates which move relative to each other with the speed of 1–10 cm/year. A lithospheric plate has a large creep strength at the Earth's surface, and deformation within a plate is minor except for continental collision zones such as Tibet. Most of the geological or geophysical processes such as earthquakes, volcanism, and mountain building occur at the boundaries of plates.

Plates that carry mostly oceans (such as the Pacific plate) are created at ocean ridges and subduct into the Earth's deep interior at ocean trenches. Subducting plates are marked by the planar zone of high seismic activity called the Wadati–Benioff zone which extends to ~ 700 km in some areas (e.g., beneath the Japan, Tonga, and Kermadic trenches).

4.3.2 Seismic Tomography and Deep-Mantle Dynamics. Although the surface motion of plates is well established, the relation between plate motion and mantle convection is

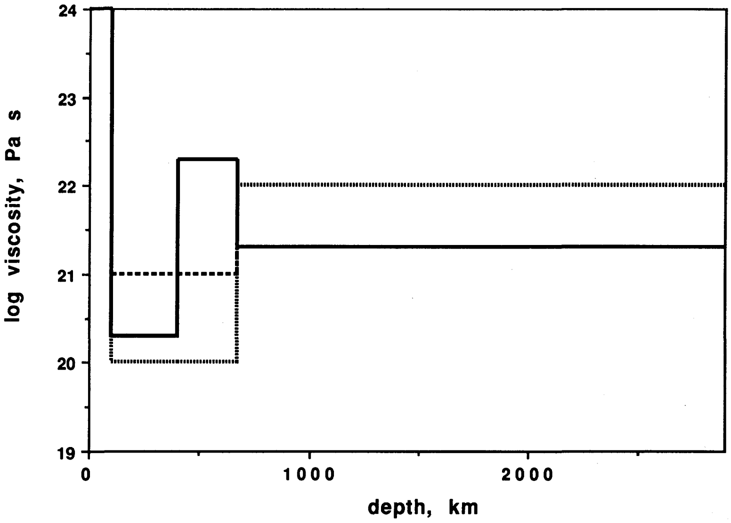

FIG. 9. Depth variation of viscosity of the mantle. The three curves are the viscosity-depth profiles consistent with various time-dependent deformation (such as post-glacial rebound and geoid anomalies), and show the range of uncertainties involved in these studies. Average viscosity of the mantle is determined to be $\sim 10^{21}$ Pa s by the analysis of the post-glacial rebound. The existence of a low-viscosity layer at $\sim$100–200 km and of relatively high-viscosity layer(s) in the deep mantle are suggested by laboratory studies and from the analysis of geoid anomalies. However, the viscosities in the deep mantle (below 400 km) are poorly constrained.

not well known. One of the major issues is the depth extent of subduction and the style of mantle convection. One hypothesis is that the subducting slabs penetrate into the lower mantle and the convection occurs in a single cell as a whole mantle. The other alternative is that the subducting slabs do not penetrate deeper than ~700 km, implying that the upper mantle (including the transition zone) and the lower mantle convect separately.

One of the powerful ways to distinguish these hypotheses is to study the fine structure of the mantle using seismological techniques. This is called seismic tomography, in which fine details of three-dimensional structures of the Earth can be revealed. In seismic tomography, deviation from a simplified layered Earth model such as PREM is mapped. There are two different types of deviations: lateral heterogeneity and anisotropy.

The lateral variations of temperature and chemistry are the most likely causes of lateral heterogeneities. When the major source of lateral heterogeneity is the heterogeneity in temperature, low velocity reflects high temperature and high velocity reflects low temperature.

The correlation of lateral heterogeneity with tectonic setting is clear in the shallow portions [Plate I(a)]. Ocean ridges and young oceanic lithosphere show up as low-velocity anomalies, and old continents such as North America and Africa show up as high-velocity anomalies. The origin of velocity anomalies in the deep portions is not clear, but an important example is the depth extent of subducting slabs [Plate I(b)]. A cold slab is expected to show up as a high-velocity anomaly. The recent studies on slab penetration suggest that slabs do penetrate deep into the lower mantle in some areas (e.g., Kamchatka-Kuril) but they are deflected at the ~670-km boundary and do not penetrate into the lower mantle in other areas (e.g., under the Izu Bonin island arc in western Pacific).

Another fine structure that is related to Earth's dynamics is anisotropy. Most minerals have anisotropic elastic properties, and therefore if they are aligned nonrandomly in a rock, the rock will show anisotropy. Such anisotropy was first observed in the oceanic upper mantle in 1960s and was interpreted in terms of the preferential alignment of olivine as a result of mantle flow related to plate tectonics. Such studies have been extended during the past few years to probe the anisotropy in the deeper portions of the Earth and on a global scale. In most cases, the direction of the fast seismic-wave velocity in the upper mantle is parallel to the flow direction. Therefore an anisotropy map such as Plate I(c) is considered to show the pattern of mantle convection.

Computer modeling of mantle convection plays an important role in the study of mantle dynamics. The equation of motion of a viscous fluid (Navier–Stokes equation) for a given set of material parameters (such as viscosity, density, and thermal expansion) may be solved with appropriate boundary conditions. The results of such studies can be used to predict possible convection patterns for a given set of material parameters and boundary conditions. Alternatively, such calculations may be used to put constraints on some material parameters such as depth variation of viscosities, by comparing the predicted results for various parameter sets with geophysical observations. These studies indicate that the material properties such as viscosity vary significantly with depth and that these variations in physical properties with depth have critical effects on the convection pattern.

4.3.3 Geochemical Observations on Mantle Heterogeneities. The best documented large-scale chemical heterogeneity of the Earth's mantle is the differences in trace-element abundance pattern between the source of the mid-ocean ridge basalt (MORB) and ocean island basalt (OIB). MORBs' trace-element abundance pattern shows that they originate in a region where incompatible elements are depleted. This indicates that MORBs are formed by the partial melting of rocks that have already undergone significant differentiation. In contrast, OIBs originate from a region which is enriched (or undepleted) in incompatible elements. Geochemical observations also indicate that these two regions have been separate for the time scale of a billion years. However, geochemical observations do not tell where these source regions are located. One possibility is that the depleted region corresponds to the upper mantle, and the undepleted region to the lower mantle. An alternative hypothesis is that these two regions are mixed like raisin bread, and that the mantle is homogeneous on the large scale.

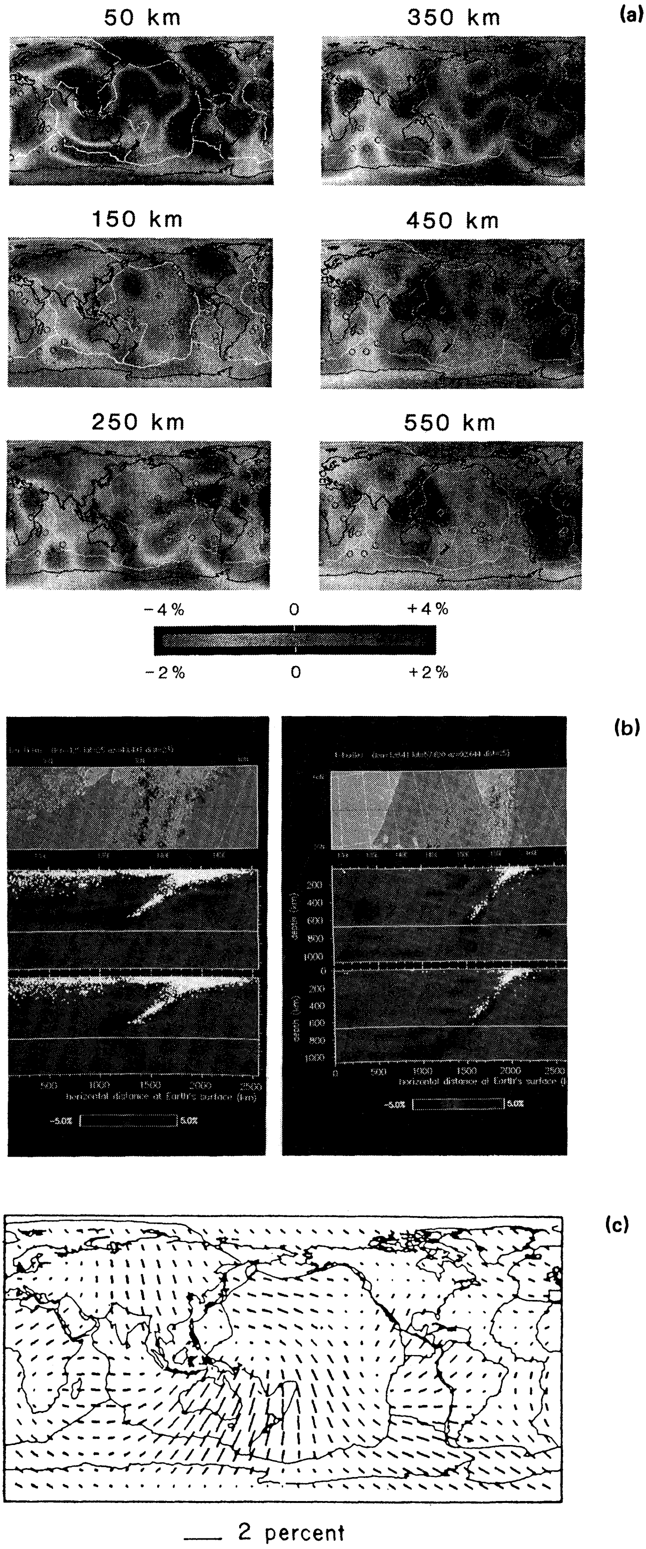
(a)
50 km
350 km
150 km
450 km
250 km
550 km
−4%
0
+4%
−2%
0
+2%
(b)
depth (km)
200
400
600
800
1000
0
500
1000
1500
2000
2500
horizontal distance at Earth's surface (km)
−5.0%
5.0%
(c)
2 percent

PLATE 1. Some results of seismic tomography. (a) Global maps of lateral heterogeneity of seismic wave velocities (distribution of anomalies in shear wave velocities at various depths in the upper mantle and transition zone; from Dziewonski and Woodward, 1991). Note that the scale for the velocity anomalies varies with depth. (b) Anomalies of compressional wave velocities in the vicinities of subduction zones (Izu-Bonin and Kurile). Top, locations of earthquakes; middle and bottom, velocity anomalies based on two different standard models together with the foci of earthquakes (shown by white dots) (van der Hilst *et al.*, 1991). High velocity anomalies (i.e., a subducting slab) appear to be deflected horizontally at ~670 km in the Izu-Bonin arc but they seem to be continuous down to ~1000 km in the Kurile arc. (c) Anisotropy of Rayleigh-wave (with period 167 sec) velocities in the asthenosphere. The directions of the bars indicate the directions of the fast Rayleigh waves, which is parallel to the orientation of convective motion (Montagner and Tanimoto, 1990).

The whole-mantle convection is consistent with the latter idea, whereas the layered-mantle convection is consistent with the former idea.

While the chemical heterogeneity is evident in incompatible elements, the abundance pattern of compatible elements (Mg, Si, Ca, Al, Ti, and heavy rare-earth elements) of mantle xenoliths, sources of MORBs and OIBs, show little heterogeneity. This indicates that major fractionation involving melting in the transition zone or in the lower mantle did not occur, or else vigorous convective mixing occurred after or during the differentiation.

5. SUMMARY

The Earth is a chemically differentiated and dynamically evolving planet. The most significant differentiation processes are the formation of the crust and the core. The crust is composed of silica-rich rocks, the core is composed mainly of iron-rich alloys, and the mantle is composed of (Mg,Fe)O–SiO_2 rich rocks. There are some fine details in the structure of each layer, an important example of which is the discontinuities in the mantle at 400- to 670-km depth range. High-pressure experiments show that most silicate minerals undergo a series of phase transformations at this depth range. Thus a significant part of these discontinuities is due to phase transformations. A remaining question is whether all of these discontinuities are attributed to the phase transformations or some of them are due to changes in chemical composition. Geophysical and geochemical studies are under way to solve this problem.

The Earth is stratified also in terms of rheological properties: strong near-surface layer (lithosphere), solid but relatively weak underlying mantle, liquid outer core, and solid inner core. Convection occurs both in the mantle and in the outer core (and presumably in the inner core). Convection in the mantle is responsible for plate tectonics, and convection in the outer core is responsible for the generation of the geomagnetic field. The nature of convection in the mantle is determined by the balance of the driving force and the resistance force for convection, both of which are determined by the physical properties of rocks such as density and rheological properties. These physical properties change significantly with depth as a result of phase transformations and the increase in pressure and temperature, leading to complicated convective motion in the Earth. A realistic picture of convective motion in the mantle is beginning to be inferred through the development of geophysical (seismic tomography) and geochemical observations combined with materials science studies of the properties of minerals under high pressures and temperatures.

GLOSSARY

Adams-Williamson Equation: The equation that describes the compression of a homogeneous body in a gravity field.

Archean: The period earlier than 2.5 billion years in the Earth's history. The earliest period from 4.6 to 4.0 billion years ago is sometimes called Hedean. If we adopt this classification, the Archean is defined as the period from 4.0 to 2.5 billion years.

Asthenosphere: A soft layer below the lithosphere (50- to 200-km depth).

Basalt: A fine-grained, mafic igneous rock composed largely of plagioclase feldspar and pyroxenes.

Brittle Deformation: Deformation due to fracture. Brittle deformation occurs at low temperatures and/or low pressures.

Carbonaceous Chondrite: One type of meteorite, which is mainly composed of silicates. It contains various minerals that are chemically in disequilibrium. This type of meteorite is believed to be the most primitive material in the solar system.

Ductile Deformation: Deformation due to plastic flow. Ductile deformation of rocks occurs mostly at high temperatures (and high pressures).

Geoid: The equipotential surface of the Earth's gravity field.

Incompatible Elements: The elements which tend to be partitioned into magmas relative to coexisting minerals upon partial melting. These elements are not accommodated in lattice sites of the minerals because of mismatching of the ionic radii with the lattice site.

Inhomogeneity Parameter: A parameter that indicates the deviation from adiabatic compression of a homogeneous body.

Komatiite: An igneous rock with ultramafic composition. The eruption of such ultramafic lavas is mainly observed in the Archean period (more than 2.5 billion years ago).

Lithosphere: Near-surface strong layer of the Earth (about 50–100 km thick). The (old) oceanic lithosphere is strong and moves without much deformation. The continental lithosphere is weak and is deformed when a significant force is applied.

Magma Ocean: A hypothetical thick layer of magmas that may have occurred on the Moon or on the Earth in their early history.

MORB: Mid-Oceanic Ridge Basalt. The basalt generated at the mid-ocean ridges. The oceanic crust is mainly composed of this basalt.

OIB: Ocean Island Basalt, which forms volcanoes in ocean islands (e.g., Hawaii). Its major-element composition is similar to MORB, but its trace-element composition is characterized by the enrichment of incompatible elements.

Peridotite: A ultramafic rock composed mainly of olivine and pyroxenes.

Plate Tectonics: A unified theory of the Earth's tectonics in which the relative motion of the near-surface layer (lithospheric plates) is responsible for major tectonic activities.

Pyrolite: A hypothetical primitive mantle rock proposed by A. E. Ringwood. Its composition corresponds to a mixture of basalt and peridotite. The mixing ratio is rather arbitrary but a typical one is one part basalt and three parts peridotite.

Q: The measure of energy dissipation. Q^{-1} represents the fraction of energy dissipated per unit cycle of deformation.

Seismic Tomography: Seismological technique to reveal three-dimensional fine structures of the Earth.

Wadati–Benioff Zone: A planar zone beneath ocean trenches where seismic activity is high. The Wadati–Benioff zone represents a subducting oceanic lithosphere.

Xenolith: A fragment of rock that is carried up to the surface by volcanoes. Xenoliths provide useful information as to the temperature and chemical composition of the Earth's deep interior.

List of Works Cited

Anderson, D. L. (1989), *Theory of the Earth,* Boston: Blackwell.

Dziewonski, A. M., Woodward, R. L. (1991), *Acoust. Imaging* **19**, 785–797.

Jeanloz, R. (1990), *Annu. Rev. Earth Planet. Sci.* **18**, 357–386.

Jeanloz, R., Morris, S. (1986), *Annu. Rev. Earth Planet. Sci.* **14**, 377–415.

Lay, T. (1989), *EOS, Trans Am. Geophys. Union* **70**, 40–59 .

Montagner, J.-P., Tanimoto, T. (1990), *J. Geophys. Res.* **95**, 4797–4819.

Ringwood, A. E. (1979), *Origin of the Earth and Moon,* New York: Springer-Verlag.

van der Hilst, R., Engdahl, R., Spakman, W., Nolet, G. (1991), *Nature* **353**, 37–43.

Further Reading

Anderson, D. L. (1989), *Theory of the Earth,* Boston: Blackwell. A unique synthesis of geophysical and geochemical observations on the Earth's structure and evolution.

Fowler, C. M. R. (1989), *The Solid Earth,* Cambridge, U.K.: Cambridge Univ. Press. An introductory account of solid earth geophysics with emphasis on geodynamics.

Jeanloz, R., and Morris, S. (1986), *Annu. Rev. Earth Planet. Sci.* **14**, 377–415. An overview of temperature profiles of the Earth's interior from experimental and theoretical viewpoints.

Peltier, W. R. (Ed.) (1989), *Mantle Convection,* New York: Gordon and Breach. A collection of review papers on the Earth's global dynamics including theoretical as well as experimental studies.

Poirier, J.-P. (1991), *Introduction to the Physics of the Earth's Interior,* Cambridge, U.K.: Cambridge Univ. Press. A succinct review of fundamental condensed matter physics to understand the internal structure of the Earth.

Shankland, T. J., Bass, J. D. (1988), *Elastic Properties and Equations of State,* Washington D.C.: American Geophysical Union. A collection of classical papers on elasticity and equation of state and the internal constitution of the Earth.

Verhoogen, J. (1980), *Energetics of the Earth,* Washington D.C.: National Academy. A brief discussion of physical processes that determine the Earth's thermal regimes.

ELECTRIC DISCHARGES

See SPARKS, ARCS, AND OTHER ELECTRIC DISCHARGES

ELECTRIC POWER ENGINEERING

SY ALPERT, BRENT BARKER, RALPH SAMM, AND ED ZEBROWSKI, *Electric Power Research Institute, Palo Alto, California, U.S.A.*

Introduction 1
1. **Role of Electric Power in the Energy Economy** 5
2. **Structure of the Electricity Industry** 6
2.1 New Trends in the Electric Power Industry 7
3. **Supplying Electricity** 8
3.1 The Pulverized-Coal–Fired Plant 10
3.2 The Atmospheric Fluidized-Bed Combustion Plant 10
3.3 The Pressurized Fluidized-Bed Combustion Plant 10
3.4 The Integrated Coal Gasification Combined-Cycle Plant 11
3.5 The Integrated Gasification Molten-Carbonate Fuel Cell 12
4. **Nuclear Power** 13
4.1 The Boiling-Water Reactor 14
4.2 The Pressurized-Water Reactor 16
4.3 The Future of Nuclear Power 17
5. **Transmission and Distribution** 17
5.1 Flexible ac Transmission Systems 17
5.2 Energy Efficiency in Transmission and Delivery 18
6. **Power Applications of Superconductivity** 18
6.1 Generators and Motors 19
6.2 Transmission Lines 19
6.3 Superconducting Magnetic Storage 19
6.4 Transformers 20
6.5 Other Applications of Superconductors 20
7. **End-Use Technologies** 20
7.1 Electric Motors 20
7.2 Lighting 21
7.3 Cooling and Refrigeration 21
7.4 Space and Water Heating 22
7.5 Other End-Use Technologies 22
8. **Environmental Issues** 22
8.1 Acid Rain 22
8.2 Air Quality and Human Health 23
8.3 CO_2 and Global Climate 23
8.4 Electric and Magnetic Fields 23
8.5 Nonradioactive Waste Disposal 24
8.6 Nuclear Waste Disposal 24
9. **Storage Technologies** 24
9.1 Lead-Acid Battery 25
9.2 Pumped Hydroelectric Storage 25
9.3 Compressed-Air Energy Storage 25
9.4 Superconducting Magnetic Energy Storage 26
9.5 The Future of Energy Storage 26
10. **Summary** 26
Acknowledgments 27
Glossary 27
Further Reading 28

INTRODUCTION

Electricity is a unique energy intermediary. In most countries it is produced by conversion of chemical energy to thermal energy and then to mechanical energy to drive electric generators. The thermal energy is extracted mainly from a variety of carbon-based fuels that occur naturally. In several countries the thermal energy is derived mainly from uranium (France, Belgium, Spain, Sweden, Taiwan). Solar energy is used via four methods: hydroelectric, photovoltaic, and solar thermal generation, and heat pumps. Wind power

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

and wave power are used to capture the mechanical energy of wind or waves to drive electric generators.

The conversion of various forms of energy to electricity involves a wide variety of physical and chemical processes. Thermal cycles produce heat energy at high temperatures by the combustion of fossil fuels, by the fission of uranium and plutonium atoms, by the fusion of light elements, or by use of mirror concentrators for solar thermal energy. Still another variety of solar energy conversion accumulates solar heating in density-stratified brine pools. Heat energy can also be extracted from the earth as geothermal energy or from the ocean by using the temperature difference between warm surface layers and cooler deep water (ocean thermal).

Heat energy is converted to mechanical energy by the expansion of a working fluid, primarily through use of the Carnot cycle. The Brayton, Rankine, and Heber cycles have been used in special applications, depending on the working fluid and the temperature differences available.

Photovoltaic electricity involves the direct conversion of solar radiation to elevated electron energies and then to electric current. While there was some decline in commercial developments and prospects during 1985–1990, there are now important new developments that are reviving interest and prospects. These developments include multilayered solid-state structures that use a larger portion of the incident solar radiation spectrum, high-temperature concentrator systems with initial conversion efficiencies more than twice as high as conventional solar cells, and low-cost mass production of thin-film devices. As the performance and lifetime testing of such systems proceeds, there is a good prospect that solar energy will begin to fulfill some of the hopes that it has generated for many decades.

Magnetohydrodynamic generation involves the direct conversion of the mechanical energy of a conductive gas, plasma, or liquid moving through a magnetic field to a potential difference and electric current. When fusion devices become practical, a combination of magnetohydrodynamic (MHD) ducts for direct conversion of the plasma energy to electricity almost certainly will be used, plus thermodynamic "bottoming" cycles to make use of the by-product heat.

Mechanical energy is captured from the motion of wind or waves to drive electric generators. The chemical energy of hydrocarbon fuels is also converted directly into electric current in electrolytic cells (fuel cells). Electrolytic cells also mediate as a storage and transport medium in a wide variety of chemical batteries under active development for electric-powered vehicles.

Three types of electricity production dominate in industrialized countries: thermodynamic cycles with combustion of fossil fuels, thermodynamic cycles with fission heat sources, and hydroelectric dams. However, nearly all of the other approaches mentioned earlier have found some economic niches in which they can be economically useful. Strong economic and environmental incentives drive active developments in all of the technologies involved in electricity production, distribution, and use. Despite the seeming scientific and technological maturity of this field, there remain many important scientific problems in the basic physical and chemical processes, as well as in the applied technologies themselves.

The classical domain of power engineering has been the development and large-scale application of conversion and distribution technologies. Increasingly, these applications depend on collateral developments in the science and technology of the conservation of energy by end users; the refinement of designs, materials, and processes to minimize adverse environmental effects; and the improvement of the efficiency of utilization of electricity in residential, commercial, and industrial applications. As a uniquely convenient and controllable form of energy, electricity can be made available almost anywhere and in any amount through distribution cables. Most control actions can be completed in a few milliseconds, and much faster when necessary. Environmental impacts are largely associated with the generation processes, less so with distribution, and least of all at the end-user locations.

The convenience and control attributes often have led to new types of production techniques, and to an enormous array of consumer products, communication products, and a wide variety of convenience products that are taken-for-granted amenities in industrialized and developing countries. The almost universal use of refrigeration in food trans-

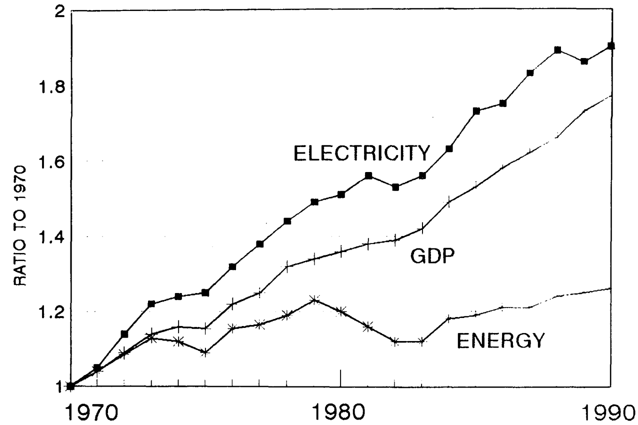

FIG. 1. Relative growth of energy and electricity use vs gross domestic product, for OECD countries, 1970–1990.

port and preservation has the effect of extending the supply of usable food for a given amount of farmland by roughly one-third, relative to the supply-to-land ratio without refrigeration. On the other hand, there is serious basis for concern that the productivity of agriculture is increasingly fossil-fuel intensive for both fertilizers and production. The concern is that this trend, together with expected escalation in oil and gas prices, will make food much more costly after the turn of the century, and will end the era of the United States as a net exporter of food (Gever *et al.*, 1989).

The application of comfort air conditioning has changed the economic geography of the United States by making living and working in warm, humid, or subtropical climates more feasible and attractive. The great demographic shifts from north to south in the United States are largely attributable to this factor.

The conversion and transport processes for electricity produced from heat energy or potential energy are costly relative to direct use of fuel for space heating or process heat. Yet the unique attributes of convenience, controllability, and cleanliness for the end user often make electricity preferable and also economically competitive for many applications that also could be accomplished by direct use of combustion fuels. The pervasiveness of uses

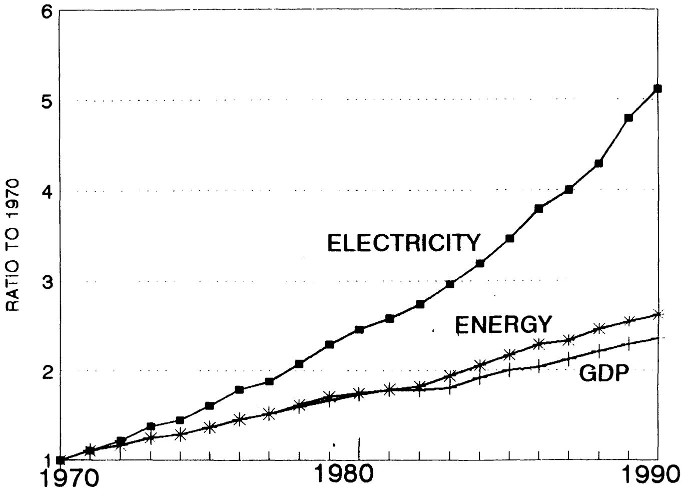

FIG. 2. Relative growth of energy and electricity use vs gross domestic product, for developing countries, 1970–1990.

of electricity in residential, commercial, communications, and industrial sectors is reflected in the close proportionality of electricity output to gross national product (GNP) for many decades, for most countries, and also on a world aggregate basis (see Figs. 1 and 2). In contrast, total energy consumption has generally declined relative to GNP, especially in industrialized countries.

In the United States, the production of electricity accounts for 37% of all energy consumed. The percentage is trending upwards towards an estimated limiting value of about 45% for the United States early in the next century, and estimated limiting values over 60% in Europe. Figure 3 shows the present annual rates of consumption of electricity in the United States by three main sectors of the economy: homes and apartments, commercial buildings, and industrial operations. The forecasts for the years 2000 and 2010 are the median-value forecasts, taking into account aggressive conservation measures, improved efficiencies for many end uses, and population growth. The remaining fuel consumption is divided between space and process heating and transportation. Europe differs from the United States in the shorter distances involved in highway transport, the high degree of electrification of rail, urban, and interurban transport, and the larger fraction of supply from nuclear energy. In the United States, 1.6×10^6 km (nearly 10^6 miles) of transmission lines link every region of the country, distributing some 2.5×10^{12} kW h of electricity each year.

As the scope and complexity of the electric power industry have grown, so has the field of power engineering. Traditionally, power engineering has consisted mainly of the classical hardware-oriented disciplines of electrical, mechanical, and civil engineering, resting on classical scientific foundations in physics, chemistry, and materials science. The need to adapt to changing requirements and growth in the variety of uses and in the demand for electricity has widened the field to include a much wider range of applied and basic science disciplines. These are as diverse as environmental and atmospheric sciences, network analyses, safety and risk assessment, artificial intelligence and neural networks, biological science, applied physics and engineering, and data, information, and simulation sciences for operations and for education and training. In the last decade there has been an increasing rate of development in analysis-based management control disciplines and in the institutional structures to track, coordinate, and regulate the huge array of disparate activities. These are required to assure both economic operations and compliance with environmental and social goals.

Electric power is unique in that, unlike other energy forms, bulk storage has been impractical except indirectly (for example, pumped storage of potential energy of water in elevated reservoirs; see STORAGE OF EN-

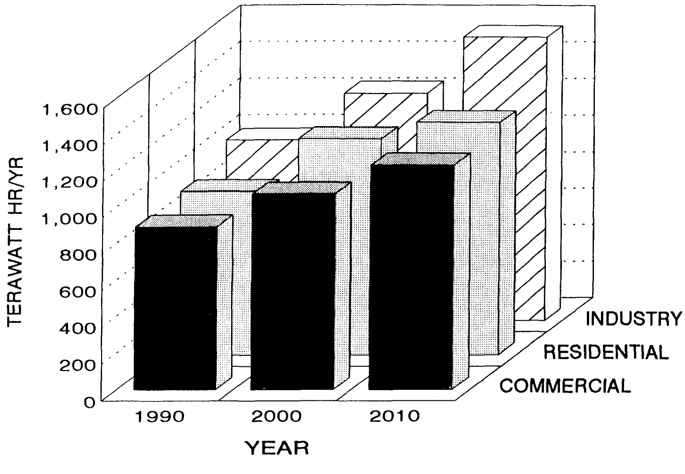

FIG. 3. Annual consumption of electricity by sector, U.S., 1990–2010.

ERGY). Unlike other commodities, it is consumed virtually at the instant of production. As world societies become increasingly electrified and computerized, reliability of supply is becoming recognized as an important factor in the productivity of an economy rather than just a convenience. Economic analyses of the consequences of occasional widespread blackout events have indicated economic losses to the regions involved ranging upwards of several hundred million dollars per event. One of the key socio-economic issues of future development of electricity supply infrastructure is the need to address the valuation of what economists call "consumer's surplus," that is, the unrecognized economic value of high reliability of supply.

Electricity production and distribution have traditionally been highly regulated in virtually all aspects, in respect to technology, location, price, allowable return on investment, and environmental and conservation practices. This approach was rationalized at a time when the production and distribution of electricity was viewed as a highly arcane technology. In return for being a designated regional monopoly, producers have a legal "obligation to serve" that has yielded a highly reliable system of supply. The nominal target of 1 hour in 10 000 unavailability of supply is usually attained.

In the last decade, the formation of independent and relatively unregulated producers has been encouraged by legislation. This has produced lively growth rates for the capacity of independent producers. This remains a small percentage of total supply and consumption, but is important in providing some degree of competition for lowest marginal costs of added capacity. Together with effective conservation measures this has significantly reduced the demand-driven pressure for building new capital-intensive base-load generating capacity. A large proportion of the independent producer capacity is fueled with natural gas that permits low capital costs. Prior to the development of highly efficient gas turbines, independent production had been at the expense of relatively high fuel costs and increased rates of depletion of a premium fossil fuel. Other independent producers, including those working on many solar- and wind-power projects, have also been encouraged by limited periods of substantial subsidies. Wind power is unfirm (that is, not reliable), and so the economics depend on how it is integrated into an overall power system that provides the needed reliability. Experience on the effective durability and lifetime of some designs is accumulating to support favorable economics in special locations with steady winds.

The rise of more competitive situations for electricity supply has changed the basis of electricity supply decisions. Where reliability of supply was once the primary factor for justifying the maintenance of a margin of base-load capacity, many other complex considerations are now involved. These include the compound goals of providing energy supply at the lowest cost while meeting a wide range of environmental and financial regulations set by federal, state, regional, and local government agencies. The trend to frequent disallowance of some capital investments from utility rate bases has often made the return on investment in new capacity risky and unattractive. Along with the surplus of capacity that developed in most regions of the United States in the 1980s and the effects of investments in energy-conserving technologies, little new generating capacity is being planned. As the average age of fossil-fuel plants rises above 30 years, replacements for obsolescence, as well as to meet modest rates of further growth, are expected to increase rapidly in the years after 1995.

1. ROLE OF ELECTRIC POWER IN THE ENERGY ECONOMY

As power lines reached through cities and across the countryside, electricity transformed the way we live and work. In the home, electric devices took over many back-breaking household tasks and created a level of comfort, convenience, and degree of emancipation for housekeepers undreamed of three or four decades earlier. In industry, electricity has helped to produce an unprecedented revolution in productivity. With the widespread transition to industrial electrification in the 1920s, natural productivity increased an average 5.2% a year, in sharp contrast to increases of less than 1% annually that marked the previous two decades. During the same period, the intensity of total energy use declined sharply in manufacturing. (Intensity here refers to the ratio of total energy input to manufacturing output.) This ratio fell by an average value of 4.3% per year during the 1920s.

Declines in energy intensity have continued with increased use of electricity by industrial societies. The greater efficiency and controllability of electrified production, even accounting for the thermodynamic and conversion losses inherent in converting primary fuels to electricity, allowed manufacturers to do more with less energy. In 1900, only 5% of industry used electric motors; today, more than 90% depends on them. Over that time, the raw energy input per unit of GNP has been roughly cut in half.

More recently, the widespread use of microprocessors has produced another revolution in manufacturing, allowing far more precise and flexible control of many kinds of industrial and commercial operations. The advent of inexpensive microprocessors has also created rapid growth and diversification in the communications, information management, and entertainment industries.

2. STRUCTURE OF THE ELECTRICITY INDUSTRY

Further development of efficiency in production and use are central to the growth and competitiveness of modern industrial economies. The pace of application of new developments contends with the environment of an extremely large U.S. base of existing capacity, namely 700 GW (7×10^{11} W). For several decades, the construction of electricity-generating and -distribution capacity represented 20%–30% of the total annual investments in national industrial infrastructure. Understanding the likely technological options, the possible rates of change, and probable impacts of new developments depends on some knowledge of the structure of the industry and its infrastructure.

Some countries have national electricity agencies. These have commonly been organized as semiautonomous industries—for example, in France, the former Soviet Union, South Korea, and many developing countries. A more common pattern is an electrical supply industry with several or many independent units, generally only partially government owned—for example, Belgium, Canada, Finland, Germany, Great Britain, Italy, Japan, Spain, Sweden, Taiwan, and the United States. Except for its size, the complexity of the U.S. industry is not atypical of the complexity evident worldwide.

Many diverse organizations form the structure of the electricity industry in the United States. This includes more than 3000 separate electric utilities, including private, municipal, and federal. Large networks of suppliers provide the equipment, parts, and a wide variety of support services and fuels to these companies. The engineering and specialized equipment and technical support services are now international in scope with companies partly or wholly owned by French, Canadian, German, and British parent organizations active in the United States, and U.S. companies active in many other countries. Electric utilities are generally categorized according to form of ownership, as listed in Table 1. Each state has some type of regulatory agency that controls the price, or rates, that utilities are permitted to charge their customers for electricity. In all cases they have jurisdiction over investor-owned utilities; in some states they also control rates charged by state- or municipal-owned systems and by cooperatives. Members of these commissions are sometimes elected, but are commonly political appointees.

The development of new technology to support the electric power industry is conducted by suppliers and individual utilities, and to a large extent is supported by the Electric Power Research Institute (EPRI) in the United States, and similar organizations in other leading countries, such as CRIEPI (Central Research Institute for the Electric Power Industry) in Japan. Founded by a consortium

Table 1. Ownership of U.S. utilities (Macdonald, 1991).

Form of ownership	Number	1989 market share (kW h basis)
Investor (shareholder) owned	208	81.1
Municipal and regional government owned	1899	7.8
Consumer owned (cooperative)	934	5.0
Federal government owned	37	6.1

of both private and public utilities in 1973, EPRI conducts research and development on a collaborative basis for electric utilities and electricity consumers. The U.S. national laboratories also conduct some relevant basic and some applied work. Except for fusion energy, their major focus is on regulatory and environmental science and technologies. The utility industry is represented by several trade associations, including the American Public Power Association for municipally owned utilities, the Edison Electric Institute for stockholder-owned companies, and the National Rural Electric Cooperative Association for cooperatives.

2.1 New Trends in the Electric Power Industry

The 1990s are a period of many rapid changes for the U.S. electric utility industry. These trends often dictate new or intensified incentives (or sometimes disincentives) for specific technological and scientific developments. Fossil-fuel costs have fluctuated dramatically during the last two decades, and long-term trends are slowly but inevitably upward as the world population increases. Increasing costs are driven by increasing demands from growth of populations, the increasing demands from the industrialization of developing countries, and the effects of increasing recognition of environmental and regulatory constraints in many countries. Many constraints now affect costs or tend to limit the recovery and production of coal, oil, natural gas, and uranium, or the construction of hydroelectric dams. Average capital costs of new or refurbished generating capacity have risen because of environmental and regulatory requirements and because of the regulatory and litigation delays now common in most large-scale industrial activities. The administrative delays commonly experienced in the construction of nuclear plants and now also for many fossil-fuel units have been a major factor driving increased capital costs. (The description of the dominant influence of such "transaction costs" led to a Nobel prize in economics in 1991.) The U.S. economy still has the benefit of a large fleet of coal (55% of output) and nuclear generating plants (21% of output) that produce electricity at low marginal costs. Many of these were built at relatively low capital costs before or early in the period of the dramatic escalation of "transaction costs" that started in the mid-1970s.

Along with society generally, the electrical industry has experienced a rapid increase in the number and complexity of environmental regulations, virtually all of which increase operating costs and involve new capital investment. For example, a single piece of legislation that sets the timing of reductions in the allowable emissions of sulfur compounds has estimated cost impacts of 50 to 200 billion dollars. These are obviously desirable enhancements for the environment, but also involve obvious major costs to the economy. In some cases, producers are squeezed financially and thus limited in the ability to implement the best available technologies.

At the same time, the general trend in the United States toward selective areas of deregulation, along with legislation aimed at fostering alternative and more efficient energy systems, has led to a number of new approaches to meeting growth in electricity demand. For example, an increasing number of industrial companies and some institutions such as hospitals and universities generate their own electricity. This is commonly called "co-generation" and combines electricity production with raising steam for process use. The electricity is used in the institution and any excess is sold to the local utility. This pattern is especially attractive to producers that have a steady year-round use for the steam production—for example, for steam injection commonly used to enhance production from depleted oil fields.

The number of independent power producers (IPPs) has also grown. Some utilities now routinely use competitive bidding from IPPs to get new blocks of power output. Others have even established nonregulated affiliates as IPPs, which compete to supply power in the service territories of other utilities—a strategy that remains controversial among utilities and regulators. The rise of IPPs and increasing competition from inside and outside the industry have given rise to other issues related to the use of distribution networks.

High-voltage transmission systems allow electric energy to be moved ("wheeled") for hundreds of miles. This is used routinely to compensate for shortfalls in available capacity in one region by temporary borrowing from another region. The borrowing is preferably temporary because there are significant losses of energy as the distance of transmission is increased. The ability to exchange power with nearby regions reduces by as much as 15% the total amount of generating capacity required to maintain a desired level of reliability of supply. These exchanges are covered by both short-term and long-term sale-purchase agreements between utilities. At present, most transmission lines are owned and controlled by the utilities that built them, many of which have formed power pools to dispatch power in a coordinated way from generating stations in the pool. Allowable contributions to the pool involve an agreed-upon economic rank order. The overall economics of the pool is optimized by sets of prioritization algorithms that take account of the generating units with the lowest marginal costs of production, tempered by the increasing costs and electrical losses as average transmission distances increase. The unifying principles shared by utilities in a power pool are the "obligation to serve" and the need to demonstrate lowest achievable costs.

Legislation has been proposed to treat transmission lines as common carriers. This has aroused concerns that if use by IPPs becomes large, system reliability may suffer, since IPPs are generally exempt from the obligation to serve and would experience appreciably higher costs if this obligation were to be imposed.

Most utilities now actively promote and subsidize measures to reduce electricity consumption—for example, by improved thermal insulation, more efficient lighting systems, better automatic controls to eliminate unneeded uses, and voluntary demand-side controls that reduce peak loads. As has been more common in other industries, mergers are now being more often considered where these can provide improved economies of scale, avoidance of some kinds of duplication, and better matches of assets and needs.

3. SUPPLYING ELECTRICITY

The classical electric generator consists of a shaft on which an array of coils is mounted, rotating within a magnetic field. The mechanical energy that provides the motive force to spin the shaft is extracted from a working fluid, typically steam or air, at high temperature and pressure, by allowing it to expand through a series of turbine stages.

In hydroelectric generating systems, the potential energy of water stored behind dams is converted to mechanical energy in hydraulic turbines. Most electricity worldwide is produced in thermal power stations using steam as the working fluid. Fuel is burned in boilers to produce heat, which is used to convert water at high pressure into steam at high pressure. The expansion of high-pressure steam through modern multistage turbines extracts mechanical energy at near-theoretical efficiency. The steam condenses in a water-cooled condenser to be recycled as feedwater for the boiler.

This cycle, known as the Rankine cycle, is most often fueled by coal, oil, or gas. Almost any fuel—wood, peat, or refuse, for example—can be used in an appropriately designed boiler. Efficiency depends primarily on the temperature range in which the steam is expanded and condensed. It can be improved if heat is recovered from exhaust streams and that energy is ultimately recovered in the steam turbine. The overall thermal efficiency of this cycle of converting the chemical energy in fuel to electricity has increased from average values near 20% as late as 1950 to over 33% currently, and leading units reach well over 40%.

Gas turbines (see TURBINES, GAS), on the other hand, use the energy contained in hot pressurized combustion gases that result from burning pressurized liquid or gaseous fuels with air at high pressure. Expansion of the gas through the gas turbine section provides the energy to turn the turbine and compressor shaft. The excess energy over the amount required by the compressor is available to turn a generator shaft to produce electricity. This cycle, known as a Brayton cycle, has an efficiency of 25% to 35%, depending on the temperature of the combustion gas entering the expansion section. By utilizing a heat-recovery steam generator in a steam turbine cycle, additional energy can be recovered from the exhaust gas from the gas turbine, which normally exits at 426.7 °C (800 °F) to 565.5 °C (1050 °F). The hot exit gas is used to heat water and produce steam in a boiler. This feeds a second cycle ("a bottoming cycle" since it recovers some of the lower-temperature energy). Such combined-cycle systems can achieve efficiencies of up to 50%. While there is added complexity and capital cost for a combined cycle, the added efficiency can offset the added cost of liquid fuels relative to coal or nuclear fuel.

Gas turbines with simple Brayton cycles have been used for several decades for "peaking units." These are low–capital-cost and high–fuel-cost units. They are economic if operated only at times of peak loads, since they serve to reduce the amount of base-load capacity that is generally much higher in capital costs. As high-temperature materials have been improved in the last decade, the practical size of gas turbine units has increased from the 10- to 15-MW range to 150–200 MW and larger. The efficiencies of simple cycle unit have also increased. Gas turbines combined with steam cycles are now increasingly widely used to generate electricity.

An important new development is to avoid the costs of premium liquid fuels by adapting gas turbines to use coal as fuel. This is done in conjunction with coal gasification plants, which produce clean gas from coal. The gas produced is lower in energy content per unit volume than natural hydrocarbon fuels ("low-BTU gas"), requiring special design considerations for turbines and boilers. However, the fluid-bed gasification process permits the use of low grades of sulfur-containing coals, while at the same time providing a major environmental advantage by eliminating the emission of sulfur compounds from the plant. Sophisticated systems known as integrated coal gasification combined-cycle power plants have been developed that produce electricity from coal at thermal efficiencies of 35% to 45%.

Electrochemical cycles can also be utilized to make electricity directly from fuel gases (see ELECTROCHEMISTRY). Methane or hydrogen from natural hydrocarbons, or from coal gasification or coal, is the reactant supplied to the anode of a fuel cell. Air or oxygen is supplied to the cathode. Electrons freed at the anode travel through an external circuit, generating direct-current electricity, and then combine with the oxidant at the cathode. Unlike steam and gas turbine cycles, the fuel-cell system does not require any mechanical moving parts to produce electricity, and there are virtually no environmental emissions to be controlled. Initial applications include urban substations where the absence of noise, vibration, or environmental effects can justify the higher costs. Such electrochemical cycles can achieve thermal efficiencies of 50% to 60% including the conversion of dc to ac, and there is potential for further improvements.

The following sections describe five typical configurations for modern electricity-generating stations.

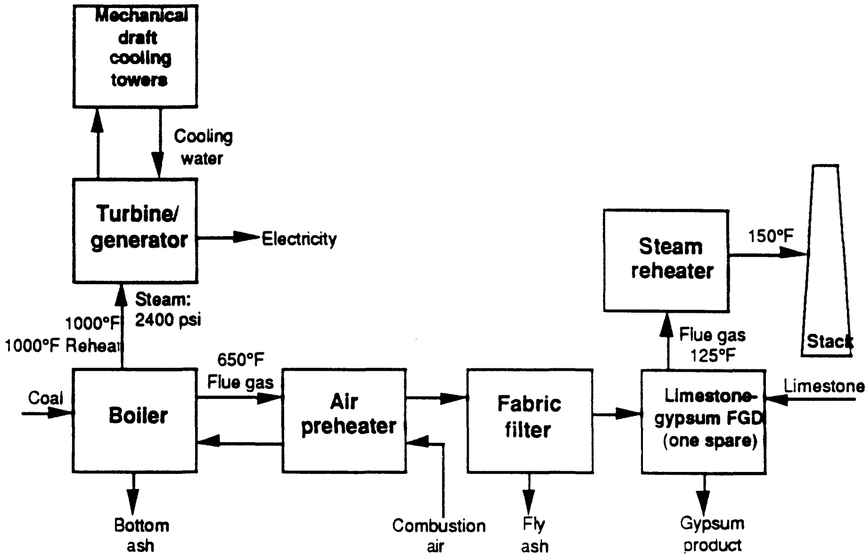

FIG. 4. Conventional pulverized-coal power plant with wet limestone flue-gas desulfurization (FGD). Primary steam outlet conditions are 1000 °F, 2400 psi; reheat to 1000 °F (538 °C).

3.1 The Pulverized-Coal–Fired Plant

Figure 4 depicts the typical pulverized-coal plant (see COAL). It can be fueled by high-sulfur bituminous coal with steam conditions of 16.5 MPa (2400 psig) and 538 °C (1000 °F) and reheat to 538 °C (1000 °F). In this cycle, steam leaving from the high-pressure section of the steam turbine is reheated in the boiler before entering the intermediate-pressure section of the turbine. A limestone scrubber removes 90% to 95% of the sulfur. This process is referred to as flue-gas desulfurization (FGD). Combustion modifications that reduce peak local temperatures are used to limit the formation of nitrogen oxides. The heat rate for such a plant is between 2343 and 2520 kcal per kW h (9300 to 10 000 Btu/kW h). The next generation of pulverized-coal plants is scheduled to use supercritical double-reheat steam cycles at 31.0 MPa and 593.3 °C (4500 psig and 1100 °F) and advanced stack-gas cleanup processes. Such systems will reach heat rates of approximately 2268 kcal per kW h (9000 Btu/kW h) or 37.9% thermal efficiency.

3.2 The Atmospheric Fluidized-Bed Combustion Plant

Atmospheric fluidized-bed combustion (AFBC) (see COAL), shown in Fig. 5, generates power using a steam cycle with the same steam conditions as the pulverized-coal plant. Their heat rates, 2369–2470 kcal/kW h (9400 to 9800 Btu/kW h), are similar. The combustion system is a bed of coal (either bubbling or circulating) fluidized by the incoming stream of air that supports combustion. The beds operate at fairly uniform temperatures between 1500 and 1700 °F (815–927 °C). Steam is generated and superheated in conventional water-wall and convection-pass tube banks, although in-bed evaporation tubes are used in some configurations. Economizing and air preheating are conventional. Limestone sorbing is added to the bed of coal to capture 90% to 95% of the potential SO_2 emissions. Because FBC units require almost twice as much sorbent as that required for the scrubbers for pulverized-coal plants, the amount of solid waste is correspondingly larger. NO_x emissions, however, are substantially lower because of the fluidized bed's relatively low and uniform combustion temperature, compared with peak temperatures over 1038 °C (1900 °F) for combustion in conventional boilers and superheaters.

3.3 The Pressurized Fluidized-Bed Combustion Plant

Pressurized fluidized-bed combustion (PFBC) (see COAL), shown in Fig. 6, is similar in many respects to AFBC, except that the boiler operates at an elevated pressure of 0.62–1.72 MPa (90–250 psig). The hot pressurized gas exiting from the boiler is expanded in

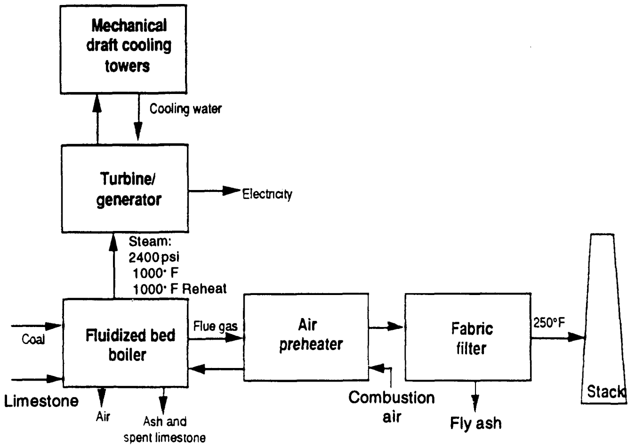

FIG. 5. Fluidized-bed boiler with *in situ* desulfurization. Primary steam outlet conditions are 1000 °F, 2400 psi; reheat to 1000 °F (538 °C).

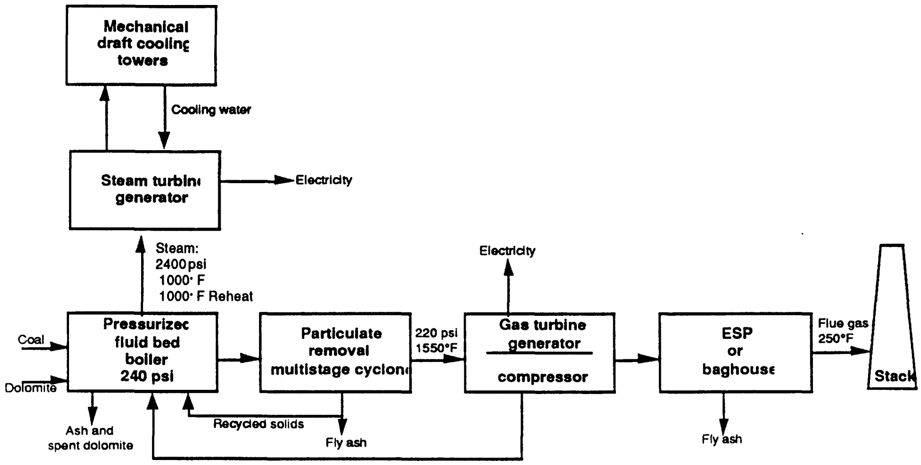
Mechanical draft cooling towers
Cooling water
Steam turbine generator
Electricity
Steam:
2400 psi
1000° F
1000° F Reheat
Electricity
Coal
Dolomite
Pressurized fluid bed boiler 240 psi
Particulate removal multistage cyclone
220 psi
1550°F
Gas turbine generator
compressor
ESP or baghouse
Flue gas 250°F
Stack
Ash and spent dolomite
Recycled solids
Fly ash
Fly ash

FIG. 6. Pressurized fluid-bed combustion combined-cycle power plant.

a gas turbine. This turbine provides all the pressurized fluidizing air for the boiler and also generates up to 20% of the gross plant power output. Since PFBC extracts part of its energy at the higher temperatures acceptable in modern gas turbines, as well as the lower temperatures practical for steam cycles, heat rates in the range of 2192–2520 kcal/kW h (8700–10 000 Btu/kW h) are calculated from the known separate efficiencies of the two cycles involved. For the scrubbers for pulverized-coal plants, the amount of solid waste is correspondingly larger. In this case, also, the emissions of NO_x are substantially lower.

3.4 The Integrated Coal Gasification Combined-Cycle Plant

Figure 7 illustrates the integrated coal gasification combined cycle (IGCC), which uses coal reacted with oxygen in a gasifier to produce a clean combustible synthesis gas

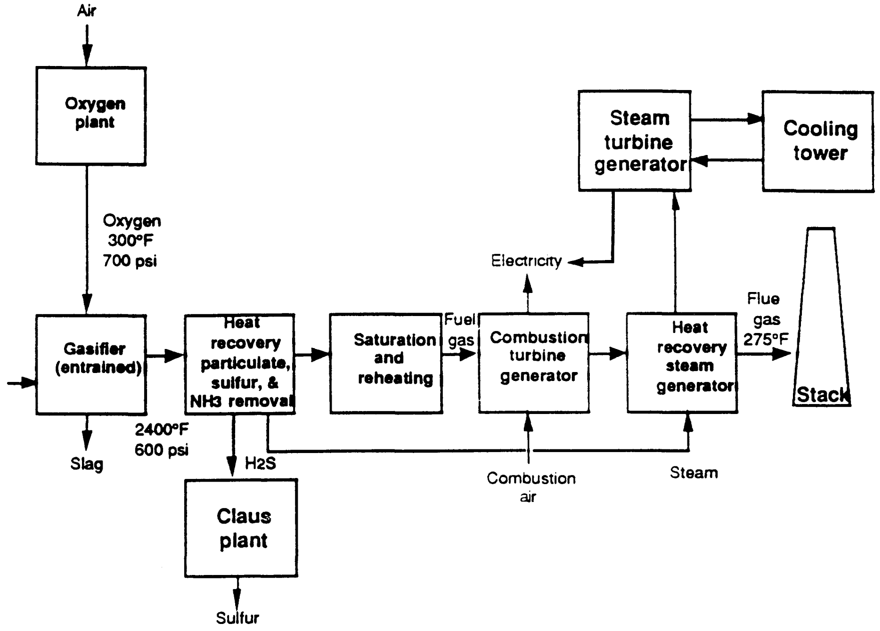
Air
Oxygen plant
Oxygen
300°F
700 psi
Gasifier (entrained)
2400°F
600 psi
Slag
Heat recovery particulate, sulfur, & NH3 removal
H2S
Claus plant
Sulfur
Saturation and reheating
Fuel gas
Combustion turbine generator
Combustion air
Electricity
Steam turbine generator
Cooling tower
Heat recovery steam generator
Steam
Flue gas 275°F
Stack

FIG. 7. Coal gasifier with combined gas-turbine and steam-turbine cycle for increased thermal efficiency.

consisting mostly of hydrogen and carbon monoxide (see COAL). The gas first passes to a gas turbine, producing part of the output of the plant. If gas leaving the gasifier is at a high temperature, as is the case in some gasification processes, the remaining available heat is recovered by heat transfer in a boiler to generate steam. The steam is used in a steam turbine to produce the remainder of the plant output. This part of the cycle is called a "bottoming" cycle since it recovers the lower-temperature portion of the heat energy available in the synthesis gas. This part of the cycle operates at nominal values of 16.5 MPa and 537.8 °C (2400 psig and 1000 °F) and with reheat to 537.8 °C (1000 °F). Particulates are completely removed from the gas in a water wash, and the gas is further cooled to allow removal of sulfur and nitrogen. Conventional chemical- or refining-industry technology is used to remove virtually all sulfur-containing compounds and to recover elemental sulfur. Ammonia formed in the gasifier from some of the nitrogen present in coal (which could contribute to NO_x emissions) is also removed. The gas is then reheated and saturated with water vapor utilizing low-level heat that was recovered as the gas was cooled. The water vapor further limits NO_x emissions in the subsequent combustion step. Finally, the gas is burned in the combustion turbine of a conventional combined cycle. The combustion turbine produces approximately 60% of the plant's power output.

If a modern gas-turbine 1260 °C (2300 °F) firing temperature is used, IGCC plant thermal efficiencies are from 37.5% to 39.7%, depending on local ambient temperatures of air and of the heat sinks available. Advances in gasification, gas-turbine technology, and plant configuration—including tight integration of the oxygen plant with the gas-turbine compressor and fuel system, 1371 °C (2500 °F) firing temperature turbines, and hot or warm gas cleanup—are expected to lead to IGCC thermal efficiencies up to 42.6%.

3.5 The Integrated Gasification Molten-Carbonate Fuel Cell

In Fig. 8, high-pressure clean, preheated, moist synthesis gas ("syngas"), much as it might come from an IGCC plant's saturator, is first reduced to atmospheric pressure through an expansion turbine that produces about 10% of the plant's gross power output. The fuel gas is then fed to the fuel cell anode. The electrolyte is a molten mixture of alkali-metal carbonates that provide a convenient range of temperatures and low vapor pressures. About 35% of the heating value of the gas is converted directly into electricity, by the electrochemical processes, producing 55% of the total power output. Because fuel-cell power output is in the form of direct current, inverters are used to produce alternating current. The remaining unconverted fuel in the exhaust gas from the anode is catalytically

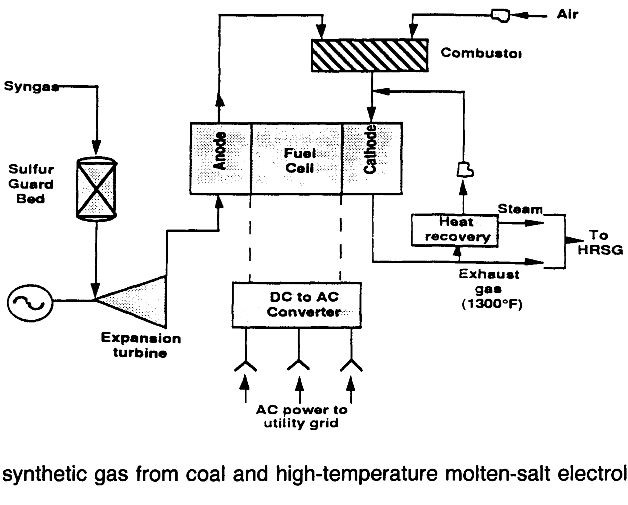

FIG. 8. Fuel cell using synthetic gas from coal and high-temperature molten-salt electrolyte.

burned, mixed with the incoming combustion air, and returned to the fuel cell's cathode. The anode exhaust plus the cathode recycle provides carbon dioxide to the cathode. The cathode exhaust gas goes to a heat recovery steam generator (HRSG), where heat is recovered for a conventional steam bottoming cycle, providing 35% of the total power output. The fuel cell operates at about 1200 °F and the catalytic combustor only slightly higher. At these low temperatures, essentially no nitrogen oxides are produced. Sulfur compounds must be completely removed from the fuel gas since they poison the electrodes of the fuel cell. The emissions of SO_x and NO_x are negligible.

4. NUCLEAR POWER

Fission reactions, which involve the breakup of the nucleus of high-mass atoms, yield energy per atom reacting more than a millionfold greater than that for chemical reactions involving the burning of fuel. The oxidation of an atom of carbon releases only 4 eV, for instance; the fissioning of an atom of uranium yields about 200×10^6 eV (i.e., 200 MeV). U.S. deposits of uranium, and the availability of unused uranium enrichment facilities from the Manhattan Project, led the federal government to provide incentives and virtual directives to electric utilities and to major suppliers to develop and adopt nuclear power for baseload generation of electricity. First-generation nuclear units, a few of them still operating, were built in three to four years from time of commitment to time of practical output. Capital costs were less than 50% higher than equivalent coal units, and overall economics were favorable for many projects because of the lower costs of nuclear fuel. However, many of the units completed after the mid-1970s experienced extensive regulatory delays, up to 17 years in one case, and the period costs (the cost of servicing the bonds that finance the construction) served to double or triple the total capital costs. Regulatory requirements have continued to escalate and are still subject to rapid change. This situation, together with the excess capacity that was available in most regions prior to 1991, makes investment in added nuclear units in the United States unlikely until and unless these circumstances change substantially. Meanwhile, France, Belgium, and Sweden get over half of their electricity from nuclear units, and Japan continues to build essentially identical types of plants economically in three to four years. Since 1991 approximately 21% percent of the U.S. generation has been supplied by over 100 nuclear power units. A similar percentage exists in Great Britain.

The only fissionable material found in nature is uranium-235. Nuclear fuel is typically enriched to about 5% in uranium-235. Part of the "fertile material" uranium-238 is converted into fissile plutonium-239 by the absorption of neutron and two subsequent β decays. The fission of the *in situ*-produced plutonium adds about one-third to the energy output available per unit of fuel. In several countries, notably France, Japan, and to some extent Belgium and Germany, the plutonium in spent fuel is recovered by chemical processes, and it is recycled by adding it to new fuel, reducing the amount of uranium-235 required. The dominant fuel cycle is the uranium-plutonium cycle in light-water reactors that are the predominant source of nuclear electric power in the world today. India, with abundant reserves of thorium, carries on development of a thorium-232 and uranium-233 fuel cycle.

Other reactor types have been developed and have reached practical commercial use. These include heavy-water natural-uranium reactors in Canada, graphite-moderated light-water–cooled reactors in the former Soviet Union, and gas-cooled, graphite-moderated reactors in Great Britain and for a time in France. None of these has had significant market penetration outside the country of origin, and capital costs are generally lower for light-water reactors in stable regulatory environments. Prototypes of other reactor types have been developed and built, notably sodium-cooled reactors without moderators to slow down the fast neutrons resulting from fission, that operate in a fast-neutron spectrum. These have been built in small numbers in France, Russia, Japan, India, and the United States. The attraction for development of the fast-neutron spectrum reactors is that highly extended reactivity lifetimes are feasible (ten or more years between refuelings), or else they can be configured to make excess fuel, as so-called "breeder" reactors. The early perceptions of limited supplies of natural uranium have been superseded by a picture of a glut of uranium supply and very large known reserves in the United States, Australia, Canada, and Russia.

The possible incentive for the breeder reactor as a producer of extra fuel is now generally viewed as arriving only in the third or fourth decade of the 21st century, since light-water reactors can be fueled with available uranium for the next 30 to 40 years. However, the breeder-reactor development continues to have the attraction of providing security of fuel supply for countries such as Japan, France, and India, which have only minor reserves of uranium. The use of fast-neutron "burner" reactors for energy production and for disposal of decommissioned nuclear weapons materials is also under discussion.

The two basic types of present commercial nuclear power reactors differ mainly in whether the coolant water is allowed to boil. These are the boiling-water reactors and pressurized-water reactors, illustrated in Figs. 9 and 10. Shielding is provided to limit external radiation levels, and pools of water are used for fuel storage and for transferring fuel in and out of the reactor core. (See NUCLEAR ENERGY, FISSION.)

Containment buildings surround the reactor system and most auxiliary systems except for the turbine-generator and the feedwater systems. The structures are designed and tested to hold safely the total thermal and pressure energy available if a complete rapid depressurization of the coolant system should occur by rupture of the primary piping. The containment also serves to provide a secure enclosure to minimize potential escape of radioactive material to the surroundings. The effectiveness of containment was evident in the accident at Three Mile Island, where the core had a complete meltdown and release of roughly 20 MCi of radioactivity to the containment, but only minor release of rare gases to the atmosphere. In contrast, the reactor at Chernobyl, without a full containment structure, released massive amounts of radioactivity, seriously damaging the environment for more than a 20-km radius, and with detectable radioactivity deposition over much of the world.

4.1 The Boiling-Water Reactor

The boiling-water reactor (BWR) design depicted in Fig. 9 can produce about 3580-MW thermal and 1220-MW net electrical power. Each fuel assembly typically contains 63 fuel rods, and there are 732 fuel assemblies. The neutron reactivity is controlled and shut down by means of control blades arranged in cruciform cross section. The blades move up from the bottom of the reactor in the gaps between four fuel assemblies. The water coolant is circulated up through the fuel assemblies by 20 jet pumps at about 7 MPa (70 atm), and net boiling occurs within the core. Fine control of the nuclear reaction is provided by varying the speed of the pumps. Because of the strong negative void coefficient of the reactor power, the power level (for a given setting of the control blades) stabilizes at constant volume fraction of steam in the core. The negative void coefficient arises because the reactor lattice is designed to be slightly undermoderated. That means that removing water (producing steam void) lowers the reactivity (the neutron multiplication rate) of the reactor. Conversely, adding water tends to increase reactivity, which tends to produce more heat and thus more steam void. This situation permits convenient control of reactor power by varying pump speed and incoming water flow. When pump flow is increased, the reactivity increases until a new higher power level is reached. If pump flow decreases, the power level decreases until the same volume fraction of steam is reached. The steam is separated from the steam–water froth by centrifugal and vane separators and then is fed to the turbine through four steam lines. About one-third of the energy released by fission is converted into electrical energy by the turbine generator, and the remaining low-grade heat is rejected in the condenser, for a thermal efficiency of about 33%. This part of the cycle is the same in both fossil-fuel and nuclear power plants requiring the dissipation of low-grade heat to the environment. Some limited use of the energy rejected by the condenser is possible but is not as yet economic. Because the steam produced is at lower temperatures and pressures, nuclear plant cycles are several percent lower in thermal efficiency than those in the most modern fossil-fuel plants, and require slightly greater

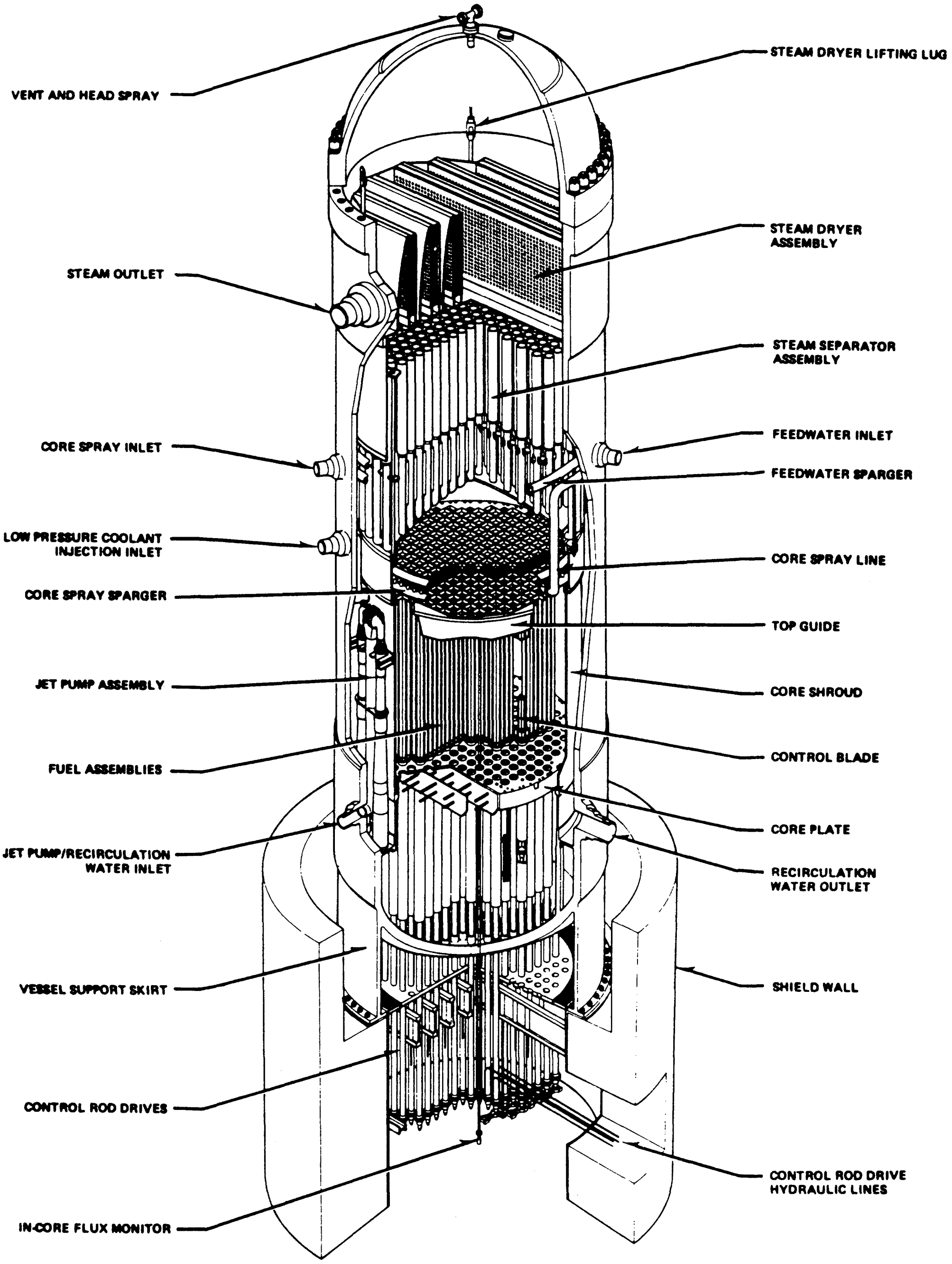
VENT AND HEAD SPRAY
STEAM DRYER LIFTING LUG
STEAM DRYER ASSEMBLY
STEAM OUTLET
STEAM SEPARATOR ASSEMBLY
CORE SPRAY INLET
FEEDWATER INLET
FEEDWATER SPARGER
LOW PRESSURE COOLANT INJECTION INLET
CORE SPRAY LINE
CORE SPRAY SPARGER
TOP GUIDE
JET PUMP ASSEMBLY
CORE SHROUD
FUEL ASSEMBLIES
CONTROL BLADE
CORE PLATE
JET PUMP/RECIRCULATION WATER INLET
RECIRCULATION WATER OUTLET
VESSEL SUPPORT SKIRT
SHIELD WALL
CONTROL ROD DRIVES
CONTROL ROD DRIVE HYDRAULIC LINES
IN-CORE FLUX MONITOR

FIG. 9. Cutaway view of 1300-MW boiling-water reactor core and pressure vessel.

heat rejection to the environment per unit of electricity produced.

The BWR containment structure uses a pool of water, called a suppression pool, to condense the steam produced by the depressurization of the primary coolant system. Other engineered safety features include the emergency core-cooling system, the containment spray system, and the high-efficiency particulate and activated carbon filters for removing

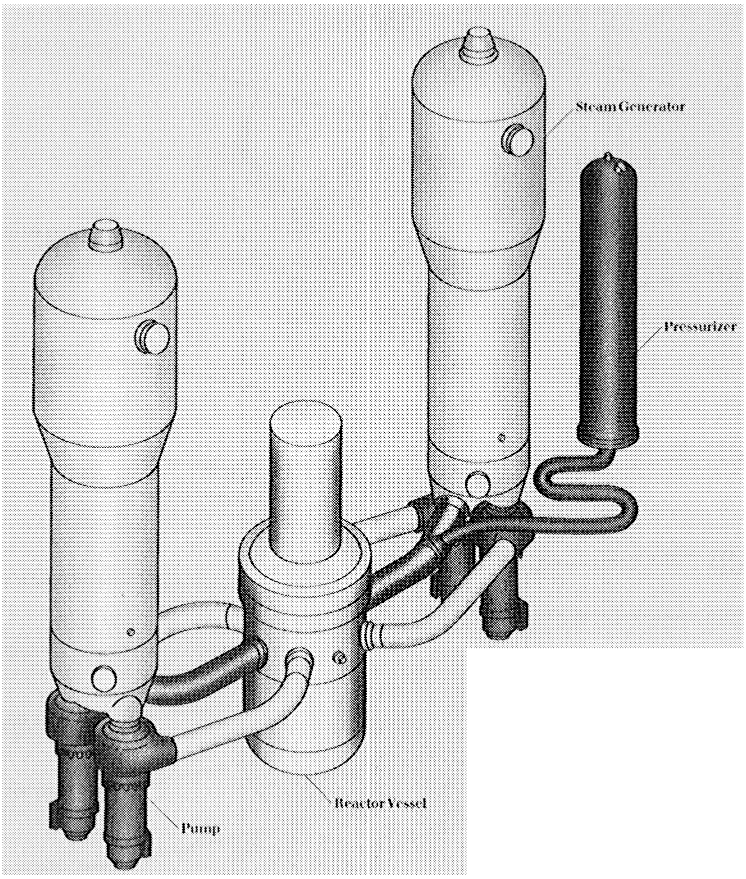

FIG. 10. Pressurized-water reactor system layout—advanced design.

radioactivity from the atmosphere of the containment buildings and from the turbine air ejector stream.

4.2 The Pressurized-Water Reactor

Unlike the BWR, in which steam from the reactor is fed directly to the turbine, the pressurized-water reactor (PWR) shown in Fig. 10 uses a second closed system. The water coolant in the primary system is pumped through the reactor vessel, transporting the heat to a steam generator, and is recirculated in a closed primary system. A separate secondary water system is used on the shell side of the steam generator to produce steam, which is fed to the turbine, condensed, and recycled back to the steam generator. A pressurizer is used in the primary system to maintain about 15 MPa (150 atm) pressure to suppress boiling in the primary coolant. Two circulating primary loops are shown here, but up to five loops have been used. Different fuel assembly configurations are used by different manufacturers. One type of assembly, for instance, contains 264 fuel rods, with 193 fuel assemblies used for the 3411-MW (thermal), four-loop plant. In this arrangement, the control rods are grouped in clusters of 24 rods, with 61 clusters. In PWRs, the control rods enter from the top of the core. Reactor operations are carried out by using both the control rods and a system to increase or decrease the boric acid content of the primary coolant. The latter is used to compensate for long-term changes in the reactivity of the core, while the former allows fine tuning of the power level and rapid shutdown of the nuclear reaction when needed.

4.3 The Future of Nuclear Power

With over 25 years of favorable operating experience worldwide, as well as extensive development and testing programs, light-water reactors such as those described above are likely to remain the dominant nuclear power technology for the next several decades. Well-tested and mature designs are offered and sold by suppliers in France, Germany, Japan, and the United States, and the capability for supplying most of the systems exists in many more countries—for example, England, Spain, Belgium, Sweden, and Russia. Currently, the U.S. Department of Energy and the utility industry are conducting an advanced light-water reactor program to develop the designs, safety standards, and regulatory requirements for a possible next generation of nuclear power plants. Substantially simpler plant design will help increase lifetime reliability, reduce capital costs, shorten construction times, lower operating and maintenance costs, and increase safety margins by a proposed factor of 10, as measured by the disciplines of probabilistic risk analysis (PRA).

Next-generation designs have already been committed to construction in France (1450-MW rating), and a joint U.S.–Japan design of an advanced BWR design (1350-MW rating) is under construction in Japan.

Breeder-reactor development continues at a modest pace in France, Japan, and India, and on a small scale in the United States and Great Britain. At one time the world uranium reserves, even used in breeder reactors, were considered to be merely about equal in energy content to the world reserves of oil and gas. Subsequently, world uranium reserves of over 10×10^6 tons are reasonably assured at moderate prices (for breeder-reactor use). This represents an energy resource larger than the total reserves of all fossil fuels including coal, lignite, tar sands, etc. However, there is no pressing economic need to deploy a fleet of breeder reactors for at least 30 years, since present low-conversion reactors do not seriously deplete low-cost uranium reserves until then. The present approach in countries carrying on breeder-reactor development may be described as maintaining the breeder reactor as a contingency against unexpected rises in the market price of uranium, and as a contingency for when fossil fuels plus alternative sources such as renewable resources and fusion are not able to fully satisfy world needs after the first quarter of the 21st century.

5. TRANSMISSION AND DISTRIBUTION

The power delivery system of U.S. utilities comprises the largest interconnection of electrical transmission, distribution, and control equipment in the world—a vast network made up of transformers, circuit breakers, fuses, overhead and underground lines, computers, relays, and data recorders. Electric power is delivered on demand, reliably and safely, through the voluntary cooperation of over 3000 utilities, which own and operate thousands of power plants, over 4.8×10^6 km of overhead transmission and delivery lines, over 300 000 km of underground cable, tens of thousands of substations, and a myriad of associated hardware, computers, and software components.

Fundamental changes in both the delivery and use of power are now under way. There is growing demand for more and better-quality power, with fewer outages, even momentary ones. This demand is expected to increase as consumers employ an increasingly wide range of new technologies. Utilities, too, are changing. Engineers are employing an unparalleled range of new technologies to meet changing demand—including the application of a variety of basic innovations: versatile, high-power semiconductors; higher-speed computer processors and systems capable of real-time control of complex systems; high-energy surge arrestors; stronger and more durable materials; advanced automation techniques; and simplified but more effective communications systems.

5.1 Flexible ac Transmission Systems

With few new power plants or lines being built, utilities have to route more power over longer distances on available lines. Because these lines were never designed for such heavy loadings, the on-line control of power flow, voltage, and stability becomes more of a challenge. Operators must rely largely for control on mechanical devices that do not respond quickly to changing conditions.

A comprehensive control system governed by high-power thyristors and advanced microprocessors—a flexible ac transmission system, or FACTS—is being developed to enable utilities to direct power with unprecedented precision. Unlike mechanical devices, thyristor-based controllers can easily manipulate power flow and voltage while maintaining stability. Thyristors are essentially high-speed switches arranged in a variety of configurations to switch and adjust electrical components on high-voltage, high-current power lines. Single solid-state thyristors are capable of handling power-class voltages up to 8500 V and currents up to 3500 A. They can be connected in series and parallel combinations to achieve very high power levels.

These controllers—which include phase-angle regulators, modular series capacitors, static VAR (volt-amperes reactive) compensators, damping devices, dynamic load brakes, circuit breakers, load tap changers, surge suppressors, and fault-current limiters—manipulate power flow by adjusting voltage, optimizing power angle (the phase difference between voltages at opposite ends of a line), changing impedance, or some combination of the three. To compensate for electrical phenomena that inhibit power flow in ac networks, they can direct power flow by manipulating resistance, capacitance, and inductance. They can also minimize potentially damaging power oscillations by damping unwanted frequencies. Because thyristor-based devices guided by microprocessors respond so speedily to changing system conditions, they will enable operators to use transmission lines consistently to more nearly their full thermal-electrical capacity while maintaining required stability margins.

5.2 Energy Efficiency in Transmission and Delivery

In some utility transmission and delivery systems that involve long distances, energy losses can amount to as much as 10% of the energy produced. Even a 1% energy-loss reduction could save as much as $1.5 billion a year nationwide. Currently, approximately 1 million new distribution transformers are installed annually on U.S. utility systems. The use of amorphous steel alloys that can reduce the core loss of distribution transformers by approximately 60% to 70% could dramatically decrease transformer load losses. Replacement by amorphous core units could result in an energy savings of nearly 6×10^8 kW h per year.

Meanwhile, other developments promise to improve energy efficiency further in transmission and delivery. A low-loss winding developed for larger power transformers, for instance, can significantly lower not only losses, but size and weight as well—reducing energy operating losses up to 30%.

6. POWER APPLICATIONS OF SUPERCONDUCTIVITY

The discovery of high-temperature superconductivity (HTSC) (*q.v.*) in ceramic materials in 1986 has led to an international research program to understand the physics, chemistry, and physical processing methods to evaluate their characteristics and possible methods of manufacture. The materials carry electric currents with no resistance losses at liquid-nitrogen temperatures (77 K). Considerable progress has been made, including the production of tapes, wires, and films that have useful engineering characteristics. Both continuing basic research and the development of practical applications of high-temperature superconductors remain one of the most important scientific challenges of power engineering.

Because superconductors can transmit direct current without resistive loss and alternating current with extremely low loss, superconducting devices could generate, transmit, and store electricity far more efficiently than conventional devices. Moreover, because superconductors can carry large currents per unit cross-sectional area, the iron in conventional electromagnets can be removed and the devices operated at higher field strengths, yielding more compact motor and generator designs and higher efficiencies. The key attributes—and the limiting properties so far—have been the current-carrying capacity of HTSC materials and the limited ability to withstand magnetic fields without collapse to the nonsuperconducting state. The limitations of brittle ceramic materials have been largely overcome by the use of metal-ceramic composites in the form of wires or ribbons.

Table 2. Cost and performance comparison for a 300-MW generator (millions of 1989 dollars).

Parameter	Conventional generator	Using LTSCs	Using HTSCs
Efficiency (%)	98.6	99.6	99.7
Capital cost	880	1240	740
Annualized capitalized losses	1400	430	280
Annual capital costs	165	230	140
Total annual costs	1565	660	420

Several superconducting devices—including generators, motors, transmission lines, magnetic storage transformers, fault-current limiters, and power electronics—have been proposed, and some working models have been built for use in the electric utility sector. Because HTSCs are still in the early stages of development, their ultimate cost and performance is highly sensitive to the degree to which the key attributes mentioned above can be improved. The incentive remains strong for improved materials since such superconductors may yield a variety of more efficient, more compact, lighter, and less expensive power applications devices.

6.1 Generators and Motors

Utility generators range from less than 100 to 1350 MW and already have a mechanical-to-electric energy conversion efficiency of between 97% and 99%. Even so, small efficiency improvements could yield significant cost savings. The two largest losses in ac generators, in the rotor and the armature, represent approximately 30% each. During operation, the rotor's dc field coil rotates with an essentially constant field, surrounded by an ac armature. A recent study of cost and performance for 300-MW generators suggests that HTSCs could improve the efficiency of conventional generators by 1% and reduce total annual capital costs to less than one-third, as Table 2 illustrates.

Table 3. Comparison of costs and efficiency for 10 000-MVA ac transmission lines (millions of 1989 dollars).

Parameter	Aerial/ Underground	HPOF	Underground LTSC	HTSC
Efficiency (%)	98.3	96.4	97.3	99.3
Capital cost	852	2546	1821	1759
Capitalized losses	419	986	713	190
Total costs	1271	3532	2534	1949

Although generators employing low-temperature superconductors (LTSCs) are estimated to have a capital cost 40% higher than the conventional generators, the HTSC, because of its reduced insulation and refrigeration requirements, may actually have a lower annual capital cost. As Table 2 illustrates, even a small gain in efficiency can yield a substantial reduction in capitalized losses. The lower effective capital cost and higher efficiency of the HTSC could eventually lower the cost-effective size of generators and motors from 300 MVA to perhaps several MVA, greatly enhancing the HTSC's market potential.

6.2 Transmission Lines

Superconducting power transmission lines may be able to transmit electricity more efficiently, at lower voltages, and at lower cost than conventional, underground, high-pressure, oil-filled pipe transmission lines (HPOF). Both ac and dc versions have been proposed. In each case, a set of coaxial cables would be contained in an insulated enclosure surrounded by a vacuum chamber. A 115-m superconducting ac transmission line rated at 100 MVA at liquid-helium temperatures was designed, constructed, and tested at Brookhaven National Laboratory during the 1970s.

As Table 3 demonstrates, the aerial and underground system has by far the lowest total cost (including capitalized losses), but the HTSC system is the most efficient. Because the HPOF system is the most expensive and least efficient, it is used only when aerial systems cannot be used, such as in heavily congested urban areas or waterway environments. Some laboratories believe that further gains in reducing current-related losses in niobium-tin tapes could reduce the cost of the LTSC system, offsetting much of the efficiency and cost advantage of HTSCs.

6.3 Superconducting Magnetic Storage

Superconducting magnetic energy storage (SMES) efficiently stores energy in a magnetic field associated with a circulating current. An inverter or convertor is used to transform ac power to direct current, which is then used to charge a large solenoid magnet. Upon discharge, power is withdrawn from the magnet and converted to ac power. SMES has a round-trip (charge to discharge cycle) efficiency between 90% and 93%—the highest of all proposed utility-side storage devices—and its ability to switch from charge to discharge mode in less than a second provides valuable system-stabilization characteristics. Large-scale SMES devices (500–10 000 MW h) would provide the most economical form of storage.

An 8.3-kW h (30-MJ), 10-MW SMES was designed, constructed, and installed for operation in the Tacoma substation in the Bonneville Power Administration utility system, where it was operated for about one year, during 1983 and 1984, at frequencies from 0.1 to 1 Hz, for more than 1×10^6 cycles. In 1982, the Japanese began a three-year design study for a 5-GW h device. In 1986, work was begun on a 5-GJ (1.4-MW h) SMES test facility, which is expected to be operational by 1992. Japanese interest has concentrated on small-scale SMES devices for load leveling of commercial building loads, although there is a long-range plan to start building a 500-MW h utility-scale SMES in 2004.

HTSCs could also be used to replace conventional leads in an LTSC SMES that has current densities of hundreds of A/cm^2 and operates in regions of low magnetic field and temperatures below 77 K, where HTSCs perform best. An HTSC current lead, designed and fabricated by Argonne National Laboratory, is being tested at Westinghouse. Such current leads could also be useful for other LTSC devices, including generators, motors, transformers, and transmission lines.

6.4 Transformers

Because conventional large-scale transformers are already 99.7% efficient and their current is always ac, which results in losses in the superconductor, the use of superconductors in transformers poses a challenge. As Table 4 indicates, LTSC and HTSC transformers have 38% and 60% lower total costs, respectively, than conventional transformers. The table enumerates the costs of operating a conventional, LTSC, and HTSC transformer for a 30-year period. Because losses are so low in conventional transformers, HTSCs must have very low ac losses. For this design, materials that can sustain a minimum current density of 10^5 A/cm^2 in a field of 0.2 T are required.

Table 4. Cost and performance comparison for 1000-MV A transformers (millions of 1989 dollars).

Parameter	Conventional transformer	LTSC	HTSC
Efficiency (%)	99.7	99.85	99.9
Capital cost	8	16	14
Capitalized losses	92	46	26
Total annual costs	100	62	40

6.5 Other Applications of Superconductors

Superconductivity may be applicable to power electronics, power switching, current limiters, sensors, and computers. For example, HTSC could be used to create a transistor that changes states from on to off, for use in motor controllers. A superconducting switch could have switching speeds as fast as 1 ns, no losses in the on state, and losses in the off state proportional to the product of the normal-state resistivity and the critical current density, ρJ_C. Careful design of the interface between the power leads and the superconducting switch would be crucial to minimize heat leaks and the associated refrigeration power.

7. END-USE TECHNOLOGIES

Electricity is used in a wide range of applications across the residential, commercial, industrial, and transportation sectors, as Table 5 illustrates. Electric motors represent the largest users of electricity across all sectors, accounting for over 50% of electricity use. Lighting accounts for about 20%. About 20% of total electricity goes to cooling and refrigeration, while about 15% goes to space and water heating. (The total exceeds 100% because of the double counting of electric motors contained in refrigeration, cooling, and heating equipment.)

7.1 Electric Motors

Electric motors have become essential to a wide array of end-use applications. The efficiency of today's electric motors ranges from 65% in the smaller fractional horsepower sizes to over 95% in motors with over 125 hp.

Table 5. Electricity usage.

	1987 (GW h)	Percent of sector use	Percent of total use
	Residential		
Refrigerator + freezer	206 391	24.3	8.2
Space heating	159 824	18.8	6.3
Lighting	127 350	15.0	5.1
Water heating	103 499	12.2	4.1
Space cooling	83 381	9.8	3.3
Cooling	30 390	3.6	1.2
Other	138 165	16.3	5.4
Total residential	849 000	100.0	33.6
	Commercial		
Lighting	238 488	31.5	9.4
Space cooling	154 299	20.4	6.1
Information systems	100 000	13.2	4.0
Space heating	77 245	10.2	3.1
Ventilation	76 959	10.2	3.1
Refrigeration	60 883	8.0	2.4
Water heating	24 068	3.2	1.0
Cooking	16 172	2.1	0.6
Other	7 886	1.2	0.2
Total commercial	756 000	100.0	29.9
	Industrial		
Motor drives	570 934	67.5	22.7
Electrolytics	98 193	11.6	4.0
Lighting	84 527	10.0	3.6
Process heating	83 008	9.8	3.2
Other	···	0.8	
Total industrial	845 000	100.0	33.5
	Other		
	76 000	···	3.0
Total consumption	2 526 000	···	100.0

Motors in the 5- to 100-hp range are typically 80% to 90% efficient. Since the 2% of motors over 5 hp use over 45% of all electricity consumed in the United States, small increases in their efficiency yield significant energy savings.

Motor efficiency can be improved either by reducing the losses within the motor itself or by electronically varying the motor speed to match the demand (rather than using a variety of backpressuring devices, such as valves or vanes, which overload the motor and waste energy). With a wide variety of advanced, highly efficient motors becoming available—including magnet, homopolar, and reluctance motors—efficiency improvements of 2% to 4% (absolute) are achievable over the next decade. Moreover, adjustable-speed drives can reduce the annual electricity consumption in a typical motor drive application by 20%. Taken together, more efficient motors coupled with adjustable-speed drives could save 10% of U.S. electricity consumption. Application of such improvements has been limited to date because of the added first cost of the motors and control systems, and the huge existing fleet of installed motors. However, many new applications, such as highly efficient heat pumps, air conditioners, and refrigeration systems, are becoming commercially viable.

7.2 Lighting

Energy-saving compact fluorescent lights are slowly but steadily replacing traditional incandescent lamps. A wide variety of lighting sensors and controls are also available that can automatically dim or turn off lights as appropriate. It is estimated that energy-efficient lighting technology available today could save over 10% of all the electricity now used in the United States.

7.3 Cooling and Refrigeration

The energy efficiency of cooling and refrigeration equipment continues to increase, as Fig. 11 demonstrates, creating the potential for energy savings of as much as 10%. However, because they deplete stratospheric ozone, the chlorofluorocarbons (CFCs) and hydrochlorofluorocarbons (HCFCs) used exten-

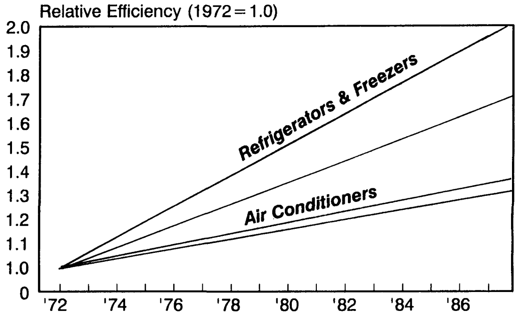

FIG. 11. Trends in energy utilization efficiency of major appliances.

sively as refrigerants in cooling and refrigeration equipment are scheduled to be phased out of production. CFCs will be eliminated by the year 2000; HCFCs will be phased out between 2020 and 2040. The shift to alternative refrigerants may limit future efficiency gains. Figure 12 shows the effects of the near-term phaseout of CFCs on both equipment and cost. Note that the increased cost of equipment may be of even greater concern than uncertainties in efficiency.

7.4 Space and Water Heating

Over the past decade, highly efficient heat pumps have been developed for a wide range of residential, commercial, and industrial applications. Heat pumps remove heat from the environment in order to heat space or water. The efficiency of a heat pump is defined by its coefficient of performance, which is the ratio of the total energy produced as heat divided by the electric energy input. By drawing on heat from ambient air, water, or ground source, heat pumps typically produce 2 to 3 times more heat than the electricity they consume. In that sense, they represent an impressive family of solar-energy devices. Heat pumps used in space and water heating have the potential to save significant amounts of energy—between 40% and 50% for water heating, for instance, and as much as 30% for space heating.

7.5 Other End-Use Technologies

A wide variety of information technologies—from office equipment to computers, printers, and storage and display devices—now account for about 4% of total electricity use. While the number of such devices in use is certain to grow, the improved energy efficiency from increasingly sophisticated computer control is expected to result in a larger reduction in total electricity consumed than is used by computers and their peripherals.

The role of electric-powered transportation is also likely to increase, both for vehicles and for mass transit. This is in response to growing concerns over the environmental and geopolitical impacts of the growing consumption of oil, gasoline, and natural gas. A typical electric vehicle might consume 10 to 30 kW h each day. Large transit systems could increase electricity use by as much as 50 000 GW h annually.

8. ENVIRONMENTAL ISSUES

Like every human activity, the generation and transmission of electricity have an impact on the environment (see AIR POLLUTION: COMPONENTS, CAUSES, and CURES). The primary environmental concerns related to electric-power generation and transmission focus on acid rain, air quality and human health, CO_2 and global climate, electric and magnetic fields, nonradioactive waste disposal, and nuclear waste disposal.

8.1 Acid Rain

Acid rain, a dilute solution of sulfuric acid and nitric acids with a *p*H generally below 5.0, is common in many industrial regions, most notably in western Europe and eastern North America. All rain is naturally acidic because

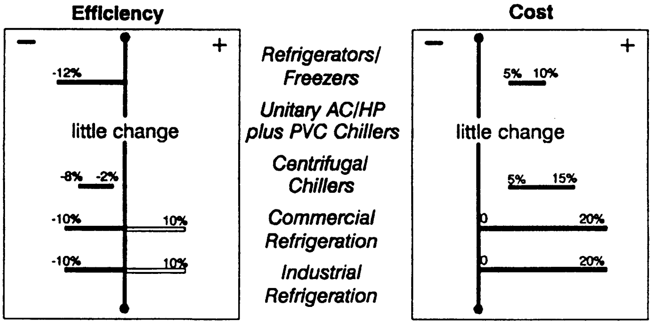

FIG. 12. Distribution of impacts of the phaseout of chlorofluorocarbons (CFCs) on equipment cost and energy efficiency.

of the equilibration with atmospheric carbon dioxide. When anthropogenic emissions of sulfur and nitrogen oxides are oxidized to sulfuric and nitric acids, respectively, they increase rain acidity considerably. Such rains can have a harmful effect on the environment through lake acidification, forest damage, reduced crop yields, material damage, visibility degradation, and impact on human health.

Over the past 15 years, massive research efforts in the United States and Europe have increased our understanding of the acid rain phenomenon. The most visible and serious damage has been lake acidification and resultant damage to fisheries. Fortunately, impact is generally limited to areas where soils have low neutralizing capacity, such as glaciated areas of parts of eastern Canada, the northeastern United States, and northern Europe. Even here, the percentage of known acidified lakes is small. In northern New York state, the most heavily affected area in the United States, for instance, well under 10% of lakes (by surface area) are acidified. Still, there is concern that continued emissions of sulfur and nitrogen oxides at their present levels could cause more widespread and permanent damage to the ecosystem.

Forest decline is another concern. Although impact in North America has been limited, forest declines in some regions of western Germany have been considerable. Acidic deposition appears to be just one factor causing forest decline. However, overall forest mass has increased in Europe.

Sulfate particles derived from sulfur dioxide emissions contribute to reduced visibility, although the relation between visibility impairment and emissions is significantly nonlinear; hence, large emissions reductions are probably needed to effect any meaningful improvement in visibility. Impact on crop productivity and yield appears to be negligible. Because of atmospheric and other factors, it is difficult to assess the role acidic deposition plays in material damage. Indirect health concerns from acidic deposition are minimal.

8.2 Air Quality and Human Health

As the severe episodes in London of the 1950s demonstrated, air pollution can cause increases in morbidity and mortality. In terms of electric-power generation, concern centers on the health effects of atmospheric concentrations of sulfur dioxide, nitrogen oxides, ozone, and acid aerosols. Because of many confounding factors, however, epidemiologic studies have not been able to relate trends in morbidity and mortality or any specific dysfunction to any of these constituents at low levels.

8.3 CO_2 and Global Climate

The use of coal-powered electricity generation has contributed to increasing atmospheric concentrations of CO_2, which some experts believe could lead to dramatic global climate changes. Electricity generation currently accounts for about 30% of CO_2 emissions in the United States. World wide, power production accounts for less than 10% of the CO_2 entering the atmosphere.

Carbon dioxide emissions could be reduced in a number of ways, including scrubbing CO_2 from flue gas, switching from coal to natural gas, increasing the use of nuclear energy, adopting renewable generating technologies, and improving end-use efficiency. Unfortunately, conventional and advanced generating technologies do not seem to offer much hope for reducing or even stabilizing CO_2 emissions well into the next century. The general use of scrubbing can double the cost of electricity, for instance, and it creates large amounts of liquid or solid wastes that must be disposed of at as yet undefined disposal sites. Solar technologies, while promising, probably will not have any meaningful impact for decades. Energy efficiency, on the other hand, remains an attractive means to reduce atmospheric emissions. In the United States, a 10% improvement in energy efficiency would stabilize utility emissions at the 1990 level.

8.4 Electric and Magnetic Fields

Extensive animal studies have minimized earlier concerns about the health impact of exposure to electric fields. But recent studies have raised new concerns about the effects of magnetic field exposure. While several epidemiologic studies showed marginal statistical links between residential electricity distribution lines and cancer, especially childhood leukemia, others have found no correlation.

The electric utility industry has developed more sensitive measurement technologies and is currently funding an extensive epidemiologic study of childhood leukemia and magnetic field exposure in the Los Angeles area. Early results show a weak but not statistically significant association of leukemia with the highest measured fields. However, childhood leukemia is also weakly associated with home use of pesticides and parental occupation during the time of the mother's pregnancy, making interpretation difficult. Epidemiologic studies found an association of cancer with the electric wiring configurations outside the home, but the wiring configuration is only poorly correlated with magnetic field.

As these findings suggest, the impact of electric or magnetic fields at typical domestic levels on human health is far from clear. The electric utility industry continues to pursue research to assess the possibility of health effects of electric and magnetic fields and their potential risks, and possible remedies in designs.

8.5 Nonradioactive Waste Disposal

U.S. utilities produce over 80 million tons of solid waste, mostly ash from coal burning. Although ash is not classified as a hazardous material, it contains constituents that could pose a threat to human health through either the food chain or drinking water. Groundwater sampling at some sites has shown concentrations of sulfate, boron, and iron that exceed proposed water quality standards. Although there have been no cases of drinking-water contamination from these constituents, the utility industry is developing techniques to minimize leaching from disposal sites and to predict the fate of contaminants in ground water.

8.6 Nuclear Waste Disposal

Providing for safe disposal of nuclear waste materials is essential if nuclear energy is to maintain or increase its role in electricity generation. Low-level waste disposal is not a serious concern. Typically, wastes are placed in drums or in lined plastic containers and buried in shallow pits. Over the past few years, because of changes in management practices, the amount of low-level waste produced in the United States has decreased by nearly 50%, despite an increase in nuclear power generation. High-level wastes—including spent fuel with its uranium, transuranics, and other fission products—pose a more serious concern. These wastes are stored on site as a temporary measure until a decision is made about long-term, permanent storage. One proposed approach is deep subsurface storage in salt strata or domes, tuffs, or basalt beds. (See NUCLEAR WASTE DISPOSAL.)

9. STORAGE TECHNOLOGIES

Managing electricity supply is an increasingly difficult challenge for the utility planner and engineer. Pressures to reduce costs are forcing utilities to improve the productivity of both generation and electrical supply systems. Growing use of complex electronic equipment has placed greater emphasis on power quality. To help match electricity supply and demand on a daily basis—and to assure power stability on a millisecond basis—many utilities are investigating energy storage systems (see STORAGE OF ENERGY).

Storage technologies now available convert ac electricity into a suitable form for storage: the potential energy of elevated water or compressed air, the chemical energy of reactive molecules such as hydrogen, the kinetic energy of a flywheel, or the energy stored in a magnetic field. Stored energy can be recovered (with some inevitable losses) and converted again to electricity that can be fed to distribution systems.

Storage systems are comprised of two major subsystems: the power conversion required to change ac electricity into another form of energy and the reservoir to store that energy. Selecting a storage technology is primarily determined by the required response time of the power conversion system and the capacity requirements for the duration of power recovery (and therefore cost) of the power conversion and reservoir systems. In most applications, the power converter and reservoir are separately sized to meet the demands of a given application. Thus a storage device might be sized to provide 1000 MW for 5 s for transmission stability or 100 MW for 10 h for leveling the daily demand for power on the utility's generating equipment.

As Table 6 shows, storage systems being deployed for utility application include lead-acid batteries, pumped hydroelectric storage, compressed air, and superconducting mag-

Table 6. Key storage systems.

Technology	Batteries	Pumped storage	Compressed air	Superconducting magnet
Power converter	ac–dc converter and electrochemical reactor	motor-pump–turbine generator	motor-compressor and expander-generator	ac–dc converter
Reservoir	Reactive chemicals	Elevated reservoir	Underground cavern	Superconducting coil
Storage equation	$\eta F\epsilon m/\mathrm{MW}$ η=electron change F=Faraday's constant ϵ=potential m=mass of chemicals MW=molecular weight	pVh p=density of water V=volume h=head or height of reservoir	PV P=pressure V=volume	$\frac{1}{2}Li^2$ L=inductance i=current
Response (time to full power)	0.02 s	10 min	10 min	0.02 s
1991 U.S. capacity (MW)	10	18 000	100	400 planned for 1996
Capital cost (1991 $)				
Power ($/kW)	150	1000	500	150
Reservoir ($/kW h)	200	10–50	1–5	300
Efficiency (input/ kW h of output)	1.4 kW h	1.4 kW h	0.8 kW h 4100 Btu gas	1.1 kW h

nets. Flywheels have been considered extensively, but the energy storage density and the cost are limited by the strength of materials.

9.1 Lead-Acid Battery

One of the oldest and best known devices to store energy, the lead-acid battery, uses the reduction of lead dioxide to lead sulfate at the positive electrode and the simultaneous oxidation of lead to lead sulfate at the negative electrode. The electrolyte, sulfuric acid, is consumed, and energy is discharged. Energy is stored by reversing the reaction; that is, charging the battery. Under suitable operating conditions, a well-designed and -manufactured lead-acid battery can achieve 2000 cycles and last 10 years. Individual cells with capacities ranging up to a few kW h can be combined in lead-acid battery systems. The cells are arranged in series strings to achieve the desired dc voltage for conversion or for a given application. The number of strings depends upon the energy requirements. Several lead-acid battery systems have been installed for utility energy management, most notably Southern California Edison's 10-MW,40-MW h system in Chino, CA, which has been operating successfully since 1988.

9.2 Pumped Hydroelectric Storage

Pumped hydroelectric storage (or pumped storage, as it is often called) has been used in electric utility systems for over 50 years and now provides 18 000 MW of U.S. capacity. Electricity is used to drive a motor-pump combination that pumps water to an elevated reservoir. When energy is required, water is allowed to flow back to a lower reservoir through the same equipment, which then serves as a turbine-generator. In areas where topographic or ecological considerations rule out conventional pumped hydroelectric plants, the upper reservoir may be at or near the surface, and the lower reservoir can be located about a mile below the earth's surface in a hard-rock cavern.

9.3 Compressed-Air Energy Storage

Compressed-air energy storage (CAES) stores energy by compressing air into an underground cavern or reservoir (excavated rock—for instance, solution-mined salt—or aquifer). Favorable geological sites exist in about 75% of the United States. For the maximum recovery of energy, the energy of compression is augmented by heating the air with oil and gas combustion and expanding the hot gas through a gas turbine to drive a generator. CAES technology has recently been commercialized for electric utility combustion. The commercial units incorporate modified combustion turbines in which the compressor and expander are physically separated by clutches and a motor-generator. The first U.S. plant was built for Alabama Electric Cooperative in 1991 and produces 110 MW for 26 h. A 290-MW CAES plant using two underground salt caverns is currently operating in Huntorf, Germany.

9.4 Superconducting Magnetic Energy Storage

The magnetic energy storage system uses an ac–dc–ac converter system, like batteries, and a large coil of a superconductor such as niobium-titanium alloy. To maintain the superconductive state, the conductor is bathed in liquid helium at about 1.8 K. Superconducting magnetic energy storage (SMES) is particularly attractive for applications where high power is required for very short periods. Best known, of course, is the superconducting magnet at the Fermi National Accelerator Laboratory near Chicago, which is used to focus particles in the accelerator. A small 10-MW 3-s experimental system was recently built to inhibit subsynchronous resonance in a high-voltage ac line from the state of Washington to California. Small SMES systems are now becoming commercially available for improving utility-customer power quality and assuring transmission stability.

9.5 The Future of Energy Storage

Storage systems offer many advantages to electric utilities. Underutilized plants can be used to charge storage systems, thus increasing the capacity factor of base-load plants. Distributed, fast-responding storage systems can level the transmission load and help ensure stability, improving throughput of transmission facilities as well as overall power quality. Because the cleanest plants can be used to charge storage systems, these technologies provide a novel way to manage emissions of generating plants. Storage technologies can reduce the impact of fuel price changes, leverage purchases of wholesale energy, minimize cycling costs of generating plants, and provide economical emergency capacity. Moreover, the ability to store energy will facilitate the use of intermittent renewable energy systems like solar and wind power systems.

Even so, utilities have been relatively slow to deploy storage technologies, in part because the economic advantages have been difficult to evaluate. Conventional utility economic analyses, for instance, while sophisticated and complex, treat storage as electrical generation, not as energy management. Moreover, these analyses usually arrive at single numbers that represent the present value of adding energy storage facilities without taking into account corporate strategies and risks of change (load growth—for instance, load shape—, regulations, and fuel costs), as well as novel applications and opportunities. As the advantages of energy storage become more widely recognized, however—and economic planning tools become more refined—it is expected that energy storage will become an essential part of electric utility systems. With the appropriate support of regulatory agencies, utilities in the future will have practical economic options of maintaining 10% to 20% of their capacity in stored energy. This can become a significant plus for the national economy by adding a highly reliable element of power supply systems that also enables system management options to maximize the productivity of generating and distribution systems.

10. SUMMARY

In the course of 100 years, the field of power engineering has changed dramatically. To the classical disciplines of electrical, mechanical, and civil engineering has been added a wide range of other research fields, from nuclear engineering and data and information analysis to the environmental and atmospheric sciences. New ways to generate electricity have been developed. A vast distribution network that increasingly uses power electronics has been created. The uses of electricity have grown from what first appeared to be a few novelty items to an immense variety of residential, commercial, communications, and industrial end uses. Electricity and its manifold applications have done more, perhaps, than any other single technological factor in recent history to change the shape of human activities.

The future is likely to bring other changes. Environmental concerns, which have already had a profound effect on the electric utility industry, are expected to intensify, and applications of electricity will commonly be part of the remedial and control measures. Advances in superconductors and other materials research and development will offer new and better ways to generate and deliver electric power. The demand for greater energy efficiency will drive the trend toward increased electrification of industrial processes. New end uses—from electric vehicles to increasingly computerized sensors and control devices—will shape the way we use electricity.

A highly reliable and efficient system of generating and delivering electric power wherever and whenever it is needed will continue to be a driving force for both national and international economic stability and growth.

ACKNOWLEDGMENTS

We would like to thank the following individuals from EPRI who contributed to this article: Jim Birk, Milt Klein, Ted Marston, Ralph Perhac, Frank Rahn, Dave Sharma, Dr. Chauncey Starr, and Ron Wolk.

GLOSSARY

Bottoming Cycles: Recovery and use of the rejected low-temperature thermal energy from high-temperature cycles to add to the net energy conversion efficiency.

Breeder Reactor: A reactor cooled by liquid metal such as sodium or lead. It sustains a chain reaction almost entirely with fast neutrons (> 0.5 MeV). With appropriate blanket arrangements, up to 1.5 new fissile atoms are produced for each atom fissioned, for a conversion ratio of 1.5.

Burner Reactor: A reactor similar in physics to a breeder reactor but designed for no net production of fissile material. The high conversion ratios are used internally to provide very long effective reactivity and fuel lifetimes, resulting in low fuel costs.

Combined Cycles: The combination of a primary high-temperature energy conversion cycle with a low-temperature cycle. The latter captures the rejected heat from the high-temperature cycle and commonly uses a different working fluid to attain higher overall conversion efficiency of heat energy to electricity.

End User: In the literal sense, a person or agency using electricity; alternatively, the device or appliance or system that uses electricity.

Fluid Bed: A bed of finely divided solid material held in suspension at lower density by the upward flow of a stream of gas. Fluid beds are used to carry out solid-to-gas heat transfer, chemical reactions, and mass transfer. Examples of applications of fluid beds include gas scrubbers, coal gasifiers, and combustion furnaces.

Megawatt: 10^6 watts; used for either thermal or electrical energy; abbreviated MW. For example, MWt refers to thermal megawatts; MWe refers to electrical megawatts.

Network Analysis: The mathematical description of current flows, especially of transients and abnormal oscillations in high-voltage power distribution networks.

NO_x: The mixture of nitrogen oxides that are produced as a by-product of combustion at high temperatures.

Negative Void Coefficient: Term describing a system that is designed so that an increase in void (or steam) content from a given steady-state condition reduces the reactivity or the multiplication constant of the chain reaction.

Neutron Reactivity: The multiplication constant of a chain reaction, which is 1.000 000 at steady state.

Ocean Thermal Cycle: A low-temperature, low-efficiency Carnot cycle that uses the difference between surface-water and deep-water temperatures.

Power Pool: A regional group of public and private utility generating companies that are linked by a transmission network. The network is operated subject to defined algorithms for the dispatch of power from different producers as a function of the system loads and spatial distributions and producer costs.

Primary System: The cooling circuit that passes through the core of a nuclear reactor.

Quality of Power: A group of attributes of electricity supply, including the annual rates and durations of power outages, the frequency of momentary interruptions, the stability of voltages and electrical frequencies, the degree of purity of wave form and absence of transients, the damping of harmonic and subharmonic oscillations, and phase-stability characteristics of multiphase circuits.

Rate Base: The amount of investment considered in setting the rates charged by a regulated utility. The rate base is governed by rules on what costs are allowed to be counted in the rate base and by the rates of depreciation of the allowed investments.

Reheat: Addition of heat to ("reheating") the working fluid while it is at an intermediate pressure; produces an improvement in the efficiency of the overall cycle.

Renewables: Energy sources that do not deplete their supply—for example, tidal or wind power, various forms of solar-derived energy including biomass for direct combustion or fuel alcohol or methane production, photovoltaic energy, hydroelectricity, and most of the output of heat pumps.

Risk Assessment (Also Called Safety Assessment): The discipline of developing systematic fault trees and event trees to describe known or postulated failure and recovery sequences in complex engineered systems. This process commonly also involves the quantification of the relative and absolute risk levels of the various sequences for purposes of setting priorities for prevention and mitigation measures, and for estimating their relative effectiveness. The quantification process is called probabilistic risk analysis (PRA).

SO_x: Oxides of sulfur, predominantly SO_2.

Space Heating: Comfort heating of habitations, as distinguished from process heating.

Superheat: The condition of a working fluid (commonly steam) that is heated to a temperature higher than the saturation temperature at the given pressure.

Thermal Efficiency: The conversion ratio to characterize any given energy cycle. In the electricity producing industry this is typically expressed as the ratio of energy in units of heat input divided by the output measured in electrical units (e.g., Btu per kilowatt hour). In both engineering and scientific terms it is often expressed as a percentage, the quotient of energy units output from a process divided by the energy input required (×100), both expressed in the same units.

Water Wall: A closely packed planar array of tubes that constitute the walls surrounding the combustion chamber of a boiler and that provide heat transfer from the flame region to water and steam by radiation and convection.

List of Works Cited

Gever, J., Kaufman, R., Skole, D., Vörösmarty, V. (1989), *Beyond Oil: The Threat to Food and Fuel in the Coming Decades*, Cambridge, MA: Ballinger.

Macdonald, E. (Ed.) (1991), *Electrical World Directory of Electric Utilities*, 99th ed. New York: McGraw-Hill.

Further Reading

Committee on Electricity in Economic Growth, National Research Council (1986), *Electricity in Economic Growth*, Washington, D.C.: National Academy Press.

Frisch, J. R., Bendow, K., Saunders, R. (Eds.) (1989), *World Energy Horizons 2000–2020*, Paris: Éditions Technip.

Gever, J., Kaufman, R., Skole, D., Vörösmarty, V. (1989), *Beyond Oil: The Threat to Food and Fuel in the Coming Decades*, Cambridge, MA: Ballinger.

Häfele, W., *et al.* (1990), *Energiesysteme im Übergang—Unter den Bedingungen der Zukunft*, Landsberg am Lech: verlag moderne industrie A.G.

Helm, J. (Ed.) (1990), *Energy Production, Consumption, and Consequences*, Washington, D.C.: National Academy Press.

Pickett, A. P., Gellings, C. W., Lovins, A. B. (1990), **263** (3), *Sci. Am.* 65–74.

Shurr, S. H., Sonnenblum, S., Wood, D. O. (Eds.) (1983), *Energy, Productivity, and Economic Growth*, Cambridge, MA: Oelgeschlager, Gunn & Hain.

ELECTRICAL CONDUCTION

See DIFFUSION AND IONIC CONDUCTION IN LIQUIDS; ELECTRONIC TRANSPORT IN SOLIDS;
INSULATORS AND SEMICONDUCTORS, CONDUCTIVITY IN;
INSULATORS AND SEMICONDUCTORS, SURFACE CONDUCTIVITY IN;
IONIC CONDUCTION AND DIFFUSION IN IONIC SOLIDS;
METALS AND ALLOYS, CONDUCTIVITY IN; TRANSPORT PHENOMENA

ELECTRICAL CONDUCTION AND DIELECTRIC BEHAVIOR IN BIOLOGICAL SYSTEMS

HERMAN P. SCHWAN AND SHIRO TAKASHIMA, *Department of Bioengineering, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.*

	Introduction	1
1.	**Some History**	1
2.	**Elements of Dielectric Spectroscopy**	2
3.	**Relaxation Mechanism**	3
3.1	Maxwell–Wagner Effects	3
3.2	Counterion Relaxation	5
3.3	Membrane Relaxation	5
3.4	Permanent Dipole Effects (Debye Theory)	6
4.	**Experimental Data**	7
4.1	Electrolytes	7
4.2	Biological Macromolecules	8
4.2.1	Polypeptides, Amino Acids	8
4.2.2	Proteins	9
4.2.3	Dielectric Properties of DNA	9
4.3	Biological Membranes	12
4.4	Cells	12
5.	**Tissue Impedance**	13
5.1	Frequency Dependence	13
5.2	Special Topics	15
5.2.1	Anisotropy	15
5.2.2	Live and Excised Tissues	15
5.2.3	Influence of Tissue Morphology	15
5.2.4	Temperature Coefficients	15
5.2.5	Diagnostic Applications	15
5.2.6	Nonlinear Properties	15
5.2.7	Brain Tissue Studies	16
6.	**Field Effects**	16
6.1	Cell Manipulation Techniques	16
6.2	Nerve Conduction	17
6.3	Muscle Contraction and Shock	17
7.	**Techniques**	17
7.1	Suspension Techniques	17
7.1.1	Bridges and Transmission Lines	18
7.1.2	Modern Commercial Equipment	18
7.1.3	Advantages and Limitations of Suspension Techniques	18
7.1.3.1	Electrode Polarization	18
7.1.3.2	Resolution	18
7.1.3.3	Drift	19
7.1.3.4	Mixture Theories	19
7.1.3.5	Membrane Conductance	19
7.2	Single-Cell Techniques	20
7.2.1	Microelectrode Technique	20
7.2.1.1	Microelectrodes	20
7.2.1.2	Voltage Clamp Method	20
7.2.2	Electrorotation	21
7.2.3	Levitation	21
7.2.4	Microspectroscopy	22
	Works Cited	22
	Further Reading	24

INTRODUCTION

The electrical properties of biological materials are most unusual: Their dielectric constants often reach extremely high values, in excess of 10^6 relative to free space, as the frequency decreases below 1 kHz. Furthermore, they change in distinct steps as functions of frequency. We summarize these and other characteristics with the intent to identify relevant conduction mechanisms. We do this by explaining observed properties in terms of "dispersions" caused by "relaxation mechanisms" as revealed by "dielectric spectroscopy."

1. SOME HISTORY

Interest in the response of biological materials to electricity developed early. These efforts may be considered the first of a biophysical nature and mark the beginnings of electrophysiology. This field originated some 200 years ago with the pioneering work of Galvani and Volta. In 1848 du Bois Reymond

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

published his important two-volume text *Untersuchungen ueber die Tierische Elektrizitaet.* Others of the Berlin group include Mueller and Helmholtz. Helmholtz combined expertise in physiology and anatomy and physics. During the 1870s important work was done by Herman (1871). He determined, for the first time, the resistance anisotropy ratio of muscle and nervous tissues at dc and studied excitability. His results were found to be consistent with the assumption that cells are surrounded by an envelope. The envelope, termed membrane, was thought to contain a "polarization" element. It also became clear that excitability threshold currents are lower if applied in the direction of the fibers instead of across them. Limited instrumentation was applied with great ingenuity at the time.

The early electrophysiology work included study of the electrical conductance of blood and tissues and study of excitability and contractility. First determinations of the electrical properties of blood were carried out shortly after the introduction of the Wheatstone bridge and the precision conductivity determinations of electrolytes by Kohlrausch. Hoeber (1910, 1912) compared the conductivities of erythrocytes at low and high frequencies and deduced, from the difference, the existence of a cell membrane. This was the first work to do so from purely electrical data.

Interest in the electrical properties of cells and tissues increased rapidly during the period separating the two world wars. H. Fricke, K. S. Cole, H. J. Curtis, and H. Daenzer were the first to apply potential theory based on Maxwell's early work (Maxwell, 1873) to cell suspensions. Advanced instrumentation was applied to collect data spanning an ever increasing frequency range extending from some kHz to about 100 kHz. This work resulted in an explanation of the dispersive behavior in the rf range (termed β dispersion by Schwan, 1957). A detailed survey of the earlier work was given by Rajewsky (1938) and later in part I of Cole's text (Cole, 1968). After World War II the frequency range was extended to include data well below the kHz range, and upwards into the microwave frequency range. Additional and partially unexpected dispersion phenomena were discovered (α, δ, and γ; Schwan, 1957). Extension into the nonlinear range led to important contributions to membrane biophysics and identification of the mechanism responsible for excitability (Cole, 1968).

2. ELEMENTS OF DIELECTRIC SPECTROSCOPY

The linear electrical properties of matter are defined as follows:

Conductivity σ=conductance G of a unit volume of sample.

Permittivity (or dielectric constant) $\epsilon=\epsilon_r\epsilon_v$ =capacitance C of a unit volume.

ϵ_r is relative permittivity ($=\epsilon/\epsilon_v$) where ϵ_v is the permittivity of free space (8.85×10^{-12} F/m).

Conductivity and permittivity can be combined into complex forms: $Y=\sigma+j\omega\epsilon$ =admittivity (or specific admittance); $\epsilon^*=\epsilon'-j\epsilon''=\epsilon-j\sigma/\omega$=complex dielectric constant. ($\omega=2\pi f$ with f frequency and $j=\sqrt{-1}$. Often ϵ is used in place of ϵ' to represent the real part of complex permittivity.)

Admittivity Y and complex permittivity are related by $Y=j\omega\epsilon^*$. ϵ and σ are usually scalar quantities. Very rarely, however, they must be treated as tensors.

ϵ and σ often change with frequency. In the simplest case such frequency dependence can be expressed by the equations

$$\epsilon^*=\epsilon_\infty+\frac{\epsilon_0-\epsilon_\infty}{1+j\omega\tau}-\frac{j\sigma_0}{\omega}, \tag{1}$$

$$\sigma^*=\sigma_0+\frac{\sigma_0-\sigma_\infty}{1+j\omega\tau}+j\omega\epsilon_\infty. \tag{2}$$

The subscripts 0 and ∞ indicate limit values at low and high frequencies, and the third terms, $j\omega\epsilon_\infty$ and $-j\sigma_0/\omega$, have been added to include low- and high-frequency limits of σ and ϵ, caused by other mechanisms. The "relaxation" mechanism responsible for this dispersive behavior is characterized by the time constant τ and the dielectric increment $\epsilon_0-\epsilon_\infty$.

The frequency dependency stated in Eq. (1) corresponds to an exponential time response of dielectric displacement **D** to a sudden change in field strength and that in Eq. (2) to a change in current density J. Equations (1) and (2) are interrelated through

$$\epsilon_0-\epsilon_\infty=(\sigma-\sigma_0)\tau. \tag{3}$$

Equations (1) and (2) can be separated into

$$\epsilon=\epsilon_\infty+\frac{\epsilon_0-\epsilon_\infty}{1+(\omega\tau)^2},\quad \sigma=\sigma_0+\frac{\sigma_\infty-\sigma_0}{1+(\omega\tau)^2}. \tag{4}$$

If the time response is not exponential, the behavior may be characterized by a sum or integral of exponential responses. A more generalized equation has been formulated for multiple relaxation times as

$$\epsilon^*=\epsilon_\infty+(\epsilon_0-\epsilon_\infty)\int_0^\infty\frac{G(\tau)d\tau}{1+(\omega\tau)^2}, \tag{5}$$

where $G(\tau)$ is a distribution function. A popular distribution of such responses is the logarithmic symmetrical Cole–Cole distribution of time constants. It yields simple extensions of Eqs. (1) and (2):

$$\epsilon^*=\epsilon_\infty+\frac{\epsilon_0-\epsilon_\infty}{1+(j\omega\tau)^\alpha}-\frac{j\sigma_0}{\omega}, \tag{6}$$

$$Y=\sigma_0+\frac{\sigma_\infty-\sigma_0}{1+(j\omega\tau)^\alpha}+j\omega\epsilon_\infty. \tag{7}$$

Equation (6) yields circles in the dielectric plane (plot of ϵ'' vs ϵ') if $\sigma_0=0$, and Eq. (7), likewise, produces circles in the admittance plane ($\omega\epsilon'$ vs σ) if $\sigma_\infty=0$. The circle centers are below the real axis for $\alpha<1$. Plots in dielectric and admittance planes are useful in order to determine limit values at low and high frequencies. Figure 1 summarizes the above discussed frequency dependences. For details see Cole and Cole (1941), Boettcher (1952), and Froehlich (1949).

3. RELAXATION MECHANISM

The relaxation mechanism responsible for dispersive behavior may be caused by either permanent or field-induced dipoles.

3.1 Maxwell–Wagner Effects

A strong contribution to dielectric relaxation in biological cell suspensions and tissues is caused by the charging of interfaces separating different cellular compartments. Since $D=\epsilon E$ and $J=\sigma E$ (D is dielectric displacement, J current density, and E electrical field strength), D/J is equal to ϵ/σ. In a system free of electrical charges, D is continuous across any given interface. Hence the current density component normal to the interface will vary unless the ratio ϵ/σ is the same on both sides of the interface. Thus charges accumulate at the interface with time until J becomes continuous. The speed of adjustment depends on geometry and electrical properties. Simple models were first treated by Wagner (1914, 1924), who extended relevant dc calculations by Maxwell (1873) to alternating fields. These include an arrangement of two slabs of material in series and suspensions of spherical particles. The admittivity Y of a suspension of spherical particles is given by

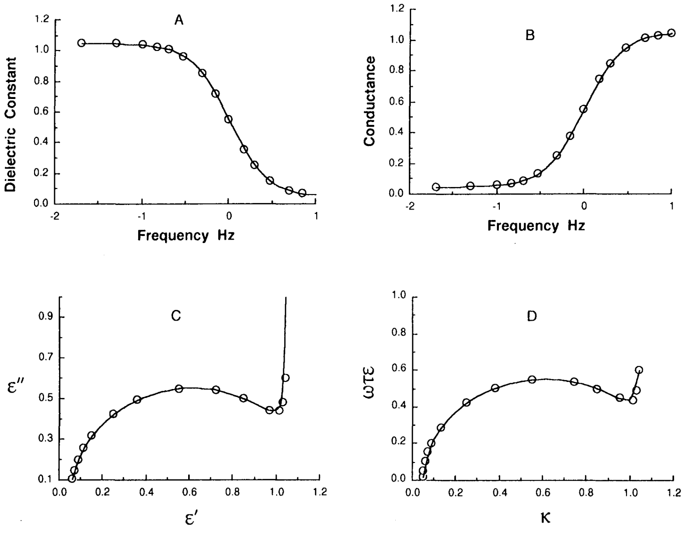

FIG. 1. Dispersion behavior of the permittivity ϵ and conductivity σ as function of frequency (curves A and B) and in the complex dielectric and admittance planes (curves C and D).

$$\frac{Y-Y_a}{Y+2Y_a}=p\,\frac{Y_p-Y_a}{Y_p+2Y_a}\,. \tag{8}$$

Here Y_a and Y_p indicate the admittivity of external medium and particle, and p is the particle volume fraction. This equation can be transformed into Eq. (4) with the parameters (Pauly and Schwan, 1959)

$$\tau=\frac{(1-p)\epsilon_i+(2+p)\epsilon_a}{(1-p)\sigma_i+(2+p)\sigma_a}\,, \tag{9}$$

$$\epsilon_0-\epsilon_\infty=\frac{9p(1-p)(\epsilon_a\sigma_i-\epsilon_i\sigma_a)^2}{[(1-p)\epsilon_i+(2+p)\epsilon_a][(1-p)\sigma_i+(2+p)\sigma_a]^2}\,, \tag{10}$$

where ϵ_i and σ_i are the permittivity and conductivity of the cell interior. The dielectric increment $\epsilon_0-\epsilon_\infty$ is usually of the order of a few dielectric units relative to free space, i.e., of small magnitude.

We consider next a spherical model for a biological cell, assuming a homogeneous interior, the cytoplasm, surrounded by a membrane. In this case Maxwell's "equivalent homogeneous" particle admittivity equation

$$\frac{Y_p-Y_m}{Y_p+2Y_m}=\frac{R^3}{(R+d)^3}\,\frac{Y_i-Y_m}{Y_i+2Y_m} \tag{11}$$

can be used to reduce the shell-surrounded particle to a homogenous one (Y_p is the equivalent particle admittivity, R core radius, and d membrane thickness; i and m indicate cell interior and membrane). Extension of Maxwell's dc equations to the complex case is justified since all boundary conditions such as continuity of potentials and current density normal to the interfaces remain valid.

Combination of Eqs. (8) and (11) gives the electrical properties of a suspension of spherical biological cells. The result is given by the sum of two relaxation equations of the type stated in Eqs. (4). For $d \ll R$ and parameters typical of biological cells, the two time constants differ by a factor of nearly 100. Then the following excellent approximations for the parameters result for the low-frequency dispersion (Pauly and Schwan, 1959):

$$\tau=RC_m\,\frac{(2+p)\rho_i+(1-p)\rho_a}{2+p+RG_m[(2+p)\rho_i+(1-p)\rho_a]}\,, \tag{12}$$

$$\epsilon_0-\epsilon_\infty=\frac{9p}{(2+p)^2}\,\frac{RC_m}{[1+RG_m\{\rho_i+\rho_a(1-p)/(2+p)\}]^2}\,. \tag{13}$$

$C_m=\epsilon_m/d$ and $G_m=\sigma_m/d$ are membrane capacitance and conductance per unit area, $\rho_i=1/\sigma_i$ and $\rho_a=1/\sigma_a$ are the resistivities of internal and external fluids, and p is volume concentration. The parameters for the high-frequency dispersion are identical with Eqs. (9) and (10).

Maxwell's derivations of Eqs. (8) and (11) assume that charges only accumulate at surfaces, i.e., that there are no space charges. Thus Laplace's potential equation can be assumed to be valid in all phases. This is not quite correct for charged particles suspended in aqueous media, and Poisson's equation ought to be solved. However, the error has been demonstrated to be small for cellular dimensions which are large compared to the Debye screening length (Garcia *et al.*, 1985).

In Table 1 are listed values for the case of typical cell suspensions. The data demonstrate the following:

1. The magnitude of low-frequency dispersion is very large, resulting in ϵ values above 10^3. Thus, the Maxwell–Wagner mechanism predicts a large magnitude dispersion at frequencies between 0.1 and 10 MHz.
2. The high-frequency dispersion is independent of C_m and G_m. It is, therefore, simply that of a suspension of core particles without a membrane and its magnitude is small.

Maxwell–Wagner dispersion effects occur if the composition of matter is heterogeneous. The accumulation of charges of opposite sign

Table 1. Maxwell–Wagner β dispersion parameters are given for two cell sizes. The following typical cell properties are used: $C_m = 1$ μF/cm^2, $\rho_i = 120$ Ω cm, $\rho_a = 60$ Ω cm, and $p = 0.3$. The effect of the membrane conductance has been neglected since it is usually small. The data are theoretically predicted [Eqs. (3) and (9)–(13)]. The low-frequency β_1 dielectric increment (relative to free space) is many orders of magnitude stronger than the high-frequency β_2 dispersion. The conductivity increments (in mS/cm) differ by a factor of nearly 100.

β dispersion	R(μm)	f_c(MHz)	$\epsilon_0 - \epsilon_\infty$	$\kappa_\infty - \kappa_0$
1	10	1.15	5760	3.68
	1	11.5	576	3.68
2	10	370	0.20	0.04
	1	370	0.20	0.04

on both sides of the interfaces create induced dipole moments. Thus the dispersion effects reflect frequency-dependent induced dipoles.

3.2 Counterion Relaxation

Another induced dipole relaxation is caused by the movement of counterions. Frequently particles carry fixed charges and are surrounded by an atmosphere of counterions of opposite charge. The application of an electrical field results in a displacement of this atmosphere creating an induced dipole moment. Colloidal suspensions display very large $\epsilon_0 - \epsilon_\infty$ increments and low characteristic frequencies, as reported by us (Schwan, 1957; Schwan *et al.*, 1962). Figure 2 presents some typical data. The time constant τ of this strong effect was found to be proportional to $1/R^2$.

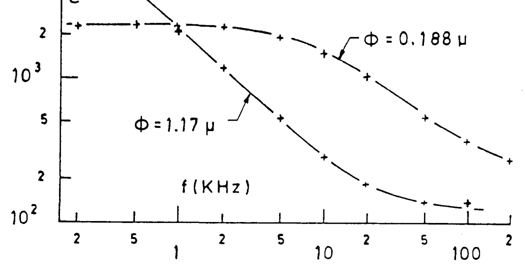

FIG. 2. Dielectric dispersion of polystyrene particles suspended in aqueous media. Numbers for each curve indicate the radius of particles in μm. [From Schwan (1957); reprinted by permission of Academic Press.]

Schwarz (1962) developed a first model for counterion relaxation. His equations for time constant τ and dielectric increment are

$$\tau = R^2/2ukT, \tag{14}$$

$$\epsilon_0 - \epsilon_\infty = \frac{9p}{(2+p)^2} \frac{e_0^2 R \delta_0}{\epsilon_v kT}, \tag{15}$$

where e_0 and u are charge and mobility of counterion; R, particle radius; p, particle volume concentration; T, absolute temperature; δ_0, counterion charge density; and k, Boltzmann's constant. Schwarz's theory was generalized to apply to prolate ellipsoids and further modified for long rodlike molecules (Takashima, 1989):

$$\epsilon^* = \frac{e_0^2 \delta_0 a^2}{b \epsilon_v kT} \frac{9p}{(1+p)^2} \frac{1}{1+j\omega\tau}. \tag{16}$$

Here, a and b are major and minor axes of very long rods.

Refinements of the Schwarz theory have been formulated recently (Fixman, 1983; Chew, 1984; Grosse and Foster, 1987). Since biological cells carry many fixed charges, counterion movement does contribute to their dispersive behavior at low frequencies.

3.3 Membrane Relaxation

The electrical properties of biological membranes are characterized by membrane capacitance per unit area C_m and membrane conductance G_m. Typical membrane capacitance values are about 1 μF/cm^2 and appear to be fairly frequency independent above some kHz. This value corresponds to a lipoprotein film of about 40 Å with a dielectric constant of about 4.

Membrane conductance values range from 0.001 to 1 S/cm^2. They are caused by membrane pores, "channels," responsible for ion exchange. Membrane conductance values at low frequency change with the applied field strength and frequency. The model equations by Hodgkins and Huxley (1952) describe experimental data. The linearization of these equations for small applied fields results in dispersion equations which fit experimental data (Fishman, 1985). Time constants are of the order of ms or less. Thus dispersive behavior is observed below 100 Hz. The mechanisms which cause the frequency dependence of C_m and G_m have not been clarified so far.

Since the membrane conductance changes with frequency, the membrane capacitance must also change as predicted by Eq. (3). Furthermore, since dielectric increments of cell suspensions are proportional to the membrane capacitance [Eq. (13)], membrane properties must cause changes in the low-frequency properties of cell suspensions. This membrane effect is superimposed on the counterion effect discussed above.

3.4 Permanent Dipole Effects (Debye Theory)

Many molecules possess electrical dipoles caused by an asymmetric distribution of positive and negative charges. These polar molecules tend to orient with an applied field, while thermal motion opposes preferential orientation. The development of the theory of polar molecules by Debye (1929) had a far-reaching impact. On the basis of the concept of orientational polarization, Debye derived an equation for the static electric polarization of molecules so dilute that intermolecular interaction can be neglected:

$$\frac{\epsilon_0-1}{\epsilon_0+2}V=\frac{4\pi N_A}{3}\left(\alpha+\frac{\mu^2}{3kT}\right). \tag{17}$$

Here ϵ_0 is the static (low-frequency) dielectric constant, V the molar specific volume, N_A Avogadro's number, and μ the dipole moment. The term α represents electronic and atomic distortions. It determines the high-frequency limit ϵ_∞ since these processes are very fast. For dilute molecules, the dielectric constant ϵ is not much larger than 1 and under these conditions Eq. (17) yields satisfactory results. It soon became apparent that this theory has limited applicability for polar substances having large dielectric constants. The shortcoming was found to stem from an insufficient analysis of internal fields. Subsequently, Onsager (1936) introduced the concept of a "direction field" and successfully eliminated the inconsistency of Debye's equation. Onsager's equation is

$$\frac{(\epsilon-n^2)(2\epsilon+n^2)}{\epsilon(n^2+2)^2}=\frac{4\pi N_A\mu_0^2}{9VkT}, \tag{18}$$

where n is the refractive index and μ_0 the dipole moment. Onsager's equation correctly predicts the dielectric constant of many polar substances. However, Onsager's equation was not able to account for the electric polarization of polar liquids such as water with large dielectric constant. The distribution of polar molecules in liquids is more ordered than in gases. Consequently, the mean dipole moment of liquids is not necessarily zero in the absence of an electrical field. Thus two reference axes are needed to define the orientation of individual molecules in liquids. One axis is parallel to the net dipole moment of the liquid with zero field, and the other to the vector of the applied field. Kirkwood (1939) derived

$$\frac{(\epsilon-1)(2\epsilon+1)}{9\epsilon}V=\frac{4\pi N_A}{3}\left\{\alpha+\frac{g\mu^2}{3kT}\right\}, \tag{19}$$

where g is the "correlation parameter" defined by

$$g=1+z\langle\cos\theta\rangle; \tag{20}$$

z is the number of nearest neighbors, and θ is the angle between adjacent molecular axes. For example, in liquid water with tetrahedral structure, $z=4$ and $\theta=45°\text{–}50°$. Satisfactory agreement with experimental results, however, demands not only the nearest neighbors but also additional layers of molecules to be included. Further refinements assume bending (Pople, 1951) or breaking of hydrogen bonds (Haggis *et al.*, 1952) between adjacent water molecules.

Dielectric relaxation of polar molecules is caused by the phase lag between the applied sinusoidal field and the dipole orientation. Dipolar molecules are unable to follow closely changes in the direction of the applied ac field. This causes phase shift and energy dissipation, resulting in a decrease in dielectric constant. If the relaxation is characterized by only one relaxation time, the dispersion of the dielectric constant is given by Eq. (1) (Debye, 1929), with τ now a cubic function of molecular radii. Distributions of relaxation times are often caused by molecular inhomogeneity and nonspherical shape. They are fairly common with synthetic polymers and biopolymers such as DNA. The origin of the distribution of relaxation times has been discussed by various authors.

The time constant of dielectric relaxation is, as mentioned before, proportional to the cube of the radii of molecules. For example, the dielectric dispersion is observed at 15–20 GHz for water, and 400–500 MHz for simple amino acids. The dispersion curves of some spherical proteins fit nicely the Debye theory of single relaxation time. However, the dispersion curves of other proteins deviate from it, indicating the presence of more than one relaxation time. For ellipsoidal molecules, there are separate orientations along major and minor axes. Then, Eq. (5) must be rewritten using two dispersion terms (Oncley, 1943):

$$\epsilon-\epsilon_\infty=\frac{\Delta\epsilon_1}{1+(\omega\tau_1)^2}+\frac{\Delta\epsilon_2}{1+(\omega\tau_2)^2}, \qquad (21)$$

where $\Delta\epsilon_1$ and $\Delta\epsilon_2$ are the dielectric increments due to the orientation along the major and minor axes. τ_1 and τ_2 are relaxation times associated with these orientations. The rotation of ellipsoidal molecules can be analyzed using Perrin's theory of rotary diffusion (see Oncley, 1943). This theory relates the rotary diffusion constants of ellipsoidal particles to their axes a and b. The relaxation time is calculated as the reciprocal of the rotary diffusion constant. Therefore, for a certain value for a/b, the ratio τ_1/τ_2 can be calculated. The ratio of $\Delta\epsilon_1$ and $\Delta\epsilon_2$ is determined by dipole angle, i.e., the angle between dipole vector and molecular axes. Therefore, Eq. (21) can be solved from assumed axial ratio and dipole angle θ. By best fit, we can obtain satisfactory agreement between calculation and experiment, and determine the shape of the protein molecule. This approach is useful for proteins with large axial ratios, i.e., 5–10. However, the theory is of limited use for proteins with an axial ratio of less than 2–3.

4. EXPERIMENTAL DATA

Figure 3 displays the dielectric behavior typical for many tissues and cell suspensions. Three relaxation regions of large magnitude are apparent at low, medium, and very high frequencies, identified as α, β, and γ (Schwan, 1983). Figure 4 details the same principal behavior for a protein solution. In this case there is no α effect and the magnitude of the β effect is less pronounced. There is evidence for a weak and broad additional relaxation effect labeled δ. Clearly, magnitude and time constants of observed dispersions vary considerably. A summary of these dispersions is provided in Table 2. We now consider the various components of cell suspensions and tissues and their individual dielectric properties.

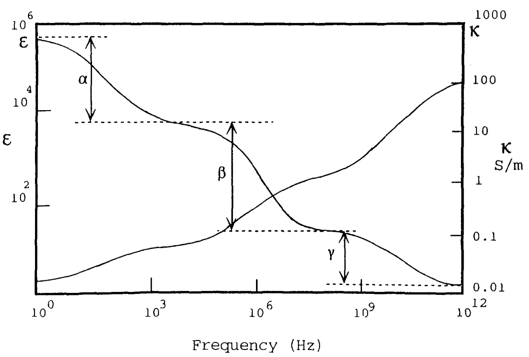

FIG. 3. Stepwise dispersions of biological materials. α dispersion is due to ionic processes, β dispersion is due to charging up of the membrane or orientation of permanent dipoles. γ dispersion is due to the orientational relaxation of water.

4.1 Electrolytes

Water is in many ways unique: It has high heat capacity, a high boiling point, and a large dielectric constant. The dipole moment of water is 1.84 D bisecting the angle between two OH bonds. Formation of intermolecular hydrogen bonds makes water a highly ordered liquid. Calculations of its dipole moment using Kirkwood's equation (1939) did not yield good results. Better agreement was obtained by Pople (1951) by allowing the bending of hydrogen bonds. Even better agreement was obtained by Haggis *et al.* (1952) by assuming the breaking and reforming of hydrogen bonds and including several layers of molecules. It was also found experimentally that the energy barrier of dipole rotation of water is about 3–4 kcal/mol (Hasted *et al.*, 1953). This observation indicates that the orientation of a water molecule requires a breakage of only one hydrogen bond.

The dielectric relaxation is characterized by only one time constant corresponding to a characteristic frequency $f_0=1/2\pi\tau$ near 20 GHz at room temperature. The dielectric constant decreases from about 78 to 4 as the frequency increases. The low-frequency term σ_0 is about 10^{-5} S/m. The dielectric response of electrolytes is quite similar, except for a small decrease of ϵ_0 and a larger conductivity σ_0. The low-frequency conductance is caused

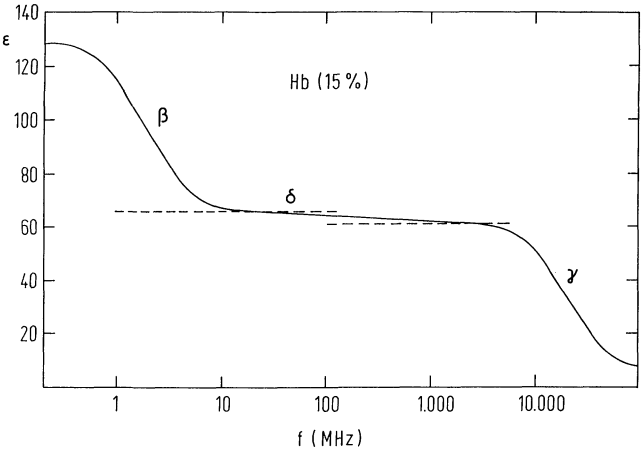

FIG. 4. Schematic presentation of the dielectric dispersion of hemoglobin solution. δ dispersion is due to the water bound to the surface of protein molecules. Abscissa, logarithmic scale of relative frequencies (f/f_r, where f_r is the center frequency). [From Schwan (1983); reprinted by permission of Springer-Verlag.]

by the mobile ions and is approximately proportional to the ion concentration.

The behavior of water in biological cells has been debated vigorously (Whipple, 1965; Hazelwood, 1984). In spite of the gel-like nature of cytoplasm, the dielectric behavior of intracellular water is amazingly similar to that of bulk water except for water tightly bound to macromolecules and membranes. Protein-bound water appears to display an entirely different behavior (Schwan, 1957 and 1965; Grant, 1965). The relaxation frequency is much lower than that of normal water. In addition, the behavior indicates a broad spectrum of relaxation times, corresponding to frequencies ranging from some hundred MHz to a few GHz. Permittivity limit values ϵ_0 and ϵ_∞ are comparable to those of water and ice.

4.2 Biological Macromolecules

4.2.1 Polypeptides, Amino Acids. Polypeptides, proteins, and nucleic acids are the most investigated biopolymers. Early dielectric constant measurements were performed with amino acids and simple peptides. It was found that amino acids possess large dipole moments. This dipole moment is due to the charge pair of COO^- and NH_3^+ groups, separated by about 3.33 Å in α amino acids and

Table 2. Relaxation mechanisms which determine the dielectric response of electrolytes, biological macromolecules, and cells. Strong ELF, rf, and MW responses are labeled α, β, and γ. γ stands for the water dispersion, β is the external cell membrane charging response, β_1 and β_2 the weaker protein response and that from intracellular organelle membranes. α_t and α_1 are contributions from membrane systems connecting with the outer cell membrane and counterion displacement near fixed surface charges of the membrane glycocalix, and α_m is caused by membrane channels. The weaker δ response at UHF is caused by protein-bound water and submacromolecular partial rotation. The table demonstrates participation of an increasing number of relaxation mechanisms as we proceed from simple water to the complex cell.

Water and electrolytes	γ
Biological macromolecules	
Amino acids	$\delta+\gamma$
Proteins	$\beta_1+\delta+\gamma$
Nucleic acids	$\alpha_1+\beta_1+\delta+\gamma$
Biological membranes	α_m
Vesicles	$\beta+\gamma$
Organelles (no surface charge)	$\beta_2+\delta+\gamma$
(with surface charge)	$\alpha_1+\beta_2+\delta+\gamma$
Cells with proteins	$\beta+\beta_2+\delta+\gamma$
+organelles	$\beta+\beta_1+\beta_2+\delta+\gamma$
+surface charge	$\alpha_1+\beta+\beta_1+\beta_2+\delta+\gamma$
+membranes+channels	$\alpha_m+\alpha_1+\beta+\beta_1+\beta_2+\delta+\gamma$
+tub. system, etc.	$\alpha_t+\alpha_m+\alpha_1+\beta+\beta_1+\beta_2+\delta+\gamma$

4.16 Å in β amino acids (Wyman, 1936). The formation of peptides increases the distance between positive and negative charges, hence causing a larger dipole moment. There are two known conformations for highly polymerized polypeptides and/or polyamino acids. These are the random coil and alpha helix structures (Pauling *et al.*, 1951). The former is an amorphous structure without an internal regularity, and the peptide chain configuration is random and flexible. The net dipole moment of random-coil polypeptides is usually very small. The latter is in a crystalline state characterized by a highly ordered secondary structure having polar groups aligned in one direction along the helical axis. The sum of the dipole moments of these polar groups turns out to be very large and proportional to the degree of polymerization (Wada, 1962). In addition, there is another structure called β-pleated sheet. However, no information is available as to the dipole moment of this configuration.

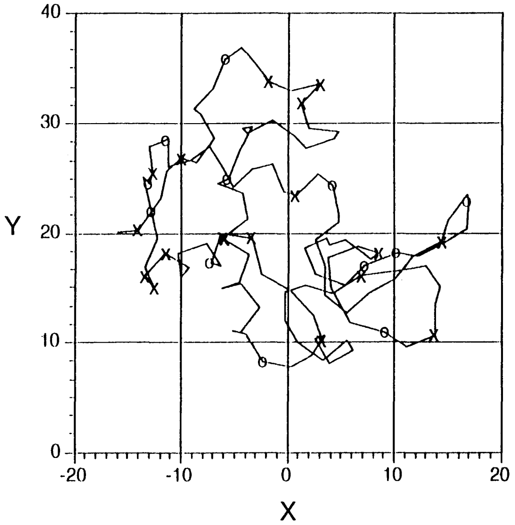

FIG. 5. Two-dimensional presentation of a protein molecule, lysozyme, in the *X*-*Y* plane. Crosses represent positive charges and open circles represent negative charges.

4.2.2 Proteins. There are two types of protein molecules: fibrous proteins such as collagen and muscle proteins, and globular proteins such as hemoglobin, albumin, and insulin. Although the internal structure of these proteins is very complex, the coordinates of their amino acids are accurately known for many proteins from x-ray crystallography. The folding of the peptide chains in proteins often appears to be structureless, except for helical regions, because of their extreme complexity. The conformation of protein molecules is now known to be the result of an intricate interplay among hydrophobic and electrostatic forces. Protein molecules carry positive and negative charges on the surface, and their locations are well known (see Fig. 5). The dipole moment of proteins is largely determined by the distribution of these charges. The distribution of surface charges has been found to be rather uniform in many proteins and the net moment is smaller than expected for their large size. Calculations of the net dipole moment based on the knowledge of the distribution of surface charges have been conducted for many proteins as shown in Table 3 (Schlecht, 1969; South *et al.*, 1972; and Barlow *et al.*, 1983). The calculations consist of two parts, the dipole moment due to fixed surface charges (column 1), and the vector sum of bond dipole moments of individual amino acids (core moment, column 2). The core moment is considerably smaller in magnitude than that due to surface charges, and its vector is often opposite to that of the fixed charges. Column 4 shows the results of the measurements of the dipole moments of these proteins. In general, agreements are reasonably good for small proteins (see myoglobin and cytochrome *C*, e.g.), and poor for large and complex proteins which often have subunits.

4.2.3 Dielectric Properties of DNA. DNA is a very large fibrous polymer which stores all genetic codes using the unique sequences of its base molecules. It has been postulated that the origin of the dipole moment of polyelectrolytes stems from the fluctuation of mobile counterions caused by an electrical field. Some exemplary results of dielectric-constant measurements with DNA solutions are presented below. Because of the double helical structure, DNA molecules are fairly rigid even in solution. Moreover, polynucleotide chains are negatively charged because of the presence of phosphate groups. Because of their large sizes, the dielectric dispersion occurs at very low frequencies, namely around 5–8 Hz. Figure 6 shows the frequency profile of the dielectric constant of a dilute solution (0.05%) of a highly polymerized DNA (molecular

Table 3. Calculated and measured dipole moments of globular proteins (in Debye units). [From Takashima and Asami (to be published).]

		Calculated dipole moment				
Name	M.W.	Charge	Core	$\langle\theta\rangle^a$	Total	Measured
Myoglobin	17 000	247	57	148	200	172
Cytochrome C	13 000	255	72	142	203	227
Lysozyme	14 600	130	61	90	141	122
Trypsin^b	23 000	356			356	345
Ribonuclease	13 700	330	19	85	270	281
Carboxypeptidase	34 000	435	152	2	575	760
Alcohol dehydrogenase	79 000	710	174	122	633	531

[a] Angle between charge and core moment vectors (in degrees).
[b] From Barlow and Thornton (1986).

weight about 3×10^6) (Hayakawa *et al.*, 1975). We first point out that the static dielectric constant of DNA solution reaches a high value of 4×10^4 at a concentration of only 0.05%. This corresponds to a dipole moment of 1.5×10^6 D. This large dipole moment was found to be proportional to the degree of polymerization (Takashima, 1966), indicating that the dipole moment of DNA exists along the major axis of the double helix, in contrast to the early concept that DNA has a transverse moment (Allgen, 1949). Structural considerations, i.e., the antiparallel double helical configuration, preclude the possibility of the DNA molecule having a large permanent dipole moment. This is because of cancellation of possible dipole moments in each strand.

Left for consideration is counterion polarization. For DNA molecules, the magnitude of major axis a is of the order of 5–6 μm, the minor axis b is of the order of 20–30 Å, and charge density δ_0 about 10^{13} cm^2. Thus, Eq. (16) predicts an exceedingly large dielectric increment as observed.

The theory above uses a model in which charged molecules are surrounded by a continuous distribution of counterions. This model does not consider the specific interaction between ions and charged sites. Binding of counterions to charged sites on the polyion's surface has a profound effect on the dielectric behavior of polyelectrolytes. For example, binding of counterions causes a drastic reduction of the dielectric constant of polyelectrolytes. In order to account for this effect, the model must have a feature which can explain the observed effects of ion binding. A model which may be termed "a discrete potential well model" has been used for some time for linear charged polymer molecules. Figure 7 is a schematic presentation of this model (Mandel, 1961). It consists of an array of potential wells which are separated from one another by an energy barrier h. If ions are found in these potential wells, they are considered bound to charged sites, i.e., neutralization of the charge, otherwise they are considered unbound. The dipole moment of linear poly-

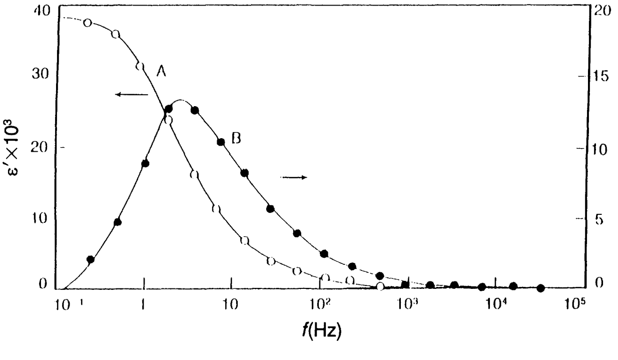

FIG. 6. Dielectric dispersion of the DNA solution measured with a four-terminal technique. A is the dielectric constant, and B is the dielectric loss: dissipation of energy.

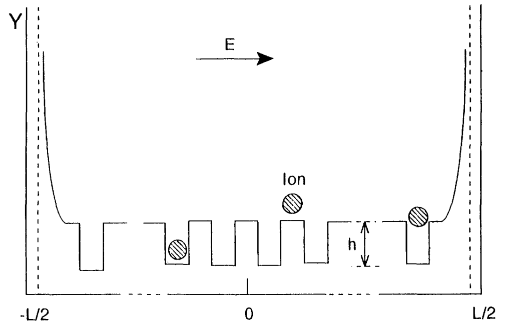

FIG. 7. Schematic presentation of a one-dimensional discrete potential well model. *h* is the height of potential barriers, and shaded circles show counterions. *L* is the length of linear polymers.

ions is calculated by computing the distribution of unbound as well as bound counterions along the fiber in the presence of an electrical field. This discrete potential well model has been used by several investigators (Mandel, 1961; McTague and Gibbs, 1966; Oosawa, 1970; Warashina *et al.*, 1973; and Minakata *et al.*, 1972). According to these investigations, the binding of counterions causes an increase in dielectric constant when charged sites are only partially occupied. However, further ion binding causes a marked decline of the dielectric constant (Minakata *et al.*, 1972). This prediction is in agreement with the results with synthetic polyelectrolyte molecules. Counterion binding to charged DNA-phosphate groups causes a marked decrease in dielectric constant. However, the initial rise in $\Delta\epsilon$ which is predicted by aforesaid theories has not been observed (Takashima, 1966).

Macromolecular solutions display a variety of responses as indicated in Table 2 and Fig. 3. They all share the γ response due to the presence of water. In addition they display a weak δ response which is probably mostly caused by bound water. The amount of bound water is about 0.3 g of water per g protein for globular proteins. Macromolecular solutions frequently display a γ dispersion which is shifted to slightly lower frequencies and involves a broadening from one to a small distribution of relaxation times. The mechanism of this effect remains unclear. It can result from the interaction of the high-frequency tail of the δ spectrum with the γ dispersion. We support this by two arguments: (a) The effect apparently increases with macromolecular concentration. (b) It appears to be unlikely that all free water shifts its properties in common as proteins are added to the pure water. Relaxation contributions depend on the distance of the water from the macromolecule, with "distant" water less affected. The δ dispersion of macromolecular solutions is about one dielectric unit per g protein per 100 cm^3, i.e., of fairly small magnitude. Its cause is mostly due to bound water as discussed above. But partial rotation of macromolecular subunits probably also contributes.

The high-frequency limits ϵ_0 of the δ dispersion are not precisely known since the dispersion extends into the γ range. However, limit values ϵ_∞ of the δ dispersion are about as expected from the mixture equation if the bound water surrounding protein is of low permittivity compared with that of the nonbound water.

At radio frequencies, about 0.1 to 10 MHz, protein solutions display another relaxation (β). Its mechanism is most probably caused by the permanent dipole moment of the macromolecules as discussed earlier. An alternative model was proposed by Kirkwood and Shumaker (1952). They pointed out the importance of the role of mobile protons as the origin of the large dipole moment of protein molecules. Protons can migrate from one amide group to another when an electrical field is applied. The fluctuation of the proton distribution can produce mean square fluctuation of the dipole moment even if the mean permanent dipole moment of the protein is zero, namely,

$$\langle\mu^2\rangle=\langle\mu\rangle^2+\Delta\mu^2. \tag{22}$$

Here $\Delta\mu^2$ is the moment produced by the fluctuating protons and is given by

$$\Delta\mu^2=e^2f^2b^2\sum_\alpha\frac{n_\alpha}{2+K_\alpha/[\mathrm{H}^+]+[\mathrm{H}^+]/K_\alpha}, \tag{23}$$

where e, b, and n_α are elementary charge, radius of equivalent sphere, and number of αth charge. f is a shape factor and K_α is the dissociation constant of the αth group. This equation indicates that $\Delta\mu^2$ increases when the *p*H of the protein solution approaches the *pK* value of the αth group on the surface. Numerical calculations by the authors demonstrate that the dipole moment of proteins can be accounted for by the proton fluctuation mechanism. However, theoretical and experimental analyses (Scheider, 1965; Takashima, 1965) demonstrated that the proton fluctuation moment is small for proteins with a molecular weight below 10^5. Kirkwood *et al.* were, nevertheless, the first to point out the importance of the fluctuating proton distribution to generate a mean square dipole moment. This mechanism is now considered unimportant for small proteins because of their small values. However, the concept of nonvanishing mean square dipole moments due to the fluctuation of mobile surface ions played an important role in the development of counterion polarization theories for polyelectrolytes, as discussed earlier. Amino acids display a dielectric dispersion around a few hundred MHz. Because of this, the dispersion should be termed a δ dispersion. Their δ behavior is of polar origin, rather than the relaxation of bound water. Neither amino acids nor proteins appear to show any α effect, i.e., their permittivities are frequency independent in the low-frequency range.

4.3 Biological Membranes

Biological membranes consist of lipids and proteins. Attempts to reconstitute biological membranes using purified lipids and proteins have not met, so far, with much success, except for a few special cases. It is, however, possible to form pure artificial bilayer membranes. Electrical membrane properties can be measured using metal electrodes placed on both sides of an aperture on which the membrane is formed. The membrane capacitance of lipid bilayers is of the order of 0.4–0.6 μF/cm^2 and no frequency dependence has been observed (Hanai *et al.*, 1964). The conductances of lipid bilayer membranes are extremely small. Only when lipid bilayers are doped with certain peptide molecules, called ionophores, do membrane conductances increase manyfold and the membranes become permeable to ionic species.

The membrane admittance of nerve cells has been investigated using internal–external electrode arrangements. Nerve cells are long cylinders, and those of some sea animals have a large diameter of 300–700 μm. This permits insertion of a wire electrode (Pt wire with a diameter of 50–70 μm) axially inside the cell, and a larger electrode in the medium. The interpretation of the results is, however, difficult because of membrane instability, systematic errors due to poor contact between the membrane and electrodes (Takashima *et al.*, 1974), and parasitic capacitance due to fringe effects at the tip of internal electrodes (Schwan, 1981).

Biological membranes also contain some protein molecules which are not essential to the transmembrane ion flow but cannot be eliminated. In addition, time-dependent nonlinear ionic currents produce reactive components which mask the real admittance of the membrane: Sodium currents produce a capacitive reactance and potassium currents an inductive one. They are the cause of the membrane-induced α dispersion effects discussed above. Nerve membrane capacitances and conductances, free of these additional α effects, can be obtained if all ionic currents are blocked using pharmacological agents. With these precautions, the membrane capacitance of nerve cells was found to be slightly frequency dependent, decreasing from 1 to 0.8 μF/cm^2 in the frequency range between 100 Hz and 1 kHz, and independent of membrane potential. Only in the presence of ionic currents do nerve membrane capacitance and conductance become dependent on the membrane potential. The phenomenological admittance due to nonlinear ionic currents has been investigated by Fishman *et al.* (1977), using linear approximations as indicated in Sec. 3.3.

4.4 Cells

Cells are surrounded by membranes, which separate the cytoplasm from the medium surrounding the cell, i.e., interior phase. There exists an immense variety of biological cells. Prokaryotes are primitive cells which emerged much earlier than the highly differentiated eukaryotes. The cytoplasm of the more highly developed eukaryotes contains many smaller celllike structures including the cell nucleus, mitochondria, lysosomes, Golgi apparatus, ribosomes, and endoplasmic reticulum, each with a specialized function. The nucleus contains nucleic acids including DNA, carrying the genetic code. The cell nucleus and other "subcellular organelles" such as mitochondria are surrounded by membranes. Some membrane systems interconnect. The cytoplasm contains many proteins. Part of them are arranged in the cytoskeleton, a web of string-like structures which appear to provide some structural rigidity to the cell. It is linked to a similar membrane structure which connects part of the membrane proteins, while others have been shown to move under the influence of electrical fields. The glycocalyx, a loose brushlike structure fixed to the outer surface of the membrane, consists of glycoproteins and sugars. It is highly permeable and rich in fixed charges. Its thickness is about 100 Å, i.e., comparable to or larger than the Debye screening length for the medium outside the cell.

The electrical properties of cell suspensions have been studied by many investigators, particularly Fricke, Cole, Schwan, Carstensen, and Hanai and their colleagues, as reviewed by Cole (1968) and Foster and Schwan (1989). This work led to an understanding of the important role of cell membranes in the rf range. The principal mechanism responsible for the β dispersion is the accumulation of charges at membranes from extracellular and intracellular fluids (Maxwell–Wagner effect). In addition, cell suspensions display the above discussed relaxations, since they contain water and macromolecules. Furthermore, they often display strong dispersions at low frequencies (α dispersions). We summarize these cases in Table 2.

First, we consider suspensions of cells or vesicles without macromolecules in the fluids inside and outside the cell. Then Eqs. (12) and (13) are applicable. Assuming physiological conditions, these equations predict time constants corresponding to relaxation frequencies between 0.1 and 10 MHz for micron sized cells. Dielectric increments for volume fractions near 30% range in the thousands. Therefore suspensions display strong Maxwell–Wagner β effects in the rf range. Their magnitudes are considerably greater than that of the β effect of proteins. In addition, the cell suspensions display a γ effect since they contain water. The magnitude of the γ effect corresponds to the volume taken up by water.

Cells containing proteins (about 30% in the cytoplasm and 5%–10% in the medium outside the cells) display the β and γ effects listed above. In addition, the proteins add their own β effect and cause the δ effect due to their bound water. Relaxation frequencies of the β dispersion of proteins are somewhat higher than those responsible for that of cells, but of much smaller magnitude. This results in a slight lifting of the high-frequency tail of the β dispersion. Cells with surface charges and/or dispersive membranes display α relaxation effects caused by the counterion atmosphere and frequency-dependent membrane properties.

Many cells possess membrane systems which extend from the cell membrane or the nuclear membrane into the cytoplasm. These structures can act like transmission-line sections, causing frequency-dependent contributions to the membrane properties. Characteristic frequencies are near 100 Hz for the tubular system of muscle. Thus, apparent membrane properties change at low frequencies, causing an additional α dispersion effect (Schwan, 1954; Fatt, 1964).

Organelles such as mitochondria and cell nuclei have dielectric responses identical to those of cells since they are surrounded by membranes separating electrolytes. Because of their small sizes, their dispersions occur at higher frequencies and their magnitudes are smaller. Organelles inside cells contribute therefore to the tail of the β dispersion of cell suspensions, as may be predicted by Eqs. (12) and (13) (Irimajiri *et al.*, 1979; Stoy *et al.*, 1982). The contributions of subcellular organelles diminish as the frequency decreases below the characteristic frequency of the cell suspensions. This is due to the shielding effect imposed by the cell membranes at lower frequencies. A detailed theory of a "multistratified shell" model has been applied to the case of organelles or nuclei inside the cell (Irimajiri *et al.*, 1979).

5. TISSUE IMPEDANCE

5.1 Frequency Dependence

Biological tissues are complex combinations of concentrated cells with interconnections. They therefore display many of the characteristic features typical of concentrated cell suspensions. The mechanism responsible for the electrical properties of such suspensions are largely responsible for those of tissues. Tissues display the same three main dispersion regions α, β, γ, as discussed above for cell suspensions. However, magnitudes and relaxation frequencies may be different.

Tissues can be divided into various categories:

1. Soft tissues with high water content. Most body tissues belong to this group. High cell concentrations, electrolytes inside and outside the cells, and protein concentrations at about 20% are typical. The model of a highly concentrated cell suspension is particularly appropriate to this class of tissues.
2. Soft tissues with low water content. Fatty tissues belong to this group. These tissues also have cellular structure. Fat concentrations in the fluid compartments are high.
3. Hard tissues, which contain many minerals in solid form. However, even they display dielectric properties somewhat similar to those of other tissues.

Brain, in particular white matter, also contains fat. Its electrical properties place it somewhere between the soft tissues with high and low water content. The same is true for white bone marrow, while red bone marrow exhibits properties close to those of highly concentrated blood suspensions, as might be expected from its composition.

The electrical properties of tissues have been extensively investigated. A compilation of resistivity data for many tissues at low frequencies has been given by Geddes and Baker (1967). Data in tabular form and covering the total frequency range have been published by Schwan (1963). For a detailed compilation of these data and those obtained in recent years, as well as their interpretation, see Foster and Schwan (1986).

The conductivity is the dominant part of tissues' admittivity at low frequencies. For alternating current, the corresponding frequency range is below about 1 to 10 khz. Conductivity data are difficult to interpret because of the complexity of tissue structure. However, a simple analysis based on the mixture theory is fairly accurate. Assuming a typical fraction of the extracellular fluid near 10% and conductivity of 2 S/m, and effectively nonconducting cells, a dc conductivity of 0.14 S/m is predicted by the Maxwell mixture theory. This is typical for soft, high-water-content tissues at low frequencies.

The contribution of the α dispersion to the conductivity is small. As the frequency approaches the β dispersion, the conductivity beings to increase noticeably. At frequencies above the β dispersion region, the cell membranes are short circuited. Then the mixture theory can be applied with reasonable success, assuming the excluded volume to consist of poorly conducting proteins plus its water of hydration. In this range, the conductivity of most tissues is not strongly dependent on frequency. The conductivity increases again above 100 MHz. Contributions resulting from the Debye relaxation processes of water become significant compared to the ionic frequency-independent electrolyte contribution. Several other relaxation processes indicated in Table 2 affect the electrical data. These effects have been reviewed in greater detail by Foster and Schwan (1986).

The permittivity data are displayed for muscle tissue in Fig. 3. Outstanding are the extremely high values at low frequencies. Large permittivity values are generated by ionic diffusion processes, ion channels, and subcellular structures such as the tubular systems as discussed before (Schwan, 1981a). Application of the Schwarz theory (1962) to muscle cells, with a diameter near 100 μm, gives low-frequency permittivity values comparable to those observed experimentally, i.e., above 10^6 relative to free space. But the frequency range of the α dispersion of muscle does not fit theoretical prediction. An alternative theory was suggested by Falk and Fatt (1964). They proposed that the α dispersion is caused by the sarcotubular system which extends from the membrane into the cytoplasm. At low frequencies, the current penetrates deep into this system, thereby not only charging the cell membranes but also the membranes of the tubular system. At higher frequencies the tubular system's capacitance is no longer charged. This happens when the frequency becomes higher than the inverse of the time constant given by the system's capacitance and the access resistance. The Falk–Fatt theory is generally accepted. However, other tissues, such as liver and subcutaneous fat, display similar behavior at low frequency without a tubular system. Clearly, different mechanisms exist and remain to be sorted out (Schwan, 1981a).

Application of the theory of the β dispersion explains the radiofrequency data. Finally, the γ dispersion for soft tissues of high water content is consistent with the assumption that tissue water has properties comparable, if not identical, to those of normal water.

5.2 Special Topics

We discuss next several other relevant topics which have been investigated extensively.

5.2.1 Anisotropy. The anisotropy of muscular-tissue impedance, first observed at low frequencies by Hermann (1871), then at higher frequencies by Sapegno (1930) and by Epstein and Foster (1983), results from the structure of muscle, which consists of bundles of cells which are long compared to their diameter. "Longitudinal" resistance values are observed to be much lower than "transversal" ones at low frequencies. This is anticipated from relevant mixture theory. However, at high frequencies this structural anisotropy disappears, since membranes are then shortcircuited. Indeed, Epstein and Foster demonstrate that the ratio of electrical properties in and perpendicular to the cell direction diminishes as the frequency increases. Many other soft tissues, such as liver, kidney, and heart muscle, are structurally isotropic and, hence, their electrical properties are also.

5.2.2 Live and Excised Tissues. Most tissue properties were obtained using excised samples. Are they representative of "live" tissue samples? Studies were undertaken comparing *in situ* data with excised samples. Redistribution of blood volume appears to be a significant factor contributing to changes with excision. Air strongly affects lung data, rendering excised data almost worthless (Schwan and Kay, 1956). Changes in membrane properties occur slowly after death. Thus the β dispersion disappears after one or two days, as the cell membranes disintegrate after metabolic activity stops. Differences between excised and *in situ* tissue samples are most pronounced at low frequencies. Disintegration of membranes does not affect high-frequency data, since such data, obtained at frequencies above the β range, are not affected by membranes as outlined above.

5.2.3 Influence of Tissue Morphology. The complexity of tissues contributes to a broadening of relaxational response curves. This effect has been investigated for the β dispersion range by Stoy *et al.* (1982) investigating liver tissues. Hepatocytes, the abundant liver cells, cell organelles such as nuclei, mitochondria, endoplasmic reticulum, and the tissue proteins were considered, and their relative contributions to the electrical properties evaluated. The cellular contribution is the largest. The proteins and, to a lesser extent, the cell organelles explain the experimentally observed broadening of the high-frequency tail of the β dispersion.

5.2.4 Temperature Coefficients. The dominant charge carriers in tissues are electrolyte ions. The conductances of tissue electrolytes have a frequency-independent positive temperature coefficient of about 2% per °C and about −0.5% for the permittivity. In spite of this simple behavior, the temperature dependence of the electrical properties of tissues is complex. The reason for the complications is the temperature dependence of the various time constants responsible for the relaxation effects as indicated in Table 2. Positive and negative temperature coefficients exist in different frequency ranges. This behavior is fairly well understood. For a detailed analysis see Foster and Schwan (1986).

5.2.5 Diagnostic Applications. Unusual electrical properties have been frequently suggested as indicators of diagnostic value. We list here only a few:

Extracellular electrolyte volume is responsible for low-frequency conductance properties, while high-frequency data relate to total electrolyte, since membranes are shortcircuited. A comparison of both low- and high-frequency data might permit the determination of both extracellular and intracellular water (Settle *et al.*, 1980; Thomasset *et al.*, 1984). However, the technique is complicated because of other factors such as cell shape and uncertainty of appropriate mixture theory.

Tissue impedance plethysmography has been studied by many authors, beginning with Nyboer (1970). Some of the techniques measure minute tissue impedance changes which are synchronous with blood flow. The technique has not gained widespread recognition. Other applications are concerned with the changes in the impedance of the chest with breathing and heart activity (Geddes and Baker, 1975).

5.2.6 Nonlinear Properties. So far, all properties considered above pertain to low field-strength levels. However, as the field level exceeds a certain value, properties are affected. Threshold levels for such changes are in part due to the phenomenon of dielectric saturation, the nonlinear properties of membrane ion channels, the field-dependent contractile properties of muscle cells, and, at even higher *E* values, outright membrane destruction.

Muscle nonlinearity manifests itself in a change of low-frequency conductance near 1 V/cm in the medium surrounding the cells. Corresponding field strength values in the membranes surrounding muscle cells would be approximately 10 kV/cm. The low-frequency permittivity of protein solutions approaches dielectric saturation at about 10 kV/cm in the medium. The conductance of membrane channels changes at membrane voltages near 10 mV, corresponding to field strength values in the membrane of the order of 10 kV/cm. Corresponding field-strength values in the medium are about 1 V/cm. Apparently, threshold field-strength levels in the medium vary by four orders of magnitude, while corresponding values in the cell membrane are surprisingly similar. This illustrates that in all cases the onset of nonlinearity appears to be governed by dielectric saturation in the membrane constituents, presumably membrane proteins.

5.2.7 Brain Tissue Studies. Extensive efforts were undertaken to study the electrical properties of brain tissues during the 1960s and early 1970s. Agreement between electron microscopic determinations of extracellular space and that derived from impedance data was reported to be satisfactory by van Harreveld (1970). But Adey *et al.* report that impedance changes do not parallel changes in extracellular space (Adey *et al.*, 1969). The frequency dependence was investigated (Ranck, Jr., 1963). Noteworthy is the pronounced increase in cerebral cortex impedance during spreading depression (Leave and Ferreira, 1953; Marshall, 1959; Ochs, 1962). Activity in this field decreased after 1973. Important questions about the relative role of extracellular space, cellular conductances, their changes under abnormal conditions, and the effects of fluid and ion exchanges between inside and outside of the various cellular compartments remained unanswered.

6. FIELD EFFECTS

A variety of moderate to strong field effects on biological tissues are known.

6.1 Cell Manipulation Techniques

Cells can be manipulated by alternating electrical fields for a variety of purposes.

Pearl chain formation is the term often used to describe the alignment of biological cells or particles which has been observed in fields (Schwan and Sher, 1969). The effect is used in order to assure close cellular contact, a prerequisite to cell fusion.

Cell fusion is accomplished by application of a brief field pulse to two adjacent cells, strong enough to destroy membrane structure at the point of contact and yielding the formation of a larger cell. It is a new technique of potential value in biotechnology (Arnold and Zimmermann, 1984).

Cell rotation occurs in rotating fields and can be used to study the electrical properties of single cells over a broad frequency range (Arnold and Zimmermann, 1982).

Electroporation is achieved when a single cell is exposed to a brief strong field pulse. It opens briefly a part of the membrane, permitting the entry of extracellular material into the cell. The technique is useful in order to incorporate macromolecules such as nucleic acids into the cell, and is of biotechnological interest (Neumann, 1972; 1982).

Cell deformation occurs when cells are exposed to alternating electrical fields. The effect has been used to study the mechanical properties of cell membranes (Bryant and Wolf, 1987; Engelhardt and Sackmann, 1988; Chang *et al.*, 1985; Ashe *et al.*, 1989).

Cells and particles move in inhomogeneous electrical fields. This effect can also be used to determine the electrical properties of single cells (Burt *et al.*, 1990; Kaler and Jones, 1990).

Other reported effects include cell destruction, the ejection of small vesicles, and protoplasmic streaming.

The physical basis of these phenomena is well understood. They occur when related electrical potential energy changes are larger than kT. The needed threshold field strength is given approximately by the condition

$$\epsilon E_{\mathrm{th}}^2 R^3 = kT, \tag{24}$$

where ϵ is the permittivity of the medium, E_{th} is the threshold field strength, R is the cell or particle radius, k is the Boltzmann constant, and T is absolute temperature. For cellular size, theory predicts threshold field strength levels of the order of 1–100 V/cm in the medium surrounding the cell, in agreement with experimental data available so far (Schwan, 1985a, 1985b).

6.2 Nerve Conduction

The electrical properties of nerve cells change significantly when membrane potential changes of 10 mV or more are applied. These changes have been studied extensively and are summarized by a set of nonlinear equations formulated by Hodgkin and Huxley (1952). The changes were later recognized to take place at specific sites responsible for the control of ion exchange between media inside and outside of nerve cells. These ion channels selectively control the flux for individual ions such as sodium, potassium, and calcium. The detailed mechanism responsible for these changes in the conformation of ion channels has not been clarified. Field-strength level changes of the order of 10 kV/cm correspond to 10 mV across the membrane, which is only 10 nm thick. For such high local field-strength levels, the Langevin criterion and Eq. (17) are satisfied for molecules such as membrane channel proteins, and changes in conformation can be expected.

The propagation of an electrical signal of 10 mV or more along the nerve fiber can be analyzed using the same transmission-line equations which describe the propagation of an electrical signal along a cable. However, the nerve membrane parameters to be used are nonlinear and the theory becomes accordingly complex.

6.3 Muscle Contraction and Shock

Muscle contraction can be triggered by electrical pulses and low-frequency fields. Here again thresholds exist and are comparable to those which elicit nerve impulses. The mechanism of muscle contraction is less well understood than that causing nerve stimulation. It is generally believed that induced membrane potentials are transmitted to the cell interior, where muscular macromolecules then undergo conformational changes. Muscle contraction, as well as nerve stimulation are induced only at frequencies below approximately 10 kHz.

Shock is induced by external fields applied to the human body. This hazard has been extensively studied. Threshold levels have been established for sensation (a fingertip contact with a small electrode), muscular contraction (determination of let-go current), and fibrillation (dangerous heart muscle contractions). In all cases, corresponding current density levels in the tissues are about 1 mA/cm^2, field strength levels near 1 V/cm, and induced membrane potentials near 10 mV. Another type of shock occurs when a charged object is contacted—e.g., a doorknob after walking on a carpet, thereby charging the walker by friction. Fairly large discharges can occur when large objects such as trucks located in a high field, such as exist under high voltage transmission lines, are contacted. These discharges may be brief on short contact or continue if contact is maintained. Abundant data are available. See, e.g., Dalziel (1956) and Geddes (1971).

7. TECHNIQUES

Techniques for the determination of the dielectric properties of macromolecules, cells, and tissues fall into two classes:

1. Techniques for the study of single cells. They are limited to low frequencies, except for some novel techniques introduced recently.
2. Suspension techniques. Measurements integrate over many molecules, particles, or cells. They extend over a broad frequency range (dielectric spectroscopy), but require that one can extract particle or cell properties from measured bulk data using some mixture theory.

Table 4 summarizes the advantages and limitations of two experimental methods, i.e., dielectric spectroscopy and the cell rotation technique.

7.1 Suspension Techniques

The following types of instruments have been employed.

Table 4. Advantages (indicated by +) and limitations (−) of dielectric spectroscopy (S) and electrorotation (R). The rotation technique requires high field strength. This causes undue heating unless the medium's conductivity is reduced. The levitation and rotation technique share similar advantages and limitations.

	$\lvert R\rvert$	$\lvert S\rvert$
Single cell observation	+	−
Well explored		+
Frequency range		+
Suspension theory problem	+	−
Several dispersions		+
East of application	−	+
Ease of extraction of electr. param.		+
Temperature and medium limitations	−	+
Electrode polarization problem	+	−
α dispersion	−	+
Equipment availability	−	+

7.1.1 Bridges and Transmission Lines. A variety of bridges, such as the Wheatstone bridge, are commercially available, but of limited accuracy and resolution. Special-purpose bridges of very high resolution have been reported. They are able to resolve impedance magnitudes to better than 1 part in 10^6. With such resolution, permittivities of cell suspensions can be determined down to less than 100 Hz if the cells are suspended in a physiological medium, and to lower values if the cells are suspended in a poorly conducting medium. The upper frequency limit of this technology extends to about 100 MHz.

Transmission-line techniques and waveguides have been developed to cover the frequency range from some hundred MHz to about 100 GHz. Such techniques include the evaluation of the signal which is reflected from the sample to be investigated, or the comparison of two branches of a network, one containing the sample, and the other a calibrated phase shifter–attenuator combination. Accuracies of these techniques are better than 1% for permittivities and 1%–2% for conductivities, for electrolytes and biological samples.

7.1.2 Modern Commercial Equipment. Commercially available impedance analyzers provide a fast and convenient means to scan rapidly over broad frequency ranges. Models cover the total range from MHz to GHz. Resolution of impedance magnitude ranges from 10^{-4} to 10^{-5}, i.e., it is high, but lower than provided by the special research equipment mentioned above. Time-domain technology provides even more rapid spectroscopy over broad frequency ranges. A pulse is applied and the response subjected to a time domain–frequency domain transformation using appropriate Fourier algorithms. Commercial equipment is now available. However, accuracy is somewhat more limited than that of frequency-domain techniques. For more detailed surveys of instrumentation and evaluation techniques, see Grant *et al.* (1978) and Macdonald (1987).

7.1.3 Advantages and Limitations of Suspension Techniques.

7.1.3.1 Electrode Polarization. Low-frequency limits to dielectric spectroscopy are imposed by electrode polarization (Schwan, 1963; 1968). The impedance of electrodes increases with decreasing frequency and, since in series with the suspension, adds increasingly to the measured impedance as the frequency is decreased. This is demonstrated by the following approximate equations (Schwan, 1963):

$$C=C_s+\frac{G^2}{1+\omega^2R^2C_p}, \tag{25}$$

$$R=R_s+R_p+R_s(R\omega C)^2. \tag{26}$$

R and C are measured resistance and capacitance, R_s and C_s are those of the sample, and R_p and C_p are the components of the polarization impedance $Z_p=R_p-j/\omega C_p$. The equations are valid approximations if $R\omega C<1$ and $R_p<R_s$.[1] Electrode artifacts contribute to measured capacitance values below about 3 kHz when cells are suspended in a physiological media ($\sigma\cong 10$ mS/cm).[2] Large sample size, proper electrode treatment, and methods to correct for the electrode effect can reduce this limit about tenfold. Electrode contributions to measured data decrease rapidly when the medium conductivity is reduced. They can be

---

[1] These conditions can be shown to apply if the polarization impedance components are smaller than those of the sample resistance (see Schwan, 1963). This is usually the case.

[2] Assuming a 10% error, Eq. (24) reduces to $f^2C_pd\epsilon=3\times10^{12}\sigma^2$. For example, for a Pt electrode covered well with Pt black, C_p is about 1 mF/cm^2 at 1 kHz. Thus, for $d=5$ cm, $\sigma=0.01$ S/cm, and $\epsilon=1000$ relative to free space, the frequency is $f=2.4$ kHz.

eliminated entirely by use of four-electrode technology (Schwan, 1963; Schwan and Ferris, 1968). This technology uses different current- and potential-sensing electrodes. Care is required to exclude the potential-sensing electrodes from currents and to avoid capacitive contributions which may mimic relaxation phenomena. No good commercial equipment is yet available for this purpose.

7.1.3.2 Resolution. Permittivity determinations are limited by the high dielectric loss-tangent values, $\tan \delta = \sigma/\omega\epsilon$, typical of cell suspensions at low frequencies (Schwan, 1963). Resolution is determined by the principle $d\sigma = \omega d\epsilon$. For physiological media and typical dielectric constants of cell suspensions (ϵ near 2000), dielectric constants can be resolved to 10% at 1 kHz with equipment providing a resolution of 10^{-5} in impedance magnitude. Such resolution is only obtainable with a few commercial products. Special laboratory devices providing resolutions as low as 10^{-6} and 10^{-7} have been reported (Schwan *et al.*, 1962; Schwan, 1963; Hayakawa *et al.*, 1975). Combination of four-electrode technology with very high-resolution equipment extends dielectric spectroscopy of biological solutions and suspensions to MHz frequencies (Fig. 6).

7.1.3.3 Drift. Biological conductivities have a temperature coefficient of about 2%/°C. Conductivity resolutions near 10^{-5} therefore require temperature stabilization to 10^{-3} °C. This drift problem can be reduced by alternating measurements at various frequencies with those at a fixed reference frequency and applying corrections using the data obtained at the reference frequency as time advances (Schwan, 1963; Hayakawa *et al.*, 1975).

7.1.3.4 Mixture Theories. Maxwell's mixture theory is not rigorous since necessary assumptions include the need for low particle concentrations in order to exclude the interference of other particles in the proximity. However, for low particle or cell concentrations, the Maxwell theory is appropriate and agrees well with other mixture theories. How serious are the errors introduced by the simple Maxwell equation? We first compare the prediction of Eq. (13), derived from Maxwell's equation (8) for the permittivity ϵ_0 of a concentrated cell suspension, with that of the simple model indicated in Fig. 8. This model assumes a densely packed number of cells of cubical shape and size $2R$. N such cells are arranged in a unit volume between the electrodes. Thin membrane surfaces perpendicular to the electrodes do not contribute to the total capacitance, and only the $2N$ membranes parallel to the electrodes need be considered. Therefore, $\epsilon_0 = C_m/2N = RC_m$, since $2NR = 1$. This is identical with Eq. (13) for $p = 1$, indicating that the Maxwell prediction is good for high particle volume concentrations. Experimental experience supports this conclusion (Kell, 1985). The validity of Maxwell's mixture equation has been also tested for the special case of nonconducting particles in a conducting medium, and was found to be excellent up to the highest possible concentrations (Cole *et al.*, 1969). Nevertheless, it is uncertain if Maxwell's mixture equation is the best choice under different circumstances. More sophisticated theories avoid the assumption made by Maxwell. In particular, Hanai's equation (1968) has been successfully applied to suspensions of particles and cells and appears to be an extension of Bruggemann's (1935) equation to complex admittances.

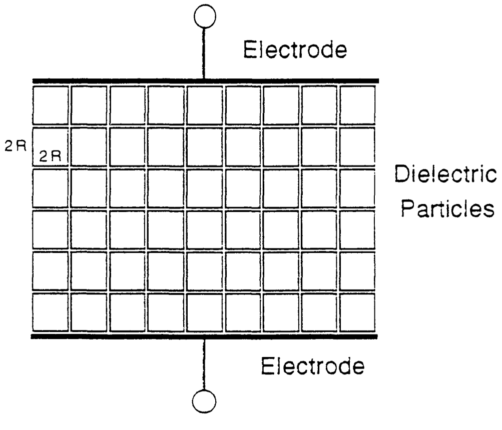

FIG. 8. A model of packed cubic dielectric particles. This model is used to prove the validity of Maxwell's mixture theory even at high volume fraction.

7.1.3.5 Membrane Conductance. Determinations of membrane conductances with the suspension technique are difficult. The reason is apparent from Eqs. (12) and (13): Typical G_m values vary between 1 and 1000 mS/cm^2, and therefore the second term in the denominator is small compared to 1 for typical cell size. Hence, the effect of the membrane conductance is usually weak and difficult to extract from measured data. This problem can only be circumvented if the second term in the denominator is made large by choice of large cells and medium resistivity. However, any increase in the resistivity of the medium may adversely affect cell properties.

7.2 Single-Cell Techniques

7.2.1 Microelectrode Technique. We discuss next the measurement of membrane capacitance and conductance of biological cells using methods which are more suitable to extract information about the membrane conductance. The membrane conductance is of greater interest than the membrane capacitance. The conductance is caused by channels and pores which regulate ion exchange and occupy only a small fraction of the membrane surface. Hence, it is functionally more important than the capacitance, which is mainly due to membrane bulk or the lipid layer.

7.2.1.1 Microelectrodes. There are two types of microelectrodes—those made of metal and of glass. Of particular importance is a glass microelectrode with a very fine tip of less than 1 μm. These electrodes are inserted into the cell interior, and transmembrane electrical properties such as the membrane potential are measured. Because of the very large tip impedance and various limitations thereof, i.e., limited frequency response and the propensity to be nonlinear, they are not suitable for the measurement of membrane admittances. In spite of these shortcomings, membrane capacitance and resistance can be measured using two such microelectrodes. With this technique, both electrodes are inserted into a cell. One of them is used for current injection, and the other for voltage recording. The membrane impedance can be obtained as a transfer function between voltage and current: $Z = V/I$. This technique was successfully used for the measurement of the transmembrane impedance of skeletal muscle between 1 Hz and 1 KHz by Falk *et al.* (1964).

7.2.1.2 Voltage Clamp Method. Originally, the voltage clamp technique was developed to investigate the current-voltage characteristics of squid axons (Cole, 1968). Using a negative-feedback control system and intracellular voltage sensing and current electrodes, the membrane potential of the nerve axon is clamped at a predetermined value, thus preventing the action potential from developing. Using this technique, a highly nonlinear current-voltage characteristic of nerve membranes was revealed (Hodgkin *et al.*, 1952). From the tangent of these current-voltage diagrams, we can calculate the conductance of nerve membranes. In addition, using the capacitive transients, the total capacitance of the membrane can be determined. A schematic diagram of the voltage clamp circuit is illustrated in Fig. 9. Finally, with the advent of the patch clamp technique (see BIOPHYSICS), it is possible to measure membrane properties of small spherical or ellipsoidal cells using a glass micropipet (Sakmann *et al.*, 1983). A glass micropipet is pressed against a small section of the cell wall with a gentle suction, in order to establish a tight seal with a very high resistance of more than 10^9 Ω. Application of a square pulse using the pipet, and subsequent current recording, permits determination of the membrane current-voltage characteristics. Of particular interest in this technique is its ability to investigate membrane "patches" and to measure single channel currents. Using a very small membrane patch, which is tightly sealed at the tip of the micropipet, spontaneous opening and closing of each channel can be recorded by monitoring small current pulses. Although channel opening and its duration are stochastic, the magnitude of single channel currents is usually of the order of 3–5 pA.

A variation of the patch clamp technique uses a glass micropipet tightly attached to the

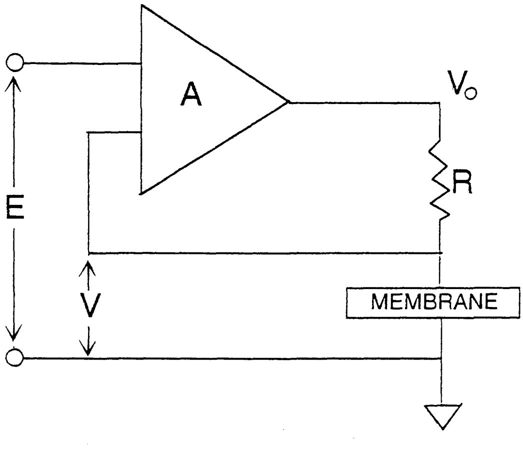

FIG. 9. A schematic diagram of the voltage clamp circuit. *E* is the command potential, and *V* is the membrane potential. *A* is the gain of differential amplifiers, and *R* is access impedance due to cytoplasm.

cell surface to measure the membrane admittance. Asami *et al.* (1990), using sinusoidal voltage inputs, successfully measured the membrane capacitance and conductance of small nearly spherical cells, such as myeloma and HeLa cells, between 1 and 1000 Hz. They found frequency-independent membrane properties $C_m \approx 0.8\ \mu\text{F/cm}^2$ and $G_m < 100\ \mu\text{S/cm}^2$.

The following single-cell technologies cover a much broader f spectrum extending from 100 Hz to more than 10 MHz. They share some disadvantages of dielectric spectroscopy, in particular the small effect of membrane conductance values on measured data, unless the second term in the denominator of Eq. (13) is made large by choice of large cells and medium resistivity.

7.2.2 Electrorotation. The rotation of particles and biological cells exposed to a rotating alternating field has been extensively investigated by Arnold and Zimmermann (Arnold, 1988; Arnold and Zimmermann, 1984, 1988; Zimmermann and Arnold, 1983) and by Glaser and Fuhr (1986). It peaks at the characteristics frequencies of the β dispersion, as may be predicted by application of the simple cell model characterized by Eqs. (9) to (13) (Schwan, 1985c). Fuhr *et al.* (1985, 1986) carried out more extensive calculations analogous to the Pauly–Schwan theory mentioned above. A rigorous discussion for arbitrary complex dielectric properties of particle and medium has been carried out by Sauer and Schloegl for cylinders and spheres (1985). The torque L which causes the particle to spin is given by $L = E\mu''$, with E field strength and μ'' the imaginary part of the induced dipole moment:

$$\mu = \mu' + j\mu'' = \epsilon_a R^3 E \frac{Y_p - Y_a}{Y_p + 2Y_a} = \epsilon_a R^3 E u, \qquad (27a)$$

where

$$u = \frac{Y_p - Y_a}{Y_p + 2Y_a}, \qquad (27b)$$

R is the particle radius, and the subscripts p and a indicate particle and outside medium. The Y's are admittivities and ϵ_a is the permittivity of the medium. If the suspending medium is an electrolyte, ϵ_a is frequency independent up to more than 1 GHz. Thus the frequency dependence of the rotation speed is directly proportional to the quantity u'', i.e., the imaginary part of u.

7.2.3 Levitation. If a particle is exposed to a divergent electrical field, it experiences a force given by

$$F = \epsilon_a u' \text{ grad } E^2, \qquad (28)$$

where u' is now the real part of the ratio u defined in Eq. (27b). u' can be either positive or negative. Thus, the force may be directed towards or away from the region of higher field intensity. This force may be used to balance gravitational forces, thus providing a null technique of potentially great accuracy. Stability criteria have been worked out, and the technique has been demonstrated to be useful for the investigation of particles and biological cells (Kaler and Jones, 1990). The relevant theory is very similar to the one for the rotational response. However, the frequency response of the gradient force is similar to the dispersion effect discussed for dielectric spectroscopy. For example, in the case of uniform particles, the force changes with frequency as expressed by the dispersion equations discussed above in Sec. 7.2. The magnitude of the dispersion is now

$$\Delta u' = \frac{3(\epsilon_a \sigma_i - \epsilon_i \sigma_a)}{(\sigma_i + 2\sigma_a)(\epsilon_i + 2\epsilon_a)} \qquad (29)$$

and, hence, can be either positive or negative. The time constant is identical to the one obtained for rotation and dielectric spectroscopy. In the case of shell-surrounded particles, two dispersions result with the same time constants as in the cases of rotation and spectroscopy.

Rotation and levitation techniques have much in common with dielectric spectroscopy. The same equations are used to obtain cellular data from the particle admittivity Y_p. In both cases u transformations must be used, as seen from the equations for torque L and force F. The difference is that u'' directly determines the rotational response, and u' that of levitation, while the transformation on the left-hand side of Eq. (8) needs to be made in addition, in order to relate suspension data to particle property. Thus the rotation and levitation techniques are one step more direct and not subject to the limitations of Maxwell's mixture equation (8). However, it is not possible to extract from either u' or u'' alone the dielectric properties ϵ and σ of the particle observed. Additional information must be provided. This can be done, e.g., by variation of the conductivity of the medium (Zimmermann and Arnold, 1983). The suspension technique provides both components of u from observed ϵ and κ data. It is also faster and easier to apply, and covers a much broader frequency range; and commercial instrumentation is available. The α dispersion observed with charged particles and data obtained from electrorotation do not agree (Arnold *et al.*, 1987). It appears that rotating and nonrotating fields interact with the counterion atmosphere in a different manner.

Table 4 attempts to summarize the advantages and limitations of various techniques.

7.2.4 Microspectroscopy. This technology appears to be just emerging as a byproduct of advances in the fabrication of very small electrode assemblies. Miniature electrode arrangements can be used to manipulate single cells and large macromolecules by electrical fields (Washizu and Kurosawa, 1990; Washizu *et al.*, 1990). This can be possibly done over a broad frequency range. Advantages combine those of suspension techniques with single-object assessment. Disadvantages include the complexity of operation necessitated by the small dimensions. Electrode polarization problems are much more pronounced with microelectrode arrangements unless four-electrode techniques are used or polarization impedances are reduced by use of high fields.

List of Works Cited

Adey, W. R., Bystrom, B. G., Costin, A., Kato, R. D., and Tarby, T. J., (1969) *Exptl. Neurol.* **23**, 29–50.

Allgen, L. G. (1950), *Acta Physiol. Scand. Suppl.* **22**, 32.

Arnold, W. M. (1988), *Ferroelectrics* **86**, 225–244.

Arnold, W. M., Zimmermann, U. (1982), *Z. Naturforsch. C* **37**, 908–915.

Arnold, W. M., Zimmermann, U. (1984), In: *Biological Membranes*, Vol. 5, D. Chapman (Ed.), New York: Academic.

Arnold, W. M., Zimmermann, U. (1984), In: *Biological Membranes*, Vol. 5, London: Academic, 389.

Arnold, W. M., Zimmermann, U. (1988), *J. Electrost.*, **21**, 151.

Arnold, W. M., Schwan, H. P., Zimmermann, U. (1987), *J. Phys. Chem.*, **91**, 5093–5098.

Asami, K., Takahashi, Y., Takashima, S. (1990), *Biophys. J.* **58**, 143–148.

Ashe, J. W., Bogen, D. K., Takashima, S. (1989), *Ferroelectrics* **86**, 311–324.

Barlow, D. J., Thornton, J. M., (1986), *Biopolymers*, **25**, 1717–1733.

Boettcher, C. J. F., (1952), *Theory of Electric Polarization*, Amsterdam: Elsevier.

Bryant, G., Wolfe, J. (1987), *J. Membrane Biol.* **96**, 129–139.

Bruggeman, D. A. G. (1935), *Ann. Phys. (Leipzig)* **24**, 636.

Burt, J. P. H., Pethig, R., Gascoyne, P. R. C., and Becker, F. F., (1990) *Biochim. Biophys. Acta*, **1034**, 93–101.

Chang, S., Takashima, S., Asakura, T. (1985), *J. Bioelectricity* **4** (2), 301–315.

Chew, W. C. (1984), *J. Chem. Phys.* **81**, 4541.

Cole, K. S., Cole, R. H. (1941), *J. Chem. Phys.* **9**, 341.

Cole, K. S., Li, C. L., and Bak, A. F. (1969), *Exper. Neurol.* **24**, 459–473.

Dalziel, C. F. (1972), *IRE Trans. Med. Electron.* **PGME-5**, 44–62.

Dalziel, C. F. (1972), IEEE Spectr. **Feb**, 41–50.

Debye, P. (1929), *Polar Molecules*, New York: Dover.

du Bois-Reymond, E. (1848–1860), *Untersuchungen ueber die Tierische Elektrizitaet*, Berlin: Reimer.

Engelhardt, H., Sackmann, E. (1988), *Biophys. J.* **54**, 495–508.

Epstein, B. R., Foster, K. R. (1983), *Med. Biol. Eng. Comput.* **21**, 51.

Falk, G., Fatt, P. (1964), *Proc. R. Soc. B* **160**, 69–123.

Fishman, H. M. (1985), in: A. Chiabrera, C. Nicolini, and H. P. Schwan (Eds.), *Interactions between Electromagnetic Fields and Cells*, NATO ASI Series A, Vol. 97, New York: Plenum, pp. 339–355.

Fishman, H. M., Poussart, D., Moore, L. E., Siebenga, E. (1977), *J. Membrane Biol.* **32**, 255–290.

Fixman, M. (1983), *J. Chem. Phys.* **78**, 1483.

Foster, K. R., Schwan, H. P. (1986), in: Ch. Polk and E. Postow (Eds.), *CRC Handbook of Biological Effects of Electromagnetic Fields*, Boca Raton, FL: CRC.

Foster, K. R., Schwan, H. P. (1989), *CRC Crit. Rev. Biomed. Eng.* **17:1**, 25–104.

Froehlich, H. (1949), *Theory of Dielectrics*, Oxford: Clarendon.

Fuhr, G. R. (1985), Dissertation, Humboldt-Universität, Berlin.

Fuhr, G., Glaser, R., Hagedorn, R. (1986), *Biophys. F.* **49**, 395–402.

Garcia, A., Barchini, R., Grosse, C. (1985), *J. Phys. D* **218**, 1891–1896.

Geddes, L. A., Baker, L. E. (1967), *Med. Biol. Eng.* **5**, 271.

Geddes, L. A., Baker, L. E. (1971), *J. Assoc. Adv. Med. Instrum.* **5**, 13–18.

Geddes, L. A., Baker, L. E. (1975), *Principles of Applied Biomedical Instrumentation*, New York: Wiley.

Glaser, R., Fuhr, G. (1986), In: *Electric Double Layers in Biology*, New York: Plenum.

Grant, E. H. (1965), *Ann. N. Y. Acad. Sci.* **125**, 418–442.

Grant, E. H., Sheppard, R. J., South, G. P. (1978), *Dielectric Behavior of Biological Molecules in Solution*, Oxford: Oxford University.

Grosse, C., Foster, K. R. (1987), *J. Phys. Chem.* **91**, 3073.

Haggis, G. H., Hasted, J. B., Buchanan, T. J. (1952), *J. Chem. Phys.* **20**, 1452.

Hanai, T., Haydon, D. A., Taylor, J. (1964), *Proc. R. Soc. A* **281**, 377–391.

Hanai, T. (1968), *Emulsion Science*, P. Sherman (Ed.), New York: Academic, Chap. 5.

Hasted, J. B., El-Sabeh, S. H. M. (1953), *Trans. Farad, Soc.* **49**, 1003–1011.

Hayakawa, R., Kanda, H., Sakamoto, M., Wada, Y. (1975), *Jpn. J. Appl. Phys.* **14**, 2039–2052.

Hazlewood, C. F. (Ed.) (1984), *Biophysical Correlates of Cellular Function (Unesco)*, Tee Bedner/Katherine Sturm, Evans Lithography.

Hermann, L. (1871), *Pfluegers Arch. Ges. Physiol.* **5**, 223–275.

Hille, B. (1971), *Biophysics and Physiology of Excitable Membranes*, W. J. Adelman (Ed.), New York: Van Nostrand and Reinhold.

Hodgkin, A. L., Huxley, A. F. (1952), *J. Physiol.* **117**, 500–544.

Hoeber, R. (1910), *Arch. Ges. Physiol.* **133**, 237–259.

Hoeber, R. (1912), *Arch. Ges. Physiol.* **148**, 189–221.

Irimajiri, A., Hanai, T. Inouye, A. (1979), *J. Theor. Biol.* **78**, 251–269.

Kaler, K. V. I. S., Jones, T. B. (1990), *Biophys. J.* **57**, 173–182.

Kell, D. (1985), Personal communication to H. P. Schwan.

Kirkwood, J. G. (1939), *J. Chem. Phys.* **7**, 911.

Kirkwood, J. G., Shumaker, J. B. (1952), *Proc. Natl. Acad. Sci.* **38**, 855–862.

Leav, A. A. A. P., Ferreira, M. (1953), *Anais Acad. Brasil. Cienc.* **25**, 259–266.

Mandel, M. (1961), *Mol. Phys.*, **4**, 489–496.

Marshall, W. H. (1959), *Physiol. Rev.* **39**, 239–279.

Maxwell, J. C. (1873), *Treatise on Electricity and Magnetism*, London: Oxford University Press.

Mcdonald, J. R. (Ed.) (1987), *Impedance Spectroscopy—Emphasizing Solid Materials and Systems*, New York: Wiley-Interscience.

McTague, J. P., Gibbs, J. H. (1966), *J. Chem. Phys.* **44**, 4295–4301.

Minakata, A., Imai, N., Oosawa, F. (1972), *Biopolymers* **11**, 347–359.

Neumann, E. (1972), *J. Membrane Biol.* **10**, 279–290.

Neumann, E. (1982), in: U. Zimmermann (Ed.), *Electric Field Effects in Macromolecules and Membranes*, Berlin: Springer, pp. 95–193.

Nyboer, J. (1970), *Electrical Impedance Plethysmography*, Springfield, IL: Charles C. Thomas.

O'Konski, C. T. (1981), in: S. Krause (Ed.), *Molecular Electro-Optics*, New York: Plenum, pp. 1–25.

Ochs, S. (1962), *Int. Rev. Neurobiol.* **4**, 1–69.

Oncley, J. L. (1943), In: *Proteins, Amino acids and Peptides as Ions and Dipolar Ions*, E. J. Cohn and J. T. Edsall (Eds.) New York: Reinhold.

Onsager, L. (1936), *J. Am. Chem. Soc.* **58**, 1486.

Oosawa, F. (1970), *Biopolymers* **9**, 671–688.

Pauling, L., Corey, R. B. (1951), *Proc. Natl. Acad. Sci. U.S.A.* **37**, 205, 729.

Pauly, H., Schwan, H. P. (1959), *Z. Naturforsch. B* **14**, 125.

Pethig, R. (1978), *Dielectric and Electronic Properties of Biological Materials*, New York: Wiley.

Pohl, H. A. (1978), *Dielectrophoresis*, London: Cambridge University Press.

Pople, J. A. (1951), *Proc. R. Soc. A.* **205**, 163.

Rajewsky, B. (1938), *Ergebnisse der Biophysikalischen Forschung*, Vol. 1, Leizig: Thieme.

Ranck, Jr., J. B. (1963), *Exptl. Neurol.* **7**, 144–152.

Sakmann, B., Neher, E. (1983), *Single Channel Recording*, New York: Plenum.

Sauer, F. A. (1983), in: H. Froehlich and F. Kremer (Eds.), *Coherent Excitations in Biological Systems*, New York: Springer.

Sauer, F. A., Schloegl, R. W. (1985), in: A. Chiabrera, C. Nicolini, and H. P. Schwan (Eds.), *Interactions between Electromagnetic Fields and Cells*, NATO ASI Series A, Vol. 97, New York: Plenum, pp. 203–251.

Scheider, W. (1965), *Biophys. J.* **5**, 617–628.

Schlecht, P. (1969), *Biopolymers* **8**, 757–765.

Schwan, H. P. (1954), *Z. Naturforsch. B* **9**, 245–251.

Schwan, H. P. (1957), in: J. H. Lawrence and C. A. Tobias (Eds.), *Advances in Biological and Medical Physics*, Vol. 5, New York: Academic, pp. 147–209.

Schwan, H. P. (1963), *Biophysik.* **1**, 198.

Schwan, H. P. (1965), In: *Forms of Water in Biologic Systems, Ann. N. Y. Acad. Sci.* **125**, 344–354.

Schwan, H. P. (1968), *Ann. N. Y. Acad. Sci.* **148**, 191–209.

Schwan, H. P. (1981), in: W. S. Adelman and D. Goldman (Eds.), *Biophysical Approach to Excitable*

Systems, New York: Plenum.

Schwan, H. P. (1983), *Blut* **46**, 185–197.

Schwan, H. P. (1985a), in: M. Grandolfo, S. M. Michaelson, and A. Rindi (Eds.), *Biological Effects and Dosimetry of Static and Electromagnetic Fields*, Vol. 19, New York: Plenum, pp. 243–271.

Schwan, H. P. (1985b), in: A. Chiabrera, C. Nicolini, and H. P. Schwan (Eds.), *Interactions Between Electromagnetic Fields and Cells*, NATO ASI Series A, Vol. 97, New York: Plenum, pp. 371–389.

Schwan, H. P. (1985c), *Studia Biophysica*, Akademie-Verlag.

Schwan, H. P., and Ferris, C. D. (1968), *Rev. Sci. Instrum.* **39**, 481–485.

Schwan, H. P., Kay, C. F. (1956), *Circ. Res.* **4**, 664.

Schwan, H. P., Sher, L. D. (1969), *J. Electrochem. Soc.* 116–170.

Schwan, H. P., Schwarz, G., Maczuk, J. Pauly, H. (1962), *J. Phys. Chem.* **66**, 2626–2635.

Schwarz, G. (1962), *J. Phys. Chem.* **66**, 2636–2642.

South, G. P., Grant, E. H. (1972), *Proc. R. Soc. London A* **328**, 371–387.

Settle, R. G., Foster, K. R., Epstein, B. R., Mullin, J. L. (1980), *Nutrition and Cancer* **2**, 72–79.

Sapegno, E. (1930), *Pfluegers Arch. Ges. Physiol.* **224**, 187–211.

Stoy, R. D., Foster, K. R., Schwan, H. P. (1982), *Phys. Med. Biol.* **27**, 501.

Takashima, S. (1965), *J. Phys. Chem.* **69**, 2281–2286.

Takashima, S. (1966), *J. Phys. Chem.* **70**, 1372–1380.

Takashima, S. (1989), In: *Electrical Properties of Biopolymers and Membranes*, Bristol, U.K.: Adam Hilger, Chap. 7.

Takashima, S., Schwan, H. P. (1974), *J. Membrane Biol.* **17**, 51–68.

Takashima, S., Yantorno, R. (1977), *Ann. N. Y. Acad. Sci.* **303**, 306–321.

Thomasset, A. L., Roullet, C., Lenoir, J. (1984), In: *Marker Protein Inflammation*, Vol. 2, New York: Walter de Gruyter.

Van Harreveld, A., Steiner, J. (1970), *Anat. Rev.* **166**, 117–129.

Wada, A. (1962), in: M. A. Stahman (Ed.), *Polyamino Acids, Polypeptides and Proteins*, Madison, WI.

Wagner, K. W. (1914), *Arch. Elektrotech.* **2**, 371; **3**, 67.

Wagner, K. W. (1924), In: *Die Isolierstoffe der Elektrotechnik*, Berlin: Springer.

Warashina, A., Minakata, A. (1973), *J. Chem. Phys.* **58**, 4743–4749.

Washizu, M., Kurosawa, O. (1990), *IEEE Trans. Ind. Appl.* **26**, 1165–1172.

Washizu, M., Nanba, T., Masuda, S. (1990), *IEEE Trans.* **1**, 25.

Whipple, H. E. (1965) (Ed.), *Forms of Water in Biological Systems, Ann. N. Y. Acad. Sci.* **125**.

Wyman, J., Jr. (1936), *Chem. Rev.* **19**, 213.

Zimmermann, U., and Arnold, W. M. (1983), in: H. Fröhlich and F. Kremer (Eds.), *Coherent Excitations in Biological Systems*, Berlin: Springer, pp. 211–221.

Further Reading

Boettcher, C. J. F., Bordewijk, P. (1987) *Theory of Electric Polarization*, 2nd Ed. Amsterdam: Elsevier.

Schwan, H. P. (1957), in: J. H. Lawrence and C. A. Tobias (Eds.), *Advances in Biology and Medical Physics*, Vol. 5, New York: Academic.

Schwan, H. P. (1963), In: *Physical Techniques in Biological Research*, Vol. VI, W. L. Nastuk (Ed.), New York: Academic.

Cole, K. S. (1968), *Membranes, Ions and Impulses*, Berkeley: University of California.

Grant, E. H., Sheppard, R. J., South, G. P. (1978), *Dielectric Behavior of Biological Molecules in Solution*, Oxford, U.K.: Oxford Univ. Press.

Hasted, J. B. (1973), *Aqueous Dielectrics*, London: Chapman and Hall.

Pethig, R. (1979), *Dielectric and Electronic Properties of Biological Materials*, New York: Wiley.

Takashima, S. (1989), *Electrical Properties of Biopolymers and Membranes*, Bristol, U.K: Adam Hilger.

ELECTRICAL INSTRUMENTATION

James O. Hylton, Allen V. Blalock, and William H. Andrews, *Instrumentation and Controls Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A.*

	Introduction	1
1.	**Basic Electrical Units**	2
2.	**Measuring Electrical Quantities**	2
2.1	General Principles	3
2.1.1	Electrical Voltage and Current Levels	3
2.1.2	Instrument Components	3
2.1.3	Instrument Resolution	4
2.2	Charge Measurement	4
2.3	Voltage Measurement	5
2.3.1	dc Measurements	5
2.3.2	ac Measurements	5
2.4	Current Measurement	6
2.4.1	Resistance Shunts	7
2.4.2	Hall-Effect Transducers	7
2.4.3	Current Transformers	8
2.4.4	Active Shunts	8
2.5	ac Impedance Measurement	8
2.5.1	Bridge Measurements	9
2.5.1.1	dc Bridges	9
2.5.1.2	ac Bridges	10
2.5.2	Active Guard Circuits	11
2.5.3	Electronic Impedance Instruments	12
2.6	Electrical Power Measurement	12
2.6.1	Electrodynamometer Power Measurements	12
2.6.2	Watt-Hour Meters	13
2.6.3	Electronic Power Measuring Instruments	13
3.	**Instrument Interfaces**	13
3.1	Instrument-to-Human Interfaces	13
3.1.1	Visual Display	13
3.1.2	Audio Signaling	14
3.2	Human-to-Instrument Interfaces	14
3.3	Instrument-to-Equipment Interfaces	14
3.3.1	Multiplexing	15
3.3.2	Analog-to-Digital Conversion	15
3.3.3	Digital-to-Analog Conversion	16
3.4	Recording	17
3.4.1	Analog Recorders	17
3.4.2	Digital Recorders	18
4.	**Noise and Other Sources of Error**	18
4.1	Induced Noise	18
4.2	Inherent Noise	19
5.	**Conclusion**	20
	Glossary	20
	Works Cited	21
	Further Reading	21

INTRODUCTION

An instrument is a device that is used in observing, measuring, or controlling a physical parameter or event. Almost every instrument (or instrument system) used today involves the detection of electrical signals at some point in the process. The parameter being measured produces an observable effect in a sensor or transducer that generates a corresponding electrical signal. The electrical signal is then amplified or "conditioned" and transmitted to an electronic instrument for measurement. This article will focus upon the various kinds of electronic instruments commonly used for signal measurement. The transducer and the signal conditioner are important parts of the measurement but are outside the scope of this discussion.

It is not necessary for the transducer to produce an electrical signal to make the desired measurement. A familiar example is the thermometer, which uses the expansion of mercury in a glass capillary as a measure of temperature. Although the mercury thermometer is an excellent instrument, its use has

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

greatly decreased in recent years because of the availability of low-cost electronic thermometers. This is also true of most other kinds of nonelectrical transducers. The convenience and versatility offered by modern electronics, especially with microprocessors and digital technology, have resulted in a drastic increase in the use of electronic-based sensors and instruments.

The instruments and techniques for making accurate measurements of electrical quantities have been common laboratory tools for several decades. Advances in modern electronics, incorporating large-scale integrated circuitry and embedded microprocessors, have not drastically improved the ultimate capabilities of making precise electrical measurements, but they have created enormous changes in the cost, convenience, and availability of precision electrical measurement devices. Many of the time-consuming and carefully controlled adjustments, which used to be done by the operator, are now accomplished by internal microprocessors. More compact electronics and integrated circuits have resulted in much smaller instruments that can perform multiple functions, thus replacing several of the older devices. Digital recording and display devices have greatly extended the available options for storage and analysis of electrical signals.

Basic methods for measuring various electrical quantities and properties will be discussed, followed by a review of common laboratory instruments for display, storage, and analysis of electrical signals.

1. BASIC ELECTRICAL UNITS

The fundamental Système International d'Unites (SI) electrical unit is the ampere, which is defined as the current in two straight, parallel, infinitely thin and infinitely long conductors, separated by 1 m, which produces a force per unit length of 2×10^{-7} N/m. This force is related to the current by the expression

$$\frac{F}{L}=\mu_0\frac{I_1I_2}{2\pi D}, \tag{1}$$

where μ_0 is the magnetic permeability of free space, I_1 and I_2 are the currents in the conductors, and D is the separation. It can be seen that the SI definition is equivalent to setting the value of μ_0 equal to $4\pi\times10^{-7}$.

The SI volt and ohm are derived from the ampere. The volt is the potential drop that produces a power dissipation of 1 W at a current of 1 A. The ohm is the resistance that produces a potential drop of 1 V at a current of 1 A. These definitions are sufficient to establish the values of the electrical units, but it has not been possible to devise a real experimental apparatus to establish them with sufficient precision to serve as measurement standards. For this reason the volt and the ohm are represented independently by the Josephson effect and the quantum Hall effect (Taylor, 1987).

The Josephson effect allows the voltage drop V in a system of superconductors and oxide layers to be related to the frequency f of a microwave excitation source by a quantity that depends only upon fundamental physical constants. The relationship is shown by

$$V_h=\frac{h}{2e}f=\frac{f}{k_j}, \tag{2}$$

where h is Planck's constant, e is the electronic charge, and K_j is the Josephson constant (483 597.9 GHz/V).

The quantum Hall effect allows the Hall voltage observed in certain metal oxide semiconductor junctions at cryogenic temperatures (below 2 K) to be related to the current by a quantity that also depends only upon Planck's constant (h) and the electronic charge (e). The relationship is shown by

$$V=\left(\frac{h}{e^2}\right)\left(\frac{I}{n}\right)=R_h\left(\frac{I}{n}\right). \tag{3}$$

R_h is the von Klitzing constant (25 812.817 Ω) and n is a small integer.

2. MEASURING ELECTRICAL QUANTITIES

The advent of electronic-based measuring instruments has greatly enhanced our ability to measure without perturbing the medium being measured. The energy drawn from the system being measured by an electronic instrument is typically 1000-fold less than that required by electromechanical instruments. Such instruments were usually based on the d'Arsonval galvanometer (Sydenham, 1983) and required 50 μA (or more) full-scale operating current. The d'Arsonval galvanometer consists of a current-carrying coil suspended in a magnetic field by a torsion spring that opposes the torque produced by the current. The coil assembly is attached to a pointer that indicates the level of current on a scale. Electronic versions of such instruments can be designed to draw no more than a few nanoamperes (digital multimeters) and as little as 1 fA (electrometers).

2.1 General Principles

2.1.1 Electrical Voltage and Current Levels. Voltage and current signals can convey information by any combination of attributes such as amplitude, repetitive rate, time waveform, phase angle, and power spectrum. These signals can be classified as having either alternating-current (ac) or direct-current (dc) forms or a combination of both. These terms were originally associated with electrical wave forms produced by particular types of electrical power-generation equipment. In more recent times, these terms have taken on a broader connotation as their use persisted within rapidly expanding electronic technologies. Now, these terms are often used to classify electrical levels totally unrelated to power generation. In the world of electronic instruments, one usually thinks of a dc level as having an instantaneous value that can change significantly only at rates slow enough for direct observation by a human. Pure ac levels have instantaneous values that change at much faster rates, or with frequencies greater than several cycles per second (hertz), and with an average value equal to zero. An ac level is usually specified by a characteristic value obtained by squaring its time waveform point by point, taking the average of the resulting all-positive wave form, and taking the square root of the resulting value. This characteristic value is called the root mean square or rms value. If the wave form is sinusoidal, other representations such as amplitude or peak-to-peak are sometimes used; however, rms is generally preferred because of its application to power measurements as discussed in Sec. 2.6.

2.1.2 Instrument Components. Regardless of the particular electrical quantity to be measured, most modern electronic measurement instruments have several functional blocks in common. In the most modern designs, usually microprocessor-based digital instruments, the front-end or input circuit consists of some kind of excitation and/or linear preamplifier, the particular arrangement depending upon whether the source of the electrical parameter to be measured is active or passive. An active source is capable of imparting power to the measuring instrument. A passive source can only dissipate power and, therefore, requires electrical excitation from the instrument to produce a measurement. A linear preamplifier is usually required for converting the charge, current, voltage, impedance, or whatever is to be measured to a proportional, scaled internal signal level, usually voltage. To maximize the resolution of the measurement, the full dynamic range of the instrument should be used. This is accomplished by matching the maximum preamplifier output to the full input range of the next stage, typically an analog-to-digital conversion (ADC) circuit.

The ADC, in turn, supplies digitized measurement data to some combination of external communication interface, data storage device, and data display device depending upon the specific application for which the instrument is intended. In the case of an oscilloscope or strip chart recorder, the ADC result is periodically obtained and keyed to time when stored for later production of an amplitude-versus-time display of the measurement. Rapid multiplexing between the taking of data and production of displays is used to simulate real-time instrument operation. Common external communication interfaces are digital single-bit or multibit serial data ports conforming to industry standards such as RS-232, RS-422, and GPIB. Data storage devices such as any form of integrated-circuit (IC) random-access memory (RAM), disk drives, and tape drives are often used. Display devices range from multidigit numeric displays for digital multimeters (DMMs) to cathode ray tubes (CRTs) for oscilloscopes. Even in the most economical digital instruments, the operations of these functional blocks as well as its ADC are usually controlled by a microprocessor.

2.1.3 Instrument Resolution. All ADC circuits function to produce a binary code that represents an input signal level. The form of code most frequently used is a binary-base numeric representation because of the simple bistate (high or low) possibility for each digit of the code. Many different types of ADC circuits are in use today. Dual-slope integrator or successive-approximation types are examples of conversion techniques that might be utilized for high resolution but usually are not suitable for the fastest signal measurement applications requiring millions or more of conversions per second. Digital oscilloscopes are available with flash ADC circuits that can sample electrical signals at rates exceeding a billion samples per second, but at the cost of reduced resolution. The operation of ADCs is discussed further in Sec. 3.3.2.

Resolution requirements, specified by the user and depending upon the particular application, determine the number of binary bits the ADC must use to code the voltage level from the preamplifier. Resolution is defined as the smallest change in voltage that can be detected by the measurement system. For example, an 8-bit binary code gives a maximum possible resolution of 1 part in 255, or 0.4% of the full-scale range. In general, an N-bit binary code gives a maximum number of increments within the ADC range given by

$$2^N - 1. \tag{4}$$

The resolution in volts for a given range is then this number of steps divided into the largest magnitude voltage or full-scale range.

The nature and purpose of a measurement determine the resolution requirements of a digital conversion system to be used for displaying or recording data. For example, a temperature detection instrument with a 0–5-V full-scale output corresponding to a 0–100 °C temperature detection range requires at least a 10-bit ADC if the temperature is to be known to within 0.1 °C. On the other hand, a switch used to detect the closure of a cover might be operated so that 5 V is connected to the input of a measurement system only when the cover is secured. The measurement system will receive either 5 or 0 V in this case for the cover closed or open, respectively. Such a measurement system need have only a 1-bit digital conversion code to utilize fully the available signal. Individual signal inputs within any electronic binary logic circuit can be thought of as inputs of 1-bit ADCs.

2.2 Charge Measurement

One coulomb of electrical charge is the quantity of electricity transported in 1 s by 1 A of electrical current. Charge is typically measured by an electrical circuit with an integrating mathematical characteristic.

Electronic instruments specifically for measuring electrical charge are referred to as electrometers. Conceptually, such an instrument consists of an extremely high-input-impedance electronic differential amplifier with a capacitive negative-feedback path and grounded positive-input terminal as shown in Fig. 1.

The capacitance of the feedback path determines a constant of proportionality between the amplifier output voltage and the net charge that has flowed from the amplifier negative-input node during the time since the amplifier output voltage was last zero. The high differential gain of the amplifier forces nearly all of the measured charge to be stored by the feedback capacitor, giving rise to a potential across the capacitor, which appears at the amplifier output. Under ideal circumstances, the voltage output is determined by the familiar equation (5), where Q is the measured charge in coulombs, V is the amplifier output in volts, and C is the feedback capacitance in farads:

$$V = -Q/C. \tag{5}$$

A measurement with such a circuit will be accurate as long as several conditions are

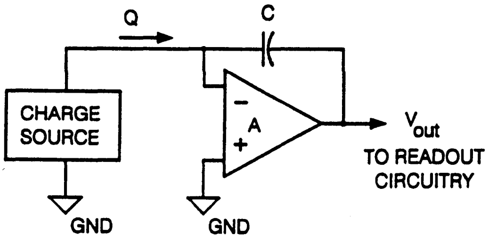

FIG. 1. Typical electrometer (charge amplifier) configuration; the large gain A of the amplifier forces any charge Q entering the input node to be stored on the capacitor C, forcing the magnitude of V_{out} equal to Q/C.

satisfied. First, the charge to be measured must be much greater in magnitude than stray charge introduced by amplifier input bias current and feedback capacitor leakage current. Second, the amplifier output must at all times during the measurement interval remain within its linear range. Third, the variations with time in the measured charge flow rate must not be faster than the gain–bandwidth product (small output signal) or slew rate (large output signal) of the operational amplifier. Otherwise, the amplifier could not "keep up" and would become nonlinear in its response.

2.3 Voltage Measurement

One volt is classically defined as the difference in potential between two points along a conductor with 1 A of electrical current flowing through the conductor and 1 W of power dissipated between the two points. Contrary to this basic definition, a voltage measurement usually does not involve a power measurement except for some types of true-rms instruments described later.

2.3.1 dc Measurements. Numerous methods of measuring voltage are in use today for applications ranging from the precise measurement of millivolt analog sensor outputs to measurement of high-voltage ac power bus levels to bistate digital level sensing. The most modern methods nearly always involve a high-input-impedance preamplifier to sense the voltage without loading the source of the voltage; i.e., without drawing significant current from the source. Typical input impedance values for modern digital voltmeters are over 10 MΩ.

Electronic voltage measurements can be classified into one of two broad categories: analog and digital. Both analog and digital meters require an analog preamplifier at the input for sensing the voltage to be measured. In the case of the analog meter, it is the presence of the preamplifier that has modernized the meter relative to purely electromechanical analog meters and placed new analog meters into the class of electronic instruments.

For analog measurement and display of voltage levels, the preamplifier referred to above is used to sense the voltage and produce a proportional analog current signal suitable for driving an analog display such as the galvanometric needle movement of a d'Arsonval meter. This type of measurement representation is classified as analog because of its continuous indication throughout the measurement range. The use of analog meters has declined in recent years because of readily available digital voltmeters that have greater accuracy, lower power consumption, and easier reading, all at comparable cost. The advantages and disadvantages of analog versus digital instruments are discussed further in Sec. 3.4.

For digital (numeric) display of voltage levels, the front-end circuitry again comprises a linear preamplifier as in the analog case, but now is used to drive an ADC circuit instead of the galvanometric movement (see Sec. 3.3.2 for information on ADCs). The measurement is then displayed from the ADC output code by appropriate logic circuitry driving a display device such as a liquid-crystal display (LCD) multidigit numeric display. This type of measurement is classified as digital because of its discrete nature, as only a finite number of readings are possible.

Measurement ranging is typically accomplished with a resistive voltage divider ahead of the preamplifier for high-level (greater than 1 V) dc measurements; with a divider and/or transformer for high-level ac voltage measurements; and with variable preamplifier gain or sensitivity. When a resistive divider is used, the total resistance of the divider circuit is maximized, again, to minimize measurement error due to loading effects upon the voltage source. Resistance values as large as the low-megohm range typically can be used without introducing other errors due to preamplifier imperfections such as bias currents, noise, and dc drift.

2.3.2 ac Measurements. The measurement of ac (rms) voltage pertains to the measurement of a characteristic value that is proportional to the root of the power in a signal. The rms value of a signal is quite different from the direct average value of the signal, although both determine power dissipated when the signal is applied to a resistive load. The direct average value is the dc component of a signal and is readily obtained by an averaging low-pass filter, a rather simple operation to implement electronically. The true rms value is much more difficult to obtain electronically because of the required nonlinear mathematical operations as well as the need for high electronic speed. Great speed is needed in certain true rms measurements because many signals of interest have an energy spectrum that extends into the megahertz range.

One of the best techniques for measurement of the true rms of an arbitrary periodic wave form with high-frequency spectral components is based upon establishing a balance between two temperatures, one due to the heating power of the signal under test and another due to the heating power of a dc test signal generated by the instrument. This instrument is called a thermal rms voltmeter and is illustrated in Fig. 2.

The thermal meter illustrated in Fig. 2 operates as follows. The input side of the balance circuit consists of an amplifier that provides drive current as required to reproduce precisely the signal under test across a heater element (resistive load) in direct thermal contact with a thermocouple. The output of this, the test thermocouple, is compared to a second, identical thermocouple and heater element. A balance control circuit produces a dc signal that is applied to the second heater element at whatever amplitude is required to make the second thermocouple output equal to the first. Under this balanced condition, the dc control signal applied to the second heater element is equivalent to the rms of the signal under test as long as several conditions are satisfied. First, the amplifier of the signal under test must exactly reproduce the wave form of the test signal across the first heater element. This means that the amplifier's bandwidth and slew rate must not be exceeded by the input signal, thus placing constraints upon the signal's crest factor and spectral extent. Also, the two heater element and thermocouple assemblies must share the same thermal ambient conditions, yet be thermally isolated from each other and have power responses that match very closely. Accurate (5% of full scale) rms measurements can be made of signals with frequency range to 10 MHz, full-scale amplitude range from 1 mV to 300 V, and crest factor (ratio of signal peak to rms values) of 10 to 1 with commercially available instruments. Thermal rms calibration standards for laboratory use are capable of accuracies on the order of a few parts per million.

2.4 Current Measurement

Electrical current is quantified in engineering units of amperes. One ampere is defined classically as the quantity of electrical current that, if flowing in two infinitely long parallel wires spaced 1 m apart in a vacuum, would produce a force of 2×10^{-7} N/m of length between the wires. Just as voltage measurement usually does not involve a power measurement, current measurement by electronic instruments typically does not involve a force measurement.

Most electronic current measurement techniques involve various operations that result in the production of a voltage signal proportional to the current signal to be measured. The resulting voltage signal is measured and scaled according to the current-to-voltage constant of proportionality of the conversion operation so that a reading calibrated in engineering units of amperes is obtained. These voltage signals can be measured by the techniques discussed in Sec. 2.3.

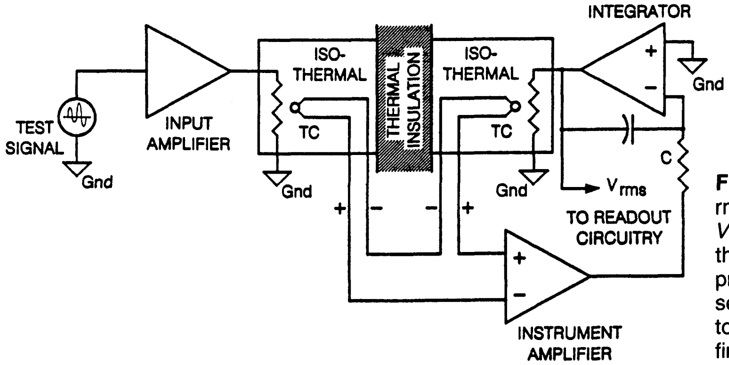

FIG. 2. Example of typical thermal rms converter; the integrator output V_{rms} is a dc voltage necessary to zero the instrument amplifier output, thus producing a heating power in the second isothermal box exactly equal to the heating power produced in the first isothermal box by the test signal.

The particular method of current-to-voltage conversion used depends upon the form of the current signal to be measured. If the current signal is ac (see Sec. 2.1.1), then linear transducers such as a resistive shunt, Hall device, or current transformer may be employed to detect the current and produce a proportional, similarly shaped voltage wave form so long as the power spectral extent of the ac current to be measured does not exceed the frequency response of the transducer. dc current signals (see Sec. 2.1.1) can be detected only by shunt or Hall-effect transducers since the voltage output of the current transformer is proportional to the derivative of the current signal.

2.4.1 Resistance Shunts. A shunt transducer is simply a resistance with a very stable and well-known value, usually very small (< 1 Ω). The shunt must be serially inserted into the circuit within which the current to be measured flows. The fact that the circuit must be broken when connecting or disconnecting the shunt is a potential disadvantage of the shunt transducer. The resistance of the shunt must also be small enough not to affect the circuit operation. When current flows through the shunt, a voltage is developed across the shunt. This voltage is usually very small because of the small resistance of the shunt. This voltage is then measured by one of the systems described in Sec. 2.3.

Note that, depending upon where within the circuit the shunt is placed, large voltages may be present relative to some reference point such as earth ground. Therefore floating instruments such as battery-powered digital multimeters are typically used with the shunt to avoid inadvertently affecting circuit current being measured by the measurement itself. Another precaution is that the current in the shunt must not be so large that the power dissipation produces thermal gradients or temperature rises that affect the shunt resistance.

2.4.2 Hall-Effect Transducers. The Hall-effect transducer operates by detecting the magnetic field produced by and proportional to the current to be measured. A Hall transducer, illustrated in Fig. 3, is a semiconductor with two pairs of plates electrically connected (resistive path) to surfaces of the semiconductor block. The plates are positioned such that a drift current due to a bias voltage between

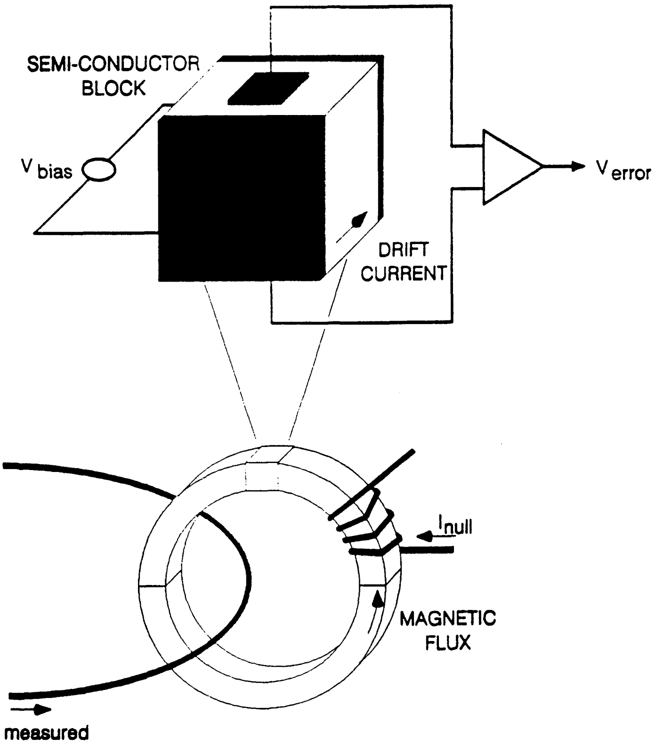

FIG. 3. Hall current transducer in a feedback control system; I_{null} is electronically controlled such that V_{error} is kept equal to zero, forcing I_{null} to be proportional to $I_{measured}$.

one pair flows orthogonally between the other pair. With no applied magnetic field, there will be no change in charge density between the second pair of plates due to the drift current. When a magnetic field resulting from proximal current flow is applied, the path of the drift current is altered, resulting in a charge density gradient and an associated voltage drop between the second pair of plates. This differential voltage can be used in a feedback system with a magnetic toroid encircling the conductor carrying the applied field current to create another magnetic field within the Hall device equal and opposite to the applied field so that a signal proportional to only the current creating the applied magnetic field is obtained.

2.4.3 Current Transformers. A current transformer is also based on magnetic field theory but is a passive device requiring no applied potential to operate. The current transformer consists of a magnetic material formed usually in a toroid shape with insulated wire wound to form a number of turns around the toroid body. This winding serves as the secondary circuit of the transformer. The conductor carrying the current to be measured is passed through the toroid center and represents a single-turn primary circuit. The flux of the magnetic field resulting from the conductor current concentrates within the magnetic material of the toroid. ac conductor current produces alternating magnetic flux within the toroid which, in turn, induces an alternating current in the secondary winding when a small resistance such as a shunt is connected between the winding terminals. If the secondary terminals are not connected to a low-resistance load, very large ac voltages can appear between them when current flows in the primary circuit, making the current transformer very sensitive but potentially very dangerous.

Current transformers have a long history of application in power measurements where they function to facilitate the measurement of very large, low-frequency sinusoidal ac currents (>10 A) by producing a much smaller but proportional ac current suitable for direct reading by ammeter instruments, historically the electromechanical d'Arsonval type. Secondary current is related to primary current by the inverse of the number of secondary turns. Since the primary is usually only a single turn, the secondary current can be no larger than the primary current. This makes the current transformer unsuitable for the measurement of low-level ac currents, although it does find some application in milliampere range measurements.

2.4.4 Active Shunts. Low-level current signals (<1 mA), both ac and dc, are typically measured by either Hall devices or active shunts. The lowest-level measurements (<1 μA) are typically made with active shunts that are similar to basic shunts in operation except that much larger proportional voltage levels can be obtained without increase in the effective resistance of the shunt. Instead of a discrete resistance, active shunts are composed of an extremely-low-noise feedback amplifier circuit similar to that described in Sec. 2.2, except that a large resistance is used in the amplifier feedback path in place of the capacitor. A typical arrangement is illustrated in Fig. 4.

The output voltage V_{out} is the product of the current flowing out of the shunt's positive terminal and the feedback resistance R except for the polarity change. Active shunts are also known as current amplifiers or picoammeters.

2.5 ac Impedance Measurement

The impedance of a circuit element is defined as the ratio of the voltage drop across the element to the current flowing through the element. In dc circuits, this ratio is the resistance of the element. In ac circuits, some elements produce not only a voltage drop, but also a phase shift between the voltage and current. In this case, the ratio (impedance) is represented by a complex number that defines both the magnitude and the phase shift. An element that produces only a voltage drop is called *resistive*. An element that produces a phase shift is called *reactive*. There are two common types of reactive elements: *capacitance*, in which the current leads the voltage by a phase angle of 90°, and *inductance*, in which the current lags the voltage by 90°.

The complex-number representation of an impedance element is given by

$$Z=R+jX, \tag{6}$$

where the resistive element (R) is in series with a reactive element (X). The "reactance"

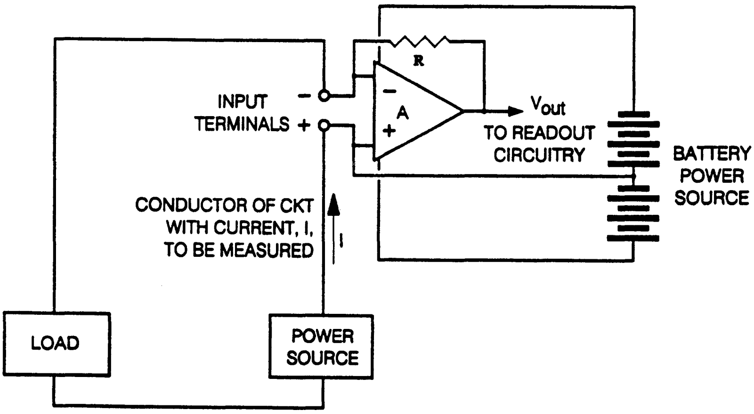

FIG. 4. Active shunt for measuring low-level currents I; the amplifier gain A and the feedback resistance R are made very large to give a large output voltage V_{out} (equal to IR) but with a negligible voltage drop between the input terminals.

of a capacitance element (X_c) or an inductance element (X_l) is

$$X_c = -1/\omega C, \tag{7}$$

$$X_l = \omega L, \tag{8}$$

where ω is the angular frequency of the signal.

dc resistance can be accurately measured by digital multimeters using the techniques described in Secs. 2.3 and 2.4. These involve measuring the voltage developed across the resistor when a fixed dc current is applied. For the most accurate dc measurements, the same current flows through the resistor being measured and a known (standard) resistor. The ratio of the two voltage measurements is the same as the ratio of resistance values. A network of resistors in which voltage measurements can be used to determine resistance values accurately is referred to as a *bridge* circuit. Bridge circuits can also be used to measure ac impedance, but the procedures are complicated by the phase shifts between the voltage and current that occur in reactive impedance elements. Bridge circuits for measuring dc and ac impedance will be described in the following section.

2.5.1 Bridge Measurements. Various forms of bridges have been used for many years to make high-precision measurements of ac and dc circuit elements. Although their use has diminished somewhat because of the capabilities available in modern digital multimeters, they are still needed for some kinds of specialized measurements, such as high-frequency impedance, or extremely precise resistance measurements. Bridge circuits still form the basis of many of the modern digital instruments, in which the operation of the bridge is automated and the associated calculations are done by an internal microprocessor. The basic principles of bridge measurements are reviewed here along with the concepts of guarded circuits and three-terminal measurements.

2.5.1.1 dc Bridges. Figure 5 shows the basic configuration of a *Wheatstone bridge.* The bridge has four resistive arms and is excited across a pair of opposite corners by a dc voltage source, usually 5–10 V. The unknown resistor is connected in one leg as shown, and the standard resistor is adjusted to give a null (0 V) across the galvanometer

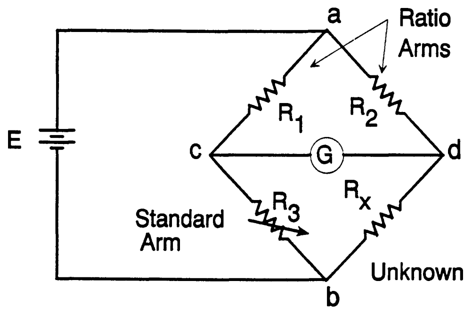

FIG. 5. The Wheatstone bridge is the basic configuration for precision measurement of dc resistance for values ranging from 1 Ω to several megohms. Lower resistance values (down to 0.000 01 Ω) can be measured with a modification called a Kelvin bridge.

(null detector). The unknown resistor can be determined from the other bridge elements as follows:

$$R_x = R_3\,(R_2/R_1). \qquad (9)$$

The measurement of the unknown resistor is independent of the excitation voltage or the calibration of the null detector. The Wheatstone bridge has been widely used in precision dc resistance measurements for values between approximately 1 Ω and 2 or 3 MΩ.

For lower resistance values, the errors due to lead resistance and contact resistance become significant for the Wheatstone bridge. A variation known as the *Kelvin bridge* (Helfrick and Cooper, 1990) can be used to measure resistance values as low as 0.000 01 Ω.

For higher resistance values (above 1 MΩ), the shunting effects of insulation resistance and leakage current can become significant. The concepts of guarded measurements and three-terminal resistance measurements were developed to overcome these problems with bridge measurements. These concepts are also used in other kinds of modern electronic instruments for measuring high resistance or ac impedance. Figure 6 illustrates a guarded bridge circuit and a three-terminal resistance measurement.

The unknown resistor R_x is connected to a pair of terminals mounted on insulating posts attached to a metal plate. A third terminal, called the guard, is connected to the metal plate. The main terminals of the resistor are connected to the bridge in the usual manner, sometimes using shielded leads whose shields are tied to the metal plate. R_1 and R_2 represent leakage paths and insulation resistance of the insulators and connecting cables. These resistances may be very high, but the shunting effect will be significant if R_x is also very high. To make the bridge measurement, the guard terminal of the resistor is connected to the guard terminal of the bridge as shown. In the bridge circuit R_2 is in parallel with R_a, but since R_a is so much smaller, the effect is negligible. Similarly R_1 is in parallel with the null detector whose resistance is also very much smaller. This effectively eliminates the effects of external leakage paths from the bridge balance condition.

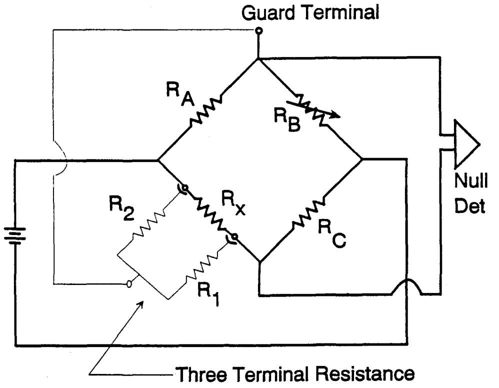

FIG. 6. Guarded three-terminal resistance measurements reduce the shunting effects of terminal insulation resistances (R_1 and R_2) in a high-voltage megohm bridge.

2.5.1.2 ac Bridges. The bridge configurations discussed in Sec. 2.5.1.1 are used to measure resistance. They are normally excited with a dc voltage source or an ac voltage source at a low frequency where the effects of reactive elements are negligible. At higher frequencies, the reactive phase shifts can produce significant effects in the operation of the bridge. By adding known reactive elements to the standard and ratio arms of the bridge (Fig. 5), the magnitude and phase of an unknown impedance can be determined by the balance condition of Eq. (9) using the complex impedance values in place of the resistance values. This, of course, means that both magnitude and phase conditions must be satisfied to balance the bridge. It also means that there may be component values for which there is no balance condition. For example, in Fig. 5, an inductive element in the R_x leg could not be balanced by an inductive element in the R_1 leg because the phase shifts are in the same direction. A capacitive element would be required in one of the legs.

Because the null condition requires balancing of both magnitude and phase, different configurations of ac bridges are used for measuring inductive and capacitive impedance values. The *Maxwell bridge* configuration is used for measuring inductance coils with phase shifts ranging from about 45° to 85° ($1 < Q < 10$). The *Hay bridge* is used for inductors with phase angles approaching 90° (high Q). The Q value of a reactive element is a measure of the ratio of stored to dissipated power and can be shown to be equal to the tangent of the phase angle of the impedance. From Eqs. (6) and (8) it is seen that the Q value of a series resistance and inductance element is

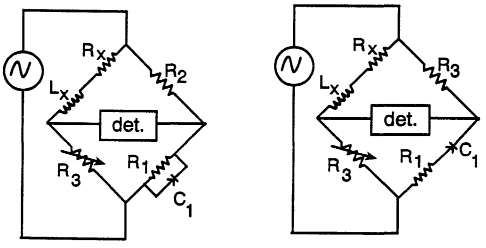

MAXWELL BRIDGE

HAY BRIDGE

FIG. 7. ac bridge configurations used for low-Q (Maxwell) and high-Q (Hay) inductance measurements.

$$Q = \omega L/R. \tag{10}$$

The *Schering bridge* is a widely used configuration for measuring capacitance values, particularly for well-insulated low-loss capacitors. Figures 7 and 8 illustrate the configurations of these three common ac bridge circuits.

2.5.2 Active Guard Circuits. The problem of leakage currents in measuring very high resistance values was discussed in Sec. 2.5.1. This problem is even more troublesome when trying to measure small capacitance values with an ac bridge. In many cases the capacitance of the connecting leads and other stray capacitances are much larger than the capacitor being measured. The *Wagner ground connection* (Fig. 9) is used for elimination of stray capacitance in ac bridges. The Wagner ground circuit consists of the resistor R_W and the capacitor C_W, which are adjusted manually until no change in the bridge balance is observed when switch S is moved from position 1 to position 2. In this condition, point A of the bridge is at ground potential, and essentially no current can flow through the stray capacitances represented by C_1 and C_2. There are also capacitances from the excited corners of the bridge to the Wagner ground, but the balancing procedure minimizes their effect. Many electronic instruments use the method shown in Fig. 10 to reduce the effect of stray capacitance. The critical parts of the circuit are surrounded by a guard (or shield) structure. The amplifier circuit serves to maintain the guard at the same potential as some point in the bridge, which eliminates any current flow through the stray capacitance to the guard. This concept is referred to as an active (or driven) shield and is often used to eliminate the effects of cable capacitance.

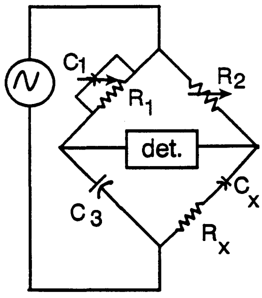

FIG. 8. The Schering bridge is used for low-loss capacitance measurements.

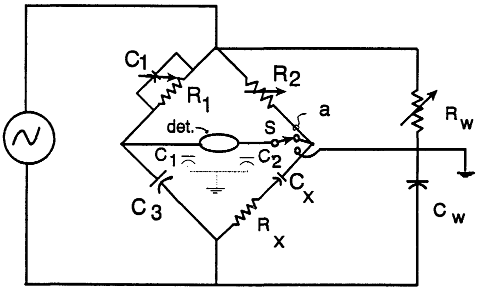

FIG. 9. The Wagner ground connection between R_W and C_W eliminates the effect of stray capacitance whenever the bridge is balanced so that no change is observed whenever switch S is moved.

2.5.3 Electronic Impedance Instruments. General purpose bridges are seldom used today because of the availability of microprocessor-based electronic instruments. Bridge instruments are still used, but the bridge functions are completely automated. Measurement results are computed and displayed in the desired units by the internal microproces-

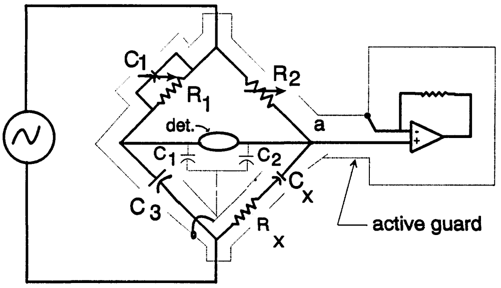

FIG. 10. Illustration of an active (driven) guard circuit. The amplifier acts to drive the guard to the same potential as point a, thereby greatly reducing the current flow through stray capacitances C_1 and C_2.

sor. There are other kinds of electronic instruments that measure impedance and phase relationships over a wide range of frequencies. The *vector impedance meter* measures impedance at frequencies below 100 kHz by direct measurement of the voltage and current magnitude and phase difference. The phase difference is measured by means of timing circuits that determine the time between zero crossings of the two signals. For higher frequencies the *vector voltmeter* converts the input signals to two heterodyne intermediate frequencies (IF) that have the same amplitude and phase relationships as the original signals. Ratio and phase measurements are made on the low-frequency (IF) signals.

2.6 Electrical Power Measurement

The current, I in amperes, through a resistive load, R in ohms, and the voltage, V in volts, across the load are related by

$$V=IR, \tag{11}$$

which arises from the basic definition of resistance. The power dissipated (or consumed) by the load is determined by the product of the current and the voltage drop. Using Eq. (11) the following simple relationships for power P are obtained:

$$P=VI=V^2/R=I^2R. \tag{12}$$

The power can be determined by measuring the voltage across the load and the current through the load. In a dc circuit, the power is calculated by the product of the voltage and current (VI) as shown in Eq. (12). The VI product represents the instantaneous power flow to the load. In an ac circuit, the VI product must be averaged over a cycle to obtain the net power dissipated by the load. If the load has a reactive component [see Eq. (6)] the voltage and current will not be in phase. In this case, the average power over a cycle is

$$P_{\mathrm{av}}=V_{\mathrm{rms}}I_{\mathrm{rms}}\cos(\phi), \tag{13}$$

where ϕ is the phase angle between the voltage and the current. The cosine of the phase angle in Eq. (13) is called the power factor.

For purely reactive loads the phase angle is 90° and the average power is zero. In this case, the load is storing energy for part of the cycle and returning it to the circuit during the remaining part. This part of the power is called reactive power and is given the units of vars (volt-amperes reactive). The quantity calculated by Eq. (13) is called the active power (watts) and represents the power actually consumed by the load. The product of rms voltage and rms current is called the apparent power. The relationship between apparent power (S), active power (P, watts), and reactive power (Q, vars) is

$$S=\sqrt{P^2+Q^2}. \tag{14}$$

ac power measurements and the application of Eqs. (13) and (14) are fairly simple for single-phase circuits with a single load. The relationships are more complex for multiple-phase, multiple-load circuits. For further information the reader should consult a textbook on electrical power engineering (e.g., Wildi, 1981).

The terminology and concepts described above are more widely used with electrical power applications than with electronic instrumentation, but there are a variety of instruments available for measuring these quantities.

2.6.1 Electrodynamometer Power Measurements. The electrodynamometer movement is used extensively in making direct-reading wattmeters. This movement is similar to the familiar d'Arsonval movement, which uses a moving armature coil in a permanent-magnet stator field. In the electrodynamometer movement, the stator field is generated by a set of windings that are excited by the load current. The armature coil is excited by the voltage drop across the load, connected with a fairly large current-limiting resistance in series. The torque on the armature coil is proportional to the product of the load voltage and the load current. The slow dynamic response for the movement provides averaging, so that the steady-state position of the coil is an indication of the power being measured. By changing the configuration of the stator coils or shifting the phase of the armature voltage, these instruments can be made to indicate other parameters such as power factor or reactive power (vars).

2.6.2 Watt-Hour Meters. One of the most common and familiar instruments for electrical power measurement is the watt-hour meter, which is used by electric utilities for metering power to customers. This instrument is essentially a small induction motor. The rotor is an aluminum disk in which eddy currents are induced by a fixed coil excited by the voltage signal. The stator field is generated by another set of fixed coils that are excited by the load current. The torque on the disk is proportional to the voltage–current product. The disk is damped by permanent magnets, which induce eddy currents as the disk rotates. This gives an arrangement in which the speed of rotation of the disk is proportional to the power. A set of gear-driven mechanical counters provide a totalized measurement of the electrical energy (watt-hours) delivered to the load.

2.6.3 Electronic Power Measuring Instruments. Electronic power transducers based on the Hall effect described in Sec. 2.4.2 are used in many industrial or laboratory instrument systems. Referring to Fig. 3, if I_{measured} is proportional to the load current and V_{bias} is proportional to the load voltage, then V_{error} will be proportional to the power. This provides an analog multiplication of the two signals in a manner similar to the electrodynamometer instruments described above.

Microprocessor-based digital power analyzers are available that can make measurements in a variety of multiphase and multiload circuits. These instruments digitize and store the voltage and current wave forms at a very fast sampling rate. All of the desired parameters are calculated and displayed by the internal computer.

3. INSTRUMENT INTERFACES

For a measurement made by an instrument to be useful, its data must be presented for interpretation by a person or by another piece of equipment.

3.1 Instrument-to-Human Interfaces

3.1.1 Visual Display. Instruments present data to a person through the use of one or more of the person's senses. The most often used sense is that of sight. In addition to the wide variety of precise and nonambiguous means to convey information visually, people have remarkable abilities to filter spatially (i.e., to focus attention on a single visual indication while ignoring other indications and distractions in the vicinity).

Some instruments are inherently sight-oriented. Galvanometers, mercury-in-glass thermometers, and Bourdon tubes exhibit deflections that reflect the value of the parameter being measured and can be observed visually. Electronic instrumentation usually involves the use of a transducer to convert the parameter being measured to an electrical signal and requires a means of displaying the amplitude of this signal.

Analog electrical interfaces usually communicate the measured value by controlling the displacement of one or more mechanical (or simulated mechanical) indicators along a scale. For many situations this type of indication is preferable to a purely digital (numeric) display. These include instances in which

1. the relative value (e.g., percentage of the full scale or relationship to alarm points) of the measured variable is more important than the precise value of the variable or the units in which it is being measured (as in an automobile fuel gauge);
2. the polarity and/or rate of change of the measured variable is important (as in meters used for tuning musical instruments);
3. fluctuations in the measured variable are so rapid that reading a numeric display is difficult (as in record level indicators on audio equipment).

Although analog indications typically have lower resolution than digital displays, this is not always the case. The hour, minute, and second hands on a clock, for instance, together provide a time measurement with nearly 0.001% resolution; yet an experienced user can read and interpret the measurement with only a casual glance. A commonly used analog visual interface that does not employ or simulate mechanical indicators is the cathode ray tube. In such a device, the amplitude of the input electrical signal controls the deflection of an electron beam. The amount of deflection is observed by placing a phosphor-covered sheet of glass in the path of the beam so that light is given off at the point of impact. Up to three measurements can be displayed by a single beam if the vertical deflection, horizontal deflection, and intensity of the beam are controlled simultaneously. This interface device is useful for signals that change too rapidly for display with mechanical devices. It is also used to display simple graphs (e.g., histograms or frequency spectra) in real time.

Digital indicators are used when high resolution is required, when errors inherent in the interpretation of analog displays (e.g., parallax and interpolation) are undesirable, when data are inherently discrete in nature (e.g., event counts), or when digitized data are required for other reasons (e.g., for transfer to a computer). With recent advances in analog-to-digital conversion and numeric display technologies, simple digital displays may cost no more than inexpensive analog mechanisms with comparable resolution and accuracy. The simplest and lowest-resolution digital display, the two-state (on–off) indicator lamp, is usually reserved for status displays and alarm indicators.

Displays of digitized data often mimic analog instruments and can provide valuable features their purely analog counterparts cannot. Bar graphs, for instance, can include peak-hold functions, or can change colors under predetermined conditions. And while numeric displays become difficult to read when updated too rapidly, the readability of digitally synthesized analog indications usually improves as update rates increase.

The use of imaging as a real-time operator interface is growing in many industrial, commercial, and research areas. Two-dimensional maps (using color to indicate measured values), surface plots, and other types of graphs can be displayed on data terminals with the assistance of computer-based data acquisition systems.

3.1.2 Audio Signaling. Audio transducers are also used as operator interfaces but are usually reserved for alarms and discrete event indications in order to avoid overwhelming the operator with more audio input than humans can readily process.

Information can be conveyed audibly by controlling a sound's amplitude, repetition rate, or frequency content. An alarm buzzer is turned on (a form of discrete amplitude control) to indicate a condition that requires the operator's attention. The repetition rate of clicks from a Geiger counter indicates the intensity of an ionizing radiation field. The varying pitch of a metal detector's audio output is a function of the distance to and size of a metal object. In speech synthesis, control over amplitude, spectral content, and temporal duration are combined to create intelligible phrases.

3.2 Human-to-Instrument Interfaces

The interfaces discussed in Sec. 3.1 address the communication of measurement information to people. Most instruments also require instruction or adjustment by their operators. Input devices seldom use visual or audio techniques, relying instead on some sort of mechanical-to-electronic transducer. Knobs, switches, touchscreens, keypads and keyboards, mice, joysticks, and trackballs all provide this function. Input devices can be grouped into analog and digital categories, and each has advantages and disadvantages for any given application. Some have built-in visual feedback of their current setting, and some retain their settings after instrument power has been removed and restored.

3.3 Instrument-to-Equipment Interfaces

When a signal from an electronic instrument is to be used by another piece of equipment, it must be in a format compatible with the target system's input. Data can be communicated in analog form using signal voltage, current, phase, and frequency, or in digital form by a wide variety of different formats and protocols. Fortunately, standards exist in the research, industrial, and consumer markets that simplify the chore of electronic interfacing. In the analog domain, 0–1-V, 0–5-V, 0–10-V, 0–20-mA, 4–20-mA, 10–50-mA, and 0–10-kHz ranges are used, and devices are available for converting from one standard range to another. Few applications will require custom-developed data transmission techniques.

The term *signal conditioning* is used to refer to the processing usually required to translate the signal from the transducer, which is usually much lower in level, into one of these commonly used ranges and may also include filtering or other manipulation of the transducer's output.

Data can also be transmitted digitally in bit-serial formats (e.g., RS-232, RS-422, RS-423, or RS-485) or in byte- or word-serial formats (e.g., general-purpose instrument bus, parallel binary or binary-coded decimal, or Centronics). Some standards cover little more than signal levels or transmission media, and the user must know how much of the interfacing task is covered by the particular standard being considered.

Digital data streams have the advantage of being able to include information other than signal amplitudes (e.g., engineering units) if their data bandwidths are sufficient. Some industrial instruments now feature hybrid data interfaces that superimpose digital data on top of the analog measured-variable signal.

When harsh electrical environments, long transmission distances, or other factors make interfacing via electrical conductors impractical or unreliable, optical data interfaces (usually via optical fibers) provide an alternative that can solve attenuation and noise problems while providing high data throughput.

3.3.1 Multiplexing. To reduce the reliability and expense penalties associated with extensive field wiring, measurements from several instruments can be transmitted over a single signal path using one of several techniques collectively referred to as multiplexing. With time-division multiplexing, signal streams (whether analog or digital) are successively connected to the transmission medium in a predetermined sequence. A demultiplexer at the other end of the cable separates and sorts the signals and applies them to the appropriate outputs. Two conditions must be met for proper time-division multiplexing. First, the demultiplexer unit must be synchronized carefully to the multiplexer so that the incoming signals are routed to the correct outputs at the correct times. Secondly, because each of the signals transmitted is now being sampled, the Nyquist sampling criteria must be met: each signal must be sampled at a rate at least twice the frequency of the highest-frequency component in that signal (Oppenheim and Schafer, 1989). Otherwise, an error-producing effect known as *aliasing* will occur. Antialias filtering may be required to limit the frequency content of the signals to be digitized. Latches (for digital signals) and sample-and-hold amplifiers (for analog signals) can be used to eliminate the time gaps in the sampled data at the output of the demultiplexer.

Frequency multiplexing permits the simultaneous transmission of several continuous signals. Baseband signals modulate carrier signals of different and properly spaced frequencies. These composite signals are combined by a mixer. At the receiving end of the transmission system, bandpass filters separate the carrier signals and feed each to a demodulator. If the system has been properly designed and implemented, the action of the multiplexer and demultiplexer circuitry is completely transparent to the end devices.

Optical data transmission systems can use wavelength multiplexing to send several signals down a single optical fiber. Input signals modulate optical sources (light-emitting diodes or lasers) having different output wavelengths. These light signals are combined using optical mixers and launched into one end of the fiber. Optical splitters (filters or prism-like devices) are used at the receiving end to separate the colors, and optoelectronic devices (usually light-sensitive diodes) convert the optical energy back into electronic signals.

In broadband transmission systems, time-division multiplexing is combined with either frequency or wavelength multiplexing to increase the number of channels of information that can be transmitted simultaneously.

3.3.2 Analog-to-Digital Conversion. Continuous electronic signals are digitized (converted to digital form) by analog-to-digital converters (ADCs). An ADC operates on one of several different principles, each of which has its unique advantages and characteristics. These features are summarized in Table 1; the following is a brief description of the operating principles.

Successive approximation ADCs determine the digital value corresponding to the analog input voltage by repetitively comparing the input to the output of an internal digital-to-analog converter (DAC) and adjusting the code fed to the DAC for a better match. An N-bit ADC requires N compare–adjust cycles to obtain the optimum digital representation. That code is then applied to the device's digital output and a conversion-complete indication is given. Accuracy depends on the accuracies of the internal comparator and DAC. By making the reference input to the internal DAC an

Table 1. Comparison of the features of common analog-to-digital conversion techniques.

Analog-to-digital technique	Accuracy	Speed	Power Use	Cost
Successive approximation	Medium	Medium	Low to medium	Low
Integrating	High	Low	Medium	Medium
Flash	Low	High	High	Medium to high
Sigma-Delta	Medium	Medium	Medium	Medium

analog input to the device, a multiplying converter is created—the output count will be proportional to the product of the analog input voltage and the DAC reference voltage.

Integrating ADCs contain a constant-current source connected to the input of an analog integrator. An internal counter measures the time required for the integrator to ramp up to a voltage matching the input signal. The required time is proportional to the input voltage. Variations on the basic integrating converter design can compensate for internal offsets and integrator leakage. Errors introduced by noise having frequency content higher than 1/(integrating period) are reduced with integrating ADCs.

Flash ADCs consist of a 2^N-element voltage divider string with one of the inputs of a differential comparator connected to each node of the string. One end of the string is grounded and the other is connected to a voltage reference. The remaining inputs of the comparators are connected to the input signal. Thus the output of all comparators with string-connected inputs at a voltage higher than the input signal will assume one output state and the outputs of remaining comparators will assume the other. Decoding circuitry converts the pattern of converter outputs into an N-bit output word. Flash converter resolution is limited by the number of comparators that can be incorporated into a single device. Power considerations presently limit output resolution to 8 bits or less.

Sigma-delta conversion is a relatively new technology. Although the technique has been known for approximately 25 years, only recently have advances in digital signal processing technology made it practical to fabricate these converters in integrated-circuit form. Sigma-delta converters oversample the input signal by a high factor (usually 32 or 64), produce a one-bit stream of digital data at this high rate, and then process this stream with digital decimation filters that produce high-resolution (14-, 16-, or 18-bit) data streams at reduced sample rates. The highly oversampled inputs reduce the requirements on antialias filters substantially and usually eliminate requirements for sample-and-hold amplifiers.

Digitization techniques also differ in their requirements for antialias filtering and for external control signals. Some must follow sample-and-hold amplifiers for optimum performance while others require additional external electronic components. Some are easy to implement in miniature, integrated-circuit form and others must be built with larger hybrid circuits and/or discrete components. The proper selection and application of ADCs are among the most critical and challenging issues in high-performance instrument design.

3.3.3 Digital-to-Analog Conversion. Performing the inverse of the ADC function, digital-to-analog converters (DACs) accept data from digital devices (conventional computers, digital signal processors, or more specialized circuits) and transform them into a continuous analog output signal. The vast majority of DACs are based on the same internal architecture: a resistor "ladder" network (Fig. 11) in which resistors or capacitors are switched in and out of a current-summing network according to the digital input code (Sheingold, 1986). DACs have either current or voltage outputs and are usually less complex and less expensive than ADCs of comparable resolution, accuracy, and speed.

3.4 Recording

We will use the term *recording* to mean the preservation of experimental data for subsequent inspection, analysis, or presentation. A recorder, therefore, is a device for copying measurements made by electrical instrumentation to a physical medium in a useful for-

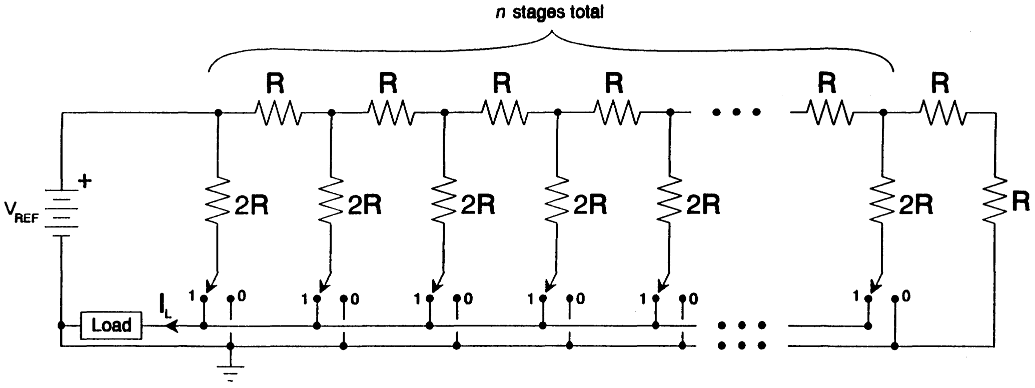

$$I_L = \frac{V_{REF}}{R} \cdot \frac{x}{2^n}$$

FIG. 11. *n*-stage *R*-2*R* resistor ladder as applied to digital-to-analog conversion. The term *x* in the equation is an integer determined by summing the binary-weighted values of the switch settings.

mat. The resulting record can be read (interpreted) by a person or by a processing system or translated into a format suitable for this purpose. Recordings, like the user interfaces described above, can be in either analog or digital form.

3.4.1 Analog Recorders. True analog recorders make continuous time records of measured variables. Strip chart recorders make such records by passing long sheets of paper beneath one or more marking pens that are deflected in a direction perpendicular to the direction of paper travel by an electrical signal corresponding to the magnitude of the variables. In this manner, the pens make traces on the chart that show how the measurements vary with time. Strip chart recorders also serve as useful real-time interfaces because data trends and rates of variable changes are readily apparent. These recorders have accuracy and resolution limitations similar to mechanical meter movements described above (the pens are often mounted on the end of galvanometer "needles") with additional limits imposed by the accuracies of the chart grid and the paper transport mechanism. Inertia in the pens and their deflection mechanisms restrict the use of strip chart recorders to signals that have no frequency content above a few hundred hertz. Electrocardiographs and seismographs are familiar examples of strip chart recorders.

Short segments of rapidly changing analog signals can be captured on photographic film by using especially designed cameras to take pictures of cathode ray tube displays or by using film recorders. Magnetic tape recorders can be used to record experimental data just as they are used to capture wave forms generated by microphones and electronic musical instruments in audio recording. Low-frequency limitations inherent in direct recording of signals can be circumvented by using the input signals to frequency-modulate carrier signals. In this technique, the deviation of the carrier signal from its center frequency corresponds to the instantaneous amplitude of the signal being recorded, and signals with frequency components down to and including dc can be preserved. Pulse-width modulation can also be used for recording signals with important low-frequency information. Both of these modulation techniques require more high-frequency bandwidth than would be required by direct recording.

Recent announcements of semiconductor devices (analog electrically erasable programmable read-only memories) that capture fragments of analog signals and preserve them indefinitely foretell of a new technology that could have a major effect in analog signal processing (Pell, 1991).

3.4.2 Digital Recorders. Digitized data can be recorded using techniques similar to the analog methods described above and can take advantage of other methods not applicable to analog signals. Conventional data loggers are the digital counterparts to analog strip chart recorders. Numeric values and/or simulated analog curves representing input signals are printed on long strips of paper.

Data can be preserved magnetically on tape, disks, and cards. Because digital data can be stored temporarily (buffered) and recorded in bursts of high-frequency signals, special modulation techniques are not required for recording low-frequency data.

Equipment for recording data on optical disks has recently become available for high-density recording of digitized experimental data. Some of this equipment has analog inputs and outputs—the digitization of the stored data is transparent to the user. Once written, data on most optical disks cannot be altered, but new devices incorporating magneto-optical technology allow users to erase and modify stored information.

Semiconductor memory preserves digital data in memory cells made up of transistors and diodes. Volatile memory can retain information only as long as power is applied to the devices. Nonvolatile memory preserves the information after power has been removed and even after the devices have been extracted from the circuitry that supplied the information. Alhough it lacks the large storage capacities of some magnetic and optical storage devices, random-access semiconductor memory can be written to and read from more quickly than any other type of storage media.

Magnetic bubble memory is an alternative technique for nonvolatile storage of data. Bubble memory devices may be similar to large integrated circuits in appearance and in fabrication techniques, but they are not semiconductor devices. Data cannot be accessed randomly as in semiconductor memories, but access times rival that of the fastest fixed-disk storage systems. Data can be stored indefinitely with or without power applied and the devices are more immune to shock, vibration, and other environmental effects than other magnetic storage systems.

4. NOISE AND OTHER SOURCES OF ERROR

Many of the electronic instruments discussed in Sec. 2 are designed to measure signals of less than 1 μV. High-gain amplifiers enable these signals to be displayed on an oscilloscope. It is usually the case that whenever the most sensitive ranges of an electronic instrument such as an oscilloscope are used, the signal becomes corrupted with "noise" signals, which may appear spurious or random in nature. The level of the noise limits the minimum signal that can be measured with the instrument; therefore, it is desirable to minimize the noise as much as possible. There are two basic sources of noise that can occur in electronic systems: induced noise and inherent noise.

4.1 Induced Noise

Induced noise is picked up in an electronic circuit from external sources, such as 60-Hz power lines or high-frequency (rf) equipment. This type of noise is referred to as electromagnetic interference (EMI) and can be quite large. Fortunately EMI can be reduced or eliminated by proper shielding and grounding of circuit components and signal leads. Shielding and grounding considerations can become quite complex; however, a few simple rules can help reduce or eliminate many common problems.

Many modern electronic instruments are designed with shielded circuits that reduce both noise pickup and noise emissions. In most cases the use of shielded signal cables and proper connections to the instrument as recommended by the manufacturer will produce acceptable results, especially in a laboratory environment. Two common types of input configurations are used for test instruments: differential and single-ended. A differential input has two terminals (plus and minus) for the signal leads and, sometimes, a third terminal for the shield. A single-ended input usually has one coax connector terminal with the center conductor for the signal lead and the outer shell for the shield. The shield serves as the "common" or "negative" lead for the signal and the measurement circuit.

For single-ended measurements, the shield is part of the circuit and is always connected at both ends. For differential measurements, the user must make a choice as to how to connect the shield. In most cases the best practice is to connect the shield to the test instrument shield or ground terminal and leave it unconnected ("floating") at the other end. Connecting the shield at only one point serves the purpose of breaking any possible circuits (called *ground loops*) through which large noise currents might flow. These currents, if present, might induce additional noise voltages into the measuring circuit. For additional information on shielding and grounding techniques, the reader is referred to textbooks [see, e.g., Morrison (1986) or Ott (1976)].

Induced noise is usually coupled into the measuring circuit by capacitive or inductive impedances, which implies that it is alternating or transient in nature. It is very difficult to eliminate induced signals completely, especially at the ac power line frequency (50 or 60 Hz). Instrument designers use various methods to reduce the effects of the noise. A common practice in instruments that make low-level dc measurements is to integrate (or average) the signal over a time interval that is a multiple of the period of the power line frequency. The induced noise at this frequency will then average to zero. Averaging also helps reduce the effects of any random noise that might be present since it also tends to average to zero.

There is another form of induced noise that can be particularly troublesome in making low-level dc measurements. These are small dc voltages which are produced by the Seebeck effect whenever conductors of different materials are joined together in the presence of temperature gradients. This is the principle upon which a thermocouple functions as a temperature measuring device. This voltage (called thermal emf) is very difficult to eliminate completely and therefore must be considered when errors on the order of a microvolt might be significant. It is easy to detect the presence of a thermal emf by simply reversing the polarity of the signal leads at the test instrument. Since the polarity of the thermal emf does not change, it will add to the measurement signal in one case and subtract in the other.

4.2 Inherent Noise

Inherent noise is generated by random molecular and atomic processes within individual circuit elements such as resistors, diodes, transistors, or any other active device. It appears in several different forms depending upon the dominant mechanism of generation. It is inherent in all electronic devices and will always be present at some level. It can be controlled to a certain extent by circuit designers with the use of low-noise components and careful placement in the circuit. This section will present a brief description of the basic types of inherent noise that might be observed when making measurements with certain kinds of electronic instruments.

Inherent noise is generally random in nature and is present over a broad range of frequencies. One of the most commonly observed is *thermal noise,* also referred to as *Johnson noise* or *Nyquist noise.* Thermal noise results from the motion of free charge carriers (electrons) due to thermal energy and is present in all resistors. The basic description of thermal noise was originated by Nyquist in 1928. Using the second law of thermodynamics, he showed that the noise power spectral density of a resistor R in thermal equilibrium at temperature T was given by

$$S_v(f)=\frac{4hfR}{\exp(hf/kT)-1}, \tag{15}$$

where

h=Planck's constant=$6.624\text{x}10^{-34}$ J s

k=Boltzmann's constant

=1.38×10^{-23} J/K

f=frequency (Hz)

S_v=power spectral density

(mean square V/Hz).

For frequencies below approximately 10^{11} Hz, Eq. (15) reduces to

$$S_v=4kTR. \tag{16}$$

Note that in Eq. (16), the noise power is constant with frequency, a condition that is referred to as "white" noise. As an example of the order of magnitude of thermal noise, a 10 000-Ω resistor at room temperature (27 °C) generates an rms signal of 1.3 μV over a bandwidth of 10 kHz.

Other types of noise are observed in different electronic devices. "Shot" noise, first studied by Schottky, is caused by the motion of charge carriers in a drift current as in a vacuum tube or semiconductor diode. Shot noise, like thermal noise, also has a white spectrum. In vacuum tubes and semiconductor devices the observed noise at low frequencies is much higher than that predicted by the Nyquist equation (15) for thermal noise. This is called "flicker" noise and has a spectrum characterized by

$$S_f(f) = K I_{\mathrm{dc}}^n / f^\alpha, \tag{17}$$

where K is a constant dependent on the material, n is approximately 2, and α varies between 0.5 and 2. S_f is the noise (mean square A/Hz) observed on the dc current flowing through the device. Flicker noise is sometimes called $1/f$ noise or pink noise. It has been observed at frequencies as low as 10^{-6} Hz, and, in carbon-composition resistors, can exhibit power spectral densities over 10 times greater than thermal noise below 100 Hz (Kennedy, 1988). The major cause of $1/f$ noise in semiconductor devices appears to be traceable to surface properties of the material. The generation and recombination of carriers in surface energy states and the density of surface states are important factors. Improved surface treatment during manufacturing has been able to reduce $1/f$ noise in semiconductors.

Other noise-generation mechanisms have been observed in various kinds of semiconductor devices, but the ones described above are the most common. Although electronic noise is of primary interest to electronics designers and researchers, the user of electronic instruments will find it useful to be aware that it is always present, and at what level it might start to affect the measurements being made.

5. CONCLUSION

Many changes in electronic instrumentation have resulted from advances in integrated-circuit and microprocessor technology. The pocket-sized digital multimeters in common use today combine many of the capabilities that required several specialized laboratory devices only a few years ago. Modern laboratory instruments also provide many combined functions in a single device able to make routine measurements at levels that formerly required careful and skillful techniques. Many laboratory instruments are becoming available as plug-in cards for personal computers. The general trend is toward greater levels of capability being available in simpler and easier to use forms. This permits the user to make difficult measurements without having to become an electronics expert. It will always be necessary for the user to have an understanding of the basic principles and the possible "pitfalls" involved in these measurements. It is hoped that the information provided herein is useful in serving that purpose.

GLOSSARY

ac: Originally an initialism for *alternating current.* Modern usage as an adjective indicates the presence of periodic variations.

Aliasing: A type of distortion caused by sampling a signal at a rate lower than the Nyquist frequency.

Charge: A quantity of electricity. The smallest identified increment of charge is that amount associated with a single electron.

Crest Factor: The ratio of the peak value to the rms value in a periodic wave form.

Current: The net flow of electronic charge past a certain point in a circuit.

dc: Originally an initialism for *direct current.* Modern usage as an adjective means nonvarying or varying slowly in a nonperiodic manner. dc often is used to refer to a frequency of zero.

Instrument: A device used for quantifying a measured variable.

Load: A device that draws current from an electrical circuit.

Noise: Signal corruption resulting from the addition of components not related to the desired information being transmitted.

Nyquist Frequency: A frequency equal to twice the frequency of the highest-frequency significant spectral component of a signal. Signals sampled at rates higher than the Nyquist frequency theoretically can be accurately reconstructed from the sampled data.

Sine wave: A single-frequency wave form that varies according to the general function

$$F(x) = A \sin(2\pi f x),$$

where A is the peak amplitude of the function and f is the frequency (in cycles per unit of x). In electronic instrumentation, x usually represents time and f is in units of hertz (cycles per second).

Slew Rate: The rate of change of a measured variable. A common, but not strictly correct, usage of slew rate refers to the maximum rate of change of an output that a device is capable of.

Transducer: A device that converts energy from one form to another. In electronic instrumentation, transducers convert measured physical stimuli into electrical signals or electrical signals into radiated energy.

Voltage: The difference in electromotive force (also called electrical potential) between two points in a circuit or an environment.

List of Works Cited

Helfrick, Albert D., Cooper, William D. (1990), *Modern Electronic Instrumentation and Measurement Techniques,* Englewood Cliffs, NJ: Prentice-Hall, pp. 101–129.

Kennedy, E. J. (1988), *Operational Amplifier Circuits Theory and Applications,* New York: Holt, Rinehart and Winston, pp. 118–147.

Morrison, Ralph (1986), *Grounding and Shielding Techniques in Instrumentation,* New York: Wiley.

Oppenheim, Alan V., Schafer, Ronald W. (1989), *Discrete Time Signal Processing,* Englewood Cliffs, NJ: Prentice-Hall, pp. 80–131.

Ott, Henry (1976), *Noise Reduction Techniques in Electronic Systems,* New York: Wiley.

Pell, Richard (1991), "First Nonvolatile Solid State Memory for Analog Waveforms," *Electron. Prod. Mag.* **33** (9), 53.

Sheingold, Daniel H. (1986), *Analog-Digital Conversion Handbook,* Englewood Cliffs, NJ: Prentice-Hall, pp. 201–204.

Sydenham, P. H. (1983), *Handbook of Measurement Science,* Vol. 2. New York: Wiley, pp. 832–837.

Taylor, B. N. (1987a), Special Report on Electrical Standards, *J. Res. Natl. Bur. Stand.* **92**, No. 1.

Taylor, B. N. (1987b), "History of the Present Value of $2e/h$," *IEEE Trans. Instrum. Meas.* **IM-36**, No. 2.

Wildi, Theodore (1981), *Electrical Power Technology,* New York: Wiley, Chaps. 8 and 9.

Further Reading

Brown, Paul B., Franz, Gunter N., Moraff, Howard (1982), *Electronics for the Modern Scientist,* New York: Elsevier.

Doebelin, Ernest O. (1983), *Measurement Systems, Application and Design,* New York: McGraw-Hill.

Gregory, B. A. (1981), *An Introduction to Electrical Instrumentation and Measurement Systems,* 2nd ed. New York: Wiley, Halsted Press.

O'Higgens, Patrick J. (1966), *Basic Instrumentation,* New York: McGraw-Hill.

White, Donald R. J. (1980), *Electromagnetic Shielding—Materials and Performance,* Gainesville, VA: Interference Control Technologies.

White, Donald R. J. (1986), *Shielding Design—Methodology and Procedures,* Gainesville, VA: Interference Control Technologies.

ELECTROCHEMISTRY

RADOSLAV R. ADŽIĆ, *Institute of Electrochemistry and Center for Multidisciplinary Studies, University of Belgrade, Belgrade, Yugoslavia*

ERNEST B. YEAGER, *Case Center for Electrochemical Sciences and Chemistry Department, Case Western Reserve University, Cleveland, Ohio, U.S.A.*

Introduction 2
1. **The Structure of the Electrode–Electrolyte Interface** 2
1.1 Potential Concepts in Electrochemistry 2
1.2 The Electrical Double Layer 4
1.2.1 New Models 6
2. **Charge-Transfer Reactions** 7
2.1 Phenomenological Description of the Kinetics of Charge-Transfer Processes 7
2.2 Theories of Electron Transfer 8
2.3 Theories of Proton Transfer 10
2.4 Electrosorption Reactions 11
2.4.1 Partial Charge Transfer During Specific Ionic Adsorption 11
3. **Electrochemistry on Well-Defined Single-Crystal Surfaces** 12
3.1 Double-Layer Capacitance Studies 12
3.2 Electrosorption of Hydrogen and Oxygen on Noble-Metal Electrodes 12
3.3 Formation of Metal Monolayers 13
3.4 Electrocatalytic Reactions 14
3.4.1 Electronic Factors 15
3.4.2 Structural Factors in Electrocatalysis 17
4. **Semiconductor Electrodes** 19
4.1 Space-Charge Effects 19
4.2 Electrons, Holes, and Surface States in Electrode Reactions 20
4.3 Photoeffects at Semiconductor Electrodes 21
5. **Techniques for the Study of Electrochemical Interfaces and Electrode Processes** 22
5.1 Electrochemical Techniques 22
5.2 *In Situ* and *Ex Situ* Spectroscopic and Surface-Physics Techniques 22
5.2.1 uv–Visible Reflectance Spectroscopy and Ellipsometry 22
5.2.2 Infrared Reflectance Spectroscopy and Raman Scattering from Electrode Surfaces 24
5.2.3 EXAFS and Surface X-Ray Diffraction 26
5.2.4 Photoemission of Electrons from a Metal into Solution 26
5.2.5 Scanning Tunneling Microscopy 27
5.3 *Ex Situ* Studies by Surface-Physics Techniques 27
6. **Applied Electrochemistry** 28
6.1 Batteries 28
6.2 Fuel Cells 28
6.3 Large-Scale Industrial Processes 29
6.4 Inorganic and Organic Electrosynthesis 29
6.5 Electroplating 29
6.6 Electrochemical Sensors 30
6.7 Corrosion and Corrosion Protection 30
7. **Challenging Problems for the Future** 30
8. **Challenging Opportunities in Electrochemistry** 30
Works Cited 31
Further Reading 33

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

Electrochemistry, one of the oldest areas of the physical sciences, is undergoing rapid development. With important advances in concepts, instrumentation, and methodology, electrochemistry now extends into many other diverse fields. In general terms, electrochemistry is concerned principally with chemical reactions that are associated with the transfer of electrons or ions across interfaces. The most typical electrochemical interface is that between a solid metal and a liquid solution of an electrolyte (ionic conductor), but any interface involving a predominantly electronic conductor and a predominantly ionic conductor is considered to be an electrochemical interface.

Interfacial reactions are governed by the potential distribution across the interface in a so-called electric double layer. It is formed by the accumulation of excess charges of equal but opposite signs on the two sides of the interface. Many models of the double-layer structures and theories for charge-transfer kinetics across electrochemical interfaces have been developed since the fundamental aspects of electrochemical reactions were first studied in 1796 by Alessandro Volta.

In this short article only essential aspects of electrochemistry are covered, with emphasis on interfacial phenomena. These phenomena are currently attracting much attention of electrochemists and also of workers from other fields, including physics and biology. Several important areas of electrochemistry such as the conductivity of electrolytes, mass transport phenomena, electrodeposition of metals, electrochemical corrosion, electrochemical machining, and some aspects of thermodynamics have been omitted.

Electrochemistry has important applications in various areas of technology such as energy storage and conversion, large-scale production and refining of almost all nonferrous metals, production of inorganic chemicals, organic synthesis, corrosion prevention, electrochemical sensors, and biomedical applications.

1. THE STRUCTURE OF ELECTRODE–ELECTROLYTE INTERFACES

1.1 Potential Concepts in Electrochemistry

Consider a metal electrode in equilibrium with an electrolytic solution containing the ion $M_1^{z_1^+}$ resulting from the dissolution of the metal; e.g.,

$$M_1 \rightleftarrows M_1^{z_1^+} + z_1 e^-, \tag{1}$$

where the electrons then accumulate on or come from the metal M_1. On the basis of thermodynamics the electrical work W_e available from this process is

$$W_e = E z_1 F = -\Delta G, \tag{2}$$

where ΔG is the Gibbs free energy change per mole of M_1 converted to $M_1^{z_1}$, F is the Faraday, z_1 is the number of electrons per metal ion in Eq. (1), and E is the thermodynamic electrode potential. Reaction (1), however, does not fully represent the situation. The cationic charge must be compensated by the generation of an equal amount of charge of opposite sign or the consumption of an equal charge of the same sign. Thus a second reaction must attend reaction (1); e.g.,

$$M_2^{z_2^+} + z_2 e^- \rightleftarrows M_2. \tag{3}$$

Multiplying Eq. (3) by z_1/z_2 and adding Eq. (1), one obtains

$$M_1 + (z_1/z_2) M_2^{z_2^+} \rightleftarrows M_1^{z_1^+} + (z_1/z_2) M_2. \tag{4}$$

Virtually no net charge can be generated in or on the metal or electrolyte phases.

The difference in the electrode potentials for Reactions (1) and (3) is determined experimentally with a voltage measuring device (e.g., a voltameter or potentiometer) connected by two metal leads M and M' (usually copper) to electrodes M_1 and M_2. This is denoted symbolically by

$$M \,|\, M_1 \,|\, s \,|\, M_2 \,|\, M'; \tag{5}$$

Then $\Delta_{m'}^{M}\phi = \Delta_{M_1}^{M}\phi + \Delta_{s}^{M_1}\phi + \Delta_{M_2}^{s}\phi + \Delta_{M'}^{M_2}\phi$, where $\Delta_b^a\phi$ is the potential difference between the bulk of the indicated phases. The voltage difference $\Delta_{M'}^{M}\phi$ on the two identical metal leads M and M' can be shown to correspond to the difference in the thermodynamic electrode potentials E_1 and E_2 (see, e.g., Trasatti, 1980). The individual values of E_1 and E_2 cannot be determined by direct measurement.

Some researchers have considered the potential difference $\Delta_s^M\phi$ between the bulk of the metal and bulk of the solution to correspond to the absolute value of the electrode potential on the thermodynamic scale. This is conceptually misleading. This potential difference term ($\Delta_s^M\phi$) is known as the Galvani potential and corresponds to the electric work to transport a unit positive test charge from the bulk of the solution phase into the metal. The electrode work term involved with the thermodynamic electrode potential, however, includes additional energy terms such as the charging of the ions in a dielectric medium and other chemical interactions associated with the accumulation of the electrons in the electron orbitals of the metal near the Fermi level.

The Galvani potential is of interest in its own right in understanding the structure of electrochemical interfaces. The potentials involved in the Galvani potential correspond to the *inner* potentials of the metal (ϕ_M) and solution (ϕ_s) phases (see Fig. 1). The inner potential is in general defined as the work to transport a unit positive test charge from an infinite distance away in a vacuum to a position within a particular phase a. The inner potential is given by

$$\phi_a=\psi_a+\chi_a, \qquad (6)$$

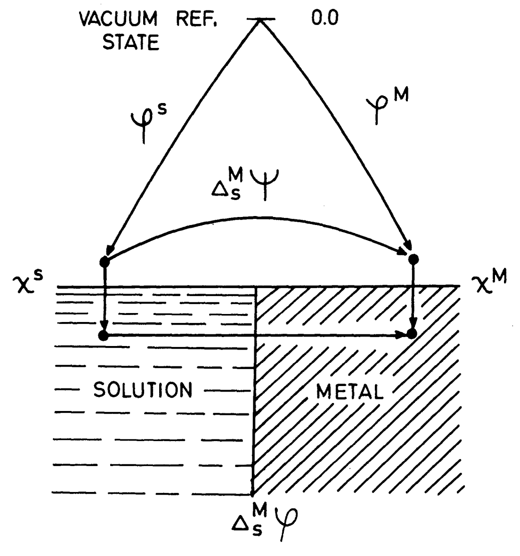

FIG. 1. Galvani ($\Delta_s^M\phi$), Volta ($\Delta_s^M\psi$), and surface (χ^s and χ^M) potentials at the metal–solution interface.

where ψ_a is the outer potential and χ_a is the surface potential. The former is a function of the net charge on the phase a and the latter is a function of dipoles at the surface of phase a. In turn, the Galvani potential may be given as

$$E_G=\Delta_s^M\phi=\Delta_s^M\psi+\Delta_s^M\chi=E_V+\Delta_s^M\chi, \qquad (7)$$

where $E_V=\Delta_s^M$ is the Volta potential. The Volta potential E_V can be determined experimentally while the surface potential difference $\Delta_x^M\chi$, and hence the Galvani potential, cannot. Any potential measuring device introduces a new interface whose Galvani potential combines with the potential difference being measured. The difference of the inner potentials between two metal leads of the same metal $\Delta_M^{M'}\chi$, however, can be measured since the surface potentials are the same for each lead. Then the Galvani potential difference reduces to a Volta potential difference which is measurable (Bockris and Reddy, 1970; Trasatti, 1980).

The difference in the thermodynamic electrode potentials ($E-E_2$) of two electrodes in a galvanic cell $M|M_1|s|M_2|M'$ corresponds to $\phi_{M'}-\phi_M$ [see Eq. (5) and the cell with which it corresponds]. Therefore it is possible to set up a relative scale of thermodynamic electrode potentials on which one electrode system is taken as the zero of the scale. For protic solvent systems such as water, the reference electrode establishing the zero of the electrode potential scale is the reversible hydrogen electrode involving the process

$$H_2 \rightleftarrows 2H^+ + 2e^-, \qquad (8)$$

occurring on a metal surface such as platinum which catalyzes this reaction. The H_2 and H^+ activities for the hydrogen reference electrode are usually not for standard-state conditions but the measured values of $E-E_{ref}$ can be easily corrected to standard state conditions for E_{ref}. Therefore we deal with

$$M|M_1|s|M_{ref}|M'$$

and

$$\Delta_{M'}^M\phi=E_1-E_{ref}.$$

For example, for a zinc electrode with a standard hydrogen reference electrode

$$M|Zn|Zn^{2+},H^+|H_2,Pt|M',$$

$$\Delta_{M'}^{M}\phi = E_{Zn/Zn^{2+}} - E_{SHE}, \qquad (9)$$

where E_{SHE} corresponds to the standard hydrogen reference electrode and the cell is isothermal and thermodynamically reversible. The electrochemical reactions are

$$Zn \rightleftharpoons Zn^{2+} + 2e^- \qquad (10)$$

$$\underline{2H^+ + 2e^- \rightleftharpoons H_2(Pt)} \qquad (11)$$

$$Zn + 2H^+ \rightleftharpoons Zn^{2+} + H_2 \qquad (12)$$

where Reactions (10) and (11) are often referred to as the half-cell reactions and Reaction (12) is the cell reaction.

The potential of this cell ($E_{Zn/Zn^{+2}} - E_{H_2/H^+}$) is a function of the concentrations of Zn^{2+} and H^+ and also the saturation pressure of H_2. From thermodynamics, it can be shown that

$$E_{Zn/Zn^{2+}} - E_{H_2/H^+} = E^0_{Zn/Zn^{2+}} - E^0_{H_2/H^+} - \frac{RT}{2F} \ln \frac{a_{Zn^{2+}} P_{H_2}}{a_{Zn}(a_{H^+})^2}, \qquad (13)$$

where R is the gas constant, T is the absolute temperature, $a_{Zn^{2+}}$, a_{Zn}, and a_{H^+} are the activities (effective concentrations) of the indicated species, and P_{H_2} is the H_2 saturation pressure. The quantities with the superscript 0 corresponds to the standard-state values.

The potential difference between the conventional electrochemical scale with $E_{SHE} = E^0_{H^+/H_2} = 0$ and the physicists' scale with a vacuum reference state for $\phi = 0$ can be calculated from various models of the electrode process occurring at the SHE reference electrode. The difference between the two scales corresponds to the removal of an electron from the Fermi level of the reference electrode to the vacuum level and for the standard hydrogen reference is about 4.5 eV. Figure 2 gives a comparison of the two scales of potentials (Gerischer, 1990).

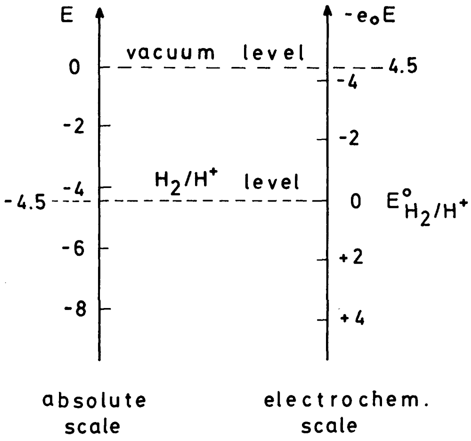

FIG. 2. Comparison of the electrochemical and absolute (vacuum) scale of potentials [adapted from Gerischer (1990), reproduced with permission of Pergamon Press].

1.2 The Electrical Double Layer

To a first approximation, the electrode–electrolyte interface can be represented as a parallel-plate condenser with one plate through the centers of the ions separated by a monolayer of solvent molecules from the electrode surface, and the second plate at the metal surface. The solvated ions, whose centers define the outer Helmholtz plane (OHP), interact with the electrode surface through electrostatic forces only. Most of the cations, e.g., Na^+, K^+, and Li^+, and anions, e.g., F^- and ClO_4^-, which have solvation shells, generally do not approach the electrode surface closer than the OHP. Some weakly solvated anions, such as Cl^-, Br^-, and I^-, are chemisorbed on the electrode surface, undergoing a chemical bonding to the surface. A partial charge transfer takes place in this interaction. The coverage of these "specifically" adsorbed ions can be substantial if a large charge transfer takes place and strong chemical bonding occurs. Their centers define the inner Helmholtz plane (IHP). The coverage of electrostatically adsorbed ions does not usually exceed 0.1–0.2 of a full monolayer.

Figure 3 gives a schematic representation of the double layer. The potential at which the net charge of the metal is zero is termed the potential of zero charge (PZC) and is characteristic of the metal as well as the electrolytic solution. The field across the double layer at the PZC is dependent on the orientation of dipoles, including water molecules at the surface. The model of an electrical double layer as a condenser with a fixed plane of charge in the electrolyte (inner layer) is due to Helmholtz (1853). Gouy (1903) and Chapman (1913) took into account the ion concentration gradient extending from the electrode surface into the solution (diffuse layer) caused by the charge in the double layer. Stern (1924)

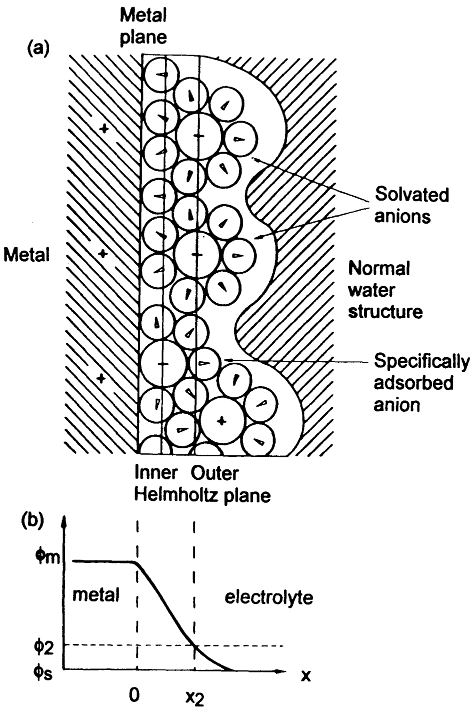

FIG. 3. Model of the electrochemical double layer.

modified this model by taking into account that the finite size of ions results in a limiting distance of closest approach to the electrode surface. The double-layer capacitance is a series combination of the capacitance of the compact and diffuse layers, and the total capacitance of the metal–solution interface, $C_{M\text{-}s}$, is given by

$$1/C_{M\text{-}s} = 1/C_{M\text{-}2} + 1/C_{2\text{-}s}, \qquad (14)$$

where $C_{M\text{-}2}$ and $C_{2\text{-}s}$ are the capacitances between the metal and the outer Helmholtz plane and for the diffuse ionic layer, respectively. The capacitance of the diffuse ionic layer depends on the potential drop $\phi_2 - \phi_s$, [see Fig. 3(b)] where ϕ_2 is the potential of the outer Helmholtz plane and ϕ_s is the potential of the bulk solution (Delahay, 1965). For a symmetrical electrolyte $C^{z+}A^{z-}$,

$$C_{2\text{-}s} = |z| fA \cosh[|z| f(\phi_2 - \phi_s)/2], \qquad (15)$$

where $A = (RT\epsilon c_s/2\pi)^{1/2}$, $f = F/RT$, F is the faraday, T is the absolute temperature, R is the gas constant, c_s is the ion concentration, and ϵ is the dielectric constant of the solvent. At potentials not far from E_{PZC} the equation predicts the shape of the capacitance potential curve for low concentrations (Fig. 4). $C_{2\text{-}s}$ passes through a minimum as ϕ_2-ϕ_s approaches zero. The measurement of the differential double-layer capacitance is used for the determination of E_{PZC}. The pronounced minimum at the PZC for low concentrations disappears at higher electrolyte concentrations because the value of the term $1/C_{2\text{-}s}$ still remains small compared to $1/C_{M\text{-}2}$ even at the PZC. In the presence of specifically adsorbed anions, the PZC is shifted to more negative values. At high positive charges on the electrode the anions become partially desolvated, and this reduces the distance of closest approach to the surface, causing a large increase of $C_{M\text{-}2}$. At high negative charges, the water structure around only the cations is changed, causing a relatively small increase of the capacitance. The Gouy–Chapman–Stern model only partially accounts for some features of the capacitance–potential curves. Some of the refinements of this model take into account the structure of the dielectric in the compact layer, dielectric saturation in the strong electric field, differences in the outer Helmholtz plane for cations and anions, and the penetration of the electric field into the metal electrode.

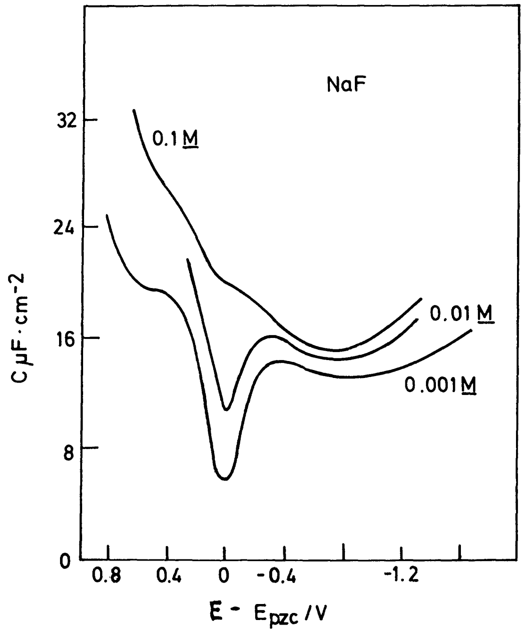

FIG. 4. Double-layer capacitance for Hg in NaF solution at 25 °C, vs $E - E_{PZC}$ (rational potential) [reproduced from Conway (1965) with permission of Ronald Press].

The potential of zero charge is related to the modified (electrochemical) work function for electrons in the electrode since zero charge conditions are required for determination of both quantities. (The Kelvin probe measures the disappearance of excess charge.) The Fermi level of an electrode at E_{PZC} correlates with the work function Φ. The relationship

$$E_{\mathrm{PZC}} = -\Phi + \mathrm{const} \qquad (16)$$

with a unit slope holds for the transition metals (Fig. 5). For *sp* metals the slope is not unity because of the metal-dependent water orientation (Trasatti, 1977). Deviations from Eq. (16) may arise from different dipole structure and ion adsorption at the metal–electrolyte interface, particularly the latter. The work function as well as the PZC depends on the crystal orientation. This means that even the concept of a PZC for a polycrystalline electrode may be undefined. The majority of more recent double-layer studies with solid metal electrodes have been done with single-crystal surfaces (Hamelin, 1985).

1.2.1 New Models. The theoretical analysis of the metal/liquid-electrolyte interface has been the subject of much study over the last two decades. In general, the models avoid some of the oversimplifications of earlier treatments. They take into account the ion size, the effect of solvent molecule size and its dipole moment, and the contribution of the metal valence-electron orbitals to the double-layer behavior. The full role of the metal in this context was not realized before recent calculations using the methods of surface physics. The penetration of the field into the metal electrode (the Thomas–Fermi screening distance) appears more important than heretofore realized.

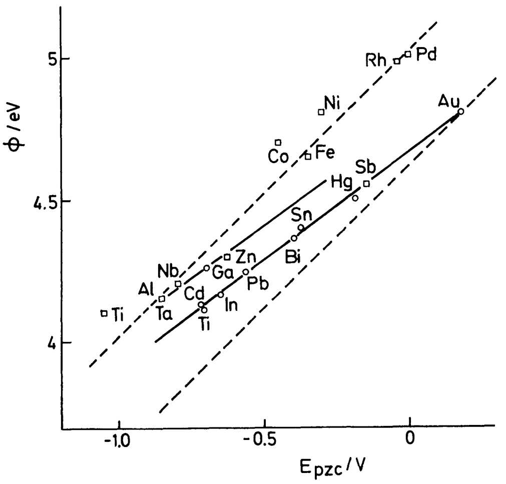

FIG. 5. Correlation of E_{PZC} with work function for various metals [reproduced from Trasatti (1977) with permission of John Wiley and Sons].

A model using hard spheres for ions and solvent with a point dipole at the center of the solvent spheres gives the following expression for the total interfacial capacitance at low concentrations (Blum and Henderson, 1981):

$$\frac{1}{C} = \frac{L_D}{\epsilon} + \frac{1}{\epsilon}\left[\frac{d}{2} + \left(\frac{\epsilon}{\epsilon_0} - 1\right)\frac{1}{\lambda}\frac{d_s}{2}\right] + \cdots , \qquad (17)$$

where d is the ionic diameter (assumed to be the same for the cations and anions), d_s is the solvent diameter, L_{D} is the Debye length, and λ is a function of the permittivities of the solution and of free space. The result may be viewed as a Gouy–Chapman capacitance (ϵ/L_{D}) in series with a second capacitance which takes into account changes in the local dielectric constant that are underestimated in this model.

Several variations of a lattice-gas model, which also allow for the effects of ionic size and solvent-molecule dipoles, produce a reasonable capacitance–potential curve if a lower dipole moment is taken for the first layer of water than for bulk water. A treatment of the monolayer of solvent separately is able to account for the shapes of capacitance curves on different metals, the temperature dependence, the H/D isotopic effect, and the effects of organic molecules. These results have been reviewed by Parsons (1990).

The role of the metal phase, ignored in the above models, has been considered with the jellium model for the metal. A substantial metal contribution to the capacitance arises from the tailing off of the electron density versus distance into the interface. The metal phase also modifies the polarization of the solvent in the IHP layer by affecting the distance of closest approach, d_i (determined by metal–solution coupling), and the relaxation of the electronic structure with charge on the metal which gives the series capacitance component C^M(dip). At the PZC, for several models, C can be approximated by

$$1/C \approx 1/C_0 + 4\pi d + 1/C^M(\mathrm{dip}), \qquad (18)$$

where C_0 is the capacitance corresponding to the solution side. The comparison with experimental data has so far been made at the PZC and not with the capacitance–potential curve over a wide range of potentials (Badialli, 1986).

2. CHARGE-TRANSFER REACTIONS

Electrochemical reactions can be divided into two classes: (I) outer-sphere one-electron transfer with the solution-phase electron donors or acceptors in the OHP when the electron transfer occurs, and (II) more complex electron-transfer processes where more than one electron may be transferred. The adsorption of reactants, products, and/or intermediates may be involved and the reactions are generally more complex. Class I electron transfer involves a simple ionic redox process in which only a change of the oxidation state of reactants positioned in the OHP is involved. Class II electrochemical reactions are often multiple steps with some chemical steps taking place before, between, or after the electron transfer steps. Under some circumstances, one step, called the rate-determining step (RDS), limits the rate, while all other preceding and following steps can be considered to be in equilibrium. Many electrochemical reactions of organic molecules and reactions accompanied by gas evolution or dissolution are in class II.

Just as for chemical reactions in general, the charge transfer is controlled by the existence of an energy barrier between the oxidized and reduced states. A unique feature of electrode reactions is that the height of this barrier can be decreased or increased by changing the potential across the metal–electrolyte interface.

2.1 Phenomenological Description of the Kinetics of Charge-Transfer Processes

Because of the heterogeneous nature of electrode reactions, the rate of the charge transfer across the electrochemical interface depends not only on the potential but also on the double-layer structure and the adsorption of reactants, intermediates, products, and other constituents of the solution either intentionally or nonintentionally present. Mass-transport limitations will not be considered here. The expression relating the current (the rate of reaction) to the electrode potential can be obtained from the absolute rate theory applied to the electrochemical interface. For electrochemical reactions the heights of the free-energy barriers are functions of the potential drop across the interface in accordance with absolute rate theory. For a one-electron transfer reaction,

$$R \rightleftarrows O + e^-, \tag{19}$$

a shift in the electrode potential from 0 to a value E causes the changes depicted in Fig. 6. The barrier for the oxidation $\Delta G_a^\ddagger$ is decreased by a fraction α of the energy change nFE, while the barrier for reduction is increased by $(1-\alpha)nFE$. The current–potential relationship is given by

$$\begin{aligned} i &= i^{\rightarrow} - i^{\leftarrow} \\ &= nF\{k^{\rightarrow} C_R \exp(+\alpha nfE) \\ &\quad - k^{\leftarrow} C_O \exp[-(1-\alpha)nfE]\}, \end{aligned} \tag{20}$$

where $i^{\rightarrow}$ and $i^{\leftarrow}$ are the anodic (oxidation) and cathodic (reduction) current densities (A/cm^2), C_R and C_O are the concentrations of reactant and product, respectively, $k^{\rightarrow}$ and $k^{\leftarrow}$ are the corresponding rate constants, E is the electrode potential versus a reference electrode, α is the anodic transfer coefficient, and $1-\alpha$ is the cathodic transfer coefficient. For a one-step single-electron transfer, α is equal to the symmetry factor β, which reflects the symmetry of the energy barrier (Fig. 6). When the slopes of the free-energy curves at their intersection are equal numerically but opposite in sign, then $\alpha=\beta=\frac{1}{2}$.

At equilibrium, the net current density is zero, and the anodic and cathodic currents are of equal magnitude. This current density at the reversible potential is termed the exchange-current density, i_0. If the potentials are taken relative to the reversible value, one obtains the current-density–overpotential equation (the Butler–Volmer equation):

$$i = i_0\{\exp(+\alpha nf\eta) - \exp[-(1-\alpha)nf\eta]\}, \tag{21}$$

where η is the overpotential $\eta = E - E_{\text{rev}}$. This is one of the fundamental relationships in electrochemistry. For $|\eta| > 0.118/n$ V the anodic or the cathodic current density becomes predominant and Eq. (21) reduces to one term. Overpotentials of 1 V are frequently

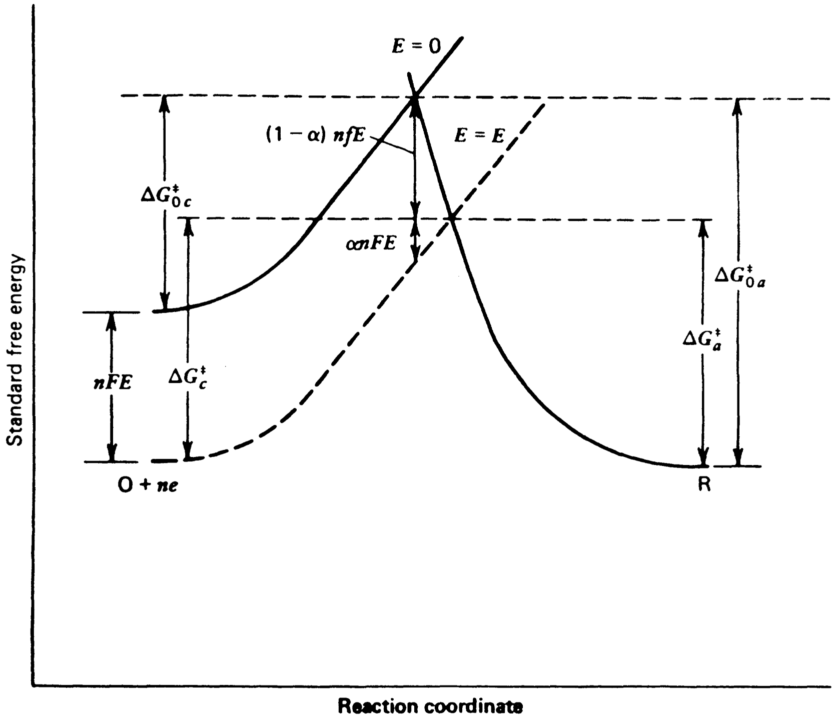

FIG. 6. A detailed picture of the change of free energy of activation caused by change of potential [reproduced from Bard and Faulkner (1980) with permission of John Wiley and Sons].

encountered experimentally. Over the 1.0-V range, the reaction rate may change by a factor 10^8. Such large changes in rates are a unique feature of electrochemical reactions. Such large changes for chemical reactions usually require extreme changes in temperature and pressure.

The double-layer structure can affect the kinetics of electrode reactions in two ways for a class I reaction. First, because of the presence of the diffuse ionic layer, the concentration of ions C_2 in the outer Helmholtz plane differs from the value of C_s in the bulk solution. Second, the overpotential η in Eq. (21) should be corrected for changes in the potential drop across the diffuse layer, $\phi_2 - \phi_s$. This effect is known as the Frumkin effect (Frumkin, 1933). With a concentrated supporting electrolyte and reactants and products of low charge (0 or ±1), these double-layer effects may be quite small. For class I reactions, corrections for these double-layer effects are relatively easily calculated, particularly when a large concentration of a supporting electrolyte is used to reduce the double-layer effects.

2.2 Theories of Electron Transfer

Theoretical descriptions of charge transfer (electrons or protons) in electrochemical reactions have been a subject of much interest since the work of Gurney (1931) who suggested that electron-transfer reactions at electrodes as well as between redox couples in solutions can be interpreted in terms of electron tunneling. Microscopic aspects of the charge transfer in the phenomenological treatment are contained in the rate constants k and the transfer coefficient α.

There have been two main approaches to the problem. In the first, which can be termed "molecular," attention is focused on the behavior of one chemical bond that is modified in the interfacial reaction. The theory assumes that the energy of activation is determined by the distribution of thermal energy in the various internal modes of the reacting species, but ignores the dynamic behavior and dielectric relaxational properties of the solvent (Horiuti and Polanyi, 1935; Bockris and Matthews, 1966; Christov, 1961).

The second approach, referred to as "continuum" theory, has been pursued by various theorists using classical and quantum statistical mechanics with a common focus on solvent dipole fluctuations as a major factor controlling the rate of charge transfer. According to the Franck–Condon principle, the electronic transition occurs in a time too short for changes in the nuclear positions and moments within the restraints of the uncertainty principle. The electron is transferred by tunneling without any change in the total energy. This is accomplished by thermal fluctuations in the solvent molecules surrounding the central ions. If, as a result of solvent fluctuations, the dipole distribution around the reacting particle is such that at the same time the energy is equal for the initial and final states, the quantum-mechanical tunneling of electrons from the metal to the reacting species can take place. Following the electron transfer, the system then relaxes to the equilibrium dipole distribution. Therefore, for simple reactions, the activation energy should be determined by the reorganization of the solvent dipoles required for quantum-mechanical electron transfer.

A particular feature of electrode reactions is that the electron energy level can be varied over a wide range by changing the electrode potential. Levich and Dogonadze (1959), using a full quantum-mechanical treatment, have shown that the variation of potential energy with displacement is parabolic, equivalent to a simple harmonic oscillation of the solvent. The solvent contribution to the potential energy is of the form

$$U_n = U_0 + \tfrac{1}{2}(h\nu_s) \sum (x_n - x_{n,0})^2, \qquad (22)$$

where x_n is a solvent coordinate, $x_{n,0}$ is the equilibrium value, U_0 is the energy when all x_n have the values $x_{n,0}$, and ν_s is the frequency of the solvent motion. The three-dimensional representation of Eq. (24) is given in Fig. 7. The electron transfer occurs at the lowest energy at which the paraboloids intersect. The intersection gives the same picture as that with the two coordinates also given in Fig. 7. The energy of activation is required for the movement of the system along the solvent coordinate from A to B. The Levich–Dogonadze–Kuznetsov treatment [see, e.g., Levich (1970)] for the heterogeneous reduction process [Eq. (21)] gives

$$i = \int \int c_0(x)\,[n(E)\rho(E)]\,W(E,x)\,de\,dx, \qquad (23)$$

where $c(x)$ is the concentration of the reacting ion at the distance x from the electrode, $\rho(e)$ is the electronic density of states in the metal, $n(E)$ is the Fermi distribution function, and $W(E,x)$ is the transition probability for the system to pass from the initial (i) to the final (f) state.

A simpler and easier-to-follow treatment of the role of solvent fluctuations in electron-transfer reactions at interfaces is due to Marcus (1964), who also gave much attention to the homogeneous redox reactions. He used only classical statistical thermodynamics. The theory gives an expression similar to Eq. (23) but based on free-energy considerations. The Marcus treatment works reasonably well for homogeneous electron-transfer reactions involving a single-electron transfer and to a somewhat lesser extent for heterogeneous redox reactions (Albery, 1975).

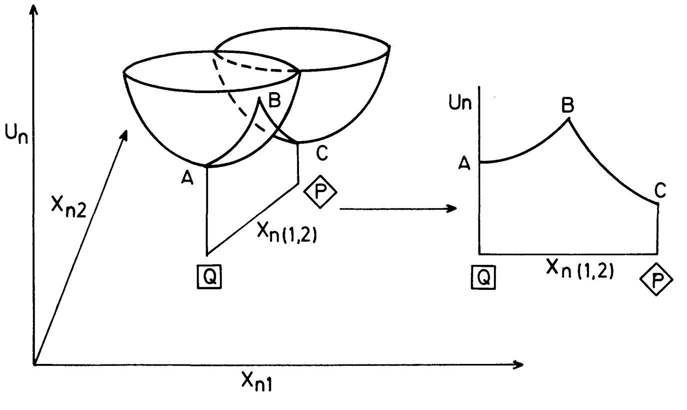

FIG. 7. Three-dimensional representation of Eq. (14). The electron transfer occurs at the lowest energy at which the two parabolic bowls intersect [reproduced from Albery (1975) with permission of Clarendon Press].

Neither Marcus nor the Levich–Dogonadze group in their initial work considered the contributions of the vibrational modes of the inner coordination sphere ligands to the free energy of activation. Marcus rather quickly recognized the importance of this contribution, which is usually considerably larger than that of the solvent continuum, and added a term to the activation free energy to take this contribution into account. The Levich–Dogonadze group felt that the vibrational modes associated with the nearest neighbors of the reacting ions would require far too large a quantum of energy to be excited at room temperatures. This was not true, and in more recent years this contribution to the reorganizational energy has been taken into account.

For a heterogeneous one-electron transfer from the metal electrode to a redox species in solution, the Marcus treatment indicates the potential dependence of the rate constant to have the form

$$i=\kappa(kT/2\pi\rho)^{1/2}c\,\exp(-\Delta G^{*}/RT), \tag{24}$$

$$\begin{aligned}\Delta G^{*}=&(W_R+W_p)/2+(\lambda_i+\lambda_0)/4\\&+(E-E_r^0)e/2+(eE-eE_r^0+W_P\\&-W_R)^{2/4}(\lambda_i+\lambda_0),\end{aligned} \tag{25}$$

where ρ is the reduced mass in the activated state, λ_0 and λ_i are reorganizational energy terms, W_P and W_R are the work to the bring reactants and products to the most probable charge-transfer distance (transmission probability), E and E_r are the electrode potential and the reversible value, and κ is the electron tunneling probability.

At low overpotentials the last term in Eq. (25) is small. In this case Eq. (25) gives $\alpha=\frac{1}{2}$ and a linear η-log$i$ curve with a slope (Tafel slope) of $2RT/F$. At higher overpotentials $\alpha\rightarrow1$ and the curve departs from linearity. The experimental results of simple redox reactions confirm the predictions of Eqs. (24) and (25) in regard to $\alpha=\frac{1}{2}$ and the existence of barrierless and activationless processes (Krishtalik, 1983). The finding of a linear η-log$i$ curve as evidence in support of the theoretical treatments has been questioned by Bockris and Khan (1979) on the grounds that most electrode reactions give linear η-log$i$ plots over a wide range of overpotentials, with a slope of 0.12 V/decade at 25 °C, corresponding to $\alpha=\frac{1}{2}$, even when the reactions do not meet the conditions for Eqs. (24) and (25) to be applicable. The Marcus theory, however, also has yielded interesting interrelations between the rate constants for similar homogeneous and heterogeneous reactions.

2.3 Theories of Proton Transfer

Of special interest to electrochemists is the discharge of the hydronium ion to form H_2 gas at the cathode surface. The rate-determining step (rds) on *sp* metal surfaces such as mercury is the formation of adsorbed H; i.e.,

$$M+H_3O^{+}+e^{-}\rightleftarrows M-H_{ads}+H_2O. \tag{26}$$

This reaction is then followed by either a recombination reaction

$$2M-H_{ads}\rightarrow 2M+H_2 \tag{27}$$

or an electrochemical desorption reaction

$$M-H_{ads}+H_3O^{+}+e^{-}\rightleftarrows M+H_2+H_2O. \tag{28}$$

The idea of a quantum-mechanical model involving tunneling for the proton transfer can be traced back to Gurney (1931). Models based on Gurney's approach assume the existence of a classical distribution of states in the reacting species and the surrounding solution. On the basis of the changes in the isotopic separation factors for H_2 and D_2 evolution on the Hg electrode over the wide potential range in which the transfer coefficient α is constant, Bockris and Matthews (1966) concluded that up to 70% of the observed current is due to proton tunneling at low overpotentials. Proton transfer takes place from the H_3O^{+} ion in the OHP to an adsorbed state. Bockris and Matthews (1966) used an asymmetric Eckart potential-energy barrier for the tunneling. The calculation of the potential-energy profile for this model is highly dependent on the models for H_2O and H_3O^{+} at the electrode surface and particularly the potential-energy contours in reaction-coordinate space as discussed by Marcus (1964). In the tunneling calculation, the assumption is made that the initial and final states do not fluctuate during the tunneling. This work has been reviewed by Bockris and Kahn (1979).

A quantum-statistical-mechanical treatment of proton transfer has been carried out by Dogonadze *et al.* (1968). They extended the formalism of outer-sphere electron transfer to proton transfer. The reaction is considered to take place as a result of a continuum of solvent polarization fluctuations. These theorists assumed that activation by stretching of the $(\text{H–OH}_2)^+$ and *M*–H bonds is sufficiently improbable to be neglected on the basis that $h\nu \gg kT$, where ν is the frequency of the vibrational modes involved in the reacting species. They take into account only the outer-sphere reorganization of the solvent and assume that activation can only occur because of nonequilibrium solvent polarization in this treatment.

This treatment reveals three distinct potential regions with different transfer coefficients, viz., $\alpha=1$ for low η (difficult to check experimentally), $\alpha \approx \frac{1}{2}$ for intermediate η (supported by experiments), and at high η a proposed barrierless region where $\alpha=0$, for which Krishtalik (1983) claims experimental evidence support.

A particularly questionable feature of this model is the assumption that the activation occurs only through the phonon interactions in a solvent volume large compared to that of H_3O^+, and particularly the neglect of internal modes involved directly in the proton transfer (H–OH_2 and $M\text{–H}_{\text{ads}}$). Kuznetsov (1987) has further developed this model by suggesting a charge-variation model for electrochemical discharge of proton donors. The model takes into account the variation of charge of the adsorbed atom in the activation–deactivation process. Experimental isotopic effects with various metals are in line with the theory.

2.4 Electrosorption Reactions

Adsorption of species from solution on an electrode surface is a replacement process in which usually the solvent is replaced by the adsorbate. The solvent molecules, however, will also interact with the adsorbate. The adsorption depends on the interaction energies at the electrode surface, the composition of the electrolyte, and the electrode potential. A strong interaction of the adsorbed species with the electrode surface usually involves direct orbital interactions with the electrode and is referred to as chemisorption. For many ions, especially anions, the adsorption energy exceeds that expected on the basis of just electrostatics. Such adsorption is referred to as specific ionic adsorption. Chemisorption is often found also with atoms and molecules and on surfaces in general.

For the ideal polarized interface, there is no electrochemical reaction occurring at a significant rate. For such an electrochemical interface the sum of the charge on the metal electrode, σ_M, and on the solution side of the interface, σ_s, must be zero because of the neutrality restriction; therefore, $\sigma_M=\sigma_s$. The charge density (charge per unit area of the electrochemical interface) can be calculated from measurements of the interfacial capacitance per unit area (C) as a function of the potential of the metal electrode E relative to a reference electrode whose potential is constant; i.e.,

$$\sigma_M=\int_{E_{\text{PZC}}}^{E} C\,dE, \tag{29}$$

where E_{PZC} corresponds to the PZC at which $\sigma_M(\text{PZC}) = 0$. The charge density on the metal can also be determined from the interfacial surface tension γ. From the well-known Gibbs adsorption equation, it can be shown that

$$\sigma_M=\left(\frac{\partial\gamma}{\partial E}\right)_{\mu,p,t}, \tag{30}$$

where E is the potential of the metal relative to a reference electrode, μ is the chemical potential of any solution phase component, p is the pressure, and t is the temperature. The derivative $(\partial\gamma/\partial E)_{\mu,p,t}$ is relatively easily obtained with a liquid-metal electrode such as mercury but is somewhat more difficult to obtain for a solid metal electrode. At the PZC, σ_M and hence $(\partial\gamma/\partial E)$ go to zero and γ passes through a maximum.

2.4.1 Partial Charge Transfer During Specific Ionic Adsorption.

When orbital overlap occurs between the electron orbitals of the adsorbed ion and the metal electrode, the bonding of the ion to the electrode is not purely electrostatic and partial charge transfer is expected. Lorenz and Salie (1961) proposed that

$$x^z \rightarrow x_{\text{ads}}^{z(1-\lambda)} + \lambda z e^-, \tag{31}$$

where λ is the partial-charge-transfer coefficient. If the adsorption is purely electrostatic, $\lambda=0$, while if the adsorption is almost entirely through covalent bonding, $\lambda=1$. The term λ cannot be determined by just measuring the charge through the external circuit since the double-layer charge also is involved.

Lorenz and Salie (1961) have attempted to derive equations relating λ to experimentally measurable quantities, principally the high-frequency differential capacitance of the interface and the excess charge on the metal when $\lambda = 0$. Despite serious questions and criticisms of the Lorenz approach to the measurement of λ, the qualitative concept of partial charge transfer between the chemisorbed ion and the metal has gained some acceptance.

An alternative approach proposed by Vetter and Schultze (1972) defines the concept of "electrosorption valency," given by the relation

$$\gamma' = -\frac{1}{F}\left(\frac{\partial \sigma_m}{\partial \Gamma}\right)_{\Delta\phi} = \frac{1}{F}\left(\frac{\partial \mu_{\mathrm{ad}}}{\partial \Delta\phi}\right)_{\Gamma}, \qquad (32)$$

where σ_m is the charge provided by the external circuit, $\Delta\phi$ is the potential drop $\phi_M - \phi_2$, and Γ is the surface excess of the absorbate per unit area of the interface.

The quantities in the first term of Eq. (32) are all directly measurable. Experimentally, Schultze and Vetter (1973) obtained, for Hg at $\Delta\phi = 0$, the values $\gamma'_{\mathrm{Cl}^-} = -0.16$, $\gamma'_{\mathrm{Br}^-} = -0.33$, and $\gamma'_{\mathrm{I}^-} = -0.40$. Despite some approximations and estimations the electrosorption valency clearly provides evidence for partial charge transfer. The quantum-mechanical calculations of Kornyshev and Schmickler (1985) support the partial-charge-transfer concept while calculations based on the "jellium" model by Lang (1983) do not. The latter is not surprising since the jellium model does not handle properly the localized covalent binding of the adsorbed ion with the electron orbitals of the metal.

3. ELECTROCHEMISTRY ON WELL-DEFINED SINGLE-CRYSTAL SURFACES

Very substantial progress has been made in the last several years in studying electrochemical reactions on single-crystal surfaces. This parallels the progress in surface science and the surface characterization of solid–vacuum and solid–gas interfaces. Double-layer properties, electrosorption reactions, and the kinetics and pathways of electrocatalytic reactions have been found to be dependent on the crystallographic orientation of the electrode, i.e., on the local symmetry of surface atoms. Single-crystal electrode studies are essential to an understanding of the various processes occurring at electrode surfaces.

3.1 Double-Layer Capacitance Studies

A large body of experimental data on the double-layer properties of single-crystal electrode surfaces exists for the *sp* metals gold, silver, and copper but only one *d* metal, iron. Striking differences in the double-layer properties of the low-index planes have been found for Au, Ag, Bi, Zn, and Sn in weakly adsorbed electrolytes as well as with specifically adsorbed halides (Hamelin, 1985). The potential of zero charge differs by several hundred millivolts from one face to the other, indicating that the kinetics of various electrochemical reactions should also be strongly affected by the crystallographic orientation of the electrode surface. For example, the PZC values for Ag(111), Ag(100), and Ag(110) are −0.45, −0.67, and −0.77 V, respectively. This indicates that even the concept of the PZC may not be valid for a polycrystalline electrode.

Note should be taken, however, of the strong possibility that restructuring of the surface occurs after the metal electrode is introduced into the electrolyte, particularly with the more active metals such as Bi, Zn, Sn, Cu, and Fe.

The linear relationship between the work function and the PZC values holds also for single-crystal electrodes* [see Eq.(17)]. The close-packed planes have higher work functions and more positive PZC values. *In situ* studies of the vibrational properties of water molecules in the compact layer on single-crystal surfaces would help to advance the understanding of the compact layer and the symmetry elements.

---

*The work function of a phase *a* is defined as the work necessary to transport an electron from within the phase to an infinite distance away *in vacuo* (vacuum reference) when there is no net charge on the surface through which the electron is removed. Since the potentials of zero charge differ for the various crystal surfaces, their work functions will also differ.

3.2 Electrosorption of Hydrogen and Oxygen on Noble-Metal Electrodes

The electrosorption of hydrogen occurs on platinum metal in a reversible manner at potentials more positive than the reversible potential for hydrogen gas evolution. This is an example of so-called underpotential deposition (UPD) (Sec. 3.3). Hydrogen adsorption can be represented by the reaction (26). A monolayer of adsorbed hydrogen forms before H_2 evolution commences, according to reaction (27). Anion adsorption has a substantial effect on the hydrogen adsorption reaction. Hydrogen adsorption exhibits a pronounced dependence on the crystallographic orientation of the electrode surface. This has been established with well-ordered single-crystal surfaces, particularly platinum. Recent data for surfaces prepared in ultrahigh-vacuum (UHV) systems, characterized with low-energy electron diffraction (LEED) and Auger electron spectroscopy (AES), and transferred to a thin-layer electrochemical cell [see Yeager *et al.* (1978) and Ross (1979)], agree reasonably well with the data obtained by a special flame annealing method of Clavilier *et al.* (1980).

Linear-sweep voltammetry curves are shown in Fig. 8 for the adsorption and desorption of hydrogen atoms on the three low-index planes of platinum in sulfuric acid. The potential is linearly swept in the positive and then the negative directions, and the current is recorded as a function of the electrode potential. By integrating the current–voltage curves, it is possible to calculate the adsorption iso-

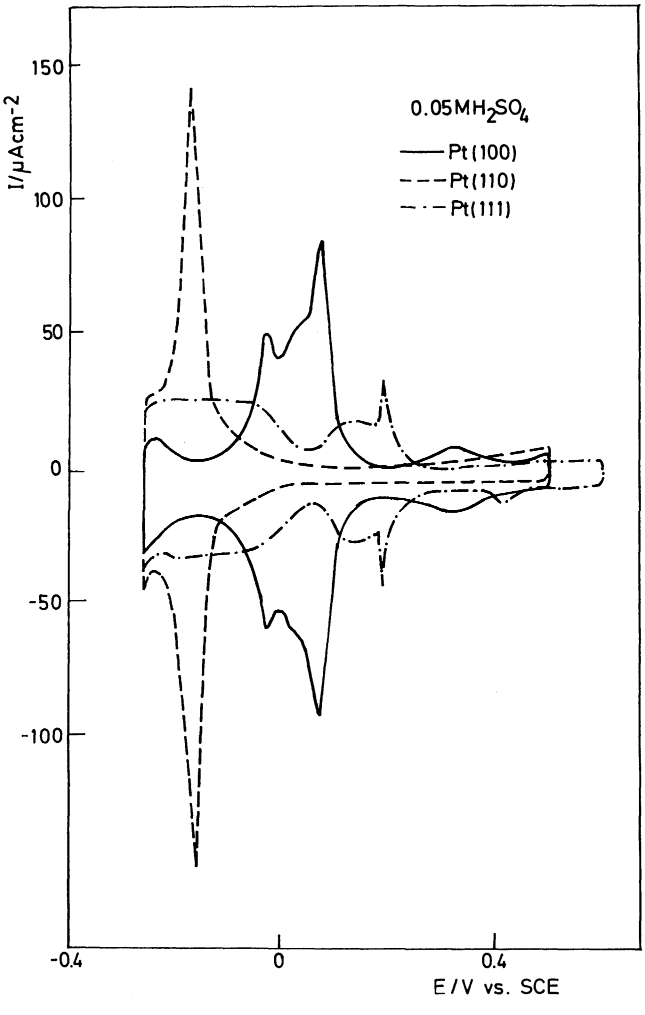

FIG. 8. Electrosorption of H on Pt single-crystal electrodes in 0.50*M* H_2SO_4. Sweep rate 50 mV s^{-1}. The potential scale is given vs saturated calomel reference electrode [adapted from Markovic *et al.* (1988) with permission of the Americal Chemical Society].

therms (surface coverage versus electrode potential) provided the electrosorption valence is known. The strong dependence of the electrochemical properties on the crystallographic orientation of the electrode surface is also supported by results obtained on high-index stepped surfaces which show a dependence on the step orientation and step density.

The interpretation of the hydrogen adsorption curves for platinum is a matter of some controversy. The behavior of the Pt(111) surface is less understood than the other two low-index planes—in particular, the more positive peak in Fig. 8. The exact location of the hydrogen atoms on the platinum surfaces is not yet well established. This was a matter of controversy for many years in gas-phase studies (Christmann, 1988). The infrared studies of Bewick *et al.* (1988) provide some evidence that on the Pt(111) and polycrystalline Pt surfaces, the bonding of H with Pt may be similar to that in metal hydrides.

The adsorption of *M*-OH and *M*-O occurs on noble metals such as platinum at potentials much less positive than the potential for oxygen evolution and leads to the formation of submonolayers or monolayers of these species. Thicker layers can be formed at such potentials on most less noble metals, such as ruthenium. The reaction is very irreversible except at the onset of the oxidation. It is also strongly dependent on the crystallographic orientation of the surface. The differences in the voltammetry curves in the early stages of the formation of the *M*-OH and *M*-O layers have been proposed by Angerstein-Kozlowska *et al.* (1987) to be due to the different strengths of adsorption and the extent to which these processes are coupled with the desorption–adsorption process of anions having tetrahedral structure (SO_4^{2-}, ClO_4^-).

3.3 Formation of Metal Monolayers

In the electrodeposition of metals onto foreign metal substrates, e.g., Pb on Pt, the formation of a monolayer of metal may take place at potentials more positive than the potential required for the reversible deposition. This phenomenon is another illustration of underpotential deposition (UPD) (see Sec. 3.2). This occurrence of UPD implies that the chemical potential of the adatoms deposited in the UPD layer (μ_{ads}) is different from that of the corresponding bulk metal (μ_{bulk}). The adatoms are more strongly bound to the surface of the foreign substrate than to the surface of their own species. The difference in the electrode potentials for reversible UPD and reversible deposition of the bulk metal corresponds to the differences in chemical potentials

$$[\mu(\theta)]_{\mathrm{ads}}-\mu_{\mathrm{bulk}}=-ze[E(\theta)_{\mathrm{ads}}-E_r] \\ =-ze\Delta E_p, \qquad (33)$$

where E_r is the reversible electrode potential for the bulk metal and $E(\theta)_{\mathrm{ads}}$ that for the UPD layer at the coverage θ. The difference $[E(\theta)_{\mathrm{ads}}-E_r]$ is a measure of the binding energy of the UPD layer to the foreign metal substrate. The electrosorption valency has been taken to be close to the Faradaic valence.

On the basis of data for polycrystalline electrodes, Kolb *et al.* (1974) found a linear relationship between the UPD effect at $\theta=0.2$ and the difference in work functions of the substrate and the adsorbate bulk metal, of the form $\Delta E_p(\theta=0.2)=s\Delta\Phi$, where $s\cong\frac{1}{2}$. This correlation does not seem to be applicable in this simple form for single-crystal electrodes (Bewick and Thomas, 1975). A necessary condition for UPD to occur is $\Phi_{\mathrm{substrate}}>\Phi_{\mathrm{adsorbate}}$. Trasatti (1975) found a linear correlation also between $\Delta E_p(\theta=0)$ and $\Delta\Phi$ by taking the potential onset ΔE_p at the UPD, i.e., $\theta=0$. The electrosorption valency of foreign adatoms is in the majority of the cases close to the Faradaic valence, i.e., only a small charge may be retained on the adsorbate. Table 1 gives a list of the relative binding energies and electrosorption valencies for various pairs of substrate and adsorbate metals.

Figure 9 shows the voltammetry curves for the formation and dissolution of a monolayer of Pb on gold single-crystal electrodes. Adžić *et al.* (1974) have shown that the UPD of metals strongly depends on the crystallographic orientation of the electrode surface. The coverage of foreign adatoms is determined by the electrode potential and concentration of cations in solution. The properties of adatoms have been studied at a constant potential by various *in situ* optical techniques, scanning tunneling and atomic force microscopies, and *ex situ* surface-physics techniques (Sec. 5.3). Kolb (1978) has reviewed the work on the UPD of metals.

Various overlayer lattices have been identified for different coverages of adatoms cor-

Table 1. Relative binding energies and the electrosorption valencies for some metal adatoms on various polycrystalline substrates.[a]

System, substrate/ adatom	Relative binding-energy valence, $ze_0\Delta E_p$ (eV)	Electrosorption	
		γ^1	γ^1/z
Pt/Pb	1.74	2	1
Pt/Bi	1.77	3	1
Pt/Cu	0.82	1.8	0.9
P/Tl	1.3	0.85	0.85
Pt/Cd	1.92	0.3	0.15
Au/Pb	0.80	1.95	0.97
Au/Tl	0.69	1	1
Au/Cu	0.44	2	1
Au/Bi	0.75	3	1
Ag/Pb	0.160	1.9	0.95
Ag/Tl	0.28	1	1

[a]Relative binding energies obtained from ΔE_p at $\theta \sim 0.1$–0.2, and electrosorption valences obtained at $\theta \sim 0.1$. Reproduced from Adžić (1984) with permission of John Wiley and Sons.

responding to the multiple peaks observed in many systems. Half-widths of the voltammetry peaks smaller than (90 mV)/z indicate attractive interaction between adatoms, while larger values indicate repulsion or irreversibility. The sharp UPD peaks observed with some systems at certain coverages have been interpreted by Adžić *et al.* (1974) in terms of structural phase transitions in the UPD layer, but attractive terms in the adsorption isotherm may also result in narrow peaks.

The UPD of metals offers a unique opportunity to modify in a controlled way the electronic and catalytic properties of electrode surfaces by changing the adatoms and their coverage from 0 to 1. Large catalytic effects occur with foreign adatoms in the oxidation of organic molecules and oxygen reduction on some noble metals; see the review by Adžić (1984) and Sec. 3.4.

3.4 Electrocatalytic Reactions

Electrocatalytic reactions involve the strong interaction of reactants and/or intermediates with the electrode surface and usually are very dependent on the nature of the electrode material. A good example is H_2 evolution, for which the exchange-current density of various metals varies by 11 orders of magnitude. Therefore the subject of electrocatalysis is principally concerned with the role of the electrode material in determining the kinetics of electrode reactions and is a subpart of heterogeneous catalysis. Electrocatalysis is a complex area since electrocatalytic reactions are often multistep reactions which are additionally complicated by the role of the electrode material and its surface properties. In general, a detailed understanding at a molecular level of the elementary steps in electrocatalytic reactions has not yet been attained. Recent advances in the *in situ* identification of reaction intermediates and products and the proper characterization of electrode surfaces are improving this situation (Sec. 5).

Many electrocatalytic reactions have great technological importance and an even larger number have the potential for such. Consequently, they have been studied in detail. These include hydrogen, oxygen, and chlorine evolution; oxygen reduction; the oxidation of small organic molecules suitable for energy conversion (e.g., alcohols, hydrocarbons, carbon monoxide); and various organic reactions for synthetic purposes. Electrocatalytic studies can be divided into studies of the electronic factor, which depends on the nature of the electrode material, and the structural factor, including the symmetry of surface sites [see Bockris and Reddy (1970), Yeager (1981), and Appleby (1983)].

3.4.1 Electronic Factors. Much effort has been directed to the correlation of the kinetics of the H_2 evolution reaction with the nature of the electrode material. Proposed mechanisms for hydrogen generation on various metals are still open to some questions.

It is believed that for Pt and platinum metals the following mechanism is operative:

$$H_3O^+ + e^- \rightleftarrows H_{ads} \quad \text{(fast)}, \tag{34}$$

$$2H_{ads} \rightarrow H_2 \quad \text{(rds)}. \tag{35}$$

On the other hand, for metals such as Hg and Pb, reaction (34) is the rds, while for Au and Ni the rds is the reaction (Bockris and Reddy, 1970)

$$H_{ads} + H_3O^+ + e^- \rightarrow H_2 + H_2O. \tag{36}$$

The kinetics of these electrocatalytic reactions have been correlated with the free energies of adsorption of the reactants. This results in the so-called "volcano" plots, discussed for heterogeneous catalysis by Balandin (1958) and by Parsons (1958) and Gerischer (1958)

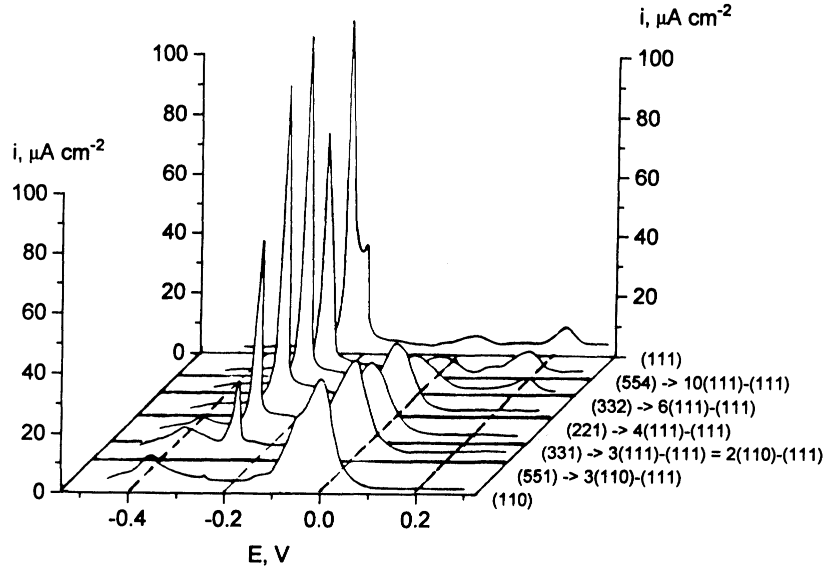

FIG. 9. UPD (anodic sweeps) of Pb on Au single-crystal electrodes in 10 m*M* $HClO_4$+1 m*M* $Pb(ClO_4)_2$ [reproduced from Hamelin and Lipkovski (1984) with permission of Elsevier Sequoia Press].

for electrocatalysis. Figure 10 shows a plot of logi_0 vs *M*–H bond strength calculated from the overvoltage data of Krishtalik (1970) (i_0 corresponds to the exchange-current densities for the overall reaction). For the left-hand branch of the volcano curve, the rate of reaction increases with the *M*–H bond strength for nontransition metals, for which the reaction (34) is the rds. For the right-hand branch, the rate decreases with the bond strength for transition metals, for which reaction (35) and, in some cases, reaction (36) are the rds. According to Khan (1985), these volcano plots support the bond-stretching model for the proton transfer (Sec. 2.3).

The oxygen evolution and reduction reactions also are electrocatalytic reactions that are of critical importance for a number of applied aspects of electrochemistry. The mechanisms of these reactions are still uncertain

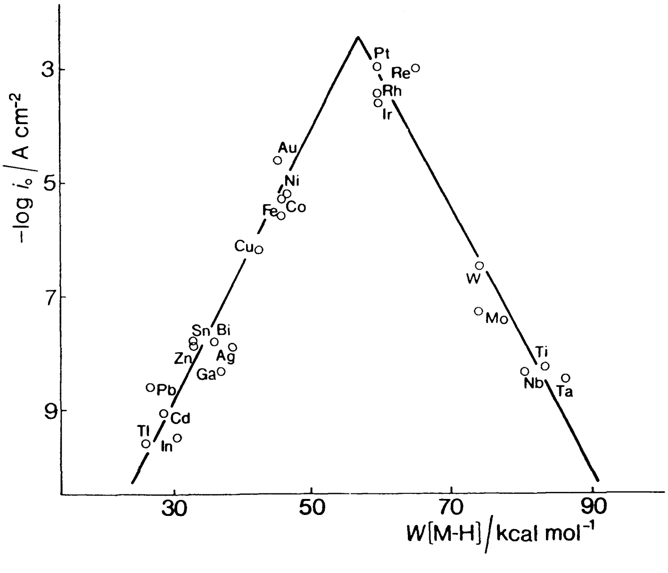

FIG. 10. Equilibrium rate for hydrogen evolution as a function of *M*–H bond strength [reproduced from Trasatti (1977) with *M*–H from Krishtalik (1970) with permission of John Wiley and Sons].

on most electrode surfaces, despite extensive studies. This is a consequence of the pronounced irreversibility at moderate temperatures and the fact that the reaction may proceed through a large number of pathways. Furthermore, the pathways for the O_2 reduction and generation are usually quite different and not just the reverse of each other. The range of potentials over which these two processes occur on a given electrode are widely separated from each other, and consequently the states of the electrocatalytic surfaces also are usually quite different for the O_2 reduction and generation processes.

The reduction of O_2 in aqueous solution, depending on the electrode material, appears to occur by two pathways overall. For acid solutions, these are

$$O_2+4H^+ +4e^- \rightarrow 2H_2O \tag{37}$$

and

$$O_2+2H^+ +2e^- \rightarrow H_2O_2. \tag{38}$$

The four-electron pathway may involve H_2O_2 as an intermediate, while for the two-electron pathway H_2O_2 is a product (Yeager, 1986). In general, little H_2O_2 is observed on Pt-group metals as a reaction product. As the peroxide builds up in concentration in the electrolyte, it undergoes further reduction,

$$H_2O_2+2H^+ +2e^- \rightarrow 2H_2O, \tag{39}$$

and/or decomposition,

$$2H_2O_2 \rightarrow 2H_2O+O_2. \tag{40}$$

The extent to which each of these pathways is involved at a particular electrode surface is usually established with the rotating-disk-ring electrode. Both the four-electron reduction and the reduction beyond the peroxide stage require the rupture of the O–O bond. A sufficiently strong interaction with a proper surface spacing of adsorbent sites is necessary for these reactions to occur (Terasevich *et al.*, 1983). A volcanolike relationship has been reported in plots of η vs ΔG_{ads} at constant current for oxygen reduction in acid solutions (Appleby, 1983). Electrocatalysis for oxygen reduction involves a number of materials, which are listed in Table 2 (Yeager, 1986).

Table 2. Electrode surfaces for O_2 reduction in aqueous solutions.[a]

Class I	Reduction predominantly through HO_2^--formation 1. Graphite, most carbons 2. Metals: gold, mercury, most oxide-covered metals (e.g., Ni, Co) 3. Oxides: most transition-metal oxides (e.g., NiO, spinels) 4. Transition-metal macrocycles: e.g., Co-TsPc, Co-TMPP
Class II	Reduction by both $4e^-$ and peroxide pathways 1. Platinum, platinum alloys, platinum-family metals 2. Other metals: silver, palladium 3. Oxides: pyrochlores (e.g., lead ruthenate) 4. Some transition-metal macrocycles (e.g., dicobalt face-to-face porphyrins, dicobalt pyridyl planar complexes, Fe-TsPc, Fe-TPyPz)
Class III	Heterogeneous peroxide decomposition catalysts 1. Metals: silver, platinum 2. Oxides: spinels, perovskites, MnO_2 3. Transition-metal macrocycles

[a] Reproduced from Yeager (1986) with permission of Elsevier Sequoia Press.

An interesting class of catalytic materials for O_2 reduction comprises the transition-metal macrocycle complexes attached to high-surface-area carbon electrodes. Their activity, but not stability, may come close to the activity of highly dispersed platinum, the best catalyst for O_2 reduction. They provide an opportunity to study the interaction of O_2 molecules with the transition metals. The various types of interactions are shown in Fig. 11 in terms of the molecular orbitals of the O_2 and the transition metal in the complex (Yeager 1986). The side-on and bridge-type interactions should be conductive to the dissociation of O_2 and hence the direct four-electron reduction. There is evidence for such with some complexes.

The oxidation of small organic molecules is of interest for electrochemical energy conversion in fuel cells. The most studied reactions have been methanol and formic acid oxidation. Volcano relations have been observed for formic acid and ethylene oxidation. The reactions are slow, even with the best catalyst, platinum. Strongly bound intermediates block the catalyst sites and make the surfaces inactive. The modification of the platinum surface

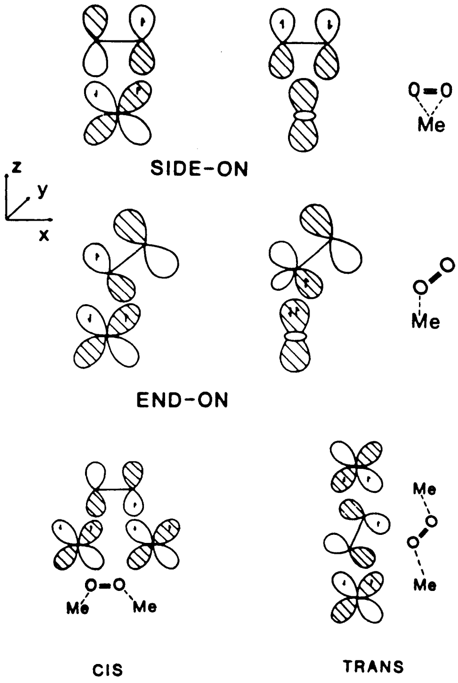

FIG. 11. Various types of interactions of O_2 with transition-metal species [reproduced from Yeager (1986) with permission of Elsevier Sequoia Press].

by alloying, UPD of foreign atoms, and the designing of bifunctional catalysts do increase somewhat the kinetics of these reactions. These modifications prevent the formation of high coverage of such intermediates as CO or facilitate their oxidation at potentials less positive than those required on platinum (Adžić, 1984; Parsons and VanderNoot, 1988).

3.4.2 Structural Factors in Electrocatalysis. The importance of the surface structure was controversial in heterogeneous catalysis as a whole, including electrocatalysis, for many years. During the last ten years or so, the results obtained with well-ordered single-crystal surfaces have demonstrated clearly the importance of structural factors. The coordination of surface atoms, interatomic distances, distance between adsorption sites, and surface energies all depend on the plane exposed. For a typical electrocatalytic reaction, the adsorption of reactants and intermediates, adsorbate surface migration, surface chemical reactions, and product desorption depend on the atomic arrangement of the electrode surface and not just on the electronic properties.

The oxidation of formic acid on single-crystal platinum surfaces is a good illustration of a striking structural effect. Linear-sweep voltammetry curves for the low-index planes of platinum are shown in Fig. 12. The reaction proceeds through two pathways after the rupture of the C–H bond:

$$\mathrm{HCOOH} \rightarrow \text{-}\mathrm{COOH}_{\mathrm{ads}} + \mathrm{H}^+ + e^-, \tag{41}$$

$$\text{-}\mathrm{COOH}_{\mathrm{ads}} \begin{cases} \nearrow \mathrm{CO}_2 + \mathrm{H}^+ + e^- & (42) \\ \searrow \text{strongly bound species.} & (43) \end{cases}$$

This strongly bound species is a matter of some controversy. On the basis of *in situ* Fourier-transform infrared spectroscopic measurements, it appears that the species is linearly bonded and bridge-bonded CO, as has also been found to be formed during the oxidation of CH_3OH. This species causes a complete blocking of the Pt(100) surface in the anodic scan. The least adsorption of CO takes place on the Pt(111) surface, which as a consequence appears to be the most active surface (see Adžić, 1990).

A strong structural dependence is characteristic of O_2 reduction on Au(*hkl*). In alkaline solutions not only the rate but also the reaction mechanism depends on the crystallographic orientation of the electrode surface. The four-electron reduction occurs on the Au(100) surface, but only the two-electron reduction takes place on Au(110) and Au(111) surfaces. The Au(100) surface is the best catalyst for O_2 reduction in alkaline solutions (Adžić *et al.*, 1984).

Many organic reactions, such as the oxidation of methanol, formaldehyde, ethylene glycol, glucose, and others, depend strongly on the crystallographic orientation of the surface. Hydrogen evolution in general has only a small dependence on the surface structure, as well as oxygen reduction on platinum [see the reviews of Adžić (1990) and Parsons (1990)]. The structural effects can be very important and as large as the electronic effects. Current research with stepped surfaces will help to rationalize the concept of "active sites." The data with stepped surfaces show that each crystallographic orientation gives the electrode surface a distinct electrochemical identity. This suggests that most of studies

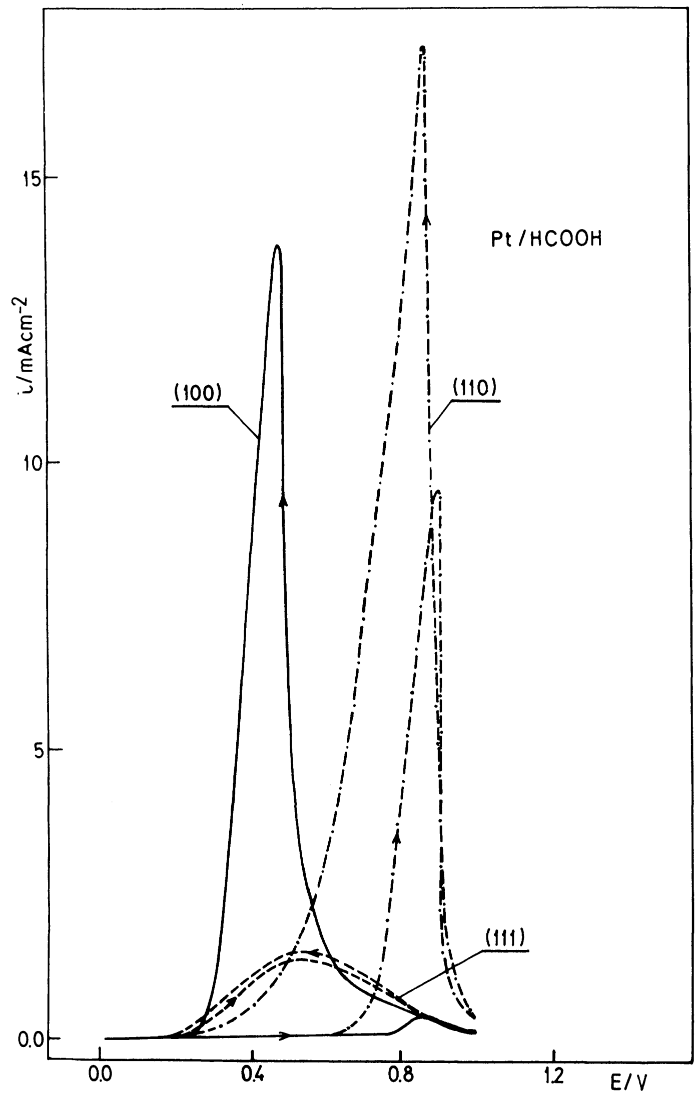

FIG. 12. Oxidation of HCOOH on Pt single-crystal electrodes in 0.1*M* $HClO_4+0.3M$ HCOOH. Sweep rate 50 mV s^{-1} [reproduced from Adžić (1990) with permission of Plenum Press].

in electrocatalysis should involve single-crystal surfaces.

4. SEMICONDUCTOR ELECTRODES

4.1 Space-Charge Effects

In general, the electrochemical properties of semiconductor electrodes are the result of a low concentration of free electrons and the presence of two types of charge carriers, conduction-band electrons (negative) and holes (positive) in the valence bands. The excess charge in the semiconductor is not located just within a fraction of an angstrom from the surface, as it is in a metal. Instead, its distribution extends into the bulk, thus making a space-charge region like the diffuse part of the ionic double layer at a metal–solution interface. The electrode potential at which the space charge in the semiconductor is zero corresponds to the PZC. For semiconductor electrodes it is called the flat-band potential, E_{fb}, since under these conditions there is no electric field and no band bending. Figure 13 shows the energy levels of the band edges and potential distributions for an *n*-type semiconductor in contact with an electrolyte for (a) no excess charge corresponding to $E=E_{fb}$, (b) positive excess charge (electron depletion); and (c) negative excess charge (electron accumulation).

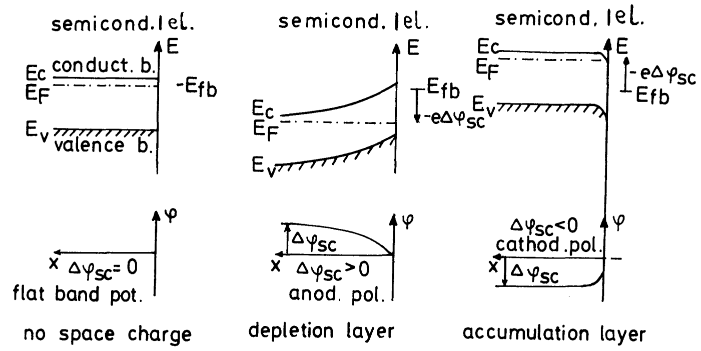

FIG. 13. Effects of space charge in an *n*-type semiconductor on the course of band edges and electric potential [adapted from Gerischer (1990) with permission of Pergamon Press].

Differential-capacitance measurements provide information on the charge distribution across the interface. When specific adsorption does not occur, the overall capacitance of the interface C is given by

$$C^{-1}=(C_{sc})^{-1}+(C_{-2})^{-1}+(C_{dif})^{-1}, \qquad (44)$$

where C_{sc} is the space-charge capacitance in the semiconductor and C_{-2} and C_{dif} are the capacitances of the outer Helmholtz layer and the diffuse ionic layers on the electrolyte side of the interface, respectively. For low and moderate carrier concentrations ($n_0<10^{16}$ carriers/cm^3) the predominant term on the right-hand side of Eq. (44) is $(C_{sc})^{-1}$ and hence $C \cong C_{sc}$.

For an extrinsic semiconductor the space-charge capacitance, C_{sc}, is given *mutatis mutandis* by Eq. (6) for the metal–diffuse-layer capacitance. For an extrinsic (e.g., *n*-type) semiconductor under depletion-layer conditions [Fig. 13(b)], C_{sc} is given by the Mott–Schottky equation, which was first derived for the metal–semiconductor interface:

$$1/C_{sc}^2=[2/(e\epsilon\epsilon_0 N_D)])(-\Delta\phi-kT/e), \qquad (45)$$

where $\Delta\phi$ is the potential drop across the space-charge layer and N_D is the density of donors. From the linear plots of $1/C_{cs}^2$ vs E one obtains E_{fb} and the doping level N_D, since $-\Delta\phi=E-E_{fb}$.

Specific ion adsorption has a pronounced effect on semiconducting oxides. The potential drop in the inner Helmholtz layer varies with the *p*H of the solution because of the interaction of the surface with H$^+$ and OH$^-$ ions.

4.2 Electrons, Holes, and Surface States in Electrode Reactions

A special feature of the semiconductor–electrolyte interface is that the charge transfer can occur to energy levels in the conduction band and in the valence band of the semiconductor. This is distinctly different than with metal electrodes, for which only the electronic levels near the Fermi level are involved. Electron transfer across the semiconductor–electrolyte interface can be treated using the theories of the electron-transfer discussion earlier in Sec. 2.2. When any change in the electrode potential occurs entirely across the space-charge layer, $\Delta\phi_{sc}$ and with not too high a doping level, then according to Gerischer (1958, 1990), the electron-transfer kinetics can be described by the current density via the conduction band:

$$i_c=k_c^+N_cN_{red}-k_cn_sN_{ox}, \qquad (46)$$

where $n_s=n_0\exp(e\Delta\phi_{sc}/kT)$ is the surface concentration of electrons and n_0 is the bulk concentration. The current density via the valence band is

$$i_v=k_v^+p_sN_{ox}-k_v^-N_vN_{ox}, \qquad (47)$$

where $p_s=p_0\exp(e\Delta\phi_{sc}/kT)$ is the surface concentration of holes, p_0 is the bulk value, N_{ox} and N_{red} are the number of solution-phase reactants per unit volume, N_c and N_v are effective densities of states in the conduction and valence bands, respectively, k_c^+ and k_v^- are the respective rate constants of the anodic and cathodic reactions involving electrons, and k_v^+ and k_v^- are the respective rate constants involving holes. In contrast to metals, all of the change in electrode potential is across the space-charge region and none between the semiconductor surface and the in-

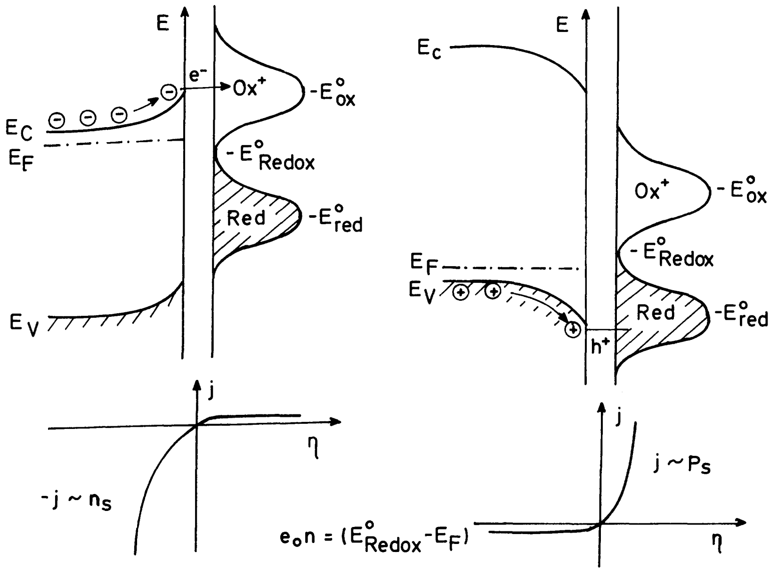

n - type
(a)

p - type
(b)

FIG. 14. Energy correlation at a particular overvoltage η and current–voltage curves for redox reactions at semiconductor electrode [reproduced from Gerischer (1990) with permission of Pergamon Press].

ner or outer Helmholtz plane when the carrier concentration is low ($<10^{-16}/\text{cm}^3$) in the semiconductor. Consequently, there is no variation of the activation barrier for electron transfer. The applied potential causes only a change in concentration of electrons or holes in the space-charge region in the semiconductor. The surface concentration of the electrons determines the rate of electron transfer from an *n*-type semiconductor to the solution-phase redox system. The rate of electron injection into the conduction band from a redox system is constant and independent of the potential. For a *p*-type semiconductor and hole transfer, the relation are reversed. This is shown schematically in Fig. 14 together with the current–potential curves. The Helmholtz layer may change with potential, as a result of the presence of surface states or chemical changes of the composition of the interface. The current–potential curves approach to a certain extent those of metal electrodes. The electrolytic decomposition and photodecomposition of semiconductors are to a large extent controlled by the electron and hole concentrations (see Gerischer, 1990).

Surface states are a well-established feature of semiconductors, resulting from dangling bands or different bond structures at the surface. A large number of surface states in a narrow energy range can be detected by means of capacitance measurements. They can act as a bridge for electron transfer and accelerate the rate enormously.

4.3 Photoeffects at Semiconductor Electrodes

Light of energy greater than the band gap ($h\nu > E_g$) creates electron–hole pairs, the majority of which are separated into individual electrons and holes by the space-charge field. For *n*-type semiconductors the holes may reach the electrode surface at a potential equivalent to the valence-band edge and there oxidize a solution-phase species. Photo-oxidation processes are promoted by irradiation of an *n*-type semiconductor electrode. For a *p*-type semiconductor electrode, the electric field is of opposite sign in the space-charge region, and the electrons produced by irradiation migrate toward the surface, while the holes migrate toward the bulk. At the surface, the electron can cause a photoassisted reduction of species in the solution phase.

The generation of photocurrents and photovoltages at semiconductor–solution interfaces has facilitated the development of three types of photoelectrochemical cells (Bard and Faulkner, 1980). These are photovoltaic, photoelectrosynthetic, and photocatalytic cells. In photovoltaic cells the photovoltaic effect is used to generate electric power as in solid-state photovoltaic cells. The reaction at the counterelectrode is simply the reverse of the photoassisted reaction at the semiconductor electrode. Corrosion problems and limited efficiency restrict the applicability of these cells. In photoelectrosynthetic cells the reaction at the counterelectrode differs from that at the semiconductor. In a typical cell with an *n*-type semiconductor, H_2O can be oxidized; or CO_2 can be reduced with a *p*-type semiconductor. In the first reaction, O_2 evolution takes place at the semiconductor electrode and H_2 evolves at the counterelectrode. The energy is stored in the form of H_2. The report by Fujishima and Honda (1972) on the limited photoelectrolysis of water with *n*-type TiO_2 attracted much attention. The photoreduction of H_2O can be performed with *p*-type semiconductors.

The semiconductors suitable for producing these reactions are poor electrocatalysts. Even so, these systems have a strong competition from the combination of solid-state photovoltaic cells with conventional electrochemical H_2 generation by water electrolysis. Colloidal semiconductor particles are being studied for use in these cells (Bard, 1980; Grätzel, 1981). This research led to the discovery of a quantum size effect, a new phenomenon in small dimensions reported by Rosetti *et al.* (1984) and Fojtik *et al.* (1984). Photocatalytic cells are similar to the photoelectrosynthetic cells, except that the reaction usually proceeds slowly even in the dark and the light energy is used to overcome the energy of activation of the reaction. The efficiency of these cells is also very low.

In overall perspective, photoelectrochemical cells are not yet ready for practical applications because of low efficiency, limited operating life, and high costs.

5. TECHNIQUES FOR THE STUDY OF ELECTROCHEMICAL INTERFACES AND ELECTRODE PROCESSES

5.1 Electrochemical Techniques

Electrochemical techniques will be only briefly summarized. A good description of these techniques can be found in the text by Bard and Faulkner (1980). Electrochemical techniques form the basis of the broad field of electroanalytical chemistry. Most of the methods involve the perturbation of the electrode system from its equilibrium state and the measurement of the steady-state or nonsteady-state response. The perturbation methods include

1. displacement of the electrode potential from the reversible value,
2. passage of a net current or charge,
3. changes in the concentrations of species involved in the electrode reaction, and
4. pressure, temperature, and area changes.

Upon perturbation, the potential, current, or impedance of the electrode systems is controlled. The transport of the reacting species from the bulk of the solution to the electrode surface is well defined and easy to treat quantitatively for the rotating-disk electrode (Levich, 1962; Pleskov and Filinovskii, 1976).

A common characteristic of most electrochemical techniques is high sensitivity. One percent of a monolayer can be easily determined, linear sweep rates of 10^6 V/s can be used with ultramicroelectrodes, and the modulation of the potential can extend into the megahertz range. Very high pressures and low temperatures have been used with electrochemical cells. Rate constants have been examined on a submicrosecond time scale directly and even on a nanosecond time scale with redox catalyst techniques (Andrieux *et al.*, 1990).

5.2 *In Situ* and *Ex Situ* Spectroscopic and Surface-Physics Techniques

Although very sensitive, electrochemical techniques often are not sufficient to determine the mechanisms of electrode reactions and the properties of the electrode–solution interface. These techniques lack the molecular-level specificity to identify adsorbed intermediates and products of the reaction. In order to overcome this problem, a large effort is being made to combine various spectroscopic techniques, capable of providing molecular-level information, with electrochemical ones. While optical spectroscopic techniques (e.g., uv–visible, infrared, Raman) have been used to examine various species adsorbed on single-crystal electrode surfaces in electrolytes (i.e., *in situ*), this is not true of many of the surface-sensitive electron physics techniques which require a vacuum environment, e.g., x-ray photoelectron spectroscopy, Auger electron spectroscopy, or low-energy electron diffraction (LEED). Special transfer systems have been developed to facilitate the transfer of single-crystal electrode surfaces between the electrochemical and UHV environments, but surface changes can occur even with the noble metals. This makes the *ex situ* approach less attractive. Some examples of both approaches are discussed below.

5.2.1 uv–Visible Reflectance Spectroscopy and Ellipsometry. Studies of electrode surfaces with reflectance spectroscopy involve the measurements of the intensity of light reflected from the electrode surface as a function of the wavelength. There are several variations of this technique, such as the wavelength dependence of the electroreflectance coefficient $R^{-1}(dR/dE)$, differential reflectance studies of surface plasmon excitation, and Fourier-transform infrared reflectance. Several properties of the electrode–electrolyte interface, not accessible by standard electrochemical techniques, can be probed by reflectance techniques. These include the electronic properties of metal surfaces and adsorption-induced changes; structural and electronic properties of adsorbates and thin oxide films; surface states on metal electrodes; and surface reconstruction. See, e.g., the reviews by McIntyre (1973) and Kolb (1988). The use of single-crystal electrodes has given considerable impetus to some of these studies.

The reflectivity of a metal–electrolyte interface depends on the electrode potential, even in the absence of charge-transfer or partial-charge-transfer reactions. This change of reflectance with potential, $R^{-1}(dR/dE)$, is referred to as the electroreflectance. Changes in electrode potential change the electron density extending into the interface. This transition region acts like an optical adlayer of thickness small compared to the wavelength, and its optical properties are functions of the surface charge which in turn depends on the electrode potential. The technique is surface sensitive and suitable for studying the electronic properties of metal surfaces. Figure 15 shows the normal-incidence electroreflectance spectra of Au(100) for various bias potentials. The potential dependence of the spectra has been interpreted in terms of the optical transition from an occupied bulk state into an empty surface state. The transition energy increases as the potential is changed in the positive direction.

Reflectance studies of adsorbed layer usually consider the interface to be composed of three regions: the bulk electrolyte, the interfacial region, and the bulk electrode. Quantitative treatments involve the calculation of the relative change of reflectance,

$$\Delta R/R=[R-R(0)]/R(0), \tag{48}$$

where $R(0)$ and R are the reflectances of the bare and adsorbate-covered surfaces. In a three-region model, $\Delta R/R$ for *s*-polarized light is given by

$$(\Delta R/R)_{\perp}=\frac{8\pi d(\epsilon_1)^{1/2}\cos\phi_1}{\lambda}\,Im\left(\frac{\epsilon_2-\epsilon_3}{\epsilon_1-\epsilon_3}\right), \tag{49}$$

where ϵ_1, ϵ_2, and ϵ_e are the dielectric constants of the bulk solution, adsorbed layer, and electrode substrate, respectively; d is the effective thickness of the layer; λ is the vacuum wavelength; and ϕ_1 is the angle of incidence. The significance of ϵ_2 and d for a monolayer or submonolayer is open to question and Eq. (49) is only qualitatively significant when applied to such layers.

Typical systems studied by reflectance spectroscopy include metal monolayers obtained by UPD, oxygen adsorption and oxide formation, halide adsorption on metals, hydrogen adsorption on platinum, and the adsorption of transition-metal complexes and dyes on metals, graphite, and semiconductors (McIntyre, 1973; Kolb, 1988).

A related technique involves the transmission of light through an electrochemical cell with optically transparent electrodes (electronically conducting SnO_2 or thin layers of Au or Pt on quartz or minigrids) and mea-

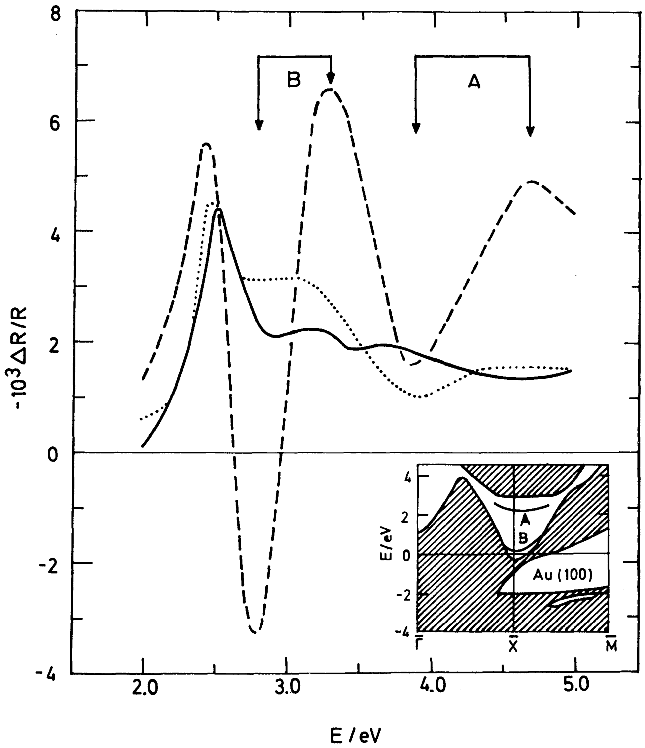

FIG. 15. Normal-incidence ER spectra of reconstructed Au(100)-(5×20) (solid), Au(100)-(1×1) (dashed), and Au(111) (dotted) in 0.01*M* $HClO_4$ caused by potential change from −0.2 to +0.3 V vs standard calomel reference electrode. Inset: projected band structure and surface states A and B of the unreconstructed Au(100) [reproduced from Kolb (1988) with permission of Elsevier Sequoia Press].

surements of the absorbance changes resulting from species produced or consumed at the electrode.

Ellipsometry and ellipsometric spectroscopy have been used to investigate the same or similar systems as reflectance spectroscopy. This work has been reviewed by Muller (1973) and Gottesfeld (1989). Reflection of linearly polarized light from a surface produces elliptically polarized light because of different amplitude reduction of the parallel and perpendicular optical components and phase retardation. The difference in phase angle defines one ellipsometric parameter, Δ, and the ratio of the amplitudes of parallel and perpendicular components of the electric field defines $\tan\psi$, the second parameter. From the changes in Δ and ψ upon reflectance, the two optical constants of a film between two bulk phases can be determined if the film thickness is known. Often the film thickness as well as the two optical constants (the real and imaginary components of the refractive index of the film) need to be determined. This problem has been solved with the automatic ellipsometer developed by Cahan and Spanier (1969), which measures the three parameters Δ, ψ, and $\Delta R/R$. Ellipsometry is widely used to study film growth on electrode surfaces. Its sensitivity to a small fraction of a monolayer makes it attractive for such studies even though the quantitative interpretation of optical data for submonolayers is a problem. Automatic instruments can be used for spectroscopic measurements, i.e., the dependence of the optical constants on wavelength (Wheeler *et al.,* 1978).

5.2.2 Infrared Reflectance Spectroscopy and Raman Scattering from Electrode Surfaces. Vibrational properties of species adsorbed at the electrode–aqueous-solution interface can be obtained by infrared reflectance spectroscopy. This was shown by Bewick (1983), who reduced the light path through the aqueous solution to approximately 1 μm by moving the reflecting electron surface close to the infrared-transparent window. Modulation of the electrode potential was necessary with a dispersive infrared spectrometer in order to achieve sufficient sensitivity. This technique is known as electrochemically modulated infrared spectroscopy (EMIRS). The use of potential-modulation techniques is difficult with Fourier-transform techniques because of the overlaps of the band pass of the

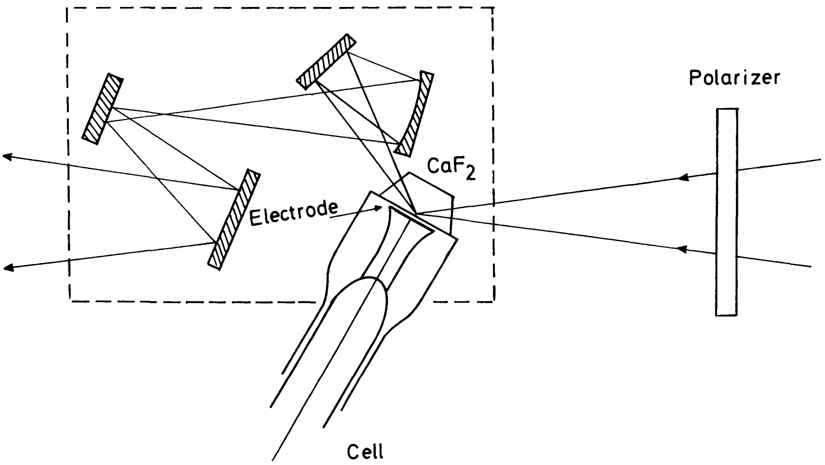

FIG. 16. Schematic diagram of ir spectroelectrochemical cell [reproduced from Kunimatsu *et al.* (1985) with permission of The American Chemical Society].

frequencies needed for the potential modulation and the interferometer. With Fourier-transform infrared (FTIR) spectroscopy, the signal-to-noise ratio is satisfactory even without such potential modulation(Bewick and Pons, 1985). One type of ir spectroelectrochemical cell using external reflectance is shown in Fig. 16. In order to eliminate the unwanted water bands from the ir spectrum, two series of interferograms at two different potentials are normalized relative to each other. This is the SNIFTIRS (subtractively normalized Fourier-transform infrared spectroscopy) technique, which has proved useful for the detection of the species generated or adsorbed on electrode surfaces.

These infrared techniques show great promise for obtaining information on a molecular level about species adsorbed at the electrode–electrolyte interface as well as in the thin electrolyte layer, and are helpful in establishing pathways for various electrode processes. Typical applications include studies of intermediates and products in the oxidation of organic molecules such as methanol, formic acid, and glucose; the reduction of CO_2; and the adsorption of CO, hydrogen, and anions. The identification of adsorbed intermediates and products has been reported in several instances. The interaction of adsorbates with the electrode surface and the potential dependence have been followed. Figure 17 gives the FTIR spectra obtained for the reduction of CO_2 on single-crystal platinum surfaces. The negative band for solution-phase CO_2 results from the consumption of this species in the thin layer of solution during its reduction to CO. The bands for bridge-bonded and linearly bonded CO are seen, i.e.,

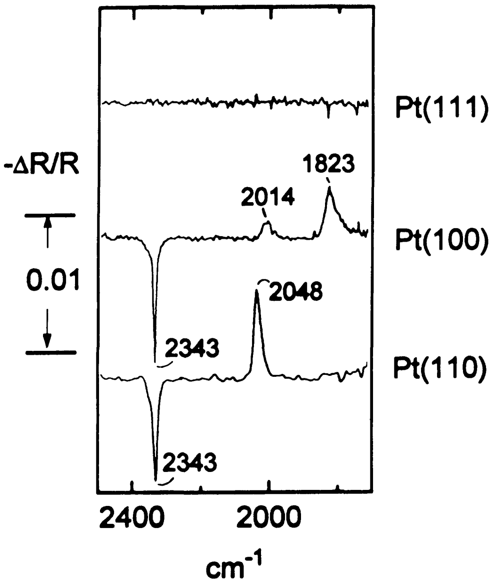

FIG. 17. Potential-difference FTIR spectra for CO_2 reduction at Pt single-crystal electrodes in CO_2-saturated 0.1 *M* $HClO_4$. No reduction occurs on Pt(111); bridge-bonded and linearly bonded CO are reduction products on Pt (100) while linearly bonded CO predominates on Pt(110) [reproduced from Nikolic *et al.* (1990) with permission of Elsevier Sequoia Press].

Pt
    C=O   and   Pt=C=O
Pt
  bridged   linear

CO is the product of the CO_2 reduction in the hydrogen adsorption region on Pt(100) and Pt(110). The Pt(111) surface is inactive for CO_2 reduction (Nikolic *et al.* 1990).

The Raman signals from adsorbed species at monolayer levels on most substrates are ordinarily very low, in fact too low for Raman scattering to be very useful in routine studies of adsorbed species. Amazingly large Raman scattering, however, was obtained by Fleischmann *et al.* (1973), and subsequently recognized as a tremendous enhancement of the Raman signals for adsorbates on silver, copper, and gold by Jeanmaire and VanDuyne (1975). This effect is known as surface-enhanced Raman scattering (SERS) and requires that the metal have very high conductivity at the optical frequency of the laser beam used to produce the Raman scattering. With gold and copper this means that the enhancement is large only at laser frequencies in the red below that of the interband transition responsible for their color. A second requirement is that the surface have a certain amount of roughness. Under optimum conditions, the enhancement can be by 10^6-fold on silver.

During the 1970s and early 1980s SERS attracted the attention of a large number of workers from various fields. This work has been reviewed by Fleischmann and Hill (1982) and by Chang and Laube (1984). There is not yet a consensus on the theory of the SERS effect. In general, roughened surfaces are used, but in some cases intentional roughening of the silver single-crystal surfaces is not necessary to achieve very substantial enhancement. Figure 18 shows the SERS spectra of Fe(III)-tetrasulfonated phthalocyanines (FeTsPc) adsorbed on the single-crystal silver electrode surfaces. These spectra have similarity with the solution-phase spectra, indicating that there is no change in the symmetry of the molecule upon adsorption and that the planes of the TsPc ligands are not oriented parallel to the surface (Adžić *et al.*, 1984). This information is of considerable help in understanding how the adsorbed FeTsPc functions as a four-electron O_2-reduction electrocatalyst.

5.2.3 EXAFS and Surface X-Ray Diffraction.

EXAFS, or extended x-ray absorption fine-structure, spectroscopy yields information concerning the radial distribution func-

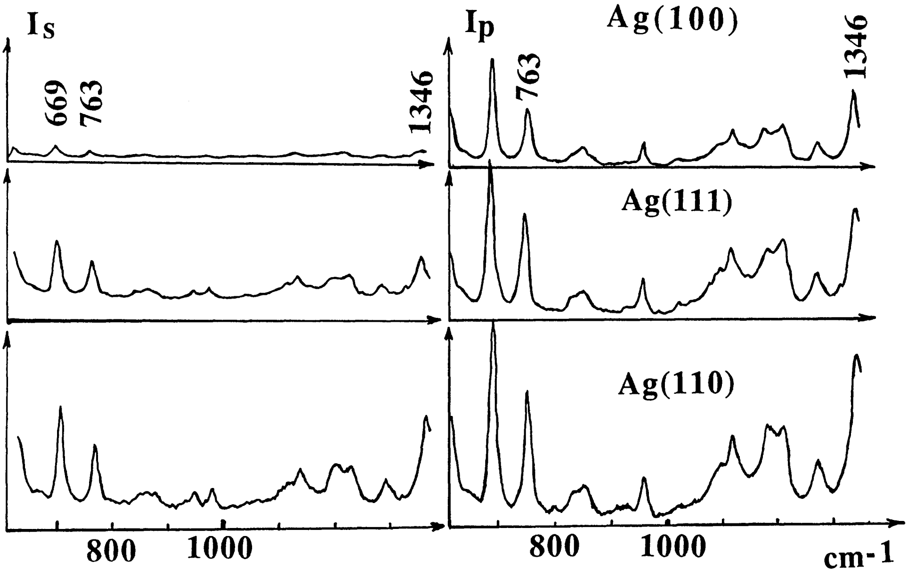

FIG. 18. Resonant surface-enhanced Raman spectra obtained from Fe-TsPc adsorbed on silver single-crystal interfaces. Electrolyte 0.1*M* $HClO_4+H_2O$ saturated with argon gases at $T=20\,°C$. Laser illumination at 632.8 nm with 15-mW output power. Spectra recorded with Triplemate monochromator and photodiode-array detection system (Adžić *et al.*, 1984).

tion for neighboring atoms; i.e., the number and types of neighboring atoms and the corresponding interatomic distances can be determined. In some instances the oxidation state can also be obtained. This information is obtained by analyzing the oscillatory fine structure in the x-ray absorption spectra.

The x-ray absorption can be obtained directly from the measurement of absorption, but it can also be obtained from the fluorescence radiation or electron yield. *In situ* applications in electrochemical systems are possible since the aqueous solutions are relatively transparent to the "hard" x rays. Generally, thin layer cells are used. Synchrotron radiation is preferred to conventional x-ray sources because of the much higher intensities obtained with the former. In electrochemical experiments, EXAFS permits the changes in bond length and coordination number with change in oxidation state to be examined *in situ*. A variety of electrochemical systems has been studied. These include dispersed electrocatalysts, polymeric films on electrodes, ions in solution, ion-exchange polymers, passivation and oxide films on metals, battery oxides, and UPD layers on gold and silver (see the review by Sharpe *et al.*, 1990).

Surface x-ray scattering, using synchrotron radiation, has been recently applied for *in situ* determination of structures of the electrochemical interface. This technique provides a direct determination of the position of atoms at the electrode surface and shows a considerable potential for improving our knowledge of electrochemical interfaces. Surface reconstruction phenomena and structures of UPD metal monolayers have been reported (Toney and Melroy, 1991; Ocko *et al.*, 1990).

5.2.4 Photoemission of Electrons from a Metal into Solution. Photoelectronic emission from metals into solutions can also provide information on the structure of the electrode–electrolyte interface. Photoemission of electrons occurs upon illumination of a metal electrode with light whose energy exceeds the work function of the metal–solution interface (Barker *et al.*, 1965). After injection of the electrons into the liquid phase, the subsequent interactions lead to the thermalization and solvation of the electron. This is followed by reaction with electron acceptors at distances of the order of 1–10 nm from the electrode. Suitable scavengers include H^+, CO_2, O_2, N_2O, NO_3^-, etc. Cathodic polarization of the interface decreases the work function and consequently increases the photoemission. Brodsky *et al.* (1974) gave the following equation for the total photoemission current for dilute solutions when the de Broglie wavelength of the electron is less than the thickness of the diffuse part of the double layer:

$$I=A[h\nu-h\nu_0-e(E-\phi_2)]^{5/2}, \tag{50}$$

where A is a constant, $h\nu$ is the light quantum energy, $h\nu_0$ is the work function of the metal–solution interface at the potential $E=0$ versus the reference electrode, and ϕ_2 is the potential of the outer Helmholtz plane. This equation, referred to as the "$\frac{5}{2}$-power law," has been found to be followed experimentally for a series of metals. At the PZC, $\phi_2=0$ and the photocurrent–potential curves for solutions of different concentrations should intersect. This method of determination of the PZC yields results in good agreement with PZC data obtained from the differential-capacitance minima within a possible error of 0.02–0.03 V (Frumkin *et al.*, 1980).

5.2.5 Scanning Tunneling Microscopy. Scanning tunneling microscopy (STM), developed by Binnig *et al.* (1982), is operative not only in vacuum but also with the samples in air or covered with electrolytes. A sharp probe of a metal such as W or Pt is scanned parallel to the surface using the electron tunneling current to establish the surface profile with atomic-level or near–atomic-level resolution. Since the electrolytic solutions are not electronic conductors, electron tunneling can occur between the tip and electrode surface that is being examined. Images with close to atomic resolution have been reported for single-crystal Au surfaces and also for the initial stages of metal deposition on metals and graphite. This work has been reviewed by Sonnenfeld *et al.* (1990). Atomic-resolution images have been reported for Au(100) and the UPD of Cu on Au(100) by Magnussen *et al.* (1990). Ordered structures of Cu adatoms have been found and the structure-determining role of the coadsorbed anions has been discussed. Figure 19 shows the ordered structure of a Cu adlayer on Au(100).

Manne *et al.* (1991) have used the atomic force microscope (AFM) to image a UPD layer of Cu at Au(111) with atomic resolution. The AFM relies on the intrinsic forces be-

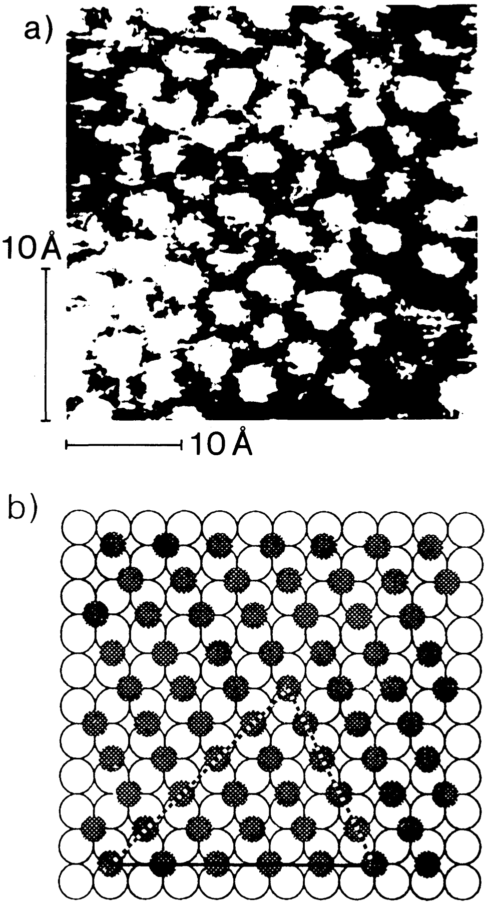

FIG. 19. (a) STM image of ordered structure of a Cu adlayer on Au(100) in 0.05*M* H_2SO_4+5 m*M* $CuSO_4$. (b) Model of the adlayer structure. Filled circles, Cu adatoms; solid line, well-aligned along [110]; dashed lines, less-ordered adlayer lattice directions [reproduced from Magnussen *et al.* (1990) with permission of The American Physical Society].

tween the tip and the surface to cause the tip to be displaced as it is scanned parallel to the surface. The displacements perpendicular to the surface are then sensed with an optical lever (see Binnig *et al.*, 1986). These results confirmed the STM data and established that the Cu atoms are in a close-packed lattice with a spacing of 0.29±0.02 nm in perchloric acid but 0.49±0.02 nm in a sulfate electrolyte.

5.3 *Ex Situ* Studies by Surface-Physics Techniques

Techniques that require vacuum or ultrahigh vacuum (UHV) cannot be applied directly to electrode–liquid-solution interfaces. In order to make use of several UHV techniques for the characterization of electrode surfaces, Hubbard *et al.* (1988), Yeager *et al.* (1978), and Ross (1979) designed systems for the transfer of the electrode from the electrolyte environment into a UHV chamber and vice vera. These systems have been designed to minimize restructuring and contamination of the single-crystal surfaces during transfer. The UHV techniques used most often by electrochemists include Auger electron spectroscopy (AES), low-energy electron diffraction (LEED), x-ray photoelectron spectroscopy (XPS), and electron energy-loss spectroscopy (EELS).

In the transfer system developed at Case Western Reserve University (Yeager *et al.*, 1978), the electrode is cleaned and annealed in a UHV chamber, characterized with AES-LEED-XPS, transferred into a second chamber, and made part of a thin-layer electrochemical cell. Following the electrochemical measurements, the electrode is transferred back into the vacuum chamber for reexamination with UHV techniques.

There is, however, always a question of the relevance of measurements in which the electrode surface is in high vacuum to those in the electrochemical environment (see Sec. 5.2). The importance of knowing the crystallographic orientation of the surface from LEED measurements before and after the electrochemical measurements is not questioned, but restructuring of the surface is possible during the transfers. This type of measurement has helped to establish the voltammetric profiles of well-ordered single-crystal surfaces for various adsorption reactions. LEED measurements in UHV have been used to confirm the validity of various methods of surface preparation. One such method is the annealing of noble-metal crystals in a hydrogen–air flame followed by quenching of the crystal in ultrapure water (Clavilier *et al.*, 1981). An improvement in this method is the cooling of the hot crystal in H_2 gas before contact with the ultrapure water (Markovic *et al.*, 1986). These methods apparently give well-ordered platinum and gold single-crystal surfaces. Zurawski *et al.* (1987) have suggested the adsorption of CO to protect the Pt surfaces from other impurities during the transfer from the ultrahigh-vacuum environment into the electrolytic solutions. The CO is easily removed from the surface by oxidative desorption.

This approach has been used for the study of hydrogen adsorption on the low-index planes of Pt (Sec. 3.2), UPD of several metals on Pt and Au, and the oxidation of a large number of organic molecules. Ordered layers were found with LEED-AES techniques for the UPD of Ag, Cu, and Pb on iodine-pretreated Pt(*hkl*) (Stickney *et al.,* 1986), Pb on Ag(*hkl*) (Hanson and Yeager, 1988), and Cu on Pt(111) (Aberdam *et al.,* 1985). Using EELS, Hubbard *et al.* (1988) have identified products and fragments of the oxidation of a large number of organic molecules on Pt(*hkl*).

Bange *et al.* (1987) have recently studied Br adsorption on Ag(110) by progressively replacing a monolayer of H_2O by Br. This approach allows a direct comparison of the potential drop in the double layer at the metal–electrolyte and the metal–vacuum interfaces.

6. APPLIED ELECTROCHEMISTRY

6.1 Batteries

Batteries are electrochemical devices for energy storage in the form of chemical energy. The main feature of the electrochemical energy conversion is its high efficiency, which explains the widespread use of batteries and continuing interest in fuel cells. A variety of electrochemical systems are used in a large number of primary and secondary (rechargeable) batteries. These include the commonly used lead–acid, nickel–cadmium, silver–zinc, nickel–iron, nickel–zinc secondary and the manganese dioxide–zinc, air–zinc and lithium primary batteries. Various metal–air systems are being considered as future electrochemical storage devices. During the last ten years primary lithium batteries, with aprotic nonaqueous solvents, have become widely used principally because of their high energy per unit weight, their high cell voltage (3.0 V), and their long life. A large effort is under way in the U.S. to improve the performance of several existing batteries and to develop new storage batteries for use in electric cars. A good survey of the battery systems is available in the book edited by Bockris (1981).

6.2 Fuel Cells

Fuel cells are unique devices that convert chemical energy of fuels (H_2, methanol, hydrocarbons) directly into electric energy. The fuel is fed continuously to the anode, while the oxidant, usually oxygen from the air, reacts at the cathode. The practical conversion efficiency can be very high because they are not limited by Carnot-cycle efficiency. This makes fuel cells very attractive energy conversion devices.

Five types of fuel cells are under development:

1. *Phosphoric acid* fuel cells with concentrated phosphoric acid as the electrolyte and operating with H_2 and air at a temperature of ~200 °C. The H_2 is usually obtained from reformer operating on a hydrocarbon such as natural gas.
2. *Solid proton-exchange membrane* fuel cell with a solid ionic conducting polymer as the electrolyte and operating on very pure H_2 and air at a temperature of 80–100 °C. The H_2 can be prepared from a hydrocarbon using a reformer, but the residual carbon monoxide must be almost completely removed. Efforts are in progress to develop new ionic-conducting polymer electrolytes that will permit this fuel cell to operate at much higher temperatures (> 150 °C).
3. *Alkaline electrolyte fuel cells* using a concentrated alkali-metal hydroxide as the electrolyte and operating at ~200 °C on H_2 and O_2 all of which must be free of CO_2. The principal application for the cell has been the Apollo space mission and other manned space missions.
4. *Molten carbonate fuel cells* using molten alkali-metal carbonates as the electrolyte at ~650 °C and operating on H_2 and air. The anode can be modified to permit direct operation of the cells on natural gas and other hydrocarbons.
5. *Solid oxide fuel cells* with solid ionic conducting oxides as electrolytes at temperatures of ~1000 °C and operating on hydrocarbons or other fuels and air.

The phosphoric acid cells are in an advanced state of development for electric utility applications but suffer from relatively low overall system conversion efficiency, principally because of the large overpotential of the air-consuming cathodes due to the inadequacy of the electrocatalysts. Phosphoric acid fuel-cell plants of several tens of megawatts have been built.

6.3 Large-Scale Industrial Processes

The electrochemical production of chlorine and sodium hydroxide is the largest tonnage production of any single process in the chemical industry. Most of the technologically important metals, except iron, are either obtained or refined electrochemically. Aqueous solutions are used in many cases, but aluminum, magnesium, titanium, zinc, and alkali and alkaline-earth metals are obtained from molten salts. The electrochemical production of aluminum metal alone involves $\sim 5\%$ of the total electric generating capacity of the United States. The operating cell voltages of most of these processes greatly exceed the thermodynamic cell potentials, and the redesign of many of these cells could save considerable electricity.

6.4 Inorganic and Organic Electrosynthesis

Hydrogen and a whole range of oxo compounds, such as chlorate, perchlorate, hypochlorite, hydrogen peroxide, perborate, etc., are produced electrochemically. Some industrially important organic compounds and pharmaceuticals are produced by organic electrosynthesis. Electrochemical processes have usually been developed in the cases where a standard organic synthesis was unavailable or required several steps, or the electrochemical process was more attractive from the energy or ecology points of view. The dimerization of acrylonitrile to adiponitrile which is used in nylon manufacture is the largest volume electro-organic process. Many electro-organic processes require the use of organic solvents, and this results in relatively poor conductivity of the electrolytic solutions. Furthermore, solubility and mass transport may limit the current densities to relatively low values. New types of electrochemical cells have been designed in order to optimize performance under these rather adverse conditions.

A particularly promising area for the use of electrochemical synthesis or organic molecules is in the pharmaceutical industry when the relatively high cost associated with electrosynthesis may be small compared to other costs, where there is a special premium on selectivity and yield, and where the total amounts of the final products are also relatively small.

6.5 Electroplating

Protective and decorative coatings of metals and alloys are obtained by an electrodeposition process called electroplating. Not only metal objects, but also nonconductors such as plastics, glass, ceramics, etc., can be electroplated. This is accomplished by first depositing a thin metal layer by a chemical reduction (electrodeless) technique and then building up the metal deposit by conventional electroplating. The corrosion resistance of electroplated objects is usually considerably enhanced. Electroplating processes play a major role in the electronics industry in the manufacture of circuit boards, plating of gold contacts, and memory devices.

6.6 Electrochemical Sensors

Electrochemical cells are basic to many sensor devices which can monitor concentrations of various species in the environment and in living organisms. From an electrochemical point of view they can be classified as potentiometric, amperometric, and conductometric. The size of the electrode depends on the sensor application. Miniaturization of various sensors takes advantage of integrated-circuit technology. Such sensors are under development for such applications as the monitoring of blood gases and ionic constituents including pH, O_2, and CO_2. The application of sensors span such diverse fields as industrial chemical processes and electrophysiology.

6.7 Corrosion and Corrosion Protection

Most corrosion processes involve electrochemical reactions. Most of the metals used for construction would be far too unstable in contact with water or moist air if it were not that they are covered by an oxide layer which blocks the corrosion. The formation and structure of this passivation layer have been intensely studied by electrochemists using the *in situ* optical and other surface techniques described earlier in this review. Even so, they are only partially understood.

Electrochemical techniques are used to suppress corrosion. The most common illustration is the cathodic protection of steel structures such as bridges, pipe lines, and ship hulls. This involves maintaining the potential of the steel structure at a negative value where the corrosion no longer occurs, using a second electrode as the anode. Another approach is electrogalvanization, which involves the electrodeposition of a layer of zinc on the steel surface. Zinc is a more active metal than iron or steel and it dissolves, while maintaining a potential at which the steel is cathodically protected. Electrogalvanization has greatly suppressed corrosion in the automobile. Even with these various ways to suppress corrosion, 10% to 15% of the ferrous metals produced annually in the U.S. are destroyed by corrosion.

7. CHALLENGING PROBLEMS FOR THE FUTURE

Electrochemistry today enters into various fields that were heretofore outside the traditional electrochemical science. For example, electrochemical mechanisms are operative in many biological processes from blood clotting to the transfer of nerve impulses, and the field of bioelectrochemistry is developing rapidly. Interfacial electrochemistry is expanding rapidly as a part of surface science as a whole. Much of this rapid growth is the result of the development of *in situ* optical and other techniques for the study of electrochemical interfaces.

Two particularly challenging areas of electrochemistry are the Fleischmann–Pons phenomenon and the electrochemistry of superconductor–electrolyte interfaces. Fleischmann *et al.* (1989) have reported evidence for nuclear fusion of deuterium electrochemically injected into palladium at room temperature. This evidence includes the apparent generation of tritium, excess heat in calorimetric measurements, and the generation of neutrons. The palladium cathode is known to be able to take up to at least one deuterium per Pd atom. Whether a nuclear reaction can occur when deuterium is electrochemically injected in Pd remains a matter of great controversy. Further work with well-characterized systems should provide the answer to this question of paramount importance.

High-temperature superconductors offer promise for great advances in various technologies. Electrochemical characterization of superconductor–liquid-electrolyte-solution and superconductor–solid-electrolyte interfaces and electrochemical synthesis of superconducting compounds will be of great electrochemical interest in the future.

8. CHALLENGING OPPORTUNITIES IN ELECTROCHEMISTRY

Some of the challenging fundamental and technological problems in electrochemistry are the following:

Fundamental Problems

1. More definitive and comprehensive models for electrochemical interfaces, taking into account the discrete structure of the electrode surface and the electronic properties as well as the structure of the compact and diffuse ionic layers.
2. Techniques capable of identifying short-lived intermediates in electrochemical reactions, both in solution and adsorbed on the electrode surface.
3. The development of a quantitative understanding of surface-enhanced Raman scattering, including the selection rules and the relation of the enhancement to the surface properties of the electrode.
4. The refining of scanning tunneling microscopy so as to achieve routinely atomic-level resolution in *in situ* studies (particularly improved methods for preparing scanning tips).
5. Studies of the electrochemical properties of superconductor–electrolytic-solution interfaces and particularly quantum-mechanical phenomena that will result in unusual electrochemical effects (e.g., barrierless electron-exchange processes, unusual double-layer properties, prevention of any penetration of the electric field into the electrode phase).

Technological Problems

1. Fuel cells suitable for terrestrial applications, such as for vehicle propulsion and not only for space missions; problem areas

include
-more active catalysts for O_2 reduction and for the oxidation of practical fuels (e.g., hydrocarbons, alcohols),
-new polymer electrolytes suitable for use in fuel cells at much higher temperatures than presently possible.

2. New and improved rechargeable batteries for vehicle propulsion such as
-lithium–polymer-electrolyte systems,
-metal–air cells (e.g., Zn–air).

3. Photoelectrochemical systems affording high conversion efficiency at reasonable costs and with long lives.

4. Implantable microsensors for biomedical applications:
-blood gases—CO_2, O_2;
-*p*H and the concentrations of other ionic constituents (e.g., KI^+, Na^+, Ca^{2+});
-glucose (diabetes);
-neurotransmitters (dopamine).

5. New surface treatments of steel affording more effective inhibition of corrosion; for example,
-surface alloys,
-amorphous surface structures,
-coatings or deposits.

6. Definitive studies to establish whether any nuclear phenomena are involved during the electrochemical charging of palladium with deuterium, as has been reported by Fleischmann *et al.* (1989); i.e., the production of excess heat, neutrons, γ radiation, and tritium. The claim by these researchers that a nuclear fusion reaction is involved has produced much controversy among electrochemists and also between electrochemists and nuclear physicists. Some of the observations of Fleischmann *et al.* (1989) have also been reported by some other electrochemists including the authors of this article (Adžić *et al.*, 1990) but it is far from clear whether the effects are the result of nuclear fusion or some other nuclear process.

List of Works Cited

Aberdam, D., Durand, R., Faure, R., El Omar, F. (1985), *Surf. Sci.* **162**, 782–788.

Adžić, R. (1984) in: H. Gerischer, C. Tobias (Eds.), *Advances in Electrochemistry and Electrochemical Engineering*, Vol. 13, New York: Wiley, Chap. 4.

Adžić, R. (1990), in: R. E. White, J. O'M. Bockris, B. E. Conway (Eds.), *Modern Aspects of Electrochemistry*, Vol. 21, New York: Plenum, Chap. 5.

Adžić, R., Gervagio, D., Bae, I., Cohan, B., Yeager, E. (1990), in *Proceedings of the First Annual Conference on Cold Fusion*, Salt Lake City: National Cold Fusion Institute, pp. 261–271.

Adžić, R., Yeager, E., Cahan, B. (1974), *J. Electrochem. Soc.* **121**, 474–484.

Adžić, R., Simic-Glavaski, B., Yeager, E. (1984), *National Meeting, The Electrochemical Society, Cincinnati, May 1984, Extended Abstracts* **84-1**, Abst. 433.

Albery, J. (1975), *Electrode Kinetics*, Oxford: Clarendon.

Andrieux, C. P., Hapiot, Ph., Saveant, J.-M. (1990), Chem Rev. **90**, 723–738.

Angerstein-Kozlowska, H., Conway, B. E., Hamelin, A., Stoicoviciu, L. (1987), *J. Electroanal. Chem.* **228**, 429–453.

Appleby, A. J. (1983), in: B. E. Conway, J. O'M. Bockris, E. Yeager, S. U. M. Khan, R. E. White (Eds.), *Comprehensive Treatise of Electrochemistry*, Vol. 7, New York: Plenum, Chap. 4.

Badialli, J. P. (1986), *Electrochim. Acta* **31**, 149–154.

Balandin, A. A. (1958), *The Problems of Chemical Kinetics, Catalysis and Reactivity*, Moscow: Academy of Sciences.

Bange, K., Stracler, B., Sass, J. K., Parsons, R. (1987), *J. Electroanal. Chem.* **229**, 87–98.

Bard, A. J. (1980), *Science* **207**, 139–144.

Bard, A. J. Faulkner, L.R. (1980), *Electrochemical Methods*, New York: Wiley.

Barker, G., Gardner, A., Sammon, D. (1966), *Electrochem. Soc.* **113**, 1182–1199.

Bewick, A. (1983), *J. Electroanal. Chem.* **150**, 481–493.

Bewick, A., Nicols, P. (1988), *J. Electroanal. Chem.* **243**, 445–453.

Bewick, A., Pons, S. (1985), in: R. J. H. Clark, R. E. Hester (Eds.) *Advances in Infrared and Raman Spectroscopy*, Vol. 12, London: Wiley-Heyden, pp. 1–63.

Bewick, A., Thomas, B. (1975), *J. Electroanal. Chem.* **65**, 911–931.

Binnig, G., Rohrer, H., Gerber, C., Weibel, E. (1982), *Phys. Rev. Lett.* **49**, 57–61.

Binnig, G., Quate, C. F., Gerber, Ch. (1986), *Phys. Rev. Lett.* **56**, 930–933.

Blum, L., Henderson, D. J. (1981), *J. Chem. Phys.* **74**, 1902–1910.

Bockris, J. O'M. (1981), *Comprehensive Treatise of Electrochemistry*, Vol. 2, New York: Plenum.

Bockris, J. O'M., Khan, S. U. M. (1979), *Quantum Electrochemistry*,New York: Plenum.

Bockris, J. O'M., Mathews, D. (1966), *J. Chem. Phys.* **44**, 298–309.

Bockris, J. O'M., Reddy, A. K. (1970), *Modern Electrochemistry*, Vol. 2, New York: Plenum.

Brodsky, A., Gurevich, Yu., Pleskov, Yu., Rotenberg, Z. (1974), *Modern Photoelectrochemistry, Photoemission Phenomena*, Moskow: Nauka.

Cahan, B. D., Spanier, R. (1969), *Surf. Sci.* **16**, 166–176.

Chang, R., Laube, K. (1984), in: *Critical Review of Solid State and Materials Sciences*, Vol. 12, Cleveland: CRC.

Chapman, D. L. (1913), *Philos. Mag.* **25**, 475–481.

Christmann, K. (1988), *Surf. Sci. Rep.* **9**, 1–163.

Christov, G. St. (1964), *Electrochim. Acta* **9**, 575–585.

Clavilier, J., Durand, R., Guinet, G., Faure, R. (1981), *J. Electroanal. Chem.* **127**, 281–287.

Conway, B. E. (1965), *Electrode Processes*, New York: Ronald.

Delahay, P. (1965), *Double Layer and Electrode Kinetics*, New York: Interscience.

Dogonadze, R. R., Kuznetsov, A. M., Levich, V. G. (1968), *Electrochim. Acta* **13**, 1025–1044.

Fleischmann, M., Hill, I. R. (1982), in: R. K. Chang, T. E. Furtak (Eds.), *Electrochemical Effects Surface Enhanced Raman Scattering*, New York: Plenum.

Fleischmann, M., Hendra, P. J, McQuillan, A. J. (1973), *J. Chem. Soc. Chem.* 80–81.

Fleischmann, M., Pons, S., Hawkings, X. (1989), *J. Electroanal. Chem.* **261**, 301–308; Errata, **263**, 187.

Fojtik, A., Weller, H., Koch, U., Henglein, A. (1984), *Ber. Bunsenges. Phys. Chem.* **88**, 969–977.

Frumkin, A. N. (1933), *Z. Phys. Chem.* **164A**, 121–133.

Frumkin, A. N., Petrii, O. A., Damaskin, B. B. (1980), in: J. O'M. Bockris, B. E. Conway, E. Yeager (Eds.), *Comprehensive Treatise of Electrochemistry*, Vol. 1, New York: Plenum, Chap. 5.

Fujishima, A., Honda, K. (1972), *Nature*, **238**, 37–38.

Gerischer, H. (1958), *Bull. Soc. Chim. Belg.* **67**, 506–527.

Gerischer, H. (1990), *Electrochim. Acta* **35**, 1677–1699.

Gottesfeld, S. (1989), in: A. Bard (Ed.), *Electroanalytical Chemistry*, Vol. 15, New York: Marcel Dekker.

Gouy, M. (1903), *Ann. Chim. Phys.* **29**, 145.

Grätzel, M. (1981), *Acc. Chem. Res.* **14**, 376–384.

Gurney, R. W. (1931), *Proc. R. Soc. London, Ser. A* **134**, 137–154.

Hamelin, A., Lipkowski, J. (1984), *J. Electroanal. Chem.* **171**, 317–330.

Hamelin, A. (1985), in: B. E. Conway, J. O'M. Bockris (Eds.), *Modern Aspects of Electrochemistry*, Vol. 16, New York: Plenum.

Hanson, M., Yeager, E. (1988), in: M. Soriaga (Ed.), *Electrochemical Surface Science*, ACS Symposium Series No. 378, Washington, D.C.: American Chemical Society, Chap. 10.

Helmholtz, H. L. F. (1853), *Ann. Phys.* (Leipzig) **89**, 211.

Horiuti, J., Polonyi, M. (1935), *Acta Physicochim. SSSR* **2**, 505.

Hubbard, A. T., Stern, D. A., Salaita, G. N., Frank, D. G., Frank, L., Laguren-Davidson, L., Batina, N., Zapien, D. C. (1988), in: M. Soriaga (Ed.), *Electrochemical Surface Science*, ACS Symposium Series No. 378, Washington, D.C.: American Chemical Society, Chap. 1.

Jeanmaire, D. L., Van Duyne, R. P. (1977), *J. Electroanal. Chem.* **84**, 1–20.

Khan, S. U. M. (1985) in: J. O'M. Bockris, B. E. Conway (Eds.), *Modern Aspects of Electrochemistry*, Vol. 15, New York: Plenum.

Kolb, D. M. (1978), in: H. Gerischer, C. W. Tobias (Eds.), *Advances in Electrochemistry and Electrochemical Engineering*, Vol. 11, New York: Wiley.

Kolb, D. M. Przasnyski, M., Gerischer, H. (1974), *J. Electroanal. Chem.* **54**, 25–38.

Kolb, D. M. (1988), in: R. J. Gale (Ed.), *Spectroelectrochemistry, Theory and Practice*, New York: Plenum, Chap. 4.

Kornyshev, A., Schmickler, W. J. (1985), *J. Electroanal. Chem.* **185**, 253–261.

Krishtalik, L. I. (1970), in: P. Delahay, C. Tobias (Eds.), *Advances in Electrochemistry and Electrochemical Engineering*, Vol. 7, New York: Wiley.

Krishtalik, K. I. (1983), in: B. E. Conway, J. O'M. Bockris, E. Yeager, S. U. M. Khan, R. E. White (Eds.), *Comprehensive Treatise of Electrochemistry*, Vol. 7, New York: Plenum, Chap. 3.

Kunimatsu, K., Goden, W. G., Seki, H., Philpott, M. R. (1985), *Langmuir* **1**, 245–250.

Kuznetsov, A. M. (1987), *Electrochim. Acta*, **32**, 1271–1278.

Lang, N. D. (1983), *Surf. Sci.* **127**, L118–L122.

Levich, V. G. (1962), *Physicochemical Hydrodynamics*, Englewood Cliffs, NJ: Prentice-Hall.

Levich, V. V. (1970), in: H. Eyring, D. Henderson, J. Jost (Eds.), *Advanced Treatise of Physical Chemistry*, Vol. 9B, New York: Academic.

Levich, V. G., Dogonadze, R. R. (1959), *Dokl. Akad. Nauk SSSR* **125**, 123.

Lorenz, W., Salie, G. (1961), *Z. Phys. Chem.* **218**, 259–271.

Magnussen, O. M., Hotlos, J., Nichols, R. J., Kolb, D. M., Behm, R. J. (1990), *Phys. Rev. Lett.* **64**, 2929–2932.

Manne, S., Hansma, P. K., Massie, J., Elings, V. B., Gewirth, A. A. (1991), *Science* **251**, 183–186.

Markovic, N., Hanson, M., McDougall, G., Yeager, E. (1986), *J. Electroanal. Chem.* **214**, 555–566.

Markovic, N., Marinkovic, N., Adžić, R. R. (1988), *J. Electroanal. Chem.* **241**, 309–328.

Marcus, R. (1964), *Annu. Rev. Phys. Chem.* **15**, 155–196.

McIntyre, J. D. E. (1973), in: R. H. Muller (Ed.), *Advances in Electrochemistry and Electrochemical Engineering*, Vol. 9, New York: Wiley.

Muller, R. (1973), in: R. H. Muller (Ed.), *Advances in Electrochemistry and Electrochemical Engineering*, Vol. 9, New York: Wiley.

Nikolic, B. Z., Huang, H., Gervasio, D., Lin, A., Fierro, C., Adžić, R. R., Yeager, E. B. (1990), *J. Electroanal. Chem.* **295**, 415–423.

Ocko, B. M., Wang, J., Davenport, A., Isaacs, H. (1990), *Phys. Rev. Lett.* **65**, 1466–1469.

Parsons, R. (1958), *Trans. Faraday Soc.* **54**, 1053–1063.

Parsons, R. (1990), *Chem. Rev.* **90**, 813–826.

Parsons, R., VanderNoot, T. (1988), *J. Electroanal. Chem.* **257**, 9–45.

Pleskov, Yu. V., Filinovskii, V. Yu. (1976), *The Rotating Disk Electrode*, New York: Consultants Bureau.

Rosetti, R., Ellison, J. L., Gibson, J. M., Brus, L. E. (1984), *J. Chem. Phys.* **80**, 4464–4469.

Ross, P. (1979), *J. Electrochem. Soc.* **126**, 67–77.

Schultze, W. J., Vetter, K. J. (1973), *J. Electroanal. Chem.* **44**, 63–81.

Sharpe, L. R., Heinbeman, W. R., Elder, R. C. (1990), *Chem. Rev.* **90**, 705–722.

Sonnenfeld, R., Schneir, J., Hansma, P. K. (1990), in: H. Gerischer, C. Tobias (Eds.), *Advances in Electrochemistry and Electrochemical Engineering*, Vol. 13, New York: Wiley, Chap. 1.

Stern, O. (1924), *Z. Elektrochem.* **30**, 508–516.

Tarasevich, M. R., Sadkowski, A., Yeager, E. (1983), in: B. E. Conway, J. O'M. Bockris, E. Yeager, S. U. M. Khan, R. E. White (Eds.), *Comprehensive Treatise of Electrochemistry*, Vol. 7, New York: Plenum, Chap. 6.

Stickney, L. J., Schardt, B. C., Stern, D. A., Wieckowski, A., Hubbard, A. T., (1986), *J. Electrochem. Soc.*, **133**, 648–650.

Tilak, B. V., Yeo, R. S., Srinivasan, S. (1981), in: J. O'M. Bockris, B. E. Conway, E. Yeager, R. E. White (Eds.), *Comprehensive Treatise of Electrochemistry*, Vol. 3, New York: Plenum, Chap. 2.

Toney, M., Melroy, O. (1991), in: H. D. Abruña (Ed.), *Electrochemical Interfaces: Modern Techniques for* in situ *Interface Characterization*, New York: VCH.

Trasatti, S. (1977), in: H. Gerischer, C. Tobias (Eds.), *Advances in Electrochemistry and Electrochemical Engineering*, Vol. 10, New York: Wiley, Chap. 3.

Trasatti, S. (1980), in: J. O'M. Bockris, B. E. Conway, E. Yeager (Eds.), *Comprenhensive Treatise of Electrochemistry*, Vol. 1, New York: Plenum, Chap. 2.

Vetter, J. J., Schultze, J. W. (1972), *Ber. Bunsenges. Phys. Chem.* **76**, 920–927.

Wheeler, D., Cahan, B., Chen, C., Yeager, E., (1978), in: R. Frankenthal, J. Kruger (Eds.), *Passivity of Metals*, Princeton, NJ: The Electrochemical Society.

Yeager, E. B. (1981), *J. Electrochem. Soc.* **128**, 160C–171C.

Yeager, E. (1986), *J. Mol. Catal.* **38**, 5–25.

Yeager, E. B., Salkind, A. (Eds.) (1972), *Electrochemical Measurements*, Vol. 1, New York: Wiley.

Yeager, E. B., O'Grady, W., Woo, M. Y. C., Hagans, P. (1978), *J. Electrochem. Soc.* **125**, 348–349.

Zurawski, D., Rice, L., Hourani, M., Wieckowski, A. (1987), *J. Electroanal. Chem.* **230**, 221–231.

Further Reading

Textbooks and Monographs

Abruña, H. D. (Ed.) (1991), *Electrochemical Interfaces: Modern Techniques for* In Situ *Interface Characterization*, New York: VCH.

Albery, W. J. (1975), *Electrode Kinetics*, Oxford: Clarendon Press.

Bard, A. J., Faulkner, L. R. (1980), *Electrochemical Methods*, New York: Wiley.

Bard, A. J., Parsons, R., Jordon, J. (Eds.) (1985), *Standard Potentials in Aqueous Solution*, Monographs in Electroanalytical Chemistry and Electrochemistry, IUPAC, New York: Marcel Dekker.

Bard, A. J., Lund, H. (Eds.) (1973–1980), *Encyclopedia of Electrochemistry of the Elements*, New York: Marcel Dekker.

Baizer, M. M. (1973), *Organic Electrochemistry*, New York: Marcel Dekker.

Bloom, I., Gutmann, F. (Eds.) (1977), *Electrochemistry, the Past Thirty Years and the Next Thirty Years, a Volume in Honor of J. O'M. Bockris*, New York: Plenum.

Bockris, J. O'M., Conway, B. E., Yeager, E. B. (Eds.) (1980), *Comprehensive Treatise of Electrochemistry: The Double Layer*, Vol. 1, New York: Plenum.

Bockris, J. O'M., Conway, B. E., Yeager, E. B., White, R. E. (Eds.) (1981), *Comprehensive Treatise of Electrochemistry: Electrochemical Processing*, Vol. 2, New York: Plenum.

Bockris, J. O'M., Conway, B. E., Yeager, E. B., White, R. E. (Eds.) (1981), *Comprehensive Treatise of Electrochemistry: Electrochemical Energy Conversion and Storage*, Vol. 3, New York: Plenum.

Bockris, J. O'M., Conway, B. E., Yeager, E. B., and White, R. E. (Eds.) (1981), *Comprehensive Treatise of Electrochemistry: Electrochemical Materials Science*, Vol. 4, New York: Plenum.

Bockris, J. O'M., Khan, S. U. M. (1979), *Quantum Electrochemistry*, New York: Plenum.

Bockris, J. O'M., Reddy, A. K. N. (1970), *Modern Electrochemistry*, (in two volumes), New York: Plenum.

Conway, B. E. (1965), *Theory and Principles of Electrode Processes*, New York: Ronald Press.

Conway, B. E., Bockris, J. O'M., Yeager, E. B. (Eds.) (1983), *Comprehensive Treatise of Electrochemistry: Thermodynamic and Transport Proper-*

ties of Aqueous and Molten Electrolytes, Vol. 5, New York: Plenum.

Conway, B. E., Bockris, J. O'M., Yeager, E. B., Khan, S. U. M., White, R. E. (Eds.) (1983) *Comprehensive Treatise of Electrochemistry: Kinetics and Mechanisms of Electrode Processes*, Vol. 7, New York: Plenum.

Delahay, P. (1965), *Double Layer and Electrode Kinetics*, New York: Wiley.

Fleischmann, M., Pons, S., Rolison, D. R., Schmidt, P. P. (Eds.) (1987), *Ultramicroelectrodes*, Morganton NC: Datatech Systems.

Gale, R. J. (Ed.) (1988), *Spectroelectrochemistry: Theory and Practice*, New York: Plenum.

Gileadi, E., Kirowa-Eisner, E., Penciner, J. (1975), *Interfacial Electrochemistry*, Reading, MA: Addison-Wesley.

Goodisman, J. (1987), *Electrochemistry: Theoretical Foundations*, New York: Wiley.

Gutierrez, C., Melendres, C. (Eds.) (1990), *Spectroscopic and Diffraction Techniques in Interfacial Electrochemistry*, Boston: Kluwer.

Hine, F. (1985), *Electrode Processes and Electrochemical Engineering*, New York: Plenum.

Kissinger, P. T., Heineman, W. R. (Eds.) (1984), *Laboratory Techniques in Electroanalytical Chemistry*, New York: Marcel Dekker.

Koryta, J. (1992), *Ions, Electrodes and Membranes*, New York: Wiley.

Koryta, J., Dvorak, J., Bohackova, V. (1966), *Electrochemistry*, London: Methuen.

Koryta, J., Dvorak, J., Bohackova, V. (1987), *Principles of Electrochemistry*, New York: Wiley.

Macdonald, D. D. (1977), *Transient Techniques in Electrochemistry*, New York: Plenum.

MacInnes, D. A. (1961), *The Principles of Electrochemistry*, New York: Dover.

Milazzo, G., Caroli, S. (Eds.) (1977), *Tables of Standard Electrode Potentials*, New York: Plenum.

Newman, J. S. (1972), *Electrochemical Systems*, Englewood Cliffs, NJ: Prentice-Hall.

Parsons, R. (1974), *Pure Appl. Chem.* **37**, 501–516 (Manual of Symbols and Terminology for Physicochemical Quantities and Units, Appendix III, Electrochemical Nomenclature).

Pletcher, D., Walsh, F. C. (1989), *Industrial Electrochemistry*, London: Chapman and Hall.

Silva, A. F. (Ed.) (1984), *Trends in Interfacial Electrochemistry*, New York: Reidel.

Soriaga, M. P. (Ed.) (1988), *Electrochemical Surface Science, Molecular Phenomena at Electrode Surfaces*, ACS Symposium Series, Vol. 378, Washington, D.C.: American Chemical Society.

Srinivasan, S., Chizmadzhev, Y. A., Bockris, J. O'M., Conway, B. E., Yeager, E. B. (Eds.) (1985), *Comprehensive Treatise of Electrochemistry: Bioelectrochemistry*, Vol. 10, New York: Plenum.

Stock, J. T., Orna, M. V. (Eds.) (1989), *Electrochemistry, Past and Present*, ACS Symposium Series, Vol. 390, Washington, D.C.: American Chemical Society.

Varma, R., Selman, J. R. (1991), *Techniques for Characterization of Electrodes and Electrochemical Processes*, New York: Wiley.

Vetter, K. J. (1967), *Electrochemical Kinetics*, New York: Academic.

White, R. E., Bockris, J. O'M., Conway, B. E., Yeager, E. B. (Eds.) (1984), *Comprehensive Treatise of Electrochemistry: Experimental Methods in Electrochemistry*, Vol. 8, New York: Plenum.

Yeager, E. B., Bockris, J. O'M., Conway, B. E., Sarangapani, S. (Eds.) (1983), *Comprehensive Treatise of Electrochemistry: Electrodics: Transport*, Vol. 6, New York: Plenum.

Yeager, E. B., Bockris, J. O'M., Conway, B. E., Sarangapani, S. (Eds.) (1984), *Comprehensive Treatise of Electrochemistry: Electrodics: Experimental Techniques*, Vol. 9, New York: Plenum.

Review Series

Bard, A. J. (Ed.) *Electroanalytical Chemistry*, Vols. 1–6, New York: Marcel Dekker.

Bockris, J. O'M., Conway, B. E. (Eds), *Modern Aspects of Electrochemistry*, Vols. 1–22, New York: Plenum.

Delahay, P., Tobias, C. W. (from Vol. 10, Gerischer, H., Tobias, C. W.) (Eds.), *Advances in Electrochemistry and Electrochemical Engineering*, Vols. 1–12, New York: Wiley.

Gerischer, H., Tobias, C. (Eds.) (1990, 1992), *Advances in Electrochemical Science and Technology*, Vols. 1–2, New York: VCH.

Lipkowski, J., Ross, P. (Eds.) (1992), *Frontiers in Electrochemistry*, Vol. 1, New York: VCH.

Specialist Periodical Reports (1971–1985), *Electrochemistry*, Vols. 1–10, London: The Chemical Society.

Yeager, E. B., Salkind, A. J. (Eds.) (1974–1978), *Techniques of Electrochemistry*, Vols. 1–3, New York: Wiley.

ELECTRODYNAMICS, CLASSICAL

J. D. JACKSON, *Department of Physics, University of California at Berkeley, Berkeley, California, U.S.A.*

Introduction 1

1. Units, Symbols, and Conventions 3

2. The Maxwell Equations in Vacuum in Differential and Integral Form 4
2.1 Coulomb's Law 4
2.2 Ampère–Maxwell Law 4
2.3 Faraday's Law 5
2.4 Law of Magnetic Flux 8
2.5 Limitations 8

3. Material Media, Constitutive Relations, Boundary Conditions, and Simple Dielectric and Magnetic Properties of Matter 8
3.1 Macroscopic Maxwell Equations 8
3.2 Constitutive Relations 9
3.3 Boundary Conditions at Interfaces between Different Media 11
3.4 Frequency Dependence of the Dielectric Constant 12
3.5 Causality and Analytic Properties of $\epsilon(\omega)$ 14

4. Electrostatic and Magnetostatic Energies, Capacitance, and Inductance 8
4.1 Electrostatic Energy 8
4.2 Capacitance and Coefficients of Capacitance 9
4.3 Magnetostatic Energy 11
4.4 Inductance 12
4.5 Symmetry and Reciprocity Properties 14

5. Potentials, Gauge Transformations, and Wave Equations 15
5.1 Scalar and Vector Potentials and Gauge Transformations 15
5.2 Green's Functions for the Wave Equation 16
5.3 Harmonic Fields and Potentials in Uniform, Linear, Isotropic, Dispersive Media 18

6. Conservation Laws for Energy and Momentum 21
6.1 Poynting's Theorem of Conservation of Electromagnetic Energy 21
6.2 Poynting's Theorem for Harmonic Fields and Field Definitions of Impedance and Admittance 22
6.3 Conservation of Momentum, Maxwell Stress Tensor, and Radiation Pressure 24

7. Special Relativity and Classical Field Theory 25
7.1 Summary of Results in Special Relativity 25
7.2 Relativistic Formulation of Electrodynamics 26
7.3 Lagrangian Formulation of Electrodynamics 28
7.4 Further Ramifications 29

Glossary 33

Works Cited 36

Further Reading 36

INTRODUCTION

Electromagnetism is ultimately at the heart of almost everything we are and do. Indeed, our world and the worlds of other animate and inanimate objects function almost exclusively on electromagnetic interactions. Atoms, the basic building blocks, are bound in their unique structures by electric forces (with dynamics subject to the laws of quantum mechanics). The admittedly nonelectromagnetic energy source, the Sun, transmits its energy to us by electromagnetic radiation.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

That radiation is converted by photosynthesis into other forms of energy, stored electromagnetic energy inside complex molecules. More immediately, the process of reading this page occurs through the scattering of light to the reader's eye, where it is absorbed and converted to electrical signals sent to the brain. We human beings and the plants and rocks around us are basically electromagnetic devices, interacting and communicating by means (whether chemical or physical) that are fundamentally electric in nature. Our societies are increasingly governed overtly by electricity and electronics.

The history of the development of our understanding of electricity and magnetism and of light and other electromagnetic radiation is a fascinating subject in itself. Here only the briefest account, with many omissions, can be given. The reader wishing to read a full and authoritative history should consult Whittaker (1951).

The properties of amber (ηλεκτρον, Greek) and lodestone were known in ancient times, but the creation of electrodynamics as a quantitative, mathematical subject occurred in about one hundred years (1766–1864). William Gilbert (1544–1603), in his book *De magnete,* published in 1600, clearly differentiated between electrical and magnetic phenomena. By 1733 the distinction had been made between resinous and vitreous electricity (surplus and deficiency of electrons, respectively). Forty years later, electrostatics was developed rapidly and quantitatively in about twenty years. Benjamin Franklin (1706–1790) made a crucial observation that was recognized by Joseph Priestley (1733–1804) as implying the inverse square law for electrical forces (1766). Henry Cavendish (1731–1810) conducted his remarkable experiments in electrostatics from 1771 to 1773. Charles Augustin Coulomb (1736–1806) began publishing his monumental researches in 1785.

These years marked the beginning of quantitative research in electricity and magnetism on a worldwide scale. Technical developments such as the invention of the battery in 1800 by Alessandro Volta (1745–1827) hastened progress in the study of electrical currents and their interactions. By 1820 Hans Christian Oersted (1777–1851) was observing the deflection of a compass needle by the current in a nearby wire, and André Marie Ampère (1775–1836) began publishing his results on the mechanical forces between current-carrying wires. By 1825 he had a mathematical description. In 1831 Michael Faraday (1791–1867) discovered his law of electromagnetic induction; in 1836 he used the famous Faraday cage to establish the law of conservation of electric charge. Faraday's many researches paved the way for the brilliant synthesis by James Clerk Maxwell (1831–1879), a dynamical theory of the electromagnetic field, published in 1864. Along the way, Karl Friedrich Gauss (1777–1855) and Simeon Denis Poisson (1781–1840) contributed importantly to the mathematical development, while practical concepts such as resistance were discovered (1827) by Georg Simon Ohm (1787–1854).

While Maxwell's paper of 1864 provided a complete classical description of electricity, magnetism, and light, its correctness and far-reaching applicability were not understood for some years. Not until 1888, when Heinrich Rudolf Hertz (1857–1894) gave a brilliant demonstration of the generation, transmission, polarization, reflection, and refraction of what we now know as radio waves, were the unity of light, Hertz's waves, and static electricity and magnetism fully appreciated. By the 1890s the basic principles of electromagnetism were thoroughly understood and ripe for the development of the myriad applications in the last hundred years.

This article is concerned with the fundamental principles of classical electrodynamics. Some derivations are sketched, but space limitations prevent lengthy proofs. Many results are just stated. References for details or elaboration are given in the text and also in the Further Reading section at the end. Applications are discussed elsewhere, under Electromagnetic Technology or Optics, for instance, as are more detailed treatments of special aspects, for example, ELECTROSTATICS, MAGNETOSTATICS, and ELECTROMAGNETIC RADIATION. The motion of charged particles under the action of external electric and magnetic fields is an important topic in its own right. The reader interested in these aspects should consult the articles on CHARGED-PARTICLE OPTICS; ACCELERATORS, LINEAR; CYCLOTRONS; SYNCHROTRONS; BETATRONS; PLASMA PHYSICS; INTENSE PARTICLE BEAMS; and FUSION, MAGNETIC CONFINEMENT.

The four laws that comprise the Maxwell equations are described physically and then mathematically in differential and integral form, with brief comments on the limitations imposed by the discreteness of physical charges and of quantum-mechanical modifications. The equations governing macroscopic fields inside material media, constitutive relations, and boundary conditions on fields at media interfaces precede a treatment of the frequency dependence of simple dielectric constants. The concept of causality is used to develop relations between the absorptive and dispersive parts of the dielectric constant (dispersion relations). The energies associated with electric and magnetic fields are exploited to define coefficients of capacitance and inductance, quantities useful in electric circuit analysis, and some examples of both are given. The scalar and vector potentials are introduced and their arbitrariness discussed (gauge transformations). Solutions of the wave equations in terms of the potentials are derived and the appropriateness of advanced and retarded solutions discussed. A general formula for radiation by a harmonically varying source is given. Poynting's theorem on the conservation of energy for fields and sources is derived, as is the conservation of momentum. The lumped-circuit concepts of impedance and admittance are discussed in terms of the fields by means of Poynting's theorem for harmonic fields. Because of its close connection to electrodynamics, the main ideas and results of special relativity are presented, along with a relativistic formulation of electrodynamics and a brief treatment of classical Lagrangian field theory.

1. UNITS, SYMBOLS, AND CONVENTIONS

The dimensions and units of quantities in electricity and magnetism are often viewed as contentious, with much heat and little light generated in polemics on the subject. The truth is that convenience of use is the only sensible criterion. Provided the schema is consistent, the choices of dimensions and units of magnitude of electromagnetic quantities are largely arbitrary, with no set more fundamental than any other. The interrelations among the various physical quantities through the laws of electrodynamics means, of course, that one does not have complete freedom of choice. The interested reader may consult the appendix of Jackson (1975) for a nonpartisan discussion. Traditionally (because mechanics preceded electromagnetism as a coherent subject), mass, length, and time are viewed as distinct in their dimensions. Systems of units differ in the amounts defined as unit magnitudes of each.

In this article the Système International (SI) of units is used, except when explicitly stated otherwise. In electromagnetism, the units of mass (the kilogram, abbreviated as kg), length (the meter, abbreviated as m), and time (the second, abbreviated as s) are augmented by a fourth, the unit of charge (the coulomb, abbreviated as C). The second is defined in terms of the frequency of a specific atomic transition, while the meter is defined as the distance light travels in vacuum in 1/299 792 458 seconds, or equivalently such that the speed of light in vacuum is $c = 299\,792\,458$ m/s. The electromagnetic parameters, the permittivity (ϵ_0) and the permeability (μ_0) of free space, are therefore both without experimental uncertainty. The permeability of free space is defined to be $\mu_0 = 4\pi \times 10^{-7}$ H/m, while the permittivity is $\epsilon_0 = 10^7/(4\pi c^2)$ F/m, where the henry (H) and the farad (F) are units of inductance and capacitance, respectively (see below). With the speed of light as a nine-digit number, $(4\pi\epsilon_0)^{-1} = 10^{-7}c^2$ is an exact 17-digit number (times a power of 10), the first six of which are $8.987\,55 \times 10^9$. To the same accuracy, $\epsilon_0 = 8.854\,19 \times 10^{-12}$ F/m. The quantity $\sqrt{\mu_0/\epsilon_0} = 376.730\ \Omega$ is called the impedance of free space. (See METROLOGY; ELECTROMAGNETIC QUANTITIES, MEASUREMENT OF BASIC.)

Harmonically varying quantities, or superpositions of them, play an important role in electromagnetism. Time-averaged entities such as power and energy are conveniently handled by means of complex numbers. The conventions are as follows.

Complex numbers:

$z = x + iy$, where $i = \sqrt{-1}$

$\mathrm{Re}(z) = x$, $\mathrm{Im}(z) = y$,

$z = \rho e^{i\theta}$, where $\rho = |z| = \sqrt{x^2 + y^2}$ and

$\theta = \tan^{-1} y/x$

$z^* = x - iy = \rho e^{-i\theta}$.

Harmonic quantities:

$$A(t)=A_0\cos(\omega t-\alpha)$$
$$=\mathrm{Re}(A_0e^{i\alpha-i\omega t})$$
$$\to A_0e^{i\alpha-i\omega t}.$$

Time averages:

$$\langle A(t)B(t)\rangle=\tfrac{1}{2}\mathrm{Re}[A(t)B^*(t)],$$

where $A(t)$ and $B(t)$ are understood to be complex amplitudes whose real parts correspond to the real harmonically varying physical quantities. It is often useful to discuss phenomena in either the time (space) or the frequency (wave-number) domain. An important tool is the Fourier integral, written here for time–frequency conversion, but easily adapted to space–wave-number domains in more than one dimension.

Fourier integrals:

$$A(\omega)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}dt\,A(t)e^{i\omega t}, \tag{1}$$

$$A(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}d\omega\,A(\omega)e^{-i\omega t}.$$

If $A(t)$ is real, then $A(-\omega)=A^*(\omega)$. Subject to appropriate existence and convergence properties, the following relation holds between $A(t)$ and $A(\omega)$:

$$\int_{-\infty}^{\infty}|A(\omega)|^2d\omega=\int_{-\infty}^{\infty}|A(t)|^2\,dt. \tag{2}$$

If $A(t)$ is real, the frequency integral can be written as an integral over only positive frequencies of $2|A(\omega)|^2$.

2. THE MAXWELL EQUATIONS IN VACUUM IN DIFFERENTIAL AND INTEGRAL FORM

The basic equations governing electromagnetic phenomena are the Maxwell equations, supplemented by constitutive relations describing the responses of material media. We first summarize the Maxwell equations in vacuum in terms of the electric and magnetic fields, $\mathbf{E}(\mathbf{r},t)$ and $\mathbf{B}(\mathbf{r},t)$, and electric charge and current densities, $\rho(\mathbf{r},t)$ and $\mathbf{J}(\mathbf{r},t)$.

2.1 Coulomb's Law

The first of the Maxwell equations stems from the 18th-century observations of Cavendish, Coulomb, Franklin, and others on the forces between electrically charged objects. The conclusion that the force between small charged bodies ("point charges") varies as the product of the charges, inversely as the square of the distance between them, and along the line joining them is expressed mathematically by the divergence equation,

$$\nabla\cdot\mathbf{E}=\rho/\epsilon_0, \tag{3}$$

where $\mathbf{E}(\mathbf{r},t)$ is the electric field strength, defined as the force per unit charge, with units of volts per meter (V/m), while $\rho(\mathbf{r},t)$ is the electric charge density in units of coulombs per cubic meter (C/m^3). By means of the divergence theorem of vector calculus, Eq. (3) can be cast into integral form,

$$\int_S \mathbf{n}\cdot\mathbf{E}\,da=\frac{1}{\epsilon_0}\int_V \rho\,d^3x. \tag{4}$$

The volume V is bounded by the closed surface(s) S, with $\mathbf{n}$ being the outwardly directed normal to its surface. The left-hand integral is thus the total or integral flux of electric field emerging from the volume; the right-hand integral is the total charge inside, divided by ϵ_0. Equation (4) is the integral electric flux theorem, or integral form of Coulomb's law. The simplest application of Eq. (4) is a point charge Q at the center of a sphere of radius r. The electric field points radially outward and has magnitude $|\mathbf{E}(\mathbf{r})|=Q/4\pi\epsilon_0r^2$. (See ELECTROSTATICS.)

2.2 Ampère–Maxwell Law

The second Maxwell equation concerns magnetic fields caused by currents and is known as the Ampère–Maxwell equation. Ampère's researches in the early 19th century were the study of forces between current loops carrying time-independent currents; Maxwell's contribution in mid-19th century was to recognize that Ampère's equation needed modification for time-varying sources and fields. The abstraction of a magnetic flux density **B** from the forces between currents is inherently more complex than that for the electric field **E** because there are no "point magnetic charges" [see the fourth Maxwell equation, Eq. (18)]. The magnetic force at a point (on a moving point electric charge, for example) is caused by a distribution of current; furthermore, the force is not central. The increment of magnetic flux density $d\mathbf{B}$ caused by an infinitesimal line element of current $I\, d\mathbf{l}$ is

$$d\mathbf{B}=\frac{\mu_0}{4\pi}\frac{I\, d\mathbf{l}\times\mathbf{r}}{r^3}, \tag{5}$$

where $\mathbf{r}$ is the coordinate vector from the current element to the observation point and I is the current (units of amperes=coulombs per second, abbreviated as A). The units of the magnetic induction are tesla (abbreviated as T). In reality, one must sum over all the current elements in a closed circuit, but the infinitesimal expression shows the parallels and differences with the electric field of a point charge q, $\mathbf{E}=q\mathbf{r}/4\pi\epsilon_0 r^3$. The magnetic flux density $\mathbf{B}$ of a long straight wire of circular cross section carrying a current I points in the azimuthal direction (circles centered on the wire, with sense given by the right-hand rule) and has a magnitude $|\mathbf{B}|=\mu_0 I/2\pi d$, where d is the perpendicular distance from the wire (Biot–Savart law).

If an infinitesimal current element $I\, d\mathbf{l}$ is in a magnetic flux density $\mathbf{B}$, the infinitesimal force on it is $d\mathbf{F}=I\, d\mathbf{l}\times\mathbf{B}$, a relation that may be viewed as an empirical result, but that has a more fundamental basis in the Lorentz force on a charge q in motion with a velocity $\mathbf{v}$, $\mathbf{F}=q(\mathbf{E}+\mathbf{v}\times\mathbf{B})$. The combination of Eq. (5) with the expression for the force on a current in a magnetic flux density permits the definition of the unit of current (1 A) in terms of the mechanical force per unit length between two very long, straight, and parallel current-carrying wires with separation d, namely,

$$\frac{dF}{dl}=\frac{\mu_0}{2\pi}\frac{I_1 I_2}{d}. \tag{6}$$

The force is attractive if the currents flow in the same direction and repulsive for opposite flow. Since μ_0 is defined to have a definite numerical value, measurements of the mechanical force per unit length and the separation of the wires permit the current to be defined without reference to the magnetic flux density $\mathbf{B}$. Note, however, that electrodynamics is involved, because the meter is defined in terms of the speed of light. It is amusing that one of Ampère's basic observations in the 1820s is the basis for the accepted definition of the SI unit of current.

The general expression for the force on a closed circuit C_1, fixed in position and carrying a steady current I_1, caused by a steady current I_2 flowing in a fixed closed circuit C_2 is

$$\mathbf{F}_{12}=-\frac{\mu_0 I_1 I_2}{4\pi}\oint\oint\frac{(d\mathbf{l}_1\cdot d\mathbf{l}_2)\mathbf{r}_{12}}{r_{12}^3}, \tag{7}$$

where the integration is over both of the closed circuits.

In differential form, the Ampère–Maxwell equation relating the magnetic flux density to the current density is

$$\nabla\times\mathbf{B}=\mu_0\mathbf{J}+\mu_0\epsilon_0\frac{\partial\mathbf{E}}{\partial t}, \tag{8}$$

where the right-hand side is the sum of the conduction current density (A/m^2) times μ_0 and the Maxwell "displacement current" density. The displacement-current term is necessary to make Eq. (8) compatible with Eq. (3), Coulomb's law, and the conservation of charge and current densities,

$$\nabla\cdot\mathbf{J}+\frac{\partial\rho}{\partial t}=0. \tag{9}$$

Use of Stokes's theorem from vector calculus converts Eq. (8) into the integral statement of Ampère's law:

$$\int_C \mathbf{B}\cdot d\mathbf{l}=\int_S \mathbf{n}\cdot\left(\mu_0\mathbf{J}+\mu_0\epsilon_0\frac{\partial\mathbf{E}}{\partial t}\right)da. \tag{10}$$

The left-hand side is the line integral of $\mathbf{B}$ along a closed path C, while the right-hand side is μ_0 times the total conduction and displacement current flowing through the open area bounded by C. (For more on Ampère's law and its use, see MAGNETOSTATICS.)

2.3 Faraday's Law

The third of the Maxwell equations is Faraday's law, codifying the observations of Faraday in the first half of the 19th century that time-varying magnetic fields produce electromotive forces. In differential form, Faraday's law reads

$$\nabla\times\mathbf{E}+\frac{\partial\mathbf{B}}{\partial t}=\mathbf{0}. \tag{11}$$

Its integral form, obtained by application of Stokes's theorem for an open surface S bounded by a closed curve C, is

$$\int_C \mathbf{E}\cdot d\mathbf{l}=-\int_S \frac{\partial\mathbf{B}}{\partial t}\cdot\mathbf{n}\; da. \tag{12a}$$

If the circuit C is stationary, the time derivative can be taken outside the integral sign to give

$$\int_C \mathbf{E}\cdot d\mathbf{l}=-\frac{d}{dt}\int_S \mathbf{B}\cdot\mathbf{n}\; da. \tag{12b}$$

In words, the electromotive force (emf) or line integral of the electric field around the closed circuit C (left-hand side) is equal to the negative of the time rate of change of the magnetic flux through the surface S (right-hand side). The negative sign reflects the empirical observation called Lenz's law, whereby the current induced by the emf tends to counteract the change in magnetic field.

Application of Faraday's law often involves nonstationary circuits. Care must be taken in using Eqs. (12a) or (12b). To understand the exact meaning of Faraday's law, we go back to Faraday's observations of 1831. He discovered that transient current flow occurred in a circuit when conditions were changed in adjacent current-carrying circuits. Specifically, he found a transient current induced when (a) the steady current flowing in a nearby circuit was turned on or off, (b) a nearby circuit with a steady current was moved relative to the first, or (c) a permanent magnet was thrust into or out of the circuit. No transient occurred unless the adjacent circuit's current changed or there was relative motion of the two circuits. Other observations showed that alteration of the shape of the first circuit in the presence of a circuit with a steady current caused a transient current in the first while its shape was being altered. Faraday interpreted the transient current in every situation as being caused by a changing magnetic flux linking the observed circuit. The current in the circuit is driven, according to Ohm's law, by the force acting on the charge carriers in the wire. We are thus led to consider the current to be proportional to the electromotive force $\mathscr{E}$, defined for a stationary circuit as the line integral of the electric field around the path of the circuit, as given by the left-hand sides of Eqs. (12a) and (12b). With the flux through the circuit defined as

$$\Phi_M=\int_S \mathbf{B}\cdot\mathbf{n}\; da, \tag{13}$$

where S is any open surface spanning the circuit C, Faraday's law in integral form is

$$\mathscr{E}=-\frac{d\Phi_M}{dt}. \tag{14}$$

To accommodate the possibility of movement of the circuit in time, we must generalize the definition of the electromotive force $\mathscr{E}$. Motion of the first circuit in the presence of a stationary second circuit or the same relative motion with the first circuit stationary produces the same transient current. But if an element $d\mathbf{l}$ of the first circuit is in motion with velocity $\mathbf{v}$, the charge carriers comoving with that element experience a force per unit charge equal to $\mathbf{E}+\mathbf{v}\times\mathbf{B}$. This means that the total electromotive force $\mathscr{E}$ causing the current flow in a circuit C whose path elements have velocity $\mathbf{v}(\mathbf{l})$ must be

$$\mathscr{E}=\oint_C (\mathbf{E}+\mathbf{v}\times\mathbf{B})\cdot d\mathbf{l}. \tag{15}$$

With Eq. (15) for the electromotive force around a circuit in motion (translating and perhaps changing shape if $\mathbf{v}$ is different for different parts of the path), Eq. (14) generalizes Eq. (12b) and is applicable to any situation. One sometimes considers the electromotive force between two points A and B in space. Then the integral in Eq. (15) is an open line integral between the points. In contrast to the electrostatic situation, the emf is in general path dependent. Care must be exercised in distinguishing between electrostatic potential differences and electromotive forces. With a closed path through which the magnetic flux is changing in time, parts of the path may be in regions where $\partial\mathbf{B}/\partial t=0$. Over such a segment the line integral of $\mathbf{E}$ can be interpreted as a potential difference without ambiguity, since $\nabla\times\mathbf{E}=\mathbf{0}$ there.

To connect Faraday's law, derived rather directly from the transient current flow in resistive circuits and embodied in Eq. (14), with the Maxwell equation, Eq. (11), one can consider how the magnetic flux Φ_M, Eq. (13), changes in time if the circuit C is in motion.

Whether C is in motion or not, the flux changes if $\mathbf{B}$ changes in time. There will therefore be a contribution to $d\Phi_M/dt$ from $\mathbf{n}\cdot\partial\mathbf{B}/\partial t$ at each point of the surface S. If the circuit C is in motion, there is an additional contribution whose nature can be seen from Fig. 1(a). In an infinitesimal time Δt the path $C(t)$ changes into $C(t+\Delta t)$. The surface spanning $C(t+\Delta t)$ can be taken as the surface spanning $C(t)$, minus the infinitesimal area around the sides, defined by the directed element of area, $d\mathbf{l}\times\mathbf{v}\Delta t$. Per unit time, the total change in magnetic flux is therefore

$$\frac{d\Phi_M}{dt}=\int_S \frac{\partial \mathbf{B}}{\partial t}\cdot\mathbf{n}\,da-\oint_C \mathbf{B}\cdot(d\mathbf{l}\times\mathbf{v}). \qquad (16)$$

Rearranging the triple scalar product and substituting Eqs. (15) and (16) into (14), one finds

$$\oint_C (\mathbf{E}+\mathbf{v}\times\mathbf{B})\cdot d\mathbf{l}=-\int_S \frac{\partial \mathbf{B}}{\partial t}\cdot\mathbf{n}\,da + \oint_C (\mathbf{v}\times\mathbf{B})\cdot d\mathbf{l}. \qquad (17)$$

The contribution to the time rate of change of the flux from the motion of the path C cancels the magnetic part of the electromotive force; Eq. (17) becomes Eq. (12a). We have thus returned to the third Maxwell equation, (11). It should be noted that these equations contain no reference to apparatus. Thus the path C and its spanning surface S can be viewed as mathematical abstractions. The integral and differential forms of Faraday's law in terms of $\mathbf{E}$ and $\mathbf{B}$ have a validity and generality far beyond their concrete origins.

A simple illustration of Faraday's law and its relation to other approaches is afforded by the situation sketched in Fig. 1(b). A conducting right circular cylinder of radius a rotates about its axis at angular speed $\omega=2\pi f$. There is a uniform magnetic induction parallel to the axis. Consider the open radial path OA, which is of length r and is rotating with the conductor. From Eq. (15), the emf from the axis at O to the point A is

$$\mathscr{E}_{OA}=\int_O^A (\mathbf{v}\times\mathbf{B})\cdot d\mathbf{r}=\tfrac{1}{2}\,\omega B r^2$$
$$=f(\pi r^2)B$$
$$=f\Phi_M(r).$$

Here $\Phi_M(r)$ is the magnetic flux contained within the radius r. The emf $\mathscr{E}_{OA}$ is equal to the rate at which the flux is cut by the rotating radial line.

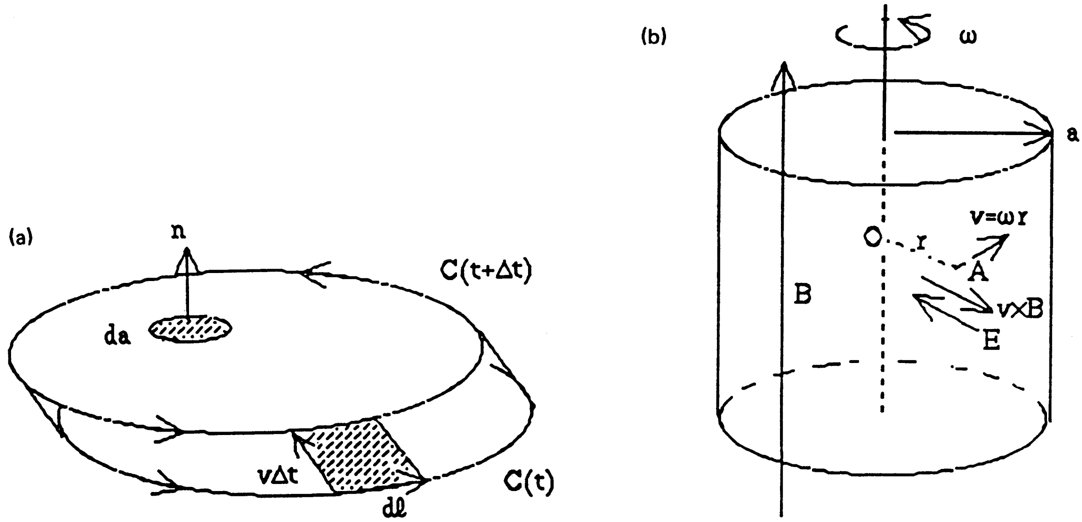

FIG. 1. (a) Closed circuit in Faraday's law. The closed path $C(t)$ changes in time Δt to $C(t+\Delta t)$. The surface spanning $C(t+\Delta t)$, with directed area element $\mathbf{n}da$, can be taken as that spanning $C(t)$, minus the "sides" defined by the directed area elements $d\mathbf{l}\times\mathbf{v}\Delta t$ around the path. (b) A long conducting right circular cylinder of radius a rotates about its axis at constant angular frequency ω in a uniform magnetic field $\mathbf{B}$ parallel to the axis.

An alternative way of approaching the problem is to consider Ohm's law in the rotating conductor. The current density is given by $\mathbf{J}=\sigma(\mathbf{E}+\mathbf{v}\times\mathbf{B})$. In the steady state no current flows. Within the conductor, there is an electrostatic field developed to oppose the magnetic force. At the typical point A (now considered as fixed in space, rather than rotating with the conductor), the electric field is radial and given by $E_r=-\omega Br$. There is a uniform volume charge density, $\rho=\epsilon_0\nabla\cdot\mathbf{E}=\epsilon_0(1/r)\partial(rE_r)/\partial r=-2\omega B\epsilon_0$. If the long cylinder is uncharged, there will be a surface charge density, $\sigma=\omega Ba\epsilon_0$. (There is a small magnetic field generated by the motion of these charge densities, but it is of higher order in smallness for surface speeds small compared to the speed of light.)

If sliding contacts fixed in space are made with the axis of the cylinder and its surface, a potential difference equal to the emf $\mathscr{E}_{OA}(a)$ or to the line integral of the electrostatic field,

$$V_{OA}=-\int_O^a E_r\,dr,$$

will be measured by a voltmeter. This example is qualitatively similar to one of Faraday's original experiments with a rotating copper disk.

2.4 Law of Magnetic Flux

The final Maxwell equation embodies the fact that, as far as is presently known, there are no magnetic charges or magnetic charge density. Its differential and integral forms parallel the first Maxwell equation, namely,

$$\nabla\cdot\mathbf{B}=0, \qquad (18)$$

$$\int_S \mathbf{B}\cdot\mathbf{n}\,da=0. \qquad (19)$$

Equation (19) is the integral magnetic flux theorem, stating that the total magnetic flux emerging from a closed surface S vanishes.

2.5 Limitations

The equations listed above are the Maxwell equations of *classical* electrodynamics in vacuum, an accurate and wonderfully successful description of myriad phenomena at the classical level. At the atomic and subatomic levels, there are quantum modifications. For example, basic to the discussion so far is the assumption that the charge and current densities are continuous, corresponding to infinitely divisible charges. But Nature supplies charges in discrete units, integral multiples of the electronic charge ($e=-1.602\times10^{-19}$ C). In many practical macroscopic applications, this unit is so small that the discreteness of charge densities can be ignored. An important relativistic quantum effect is the modification of the inverse square law of electrostatics at short distances ($r<\hbar/m_ec$, the electronic Compton wavelength). This modification, causing the force to be stronger at short distances and resulting in a tiny shift of atomic energy levels, is the result of "vacuum polarization" (see QUANTUM ELECTRODYNAMICS). A discussion of the idealizations and limitations of classical electrodynamics can be found in the introductory chapter of Jackson (1975).

3. MATERIAL MEDIA, CONSTITUTIVE RELATIONS, BOUNDARY CONDITIONS, AND SIMPLE DIELECTRIC AND MAGNETIC PROPERTIES OF MATTER

3.1 Macroscopic Maxwell Equations

The basic equations of electromagnetism of the previous section can be viewed as a description at the microscopic level, with the charge and current densities representing all the sources, whether free or bound. At length and time scales large compared to the atomic ($L\gg10^{-10}$ m, $T\gg10^{-17}$ s), the rapid fluctuations of the microscopic fields are unobservable. Only suitable averages over microscopically large, but macroscopically small, regions of space and/or time are relevant. Such averaged sources and fields are called macroscopic fields and sources. They satisfy the *macroscopic Maxwell equations* given below, and reflect the response of the material media to applied fields and sources.

In the process of averaging the microscopic equations, a distinction is made between bound charges (and atomic circulating currents) and the free charges (conduction electrons) whose averages define the familiar macroscopic charge and current densities, $\rho(\mathbf{r},t)$ and $\mathbf{J}(r,t)$:

$$\rho(\mathbf{r},t)=\langle\rho_{\text{free}}(\mathbf{r},t)\rangle,\quad \mathbf{J}(\mathbf{r},t)=\langle\mathbf{J}_{\text{free}}(\mathbf{r},t)\rangle, \qquad (20)$$

where the brackets ⟨ ⟩ indicate spatial averaging or its equivalent (Jackson, 1975, Sec. 6.7; Robinson, 1973, Chaps. 6–8). The contributions of the bound charges and currents are systematized in terms of the electric polarization $\mathbf{P}(\mathbf{r},t)$ [dominated by the averaged electric dipole moment density, but including in principle the electric quadrupole density $Q'_{jk}(\mathbf{r},t)$ and higher multipoles], and the magnetization $\mathbf{M}(\mathbf{r},t)$ (averaged magnetic dipole moment density, plus higher magnetic multipoles). These polarizations require the introduction of two new phenomenological fields $\mathbf{D}(\mathbf{r},t)$ and $\mathbf{H}(\mathbf{r},t)$, called the electric displacement and magnetic field strength, respectively, in the equations involving sources [Eqs. (3) and (8)]. The local macroscopic Maxwell equations are

$$\begin{aligned}
&\nabla\cdot\mathbf{D}=\rho,\\
&\nabla\times\mathbf{H}=\mathbf{J}+\frac{\partial\mathbf{D}}{\partial t},\\
&\nabla\times\mathbf{E}+\frac{\partial\mathbf{B}}{\partial t}=\mathbf{0},\\
&\nabla\cdot\mathbf{B}=0.
\end{aligned}\tag{21}$$

Here the electric field $\mathbf{E}$ and the magnetic induction $\mathbf{B}$ of the two homogeneous equations are just the averaged quantities of the corresponding microscopic equations. The new macroscopic fields that include the responses of the bound charges are

$$D_i=\epsilon_0E_i+P_i \tag{22}$$

and

$$H_i=\frac{1}{\mu_0}B_i-M_i. \tag{23}$$

Since $\mathbf{B}$ is the fundamental magnetic field quantity and $\mathbf{H}$ is phenomenological, Eq. (23) is the rational way to express the relationship. The minus sign associated with the magnetization is a natural one. It is conventional, however, to rearrange Eq. (23) and write the connections between the four fields as

$$\begin{aligned}
&\mathbf{D}=\epsilon_0\mathbf{E}+\mathbf{P},\\
&\mathbf{B}=\mu_0(\mathbf{H}+\mathbf{M}),
\end{aligned}\tag{24}$$

even though the basic physics is thereby obscured.

Integral statements of the Maxwell equations can be written down from Eqs. (21) in complete analogy with the development in Sec. 2.

3.2 Constitutive Relations

The Maxwell equations (21) are eight coupled partial-differential equations in space and time for the twelve field quantities **E, B, D**, and **H**. The system cannot be solved without the *constitutive relations* relating **D** and **H** to **E** and **B**:

$$\mathbf{D}=\mathbf{D}[\mathbf{E},\mathbf{B}],\quad \mathbf{H}=\mathbf{H}[\mathbf{E},\mathbf{B}]. \tag{25}$$

For conducting media, there is a generalized Ohm's law,

$$\mathbf{J}=\mathbf{J}[\mathbf{E},\mathbf{B}]. \tag{26}$$

The square brackets in Eqs. (25) and (26) signify that the connection may be nonlinear, nonlocal, and dependent on past history (hysteresis). The connections (24) show that the electric polarization and the magnetization of matter determine the nature of the constitutive relations. In general, these relations are far from simple, especially in crystalline solids, with ferroelectric and ferromagnetic materials having nonzero **P** or **M** even in the absence of applied fields. The study and elucidation of these properties is the province of condensed-matter physics. Only the simplest aspects are discussed here (for more detail, see Kittel, 1986; OPTICAL PROPERTIES OF SOLIDS; DIELECTRIC PROPERTIES OF INSULATORS; MAGNETOSTATICS; MAGNETIC MATERIALS).

In substances other than ferroelectrics or ferromagnets, a weak applied electric or magnetic field induces a polarization proportional to the magnitude of the applied field. The response of the medium is then said to be linear, and the constitutive relations take the form

$$D_i=\sum_{k=1}^{3}\epsilon_{ik}E_k,\quad B_i=\sum_{k=1}^{3}\mu_{ik}H_k. \tag{27}$$

The tensors ϵ_{ik} and μ_{ik} are called the electric permittivity or dielectric tensor and the magnetic permeability tensor, respectively. They are generally dependent on the frequency of the applied field, and on bulk properties of the material, such as density and temperature, as well as on details of molecular and perhaps crystalline structure. For simple materials (gases, liquids, some solids) the linear response is isotropic in space. Then $\epsilon_{ik}=\epsilon\delta_{ik}$ and $\mu_{ik}=\mu\delta_{ik}$; $\mathbf{D}=\epsilon\mathbf{E}$ and $\mathbf{B}=\mu\mathbf{H}$; $\epsilon/\epsilon_0=\epsilon_r$ is called simply the (relative) dielectric "constant," and $\mu/\mu_0=\mu_r$ the (relative) magnetic permeability. Many crystals have anisotropies in their dielectric constant (or magnetic permeability). The elements of the tensor ϵ_{ik} form a nonsingular matrix with eigenvalues ϵ_1, ϵ_2, ϵ_3, and associated principal axes in the crystal (which may or may not be along the axes of symmetry of the crystal). When expressed with respect to these axes, the tensor ϵ_{ik} takes the diagonal form $\epsilon_{ik}=\epsilon_i\delta_{ik}$. Unless the eigenvalues are all equal, such crystals (e.g., calcite or mica) exhibit birefringence and double refraction. See Hecht and Zajac (1974), Sec. 8.2.

For orientation, we mention that at low frequencies ($\nu<10^6$ Hz), where all charges, regardless of their masses, respond to an applied electric field, solids have dielectric constants typically in the range $\epsilon_r=\langle\epsilon_{ii}\rangle/\epsilon_0 \approx 2$–20, with larger values not uncommon. Systems with permanent molecular dipole moments can have much larger and temperature-sensitive dielectric constants. Distilled water, for example, has a static ($\nu=0$) dielectric constant of $\epsilon_r=80$ at 0 °C and $\epsilon_r=56$ at 100 °C. At optical frequencies only the electrons can respond significantly. The dielectric constants are frequency dependent in the range $\epsilon_r\approx 1.7$–10, with $\epsilon_r\approx 2$–3 for most solids. Water has $\epsilon_r\approx 1.77$–1.80 over the visible range, essentially independent of temperature from 0 to 100 °C.

Applied magnetic fields induce different responses, depending upon the properties of the individual atoms and also on their interactions. *Diamagnetic* materials consist of atoms or molecules with no net angular momentum. The response to an applied magnetic field is the creation of circulating atomic currents that produce a very small bulk magnetization opposed to the applied field ($\mu_{ii}<\mu_0$). Bismuth, the most diamagnetic substance known, has $1-\mu_r=1.8\times10^{-4}$. Diamagnetism is clearly a very small effect. If the basic atomic unit of the material has a net angular momentum from an unpaired electron, the substance is *paramagnetic*. The magnetic moment of the odd electron is aligned parallel to the applied field and $\mu>\mu_0$. Typical values are in the range $\mu_r-1\approx10^{-2}$–10^{-4} at room temperature, decreasing at higher temperatures because of the randomizing effect of thermal excitations. (See MAGNETOSTATICS.)

Ferromagnetic materials are paramagnetic (unpaired electrons), but show drastically different behavior because of interactions among the atoms. Below the Curie temperature (1040 K for iron, 630 K for nickel), ferromagnetic substances show spontaneous magnetization; all the atomic magnetic moments within a microscopically large region called a domain are aligned. Neighboring domains have random directions of alignment, but the application of a very modest external magnetic field tends to cause domains to merge or to line up in the same direction, leading to the saturation of the bulk magnetization. Removal of the field leaves a considerable fraction of the moments still aligned. The residual "permanent" magnetization can be as large as $\mu_0|\mathbf{M}|\gtrsim 1$ T. (See MAGNETS; MAGNETIC MATERIALS.)

For data on dielectric and magnetic properties of materials, the reader may consult Weast, *Handbook of Chemistry and Physics*, 69th ed. (1989).

Materials that show a linear response in weak fields eventually show *nonlinear behavior* at high enough field strengths as the electronic or ionic oscillators are driven to large amplitudes. The linear relations (27) are modified to, for example,

$$D_i=\sum_k \epsilon_{ik}^{(1)}E_k+\sum_{k,m}\epsilon_{ikm}^{(2)}E_kE_m+\cdots. \qquad (28)$$

For static fields the consequences are not dramatic, but for time-varying fields it is another matter. A large-amplitude wave of two frequencies ω_1 and ω_2 generates waves in the medium with frequencies 0, $2\omega_1$, $2\omega_2$, $\omega_1+\omega_2$, $\omega_1-\omega_2$, as well as the original ω_1 and ω_2. From cubic and higher nonlinear terms an even richer spectrum of signals can be generated. Lasers can produce peak electric fields $\gtrsim10^{12}$ V/m, comparable to or larger than the internal electric fields in atoms ($e/4\pi\epsilon_0a_0^2\approx5\times10^{11}$ V/m). The regime of nonlinearity is readily accessible for research in the laboratory and provides a vast array of applications. (See LASER PHYSICS; OPTICS, NONLINEAR.) The treatment here is restricted to linear phenomena.

3.3 Boundary Conditions at Interfaces between Different Media

At interfaces between different media, with different constitutive relations between **E** and **D** and **B** and **H**, and perhaps with surface charge and current densities, the various fields obey boundary conditions that can be inferred from the integral forms of the four Maxwell equations. Application of the divergence theorem to the first and last equations in (21) yields the integral statements

$$\int_S \mathbf{D}\cdot\mathbf{n}\, da = \int_V \rho\, d^3x \tag{29a}$$

and

$$\int_S \mathbf{B}\cdot\mathbf{n}\, da = 0. \tag{29b}$$

Here S is the closed surface bounding the volume V. Similarly, use of Stokes's theorem on the second and third equations in (21) yields the integral relations

$$\int_C \mathbf{H}\cdot d\mathbf{l} = \int_{S'} \mathbf{t}\cdot\left(\mathbf{J}+\frac{\partial \mathbf{D}}{\partial t}\right) da \tag{30}$$

and

$$\int_C \mathbf{E}\cdot d\mathbf{l} = -\int_{S'} \mathbf{t}\cdot\frac{\partial \mathbf{B}}{\partial t} da. \tag{31}$$

In (30) and (31) the surface S' is an open surface bounded by the closed curve C, and **t** is a unit normal to the surface S'.

For an interface between different media the appropriate geometrical arrangement of volumes and surfaces is shown in Fig. 2. The Gaussian pillbox of volume V, with top and bottom surfaces just inside each medium, and normal **n** defined as going from medium 1 to medium 2, is suitable for Eqs. (29a) and (29b). In the limit of vanishing height of the pillbox, these equations yield relations between the normal components of **D** and **B** on either side of the boundary surface, namely,

$$(\mathbf{D}_2-\mathbf{D}_1)\cdot\mathbf{n}=\sigma, \quad (\mathbf{B}_2-\mathbf{B}_1)\cdot\mathbf{n}=0. \tag{32}$$

In words, the normal component of **B** is continuous, while the discontinuity of the normal component of **D** is equal to the surface charge density σ at that point.

Applying Eqs. (30) and (31) to the infinitesimal Stokesian loop in Fig. 2, with sides of negligible length perpendicular to the surface, gives relations between the tangential components of **E** and **H** on either side of the interface,

$$\mathbf{n}\times(\mathbf{E}_2-\mathbf{E}_1)=\mathbf{0}, \quad \mathbf{n}\times(\mathbf{H}_2-\mathbf{H}_1)=\mathbf{K}. \tag{33}$$

The surface current **K** (with dimensions of current per unit length and units A/m) is assumed to have only components parallel to the interface. The tangential component of **E** is continuous across the interface, while the tangential component of **H** is discontinuous by an amount whose magnitude is equal to the magnitude of the surface current **K** and whose direction is parallel to $\mathbf{K}\times\mathbf{n}$. The discontinuity equations (32) and (33) for the normal and tangential components of the fields are useful in solving the Maxwell equations in different regions and then splicing them together to provide a solution throughout all space.

For completeness it can be noted that Eq. (9), the differential statement of charge–current-density conservation, leads to a corresponding boundary condition for the volume current density and the surface current and charge densities,

$$(\mathbf{J}_2-\mathbf{J}_1)\cdot\mathbf{n}+\nabla_{\parallel}\cdot\mathbf{K}+\frac{\partial\sigma}{\partial t}=0. \tag{34}$$

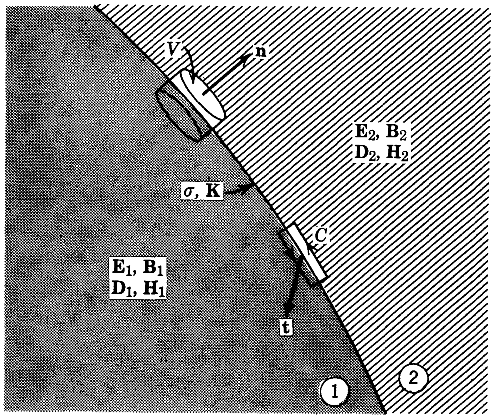

FIG. 2. Schematic diagram of boundary surface (heavy line) between different media. The boundary is assumed to carry idealized surface charge and current densities, σ and **K**. The volume V is a pillbox with its plane surfaces parallel to the interface. The normal **n** to the planes points from medium 1 to medium 2. The rectangular contour C, half in one medium and half in the other, is oriented so that the normal t to its spanning surface is tangent to the interface. [From Jackson (1975), reproduced by permission of John Wiley & Sons, Inc.]

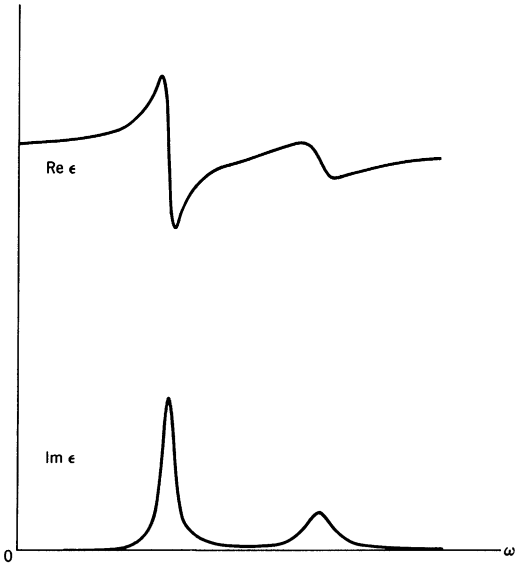

FIG. 3. Real and imaginary parts of the dielectric constant $\epsilon(\omega)$ as functions of frequency ω in the neighborhood of two resonances. [From Jackson (1975), reproduced by permission of John Wiley & Sons, Inc.]

3.4 Frequency Dependence of the Dielectric Constant

A completely correct description of the linear response of a material medium to applied electromagnetic fields requires the use of quantum mechanics, but the behavior of the dielectric constant can be inferred from largely classical considerations. Consider a model of the atom as a collection of Z electrons, each of mass m and charge $-e$, bound harmonically, with f_k of them having circular binding frequency ω_k and damping constant γ_k $(\Sigma_k f_k = Z)$. In the low-density limit in which the applied fields can be assumed to be those acting on the individual atoms, each electron obeys an equation of motion of the form

$$m(\ddot{\mathbf{x}}+\gamma_k\dot{\mathbf{x}}+\omega_k^2\mathbf{x}) = -e\mathbf{E}(0,t),$$

where the amplitude of oscillation is assumed small enough that the applied electric field can be evaluated at the average position of the charge (dipole approximation). For sinusoidal time dependence at frequency ω, the atomic dipole moment can be found as a sum over the contributions of all the electrons. With N atoms per unit volume, the dielectric constant at frequency ω is

$$\epsilon(\omega) = \epsilon_0 + \frac{Ne^2}{m} \sum_k f_k(\omega_k^2 - \omega^2 - i\omega\gamma_k)^{-1}. \quad (35)$$

With suitable quantum-mechanical definitions of f_k (called the oscillator strengths), γ_k, and ω_k, Eq. (35) is an accurate description of the electronic contributions to the dielectric constant. The relation $\Sigma_k f_k = Z$ is called the oscillator strength sum rule. Classically, it merely records the fact that there are Z electrons in the atom; quantum mechanically, it is a statement about the sum of the squares of the transitional dipole moments, weighted by the transitional frequencies, from a given quantum state to all others.

The damping constants γ_k in Eq. (35) are generally very small compared to the resonance frequencies ω_k. This means that $\epsilon(\omega)$ is approximately real for most frequencies, and, despite the fact that the overall trend with increasing frequency is downward, the most common behavior, called *normal dispersion,* is for $\mathrm{Re}[\epsilon(\omega)]$ to increase with frequency. The overall trend is produced by rather violent behavior in narrow regions around $\omega \approx \omega_k$, called regions of *anomalous dispersion,* in which $\mathrm{Re}[\epsilon(\omega)]$ decreases rapidly with frequency. A typical situation is illustrated in Fig. 3, which shows the behavior of the real and imaginary parts of $\epsilon(\omega)$ in the neighborhood of two resonances. In a region of anomalous dispersion, the imaginary part of $\epsilon(\omega)$ shows a typical resonant form. A significant $\mathrm{Im}[\epsilon(\omega)]$ represents dissipation of electromagnetic energy from the wave into the medium—such regions are called regions of resonant absorption. The fractional decrease in intensity $(\Delta I/I)$ of a monochromatic electromagnetic wave, per wavelength divided by 2π $(\Delta z = \lambda/2\pi)$, can be shown to be given by $\mathrm{Im}[\epsilon(\omega)]/\mathrm{Re}[\epsilon(\omega)]$.

Ohm's law is the statement that the electric field strength and the current density are related by $\mathbf{E} = \rho\mathbf{J}$, where the proportionality constant ρ, characteristic of the material, is called the electrical resistivity (units of ohms times meters, abbreviated as Ω m). Ohm's law holds in a wide range of practical situations, although it is clearly an approximation and sometimes a gross oversimplification. If it is inserted in the second Maxwell equation in Eq. (21) and the medium is linear, the right-hand side of the equation is linear in the electric field (and its time derivative). This suggests that electrical conductivity and dielectric response are actually two facets of the same phenomenon. Conduction is caused by the motion of quasifree electrons (and holes). Suppose a fraction f_0 of the "bound" electrons in the model for $\epsilon(\omega)$ are actually free, that is, they have no binding—their resonant frequency is zero. Their contribution to $\epsilon(\omega)$ in Eq. (35) can be separated off to give

$$\epsilon(\omega)=\epsilon_{\text{bound}}+i\frac{Ne^2f_0}{m\omega(\gamma_0-i\omega)}. \tag{36}$$

Because it is $-i\omega\epsilon(\omega)$ that enters the right-hand side of the Ampère–Maxwell equation, the singular behavior of the contribution from the free electrons represents a well-behaved resistivity,

$$\rho=m(\gamma_0-i\omega)/Ne^2f_0. \tag{37}$$

This expression is essentially the Drude (1900) expression for electrical resistivity, with f_0N being the number of free electrons per unit volume. The damping constant γ_0/f_0 can be determined empirically. For copper, $N\approx 8\times 10^{22}$ atoms/cm^3, and at normal temperatures the low-frequency resistivity is $\rho\approx 1.8\times 10^{-8}$ Ω m. This gives $\gamma_0/f_0\approx 3\times 10^{13}$ s^{-1}. With the assumption that $f_0=O(1)$, Eq. (37) shows that up to frequencies well beyond the microwave region ($\omega\leqslant 10^{11}$ s^{-1}) resistivities of metals are essentially real and independent of frequency. At higher frequencies the simple model fails in detail, but still describes the qualitative behavior. The reader may consult Wilson (1953) or Beam (1965) for details on the role of lattice vibrations, lattice imperfections, and impurities in determining the temperature dependences and magnitudes of resistivities.

At frequencies far above the highest atomic resonance, the expression (35) for $\epsilon(\omega)$ simplifies to

$$\epsilon_r(\omega)\approx 1-\omega_p^2/\omega^2,$$

where

$$\omega_p^2=NZe^2/\epsilon_0 m. \tag{38a}$$

The frequency ω_p, which depends only on the total number NZ of electrons per unit volume, is called the *plasma frequency* of the medium. In a metal, if $\omega\gg\gamma_0$, the conduction electrons give a term like the second one in Eq. (38a), so that one has

$$\epsilon_r(\omega)\approx\epsilon_{r,\text{bound}}-\omega_p^2/\omega^2,$$

where

$$\omega_p^2=ne^2/\epsilon_0 m^*. \tag{38b}$$

The plasma frequency is now that of the conduction electrons only, with an effective mass m^* to include effects of the binding (band structure). At frequencies such that $\omega\ll\omega_p$, the last term dominates; ϵ_r is negative; the index of refraction is purely imaginary; waves are reflected by the metal. When ω is sufficiently large that $\epsilon_r(\omega)>0$, waves can penetrate and propagate through the metal. Typically, this transition occurs in the ultraviolet and one speaks of the ultraviolet transparency of metals. Aluminum, for example, shows a sharp drop in reflectance for wavelengths shorter than about 80 nm, corresponding to a conduction-electron plasma energy of $\hbar\omega_p\approx 15$ eV compared to a theoretical value of 15.7 eV from Eq. (38b) with three conduction (valence) electrons per atom and $m^*=m$. For further discussion of this transition and other dielectric properties of metals in the optical and ultraviolet regions, see Pines (1963), Wooten (1972), and OPTICAL PROPERTIES OF SOLIDS.

We mentioned above that the free-electron contribution to the dielectric constant could be viewed alternatively as a representation of Ohm's law, while the bound electrons gave the usual dielectric constant. The dominance of one or the other contribution governs the behavior of fields in matter. Consider for simplicity a uniform isotropic linear medium with resistivity ρ, dielectric constant ϵ, and magnetic permeability μ, all real. For harmonic fields, the second and third Maxwell equations in Eq. (21), combined with the fourth and Ohm's law, lead to a Helmholtz wave equation for **H**,

$$\nabla^2\mathbf{H}+k^2\mathbf{H}=\mathbf{0},$$

where

$$k^2=i\omega\mu(\rho^{-1}-i\omega\epsilon)$$

is the square of the (complex) wave number. If the medium is poorly conducting, the Maxwell displacement current dominates and $k^2\approx\mu\epsilon\omega^2$. The wave vectors are approximately real, and the electric field is in time phase with the magnetic field. If, on the other hand, the resistivity is small enough (i.e., $\rho\omega\epsilon\ll 1$), the conduction current dominates and $k^2\approx i\omega\mu/\rho$. Then the wave numbers are complex,

$$k \approx \pm [(1+i)/\sqrt{2}]\,\sqrt{\omega\mu/\rho},$$

with equal real and imaginary parts. The electric field is 135° out of phase with the magnetic field. For the plane surface of a good conductor, the field components vary with distance x into the conductor as $e^{-x/\delta} \times \cos(x/\delta+\beta)$, where the skin depth δ is

$$\delta = \sqrt{2\rho/\omega\mu}. \qquad (39)$$

The behavior in other geometries is similar—for cylinders, for example, the fields are proportional to Kelvin's ber and bei functions, Bessel functions of complex argument. For copper ($\rho \approx 1.7\times 10^{-8}$ Ω m at 20 °C) the skin depth is numerically $\delta \approx 6.6/\sqrt{\nu(\mathrm{Hz})}$ cm. At 10 GHz, the skin depth in copper is less than 1 μm. The concept of skin or penetration depth in conducting media applies in several domains, from quasistatic induction heating to microwave components.

3.5 Causality and Analytic Properties of $\epsilon(\omega)$

The dielectric constant, Eq. (35), viewed as a function of complex ω, has a pattern of singularities and domain of analyticity that are a consequence of the causal nature of the response of the charged atomic particles to applied fields. If a single-frequency component of the field is considered, **D** and **E** at a given point in space are related by

$$\mathbf{D}(\mathbf{r},\omega) = \epsilon(\omega)\mathbf{E}(\mathbf{r},\omega). \qquad (40)$$

If use is made of the Fourier integrals in time and frequency [Eq. (1)], $\mathbf{D}(\mathbf{r},t)$ can be expressed nonlocally (in time) in terms of $\mathbf{E}(\mathbf{r},t)$ through

$$\mathbf{D}(\mathbf{r},t) = \epsilon_0\left(\mathbf{E}(\mathbf{r},t) + \int_{-\infty}^{\infty} G(\tau)\mathbf{E}(\mathbf{r},t-\tau)\,d\tau\right), \qquad (41)$$

where $G(\tau)$ is the Fourier transform of $\epsilon_r(\omega)-1$,

$$G(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} [\epsilon_r(\omega)-1]e^{-i\omega\tau}\,d\omega. \qquad (42)$$

$G(\tau)$ is called the susceptibility kernel. A simple, single-oscillator model of $\epsilon(\omega)$ is sufficient to illustrate its general features. If $\epsilon_r(\omega)$ is given by

$$\epsilon_r(\omega) = 1 + \omega_p^2(\omega_0^2-\omega^2-i\omega\gamma)^{-1}, \qquad (43)$$

then

$$G(\tau) = \omega_p^2 e^{-\gamma\tau/2}\,\frac{\sin \nu_0\tau}{\nu_0}\,\theta(\tau), \qquad (44)$$

with

$$\nu_0 = \sqrt{\omega_0^2 - \tfrac{1}{4}\gamma^2},$$

where $\theta(\tau)$ is the unit step function [$\theta(x)=1$ for $x>0$; $\theta(x)=0$ if $x<0$]. Evidently the susceptibility kernel for Eq. (35) is just a linear superposition of terms like Eq. (44). There are three major features: (a) $G(\tau)$ is causal, that is, it vanishes for $\tau<0$. This is a general feature, not restricted to the particular model used here. (b) $G(\tau)$ is oscillatory, at a frequency that is closely the resonant frequency ω_0 (since $\gamma \ll \omega_0$). (c) $G(\tau)$ is damped in time for $\tau>0$ with the damping constant of the electronic oscillator. These damping constants, quantum mechanically the widths in frequency of the relevant spectral lines, are typically 10^7–10^9 s^{-1}. The nonlocality in time in Eq. (41) is therefore of the order of 10^{-7}–10^{-9} s. For frequencies above the microwave region many earlier cycles of the electric field oscillations contribute to the value of **D** at a given instant of time. [Actually, the response of the medium is nonlocal in space as well as time, but for insulators the spatial nonlocality is only of the order of atomic sizes. The concept of a dielectric constant $\epsilon(\omega)$ or index of refraction $n(\omega)$ therefore applies far beyond the visible into the soft-x-ray region. For metals, especially at low temperatures, where electrons have very long mean free paths, the nonlocal region can be macroscopic in size. See Pippard (1960).]

Viewed as a function of complex ω, the dielectric constant, Eq. (35), or its special case, Eq. (43), has poles only in the lower half-plane. The absence of singularities above the real axis is directly responsible for the causal behavior of $G(\tau)$. The relation inverse to Eq. (42) expresses $\epsilon(\omega)$ in terms of $G(\tau)$ as

$$\epsilon_r(\omega) = 1 + \int_0^{\infty} G(\tau)e^{i\omega\tau}\,d\tau. \qquad (45)$$

From the reality of $G(\tau)$ it can be inferred from Eq. (45) that (a) for complex ω, $\epsilon(-\omega) = \epsilon^*(\omega^*)$; (b) $\epsilon(\omega)$ is an analytic function of ω in the upper half-plane and on the real axis ($\omega \neq 0$), provided $G(\tau)$ is finite for all τ. At the origin it is necessary to invoke the requirement that $G(\tau)\to 0$ as $\tau\to\infty$ to assure that $\epsilon(\omega)$ is also analytic there. This is true for insulators, but not for conductors. For the latter, $G(\tau)\to$const as $\tau\to\infty$ and $\epsilon(\omega)$ has a pole at $\omega=0$, as already discussed in connection with Eq. (36).

The analyticity of $\epsilon(\omega)$ in the upper half-plane permits us to use Cauchy's theorem to deduce "dispersion relations" or Kramers–Kronig relations (1926–27), as they are called. [See Jackson (1975), p. 310ff, for a derivation.] These relations express the real part of $\epsilon_r(\omega)$ as an integral over the imaginary part, and vice versa. For insulators (without the pole at $\omega=0$), the Kramers–Kronig relations are

$$\mathrm{Re}\ \epsilon_r(\omega)=1+\frac{2}{\pi}\ \mathrm{P}\int_0^\infty \frac{\omega'\ \mathrm{Im}\ \epsilon_r(\omega')}{\omega'^2-\omega^2}\,d\omega',$$

$$\mathrm{Im}\ \epsilon_r(\omega)=-\frac{2}{\pi}\ \mathrm{P}\int_0^\infty \frac{[\mathrm{Re}\ \epsilon_r(\omega')-1]}{\omega'^2-\omega^2}\,d\omega'. \tag{46}$$

Here the symbol P denotes "principal part" of the integral, namely, the integral over ω', omitting the symmetric interval around $\omega'=\omega$ from $\omega'=\omega'-\epsilon$ to $\omega'=\omega+\epsilon$, in the limit $\epsilon\to 0$. The power of these relations lies in their generality. Experimental information on absorption [Im $\epsilon(\omega)$] permits calculation of Re $\epsilon(\omega)$, etc. The asymptotic behavior of $\epsilon(\omega)$ for large ω, Eq. (38a), can be used as a definition of the plasma frequency, i.e., $\omega_p^2=\lim_{\omega\to\infty}\{\omega^2[1-\epsilon_r(\omega)]\}$. Application of this limit to the first equation in Eqs. (46) leads to a "sum rule" for the plasma frequency, equivalent to the sum rule for the oscillator strengths,

$$\omega_p^2=\frac{2}{\pi}\int_0^\infty \omega\ \mathrm{Im}\ \epsilon_r(\omega)d\omega. \tag{47}$$

This optical sum rule is one of many that can be developed (Altarelli *et al.*, 1972; Wooten, 1972).

The analyticity of $\epsilon(\omega)$ translates into analytic properties for its square root, the index of refraction, relevant for the propagation of electromagnetic waves in matter. In two historical papers in 1914, Sommerfeld and his student Brillouin (a) showed that no electromagnetic disturbance in matter can propagate faster than the speed of light in vacuum, despite the fact that the phase velocity of the wave in the material can exceed that speed; and (b) analyzed in detail the rather complex arrival of the signal at a distant point. See Sec. 7.11 of Jackson (1975) for a simplified account, including discussion of an experimental demonstration, and Brillouin (1960) for translations of the original papers and subsequent work.

4. ELECTROSTATIC AND MAGNETOSTATIC ENERGIES, CAPACITANCE, AND INDUCTANCE

(See also ELECTROSTATICS; MAGNETOSTATICS; CIRCUIT ELEMENTS; ELECTROSTATIC CAPACITIVE STORAGE OF ENERGY; INDUCTORS.)

4.1 Electrostatic Energy

For static fields Faraday's law states that the electric field has zero curl. It can therefore be written as the gradient of a scalar potential function $\phi(\mathbf{r})$, called the electric potential. Conventionally, the signs are chosen so that

$$\mathbf{E}=-\nabla\phi. \tag{48}$$

If an infinitesimal point charge δq is moved from infinity to a point at $\mathbf{r}$ through a preexisting electric field $\mathbf{E}$ whose potential vanishes at infinity, the mechanical work done on the charge (the negative of the work done by the field) is $\delta W=\delta q\phi(r)$. This work represents the electrostatic energy of δq. Suppose now that the preexisting field is produced by a charge density $\rho(\mathbf{r})$ and that, instead of an infinitesimal point charge δq being brought in from infinity, there is an infinitesimal change $\delta\rho$ in the charge density throughout space (think of all the little bits of charge making up $\delta\rho$ being brought into their places from infinity). The work done to accomplish the change is

$$\delta W=\int \delta\rho(\mathbf{r})\phi(\mathbf{r})d^3x, \tag{49}$$

where the integration is over all regions where $\delta\rho\neq 0$. Without any assumptions about linearity, uniformity, isotropy, etc., the infinitesimal change in the charge density $\delta\rho$ can be related to a change in the electric displacement $\delta\mathbf{D}$ through the first Maxwell equation in Eq. (21):

$$\delta\rho=\nabla\cdot(\delta\mathbf{D}).$$

An integration by parts then leads to

$$\delta W = \int \mathbf{E}\cdot\delta\mathbf{D}\, d^3x,$$

where it has been assumed that $\rho(\mathbf{r})$ is a localized charge distribution. The total electrostatic energy (total work done in assembling the charge density) can be written formally as

$$W = \int d^3x \int_0^D \mathbf{E}\cdot\delta\mathbf{D}[\mathbf{E}]. \tag{50}$$

The symbolic integration over $\delta\mathbf{D}$ from zero to a final value contains all the information in the constitutive relations. For linear media, $2\mathbf{E}\cdot\delta\mathbf{D}=\delta(\mathbf{E}\cdot\mathbf{D})$. The electrostatic energy for linear media is thus

$$W = \frac{1}{2}\int \mathbf{E}\cdot\mathbf{D}\, d^3x, \tag{51}$$

where the integration is over all space. The integrand in Eq. (51) may be identified as an electrostatic energy density distributed in space. Use of Eq. (48) and the first equation in (21) will transform Eq. (51) into

$$W = \frac{1}{2}\int_V \rho(\mathbf{r})\phi(\mathbf{r})d^3x - \int_S \phi(\mathbf{r})\mathbf{D}\cdot\mathbf{n}\, da. \tag{52}$$

Here a formal device of finite volume V and surrounding surface S has been used to show explicitly the integration by parts. If the charge density is localized, the volume V can be expanded to infinity. The potential decreases inversely or faster with distance and the field $\mathbf{D}$ falls off with one higher power. The surface integral therefore manifestly vanishes as S goes to infinity. [It must vanish as soon as S is large enough to contain all of $\rho(\mathbf{r})$, but that is not immediately obvious.]

The last expression for the energy is what occurs for the microscopic Maxwell equations, where all the charges, both bound and free, are included in the computation of the work. In macroscopic electromagnetism, it is a valid result only for linear media. In general, Eq. (50) must be used.

4.2 Capacitance and Coefficients of Capacitance

A set of n conductors is embedded in a linear (but not necessarily isotropic or uniform) medium, with each conductor having a total charge Q_j and a corresponding potential V_j $(j=1,2,\ldots,n)$. The electrostatic energy of the system can be expressed in terms of the potentials V_j alone and certain geometrical quantities called coefficients of capacitance. The simplest example is an isolated metal sphere of radius R with a total charge Q in an isotropic, uniform medium with dielectric constant ϵ. The electrostatic potential at the surface of the sphere is $V=Q/4\pi\epsilon R$ and the energy is $W=QV/2$. Elimination of Q gives $W=CV^2/2$, where the coefficient $C=4\pi\epsilon R$ is called the capacitance of a sphere. This example illustrates the geometrical aspect of capacitance. Apart from the conventional factor $4\pi\epsilon$, the dimensions of capacitance are those of length. For the isolated sphere, capacitance is defined as the ratio of charge to potential, $C=Q/V$. A closely parallel definition applies for each set of n conductors, as is immediately apparent.

For a given configuration of conductors in a linear medium, the linear connection of the fields and potential with the charge density implies that the potential of the jth conductor is related linearly to the charges on all the conductors, that is,

$$V_j = \sum_{k=1}^{n} p_{jk}Q_k,$$

where the coefficients p_{jk} depend on the geometry of the conductors and the medium (if nonisotropic and/or nonuniform). These n equations can be inverted to yield the charge on the jth conductor in terms of all the potentials, with coefficients that are purely geometrical and independent of field strengths,

$$Q_j = \sum_{k=1}^{n} C_{jk}V_k, \quad j=1,2,\ldots,n. \tag{53}$$

The diagonal coefficients C_{jj} are called *capacitances*, while the C_{jk} $(j\neq k)$ are called coefficients of capacitance. From Eq. (53) it is evident that the capacitance of a conductor is the total charge on the conductor when it is maintained at unit potential, all other conductors being held at zero potential.

From Eqs. (52) and (53) it is seen that the electrostatic energy of the system of n conductors is

$$W = \frac{1}{2}\sum_{j=1}^{n} Q_jV_j = \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} V_jC_{jk}V_k. \tag{54}$$

This expression for the energy, equal to Eq. (51), can be used to find capacitances of specific configurations, either exactly or approximately (by variational methods, for example). The coefficients of capacitance are symmetric, i.e., $C_{jk}=C_{kj}$, as can be proven by means of Green's reciprocation theorem (Sec. 4.5).

Some common example of capacitances (with the surrounding medium isotropic, linear, and uniform) are

1. Isolated conducting sphere of radius R:

   $C=4\pi\epsilon R.$

2. Isolated thin flat circular conducting disk of radius R:

   $C=8\epsilon R.$

3. Two equal flat parallel conducting sheets of area A located one above the other with spacing d (with the smallest dimension across any part of the sheets large compared to d), as shown in Fig. 4, upper left:

   $$C=A\epsilon/d \quad +\text{fringe-field corrections } O(\epsilon d).$$

   To the neglect of the fringing field, the same capacitance applies for a flat sheet of area A positioned parallel to and a distance d away from any larger parallel flat sheet of conductor that extends beyond the projection of the smaller one.
4. Two coaxial cylindrical surfaces, the inner one of radius a, the outer of radius $b>a$, and length l ($l\gg b$), as shown in Fig. 4, upper right, with the potential difference between them:

   $$C=\frac{2\pi\epsilon l}{\ln(b/a)} \quad +\text{end-field corrections } O(\epsilon(b-a)).$$

5. Two long straight parallel right circular cylinder wires of radii a_1 and a_2 and common length l, whose central axes are separated by a distance d as shown in Fig. 4, lower left:

   $$C=2\pi\epsilon l\left[\cosh^{-1}\left(\frac{d^2-a_1^2-a_2^2}{2a_1a_2}\right)\right]^{-1} \approx\frac{\pi\epsilon l}{\ln(d/\sqrt{a_1a_2})}.$$

   The approximate result applies when $d\gg a_1,a_2$. End-field corrections are of the relative order d/l or less.

For conducting surfaces of not too severe shapes, capacitances can be estimated by comparison with known results. It can be proved that the capacitance of a closed surface S is bounded between those of the surfaces $S_<$ and $S_>$, where $S_<$ ($S_>$) lies totally inside (outside) S, as sketched in Fig. 4, lower right. For example, the capacitance of a conducting oblate spheroid with axes $a=b=R$ and $c<R$ has a capacitance between those of a sphere and a disk of radii R. The capacitance of a cube with sides of length a lies between those of the inscribed and circumscribed spheres, $C_<=0.5(4\pi\epsilon a)$ and $C_>=(\sqrt{3}/2)(4\pi\epsilon a)$. The arithmetic average of these bounds, $0.683(4\pi\epsilon a)$, is close to the numerically determined result, $0.655(4\pi\epsilon a)$.

Other examples of capacitance can be found in Sec. 5.b of Gray, *American Institute of Physics Handbook* (1972) and in ELECTROSTATIC CAPACITIVE STORAGE OF ENERGY.

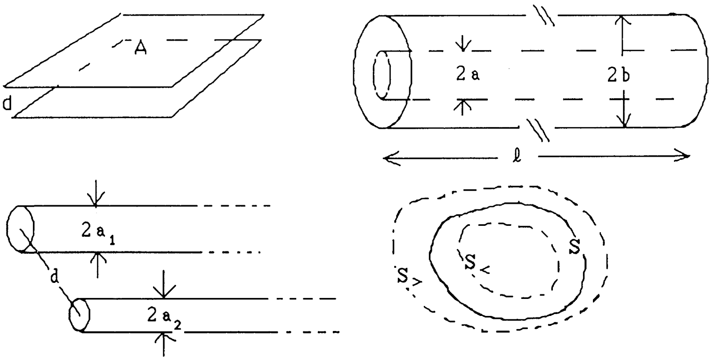

FIG. 4. Some capacitance geometries: parallel plates (upper left), coaxial cylinders (upper right), parallel circular wires (lower left), and closed conducting surface S, with inscribed ($S_<$) and circumscribed ($S_>$) surfaces (lower right).

4.3 Magnetostatic Energy

In deriving the electrostatic energy it is not necessary to consider time variation of the fields explicitly. Charges can be assumed to be moved arbitrarily slowly from infinity to their final locations. For magnetic fields, however, the variation in time of the fields as they are built up to their final magnetostatic configuration (defined by $\nabla\cdot\mathbf{J}=0$, or equivalently, $\partial\rho/\partial t=0$) produces electromotive forces (Faraday's law) that do work on the charges and currents. A treatment paralleling that of Sec. 4.1 [see Jackson (1975), Sec. 6.2] leads to a result for the infinitesimal change in magnetic energy caused by an infinitesimal change in the magnetic flux density of the form

$$\delta W=\int \mathbf{H}\cdot\delta\mathbf{B}\, d^3x. \tag{55}$$

This expression is the magnetic equivalent to the equation above Eq. (50). The constitutive relation between $\mathbf{B}$ and $\mathbf{H}$ is necessary to complete the determination of the total magnetostatic energy. For linear media, the energy is

$$W=\frac{1}{2}\int \mathbf{H}\cdot\mathbf{B}\, d^3x, \tag{56}$$

where the integration is over all space. In analogy with electrostatics, the integrand can be thought of as a magnetic energy density.

An expression for the magnetic energy analogous to Eq. (52) can be written by introduction of the vector potential $\mathbf{A}$, an auxiliary vector field that takes into account the fourth Maxwell equation, $\nabla\cdot\mathbf{B}=0$. Any divergenceless vector field can be written as the curl of another vector field. The magnetic flux density can therefore be defined as

$$\mathbf{B}=\nabla\times\mathbf{A}. \tag{57}$$

Substitution of this definition into Eq. (56), followed by an integration by parts and use of the Ampère–Maxwell equation, leads to the magnetic energy of a linear medium as

$$W=\frac{1}{2}\int \mathbf{J}\cdot\mathbf{A}\, d^3x. \tag{58}$$

The sources $\mathbf{J}$ are assumed to be localized in space; otherwise, there would be an additional surface integral at infinity. [For a localized steady current source, the vector potential decreases asymptotically inversely with distance and the fields decrease as the cube of the distance. The integrand in Eq. (56) therefore vanishes rapidly enough to make (56) equal (58) without a surface term.]

4.4 Inductance

A common situation in magnetostatics is a set of n currents I_j flowing in conductors C_j embedded in a nonconducting, magnetically linear (although not necessarily isotropic or uniform) medium. The conductors, which are also taken to be magnetically linear, are not assumed to be infinitesimally thin wires, although this may be a permissible approximation in some computations. The magnetic energy of the configuration can be expressed as a quadratic form in the currents, with geometrical coefficients of inductance, just as with the electrostatic energy and the coefficient of capacitance. The starting point is Eq. (58), which can be written as

$$W=\frac{1}{2}\sum_{j=1}^{n}\int_{V_j}\mathbf{J}_j\cdot\mathbf{A}_j\, d^3x + \sum_{j=1}^{n}\sum_{k>j}^{n}\int_{V_j}\mathbf{J}_j\cdot\mathbf{A}_k\, d^3x, \tag{59}$$

where $\mathbf{J}_j$ is the current density in the jth conductor (proportional to the total current I_j) and $\mathbf{A}_j$ is the vector potential resulting from that current density. The integrations are over the volumes V_j of the set of conductors. With a linear medium, the magnetic fields $\mathbf{B}_j$ and $\mathbf{H}_j$ and also the vector potential $\mathbf{A}_j$ throughout space are linear in the current I_j. Equation (59) may therefore be expressed in the form

$$W=\frac{1}{2}\sum_{j=1}^{n} L_j I_j^2+\sum_{j=1}^{n}\sum_{k>j}^{n} M_{jk} I_j I_k, \tag{60}$$

where the coefficients L_j and M_{jk} are purely geometrical quantities, dependent upon the configuration of conductors and intervening medium and how the current flows through the conductors, but not dependent on the magnitudes of the currents I_j. The quantities L_j (M_{jk}) are called coefficients of self- (mutual) inductance. Like the coefficients of capacitance, the mutual inductances are symmetric, i.e., $M_{jk}=M_{kj}$.

The approximation of infinitesimally thin wires can sometimes be used in the calculation of mutual inductances (but not for self-inductances, where an infinite value is the result). Consider such a conducting wire of circuit C_1 in the presence of another such circuit C_2. From Eq. (58) the mutual energy is

$$W_{12}=I_1 \oint_{C_1} \mathbf{A}_2\cdot d\mathbf{l}_1. \tag{61}$$

Stokes's theorem, together with Eq. (57) connecting **B** and **A**, can be used to transform W_{12} into

$$W_{12}=I_1 \int_{S_1} \mathbf{B}_2\cdot\mathbf{n}\; da_1.$$

The open surface S_1 spans the circuit C_1. The integral is the magnetic flux Φ_{12} (caused by the second circuit with its current I_2) through the first circuit C_1. The mutual inductance M_{12} is therefore $M_{12}=\mathcal{F}_{12}$, where $\mathcal{F}_{12}$ is the magnetic flux through circuit C_1 caused by unit current flowing in circuit C_2. For wire circuits in a uniform, isotropic medium of permeability μ, it can be shown, from Eq. (61) and the expression for the vector potential in terms of the current, that M_{12} can be written in this approximation in the purely geometric form

$$M_{12}=\frac{\mu}{4\pi} \oint_{C_1} \oint_{C_2} \frac{d\mathbf{l}_1\cdot d\mathbf{l}_2}{r_{12}}. \tag{62}$$

Clearly such an expression is meaningless when applied to self-inductance.

Equation (56) or (58) for the energy can be used together with Eq. (60) to calculate the self-inductances of specific configurations of currents. For mutual inductances, the same approach can be employed, although in some circumstances the linked flux approach or Eq. (62) can be substituted.

Some examples of self-inductances (circuits are in vacuum and at low frequency unless stated otherwise) are the following.

1. A straight right circular wire of radius a, length l ($l \gg a$), and relative permeability μ_r:

   $$L\approx\frac{\mu_0}{2\pi}\, l\left[\ln\left(\frac{2l}{a}\right)-1+\frac{\mu_r}{4}\right].$$

   The contribution $\mu_r/4$ comes from the interior of the wire. At very high frequencies, when the skin depth (39) $\delta \ll a$, this interior contribution is absent. In actuality, the argument of the logarithm is uncertain, or rather, it depends upon the actual way in which the wire of length l and radius a is connected into a complete circuit.
2. Two parallel long straight circular nonpermeable wires of length l with radii a_1 and a_2 and a separation d between their axes, with current flowing down one wire and back the other, as shown in Fig. 5, upper left:

   $$L=\frac{\mu_0 l}{\pi}\left[\ln\left(\frac{d}{\sqrt{a_1 a_2}}\right)+\frac{1}{4}\right].$$

   At very high frequencies the "interior" contribution, $\frac{1}{4}$, is absent.
3. Two coaxial conducting right circular cylinders of length l with current flow in opposite directions, the inner one solid with radius a and relative permeability μ_1, the outer one hollow, with inner radius b, thickness t ($t \ll b$), and relative permeability μ_3, and the space between the cylinders having relative permeability μ_2, as sketched in Fig. 5, upper right:

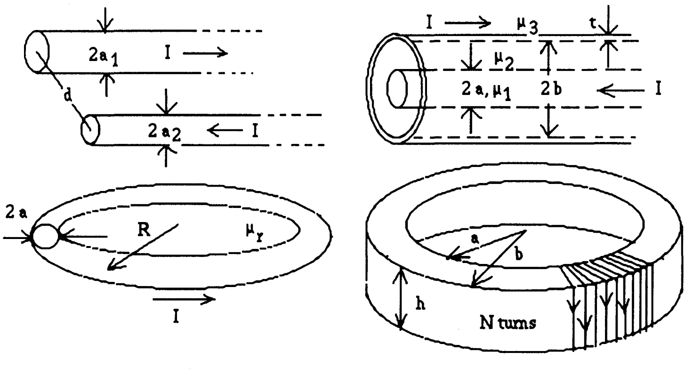

FIG. 5. Some inductance geometries: parallel circular wire transmission line (upper left), coaxial cylinder transmission line (upper right), circular loop of wire (lower left), circular toroid of rectangular cross section (lower right).

$$L\approx\frac{\mu_0}{2\pi}\,l\left[\frac{\mu_1}{4}+\mu_2\ln\left(\frac{b}{a}\right)+\frac{\mu_3 t}{3b}\right].$$

At very high frequencies the terms in μ_1 and μ_3 are absent, as they would be if the inner and outer cylinders were both very thin cylindrical shells.

4. A circular wire of radius a and relative permeability μ_r bent into a circle of mean radius R ($R\gg a$), as shown in Fig. 5, lower left:

$$L\approx\mu_0 R\left[\ln\left(\frac{8R}{a}\right)-2+\frac{\mu_r}{4}\right].$$

There are corrections of order a^2/R^2 and $(a^2/R^2)\ln(R/a)$, and the same comment as above applies about the absence of the interior contribution at high enough frequencies. For a coil of N turns wound closely together such that the bundle has an approximately circular cross section of radius a, the self-inductance is approximately N^2 times that of the single turn, provided the mean radius is used for R.

5. A toroid of rectangular cross section, with inner (outer) radius a (b) and height h, and permeability μ, wound uniformly with N turns, as sketched in Fig. 5, lower right:

$$L\approx\frac{\mu N^2 h}{2\pi}\ln\left(\frac{b}{a}\right).$$

It is assumed that the wire size is very small compared to the dimensions of the toroid and that the number of turns is large enough that there is negligible spacing between turns.

6. A toroid of circular cross section of radius a whose center is a radius R ($R>a$) from the axis of the toroid, with permeability μ, wound uniformly with N turns:

$$L\approx\mu N^2(R-\sqrt{R^2-a^2}).$$

The same assumptions about wire size, etc. apply here as in example **4**.

For examples **4–6**, approximate mutual inductances can be obtained for coils wound closely on top of each other by replacing N^2 by N_1N_2 and using mean dimensions. With the idealized toroids of examples **5** and **6**, the second coil of N_2 turns need not extend around the complete toroid because the magnetic flux is contained entirely within the toroid.

Further examples and more precise expressions for self- and mutual inductances can be found in Sec. 5.b of Gray, *American Institute of Physics Handbook* (1972), and in INDUCTORS.

4.5 Symmetry and Reciprocity Properties

There are useful symmetries and reciprocity relations based on Green's reciprocation theorem of electrostatics and its generalization to magnetostatics. Consider a volume V defined by a set of conducting boundary surfaces collectively labeled S (one of which may be at infinity). Let $\phi(\mathbf{r})$ be the electrostatic potential produced by a volume charge density ρ and a surface charge density σ on the surface(s) S, and let $\phi'(\mathbf{r})$ be the potential produced by another charge distribution, ρ' and σ'. Green's reciprocation theorem states that

$$\int_V \rho\phi'\,d^3x+\int_S \sigma\phi'\,da=\int_V \rho'\,\phi\,d^3x+\int_S \sigma'\phi\,da. \tag{63}$$

Physically, the two sides of Eq. (63) are just different expressions for the mutual electrostatic potential energy of one charge distribution in the presence of the electric potential caused by the other. The equality can be traced back to Newton's third law. For a set of conducting surfaces raised to potentials V_j (V_j') by the placement of charges Q_j (Q_j') on them, Eq. (63) becomes

$$\sum_j Q_j' V_j=\sum_k Q_k V_k'. \tag{64}$$

Suppose that in one situation $Q_1=q$, but all other $Q_j=0$, while in the other, $Q_2'=q$ and all other $Q_j'=0$. Equation (64) shows that the uncharged surface S_2 is raised by the presence of q on S_1 to the same potential as is the uncharged surface S_1 when the same charge q is placed on S_2. In another application, the symmetry of the coefficients of capacitance C_{ij} follows from Eq. (64) if Eq. (53) is used to express the charges in terms of the potentials on both sides of the equation.

The reciprocation theorem can be used to answer questions about induced charges. Consider, for example, a point P_1 a distance r from the center of a conducting sphere of radius a

on which a charge q has been placed. The potential at P_1 is $V_1 = q/4\pi\epsilon r$. If, instead, a point charge q is placed at P_1, the uncharged sphere will be raised to the same potential, $V_2' = V_1$. With the point charge still at P_1, suppose that the sphere is now grounded. Enough charge q' flows in from (out to) infinity to make the sphere's potential zero, i.e., $q'/4\pi\epsilon a + V_2' = q'/4\pi\epsilon a + q/4\pi\epsilon r = 0$, or $q' = -aq/r$, the well-known result for the charge induced on a grounded sphere by a nearby point charge.

In magnetostatics there is a corresponding, although less familiar, reciprocation theorem involving currents in a set of circuits and the resulting magnetic fluxes through them.

5. POTENTIALS, GAUGE TRANSFORMATIONS, AND WAVE EQUATIONS

5.1 Scalar and Vector Potentials and Gauge Transformations

In Sec. 4 the scalar potential ϕ and the vector potential $\mathbf{A}$ were introduced in the context of time-independent fields. For time-dependent phenomena they are equally useful. The definition, Eq. (57), of the magnetic flux density in terms of the vector potential is applicable to the more general situation because it stems from the fourth Maxwell equation, $\nabla \cdot \mathbf{B} = 0$, but the connection of the electric field to the potentials, Eq. (48), must be generalized. Recall that the curl of a gradient vanishes. In contrast, Faraday's law shows that

$$\nabla \times \left(\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} \right) = \mathbf{0}.$$

With time-dependent fields, it is not $\mathbf{E}$, but $\mathbf{E} + \partial \mathbf{A}/\partial t$ that can be expressed as a gradient of the scalar function. The definitions of the fields $\mathbf{E}$ and $\mathbf{B}$ in terms of the potentials ϕ and $\mathbf{A}$ are thus

$$\mathbf{E} = -\nabla \phi - \frac{\partial \mathbf{A}}{\partial t},$$

$$\mathbf{B} = \nabla \times \mathbf{A}. \tag{65}$$

With these definitions, the two homogeneous Maxwell equations are automatically satisfied.

The electric and magnetic fields are fundamental; the potentials are useful auxiliary fields, but they are not unique. The magnetic flux density is unchanged if the vector potential is modified by adding the gradient of some scalar function $\chi(\mathbf{r},t)$ to it. To keep the electric field unchanged, the scalar potential must also be modified. The simultaneous change of the two potentials,

$$\mathbf{A} \rightarrow \mathbf{A}' = \mathbf{A} + \nabla \chi,$$

$$\phi \rightarrow \phi' = \phi - \frac{\partial \chi}{\partial t} \tag{66}$$

is called a gauge transformation. It preserves the electric and magnetic fields, but can modify the equations satisfied by the potentials.

Consider the Maxwell equations in free space. If $\mathbf{E}$ and $\mathbf{B}$ are eliminated by means of Eq. (65) from the inhomogeneous pair of the Maxwell equations, they become equations for the potentials,

$$\nabla^2 \phi + \frac{\partial}{\partial t}(\nabla \cdot \mathbf{A}) = \frac{1}{\epsilon_0} \rho, \tag{67}$$

$$\nabla^2 \mathbf{A} - \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{A}}{\partial t^2} - \nabla \left(\nabla \cdot \mathbf{A} + \mu_0 \epsilon_0 \frac{\partial \phi}{\partial t} \right) = -\mu_0 \mathbf{J}. \tag{68}$$

The four coupled first-order partial-differential equations of Maxwell have been replaced by two coupled second-order partial-differential equations for the potentials. The freedom to make gauge transformations without affecting the fields means that Eqs. (67) and (68) can be decoupled or otherwise put into forms that are simpler or more convenient. For example, potentials can be chosen such that they satisfy the Lorentz condition,

$$\nabla \cdot \mathbf{A} + \mu_0 \epsilon_0 \frac{\partial \phi}{\partial t} = 0. \tag{69}$$

Then the potentials satisfy the standard second-order wave equation (with $\mu_0 \epsilon_0 = 1/c^2$)

$$\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = -\frac{1}{\epsilon_0} \rho, \tag{70}$$

$$\nabla^2 \mathbf{A} - \frac{1}{c^2} \frac{\partial^2 \mathbf{A}}{\partial t^2} = -\mu_0 \mathbf{J}. \tag{71}$$

Potentials that satisfy Eq. (69) are said to be in the class of Lorentz gauge potentials. If the original set did not satisfy Eq. (69), one can make a gauge transformation, Eq. (66), with χ a solution of

$$\nabla^2\chi-\frac{1}{c^2}\frac{\partial^2\chi}{\partial t^2}=-\left(\nabla\cdot\mathbf{A}+\frac{1}{c^2}\frac{\partial\phi}{\partial t}\right). \tag{72}$$

The new potentials will satisfy the Lorentz condition. Since the solution of Eq. (72) is not unique (any solution of the homogeneous wave equation can be added), it is clear that there is an infinite set of potentials in the Lorentz gauge class. Potentials in the Lorentz gauge are convenient because the equations for ϕ and $\mathbf{A}$ are simple decoupled wave equations with source terms given by the charge density ρ and the current density $\mathbf{J}$, respectively.

Another useful choice of gauge is the gauge in which $\nabla\cdot\mathbf{A}=0$, called the Coulomb, transverse, or radiation gauge. Equation (73) for the scalar potential becomes Poisson's equation, with the solution

$$\phi(\mathbf{r},t)=\frac{1}{4\pi\epsilon_0}\int\frac{\rho(\mathbf{r}',t)}{|\mathbf{r}-\mathbf{r}'|}\,d^3x'. \tag{73}$$

The scalar potential is just the *instantaneous* Coulomb potential caused by the charge density $\rho(\mathbf{r},t)$. The vector potential satisfies an inhomogeneous wave equation,

$$\nabla^2\mathbf{A}-\frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2}=-\mu_0\mathbf{J}_t,$$

where

$$\mathbf{J}_t=\mathbf{J}-\epsilon_0\nabla\frac{\partial\phi}{\partial t} \tag{74}$$

is the transverse current density, so named because it is the part of the current that has zero divergence, $\nabla\cdot J_t=0$. The name "transverse gauge" comes from this circumstance.

The Coulomb gauge is often used when no sources are present. The scalar potential can be taken to be zero, while $\mathbf{A}$ satisfies the homogeneous wave equation and $\nabla\cdot\mathbf{A}=0$. Then the fields are given solely in terms of $\mathbf{A}$ by Eq. (65) with $\phi=0$. These fields propagate with the speed of light in vacuum, c. A plane wave, harmonic in time with frequency ω and wave vector $\mathbf{k}$, can be represented by a vector potential, $\mathbf{A}=A_0\boldsymbol{\epsilon}\exp(i\mathbf{k}\cdot\mathbf{r}-i\omega t)$, where $|\mathbf{k}|=\omega/c$ and $\mathbf{k}\cdot\boldsymbol{\epsilon}=0$. From Eq. (65), the fields are $\mathbf{E}=i\omega\mathbf{A}$ and $\mathbf{B}=i\mathbf{k}\times\mathbf{A}$. The transversality of the fields ($\mathbf{k}\cdot\mathbf{E}=\mathbf{k}\cdot\mathbf{B}=0$) is manifest, as well as the orthogonality of $\mathbf{E}$ and $\mathbf{B}$. If $\boldsymbol{\epsilon}$ is a real vector, the plane wave is said to be plane polarized, with the directions of $\mathbf{E}$ and $\mathbf{B}$ fixed in space. If $\boldsymbol{\epsilon}$ is complex, e.g., $\boldsymbol{\epsilon}=(\mathbf{e}_1+i\mathbf{e}_2)/\sqrt{2}$, where $\mathbf{e}_1$ and $\mathbf{e}_2$ are real unit vectors, perpendicular to each other and to $\mathbf{k}$, then the wave is elliptically polarized. (The example is a right circularly polarized or positive helicity wave. See OPTICS, LINEAR.)

Since the electric field strength $\mathbf{E}$ has the dimensions of electric potential per unit length (units of V/m) while the magnetic field strength $\mathbf{H}$ has dimensions of current per unit length (units of A/m), their ratio has the dimensions of impedance (V/A or Ω). For a plane wave in free space this ratio is $Z_0=\mu_0\omega/k=\sqrt{\mu_0/\epsilon_0}$, called the impedance of free space, and is equal to 376.730 Ω. Analogous ratios of electric and magnetic fields in transmission lines and wave guides permit the use of lumped-circuit concepts in systems large compared to the free-space wavelength $\lambda=2\pi c/\omega$. (See ELECTROMAGNETIC WAVE PROPAGATION.)

5.2 Green's Functions for the Wave Equation

The first step in solving the wave equation, Eq. (70) or (71) or (74), is to find a Green's function, that is, a solution of the scalar wave equation with a unit impulse source in space and time. If the source point is at $(\mathbf{r}',t')$, the Green's function $G(\mathbf{r},t;\mathbf{r}',t')$ satisfies

$$\nabla^2G(\mathbf{r},t;\mathbf{r},'t')-\frac{1}{c^2}\frac{\partial^2G(\mathbf{r},t;\mathbf{r}',t')}{\partial t^2}$$
$$=-\delta(\mathbf{r}-\mathbf{r}')\delta(t-t'). \tag{75}$$

If there are no boundary surfaces at finite distances in space and time, the Green's function can only be a function of the relative coordinates, $\mathbf{R}=\mathbf{r}-\mathbf{r}'$ and $\tau=t-t'$. Furthermore, the source is spherically symmetric. The Green's function can be a function only of $R=|\mathbf{r}-\mathbf{r}'|$ and τ. Now introduce a frequency-domain Green's function through the Fourier transforms

$$\delta(\tau)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-i\omega\tau}\,d\omega \tag{76}$$

and

$$G(R,\tau)=\frac{1}{2\pi}\int_{-\infty}^{\infty} G(R,\omega)e^{-i\omega\tau}\,d\omega. \qquad (77)$$

The equation obeyed by $G(R,\omega)$ is the inhomogeneous Helmholtz wave equation,

$$\nabla_{\mathbf{R}}^2 G(R,\omega)+k^2G(R,\omega)=-\delta(\mathbf{R}), \qquad (78)$$

where $k^2=\omega^2/c^2$ is the square of the wave number associated with the frequency ω. Because of the spherical symmetry, only the radial part of the Laplacian gives a nonvanishing contribution. For all points with $R\neq 0$, the right-hand side is zero and the equation reads

$$\frac{d^2}{dR^2}RG(R,\omega)+k^2RG(R,\omega)=0$$

with the solution

$$RG(R,\omega)=Ae^{ikR}+Be^{-ikR}.$$

Furthermore, the δ function in Eq. (78) influences the solution only at the origin, where the second term on the left-hand side is negligible compared with the first. In these circumstances, Eq. (78) becomes Laplace's equation of electrostatics; G must reduce to the potential for a point charge and therefore

$$\lim_{kR\to 0} RG(R,\omega)=1/4\pi.$$

The coefficients A and B above must satisfy $4\pi(A+B)=1$.

It is useful to define the so-called retarded and advanced Green's functions, $G_{\pm}(R,\omega)$, to be

$$G_{\pm}(R,\omega)=\frac{1}{4\pi}\frac{e^{\pm ikR}}{R}. \qquad (79)$$

Given the convention of $e^{-i\omega t}$ for time dependence, G_+ (G_-) represents a spherically diverging (converging) wave propagating outwards from (inwards towards) the origin. To understand the designations, retarded and advanced, it is only necessary to insert $G_{\pm}(R,\omega)$ into Eq. (77) to determine the time-domain Green's function,

$$G_{\pm}(R,\tau)=\frac{1}{4\pi R}\,\delta\!\left(\tau\mp\frac{R}{c}\right)$$
$$=\frac{1}{4\pi|\mathbf{r}-\mathbf{r}'|}\,\delta\!\left(t'-\left[t\mp\frac{|\mathbf{r}-\mathbf{r}'|}{c}\right]\right). \qquad (80)$$

Evidently the retarded Green's function, G_+, represents an effect at $(\mathbf{r},t)$ caused by the action of an impulsive source a distance R away acting at an earlier time $t'=t-R/c$. The time difference R/c is just the time of propagation of the disturbance from one point to the other. Similarly, the advanced Green's function G_- represents sources acting at *later* times.

Because causality is observed in natural phenomena, it is customary to express the action of sources only in terms of the retarded Green's function, but that is not necessarily appropriate. Let $\Psi(\mathbf{r},t)$ and $f(\mathbf{r},t)$ represent ϕ (or a component of $\mathbf{A}$) and ρ/ϵ_0 (or a component of $\mu_0\mathbf{J}$), respectively, in the wave equations, Eqs. (70) and (71). It is assumed that $f(\mathbf{r},t)$ is localized in space and time. At remote times, $t\to\pm\infty$, the wave amplitude Ψ is a solution of the homogeneous wave equation. If Ψ is specified to be equal to a given initial wave, $\Psi_{\text{in}}(\mathbf{r},t)$, at $t\to-\infty$, the solution at all times is given in terms of the retarded Green's function,

$$\Psi(\mathbf{r},t)=\Psi_{\text{in}}(\mathbf{r},t)$$
$$+\int\int G_+(\mathbf{r},t;\mathbf{r}',t')f(\mathbf{r}',t')d^3x'\,dt. \qquad (81)$$

With the sources localized in time, it is clear that at early enough times the integral will be zero; the wave will be given by its initial free-wave form. At late times the sources will have acted; Ψ will consist of the initial wave Ψ_{in} plus waves causally radiated outwards from the sources.

If, on the other hand, the wave is specified to be $\Psi_{\text{out}}(\mathbf{r},t)$ for $t\to+\infty$, the solution is in terms of the advanced Green's function, that is, Eq. (81) with "in" replaced by "out" and G_+ replaced by G_-. The radiation field Ψ_{rad} can be defined formally as the difference, $\Psi_{\text{out}}-\Psi_{\text{in}}$:

$$\Psi_{\text{rad}}(\mathbf{r},t)=\int\int [G_+(\mathbf{r},t;\mathbf{r}',t')$$
$$-G_-(\mathbf{r},t;\mathbf{r}',t')]f(\mathbf{r}',t')d^3x'\,dt. \qquad (82)$$

For the usual situation in which the radiation fields are considered only after the sources have acted, G_+ alone is effective; the familiar retarded solution for the radiation fields emerges.

5.3 Harmonic Fields and Potentials in Uniform, Linear, Isotropic, Dispersive Media

The dispersive nature of dielectric constants and the resulting nonlocality in time for the relation between **D** and **E** has been discussed in Secs. 3.4 and 3.5. It is therefore only approximate to generalize the free-space wave equations such as Eqs. (70) and (71) merely by the replacement of μ_0 and ϵ_0 by μ and ϵ. Only if $\epsilon(\omega)$ and $\mu(\omega)$ are sensibly constant over the range of frequencies of interest can one make such a simple substitution. Instead, one must consider one frequency component of the fields at a time. Assume that all the fields and sources oscillate in time according to $e^{-i\omega t}$ and write $\mathbf{E}(\mathbf{r},t)=\mathbf{E}(\mathbf{r})\,e^{-i\omega t}$, etc., where $\mathbf{E}(\mathbf{r})$ is a complex vector field, and similarly for the other fields and sources. With the generally dispersive connections, $\mathbf{D}=\epsilon(\omega)\mathbf{E}$, $\mathbf{B}=\mu(\omega)\mathbf{H}$, appropriate to a linear, isotropic, uniform medium, the Maxwell equations (21) become

$$\nabla\cdot\mathbf{E}=\rho/\epsilon,$$

$$\nabla\times\mathbf{H}+i\epsilon\omega\mathbf{E}=\mathbf{J},$$

$$\nabla\times\mathbf{E}-i\omega\mu\mathbf{H}=\mathbf{0},$$

$$\nabla\cdot\mathbf{H}=0. \tag{83}$$

The continuity equation for charge and current densities now reads $\nabla\cdot\mathbf{J}-i\omega\rho=0$. If the potentials are similarly decomposed into spatial factors times $e^{-i\omega t}$, the fields are related to them via $\mathbf{E}=-\nabla\phi+i\omega\mathbf{A}$ and $\mu\mathbf{H}=\nabla\times\mathbf{A}$. The Lorentz condition, Eq. (70), is generalized for harmonic time dependence, but dispersive media, to $\nabla\cdot\mathbf{A}=i\mu\epsilon\omega\phi$. In the Lorentz class of gauges, the scalar potential is not an independent quantity. Once **A** is known, ϕ can be computed, but it is not needed—fields can be expressed in terms of **A** alone. The equation satisfied by the vector potential is an inhomogeneous Helmholtz wave equation,

$$\nabla^2\mathbf{A}+\mu\epsilon\omega^2\mathbf{A}=-\mu\mathbf{J}. \tag{84}$$

The solution of Eq. (84) can be written down in terms of the frequency-domain retarded Green's function $G_+(R,\omega)$, Eq. (79):

$$\mathbf{A}(\mathbf{r})=\mathbf{A}_{\text{in}}(\mathbf{r})+\mu\int\frac{e^{ikR}}{4\pi R}\,\mathbf{J}(\mathbf{r}')d^3x', \tag{85}$$

where $k=\omega\sqrt{\mu\epsilon}=\omega\sqrt{\mu_r\epsilon_r}/c$ is the wave number at frequency ω in the dispersive medium. This formula can be used to compute the radiation from harmonically driven sources such as antennas. (See ELECTROMAGNETIC RADIATION.)

If there are no external sources of charge and current, but the medium is conducting, the Ohm's-law contribution ($\mathbf{J}=\mathbf{E}/\rho$) can be written explicitly or included in a complex dielectric constant $\epsilon(\omega)$, as described in Sec. 3.4. With the latter choice, the wave equation (84) becomes a homogeneous equation with a complex wave number.

Auxiliary potentials, called electric and magnetic Hertz vectors, $\mathbf{Z}_e$ and $\mathbf{Z}_m$, are sometimes introduced. For harmonic fields and with any Ohm's-law contribution included in $\epsilon(\omega)$, the electric Hertz vector $\mathbf{Z}_e$ is just proportional to **A**. The connections to ϕ and **A** are

$$\phi=-\nabla\cdot\mathbf{Z}_e,\quad \mathbf{A}=-i\omega\mu\epsilon\mathbf{Z}_e. \tag{86}$$

It may be seen that these definitions yield potentials in the Lorentz class of gauges. In the absence of external sources, $\mathbf{Z}_e$ satisfies the homogeneous wave equation and the fields are given by

$$\nabla^2\mathbf{Z}_e+\mu\epsilon\omega^2\mathbf{Z}_e=0,$$

$$\mathbf{E}=\nabla\times\nabla\times\mathbf{Z}_e,$$

$$\mathbf{H}=-i\omega\epsilon\nabla\times\mathbf{Z}_e. \tag{87}$$

Since the Maxwell equations (83) transform into themselves under the simultaneous substitutions $\mu\leftrightarrow-\epsilon$, $\mathbf{E}\leftrightarrow\mathbf{H}$, in the absence of sources, the same transformation can be made in Eq. (87), leading to the magnetic Hertz vector $\mathbf{Z}_m$, which obeys the same wave equation and has the fields given by

$$\mathbf{E}=i\omega\mu\nabla\times\mathbf{Z}_m,$$

$$\mathbf{H}=\nabla\times\nabla\times\mathbf{Z}_m. \tag{88}$$

Hertz originally introduced these "superpotentials" in a treatment of the macroscopic Maxwell equations that kept explicit the polarization and magnetization densities, **P** and **M**, treating them as given external sources. The two Hertz vectors then satisfied inhomogeneous Helmholtz wave equations with **P** and **M** on the right-hand sides, with solutions like

Eq. (85), but with **J** replaced by **P** or **M**. In all but a few circumstances, it is simpler to deal with the usual scalar and vector potentials.

The plane waves described at the end of Sec. 5.1 have their counterparts in uniform, isotropic, linear media. It is only necessary to replace μ_0 and ϵ_0 everywhere by $\mu(\omega)$ and $\epsilon(\omega)$, respectively. The wave number becomes $k=n(\omega)\omega/c$, where $n(\omega)=\sqrt{\mu_r\epsilon_r}$ is the index of refraction of optics.

6. CONSERVATION LAWS FOR ENERGY AND MOMENTUM

6.1 Poynting's Theorem of Conservation of Electromagnetic Energy

An important theorem of Poynting (1884) concerns the continuity equation for the flux of electromagnetic energy and the energy content in the fields. If an infinitesimal point charge δq is acted on by electric and magnetic fields, **E** and **B**, the rate at which the fields do work on the charge is $\delta W=\delta q\,\mathbf{v}\cdot\mathbf{E}$, where **v** is the velocity of the charge. (The magnetic field does no work because the magnetic force is perpendicular to the velocity.) If there is a continuous distribution of charge and current, the total rate of doing work by the fields in a finite volume V is

$$\int_V \mathbf{J}\cdot\mathbf{E}\,d^3x.$$

This power represents a conversion of electromagnetic energy into mechanical or thermal energy and must result in a decrease in the total electromagnetic energy within V. The necessary conservation law is the content of Poynting's theorem. To find the explicit statement, one substitutes for **J** from the second line in Eq. (21):

$$\int_V \mathbf{J}\cdot\mathbf{E}\,d^3x=\int_V \left(\mathbf{E}\cdot(\nabla\times\mathbf{H})-\mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t}\right)d^3x. \tag{89}$$

The vector identity

$$\nabla\cdot(\mathbf{E}\times\mathbf{H})=\mathbf{H}\cdot(\nabla\times\mathbf{E})-\mathbf{E}\cdot(\nabla\times\mathbf{H})$$

may be used, together with Faraday's law [third line in Eq. (21)], to transform the right-hand side of Eq. (89) and obtain

$$\int_V \mathbf{J}\cdot\mathbf{E}\,d^3x=-\int_V \left(\nabla\cdot(\mathbf{E}\times\mathbf{H})+\mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t}+\mathbf{H}\cdot\frac{\partial\mathbf{B}}{\partial t}\right)d^3x. \tag{90}$$

Three assumptions are needed to proceed further, namely, that the medium is linear, that it is nondispersive, and that the sum of Eqs. (51) and (56), derived in stationary situations, gives the total electromagnetic energy when the fields are time dependent. With these assumptions, the second and third terms on the right-hand side of Eq. (90) represent the time derivative of the total electromagnetic energy density u, where

$$u=\tfrac{1}{2}(\mathbf{E}\cdot\mathbf{D}+\mathbf{B}\cdot\mathbf{H})=\tfrac{1}{2}(\epsilon E^2+\mu B^2). \tag{91a}$$

If the medium is dispersive then **D** and **B** are nonlocal in time and the conversion of the last two terms in Eq. (90) is more complicated. As first shown by Brillouin in 1921 (Brillouin, 1960; Landau *et al.*, 1984, Sec. 80), if the fields are concentrated in frequency around ω, with slow modulation in time, the last two terms in Eq. (90) correspond to the steady time derivative of

$$\bar{u}=\frac{1}{2}\left(\frac{d(\omega\epsilon)}{d\omega}\overline{E^2}+\frac{d(\omega\mu)}{d\omega}\overline{B^2}\right), \tag{91b}$$

where bars denote time averages over the rapid oscillations. If the medium is nondispersive, Eq. (91a) is recovered.

The first term on the right-hand side of Eq. (90) can be converted to a surface integral by means of the divergence theorem. It evidently represents a flow of energy out through the surface S bounding the volume V. It is therefore natural to define an electromagnetic energy flux density, called the *Poynting vector*,

$$\mathbf{S}=\mathbf{E}\times\mathbf{H} \tag{92}$$

with dimensions of energy per unit area per unit time. Equation (90) can then be written as a differential or integral statement of conservation of electromagnetic energy,

$$\frac{\partial u}{\partial t}+\nabla\cdot\mathbf{S}+\mathbf{J}\cdot\mathbf{E}=0,$$

$$\int_V \frac{\partial u}{\partial t}\,d^3x+\int_S \mathbf{n}\cdot\mathbf{S}\,da+\int_V \mathbf{J}\cdot\mathbf{E}\,d^3x=0. \tag{93}$$

The first two terms correspond to the rate of change of electromagnetic energy, either from doing work on charges and currents or from outward flow, while the third is the energy gained by the charges and currents from the fields. Equation (93) is Poynting's theorem.

6.2 Poynting's Theorem for Harmonic Fields and Field Definitions of Impedance and Admittance

A useful form of Poynting's theorem is that of time-averaged quantities with harmonic fields of a single frequency. For such fields the Maxwell equations take the form

$$\nabla\cdot\mathbf{B}=0,$$

$$\nabla\cdot\mathbf{D}=\rho,$$

$$\nabla\times\mathbf{E}-i\omega\mathbf{B}=\mathbf{0},$$

$$\nabla\times\mathbf{H}+i\omega\mathbf{D}=\mathbf{J}, \tag{94}$$

where all the quantities are complex functions of **r** with time dependence, $e^{-i\omega t}$, understood. In keeping with the conventions of Sec. 1 concerning time averages, the time-average rate of doing work on the charge and current distribution is

$$\frac{1}{2}\mathrm{Re}\left(\int_V \mathbf{J}^*\cdot\mathbf{E}\,d^3x\right).$$

If the steps from Eqs. (89)–(93) are repeated with obvious changes, the result is the pair of complex equations,

$$2i\omega(w_e-w_m)+\nabla\cdot\mathbf{S}+\tfrac{1}{2}\,\mathbf{J}^*\cdot\mathbf{E}=0,$$

$$2i\omega\int_V (w_e-w_m)d^3x+\int_S \mathbf{n}\cdot\mathbf{S}\,da+\frac{1}{2}\int_V \mathbf{J}^*\cdot\mathbf{E}\,d^3x=0, \tag{95}$$

where the complex Poynting vector and harmonic electric and magnetic energy densities are

$$\mathbf{S}=\tfrac{1}{2}\,\mathbf{E}\times\mathbf{H}^*,\quad w_e=\tfrac{1}{4}\,\mathbf{E}\cdot\mathbf{D}^*,\quad w_m=\tfrac{1}{4}\,\mathbf{B}\cdot\mathbf{H}^*. \tag{96}$$

Equations (95) and (96) are the harmonic analog of Eq. (93). The real part of Eq. (95) is the statement of conservation of energy for the time-averaged quantities. Its imaginary part relates the reactive or stored energy to its alternating flow.

The complex form of Poynting's theorem, Eq. (95), can be exploited to define the input (junction) impedance of a general, two-terminal, linear, passive, electromagnetic system. Imagine the isolated system within a volume V with boundary surface S, as shown in Fig. 6. Only the two input–output terminals are protruding. Let the complex harmonic current and voltage at the terminals be I_i and V_i. Then the complex power input is $I_i^* V_i/2$. This power is assumed to flow *into* the volume V through the surface S_i, e.g., the cross section of the coaxial line shown in the bottom half of Fig. 6. In terms of the Poynting vector, one has

$$\frac{1}{2}\,I_i^* V_i=-\oint_{S_i} \mathbf{S}\cdot\mathbf{n}\,da,$$

where **n** is the outwardly directed normal to S, as indicated in Fig. 6. Now applying Eq. (95) to the volume V and total surface S, one obtains

$$2i\omega\int_V (w_e-w_m)d^3x+\int_{S-S_i} \mathbf{n}\cdot\mathbf{S}\,da+\frac{1}{2}\int_V \mathbf{J}^*\cdot\mathbf{E}\,d^3x=\frac{1}{2}\,I_i^* V_i. \tag{97}$$

The surface integral on the left represents the energy flow out of the volume V. If the surface $S-S_i$ is taken to infinity, this integral is real and describes escaping radiation. At low frequencies radiation is generally negligible; no distinction need be made between S and S_i; the upper diagram in Fig. 6 is adequate.

The input impedance $Z=R-iX$ (engineers, please recall that the conventions of Sec. 1 imply $j=-i$) of the system follows from Eq. (97) with its definition, $V_i=ZI_i$. The right-hand side of Eq. (97) is $Z|I_i|^2/2$.

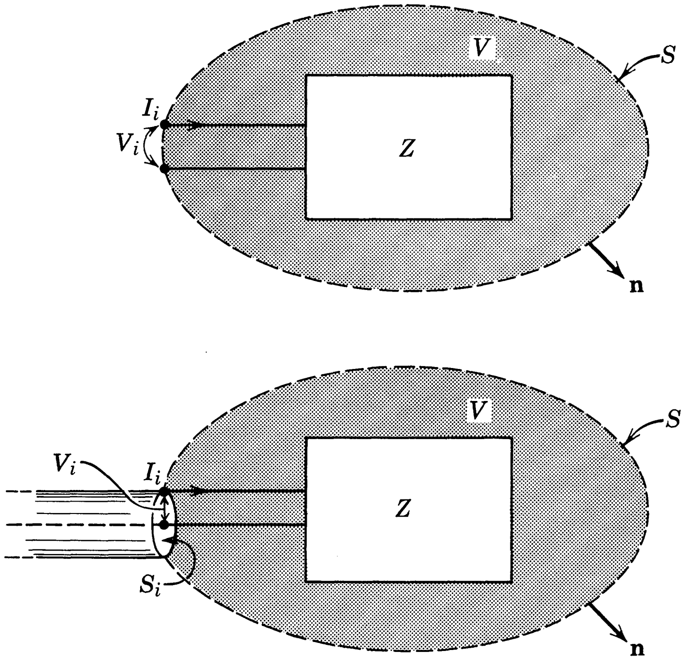

FIG. 6. Schematic diagrams of arbitrary, two-terminal, linear, passive electromagnetic systems. The surface S completely surrounds the system; only the input terminals protrude. At the terminals, the input current and voltage are I_i and V_i, with the input impedance defined through $V_i = ZI_i$. The upper diagram pertains to low frequencies, the lower, with its coaxial feed, to high frequencies where radiation resistance may enter. [From Jackson (1975), reproduced by permission of John Wiley & Sons, Inc.]

The real and imaginary parts of Z are

$$R = \frac{1}{|I_i|^2}\left[\mathrm{Re}\left(\int_V \mathbf{J}^*\cdot\mathbf{E}\,d^3x\right) + 4\omega\,\mathrm{Im}\left(\int_V (w_m - w_e)\,d^3x\right) + 2\int_{S-S_i} \mathbf{n}\cdot\mathbf{S}\,da\right],$$

$$X = \frac{1}{|I_i|^2}\left[4\omega\,\mathrm{Re}\left(\int_V (w_m - w_e)\,d^3x\right) - \mathrm{Im}\left(\int_V \mathbf{J}^*\cdot\mathbf{E}\,d^3x\right)\right]. \tag{98}$$

Here it has been assumed that the power flow out through $S - S_i$ is real. The last term in the expression for R is the *radiation resistance*, important at high frequencies. At low frequencies, the Ohmic losses are generally the only dissipation (**J** in phase with **E**); the electric and magnetic energies densities are real throughout V. Then the expressions simplify to

$$R \approx \frac{1}{|I_i|^2}\int_V \rho|\mathbf{J}|^2\,d^3x,$$

$$X \approx \frac{4\omega}{|I_i|^2}\int_V (w_m - w_e)\,d^3x. \tag{99}$$

Here ρ is the real (low-frequency) resistivity. The expression for the resistance R exhibits the Ohmic heat loss throughout the system. Similarly, the expression for the reactance X shows the expected behavior. If stored magnetic energy dominates, as for a lumped inductance, the reactance is positive, etc. The different dependences on frequency of the reactance for inductances ($X = \omega L$) and capacitances ($X = -1/\omega C$) can be traced to the definition of L in terms of current and voltage ($V = L\,dI/dt$) on the one hand, and of C in terms of charge and voltage ($V = Q/C$) on the other.

The admittance $Y = G - iB$, where G is the conductance and B the susceptance, can be defined via Eq. (97) using $I_i = YV_i$. The low-frequency forms analogous to Eq. (99) are

$$G \approx \frac{1}{|V_i|^2}\int_V \frac{1}{\rho}|\mathbf{E}|^2\,d^3x,$$

$$B \approx -\frac{4\omega}{|V_i|^2}\int_V (w_m - w_e)\,d^3x. \tag{100}$$

It should be noted that the situation depicted in Fig. 6 and the above definitions of impedance or admittance are not immediately applicable to circumstances in which energy input is distributed in space, as, for example, in the exchange of energy between an electron beam and the cavity structure in a magnetron or free-electron laser. Poynting's theorem is applicable and circuit concepts can be used, but care must be taken in the definitions. For example, the voltage V_i in the definition of admittance is replaced by an appropriate line integral of the electric field. For discussions in the microwave domain, see Chap. IX of Slater (1950) or Chap. 2 of Collins (1948).

6.3 Conservation of Momentum, Maxwell Stress Tensor, and Radiation Pressure

The conservation of momentum can be similarly addressed. A particle with infinitesimal charge δq in the presence of fields **E** and **B** experiences a force, called the Lorentz force,

$$\delta\mathbf{F}=\delta q(\mathbf{E}+\mathbf{v}\times\mathbf{B}).$$

If $\mathbf{P}_{\text{mech}}$ is the sum of the mechanical momenta of all the particles within a volume V, Newton's second law can be expressed as

$$\frac{d\mathbf{P}_{\text{mech}}}{dt}=\int_V(\rho\mathbf{E}+\mathbf{J}\times\mathbf{B})d^3x, \qquad (101)$$

where a discrete sum over charged particles is formally replaced by integration over the charge and current densities times the appropriate fields. For simplicity it is assumed that the charges described by ρ and **J** are in otherwise free space, i.e., $\mathbf{D}=\epsilon_0\mathbf{E}$ and $\mathbf{B}=\mu_0\mathbf{H}$, and that the volume V is such that there is no flow of particles into or out of it. If the first two of the Maxwell equations in Eq. (21) are used to eliminate ρ and **J** from Eq. (101), the integrand can be cast in the form

$$\begin{aligned}\rho\mathbf{E}+\mathbf{v}\times\mathbf{B}=&\epsilon_0[\mathbf{E}(\nabla\cdot\mathbf{E})-\mathbf{E}\times(\nabla\times\mathbf{E})]\\&+\mu_0[\mathbf{H}(\nabla\cdot\mathbf{H})-\mathbf{H}\times(\nabla\times\mathbf{H})]\\&-\mu_0\epsilon_0\frac{\partial}{\partial t}(\mathbf{E}\times\mathbf{H}).\end{aligned} \qquad (102)$$

In Cartesian components, the square brackets have the form

$$[\mathbf{E}(\nabla\cdot\mathbf{E})-\mathbf{E}\times(\nabla\times\mathbf{E})]_j=\sum_{k=1}^{3}\frac{\partial}{\partial x_k}(E_jE_k-\tfrac{1}{2}\delta_{jk}\mathbf{E}\cdot\mathbf{E}). \qquad (103)$$

and similarly for the second one. The last term in Eq. (102) has the form of the time derivative of a momentum density ($\mu_0\epsilon_0\,\mathbf{E}\times\mathbf{H}=\mathbf{S}/c^2$). The total electromagnetic momentum within V is thus defined as

$$\mathbf{P}_{\text{em}}=\mu_0\epsilon_0\int_V\mathbf{E}\times\mathbf{H}\,d^3x. \qquad (104)$$

Equation (101) can then be written in component form to express conservation of the sum of mechanical and electromagnetic momentum,

$$\frac{d}{dt}(\mathbf{P}_{\text{mech}}+\mathbf{P}_{\text{em}})_j=\sum_{k=1}^{3}\int_V\frac{\partial}{\partial x_k}T_{jk}\,d^3x, \qquad (105)$$

where the *Maxwell stress tensor* T_{jk} is

$$T_{jk}=\epsilon_0E_jE_k+\mu_0H_jH_k-\tfrac{1}{2}\delta_{jk}(\epsilon_0\mathbf{E}\cdot\mathbf{E}+\mu_0\mathbf{H}\cdot\mathbf{H}). \qquad (106)$$

Application of the divergence theorem to the volume integral in Eq. (105) gives Newton's second law for the combined system of particles and fields within the volume V,

$$\frac{d}{dt}(\mathbf{P}_{\text{mech}}+\mathbf{P}_{\text{em}})_j=-\int_S\sum_k T_{jk}n_k\,da, \qquad (107)$$

where **n** is the *inward*-directed normal to the closed surface S surrounding V. The integrand on the right-hand side of Eq. (107) is evidently the force per unit area transmitted across an element of the surface S and acting on the combined system inside V.

The force per unit area or pressure, $-\Sigma_kT_{jk}n_k$, can be used to calculate forces on material objects such as conductors by integration over an appropriate surface. A simple example of the ponderomotive nature of electromagnetic radiation is *radiation pressure.* A plane wave with wave vector **k** has both **E** and **H** perpendicular to **k**, in the plane of constant phase, with $|\mathbf{E}|=\sqrt{\mu_0/\epsilon_0}|\mathbf{H}|$. The force per unit area in the direction of **k** then comes only from the δ_{jk} terms in Eq. (106). The radiation pressure in the direction of propagation is therefore

$$P_{\text{rad}}=-\sum_{i,j}\hat{k}_iT_{ij}\hat{k}_j=\frac{1}{c}|\mathbf{S}|=u. \qquad (108)$$

Here u is the energy density and **S** the Poynting vector. The pressure (flux of momentum) has the relation to the Poynting vector (flux of energy) expected for relativistically moving (massless) particles (photons). If a plane wave in vacuum is incident and reflected at angle θ to the normal by a perfectly reflecting plane, the radiation pressure on the plane is easily found to be $P_{\text{rad}}=2\cos^2\theta|\mathbf{S}_{\text{inc}}|/c$, where $\mathbf{S}_{\text{inc}}$ is the Poynting vector of the incident wave.

Radiation pressure is the reason why most comets' tails point away from the Sun at perihelion—the dust particles that make up the tail are pushed away from the Sun by the radiation pressure of the solar radiation [a few microjoules per cubic meter or 10^{-11} atmospheres at the radius of the Earth and increasing as $(R_E/R)^2$ at smaller R].

For material media, care must be exercised in defining the purely electromagnetic energy density u, energy flow **S**, momentum flow, and stress tensor T_{jk} because what is considered electromagnetic and what is mechanical is to some extent arbitrary. After all, the mechanical properties of matter are governed by electromagnetic forces, even if the mass is not. It is generally agreed that in matter the energy flow is still given by the **S** of Eq. (92) and that the momentum density is $\mathbf{S}/c^2$. For detailed discussions, see deGroot (1969), and also Landau *et al.* (1984) or Robinson (1973).

7. SPECIAL RELATIVITY AND CLASSICAL FIELD THEORY

7.1 Summary of Results in Special Relativity

The special theory of relativity is based on two postulates:

1. There exists a multiply infinite set of equivalent Euclidean reference frames or coordinate systems moving with constant velocities in rectilinear paths relative to one another in which all physical phenomena occur in an identical matter. For brevity, these frames of reference are called inertial frames. This principle of relativity is very ancient, dating from Copernicus's time or earlier. It conforms with all known observations of ordinary mechanical systems, as well as the Michelson–Morley experiment (failure to detect motion through the putative aether).
2. There is a limiting speed of an object, regardless of the inertial frame from which it is observed. Experimentally, the speed is the speed of light *in vacuo.*

This second postulate, simplicity itself, has radical consequences, changing the concept of time as something absolute and putting it on an equal footing with the spatial coordinates. In high-energy physics, special relativity is verified routinely in very practical ways in the operation of accelerators, in the transport of beams of unstable particles over long distances, etc. See Sec. 11.2 of Jackson (1975) for a discussion of some of the experimental tests.

It is convenient to introduce the time coordinate $x_0=ct$ to give time the same dimensions as the spatial coordinates, $x_1=x$, $x_2=y$, $x_3=z$. Let the unprimed coordinates (x_0,x_1,x_2,x_3) be the coordinates of a point P in space-time with respect to axes fixed in an inertial frame K, and let the primed coordinates (x'_0,x'_1,x'_2,x'_3) be the coordinates of P with respect to axes fixed in the inertial frame K', moving with speed βc parallel to the x_1 axis of K. The axes in the two frames are defined to be parallel to each other and coincident at $x_0=x'_0=0$. The second postulate is equivalent to the invariance of the quantity $x_0^2-x_1^2-x_2^2-x_3^2$. The first postulate implies the isotropy and homogeneity of space-time. Together they lead to the *Lorentz transformation* among the coordinates in K and K',

$$x'_0=\gamma(x_0-\beta x_1),\quad x'_1=\gamma(x_1-\beta x_0),$$
$$x'_2=x_2,\quad x'_3=x_3. \tag{109a}$$

The Lorentz factor γ is

$$\gamma=1/\sqrt{1-\beta^2}. \tag{109b}$$

If the coordinate axes in K and K' remain parallel, but the velocity $\mathbf{v}=\boldsymbol{\beta}c$ of frame K' with respect to frame K is in an arbitrary direction, the Lorentz transformation (109a) has the generalization

$$x'_0=\gamma(x_0-\boldsymbol{\beta}\cdot\mathbf{x}),$$
$$\mathbf{x}'=\mathbf{x}+\frac{\gamma-1}{\beta^2}(\boldsymbol{\beta}\cdot\mathbf{x})\boldsymbol{\beta}-\gamma\boldsymbol{\beta}x_0. \tag{110}$$

Energy and momentum of a particle transform in the same manner as the time and space coordinates. If E is the particle's total energy (including its rest energy, mc^2) and $\mathbf{p}$ its momentum, then $(p_0=E/c, p_1, p_2, p_3)$ transform in the same way as (x_0, x_1, x_2, x_3). One calls the four quantities a four-vector and sometimes writes $x^\mu=(x_0; \mathbf{x})$. Similarly, $p^\mu=(p_0; \mathbf{p})$ is the energy–momentum four-vector. For electromagnetic radiation, $k_0=\omega/c$ and $\mathbf{k}$ form the frequency–wave vector four-vector. The four quantities G^μ $(\mu=0,1,2,3)$ form a four-vector if they transform under the above Lorentz transformation in the same way as the space and time coordinates, namely,

$$G_0'=\gamma(G_0-\beta G_1), \quad G_1'=\gamma(G_1-\beta G_0),$$

$$G_2'=G_2, \quad G_3'=G_3, \tag{111}$$

or the relations equivalent to Eq. (110) for a general velocity $\mathbf{v}$ of K' in K. The inverse transformation is obtained by interchanging primed and unprimed coordinates and changing the sign of β or $\boldsymbol{\beta}$.

Application of Eq. (111) to frequency and wave number yields the *relativistic Doppler formulas*,

$$\omega'=\gamma\omega(1-\beta\cos\theta),$$

$$\tan\theta'=\sin\theta/\gamma(\cos\theta-\beta), \tag{112}$$

where θ and θ' are the angles of $\mathbf{k}$ and $\mathbf{k}'$ relative to the direction of $\mathbf{v}$, the relative velocity of the inertial frame K' with respect to K. The inverse relations can be obtained by interchanging primed and unprimed quantities and reversing the sign of β. The more familiar nonrelativistic Doppler shift has $\gamma\approx 1$ in Eq. (112). The presence of $\gamma\geqslant 1$ shows that there is a frequency shift even for radiation emitted perpendicular to the velocity. This relativistic effect has been detected for rapidly rotating sources (but still $\beta\ll 1$) by means of the Mössbauer effect.

The relativistic law of transformation for velocity is not that of a four-vector. Let $\mathbf{u}'$ be a velocity in the frame K', and let the velocity of K' with respect to K be $\mathbf{v}$. Then the velocity $\mathbf{u}$ in K has components parallel $(u_\parallel)$ and perpendicular $(\mathbf{u}_\perp)$ to $\mathbf{v}$ given by

$$u_\parallel=\frac{u_\parallel'+v}{1+(\mathbf{v}\cdot\mathbf{u}')/c^2}, \quad \mathbf{u}_\perp=\frac{\mathbf{u}_\perp'}{\gamma_v[1+(\mathbf{v}\cdot\mathbf{u}')/c^2]}. \tag{113}$$

Here γ_v is the Lorentz factor associated with the velocity $\mathbf{v}$. The inverse relation is obtained by interchanging primed and unprimed quantities and changing the sign of $\mathbf{v}$. For parallel velocities, the first equation in Eq. (113) shows that, while the intuitive sum of velocities holds for nonrelativistic speeds, the factor in the denominator assures that the second postulate is not violated. In particular, if $u'=c(1-\epsilon)$ and $v=c(1-\epsilon_v)$, then $u\approx c(1-\epsilon\epsilon_v)$, correct to second order in small quantities.

While velocity does not transform in the same way as energy and momentum, a closely related four-vector, U^μ, the four-velocity, can be defined through the ratio of a particle's four-momentum to its mass, i.e., $U^\mu=(\gamma c; \gamma\mathbf{u})$. The four-vector law of transformation, Eq. (111) can be applied to U^μ and then expressions for the parallel and perpendicular components of $\mathbf{u}$ in terms of $\mathbf{u}'$ and $\mathbf{v}$ can be found; Eq. (113) results.

Because of the possibility of interpreting the invariant quantity, $s^2=x_0^2-x_1^2-x_2^2-x_3^2$, as the square of a "length" in a non-Euclidean space, it is useful to introduce the ideas of upper and lower indices, of a metric tensor, and the Einstein summation convention of implicit summation over any repeated index, one lower and one upper. Thus the familiar coordinates (ct,x,y,z) form a contravariant four-vector $x^\mu=(x^0=ct, x^1=x, x^2=y, x^3=z)$. The covariant four-vector $x_\mu=(x_0=ct, x_1=-x, x_2=-y, x^3=-z)$ is obtained from x^μ through multiplication by the metric tensor $g^{\mu\nu}$. Thus $x_\mu=g_{\mu\nu}x^\nu$, where $g^{\mu\nu}=g_{\mu\nu}$ and the only nonvanishing elements are

$$g^{00}=1, \quad g^{11}=g^{22}=g^{33}=-1. \tag{114}$$

The invariant "length" can then be written $s^2=x_\mu x^\mu$. The differential invariant length element in the space described by the metric tensor, Eq. (114) (called Minkowski space), is $(ds)^2=g_{\alpha\beta}\,dx^\alpha\,dx^\beta=dx_\beta\,dx^\beta=dx^\alpha\,dx_\alpha$. The scalar product of two four-vectors is defined as $A\cdot B=A_\mu B^\mu=A^0B^0-\mathbf{A}\cdot\mathbf{B}$, where boldface type means an ordinary vector in three-space. It should be noted that the Jacobian

$$J(x,x')=\frac{\partial(x^0,x^1,x^2,x^3)}{\partial(x'^0,x'^1,x'^2,x'^3)}$$

connecting the four-dimensional volume element $d^4x\equiv dx^0dx^1dx^2dx^3$ and $d^4x'\equiv dx'^0\,dx'^1dx'^2dx'^3$ is equal to unity. The four-volume element is therefore "invariant."

In Minkowski space, differential operators are defined appropriately as

$$\partial^\alpha \equiv \frac{\partial}{\partial x_\alpha} = \left(\frac{\partial}{\partial x^0}, -\nabla\right),$$

$$\partial_\alpha \equiv \frac{\partial}{\partial x^\alpha} = \left(\frac{\partial}{\partial x^0}, \nabla\right),$$

$$\Box \equiv \partial_\alpha \partial^\alpha = \frac{\partial^2}{\partial x^{02}} - \nabla^2. \qquad (115)$$

The four-divergence of a four-vector A^μ is the invariant quantity

$$\partial^\alpha A_\alpha = \partial_\alpha A^\alpha = \frac{\partial A^0}{\partial x^0} + \nabla \cdot \mathbf{A}, \qquad (116)$$

an equation familiar from the continuity of charge and current densities ρ and $\mathbf{J}$, and the Lorentz condition for the potentials ϕ and $\mathbf{A}$. (Here is the first suggestion of specific Lorentz transformation properties of electromagnetic quantities.) The second line in Eq. (115) shows that the wave-equation operator is a Lorentz-invariant four-dimensional Laplacian (denoted by $\Box$).

7.2 Relativistic Formulation of Electrodynamics

If $c\rho$ and $\mathbf{J}$ are interpreted as the time and space components of a four-vector density J^μ, the continuity equation for charge and current, Eq. (9), can be expressed in the invariant divergence form of Eq. (116):

$$\partial_\mu J^\mu = 0. \qquad (117)$$

The validity of such an identification can be inferred from the fact that electric charge is known experimentally to be a Lorentz-invariant quantity. [See Sec. 11.9 of Jackson (1975) for a brief discussion.] Since $\rho\, dx\, dy\, dz = \rho\, dx^1\, dx^2\, dx^3$ is such a quantity of charge and since the four-dimensional volume element d^4x is invariant, one can infer that the charge density ρ transforms as dx^0, that is, as the time component of a four-vector. The requirement that the continuity equation must be true in all inertial frames then forces $\mathbf{J}$ to be the space part of the same four-vector $J^\mu = (c\rho; \mathbf{J})$.

The Lorentz condition, Eq. (69), takes the invariant form of Eq. (116) if the potentials ϕ/c and $\mathbf{A}$ are identified as the components of a four-vector potential $A^\mu = (\phi/c; \mathbf{A})$. The wave equations, Eqs. (70) and (71), are seen to involve the invariant four-dimensional Laplacian operator, with components of the four-vector potential on one side and the four-current density on the other. They therefore provide a properly Lorentz-covariant description summarized in the wave equation

$$\Box A^\mu = \mu_0 J^\mu \qquad (118)$$

with the invariant subsidiary condition,

$$\partial_\nu A^\nu = 0. \qquad (119)$$

The definitions of the electric and magnetic fields in terms of the potentials [Eq. (65)] can be shown to imply that the field strengths $(E_x, E_y, E_z, B_x, B_y, B_z)$ are the elements of a second-rank, antisymmetric, field-strength tensor $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$. The explicit form as a matrix array is

$$F^{\mu\nu} = \begin{pmatrix} 0 & -E_x/c & -E_y/c & -E_z/c \\ E_x/c & 0 & -B_z & B_y \\ E_y/c & B_z & 0 & -B_x \\ E_z/c & -B_y & B_x & 0 \end{pmatrix}. \qquad (120)$$

The Maxwell equations can be written in covariant form as

$$\partial_\alpha F^{\alpha\beta} = \mu_0 J^\beta,$$

$$\partial^\alpha F^{\beta\delta} + \partial^\beta F^{\delta\alpha} + \partial^\delta F^{\alpha\beta} = 0. \qquad (121)$$

The first equation embodies Coulomb's law ($\beta=0$) and the three Ampère–Maxwell equations ($\beta = 1,2,3$); the second, in which α, β, and δ are any three of (0,1,2,3), the four source-free equations. Note that current conservation is implicit in the first equation in (121). If the four-divergence is taken of both sides, the antisymmetric nature of $F^{\alpha\beta}$ means that $\partial_\beta \partial_\alpha F^{\alpha\beta} = 0$ and Eq. (117) is recovered. The completely antisymmetric unit tensor $\epsilon_{\alpha\beta\gamma\delta} = -\epsilon^{\alpha\beta\gamma\delta}$ (defined to be zero if two or more indices are the same, and equal to ± 1 depending on the even or odd number of permutations of the indices needed to go from $\epsilon_{0123} = +1$) can be used to define the dual field-strength tensor, $\mathscr{F}^{\alpha\beta} = \frac{1}{2}\epsilon^{\alpha\beta\gamma\delta} F_{\gamma\delta}$, which has the general appearance of Eq. (120), but with the roles of $\mathbf{E}$ and $\mathbf{B}$ interchanged (and some signs different). The contraction of $F^{\alpha\beta}$ with itself and with its dual tensor yield two Lorentz-invariant combinations, quadratic in the field strengths:

$$F_{\alpha\beta}F^{\alpha\beta} = -2(\mu_0\epsilon_0\mathbf{E}\cdot\mathbf{E} - \mathbf{B}\cdot\mathbf{B}),$$

$$F^{\alpha\beta}\mathscr{F}_{\alpha\beta} = -4\sqrt{\mu_0\epsilon_0}\,\mathbf{E}\cdot\mathbf{B}. \qquad (122)$$

The Lorentz transformation properties of the field strengths as a second-rank tensor can be expressed in vectorial form as

$$\mathbf{E}' = \gamma(\mathbf{E} + \mathbf{v}\times\mathbf{B}) - \frac{\gamma^2}{\gamma+1}\,\boldsymbol{\beta}(\boldsymbol{\beta}\cdot\mathbf{E}),$$

$$\mathbf{B}' = \gamma(\mathbf{B} - \boldsymbol{\beta}\times\mathbf{E}/c) - \frac{\gamma^2}{\gamma+1}\,\boldsymbol{\beta}(\boldsymbol{\beta}\cdot\mathbf{B}). \qquad (123)$$

The fields $\mathbf{E}'$ and $\mathbf{B}'$ are those in the inertial frame K', which is moving with velocity $\mathbf{v} = \boldsymbol{\beta}c$ with respect to the frame K. The inverse transformation is obtained, as usual, by interchanging primed and unprimed quantities and changing the sign of $\mathbf{v}$ (and $\boldsymbol{\beta}$).

Consider a point charge q at rest in K', with electric field $\mathbf{E}' = q\mathbf{r}'/4\pi\epsilon_0 r'^3$ and $\mathbf{B}' = \mathbf{0}$ in that frame. In the frame K the charge is moving with velocity $\mathbf{v}$. The inverse of Eq. (123) shows that, to first order in β (i.e., the nonrelativistic approximation), the electric field in K is the same as before, but there is now a magnetic field,

$$\mathbf{B} \approx \frac{\mu_0}{4\pi}\,\frac{q\mathbf{v}\times\mathbf{r}}{r^3}, \qquad (124)$$

which is just the Ampère–Biot–Savart expression for the magnetic field of a moving charge.

The above considerations are restricted to the Maxwell equations in vacuum, apart from the source four-vector, J^{μ}. For a description of the relativistic formulation of moving media, where $(\mathbf{E}/c,\mathbf{B})$, $(c\mathbf{P},\mathbf{M})$, and $(c\mathbf{D},\mathbf{H})$ are separately second-rank tensors (with the electric and magnetic properties of the medium defined in its rest frame), see Pauli (1958) or Penfield and Haus (1967).

7.3 Lagrangian Formulation of Electrodynamics

The powerful methods of Lagrangians, familiar in the classical mechanics of systems with a finite number of degrees of freedom, can be applied to continuous systems (with an infinite number of degrees of freedom). The finite number of coordinates and velocities, $q_i(t)$, $\dot{q}_i(t)$, $i = 1,2,3,\ldots,n$, are replaced by continuous fields and their derivatives. The discrete index i is replaced by the continuous space-time coordinate x^{μ} and perhaps by a finite number of discrete values at each x^{μ}. The correspondences are $q_i(t) \to \phi_k(x)$, $\dot{q}_i(t) \to \partial^{\beta}\phi_k(x)$. Here x stands for the space and time coordinates, and there are N different fields ϕ_k, $k = 1,2,3,\ldots,N$, defined at each point x. The dynamics is defined by a Lagrangian density $\mathscr{L}$ whose integral over all three-space is the Lagrangian L. The Euler–Lagrange equations follow from the stationary property of the action integral (of L over time) with respect to variations $\delta\phi_k$ and $\delta(\partial^{\beta}\phi_k)$ around the physical values. There are thus the following correspondences:

$$i \to x^{\mu},k,$$

$$q_i \to \phi_k(x),$$

$$\dot{q}_i \to \partial^{\beta}\phi_k(x),$$

$$L = \sum_i L_i(q_i,\dot{q}_i) \to \int \mathscr{L}(\phi_k,\partial^{\beta}\phi_k)\,d^3x,$$

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) = \frac{\partial L}{\partial q_i} \to \partial^{\beta}\frac{\partial\mathscr{L}}{\partial(\partial^{\beta}\phi_k)} = \frac{\partial\mathscr{L}}{\partial\phi_k}. \qquad (125)$$

The action integral S takes the form

$$S = \int dt \int d^3x\,\mathscr{L} = \int \mathscr{L}\,d^4x. \qquad (126)$$

The necessary Lorentz-invariant nature of the action is preserved provided the Lagrangian density $\mathscr{L}$ is a Lorentz scalar. By analogy with discrete-particle dynamics it is expected that the kinetic energy part of the Lagrangian density is quadratic in the velocities, i.e., in $\partial^{\beta}\phi_k$. The combination, $\partial^{\beta}\phi_k\partial_{\beta}\phi_k$, for each k has the required invariance under Lorentz transformations. Additional terms in $\mathscr{L}$ depend on the physical properties of the fields and their interactions.

For the electromagnetic field, the four-vector potential A^{β} plays the role of the set of ϕ_k. The Lagrangian density is expected to be quadratic in the derivatives $\partial^{\alpha}A^{\beta}$ or equivalently in $F^{\alpha\beta}$. Equation (122) displays the only Lorentz-invariant quadratics in $F^{\alpha\beta}$. Restriction of $\mathscr{L}$ to a scalar under spatial inversions eliminates the second form and leaves a multiple of $F^{\alpha\beta}F_{\alpha\beta}$ as the only possibility. If one recalls that the Hamiltonian (energy) density involves subtraction of $\mathscr{L}$ and compares Eq. (91) with the first term in Eq. (122), it is apparent that the Lagrangian density must be

$$\mathscr{L}=-\frac{1}{4\mu_0}F_{\alpha\beta}F^{\alpha\beta}-J_\alpha A^\alpha. \tag{127}$$

The interaction of the potentials with the external sources J^β has also been included. In order to find the Euler–Lagrange equations of motion [last line in Eq. (125)], the Lagrangian density must be written in terms of the potentials:

$$\mathscr{L}=-\frac{1}{4\mu_0}g_{\lambda\mu}\,g_{\nu\sigma}(\partial^\mu A^\sigma-\partial^\sigma A^\mu)(\partial^\lambda A^\nu-\partial^\nu A^\lambda) \\ -J_\alpha A^\alpha. \tag{128}$$

The ingredients of the Euler–Lagrange equation of motion are

$$\frac{\partial\mathscr{L}}{\partial(\partial^\beta A^\alpha)}=-\frac{1}{\mu_0}F_{\beta\alpha}=\frac{1}{\mu_0}F_{\alpha\beta},\quad \frac{\partial\mathscr{L}}{\partial A^\alpha}=-J_\alpha$$

with the equation of motion taking the form

$$\partial^\beta F_{\beta\alpha}=\mu_0 J_\alpha, \tag{129}$$

equivalent to the first equation in Eq. (121). The homogeneous Maxwell equations do not emerge as Euler–Lagrange equations of motion—they are already implicit in the definition of the fields in terms of the potentials (See Sec. 5.1).

7.4 Further Ramifications

The reader who wishes to explore further ramifications of the relativistic formulation of classical field theory may consult Landau and Lifshitz (1975), Rohrlich (1990), or Soper (1976). Construction of the relativistic generalization of the Hamiltonian density (called the stress–energy–momentum tensor) and its connection to a unified description of the conservation of energy and momentum are discussed briefly in Jackson (1975), Sec. 12.10, and in the just cited references.

GLOSSARY

Admittance: Symbol, $Y=G-iB$ (engineers: read $Y=G+jB$); reciprocal of impedance (Z).

Ampere: The SI unit of electric current, abbreviated A, equal to one coulomb per second.

Capacitance: Amount of charge carried by a conductor raised to a potential difference of one volt.

Capacitance, Coefficients of: Generalization of capacitance; geometrical coefficients for a given configuration of conducting surfaces and linear dielectric media between, which permit the total electrostatic energy to be expressed in terms of the potentials of the conductors—see Eq. (54).

Causality: Loosely, cause precedes effect; the response of a system to an external stimulus always follows in time the application of the stimulus. In relativity, it also involves the concept that signals do not travel faster than the speed of light in vacuum.

Conductivity: Symbol, σ; local property of conducting media, reciprocal of resistivity.

Constitutive Relations: Relations specifying the dielectric and magnetic properties of a medium.

Coulomb: The SI unit of electric charge, abbreviated C. The magnitude of the charge on the electron or the proton in this unit is $e=1.6022\times10^{-19}$ C.

Dielectric Constant: Symbol, $\epsilon(\omega)$ [or $\epsilon_r(\omega)$, for $\epsilon(\omega)/\epsilon_0$]; frequency-dependent proportionality between electric displacement and electric field in simple linear media. More generally a tensor, $\epsilon_{jk}(\omega)$.

Dispersion: Property of dielectrics or magnetic media that the product of the wavelength and the frequency of electromagnetic waves depends on the frequency (in vacuum, $\lambda\nu=c$).

Dispersion Relations: Relations between the real and imaginary parts of the dielectric constant; consequences of causality—see Eq. (46). Have applications in other fields.

Displacement, Electric: Symbol **D**; field variable in dielectrics, related to electric field and electric polarization—see Eq. (24). The SI unit is coulomb per square meter.

Electric Field: Symbol, **E**; a vectorial quantity, defined at a given space-time point as the mechanical force per unit charge experienced by an infinitesimal test charge placed at the point. The SI unit is volts per meter.

Electromotive Force: Line integral of the electric field around a closed path in space; accompanies a magnetic flux density changing in time.

Farad: The SI unit of capacitance, abbreviated F; one coulomb per volt.

Gauge, Gauge Transformation: Particular choice of scalar and vector potentials; modification of choice of potentials, without change in the physically observable quantities.

Henry: The SI unit of inductance, abbreviated H. One henry equals one volt–second per ampere.

Hertz Vectors: Auxiliary potentials used in solution of the wave equation—see Sec. 5.3.

Impedance: Symbol, $Z=R-iX$ (engineers: read $R+jX$). Complex quantity with dimensions of resistance (ohms) that describes the magnitude and phase of the voltage across the terminals of an electrical device with respect to an input current.

Impedance of Free Space: Constant ($\sqrt{\mu_0/\epsilon_0}$) with dimensions of resistance; ratio of electric to magnetic field strengths in a plane electromagnetic wave in vacuum. Numerically, $\sqrt{\mu_0/\epsilon_0}=376.730\ \Omega$.

Inductance: Symbol, L (self-inductance) or M_{12} (mutual inductance); geometrical property of a coil of wire or other circuit describing its ability to develop a voltage difference across it when the current through it (or a neighboring circuit) changes with time—a manifestation of Faraday's law. Symbolically, $V=L\,dI/dt$. The SI unit is the henry.

Inductance, Coefficients of: Totality of self- and mutual inductances of a configuration of circuits, permitting the total energy of the configuration to be expressed in terms of the currents in the circuits—see Eq. (60).

Kramers–Kronig Relations: Another name for dispersion relations.

Linear Media: Substances whose induced electric or magnetic polarizations are linear in externally applied fields.

Lorentz Transformation: Linear transformation in special relativity of space-time coordinates from one coordinate frame to another moving at uniform velocity with respect to the first. More generally, the transformation of four-vectors and higher-rank four-tensors.

Magnetic Field: Symbol, **H**; field variable related to the magnetic flux density and magnetization—see Eq. (24). The SI unit is ampere-turn per meter.

Magnetic Flux Density: Symbol, **B**; analog for currents of electric field. Defined in magnitude as force per unit current, but directions of current flow, **B**, and force are not parallel—see discussion above Eq. (6). The SI unit is the tesla (sometimes, weber per square meter).

Magnetization: Symbol, **M**; macroscopic magnetic property of a substance, the average molecular or atomic magnetic dipole moment density—see Sec. 3.1. The SI unit is ampere per meter.

Maxwell Stress Tensor: Symbol, T_{jk}; tensor bilinear in the electric and magnetic field strengths, used to calculate electromagnetic forces—see Eq. (106).

Minkowski Space: Non-Euclidean space-time—see Eq. (114).

Ohm: The SI unit of resistance, abbreviated Ω; one ohm is equal to one volt per ampere.

Ohm's Law: Law obeyed by many conducting substances: current density locally is proportional to electric field. Generally, there is a tensor of proportionality; often it is a scalar quantity.

Plasma Frequency: Symbol, ω_p; material property governing the high-frequency behavior of dielectric constants—see Eq. (38); depends only on density of electrons (not details of atomic structure).

Polarization, Electric: Symbol, **P**; macroscopic electric property of a substance, the average molecular or atomic electric dipole moment density—see Sec. 3.1. The SI unit is coulomb per square meter.

Pressure, Radiation: Electromagnetic momentum per unit area per unit time.

Resistance: Symbol, R; property of conductor or circuit element to limit the flow of current for a given applied voltage. Symbolically, $V=IR$. The SI unit is the ohm.

Resistance, Radiation: Effective resistance of a radiating element; ratio of radiated power to the square of the driving current.

Resistivity: Symbol, ρ; local property of a conducting substance; enters Ohm's law: $\mathbf{E}=\rho\mathbf{J}$.

Permeability of Free Space: Constant (μ_0) entering the relation between electric current and magnetic flux density [see Eq. (8)]. Defined to be $\mu_0=4\pi\times10^{-7}$ henrys per meter.

Permittivity of Free Space: Constant (ϵ_0) entering the relation between charge and electric field [see Eq. (3)]. Its inverse, divided by 4π, is a 17-digit number related to the speed of light in vacuum. Numerically, $\epsilon_0=8.854\ 19...\times10^{-12}$ farads per meter.

Potential, Four-Vector: Symbol, A^{μ}; set of four potentials, ϕ and the three components of **A**, which transform relativistically in the same way as $x^{\mu}=(x^0,\mathbf{x})$.

Potential, Scalar: Symbol, ϕ; in static situations, the difference $\phi(\mathbf{r}_1)-\phi(\mathbf{r}_2)$ represents the line integral of the electric field from the point $\mathbf{r}_1$ to the point $\mathbf{r}_2$.

Potential, Vector: Symbol, **A**; an auxiliary field related to the magnetic flux density through $\mathbf{B}=\nabla\times\mathbf{A}$.

Skin Depth: Symbol, δ; depth of penetration of fields into a good conductor—see Eq. (39).

Système Internationale (SI): An internationally agreed system of weights and measures; roughly, the metric system.

Tesla: The SI unit of magnetic flux density, abbreviated T; one tesla is equal to one newton per ampere–meter.

Volt: The SI unit of electric potential difference, abbreviated V; one volt equals one joule per coulomb.

List of Works Cited

Altarelli, M., Dexter, D. L., Nussenzveig, H. M., Smith D. Y. (1972), *Phys. Rev. B* **6**, 4502–4509.

Beam, W. R. (1965), *Electronics of Solids,* New York: McGraw-Hill.

Brillouin, L. (1960), *Wave Propagation and Group Velocity,* New York: Academic.

Collins, G. B. (Ed.) (1948), *Microwave Magnetrons,* MIT Radiation Laboratory Series, Vol. 6. New York: McGraw-Hill.

deGroot, S. R. (1969), *The Maxwell Equations,* Studies in Statistical Mechanics, Vol. IV. Amsterdam: North Holland.

Gray, D. E. (Ed.), *American Institute of Physics Handbook* (1972), 3rd ed., New York: McGraw-Hill.

Hecht, E., Zajac, A. (1974), *Optics,* 2nd ed., Reading, MA: Addison-Wesley.

Jackson, J. D. (1975), *Classical Electrodynamics,* 2nd ed., New York: Wiley.

Kittel, C. (1986), *Introduction to Solid State Physics,* 6th ed., New York: Wiley.

Landau, L. D., Lifshitz, E. M. (1975), *Classical Theory of Fields,* 4th rev. Eng. ed., Oxford: Pergamon.

Landau, L. D., Lifshitz, E. M., Pitaevskii, L. P. (1984), *Electrodynamics of Continuous Media,* 2nd ed., Oxford: Pergamon.

Pauli, W. (1958), *Theory of Relativity,* Oxford: Pergamon.

Penfield, P., Haus, H. A. (1967), *Electrodynamics of Moving Media,* Cambridge, MA: MIT Press.

Pines, D. (1963), *Elementary Excitations in Solids,* New York: Benjamin.

Pippard, A. B. (1960), *Rep. Progr. Phys.* **23**, 176–266.

Robinson, F. N. H. (1973), *Macroscopic Electromagnetism,* Oxford: Pergamon.

Rohrlich, F. (1990), *Classical Charged Particles,* 2nd ed., Redwood City, CA: Addision-Wesley.

Slater, J. C. (1950), *Microwave Electronics,* New York: Van Nostrand.

Soper, D. E. (1976), *Classical Field Theory,* New York: Wiley.

Weast, R. C. (Ed.), *Handbook of Chemistry and Physics* (1989), 69th ed., Boca Raton, FL: CRC.

Whittaker, E. T. (1951), *A History of the Theories of Aether and Electricity,* Vol. 1, revised and enlarged, London: Nelson; reprinted (1960) by New York: Harper Torchbooks.

Wilson, A. H. (1953), *Theory of Metals,* 2nd ed., Cambridge: Cambridge University Press.

Wooten, F. (1972), *Optical Properties of Solids,* New York: Academic.

Further Reading

Electromagnetism is part and parcel of almost everything that exists or happens. Its applications are everywhere. Recommendations for further reading can only touch on a few areas or subjects, largely reflecting the writer's knowledge and prejudices.

For more leisurely or elementary accounts of electricity and magnetism, consult the physics books,

Griffths, D. J. (1989), *Introduction to Classical Electrodynamics,* 2nd ed., New York: Prentice-Hall.

Purcell, E. M. (1985), *Electricity and Magnetism,* 2nd ed., New York: McGraw-Hill.

For a more applied, but still elementary, account, see

Kraus, J. D. (1984), *Electromagnetics,* 3rd ed., New York: McGraw-Hill.

For applications in engineering, particularly transmission and radiation, see

Collin, R. E. (1966), *Field Theory of Guided Waves,* New York: McGraw-Hill.

Collin, R. E. (1985), *Antennas and Radiowave Propagation,* New York: McGraw-Hill.

Ramo, S., Whinnery, J. R., Van Duzer, T. (1984), *Fields and Waves in Communication Electronics,* 2nd ed., New York: Wiley.

The subjects of radiation, scattering, and diffraction are treated in

Jackson, J. D. (1975), *Classical Electrodynamics,* 2nd ed., New York: Wiley, Chaps. 9 and 16.

Landau, L. D., Lifshitz, E. M. (1975), *Classical Theory of Fields,* 4th rev. Eng. ed., Oxford: Pergamon.

Special relativity is a topic far broader than electromagnetism. For discussion of its place in mechanics, see

Møller, C. (1972), *The Theory of Relativity*, 2nd ed., Oxford: Clarendon.

Sard, R. D. (1970), *Relativistic Mechanics*, New York: Benjamin.

For the relativistic theory of classical charged particles and fields, consult

Rohrlich, F. (1990), *Classical Charged Particles*, 2nd ed., Redwood City, CA: Addison-Wesley.

Soper, D. E. (1976), *Classical Field Theory*, New York: Wiley.

Discussion of relativistic radiation phenomena, including synchrotron radiation, transition radiation, classical and semiclassical bremsstrahlung, etc., may be found in

Jackson, J. D. (1975), *Classical Electrodynamics*, 2nd ed., New York: Wiley, Chaps. 14 and 15.

ELECTRODYNAMICS, QUANTUM

See QUANTUM ELECTRODYNAMICS

ELECTROLUMINESCENCE

YOSHIMASA A. ONO, *Hitachi Research Laboratory, Hitachi, Ltd., Ibaraki-ken, Japan*

Introduction 1
1. **EL Device Structures and EL Emission Mechanisms** 3
1.1 ac Powder EL 3
1.2 dc Powder EL 4
1.3 ac Thin-Film EL 5
1.4 dc Thin-Film EL 6
1.5 Organic Thin-Film EL 6
2. **Physics of ac Thin-Film EL** 7
2.1 Equivalent-Circuit Model and EL Emission Mechanism 7
2.2 Electro-optical Characteristics 9
2.3 Physical and Chemical Properties of Phosphor Materials and Luminescent Centers 11
3. **Materials Requirements** 12
3.1 Glass Substrates 13
3.2 Transparent Conducting Films 13
3.3 Insulating Layers 13
3.4 Phosphor Layers 15
3.5 Rear Electrodes 15
4. **Material Deposition Methods** 15
4.1 PVD Methods 16
4.1.1 EBD Method 16
4.1.2 MSD Method 16
4.1.3 Sputtering Method 16
4.2 CVD Methods 16
4.2.1 ALE Method 16
4.2.2 MOCVD Method 17
4.2.3 HT-CVD Method 17
4.3 Summary of Deposition Methods 18
5. **Color Thin-Film Electroluminescent Devices** 18
5.1 Color Thin-Film EL Devices Based on ZnS Phosphors 19
5.1.1 Orange-Yellow (Amber) Emitting ZnS:Mn 19
5.1.2 Multicolor-emitting ZnS:RE,F (RE: Rare-Earth Elements) 19
5.2 Color Thin-Film EL Devices Based on CaS and SrS Phosphors 21
5.2.1 Red-Emitting CaS:Eu 21
5.2.2 Green-Emitting CaS:Ce and Blue-Green–Emitting SrS:Ce 21
5.3 Summary of Color Thin-Film EL Devices 22
6. **Multicolor Thin-Film EL Panels** 22
6.1 Stacked-Phosphor Structure 22
6.2 Patterned-Phosphor Structure 23
6.3 White or Broadband-Spectrum Phosphor with Color Filters 23
7. **Reliability** 24
7.1 Electrical Reliability 24
7.2 Long-Time Reliability 25
8. **Future Prospects** 26
Glossary 28
List of Abbreviations 28
List of Works Cited 28
Further Reading 31

INTRODUCTION

Electroluminescence (EL) is a nonthermal generation of light resulting from the application of an electric field to a substance. Although this definition includes light-emitting diodes (LEDs), they are not discussed here; rather they are considered elsewhere in the Encyclopedia (see LIGHT-EMITTING DIODES). EL devices are divided into four types according to the phosphor configuration and the drive-voltage wave forms: ac powder EL, dc powder EL, ac thin-film EL, and dc thin-film EL. Besides these, organic thin-film EL (Tang and VanSlyke, 1987) is of interest, although this belongs to the LED family. Among these,

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

two types are now commercially available: ac thin-film EL devices (ZnS:Mn) are used as flat screens of laptop personal computers and word processors, and ac powder EL devices are used as backlights for liquid-crystal displays.

High-field electroluminescence was discovered by Destriau in 1936 (Destriau, 1936). Light emission from a ZnS phosphor powder dispersed in an insulator and sandwiched between two electrodes was observed when a high ac voltage was applied. Until 1950, when transparent electrically conductive films made of SnO_2 were developed, no effort was made to develop practical devices. With use of these SnO_2 transparent conductive films, ac powder EL devices were developed, which triggered worldwide research and development of EL devices.

During the period from 1950 to 1960, basic studies on the ac powder EL devices were carried out actively, aiming at flat light sources for wall illumination. However, it was not possible to overcome the weaknesses of these ac powder EL devices: low luminance, low luminous efficiency, and short lifetime. Thus, the first era of EL development gradually drew to a close.

In 1968, the second period of EL development began with dc powder EL devices (Vecht *et al.*, 1968) and ac thin-film EL devices using Lumocen (luminescence from molecular centers) (Kahng, 1968; Chase *et al.*, 1969). The former were made possible by coating powder EL phosphor surfaces with copper ions. These devices were intensively studied because technical advantages of producing large-area multiplexed panels were expected at relatively low production cost. In the latter, rare-earth fluorides as molecular centers were doped into ZnS thin films. These developments were made possible by advances in electronics and materials science, especially in new process technologies and device structures. Thin-film EL devices benefitted, in particular, from development of the double-insulating-layer structure (Russ and Kennedy, 1967), which overcame the problems of short lifetime and low luminance. Inoguchi *et al.* (1974) obtained stable high-luminance thin-film EL panels with this structure, and Mito *et al.* (1974) showed that these EL panels could be used as a TV imaging system. Since then, this type of EL device has been investigated most intensively among the four types because of the highest possibility of practical use.

In regard to the production of thin-film EL panels, Ueda *et al.* (1981) reported on practical EL display units, and Takeda *et al.* (1981) reported on practical application technologies of ZnS:Mn EL display panels. With use of these technologies, mass production of 6-in.-diagonal matrix EL panels started in 1983 (Takeda *et al.*, 1983).

At this point, we consider dc thin-film EL development. Cusano (1962) showed that the luminance of a ZnS:Mn thin film coated with copper gradually increases during operation as a result of an electrical forming process. Although the fabrication process was improved, luminance and reliability improvements were not realized, and interests in dc thin-film EL faded away. It should be noted, however, that low-voltage ($\sim$10 V) driven ZnSe:Mn dc thin-film EL devices were studied using the molecular-beam deposition method (Mishima *et al.*, 1983; Mishima and Takahashi, 1983).

The following treatment concentrates on ac thin-film EL devices, which are all-solid and emissive displays with the following characteristics: fast response, wide viewing angles, high resolution, wide operating temperatures, light weight, and good display quality. Displays with 640×400 dots are now in mass production, and higher-resolution A4-sized display devices with 1024×800 dots (Schmachtenberg *et al.*, 1989) are being developed. These display devices are compatible with high-resolution cathode ray tubes (CRTs). Table 1 lists the specifications of commercially available thin-film EL display devices.

To show more information on the screen, multicoloring of EL displays is necessary. For this purpose, phosphor materials that emit three primary colors, i.e., red, green, and blue, are necessary. Beginning with the pioneering work by Kahng and co-workers (Kahng, 1968; Chase *et al.*, 1969), extensive studies have been conducted on materials and processing technologies. Most efforts were concerned with the ZnS host material doped with various rare-earth luminescent centers. Green-emitting ZnS:Tb,F EL devices were obtained with high luminance and high luminous efficiency, close to those of ZnS:Mn (Ogura *et al.*, 1986; Ohnishi *et al.*, 1987). Despite all these efforts, however, the maximum luminances of blue-emitting ZnS:Tm,F EL devices and red-emit-

Table 1. Specifications of commercial EL displays (ZnS:Mn), listed by manufacturer and model number.

	Sharp		Planar	Lohja[a]
Items	LJ024U33	LJ640U48	EL751214M	MD640.400
Display area (mm^2)	256×200	192×144	345×292	195×122
Inch diagonal	12.8	9.4	18	9
Display format (dot)	1024×800	640×480	1024×800	640×400
Dot pitch (mm)	0.25	0.3	—	0.3
Luminance (cd/m^2)	100	100	70	110
Luminance (ft. L)	30	30	20	32
Viewing angle (°)	>120	>120	>160	>140
Frame frequency (Hz)	60	60	—	60
Operating temperature (° C)	0–55	0–55	10–55	0–55
Power consumption (W)	22 (typ.)	14 (typ.)	60 (max.)	14 (typ.)
Weight (g)	1000	660	6800	480

[a] Now Planar Systems, International.

ting ZnS:Sm,F EL devices are low compared with those of green-emitting ZnS:Tb,F EL devices (Kobayashi *et al.*, 1985). But recently, bright red EL emission has been obtained by using ZnS:Mn phosphor with a thin-film red filter made of CdSSe (Tuenge and Kane, 1991). And ZnS-based phosphors have been utilized in the fabrication of prototype multicolor thin-film EL devices (Yamauchi *et al.*, 1987; Yamamoto *et al.*, 1989).

In order to obtain efficient color EL phosphors, many studies have been directed to new host materials and luminescent centers. It was found that alkaline-earth sulfides, such as SrS and CaS, doped with rare-earth luminescent centers were promising phosphor materials (Barrow *et al.*, 1984; Tanaka *et al.*, 1985). As these phosphors were good cathodoluminescence phosphors (Lehmann, 1972), they were used for dc power EL phosphors (Higton *et al.*, 1978; Vecht, 1980). The thin-film EL phosphors CaS:Eu and SrS:Ce also seemed promising, and prototype multicolor display panels were fabricated with them (Barrow *et al.*, 1986).

The present review discusses the characteristics of EL devices, with an emphasis on double-insulating-layer type thin-film EL devices, materials requirements, material deposition methods, color EL materials, multicolor EL devices, and reliability. For other topics, readers are referred to the review articles by Howard (1981), Mach and Müller (1982), Kobayashi (1983), Alt (1984), and Ohnishi (1989).

1. EL DEVICE STRUCTURES AND EL EMISSION MECHANISMS

In this section device structures and emission mechanisms of four kinds of EL devices are described. The characteristics of organic thin-film EL diode devices are also discussed.

1.1 ac Powder EL

This type of EL device was first developed by Sylvania in the first era of EL devices. Now they are used as backlights of liquid-crystal displays. The device structure is shown in Fig. 1. The phosphor layer consists of a suitably doped ZnS powder (5–20-μm grains) suspended in a dielectric, which acts as a binder as well. The phosphor layer is sandwiched between two electrodes, one of which is transparent, and is supported by a substrate, consisting of either glass or flexible plastic. As a dielectric, organic material, such as cyanoethylcellulose, or low-melting glass is used. An insulating layer may be placed between the phosphor layer and the rear electrode to avoid catastrophic dielectric breakdown. The important point is that the ZnS grains should be isolated from each other, as shown in the enlarged view of the phosphor layer. In the case of good isolation, an insulating layer is not necessary.

Electric fields of 10^4–10^5 V/cm are necessary to obtain EL emission. Luminance increases proportionally with frequency up to 10 kHz but lifetime is decreased in the same proportion. The luminance level is about 100

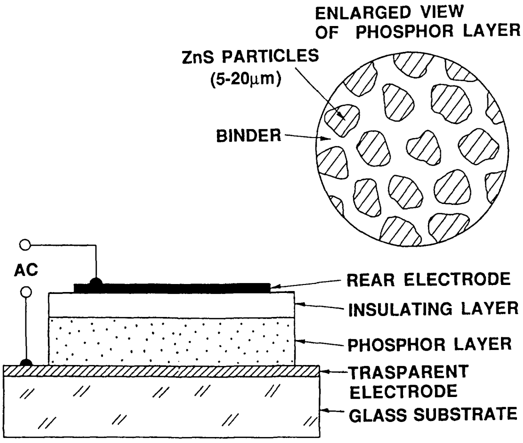

FIG. 1. Cross-sectional device structure of ac powder EL device. Enlarged view of phosphor layer is also shown.

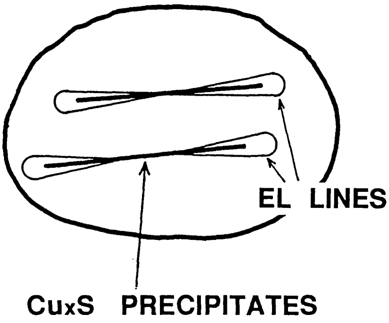

FIG. 2. EL emission mechanism of ac powder EL device (after Fischer, 1963).

cd/m^2 at a voltage of 300 V and frequency of 400 Hz and luminous efficiency is about 1 lm/W.

Emission colors can be controlled by adding different kinds of luminescent centers. A combination of Cu and Cl (ZnS:Cu,Cl) gives either blue (~460 nm) or green (~510 nm) emission, depending on the relative amount of Cl. This EL emission is caused by the recombination transition of *D-A* (donor–acceptor) pairs, where Cu constitutes an acceptor and Cl constitutes a donor. The combination of Cu and Al (ZnS:Cu,Al) gives green and the addition of Mn to Cu and Cl (ZnS: Cu,Cl,Mn) gives yellow (~590 nm). An advantage of this type of EL device is easy control of emission colors by mixing different phosphors with different emission colors.

The EL emission mechanism is considered to be as follows (Fischer, 1962, 1963). EL powders fired at high temperatures have the hexagonal structure but subsequent cooling transforms them into the cubic (zinc blende) structure. Then copper exceeding the solubility limit precipitates on defects. The result is embedded Cu_xS conducting needles, indicated by Cu_xS precipitates in Fig. 2. These precipitates act to concentrate an applied electric field at their tips, and therefore an applied field of 10^4–10^5 V/cm can induce a local field of 10^6 V/cm or more. This electric field is strong enough to induce tunneling of holes from one end of the needle and electrons from the other. The holes are trapped on Cu recombination centers, and upon reversal of the fields the emitted electrons recombine with the trapped holes to produce light. Thus EL emission occurs along the Cu_xS precipitates, as indicated by the EL lines. Larger particle sizes lead to longer needles and greater field enhancement.

Lifetime is a key issue for the application of this EL device. The half-life, defined by the time when the luminance becomes $\frac{1}{2}$ the initial value, now exceeds 10 000 hours through the improvement of phosphor treatment conditions. However, it is very difficult to have both long lifetime and high luminance simultaneously because they are trade-off characteristics.

1.2 dc Powder EL

Since the pioneering work by Vecht *et al.* (1968), dc powder EL devices have been extensively studied in Great Britain. Now commercial panels with the configuration of 640×200 dots are available. In the dc powder EL device, a fine (0.5–1 μm) Mn-doped ZnS powder is deliberately prepared with a Cu_xS coating, and a layer of the powder with a small amount of binder is formed into a device, as shown in Fig. 3. In contrast to the ac powder EL case, there is no insulating layer. To obtain stability and emission uniformity, a forming process is necessary. When a voltage is first applied, a large current flows, the layer heats up, and gradually a narrow region ~1 μm thick adjacent to the SnO_2 anode, begins to luminesce (see Fig. 4).

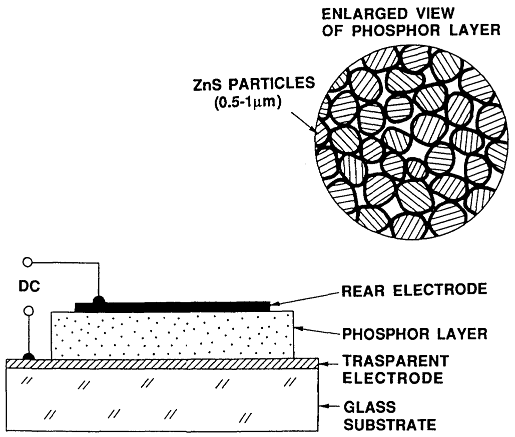
ENLARGED VIEW
OF PHOSPHOR LAYER
ZnS PARTICLES
(0.5-1μm)
DC
REAR ELECTRODE
PHOSPHOR LAYER
TRASPARENT
ELECTRODE
GLASS
SUBSTRATE

FIG. 3. Cross-sectional device structure of dc powder EL device. Enlarged view of phosphor layer is also shown.

The emission color of ZnS:Mn,Cu is yellow (~590 nm). In ZnS-based EL devices activated by rare-earth ions, Tm^{3+} gives blue emission, Tb^{3+} and Er^{3+} give green emission, and Nd^{3+} and Sm^{3+} give red emission. In CaS- and SrS-based EL devices, SrS:Ce,Cl gives blue-green emission, CaS:Ce,Cl gives green emission, and CaS:Eu,Cl gives red emission. Luminous efficiency, however, is not high with values of 0.2–0.3 lm/W.

Figure 4 shows the microscopic model of dc powder EL. In the forming process, electromigration drives all copper out of a high-field region near the anode. As the width of this copper-free region grows, the field is progressively reduced. Under a high electric field, electrons, which have tunneled out of the Cu_xS at the edge of the formed region, are accelerated and excite Mn^{2+} luminescent centers by impact. When the excited electrons make transitions radiatively to the ground state, EL emission is realized.

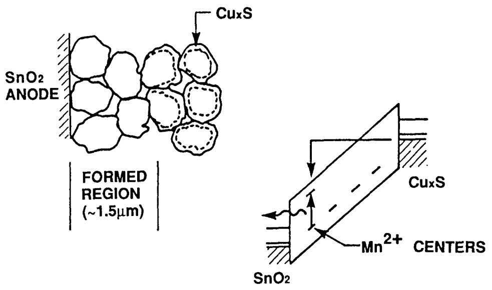
CuxS
SnO2
ANODE
FORMED
REGION
(~1.5μm)
CuxS
Mn2+ CENTERS
SnO2

FIG. 4. Microscopic model of dc powder EL device.

1.3 ac Thin-Film EL

Since Inoguchi *et al.* (1974) reported on high-luminance, long-lifetime EL devices, ac thin-film EL with the double-insulating-layer structure has been the center of research and development. A schematic structure of this type of EL device is shown in Fig. 5, where a thin-film phosphor layer is sandwiched between thin-film insulating layers. These thin films are deposited on the ITO (indium tin oxide)-coated glass substrate, and Al rear electrodes are deposited on the second insulating layer. The thicknesses of the phosphor and the insulating layers are 0.5–1 and 0.3–0.5 μm, respectively. When an ac voltage of about 200 V is applied between the ITO and Al electrodes, EL emission is realized.

The double-insulating-layer type EL devices have the following advantages:

1. Double insulating layers can protect the phosphor layer against impurities and moisture from outside, to ensure higher stability.
2. In this structure, direct electron flows from electrodes to the phosphor layer are prevented, resulting in a high breakdown electric field.
3. Trapped charges at the phosphor layer/insulating layer interfaces cause internal

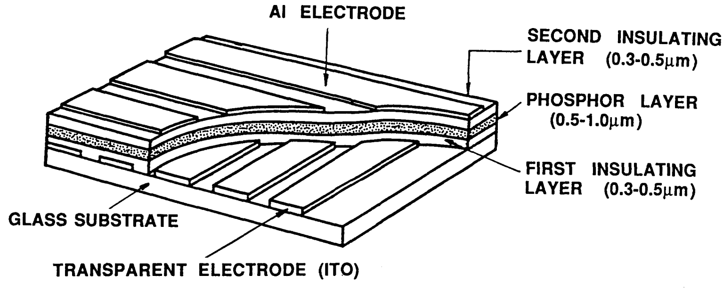
Al ELECTRODE
SECOND INSULATING
LAYER (0.3-0.5μm)
PHOSPHOR LAYER
(0.5-1.0μm)
FIRST INSULATING
LAYER (0.3-0.5μm)
GLASS SUBSTRATE
TRANSPARENT ELECTRODE (ITO)

FIG. 5. Structure of double-insulating type thin-film EL device.

polarization, which increases the effective electric field under ac drive conditions. As a result, luminous efficiency and luminance are increased.

Emission colors can be controlled by adding different luminescent centers. ZnS:Mn gives orange-yellow emission with the highest luminance. In the ZnS-based EL devices activated by rare-earth ions, ZnS:Tb,F gives green emission, ZnS:Tm,F gives blue emission, and ZnS:Sm,F gives orange-red emission. In the CaS- and SrS-based EL devices, SrS:Ce gives blue-green emission, CaS:Ce gives green emission, and CaS:Eu gives red emission. It should be noted that copper, which is used in powder EL, is not used in ac thin-film EL because copper ions migrate under a high electric field ($\sim 10^6$ V/cm), causing deterioration of the EL device characteristics.

The details of the equivalent-circuit model and the EL emission mechanism are given in Sec. 2.1.

1.4 dc Thin-Film EL

This type of EL device has the simplest structure: a thin film of phosphor layer with electrodes on both sides. Several attempts have been carried out by depositing a thin-film phosphor layer of ZnS:Mn, ZnS:Cu,Cl, and ZnS:Mn,Cu to increase stability and luminance, but good results have not been obtained. Since the basic problem with the device is tendency toward catastrophic failure, employment of a current-limiting layer should be considered. Kobayashi *et al.* (1989) succeeded in fabricating a stable ZnS:Mn EL device by inserting a thick resistive powder layer made of MnO_2 between the phosphor layer and rear electrode. They called this device dc thin-film-powder hybrid EL. The luminous efficiency was 0.8 lm/W and the lifetime was more than 20 000 h.

1.5 Organic Thin-Film EL

Figure 6 shows the structure of an organic thin-film EL diode device (Tang and Van Slyke, 1987; Tang *et al.*, 1989). The substrate is an ITO-coated glass and the first organic layer (about 75 nm) is a hole-transport layer made of amorphous diamine film of molecular structure shown in Fig. 6. The second organic layer (about 60 nm) is the luminescent layer as well as an electron-transport layer, made of 8-hydroxyquinoline aluminum (Alq_3) shown in Fig. 6. The rear electrode is a mixture of Mg and Ag with an atomic ratio of 10:1. The emission color of undoped Alq_3 was green, and a luminance of more than 1000 cd/m^2 and a luminous efficiency of 1.5 lm/W were achieved at a dc pulse driving voltage of 10 V. In the device with an Alq_3 layer doped with fluorescent molecules, the EL quantum efficiency was about 0.025 photon/electron. The EL color was tuned from blue-green to orange-red by a suitable choice of dopants as well as by changing the concentration of the dopant. The electron–hole recombination and emission zones were confined to about 5 nm near the hole-transport interface.

Adachi *et al.* (1988) introduced a three-layer structure made up of a hole-transport layer, a luminescent layer, and an electron-transport layer. By separating the luminescent function from the hole- or electron-transport layers, it became possible to select luminescent-layer materials freely to obtain different emission colors. Figure 7 shows the

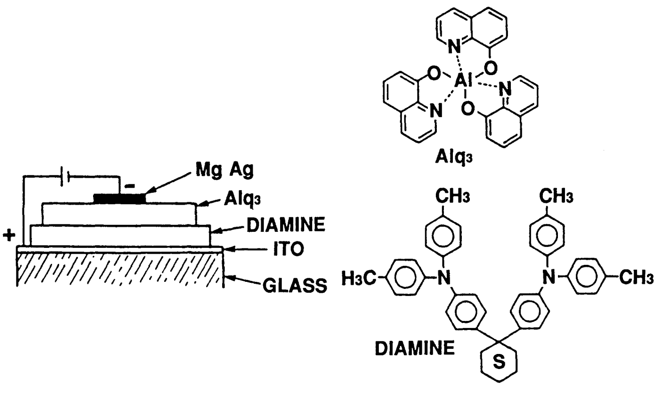

FIG. 6. Configuration of organic EL diode device and molecular structures of Alq_3 and diamine (after Tang and VanSlyke, 1987).

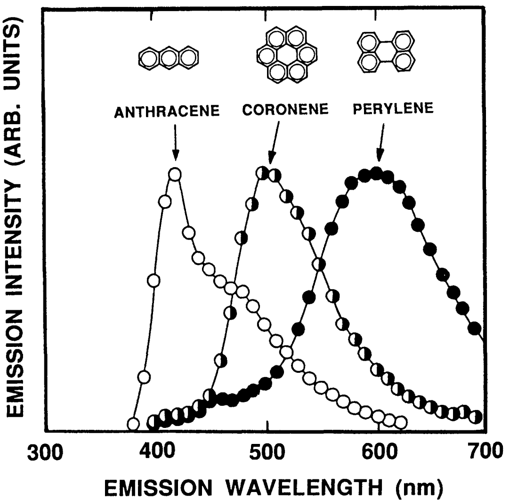

FIG. 7. EL emission intensities of three layered organic EL diode devices with emitter materials; open circles are for anthracene, half-black and half-open circles are for coronene, and black circles are for perylene. Chemical structures of these are shown above (after Adachi *et al.*, 1988).

results: By changing the luminescent-layer material from anthracene to coronene to perylene, the emission color changed from blue to green to red. Here an aromatic diamine and a perylene tetracarboxylic acid derivative were used for the hole- and electron-transport materials, respectively.

Lifetime is a key issue for the application of this EL diode device.

2. PHYSICS OF ac THIN-FILM EL

In this section the physics of ac thin-film EL devices with double-insulating-layer structure is discussed. First, based on the current–voltage (I-V) characteristics, an equivalent-circuit model is introduced and the EL emission mechanism is discussed by use of an energy-band diagram. Then detailed treatments of the electro-optical characteristics are given, and the physical and chemical properties of phosphor materials and luminescent centers of color thin-film EL are described.

2.1 Equivalent-Circuit Model and EL Emission Mechanism

Figure 8 shows the current density vs voltage (I-V) and phase difference vs voltage (ϕ-V) characteristics under 1-kHz sinusoidal-wave drive conditions. The I-V profile is given by two straight lines linked with a kink, the voltage of which corresponds to the threshold voltage. The profile of the phase difference ϕ between current and voltage indicates that below the threshold voltage, only the displacement current flows with 90° phase difference. On the other hand, above the threshold voltage, conduction current in phase with the voltage starts to flow, and with the increase in voltage the phase difference decreases from 90°, indicating that more in-phase conduction current flows. The impedance of the EL device can be estimated from the I-V characteristics. The results are as follows: Below the threshold voltage, the estimated capacitance is equal to the total capacitances of the insulating layer, the phosphor layer, and the insulating layer in series, with the phosphor layer being a capacitor. This indicates that the phosphor layer has high resistivity and acts as an insulator. Above the threshold voltage, the estimated capacitance is equal to the total capacitances of only the insulating layers in series, with the phosphor layer being shunted. This indicates that the phosphor layer loses its insulting characteristics, and electrons at the interface states between the phosphor layer and the insulating layers are injected into the conduction band through a tunneling effect.

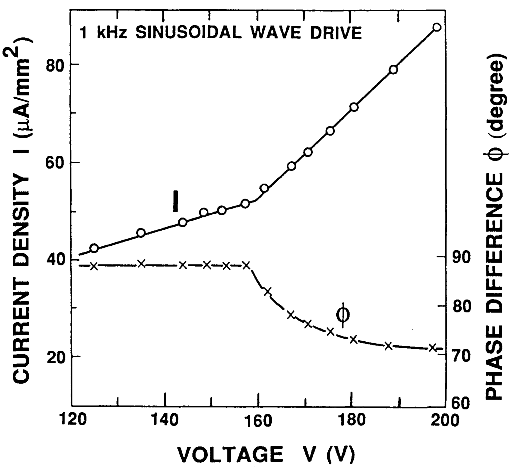

FIG. 8. Current density vs voltage (I-V) and phase difference vs voltage (ϕ-V) characteristics of an ac thin-film EL device with a double-insulating-layer structure.

When the drive voltage becomes larger than the threshold voltage, the average electric field in the phosphor layer becomes (1.3–

1.8) × 10^6 V/cm. Only above this electric field can electrons flow in the phosphor layer. Since EL emission occurs only above the threshold voltage, EL excitation can be considered to be due to this electron flow. It has been verified experimentally (Chen and Krupka, 1972; Smith, 1981; Ono *et al.,* 1989) that the electric field of the phosphor layer is kept at a constant value above the threshold voltage (clamp effect). Therefore, the excess voltage is placed on the insulating layers. This indicates the importance of insulating layers against dielectric breakdown of the EL device.

The above arguments lead to the equivalent-circuit model shown in Fig. 9. Here C_I indicates the effective capacitance of the insulating layers, $C_I = C_{I1}C_{I2}/(C_{I1}+C_{I2})$, with C_{I1} and C_{I2} being the capacitances of the first and second insulating layers, respectively. C_{EL} is the capacitance of the EL phosphor layer, and D_I indicates a bidirectional switching device, which is turned on when the electric field in the phosphor layer becomes larger than the threshold electric field. In an ideal case, this switch can be simulated by back-to-back Zener diodes.

On the basis of the above discussion, the EL emission mechanism is indicated in Fig. 10, which shows an energy-band diagram of the EL device. The EL emission occurs in the following way.

1. Above the threshold voltage, electrons are injected from the interface states between the phosphor layer and insulating layer by high-field–assisted tunneling.
2. The injected electrons are accelerated and gain kinetic energy large enough to excite luminescent centers or the host lattice.
3. High-energy electrons, which are called hot electrons, directly excite luminescent centers through the impact-excitation mechanism. When these electrons in the excited states of the luminescent centers make radiative transitions to the ground state, EL emission is realized.
4. The hot electrons travel through the phosphor layer and are finally trapped at the phosphor–insulator interface states on the anode side, causing polarization.
5. When the polarity of the ac voltage wave forms is reversed, the same process takes place in the opposite direction in the phosphor layer.

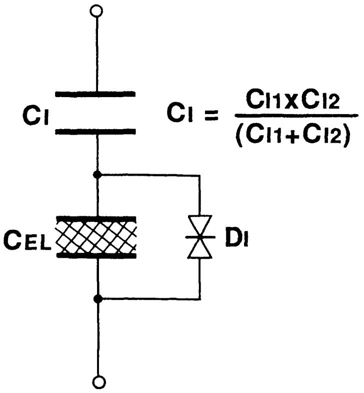

FIG. 9. Equivalent-circuit model of double-insulating-layer type ac thin-film EL device. C_I is the total capacitance of insulating layers in series, with C_{I1} and C_{I2} being capacitances of the first and second insulating layers. C_{EL} is the capacitance of EL phosphor layer, and D_I indicates a bidirectional switch which is on above the threshold electric field.

Since the tunnel injection is temperature insensitive, ac thin-film EL devices can be used over a wide temperature range from −100 to 100 °C.

This mechanism is generally accepted for ZnS-based EL devices (Smith, 1981; Alt, 1984; Okamoto and Miura, 1986; Mikami *et al.,* 1988). In CaS- or SrS-based EL devices, however, an additional mechanism has been proposed (Crandall, 1987a; Tanaka *et al.,* 1988a; Yoshiyama *et al.,* 1989). This additional mechanism is characterized by a field-induced ionization of luminescent centers and subsequent trapping of low-energy electrons, resulting in EL emission.

Since the luminance is approximately proportional to the transferred charge within the phosphor layer (Alt, 1984; Ono *et al.,* 1987), an increase in the transferred charge is necessary to improve luminance. Furthermore, the excitation probability of the luminescent centers by hot electrons should be improved, where the excitation probability is a function of the energy received by hot electrons from the electric field, i.e., a function of the product of the electric field and mean free path. The electric field is limited by the clamp electric field, while the mean free path depends on the crystallinity of the phosphor-layer thin films. From these arguments, an increase in transferred charge density and better crystallinity are necessary to improve luminance.

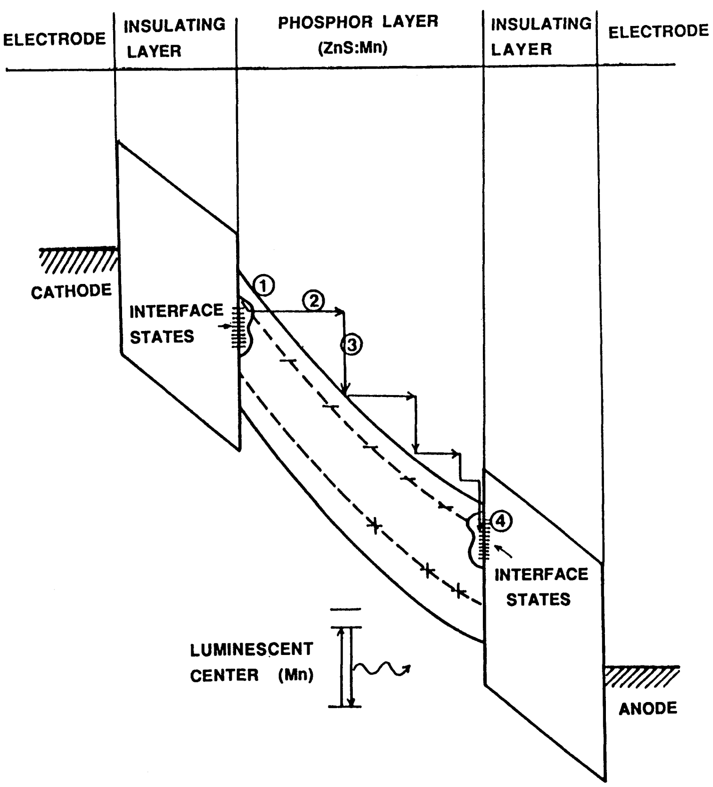
ELECTRODE
INSULATING LAYER
PHOSPHOR LAYER (ZnS:Mn)
INSULATING LAYER
ELECTRODE
CATHODE
INTERFACE STATES
①
②
③
④
INTERFACE STATES
LUMINESCENT CENTER (Mn)
ANODE

FIG. 10. Schematic energy-band diagram of the double-insulating-layer type ac thin-film EL device and EL emission mechanism. Circled numbers refer to features described in the text.

2.2 Electro-Optical Characteristics

Figure 11 shows schematic characteristics of the voltage wave form *V*, the current wave form *I*, and the EL-emission wave form *L* under triangular-wave drive conditions. Below the threshold voltage V_{th}, only a constant displacement current flows and no EL emission results. On the other hand, above V_{th} an additional current flows in the time period when the voltage is larger than V_{th}. This additional current corresponds to an in-phase current that flows through the phosphor layer, causing EL emission, as indicated in the EL-emission wave form *L*. The rise time and decay time of *L* for a ZnS:Mn EL device are of the order of several μs and several ms, respectively. As shown in Fig. 11, there are two EL emissions in one cycle. Therefore, luminance

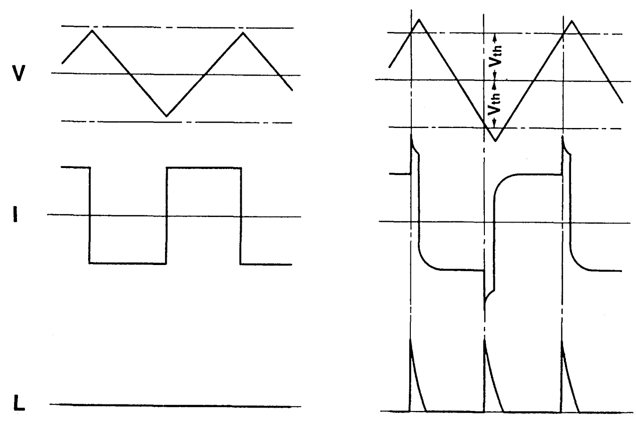

(a) V < Vth

(b) V > Vth

FIG. 11. Schematic characteristic diagrams of voltage wave form *V*, current wave form *I*, and EL emission waveform *L* under triangular-wave drive condition for double-insulating-layer type ac thin-film EL devices. (a) $V < V_{th}$; (b) $V > V_{th}$.

is proportional to the drive frequency up to several kHz.

Figure 12 shows typical luminance vs voltage (L-V) and luminous efficiency vs voltage (η-V) characteristic curves of ZnS:Mn EL devices with the thickness of the phosphor layer being 0.65 μm. Light emission starts when the voltage becomes larger than the threshold voltage V_{th}, and luminance L increases very sharply with the increase in voltage. The threshold voltage is defined by the voltage corresponding to a luminance of 1 cd/m^2. As indicated in the figure, L_{30} is defined by the luminance at 30 V above V_{th}. This value is very important because under practical drive conditions, the drive voltage is usually set at 30 or 40 V above V_{th}. The sharp increase in luminance is due to tunnel injection of electrons from trap states at the phosphor layer/insulating layer interface. On the other hand, the luminous efficiency η increases just above V_{th} and has a maximum in the voltage region where a steep increase of luminance takes place. Above this voltage region, η decreases slowly. To calculate η [lm/W], luminance L and input power density into the EL devices, P_{in} [W/cm^2], must be measured because η is defined by the following equation (Ono *et al.*, 1987):

$$\eta[\mathrm{ln/W}] = (\pi L[\mathrm{cd/m^2}]/P_{in}[\mathrm{W/cm^2}]) \times 10^{-4}.$$

An estimation method of input power density P_{in} is described by Smith (1981), Tiku and Smith (1984), and Ono *et al.* (1987), based on the Sawyer–Tower circuit shown in Fig. 13. Here a sinusoidal-wave drive condition is employed, and C_t and C_s are the total capacitance of the EL device and the capacitance of sense capacitor, respectively. A schematic charge vs voltage (Q-V) characteristic diagram is shown on an oscilloscope, where charge Q is given by the product of C_s and V_y. In order that the partial voltage of the sense capacitor V_y be much smaller than the total voltage V, the condition $C_t/C_s \ll 1$ must be satisfied. Under this condition accurate Q-V characteristic diagrams can be obtained. In Fig. 14, an enlarged Q-V characteristic diagram is shown together with physical quantities: V_{th} is the threshold voltage of the whole EL device, $V_{th,EL}$ is the threshold voltage of the phosphor layer, ΔQ is the transferred charge,

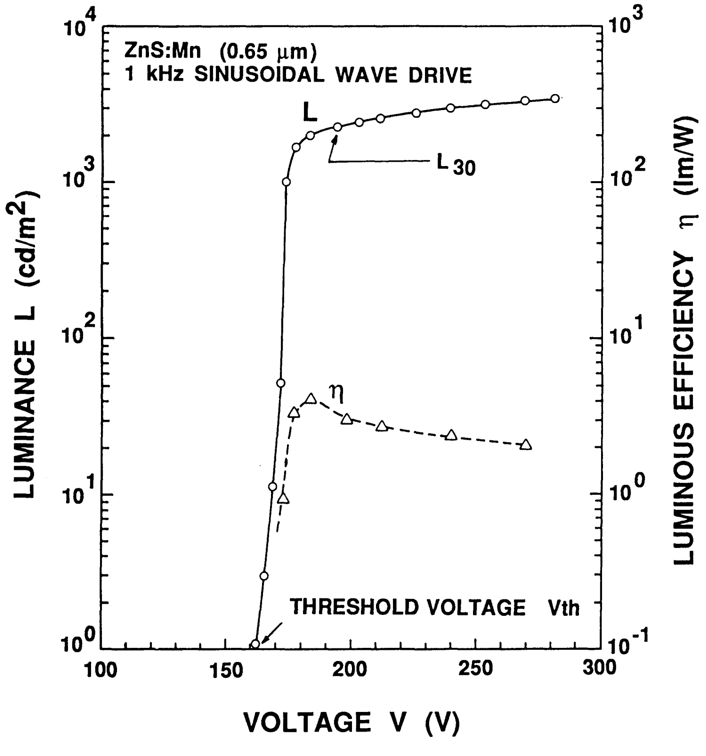

FIG. 12. Typical luminance vs voltage (L-V) and luminous efficiency vs voltage (η-V) characteristic curves of thin-film EL devices and definition of the threshold voltage (after Mikami *et al.*, 1991).

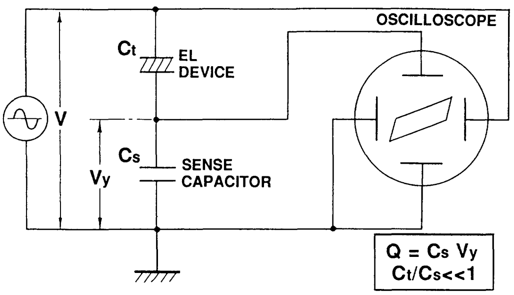

FIG. 13. Charge vs voltage (*Q-V*) characteristic measuring circuit based on Sawyer–Tower circuit (after Ono *et al.*, 1987).

and Q' is the charge remaining in the EL device when V becomes 0 V. Here $V_{th,EL}$ is given by $V_{th}C_I/(C_I+C_{EL})$. The input power density P_{in} is given by the product of drive frequency f and the input energy density per cycle, E_{in}, given by the area encompassed by the Q-V diagram divided by the pixel area A. On the assumption of a parallelogram form for the Q-V diagram, P_{in} is given by

$$P_{in}=fE_{in}=2fV_{th,EL}\Delta Q/A=4fV_{th}Q'/A.$$

As luminance depends on the drive frequency and drive wave forms, it is recommended that the frequency and drive wave form be indicated when luminance values are discussed. Recommended drive wave forms are sinusoidal waves and pulse waves of alternate polarity with frequencies of 50 or 60 Hz or 1 kHz. Furthermore, voltage should be indicated by zero-to-peak values, rather than the rms values, since luminance depends on the peak voltage values.

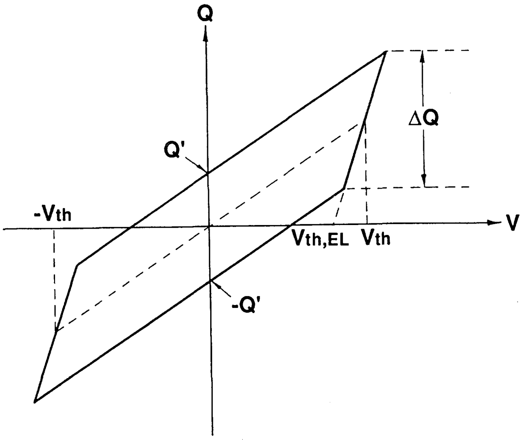

FIG. 14. Charge vs voltage (*Q-V*) characteristic diagram with definitions of physical quantities: V_{th} is the threshold voltage of the EL device, $V_{th,EL}$ is the threshold voltage of the phosphor layer, ΔQ is the transferred charge, and Q' s is the remaining charge in the EL device when V becomes 0.

In regard to increasing the luminance, the light-trapping effect in a ZnS:Mn ac thin-film EL device with a double-insulating-layer structure is discussed. As shown in Fig. 15, a ZnS:Mn phosphor layer with a refractive index n of 2.3 is sandwiched by insulating layers with smaller values of refractive indices, such as Y_2O_3 ($n=1.9$), Al_2O_3 ($n=1.63$), Si_3N_4 ($n=2.0$), or SiO_2 ($n=1.46$). Furthermore, the refractive index of glass is 1.53, larger than the refractive index of air, 1.0. Therefore, a large part of the EL emissions from the phosphor layer do not go through the glass substrate into air; rather they are trapped within the EL device. Various estimates indicate that only 10% of the generated light exits through the front glass surface. Scattering in thin films also affects light output. Increased light output via scattering can be bought at the expense of contrast.

On the other hand, when viewed from the edge of the EL devices, a very bright EL emission can be observed. An application of this edge emission has been proposed (Kun *et al.*, 1986) in the field of electrophotographic printers, i.e., thin-film edge-emitter arrays for the optical image bar in place of a scanning laser with a rotating mirror or LED arrays.

2.3 Physical and Chemical Properties of Phosphor Materials and Luminescent Centers

Phosphor materials of thin-film EL devices are required to have EL emission from the doped luminescent centers and to hold a high

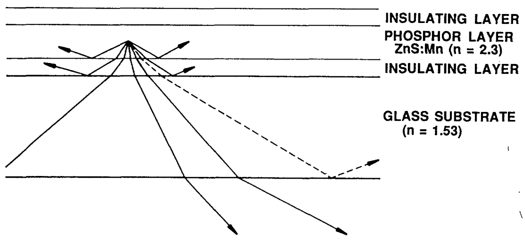

FIG. 15. Schematic representation of light-trap effect in ZnS:Mn ac thin-film EL devices with double-insulating-layer structure. Refractive indices of typical insulating-layer materials are as follows: Y_2O_3, $n=1.9$; Al_2O_3, $n=1.63$; Si_3N_4, $n=2.0$; and SiO_2 $n=1.46$.

electric field of about 10^6 V/cm. For these purposes, wide-gap semiconductors of II-VI compounds are used—IIb-VIb compounds such as ZnS and IIa-VIb compounds such as CaS and SrS. The IIb-VIb compounds have the cubic zinc-blende or hexagonal wurtzite crystal structure. On the other hand, the more ionic IIa-VIb compounds have the rocksalt or sodium chloride structure. Thus alkaline-earth sulfides, such as CaS and SrS, have different properties from ZnS. Table 2 summarizes the material parameters of typical phosphor host materials. The ionic radii of Ca^{2+} and Sr^{2+} are very close to those of rare-earth ions (Table 3). Furthermore, their chemical properties are very similar. Therefore, CaS and SrS are expected to be good candidates for accepting rare-earth ions as luminescent centers. However, since these phosphor materials have hygroscopic characteristics, encapsulation is necessary.

For the luminescent centers, luminescence of the donor–acceptor type is not plausible because a very high electric field is necessary to obtain EL emission. Therefore only isolated luminescent centers are suitable. As indicated in Sec. 2.1, EL emission results from impact excitation of luminescent centers by high-energy electrons accelerated by a high electric field. Because of this, luminescent centers weakly coupled to host materials, i.e., isolated luminescent centers, are expected to have high luminous efficiency. From these discussions, transition-metal ions, such as Mn, and rare-earth ions, such as Tb, Sm, Tm, Eu, and Ce, are found to be suitable luminescent centers.

The rare-earth ions have an open $4f$ shell, which is screened by the outer electrons in the $5s^2 5p^6$ closed shells. Therefore the $4f$ electrons of rare-earth ions in the host materials are not greatly influenced by crystal fields. Two different types of luminescent transitions can be observed: a parity-forbidden f-f intrashell transition, and the parity-allowed f-d transition. Examples of EL emission due to the f-f transition are the green emission from Tb^{3+}, red emission from Sm^{3+}, blue emission from Tm^{3+}, and white emission from Pr^{3+}. Eu^{2+} and Ce^{3+} produce EL emission due to f-d transitions. The $5d$ electrons are easily affected by the crystal field. Therefore the excited energy levels depend on the host material, and the EL emission color can be varied by changing it.

To dope with transition metals or rare-earth elements effectively, matching of chemical properties and ionic radii between the cations of the host phosphor materials and the dopant luminescent centers is very important. Table 3 lists the ionic radii of Zn^{2+}, Ca^{2+}, Sr^{2+}, and rare-earth luminescent-center ions. Since rare-earth ions have different valencies and ionic radii from those of Zn^{2+}, it is difficult to dope them uniformly into ZnS hosts at high concentration. To overcome this difficulty, doping methods featuring addition of charge compensators (Tanaka *et al.*, 1987a) or molecular luminescent centers (Chase *et al.*, 1969) have been proposed, and fluorides or sulfides of these luminescent centers have been used.

3. MATERIALS REQUIREMENTS

In this section, the requirements of materials used in thin-film EL devices with double-insulating-layer structure are discussed in terms of optical, electrical, and physical characteristics.

Table 2. Material parameters of phosphor host materials.

Items	ZnS	CaS	SrS
Melting point (°C)	1000–1900	2400	>2000
Band gap (eV)	3.6	4.4	4.3
Type of transition	Direct	Indirect	Indirect
Refractive index	2.3	—	2.1
Relative dielectric constant	8.3	9.3	9.4
Crystal structure	Zinc blende	NaCl type	NaCl type
Lattice constant (Å)	5.409	5.697	6.019
Ionic radius of cation (Å)	0.74	0.99	1.13
Ionicity	0.623	>0.785	>0.785

Table 3. Ionic radii of Zn^{2+}, Ca^{2+}, Sr^{2+}, and rare-earth luminescent-center ions.

Ion	Ionic radius (Å)
Zn^{2+}	0.74
Ca^{2+}	0.99
Sr^{2+}	1.13
Ce^{3+}	1.034
Pr^{3+}	1.013
Sm^{3+}	0.964
Eu^{2+}	1.09
Eu^{3+}	0.950
Tb^{3+}	0.923
Tm^{3+}	0.869
Mn^{2+}	0.80

3.1 Glass Substrates

The characteristics and surface conditions of the glass substrates control the properties and reliability of EL devices. During EL fabrication, severe processing conditions are used, such as high-temperature (~550 °C) annealing and immersion in acid or alkaline fluids, and so the glass substrates must be heat resistant and acid and alkali resistant. Furthermore, diffusion of alkali-metal ions from the glass into the phosphor layer must be prevented, because they deteriorate the EL characteristics. This has led to nonalkaline phosphosilicate glasses being used for EL devices. A final requirement is that the glass surface be as smooth as that for photomasking.

Physical properties required of the glass substrates are as follows:

1. a high softening temperature, i.e., above 650 °C;
2. low or no alkali-metal content;
3. a high electrical resistivity; and
4. an expansion coefficient that matches that of the deposited films.

Much of the thin-film EL work done to date has used Corning 7059 glass; it is a nonalkaline phosphosilicate glass with high thermal resistivity (Faulkner *et al.,* 1983). The softening temperature of this glass is, however, 598 °C, and glass substrates with higher softening temperature are desired.

3.2 Transparent Conducting Films

Large conductivity values and high transmittance are necessary for transparent conducting films for EL devices, and indium tin oxide (ITO: In_2O_3–10 wt % SnO_2) is best suited. Because of the good characteristics of the thin films obtained and the high production yield, electron-beam deposition (EBD) and sputtering methods are employed to deposit ITO films (Shimizu and Matsudaira, 1985).

A small electrical resistance is required for EL devices. Presently the sheet resistance of typical ITO thin films is about 10–20 Ω/□. The sheet resistance can be decreased by increasing the film thickness, which in turn increases the slope of the etching edge of the ITO stripe electrodes, causing a decrease in the breakdown electric field of the first insulating layer deposited just above the ITO electrode. From experiments it has been found that the film thickness should be less than 150 nm. Then the resistivity must be smaller than 10^{-4} Ω cm, which is the smallest value obtainable at present.

3.3 Insulating Layers

The most important role of the insulating layers is to protect the phosphor layers from electric breakdown. Additionally they must prevent metal-ion diffusion into the phosphor layer and provide interface states at the boundary of the phosphor layer and the insulating layer. From these interface states, electrons are injected into the conduction band of the phosphor layer through tunneling when the applied electric field is larger than the threshold electric field.

For these purposes, the following properties are required (Howard, 1977; Alt, 1984; Tiku and Smith, 1984):

1. high breakdown electric field (E_{BD});
2. high dielectric constant ($\epsilon_0\ \epsilon_r$);
3. small number of pinholes;
4. small tan δ; and
5. good adhesiveness.

Here ϵ_0 and ϵ_r are the dielectric constant of a vacuum and relative dielectric constant, respectively; tan δ is a measure of dissipative characteristics defined by $1/2\pi fCR$, where f is the drive frequency, C is the capacitance, and R is the resistance.

There are two groups of dielectric materials used in EL devices: (a) amorphous oxides and nitrides, such as Al_2O_3, SiO_2, Y_2O_3, TiO_2, Si_3N_4, and Ta_2O_5; and (b) ferroelectric materials, such as $BaTiO_3$ and $PbTiO_3$. Group-(a) dielectric materials have a small dielectric constant and high breakdown electric field. On the other hand, group-(b) dielectric materials have a large dielectric constant and low breakdown electric field. To characterize dielectric materials for insulating layers, a figure of merit is introduced, which is the product of dielectric constant and breakdown electric field, $\epsilon_0\epsilon_r E_{BD}$. This figure of merit is equal to the maximum trapped charge density Q_{max} at the phosphor layer/insulating layer interface. Table 4 lists figures of merit for typical insulating-layer materials and also relative dielectric constants ϵ_r and breakdown electric fields E_{BD} (Fujita *et al.*, 1984). The numbers should be taken as typical values because there are variations, depending on the deposition conditions and methods. Figure 16 relates ϵ_r and E_{BD} (Fujita *et al.*, 1984).

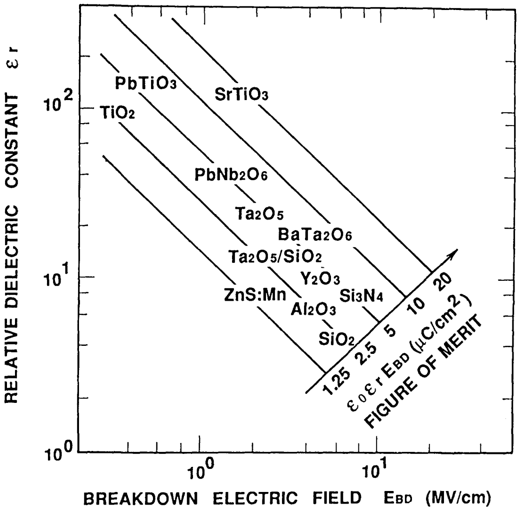

FIG. 16. Relative dielectric constant vs breakdown electric field (ϵ_r–E_{BD}) characteristics of typical dielectric-layer materials for thin-film electroluminescent devices. Straight lines are figures of merit defined by $\epsilon_0\epsilon_r E_{BD}$ in units of $\mu C/cm^2$ (after Fujita *et al.*, 1984).

One more important characteristic for EL devices is the electric breakdown mode, which may be a propagating breakdown mode (PB) or a self-healing breakdown mode (SB). In the former mode, electric breakdown, once occurring somewhere within an insulating layer, propagates to neighboring places, resulting in a large-area electric breakdown. On the other hand, in the latter mode, electric breakdown does not propagate and stays at the initial place, resulting in a small-area electric breakdown. These modes are also indicated in Table 4. Insulating thin films for EL devices must have a self-healing breakdown mode. Generally, dielectric thin films with large dielectric constants or not fully oxidized or not fully nitrified dielectric thin films tend to have a propagating breakdown mode. Coupled with other dielectric thin films, insulating layers with the propagating break-

Table 4. Relative dielectric constant ϵ_r, breakdown electric field E_{BD}, figure of merit defined by the product $\epsilon_0\epsilon_r E_{BD}$ (equal to the maximum trapped charge density), and breakdown mode of typical insulating-layer materials (after Fujita *et al.*, 1984).

Material	Deposition method	ϵ_r	E_{BD} (MV/cm)	$\epsilon_0\epsilon_r E_{BD}$ ($\mu C/cm^2$)	Breakdown mode[a]
Y_2O_3	EB	12	3–5	3–5	SB
Sm_2O_3	EB	15	2–4	3–6	SB
Al_2O_3	Sputtering	8	5	4	SB
SiO_2	Sputtering	4	6	2	SB
Si_3N_4	Sputtering	8	6–8	4–6	SB
Ta_2O_5	Sputtering	23–25	1.5–3	4–6	SB
SiAlON	Sputtering	8	8–9	5–6	SB
$PbTiO_3$	Sputtering	150	0.5	7	PB
$BaTa_2O_6$	Sputtering	22	3.5	7	SB
$PbNb_2O_6$	Sputtering	41	1.5	6	SB
$SrTiO_3$	Sputtering	140	1.5–2	19–25	PB
$Sr(Zr,Ti)O_3$	Sputtering	100	3	27	PB

---

[a] SB: self-healing breakdown mode; PB: propagating breakdown mode.

down mode can be converted into the self-healing type.

Originally, oxide thin films such as Y_2O_3, Al_2O_3, Ta_2O_5, and SiO_2 were used as EL insulating layers (Okamoto *et al.*, 1981; Kozawaguchi *et al.*, 1982; Tiku and Smith, 1984), but they tended to have deficiencies such as too much porosity and a low breakdown electric field. Furthermore, some of these oxide thin films tended to crystallize and cause nonuniformity in the phosphor thin films deposited on top of them, resulting in nonuniform luminance. To overcome these difficulties, amorphous nitride thin films, such as Si_3N_4, were introduced (Takeda *et al.*, 1981).

Recently, insulating-layer materials with a large figure of merit, such as $SrTiO_3$ and $Ba_2Ta_2O_6$, have been introduced to reduce the drive voltage to approximately 100 V (Fujita *et al.*, 1984). Using $Ba_2Ta_2O_6$ thin films ($\epsilon_r = 22$) for the first insulating layer and $SrTiO_3$ thin films ($\epsilon_r = 140$) for the second one, highly reliable EL devices have been obtained.

3.4 Phosphor Layers

The requirements for the phosphor layers are as follows:

1. Their host materials can be excited by a high electric field, on the order of 10^6 V/cm, to cause electron avalanche.
2. They can maintain a high electric field without electric breakdown and they must have insulating characteristics below the threshold voltage.
3. They can accept suitable luminescent centers that emit light in visible colors.

Phosphor materials satisfying the above requirements are limited to II-VI compounds: IIb-VIb compounds such as ZnS and ZnSe, and IIa-VIb compounds such as CaS, SrS, and SrSe. Detailed discussions of them have already been given in Sec. 2.3.

3.5 Rear Electrodes

The requirements for the rear-electrode materials are as follows:

1. good adhesiveness to the second insulating layer;
2. no metal-ion migration at high electric field;
3. an ability to prevent breakdown spread when dielectric breakdown of the phosphor layer or insulating layers occurs; and
4. low resistivity.

Among the materials satisfying these conditions, Al is considered to be the best. Deposition of Al thin films can be done by thermal heating, electron-beam deposition (EBD), and sputtering methods. One disadvantage in their use is their high reflectivity of light, so that mirror-type reflection occurs. This reflection can be suppressed by a smoke filter of 50%–70% or by circularly polarizing filters. Since Al films have specular reflection, the latter works very well. Then no incident light going into the EL cell can come out from the glass plate and only the EL emissions are effectively taken out from the EL devices.

4. MATERIAL DEPOSITION METHODS

In this section deposition methods used in fabricating phosphor layers of ZnS-based thin-film EL devices (see THIN-FILM DEPOSITION), especially ZnS:Mn EL devices, are discussed, with emphasis on the deposition conditions, crystal structures, and their relations to electro-optical characteristics (Vecht, 1989; Ono, 1990).

To increase the luminance of EL displays, better crystallinity of the phosphor layers is necessary, and a large number of deposition methods have been studied. These deposition methods are divided into two groups: physical vapor deposition (PVD) methods and chemical vapor deposition (CVD) methods [see CVD (CHEMICAL VAPOR DEPOSITION)]. The former include electron-beam deposition (EBD) (Sasakura *et al.*, 1981), multisource deposition (MSD) (Nire *et al.*, 1989), and sputtering (Miura *et al.*, 1983; Ohnishi *et al.*, 1987) methods. The latter include atomic-layer epitaxy (ALE) (Suntola *et al.*, 1980; Törnqvist *et al.*, 1983; Leskelä, 1989), metal-organic chemical vapor deposition (MOCVD) (Cattel *et al.*, 1983; Hirabayashi and Kozawaguchi, 1986; Migita *et al.*, 1988), and low-pressure hydride-transport chemical vapor deposition (HT-CVD) with halogen transport of Mn (Mikami *et al.*, 1991). Among these, the

EBD and ALE methods are now used to fabricate commercially available EL displays.

4.1 PVD Methods

4.1.1 EBD Method. The EBD method is one of the flash evaporation methods suitable for obtaining stoichiometry and reproducibility of the film, since both ZnS and Mn are sublimate materials. The substrate temperature is set at approximately 200 °C. Thermal annealing is performed in vacuum at approximately 550 °C after deposition of the ZnS:Mn thin film. Thermal annealing improves the phosphor layers because it removes strains in the films, enhances crystallinity of the grains, and accelerates Mn diffusion.

For CaS- and SrS-based EL devices, the EBD method is mainly used (Tanaka *et al.*, 1987a).

4.1.2 MSD Method. In this method, evaporation sources for Zn, S, and Mn are placed in a vacuum chamber having a pressure of 10^{-5} to 10^{-6} Torr. By independently controlling the temperatures of these sources, highly crystalline and stoichiometric ZnS:Mn films can be obtained (Nire *et al.*, 1989).

From a cross-sectional transmission electron microscopic (TEM) picture of deposited ZnS:Mn film on a glass substrate, it was found that columnar polycrystals of ZnS extended from the bottom to the top surface. The grain size of these columns was approximately 100 to 200 nm. X-ray diffraction patterns and transmission electron diffraction measurements showed that the crystal structure was hexagonal (wurtzite) with a (002) preferred orientation. The luminance L_{80} was 5300 cd/m^2 and the luminous efficiency η was 3.18 lm/W under a 5-kHz sinusoidal excitation.

SrS:Ce thin-film fabrication has been carried out with MSD, and good crystallinity was obtained (Tanda *et al.*, 1988).

4.1.3 Sputtering Method. The sputtering method (see SPUTTERING) has high deposition rates and is suitable for in-line processes, because both the insulating layers and the phosphor layer can be grown by the same deposition method (Miura *et al.*, 1983). In the case of ZnS:Mn, however, the crystallite diameter of the deposited films is not large, and stoichiometric ZnS films are difficult to obtain (Matsuoka *et al.*, 1988). As a result, a bright ZnS:Mn EL device has not been realized yet. On the other hand, for ZnS:Tb,F, the brightest green-emitting EL devices come from the sputtering method (Ohnishi *et al.*, 1987).

4.2 CVD Methods

4.2.1 ALE Method. The ALE method is an epitaxial deposition method in which the reactants are alternately supplied to the substrate (Suntola *et al.*, 1980; Leskelä, 1989). As shown in Fig. 17, the ALE growth process for ZnS is based on the exchange reaction $ZnCl_2(ads) + H_2S(g) \rightarrow ZnS(s) + 2HCl(g)$. On introduction of $ZnCl_2$ vapor, a single layer of $ZnCl_2$ is formed on a surface covered by oxygen or sulfur, by controlling the substrate temperature. Excess $ZnCl_2$ vapor is flushed away and then H_2S is introduced. Through

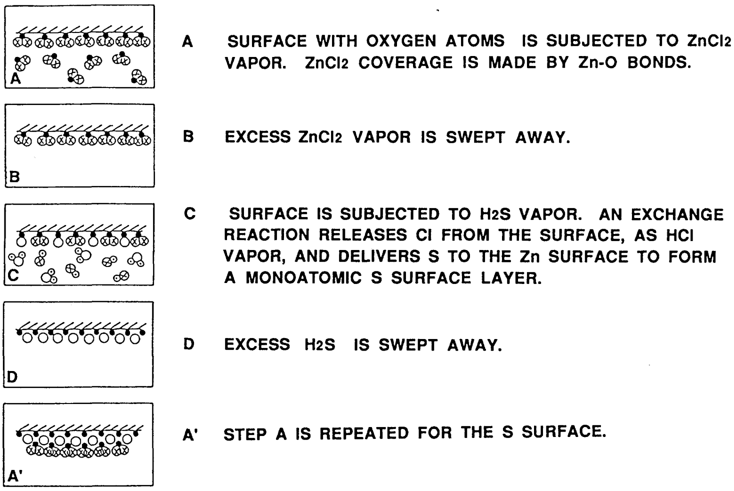

A SURFACE WITH OXYGEN ATOMS IS SUBJECTED TO ZnCl2 VAPOR. ZnCl2 COVERAGE IS MADE BY Zn-O BONDS.

B EXCESS ZnCl2 VAPOR IS SWEPT AWAY.

C SURFACE IS SUBJECTED TO H2S VAPOR. AN EXCHANGE REACTION RELEASES Cl FROM THE SURFACE, AS HCl VAPOR, AND DELIVERS S TO THE Zn SURFACE TO FORM A MONOATOMIC S SURFACE LAYER.

D EXCESS H2S IS SWEPT AWAY.

A' STEP A IS REPEATED FOR THE S SURFACE.

FIG. 17. Growth mechanism of the atomic-layer epitaxy (ALE) process for ZnS using $ZnCl_2$ and H_2S as reactants. The substrate temperature must be high enough to prevent condensation of the reactants (after Suntola *et al.*, 1980).

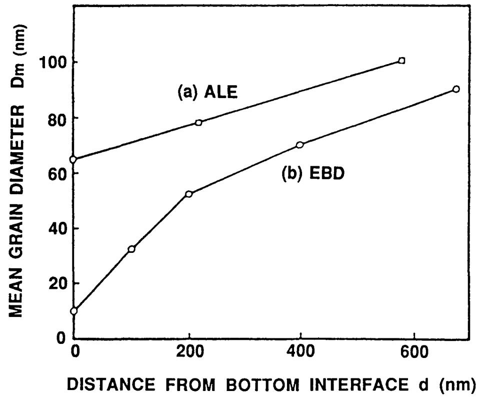

FIG. 18. Mean grain diameter of polycrystalline ZnS:Mn phosphor layer vs the distance from the bottom interface. (a) ALE and (b) EBD processes (after Theis *et al.*, 1983).

reaction of H_2S with the adsorbed $ZnCl_2$, ZnS is formed and HCl is liberated. After flushing away excess HCl, the cycle is repeated. The growth is insensitive to pressure and substrate materials, but the substrate temperature is important.

From x-ray measurements of ZnS films (Tanninen *et al.*, 1983) it was found that thin films grown at 350 °C have the cubic (zinc-blende) crystal structure, and films grown at 500 °C have the hexagonal (wurtzite) structure. Figure 18 shows the mean grain diameter parallel to the film surface vs the distance from the bottom interface (Theis *et al.*, 1983). In the EBD films small grains are present in the so-called dead layer at the beginning of deposition, while in the ALE films, large grains are formed from the beginning. The average luminous efficiency is 3 lm/W, with the maximum being 8 lm/W.

This method has been successfully used to deposit SrS- and CaS-based EL devices (Leskelä, 1989).

4.2.2 MOCVD Method.

The MOCVD method offers several advantages, such as lower growth temperature and easy control of impurity concentration. For Zn, dimethylzinc (DMZ) is used, and for sulfur, H_2S with 5% hydrogen or diethylsulfide (DES) is used. Hydrogen serves as a carrier gas for the reactants. For Mn, tricarbonylmethylcyclopentadienyl-Mn (TCM) is used. X-ray diffraction analysis revealed that EBD films have the zinc-blende crystal structure, while the MOCVD films are wurtzite crystal.

Recently, new Mn sources, di-π-cyclopentadienyl-Mn(CPM) and bis-methylcyclopentadienyl-Mn (BCPM), were studied and both were efficiently decomposed at the optimum growth temperature of ZnS, 280–350 °C (Migita *et al.*, 1988). Figure 19 shows crystallite size vs film thickness of undoped ZnS films prepared by the MOCVD and EBD methods. The former leads to large grains and good crystallinity from the start of growth. The EL devices with CPM have the largest luminance. The saturation luminance L_{sat} is 4300 cd/m^2, and the maximum luminous efficiency is 4.8 lm/W under 1-kHz sinusoidal drive conditions.

Color EL devices based on a ZnS phosphor layer have been fabricated by the MOCVD method; use of $SmCl_3$ as a source material of luminescent centers gave a much brighter red EL emission than SmF_3 (Hirabayashi *et al.*, 1987).

4.2.3 HT-CVD Method.

Good-quality ZnS:Mn phosphor layers can be prepared by using a low-pressure hydride-transport CVD system with halogen transport of Mn and hydrogen transport of ZnS. This system with a horizontal reactor is schematically shown in Fig. 20 (Mikami *et al.*, 1991). The component elements, Zn and S, are transported as vapor and/or gaseous hydride by passing H_2 over the ZnS powder source. The Mn is transported as a gaseous halide by passing HCl over the Mn source. These gases are brought together

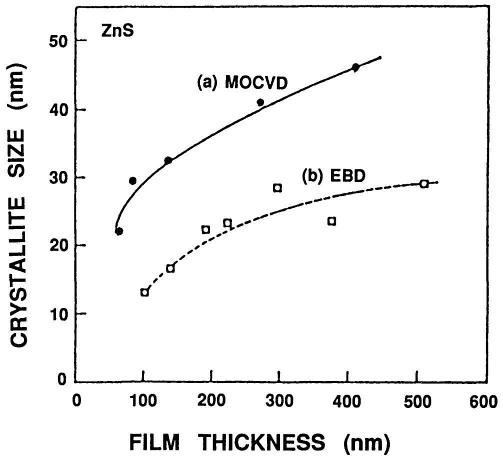

FIG. 19. Crystallite size vs film thickness of undoped ZnS thin films. (a) MOCVD and (b) EBD processes (after Migita *et al.*, 1988).

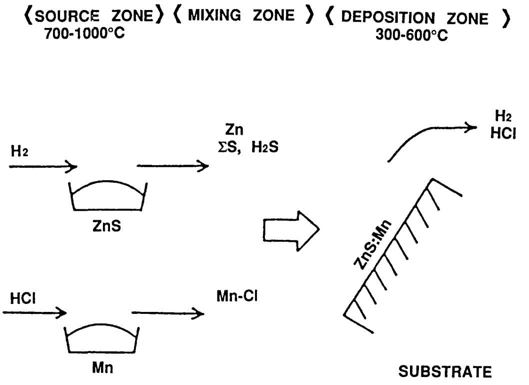

FIG. 20. Growth process of low-pressure hydride-transport CVD (HT-CVD) system with halogen transport of Mn (after Mikami *et al.*, 1991).

in the deposition zone, where the substrates have been placed on a quartz holder.

From the results of x-ray diffraction and reflection high-energy electron diffraction (RHEED) analysis, the following results have been obtained. The undoped ZnS films have the cubic (zinc-blende) crystal structure with a strong (111) orientation. The Mn-doped ZnS films have hexagonal (wurtzite) crystals with a (001) preferred orientation.

Figure 21 shows luminance vs voltage (*L-V*) characteristic curves under a 1-kHz sinusoidal excitation. In the inset, film-thickness dependences of luminance L_{50} and luminous efficiency η are shown. EL devices with film thickness as small as 0.26 μm show steep *L-V* characteristics, and the maximum luminous efficiency is 4.4 lm/W in an EL device with film thickness of 0.5 μm.

4.3 Summary of Deposition Methods

Table 5 summarizes deposition conditions of phosphor layers and luminescent characteristics of ZnS:Mn EL devices. The EL-emitted color differs slightly depending on the crystal structure, and luminous efficiency is larger in the CVD-prepared EL devices.

5. COLOR THIN-FILM ELECTROLUMINESCENT DEVICES

As described in Sec. 2.3, II-VI compounds are used mostly as EL phosphor materials. In the following, recent developments of color EL devices based on ZnS and alkaline-earth sulfides, such as CaS and SrS, are discussed in detail.

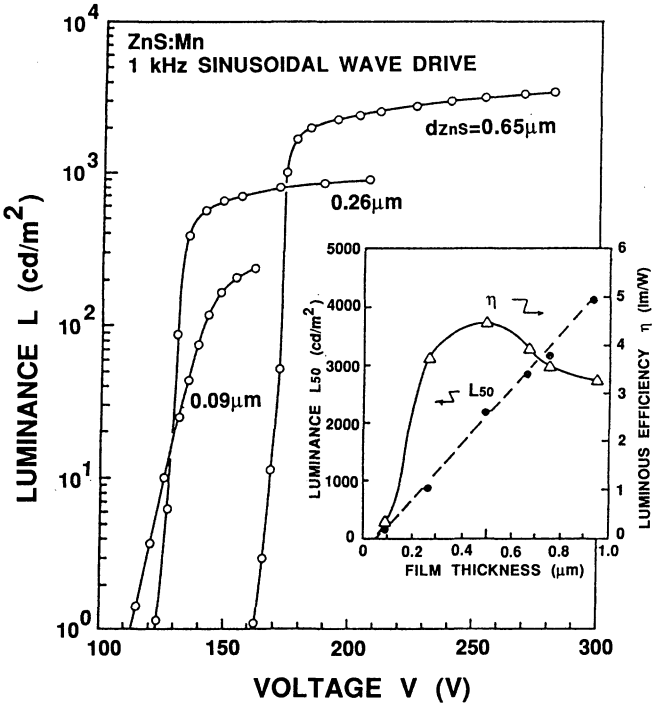

FIG. 21. Luminance vs voltage (*L-V*) characteristics and film-thickness dependence of luminance L_{50} and the luminous efficiency of an EL device fabricated by the HT-CVD method (after Mikami *et al.*, 1991).

Table 5. Comparison of deposition methods used for ZnS:Mn EL devices.

Conditions and characteristics	Physical vapor deposition			Chemical vapor deposition		
	Sputtering	EBD	MSD	MOCVD	HT-CVD	ALE
Deposition rate (Å/min)	>100	>1000	~300	>100	>100	~100
Substrate temperature (°C)	200	200	100–300	300–500	500	500
Crystal structure of ZnS[a]	z.b.	z.b.	wur.	wur.;z.b.	wur.	wur.
Emission wavelength (nm)	585	585	580	580;585	580	580
Luminance (cd/m^2)[b]	>1500	>3000	>3000	>3000	>3000	>3000
Luminous efficiency (lm/W)	~2	~3	~3	>4	>4	3–8

[a] wur.: wurtzite; z.b.: zinc-blende.
[b] 1-kHz excitation.

5.1 Color Thin-Film EL Devices Based on ZnS Phosphors

5.1.1 Orange-Yellow (Amber) Emitting ZnS:Mn. The ZnS:Mn EL device shows the highest luminance and luminous efficiency (Table 6). This can be explained by the following discussion. In this device, Mn^{2+} occupies the lattice site of Zn, having the same valence as the dopant. This valence matching leads to high solid solubility without using charge compensators. Furthermore, as shown in Table 3, the ionic radius of Mn^{2+} is very close to that of Zn^{2+} (only 8% difference), so that it is not difficult for Mn^{2+} to replace Zn^{2+}. These conditions lead to high luminance and high luminous efficiency. The $3d^5$ energy-level configurations of Mn^{2+} are the ^6S ground state and the ^4G first excited state. As shown in Fig. 22(a), a broad orange-yellow (amber) emission with a peak around 585 nm for cubic ZnS:Mn is realized when the transition from ^4G to ^6S occurs. This *d-d* electron transition is partially allowed under perturbation of the crystal field around Mn^{2+}. The CIE (Commission Internationale de l'Eclairage) color coordinates of this emission are $x=0.50$ and $y=0.50$.

Since the emission spectrum of ZnS:Mn is very wide, ranging from green to red, a bright red emission can be obtained by using a red filter. Recently, Tuenge and Kane (1991) reported a bright red EL emission by using an inorganic thin-film, long-wavelength pass filter made of cadmium sulfoselenide. They obtained a luminance of 68.5 cd/m^2 [20 foot-lamberts (ft. L)] under a 60-Hz pulse-wave drive. The color gamut is close to that of the Y_2O_2S:Eu color-TV red phosphor, and the CIE coordinates are $x=0.65$ and $y=0.35$.

5.1.2 Multicolor-Emitting ZnS:RE,F (RE: Rare-Earth Elements). Rare-earth ions have been the most important luminescent centers for multicolor electroluminescent devices since Lumocen devices were reported (Kahng, 1968; Chase *et al.*, 1969). Most rare-earth–doped ZnS EL devices show EL emissions in the visible-light region (Yoshida *et al.*, 1980; Kobayashi *et al.*, 1985; Hamakawa *et al.*, 1988). The valences of rare-earth ions are usually +3, so that charge compensation is necessary. As shown in Fig. 22(b), the EL emission spectra consist of line spectra due to (4*f*) inner-shell transitions. For the three primary-color rep-

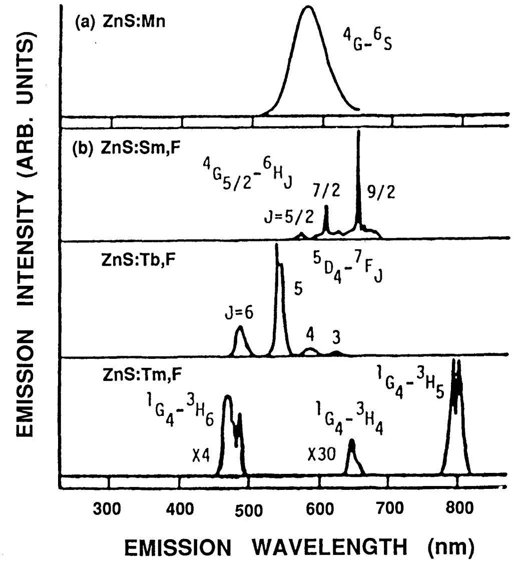

FIG. 22. Emission spectra of ZnS EL devices. (a) ZnS:Mn; (b) ZnS:RE,F (RE: rare-earth elements).

Table 6. Luminance and luminous efficiency of color thin-film EL devices.

Phosphor layer material	Emission color	CIE color coordinates		Luminance L (cd/m^2)		Luminous efficiency
		x	y	1 kHz	60 Hz	η (lm/W)(1 kHz)
ZnS:Mn	Orange-yellow	0.50	0.50	5000	200	2–8
ZnS:Sm,F	Orange red	0.60	0.38	120	8	0.05
ZnS:Sm,Cl	Red	0.64	0.35	200	12	0.08
CaS:Eu	Red	0.68	0.31	200	12	0.05
ZnS:Mn/filter	Red	0.65	0.35	1100	65	0.8
ZnS:Tb,F	Green	0.28	0.62	1500	100	0.5–1
CaS:Ce	Green	0.27	0.52	150	10	0.1
ZnS:Tm,F	Blue	0.11	0.09	2	0.15	<0.01
SrS:Ce	Blue-green	0.19	0.38	1150	70	0.4
SrS:Ce/filter	Blue	0.10	0.19	200	12	0.07

resentations, ZnS:Sm,F is used for red, ZnS:Tb,F is used for green, and ZnS:Tm,F is used for blue. Table 6 lists the maximum luminances and luminous efficiencies of these EL devices.

The ZnS:Tb,F green-emitting EL devices are second in luminance after ZnS:Mn (Ohnishi *et al.*, 1987). Sputtering is usually used to obtain high luminance because of uniform doping of luminescent centers into the ZnS host is possible (Miura *et al.*, 1983). The Tb^{3+} ion has an ionic radius of 0.923 Å, which is larger than that of Zn^{2+} ion by 31% (Table 3). This size mismatch makes thermal diffusion of Tb ions into the ZnS lattice difficult. In this respect, the sputtering method is more effective than the EBD method with subsequent annealing (Mita *et al.*, 1986).

The charge-compensation problem has been studied extensively in order to increase luminance, and the role of F in ZnS:Tb,F has received particular attention (Ogura *et al.*, 1986; Okamoto and Watanabe, 1986; Mikami *et al.*, 1987). In these experiments, the luminance of sputtered ZnS:Tb,F was studied as a function of the luminescent-center concentration in the sputtering targets and the sputtered phosphor films. It was found that not only the Tb concentration but also the F concentration is important to obtain high luminance. As shown in Fig. 23, the highest luminance is obtained when the [F]/[Tb] ratio is close to unity. This is reasonable considering that a Tb^{3+} ion replaces a Zn^{2+} ion and an F^- ion acts as a charge compensator for the excess of +1 charge. Furthermore, recent studies showed that TbOF-complex luminescent centers are more effective in increasing luminance (Okamoto *et al.*, 1989a). Prototypes of green-emitting EL devices have been fabricated (Okamoto *et al.*, 1989b).

In order to increase the luminance of red-emitting ZnS:Sm,F EL devices, the ionic radius difference and valence difference must also be taken into account. Since the emission color is orange-red and the luminance is low,

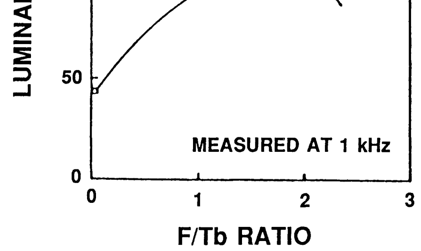

FIG. 23. [F]/[Tb] dependence of luminance in ZnS:Tb,F devices with T_a being the annealing temperature (1ft L=3.43 cd/m^2) (after Ogura *et al.*, 1986).

ZnS:Sm,F is not a good candidate for a red EL phosphor (Kobayashi *et al.*, 1985). Color purity and luminance improvements have been reported from replacing SmF_3 by $SmCl_3$ (Hirabayashi *et al.*, 1987).

For a blue EL phosphor, only ZnS:Tm,F has been reported (Kobayashi *et al.*, 1985). However, the luminance level is very low, because the ionic radius and valence matching between Zn^{2+} and Tm^{3+} is not very good. Furthermore, there is a very strong infrared emission at 800 nm [Fig. 22(b)]. Therefore ZnS:Tm,F is not a satisfactory blue phosphor, although the emission color is pure blue.

5.2 Color Thin-Film EL Devices Based on CaS and SrS Phosphors

Alkaline-earth sulfides, such as CaS and SrS, are known as dc powder multicolor EL phosphors (Higton *et al.*, 1978; Vecht, 1980). The application to thin-film EL devices, however, has not been easy because of the difficulty in obtaining good-quality thin films. With recent development in thin-film fabrication technologies, it is now possible to obtain red and blue EL devices with high luminance (Barrow *et al.*, 1984; Tanaka *et al.*, 1985, 1987a).

By doping rare-earth elements as luminescent centers into CaS and SrS, primary-color EL emissions can be obtained. The EL emission spectra are shown in Fig. 24, and the maximum luminances and luminous efficiencies are given in Table 6. Since EL emissions from Eu^{2+} and Ce^{3+} luminescent centers are due to *f-d* allowed transitions, the emission spectra strongly depend on host phosphor materials.

5.2.1 Red-Emitting CaS:Eu. As shown in Fig. 24(a), EL emission from CaS:Eu is deep red with an emission peak around 650 nm. The CIE coordinates are $x=0.68$ and $y=0.31$. The red purity of CaS:Eu is better than that of ZnS:Sm,F. Since Eu^{2+} is a stable ion and the ionic radius difference from Ca^{2+} is only 10.1%, Eu can easily be doped into a CaS host. These facts make CaS:Eu a promising candidate for a red EL phosphor.

Improvements of luminance and emission response characteristics have been extensively studied. An increase in luminance was achieved by increasing the substrate temperature to 400–500 °C (Tanaka *et al.*, 1987a; Ono *et al.*, 1989). This improvement was found to be related to crystalline improvement; e.g., in CaS:Eu EL, an increase in temperature changed the crystalline alignment plane from (200) to (220) to (111). This change was attributed to the density of dangling bonds on the alignment plane and a reevaporation effect of the deposited atoms (Yoshiyama *et al.*, 1988).

In electron-beam (EB) deposited CaS:Eu devices, emission and charge density responses were found to be slow (Crandall, 1987b), which must be remedied before CaS:Eu can be used commercially as a red phosphor. Recently, it was found that this slow response could be overcome without changing the color gamut by coactivating CaS:Eu with a small amount (0.05 mol %) of Ce (Ando and Ono, 1989).

5.2.2 Green-Emitting CaS:Ce and Blue-Green–Emitting SrS:Ce. As shown in Fig. 24(b), the EL emission from CaS:Ce is green, and the spectrum is broad and has peaks at 505 and 570 nm. The luminance of CaS:Ce is only 10% that of ZnS:Tb,F and therefore, insufficient.

As shown in Fig. 24(c), the EL emission from SrS:Ce is blue-green and the spectrum is broad and has peaks at 475 nm (blue) and 530 nm (green). This spectrum makes it possible to obtain only blue EL emission with good color purity by using a blue filter. In this case

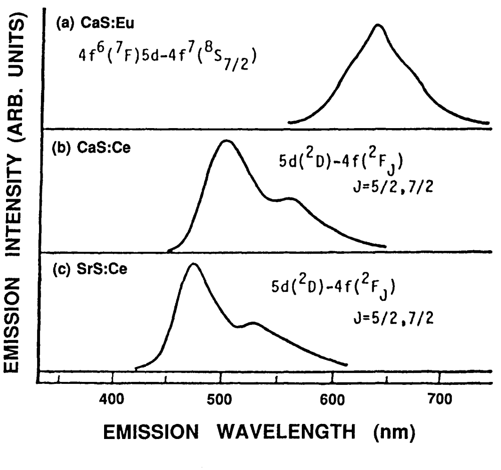

FIG. 24. Emission spectra of CaS and SrS EL devices. (a) CaS:Eu; (b) CaS:Ce; (c) SrS:Ce.

the luminance is reduced to approximately $\frac{1}{5}$ the original luminance with blue-green emission (Ono *et al.,* 1989). This luminance level is about 100 times higher than that of ZnS:Tm,F (Table 6), since the maximum luminance to date is 1150 cd/m^2 under the 1-kHz drive condition (Mita *et al.,* 1991). Therefore SrS:Ce is promising as a blue EL phosphor.

The SrS thin films tend to be sulfur deficient and easily oxidized (Onisawa *et al.,* 1988; Okamoto and Hanaoka, 1988) as well as being hygroscopic. To cope with these problems, sulfur coevaporation and reduction of oxygen in the target materials have been tried, and an increase in luminance has been realized. To prevent moisture from coming into the phosphor layer, device structure optimization, including different insulating-layer material combinations and process conditions, must be studied.

5.3 Summary of Color Thin-Film EL Devices

The above results indicate that the following phosphors are suitable for multicolor EL devices based on considerations of luminance and color gamut: ZnS:Mn with red color filter or CaS:Eu or ZnS:Sm,Cl for red emission; ZnS:Tb,F for green emission; and SrS:Ce for blue emission. The color coordinates of these EL devices are shown in Fig. 25. In the figure, R, G, and B (denoted by black squares) represent the color gamut positions of red, green, and blue of color CRTs, respectively, i.e., R for Y_2O_3:Eu, G for ZnS:Cu,Al, and B for ZnS:Ag. The CaS:Eu red EL emission and ZnS:Tb,F green EL emission are close to the respective CRT emissions. On the other hand, the SrS:Ce blue-green EL emission is far from the blue CRT emission. By using a blue filter, higher-purity blue emission (indicated by the triangle) can be realized. Luminance, however, then falls below the standards required for commercial application.

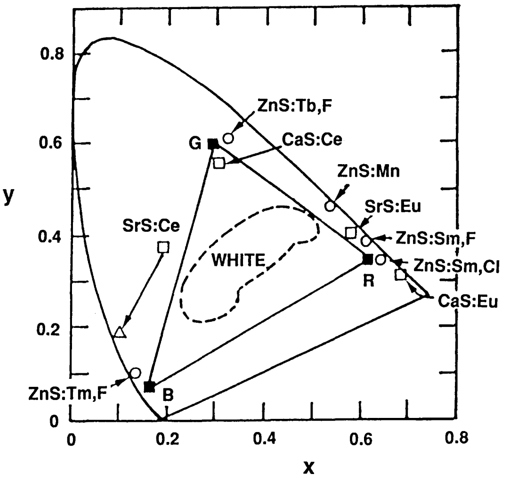

FIG. 25. CIE color coordinates of color EL devices (after Ono *et al.,* 1989).

6. MULTICOLOR THIN-FILM EL PANELS

Three types of multicolor thin-film EL panel structures have been investigated (Hamakawa *et al.,* 1988; King, 1989):

1. stacked-phosphor structure,
2. patterned-phosphor structure, and
3. white or broadband-spectrum phosphor with color filters.

Among these, the patterned-phosphor structure is most easily realizable, considering deposition conditions and drive circuitry.

6.1 Stacked-Phosphor Structure

Figure 26 shows a typical stacked-phosphor device structure, where two EL devices of different colors are stacked to produce

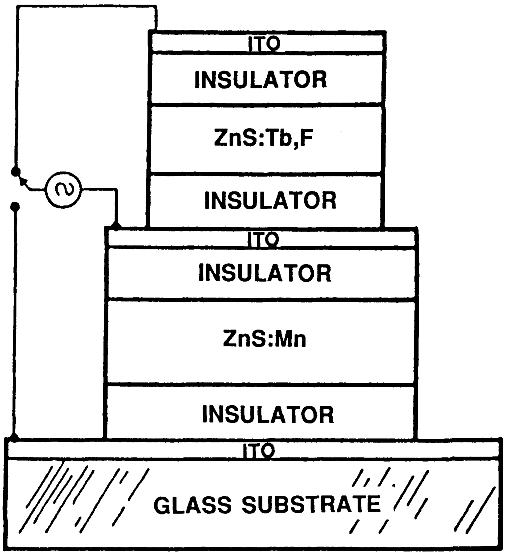

FIG. 26. Structure of a stacked-phosphor type multicolor EL device (after Coovert *et al.,* 1982).

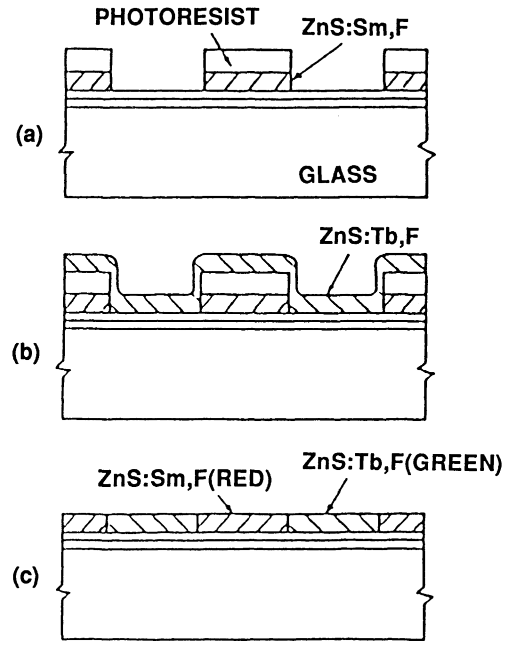

FIG. 27. Process procedure of a patterned-phosphor type multicolor EL device (after Yamauchi *et al.*, 1987).

multicolor emission (Coovert *et al.*, 1982). Advantages of this structure are that the resolution of each EL device can be retained, and there is no need for patterning of phosphor layers. On the other hand, disadvantages are that thin-film deposition becomes harder at the upper layers and the drive circuitry becomes complicated. Barrow *et al.* (1986) fabricated a prototype three-color (green, yellow, and red) EL device by stacking green and red EL devices.

6.2 Patterned-Phosphor Structure

In this structure, photolithographically patterned phosphor layers corresponding to three primary colors are placed on the same plane. Its advantages are a drive circuitry that is essentially the same as the monochromatic EL devices, and simpler deposition processes compared with those of the stacked pattern. Disadvantages are a reduction of resolution and difficulty in phosphor-layer patterning. Examples are shown in Figs. 27 and 28. Figure 27 depicts the process for obtaining a three-color EL device consisting of self-aligned phosphors of red (ZnS:Sm,F) and green (ZnS:Tb,F) (Yamauchi *et al.*, 1987). Barrow *et al.* (1988) fabricated a prototype full-color EL display panel by patterning-phosphor layers of red (ZnS:Sm,F), green (ZnS:Tb,F), and blue (SrS:Ce).

Yamamoto *et al.* (1989) fabricated a prototype 10-in.-diagonal three-color (orange-yellow, yellow, and green) EL display panel. Its structure is shown in Fig. 28. In order to protect the first phosphor layer of ZnS:Mn from being etched, an intermediate layer was placed between the first phosphor layer and the second phosphor layer of ZnS:Tb,F.

6.3 White or Broadband-Spectrum Phosphor with Color Filters

By combining a white-light–emitting EL phosphor having red, green, and blue spectral components, or a broadband-spectrum phosphor, with color filters, it becomes possible to obtain multicolor EL devices. For this purpose, white-light–emitting EL devices using SrS and CaS have been investigated (Tanaka *et al.*, 1987b, 1988b; Ono *et al.*, 1989). Figure 29 shows color coordinates of several white-light–emitting EL devices. In particular, SrS:Ce/SrS:Eu and SrS:Ce,Eu devices (Tanaka *et al.*, 1988b) produced white emissions because blue-green from Ce^{3+} and orange-red from Eu^{2+} are complementary colors. As expected from Fig. 24, the emission spectrum of SrS-based white-emitting EL devices is broad

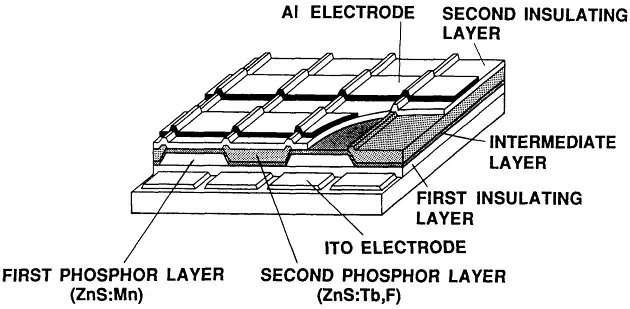

FIG. 28. Structure of a patterned-phosphor type multicolor EL device (after Yamamoto *et al.*, 1989).

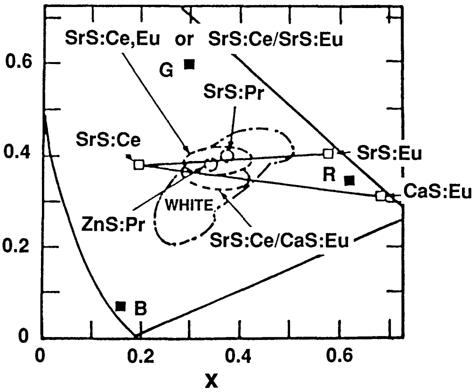

FIG. 29. CIE color coordinates of white EL devices.

so that they are suitable for multicolor EL phosphors. Luminance, however, is still low, at most 1150 cd/m^2 under 1-kHz drive conditions (Mita *et al.,* 1991). For the present purpose, at least 3000 cd/m^2 under 1-kHz drive conditions is necessary. On the other hand, from an ergonomics viewpoint, a paper-white emission is more desirable than the eggshell-white emission from SrS:Ce,Eu or SrS:Ce/SrS:Eu. Therefore, stacked-phosphor layers of SrS:Ce/CaS:Eu were studied (Ono *et al.,* 1989), and good color balance of R:G:B =3:6:1 was obtained by using color filters.

A prototype full-color EL device using CaS- and SrS-based phosphors and interference color filters has been fabricated (Tanaka *et al.,* 1987b). Its cross-sectional structure is shown in Fig. 30. The emission colors were very close to those of a CRT. However, luminance was very low and at least one order of magnitude increase in luminance is necessary.

A candidate for a broadband-spectrum phosphor with high luminance is ZnS:Mn, as shown in Fig. 22(a). Okibayashi *et al.* (1991) fabricated a prototype multicolor (green–red) EL device using ZnS:Mn with organic color filters. The luminances and color coordinates of filtered EL emission under the frame frequency of 60-Hz drive conditions were as follows: $L(\text{green})=80$ cd/m^2 with $x=0.45$ and $y=0.55$, and $L(\text{red})=30$ cd/m^2 with $x=0.66$ and $y=0.34$. Here the original orange-yellow luminance under the same drive conditions was 200 cd/m^2.

7. RELIABILITY

In considering the reliability of EL devices, two important factors must be taken into account:

1. electrical reliability with respect to electrical breakdown, and
2. long-time reliability.

Electrical breakdown properties depend strongly on the electrical properties of the dielectric-layer materials and the thin-film fabrication processes. For long-time reliability, the hygroscopic, adhesive, and photolithographic resistant properties of phosphor and dielectric-layer materials must be taken into account, and proper choices of passivation and sealing materials are essential.

7.1 Electrical Reliability

As discussed in Sec. 3.3, the requirements of the insulating layers are high dielectric constant $\epsilon_0\epsilon_r$ and high breakdown electric field E_{BD}, i.e., a large figure of merit defined by $\epsilon_0\epsilon_r E_{\text{BD}}$, which is equal to the maximum storage charge density, $Q_{\max}$. When insulators with high $Q_{\max}$ are used, a low-voltage drive becomes possible. However, insulators with

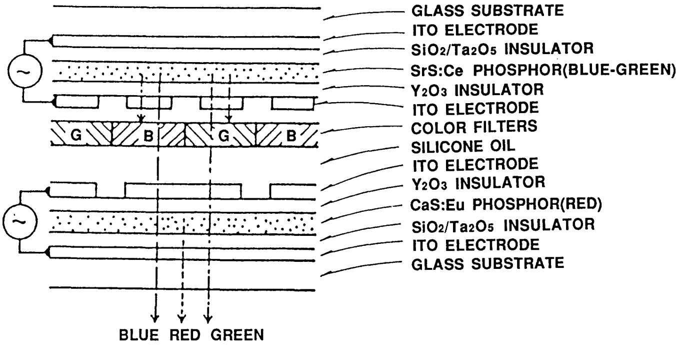

FIG. 30. Cross-sectional view of full-color EL device with stacked EL cells and color filters (after Tanaka *et al.,* 1987b).

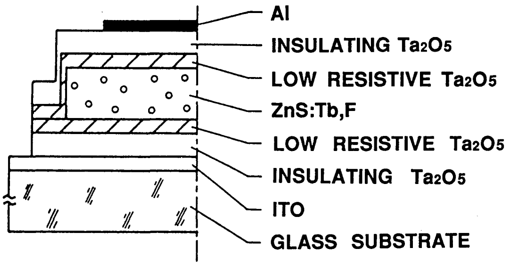

FIG. 31. EL device structure employing insulating and conducting Ta_2O_5 thin films (after Mikami and Ando, 1985).

high Q_{max} tend to have a low reliability, and propagating-type breakdown often occurs. Several multilayered structures of insulating layers have been proposed to overcome this difficulty (Mikami and Ando, 1985; Mita *et al.*, 1987). In Fig. 31, insulating and low-resistive Ta_2O_5 insulators are stacked in the order Al/insulating Ta_2O_5/low-resistive Ta_2O_5/ phosphor layer/low-resistive Ta_2O_5/insulating Ta_2O_5/ITO/glass substrate (Mikami and Ando, 1985). Low-resistive and insulating Ta_2O_5 films can be fabricated by changing deposition conditions. In this structure, insulating Ta_2O_5 films provide electrical insulation and low-resistive Ta_2O_5 films give increased charge generation and better adhesiveness between the phosphor layer and insulating layers. Another combination of the insulating layers is a SiO_2/Ta_2O_5 stacked structure in the order Al/SiO_2/Ta_2O_5/phosphor layer/Ta_2O_5/SiO_2/ ITO/glass substrate (Mita *et al.*, 1987). Here the SiO_2 layers act as electrically insulating layers and the Ta_2O_5 layers serve as partly conductive layers. In addition to these examples, SiO_2, TiO_2, and CaS are used as buffer thin films between electrodes and insulating layers or between insulating layers and the phosphor layer to stabilize EL operation (Nishikawa *et al.*, 1988; Khormaei *et al.*, 1989).

In commercial EL devices, stacked insulating layers are employed to increase the breakdown electric field. This is shown in Fig. 32 (Takeda *et al.*, 1983), where the first insulating layer consists of silicon oxinitride/SiO_2 stacked layers and the second insulating layer, of Al_2O_3/silicon oxinitride. There are two advantages for this stacked layer structure. The first is better adhesiveness; Al_2O_3 is introduced between the rear Al electrode and the second insulating layer of silicon oxinitride. The second advantage is an increase in the breakdown electric field by a decrease of the number of pinholes.

7.2 Long-Time Reliability

The lifetime-determining factors of EL devices are the peeling of thin films and the formation of very small dielectric breakdown holes. These small breakdown holes result from local defects in the thin films during the first stage of the drive annealing process. These holes are approximately 30 μm in diameter and their number does not increase with annealing time. Therefore, they do not cause any practical trouble. Proper control of process conditions is important to decrease the defects.

On the other hand, peeling of the thin films is caused by moisture (water) leaking into the phosphor layer/insulating layer interface through pinholes (Okamoto *et al.*, 1983). During EL drive operation, ionization of the water occurs from hot-electron impact in the phosphor layer, generating H^+ and OH^- ions. The H^+ ions will combine at the interface to form hydrogen gas. The gas pressure lifts the second insulating layer from the phosphor layer, leading to peeling of the thin films. With longer drive periods, peeling proceeds rapidly, and it affects the lifetime of EL devices. To avoid this, proper control of atmospheric

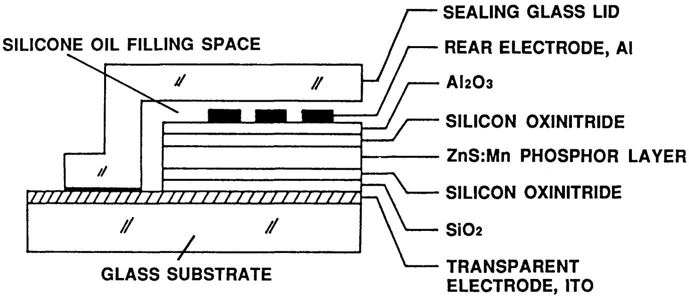

FIG. 32. Cross-sectional structure of commercially available EL devices with double insulating layers (after Takeda *et al.*, 1983).

conditions, especially humidity control, during and after the deposition process is necessary. In particular, encapsulation of EL devices by sealing is very effective to prevent moisture from entering into them. Sealing methods include oil sealing, gas sealing, vacuum sealing, and resin sealing. Among these, the oil sealing method is the best. In commercial EL display devices, silicone oil and silica gel powders are placed in the glass-sealed EL devices as shown in Fig. 32. Furthermore, since in the double-insulating-layer structure the phosphor layer is sandwiched by insulating layers, it is possible to prevent moisture from entering the phosphor layer. By employing the structures discussed above, a long lifetime, more than 20 000 hours, has been realized.

The drive method is also important to obtain long-time reliability (Watanabe *et al.*, 1987; Flegal and King, 1986). In the beginning, a field-refresh drive method was employed (Takeda *et al.*, 1981). Schematic pulse drive wave forms for scan lines are shown in Fig. 33. After all the write-in procedures for the scan lines in the line-at-a-time drive scheme are finished, field-refresh pulses with opposite polarity to the write-in pulse are applied to all the scan lines at the same time, ending a cycle of an ac drive. Therefore there are two EL emissions in one cycle. This drive method utilizes the polarization characteristics described in Sec. 1.3 to obtain high luminance. However, unsymmetrical drive voltage wave forms are created at different parts of the EL devices so that the luminance versus voltage characteristics deteriorate and latent images are created. The deterioration is characterized by decreases in both threshold voltage and luminance within the high-luminance region (Watanabe *et al.*, 1987; Flegal and King, 1986; Müller *et al.*, 1988).

To overcome this difficulty, the *p-n* symmetric drive method and the *p-p* symmetric drive method have been introduced (Harada *et al.*, 1986; Shoji *et al.*, 1989). Schematic pulse drive wave forms for scan lines in the *p-n* symmetric drive method are shown in Fig. 34. This method is also called a field-reversal method, because the polarity of the write-in pulses is reversed in every other field. Therefore, one cycle is composed of two fields and only one EL emission in one field. However, the symmetric drive method allows a higher voltage to be applied to EL devices than that in the field-refresh method, so that saturation regions of the luminance–voltage characteristic curves can be utilized. This makes it possible to realize a uniform luminance over the whole area of the EL device, and the luminance change with time is substantially reduced.

8. FUTURE PROSPECTS

Monochromatic EL display devices are now commercially available from three companies; Sharp Corporation, Planar Systems, and Lohja Corporation (now Planar Systems, International). These EL display devices are used as flat screens of laptop personal computers, workstations, and word processors (see Plates 1 and 2). However, volume production is small compared with that of liquid-crystal displays (LCDs) because of the high cost, and so price reduction is necessary. Lower cost can only be realized if cheaper high-voltage driver integrated circuits (ICs) become available. Thanks to innovations in semiconductor technology, this will be possible in the near future.

A second reason why EL display devices are not so popular is a lack of awareness about

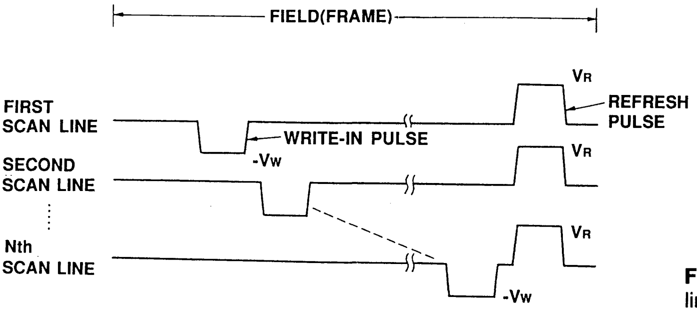

FIG. 33. Pulse drive wave forms for scan lines in the field-refresh drive method.

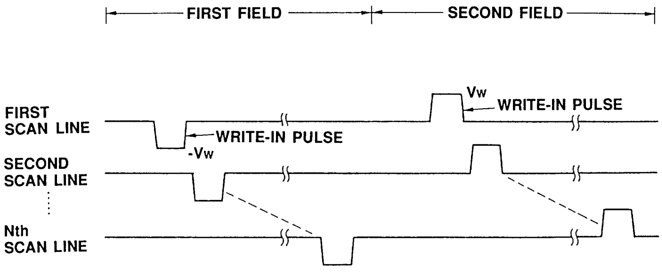

FIG. 34. Pulse drive wave forms for scan lines in the *p-n* symmetric drive method.

their good characteristics. In comparison with other flat-panel displays (See DISPLAY TECHNOLOGY), such as LCDs, plasma display panels (PDPs), and vacuum fluorescent displays (VFDs), EL display devices have the following superior characteristics, and their applications should be encouraged:

1. good display qualities from an ergonomics viewpoint
   - an emissive display with large viewing angle
   - a possibility of resolution higher than 8 lines/mm
2. temperature-independent emission
   - suitable for automotive equipment used under very severe temperature requirements
3. all-solid devices
   - rugged and vibration resistant
   - suitable for personal computers used under very severe vibrational conditions, such as in military equipment or spacecraft
4. low-power, thin, lightweight devices
   - lowest power consumption among emissive flat-panel displays
   - thickness of less than 24 mm and weight of less than 1 kg in practical panels

This review has discussed the present status of EL with an emphasis on the double-insulating-layer type thin-film EL, because of its high possibility to be commercialized as flat-panel displays. The most important research and development target is realization of multicolor and full-color EL display devices. As discussed in Sec. 6, several prototype multicolor EL display devices have been fabricated (see Plates 3 and 4). However, for full-color EL display devices, bright red, green, and blue EL emissions with good purities and balanced luminance levels are necessary. To realize this, considerations of color phosphor materials and color EL drive methods have to be taken into account.

First, consider phosphor materials. For a green phosphor, ZnS:Tb,F is suitable with respect to luminance and purity. For a red phosphor, ZnS:Mn with a red color filter or ZnS:Sm,Cl or CaS:Eu is a good candidate, but luminance improvement is still necessary. For a blue phosphor, SrS:Ce is a possible candidate, but the emission color is blue-green because of the broad spectrum as shown in Fig. 24(c). Therefore, color filtering is necessary to obtain pure blue, the luminance of which is still low (one order of magnitude increase is desired). However, at this point of writing, no good improvement method for blue emission has been proposed.

The task now should be to find suitable combinations of phosphor materials and luminescent centers suitable for full-color representation. After finding suitable combinations, luminance improvement of each color should be pursued.

Next, consider drive methods. It is possible to obtain memory EL devices by incorporating thin-film transistors (TFTs) at each pixel, which has already been done in color LCD panels. This method was first applied to EL devices, but no practical display panels were fabricated because of difficulties in obtaining large-area, high-voltage TFTs without defects. However, recent success in large-area TFTs for LCDs together with substantial improvements in fabrication technologies should give a strong impetus to the EL community in trying to fabricate TFT-driven EL display devices. One of the advantages of three-terminal TFTs over the present two-terminal devices is that the circuit system with TFTs is easier to handle. Once high-voltage TFT arrays attain a practical level of production, TFT-driven EL display devices should become the mainstream of memory devices.

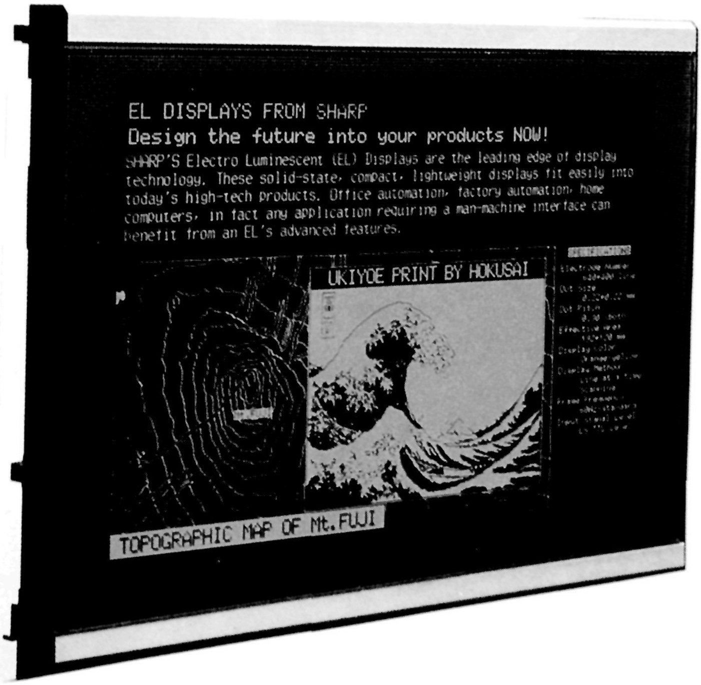

EL DISPLAYS FROM SHARP

Design the future into your products NOW!

SHARP'S Electro Luminescent (EL) Displays are the leading edge of display technology. These solid-state, compact, lightweight displays fit easily into today's high-tech products. Office automation, factory automation, home computers, in fact any application requiring a man-machine interface can benefit from an EL's advanced features.

Vol. 5 – PLATE 1. Photograph of orange-yellow emitting ZnS:Mn EL display unit with 640·400 dots in normal-mode representation (courtesy of Sharp Corporation).

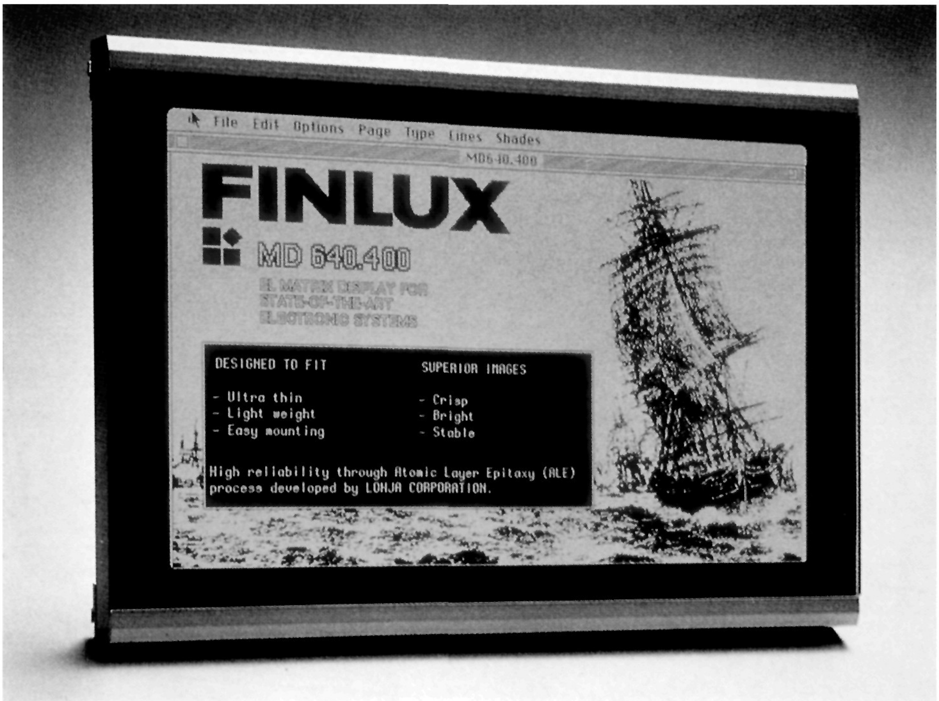

FINLUX

MD 640.400

EL MATRIX DISPLAY FOR STATE-OF-THE-ART ELECTRONIC SYSTEMS

DESIGNED TO FIT

- Ultra thin
- Light weight
- Easy mounting

SUPERIOR IMAGES

- Crisp
- Bright
- Stable

High reliability through Atomic Layer Epitaxy (ALE) process developed by LOHJA CORPORATION.

Vol. 5 – PLATE 2. Photograph of orange-yellow emitting ZnS:Mn EL display unit with 460·400 dots in reverse-mode representation [courtesy of Lohja Corporation (now Planar Systems, International)].

Vol. 5 – PLATE 3. Photograph of prototype three-color (red, orange, and green) EL display in stacked phosphor structure with 320·240 dots (Courtesy of Planar Systems.)

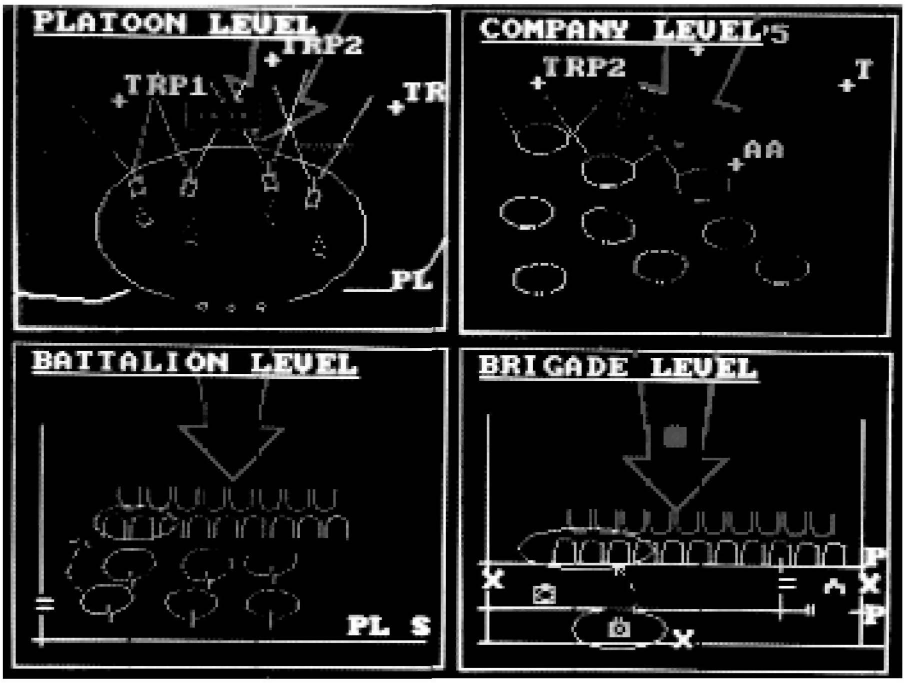
PLATOON LEVEL
TRP2
TRP1
TR
PL
COMPANY LEVEL
5
TRP2
T
AA
BATTALION LEVEL
PL S
BRIGADE LEVEL
P
X
X
P
X

Vol. 5 – PLATE 4. Photograph of prototype full-color display in patterned phosphor structure with (320·240 dots (courtesy of Planar Systems.)

GLOSSARY

Contrast Ratio: The ratio of the "on" luminance with respect to the background luminance or "off" luminance. Pixel visibility depends on this quantity.

Impact Excitation: Excitation of ions in a crystal field via a strong interaction with high-energy electrons or hot electrons.

Luminance: Photometric term indicating the radiated power in terms of human eye sensitivity, i.e., a physical measure of luminous intensity per unit area, in units of cd/m^2. Other units, ft L (foot-lambert), are also used: $1\ ft\ L = 3.43\ cd/m^2$.

Luminescence: Physical phenomenon indicating emission of light.

Luminous Efficiency: Number of lumens emitted per watt electrical power.

Pixel (Picture Element): Basic building blocks for displays. One monochromatic pixel is composed of a single dot and one color pixel is composed of three primary color dots, i.e., in red, green, and blue triads.

Powder EL Device: ZnS phosphor powder dispersed in an insulator with a binder, sandwiched between two electrodes.

Primary Colors: Mathematical abstraction of spatial energy distributions which correspond to human sensation of red, green, and blue. The primary colors can be combined to give any desired color. [see the chromaticity diagram of the CIE (Commission Internationale de l'Eclairage), 1931.]

Threshold Voltage: The voltage at which luminance becomes $1\ cd/m^2$.

LIST OF ABBREVIATIONS

ALE: Atomic layer epitaxy
BCPM: bis-methylcyclopentadienyl-Mn
CIE: Commission International de l'Eclairage
CPM: di-π-cyclopentadienyl-Mn
CRT: Cathode ray tube
CVD: Chemical vapor deposition
DES: Diethylsulfide
DEZ: Diethylzinc
EBD: Electron-beam deposition
EL: Electroluminescence, electroluminescent
HT-CVD: Hydride-transport chemical vapor deposition
ITO: Indium tin oxide
LCD: Liquid-crystal display
LED: Light-emitting diode
Lumocen: Luminescence from molecular centers
MOCVD: Metal-organic chemical vapor deposition
MSD: Multisource deposition
PB: Propagating mode breakdown mode
PDP: Plasma display panel
PVD: Physical vapor deposition
RHEED: Reflection high-energy electron diffraction
SB: Self-healing breakdown mode
TCM: Tricarbonylmethylcyclopentadienyl-Mn
TEM: Transmission electron microscopy
TFT: Thin-film transistor

List of Works Cited

Adachi, C., Tokito, S., Tsutsui, T., Saito, S. (1988), *Jpn. J. Appl. Phys.* **27**, L269–L271.

Alt, P. M. (1984), *Proc. Soc. Inf. Disp.* **25**, 123–146.

Ando, M., Ono, Y. A. (1989), *J. Appl. Phys.* **65**, 3290–3292.

Barrow, W. A., Coovert, R. E., King, C. N. (1984), in: *Digest of the 1984 SID International Symposium*, Los Angeles: Society for Information Display.

Barrow, W. A., Tuenge, R. T., Ziuchkovski, M. J. (1986), in: *Digest of the 1986 SID International Symposium*, Playa del Rey, CA: Society for Information Display, pp. 25–28

Barrow, W. A., Coovert, R. E., King, C. N., Ziuchkovski, M. J. (1988), in: *Digest of the 1986 SID International Symposium*, Playa del Rey, CA: Society for Information Display, pp. 284–285.

Cattel, A. R., Cockayne, B., Dexter, K., Kirton, J., Wright, P. J. (1983), *IEEE Trans. Electron Devices* **ED-30**, 471–475.

Chase, E. W., Heppelwhite, R. T., Krupka, D. C., Kahng, D. (1969), *J. Appl. Phys.* **40**, 2512–2519.

Chen, Y. S., Krupka, D. C. (1972), *J. Appl. Phys.* **43**, 4089–4096.

Coovert, R. E., King, C. N., Tuenge, R. T. (1982), in: *Digest of the 1982 SID International Symposium*, Los Angeles: Society for Information Display, pp. 128–129.

Crandall, R. S., (1987a), *Appl. Phys. Lett.* **50**, 551–553.

Crandall, R. S. (1987b), *Appl. Phys. Lett.* **50**, 641–643.

Cusano, D. A. (1962), in: H. P. Kallmann, G. M. Spruch (Eds.), *Luminescence of Organic and Inorganic Materials*, New York: Wiley.

Destriau, G. (1936), *J. Chim. Phys.* **33**, 587–625.

Faulkner, E. K., Whitney, R. K., Zeman, J. E. (1983), *Proc. Soc. Inf. Disp.*, **24**, 205–208.

Fischer, A. G. (1962), *J. Electrochem. Soc.* **109**, 1041–1049.

Fischer, A. G. (1963), *J. Electrochem. Soc.* **110**, 733–748.

Flegal, R. T., King, C. N. (1986), in: *Digest of the 1986 SID International Symposium,* Playa del Rey, CA: Society for Information Display.

Fujita, Y., Kuwata, J., Nishikawa, M., Tohda, T., Matsuoka, T., Abe, A., Nitta, T. (1984), *Proc. Soc. Info. Displ.* **25**, 177–182.

Hamakawa, Y., Fukao, R., Fujikawa, H. (1988), *Optoelectron. Devices Technol.* **3**, 31–46.

Harada, S., Ohba, T., Kanatani, Y., Uede, H. (1986), in: K. Miyaji, C. Suzuki, S. Kobayashi (Eds.), *Proceedings of the 6th International Display Research Conference, Japan Display '86,* Playa del Rey, CA: Society for Information Display, and Minatoku, Tokyo: Institute of Television Engineers of Japan, 238–241.

Higton, M., Vecht, A., Mayo, J. (1978), in: *Digest of the 1978 SID International Symposium,* Los Angeles: Society for Information Display.

Hirabayashi, K., Kozawaguchi, H. (1986), *Jpn. J. Appl. Phys.* **25**, 711–713.

Hirabayashi, K., Kozawaguchi, H. Tsujiyama, B. (1987), *Jpn. J. Appl. Phys.* **26**, 1472–1476.

Howard, W. E. (1977), *IEEE Trans. Electron Devices* **ED-24**, 903–908.

Howard, W. E. (1981), *Proc. Soc. Inf. Disp.* **22**, 47–56.

Inoguchi, T., Takeda, M., Kakihara, Y., Nakata, Y., Yoshida, M. (1974), in: *Digest of the 1974 SID International Symposium,* Los Angeles: Society for Information Display, pp. 84–85.

Kahng, D. (1968), *Appl. Phys. Lett.* **13**, 210–212.

Khormaei, R., Wager, J. F., King, C. N. (1989), in: *Digest of the 1989 SID International Symposium,* Playa del Rey, CA: Society for Information Display, pp. 65–67.

King, C. N. (1989), in: A. Sasaki, T. Kojima, H. Uchiike (Eds.), *Proceedings of the 9th International Display Research Conference, Japan Display '89,* Playa del Rey, CA: Society for Information Display, and Minatoku, Tokyo: Institute of Television Engineers of Japan, pp. 224–226.

Kobayashi, H. (1983), in: M. A. Herman (Ed.), *Optoelectronic Materials and Devices,* Warszawa: PWN–Polish Scientific Publishers, Chap. 13.

Kobayashi, H., Tanaka, S., Shanker, V., Shiiki, M., Kunou, T., Mita, J., Sasakura, H. (1985), *Phys. Status Solidi A* **88**, 713–720.

Kobayashi, S., Aoki, Y., Enjoji, K., Johnson, I. P., Higton, M. H., Huntley, P. J. (1989), in: A. Sasaki, T. Kojima, H. Uchiike (Eds.), *Proceedings of the 9th International Display Research Conference, Japan Display '89,* Playa del Rey, CA: Society for Information Display, and Minatoku, Tokyo: Institute of Television Engineers of Japan, pp. 242–245.

Kozawaguchi, H., Ohwaki, J., Tsujiyama, B., Murase, K. (1982), *Proc. Soc. Inf. Disp.* **23**, 181–186.

Kun, Z. K., Leksell, D., Malmberg, P. R., Asars, J., Brandt, G. B. (1986), in: *Digest of the 1986 SID International Sympposium,* Playa del Rey, CA: Society for Information Display, pp. 270–272.

Lehmann, W. (1972), *J. Lumin.* **5**, 87–107.

Leskelä, M. (1989), in: S. Shionoya, H. Kobayashi (Eds.), *Electroluminescence,* Springer Proceedings in Physics, Vol. 38. Heidelberg: Springer, pp. 218–223.

Mach, R., Müller, G. O. (1982), *Phys. Status Solidi A* **69**, 11–66.

Matsuoka, T., Kuwata, J., Nishikawa, M., Fujita, Y., Tohda, T., Abe, A. (1988), *Jpn. J. Appl. Phys.* **27**, 592–596.

Migita, M., Kanehisa, O., Shiiki, M., Yamamoto, H. (1988), *J. Cryst. Growth* **93**, 686–691.

Mikami, A., Ando, K. (1985), *Trans. Inst. Electron Commun. Eng. Jpn.* **J68-C**, 285–290 [in Japanese].

Mikami, A., Ogura, T., Tanaka, K., Taniguchi, K., Yoshida, M., Nakajima, S. (1987), *J. Appl. Phys.* **61**, 3028-3034.

Mikami, A., Ogura, T., Taniguchi, K., Yoshida, M., Nakajima, S. (1988), *J. Appl. Phys.* **64**, 3650–3657.

Mikami, A., Terada, K., Okibayashi, K., Tanaka, K., Yoshida, M., Nakajima, S. (1991), *J. Cryst. Growth* **110**, 381–394.

Mishima, T., Takahashi, K. (1983), *J. Appl. Phys.* **54**, 2153–2155.

Mishima, T., Konagai, M., Takahashi, K. (1983), *IEEE Trans. Electron Devices* **ED-30**, 282–284.

Mita, J., Koizumi, M., Hayashi, T., Sekido, Y., Kazama, M., Nihei, K. (1986), in: K. Miyaji, C. Suzuki, S. Kobayashi (Eds.), *Proceedings of the 6th International Display Research Conference, Japan Display '86,* Playa del Rey, CA: Society for Information Display, and Minatoku, Tokyo: Institute of Television Engineers of Japan, pp. 250–253.

Mita, J. Koizumi, M., Kanno, H., Hayashi, T., Sekido, Y., Abiko, I., Nihei, K. (1987), *Jpn. J. Appl. Phys.* **26**, L541–L543.

Mita, J., Gao, Q. Z., Kobayashi, M., Kawamura, K. (1991), in: *Digest of the 1991 SID International Symposium,* Playa del Rey, CA: Society for Information Display.

Mito, S., Suzuki, C., Kanatani, Y., Ise, M. (1974), in: *Digest of the 1974 SID International Symposium,* Los Angeles: Society for Information Display.

Miura, S., Okamoto, K., Sato, S., Andoh, S., Ohnishi, H., Hamakawa, Y. (1983), in: K. Miyaji, I. Ohishi, A. Sasaki (Eds.), *Proceedings of the 3rd International Display Research Conference, Japan Display '83,* Los Angeles: Society for Information Display, and Minatoku, Tokyo: Institute of Television Engineers of Japan, pp. 84–87.

Müller, G. O., Mach, R., Reets, R., Reinsperger, G. U. (1988), in: *Digest of the 1988 SID International Symposium,* Playa del Rey, CA: Society for Information Display.

Nire, T., Watanabe, T., Tsurumaki, N., Miyakoshi, A., Tanda, S. (1989), in: S. Shionoya and H. Kobayashi (Eds.), *Electroluminescence,* Springer Proceedings in Physics, Vol. 38. Heidelberg: Springer.

Nishikawa, M., Matsuoka, T., Tohda, T., Fujita, Y., Kuwata, J., Abe, A. (1988), *Proc. Soc. Inf. Disp.* **29**, 301–304.

Ogura, T., Mikami, A., Tanaka, K., Taniguchi, K., Yoshida, M., Nakajima, S. (1986), *Appl. Phys. Lett.* **48**, 1570–1571.

Ohnishi, H. (1989), *Annu. Rev. Mater. Sci.* **19**, 83–101.

Ohnishi, H., Yamasaki, Y., Iwase, R. (1987), *Proc. Soc. Inf. Disp.* **28**, 345–350.

Okamoto, K., Hanaoka, K. (1988), *Jpn. J. Appl. Phys.* **27**, L1923–L1925.

Okamoto, K., Miura, S. (1986), *Appl. Phys. Lett.* **49**, 1596–1598.

Okamoto, K., Watanabe, K. (1986), *Appl. Phys. Lett.* **49**, 578–580.

Okamoto, K., Nasu, Y., Hamakawa, Y. (1981), *IEEE Trans. Electron Devices* **ED-28**, 698–702.

Okamoto, K., Wakitani, M., Sato, S., Miura, S. Andoh, S., Umeda, S. (1983), in: *Digest of the 1983 SID International Symposium,* Los Angeles: Society for Information Display.

Okamoto, K., Yoshimi, T., Miura, S. (1989a), in: S. Shionoya and H. Kobayashi (Eds.), *Electroluminescence,* Springer Proceedings in Physics, Vol. 38. Heidelberg: Springer.

Okamoto, K., Yoshimi, T., Nakamura, K. Kobayashi, T., Sato, S., Miura, S. (1989b), *Jpn. J. Appl. Phys.* **28**, 1378–1384.

Okibayashi, K., Ogura, T., Terada, K., Taniguchi, K., Yamashita, T., Yoshida, M., Nakajima, S. (1991), in: *Digest of the 1991 SID International Symposium,* Playa del Rey, CA: Society of Information Display, pp. 275–278.

Onisawa, K., Fuyama. M., Taguchi, K., Tamura, Y., Ono, Y. A. (1988), *J. Electrochem. Soc.* **135**, 2631–2634.

Ono, Y. A. (1990), *Acta Polytech. Scand.* **170**, Applied Physics Series No. 170, 41–47.

Ono, Y. A., Kawakami, H., Fuyama, M., Onisawa, K. (1987), *Jpn. J. Appl. Phys.* **26**, 1482–1492.

Ono, Y. A., Fuyama, M., Onisawa, K., Tamura, K., Ando, M. (1989), *J. Appl. Phys.* **66**, 5564–5571.

Russ, M. J., Kennedy, D. I. (1967), *J. Electrochem. Soc.* **114**, 1066–1072.

Sasakura, H., Kobayashi, H., Tanaka, S., Mita, J., Tanaka, T., Nakayama, H. (1981), *J. Appl. Phys.* **52**, 6901–6906.

Schmachtenberg, R., Jenness, T., Ziuchkovski, M., Flegal, T. (1989), in: *Digest of the 1989 SID International Symposium,* Playa del Rey, CA: Society for Information Display, pp. 58–60.

Shimizu, Y., Matsudaira, T. (1985), in: P. A. Penz (Ed.), *Conference Record of the 1985 International Display Research Conference,* New York: IEEE, and Playa del Rey, CA: Society for Information Display, pp. 101–105.

Shoji, K., Ohba, T., Kishishita, H., Uede, H. (1989), in: S. Shionoya and H. Kobayashi (Eds.), *Electroluminescence,* Springer Proceedings in Physics, Vol. 38. Heidelberg: Springer, pp. 324–330.

Smith, D. H. (1981), *J. Lumin.* **23**, 209–235.

Suntola, T., Anston, J., Pakkala, A., Linfors, S. (1980), in: *Digest of the 1980 SID International Symposium,* Los Angeles: Society for Information Display, pp. 108–109.

Takeda, M., Kanatani, Y., Kishishita, H., Inoguchi, T., Okano, K. (1981), *Proc. Soc. Inf. Disp.* **22**, 57–62.

Takeda, M., Kanatani, Y., Kishishita, H., Uede, H. (1983), in: *Advances in Display Technology III,* SPIE Proceedings Vol. 386, Bellingham, WA: SPIE, pp. 34–38.

Tanaka, S., Shanker, V., Shiiki, M., Deguchi, H., Kobayashi, H. (1985), *Proc. Soc. Inf. Disp.* **26**, 255–258.

Tanaka, S., Deguchi, H., Mikami, Y., Shiiki, M., Kobayashi H. (1987a), *Proc. Soc. Inf. Disp.* **28**, 21–25.

Tanaka, S., Mikami, Y., Nishiura, J., Ohshio, S., Yoshiyama, H., Kobayashi, H. (1987b), *Proc. Soc. Inf. Disp.* **28**, 357–363.

Tanaka, S., Yoshiyama, H., Mikami, Y., Nishiura, J., Ohshio, S. Kobayashi, H. (1988a), *Proc. Soc. Inf. Disp.* **29**, 77–81.

Tanaka, S., Yoshiyama, H., Nishiura, J., Ohshio, S., Kawakami, H., Kobayashi, H. (1988b), *Proc. Soc. Inf. Disp.* **29**, 305–310.

Tanda, S, Miyakoshi, A., Nire, T. (1988), in: L. F. Weber (Ed.), *Conference Record of the 1988 International Display Conference,* New York: IEEE, and Playa del Rey, CA: Society for Information Display, pp. 122–124.

Tang, C. W., VanSlyke, S. A. (1987), *Appl. Phys. Lett.* **51**, 913–915.

Tang, C. W., VanSlyke, S. A., Chen, C. H. (1989), *J. Appl. Phys.* **65**, 3610–3616.

Tanninen, V.-P., Oikkonen, M., Tuomi, T. (1983), *Thin Solid Films* **109**, 283–291.

Theis, D., Oppolzer, H., Ebbinghaus, G., Schild, S. (1983), *J. Cryst. Growth* **63**, 47–57.

Tiku, S. K., Smith, G. C. (1984), *IEEE Trans. Electron Devices* **ED-31**, 105–108.

Törnqvist, R., Antson, J., Skarp, J., Tanninen, V. P. (1983), *IEEE Trans. Electron Devices* **ED-30**, 468–471.

Tuenge, R. T., Kane, J. (1991), in: *Digest of the 1991 SID International Symposium,* Playa del Rey, CA: Society for Information Display.

Uede, H., Kanatani, Y., Kishishita, H., Fujimori, A., Okano, K. (1981), in: *Digest of the 1981 SID International Symposium,* Los Angeles: Society for Information Display, pp. 28–29.

Vecht, A. (1980), in: *Digest of the 1980 SID International Symposium,* Los Angeles: Society for Information Display, pp. 110–111.

Vecht, A. (1989), in: *Digest of the 1989 SID International Symposium,* Playa del Rey, CA: Society for Information Display, pp. 304–308.

Vecht, A., Werring, N. J., Smith, P. J. F. (1968), *Brit. J. Appl. Phys.* **1**, 134–136.

Watanabe, J., Wakitani, M., Sato, S., Miura, S. (1987), in: *Digest of the 1987 SID International Symposium,* Playa del Rey, CA: Society for Information Display, pp. 288–291.

Yamamoto, M., Inoue, M., Tohada, T., Matsuoka, T., Abe, A. (1989), in: A. Sasaki, T. Kojima, H. Uchiike (Eds.), *Proceedings of the 9th International Display Conference, Japan Display '89,* Playa del Rey, CA: Society for Information Display, and Minatoku, Tokyo: Institute of Television Engineers of Japan, pp. 228–231.

Yamauchi, N., Kozawaguchi, H., Kogure, O., Tsujiyama, B. (1987), in: *Digest of the 1987 SID International Symposium,* Los Angeles: Society for Information Display, pp. 230–233.

Yoshida, M., Tanaka, K., Taniguchi, K., Yamashita, T., Kakihara, Y., Inoguchi, T. (1980), in: *Digest of the 1980 SID International Symposium,* Los Angeles: Society for Information Display, pp. 106–107.

Yoshiyama, H., Tanaka, S., Mikami, Y., Ohshio, S., Nishiura, J. Kawakami, H., Kobayashi, H. (1988), *J. Cryst. Growth* **88**, 56–60.

Yoshiyama, H., Sohn, S. H., Tanaka, S., Kobayashi, H. (1989), in: S. Shionoya and H. Kobayashi (Eds.), *Electroluminescence,* Springer Proceedings in Physics, Vol. 38. Heidelberg: Springer, pp. 48–55.

Further Reading

Alt, P. M. (1984), *Proc. Soc. Inf. Disp.* **25**, 123–146.

Howard, W. W. (1981), *Proc. Soc. Inf. Disp.* **22**, 47–56.

Kobayashi, H. (1983), in: M. A. Herman (Ed.), *Optoelectronic Materials and Devices,* Warszawa: PWN–Polish Scientific Publishers, Chap. 13.

Läskelä, M., Nykänen, E. (Eds.) (1990), *Acta Polytech. Scand.* **170**, Applied Physics Series No. 170, Fifth International Workshop on Electroluminescence, Helsinki: Finnish Academy of Technology.

Mach, R., Müller, G. O., (1982), *Phys. Status Solidi A* **69**, 11–66.

Ohnishi, H. (1989), *Annu. Rev. Mater. Sci.* **19**, 83–101.

Ono, Y. A., Kawakami, H., Fuyama, M., Onisawa, K. (1987), *Jpn. J. Appl. Phys.* **26**, 1482–1492.

Shionoya, S., Kobayashi, H. (Eds.) (1989), *Electroluminescence—Proceedings of the Fourth International Workshop,* Springer Proceedings in Physics, Vol. 38. Heidelberg: Springer.

Takeda, M., Kanatani, Y., Kishishita, H., Inoguchi, T., Okano, K. (1981), *Proc. Soc. Inf. Disp.* **22**, 57–62.

8

Measurement of Basic Electromagnetic Quantities

Bryan P. Kibble

8.1 Introduction 245
8.2 Realizing the Base Units 246
8.2.1 The Farad, Realized by a Calculable Capacitor 246
8.2.1.1 The Ohm, Derived from the Farad 247
8.2.1.2 Realizing the Ohm via Measurements of the Fine-Structure Constant 249
8.2.2 The Volt or Watt 251
8.2.2.1 The Watt Balance 251
8.2.2.2 Electrometers 252
8.2.2.3 Defining the Kilogram via Avogadro's Constant 253
8.2.2.4 Self-Consistency of the SI 254
8.3 Working Standards of Voltage and Resistance 255
8.3.1 A Standard for the Volt 255
8.3.2 A Standard for the Ohm 256
8.3.3 Conventional Values Adopted for KJ RK to Represent the SI Volt and Ohm 260
8.4 Other Primary Standards of National Measurement Institutes 260
8.4.1 dc Voltage 260
8.4.2 dc Resistance 260
8.4.2.1 The ac Behavior of Standard Resistors 260
8.4.3 Capacitance 261
8.4.4 Inductance 261
8.4.4.1 The Frequency Dependence of Standard Inductors 261
8.4.5 Electric Fields and Dielectric Constants 263
8.4.6 Magnetic Quantities 263
8.4.6.1 Total Magnetic Flux 263
8.4.6.2 Magnetic Flux Density 263
8.4.7 ac Current and Voltage 263
Glossary 265
References 265
Further Reading 266

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

8.1
Introduction

Electromagnetism deals with the relationships between two fields, electric and magnetic. Therefore, two independent electromagnetic units must be defined in terms of the meter, kilogram, and second to ensure that the electrical units are part of the International System of Units (SI) that are now used universally for practically all scientific and engineering measurements. The two necessary connections between the electrical and the mechanical units are made at present by

1. implicitly defining mechanical and electrical power to be equal by using the same unit (watt) for both; and
2. explicitly defining the ampere as "the constant current which, if maintained in two straight parallel conductors of infinite length and negligible circular section placed 1 m apart in vacuum, will produce a force equal to 2×10^{-7} N m^{-1} between these conductors." (This definition of the ampere is equivalent to defining the magnetic permeability of free space μ_0 as $4\pi \times 10^{-7}$ H m^{-1} exactly.)

These abstract definitions have to be put into practice (or realized, in the jargon of metrology). Realizations are carried out by more than one method and in more than one national measurement laboratory to secure confidence in their correctness. To realize an electrical unit, it is necessary to assume the complete accuracy of Maxwell's electromagnetic theory and to use an apparatus in a way that is, according to this theory, consistent with the definitions of the watt and the ampere. Up to about 40 years ago, the two apparatuses that realized the two independent electrical units were

1. some form of current balance in which the force between circular current-carrying coils was weighed as a direct extension of the ampere definition; and
2. a mutual inductor whose carefully measured dimensions established a link between the meter and the ohm.

Neither the current balance nor the mutual inductor is accurate enough for the present-day routine needs of science and commerce at the highest level of accuracy of a part per million (ppm) or better. They have been replaced by electrometers or watt balances that realize the volt or watt and the Thompson–Lampard calculable capacitor that realizes the ohm via the farad (Thompson and Lampard, 1956; see also Section 8.2.1). An alternative route for the ohm is via a measurement of the fine-structure constant and the quantum Hall effect (see Section 8.2.1.2).

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

The correctness of these realizations has been confirmed by examining the relationships between those fundamental physical constants whose measured values involve mass, length, and time as well as electrical quantities.

Once they have been realized by these difficult and time-consuming measurements (each of which may take many years), the electrical units are preserved for everyday use in national standardizing laboratories by ascribing values to the constants that govern the Josephson (1962) and quantum Hall effects (von Klitzing, Dorda, and Pepper, 1980). These precise effects of quantum physics, which occur at cryogenic temperatures, reproduce voltages and resistances very exactly (to 0.005 ppm at present).

Values of standards to represent other derived, electrical units must then be determined in terms of the ohm and volt. There is usually only one preferred method of establishing each derived unit. This method yields the most accurate measurement in volts and ohms of an artifact standard representing the derived unit.

At present, the stability and reproducibility of volt and ohm units preserved by means of the Josephson and quantum Hall effects greatly exceed the accuracy with which these units can be realized in terms of the meter, kilogram, and second. This situation suggests a possible future scenario: The present link to the meter, kilogram, and second could be replaced by a pair of basic electrical units (voltage defined in terms of the Josephson effect and impedance in terms of the quantum Hall effect) and then the kilogram could be derived from these. μ_0 would become an experimentally measurable physical constant, rather than a defined quantity, and the present prototype kilogram, which is the last unit based on a material artifact, would become a maintained standard whose value would have to be realized either from these new definitions of the volt and ohm via a watt balance or via a measurement of Avogadro's constant (see Section 8.2.2.3).

At the time of writing, the General Conference on Weights and Measures, which recommends changes to the SI for worldwide acceptance, is actively seeking to change the basis of the SI along these lines with a target date of 2011 (Mills *et al.*, 2006).

8.2
Realizing the Base Units

8.2.1
The Farad, Realized by a Calculable Capacitor

The Thompson–Lampard theorem (Thompson and Lampard, 1956) applied to a system of four electrodes of constant cross section in vacuum (Figure 8.1) states that the cross capacitances C_1 and C_2 per meter length are related by the expression

$$\exp(-\pi C_1/\varepsilon_0) + \exp(-\pi C_2/\varepsilon_0) = 1 \tag{8.1}$$

If C_1 is nearly equal to C_2(i.e., $C_1 - C_2 = \Delta C$), this simplifies to give the mean cross capacitance $C = (C_1 + C_2)/2$ as

$$\begin{aligned} C &= \left[\epsilon_0 (ln2)/\pi\right]\left[1 + 0.086(\delta C/C)^2\right] F/m, \\ &= \left[(ln2)/\pi\mu_0 c^2\right]\left[1 + 0.086(\delta C/C)^2\right] \\ &= 1.953\ 549\ 043\{1 + 0.086(\delta C/C)^2\} pF/m \end{aligned} \tag{8.2}$$

The farad can therefore be realized from a single length measurement and the defined values of μ_0 and the speed of light in vacuum c.

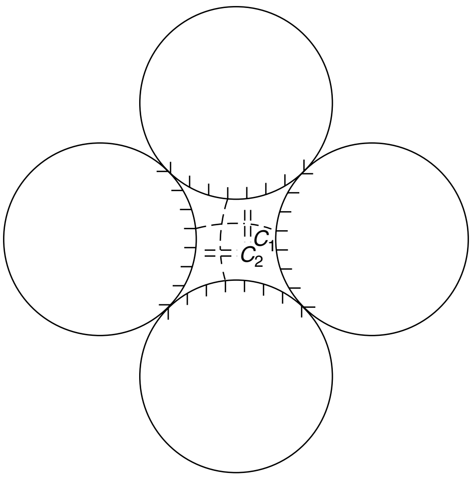

Fig. 8.1 Illustration of the Thompson–Lampard theorem, showing a cross section through four electrodes.

Usually, the cross capacitances are between the inward-facing surfaces of four cylindrical electrodes placed with their axes on the corners of a square (Figure 8.2). The small gaps between the electrodes cause negligible error. The whole structure is surrounded by a guard electrode. To avoid having to calculate the fringing fields at either end of the electrodes, guarding cylinders are inserted along the axis of the structure. The effect of the fringing fields between the ends of these guards and the cylinders is eliminated by observing the differences in the cross capacitances as the guards are moved a measured distance *L* toward each other. They must not approach each other enough to cause the fringing fields to overlap appreciably.

The accuracy of this realization of the farad is limited by the departure of real machined cylindrical electrodes from the constant cross section assumed by the Thompson–Lampard theorem. The fringing fields can be shaped by attaching extensions of lesser diameter to the ends of the inserted guards to reduce this limitation, but it is impossible to eliminate entirely the effect of changes in cross section over a length comparable with these extensions. The accuracy with which the cylinders can be made, therefore, sets a limit of the order of 0.01 ppm to the achievable accuracy in realizing the farad. A similar uncertainty arises from the small inductance of the electrodes, which causes the surfaces of the electrodes to be not quite equipotentials, because there are currents flowing along them.

8.2.1.1 The Ohm, Derived from the Farad

The ohm is obtained from the farad by scaling up the fraction of a picofarad of capacitance of a calculable capacitor to, usually, 1000 pF and then equating its impedance at a frequency of 10^4 rad s^{-1} to a resistance of 10^5 Ω, using ac bridges. This resistance is further scaled to yield the ac resistance of a 1000-Ω resistor so constructed (see Figure 8.3) that the difference in its resistance between ac and dc conditions is small and calculable.

Finally, the 1000-Ω dc SI resistance derived from a calculable capacitor by this long chain of measurements is scaled up, either by the reconfigurable resistor networks known as *Hamon resistors* (Hamon, 1954) or by a superconducting (cryogenic) current comparator (Harvey,

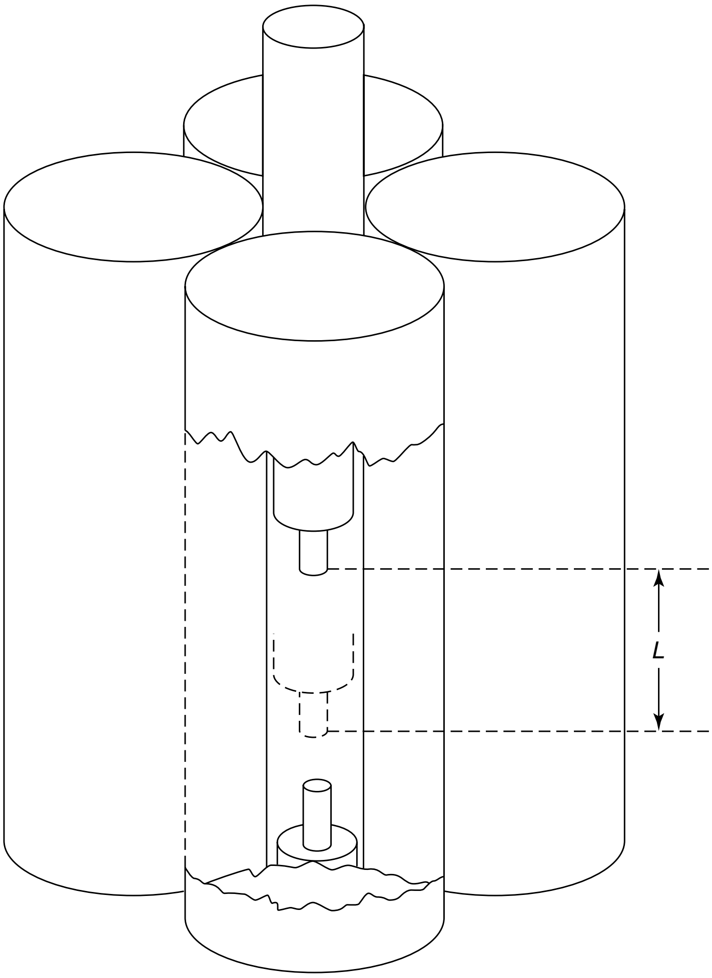

Fig. 8.2 A practical calculable capacitor.

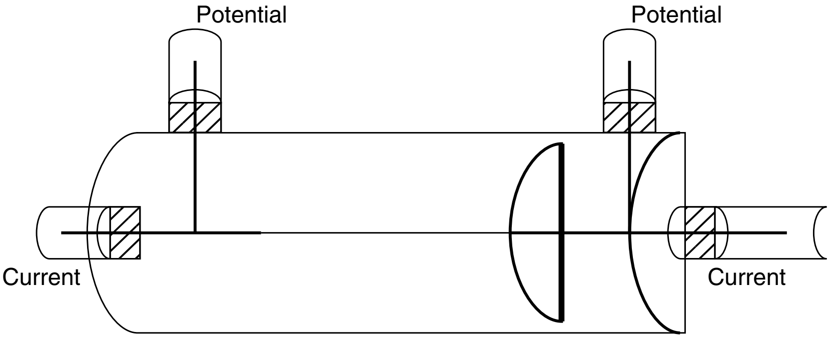
Potential
Potential
Current
Current

Fig. 8.3 A section through a coaxial resistor whose ac value is calculable from its geometry and dc value.

1972), to a value suitable for comparison with the quantum Hall effect.

A superconducting current comparator is also used to scale dc resistance ratios in major standardizing laboratories possessing liquid-helium cryogenic facilities to derive their maintained representation of the ohm from the quantum Hall effect. The construction and use of a superconducting current comparator is described in Section 8.3.2.

The parts-per-billion accuracy achieved in the AC measurements is only possible because the measuring conditions applying at the terminations of an ac standard are carefully defined. An example of these conditions is the four-terminal pair definition illustrated in Figure 8.4, where the *impedance of the standard* is defined as the ratio of the voltage appearing across the P_H coaxial connector when zero current flows through either inner or outer contact to the current flowing from the inner contact of the C_L coaxial connector. The measurement circuit must be so arranged that an equal and opposite current flows into the outer contact of the C_L coaxial connector, and that there is zero voltage between the inner and outer contacts of the associated P_L connector together with zero current flowing through either. The remaining C_H connector has a source connected between its inner and outer contacts, which is adjusted, together with details of the measurement circuitry, to set up these defining conditions. This complete and precise definition, while being somewhat strange at first sight, naturally leads to measuring networks ("bridges") of components joined by coaxial conductors. The definition can be readily understood by recognizing that the impedance Z in the inner circuit is a familiar four-terminal impedance having separate current and potential terminals and that the outer circuit is also a small four-terminal impedance z. The effect of the defining conditions is to make the total impedance of the device as the series combination, $Z + z$. Standard impedances defined in this way are compared in networks in which the current in the coaxial interconnecting conductor of each cable is arranged to be as nearly as possible equal and opposite to that in the outer conductor. Circuitry of this kind has the two important and related properties that the measurements are affected neither by the disposition of cables and components nor by externally generated electric and magnetic fields, including those at the measurement frequency. Further details and an introduction to the literature of this subject have been given by Kibble and Rayner (1984).

8.2.1.2 Realizing the Ohm via Measurements of the Fine-Structure Constant

If certain assumptions are made, the ohm may be derived directly from the quantum Hall effect. Under proper experimental

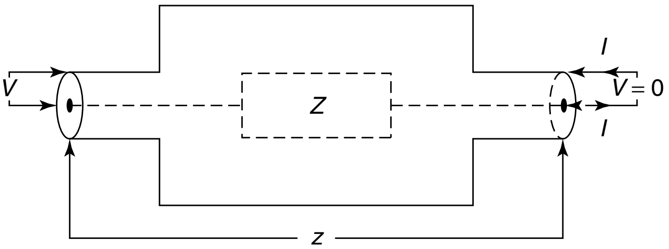

Fig. 8.4 Four-terminal pair definition of an impedance. $Z' = V/I = Z + z$.

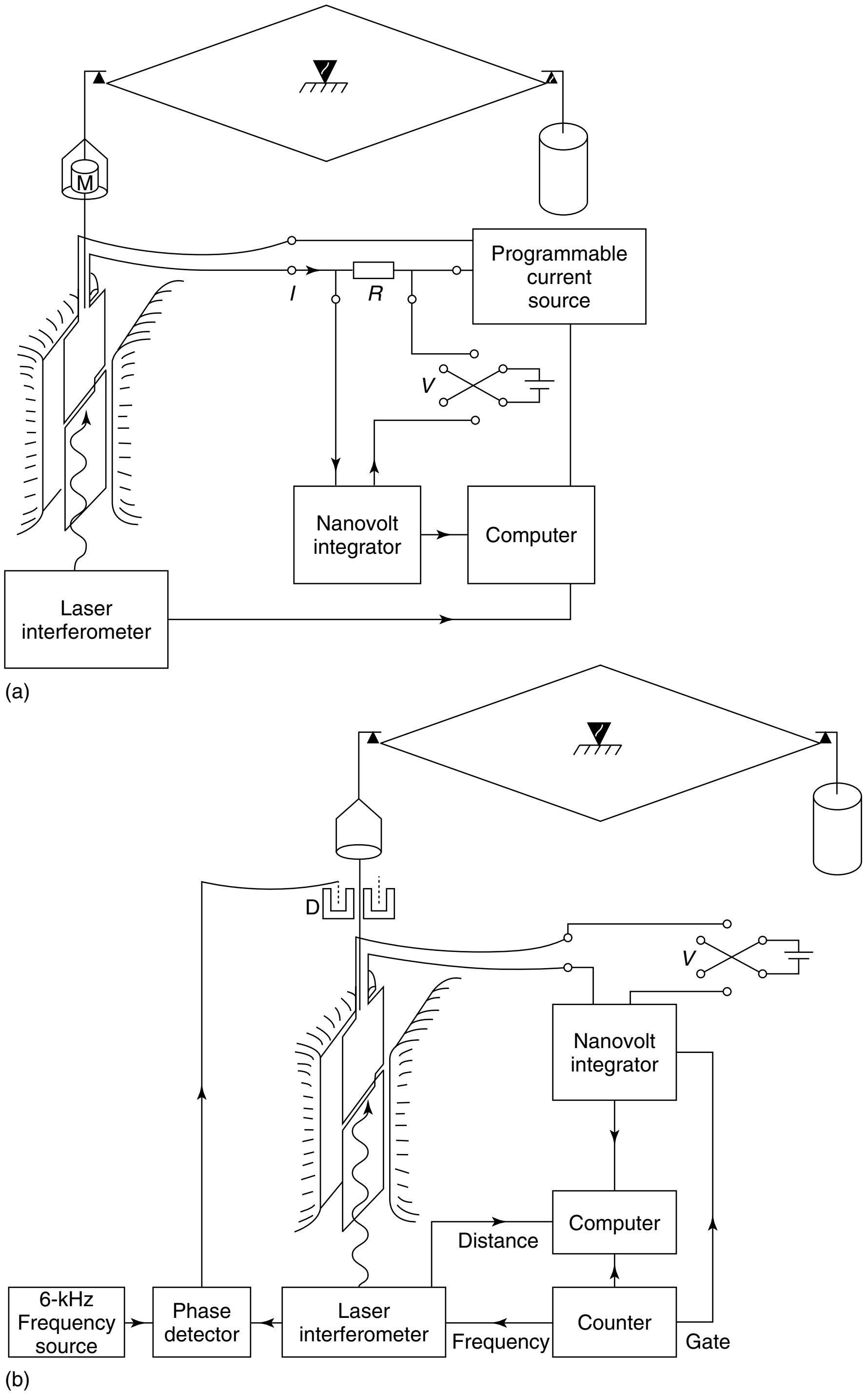
M
Programmable current source
I
R
V
Nanovolt integrator
Computer
Laser interferometer
(a)
D
V
Nanovolt integrator
Computer
Distance
6-kHz Frequency source
Phase detector
Laser interferometer
Frequency
Counter
Gate
(b)

◄ **Fig. 8.5** The moving-coil watt balance. (a) The force on a current from the programmable current source flowing in the coil is opposed by a mass *M*. Any difference between the voltage *V* of a Josephson voltage standard and the voltage across the resistor due to the current is measured by the nanovolt integrator. (b) The nanovolt integrator measures any small difference between the voltage *V* and the voltage induced in the coil when it is moved with constant velocity by the coil-and-magnet drive motor D.

conditions, the current-to-voltage ratio of a device exhibiting this effect is believed to be a multiple of e^2/h, where e is the elementary charge and h is the Planck constant. A measurement of the fine-structure constant, α, in SI units, [$\alpha = (\mu_0 c/2)(e^2/h)$], would therefore constitute a realization of the SI ohm, because the values of μ_0 and c are defined. A very accurate value for α can be obtained from the measurements of the departure of the value of the magnetic moment of the electron from 2 exactly. This departure can be measured as the ratio of two frequencies and calculated (by a very long and involved calculation that requires many hundreds of hours on a powerful computer) (Kinoshita, 2006). The accuracy of the value of e^2/h, and therefore the ohm, obtained in this way is approaching 0.001 ppm, which is more accurate than that obtainable from calculable capacitors and associated bridges. However, this realization does assume that the theories of the quantum Hall effect and the "anomalous" magnetic moment of the electron are accurate to the uncertainty claimed.

8.2.2
The Volt or Watt

Realizing either of these units involves deriving a force from a mass M subject to gravitational acceleration g to balance against an electromagnetic force. g can be measured with a "free-fall" gravimeter (Niebauer *et al.*, 1995) to approximately 0.005 ppm. Mass was involved in realizations of the ampere using a current balance, and indeed this instrument inspired the definition of the ampere. However, because of experimental difficulties primarily associated with measuring the dimensions of the coils, the accuracy of a current balance was limited to about 5 ppm. The instruments described below have replaced it.

8.2.2.1 **The Watt Balance**

In Figure 8.5a, the force produced when a current I flows through a coil suspended from a balance so that the coil lies partly in the field of a large permanent magnet is made equal to Mg. In a separate measurement, the coil is moved vertically with velocity u and an electromotive force V is induced, as illustrated in Figure 8.5b. Making the measurements separately ensures that virtual, rather than real, power is involved and therefore that errors associated with real energy expenditure such as Joule heating in the coil are eliminated. Equality of mechanical and electrical power ensures that

$$IV = Mgu \tag{8.3}$$

If I is related to V via the voltage drop that it produces across a resistor R, then

$$V = \sqrt{MguR} \tag{8.4}$$

If the voltage is measured in terms of the Josephson effect voltage $\nu = hf/2e$ and the current I in terms of v and the quantum Hall resistance $R = h/\,ne^2$ so that $I = f'\,ne/2$, where n is a small integer and f, f' are frequencies, the Planck constant

h can be measured (see Sections 8.3.1 and 8.3.2 below).

Neither the dimensions of the coil nor the density and distribution of the magnetic flux need to be measured. In principle, instantaneous values are required for u and V as the coil passes through the weighing position, but in practice the magnetic flux and coil can be so shaped that Vu^{-1} is almost constant and a value interpolated from measurements made over a finite distance is adequate. If the mean velocity is measured between two instants by counting optical fringes with a gated timer/counter and the mean voltage between the same two instants with a gated integrator, the correlated fluctuations in the actual velocity and voltage caused by vibration are removed from the data.

In the National Physical Laboratory (NPL) realization (Robinson and Kibble, 2007), a current of about 10 mA is controlled to keep the balance in equilibrium. The current is then reversed and a mass M of about 1 kg is placed on the same side of the balance as the coil. The knives and planes are not disturbed, so that the force caused by the current reversal is measured by substitution. Under good conditions, 0.003-ppm repeatability of the mean is obtained.

Then the coil is moved successively upward and downward some 15 times and the average value of uV^{-1} is measured with an accuracy of 0.015 ppm or better as the coil passes through the weighing position. Taking the mean for the two directions of travel eliminates extraneous voltages. The velocity and weighing parts of the realization each contribute to the total uncertainty, and the reproducibility of the result for the watt or for h is about 0.02 ppm.

Another version of this principle of watt realization is in progress at the National Institute of Standards (NIST) (Steiner *et al.*, 2007) and results accurate to 0.036 ppm have been obtained so far.

8.2.2.2 **Electrometers**

Another direct method of relating the mechanical and electrical SI units is to measure the force exerted on the electrode surface of a capacitor when a voltage V is applied to it. Clothier *et al.* (1989) measured the change in force per unit area as a hydrostatic pressure ρgh on a pool of mercury of density ρ when an electric field applied by a plate a distance d above the surface of the mercury was changed from V_1/d_1 to V_2/d_2 (see Figure 8.6a). The mercury pool is surrounded by a moat connected electrically to a guard electrode above it, to establish the unperturbed mercury level corresponding to $V = 0$ from which the height h of the surface of the mercury is raised and is measured. The density of mercury, ρ, was known from the measurements of Cook (1961) at NPL. No capacitance measurements were necessary, because the change in voltage may be calculated from the equation

$$V_2/\sqrt{h_2} - V_1/\sqrt{h_1} = \left(2\rho g/\varepsilon\right)\left(d_2 - d_1\right) \tag{8.5}$$

where $V_{1,2}$ and $d_{1,2}$ are so chosen that $h_1 \approx h_2$; ε is the relative permittivity of the residual vapor above the mercury surface. h was only about 0.6 mm (equivalent to 2000 optical fringes), but nevertheless this realization reached an accuracy of ±0.32 ppm.

Other realizations of similar accuracy have involved an electrode suspended from a balance, as shown in Figure 8.6b. The force F_x and capacitance changes resulting from the displacement of a second electrode by a measured distance yield V in SI volts, because

$$V^2 = \frac{2F_x}{\partial C/\partial x} \tag{8.6}$$

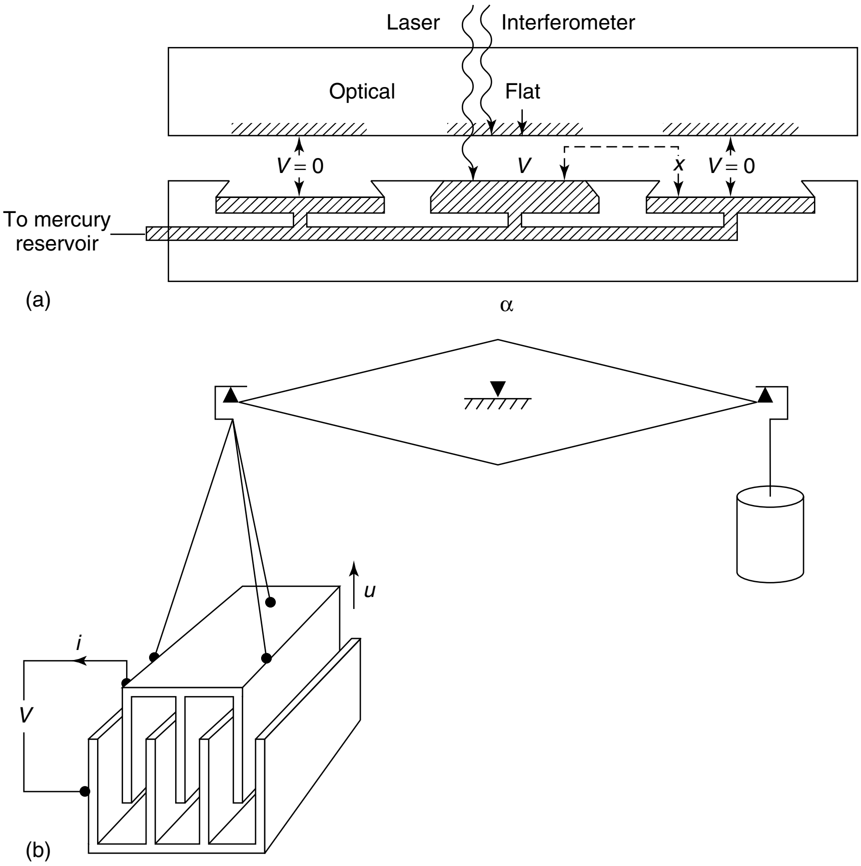

Fig. 8.6 (a) A mercury-pool electrometer. (b) An interleaved plate electrometer.

$\partial C/\partial x$ is obtained either from the dimensions of the apparatus by calculation (Elnekave, Fau, and Delahaye, 1976) or by measurement of capacitance (Bego, 1988; Funck and Sienknecht, 1988), in terms of a calculable capacitor. Because a watt balance seems to offer better accuracy, work on electrostatic balances has ceased for the present but it is also possible to proceed via the virtual work principle that lies behind the watt balance realization. The magnitudes of the quantities involved are less favorable than in a watt balance apparatus, and a measurement having useful accuracy has yet to be made by this technique.

8.2.2.3 **Defining the Kilogram via Avogadro's Constant**

A seemingly straightforward way to define the kilogram in terms of a fundamental constant would be to specify it as the mass of a (very large) number of atoms. Counting such a large number can be accomplished if the atoms are arranged in a crystal whose structure is as nearly perfect as possible (Becker, 2003). The refining techniques that have been developed for silicon make it the best candidate at present. The number of atoms in an almost perfectly spherical artifact of silicon can be calculated from the distance between atoms and the mean diameter of the sphere. The distance between atoms is measured by a device in which a laminar plate is moved a measured distance between two fixed plates while the passage of planes of atoms is counted by observing Moiré fringes in the passage of a collimated X-ray beam that passes perpendicularly through the plates. Obtaining and maintaining subatomic alignment of the planes while translating the middle plate through a few millimeters demands the ultimate precision from strip-hinge technology.

This atom-counting procedure constitutes a measurement of Avogadro's constant, N_A which is connected to Planck's constant, and therefore to the watt balance approach to the kilogram, by the expression

$$N_A = c_0\alpha^2 A_r(\mathrm{e})M_u/(2R_\infty h) \tag{8.7}$$

where the other constants, c_0, the velocity of light in free space, α, the fine-structure constant, $A_r(e)$, the relative atomic mass of the electron, M_u, the molar mass constant and R_∞, the Rydberg constant, have either defined or more accurately known values.

There are two major problems in this approach. First, the mass of a silicon atom is not unique because naturally occurring silicon consists of three isotopes in ratios of approximately 92 : 5 : 3. Determining the ratio of these amounts accurately enough is very difficult, but work is in progress to create artifacts from crystals consisting almost entirely of a single isotope. Second, the surface of a silicon sphere is covered with a layer a few atoms thick of silicon oxides and the thickness of this layer has to be accurately measured by ellipsometry to account for the mass discrepancy it produces.

At the time of writing (2007), there is up to one part-per-million discrepancy between the silicon crystal and watt balance approaches to redefining the kilogram, but it is hoped that this discrepancy will be resolved in the next few years.

8.2.2.4 **Self-Consistency of the SI**

The continuing progress toward basing the SI on the (presumed) unchanging values of fundamental physical constants has raised the question as to whether the von Klitzing constant is exactly h/e^2, the Josephson constant is exactly $2e/h$ and whether electron-counting experiments transport charge in elements of e exactly. Strong evidence for the validity of these assumptions would be obtained from experiments, which examine the closure of the so-called ''metrological triangle'' in which a current measured in counted electrons is passed through a resistance measured by the quantum Hall effect in terms of the von Klitzing constant and the resulting voltage drop measured against a Josephson effect voltage standard. That is, whether

$$(e) \times (2e/h) \times (h/e^2) = 2 \tag{8.8}$$

within experimental error. In principle, a single experimental arrangement in which electrons are counted, pass through a quantum Hall device and the resulting voltage is opposed to a Josephson junction voltage, will examine this. No other units are required provided that the radio frequencies needed to count out the electrons and generate the Josephson voltage are scaled from the same frequency source. At present, the 1-ppm accuracy with which electrons can be counted is a limitation, but this is likely to improve substantially in the future.

8.3 Working Standards of Voltage and Resistance

8.3.1 A Standard for the Volt

We can establish a representation of the SI volt from the ac Josephson effect in which microwaves of frequency f are shone on a junction formed from an insulating barrier a few nanometers thick between two superconductors (see Figure 8.7). A current of electron pairs tunneling through the barrier gives rise to a voltage $V = nf(h/2e) = nf/K_J$ across this insulated junction of superconductors, where n is a small integer, h is the Planck's constant, e is the electronic charge, and $K_J(= 2e/h)$ is the Josephson constant. A measurement of f and knowledge of n gained from the approximate value of the voltage enables V to be derived in terms of a conventional value assumed for $2e/h$. This value is so chosen that V is then expressed in terms of the SI volt as given by the best realizations available. If a single junction is used, V is only a few millivolts and has to be scaled up to 10 V for practical dissemination either by a potentiometric ratio method often involving Hamon resistors or by using a superconducting current comparator (described below in connection with the quantum Hall effect). The technology of connecting many Josephson junctions in a series array so that each is subjected to nearly the same incident microwave power has been developed (Niemeyer, Hinken, and Kautz, 1984; Hamilton *et al.*, 1985), and any multiple of 150 μV up to a few volts can be established. Hundred and fifty microvolts corresponds to the 72.5-GHz microwave radiation for which the array is designed. An array provides a greatly simplified and more accurate voltage standard than the use of a single junction. Modern arrays employ a superconductor fabrication technology, which leads to a nonhysteretic $I - V$ characteristic. The integer n, equal to the number of junctions, is stable and only

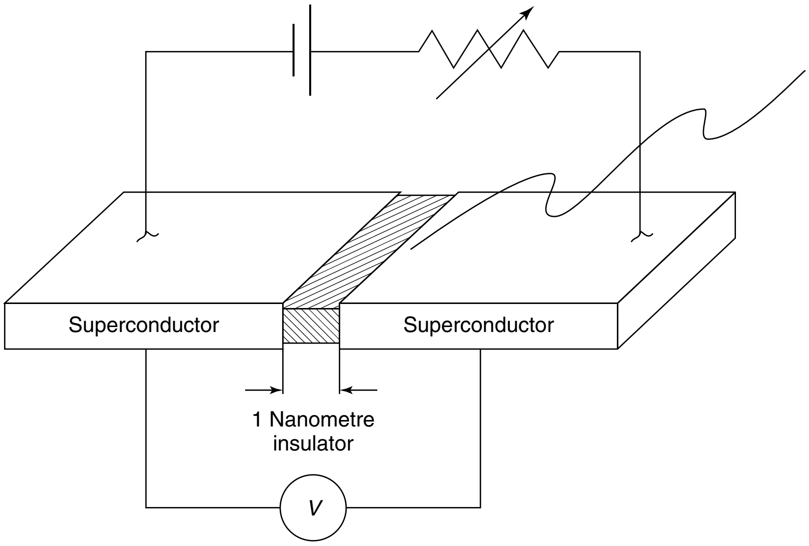

Fig. 8.7 A single Josephson junction.

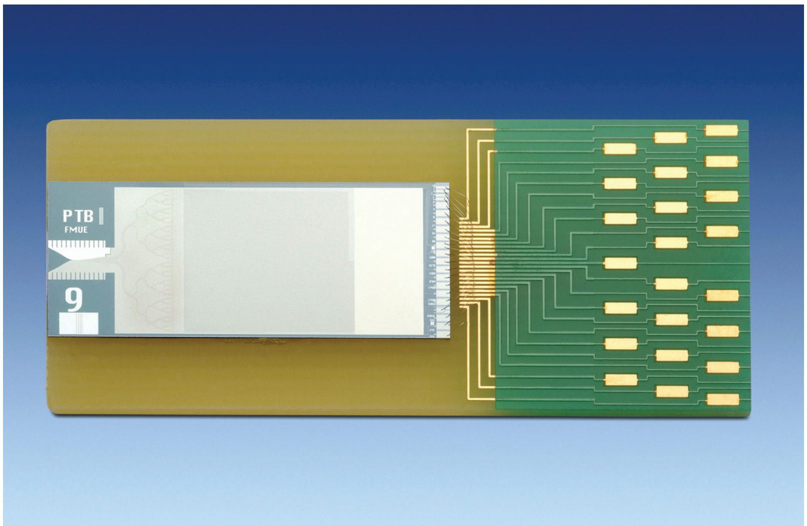

Fig. 8.8 A Josephson multijunction array.

needs to be determined once. Arrays consisting of subarrays with junction numbers forming a binary sequence can be used as programmable quantum digital-to-analog converters and provide a quantum standard for low-frequency ac voltages. (Hamilton *et al.*, 1989; Pöpel *et al.*, 1990) Figure 8.8 shows a modern 10-V array manufactured at the Physikalisch-Technische Bundesanstalt (PTB), Germany.

Using an array of Josephson junctions, calibrations reproducible to 0.01 ppm are readily achieved, the principle limitation being the variability of the thermal emfs arising from the thermal gradients along the conductors connecting the array, which is at a cryogenic temperature, to the room-temperature terminals. To eliminate thermal and other spurious emfs, both voltage sources being compared must be capable of reversal. Taking the mean of the comparison before and after reversal then eliminates all unwanted emfs, provided they do not vary significantly during this short time. A Josephson array can be made to give a voltage of either polarity by presenting its terminals with a voltage of the required magnitude and polarity while increasing the microwave power from zero.

The resolution with which voltage sources can be compared has recently been greatly increased by the availability of electronic null detectors. The noise these contribute to the measurement does not add significantly to the irreducible Johnson noise arising from source resistance. The semiconducting transistor and diode junctions in these electronic detectors are, however, prone to rectify any interference voltage appearing at their input terminals, leading to systematic errors.

8.3.2
A Standard for the Ohm

The quantum Hall effect affords precisely reproducible voltage-to-current ratios, and these constitute the best representation of the SI ohm now available. Both silicon MOSFETs and heterostructures of gallium arsenide-gallium aluminum arsenide at liquid helium temperatures provide the necessary conditions to establish a two-dimensional electron gas in a boundary layer of a few nanometers thick. When a strong $(1 - 14T)$ magnetic field is applied in a direction perpendicular to this layer, the ratio of the voltage between the sides of the layer to the current flowing across the layer (i.e., the Hall "resistance" R_H) is found to take one of a number of discrete values depending on the strength of the applied magnetic field (see Figure 8.9). The peculiar quantum-mechanical nature of the device is such that any two-terminal measurement of the voltage-to-current ratio between any two points on the boundary of the layer reveals a true resistance equal to the step resistance plus that any contact resistances involved. By contrast, in a four-terminal measurement when the current and potential terminals alternate around the boundary, the voltage-to-current ratio equal to R_H is observed. The theory of the effect indicates that $R_H = h/ie^2$, where i is an integer, as the temperature approaches absolute zero for certain ranges of magnetic field values. Measurements made so far support this theory, at least to an accuracy of 0.01 ppm. Moreover, the values of R_H have been measured to be independent of the particular device used to generate them to within 0.00035 ppm.

Despite its purely superficial resemblance to the Josephson effect used as a voltage standard, the physical origins of the quantum Hall effect are quite different (Hartland, 1988), and in particular it is important to realize that the simple expression for R_H given above is exact only at the unattainable absolute zero of temperature, $T = 0$. However, fortunately, the approximation is sufficiently good for all practical purposes at temperatures between 0.3 and 4.2 K, depending on the particular device. A measure of the error involved is simply obtained by measuring the voltage drop along the direction of current flow, using extra connections made on the

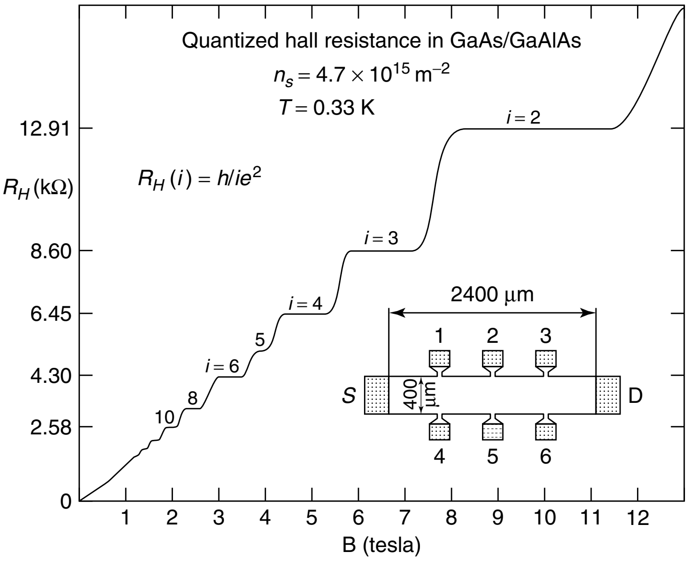

Fig. 8.9 A graph showing the variation of the voltage-to-current ratio as a function of magnetic field for the quantum Hall device shown in the inset.

same side of the layer between the current connections. This voltage drop is zero only at $T = 0$, but at finite temperatures its ratio to the perpendicular Hall voltage is of the same order of magnitude as the departure from the above expression for R_H. In practice, the temperature can usually be made low enough for the error to be less than 0.01 ppm.

The best resolution currently being obtained in relating R_H to the value of a conventional standard resistor at room temperature is a few times 0.001 ppm.

The measurement circuit used to attain the greatest accuracy is based on a superconducting current comparator (Harvey, 1972), a device whose accuracy depends on the Meissner effect of exclusion of magnetic flux from a superconductor. A long (in principle, infinite), superconducting tube is threaded by two windings of n_1 and n_2 turns carrying the two currents to be compared. The Meissner effect then ensures that any imbalance in their ampere-turns appears as a current flowing over the whole inside and outside surface of the tube so that the total current flowing through the inside of the tube is zero. The magnetic flux from the imbalance current $(n_1 I_1 - n_2 I_2)$ that flows uniformly back around the outside of the tube is independent of the positions of the windings within the tube and can be sensed with great sensitivity by a pickup coil coupled to a SQUID flux detector . A dc SQUID detector circuit contains a Josephson junction near which is a radio-frequency coil that can sense the changing impedance of the circuit when quanta of magnetic flux thread it. An ac SQUID circuit contains two Josephson junctions connected into a superconducting loop and the voltage across the loop is periodic in the number of flux quanta threading it. A dc SQUID is now more frequently employed, as it has greater sensitivity. By registering the state of the SQUID circuit when both the currents in n_1 and n_2 are zero, and maintaining this state with a feedback locking circuit as these currents are increased to their final value, flux balance from these currents is ensured so that $n_1 I_1 = n_2 I_2$. The need for an infinitely long superconducting tube is avoided by arranging its

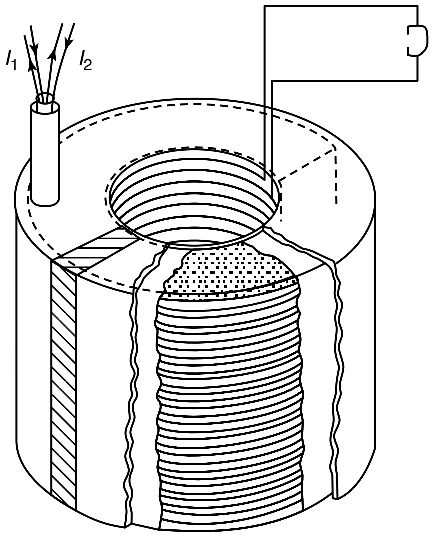

Fig. 8.10 A superconducting current comparator.

topology to be that of a snake swallowing its own tail for about three convolutions. The construction of a practical device is illustrated in Figure 8.10. The n_1 and n_2 turns carrying the currents I_1 and I_2 are wound on a bobbin, which is enclosed in a convoluted superconducting lead-and-insulation shield. The ends of the windings are taken out through a similarly constructed side tube whose layers are connected to the corresponding layers of the convoluted shield. The magnetic field caused by imbalance in the ampere-turns of the windings threads the center of the insulated bobbin where it is sensed by a coil connected to the SQUID.

The resistor and the quantum Hall device to be compared are connected to the windings of the current comparator as shown in Figure 8.11 with a direct connection between one set of potential terminals and a sensitive null voltage detector D between the others. When the voltage at the detector is nulled, either by a shim current in an auxiliary winding on the current comparator or by shunting one of the resistors with an adjustable high resistance, $I_1 R_H = I_2 R$ or $n_1/n_2 = R_H/R$.

As R_H can be equal to 25 812 Ω, 6453 Ω, and so on, 100 Ω is a good choice for the value of R_2. Hundred-ohm resistors are better than the 1-Ω resistors used hitherto as the starting point for comparing with the decade values of resistance in routine calibration work, both because of the relatively smaller effect that thermoelectric potentials have on the measurement circuit and because matching to the noise properties of null detectors is easier.

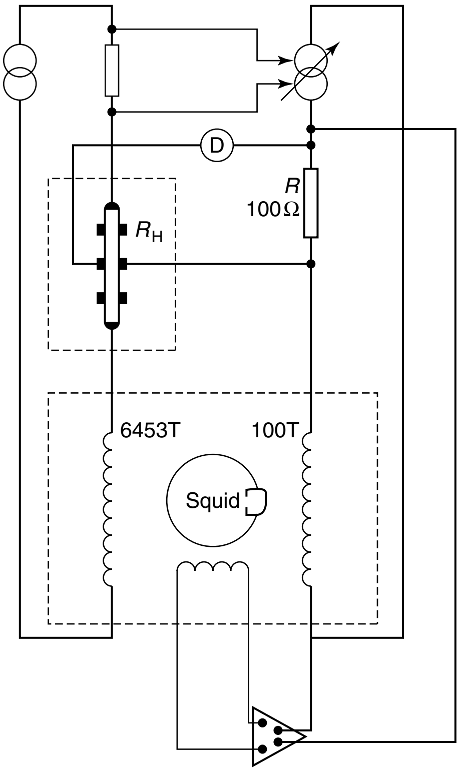

Fig. 8.11 A measurement circuit for comparing R_H with a resistor at room temperature.

8.3.3
Conventional Values Adopted for K_J R_K to Represent the SI Volt and Ohm

From 1 January 1990, the General Conference on Weights and Measures agreed to adopt exact conventional values $K_{J-90} = 483597.9\,\mathrm{GHz\,V^{-1}}$ and $R_{K-90} = 25812.807\ \Omega$, respectively, for the Josephson and von Klitzing constants to represent the SI volt and Ohm for international conformity in ordinary measurements. It must be understood that if the actual value of K_J or R_K in SI units is required, for example, for the comparison of measured values of fundamental constants, an additional uncertainty must be ascribed to the measurements.

8.4
Other Primary Standards of National Measurement Institutes

8.4.1
dc Voltage

Other primary working voltage standards, such as Zener diodes and Weston standard cells, can be measured against an array of Josephson junctions. An array yields a representation of the SI volt by using the conventionally agreed value for the frequency-to-voltage conversion constant governing the Josephson effect. Higher and lower dc voltages are easily measured using the measured resistance ratios of resistive voltage dividers.

8.4.2
dc Resistance

Stable wire resistance standards of (usually) 100-Ω value can be calibrated directly in terms of the quantum Hall effect using the verifiably accurate current ratios created by a superconducting current comparator. Standards of other decade values can be derived either by a superconducting current comparator or by Hamon resistors (see Section 8.2.1.1), or simply by connecting 10 resistors in series to generate standards of 10 kΩ and greater. Typical accuracies of decade values derived in this way are given in Table 8.1.

8.4.2.1 The ac Behavior of Standard Resistors

An ideal resistive circuit component under ac conditions will have the potential across it exactly in phase with the current passing through it. Its resistance will be independent of the ac frequency. The value of a real resistor $R(f)$ will be its resistance at the frequency f, and will depend only slightly on f through causes such as the tendency of the current to be deflected outwards to a path of smaller cross section at the surface of the conductor by its associated magnetic field (the ‘‘skin effect’’) or the ‘‘proximity effect’’ caused by additional loss mechanisms associated with currents magnetically induced in neighboring conductors. The magnetic field of these induced currents couples back into the original conductor and generates an opposing in-phase emf.

A practical resistor will also have small self-generated modifying currents and

Tab. 8.1 Certified accuracies in parts per million of dc resistances.

Resistance (Ω)	10^{-3}	10^{-2}	10^{-1}	1	10	10^2	10^3	10^4	10^6	10^9	10^{12}
Accuracy	0.8	0.6	0.3	0.15	0.25	0.07	0.13	0.15	1.3	10	500

Tab. 8.2 Certified accuracies in parts per million of capacitances at 1 kHz.

Capacitance	1 pF	10 pF	1 nF	10 nF	1 μF	10 μF	1 mF	1 F
Accuracy	1	0.7	3	30	60	100	500	10 000

voltages at 90°-phase associated with its accompanying magnetic and electric fields. These effects are quantified as the self inductance, L, and self capacitance, C, of the resistor and their effects can, up to quite high frequencies, be described by a single parameter, the time constant τ of the resistor. The ac impedance of the resistor is

$$Z(f) = R(f) + j \times 2\pi f \tau \tag{8.9}$$

τ is equal to LR^{-1}, or $-RC$ if the self capacitance or inductance, respectively, is negligible compared to the other. τ can be as small as 10^{-9} s, particularly for resistors in the range of values from 10 to 1000 Ω when the inductive and capacitive effects tend to cancel. τ is measured in a bridge network by comparison or substitution with standards having known or negligible values of τ, measured in quadrature bridge networks from well-designed gas-filled capacitors having very small departures from 90°-phase between current and voltage.

8.4.3
Capacitance

Primary standards of capacitance are calibrated either via an artifact – a very stable fixed-value capacitor whose value is established from time to time directly in terms of the Thompson–Lampard calculable capacitor – or from a resistance measured in terms of the quantum Hall effect, using a resistance-capacitance ac bridge (Kibble and Rayner, 1984). The value of the best capacitance standards can be constant to within 0.1 ppm over a few years. Either route depends on coaxial ac bridge techniques, which imply complete electrical definition of impedance standards at kilohertz frequencies. Coaxial ac bridge techniques are also used to scale the value of a primary capacitance standard to values of other standards at decade intervals from 1 pF to 1 F with the accuracies given in Table 8.2.

8.4.4
Inductance

Inductance is derived from capacitance. Either the parallel-resonance technique illustrated in Figure 8.12a is employed, in which the reactances of a capacitor and an inductor are observed to be equal at a particular measured frequency, or the ratio of these reactances is measured using the Maxwell–Wien ac bridge illustrated in Figure 8.12b. An inductive ratio-arm ac bridge can scale inductance to other decade values with little further loss of accuracy, and the typical 100-ppm or better accuracies obtained are sufficient for all present scientific or commercial needs.

8.4.4.1 The Frequency Dependence of Standard Inductors

Skin and proximity effects, by excluding magnetic flux from the conductors and any constructional metalwork of an inductor, will alter its value slightly in a frequency-dependent manner. Its reactance will also be modified by the capacitance associated with its surrounding electric field. These effects are investigated in calibrating standard inductors by comparing them with specially constructed inductors wound with thin, spaced conductors

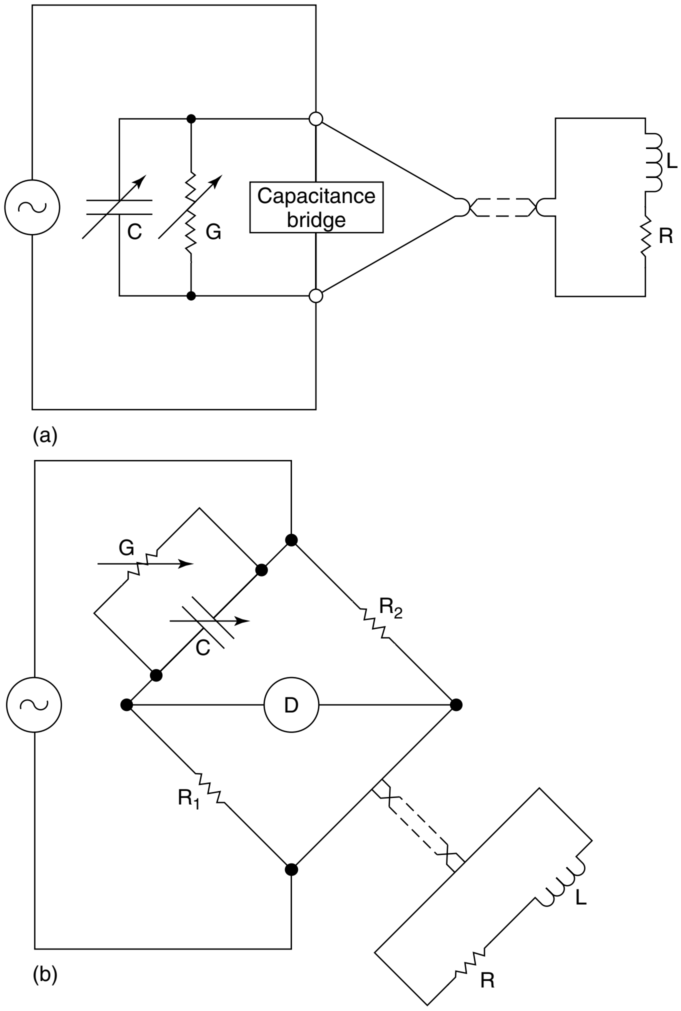
C
G
Capacitance bridge
L
R
(a)
G
R$_2$
C
D
R$_1$
L
R
(b)

Fig. 8.12 (a) The parallel-resonance technique for establishing inductance *L* and associated series resistance *R* in terms of capacitance *C* and conductance *G*. (b) A Maxwell–Wien bridge. The screening that eliminates stray capacitances from the measurement has been omitted from the diagrams.

whose self capacitance has been measured by observing their self-resonant frequency when the terminals are shorted.

8.4.5
Electric Fields and Dielectric Constants

Electrical field strengths can be inferred directly in volts per meter from the voltages on conducting surfaces between which the electric field is generated, because conducting surfaces define the field-boundary conditions accurately.

The relative permittivity of materials is measurable in terms of the change of capacitance that a sample of the material produces when inserted in a suitable test capacitor. Permittivities are also needed at frequencies as high as in the microwave region, and are then measured by a waveguide insertion or an interferometric technique (Kemp *et al.*, 1996).

8.4.6
Magnetic Quantities

In contrast to electric flux and permittivity, magnetic flux cannot in general be deduced with any accuracy from the currents generating it if the boundary conditions at the surfaces of any magnetic media are involved (the superconductivity Meissner effect excluded), and the units and basic measurement of total flux and flux density of magnetic fields must be provided in other ways.

8.4.6.1 Total Magnetic Flux

The SI unit is the weber, and it is defined and realized in terms of the emf V produced around a circuit resulting from the rate of change of the total magnetic flux in webers threading it. That is,

$$V = -d\Phi/dt \qquad (8.10)$$

which may be integrated to yield

$$\Phi = \int V dt \qquad (8.11)$$

Therefore, the weber can be realized by directly measuring total flux by integrating the voltage induced in a search coil, which encompasses all the flux. The voltage integral is measured as the flux is established, or as the coil is moved into position from a region where the flux is negligible, with an electronic integrating circuit that is calibrated in terms of a known direct current charging a capacitor of known value. If the flux is caused by an alternating current of known wave form, the voltage integral can either be measured directly as a mean rectified value or be inferred from the wave form and a measurement of the peak or root mean square voltage.

8.4.6.2 Magnetic Flux Density

This can be measured using a similar search coil method in which the search coil is small enough to encompass only a region over which the flux is uniform. The flux density measurements made by the search coil can be calibrated either by measuring the voltage induced in it when it is inserted into, or its sense is reversed in, a known flux density (derived from a current-carrying coil of measured geometry).

A much more accurate indirect method is to observe the precessional frequency of protons in the flux (Williams *et al.*, 1985).

8.4.7
ac Current and Voltage

These are not ‘‘basic’’ quantities in that they do not require separate definition of their units, but relating them at the highest accuracy level to the corresponding dc

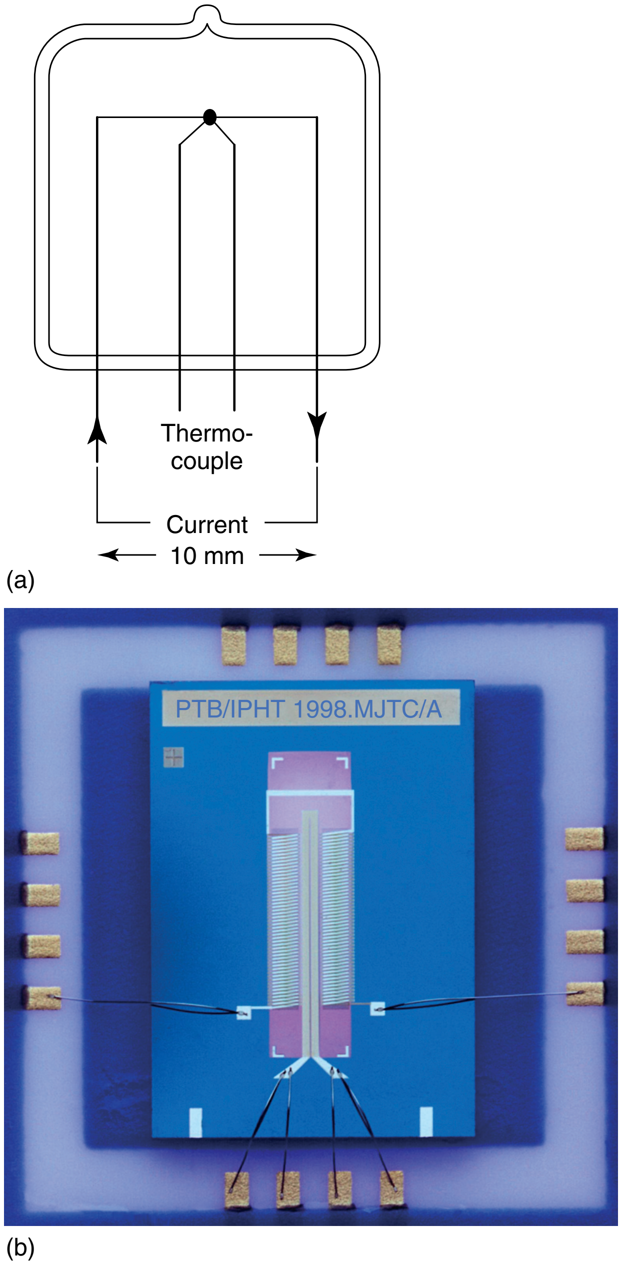
Thermo-
couple
Current
10 mm
(a)
PTB/IPHT 1998.MJTC/A
(b)

Fig. 8.13 A thermal convertor for relating ac and dc power: (a) design and (b) picture.

quantities does need a specialized measurement technique that is only to be found in national measurement laboratories.

The basic concept is to equate the heating of a dissipating element by a direct current with the same heating power generated by an alternating current. To ensure that the distribution of heat in the dissipating element is independent of frequency up to at least 100 kHz (i.e., eddy-current, capacitive, and inductive effects are negligible), the physical dimensions of the element are kept very small – usually by making it from a fine, straight length of resistance wire. The design of a typical single-element device is illustrated in Figure 8.13 where the temperature of a heater wire is sensed by a thermocouple attached to its center by an insulating glass bead. Sensitivity and accuracy can be somewhat enhanced, at the expense of high-frequency performance, by putting several thermocouple temperature sensors in series along a heater element (Dix and Bailey, 1975). Devices deposited on a substrate by integrated circuit techniques have now superseded these wire-and-bead convertors (Klonz and Weimann, 1989; Scarioni, Klonz, and Funck, 2005).

If the two currents are also compared via the voltages they produce in turn across a resistor whose value is independent of frequency, then the alternating and direct currents and voltages can be separately related.

Glossary

ac Josephson Effect: A phenomenon in which microwaves of frequency f are shone on a thin insulating barrier between two superconductors. A current of electron pairs tunneling through the barrier interacts with the nth harmonic of the microwave frequency, resulting in a precise voltage equal to $nf(h/2e)$ appearing across the barrier, where e is the electronic charge and h is the Planck constant.

Farad: The electrical capacitance between electrodes having a potential difference of 1 V when charged with 1 C.

Quantum Hall Effect: This effect gives rise to fixed ratios of transverse voltage to the current flowing through a semiconducting device in which the electrons are confined to two-dimensional motion at very low temperatures in a strong perpendicular magnetic field. These ratios are often termed as *quantized Hall resistances*.

Realize (a Unit): Use appropriate apparatus to derive a value for a standard consistent with the abstract definition of a unit. This standard can then be used to calibrate further standards and instruments.

References

Becker, P. (**2003**) Determination of the Avogadro constant via the silicon route. *Metrologia*, **40**, 366–375.

Bego, V. (**1988**) Determination of the volt by means of voltage balances. *Metrologia*, **25**, 127–133.

Clothier, W.K., Sloggett, G.J., Bairnsfather, H., Currey, M.F., and Benjamin, D.J. (**1989**) A determination of the volt with improved accuracy. *Metrologia*, **26**, 9–46.

Cook, A.H. (**1961**) A determination of the density of mercury. *Philos. Trans. R. Soc.*, **A254**, 125–154.

Dix, C.H. and Bailey, A.E. (**1975**) Electrical standards of measurement: 1. DC and low frequency standards. *Proc. IEEE*, **122**, 1018–1036.

Elnekave, N., Fau, A., and Delahaye, F. (**1976**) Evaluation provisoire du coefficient 2e/h à Partir de la dètermination absolute de volt. *BNM*, **7**, 3–8.

Funck, V.T. and Sienknecht, V. (**1988**) Darstellung der spannungseinheit volt mit der PTB-Spannungswaage. *Elecktrizitätswirtschaft*, **87**, 104–106.

Hamilton, C.A., Kautz, R.L., Steiner, R.L., and Lloyd, F.L. (**1985**) A practical Josephson voltage standard at 1*V*. *IEEE Device Lett.*, **EDL-6**, 623.

Hamilton, C.A., Lloyd, F.L., Chieh, K., and Goeke, W.C. (**1989**) A 10-V Josephson voltage standard. *IEEE Trans. Instrum. Meas.*, **38** (2), 314–316.

Hamon, B.V. (**1954**) A 1–1000~Ohm built-up resistor for the calibration of standard resistors. *J. Sci. Instrum.*, **31**, 450–453.

Hartland, A. (**1988**) Quantum standards for electrical units. *Contemp. Phys.*, **29**, 477–498.

Harvey, I.K. (**1972**) A precise low temperature DC ratio transformer. *Rev. Sci. Instrum.*, **43**, 1626–1629.

Josephson, B.D. (**1962**) Possible new effects in superconductive tunnelling. *Phys. Lett.*, **1**, 251–253.

Kemp, I. *et al.* (eds) (**1996**) Special issue on measurement techniques. IEEE Proceedings of the Seventh International Conference on Dielectric Materials, Measurements and Applications (Conf. Publ. No.430).

Kibble, B.P. and Rayner, G.H. (**1984**) *Coaxial AC Bridges*, Adam Hilger, Bristol.

Kinoshita, T. (**2006**) Improved α^4 term of the electron anomalous magnetic moment. *Phys. Rev. D*, **73**, 013003-1–013003-28.

von Klitzing, K., Dorda, G., and Pepper, M. (**1980**) New method for high accuracy determination based on quantized Hall resistance, *Phys. Rev. Lett.* **45**, 494–497.

Klonz, M. and Weimann, T. (**1989**) Accurate thin film multijunction thermal converter on a silicon chip [AC-DC standard]. *IEEE Trans. Instrum. Meas.*, **38** (2), 335–337.

Mills, I.A., Mohr, P.J., Quinn, T.J., Taylor, B.N., and Williams, E.R. (**2006**) Redefinition of the kilogram, ampere, Kelvin and mole: a proposed approach to implementing CIPM recommendation 1 (CI-2005). *Metrologia*, **43**, 227–246.

Niebauer, T.M., Sasagawa, G.S., Faller, J.E., Hilt, R., and Klopping, F. (**1995**) A new generation of absolute gravimeters. *Metrologia*, **32**, 159–180.

Niemeyer, J., Hinken, J.H., and Kautz, R.L. (**1984**) Microwave-induced constant voltage steps at one volt from a series array of Josephson junctions. *Appl. Phys. Lett.*, **45**, 478–480.

Pöpel, R., Niemeyer, J., Fromknecht, R., Meier, W., and Grimm, L. (**1990**) 1- and 10-V series array Josephson voltage standards in Nb/Al2O3/Nb technology. *J. Appl. Phys.*, **68**, 4294–4303.

Robinson, I.A. and Kibble, B.P. (**2007**) An initial measurement of Planck's constant using the NPL Mark II watt balance. *Metrologia*, **44**, 427–440.

Scarioni, L., Klonz, M., and Funck, T. (**2005**) Quartz planar multijunction thermal convertor as a new ac-dc current transfer standard up to 1 MHz. *IEEE Trans. Instrum. Meas.*, **54** (2), 799–802.

Steiner, R.L., Williams, E.R., Newell, D.B., and Liu, R. (**2007**) Uncertainty improvements of the NIST electronic kilogram, *IEEE Trans. Instrum. Meas.* **56** (2), 592–596.

Thompson, A.M. and Lampard, D.G. (**1956**) A new theorem in electrostatics and its application to calculable standards of capacitance. *Nature*, **177**, 888.

Williams, E.R., Jones, G.R., Song, J.S., Phillips, W.D., and Olsen, P.T. (**1985**) Report on the new NBS determination of the Proton Gyromagnetic Ratio. *IEEE Trans. Instrum. Meas.*, **IM-34**, 163–167.

Further Reading

Deslattes, R.D. and Kessler, E.G. (**1991**) Status of a silicon lattice measurement and dissemination exercise. *IEEE Trans. Instrum. Meas.*, **IM40**, 92–97.

Kibble, B.P. (**1981**) New amps for old. *Electron. Power*, **21**, 136–139.

Kibble, B.P. and Hunt, G.J. (**1979**) A measurement of the gyromagnetic ratio of the proton in a strong magnetic field. *Metrologia*, **15**, 5–30.

Petley, B.W. (**1985**) *The Fundamental Physical Constants and the Frontier of Measurement*, Adam Hilger, Bristol.

Samuel, M. (**1988**) Samuel Replies (He replies to a comment on his original article in PRL57). *Phys. Rev. Lett.*, **61**, 2899.

Seyfried, P. *et al.* (**1987**) The Avogadro Constant – recent results on the Molar Volume of Silicon. *IEEE Trans. Instrum. Meas.*, **IM-36**, 201–204.

1
Electromagnetic Radiation

Wendell T. Hill, III

1.1 Introduction 3
1.2 The Spectrum of Light 7
1.3 Basics of Electromagnetic Waves 8
1.3.1 Maxwell's Equations 9
1.3.2 Wave Equation 10
1.3.2.1 Plane Waves 10
1.3.2.2 Scalar Harmonic Waves 11
1.3.2.3 Waves with Curved Phase Fronts 12
1.4 Energy, Intensity, Power, and Brightness 12
1.5 Polarization 13
1.5.1 Polarization Bookkeeping 15
1.5.2 Jones Matrices 15
1.5.3 Mueller Matrices 15
1.6 Longitudinal Field Component 16
1.7 Diffraction 17
1.8 Interference 18
1.8.1 Superposition: Single Frequency 18
1.8.1.1 Interferometry 20
1.8.2 Superposition: Multiple Frequencies 20
1.8.3 Short Pulses 22
1.9 Photons and Particles 24
References 24
Further Reading 25

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

> *In the beginning … darkness was upon the face of the deep.*
> *And God said, Let there be light: and there was light.*
>
> Genesis 1:1-3

1.1
Introduction

Light has been a trusted probe of a variety of aspects of the universe since the beginning of scientific inquiry. Today, regardless of whether searching for gravitational waves, exploring the fundamental properties of quantum mechanics, or designing metamaterial, light continues to play a critical role in revealing nature and engineering tools to enhance life. In this chapter, we review a few key elements of classical and quantum light. A comprehensive review of light is well beyond the scope of this chapter. Thus, we have chosen to focus on the properties that are most often encountered in the laboratory while providing some context and history.

Light, an electromagnetic (EM) field, is an intimate coupling between time-dependent electric and magnetic fields. Classically, the EM field is described quantitatively through Maxwell's equations (Section 1.3.1), where it can be viewed as a wave1) – a disturbance – satisfying the wave equation,

$$\nabla^2\Psi = \varepsilon\mu\frac{\partial^2\Psi}{\partial t^2} \qquad (1.1)$$

In Eq. (1.1), ε and μ are the permittivity and permeability of the medium through which the light is traversing ($\sqrt{1/\varepsilon\mu} = v$ is the speed of light or phase velocity in the medium), and ∇^2 is the Laplacian.2) The form of the solution depends on the coordinate system – rectangular, spherical, and so on – in which ∇^2 is expressed, as we discuss in Section 1.3.2. However, in general, the solution for the so-called running wave in one dimension is

$$\Psi(\mathbf{x}, t) = f(\mathbf{x} \pm vt) \qquad (1.2)$$

---

1) Christiaan Huygens, a contemporary of Isaac Newton, viewed light as a wave prior to the mathematical formulation as we now know it. Newton, on the other hand, was convinced that light was a stream of corpuscles.

2) The Laplacian is shorthand for

$$\nabla \cdot \nabla$$

which can be written as

$$\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

in rectangular coordinates.

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

where **x** describes the distance that the disturbance moves as t increases. The "−"("+") sign indicates motion in the positive (negative) **x** direction.

In vacuum, $\varepsilon \to \varepsilon_0$, $\mu \to \mu_0$, and $v \to c = \sqrt{1/\varepsilon_0\mu_0}$, the vacuum light speed. Special relativity tells us that c sets a "speed boundary" across which information cannot flow. Specifically, those of us living in a "sub-c" universe are prohibited from achieving speeds equal to or larger than c as well as those living in a "super-c" universe from speeds lower than c. Light waves in vacuum are very special and differ from other waves that we encounter in everyday life. There is no rest frame for light and light travels at the same speed in all frames.

As is true of all waves, light is characterized by a wavelength (λ) and a frequency (ν). In vacuum, these quantities are linked by c,

$$\nu \equiv \frac{c}{\lambda} \tag{1.3}$$

In media different from vacuum $v = c/n$ where,

$$n = \sqrt{\left(\frac{\varepsilon}{\varepsilon_0}\right)\left(\frac{\mu}{\mu_0}\right)} \tag{1.4}$$

is the index of refraction and the quantities in parentheses are the electric and magnetic dielectric constants respectively; μ differs slightly for μ_0 for most cases of interest. The more general relationship between λ and ν is

$$\nu = \frac{v}{\lambda} = \frac{c}{\lambda_0} \tag{1.5}$$

where λ_0 is defined as the vacuum wavelength. By definition, ν is medium independent and maintains its vacuum value so

$$\lambda = \frac{\lambda_0}{n} \tag{1.6}$$

We note that it is possible for the speed of massive particles to exceed the speed of light in media. Shock waves that result are similar to a sonic boom for sound waves. In the case of light, it is called *Cherenkov radiation*.[3] The emitted light is confined to a cone, the half angle of which is defined by

$$\alpha = \cos^{-1}\left(\frac{c}{nv_p}\right) \tag{1.7}$$

where v_p is the particle speed.

The wavelength spectrum of light is vast, ranging from radio waves to γ-rays, with characteristic wavelengths as large as astrophysical objects to as small as nuclei, respectively (see Section 1.2). A narrow light source, such as a line-narrowed continuous wave (CW) laser, is often said to emit a "single frequency;" such light is termed *monochromatic light*. Monochromanicity, however, is a relative statement. Monochromatic as compared to what? All known sources of light emit within some bandwidth – a spread in wavelength ($\Delta\lambda$) or frequency ($\Delta\nu$) – be it as broad as the solar spectrum or as narrow as the resonance line of an atom.[4] While λ and ν are inversely proportional, it is helpful to

---

[3] The 1958 Nobel Prize in Physics went to Pavel A. Cherenkov for his discovery in 1934 that bears his name, which he shared with Ilya M. Frank and Igor Y. Tamm for their explanation (Cherenkov, Frank and Tamm, 1958).

[4] Microwave sources and state-of-the-art ultra-stable lasers can have widths of a fraction of a Hertz. Even still, $\Delta\nu \neq 0$!

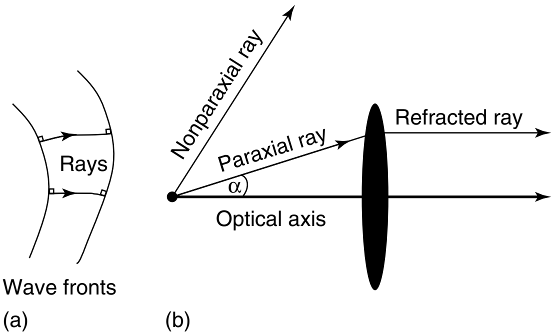

Fig. 1.1 Definition of rays and wave fronts (a) and paraxial rays (b) where $\alpha \ll 1$ rad.

recognize that

$$\left|\frac{\Delta\lambda}{\lambda}\right| = \left|\frac{\Delta\nu}{\nu}\right| \tag{1.8}$$

Classically, the treatment of light falls into two categories: geometrical and physical optics. In geometrical optics, the wave properties (e.g., diffraction) are ignored. Conceptually, we let $\lambda \to 0$ and instead discuss rays. While we do not review the usage of rays in this chapter, we do point out that ray tracing is employed extensively for designing optical systems. Rays are related to waves in that they are perpendicular lines joining the wave fronts (see Figure 1.1). The wave fronts turn out to be the surfaces of constant phase and so rays point in the direction of energy flow. Thus, rays from a point source are radial lines perpendicular to spherical surfaces. Typically, when dealing with rays, we focus on a subset of all the rays called *paraxial rays*. These rays are nearly parallel or form a small angle about a preferred direction. In the example shown in Figure 1.1, the ray traversing the center of the lens is the preferred direction and is called the *optical axis*. Paraxial rays deviate from the optical axis by such a small amount that $\sin\alpha \simeq \tan\alpha \simeq \alpha$. Rays are useful for describing refraction, the bending or redirection of light at the interface between two media with different indices of refraction, and reflection. Refraction, responsible for focusing of light by lenses and the angular spread of the $\Delta\lambda$ components after passing through prisms, is a result of momentum conservation and is succinctly stated through Fermat's principle: *light traverses a path from A to B that is an extremum of the optical path length (OPL)*.[5] That is,

$$\delta(OPL) \equiv \delta\left(\int_A^B n(s)ds\right) = 0 \tag{1.9}$$

where s is the geometric path. Fermat's Principle leads to two important properties of light. First, the law of reflection,

$$\theta_i = \theta_r \tag{1.10}$$

---

[5] The principle is often stated as the shortest path, that is, $\delta(OPL)$ would be a minimum. However, the calculus of variation only uses the fact that OPL is stationary; the second derivative is not considered. Thus, while usually the case, the path taken is not necessarily the minimum optical path.

where θ_i (θ_r) is the incident (reflected) angle. The second is Snell's Law,

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \qquad (1.11)$$

relating the incident (θ_1) and refracted (θ_2) angles for a ray refracted (i.e., bent) as it passes through an interface between two media with different indices of refraction.

It is interesting to note that Fermat's principle (ca 1657) is closely related to Maupertu's principle in mechanics (ca 1744) for self-contained systems obeying conservation laws,

$$\delta\left(\int_A^B p ds\right) = 0 \qquad (1.12)$$

where p is the momentum. Equation (1.12) is the principle of least action[6] when formulated more generally as

$$\delta\left(\int_A^B \mathcal{L}\, dt\right) = 0 \qquad (1.13)$$

with $\mathcal{L}$ being the Lagrangian.[7] These equations show that particles and rays of light assume rectilinear motions in free space or when there are no forces (fields[8]) and the index of refraction is constant. In general, the trajectories of particles and light are stationary. The index of refraction plays the role of a field causing rays to deviate from linearity when not constant just like forces (potentials) cause particle trajectories to bend.

The geometrical approximation is good when the variation of the physical features of the media are large in comparison to λ. When they become comparable to λ, the wave properties of the light must be considered. The realm of physical optics allows descriptions of elements such as apertures and grating. It further provides a framework to discuss fundamental concepts such as diffraction (the angular spread of a beam of light and the bending of light around obstacles Section 1.7), interference (the superposition of two or more waves, leading to constructive and destructive sums depending on the relative phase of the waves Section 1.8), and coherence (issues associated with how stable the phase is in time and across wave fronts).

The smallest unit of light is called the *photon*, light quanta after the German *Lichtquanten* meaning portions of light.[9] While centuries before the age of quantum physics, Isaac Newton championed the idea of light as a stream of corpuscles, photons are quantum entities whose behavior under certain conditions are well known. However, the answer to the question "*What is a photon precisely*" continues to be illusive. Over the years, the *definitions* tend to fall into one of three distinct categories:

- a fundamental particle;
- an elementary excitation of the EM field; or
- something registered by a photodetector.

---

[6] Like with Fermat's principle, least action is a bit of a misnomer; stationary action would be more appropriate as again only the first derivative is considered.

[7] We point out that Hamilton, Lagrange, Euler and others played a role in the development of the principle as well.

[8] Even in vacuum, the trajectory of light is deflected by a gravitational field. See, for example, Refs. Misner, Thorne and Wheeler (1973) and Hartle (2003) for a discussion of light in a gravitational field.

[9] Gilbert N. Lewis is given credit for coining this name (Lewis, 1926).

In this chapter, we do not argue for or against one view over another.

Massless photons, like massive particles, carry both energy, $h\nu$, and momentum, h/λ, where h is Planck's constant. However, the photon wavefunction must be constructed with care. There have been suggestions that the photon can be understood as simply a classical field plus vacuum fluctuations[10] – a semiclassical approach if you will. There are cases, however, where such an approach gives the wrong answer (as determined by experimental observation). Thus, we have two regimes: classical light and quantum light. By definition, quantum light is any behavior of light that cannot be explained by classical fields, that is, solutions to the wave equation. An example would be squeezed light (Henry and Glotzer, 1988).

The photon is considered to be a fundamental particle. It has an intrinsic spin, which is an integer of unit magnitude. Thus, it obeys Bose statistics, but it has only two states of helicity (aligned or antialigned with its direction of propagation) because being massless, it has no vacuum rest frame. In addition to spin, light has other nonclassical features, typically revealing themselves through intensity noise, correlations, and counting statistics. Finally, both classical fields as well as photons can carry orbital angular momentum and support vortices and solitons (Desyatnikov, Kivshar and Torner, 2005; Kivsha and Agrawol, 2003; Pismen, 1999).

We conclude this introduction by pointing out that if we substitute h/λ for p into Eq. (1.12), we get a different formulation of Fermat's principle,

$$\delta\left(h\int_A^B \frac{ds}{\lambda}\right) = 0 \tag{1.14}$$

The close analogy between Eqs (1.12) and (1.14) suggests an intimate connection between matter and light, from which one can postulate a wave equation for matter similar to that for light. As we discuss in Section 1.9, if we identify the wavelength of the particle as h/p (the de Broglie wavelength) and the index of refraction with $(U - V)/U$, where U and V are the total and potential energies, respectively, the time-independent Schrödiger equation emerges in a form that is not very different from the wave equation for light. Thus, photons, like matter, exhibit both wave and particle behavior.

With this overview as a backdrop, the remainder of this chapter is devoted to the details of selected characteristics of light. We start with the description of EM spectrum in Section 1.2 followed by a review of the wave equation and its solutions in Section 1.3. In Section 1.4, we consider radiometric issues and address the vector nature of light in Sections 1.5 and 1.6. We cover diffraction and interference in Sections 1.7 and 1.8 and conclude the chapter by further discussing the photon matter analogy in Section 1.9.

1.2
The Spectrum of Light

The EM spectrum is traditionally divided into the seven regions shown in Figure 1.2. It should be understood that the boundaries between these regions as well as those between subregions are

---

[10] Vacuum fluctuations refer to the photons that are created spontaneously from the vacuum.

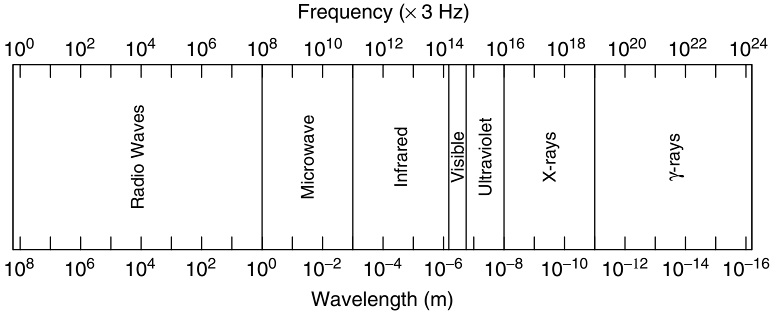
Frequency (× 3 Hz)
10^0
10^2
10^4
10^6
10^8
10^{10}
10^{12}
10^{14}
10^{16}
10^{18}
10^{20}
10^{22}
10^{24}
Radio Waves
Microwave
Infrared
Visible
Ultraviolet
X-rays
γ-rays
10^8
10^6
10^4
10^2
10^0
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
10^{-16}
Wavelength (m)

Fig. 1.2 The electromagnetic radiation spectrum.

not hard and fast, nor are the number of subregions unique. The most familiar region of the spectrum, the visible region, consists of wavelengths that range from about 0.40 μm at the blue end to 0.78 μm at the red end. Table 1.1 shows the corresponding colors for the wavelengths between. The visible subregions are a good example of the nonuniqueness of subbands of regions; for example, some references include cyan between green and blue while others insert indigo between blue and violet. Breaking the spectrum into six rather than seven or eight bands is of little consequence typically, because most objects emit a range of colors (i.e., $\Delta\lambda$ is relatively broad) or multiple colors (e.g., $\lambda_1 + \lambda_2 + \cdots + \lambda_n$ again spanning a large $\Delta\lambda$), making the identification of a pure color a rare event. Of course, when $\Delta\lambda$ is small as it often is for some lasers, our eyes in fact do perceive a pure color. For example, consider the red Helium–Neon laser at 632.8 nm or the green doubled Nd : YAG laser at 532 nm.

Subbands also exist for the other regions of the EM spectrum. Tables 1.2–1.4 give some of the more familiar subbands for the other regions. More about the spectrum of light can be found in Ref. (HyperPhysics, 2006).

1.3
Basics of Electromagnetic Waves

As mentioned in the introduction, physical optics is concerned with the

Tab. 1.1 The approximate wavelength, frequency, and energy ranges for six primary visible color bands. Energies increase from left to right.

Color	Wavelengths (nm)	Frequencies (×10^{14} Hz)	Energies (eV)
Red	780–625	3.8–4.8	1.6–2.0
Orange	625–590	4.8–5.1	2.0–2.1
Yellow	590–565	5.1–5.3	2.1–2.2
Green	565–500	5.3–6.0	2.2–2.4
Blue	500–435	6.0–6.9	2.4–2.8
Violet	435–380	6.9–7.9	2.8–3.6

Tab. 1.2 The approximate wavelengths, frequencies, and energies of key radio bands.

Designation	Wavelengths (m)	Frequencies (MHz)	Energy (meV)
AM radio	560–190	0.540–1.600	$(2.33–6.62) \times 10^{-3}$
TV	5.5–0.33	54–890	0.223–3.68
FM radio	3.40–2.78	88.1–108.1	0.364–0.447
Cell phone	0.43, 0.35, 0.18, 0.16, and 0.14	700, 850, 1700 1900, and 2100	2.89, 3.52, 7.03 7.85, and 8.68
Satellite radio	0.129–0.128	2320–2345	9.59 and 9.70
WiFi[a]	0.124–0.121	2412–2480	9.97–10.1
Radar	300–0.001	1–300 000	0.00414–1241

[a] The IEEE 802.11b/g/n standards communicate at 2.4 GHz. The IEEE 802.11a standard communicates at 5 GHz, but is essentially obselete.

Tab. 1.3 The approximate wavelength, frequency, and energy ranges of the three primary infrared (IR) bands.

Color	Wavelengths (μm)	Frequencies ($\times 10^{14}$ Hz)	Energies (eV)
Far IR	1000–10	3×10^{-3}–0.3	1.24×10^{-3}–0.124
IR-C	10–3.0	0.3–1.0	0.124–0.414
IR-B	3.0–1.4	1.0–2.14	0.414–0.886
IR-A	1.4–0.7	2.14–4.28	0.886–1.77

wave properties of light, which we examine in this section. We first review how Maxwell's equations lead to the wave equation for the components of the E and B fields. We examine EM wave solutions and their properties, primarily focusing on laser light in homogeneous media in the absence of charge and current sources.

1.3.1
Maxwell's Equations

As we alluded to in the introduction, a wave theory of light predates Maxwell's treatment. In the late 1800s, it was believed that the *aether*, an omnipresent elastic medium, supported light propagation much like the atmosphere or material supports sound waves. Although the theory based on the *aether* had consistency issues and required ad hoc assumptions to sustain it, its belief was so pervasive that even the null result of the experiments by Michelson and Morley could not easily dethrone it.11) It was not until Maxwell showed that his

---

11) The Michelson-Morley experiment was designed to measure a shift in interference fringes (see Section 1.8.1 for a discussion of interference) caused by a change in the speed of light moving with and perpendicular to the *aether*. The experiment is discussed in Refs. Michelson (1881) and Michelson and Morley (1887a,b).

Tab. 1.4 The approximate wavelength, frequency, and energy ranges of the three primary ultraviolet (UV) bands.

Designation	Wavelengths (nm)	Frequencies ($\times 10^{14}$ Hz)	Energies (eV)
UVA	400–320	7.45–9.37	3.10–3.88
UVB	320–280	9.37–10.7	3.88–4.43
UVC	280–200	10.7–15.0	4.43–6.2
Vacuum UV	200–50	15.0–60.0	6.2–24.8
Extreme UV	80–2.5	$37.5–1.2 \times 10^3$	15.5–500
Soft X-ray	4.5–0.15	$(0.7–2) \times 10^4$	$275–8.3 \times 10^3$

mathematical formulation of Gauss's, Ampere's and Faraday's laws (Maxwell's equations) lead naturally to an EM wave equation that could account for observed phenomena without resorting to arbitrary assumptions that the *aether* idea was abandoned.

Maxwell's equations in a medium with sources (represented by ρ, the charge density) and currents (represented by $\mathbf{j}$, the current density) can be written as

$$\nabla \cdot \mathbf{D} = \rho \tag{1.15}$$

$$\nabla \cdot \mathbf{H} = 0 \tag{1.16}$$

$$\nabla \times \mathbf{E} = \mathbf{j} - \frac{\partial \mathbf{B}}{\partial t} \tag{1.17}$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} \tag{1.18}$$

where

$$\mathbf{D} = \varepsilon \mathbf{E} \tag{1.19}$$

$$\mathbf{B} = \mu \mathbf{H} \tag{1.20}$$

These equations are simplified in the absence of sources, $\rho \to j \to 0$, and are the only cases we consider in this chapter.

1.3.2
Wave Equation

It is very straightforward to show that $\mathbf{E}$ and $\mathbf{H}$ of the EM field satisfy Maxwell's equations and the wave equation simultaneously in a homogeneous medium where $\nabla\varepsilon = \nabla\mu = 0$, by taking the curl of Eqs (1.17) and (1.18). Using the vector identity, $\nabla \times \nabla \times \mathbf{V} = \nabla(\nabla \cdot \mathbf{V}) - \nabla^2 \mathbf{V}$, and the fact that $\nabla \cdot \nabla \times \mathbf{V} = 0$ leads to

$$\nabla^2 \mathbf{E} = \varepsilon\mu \frac{\partial^2 \mathbf{E}}{\partial t^2} \tag{1.21}$$

$$\nabla^2 \mathbf{H} = \varepsilon\mu \frac{\partial^2 \mathbf{H}}{\partial t^2} \tag{1.22}$$

which, by inspection, are the same as Eq. (1.1) because $\varepsilon\mu = n^2\varepsilon_0\mu_0 = 1/v^2$. Of course, the general solution is of the form of Eq. (1.2), however, the specifics depend on the geometry and constraints of the problem.

1.3.2.1 Plane Waves

The simplest solution for EM waves is that of a plane wave, where each component of $\mathbf{E}$ and $\mathbf{H}$ is a function of $\xi = \hat{u} \cdot \mathbf{r} - vt$. Recall that $\hat{u} \cdot \mathbf{r} =$ constant defines a plane with $\hat{u}$ being a dimensionless unit vector perpendicular to that plane. It is straightforward then to show that

$$\frac{\partial \mathbf{E}}{\partial t} = -v\frac{\partial \mathbf{E}}{\partial \xi} \tag{1.23}$$

$$\nabla \times \mathbf{E} = \hat{u} \times \frac{\partial \mathbf{E}}{\partial \xi} \tag{1.24}$$

and similarly for **H**. Because **E** and **H** must satisfy Eqs (1.15)–(1.18) (again assuming a homogeneous medium), we can further write

$$\hat{u} \times \frac{\partial \mathbf{E}}{\partial \xi} = \sqrt{\frac{\mu}{\varepsilon}}\frac{\partial \mathbf{H}}{\partial \xi} \tag{1.25}$$

$$\hat{u} \times \frac{\partial \mathbf{H}}{\partial \xi} = -\sqrt{\frac{\varepsilon}{\mu}}\frac{\partial \mathbf{E}}{\partial \xi} \tag{1.26}$$

Integrating Eqs (1.25) and (1.26) and setting the constant to zero (no contribution from the background) leads to

$$\mathbf{E} = -\sqrt{\frac{\mu}{\varepsilon}} \quad \hat{u} \times \mathbf{H} \tag{1.27}$$

$$\mathbf{H} = \sqrt{\frac{\varepsilon}{\mu}} \quad \hat{u} \times \mathbf{E} \tag{1.28}$$

This implies that **E**, **H**, and $\hat{u}$ form a right-handed orthogonal triad and that light is a transverse field, that is, **E** and **H** oscillate in a plane normal to the propagation direction, in homogeneous media (including vacuum) without sources.

1.3.2.2 Scalar Harmonic Waves

The most common building block for the EM wave is a wave that is harmonic in both time and space. These exhibit sinusoidal variation. Typically, a scalar wave can be expressed as either a real quantity[12)]

$$\Psi(\mathbf{r}, t) = A(\mathbf{r}, t)\cos(\mathbf{k}\cdot\mathbf{r} \pm \omega t + \varphi) \tag{1.29}$$

or a complex quantity,

$$\Psi(\mathbf{r}, t) = A(\mathbf{r}, t)e^{i(\mathbf{k}\cdot\mathbf{r}\pm\omega t+\varphi)} \tag{1.30}$$

where A is an amplitude that is a slowly varying function of position and time (compared with the rapid variation of the sinusoidal arguments), **k** is the wavevector ($|\mathbf{k}| = 2\pi/\lambda$) and ω ($= 2\pi\nu$) is the angular frequency. Now it should be clear that $\hat{u}$. The harmonic time dependence of the wave allows the wave equation to be written as

$$\nabla^2\psi + k^2\psi = 0 \tag{1.31}$$

Because **E** and **B** are vectors, the EM wave is actually a vector wave. Generally, each component of the field satisfies the wave equation (Eq. 1.31) and has solutions like those in Eq. (1.29) and (1.30).

The argument of the harmonic wave consists of two phase terms, $\overline{\xi}_\pm \equiv \mathbf{k}\cdot\mathbf{r} \pm \omega t$, the dimensionless version of ξ, and φ. A constant $\overline{\xi}_-$ ($\overline{\xi}_+$) defines a profile or phase of the wave that moves toward more positive (negative) r as time evolves at a speed

$$v = \frac{\omega}{|\mathbf{k}|} \tag{1.32}$$

known as *the phase velocity*. The second phase term is often referred to as the *relative phase of the wave*. It can be a fixed constant or time dependent. When it is a constant or has a well defined time dependence, it gives rise

---

12) Note, we have multiplied ξ by $|\mathbf{k}|$ ($|\mathbf{k}|v = \omega$ because $|\mathbf{k}| = n\omega/c$) to make the argument dimensionless.

to coherence. When it varies randomly with time, the light is said to be incoherent. Furthermore, when $\varphi(t) = \text{const}{\cdot}t^n$, the frequency changes with time. A linear dependence simply shifts the frequency while higher powers chirp the frequency – as the wave passes the frequency either increases or decreases, depending on the sign of the constant. More complicated functions are possible as well.

1.3.2.3 Waves with Curved Phase Fronts

Although plane waves are highly convenient to use, they are appropriate only when dealing with light that is effectively far from its source.[13] Many situations do not fall into this category. It is beyond the scope of this chapter to discuss nonplanar waves extensively, but we will give two examples. For a more extensive discussion, the reader is directed to the text by Cowan, 1968. First, when the fronts are not planes, the solutions in Eq. (1.30) must be modified to correspond to the Laplacian being expressed in a different coordinate system. For example, a spherical wave takes the form

$$\Psi(r, t) = \frac{A}{r} e^{i(kr \pm \omega t + \varphi)} \tag{1.33}$$

where kr = constant. The phase fronts are clearly spheres. A bit more complicated example would be a cylindrical wave, which follows

$$\Psi(\rho, z, \theta, t) = A J_m(k\rho) e^{\pm i k_z z} e^{\pm i m\theta} e^{-i(\omega t + \varphi)} \tag{1.34}$$

where $J_m(k\rho)$ is the mth order Bessel function of the first kind (which are regular at the origin)[14] with m being a positive integer, and

$$k^2 = \left(\frac{\omega}{c}\right)^2 + k_z^2 \tag{1.35}$$

The surfaces of constant phase are just cylinders in this case.

1.4
Energy, Intensity, Power, and Brightness

Because the EM field is composed of E and B fields, its *Energy Density* is given by

$$u = \frac{1}{2}\left(\varepsilon|\mathbf{E}|^2 + \mu|\mathbf{H}|^2\right) \tag{1.36}$$

As a wave, this energy flows as described by the Poynting Vector,

$$\mathbf{S} = \mathbf{E} \times \mathbf{H} \tag{1.37}$$

The *Intensity* of the light is defined as the time average of **S**,

$$I = |\langle \mathbf{S} \rangle| \equiv \frac{1}{2}|\mathbf{E} \times \mathbf{H}| = \frac{n}{2\mu c}|\mathbf{E}|^2 \tag{1.38}$$

which has dimensions of watt per square centimeter.[15] In Eq. (1.38) we used the fact that $\omega/|\mathbf{k}| = c/n$. In vacuum, using the fact that $\varepsilon_0\mu_0 = 1/c^2$, we can write

$$I = \frac{1}{2\mu_0 c}|\mathbf{E}|^2 = \frac{1}{2}\varepsilon_0 c|\mathbf{E}|^2 \tag{1.39}$$

$$\simeq \frac{|\mathbf{E}|^2}{240\pi} \tag{1.39a}$$

---

[13] By far field, we mean that the phase fronts are planes. This can be achieved near the source with lenses.

[14] The boundary conditions of the problem might dictate a different Bessel solution. For example, if the origin were excluded, Bessel functions of the second kind, which are singular at the origin, would have to be considered as well.

[15] Technically, the SI unit is watt per square meter but in the United States, it is typically expressed as watt per square centimeter .

In this form, I has dimensions of watt per square centimeter (watt per square meter) when the dimensions of $\mathbf{E}$ are volt per centimeter (volt per meter). The *Power, P*, delivered is the integrated intensity over the exposed area,

$$I = \frac{P}{A} \tag{1.40}$$

where A is the area.[16] A related quantity, the *Brightness*, which is sometimes referred to as the *Radiance*, takes into account the solid angle, $\Delta\Omega$, through which the intensity is delivered and is given by

$$B = \frac{I}{\Delta\Omega} \tag{1.41}$$

which has dimensions watts per steradian per square centimeter. It is interesting to note that an unfocused laser delivering 1 mW of power at 780 nm is considerably brighter than a 100 W light bulb, 1.7×10^7 W/sr-cm^2 for a typical laser beam[17] with $w_0 = 1$ mm and $\Delta\Omega = 2 \times 10^{-7}$ sr compared with 0.6 W/sr-cm^2 for a light bulb at a distance of 1 m radiating into 4π. Thus, a laser is considered very bright, which can do real damage to an unprotected eye. Finally, laser light can be further characterized by its *spectral brightness*, the brightness per unit optical bandwidth,

$$SB = \frac{B}{\Delta\nu} \tag{1.42}$$

with units as watts per steradian per square centimeter hertz. The *brightness* and *spectral brightness* are often confused with each other as well as with the *Luminance*, a photometric quantity referring to a perceived brightness related more to how the eye responds.

1.5
Polarization

As mentioned earlier, EM waves are actually vector waves, because $\mathbf{E}$ and $\mathbf{B}$ point in specific directions. *Polarization* captures this feature, and is defined in terms of the direction of $\mathbf{E}$.[18] The most general case is elliptical polarization, which has two limiting cases, linear and circular polarization. These names are so chosen because they describe the geometric shapes $\mathbf{E}$ that sweeps out while looking at the light along (parallel or antiparallel to) $\mathbf{k}$. We have already discussed that $\mathbf{E}$, $\mathbf{B}$, and $\mathbf{k}$ form a right-handed Cartesian triad so polarization also specifies the direction of $\mathbf{B}$. We will take $\hat{k} \equiv \hat{z}$ and focus on light that is perfectly polarized in the discussion that follows.

In general $\mathbf{E}$ will have two orthogonal components,

$$\begin{aligned} \mathbf{E}_1 &= \hat{x}E_{01}e^{i(kz-\omega t+\varphi_1)} \\ &= \hat{x}E_{01}e^{i(\bar{\xi}+\varphi_1)} \end{aligned} \tag{1.43}$$

$$\begin{aligned} \mathbf{E}_2 &= \hat{y}E_{02}e^{i(kz-\omega t+\varphi_2)} \\ &= \hat{y}E_{02}e^{i(\bar{\xi}+\varphi_2)} \end{aligned} \tag{1.44}$$

We will first consider the case where E_{01}, E_{02}, φ_1, and φ_2 are all real and

---

[16] The area of a laser beam is given by πw_0^2, where w_0 is the beam radius.

[17] For a diffraction limited laser beam, $\Delta\Omega = \pi\theta_d^2$ where $\theta_d = \lambda/\pi w_0$.

[18] Another reason for considering only $\mathbf{E}$ is the magnitude of $\mathbf{B}$ relative to $\mathbf{E}$ is down by a factor of c. Thus at low intensities, $< 10^{14}$ W/cm^2, $\mathbf{E}$ dominates the physics.

Tab. 1.5 Various electromagnetic field quantities.

Quantity	Name	SI Unit
$c = 2.99792458 \times 10^8$	Light vacuum speed[a]	m/s
$\mu_0 = 4\pi \times 10^{-7}$	Vacuum permeability[a]	T-m/A (kg-m/A^2-s^2)
$\varepsilon_0 = 8.854187817\ldots \times 10^{-7}$	Vacuum permittivity	F/m (A^2-sec^4/kg-m^3)
$\mathbf{E}$	Electric field[b]	V/m (kg-m/A-s^3)
$\mathbf{D}$	Electric displacement	C/m^2
$\mathbf{B}$	Magnetic induction[c]	T (kg/A-s^2)
$\mathbf{H}$	Magnetic field	A/m
ρ	Charge density[b]	C/m^3
$\mathbf{j}$	Current density[b]	A/m^2
P	Power	W (kg/m^2-s^3)
$\mathbf{S} \equiv \mathbf{E} \times \mathbf{H}$	Poynting vector	W/m^2 (kg/m^4-s^3)
$I \equiv \langle \lvert\mathbf{S}\rvert \rangle$	Intensity[b]	W/cm^2

[a] All defined to be exact.

[b] In the US, the explicit length measures for these quantities are given in centimeters, for example, volt per centimeter, watt per square centimeter, and so on.

[c] Sometimes called the *magnetic-flux density*.

time independent. Taking the real part of these fields,

$$\mathbf{E}_1 = \hat{x}E_{01}\cos(\bar{\xi} + \varphi_1) \tag{1.45}$$

$$\mathbf{E}_2 = \hat{y}E_{02}\cos(\bar{\xi} + \varphi_2) \tag{1.46}$$

leads to an equation of a conic section

$$\left(\frac{|\mathbf{E}_1|}{E_{01}}\right)^2 + \left(\frac{|\mathbf{E}_2|}{E_{02}}\right)^2 - 2\left(\frac{|\mathbf{E}_1|}{E_{01}}\right)\left(\frac{|\mathbf{E}_2|}{E_{02}}\right)\cos\varphi = \sin^2\varphi \tag{1.47}$$

and $\varphi = \varphi_2 - \varphi_1$. Equation (1.47) describes an ellipse when

$$\frac{\sin^2\varphi}{E_{01}^2 E_{02}^2} \geq 0 \tag{1.48}$$

Because the numerator and denominator are positive definite, Eq. (1.48) is always true.

Special Linear Case # 1: $\varphi = 0$ or $\varphi = \pi$
Equation (1.47) reduces to

$$\frac{|\mathbf{E}_1|}{E_{01}} = \frac{|\mathbf{E}_2|}{E_{02}} \tag{1.49}$$

which describes a straight line. Because the two component oscillate in phase, this case leads to linear polarization. The case for $\varphi = 0$ and $\varphi = \pi$ are orthogonal to each other.

Special Circular Case # 2: $\varphi = \pm\pi/2$ and $E_{01} = E_{02}$
The equation reduces to

$$|\mathbf{E}_1|^2 + |\mathbf{E}_2|^2 = E_{01}^2 \tag{1.50}$$

the equation of a circle of radius E_{01}. The two components are out of phase by half a wavelength (or period) but the magnitude of the resultant, E_{01}, is constant but sweeps out a circle leading to circular polarization. The sense of rotation depends on the sign of φ, with the minus (plus) sign producing light with positive (negative) helicity, where positive helicity obeys the right-hand rule, so if you look in the direction of propagation, the E-field rotates clockwise.[19]

Special Elliptical Case # 3: $\varphi = \pm\pi/2$ *and* $E_{01} \neq E_{02}$

Equation (1.47) reduces to

$$\left(\frac{|\mathbf{E}_1|}{E_{01}}\right)^2 + \left(\frac{|\mathbf{E}_2|}{E_{02}}\right)^2 = 1 \tag{1.51}$$

which is an ellipse with the major axis aligned with the horizontal (vertical) axis when $E_{01} > E_{02}$ ($E_{01} < E_{02}$). The sense of rotation is the same as in special case # 2.

General Elliptical Case $E_{01} \neq E_{02}$

In the general elliptical polarization case, one has a rotated ellipse where the angle, α, of the major axis away from the $\hat{E}_1$ direction is given by

$$\tan 2\alpha = \frac{2E_{01}E_{02}}{E_{01}^2 - E_{02}^2}\cos\varphi \tag{1.52}$$

Note, when $E_{01} = E_{02}$, Eq. (1.52) cannot be used and one must go back to Eq. (1.47) to determine α.

1.5.1
Polarization Bookkeeping

There are several approaches to keeping track of the polarization of light, which is particularly important when light interacts with media that can either decrease the intensity or delay the transit time of one polarization or helicity relative to the other. Here, we mention two matrices, namely, the Jones and Mueller matrices. A more in depth discussion can be found in Ref. Goldstein, 2003. The Jones approach involves a set of 2 × 2 matrices with complex elements that transform two-element vectors that describe the complex amplitude and phase of the light. Every Jones vector corresponds to a physically realizable polarization configuration. The Mueller approach uses a set of 4 × 4 matrices with real elements, which have values of either 0 or ±1, to transform the Stokes vectors (see Section 1.5.3). However, some matrices do not represent real configurations.

1.5.2
Jones Matrices

In the Jones calculus Clark Jones, 1941, an initial complex field **E** is transformed to the final complex field **E**′ via matrix multiplication

$$\begin{pmatrix} E'_x e^{i\phi'_x} \\ E'_y e^{i\phi'_y} \end{pmatrix} = \begin{pmatrix} j_{11} & j_{12} \\ j_{21} & j_{22} \end{pmatrix} \cdot \begin{pmatrix} E_x e^{i\phi_x} \\ E_y e^{i\phi_y} \end{pmatrix} \tag{1.53}$$

where the components of the matrix **J** for common elements are given in Table 1.6.

1.5.3
Mueller Matrices

The general Muller calculus is also a matrix operation,

$$\mathbf{S}' = \mathbf{M} \cdot \mathbf{S} \tag{1.54}$$

where **M** is a 4 × 4 matrix and **S** is the four-element Stokes vector, credited to Sir George Gabriel Stokes for their invention. Given the electric fields in Eqs (1.45) and (1.46), the four

---

[19] It should be noted that some references define circular polarization in terms of right-hand and left-hand circular polarization. This definition traditionally corresponds to looking antiparallel to **k** so $\varphi = -\pi/2$ would lead to left-hand circular polarization.

Tab. 1.6 Jones matrices for common optical elements.

Optical element	Jones matrix
Linear polarizer $\parallel\hat{x}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$
Linear polarizer $\parallel\hat{y}$	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$
Linear polarizer at ±45°	$\frac{1}{2}\begin{pmatrix} 1 & \pm 1 \\ \pm 1 & 1 \end{pmatrix}$
$\frac{1}{4}$-Wave plate, Fast axis $\parallel \begin{matrix} \hat{x} & (+) \\ \hat{y} & (-) \end{matrix}$	$e^{i\pi/4}\begin{pmatrix} 1 & 0 \\ 0 & \pm i \end{pmatrix}$
Circular polarizer, ± Helicity	$e^{i\pi/4}\begin{pmatrix} 1 & \mp i \\ \pm i & 1 \end{pmatrix}$

Tab. 1.7 Mueller matrices for common optical elements.

Optical element	Mueller matrix
Linear polarizer $\parallel \begin{matrix} \hat{x} & (+) \\ \hat{y} & (-) \end{matrix}$	$\frac{1}{2}\begin{pmatrix} 1 & \pm 1 & 0 & 0 \\ \pm 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$
Linear polarizer at ±45°	$\frac{1}{2}\begin{pmatrix} 1 & 0 & \pm 1 & 0 \\ 0 & 0 & 0 & 0 \\ \pm 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$
$\frac{1}{4}$-Wave plate, Fast axis $\parallel \begin{matrix} \hat{x} & (+) \\ \hat{y} & (-) \end{matrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \pm 1 \\ 0 & 0 & \pm 1 & 0 \end{pmatrix}$
Circular polarizer, ± Helicity	$\frac{1}{2}\begin{pmatrix} 1 & 0 & 0 & \pm 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \pm 1 & 0 & 0 & 1 \end{pmatrix}$

components of **S** are defined as

$$S_0 = |E_{01}|^2 + |E_{02}|^2 \tag{1.55}$$

$$S_1 = |E_{01}|^2 - |E_{02}|^2 \tag{1.56}$$

$$S_2 = |2E_{01}E_{02}\cos\varphi| \tag{1.57}$$

$$S_3 = |2E_{01}E_{02}\sin\varphi| \tag{1.58}$$

The Mueller matrices are given in Table 1.7.

1.6
Longitudinal Field Component

Another manifestation of the vector nature of light is that it has a longitudinal component, even though it is customary to ignore it. It must be noted that a pure plane wave exists over all space and has no transverse variation. The finite extent of the field turns out to be acceptable in many cases but one runs into problems with very intense light, particularly when focused. Lax, Louisell, and McKnight (1975) showed that a purely transverse field is not an exact solution to Maxwell's equations, but rather it is the zeroth-order solution to the paraxial approximation to Maxwell's equations. The exact solutions require a longitudinal component given by

$$E_z(x, y, z) = \frac{1}{ik}\nabla_\perp \cdot \mathbf{E}_\perp \tag{1.59}$$

where $k = |\mathbf{k}| = \sqrt{k_x^2 + k_y^2 + k_z^2}$ and the $\perp$ symbol indicate the transverse components of the ∇ operator and **E** field. Using Fourier analysis of the fields, Scully and Zubairy (1991) showed a field obeying the paraxial approximation,

$$k_x, k_y \ll k \tag{1.60}$$

$$k_z \simeq k\left(1 - \frac{1}{2}\frac{k_x^2 + k_y^2}{k^2}\right) \tag{1.61}$$

implies Eq. (1.59).

1.7
Diffraction

When light passes a sharp edge, it does not produce a sharp shadow. Also, when it passes through a circular hole, it does not produce a disk of the same size. Under the right conditions, it produces not only a larger spot but also rings. Furthermore, the transverse size of a laser beam expands as it propagates. These observations are elegantly described by diffraction theory. Diffraction falls into two classes – Fraunhofer and Fresnel. Fraunhofer diffractions describes what happens when the phase fronts are near plane waves, where the curvature of the field can be ignored. Fresnel diffraction takes curvature into account.

Huygens, in the late seventeenth century, suggested a description for wave propagation as a collection of individual spherical sources called *secondary sources*, the sum of which would make up the wavefront. It is a straightforward exercise to convince oneself that Huygens's principle can be used to construct a plane as well as other simple geometries. When applied to a hole, Huygens's approach leads to an emerging spherical wave, because part of the plane wave is blocked. This would appear to account for the observed spread. However, there is a difficulty. If the secondary waves are spherical, then there should also be part of the wave going backward. Huygens had to ignore this part of the wave. It turns that when considered more mathematically, this problem is corrected by what is call the *obliquity factor*.

The mathematical statement of the principle for a wave propagating in free space is the Fresnel-Kirchhoff integral formula,[20)]

$$\psi_P = -\frac{ik}{4\pi}\psi_0 \iint \frac{e^{ik(r+\bar{r})}}{r\bar{r}} \times [\cos\theta(\mathbf{n},\mathbf{r}) - \cos\theta(\mathbf{n},\bar{\mathbf{r}})]\,dA \tag{1.62}$$

where the integral is over the area of the aperture. The distances, r and $\bar{r}$, between the aperture and observation point and aperture and source, respectively, are defined in Figure 1.3 as is $\mathbf{n}$, the normal to the surface, pointing toward the source. The angles between the vectors and the normal are represented by $\theta(\mathbf{n},\mathbf{r})$ and $\theta(\mathbf{n},\bar{\mathbf{r}})$.

Let's consider an example of an aperture. In the Fraunhofer limit, s and p are effectively a long way from the aperture. In this case, we can take the surface of the aperture to be a spherical cap such that F is constant. Thus, $\bar{r}$ and $\mathbf{n}$ are antiparallel always and $\cos\theta(\mathbf{n},\bar{\mathbf{r}}) = -1$.[21)] Equation (1.62) then reduces to

$$\psi_P = -\frac{ik}{4\pi}A\psi_0 \int \frac{e^{ik(r+\bar{r})}}{r\bar{r}} \times [\cos\theta(\mathbf{n},\mathbf{r}) + 1]\,dA \tag{1.63}$$

[20)] The Fresnel-Kirchhoff integral formula of Eq. (1.62) can be derived from Green's theorem (see, for example, Fowles, 1968) for two functions that are continuous, integrable and satisfy the wave equation,

$$\iint (V\nabla_\perp U - U\nabla_\perp V)\,dA = \iiint (V\nabla^2 U - U\nabla^2 V)\,dV$$

where the first integral is over any closed surface and the second is over the volume enclosed.

[21)] When $\bar{r}$ and r are much larger than the aperture size, a spherical surface is not much different from a flat surface.

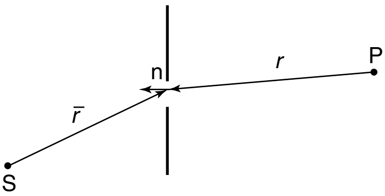

Fig. 1.3 Geometry for the Fresnel-Kirchhoff integral formula.

where $\cos\theta(\mathbf{n},\mathbf{r})+1$ is the obliquity factor mentioned above, which is zero for the wave that is going backwards, towards the source. Table 1.8 gives a few key results derived from Eq. (1.63).

From Table 1.8, we draw an important conclusion. The smaller the aperture, the larger the diffraction. In general, the diffraction angle is given by

$$\text{Slit} \longrightarrow \theta \sim \lambda/b \tag{1.64}$$

$$\text{Circle} \longrightarrow \theta \sim 1.22\lambda/2r \tag{1.65}$$

1.8
Interference

Interference is concerned with the superposition of waves. In general, the sum of wave solutions is also a solution to the wave equation. Here, we consider three examples: summing two waves of the same frequency, the general case of summing multiple waves of different frequencies and generation of short pulses of light by summing many frequencies with a precise phase relationship.

Tab. 1.8 Diffraction patterns from key apertures.*

Aperture	**Intensity**
Slit, width b	$I_0\,(\sin\beta/\beta)^2$
Rectangular slit, area $a \times b$	$I_0\,(\sin\alpha/\alpha)^2$ $(\sin\beta/\beta)^2$
Circular aperture, radius r	$I_0\left(2J_1(\rho)/\rho\right)^2$
N slits, h spacing	$N^2 I_0\,(\sin\beta/\beta)^2$ $(\sin N\gamma/N_{\sin}\gamma)^2$

*$\alpha = (\pi a/\lambda)\sin\phi$
$\beta = (\pi b/\lambda)\sin\theta$
$\gamma = (\pi h/\lambda)\sin\theta$
$\rho = (2\pi r/\lambda)\sin\theta$
ϕ diffraction angle, y direction.
θ diffraction angle, x direction.

1.8.1
Superposition: Single Frequency

This is straightforward to see with plane waves. Consider first two plane waves of the same frequency but with different real amplitudes and a relative phase between them,

$$A_T(\mathbf{r},t,\varphi_o) = a_1 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t+\varphi_1)} + a_2 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t+\varphi_2)} \tag{1.66}$$

The sum produces a new sinusoidal wave,

$$A_T(\mathbf{r},t,\varphi) = A_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t+\varphi_0)} \tag{1.67}$$

where

$$A_0 e^{i\varphi_0)} = a_1 e^{i\varphi_1} + a_2 e^{i\varphi_0} \tag{1.68}$$

We can determine A_0 and φ_0 in terms of $a_{1,2}$ and $\varphi_{1,2}$ by expanding the exponentials such that

$$\begin{aligned} A_0(\cos\varphi_0 + i\sin\varphi_0) &= a_1(\cos\varphi_1 + i\sin\varphi_1) \\ &\quad + a_2(\cos\varphi_2 + i\sin\varphi_3) \end{aligned} \tag{1.69}$$

Equating the cosine (sine) terms on the left with those on the right and then dividing the sine terms by the cosine terms leads to

$$\tan\varphi_0 = \frac{a_1\sin\varphi_1 + a_2\sin\varphi_2}{a_1\cos\varphi_1 + a_2\cos\varphi_2} \tag{1.70}$$

At the same time, taking the modulus squared of Eq. (1.68) produces

$$\begin{aligned}|A_0|^2 &= |a_1|^2 + |a_2|^2 \\ &\quad + \left(a_1 a_2^* e^{i\Delta\varphi_0} + c.c.\right) \\ &= a_1^2 + a_2^2 + 2a_1a_2\cos\Delta\varphi_0 \end{aligned}\tag{1.71}$$

where $\Delta\varphi_0 = \varphi_1 - \varphi_2$. Given $a_{1,2}$ and $\varphi_{1,2}$, A_0 (the intensity) and φ_0 can be found from Eqs (1.70) and (1.71). Equation (1.71) is known as the *coherent sum of the two waves.* That is, one adds the amplitudes before squaring to get the total intensity. The intensity is proportional to the square of the amplitude so it is also possible to write Eq. (1.71) as

$$I_T = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\Delta\varphi_0 \tag{1.72}$$

The third term in (Eqs 1.71 and 1.72) is sometimes called the *interference term* and plays an important role in describing the intensity of the resultant wave. Consider the case where $a_1 = a_2$ so $I_1 = I_2 = I$. When $\Delta\varphi_0 = 2m\pi$ $(m = 0, 1, 2, \ldots)$, the two waves are said to be in phase, in which case

$$I_T = (a_1 + a_2)^2 = 4I \tag{1.73}$$

When $\Delta\varphi_0 = (2m+1)\pi/2$, we have the opposite extreme,

$$I_T = (a_1 - a_2)^2 = 0 \tag{1.74}$$

When $a_1 \neq a_2$, the two extremes give resultants with maximum and minimum I_T respectively.

In the more general case of many waves, all with the same frequency, we have

$$\begin{aligned}A_T(\mathbf{r}, t, \varphi_0) &= A_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t + \varphi_0)} \\ &= \sum_{j=1}^{N} a_j e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t + \varphi_j)}\end{aligned}\tag{1.75}$$

where

$$\begin{aligned}I_T = |A_0|^2 = \sum_{j=1}^{N} |a_j|^2 + \frac{1}{2}\sum_{j\neq k}^{N} \\ \left(a_j a_k^* e^{i(\varphi_j - \varphi_k)} + c.c.\right)\end{aligned}\tag{1.76}$$

and

$$\tan\varphi_0 = \frac{\sum_{j=1}^{N} a_j \sin\varphi_j}{\sum_{j=1}^{N} a_j \cos\varphi_j} \tag{1.77}$$

Again, the resultant is a sinusoidal wave with an intensity given by a coherent sum. In the case where all the amplitudes are the same so that each wave has an intensity I,

$$I_T = N^2 I \tag{1.78}$$

In the case where $\Delta_{j,k} = \varphi_j - \varphi_k$ is not well defined but varies randomly with time, it is straightforward to show that the second sum in Eq. (1.76) vanishes by writing the exponentials in terms of sines and cosines and using the fact that the time average of $\sin\Delta_{j,k} \to 0$ as does that of $\cos\Delta_{j,k}$. Thus, the interference terms vanish. In the case where all amplitudes are the same, the resultant wave corresponds to an incoherent sum of the contributors,

$$I_T = NI \tag{1.79}$$

For an incoherent sum, one squares first and then adds the intensities.

1.8.1.1 Interferometry

An entire field of study with industrial applications is built upon an equation similar to Eq. (1.72). The most general situation is where a beam of light is divided into two with each traveling different paths and brought back together. Because the two beams came from the same source, and if the path length difference is not too large, so that the two beams are still in phase, the resultant intensity will be the same as Eq. (1.72) except that $\Delta\varphi_0 \to \delta$ in the argument of the interference term where

$$\delta = k\Delta l \tag{1.80}$$

with Δl being the path length difference between the two arms. In this case, constructive interference occurs when $\Delta l = n\lambda$, whereas destructive interference occurs when $\Delta l = (2n+1)\lambda/2$, where n is a positive integer. Two-beam interferometry exploits interference patterns to measure inhomogeneities and defects in material.

1.8.2
Superposition: Multiple Frequencies

Superposition involving waves of different frequencies leads to some very interesting possibilities such as ultrashort busts of light. The general principle of summing waves with different frequencies can be understood in the special case where the amplitude and phase are the same for each wave:

$$\begin{aligned} A_T(\mathbf{r}, t) &= A_0 \exp[i(\mathbf{k}_1 \cdot \mathbf{r} - \omega_1 t)] \\ &\quad + A_0 \exp[i(\mathbf{k}_2 \cdot \mathbf{r} - \omega_2 t)] \\ &= 2A_0 \exp\left[\frac{i}{2}(\Delta\mathbf{k} \cdot \mathbf{r} - \Delta\omega t)\right] \\ &\quad \times \exp\left[\frac{i}{2}(\mathbf{k}_m \cdot \mathbf{r} - \omega_m t)\right] \end{aligned} \tag{1.81}$$

where

$$\Delta\mathbf{k} = \frac{1}{2}(\mathbf{k}_1 - \mathbf{k}_2) \tag{1.82}$$

$$\mathbf{k}_m = \frac{1}{2}(\mathbf{k}_1 + \mathbf{k}_2) \tag{1.83}$$

$$\Delta\omega = \frac{1}{2}(\omega_1 - \omega_2) \tag{1.84}$$

$$\omega_m = \frac{1}{2}(\omega_1 + \omega_2) \tag{1.85}$$

Equation (1.81) represents a wave oscillating at the mean of the two frequencies, ω_m, and modulated by a temporal and spatial envelope given by $2A_0 \exp[\frac{i}{2}(\Delta\mathbf{k} \cdot \mathbf{r} - \Delta\omega t)]$. Figure 1.4 shows examples of adding two waves with different frequencies. Unlike the case of equal frequencies, in this case, the two sinusoidal waves produce a wave that is periodic but not sinusoidal. Such waves are called *anharmonic*. For the sum of two waves, we have two different speeds. As with a single frequency, we again have a phase velocity – the ratio between the average frequency and wavenumber, $v_{ph} = \omega_m/|\mathbf{k}|$. But, we have a new speed that goes by the name of the group velocity, the speed with which the envelope moves, $v_g = \Delta\omega/\Delta|\mathbf{k}|$.

When a wave is composed of many frequencies, $\omega \to \omega(k)$. Typically, the frequencies are grouped around a central frequency, $\omega(k_0)$, allowing $\omega(k)$ to be expanded into a Taylor series,

$$\omega(k) = \omega(k_0) + (k - k_0) \left.\frac{d\omega}{dk}\right|_{k_0} + \cdots \tag{1.86}$$

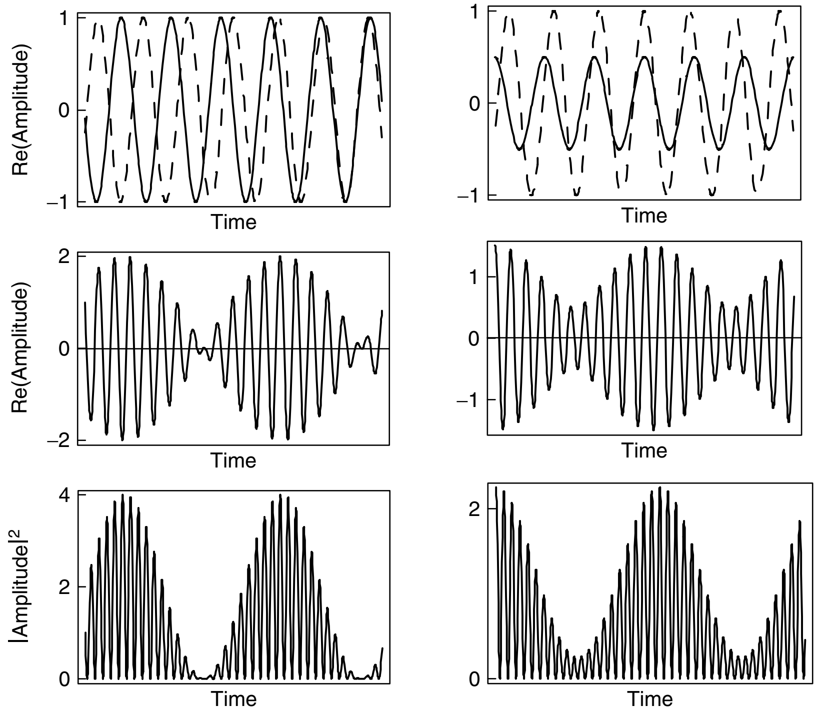

Fig. 1.4 Superposition of two waves, $A_1 \exp[i(\mathbf{k}_1 \cdot \mathbf{r} - \omega_1 t) + \varphi_1] + A_2 \exp[i(\mathbf{k}_2 \cdot \mathbf{r} - \omega_2 t) + \varphi_2]$ where the top row corresponds to the individual amplitudes (A_1 solid curves), the middle row to the resultant sum of amplitudes and the bottom row to the square of the modulus of the resultant amplitudes: (left column) $A_1 = A_2$, $\Delta\varphi = \pi/2$ and $\omega_1/\omega_2 = 0.9$; (right column) $A_1 = A_2/2$, $\Delta\varphi = 0$ and $\omega_1/\omega_2 = 0.9$, where $\Delta\varphi = \varphi_1 - \varphi_2$.

In this case, the addition of the various frequency components is more easily handled via Fourier analysis (see, for example Arfken and Weber (2000) and Boas (2006) for a review), where the amplitude $A(t)$ $(A(z))$ in the time (coordinate space) domain is linked to $\tilde{A}(\omega)$ $(\tilde{A}(k))$ in the frequency (spatial frequency) domain through

$$A(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \tilde{A}(\omega) e^{-i\omega t} d\omega \quad (1.87)$$

$$\tilde{A}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} A(t) e^{i\omega t} dt \quad (1.88)$$

for time-frequency and

$$A(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \tilde{A}(\mathbf{k}) e^{-ikz} dk \quad (1.89)$$

$$\tilde{A}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} A(z) e^{ikz} dz \quad (1.90)$$

for coordinate space - frequency space. Thus, in Fourier components, the electric field for a scalar wave propagating in the z-direction can be written as

$$E(z, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \tilde{E}(k) e^{-i[kz + \omega(k)t]} dk \quad (1.91)$$

In many cases, $d\omega/dk$ is the appropriate and more general expression for the group velocity. This can be seen by substituting the first two terms of Eq. (1.86) into Eq. (1.91),

$$E(z,t) = \frac{1}{\sqrt{2\pi}} e^{i[k_0(d\omega/dk)|_{k_0} - \omega(k_0)]t} \times \int_{-\infty}^{\infty} \tilde{E}(k) e^{-i[z+(d\omega/dk)|_{k_0} t]k} dk \tag{1.92}$$

However, Eq. (1.90) implies

$$\tilde{E}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E(z, t=0) e^{ikz} dz \tag{1.93}$$

which allows Eq. (1.92) to be written as

$$E(z,t) = \frac{e^{i[k_0(d\omega/dk)|_{k_0} - \omega(k_0)]t}}{2\pi} \times \int_{-\infty}^{\infty} E(z', 0) dz' \times \int_{-\infty}^{\infty} e^{i(z' - z - (d\omega/dk)|_{k_0} t)k} dk \tag{1.94}$$

where we do the k integration first. The last integral is just $\delta\left(z' - z - \frac{d\omega}{dk}\Big|_{k_0} t\right)$, from which we get

$$E(z,t) = \frac{1}{2\pi} E(z + d\omega/dk|_{k_0} t, 0) \times e^{-i[\omega(k_0) - k_0(d\omega/dk)|_{k_0}]t} \tag{1.95}$$

By inspection, it is clear that the envelope in Eq. (1.95) moves with speed $d\omega/dk|_{k_0}$ and the carrier oscillates with frequency $\omega(k_0) - k_0 \frac{d\omega}{dk}|_{k_0}$ under the envelope. Thus, we define the group velocity as

$$v_g = \frac{d\omega}{d|\mathbf{k}|} \tag{1.96}$$

In vacuum $v_{ph} = v_g$. However, if the medium through which the wave propagates is dispersive, $n \to n(\lambda)$ so that $dn/d|\mathbf{k}| \neq 0$, the two velocities can be very different. Thus, it is often convenient to write v_g in a form that includes the dispersion explicitly,

$$v_g = \frac{c}{n}\left(1 - \frac{|\mathbf{k}|}{n}\frac{dn}{d|\mathbf{k}|}\right) \tag{1.97}$$

The group velocity is typically the speed with which information is transmitted. It is important to remember that the group velocity is actually only the first term in a series and in cases where $dn/d|\mathbf{k}|$ changes very rapidly or is anomalous (i.e., negative), higher order terms must be kept to determine the speed with which information travels correctly.

1.8.3
Short Pulses

Figure 1.4 shows the basic idea for generating pulses of light of short duration. Specifically, in this case, two frequencies with well-defined relative phase (i.e., fixed in time) are summed in the frequency domain to provide a new wave with beats in the time domain. As additional frequencies are added, the temporal width of the beat envelope narrows. To gain a better understanding of the relationship between the length of the pulse train and its *bandwidth* or number of frequencies required to sum in order to produce it, we will turn the problem around and start with an idealized pulse train in the time domain. Figure 1.5, for example, shows two finite length, idealized, pulse trains, one with three cycles and the other with six cycles.

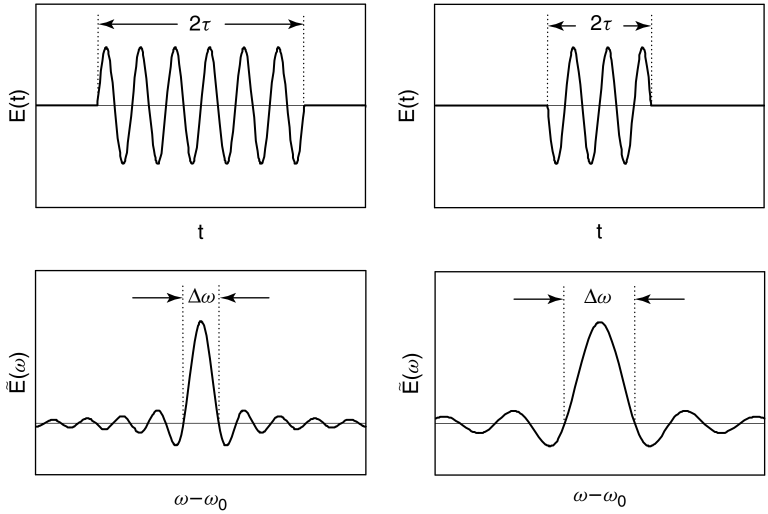

Fig. 1.5 Idealized short pulses formed by finite unit amplitude N-cycle pulse trains (top) with $N = 6$ (left) and $N = 3$ (right). Their respective Fourier transforms appear below with peak amplitudes of $\sqrt{\pi/2}N/\omega_0$ and the first zeros occurring at $\omega = \omega_0(1 \pm 1/N)$.

Mathematically, these obey

$$E(t) = \begin{cases} E_0 \sin \omega_0 t & \text{for } -\tau \le t \le \tau \\ 0 & \text{at other times.} \end{cases} \tag{1.98}$$

The length of this pulse is 2τ, where $\tau = N\pi/\omega_0$ with N being the number of cycles in the train. Using a Fourier analysis similar to that described above, the frequency spectrum is given by

$$\tilde{E}(\omega) = \frac{E_0}{\sqrt{2\pi}}\left[\frac{\sin \tau(\omega - \omega_0)}{\omega - \omega_0} - \frac{\sin \tau(\omega + \omega_0)}{\omega + \omega_0}\right] \tag{1.99}$$

At optical or near IR frequencies, because the second term is much smaller than the first, we can apply the Fourier transform to just the first term, which is also plotted in Figure 1.5. Clearly, the number of frequencies involved in the shorter pulse is larger than the number needed for the longer pulse. This inverse relationship between the length of the pulse in the time domain and the spread in the frequency domain is conveniently captured in the time-bandwidth product, $\tau\Delta\nu$. From Figure 1.5 and Eqs (1.98) and (1.99), it is clear that $\tilde{E}(\omega) = 0$ when $N\pi(\omega - \omega_0)/\omega_0 = \pm\pi$. Thus, $\Delta\omega = \omega_+ - \omega_- = 2\omega_0/N = 2\pi/\tau$, where $\omega_\pm = \omega_0 (1 \pm 1/N)$, which leads to

$$\tau\Delta\nu = 1 \tag{1.100}$$

It is interesting to note that if we multiply Eq. (1.100) by h, this time-bandwidth product satisfies the Heisenberg uncertainty principle,

$$\Delta t \Delta E \ge \frac{\hbar}{2} \tag{1.101}$$

where $\Delta E = h\Delta\nu$, $\hbar = h/2\pi$ and we substituted Δt for τ. The minimum is reached when the so-called minimum uncertainty wavepacket is prepared.[22]

Ultrashort pulses are achieved by "locking" the frequency components that extends over a wide frequency range. The minimum width achievable by this technique corresponds to one complete cycle of light. At 800 nm, near the peak of the Ti : Sapphire laser, this is ~ 2.7 femtoseconds. For a more complete discussion on mode locking and the generation of ultrashort pulses, the reader is directed to the classic text by Siegman Siegman, 1986

1.9
Photons and Particles

We conclude by discussing the analogy between light and particle waves a bit further. Equations (1.13), (1.14) and (1.31) can be used to motive the time-independent Schrödinger wave equation,

$$-\frac{\hbar^2}{2m}\nabla^2\psi + V\psi = \mathcal{E}\psi \qquad (1.102)$$

$$\nabla^2\psi + \frac{2m}{\hbar^2}(\mathcal{E} - V)\psi = 0 \qquad (1.103)$$

where $\hbar = h/2\pi$. Because λ for the particle is h/p, in free space, we postulate that k $(= 2\pi n/\lambda)$ in Eq. (1.31) must be proportional to p. Thus, in the absence of a potential (when $n = 1$)

$$k^2_{n=1} = \frac{p^2}{h^2/4\pi^2} \qquad (1.104)$$

But,

$$p^2 = 2m\mathcal{E} \qquad (1.105)$$

so

$$k^2_{n=1} = \frac{2m}{\hbar^2}\mathcal{E}, \quad \text{which lead to Eq. 1.102.} \qquad (1.106)$$

To account for the potential we let $n^2 = (\mathcal{E} - V)/\mathcal{E} \neq 1$ so that $k^2 = n^2 k^2_{n=1}$ is just the coefficient of the second term in Eq 1.103.

Space does not permit a more in depth discussion of the quantum nature of light. The interested reader is directed to a recent review Smith and Raymer, 2007 and references therein.

---

[22] The minimum spread criterion applies to conjugate variables such as time frequency and position momentum.

References

Arfken, G. and Weber, H. (2000) *Mathematical Methods for Physicists*, 5th edn, Academic Press, New York. ISBN 0-12-059825-6.

Boas, M.L. (2006) *Mathematical Methods in the Physical Sciences*, 3rd edn, Wiley, New York. ISBN 978-0-471-19826-0.

Cherenkov, P.A., Frank, I.M. and Tamm, I.Y. (1958) Nobel Prize. http://nobelprize.org/nobel_prizes/physics/laureates/1958/.

Clark Jones, R. (1941) New calculus for the treatment of optical systems. I. Description and discussion of the calculus. *J. Opt. Soc. Am.*, **31**, 488.

Cowan, E.W. (1968) *Basic Electromagnetism*, Academic Press, New York. ISBN 0-12-193950-2.

Desyatnikov, A.S., Kivshar, Y.S. and Torner, L. (2005) *Progress in Optics*, Vol. **47**, North-Holland, Publishing Company Amsterdam, 291–391.

Fowles, G.F. (1968) *Introduction to Modern Optics*, Holt, Rinehart and Winston, Inc, New York. ISBN 0-03-065365-7.

Glauber, R.J. (2005) Nobel Prize. http://nobelprize.org/nobel_prizes/physics/laureates/2005/glauber-lecture%.html.

Goldstein, D. (2003) *Polarized Light*, 2nd edn, Marcel Dekker, Inc., New York. ISBN 0-8247-4053-X, Revised and Expanded.

Hartle, J.B. (2003) *Gravity: An Introduction to Einstein's General Relativity*, Addison Wesley, San Francisco. ISBN 0805386629.

Henry, R.W. and Glotzer, S.C. (1988) A squeezed-state primer. *Am. J. Phys.*, **56**, 318.

Kivsha, Y.S. and Agrawol, G.P. (2003) *Optical Solitons*, Academic Press, New York.

Lax, M., Louisell, W.H. and McKnight, W.B. (1975) From Maxwell to paraxial wave optics. *Phys. Rev. A*, **11**, 1365.

Lewis, G.N. (1926) *Nature*, **118**, 874.

Michelson, A.A. (1881) The relative motion of the earth and the luminiferous aether. *Am. J. Sci.*, **22**, 120.

Michelson, A.A. and Morley, E.W. (1887) On the relative motion of the earth and the luminiferous ether. *Am. J. Sci.*, **34**, 333.

Michelson, A.A. and Morley, E.W. (1887) On the relative motion of the earth and the luminiferous ether. *Philos. Mag., series 5*, **524**, 449.

Misner, C.W., Thorne, K.S. and Wheeler, J.A. (1973) *Gravitation*, W. H. Freeman and Company, San Francisco. ISBN 0-7167-0344-0.

Nave, C.R. (2006) *HyperPhysics: Electricity and Magnetism; EM Waves; Electromagnetic Spectrum*, Georgia State University, http://hyperphysics.phy-astr.gsu.edu/Hbase/ligcon.html#c1.

Pismen, L.M. (1999) *Vortices in Nonlinear Fields*, Clarendon Press, Oxford. ISBN 0-935702-11-3.

Scully, M.O. and Zubairy, M.S. (1991) Simple laser accelerator: optics and particle dynamics. *Phys. Rev. A*, **44**, 2656.

Siegman, A.E. (1986) *Lasers*, University Science Books, Sausalito. ISBN 0-935702-11-3.

Smith, B.J. and Raymer, M.G. (2007) Photon wave functions, wave-packet quantization of light, and coherence theory. *New J. Phys.*, **9**, 414.

Further Reading

Bockasten, K. (1974) *Phys. Rev. A*, **9**, 1087.

Born, M. and Wolf, E. (1999) *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, 7th edn, Cambridge University Press, New York. ISBN 0 521 64222 1.

Cowan, R.D. (1981) *The Theory of Atomic Structure and Spectra*, University of California Press, Los Angeles. ISBN 0-520-03821-5.

Ditchburn, R.W. (1961) *Light*, Dover Publications, New York. ISBN.

Durnin, J. (1987) Exact solutions for diffracting beams. I. The scalar theory. *J. Opt. Soc. Am. A*, **4**, 651.

Edmonds, A.R. (1974) *Angular Momentum in Quantum Mechanics*, Princeton University Press, Princeton. ISBN 0-691-07912-9, Third Printing with Corrections.

Einstein, A. (1905) *Ann. Phys.*, **17**, 132.

Einstein, A. (1921) Nobel Prize.

http://nobelprize.org/nobel_prizes/physics/laureates/1921/.

Faisal, F.H. (1987) *Theory of Multiphoton Processes*, Plenum Press, New York. ISBN 0-306-42317-0.

Gallagher, T.F. (1994) *Rydberg Atoms, Atomic, Molecular and Chemical Physics*, Cambridge University Press, New York. ISBN 0-521-38531-8 hardback Engin QC454.A8S27 1994.

Herzberg, G. (1950) *Molecular Spectra and Molecular Structure I: Spectra of Diatomic Molecules*, Van Nostrand Reinhold Company, New York.

Johnson, C.S. Jr. and Pedersen, L.G. (1986) *Problems and Solutions in Quantum Chemistry and Physics*. Dover Publications, New York.

Kim, M.S. (2008) Recent developments in photon-level operations on travelling light fields. *J. Phys. B At. Mol. Opt. Phys.*, **41**, 8, 133001.

Planck, M.K.E.L. (1918) Nobel Prize. http://nobelprize.org/nobel_prizes/physics/laureates/1918/planck-bio.html.

Messiah, A. (2000) *Quantum Mechanics*, Dover Publications, New York. ISBN 0-48-640924-4.

Planck, M. (1900) *Verh. dt. phys. Ges.*, **2**, 202.

Planck, M. (1900) *Verh. dt. phys. Ges.*, **2**, 237.

Rayeigh, L. (1900) *Phil. Mag.*, **49**, 539.

Sakurai, J.J. (1973) *Advanced Quantum Mechanics*, Addison-Wesley Publishing Company, Inc, Menlo Park.

Salpeter, E.E. and Bethe, H.A. (1977) *Quantum Mechanics of One- and Two-Electron Atoms*, Plenum Publishing Corporation, New York. ISBN 0-306-20022-8, First paperback printing.

Scully, M.O. and Suhail Zubairy, M. (1999) *Quantum Optics*, Cambridge University Press, New York. ISBN 0-521-43595-1.

Sobelman, I.I. (1979) *Atomic Spectra and Radiative Transitions*, Springer Series in Chemical Physics. Spring-Verlag, New York. ISBN 0-387-09082-7.

ter Haar, D. (1967) *The Old Quantum Theory*, Pergamon, New York. ISBN.

Turunen, J., Vasara, A. and Friberg, A.T. (1989) Realization of general nondiffracting beams with computer-generated holograms. *J. Opt. Soc. Am. A*, **6**, 1748.

Walker Jon Mathews, R.L. (1970) *Mathematical Methods of Physics*, 2nd edn, Benjamin Cummings, San Francisco. ISBN 0-8053-7002-1.

ELECTROMAGNETIC WAVE PROPAGATION

DAVID C. CHANG AND EDWARD F. KUESTER, *University of Colorado at Boulder, Boulder, Colorado, U.S.A.*

	Introduction	1
1.	**Background**	1
2.	**Plane Waves in Media**	2
2.1	Propagation in Linear, Nondispersive, and Isotropic Media	2
2.2	Wave Propagation in Dispersive Media	5
2.3	Wave Propagation in Anisotropic Media	6
2.4	Wave Propagation in Chiral Media	8
3.	**Propagation of Gaussian Beams**	9
4.	**Reflection and Transmission of Plane Waves at an Interface**	10
4.1	Total Reflection, Critical Angle, and Brewster Angle	11
4.2	Reflection of a Gaussian Beam: Goos–Hänchen Shift	12
4.3	Wave Propagation in a Multilayer Medium	13
5.	**Guided Waves (I)**	14
5.1	Parallel-Plate Waveguides: TE, TM, and TEM Modes	14
5.2	Rectangular Waveguides	17
6.	**Guided Waves (II)**	19
6.1	Dielectric Waveguides: Surface-Wave Modes	19
6.2	Dielectric-Rib Waveguide and Other Surface-Wave Structures	21
7.	**Multiconductor Transmission Lines**	22
7.1	TEM-Mode Transmission Lines	22
7.2	Quasi-TEM-Mode Transmission Lines	24
	Glossary	25
	Works Cited	26
	Further Reading	26

INTRODUCTION

Rapid advances have been made in recent years in applying electromagnetic wave propagation in natural or man-made environments to telecommunication and to remote sensing. The purpose of this article is to familiarize readers with the fundamental concepts and theories regarding both guided and unguided electromagnetic waves.

1. BACKGROUND

Consider first the macroscopic Maxwell equations governing the behavior of the vector electric field $\mathbf{E}(\mathbf{r},t)$ and the vector magnetic field $\mathbf{H}(\mathbf{r},t)$ in free space due to a known distribution of current and charge sources (Stratton, 1941; Jackson, 1975; Kraus, 1984; Johnk, 1988; see also ELECTRODYNAMICS, CLASSICAL):

$$\epsilon_0\nabla\cdot\mathbf{E}=\rho, \tag{1a}$$

$$\nabla\times\mathbf{E}=-\mu_0\frac{\partial}{\partial t}\mathbf{H}, \tag{1b}$$

$$\mu_0\nabla\cdot\mathbf{H}=0, \tag{1c}$$

$$\nabla\times\mathbf{H}=\mathbf{J}+\epsilon_0\frac{\partial}{\partial t}\mathbf{E}. \tag{1d}$$

Here μ_0 and ϵ_0 are, respectively, the permeability and permittivity of free space: in MKS units, $\mu_0 = 4\pi \times 10^{-9}$ H/m and $\epsilon_0 \simeq 1/36\pi \times 10^{-9}$ F/m. The terms $\mathbf{J}(\mathbf{r},t)$ and $\rho(\mathbf{r},t)$, on the other hand, are respectively the volume densities of electric current and charge which satisfy the charge continuity equation:

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

$$\nabla\cdot\mathbf{J}+\frac{\partial}{\partial t}\rho=0. \tag{2}$$

Of course various types of wave propagation represent possible solutions of the Maxwell equations in a source-free region. In most applications, we are interested in the steady-state solution for harmonic sources operating at an angular frequency of ω. Since the field responses also have to be harmonic in nature, we can for each component of the vector field or each scalar function [say $A(\mathbf{r},t)$] designate a complex function $A(\mathbf{r};\omega)$, such that

$$A(\mathbf{r},t)=A_0\cos(\omega t+\Phi_0)=\mathrm{Re}\{A(\mathbf{r};\omega)e^{j\omega t}\};$$

$$A(\mathbf{r};\omega)=A_0(\mathbf{r};\omega)e^{j\Phi_0(\mathbf{r};\omega)}; \tag{3}$$

so that Maxwell's equations in the frequency domain now take the form

$$\epsilon_0\nabla\cdot\mathbf{E}=\rho, \tag{4a}$$

$$\nabla\times\mathbf{E}=-j\omega\mu_0\mathbf{H}, \tag{4b}$$

$$\mu_0\nabla\cdot\mathbf{H}=0, \tag{4c}$$

$$\nabla\times\mathbf{H}=\mathbf{J}+j\omega\epsilon_0\mathbf{E}. \tag{4d}$$

The time dependence of $\exp(j\omega t)$ is then suppressed and is understood implicitly hereafter. Even in the case of time-varying sources, we often find that it is convenient to convert the problem from the time domain to the frequency domain via a Fourier transform defined as

$$A(\mathbf{r};\omega)=\frac{1}{\pi}\int_{-\infty}^{\infty}A(\mathbf{r},t)e^{-j\omega t}\,dt, \tag{5a}$$

and then convert the solution back to the time domain by an inverse Fourier transform,

$$A(\mathbf{r},t)=\frac{1}{2}\int_{-\infty}^{\infty}A(\mathbf{r};\omega)e^{+j\omega t}\,d\omega. \tag{5b}$$

The constraint that a physically observable function must be both casual and real can be restated in the frequency domain as follows (Nussenzveig, 1972):

$A(\mathbf{r};\omega)=A^*(\mathbf{r};-\omega)$;

$A(\mathbf{r};\omega)$ is analytic in the upper half of the complex ω plane. (6)

In Sec. 2, we shall examine in some detail how different materials can influence the propagation of a plane wave.

2. PLANE WAVES IN MEDIA

2.1. Propagation in Linear, Non-dispersive, and Isotropic Media

When current and charge sources are situated inside a material region, they act as primary (impressed) sources to generate electric and magnetic fields which are capable of polarizing, or reorienting the charges in, the molecules or atoms of the medium, or forcing the free electrons to flow in the same or, in some cases, a different direction depending upon the nature of the material. The net effect, on the macroscopic scale, is to produce continuously distributed current and charge densities, which then act as secondary sources and which, in turn, set up secondary fields. In the case of a dielectric material, the volume density of electric polarization **P** is proportional to the applied electric field, i.e., $\mathbf{P}=\chi_e\epsilon_0\mathbf{E}$, where χ_e is the dielectric susceptibility of the medium. It is not difficult to show that spatial and time variations of the polarization density are in fact equivalent to having charge and current densities equal to $\nabla\cdot\mathbf{P}$ and $j\omega\mathbf{P}$, respectively. In order to incorporate these secondary sources into the Maxwell equations properly, we find it convenient to introduce an electric flux density **D** and the permittivity of a material ϵ, as follows:

$$\mathbf{D}=\epsilon_0\mathbf{E}+\mathbf{P}=\epsilon_0(1+\chi_e)\mathbf{E}=\epsilon\mathbf{E}. \tag{7a}$$

Likewise in the case of magnetic material, we can define a volume density of magnetization **M**, and a magnetic susceptibility χ_m, according to $\mathbf{M}=\chi_m\mathbf{H}$. While there is no net charge density induced by the magnetization, the equivalent current density can be shown to be $\nabla\times\mathbf{M}$. The effect of such a secondary current on the Maxwell equations can be accounted for by introducing a magnetic flux density **B** and a permeability μ so that

$$\frac{\mathbf{B}}{\mu_0}-\mathbf{M}=\mathbf{H},\quad\text{or}\quad\mathbf{B}=\mu_0(1+\chi_m)\mathbf{H}=\mu\mathbf{H}. \tag{7b}$$

Finally, in the case of conducting material, the induced current density is simply characterized by $\sigma\mathbf{E}$, where σ is the conductivity of the material. Using the definitions for the material parameters ϵ, μ, and σ and their associated equivalent sources in Eq. (4) immediately yields the appropriate Maxwell equations in a material region in the form

$$\nabla\cdot\epsilon\mathbf{E}=\rho, \tag{8a}$$

$$\nabla\times\mathbf{E}=-j\omega\mu\mathbf{H}, \tag{8b}$$

$$\nabla\cdot\mu\mathbf{H}=0, \tag{8c}$$

$$\nabla\times\mathbf{H}=\mathbf{J}+(\sigma+j\omega\epsilon)\mathbf{E}. \tag{8d}$$

Often, we simplify these expressions through the introduction of a complex permittivity $\hat{\epsilon}=\epsilon-j\sigma/\omega$.

To see what basic types of waves can exist, we first consider the case where the only significant variation of the field is along a given, locally constant, direction (denoted here as the z direction in a rectangular coordinate system). We look at a region in which sources are absent, the ones which produce the field being assumed to lie either to the left or to the right of this region. In this case, the Maxwell curl equations (8b) and (8d) reduce to

$$(\sigma+j\omega\epsilon)E_z=0; \quad j\omega\mu H_z=0, \tag{9a}$$

$$\begin{aligned}&\frac{\partial}{\partial z}(\mathbf{e}_z\times\mathbf{E}_t)=-j\omega\mu\mathbf{H}_t;\\&\frac{\partial}{\partial z}(\mathbf{e}_z\times\mathbf{H}_t)=(\sigma+j\omega\epsilon)\mathbf{E}_t.\end{aligned} \tag{9b}$$

Here $\mathbf{e}_z$ is a unit vector in the z-direction, $\mathbf{E}=\mathbf{E}_t+\mathbf{e}_zE_z$, and $\mathbf{H}=\mathbf{H}_t+\mathbf{e}_zH_z$. Equation (9a) indicates that the longitudinal components of both the electric and magnetic field will have to vanish everywhere if the frequency is not zero. Assuming that we can cast the solution in the form of $\mathbf{E}_t=\mathbf{E}_0\exp\{\pm\gamma z\}$ where γ is a yet undetermined constant, the two equations in Eq. (9b) can be used together to obtain

$$\gamma\times\gamma\times\mathbf{E}_t+j\omega\mu(\sigma+j\omega e)\mathbf{E}_t=0, \quad \gamma=\mathbf{e}_z\gamma, \tag{9c}$$

whose solution is readily found to be

$$\begin{aligned}&\mathbf{E}_t=\mathbf{E}_0^+e^{-\gamma z}+\mathbf{E}_0^-e^{+\gamma z};\\&\mathbf{H}_t=\eta^{-1}\mathbf{e}_z\{\mathbf{E}_0^+e^{-\gamma z}-\mathbf{E}_0^-e^{+\gamma z}\},\\&\gamma=[j\omega\mu(\sigma+j\omega\epsilon)]^{1/2}, \quad \eta=j\omega\mu/\gamma,\end{aligned} \tag{10a}$$

$$\text{where } \mathrm{Re}(\gamma)\geqslant 0 \text{ and } \mathrm{Re}(\eta)\geqslant 0, \tag{10b}$$

and $\mathbf{E}_0^\pm$ are two yet-undetermined complex amplitudes. We note that γ is often referred to as the complex propagation constant and η, the characteristic impedance.

Equations (10) underlie some of the most fundamental aspects regarding the propagation of electromagnetic waves. First and foremost, we see that the field components of these waves are basically transverse in nature; in fact, not only are the electric field and the magnetic field orthogonal to each other, but they are also orthogonal to the direction of the propagation as well. Furthermore, the characteristic impedance of these waves, defined as the ratio of the electric to the magnetic field amplitude, remains constant everywhere, and depends only upon the properties of the medium the wave is propagating through and the operating frequency of the source(s). Expressing the complex propagation constant γ in terms of its real and imaginary parts,

$$\gamma=\alpha+j\beta,$$

$$\alpha=\frac{\omega\sqrt{\mu\epsilon}}{\sqrt{2}}\left[\sqrt{1+\left(\frac{\sigma}{\omega\epsilon}\right)^2}-1\right]^{1/2},$$

$$\beta=\frac{\omega\sqrt{\mu\epsilon}}{\sqrt{2}}\left[\sqrt{1+\left(\frac{\sigma}{\omega\epsilon}\right)^2}+1\right]^{1/2},$$

we can readily write down the form of the solution for the electric field in the time domain as

$$\begin{aligned}\mathbf{E}_t^+(z,t)=&\,\mathbf{e}_xE_{0x}^+e^{-\alpha z}\cos(\omega t-\beta z+\Phi_x^+)\\&+\mathbf{e}_yE_{0y}^+e^{-\alpha z}\cos(\omega t-\beta z+\Phi_y^+),\end{aligned} \tag{11a}$$

$$\begin{aligned}\mathbf{E}_t^-(z,t)=&\,\mathbf{e}_xE_{0x}^-e^{+\alpha z}\cos(\omega t+\beta z+\Phi_x^-)\\&+\mathbf{e}_yE_{0y}^-e^{+\alpha z}\cos(\omega t+\beta z+\Phi_y^-).\end{aligned} \tag{11b}$$

The term involving E_{0x}^+ is plotted in Fig. 1, as a function of βz, for a fixed instant of time t, with an indication of the movement of the wave as t increases. As we will discuss in more detail later, a wave which possesses only an x

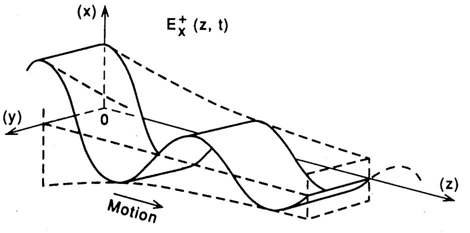

FIG. 1. A forward-propagating wave in a dissipative medium.

component of electric field is often referred to as a wave polarized linearly in the x direction. The velocity of the forward-moving pattern for such a wave can be determined easily by holding the phase constant, giving

$$\omega t-\beta z+\Phi_x^+=\text{constant}\Rightarrow$$

$$\text{phase velocity } v_p=\frac{\partial z}{\partial t}=\frac{\omega}{\beta}. \tag{12}$$

On the other hand, the distance at which the pattern repeats itself is called the wavelength, and is found from

$$\beta\lambda=2\pi\Rightarrow\text{wavelength } \lambda=2\pi/\beta. \tag{13}$$

These two results, when combined together, yield $v_p = \lambda f$, where f is the frequency of the propagating wave. In addition to propagation, the wave may also decay exponentially in the forward direction. The distance $\delta=1/\alpha$ at which the wave has decayed by a factor of e^{-1} is often referred to as the skin depth. This notion is often applied in the case of a good conductor, $\sigma\gg\omega\epsilon$, for which $\alpha\approx\beta\approx[\omega\mu\sigma/2]^{1/2}$.

From the above discussion, it is not difficult to construct a physical picture of a plane wave propagating either in the $+z$ or $-z$ direction with a phase velocity of v_p. For a lossless and dispersionless dielectric material, $\sigma = 0$ and ϵ is a constant, and the phase velocity reduces to the simple expression $[1/\mu\epsilon]^{1/2}$, which is independent of the operating frequency; we sometimes refer to it as the speed of light in the medium. It is of interest to note that when both forward- and backward-propagating waves are present, they can be combined together to form a standing wave pattern. In particular, the magnitude of the total electric field linearly polarized in the x direction is given as

$$|E_x(z)|=|E_{0x}^+|\{1+|\Gamma|^2+2|\Gamma|\cos(2\beta z+\Phi_\Gamma)\}^{1/2}, \tag{14a}$$

where

$$\Gamma=|\Gamma|e^{j\Phi_\Gamma}=E_{0x}^-/E_{0x}^+. \tag{14b}$$

When the backward-traveling wave arises as the result of a reflection of the forward-traveling wave, the complex quantity Γ, defined as the ratio of backward and forward wave amplitudes, is referred to as the reflection coefficient. As seen in Fig. 2, the magnitude of the total electric field oscillates between $1+|\Gamma|$ and $1-|\Gamma|$. The ratio of maximum to minimum values in the standing wave pattern thus provides a measure of how strong the reflection is, and it is called the standing wave ratio (SWR):

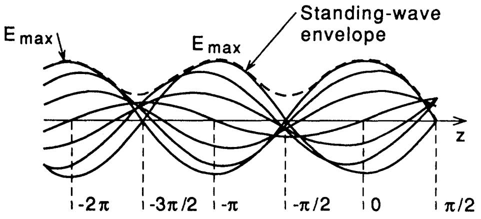

FIG. 2. Standing wave pattern.

$$\text{SWR}=\frac{1+|\Gamma|}{1-|\Gamma|}. \tag{15}$$

The discussion so far has been restricted to linearly polarized waves, when only one of the two components of the electric field is excited. For many applications, both components are often present, and depending upon their relative magnitudes and phases this results in various kinds of polarization. To see how this may occur, we now assume that the solution in the region of interest consists of only a forward-propagating wave, so that at a fixed observation point, say $z = 0$, the electric field in the frequency domain is given as

$$\mathbf{E}_t=\mathbf{e}_xE_x+\mathbf{e}_yE_y,$$

$$E_x=E_{0x}\,e^{j\Phi_x} \text{ and } E_y=E_{0y}\,e^{j\Phi_y}, \tag{16a}$$

or in the time domain,

$$\mathbf{E}_t=\mathbf{e}_xE_{0x}\cos(\omega t+\Phi_x) +\mathbf{e}_yE_{0y}\cos(\omega t+\Phi_y). \tag{16b}$$

We can clearly see from Eq. (16b) that both the direction and magnitude of the electric field vary as functions of ωt. In fact, the equation that governs the relationship between the x and y components can be derived from Eq. (16b) as

$$\left(\frac{E_x}{E_{0x}}\right)^2+\left(\frac{E_y}{E_{0y}}\right)^2-2\cos(\Phi_x-\Phi_y)\left(\frac{E_x}{E_{0x}}\right)\left(\frac{E_y}{E_{0y}}\right) =\sin^2(\Phi_x-\Phi_y), \tag{17a}$$

which describes a tilted ellipse with major and minor axes L_0 and l_0, respectively, where

$$2L_0^2=E_{0x}^2+E_{0y}^2+[E_{0x}^4+E_{0y}^4+2E_{0x}^2E_{0y}^2\cos 2(\Phi_x-\Phi_y)]^{1/2}, \qquad (17b)$$

$$2l_0^2=E_{0x}^2+E_{0y}^2-[E_{0x}^4+E_{0y}^4+2E_{0x}^2E_{0y}^2\cos 2(\Phi_x-\Phi_y)]^{1/2}, \qquad (17c)$$

and with a tilt angle τ given by

$$2\tau=\pi-\tan^{-1}\left[\frac{2E_{0x}E_{0y}\cos(\Phi_x-\Phi_y)}{E_{0x}^2+E_{0y}^2}\right]. \qquad (17d)$$

We should note that although Eqs. (17a)–(17c) depend only upon the magnitude, but not upon the sign, of the phase difference $\Phi_x-\Phi_y$, it is not difficult to show that the $+$sign traces the ellipse in the clockwise, or right-hand, direction, while the $-$ sign traces the ellipse in the counterclockwise, or left-hand, direction. The polarization state of a plane wave, thus, depends not only on the ratio of E_{0x}^2 to E_{0y}^2, but on the phase difference of the two components as well. Figure 3, for instance, shows a right-hand circularly polarized wave vs a left-hand circularly polarized wave when $E_{0x}^2=E_{0y}^2$ and $\Phi_x-\Phi_y=\pm\pi/2$.

2.2 Wave Propagation in Dispersive Media

Our discussion on plane-wave propagation so far has been confined to harmonic sources with a single operating frequency. In principle, we can always obtain the field response due to arbitrarily time-varying sources by superposition so long as the medium in which it propagates is linear. However, because characteristics of plane waves such as the phase velocity are dependent on the operating frequency, the frequency content in a resultant field can vary from one location to another and, consequently, cause dispersion in the signal we intend to transmit. Before we can see how this could happen, let us first look at the case of a lossless dielectric medium and familiarize ourselves initially with the propagation of an impulse signal which does not exhibit the dispersion phenomenon. Suppose that we have a linearly polarized wave propagating in the forward z direction and behaving like an impulse function initially at $z=0$, i.e., $E_x(0,t)=\delta(t)$. We can find an expression for this wave at other locations from Eqs. (5a), (5b), and (10) for a lossless medium as

$$E_x(z,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-j\beta z+j\omega t}\,d\omega=\delta\left(t-\frac{z}{v_p}\right), \qquad (18)$$

since $\sigma=0$, $\gamma=j\beta=j\omega/v_p$, $\alpha=0$, and $v_p=(\mu\epsilon)^{-1/2}$. What Eq. (18) says is that we can expect the impulse signal to remain essentially intact when it propagates in a lossless medium, assuming that both the permittivity ϵ and the permeability μ are not functions of frequency. However, for a lossy medium where the conductivity is nonvanishing, the corresponding solution for the electric field becomes

$$E_x(z,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\gamma z+j\omega t}d\omega, \qquad (19a)$$

$$\gamma=\frac{j}{v_0}[\omega(\omega-j\Delta)]^{1/2},$$

where $v_0=(\mu\epsilon)^{-1/2}$ and $\Delta=\sigma/\epsilon$. By deforming the integration contour in Eq. (19a) from the real axis, we can readily show (Stratton, 1941) that for $t<z/v_0$, $E_x(z,t)$ is zero, and for $t>z/v_p$,

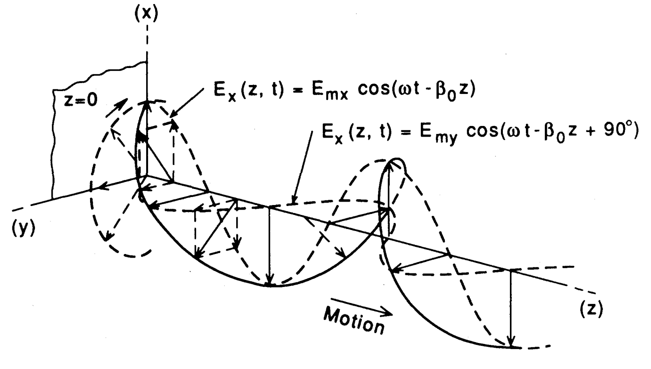
(x)
z=0
$E_x(z, t) = E_{mx}\cos(\omega t-\beta_0 z)$
$E_x(z, t) = E_{my}\cos(\omega t-\beta_0 z + 90°)$
(y)
Motion
(z)

FIG. 3. A right-hand circularly polarized wave.

$$E_x(z,t)=\frac{j}{2\pi}\int_0^{\Delta}\left[\exp\left(-j\frac{z}{v_0}[\nu(\Delta-\nu)]^{1/2}\right)-\exp\left(+j\frac{z}{v_0}[\nu(\Delta-\nu)]^{1/2}\right)\right]e^{-\nu t}\,d\nu$$
$$=\left(\frac{\Delta}{2v_0}\right)\frac{z/v_0}{[t^2-(z/v_0)^2]^{1/2}}I_1\left(\frac{\Delta}{2}[t^2-(z/v_0)^2]^{1/2}\right)e^{-\Delta t/2},\qquad (19b)$$

where I_1 is the modified Bessel function of the first kind and order one. While the behavior of such a field response is rather complicated upon its arrival, i.e., at $t\approx z/v_0$, it turns out that, at later times when $t^2 \gg (z/v_0)^2$ and $t\Delta \gg 1$, then $E_x(z,t) \sim t^{-3/2}$ and thus the field decays only algebraically as the result of the dispersion in a conducting medium.

The dispersion phenomenon we have discussed so far has assumed very particular forms for the frequency dependences of the attenuation and phase velocity of the constituents of the transient waves. Specifically, we have assumed that the material parameters ϵ, σ, and μ remain constant throughout the entire frequency spectrum of interest. It turns out that these assumptions, while valid for signals of mainly low-frequency content, become less accurate at higher frequencies as a result of molecular and other resonances in the material. As an example, the complex relative permittivity $\hat{\epsilon}_r=\hat{\epsilon}/\epsilon_0=\epsilon_r'-j\epsilon_r''$ for a dielectric material that can be characterized by only a single molecular resonance frequency is given as

$$\epsilon_r'=1+\left(\frac{N_e q^2}{\epsilon_0 m}\right)\frac{(\omega_0^2-\omega^2)}{(\omega_0^2-\omega^2)^2+(\omega\nu/m)^2},$$
$$\epsilon_r''=\left(\frac{N_e q^2}{\epsilon_0 m}\right)\frac{(\omega\nu/m)}{(\omega_0^2-\omega^2)^2+(\omega\nu/m)^2},\qquad (20)$$

where q, m, and ν are, respectively, the charge, effective mass, and damping coefficient of a polarized molecule; N_e is the number of polarized molecules per unit volume. The frequency ω_0, on the other hand, is the resonant frequency of the molecules, and it is related to the elastic coefficient s and the effective mass m according to $\omega_0 = (s/m)^{1/2}$. The investigation of wave dispersion in media where the constitutive parameters cannot be regarded as frequency-independent follows the same principles as laid out earlier in this section, but the more cumbersome analytical details are beyond the scope of this article.

Regardless of the specific form of their frequency dependence, the permittivity and permeability characterize physical systems whose inputs and outputs are physical quantities like **E**, **B**, etc., which must satisfy the constraints listed in Eq. (6), and it is not difficult to show (Nussenzveig, 1972) from causality considerations that their real and imaginary parts, as functions of frequency, must satisfy the so-called Kramers–Kronig relations, which can be stated in the following form:

$$\kappa'(\omega)=1+\frac{2}{\pi}\int_0^{\infty}\frac{\omega'\kappa''(\omega')}{(\omega')^2-\omega^2}d\omega',$$
$$\kappa''(\omega)=\frac{2\omega}{\pi}\int_0^{\infty}\frac{1-\kappa'(\omega')}{(\omega')^2-\omega^2}d\omega',\qquad (21)$$

where $\kappa=\kappa'+j\kappa''$ stands for either the relative permittivity ϵ_r $(=\epsilon/\epsilon_0)$ or the relative permeability. We have also assumed that $\epsilon \rightarrow \epsilon_0$ and $\mu \rightarrow \mu_0$ as $\omega \rightarrow \infty$, on the basis that a real medium will not be able to respond substantially to a field of high enough frequency.

2.3 Wave Propagation in Anisotropic Media

In addition to dispersion, complication also arises when electromagnetic waves propagate in a medium which is anisotropic (i.e., where **B** and **H** and/or **D** and **E** are not parallel in general). In such media, the permeability and/or permittivity is given as a tensor. The Maxwell equations in Eqs. (8a)–(8d) are modified for a dielectrically anisotropic material to read

$$\nabla\times\mathbf{E}=-j\omega\mu_0\mathbf{H},\quad \nabla\times\mathbf{H}=j\omega\epsilon_0\epsilon_r\cdot\mathbf{E},\qquad (22)$$

so that the sum of the displacement and polarization currents, $j\omega\epsilon_0\epsilon_r\cdot\mathbf{E}$, is no longer necessarily parallel to the direction of the applied electric field. Many types of materials exhibit anisotropy: ferrites, crystalline materials such as calcite or fluorspar, and plasmas in the presence of a dc magnetic field, for example (Yariv and Yeh, 1984; Kong, 1986).

A plasma medium biased by the dc magnetic field $\mathbf{e}_z H_0$ exhibits the tensor relative permittivity (Ishimaru, 1991)

$$\epsilon_r=[\epsilon_r]=\begin{bmatrix}\epsilon_t & ja & 0\\ -ja & \epsilon_t & 0\\ 0 & 0 & \epsilon_z\end{bmatrix}, \tag{23a}$$

where

$$\epsilon_t=1-\frac{(\omega_p/\omega)^2[1-j(\nu/\omega)]}{[1-j(\nu/\omega)]^2-(\omega_c/\omega)^2},$$

$$a=\frac{-(\omega_p/\omega)^2(\omega_c/\omega)}{[1-j(\nu/\omega)]^2-(\omega_c/\omega)^2}, \tag{23b}$$

$$\epsilon_z=1-\frac{(\omega_p/\omega)^2}{1-j(\nu/\omega)}$$

and ω_p and ω_c are, respectively, the plasma and cyclotron frequencies, given by

$$\omega_p=\left(\frac{N_e q^2}{m\epsilon_0}\right)^{1/2}, \quad \omega_c=\frac{q\mu_0 H_0}{m}. \tag{23c}$$

To see how waves propagate in such a material, assume that a plane wave has the form $\exp(j\boldsymbol{\beta}\cdot\mathbf{r})$, where $\boldsymbol{\beta}$ is a yet-undetermined vector propagation constant pointing in the direction of propagation. Substitution of this expression into Eq. (22) yields the following governing equation:

$$\boldsymbol{\beta}\times\boldsymbol{\beta}\times\mathbf{E}+\omega^2\mu_0\epsilon_0\epsilon_r\cdot\mathbf{E}=0. \tag{24}$$

We will consider two special categories of solutions to this equation:

(i) those which propagate parallel to the direction of the dc magnetization field,

(ii) those which propagate perpendicularly to the direction of the dc magnetization field.

General oblique propagation involves considerably more cumbersome expressions, and will not be considered here. In this case, the propagation direction as defined by the propagation of a narrow beam or by the vector group velocity (cf. Secs. 3 and 5.1) is not necessarily the same as the normal to planes of constant phase, a common occurrence in anisotropic media.

In the first case, $\boldsymbol{\beta}=\mathbf{e}_z\beta$, and Eq. (24) in matrix form becomes

$$\begin{bmatrix}-\beta^2+k_0^2\epsilon_t & jk_0^2a & 0\\ -jk_0^2a & -\beta^2+k_0^2\epsilon_t & 0\\ 0 & 0 & k_0^2\epsilon_t\end{bmatrix}\begin{bmatrix}E_x\\ E_y\\ E_z\end{bmatrix}=0,$$

$$k_0^2=\omega^2\mu_0\epsilon_0,$$

which can have a nontrivial solution only when the determinant of the matrix is zero. Solving for β yields two sets of acceptable solutions:

Ordinary wave, $\beta=\beta_+=k_0(\epsilon_t-a)^{1/2}$,

$$E_x=-jE_y, \quad E_z=0; \tag{25a}$$

Extraordinary wave, $\beta=\beta_-=k_0(\epsilon_t+a)^{1/2}$,

$$E_x=+jE_y, \quad E_z=0. \tag{25b}$$

Thus, instead of one, there are actually two possible types (ordinary and extraordinary) of plane waves that may exist. The ordinary wave is left-hand circularly polarized, while the extraordinary wave is right-hand circularly polarized. Indeed, the phase velocity is seen to depend on the polarization of the wave, as is also often the case in anisotropic media.

In the case of a collision-free plasma ($\nu=0$), the ordinary wave can propagate in the forward z direction only if $\epsilon_t-a>0$, or $\omega>\omega_1$, where $\omega_1=[\omega_p^2+(\omega_c/2)^2]^{1/2}-\omega_c/2$. The extraordinary wave, on the other hand, can propagate in the forward z direction only if $\epsilon_t+a>0$, or for frequencies in the range of either $0<\omega<\omega_c$ or $\omega>\omega_2$, where $\omega_2=[\omega_p^2+(\omega_c/2)^2]^{1/2}+\omega_c/2$. It is of interest to note that such a phenomenon can be used advantageously in many applications. Figure 4 shows regions in a β–ω diagram for each type of wave discussed above.

A second category of waves occurs when the propagation is orthogonal to the dc bias

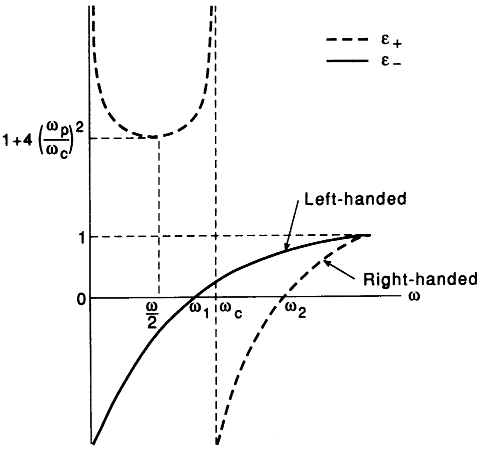

FIG. 4. The ω–β diagram for the left- and right-hand polarized plane waves in an anisotropic medium.

field. Substituting $\boldsymbol{\beta}=\mathbf{e}_x\beta$ into Eq. (24), the matrix equation becomes

$$\begin{bmatrix} k_0^2\epsilon_t & jk_0^2 a & 0 \\ -jk_0^2 a & k_0^2\epsilon_t-\beta^2 & 0 \\ 0 & 0 & k_0^2\epsilon_z-\beta^2 \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} = 0,$$

which again has two acceptable types of solutions:

$$\text{Transverse wave,}\quad \beta=\beta_t=k_0\epsilon_z^{1/2},$$

$$E_x=E_y=0, \quad\text{and}\quad E_z\neq 0; \tag{26a}$$

$$\text{Longitudinal wave,}\quad \beta=\beta_l=k_0\left(\frac{\epsilon^2-a^2}{\epsilon}\right)^{1/2},$$

$$E_x=\left(\frac{-ja}{\epsilon}\right)E_y, \quad\text{and}\quad E_z=0. \tag{26b}$$

The transverse wave is in fact similar to an ordinary plane wave in an isotropic medium since its propagation constant is independent of the dc magnetization field. This is hardly surprising since in this case the sum of the polarization and displacement currents, $j\omega\epsilon_0\epsilon_z E_z$, is in the direction of the dc bias and is thus unaffected by its presence. The longitudinal wave, on the other hand, can exist only when both $\epsilon_t + a$ and $\epsilon_t - a$ are greater than zero. Unlike a plane wave in an isotropic medium, the electric field of this wave has a component in the direction of propagation and is elliptically polarized.

2.4 Wave Propagation in Chiral Media

In recent years a class of materials of considerable interest has been the so-called chiral materials (Lakhtakia *et al.*, 1989) which exhibit a bi-isotropic behavior. Typically, a chiral material is characterized by long molecules in the form of a helix. When the pitch of the helix is neither too great nor too small (so that it does not resemble either a circular loop or a string), an applied electric field or magnetic field alone produces volume densities of both polarization and magnetization—the "handedness" of the helix results in terms proportional to the applied field itself, but also in terms proportional to its curl. Such a property can best be described in terms of how the electric and magnetic flux densities **D** and **B** are related to the electric and magnetic field intensities **E** and **H**, even though the detailed constituent relationships are very involved, and not as well understood as those of isotropic or conventional anisotropic media:

$$\mathbf{D}=(\epsilon+\kappa^2\mu)\mathbf{E}-j\kappa\mu\mathbf{H}, \quad \mathbf{B}=j\kappa\mu\mathbf{E}+\mu\mathbf{H}, \tag{27a}$$

where κ is called the chirality parameter of the medium. Clearly, when $\kappa=0$, Eq. (27a) reduces to the constitutive relations for an isotropic medium. The corresponding Maxwell equations in a source-free region for this case become

$$\begin{aligned} \nabla\times\mathbf{E}&=-j\omega(j\kappa\mu\mathbf{E}+\mu\mathbf{H}), \\ \nabla\times\mathbf{H}&=j\omega[(\epsilon+\kappa^2\mu)\mathbf{E}-j\kappa\mu\mathbf{H}]. \end{aligned} \tag{27b}$$

Thus for a plane wave of the form $\exp(j\boldsymbol{\beta}\cdot\mathbf{r})$, the governing equation readily reduces to

$$\boldsymbol{\beta}\times\boldsymbol{\beta}\times\mathbf{E}+j2\omega\kappa\mu\boldsymbol{\beta}\times\mathbf{E}+\omega^2\mu_0\epsilon_0\mathbf{E}=0, \tag{28}$$

which, upon assuming $\boldsymbol{\beta}=\mathbf{e}_z\beta$, again yields a matrix equation of the following form:

$$\begin{bmatrix} -\beta^2+k_0^2 & j2\omega\kappa\mu\beta \\ -j2\omega\kappa\mu\beta & -\beta^2+k_0^2 \end{bmatrix} \begin{bmatrix} E_x \\ E_y \end{bmatrix} = 0.$$

Setting the determinant equal to zero and solving for β gives two sets of acceptable solutions:

right-hand circularly polarized wave,

$$\beta=\beta_+=\omega\mu\kappa+[(\omega\mu\kappa)^2+k_0^2]^{1/2},$$

$$E_x=jE_y, \quad E_z=0; \tag{29a}$$

left-hand circularly polarized wave,

$$\beta=\beta_-=-\omega\mu\kappa+[(\omega\mu\kappa)^2+k_0^2]^{1/2},$$

$$E_x=-jE_y, \quad E_z=0. \tag{29b}$$

The behavior of these two types of waves in many ways resembles that of the ordinary and extraordinary waves in an anisotropic medium, except that no dc bias field is required. While new device concepts have actively been explored in recent years, the major challenge for microwave applications appears to be the lack of either natural or synthetic materials which exhibit a strong chirality at these frequencies.

3. PROPAGATION OF GAUSSIAN BEAMS

In Sec. 2, propagation of electromagnetic waves was discussed in the context of plane waves under the assumption that field quantities vary only in one direction. This may be a valid assumption when the region of interest is sufficiently remote from the excitation sources. However, the infinite extent of a plane wave in the transverse direction naturally limits our ability to apply it to situations where the field distribution is initially highly collimated in a cross section of finite extent. A good example is a linearly polarized "beam" which initially at $z=0$ has a Gaussian envelope (Arnaud, 1976; Solimeno *et al.*, 1986; Ishimaru, 1991):

$$E_x(x,y;0)=E_0 \exp\{-(x^2+y^2)/2w^2\}. \qquad (30)$$

Here w is the beamwidth used to specify the spatial confinement in the transverse direction. Such a distribution is typically encountered for waves emanating from a laser, in the near field of a reflector antenna, or in a lens waveguide. What we need to know is the behavior of this beam as it propagates in the forward z direction.

For convenience, we first consider the case of a one-dimensional beam in free space: $E_x(x;0)=E_0\exp[-(x^2/2w^2)]$. This aperture field distribution can be rewritten as an inverse Fourier transform integral with respect to the x variable:

$$E_x(x;0)=\left(\frac{1}{2\pi}\right)^{1/2} E_0 w \times \int_{-\infty}^{\infty} e^{-\beta_x^2 w^2/2} e^{-j\beta_x x}\, d\beta_x.$$

We interpret this expression as a superposition of spatial harmonics in the x direction, analogous to the representation of an arbitrary function of time as a superposition of temporal harmonics. For each spatial harmonic β_x, the propagation vector assumes the form $\boldsymbol{\beta}=\mathbf{e}_x\beta_x+\mathbf{e}_z\beta_z$. Substituting into the governing equations in Eq. (9) immediately yields $\beta_z=(k_0^2-\beta_x^2)^{1/2}$ for $\beta_x<k_0$, and $=-j(\beta_x^2-k_0^2)^{1/2}$ for $\beta_x>k_0$. Here, $k_0=\omega(\mu_0\epsilon_0)^{1/2}$ is the wave number in free space. Consequently, we can formally write down the solution in the region $z>0$ as

$$E_x(x;z)=\left(\frac{1}{2\pi}\right)^{1/2} E_0 w \int_{-\infty}^{\infty} e^{-\beta_x^2 w^2/2} \times e^{-j(\beta_x x+\beta_z z)}\, d\beta_x. \qquad (31)$$

What this means is that the wave solution in general consists of a continuous spectrum of homogeneous and inhomogeneous plane waves. In the range $-k_0<\beta_x<k_0$, the waves are propagating without attenuation at an angle $\tan^{-1}(\beta_x/\beta_z)$ with respect to the z axis, and they are called homogeneous plane waves. For $|\beta_x|>k_0$, the waves are propagating in the x direction while decaying exponentially in the z direction and they are called inhomogeneous plane waves. Since the amplitude spectrum of these waves has a Gaussian distribution in β_x, the most important contribution to the spectral integral comes from sufficiently small values of β_x. Thus a beam wave which has a large beamwidth, i.e., $k_0 w \gg 1$, is composed mainly of plane waves which propagate almost parallel to the z axis. But since these constituents are not all propagating exactly parallel to z, as the beam wave moves forward it also has a tendency to lose its original property of being concentrated near the z axis; the beam is said to diverge.

To put the above observations on a more analytical foundation, we use the so-called paraxial approximation, $\beta_z \approx k_0-\beta_x^2/2k_0$, in Eq. (31) to obtain

$$E_x(x;z)=E_0\left(\frac{w}{w_e}\right)\exp\left(-\frac{x^2[1+jz/(k_0w^2)]}{2w_e^2}\right) \times e^{-jk_0 z},$$

$$w_e=w\left[1+\left(\frac{z}{k_0w^2}\right)^2\right]^{1/2}. \qquad (32)$$

Thus, the width of the beam wave is broadened from w to w_e as it propagates with a propagation constant equal to k_0. In addition, the constant-phase front which is usually associated with the propagation of a plane wave also needs to be suitably modified.

The question of how long a beam wave can remain reasonably intact without being noticeably distorted is answered by examining further the validity of the paraxial approximation. The two-term Taylor series expansion is in principle a good approximation if $\beta_x^2 z/2k_0 \ll 1$. Since the values of β_x are confined to $\beta_x w<5$, say, we can define a characteristic length $D_0 = k_0w^2$, so that the paraxial approximation applies so long as $z < D_0$. We sometimes refer to this region as the near-field or the Fresnel zone even though it may be very large compared with the free-space wavelength.

A topic of considerable interest in recent years is the development of a focused beam in free space without the usual divergence effect. It is not difficult to show that if, in addition to a Gaussian amplitude, we also introduce a quadratic phase front in the initial aperture distribution, i.e.,

$$E_x(x;0)=E_0\exp\left\{-\frac{x^2}{2w^2}+j\frac{k_0x^2}{2R_0}\right\} \tag{30a}$$

the effective beamwidth w_e after a distance z is then

$$w_e=w\left[\left(1-\frac{z}{R_0}\right)^2+\left(\frac{z}{D_0}\right)^2\right]^{1/2},\quad D_0=k_0w^2. \tag{31}$$

It is in fact not difficult to show that the smallest beamwidth then occurs at $z=z_{\min}$, where

$$z_{\min}=\left(\frac{D_0^2}{R_0^2+D_0^2}\right)R_0,$$

and the ratio of w_e to w is given by $R_0/(R_0^2+D_0^2)^{1/2}$ in this case. Thus, the challenge of producing a focused beam lies in one's ability to produce an aperture field with a suitable, yet physically realizable, phase distribution.

4. REFLECTION AND TRANSMISSION OF PLANE WAVES AT AN INTERFACE

Our discussion up until now has been confined to various types of waves that can exist in different media. We come now to the question of what a wave will do when it encounters a different medium along the path of propagation. For a solution to be physically acceptable, certain boundary conditions must be met at an interface between two dissimilar media. They can be derived directly from applying the integral form of Maxwell's equations to differentially small volumes and surfaces that enclose both sides of the interface. Defining a unit vector $\mathbf{e}_n$ normal to the interface and pointing from medium No. 2 to medium No. 1, we can write down the boundary conditions as follows:

$$\mathbf{e}_n\cdot(\epsilon_1\mathbf{E}_1-\epsilon_2\mathbf{E}_2)=\rho_s,\quad \mathbf{e}_n\times(\mathbf{E}_1-\mathbf{E}_2)=0, \tag{32a}$$

$$\mathbf{e}_n\cdot(\mu_1\mathbf{H}_1-\mu_2\mathbf{H}_2)=0,\quad \mathbf{e}_n\times(\mathbf{H}_1-\mathbf{H}_2)=\mathbf{J}_s, \tag{32b}$$

where $\mathbf{E}_{1,2}$ and $\mathbf{H}_{1,2}$ are the electric and magnetic fields on the two sides of the interface; ρ_s and $\mathbf{J}_s$ are, respectively, the surface charge density and surface current density at the interface, which are nonvanishing only when one of the media is perfectly conducting.

Suppose that a linearly polarized plane wave in medium No. 1 is incident obliquely from above onto a plane interface at an angle θ with respect to the y axis (which is normal to the interface). As shown in Fig. 5, two types of polarization can exist; in one case, the electric field is in the plane of incidence (the xy plane), and in the other, perpendicular to the plane of incidence. We now further define in medium No. 1 an incident vector propagation constant as γ_i, a reflected one as γ_r, and in medium No. 2, a transmitted one as γ_t, so that

$$\gamma_i=\gamma_1\mathbf{e}_i,\quad \gamma_r=\gamma_1\mathbf{e}_r,\ \text{and}\ \gamma_t=\gamma_2\mathbf{e}_t,$$

where $\gamma_{1,2}=[j\omega\mu_{1,2}(\sigma_{1,2}+j\omega\epsilon_{1,2})]^{1/2}$, and $(\mathbf{e}_i,\mathbf{e}_r,\mathbf{e}_t)$ are unit vectors in the directions of the incident, reflected, and transmitted wave, respectively, i.e.,

$$\mathbf{e}_i=\mathbf{e}_x\sin\theta_i-\mathbf{e}_y\cos\theta_i,$$

$$\mathbf{e}_r=\mathbf{e}_x\sin\theta_r+\mathbf{e}_y\cos\theta_r,$$

$$\mathbf{e}_t=\mathbf{e}_x\sin\theta_t-\mathbf{e}_y\cos\theta_t.$$

The solution in each region is then, for *perpendicular polarization,*

$$\begin{aligned} y>0:\quad &\mathbf{E}_1(\mathbf{r})=\mathbf{e}_z\{E^ie^{-\gamma_i\cdot\mathbf{r}}+E^re^{-\gamma_r\cdot\mathbf{r}}\},\\ &\mathbf{H}_1(\mathbf{r})=\frac{1}{\eta_1}\{(\mathbf{e}_i\times\mathbf{e}_z)E^ie^{-\gamma_i\cdot\mathbf{r}}\\ &\qquad+(\mathbf{e}_r\times\mathbf{e}_z)E^re^{-\gamma_r\cdot\mathbf{r}}; \end{aligned} \tag{33a}$$

$$\begin{aligned} y<0:\quad &\mathbf{E}_2(\mathbf{r})=\mathbf{e}_zE^te^{-\gamma_t\cdot\mathbf{r}},\\ &\mathbf{H}_2(\mathbf{r})=\frac{1}{\eta_2}(\mathbf{e}_t\times\mathbf{e}_z)E^te^{-\gamma_t\cdot\mathbf{r}}; \end{aligned} \tag{33b}$$

and for *parallel polarization,*

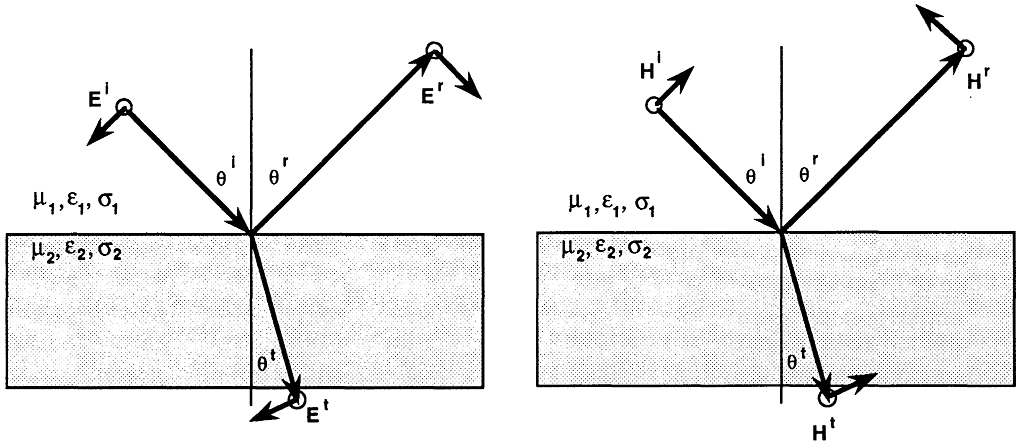

E^i
E^r
θ^i
θ^r
$\mu_1, \varepsilon_1, \sigma_1$
$\mu_2, \varepsilon_2, \sigma_2$
θ^t
E^t
H^i
H^r
$\mu_1, \varepsilon_1, \sigma_1$
$\mu_2, \varepsilon_2, \sigma_2$
θ^i
θ^r
θ^t
H^t

FIG. 5. Plane wave obliquely incident onto an interface.

$$y>0: \quad \mathbf{E}_1(\mathbf{r}) = \{(\mathbf{e}_i \times \mathbf{e}_z) E^i e^{-\gamma_i \cdot \mathbf{r}} + (\mathbf{e}_r \times \mathbf{e}_z) \times E^r e^{-\gamma_r \cdot \mathbf{r}}\},$$

$$\mathbf{H}_1(\mathbf{r}) = -\frac{\mathbf{e}_z}{\eta_1}\{E^i e^{-\gamma_i \cdot \mathbf{r}} + E^r e^{-\gamma_r \cdot \mathbf{r}}\}, \tag{34a}$$

$$y<0: \quad \mathbf{E}_2(\mathbf{r}) = (\mathbf{e}_t \times \mathbf{e}_z) E^t e^{-\gamma_t \cdot \mathbf{r}},$$

$$\mathbf{H}_2(\mathbf{r}) = -\frac{\mathbf{e}_z}{\eta_2} E^t e^{-\gamma_t \cdot \mathbf{r}}. \tag{34b}$$

The requirement that the tangential electric and magnetic field be continuous at $y=0$ provides the equations for obtaining the following solution:

$$\gamma_1 \sin \theta_i = \gamma_1 \sin \theta_r = \gamma_2 \sin \theta_t \quad \text{(Snell's law)}. \tag{35}$$

For perpendicular polarization,

$$\Gamma_\perp = \frac{E^r}{E^i} = \frac{\eta_2 \cos \theta_i - \eta_1 \cos \theta_t}{\eta_2 \cos \theta_i + \eta_1 \cos \theta_t},$$

$$T_\perp = \frac{E^t}{E^i} = \frac{2\eta_2 \cos \theta_i}{\eta_2 \cos \theta_i + \eta_1 \cos \theta_t}. \tag{36}$$

For parallel polarization,

$$\Gamma_\parallel = \frac{E^r}{E^i} = \frac{\eta_2 \cos \theta_t - \eta_1 \cos \theta_i}{\eta_2 \cos \theta_t + \eta_1 \cos \theta_i},$$

$$T_\parallel = \frac{E^t}{E^i} = \frac{2\eta_2 \cos \theta_i}{\eta_2 \cos \theta_t + \eta_1 \cos \theta_i}. \tag{37}$$

Often, and especially for nonmagnetic media where $\mu=\mu_0$, we introduce the refractive index of the medium as $n=\epsilon_r^{1/2}$ and replace the various characteristic impedances η in Eqs. (36) and (37) by $n^{-1}\eta_0$. The fact that the reflection coefficients Γ and the transmission coefficients T are different for the two polarizations, even though they are both linearly polarized waves, may have many interesting applications. For instance, the amount of depolarization in the reflected wave for a circularly polarized incident wave can be used as a sensing mechanism for finding the electrical properties of an unknown material (see ELLIPSOMETRY).

4.1 Total Reflection, Critical Angle, and Brewster Angle

There are two especially interesting physical phenomena associated with a plane wave obliquely incident onto an interface. In one case, the waves are totally reflected with no power transmitted to the second medium, and in the second case, the waves are totally transmitted into the second medium with no power reflected. We shall discuss in some detail below the conditions for the occurrence of these effects.

Let us first look at how the angle of the transmitted wave varies according to the incident angle. According to Snell's law in Eq. (35), a transmitted wave will become more grazing when a wave is incident from a denser medium, i.e., $n_1>n_2$, since $\theta_t=\sin^{-1}[(n_1/n_2)\sin \theta_i]$. This means that when the incident angle is increased beyond the so-called critical angle,

$$\theta_c = \sin^{-1}(n_1/n_2), \tag{38}$$

the transmitted angle reaches $\pi/2$ and thereafter must become complex: the wave can no longer propagate freely in the second medium in the form of a homogeneous plane wave. The complex wave vector corresponding to this angle of transmission is given by

$$\begin{aligned}\gamma_t &= jk_0 n_2 \mathbf{e}_t \\ &= j\mathbf{e}_x k_0 n_1 \sin\theta_i - \mathbf{e}_y k_0 [n_1^2 \sin^2\theta_i - n_2^2]^{1/2}.\end{aligned}$$

The transmitted wave, as given by Eqs. (33b) and (34b) for either polarization, becomes an inhomogeneous plane wave. In other words, the transmitted wave, while propagating with the same phase velocity as the incident and reflected waves, nevertheless decays exponentially away from the interface; it is said to be evanescent. For instance, for the perpendicular polarization, we have

$$\begin{aligned}\mathbf{E}_2(\mathbf{r}) &= \mathbf{e}_z T_\perp E^i \exp\{k_0[n_1^2 \sin^2\theta_i - n_2^2]^{1/2} y\} \\ &\quad \times \exp(-jk_0 n_1 x \sin\theta_i);\end{aligned} \tag{39a}$$

$$\begin{aligned}\mathbf{H}_2(\mathbf{r}) &= \frac{1}{\eta_2}\left\{\mathbf{e}_x\left(\frac{n_1}{n_2}\right)\sin\theta_i \right. \\ &\quad \left. + j\mathbf{e}_y\left[\left(\frac{n_1}{n_2}\right)^2 \sin^2\theta_i - 1\right]^{1/2}\right\} \times \mathbf{E}_2(\mathbf{r}) .\end{aligned} \tag{39b}$$

A similar expression can be derived for the parallel polarization. We should note that this physical phenomenon is often referred to as total reflection, not because there is no longer any electromagnetic field in medium No. 2, but because it is no longer possible to transmit any time-average power into this medium, i.e., $P_{\text{av}} = \frac{1}{2}\text{Re}\{E_z H_x^*\} = 0$.

The second phenomenon of interest is that of total transmission. By setting the reflection coefficient $\Gamma_\parallel$ equal to zero for the case of parallel polarization, we can find explicitly the angle of incidence necessary for this to occur to be

$$\theta_b = \sin^{-1}\left\{\frac{n_2}{(n_1^2 + n_2^2)^{1/2}}\right\} \tag{40}$$

and this angle is called Brewster's angle. It is not difficult to show that the angle for the transmitted wave under this condition is given by $\pi/2 - \theta_b$. This means that the dipoles (molecules) in the second medium, which either are polarized or reoriented in the same direction as the electric field in that medium, are all aligned parallel to the direction the reflected wave would be propagating in (were it not zero in amplitude), and hence do not produce any net radiation in that direction. This helps explain why the reflection is zero at this angle, and, for that matter, why such a phenomenon can occur only in the case of parallel polarization, and not when the incident wave is perpendicularly polarized. The general behaviors for both types of polarization are sketched in Fig. 6 for comparison.

4.2 Reflection of a Gaussian Beam: Goos–Hänchen Shift

The total reflection phenomenon for a wave incident at an angle greater than the critical angle also brings about another interesting physical consequence when the incident wave is in the form of a Gaussian beam. As we explained in Sec. 3, a plane wave with a Gaussian envelope can be viewed as a collimated collection of plane waves propagating at angles slightly different from each other and, as a result, has the tendency to diverge after a certain propagation distance. Recognizing that the direction of propagation for the incident wave is now at an angle θ_i with respect to the direction normal to the inter-

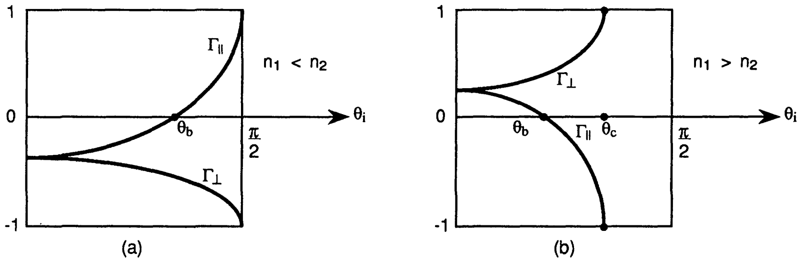

FIG. 6. Reflection coefficients as a function of incident angle.

face in the xy plane, we can rewrite the spectral representation (31) for a perpendicularly polarized incident Gaussian beam by replacing z with $x \sin \theta_i - y \cos \theta_i$, and replacing x with $x \cos \theta_i + y \sin \theta_i$, to yield

$$\mathbf{E}^i(\mathbf{r}) = \mathbf{e}_z \frac{E_0 w}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} e^{-\beta_x^2 w^2/2} \times \exp\{-jk_0 n_1 [x \sin(\theta_i + \theta_{\beta_x}) - y \cos(\theta_i + \theta_{\beta_x})]\} d\beta_x, \tag{41a}$$

where $\theta_{\beta_x} = \sin^{-1}(\beta_x / k_0 n_1)$ represents the angle of propagation for a plane-wave spectral component β_x. A similar expression can be obtained for the parallel polarization case.[1] Now since the angle of spread is typically very small, we can use a small-angle approximation to obtain $\theta_{\beta_x} \simeq \beta_x / k_0 n_1$, so that at the interface $y = 0$, the incident electric field is given approximately as

$$\mathbf{E}^i(\mathbf{r}) = \mathbf{e}_z \frac{E_0 w}{(2\pi)^{1/2}} \times \exp\{-jk_0 n_1 x \sin \theta_i\} W_i(x, 0), \tag{41b}$$

where

$$W_i(x, 0) = \int_{-\infty}^{\infty} e^{-\beta_x^2 w^2/2} e^{-j\beta_x x \cos \theta_i} \, d\beta_x \tag{41c}$$

is the projection of the Gaussian beam along the interface. The projection of the reflected Gaussian beam, on the other hand, is given by

$$W_r(x, 0) = \int_{-\infty}^{\infty} \Gamma_\perp(\theta_i + \theta_{\beta_x}) \times e^{-\beta_x^2 w^2/2} e^{-j\beta_x x \cos \theta_i} \, d\beta_x, \tag{42}$$

where $\Gamma_\perp(\theta_i + \theta_{\beta_x})$ is the reflection coefficient given in Eq. (36). For waves which are incident at an angle greater than the critical angle, the magnitude of $\Gamma_\perp$ is unity, i.e.,

$$\Gamma_\perp(\theta_i + \theta_{\beta_x}) = \exp\{-j\Psi(\theta_i + \theta_{\beta_x})\},$$

$$\Psi(\theta) = 2 \tan^{-1} \left\{ \frac{(n_1^2 \sin^2 \theta - n_2^2)^{1/2}}{n_1 \cos \theta} \right\}.$$

Expanding in a Taylor series with respect to the incident angle θ_i,

$$\Gamma_\perp(\theta_i + \theta_{\beta_x}) \approx \Gamma_\perp(\theta_i) e^{-j\beta_x \Delta \cos \theta_i},$$

$$\Delta = \left(\frac{1}{k_0 n_1 \cos \theta_1} \right) \frac{\partial \Psi}{\partial \theta} \bigg|_{\theta = \theta_1}, \tag{43}$$

and substituting Eq. (43) into Eq. (42) yields the following result:

$$W_r(x, 0) = \Gamma_\perp(\theta_i) W_i(x + \Delta, 0).$$

The reflected Gaussian beam therefore has an amplitude of $\Gamma_\perp(\theta_i)$ as might be expected. It also has a projected envelope similar to that of the incident beam, except that it is shifted by an amount Δ in the x direction, commonly referred to as the Goos–Hänchen shift (Arnaud, 1976; Solimeno *et al.*, 1986; Ishimaru, 1991). This shift is caused by the slightly different phases of the spectral components making up the reflected wave, even though they are all totally reflected from the interface.

4.3. Wave Propagation in a Multilayer Medium

Analysis of plane waves in a multilayer medium follows basically the same approach as in the two-layer case, except that the solutions in the intermediate layers contain both forward and backward propagating waves bouncing between the two interfaces. For convenience, we will consider only the case of a linearly polarized plane wave incident normally onto a multilayer medium as shown in Fig. 7 so that for the νth layer, having a thickness d_ν and refractive index n_ν, the general plane-wave solution can be written in the following form:

$$\mathbf{E}_\nu(z) = \mathbf{e}_x \{E_\nu^+ \, e^{-j\beta_\nu z} + E_\nu^- \, e^{+j\beta_\nu z}\},$$

$$\mathbf{H}_\nu(z) = \mathbf{e}_y \frac{1}{\eta_\nu} \{E_\nu^+ e^{-j\beta_\nu z} - E_\nu^- e^{+j\beta_\nu z}\}, \tag{44}$$

where $\beta_\nu = n_\nu k_0$ is the propagation constant and $\eta_\nu = \eta_0 / n_\nu$ is the characteristic impedance of the νth layer. Before proceeding to find the wave amplitudes $E_\nu^\pm$ it is sometimes more convenient first to define a reflection coefficient $\Gamma_\nu(z)$ as the ratio of the backward to the forward wave, and a total-field impedance $Z_\nu(z)$ as the ratio of the total electric field to the total magnetic field, so that

---

[1] The reader is reminded that we have changed the coordinate system for the plane of incidence from the xz plane in Eq. (31) to the xy plane in Eq. (41a) in order to be consistent with the equations we have derived in this section.

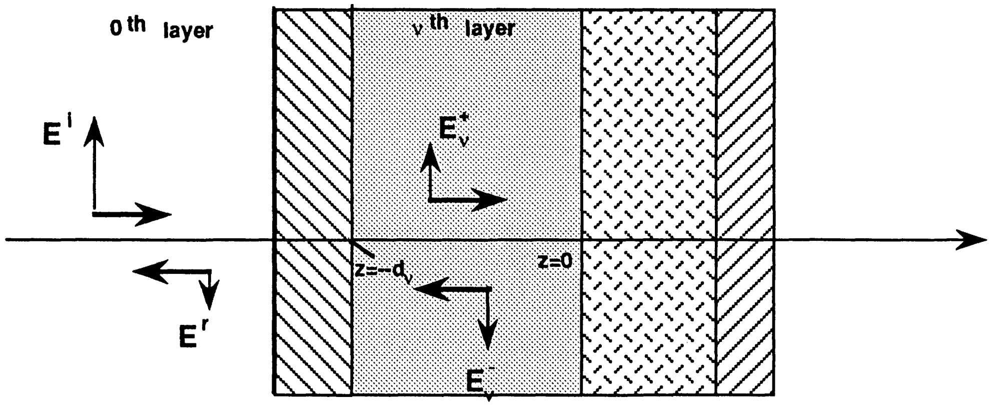

FIG. 7. Wave propagation in a multilayer medium.

$$\begin{aligned}&\Gamma_v(z_1)=\Gamma_v(z_2)e^{j2\beta_v(z_1-z_2)},\\&Z_v(z)=\eta_v\left\{\frac{1+\Gamma_v(z)}{1-\Gamma_v(z)}\right\}.\end{aligned}\tag{45}$$

Equation (45) provides us a convenient way to relate the solution in the v th layer to those in the $(v\pm1)$th layers. For instance, using a local coordinate system with $z=0$ set at the right-hand side of each layer shown in Fig. 7, it is not difficult to show from Eq. (45) that at the interface on the left, i.e., $z=-d_v$, the total-field impedance is given by

$$Z_v(-d_v)=\eta_v\left\{\frac{Z_v(0)+j\eta_v\tan\beta_vd_v}{\eta_v+jZ_v(0)\tan\beta_vd_v}\right\}.\tag{46}$$

Since the boundary conditions that the tangential electric and magnetic fields at an interface between two layers be continuous imply the continuity of the total-field impedance there, i.e., $Z_v(0)=Z_{v+1}(-d_{v+1})$ and $Z_{v-1}(0)=Z_v(-d_v)$, we can therefore obtain the value of Z in any layer if it is known in another layer. In particular, since there is no reflected wave in the last (say the Mth) layer, $\Gamma_M=0$ and $Z_{M-1}(0)=\eta_M$, and we can then proceed to track the value of Z, and hence the reflection coefficient Γ , back to the incident region, and then obtain the amplitude of the reflected wave in that region.

A useful device that illustrates these principles is the quarter-wave transformer. As seen from Eq. (45), the relation between $Z_v(0)$ and $Z_v(-d_v)$ reduces to a very simple one when the thickness d_v equals a quarter of the wavelength in that medium, i.e., $\beta_vd_v=2\pi d_v/\lambda_v=\pi/2$. In this situation, $Z_v(-d_v)=\eta_v^2/Z_v(0)$. This means we can match the total-field impedance at the input end to any positive real value, if the total-field impedance at the other end is positive real and if we can suitably choose the material parameters and thus the characteristic impedance η_v in that medium. For instance, we may wish to minimize the amount of reflection when a plane wave is incident from one medium onto another. A possible solution is to insert an intermediate layer with material parameters chosen in such a way that the characteristic impedance in that layer is given by $\eta_2=(\eta_1\eta_3)^{1/2}$, where η_1 and η_3 are, respectively, the characteristic impedances for the incident and transmitted regions. No reflection will then be encountered when the thickness of the matching layer thickness is 1/4 of the wavelength in that layer, since according to Eq. (44), $\Gamma_1=0$ when $Z_1(0)=\eta_1$.

5. GUIDED WAVES (I)

5.1 Parallel-Plate Waveguides: TE, TM, and TEM Modes

In Secs. 2–4, we have discussed how plane waves propagate in different environments; in most cases, we assumed that the regions of interest were unbounded in extent perpendicular to the direction of propagation. Obviously such an assumption restricts the kind of problems we can tackle, particularly those where we wish to guide the wave propagation in some prescribed manner. In this and the following section, we will discuss physical phenomena associated with guided waves which are supported by guiding structures in either a closed or an open environment (Ramo *et al.*, 1984; Pozar, 1990).

A parallel-plate waveguide consists of a region bounded by two infinite parallel metallic plates separated by a distance a as shown in Fig. 8. Although we are still interested in regions away from any source, possible solutions of the Maxwell equation are now subject to the boundary conditions on the metallic walls; e.g., the tangential electric field must vanish if we assume the two walls are perfectly conducting.

We start by examining the case of a plane wave propagating in the z direction, parallel to the walls. Assuming that the parallel-plate waveguide, i.e., the region between the two walls, is filled with a lossless dielectric material of permittivity ϵ, and at an operating frequency ω, the field of a forward-propagating plane wave is

$$E_x(z)=E_0e^{-j\beta z}, \quad H_y(z)=\frac{E_0}{\eta}e^{-j\beta z}, \tag{47}$$

where $\beta=k=\omega(\mu\epsilon)^{1/2}$ is the propagation constant, and $\eta=(\mu/\epsilon)^{1/2}$ is the characteristic impedance of the material. Now since the only electric field component here is in the x direction, the boundary condition of vanishing tangential electric field is automatically satisfied. In other words, the insertion of the two walls does not physically perturb the propagation of a plane wave propagating in the z direction. We sometimes refer to this type of solution as the TEM mode since both E_x and H_y are perpendicular or transverse to the direction of propagation.

The reason why the TEM mode is an acceptable solution can also be explained by the fact that the direction of propagation, often defined as a ray in geometrical theories of electromagnetic and optical wave propagation, is parallel to the walls, and thus will bounce neither toward nor away from them. This of course will change if the rays are propagating in some other direction. For instance, let us consider the case when a plane wave of amplitude A_1 is perpendicularly polarized, incident onto the lower wall at an angle θ. According to the analysis given in Sec. 4.1 for the case when $\sigma_2 \to \infty$ and $\eta_2 = 0$, a reflected wave will be generated with an amplitude equal but opposite to that of the incident wave so that the total field in the region is given as

$$E_y(x,z)=A_1\, e^{-jk(-x\cos\theta+z\sin\theta)} + A_2\, e^{-jk(x\cos\theta+z\sin\theta)}, \tag{48}$$

where $A_2=-A_1$. The wave or ray that bounces from the lower wall will become an incident wave onto the upper wall at $x=a$ at the same angle θ, which in turn, generates a reflected wave in the same direction as the original incident wave. A constructive interference of all those subsequent waves that are propagating up and of all those propagating down can occur only if the two types of waves also satisfy the boundary condition at $x=a$ simultaneously:

$$A_2=-A_1, \quad A_1\, e^{jka\cos\theta}=-A_2\, e^{-jka\cos\theta},$$

which implies

$$\theta_m=\cos^{-1}\left(\frac{m\pi}{ka}\right), \quad m=1,2,3,\ldots. \tag{49}$$

In other words, only those waves that are propagating at an angle $\theta=\theta_m$, $m=1,2,3\ldots$, can form acceptable solutions in a parallel-plate waveguide. The substitution of Eq. (49) into Eq. (48) yields the following expression:

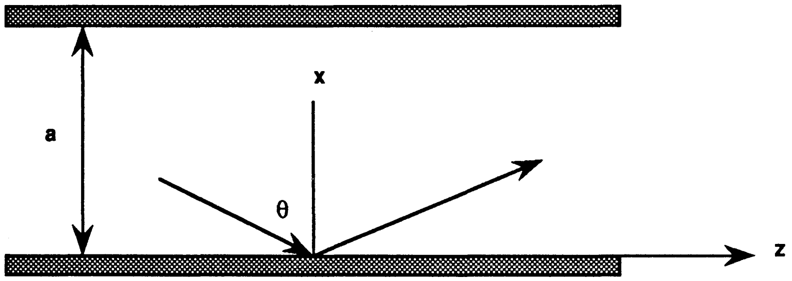

FIG. 8. A ray description for a TE mode in a parallel waveguide.

$$E_y(x,z) = \sum_{m=1}^{\infty} E_{ym} \sin\left(\frac{m\pi}{a} x\right) e^{-\gamma_m z},$$

$$\gamma_m = j\beta_m = j\left[k^2 - \left(\frac{m\pi}{a}\right)^2\right]^{1/2}. \tag{50a}$$

In deriving Eq. (50a), we have replaced the term $-j2A_1$ by E_{ym} to stress the point that the amplitude of each solution is yet undetermined because we do not know the nature of excitation. Moreover, since the electric field component in this case is always transverse to the direction of propagation, we designate this type of solution as TE$_m$ modes. It is not difficult to show that solutions corresponding to parallel-polarized plane waves are characterized by

$$H_y(x,z) = \sum_{m=0}^{\infty} H_{ym} \cos\left(\frac{m\pi}{a} x\right) e^{-\gamma_m z},$$

$$\gamma_m = j\beta_m = j\left[k^2 - \left(\frac{m\pi}{a}\right)^2\right]^{1/2}, \tag{50b}$$

and are designated as the TM$_m$ modes. The special case of $m = 0$ clearly corresponds to the TEM mode referred to earlier. While not written out explicitly here, we should nevertheless also note that the magnetic field of a TE$_m$ mode can be obtained directly from the Maxwell equation: $\mathbf{H} = (-j\omega\mu)^{-1}\nabla E_y \times \mathbf{e}_y$, and likewise for the electric field of a TM$_m$ mode.

Several important aspects of TE and TM modes are the following

1. Cutoff frequency: As suggested by Eqs. (50a) and (50b), each mode in a parallel-plate waveguide propagates with a different complex propagation constant γ_m. For those modes where $k > (m\pi/a)$, γ_m is purely imaginary, i.e., $\gamma_m = j\beta_m$; the wave behaves like $\cos(\omega t - \beta_m z)$ and thus propagates down the waveguide without decay. However, for those modes where $k < (m\pi/a)$, γ_m is purely real, i.e., $\gamma_m = \alpha_m$; the wave behaves like $\exp(-\alpha_m z)\cos \omega t$, and thus decays exponentially in the z direction like an inhomogeneous plane wave. From the point of view of signal transmission, it is not difficult to show from Poynting's theorem that no time-average power can be transmitted from one cross section z=constant to another in the case of a single exponentially decaying wave. Thus, for any operating frequency ω, there are only a finite number of modes that can propagate. The frequency below which a mode ceases to propagate is called the cutoff frequency and is given by $\omega_{cm} = m\pi/a\sqrt{\mu\epsilon}$.
2. Fast waves: A propagating mode in a parallel-plate waveguide is often called a fast-wave mode. The reason for this designation comes from the observation that the phase velocity along the direction of propagation is given by $v_{pz} = \omega/\beta_m$, which is always greater than the speed of light $c = (\mu\epsilon)^{-1/2}$, since $\beta_m < k$ and $k = \omega/c$. This does not mean that any known physical principle has been violated. Rather it is because the phase velocity defined in this manner overlooks the fact that the underlying plane waves are all propagating in directions other than along the z axis. Nevertheless, such an observation does underline an important physical realization, and that is: the phase velocity associated with either a mode or a single plane wave is only a measure of how fast a phase front is moving as observed along a certain direction. For instance, the phase velocity would have been infinity had we chosen to observe along a constant phase front, no matter how fast the wave were actually moving. In fact, since the component plane waves travel down the waveguide in a zig-zag manner, it can be shown that the signal velocity (as defined by the so-called group velocity $v_g = \partial\omega/\partial\beta_m$) is slower rather than faster than the speed of light for all propagating modes.
3. Characteristic impedance of a mode: We have seen that it is often convenient to introduce the concept of impedance to characterize the relationship between the electric field and the magnetic field of a plane wave; the value of this impedance depends only upon the material properties, but not upon where the field is observed. This is also true in the case of TEM, TE, and TM modes, if we define the characteristic impedance to be the ratio of the transverse electric field $\mathbf{E}_t$ to the transverse magnetic field $\mathbf{H}_t$:

$$\mathbf{H}_t = \frac{1}{\eta_m}(\mathbf{e}_z \times \mathbf{E}_t), \tag{51a}$$

where

$$\eta_m = j\omega\mu/\gamma_m, \quad (\text{TE}_m \text{ modes}),$$

$$\eta_m = \gamma_m/j\omega\epsilon \quad (\text{TM}_m \text{ modes}). \qquad (51\text{b})$$

In contrast to the case of a plane wave, however, the characteristic impedance of a mode is purely real if the mode is above cutoff but is purely imaginary below cutoff.

4. Dispersion: The parallel-plate waveguide, in either a TE or TM mode of operation, is a dispersive structure. This is because the propagation constant β_m as given by Eq. (49), and consequently the phase velocity v_{pz}, exhibit a strong dependence on frequency. A signal which is rich in frequency content will disperse after traveling a sufficient distance along the waveguide because the phase relation between the various spectral components is no longer the same. Consider, for instance, the electric field of a TE_1 mode which is amplitude-modulated so that initially, at the $z=0$ plane, we have

$$8qE_y(x,0;t) = E_0\{1+\Delta_m \cos \omega_m t\} \times \sin(\pi x/a) \cos \omega_0 t,$$

where Δ_m is the amplitude of modulation, and ω_m and ω_0 are the modulation and carrier frequencies, respectively. Its envelope will eventually evolve into a different shape because its carrier frequency ω_0 and the two sideband frequencies $\omega^\pm = \omega_0 \pm \omega_m$ will have different phase shifts:

$$E_y(x,z;t) = E_0\left\{\cos(\omega_0 t - \beta_0 z) + \frac{\Delta_m}{2}[\cos(\omega^+ t - \beta^+ z) + \cos(\omega^- t - \beta^- z)]\right\} \sin\left(\frac{\pi x}{a}\right).$$

For many applications, the sidebands are very narrow, i.e., $\omega_m \ll \omega_0$. A good approximation for $\beta^\pm$ can be obtained by keeping only the first two terms in the Taylor series: $\beta^\pm z \simeq (\beta_0 \pm \omega_m/v_g)z$, where $v_g = \partial\omega/\partial\beta$, evaluated at $\omega = \omega_0$. We can then further simplify the expression for E_y to obtain

$$E_y(x,z;t) = E_0\left\{1+\Delta_m \cos \omega\left(t - \frac{z}{v_g}\right)\right\} \times \sin\left(\frac{\pi x}{a}\right) \cos \omega\left(t - \frac{z}{v_{pz}}\right),$$

$$v_{pz} = \frac{\omega_0}{\beta_0}, \qquad (52)$$

which clearly shows that the carrier is propagating at the phase velocity v_{pz}, while the envelope or modulation is moving at the group velocity v_g which according to Eq. (49) equals c^2/v_{pz} (see Fig. 9). Thus, so long as the bandwidth of a signal is narrow, we see that the basic characteristics of the signal are retained, at least initially, before it begins to be dispersed. The dispersion distance, in the case of amplitude modulation, can be shown to be proportional to $(\omega_0/\omega_m)^2$.

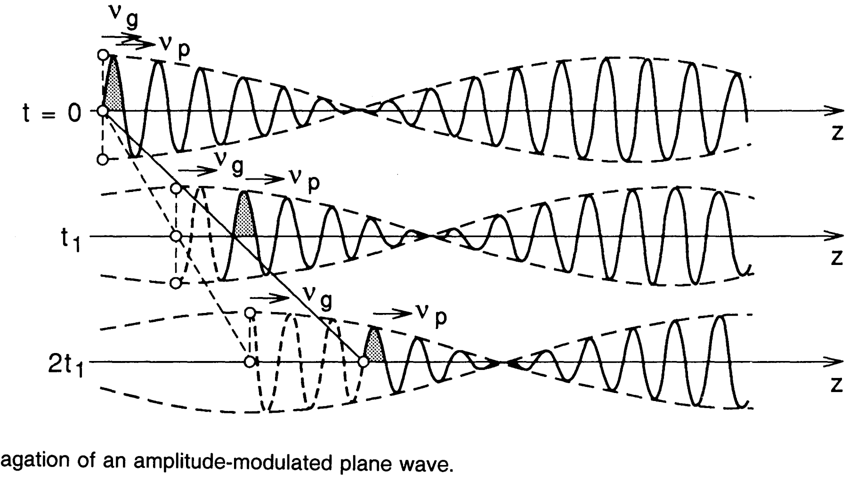

FIG. 9. Propagation of an amplitude-modulated plane wave.

5.2 Rectangular Waveguides

A rectangular waveguide differs physically from a parallel-plate waveguide in that the cross section is now finite in both the x and y directions as shown in Fig. 10. Recalling that a parallel-plate waveguide is capable of propagating a TEM mode without cutoff basically because the insertion of the two walls does not perturb the field distribution of that mode, we see that this is no longer true in the case of a rectangular waveguide. However, the ray approach we used is still applicable. For instance, for a TM mode characterized by the longitudinal component of the electric field E_z, we can first assume a set of forward-propagating plane waves moving at an angle θ with respect to the z axis, which bounce around between the four surrounding walls at $x=0,a$ and $0<y<b$, and at $y=0,b$ and $0<x<a$.

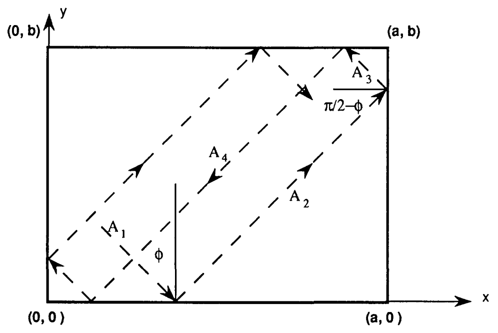

FIG. 10. Rays in a rectangular waveguide.

We further assume that the projection of the incident angle of one wave (one impinging on the wall at $y=0$, say) in the xy plane is ϕ, so that the sum of all possible waves resulting from repeated reflections of this original wave is

$$E_z(x,y,z)=\{A_1\exp[-jk\sin\theta(x\sin\phi-y\cos\phi)]+A_2\exp[-jk\sin\theta(x\sin\phi+y\cos\phi)]$$
$$+A_3\exp[+jk\sin\theta(x\sin\phi-y\cos\phi)]+A_4\exp[+jk\sin\theta(x\sin\phi+y\cos\phi)]\}$$
$$\times e^{-jkz\cos\theta}. \tag{53}$$

Again, for these waves to interact constructively with each other, they must be oriented in such a way to produce a vanishing tangential field at all four walls. For example, as shown in Fig. 10, for the wall at $y=0$, ray No. 1 and ray No. 4 are incident waves, while ray No. 2 and ray No. 3 are reflected rays, so that $A_2=-A_1$ and $A_4=-A_3$. Proceeding with the conditions at the remaining walls, the following relationships are obtained:

$$A_1=-A_2=A_3=-A_4;$$
$$e^{-jka\sin\theta\sin\phi}=e^{+jka\sin\theta\sin\phi}, \tag{54}$$
$$e^{-jkb\sin\theta\cos\phi}=e^{+jkb\sin\theta\cos\phi},$$

which can then be translated into a set of solutions for θ and ϕ:

$$\theta_{mn}=\sin^{-1}\left[\left(\frac{m\pi}{a}\right)^2+\left(\frac{n\pi}{b}\right)^2\right]^{1/2};$$
$$\phi_{mn}=\sin^{-1}\left(\frac{m\pi}{a\sin\theta_{mn}}\right), \tag{55}$$

where $m,n=1,2,3\ldots$. Using these values of θ and ϕ, we can rearrange Eq. (53) into a more familiar form:

$$E_z(x,y,z)=\sum_{m,n}E_{mn}\sin\left(\frac{m\pi}{a}\right)\sin\left(\frac{n\pi}{b}\right)e^{-\gamma_{mn}z}, \tag{56a}$$

where

$$\gamma_{mn}=j\beta_{mn}=j\left\{k^2-\left(\frac{m\pi}{a}\right)^2-\left(\frac{n\pi}{b}\right)^2\right\}^{1/2}; \tag{56b}$$

and each such solution is called a TM$_{mn}$ mode. The amplitude of each mode, i.e., E_{mn}, is again yet to be determined from the actual excitation of the waveguide. The derivation for TE$_{mn}$ modes follows in essentially the same manner, except that either m or n, but not both, can be zero:

$$H_z(x,y,z)=\sum_{m,n}H_{mn}\cos\left(\frac{m\pi}{a}\right)$$
$$\times\cos\left(\frac{n\pi}{b}\right)e^{-\gamma_{mn}z}, \tag{57a}$$

where

$$\gamma_{mn}=j\beta_{mn}=j\left\{k^2-\left(\frac{m\pi}{a}\right)^2-\left(\frac{n\pi}{b}\right)^2\right\}^{1/2}. \tag{57b}$$

As in the case of a parallel-plate waveguide, we can again conclude that only a finite number of propagating modes exist for which $\gamma_{mn}=j\beta_{mn}$ is purely imaginary. For most applications, the operating frequency is set to be above the cutoff frequency of a TE_{10} mode, but below the cutoff frequencies of all the other modes, so that a single mode of operation is assured. To complete our discussion on waveguide modes, we should also point out that both the transverse electric and transverse magnetic field components can be readily obtained once the longitudinal components are known. This is because the two Maxwell curl equations decouple when a functional dependence of $\exp(-\gamma z)$ is assumed:

$$-\gamma\mathbf{e}_z\times\mathbf{E}_t+\nabla_tE_z\times\mathbf{e}_z=-j\omega\mu\mathbf{H}_t,$$

$$-\gamma\mathbf{e}_z\times\mathbf{H}_t+\nabla_tH_z\times\mathbf{e}_z=j\omega\epsilon\mathbf{E}_t,$$

which after some manipulation yield the following result for a TE mode:

$$\mathbf{E}_t=\left(\frac{-\gamma}{k^2+\gamma^2}\right)\nabla_tE_z,\quad \mathbf{H}_t=\left(\frac{1}{\eta_{TE}}\right)\mathbf{e}_z\times\mathbf{E}_t,$$

where $\eta_{TE}=\gamma/(j\omega\epsilon)$ is the characteristic impedance of a TE mode. A similar expression also exists for the TM mode. Without further elaboration, it suffices for us to state that these modes share the fast-wave, dispersive properties of the TE and TM modes in a parallel-plate waveguide.

6. GUIDED WAVES (II)

6.1 Dielectric Waveguides: Surface-Wave Modes

The waveguides we have discussed so far are closed waveguides, i.e., the electromagnetic field exists only inside the metallic waveguide structure. The dimensions of such waveguides are made comparable to the wavelength of a plane wave if they are to be operated in the single-mode regime. This makes fabrication difficult when the operating frequency is in the millimeter-wave or optical range, and prompts the question of whether electromagnetic waves can be guided in single-mode fashion on a structure of larger dimension, possibly in an open environment without a metallic wall. Consider a dielectric slab in air, of thickness $2d$ and refractive index n, as shown in Fig. 11, which is an analog of the parallel-plate waveguide discussed in Sec. 5.1. We again start with a perpendicularly polarized plane wave inside the dielectric region and incident onto the air-dielectric interface at $z=d$ at an angle θ . This plane wave of course will normally be partially transmitted and partially reflected, unless the incident angle is greater than the critical angle, in which case the wave will be totally reflected from the boundary. Assuming that to be the case, we can write down the general expression for the transverse electric field as

$$E_y(x,z)=A_1\,e^{-jk_0n[(x-d)\cos\theta+z\sin\theta]}+A_2\,e^{-jk_0n[(d-x)\cos\theta+z\sin\theta]}, \tag{58}$$

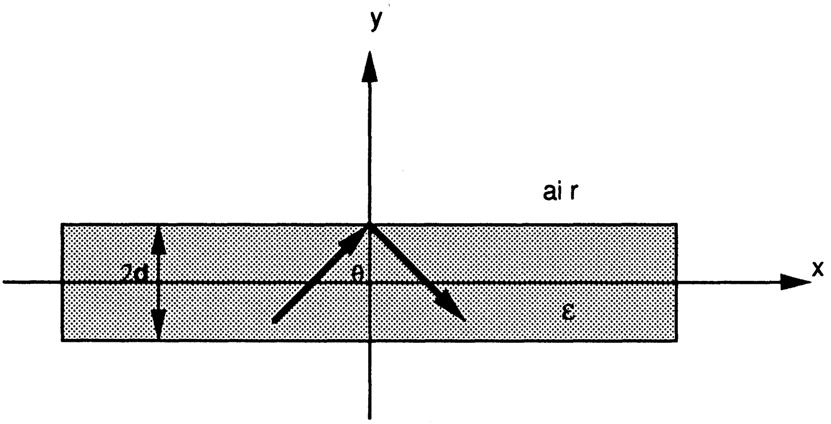

FIG. 11. Geometry of a dielectric slab waveguide.

where

$$A_2=A_1\, e^{-j2\Psi},$$

$$\Psi(\theta)=\tan^{-1}\left\{\frac{[n^2\sin^2\theta-1]^{1/2}}{n\cos\theta}\right\}.$$

Now since the structure is symmetric in the x direction, we can divide the solutions of this problem into even and odd modes. For the even modes, we require that $A_1\, e^{jk_0 n\cos\theta} = A_2\, e^{-jk_0 n\cos\theta}$ so that the field distribution will be symmetric. Constructive interference of waves inside the dielectric region occurs when

$$nk_0 d\cos\theta=m\pi-\Psi(\theta)\Rightarrow$$

$$\tan(nk_0 d\cos\theta)=\frac{[n^2\sin^2\theta-1]^{1/2}}{n\cos\theta} \qquad (59a)$$

which can be used to solve for θ . In order to discuss the nature of the solution more fully, it is convenient to define two new parameters: $h(\theta)=nk_0\cos\theta$, and $p(\theta)=k_0(n^2\sin^2\theta-1)^{1/2}$. As seen from Eq. (58), $h(\theta)$ is a transverse wave number telling how fast the field varies in the x direction inside the dielectric slab. On the other hand, $p(\theta)$ is a measure of the rate of field decay with x in the air. The equation to solve for θ in Eq. (59a) now becomes

$$h\tan hd=p, \quad (hd)^2+(pd)^2=V^2, \qquad (59b)$$

where $V=k_0(n^2-1)^{1/2}$ is often referred to as the normalized frequency of the dielectric slab waveguide. Equation (59b) suggests that a solution occurs whenever the curve $pd=hd\tan hd$ intersects a circle whose radius is V centered at the origin of the hd–pd plane. As shown in Fig. 12, only a finite number of solutions can exist for a given value of V, and each solution corresponds to a TE mode of operation. Using the definitions of h and p, it is not difficult to show that the transverse electric field in this case can be rewritten in a more familiar form:

$$E_y(x,z)=\sum_m E_m\Phi_m(x)e^{-j\beta_m z};$$

$$\beta_m=[(nk_0)^2-h_m^2]^{1/2}=[p_m^2+k_0^2]^{1/2}, \qquad (60)$$

and

$$\Phi_m(x)=\cos h_m x/\cos h_m d \text{ for } 0<x<d$$

and

$$\Phi_m(x)=e^{-p_m(x-d)} \text{ for } x>d.$$

We shall now discuss some of the most important features of the modes supported by a dielectric slab.

1. Surface-wave modes: The guided-wave modes that propagate along a dielectric slab are often referred to as surface-wave modes, and the dielectric slab is also known as a slow-wave structure. The reason for

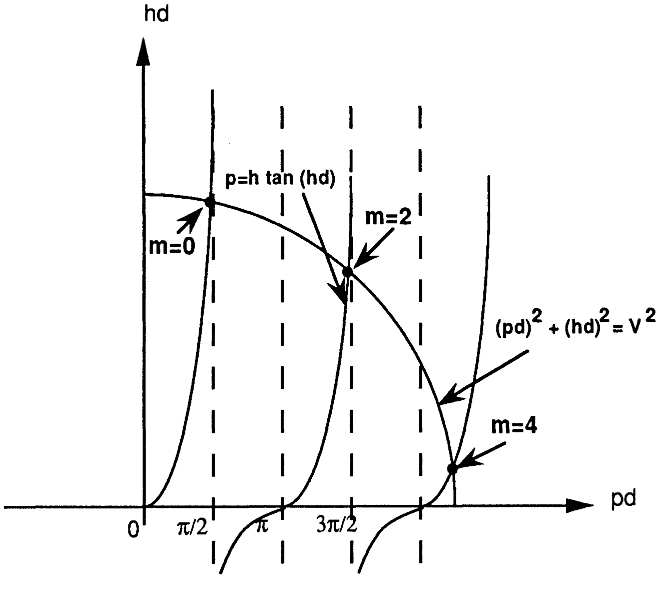

FIG. 12. Surface wave solution(s) for a dielectric waveguide.

the term surface wave comes from the recognition that the field decays exponentially as we move away from the surface outside of the dielectric slab. Thus, it has the appearance of existing only in the vicinity of that surface. For the same reason, the phase velocity of the *m*th surface-wave mode is given by the expression $v_{pz} = \omega/\beta_m = c/[1 + (cp/\omega)^2]^{1/2}$, where c is the speed of light measured in air. Thus, it is a slow wave compared to an ordinary plane wave in air. It is important to realize, however, that even though such a designation is semantic in nature, it does underline a very important difference between guided waves supported by a metallic structure and those supported by a dielectric structure. In a metallic waveguide, waves are bouncing back completely from its walls with the phase of a reflected wave differing by either 0 or π from that of the incident wave. In a dielectric waveguide, waves are totally reflecting because of the critical-angle phenomenon, and the phase of the reflected wave varies according to its incident angle. As a result, the field distribution of a surface-wave mode depends in a complicated manner on the operating frequency via the expression for the normalized frequency V. This of course is in contrast to the modes in a metallic waveguide, whose transverse field structure is independent of frequency.

2. Cutoff frequency: The number of surface-wave modes that can be supported by an open dielectric waveguide depends upon the value of the normalized frequency, $V = (n^2 - 1)^{1/2}k_0d$. The frequency for which a mode ceases to exist is called the cutoff frequency. As seen in Fig. 12, the TE_0 mode clearly does not have a cutoff, even though as V decreases, the value of pd also decreases and thus the mode is less confined in the air region. The second mode, on the other hand, no longer exists when $V \leqslant \pi$, giving a cutoff frequency of $\omega_{c,2} = \pi c/(n^2 - 1)^{1/2}d$. This means that the operating frequency must be lower than $\omega_{c,2}$ if single-mode operation is desired. It is of interest to note that, since the value of V depends on the contrast in refractive index between the dielectric and the air, we often can maintain single-mode operation even when the dimension of the dielectric slab is much greater than the wavelength of a plane wave in either region.

The above discussion points to an important difference between the discrete modes in a metallic waveguide and those of a dielectric waveguide regarding the concept of cutoff. In a metallic waveguide, higher-order modes can still exist even though they are under cutoff: they merely decay exponentially along the waveguide. In a dielectric waveguide, a mode under cutoff simply cannot exist physically. The general solution for a dielectric waveguide turns out to contain, instead of discrete cutoff modes, a continuous spectrum of modes. These modes are commonly referred to as radiation modes since they form a spectrum of plane waves in air and thus are not confined to the vicinity of the guiding structure. To complete our discussion on the dielectric-slab waveguide, we should point out that the structure is also dispersive. The analysis of this, however, is basically similar to that for the metallic waveguide and will not be repeated here.

6.2 Dielectric-Rib Waveguide and Other Surface-Wave Structures

Figure 13 shows the geometry of the so-called dielectric-rib waveguide commonly encountered in integrated optical and millimeter-wave circuits. The analysis of the modes for this structure is substantially more complicated than for the slab structure; however, we can once again follow the basic concept of total reflection to gain some insight as to how surface waves can be supported. This can be done by first dividing the cross section into three separate regions consisting of dielectric slabs of different thicknesses. Assuming that only one surface-wave mode of TE type can exist in each of the separate slabs, we can write down the general expression for a surface wave in any of the slab regions as

$$\mathbf{E}(x,y,z) = \mathbf{a}_e\{A_1\, e^{-j\beta(x\cos\theta + z\sin\theta)} + A_2\, e^{-j\beta(-x\cos\theta + z\sin\theta)}\}\Phi(y), \tag{61}$$

where $\Phi(y)$ is the modal electric field distribution of the surface wave and β is its propagation constant as determined from $\beta = (k_0^2 + p^2)^{1/2}$; $p > 0$ in Eqs. (59) and (60). The surface wave in this case is assumed to be

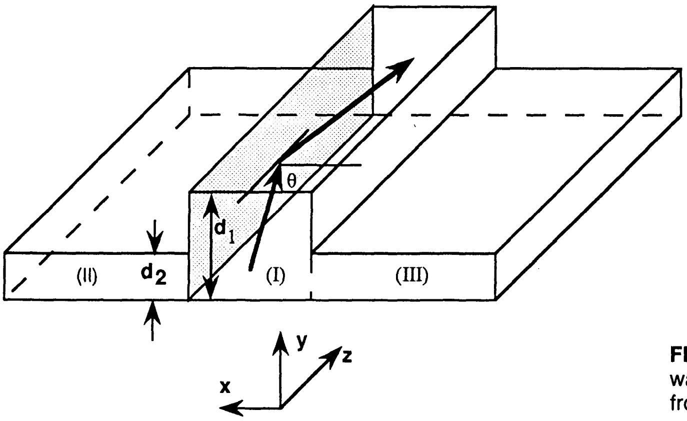

FIG. 13. A ray description of a surface-wave mode in the core region reflected from the interface on the left.

propagating in the lateral direction (xz plane) at an angle θ with respect to the x axis. The unit vector $\mathbf{a}_e$ is perpendicular to both the direction of propagation and the y axis. The situation is not unlike that of an obliquely incident plane wave, particularly if we choose to observe the field inside the dielectric region. The requirement that the phase variations in the z direction of waves on either side of the rib interface should match each other again results in Snell's law, $\beta_1 \sin \theta_1 = \beta_2 \sin \theta_2 = \beta_3 \sin \theta_3$ for the three slab regions. Now since we know that the value of p increases as the normalized frequency $V = k_0 d(n^2 - 1)^{1/2}$ increases, we can readily conclude that $\beta_1 > \beta_{2,3}$. Thus, for some incident angle θ_1, it is indeed possible to produce the condition $\beta_{2,3} \cos \theta_{2,3} = j[(\beta_1 \sin \theta_1)^2 - \beta_{2,3}^2]^{1/2}$ so that the waves in regions No. 2 and No. 3 will decay away from the rib region in the x direction as well. What this means is that at some appropriate angle, a total reflection indeed can occur, when the surface waves in the core region, which are already confined in the y direction, impinge onto the interface separating it from the surrounding slabs. The governing equation for finding the angles corresponding to the surface-wave modes supported by the rib waveguide can be derived approximately if we use the same expression for the reflection coefficient derived for plane waves.

The procedure described above works well for a commonly encountered class of planar dielectric waveguide structures. Some of those are shown in Fig. 14. A more detailed discussion is beyond the scope of this review.

7. MULTICONDUCTOR TRANSMISSION LINES

7.1 TEM-Mode Transmission Lines

The metallic waveguide modes we have discussed so far are of either TE or TM type with specific cutoff frequencies depending upon the dimensions of the waveguide and the properties of the material in the waveguiding region. The only exception to this was the TEM mode propagating in a parallel-plate waveguide, which does not have a cutoff. It turns out that the existence of one or more TEM modes is in fact a general property of multiconductor systems immersed in a homogeneous medium. Other than having the ability to propagate at all frequencies, these modes also have the interesting property that their fields behave like static fields in any given transverse cross section, while exhibiting a wave behavior along the direction of propagation. In fact, we often call these modes transmission-line modes, because many of the physical phenomena associated with these modes can be derived from an equivalent circuit representation consisting of distributed series inductances and shunt capacitances. Nevertheless, the conductors are merely the boundary of a waveguide in which power transmission is actually taking place.

Assuming TEM fields behaving like $\exp(-\gamma z)$, where γ is a yet undetermined TEM-mode propagation constant, the two curl equations become $\gamma \mathbf{e}_z \times \mathbf{E}_t = -j\omega\mu \mathbf{H}_t$ and $-\gamma \mathbf{e}_z \times \mathbf{H}_t = j\omega\epsilon \mathbf{E}_t$. Elimination of either $\mathbf{E}_t$ or $\mathbf{H}_t$ from these equations yields

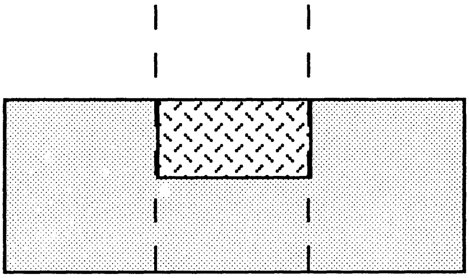

dielectric channel waveguide

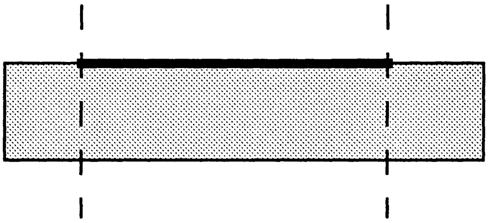

microguide

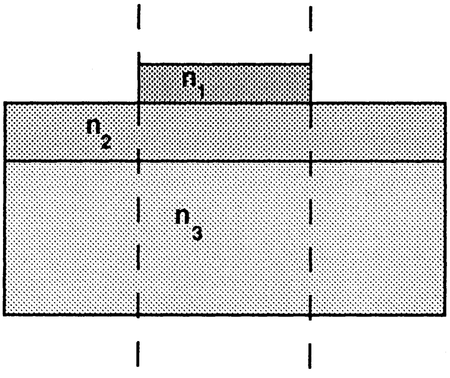

dielectric strip guide

FIG. 14. Some planar dielectric waveguides.

$$\mathbf{H}_t=\left(\frac{1}{\eta}\right)\mathbf{e}_z\times\mathbf{E}_t,\quad \gamma=j\beta=j\omega(\mu\epsilon)^{1/2},\ \text{and}$$
$$\eta=\left(\frac{\mu}{\epsilon}\right)^{1/2}. \tag{62}$$

Thus, a TEM mode, if one exists, will behave exactly like a plane wave, except that both $\mathbf{E}_t$ and $\mathbf{H}_t$ may have some yet-unspecified transverse distribution.

The reason why a TEM does not have a nonzero cutoff frequency can be seen directly from the Maxwell equations. To find out whether the TEM mode can or cannot exist for a given structure, we now take the z component of $\nabla\times\mathbf{E}=-j\omega\mu\mathbf{H}$ to obtain $\mathbf{e}_z\cdot\nabla_t\times\mathbf{E}_t=0$. Thus, the transverse electric field in a given cross section can always be obtained from the gradient of a potential function, i.e., $\mathbf{E}_t=-\nabla_t\Phi$, which (as far as x and y dependences are concerned) is a static field. Now since we know that the static problem has exactly $M-1$ independent solutions for Φ corresponding to various voltage differences applied among the conductors in an M-conductor system, we can conclude that the same is true in the case of TEM modes supported by the same structure. This is why a coaxial cable or a two-wire transmission is so often used to transmit electromagnetic waves at lower frequencies.

Consider the example, shown in Fig. 15, of two infinitely long conductors of arbitrary cross section. The circuit interpretation can be demonstrated if we use the voltage $V(z)$ across this pair of conductors, and the total current $I(z)$ flowing on one of the two conductors, as a characterization of a transmission-line mode. For a forward-propagating TEM wave, the voltage is given by

$$V(z)=\left\{\int_P \mathbf{E}_t\cdot d\mathbf{l}\right\}e^{-j\beta z},$$

where P is a path in a plane of constant z between the two conductors. Using the relationship between $\mathbf{E}_t$ and $\mathbf{H}_t$ in Eq. (62), we can show that the voltage between the two conductors is related to the total magnetic flux per unit length Ψ_m cutting across the integration path P according to $V=v_p\Psi_m$, where $v_p=(\mu\epsilon)^{-1/2}$ is the phase velocity. As in a magnetostatic problem, we can define a distributed series inductance L_e per unit length as the ratio of Ψ_m to the current I so that the voltage and the current are now related to each other by $V/I=v_pL_e$. If, on the other hand, we start with the current as equal to the integration of the tangential magnetic field around a conductor, we can show again from Eq. (62) that $I=v_pQ$, where Q is the total charge per unit length on the conductor. As in an electrostatic problem, we can define a distributed shunt capacitance C_e per unit length as the ratio of Q to the voltage V between the two conductors so that $I/V=v_pC_e$. Combining the two equations relating V and I finally gives

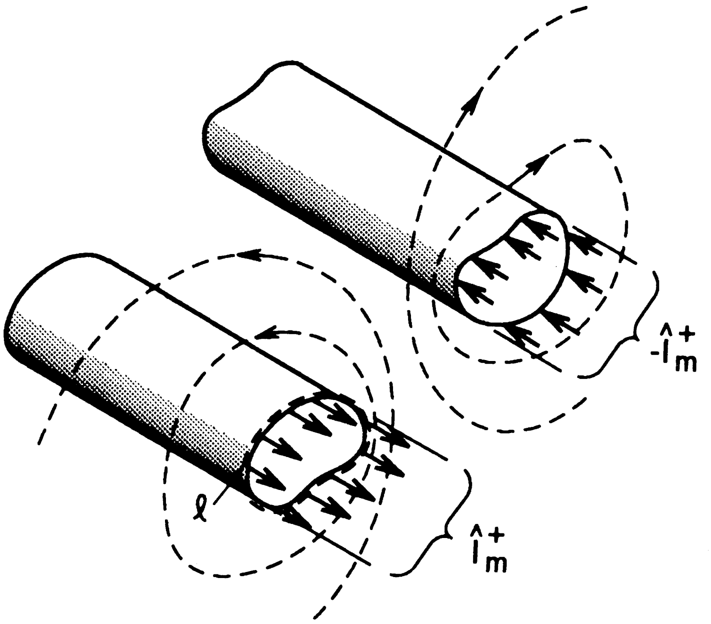

FIG. 15. A two-wire transmission line.

us the well-known result that $L_e C_e = \mu\epsilon$ for a TEM mode. To complete the circuit analogy, we now define the characteristic impedance Z_c of a transmission-line mode as the ratio of the voltage to current for a forward propagating wave, $Z_c = [L_e/C_e]^{1/2}$. The voltage and current waves in the presence of both forward and backward waves can then be cast in the most general form,

$$\begin{aligned} V(z) &= V_m^+ e^{-j\beta z} + V_m^- e^{+j\beta z}, \\ I(z) &= \frac{1}{Z_c}\{V_m^+ e^{-j\beta z} - V_m^- e^{+j\beta z}\}, \end{aligned} \tag{63}$$

where $V_m^\pm$ are the amplitudes of the forward- and backward-traveling waves. In most applications, the design of a transmission-line circuit can be carried out on the basis of the voltage and current waves, without referring back to the underlying electric and magnetic field distributions. It is also of interest to note that the governing equations (63) resemble those of a plane wave. Thus it is not surprising that many of the techniques we used to analyze plane-wave propagation in a multilayer problem are equally applicable here.

7.2 Quasi-TEM–Mode Transmission Lines

Microstrip line is one of the most common forms of transmission line in microwave integrated circuits because of its broadband nature, and because it is easy to fabricate with

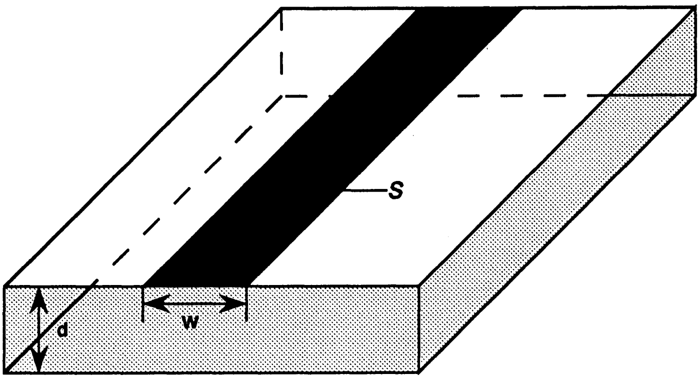

FIG. 16. A microstrip transmission-line structure.

modern integrated- or printed-circuit technology (Hoffmann, 1987). The cross section of such a line is shown in Fig. 16. Basically it consists of a thin metal strip of finite width on a dielectric substrate on a ground plane. The transmission line formed by the strip and its image is of the same type as the two-wire transmission line, and thus would support a TEM mode if the dielectric substrate were replaced by air. However, in the presence of a substrate, a pure TEM mode can no longer exist: It becomes impossible to satisfy the boundary conditions at the air-dielectric interface. However, if both the strip width and substrate thickness are very small compared with a wavelength, a quasi-TEM mode will still exist, whose E_z and H_z are small compared to their transverse counterparts, which are found (at least approximately) from quasistatic solutions in the cross-sectional plane in a manner similar to that for TEM modes.

As for the case of TEM modes, the transverse electric field (which is a static field in x and y) of a quasi-TEM mode can be associated with a capacitance C_e per unit length, and the transverse magnetic field with an inductance L_e per unit length. As these two static fields are completely decoupled, the capacitance depends on the spatial distribution of ϵ, but not of μ; and conversely for the inductance, which is affected by μ and not ϵ. The importance of this observation is that relations (62) do not hold, and therefore $L_eC_e \neq \mu\epsilon$, if $\mu\epsilon$ is not independent of x and y. Nevertheless, within the quasi-TEM approximation, we still have

$$\gamma = j\beta = j\omega(L_eC_e)^{1/2} \text{ and } Z_c = [L_e/C_e]^{1/2}, \quad (64)$$

and thus such transmission lines can still be treated by much the same methods as are used for true TEM lines.

In addition to microstrip, other quasi-TEM structures such as suspended microstrip, slotline, coplanar waveguide, finline, etc., also submit to the same treatment. Figure 17 shows the geometry of some of these structures. Each has its own practical advantages and disadvantages which make it suitable for certain types of application.

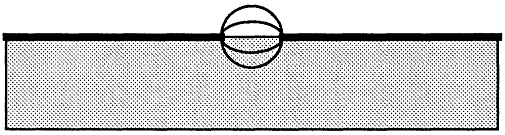

slot line

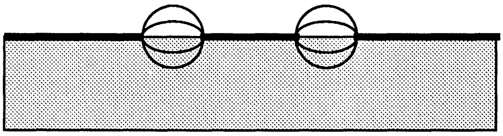

coplanar waveguide

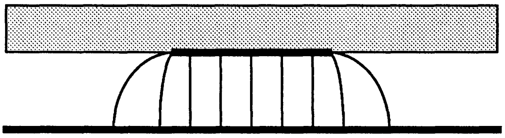

suspended microstrip

FIG. 17. Some quasi-TEM structures.

GLOSSARY

Brewster Angle: Angle of incidence at which a perpendicularly polarized electromagnetic plane wave will be transmitted through a plane dielectric interface without reflection.

Chiral Medium: A medium in which **D** is proportional to a linear combination of **E** and $\nabla \times \mathbf{E}$, and similarly for **B** and **H**.

Critical Angle: Angle of incidence at which an electromagnetic plane wave will be reflected at a plane dielectric interface without any transmission of power through the interface.

Dispersion: Distortion of a pulsed wave due to a dependence of wave speed on frequency.

Group Velocity: Velocity of the envelope of a modulated wave.

Phase Velocity: Velocity of the phase of the carrier of a modulated plane wave.

Skin Depth: Depth at which a plane wave has decayed in amplitude to e^{-1} of its original value in a lossy medium.

Surface Wave: A wave which is confined close to a surface or surfaces of a waveguiding structure.

Wavelength: Distance between points of a traveling wave whose phases differ by 2π.

List of Works Cited

Arnaud, J. A. (1976), *Beam and Fiber Optics*, New York: Academic.

Hoffmann, R. K. (1987), *Handbook of Microwave Integrated Circuits*, Norwood, MA: Artech House.

Ishimaru, A. (1991), *Electromagnetic Wave Propagation, Radiation, and Scattering*, Englewood Cliffs, NJ: Prentice-Hall.

Jackson, J. D. (1975), *Classical Electrodynamics*, New York: Wiley.

Johnk, C. T. A. (1988), *Engineering Electromagnetic Fields and Waves*, New York: Wiley.

Kong, J. A. (1986), *Electromagnetic Wave Theory*, New York: Wiley.

Kraus, J. D. (1984), *Electromagnetics*, New York: McGraw-Hill.

Lakhtakia, A., Varadan, V. K., Varadan, V. V. (1989), *Time-Harmonic Electromagnetic Fields in Chiral Media*, Lecture Notes in Physics No. 335, Berlin: Springer.

Nussenzveig, H. M. (1972), *Causality and Dispersion Relations*, New York: Academic.

Pozar, D. M. (1990), *Microwave Engineering*, Reading, MA: Addison-Wesley.

Ramo, S., Whinnery, J. R., van Duzer, T. (1984), *Fields and Waves in Communication Electronics*, New York: Wiley.

Solimeno, S., Crosignani, B., DiPorto, P. (1986), *Guiding, Diffraction and Confinement of Optical Radiation*, Orlando, FL: Academic.

Stratton, J. A. (1941), *Electromagnetic Theory*, New York: McGraw-Hill.

Yariv, A., Yeh, P. (1984), *Optical Waves in Crystals*, New York: Wiley.

Further Reading

Born, M., Wolf, E. (1975), *Principles of Optics*, Oxford: Pergamon.

Collin, R. E. (1991), *Field Theory of Guided Waves*, New York: IEEE.

Felsen, L., Marcuvitz, N. (1973), *Radiation and Scattering of Waves*, Englewood Cliffs, NJ: Prentice-Hall.

Harrington, R. F. (1961), *Time-Harmonic Electromagnetic Fields*, New York: McGraw-Hill.

Jones, D. S. (1964), *The Theory of Electromagnetism*, Oxford: Pergamon.

ELECTRON LEVEL SPLITTING

SERGIO RODRIGUEZ, *Department of Physics, Purdue University, West Lafayette, Indiana, U.S.A.*

	Introduction	1
1.	**Quantum Mechanics of Complex Atoms and Ions**	2
1.1	General Considerations: The Central-Field Approximation	2
1.2	The Ground State of Free Ions: Hund's Rules	2
2.	**Magnetic Properties of Free Ions**	5
2.1	Magnetic Susceptibilities of Free Ions: Theories of Langevin and Brillouin	5
2.2	Van Vleck Paramagnetism: The Behavior of Europium and Samarium	6
3.	**The Effect of the Crystal Field on Magnetic Atoms**	7
3.1	The Crystal Field	7
3.2	Weak Crystal Field: Salts of the Rare-Earth Elements	9
3.3	The Method of Operator Equivalents	10
3.4	Crystal Field of Intermediate Strength: The Spin-Hamiltonian Formalism	11
3.5	Fe^{2+} in a Tetrahedral Environment	12
3.6	Co^{2+} in a Tetrahedral Field	14
4.	**Energy-Level Splittings in Crystals under External Stress**	15
4.1	Deformation Potentials of Band Extrema in Semiconductors	15
4.2	Relative Intensities of Stress-Induced Components of an Optical Transition	18
5.	**Concluding Remarks**	21
	Glossary	21
	Works Cited	22
	Further Reading	22

INTRODUCTION

In atomic spectroscopy the application of external perturbations such as electric and magnetic fields provides considerable insight into the properties of atoms. In the study of the properties of matter in the solid state of aggregation, spectroscopic techniques have also been crucial in the elucidation of the properties of such systems. In this area we have at our disposal, in addition to the external perturbations mentioned above, a perturbation unique to crystals, namely, the use of uniaxial stresses which alter the symmetry of the environment of the individual atoms in the solid. Furthermore, the presence of microscopic, nonuniform distributions of electric charge and magnetic moments modifies the electronic states of the atoms within the solid.

Free atoms or ions possess the complete rotation-inversion symmetry described by the three-dimensional orthogonal group O(3) (see Koster *et al.*, 1966). This symmetry is reduced by the application of external perturbations or by the electric and magnetic fields present in the solid phase. This article is devoted to a study of the excitation spectra of atoms due to the reduction of their symmetry as mentioned above. In Sec. 1 we deal with the properties of free ions. The paramagnetism of ions with incomplete shells, ordinarily present in the form of compounds, is discussed in Sec. 2. Section 3 is devoted to the effect of the crystal field on the energy-level spectra of atoms, while Sec. 4 deals with the effect of uniaxial stress on shallow impurities in semiconductors. Topics such as the Jahn-Teller effect are outside the scope of this article. Throughout we give a general discussion of the subject with a few illustrative examples. There is a large literature in this field, for the details of which the reader is referred to treatises like those by Abragam and Bleaney (1970) and by

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Zeiger and Pratt (1973) and to the original papers cited.

1. QUANTUM MECHANICS OF COMPLEX ATOMS AND IONS

1.1 General Considerations: The Central-Field Approximation

The dynamical properties of an ion containing N electrons in the field of a heavy nucleus of charge Ze are described, in a first approximation, by the nonrelativistic Hamiltonian operator

$$H=\sum_{i=1}^{N}\frac{p_i^2}{2m}-\sum_{i=1}^{N}\frac{Ze^2}{4\pi\epsilon_0 r_i}+\sum_{i<j}\frac{e^2}{4\pi\epsilon_0 r_{ij}}, \quad (1)$$

where r_{ij} is the distance between the ith and jth electrons, r_i is the distance between the ith electron and the nucleus, $-e$ and m are the electric charge and mass of the electron, and ϵ_0 is the permittivity of free space. In this approximation the nucleus is assumed so massive that it can be regarded as at rest.

We note that **L**, the total orbital angular momentum, commutes with H, the latter being invariant under the operations of the rotation-inversion group (see SYMMETRIES AND CONSERVATION LAWS). In addition, since the spin operators of the electrons are absent from H, the total intrinsic spin **S** is, like **L**, a constant of the motion. Thus, H, L^2, S^2, L_z, and S_z, where the latter two are projections of **L** and **S** on an arbitrary axis, $\hat{\mathbf{z}}$, can be simultaneously diagonalized. The energy spectrum of the Hamiltonian (1) cannot be obtained by separation of variables because of the Coulomb interaction between the electrons. We can, however, take their effect into account as a perturbation by introducing an average central potential $V(r)$ and express the Hamiltonian operator as the sum of two terms, H_0 and H_1, where H_0 includes the kinetic energy of the electrons, the Coulomb attraction of the central nucleus, and the central potential $\Sigma_{i=1}^{N}V(r_i)$ for the electrons, and H_1, the effect of the residual Coulomb repulsion between the electrons. The Schrödinger equation for H_0 can now be solved by separation of variables. The one-electron wave functions are of the form

$$\Psi_i(nlm_lm_s)=R_{nl}(r_i)Y_l^{m_l}(\theta_i,\phi_i)\chi_{m_s}^{(i)},$$

where n and l are the principal and angular momentum quantum numbers ($n=1,2,3,\ldots$; $l=0,1,2,\ldots,n-1$); m_l and m_s are the eigenvalues of the projection on $\hat{\mathbf{z}}$ of the orbital and spin angular momenta ($m_l=-l, -l+1, \ldots, l-1, l$; $m_s=\pm 1/2$); r_i, θ_i, and ϕ_i are the polar coordinates of the ith electron, $Y_l^m(\theta,\phi)$ is a spherical harmonic, and $\chi_{m_s}^{(i)}$ is a spin function. The states $|nlm_lm_s\rangle$ are filled in order of increasing energy E_{nl} to form the lowest-energy state of the atom or ion. Often the semiempirical rule stating that the energy eigenvalues occur in the order of increasing $n+l$ is applied. The disposition of the electrons in the one-electron states is called a *configuration.*

We have seen that the eigenstates of the Hamiltonian (1) can be selected so that H, L^2, S^2, L_z, and S_z are simultaneously diagonal so that its levels are labeled by the values of L and S which represent the eigenvalues of the operators **L·L** and **S·S**, namely, $L(L+1)$ and $S(S+1)$, respectively. The states $|nLSM_LM_S\rangle$ have a degeneracy of order $(2L+1)(2S+1)$. These levels are designated by the notations $S,P,D,F,G,H,I,K,L,M,N,\ldots$ corresponding to $L=0,1,2,3,\ldots,10,\ldots$. The multiplicity $2S+1$ is placed as a superindex on the left of this symbol. Thus, a ^2P level corresponds to $L=1$, $S=1/2$. These levels are called *terms.* Group theory cannot tell us what is the order of the terms. An analysis of the configurations on the basis of the central-field approximation complying with the requirement of antisymmetry under exchange of identical particles is necessary.

1.2 The Ground State of Free Ions: Hund's Rules

We consider the nature of the lowest term arising from a given configuration. For example, the configuration of Ne is $(1s)^2(2s)^2(2p)^6$ in its lowest state. We can form only one antisymmetric wave function with this configuration, i.e., only one Slater determinant with the ten orbitals in the configuration. On the other hand, the Si atom has the configuration $(1s)^2(2s)^2(2p)^6(3s)^2(3p)^2$. Including the spin degeneracy, there are six $3p$ orbitals with which one can form 15 Slater determinants constructed with linear combinations of states $(m_{1l},m_{2l},m_{1s},m_{2s})$. They are classified as ^1S, ^3P, and ^1D, showing that the configuration of

Si splits into three *terms* which are onefold, ninefold, and fivefold degenerate, respectively.

We can derive the same result as follows using the irreducible representations $D^{(J)}$ of the group SO(3) (see GROUP THEORY). The combination of the orbital angular momenta gives $D^{(1)} \times D^{(1)} = D^{(0)} \oplus D^{(1)} \oplus D^{(2)}$ while that of the spin states corresponds to $D^{(1/2)} \times D^{(1/2)} = D^{(0)} \oplus D^{(1)}$. The state $L=2$, $M_L=2$ has two electrons in the $m_l=1$ state. Thus, its spin state must be antisymmetric, i.e., it belongs to $D^{(0)}$. Successive application of $L_- = L_x - iL_y$ gives the five states $|L,S,M_L,M_S\rangle$ with $L=2$, $S=0$, $M_L=2,1,0,-1,-2$, and $M_S=0$, forming the ^1D term with symmetric orbitals and antisymmetric spin states. The orbital part of the P states ($L=1$) is orthogonal to the D states with $M_L=1$. Both are linear combinations of $|1,0\rangle = |m_{l1}=1, m_{l2}=0\rangle$ and $|0,1\rangle = |m_{l1}=0, m_{l2}=1\rangle$, namely, $(|1,0\rangle + |0,1\rangle)2^{-1/2}$ and $(|1,0\rangle - |0,1\rangle)2^{-1/2}$. The first, being symmetric, is part of the ^1D term. The second can only be combined with the symmetric spin triplet, giving rise to nine states in a ^3P term. There remains the $L=S=0$ term ^1S.

Two $3d$ electrons $[(3d)^2]$ can be combined in the same way:

$$D^{(2)} \times D^{(2)} = D^{(0)} \oplus D^{(1)} \oplus D^{(2)} \oplus D^{(3)} \oplus D^{(4)}$$

and

$$D^{(1/2)} \times D^{(1/2)} = D^{(0)} \oplus D^{(1)}.$$

This yields the terms ^1S, ^3P, ^1D, ^3F, ^1G. The next task is to find the order of the terms. Their energies are split by the Coulomb repulsion; their order is established using a method due to Slater (1929) which makes use of the invariance of the trace of an operator under unitary transformations. The lowest state obtained by that method can be determined by the following rules known as Hund's rules:

1. The ground-state term has the highest spin multiplicity consistent with the Pauli principle. This rule is a consequence of the Pauli principle because electrons with parallel spins avoid one another.
2. The ground term has the largest possible value of L consistent with 1 and with the Pauli principle.

After the terms (L,S) have been obtained we introduce relativistic corrections to the Hamiltonian. The most important is the spin-orbit interaction; an accurate approximation has the form

$$H_2 = \sum_i \zeta(r_i)\mathbf{l}_i \cdot \mathbf{s}_i, \tag{2}$$

where $\zeta(r_i)$ is a function of the distance of the ith electron from the nucleus. We assume here that $H_2 \ll H_1$, the so-called *L-S* or Russell-Saunders coupling. If $H_2 \gg H_1$, we must form $\mathbf{j}_i = \mathbf{l}_i + \mathbf{s}_i$ and then combine the total angular momenta of the individual electrons (*j-j* coupling). This, and the intermediate case, are important in heavy elements but we shall not deal with them here.

The Hamiltonian $H = H_0 + H_1 + H_2$ is invariant under rotations so that $[H,\mathbf{J}]=0$ where $\mathbf{J} = \mathbf{L} + \mathbf{S} = \Sigma_i(\mathbf{l}_i + \mathbf{s}_i)$ is the *total* angular momentum of the system.

To proceed further we recall the Wigner-Eckart theorem which simplifies the calculation of matrix elements of tensor operators. We express the components of a tensor in such a way that, under rotations, they transform in the same manner as spherical harmonics. A tensor of rank k has components $T_\kappa^{(k)}$ where $\kappa = -k, -k+1, \ldots, k-1, k$. When the components of a tensor are given in this form they are said to be in irreducible form because they generate the irreducible representation $D^{(k)}$ of SO(3). The Wigner-Eckart theorem is applicable to expressions of the matrix elements of irreducible tensor components between states $|njm\rangle$ of J^2 and J_z, where $\mathbf{J}$ is the total angular momentum of the system. The index n collects, in a single symbol, all the other quantum numbers necessary to specify the states. The theorem states that

$$\begin{aligned}\langle n'j'm' | T_\kappa^{(k)} | njm\rangle &= (-1)^{2k}(2j'+1)^{-1/2}\langle n'j' \| T^{(k)} \| nj\rangle \\ &\quad \times \langle jk;j'm' | jk;m\kappa\rangle,\end{aligned} \tag{3}$$

where $\langle jk;j'm' | jk;m\kappa\rangle$ is a Clebsch-Gordan coefficient. The quantity $\langle n'j' \| T^{(k)} \| nj\rangle$ is a number depending only on $n'j'nj$ and k and not on the "angular" variables m', m, and κ. All angular dependence is thus contained in the Clebsch-Gordan coefficient which vanishes if $m' \neq m + \kappa$. For states with $j'=j$, the following special case is a consequence of the Wigner-Eckart theorem:

$$\langle n'jm'|V_\kappa|njm\rangle=\frac{\langle n'jm|\mathbf{V}\cdot\mathbf{J}|njm\rangle}{j(j+1)}\langle njm'|J_\kappa|njm\rangle. \tag{4}$$

Since **V·J** is a scalar, $\langle n'jm|\mathbf{V}\cdot\mathbf{J}|njm\rangle$ is independent of m. This shows that the matrix elements of any vector operator **V** between states $|jm'\rangle$ and $|jm\rangle$ (with all other quantum numbers the same) are proportional to those of **J**. We consider now a given term (L,S) of an atom. Then

$$\begin{aligned}\langle nLSM'_LM'_S|H_2|nLSM_LM_S\rangle&=\sum_i\frac{\langle nLM_L|\zeta(r_i)\mathbf{l}_i\cdot\mathbf{L}|nLM_L\rangle\langle SM_S|\mathbf{s}_i\cdot\mathbf{S}|SM_S\rangle}{L(L+1)S(S+1)}\\&\quad\times\langle LM'_L|\mathbf{L}|LM_L\rangle\cdot\langle SM'_S|\mathbf{S}|SM_S\rangle\\&=\lambda\langle nLSM'_LM'_S|\mathbf{L}\cdot\mathbf{S}|nLSM_LM_S\rangle,\end{aligned} \tag{5}$$

where

$$\lambda=\sum_i\frac{\langle nLM_L|\zeta(r_i)\mathbf{l}_i\cdot\mathbf{L}|nLM_L\rangle\langle SM_S|\mathbf{s}_i\cdot\mathbf{S}|SM_S\rangle}{L(L+1)S(S+1)}. \tag{6}$$

This means that if we consider only one term, H_2 can be replaced by

$$H_2=\lambda\mathbf{L}\cdot\mathbf{S}. \tag{7}$$

The function $\zeta(r)$ is positive for attractive potentials. For a particular choice of configuration, electrons with given values of n and l can be described by the parameter $\zeta(n,l)$ when dealing with the spin-orbit interaction, this parameter being defined by

$$\zeta(n,l)=\int_0^\infty\zeta(r)R^2_{nl}(r)r^2\,dr. \tag{8}$$

Taking $M'_L=M_L$ and $M'_S=M_S$, we have

$$\begin{aligned}&\left\langle nLSM_LM_S\middle|\sum_i\zeta(r_i)\mathbf{l}_i\cdot\mathbf{s}_i\middle|nLSM_LM_S\right\rangle\\&\quad=\langle nLSM_LM_S|\lambda\mathbf{L}\cdot\mathbf{S}|nLSM_LM_S\rangle\\&\quad=\lambda M_LM_S.\end{aligned}$$

We take for $|nLSM_LM_S\rangle$ the single Slater determinant with $M_S=S$. Then

$$\lambda M_LS=\zeta(n,l)\sum_i m_{li}m_{si}.$$

If the shell is less than half-filled, all m_{si} are the same, say, $m_{si}=1/2$ for $M_S=S$, and

$$\lambda=\frac{\zeta(n,l)}{M_LS}\frac{1}{2}\sum_i m_{li}=\frac{\zeta(n,l)}{2S}>0, \tag{9}$$

and the lowest state in the term corresponds to **L·S** negative and having the largest absolute value. When the shell is more than half-filled some of the spins must be reversed so that

$$\begin{aligned}\lambda&=\frac{\zeta(n,l)}{M_LS}\left(\frac{1}{2}\sum_{m_{li}=-l}^{l}m_{li}-\frac{1}{2}M_L\right)\\&=-\frac{\zeta(n,l)}{2S}<0.\end{aligned} \tag{10}$$

This leads to the third Hund rule:

3. The lowest level in a term of an atom is that for which $J=|L-S|$ if the incomplete shell is less than half-full and $J=L+S$ if the shell is more than half-full. Of course, if the shell is exactly half-filled $J=S$.

Once the lowest term has been found, the spin-orbit interaction H_2 splits the $(2L+1)(2S+1)$ degeneracy into states with $J=|L-S|,|L-S|+1,\ldots,L+S-1,L+S$. The additional energies are given by

$$\begin{aligned}E_J&=\langle nLSJM_J|\lambda\mathbf{L}\cdot\mathbf{S}|nLSJM_J\rangle\\&=\frac{\lambda}{2}[J(J+1)-L(L+1)-S(S+1)],\end{aligned} \tag{11}$$

which are, of course, independent of M_J. The separation of these levels is $E_J-E_{J-1}=\lambda J$, a result known as the Landé interval rule.

The Zeeman interaction in a magnetic field **B** is

$$H_Z=\mu_B\mathbf{B}\cdot(\mathbf{L}+g_e\mathbf{S}), \tag{12}$$

where $g_e \approx 2$. By virtue of the Wigner-Eckart theorem in the form given by Eq. (4), the matrix elements of H_Z for states in an atomic level characterized by $|nLSJM_J\rangle$ are

$$\langle nLSJM'_J|H_Z|nLSJM_J\rangle$$
$$=g\mu_B\mathbf{B}\cdot\langle nLSJM'_J|\mathbf{J}|nLSJM_J\rangle, \tag{13}$$

where g, the so-called Landé g factor, depends on the particular level, i.e., on L, S, and J, and is equal to

$$g=\frac{1}{2}(g_e+1)-\frac{1}{2}(g_e-1)\frac{L(L+1)-S(S+1)}{J(J+1)}. \tag{14}$$

For more details on the spectroscopy of atoms, the reader is referred to Condon and Shortley (1953) and to Heckmann and Träbert (1989); see also ATOMIC SPECTROSCOPY.

2. MAGNETIC PROPERTIES OF FREE IONS

2.1 Magnetic Susceptibilities of Free Ions: Theories of Langevin and Brillouin

In the previous section we have described the energy levels of free atoms and ions. In the presence of a magnetic field **B**, each J multiplet is split into $2J+1$ levels with additional energies $g\mu_B BM_J$ ($M_J=-J, -J+1,..., J-1,J$) where g is given by Eq. (14). These energy separations are observed, e.g., in magnetoabsorption experiments (Zeeman effect) and manifest themselves in the phenomenon of paramagnetism. If the energy separation between the lowest and the excited J multiplets is large compared to the Zeeman and thermal energies, the magnetization of an ensemble of N atoms or ions per unit volume is

$$M=Ng\mu_B JB_J\left(\frac{g\mu_B JB}{k_BT}\right), \tag{15}$$

where

$$B_J(x)=\frac{2J+1}{2J}\coth\left(\frac{2J+1}{2J}x\right)-\frac{1}{2J}\coth\left(\frac{x}{2J}\right) \tag{16}$$

is called the Brillouin function. Here J is the quantum number of the total angular momentum in the ground state of the ion and k_B is the Boltzmann constant. At ordinary temperatures and magnetic fields, $g\mu_B JB \ll k_BT$, and the magnetic susceptibility is

$$\chi=N(p\mu_B)^2/3k_BT, \tag{17}$$

where $p=g[J(J+1)]^{1/2}$ is called the effective number of Bohr magnetons. Table 1 gives a set of experimental and calculated values for

Table 1. Rare-earth ions, their ground-state levels, g factors, and calculated and experimental numbers of Bohr magnetons p (after Refs. 7 and 8).

	Configuration	Ground state	g	p calculated	p experimental
La^{3+}	$4f^0$	$^1S_0$		0.00	Diamagnetic
Ce^{3+}	$4f^1$	$^2F_{5/2}$	6/7	2.54	2.4
Pr^{3+}	$4f^2$	$^3H_4$	4/5	3.58	3.5
Nd^{3+}	$4f^3$	$^4I_{9/2}$	8/11	3.62	3.5
Pm^{3+}	$4f^4$	$^5I_4$	3/5	2.68	···
Sm^{3+}	$4f^5$	$^6H_{5/2}$	2/7	0.84	1.5
Eu^{3+}	$4f^6$	$^7F_0$		0.00	3.5
Gd^{3+}	$4f^7$	$^8S_{7/2}$	2	7.94	8.0
Tb^{3+}	$4f^8$	$^7F_6$	3/2	9.72	9.5
Dy^{3+}	$4f^9$	$^6H_{15/2}$	4/3	10.63	10.6
Ho^{3+}	$4f^{10}$	$^5I_8$	5/4	10.60	10.4
Er^{3+}	$4f^{11}$	$^4I_{15/2}$	6/5	9.59	9.5
Tm^{3+}	$4f^{12}$	$^3H_6$	7/6	7.57	7.3
Yb^{3+}	$4f^{13}$	$^2F_{7/2}$	8/7	4.54	4.5
Lu^{3+}	$4f^{14}$	$^1S_0$		0.00	Diamagnetic

representative rare-earth ions (having incomplete $4f$ shells) obtained at room temperature in several salts of these elements. These are collected from several references [see Van Vleck (1932) and Kittel (1986)]. Table 2 gives similar results for the transition-metal elements of the iron group. In this table values of p calculated with $g\sqrt{J(J+1)}$ and with $2\sqrt{S(S+1)}$ are listed. We note the following trends:

1. For the rare-earth salts there is good agreement between the theoretical and measured values of the susceptibility expressed in terms of p, the effective number of Bohr magnetons, except for Eu and Sm.
2. In the case of transition-metal ions we note that the agreement between the theoretical and experimental numbers of Bohr magnetons is unsatisfactory but there is improved agreement if, instead of using $p=g\sqrt{J(J+1)}$, we use $p=2\sqrt{S(S+1)}$. This suggests that the orbital contribution to the magnetic moment is zero or, as we usually say, that the orbital contribution to the magnetic moment is *quenched.*
3. There is a significant deviation between $p = 2\sqrt{S(S+1)}$ and the experimental value of p for Fe^{2+}, Co^{2+}, Ni^{2+}, and Cu^{2+}. This seems to indicate that there is some contribution to p due to the orbital motion of the electrons in the $3d$ shell.

Item **1** can be understood in terms of the properties of free ions. Items **2** and **3** require for their analysis a study of the effect of the crystal environment on the magnetic atoms. This will be discussed in Sec. 3. In the remainder of this section we discuss the "anomalous" behavior of europium and samarium.

2.2 Van Vleck Paramagnetism: The Behavior of Europium and Samarium

The magnetic properties of Eu^{3+} and Sm^{3+} are understood if we recognize that the splitting of the lowest term by the spin-orbit coupling is small. Consider, for example, Eu^{3+} which has a ^7F ground term. Since the $4f$ shell is less than half-filled the lowest level is $^7F_0$ corresponding to $J=0$. According to the simple theory, Eu^{3+} should be diamagnetic. But measurement of the magnetic susceptibility at room temperature shows that p, the number of effective Bohr magnetons, is 3.5. The ^7F term splits into the levels $^7F_0$, $^7F_1$,...,$^7F_6$ by the spin-orbit interaction $\lambda\mathbf{L}\cdot\mathbf{S}$. All levels, except the lowest, have $g=3/2$ and $\lambda\approx 255$ cm^{-1}.

If we denote the energy states of the atoms by $|n\rangle$ and the probability that the state $|n\rangle$ be occupied at temperature T by p_n (in the absence of an external magnetic field), the linear magnetic susceptibility is

$$\chi=\frac{N\mu_B^2}{k_BT}\left[\sum_n p_n|\langle n|L_z+2S_z|n\rangle|^2 -\left(\sum_n p_n\langle n|L_z+2S_z|n\rangle\right)^2\right]$$

Table 2. Transition-metal ions of the iron group, their ground-state levels, g factors, and calculated and experimental numbers of Bohr magnetons, p (after Refs. 7 and 8).

	Configuration	Ground state	g	$p = g\sqrt{J(J+1)}$	$p = 2\sqrt{S(S+1)}$	p experimental
Ti^{3+}	$3d^1$	$^2D_{3/2}$	4/5	1.55	1.73	···
V^{4+}	$3d^1$	$^2D_{3/2}$	4/5	1.55	1.73	1.8
V^{3+}	$3d^2$	$^3F_2$	2/3	1.63	2.83	2.8
V^{2+}	$3d^3$	$^4F_{3/2}$	2/5	0.77	3.87	3.8
Cr^{3+}	$3d^3$	$^4F_{3/2}$	2/5	0.77	3.87	3.7
Mn^{4+}	$3d^3$	$^4F_{3/2}$	2/5	0.77	3.87	4.0
Cr^{2+}	$3d^4$	$^5D_0$		0.00	4.90	4.8
Mn^{3+}	$3d^4$	$^5D_0$		0.00	4.90	5.0
Mn^{2+}	$3d^5$	$^6S_{5/2}$	2	5.92	5.92	5.9
Fe^{3+}	$3d^5$	$^6S_{5/2}$	2	5.92	5.92	5.9
Fe^{2+}	$3d^6$	$^5D_4$	3/2	6.70	4.90	5.4
Co^{2+}	$3d^7$	$^4F_{9/2}$	4/3	6.63	3.87	4.8
Ni^{2+}	$3d^8$	$^3F_4$	5/4	5.59	2.83	3.2
Cu^{2+}	$3d^9$	$^2D_{5/2}$	6/5	3.55	1.73	1.9

$$+N\mu_{\mathrm{B}}^{2}\sum_{n\neq n'}\frac{p_{n}-p_{n'}}{E_{n'0}-E_{n0}}|\langle n'|L_{z}+2S_{z}|n\rangle|^{2}$$

$$-\frac{Ne^{2}}{4mc^{2}}\sum_{n}p_{n}\left\langle n\left|\sum_{i}(x_{i}^{2}+y_{i}^{2})\right|n\right\rangle. \quad (18)$$

Here E_{n0} is the energy eigenvalue of the $|n\rangle$ eigenstate in the absence of the Zeeman interaction. The first term in Eq. (18) is the ordinary, temperature-dependent paramagnetic susceptibility; the second, which becomes important at low temperatures, owes its origin to mixing of states by the Zeeman interaction. It is particularly important in ions like Eu^{3+} whose lowest level has zero total angular momentum; at the lowest temperatures it leads to a temperature-independent paramagnetic susceptibility known as Van Vleck paramagnetism. The third term in Eq. (18) is the diamagnetic contribution to the susceptibility. It will be neglected from now on. Another expression for χ, equivalent to Eq. (18), is

$$\chi=\frac{N}{k_{\mathrm{B}}T}\left[\left\langle\left(\frac{\partial E_{n}}{\partial B}\right)^{2}_{B=0}\right\rangle-\left(\left\langle\frac{\partial E_{n}}{\partial B}\right\rangle_{B=0}\right)^{2}\right]$$

$$-N\left\langle\left(\frac{\partial^{2}E_{n}}{\partial B^{2}}\right)_{B=0}\right\rangle, \quad (19)$$

where $E_{n}=E_{n}(B)$ is the energy of the n quantum-mechanical state as a function of the magnetic induction B; the averages are taken over the statistical distribution in zero magnetic field. The term $-N(\partial^{2}E_{n}/\partial B^{2})_{B=0}$ gives the Van Vleck paramagnetism. The results of the calculation of $E_{n}(B)$ to second order in B for Eu^{3+} are given in Table 3 for the three lowest J multiplets. We find

$$\chi=N\mu_{\mathrm{B}}^{2}\left[\frac{8}{\Delta}+\left(\frac{9}{2k_{\mathrm{B}}T}-\frac{1}{2\Delta}\right)e^{-\Delta/k_{\mathrm{B}}T}\right.$$

$$\left.+\left(\frac{45}{2k_{\mathrm{B}}T}-\frac{5}{6\Delta}\right)e^{-3\Delta/k_{\mathrm{B}}T}+\cdots\right]$$

$$\times[1+3e^{-\Delta/k_{\mathrm{B}}T}+5e^{-3\Delta/k_{\mathrm{B}}T}+\cdots]^{-1}, \quad (20)$$

where Δ is the energy separation between the $^{7}F_{1}$ and $^{7}F_{0}$ levels, i.e., $\Delta=\lambda$. Using the experimental value of the spin-orbit interaction, we find $p\approx3.5$ at $T=293$ K, in good agreement with the measurements of the susceptibility. We note that, as the temperature is lowered toward zero, χ approaches $8N\mu_{\mathrm{B}}^{2}/\Delta$.

Table 3. Energies of the low-lying states of Eu^{3+} in an external magnetic field.

J	M_J	Energy
0	0	$-4(\mu_{\mathrm{B}}B)^{2}/\Delta$
1	-1	$\Delta-(3/2)\mu_{\mathrm{B}}B-9(\mu_{\mathrm{B}}B)^{2}/8\Delta$
	0	$\Delta+5(\mu_{\mathrm{B}}B)^{2}/2\Delta$
	1	$\Delta+(3/2)\mu_{\mathrm{B}}B-9(\mu_{\mathrm{B}}B)^{2}/\Delta$
2	-2	$3\Delta-3\mu_{\mathrm{B}}B-10(\mu_{\mathrm{B}}B)^{2}/21\Delta$
	-1	$3\Delta-(3/2)\mu_{\mathrm{B}}B+61(\mu_{\mathrm{B}}B)^{2}/168\Delta$
	0	$3\Delta+9(\mu_{\mathrm{B}}B)^{2}/14\Delta$
	1	$3\Delta+(3/2)\mu_{\mathrm{B}}B+61(\mu_{\mathrm{B}}B)^{2}/168\Delta$
	2	$3\Delta+3\mu_{\mathrm{B}}B-10(\mu_{\mathrm{B}}B)^{2}/21\Delta$

Samarium exhibits a similar anomaly. The ground state of Sm^{3+} is $^{6}H_{5/2}$, which is separated from the next level, $^{6}H_{7/2}$, by $\Delta=(7\lambda/2)=932\ \mathrm{cm}^{-1}$. A calculation for Sm^{3+} similar to that for Eu^{3+} makes use of the energy levels given in Table 4. The magnetic susceptibility is

$$\chi=N\mu_{\mathrm{B}}^{2}\left[\frac{5}{7k_{\mathrm{B}}T}+\frac{60}{7\Delta}+\left(\frac{2704}{189k_{\mathrm{B}}T}\right.\right.$$

$$\left.\left.+\frac{3256}{1701\Delta}\right)e^{-\Delta/k_{\mathrm{B}}T}+\cdots\right]$$

$$\times[3+4e^{-\Delta/k_{\mathrm{B}}T}+\cdots]^{-1}. \quad (21)$$

At $T=293$ K we find $p=1.6$ which compares favorably with the experimental value of 1.5.

3. THE EFFECT OF THE CRYSTAL FIELD ON MAGNETIC ATOMS

3.1 The Crystal Field

We consider an atom or ion in a crystal field. The outer electrons usually form covalent bonds with those of the neighboring atoms and, hence, produce no magnetic effects other than the ubiquitous diamagnetism. However, if the atom has incomplete d or f shells, the electrostatic field produced by the total charge distribution around the ion has significant influence on its magnetic properties. This field, $\Phi(\mathbf{x})$, must satisfy Laplace's equation, and, having no singularities in the region of the ion itself, it is a linear combination of functions of the form $r^{l}Y_{l}^{m}(\theta,\phi)$ where $r=|\mathbf{x}|$ is measured from the center of the ion, and θ, ϕ are the polar angles with respect to a

Table 4. Energies of the low-lying states of Sm^{3+} in an external magnetic field.

J	M_J	Energy
5/2	−5/2	$-(5/7)\mu_B B-45(\mu_B B)^2/49\Delta$
	−3/2	$-(3/7)\mu_B B-75(\mu_B B)^2/49\Delta$
	−1/2	$-(1/7)\mu_B B-90(\mu_B B)^2/49\Delta$
	1/2	$(1/7)\mu_B B-90(\mu_B B)^2/49\Delta$
	3/2	$(3/7)\mu_B B-75(\mu_B B)^2/49\Delta$
	5/2	$(5/7)\mu_B B-45(\mu_B B)^2/49\Delta$
7/2	−7/2	$\Delta-(26/9)\mu_B B-2548(\mu_B B)^2/3645\Delta$
	−5/2	$\Delta-(130/63)\mu_B B-54\,466(\mu_B B)^2/178\,605\Delta$
	−3/2	$\Delta-(26/21)\mu_B B-838(\mu_B B)^2/19\,845\Delta$
	−1/2	$\Delta-(26/63)\mu_B B+3184(\mu_B B)^2/35\,721\Delta$
	1/2	$\Delta+(26/63)\mu_B B+3184(\mu_B B)^2/35\,721\Delta$
	3/2	$\Delta+(26/21)\mu_B B-838(\mu_B B)^2/19\,845\Delta$
	5/2	$\Delta+(130/63)\mu_B B-54\,466(\mu_B B)^2/178\,605\Delta$
	7/2	$\Delta+(26/9)\mu_B B-2548(\mu_B B)^2/3645\Delta$

suitably chosen coordinate system. The Hamiltonian of the ion in question becomes

$$H=H_0+H_1+H_2+H', \tag{22}$$

where

$$H'=-e\sum_i \Phi(\mathbf{x}_i)$$
$$=\sum_i \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_l^m r_i^l Y_l^m(\theta_i,\phi_i) \tag{23}$$

and H_0, H_1, and H_2 are the central-field approximation Hamiltonian, the Coulomb repulsion between the electrons, and the spin-orbit interaction, respectively. The sum over i extends over all electrons in the atom.

Restrictions on the coefficients A_l^m are imposed by the requirements that H' be real and that it remain invariant under the symmetry operations leaving the site of the ion unchanged. In order for H' to be real, we must have

$$A_l^{-m}=(-1)^m A_l^{m*}. \tag{24}$$

If $\mathbf{x}=0$ is a center of inversion, $\Phi(\mathbf{x})$ must be even in x, y, and z so that only even-l terms appear in its expansion. If the (x,y) plane is a mirror plane, σ_h, then, since $\Phi(\mathbf{x})$ must be invariant under the transformation $z\to z'=-z$ (or $\theta\to\theta'=\pi-\theta$), $A_l^m=0$ unless $l+m$ is even. If the z axis is an n-fold axis of rotational symmetry, then m must be a multiple of n. If the xz plane is a vertical mirror plane, σ_v, since $\Phi(\mathbf{x})$ is invariant under $\theta\to\theta'=\theta$, $\phi\to\phi'=2\pi-\phi$, then $A_l^{-m}=(-1)^m A_l^m$ and A_l^m is real.

The effect of the crystal field on the energy levels depends on the relative importance of H' compared to the remaining terms in the Hamiltonian of the ion. We consider three possible situations:

1. Weak field: $H_1 \gg H_2 \gg H'$. Here H' causes a small splitting of the J multiplets of the ion. The magnetic properties of ions for which this approximation holds are similar to those of the free ions except when $k_B T \lesssim H'$. Many salts of the rare earths behave in this way. These salts contain atoms with incomplete nf shells. The $(n+1)s$ and $(n+1)p$ shells are filled so that the f-shell electrons are partially screened from the outside electrostatic field by the eight electrons in the s and p shells above the incomplete f shell.
2. Intermediate field: $H_1 \gg H' \gg H_2$. In this case H' splits the Russell-Saunders terms. We note that, since H' is independent of the spin, if we disregard H_2, the spin is a constant of the motion. In this case, if the lowest level is nondegenerate, there is no orbital contribution to the magnetic moment of the ion and the material behaves, in a first approximation, as a collection of spins with angular momentum $\mathbf{S}$. This situation often occurs in compounds of the iron group elements which have an incomplete $3d$ shell. The outer $(4s)^2$ electrons participate in the binding of the system, and do not contribute significantly to the screening of the $3d$ shell. This is the reason why H' is usually stronger than the spin-orbit interaction H_2. We say that the orbital

angular momentum **L** is quenched. This result is proved as follows. In the absence of an external magnetic field $[H,T]=0$ where T is the time-reversal operator. If $H\Psi_0 = E_0\Psi_0$ where Ψ_0 is nondegenerate, then $T\Psi_0 = c\Psi_0$ where $|c|=1$. Now

$$\langle T\Psi_0|\mathbf{L}|T\Psi_0\rangle = -\langle\Psi_0|\mathbf{L}|\Psi_0\rangle.$$

But

$$\langle T\Psi_0|\mathbf{L}|T\Psi_0\rangle = c^*c\langle\Psi_0|\mathbf{L}|\Psi_0\rangle = \langle\Psi_0|\mathbf{L}|\Psi_0\rangle$$

so that $\langle\Psi_0|\mathbf{L}|\Psi_0\rangle=0$.

3. Strong field: $H' \gg H_1 \gg H_2$. In this case H' is so strong that the description of the electrons in terms of the central-field approximation breaks down. In a first approximation the electron states are described by quantum numbers appropriate to the crystal potential alone and H_1 and H_2 are treated as perturbations. This situation occurs in the cyanides of elements of the iron group, in many compounds of elements of the palladium and platinum groups, and in some salts of the trans-uranic elements.

3.2 Weak Crystal Field: Salts of the Rare-Earth Elements

We study first case **1**. The ground term of the ion is characterized by L and S and is split by H_2 into several J multiplets with $J=|L-S|,|L-S|+1,\ldots,L+S-1,L+S$. The states corresponding to J generate one of the irreducible representations $D_{\pm}^{(J)}$ of O(3). The group of H' is, in general, a finite subgroup of O(3). The new level scheme is, therefore, obtained by reducing the representation of the group of H' generated by the $2J+1$ states $|JM_J\rangle$.

For example, the symmetry of the Ce^{3+} ion in cerium ethyl sulfate, $Ce(C_2H_5SO_4)_3\cdot 9H_2O$, is $C_{3h}=\bar{6}$. The symmetry elements of this group are

$$C_{3h}=\{E,C_3,C_3^{-1},\sigma_h,\sigma_hC_3,\sigma_hC_3^{-1}\}.$$

It is an Abelian group so that the double group $\bar{C}_{3h}$ has 12 elements and 12 irreducible representations. Since Ce^{3+} contains a single $4f$ electron its levels are classified according to the double-valued representations of $\bar{C}_{3h}$. The reduction of the $D_{-}^{(5/2)}$ representation is $D_{-}^{(5/2)}=\Gamma_7\oplus\Gamma_8\oplus\Gamma_9\oplus\Gamma_{10}\oplus\Gamma_{11}\oplus\Gamma_{12}$. Thus, since the ground state of Ce^{3+} is $^2F_{5/2}$ its ground state splits into three doublets: $\Gamma_7+\Gamma_8$, $\Gamma_9+\Gamma_{10}$, and $\Gamma_{11}+\Gamma_{12}$ corresponding to $M_J=\pm 1/2,\pm 3/2,\pm 5/2$, respectively. We recall that time-reversal invariance requires the double degeneracies of the Kramers doublets $\Gamma_7+\Gamma_8$, $\Gamma_9+\Gamma_{10}$, $\Gamma_{11}+\Gamma_{12}$. To obtain numerical results we must go beyond group theory and make use of perturbation theory. The potential H' can be written as

$$H'=\sum_{l=0}^{\infty}\sum_{m=-l}^{l}A_l^mT_m^{(l)}, \tag{25}$$

where

$$T_m^{(l)}=\sum_i r_i^lY_l^m(\theta_i,\phi_i). \tag{26}$$

The operators $T_m^{(l)}$ for $m=-l,-l+1,\ldots,l-1,l$ are the irreducible components of a tensor operator of rank l. Thus, by virtue of the Wigner-Eckart theorem

$$\langle J'M_J'|T_m^{(l)}|JM_J\rangle=(2J'+1)^{-1/2}\langle J'\|T^{(l)}\|J\rangle \times\langle J,l;J'M_{J'}'|J,l;M_Jm\rangle. \tag{27}$$

But the Clebsch-Gordan coefficient vanishes unless $M_{J'}'=M_J+m$, and l is restricted to the range $|J'-J|,|J'-J|+1,\ldots,J'+J-1,J'+J$. Hence

$$\langle J'M_{J'}'|H'|JM_J\rangle = \sum_{l=|J'-J|}^{J'+J}A_l^{M_{J'}'-M_J}\frac{\langle J'\|T^{(l)}\|J\rangle}{(2J'+1)^{1/2}} \times\langle J,l;J'M_{J'}'|J,l;M_J,M_{J'}'-M_J\rangle. \tag{28}$$

We return now to the example of Ce^{3+} whose lower term is split into $^2F_{5/2}$ and $^2F_{7/2}$ by the spin-orbit interaction. To determine the number of parameters required in the expansion of the potential we recall that m must be a multiple of 3 and that $l+m$ must be even. Thus, we need $A_2^0,A_3^0,A_3^{\pm 3},A_4^0,A_5^{\pm 3},A_6^0,A_6^{\pm 6},\ldots$. Since only terms with l between 1 and 6 appear in the matrix elements we can stop after the list above. If we consider only the lowest term, i.e., if we neglect hybridization with terms above ^2F, there is an additional simplification. In fact, the relevant matrix elements of $T_m^{(l)}$ are between $4f$ states which have negative parity. Thus, only terms with l

even yield nonvanishing matrix elements. Further, $A_l^{-m}=(-1)^m A_l^{m*}$. By taking the $\Phi=0$ fiducial line in a suitable way we can require $A_6^6 = A_6^{-6}$ to be real. Thus

$$\begin{aligned}\Phi(\mathbf{x}) &= A_2^0 r^2 Y_2^0(\theta,\phi) + A_4^0 r^4 Y_4^0(\theta,\phi)\\ &\quad + A_6^0 r^6 Y_6^0(\theta,\phi)\\ &\quad + A_6^6 r^6 [Y_6^6(\theta,\phi) + Y_6^{-6}(\theta,\phi)] + \cdots, \qquad (29)\end{aligned}$$

where the additional terms give zero contributions to the matrix elements of H'. Experimentally, it is known that the separation of $^2F_{7/2}$ from $^2F_{5/2}$ is of order 2000 cm^{-1} while the effect of H' is about 200 cm^{-1}. Thus, in a first approximation we will consider only $J'=J$ matrix elements. Since $2J=5$, we can disregard the A_6^0 and A_6^6 contributions to $\Phi(\mathbf{x})$. Thus we use

$$\Phi(\mathbf{x}) = B_2^0(3z^2-r^2) + B_4^0(35z^4-30r^2z^2+3r^4) + \cdots \qquad (30)$$

with

$$B_2^0 = A_2^0(5/16\pi)^{1/2}, \quad B_4^0 = A_4^0(3/16\sqrt{\pi}). \qquad (31)$$

Since we are dealing here with a single $4f$ electron it is possible to calculate in a simple way the matrix elements of Φ between the $4f$ states. We can describe the states of the ion by Slater determinants with all levels filled. The several states differ only in the particular $4f$ eigenfunction that appears in each. In the diagonal matrix elements the contributions of the filled shells give zero so that we finally have an integral over the position of the $4f$ electron. But this procedure becomes more complicated with the configurations $(4f)^n$ for $n=2,3,\ldots$ so that we use a more powerful approach developed by Stevens (1952) and reviewed by Hutchings (1964) called the method of "operator equivalents." This method is based on the Wigner-Eckart theorem already stated in Eq. (3).

3.3 The Method of Operator Equivalents

The matrix element of $T_\kappa^{(k)}$, the κ-irreducible component of a tensor of rank k, between two states $|njm\rangle$ and $|njm'\rangle$ having the same angular momentum j is given in Eq. (3) with $n'=n$ and $j'=j$. Suppose now that a second irreducible component of a tensor of rank k, $O_\kappa^{(k)}$, has matrix elements which are easier to calculate than those of $T^{(k)}$. Those are again proportional to the same Clebsch-Gordan coefficient as for $T_\kappa^{(k)}$. Thus, the matrix elements of $T_\kappa^{(k)}$ are simply proportional to those of $O_\kappa^{(k)}$, the constant of proportionality being independent of m, m', and κ. We form operators $O_\kappa^{(k)}$ by taking combinations of powers of degree k of components of $\mathbf{J}$ with zero trace. The operator equivalent of $3z^2-r^2=2z^2-x^2-y^2$ is, clearly, $3J_z^2 - \mathbf{J}\cdot\mathbf{J} = 2J_z^2 - J_x^2 - J_y^2$. For more complicated operators we must write the powers of J_x, J_y, J_z in a completely symmetrical fashion. For example, the operator equivalent to

$$\sum_i (35z_i^4 - 30z_i^2r_i^2 + 3r_i^4)$$

is obtained by writing

$$\begin{aligned}35z^4-30z^2r^2+3r^4 &= 35z^4-30z^2(x^2+y^2+z^2)\\ &\quad +3(x^4+y^4+z^4+2y^2z^2\\ &\quad +2z^2x^2+2x^2y^2).\end{aligned}$$

The operator equivalents to x^4, y^4, and z^4 are, simply, J_x^4, J_y^4, and J_z^4, respectively. The operator equivalent of x^2y^2 is

$$\begin{aligned}&\tfrac{1}{6}(J_x^2J_y^2 + J_y^2J_x^2 + J_xJ_yJ_xJ_y + J_yJ_xJ_yJ_x + J_xJ_y^2J_x\\ &\quad + J_yJ_x^2J_y) = \tfrac{1}{6}(3J_x^2J_y^2 + 3J_y^2J_x^2 - 2\mathbf{J}\cdot\mathbf{J} + 5J_z^2).\end{aligned}$$

From these results we conclude that

$$\begin{aligned}O_0^{(4)} &= 35J_z^4 - 30J(J+1)J_z^2 + 25J_z^2 - 6J(J+1)\\ &\quad + 3J^2(J+1)^2. \qquad (32)\end{aligned}$$

Therefore,

$$\begin{aligned}&\left\langle JM_J' \middle| \sum_i (3z_i^2 - r_i^2) \middle| JM_J \right\rangle\\ &\quad = a\langle JM_J' | 3J_z^2 - \mathbf{J}\cdot\mathbf{J} | JM_J\rangle\\ &\quad = a[3M_J^2 - J(J+1)]\delta_{M_JM_J'} \qquad (33)\end{aligned}$$

and

$$\begin{aligned}&\left\langle JM_J' \middle| \sum_i (35z_i^4 - 30z_i^2r_i^2 + 3r_i^4) \middle| JM_J \right\rangle\\ &\quad = b\langle JM_J' | O_0^{(4)} | JM_J\rangle\\ &\quad = b[35M_J^4 - 30J(J+1)M_J^2 + 25M_J^2\\ &\qquad - 6J(J+1) + 3J^2(J+1)^2]\delta_{M_JM_J'}, \qquad (34)\end{aligned}$$

where a and b are appropriate numerical coefficients. In this approximation, the shifts in the energy levels are

$$\begin{aligned} \Delta E_{\pm 1/2} &= -8aB_2^0 + 120bB_4^0, \\ \Delta E_{\pm 3/2} &= -2aB_2^0 - 180bB_4^0, \\ \Delta E_{\pm 5/2} &= 10aB_2^0 + 60bB_4^0. \end{aligned} \tag{35}$$

The coefficients B_2^0 and B_4^0 are simply related to A_2^0 and A_4^0. The latter can be calculated accurately for ionic crystals and for crystals in which the electron density is known as a function of position. The parameters a and b are determined using the radial wave function $R(r)$ of the $4f$ electron. We find

$$a = -\tfrac{2}{35}\,\overline{r^2}, \quad \overline{r^2} = \int_0^\infty R^2(r) r^4 \, dr, \tag{36}$$

and

$$b = \tfrac{1}{60}\,\tfrac{8}{21}\,\overline{r^4}, \quad \overline{r^4} = \int_0^\infty R^2(r) r^6 \, dr. \tag{37}$$

3.4 Crystal Field of Intermediate Strength: The Spin-Hamiltonian Formalism

The study of case **2** requires the introduction of the so-called *spin-Hamiltonian* formalism (Pryce, 1950; Abragam and Pryce, 1951). It is applicable when $H_1 \gg H' \gg H_2$. We suppose that we have obtained the solution of the Schrödinger equation of the ion in the crystal field without spin-orbit interaction, i.e.,

$$(H_0 + H_1 + H') \, |nr\rangle = E_n \, |nr\rangle. \tag{38}$$

The label n denotes the orbital levels and $r = 1,2,\ldots,l_n$, the different orthonormal states with energy E_n. We note that, since $H_0 + H_1 + H'$ does not contain the spin operator, $\mathbf{S}$ is a constant of the motion. Taking account of the spin, the states are $|nrM_S\rangle$ having degeneracy $l_n(2S+1)$. The ground state is denoted by $n=0$. We introduce now the spin-orbit interaction H_2 in the form $\lambda\mathbf{L}\cdot\mathbf{S}$ appropriate to the case in which a single term is taken into account. For additional generality we assume that a magnetic field of induction $\mathbf{B}$ has been applied. Thus the perturbation on the level E_n characterized by the multiplet $\{|nrM_S\rangle\}$ is

$$H'' = \lambda\mathbf{L}\cdot\mathbf{S} + \mu_B \mathbf{B}\cdot(\mathbf{L} + 2\mathbf{S}). \tag{39}$$

The splittings are obtained to second order in H'' by diagonalizing a matrix whose elements are

$$\begin{aligned} \langle r'M'_S|H_S|rM_S\rangle &= \lambda\langle nr'|\mathbf{L}|nr\rangle\cdot\langle M'_S|\mathbf{S}|M_S\rangle + \mu_B\mathbf{B}\cdot(\delta_{M'_S M_S}\langle nr'|\mathbf{L}|nr\rangle + 2\delta_{r'r}\langle M'_S|\mathbf{S}|M_S\rangle) \\ &\quad + \sum_{n'r''M''_s}{}' \frac{\langle nr'M'_S|H''|n'r''M''_S\rangle\langle n'r''M''_S|H''|nrM_S\rangle}{E_n - E_{n'}} + \cdots. \end{aligned} \tag{40}$$

Here the prime means that the sum extends over all n' except $n'=n$. The matrix elements of H'' are of the form

$$\begin{aligned} &\langle n'r''M''_S|H''|nrM_S\rangle \\ &\quad = \lambda\langle n'r''|\mathbf{L}|nr\rangle\cdot\langle M''_S|\mathbf{S}|M_S\rangle + \mu_B\mathbf{B}\cdot\langle n'r''|\mathbf{L}|nr\rangle\delta_{M''_S M_S}. \end{aligned} \tag{41}$$

The term in $\mathbf{S}$ drops out because of the orthogonality of $|nr\rangle$ and $|n'r''\rangle$ for $n' \neq n$. The second-order contribution to $\langle r'M'_S|H_S|rM_S\rangle$ is $(i,j = x,y,z)$

$$\begin{aligned} &\sum_{ij}\sum_{n'}{}' \frac{1}{E_n - E_{n'}} \sum_{r''M''_S} (\lambda\langle nr'|L_i|n'r''\rangle\langle M'_S|S_i|M''_S\rangle \\ &\quad + \mu_B B_i\langle nr'|L_i|n'r''\rangle\delta_{M'_S M''_S})(\lambda\langle n'r''|L_j|nr\rangle\langle M''_S|S_j|M_S\rangle + \mu_B B_j\langle n'r''|L_j|nr\rangle\delta_{M''_S M_S}). \end{aligned} \tag{42}$$

We note that the sum over M''_S can be carried out:

$$\sum_{M''_S} \langle M'_S|S_i|M''_S\rangle\langle M''_S|S_j|M_S\rangle = \langle M'_S|S_iS_j|M_S\rangle. \tag{43}$$

For each orbital level we define a matrix of dimension l_n by means of the equation

$$\langle r'|D_{ij}^{(n)}|r\rangle=\lambda^2 \sum_{n'}{}' \sum_{r''} \frac{\langle nr'|L_i|n'r''\rangle\langle n'r''|L_j|nr\rangle}{E_n-E_{n'}}. \tag{44}$$

The second-order contribution to H_S is, therefore,

$$\begin{aligned}&\langle r'M'_S|H_S|rM_S\rangle_2\\ &=\sum_{ij}\langle r'|D_{ij}^{(n)}|r\rangle(\langle M'_S|S_iS_j|M_S\rangle\\ &\quad+(\mu_B/\lambda)\langle M'_S|S_iB_j+S_jB_i|M_S\rangle\\ &\quad+(\mu_B/\lambda)^2B_iB_j\delta_{M'_SM_S}).\end{aligned} \tag{45}$$

Then, if we denote the matrix $D_{ij}^{(n)}$ simply by this symbol and $\langle nr'|\mathbf{L}|nr\rangle$ by $\mathbf{M}$ we can write

$$\begin{aligned}H_S=&\lambda\mathbf{M}\cdot\mathbf{S}+\mu_B\mathbf{B}\cdot(\mathbf{M}+2\mathbf{S})+\sum_{ij}D_{ij}^{(n)}S_iS_j\\ &+(\mu_B/\lambda)\sum_{ij}(D_{ij}^{(n)}+D_{ji}^{(n)})S_iB_j\\ &+(\mu_B/\lambda)^2\sum_{ij}D_{ij}^{(n)}B_iB_j.\end{aligned} \tag{46}$$

The last term is understood as the direct product of $D_{ij}^{(n)}$ by the unit matrix of dimension $(2S+1)$. For an orbital nondegenerate level, $\mathbf{M}=0$ so that

$$\begin{aligned}H_S=&2\mu_B\mathbf{B}\cdot\mathbf{S}+\sum_{ij}D_{ij}S_iS_j\\ &+(\mu_B/\lambda)\sum_{ij}(D_{ij}+D_{ji})S_iB_j\\ &+(\mu_B/\lambda)^2\sum_{ij}D_{ij}B_iB_j,\end{aligned} \tag{47}$$

where

$$D_{ij}=\lambda^2\sum_{n',r'}{}'\frac{\langle n|L_i|n'r'\rangle\langle n'r'|L_j|n\rangle}{E_n-E_{n'}} \tag{48}$$

is a number for each set of indices i, j. The spin Hamiltonian is often written as

$$\begin{aligned}H_S=&\sum_{ij}D_{ij}S_iS_j+\mu_B\sum_{ij}g_{ij}S_iB_j\\ &+(\mu_B/\lambda)^2\sum_{ij}D_{ij}B_iB_j,\end{aligned} \tag{49}$$

where

$$g_{ij}=2\delta_{ij}+(D_{ij}+D_{ji})/\lambda. \tag{50}$$

3.5 $\mathbf{Fe^{2+}}$ in a Tetrahedral Environment

We consider, as an example, the energy-level splitting of the ^5D term of Fe^{2+} in a crystal environment of symmetry T_d. These results were first obtained by Slack *et al.* (1969) using a procedure different from that given here following the method of Villeret *et al.* (1990a,b). The lowest configuration of Fe^{2+} is $(3d)^6$. The ion has, therefore, an even number of electrons so that its energy levels are classified according to the single-valued representations of $\overline{T}_d$. The crystal potential H' is of the form

$$\begin{aligned}H'=&A_3\sum_i x_iy_iz_i\\ &-A_4\sum_i(x_i^4+y_i^4+z_i^4-\tfrac{3}{5}r_i^4)+\cdots.\end{aligned} \tag{51}$$

A constant term has been omitted since it shifts all the energy levels by the same amount. The term in xyz gives zero matrix elements between the even $3d$ states. There remains the contribution containing fourth powers of the coordinates. Its operator equivalent is

$$\begin{aligned}O^{(4)}=&\tfrac{1}{8}(L_+^2+L_-^2)^2+\tfrac{3}{2}L_z^4-L(L+1)L_z^2\\ &-\tfrac{1}{10}L^2(L+1)^2+\tfrac{1}{5}L(L+1).\end{aligned} \tag{52}$$

The constant before the potential we denote by $\pm a$ $(a>0)$. If the configuration $(3d)^n$ is more than half-full $(n>5)$

$$H'=\pm aO^{(4)}, \tag{53}$$

and we take the positive sign for n odd and the negative sign for n even. For $n<5$ the opposite choice must be made. Thus for Fe^{2+}: $(3d)^6$,

$$H'=-a[\tfrac{1}{8}(L_+^2+L_-^2)^2+\tfrac{3}{2}L_z^4-6L_z^2-\tfrac{12}{5}]. \tag{54}$$

The states of H_0+H_1 which include the central-field Hamiltonian and the Coulomb repulsion between the electrons (H_1) can be characterized by $|M_L\rangle$, $M_L=2,1,0,-1,-2$, with $L_z|M_L\rangle=M_L|M_L\rangle$. These states generate the representation $D_+^{(2)}$ of O(3). The level splitting is obtained, reducing this representation when these functions generate a representation of

T_d. We obtain $D_+^{(2)} = \Gamma_3 + \Gamma_5$. This result can be deduced directly because the states $|M_L\rangle$ transform under the operations of SO(3) as the spherical harmonics $Y_L^{M_L}(\theta,\phi)$ with $L=2$. The combinations of the states $|M_L\rangle$ which belong to Γ_3 and Γ_5 will be denoted by

$$\begin{aligned}\Gamma_3:&\quad u_1=|0\rangle,\quad u_2=2^{-1/2}(|2\rangle+|-2\rangle),\\ \Gamma_5:&\quad v_+=|-1\rangle,\\ &\quad v_0=2^{-1/2}(|2\rangle-|-2\rangle);\quad v_-=-|1\rangle.\end{aligned}\tag{55}$$

We find the splitting between the Γ_3 and Γ_5 levels by taking the expectation value of H' for one state from each irreducible representation. They are $-18a/5$ for Γ_3 and $12a/5$ for Γ_5. Thus, the separation between the levels is $\Delta=6a$. We now obtain the spin Hamiltonian for the orbital Γ_3 level. We recall that $S=2$ so that the spin quantum number takes the values $M_S=2,1,0,-1,-2$. The orbital Γ_3 level is twofold degenerate. Thus, the spin Hamiltonian is tenfold dimensional. It is, however, preferable to write the complete 25×25 matrix which includes also the Γ_5 orbital levels. We note that the matrix elements of $\mathbf{L}$ in the states u_1 and u_2 vanish. We need to obtain the matrix elements of H'' in the 25 states ${}^5\Gamma_3+{}^5\Gamma_5$ written in the order $|u_i,M_S\rangle$, $|v_\kappa,M_S\rangle$; $i=1,2$; $\kappa=+,0,-$; $M_S=2,1,0,-1,-2$. The appropriate spin Hamiltonian is

$$H_S=\begin{pmatrix}2\mu_B\mathbf{B}\cdot\mathbf{S} & \mathbf{U}\cdot(\lambda\mathbf{S}+\mu_B\mathbf{B})\\ \mathbf{U}^\dagger\cdot(\lambda\mathbf{S}+\mu_B\mathbf{B}) & \Delta-\lambda\mathbf{I}\cdot\mathbf{S}+\mu_B\mathbf{B}\cdot(2\mathbf{S}-\mathbf{I})\end{pmatrix},\tag{56}$$

where $\mathbf{U}$ is a 2×3 matrix whose elements are those of the angular momentum operator between the states u_i and v_κ ($i=1,2$; $\kappa=+,0,-$) and $\mathbf{I}$ is the matrix whose elements are $-\langle v_{\kappa'}|\mathbf{L}|v_\kappa\rangle$. Its components are identical to the angular momentum matrices for $I=1$. The matrix $\mathbf{U}$ is

$$\mathbf{U}=\begin{pmatrix}(\frac{3}{2})^{1/2}(\hat{\mathbf{x}}-\hat{\imath}\hat{\mathbf{y}}) & 0 & -(\frac{3}{2})^{1/2}(\hat{\mathbf{x}}+\hat{\imath}\hat{\mathbf{y}})\\ 2^{-1/2}(\hat{\mathbf{x}}+\hat{\imath}\hat{\mathbf{y}}) & 2\hat{\mathbf{z}} & -2^{-1/2}(\hat{\mathbf{x}}-i\,\hat{\mathbf{y}})\end{pmatrix}.\tag{57}$$

Thus the Hamiltonian is a supermatrix with diagonal blocks of dimensions 10 and 15 and off-diagonal rectangular matrices.

The Schrödinger equation

$$H_S\begin{pmatrix}\psi_3\\ \psi_5\end{pmatrix}=E\begin{pmatrix}\psi_3\\ \psi_5\end{pmatrix},\tag{58}$$

where ψ_3 and ψ_5 have 10 and 15 components, respectively, is equivalent to the coupled relations

$$2\mu_B\mathbf{B}\cdot\mathbf{S}\psi_3+\mathbf{U}\cdot(\lambda\mathbf{S}+\mu_B\mathbf{B})\psi_5=E\psi_3\tag{59}$$

and

$$\mathbf{U}^\dagger\cdot(\lambda\mathbf{S}+\mu_B\mathbf{B})\psi_3+[\Delta+2\mu_B\mathbf{B}\cdot\mathbf{S}-\mathbf{I}\cdot(\lambda\mathbf{S}+\mu_B\mathbf{B})]\psi_5=E\psi_5.\tag{60}$$

Eliminating ψ_5, we find

$$\begin{aligned}&[2\mu_B\mathbf{B}\cdot\mathbf{S}-(\lambda\mathbf{S}+\mu_B\mathbf{B})\cdot\mathbf{U}\\ &\times\{\Delta-E+2\mu_B\mathbf{B}\cdot\mathbf{S}-\mathbf{I}\cdot(\lambda\mathbf{S}+\mu_B\mathbf{B})\}^{-1}\mathbf{U}^\dagger\cdot(\lambda\mathbf{S}\\ &+\mu_B\mathbf{B})]\psi_3=E\psi_3.\end{aligned}\tag{61}$$

We now solve this equation by iteration assuming $\Delta \gg \overline{H''}$. We obtain

$$\begin{aligned}&\Big(-\frac{\lambda^2}{\Delta}\mathbf{S}\cdot\mathbf{U}\mathbf{U}^\dagger\cdot\mathbf{S}+2\mu_B\mathbf{B}\cdot\mathbf{S}\\ &-\frac{\lambda\mu_B}{\Delta}(\mathbf{S}\cdot\mathbf{U}\mathbf{U}^\dagger\cdot\mathbf{B}+\mathbf{B}\cdot\mathbf{U}\mathbf{U}^\dagger\cdot\mathbf{S})\\ &-\frac{\mu_B^2}{\Delta}\mathbf{B}\cdot\mathbf{U}\mathbf{U}^\dagger\cdot\mathbf{B}\Big)\psi_3=E\psi_3.\end{aligned}\tag{62}$$

Consider first $H_{S,3}^{(0)}$, the term in $H_{S,3}$ independent of $\mathbf{B}$. It is

$$\begin{aligned}H_{S,3}^{(0)}&=-\frac{\lambda^2}{\Delta}\mathbf{S}\cdot\mathbf{U}\mathbf{U}^\dagger\cdot\mathbf{S}\\ &=-\frac{\lambda^2}{\Delta}\begin{pmatrix}3(\mathbf{S}\cdot\mathbf{S}-S_z^2) & \sqrt{3}(S_x^2-S_y^2)\\ \sqrt{3}(S_x^2-S_y^2) & \mathbf{S}\cdot\mathbf{S}+3S_z^2\end{pmatrix}.\end{aligned}\tag{63}$$

The terms linear and quadratic in $\mathbf{B}$ are

$$H_{S,3}^{(1)}=2\mu_B\begin{pmatrix}\mathbf{B}\cdot\mathbf{S}-\frac{3\lambda}{\Delta}(\mathbf{S}\cdot\mathbf{B}-S_zB_z) & -\sqrt{3}\frac{\lambda}{\Delta}(S_xB_x-S_yB_y)\\ -\sqrt{3}\,\frac{\lambda}{\Delta}(S_xB_x-S_yB_y) & \mathbf{B}\cdot\mathbf{S}-\frac{\lambda}{\Delta}(\mathbf{S}\cdot\mathbf{B}+3S_zB_z)\end{pmatrix}\tag{64}$$

and

$$H_{S,3}^{(2)}=-\frac{\mu_B^2}{\Delta}\begin{pmatrix} 3(B^2-B_z^2) & \sqrt{3}(B_x^2-B_y^2) \\ \sqrt{3}(B_x^2-B_y^2) & B^2+3B_z^2 \end{pmatrix}, \qquad (65)$$

respectively. The operator $H_{S,3}^{(0)}$ can be rewritten in the form

$$H_{S,3}^{(0)}=-\frac{2\lambda^2}{\Delta}S(S+1)-\frac{\lambda^2}{\Delta}\Omega, \qquad (66)$$

where

$$\Omega=(\mathbf{S}\cdot\mathbf{S}-3S_z^2)\sigma_z+\tfrac{1}{2}\sqrt{3}(S_+^2+S_-^2)\sigma_x. \qquad (67)$$

We denote the states $|u_1,M_S\rangle$ by $|M_S^+\rangle$ and $|u_2,M_S\rangle$ by $|M_S^-\rangle$ and $\sigma_z u_1 = u_1$, $\sigma_z u_2 = -u_2$, $\sigma_x u_1 = u_2$, $\sigma_x u_2 = u_1$. This means that the state $|2^+\rangle$ has nonvanishing matrix elements only with itself and with $|0^-\rangle$. But $|0^-\rangle$ is connected by Ω with itself and with $|-2^+\rangle$. Thus, after similar considerations, the matrix of Ω can be written in block form with two submatrices of dimension three and two of dimension two. The diagonalization of Ω is a trivial exercise. The eigenvectors and eigenvalues of $H_{S,3}^{(0)}$ are

$$-\frac{24\lambda^2}{\Delta}:\quad \frac{1}{2}(|2^-\rangle+\sqrt{2}|0^+\rangle+|-2^-\rangle) \quad (\Gamma_1),$$

$$-\frac{18\lambda^2}{\Delta}:\quad \begin{cases} \frac{1}{\sqrt{2}}(|2^-\rangle-|-2^-\rangle) \\ \frac{1}{2}(\sqrt{3}|1^+\rangle+|-1^-\rangle) \quad (\Gamma_4), \\ \frac{1}{2}(|1^-\rangle+\sqrt{3}|-1^+\rangle) \end{cases}$$

$$-\frac{12\lambda^2}{\Delta}:\quad \begin{cases} \frac{1}{2}(|2^+\rangle+\sqrt{2}|0^-\rangle+|-2^+\rangle) \\ \frac{1}{2}(|2^-\rangle-\sqrt{2}|0^+\rangle+|-2^-\rangle) \end{cases} \quad (\Gamma_3),$$

$$-\frac{6\lambda^2}{\Delta}:\quad \begin{cases} \frac{1}{\sqrt{2}}(|2^+\rangle-|-2^+\rangle) \\ \frac{1}{2}(|1^+\rangle-\sqrt{3}|-1^-\rangle) \quad (\Gamma_5), \\ \frac{1}{2}(\sqrt{3}|1^-\rangle-|-1^+\rangle) \end{cases}$$

$$0:\quad \frac{1}{2}(|2^+\rangle-\sqrt{2}|0^-\rangle+|-2^+\rangle) \quad (\Gamma_2).$$

These ten states belong to $\Gamma_3\times(\Gamma_3\oplus\Gamma_5) = \Gamma_1\oplus\Gamma_2\oplus\Gamma_3\oplus\Gamma_4\oplus\Gamma_5$ since the states $|M_S\rangle$ belong to $D_+^{(2)} \to \Gamma_3\oplus\Gamma_5$.

Using these wave functions, assuming that $6\lambda^2/\Delta \gg \mu_B B$, we regard $H_{S,3}^{(1)}$ and $H_{S,3}^{(2)}$ as perturbations. We obtain, in this way, the different energy eigenvalues. We quote here only the value for the lowest level, Γ_1:

$$E_{\Gamma_1}=-\frac{24\lambda^2}{\Delta}-\frac{4\Delta(\mu_B B)^2}{3\lambda^2}-\frac{2(\mu_B B)^2}{\Delta}.$$

The remaining eigenvalues are obtained in a straightforward way. Thus, at 0 K, the magnetic susceptibility is

$$\chi\cong\frac{N\mu_B^2}{3}\frac{8\Delta}{\lambda^2}. \qquad (68)$$

For Fe^{2+} in CdTe, this is equivalent to $\chi = N\mu_B^2/3k_B T_{\text{eff}}$ with $T_{\text{eff}}=0.6$ K. A more accurate calculation gives

$$\chi=\frac{N\mu_B^2}{3}\frac{8\Delta}{\lambda^2}\left(1+\frac{13|\lambda|}{\Delta}+100\frac{\lambda^2}{\Delta^2}+\cdots\right). \qquad (69)$$

3.6 Co^{2+} in a Tetrahedral Field

As a second example we consider Co^{2+} in a T_d environment. The ground-state configuration is $(3d)^7$ so that, according to Hund's rules, the lowest term is ^4F. An orbital F state generates the representation $D_+^{(3)}$ of SO(3) which reduces to $\Gamma_2\oplus\Gamma_4\oplus\Gamma_5$ of T_d. The crystal potential is

$$H'=a[\tfrac{1}{8}(L_+^2+L_-^2)^2+\tfrac{3}{2}L_z^4-12L_z^2-12] \qquad (70)$$

with $a>0$. The reduction of $D_+^{(3)}$ can be carried out in a variety of ways. The most convenient one for our purposes is to note that the states $|M_L\rangle$ with $M_L=3,2,1,0,-1,-2,-3$ transform according to the representation $D_+^{(3)}$ of O(3). Thus, they transform as the f levels of a single electron except that the parity of $|M_L\rangle$ is positive since they are formed by combinations of $3d$ states.

The reduction of $D_+^{(3)}$ is given by the following states:

$$\beta=2^{-1/2}(|2\rangle-|-2\rangle),\quad \Gamma_2$$

$$\left.\begin{array}{l}\epsilon_+=(\tfrac{5}{8})^{1/2}|-1\rangle-(\tfrac{3}{8})^{1/2}|3\rangle,\\ \epsilon_0=2^{-1/2}(|2\rangle+|-2\rangle),\\ \epsilon_-=(\tfrac{5}{8})^{1/2}|1\rangle-(\tfrac{3}{8})^{1/2}|-3\rangle,\end{array}\right\}\Gamma_5$$

$$\left.\begin{array}{l}\delta_+=-(\tfrac{3}{8})^{1/2}|1-(\tfrac{5}{8})^{1/2}|-3,\\ \delta_0=|0,\\ \delta_-=-(\tfrac{3}{8})^{1/2}|-1-(\tfrac{5}{8})^{1/2}|3;\end{array}\right\}\Gamma_4$$

The energy eigenvalues can now be calculated using these wave functions and the expression for H'. We find

$$E_{\Gamma_5}-E_{\Gamma_2}=\Delta=30a, \tag{71}$$

and

$$E_{\Gamma_4}-E_{\Gamma_2}=9\Delta/5=54a. \tag{72}$$

Since here the orbital ground state is nondegenerate, the spin Hamiltonian for the lowest state $\Gamma_2\times\Gamma_8=\Gamma_8$ is simply given by Eqs. (49) and (50) with

$$D_{ij}=\lambda^2\sum_{nr}{}' \frac{\langle 0|L_i|nr\rangle\langle nr|L_j|0\rangle}{E_0-E_n}. \tag{73}$$

The sum over nr is over all states other than $|0\rangle$ including all degenerate orbital states. But D_{ij} is a second-rank tensor with symmetry T_d so that $D_{xx}=D_{yy}=D_{zz}=D$, and $D_{ij}=0$ if $i\neq j$. Thus

$$D=\lambda^2\sum_{nr}{}' \frac{|\langle nr|L_z|0\rangle|^2}{E_0-E_n}. \tag{74}$$

The ground (orbital) state is β and $L_z|\beta\rangle=2\epsilon_0$ so that $\langle\epsilon_0|L_z|\beta\rangle=2$ and, neglecting higher terms,

$$D=-4\lambda^2/\Delta. \tag{75}$$

Thus

$$H_S=-\frac{15\lambda^2}{\Delta}+g\mu_B\mathbf{B}\cdot\mathbf{S}-\frac{4\mu_B^2}{\Delta}B^2 \tag{76}$$

with

$$g=2-8\lambda/\Delta. \tag{77}$$

These examples suffice to explain the procedures for the calculation of the energy eigenvalues and g factors of the low-lying levels of magnetic ions in crystals. The analysis is often simplified by introducing the operator

$$\mathbf{F}=\mathbf{M}+\mathbf{S}, \tag{78}$$

where $\mathbf{M}$ is defined in Eq. (46). The operator $\mathbf{M}$ has the same commutation rules as any angular momentum operator and the common eigenvectors of $\mathbf{F}\cdot\mathbf{F}$ and F_z are found by the usual rules of composition of angular momenta.

4. ENERGY-LEVEL SPLITTINGS IN CRYSTALS UNDER EXTERNAL STRESS

4.1 Deformation Potentials of Band Extrema in Semiconductors

We consider the energy shifts under stress of orbitally nondegenerate states of tetrahedrally coordinated elemental and compound semiconductors. These considerations are applicable to a large number of similar materials. We recall that, by virtue of Kramers's theorem, the Bloch states of electrons in a crystal always possess an evenfold degeneracy. The main consequence of this theorem is that a state at wave vector **k** is degenerate with one at $-\mathbf{k}$, their spin states being orthogonal. However, if the crystal has a center of inversion there is a spin degeneracy for each value of **k**. Thus, in Si and Ge the Bloch states have twofold and fourfold degeneracies at each point in the Brillouin zone. Application of stress cannot remove the twofold degeneracy but will, in general, split the quadruplet levels into two doublets.

We first discuss the energy shifts of the electronic levels at the extrema of the conduction band of the elemental group-IV semiconductors. The absolute minima of the conduction band occur along Δ in Si and at the L point in Ge, i.e., along the ⟨100⟩ and ⟨111⟩ directions. The minimum of the conduction band of Ge is at the Brillouin zone edge along ⟨111⟩. There are local minima at $\Gamma(\mathbf{k}=0)$ whose classification is Γ_6^- in Si and Γ_7^- in Ge. In both cases we rewrite the deformation potential in the form

$$\begin{aligned}\sum_{ij} D_{ij}\epsilon_{ij}=\frac{1}{3}\Big(\sum_{i} D_{ii}\Big)\Big(\sum_{j} \epsilon_{jj}\Big)+\frac{1}{6}[(2D_{zz}&\\ -D_{xx}-D_{yy})(2\epsilon_{zz}-\epsilon_{xx}-\epsilon_{yy})&\\ +3(D_{xx}-D_{yy})(\epsilon_{xx}-\epsilon_{yy})]&\\ +2\sum_{i<j} D_{ij}\epsilon_{ij},& \qquad (79)\end{aligned}$$

where x, y, z denote the cubic axes. The quantity $\Sigma_i D_{ii}$ belongs to the totally symmetric representation Γ_1^+ of O_h, $2D_{zz}-D_{xx}-D_{yy}$, and $\sqrt{3}(D_{xx}-D_{yy})$ generate Γ_3^+ and D_{yz}, D_{zx}, D_{xy} generate Γ_5^+. Using results from the theory of groups, the Γ_6^- and Γ_7^- levels shift by

$$\Delta E(\Gamma_{6,7})=\tfrac{1}{3}\Big\langle\Gamma_{6,7}\Big|\sum_{i} D_{ii}\Big|\Gamma_{6,7}\Big\rangle\sum_{j}\epsilon_{jj}, \qquad (80)$$

i.e., proportionally to the dilatation.

The conduction-band edge of Si is a Δ_1 state. The group of Δ, often designated by this symbol, is C_{4v} and the decomposition of the deformation potential operator appropriate for this case is

$$\begin{aligned}\sum_{ij} D_{ij}\epsilon_{ij}=\tfrac{1}{2}(D_{xx}+D_{yy})(\epsilon_{xx}+\epsilon_{yy})+D_{zz}\epsilon_{zz}&\\ +\tfrac{1}{2}(D_{xx}-D_{yy})(\epsilon_{xx}-\epsilon_{yy})&\\ +2D_{yz}\epsilon_{yz}+2D_{zx}\epsilon_{zx}+2D_{xy}\epsilon_{xy}.& \qquad (81)\end{aligned}$$

Here $D_{xx}+D_{yy}$ and D_{zz} belong to Δ_1, D_{xy} to Δ_2, $D_{xx}-D_{yy}$ to Δ_2', and D_{yz} and D_{zx} to Δ_5. [Note that Δ_1, Δ_1', Δ_2, Δ_2', and Δ_5 correspond to Γ_1, Γ_2, Γ_4, Γ_3, and Γ_5, respectively, in the notation of Koster *et al.* (1966)]. We immediately conclude that the energy shift of a specific Δ_1 is given by

$$\begin{aligned}\Delta E=\Big\langle\Delta_1\Big|\sum_{ij} D_{ij}\epsilon_{ij}\Big|\Delta_1\Big\rangle&\\ =\Xi_d(\epsilon_{xx}+\epsilon_{yy})+(\Xi_d+\Xi_u)\epsilon_{zz},& \qquad (82)\end{aligned}$$

where

$$\Xi_d=\tfrac{1}{2}\langle\Delta_1|D_{xx}+D_{yy}|\Delta_1\rangle \qquad (83)$$

and

$$\Xi_u=\tfrac{1}{2}\langle\Delta_1|2D_{zz}-D_{xx}-D_{yy}|\Delta_1\rangle \qquad (84)$$

are called the dilatational and shear deformation potential constants, respectively.

A similar treatment can be given for the $L_6^+(\Gamma_4^+)$ minimum of the lowest conduction band of germanium. Here, the deformation potential is best expressed in terms of invariants in which the $\hat{\zeta}$ axis is the axis of highest rotational symmetry of the group of L, namely [111]. It is sometimes convenient, as we do here, to select $\hat{\xi}$ and $\hat{\eta}$ parallel to [11$\bar{2}$] and [$\bar{1}$10], respectively. Then

$$\begin{aligned}\sum_{ij} D_{ij}\epsilon_{ij}=\tfrac{1}{2}(D_{\xi\xi}+D_{\eta\eta})(\epsilon_{\xi\xi}+\epsilon_{\eta\eta})+D_{\zeta\zeta}\epsilon_{\zeta\zeta}&\\ +\tfrac{1}{2}(D_{\xi\xi}-D_{\eta\eta})(\epsilon_{\xi\xi}-\epsilon_{\eta\eta})&\\ +2D_{\xi\eta}\epsilon_{\xi\eta}&\\ +2D_{\zeta\xi}\epsilon_{\zeta\xi}+2D_{\zeta\eta}\epsilon_{\zeta\eta}.& \qquad (85)\end{aligned}$$

Now $D_{\xi\xi}+D_{\eta\eta}$ and $D_{\zeta\zeta}$ belong to L_1 while the pairs $(D_{\xi\xi}-D_{\eta\eta}, D_{\xi\eta})$ and $(D_{\zeta\xi}, D_{\zeta\eta})$ generate L_3. Using the orthogonality theorem of the theory of group representations and time-reversal symmetry, we find

$$\begin{aligned}\Big\langle L_6^+\Big|\sum_{ij} D_{ij}\epsilon_{ij}\Big|L_6^+\Big\rangle&\\ =\Xi_d(\epsilon_{\xi\xi}+\epsilon_{\eta\eta})+(\Xi_d+\Xi_u)\epsilon_{\zeta\zeta}&\\ =(\Xi_d+\tfrac{1}{3}\Xi_u)(\epsilon_{xx}+\epsilon_{yy}+\epsilon_{zz})&\\ +\tfrac{2}{3}\Xi_u(\epsilon_{yz}+\epsilon_{zx}+\epsilon_{xy}).& \qquad (86)\end{aligned}$$

Here

$$\Xi_d=\tfrac{1}{2}\langle L_6^+|D_{\xi\xi}+D_{\eta\eta}|L_6^+\rangle \qquad (87)$$

and

$$\Xi_u=\tfrac{1}{2}\langle L_6^+|2D_{\zeta\zeta}-D_{\xi\xi}-D_{\eta\eta}|D_6^+\rangle. \qquad (88)$$

As an example in which the levels are orbitally degenerate, we consider here the deformation potential theory for the Γ_5^+ (Γ_{25}') states at the top of the valence band of elemental semiconductors having the diamond structure. We disregard the spin of the electron for the time being and designate the three degenerate states associated with the top of the valence band by the symbols ϵ_1, ϵ_2, ϵ_3. They transform under the operations of O_h as yz, zx, and xy where x, y, z are the components of a vector referred to the cubic axes. The changes in energy upon deformation are obtained diagonalizing the 3×3 matrix

$$\sum_{ij}\epsilon_{ij}\langle\epsilon_\mu|D_{ij}|\epsilon_\nu\rangle,$$

$\mu,\nu=1,2,3$. We can rewrite the quantity $\Sigma_{ij}\,\epsilon_{ij}D_{ij}$ as a linear combination of

$$(\epsilon_{xx}+\epsilon_{yy}+\epsilon_{zz})(D_{xx}+D_{yy}+D_{zz}),$$

$$(2\epsilon_{zz}-\epsilon_{xx}-\epsilon_{yy})(2D_{zz}-D_{xx}-D_{yy})$$
$$+3(\epsilon_{xx}-\epsilon_{yy})(D_{xx}-D_{yy}),$$

and

$$\epsilon_{yz}D_{yz}+\epsilon_{zx}D_{zx}+\epsilon_{xy}D_{xy}.$$

We recall that $D_{xx}+D_{yy}+D_{zz}$ belongs to the totally symmetric representation Γ_1^+ of the cubic group O_h, that $2D_{zz}-D_{xx}-D_{yy}$ and $\sqrt{3}\,(D_{xx}-D_{yy})$ belong to $\Gamma_3^+\,(\Gamma_{12})$, and D_{yz}, D_{zx}, D_{xy} to $\Gamma_5^+\,(\Gamma_{25}')$. $D_{xx}+D_{yy}+D_{zz}$ has only diagonal matrix elements in the basis $\{\epsilon_\mu\}$ $(\mu=1,2,3)$ and all its elements are equal. The matrices associated with $2D_{zz}-D_{xx}-D_{yy}$ and $\sqrt{3}\,(D_{xx}-D_{yy})$ are proportional to $2I_z^2-I_x^2-I_y^2$ and $\sqrt{3}(I_x^2-I_y^2)$, respectively, where

$$I_x=\begin{pmatrix}0&0&0\\0&0&-i\\0&i&0\end{pmatrix},\quad I_y=\begin{pmatrix}0&0&i\\0&0&0\\-i&0&0\end{pmatrix},$$

$$I_z=\begin{pmatrix}0&-i&0\\i&0&0\\0&0&0\end{pmatrix}$$

are the components of the angular momentum operator for $I=1$ on the cubic axes. In the same way the matrix elements of D_{yz}, D_{zx}, and D_{xy} in the same basis are proportional to $\{I_y,I_z\}$, $\{I_z,I_x\}$, and $\{I_x,I_y\}$, respectively. Thus, the operator $\Sigma_{ij}\,\epsilon_{ij}D_{ij}$ in the basis generating the $\Gamma_5^+\,(\Gamma_{25}')$ states has the form of the matrix

$$-a\sum_i\epsilon_{ii}-3b\sum_i\epsilon_{ii}(I_i^2-\tfrac{1}{3}I^2)$$
$$-\sqrt{3}d\sum_{i<j}\epsilon_{ij}\{I_i,I_j\}. \tag{89}$$

We consider now the effect of the spin-orbit interaction. The deformation potentials D_{ij} depend, of course, on the spin of the electron. However, in most circumstances this correction is negligible so that we can take as the deformation potential contribution to the effective second-order Hamiltonian the matrix (89) multiplied (in the sense of the direct product of matrices) by the unit 2×2 matrix. However, the spin-orbit coupling splits the sixfold degeneracy of the $\epsilon_\mu\alpha$, $\epsilon_\mu\beta$ states $(\mu=1,2,3)$ into a Γ_8^+ quadruplet and a Γ_7^+ doublet (α, β designate the two orthogonal spin states, taken, to fix the ideas, so that σ_z is diagonal, $\hat{\mathbf{z}}$ being one of the cubic axes).

The matrix elements of the spin-orbit interaction in the representation $\epsilon_\mu\chi_\pm$ $(\mu=1,2,3)$ where $\chi_+=\alpha$ and $\chi_-=\beta$ are

$$\frac{\hbar}{4m^2c^2}\langle\epsilon_{\mu'}|\nabla U\times\mathbf{p}|\epsilon_\mu\rangle\cdot\langle\chi_{\sigma'}|\boldsymbol{\sigma}|\chi_\sigma\rangle,$$

where $U(\mathbf{x})$ is the crystal potential. Now the components of the pseudovector $\nabla U\times\mathbf{p}$ generate the $\Gamma_4^+\,(\Gamma_{15}')$ representation of O_h so that its matrix representation is simply a constant times $\mathbf{I}$. Thus, we write the spin-orbit interaction in the form

$$\tfrac{1}{3}\Delta\mathbf{I}\cdot\boldsymbol{\sigma}.$$

The Hamiltonian matrix for the six $\Gamma_5^+\times\Gamma_6^+$ states for $\mathbf{k}$ near the center of the Brillouin zone and in the presence of strain in the diamond structure is of the form

$$H=Ak^2-a\sum_i\epsilon_{ii}$$
$$-3\sum_i(Bk_i^2+b\epsilon_{ii})(I_i^2-\tfrac{1}{3}I^2)$$
$$-\sqrt{3}\sum_{i<j}(Dk_ik_j+d\epsilon_{ij})\{I_i,I_j\}+\tfrac{1}{3}\Delta\mathbf{I}\cdot\boldsymbol{\sigma}. \tag{90}$$

The calculation of the eigenvalues of H is simplified if we notice that H is invariant under the time-reversal operator, $\mathbf{I}\rightarrow-\mathbf{I}$, and $\boldsymbol{\sigma}\rightarrow-\boldsymbol{\sigma}$. Thus, the eigenvalues of H in Eq. (90) are doubly degenerate (Kramers's theorem).

The eigenvalues of this matrix to first order in $\boldsymbol{\epsilon}$ and second order in $\mathbf{k}$ are

$$-\tfrac{2}{3}\Delta+Ak^2-a\sum_i\epsilon_{ii} \tag{91}$$

and

$$E_\pm(\mathbf{k},\boldsymbol{\epsilon})=Ak^2-a\sum_i\epsilon_{ii}+\tfrac{1}{3}\Delta$$
$$\pm(R_k+R_{k\epsilon}+R_\epsilon)^{1/2} \tag{92}$$

with

$$R_k=B^2k^4+(D^2-3B^2)\sum_{i<j}k_i^2k_j^2, \tag{93}$$

$$R_{k\epsilon} = -Bbk^2 \sum_i \epsilon_{ii} + 3Bb \sum_i k_i^2 \epsilon_{ii} + 2Dd \sum_{i<j} k_i k_j \epsilon_{ij}, \qquad (94)$$

and

$$R_\epsilon = b^2 \sum_i \epsilon_{ii}^2 + \sum_{i<j} (d^2 \epsilon_{ij}^2 - b^2 \epsilon_{ii} \epsilon_{jj}). \qquad (95)$$

In the elemental semiconductors Si and Ge, the spin-orbit splitting of the $\Gamma_{25}' \times \Gamma_6$ valence band is positive so that, in the absence of stress, the top of the valence band is a quadruplet Γ_8^+. The split-off band, Γ_7^+, lies at an energy Δ below the fourfold level. In the presence of stress the Γ_7^+ shifts by an amount proportional to the dilatation. A similar shift occurs for the Γ_8^+ level but, in addition, it splits into two doublets under shear stress. The deformation potential operator of the Γ_8^+ level is

$$D = -a \sum_i \epsilon_{ii} - b \sum_i \epsilon_{ii} (J_i^2 - \tfrac{1}{3} \mathbf{J} \cdot \mathbf{J}) - (d/\sqrt{3}) \sum_{i<j} \epsilon_{ij} \{J_i, J_j\}. \qquad (96)$$

Simple examples for uniaxial stress along three high-symmetry directions are given below. The components of the strain tensor resulting from the application of a stress generated by a force **F** along [001] are $\epsilon_{xx} = \epsilon_{yy} = s_{12}T, \epsilon_{zz} = s_{11}T, \epsilon_{yz} = \epsilon_{zx} = \epsilon_{xy} = 0$. Here T is the force per unit area, taken positive for extension and negative for compression. The dilatation produces a shift as stated above but the associated shear splits the levels associated with $J_z = \pm 1/2$ from those for which $J_z = \pm 3/2$ by $\Delta E_{100} = 2b(s_{11} - s_{12})T$, the $J_z = \pm 1/2$ level lying above $J_z = \pm 3/2$ if $bT(s_{11} - s_{12}) > 0$. For a force along [111], $\epsilon_{xx} = \epsilon_{yy} = \epsilon_{zz} = (s_{11} + 2s_{12})T/3$ and $\epsilon_{yz} = \epsilon_{zx} = \epsilon_{xy} = Ts_{44}/6$, the level corresponding to $J_n = \pm 1/2$ lying above $J_n = \pm 3/2$ if $Tds_{44} > 0$. Here J_n is the projection of **J** on the [111] direction.

For a force directed parallel to [110], $\epsilon_{xx} = \epsilon_{yy} = T(s_{11} + s_{12})/2, \epsilon_{zz} = s_{12}T, \epsilon_{yz} = \epsilon_{zx} = 0, \epsilon_{xy} = Ts_{44}/4$.

In general, the splitting of the Γ_8^+ level under uniaxial stress applied along an arbitrary direction $\hat{\mathbf{e}}$ is given by

$$(\Delta E_e)^2 = (\Delta E_{100})^2 + 3(\Delta E_{111}^2 - \Delta E_{100}^2) f(\hat{\mathbf{e}}),$$

where

$$f(\hat{\mathbf{e}}) = e_y^2 e_z^2 + e_z^2 e_x^2 + e_x^2 e_y^2.$$

The quantity $f(\hat{\mathbf{e}})$ has extremal values of 1/3, 1/4, and 0 along ⟨111⟩, ⟨110⟩, and ⟨100⟩, respectively, the first being an absolute maximum and the third an absolute minimum.

These considerations are applicable not only to the band states but also to the localized levels of electrons or holes in semiconductors. Under such conditions the energy levels are classified according to the irreducible representations of the symmetry group of the site of the localized state. The results described above hold, for example, for donor and acceptor states except that there is no question of a wave-vector dependence. The single-electron or -hole states of donors and acceptors in elemental semiconductors are classified according to the irreducible representations of the double group $\overline{T}_d$ but the results are the same as in Eq. (96). Of course, the parameters a, b, and d are different for different Γ_8 levels. However, for the ground states of shallow donors and acceptors, these parameters should be approximately equal to those of the band levels.

4.2 Relative Intensities of Stress-Induced Components of an Optical Transition

Spectroscopic investigations provide information on the symmetry assignments of the quantum-mechanical states associated with the initial and final states for given optical transitions. The energies of the transitions allow us to obtain band parameters and, in the presence of strain, values for the deformation potential constants.

The magnitude and sign of the deformation potentials are obtained experimentally in conjunction with theoretical studies of the symmetry characteristics of the initial and final states of the stress-induced lines associated with specific optical transitions carried out with incident radiation of varied polarization. But this is often insufficient to obtain unambiguous results. In some cases it is useful to compare the measured relative intensities of the stress-induced components of the absorption lines with the predictions of a symmetry analysis. The relative intensities of these components can be deduced, in the absence of significant hybridization with adjacent levels, by calculating the matrix elements of the electric dipole moment **d** between initial and final states. These calculations provide both the selection rules and expressions for the relative intensities, either directly or as functions of at most two adjustable parameters.

We recall that d_x, d_y, and d_z, the components of **d** along the cubic axes, generate the Γ_5 representation of T_d and the Γ_4^- representation of O_h. For transitions between Γ_8 states, d_x, d_y, d_z are 4×4 matrices. We consider first transitions between localized states of levels where the symmetry is $\overline{T}_d$. The matrix elements of d_x, d_y, d_z for a transition between two Γ_8 levels are proportional to suitable combinations of products of the components of **J**, the angular momentum operator matrix for $j=3/2$, which transform in the same way as d_x, d_y, d_z under the operations of T_d. The two sets of matrices

$$U_x=\{J_y,J_z\},\quad U_y=\{J_z,J_x\},\quad U_z=\{J_x,J_y\} \tag{97}$$

and

$$V_x=\{J_x,J_y^2-J_z^2\},\quad V_y=\{J_y,J_z^2-J_x^2\},$$
$$V_z=\{J_z,J_x^2-J_y^2\} \tag{98}$$

transform as Γ_5. Thus, the components of **d** are linear combinations of **U** and **V** with suitable, complex valued, coefficients. It is important to recognize that the shorthand notation for these operators does not imply that they are vectors under O(3). For transitions between Γ_8 and Γ_6 or Γ_7 levels the matrix elements of **d** are proportional to

$$\begin{pmatrix} \hat{\mathbf{x}}-i\hat{\mathbf{y}} & 0 & \sqrt{3}(\hat{\mathbf{x}}+i\hat{\mathbf{y}}) & 2\hat{\mathbf{z}} \\ 2\hat{\mathbf{z}} & -\sqrt{3}(\hat{\mathbf{x}}-i\hat{\mathbf{y}}) & 0 & -(\hat{\mathbf{x}}+i\hat{\mathbf{y}}) \end{pmatrix} \tag{99}$$

for $\Gamma_8\rightarrow\Gamma_7$ transitions and to

$$\begin{pmatrix} -\sqrt{3}(\hat{\mathbf{x}}+i\hat{\mathbf{y}}) & 2\hat{\mathbf{z}} & \hat{\mathbf{x}}-i\hat{\mathbf{y}} & 0 \\ 0 & -(\hat{\mathbf{x}}+i\hat{\mathbf{y}}) & 2\hat{\mathbf{z}} & \sqrt{3}(\hat{\mathbf{x}}-i\hat{\mathbf{y}}) \end{pmatrix} \tag{100}$$

for $\Gamma_8\rightarrow\Gamma_6$ transitions.

These matrix elements between eigenvectors of J_z corresponding to the values 3/2, 1/2, −1/2, −3/2 for Γ_8 and 1/2, −1/2 for Γ_6 and Γ_7, are written in the above order.

For transitions between states whose symmetry is O_h a slightly different situation occurs. Here d_x, d_y, d_z transform as Γ_4^-. For transitions from a $\Gamma_8^\pm$ level to a $\Gamma_8^\mp$ level, the matrix elements of d_i $(i=x,y,z)$ are proportional to linear combinations of J_i and J_i^3 because the sets (J_x,J_y,J_z) and (J_x^3,J_y^3,J_z^3) transform as Γ_4^+. Note that for systems having a center of symmetry, dipole transitions can occur only between states of different parity. The matrix elements for electric dipole transitions from $\Gamma_8^\pm$ to $\Gamma_6^\mp$ are proportional to those in expression (100), while those for $\Gamma_8^\pm\rightarrow\Gamma_7^\mp$ are proportional to the quantities appearing in expression (99). For transitions between Γ_6 and Γ_7 states the matrices associated with electric dipole transitions are proportional to the components of the Pauli spin matrix $\boldsymbol{\sigma}$.

We emphasize that to obtain the relative intensities for a given set of stress-induced components of an absorption line, it is first required to obtain the energy eigenfunctions under stress and carry out the unitary transformation between the eigenstates of J_z and the actual eigenfunctions. These transformations are trivial to carry out for uniaxial stresses along high symmetry directions.

The energy eigenvalues as functions of **k** for states near the top of the valence band of the elemental semiconductors having the diamond structure have been given in Eq. (92). If Δ is large compared to the kinetic-energy terms, these levels are the eigenvalues of

$$H=Ak^2-B\sum_i k_i^2(J_i^2-\tfrac{1}{3}\mathbf{J}\cdot\mathbf{J})$$
$$-(D/\sqrt{3})\sum_{i<j} k_ik_j\{J_i,J_j\} \tag{101}$$

to which we add the deformation potential (96) in the presence of strain. For the zincblende semiconductors, lacking inversion symmetry, there is an additional term H_A, linear in **k**, namely,

$$H_A=(C/\sqrt{3})\sum_i k_iV_i, \tag{102}$$

where V_x, V_y, V_z are given by Eq. (98). However, the constant C is usually so small that for many purposes H_A can be neglected. It originates in second-order perturbation theory as a term proportional to the product of the perturbation $(\hbar/m)\mathbf{k}\cdot\mathbf{p}$ with the spin-orbit interaction so that it is expected to be small.

In general, application of uniaxial stress splits the $\Gamma_8(\Gamma_8^+)$ level at the top of the valence band of a zinc-blende (diamond) semiconductor into two Kramers doublets. The same occurs for the Γ_8 levels of an acceptor in tetrahedrally bonded semiconductors. For a force $\mathbf{F}$ applied along one of the cubic axes a Γ_8 level splits into Γ_6 and Γ_7 levels of $\overline{D}_{2d}$, the group D_{2d} characterizing the site symmetry after application of the stress. If the force $\mathbf{F}$ is applied along the [001] axis, and the deformation potential is expressed in the basis generated by the eigenvectors of J_z, namely, ϕ_μ $(\mu=3/2,1/2,-1/2,-3/2)$, the functions $\phi_{\pm 3/2}$ belong to $\Gamma_6(\overline{D}_{2d})$ while $\phi_{\pm 1/2}$ generate $\Gamma_7(\overline{D}_{2d})$. Under uniaxial stress resulting from an applied force directed along [111] the symmetry of the site becomes C_{3v} and the Γ_8 levels split into $\Gamma_4(\overline{C}_{3v})$ corresponding to $\hat{\zeta}\cdot\mathbf{J}=\mathbf{J}_\zeta = \pm 1/2$, where ζ is a unit vector parallel to the [111] direction, and $\Gamma_5(\overline{C}_{3v}) \oplus \Gamma_6(\overline{C}_{3v})$ associated with $J_\zeta = \pm\ 3/2$. Γ_5 and Γ_6 are one-dimensional irreducible representations which are each other's complex conjugate and, hence, correspond to states of equal energy eigenvalues. We represent this level by Γ_{5+6}.

For the analysis of the relative intensities of stress-induced components referred to above we use the following considerations. For transitions between Γ_8 levels the matrix elements of the i component of the electric dipole operator $(i=x,y,z)$ are linear combinations of the components of the quantities U_i and V_i, namely,

$$d_i = \frac{i}{\sqrt{3}} d_1 U_i - \frac{2}{\sqrt{3}} d_2 V_i, \quad i=x,y,z. \qquad (103)$$

We find the relative intensities of the stress-induced absorption lines by reading off the matrix elements and adding the square of the magnitudes of matrix elements of d_i between degenerate levels. The form of the matrices U_i and V_i permits the direct determination of the relative intensities for uniaxial stress along a cubic axis. The results of this analysis are given in Table 5 for incident radiation polarized parallel or perpendicular to the applied force. These results are accurate as long as hybridization with neighboring levels can be neglected. Otherwise the possibility of a transfer of intensity from one set of lines to another occurs and can also be incorporated into the formalism. Results are also displayed for transitions from a Γ_8 level to Γ_6 and Γ_7 states using the rectangular matrices in Eqs. (99) and (100). We note that the relative intensities can be expressed in terms of d_1/d_2, the ratio

Table 5. Relative intensities for stress-induced components of acceptors in tetrahedrally bonded semiconductors with $\mathbf{F}\|[001]$ and incident radiation polarized parallel to $\mathbf{E}$.[a]

Zero-stress transition $\overline{T}_d$	Stress-induced components $\overline{D}_{2d}$	Sum of the squared amplitudes for the stress-induced components		Relative intensities	
		$\mathbf{E}\perp(\hat{x},\hat{y})$	$\mathbf{E}\|\hat{z}$	$\mathbf{E}\perp(\hat{x},\hat{y})$	$E\|\hat{z}$
$\Gamma_8\rightarrow\Gamma_8$	$\Gamma_7(\pm\frac{1}{2})\rightarrow\Gamma_7(\pm\frac{1}{2})$	$6\|d_2\|^2$	0	$\frac{3}{8}u$	0
	$\Gamma_7(\pm\frac{1}{2})\rightarrow\Gamma_6(\pm\frac{3}{2})$	$2\|d_1+d_2\|^2$	$2\|d_1-2d_2\|^2$	$\frac{1}{2}-\frac{3}{8}u-\frac{1}{2}v$	$\frac{1}{2}+v$
	$\Gamma_6(\pm\frac{3}{2})\rightarrow\Gamma_7(\pm\frac{1}{2})$	$2\|d_1-d_2\|^2$	$2\|d_1+2d_2\|^2$	$\frac{1}{2}-\frac{3}{8}u+\frac{1}{2}v$	$\frac{1}{2}-v$
	$\Gamma_6(\pm\frac{3}{2})\rightarrow\Gamma6(\pm\frac{3}{2})$	$6\|d_2\|^2$	0	$\frac{3}{8}u$	0
$\Gamma_8\rightarrow\Gamma_6$	$\Gamma_7(\pm\frac{1}{2})\rightarrow\Gamma_6$	$2\|d_0\|^2$	$8\|d_0\|^2$	1	4
	$\Gamma_6(\pm\frac{3}{2})\rightarrow\Gamma_6$	$6\|d_0\|^2$	0	3	0
$\Gamma_8\rightarrow\Gamma_7$	$\Gamma_7(\pm\frac{1}{2})\rightarrow\Gamma_7$	$6\|d_0'\|^2$	0	3	0
	$\Gamma_6(\pm\frac{3}{2})\rightarrow\Gamma_7$	$2\|d_0'\|^2$	$8\|d_0'\|^2$	1	4

[a] The parameters d_1 and d_2 are defined in Eq. (103). d_0 and d_0' are the coefficients of the matrices (99) and (100) for $\Gamma_8 \rightarrow \Gamma_6$ and $\Gamma_8 \rightarrow \Gamma_7$ transitions, respectively.

Table 6. Relative intensities for stress-induced components of acceptors in tetrahedrally bonded semiconductors with **F**‖[111] and incident radiation polarized parallel to **E**.

Zero stress transition $\overline{T}_d$	Stress-induced components $\overline{C}_{3v}$	Sum of the squared amplitudes for the stress-induced components		Relative intensities	
		$\mathbf{E}\perp(\hat{\xi},\hat{\eta})$	$\mathbf{E}\parallel\hat{\zeta}$	$\mathbf{E}\perp(\hat{\xi},\hat{\eta})$	$E\parallel\hat{\zeta}$
$\Gamma_8\rightarrow\Gamma_8$	$\Gamma_4(\pm\frac{1}{2})\rightarrow\Gamma_4(\pm\frac{1}{2})$	$8\|d_2\|^2$	$2\|d_1\|^2$	$\frac{1}{2}u$	$\frac{1}{2}-\frac{1}{2}u$
	$\Gamma_4(\pm\frac{1}{2})\rightarrow\Gamma_{5+6}(\pm\frac{3}{2})$	$2\|d_1\|^2+4\|d_2\|^2$	0	$\frac{1}{2}-\frac{1}{4}u$	0
	$\Gamma_{5+6}(\pm\frac{3}{2})\rightarrow\Gamma_4(\pm\frac{1}{2})$	$2\|d_1\|^2+4\|d_2\|^2$	0	$\frac{1}{2}-\frac{1}{4}u$	0
	$\Gamma_{5+6}(\pm\frac{3}{2})\rightarrow\Gamma_{5+6}(\pm\frac{3}{2})$	0	$2\|d_1\|^2+16\|d_2\|^2$	0	$\frac{1}{2}+\frac{1}{2}u$
$\Gamma_8\rightarrow\Gamma_6$	$\Gamma_4(\pm\frac{1}{2})\rightarrow\Gamma_4$	$2\|d_0\|^2$	$8\|d_0\|^2$	1	4
	$\Gamma_{5+6}(\pm\frac{3}{2})\rightarrow\Gamma_4$	$6\|d_0\|^2$	0	3	0
$\Gamma_8\rightarrow\Gamma_7$	$\Gamma_4(\pm\frac{1}{2})\rightarrow\Gamma_4$	$2\|d_0'\|^2$	$8\|d_0'\|^2$	1	4
	$\Gamma_{5+6}(\pm\frac{3}{2})\rightarrow\Gamma_4$	$6\|d_0'\|^2$	0	3	0

of two complex numbers, and thus depends on two real parameters u and v defined by

$$u=[1+|d_1/2d_2|^2]^{-1},$$

$$v=-\frac{1}{4}\left(\frac{d_1}{d_2}+\frac{d_1^*}{d_2^*}\right)u.$$

To investigate the relative intensities of the stress-induced lines we transform U_i and V_i so that the matrix elements of **d** are referred to a coordinate system $\hat{\mathbf{z}}$, $\hat{\xi}$, $\hat{\zeta}$ in which $\hat{\zeta}$ is directed along [111]. We select $\hat{\xi}$ parallel to [11$\bar{2}$] and $\hat{\eta}$ along [$\bar{1}$10] and transform the Γ_8 states by means of the rule

$$\psi_\mu=\sum_{\mu'}\phi_{\mu'}D^{(3/2)}_{\mu'\mu}(\alpha,\beta,\gamma),$$

where α, β, and γ are the Euler angles of the rotation from $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$ to $\hat{\xi}$, $\hat{\eta}$, $\hat{\zeta}$ ($\alpha=\pi/4$, $\cos\beta=3^{-1/2}$, $\gamma=0$) and $D^{(3/2)}(\alpha,\beta,\gamma)$ is the rotation operator for angular momentum 3/2. The transformations are straightforward and the results are displayed in Table 6. We note that in this case the relative intensities depend only on the parameter u defined above.

5. CONCLUDING REMARKS

In this article we have discussed the splitting of electron levels of atoms due to the presence of external fields or to the interactions arising when they are embedded in a crystal. Topics such as the static and dynamic Jahn-Teller effects and the interaction with nuclei (hyperfine coupling) lie outside the scope of this article. Even in the area selected we have only been able to include the main relevant subjects and have illustrated them with a few significant examples.

The author is grateful to Professor A. K. Ramdas and to Dr. M. Villeret for their comments on the manuscript and for many useful discussions. Part of the examples presented here are the results of research supported by the National Science Foundation (Grants No. DMR-89-21717 and No. DMR-89-13706). The preparation of the manuscript was completed while the author was a guest of Scuola Normale Superiore (Pisa, Italy). The author is grateful to his colleagues in Pisa for the opportunity of working at the Scuola Normale. In particular, he wishes to express his appreciation to Professor F. Bassani, to Professor L. Foà, and to Professor L. A. Radicati for their kind hospitality.

GLOSSARY

Configuration: Atomic energy levels in the central-field approximation. For example, the configuration of Na in its lowest state is $(1s)^2(2s)^2(2p)^6(3s)$. (See definitions of *Term* and *Level*.)

Crystal Field: Electrostatic potential of an atom or ion in a crystal lattice.

Deformation Potential: Potential energy of a charge carrier in a crystal due to the application of external stress.

Hund's Rules: Rules for the determination of the lowest term and ground state of an atom or ion. (See text or Condon and Shortley, 1953).

Level: Energy eigenstate of an atom in a term with given orbital (L) and spin (S) angular momenta and the total angular momentum J determined using Hund's rules. (See definition of *Term* below.)

Operator Equivalent: The crystal-field potential expressed in terms of powers of angular momentum operators.

Quenching of the Orbital Angular Momentum: Expression describing the situation when the expectation value of the orbital angular momentum in the ground state of an atom vanishes as a consequence of its presence in a crystal.

Spin Hamiltonian: Operator describing the effect of the crystal potential and of an external magnetic field on an atom or ion in a crystal.

Term: Quantum-mechanical states resulting from the splitting of the configuration of an atom or ion due to the Coulomb repulsion between electrons prior to the consideration of the spin-orbit interaction and other relativistic corrections.

List of Works Cited

Abragam, A., Bleaney, B. (1970), *Electron Paramagnetic Resonance of Transition Ions*, Oxford, England: Clarendon.

Abragam, A., Pryce, M. H. L. (1951), *Proc. R. Soc. London Ser. A* **205**, 135–153.

Condon, E. U., Shortley, G. H. (1953), *The Theory of Atomic Spectra*, Cambridge, England: Cambridge University Press.

Heckmann, P. H., Träbert, E. (1989), *Introduction of the Spectroscopy of Atoms*, Amsterdam: North-Holland.

Hutchings, M. T. (1964), in: F. Seitz and D. Turnbull (Eds.), *Solid State Physics*, Vol. 16. New York: Academic, pp. 227–273.

Kittel, C. (1986), *Introduction to Solid State Physics*, 6th ed. New York: Wiley.

Koster, G. F., Dimmock, J. O., Wheeler, R. G., Statz, H. (1966), *Properties of the Thirty-Two Point Groups*, Cambridge, MA: MIT Press.

Pryce, M. H. L. (1950), *Phys. Rev.* **80**, 1107–1108.

Slack, G. A., Roberts, S., Vallin, J. T. (1969), *Phys. Rev.* **187**, 511–524.

Slater, J. C. (1929), *Phys. Rev.* **34**, 1293–1322.

Stevens, K. W. H. (1952), *Proc. Phys. Soc. London* **65**, 209–215.

Van Vleck, J. H. (1932), *The Theory of Electric and Magnetic Susceptibilities*, London: Oxford University Press.

Villeret, M., Rodriguez, S., Kartheuser, E. (1990a), *J. Appl. Phys.* **67**, 4221–4232.

Villeret, M., Rodriguez, S., Kartheuser, E. (1990b), *Phys. Rev. B* **41**, 10 028–10 042.

Zeiger, H. J., Pratt, G. W. (1973), *Magnetic Interactions in Solids*, Oxford, England: Clarendon.

Further Reading

Abragam, A., Bleaney, B. (1970), *Electron Paramagnetic Resonance of Transition Ions*, Oxford: Clarendon.

Bir, G. L., Pikus, G. E. (1974), *Symmetry and Strain-Induced Effects in Semiconductors*, New York: Wiley.

Condon, E. U., Shortley, G. H. (1953), *The Theory of Atomic Spectra*, Cambridge, England: Cambridge University Press.

Griffith, J. S. (1961), *Theory of Transition Metal Ions*, London: Cambridge University Press.

Heine, V. (1960), *Group Theory in Quantum Mechanics*, Oxford: Pergamon.

Van Vleck, J. H. (1932), *The Theory of Electric and Magnetic Susceptibilities*, London: Oxford University Press.

Weissbluth, M. (1978), *Atoms and Molecules*, London: Academic.

Zeiger, H. J., Pratt, G. W. (1973), *Magnetic Interactions in Solids*, Oxford, England: Clarendon.

ELECTRON PARAMAGNETIC RESONANCE

Richard G. Barnes, *Ames Laboratory, USDOE, and Department of Physics, Iowa State University, Ames, Iowa, U.S.A.*

	Introduction	1
1.	**Resonance Fundamentals**	2
1.1	Classical and Quantum-Mechanical Descriptions of Resonance	2
1.2	Relaxation Phenomena	3
1.3	Experimental Arrangements	4
2.	**Basic Features of EPR Spectra**	6
2.1	Introduction: The Spin Hamiltonian Concept	6
2.2	The *g* factor	6
2.3	Fine Structure	8
2.4	Hyperfine Structure	8
2.5	Resonance Line Shapes and Intensity	9
3.	**Sources of EPR Spectra: Their Interpretation and Applications**	10
3.1	Free Radicals and Triplet States	10
3.2	Transition-Metal Ions	12
3.3	Other Systems	14
3.3.1	Nonmetallic Solids	14
3.3.1.1	Amorphous Silicon (*a*-Si)	14
3.3.1.2	Radiation-Induced Defects	14
3.3.2	Metallic Solids	15
4.	**Spin Relaxation Phenomena and Applications**	16
4.1	Spin–Lattice Relaxation	16
4.2	Spin–Spin Relaxation	17
4.3	Exchange Interactions	17
4.4	Chemically Induced Dynamic Electron Polarization	18
5.	**Advanced Techniques and Applications**	19
5.1	Electron-Nuclear Double Resonance (ENDOR)	19
5.2	Optically Detected Magnetic Resonance (ODMR)	19
6.	**Summary**	20
	Glossary	21
	Works Cited	22
	Further Reading	23

INTRODUCTION

Electron paramagnetic resonance or EPR (also frequently called electron spin resonance, ESR) is a radiofrequency spectroscopy in which transitions among the energy levels of a system of weakly coupled *electronic* magnetic dipole moments in a magnetic field are detected. Such *paramagnetic* electronic magnetic moments occur in solids, liquids, and gases and may belong to conduction electrons in metals, impurity ions in both insulators and metals, organic free radicals, and radiation-induced defects, to mention a few. In classical terms, in a fixed (static) magnetic field, the magnetic moments precess around the field direction at a frequency (the *Larmor frequency)* determined by the strengths of both the moment and the field. When the frequency of the applied radiofrequency (rf) magnetic field matches the Larmor frequency, transitions between the energy levels are induced. When this occurs, the frequency of the applied radiation and the precession frequency are said to be *in resonance.*

Historically, the first experimental observations were made by Zavoisky in the USSR (Zavoisky, 1945). These experiments were performed in rather weak applied fields, so that the frequency of the corresponding radiation field was essentially in the radiofrequency domain. Subsequently, Bagguley and Griffiths in England (Bagguley and Griffiths, 1947) detected the resonance of chromium ions in a chrome alum crystal using microwave radiation at a frequency of 9.43 GHz (3.18-cm wavelength), thereby initiating the application to EPR of microwave (i.e., radar) techniques developed during World War II. Because the inherent spectroscopic sensitivity

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

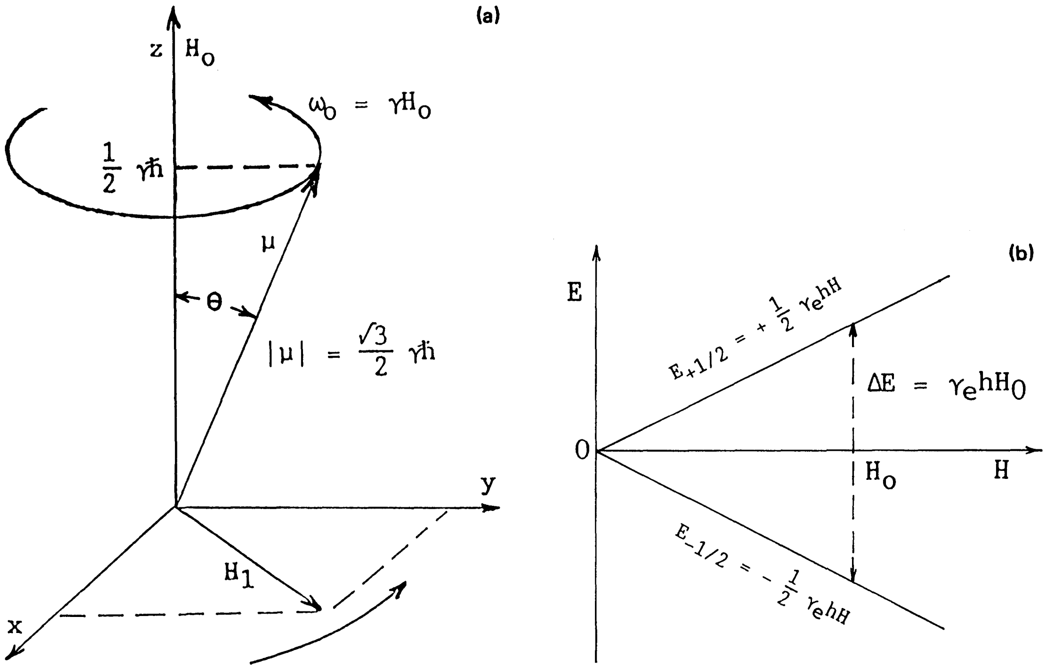

FIG. 1. (a) Classical depiction of the electron paramagnetic resonance phenomenon for spin $S=1/2$. The magnetic moment μ precesses around the direction of H_0 with angular frequency $\omega_0=\gamma H_0$, while the angle θ remains constant. (b) Energy levels and transition frequency for spin 1/2 as functions of applied field H_0.

increases with increasing frequency of the radiation (therefore increasing the energy quantum), it is advantageous to work in the microwave (GHz) region of the electromagnetic spectrum, and most measurements are now made in the range 5–40 GHz in which the needed microwave components are readily available.

1. RESONANCE FUNDAMENTALS

1.1 Classical and Quantum-Mechanical Descriptions of Resonance

The classical representation of the resonance phenomenon is shown in Fig. 1(a). An electron having angular momentum $\mathbf{S}\hbar$ and magnetic dipole moment $\boldsymbol{\mu}$ experiences a torque, $\boldsymbol{\tau}=\boldsymbol{\mu}\times\mathbf{H}_0$, in a magnetic field $\mathbf{H}_0$. This torque causes the direction of the angular momentum vector to change. Since

rate of change of angular momentum

=torque,

or

$$\frac{d(\mathbf{S}\hbar)}{dt}=\boldsymbol{\tau}=\boldsymbol{\mu}\times\mathbf{H}_0, \tag{1}$$

and since $\boldsymbol{\tau}$ is constantly perpendicular to the plane of $\boldsymbol{\mu}$ and $\mathbf{H}_0$, the magnitude of $\boldsymbol{\mu}$ cannot change. Instead, only the direction of $\boldsymbol{\mu}$ changes as it *precesses* around the direction of $\mathbf{H}_0$. Because the electron charge is negative, its dipole moment is directed oppositely (antiparallel) to its angular momentum, so that $\boldsymbol{\mu}$ and $\mathbf{S}$ are related by $\boldsymbol{\mu}=-\gamma\hbar\mathbf{S}$, and the angular frequency of precession is $\omega_0=\gamma H_0$, known as the Larmor frequency.

In the resonance experiment, the magnetic field $\mathbf{H}_1$ of the microwave radiation ($\mathbf{H}_1\ll\mathbf{H}_0$) rotates with the angular frequency ω of the microwaves in the xy plane, perpendicular to $\mathbf{H}_0$. Here, $\mathbf{H}_1$ also exerts a torque, $\boldsymbol{\mu}\times\mathbf{H}_1$, on $\boldsymbol{\mu}$, but unless $\omega=\omega_0$, the time-average value of this torque is zero. However, when $\omega=\omega_0$, the resonance condition is established, and $\boldsymbol{\mu}$ precesses around $\mathbf{H}_1$ at frequency $\omega_1=\gamma H_1$ ($\omega_1\ll\omega_0$), as well as continuing to precess around $\mathbf{H}_0$ at the much faster frequency ω_0.

The result is that μ slowly changes its orientation in the field $\mathbf{H}_0$, tipping down until it reaches the position opposite to its original orientation. In this process, work is done on the dipole moment by the microwave field, or in other words, energy is absorbed from the microwave field $\mathbf{H}_1$ during resonance.

On the other hand, in quantum mechanics the allowed energies of the dipole μ, having magnitude $\gamma\hbar[S(S+1)]^{1/2}$, in the field $\mathbf{H}_0$ are determined by the energy operator (Hamiltonian),

$$\mathscr{H} = -\boldsymbol{\mu}\cdot\mathbf{H}_0 = \gamma\hbar\mathbf{S}\cdot\mathbf{H}_0. \tag{2}$$

By convention, $\mathbf{H}_0$ is taken in the $+z$ direction, so that $\mathbf{S}\cdot\mathbf{H}_0 = H_0 S_z$. The result is that the allowed energies are

$$E_M = \gamma\hbar H_0 M, \tag{3}$$

where the quantum number M specifies the allowed values of the z component of $\mathbf{S}$. These are $-S$, $-S+1,\dots,$ $S-1$, S, so that in the simplest case, that of a single "free" electron, for which $S=1/2$, we have $M=\pm 1/2$, and the allowed energies are just two,

$$E_{\pm 1/2} = \pm\tfrac{1}{2}\gamma_e\hbar H_0, \tag{4}$$

as shown in Fig. 1(b). Transitions between these levels due to magnetic dipole radiation require that $\Delta M = \pm 1$, so that in the resonance experiment the energy quantum $\hbar\omega$ of the microwave magnetic field must satisfy the Bohr condition that

$$\hbar\omega = E_{+1/2} - E_{-1/2} = \gamma_e\hbar H_0, \tag{5}$$

which means that $\omega = \gamma H_0 = \omega_0$, just as in the classical picture.

An order-of-magnitude indication of ω_0 results from the basic *free-electron magnetogyric ratio*:

$$\gamma_e = 2\pi(28.0246\ \text{GHz/Tesla}).$$

1.2 Relaxation Phenomena

Although the preceding section concerned the resonance behavior of a single electronic moment, or spin, real materials contain many spins. In thermal equilibrium, the relative numbers of moments in the two energy levels in Fig. 1(b) are determined by the Boltzmann distribution,

$$N^+/N^- = \exp(-\gamma\hbar H_0/k_B T) \tag{6}$$

where N^+ and N^- are the numbers of moments in the higher and lower energy states and k_B is Boltzmann's constant. This assures that $N^- > N^+$, i.e., more moments are in the low-energy state. And except at very low temperature, $\gamma\hbar H_0 \ll k_B T$ (for the "free" electron in a field of 1 T, $\gamma\hbar H_0 = k_B T$ when $T = 1.35$ K). The microwave radiation field results in both absorption and stimulated (induced) emission by the electronic moments. The transition probabilities are the same for both processes, but since $N^- > N^+$, energy absorption will predominate until $N^+ = N^-$. When this occurs the *spin system* is said to be *saturated.* No further energy can be absorbed from the microwave field, and resonance would not be detectable.

Since EPR is usually readily detected, other mechanisms acting within the material (liquid or solid) must effect energy transfer from the magnetic moments to their surroundings. These mechanisms contribute to *spin–lattice relaxation* and result from interactions of the moments with each other and with other electrons of the host molecules or solid, referred to collectively as the lattice. The magnetic fields of these interactions often fluctuate randomly in time because of thermally activated motions. In a solid, the cause may be lattice vibrations (phonons), and in a liquid, fast random tumbling of the molecules. The spectrum of the resultant fluctuating fields at a specific electronic moment will contain a component at the resonance (Larmor) frequency that is also capable of causing transitions between the energy levels in Fig. 1(b), thereby transferring energy from the moments (the spin system) to the lattice. The relaxation process usually depends exponentially on time, and its rate is characterized by the *spin–lattice relaxation time,* T_1, also called the *longitudinal relaxation time* because it concerns changes in the component of μ parallel to $\mathbf{H}_0$.

If T_1 is sufficiently short, it affects the width of the resonance line through the uncertainty-principle relation, $\Delta E\ \Delta\tau > \hbar$, where ΔE is the uncertainty (spread) of an energy level and $\Delta\tau$ is the lifetime of a state belonging to that level, i.e., T_1. If T_1 is very short, ΔE becomes large, and the transition (resonance) frequency is no longer sharp, but broadened.

In addition, the various interactions referred to above cause different moments to "see" slightly different local fields so that all moments do not precess at exactly the same Larmor frequency, resulting in broadening of the resonance. If all the moments could somehow be caused to precess in phase at exactly the same frequency, the existence of the different local fields would cause them to gradually fall out of phase. The characteristic time for this to occur is the *transverse* or *phase-coherence time*, T_2, also called the *spin–spin relaxation time.* Both times, T_1 and T_2, influence the rate at which the magnetization of the entire sample returns to its equilibrium value after a disturbance: T_1 that of the component parallel to H_0, and T_2 that of the components perpendicular to H_0. These times, and relaxation processes, are treated more quantitatively in Secs. 2.5 and 4.

1.3 Experimental Arrangements

The essential elements of a typical EPR spectrometer are a magnet to provide a homogeneous field H_0 over the dimensions of the sample, a source of microwave radiation at frequency ω_0, and a means of detecting the absorption of the radiation at resonance. By far the most frequently used microwaves are those in the so-called X band, having a nominal frequency in the range 9–10 GHz and wavelength $\lambda \simeq 3$ cm, for which the required $H_0 \simeq 0.32$ T. Since resonances can be as narrow as 5×10^{-6} T (or ~ 150 kHz), both the field and frequency must be stable to better than 2 in 10^5 in order for the resonance to be resolved without distortion.

EPR spectrometers are of two basic, complementary types, *steady state* or *continuous wave* (cw), and *pulsed* or *Fourier transform* (FT). With a cw spectrometer, the equilibrium energy-level populations, N^+ and N^-, of Fig. 1(b) are only slightly disturbed from their thermal-equilibrium values, whereas with a pulse spectrometer N^+ and N^- may be equalized or even inverted. Thus, whereas the cw system observes an EPR resonance under thermal equilibrium conditions, the FT system detects the resonance by observing the recovery of the spin (moment) system from a condition of extreme nonequilibrium. Historically, the overwhelming majority of EPR studies have been made using cw spectrometers, both because of their earlier development and because of the lesser technical demands on the microwave components.

In order to "see" a resonance using a cw spectrometer, some range of field or frequency enclosing the resonance condition must be *scanned.* For technical reasons, the microwave source, which is usually a klystron, is held fixed by *locking* its frequency to that of the sample cavity or to a high harmonic of a quartz crystal oscillator, and the magnetic field is scanned through the resonance. Additionally, a weak sinusoidal *modulation* of the field H_0 at a frequency on the order of 100 kHz is normally utilized. This is provided by small coils affixed to the cavity walls. When the amplitude of this modulation field is less than about $0.2 \times$ linewidth, the envelope of the detected signal is proportional to the first derivative of the lineshape. The derivative signal has the advantage that small changes in slope of the lineshape are more readily apparent, especially for powder samples. It is the derivative signal, rather than the lineshape itself, that is conventionally presented in the literature.

A schematic block diagram of a conventional cw spectrometer system is shown in Fig. 2 (all of the microwave waveguide circuitry is not shown). The central element of the microwave bridge is a circulator or magic tee. This yields the sensitivity of a bridge circuit by enabling the detector to respond to the signal from the sample cavity without interference from the incident microwave power. When this element is properly matched to the characteristic impedance of the waveguide, none of the input microwave power from the klystron appears at the detector (see MAGNETIC RESONANCE SPECTROMETERS and MICROWAVE CIRCUITS for additional detail). The magnetic field is stabilized and controlled by a Hall-effect probe located in the magnet gap. A nuclear magnetic resonance (NMR) fluxmeter ("gaussmeter") is employed to measure the field, and may also be used for field stabilization. In the system shown, the klystron frequency is locked to the cavity resonance frequency through auxiliary circuitry, and its frequency is measured by comparison with a quartz-crystal standard in the frequency counter. Accurate measurements of both field and frequency are essential to ensure 5-significant-figure accuracy in values of the positions of resonance features, especially the effective magnetogyric ratio γ. The voltage output of the lock-in amplifier is changed to a pulse train by a voltage-to-frequency converter

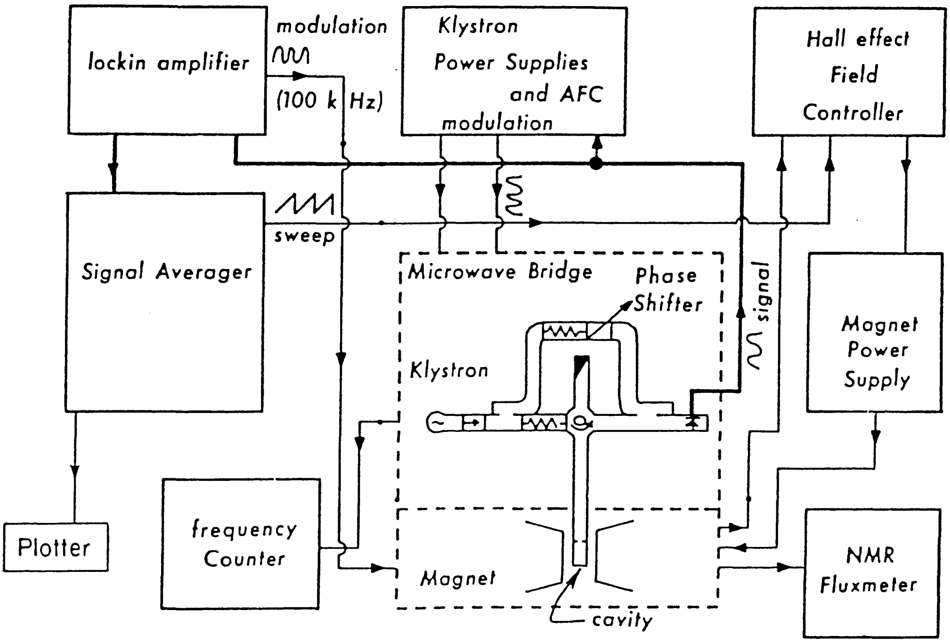
lockin amplifier
modulation
(100 kHz)
Klystron
Power Supplies
and AFC
modulation
Hall effect
Field
Controller
Signal Averager
sweep
Microwave Bridge
Phase
Shifter
Klystron
signal
Magnet
Power
Supply
Plotter
frequency
Counter
Magnet
cavity
NMR
Fluxmeter

FIG. 2. Schematic block diagram of a conventional continuous-wave EPR spectrometer system.

circuit, and the pulses are counted and stored in the channels of the signal averager. The signal averager enables repetitive scans to be made through the spectrum to improve the signal/noise performance. Final output is to a plotting device.

Two different methods for detection of a spectrum may be distinguished in pulsed EPR. In the first, the magnetization recovery following a pulse of microwave radiation that equalizes the populations N^+ and N^- is Fourier transformed from the time domain to the frequency domain. Such a pulse is referred to as a "$\pi/2$ or 90° pulse," since it corresponds to changing the z component of the electronic magnetization from parallel to H_0 to perpendicular. The length t (in time) of the required pulse follows from the condition $\omega_1 = \gamma H_1$ (preceding section), and the relation, $\theta = \omega_1 t$, for the angle through which μ precesses (nutates), yielding

$$\theta = \gamma H_1 t. \tag{7}$$

A typical 90° pulse length of 20 ns requires that the amplitude of $H_1 \simeq 0.45$ mT (4.5 Gauss). The magnetization recovery following a 90° pulse is called the *free-induction decay* or *FID* signal, and its Fourier transform yields the spectrum in the frequency domain. As examples, the FID signal and resulting FT spectrum of perylene ($C_{20}H_{12}$) are shown in Fig. 3. The multiplicity of lines in the spectrum results from the hyperfine interaction of the unpaired ring-electron density with the attached protons (see Sec. 2.4). In the second method, the integrated *echo* signal (see Sec. 4.2) following a 90°–180°-pulse sequence is recorded as the magnetic field is scanned. This method is free of bandwidth limitations which affect the applicability of the FT technique to broad resonances. Both techniques yield the absorption signal itself rather than its derivative.

The block diagram of a simplified pulse spectrometer is shown in Fig. 4. The pulse gate passes and shapes the microwave pulses which are then amplified by the traveling-wave-tube (TWT) amplifier unit and sent to the resonance cavity in the magnet gap. The in-phase and out-of-phase components of the magnetization recovery signal are obtained simultaneously with separate detection channels and combined in the final Fourier-transform output.

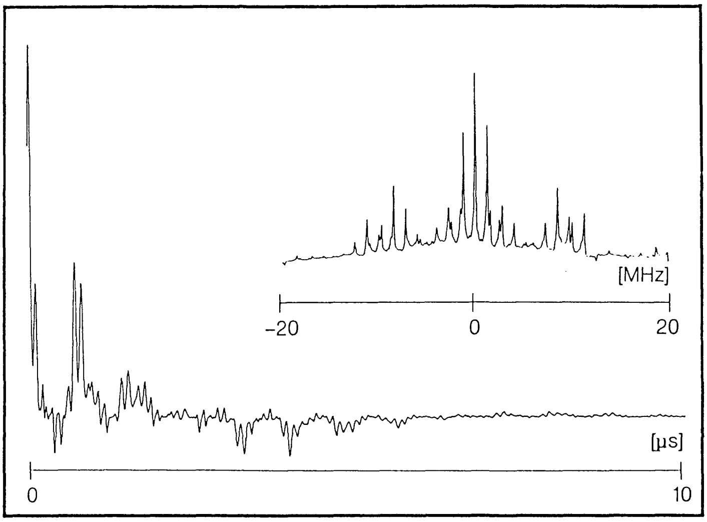

FIG. 3. Free-induction decay (FID) signal of perylene [$C_{20}H_{12}$]. The inset shows the resulting Fourier-transform spectrum. The multiplicity of lines results from the hyperfine interaction of the unpaired electron density with the attached protons.

2. BASIC FEATURES OF EPR SPECTRA

2.1 Introduction: The Spin Hamiltonian Concept

Electrons in atoms, molecules, and solids experience a variety of electromagnetic interactions that shift and split the simple energy level scheme expected solely from the electron's interaction with the applied field H_0. These effects are reflected in the EPR spectrum. In practice, it turns out that spectra can be described fairly simply in terms of transitions between energy levels that are derived from a quantum mechanical energy operator (Hamiltonian) that depends only on spin operators. The coefficients of the various energy terms may be regarded, on the one hand, as parameters derived from the measured EPR spectrum, and, on the other, as values to be calculated from theoretical (first-principles) models of the appropriate electronic structure. This so-called *spin Hamiltonian* constitutes the meeting ground between experimental measurements and basic theory. Virtually all EPR measurements are "fitted" first by an energy level scheme that results from such a spin Hamiltonian.

The interactions that enter into the spin Hamiltonian are basically three:

1. the direct magnetic interaction of the electronic dipole moment μ with the applied field $\mathbf{H}_0$,
2. the dipolar interaction of unpaired electrons with one another, and
3. the electric and magnetic interactions of the unpaired electron(s) with atomic nuclei.

The parameters that characterize these three interactions are

1. the g factor, or g tensor,
2. the fine-structure parameters D and E, and
3. the hyperfine interaction parameter A.

Each of these furnishes its own "signature" that contributes to identifying and characterizing the origin and environment of the paramagnetic entity.

In the simplest cases, only one or two parameters are needed to determine the spectrum from the spin Hamiltonian. But even quite complicated spectra frequently require not more than 8, or at most 10, parameters for their characterization.

2.2 The g factor

Quantum mechanically, μ for a single electronic entity having spin angular momentum $\mathbf{S}\hbar$ is related to $\mathbf{S}$ by

$$\mu = -g(|e|\hbar/2m)\mathbf{S} = -g\beta\mathbf{S} = \gamma\hbar\mathbf{S}, \qquad (8)$$

where $\beta = |e|\hbar/2m$ is the Bohr magneton, g is the spectroscopic splitting factor, or simply "g factor," and μ and $\mathbf{S}$ are vector operators. A truly "free" electron has only its intrinsic spin angular momentum and so $g = g_e = 2.0023$, the free-electron g value. The g factor determines the location of the resonance in the field-

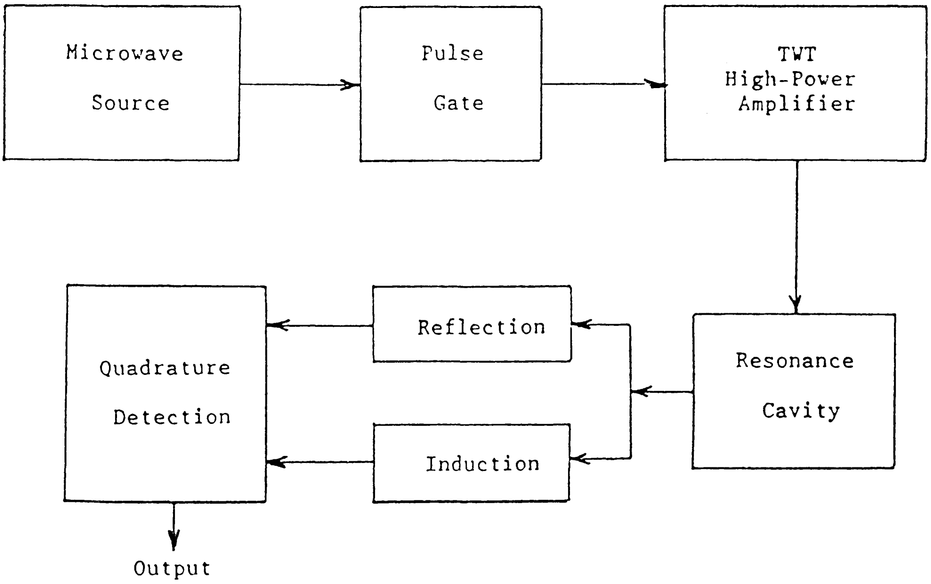

FIG. 4. Schematic block diagram of a simplified pulse EPR spectrometer.

frequency domain. It provides the first of several signatures that enable the origin of the resonance to be identified, and/or interpreted in terms of atomic (or molecular) bonding and other considerations.

Generally, the g factor is not a single number and may differ significantly from g_e. The electrostatic interactions referred to above cause the electronic environment to be determined in part by orbital as well as spin contributions. Moreover, because of the orbital contribution, μ and $\mathbf{S}$ may no longer be collinear, and g may not be isotropic, i.e., independent of the orientation of the field $\mathbf{H}_0$ with respect to the sample. Then, g becomes a tensor $\tilde{g}$ with components g_{xx}, g_{yy}, g_{zz} in its principal-axis coordinate system which is fixed in the molecule or solid. As a result, $\mu = -\beta\tilde{g}\cdot\mathbf{S}$, and the resonance field for a given microwave frequency depends on the orientation of the sample in the field. An important simplification arises, however, when the molecule or ion is in highly fluid motion. The position of the resonance is then determined by the average of g_{xx}, g_{yy}, g_{zz}, since the sum of the diagonal elements of a tensor (the trace) is invariant with respect to coordinate rotations. This average g is often called the g factor to distinguish it from the g tensor.

Shown in Fig. 5 is the orientation dependence of the spectrum of Pt^{3+} impurities present in a $ZnWO_4$ crystal grown in a Pt crucible (Watterich *et al.*, 1991). The spectrum was recorded at a temperature of 6 K as $\mathbf{H}_0$ was rotated in the (100) plane of the crystal. The three-line EPR pattern includes a central line arising from Pt^{3+} ions whose nuclei have zero spin, and two symmetrically placed hyperfine satellite lines due to Pt^{3+} ions whose nuclei (^{195}Pt) have spin 1/2 (see Sec. 2.4). In this example, $g_{xx} \neq g_{yy} \neq g_{zz}$, indicating that the site symmetry is low. Single crystals of $ZnWO_4$ have considerable interest

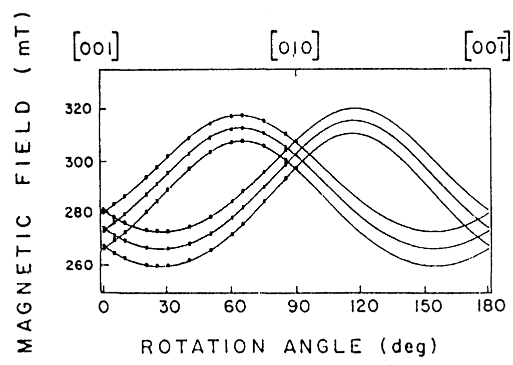

FIG. 5. Angular variation of Pt^{3+} EPR spectrum in $ZnWO_4$ measured at ~6 K and 9.2 GHz. Experimental data are shown by points, and the solid lines represent the calculated angular variation based on computer-fit spin-Hamiltonian parameters (Watterich *et al.*, 1991).

because of their potential application for x-, γ-, and particle-ray scintillation detectors. For such applications, it is desirable to have crystals that do not exhibit coloration due to impurity ions. EPR studies have identified and characterized the location of low levels of various paramagnetic impurity ions in such crystals.

In many cases, the g tensor has axial symmetry. In a powder sample, crystallites of all orientations θ contribute to the resonance, where θ is the angle between the site symmetry axis and H_0. An example of the resulting *powder pattern* is furnished by the dangling-bond EPR signal in amorphous silicon (Biegelson, 1981), shown in Fig. 6. The stronger feature in the spectrum occurs at $\theta=90°$, i.e., $g_\perp$, and the weaker at $\theta=0$, i.e., $g_\parallel$. Thus, even with a powder sample, the elements of the g tensor can be determined, although not its orientation within the material.

2.3 Fine Structure

The EPR spectra that occur when $S>1/2$ often show up to $2S$ different features, rather than a single resonance, indicating that the separations between the $2S+1$ energy levels are not equal. This *fine structure* in EPR spectra has different origins depending on the nature of the paramagnetic center. The principal characteristics of the interactions that give rise to fine structure are

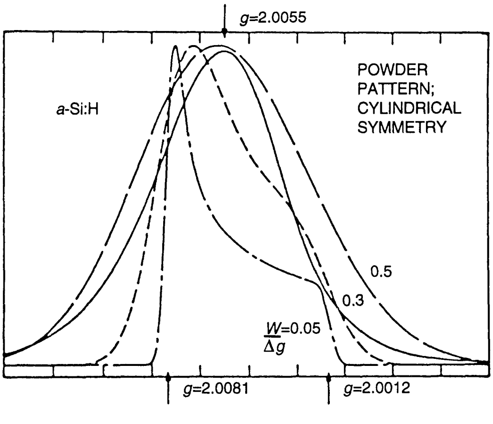

FIG. 6. Powder pattern for dangling-bond EPR using Gaussian broadening with half-width at half-height W. The solid curve is the EPR absorption signal in hydrogenated amorphous silicon (Biegelson, 1981).

1. their dependence on μ^2 (or S^2), in contrast to the simple magnetic interaction with the field H_0 which depends on μ (or S_z), and
2. the fact that they are independent of H_0.

The dependence on μ^2 (or S^2) means that the interactions are indifferent to the signs of the magnetic sublevels, so that the $M=\pm 1/2$ levels of $S=1/2$ experience the same shift in energy, for example. Consequently, fine structure only arises when $S>1/2$. When $S=1$, the $M=\pm 1$ sublevels are split away from the $M=0$ sublevel; when $S=3/2$, there will be a splitting into $M=\pm 3/2$ and $\pm 1/2$ sublevels. The second characteristic means that the splitting persists in the absence of H_0—it is "built into" the molecule or solid—and for this reason is frequently referred to as *zero-field splitting.*

In the framework of the spin-Hamiltonian approach, the interactions are represented by an energy term of the form

$$E_{\text{fs}}=D[S_z^2-S(S+1)/3]+E(S_x^2-S_y^2). \qquad (9)$$

When the interaction is isotropic or nearly so ($E \ll D$), and sufficiently weak in comparison to the magnetic dipole energy $-\boldsymbol{\mu}\cdot\mathbf{H}_0$, to a first approximation the effect on the EPR spectrum takes the form

$$E_{\text{fs}}\simeq (D/2)[3S_z^2-S(S+1)](3\cos^2\theta-1), \qquad (10)$$

where θ is the angle between the z axis of the D, E tensor and $\mathbf{H}_0$.

Briefly, fine structure in triplet states of organic molecules and biradicals arises from the dipolar interaction between two electrons whose total spin $S=1$. This is the simplest case. For EPR of complexes of transition-metal ions, fine structure is generally due to the spin–orbit interaction, but the resulting spin-Hamiltonian form is the same as for the dipolar interaction. (See Sec. 3.1 for an example.)

2.4 Hyperfine Structure

Many species of atomic nuclei (nuclides) also possess magnetic dipole moments. When such nuclides are the nuclei of paramagnetic ions (or even the nuclei of neighboring atoms or ions), the magnetic dipole–dipole interaction between nucleus and electron moments affects the energies of both. This effect is given the name *hyperfine interaction.* The small magnitude of the nuclear moment in comparison to that of the electron ($\sim 10^3$ smaller) means that the electron experiences very weak magnetic fields, on the order of 1 mT, in addition to the applied field H_0. In contrast, the nucleus experiences a strong field (the hyperfine field) due to the strong electron moment, frequently much stronger than any H_0 used in NMR experiments.

In the simplest cases, the hyperfine interaction is isotropic, and perturbation theory shows that the interaction energy is represented by a term aMm_I, where m_I is the z-component angular momentum quantum number of the nuclear spin I, and a is proportional to the product of the nuclear and electron moments and to the electronic spin density at the nucleus. Since there can be a total of $2I+1$ values of m_I, corresponding to the allowed orientations of the spin I in the field H_0, each electronic level characterized by M splits into $2I+1$ closely spaced levels. In the EPR spectrum resulting from transitions in which $\Delta M = \pm 1$, the nuclear orientation, i.e., m_I, remains fixed, so that each line in the spectrum splits into $2I$ components which are referred to as *hyperfine structure* (hfs). The hyperfine interaction often provides direct identification of the atom or locale of the paramagnetic entity. Figures 3, 5, 7, and 11 provide examples of hyperfine structure in EPR spectra.

More generally, the hyperfine interaction contains contributions from both the electron's orbital motion (*dipolar* contribution) with respect to the nucleus while in p or d orbitals, and its direct (*contact*) interaction with the nucleus while in an s orbital. The dipolar part, A, contributes anisotropically whereas the contact interaction, a, contributes only isotropically.

2.5 Resonance Line Shapes and Intensity

In bulk matter (solid or liquid), the magnetic moments interact with the static field $\mathbf{H}_0$, with one another, and also with the random thermal motions of the lattice (all other atoms). Except at very low temperatures, the lattice serves effectively as an infinite thermal reservoir at temperature T since the thermal energy $k_BT \gg \gamma\hbar H_0$ even in moderately strong magnetic fields, where k_B is the Boltzmann constant. In thermal equilibrium, assuming interactions between the moments are also weak, the equilibrium magnetization is proportional to $\mathbf{H}_0$ and is given by the Curie formula for the magnetic susceptibility χ_0,

$$\mathbf{M}_0 = \chi_0 \mathbf{H}_0 = \frac{N\gamma^2\hbar^2 J(J+1)}{3k_B\mathrm{T}} \mathbf{H}_0 \qquad (11)$$

where N is the density of magnetic moments per unit volume.

In the resonance experiment the total applied field is the sum of $\mathbf{H}_0$ and the field $\mathbf{H}_1$ perpendicular to $\mathbf{H}_0$ and oscillating at frequency ω. To describe the behavior of the macroscopic magnetization vector $\mathbf{M}$ in such a field, where $\mathbf{M}$ is the sum of all the individual moments, i.e., $\mathbf{M} = \Sigma_i \boldsymbol{\mu}_i$, the basic torque equation

$$\frac{d\mathbf{M}}{dt} = \gamma \mathbf{M} \times \mathbf{H}, \qquad (12)$$

which describes the precessional and nutational motions of the magnetization $\mathbf{M}$, is added to equations describing the transient damping out of nonequilibrium components of $\mathbf{M}$. Taking $\mathbf{H}_0$ in the z direction,

$$\frac{dM_z}{dt} = -\frac{(M_z - M_0)}{T_1} \qquad (13)$$

defines the longitudinal, or spin–lattice, relaxation time T_1 with which M_z approaches its equilibrium value $M_0 = \chi_0 H_0$. Similarly,

$$\frac{dM_{x,y}}{dt} = -\frac{M_{x,y}}{T_2} \qquad (14)$$

defines the transverse relaxation time T_2 with which M_x and M_y are reduced to their equilibrium value of zero.

Under these assumptions the motion of $\mathbf{M}$ is described by the equations [the *Bloch equations* (Bloch *et al.*, 1946)]

$$\frac{d\mathbf{M}}{dt} = \gamma \mathbf{M} \times \mathbf{H} - \frac{M_x \mathbf{i} + M_y \mathbf{j}}{T_2} + \frac{M_z - M_0}{T_1} \mathbf{k}, \qquad (15)$$

where $\mathbf{H} = \mathbf{H}_0 + \mathbf{H}_1$, and $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are unit vectors in the x, y, z directions in the laboratory coordinate frame.

The steady-state solution of these equations has a component of $\mathbf{M}$ in phase with the applied oscillating field $\mathbf{H}_1$ and a component out of phase with $\mathbf{H}_1$. Conventionally, that part of $\mathbf{M}$ parallel to $\mathbf{H}_1$ is written as a complex quantity $\mathbf{M}_1 = (\chi' - i\chi'')\mathbf{H}_1$. The components of the rf susceptibility $\chi' - i\chi''$ are

$$\chi'=\frac{1}{2}\chi_0\omega_0\frac{\Delta\omega T_2^2}{1+(\Delta\omega)^2T_2^2+(\gamma H_1)^2T_1T_2}, \qquad (16)$$

$$\chi''=\frac{1}{2}\chi_0\omega_0\frac{T_2}{1+(\Delta\omega)^2T_2^2+(\gamma H_1)^2T_1T_2}, \qquad (17)$$

where $\Delta\omega=(\omega_0-\omega)$. These susceptibilities exhibit a resonance character similar to those of the complex impedance of a tuned circuit. When the third term in the denominators becomes significant, both χ' and χ'' decrease in magnitude. This *saturation* term $\gamma^2H_1^2T_1T_2$ is usually made smaller than the other terms in cw experiments by reducing H_1 sufficiently.

For a sample in a resonant microwave cavity, the changes in the cavity quality factor Q and resonance frequency ω in passing from off resonance to on resonance (e.g., by varying the applied field H_0) depend on χ' and χ''. Resonance can be detected by the change in Q or by the shift in the resonant frequency of the circuit. These signals are referred to as the *absorption* and *dispersion modes,* respectively, and it is the absorption-mode (χ'') signal that is ordinarily observed. For this,

$$\Delta Q/Q_0=4\pi f\chi''Q_0, \qquad (18)$$

where Q_0 is the Q of the empty cavity and f (filling factor) represents the ratio of magnetic energy within the sample volume to the total magnetic energy stored in the resonant cavity.

Without saturation, the shape of $\chi''(\omega)$ is *Lorentzian* [after the function $f(x)=a/(1+x^2)$]. The half width at half-maximum intensity (HWHM) is $\Delta\omega_{1/2}=1/T_2$, so that measurement of the linewidth determines T_2. If T_1 is very short, the linewidth is dominated by lifetime broadening (Sec. 1.2), and $T_2=T_1$, so that in such cases measurement of the linewidth determines T_1.

The other commonly encountered line shape is the *Gaussian* function, for which $\chi''(\omega)\propto\exp[-(\Delta\omega T_2)^2/2]$. Such a line shape, which is an envelope of contributions from spins in slightly different magnetic field environments, is *inhomogeneously* broadened. It is frequently possible to *burn a hole* in such a resonance by subjecting a portion of it to strong microwave radiation whose frequency spread is less than the total linewidth, as then only those spins within a narrow range of local fields experience saturation. EPR spectra of color centers provide examples of such resonances (see Sec. 3.3.1.2 and Fig. 11). In contrast, *homogeneous* broadening occurs when the resonance results from a transition between energy levels that are not sharply defined, but are broadened somewhat by the dipolar interaction between moments, or by a short T_1, or by diffusive motion, for example.

3. SOURCES OF EPR SPECTRA: THEIR INTERPRETATION AND APPLICATIONS

EPR signals may arise whenever unpaired electron spins are present. This may happen in solids, liquids, and gases, and may be a natural consequence of the electronic structure of the material or may be induced by a variety of changes and external influences. The following sections illustrate EPR spectra in several categories of materials. The application and utilization of these spectra are described.

3.1 Free Radicals and Triplet States

The analysis and interpretation of the hyperfine structure (hfs) in the spectra of free radicals in solution have yielded some of the most important EPR results. First, the electrons in covalent bonds are strongly constrained to the molecular framework, so that there is a large energy splitting of the orbital levels. The isotropic g factor is always close to g_e, and the g tensor is close to isotropic. Put simply, the hfs reflects directly the unpaired-electron distribution in the radical. The fast random tumbling of a radical in solution averages the dipolar (anisotropic) hyperfine contribution to zero, so that the observed hfs results from the contact interaction with electrons in s orbitals. The isotropic interaction parameter a is directly related to the spin-density distribution of the unpaired electron at nuclei in the radical.

The organic salt $(C_6H_5)_2N{-}NC_6H_2(NO_2)_3$, diphenyl picryl hydrazine, usually known as DPPH, was among the first free radicals studied by EPR. Because of its narrow resonance, $g=2.0036$, close to g_e, and strong signal (<10 μg yields a readily detected signal at 9 GHz), it has been extensively used to provide magnetic field calibration by including a minute crystal in the resonance cavity adjacent to the material under investigation. When DPPH is dissolved in benzene, a well-defined hyperfine pattern appears having a central peak flanked by two subsidiary lines on each side. Since $I=1$ for ^{14}N, the total nuclear spin of the two central N atoms is 2, so that $2\times2+1=5$ orientations of the total nuclear spin can occur. This shows that the unpaired electron wave function is distributed equally over the two *N* atoms.

Free radicals have been extensively used in the form of *spin labels* to study bimolecular structure. In many cases the free radical is the relatively unreactive nitroxide (N–O) group containing an unpaired electron localized on the N, so that the EPR signal shows the characteristic three-line hyperfine pattern due to the spin-1 ^{14}N nucleus. A reactive functional group on the other end of the molecule containing the N–O group enables the spin label to be attached to the molecules or systems under study. Information about the host is derived from the effects of motion and *chemical exchange* (see Sec. 4.3) on the spectrum. The effects of motion are illustrated in Fig. 7 which shows the spectrum of the nitroxide radical $[(CH_3)_3C]_2N–O$ dissolved in ethylene glycol at temperatures ranging from 25 °C in (a) down to −150 °C in (d). At 25 °C the fast random tumbling motion in the nonviscous solvent averages the anisotropic hyperfine contribution to zero, leaving a sharp, symmetric spectrum. With decreasing temperature, the increasing viscosity of the solvent slows down the tumbling motion of the radical, and the spectrum becomes increasingly asymmetric as the anisotropy of the hyperfine interaction begins to contribute. In (d) the solvent has essentially become a glass, so that all orientations of N–O groups contribute to the hyperfine *powder pattern* spectrum.

Both spin labels and naturally occurring free radicals have been extensively used in studies of natural organic matter, including applications to metabolic activity, enzymes, photosynthesis, protein structure, etc. Other applications have included the study of coal and coal derivatives and many other forms of charred materials—even cigarette tar and roasted coffee.

The triplet state of an organic molecule arises from two electrons, $S_1=S_2=1/2$, separated by a distance *r*, which yield a total spin $S=1$. This may occur when one electron from an electron-pair bonding orbital is excited to a higher energy level with its spin aligned parallel to that of the remaining electron. The magnetic dipole–dipole interaction between them is responsible for the fine structure energy parameters *D* and *E* of Eq. (11). Although the interaction is not generally isotropic, both *D* and *E* are proportional to $g^2\beta^2/r^3$. For $S=1$, the zero-field splitting ($H_0=0$) is just *D* in the isotropic case ($E=0$), and the EPR spectrum consists of two lines symmetrically displaced on either side of the nominal resonance position. Their spacing depends on sample orientation *via* Eq. (10).

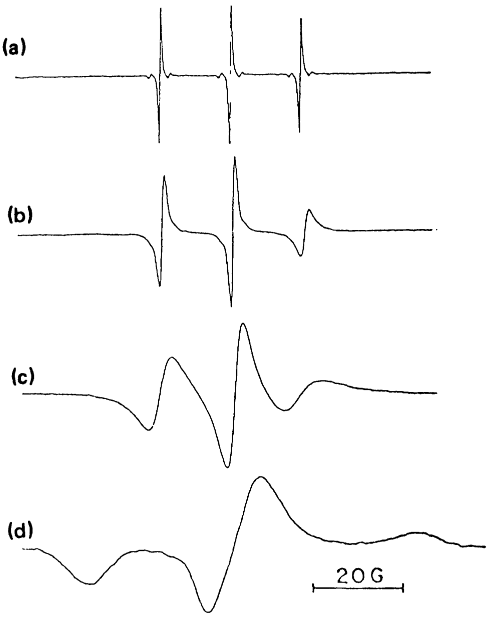

FIG. 7. EPR spectra of the nitroxide free radical $[(CH_3)_3C]_2N–O$ dissolved in ethylene glycol at (a) 25 °C, (b) −25 °C, (c) −80 °C, and (d) −150 °C, showing the progression of the three-line ^{14}N hyperfine pattern from fast fluid motion to a random rigid solid (Griffith and Waggoner, 1969).

As an example, Fig. 8 shows the angular variation of the triplet state spectrum of the naphthalene molecule, $C_{10}H_8$, diluted in a crystal of durene, 1,2,4,5-$(CH_3)_4C_6H_2$ (Hutchison and Mangum, 1961). The triplet state was excited by irradiating the crystal with light from a mercury arc lamp. Two sets of lines appear in Fig. 8 because the durene crystal contains two molecules per unit cell. The spacing between the lines of a pair is as great as 0.2 T (2000 Gauss), a substantial fraction of the nominal resonance field of 0.33 T. Detailed analysis of the spectra yields $g=2.0023$,

$D=3.0$ GHz, and $E=0.41$ GHz. The strong fine structure interaction mixes the $M=0$, ± 1 levels to such a degree that it becomes possible to observe the normally *forbidden*, $\Delta M=\pm 2$, transition between the $M=\pm 1$ levels, which occurs at approximately half the normal resonance field, i.e., at $g \simeq 4$. This *half-field* transition is relatively insensitive to the sample orientation and can be detected in materials containing a random orientation of molecules (see example in Sec. 5.2).

3.2 Transition-Metal Ions

For transition-metal ions, the observed EPR spectrum results from transitions between a few energy levels of the ion which have lower total energies than any others. In many cases, only two levels are involved, and the total number of levels almost never exceeds 8. For the purpose of characterizing such spectra, these ground-state levels may be taken to constitute an isolated set with properties described independently of other levels at higher energies. In so doing, the concept of a fictitious or *effective spin*, S_{eff}, or simply S, is introduced, defined by setting the number of ground state levels (i.e., the multiplicity) equal to $(2S+1)$. If the ground state is a doublet, then $S=1/2$, and the energies of the two levels in an applied field H_0 are $g\beta H_0 M$, with $M=\pm 1/2$. If the ground state is 4-fold degenerate, $S=3/2$, and at least one more parameter is needed to characterize the zero-field splitting. Any hyperfine interaction can also be described as occurring between S and the nuclear spin I.

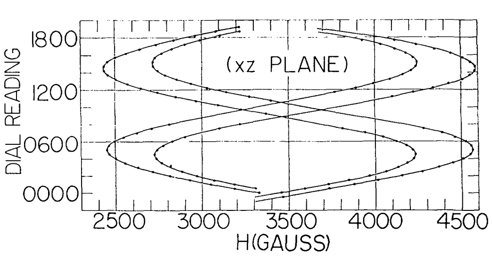

FIG. 8. Angular variation of the EPR spectrum of the excited triplet state of naphthalene molecules in a crystal of durene. Two sets of spectra appear because each unit cell of the crystal contains two molecules. (Note: 1000 divisions on the dial reading = 90°) (Hutchison and Mangum, 1961).

For ions in solids, it is necessary to take account of the effect of the host lattice. For low concentrations ($<0.1\%$) of paramagnetic ions present as impurities or deliberately "doped" into a diamagnetic host crystal, the usual method of estimating the effect the crystal has in modifying the magnetic properties of the ion is the *crystalline electric field* (*CEF*) approximation. The influence of neighboring ions is assumed to occur entirely through the electric field they produce at the site of the magnetic ion. This field exerts forces on the unpaired electrons of the ion and modifies their usual orbital motion. An order-of-magnitude estimate of the energy involved is given by that of two electrons separated by a typical nearest-neighbor distance (~ 2 Å): $e^2/4\pi\epsilon_0 r^2 \simeq 7.2$ eV $\simeq 8.4\times 10^4$ K. This is comparable to electron–electron energies within the ion (2–20 eV), and hence exerts a major influence on ionic properties.

This estimate neglects electrostatic screening of the ion's unpaired electrons if the ion's outer shells are also populated. This is most significant for ions of the rare-earth series in which the unpaired *4f* electrons lie within the closed *5s* and *5p* shells, and reduces the CEF interaction energy to roughly 10^{-2}–10^{-3} eV, which is less than typical spin–orbit interaction ($\lambda\mathbf{L}\cdot\mathbf{S}$) energies and is not comparable with electron–electron interactions. Conversely, the unpaired *3d* electrons in ions of the iron group elements are the outermost electrons, and CEF energies on the order of 5–10 eV occur.

Taking account of these factors, three cases are usually considered:

1. The weak-CEF case applies to the rare-earth and actinide (*5f*) ions in which the spin-orbit interaction dominates so that the ion's magnetic moment is determined by the total angular momentum, $\mathbf{J}=\mathbf{L}+\mathbf{S}$. The g factor is given by the Lande formula appropriate to the free ion,

$$g_L=1+\frac{J(J+1)-L(L+1)+S(S+1)}{2J(J+1)}. \qquad (19)$$

2. The intermediate-CEF case applies to the iron-group ions in which the CEF dominates. The orbital contribution to the moment is "quenched" by the CEF, and the ion's moment is due almost entirely to

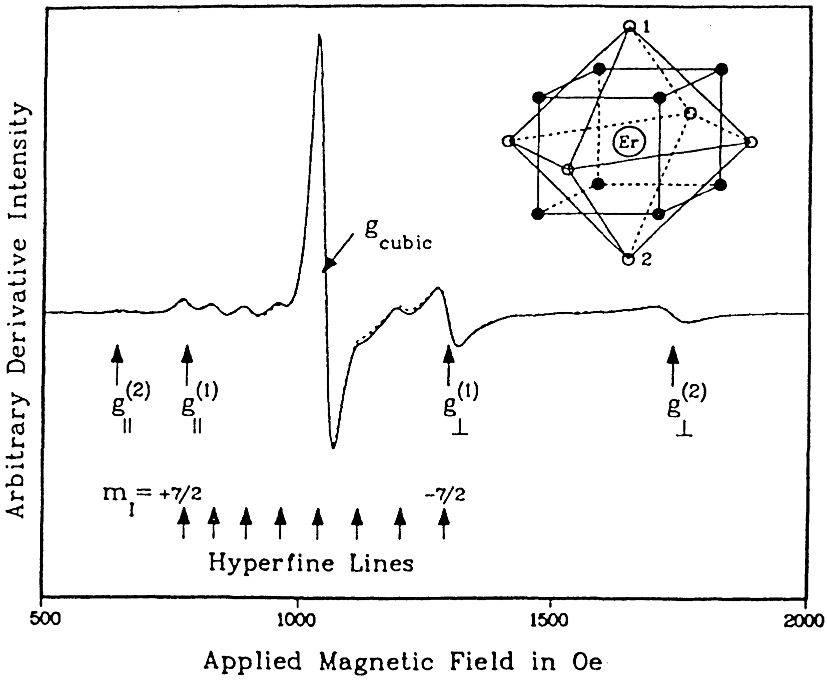

FIG. 9. Experimental (solid) and fitted (dashed) first derivatives of the Er^{3+} EPR in scandium dihydride powder. Positions of the hyperfine lines for the main (cubic site) resonance, as well as of the axial g-tensor resonances ($g_{\parallel}$ and $g_{\perp}$), are shown. The inset shows the tetrahedral (solid circles) and octahedral (open circles) hydrogen sites surrounding Er ions in the face-centered-cubic host lattice (Venturini and Richards, 1979).

electron spin. Consequently, the g factor is close to the free-electron value, and the magnetic properties of the ion in the host lattice bear little resemblance to those of the free ion.

3. The strong-CEF case applies mainly to ions of the platinum- and palladium-group elements which tend to form covalent bonds with neighbor ions.

Ions of the rare-earth element erbium, Er^{3+}, have been used to determine hydrogen site occupancies in metallic hydrides. An example is provided by the study (Venturini and Richards, 1979) of the Er^{3+} EPR in scandium dihydride, ScH_x ($1.85 \leqslant x \leqslant 1.99$), which has the cubic CaF_2 structure with Sc in the Ca positions and H in the F locations. The EPR spectrum of 0.1% Er^{3+}, which substitutes for Sc, at 2 K and 9.8 GHz, is shown in Fig. 9. This is a powder pattern resulting from Er^{3+} ions in three different CEF environments that result from the ionic charges on the adjacent hydrogens. As shown in Fig. 9, the tetrahedral (T) site hydrogens form a simple cube around an Er^{3+} ion. If there are no hydrogens in octahedral (O) sites, the cubic CEF results in an isotropic g tensor for the ground-state doublet, for which the effective spin, $S=1/2$. Since $S=1/2$, there can be no fine structure (Sec. 2.3); however, g-tensor anisotropy and hyperfine interactions can occur.

The strong resonance at ~0.1 T in the spectrum has $g=6.77$, consistent with CEF theoretical expectations and negative charges on the H ions. Weak hyperfine lines surround this resonance. These result from the 23%-abundant ^{167}Er isotope with spin $I=7/2$. The strong unsplit central line results from Er^{3+} ions whose nuclei have spin zero and no magnetic moments($^{162,164,166,168,170}Er$).

The remaining four weak lines in the spectrum are from Er ions in axially symmetric (less than cubic) CEF environments that result from the presence of one or two hydrogens in adjacent O sites, as shown in the inset in Fig. 9. The measured values are $g_{\parallel}^{(1)} = 9.0$, $g_{\perp}^{(1)}=5.42$, corresponding to a single O-site hydrogen on site 1 in the inset, and $g_{\parallel}^{(2)}=11.0$ and $g_{\perp}^{(2)}=4.05$ for 2 hydrogens on sites 1 and 2 in the inset. By measuring the integrated intensities of these lines relative to the isotropic line, the relative probability of having 1 or 2 adjacent O-site hydrogens was determined, and used to evaluate the success of theoretical calculations for site occupancy.

Finally, in this example, the shape of the strong central resonance derivative is not symmetric. This is a consequence of the metallic character of the sample (see Sec. 3.3.2).

3.3 Other Systems

3.3.1 Nonmetallic Solids. The electronic structure of nonmetallic solids (insulators and semiconductors) is characterized in terms of a band structure having a filled valence band (VB) and an empty conduction band (CB), separated from the VB by an energy gap (band gap). The chemical bonding of the valence electrons (density$\simeq 10^{23}$ cm^{-3}) normally assures that their spins are paired according to the Pauli principle, so that such a solid is diamagnetic and gives no EPR spectrum.

However, imperfections such as vacancies, interstitials, and impurities may have localized states and energy levels within the bandgap. Trapping of an unpaired electron in one of these states depends on the position of the Fermi level in the case of semiconductors, and on the availability of electrons (or holes) in the case of insulators. Exposure to light or ionizing radiation, impurity doping, and temperature, for example, may cause existing defects to become paramagnetic. Defects can also arise during crystal growth (see Sec. 2.2 and Fig. 5), ion implantation, and laser annealing (see Sec. 2.4 and Fig. 7), as well as other processes.

3.3.1.1 Amorphous Silicon (a-Si). EPR has been used to study the dominant states in the band gap as well as recombination processes and doping in this technologically important prototypical amorphous semiconductor. The EPR line shape in undoped *a*-Si films is very nearly independent of the conditions under which the film is deposited (evaporation, sputtering, glow discharge, etc.), being an essentially featureless line at $g=2.0055$. The insensitivity to large variations in local defect structures indicates a population of isolated, strongly paramagnetic centers resulting from so-called dangling Si bonds. The line shape has been shown to be the powder spectrum resulting from an assumed Gaussian distribution of g-tensors with $g_{\parallel} \simeq 2.0012$ and $g_{\perp} \simeq 2.0081$. An observed resonance and computer-simulated shapes are compared in Fig. 6 (Biegelson, 1981).

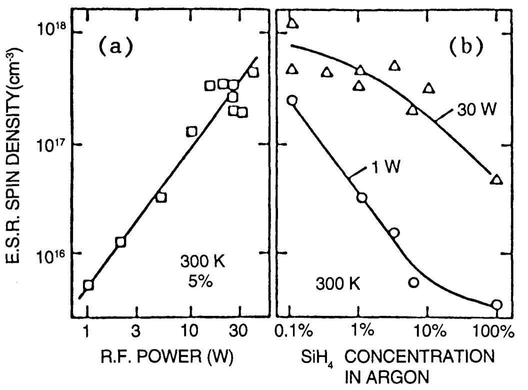

FIG. 10. Dangling-bond spin density in a-Si:H as a function of deposition parameters (a) radiofrequency power level and (b) SiH$_4$ concentration in the argon carrier gas. (Biegelson, 1981).

Although the line shape is independent of deposition conditions, the dangling-bond spin density can vary greatly, from a maximum of $\sim 3 \times 10^{19}$ cm^{-3} to an undetermined low density. Figure 10 shows the dependence of this spin density on two deposition parameters, the radiofrequency power level and the SiH$_4$ (silane) concentration in the argon carrier gas. These graphs (Biegelson, 1981) show that material with the lowest defect concentration is grown in pure SiH$_4$ at the lowest power level. Thus, EPR serves to characterize sample quality and leads to improvements in the deposition process.

3.3.1.2 Radiation-Induced Defects. Among the first defects studied with EPR, *color centers* result from an excess of metal ions in alkali halide (and other) crystals (e.g., NaCl, CaF$_2$, etc.) when heated in alkali metal vapor or when subjected to x-ray, γ-ray, etc., irradiation. The electronic structure of the simplest color center, the *F* center (from the German word for color, *Farbe*), was identified by EPR as that of an electron bound at a negative-ion vacancy. The *F*-center resonance is a Gaussian envelope of the hyperfine lines resulting primarily from interaction with the nuclear magnetic moments of the alkali-metal ions adjacent to the vacancy. The cw absorption derivative record of the *F*-center resonance in an x-rayed LiF crystal (Holton and Slichter, 1960) is shown in Fig. 11. This spec-

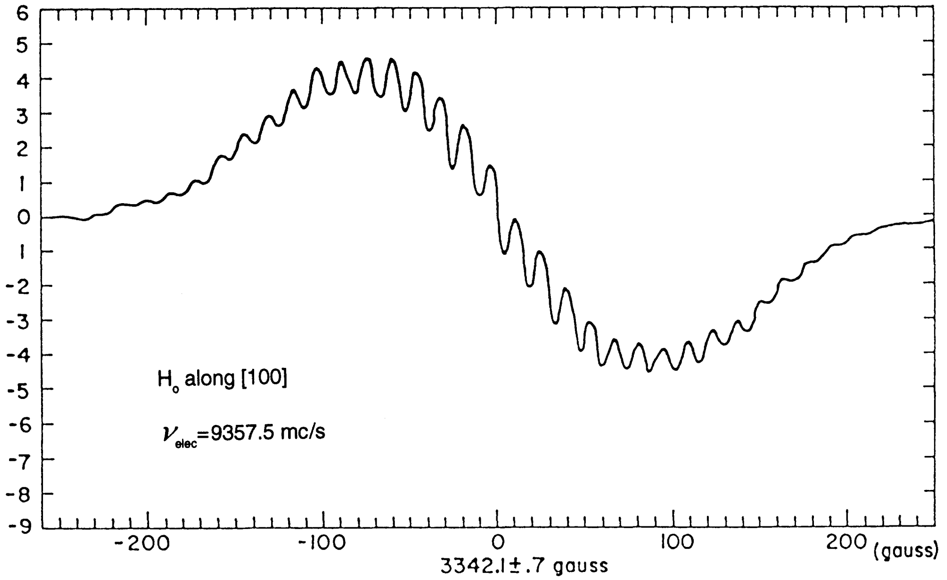

FIG. 11. EPR absorption derivative curve for x-ray irradiated single-crystal LiF at 1.3 K, showing 35 resolved hyperfine lines (Holton and Slichter, 1960).

trum was obtained at $T = 1.3$ K with H_0 parallel to the [100] direction in the crystal. The resonance is centered at $g = 2.0006 \pm 0.0006$, close to the free-electron value, and the width between derivative maxima is ~170 Gauss.

The peaks of 35 resolved hyperfine lines are evident in Fig. 11, whereas the number of hyperfine lines due solely to the six nearest-neighbor ^{7}Li nuclei (spin 3/2) is 19, corresponding to the 19 possible orientations of the total nuclear spin, $6 \times 3/2 = 9$, of the six ^{7}Li nuclei acting together $[2(6 \times 3/2) + 1 = 19]$. This shows that the wave function of the trapped electron extends well-beyond the nearest-neighbor metal ions and interacts with some distant nuclei, both ^{7}Li and ^{19}F, in this case. A detailed ENDOR study (see Sec. 5.1) of this spectrum identified hyperfine interactions from ^{7}Li and ^{19}F nuclei in the first seven neighbor shells around the vacancy. (See Sec. 5.1 and Fig. 14 for an example of *F*-center hyperfine interactions detected by ENDOR.)

A number of other color centers resulting from combinations of *F* centers have also been studied by EPR, the simplest being the *M* center formed by two adjacent *F* centers. A hole in the normally filled p^6 shell of a halogen ion may also be trapped to form a color center. The best-known one is the V_K center, identified by EPR.

3.3.2 Metallic Solids. Metallic solids are characterized by the presence of delocalized electrons (*n* per unit volume) in conduction-band states. At low temperatures, essentially all states below the Fermi energy E_F are filled and those above are empty. With increasing temperature, energy states within a range $\pm k_B T$ of E_F become partially occupied. The net magnetism of these *unpaired* conduction electrons results in the susceptibility of the entire electron gas given by $\chi_P = 3n\beta^2/2E_F$, the *Pauli spin susceptibility* (per unit volume), which is independent of temperature. The EPR signal from the unpaired electrons is readily observed in simple metals such as Li and Na. For most heavier metals the extremely short spin-lattice relaxation time T_1 broadens the resonance beyond detectability. An additional complicating feature of EPR in metallic materials results from the high electrical conductivity which severely attenuates the microwave field within a thin surface region (the *skin depth*), so that extremely small particles are required as samples. If the particles are too large, electrons moving in and out of the skin depth region result in an asymmetric resonance lineshape (see example in Sec. 3.2).

The EPR of magnetic ions (especially rare earths) dissolved in metals (including alloys and intermetallics) yields information about

1. the CEF-split ground state of the impurity ion,
2. the dynamics of the relaxation processes,
3. the sign and form of the coupling between the impurity moments and the conduction electrons (and therefore about the electronic band structure of the host metal), and
4. the onset of magnetic ordering.

The principal measured quantities in such EPR studies are the *g* value for the magnetic ions and the resonance linewidth. Both are usually functions of temperature and ion concentration in the metal. Because the ion moment interacts with the conduction electrons, the *g* value departs somewhat from the free-ion value by an amount Δg, given in first approximation by $\Delta g = JN(E_F)$, where J is the ion moment–conduction electron exchange integral and $N(E_F)$ is the conduction-electron density of states at E_F. EPR is one of the best methods for the measurement and interpretation of J.

4. SPIN RELAXATION PHENOMENA AND APPLICATIONS

As noted in Sec. 1.2, the interaction of spins with their surroundings, the "lattice," maintains the system of magnetic moments (the *spin system*) in thermal equilibrium at a temperature only slightly higher than that of the lattice while subjected to the microwave radiation field. If this did not occur, the resonance would quickly become saturated and cease to be observable. In some systems that have long spin–lattice relaxation times, it is possible to shift the field H_0 suddenly from off resonance to resonance and to observe the signal decay to zero as saturation occurs. In general, information gained from measurements of the relaxation times, T_1 and T_2, relates to the *dynamic* (fluctuating in time) components of the same interactions responsible for the *static* features of the spectrum.

4.1 Spin–Lattice Relaxation

Both steady-state (cw) and pulse methods are used to measure T_1. When the resonance linewidth, ΔH, is dominated by T_1, i.e., lifetime broadening, we have $\Delta H \propto 1/T_1$, and measurement of the temperature dependence of ΔH yields that of T_1. This method is useful for T_1's in the approximate range 10^{-8}–10^{-10} s, since otherwise either the resonance is too broad to detect or ΔH no longer depends on T_1.

In the *progressive saturation* method, the dependence of the absorption signal amplitude χ'' of Eq. (17) and its first derivative, $d\chi''/dH$, on H_1 is monitored. The presence of the saturation term $\gamma^2 H_1^2 T_1 T_2$ in the denominator of Eq. (17) leads to a linear dependence of $d\chi''/dH$ on H_1 (hence on the square root of the microwave power) for small H_1, after which $d\chi''/dH$ passes through a maximum and decreases as H_1 increases. For a homogeneously broadened line, for which T_2 may be determined from the linewidth, T_1 is then determined from the dependence of $d\chi''/dH$ on $\gamma^2 H_1^2 T_1 T_2$.

Pulse methods for measuring T_1 are based on Eq. (13) for the return of M_z to its equilibrium value M_0 following a sudden change. The simplest scheme utilizes a 90°–τ–90°–FID pulse sequence in which the first 90° pulse [see Eq. (7)] reduces M_z to zero. After a delay time τ, the recovered magnetization is given by the amplitude of the FID following the second 90° pulse. By varying τ, the time evolution of the recovery is recorded, determining T_1.

For paramagnetic ions in solids, spin–lattice relaxation (SLR) (see Sec. 1.2) results from fluctuations in the crystalline electric field (CEF) (see Sec. 3.2) that couple electrostatically to the orbital motion of the ion's magnetic electrons. CEF fluctuations are caused by lattice vibrations (phonons) that displace ions from their equilibrium lattice sites (both the magnetic ions and their nonmagnetic neighbors). The component of these fluctuations at the Larmor frequency induces transitions that transfer energy to the lattice. When an energy quantum $\hbar\omega_0$ is transferred to the lattice, a phonon of that energy is created. Finally, lattice vibrations also transfer energy to the sample's surroundings (the *bath*) which may be a cryogenic bath (e.g., liquid helium), a temperature-controlled chamber, or simply the ambient temperature in the laboratory. The rate of energy transfer from lattice to bath must exceed that from spins to lattice. If this is not the case, e.g., when the concentration of paramagnetic ions is too great, a *phonon bottleneck* occurs, and T_1 cannot be reliably determined.

The simplest relaxation mechanism is the *direct*, or *one-phonon* process in which $T_1 \propto 1/T$. This process is most important at temperatures in the liquid helium range (~ 4 K), since then the phonon distribution peaks in the range of microwave frequencies. *Two-phonon* processes become increasingly important at higher temperatures. In one of these, the *Raman* process, a phonon at frequency ω_1 is absorbed while a second at ω_2 is emitted, the difference $(\omega_2 - \omega_1)$ equaling the Larmor frequency ω_0. Although the intrinsic probability for two-phonon interaction is much less than for the direct process, there are a far greater number of ways in which the two phonons can be chosen. The Raman process yields a temperature dependence of the form $T_1 \propto 1/T^9$. The *Orbach* process is a resonant two-phonon process, involving an excited level of the magnetic ion at an energy Δ above the ground state. The temperature dependence has the form $T_1 \propto \exp(\Delta/k_B T) - 1$. This process can dominate the direct process even at low temperatures.

To summarize, the temperature dependence of the relaxation rate $1/T_1$ consists generally of three terms, the direct, Raman, and Orbach processes:

$$\frac{1}{T_1} = AT + BT^9 + C[\exp(\Delta/k_B T) - 1]^{-1}. \quad (20)$$

4.2 Spin–Spin Relaxation

With a pulse spectrometer, the spin–spin relaxation time T_2 may be measured using the *spin-echo* technique. A 90° pulse [Sec. 1.3 and Eq. (7)] tips the magnetization into the xy plane where the slightly different precession frequencies of the individual moments, due to inhomogeneities in H_0, cause them to fan out as time progresses. Application of a 180° pulse after a time interval τ reverses this process, so that the moments become *refocussed* at time 2τ, yielding an *echo* signal. As the individual moments precess in the xy plane, they experience random relaxation processes that prevent their becoming fully refocussed. The longer the time τ, the greater the loss of coherence among the spins, and the weaker the echo. This process can be followed by applying a succession of 180° pulses at equal intervals 2τ after the initial 180° pulse, as shown in Fig. 12. Pulse lengths in typical electron spin-echo (ESE) experiments range from 10 to 100 ns.

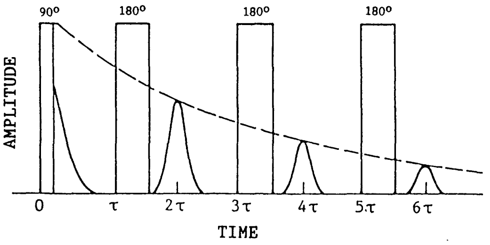

FIG. 12. Spin-echo pulse sequence consisting of an initial 90° pulse followed at equal time intervals by a succession of 180° pulses, each of which yields an echo signal. The envelope of the echo amplitudes (dashed line) decays exponentially with time constant T_2.

4.3 Exchange Interactions

The linewidth ΔH of the EPR signal of solid DPPH (Sec. 3.1) is about 0.2 mT (2 Gauss), although electrons on neighboring radicals are only ~ 6 Å apart and would be expected to give rise to a $\Delta H \propto \beta/r^3 \simeq 10$ mT. Similar reductions in ΔH are also typical of undiluted concentrations of paramagnetic ions in solids. This phenomenon of *exchange narrowing* may be understood in terms of the simultaneous reversal of antiparallel spins on adjacent sites. This corresponds to rapid movement of spins of given orientation through the solid lattice (*spin diffusion*), which averages the dipolar field seen by any specific electron moment to zero. Exchange narrowing results in a nearly Lorentzian line shape, rather than Gaussian, with intensity shifted from the center to the wings of the line.

No energy transfer occurs between spins and lattice, and so exchange narrowing is a T_2 process. Coulomb interactions between the electrons of two neighboring ions are responsible, leading through the Pauli exclusion principle to an effective interaction between spins of the form $-2J_{ij}\mathbf{S}_i \cdot \mathbf{S}_j$, where J_{ij}, the exchange interaction due to electrons on adjacent ions, is inversely proportional to the distance between the electrons.

The width and position of an EPR line are sensitive to *chemical exchange* if the resonance frequency of the unpaired electron changes when the exchange occurs. For free radicals, this may be caused by changes in the g factor, hyperfine splitting, or both. A chemical reaction in which a paramagnetic entity P associates and disassociates with a diamagnetic entity D can be characterized by the mean lifetimes, τ_P and τ_{PD}, of the two species P and PD. The exchanges themselves are assumed to occur instantaneously.

When τ_P and τ_{PD} are very long (slow exchange), the EPR spectrum consists of two lines, for P and for PD, respectively, with intensities proportional to the concentrations of P and PD. As the lifetimes become shorter, the lines acquire an additional broadening, $\Delta\omega$, over their natural widths, due to the exchange. This is a form of lifetime broadening (Sec. 1.2) with $\Delta\omega_P=\tau_P^{-1}$ and $\Delta\omega_{PD}=\tau_{PD}^{-1}$. For free radicals in solution, such linewidth changes may become apparent if $\tau < 10^{-5}$ s. At the other extreme of very short τ values (fast exchange), the unpaired electrons are unable to follow the changes in the local fields, and a single line occurs at the average resonance frequency determined by the average (i.e., population weighted) g factor, with a linewidth determined by the average of the natural linewidths. As the exchange rate slows down, this line is broadened by the changing g value, with the contribution to the width given by

$$(\Delta\omega)_{\text{exch}} \propto (\Delta\omega)^2\tau \tag{21}$$

with

$$\Delta\omega=\omega_P-\omega_{PD}=(\beta H/\hbar)(g_P-g_{PD}), \tag{22}$$

and mean lifetime τ given by

$$\tau^{-1}=\tau_P^{-1}+\tau_{PD}^{-1}. \tag{23}$$

4.4 Chemically Induced Dynamic Electron Polarization

In *chemically induced dynamic electron polarization* (CIDEP), radicals are produced with nonequilibrium spin-state populations, so that the intensity ratios of hyperfine components differ from the normal binomial-type distribution. This can happen when an excited triplet state is produced having a preferential population of certain M levels. Populations may depart so far from the Boltzmann distribution that some EPR lines appear in emission. In a typical CIDEP experiment a short (20 ns–2 μs) pulse of radiation generates free radicals in a sample located in the resonance cavity of an EPR spectrometer, which serves to monitor changes in the dynamic magnetic susceptibility of the sample as a function of time as the radicals are created and destroyed. In *photolytic* experiments the radiation pulse is electromagnetic (e.g., light, x rays, etc.), and in *radiolytic* experiments it is particle radiation (e.g., an electron beam). The resulting polarized transient signals are detected in a time-resolved experiment using a cw EPR spectrometer having a sufficiently fast response time, or by the electron spin-echo technique (ESE) (Sec. 4.3 and Fig. 12) using a pulse spectrometer.

Schematic examples of polarized transients detected in a time-resolved experiment are shown in Fig. 13. The signals decay quickly as the polarized spin system returns to equilibrium in a time that is essentially the spin–lattice relaxation time T_1 followed afterward by the slower decay of the normal absorption signal as the radicals are destroyed by chemical reaction. The polarization factor γ is defined as

$$\gamma=(|S|-|S_0|)/|S_0| \quad \text{for enhanced absorption} \tag{24}$$

and as

$$\gamma=-(|S|+|S_0|)/|S_0| \quad \text{for enhanced emission.} \tag{25}$$

Radicals may be detected with three different polarization patterns: mixed emission–enhanced absorption, pure emission, and pure enhanced absorption.

When using the ESE technique, the 90° pulse of the echo sequence (Fig. 12) occurs after a delay time t_d following a laser light flash which initiates the photolytic reaction. The echo signal following the first 180° pulse can be used to monitor the evolution of the polarization by varying the delay time t_d, keeping the pulse spacing τ constant. This evolution is determined by the rates of creation and decay of the paramagnetic species and the T_1's among the spin sublevels. This measurement determines the preferential routes by which the spin levels are populated as the photoexcited triplet state is formed.

5. ADVANCED TECHNIQUES AND APPLICATIONS

5.1 Electron-Nuclear Double Resonance (ENDOR)

In electron-nuclear double resonance (ENDOR), an EPR transition is driven nearly to saturation by applying a high microwave power level. This condition is monitored while simultaneously exciting another (nuclear) transition with a so-called pumping field. At pump frequencies related to the nuclear hyperfine interaction, the saturation condition will be relaxed and a nuclear resonance detected. In one method, the intensity of an EPR line is monitored while radio-frequency (rf) power is applied to the sample via a small auxiliary coil wrapped around the sample inside the microwave cavity. When the rf matches the separation of a pair of hyperfine levels, it induces a *nuclear* transition, $\Delta m_I = \pm 1$ (M remains unchanged). This tends to equalize the populations of the two hyperfine levels, and restores a population difference between the originally saturated levels, causing an increase in the EPR signal strength. This is an ENDOR signal. As the radiofrequency is scanned, an ENDOR signal appears on each side of the pure nuclear magnetic resonance (NMR) frequency, $\omega_N = \gamma_n H_0$, and the spacing in frequency of the ENDOR signals is exactly the hyperfine interaction frequency a.

Figure 14 shows the ENDOR lines of the nearest ^{19}F neighbor nuclei to an F center in the alkaline earth fluorohalide SrFCl (Spaeth and Koschnick, 1991). The ENDOR signals appear as increases above the noise baseline of the nearly saturated normal EPR line. Crystals such as SrFCl are important as x-ray storage phosphor materials, and have scientific interest because, since both halide vacancies can trap electrons, two different F centers having different local symmetries can be formed in the same crystal. In the example shown, the F center electron is trapped in a Cl$^-$ ion vacancy, and its hyperfine interaction is with the nearest ^{19}F nuclei.

5.2 Optically Detected Magnetic Resonance (ODMR)

A variation in the intensity of light absorbed or emitted by a material while scanning through an EPR absorption indicates that the optical transition involves particular magnetic sublevels. The variation in light intensity can be utilized to detect EPR, and the technique is known as *optically detected magnetic resonance (ODMR)*. For example, this

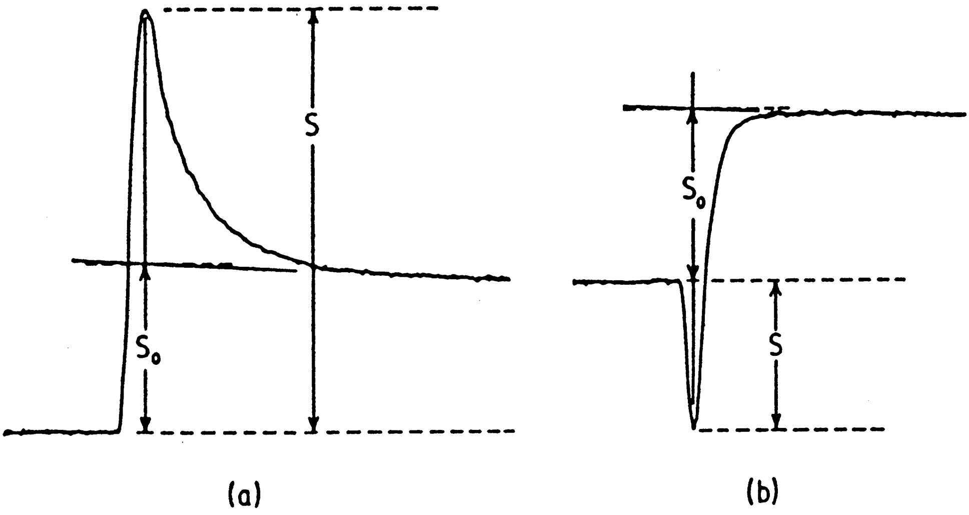

FIG. 13. Schematic (a) enhanced absorption and (b) emission transient CIDEP signals. The signal amplitudes S and S_0 are evaluated at the same point in time, so that the decay curve of the radical is extrapolated back to allow for radical decay.

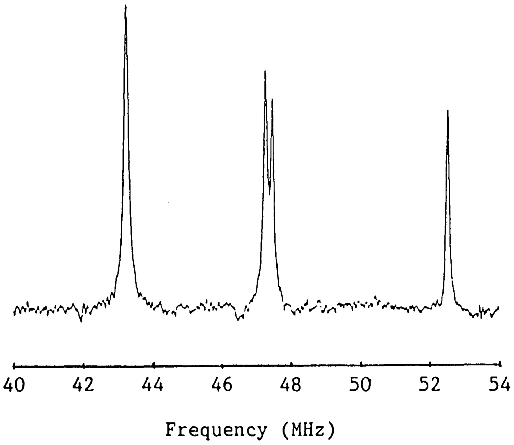

FIG. 14. ENDOR spectrum of nearest ^{19}F neighbors of the F(Cl^{-}) center in SrFCl. The frequency scale is that of the applied radio-frequency radiation, and the signals appear as increases above the noise baseline of the nearly saturated EPR signal (Spaeth and Koschnick, 1991).

may be done by monitoring an optical emission line of the material as the resonance condition is scanned, in which case the experiment yields the spin-Hamiltonian parameters of the excited optical state. Since optical spectroscopic techniques have intrinsically far greater sensitivity than microwave techniques, ODMR can greatly increase sensitivity, so that as few as 10^4 excited state spins may be detected. Optical detection may also be combined with ENDOR, yielding ODENDOR.

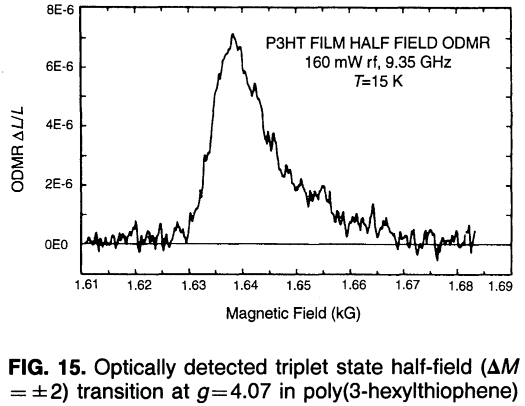

FIG. 15. Optically detected triplet state half-field ($\Delta M = \pm 2$) transition at $g = 4.07$ in poly(3-hexylthiophene) film at 15 K. The signal is detected by the enhanced photoluminescence, $\Delta L/L$, as the resonance condition is traversed by scanning the applied field, H_0 (Swanson *et al.*, 1990).

Shown in Fig. 15 is the $\Delta M = \pm 2$ (half-field) transition (see Sec. 3.2) of the triplet state exciton in the semiconducting polymer, poly(3-hexylthiophene) (P3HT), obtained by ODMR at 15 K (Swanson *et al.*, 1990). These polymers are photoluminescent, emitting light in the red region of the spectrum when illuminated with blue laser light. The resonance was recorded by monitoring the enhanced photoluminescence, $\Delta L/L$, of the sample as the magnetic field was scanned through the resonance condition at $g = 4.07$.

6. SUMMARY

Electron paramagnetic resonance (EPR) is a spectroscopic technique in which transitions between energy levels of weakly coupled electronic magnetic dipole moments in a magnetic field are detected. Such electronic moments occur in solids, liquids, and gases, and may belong to conduction electrons in metals, impurity ions in various hosts, organic free radicals, paramagnetic molecular species, and radiation-induced defects, for example. In typical laboratory magnetic field strengths on the order of a few teslas, these transitions (resonances) occur in the microwave region (5–50 GHz) of the radiofrequency spectrum. Electromagnetic interactions shift and split the simple energy-level scheme expected solely from interaction with the applied magnetic field. These include electron–electron and electron–nuclear interactions, as well as interactions of electronic orbital moments with the local electrostatic potential, each furnishing its own "signature" to the EPR spectrum that contributes to identifying and characterizing the origin and environment of the paramagnetic entity. The dynamic properties of the system of electronic moments are reflected in the rates at which energy is exchanged among the moments and between the moments and the host substance. These rates may also be measured and yield information on the nonstatic components of the interactions listed above. Examples discussed in this review include applications of EPR to the study of free radicals, impurity ions in crystalline solids, color centers, hydrogenated amorphous silicon, hydrogen in metals, and conducting polymers, among others.

GLOSSARY

Absorption Mode: Signal detected in an EPR spectrometer due to absorption of radio-frequency energy by electronic moments at resonance.

Bloch Equations: Phenomenological equations describing motion of magnetization vector in a magnetic field and subject to relaxation processes.

Chemical Exchange: Association and dissociation of paramagnetic and diamagnetic entities in a chemical reaction resulting in changes in the shift and/or width of the resonance.

CIDEP: Acronym for chemically induced dynamic electron polarization, experiments in which free radicals are produced with nonequilibrium spin-state populations.

Color Center: An electron trapped at a vacancy in an alkali halide (or similar) crystal lattice.

Contact Interaction: Magnetic hyperfine interaction of a nuclear dipole moment with the magnetic moment of an electron in an *s*-orbital which penetrates the nucleus.

Crystalline Electric Field: The electric field at a particular site in a solid arising from the electric charges on all surrounding ions.

Dipole–Dipole Interaction: Energy of a magnetic dipole moment in the magnetic field of a second dipole moment.

Dispersion Mode: Signal detected in an EPR spectrometer due to the frequency shift of the resonant circuit element by electronic moments at resonance.

ENDOR: Acronym for Electron-Nuclear DOuble Resonance, an experimental arrangement in which hyperfine structure in an EPR spectrum is detected by inducing a nuclear spin transition while saturating and monitoring the EPR transition.

Exchange Narrowing: Narrowing of the resonance linewidth due to the simultaneous reversal of antiparallel spins on adjacent sites, corresponding to spin diffusion.

Fine Structure: Splitting of an EPR transition into several transitions due to electronic spin–spin or spin–orbit interactions.

Fourier Transform Spectroscopy: Spectroscopy accomplished by Fourier transforming the time-domain response of a system to a short pulse of radiation at the resonance frequency into the frequency domain.

Free-Induction Decay: Signal induced in a resonant circuit element (cavity) by the precessing magnetization of a system of electronic moments following the application of a pulse of resonance radiation that turns the magnetization through an angle of 90°.

***g* Factor:** Ratio of the magnetic dipole moment in units of the Bohr magneton to its associated mechanical angular momentum in units of Planck's constant over 2π.

Half-Field Transition: A normally forbidden transition that occurs between the $+1$ and -1 levels of a spin-1 system, characterized by an effective g value of 4, hence occurring at half the normal resonance field strength.

Homogeneous Broadening: Broadening of an EPR line wherein all parts of the resonance line are strongly coupled to one another so that strong radio-frequency irradiation near the line center saturates the entire line.

Hyperfine Interaction: Mutual interaction energy of an electronic magnetic moment and a nuclear magnetic moment.

Hyperfine Structure: Splitting of an EPR transition into several transitions caused by a hyperfine interaction.

Inhomogeneous Broadening: Broadening of an EPR line that is an envelope of many narrow, unresolved homogeneous component lines, so that strong radio-frequency irradiation saturates only those components at the applied frequency (hole burning).

Larmor Frequency: The frequency at which a magnetic dipole moment precesses in a magnetic field.

Lattice: Term applied to all energies and degrees of freedom of a host material except for those of the system of electronic moments.

Longitudinal Relaxation Time: Characteristic time in which the component of the electronic magnetization parallel to the applied field returns to its thermal equilibrium value following a disturbance.

Magnetogyric Ratio: Ratio of the magnetic dipole moment to its associated mechanical angular momentum.

ODMR: Acronym for Optically Detected Magnetic Resonance, an experimental arrangement in which EPR is detected by changes in optical absorption or emission of the sample.

One-Phonon Process: Spin–lattice relaxation process in which a single energy quantum at the Larmor frequency is transferred from the electronic moment system to a lattice vibration.

Orbach Process: A resonant two-phonon spin-lattice relaxation process involving an excited level of the magnetic ion at an energy splitting above the ground state.

Pauli Spin Susceptibility: Magnetic susceptibility of the conduction electrons in a metal.

Phase-Coherence Time: Characteristic time for two magnetic moments initially precessing in phase with one another to become out of phase.

Phonon Bottleneck: Inability of lattice vibrations (phonons) to transfer energy to the sample's surroundings as fast as they receive it from relaxing electronic moments.

Photolytic CIDEP: CIDEP experiments in which an electromagnetic radiation pulse generates a nonequilibrium spin-state population.

Powder Pattern: EPR spectrum obtained from a polycrystalline (or amorphous) sample in which resonances arise simultaneously from crystallites at all orientations with respect to the applied field.

Precession: Regular rotation of the spin axis of angular momentum at a fixed orientation around the direction of the applied field.

Radiolytic CIDEP: CIDEP experiments in which a particle radiation pulse generates a nonequilibrium spin-state population.

Raman Process: Spin–lattice relaxation process in which a lattice vibration quantum (phonon) of one frequency is absorbed while a second at another frequency is emitted, the difference frequency of the two being equal to the Larmor frequency.

Resonance: Condition when frequency of applied electromagnetic radiation matches the natural (Larmor) precession of the magnetic dipole moment.

Saturation: Condition that results when the populations of two energy levels between which a transition occurs become equalized so that no further net absorption or emission of energy can occur.

Skin Depth: Characteristic distance in which an electromagnetic wave in a metal is attenuated to $1/e$ of its incident intensity.

Spin Diffusion: Spatial diffusion of spin orientation resulting from spin flips of adjacent spins caused by their mutual dipole–dipole interaction.

Spin-Echo: EPR signal that results when precessing spins (moments) are refocussed following a particular sequence of radiofrequency pulses.

Spin Hamiltonian: Quantum-mechanical energy operator constructed from spin operators, yielding an energy level scheme that accounts for an observed EPR spectrum.

Spin Label: Chemical group containing an unpaired electron spin that can be attached to other groups or molecules.

Spin–Lattice Relaxation: Restoration of thermal equilibrium populations of spin levels caused by exchange of energy between spins (moments) and host structure (lattice) following a disturbance.

Spin–Spin Relaxation: Restoration of thermal equilibrium values of the transverse (perpendicular to applied field) components of the magnetization following a disturbance.

Steady-State Spectroscopy: Detection of EPR transitions by methods involving negligible disturbance of equilibrium energy level populations.

Zero-Field Splitting: Any splitting of electronic moment energy levels arising from interactions that are independent of the applied magnetic field strength, e.g., crystalline electric field interactions or dipole–dipole interactions.

Works Cited

Bagguley, D. M. S., Griffiths, J. H. E. (1947), *Nature* **160**, 532–540.

Bloch, F., *et al.* (1946), *Phys. Rev.* **69**, 127–145.

Biegelson, D. K (1981), in: E. N. Kaufman and G. K. Shenoy (Eds.), *Nuclear and Electron Resonance Spectroscopies Applied to Materials Science,* Vol. 3, New York: Elsevier North Holland, pp. 85–94.

Griffith, O. H., Waggoner, A. S. (1969), *Accts. Chem. Res.* **2**, 17–24.

Holton, W. C., Slichter, C. P. (1960), *Phys. Rev. Lett.* **5**, 197–200.

Hutchison, C. A., Jr., Mangum, B. W. (1961), *J. Chem. Phys.* **34**, 908–922.

Spaeth, J.-M., Koschnick, F. K. (1991), *J. Phys. Chem. Solids* **52**, 1–33.

Swanson, L. S., *et al.* (1990), *Phys. Rev. Lett.* **65**, 1140–1143.

Venturini, E. L., Richards, P. M. (1979), *Solid State Commun.* **32**, 1185–1188.

Watterich, A., *et al.* (1991), *J. Phys. Chem. Solids* **52**, 449–455.

Zavoisky, J. (1945), *J. Phys. USSR* **9**, 211–214.

Further Reading

Introductory

Melissinos, A. C. (1966), *Experiments in Modern Physics,* New York: Academic, Chap. 8.

General Texts

Atherton, N. M. (1973), *Electron Spin Resonance: Theory and Applications,* New York; Wiley.

Bolton, J. R., Wertz, J. E. (1972), *Electron Spin Resonance, Elementary Theory and Practical Applications,* New York: McGraw-Hill.

Pake, G. E., Estle, T. (1973), *The Physical Principles of Electron Spin Resonance,* Reading, MA: Addison-Wesley.

Slichter, C. P. (1978), *Principles of Magnetic Resonance, 2nd Ed.,* Berlin: Springer-Verlag.

Specialized

Abragam, A., Bleaney, B. (1970), *Electron Paramagnetic Resonance of Transition Ions,* Oxford: Clarendon.

Berlinger, L. J. (Ed.) (1976), *Spin-Labeling, Theory and Practice,* New York: Academic.

Bolton, J. R., Borg, D. C., Schwartz, H. M. (1971), *Biological Applications of Electron Spin Resonance,* New York: Wiley-Interscience.

Brey, W. S. (Ed.) (1969), *The Journal of Magnetic Resonance,* New York, Academic.

Kaufman, E. N., Shenoy, G. K. (Eds.) (1981), *Nuclear and Electron Resonance Spectroscopies Applied to Materials Science,* New York: Elsevier North Holland.

Clarke, R. H. (Ed.) (1982), *Triplet State ODMR Spectroscopy,* New York: Wiley-Interscience.

Kevan, L. Schwartz, R. N. (Eds.) (1979), *Time Domain Electron Spin Resonance,* New York: Wiley-Interscience.

Lepley, A. R., and Closs, G. L. (Eds.) (1973), *Chemically Induced Magnetic Polarization,* New York: Wiley-Interscience.

Muus, L. T., Atkins, P. W. (Eds.) (1972), *Electron Spin Relaxation in Liquids,* New York: Plenum.

Poole, C. P., Jr. (1983), *Electron Spin Resonance, A Comprehensive Treatise on Experimental Techniques,* New York: Wiley-Interscience.

Poole, C. P., Jr., Farach, H. A. (1972), *Theory of Magnetic Resonance,* New York: Wiley-Interscience.

Poole, C. P., Jr., Farach, H. A. (1971), *Relaxation in Magnetic Resonance,* New York: Academic.

Standley, K. J., Vaughan, R. A. (1969), *Electron Spin Relaxation Phenomena in Solids,* New York: Plenum.

Warren, W. S. (Ed.) (1965), *Advances in Magnetic and Optical Resonance,* New York: Academic.

8
Electron Scattering by Atoms, Ions, and Molecules

Klaus Bartschat, Philip G. Burke and Albert Crowe

8.1 Introduction 211
8.2 Review of Experimental Methods 212
8.2.1 The Electron Spectrometer 212
8.2.2 Elastic Scattering 213
8.2.2.1 Differential Cross Sections 213
8.2.3 Excitation of Discrete States 215
8.2.3.1 Differential Cross Sections 215
8.2.3.2 Integral Cross Sections 215
8.2.3.3 Angular and Polarization Correlations 216
8.2.4 Ionization 220
8.2.4.1 Total Cross Sections 220
8.2.4.2 Double Differential Cross Sections 221
8.2.4.3 Higher Order, including, Fully Differential Cross Sections 222
8.2.4.4 (e,2e) Studies of Direct Single Ionization 223
8.2.4.5 (e,3e) Studies of Double Ionization 226
8.2.4.6 Excitation–Ionization 227
8.2.4.7 Excitation–Autoionization 229
8.2.4.8 Inner-Shell Ionization 230
8.2.5 Electron–Molecule/Biomolecule Interactions 230
8.2.6 Electron–Ion Interactions 231
8.2.7 Spin-Polarized Electron Studies 233
8.2.7.1 Exchange Interaction 234
8.2.7.2 Spin–Orbit Interaction 235
8.2.7.3 A Typical Apparatus and Some Illustrative Results 236
8.2.8 Positron Scattering 238
8.3 Review of Theoretical Methods 238
8.3.1 The Close-Coupling Expansion 238
8.3.1.1 The Wave Equation, Scattering Amplitude, and Cross Section 238
8.3.1.2 Extension of the Close-Coupling Method to Intermediate and High Energies 242
8.3.1.3 Inclusion of Relativistic Effects 243

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

8.3.1.4 Illustrative Results 244
8.3.1.5 Resonances and Quantum Defect Theory 247
8.3.1.6 Dielectronic Recombination 250
8.3.2 Born Series Methods 251
8.3.2.1 Elastic Scattering and Excitation of Atoms and Ions at Intermediate and High Energies 251
8.3.2.2 The Born Series 251
8.3.2.3 The Distorted-Wave (DW) Method 252
8.3.3 Initial-Value Methods 254
8.3.3.1 Time-Dependent Close-Coupling (TDCC) 254
8.3.3.2 Exterior Complex Scaling (ECS) 255
8.3.4 Ionization of Atoms and Ions 256
8.3.4.1 The Ionization Amplitude and Cross Section 256
8.3.4.2 Threshold Behavior of the Ionization Cross Section 258
8.3.4.3 Ionization at Intermediate and High Energies 259
8.3.4.4 Excitation–Autoionization 261
8.3.5 Scattering by Molecules 262
8.3.5.1 Laboratory Frame Representation 262
8.3.5.2 Molecular Frame Representation 263
8.3.5.3 Inclusion of the Nuclear Motion 265
8.3.5.4 Illustrative Results for Molecules 270
8.4 Conclusions and Future Work 274
References 276
Further Reading 281

8.1
Introduction

The experimental study of electron–atom and electron–molecule collisions dates back to the beginning of the 20th century Lenard (1903). Later in 1914, Nobel prize winners Franck and Hertz (1914) demonstrated the energy loss of electrons scattered from mercury vapor. Theoretically, electron scattering by atoms and molecules has attracted considerable attention since the earliest days of quantum mechanics, because such processes provide a means of investigating the dynamics of many-particle systems at a fundamental level. In addition, a detailed understanding of these processes is required in many other fields, particularly in astrophysics, plasma physics, laser physics, electrical discharges, ionospheric processes in planetary atmospheres, isotope separation, and radiation chemistry and physics.

In recent years, a number of important advances have been made both on the experimental and on the theoretical side. On the experimental side, these include absolute measurements of cross sections, experiments using coincidence techniques, spin-polarized projectile beams and targets, high-energy-resolution electron beams, and bright beams of positrons. On the theoretical side, many new developments have been stimulated by the increasing availability of powerful computers that allow for detailed calculations to be carried out for complex atoms, ions, and molecules.

This review is an update to the article by Burke and Crowe (1993). As previously, we first describe recent experimental work in Section 8.2, which is followed by the theory of electron scattering by atoms, ions, and molecules in Section 8.3. We further compare theoretical predictions with recent experimental data for a few representative cases. Within the page limit of this article, we can mention only a small fraction of the ongoing work in this field. Interested readers are referred to journals such as The Physical Review A or Journal of Physics B, which contain extensive sections on electron collisions, as well as to the Proceedings and Book of Abstracts of the biannual “International Conference on Photonic, Electronic, and Atomic Collisions” (ICPEAC). Finally, in Section 8.4, we draw some conclusions and mention the areas where future work is needed.

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

8.2
Review of Experimental Methods

8.2.1
The Electron Spectrometer

The basic workhorse of a large number of electron scattering studies is the electron spectrometer. In principle, free electrons can be formed into a beam and their energy selected using various combinations of electrostatic and magnetic fields. In practice, the use of electrostatic fields is more common, these being more easily controlled and shielded than their magnetic counterparts. This is particularly important where it is essential to preserve the direction of low-energy electrons following the collision process.

We show in Figure 8.1 a recent example of a spectrometer that combines the characteristics of a conventional electrostatic device with a recent innovation Allan (2004). Basically, the electron gun consists of a source of electrons produced by thermionic emission from a heated filament. The electrons are collimated and focused by an electrostatic lens system onto the input aperture of a double hemispherical energy selector. Those electrons within a narrow band of energies satisfying the criteria for transmission through the selector are then focused on the gas beam produced by a nozzle arrangement. Scattered electrons from the interaction region traveling in the direction of the scattered electron analyzer are similarly focused onto the input aperture of its double hemispherical analyzer, the transmitted electrons being focused onto a single-channel electron multiplier detector.

One drawback of conventional electron spectrometers is that the angular range of the electron analyzer is limited by the physical presence of other components of the spectrometer. This was overcome by Read and Channing (1996) who applied a localized static magnetic field to the interaction region of a conventional spectrometer. The incident electron beam and the scattered electrons are respectively steered to and from the interaction region through angles set by the field (hence the common name "magnetic angle changer" ("MAC")). This steering means that electrons normally scattered into inaccessible scattering angles are rotated into the accessible angular range of the electron analyzer while the magnetic field design is such that it leaves the angular distribution of the electrons undistorted. The spectrometer shown in Figure 8.1 has a MAC fitted, thereby enabling the full angular range $0^\circ - 180^\circ$ to be accessed.

Another important development in the study of small and/or multidifferential cross sections is the increasing use of multidetection methods, usually in conjunction with toroidal analyzers. These enable simultaneous detection of scattering over a wide range of angles. Recent examples are the spectrometers of Catoire *et al.* (2007) and of Lower *et al.* (2007).

Perhaps the most novel and exciting recent experimental development in electron–atom/molecule scattering is the reaction microscope described by Ullrich *et al.* (2003). In contrast to conventional electron spectrometers, it uses recoil ion and electron momentum spectroscopy to measure the vector momenta of outgoing charged particles over large (close to 4π) solid angles.

These recent experimental innovations have been applied in studies of

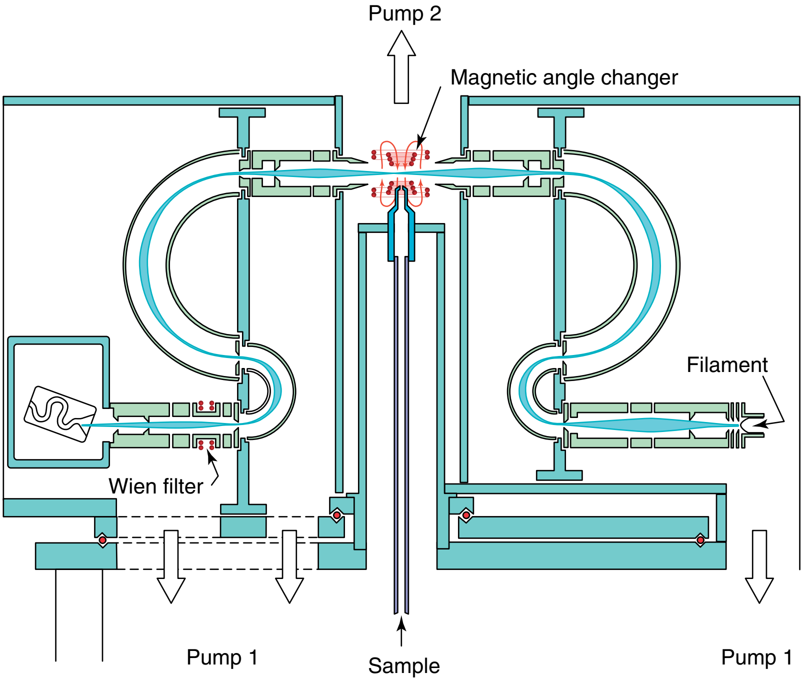

Fig. 8.1 The electron spectrometer of Allan (2004).

ionization and are discussed further in Section 8.2.4.3.

8.2.2
Elastic Scattering

Elastic scattering is characterized by the absence of energy transfer from the incident electron to the internal energy of the atom.

8.2.2.1 Differential Cross Sections

Differential elastic scattering studies are normally carried out using an electron spectrometer broadly similar to that shown in Figure 8.1. The electron analyzer is tuned to transmit only the electrons having an energy equal to the incident electron energy. Relative angular differential cross sections are determined by observation of the scattered electron signal, as the electron analyzer is rotated about the reaction region. The elastic scattering differential cross section $d\sigma\ (E,\theta)/d\Omega$ can be written as

$$\frac{d\sigma}{d\Omega}(E,\theta) = \frac{dN_S(E,\theta)}{N_i\, n\, l\, d\Omega} \qquad (8.1)$$

where $dN_S(E,\theta)$ is the number of electrons with energy E per unit time scattered into a solid angle $d\Omega$ at a scattering angle θ with respect to the incident electron beam direction, and N_i is the number of electrons per unit time incident on the target gas of number density n in a scattering cell of length l.

Various techniques have been used to place the relative cross sections on an absolute cross-section scale. The most direct method of making the cross section absolute is to measure all of the quantities on the right-hand side of Eq. (8.1). However, this is rarely done owing to the extreme difficulties in measuring both these quantities and the various efficiencies involved.

In recent years, the relative flow technique has been the most widely used method for placing measured relative cross sections on an absolute scale. The unknown cross sections are determined by comparing the measured electron scattering intensities against those of another species with known cross sections under identical experimental conditions. Details of the technique are given by Brunger and Buckman (2002) while the most recent developments are described by Khakoo *et al.* (2007).

Helium elastic scattering cross sections are the most commonly used standards. They have been accurately measured and put on an absolute cross-section scale using a phase shift analysis; see, for example, Andrick (1973). The measured values are accurately reproduced by the convergent close-coupling (CCC) theory of Fursa and Bray (1997) and the *R*-matrix with pseudostates (RMPS) method of Bartschat *et al.* (1996a). For energies up to 19 eV, that is, below the first excitation threshold, accurate variational calculations for the s-wave and p-wave phaseshifts were already performed 30 years ago by Nesbet (1979).

In Figure 8.2, we show the elastic differential scattering cross section for xenon at an electron energy of 30 eV. One of the two experiments reported by Cho *et al.* (2006) uses a MAC to cover the full range of scattering angles and the relative flow method has been used to put the data on an absolute cross-section scale. Their theories show

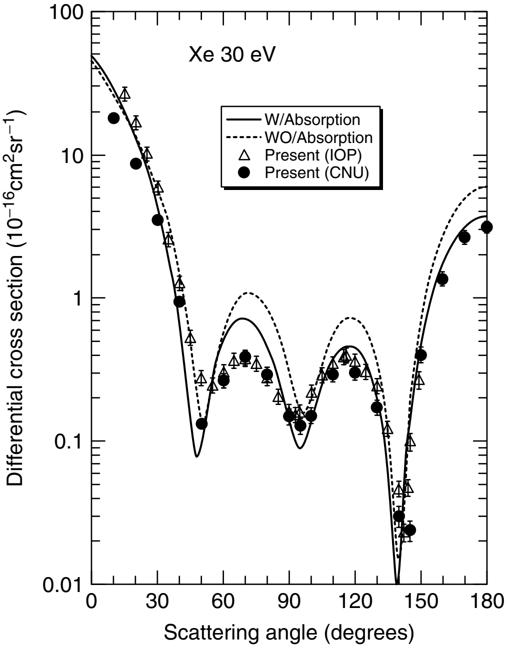

Fig. 8.2 The absolute differential elastic scattering cross section for xenon at an incident electron energy of 30 eV. (From Cho *et al.* (2006).)

the importance of including absorption effects in the theoretical description.

8.2.3
Excitation of Discrete States

The excitation process for an atomic target can be written as

$$e^-(E_i) + A_i \longrightarrow e^-(E_j) + A_j \qquad (8.2)$$

where, in atomic units, $E_i = \frac{1}{2}k_i^2$ is the incident electron energy while $E_j = \frac{1}{2}k_j^2$ is the scattered electron energy; k_i and k_j are the respective wave numbers. (In this and later equations, we use atomic units, $\hbar = m = e = 1$, where m is the mass of the electron and $-e$ is its charge.)

8.2.3.1 **Differential Cross Sections**

The scattered electron studies are conducted as for elastic scattering, except that the electron analyzer is tuned to accept electrons with energy E_j. Angular differential cross sections can then be measured for a particular excited state. An example of an angular differential cross section is shown in Figure 8.6 for the 2^1P state of helium.

For spectroscopic studies, the measurement of energy-loss spectra, where the energy of the detected electrons is varied for a fixed incident electron energy or alternatively the detected electron energy is fixed for a varying incident electron energy, is a useful technique. Electron spectroscopy is particularly valuable for the study of optically forbidden transitions.

8.2.3.2 **Integral Cross Sections**

When the excited atom is metastable, information on the excitation process can be obtained in a relatively straightforward way by making use of the fact that it can eject an electron from a surface whose work function is less than the internal energy of the atom. Time-of-flight (TOF) techniques are normally used to isolate the neutral-atom signal from that due to photons. The method has the advantage that the energy resolution of the measurements is determined by the incident electron beam resolution only. Hence, it has been used as a convenient method for the study of resonance features contributing to the excitation process, see Section 8.3.1.4, Figure 8.21. It has the disadvantage that it is difficult to isolate individual metastable states that are excited. The TOF technique can also be used to determine the energy distributions of metastable fragment atoms from molecular dissociation.

The traditional method of studying short-lived excited states is to monitor the photon intensity arising from the optical decay of the excited state as a function of the incident electron energy, giving the so-called "excitation function" of the state. Here individual states, particularly those radiating in the visible region of the spectrum, can often be readily isolated using high-resolution optical spectrometers. However, even for simple atoms, there is great difficulty relating the measured excitation function to the cross section for excitation of the state because of, often unknown, cascade contributions from higher excited states, inherent and instrumental polarization effects, or trapping of the emitted radiation prior to reaching the detector, in addition to the usual difficulties of making cross sections absolute. Again, many studies have concentrated on resonance features in the excitation functions; see, for example, the recent work of Stepanovic *et al.* (2006) in Figure 8.20 of Section 8.3.1.4. For a review of optical excitation functions for electron–atom and electron–ion excitation, see Heddle and Gallagher (1989).

Another way of measuring total cross sections takes advantage of the rapidly developing trap technology in atomic physics. In this technique, a laser is used to trap the atoms of interest, and absolute cross sections can be obtained directly from the trap loss when the laser is switched off. Examples of such measurements are the work of Schappe *et al.* (1995) and MacAskill *et al.* (2002). Lukomski *et al.* (2006) showed that the technique can, in principle, also be used to obtain total cross sections from laser-excited initial states that would be hard to measure otherwise. Another trap measurement involving metastable He atoms was carried out by Uhlmann *et al.* (2005).

Information on the excited state is also accessible through the polarization of the emitted photons. Many studies of the polarization fraction

$$P = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}} \tag{8.3}$$

where $I_{\parallel}$ and $I_{\perp}$ are the radiation intensities observed following transmission through a polarizer with optic axis parallel and perpendicular to the incident electron beam, have been carried out.

8.2.3.3 **Angular and Polarization Correlations**

The information from polarization fraction data has limited value because it relates to the incoherent mixture of the magnetic sublevels of the excited state due to integration over all possible kinematical processes. On the other hand, if identically prepared states are produced, a coherent mixture of the magnetic sublevels of the excited state is observed. For example, an excited $^1\mathrm{P^o}$ state can be described as

$$|\psi\,^1\mathrm{P}\rangle = a_{+1}|11\rangle + a_0|10\rangle + a_{-1}|1-1\rangle \tag{8.4}$$

where the a_M are the excitation amplitudes for the sublevels $|LM\rangle$ with magnetic quantum number M. Theoretical analyses of coherent excitation can be found in Blum (1996), Andersen, Gallagher and Hertel (1988), and Andersen and Bartschat (2001).

Experimentally, the conditions for coherent excitation are produced by observation of the scattered electron responsible for excitation of the particular states at a well-defined scattering angle, in coincidence with the emitted photon. This ensures that only identically prepared excited states are observed. Information on the excited state is then determined by measuring the polarization state of the emitted radiation. This can be done directly, at least for simple systems such as $^1\mathrm{S} \to {}^1\mathrm{P^o}$ excitation in helium, by measuring the linear and circular polarization of the radiation emitted perpendicular to the scattering plane. Alternatively, the information can be deduced from the angular correlation between the scattered electron and the emitted photon.

Eminyan *et al.* (1974) carried out the first electron–photon angular correlation experiment for the $2^1\mathrm{P}$ state of helium. In this experiment, those electrons that lost 21.2 eV (the excitation energy of the $2^1\mathrm{P}$ state of helium) were detected by the electron analyzer at a fixed scattering angle. Angular correlations were then determined by observing the coincidence signal as the photon detector was rotated about the collision center.

The first polarization correlation measurement was reported for the 3^1P state of helium by Standage and Kleinpoppen (1976). Here, the linear polarizations P_1 and P_2 and the circular polarization P_3 are measured in coincidence with the scattered electrons for radiation emitted perpendicular to the scattering plane (z-direction). In the "natural coordinate system" with the z-axis defined perpendicular to the scattering plane and the x-axis chosen along the incident beam direction, these "Stokes parameters" are defined by

$$I_z P_1 = I_z(0^\circ) - I_z(90^\circ) \tag{8.5}$$

$$I_z P_2 = I_z(45^\circ) - I_z(135^\circ) \tag{8.6}$$

$$I_z P_3 = I_z(\mathrm{RHC}) - I_z(\mathrm{LHC}) \tag{8.7}$$

where $I_z(\theta)$ is the signal transmitted by a linear polarizer with optic axis at an angle θ to the electron beam, I is the total signal in that direction, and RHC and LHC refer to right-hand and left-hand circularly polarized radiation. A more complete description of excitation of states with higher angular momentum requires a further linear polarization measurement in the scattering plane (usually the y-direction),

$$I_y P_1 = I_y(0^\circ) - I_y(90^\circ) \tag{8.8}$$

This is also true for heavier targets in which relativistic effects such as the spin–orbit interaction can no longer be ignored.

Note that many other definitions of the Stokes parameters have been used in the literature, as well as a different coordinate system. The "collision frame," often used for convenience in numerical calculations, is defined with the z-axis along the incident beam and the y-axis perpendicular to the scattering plane. For a translation between the various conventions, see Andersen, Gallagher and Hertel (1988).

We show in Figure 8.3 a schematic diagram of the polarization analysis system used in Newcastle. By rotating the in-plane photon detector, the apparatus can also measure angular correlations.

We show the most recent experimental data and theory for the reduced (after fine-structure depolarization has been taken into account) Stokes parameters for 1s − 2p excitation in e − H scattering at 54.4 eV in Figure 8.4. Note that the angular correlation data of Yalim, Cvejanovic and Crowe (1997) and the polarization correlation data of Williams and Mikosza (2006) lie consistently within one standard deviation of theory.

The measured angular correlations have been analyzed in different ways. The analysis of Andersen, Gallagher and Hertel (1988) in terms of the shape and dynamics of the excited state charge cloud gives a particularly straightforward physical picture. They define the following parameters, sometimes referred to as *electron impact coherence parameters*: the alignment angle γ, the length l and width w in the scattering plane, and the height h of the charge cloud along the axis perpendicular to the scattering plane. Furthermore, $L_\perp$ is the expectation value of the angular momentum transfer perpendicular to the scattering plane while the degree of linear polarization, P_l, of the emitted light is defined in terms of l and w. As shown in Figure 8.5, these

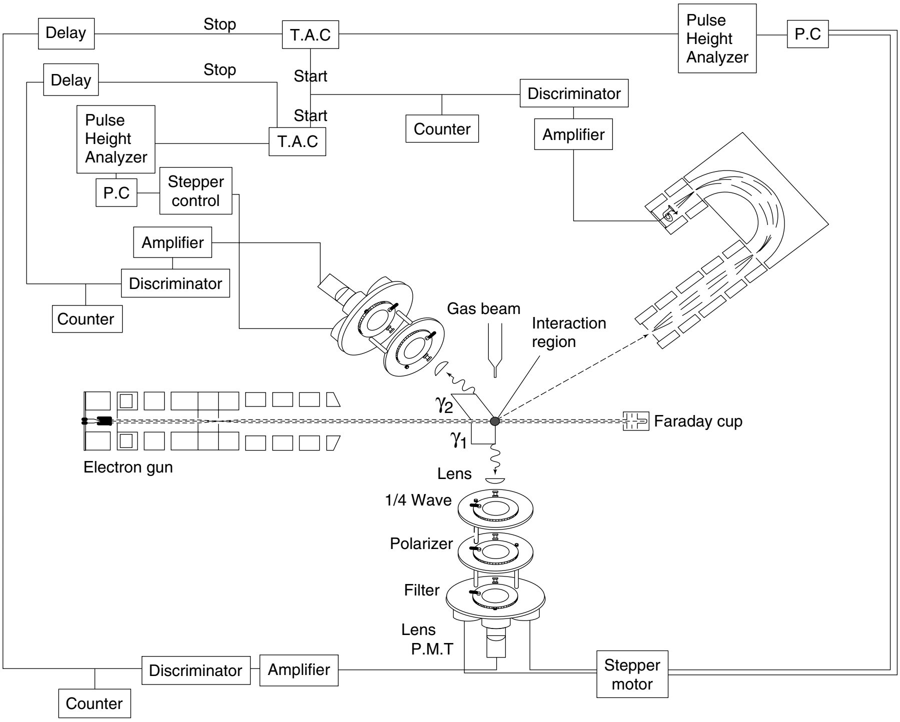
Delay
Stop
T.A.C
Pulse Height Analyzer
P.C
Delay
Stop
Start
Discriminator
Start
Counter
Amplifier
Pulse Height Analyzer
T.A.C
P.C
Stepper control
Amplifier
Discriminator
Counter
Gas beam
Interaction region
γ2
γ1
Faraday cup
Electron gun
Lens
1/4 Wave
Polarizer
Filter
Lens
P.M.T
Discriminator
Amplifier
Stepper motor
Counter

Fig. 8.3 The Newcastle scattered electron-polarized photon correlation spectrometer.

parameters are readily related to the measured Stokes parameters and angular correlations. They have the added advantage that they can be used to describe a wide range of excited states in different atoms.

The variations of $L_{\perp}$, P_l, and γ with scattering angle are shown in Figure 8.6 for the 2^1P state in helium at an incident electron energy of 50 eV. Once again, there is excellent agreement between experiment and the most sophisticated theories currently available for this problem.

An alternative method giving the same complete information on excitation is superelastic scattering from laser-excited states. The technique involves exciting the state of interest with a laser. Electrons are then scattered from this excited state and those scattered electrons that have gained energy corresponding to the energy of the excited state (superelastically scattered) are observed as a function of the polarization of the laser photons. Compared with the correlation methods, the technique is relatively fast because it does not involve coincidence methods. On the negative side, it can be applied only to the states that are accessible by a single laser photon. The latest developments and results from excitation of $4P_1$ state of calcium can be found in Hussey *et al.* (2008) and for the $(6p)^2P_{3/2}$ state of cesium in Slaughter *et al.* (2007). Interestingly, the latter study shows that the measured Stokes parameters are insensitive to inclusion of core and relativistic effects for a heavy atom.

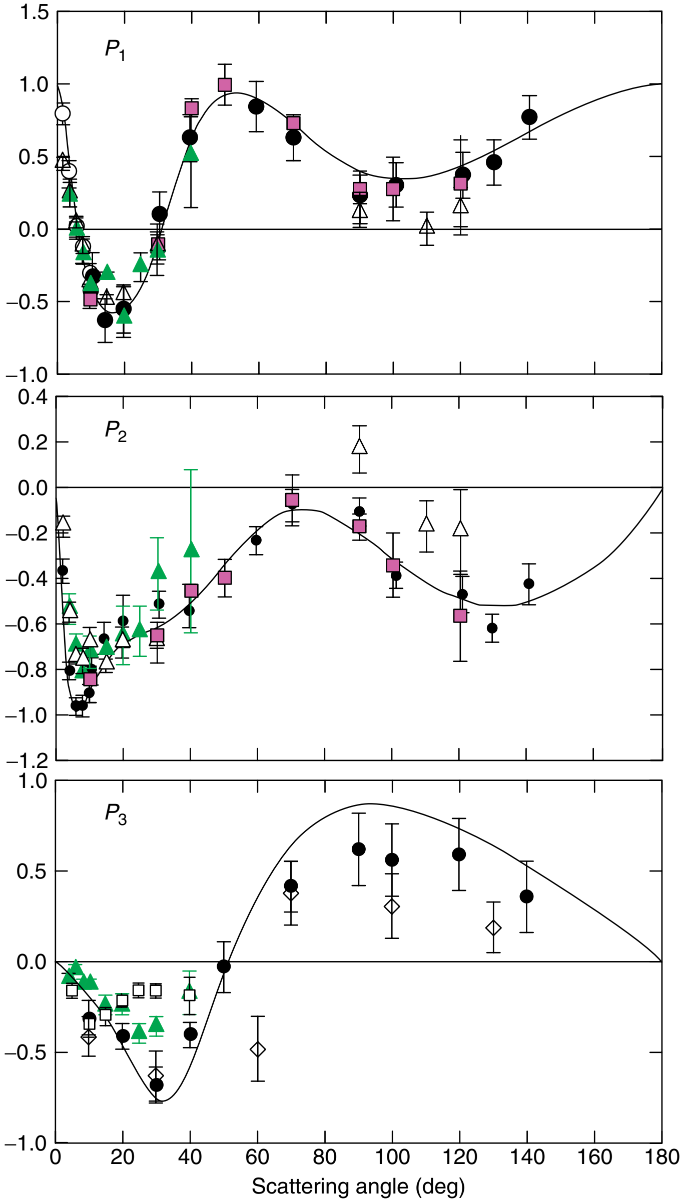

Fig. 8.4 The reduced Stokes parameters, P_1, P_2 and P_3 for the H(2p) state at an incident electron energy of 54 eV. Closed squares, Yalim, Cvejanovic and Crowe (1997); closed and open circles, Williams and Mikosza (2006); closed triangles, Gradziel and O'Neill (2004); open triangles, O'Neill *et al.* (1998); open squares, Nic Chormaic, Chwirot and Slevin (1993); open diamonds, Williams (1986); full line, the PECS theory of Bartlett *et al.* (2005). Adapted from Williams and Mikosza (2006).

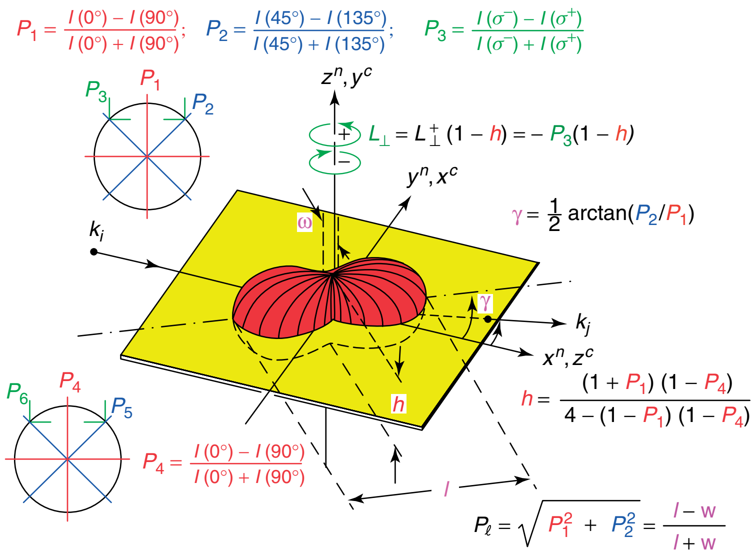
$P_1 = \frac{I(0°) - I(90°)}{I(0°) + I(90°)}$;
$P_2 = \frac{I(45°) - I(135°)}{I(45°) + I(135°)}$;
$P_3 = \frac{I(\sigma^-) - I(\sigma^+)}{I(\sigma^-) + I(\sigma^+)}$
P_3
P_1
P_2
z^n, y^c
$L_\perp = L_\perp^+(1 - h) = - P_3(1 - h)$
+
–
y^n, x^c
$\gamma = \frac{1}{2}\arctan(P_2/P_1)$
k_i
ω
γ
k_j
x^n, z^c
$h = \frac{(1 + P_1)(1 - P_4)}{4 - (1 - P_1)(1 - P_4)}$
h
P_6
P_4
P_5
$P_4 = \frac{I(0°) - I(90°)}{I(0°) + I(90°)}$
l
$P_\ell = \sqrt{P_1^2 + P_2^2} = \frac{l - w}{l + w}$

Fig. 8.5 The charge cloud parameters of Andersen *et al.* (see text).

More detailed reviews of angular and polarization correlation and superelastic scattering studies were given by Andersen, Gallagher and Hertel (1988) Andersen *et al.* (1997), and Andersen and Bartschat (2001).

8.2.4
Ionization

The simplest electron-impact ionization process arises from the ejection of a single atomic electron from an outer shell:

$$e^-(E_i) + A_i \rightarrow A_j^+ + e^-(E_A) + e^-(E_B) \tag{8.9}$$

where E_i, E_A, and E_B are the incident, scattered, and ejected electron energies, respectively. They are related through energy conservation by

$$E_i = E_A + E_B + \varepsilon \tag{8.10}$$

where ε is the binding energy of the ejected atomic electron.

8.2.4.1 Total Cross Sections

These measurements are conventionally carried out by observation of the positive ions produced in the ionization process:

$$i^+ = i_0 N \sigma_i x \tag{8.11}$$

where i_0 is the incident electron beam current; N, the number density of the target gas; x, the effective target thickness; i^+, the measured positive-ion current; and σ_i, the total ionization cross section. The work of Rapp and Englander-Golden (1965) for a range of stable target atoms and molecules is widely used as a standard. For unstable targets, crossed modulated-beam techniques have been developed involving both thermal energy and fast atomic beams. Components of the total ionization cross section, for

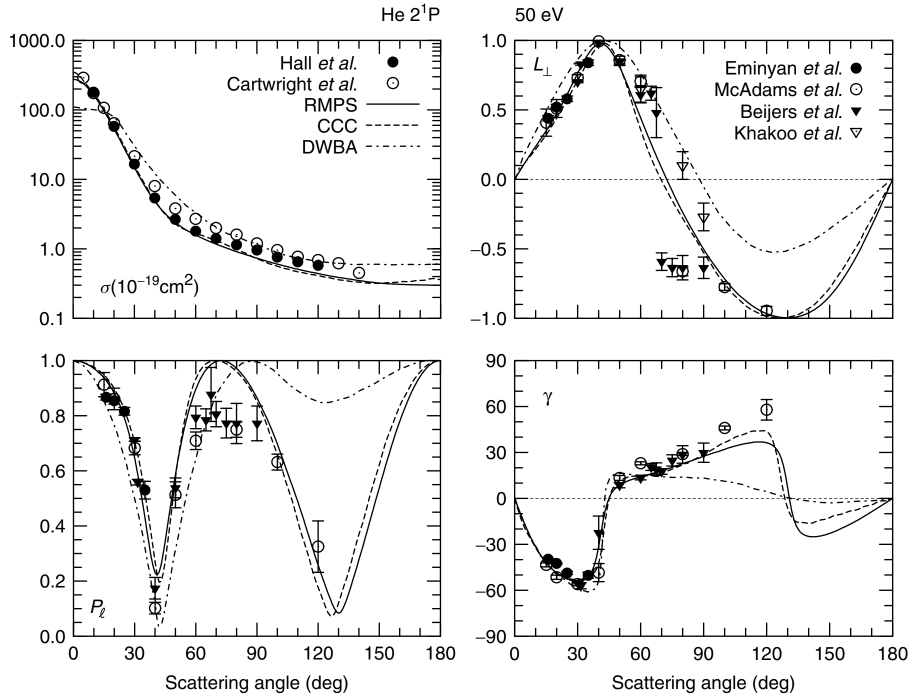

Fig. 8.6 Differential cross section (DCS) and electron impact coherence parameters $L_\perp$, γ and P_ℓ for electron impact excitation of the 2^1P^o state in helium from the ground state 1^1S at an incident electron energy of 50 eV. The theoretical curves are from: RMPS, Bartschat *et al.* (1996a); CCC, Fursa and Bray (1995); DWBA, Beijers *et al.* (1987). The sources for the experimental data are: Hall *et al.* (1973); Cartwright *et al.* (1992); Eminyan *et al.* (1974); Beijers *et al.* (1987); McAdams *et al.* (1980); Khakoo *et al.* (1986).

example, multiple ionization and dissociative ionization, can be identified by incorporating mass spectroscopic analysis in the apparatus. Accurate total and partial ionization cross sections for a wide range of atoms and molecules were recently reported by a group from Rice University; see Lindsay and Mangan (2003) and references therein.

8.2.4.2 **Double Differential Cross Sections**

The detection of a single outgoing electron as a function of its energy and angle is less valuable than in the case of excitation, because the available excess energy of the incoming electron is shared between the scattered and ejected electrons, thus leading to a continuum of energies for both electrons. Note that single differential cross sections as a function of one variable are difficult to measure.

Two distinct double differential cross sections (DDCS) $d^2\sigma/d\Omega dE$ for ionization can be measured. The angular distributions of outgoing electrons of fixed energy may be observed for a particular incident electron energy. For incident electron energies well above the ionization threshold, the general behavior of these cross sections is readily characterized. When the observed electron energy is high, the measured cross section is typical of a "scattered" electron, strongly peaked in the forward direction and falling by orders of magnitude as the scattering angle is increased. As the observed electron energy is decreased, the forward peak becomes less pronounced and the cross section is dominated by a peak typically around 60° scattering angle, due to the increasing dominance of "ejected" electrons.

Alternatively, the energy distribution of the outgoing electron can be monitored at particular scattering angles. Extensive DDCS data were recently reported at low electron energies for atomic hydrogen and helium by Childers *et al.* (2004) and Schow *et al.* (2005).

8.2.4.3 **Higher Order, including, Fully Differential Cross Sections**

In recent years, partly due to the development of more efficient coincidence electron spectrometers, a wide range of multidifferential cross sections are being measured for most direct and indirect ionization processes. These include single ionization to the ground and specific excited ion states, double excitation–autoionization, double ionization, and inner-shell ionization. Each of these is discussed in detail in the following sections.

Experimentally, these measurements are performed using electron spectrometers with some combination of electron analyzers to observe the scattered and/or ejected electrons and, if appropriate, a photon detector to observe the decay of the excited ion state. Coincidence techniques allow for identification of the ionization process of interest.

Two very different techniques have been used to dramatically increase the efficiency of spectrometers that are used to detect electrons. Catoire *et al.* (2007) describe a spectrometer using three toroidal electron analyzers in conjunction with time- and position-sensitive detectors to observe the three outgoing electrons. This setup enables the *simultaneous* observation of both the ejected electrons from double ionization over most of the scattering plane, defined by the incident electron momentum $\boldsymbol{k}_i$ and the scattered electron momentum $\boldsymbol{k}_A$, and for a reasonable range of scattering angles. Figure 8.7 shows a schematic illustration of their triple coincidence spectrometer.

A typical reaction microscope used by Dürr *et al.* (2008) to study single ionization processes is shown in Fig. 8.8. It operates on entirely different principles from conventional electron spectrometers. Briefly, a pulsed beam of electrons crosses a supersonic atom beam. The ejected electrons and the recoiling ions are extracted in opposite directions by a weak uniform electric field that is parallel to the direction of the incident electron beam. A uniform magnetic field is also applied in this direction to confine electrons that are emitted perpendicular to the electric field. After passing through field-free drift regions, the slow ejected electrons are then detected in two time- and position-sensitive multihit detectors, allowing for the vector momenta of the two particles to be calculated. Unlike most conventional coincidence electron spectrometers, including the toroidal based systems, which only enable measurements in a single plane at any one time, this technique allows for data to be collected over a large part of the full solid angle simultaneously.

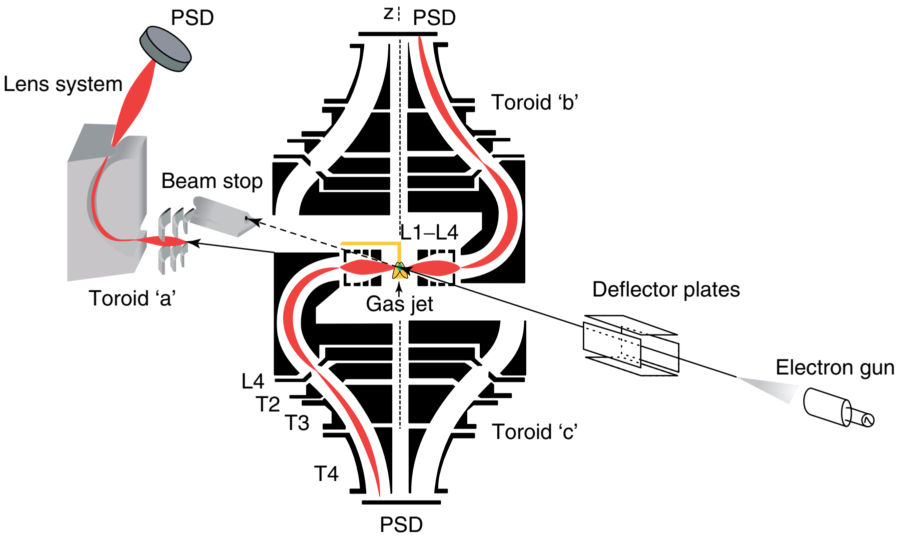
PSD
Lens system
Beam stop
Toroid ‘a’
z
PSD
Toroid ‘b’
L1–L4
Gas jet
Deflector plates
Electron gun
L4
T2
T3
T4
Toroid ‘c’
PSD

Fig. 8.7 The (e,2e)/(e,3e) spectrometer of Catoire *et al.* (2007).

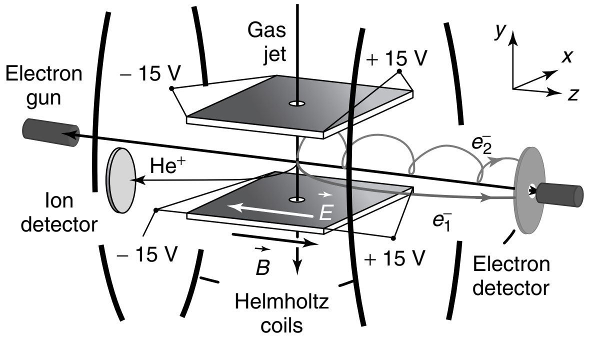
Electron gun
– 15 V
Gas jet
+ 15 V
y
x
z
e_2^-
He$^+$
Ion detector
$\vec{E}$
e_1^-
– 15 V
$\vec{B}$
+ 15 V
Electron detector
Helmholtz coils

Fig. 8.8 The reaction microscope of Dürr *et al.* (2008).

8.2.4.4 (e,2e) Studies of Direct Single Ionization

The most detailed information on ionizing collisions involving ejection of a *single* electron and leaving the ion in its ground state comes from the measurement of triple differential cross sections (TDCS), $d^3\sigma_i/d\Omega_A d\Omega_B dE$. They are also referred to as (e,2e) cross sections.

The momentum transfer vector

$$\boldsymbol{K} = \boldsymbol{k}_i - \boldsymbol{k}_A \tag{8.12}$$

is an important parameter in discussing this work. Two major and distinct groups of study have been developed depending on the magnitude of the momentum transfer vector $\boldsymbol{K}$.

The magnitude of $\boldsymbol{K}$ is maximized in experiments in which the two outgoing electrons from a high-energy collision are observed with equal energies at equal angles $\theta = \theta_A = \theta_B$ relative to the incident electron beam. In such “high-energy symmetric kinematics,” the interaction approximates a binary collision between two free electrons. Conservation of momentum requires

$$\boldsymbol{k}_i + \boldsymbol{p} = \boldsymbol{k}_A + \boldsymbol{k}_B \tag{8.13}$$

where $\boldsymbol{p}$ is the momentum of the atomic electron prior to the collision. The magnitude of the struck electron momentum is then given by

$$p = [(2k_B \cos\theta - k_i)^2 + 4k_B \sin^2\theta \times \sin^2 \varphi_A/2]^{1/2} \tag{8.14}$$

Here, in principle, variation of any of the parameters k_B, θ, or φ_A allows for the electron momentum p to be scanned. It can be shown that the (e,2e) cross section is proportional to the spherically averaged electron momentum distribution of the atomic electron. Hence a measurement of the TDCS as a function of $\pm p$ reveals the atomic electron momentum distribution, which is related to the spatial wave function through a Fourier transform. Measurements of this type were extensively reviewed, for example, by Weigold and McCarthy (1999).

Measurements of the TDCS for a range of small momentum transfers $\boldsymbol{K}$ were pioneered by Ehrhardt *et al.* (1969). Using low-energy incident electrons ($< 600\,\text{eV}$), they measured angular correlations between the outgoing scattered and ejected electrons under conditions where a slow ejected electron is observed in coincidence with a fast electron scattered through small angles. Most ionizing collisions fall into this category of "low-energy asymmetric kinematics."

The early experiments by Ehrhardt's group, mostly in helium, quickly established the general trends of relative low-energy asymmetric (e,2e) cross sections measured in the scattering plane. Typical results (see Figure 8.9 for an example) are dominated by a peak in a direction close to the momentum transfer direction, referred to as the *binary encounter* peak, and a peak in the opposite direction, referred to as the *recoil peak*. Earlier advances in (e,2e) and related studies were reviewed by Lahmam-Bennani (1991).

Figure 8.10 shows the current state of the art for electron-impact ionization of atomic hydrogen close to the threshold energy and for equal sharing of the excess energy by the two outgoing electrons. The most sophisticated currently available theories for this two-electron process, convergent close-coupling (CCC), time-dependent close-coupling (TDCC), and exterior complex scaling (ECS), all of which are discussed in the theory section, yield predictions in excellent agreement with each other. These predictions also agree very well with most of the available experimental data. In fact, for the discrepancies which remain, it is not obvious that the problems lie with theory for this benchmark process.

Stevenson and Lohmann (2008) reported the first (e,2e) measurements where one of the outgoing electrons is observed over the full angular range, $0^\circ - 360^\circ$, using a MAC device. Their new data for ionization of the 3p electron in argon reveal structures in the previously unobserved angular ranges that are generally not well reproduced by the state-of-the-art calculations of Prideaux, Madison and Bartschat (2005) and Bartschat and Vorov (2005).

Recent reaction microscope measurements in three dimensions also revealed previously unobserved features in the helium TDCS. Typical data of Dürr *et al.* (2006) are shown in Figure 8.11. In the

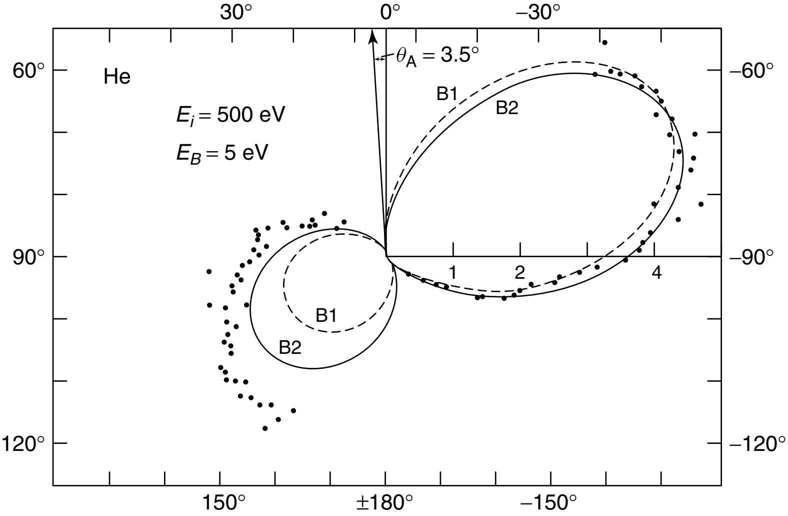

Fig. 8.9 The triple-differential cross section (in atomic units) for helium as a function of the ejected electron angle θ_B for an incident electron energy of 500 eV, an ejected electron energy of 5 eV, and a fixed detection angle of 3.5° for the fast electron. The dots are experimental data, while the curves B1 and B2 represent first-order and second-order plane-wave Born approximations. (Adapted from Ehrhardt *et al.* (1986).)

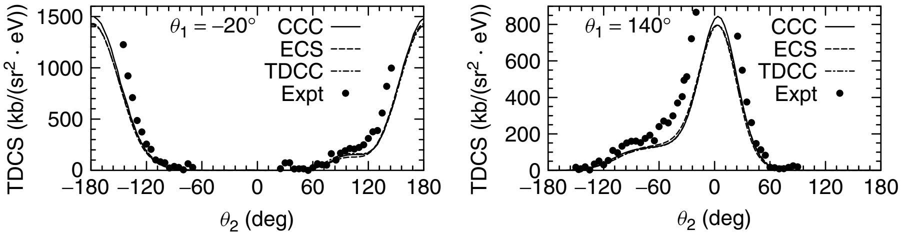

Fig. 8.10 TDCS for electron-impact ionization of H at an incident electron energy of 17.6 eV. Both outgoing electrons have an energy of 2 eV. The detection of one electron is fixed at −20° or 140°, respectively. The absolute experimental data of Röder, Baertschy and Bray (2003) are compared with predictions from CCC, time-dependent close-coupling, and exterior complex scaling models. (Adapted from Colgan and Pindzola (2006).)

scattering plane, the data are consistent with previous experiments and, after taking account of the postcollision interactions between the charged particles, with the dominance of first-order processes. However, in the plane perpendicular to the momentum transfer vector, there is a sizable cross section with maxima at angles of 70° and 290°. These are not reproduced by first-order theories Dürr *et al.* (2008), even at an incident energy as high as 1 keV.

There is an increasing interest in the use of the (e,2e) technique to study the dynamics of molecular ionization, including molecules of biological interest,

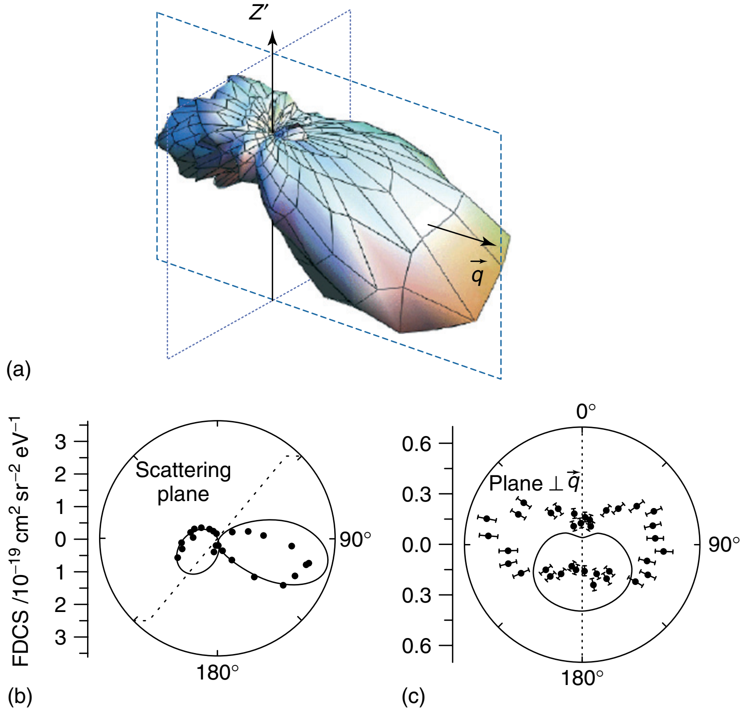

Fig. 8.11 The three-dimensional (e,2e) cross sections of Dürr *et al.* (2006) for helium at an incident electron energy of 102 eV, a scattering angle of 20° and slow outgoing electron energy of 10 eV. (a) 3-D cross section. (b) Cut in the scattering plane [dashed line in (a)]. (c) Cut perpendicular to the scattering plane [dotted line in (a)]. The dotted lines in (b) and (c) show the incident electron direction. The solid curves are the three Coulomb wave function (3C) theory. (From Dürr *et al.* (2006).)

Milne-Brownlie *et al.* (2004). This is despite the experimental complexities associated with the additional degrees of freedom arising from the nuclear motion. Even overlapping electronic states may remain unresolved in these experiments, and so far experiments have only been carried out for random orientations of the internuclear axis.

8.2.4.5 (e,3e) Studies of Double Ionization

Double electron ejection has been studied using the (e,3e) method, which involves detection of all three outgoing electrons in coincidence. The process can be written as

$$e^-(E_i) + A_i \longrightarrow A_j^{++} + e^-(E_A) + e^-(E_B) + e^-(E_C) \qquad (8.15)$$

These experiments yield fivefold differential cross sections, $d^5\sigma_i/dE_A dE_B d\Omega_A d\Omega_B d\Omega_C$.

Although the (e,3e) technique was pioneered by Lahmam-Bennani, Dupré and Duguet (1989) two decades ago, it can still be considered to be in its infancy due to the small cross sections

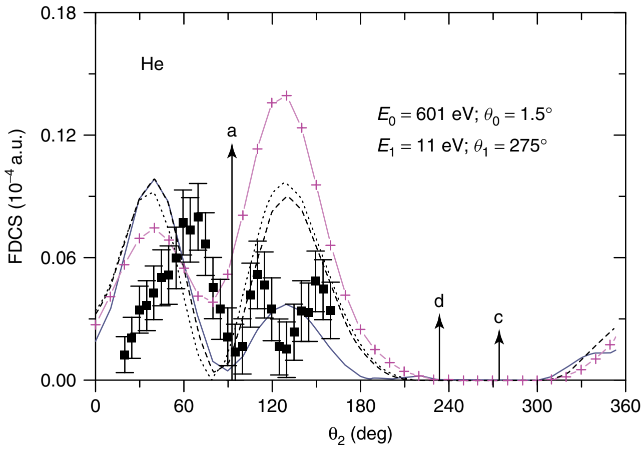

Fig. 8.12 The fivefold coplanar differential cross section in atomic units for the double ionization of helium at an incident electron energy of 601 eV. The scattered electron with an energy of 500 eV and scattered through 1.5° is observed along with an 11 eV ejected electron at 275° and the other ejected electron as a function of angle. The experimental data are from Lahmam-Bennani *et al.* (2003). Theory: first Born, dotted line; second Born, solid line with crosses; approximate 6C models, solid and dashed lines. According to general selection rules (Berakdar, Lahmam-Bennani and Dal Cappello (2003)), there should be minima at the angles denoted by "a," "c," and "d." (From Elazzouzi *et al.* (2005).)

for double ionization, the wide variation of kinematical situations arising from the presence of the second ejected electron, and the experimental complexities involved. The experimental efficiencies have been greatly enhanced by the approaches of both Catoire *et al.* (2007) and Dürr *et al.* (2007). Figure 8.12 shows typical (e,3e) data for the kinematical conditions shown. While the minima predicted by fundamental selection rules (Berakdar, Lahmam-Bennani and Dal Cappello (2003)) are reproduced by experiment, the overall agreement with the calculations shown is not good.

8.2.4.6 Excitation–Ionization

The study of ionization, where the ion is formed in an excited state, has been the subject of considerable attention recently. At least some of this work was stimulated by the desire to carry out the complete ionization experiment, analogous to the complete excitation experiments discussed in Section 8.2.3.3. Such a complete experiment would require a triple coincidence experiment, (e, 2eγ), in which the polarization of the decay photon is determined for specific kinematics by its observation in coincidence with the two outgoing electrons. Such an experiment has not been performed to date, although some progress toward its realization has been made.

Schematically, the process is written as

$$e^-(E_i) + A_i \longrightarrow A_j^{+*} + e^-(E_A) + e^-(E_B) \qquad (8.16)$$

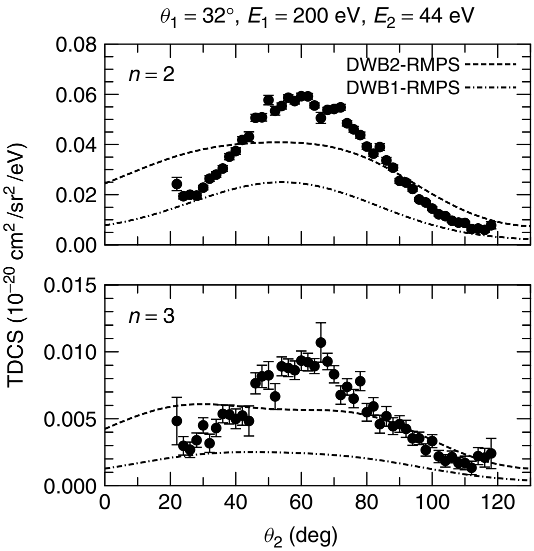

Fig. 8.13 Experimental TDCS for excitation-ionization to the $n = 2$ and $n = 3$ states of He$^+$ compared with hybrid R-matrix calculations. In each case, the outgoing scattered electron has an energy of 200 eV and the scattering angle is 32°. The ejected electron energy is 44 eV. The experimental data are from Bellm, Lower and Bartschat (2006), after combining their measured ratios with theoretical predictions for ionization without excitation. (From Bartschat *et al.* (2007).)

The excited ion state, A_j^{+*}, then decays by optical emission to a lower lying ion state,

$$A_j^{+*} \longrightarrow A_k^+ + h\nu \tag{8.17}$$

Three related types of studies have been reported. Stefani, Avaldi and Camilloni (1990) first reported an (e,2e) study of the He$^+$(n = 2) states. This produces a TDCS with the disadvantages that the 2s and 2p ion states are not resolved and that the detailed dynamical information gained from the polarization of the photon is absent. Hayes and Williams (1996) first applied the single electron–photon, (e,eγ), technique to the same system. This type of study enables the 2p state to be isolated in the study and angular correlation parameters to be determined, integrated over the slow ejected electron emission angles (Schwienhorst *et al.*, 1996). A single-electron-polarized-photon study of the 4p^{2}P$_{3/2}$ excited state of Ca$^+$ was reported by Stevenson and Crowe (2004). Sakhelashvili *et al.* (2005) carried out an (e,2eγ) experiment in helium without resolving the different He$^+$(np)^{2}P states and without regard to the direction or polarization of the emitted photons. If any isotropy in the emitted radiation is neglected as well, the data correspond to the TDCS for the (np)^{2}P states, with approximately 90% of the signal believed to be coming from the He$^+$(2p) state.

Figure 8.13 shows TDCS for the excitation-ionization to the n = 2 and n = 3 states of He$^+$ compared with hybrid R-matrix calculations. The absolute "experimental data" presented by Bartschat *et al.* (2007) were generated by combining the measured ratios of these TDCS to the n = 1 ion state by Bellm, Lower and Bartschat (2006) with CCC results for n = 1. The data show broad peaks in the forward direction. Although the kinematics are not identical, the TDCS of Dogan and Crowe (2000) for these states suggest that more interesting features in the TDCS exist outside the angular range of the data shown here. The need for higher order effects to be taken into account in theoretically describing these two-active atomic electron processes is demonstrated by comparing theoretical predictions obtained through first-order (DWB1-RMPS) or second-order (DWB2-RMPS) treatments of the projectile-target interaction, combined with a convergent *R*-matrix with pseudostates (RMPS) model for the inital bound state and the ejected-electron–residual-ion interaction.

8.2.4.7 Excitation–Autoionization

A major contribution to the ionization cross section for complex atoms and ions often arises from the process known as excitation–autoionization. In this process the target is first excited to an autoionizing or resonant state lying above the ionization threshold by the incident electron,

$$e^- + A_i \rightarrow e^- + A_j^* \qquad (8.18)$$

and the resultant resonant state A_j^* subsequently autoionizes,

$$A_j^* \rightarrow A_k^+ + e^- \qquad (8.19)$$

Hence, the total effect of these two processes is to ionize the original atom or ion. A recent example involving inner-shell excitation of Na was reported by Borovik, Zatsarinny and Bartschat (2008).

In addition, the incident electron may sometimes be captured into a resonant state that decays by the emission of two (or occasionally more) electrons:

$$e^- + A_i \rightarrow A_l^{-*} \rightarrow A_k^+ + 2e^- \qquad (8.20)$$

This process, known as resonant excitation double autoionization (REDA), interferes with the direct process given by Eqs. (8.16) and (8.17) and often dominates the cross section close to threshold. Excitation–autoionization was first observed in the ionization of Ba$^+$ by Peart and Dolder (1968) and has since been shown to be a dominant ionizing mechanism for many ions of laboratory and astrophysical interest.

Excitation–autoionization can be studied experimentally by observation of either the ejected or the scattered electron spectra. Observation of the ejected electron has the advantage that the observed spectra are independent of the inherent inhomogeneity in the incident electron energy E_i.

Autoionization spectra are characterized by asymmetric line profiles due to interference between the direct ionization and the excitation–autoionization process. The majority of single-electron spectra have been analyzed to yield the energies and widths of the resonance states.

Excitation–autoionization processes involving the $2l2l'$ states of helium have also been studied using the (e,2e) technique (see McDonald and Crowe (1993) and references therein). In these experiments, the angular variation of the Shore resonance parameters (Shore (1967)) of each of the autoionizing states was determined. In the most recent study, deHarak, Bartschat and Martin (2008) measured out-of-plane TDCS for each of these singlet states. Dogan and Crowe (2002) used the electron–photon coincidence method to observe the influence of the $3lnl'$ autoionizing states of helium on simultaneous excitation-ionization of the He$^+$2p state.

8.2.4.8 Inner-Shell Ionization

We are interested in processes of the type

$$e^- + A_i \longrightarrow A^+(J, M) + 2e^- \quad (8.21)$$

where the ion $A^+(J, M)$ has an inner-shell vacancy and can decay by either X-ray

$$A^+(J, M) \longrightarrow A^+(J'', M'') + hv \quad (8.22)$$

or Auger-electron emission

$$A^+(J, M) \longrightarrow A^{++}(J', M') + e_{\text{Auger}} \quad (8.23)$$

Except for very heavy atoms, the dominant decay mechanism is by Auger-electron emission, and hence we will concentrate on that mode of decay.

Most studies of inner-shell ionization–Auger-electron decay have involved observation of the energy-loss spectrum of a single outgoing electron. These measurements yield a spectrum of Auger-electron lines superimposed on the continuum of scattered and ejected electrons from all ionization processes. A second type of experimental study has involved measurement of the angular distribution of the Auger electrons. When an inner-shell vacancy is created with $j > \frac{1}{2}$ by a beam of particles, the initial ion A^+ will be aligned with respect to the incident beam direction, the alignment manifesting itself in an anisotropic angular distribution of the decay product. More detailed information on the alignment of the initial ion can be obtained from angular correlation measurements between the Auger electrons and a fast scattered electron, where only specific momentum transfer processes are studied. Such studies were first reported by Sewell and Crowe (1982). Recently, both the (e,2e) and (e,3e) techniques were used by Naja *et al.* (2007) to study inner-shell ionization, taking advantage of the increased efficiency of the electron detectors discussed above.

8.2.5
Electron–Molecule/Biomolecule Interactions

The experimental techniques used to study electron scattering from molecules are the same as for atoms and do not merit special attention. For example, the spectrometers of Allan (2004) and Catoire *et al.* (2007) discussed earlier have been used to study different aspects of electron–molecule scattering. In this case, the nuclear motion opens up a wealth of additional possibilities, in particular in low-energy collisions that often reveal resonance effects. Such resonance and threshold phenomena in low-energy electron collisions with molecules and clusters were reviewed by Hotop *et al.* (2003).

Experiments to determine electronic excitation cross sections, with typical energy resolutions in the range of 30 – 50 meV, can be very problematic due to the nuclear degrees of freedom. In this case, spectral deconvolution techniques are often necessary to obtain well-defined results, especially for polyatomic molecules. Brunger *et al.* (2008) give an excellent recent example of the use of deconvolution methods to obtain differential excitation cross sections for each of the six lowest lying electronic states of water.

One recent development that justifies comment is the study of biomolecules. This interest is largely generated by the important role played by low-energy electrons in radiobiology. Such low-energy ($\leq$30 eV) electrons are produced by the

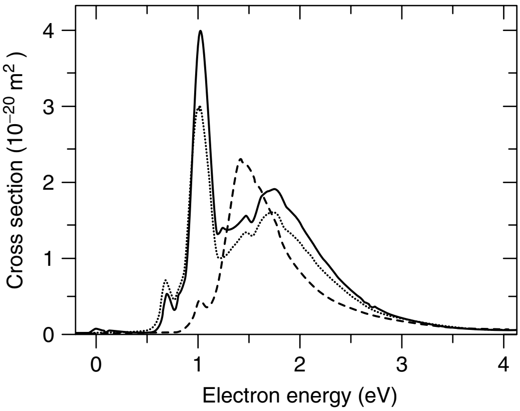

Fig. 8.14 The cross sections for formation of the most abundant negative ions formed by electron attachment to each of uracil (U), thymine (U), and cytosine (C). (U-H)$^-$ from U (dotted line); (T-H)$^-$ from T (full line and × 0.33); (C-H)$^-$ from C (dashed line). (From Denifl *et al.* (2004).)

interaction of fast ionizing radiation with cellular constituents. The vast majority of these electrons are responsible for the production of reactive negative ions and radical fragments through the formation and decay of resonances.

Advances in understanding the role of these low-energy electrons have relied to a considerable extent on the interplay between scattering from gaseous- and solid-state targets. For example, while Boudaiffa *et al.* (2000) observed single- and double-strand breaks in supercoiled DNA deposited on a cold surface, they argued that the resonances observed were similar to those seen in gas-phase electron attachment experiments. More generally the gas-phase experiments are important in distinguishing between environmental and intrinsic molecular effects.

The first gas-phase electron scattering studies on uracil and the DNA bases were performed by Aflatooni, Gallup and Burrow (1998) using their electron transmission method developed for electron–atom scattering. Another important series of systematic studies of the stable negative ions produced from uracil and the DNA bases, detected using a high-resolution quadrupole mass spectrometer, have provided considerable insight into the mechanisms of damage to DNA by low-energy electrons; see, for example, Ptasińska *et al.* (2005). Electron attachment to these biomolecules leads to dissociation into various fragments without any measurable amount of the stable parent ions. Figure 8.14 shows the cross sections for the formation of the most dominant negative ions produced from the RNA base uracil and the DNA bases, thymine and cytosine reported by Denifl *et al.* (2004). The reader is referred to the original publication for details and to the review of Sanche (2005) for a broad discussion of the effects arising from the interaction of low-energy electrons with biomolecules.

8.2.6
Electron–Ion Interactions

Experimental studies of electron-ion interactions have traditionally been sparse compared with their theoretical counterparts. Recently, this has been transformed by the adoption of heavy-ion storage rings for electron–ion studies. The technique can be considered

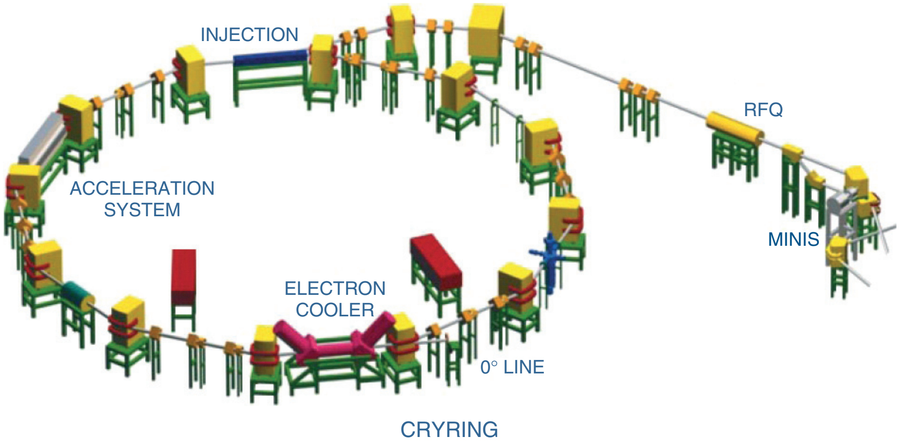

Fig. 8.15 The Stockholm ion storage ring (51.6 m circumference). (From Larsson and Thomas (2001).)

as an extension of the merged beams method, in which electron and ion beams are brought together and interact over a common path length, (Phaneuf *et al.*, 1999). One of a number of very significant advantages of the storage rings is that the ions pass through the interaction region more than 10^5 times per second, whereas the merged beams machines are single-pass devices. One of the storage rings, which has been used in atomic/molecular physics, the CRYRING device, is shown in Figure 8.15 (Larsson, 2001). Briefly, ions are extracted from the ion source MINIS, mass selected, and then injected into the storage ring and further accelerated to the required beam energy, typically $\sim$ MeV. In the straight "electron cooler" section, the stored ion beam is merged with an electron beam with the same velocity and cooled, typically for a few seconds. Electron cooling is based on the fact that the Coulomb interaction between the electrons and ions leads to heat transfer from the warm ion beam to the cold and continuously renewed electron beam. Amongst the outcomes of cooling are an ion beam with an extremely narrow momentum spread and ions in the lowest vibrational state of a molecular ion. The electron beam in the cooler also has excellent properties for electron–ion collision studies. The electron–ion interaction energy is given by the relative velocities of the two beams. In principle, this can be made zero, but in practice the lowest energy is a few milli electron-volts due to the velocity distributions and finite angular resolutions of the beams. Overall energy resolutions are typically $\sim 1 - 2$ meV. The more recent addition of a supersonic expansion ion source has also led to studies involving rotationally cold molecular ions (McCall *et al.*, 2003). A range of detection devices can be used to detect the collision products depending on the process studied.

Storage rings have been used to study a wide range of low-energy $(0 - 100\,\text{eV})$ electron–ion collisions with unprecedented resolution and often of fundamental importance in fusion and astrophysical plasmas. The cross

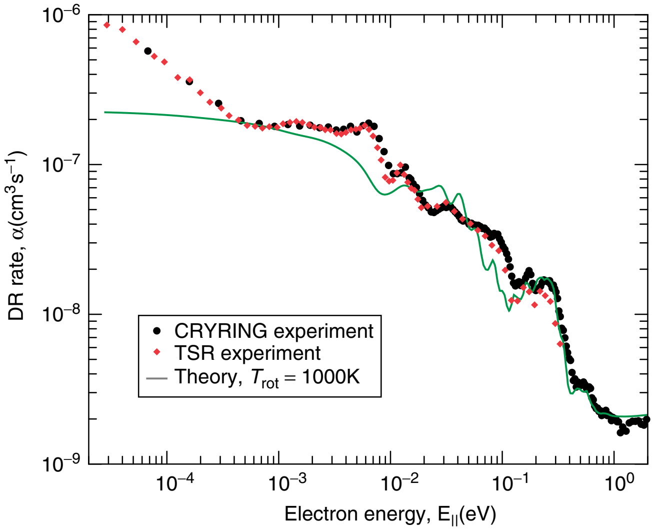

Fig. 8.16 Dissociative recombination rate of H_3^+ as a function of the incident electron energy. The theoretical predictions of dos Santos, Kokoouline and Greene (2007) are compared with experimental data from the Stockholm (CRYRING) and Heidelberg (TSR) groups.

section measurements of recombination processes are particularly important. Figure 8.16 shows an example of dissociative recombination. For many years, theory had been unable to explain the experimentally observed large rates for this process in H_3^+, which is of fundamental importance in diffuse interstellar clouds. Kokoouline and Greene (2003a) finally identified Jahn–Teller coupling as the critical mechanism that must be included in a theoretical treatment. As seen in Figure 8.16, the latest calculations of dos Santos, Kokoouline and Greene (2007) finally obtained good agreement with recent storage ring data.

8.2.7
Spin-Polarized Electron Studies

Previous discussions have involved experiments in which the electron beam was produced by thermionic emission with the spins of the electrons having arbitrary directions. Hence, for example, in discussing excitation of the 2^2P state of atomic hydrogen in Section 8.2.3.3, the analysis involves an average over all spin orientations.

It is possible to produce electron beams with a preferential orientation of the electron spin (Kessler (1985)). In general, the use of spin-polarized electron beams in electron–atom scattering provides a further definition of the initial and/or final state of the system, leading toward the "perfect" or "complete" scattering experiment. In such an experiment, all parameters that can be determined according to the limits set by quantum mechanics, are measured for a given process. This information is contained in the magnitudes and (relative) phases of the complex-valued scattering amplitudes. The concept of complete experiments was introduced by Bederson (1969). More information can be found in Andersen and Bartschat (1996); Andersen and Bartschat (2001).

Spin-polarized electron and atom beams provide a means of studying spin-dependent interactions, the details of which are often masked in experiments involving unpolarized electrons by the stronger Coulomb interaction. Two specific spin-dependent interactions have been the subject of extensive study – the exchange interaction and the spin–orbit interaction. Although both these effects are important for scattering from a wide range of atomic targets, it is useful to consider the extreme situations where only one is important.

8.2.7.1 Exchange Interaction

This is particularly important for spin$-\frac{1}{2}$, low-Z targets, including atomic hydrogen and the alkali atoms. Assuming that the total spin of the collision system (singlet and triplet) is conserved, the cross sections for scattering from such targets depend on the relative orientations of the electron and target spins, that is, $\sigma(\uparrow\downarrow) \neq \sigma(\uparrow\uparrow)$, where the arrows indicate the electron and atom spins. As an example, we give in Table 8.1 the possible scattering modes for the elastic scattering of electrons with spin parallel or antiparallel to an atom with a given spin direction ($A\uparrow$ or $A\downarrow$, together with the corresponding scattering amplitudes and cross sections. These expressions can readily be generalized to inelastic scattering (Kessler, 1985).

Scattering of spin-polarized electrons from spin-polarized atoms with spin analysis of the products yields the cross sections $|f|^2$, $|g|^2$, and $|f-g|^2$. This constitutes an extremely difficult experiment. Some information can already be obtained if either the initial electron or atom beam is spin polarized, and if either the electron or the atom polarization is measured after the collision process. However, a complete experiment requires the measurement of the relative phase of f and g as well.

Most experimental studies do not set out to obtain complete information on the scattering amplitudes. Several experiments have been carried out by scattering polarized electrons from polarized atoms and measuring the asymmetry

$$A = \frac{\sigma(\uparrow\downarrow) - \sigma(\uparrow\uparrow)}{\sigma(\uparrow\downarrow) + \sigma(\uparrow\uparrow)} \tag{8.24}$$

where $\sigma(\uparrow\downarrow)$ is the cross section for antiparallel electron and atom spins and $\sigma(\uparrow\uparrow)$ is the cross section for parallel electron and atom spins. Of particular importance for the further confirmation of the CCC theory for hydrogen-like systems (Bray (1994)) were several experiments performed by the

Tab. 8.1 Amplitudes and cross sections for scattering of spin-polarized electrons and spin-$\frac{1}{2}$ targets.

Process	Amplitude	Cross section
$e\downarrow + A\uparrow \rightarrow e\downarrow + A\uparrow$	f	$\|f\|^2$
$e\downarrow + A\uparrow \rightarrow e\uparrow + A\downarrow$	$-g$	$\|g\|^2$
$e\uparrow + A\uparrow \rightarrow e\uparrow + A\uparrow$	$f-g$	$\|f-g\|^2$

NIST group for elastic and inelastic electron scattering from Na atoms (see McClelland, Kelley and Celotta (1989), Lorentz *et al.* (1993), and references therein).

Exchange effects are also very important in spin-forbidden transitions, such as $^1S^e \rightarrow {}^3P^o$ in He, the heavy noble gases, and even mercury. If such a transition is achieved by an incident polarized electron beam, the change of the beam polarization is a measure of the relative cross section. Furthermore, an unpolarized electron beam can become polarized if an individual fine-structure level of the target is resolved. This so-called "fine-structure effect," originally suggested by Hanne (1976), has been seen in both excitation and ionization Hanne (1983), Jones, Madison and Hanne (1994), and Andersen and Bartschat (2001).

8.2.7.2 Spin–Orbit Interaction

This is particularly important for high-Z targets (e.g., Cs or Hg). In this case, relativistic effects cause a spin–orbit interaction between the incident electron and the target atom characterized by the scalar product $\boldsymbol{l} \cdot \boldsymbol{s}$, where $\boldsymbol{l}$ is the orbital and $\boldsymbol{s}$ is the spin angular momentum.

In the case of elastic scattering from closed-shell targets (like Hg), the scattering process can be described by two complex amplitudes, f, due to the Coulomb interaction, and g, due to the spin–orbit interaction (not to be confused with the direct and exchange amplitudes above). Complete determination of f and g, apart from an overall phase, requires three independent measurements. One of these is the so-called "polarization function" S_P that determines the spin polarization of an initially unpolarized electron beam after the scattering. Owing to parity conservation, the polarization vector must be perpendicular to the scattering plane. Another possibility is to measure the spin-asymmetry function S_A, which determines the left-right asymmetry in the DCS if the incident electron beam is already spin polarized. Owing to time-reversal invariance, $S_P = S_A \equiv S$ for elastic scattering (Bartschat, 1989). The other parameters, called T and U for elastic scattering from closed-shell targets, determine the change of an initial electron polarization due to the scattering process. Such experiments are highly complicated but have been performed successfully in the Münster group; see, for example, Berger and Kessler (1986). For inelastic collisions, "generalized" STU-parameters need to be defined (Bartschat, 1989), which have also been measured in several cases; see Andersen *et al.* (1997) for a summary of available data.

Using a spin-polarized incident electron beam also vastly increases the possibilities for the polarization properties of the emitted radiation. Not surprisingly, "generalized" Stokes parameters (Andersen and Bartschat (2001)) have to be defined that depend on the electron polarization as well as the position and the setting of the light analyzers. In addition to performing very difficult angle-resolved electron–photon coincidence experiments, it has become evident that some interesting information can already be obtained by only observing the polarization of the emitted light. For properly chosen transitions, observing the circular polarization of the emitted light can be used to optically measure the electron polarization Eminyan and Lampel (1980), Gay (1983), Goeke, Kessler and Hanne (1987), while observation of the linear polarization P_2 can give an indication about the importance of explicitly spin-dependent effects on the collision process Bartschat *et al.* (1981), Bartschat and Blum (1982). Recent data for the "integrated Stokes parameter" in spin-polarized electron-impact excitation of Zn were reported by Pravica *et al.* (2007).

8.2.7.3 A Typical Apparatus and Some Illustrative Results

Figure 8.17 shows the apparatus of Sohn and Hanne (1992) used to determine generalized Stokes parameters for electron-impact excitation of mercury. More importantly, from the point of view of this review, it illustrates the essential features of many polarized-electron studies. Today's sophisticated experiments have only become feasible with the development of highly efficient GaAs polarized-electron sources. The polarized electrons are produced from the GaAsP cathode by photoemission using circularly polarized light, the direction of polarization coinciding with the axis of the incident laser beam. The electron beam is then guided through an electron-optical system, which also includes the possibility of rotating the polarization vector via a Wien filter. After scattering from a beam of target atoms, the scattered electrons and the emitted photons (with their polarization analyzed as well) are observed in coincidence. Finally, the polarization of the incident beam is determined by measuring the left-right asymmetry of the accelerated outgoing electrons (without a target beam) scattered from a gold foil, making use of the spin–orbit effect in a so-called "Mott detector."

Similar apparatus were built by Berger and Kessler (1986) to measure the (generalized) STU-parameters for electron collisions with mercury and heavy noble gases, and by the Bielefeld group (Baum *et al.* (2002); Baum *et al.* (2004))

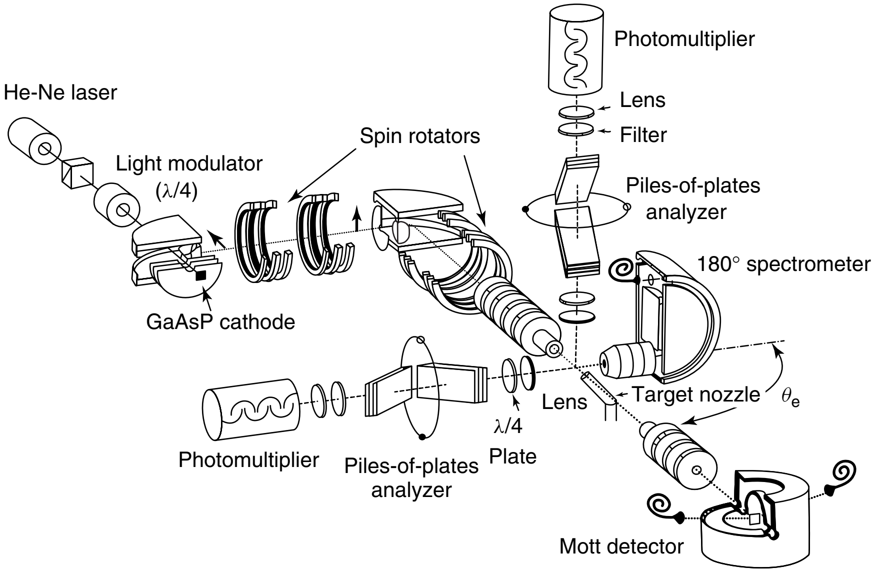
Photomultiplier
He-Ne laser
Lens
Filter
Spin rotators
Light modulator
(λ/4)
Piles-of-plates
analyzer
180° spectrometer
GaAsP cathode
Lens
Target nozzle
θ_e
λ/4
Plate
Photomultiplier
Piles-of-plates
analyzer
Mott detector

Fig. 8.17 The polarized-electron apparatus of Sohn and Hanne (1992).

to measure several spin asymmetries for elastic and inelastic e–Cs collisions. For such heavy targets, both exchange effects and the spin–orbit interaction are often important, and it is by no means trivial to disentangle the various effects. Some sophisticated recipes for performing complete, or almost complete, experiments were discussed by Andersen and Bartschat (1996;2001). The general idea is to look for parameters that would vanish if a particular interaction is not important for the outcome of the collision process. If such an observable is found to be nonzero experimentally, it is generally a good candidate to test this particular aspect of a theoretical model.

Figure 8.18 shows two such examples, namely, the integrated Stokes parameter P_2 for e–Hg excitation and the spin-asymmetry function S_A for elastic e–Cs scattering. Both observables would be zero in any nonrelativistic calculation. In both cases, we note excellent agreement with theoretical predictions that include relativistic effects. However, numerous examples remain where the agreement between theory and experiment is much less satisfactory, in particular for more complex targets such as the heavy noble gases.

Spin-polarized electron beams have also been used in the study of ionization processes, either in ionization asymmetries using a spin-polarized atomic beam as well (see, for example, Baum *et al.* (1993;1992)), or in studies where a p-electron is ejected from a noble gas such as Xe and the fine-structure of the residual ion is resolved (see Bellm *et al.* (2008) and references therein). Finally, relatively few polarized-electron experiments have been performed with molecular targets to date. In many cases, spin-effects are largely masked by the averaging over molecular orientations. However, Green *et al.* (2004) recently found an example of circularly polarized molecular fluorescence induced

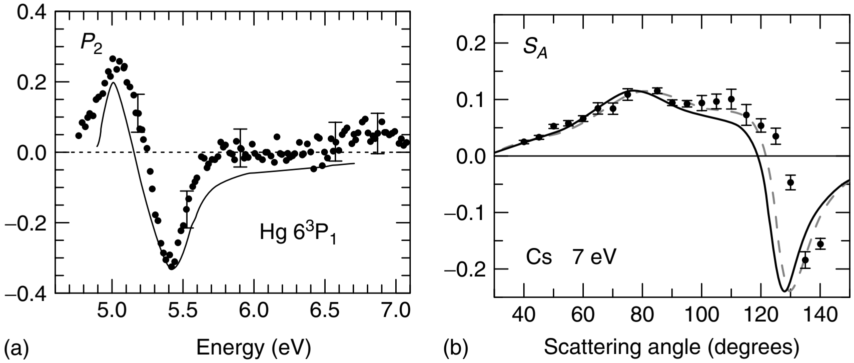

Fig. 8.18 (a) Angle-integrated Stokes parameter P_2 for polarized-electron-impact excitation of the $(6s6p)^3P_1$ state of Hg as a function of the incident electron energy. The experimental data of Wolcke *et al.* (1983) are compared with theoretical predictions from a Breit–Pauli *R*-matrix approach developed by Bartschat *et al.* (1984).
(b) Spin-asymmetry function S_A for elastic e–Cs scattering at an electron energy of 7 eV. The experimental data of Baum *et al.* (2002) are compared with fully relativistic CCC (solid line, Fursa and Bray (2008)) and fully relativistic Dirac *B*-spline *R*-matrix results (dashed line, Zatsarinny and Bartschat (2008)).

by spin exchange. Another pioneering experiment was performed by Mayer and Kessler (1995) who verified electron optic dichroism.

8.2.8
Positron Scattering

The scattering of positrons from atoms and molecules is an interesting area in itself, but is also important in testing some aspects of the theory of electron collisions as a result of the various comparisons and contrasts between the two situations. For example, the static potential associated with the Coulomb interaction is attractive for electrons and repulsive for positrons, while the polarization potential is attractive for both particles. Exchange effects occur only for electron collisions. On the other hand, for positrons, the rearrangement channel of positronium formation can become very important during the interaction.

A detailed discussion of positron-atom/molecule scattering lies outside the scope of this review. The reader is referred to the review by Surko, Gribakin and Buckman (2005) for details of the numerous recent experimental and theoretical developments in this field.

8.3
Review of Theoretical Methods

The ongoing rapid development of computational resources has resulted in a variety of new methods to treat electron/positron collisions with atoms, molecules, and ions. Some of these approaches are, at least in principle, universally applicable over the entire energy range of interest. Consequently, we will no longer use the traditional classification in terms of "low-energy" and "high-energy" methods. Instead, our first separation will be made according to whether an approach is based on the close-coupling expansion or on the Born series. While both expansions, if carried to infinity, would lead to exact results (within limits such as a nonrelativistic or a relativistic formulation), these expansions are cut off or approximated in practical applications and, via the details of the procedure, result in different classifications of a particular method.

Another important aspect concerns the traditional distinction between excitation, including elastic scattering and ionization. As is shown below, theoretical methods originally developed for excitation processes have been very successful in treating ionizing collisions as well. This is sometimes achieved by reinterpreting the results obtained for excitation of a positive-energy pseudostate as an ionization process. However, in that case, special measures are required to extract the information that is actually measured in the corresponding experiments.

8.3.1
The Close-Coupling Expansion

8.3.1.1 The Wave Equation, Scattering Amplitude, and Cross Section

We start by considering elastic scattering and excitation of atoms and atomic ions by electron impact in a nonrelativistic framework and consider the process

$$e^- + A_i \longrightarrow e^- + A_j \qquad (8.25)$$

where A_i and A_j are the initial and final N-electron bound states of the target.

The Schrödinger equation for the $(N+1)$-electron collision system (target electrons plus projectile) for a target with nuclear charge Z is given by

$$H_{N+1}\Psi = E\Psi \tag{8.26}$$

where E is the total energy of the system. In atomic units, the $(N+1)$-electron Hamiltonian H_{N+1} is

$$H_{N+1} = \sum_{i=1}^{N+1}\left(-\tfrac{1}{2}\nabla_i^2 - \frac{Z}{r_i}\right) + \sum_{i>j=1}^{N+1}\frac{1}{r_{ij}} \tag{8.27}$$

Here $r_{ij} \equiv |\boldsymbol{r}_i - \boldsymbol{r}_j|$, with $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ denoting the vector coordinates of the electrons i and j, respectively. The origin of the coordinate system is taken to be the target nucleus, which is assumed to have infinite mass.

Next, we introduce the target eigenfunctions Φ_i and corresponding eigenenergies w_i by

$$\langle\Phi_i|H_N|\Phi_j\rangle = w_i\delta_{ij} \tag{8.28}$$

where H_N is the target Hamiltonian defined by Eq. (8.27) with $N+1$ replaced by N.

We look for a solution of Eq. (8.26) corresponding to the process defined by Eq. (8.25), where an electron is incident upon the target in some state Φ_i and is scattered leaving the target in some other state Φ_j. The asymptotic form of the wave function in the case of a neutral target is

$$\Psi_i \xrightarrow[r\to\infty]{} \Phi_i \chi_{m_i} e^{ik_i z} + \sum_j \Phi_j \chi_{m_j} \times f_{ji}(\theta,\varphi)\frac{e^{ik_j r}}{r} \tag{8.29}$$

Here χ_{m_i} and χ_{m_j} are the spin eigenfunctions of the incident and scattered electrons and f_{ji} is the scattering amplitude. In numerical calculations, the direction of spin quantization is usually taken to be the incident beam direction (the so-called "collision system"). The wave numbers k_i and k_j of the incident and scattered electrons are related to the total energy E of the collision system by

$$E = w_i + \tfrac{1}{2}k_i^2 = w_j + \tfrac{1}{2}k_j^2 \tag{8.30}$$

The outgoing-wave term on the right of Eq. (8.29) contains contributions from all target states that are energetically allowed, that is, for which $k_j^2 \geq 0$. The remaining states, for which $k_j^2 < 0$, can only occur virtually in the collision. Although they do not contribute to the asymptotic form, they can play an important role during the process giving rise, for example, to intermediate resonance states. Finally, when the target is an ion, the exponents in Eq. (8.29) must be modified by including logarithmic phase factors due to the distortion by the Coulomb field. These factors introduce no essential complications, and hence we shall not include them explicitly below.

The angle-differential cross section for a transition from a state $|\boldsymbol{k}_i, \Phi_i, \chi_{m_i}\rangle$ to a state $|\boldsymbol{k}_j, \Phi_j, \chi_{m_j}\rangle$ can be calculated by considering the incident and scattered fluxes in Eq. (8.29). One finds

$$\frac{d\sigma_{ji}}{d\Omega} = \frac{k_j}{k_i}|f_{ji}(\theta,\varphi)|^2 \tag{8.31}$$

The fundamental problem is to solve the Schrödinger equation (8.26) to obtain the scattering amplitude and the cross section corresponding to the process given by Eq. (8.25). We start by expanding the total wave function in the form

$$\Psi_i^\Gamma(\boldsymbol{X}_{N+1}) = \mathcal{A}\sum_j \overline{\Phi}_j^\Gamma(\boldsymbol{X}_N;\widehat{\boldsymbol{r}}_{N+1}\sigma_{N+1})F_{ji}^\Gamma(r_{N+1}) + \sum_j \chi_j^\Gamma(\boldsymbol{X}_{N+1})a_{ji}^\Gamma \tag{8.32}$$

Here $\boldsymbol{X}_{N+1} = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N+1}$ represents the space and spin coordinates of all $N+1$ electrons, $\boldsymbol{x}_i \equiv \boldsymbol{r}_i\sigma_i$ denotes the space and spin coordinates of the ith electron, and $\mathcal{A}$ is the operator that antisymmetrizes the first summation with respect to any pair of electrons in accordance with the Pauli exclusion principle. The channel functions $\overline{\Phi}_j^\Gamma$ are obtained by coupling the orbital and spin angular momenta of the target with those of the scattered electron to form eigenstates of $\Gamma = LSM_LM_S$ and parity π that are conserved in the collision. The role of the square-integrable, fully antisymmetrized correlation functions χ_j^Γ is further discussed below.

Equation (8.32) is the so-called "close-coupling expansion." To be complete, it should include all discrete bound states as well as the continuum states of the target. In practice, of course, exact completeness is impossible to achieve in a numerical treatment, and hence the expansion has to be cut off. In the simplest case, only one target state might be retained, allowing for elastic scattering from this state to be described in the "static exchange approximation." Traditionally, the expansion was cut after a few discrete target states, thereby enabling the calculation of transitions between all the states retained in the expansion. However, it was soon realized that such a severe cut in the expansion typically limits the applicability of the method to "low" energies, generally below the threshold of the highest target state kept in the expansion. Consequently, significant work has been devoted in recent years to account in some approximate way for the high-lying discrete and even the continuum states of the target. This resulted in the convergent close-coupling (CCC) method, originally developed by Bray and Stelbovics (1992), and the R-matrix with Pseudostates (RMPS) method proposed by Bartschat *et al.* (1996b) based on the same idea. Both methods are further discussed below.

By substituting Eq. (8.32) into the Schrödinger equation (8.26), projecting onto the channel functions $\overline{\Phi}_j^\Gamma$ and onto the square-integrable functions χ_j^Γ, and eliminating the coefficients a_{ji}^Γ, one obtains coupled integro-differential equations for the reduced radial functions F_{ji}^Γ representing the motion of the scattered electron. These are of the form

$$\left(\frac{d^2}{dr^2} - \frac{l_j(l_j+1)}{r^2} + \frac{2Z}{r} + k_j^2\right)F_{ji}^\Gamma(r) = 2\sum_l \left(V_{jl}^\Gamma + W_{jl}^\Gamma + X_{jl}^\Gamma\right)F_{li}^\Gamma(r) \tag{8.33}$$

where l_j is the orbital angular momentum of the scattered electron and V_{jl}^Γ, W_{jl}^Γ, and X_{jl}^Γ are partial-wave decompositions of the local direct, nonlocal exchange, and nonlocal correlation potentials, respectively.

The direct potential can be written as

$$V_{ij}^\Gamma(r_{N+1}) = \left\langle \overline{\Phi}_i^\Gamma(\boldsymbol{X}_N;\, \widehat{\boldsymbol{r}}_{N+1}\sigma_{N+1}) \right| \sum_{i=1}^N \frac{1}{r_{i\,N+1}} \left| \overline{\Phi}_j^\Gamma(\boldsymbol{X}_N;\widehat{\boldsymbol{r}}_{N+1}\sigma_{N+1}) \right\rangle \tag{8.34}$$

where the integral is taken over all $N+1$ electron space and spin coordinates, except for the radial coordinate of electron "$N+1$." This potential has the asymptotic form

$$V_{ij}^{\Gamma}(r) = \frac{N}{r}\delta_{ij} + \sum_{\lambda=1}^{\lambda_{\max}} a_{ij}^{\lambda} r^{-\lambda-1}, \quad r > a \tag{8.35}$$

where a is the range beyond which the bound orbitals in $\overline{\Phi}_i^{\Gamma}$ and $\overline{\Phi}_j^{\Gamma}$ are negligible. The long-range potentials with $\lambda \geq 1$ play an important role in low-energy electron–atom scattering.

The exchange and correlation potentials, unlike the direct potential, are both nonlocal, with the exchange potential vanishing exponentially for large r. Explicit expressions for these potentials are complicated and, except for hydrogen-like targets, are not written down explicitly in practice. Instead, they are calculated by general computer programs.

To determine the cross section, we look for solutions of Eqs. (8.33) satisfying $\boldsymbol{K}$-matrix asymptotic boundary conditions:

$$F_{ji}^{\Gamma}(r) \xrightarrow[r\to\infty]{} k_j^{-1/2}(\sin\theta_j\delta_{ji} + \cos\theta_j K_{ji}^{\Gamma})$$

for open channels ;

$$F_{ji}^{\Gamma}(r) \xrightarrow[r\to\infty]{} 0,$$

for closed channels (8.36)

Here

$$\theta_j = k_j r - \tfrac{1}{2} l_j \pi + \frac{z}{k_j}\ln 2k_j r + \sigma_j \tag{8.37}$$

with $z = Z - N$ as the residual charge on the ion and $\sigma_j = \arg\Gamma(l_j + 1 - iz/k_j)$ denoting the Coulomb phase shift. The (scattering) $\boldsymbol{S}$-matrix is related to the (transition) $\boldsymbol{T}$-matrix and the (reactance) $\boldsymbol{K}$-matrix with elements defined by Eqs. (8.36) by the matrix equations

$$\boldsymbol{S}^{\Gamma} = \boldsymbol{T}^{\Gamma} - \boldsymbol{1} = \frac{\boldsymbol{1} + i\boldsymbol{K}^{\Gamma}}{\boldsymbol{1} - i\boldsymbol{K}^{\Gamma}} \tag{8.38}$$

The hermiticity and time-reversal invariance of the Hamiltonian ensure that $\boldsymbol{K}^{\Gamma}$ is real and symmetric while $\boldsymbol{S}^{\Gamma}$ is unitary and symmetric.

The scattering amplitude defined by Eq. (8.29) and hence the differential cross section can be readily related to the $\boldsymbol{S}$-matrix. The resultant expression is given by Burke (1985). Computer programs to calculate scattering amplitudes, the DCS, and many other observables were published by Bartschat and Scott (1983), Bartschat (1983); Bartschat (1996), and Grum-Grzhimailo (2003). The total cross section is obtained by averaging the differential cross section over initial spin states, summing over final spin states, and integrating over all scattered electron angles. One obtains

$$\sigma_{\text{tot}}(i \to j) = \frac{\pi}{k_i^2} \sum_{\substack{LS\pi \\ l_i l_j}} \frac{(2L+1)(2S+1)}{2(2L_i+1)(2S_i+1)} \times |S_{ji}^{\Gamma} - \delta_{ji}|^2 \tag{8.39}$$

for a transition from an initial target state $\alpha_i L_i S_i$ to a final target state $\alpha_j L_j S_j$, where α_i and α_j represent any additional quantum numbers that are required to define the initial and final target states. It is also useful to define the collision strength by

$$\Omega(i,j) = k_i^2(2L_i+1)(2S_i+1)\sigma_{\text{tot}} \times (i \to j) \tag{8.40}$$

It is dimensionless and symmetric with respect to interchange of the initial and final states denoted by i and j.

There are many different ways to solve the close-coupling equations in practice. These include packages based on the R-matrix method, on the reduction of Eqs. (8.33) to linear algebraic equations, on the noniterative integral equations method, and on the Kohn variational method. These computational methods and early associated program packages were reviewed by Burke and Seaton (1971) and by Burke and Eissner (1983). The R-matrix method, in particular, has been applied in numerous calculations. An overview of the method and key papers in the development can be found in Burke and Berrington (1993) and in Burke, Noble and Burke (2006). The most important recent developments of the R-matrix method include the RMPS formulation mentioned above Bartschat *et al.* (1996a), Gorczyca and Badnell (1997), and an entirely independent formulation developed by Zatsarinny and coworkers. The program package BSR (Zatsarinny (2006)) is based on the use of basis splines ("B-splines") as a universal and effectively complete basis to expand the wave function of the projectile inside the R-matrix box. In addition, great flexibility is achieved by allowing for the use of term-dependent, nonorthogonal sets of orbitals to describe the target structure. As is shown by the examples below, the method is particularly advantageous in the treatment of electron collisions with complex atoms, such as noble gases and target states involving several open shells. In these cases, it is well worth to pay the price of a more complicated setup of the Hamiltonian matrix, in return for a highly accurate target description with a relatively small number of individually optimized configurations.

8.3.1.2 **Extension of the Close-Coupling Method to Intermediate and High Energies**

As mentioned above, a severe limitation to the applicability of the close-coupling method was the traditional cut-off of the expansion after a few low-lying target states. Early attempts to address this problem augmented the expansion (8.32) by adding a few well-chosen pseudostates to represent the remaining infinity of omitted eigenstates (Burke and Schey (1962); Damburg and Karule (1967); Burke and Webb (1970); Callaway and Wooten (1973)). These pseudostates are not target eigenstates but instead are chosen to optimize the representation of the short-range correlation effects or long-range polarization effects. In this way, part of the continuum is included in expansion (8.32), and the pseudostates allow approximately for loss of flux into all open channels including the ionizing channels.

Although a number of calculations had been carried out using this approach, particularly for e–H scattering, early calculations suffered from the appearance of unphysical pseudostate thresholds and pseudoresonances in the intermediate energy region of interest. Burke, Noble and Scott (1987) proposed an intermediate-energy R-matrix (IERM) theory, in which two electrons are allowed to occupy continuum orbitals in the wave function expansion. This method gave accurate results for electron and positron scattering by hydrogen atoms, although special care had to be taken in the averaging of unphysical structures in the $\boldsymbol{T}$-matrix elements at intermediate energies.

Bray and Stelbovics (1992) finally achieved a breakthrough by developing the CCC method. The basic idea is to include a large number of systematically generated square-integrable pseudostates in the close-coupling expansion and check for convergence of the final results. Such L^2 methods were already suggested and reviewed by Reinhardt (1979) and Broad (1983), but they met with various difficulties at the time. The CCC method was formulated in momentum space, based on the Lippman-Schwinger equation for the $\boldsymbol{T}$-matrix. To date, the method has been applied with great success to electron and positron collisions with hydrogen-like Bray and Stelbovics (1995) and helium-like Fursa and Bray (1995) systems, including electron-impact ionization as well as single and double ionization by photon impact Bray *et al.* (2002). If necessary, inner shells are represented by sophisticated core potentials. The most recent extension of the method is a fully relativistic formulation for scattering from quasi-one-electron systems such as Cs Fursa and Bray (2008).

An equivalent method based on the R-matrix approach in coordinate space is the "R-matrix with Pseudostates" (RMPS) method proposed by Bartschat *et al.* (1996b) and further developed by Gorczyca and Badnell (1997). Once again, many square-integrable states, obtained by diagonalizing the target hamiltonian in a given basis (e.g., Sturmian or Laguerre type), are included in the close-coupling expansion. When used in connection with the standard expansion of the projectile wave function in terms of a basis set, the method is generally restricted to low and intermediate energies. However, the computer program is general and hence accurate results can readily be obtained for complex targets as well.

The above developments have, to a large extent, replaced the use of optical-potential methods based on the close-coupling expansion. However, semiempirical potentials are sometimes employed to represent the effect of inner closed shells, core polarization, and even exchange. Some examples are discussed below.

8.3.1.3 Inclusion of Relativistic Effects

As the nuclear charge Z of the target increases, relativistic effects become important even for low-energy electron scattering. There are two ways in which relativistic effects play a role. First, there is a direct effect corresponding to the relativistic distortion of the wave function describing the scattered electron by the strong nuclear Coulomb potential. Second, there is an indirect effect caused by the change in the charge distribution of the target due to the use of relativistic target wave functions.

For atoms and ions with small and intermediate Z values, the scattering calculations can first be performed in LS-coupling; the $\boldsymbol{K}$-matrices are subsequently recoupled to yield transitions between fine-structure levels. For atoms and ions with high$Z$ values, relativistic terms must be retained in the Hamiltonian. One method of achieving this is to use the Breit–Pauli Hamiltonian

$$H_{N+1}^{\mathrm{BP}} = H_{N+1}^{\mathrm{NR}} + H_{N+1}^{\mathrm{REL}} \tag{8.41}$$

where H_{N+1}^{NR} is the nonrelativistic Hamiltonian given by Eq. (8.27) and H_{N+1}^{REL} contains the one-body spin–orbit, Darwin, and mass correction terms, as well as possibly two-body relativistic terms. The total wave function is expanded in a form analogous to Eq. (8.32), except that the target states must also be generated using the Breit–Pauli Hamiltonian, and the wave function is an eigenstate of the total electronic angular momentum J, its z-component M_J, and the parity π rather than of LSM_LM_S and π.

Alternatively, for high-Z atoms and ions, the Dirac-Coulomb Hamiltonian

$$H^D_{N+1} = \sum_{i=1}^{N+1}\left(c\boldsymbol{\alpha}\cdot\boldsymbol{p}_i + \beta c^2 - \frac{Z}{r_i}\right) + \sum_{i>j=1}^{N+1}\frac{1}{r_{ij}} \tag{8.42}$$

may be used, where $\boldsymbol{\alpha}$ and β are the usual Dirac matrices and c is the speed of light.

Figure 8.18 showed two examples of close-coupling results obtained with the Breit–Pauli R-matrix method for e − Hg scattering by Bartschat *et al.* (1984), who included the $(6s^2)^1S^0$, $(6s6p)^3P^o_{0,1,2}$, and $(6s6p)^1P^o_1$ target states in the expansion of the total wave function, and with the fully relativistic CCC (Fursa and Bray (2008)) and the fully relativistic B-spline R-matrix method (Zatsarinny and Bartschat (2008)) for elastic e − Cs collisions. The good agreement between experiment and theory indicates that the three main effects of strong channel coupling, electron exchange, and the spin–orbit interaction are correctly represented by the theories in these cases. However, these are still comparatively simple targets resembling quasi-one-electron and quasi-two-electron systems. Significantly, more challenges are expected for heavy noble-gas targets such as Kr and Xe, or for cases in which inner-shell excitation plays an important role.

8.3.1.4 **Illustrative Results**

Figure 8.19 shows the angle-integrated cross section for elastic scattering and electron-impact excitation from the $(1s)^2S$ state of atomic hydrogen. The excellent agreement between predictions from the CCC, RMPS, and IERM methods, as well as with the experimental data of Williams (1988), indicates that these sophisticated methods can indeed solve this problem to a very high degree of accuracy.

Stepanovic *et al.* (2006) recently measured the angle-integrated cross sections for electron-impact excitation of the 3^3S and 3^1S states of helium shown in figure 8.20. For a direct comparison with the experimental data, the 69-state B-spline R-matrix (BSR-69) results, originally generated on a very fine energy mesh with a stepsize of 10^{-5} Ry, were convoluted with a Gaussian function of width 37 meV (FWHM), corresponding to the energy resolution in the experiment. The dashed line in the top graph for the 3^3S state represents the theoretical predictions for direct excitation, while the solid lines represent the emission cross sections, that is, including the cascade contributions from the higher-lying $n = 3, 4, 5$ states. The experimental data were normalized to the BSR-69 results with cascade contribution at 23.20 eV for both the 3^3S and 3^1S states, because the cross sections exhibit only a smooth energy dependence around this energy and hence the energy resolution does not play a role. For a detailed analysis of the resonances (see Stepanovic *et al.* (2006)), the right part of the figure exhibits the theoretical excitation cross sections without convolution with the experimental energy resolution.

Figure 8.21 exhibits results for the production of metastable neon atoms.

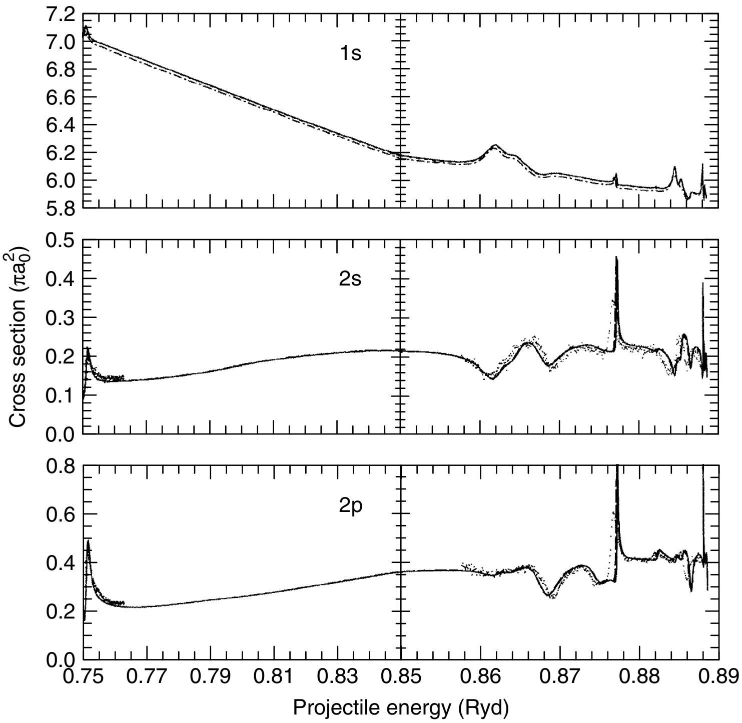

Fig. 8.19 Elastic and excitation cross sections for e–H scattering at incident electron energies between the $n = 2$ and $n = 3$ thresholds. The RMPS (solid line), IERM (dashed line), and CCC (dash-dotted line) curves are barely distinguishable. The small dots represent the experimental data of Williams (1988). (From Bartschat *et al.* (1996c).)

After renormalizing the published experimental data of Buckman *et al.* (1983) (within the given overall experimental uncertainty) and including the cascade contributions, the semirelativistic BSR method of Zatsarinny and Bartschat (2004) obtains essentially perfect agreement with experiment regarding the energy dependence in the near-threshold region. These results represent a considerable improvement over previous theoretical predictions. The predicted metastable cross sections were then used to calibrate and interpret new experimental data obtained with an energy resolution of ≈4 meV by Bömmels *et al.* (2005). For example, a very sharp Feshbach resonance was theoretically predicted and experimentally confirmed at 18.527 eV.

The enormous importance of FeII, the singly ionized species of the iron element, for astronomical observations is well known. A meaningful interpretation of the spectral lines critically depends on an accurate knowledge of the electron-induced collision strengths, as well as the so-called "effective collision strengths," which are obtained by integrating the energy-dependent results over a Maxwellian distribution for a range of electron temperatures. Because

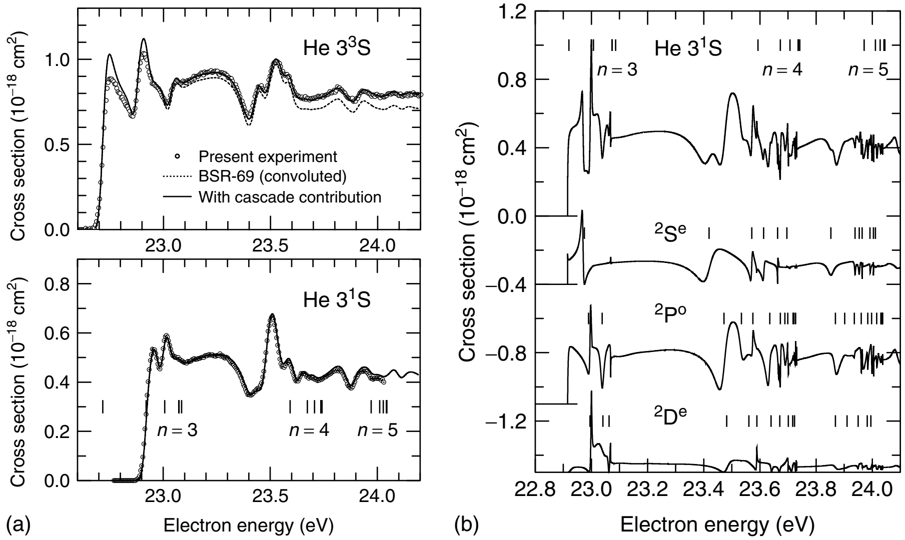

Fig. 8.20 (a) Angle-integrated cross sections for electron-impact excitation of the 3^3S and 3^1S states in helium. The theoretical results without (dashed line) and with (solid line) cascade contributions were convoluted with a Gaussian of 37 meV (FWHM). The vertical bars in the lower panel represent the thresholds for the He target states. (b) Angle-integrated cross section for electron-impact excitation of the 3^1S state. The thick solid line at the top exhibits the predictions from a 69-state BSR model summed over all partial waves, while the remaining thin solid lines (offset to increase the visibility) show the most important individual partial-wave contributions. The vertical bars represent the thresholds for the He target states at the top and prominent He$^-$ resonances found in a detailed partial-wave analysis. (From Stepanovic *et al.* (2006).)

of the importance of these data and the complete lack of experiments, major theoretical and computational efforts have been devoted to this system for more than two decades.

Figure 8.22 exhibits results for the collision strength for the (3d^{6}4s)a^{6}D – (3d^7)a^{4}F transition. Predictions of Ramsbottom *et al.* (2004) and Zatsarinny and Bartschat (2005) for contributions to the collisions strength from selected partial-wave symmetries are compared and exhibit overall satisfactory agreement. The difference in the intensity of some narrow resonances may be caused by the fact that finer energy steps of 10^{-5} Ry were used in the BSR calculations than the 10^{-4} Ry energy mesh chosen by Ramsbottom *et al.* (2004). A notable exception is a broad structure in the ^{5}D^e partial wave around 0.15 Ry. While there are differences at such detailed level, good overall agreement among the calculations was obtained after summing up the individual partial-wave contributions and integrating over a Maxwellian distribution. Because there are no experimental data available for these cases, such independent checks are very important in assessing the reliability of the data before providing them to the astrophysical community.

Having established reliable results in a nonrelativistic framework, an important next step involves fine-structure

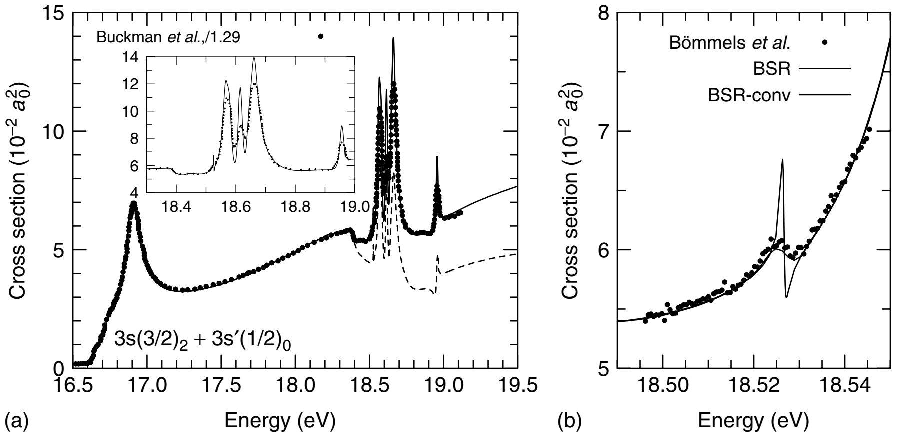

Fig. 8.21 (a) Angle-integrated cross section for production of neon atoms in the metastable $3s[3/2]_2$ and $3s'[1/2]_0$ states (from Zatsarinny and Bartschat (2004).) The experimental data of Buckman *et al.* (1983) (thick dots) were re-normalized by a factor of 1.29 to provide a good visual fit to the BSR theory at energies just above the excitation threshold. The solid line includes the cascade contributions from all the states included in the model, while the thin dotted line (starting around 18.4 eV) represents the results without cascades. (b) A very narrow resonance predicted by Zatsarinny and Bartschat (2004) and later found in the experiment of Bömmels *et al* (2005).

transitions. Ramsbottom *et al.* (2007) recently carried out a 262-state Breit–Pauli *R*-matrix calculation with about 1800 coupled channels to study fine-structure forbidden transitions in e–FeII collisions. This knowledge is required in many astrophysical applications.

FeII is just one example of a target with an open d-shell, which represents a major challenge to atomic collision theory. On the other hand, collisions systems involving such targets are very important in practical applications. An example is e – Mo, which may be a suitable candidate for a mercury-free lighting source (Petrov *et al.* (2004); Bartschat *et al.* (2004a)). Another case of interest involves electron collisions with tungsten and its ions, which are important impurities in fusion devices such as ITER.

Finally, as mentioned above, an important development has been the extension of the close-coupling method to intermediate and even high energies, and the inclusion of relativistic effects, either by recoupling of nonrelativistic results, perturbatively through the Breit–Pauli hamiltonian, or in a fully relativistic treatment. Figure 8.23 shows the total cross section for electron collisions with Cs atoms. Overall, there is good agreement between results from a nonrelativistic CCC model, a semirelativistic Breit–Pauli *R*-matrix approach, and a fully relativistic Dirac-based *B*-spline *R*-matrix treatment.

8.3.1.5 **Resonances and Quantum Defect Theory**

We have already seen in previous sections that resonances play an important role in low-energy scattering of electrons by atoms and ions. The general theory of resonances has been discussed by Fano (1961), Fano (1983),

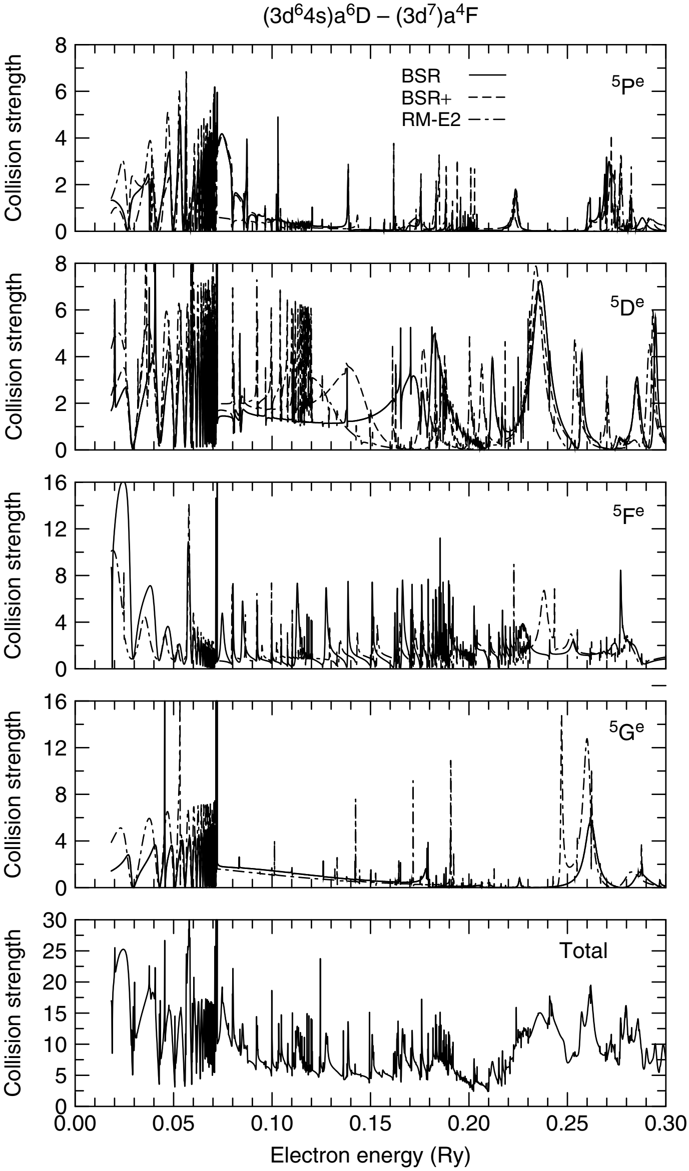

Fig. 8.22 Important partial-wave contributions, indicated in the top four panels, and total (bottom) collision strength for the electron-induced transition (3d^{6}4s)^{6}D – (3d^7)^{4}F in Fe$^+$. (From Zatsarinny and Bartschat (2005).)

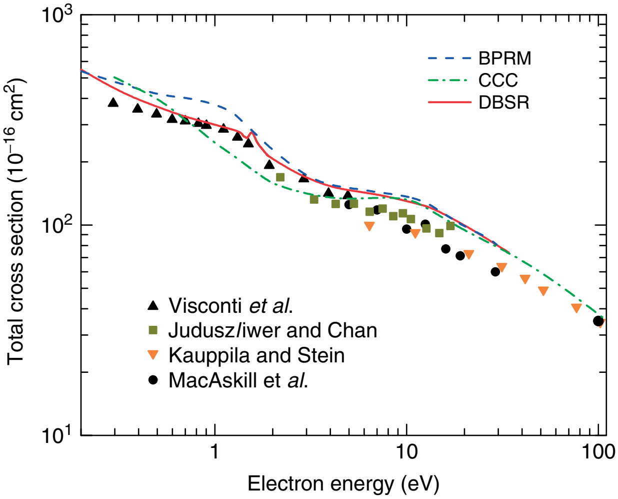

Fig. 8.23 Total electron scattering cross section from the ground $(6s)^2S$ ground state of Cs. The 12-state DBSR results of Zatsarinny and Bartschat (2008) are compared with various sets of experimental data (Visconti, Slevin and Rubin (1970), Jaduszliwer and Chan (1992), Kauppila and Stein (2001), MacAskill *et al.* (2002)), as well as predictions from a 40-state semirelativistic Breit–Pauli *R*-matrix (BPRM)calculation by Bartschat and Fang (2000) and a nonrelativistic CCC model presented by MacAskill *et al.* (2002). (From Zatsarinny and Bartschat (2008).)

Burke (1968), Schulz (1973a), and Schulz (1973b). Here, we review their principal features.

It is useful to distinguish two types of resonances, namely, open-channel or "shape" resonances, and closed-channel or "Feshbach" resonances. Shape resonances occur when the effective potential between the electron and the target has a characteristic shape with an inner attractive well and an outer repulsive barrier, the latter usually caused by the centrifugal repulsion. Examples of such resonances include a dominant *p*-wave shape resonance in low-energy elastic e – Mg collisions Bartschat *et al.* (2004b) and a broad *d*-wave shape resonance in e – CH_4 scattering Gianturco and Scialla (1987).

On the other hand, Feshbach resonances occur for incident electron energies just below excitation thresholds. The scattered electron is captured temporarily into a bound state with the target in an excited state. We have seen several examples of such resonances in the previous sections.

Calculations based on expansion (8.32) are a very convenient and appropriate way of describing Feshbach resonances, provided that the relevant open and closed channels are included in the expansion. For example, bound states caused by the long-range attractive Coulomb potential in the closed channels for electron–ion scattering give rise to Rydberg series of resonances when the coupling to the open channels is included.

One of the motivations for the development of multichannel quantum defect theory (MQDT) (Seaton (1958); Seaton (1983)) is to treat entire series of resonances rather than an individual resonance. Seaton showed that the $\boldsymbol{K}$-matrix defined by Eq. (8.36) can be written as

$$\boldsymbol{K}^{\Gamma} = \boldsymbol{K}^{\Gamma}_{\mathrm{oo}} - \boldsymbol{K}^{\Gamma}_{\mathrm{oc}}(\boldsymbol{K}^{\Gamma}_{\mathrm{cc}} + \tan\pi\boldsymbol{\nu}_c)^{-1}\boldsymbol{K}^{\Gamma}_{\mathrm{co}} \tag{8.43}$$

where the elements of the submatrices $\boldsymbol{K}^{\Gamma}_{\mathrm{oo}}$, $\boldsymbol{K}^{\Gamma}_{\mathrm{oc}}$, $\boldsymbol{K}^{\Gamma}_{\mathrm{co}}$, and $\boldsymbol{K}^{\Gamma}_{\mathrm{cc}}$ are obtained by analytically continuing the $\boldsymbol{K}$-matrix from above the threshold to which the resonances converge to below this threshold, where "o" refers to the channels that remain open below this threshold, and "c" refers to the channels that become closed. In these closed channels ($k_i^2 < 0$), the components ν_i of the vector $\boldsymbol{\nu}_c$ are defined by

$$k_i^2 = -(Z-N)^2/\nu_i^2 \tag{8.44}$$

The factor $\tan\pi\boldsymbol{\nu}_c$ causes the last term in Eq. (8.43) to have an infinite series of poles that give rise to the Feshbach resonances. When the $\boldsymbol{K}$-matrix is calculated by solving Eqs. (8.36) at a few energies above threshold, it can be extrapolated to yield an infinite series of resonances below threshold. This approach was extended by Gailitis (1963) to predict the average behavior of electron–ion cross sections in resonance regions.

8.3.1.6 **Dielectronic Recombination**

It was pointed out by Burgess (1964); Burgess (1965) that Rydberg resonance series play a fundamental role in dielectronic recombination. This process is illustrated, for example, in the case of $\mathrm{e} - \mathrm{C}^{3+}$ scattering by the equation

$$\begin{aligned}\mathrm{e}^- + \mathrm{C}^{3+}(1\mathrm{s}^2 2\mathrm{s}) &\rightarrow \mathrm{C}^{2+*}(1\mathrm{s}^2 2\mathrm{p}\; nl)\\ &\rightarrow \mathrm{C}^{2+}(1\mathrm{s}^2 2\mathrm{s}\; nl) + h\nu\end{aligned} \tag{8.45}$$

The electron is captured into a member of a Rydberg series of resonances converging to the $1\mathrm{s}^2 2\mathrm{p}$ excited state of the C^{3+} ion. These resonances can stabilize by radiative decay, thereby leaving the C^{2+} ion in an excited bound state. This process can often yield dielectronic recombination coefficients α^{DR} that are two orders of magnitude larger than radiative recombination coefficients, basically because of the very large number of resonance levels that contribute.

In the isolated-resonance approximation, α_i^{DR} is given by

$$\alpha_i^{\mathrm{DR}} = \left(\frac{2\pi}{kT}\right)^{3/2} \sum_n \frac{g_n}{2g_i} \frac{\Gamma_n^a \Gamma_n^r}{\Gamma_n^a + \Gamma_n^r} e^{-E_n/kT} \tag{8.46}$$

where i refers to the initial state of the target ion and n to the intermediate resonance state. The coefficient g is the statistical weight of the state, Γ_n^a is the autoionization width, Γ_n^r is the radiative width of the resonance state, and E_n is the energy of the resonance state relative to the ground state of the target ion. The velocity distribution of the electrons is assumed to be Maxwellian corresponding to a temperature T. The summation over n in Eq. (8.46) converges for large n because Γ_n^a behaves as n^{-3} along a Rydberg resonance series while Γ_n^r becomes independent of n. In practice, very high n values often contribute.

Dielectronic recombination measurements have been made on a number of ions, while many calculations have been carried out illustrating the importance of this recombination process in hot gaseous plasmas. More information, including an extensive list of references, can be found in Pindzola, Griffin and Badnell (2006).

8.3.2
Born Series Methods

8.3.2.1 Elastic Scattering and Excitation of Atoms and Ions at Intermediate and High Energies

We now consider electron–atom and electron–ion scattering at electron impact energies greater than the ionization threshold of the target. In this case, an infinite number of channels are open, so that they cannot all be included in the expansion of the total wave function. Several approaches have been developed to treat scattering at these energies. Before the recent extensions of traditionally low-energy methods based on the close-coupling expansion, variants of the Born approximation were developed to address such problems. Being computationally much less demanding than CCC, RMPS, and the "initial-value methods" to be discussed below, the "distorted-wave"(DW) method in particular continues to be of practical importance for applications to complex targets, especially for ionization processes. The method has also proven to be an accurate way of calculating partial-wave cross sections for high angular momenta, and thus to obtain converged results with the number of partial waves. Finally, by studying the effect of different distortion potentials, including varying orders of the perturbation, and moving (parts of) the perturbation to the exact treatment by including it into the calculation of the scattering wave function, it is possible to study the importance of individual physical effects on the outcome of a collision process. Such effects may include the polarization of the atomic charge cloud due to the incident projectile, exchange, absorption (i.e., loss of flux into inelastic channels), or the long-range Coulomb interaction between three (or more) charged particles in the final state.

8.3.2.2 The Born Series

In the "high-energy" domain, which extends from several times the ionization threshold upward, methods based on the Born series are appropriate. Ignoring exchange for the moment, the Born series for the direct scattering amplitude can be written as (Walters (1984); Joachain (1987))

$$f = \sum_{n=1}^{\infty} f_{\mathrm{B}n} \tag{8.47}$$

The nth Born term $f_{\mathrm{B}n}$ contains the interaction V_{pt} between the scattered electron and the target atom n times and the direct Green's operator, $G_0^+ = (E - K_{\mathrm{p}} - H_N + i\varepsilon)^{-1}$, $n-1$ times. Here H_N is again the N-electron target hamiltonian while K_{p} is the kinetic energy operator for the projectile.

It is important to retain consistently all terms in the Born series with similar energy and momentum-transfer ($\Delta = |\boldsymbol{k}_i - \boldsymbol{k}_j|$) dependences. For elastic scattering, the scattering amplitude converges to the first Born approximation for all Δ, but at lower energies, it is necessary to include $\mathrm{Re}\{f_{\mathrm{B3}}\}$ as well as the second Born terms to obtain the cross section correct to k^{-2}. In the forward direction, convergence to the first Born approximation is slow because of contributions from $\mathrm{Im}\{f_{B2}\}$, which allow for loss of flux into the open channels.

For inelastic scattering, the first Born approximation does not give the correct high-energy limit for large Δ. Instead, this comes from $\mathrm{Im}\{f_{B2}\}$. Physically we can understand this by noting that inelastic scattering into large scattering angles involves a collision with the nucleus, followed or preceded by an inelastic collision with the bound electron to give excitation. In the forward direction, $\mathrm{Re}\{f_{B3}\}$ must again be retained to yield the high-energy limit correct to k^{-2}.

Because the third Born term is difficult to calculate directly, Byron and Joachain (see Joachain (1987)) suggested that, to third order, the scattering amplitude should be calculated using the eikonal Born series (EBS)

$$f_{\mathrm{EBS}} = f_{\mathrm{B1}} + f_{\mathrm{B2}} + f_{\mathrm{G3}} + g_{\mathrm{Och}} \tag{8.48}$$

where f_{G3} is the third-order term in the expansion of the Glauber amplitude (Glauber (1959)) in a power series of the projectile–target interaction V_{pt}, while g_{Och} is the Ochkur electron exchange amplitude (Ochkur (1964)). The EBS method has been very successful when perturbation theory converges rapidly, that is, at high energies, at small and intermediate angles, and for light atoms.

8.3.2.3 The Distorted-Wave (DW) Method

Distorted-Wave (DW) methods are characterized by the fact that the interaction is separated into two parts, one of which is treated exactly while the other is included using first-order perturbation theory. A semirelativistic first-order DW approximation (DWBA) was introduced by Madison and Shelton (1973) and later modified by Bartschat and Madison (1987) to treat electron collisions with heavy noble gases and also with mercury. Other variants include the first-order many-body theory (FOMBT) of the Los Alamos group Machado, Leal and Csanak (1982), the semirelativistic version developed and described by Dasgupta, Blaha and Giuliani (2000), and the relativistic distorted-wave (RDW) method of the Toronto group Zuo, McEachran and Stauffer (1991). The basic idea behind all these methods is to calculate distorted waves for the projectile electron by solving the differential equation

$$\left[\frac{d^2}{dr^2} - \frac{l(l+1)}{r^2} - 2\left\{U(r) + V_r(r) - E\right\}\right] \chi_{E,l}(r) = 0 \tag{8.49}$$

Here $U(r)$ is the static Coulomb potential, often modified by including terms to account for electron exchange and the charge distortion (polarization) of the target, while V_r contains relativistic effects. [Instead of (8.49), the fully relativistic RDW solves the corresponding Dirac equation.] The distorted waves are then used to calculate transition matrix elements, treating to first order the part of the projectile–target interaction not explicitly included in the distortion potentials.

A detailed description of the DWBA method, including a computer program, was provided by Madison and Bartschat (1996). Here we only mention the principal advantages and disadvantages. The DWBA approach has proven particularly useful for electron scattering by highly ionized ions where it is assumed that only the two channels corresponding to

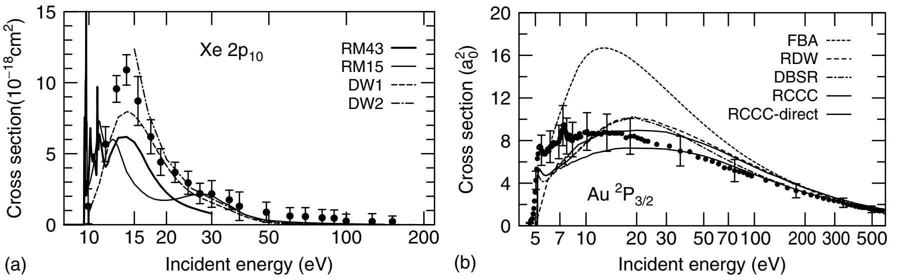

Fig. 8.24 Angle-integrated cross section for electron-impact excitation of the ($5p^5 6p$) "$2p_{10}$" state in Xe (a) and the $(6p)^2 P_{3/2}$ state in Au (b). The figures show the comparison of various DW and Breit–Pauli *R*-matrix (RM) predictions for Xe and results from a plane-wave first-order Born approximation (FBA), as well as the relativistic distorted-wave RDW, relativistic Dirac *B*-spline *R*-matrix (DBSR), and relativistic convergent close-coupling (RCCC) for Au with experimental data from the Wisconsin (Xe) and Flinders (Au) groups. The difference between the RCCC and RCCC-direct results is due to cascade effects. (Adapted from Bartschat, Dasgupta and Madison (2004) for Xe and from Maslov *et al.* (2008) for Au.)

the initial and final states in Eq. (8.32) need to be considered. The dominant diagonal potentials are treated exactly and the off-diagonal potentials coupling these channels are accounted for to first order. The method becomes more accurate at intermediate energies as $z = Z - N$ increases. As mentioned previously, the DWBA has also proven to be an accurate way of calculating partial-wave cross sections for high angular momenta. The angular momentum barrier in this case depresses the importance of the coupling potential.

The DWBA approach can also be a useful way of predicting the total and differential cross sections of inelastic electron–atom at intermediate energies well removed from threshold. Furthermore, the method is fast and relatively easy to implement. In principle, it allows for significant flexibility in the target description for the initial and final states because the orbitals can be specifically optimized for the states of interest in a given transition.

The major problem with the DW approach is the fact that channel coupling is neglected. Consequently, the method cannot account for near-threshold Feshbach-type resonances, it is generally much less reliable for optically forbidden than for optically allowed transitions, and further problems can occur if the calculation is performed without explicit unitarization of the scattering matrix; see Dasgupta *et al.* (2001).

Figure 8.24 exhibits a comparison of results for angle-integrated cross sections obtained in a variety of Born-based and close-coupling-based approximations. In both cases, the first-order DW models do quite well compared to the close-coupling approximations, while the first-order plane-wave approach generally overestimates the cross section near the maximum. However, in many cases of practical interest, the (plane-wave) FBA can still be used to get reasonably accurate results for excitation of optically allowed transitions by performing a rescaling transformation of the results as suggested by Kim (2001).

8.3.3
Initial-Value Methods

In this section, we briefly discuss the basic idea behind two highly successful methods developed using a somewhat different philosophy. These are the TDCC and the ECS approaches. The goal of both methods is the construction of a highly accurate numerical solution of the Schrödinger equation by propagating a known initial state. The matching to sometimes highly complicated boundary conditions, especially for the ionization processes to be discussed below, is avoided through a different way of extracting the relevant information from the wave function after large propagation times or sufficiently far away from the interaction region.

8.3.3.1 Time-Dependent Close-Coupling (TDCC)

With the recent development of massively parallel supercomputers, the use of more direct numerical approaches has become very popular. One example is the TDCC method recently reviewed by Pindzola *et al.* (2007). For a target with one or two electrons outside a closed shell, they derive the time-dependent close-coupling equations for the radial parts of the wave function describing the projectile and the active target electron or electrons. For each partial-wave symmetry, these equations are propagated in time over a two- or three-dimensional lattice.

The initial wave function is written as the properly symmetrized product of a bound state, represented on the lattice, and a Gaussian wavepacket for the projectile. The width of the package is chosen in such a way that any overlap between the projectile and the target wave function before the collision can be neglected. The electron wave function is then propagated for a sufficiently long time to ensure that the final state after the collision has been reached. However, note that the propagation time has to be small enough to ensure that boundary effects, due to collisions with the wall of the lattice, do not disturb the result. Cross sections for elastic scattering, excitation and ionization are obtained from projections of the final-state wave function on individual asymptotic channel functions.

The principal advantage of the method lies in the fact that the radial part of the wave function, which contains all the dynamics of the collision, is not described by a finite basis set as in the time-independent close-coupling approaches discussed above. On the other hand, the computational resources needed for such calculations are extensive. Hence, the method is so far effectively restricted to (quasi-)one- and (quasi-)two-electron targets. Also, often only the lowest few partial waves are treated in this way, while contributions from higher partial waves are handled by simpler methods, such as the first-order DWBA approximation mentioned above.

Figure 8.25 shows the development of a wavepacket propagated in the TDCC approach. This particular example is for the $^1S^e$ partial-wave symmetry in $e - Be^+$ collisions at an incident projectile energy of 50 eV. Initially, we have a properly symmetrized product of a Gauss packet for the projectile and the active 2s electron in Be^+, which then develops into a complex object containing contributions from elastic scattering, excitation, and ionization. The latter can be identified by the wavepacket having nonzero magnitude in the region where the radial

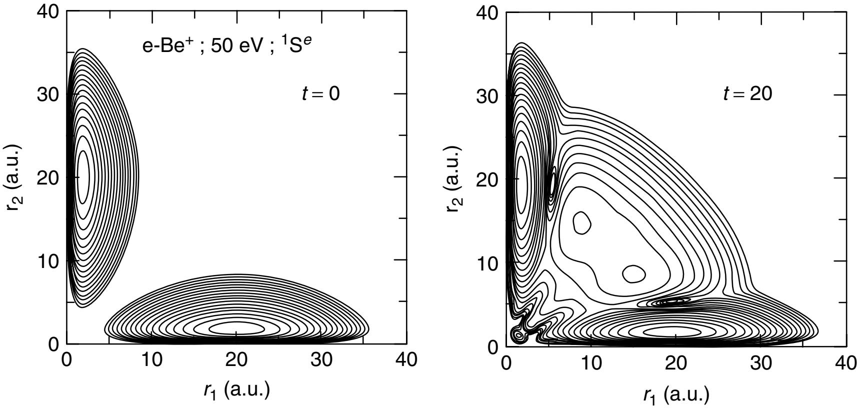

Fig. 8.25 Magnitude of the wave function for e − Be$^+$ collisions in the $^1S^e$ partial wave as a function of the radial coordinates of the projectile and the active electron for an incident projectile energy of 50 eV. After developing for 20 a.u. of time, the initial wavepacket on the left contains contributions from elastic scattering, excitation, and ionization. (Adapted from Pindzola *et al.* (1997)).

coordinates of both electrons (r_1 and r_2) are large.

8.3.3.2 Exterior Complex Scaling (ECS)

The exterior complex scaling method is another attempt to obtain the full wave function more directly than in the standard close-coupling or Born series approaches. The basic idea is to separate the wave function into two parts, namely, an incident part Φ_{in} and a scattered part Ψ_{sc}. The Schrödinger equation is rewritten as

$$(E - \boldsymbol{H})\Psi_{\text{sc}} = (\boldsymbol{H} - E)\Phi_{\text{in}} \tag{8.50}$$

In other words, the incident wave function is the driving term in the equation for the scattering wave function.

The critical point in what follows is the fact that Ψ_{sc} *only* contains outgoing waves. Hence, by making the variable transformation

$$z(r) \to \begin{cases} r & \text{if } < R_0 \\ R_0 + (r - R_0)\, e^{i\theta} & \text{if } r \geq R_0 \end{cases} \tag{8.51}$$

for all radial coordinates beyond some value R_0, any outgoing oscillating wave is transformed to an exponentially decreasing function. In this way, the complicated three-body Coulomb boundary conditions can be avoided in the solution of the driven Schrödinger equation. The physical information can be extracted inside the region defined by R_0 by using a variety of formulas for the scattering amplitude and the cross section. Of course, R_0 has to be chosen sufficiently large to ensure converged and stable results.

The ECS method was applied with great success to the electron-impact ionization of atomic hydrogen by Rescigno *et al.* (1999). As an example, Figure 8.26 shows the scattering wave function on a 2D radial grid for a particular set of angular momentum quantum numbers. A review of other applications was given by McCurdy, Baertschy and Rescigno (2004). The numerical efficiency of the method was greatly enhanced by a propagation algorithm (the PECS approach) described by Bartlett

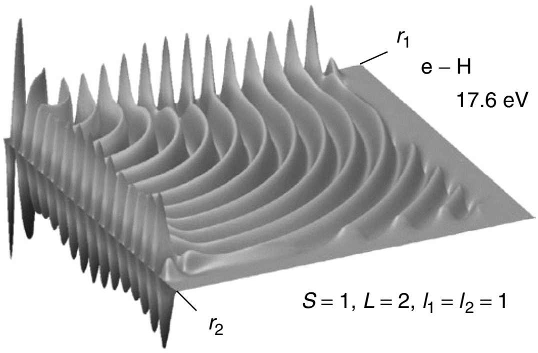

Fig. 8.26 Real part of the wave function for e – H collisions for an incident projectile energy of 17.6 eV in the triplet spin channel for individual angular momenta $l_1 = l_2 = 1$ and a coupled angular momentum $L = 2$. Note how the oscillatory parts of the scattering wave function vanish at large values of the radial coordinates r_1 and r_2. (Adapted from McCurdy, Baertschy and Rescigno (2004)).

(2006). The ECS method has recently been combined with a variety of numerical schemes, including *B*-splines and finite-element discrete-variable representations (FEDVR). Calculations for He, in the so-called "*S*-wave model" with all orbital angular momenta being zero, were reported by Bartlett *et al.* (2007), thus suggesting that the full problem with three active electrons will likely be treated in the near future.

8.3.4
Ionization of Atoms and Ions

In this section, we discuss the process whereby electrons are ejected from the target giving rise to ionization. We are mainly concerned with single ionization or (e,2e) reactions of the type discussed in Section 8.2.4, that is,

$$e_i^- + A_i \longrightarrow A_j^+ + e_A^- + e_B^- \qquad (8.52)$$

where A_i is the initial bound state of the target atom (or ion) and A_j^+ is the final state of the ion.

8.3.4.1 The Ionization Amplitude and Cross Section

The scattering amplitude for the direct ionization process is given by the integral expression Rudge and Seaton (1965)

$$f_i(\boldsymbol{k}_A, \boldsymbol{k}_B) = -(2\pi)^{1/2} \times \exp[i\chi(\boldsymbol{k}_A, \boldsymbol{k}_B)] \times \langle\!\langle (H_{N+1} - E)\Phi|\Psi_i^{(+)}\rangle \qquad (8.53)$$

Here $\Psi_i^{(+)}$ is an exact solution of Eq. (8.26) satisfying plane-wave plus outgoing-wave boundary conditions (8.29) and Φ is a solution of the equation

$$(H_{N+1} - V - E)\,\Phi = 0 \qquad (8.54)$$

satisfying ingoing-wave boundary conditions. Also, $\boldsymbol{k}_A$ and $\boldsymbol{k}_B$ are the momenta of the two outgoing electrons defined in terms of E_A and E_B in Eq. (8.9) by $\frac{1}{2}k_A^2 = E_A$ and $\frac{1}{2}k_B^2 = E_B$.

In practice, V is chosen so that Φ has a simple form. Thus, for electron scattering by H-like ions with nuclear charge Z, V is often chosen as

$$V = -\frac{Z - Z_A}{r_1} - \frac{Z - Z_B}{r_2} + \frac{1}{r_{12}} \qquad (8.55)$$

An approximate final-state wave function is then defined by

$$\Phi(\boldsymbol{r}_1, \boldsymbol{r}_2) = \psi_C^{(-)}(Z_A, \boldsymbol{k}_A, \boldsymbol{r}_1) \times \psi_C^{(-)}(Z_B, \boldsymbol{k}_B, \boldsymbol{r}_2) \qquad (8.56)$$

where $\psi_C^{(-)}$ is a one-electron Coulomb wave function satisfying ingoing-wave boundary conditions while Z_A and Z_B are the effective charges "seen" by the outgoing electrons. In order that the scattering amplitude not contain a divergent phase factor, Z_A and Z_B must satisfy the condition

$$\frac{Z_A}{k_A} + \frac{Z_B}{k_B} = \frac{Z}{k_A} + \frac{Z}{k_B} - \frac{1}{|\boldsymbol{k}_A - \boldsymbol{k}_B|} \tag{8.57}$$

in which case the phase $\chi(\boldsymbol{k}_A, \boldsymbol{k}_B)$ is given by

$$\chi(\boldsymbol{k}_A, \boldsymbol{k}_B) = \frac{Z_A}{k_A} \ln \frac{k_A^2}{k_A^2 + k_B^2} + \frac{Z_B}{k_B} \ln \frac{k_B^2}{k_A^2 + k_B^2} \tag{8.58}$$

The exchange ionization amplitude can be obtained from the theorem

$$g_i(\boldsymbol{k}_A, \boldsymbol{k}_B) = f_i(\boldsymbol{k}_B, \boldsymbol{k}_A) \tag{8.59}$$

first derived by Peterkop (1961). This result follows from the fact that the two amplitudes $f_i(\boldsymbol{k}_B, \boldsymbol{k}_A)$ and $g_i(\boldsymbol{k}_A, \boldsymbol{k}_B)$ describe the same physical process, in which the electron at $\boldsymbol{r}_N$ has momentum $\boldsymbol{k}_B$ and the electron at $\boldsymbol{r}_{N+1}$ has momentum $\boldsymbol{k}_A$. In practice, the Peterkop theorem will not be valid if approximate wave functions are used to calculate the ionization amplitudes. Therefore, a relative phase $\tau_i(\boldsymbol{k}_A, \boldsymbol{k}_B)$ between the two amplitudes is sometimes introduced by

$$g_i(\boldsymbol{k}_A, \boldsymbol{k}_B) = \exp[i\tau_i(\boldsymbol{k}_A, \boldsymbol{k}_B)] f_i(\boldsymbol{k}_B, \boldsymbol{k}_A) \tag{8.60}$$

Note that different choices for $\tau_i(\boldsymbol{k}_A, \boldsymbol{k}_B)$ can lead to significant variations in the predicted cross sections.

An important development in ionization theory was the adaptation of the continuum distorted-wave eikonal-initial-state (CDW-EIS) approach, which was put on a firm mathematical footing for heavy-particle impact by Crothers and McCann (1983), to electron impact. In particular, Brauner, Briggs and Klar (1989) used the so-called 3C wave function, which has the correct asymptotic behavior originally proposed by Redmond (1973), when all three particles are infinitely far apart from each other. In fact, Jones and Madison (2000) showed that the wavefunction is accurate when at least one particle is far away from the other two. The 3C function for a system of three free Coulomb particles (specifically two electrons and a proton) for the final state of an (e,2e) process on atomic hydrogen is given by

$$\begin{aligned} \Phi^-(\boldsymbol{r}_1, \boldsymbol{r}_2) = {} & \\ & (2\pi)^{-3} \exp i \times (\boldsymbol{k}_1 \cdot \boldsymbol{r}_1 + \boldsymbol{k}_2 \cdot \boldsymbol{r}_2) \\ & \times C(\alpha_1, \boldsymbol{k}_1, \boldsymbol{r}_1) \times C(\alpha_2, \boldsymbol{k}_2, \boldsymbol{r}_2) \\ & \times C(\alpha_{12}, \boldsymbol{k}_{12}, \boldsymbol{r}_{12}) \end{aligned} \tag{8.61}$$

Here $\boldsymbol{k}_{12} \equiv \mu(\boldsymbol{k}_1 - \boldsymbol{k}_2)$ is the relative wave vector of the two particles while $\mu = 1/2$ is the reduced mass of the two electrons. Furthermore, the Sommerfeld parameters are given by $\alpha_1 = -1/k_1, \alpha_2 = -1/k_2$, and $\alpha_{12} = \mu/k_{12}$, respectively. Distortion effects due to the Coulomb potential are contained in the function

$$\begin{aligned} C(\alpha, \boldsymbol{k}, \boldsymbol{r}) \equiv {} & \Gamma(1 - i\alpha) \exp(-\pi\alpha/2)_1 \\ & \times F_1(i\alpha, 1; -ikr - i\boldsymbol{k} \cdot \boldsymbol{r}) \end{aligned} \tag{8.62}$$

where ${}_1F_1$ and Γ are the hypergeometric and gamma functions, respectively.

Since 1989, several further modifications to the 3C function have been proposed. These include a suggestion by Alt and Mukhamedzhanov (1993) to use locally defined rather than global wave vectors, Berakdar's "dynamically screened" DS3C function (Berakdar (1996)), the use of the full CDW-EIS function to account for initial-state effects by Jones and Madison (1998), and the development of a DW version by Prideaux, Madison and Bartschat (2005). While improvements have been achieved in some cases, especially regarding the account of the postcollision interaction, the major drawback of the approach has been the fact that 3C-type wave functions are generally not sufficiently accurate at short distances from the target center. Here "short" is typically only a few atomic units (Foster *et al.* (2006); Bartschat, Weflen and Guan (2007)), but that is where the actual ionization process occurs.

Ionization cross sections are obtained by taking the ratio of the number of ionization events per unit time and per unit target atom to the incident electron flux. For random electron spin orientations in targets with one active electron, for example, the triple-differential cross section (TDCS) for ionization from an initial state $|i\rangle$ is given by

$$\frac{d^3\sigma_i}{d\Omega_A d\Omega_B dE} = \frac{k_A k_B}{k_i} \times \left[\frac{1}{4}|f_i + g_i|^2 + \frac{3}{4}|f_i - g_i|^2\right] \tag{8.63}$$

As discussed in Sections 8.2.4.2 and 8.2.4.3, various double and single differential cross sections can be formed by integrating the TDCS with respect to $d\Omega_A$, $d\Omega_B$, and/or dE. By integrating Eq. (8.63) over all outgoing electron scattering angles and energies, we obtain the total ionization cross section

$$\sigma_i = \frac{1}{k_i}\int_0^{E/2} dE\, k_A k_B \int d\Omega_A \int d\Omega_B \times \left[\frac{1}{4}|f_i + g_i|^2 + \frac{3}{4}|f_i - g_i|^2\right] \tag{8.64}$$

Because the two electrons are indistinguishable, the upper limit of integration over the energy variable is $E/2$ to avoid double counting. For atomic targets, the plane-wave Born program of Bartlett and Stelbovics (2003) allows for a fast way of obtaining reasonably accurate total ionization cross sections that are needed in many applications. The "Binary Encounter Bethe" model of Kim and Rudd (1994), Kim *et al.* (2008) has also been highly successful in the production of ionization cross sections for complex atoms and molecules, including radicals, that are important for the modeling of many processes in plasmas.

8.3.4.2 Threshold Behavior of the Ionization Cross Section

In a fundamental paper, Wigner (1948) pointed out that the derivation of threshold laws does not require a detailed knowledge of the collision dynamics in the "reaction zone," where all the particles are close together. Instead, they are determined by the asymptotic form of the wave function as the particles move apart. Wigner used this result to obtain threshold laws for two particles escaping from each other.

Wannier (1953) extended Wigner's theory to single ionization of atoms or ions by electrons. He considered the division of configuration space into three regions. In analogy with Wigner's analysis, it is not necessary to know the detailed behavior of the electrons in the reaction zone ($R < R_0$), but only to assume that the distribution in phase space of the two escaping electrons is uniform (i.e., quasiergodic) as they enter the Coulomb zone ($R_0 < R < R_1$). Wannier then assumed that for large enough R_0 the Coulomb potential varied sufficiently slowly for classical mechanics to be valid in the Coulomb zone even when the total energy $E = E_A + E_B$ tends to zero. Finally, at very large $R > R_1$, the combined kinetic energies of the electrons is larger than the potential energy, and hence the electrons move freely. As $E \to 0$, the free zone recedes and the Coulomb zone extends to infinity. The threshold behavior of the ionization cross section is therefore determined by the motion of the two electrons in the Coulomb zone.

Wannier showed that the threshold ionization cross section for hydrogenlike ions is dominated by the region in the neighborhood of a saddle point in the potential energy surface at $\alpha = \pi/4$ and $\theta_{12} = \pi$, where $\tan\alpha = r_1/r_2$ and θ_{12} is the angle between the two electron vectors $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$. By expanding the potential around this saddle point, he showed that for total angular momentum $L = 0$, the threshold law for the total ionization cross section is given by

$$\sigma_{\text{ion}} \sim E^m \tag{8.65}$$

where

$$m = -\frac{1}{4} + \frac{1}{4}\left(\frac{100Z - 9}{4Z - 1}\right)^{1/2} \tag{8.66}$$

When $Z = 1$, then $m = 1.127$; as $Z \to \infty$, $m \to 1$. A further important result, obtained by Vinkalns and Gailitis (1967), is that the width of the angular distribution in θ_{12} is

$$\Delta\theta_{12} \sim E^{1/4} \tag{8.67}$$

with a maximum at $\theta_{12} = \pi$. Hence the angular distribution becomes more concentrated as $E \to 0$.

Greene and Rau (1982, 1983) extended the Wannier theory to other $LS\pi$ values. They showed that in all states, except for $^3\text{S}^\text{e}$ and $^1\text{P}^\text{e}$, the wave function can remain finite at the saddle point, and hence the Wannier threshold law will pertain. In the $^3\text{S}^\text{e}$ and $^1\text{P}^\text{e}$ states, on the other hand, the symmetry of the wave function forces a node to occur at the saddle point, thereby giving rise to an exponent of $3m$ rather than m in Eq. (8.65).

Figure 8.27 shows the cross section and the spin asymmetry for electron impact ionization of atomic hydrogen in the near-threshold regime. The PECS results of Bartlett (2006) are compared to a variety of other theoretical and experimental results. As seen from the figure, the PECS theory agrees particularly well with the analytical threshold law and the experimental data. This is mostly due to the fact that the wave function can be integrated out in a stable way to very large distances.

8.3.4.3 **Ionization at Intermediate and High Energies**

The integral expression (8.53) for the direct ionization amplitude provides a starting point for the calculation of the cross section when some approximation is made for $\Psi_i^{(+)}$. The simplest approximation is to represent $\Psi_i^{(+)}$ by the incident plane-wave term. This gives a variety of first Born approximations, depending on the choice of phase in Eq. (8.60). However, this approximation can only be expected to be accurate at high energies.

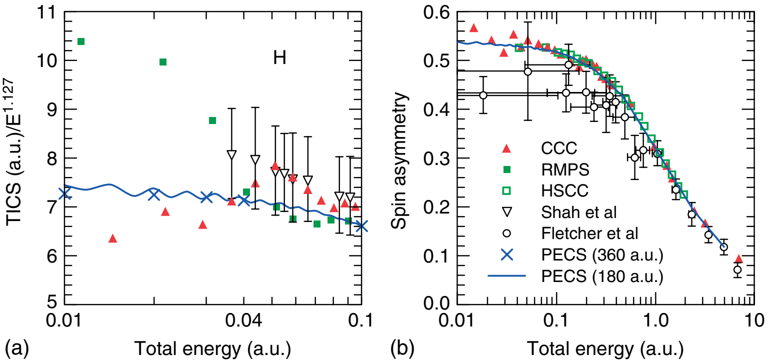

Fig. 8.27 Cross section (a) and spin asymmetry (b) for electron impact ionization of atomic hydrogen in the near-threshold regime. The PECS results of Bartlett (2006) are compared to a variety of other theoretical and experimental results. (Adapted from Bartlett (2006)).

To obtain cross sections that are reliable at lower energies, more sophisticated approximations must be made for $\Psi_i^{(+)}$. For example, $\Psi_i^{(+)}$ in Eq. (8.53) has been represented by coupled eigenstate plus pseudostate expansions of the form of Eq. (8.32). This is the basic idea behind the highly successful application of the CCC method to ionization processes, and it could be used in the RMPS and IERM methods as well. Alternatively, the ionization cross section can be obtained directly from the asymptotic form of $\Psi_i^{(+)}$. This is done, for example, in the TDCC and ECS approaches mentioned above.

Another approximation is a hybrid method introduced by Jakubowicz and Moores (1981) and later implemented into an R-matrix framework by Bartschat and Burke (1987). Starting from Eq. (8.53), the N-electron target in the initial state and the $(N-1)$-electron residual ion plus ejected electron in the final state are both expanded as in Eq. (8.32), whereas the ionizing electron is represented by plane or distorted waves in both the initial and final states. This approximation is appropriate for situations in which the ejected electron moves much slower than the ionizing electron, and it becomes increasingly accurate at all energies for ionization of ions as Z increases. After further development, it has become possible to account, at least approximately, for second-order effects in the Projectile–target interaction (Reid, Bartschat, and Raeker (1998)) and to represent the initial bound-state and the ejected-electron–residual-ion interaction by a convergent R-matrix with pseudostates expansion. As seen in Figure 8.13, the method has been fairly successful in describing simultaneous ionization plus excitation of He and other quasi-two-electron systems. Bartschat and Vorov (2005) also applied it with some success to direct ionization of the outermost p and s electrons in heavy noble gases. However, the most recent experimental data of Stevenson and Lohmann (2008) suggest that further improvements are necessary to obtain good agreement with experiment over the full angular range.

At high energies, methods based on the plane-wave Born series become valid. The TDCS for $e-\text{He}$ in Figure 8.9

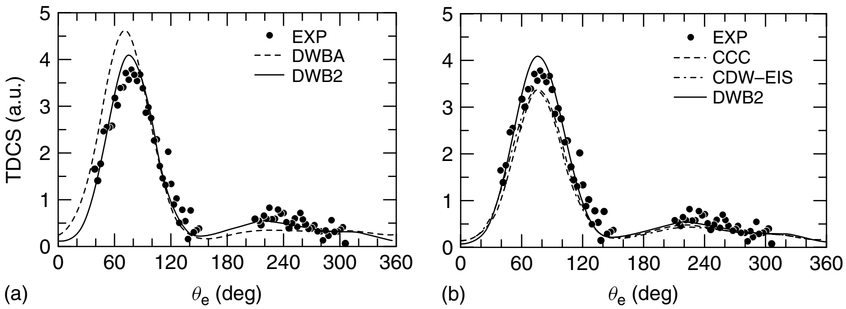

Fig. 8.28 The triple-differential cross section (in atomic units) for atomic hydrogen as a function of the ejected electron angle θ_e for an incident electron energy of 250 eV, an ejected electron energy of 5 eV, and a fixed detection angle of 8° for the fast electron. The dots are experimental data of Ehrhardt *et al.* (1986), while the various theoretical curves represent first-order and second-order B1 and B2 DW Born approximations, CCC and continuum distorted-wave plus eikonal-initial-state (CDW-EIS) model. (Adapted from Chen *et al.* (2004)).

shows that the second Born approximation correctly predicts the change in height and shift of the binary and recoil peaks from that given by the first Born approximation.

Replacing the plane waves by distorted waves allows for the Born-type methods to reach lower energies. This is shown in Figure 8.28 for 250 eV electron-impact ionization of atomic hydrogen.

8.3.4.4 **Excitation–Autoionization**

Excitation–autoionization discussed in Section 8.2.4.7 can be described using the hybrid method described above. In this case, it is necessary that the relevant resonant states are included in the expansion describing the scattering of the ejected electron from the $(N-1)$-electron residual ion. These resonances occur in Rydberg series converging to the excited states of the residual ion. However, this theory is appropriate only if the scattered electron is sufficiently fast that it can be adequately described by a distorted wave. In particular, this theory will fail if the incident electron forms a resonant state with the target ion as in Eq. (8.20).

When the incident electron interacts strongly with the target, it is often possible to obtain an accurate estimate of this cross section by assuming that the two processes defined by Eqs. (8.18) and (8.19) occur independently and can thus be calculated separately. This is equivalent to assuming that the lifetimes of the autoionizing states are large compared with their excitation time, and hence that these states can be treated as if they were bound states lying in the continuum. The total wave function describing the excitation process is then expanded as in Eq. (8.32), where the first summation includes the initial state of the target as well as all important autoionizing states. The ionization cross section is then obtained by solving the coupled integro-differential equations (8.33) either directly or by using a DW approximation and then summing the excitation cross section to all autoionizing states that decay into the continuum. This theory also describes the REDA process, because the intermediate states in Eq. (8.20) appear as resonances in the scattering process defined by Eq. (8.18). This approach has been widely used to obtain excitation–autoionization cross sections for complex ions (Müller, 1990).

8.3.5
Scattering by Molecules

The processes that occur in electron scattering by molecules are more varied than those that occur in electron scattering by atoms and ions because of the possibility of exciting degrees of freedom associated with the motion of the nuclei as well as electronic degrees of freedom. In addition, the multicentered and nonspherical nature of the electron–molecule interaction considerably complicates the solution of the scattering problem by reducing its symmetry.

There are two other distinctive features of electron–molecule scattering. First, there is the crucial role that resonances play in vibrational excitation and dissociative attachment (Hotop *et al.*, 2003). For example, it is mainly near resonances that the scattered electron spends sufficient time in the neighborhood of the molecule to be able to excite the nuclear motion with significant probability. Second, the long-range interaction between an electron and a molecule is more complicated than that between an electron and an atom. Thus, for a linear molecule, the potential has the asymptotic form

$$V(\boldsymbol{r}, \widehat{\boldsymbol{R}}) \xrightarrow[r\to\infty]{} -\frac{\mu}{r^2}P_1(\cos\theta) - \frac{Q}{r^3}P_2 \times (\cos\theta) - \frac{\alpha_0}{2r^4} - \frac{\alpha_2}{2r^4}P_2(\cos\theta) \qquad (8.68)$$

where $\boldsymbol{R}$ is a unit vector along the internuclear axis, $\boldsymbol{r}$ is the coordinate of the scattered electron with respect to the center of gravity of the molecule, and $\cos\theta = \widehat{\boldsymbol{r}} \cdot \widehat{\boldsymbol{R}}$. For polar molecules, the dipole moment μ is nonzero and the corresponding term in Eq. (8.68) dominates the angular distribution in the forward direction and gives rise to bound and virtual states that often lead to threshold peaks in the vibrational excitation cross sections. Furthermore, the long-range nature of the quadrupole (Q) and polarization (α_0 and α_2) terms in the potential lead to enhanced rotational excitation cross sections.

8.3.5.1 Laboratory Frame Representation

We first consider the derivation of the basic equations describing electron–molecule scattering in the laboratory frame of reference. Our discussion closely parallels that given in Section 8.3.1.1 for electron–atom scattering. The Schrödinger equation describing the electron–molecule system is

$$(H_m + K + V)\Psi = E\Psi \qquad (8.69)$$

where H_m is the Hamiltonian for the target molecule, K is the kinetic energy of the scattered electron, and V is the electron–molecule interaction potential

$$V(\boldsymbol{R}, \boldsymbol{r}_m, \boldsymbol{r}) = \sum_j \frac{1}{|\boldsymbol{r} - \boldsymbol{r}_j|} - \sum_j \frac{Z_j}{|\boldsymbol{r} - \boldsymbol{R}_j|} \qquad (8.70)$$

Here $\boldsymbol{R}$ stands symbolically for the positions $\boldsymbol{R}_j$ of all the nuclei, $\boldsymbol{r}_m$ for the set of coordinates $\boldsymbol{r}_j$ of the electrons in the target molecule, and $\boldsymbol{r}$ for the coordinate of the scattered electron. The total energy E refers to the frame of reference where the center of mass of the whole system is at rest.

We again introduce the target eigenstates and possible pseudostates Φ_i by the equation

$$\langle \Phi_i | H_m | \Phi_j \rangle = w_i \delta_{ij} \tag{8.71}$$

and then expand the total wave function Ψ, in analogy with Eq. (8.32) for atoms and ions, in the form

$$\Psi_i = \mathcal{A} \sum_j \Phi_j(\boldsymbol{R}, \boldsymbol{r}_m) \mathcal{F}_{ji}(\boldsymbol{r}) + \sum_j \chi_j(\boldsymbol{R}, \boldsymbol{r}_m, \boldsymbol{r}) \alpha_{ji} \tag{8.72}$$

Here, we have suppressed the spin variables for notational simplicity and have not carried out a partial-wave decomposition of the wave function $\mathcal{F}_{ji}$ representing the scattered electron. The subscripts i and j now represent the ro-vibrational and electronic states of the target molecule.

Coupled equations for the functions $\mathcal{F}_{ji}$ can be obtained by substituting expansion (8.72) into Eq. (8.69) and projecting onto the target states Φ_i and the square-integrable function χ_j. After eliminating the coefficients α_{ji}, we obtain the following coupled integro-differential equations that are satisfied by the functions $\mathcal{F}_{ji}$:

$$(\nabla^2 + k_j^2) \mathcal{F}_{ji}(r) = 2 \sum_l (V_{jl} + W_{jl} + X_{jl}) \mathcal{F}_{li}(\boldsymbol{r}) \tag{8.73}$$

Here

$$k_j^2 = 2(E - w_j) \tag{8.74}$$

and V_{jl}, W_{jl}, and X_{jl} are again the direct, nonlocal exchange, and nonlocal correlation potentials. By expanding the $\mathcal{F}_{ji}$ in partial waves we obtain a set of coupled radial integro-differential equations, analogous to Eqs. (8.33) for atoms and ions, which can, in principle, be solved numerically.

The scattering amplitude and cross section can be obtained from the asymptotic form of the total wave function, which is written in a form analogous to Eq. (8.29). The differential cross section for a transition from an initial state $|\boldsymbol{k}_i, \Phi_i, \chi_m\rangle$ to a final state $|\boldsymbol{k}_j, \Phi_j, \chi_m\rangle$ is then given by Eq. (8.31), except that the scattering amplitude $f_{ji}(\theta, \phi)$ now represents transitions between ro-vibrational states in addition to electronic states.

8.3.5.2 **Molecular Frame Representation**

The theory described in the previous section is completely general and has been the basis of a number of early calculations for light diatomic molecules. However, a major difficulty arises because of the very large number of ro-vibrational and electronic channels that need to be included in expansion (8.72) for all but the simplest low-energy calculations.

This difficulty can, to a large extent, be overcome by adopting an approximation similar to that used in molecular bound-state calculations, where a Born-Oppenheimer separation of the electronic and nuclear motion is made. The electronic motion is first determined with the nuclei held fixed. This is referred to as the *fixed-nuclei approximation*. The molecular rotational and vibrational motion is then included in a second step in the calculation. The entire procedure is called the *adiabatic-nuclei approximation*. It owes its validity to the large ratio of the nuclear mass to the electronic mass

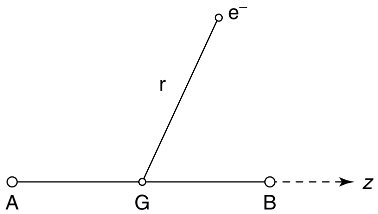

Fig. 8.29 Molecular frame for a diatomic molecule.

and can be adopted in electron–molecule scattering processes when the collision time is much shorter than the period of molecular rotation and/or molecular vibration. The approximation is thus expected to be valid when the scattered electron energy is not close to threshold or when the energy does not coincide with that of a narrow resonance. In these cases, further developments are needed to obtain reliable cross sections.

The fixed-nuclei approximation is not new, having been used in the 1930s to describe the scattering of electrons by homonuclear diatomic molecules by Stier (1932) and Fisk (1936). However, since the early 1970s, it has gained general acceptance as the basis of *ab initio* computational methods that are yielding the most accurate cross sections for complex molecules.

To formulate the collision in this representation, we adopt a frame of reference that is rigidly attached to the molecule, as illustrated in Figure 8.29 for a diatomic molecule. In this figure, A and B are the two nuclei, G is the center of gravity that is taken as the origin of coordinates, and the z-axis is chosen along the internuclear axis.

The fixed-nuclei approximation starts from the Schrödinger equation

$$(H_{el} + K + V)\psi = E\psi \tag{8.75}$$

where H_{el} is the electronic part of the target Hamiltonian defined by assuming that the target nuclei have fixed coordinates denoted by $\boldsymbol{R}$. It follows that H_{el} is related to H_m in Eq. (8.69) by

$$H_m = H_{el} + K_R \tag{8.76}$$

where K_R is the kinetic energy operator for the rotational and vibrational motion of the nuclei. The remaining quantities K and V are the same as in Eq. (8.69).

The solution of Eq. (8.75) proceeds in an analogous way to the solution of Eq. (8.26) for electron–atom or electron–ion scattering. We adopt an expansion similar to Eq. (8.32), where we now expand the function representing the motion of the scattered electron in terms of symmetry-adapted angular functions that transform as an appropriate irreducible representation (IRR) of the molecular point group (Burke, Chandra and Gianturco, 1972). Substituting this expansion into Eq. (8.75) and projecting onto the corresponding channel functions and the square-integrable functions yields the coupled integro-differential equations

$$\left(\frac{d^2}{dr^2} - \frac{l_j(l_j+1)}{r^2} + k_j^2\right) F_{ji}^{\Delta}(r) = 2\sum_l (V_{jl}^{\Delta} + W_{jl}^{\Delta} + X_{jl}^{\Delta}) F_{li}^{\Delta}(r) \tag{8.77}$$

Now the channel indices i, j, and l represent the component of the IRR as well as the electronic state of the target, Δ refers to the conserved quantum numbers that include the IRR and the total spin, l_j is the angular momentum of the scattered electron, and

$$k_j^2 = 2(E - \varepsilon_j) \tag{8.78}$$

Finally, V_{jl}^{Δ}, W_{jl}^{Δ}, and X_{jl}^{Δ} are the direct, nonlocal exchange, and nonlocal correlation potentials that depend parametrically on the internuclear coordinates $\boldsymbol{R}$.

The final step is to solve Eqs. (8.77), for each set of coordinates $\boldsymbol{R}$, subject to $\boldsymbol{K}$-matrix boundary conditions analogous to Eqs. (8.36). The scattering matrix in the molecular frame of reference, and hence the cross section, can then be obtained in a straightforward way.

A number of general computer-program packages have been developed to solve Eqs. (8.77). As for electron–atom or electron–ion scattering, these include packages based on the reduction of these equations to linear algebraic equations, on the Schwinger variational method, on the R-matrix method, and on the complex Kohn method. All of these approaches use standard quantum-chemistry structure packages that have been appropriately modified for scattering. More details about the methods can be found in the book edited by Huo and Gianturco (1995).

For scattering calculations involving only the ground electronic state, a number of approaches have been developed that replace the nonlocal exchange and correlation potentials by local potentials and hence make the resultant equation easier to solve. See, for example, Gianturco and Jain (1986), Gianturco and Lucchese (2004), or Tonzani and Greene (2005).

8.3.5.3 Inclusion of the Nuclear Motion

We now discuss how observables involving the nuclear motion, such as rotational and vibrational excitation cross sections and dissociative attachment cross sections, can be obtained from the solutions of the fixed-nuclei equations.

One of the most widely used approaches is the adiabatic-nuclei approximation proposed by Drozdov (1955, 1956) and by Chase (1956) in studies of neutron scattering by nuclei. As an illustration of this approach, we write the relevant scattering amplitude for a diatomic molecule in a $^1\Sigma$ state as

$$\begin{aligned} &A_{ivjm,i'v'j'm'}(\widehat{\boldsymbol{k}}, \widehat{\boldsymbol{r}}) = \\ &\quad \langle \chi_{iv}(R) Y_{jm}(\widehat{\boldsymbol{R}}) \times |A_{ii'}(\widehat{\boldsymbol{k}} \cdot \widehat{\boldsymbol{r}}; \boldsymbol{R})| \\ &\quad \times \chi_{i'v'}(R) Y_{j'm'}(\widehat{\boldsymbol{R}})\rangle \end{aligned} \tag{8.79}$$

where $A_{ii}''(\widehat{\boldsymbol{k}} \cdot \widehat{\boldsymbol{r}}; \boldsymbol{R})$ is the fixed-nuclei scattering amplitude obtained by solving Eqs. (8.77) and $\chi_{iv}(R)$ and $Y_{jm}(\widehat{\boldsymbol{R}})$ are the molecular vibrational and rotational eigenfunctions. This approximation is valid provided that the collision time is short compared with the vibration and/or rotation time. It has been widely used for nonresonant scattering away from threshold.

The adiabatic-nuclei approximation breaks down close to threshold or in the neighborhood of narrow resonances as discussed by Morrison (1988). A rather straightforward way of including nonadiabatic effects that arise in vibrational excitation is to retain the vibrational terms in the Hamiltonian but still to treat the rotational motion adiabatically. Hence, instead of Eq. (8.75), we solve the equation

$$(H_{\mathrm{el}} + K_{\mathrm{vib}} + K + V)\widetilde{\psi} = E\widetilde{\psi} \tag{8.80}$$

Here K_{vib} is the kinetic energy operator for the nuclear vibrational motion, while the other quantities have the same meaning as in Eq. (8.75). Adopting a frame of reference in which the molecule has fixed spatial orientation, and separating out the angular variables of the scattered electron, we can derive a set of integrodifferential equations coupling the target vibrational as well as electronic eigenstates. This approach has been used with success by several authors, but it is computationally demanding because the number of coupled channels can become very large.

Earlier, we noted that vibrational excitation and dissociative attachment are particularly important in resonance regions. As a result, a number of approaches have been developed for describing these processes based on electron–molecule resonance theories (Herzenberg (1984), Domcke (1991)). The basic idea is to introduce a series of fixed-nuclei resonance states ψ_n^r for a range of values of R either by imposing Siegert (1939) resonance boundary conditions or by introducing Feshbach projection operators. The amplitude for a transition from an initial electronic-vibrational state iv to a final state $i'v'$ is then given by

$$T_{iv,i'v'} = \sum_n \langle \chi_{iv}(R)\zeta_{ni}(R)|G_n^r(R, R')| \times \zeta_{ni'}(R')\chi_{i'v'}(R')\rangle \qquad (8.81)$$

where $\chi_{iv}(R)$ are the vibrational eigenfunctions, $\zeta_{ni}(R)$ are the "entry amplitudes" from the resonance states ψ_n^r into the initial or final electronic states, and $G_n^r(R, R')$ are the Green's functions that describe the propagation in the intermediate resonance states ψ_n^r.

We now consider the frame-transformation theory of Chang and Fano (1972) which has been influential in electron–atom scattering as well as laying the foundations of work on nuclear motion effects in electron–molecule scattering. Following earlier work by Fano (1970), who considered the extension of the quantum defect theory of Seaton (1958, 1983) to molecular photoabsorption, Chang and Fano pointed out that the electron–molecule interaction exhibits qualitatively different physical features when the distance r of the scattered electron from the center of gravity of the molecule lies in different regions.

We illustrate the main features of frame-transformation theory in the upper part of Figure 8.30. We see in this figure that configuration space is partitioned into two main regions labeled A and B. If vibrational motion is not considered, then in region A, where $r \leq a_2$, the molecular frame of reference can be used to describe the scattering process and the electron–target atom complex can be described as in molecular bound-state calculations. However, at larger distances in region B, where $r > a_2$, the coupling of the scattered electron to the molecular axis no longer dominates and the laboratory frame of reference is appropriate. In this case, the rotational quantum number j of the molecule and the orbital and spin quantum numbers (ℓ, s) of the scattered electron label the scattering channels, where the total angular momentum $\boldsymbol{J} = \boldsymbol{j} + \boldsymbol{\ell} + \boldsymbol{s}$ is conserved in the collision. When the vibrational motion of the molecule is being considered, region A must be partitioned into two subregions Aa and Ab. In subregion Aa, where $r \leq a_1$, the scattered electron spends only a short time compared with the vibration time and

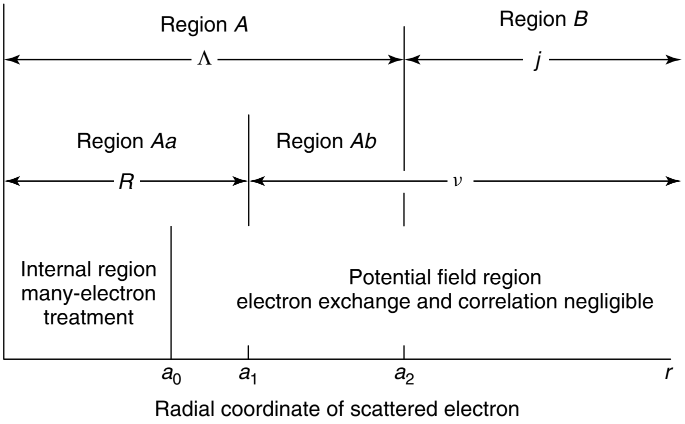
Region A
Λ
Region B
j
Region Aa
R
Region Ab
ν
Internal region many-electron treatment
Potential field region electron exchange and correlation negligible
a_0
a_1
a_2
r
Radial coordinate of scattered electron

Fig. 8.30 The upper part of the figure shows the partitioning of configuration space in frame-transformation theory of electron–molecule scattering with the quantum numbers relevant to different ranges of the electron–molecule distance r. The lower part of the figure shows the partitioning of configuration space adopted in electron–molecule scattering calculations.

hence the Born-Oppenheimer separation of the nuclear and electronic motion can be adopted and the collision wave function can be determined as a function of the internuclear coordinate R. On the other hand, in regions Ab and B, where $r > a_1$, the vibrational motion must be included nonadiabatically and the vibrational channels included explicitly in the expansion of the electron–molecule scattering wave function.

In the lower part of Figure 8.30 we compare the frame-transformation partitioning of configuration space with the partitioning of configuration space adopted in electron–molecule scattering calculations using one of the general computer-program packages discussed in Section 8.3.5.2. In this case, the subdivision is determined by the need to include electron–electron exchange and correlation effects in an internal region $r \leq a_0$ where the scattered electron penetrates the electronic cloud corresponding to the target states of interest. A multicenter configuration-interaction expansion of the total electron–molecule wave function in the molecular frame of reference can be adopted in this region. However, when $r > a_0$, the scattered electron lies outside the target electronic cloud and moves in the long-range local multipole potential of the residual molecule, including molecular polarization terms, which can be conveniently represented by a single-center expansion. If the rotational and vibrational motions of the molecule are being considered, a transformation from the molecular frame of reference to the laboratory frame of reference is applied to the single-center wave function at a_0, a_1, or a_2 as appropriate. In addition, we observe that, although a_0 is shown to be less than a_1 in Figure 8.30, this is not necessarily always the case. In particular, in applications where electron collisions with excited molecular states are considered, it may be necessary for a_0 to be greater than a_1 to

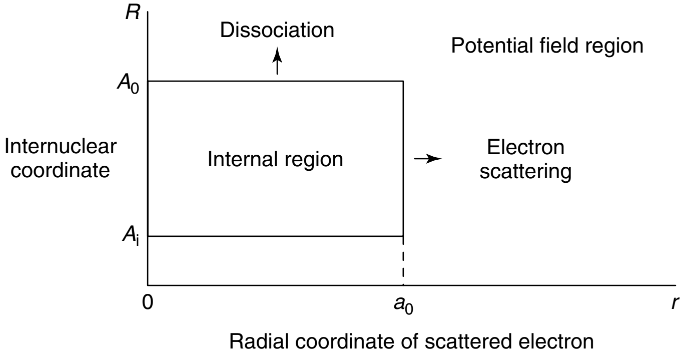

Fig. 8.31 Partitioning of configuration space in electron–molecule scattering calculations allowing for dissociative processes.

include electron exchange and correlation effects arising from the long-range tail of the wave functions corresponding to excited states. The solutions in the internal region of the potential field region can then be determined and matched on the boundary $r = a_0$, using one of the general electron–molecule computer packages discussed in Section 8.3.5.2.

The theory must be further extended to enable dissociative processes as well as rotational, vibrational, and electronic processes to be calculated. In the case of diatomic molecules, we consider the following processes:

$$\begin{array}{llll} & & AB_{i'v'} + \mathrm{e}^- & \text{vibronic excitation} \\ & \nearrow & & \\ \mathrm{e}^- + AB_{iv} & \longrightarrow & A_{i'} + B^-_{i''} & \text{dissociative attachment/recombination} \\ & \searrow & & \\ & & A_{i'} + B_{i''} + \mathrm{e}^- & \text{dissociation} \end{array} \quad (8.82)$$

where i, i' and i'' label the electronic states, v and v' label the vibrational states and where, for notational convenience, we have omitted the rotational quantum numbers. The partitioning of configuration space adopted in electron–molecule scattering calculations allowing for dissociative processes, which generalizes the lower part of Figure 8.30, is shown in Figure 8.31. The internal region is taken to be a rectangle defined by $0 \le r \le a_0$ and $A_i \le R \le A_0$, where a_0 is defined as in Figure 8.30, A_i is chosen to exclude the nuclear Coulomb repulsion singularity at $R = 0$ where the wave function describing the nuclear motion is negligible, and A_0 is chosen so that the target vibrational states of interest have negligible amplitude for $R > A_0$. For $r > a_0$ the collision complex separates into an electron plus a residual molecule, which may be rotationally, vibrationally, and electronically excited. For $R > A_0$ the molecule separates into an atom plus negative ion or into two atoms corresponding to dissociative attachment or dissociation. The solutions in the internal region and the two external regions are then matched on the boundaries between them. In the internal region, the calculation can be carried out in the molecular frame using one of the general computer program packages discussed in Section 8.3.5.2, while the calculations for the external regions can be carried out in the laboratory frame. The corresponding procedure using *R*-matrix theory was introduced by Schneider, Le Dourneuf and Burke (1979); Schneider, Le Dourneuf and Vo Ky Lan (1979) and has been reviewed by Burke and Tennyson (2005).

The extension by Fano (1970) of MQDT to molecular photoabsorption has led to an increasing interest in recent years in its application to the study of electron–molecule scattering and molecular photoionization and dissociation processes. As we saw in Figures. 8.30 and 8.31, configuration space can be partitioned into different regions depending on the distance of the electron from the molecule or the distance apart of the dissociating atoms or ions. In each of these regions the collision process can be analyzed in a frame of reference that reflects the physical characteristics of the region, for example the relative importance of electron–electron exchange and the relative motion of the electron and nuclear components. This enables the calculation and analysis of the process to be carried out in each region using the most appropriate frame of reference and, using MQDT, allows for the *S*-matrix describing the scattering process to be expressed in terms of analytic functions of the energy, thereby enabling it to be continued through thresholds as discussed in Section 8.3.1.5.

Following the application by Fano (1970) of MQDT to the high-resolution photoabsorption spectrum of H_2 near threshold, observed by Herzberg (1969), Jungen and Atabec (1977) developed and applied the method to rovibronic interactions in the photoabsorption spectrum of H_2 and D_2. This work was later extended by Jungen and Dill (1980) to treat rotational and vibrational preionization channels of H_2, obtaining good agreement with photoionization data of Dehmer and Chupka (1976). Further theoretical developments were made by Giusti (1980), Giusti-Suzor and Jungen (1984), Jungen (1984), and Stephens and Greene (1995). A review of molecular applications of MQDT was written by Greene and Jungen (1985). An *ab initio* study of vibrational inelastic processes in H_2 scattering below 5 eV was carried out by Robicheaux (1991), using the frame-transformation method proposed by Greene and Jungen (1985). Also, a noniterative eigenchannel *R*-matrix approach combined with MQDT was developed by Gao, Jungen and Greene (1993). It was applied to predissociation of H_2 in the $3p\pi D\ {}^1\Pi_u^+$ state, and a unified MQDT treatment of both molecular ionization and dissociation was developed by Jungen and Ross (1997).

Finally, we briefly discuss recent theoretical dissociative recombination studies of the triatomic ion H_3^+ by Kokoouline and Greene (2003a, 2003b) and by dos Santos, Kokoouline and Greene (2007). Kokoouline and Greene (2003b, 2005) also considered dissociative recombination of the triatomic ions D_3^+ H_2D^+ and D_2H^+. As mentioned in Section 8.2.6, where complementary experimental studies from the Stockholm (CRYRING) and the Heidelberg (TSR) groups were discussed, dissociative recombination of the H_3^+ ion is a fundamental process in diffuse interstellar clouds. Also, as the simplest triatomic ion, detailed theoretical studies can be seen as a prototype for the study of electron scattering by more complex polyatomic atoms and ions.

In contrast to dissociative attachment/recombination in diatomic molecules and ions given by Eq. (8.82), there is an additional three-body dissociative pathway for H_3^+. Thus we have the following possibilities

$$
\begin{array}{ccc}
 & & H_3^+ + e^- \\
 & \nearrow & \\
e^- + H_3^+ & \longrightarrow & H_2 + H \\
 & \searrow & \\
 & & H + H + H
\end{array}
\tag{8.83}
$$

Here we have omitted the electronic, vibrational, and rotational labels, as well as processes in which H or H_2 are ionized. We see from Eq. (8.83) that dissociative recombination of H_3^+ is inherently a four-body problem. To represent the Jahn–Teller coupling between the electronic and vibrational motion, it is in principle necessary to take into account at least two degrees of freedom of vibrational motion. Finally, the third vibrational coordinate should be introduced to describe the dissociative channel.

The theoretical approach developed by Kokoouline and Greene (2003a, 2003b) combined MQDT to represent the closed channels, the adiabatic hyperspherical coordinate approach to represent the motion of the nuclei, which is discussed in detail by Kokoouline and Masnou-Seeuws (2006), and included outgoing-wave Siegert (1939) pseudostates to represent the vibrational continuum. The Siegert pseudostates, which are analogous to the pseudostates discussed in Section 8.3.1.1, are introduced to let dissociative flux escape if it reaches the hyper-radial boundary. In the later work by dos Santos, Kokoouline and Greene (2007), accurate vibrational wave functions were used and a larger number of possible rotational states of the H_3^+ ground state were included. This resulted in better agreement with the experimental storage-ring data shown in Figure 8.16. It was also shown that the Jahn–Teller coupling between the electronic and vibrational motion plays an important role in this process. In conclusion, this work has shown that recent state-of-the-art *ab initio* calculations on dissociative recombination in simple polyatomic molecules are now capable of accurately describing this complex process.

8.3.5.4 **Illustrative Results for Molecules**

In this section we present a few examples of recent work that illustrate what theory is capable of in this field. We will show a variety of angle-integrated and angle-differential cross sections, for elastic scattering, excitation, and finally ionization.

To begin with, Figure 8.32 shows the angle-differential cross section for vibrationally elastic electron scattering from water molecules at an energy of 4 eV. There is very good agreement between the predictions of Faure, Gorfinkiel and Tennyson (2004) and the experimental data of Cho, Lee and Park (2003). The theoretical results were obtained by the *R*-matrix method within the fixed-nuclei approximation. To compare directly with experiment, they represent the sum over rotational excitations.

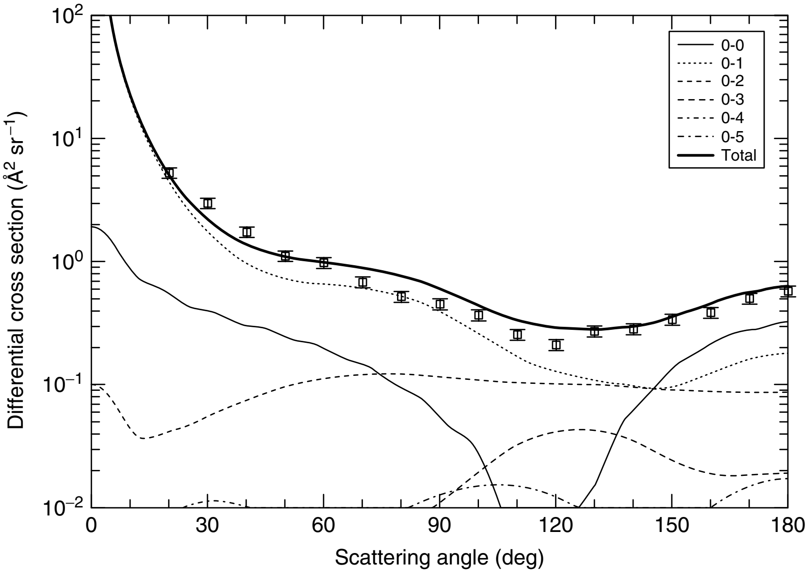

Fig. 8.32 Angle-differential cross section for vibrationally elastic electron scattering from water at an incident electron energy of 4 eV. The thick solid line is the rotationally summed result of Faure, Gorfinkiel and Tennyson (2004) while the squares represent the experimental data of Cho, Lee and Park (2003). Also shown are partial cross sections for individual rotational transitions. (From Faure, Gorfinkiel and Tennyson (2004).)

Next, Figure 8.33 exhibits results for elastic scattering and electron-induced vibrational excitation of NO. For this problem, Trevisan *et al.* (2005) employed a nonlocal treatment of the nuclear dynamics, which significantly improved agreement with experiment compared to results obtained in earlier treatments with a local complex potential. Nevertheless, the agreement between theory and experiment is still far from perfect, indicating the remaining difficulties even for diatomic molecules. An extensive review on the subject was published by Brunger and Buckman (2002).

The situation becomes even more complicated for polyatomic molecules. Most calculations are restricted to elastic scattering. Examples include elastic electron scattering from C_2H_4 by Winstead, McKoy and Bettega (2005) and very large molecules such as uracil treated by Gianturco and Lucchese (2004), by Tonzani and Greene (2005), and by Winstead and McKoy (2006). Examples of recent experimental work on molecules of biological interest include the papers by Allan (2007) on tetrahydrofuran and by Khakoo *et al.* (2008) on methanol amd ethanol.

Figure 8.34 exhibits the challenges associated with electron-impact excitation of electronic states in di-atomic and even more so poly–atomicmolecules. Both experiment and theory are very difficult. For the CO transition, there is qualitative agreement between theory and experiment, whereas the predictions and the measurements are orders of magnitude apart for the $\tilde{B}^1A_1$

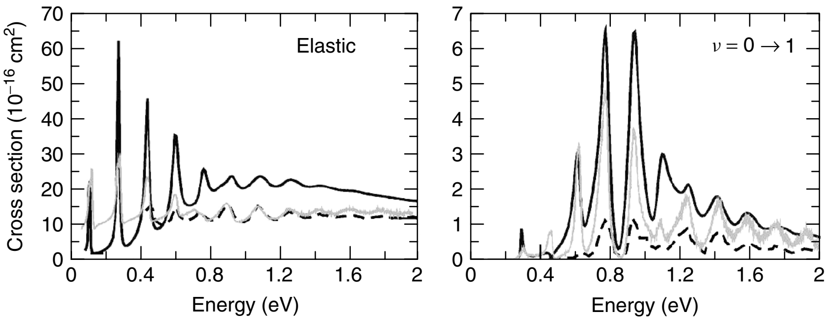

Fig. 8.33 Angle-integrated cross section for elastic scattering and excitation of the $v = 0 \rightarrow 1$ transition in NO. The theoretical predictions of Trevisan *et al.* (2005) (dark solid line) are compared with two sets of experimental data by Jelisavcic, Panajotovic and Buckman (2003) and Allan (2005). (Adapted from Trevisan *et al.* (2005).)

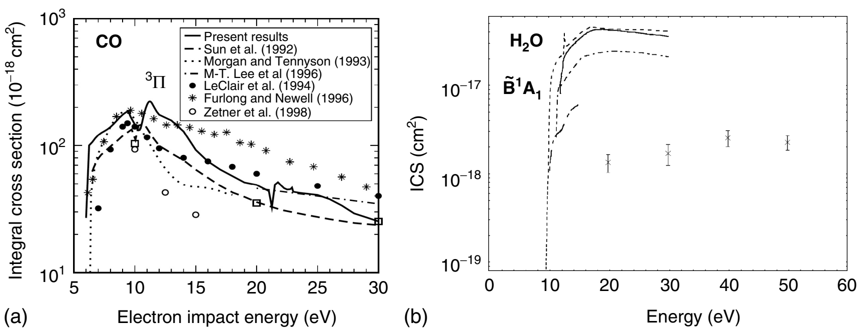

Fig. 8.34 Angle-integrated cross section for electron-impact excitation of the $^3\Pi$ state of CO (a) and the $\tilde{B}^1A_1$ state of water (b). (Adapted from da Costa *et al.* (2007) (CO) and Thorn *et al.* (2007) (H_2O).)

state in water. While there are obvious problems in reconciling the experimental and theoretical magnitudes, Gil *et al.* (1994) achieved impressive qualitative agreement with the (relative) electron energy loss spectra measured at a few fixed angles by Trajmar, Williams and Kuppermann (1973) for water. Further progress was made recently in the fully *ab initio* *R*-matrix calculation on methanol reported by Bouchiha *et al.* (2007). Good agreement with experiment was achieved regarding the total (elastic plus excitation) cross section. Unfortunately, no experimental data are available to test the individual predictions for the excitation processes, including some Feshbach resonances.

For optically allowed transitions, the semiempirical "Binary Encounter f-scaling" method of Kim (2001) often allows for a quick and accurate estimate of the excitation cross section. This is shown for excitation of CO in Figure 8.35.

Figure 8.36 shows the results of Gorfinkiel and Tennyson (2005) for ionization of H_2 by electron impact. These *ab initio* results were obtained using the molecular *R*-matrix with pseudostates

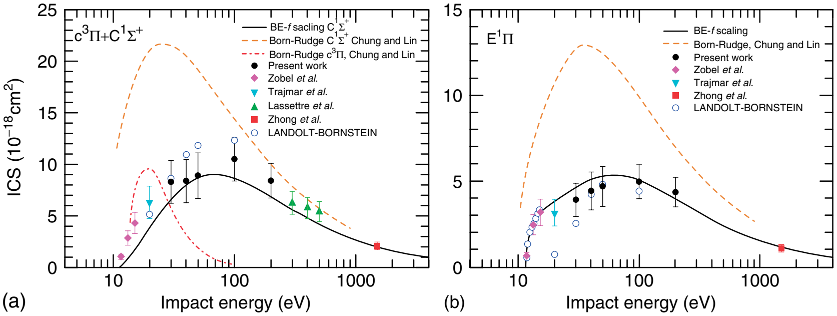

Fig. 8.35 Angle-integrated cross section for electron-impact excitation of the $C^1\Sigma + c^3\Pi$ (a) and the $E^1\Pi$ (b) states of CO. (Adapted from Kawahara *et al.* (2008).)

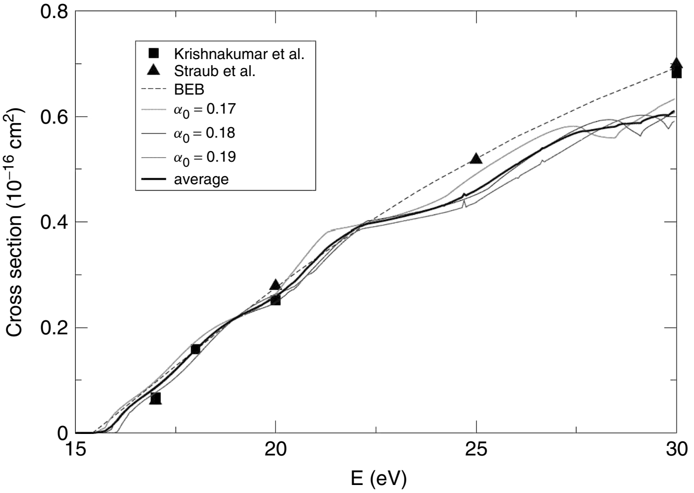

Fig. 8.36 Angle-integrated cross section for electron-impact ionization of H_2. The various values of α_0 refer to the basis used. See Gorfinkiel and Tennyson (2005) for details.

(MRMPS) method. Clearly, the results for the ionization problem are very encouraging. On the other hand, the very same model predicted a resonance structure in the excitation cross section that was not seen in most other calculations, nor in the few available experimental data.

We finish this section with the triple-differential cross section of electron-impact ionization of N_2. For this, and certainly for large molecules, such calculations are extremely difficult due to the complex molecular structure and the need to average over the molecular orientations that are usually not resolved in experiments with molecules in the gas phase. To address this problem, Gao, Madison and Peacher (2005a) suggested the "orientation averaged molecular orbital" (OAMA) approach. The major advantage of the method is the fact that it

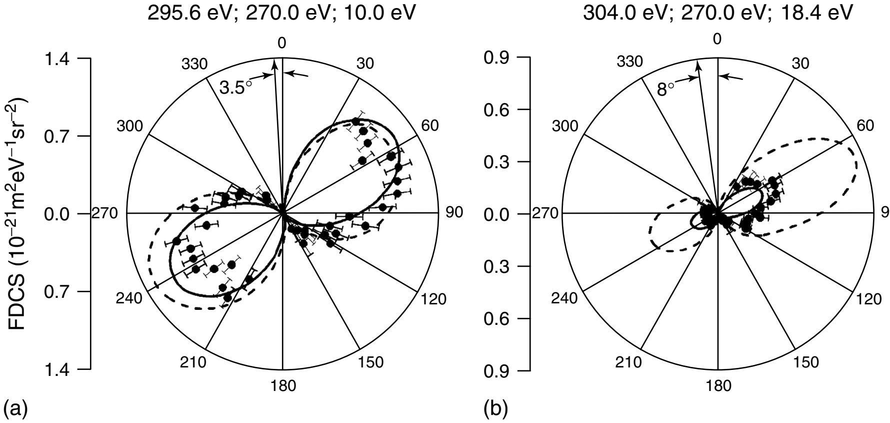

Fig. 8.37 Fully differential cross section for electron-impact ionization of the $3\sigma_g$ state in N_2 for coplanar asymmetric scattering. The incident electron energy is 295.6 eV (a) and 304.0 eV (b). The faster of the two outgoing electrons with energy 270 eV is detected at a scattering angle $\theta_a = 3.5^\circ$ (a) or $\theta_a = 8.0^\circ$ (b), while the detection angle of the slower electron (10.0 eV or 18.4 eV) is varied. The absolute experimental data are from Avaldi *et al.* (1992). The solid and dashed lines represent the M3DW and DWBA results, respectively, see Section 8.3.5.4, normalized to the experiment at the single point at 295.6 eV where $\theta_a = 3.5^\circ$ and $\theta_b = 68^\circ$. (From Gao, Madison and Peacher (2005b).)

is adaptable to the DW theory developed for (e,2e) processes with atoms. However, note that it is restricted to molecular orbitals of appropriate symmetry.

Figure 8.37 shows results for electron-impact ionization of the $3\sigma_g$ state in N_2 for the coplanar asymmetric geometry. For this system, the Rolla group extended the orbital average method further and also included the proper asymptotic three-body Coulomb boundary condition. Comparison between their results obtained in the standard DWBA approach and their molecular 3DW (M3DW) method shows that both calculations reproduce in a satisfactory way the shape of the angular distributions. However, the experimental data of Avaldi *et al.* (1992) are cross-normalized, that is, only a single normalization factor is necessary (and allowed) to put the experimental data on an absolute scale. The figure shows that the M3DW is far superior to the DWBA in reproducing the proper cross normalization. Further applications, for example, to water Kaiser *et al.* (2007) showed some encouraging qualitative agreement between theory and experiment, although discrepancies remained that call for a more sophisticated theoretical treatment.

8.4
Conclusions and Future Work

We have presented an overview of electron scattering by atoms, ions and molecules both from an experimental and from a theoretical standpoint. These areas have seen major advances during the past decade, and our knowledge of these processes is now much more secure than only ten years ago. In particular, advances in experimental methods involving multihit coincidence techniques, high-resolution electron beams, the MAC, spin-polarized projectile and target beams, and the reaction microscope are now providing critical benchmark data against which theory can be tested.

However, much work needs to be done before a complete picture emerges. In the case of electron–atom scattering, there is still an incomplete understanding of collisions involving highly excited states. While a full quantal description of low-energy ionization has been achieved for the atomic hydrogen target and also for helium without excitation, this is not the case for more complex targets or for more correlated processes such as ionization with simultaneous excitation, double ionization, and out-of-plane ionization. Another interesting example is the need to develop a complete quantal theory of excitation—autoionization near threshold. More work is also required to fully understand collisions with heavy atoms and ions, where relativistic effects are important. Examples include fine-structure transitions in collisions of astrophysical importance, such as the $e - Fe^+$ system, and generally atomic and ionic targets with open d-shells. Important examples include e—Mo or e—W collisions for mercury-free alternative lighting concepts or understanding and diagnosing the effects of impurities in fusion devices.

Turning to electron–molecule scattering, we have, in addition to the problems mentioned above for atoms and ions, the need to account fully for the nuclear motion, particularly in the role that it plays during electronic excitation. Experimentally, the resolution of well-defined states will remain a challenge while exciting new information can be expected from studies with "fixed in space" nuclear axes. Also, while new computer-program packages and rapidly increasing computational power have allowed for some *ab initio* calculations on polyatomic molecules, going beyond elastic scattering from biological molecules remains a significant challenge for the future. Some *ab initio* calculations, in the fixed-nuclei approximation, have recently become possible. The inclusion of vibrational excitation in such calculations is the next grand challenge in this field.

It is also worth pointing out that many of the techniques originally developed for steady-state electron collisions with atoms, ions, molecules, and even surfaces are now being used to treat such collisions on ultrashort time scales, involving strong-field laser-atom interactions that lead to excitation, single and multiple ionization. For recent examples, see Guan *et al.*, (2007); Guan, Bartschat and Schneider (2008), Feist *et al.*, (2008), and references therein.

The need for electron–atom, electron–ion, and electron–molecule collision cross sections in many applications provides an important stimulus for experimental and theoretical studies of electron collisions. Important questions involving highly ionized ions in hot plasmas that are difficult to resolve otherwise have already been theoretically explained. However, there are still plenty of unresolved topics, particularly in low-temperature discharges involving neutral atomic and molecular species, where more research is needed before final answers can be given. These questions from researchers in adjacent fields will continue to provide the motivation for work in this field for many years to come.

References

Aflatooni, K., Gallup, G.A., and Burrow, P.D. (1998) *J. Phys. Chem. A*, **102**, 6205.

Allan, M. (2004) *Phys. Scripta*, **T110**, 161.

Allan, M. (2005) *J. Phys. B*, **38**, 603.

Allan, M. (2007) *J. Phys. B*, **40**, 3531.

Alt, E.O. and Mukhamedzhanov, A.M. (1993) *Phys. Rev. A*, **47**, 2004.

Andersen, N. and Bartschat, K. (1996) *Adv. At. Mol. Opt. Phys.*, **36**, 1.

Andersen, N. and Bartschat, K. (2001) *Polarization, Alignment and Orientation in Atomic Collisions*, Springer, New York.

Andersen, N., Bartschat, K., Broad, J.T., and Hertel, I.V. (1997) *Phys. Rep.*, **279**, 251.

Andersen, N., Gallagher, J.W., and Hertel, I.V. (1988) *Phys. Rep.*, **165**, 1.

Andrick, D. (1973) *Adv. Atom. Mol. Phys.*, **9**, 207.

Avaldi, L., Camilloni, R., Fainelli, E., and Stefani, G. (1992) *J. Phys. B*, **25**, 3551.

Bartlett, P.L. (2006) *J. Phys. B*, **39**, R379.

Bartlett, P.L. and Stelbovics, A.T. (2003) *Comp. Phys. Commun.*, **154**, 159.

Bartlett, P.L., Stelbovics, A.T., Lee, G.M., and Bray, I. (2005) *J. Phys. B*, **38**, L95.

Bartlett, P.L., Stelbovics, A.T., Rescigno, T.N., and McCurdy, C.W. (2007) *J. Phys.: Conf. Ser.*, **88**, 012011.

Bartschat, K. (1983) *Comp. Phys. Commun.*, **30**, 383.

Bartschat, K. (1989) *Phys. Rep.*, **180**, 1.

Bartschat, K. (1996) in *Computational Atomic Physics* (ed. K. Bartschat), chaps. 10 and 11, Springer, Heidelberg and New York.

Bartschat, K. and Blum, K. (1982) *Z. Phys. A*, **304**, 85.

Bartschat, K., Blum, K., Hanne, G.F., and Kessler, J. (1981) *J. Phys. B*, **14**, 3761.

(a) Bartschat, K., Bray, I., Fursa, D.V., Stelbovics, A.T. (2007) *Phys. Rev. A*, **76**, 024703; (b) corrigendum (2008) *Phys. Rev. A*, **77**, 029903.

Bartschat, K., Hudson, E.T., Scott, M.P., Burke, P.G., and Burke, V.M. (1996a) *J. Phys. B*, **29**, 2875.

Bartschat, K., Hudson, E.T., Scott, M.P., Burke, P.G., and Burke, V.M. (1996b) *J. Phys. B*, **29**, 115.

Bartschat, K., Bray, I., Burke, P.G., and Scott, M.P. (1996c) *J. Phys. B*, **29**, 5493.

Bartschat, K. and Burke, P.G. (1987) *J. Phys. B*, **20**, 3191.

Bartschat, K., Dasgupta, A., Petrov, G.M., Giuliani, J.L., and Pecharek, R.E. (2004a) *New J. Phys.*, **6**, 145.

Bartschat, K., Zatsarinny, O., Bray, I., Fursa, D.V., and Stelbovics, A.T. (2004b) *J. Phys. B*, **37**, 2617.

Bartschat, K., Dasgupta, A., and Madison, D.H. (2004) *Phys. Rev. A*, **69**, 062706.

Bartschat, K. and Fang, Y. (2000) *Phys. Rev. A*, **62**, 052719.

Bartschat, K. and Madison, D.H. (1987) *J. Phys. B*, **20**, 5839.

Bartschat, K. and Scott, N.S. (1983) *Comp. Phys. Commun.*, **30**, 369.

Bartschat, K., Scott, N.S., Blum, K., and Burke, P.G. (1984) *J. Phys. B*, **17**, 269.

Bartschat, K. and Vorov, V. (2005) *Phys. Rev. A*, **72**, 022728.

Bartschat, K., Weflen, D., and Guan, X. (2007) *J. Phys. B*, **40**, 3231.

Baum, G., Blask, W., Freienstein, P., Frost, L., Hesse, S., Raith, W., Rappolt, P., and Streun, M. (1992) *Phys. Rev. Lett.*, **69**, 3037.

Baum, G., Förster, S., Pavlovic, N., Roth, B., Bartschat, K., and Bray, I. (2004) *Phys. Rev. A*, **70**, 012707.

Baum, G., Granitza, B., Grau, L., Leuer, B., Raith, W., Rott, K., Tondera, M., and Witthuhn, B. (1993) *J. Phys. B*, **26**, 331.

Baum, G., Pavlovic, N., Roth, B., Bartschat, K., Fang, Y., and Bray, I. (2002) *Phys. Rev. A*, **66**, 022705.

Bederson, B. (1969) *Comm. Atom. Mol. Phys.*, **1**,, 41 and 65.

Beijers, J.P.M., Madison, D.H., van Eck, J., and Heideman, H.G.M. (1987) *J. Phys. B*, **20**, 167.

Bellm, S., Lower, J., and Bartschat, K. (2006) *Phys. Rev. Lett.*, **96**, 223201.

Bellm, S., Lower, J., Stegen, Z., Madison, D.H., and Saha, H.P. (2008) *Phys. Rev. A*, **77**, 032722.

Berakdar, J. (1996) *Phys. Rev. A*, **53**, 2314.

Berakdar, J., Lahmam-Bennani, A., and Dal Cappello, C. (2003) *Phys. Rep.*, **374**, 91.

Berger, O. and Kessler, J. (1986) *J. Phys. B*, **19**, 3539.

Blum, K. (1996) *Density Matrix Theory and Applications*, 2nd edn, Plenum, New York.

Bömmels, J., Franz, K., Hoffmann, T.H., Gopalan, A., Zatsarinny, O., Bartschat, K., Ruf, M.-W., and Hotop, H. (2005) *Phys. Rev. A*, **71**, 012704.

Borovik, A., Zatsarinny, O., and Bartschat, K. (2008) *J. Phys. B*, **41**, 035206.

Bouchiha, D., Gorfinkiel, J.D., Caron, L.G., and Sanche, L. (2007) *J. Phys. B*, **40**, 1259.

Boudaiffa, B., Cloutier, P., Hunting, D., Huels, M.A., and Sanche, L. (2000) *Science*, **287**, 1658.

Brauner, M., Briggs, J.S., and Klar, H. (1989) *J. Phys. B*, **22**, 2265.

Bray, I. (1994) *Phys. Rev. A*, **49**, 1066.

Bray, I., Fursa, D.V., Kheifets, A.S., and Stelbovics, A.T. (2002) *J. Phys. B*, **35**, R117.

Bray, I. and Stelbovics, A.T. (1992) *Phys. Rev. Lett.*, **69**, 53.

Bray, I. and Stelbovics, A.T. (1995) *Adv. At. Mol. Phys.*, **35**, 209.

Broad, J.T. (1983) in *Electron-Atom and Electron-Molecule Collisions* (ed. J. Hinze), Plenum, New York, p. 91.

Brunger, M.J. and Buckman, S.J. (2002) *Phys. Rep.*, **357**, 215.

Brunger, M.J., Thorn, P.A., Campbell, L., Diakomichalis, N., Kato, H., Kawahara, H., Hoshino, M., Tanaka, H., and Kim, Y.-K. (2008) *Int. J. Mass Spectrom.*, **271**, 80.

Buckman, S.J., Hammond, P., Read, F.H., and King, G.C. (1983) *J. Phys. B*, **16**, 4219.

Burgess, A. (1964) *Astrophys. J.*, **139**, 776.

Burgess, A. (1965) *Ann. d'Astrophys.*, **28**, 774.

Burke, P.G. (1968) *Adv. Atom. Mol. Phys.*, **4**, 173.

Burke, P.G. (1985) in *Fundamental Processes in Atomic Collision Physics* (eds H. Kleinpoppen, J.S. Briggs, and H.O. Lutz), Plenum, New York, p. 51.

Burke, P.G. and Berrington, K.A. (1993) *Atomic and Molecular Processes: An R-Matrix Approach*, Institute of Physics Publishing, Bristol.

Burke, P.G., Chandra, N., and Gianturco, F.A. (1972) *J. Phys. B*, **5**, 2212.

Burke, P.G. and Crowe, A. (1993) in *Encyclopedia of Applied Physics*, Vol. 5, VCH Publishers, Weinheim and New York, p. 499.

Burke, P.G. and Eissner, W.B. (1983) in *Atoms in Astrophysics* (eds P.G. Burke, W.B. Eissner, D.G. Hummer, and I.C. Percival), Plenum, New York, p. 1.

Burke, P.G., Noble, C.J., and Burke, V.M. (2006) *Adv. At. Mol. Opt. Phys.*, **54**, 237.

Burke, P.G., Noble, C.J., and Scott, M.P. (1987) *Proc. Roy. Soc. A*, **410**, 289.

Burke, P.G. and Schey, H.M. (1962) *Phys. Rev.*, **126**, 147.

Burke, P.G. and Seaton, M.J. (1971) *Meth. Comp. Phys.*, **10**, 1.

Burke, P.G. and Tennyson, J. (2005) *Mol. Phys.*, **103**, 2537.

Burke, P.G. and Webb, T.G. (1970) *J. Phys. B*, **3**, L131.

Callaway, J. and Wooten, J.W. (1973) *Phys. Lett. A*, **45**, 85.

Cartwright, D.C., Csanak, G., Trajmar, S., and Register, D.F. (1992) *Phys. Rev. A*, **45**, 1602.

Catoire, F., Staicu-Casagrande, E.M., Lahmam-Bennani, A., Duguet, A., Naja, A., and Ren, X.G. (2007) *Rev. Sci. Instrum.*, **78**, 013108.

Chang, E.S. and Fano, U. (1972) *Phys. Rev. A*, **6**, 173.

Chase, D.M. (1956) *Phys. Rev.*, **104**, 838.

Chen, Z., Madison, D.H., Whelan, C.T., and Walters, H.R.J. (2004) *J. Phys. B*, **37**, 981.

Childers, J.G., James, K.E., Bray, I., Baertschy, M., and Khakoo, M.A. *Jr.* (2004) *Phys. Rev. A*, **69**, 022709.

Cho, H., McEachran, R.P., Buckman, S.J., Filipović, D.M., Pejčev, V., Marinković, B.P., Tanaka, H., Stauffer, A.D., and Jung, E.C. (2006) *J. Phys. B*, **39**, 3781.

Cho, H., Lee, H.S., and Park, Y.S. (2003) *Radiat. Phys. Chem.*, **68**, 115.

Colgan, J. and Pindzola, M.S. (2006) *Phys. Rev. A*, **74**, 012713.

Crothers, D.S.F. and McCann, J.F. (1983) *J. Phys. B*, **16**, 3229.

Damburg, R. and Karule, E. (1967) *Proc. Phys. Soc.*, **90**, 637.

da Costa, R.F., Bettega, M.H.F., Ferreira, L.G., and Lima, M.A.P. (2007) *J. Phys.: Conf. Ser.*, **88**, 012028.

Dasgupta, A., Bartschat, K., Vaid, D., Grum-Grzhimailo, A.N., Madison, D.H., Blaha, M., and Giuliani, J.L. (2001) *Phys. Rev. A*, **64**, 052710.

Dasgupta, A., Blaha, M., and Giuliani, J.L. (2000) *Phys. Rev. A*, **61**, 012703.

deHarak, B.A., Bartschat, K., and Martin, N.L.S. (2008) *Phys. Rev. Lett.*, **100**, 063201.

Dehmer, P.M. and Chupka, W.A. (1976) *J. Chem. Phys.*, **65**, 2243.

Denifl, S., Ptasińska, S., Hanel, G., Gstir, B., Scheier, P., Probst, M., Farizon, B., Farizon, M., Matejcik, S., Illenberger, E., and Märk, T.D. (2004) *Phys. Scr.*, **T110**, 252.

Dogan, M. and Crowe, A. (2000) *J. Phys. B*, **33**, L461.

Dogan, M. and Crowe, A. (2002) *J. Phys. B*, **35**, 2773.

Domcke, W. (1991) *Phys. Rep.*, **208**, 97.

dos Santos, S.F., Kokoouline, V., and Greene, C.H. (2007) *J. Chem. Phys.*, **127**, 124309.

Drozdov, S.I. (1955) *Sov. Phys. JETP*, **1**, 591.

Drozdov, S.I. (1956) *Sov. Phys. JETP*, **3**, 759.

Dürr, M., Dimopoulou, C., Najjari, B., Dorn, A., Bartschat, K., Bray, I., Fursa, D.V., Chen, Z., Madison, D.H., and Ullrich, J. (2008) *Phys. Rev. A*, **77**, 032717.

Dürr, M., Dimopoulou, C., Najjari, B., Dorn, A., and Ullrich, J. (2006) *Phys. Rev. Lett.*, **96**, 243202.

Dürr, M., Dorn, A., Ullrich, J., Cao, S.P., Czasch, A., Kheifets, A.S., Götz, J.R., and Briggs, J.S. (2007) *Phys. Rev. Lett.*, **98**, 193201.

Ehrhardt, H., Jung, K., Knoth, G., and Schlemmer, P. (1986) *Z. Phys. D*, **1**, 3.

Ehrhardt, H., Schulz, M., Tekaat, T., and Willmann, K. (1969) *Phys. Rev. Lett.*, **22**, 89.

Elazzouzi, S., Dal Cappello, C., Lahmam-Bennani, A., and Catoire, F. (2005) *J. Phys. B*, **38**, 1391.

Eminyan, M. and Lampel, G. (1980) *Phys. Rev. Lett.*, **45**, 1171.

Eminyan, M., MacAdam, K.B., Slevin, J., and Kleinpoppen, H. (1974) *J. Phys. B*, **7**, 1519.

Fano, U. (1961) *Phys. Rev.*, **124**, 1866.

Fano, U. (1970) *Phys. Rev. A*, **2**, 353.

Fano, U. (1983) *Rep. Prog. Phys.*, **46**, 97.

Faure, A., Gorfinkiel, J.D., and Tennyson, J. (2004) *J. Phys. B*, **37**, 801.

Feist, J., Nagele, S., Pazourek, R., Persson, E., Schneider, B.I., Collins, L.A., and Burgdörfer, J. (2008) *Phys. Rev. A*, **77**, 043420.

Fisk, J.B. (1936) *Phys. Rev.*, **49**, 167.

Foster, M., Peacher, J.L., Schulz, M., Madison, D.H., Chen, Z., and Walters, H.R.J. (2006) *Phys. Rev. Lett.*, **97**, 093202.

Franck, J. and Hertz, G. (1914) *Verh. Dtsch. Phys. Ges.*, **16**, 457.

Fursa, D.V. and Bray, I. (1995) *Phys. Rev. A*, **52**, 1279.

Fursa, D.V. and Bray, I. (1997) *J. Phys. B*, **30**, 757.

Fursa, D.V. and Bray, I. (2008) *Phys. Rev. Lett.*, **100**, 113201.

Gailitis, M. (1963) *Sov. Phys. JETP*, **17**, 1328.

Gao, H., Jungen, Ch., and Greene, C.H. (1993) *Phys. Rev. A*, **47**, 4877.

Gao, J., Madison, D.H., and Peacher, J.L. (2005a) *Phys. Rev. A*, **72**,, 020701(R).

Gao, J., Madison, D.H., and Peacher, J.L. (2005b) *J. Chem. Phys.*, **123**, 204314.

Gay, T.J. (1983) *J. Phys. B.*, **16**, L553.

Gianturco, F.A. and Jain, A. (1986) *Phys. Rep.*, **143**, 347.

Gianturco, F.A. and Lucchese, R.R. (2004) *J. Chem. Phys.*, **120**, 7446.

Gianturco, F.A. and Scialla, S. (1987) *J. Phys. B*, **20**, 3171.

Gil, T.J., Rescigno, T.N., McCurdy, C.W., and Lengsfield, B.H. III (1994) *Phys. Rev. A*, **49**, 2642.

Giusti, A. (1980) *J. Phys. B*, **13**, 3867.

Giusti-Suzor, A. and Jungen, Ch. (1984) *J. Chem. Phys.*, **80**, 986.

Glauber, R.J. (1959) in *Lectures in Theoretical Physics* (ed. W.E. Britten), Interscience, New York, p. 315.

Goeke, J., Kessler, J., and Hanne, G.F. (1987) *Phys. Rev. Lett.*, **59**, 1413.

Gorczyca, T.W. and Badnell, N.R. (1997) *J. Phys. B*, **30**, 3897.

Gorfinkiel, J.D. and Tennyson, J. (2005) *J. Phys. B*, **38**, 1607.

Gradziel, M.L. and O'Neill, R.W. (2004) *J. Phys. B*, **37**, 1893.

Green, A.S., Gallup, G.A., Rosenberry, M.A., and Gay, T.J. (2004) *Phys. Rev. Lett.*, **92**, 093201.

Greene, C.H. and Jungen, Ch. (1985) *Adv. At. Mol. Phys.*, **21**, 51.

Greene, C.H. and Rau, A.R.P. (1982) *Phys. Rev. Lett.*, **48**, 533.

Greene, C.H. and Rau, A.R.P. (1983) *J. Phys. B*, **16**, 99.

Grum-Grzhimailo, A.N. (2003) *Comp. Phys. Commun.*, **152**, 101.

Guan, X., Zatsarinny, O., Bartschat, K., Schneider, B.I., Feist, J., and Noble, C.J. (2007) *Phys. Rev. A*, **76**, 053411.

Guan, X., Bartschat, K., and Schneider, B.I. (2008) *Phys. Rev. A*, **77**, 043421.

Hall, R.I., Joyez, G., Mazeau, Y., Reinhard, J., and Schermann, C.S. (1973) *J. Physique*, **34**, 827.

Hanne, G.F. (1976) *J. Phys. B*, **9**, 805.

Hanne, G.F. (1983) *Phys. Rep.*, **95**, 95.

Hayes, P.A. and Williams, J.F. (1996) *Phys. Rev. Lett.*, **77**, 3098.

Heddle, D.W.O. and Gallagher, J.W. (1989) *Rev. Mod. Phys.*, **61**, 221.

Herzberg, G. (1969) *Phys. Rev. Lett.*, **23**, 1081.

Herzenberg, A. (1984) in *Electron Molecule Collisions* (eds I. Shimamura and K. Takayanagi), Plenum, New York, p. 351.

Hotop, H., Ruf, M.-W., Allan, M., and Fabrikant, I.I. (2003) *Adv. At. Mol. Opt. Phys.*, **49**, 85.

Huo, W.M. and Gianturco, F.A. (eds) (1995) *Computational Methods for Electron-Molecule Collisions*, Plenum, New York.

Hussey, M., Murray, A., MacGillivray, W., and King, G. (2008) *J. Phys. B*, **41**, 055202.

Jaduszliwer, B. and Chan, Y.C. (1992) *Phys. Rev. A*, **45**, 197.

Jakubowicz, H. and Moores, D.L. (1981) *J. Phys. B*, **14**, 3733.

Jelisavcic, M., Panajotovic, R., and Buckman, S.J. (2003) *Phys. Rev. Lett.*, **90**, 203201.

Joachain, C.J. (1987) in *Collision Theory for Atoms and Molecules* (ed. F.A. Gianturco), Plenum, New York, p. 59.

Jones, S., Madison, D.H., and Hanne, G.F. (1994) *Phys. Rev. Lett.*, **72**, 2554.

Jones, S. and Madison, D.H. (1998) *Phys. Rev. Lett.*, **81**, 2886.

Jones, S. and Madison, D.H. (2000) *Phys. Rev. A*, **62**, 042701.

Jungen, Ch. (1984) *Phys. Rev. Lett.*, **53**, 2394.

Jungen, Ch. and Atabec, O. (1977) *J. Chem. Phys.*, **66**, 5584.

Jungen, Ch. and Dill, D. (1980) *J. Chem. Phys.*, **73**, 3338.

Jungen, Ch. and Ross, S.C. (1997) *Phys. Rev. A*, **55**, R2503.

Kaiser, C., Spieker, D., Gao, J., Hussey, M., Murray, A., and Madison, D.H. (2007) *J. Phys. B*, **40**, 2563.

Kauppila, W.E. and Stein, T.S. (2001) unpublished; cited in MacAskill *et al.* (2002).

Kawahara, H., Kato, H., Hoshino, M., Tanaka, H., and Brunger, M.J. (2008) *Phys. Rev. A*, **77**, 012713.

Kessler, J. (1985) *Polarized Electrons*, 2nd edn., Springer, Berlin.

Khakoo, M.A., Becker, K., Forand, J.L., and McConkey, J.W. (1986) *J. Phys. B*, **19**, L209.

Khakoo, M.A., Keane, K., Campbell, C., Guzman, N., and Hazlett, K. (2007) *J. Phys. B*, **40**, 3601.

Khakoo, M.A., Blumer, J., Keane, K., Campbell, C., Silva, H., Lopes, M.C.A., Winstead, C., McKoy, V., da Costa, R.F., Ferreira, L.G., Lima, M.A.P., and Bettega, M.H.F. (2008) *Phys. Rev. A*, **77**, 042705.

Kim, Y.K. (2001) *Phys. Rev. A*, **64**, 032713.

Kim, Y.K. and Rudd, M.E. (1994) *Phys. Rev. A*, **50**, 3954.

Kim, Y.K., Irikura, K.K., Rudd, M.E., Ali, M.A., Stone, P.M., Coursey, J.S., Dragoset, R.A., Kishore, A.R., Olsen, K.J., Sansonetti, A.M., Wiersma, G.G., Zucker, D.S., and Zucker, M.A. (2008) http://physics.nist.gov/PhysRefData/Ionization/Xsection.html.

Kokoouline, V. and Greene, C.H. (2003a) *Phys. Rev. Lett.*, **90**, 133201.

Kokoouline, V. and Greene, C.H. (2003b) *Phys. Rev. A*, **68**, 012703.

Kokoouline, V. and Greene, C.H. (2005) *Phys. Rev. A*, **72**, 022712.

Kokoouline, V. and Masnou-Seeuws, F. (2006) *Phys. Rev. A*, **73**, 012702.

Lahmam-Bennani, A. (1991) *J. Phys. B*, **24**, 2401.

Lahmam-Bennani, A., Duguet, A., Dal Cappello, C., Nebdi, H., and Piraux, B. (2003) *Phys. Rev. A*, **67**, 010701.

Lahmam-Bennani, A., Dupré, C., and Duguet, A. (1989) *Phys. Rev. Lett.*, **63**, 1582.

Larsson, M. (2001) *Gas Phase Ion Chem.*, **4**, 179.

Larsson, M. and Thomas, R. (2001) *Phys. Chem. Chem. Phys.*, **3**, 4471.

Lenard, P. (1903) *Ann. Phys.*, **12**, 714.

Lindsay, B.G. and Mangan, M.A. (2003) in *Landolt-Börnstein I-17C* (ed. Y. Itikawa), Springer, Berlin.

Lorentz, S.R., Scholten, R.E., McClelland, J.J., Kelley, M.H., and Celotta, R.J. (1993) *Phys. Rev. A*, **47**, 3000.

Lower, J., Panatović, R., Bellm, S., and Weigold, E. (2007) *Rev. Sci. Instrum.*, **78**, 111301.

Lukomski, M., Sutton, S., Kedzierski, W., Reddish, T.J., Bartschat, K., Bartlett, P.L., Bray, I., Stelbovics, A.T., and McConkey, J.W. (2006) *Phys. Rev. A*, **74**, 032708.

MacAskill, J.A., Kedzierski, W., McConkey, J.W., Domyslawska, J., and Bray, I. (2002) *J. Elect. Spect. Rel. Phen.*, **123**,(2002), 173.

Machado, L., Leal, E.P., and Csanak, G. (1982) *J. Phys. B*, **15**, 1773.

Madison, D.H. and Bartschat, K. (1996) in *Computational Atomic Physics* (ed. K. Bartschat), chap. 4, Springer, Heidelberg and New York.

Madison, D.H. and Shelton, W.N. (1973) *Phys. Rev. A*, **7**, 499.

Maslov, M., Brunger, M.J., Teubner, P.J.O., Zatsarinny, O., Bartschat, K., Fursa, D.V., Bray, I., and McEachran, R.P. (2008) *Phys. Rev. A*, **77**, 062711.

Mayer, S. and Kessler, J. (1995) *Phys. Rev. Lett.*, **74**, 4803.

McAdams, R., Hollywood, M.T., Crowe, A., and Williams, J.F. (1980) *J. Phys. B*, **13**, 3691.

McCall, B.J., Huneycutt, A.J., Saykally, R.J., Geballe, T.R., Djuric, N., Dunn, G.H., Semaniak, J., Novotny, O., Al-Khalili, A., Ehlerding, A., Hellberg, F., Kalhori, S., Neau, A., Thomas, R., Österdahl, F., and Larsson, M. (2003) *Nature*, **422**, 500.

McClelland, J.J., Kelley, M.H., and Celotta, R.J. (1989) *Phys. Rev. A*, **40**, 2321.

McCurdy, C.W., Baertschy, M., and Rescigno, T.N. (2004) *J. Phys. B*, **37**, R137.

McDonald, D.G. and Crowe, A. (1993) *J. Phys. B*, **26**, 2887.

Milne-Brownlie, D.S., Cavanagh, S.J., Lohmann, B., Champion, C., Hervieux, P.A., and Hanssen, J. (2004) *Phys. Rev. A*, **69**, 032701.

Morrison, M.A. (1988) *Adv. At. Mol. Phys.*, **24**, 51.

Müller, A. (1990) in *The Physics of Electronic and Atomic Collisions* (eds A. Dalgarno, R.S. Freund, P.M. Koch, M.S. Lubell, and T.B. Lucatorto), (AIP Conf. Proc. 205), p. 418.

Naja, A., Staicu-Casagrande, E.M., Ren, X.G., Catoire, F., Lahmam-Bennani, A., Dal Cappello, C., and Whelan, C.T. (2007) *J. Phys. B*, **40**, 2871.

Nesbet, R.K. (1979) *Phys. Rev. A*, **20**, 58.

Nic Chormaic, S., Chwirot, S., and Slevin, J. (1993) *J. Phys. B*, **26**, 139.

Ochkur, V.I. (1964) *Sov. Phys. JETP*, **18**, 503.

O'Neill, R.W., van der Burgt, P.J.M., Dziczek, D., Bowe, P., Chwirot, S., and Slevin, J.A. (1998) *Phys. Rev. Lett.*, **80**, 1630.

Peart, B. and Dolder, K.T. (1968) *J. Phys. B*, **1**, 872.

Peterkop, R.K. (1961) *Proc. Phys. Soc.*, **77**, 1220.

Petrov, G.M., Giuliani, J.L., Dasgupta, A., Bartschat, K., and Pecharek, R.E. (2004) *J. Appl. Phys.*, **95**, 5284.

Phaneuf, R.A., Dunn, G.H., Havener, C.C., and Müller, A. (1999) *Rep. Prog. Phys.*, **62**, 1143.

Pindzola, M.S., Griffin, D.C., and Badnell, N.R. (2006) in *Springer Handbook of Atomic, Molecular, and Optical Physics* (ed. G. W. F. Drake), chap. 55, Springer, Heidelberg.

Pindzola, M.S., Robicheaux, F., Badnell, N.R., and Gorczyca, T.W. (1997) *Phys. Rev. A*, **56**, 1994.

Pindzola, M.S., Robicheaux, F., Loch, S.D., Berengut, J.C., Topcu, T., Colgan, J., Foster, M., Griffin, D.C., Ballance, C.P., Schultz, D.R., Tinami, T., Badnell, N.R., Witthoeft, M.C., Plante, D.R., Mitnik, D.M., Ludlow, J.A., and Kleiman, U. (2007) *J. Phys. B*, **40**, R39.

Pravica, L., Cvejanovic, D., Williams, J.F., and Napier, S.A. (2007) *J. Phys.: Conf. Ser.*, **88**, 012067.

Prideaux, A., Madison, D.H., and Bartschat, K. (2005) *Phys. Rev. A*, **72**, 032702.

Ptasińska, S., Denifl, S., Grill, V., Märk, T.D., Scheier, P., Gohlke, S., Huels, M.A., and Illenberger, E. (2005) *Angew. Chem. Int. Ed.*, **44**, 1647.

Ramsbottom, C.A., Noble, C.J., Burke, V.M., Scott, M.P., and Burke, P.G. (2004) *J. Phys. B*, **37**, 3609.

Ramsbottom, C.A., Hudson, C.E., Norrington, P.H., and Scott, M.P. (2007) *Astron. Astrophys.*, **475**, 765.

Rapp, D. and Englander-Golden, P. (1965) *J. Chem. Phys.*, **43**, 1464.

(a) Redmond, P.J. (1973) unpublished; (b) quoted in Rosenberg, L. (1973) *Phys. Rev. D*, **8**, 1833.

Read, F.H. and Channing, J.M. (1996) *Rev. Sci. Instrum.*, **67**, 2372.

(a) Reid, R.H.G., Bartschat, K., and Raeker, A. (1998) *J. Phys. B*, **31**, 563; (b) corrigendum (2000) *J. Phys. B*, **33**, 5261.

Reinhardt, W.P. (1979) *Comp. Phys. Comm.*, **17**, 1.

Rescigno, T.N., Baertschy, M., Isaacs, W.A., and McCurdy, C.W. (1999) *Science*, **286**, 2474.

Robicheaux, F. (1991) *Phys. Rev. A*, **43**, 5946.

Röder, J., Baertschy, M., and Bray, I. (2003) *Phys. Rev. A*, **67**, 010702.

Rudge, M.R.H. and Seaton, M.J. (1965) *Proc. Roy. Soc. A*, **283**, 262.

Sakhelashvili, G., Dorn, A., Hörr, C., Ullrich, J., Kheifets, A.S., Lower, J., and Bartschat, K. (2005) *Phys. Rev. Lett.*, **95**, 033201.

Sanche, L. (2005) *Eur. Phys. J. D*, **35**, 367.

Schappe, R.S., Feng, P., Anderson, L.W., Lin, C.C., and Walker, T. (1995) *Europhys. Lett.*, **29**, 439.

Schneider, B.I., Le Dourneuf, M., and Burke, P.G. (1979) *J. Phys. B*, **12**, L365.

Schneider, B.I., Le Dourneuf, M., and Lan, Vo Ky (1979) *Phys. Rev. Lett.*, **43**, 1926.

Schow, E., Hazlett, K., Childers, J.G., Medina, C., Vitug, G., Bray, I., Fursa, D.V., and Khakoo, M.A. (2005) *Phys. Rev. A*, **72**, 062717.

Schulz, G.J. (1973a) *Rev. Mod. Phys.*, **45**, 378.

Schulz, G.J. (1973b) *Rev. Mod. Phys.*, **45**, 423.

Schwienhorst, R., Raeker, A., Bartschat, K., and Blum, K. (1996) *J. Phys. B*, **29**, 2305.

Seaton, M.J. (1958) *Mon. Not. R. Astron. Soc.*, **118**, 504.

Seaton, M.J. (1983) *Rep. Prog. Phys.*, **46**, 167.

Sewell, E.C. and Crowe, A. (1982) *J. Phys. B*, **15**, L357.

Shore, B.W. (1967) *Rev. Mod. Phys.*, **39**, 439.

Siegert, A.J.F. (1939) *Phys. Rev.*, **56**, 750.

Slaughter, D.A., Karaganov, V., Brunger, M.J., Teubner, P.J.O., Bray, I., and Bartschat, K. (2007) *Phys. Rev. A*, **75**, 062717.

Sohn, M. and Hanne, G.F. (1992) *J. Phys. B*, **25**, 4627.

Standage, M.C. and Kleinpoppen, H. (1976) *Phys. Rev. Lett.*, **36**, 577.

Stefani, G., Avaldi, L., and Camilloni, R. (1990) *J. Phys. B*, **23**, L227.

Stepanovic, M., Minic, M., Cvejanovic, D., Jureta, J., Kurepa, J., Cvejanovic, S., Zatsarinny, O., and Bartschat, K. (2006) *J. Phys. B*, **39**, 1547.

Stephens, J.A. and Greene, C.H. (1995) *J. Chem. Phys.*, **103**, 5470.

Stevenson, M. and Crowe, A. (2004) *J. Phys. B*, **37**, 2493.

Stevenson, M.A. and Lohmann, B. (2008) *Phys. Rev. A*, **77**, 032708.

Stier, H.C. (1932) *Z. Phys.*, **76**, 439.

Surko, C.M., Gribakin, G.F., and Buckman, S.J. (2005) *J. Phys. B*, **38**, R57.

Thorn, P.A., Brunger, M.J., Kato, H., Hoshino, M., and Tanaka, H. (2007) *J. Phys. B*, **40**, 697.

Tonzani, S. and Greene, C.H. (2005) *J. Chem. Phys.*, **124**, 054312.

Trajmar, S., Williams, W., and Kuppermann, A. (1973) *J. Chem. Phys.*, **58**, 2521.

Trevisan, C.S., Houfek, K., Zhang, Z., Orel, A.E., McCurdy, C.W., and Rescigno, T.N. (2005) *Phys. Rev. A*, **71**, 052714.

Uhlmann, L.J., Dall, R.G., Truscott, A.G., Hoogerland, M.D., Baldwin, K.G.H., and Buckman, S.J. (2005) *Phys. Rev. Lett.*, **94**, 173201.

Ullrich, J., Moshammer, R., Dorn, A., Dörner, R., Schmidt, L.Ph.H., and Schmidt-Böcking, H. (2003) *Rep. Prog. Phys.*, **66**, 1463.

Vinkalns, I. and Gailitis, M. (1967) in *Abstracts of the Vth ICPEAC* (eds I.P. Flaks and E.S. Solovyov), Nauka, Leningrad, p. 648.

Visconti, P.J., Slevin, J.A., and Rubin, K. (1970) *Phys. Rev. A*, **3**, 1310.

Walters, H.R.J. (1984) *Phys. Rep.*, **116**, 1.

Wannier, G.H. (1953) *Phys. Rev.*, **90**, 817.

Weigold, E. and McCarthy, I.E. (1999) *Electron Momentum Spectroscopy*, Kluwer, Dordrecht/Plenum, New York.

Wigner, E.P. (1948) *Phys. Rev.*, **73**, 1002.

Williams, J.F. (1986) *Aust. J. Phys.*, **39**, 621.

Williams, J.F. (1988) *J. Phys. B*, **21**, 2107.

Williams, J.F. and Mikosza, A.G. (2006) *J. Phys. B*, **39**, 4113.

Winstead, C. and McKoy, V. (2006) *J. Chem. Phys.*, **125**, 174304.

Winstead, C., McKoy, V., and Bettega, M.H.F. (2005) *Phys. Rev. A*, **72**, 042721.

Wolcke, A., Bartschat, K., Blum, K., Borgmann, H., Hanne, G.F., and Kessler, J. (1983) *J. Phys. B*, **16**, 639.

Yalim, H.A., Cvejanovic, D., and Crowe, A. (1997) *Phys. Rev. Lett.*, **79**, 2951.

Zatsarinny, O. (2006) *Comp. Phys. Commun.*, **174**, 273.

Zatsarinny, O. and Bartschat, K. (2004) *J. Phys. B*, **37**, 2173.

Zatsarinny, O. and Bartschat, K. (2005) *Phys. Rev. A*, **72**,, 020702(R).

Zatsarinny, O. and Bartschat, K. (2008) *Phys. Rev. A*, **77**, 062701.

Zuo, T., McEachran, R.P., and Stauffer, A.D. (1991) *J. Phys. B*, **24**, 2853.

Further Reading

Balashov, V.V., Grum-Grzhimailo, A.N., and Kabachnik, N.M. (2000) *Polarization and Correlation Phenomena in Atomic Collisions*, Springer, New York.

Bartschat, K. (ed.) (1996) *Computational Atomic Physics*, Springer, Heidelberg and New York.

Becker, U., Moshammer, R., Mokler, P., and Ullrich, J. (eds) (2007) *XXV International Conference on Photonic, Electronic and Atomic Collisions, J. Phys. Conf. Ser.*, **88**, (entire issue).

Bransden, B.H. (1983) *Atomic Collision Theory*, 2nd edn., Benjamin, New York.

Burke, P.G. and Joachain, C.J. (1995) *Theory of Electron-Atom Collisions. Part One: Potential Scattering*, Springer, Heidelberg and New York.

Burke, P.G. and Moiseiwitsch, B.L. (eds) (1976) *Atomic Processes and Applications*, North-Holland, Amsterdam.

Burke, P.G. and West, J.B. (eds) (1988) *Electron-Molecule Scattering and Photoionization*, Plenum, New York.

Fainstein, P.D., Lima, M.A.P., Miraglia, J.E., Montenegro, E.C., and Rivarola, R.D. (eds) (2006) *Photonic, Electronic and Atomic Collisions*, World Scientific, Singapore.

Drake, G.W.F. (ed.) (2006) *Springer Handbook of Atomic, Molecular, and Optical Physics*, Springer, Heidelberg.

Ehrhardt, H., Fischer, M., Jung, K., Byron, F.W., Joachain, C.J., and Piraux, B. (1982) *Phys. Rev. Lett.*, **48**, 1807.

Gianturco, F.A. (ed.) (1987) *Collision Theory for Atoms and Molecules*, Plenum, New York.

Grant, I.P. (2007) *Relativistic Quantum Theory of Atoms and Molecules*, Springer, Heidelberg.

Herzenberg, A. (ed.) (1990) *Aspects of Electron-Molecule Scattering and Photoionization*, American Institute of Physics, New York.

Massey, H.S.W. and Bates, D.R. (eds) (1982) *Applied Atomic Collision Physics*, Vol. 1: Atmospheric Physics and Chemistry, Academic, New York.

Massey, H.S.W. and Burhop, E.H.S. (1969) *Electronic and Ionic Impact Phenomena*, 2nd edn., Vols. 1 and 2, Clarendon, Oxford.

McCarthy, I.E. and Weigold, E. (1995) *Electron-Atom Collisions*, University Press, Cambridge.

McGuire, J.H. (1997) *Electron Correlation Dynamics in Atomic Collisions*, University Press, Cambridge.

Mott, N.F. and Massey, H.S.W. (1965) *The Theory of Atomic Collisions*, 3rd edn., Clarendon, Oxford.

Mšrk, T.D. and Dunn, G.H. (eds.) (1985) *Electron-Impact Ionization*, Springer, Berlin.

McDaniel, E.W. (1969) *Atomic Collisions: Electron and Photon Projectiles*, John Wiley & Sons, New York.

Nesbet, R.K. (1980) *Variational Methods in Electron-Atom Scattering Theory*, Plenum, New York.

Peterkop, R.K. (1977) *Theory of Ionization of Atoms by Electron Impact*, Colorado Associated University Press, Boulder.

Shimamura, I. and Takayanagi K. (eds) (1984) *Electron-Molecule Collisions*, Plenum, New York.

Ullrich, J. and Shevelko, V.P. (eds) (2003) *Many-Particle Quantum Dynamics in Atomic and Molecular Fragmentation*, Springer, Heidelberg.

Watanabe, T., Shimamura, I., Shimizu, M., and Itikawa, Y. (eds) (1990) *Molecular Processes in Space*, Plenum, New York.

ELECTRON STATES IN ZERO-, ONE-, AND TWO-DIMENSIONAL STRUCTURES

FRANK STERN, *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, New York, U.S.A.*

	Introduction	1
1.	**Dimensionality**	2
2.	**Structures and Fabrication**	3
2.1	Induced Charge	5
3.	**Electronic Structure**	5
3.1	Electronic States	5
3.1.1	Two-Dimensional Systems	6
3.1.2	One-Dimensional Systems	7
3.2	Dielectric Response	8
3.3	Bound States	8
4.	**Transport Properties**	9
4.1	Boltzmann Transport	9
4.2	Mesoscopic Systems	11
4.3	Constrictions	12
5.	**Optical Properties**	12
6.	**Electrons on Cryogenic Substrates**	14
7.	**Devices and Applications**	15
8.	**Modeling**	16
9.	**Summary**	16
	Glossary	16
	Works Cited	17
	Further Reading	18

INTRODUCTION

For several decades, structures used in condensed-matter physics research and in many device areas have been getting smaller and have incorporated many ingenious schemes for confining electrons and, in some cases, holes. This has made it possible to realize structures in which carrier motion is effectively constrained to two, one, or most recently zero dimensions, where in the last case we consider carriers with no freedom to move in any direction. Such low-dimensional systems have been considered in physics and mathematics textbooks for a long time, but now can be realized, can be studied experimentally, and can form the basis for novel devices and applications. This has been made possible by substantial advances in growth and fabrication processes and in lithography, and the resulting field is one in which technology, materials science, and physics have all played essential roles.

This article focuses on the physics of low-dimensional structures and deals mainly with structures and effects involving electrons in semiconductors. Similar effects occur for holes, but generally require more complicated analysis because of the more complicated valence-band structures of most semiconductors. Low-dimensional effects are also found in metals and semimetals, but such examples will not be considered here.

As structures in solids continue to shrink, they eventually approach atomic size. There has been in recent years a growing interest in atomic clusters (see MOLECULAR AND ATOMIC CLUSTERS), and these two regimes, one starting from an atomic picture and the other starting from a picture based on bulk materials, will begin to merge. The approach in this article is based on a modification of the properties of bulk materials when the carriers are confined.

Although the new physical effects associated with reduced dimensionality are quite dramatic, many of them—with the exception of some of the optical effects—weaken as the temperature increases. Many of the phenomena to be discussed here are, therefore, most pronounced at low temperatures.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Part of the material in this article is based on lectures given in Trieste at the 1990 Spring College in Condensed Matter, on Physics of Low-Dimensional Semiconductor Structures (Stern, 1992).

1. DIMENSIONALITY

Although we live in a world with three space dimensions, it is easy to visualize situations in which motion can be constrained to a smaller number of dimensions. A one-dimensional system, for example, has one direction in which motion is relatively unimpeded, while motion in the other two space dimensions is constrained. Structures with one-dimensional motion are often called quantum wires. Fully confined or zero-dimensional structures are called quantum dots, or quantum boxes.

Physical confinement of electrons arises when some dimension of a sample is small on a relevant physical scale. There are many such scales, of which we discuss only three. The scale important for understanding many transport properties in solids is the phase-breaking length, the average distance over which a carrier loses phase coherence. In many cases, this can be approximated by the mean distance through which a carrier diffuses between inelastic scattering events such as electron–phonon and electron–electron scattering, and is approximately given by $(2dD\tau_{\text{in}})^{1/2}$, where d is the dimensionality, D is the diffusion constant (with dimensions $\text{cm}^2\ \text{s}^{-1}$), and τ_{in} is the mean time between inelastic scattering events. At low temperatures, the diffusion constant of most semiconductors and metals is only weakly dependent on temperature, while the inelastic scattering rate, τ_{in}^{-1}, varies as a power of the absolute temperature T. The phase-breaking length gets longer as the temperature is reduced, and can be of order μm or longer at temperatures of order 1 K in many semiconductors and metals. This leads to an interesting class of transport phenomena sometimes grouped under the heading weak localization (*q.v.*; see also ELECTRONIC TRANSPORT IN SOLIDS), a subject that has been actively studied since about 1980.

Another physical length is the carrier mean free path, an effective distance traveled between collisions, both elastic and inelastic. In systems with high carrier mobility at low temperatures, this length can be of order 5 μm or larger. In a sufficiently small structure, carriers can traverse the entire device without scattering and are in the ballistic regime, with analogies to motion in free space.

The third physical scale length, and the one of direct interest in this article, is related to the wavelength of the electron, given by Planck's constant divided by the electron momentum. At low temperatures, the physical length scale for carrier confinement is best given by k_{F}^{-1}, the inverse of the wave vector at the Fermi surface. The Fermi wave vector in a simple solid with an isotropic, nondegenerate energy band is related to the carrier density by $n = k_{\text{F}}^3/3\pi^2$.

Reduced dimensionality manifests itself physically in a number of ways. Physical properties such as conductance are anisotropic and respond in an anisotropic way to external electric and magnetic fields. The energy-level structure, as reflected in electrical and optical properties, often shows a distinct signature of the dimensionality. For example, the first clear evidence for the existence of a two-dimensional electron system was the characteristic appearance of equally spaced peaks in the conductance of a silicon inversion layer in a magnetic field versus gate voltage, measured by Fowler *et al.* (1966). Collective effects, such as plasma modes, also carry a distinctive signature of the dimensionality, as in the plasmon dispersion described in Sec. 3.2.

The study of layered materials and of bulk materials with one-dimensional conducting strands has evolved separately from the study of low-dimensional structures in semiconductors. For historical reasons and because the materials covered are somewhat different, these fields have almost disjoint sets of researchers, of conferences, and of publications. (See GRAPHITE; INTERCALATION COMPOUNDS.)

Much of the theoretical work on thermodynamic properties of low-dimensional systems is concerned with phase transitions (*q.v.*; see also METAL-INSULATOR TRANSITIONS). We shall touch on one example below, in Sec. 6.

2. STRUCTURES AND FABRICATION

The methods developed for confining carriers can be grouped into several major categories, including compositional confinement, electrostatic confinement, geometrical confinement, localized damage, and selective growth. Confinement in actual structures usually involves more than one of these, and new combinations and fabrication schemes continue to be developed. Schematic pictures of three such structures are shown in Fig. 1.

Compositional confinement arises from the barrier to carrier motion imposed by a change—either abrupt or graded—in chemical composition. One example is the barrier, about 3 eV, for electrons at the interface between Si and SiO_2, an interface of great technological importance because of the dominant role of silicon in semiconductor device applications. Another is the conduction-band offset at the interface between GaAs and $Al_xGa_{1-x}As$. This interface is widely studied because there is little difference in lattice constants through the $Al_xGa_{1-x}As$ alloy system, allowing layers of various compositions to be deposited successively epitaxially with very little strain. The conduction-band offset is about 0.3 eV if the AlAs fraction x is 0.3. Thus a quantum well structure that confines a layer of GaAs between layers of $Al_{0.3}Ga_{0.7}As$ will create a well 0.3 eV deep for electrons. Holes will also be confined in such a quantum well, by about 0.15 eV. Many other structures, including multiple quantum wells and superlattices, are possible in the $Al_xGa_{1-x}As$ system and also in other materials (see HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR).

Electrostatic confinement was one of the first means used to create a true quantum-confined two-dimensional system, the inversion layer of a metal-oxide-silicon (MOS) field-effect transistor (FET) structure. Applying a voltage between the metallic gate and the silicon in an MOS structure produces an electric field that attracts carriers toward the semiconductor surface, with a corresponding charge density per unit area, but of opposite sign, on the gate. The conduction-band discon-

(a)

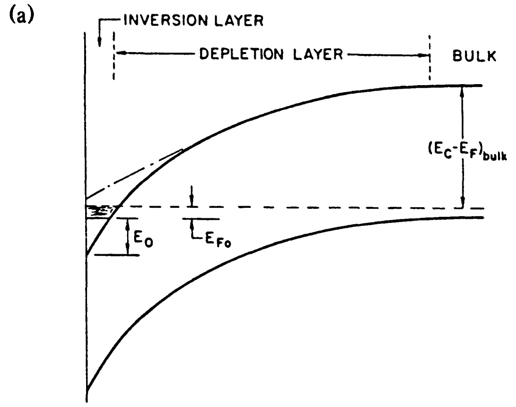

(b)

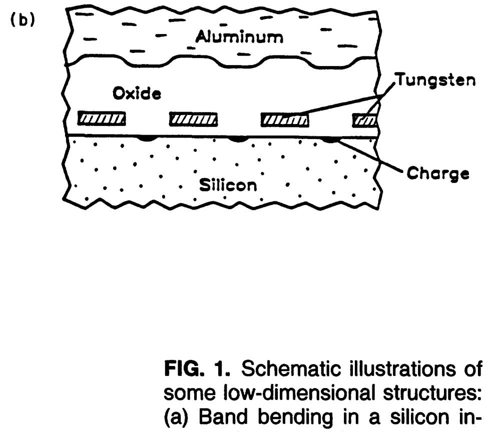

(c)

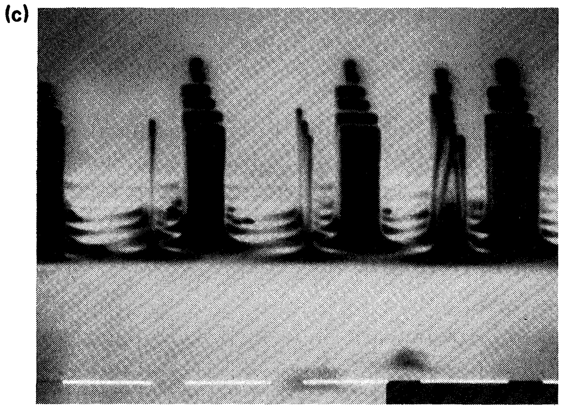

FIG. 1. Schematic illustrations of some low-dimensional structures: (a) Band bending in a silicon inversion layer, showing conduction- and valence-band edges, the Fermi level, and the energy of one subband. (b) Two-gate structure for producing an array of quantum wires. A positive voltage is applied to the upper gate and induces electrons at the $Si-SiO_2$ interface, below the openings in the lower grating gate. [After Warren *et al.* (1986).] (c) Scanning electron micrograph of GaAs nanostructures containing quantum dots. Vertical confinement is provided by a quantum well in each column. The horizontal bars are 0.5 μm wide. [After Reed *et al.* (1990), (1991).]

tinuity between silicon and SiO_2 prevents the carriers from entering the oxide, and leads—if the electric field is large enough—to a thin layer of carriers near the silicon surface. For the case when the silicon is *p* type, i.e., with holes as the majority carriers, the surface layer is called an inversion layer if the gate voltage is positive and electrons are attracted toward the Si–SiO_2 interface, and is called an accumulation layer if the gate voltage is negative and holes are attracted toward the interface. Band bending in an inversion layer is shown schematically in Fig. 1(a). Such metal-oxide–semiconductor systems underlie the field-effect transistors, or MOSFETs, that are the basis of many semiconductor devices used in computers and in other electronic equipment (see TRANSISTORS, FIELD EFFECT). Silicon inversion layers were the first widely studied two-dimensional system. Some of the extensive literature of the physics of this system is covered by Ando *et al.* (1982).

Electrostatic confinement can also be used to reduce the dimensionality further. Suitably patterned gates, such as narrow gates, split gates, grating gates, and other variants, can be used to shape the electrostatic confining potential in ways which, when combined with one or more barriers due to compositional confinement, lead to effectively one- or zero-dimensional systems. A double-gate structure used by Warren *et al.* (1986) for electrostatic confinement of arrays of quantum wires in silicon is shown in Fig. 1(b).

Geometrical confinement is the simplest mechanism to consider, although not necessarily the easiest to realize. One or more sample dimensions are made suitably small either by an appropriate growth technique or by removing material from a larger sample, for example, by etching. Posts of material with diameter of order 100 nm or less have been prepared from a GaAs-based heterostructure, and the combination of compositional and geometrical confinement has been used to fabricate structures for optical studies, including lasers, and for tunneling experiments. Figure 1(c) shows such a structure, used in quantum dot tunneling experiments in 1988 by Reed *et al.* (1990, 1991). One of the limitations of geometrical confinement is the presence of surface states that may deplete charges from a region near the surface or have adverse effects on carrier mobility, surface recombination velocity, or other properties of the device. Methods for mitigating these effects by chemical or other surface treatments continue to be investigated.

Some processes that introduce localized damage, such as bombardment by ions or electrons, lead to changes in the surface or bulk properties of the material that have been used to confine carriers. While the effectiveness of these methods has been demonstrated and can be understood in part to result from the creation of trapping states in the energy gap, work remains to be done to clarify the microscopic structure of the defects and the processes by which they are created.

Selective growth on a terraced semiconductor surface, as obtained for example by using a surface slightly tilted from a $\{001\}$ plane, can lead to a wirelike structure whose width can be controlled by the growth conditions and the terrace width and whose thickness can be as small as a monolayer. Selective growth can also be used on the etched or cleaved surface of a superlattice or other compositionally modulated structure to yield lateral confinement.

Methods have also been developed to fabricate small particles in an inert matrix—for example, small particles of semiconductor material in a glassy matrix. See, for example, Ekimov *et al.* (1989).

An ingenious method has been devised to confine excitons (*q.v.*) by use of strain effects. (An exciton, an electron and hole coupled together by their Coulomb attraction, is neutral and therefore cannot easily be confined electrostatically.) See, for example, the article by Kash (1990), which summarizes work on optical properties of quantum wires and quantum dots in III-V semiconductors.

Many of the methods we have described can be used to fabricate quantum wires or quantum dots, or arrays of wires or dots that may or may not interact with each other. Alternatively they can be used to add spatial modulation, as, for example, to introduce linear or rectangular grating modulations onto a two-dimensional electron system. Such modulated systems allow a transition from one dimensionality to another as the strength of the modulation is varied.

Structures with very fine spatial features, of order 0.1 μm or less, usually require advanced fabrication technology and tend to be made in large laboratories equipped with focused electron or ion beams or with x-ray lithography. Structures that do not require such elaborate facilities can also exhibit spatial confinement. The number of laboratories making low-dimensional structures and devices continues to increase.

2.1 Induced Charge

Fabricating a particular low-dimensional structure, as briefly described above, may suffice for structures to be studied optically, but for electrical studies it is also necessary to supply mobile carriers. In bulk semiconductors, donor impurities are added to supply electrons, or acceptor impurities to supply holes, but in a homogeneous system the impurities and the carriers occupy the same region of space, leading to strong scattering and usually to low carrier mobility. In several categories of low-dimensional systems, carriers can be introduced without impurities or they can be spatially separated from them.

The metal-oxide–semiconductor (MOS) structure briefly described in the Introduction is an example of the first mechanism for introducing charge. The semiconductor is one plate of a capacitor, and application of a voltage V across the insulator induces a charge, approximately given by $C_{ox}(V-V_{th})$, where C_{ox} is the oxide capacitance and V_{th} is the threshold voltage. At voltages below the threshold, the charge need not drop abruptly to zero; there is, for example, a thermally activated subthreshold charge, not considered further here. The maximum surface charge that can be induced is related to the breakdown electric field of the insulator. For a good insulator such as SiO_2, with a breakdown field close to 10^7 V/cm, the density of induced surface charges can exceed 10^{13} cm^{-2}.

Dopant atoms placed on the high-barrier side of a heterojunction transfer carriers to the low-barrier side to equalize the Fermi levels. If the dopant atoms are separated from the heterojunction by a spacer layer, a process called modulation doping or selective doping, the carrier mobility may be enhanced because of the increased separation between carriers and ionized impurities. The amount of charge transferred in this way is controlled by the density, location, and binding energy of the impurities, by the band offset between the two sides of the heterojunction, and by any external voltages applied. Induced charge densities in heterojunctions are usually smaller than those possible in silicon MOS structures.

Charge can also be induced in a semiconductor by a charged layer of surface states or surface atoms, which is compensated by a space-charge layer in the semiconductor. Surface states, as controlled by ambient gas atmospheres, were being investigated by Brattain and Bardeen in the early 1950s (see ELECTRONIC STRUCTURE OF SURFACES). The physics of semiconductor surfaces, a forerunner of the physics of inversion layers and related two-dimensional electron systems, has been well described in the book by Many *et al.* (1965). In some materials, for example ZnO, it is possible to create charged surface states, and therefore a compensating space-charge layer, with more than 10^{14} charges per cm^2.

At the other extreme of charge density is the system of electrons on liquid helium, discussed in Sec. 6 below.

3. ELECTRONIC STRUCTURE

The wave functions of electronic states in a perfect crystal can be expressed in the Bloch form (see ELECTRONIC STRUCTURE OF SOLIDS)

$$\psi_{\mathbf{k}}(\mathbf{r})=e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r}), \qquad (1)$$

where **k** is the wave vector of the state and $u_{\mathbf{k}}(\mathbf{r})$ has the translational periodicity of the crystal lattice. In the effective-mass approximation, which underlies the discussion in this article, the periodic function $u(\mathbf{r})$ is often omitted in describing electronic states, and the electrons are considered to have an effective mass m^* that is usually different from the free-electron mass m_0. For some semiconductors, such as silicon, it is necessary to consider the anisotropy of the effective mass and the multivalley nature of the conduction band. The complications that arise from the degenerate valence-band structure are largely ignored in this article.

3.1 Electronic States

In the effective-mass approximation approach, it is often sufficient to describe electronic states by an envelope function that includes the effects of external potentials, which are presumed to vary slowly on an atomic scale, and of boundary conditions. In this approximation the potential variations on an atomic scale are considered only through their influence on the effective mass and the dielectric constant. Although one would normally expect this approximation to be valid when the potential is slowly varying on an atomic distance scale, it holds up surprisingly well even for quantum-well structures only a few atomic layers across, as shown by calculations that include the discrete atomic structure (see, for example, Gell *et al.*, 1987).

3.1.1 Two-Dimensional Systems. The wave function for an electronic state in a two-dimensional system in the presence of a confining potential $V(z)$ can be approximated as

$$\psi_{n,\mathbf{K}}(\mathbf{R},z)=e^{i\mathbf{K}\cdot\mathbf{R}}\zeta_n(z)u(\mathbf{r}), \qquad (2)$$

where $u(\mathbf{r})$ is the periodic part of the Bloch function of Eq. (1), taken at a point in the Brillouin zone representative of the states being described, the plane-wave factor describes free-carrier motion in the x-y plane, and the envelope function $\zeta_n(z)$ is the solution of the Schrödinger equation

$$-\frac{\hbar^2}{2}\frac{d}{dz}\frac{1}{m^*}\frac{d\zeta_n}{dz}+V(z)\zeta_n(z)=E_n\zeta_n(z), \qquad (3)$$

where $\hbar$ is the Planck constant divided by 2π, with the boundary condition that the wave function vanish at large values of $|z|$. Note that we use uppercase letters to represent vectors, or lengths of vectors, in two dimensions and lower case letters for three dimensions. Thus the vector $\mathbf{R}$ has coordinates (x,y) and the vector $\mathbf{r}$ has coordinates (x,y,z). We here consider a simple case with an isotropic effective mass independent of energy, but allow for the possibility that the effective mass may have a spatial variation associated with composition changes along the z direction. Multivalley effects and mass anisotropy effects arise in silicon inversion layers, for example, and are discussed in the paper by Stern and Howard (1967). More complicated situations can arise in heterostructures. In the discussion that follows we shall usually omit the factor $u(\mathbf{r})$ from the wave function, but it is essential in some cases, such as the calculation of the strength of optical transitions.

The energy levels of the system described by Eqs. (2) and (3) form series of continua, each of which has its origin at an eigenvalue E_n of Eq. (3), with energies

$$E=E_n+(\hbar^2K^2)/2m^*. \qquad (4)$$

Each of these two-dimensional continua arises out of states of the conduction band as modified by the z-dependent potential, and is called a subband. The corresponding continuum of levels in a superlattice, where the energy depends both on the wave vector in the plane and on the wave vector along the superlattice axis, has come to be called a miniband.

The density of states per unit energy and unit area in the nth two-dimensional subband is

$$\rho(E)=g_s\, g_v m_d/2\pi\hbar^2, \quad E>E_n \qquad (5)$$

and vanishes for $E<E_n$. We have introduced m_d, the density-of-states mass per valley for motion in the K_x-K_y plane; g_v, the valley degeneracy; and g_s, the spin degeneracy, which equals 2 unless a magnetic field is present to separate spin-up and spin-down levels. Note that for a two-dimensional system the density of states is independent of energy within each subband under the assumptions used here. The total density of states is a sum of such strips of constant density, each starting at the bottom of the corresponding subband. For electrons in GaAs the valley degeneracy g_v is 1, and for electrons in the lowest subband of an inversion layer on a {001} surface of silicon it is 2.

In two dimensions, unlike one and three dimensions, the integral of the density of states times the Fermi–Dirac occupation function can be evaluated analytically, and the density of electrons per unit area in the nth subband at temperature T is

$$N_n=(g_s\, g_v m_d k_B T/2\pi\hbar^2)$$
$$\times\ln\{1+\exp[(E_F-E_n)/k_BT]\}, \qquad (6)$$

where E_F is the Fermi energy and k_B is the Boltzmann constant. If the carrier density is large enough that $(E_F-E_n)>k_BT$, then the system is said to be degenerate and the Fermi level relative to the subband edge is approximately proportional to the carrier density if only one subband is occupied. If the carrier density is very small, the Fermi level lies below the bottom of the lowest subband and the system is said to be nondegenerate.

The energy levels of a two-dimensional system in a magnetic field (strictly, magnetic induction) B perpendicular to the plane of motion are the Landau levels $(i+\frac{1}{2})\hbar\omega_c$, where

$\omega_c = eB/m^*$ is the cyclotron radian frequency and i is an integer. Thus when a magnetic field is present perpendicular to the plane of a two-dimensional electron system, the constant density of states described in Eq. (5) is rearranged, and the levels form a series of peaks separated by the cyclotron energy. The number of states per unit area in each peak is eB/h, which must be multiplied by the spin and valley degeneracies if these are not resolved. If there were no scattering or disorder, the peaks would have zero width, but in a real system there is a finite width. There is good evidence that for high magnetic fields, the states in the tails of the Landau levels are localized and do not carry a current.

The conductance of a two-dimensional system in a magnetic field shows clear evidence of the equally spaced Landau levels expected from the simple arguments just described. When a gate voltage is applied to a metal–insulator–semiconductor system, the charge in the two-dimensional system is approximately linear in gate voltage. Thus as the gate voltage is increased in the presence of a magnetic field, a fixed amount of charge is required for each Landau level. This leads to a periodic dependence of conductance on magnetic field, as shown in Fig. 2 from the work by Fowler *et al.* (1966) that first clearly demonstrated, using a silicon MOS structure, the presence of two-dimensional electrons.

Electron–electron interactions have a substantial effect on the energy levels in metals and semiconductors (see Collective Phenomena in Solids). These effects are even more important in two-dimensional systems than in bulk systems, because they can be of the same magnitude as the energy-level spacings connected with spatial quantization. Early optical experiments on inversion layers showed that the measured spectrum of transitions between subbands cannot be reconciled with calculations in the Hartree approximation, which ignores the "hole" in the electron distribution that follows each electron dynamically. Other evidence for the importance of many-electron interactions comes from the effective mass and the g factor, as inferred from magnetotransport and magneto-optical measurements [see, for example, Ando *et al.* (1982) for an account of such experiments and related theory].

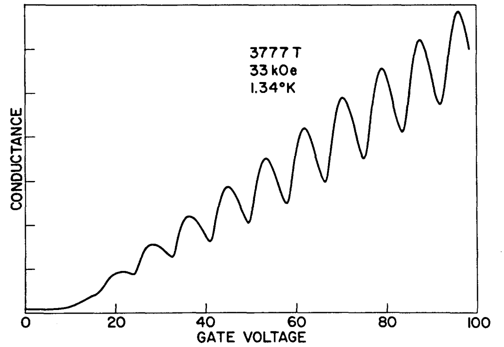

FIG. 2. Conductance of an inversion layer in a metal-oxide–silicon structure in a magnetic field vs applied gate voltage. The equally spaced oscillations, associated with the equally spaced Landau levels induced by the magnetic field, provided the first clear demonstration of a two-dimensional electron gas. [After Fowler *et al.* (1966).]

As in three dimensions, one can separate the many-electron effects into an exchange part, which reflects the energy gained by keeping electrons of like spin apart, and a correlation part, which includes all other effects (see Collective Phenomena in Solids). The exchange contribution to the energy is $-e^2k_F/3\pi^2\bar{\epsilon}$ in two dimensions, where $\bar{\epsilon}$ is the average permittivity of the media surrounding the electron plane, as compared to the value $-3e^2k_F/16\pi^2\epsilon$ found in three dimensions. Note that these values do not depend explicitly on the effective mass. Estimating the correlation effects requires elaborate calculations, for which the so-called plasmon-pole approximation to the dielectric response has been frequently used. Ando *et al.* (1982) and Czachor *et al.* (1982) give results on correlation effects in two-dimensional systems, with references to the earlier work.

The local-density formulation (see Collective Phenomena in Solids) can be used to approximate the many-electron effects in heterostructures, and adds an effective potential to the potential in the Schrödinger equation, Eq. (3). One complication in consideration of electron–electron interaction in inversion layers arises from the large difference of dielectric constants of Si and SiO_2. The resulting image potential modifies the effective interaction of two electrons near the Si–SiO_2 interface, as first treated by Ando (1976).

3.1.2 One-Dimensional Systems. The equations describing a dynamically one-dimensional case are similar to those given in the preceding section, except that the electrons are only free to move in one dimension, and the Schrödinger equation satisfied by the envelope function now involves two space dimensions. We only give here the density of states per unit energy and unit length for the one-dimensional case,

$$\rho(E)=(g_s\,g_v/2\pi)(2m_x/\hbar^2)^{1/2}(E-E_n)^{-1/2}, \tag{7}$$

with its divergence at the threshold. Here m_x is the effective mass for motion along the wire. The density of charges per unit length in the nth subband is

$$N_n=(g_s\,g_v k_B T/2\pi)(2m_x/\hbar^2)^{1/2} \times F_{-1/2}((E_F-E_n)/k_B T), \tag{8}$$

where

$$F_p(\eta)=\int_0^\infty \frac{x^p}{1+\exp(x-\eta)}\,dx \tag{9}$$

is the Fermi–Dirac integral of order p. In real systems, scattering processes will broaden the levels and remove the singularity at the bottom of each one-dimensional subband.

3.2 Dielectric Response

The self-consistent response of a two-dimensional free-electron system to a weak electric field of radian frequency ω and wave vector $\mathbf{Q}$ in the electron plane was obtained by Stern (1967) in the simplest approximation, often called the random-phase approximation, which ignores many aspects of electron–electron interaction (Ehrenreich and Cohen, 1959). The long-wavelength Fourier components of an applied static potential are screened by a factor $Q/(Q+Q_s)$, where the screening parameter is given by $Q_s=g_v m^* e^2/2\pi\bar{\epsilon}\hbar^2$ in a degenerate system and by $Q_s=e^2N/2\bar{\epsilon}k_B T$ in a nondegenerate system, where N is the two-dimensional carrier density. Note that the screening parameter is independent of carrier density in this long-wavelength limit in the degenerate case, reflecting the fact that the density of states of a two-dimensional electron gas is independent of energy within a subband. In three dimensions, the corresponding screening factor has the form $q^2/(q^2+q_s^2)$. The screened potential of a point charge falls off as $\exp(-q_s r)$ in a three-dimensional system, and falls off as R^{-3} for large values of R in two dimensions. Thus the screening in two dimensions is considerably weaker than in three dimensions, and the two-dimensional screening does not set a length scale in the same way as its counterpart in three dimensions. At low temperature and if level broadening is small, the potential also has oscillatory terms, the Friedel oscillations, which arise from the sharp drop in the screening of a two-dimensional electron gas at a wave vector equal to the diameter of the Fermi circle, the occupied region of wave-vector space.

The dynamical response of a two-dimensional electron gas leads to collective longitudinal modes, called plasma oscillations or plasmons, whose dispersion at small wave vectors is $\omega_p=(Ne^2Q/2\epsilon m^*)^{1/2}$, where N is the two-dimensional carrier density, first obtained by Ritchie in 1957, which has a square-root dependence on wave vector. This is considerably different from the three-dimensional result $\omega_p=(ne^2/\epsilon m^*)^{1/2}$, where n is the three-dimensional carrier density. The square-root dispersion of plasma oscillations in a two-dimensional system, which becomes quasilinear at small wave vectors if there are nearby gates or other metallic regions, has been verified experimentally both for electrons on liquid helium and for electrons in two-dimensional systems in semiconductors. This subject, including effects of magnetic fields, has been reviewed by Theis (1980).

Detailed studies of the dielectric response of the two-dimensional electron gas, with references to earlier work, are given by Czachor *et al.* (1982) and by Tanatar (1991). Studies for one-dimensional electron systems have been carried out by Friesen and Bergersen (1980) and others.

3.3 Bound States

Just as the hydrogen atom has proved to be a useful model system for discussing atomic spectra, its analog in bulk semiconductors has played a similar role. The energy levels of an electron bound to a center with a (positive) charge e in a semiconductor with permittivity ϵ and effective mass m^*, measured relative to the adjacent conduction-band edge, are

$$E_n=-m^*e^4/32\pi^2\epsilon^2\hbar^2n^2,\quad n=1,2,\dots, \tag{10}$$

in the hydrogenic approximation, whose validity improves for small effective mass and large dielectric constant. Corrections, sometimes called central-cell corrections, can change the energy significantly, particularly for the lowest states, whose wave functions are most localized. If mobile carriers are present, they will screen the attractive potential and reduce the binding energy. In three dimensions bound states arise only if the attractive potential is sufficiently strong. Thus for sufficiently high carrier densities, with correspondingly strong screening of the Coulomb potential, there are no bound states in three-dimensional systems.

In two dimensions the situation is quite different. The hydrogenic binding energy of an electron confined to a plane, in the presence of the attractive potential of a single (positive) charge e located in the plane is

$$E_n = -m^* e^4 / 8\pi^2 \bar{\epsilon}^2 \hbar^2 n^2, \quad n = 1,2,..., \qquad (11)$$

where $\bar{\epsilon}$ is the average permittivity of the media on the two sides of the plane in which the electron moves. If the dielectric constant is the same on both sides, the hydrogenic binding energy in two dimensions is four times as large as the bulk value. Screening effects will reduce the binding energy, as will a displacement of the attractive charge out of the plane in which the electrons move. In real two-dimensional systems, the electron wave function has a nonzero spatial extent perpendicular to the plane, and this also reduces the binding energy.

Another important result of the lower dimensionality is that there is no longer a threshold for appearance of a bound state. An arbitrarily weak attractive potential in a two-dimensional system can have a bound state, although the binding energy rapidly approaches zero and the spatial extent of the wave function becomes extremely large as the potential weakens.

Bound states in silicon inversion layers associated with sodium ions at the Si–SiO$_2$ interface have been seen in electrical measurements by Fowler and Hartstein (1980) and in optical measurements by Glaser *et al.* (1986) (see Interface Structures, Electronic Properties of; Si–SiO$_2$ Interface, Electronic Properties of). Detailed investigations of impurity levels in quantum wells have been made by a number of authors, for example, Reeder *et al.* (1988). For additional discussion, including theoretical aspects, see Heterostructures and Superlattices, Semiconductor.

Extensions of these results to dimensionality less than two are less simple in form, because the spatial extent of the wave function plays an increasingly important role. The general trends continue, however. Lower dimensionality leads to stronger binding.

4. TRANSPORT PROPERTIES

Although the general heading of transport properties includes heat transport and thermoelectricity, the discussion in this article is limited to electrical conduction. Many of the effects found in bulk systems are found in low-dimensional systems as well. For example, optical-phonon scattering is generally the dominant scattering mechanism at room temperature. There are two new elements that enter in low-dimensional systems: the states that carry current are now the constrained states of the particular structure, and the presence of interfaces may add new scattering mechanisms. Also, the size of the device in the direction of current flow may be comparable to the phase-breaking length or the mean free path, in which case the theory used to describe the transport differs from the conventional Boltzmann transport theory.

4.1 Boltzmann Transport

For systems that are large in the direction of current flow but small in one or two other directions, the effect of an electric field can be described as a perturbation on the occupation of current-carrying states, as given by the Boltzmann equation (see Electronic Transport in Solids), leading to a nonzero current flow. In some cases (for example, elastic scattering only) the Boltzmann equation can be solved explicitly and the scattering rate is given by an integration over all scattering events, weighted according to their influence on current flow. In other cases (for example, where optical-phonon scattering is important), more elaborate means for solving the Boltzmann equation must be used.

The low-temperature mobility of a two-dimensional electron gas is limited mainly by scattering from ionized impurities. In a silicon inversion layer, these are usually located near the Si–SiO$_2$ interface, and lead to peak mobilities of order 3×10^4 cm^2/V s. In a heterojunction or quantum well, the impurities may be purposely introduced to supply carriers, or may be residual impurities. In the highest-mobility heterolayers, it is the residual impurities that probably dominate the scattering. Because the cross section for Coulomb scattering decreases with increasing carrier velocity, the scattering rate in high-mobility samples decreases—and the mobility therefore increases—as the carrier density increases, because of the increase in the velocity of carriers at the Fermi surface.

Interface roughness is another scattering mechanism important at low temperatures. The roughness is usually modeled through the correlation of the excursion of the relevant interface from a plane. A Gaussian model,

$$\langle \delta z(\mathbf{R})\delta z(0)\rangle = \Delta^2 \exp(-R^2/\Lambda^2), \qquad (12)$$

where Δ and Λ are, respectively, the root mean square magnitude of the roughness and its lateral correlation length, was used at first, partly because of its simplicity. More recent work (Goodnick *et al.*, 1985) indicates that an exponential form for the roughness autocorrelation may be closer to experiment for silicon inversion layers. Interface roughness scattering becomes important in silicon inversion layers at high electron densities, and in quantum wells at small well widths. A discussion of low-temperature inversion-layer mobility that combines these mechanisms and analyzes the peak mobility versus carrier density has been given by Gold (1985).

Although the mobility at low temperatures in a degenerate electron system is expected to depend very little on temperature, since the velocity of the carriers involved in scattering processes is very close to the Fermi velocity, experiments by Cham and Wheeler (1980) and by Kawaguchi and Kawaji (1980) found an unexpectedly strong temperature dependence that could not be accounted for by known phonon-scattering mechanisms. Stern (1980) showed that the strong singularity at $2k_F$ of the wave-vector–dependent screening by a two-dimensional electron gas leads to a substantial temperature dependence even without phonon scattering. Detailed analytic expressions for the effect were given by Gold and Dolgopolov (1986).

Acoustic phonons scatter electrons both by deforming the lattice and perturbing the energy-band structure, called deformation-potential scattering, and—in polar materials—by creating electric fields, called piezoelectric scattering. At low temperatures the strength of these two scattering mechanisms is generally proportional to T. At higher temperatures, usually of order 100 K and above, optical-phonon scattering becomes important and generally increases rapidly with increasing temperature. Phonon-scattering mechanisms in semiconductor layers have been described in detail by Price (1981). Detailed analysis of high-temperature transport in low-dimensional systems is complex, both because the matrix elements for the scattering processes involve the wave functions of the confined states and because many subbands are occupied at elevated temperatures, requiring intersubband scattering processes to be treated. As the temperature increases and many subbands become occupied, the transport properties are expected to approach those of the bulk material.

An example of the temperature dependence of electron mobility in a high-quality Al$_{0.35}$Ga$_{0.65}$As/GaAs heterojunction, from the work of Pfeiffer *et al.* (1989), is shown in Fig. 3. Very high mobilities, of order 10^7 cm^2/V s, are achieved only at low temperatures and at very low current densities, and require samples with tight control over residual impurities. Figure 3 shows the great strides that have been made in improving electron mobility in GaAs-based heterojunctions.

If the electric field that induces the current is increased, the carrier temperature rises and the mobility drops. Non-Ohmic effects are seen in all two-dimensional systems as well as in three-dimensional systems, but the details depend on quantitative aspects of the scattering mechanisms and energy-level structure of the system.

When a current flows in a two-dimensional system with a magnetic field applied perpendicular to the electron plane, a current (or voltage, if current flow is blocked) develops at right angles to the direction of the current and the field. This is the Hall effect, well known in bulk systems (see ELECTRONIC TRANSPORT IN SOLIDS). The discovery (von Klitzing *et al.*, 1980) that at high magnetic fields the Hall conductance is accurately quantized in integer multiples of e^2/h, the famous quantum

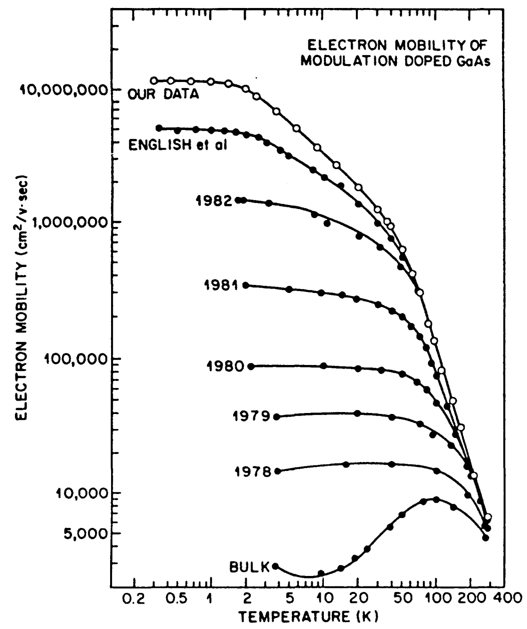

FIG. 3. Electron mobility in modulation-doped $Al_xGa_{1-x}As/GaAs$ heterojunctions, showing steady increases over the years. Note that all the mobilities approach the bulk value at 300 K. [After Pfeiffer *et al.* (1989).]

Hall effect, led to important theoretical and practical developments, which are described elsewhere (see QUANTUM HALL EFFECT, where the subsequent discovery of the fractional quantum Hall effect is also described), and to the award of the Nobel Prize in Physics to Klaus von Klitzing in 1985.

Microwave absorption between the Landau levels induced by a magnetic field B perpendicular to the plane of a two-dimensional electron gas, called cyclotron resonance, occurs at radian frequency eB/m^* if there are no impurities or defects. The experiments have a rich structure and an extensive literature. Work on cyclotron resonance in silicon inversion layers has been discussed by Ando *et al.* (1982). Richter *et al.* (1989) describe cyclotron resonance in doped heterostructures.

Sakaki (1980) suggested that the mobility in one-dimensional electron systems might be enhanced over corresponding values in two-dimensional systems because the only scattering that can limit the current is backscattering, in which the carrier direction reverses. Such events have a large momentum transfer, and this reduces the cross section for Coulomb scattering. There has been, as of this writing, no detailed confirmation of this idea. Measured mobilities for quantum wires are lower than those of the best two-dimensional structures. Quantum-mechanical corrections to a simple Boltzmann description of transport become stronger as the dimensionality is reduced, and may limit the applicability of ideas based on intuition developed for bulk transport.

Dimensionality has a strong influence on the transition between localized and delocalized states, in some ways similar to its influence on bound states. For dimensions above two, there is a threshold in the amount of disorder needed to localize states in a band. For one and, of course, for zero dimensions, all states are localized at absolute zero. In two dimensions, which represents a boundary between these two regimes, all states are weakly localized even in the limit of very weak disorder, but for sufficiently strong disorder states can be strongly (exponentially) localized. (See METAL–INSULATOR TRANSITIONS; ELECTRON STATES, LOCALIZED.)

4.2 Mesoscopic Systems

In two-dimensional systems, such as heterojunctions and inversion layers, it is possible to make samples in which one or more dimensions are small compared to the phase-breaking length, mentioned in Sec. 1. In this regime, sometimes called the mesoscopic regime, the assumptions underlying a Boltzmann transport picture are no longer valid (see MESOSCOPIC SYSTEMS). Phase coherence leads to interference effects not present at higher temperatures.

Among the phenomena that characterize the mesoscopic regime are fluctuations, often of order e^2/h, in the conductance of a sample as some physical parameter such as the voltage on a control gate, or the value of an applied magnetic field, is varied. Another effect seen in such samples is the Aharonov–Bohm effect, a periodic variation in the resistance of a ring as the magnetic flux through the ring changes by multiples of the flux quantum, h/e. A review of some of these effects has been given by Washburn and Webb (1986).

An approach to transport based on the transmission of a current past a scatterer or obstacle, originally proposed by Landauer in

1957, has been extended by him, Büttiker, and others to provide a valuable framework for considering transport in mesoscopic systems, with or without external magnetic fields present. The role of contacts as reservoirs in which equilibration takes place is an essential part of this picture. It is now recognized that simple concepts, such as the additivity of resistances of successive segments of a wire, can break down completely in sufficiently small structures. A recent discussion of transport in quantum wires has been given by Timp (1991).

4.3 Constrictions

The many methods of patterning samples, some of them described in Sec. 2, allow the fabrication of structures that cannot be described as quantum wires or dots. One such category, constrictions, has led to a number of dramatic experimental and theoretical discoveries. A constriction can be introduced by using electrostatic or other means to create a narrow neck separating two wider regions. This could take the form of a quantum wire, if the constriction is long compared to its width, or it could just be a saddle point separating two wider regions. van Wees *et al.* (1988) and Wharam *et al.* (1988) found that the conductance of a narrow constriction separating two quasi two-dimensional electron systems had a steplike structure as the voltage that controlled the size of the constriction was varied, and was approximately (within about 1%) quantized in multiples of $2e^2/h$, where the factor 2 arises from spin degeneracy.

The approximate quantization of conductance of constrictions is understood qualitatively. Far from the constriction, there are many current-carrying modes or channels. Near the constriction most of these modes are raised to such high energies that they lie well above the Fermi energy and cannot carry current past the constriction. For the few modes that can still pass a current, the transmission coefficient can be very close to unity if the constriction narrows and then widens gradually in the direction of current flow. That is expected to be the case for constrictions made electrostatically via patterned gates, because of fringing of the field lines between the gate and the layer where the electrons are located, and accounts for the experimental results. Constrictions with sharp corners would be expected to have additional structure in the dependence of conductance on gate voltage or another parameter controlling the width of the constriction. This subject has been well reviewed in the article on quantum transport in semiconductor nanostructures by Beenakker and van Houten (1991).

5. OPTICAL PROPERTIES

The changes in electronic structure associated with reduced dimensionality have significant effects on the optical properties in a one-electron picture. In addition, the effects of Coulomb interaction between the electron and hole that take part in an optical transition also enter in an important way. We saw, for example, that the binding energy of a charge attracted to an unscreened Coulomb potential increases by a factor of 4 on going from three to two dimensions. The combined effects of density of states and of Coulomb interaction serve to enhance the strength of optical transitions near the optical threshold. An important consequence is that the effects of low dimensionality on some optical properties can persist to room temperature, while the effects on transport properties generally are only seen at cryogenic temperatures. Thus some optical applications of low dimensionality have the advantage over electrical applications that low temperatures may not be required.

An early optical experiment on a two-dimensional system was the far-infrared intersubband absorption study of silicon accumulation layers by Kamgar *et al.* (1974). Because such transitions involve states within the same band of the underlying semiconductor, the optical matrix element between the Bloch function factors in the wave function is zero or very small. The optical matrix element in this case is proportional to $|\zeta_f^* \mathbf{E}\cdot\nabla\zeta_i|$, where ζ_i and ζ_f are the envelope wave functions of the initial and final states. Because the envelope functions depend only on the z coordinate, the matrix element will vanish in this simple example unless the electric field $\mathbf{E}$ of the incident light has a component polarized in the z direction. Light incident normal to the inversion layer is, therefore, not absorbed by this process. Kamgar and co-workers constructed a waveguide structure to funnel the light along the two-dimensional layer, but many other schemes for achieving the required polarization have been used. The experiments were in good agreement with theory after the effects of electron–electron interaction and so-called depolarization corrections were taken into account. See, for example, the discussion in Ando *et al.* (1982).

Parabolic potentials can be imposed on electronic motion in low-dimensional structures either by compositional grading, as in parabolic quantum wells, or as a natural consequence of electrostatic confinement in quantum wires and quantum dots. A remarkable property of electron motion in a parabolic potential is that many of the electron–electron interactions that perturb energy levels drop out in optical excitations of such a system (Brey *et al.*, 1989). That effect is exhibited in the far-infrared absorption of an array of quantum wires, studied by Hansen *et al.* (1987) and illustrated in Fig. 4. In these samples the density of a two-dimensional electron gas in a GaAs-based heterojunction is modulated by changing the voltage on a gate deposited over an insulator whose thickness varies laterally with a period of 500 nm. When a negative voltage is applied to the gate, charges induced in the GaAs by donors in the adjacent $Al_xGa_{1-x}As$ are depleted. For the upper range of gate voltages shown, the experiment measures the plasma resonance of a modulated two-dimensional electron gas, whose energy drops as the gate voltage becomes negative because the average density is decreasing. The clear break in the curve in Fig. 4 is associated with decoupling of the charge into adjacent quantum wires, whose effective width and effective charge density decrease as the gate voltage becomes more negative. The measured energy is identified as the harmonic oscillator energy of the bare potential in the wire at the given gate voltage, not as an energy difference of the one-dimensional subbands. This result is supported by separate magnetotransport measurements which show that the subband separations are considerably smaller than the measured optical transition energies.

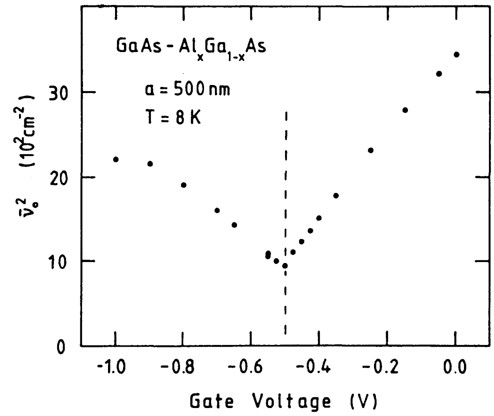

FIG. 4. Square of the resonant absorption wave number of an array of quantum wires induced in a selectively doped GaAs–$Al_xGa_{1-x}As$ heterostructure by electrostatic modulation of the carrier density using a metallic grating with a 500-nm period as a gate electrode. As the gate voltage is made more negative, the density of the two-dimensional electron gas in the GaAs is at first spatially modulated, with accompanying reduction of the average density, until at the most negative gate voltages the electrons are confined to an array of quantum wires. Note that 8 cm^{-1} corresponds to about 1 meV. [After Hansen *et al.* (1987).]

The cancellation of many-electron interactions in optical transitions also occurs when a magnetic field is present: the excitations observed are those of noninteracting electrons. In a quantum disk with a two-dimensional harmonic oscillator potential and a magnetic field perpendicular to the plane of the disk, the energy levels (first found by Fock in 1928) are

$$E=(2n+|l|+1)(E_L^2+E_O^2)^{1/2}+lE_L,$$
$$l=0,\pm1,\pm2,\ldots \qquad (13)$$

where E_O is the harmonic oscillator energy and $E_L=\hbar\omega_c/2$ is the Larmor energy. The energies of optical transitions, subject to the selection rule $\delta l=\pm1$, are given by

$$\delta E=(E_O^2+E_L^2)^{1/2}\pm E_L, \qquad (14)$$

as found by Sikorski and Merkt (1989).

Intersubband optical transitions can also be studied with inelastic light scattering. Here the polarizations of both the incident and the scattered photons enter, and one distinguishes between parallel polarizations, which are sensitive to collective effects, and crossed polarizations, which sense the one-electron states without such corrections. For a discussion of this powerful spectroscopic technique, including application to two-dimensional systems, see, for example, Abstreiter *et al.* (1984). See also RAMAN SCATTERING. Many light-scattering experiments to explore the effects of low dimensionality have been carried out in quantum wells and other heterostructures (see HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR).

Many aspects of reduced dimensionality are exhibited by the system of small particles in a glassy matrix [see, for example, Ekimov *et al.* (1989)]. In particular, excitonic effects exhibit considerable variety, depending on the relation of the particle size to the exciton radius or to the effective Bohr radii of electrons and holes. [The effective Bohr radius of a carrier with effective mass m^* in a medium with permittivity ϵ is $4\pi\epsilon\hbar^2/m^*e^2$. The exciton radius is given by the same expression, with the reduced mass $m_e^*m_h^*/(m_e^*+m_h^*)$ taking the place of the effective mass.] For particles smaller than the effective Bohr radius of both electrons and holes, the confinement effects begin to dominate the excitonic effects.

Semiconductor lasers have been optimized through the flexibility afforded by the various growth and fabrication methods available. Many kinds of structures have been made to confine the carriers, the photons, or both so as to reduce losses and threshold currents. Quantum wells— and more recently quantum wires and quantum dots—are widely used, and benefit from the enhancement effects mentioned earlier in this section (see LASERS, SEMICONDUCTOR).

Optical methods have been used to study the effects of reduced dimensionality on diluted magnetic semiconductors (see SEMICONDUCTORS, DILUTED MAGNETIC)—for example, by tailoring the wave-function confinement in the quantum well and thereby influencing the interaction between electrons and the magnetic atoms, which in some structures are located only in the barrier and not in the quantum well.

6. ELECTRONS ON CRYOGENIC SUBSTRATES

Electrons just outside the surface of liquid helium are attracted to the surface by the very weak image potential energy, $-e^2\delta/16\pi\epsilon_0 z$, where $\delta=(\epsilon_{He}-\epsilon_0)/(\epsilon_{He}+\epsilon_0)$. Here ϵ_{He} is the permittivity of liquid helium, ϵ_0 is the permittivity of free space, and z is the distance of the electron from the surface. The dielectric constant of liquid ^{4}He is $\epsilon_{He}/\epsilon_0=1.057$, and so the potential is rather weak. In most experiments, it is enhanced by an external electric field applied normal to the liquid–vapor interface. The resulting potential well leads to two-dimensional subbands with energy separations of order 1 meV. The density of electrons is limited to values below about 10^9 cm^{-2} by a hydrodynamic instability. At such low densities, the electrons are not degenerate, i.e., their energy distribution in equilibrium is described by Maxwell–Boltzmann statistics even at temperatures in the millikelvin range.

The mobility of electrons in such a system is limited by scattering from helium atoms in the vapor above the liquid helium—an effect that can be reduced by lowering the temperature and therefore the vapor pressure of the helium—and by scattering of electrons by surface excitations in the liquid, called ripplons. At very low temperatures, mobilities as high as 10^7 cm^2/V s have been achieved. Millimeter-wave absorption measurements give sharp spectral lines in good agreement with theoretical expectations provided allowance is made for the fact that the density of the helium changes from its value in the liquid to its value in the vapor over a distance of about 0.5 nm, leading to a gradual interface rather than the abrupt interface assumed in the conventional image potential as given above.

Although this system has no known applications, it has led to several beautiful experimental results. The first experimental confirmation of the predicted dispersion law of two-dimensional plasmons (see Sec. 3.2) was obtained for electrons on liquid helium. Even more dramatic, this system gave the first experimental example of an electron crystal, a structure in which the electrons form a regular lattice. Microwave measurements by Grimes and Adams in 1979 showed a series of features that could be identified as the modes of a hexagonal electron lattice [Grimes and Adams (1980)]. The melting temperature is proportional to the square root of the electron density, consistent with a Kosterlitz–Thouless (1973) melting mechanism involving the thermal excitation of dislocations in the electron lattice. The physical parameter in such a theory is the ratio of the average potential energy of interaction between electrons to the average kinetic energy. This ratio, $\Gamma=(\pi N)^{1/2}\times e^2/4\pi\epsilon_0 k_B T$, where N is the two-dimensional electron density, is found to be about 130 at the melting temperature. For a detailed account of the theory, see Nelson and Halperin (1979).

If, at some fixed and very low temperature, the electron density could be steadily increased, the electron gas would eventually become degenerate. A melting transition is again expected, but in this case the system would be an electron liquid, with no long-range order, at densities *above* the critical density, while at densities below the transition the electrons would form an ordered lattice to minimize their potential energy. This transition was predicted by Wigner in 1934 for three-dimensional systems, and the resulting crystal is called a Wigner crystal, but has eluded experimental observation. The two-dimensional density where such a transition would be expected to occur for electrons on liquid helium is of order 10^{12} cm $^{-2}$, three orders of magnitude higher than the stability limit mentioned above. A number of people have tried to achieve such densities by using cryogenic solids instead of liquids as substrates, or by using thin films of liquid helium on a substrate with higher dielectric constant. A considerable increase in density has been achieved [see, for example, Kajita (1988)] but not enough as yet to see the Wigner crystal in this system.

A magnetic field perpendicular to the electron plane can enhance the tendency to crystallize at low densities (see, for example, Lam and Girvin, 1984). Such a system, often also called a Wigner crystal but perhaps more properly called a magnetically induced electron crystal, has been investigated primarily in GaAs-based heterostructures. Impurity effects complicate the interpretation of experiments, and definitive confirmation of the presence of a magnetically induced electron crystal remains elusive.

An introduction to the physics of electrons on liquid helium is provided, for example, by the articles of Cole (1974) and Grimes (1978). Stan and Dahm (1989) have investigated the melting of the electron crystal through measurements of the temperature dependence of electron mobility and effective mass.

7. DEVICES AND APPLICATIONS

Semiconductor devices have been central to the massive growth of applications in information storage and processing and in communications. As miniaturization of these devices continues, they are coming closer to the regime where low dimensionality begins to play an important role, particularly if the big step to low-temperature operation is taken. It is therefore likely that some of the effects of reduced dimensionality, which have primarily been of research interest, will find increasing prominence in applications in coming years.

Devices that exhibit negative resistance as a result of tunneling through barriers that contain energy levels resonant with states in the contact regions have been widely investigated and are used as high-frequency oscillators. Reduced dimensionality plays a role in the physics of such devices. See RESONANCE TUNNELING for a discussion of this fascinating field.

Some of the properties of low-dimensional structures lend themselves to new kinds of devices. In the ballistic regime, for example, electronic structures are analogous to waveguides and lend themselves to manipulation in a coherent way, to influence the nature of modes, their transmission, and their direction. Many such devices have been proposed and investigated. Electrostatic focusing of electrons has been demonstrated by Spector *et al.* (1990) and Sivan *et al.* (1990). The possibility that the motion of individual electrons through small tunnel junctions can be correlated has been discussed by Likharev (1989), and a number of related ideas have been proposed and in some cases implemented. The high impedance and limited current-carrying capacity of small devices in the regime in which these novel effects occur will limit the range of applications that is possible. Landauer (1990) has emphasized conditions that devices for general-purpose computer logic should meet, and pointed out that many proposals do not meet these conditions.

The situation is somewhat different in the case of optical applications. Many of the effects associated with reduced dimensionality, such as enhanced binding energy and oscillator strength, persist to room temperature and therefore are quite promising for immediate application. Small semiconductor lasers are an example. Quantum wells have also been proposed as versatile infrared detectors, since the resonant energies can be varied over a wide range by appropriate design of the quantum-well dimensions and composition [see, for example, Goossen *et al.* (1988) and Levine *et al.* (1990)].

Optical nonlinearities in low-dimensional structures also lend themselves to device applications, such as bistability, mixing effects, modulation of amplitude or phase, switching, and beam deflection [see, for example, the articles in the book edited by Haug (1988)].

8. MODELING

Low-dimensional structures with simple shapes lend themselves to simple theoretical models such as the particle in a box (a flat potential bounded by infinitely high, steep walls) and the parabolic potential. Analytic results for some other structures, including quantum wires near and below threshold, have also been obtained [see, for example, Davies (1988)]. More complicated structures, such as electron waveguides with bends, generally cannot be treated analytically and are increasingly being studied with numerical methods. Some of the techniques are similar to those used for modeling semiconductor devices (see SEMICONDUCTOR DEVICE MODELING) or microwave devices. The models can be grouped into static models, which find the potential profiles and energy levels for a particular structure given the distribution of impurities and the voltages on electrodes, and dynamic models, which calculate the transmittance properties linking various modes of the system when a current is flowing. Examples of static calculations are the work of Laux and Stern (1986) and Kerkhoven *et al.* (1990), who solve the Schrödinger and Poisson equations self-consistently for a rather general two-dimensional semiconductor structure. An example of a dynamic calculation is the work of Baranger (1990), who finds the transmission coefficients of various modes in a semiconductor waveguide. Computational aspects of such calculations have been considered, for example, by Lent and Kirkner (1990).

Numerical methods are also required for calculations of conductance of low-dimensional structures in the high-electric-field, non-Ohmic regime. Such calculations are similar to those used for bulk semiconductors, invoking the drift-diffusion approximation or similar approximations based on moments of the Boltzmann equation, or using Monte Carlo methods to solve the Boltzmann equation itself [see for example, Jacoboni and Lugli (1989)]. Lower dimensionality, however, makes quantum corrections to Boltzmann transport increasingly important, and more powerful methods may be required. Fully quantum-mechanical treatments have not been as widely used as treatments based on the Boltzmann equation, except for the work on localization effects noted in Sec. 1 above.

9. SUMMARY

Low-dimensional structures have provided enormous scope for invention, both in physics and in potential applications. The field is about 25 years old, and some aspects are reasonably well developed and understood. The effects of reduced dimensionality on electrical and optical properties and on thermodynamic properties can be dramatic, and this has made the field exciting and productive. Technological and materials problems in fabricating ever smaller structures often limit the rate at which advances are made, but there is good reason to expect continued progress, with the ultimate limit on size set mainly by atomic dimensions. Many applications of low-dimensional structures have been proposed, and in many cases realized, and it is likely that use of such devices will increase in the next decade.

ACKNOWLEDGMENTS

I am indebted to John and Christine Davies for helpful comments.

GLOSSARY

Degenerate System: A system of carriers, governed by Fermi–Dirac statistics, at low temperatures. Apart from thermal broadening of order $k_B T$ near the Fermi energy, where k_B is the Boltzmann constant, all states from the band edge or subband edge to the Fermi level are occupied.

Fermi Wave Vector: The radius of the region in wave-vector space occupied by electrons at low temperatures; proportional to $n^{1/d}$, where n is the carrier density and d is the dimensionality.

Inversion Layer: A confined layer of carriers near the surface of a semiconductor, with carrier type opposite to that of the bulk.

Modulation Doping: Impurity placement in a heterostructure that minimizes scattering and maximizes mobility by introducing the impurities outside the region in which carriers are confined. Also called *selective doping.*

One-Dimensional System: A system in which carriers are free to move in one direction, but are constrained in the other two space dimensions. Sometimes called a *quantum wire.*

Quantum Dot: A structure in which carriers are confined in all three space dimensions. A *zero-dimensional* system.

Subband: A continuum of states in a two-dimensional system associated with quantization in one dimension and free electron motion in two dimensions, or the analogous continuum in a one-dimensional system.

List of Works Cited

Abstreiter, G., Cardona, M., Pinczuk, A. (1984), in: M. Cardona and G. Güntherodt (Eds.), *Light Scattering in Solids,* Vol. IV (*Topics in Applied Physics,* Vol. 54), Berlin: Springer-Verlag, pp. 5–150.

Ando, T. (1976), *Phys. Rev. B* **13**, 3468–3477.

Ando, T., Fowler, A. B., Stern, F. (1982), *Rev. Mod. Phys.* **54**, 437–672.

Baranger, H. U. (1990), *Phys. Rev. B* **42**, 11 479–11 495.

Beenakker, C. W. J., van Houten, H. (1991), in: H. Ehrenreich and D. Turnbull (Eds.), *Solid State Physics,* Vol. 44, New York: Academic, pp. 1–228.

Brey, L., Johnson, N. F., Halperin, B. I. (1989), *Phys. Rev. B* **40**, 10, 647–10, 649.

Cham, K. M., Wheeler, R. G. (1980), *Phys. Rev. Lett.* **44**, 1472–1475.

Cole, M. W. (1974), *Rev. Mod. Phys.* **46**, 451–464.

Czachor, A., Holas, A., Sharma, S. R. Singwi, K. S. (1982), *Phys. Rev. B* **25**, 2144–2159.

Davies, J. H. (1988), *Semicond. Sci. Technol.* **3**, 995–1009.

Ehrenreich, H., Cohen, M. H. (1959), *Phys. Rev.* **115**, 786–790.

Ekimov, A. I., Kudryavtsev, I. A., Ivanov, M. G., Efros, Al. L., (1989), *Fiz. Tverd. Tela (Leningrad)* **31**, 192–207 [*Sov. Phys. Solid State* **31**, 1385–1393].

Fowler, A. B., Hartstein, A. (1980), *Philos. Mag. B* **42**, 949–959.

Fowler, A. B., Fang, F. F., Howard, W. E., Stiles, P. J. (1966), *Phys. Rev. Lett.* **16**, 901–903.

Friesen, W. I., Bergersen, B. (1980), *J. Phys. C* **13**, 6627–6640.

Gell, M. A., Ninno, D., Jaros, M., Wolford, D. J., Kuech, T. F., Bradley, J. A. (1987), *Phys. Rev. B* **35**, 1196–1222.

Glaser, E., Czaputa, R., McCombe, B. D., Kramer, G., Wallis, R. F. (1986), *Surf. Sci.* **170**, 386–390.

Gold, A. (1985), *Phys. Rev. Lett.* **54**, 1079–1082.

Gold, A., Dolgopolov, V. T. (1986), *Phys. Rev. B* **33**, 1076–1084.

Goodnick, S. M., Ferry, D. K., Wilmsen, C. W., Lilienthal, Z., Fathy, D., Krivanek, O. L. (1985), *Phys. Rev. B* **32**, 8171–8186.

Goossen, K. W., Lyon, S. A., Alavi, K. (1988), *Appl. Phys. Lett.* **52**, 1701–1703.

Grimes, C. C. (1978), *Surf. Sci.* **73**, 379–395.

Grimes, C. C., Adams, G. (1980), *Surf. Sci.* **98**, 1–7.

Hansen, W., Horst, M., Kotthaus, J. P., Merkt, U., Sikorski, Ch., Ploog, K. (1987), *Phys. Rev. Lett.* **58**, 2586–2589.

Haug, H. (Ed.) (1988), *Optical Nonlinearities and Instabilities in Semiconductors,* San Diego: Academic.

Jacoboni, C., Lugli, P. (1980), *The Monte Carlo Method for Semiconductor Device Simulation,* New York: Springer-Verlag.

Kajita, K. (1988), *J. Phys. Soc. Jpn.* **54**, 4092–4095.

Kamgar, A., Kneschaurek, P., Dorda, G., Koch, J. F. (1974), *Phys. Rev. Lett.* **32**, 1251–1254.

Kash, K. (1990), *J. Lumin.* **46**, 69–82.

Kawaguchi, Y., Kawaji, S. (1980), *Surf. Sci.* **98**, 211–217.

Kerkhoven, T., Galick, A. T., Ravaioli, U., Arends, J. H., Asad, Y. (1990), *J. Appl. Phys.* **68**, 3461–3469.

Kosterlitz, J. M., Thouless, D. (1973), *J. Phys. C* **6**, 1181–1203.

Lam, P. K., Girvin, S. M. (1984), *Phys. Rev. B* **30**, 473–475; **31**, 613 (1985), erratum.

Landauer, R. (1990), *Physica A* **168**, 75–87.

Laux, S. E., Stern, F. (1986), *Appl. Phys. Lett.* **49**, 91–93.

Lent, C. S., Kirkner, D. J. (1990), *J. Appl. Phys.* **67**, 6353–6359.

Levine, B. F., Bethea, C. G., Hasnain, G., Shen, V. O., Pelve, E., Abbott, R. R., Hsieh, S. J. (1990), *Appl. Phys. Lett.* **56**, 851–853.

Likharev, K. K. (1989), *IBM J. Res. Develop.* **32**, 144–158.

Many, A., Goldstein, Y., Grover, N. B. (1965), *Semiconductor Surfaces,* Amsterdam: North-Holland.

Nelson, D. R., Halperin, B. I. (1979), *Phys. Rev. B* **19**, 2457–2484.

Pfeiffer, L., West, K. W., Stormer, H. L., Baldwin, K. W. (1989), *Appl. Phys. Lett.* **55**, 1888–1890.

Price, P. J. (1981), *Ann. Phys. (N.Y.)* **133**, 217–239.

Reed, M. A., Randall, J. N., Luscombe, J. H. (1990), in: E. M. Anastassakis and J. D. Joannopoulos (Eds.), *20th International Conference on the Physics of Semiconductors*, Vol. 3, Singapore: World Scientific, pp. 2451–2454.

Reed, M. A., Randall, J. N., Luscombe, J. H. (1991), in: L. L. Chang, E. E. Mendez, and C. Tejedor (Eds.), *Resonant Tunneling in Semiconductors: Physics and Applications*, New York: Plenum, pp. 441–449.

Reeder, A. A., Mercy, J.-M., McCombe, B. D. (1988), *IEEE J. Quant. Electronics* **24**, 1690–1697.

Richter, J., Sigg, H., von Klitzing, K., Ploog, K. (1989), *Phys. Rev. B* **39**, 6268–6271.

Sakaki, H. (1980), *Jpn. J. Appl. Phys.* **19**, L735–L738.

Sikorski, Ch., Merkt, U. (1989), *Phys. Rev. Lett.* **62**, 2164–2167.

Sivan, U., Heiblum, M., Umbach, C. P., Shtrikman, H. (1990), *Phys. Rev. B* **41**, 7937–7940.

Spector, J. S., Störmer, H. L., Baldwin, K. W., Pfeiffer, L. N., West, K. W. (1990), *Appl. Phys. Lett.* **56**, 1290–1292.

Stan, M. A., Dahm, A. J. (1989), *Phys. Rev. B* **40**, 8995–9005.

Stern, F. (1967), *Phys. Rev. Lett.* **18**, 546–548.

Stern, F. (1980), *Phys. Rev. Lett.* **44**, 1469–1472.

Stern, F. (1992), in: P. Butcher, N. March, and M. Tosi (Eds.), *Physics of Low-Dimensional Semiconductor Structures*, New York: Plenum.

Stern, F., Howard, W. E. (1967), *Phys. Rev.* **163**, 816–835.

Tanatar, B. (1991), *Phys. Rev. B* **43**, 14 621–14 628.

Theis, T. N. (1980), *Surf. Sci.* **98**, 515–532.

Timp, G. (1991), in: E. Weber and M. A. Reed (Eds.), *Nanostructured Systems* [R. K. Willardson, and A. C. Beer (Eds.), *Semiconductors and Semimetals*, Vol. 35], New York: Academic, pp. 113–190.

van Wees, B. J., van Houten, H., Beenakker, C. W. J., Williamson, J. G., Kouwenhoven, L. P., van der Marel, D., Foxon, C. T. (1988), *Phys. Rev. Lett.* **60**, 848–850.

von Klitzing, K., Dorda, G., Pepper, M. (1980), *Phys. Rev. Lett.* **45**, 494–497.

Warren, A. C., Antoniadis, D. A., Smith, H. I. (1986), *Phys. Rev. Lett.* **56**, 1858–1861.

Washburn, S., Webb, R. A. (1986), *Adv. Phys.* **35**, 375–422.

Wharam, D. A., Thornton, T. J., Newbury, R., Pepper, M., Ahmed, H., Frost, J. E. F., Hasko, D. G., Peacock, D. C., Ritchie, D. A., Jones, G. A. C. (1988), *J. Phys. C* **21**, L209–L214.

Further Reading

Ando, T., Fowler, A. B., Stern, F. (1982), *Rev. Mod. Phys.* **54**, 437–672. A detailed account of the electronic properties of two-dimensional systems, particularly inversion layers, with many references.

Hansen, W., Kotthaus, J. P., Merkt, U. (1990), in: E. Weber and M. A. Reed (Eds.), *Nanostructured Systems*, [R. K. Willardson and A. C. Beer (Eds.), *Semiconductors and Semimetals*, Vol. 35], New York: Academic. A comprehensive summary of work on electrons in laterally periodic nanostructures.

Many, A., Goldstein, Y., Grover, N. B. (1965), *Semiconductor Surfaces*, Amsterdam: North-Holland. This book gives an excellent account of the physics of semiconductor surfaces and surface space-charge effects just prior to the era of quantization effects described in the present article.

Surface Science, Vols. 58, 73, 98, 113, 142, 170, 196, 229, and 263. These volumes contain the Proceedings of the biennial International Conferences on Electronic Properties of Two-Dimensional Systems (including, in recent years, zero- and one-dimensional systems). They give a running survey of the development of the field and provide an entry point into the literature.

Thouless, D. (1989), in: P. Davies (Ed.), *The New Physics*, Cambridge: University Press, pp. 209–235. A valuable overview of several aspects of physics in less than three dimensions.

Weisbuch, C., Vinter, B. (1991), *Quantum Semiconductor Structures: Fundamentals and Applications*, Boston: Academic. A good introduction to quantization effects in heterostructures and to electrical and optical devices and applications.

ELECTRON STATES, LOCALIZED

C. M. SOUKOULIS, *Ames Laboratory–USDOE and Department of Physics and Astronomy, Iowa State University, Ames, Iowa, U.S.A.*

E. N. ECONOMOU, *Research Center of Crete and Department of Physics, University of Crete, Heraklion, Crete, Greece*

Introduction 1

1. **The Universal Features of Electronic Structures in Crystalline and Disordered Systems** 2

1.1 Perfect Crystals 2

1.2 Single Localized Impurity in a Perfect Crystal 3

1.3 Imperfect Crystals............ 3

2. **Effects on Density of States** .. 4

2.1 Urbach Tails 5

3. **Classical Transport** 7

4. **Interference Effects**.......... 7

5. **Localization** 10

5.1 Explicit Results and Discussion 14

5.2 Fractal Character of the Eigenstates................... 18

6. **Mott-Hubbard Insulators: Impurity Bands**.............. 19

7. **Classical-Wave Localization** .. 20

8. **Summary**.................... 20

List of Works Cited.......... 21

Further Reading............. 22

INTRODUCTION

In a perfect elemental crystal, all atoms are identical and arranged periodically in space. This periodic arrangement constitutes perfect, long-range translational order. An immediate consequence of this perfect periodicity is that there are universal features in the electronic structure of crystals. The most important of these are energy bands separated by gaps, crystal momentum as a good quantum number, and the form of the states as plane waves modulated by the lattice periodicity. These universal features form the basis of Bloch's theory of electron states in a perfect periodic lattice. Although this basic theory of perfect crystals with noninteracting particles can only explain a few observable properties of solids, it provides the conceptual foundations for most of solid-state physics, if it is supplemented by a kinetic theory in which the electrons interact with one another and with imperfections of different types in the lattice.

Disordered materials lack the periodicity of the crystals. Accordingly, one might expect that all those universal features of crystalline materials dependent upon their translational symmetry disappear as well. However, it might be possible that there are also corresponding universal features in the electronic structure of disordered materials. If so, what are they?

The theory of noninteracting electrons in static disordered systems is well developed only in the case of low concentration of impurities. In this limit, multiple scattering from different impurities can be omitted and the problem can be reduced to that of a single impurity. Even under circumstances where some appropriate measure of the disorder cannot be considered weak, a very common situation in real materials with practical interest, we usually approximate the disordered system under study by an "effective" periodic one, a situation where Bloch's theory can be readily applied.

Anderson's classic paper "Absence of Diffusion in Certain Random Lattices" in 1958 treated the motion of an electron in a random potential, outside the framework of Bloch's theory. Anderson came to the conclusion that randomness, if strong enough, can trap the electron in a finite region of space. This is known as the *Anderson localization problem.* If the disorder is weak, the wave function

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

appears as a plane wave on a short length scale and, of course, it loses phase coherence—because of scattering by the random potential—on a long length scale. The distance over which the phase of the wave function deviates appreciably from that of the plane wave is called the mean free path. As the disorder is made stronger, additional changes in the nature of the wave function occur: amplitude fluctuations appear that become stronger and of larger extent as the randomness increases. Eventually, one (or a few) of the fluctuations may dominate and the corresponding eigenfunctions undergo a qualitative transformation, becoming *localized,* or *nonpropagating.* This means that their amplitude, $|\psi(\mathbf{r})|$, decays to zero as $|\mathbf{r}| \to \infty$ (usually by acquiring an exponential envelope). Physically, one expects states to become localized when the mean free path becomes comparable to the wavelength. The central question in the Anderson localization problem is precisely how the localized states evolve from extended states as the disorder increases. The study of Anderson localization for the last 30 years has led to new concepts such as wave-function localization, mobility edges, and fractal wave functions, and to a different way of describing materials. It also marked a significant departure from the philosophy (Bloch's theorem and perturbation theory) that had dominated solid-state physics. These new concepts and ideas have expanded into many fields beyond the electronic structure of matter. Research in the physics of disordered systems has seen a substantial growth in the last decade.

The approach taken here is to review the universal features of perfect and imperfect crystals, observing the effects of increasing disorder on the electronic structure. Then the nature of disordered materials is discussed and a simple model that illustrates the main characteristics of the electronic structure is presented. The focus is on the universal features of these models and their consequences for amorphous semiconductors. Simple arguments are given showing that continuous bands of extended states with tails of localized states associated with fluctuations within the disordered material can always be expected. The significance of the disorder-induced localization transition is discussed. A distinct energy exists, called the mobility edge, that separates regions of localized from regions of extended states. At the mobility-edge energy, the character of the wave functions changes abruptly, even though the eigenvalue spectrum is continuous. Exactly at the mobility edge, the electron states are fractal in nature. The question of the dimensionality dependence of the Anderson transition and the energy dependence of the tails of the density of states (Urbach tails) is reviewed. We present examples of the general theoretical results applied to binary alloys, quasiperiodic systems, and model systems with different types of disorder. Experimental evidence is also presented that supports the localization transition. Recently, the question of classical wave localization has received increased attention and the existence or not of gaps in classical wave systems is of central interest. In conclusion, we summarize the present state of the field and indicate the unsolved problems of immediate interest.

1. THE UNIVERSAL FEATURES OF ELECTRONIC STRUCTURES IN CRYSTALLINE AND DISORDERED SYSTEMS

The electronic structures of crystals, both perfect and imperfect, show universal features. Bloch's theorem for the crystalline system and the existence of energy bands separated by gaps are two typical examples. The basic model for disordered systems is first presented as a natural extrapolation of the modifications resulting as one moves from perfect crystals to imperfect crystals.

1.1 Perfect Crystals

We consider the motion of electrons in perfect crystals in the single-particle picture. The common characteristic of perfect crystals is their periodicity. An immediate consequence of this perfect periodicity is the Bloch-Floquet theorem stating that the one-electron wave function is a periodically modulated plane wave, i.e., it must be of the form

$$\psi_{nk}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} U_{nk}(\mathbf{r}), \tag{1}$$

where the wave vector **k** is in the first Brillouin zone (see ELECTRON STRUCTURE OF SOLIDS or Ashcroft and Mermin, 1976). $U_{nk}(\mathbf{r})$ has the periodicity of the crystal structure and n is the band index. Thus, all the wave functions in the crystal are extended: the magnitude is the same in every unit cell of the crystal. Moreover, all states possess perfect phase coherence, i.e., long-range order in the phase. Given the phase of the wave function at one point, one can determine the phase at any other point in the crystal provided that one knows the wave vector **k**. The effects of the periodicity are much more severe for the density of states (DOS); the energy spectrum consists of continuous bands of allowed energy levels that may or may not overlap. In the latter case, forbidden regions, termed gaps, where no eigenenergies exist, appear. There are sharp band edges where the DOS shows square-root singularities in its dependence on energy. Within the bands there are also singularities in the DOS corresponding to various kinds of critical points in the dependence of E on **k**. As a result of these properties, an applied dc voltage will produce an ac or zero current, depending on whether or not the Fermi energy (i.e., the highest occupied energy) is inside a band. All of these features, and particularly the extended nature of the states, their perfect phase coherence, and the sharp band edges, are universal properties of the electronic structure of perfect crystals stemming from the periodicity of the lattice.

1.2 Single Localized Impurity in a Perfect Crystal

Let us consider a crystal containing a single, local imperfection such as an impurity. The problem can then be thought of as one of scattering by the imperfection, where the Bloch states of the perfect crystal play the role of the incident waves. If the local potential variation ΔV introduced by the impurity is weak, one can use the Born approximation to obtain the solution of the scattering problem. The result is that the states remain extended and retain their phase coherence. There are only minor changes in the local DOS and, in particular, the sharp band edges remain. As ΔV increases, a resonance structure develops in the DOS near the bottom of the lowest band for ΔV attractive, or near the top of this band for ΔV repulsive. Higher bands, because they are wider and with smaller gaps among them, require, in general, larger values of $|\Delta V|$ for resonances to appear. The amplitude of the states associated with the resonance is no longer spatially uniform: it is enhanced near the local imperfection, although it still extends throughout the crystal. Finally, when the strength of the potential change exceeds a certain critical value, in $d=3$, the resonance splits off from the band as a delta function in the band gap. It can be shown that the state corresponding to this delta function is bound to and consequentially localized around the impurity (for $d=1$ and $d=2$ there are always bound states independent of the strength of the potential change). Accordingly, both attractive and repulsive localized potential changes can bind an electron in a crystal, in contrast to the ordinary scattering problem where one cannot produce bound states out of repulsive potentials. The reason is that the energy spectrum in a crystal is composed of bands that possess both lower and upper bounds. A repulsive potential, if strong enough, can push a localized state above the upper bound of the continuum in the same way that an attractive potential can push a localized state below the lower bound of the continuum. These features are summarized in Fig. 1.

1.3 Imperfect Crystals

The next step is to consider what happens if we have a finite concentration of randomly distributed impurities of equal potential strength, e.g., when a crystalline semiconductor is doped with p-type impurities, which substitutionally replace host atoms. Here the disorder is due to the spatial distribution of the scatterers and is characterized by the concentration of the impurities and by the difference of the atomiclike levels between the impurity and the host atom. The disorder will affect the DOS as well as the wave functions. In general, the wave function can be written as

$$\psi(\mathbf{r})=A(\mathbf{r})e^{i\phi(\mathbf{r})}, \qquad (2)$$

where for a periodic system the amplitude $A(\mathbf{r})$ is a periodic function of $\mathbf{r}$ and the phase $\phi(\mathbf{r})$ exhibits perfect phase coherence and is given by $\phi(\mathbf{r})=\mathbf{k}\cdot\mathbf{r}$.

For a disordered system, both the phase and the amplitude of the wave function will be affected. When the disorder is weak, it is usually assumed that the amplitude is essentially unaffected, while phase incoherence appears:

$$\langle e^{i(\phi(r)-\phi(0))}\rangle \sim \exp[-|\mathbf{r}|/l], \qquad (3)$$

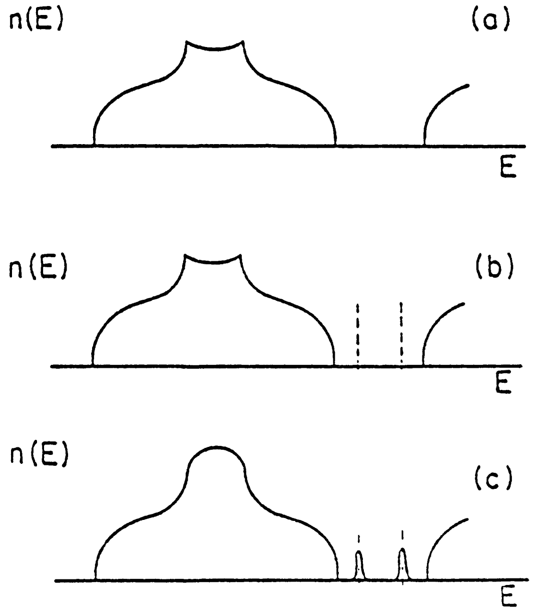

FIG. 1. Density of states *n*(*E*) as a function of energy *E* for (a) a perfect crystal, (b) a crystal containing only one localized imperfection, and (c) a crystal containing a low concentration of localized imperfections. In all three cases, there are continuous bands of energy levels separated by gaps with square-root singularities at the band edges. The square-root singularities associated with saddle points within the band in (a) and (b) are eliminated by scattering in (c). If the potential change ΔV introduced by the imperfection is strong enough in (b), localized levels split off from the bands (off the bottom for attractive ΔV or the top for repulsive), which broaden into bands in (c).

i.e., the memory of the value of the phase at a given point **r** is lost after the wave propagates a distance larger than or equal to a characteristic length called the phase-coherence length, which is of the same order of magnitude as the mean free path l. The scattering of the particle by the impurities implies a finite collision time τ. In general l depends on the eigenenergy E and the other quantum numbers characterizing the eigenfunction $\psi(r)$. Similarly τ depends on E and **k**. Through an appropriate fixed energy averaging [achieved by a Green's-function approach; see, e.g., Abrikosov *et al.* (1963), Chap. 39.2] one obtains $l(E)$ and $\tau(E)$. For weak disorder, $l(E)=v(E)\tau(E)$, where $v(E)$ is a properly averaged unperturbed velocity. If the disorder is strong, it will affect not only the phase but the amplitude as well. The wave functions exhibit strong amplitude fluctuations of variable spatial extent. It is possible, for even stronger disorder, that $A(\mathbf{r})\rightarrow 0$ as $|\mathbf{r}|\rightarrow\infty$ (usually in an exponential way). We will first discuss the effects of disorder on the DOS and then on the wave functions.

2. EFFECTS ON DENSITY OF STATES

As we have previously discussed for the case of the perfect crystal, the corresponding one-electron energies $E=E_n(\mathbf{k})$ fall into continuous bands of allowed levels usually separated by forbidden gaps. The energy $E_n(\mathbf{k})$ is a continuous function of **k** within each band and analytic except at certain high-symmetry points and lines of the Brillouin zone. There are well-defined band edges that correspond to the absolute minimum or maximum of E as a function of **k** for each band. The DOS has square-root singularities at these edges as shown in Fig. 1(a). Topological consideration shows that other critical points, saddle points, must occur within the band, also giving rise to square-root singularities in the DOS.

For the case of isolated impurities in an otherwise perfect crystal, the sharp band edges as well as the square-root singularities associated with saddle points within the band remain. The only new feature of the DOS is the introduction of impurity states in the gap that can broaden into bands if the concentration of the localized impurities becomes finite. These are called impurity bands and will be discussed later. The other effect that the finite concentration of impurities has on the DOS is the rounding off of the square-root singularities in the band. As the concentration of the localized impurities increases, one might question the entire band picture. However, there are some common experimental observations which suggest that the main features of the band picture must still be preserved. In particular, ordinary window glass, which is a disordered system, is transparent to visible light and, hence, must have a band gap of several eV, as otherwise the photons would be absorbed and the glass would appear opaque. The reason why the band picture is still valid is that short-range order is still present in disordered systems. This local order, under certain conditions, still gives rise to bands and gaps. However, the absence of long-range order has the effect of smearing the band edge into tails. We will study in Sec. 2 these tails, which are commonly called the Urbach tails. As will be discussed later, the electron states in the band tails are not the same as those within the band. The tail states are localized and, hence, electrons in those states have a very low mobility compared with those in the bands. As we will see from the discussion of the next section, the tail states are, in general, associated with local potential fluctuations.

2.1 Urbach Tails

It is well demonstrated experimentally (Cody, 1984) that the introduction of disorder causes tails to appear at the band edges in the DOS. In a systematic and thorough study, Urbach documented the existence of exponential tails in the optical adsorption of ionic crystals in 1953. Since Urbach, such exponential tails have been found in a wide range of materials, both in the optical absorption and in the DOS of individual bands (Cody, 1984). The rather general character of the exponential tails in the DOS suggests a quasiuniversal mechanism that bypasses the complexity of real materials. In the quest for uncovering such possible underlying quasiuniversality, the behavior of the DOS near the band edge can be easily understood from the following representation.

An electron moving in a disordered system can be described by a Hamiltonian of the form

$$H=H_0+H_1, \tag{4}$$

where H_0 is a nonrandom part that is assumed solvable and H_1 is random. In almost all cases, the operator $\langle H_1\rangle$, i.e., the average of H_1 over the randomness, is a constant which without loss of generality can be taken as zero. For a typical disordered system, H_0 allows us to obtain the unperturbed DOS [see Fig. 1(a)] $\rho_0(E)$ and the unperturbed Green's function $G_0(E)$; $G_0(E)$ can be expressed as an integral of $\rho_0(E')/(E-E')$. The random part H_1 in simple cases can be expressed in terms of some matrix elements ϵ_i, which are random variables possessing a probability distribution $p(\epsilon_i)$. To a first approximation, all we need to know from $p(\epsilon_i)$ is its variance w^2 and its form in the tail; in most cases, because of the many independent sources of disorder, one expects the distribution $p(\epsilon_i)$ to be Gaussian,

$$p(\epsilon_i)=\frac{1}{\sqrt{2\pi}w}\exp\left[-\frac{\epsilon_i^2}{2w^2}\right]. \tag{5}$$

Note also that lattice oscillations of not so low temperatures T produce fluctuating time-dependent local potentials with a Gaussian distribution and with an extent of atomic size. For not so low T and for fast processes, such as optical absorption, the phonon-induced fluctuations behave as static and, consequently, they give rise to another independent contribution of the form given by Eq. (5). The standard deviation (w_T) of this contribution can be obtained if the electron-phonon interaction is known; e.g., if a deformation potential coupling is assumed, i.e., if the phonon-induced electronic potential ϵ is given by

$$\epsilon=E_s\nabla\cdot\mathbf{u}, \tag{6}$$

where $\mathbf{u}$ is the lattice displacement and E_s is the so-called deformation potential, then w_T is given by

$$w_T^2=(E_s^2/B)U_L, \tag{7}$$

where B is the bulk modulus and U_L is the lattice vibration energy per unit volume due to longitudinal phonons only. For high T, $U_L\simeq k_BT/a^3$. Therefore, in the presence of both Gaussian static disorder and phonon-induced fluctuations, one obtains a local contribution of the form given by Eq. (5) with a total variance w^2 given by the sum of the static disorder variance w_S^2 and the thermal variance w_T^2,

$$w^2=w_S^2+w_T^2. \tag{8}$$

The above discussion supports the claim that the tails of the probability distribution of the local fluctuating potential exhibit, in general, a Gaussian form, which is characterized by a single parameter w^2. This parameter, together with the unperturbed DOS (and the unperturbed Green's function) near the band edges, constitute, to a first approximation, the basic inputs for calculating the main features of the electronic structure of disordered systems such as amorphous semiconductors.

The behavior of the DOS of a disordered system near the gap can be deduced from the unperturbed density of states $\rho_0(E)$ with the help of the coherent potential approximation (CPA) (Economou, 1983). The CPA is conceptually simple: the approximation consists in replacing the random matrix elements ϵ_i by appropriately chosen nonrandom complex effective matrix elements $\Sigma(E)$. The criterion that determines $\Sigma(E)$ is that the fluctuation around $\Sigma(E)$ produces no scattering on the average. This conceptually simple condition leads to a rather complicated set of equations involving $G_0(E)$, from which $\Sigma(E)$ are determined. In the weak-scattering case (w^2 small) the CPA reduces to second-order perturbation theory. It is worthwhile to note that the CPA becomes exact in the limit where the eigenstates are bound around a single, large, isolated potential fluctuation; the CPA is also very satisfactory for states extending over a large number of sites. It is only in the intermediate case where the eigenstates are trapped in a (random) cluster of a few sites that the CPA fails.

CPA (or second-order perturbation theory for small w^2) shows that the fluctuating local potential ϵ_i will produce an almost rigid shift of both the valence and the conduction bands towards the gap by amounts δE_v and δE_c, respectively, where

$$\delta E_v = |G_0(E_{v0})|w^2, \quad \delta E_c = |G_0(E_{c0})|w^2, \tag{9}$$

where $E_{v0}(E_{c0})$ is the unperturbed valence (conduction) band edge. Thus, the gap is reduced by an amount equal to $\delta E_g = \delta E_v + \delta E_c$.

In addition to the almost rigid shift of the DOS, both the valence and the conduction bands develop tails toward the gap (see Fig. 2). The tails in the DOS near the new band "edges" $E_v = E_{v0} + \delta E_v$ and $E_c = E_{c0} - \delta E_c$ are dominated by states trapped in some random clusters of atomic sites. These cluster-trapped states, under normal conditions, dominate the DOS over a very narrow range of energies. The deeper tails are usually dominated by single-site bound states. For a Gaussian probability distribution, the single-site bound states produce an exponential tail in the DOS, as observed experimentally:

$$\rho_0(E) \sim \exp[-|E|/E_0]. \tag{10}$$

Equation (10) is an immediate consequence of the fact that the binding energy E in a single potential well of depth ϵ is approximately a linear function of ϵ^2:

$$E \simeq c_1\epsilon^2 + b_2. \tag{11}$$

Equation (11) is valid over a considerable range of intermediate energies ϵ. Combining Eqs. (5) and (11), one obtains immediately Eq. (10) with $E_0 = 2c_1w^2$. This exponential behavior of the DOS could span over more than five orders of magnitude. Moreover, the calculated values of c_1 together with the estimates of w yield values of E_0 of the order of 50 meV or less, in good quantitative agreement with experiments on crystalline and amorphous semiconductors (Cody, 1984). Finally, much deeper into the tail one obtains, as one should, a gradual change of the behavior of the DOS that approaches the same functional form as the probability distribution for the disorder. This limiting behavior cannot be easily seen because it is very deep in the tail. Conditions that favor the observation of this limiting behavior are either the nearby presence of another band of similar atomic nature or the existence of large correlation lengths in the fluctuating potential responsible for the tails in the DOS.

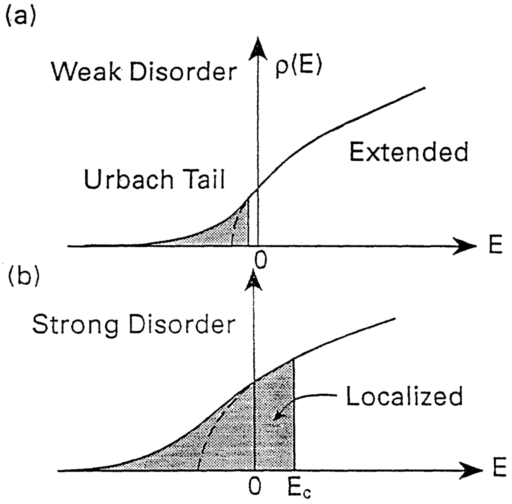

FIG. 2. (a) One-electron DOS in a random potential. For weak disorder, there is an Urbach tail of localized states below the positive-energy square-root continuum of extended states, which is shifted downward slightly by the disorder. (b) As the disorder is increased, the mobility edge eventually moves into the positive-energy regime.

The simple physical model discussed above gives clear physical insight into, and a good quantitative account of, the Urbach tail. Nevertheless, some of the assumptions of the model were untested; in particular, the correlation in the random potential was not considered explicitly. More sophisticated saddle-point (instanton) evaluation of the replica-functional-integral representation of the one-electron propagator and the Feynman path-integral formulation provide a careful treatment of DOS in a correlated Gaussian random potential (John *et al.,* 1988). The primary conclusion of this type of work is that the shape of Urbach tails in disordered materials provides a sensitive measure of the microscopic spatial autocorrelations in the random potential. The observed linear dependence of the logarithm of the tail DOS on the energy [see Eq. (10)] in a variety of materials suggests strong short-range order on the scale of the interatomic distance, but correlations that decay more rapidly than exponentially on longer length scales.

The electron–phonon (el-ph) interaction has been taken into account only in the static limit in all the evaluations described thus far. However, in slow processes such as photoluminescence, the dynamic effects of the el–ph interaction are important. These include inelastic scattering, phonon-induced hopping, and polaron formation. It is already known that disorder has a strong effect on polaron formation, and the effects of disorder are strongly affected by the dynamic aspects of the el-ph interaction. The problem is a very difficult one, and only recently have some results on the influence of lattice dynamics at nonzero temperature on band tailing been obtained (John *et al.,* 1988). The interplay between the dynamic aspects of the el-ph interaction and disorder must be addressed for slow processes. Similarly, for fast interband processes, the interplay between the electron–hole interaction and disorder must be addressed in the future.

3. CLASSICAL TRANSPORT

We know that at $T=0$, there exist neither excitations of lattice, vibrations (phonons) nor any inelastic processes. Therefore, in a perfect periodic lattice, no scattering occurs at all. This is a direct consequence of the Bloch-Floquet theorem. One is tempted to say that the collision time (mean free time) τ of the particles due to scattering of impurities, etc., is infinitely long ($\tau=\infty$) and that the mean free path $l = v_F\tau$ of the particles is also infinite ($l=\infty$); here v_F is the Fermi velocity. Actually a dc voltage will produce, in this idealized case, an ac current (or no current at all if no partially filled band exists).

In a lattice that is weakly disturbed by disorder, the situation is different. As we have discussed above, when the disorder is weak, it is usually assumed that the amplitude of the wave function is essentially unaffected, while phase incoherence appears. The scattering of the particles off the impurities implies a finite collision time τ and thereby a finite mean free path l. Then the conductivity, which is proportional to $|\int_V d^3r\psi_a^*(\mathbf{r})\hat{p}\psi_b(\mathbf{r})|^2$, becomes

$$\sigma_0 \sim \sum \left| \int_{V_l} d^3r\psi_a^*(\mathbf{r})\hat{p}\psi_b(\mathbf{r}) \right|, \tag{12}$$

where $V_l \sim l^d$, d is the dimensionality, and the summation is over all coherent regions, each of volume V_l. Mott and Davis (1979) show how one can obtain from Eq. (12) the usual formula for conductivity,

$$\sigma_0 = (e^2n/m)\tau = (e^2/p_F)nl, \tag{13}$$

where n is the electronic density, e and m are the charge and the mass of the particle, respectively, and $p_F = mv_F$ is the Fermi momentum. The quantity σ_0 is often called "Boltzmann conductivity," because Eq. (13) is also a direct consequence of the Boltzmann transport theory. In this theory, consecutive collisions of particles are assumed to be independent of each other, i.e., collisions are uncorrelated. This implies that multiple scattering of a particle from different scattering centers is not taken into account. Consequently, if such multiple scatterings become important, the basic assumption of the independence of scattering events breaks down and the validity of Eq. (13) is questionable. We will discuss in the next section what one does in this case.

4. INTERFERENCE EFFECTS

To investigate this fundamental point, we consider the diffusive behavior of a particle in a disordered system. Let the particle be in a position A at some initial time. Because of its diffusive motion, the particle will be in position B at some later time. In Feynman's formulation of quantum mechanics, the amplitude for the propagation of a particle between two points is given by the sum of classical amplitudes over all the paths connecting the two points. This is different from classical mechanics, in which one considers the most probable path only. For a free particle, the dominant trajectories have a typical width given by the de Broglie wavelength, $\lambda_{\mathrm{F}}=\hbar/mv_{\mathrm{F}}$. In the classical case ($\hbar=0$), these paths are arbitrarily sharp ($\lambda_{\mathrm{F}}=0$); in the quantum-mechanical case, however, one has that $\lambda_{\mathrm{F}} \simeq k_{\mathrm{F}}^{-1} \simeq a$, i.e., the trajectories have a finite diameter, which means that the cross section is $\lambda_{\mathrm{F}}^{d-1}$.

Now consider the propagation or transport of a particle from point A to point B. Since the transport from A to B may take place along different trajectories, there is a probability amplitude A_i connected to every path i. The total probability P of reaching point B from point A is then given by the square of the magnitude of the sum of all amplitudes, $P=|\Sigma_i A_i|^2=\Sigma_i|A_i|^2+\Sigma_{i\neq j}A_iA_j^*$. The first term describes separate noninterfering paths. This is the classical case, in which the cubes are infinitely sharp. On the other hand, the second term represents the contribution due to interference of the path amplitudes, which is a quantum-mechanical effect. In the Boltzmann theory, these interference terms were neglected. In most cases this is justified, since the trajectories have different lengths and the amplitudes A_i have different phases. On the average, this leads to destructive interference; hence, the quantum-mechanical interference is generally unimportant. In other words, the interference terms average to zero, and the addition of probabilities leads to the Boltzmann description or classical diffusion. This is no longer true, however, when one looks at closed loops—that is, when we consider the probability of returning to the starting point. In this case, starting point and ending point are identical, such that each (closed) path can be traversed in two opposite directions: forward and backward. The amplitudes for the forward and backward directions of the same path are equal; $A_{if}=A_{ib}=A_i$. For this case the quantum-mechanical probability for returning to the origin is given by $P_{\mathrm{QM}}=2\Sigma_i|A_i|^2+2\Sigma_i A_{if}A_{ib}^*=4\Sigma_i|A_i|^2$, while the classical return probability P_{cl} due to the neglect of the interference terms is given by $P_{\mathrm{cl}}=2\Sigma_i|A_i|^2$. Hence, we obtain $P_{\mathrm{QM}}=2P_{\mathrm{cl}}$. The probability for a quantum-mechanical particle to return to a starting point is seen to be twice that of a classical particle. One might argue that "quantum diffusion" is slower than classical diffusion because in the first case there exists a more effective backscattering effect. This enhanced probability of return decreases the diffusion constant and diminishes the conductivity.

To estimate the effect of the enhancement of backscattering on conductivity σ, we must consider the reduction $\delta\sigma$ of σ relative to the metallic conductivity σ_0, i.e., $\delta\sigma/\sigma_0$. Because of the expected decrease of σ, the sign of $\delta\sigma/\sigma_0$ will be negative. Furthermore, $\delta\sigma$ will be proportional to the percentage of different paths that will return to the starting point within the *inelastic* collision time, τ_{in}, after which coherence will be lost by inelastic processes. This is the same as the probability for a trajectory to intersect itself during the diffusion and for $\tau\ll\tau_{\mathrm{in}}$. In a d-dimensional tube with diameter λ_{F}, i.e., of cross section $\lambda_{\mathrm{F}}^{d-1}$, during a time interval dt, a particle moves a distance $ds=v_{\mathrm{F}}dt$; thus the corresponding volume element is given by $dV=v_{\mathrm{F}}dt\lambda_{\mathrm{F}}^{d-1}$. On the other hand, $V_{\mathrm{diff}}=(D_0t)^{d/2}$ is the d-dimensional volume into which the particle will diffuse after time t. Thus, the probability for the trajectory of the particle to intersect itself during diffusion is given by the ratio of these two volumes. We therefore have

$$P=\int_{\tau}^{\tau_{\mathrm{in}}}\frac{dV}{V_{\mathrm{diff}}}=v_{\mathrm{F}}\lambda_{\mathrm{F}}^{d-1}\int_{\tau}^{\tau_{\mathrm{in}}}\frac{dt}{(D_0t)^{d/2}}, \qquad (14)$$

where we have integrated over all times $\tau\ll\tau\ll\tau_{\mathrm{in}}$; τ is the microscopic time for single elastic collision (shorter times are not possible), while τ_{in} is the shortest inelastic time in the system. It determines the maximal time during which coherent interference of the forward-backward path amplitudes is possible. The conductivity σ_0 and the diffusion constant D_0 are related by the Einstein relation $\sigma_0=e^2N_{\mathrm{F}}D_0$, where N_{F} is the density of states at the Fermi surface. Because of $D_0\propto l$ and $\lambda_{\mathrm{F}}\propto\hbar$, one obtains, for $\tau\ll\tau_{\mathrm{in}}$, that

$$\frac{\delta\sigma}{\sigma_0}\simeq-\begin{cases}(\tau_{\mathrm{in}}/\tau)^{1/2} & d=1,\\ \ln(\tau_{\mathrm{in}}/\tau), & d=2,\\ (\tau_{\mathrm{in}}/\tau)^{-1/2}, & d=3.\end{cases} \qquad (15)$$

One can rewrite Eqs. (14) and (15) as a function of wave vectors q or lengths L instead of the different characteristic times τ and τ_{in}. Indeed, by defining $L_m=(D_0\tau)^{1/2}$, L_M

$= (D_0\tau_{\rm in})^{1/2}$, and $q=(D_0 t)^{-1/2}$, one obtains from Eq. (14)

$$P= \int_{1/L_M}^{1/L_m} \frac{q^{d-1}dq}{q^2} \sim -\frac{\delta\sigma}{\sigma_0}, \qquad (16)$$

where L_M is equal to the shortest of several upper cutoff lengths that may be present in the system, and L_m, the lower cutoff length, is believed to be very close to the mean free path l. Examples of upper cutoff lengths L_M, besides the geometrical dimension L, are the diffusion length during the inelastic relaxation time $\tau_{\rm in}$, $L_D=(D\tau_{\rm in})^{1/2}$; and the diffusion length during the time ω^{-1}, where ω is the frequency of an external ac field, $L_\omega=(D/\omega)^{1/2}$. The presence of an external magnetic field H introduces two characteristic lengths $L'_H=(D/\omega_0)^{1/2}$, where $\omega_0=eH/mc$ is the cyclotron frequency, and the cyclotron radius, $L_H=(\hbar c/eH)^{1/2}$. From Eq. (16), one finds that the contribution to the conductivity is (assuming that $L_m \ll L_M$)

$$\frac{\delta\sigma}{\sigma_0} \sim -\begin{cases} L_M/L_m, & d=1, \\ \ln(L_M/L_m), & d=2, \\ L_m/L_M, & d=3. \end{cases} \qquad (17)$$

Equation (17) is exactly equivalent to Eq. (15). For $d=1$ and $d=2$, Eq. (17) is very significant because these values show that the corrections to the conductivity do not disappear as the upper cutoff length L_M becomes very large. On the contrary, they increase with L_M until L_M becomes so large that $\sigma_0 \to 0$. Thus, a truly metallic behavior is not possible for $d \leqslant 2$. That is, of course, consistent with the statement that all eigenstates are localized for $d \leqslant 2$ (Abrahams *et al.*, 1979). However, in the presence of spin-orbit coupling, the above conclusions are modified drastically (Kramer and Schön, 1990). The nonmetallic nature of the conduction for $d \leqslant 2$ is usually masked by the fact that at high enough temperatures, L_D (and hence L_M) becomes quite short. It is clear from Eq. (17) that when $L_M \simeq L_m$, $\delta\sigma=0$.

For inelastic scattering as $T \to 0$, $\tau_{\rm in}^{-1}$ typically is proportional to T^p where the exponent p depends on which scattering, electron–phonon or electron–electron, is dominant in a given temperature range. By substituting in Eq. (15) that $\tau_{\rm in}^{-1} \sim T^p$, one obtains the following T dependence of the conductivity corrections:

$$\frac{\delta\sigma}{\sigma_0} \simeq -\begin{cases} T^{-p/2}, & d=1, \\ \ln(\hbar/\tau/k_{\rm B}T), & d=2 \\ T^{p/2}, & d=3. \end{cases} \qquad (18)$$

Thus, a ln$T$ correction is predicted for the conductivity of thin films. A decrease of the conductivity ($\delta\sigma<0$) is equivalent to an increase of the resistance ($\delta R>0$). The corresponding logarithmic T dependence of the resistance of very thin films ($d=2$) has been found in numerous experiments (Bergmann, 1984).

Finally, it should be stressed that the results obtained within the weak-localization theory and given by Eq. (18) have been derived with the assumption that a particular path and its time-reversed path are perfectly coherent. Such a "time-reversal invariance" is obeyed if there are no external magnetic fields and if the impurities are nonmagnetic.

However, the most dramatic effects of the weak-localization theory occur when a magnetic field is present, in which case the theory predicts a negative magnetoresistance, in contrast to the predictions of the Boltzmann theory. In standard transport theory, an applied external magnetic field perpendicular to the current tends to cause the carriers to drift in a direction transverse to both the current and the field, thus increasing the resistivity. Weak-localization theory, by contrast, predicts a decrease in the resistivity when a magnetic field is turned on. In the presence of a magnetic field, each path picks up an additional phase factor $\mathbf{A}\cdot\int d\mathbf{r}$, where the line integral is along the trajectory. Hence, a particular path and its time-reversed path (or conjugate) will pick up opposite phases in a magnetic field. When the paths return to the starting point, their phases will be different by $2\mathbf{A}\cdot\oint d\mathbf{r}$; that is, by twice the flux enclosed by the loop the closed path forms—and the constructive interference leading to the weak-localization effect will be destroyed. These ideas are related to the Aharonov-Bohm effect. Therefore, the resistivity will decrease when the phase difference becomes of order π. The typical area of a loop is $L_{\rm in}^2$, where $L_{\rm in} = (D_0\tau_{\rm in})^{1/2}$. This condition can be described as $HL_{\rm in}^2 \approx \phi_0$, where $\phi_0=hc/e$ is the flux quantum. Because at low T, $L_{\rm in}$ is much larger than the mean free path l, the external magnetic field H becomes very small, even as small as a few tens of Gauss, as has been seen experimentally in inversion layers of silicon MOSFET and thin metallic films (Bergmann, 1984; Lee and Ramakrishnan, 1985).

Another very important evidence of the same physics is the prediction of oscillations in the magnetoresistance of a sample shaped like a hollow cylinder. Provided that the walls of the cylinder are very thin and its perimeter is less than L_{in}, the phase difference between each path that encloses the cylinder and its conjugate path will be the same and equal to $2\phi/\phi_0$, which is the total magnetic flux through the cylinder. The resistance along the cylinder was predicted and confirmed experimentally by Sharvin and Sharvin (1981) to oscillate with increasing magnetic field with the period of the oscillation being $\phi_0/2$, which is the flux quantum corresponding to a charge $2e$.

We want to mention again that Eqs. (15)–(18) are based on the treatment of the backscattering effects, i.e., multiple scattering and the correlation between consecutive collisions. Thus they cannot be obtained within the Boltzmann transport theory. However, this interference effect argument enables us to understand the temperature dependence of the conductivity. One must rely on more sophisticated field theoretical techniques to calculate the precise prefactors in Eq. (18). Two simplified versions of such theories are the scaling theory of Anderson localization and the self-consistent theory of localization. As will be discussed later, these two approaches not only produce integrals of the type of Eqs. (14) and (16), but also give the precise prefactors in Eq. (18).

5. LOCALIZATION

The interference effects, which on a short length scale are responsible for the fluctuations in the amplitude of the eigenfunction and the resulting reduction of the conductivity, may in addition force the eigenfunction to decay to zero for very large distances as shown in Fig. 3 (Anderson, 1958). In other words, the multiple-scattering processes trap the particle in a finite region of space. This trapping, for an infinite system, implies absence of propagation and zero value for the conductivity and the diffusion coefficient. It turns out that for one-dimensional systems, all eigenstates are *localized* (which by definition means they are decaying to zero for large distances) no matter how weak the disorder is (some pathological exceptions do exist) (Mott and Twose, 1961; Economou, 1983). The decay is exponential and is characterized by the localization length L_c, which is defined through the logarithmic average of $|\psi(x)|$:

$$L_c^{-1} \equiv -\lim_{|x|\to\infty} (\langle \ln|\psi(x)|\rangle/|x|). \qquad (19)$$

Because the probability distribution of $|\psi(x)|$ has long tails, other averages of $|\psi(x)|$ would give different values of the decay length, e.g., $\langle|\psi(x)|\rangle \sim \exp(-|x|/2L_c)$. A localized eigenfunction is not necessarily centered around a single atomic site; it may be trapped around a cluster of sites, in which case the decay sets in as one moves out of it. As a result, the length L_c' beyond which the localization is visible equals $L_c + l_c$, where l_c is the linear dimension of the cluster. Usually, but not always, $l_c \ll L_c$, and $L_c' \approx L_c$.

It is widely believed, although no rigorous proof exists, that the critical dimensionality for the localization problem in the absence of magnetic scattering or external magnetic field is $d=2$. For $d<2$, any amount of disorder is enough to localize all the eigenstates (pathological exceptions exist); for $d>2$, a critical value of the disorder must be reached before an eigenstate becomes localized. For $d=2$, one distinguishes different universality classes (Kramer and Schön, 1990). The main class, characterized by time invariance and absence of spin-orbit coupling, is believed to have all eigenstates localized no matter how weak the disorder is as in the 1D case. In the presence of spin-orbit coupling or an external magnetic

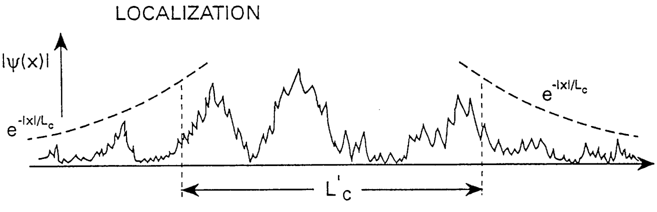

FIG. 3. Schematic plot of the amplitude of a one-dimensional eigenfunction. For large distances ($|x| \to \infty$) the eigenfunction decays on the average as $\exp(-|x|/L_c)$, where L_c is the localization length.

field, a critical value of the disorder is needed in order to localize all eigenstates, as in the 3D case.

The transition from localized (i.e., decaying) to extended (i.e., propagating) states depends on the parameters W_i ($i=1,2,\ldots$) that characterize the strength and the type of disorder and on the energy E of the eigenstate. In the $(E,\{W_i\})$ space, one can define a hypersurface separating the regions of localized from those of extended eigenstates, as shown schematically in Fig. 4 for a periodic system on which disorder is imposed. The critical energies at which the transition from localized to extended states takes place are termed *mobility edges*; their trajectory, as the $\{W_i\}$ vary, is shown schematically as a thick line in Fig. 4.

The localization problem in 1D systems has been studied extensively using various methods (Thouless, 1974; Economou, 1983). Exact analytical results and very accurate numerical data are available concerning the energy and disorder dependence of the localization length; the probability distribution of various quantities of interest, such as Green's functions; the relation between the localization length, the phase coherence length, and the DOS; etc. Of particular interest is the transmission coefficient, $|t|^2$, over a disordered segment of length L. The logarithmic average of $|t|^2$ behaves as $\exp(-2L/L_c)$ for $L/L_c \gg 1$, while other averages exhibit a different behavior (e.g., $\langle |t|^{-2}\rangle = \frac{1}{2}[1+\exp(4L/L_c)]$, while $\langle |t|^2\rangle = C_2(L_c/2L)^{3/2}\exp(-L/2L_c)$. This is due to the long tails in the probability distribution of $|t|^2$ due to very sharp resonances in $|t|^2$ vs E (where $|t|^2$ approaches almost unity). These resonances become exact localized eigenstates in the limit $L\rightarrow\infty$.

The interest in $|t|^2$ stems from its connection to the resistance R of the linear segment. Landauer (1957) has shown that

$$R=\frac{\pi\hbar}{e^2}\frac{1-|t|^2}{|t|^2}, \tag{20}$$

while Economou and Soukoulis (1981) on the basis of the Kubo formula obtained

$$R=\frac{\pi\hbar}{e^2}\frac{1}{|t|^2}. \tag{21}$$

The difference in the two expressions for R appears in the almost perfect conductor case ($L \ll L_c$), where the first expression becomes

$$R=\frac{\pi\hbar}{e^2}(e^{2L/L_c}-1)\approx\frac{\pi\hbar}{e^2}\frac{2L}{L_c}\ll\frac{\pi\hbar}{e^2}, \tag{22}$$

which is the classical result (taking into account that $L_c=4l$, where l is the mean free path), while the second tends to $\pi\hbar/e^2$. The reason for this difference is that the current-field relation is not local for ballistic transport, i.e., when $L \ll L_c$, and as a result, the resistance of the segment depends on how it is connected to the circuit and on what is happening outside the segment. A recent experiment (Timp *et al.*, 1989) in quantum lines has shown plateaus in the conductance at values equal to $e^2 n/\pi\hbar$ ($n=1,2,\ldots$) corresponding to

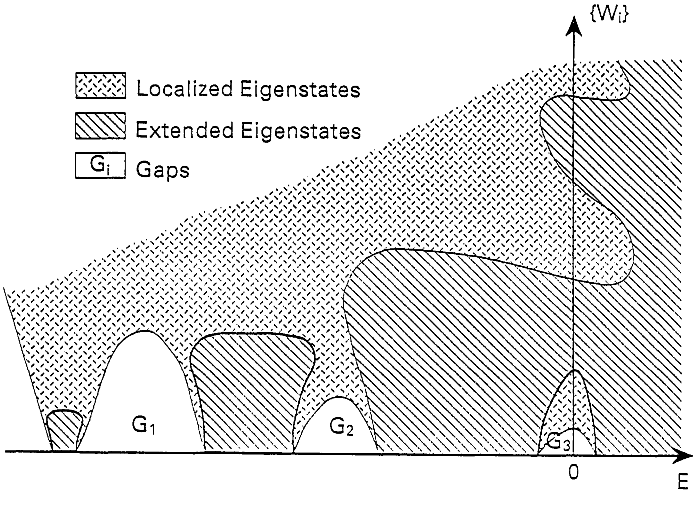

FIG. 4. The $(E,\{W_i\})$ space is separated into regions of no states (gaps), extended eigenstates, and localized eigenstates. The thin lines are the band-edge trajectories; the thick lines are the mobility-edge trajectories. The horizontal E axis ($\{W_i\}=0$) corresponds to a periodic system of alternating bands and gaps.

1,2,... one-dimensional channels reaching the saturation value $e^2/\pi\hbar$.

Quasi one-dimensional systems, of parallel-coupled $1D$ channels, have also been studied extensively because they correspond to the important and realistic cases of thin wires or quantum lines on semiconductor surfaces. Parallel-coupled $1D$ channels arranged on a plane so as to form a strip, or in space so as to form a wire of square cross section, allow reliable numerical studies of higher-dimensional systems; indeed, as the width of the strip Ma or the cross section of the wire M^2a^2 increases, the system approaches a $2D$ or $3D$ one (Kramer and Schön, 1990). It is worthwhile to mention also $1D$ quasiperiodic incommensurable systems [e.g., a simple tight-binding model with diagonal matrix elements $\epsilon_n=\epsilon_0\cos(\pi\sigma n)$ where σ is an irrational number] that exhibit a behavior intermediate between periodic and random systems (Sokoloff, 1985). In particular, the quasiperiodic $1D$ system allows either extended (when $\epsilon_0/V<2$) or localized eigenstates (when $\epsilon_0/V>2$), where V is the nearest-neighbor transfer-matrix element.

For higher-dimensional disordered systems, the localization problem is considerably more difficult and complex. Very few rigorous results exist. However, there is a plethora of theoretical approximate methods as well as some sophisticated numerical techniques.

The most systematic approach to the localization problem is based on its mapping into the sigma model (Wegner, 1976) and on subsequent use of field-theoretical techniques to establish a general scaling behavior. Through renormalization-group methods, one can obtain universal properties such as critical exponents for dimensionality d near 2. The perturbative expansion in powers of $d-2$ is then extrapolated to $d=3$.

Another approach, the so-called scaling one, starts with the observation that wave propagation in a simple tight-binding model must depend (besides the energy) on a single dimensionless parameter, namely the ratio $V/\delta\epsilon$ of the transfer-matrix element V to the level mismatch $\delta\epsilon$. The latter is proportional to the standard deviation of the probability distribution of each diagonal matrix element, ϵ_n. Next, one mentally divides the system into blocks (e.g., cubes) of ever-increasing linear dimension L. The question that arises is whether one can still define for each block two quantities $V(L)$ and $\delta\epsilon(L)$ such that the ratio $V(L)/\delta\epsilon(L)$ fully characterizes the localization properties of the system. Thouless (1974) argues that $\delta\epsilon(L)$ is essentially the level spacing, i.e., $\delta\epsilon(L)=\rho^{-1}L^{-d}$, where ρ is the DOS per unit volume, and $V(L)$ is the average change in the energy levels when the boundary condition in a block changes from periodic to antiperiodic. Furthermore, one can then show that $V(L)=\hbar/\tau'$, where τ' is the time it takes for a particle to diffuse from the center of the block to its boundary: $\tau'=(L/2)^2/D$, where D is the diffusion coefficient. Combining everything and using Einstein's relation $\sigma=2e^2\rho D$ (where σ is the conductivity), we obtain the assumed single dimensionless quantity, which fully characterizes the localization properties, as the dimensionless dc conductance $g=(\hbar/e^2)G$. The β function defined by

$$\beta\equiv\frac{d\ln g}{d\ln L}, \tag{23}$$

if positive (negative) for $L>L_0$, would imply that $G(L)\to\infty$ (zero) as $L\to\infty$, i.e., that the eigenstates are extended (localized). Abrahams *et al.* (1979) argued that β must necessarily be a function of g only (since there is a single parameter, namely g, which controls localization); furthermore, they assume that β is a monotonic function of g (an assumption that was proven wrong in the presence of spin-orbit coupling). This assumed monotonicity together with the limiting expressions $g\sim\exp(-2L/L_c)$ (when $g\to 0$) and $g\sim\sigma L^{d-2}$ (when $g\to\infty$) leads to a β vs $\ln g$ relation as shown in Fig. 5 for $d=1,2,3$. For $d=1,2$, β is always less than zero. Hence, for increasing L, g decreases, β becomes more negative, and thus we slide down toward strong localization as $L\to\infty$. In the 3D case, we distinguish two cases:

1. $g>g_c$ (g_c is roughly estimated to be 0.024). Then β is positive; g increases with increasing L, which leads to a further increase in β, and thus we move up the β curve toward extended states and classical transport behavior.
2. $g<g_c$. In this case β is negative and hence, with increasing L, we slide toward strong localization.

Another approach to the localization problem starts with the equation

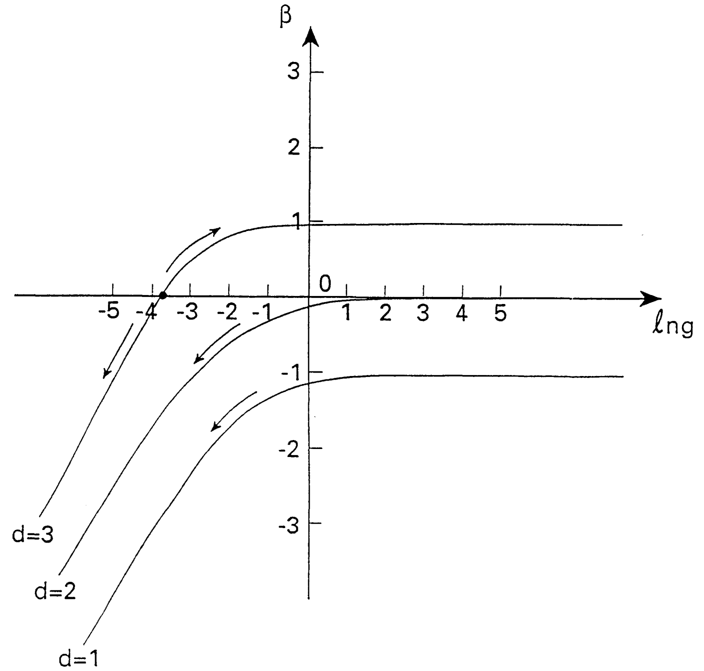

FIG. 5. The qualitative behavior of $\beta(g) = d\ln g/d\ln L$ for one, two, and three dimensions.

$$\delta\sigma \equiv \sigma - \sigma_0 = -\frac{e^2}{\pi\hbar}\frac{2}{(2\pi)^d}\int d^3q\,\frac{1}{q^2 - i\omega/D_0} \tag{24}$$

for the correction to the classical expression σ_0 for the conductivity, which, as the frequency $\omega \rightarrow 0$, reduces to Eq. (13). Equation (24) is valid for large g (near the classical regime) and is derived (Lee and Ramakrishnan, 1985) by summing up a subset of infinite terms in the perturbation expansion for σ (these terms correspond to the so-called maximally crossed diagrams). By replacing D_0 inside the integral by D, Eq. (24) can be reduced to a self-consistent one for the diffusion coefficient D. We have then

$$D(\omega) = D_0(\omega) - \frac{1}{(2\pi)^d \pi\hbar\rho}\int d^3q\,\frac{1}{q^2 - i[\omega/D(\omega)]}\,. \tag{25}$$

Note that in the localized regime $D(\omega) \rightarrow 0$ and the ratio $iD(\omega)/\omega \rightarrow L_c^2$ as $\omega \rightarrow 0$. Hence, Eq. (25) allows us to obtain the localization length (in the localized regime), the correction to the conductivity (in the extended regime), and the curve of β vs ln$g$ over the whole range.

Economou and Soukoulis (1983) and Economou *et al.* (1984) noticed that the structure of the maximally crossed diagrams and the form of Eq. (24) are mathematically identical to corresponding quantities in the simple problem of a particle being scattered or trapped by a local potential well. Thus, the localization problem (or at least the self-consistent approximation to it) can be mapped to that of a potential well (PW). The extent of the equivalent PW is proportional to the mean free path l, and its depth is proportional to $\sigma_0^{-1}l^d$. Thus the recipe for the potential-well analogy (PWA) is the following: For each particular energy and disorder, determine the mean free path l and the classical conductivity σ_0 (the best way is by use of the CPA). In terms of σ_0 and l, construct the equivalent potential well. If this PW sustains as a ground state a bound state with decay length L_c, then the eigenstates of the disordered system at this particular energy and disorder are localized with localization length L_c. If the PW has no bound states, but gives rise to a state with scattering length $|f| = \xi/\sqrt{2}$, then the eigenstates of the disordered system at this energy and disorder are extended but fluctuating up to a maximum length scale equal to ξ. (The $\sqrt{2}$ has been introduced in order that in the limit $E_r \rightarrow 0$ we have $E_r = \hbar^2/2m\xi^2$, where E_r is the resonance energy.) Furthermore, the bound or scattering eigenstate in the equivalent PW allows us to obtain the reduction of the conductivity σ_0

due to the weak or strong localization effects, i.e., to obtain $\sigma(L)$. Using the well-known results for the decay length of the lowest bound state in a PW of depth V_0 and radius R [$L_c \sim V_0^{-1}R^{-1}$ in 1D; $L_c \sim R\exp(\hbar^2/mV_0R^2)$in 2D; $L_c \sim RV_0/(V-V_0)$ for $V>V_0 = \pi^2\hbar^2/8mR^2$ in 3D], we obtain that all states are localized in 1D and 2D with $L_c \sim \sigma_0 \sim l$ and $L_c \sim l\exp[\pi^2 g_0]$, respectively. In 3D, localization appears only when $\sigma_0^{-1}l^{-1}$ exceeds a critical value, i.e., when $Sl^2 < C$ where S is the area of the surface of constant energy in k space and the universal constant C is estimated to be equal to 8.96; for isotropic systems $S=4\pi k^2$ and the condition for localization reduces to the well-known Ioffe-Regel condition

$$kl \leqslant (C/4\pi)^{1/2} \simeq 0.84. \tag{26}$$

In view of the fact that both L_c and ξ vary as $(V-V_0)^{-1}$ as $|V-V_0| \to 0$ for a PW, it follows that L_c and ξ blow up with a critical exponent equal to unity, i.e.,

$$\begin{aligned} L_c &= Al(C-Sl^2)^{-1}, \quad Sl^2 \to C^- \\ \xi &= Al(Sl^2-C)^{-1}, \quad Sl^2 \to C^+. \end{aligned} \tag{27}$$

The most reliable numerical technique (Economou, 1983; Kramer and Schön, 1990) for studying the localization problem calculates the longest localization length L_{CM} for a quasi one-dimensional system of coupled 1D channels arranged either in planar geometry (a strip of M channels for studying 2D systems) or in 3D geometry (a wire of $M \times M$ channels of square cross section in order to study 3D systems). In the numerical calculations, all lengths are measured in units of the lattice constant a; thus $Ma=M$ is the width of the strip and $M^2a^2=M^2$ is the cross-sectional area of the wire. It turns out that the quantity L_{CM}/M follows to a reasonable accuracy a single-parameter scaling law, i.e., the quantity L_{CM}/M is a function of only L_c/M or ξ/M if we are in the localized or the extended regime, respectively, where L_c and ξ are the localization and the correlation lengths, respectively, in the $M \to \infty$ limit. For two-dimensional disordered systems (with no spin-orbit coupling and no external magnetic field), there is a single branch in the curve of L_{CM}/M vs L_c/M. In most cases, for large yet numerically realizable M's, M/L_{CM} vs M becomes a straight line the slope of which determines L_c (which may be many orders of magnitude larger than M). For a three-dimensional disordered system and for $L_{\text{CM}}/M \lesssim 0.6$, there is the lower branch of the curve of L_{CM}/M vs. M, the behavior of which is similar to the one described above. For $L_{\text{CM}}/M \gtrsim 0.6$, there is an upper branch of the curve L_{CM}/M vs M corresponding to extended states, the behavior of which for large yet realizable M is $L_{\text{CM}} \approx M^2/\bar{\xi}$, where $\bar{\xi}$ is proportional to the correlation length ξ introduced earlier ($\bar{\xi} \simeq 4.82\xi$). The quantity ξ, which can be easily determined from the present numerical method, is very important because it is directly connected to the dc conductivity over the entire range of extended states up to the mobility edge (Economou *et al.*, 1985a) by the simple relation

$$\sigma = \frac{e^2}{\hbar}\left(\frac{a}{\xi} + \frac{b}{L}\right), \tag{28}$$

where L is the linear dimension of the system and a and b are numerical constants ($a \simeq 0.07 \pm 0.007$ and $b \simeq 0.05 \pm 0.03$).

It is worthwhile to point out that the probability distribution of the level spacing is different for localized states (for which there is little or no level repulsion) and for extended states (for which the level repulsion strongly reduces to zero the low end of the probability distribution). This observation allows another numerical approach to the localization problem (based on the level statistics) that has been used profitably in recent years (Kramer and Schön, 1990).

5.1 Explicit Results and Discussion

Explicit results for the strong-localization problem have been obtained mainly for simple tight-binding models (with a Hamiltonian of the form $\langle n|H|m\rangle = \epsilon_n\delta_{nm} + V[\delta_{n,m+1} + \delta_{n,m-1}]$, where $|n\rangle$ is an atomiclike orbital centered at the lattice site n) appropriate for describing isolated bands. The disorder has usually been introduced by assuming that the ϵ_n are independent (some correlated cases have also been studied) random variables, each with a probability distribution $p(\epsilon_n)$. Common choices for p are Gaussian, fully characterized by its standard deviation $\bar{\sigma}$; rectangular of total width W ($\bar{\sigma}$ in this case equals $W/\sqrt{12}$); and binary of the form $x\delta(\epsilon_n + \epsilon_0) + (1-x)\delta(\epsilon_n - \epsilon_0)$ characterized by two parameters x and ϵ_0. In Fig. 6, we show results

(Zdetsis *et al.,* 1985) for the mobility-edge trajectory, the exact band-edge trajectory (L), and the approximate band-edge trajectory (according to the CPA) in the upper half E-W plane for a simple tight-binding model with rectangular distribution for each ϵ_n.

The DOS between CPA and line L is extremely low. The solid line for the mobility-edge trajectory is based on a combination of the CPA (to obtain l and σ_0) and the PWA, while the triangular points are based on the numerical scaling method (the shaded area denotes the estimated errors of the latter). Figure 6 shows that when W exceeds a critical value W_c ($W_c \simeq 16.5$ V), the entire band consists of localized states only. This transition, where all regions of extended eigenstates in a given band disappear, is termed the Anderson transition. Anderson transition implies that the given band does not contribute at all to diffusion. In Fig. 7(a), we plot ξ and L_c vs E, i.e., as we move along the horizontal dotted line of Fig. 6; while in Fig. 7(b) we plot the same quantities versus W, i.e., as we move along the vertical line in Fig. 6. The agreement between the results of the two different methods is good (Zdetsis *et al.,* 1985). In Fig. 8, we plot the conductivity σ (together with the Born approximation result σ_{00} and the CPA result σ_0) vs W for the above-discussed model and for energy at the center of the band. Again, the agreement between the PWA and the numerical scaling method is good (Economou *et al.,* 1985a).

In Fig. 9, we show results for the binary alloy distribution (Soukoulis *et al.,* 1987; Soukoulis and Grest, 1991). For ϵ_0 larger than a critical value (which depends on x) a minority subband is split off the majority subband, creating a DOS gap (e.g., the segment ab in Fig. 9). The regions near all band edges consist of localized states; thus the mobility gap [i.e., the region where $\sigma(E)=0$ at $T=0$ K] is the segment $a'b'$, which is wider than the DOS gap. In the present case of a single band with binary distribution, increasing disorder creates first a gap which, with further increase, becomes wider and wider as shown in Fig. 9. However, usually disorder may create the opposite effect: its increase tends to make preexisting gaps narrower until they completely disappear. In this case, the mobility gaps initially become narrower and eventually either they may disappear (see, e.g., the G_3 region in Fig. 4) or they may be connected with other mobility gaps giving rise to a much wider region of localized states (e.g., the G_2 and G_1 regions in Fig. 4).

Besides the isolated bands described by tight-binding models, the simple continuum model,

$$H=p^2/2m^*+V(\mathbf{r}), \tag{29}$$

has been studied extensively. $V(\mathbf{r})$ is a Gaussian random function characterized by its autocorrelation function $\langle V(\mathbf{r})V(\mathbf{r}')\rangle-\langle V(\mathbf{r})\rangle \times\langle V(\mathbf{r})'\rangle$. The latter is assumed to possess a correlation length L_0 and a strength $\gamma=\bar{\sigma}^{\,2}L_0^d$,

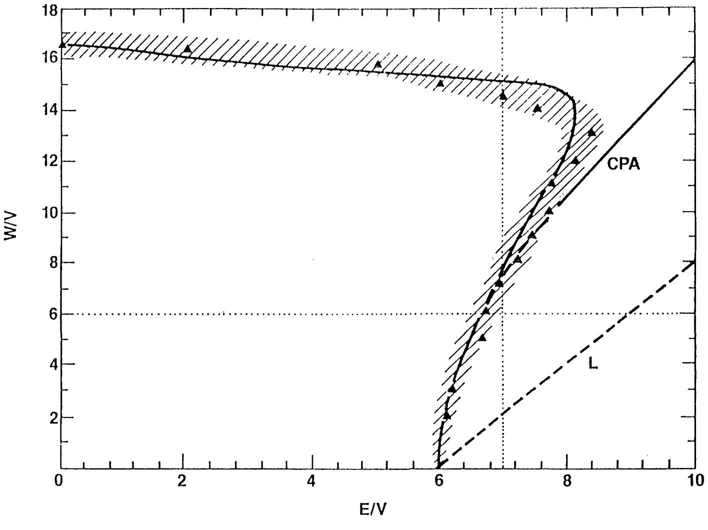

FIG. 6. Dependence of the mobility edge on the strength of the diagonal disorder W for the simple cubic lattice. Solid triangles are the numerical results. Thick solid line through the solid triangles is the prediction of the PWA with the CPA. The CPA band-edge trajectory is also indicated (solid line with the word CPA under it) together with the true (Lifshitz) band-edge trajectory (thin dashed line). The dotted straight lines indicate the independent variables for Fig. 7.

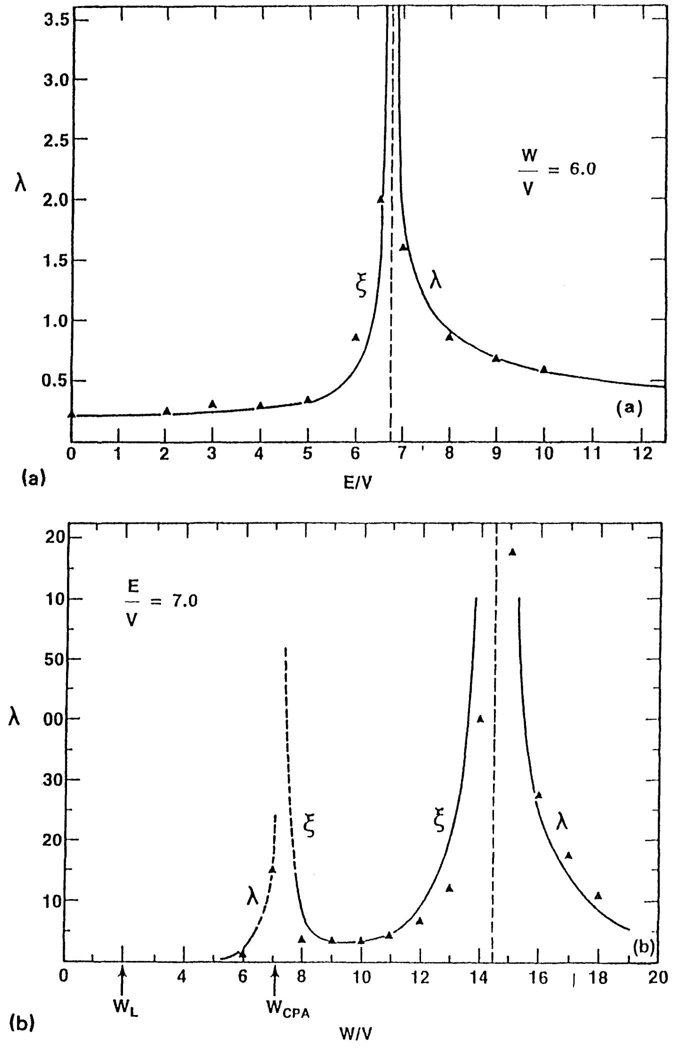

FIG. 7. (a) Localization length λ or correlation length ξ as a function of the energy E/V for a simple cubic lattice with disorder $W/V=6.0$. The solid line is the result of the PWA with the CPA, while the triangles are the numerical data. The CPA band edge is just above the critical energy, while the true band edge (Lifshitz limit) is at $E/V=9$. (b) Localization length λ or correlation length ξ versus the strength of the diagonal disorder W/V for a simple cubic lattice for energy $E/V=7.0$. The solid line is the PWA result and the triangles are the numerical data. The arrow indicates the position of the CPA band edge. The true band edge (Lifshitz limit) is lower at $W/V=2$. Near the lower critical point results have large uncertainties. All lengths are measured in units of the lattice constant a.

where $\bar{\sigma}^2$ is the variance of $V(\mathbf{r})$ and d is the dimensionality. Furthermore, most of the studies assume that the energy is close enough to the CPA band edge that the wavelength λ is much larger than L_0, and that the disorder $\bar{\sigma}$ is much smaller than $\hbar^2/|m^*|L_0^2$. These are very severe restrictions indeed, because many interesting structures appear when λ^{-1} and $\bar{\sigma}$ are comparable to or smaller than L_0^{-1} and $\hbar^2/|m^*|L_0^2$, respectively. Anyway, under the conditions $L_0 \ll \lambda$, $\lambda' \equiv (\hbar^2/|m^*|\bar{\sigma})^{1/2}$, the present model reduces essentially to the so-called white-noise model (WNM) for which several exact results exist (Thouless, 1974). It must be pointed out that the WNM (with some obvious modifications) can describe successfully a region around a CPA band edge, as long as both the energy difference from the CPA band edge and the disorder are much smaller than both the band gap and the bandwidth. Thus, for weak disorder, the region near a band edge exhibits a quasiuniversality as far as the DOS, the localization length L_c, the correlation length ξ, the mean free path l, and the conductivity σ are concerned (Economou *et al.*, 1985b).

In Fig. 10 we show results (Soukoulis *et al.*, 1989) for the mobility-edge trajectory (dashed line) obtained through an extension of the CPA-PWA method for a continuum model of nonoverlapping spheres of radius R embedded randomly within a uniform material and occupying 10% of its volume. The potential V is zero in the host material and $-\delta$ within the spheres. The unit of energy is taken as $\hbar^2/2m^*R^2$. The quasiuniversality region ends at

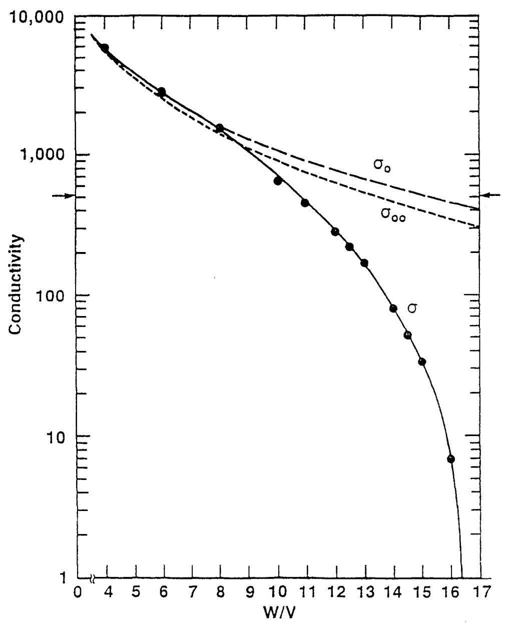

FIG. 8. Conductivity (in units of $10^{-4}\ e^2/\hbar a$), where a is the lattice spacing versus disorder for the center of the band of a simple cubic lattice. σ_{00} is the Born-approximation result, σ_0 is the CPA result; the solid line is based on Eq. (28), and the solid circles are based on the numerical results for ξ.

$\delta \approx 2$. For the interesting structure in the mobility-edge trajectory above $\delta \simeq 3$, both λ and λ' are comparable to or smaller than L_0, which in the present case equals $3.5R$. The solid line is the mobility-edge trajectory obtained through an extension of the numerical scaling method for a quasicontinuum model (consisting of $4\times4\times4$ lattice blocks of material A embedded in a host material B).

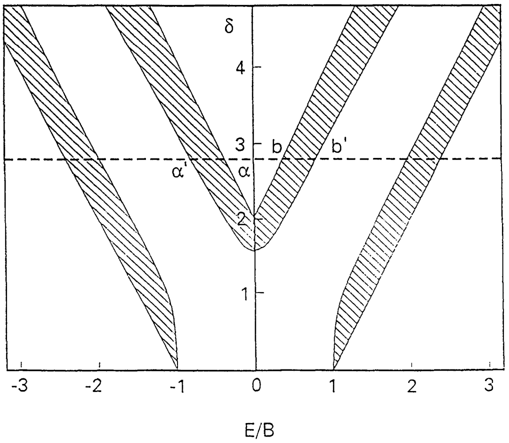

FIG. 9. Dependence of the mobility edge on the strength of the disorder $\delta = 2\epsilon_0/B$ for the binary alloy case for a particular x. B is half the bandwidth; shaded areas denote localized states.

5.2 Fractal Character of the Eigenstates

Explicit numerical results for the eigenstates in disordered d-dimensional systems ($d = 1,2,3$) show very wide fluctuations at various length scales. In a statistical sense, the eigenstates seem to be self-similar and, as a result, they can be characterized by various fractal dimensionalities (Schroeder, 1991), at least between a lower and an upper length limit. More explicitly, let a be an average interatomic distance (taken as the unit of length), N the total number of atoms in our system (so that $L \equiv N^{1/d}$ is its linear dimension), and $p_n \equiv \int |\psi(\mathbf{r})|^2 d^3r$ the probability of finding the electron in the immediate vicinity of atom n (the integral is over a volume a^3 around the atom n).

The various fractal dimensionalities D_q are defined as follows:

$$D_q = \lim_{L\to\infty} [\ln \Sigma_n p_n^q / (q-1) \ln(1/L)]. \qquad (30)$$

Of particular interest is D_2, since it is directly related with the so-called participation number $P \equiv (\Sigma_n p_n^2)^{-1}$, which is a measure of the number of sites over which the wave function is appreciable. Indeed, the above definition gives

$$P = cL^{D_2} = cN^{D_2/d}, \qquad (31)$$

where c is a constant. Equation (31) shows that P is *not* proportional to L^d, as expected for ordinary extended eigenstates, but grows with an exponent D_2 lower than the dimensionality of space. This reflects the self-similar nature of the states, which exhibit larger and larger regions where their values are almost zero. Numerical as well as field-theoretical results (Soukoulis and Economou, 1984) show that the above equation is obeyed with $D_2 \simeq 1.7 \pm 0.3$ (at the mobility edge) as long as $L_{\min} < L < L_{\max}$. The upper limiting length $L_{\max}$ equals L_c (in the localized regime) and ξ (in the extended regime), while it becomes infinite at the mobility edge. It is not clear what the lower limiting length is; possible candidates are the wavelength λ or the mean free path l (most probably λ).

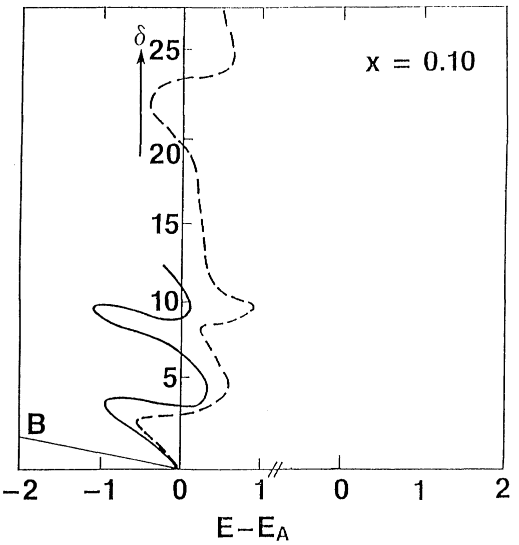

FIG. 10. Mobility-edge trajectory obtained from numerical results (solid line) and CPA results (dashed line) for the continuum binary case with $x=0.10$. The units of δ and $E-E_A$ are equal to $\hbar^2/2maR^2$ (see text).

Another interesting fractal dimension is $D_1=\lim_{L\to\infty}[\Sigma_n p_n \ln p_n/\ln(1/L)]$, which is related with the entropy measure $P_e \equiv \exp[-\Sigma p_n \ln p_n]$ of the number of sites over which the eigenfunction is appreciable,

$$P_e=c'L^{D_1}=c'N^{D_1/d}, \tag{32}$$

where c' is a constant.

It must be pointed out that the fractal exponent D_2 turns out to be the same as the so-called mass or density correlation fractal exponent D_m defined by the relation

$$\sum_{\substack{n,m \\ |\mathbf{n}-m|\leqslant r}} p_n p_m=c''r^{D_m}, \tag{33}$$

with $L_{\min} \leqslant r \leqslant L_{\max}$. The constant c'' can be obtained by setting $r=L$ (assuming, of course, that $L \leqslant L_{\max}$) in the above relation. We have then $1=c''L^{D_m}$ so that $c''=L^{-D_m}=N^{-D_m/d}$. Taking $r=L_{\min}$, we obtain $P{\sim}N^{D_m/d}$, assuming that

$$\sum_{|\mathbf{n}-\mathbf{m}|\leqslant L_{\min}} p_n p_m \sim \sum_n p_n^2.$$

Within numerical accuracy, it turns out that $D_m=D_2$. On the other hand, the various D_q are not the same, which, by definition, means that the eigenstates in disordered systems are *multifractal* objects in the range $\lambda \lesssim L \lesssim (L_c$ or $\xi)$, assuming, of course, that the parameters are such that $\lambda \ll (L_c$ or $\xi)$. An interesting physical application of the fractal exponent D_2 was proposed for the analysis of polaron formation near a mobility edge (Cohen *et al.,* 1983).

6. MOTT-HUBBARD INSULATORS: IMPURITY BANDS

The experimental investigation of the various consequences of strong localization, especially the metal–insulator transition, becomes complicated because other effects, such as thermal excitations, phonon-assisted hopping, and electron-phonon interactions, interfere. Among them probably the most important is the electron–electron (el-el) interaction because it, too, can induce a metal–insulator transition (Mott, 1974). The most pronounced effects of el-el interactions appear in narrow bands as in the case of disorder. Thus, a satisfactory theory must take into account simultaneously the effects of randomness and of el-el interactions. As in the case of disorder, the simplest system is an isolated band made up of a single atomiclike orbital per atom. Usually a further approximation is made that the Coulomb repulsion is appreciable only when the two interacting electrons are on the same site. Accordingly, an additional term is incorporated in the simple tight-binding Hamiltonian of the form

$$U \sum_i n_{i\downarrow}\, n_{i\uparrow}, \tag{34}$$

where U is the on-site Coulomb repulsion and $n_{i\uparrow}$ and $n_{i\downarrow}$ are the number operators at site i with spin up and spin down, respectively. The resulting Hamiltonian defines the so-called Hubbard model, which has been solved exactly for $d=1$. For the special, but important, case of one electron per atom (half-filled case), the Fermi level lies in the middle of a gap provided that U exceeds a critical value, U_c, of the order of half the bandwidth. Thus el-el interaction, at least in this particular case, induces a metal–insulator transition for a half-filled band. We remind the reader that even without the el-el interaction term (34), a metal–insulator transition appears when the disorder W exceeds the critical value W_c. The latter transition is termed Anderson M-I transition, while the former is known as a Mott(-Hubbard) M-I transition. The Hubbard model for a less than half-filled band has been studied extensively over the last 5 years because of its suspected connection with the magnetic and superconducting behavior of the high-T_c ceramic materials (Baskaran *et al.*, 1991).

A physical system that provides controlled conditions for studying the combined effects of both disorder and el-el interactions is the impurity band in a crystalline semiconductor. A substitutional impurity of valence $u=u_0 \pm 1$, where u_0 is the average valence of the host crystalline semiconductor, creates an impurity level just below the conduction band (if $u=u_0+1$) or just above the valence band (for $u=u_0-1$). As the concentration of impurities increases from zero, an impurity subband around the impurity level is formed. At a critical concentration, an insulator-to-metal transition takes place within the impurity subband. The phenomenon is further complicated by the nearby presence of the main band with which the impurity band eventually merges together completely. The on-site Coulomb repulsion is $U \simeq e^2/\epsilon a_B^*$, while the half-bandwidth $B \simeq U a_B^*/R = U a_B^*/n^{-1/3}$, where R^3 is the volume per impurity, $a_B^* = \hbar\epsilon/e^2 m^*$ is the effective Bohr radius, ϵ is the dielectric constant, and n is the concentration of impurities. The predicted critical concentration n_c for the I-M transition according to the Hubbard model is given by $U/B \simeq 1.2$, from which it follows that $a_B^* n_c^{-1/3}$ is constant (about 0.20). The experimental results for many combinations of impurities and host semiconductors are closely clustered around $a_B^* n_c^{-1/3} \simeq 0.26$. However, disorder, mainly due to big fluctuations in the distance between nearest neighbors, also predicts an I-M transition at $a_B^* n_c^{-1/3} \simeq 0.18$. Thus, it seems that both el-el and disorder simultaneously drive the I-M transition, pushing the critical value very close to the experimental one. By fine tuning the transition through pressure, the critical exponent for the conductivity was found to be close to 0.5, indicating that el-el interactions play a determining role at the critical point (disorder alone would give a critical exponent equal to unity).

7. CLASSICAL-WAVE LOCALIZATION

Acoustic waves in inhomogeneous (or composite) fluid materials having a constant mass density ρ_m obey a wave equation of the form

$$\nabla^2 p + [\omega^2/c(\mathbf{r})^2]p = 0, \tag{35}$$

where p is the pressure, $c^2(\mathbf{r}) = \lambda(\mathbf{r})/\rho_m$ is the sound velocity squared, and $\lambda(\mathbf{r})$ is the local bulk modulus. This equation is equivalent to Schrödinger's equation

$$\nabla^2\Psi + (2m/\hbar^2)[E - V(\mathbf{r})]\Psi = 0. \tag{36}$$

The motivation for studying classical-wave localization is twofold: on one hand, classical waves (acoustic, elastic, electromagnetic) are of interest in their own right from the point of view of basic physics as well as for possible applications. On the other hand, classical waves can easily be prepared free from self-interactions and from interactions with other degrees of freedom; their frequency can be fixed very accurately almost at will; and they are immune to complicating finite temperature effects. Thus, they present "clean" conditions for studying experimentally the localization problem. (Absorption could be a problem in some cases.) The main difficulty is that localization is inherently more difficult for classical waves than for electronic waves. To clarify this important point, we consider again a composite system consisting of nonoverlapping spheres of radius R embedded in a host material. For the electronic problem, the potential is taken as zero in the host and $-\delta$ inside each sphere. For the acoustic problem the sound velocity is c_0 in the host and c_i in each sphere ($c_0 > c_i$). By introducing the unit of energy $E_0 = \hbar^2/2mR^2$, we have the following correspondences between the acoustic and the electronic problem:

$$\begin{aligned} &R^2\omega^2/c_0^2 \leftrightarrow E/E_0, \\ &R^2\omega^2/c_i^2 \leftrightarrow (E+\delta)/E_0, \\ &c_0^2/c_i^2 \leftrightarrow (E+\delta)/E. \end{aligned} \tag{37}$$

Thus, for a constant ratio of velocities $c_0^2/c_i^2 = \mu^2 > 1$ and varying ω^2, we move in the equivalent electronic problem along the straight line

$$\delta = (\mu^2 - 1)E \tag{38}$$

in the δ vs E plane. For localized acoustic waves to appear, this straight line must cross the mobility-edge trajectory (see Fig. 10) so that the segments of $\delta=(\mu^2-1)E$ line that are in the localized regime give the frequency regions for which localization appears. As can be seen from Fig. 10, in order to have classical wave localization, it is necessary that

1. the mobility-edge trajectory cross to the right of the δ axis, which means that localized states must appear *above the highest value* of the random potential; and
2. the contrast $\mu=c_0/c_i$ must exceed a critical value.

The approximate theoretical and numerical results of Fig. 10 show that the first condition is satisfied, although under rather severe conditions (volume fraction for the spheres, x, in the range from roughly 5% to 40% and $\mu \gtrsim 1.6$).

In order to check the above approximate results and possibly to improve the localization conditions, it was suggested (John, 1988) to solve first the band structure of the problem resulting from placing the spheres in a periodic lattice. Attention was focused (Economou and Zdetsis, 1989) on finding out whether or not band gaps exist (we remind the reader that we are interested in band gaps above the maximum value of the periodic potential). If such gaps exist, it was argued that by gradually disordering the periodic system, localized states would appear in the gaps and eventually may fill the entire gap (John, 1988; Economou and Zdetsis, 1989).

The problem of the classical-wave band structure received increased attention when such periodic structures were fabricated (Yablonovitch and Gmitter, 1989) and were used to obtain experimentally the photonic band structure. It turned out (Ho *et al.*, 1990) that the vector character of the electromagnetic waves introduces significant differences from the scalar case. In addition, the band structures for acoustic (with mass density ρ_m being a periodic function as well) and elastic waves have been studied (Sigalas and Economou, 1992).

8. SUMMARY

In this article we have reviewed some of the recent developments in the theory of noninteracting electrons in disordered systems. Particular emphasis has been placed on the discussion of the universal features of the electronic density of states in disordered systems. An exponential behavior of the DOS is exhibited over many decades. The role of the disorder is, first, to shift the band edge by an amount proportional to w^2, which is given by second-order perturbation theory or by the CPA. Once the energies are measured relative to the CPA band edge, all the physical quantities of interest, including the DOS, are quasiuniversal, provided that the length and energy scaled relative to disorder-dependent units of energy and length are small. Next, the weak-interaction regime or the interference effects of the localization theory were presented. The interference between a closed Feynman path and its time-reversed path gives the unexpected result that $d=2$ is the marginal dimension for localization. Another dramatic effect of the weak localization in disordered electronic systems occurs when a magnetic field is applied, in which case the theory predicts a negative magnetoresistance. Then the scaling theory of localization, the self-consistent theory of localization, and the equivalence of localization in a random potential with the existence of a bound state in a potential well were briefly presented. The potential-well analogy (PWA) facilitates explicit calculations of the localization lengths, conductivities, mobility edges, etc., from quantities that can be obtained from mean-field theories, such as CPA. It was shown that the PWA coupled with the CPA is capable of producing results not only in qualitative but in quantitative agreement with an independent numerical approach. The advantage of the PWA-CPA approach over numerical approach is that it is analytical and gives results almost immediately, even for complicated quantities. The PWA method was applied to 3D systems by modeling the disorder using a Gaussian, a rectangular, and a binary distribution for the random site energies. The numerical results for the DOS exhibit the quasiuniversality expected from the analytical arguments (white noise model). However, the trajectory in the energy-disorder plane of the mobility edge depends on the type of the probability used. It is also discussed that the electronic wave functions in disordered systems are very complicated quantities. Explicit calculations have shown that the wave functions have strong amplitude fluctuations of various spatial extents. In particular, exactly at the mobility edge the wave functions can be characterized by fractal dimensionalities and in this sense they are fractal objects for all the length scales, while they are fractal-like for lengths up to the localization (correlation) length for localized (extended) states, respectively. Finally, a few remarks are given about the propagation of classical waves such as electromagnetic or sound waves in disordered media. What we learned about the wave nature of electrons in disordered systems should enhance our understanding of the propagation of classical waves in these systems or vice versa.

Works Cited

Abrahams, E., Anderson, P. W., Licciardello, D. C., Ramakrishnan, T. R. (1979), *Phys. Rev. Lett.* **42**, 673–676.

Abrikosov, A. A., Gorkov, L. P., Dzyaloshinski, I. E., (1963), *Methods of Quantum Field Theory in Statistical Physics*, New Jersey: Prentice Hall.

Anderson, P. W. (1958), *Phys. Rev.* **109**, 1492–1505.

Ashcroft, N. W., Mermin, N. D. (1976), *Solid State Physics*, New York: Holt, Rinehart & Winston.

Baskaran, G., Ruckenstein, A. E., Tosatti, E., Lu, Y. (Eds.) (1991), *Internat. J. Med. Phys. B* **5**, 1-417.

Bergmann, G. (1984), *Phys. Rep.* **107**, 1–58.

Cody, G. D. (1984), "Hydrogenated Amorphous Silicon," in: J. I. Pankove (Ed.), *Semiconductors and Semimetals*, Vol. 21B, New York: Academic, pp. 11–82.

Cohen, M. H., Economou, E. N., Soukoulis, C. M. (1983), *Phys. Rev. Lett.* **51**, 1202–1205.

Economou, E. N., Soukoulis, C. M. (1981), *Phys. Rev. Lett.* **46**, 618–621.

Economou, E. N. (1983), *Green's Functions in Quantum Physics*, Heidelberg: Springer-Verlag.

Economou, E. N., Soukoulis, C. M. (1983), *Phys. Rev. B* **28**, 1093–1095.

Economou, E. N., Zdetsis, A. D. (1989), *Phys. Rev. B* **40**, 1337–1340.

Economou, E. N. Soukoulis, C. M., Zdetsis, A. D. (1984), *Phys. Rev. B* **30**, 1686–1694.

Economou, E. N. Soukoulis, C. M., Zdetsis, A. D. (1985a), *Phys. Rev. B* **31**, 6483–6489.

Economou, E. N., Soukoulis, C. M., Cohen, M. H., Zdetsis, A. D. (1985b), *Phys. Rev. B* **31**, 6172–6183.

Ho, K. M., Chan, C. T., Soukoulis, C. M. (1990), *Phys. Rev. Lett.* **65**, 3152–3155.

John, S. (1988), *Comments, Cond. Matt. Phys.* **14**, 193–236.

John, S., Chou, M. Y., Cohen, M. H., Soukoulis, C. M. (1988), *Phys. Rev. B* **37**, 6963–6976.

Kramer, B., Schön, G. (Eds.) (1990), *Physica A* **167**, 1–314.

Landauer, R. (1957), *IBM J. Res. Dev.* 1, 223–231.

Lee, P. A., Ramakrishnan, T. R. (1985), *Rev. Mod. Phys.* **57**, 287–337.

Mott, N. F. (1974), *Metal-Insulator Transitions*, London: Taylor and Francis.

Mott, N. F., Davis, E. A. (1979), *Electronic Processes in Non-Crystalline Materials*, Oxford: Oxford Univ. Press.

Mott, N. F., Twose, W. D. (1961), *Adv. Phys.* **10**, 107.

Schroeder, M. (1991), *Fractal, Chaos, Power Laws*, New York: W. H. Freeman.

Sharvin, D. Yu., Sharvin, Yu. V. (1981), *JETP Lett.* **34**, 272–274.

Sigalas, M., Economou, E. N. (1992), *J. Sound Vib.* **155**, to be published.

Sokoloff, J. B. (1985), *Phys. Rep.* **126**, 189–244.

Soukoulis, C. M. Economou, E. N. (1984), *Phys. Rev. Lett.* **52**, 565–568.

Soukoulis, C. M., Economou, E. N., Grest, G. S. (1987), *Phys. Rev. B* **36**, 8649–8655.

Soukoulis, C. M., Grest, G. S. (1991), *Phys. Rev. B* **44**, 4685–4688.

Soukoulis, C. M., Economou, E. N., Grest, G. S., Cohen, M. H. (1989), *Phys. Rev. Lett.* **62**, 575–578.

Thouless, D. J. (1974), *Phys. Rep.* **13**, 93–142.

Timp, G., Behringer, R., Cunningham, J. E., Howard, R. E. (1989) *Phys. Rev. Lett.* **63**, 2268–2271.

Wegner, F. (1976), *Z. Phys. B* **25**, 327–337.

Yablonovitch, E., Gmitter, T. J., (1989), *Phys. Rev. Lett.* **63**, 1950–1953.

Zdetsis, A. D., Soukoulis, C. M., Economou, E. N., Grest, G. S. (1985), *Phys. Rev. B* **32**, 7811–7816.

Further Reading

Bergmann, G. (1984), *Phys. Rep.* **107**, 1–58. An experimental review.

Cusack, N. E. (1987), *The Physics of Structurally Disordered Matter: An Introduction*, Bristol: Adam Hilger. An introduction to the physics of disordered systems with an extensive review of the experiments.

Elliot, S. R. (1990), *Physics of Amorphous Materials*, London: Longman. A very readable account of the physics of disordered materials.

Kramer, B., Schön, G. (Eds.) (1990), *Physica* **A167**, 1–314. A recent review.

Lifshits, I. M., Gredeskul, S. A., Pastur, L. A. (1988), *Introduction to the Theory of Disordered Systems*, New York: Wiley. An introduction to the theory of disordered systems, with an emphasis on the mathematical details.

Sheng, Ping (1990), *Scattering and Localization of Classical Waves in Random Media,* Singapore: World Scientific. A recent review of the field of classical wave localization.

Vollhardt, D., Wolfle, P. (1991), "Self-consistent Theory of Anderson Localization," in W. Hanke and Ya. V. Kopaev (Eds.), *Electronic Phase Transitions,* Amsterdam: North-Holland. A detailed survey of the literature dealing with applications of the self-consistent theory of the localization of electrons.

Ziman, J. M. (1979), *Models of Disorder,* Cambridge, U.K.: Cambridge Univ. Press. A theoretical introduction to the different models of disordered systems.

ELECTRON STRUCTURE OF LIQUIDS

TAKEO FUJIWARA, *Department of Applied Physics, University of Tokyo, Tokyo, Japan*

Introduction 1
1. **Atomic and Electron Structures of Liquids** 2
1.1 Electrons in Liquids 3
1.2 Adiabatic Approximation 3
1.3 Hard-Sphere and Multipole Interactions of Rare-Gas Atoms and Molecules 3
1.4 Electron Structure Calculations in Liquids 4
1.4.1 Pseudopotential Theory 4
1.4.2 Interatomic Potentials in Simple Metals and Semiconductors 5
1.4.3 Green's-Function Theory 6
1.4.4 Effective-Medium Approximation 7
1.4.5 Recursion Method, LMTO-Recursion Method, and Supercell Method 7
2. **Electron Liquids** 8
2.1 Homogeneous Electron Liquids 9
2.2 Fermi Liquids 9
2.3 Local-Density-Functional Theory 10
2.4 Some Applications 11
2.4.1 Surfaces of Simple Metals 11
2.4.2 Two-Component Quantum Plasma 11
2.4.3 Two-Dimensional Electron Liquids: Quantum Hall Effect 11
3. **Electron Structure of Liquid Metals and Semiconductors** 12
3.1 Structure of Liquid Metals 12
3.2 Electron Spectroscopy in Liquid Metals 14
3.3 *sp*-Electron Metals 15
3.4 Group-IV Elements: Si, Ge, Sn, and Pb 16
3.5 Liquid Chalcogens: S, Se, and Te 16
3.6 Liquid Transition Metals 17
3.7 Metal–Nonmetal Transition 17
3.7.1 Mercury 17
3.7.2 Cesium 18
3.8 Electron Localization 18
4. **First-Principles Molecular Dynamical Simulation** 18
5. **Electron and Atomic Structure of Water** 20
Glossary 21
Works Cited 22
Further Reading 23

INTRODUCTION

A liquid is a random system, and quite a few investigations of it have been carried out over the past several decades. There are several different types of liquids, and they can be classified according to interparticle interactions. A liquid of rare gas atoms may be called *atomic liquid,* which is the simplest case. *Molecular liquids* are systems consisting of independent molecules, e.g., liquids of nonpolar molecules (normal liquid) and water (H_2O). Water molecules form a network of hydrogen bonds. *Liquid metals* are electron-gas systems in a positive charge background. The ionic distribution usually plays the role of a static random potential field, because the ionic motion is slower than the electron motion by a factor of a few tens or hundreds. Electrons in diluted systems are capable of undergoing a metal–nonmetal transition. Actually metal–nonmetal transitions have been observed in several liquid metals, sometimes near the critical point, as in cesium (Cs), and sometimes not, as in mercury (Hg). There are several varieties of liquid group-IV elements. Liquid germanium and silicon are metallic, and their structure is somewhere between a tetrahedrally bonded network and dense random packing of spheres. Recent optical and dc-conductivity experiments indicate that even liquid carbon is metallic. Liquid tellurium

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

and selenium are kinds of molecular liquids, forming large molecules of spiral chain structures. *Molten salts* (liquid ionic compounds) are systems of practical importance, and the interaction between ions is of the long-range Coulomb type. Electron systems in a uniform positive background are themselves liquids of exotic character (*electron liquid*).

Here, we shall discuss the interplay of electronic structures and interatomic interaction in several different liquid systems. Very recently, an important new simultaneous simulation method for the electron and atomic structures has been developed by Car and Parrinello (1985); this is called the *first-principles molecular dynamics* method (or the Car–Parrinello method).

1. ATOMIC AND ELECTRON STRUCTURES OF LIQUIDS

Liquids are characterized by a certain atomic arrangement. The atomic environment is periodic in crystals but is aperiodic in liquids. A useful accessible characterization of the local atomic environments can be given in terms of the partial pair distribution function. The partial pair distribution function $g_{\alpha\beta}(\mathbf{R})$ or the density function $\rho_{\alpha\beta}(\mathbf{R})$ between α- and β-type atoms is defined by

$$\begin{aligned}\rho_{\alpha\beta}(\mathbf{R}) &= c_\beta\, \rho_0 g_{\alpha\beta}(\mathbf{R}) \\ &= \frac{1}{c_\alpha N}\left\langle \sum_{i,j} \delta(\mathbf{R}-\mathbf{R}_{i(\alpha)}+\mathbf{R}_{j(\beta)})\right\rangle_c \\ &\quad -\delta_{\alpha\beta}\delta(\mathbf{R}), \qquad (1)\end{aligned}$$

where $\langle ... \rangle_c$ means a configurational average of atomic positions or an average over different samples, $\mathbf{R}_{i(\alpha)}$ is the position of the ith α-type atom, N is the total number of atoms, c_α is the atomic concentration of α-type atoms, and ρ_0 is the density of atoms N/V (V is the volume of the system). The number of β-type atoms in a volume element d^3R at $\mathbf{R}$, when the α-type atom is located at the origin, is given by $c_\beta \rho_0 g_{\alpha\beta}(\mathbf{R})\, d^3R$. The asymptotic value of $g_{\alpha\beta}(\mathbf{R})$ is unity in the limit of $|\mathbf{R}| \to \infty$, which implies no long-range particle–particle correlation and homogeneity of the liquid.

The Faber–Ziman (FZ)-type partial structure factor is defined by

$$a_{\alpha\beta}(\mathbf{Q}) \equiv 1+\rho_0 \int [g_{\alpha\beta}(\mathbf{R})-1]e^{i\mathbf{Q}\cdot\mathbf{R}}\, d^3R, \qquad (2)$$

which is essentially the Fourier transform of the partial pair distribution function. This is directly related to the experimental scattering intensity $I(\mathbf{Q})$, which is the square of the Fourier transform of the scattering potential and is given by

$$\begin{aligned}I(\mathbf{Q}) = N^{-1}\Bigg\{ &\sum_{\alpha,\beta} c_\alpha c_\beta a_{\alpha\beta}(\mathbf{Q}) u_\alpha(\mathbf{Q})^* u_\beta(\mathbf{Q}) \\ &+ \sum_\alpha c_\alpha |\Delta u_\alpha(\mathbf{Q})|^2 \Bigg\} \quad (|\mathbf{Q}| \neq 0), \qquad (3)\end{aligned}$$

where $u_\alpha(\mathbf{Q})$ is the Fourier transform of the scattering potential $u_\alpha(\mathbf{r}-\mathbf{R}_{i(\alpha)})$ of an α-type atom at the position $\mathbf{R}_{i(\alpha)}$ and

$$\Delta u_\alpha(\mathbf{Q}) = u_\alpha(\mathbf{Q}) - \sum_\beta c_\beta u_\beta(\mathbf{Q}).$$

The scattering vector $\mathbf{Q}$ is measured by the scattering angle θ between incident and scattered wave vectors $\mathbf{k}$ and $\mathbf{k}'$ as $|\mathbf{Q}| = 2k \sin(\theta/2)$, $k = |\mathbf{k}| = |\mathbf{k}'|$, as shown in Fig. 1. Therefore, the Fourier transform of the scattering intensity $I(\mathbf{Q})$ can provide information on atomic arrangement. For monatomic liquids, the pair distribution function $g(\mathbf{R})$ and the FZ-type structure factor $a(\mathbf{Q})$ are defined as in Eqs. (1) and (2) with no specification of α and β, and $c=1$.

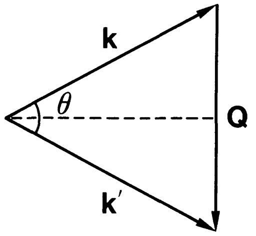

FIG. 1. The scattering vector **Q** and the scattering angle θ.

1.1 Electrons in Liquids

In periodic lattices, electronic states can be described by solutions $\psi_{\mathbf{k}}(\mathbf{r})$ of the one-electron Schrödinger equation

$$\{-(\hbar^2/2m)\nabla^2+v(\mathbf{r})\}\psi_{\mathbf{k}}(\mathbf{r})=E_{\mathbf{k}}\psi_{\mathbf{k}}(\mathbf{r}), \qquad (4)$$

where $\mathbf{r}$ is the position vector of the electron and $v(\mathbf{r})$ is a periodic potential. The Bloch theorem states that the wave functions $\psi_{\mathbf{k}}(\mathbf{r})$ are of the form

$$\psi_{\mathbf{k}}(\mathbf{r})=e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r}) \qquad (5)$$

and the functions $u_{\mathbf{k}}(\mathbf{r})$ are periodic with the periodicity of the lattice. The energy $E_{\mathbf{k}}$ is a function of the wave vector $\mathbf{k}$ or the electron momentum $\mathbf{p}=\hbar\mathbf{k}$. The relation between $E_{\mathbf{k}}$ and $\mathbf{k}$ determines the energy spectrum, the number of electron states with energies between E and $E+\Delta E$, and various physical properties. The periodic potential cannot be a scattering center of electrons, and excitations do not decay if there is no electron–electron interaction.

On the other hand, liquids are typical random aperiodic systems, and the Bloch theorem Eq. (5) does not apply. However, if the potential $v(\mathbf{r})$ is weak enough, we can start from the assumption that electrons can be described by the nearly-free-electron model and that the effect of random potential can be treated by perturbation theory. This is actually the case in alkali metals and alkaline earth metals.

In random systems, the momentum $\mathbf{p}=\hbar\mathbf{k}$ is not a good quantum number for the electronic states, but the $E_{\mathbf{k}}$–$\mathbf{k}$ relation can still be of great help in understanding electronic states. The electron transport phenomenon or the possibility of a metal–nonmetal transition is important in random systems.

1.2 Adiabatic Approximation

The electronic states are strongly influenced by the local geometry of atomic arrangement, and the electron wave function $\psi(\mathbf{r})$ is usually a function of the nuclear coordinates $(\mathbf{R}_1, \mathbf{R}_2, \ldots)$. In a crystal, the nuclear positions are fixed in a periodic lattice, and we can solve Eq. (4) in a static fixed lattice. Nuclei start to vibrate to elevated temperatures in a crystal. However, the electronic motion can be treated separately because the characteristic energies or frequencies of two different motions, electronic and nuclear, are different, by at least two orders of magnitude. The typical electron energy may be estimated as $E_0 \sim \hbar^2/ma_0^2$, where m is the electron mass and a_0 is the Bohr radius which may be taken as a typical value for the extent of the one-electron wave functions. On the other hand, if M denotes the mass of an ion, the nuclear or molecular translational (vibrational) and rotational energies are characterized by $E_t \sim (m/M)^{1/2}E_0$ and $E_r \sim \hbar^2/Ma_0^2 \sim (m/M)E_0$, respectively. Generally, $m/M \leqslant 10^{-3}$, and thus E_t and E_r are much smaller than E_0, by factors of 10^{-2} and 10^{-4}, respectively. Therefore, the effect of the nuclear kinetics on the electronic states is negligible, and the electron and nuclear motions can be treated separately. The total energy is, to a good approximation, a simple sum of energies of electron and nuclear motions. This approximation is called the *adiabatic approximation* or the *Born–Oppenheimer approximation.*

This is also the case in liquids. The characteristic time of atomic or nuclear motion in liquids is very similar to that of vibration or rotation of atoms or molecules in crystals. Electrons reach a stable state more rapidly during atomic motion. The electronic structure in a fixed aperiodic lattice can therefore be a good model of that in liquids, the aperiodic lattice being a kind of "snapshot" of changing atomic configuration in liquids.

1.3 Hard-Sphere and Multipole Interactions of Rare-Gas Atoms and Molecules

For rare-gas atoms, the electronic shell structure is closed, and the local electronic charge distribution around each nucleus is spherical. Therefore, if we neglect the effect of electronic excited states induced by interactions with nearby atoms, there are no higher multipole moments, and the interatomic potential should be of very short range. This interaction is given by the hard-sphere potential,

$$\phi(\mathbf{R}_i,\mathbf{R}_j)=\begin{cases}\infty, & \text{if } R\leqslant\sigma\,; \\ 0, & \text{if } R>\sigma,\end{cases} \qquad (6)$$

where $\mathbf{R}_i$ and $\mathbf{R}_j$ are atomic coordinates of the ith and jth atoms, $\mathbf{R}=\mathbf{R}_i-\mathbf{R}_j$, $R=|\mathbf{R}|$, and σ is the diameter of a hard sphere.

In the case of homopolar molecules without a dipole moment, such as oxygen molecules O_2, there can be a quadrupole–quadrupole interaction with a R^{-5} dependence. However, this term is usually averaged out by the rotational motion of the molecules and vanishes. On the other hand, when the electron charge distribution in one molecule is spherical, an interaction between an electron on one molecule and another electron on the other molecule has a R^{-3} dependence. Therefore, the second-order perturbation theory gives an effective intermolecular interaction with a R^{-6} dependence, which is the van der Waals interaction. The Lennard–Jones potential, an approximation for interatomic potential, is defined by the sum of the van der Waals potential plus a R^{-12}-dependent strong repulsive term and is usually written as

$$\phi(\mathbf{R}_i,\mathbf{R}_j)=\epsilon\{(\sigma/R)^{12}-2(\sigma/R)^{6}\}. \tag{7}$$

In more general cases, the interaction between two molecules at $\mathbf{R}_i$ and $\mathbf{R}_j$ can be given in series of 2^{α}-pole–2^{β}-pole interactions by

$$\phi(\mathbf{R}_i,\mathbf{R}_j)=\sum_{\alpha,\beta} C_{\alpha\beta}(\mathbf{R}/R)R^{-(\alpha+\beta+1)}, \tag{8}$$

where $C_{\alpha\beta}(\mathbf{R}/R)$ is a coefficient depending on α, β, and the configuration of two molecules relative to the intermolecular direction $\mathbf{R}/R$.

1.4 Electron Structure Calculations in Liquids

As already mentioned in Sec. 1.2, the characteristic energy of the electron motion, e.g., 1–10 eV, corresponds to about 10^4–10^5 K and is much higher than those of atomic or molecular motions, which are at most 10^3 K. Therefore, in most cases, atoms or molecules are essentially a source of a static random potential field. We now turn to the problem of surveying the electronic structures in a random array of atomic potentials. A very weak random potential field can be treated by perturbation theory. However, a nonperturbation method is required for a strong potential field.

1.4.1 Pseudopotential Theory. The one-electron Schrödinger equation is

$$\{-(\hbar^2/2m)\nabla^2+v(\mathbf{r})\}\psi(\mathbf{r})=E\psi(\mathbf{r}), \tag{9}$$

where $v(\mathbf{r})$ is a one-electron potential. (In Sec. 2.3, the one-electron Hamiltonian is derived from the many-body Hamiltonian.) We define a *pseudowave function* ϕ as

$$\phi(\mathbf{r})=\psi(\mathbf{r})-\sum_{c}\psi_c(\mathbf{r})\int d^3r'\ \psi_c^*(\mathbf{r}')\psi(\mathbf{r}'), \tag{10}$$

and a new nonlocal effective one-electron potential (*pseudopotential*) as

$$w(\mathbf{r},E)=v(\mathbf{r})+\sum_{c}(E-E_c)|\psi_c\rangle\langle\psi_c|. \tag{11}$$

Here $P_c=|\psi_c\rangle\langle\psi_c|$ is a projection operator onto a core state ψ_c, and the quantity $P_c\phi(\mathbf{r})$ is defined by

$$P_c\phi(\mathbf{r})=\psi_c(\mathbf{r})\int d^3r'\ \psi_c^*(\mathbf{r}')\phi(\mathbf{r}').$$

The summation in Eq. (11) extends over all core states ψ_c of energy E_c. Then, the Schrödinger equation for the pseudowave function ϕ is

$$\{-(\hbar^2/2m)\nabla^2+w(\mathbf{r},E)\}\phi(\mathbf{r})=E\phi(\mathbf{r}). \tag{12}$$

The pseudopotential $w(\mathbf{r},E)$ becomes shallower than $v(\mathbf{r})$ for valence orbitals of energy $E(>E_c)$, because the pseudowave function ϕ is orthogonalized to the core-electron wave functions ψ_c and is nodeless in the core region. This is the pseudopotential method (Heine, 1971). In alkali metals, alkaline earth metals, and the group-IV elements, except those in the first row in the periodic table, conduction or valence electrons are subject to weak pseudopotentials. If orbitals with a certain angular momentum quantum number l do not appear in a core-electron configuration, the pseudopotential $w(\mathbf{r},E)$ for valence orbitals with this angular momentum quantum number l cannot be appreciably shallower than $v(\mathbf{r})$. This is the reason why the free-electron model is not a good starting model for transition and rare-earth metals and the first-row elements. The spatial extent of the $4d$ wave functions of Y–Ag is wider than that of the $3d$ wave functions of Sc–Cu, and this fact means that potentials are shallower in $4d$ transition metals than in $3d$ transition metals. Conventionally, the pseudopotential has been constructed empirically so as to reproduce optical spectra of valence electrons. A method of nonempirical pseudopotentials has been proposed recently, and is called the *norm-conserving pseudopo-*

tential (Hamann *et al.*, 1979; Hamann, 1989). This method is now widely and successfully applied to solids and liquids.

1.4.2 Interatomic Potentials in Simple Metals and Semiconductors. The total energy per atom is given in the pseudopotential theory by the second-order perturbation theory as

$$\frac{E_0}{N}=\frac{1}{N}\sum_{|\mathbf{k}|<k_{\mathrm{F,spin}}} E(\mathbf{k})+\frac{1}{N}\sum_{i\neq j}\frac{1}{2}\frac{Z_iZ_je^2}{4\pi\epsilon_0R_{ij}},$$

$$\begin{aligned}E(\mathbf{k})&=\frac{\hbar^2k^2}{2m}+\sum_{\alpha}\langle\mathbf{k}|W_\alpha|\mathbf{k}\rangle+\sum_{\mathbf{q}\neq 0}\frac{\epsilon_0}{\epsilon(\mathbf{q})}\\&\quad\times\frac{\langle\mathbf{k}|W_\alpha|\mathbf{k}+\mathbf{q}\rangle\langle\mathbf{k}+\mathbf{q}|W_\beta|\mathbf{k}\rangle}{\hbar^2\{k^2-|\mathbf{k}+\mathbf{q}|^2\}/2m}\\&=E_{\mathrm{KE}}(\mathbf{k})+E_1(\mathbf{k})+E_2(\mathbf{k}),\end{aligned}\tag{13}$$

where

$$\langle\mathbf{k}|W_\alpha|\mathbf{k}+\mathbf{q}\rangle=\frac{1}{V}\int d^3r e^{-i\mathbf{k}\cdot\mathbf{r}}W_\alpha(\mathbf{r})e^{i(\mathbf{k}+\mathbf{q})\cdot\mathbf{r}},$$

$$W_\alpha(r)=\sum_i w_\alpha[\mathbf{r}-\mathbf{R}_{i(\alpha)}].$$

w_α is the pseudopotential of an α-type ion, Z_i is the valency of the ith ion, $-e$ is the electron charge, and ϵ_0 is the permittivity of vacuum. The function $\epsilon(\mathbf{q})/\epsilon_0$ is the relative dielectric constant due to electron screening. The second term in E_0/N is the ion–ion Coulomb energy. The first term of $E(\mathbf{k})$, $E_{\mathrm{KE}}(\mathbf{k})$, is the electron kinetic energy, and the second term $E_1(\mathbf{k})$ is the first-order band energy. The Fermi wave number k_{F} is defined by $k_{\mathrm{F}}=(3\pi^2n)^{1/3}$ with n the electron density. The summation on $\mathbf{k}$ can be replaced by an integral as

$$\frac{1}{N}\sum_{\mathbf{k}}E(\mathbf{k})\rightarrow\left(\frac{1}{2\pi}\right)^3\frac{V}{N}\int_{|\mathbf{k}|<k_{\mathrm{F}}}d^3kE(\mathbf{k}).$$

Using the static structure factor

$$S_\alpha(\mathbf{q})=N^{-1}\sum_{j(\alpha)}\exp(-i\mathbf{q}\cdot\mathbf{R}_j)$$

and the Fourier transform of the pseudopotential

$$w_\alpha(\mathbf{q})=\left(\frac{N}{V}\right)\int d^3r\exp(-i\mathbf{q}\cdot\mathbf{r})w_\alpha(\mathbf{r}),$$

we obtain

$$\langle\mathbf{k}|W_\alpha|\mathbf{k}+\mathbf{q}\rangle=S_\alpha(\mathbf{q})w_\alpha(\mathbf{q}).$$

By defining a function, including the spin degeneracy,

$$\begin{aligned}F_{\alpha\beta}(\mathbf{q})&=\frac{2V/N}{(2\pi)^3}\int_{|\mathbf{k}|<k_{\mathrm{F}}}\frac{\epsilon_0}{\epsilon(\mathbf{q})}\\&\quad\times\frac{w_\alpha(\mathbf{q})^*w_\beta(\mathbf{q})}{\hbar^2\{k^2-|\mathbf{k}+\mathbf{q}|^2\}/2m}d^3k,\end{aligned}\tag{14}$$

we can express the second-order band energy in terms of ion–ion pairs as

$$\begin{aligned}E_2&=\frac{1}{N}\sum_{|\mathbf{k}|<k_{\mathrm{F,spin}}}E_2(\mathbf{k})\\&=\sum_{\alpha,\beta}\sum_{\mathbf{q}\neq 0}S_\alpha(\mathbf{q})^*S_\beta(\mathbf{q})F_{\alpha\beta}(\mathbf{q})\\&=\frac{1}{2N}\sum_{i(\alpha)\neq j(\beta)}\phi_{\alpha\beta}(\mathbf{R}_{ij})\\&\quad+\sum_{\alpha}\frac{1}{N}\sum_{\mathbf{q}\neq 0}F_{\alpha\alpha}(\mathbf{q}),\end{aligned}\tag{15}$$

$$\phi_{\alpha\beta}(\mathbf{R}_{ij})=\frac{2}{N}\sum_{\mathbf{q}\neq 0}e^{i\mathbf{q}\cdot\mathbf{R}_{ij}}F_{\alpha\beta}(\mathbf{q}).$$

The first term of Eq. (15) determines the effective ion–ion two-body potential $\phi(\mathbf{R}_{ij})$. In Fig. 2, the Al–Al pair potential in face-centered cubic (fcc) aluminum is shown, which is presumably not much different from that in liquid Al. The distribution of near-neighbor

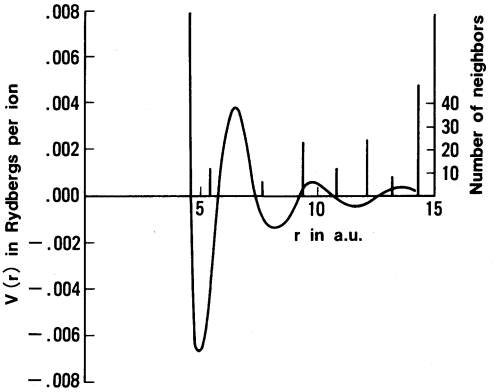

FIG. 2. The effective interaction between Al ions in the fcc structure (a continuous curve). The distribution of neighbor atoms in the fcc structure is shown as a function of distance by vertical lines. The distance is given in atomic units, i.e., 1 a.u.$=5.29\times 10^{-2}$ nm (Harrison, 1964).

atoms in fcc Al is shown as well. At large distances, the tail of the pair potential oscillates as

$$\phi(R_{ij}) \propto \frac{\cos(2k_F R_{ij})}{(2k_F R_{ij})^3}, \qquad (16)$$

and this oscillation, called the Friedel oscillation, is a consequence of oscillating screening charge density. Nearest-neighbor atoms sit at the first minimum of the pair potential and stabilize the total energy. The structure of solids and liquids may be determined so as to give the lowest total energy. Applications of the pseudopotential theory to atomic structures in liquids will be discussed in Sec. 3.1.

1.4.3 Green's-Function Theory. In a very weak potential field, we use perturbation theory and calculate the interatomic potential. In this case, plane waves can be good starting wave functions even in random systems. However, in systems with a strong potential field without translational symmetry, the plane-wave representation cannot be adopted, and a formalism of real-space representation is necessary.

We define the Green's function (or Green's-function operator) as

$$G=(E-H)^{-1}. \qquad (17)$$

Here, we write the Hamiltonian as $H=H_0+W$, where H_0 is the unperturbed part of the Hamiltonian; e.g., only the kinetic energy term. It should be noted that H and G are matrices. We assume, for simplicity, that there are no interactions between electrons. The random local potential in liquids is represented by W, i.e., $W_{ij}=w_i(\mathbf{r})\delta_{ij}$, where i denotes an individual atom. We assume, for convenience, that the zero of energy is the average potential $(1/V)\Sigma_i\int w_i(\mathbf{r})d^3r$. The spectral representation of the Green's function can be given, using the projection operator $|n\rangle\langle n|$ onto an eigenstate ψ_n of H, by

$$G=\sum_n \frac{|n\rangle\langle n|}{(E-E_n)},$$

where E_n is the eigenenergy of the state ψ_n. The Green's function can be expanded in terms of the unperturbed Green's function $G^0=(E-H_0)^{-1}$ as

$$G=G^0+G^0WG^0+G^0WG^0WG^0 +G^0WG^0WG^0WG^0+\cdots . \qquad (18)$$

The density of states $N(E)$ can be calculated in terms of Green's functions, including spin degeneracy 2, as

$$N(E)=-\frac{2}{V}\,\mathrm{Tr}\,\frac{1}{2\pi i}\lim_{\delta\to 0^+}\{G(E+i\delta) -G(E-i\delta)\}. \qquad (19)$$

The quantity $N(E)dE$ is the number of electronic states in an energy interval between E and $E+dE$. Hereafter, we assume that the system is monatomic and drop all indices α or β labeling the atom.

The self-energy can be defined in a perturbation series and calculated in a momentum representation, by the Fourier transformation, as

$$\Sigma(\mathbf{k},E)=\sum_{\mathbf{k}_1}\langle W(\mathbf{k}-\mathbf{k}_1)G^0(\mathbf{k}_1)W(\mathbf{k}_1-\mathbf{k})\rangle_c + \sum_{\mathbf{k}_1,\mathbf{k}_2}\langle W(\mathbf{k}-\mathbf{k}_1)G^0(\mathbf{k}_1)W(\mathbf{k}_1 -\mathbf{k}_2)G^0(\mathbf{k}_2)W(\mathbf{k}_2-\mathbf{k})\rangle_c+\cdots . \qquad (20)$$

The scattering potential $W(\mathbf{k})$ can be expressed as $W(\mathbf{k})=S(\mathbf{k})w(\mathbf{k})$, where

$$S(\mathbf{k})=N^{-1}\sum_j \exp(-i\mathbf{q}\cdot\mathbf{R}_j)$$

is the structure factor and $w(\mathbf{k})$ is the Fourier transform of $w_i(\mathbf{r})$. Using an experimentally observed quantity

$$a(\mathbf{k})=N\langle|S(\mathbf{k})|^2\rangle_c,$$

the second-order self-energy is derived as

$$\Sigma^{(2)}(\mathbf{k},E)=\frac{V/N}{(2\pi)^3}\int \frac{a(\mathbf{q})|w(\mathbf{q})|^2}{E-\epsilon_{\mathbf{k}+\mathbf{q}}}d^3q, \qquad (21)$$

where $\epsilon_{\mathbf{k}}$ is an eigenenergy of the unperturbed Hamiltonian H_0. If the energy E lies in a band region of $\epsilon_{\mathbf{k}}$, the integral in Eq. (21) gives both real and imaginary parts. The real part of the self-energy gives a shift of the electron energy. The imaginary part gives a broadening of the imaginary part of the Green's function or, in other words, the inverse of the imaginary part corresponds to the lifetime of an electron with a momentum $\hbar\mathbf{k}$. An electron with a momentum $\hbar\mathbf{k}$ is scattered to a state with $\hbar(\mathbf{k}+\mathbf{q})$ by the random potential field $W(\mathbf{q})$ and has a finite lifetime. The schematic energy dispersion and energy broadening of electrons in liquids are shown in Fig. 3. The shift of $E(\mathbf{k})$ has a positive sign for $k > K_p/2$ and a negative sign for $k < K_p/2$, where the wave vector K_p is a scattering vector of the strongest peak of the structure factor. Therefore, a pseudogap or dip appears in the electron density of states at an energy $E \simeq \epsilon_\mathbf{k}$ with $|\mathbf{k}| = K_p/2$. However, compared with the results obtained by a more sophisticated method, the second-order perturbation treatment Eq. (21) cannot give satisfactory results.

1.4.4 Effective-Medium Approximation. The best approximation of Eq. (18) to disordered systems on a periodic lattice is the *coherent-potential approximation* (CPA). The CPA has been widely and successfully applied to random systems on periodic lattices and extended to muffin-tin potential systems (Elliott *et al.*, 1974). We need an equivalent theory for liquids and amorphous systems. This is the *effective-medium approximation* (Roth, 1974). We consider the configurational average in the pair approximation, where the Green's function always appears together with a pair distribution function $g(\mathbf{R})$.

The single-site t matrix is defined, independent of site indices, by the equation

$$t = v + v\frac{1}{\rho_0}\int d^3R_1\, d^3R_2\, B(-\mathbf{R}_1)T(\mathbf{R}_1-\mathbf{R}_2), \tag{22}$$

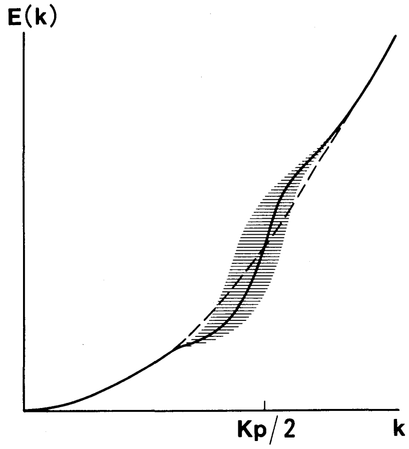

FIG. 3. Energy dispersion of electrons in liquids. The broadening of the Green's function is shown by hatching.

where v is the local scattering potential and $B(\mathbf{R})$ is the free Green's-function matrix defined by the matrix elements of $[E-(\hbar^2/2m)\Delta]^{-1}$. The single-site t matrix describes effects of repeated scattering of an electron at one site. The matrix $T(\mathbf{R})$, called the scattering path operator, describes the repeating scattering and hopping of electrons in a consistent way and should be determined by a self-consistent equation,

$$T(\mathbf{R}) = \rho_0 t\delta(\mathbf{R}) + \rho_0 t\int d^3R_1 \tilde{G}(\mathbf{R}-\mathbf{R}_1)T(\mathbf{R}_1). \tag{23}$$

Here, ρ_0 is the density of ions. The Green's function $\tilde{G}$ in the effective-medium approximation should obey the equation

$$\begin{aligned}\tilde{G}(\mathbf{R}-\mathbf{R})' &= B(\mathbf{R}-\mathbf{R}')g(\mathbf{R}-\mathbf{R}') + h(\mathbf{R}-\mathbf{R}')\\ &\times \int d^3R_1\, d^3R_2\, \tilde{G}(\mathbf{R}-\mathbf{R}_1)\\ &\times T(\mathbf{R}_1-\mathbf{R}_2)\tilde{G}(\mathbf{R}_2-\mathbf{R}'),\end{aligned} \tag{24}$$

where $h(\mathbf{R}) = g(\mathbf{R}) - 1$. These coupled equations should be solved self-consistently. The effective-medium approximation is the best approximation which consistently includes all possible single-site contributions. It has been extended to the muffin-tin potential systems and applied to liquid noble and transition metals. The electronic density of states in liquid Ni was calculated by Asano and Yonezawa (1980). The results is very close to that of solid fcc Ni, and shows that the short-range order persists in a liquid metal. (See Sec. 3.6.)

1.4.5 Recursion Method, LMTO-Recursion Method, and Supercell Method. The Green's function can be investigated by moment analysis of the spectral function, especially in cases in which the electron transfer is dominated by short-range hoppings. The *recursion method* (Haydock, 1980) is very practical for calculation of Green's functions. This method is closely related to the moment expansion method of the spectrum and the Lanczos method for matrix tridiagonalization. Let us try to calculate the diagonal element G_{00} of the Green's function of a tight-binding model, where the subscript 00 denotes the diagonal element with respect to an arbitrary state ϕ_0. We generate the whole set $\{\phi_n\}$ by the recursion relations, starting from ϕ_0,

$$b_1\phi_1 = H\phi_0 - a_0\phi_0$$
$$b_{n+1}\phi_{n+1} = H\phi_n - a_n\phi_n - b_n\phi_{n-1}. \qquad (25)$$

The coefficients a_n and b_n are determined so as to orthogonalize $H\phi_n$ (or ϕ_{n+1}) to the preceding functions ϕ_{n-1} and ϕ_n, and b_{n+1} is a coefficient which normalizes ϕ_{n+1} to unity. The resulting $\{\phi_0, \phi_1, \ldots, \phi_n, \ldots\}$ is an orthonormalized set. The matrix elements of the Hamiltonian are given by

$$\langle\phi_n|H|\phi_n\rangle = a_n,$$
$$\langle\phi_{n-1}|H|\phi_n\rangle = \langle\phi_n|H|\phi_{n-1}\rangle = b_n. \qquad (26)$$

No other states such as ϕ_{n-2} or ϕ_{n-3} appear in $H\phi_n$, and the Hamiltonian has been tridiagonalized on an orthonormalized basis set $\{\phi_0, \phi_1, \phi_2, \ldots\}$. Therefore, the diagonal element of the Green's function can be written as a continued fraction:

$$G_{00} = [(E-H)^{-1}]_{00} = \cfrac{1}{E - a_0 + \cfrac{|b_1|^2}{E - a_1 + \cfrac{|b_2|^2}{E - a_2 + \cfrac{|b_3|^2}{E - a_3 + \cdots}}}}. \qquad (27)$$

This method does not require translational periodicity. The recursion method was first applied to the tight-binding models.

The *linear muffin-tin orbital* (LMTO) method is one of the most efficient methods for the electronic band structure calculation, which can give high accuracy for energy eigenvalues and wave functions with minimal bases, i.e., one *s*-type, three *p*-type, and five *d*-type muffin-tin orbitals per atom (Andersen, 1975). The muffin-tin orbital is similar to an atomic orbital and is a solution to the Schrödinger equation in an atomic sphere at each atomic site and to the Laplace equation in an outer region where the potential is flat and the electron kinetic energy is small. By choosing localized envelope functions of each muffin-tin orbital so as to make them continuous at the boundary between two regions, we can reduce the effective Hamiltonian to a tight-binding form which we call the *first-principles tight-binding* method (Andersen and Jepsen, 1984). This method can be combined with the recursion method to give the *LMTO-recursion method*, which has been applied to amorphous and liquid systems (Fujiwara, 1984, Nowak *et al.*, 1991). Eigenvalues and eigenfunctions are very reliable even in liquid structures as well as in solids, but a structural model should be constructed by other simulation methods.

Recently, computational facilities have made great progress in terms of both computational speed and memory size. Hence, direct calculation of electronic structures in liquids can be achieved. For example, an efficient band-structure method can be applied to a system with a large unit cell containing several tens or hundreds of atoms. If we construct a cell with liquid structures of hundreds of atoms and impose a periodic boundary condition, the calculating procedure is just the same as that for crystalline solids. This is called the *supercell method* and has been applied to several liquid metals (Hafner and Jank, 1990, Hafner and Payne, 1990).

2. ELECTRON LIQUIDS

Electrons in metals form a highly degenerate liquid, subject to the Fermi–Dirac statistics, in a positive-charge background of ions. The effects of many-body correlation are quite important. The ground-state energy per electron is given in Rydberg units (1 Ry = 13.6058 eV = 2.1799×10^{-18} J), including the spin degeneracy 2, by

$$\frac{E_0}{N_e} = \frac{3}{5}\left(\frac{9\pi}{4}\right)^{2/3} \frac{1}{(r_s/a_0)^2} - \frac{3}{2}\left(\frac{3}{2\pi}\right)^{2/3} \frac{1}{r_s/a_0} + \epsilon_c, \quad \text{in Ry.} \qquad (28)$$

The parameter r_s is a mean interelectronic spacing defined by $r_s = (3/4\pi)^{1/3} n^{-1/3}$ with the electron density n, N_e is the total electron number, and a_0 is the Bohr radius (5.29×10^{-2} nm). Values of r_s/a_0 lie in normal metals between 2 and 6. The first and the second terms of Eq. (28) are the kinetic and exchange energies, respectively. The electron–electron Coulomb energy does not appear here because it exactly cancels the ion–ion and electron–ion interaction energies in a uniform system. The term ϵ_c is the remaining interaction energy, called a correlation energy, and it is a function of r_s as well. We can imagine that in the low-density region the electron liquid may crystallize, because stabilization by the exchange energy [the second term of Eq. (28)] overcomes the kinetic energy. This crystallization of the electron liquid is called *Wigner crystallization.* It is supposed that the first-order liquid–crystal phase transition occurs at about $r_s/a_0=14$–15. The correlation energy is quite important for understanding the ground-state properties in condensed phases.

2.1 Homogeneous Electron Liquids

In the interacting electron system, the N_e-electron Hamiltonian is the sum of the noninteracting part H_0 and the interaction part $H'(H=H_0+H')$. We can calculate the total energy of the ground state, using the adiabatic theorem,

$$\frac{E_0}{N_e}=\frac{E_0^{(0)}}{N_e}-\sum_{\mathbf{k}}\left\{\hbar\int_0^1\frac{d\lambda}{\lambda}\int_0^\infty\frac{d\omega}{2\pi}\,\mathrm{Im}\,\frac{\epsilon_0}{\epsilon_\lambda(\mathbf{k},\omega)}+n\frac{e^2}{4\pi\epsilon_0}\frac{2\pi}{k^2}\right\},\tag{29}$$

where $\epsilon_\lambda(\mathbf{k},\omega)/\epsilon_0$ is the dielectric function of the system $H_0+\lambda H'$. This equation gives the total energy of electrons in terms of the dielectric function.

The dielectric function $\epsilon(\mathbf{k},\omega)/\epsilon_0$ is related to $\chi(\mathbf{k},\omega)$, the density–density response function of electrons, by

$$\frac{\epsilon_0}{\epsilon(\mathbf{k},\omega)}=1+\frac{4\pi e^2}{4\pi\epsilon_0 k^2}\chi(\mathbf{k},\omega).\tag{30}$$

The density–density response function $\chi(\mathbf{k},\omega)$ can be precisely expressed as

$$\chi(\mathbf{k},\omega)=\sum_n\left|\left\langle n\left|\sum_i e^{i\mathbf{k}\cdot\mathbf{r}_i}\right|0\right\rangle\right|^2\times\left\{\frac{1}{\omega-\omega_{n0}+i\eta}-\frac{1}{\omega+\omega_{n0}+i\eta}\right\},\tag{31}$$

where $\omega_{n0}=\omega_n-\omega_0$, and $\langle n|\Sigma_i e^{i\mathbf{k}\cdot\mathbf{r}_i}|0\rangle$ denotes the matrix element of an electron density operator $\Sigma_i e^{i\mathbf{k}\cdot\mathbf{r}_i}$ between the N_e-electron ground state Ψ_0 with energy $\hbar\omega_0$ and the N_e-electron excited state Ψ_n with energy $\hbar\omega_n$. The infinitesimally small imaginary part $+i\eta$ $(\eta\to 0+)$ in the denominator is introduced for causality or the adiabatic switching of the external field.

2.2 Fermi Liquids

In case of noninteracting Fermi particle (fermion) systems, the total energy of the system can be expressed as

$$E=\sum_{\mathbf{p}} n_{\mathbf{p}}\frac{p^2}{2m},\tag{32}$$

where the occupation number of the fermions with momentum $\mathbf{p}$ is denoted by $n_{\mathbf{p}}$. When the fermion–fermion interaction is introduced adiabatically, the ground and excited states are still assumed to be specified by the particle occupation number $n_{\mathbf{p}}$. The occupation number of the ground state $n_{\mathbf{p}}^0$ is unity at energies below the chemical potential and vanishes at energies above. The excitation from the ground state can be characterized by the deviation of $n_{\mathbf{p}}$ from $n_{\mathbf{p}}^0$, i.e., $\delta n_{\mathbf{p}}=n_{\mathbf{p}}-n_{\mathbf{p}}^0$. The collision time of particles with density n and velocity v_{F} (the Fermi velocity) is approximately $\tau=n^{-1/3}/v_{\mathrm{F}}$ if all particles participate. However, because of the Fermi–Dirac statistics, the fraction of particles participating in collisions is Δ/E_{F}, where $\Delta\simeq k_{\mathrm{B}}T$ and E_{F} is the Fermi energy. Therefore, at finite temperatures T, the lifetime of excitations near E_{F} can be estimated as

$$\tau=\frac{n^{-1/3}}{v_{\mathrm{F}}}\left(\frac{E_{\mathrm{F}}}{\Delta}\right)^2\propto\frac{1}{T^2}.$$

As seen, the low-energy excitations are long lived at low temperatures, and they are called *quasiparticles.* At absolute zero temperature, the fermion occupation $n_{\mathbf{p}}$ is discontinuous at the Fermi energy.

The energy of the system is a functional of $n_{\mathbf{p}}$ and can be expanded as

$$E[n_{\mathbf{p}}]=E_0+\sum_{\mathbf{p}}\epsilon_{\mathbf{p}}\delta n_{\mathbf{p}}+\frac{1}{2}\sum_{\mathbf{p},\mathbf{p}'}f_{\mathbf{p}\mathbf{p}'}\delta n_{\mathbf{p}}\delta n_{\mathbf{p}'}+O(\delta n_{\mathbf{p}}^3),\tag{33}$$

where $E_0=\Sigma_{\mathbf{p}}n_{\mathbf{p}}^0p^2/2m$ is the ground-state energy and $\epsilon_{\mathbf{p}}$ and $f_{\mathbf{p}\mathbf{p}'}$ are the first and second derivatives of $E[n_{\mathbf{p}}]$ with respect to $\delta n_{\mathbf{p}}$. The excitation energy of a quasiparticle of momentum $\mathbf{p}$ can be written as

$$\tilde{\epsilon}_{\mathbf{p}}=\epsilon_{\mathbf{p}}+\sum_{\mathbf{p}'}f_{\mathbf{p}\mathbf{p}'}\delta n_{\mathbf{p}'}.\tag{34}$$

We assume that $\epsilon_{\mathbf{p}}$ and $f_{\mathbf{p}\mathbf{p}'}$ are continuous and that $f_{\mathbf{p}\mathbf{p}'}$ is of short range. The long-range interaction of the Coulomb term in the case of charged fermions (i.e., electrons) can be treated by adding to the free energy an average longitudinal polarization field with a momentum transfer **q** acting on a charge fluctuation,

$$\mathbf{E}_{\text{pol}} = -i\mathbf{q}\,\frac{4\pi(-e)}{4\pi\epsilon_0 q^2}\sum_{\mathbf{p}}\delta n_{\mathbf{p}}.$$

Our ground state at absolute zero temperature can now be redefined as the *vacuum* of the *quasiparticles* and *quasiholes*, i.e., absence of excitations. Equation (33) is the basic equation of the Fermi-liquid theory of Landau (see, e.g., Nozières and Pines, 1966). The short-range interaction $f_{\mathbf{pp}'}$ actually depends upon the spins, and we can define symmetric and antisymmetric parts as

$$\begin{aligned} f^s_{\mathbf{pp}'} &= \tfrac{1}{2}\,(f^{\uparrow\uparrow}_{\mathbf{pp}'} + f^{\uparrow\downarrow}_{\mathbf{pp}'}) \\ f^a_{\mathbf{pp}'} &= \tfrac{1}{2}\,(f^{\uparrow\uparrow}_{\mathbf{pp}'} - f^{\uparrow\downarrow}_{\mathbf{pp}'}), \end{aligned} \tag{35}$$

where $f^{\sigma\sigma'}_{\mathbf{pp}'}$ is the interaction between a fermion with momentum **p** and spin σ and a fermion with $\mathbf{p}'$ and σ'. These parts can be expanded in series of Legendre polynomials P_l as

$$f^{s(a)}_{\mathbf{pp}'} = \sum_{l=0}^{\infty} f^{s(a)}_l P_l(\cos\xi), \tag{36}$$

where ξ is the angle between **p** and $\mathbf{p}'$. The momenta **p** and $\mathbf{p}'$ are on the Fermi surface. It is, however, more convenient to define dimensionless parameters

$$F^{s(a)}_l = \nu(0) f^{s(a)}_l, \tag{37}$$

where $\nu(0)$ is the density of states at the Fermi energy, $\nu(0) = Vm^* p_{\mathrm{F}}(\pi^2\hbar^3)$. Macroscopic properties of the Fermi liquid can be characterized by these parameters:

$$\text{effective mass,}\quad m^* = m\left(1 + \frac{F^s_1}{3}\right); \tag{38a}$$

$$\text{compressibility,}\quad \kappa = \kappa^0\,\frac{m^*}{m}\,\frac{1}{1+F^s_0}; \tag{38b}$$

$$\text{spin susceptibility,}\quad \chi_P = \chi^0_P\,\frac{m^*}{m}\,\frac{1}{1+F^a_0}, \tag{38c}$$

where κ^0 and χ^0_P are the compressibility and susceptibility in the ideal noninteracting system.

2.3 Local-Density-Functional Theory

We shall now return to the general electron–ion systems. The total energy of the ground state is a unique functional of the electron density $n(\mathbf{r})$ (see COLLECTIVE PHENOMENA IN SOLIDS) and is extremum with respect to $n(\mathbf{r})$. The ground state energy of the electron–ion system, i.e., E_0 in Eq. (33), can be expressed as

$$\begin{aligned} E_0[n(\mathbf{r})] = {} & \int \nu(\mathbf{r}) n(\mathbf{r}) d^3r \\ & + \frac{1}{2}\frac{e^2}{4\pi\epsilon_0}\iint d^3r\, d^3r'\,\frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \\ & + G[n(\mathbf{r})], \\ G[n(\mathbf{r})] = {} & T_s[n(\mathbf{r})] + E_{xc}[n(\mathbf{r})], \end{aligned} \tag{39}$$

where $\nu(\mathbf{r})$ is the external potential including the Coulomb potential of the nuclei, the second term of $E_0[n]$ is the electrostatic interaction between electrons, $T_s[n]$ is the kinetic energy of the noninteracting system of the same density $n(\mathbf{r})$, and E_{xc} is the rest, called the exchange–correlation energy. We assume that the electron density $n(\mathbf{r})$ can be written as

$$n(\mathbf{r}) = \sum_i |\psi_i(\mathbf{r})|^2, \tag{40}$$

and that $\psi_i(\mathbf{r})$ is a one-electron wave function of the state i, and the summation encompasses all occupied one-electron states. The ground-state N_e-electron wave function can be given as the Slater determinant of $\{\psi_i\}$. Using the variational principle, we can show that the one-electron wave function $\psi_i(\mathbf{r})$ should obey the equation (the Kohn–Sham equation)

$$\left\{-\frac{\hbar^2}{2m}\nabla^2 + \nu(\mathbf{r}) + \frac{e^2}{4\pi\epsilon_0}\int d^3r'\,\frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + \frac{\delta E_{xc}[n]}{\delta n(\mathbf{r})}\right\}\psi_i(\mathbf{r}) = \epsilon_i\psi_i(\mathbf{r}). \tag{41}$$

The total energy is, of course, not a sum of the one-electron energies ϵ_i. The energy $-\epsilon_i\delta n$ corresponds to the energy cost for removing an infinitesimally small number δn of electrons of the state i.

Usually the exchange-correlation energy functional $E_{xc}[n]$ is approximated by the local quantity

$$E_{xc}[n(\mathbf{r})] = \int n(\mathbf{r})\epsilon_{xc}[n(\mathbf{r})]d^3r.$$

This is called the *local-density-functional theory* (Jones and Gunnarsson, 1989). The formulation has been generalized to the spin-polarized case and has been successfully applied to the ground-state properties of magnetic metals. Almost all electronic structure calculations in condensed matter are now based on the local-density-functional theory with great success in describing ground-state properties.

Although the density-functional theory is exact for the ground state, actual calculations are restricted by several approximations, in particular by the local approximation. Therefore, they still have several persistent and systematic errors, such as errors in the long-range tail of charge density from the surface, and incorrect prediction of metallic ground states of the transition-metal oxides and the undoped mother compounds of high-T_c cuprate superconductors. Several attempts have been made to improve or extend the local-density-functional theory in order to remove errors in the ground-state properties and to be able to predict excited-state properties such as the band gap in semiconductors and insulators, e.g., gradient expansion of the exchange correlation energy, the self-interaction correction, and many-body effects for the exchange-correlation potential.

2.4 Some Applications

2.4.1 Surfaces of Simple Metals. Several physical properties of metallic surfaces have been discussed using an electron liquid model. The ionic systems are treated as a semi-infinite positive-charge background. The exchange-correlation potential is important because in the vacuum region the electron density is very low ($r_s \rightarrow \infty$). The values of the work function and the surface energy increase with decreasing electron density in the bulk. The positive background model for solids fails in estimating the surface energy of systems with small r_s and even gives negative surface energies. However, when a model pseudopotential is assigned to each atomic position, positive and reasonable values of the surface energy can be predicted.

2.4.2 Two-Component Quantum Plasma. The theory of the one-component quantum liquid explained in Sec. 2.1 can be extended to a multicomponent quantum liquid and has been applied to problems such as screening charge around positrons in metals, electron-hole liquids, and liquid metals (Tosi *et al.*, 1974). The dielectric function Eq. (30) can be generalized as

$$\frac{\epsilon_0}{\epsilon(\mathbf{k},\omega)} = 1 + \sum_{i,j} \frac{4\pi e_i e_j}{4\pi\epsilon_0 k^2} \chi_{ij}(\mathbf{k},\omega), \tag{42}$$

where e_i is the electric charge of the ith component and $\chi_{ij}(\mathbf{k},\omega)$ is the i–j component of the density–density correlation function. In this system, there are two types of well-defined collective modes: density fluctuation of ions analogous to phonons in solids, and plasmons.

2.4.3 Two-Dimensional Electron Liquids: Quantum Hall Effect. The surface inversion layer or accumulation layer in semiconductors, and the interface between two semiconductors, are systems of typical two-dimensional electron liquids. The electron motion is quantized along the direction perpendicular to the two-dimensional layer. If we apply a magnetic field perpendicular to the two-dimensional layer, electron motion in the two-dimensional space can also be quantized. In this system, the Hall conductivity is quantized as

$$\sigma_{xy} = -n(e^2/h), \tag{43}$$

where n is an integer. This is a macroscopic appearance of quantum effects, called the quantum Hall effect, and it is now used as the standard of Hall resistance. This effect also provides an experimental method for determining the fine-structure constant $e^2/4\pi\epsilon_0\hbar c$.

Let us consider a ring or a disk with a spatial hole in the center and a magnetic field penetrating the hole (in the z direction). If the radius of an electron orbit on the disk is R, the quantization condition of the magnetic flux is

$$\pi R^2 B = m'\phi_0, \tag{44}$$

with a magnetic field B, a flux quantum $\phi_0 \equiv h/e$, and m' an integer. Generation of the Hall current is equivalent to adding another magnetic flux Φ which does not touch the metallic disk. The total magnetic flux is then given by

$$\pi R^2 B + \Phi = m''\phi_0, \tag{45}$$

where m'' is an integer, and this quantized magnetic flux appears in the phase factor of electron wave functions. If the flux Φ increases by a flux quantum ϕ_0, then R must decrease. This causes electrons to jump to an orbit with a smaller radius (along the x direction). The only change is, as a whole, that one electron moves effectively from the outer edge of the disk to the inner edge by a radial distance L. For measurement of the Hall resistance, an electric field E is applied along the radial direction (x direction) of the disk, and the resulting change of the energy is $\Delta\epsilon = eEL$. The net current along the tangential direction (y direction) can be written as

$$j_y = -\frac{1}{L}\frac{\Delta\epsilon}{\Delta\Phi} = -\frac{e^2 E}{h}, \tag{46}$$

and the Hall conductivity is $-e^2/h$. If the number of electrons shifting from one edge to the other is n, then the Hall conductivity is given by

$$\sigma_{xy} = -n(e^2/h). \tag{47}$$

This is the integral quantum Hall effect (Prange and Girvin, 1987).

So far, we have ignored the electron–electron interaction in the two-dimensional electron liquid. In a symmetric gauge where the vector potential is $\mathbf{A} = (y, -x)B/2$, the ground-state wave function of N_e electrons can be written, in a good approximation, as

$$\Psi(z_1, z_2, \ldots, z_{N_e}) \propto \prod_{j>i=1}^{N_e} (z_i - z_j)^m \times \exp\left(-\frac{1}{4}\right) \sum_{i=1}^{N_e} |z_i|^2). \tag{48}$$

Here, the electron coordinates are $z_j = x_j + iy_j$ in units of the orbital radius of the lowest Landau level $l_0 = \hbar/eB$, and m is an odd number because of the antisymmetric property of the wave function. This wave function produces a new liquid state, called the Laughlin state, with a lower energy than that of the Wigner crystal. The pair correlation function has an asymptotic behavior of the form r^{2m} for short distances, which is a more exclusive particle–particle correlation than that of the Wigner crystal. The power of a coordinate z_i in the first factor of the wave function (48) is $M = m(N_e - 1)$, and this means that the wave function (48) contains one-electron wave functions with angular momentum up to $\hbar M$. A one-electron state with angular momentum $\hbar l$ has $(l+1)$-fold degeneracy in the two-dimensional case. Therefore, this liquid state with a specific m is only possible for a specific occupancy of the Landau level, and its filling factor is $\nu = N_e/M = 1/m$ (m=odd number) when $N_e \to \infty$. Furthermore, the wave function (48) implies that the statistical behavior of the two-dimensional electron liquid is equivalent to that of the two-dimensional classical one-component plasma of charge m at the temperature $k_B T = m$. The liquid–solid transition of a two-dimensional one-component plasma has been investigated by Monte Carlo simulations, and the results indicate that the one-component plasma is a liquid when $2m < 140$ ($\nu > 1/70$) and a triangular solid when $2m > 140$ ($\nu < 1/70$). It has been shown that an excitation in the Laughlin state, a quasihole or quasiparticle, carries a fractional charge $\pm e/m$, and the Hall conductivity (the fractional quantum Hall effect) is given by

$$\sigma_{xy} = -\nu(e^2/h), \quad \nu = 1/m. \tag{49}$$

The finite density of quasiparticles and quasiholes ensures more complicated hierarchical excitation states stabilized at the electron filling factor ν;

$$\nu = \cfrac{1}{m + \cfrac{\alpha_1}{p_1 - \cfrac{\alpha_2}{(p_2 - \cdots)}}}$$

with $\alpha_i = \pm 1$, and p_i=even integers.

3. ELECTRON STRUCTURE OF LIQUID METALS AND SEMICONDUCTORS

One of the ideal structures of liquid metals is the dense random packing of spheres because of the nondirectional character of metallic cohesion. First, we will describe the structure of liquid metals and, then, several experimental and theoretical results will be presented and discussed.

3.1 Structure of Liquid Metals

The atom–atom interaction energy in liquid simple metals (*sp*-electron metals) is a sum of the two-body interactions, and the asymptotic behavior of the interatomic potential is given by Eq. (16). This is a characteristic result of the metallic behavior of electrons and weak potentials in simple metals, because the metallic bonding requires no specific directional ordering. The resulting structure of liquid metals is essentially dense random packing. The situation is very complicated in the transition metals and rare-earth metals, where the perturbation treatment for the total energy fails and multiatom interaction is quite important.

The fact is that, even in simple metals, the pair potentials show a wide variety. The pair potential is a function of the parameters $x = 2k_FR/2\pi$, $R_s = R_aZ^{-1/3}$, and R_c, where R, Z, R_a, and R_c are the interatomic distance, the valence, the atomic radius (characterizing the density), and the radius of the empty core of the pseudopotential, respectively. The parameter R_s is the electron density parameter denoted previously by r_s. The pair potential can be reduced to a factorized form

$$\Phi(R) = (Z^2/R_a)\Phi_{\mathrm{red}}(x), \qquad (50)$$

where $\Phi_{\mathrm{red}}(x)$ is shown in Fig. 4 (Hafner and Heine, 1983; Hafner, 1987). The oscillating behavior of the pair potential comes from the Friedel oscillation. The values of the phase and amplitude of the Friedel oscillations of the pair potential $\Phi_{\mathrm{red}}(x)$ are constant at a fixed R_c/R_s. The first-neighbor distance and further-neighbor configurations are determined by the position of the first minimum of the pair potential and by positions and depths of further deviations. The variation of $\Phi_{\mathrm{red}}(x)$ with R_c/R_s for a constant R_s shows a drastic change in the amplitude of the oscillation at $k_FR_c = \pi/4$ ($R_c/R_s = 0.4092$) and the damping of the oscillations accompanied by an increasing phase shift. [See Eq. (16).]

The structures of liquid metals can be classified in three different groups (Hafner, 1987). The first is the structure of I*a* and I*b* groups (Be, Mg, Al, and Pb). The structures show the following:

1. The first peaks of both the pair distribution function $g(R)$ and the static structure factor $S(Q)$ are symmetric.
2. The ratios of the positions of their second and first peaks are about $R_2/R_1 = 1.92$ and $Q_2/Q_1 = 1.86$.
3. The number of atoms n_1 in the first coordination shell is between 9.0 and 11.0.

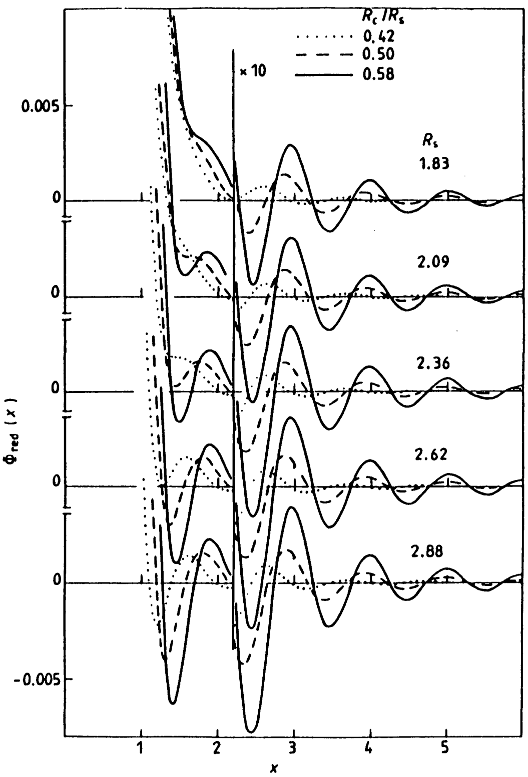

FIG. 4. Pair potential $\Phi_{\mathrm{red}}(x)$ as a function of R_c/R_s and R_s (Hafner and Heine, 1983).

These facts imply that the structure can be described by the dense random packing of hard spheres.

The second type is the group of Zn, Cd, Hg, and In, whose structures show

1. asymmetric peaks of $g(R)$ and $S(Q)$,
2. nonideal ratios of R_2/R_1 and Q_2/Q_1, and
3. metallic coordination number $n_1 \simeq 9$–11.

The last type is the structures of Ga, Si, Ge, and Sn, whose structures are greatly different from those of random packing of hard spheres and show

1. the first peak with shoulder in $S(Q)$ and with subsidiary minima in $g(R)$,
2. large deviations from the ideal values $R_2/R_1 \geqslant 2.0$ and Q_2/Q_1–2.0, and
3. low coordination number $n_1 \simeq 6.5$–7.

These three different classes can be explained by the pair potential and the systematic variation of the valency Z or R_s with a constant R_c/R_s.

3.2 Electron Spectroscopy in Liquid Metals

Electronic structures in liquid metals have been investigated quite accurately using synchrotron radiation in x-ray photoelectron spectroscopy (XPS) and ultraviolet photoelectron spectroscopy (UPS) experiments (Indlekofer *et al.*, 1990). In the photoelectron experiments, the photon energy is used to lift an electron from an occupied level to a vacuum level. The photoelectron intensity $I(E,\hbar\omega)$ is given approximately by

$$I(E,\hbar\omega) \propto (1/\omega^2) \sum_l N_l(E)\sigma_l(E,\hbar\omega), \qquad (51)$$

where E and $\hbar\omega$ are the energies of an electron and an incident photon, respectively, $N_l(E)$ is the electron partial density of states with the orbital angular momentum quantum number l, and $\sigma_l(E,\hbar\omega)$ is the partial photoionization cross section. Therefore, the intensity of photoelectrons does not give the exact density of states, but we can obtain information about it. The partial photoionization cross section depends rather sensitively on the incident photon energy and we can identify the orbital character of electrons in a given energy range by varying the photon energy. Comparison with crystals and amorphous systems provides much fundamental information on the cohesion of condensed matter. The problems of the photoelectron experiments in liquid metals may be summarized as follows:

1. The surface of a liquid metal is easily covered by oxides, and it is crucially important to obtain a clean surface.
2. The atomic composition of liquid metal alloys at a surface may be different from the bulk.
3. The vapor pressure at the melting point is, for many liquid metals, too high to be treated with the technique presently available.

A large amount of experimental data has been accumulated for monovalent metals (Li, Na) and polyvalent metals (Al, Ga, Ge, In, Sn, Hg, Tl, Pb, Bi) and their alloys (e.g., Indlekofer *et al.*, 1990). These experimental results can be summarized as follows:

1. The valence-band spectra of Li are in good agreement with a free-electron-like parabolic band, but the width (3.7 eV) is reduced significantly from the free-electron band (4.7 eV). The bandwidth of Na (3.2 eV) shows a minor reduction from the free-electron model.
2. The valence-band spectra of Al and Si show very simple free-electron-like *sp* bands. That of Ge, however, shows a small dip between *s* and *p* bands. The dip between *s* and *p* bands in Te is much enhanced and the two bands are almost separated. The spectra of these liquids show a sharp Fermi edge and are consistent with the transport properties. The calculated density of states (Hafner and Jank, 1990) will be discussed in the next section.
3. Solid Bi is a semimetal in the solid phase but a metal in the liquid phase. A band gap between 3*s* and 3*p* states is observed in Bi, Pb, and Tl, as shown in Fig. 5. The 6*s* band is located at 10, 8, and 5 eV below the Fermi energy, respectively. The 6*p* band in these liquids shows a small splitting of the peak,

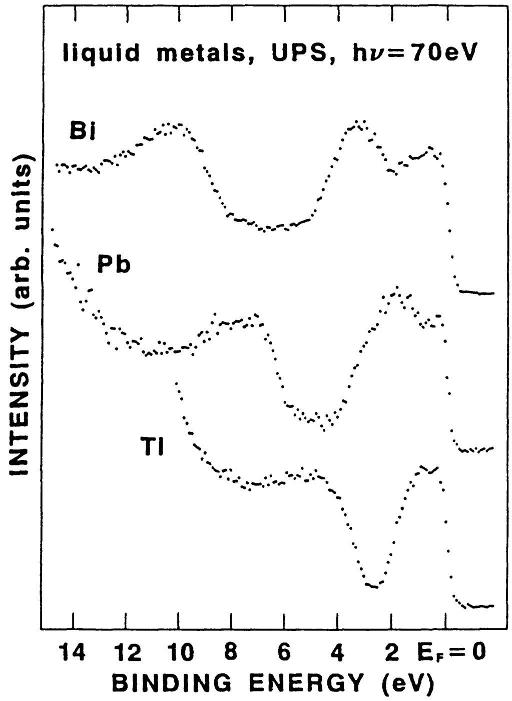

FIG. 5. Ultraviolet photoelectron spectra of liquid Bi, Pb, and Tl (Indlekofer *et al.*, 1990).

and this splitting is attributed to the spin–orbit interaction. The size of the band gap between the 3*s* and 3*p* bands and the depth of the 3*s* band change systematically.

4. The valence-band spectra of $Bi_{50}Ga_{50}$ are almost the same as those of pure Bi, while the valence-band spectra of $Bi_{50}Tl_{50}$ are different from those of both Bi and Tl. This is explained by the fact that the concentration of Bi atoms is much higher in the surface region in liquid BiGa alloys, while the atomic composition of $Bi_{50}Tl_{50}$ alloy scarcely changes in the surface region.

3.3 *sp*-Electron Metals

The systematic change in the electronic density of states for simple metals (*sp*-electron metals) has been investigated theoretically by Hafner and Jank (1990), using the LMTO-supercell method in a system of 50–70 atoms generated by the molecular-dynamics simulation. The agreement with experiments is fairly good. These results can be summarized as follows, and the calculated densities of states are shown in Fig. 6.

1. For Li and Be, a strong reduction of the occupied band and pronounced structure-induced minima in the density of states are found.
2. The densities of states of Na, Mg, Al, and Si are very close to those of the free-electron model.
3. In alkaline-earth metals, *s*-*d* hybridization and relativistic energy lowering of the *s* state induce strong deviation of the density of states from that of the free-electron models. The number of *d* electrons in liquid Ba is estimated to be 0.74/atom.
4. In the group-II*b* elements (Zn, Cd, Hg), the narrow *d*-band overlaps with the bottom of the broad *sp* band and the density of states at the Fermi energy is greatly reduced compared to the free-electron values.
5. In In and Tl, a deep minimum of the density of states appears 2–3 eV below the Fermi energy, highly analogous to the crystalline case. However, the electronic struc-

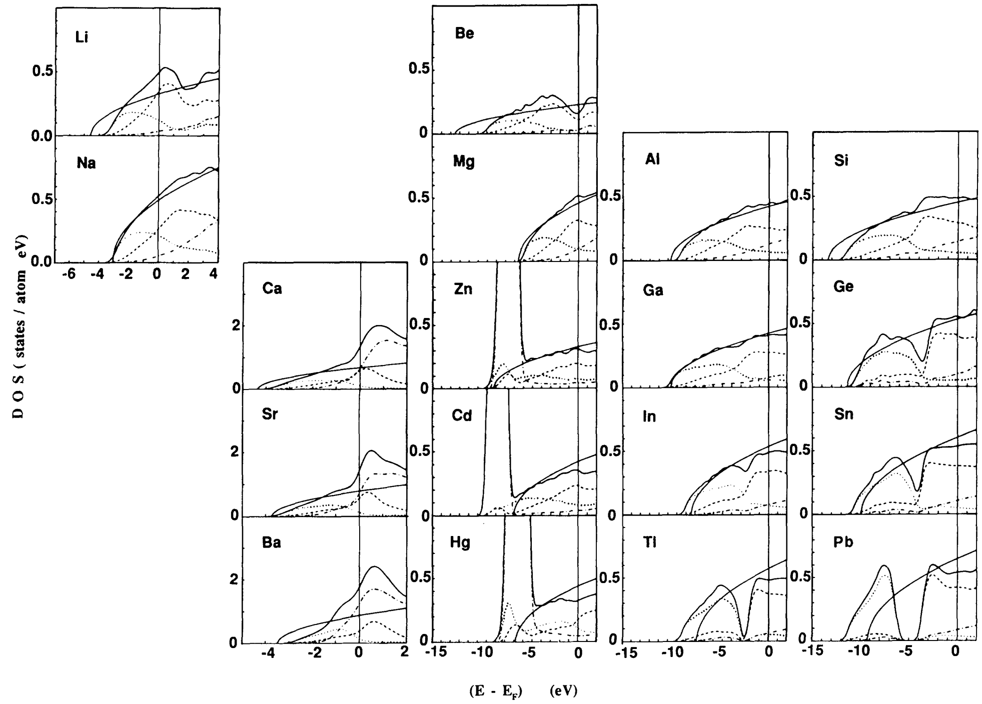

FIG. 6. Calculated density of states (DOS) in liquid *sp* metals and group-IV elements. Full curve, total density of states; dotted curve, *s* states; broken curve, *p* states; chain curve, *d* states; solid parabola, free-electron result. A vertical line shows the position of the Fermi energy. Energy origin is fixed at the Fermi energy (Hafner and Jank, 1990).

ture of Ga seems to be very different from the crystalline case whose density of states has a deep minimum at the Fermi energy.

3.4 Group-IV Elements: Si, Ge, Sn, and Pb

In the solid phase, Si and Ge are tetrahedrally bonded semiconductors with the diamond structure. The wave functions are sp^3 hybridized orbitals, and the band gap is attributed to the bonding–antibonding splitting of the hybridized orbitals. Tin has two solid phases: one is white tin (<13.2 °C) and the other is gray tin (>13.2 °C). Gray tin (α-Sn) has the diamond structure and is a zero-gap semimetal. White tin (β-Sn) has the tetragonal structure and is metallic. In heavier elements, the hybridization becomes weaker and the hybridization gap disappears. β-Sn shows a broad minimum between -4 and -7 eV below the Fermi energy which separates the 5*s* and 5*p* bands. The band gap appears between the 6*s* and 6*p* bands in Pb with the fcc structure.

Recent optical and dc-conductivity experiments show that even liquid carbon is metallic, and this is consistent with the very recent theoretical investigation of liquid carbon using first-principles molecular dynamics (Galli *et al.*, 1990). Liquid Si is metallic and has a free-electron-like density of states and no sp^3 hybridization. In liquid Ge and Sn, the *s-p* band separation leads to a broad minimum in the occupied density of states. In liquid Pb, the 6*s* and 6*p* bands are gap-separated at 4–5 eV below E_F. The structural deviation of Si and Ge from the ideal random packing is due to the Friedel oscillations of the pair potentials. In heavier elements (Sn, Pb), the separation of *s* and *p* states is enhanced by relativistic effects, and this leads to a strong damping of the Friedel oscillations as a result of smaller R_s. Therefore, the *s-p* separation stabilizes the close-packing structure. The calculated densities of states for these elements can be seen in Fig. 6.

3.5 Liquid Chalcogens: S, Se, and Te

The liquid chalcogens, sulfur, selenium, and tellurium, are other typical examples of liquid metals (Endo, 1985). The melting point of liquid sulfur is 113 °C. At the melting point, it is a molecular liquid with an eight-member molecular ring S_8 and is an insulator. A structural change occurs at 159 °C from the S_8 ring to a polymer chain with 10^6 atoms. At sufficiently high temperatures, it is a semiconductor. The polymerization leads to the appearance of a new optical band at 3.2 eV, which is assigned to the transition from a lone-pair band to a σ* band (antibonding band of σ orbitals).

Selenium has a chain structure at the melting point (217 °C) and behaves as a semiconductor, while at higher temperatures it is metallic. The valence–electron configuration is $4s^24p^4$. In the chain structure, two 4*p* electrons out of four occupy the σ bond connecting two Se atoms. The other two 4*p* electrons do not contribute to the bonding mechanism and form the lone-pair orbitals (LP) whose lobes extend perpendicular to the chain. The energy of two 4*s* electrons is much lower than that of the σ bond, and therefore the σ and the lone-pair orbitals are of almost pure 4*p* character. The schematic density of states is shown in Fig. 7. The bond angle of Se-Se-Se is about 90°, and the dihedral angle in the solid phase is 90°. The lone-pair orbital is almost perpendicular to the chain. Furthermore, the neighboring lone-pair orbitals are perpendicular to each other. These geometrical configurations are mainly determined by the repulsive interaction between the lone-pair orbitals. The lone-pair electrons essentially only contribute to the interaction between neighboring chains. The cohesive properties do not change very much at the melting point in Se. The liquid phase is characterized by fluctuation of the dihedral angle $\pm 90°$ of the spiral chain. In the high-temperature (~ 1200 °C) and high-pressure (~ 400 bar) region, liquid Se is in a metallic phase. In this region, the number of atoms per chain is reduced to 10, which has been observed by NMR experiments. At the same time, the density becomes ten times larger than that of the low-temperature semiconductor phase. The semiconductor–metal transition is also observed in the optical ab-

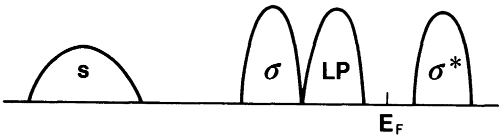

FIG. 7. Schematic energy spectra of liquid Se. LP stands for the lone-pair bands and σ* for the antibonding σ orbitals.

sorption spectrum. The absorption edge disappears, and the optical gap vanishes near the critical point.

Tellurium has a melting point at 450 °C, and the liquid phase is metallic with a coordination number of about 3. The resistivity is about 770 $\mu\Omega$ cm at 400 °C and about 360 $\mu\Omega$ cm at 800 °C. The threefold coordinated structure is an outcome of the shortened polymer chain and the interaction between the lone-pair electrons and the σ^* antibonding orbitals.

3.6 Liquid Transition Metals

The liquid transition metals are the extreme opposite of the *sp*-electron metals, and the nearly free-electron model does not work for them at all. A tight-binding model may, however, be applicable and could serve as a reasonable first-step approach. Fujiwara (1979) applied the tight-binding recursion method to liquid and amorphous transition metals Fe, Co, and Ni, using tight-binding parameters so determined as to reproduce the density of states obtained by standard band calculations in the crystalline cases. An apparent difference was found in the density-of-states curves of liquid and amorphous metals. In the amorphous transition metals, the densities of states have a double-peaked asymmetric shape, while those in liquids are single-peaked and asymmetric. These differences are attributed to the difference of the atomic density fluctuations and local environment. Asano and Yonezawa (1980) extended the effective-medium approximation to the Korringa–Kohn–Rostoker method. The density of states in liquid Ni, shown in Fig. 8, closely resembles that of fcc Ni and the result calculated by the tight-binding recursion method. This resemblance indicates the existence of a close-packed local environment in liquid structures similar to the fcc structure. Hafner and Jank (1990) presented the density of states of liquid transition metals Cr–Ni by the LMTO supercell method and reported the persistence of the double-peaked structure even in liquid systems.

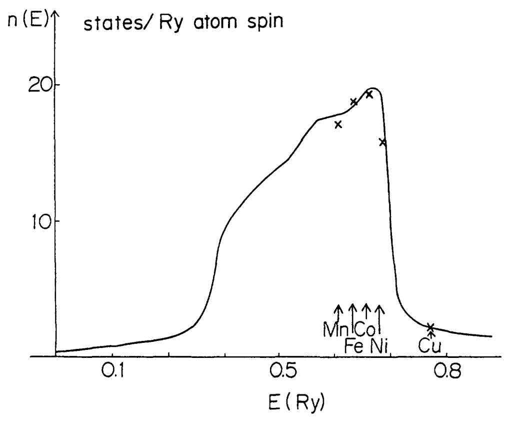

FIG. 8. Density of states of liquid Ni. Crosses denote the values estimated from susceptibility experiments. The Fermi energies of Ni, Co, Fe, Mn, and Cr are shown by arrows (Asano and Yonezawa, 1980).

3.7 Metal–Nonmetal Transition

A metal–nonmetal transition is observed in expanded Hg (higher density than the critical density) and in Cs (in the critical region). The metal–nonmetal transition in these systems is due to *s-p* band separation with changing coordination number, but not due to the scattering of electrons by the random potential.

3.7.1 Mercury. The resistivity of pure liquid Hg (density$=13.6$ g/cm^3) is about 100 $\mu\Omega$ cm (25 °C), and the corresponding mean free path is 0.7 nm. The conduction band of Hg originates from the 6*s* and 6*p* bands (see Fig. 6). Near the critical point, the density is very low; therefore the bandwidth becomes narrower because of the large separation between atoms, and the density of states at the Fermi energy $N(E_F)$ is reduced. At 9.0 g/cm^3 and 1440 °C, the resistivity is about 2500 $\mu\Omega$ cm. The system transforms to an insulating phase at a density about 1.6 times larger than the critical density d_c at the critical point ($d_c=5.7$ g/cm^3). The metal–nonmetal transition of Hg in the low-density region was confirmed by several experiments: thermoelectric power, Knight shift of NMR spectra, reflection spectra, thermal pressure coefficients, etc. These experiments show a change in the characteristics of interatomic interactions, accompanied by a drastic change in the transport properties. Yonezawa *et al.* (1977) presented the calculated density of states in expanded Hg and showed that there is a reduction in the density of states at E_F with decreasing atomic density and that a band gap opens up at some critical density around 2 to 4 g/cm^3, though this value is too low in comparison with the observed value. This metal–nonmetal transition is essentially due to the 6*s*-6*p* band separation in a diluted region.

3.7.2 Cesium. A metal–nonmetal transition is also observed in the monovalent metal Cs. The compressibility of heavier alkali metals is very large, and the density can be doubled by compression. In alkali metals, the temperature dependence of the transport properties can be well described by the nearly free-electron model in the region where the density is twice or much larger than the critical density. With increasing temperature, the density is reduced, and the resistivity increases gradually and finally reaches a value of two orders of magnitude larger. At the critical density $d_c = 0.41$ g/cm^3, the resistivity reaches a value of 3300 $\mu\Omega$ cm. The resistivity of Rb shows a behavior similar to that of Cs. The magnetic experimental data indicate that the susceptibility in the region $d > d_c$ can be explained by the Pauli-paramagnetic plus Landau-diamagnetic behavior enhanced by electron correlation effects and, on the other hand, that the susceptibility in the critical region follows the Curie law of localized spin. Therefore, electron–electron correlation effects may be important in the metal-nonmetal transition in Cs.

3.8 Electron Localization

Electron scattering by the random potentials in liquid metals leads to appreciable electrical resistivity and would therefore be expected to affect the electronic structures as well. The effect of the second-order energy shift and width may be included effectively in the self-energy, Eq. (20), by the self-consistent substitution of G for G^0, and the resulting formula is

$$\Sigma_{\mathrm{eff}}(\mathbf{k},E) = \frac{V/N}{(2\pi)^3} \int a(\mathbf{k}-\mathbf{q}) \times \frac{|w(\mathbf{k}-\mathbf{q})|}{E-\epsilon_{\mathbf{q}}-\Sigma_{\mathrm{eff}}(\mathbf{q},E)}\, d^3q. \tag{52}$$

This formula has been solved numerically, using pseudopotentials for Al, Li, Bi, etc. (Ballentine 1966; Chan and Ballentine, 1972). The density of states in these calculations was, more or less, similar to that of the free-electron model with a shallow dip at $k = K_p/2$. On the other hand, recent photoelectron experiments (Indlekofer *et al.*, 1990) show an almost free-electron-like *sp* band in Al and a separation of 6*s* and 6*p* bands in liquid Bi, but a depression of the density of states at $K_p/2$ was not observed. The supercell method was employed for calculating the electronic density of states in liquid Al and Bi, and the results can explain the experimentally observed spectra.

When electrons are scattered by several scattering centers, do wave functions of the electrons still extend over the whole system or do they localize at some specific center? The possibility of localization of the electronic wave function in a random potential field was first pointed out by Anderson in 1958 (Abraham *et al.*, 1979) and is now called Anderson localization (see ELECTRON STATES, LOCALIZED). Anderson localization is essentially due to the interference between a plane wave with wave vector **k** and the backscattered wave with wave vector −**k**. We might say that the electrons in liquids localize as a result of the randomness of the atomic configuration and that the transport in liquid metals occurs through the overlap of electron wave functions localized at each center, i.e., a density-fluctuation–assisted hopping process. However, this is very unlikely because normal metallic electron transport has been observed in amorphous metals whose atomic structure is believed to be just the quenched version of the liquid phase. The dimensionality is quite important in Anderson localization. In three-dimensional metals, even though there might be some localized states, the states at the Fermi energy are usually not localized. Furthermore, in amorphous metals at very low temperatures, phenomena premonitory to localization can be experimentally observed (Cochrane and Strom-Olsen, 1984).

4. FIRST-PRINCIPLES MOLECULAR DYNAMICAL SIMULATION

Computer simulations are a very important means of investigating the static and dynamical structure of liquids. The dynamics of atoms and molecules may be studied by solving the equations of motion. This is called the molecular dynamics method. After storing all data of atomic motions, we can observe or calculate several physical quantities of the model system. Conventional molecular dynamical simulation is based upon empirical interatomic potentials. Electron motions follow the atomic motions very faithfully, and with changing local atomic configurations, the electron wave functions are changed accordingly. It takes a great deal of computational time and effort to simulate atomic dynamics and dynamical effects on the electron wave functions by conventional methods.

Very recently, a new method was developed by Car and Parrinello (1985), wherein the electronic and atomic motions are simultaneously simulated. Usually we calculate the electronic structure in a fixed configuration of the ionic positions $\{\mathbf{R}_n\}$, and the electron charge density $n(\mathbf{r};\{\mathbf{R}_n\})$ is evaluated for this configuration. The new idea treats the electron density $n(\mathbf{r})$ and atomic configuration $\{\mathbf{R}_n\}$ as independent variables so as to find an extremum of the total energy in the extended configurational space $(n(\mathbf{r}),\{\mathbf{R}_n\})$, or more specifically $(\{\psi_i(\mathbf{r})\}, \{\mathbf{R}_n\})$, where $\psi_i(\mathbf{r})$ is the electronic wave function of the state *i*. The basic quantity is the local density functional $E_{\mathrm{LDF}}(\{\psi_i\}, \{\mathbf{R}_n\}, \{\alpha_\nu\})$, which includes the electron kinetic energy and electron–electron, electron–ion, and ion–ion interactions. The parameters α_ν are external constraints imposed on the system such as the volume of the system, symmetry, the thermostat leading the system to a canonical distribution, and so on. The second step of the new idea is the introduction of the Lagrangian

$$L=\sum_i \frac{1}{2}\mu \int d^3r|\dot{\psi}_i|^2+\sum_n \frac{1}{2}M_n|\dot{\mathbf{R}}_n|^2 + \sum_\nu \frac{1}{2}\mu_\nu\dot{\alpha}_\nu^2-E_{\mathrm{LDF}}(\{\psi_i\},\{\mathbf{R}_n\},\{\alpha_\nu\}), \quad (53)$$

with the orthogonalization condition $\int d^3r\psi_i^*\psi_j=\delta_{ij}$. The dot denotes time derivative, M_n is the ion mass, and μ and μ_ν are arbitrary parameters. The first and third terms are fictitious kinetic energies of the electrons and the external constraints. Note that the actual electron kinetic energy is included in the term E_{LDF}. The equations of motion can be derived from the Lagrangian as

$$\mu\ddot{\psi}_i(\mathbf{r},t)=-\frac{\delta E_{\mathrm{LDF}}}{\delta\psi_i^*(\mathbf{r},t)}+\sum_k \Lambda_{ik}\psi_k(\mathbf{r},t),$$

$$M_n\ddot{\mathbf{R}}_n=-\nabla_{\mathbf{R}_n}E_{\mathrm{LDF}},$$

$$\mu_\nu\ddot{\alpha}_\nu=-\frac{\delta E_{\mathrm{LDF}}}{\delta\alpha_\nu}. \quad (54)$$

The Lagrange multipliers Λ_{ik} are introduced because of the orthogonality condition for electron wave functions. The equation of the ionic motion is physical if the electron wave functions are on the Born–Oppenheimer surface, which can be specified by $\min_{(\{\psi_i\}, \{\alpha_\nu\})}E_{\mathrm{LDF}}(\{\psi_i\},\{\mathbf{R}_n\},\{\alpha_\nu\})$. The equations for the other variables $\{\psi_i\}$ and $\{\alpha_\nu\}$ are fictitious. If we successfully reach an equilibrium state for the electronic configuration, $\ddot{\psi}_i=0$, then the wave functions should satisfy the Kohn–Sham equation and be on the Born–Oppenheimer surface. The values of the parameters $\{\mu\}$ and $\{\mu_\nu\}$ should be chosen so that an equation of motion follows a trajectory not far from the Born–Oppenheimer surface.

The (norm-conserving) pseudopotential formalism is usually adopted, because the plane-wave base is very convenient for actual calculations. If we use a local basis set of electron wave functions and the centers of electron wave functions are chosen to coincide with the centers of moving atoms, additional variational forces acting on atoms and electron wave functions appear and create very serious difficulties. Furthermore, the transferability (i.e., transferable to the same ion in different local environments) of the norm-conserving pseudopotential is very desirable for treating a system with fluctuating local environments.

Up to now, several atomic liquids, C, Si, Ge, As, Al, and liquid GaAs have been investigated by this method. A standard supercomputer allows a unit cell size of about 50 atoms with periodic boundary conditions and an energy cutoff of about 10 Ry for the plane-wave expansion. As an example, we summarize the results for liquid Si (Stich *et al.*, 1989). These results show excellent agreement with the observed structure factor. The velocity autocorrelation function shows a bump at a frequency of ~40 meV, which is very close to the optical vibrational frequency of crystalline Si at temperatures just below the melting point. This may imply a persistence of solidlike atomic arrangement in the liquid state. The

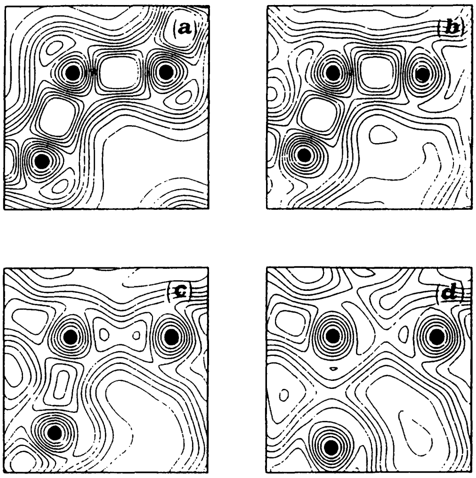

FIG. 9. Valence-charge contours of liquid Si. Black circles are Si atoms. (a) Crystalline Si in the (110) plane, (b)–(d) evolution of charge density in liquid Si at time intervals of ~0.02 ps (Stich *et al.*, 1989).

contour plot of the charge density (Fig. 9) also shows the local valence-bond character remaining in the liquid phase within a period of ~0.02 ps. For larger periods, some bonds start to break and reform. The electronic density of states in the liquid state indicates metallic behavior, and the conductivity calculated by the Kubo formula is estimated to be 0.0175 $(\mu\Omega\ \text{cm})^{-1}$, in good agreement with experimental results.

5. ELECTRON AND ATOMIC STRUCTURE OF WATER

Water is a very unique substance whose properties are determined by hydrogen bonds and their network. The melting and boiling temperatures are much higher than those of many other molecular liquids of similar molecular weight. The heat of evaporation and the heat capacity are very large.

We should also point out that the behavior of nonelectrolyte solutions such as Ar, methane (CH_4), and other nonpolar molecules with C–H base in water is anomalous. The solubility of these molecules is very low because of the hydrophobic bases C–H. The hydrophobic solute molecules have a tendency to come together and attract each other in order to reduce the area touching water molecules. This effective interaction between hydrophobic bases is called a *hydrophobic interaction* or a *solvent-separated hydrophobic pair interaction.* The ordered network of hydrogen bonds is broken in solutions, and the entropy increases. Therefore, the hydrophobic bond is stabilized at higher temperatures. Sometimes, water molecules form a more rigid hydrogen-bonded network (called an *iceberg* model) around a guest molecule (Ar, CH_4) with decreasing entropy. This is called *hydrophobic hydration.*

To understand these interesting properties of water and hydrogen bonds, we must first understand the structure of water molecules and their interactions. The distance between oxygen and hydrogen atoms in a water molecule is 0.09572 nm and the bond angle is 104.52°. The oxygen 2*s* and 2*p* orbitals form sp^3 hybridized orbitals $\phi_O(t_1)$–$\phi_O(t_4)$ as follows:

$$\phi_O(t_i)=\frac{1}{\sqrt{3+v^2}}\{v\phi_O(2s)\pm\phi_O(2p_x)\pm\phi_O(2p_y)\pm\phi_O(2p_z)\}, \tag{55}$$

where $\phi_O(2s)$ and $\phi_O(2p_x)$ are the oxygen 2*s* and $2p_x$ orbitals and v is a mixing parameter [Fig. 10(a)]. Two electrons from two hydrogen atoms and two electrons of the oxygen atom occupy one of the bonding orbitals

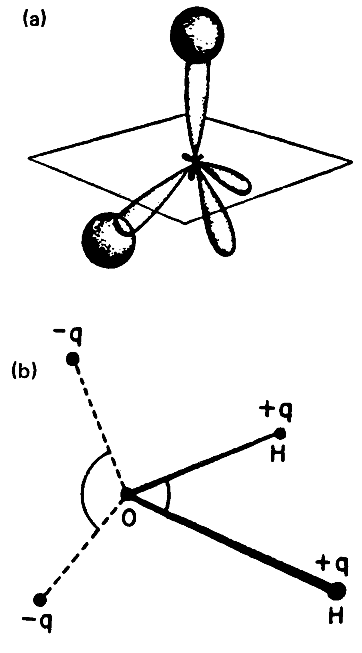

FIG. 10. (a) Extent of sp^3 orbitals of a water molecule; (b) a model structure of a water molecule.

formed by the hydrogen 1*s* orbital and either t_1 or t_2 hybridized orbitals:

$$\psi_1 = \lambda\phi_{H1}(1s) + \mu\phi_O(t_1),$$
$$\psi_2 = \lambda\phi_{H2}(1s) + \mu\phi_O(t_2), \tag{56}$$

where $\phi_{H1}(1s)$ is the 1*s* orbital of hydrogen atom 1. The remaining four valence electrons of the oxygen atom occupy the unbonded t_3 and t_4 orbitals. The bond angle 104.52° can be explained by appropriate values of the mixing parameters ν, λ, and μ and the Coulomb repulsion between electrons.

The symmetrical tetrad of point charges has been adopted as a model structure of the water molecule, based on the above bonding and unbonded orbitals, and this model ensures the existence of the tetrahedrally directed hydrogen bonds. One of the model pair potentials for water molecules is based on a rigid four-point-charge model proposed by Stillinger and Rahman (1974). The geometry involved is shown in Fig. 10(b). The positive charges $+q$ are partially shielded protons, and two negative point charges $-q$ are positioned 0.08 nm away from the oxygen ion. A vector connecting the two negative charges is perpendicular to the H-O-H plane. The water-water interaction consists of the 12-6 Lennard–Jones potential between oxygen atoms and Coulomb interactions of 16 pairs of positive and negative point charges $\pm q$.

Carravetta and Clementi (1984) calculated the electronic structure of water dimers in various geometrical configurations, considering the electron correlation effects, and constructed a revised pair potential. Each water molecule represents two positive charges $+q$ for the protons and one negative charge $-2q$ for the unbonding electrons of the oxygen ion on the bisection lines of the H-O-H bond angle. The new pair potential consists of Coulomb interactions of nine pairs of charges and numerically determined exponential terms between atoms. The correlation effect included by the configuration interaction accounts for about -4×10^3 J mol^{-1} of the total dimerization energy of about $(1.8\text{–}2.3)\times10^4$ J mol^{-1}. The correlation energy has a shallow minimum at the oxygen–oxygen distance of 0.37–0.42 nm. The potential minimum is obtained at the oxygen–oxygen distance of about 0.29 nm.

Several calculations of polymerized water molecules show that the structure of $(H_2O)_n$ is more stable for larger *n*. The pair distribution functions calculated by molecular dynamical simulation are in satisfactory agreement with experimental results, and the calculated short-range structure of water, in regard to the number of near-neighbor molecules ($\simeq$4.4) and the maximum peak position ($\simeq$0.45 nm), is quite similar to that of hexagonal ice.

The molecular dynamical simulation can also give the equation of state. A simulation by Stillinger and Rahman (1974) reproduced the density maximum at 27 °C (experimentally 1.0 g/cm^3 at 4 °C). The calculated velocity correlation function shows a structure of molecular vibrations and implies a persistence of the polymerized structure of a hydrogen-bonded network for a period of the order of 10^{-12} s. It was also suggested that the extent of the hydrogen-bonded network is limited to fewer than eight molecules.

The calculated average interaction energy was reported to be $(-3.9\pm0.02)\times10^4$ J mol^{-1} by Carravetta and Clementi (1984) and, adding the kinetic energy of molecular rotation and translation $3RT$, a total energy of -3.21×10^4 J mol^{-1} was obtained. This is in good agreement with the experimental value of -3.39×10^4 J mol^{-1}. The oxygen–oxygen radial distribution function is in good agreement with the experimental ones.

GLOSSARY

Anderson Localization: Localization due to randomness by means of quantum interference between a wave with wave vector **k** and a backscatter wave with wave vector −**k**.

Born–Oppenheimer Surface: A surface specifying the ground-state configuration and its energy as a function of the *N*-atom configuration $\{\mathbf{R}_n\}$.

Configuration Interaction: Interaction between several electron configurations, which stabilizes a system more than a state specified by a single electron configuration.

Dense Random Packing of Hard Spheres: A structure of densely packed hard spheres, which contains no internal holes large enough to accommodate another sphere and has only weak correlations between positions of spheres with intersphere distances more than five diameters.

Free-Electron Model: A model of free electrons, which is applicable to simple metals, i.e., no external potentials and no electron-electron interactions. A *nearly-free electron model,* starting from the free-electron model, can describe both periodic and aperiodic systems with weak potentials.

Friedel Oscillation: Oscillating electron density around a charged particle in condensed matter, which is due to screening of the charged particle and Pauli's exclusion principle for electrons.

Korringa–Kohn–Rostoker Method: One of the methods of band structure calculations, invented by Korringa, Kohn, and Rostoker and based on a Green's-function formalism of the Schrödinger equation in solids (Ziman, 1971).

Linear Muffin-Tin Orbital Method: A theoretical method for calculating the electronic structure in solids (Andersen, 1975) using energy-linearized partial waves

$$\psi_{\mathbf{R}l}(\mathbf{r}) = \phi_{\mathbf{R}l\nu}(\mathbf{r}) + (E - E_\nu)\dot{\phi}_{\mathbf{R}l\nu}(\mathbf{r}),$$

where $\phi_{\mathbf{R}l\nu}$ is a solution to the radial Schrödinger equation in each spherical potential region of an appropriately chosen energy E_ν and $\dot{\phi}_{\mathbf{R}l\nu}$ is its energy derivative.

Molecular Dynamics: A simulation method for the dynamics of atoms and molecules, obtained by computerized solution of the equations of motion.

Muffin-Tin Potential: Potentials in condensed matter, spherically symmetric inside nonoverlapping spheres at each ion and constant between the spheres.

NMR: Nuclear magnetic resonance, whose resonance frequency gives information about the local magnetic field and relaxation of spin motion to an equilibrium state.

Pair Distribution Function: A distribution function characterizing the atomic arrangement around one atom. Its Fourier transform can be measured by scattering intensity.

Quasiparticle: Low-energy excitation of fermion systems near the Fermi energy, which has a long life-time and behaves as an almost independent particle.

Slater Determinant: A many-electron wave function of the form of an $N_e \times N_e$ determinant:

$$\Psi(\{\mathbf{r}_i\}) = (N_e!)^{-1/2} \times \begin{vmatrix} \phi_a(\mathbf{r}_1) & \phi_a(\mathbf{r}_2) & \cdots & \phi_a(\mathbf{r}_N) \\ \phi_b(\mathbf{r}_1) & \phi_b(\mathbf{r}_2) & \cdots & \phi_b(\mathbf{r}_N) \\ \phi_c(\mathbf{r}_1) & \phi_c(\mathbf{r}_2) & \cdots & \phi_c(\mathbf{r}_N) \\ \cdots & \cdots & \cdots & \cdots \end{vmatrix},$$

which satisfies Pauli's exclusion principle.

***sp*-Electron Metals:** Metals whose valence electrons are of *s* or *p* type, with electron wave functions well approximated by plane waves.

Structure Factor: The Fourier transform of the atomic density function in the bulk.

Tight-Binding Model: A model for the electronic structure of materials, which can be described by localized atomic wave functions and short-range electron hopping between pairs of atoms. This model is an extreme opposite of the nearly free-electron model.

List of Works Cited

Abraham, E., Anderson, P. W., Licciardello, D. C., Ramakrishnan, T. V. (1978), *Phys. Rev. Lett.* **42**, 673–676.

Anderson, O. K. (1975), *Phys. Rev. B* **12**, 3060–3083.

Anderson, O. K., Jepson, O. (1984), *Phys. Rev. Lett.* **53**, 2571–2574.

Asano, S., Yonezawa, F. (1980), *J. Phys. F* **10**, 75–97.

Ballentine, L. E. (1966), *Can. J. Phys.* **44**, 2533–2552.

Car, R., Parrinello, M. (1985), *Phys. Rev. Lett.* **55**, 2471–2474.

Carravetta, V., Clementi, E. (1984), *J. Chem. Phys.* **81**, 2646–2651.

Chan, T., Ballentine, L. E. (1972), *Can. J. Phys.* **50**, 813–820.

Cochrane, R. W., Strom-Olsen, J. O. (1984), *Phys. Rev. B* **29**, 1088–1090.

Elliott, R. J., Krumhansl, J. A., Leath, P. L. (1974), *Rev. Mod. Phys.* **46**, 465–543.

Endo, H. (1985), *Prog. Theor. Phys. Suppl.* **72**, 100–124.

Fujiwara, T. (1979), *J. Phys. F* **9**, 2011–2024, amendment *op. cit.*, **10**, 181.

Fujiwara, T. (1984), *J. Non-Cryst. Solids* **61–62**, 1039–1048.

Galli, G., Martin, R. M., Car, R., Parrinello, M. (1990), *Phys. Rev. B* **42**, 7470–7482.

Hafner, J. (1987), *From Hamiltonians to Phase Diagrams,* New York: Springer-Verlag.

Hafner, J., Heine, V. (1983), *J. Phys. F* **13**, 2479–2501.

Hafner, J., Jank, W. (1990), *J. Phys. Condens. Matt.* **2**, SA239–SA243.

Hafner, J., Payne, M. C. (1990), *J. Phys. Condens. Matt.* **2**, 221–230.

Hamann, D. R. (1989), *Phys. Rev. B* **40**, 2980–2987.

Hamann, D. R., Shlueter, M., Chiang, C. (1979), *Phys. Rev. Lett.* **43**, 1494–1497.

Harrison, W. A. (1964), *Phys. Rev.* **136**, A1107–A1119.

Haydock, R. (1980), *Solid State Phys.* **35**, 215–294.

Heine, V. (1971), *Solid State Phys.* **24**, 1–36.

Indlekofer, G., Pflugi, A., Oelhafen, P., Chauveau, D., Guillot, C., Lecante, J. (1990), *J. Non-Cryst. Solids* **117/118**, 351–354.

Jones, R. O., Gunnarsson, O. (1989), *Rev. Mod. Phys.* **61**, 689–746.

Nozières, P., Pines, D. (1966), *The Theory of Quantum Liquids*, New York: Benjamin.

Nowak, H. J., Andersen, O. K., Fujiwara, T., Jepsen, O., Vargas, P. (1991), *Phys. Rev. B* **44**, 3577–3598.

Prange, R. E., Girvin, S. M. (1987), *The Quantum Hall Effects*, New York: Springer-Verlag.

Roth, L. (1974), *Phys. Rev. B* **9**, 2476–2484.

Stich, I., Car, R., Parrinello, M. (1989), *Phys. Rev. Lett.* **63**, 2240–2243.

Stillinger, F. H., Rahman, A. (1974), *J. Chem. Phys.* **60**, 1545–1557.

Tosi, M. P., Parrinello, M., March. N. H. (1974), *Nuovo Cimento* **23B**, 135–171.

Yonezawa, F., Ishida, I., Martino, F., Asano, S. (1977), *Liquid Metals 1976* (IOP Conference Ser. No. 30), London: Institute of Physics, pp. 385–397.

Ziman, J. M. (1971), *Solid State Phys.* **26**, 1–101.

Further Reading

Hafner, J. (1987), *From Hamiltonians to Phase Diagrams*, New York: Springer-Verlag.

March, N. H. (1990), *Liquid Metals*, Cambridge, U.K.: Cambridge Univ. Press.

Ziman, J. M. (1979), *Models of Disorder*, Cambridge, U.K.: Cambridge Univ. Press.

ELECTRON STRUCTURE OF SOLIDS

CHRISTIAN MAILHIOT, *Physics Department, Lawrence Livermore National Laboratory, Livermore, California, U.S.A.*

	Introduction	1
1.	**Properties of Electron States in Crystalline Solids**	3
1.1	Introduction	3
1.2	Crystal Symmetry: Space Groups	3
1.3	Direct and Reciprocal Space: Bravais Lattices and Brillouin Zones	4
1.3.1	Example: Face-Centered-Cubic Lattice	6
1.3.2	Relation with X-Ray Diffraction by a Crystal	6
1.4	Form of Electron Solutions in Crystalline Solids: Bloch States	8
1.4.1	Proof of Bloch's Theorem	8
1.4.2	Allowed Values of **k**: Periodic Boundary Conditions	9
1.5	General Properties of Bloch States and Energy Bands	10
1.5.1	General Properties of Bloch States	10
1.5.2	Energy Bands and Energy-Band Gap	11
1.5.3	Band Occupation, Fermi Surface, and Density of States	13
1.5.4	Time-Reversal Symmetry	15
2.	**Calculation of Electron States in Solids: Band Theory**	17
2.1	Introduction: Basic Approximations	17
2.2	Fixed-Basis-Set Methods	20
2.2.1	Local-Orbital Methods	20
2.2.2	Orthogonalized-Plane-Wave Method	21
2.2.3	Pseudopotential Theory	23
2.3	Scattering Methods	26
2.3.1	Cellular Method	26
2.3.2	Augmented-Plane-Wave and Linear Muffin-Tin Methods	28
2.3.3	Green's-Function Method	30
2.4	Applications of Band Theory	31
2.4.1	Pseudopotential Band Structure of Silicon	31
2.4.2	Electron-Structure Trends among the Elements	33
3.	**First-Principles Total-Energy Methods**	33
3.1	Density-Functional Theory	33
3.2	Total Energy	36
3.3	Recent Developments in Density-Functional Theory: Harris Functional	37
3.4	Applications of Total-Energy Methods	38
3.4.1	First-Principles Pseudopotentials: Covalent Materials	38
3.4.2	All-Electron LMTO Method: Metals	41
3.4.3	Comparison between Different First-Principles Methods	43
3.4.4	Applications of the Harris Functional	43
4.	***Ab initio* Molecular Dynamics**	45
4.1	Introduction	45
4.2	Lagrangian Formulation	46
4.3	Nonlinear Local Optimization	47
4.4	Applications	48
5.	**Summary**	48
	Glossary	49
	Works Cited	49

INTRODUCTION

The quantum-mechanical description of electron states in solids has been a goal that has occupied the physics community since the advent of quantum theory. The purpose of this article is to present an overview of the modern theoretical techniques that serve as the basis for such descriptions. The determination of the electron structure of materials is a topic of

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

great importance in condensed-matter physics. In this context, the term *electron structure* refers to the distribution of electrons in space, momentum, and energy (Moruzzi *et al.*, 1978). The field is very broad and highly developed and continues to grow rapidly. Consequently, our presentation is inevitably incomplete, and the interested reader is advised to consult the many excellent references on this subject matter.

Within the one-electron band approximation, the problem of calculating the energy levels and eigenfunctions of electron states in a solid essentially reduces to solving the Schrödinger equation associated with a periodic crystal potential. Consequently, the vast majority of basic theoretical tools necessary to perform such a task were introduced in the 1930s and the 1940s soon after the development of quantum theory. In fact, most of the theoretical methods outlined in this article were formulated before the middle of the century: Bloch proposed the tight-binding method in 1928, the concept of orthogonalized plane waves was formulated by Herring in 1940, the cellular method was introduced by Wigner and Seitz in 1933, the utilization of augmented plane waves was suggested by Slater in 1937, and Green's-function methods were originated by Korringa in 1937. Initial calculations of band structure were performed on the simplest high-symmetry systems and established a proof of principle. However, the advent of high-speed digital computers in the mid-1950s propelled the field of band theory to new levels by allowing workers to perform calculations with large enough basis sets to predict features of the band structure of realistic materials: The first calculation (correctly) predicting the existence of an indirect energy band gap in silicon appeared in the literature in 1954 and was performed with the orthogonalized-plane-wave method (Herman, 1954). Over the following decades, the appearance in the solid-state physics literature of numerous applications of band-structure methods clearly indicates that, to this day, the band-structure problem remains a very timely, exciting, and challenging subfield of condensed-matter theory.

Original applications of band-structure methods relied on empirical developments in which Hamiltonian matrix elements were adjusted to reproduce experimentally observed features of the energy band structures. In most cases, early band-structure calculations did not embody a self-consistency procedure, i.e., the crystal potential was not recalculated, and the effects of electronic rearrangements (screening) arising when isolated atoms are brought together to form the solid were not incorporated. However, with the formulation of density-functional theory in the mid-1960s and the accelerating developments in computer architecture, self-consistent first-principles electron-structure (band structure and total energy) calculations emerged in the early 1970s and are routinely performed nowadays.

Up until the mid-1980s, self-consistent first-principles electron-structure calculations were possible for systems containing at most a few tens of atoms. Moreover, only static configurations of atoms were considered (at zero temperature), and the modeling of highly complex materials (disordered phases, liquids, etc.) was prohibitively expensive. However, recent theoretical developments have led to the formulation of a unified combination of first-principles methods and concepts extracted from molecular-dynamics simulations. This new formulation constitutes one of the latest advances in the field of electron-structure theory and is based on a unified treatment of the electron-structure problem by coupling the local electron structure (bonding, localization, etc.) with the atomic motions at finite temperature. Such *ab initio* molecular-dynamics descriptions allow the determination of the thermodynamical properties of a system from first-principles evaluations of microscopic quantities. By allowing finite-temperature treatment of complex condensed-matter systems, the advantages introduced by these first-principles molecular-dynamics methods will bridge the gap between electron-structure theory and materials science.

The article is organized as follows: In Sec. 1, the language of group theory is introduced for the purpose of presenting theorems that are of central importance in the theory of the electronic structure of solids. The general form of an electron state in a periodic crystal potential is deduced on the basis of symmetry considerations alone. Fundamental properties of the crystal wave functions and of their associated energy eigenvalue spectrum are presented. Important elementary concepts such as energy bands, band gap, band occupation, Fermi level, and density of states are indicated. In Sec. 2, we present the elements of band theory by describing the various theoretical methods used to calculate the energy eigenvalue spectrum and eigenfunctions of electron states in a solid. For presentation purposes, these methods are divided into two approaches: fixed-basis-set and scattering methods. While this classification is certainly not unique and each electronic-structure technique is a method in its own right, it provides a natural distinction between the different classes of methods. Modern first-principles total-energy methods based on density-functional theory are presented in Sec. 3. This theory provides the basis for most first-principles calculations in the area of solid-state physics. We present results of total-energy density-functional calculations to illustrate its predictive power and to compare different implementations of this theory. Finally, in Sec. 4 we describe recent theoretical developments that go beyond static, zero-temperature electron-structure calculations by combining first-principles electron-structure descriptions with approaches and algorithms utilized in molecular-dynamics simulations.

1. PROPERTIES OF ELECTRON STATES IN CRYSTALLINE SOLIDS

1.1 Introduction

The self-consistent calculation of the energy eigenvalues and eigenfunctions of electron states in solids is essentially a many-body problem that can only be solved exactly with the use of many-body stochastic methods. Such methods are outside the scope of this article. Approximations are universally made to reduce this problem to a tractable endeavor and to obtain approximate solutions. However, the symmetry properties of the full many-body Hamiltonian can be used to derive the resulting symmetry properties of the *exact* solutions, even though these solutions are unknown. Group theory (*q.v.*) is the mathematical apparatus that affords such derivations.

The one-electron crystal Hamiltonian we will discuss in this article is of the form

$$H \equiv H(\mathbf{x}) = \frac{-\hbar^2}{2m}\nabla^2 + V(\mathbf{x}), \qquad (1)$$

where $\hbar = 6.5822 \times 10^{-16}$ eV s is Planck's constant and $m = 9.1095 \times 10^{-28}$ g is the free-electron mass. The justification for using the one-electron Hamiltonian H will be presented in the next section. The crystal potential $V(\mathbf{x})$ is invariant under all the symmetry operations that leave the crystal invariant. We assume that the crystal potential $V(\mathbf{x})$ is constructed from a linear superposition of atomic potentials such as shown in Fig. 1, where both the free-atom potentials and the total potential are indicated. The crystal potential is atomiclike in the region near the ions and flattens out in the interstitial regions between the ions.

The purpose of the present section is to indicate how essential results can be derived from the group-theoretical analysis of the crystal Hamiltonian. The language introduced here will serve as the basis for the next section, which describes actual methods of solution. The discussion that follows is necessarily incomplete, and the interested reader is advised to consult standard texts on the applications of group-theoretical methods to quantum mechanics (Tinkham, 1964; Heine, 1960; Hamermesh, 1962). In the next section we justify the utilization of the one-electron Hamiltonian of Eq. (1), and present methods of solutions of Schrödinger's equation associated with the crystal Hamiltonian H.

1.2 Crystal Symmetry: Space Groups

The elements of symmetry relevant to the study of electronic states in crystalline solids include rotations, translations, and reflections. The ensemble of such symmetry oper-

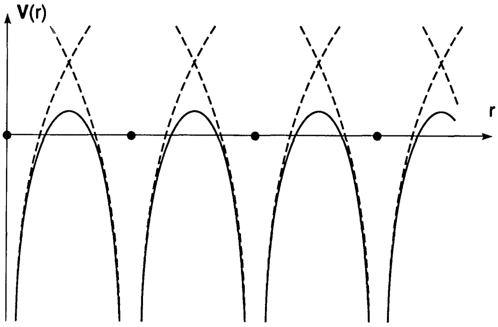

FIG. 1. Schematic representation of the crystal potential plotted along a line of ions (solid circles) in the solid. The crystal potential (solid line) is constructed from the individual ion potential (dashed line).

ations are called *space groups* (see CRYSTALLOGRAPHY). We describe below the essential features of space groups that allow us to derive the form of the crystal eigenfunctions. Such derivations can be found in standard solid-state physics texts (Callaway, 1974; Kittel, 1963; Ashcroft and Mermin, 1976).

An operation of symmetry with respect to the Hamiltonian operator $H(\mathbf{x})$ is defined as a linear transformation of the coordinates

$$\mathbf{x}'=\alpha\mathbf{x}+\mathbf{t}, \tag{2}$$

which leaves the Hamiltonian invariant,

$$H(\mathbf{x}')=H(\alpha\mathbf{x}+\mathbf{t})=H(\mathbf{x}). \tag{3}$$

In Eq. (2), α is a real 3×3 orthogonal (real unitary) matrix representing a (proper or improper) rotation and $\mathbf{t}$ is a translation vector. The symbol $\{\alpha|\mathbf{t}\}$ is used to denote an operator of the space group consisting of a rotation α and a translation $\mathbf{t}$. The unit operator is $\{\epsilon|\mathbf{0}\}$, where ϵ is a unit matrix.

We will not discuss in this article the very important subject of irreducible-representation theory (Tinkham, 1964; Heine, 1960). The basis of representation theory is the representation of a group in terms of matrices. This theory links group theory and quantum mechanics by introducing the concept of basis functions: The wave functions of a quantum system form basis functions for irreducible representations of the group of Schrödinger's equation, i.e., the group of operators which commute with the Hamiltonian, $H(\mathbf{x})$. Essentially, irreducible representation matrices indicate how basis functions transform among themselves under the operations of a group. Irreducible-representation theory can be used to classify the solutions of the Hamiltonian and is of central importance in the determination of the symmetry properties of the electronic states of the crystal: essential degeneracies, selection rules, etc.

The operator representing a pure lattice translation by $\mathbf{R}_j$ is denoted by $\{\epsilon|\mathbf{R}_j\}$, where the index j labels a particular lattice-translation vector. The set of lattice translations forms a group called the *lattice-translation group*. The lattice-translation group is Abelian, i.e., its members commute with each other. The commutativity property of the lattice-translation group implies that the irreducible representations of the lattice-translation group are one dimensional (Tinkham, 1964).

We formally define a space group as a group of operators of the form $\{\alpha|\mathbf{t}\}$ that contains an invariant subgroup of pure translations whose operators are of the form $\{\epsilon|\mathbf{R}_j\}$. In three dimensions, there are 230 possible space groups.

A space group is called symmorphic if the translation part $\mathbf{t}$ of all of the space-group operators $\{\alpha|\mathbf{t}\}$ can be written in the form

$$\mathbf{t}=n_1\mathbf{a}_1+n_2\mathbf{a}_2+n_3\mathbf{a}_3, \quad n_i=\text{integers} \tag{4}$$

where the vectors $\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3$ are noncoplanar lattice-translation vectors defining the space lattice. For a symmorphic space group, the set of operators $\{\alpha|\mathbf{0}\}$ that does not involve any translation forms a subgroup of the space group known as the *point group*. Therefore, we can symbolically write the following association:

$$\text{crystal space group} \rightarrow \text{lattice-translation group}+\text{point group}.$$

1.3 Direct and Reciprocal Space: Bravais Lattices and Brillouin Zones

A three-dimensional *Bravais lattice* consists of all points with position vectors $\mathbf{R}_j$ of the form of Eq. (4):

$$\mathbf{R}_j=n_1\mathbf{a}_1+n_2\mathbf{a}_2+n_3\mathbf{a}_3, \quad n_i=\text{integers} \tag{5}$$

where the vectors $\mathbf{a}_i$ are any three vectors not all in the same plane (Ashcroft and Mermin, 1976). The label j identifies a particular set of integers n_1,n_2,n_3. The vectors $\mathbf{a}_i (i=1,2,3)$ are called *primitive-lattice vectors*. The Bravais lattice specifies the periodic array in which the repeated units (atoms, groups of atoms, molecules, etc.) of the crystal are arranged and describes only the geometry of the underlying periodic structure. There are 14 different Bravais lattices in three dimensions. The 14 possible three-dimensional Bravais lattices are listed in Table 1. The Bravais lattices are classified into seven crystallographic systems corresponding to seven types of unit cell: triclinic, monoclinic, orthorhombic, tetragonal, cubic, trigonal, and hexagonal. The second column in Table 1 indicates the number of lattices per system. The third column indicates the axial and angular relations that describe the different cells.

When the lattice-translation vectors $\mathbf{a}_i$ $(i=1,2,3)$ are such that they define a unit cell

Table 1. Fourteen possible three-dimensional Bravais lattices. The Bravais lattices are classified into seven crystallographic systems corresponding to seven types of unit cell: triclinic, monoclinic, orthorhombic, tetragonal, cubic, trigonal, and hexagonal. The second column indicates the number of lattices per system. The third column defines the axial and angular relations that describe the different cells where α is the angle between $\mathbf{a}_2$ and $\mathbf{a}_3$, β is the angle between $\mathbf{a}_3$ and $\mathbf{a}_1$, and γ is the angle between $\mathbf{a}_1$ and $\mathbf{a}_2$.

System	Number of lattices	Restrictions on axes and angles
Triclinic	1	$a_1 \neq a_2 \neq a_3$ $\alpha \neq \beta \neq \gamma$
Monoclinic	2	$a_1 \neq a_2 \neq a_3$ $\alpha = \gamma = 90° \neq \beta$
Orthorhombic	4	$a_1 \neq a_2 \neq a_3$ $\alpha = \beta = \gamma = 90°$
Tetragonal	2	$a_1 = a_2 \neq a_3$ $\alpha = \beta = \gamma = 90°$
Cubic	3	$a_1 = a_2 = a_3$ $\alpha = \beta = \gamma = 90°$
Trigonal	1	$a_1 = a_2 = a_3$ $\alpha = \beta = \gamma < 120°, \neq 90°$
Hexagonal	1	$a_1 = a_2 \neq a_3$ $\alpha = \beta = 90°, \gamma = 120°$

containing exactly one lattice point, the cell is a *minimum-volume cell* and is called a *primitive cell.* A primitive cell is called a *Wigner–Seitz primitive cell* if it has the full symmetry of the Bravais lattice. A primitive cell, when translated through all the vectors $\mathbf{R}_j$ in a Bravais lattice, fills all space without overlapping. The volume of a Bravais lattice primitive cell is

$$\Omega_{\mathbf{x}} = |\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)|. \tag{6a}$$

Since, by definition, a primitive cell contains a single lattice point, the volume of a primitive cell is

$$\Omega_{\mathbf{x}} = V/N, \tag{6b}$$

where V is the total volume of the crystal and N is the total number of primitive cells (or lattice points) in the crystal.

From the lattice-translation vectors $\mathbf{a}_i$ ($i = 1,2,3$) we define a set of *primitive reciprocal-lattice vectors,*

$$\mathbf{b}_1 = 2\pi \frac{\mathbf{a}_2 \times \mathbf{a}_3}{\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)}, \tag{7a}$$

$$\mathbf{b}_2 = 2\pi \frac{\mathbf{a}_3 \times \mathbf{a}_1}{\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)}, \tag{7b}$$

$$\mathbf{b}_3 = 2\pi \frac{\mathbf{a}_1 \times \mathbf{a}_2}{\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)}. \tag{7c}$$

It can be verified that

$$\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}, \tag{7d}$$

where δ_{ij} is the Kronecker delta defined as

$$\delta_{ij} \equiv \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases} \tag{7e}$$

The volume of a reciprocal-lattice primitive cell is

$$\Omega_{\mathbf{k}} = \mathbf{b}_1 \cdot (\mathbf{b}_2 \times \mathbf{b}_3) = \frac{(2\pi)^3}{\Omega_{\mathbf{x}}}, \tag{8}$$

where $\Omega_{\mathbf{x}}$ is the volume of the Bravais-lattice primitive cell given by Eq. (6a).

It is convenient to define a set of *reciprocal-lattice vectors* $\mathbf{G}_i$ such that

$$\exp(i\mathbf{G}_i \cdot \mathbf{R}_j) = 1, \tag{9a}$$

and having solutions

$$\mathbf{G}_i = m_1\mathbf{b}_1 + m_2\mathbf{b}_2 + m_3\mathbf{b}_3, \quad m_j = \text{integers} \tag{9b}$$

where the label i identifies a particular set of integers m_1, m_2, m_3. The set of reciprocal-lattice vectors $\mathbf{G}_i$ generates a Bravais lattice in $\mathbf{k}$ space, just as the set of lattice translation vectors $\mathbf{R}_j$ generates the Bravais lattice of the crystal in $\mathbf{x}$ space. A Wigner–Seitz primitive cell in reciprocal space defined by the $\mathbf{b}_j$ ($j = 1,2,3$) is called the *first Brillouin zone.*

1.3.1 Example: Face-Centered-Cubic Lattice.

We illustrate the above definitions by considering the Bravais lattice and the reciprocal lattice associated with the face-centered-cubic lattice. The face-centered-cubic lattice is shown in Fig. 2(a) (Kittel, 1986) where the filled spheres indicate the lattice points defining the face-centered-cubic lattice and not the atomic coordinates. The symmetric set of primitive lattice-translation vectors for the face-centered-cubic lattice shown in Fig. 2(a) is

$$\mathbf{a}_1=\tfrac{1}{2}a(\hat{\mathbf{y}}+\hat{\mathbf{z}}), \tag{10a}$$

$$\mathbf{a}_2=\tfrac{1}{2}a(\hat{\mathbf{z}}+\hat{\mathbf{x}}), \tag{10b}$$

$$\mathbf{a}_3=\tfrac{1}{2}a(\hat{\mathbf{x}}+\hat{\mathbf{y}}), \tag{10c}$$

where a is the length of the cube edge. The primitive-lattice vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ connect the lattice point at the origin with lattice points at the face centers. The volume of the face-centered-cubic primitive cell is

$$\Omega_x=|\mathbf{a}_1\cdot(\mathbf{a}_2\times\mathbf{a}_3)|=\frac{a^3}{4}\,. \tag{11}$$

Since the primitive cell contains only one lattice point, the number of lattice points per volume is $4/a^3$ in the face-centered-cubic Bravais lattice. The primitive lattice-translation vectors of the lattice reciprocal to the face-centered-cubic lattice are found by using the construction of Eqs. (7):

$$\mathbf{b}_1=(2\pi/a)(-\hat{\mathbf{x}}+\hat{\mathbf{y}}+\hat{\mathbf{z}}), \tag{12a}$$

$$\mathbf{b}_2=(2\pi/a)(\hat{\mathbf{x}}-\hat{\mathbf{y}}+\hat{\mathbf{z}}), \tag{12b}$$

$$\mathbf{b}_3=(2\pi/a)(\hat{\mathbf{x}}+\hat{\mathbf{y}}-\hat{\mathbf{z}}). \tag{12c}$$

The first Brillouin zone of the face-centered-cubic lattice is the truncated octahedron confined by the planes normal to the reciprocal-lattice vectors,

$$\mathbf{G}=(2\pi/a)(\pm\hat{\mathbf{x}}\pm\hat{\mathbf{y}}\pm\hat{\mathbf{z}}), \tag{13}$$

at their midpoints. The corners of the octahedron are cut by the planes that are the perpendicular bisectors of six other reciprocal-lattice vectors,

$$\mathbf{G}=(2\pi/a)(\pm 2\hat{\mathbf{x}}),$$

$$\mathbf{G}=(2\pi/a)(\pm 2\hat{\mathbf{y}}), \tag{14}$$

$$\mathbf{G}=(2\pi/a)(\pm 2\hat{\mathbf{z}}).$$

The first Brillouin zone of the face-centered-cubic lattice is shown in Fig. 2(b) (Kittel, 1986).

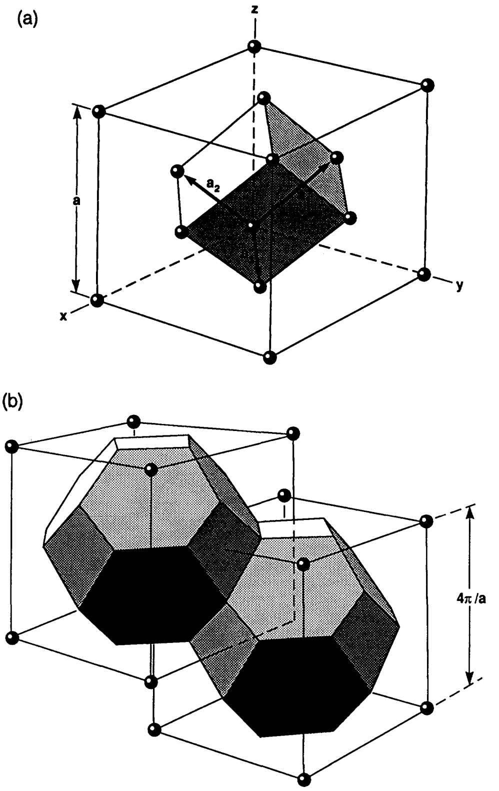

FIG. 2. (a) Primitive basis vectors of the face-centered-cubic lattice. The primitive lattice vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ connect the lattice point at the origin with lattice points at the face centers. The filled spheres indicate the lattice points defining the face-centered-cubic lattice and not atomic coordinates. (b) Brillouin zones of the face-centered-cubic lattice. Two Brillouin zones are shown. (Reproduced with permission from Kittel, 1986, pp. 42 and 43.)

1.3.2 Relation with X-Ray Diffraction by a Crystal.

We now indicate the direct relationship between the construction of the reciprocal space, particularly as described by Eq. (9a), and the condition for constructive interference deduced from the formulation of x-ray diffraction by a crystal (Ashcroft and Mermin, 1976). The condition for constructive interference derived below is closely related to the problem of the scattering of electron at the Brillouin-zone boundary by a weak periodic potential treated in the next section.

When x-ray radiation is scattered from a crystalline solid, intense peaks (known as *Bragg peaks*) are observed for certain sharply defined wavelengths and incident directions. Our purpose is to determine the relationship between the atomic coordinates in the solid and the properties of the incoming x-ray radiation (energy and direction of propagation) giving rise to constructive interference. To determine the condition for constructive interference, we first consider two scatterers separated by a displacement vector **d**. The incoming x ray is incident from very far away and is characterized by a wavelength λ and a wave vector $\mathbf{k}=2\pi\hat{\mathbf{n}}/\lambda$, where $\hat{\mathbf{n}}$ defines the incident direction of propagation. The elastically scattered radiation is observed along the direction $\hat{\mathbf{n}}'$ and is characterized by the same wavelength λ and a wave vector $\mathbf{k}'=2\pi\hat{\mathbf{n}}'/\lambda$. A schematic representation of the geometry is indicated in Fig. 3 and reveals that the path difference between the rays scattered by each of the two ions is

$$d(\cos\theta+\cos\theta')=\mathbf{d}\cdot(\hat{\mathbf{n}}-\hat{\mathbf{n}}'). \tag{15}$$

The condition for constructive interference is that this path difference equals an integral number of wavelengths:

$$\mathbf{d}\cdot(\hat{\mathbf{n}}-\hat{\mathbf{n}}')=m\lambda, \tag{16}$$

where m is an integer. The condition of constructive interference for the wave vectors **k** (incident) and **k**′ (scattered) is found from Eq. (16) by multiplying both sides by $2\pi/\lambda$,

$$\mathbf{d}\cdot(\mathbf{k}-\mathbf{k}')=2\pi m, \tag{17}$$

for integral values of m.

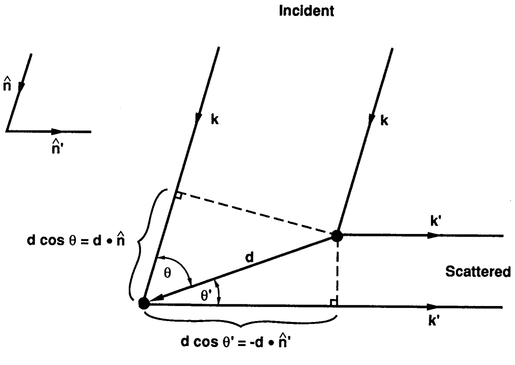

FIG. 3. Geometrical construction indicating that the path difference between incident (**k**) and scattered (**k**′) radiation is given by $d(\cos\theta+\cos\theta')=\mathbf{d}\cdot(\hat{\mathbf{n}}-\hat{\mathbf{n}}')$. (Reproduced with permission from Ashcroft and Mermin, 1976, p. 98.)

We now extend the condition for constructive interference, Eq. (17), to an array of scatterers located at the sites of a Bravais lattice. Since the displacment vectors between the scatterers are simply the Bravais-lattice vectors $\mathbf{R}_j$, the condition that all scattered rays interfere constructively is that the condition Eq. (17) holds simultaneously for all displacement vectors $\mathbf{d} = \mathbf{R}_j$, i.e.,

$$\mathbf{R}_j\cdot(\mathbf{k}-\mathbf{k}')=2\pi m, \tag{18}$$

for integral values of m and all Bravais-lattice vectors $\mathbf{R}_j$. Equation (18) can also be written as

$$\exp[i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{R}_j]=1, \tag{19}$$

for all Bravais-lattice vectors $\mathbf{R}_j$. This relation [Eq. (19)] is called the von Laue condition. A comparison between this condition and the definition of the reciprocal lattice, Eq. (9a), reveals that constructive interference will occur provided that the change of wave vector upon scattering is a reciprocal- lattice vector, $\mathbf{k}-\mathbf{k}'=\mathbf{G}$.

It is somewhat more convenient to use an alternative form of the von Laue condition entirely in terms of the incident wave vector **k**. This is done by noting that the reciprocal lattice is a Bravais lattice and, consequently, if $\mathbf{k}'-\mathbf{k}$ is a reciprocal-lattice vector, so is $\mathbf{k}-\mathbf{k}'=\mathbf{G}$. The condition that the wave vectors **k** and **k**′ have the same magnitude (elastic scattering) is

$$|\mathbf{k}|=|\mathbf{k}'|=|\mathbf{k}-\mathbf{G}|. \tag{20a}$$

Squaring both sides of Eq. (20a) yields,

$$\mathbf{k}\cdot\hat{\mathbf{G}}=\tfrac{1}{2}G, \tag{20b}$$

where $\hat{\mathbf{G}}=\mathbf{G}/G$. Thus an incident wave vector **k** will satisfy the condition for constructive interference if the component of **k** along the reciprocal-lattice vector **G** is half the length of **G** [Eq. (20b)] or, equivalently, if the tip of the vector **k** lies in a plane that is the perpendicular bisector of a line joining the origin of the reciprocal space to a reciprocal-lattice point **G**. Such planes in **k** space are called *Bragg planes*. A realization of this condition is indicated in Fig. 4.

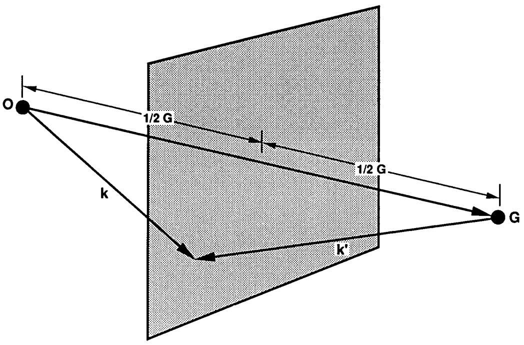

FIG. 4. Schematic representation of the von Laue condition for constructive interference. The conditions $\mathbf{k}-\mathbf{k}'=\mathbf{G}$ and $|\mathbf{k}|=|\mathbf{k}'|$ are simultaneously satisfied when the tip of **k** is equidistant from the origin *O* and the tip of the reciprocal-lattice vector **G**. The wave vector **k** therefore lies in the plane bisecting the line joining the origin *O* and the tip of **G**. Expressed in terms of the incident wave vector **k**, constructive interference occurs when the component of the incident wave vector **k** along the reciprocal-lattice vector **G** is half the length of **G**, or $\mathbf{k}\cdot\hat{\mathbf{G}}=\frac{1}{2}G$. (Reproduced with permission from Ashcroft and Mermin, 1976, p. 99.)

1.4 Form of Electron Solutions in Crystalline Solids: Bloch States

We now provide a derivation of Bloch's theorem, which is a result of central importance in the theory of solids. This result can be obtained by considering irreducible representations of the translation group. Under a pure translation of the form $\{\epsilon|\mathbf{R}_j\}$, the electron wave functions in a crystal must transform according to the irreducible representations of the lattice-translation group. As indicated in Sec. 1.2, the lattice-translation group is Abelian and, consequently, its irreducible representations are one dimensional (Tinkham, 1964). We show below that the one-dimensional irreducible representation of the lattice-translation group are labeled by a wave vector **k**, which classifies the Bloch solutions of the crystal Hamiltonian.

1.4.1 Proof of Bloch's Theorem. The central approximation in the one-electron band approximation is that the state of a solid can be specified by calculating the wave function of each individual electron. Although in principle only the full many-electron wave function is well defined, the one-electron description is a very useful concept. We assume that these one-electron wave functions are solutions of an effective (unspecified) Hamiltonian *H*, which commutes with all the operators $\{\alpha|\mathbf{t}\}$ of the space group (Callaway, 1974),

$$[H,\{\alpha|\mathbf{t}\}]=0, \tag{21}$$

where the commutator of two operators *A* and *B* is defined as $[A,B]\equiv AB-BA$. We define the translation operator

$$T(\mathbf{R}_j)\equiv\{\epsilon|-\mathbf{R}_j\}, \tag{22}$$

and define its action on a wavefunction $\psi(\mathbf{x})$ as

$$T(\mathbf{R}_j)\psi(\mathbf{x})=\psi(\mathbf{x}+\mathbf{R}_j). \tag{23}$$

The operators $T(\mathbf{R}_j)$ commute with each other,

$$[T(\mathbf{R}_j),T(\mathbf{R}_i)]=0, \tag{24a}$$

and with the Hamiltonian *H*,

$$[T(\mathbf{R}_j),H]=0. \tag{24b}$$

As a result, the one-electron wave functions $\psi(\mathbf{x})$ can be chosen to be simultaneously eigenfunctions of the Hamiltonian *H* and of the translation operator $T(\mathbf{R}_j)$ (Sakurai, 1985):

$$T(\mathbf{R}_j)\psi(\mathbf{x})=\psi(\mathbf{x}+\mathbf{R}_j)=\lambda(\mathbf{R}_j)\psi(\mathbf{x}), \tag{25a}$$

$$H\psi(\mathbf{x})=\epsilon_j\psi(\mathbf{x}), \tag{25b}$$

where $\lambda(\mathbf{R}_j)$ is the eigenvalue associated with the translation operator $T(\mathbf{R}_j)$ and ϵ_j is the energy eigenvalue associated with the Hamiltonian *H*. The eigenvalue $\lambda(\mathbf{R}_j)$ describes the action of $T(\mathbf{R}_j)$ on $\psi(\mathbf{x})$ and must be a complex number with unity modulus if we require the (one-dimensional) representation to be unitary, i.e., that the norm of $\psi(\mathbf{x})$ remains invariant under the action of $T(\mathbf{R}_j)$ (Callaway, 1974).

The form of the eigenvalue $\lambda(\mathbf{R}_j)$ can be derived by noting that the successive action of $T(\mathbf{R}_j)$ and $T(\mathbf{R}_i)$ corresponds to a single translation $T(\mathbf{R}_i+\mathbf{R}_j)$,

$$\begin{aligned}T(\mathbf{R}_i)T(\mathbf{R}_j)\psi(\mathbf{x})&=\psi(\mathbf{x}+\mathbf{R}_j+\mathbf{R}_i)\\&=\lambda(\mathbf{R}_j)\lambda(\mathbf{R}_i)\psi(\mathbf{x})\\&=\lambda(\mathbf{R}_j+\mathbf{R}_i)\psi(\mathbf{x}).\end{aligned} \tag{26}$$

It follows that

$$\lambda(\mathbf{R}_j)\lambda(\mathbf{R}_i)=\lambda(\mathbf{R}_j+\mathbf{R}_i). \tag{27}$$

The lattice-translation vector $\mathbf{R}_j$ is of the form of Eq. (5),

$$\mathbf{R}_j=n_1\mathbf{a}_1+n_2\mathbf{a}_2+n_3\mathbf{a}_3,\quad n_i=\text{integers} \tag{28}$$

where the vectors $\mathbf{a}_i$ ($i=1,2,3$) are primitive lattice-translation vectors. The complex eigenvalue $\lambda(\mathbf{R}_j)$ has unit modulus and can be taken to be of the form

$$\lambda(\mathbf{a}_i)=\exp(2\pi i x_i), \quad i=1,2,3. \tag{29}$$

We show below that the numbers x_i are related to a wave vector **k** used to classify the (one-dimensional) irreducible representations of the translation group.

Using the multiplication rule Eq. (27) resulting from successive applications of the translation operator $T(\mathbf{R}_j)$, we deduce

$$\begin{aligned}\lambda(\mathbf{R}_j)&=\lambda(n_1\mathbf{a}_1+n_2\mathbf{a}_2+n_3\mathbf{a}_3)\\&=\lambda(\mathbf{a}_1)^{n_1}\lambda(\mathbf{a}_2)^{n_2}\lambda(\mathbf{a}_3)^{n_3},\end{aligned} \tag{30a}$$

which can be written in the form indicated by Eq. (29),

$$\lambda(\mathbf{R}_j)=\exp(i\mathbf{k}\cdot\mathbf{R}_j), \tag{30b}$$

where **k** is an arbitrary wave vector that is the same for each translation. The eigenvalue $\lambda(\mathbf{R}_j)$ appearing in Eq. (30b) can be written in terms of the reciprocal-lattice primitive vectors $\mathbf{b}_j$ ($j=1,2,3$) as

$$\mathbf{k}=x_1\mathbf{b}_1+x_2\mathbf{b}_2+x_3\mathbf{b}_3, \tag{31}$$

by using the orthonormality relations $\mathbf{a}_i\cdot\mathbf{b}_j = 2\pi\delta_{ij}$. The coefficients x_i specify the allowed values of **k** and are determined by imposing suitable boundary conditions.

The wave vector **k** appearing in Eq. (30b) labels the wave function of an electron in a crystal, which we write as $\psi(\mathbf{k},\mathbf{x})$ and call a *Bloch function*. The energy eigenvalue corresponding to the crystal state $\psi(\mathbf{k},\mathbf{x})$ associated with **k** is denoted by $\epsilon(\mathbf{k})$. Combining Eqs. (25a) and (30b) leads to the formulation of Bloch's theorem:

$$T(\mathbf{R}_j)\psi(\mathbf{k},\mathbf{x})=\psi(\mathbf{k},\mathbf{x}+\mathbf{R}_j)=\exp(i\mathbf{k}\cdot\mathbf{R}_j)\psi(\mathbf{k},\mathbf{x}). \tag{32}$$

In the language of group theory, we say that the wave vector **k** labels the one-dimensional irreducible representations $\exp(i\mathbf{k}\cdot\mathbf{R}_j)$ of the translation group, or

$$\{\epsilon|-\mathbf{R}_j\} \text{ is represented by } \exp(i\mathbf{k}\cdot\mathbf{R}_j). \tag{33}$$

Equation (32) indicates the form of a Bloch state in a crystal and is the central result of this section. The form of Eq. (32) implies that the charge density remains unchanged by a lattice translation $\mathbf{R}_j$, i.e.,

$$|\psi(\mathbf{k},\mathbf{x}+\mathbf{R}_j)|^2=|\psi(\mathbf{k},\mathbf{x})|^2. \tag{34}$$

It is convenient to write the Bloch functions $\psi(\mathbf{k},\mathbf{x})$ as a product of phase factor $\exp(i\mathbf{k}\cdot\mathbf{x})$ and a *cell-periodic function* $u(\mathbf{k},\mathbf{x})$ defined as

$$\psi(\mathbf{k},\mathbf{x})=\exp(i\mathbf{k}\cdot\mathbf{x})u(\mathbf{k},\mathbf{x}), \tag{35a}$$

where the cell-periodic function $u(\mathbf{k},\mathbf{x})$ has the periodicity of the lattice,

$$u(\mathbf{k},\mathbf{x}+\mathbf{R}_j)=u(\mathbf{k},\mathbf{x}). \tag{35b}$$

1.4.2 Allowed Values of k: Periodic Boundary Conditions. We now impose boundary conditions on the wave functions $\psi(\mathbf{k},\mathbf{x})$ to determine the possible values of the wave vector **k** of the form given by Eq. (31). The possible values of **k** are determined by the application of *periodic boundary conditions.*

The lattice translation vector $\mathbf{R}_j$ is of the form of Eq. (5),

$$\mathbf{R}_j=n_1\mathbf{a}_1+n_2\mathbf{a}_2+n_3\mathbf{a}_3, \quad n_i=0,\ldots,N_i, \tag{36}$$

where we assume that the crystal has finite dimensions $N_1\mathbf{a}_1$, $N_2\mathbf{a}_2$, $N_3\mathbf{a}_3$. The integers N_i are large but finite. The total number of primitive cells in the crystal is $N=N_1N_2N_3$. The application of periodic boundary conditions requires that

$$\begin{aligned}&T(N_i\mathbf{a}_i)\psi(\mathbf{k},\mathbf{x})\\&\quad\equiv\psi(\mathbf{k},\mathbf{x}+N_i\mathbf{a}_i)\\&\quad=\psi(\mathbf{k},\mathbf{x}), \quad i=1,2,3.\end{aligned} \tag{37}$$

By application of Bloch's theorem, we deduce

$$\begin{aligned}&\psi(\mathbf{k},\mathbf{x}+N_i\mathbf{a}_i)\\&\quad=\exp[i\mathbf{k}\cdot(N_i\mathbf{a}_i)]\psi(\mathbf{k},\mathbf{x}), \quad i=1,2,3\end{aligned} \tag{38a}$$

which requires that

$$\exp[i\mathbf{k}\cdot(N_i\mathbf{a}_i)]=1, \quad i=1,2,3. \tag{38b}$$

If **k** has the form of Eq. (31), we must have

$$\exp[2\pi i(x_iN_i)]=1, \quad i=1,2,3, \tag{38c}$$

or

$$x_i=m_i/N_i, \quad m_i=0,\ldots,N_i. \tag{38d}$$

It follows that the possible values of **k** allowed by the periodic boundary conditions are of the form

$$\mathbf{k}=\frac{m_1}{N_1}\mathbf{b}_1+\frac{m_2}{N_2}\mathbf{b}_2+\frac{m_3}{N_3}\mathbf{b}_3, \quad m_i=0,\ldots,N_i. \tag{39}$$

Since the integers N_i are very large numbers, the allowed values of **k** are very closely spaced. The volume in **k** space per allowed value of **k** corresponds to the parallelepiped with edges $\mathbf{b}_i/N_i$ $(i=1,2,3)$ defined by Eq. (39) and is given by (Ashcroft and Mermin, 1976)

$$\Delta\mathbf{k}=\frac{\mathbf{b}_1}{N_1}\cdot\left(\frac{\mathbf{b}_2}{N_2}\times\frac{\mathbf{b}_3}{N_3}\right)=\frac{1}{N}\,\mathbf{b}_1\cdot(\mathbf{b}_2\times\mathbf{b}_3)=\frac{\Omega_\mathbf{k}}{N}, \tag{40a}$$

or

$$\Omega_\mathbf{k}=N\Delta\mathbf{k}, \tag{40b}$$

where $\Omega_\mathbf{k}=\mathbf{b}_1\cdot(\mathbf{b}_2\times\mathbf{b}_3)=(2\pi)^3/\Omega_\mathbf{x}$ is the volume of the reciprocal-lattice primitive cell [Eq. (8)] and $N=N_1N_2N_3$. It follows from Eq. (40b) that *the number of allowed wave vectors in the primitive cell of the reciprocal space is equal to the number of lattice sites (or primitive cells) in the crystal.* By combining Eqs. (40a) and (8), the volume per allowed value of **k** can also be written as

$$\Delta\mathbf{k}=\frac{\Omega_\mathbf{k}}{N}=\frac{(2\pi)^3}{N\Omega_\mathbf{x}}=\frac{(2\pi)^3}{V}, \tag{41}$$

where $V=N\Omega_\mathbf{x}$ is the total volume of the crystal and N is the total number of primitive cells in the crystal.

1.5 General Properties of Bloch States and Energy Bands

In this section, we briefly describe the symmetry properties of the crystal Bloch functions $\psi(\mathbf{k},\mathbf{x})$ and of their associated energy eigenvalue spectrum $\epsilon(\mathbf{k})$. We first discuss certain general properties of the Bloch states arising from the form of the crystal Hamiltonian and from Bloch's theorem. We then describe the additional symmetry properties of the crystal Bloch states due to time-reversal symmetry. Most of the discussions which follows can be found in general solid-state physics texts (Kittel, 1963; Ashcroft and Mermin, 1976).

1.5.1 General Properties of Bloch States. We have seen before that the application of Bloch's theorem led to the introduction of a crystal wave vector **k** labeling the Bloch states $\psi(\mathbf{k},\mathbf{x})$, which we write as a product of phase factor $\exp(i\mathbf{k}\cdot\mathbf{x})$ and a cell-periodic function $u(\mathbf{k},\mathbf{x})$,

$$\psi(\mathbf{k},\mathbf{x})=\exp(i\mathbf{k}\cdot\mathbf{x})u(\mathbf{k},\mathbf{x}). \tag{42}$$

The crystal wave vector **k** plays the same role in the description of the motion of an electron in a periodic potential as does the free-electron wave vector $\mathbf{k}_\text{free}$ in describing the motion of a free electron described by a plane wave of the form $\exp(i\mathbf{k}_\text{free}\cdot\mathbf{x})$. In the case of a free electron, the wave vector $\mathbf{k}_\text{free}$ is related to the momentum **p** by $\mathbf{k}_\text{free}=\mathbf{p}/\hbar$. However, for a Bloch state the wave vector **k** is not proportional to the electronic momentum. In fact, it can be directly shown that a Bloch state $\psi(\mathbf{k},\mathbf{x})$ is not an eigenstate of the momentum operator $\mathbf{p}=-i\hbar\nabla$ (Ashcroft and Mermin, 1976),

$$\mathbf{p}\psi(\mathbf{k},\mathbf{x})\equiv\frac{\hbar}{i}\,\nabla[\exp(i\mathbf{k}\cdot\mathbf{x})u(\mathbf{k},\mathbf{x})] \tag{43a}$$

$$=\hbar\mathbf{k}\psi(\mathbf{k},\mathbf{x})+\exp(i\mathbf{k}\cdot\mathbf{x})\frac{\hbar}{i}\,\nabla u(\mathbf{k},\mathbf{x}). \tag{43b}$$

As a natural extension of the free-electron case, however, the quantity $\hbar\mathbf{k}$ is called the *crystal momentum.* The wave vector **k** should be interpreted as a quantum number characterizing the translational symmetry of the crystal potential [or, equivalently, labeling the one-dimensional irreducible representations $\exp(i\mathbf{k}\cdot\mathbf{R}_j)$ of the translation group] just as the momentum **p** is a quantum number characterizing the full translational symmetry of free space (Ashcroft and Mermin, 1976).

Since the crystal wave vector **k** is used as an index (or label) for the Bloch solutions $\psi(\mathbf{k},\mathbf{x})$, it serves as a parameter that enters the crystal Hamiltonian of Eq. (1), i.e., $H\equiv H(\mathbf{k})$. For each value of the parameter **k**, the many possible eigensolutions and eigenenergies of the Schrödinger equation for the crystal are labeled by a *band index n*, and we write

$$\psi(\mathbf{k},\mathbf{x})\rightarrow\psi_n(\mathbf{k},\mathbf{x}), \tag{44a}$$

and

$$\epsilon(\mathbf{k})\rightarrow\epsilon_n(\mathbf{k}). \tag{44b}$$

For a given band index n, the set of energy levels $\epsilon_n(\mathbf{k})$ is called an *energy band.* The *bandwidth* is the energy spread between the minimum and maximum energy in a band n. The energy $\epsilon_n(\mathbf{k})$ is a continuous function of the wave vector, which has the form of Eq.

(39) imposed by the application of cyclic boundary conditions:

$$\mathbf{k}=\frac{m_1}{N_1}\mathbf{b}_1+\frac{m_2}{N_2}\mathbf{b}_2+\frac{m_3}{N_3}\mathbf{b}_3,\quad m_i=0,\ldots,N_i. \tag{45}$$

We now examine the periodic properties of the crystal electronic states and energy levels. For given band n, the energy $\epsilon_n(\mathbf{k})$ associated with a Bloch state $\psi_n(\mathbf{k},\mathbf{x})$ is periodic in reciprocal space, i.e.,

$$\epsilon_n(\mathbf{k}+\mathbf{G})=\epsilon_n(\mathbf{k}), \tag{46}$$

where $\mathbf{G}$ is a reciprocal-lattice vector of the form of Eq. (9b). The periodicity of the function $\epsilon(\mathbf{k})$ can be demonstrated by writing a Bloch state $\psi(\mathbf{k},\mathbf{x})$ in the form (Kittel, 1963)

$$\psi_n(\mathbf{k},\mathbf{x})=\exp(i\mathbf{k}\cdot\mathbf{x})u_n(\mathbf{k},\mathbf{x}) \tag{47a}$$

$$=\exp[i(\mathbf{k}+\mathbf{G})\cdot\mathbf{x}]\times[\exp(-i\mathbf{G}\cdot\mathbf{x})u_n(\mathbf{k},\mathbf{x})] \tag{47b}$$

$$=\exp[i(\mathbf{k}+\mathbf{G})\cdot\mathbf{x}]\times u_n(\mathbf{k}+\mathbf{G},\mathbf{x}) \tag{47c}$$

$$\equiv\psi_n(\mathbf{k}+\mathbf{G},\mathbf{x}) \tag{47d}$$

where the cell-periodic function defined as

$$u_n(\mathbf{k}+\mathbf{G},\mathbf{x})\equiv\exp(-i\mathbf{G}\cdot\mathbf{x})u_n(\mathbf{k},\mathbf{x}) \tag{48}$$

has the periodicity of the direct lattice since

$$\exp(i\mathbf{G}\cdot\mathbf{R})=1, \tag{49}$$

for all reciprocal-lattice vectors $\mathbf{G}$ and lattice-translation vectors $\mathbf{R}$. From the fact that $\psi_n(\mathbf{k}+\mathbf{G},\mathbf{x})$ can be constructed from $\psi_n(\mathbf{k},\mathbf{x})$, one deduces that $\epsilon_n(\mathbf{k}+\mathbf{G})=\epsilon_n(\mathbf{k})$. The periodicity of the energy eigenvalue $\epsilon_n(\mathbf{k})$ implies that the wave vector $\mathbf{k}$ introduced in Bloch's theorem can be chosen to lie in the first Brillouin zone since a wave vector $\mathbf{k}'$ not lying in the first Brillouin zone can be translated back within the first Brillouin zone by a reciprocal-lattice vector $\mathbf{G}$,

$$\mathbf{k}'=\mathbf{k}+\mathbf{G}, \tag{50}$$

where the wave vector $\mathbf{k}$ lies within the first Brillouin zone.

From the above discussion, we see that the *electronic band structure* of a crystalline solid is described in terms of a family of energy band functions $\{\epsilon_n(\mathbf{k})\}$ that are continuous in the wave-vector parameter $\mathbf{k}$ and periodic with the periodicity of the reciprocal space. We state without proof that the mean velocity $\mathbf{v}_n(\mathbf{k})$ of an electron associated with a band n at a wave vector $\mathbf{k}$ is proportional to the derivative of the energy function $\epsilon_n(\mathbf{k})$ with respect to $\mathbf{k}$ (Ashcroft and Mermin, 1976),

$$\mathbf{v}_n(\mathbf{k})=\frac{1}{\hbar}\left(\frac{\partial\epsilon_n(\mathbf{k})}{\partial\mathbf{k}}\right)\equiv\frac{1}{\hbar}\nabla_{\mathbf{k}}\epsilon_n(\mathbf{k}). \tag{51}$$

1.5.2 Energy Bands and Energy-Band Gap. Having described in somewhat general terms the mathematical properties of the crystal Bloch states $\psi_n(\mathbf{k},\mathbf{x})$ and their associated energy bands $\epsilon_n(\mathbf{k})$, we now use a simple physical picture to discuss how these quantities determine the electronic properties of a solid. The discussion which follows is qualitative and can be found, in one form or another, in general solid-state physics texts (Kittel, 1986; Ashcroft and Mermin, 1976).

We have seen that the solutions to the one-electron crystal Hamiltonian, Eq. (1), are the Bloch functions $\psi(\mathbf{k},\mathbf{x})$ of the form given by Eq. (42) and associated with the energy eigenvalue $\epsilon(\mathbf{k})$. Consequently, we can write the crystal Schrödinger equation in the form

$$[-(\hbar^2/2m)\nabla^2+V(\mathbf{x})]\psi_n(\mathbf{k},\mathbf{x})=\epsilon_n(\mathbf{k})\psi_n(\mathbf{k},\mathbf{x}), \tag{52}$$

where, as in Eqs. (44), we have classified the crystal eigenfunctions $\psi_n(\mathbf{k},\mathbf{x})$ and the eigenvalues $\epsilon_n(\mathbf{k})$ according to the band index n for a given value of the crystal wave vector $\mathbf{k}$. The single most important feature of Eq. (52) is that it leads to two classes of solutions depending on whether the Bloch state $\psi_n(\mathbf{k},\mathbf{x})$ has a propagating or evanescent character. For a fixed (real) value of the energy $\epsilon_n(\mathbf{k})$, a propagating Bloch state corresponds to a real value of the crystal wave vector $\mathbf{k}$, whereas an evanescent solution is associated with a complex value of $\mathbf{k}$. As a result, the energy eigenvalue spectrum $\epsilon_n(\mathbf{k})$ consists of sets of energy bands supporting wavelike (real values of $\mathbf{k}$) Bloch states, separated by forbidden energy regions, where the solutions are evanescent (complex values of $\mathbf{k}$) and do not propagate throughout the solid. This feature is the single most important effect arising from the presence of the periodic crystal potential $V(\mathbf{x})$ in Eq. (52). The magnitude of these forbidden energy regions separating the energy bands is of decisive importance in determining whether the solid is a metal, semiconductor, or insulator. This classification is discussed below.

We now illustrate the concept of energy bands and the physical origin of the band gap. We have seen that Bragg reflection is a characteristic feature of wave propagation in crystals. Bragg reflection of electron waves in crystals is the cause of energy gaps since, as we show below, wavelike solutions of the Schrödinger equation do not exist at Bragg reflection. This property can be most directly demonstrated by introducing the *nearly-free-electron model* (Kittel, 1986; Ashcroft and Mermin, 1976), in which we consider the effect of the crystal potential $V(\mathbf{x})$ on the otherwise free-electron solutions. For the free-electron Hamiltonian, we have $V(\mathbf{x}) \equiv 0$, and the solutions are plane waves,

$$\psi(\mathbf{k},\mathbf{x}) \sim \exp(i\mathbf{k}\cdot\mathbf{x}), \tag{53a}$$

corresponding to the free-electron energy,

$$\epsilon(\mathbf{k}) = (\hbar^2/2m)\mathbf{k}^2. \tag{53b}$$

Now consider the effect of a weak periodic potential $V(\mathbf{x})$ on the free-electron solutions. For simplicity we assume a one-dimensional model describing a linear solid of lattice constant a. The first Brillouin zone consists of the region $-\pi/a < k < \pi/a$, and the set of one-dimensional reciprocal-lattice vectors is of the form

$$G = n(2\pi/a), \quad n = \text{integer}. \tag{54}$$

Since the one-dimensional potential is periodic with the periodicity of the lattice,

$$V(x) = V(x+a), \tag{55}$$

a Fourier expansion of $V(x)$ will only contain Fourier components associated with the reciprocal-lattice vectors G,

$$V(x) = \sum_G V(G)\exp(iGx). \tag{56}$$

Consequently, the potential $V(x)$ can only admix Bloch solutions $\psi(k,x)$ whose values of k differ by a multiple of G. In the language of elastic scattering theory in a periodic potential introduced in Sec. 1.3.2, this phenomenon is described by saying that electrons are *Bragg reflected* when their wave vector $\mathbf{k}$ satisfies the condition for constructive interference [Eq. (20b)] (Kittel, 1986),

$$\mathbf{k}\cdot\hat{\mathbf{G}} = \tfrac{1}{2}G. \tag{57}$$

In one dimension, Bragg reflections occur at values of k given by

$$k = \pm\tfrac{1}{2}G = n(\pi/a), \quad n = \text{integer}. \tag{58}$$

The first Bragg reflection occurs at $k = \pm\pi/a$ and arises because the electron reflected from one atom in the linear lattice interferes constructively with a phase difference of 2π with the electron reflected from a nearest-neighbor atom. As a result, the first *energy gap* E_g opens up at $k \pm (\pi/a)$ and is due to the admixture of the free-electron solutions by the Fourier component of the periodic potential $V(\frac{1}{2}G)$ at the Brillouin-zone boundary. The magnitude of the energy gap can be shown (Ashcroft and Mermin, 1976) to be equal to the magnitude of the admixing Fourier component of the potential, $E_g = |V(\frac{1}{2}G)|$. Other energy gaps occur at other values of the integer n in Eq. (58). The opening of a gap in a free-electron spectrum due to a periodic potential is schematically indicated in Fig. 5.

We now discuss the physical origin of the energy gap. At the values $k = \pm\pi/a$, the solutions correspond to equal admixture of waves

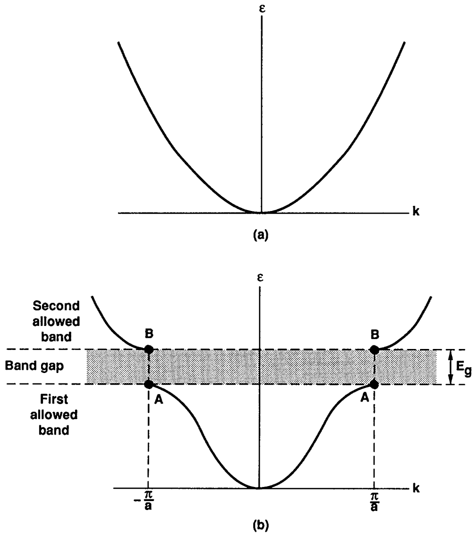

FIG. 5. (a) Parabolic dispersion relation between the energy ϵ and the wave vector k for a free electron, $\epsilon(k) \sim k^2$. (b) Dispersion relation between the energy ϵ and the wave vector k for an electron in a monoatomic linear lattice of lattice constant a. The energy gap E_g separating the first and second allowed energy bands is associated with the first Bragg reflection at $k = \pm\pi/a$. Other gaps are found at higher energies of $k = \pm n\pi/a$, where n is an integer. (Reproduced with permission from Kittel, 1986, p. 161.)

traveling to the right and to the left because the direction of travel is reversed upon each Bragg reflection (Kittel, 1986). We construct symmetric and antisymmetric standing-wave solutions at $k=\pm\pi/a$ from the plane waves $\exp[\pm i(\pi/a)x]$ as follows:

$$\psi^{(+)}(k_0,x)=\exp(ik_0x)+\exp(-ik_0x)$$
$$=2\cos(k_0x), \qquad (59a)$$

$$\psi^{(-)}(k_0,x)=\exp(ik_0x)-\exp(-ik_0x)$$
$$=2i\sin(k_0x), \qquad (59b)$$

where $k_0=\pi/a$. The standing-wave solution $\psi^{(+)}(k_0,x)$ is even under space inversion, whereas the standing-wave solution $\psi^{(-)}(k_0,x)$ is odd under space inversion.

We can now understand the physical origin of the energy gap by considering a simple electrostatic argument based on the electronic charge distribution generated by the symmetric and antisymmetric standing-wave solutions $\psi^{(+)}(k_0,x)$ and $\psi^{(-)}(k_0,x)$, respectively, at the Brillouin-zone boundary compared to the charge density associated with the free-electron solution. The electronic charge density arising from a free-electron solution is simply a constant independent of position,

$$n_0=|\exp(\pm ik_0x)|^2=1. \qquad (60a)$$

However, the symmetric standing-wave solution gives rise to an electronic charge density

$$n^{(+)}(x)=|\psi^{(+)}(k_0,x)|^2\sim\cos^2[(\pi/a)x], \qquad (60b)$$

which concentrates negative charge *on* the lattice sites occupied by the positive ion cores where the potential energy (electron scale) is lowest. The antisymmetric standing-wave solution gives rise to an electronic charge density

$$n^{(-)}(x)=|\psi^{(-)}(k_0,x)|^2\sim\sin^2[(\pi/a)x], \qquad (60c)$$

which concentrates negative charge midway *between* the lattice sites occupied by the positive ion cores. Sometimes the two linear combinations of Eqs. (59) are called s-like ($|\psi^{(+)}|^2\sim\cos^2[(\pi/a)x]$) or p-like ($|\psi^{(-)}|^2\sim\sin^2[(\pi/a)x]$). The s-like combination is maximum at the ion-core site, like an atomic s orbital. On the other hand, the p-like combination vanishes as the square of the distance from the ion site for small distances, like an atomic p orbital. The one-dimensional electron charge densities n_0, $n^{(+)}(x)$, and $n^{(-)}(x)$ are illustrated in Fig. 6 along the linear lattice formed by the ion cores.

Because of the attractive electron–ion interaction, a calculation of the average electrostatic energy indicates that the potential energy associated with the charge distribution $n^{(+)}(x)$ is lower than that associated with n_0, whereas the potential energy associated with the charge distribution $n^{(-)}(x)$ is higher than that associated with n_0. Consequently, $n^{(+)}(x)$ lowers the energy in comparison with the average potential energy "seen" by a free-electron wave, and $n^{(-)}(x)$ raises the energy in comparison with the average potential energy "seen" by a free-electron wave. The energy gap E_g is related to the energy difference between the charge distributions $n^{(-)}(x)$ and $n^{(+)}(x)$. In the case where $V(\frac{1}{2}G)<0$, the crystal wave function just ***below*** the energy gap at points A in Fig. 5(b) is the symmetric standing-wave combination $\psi^{(+)}(k_0,x)$, and the crystal wave function just ***above*** the energy gap at points B in Fig. 5(b) is the antisymmetric standing-wave combination $\psi^{(-)}(k_0,x)$.

1.5.3 Band Occupation, Fermi Surface, and Density of States.

Having schematically indicated the effect of a periodic poten-

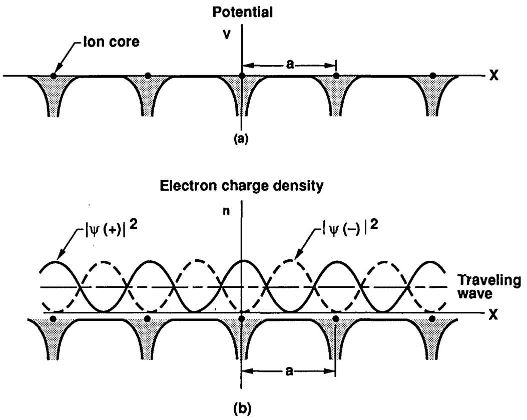

FIG. 6. (a) Variation of the potential "seen" by a conduction electron in the field of the ion cores of a linear lattice of lattice constant a. (b) Electron charge density $n(x)$ in the linear lattice for three cases: $n^{(+)}=|\psi^{(+)}|^2\sim\cos^2[(\pi/a)x]$, $n^{(-)}=|\psi^{(-)}|^2\sim\sin^2[(\pi/a)x]$ and the free-electron case $|\psi^{(0)}|^2=1$. (Reproduced with permission from Kittel, 1986, p. 162.)

tial on the free-electron solutions and energy spectrum and having given a simple physical interpretation of the concepts of energy bands and band gaps, we now turn to the important question of determining the filling of the energy bands by one-electron states. This question is central in deciding whether the solid is a metal or an insulator.

The concept of band occupation leads naturally to a discussion of Fermi level and density of states. Consider the ground state of N Bloch states. Since the energies $\epsilon_n(\mathbf{k})$ are periodic in the reciprocal space, i.e., $\epsilon_n(\mathbf{k} + \mathbf{G}) = \epsilon_n(\mathbf{k})$, we restrict the wave vector $\mathbf{k}$ to lie within the first Brillouin zone to enumerate the energy levels only once. This procedure is completely sufficient to describe the electronic band structure of the crystal.

The filling of the one-electron levels $\epsilon_n(\mathbf{k})$ determines, to a large extent, the general electronic properties of the solid. We have seen above that the number of allowed wave vectors $\mathbf{k}$ in the primitive cell of the reciprocal space is equal to the number of primitive cells in the crystal. The volume per allowed value of $\mathbf{k}$ is given by Eq. (41) as $\Delta\mathbf{k}=(2\pi)^3/V$, where V is the total volume of the crystal. As a result, the number of levels $\epsilon_n(\mathbf{k})$ in a band n is equal to the number of primitive cells N in the crystal. Since each level $\epsilon_n(\mathbf{k})$ can accommodate two electrons (one electron per spin), an energy band contains $2N$ accessible states. An energy band can either be completely or partially filled depending on the number of electrons per primitive cell. We distinguish these two important cases (Ashcroft and Mermin, 1976):

1. In certain cases, some energy bands are completely filled and all the other bands are completely empty [Fig. 7(a)]. Filled bands are sometimes referred to as *valence bands* and empty bands are called *conduction bands*. As illustrated in the previous section, the energy band gap E_g is defined as the energy difference between the lowest empty band and the highest filled band. If the band gap E_g is much larger than k_BT at room temperature ($k_B=1.3807\times10^{-23}$ J K^{-1} is Boltzmann's constant), the solid is called an *insulator*. If the band gap E_g is of the order of k_BT at room temperature, the solid is called a *semiconductor*. Since an energy band can accommodate $2N$ electrons, a band gap can arise only if the number of electrons per primitive cell is even. However, a crystal with an even number of electrons per primitive cell does not guarantee the existence of a band gap since certain energy bands may overlap in energy and, instead of one completely filled band, one has two partially filled bands and the solid is a *semimetal* [Fig. 7(b)]. The band overlap need not occur along the

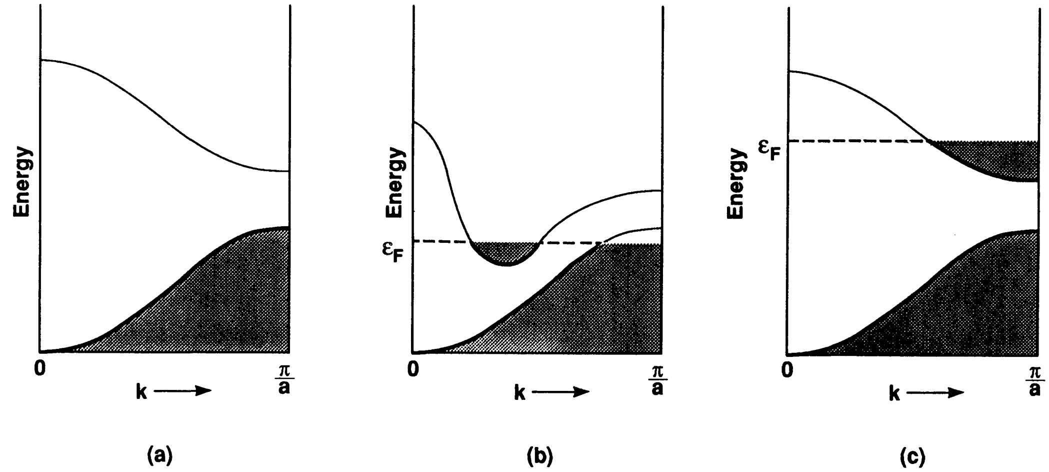

FIG. 7. The occupation of the energy bands of a solid by electrons determines its insulating or metallic behavior. (a) In an *insulator* or *semiconductor*, completely filled bands are separated from completely empty bands by an energy gap E_g. (b) In a *semimetal*, energy bands overlap in energy yielding two partially filled bands. The band overlap need not occur along the same $\mathbf{k}$ direction in the Brillouin zone. (c) In a *metal*, the topmost band is partially filled. The highest occupied state is called the *Fermi level*, ϵ_F, and lies within the energy range of one or more bands. Portions of energy bands which are filled are shaded. (Reproduced with permission from Kittel, 1986, p. 178.)

same **k** direction in the Brillouin zone.

2. In certain cases, some energy bands are partially filled. In this situation, the solid is a *metal* [Fig. 7(c)]. The highest occupied state is called the *Fermi level* ϵ_F and lies within the energy range of one or more bands. The *Fermi surface* is the surface of constant energy ϵ_F in **k** space. At absolute zero temperature, the Fermi surface separates the filled and empty levels. The branch of the Fermi surface in the band n is defined as

$$\epsilon_n(\mathbf{k}) = \epsilon_F. \tag{61}$$

A schematic representation of the above classification based on the filling of the energy levels of a solid is indicated in Fig. 7 where it is shown how the occupation of the energy bands of a solid by electrons determines its insulating, semimetallic, or metallic behavior.

The concept of *density of states* is used to enumerate the number of allowed states in a given energy range. The density of states $g_n(\epsilon)$ associated with the band n is a quantity defined such that the product $g_n(\epsilon)d\epsilon$ is equal to the number of allowed states in the band n in the energy range between ϵ and $\epsilon+d\epsilon$ per unit volume. The number of allowed states associated with the band in the energy range between ϵ and $\epsilon+d\epsilon$ is twice the volume of **k** space with $\epsilon \leqslant \epsilon_n(\mathbf{k}) \leqslant \epsilon+d\epsilon$ divided by the volume per allowed wave vector, $\Delta\mathbf{k} = (2\pi)^3/V$. Consequently, the density of states associated with the band n is (Ashcroft and Mermin, 1976)

$$\begin{aligned} g_n(\epsilon) &= \frac{2}{V}\frac{V}{(2\pi)^3}\int d^3k\, \delta^3(\epsilon-\epsilon_n(\mathbf{k})) \\ &= \frac{1}{4\pi^3}\int d^3k\, \delta^3(\epsilon-\epsilon_n(\mathbf{k})). \end{aligned} \tag{62}$$

The factor 2 arises since there are two accessible states per allowed wave vector **k**.

The integration over the volume in **k** space appearing in Eq. (62) can be transformed using the following property of δ distributions in one dimension (Bassani and Pastori Parravicini, 1975):

$$\int_a^b dx\, g(x)\delta(f(x)) = \sum_{x_0} g(x_0)\left(\left|\frac{df}{dx}\right|_{x_0}\right)^{-1}, \tag{63}$$

where x_0 are the zeroes of the function $f(x)$ on the interval (a,b). Using the property Eq. (63) in the definition Eq. (62), we obtain

$$g_n(\epsilon) = \frac{1}{4\pi^3}\int_{\epsilon=\epsilon_n(\mathbf{k})} dS_n \frac{1}{|\nabla_\mathbf{k}\epsilon_n(\mathbf{k})|}, \tag{64}$$

where dS_n is an element of surface in **k** space on the surface defined by the equation $\epsilon = \epsilon_n(\mathbf{k})$. Equation (64) gives an explicit relation between the energy eigenvalue spectrum $\epsilon_n(\mathbf{k})$ and the density of states $g_n(\epsilon)$. The total density of states is the sum of the contributions from all the bands,

$$g(\epsilon) = \sum_n g_n(\epsilon). \tag{65}$$

We now discuss some features of the density of states $g(\epsilon)$ related to the energy eigenvalue spectrum $\epsilon_n(\mathbf{k})$ (Kittel, 1963). The function $\epsilon_n(\mathbf{k})$ is periodic in reciprocal space [Eq. (46)]. Any function of more than one independent variable that is periodic in all its variables, such as $\epsilon_n(\mathbf{k})$, exhibits saddle points (van Hove, 1953). The number of saddle points depends only on the number of independent variables and is determined by topological considerations. Saddle points in the function $\epsilon_n(\mathbf{k})$ generally lead to singularities in the density of states $g(\epsilon)$. For example, in two dimensions the singularity in $g(\epsilon)$ is logarithmic, whereas in three dimensions the density of states $g(\epsilon)$ is continuous but the energy derivative $dg(\epsilon)/d\epsilon$ has an infinite discontinuity. These singularities are called *van Hove singularities* and they occur at values of ϵ on the constant-energy surface $\epsilon = \epsilon_n(\mathbf{k})$ for which the velocity vanishes, i.e., $\nabla_\mathbf{k}\epsilon_n(\mathbf{k}) = \mathbf{0}$. A characteristic profile of the density of states $g_n(\epsilon)$ associated with a band n is represented in Fig. 8. The density of states shown in Fig. 8 exhibits van Hove singularities where the density of states $g(\epsilon)$ is continuous but the energy derivative $dg(\epsilon)/d\epsilon$ has an infinite discontinuity.

1.5.4 Time-Reversal Symmetry. The Bloch solutions have symmetry properties other than those imposed by the space group of the crystal. These symmetries give rise to degeneracies in addition to those introduced by the space group of the crystal. One such important symmetry is *time-reversal symmetry*. We now briefly describe the consequences

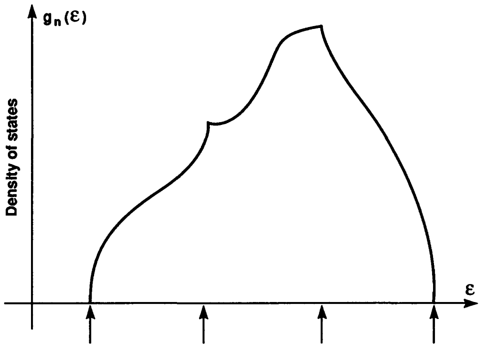

FIG. 8. Characteristic profile of the density of state $g_n(\epsilon)$ associated with the band n. van Hove singularities occur at values of the energy ϵ on the constant- energy surface $\epsilon=\epsilon_n(\mathbf{k})$ for which the velocity vanishes, i.e., $\nabla_{\mathbf{k}}\epsilon_n(\mathbf{k})=\mathbf{0}$. At these singularities, the density of states $g(\epsilon)$ is continuous but the energy derivative $dg(\epsilon)/d\epsilon$ has an infinite discontinuity. (Reproduced with permission from Ashcroft and Mermin, 1976, p. 145.)

of time-reversal symmetry in quantum mechanics (Bassani and Pastori Parravicini, 1975).

In quantum mechanics, the state of a system is specified by a wave function $\psi(\mathbf{x},t)$ whose evolution in time is determined by the time-dependent Schrödinger equation

$$H(\mathbf{x},\mathbf{p})\psi(\mathbf{x},t)=i\hbar\frac{\partial}{\partial t}\psi(\mathbf{x},t), \tag{66}$$

where $\mathbf{p}=-i\hbar\nabla$ is the momentum operator. We now perform the transformation $t\rightarrow -t$ and take the complex conjugate of Eq. (66):

$$H(\mathbf{x},-\mathbf{p})\psi^*(\mathbf{x},-t)=i\hbar\frac{\partial}{\partial t}\psi^*(\mathbf{x},-t). \tag{67}$$

In an isolated system without a magnetic field, we have $H(\mathbf{x},-\mathbf{p})=H(\mathbf{x},\mathbf{p})$, and consequently Eqs. (66) and (67) indicate that if $\psi(\mathbf{x},t)$ describes the evolution of a state, then $\psi^*(\mathbf{x},-t)$ also describes the evolution of a complex-conjugate state in reversed chronological order. If we ignore spin, the state $\psi^*(\mathbf{x})$ is called the *time-reversed conjugate* of $\psi(\mathbf{x})$. Therefore, for a spinless particle, the time-reversal operator K_0 transforms a state function $\psi(\mathbf{x})$ into its complex conjugate $\psi^*(\mathbf{x})$,

$$K_0\psi(\mathbf{x})=\psi^*(\mathbf{x}). \tag{68}$$

When spin is considered, a commonly used form for the time-reversal operator K is

$$K\equiv -i\sigma_2K_0, \tag{69}$$

where σ_2 is one of the Pauli spin operator matrices defined as

$$\sigma=\hat{\mathbf{x}}\sigma_1+\hat{\mathbf{y}}\sigma_2+\hat{\mathbf{z}}\sigma_3, \tag{70a}$$

where

$$\sigma_1=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix},\quad \sigma_2=\begin{pmatrix}0 & -i\\ i & 0\end{pmatrix},$$

$$\sigma_3=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}, \tag{70b}$$

for quantization of the electron spin along the $\hat{\mathbf{z}}$ direction.

The time-reversal operator K obeys the following commutation rules (Kittel, 1963):

$$K\mathbf{x}K^{-1}=\mathbf{x}, \tag{71a}$$

$$K\mathbf{p}K^{-1}=-\mathbf{p}, \tag{71b}$$

$$K\sigma K^{-1}=-\sigma, \tag{71c}$$

and commutes with the Hamiltonian,

$$[K,H]=0. \tag{71d}$$

Kramers's theorem states that, in the absence of an external magnetic field, if ψ is a one-electron eigenstate of H, then $K\psi$ is also an eigenstate of H belonging to the same energy eigenvalue. Moreover, the degenerate eigenstates ψ and $K\psi$ are orthogonal to each other.

The form of K given by Eq. (69), although widely used, is not unique and can be chosen on the basis that it satisfies the commutation rules given by Eqs. (71) and that the application of the time-reversal operator twice to a wave function must produce a state physically indistinguishable from the original state (Callaway, 1974). The latter condition implies that the only permissible effect of the double application of K would be the introduction of a phase factor $\exp(i\theta)$. It can further be shown that this phase factor can only be ± 1 (Wigner, 1959). Thus

$$K^2=\pm I, \tag{72}$$

where I is the identity. The particular form for K given by Eq. (69) yields the $+$ sign in Eq. (72). Other equally valid forms do not include the factor $-i$ appearing in our definition of K (Callaway, 1974).

We now consider the additional degeneracies of the crystal Bloch states introduced by time-reversal symmetry. Consider a Bloch state $\psi(\mathbf{k},\uparrow)$ of the form

$$\psi(\mathbf{k},\uparrow)=\exp(i\mathbf{k}\cdot\mathbf{x})u(\mathbf{k},\mathbf{x};\uparrow), \tag{73}$$

where ↑ indicates the direction of the spin. It can be shown that the action of the time-reversal operator K on $\psi(\mathbf{k},\uparrow)$ is

$$K\psi(\mathbf{k},\uparrow)=\psi(-\mathbf{k},\downarrow), \tag{74}$$

to within a phase factor. Therefore, the states $\psi(\mathbf{k},\uparrow)$ and $\psi(\mathbf{k},\downarrow)$ are associated with the wave vector **k**, and the time-reversed states $K\psi(\mathbf{k},\uparrow)$ and $K\psi(\mathbf{k},\downarrow)$ are associated with the wave vector $-\mathbf{k}$. From Kramers's theorem, it follows that the energies occur twice,

$$\epsilon(\mathbf{k};\uparrow)=\epsilon(-\mathbf{k};\downarrow),\quad \epsilon(\mathbf{k};\downarrow)=\epsilon(-\mathbf{k};\uparrow), \tag{75}$$

which implies that the energy bands are twofold degenerate in the sense that each energy occurs twice but not at the same wave vector **k**. However, if the crystal has inversion symmetry, the energy surface will be doubly degenerate at each **k** point.

The space inversion operator J has the following properties:

$$J\mathbf{x}J^{-1}=-\mathbf{x}, \tag{76a}$$

$$J\mathbf{p}J^{-1}=-\mathbf{p}, \tag{76b}$$

$$J\sigma J^{-1}=\sigma. \tag{76c}$$

The product operation KJ is called the *conjugation operator C* and is defined as

$$C\equiv KJ=-i\sigma_2K_0J=JK. \tag{77}$$

The conjugation operator leads to

$$C\psi(\mathbf{k},\uparrow)=KJ\psi(\mathbf{k},\uparrow)=K\psi(-\mathbf{k},\uparrow)=\psi(\mathbf{k},\downarrow), \tag{78}$$

to within a phase factor, and

$$\epsilon(\mathbf{k};\uparrow)=\epsilon(\mathbf{k};\downarrow). \tag{79}$$

As seen from Eq. (78), the conjugation operator C reverses the spin of a Bloch state, but not its wave vector. As a result, if the crystal exhibits inversion symmetry, the energy eigenvalues are doubly degenerate at each **k** point in the Brillouin zone as indicated by Eq. (79). This property will be illustrated in the next section when we consider the electronic band structure of solids with inversion symmetry.

2. CALCULATIONS OF ELECTRON STATES IN SOLIDS: BAND THEORY

2.1 Introduction: Basic Approximations

In Sec. 1, we showed that symmetry requirements imposed by the translation group required that the electronic solutions in a crystal be of the Bloch form,

$$|\psi(\mathbf{k})\rangle=\exp(i\mathbf{k}\cdot\mathbf{x})u(\mathbf{k},\mathbf{x}), \tag{80}$$

where the cell-periodic function $u(\mathbf{k},\mathbf{x})$ has the periodicity of the lattice,

$$u(\mathbf{k},\mathbf{x}+\mathbf{R}_j)=u(\mathbf{k},\mathbf{x}), \tag{81}$$

and where the wave vector **k** labels the Bloch solutions. In Eq. (80), we have used Dirac's notation, $|\psi(\mathbf{k})\rangle\equiv\psi(\mathbf{k},\mathbf{x})$, to label a Bloch state. In this section, we indicate how to calculate the Bloch function $|\psi(\mathbf{k})\rangle$ in an actual crystal. This field of solid-state physics is generally referred to as *electron band theory.*

Before we embark on a discussion of the various methods used to calculate the Bloch state $|\psi_n(\mathbf{k})\rangle$ and energy bands $\epsilon_n(\mathbf{k})$, we would like to indicate the basic set of approximations embodied within the band theory of solids (Bassani and Pastori Parravicini, 1975).

A solid contains an enormous number (of the order of 10^{23}) of interacting particles, and the theoretical treatment of the wave functions and energy levels of a solid therefore necessitates the use of simplifying approximations. A solid can be described as a collection of mutually interacting electrons and nuclei. The total Hamiltonian of the system is

$$H_t=T_e+T_N+V_{ee}+V_{eN}+V_{NN}, \tag{82}$$

where T_e and T_N represent the kinetic energy of the electrons and nuclei, respectively, and V_{ee}, V_{eN}, and V_{NN} are the electron–electron, electron–nuclei, and nuclei–nuclei interactions, respectively. Since the strongest interactions between the particles are the Coulomb interactions, we use the following expression for H_t:

$$\begin{aligned}H_t=&-\sum_i \frac{\hbar^2}{2m}\nabla_i^2-\sum_I \frac{\hbar^2}{2M_I}\nabla_I^2\\&+e^2\sum_{i<j}\frac{1}{|\mathbf{x}_i-\mathbf{x}_j|}-e^2\sum_{i,I}\frac{Z_I}{|\mathbf{x}_i-\mathbf{R}_I|}\\&+e^2\sum_{I<J}\frac{Z_IZ_J}{|\mathbf{R}_I-\mathbf{R}_J|},\end{aligned} \tag{83}$$

where the indices i, j and I, J refer to the electrons and nuclei, respectively, and Z indicates the nuclear charge. Equation (83) is only an approximation since it does not contain all possible interactions (e.g., spin-spin coupling) and assumes a nonrelativistic expression for the kinetic energy.

Band theory is concerned with the solution of the Schrödinger equation

$$H_t\Psi(\mathbf{R},\mathbf{x})=E\Psi(\mathbf{R},\mathbf{x}), \tag{84}$$

where $\mathbf{R}$ and $\mathbf{x}$ represent the space and spin coordinates of the nuclei and electrons, respectively. A basic approximation used in band theory is the *Born–Oppenheimer approximation* (Born and Oppenheimer, 1927), which consists in splitting the eigenvalue problem Eq. (84) into two separate but interdependent eigenvalue problems for electron and nuclei. This approximation is possible because the nuclei have a much greater mass than the electrons. Let us consider the electron Hamiltonian

$$H_e=T_e+V_{ee}+V_{eN}, \tag{85a}$$

or

$$H_e=-\sum_i \frac{\hbar^2}{2m}\nabla_i^2+e^2\sum_{i<j}\frac{1}{|\mathbf{x}_i-\mathbf{x}_j|} -e^2\sum_{i,I}\frac{Z_I}{|\mathbf{x}_i-\mathbf{R}_I|}. \tag{85b}$$

The Hamiltonian H_e depends parametrically on the nuclei coordinates $\mathbf{R}$. For a *given* nuclear configuration $\mathbf{R}$, we consider the eigenvalue problem

$$(T_e+V_{ee}+V_{eN})\psi_n(\mathbf{R};\mathbf{x})=E_n(\mathbf{R})\psi_n(\mathbf{R};\mathbf{x}), \tag{86}$$

where n indicates the set of quantum numbers for the electronic coordinates $\mathbf{x}$. Since the electronic mass is much smaller than the nuclear masses, we assume that the electrons will adiabatically follow the motion of the nuclei and that the eigenvalues $E_n(\mathbf{R})$ and the eigenfunctions $\psi_n(\mathbf{R};\mathbf{x})$ will be smooth and continuous functions of the nuclear coordinates $\mathbf{R}$.

The eigenfunction of the total Hamiltonian H_t is of the form

$$\Psi_{n,\mu}(\mathbf{R},\mathbf{x})=\chi_{n,\mu}(\mathbf{R})\psi_n(\mathbf{R};\mathbf{x}), \tag{87}$$

where the dependence on the electronic coordinates $\mathbf{x}$ only appears in $\psi_n(\mathbf{R};\mathbf{x})$ and μ labels the quantum numbers for the nuclear coordinates. The total eigenfunction $\Psi_{n,\mu}(\mathbf{R},\mathbf{x})$ is an eigenstate of H_t, and satisfies

$$(T_e+T_N+V_{ee}+V_{eN}+V_{NN})\chi_{n,\mu}(\mathbf{R})\psi_n(\mathbf{R};\mathbf{x}) =E_{n,\mu}\chi_{n,\mu}(\mathbf{R})\psi_n(\mathbf{R};\mathbf{x}). \tag{88}$$

Let us consider the quantity

$$\nabla_{\mathbf{R}}^2[\chi_{n,\mu}(\mathbf{R})\psi_n(\mathbf{R};\mathbf{x})]$$

required for the calculation of T_N. We have

$$\begin{aligned}\nabla_{\mathbf{R}}^2[\chi_{n,\mu}(\mathbf{R})\psi_n(\mathbf{R};\mathbf{x})] &=\psi_n(\mathbf{R};\mathbf{x})\nabla_{\mathbf{R}}^2\chi_{n,\mu}(\mathbf{R})\\ &\quad+2\nabla_{\mathbf{R}}\psi_n(\mathbf{R};\mathbf{x})\cdot\nabla_{\mathbf{R}}\chi_{n,\mu}(\mathbf{R})\\ &\quad+\chi_{n,\mu}(\mathbf{R})\nabla_{\mathbf{R}}^2\psi_n(\mathbf{R};\mathbf{x}).\end{aligned} \tag{89}$$

Because of the smooth parametric dependence of $\psi_n(\mathbf{R};\mathbf{x})$ on $\mathbf{R}$, the term $\psi_n(\mathbf{R};\mathbf{x}) \times\nabla_{\mathbf{R}}^2\chi_{n,\mu}(\mathbf{R})$ in Eq. (89) is the dominant one and we make the basic approximation of retaining only this term. Consequently, the nuclear kinetic operator T_N can be written as

$$T_N[\chi_{n,\mu}(\mathbf{R})\psi_n(\mathbf{R};\mathbf{x})]\approx\psi_n(\mathbf{R};\mathbf{x})T_N\chi_{n,\mu}(\mathbf{R}). \tag{90}$$

Substituting Eqs. (86) and (90) into Eq. (88) and dividing by $\psi_n(\mathbf{R};\mathbf{x})$, we obtain an eigenvalue equation of the nuclear eigenfunctions,

$$[T_N+V_{NN}+E_n(\mathbf{R})]\chi_{n,\mu}(\mathbf{R})=E_{n,\mu}\chi_{n,\mu}(\mathbf{R}). \tag{91}$$

Equations (86) and (91) are the fundamental equations of the Born–Oppenheimer (or adiabatic) approximation. Equation (86) gives the energy levels $E_n(\mathbf{R})$ of the electronic states when the nuclei have the fixed positions $\mathbf{R}$. The electronic energy eigenvalues $E_n(\mathbf{R})$ then appear as a *potential function* in the eigenvalue equation for the nuclear wave function $\chi_{n,\mu}(\mathbf{R})$. Equations (86) and (91) therefore permit a separate, but interdependent, evaluation of the energy eigenvalues $E_n(\mathbf{R})$ associated with the electronic states $\psi_n(\mathbf{R};\mathbf{x})$ as well as a calculation of the lattice vibrational states. Relaxation of the adiabatic approximation requires a consideration of the electron–lattice coupling and the evaluation of the second and third terms in Eq. (89).

To determine the electronic states $\psi_n(\mathbf{R};\mathbf{x})$ of a crystal, we use Eq. (86) with the nuclei at their equilibrium positions, giving rise to a one-electron potential representing the electron–nuclear Coulomb interaction. Equation (86) represents a very complicated many-body problem that cannot be exactly solved. To simplify its solution, one relies on the *one-electron band approximation.*

The central approximation of band theory is that the exact many-body problem of solving the Schrödinger equation can be meaningfully mapped into a one-electron problem for a single electron moving in an effective potential. The corresponding one-electron Schrödinger equation reduces to

$$\left(\frac{-\hbar^2}{2m}\nabla^2+V(\mathbf{x})\right)|\psi_n(\mathbf{k})\rangle=\epsilon_n(\mathbf{k})|\psi_n(\mathbf{k})\rangle, \tag{92}$$

where the crystal potential $V(\mathbf{x})$ is invariant under all the symmetry operations of the space group. We assume that the crystal potential $V(\mathbf{x})$ can be written as a linear superposition of atomic potentials $v_\mu(\mathbf{x})$ corresponding to an atomic species μ,

$$V(\mathbf{x})=\sum_{\mathbf{R}_j}\sum_{\mathbf{d}_\mu} v_\mu(\mathbf{x}-\mathbf{R}_j-\mathbf{d}_\mu), \tag{93}$$

where the vectors $\mathbf{R}_j$ label the primitive cells of the crystal and the vectors $\mathbf{d}_\mu$ indicate the positions of the atoms of type μ within a unit cell. A typical electron wave function in a crystal (Ziman, 1984) is shown in Fig. 9: Outside the ion-core region (i.e., the region near the atomic nucleus), the wave function looks very much like a plane wave, whereas it looks like an atomic orbital within the core region. Near the nucleus, the ionic potential is very attractive and the electron has a high kinetic energy (short-wavelength oscillations). The purpose of this section is to describe some of the modern theoretical schemes used to calculate the crystal Bloch states and their associated energy eigenvalue spectrum.

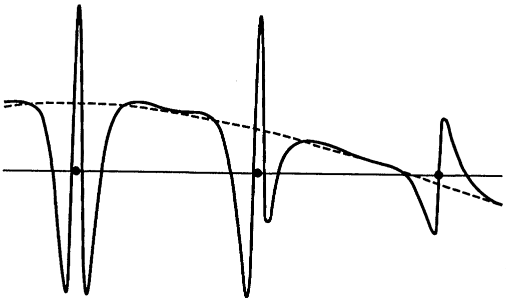

FIG. 9. Typical electron wave function in a solid. Outside the ion core region (i.e., the region near the atomic nucleus), the wave function looks very much like a plane wave, whereas it looks like an atomic orbital within the core region. Near the nucleus, the ionic potential is very attractive and the electron has a high kinetic energy (short-wavelength oscillations). The solid circles indicate the positions of the ions in the crystal and the dashed line represents a plane wave. (Reproduced with permission from Ziman, 1984, p. 92.)

The underlying idea behind the success of the band approximation is that the eigenvalues for the electronic states depend to a large extent on the symmetry properties of the lattice and are not very sensitive to small changes in the form of the crystal potential $V(\mathbf{x})$. There exist two general classes of methods for solving Eq. (92) for the energy eigenvalues $\epsilon_n(\mathbf{k})$ and wave functions $|\psi_n(\mathbf{k})\rangle$ within the band approximation. One class of methods is based on an expansion of the wave functions $|\psi_n(\mathbf{k})\rangle$ in a complete set of Bloch-type basis functions. The expansion coefficients are determined by solving the appropriate Schrödinger equation. We refer to these methods as *fixed-basis-set methods*. The other class of methods is based on an expansion of the wave functions $|\psi_n(\mathbf{k})\rangle$ in a complete set of functions, which are solutions to Schrödinger's equation within a unit cell. The resulting expansion coefficients are determined by imposing appropriate boundary conditions at the unit cell boundaries. We refer to methods based on such a construction as *scattering methods*. Band-structure calculation schemes based on fixed-basis-set methods include local-orbital methods, orthogonalized-plane-wave methods, and pseudopotential formalisms. Band-structure techniques based on scattering methods include the cellular method, the muffin-tin and augmented-plane-wave methods, and the Green's-function method. We discuss, in turn, fixed-basis-set and scattering methods. We emphasize again that the above classification is adopted for presentation purposes only and that each of the electronic band-structure techniques presented below has been highly developed over the years and constitutes a self-contained method in its own right.

The formalisms described within this section can be implemented using two different approaches that differ as to whether the solutions of Schrödinger's equation, Eq. (92), are obtained self-consistently or not. In calculations where the band equation, Eq. (92), is solved self-consistently, the crystal potential $V(\mathbf{x})$ is iteratively recomputed from the electron charge density constructed from the eigenfunctions $|\psi_n(\mathbf{k})\rangle$. This self-consistent procedure requires a simultaneous solution of Poisson's equation (relating the electrostatic potential to the charge density) and Schrödinger's equation. On the other hand, solutions that are not self-consistent do not involve a recalculation of the crystal potential.

2.2 Fixed-Basis-Set Methods

In electron-structure calculation schemes based on fixed-basis-set methods, the crystal electronic eigenfunctions $|\psi_n(\mathbf{k})\rangle$ are expanded in a complete set of fixed basis functions. The basis functions themselves are constructed to satisfy Bloch's theorem. The expansion coefficients are then determined by solving the resulting Schrödinger equation in the appropriate representation. We consider below two classes of fixed-basis-set methods depending on whether the basis functions used for the expansion of $|\psi_n(\mathbf{k})\rangle$ represent primarily *localized* or *itinerant* electrons.

2.2.1 Local-Orbital Methods. The local-orbital, or tight-binding, method was originally suggested by Bloch (1928) and uses the atomic orbitals of the atoms within a unit cell as the basis set on which the electron wave functions $|\psi_n(\mathbf{k})\rangle$ are expanded. Atomic orbitals, each centered on a particular atom, are combined to describe a Bloch state delocalized throughout the whole crystal. Let $\varphi_\mu^\alpha(\mathbf{x})$ denote the atomic orbital corresponding to the atom of type μ and characterized by the atomic quantum number α $(\alpha=n,l,m)$. These local atomic orbitals are combined into a *Bloch sum*, corresponding to the wave vector $\mathbf{k}$ and satisfying Bloch's theorem. Such a linear combination is given by

$$|\Phi_\mu^\alpha(\mathbf{k})\rangle=\frac{1}{\sqrt{N}}\sum_{\mathbf{R}_j}\exp(i\mathbf{k}\cdot\mathbf{R}_j)\varphi_\mu^\alpha(\mathbf{x}-\mathbf{R}_j-\mathbf{d}_\mu), \tag{94}$$

where N indicates the number of primitive cells in the crystal and $\varphi_\mu^\alpha(\mathbf{x}-\mathbf{R}_j-\mathbf{d}_\mu)$ is an atomic orbital centered at the site $\mathbf{R}_j+\mathbf{d}_\mu$. The electron Bloch function is then expanded as a linear combination of the Bloch sums $|\Phi_\mu^\alpha(\mathbf{k})\rangle$ as

$$|\psi_n(\mathbf{k})\rangle=\sum_{\mu,\alpha}|\Phi_\mu^\alpha(\mathbf{k})\rangle C_\mu^\alpha(n,\mathbf{k}), \tag{95}$$

where the expansion coefficients $C_\mu^\alpha(n,\mathbf{k})$ are to be determined by solving the eigensystem resulting from the substitution of Eq. (95) into Schrödinger's equation. The tight-binding expansion of the Bloch state $|\psi_n(\mathbf{k})\rangle$ in terms of the Bloch sums $|\Phi_\mu^\alpha(\mathbf{k})\rangle$ is a useful procedure in cases where a small number of basis functions $|\Phi_\mu^\alpha(\mathbf{k})\rangle$ are important in the expansion of $|\psi_n(\mathbf{k})\rangle$. This applies to low-energy atomic orbitals that are strongly localized around the nucleus and are little modified when the atoms are brought together to form the crystal (Bassani and Pastori Parravicini, 1975).

The evaluation of the expansion coefficients $C_\mu^\alpha(n,\mathbf{k})$ requires a knowledge of the Hamiltonian matrix elements between two Bloch sums,

$$H_{\mu,\mu'}^{\alpha,\alpha'}(\mathbf{k})\equiv\langle\Phi_\mu^\alpha(\mathbf{k})|H|\Phi_{\mu'}^{\alpha'}(\mathbf{k})\rangle. \tag{96}$$

The matrix elements $H_{\mu,\mu'}^{\alpha,\alpha'}(\mathbf{k})$ are of the form

$$H_{\mu,\mu'}^{\alpha,\alpha'}(\mathbf{k})=\sum_{\mathbf{R}_j}\exp(i\mathbf{k}\cdot\mathbf{R}_j)\int d^3x\,\varphi_\mu^{\alpha*}(\mathbf{x}-\mathbf{d}_\mu)$$
$$\times H\varphi_{\mu'}^{\alpha'}(\mathbf{x}-\mathbf{R}_j-\mathbf{d}_{\mu'}) \tag{97a}$$
$$\equiv\sum_{\mathbf{R}_j}\exp(i\mathbf{k}\cdot\mathbf{R}_j)H_{\mu,\mu'}^{\alpha,\alpha'}(\mathbf{R}_j). \tag{97b}$$

In practice, two approaches are followed for the evaluation of the matrix elements $H_{\mu,\mu'}^{\alpha,\alpha'}(\mathbf{R}_j)$. In one approach, one assumes an analytic form for the localized atomic orbitals $\varphi_\mu^\alpha(\mathbf{x})$ (Gaussian orbitals, Slater orbitals, etc.), and then explicitly evaluates the matrix elements $H_{\mu,\mu'}^{\alpha,\alpha'}(\mathbf{R}_j)$. This approach also requires the evaluation of the overlap matrix elements between atomic orbitals centered at different sites. This approach is sometimes referred to as the *linear combination of atomic orbitals (LCAO) method.* In another approach, the matrix elements $H_{\mu,\mu'}^{\alpha,\alpha'}(\mathbf{R}_j)$ are considered as empirical parameters to be fitted in such a way as to recover the "correct" (i.e., as extracted from experiments or calculated from first-principles methods) energy eigenvalue spectrum $\epsilon_n(\mathbf{k})$ at high-symmetry $\mathbf{k}$ points within the Brillouin zone. In this method, known as the *empirical tight-binding method,* the matrix elements $H^{\alpha,\alpha'}_{\mu,\mu'}(\mathbf{R}_j)$ are not explicitly calculated.

We now consider an example to illustrate the features of the LCAO and tight-binding methods. Consider a simple-cubic Bravais lattice of lattice constant a. Application of the LCAO method to an s band derived from a single atomic orbital of $\alpha = s$ $(l=0)$ symmetry yields an energy band of the form (Ziman, 1984)

$$\epsilon(\mathbf{k}) \approx \epsilon^s_a + 2\epsilon_{100}[\cos(k_x\, a) + \cos(k_y\, a) + \cos(k_z\, a)], \qquad (98)$$

where ϵ^s_a is the atomic energy of an s state in the isolated atom. In the dispersion relation $\epsilon(\mathbf{k})$ given by Eq. (98), the energy ϵ_{100} is a matrix element of the form $H^{\alpha,\alpha'}_{\mu,\mu'}(\mathbf{R}_j)$ (with $\mathbf{R}_j = \pm a\hat{\mathbf{x}}, \pm a\hat{\mathbf{y}}, \pm a\hat{\mathbf{z}}$), which is negative and which determines the bandwidth. The Brillouin zone is a cube and the energy surfaces within the zone are shown in Fig. 10(a) (Ziman, 1984). The corresponding energy dispersion $\epsilon(\mathbf{k})$ along a cube axis is indicated in Fig. 10(b) (Ziman, 1984) for $-\pi/a < k_x < \pi/a$, $k_y = k_z = 0$. From the tight-binding dispersion of Eq. (98), one deduces that the bandwidth is $12|\epsilon_{100}|$ and that the bottom of the band (lowest energy) is at $\epsilon^s_a + 6\epsilon_{100}$.

The band structure depends on the tight-binding integrals $H^{\alpha,\alpha'}_{\mu,\mu'}(\mathbf{R}_j)$ of the form shown in Eqs. (97): They are functions of the orbitals $\varphi^{\alpha}_{\mu}(\mathbf{x} - \mathbf{R}_j - \mathbf{d}_\mu)$ and of the crystal structure $\mathbf{R}_j$. The energy bands tend to become narrow if the atomic orbitals are strongly localized or if the lattice spacing is increased. In the limit where the atoms are infinitely far apart, the energy bands reduce to an N-fold-degenerate atomic level of the atoms brought together to form the crystal.

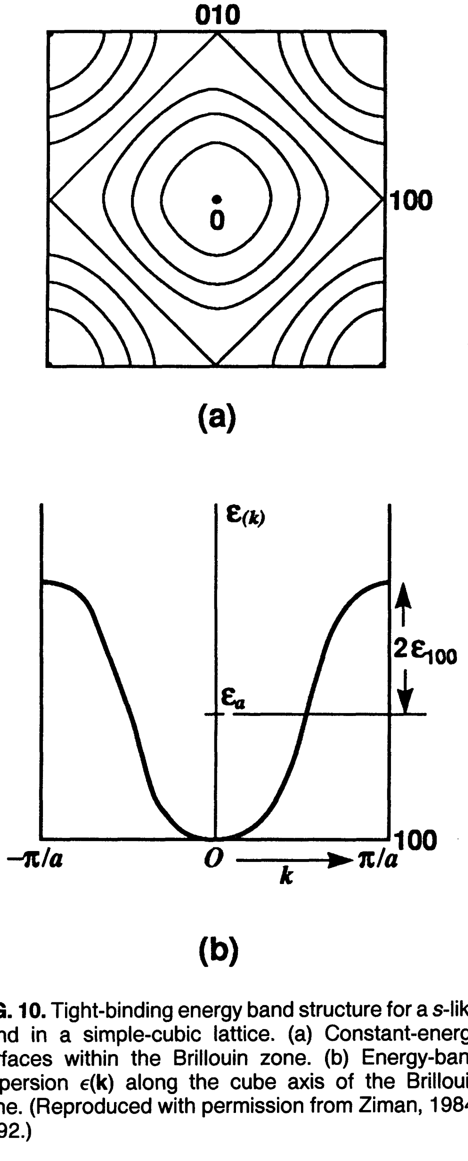

FIG. 10. Tight-binding energy band structure for a s-like band in a simple-cubic lattice. (a) Constant-energy surfaces within the Brillouin zone. (b) Energy-band dispersion $\epsilon(\mathbf{k})$ along the cube axis of the Brillouin zone. (Reproduced with permission from Ziman, 1984, p. 92.)

2.2.2 Orthogonalized-Plane-Wave Method.

We now discuss the orthogonalized-plane-wave method which, while not widely used nowadays, has enjoyed many early successes and serves as the basis to develop the very useful concept of pseudopotential. The orthogonalized-plane-wave method is based on a description of the crystal Bloch states $|\psi_n(\mathbf{k})\rangle$ in terms of plane waves that form a simple and convenient basis set. The idea of expanding the crystal electronic states in terms of plane waves was first proposed by Sommerfeld and Bethe (1933). The Bloch state is a product of a phase times a cell-periodic function,

$$|\psi_n(\mathbf{k})\rangle = \exp(i\mathbf{k}\cdot\mathbf{x})\, u_n(\mathbf{k},\mathbf{x}), \qquad (99)$$

and since the function $u_n(\mathbf{k},\mathbf{x})$ is periodic in the lattice translation vector,

$$u_n(\mathbf{k},\mathbf{x}+\mathbf{R}_j) = u_n(\mathbf{k},\mathbf{x}), \qquad (100)$$

a plane-wave expansion will only contain Fourier components associated with a reciprocal-lattice vector $\mathbf{G}$ of the form given by Eq. (9b). Therefore, the plane-wave expansion of the Bloch function $|\psi_n(\mathbf{k})\rangle$ is written as

$$
\begin{aligned}
|\psi_n(\mathbf{k})\rangle &= \exp(i\mathbf{k}\cdot\mathbf{x})u_n(\mathbf{k},\mathbf{x}) \\
&= \exp(i\mathbf{k}\cdot\mathbf{x})\left(\frac{1}{\sqrt{N\Omega_{\mathbf{x}}}}\right. \\
&\quad \times \left.\sum_{\mathbf{G}} \exp(i\mathbf{G}\cdot\mathbf{x})C_n(\mathbf{k}+\mathbf{G})\right) \\
&= \frac{1}{\sqrt{N\Omega_{\mathbf{x}}}} \sum_{\mathbf{G}} \exp[i(\mathbf{k}+\mathbf{G})\cdot\mathbf{x}] \\
&\quad \times C_n(\mathbf{k}+\mathbf{G}) \\
&\equiv \sum_{\mathbf{G}} |\mathbf{k}+\mathbf{G}\rangle C_n(\mathbf{k}+\mathbf{G}), \qquad (101)
\end{aligned}
$$

where we have defined the plane wave with momentum $\mathbf{k}+\mathbf{G}$ as

$$
|\mathbf{k}+\mathbf{G}\rangle \equiv \frac{1}{\sqrt{N\Omega_{\mathbf{x}}}} \exp[i(\mathbf{k}+\mathbf{G})\cdot\mathbf{x}], \qquad (102)
$$

and where $\Omega_{\mathbf{x}}$ is the volume of a lattice primitive cell and N is the total number of primitive cells in the crystal.

As is the case in every fixed-basis-set method, the expansion coefficients $C_n(\mathbf{k}+\mathbf{G})$ of the plane-wave expansion Eq. (101) are determined by solving the eigensystem resulting from the substitution of Eq. (101) in Schrödinger's equation. This eigensystem involves the calculation of Hamiltonian matrix elements between two plane waves,

$$
H(\mathbf{k}+\mathbf{G},\mathbf{k}+\mathbf{G}') = \langle \mathbf{k}+\mathbf{G}|H|\mathbf{k}+\mathbf{G}'\rangle. \qquad (103)
$$

It is apparent that the plane-wave expansion of Eq. (101) suffers from one important difficulty: the expansion is very slowly convergent (i.e., a very large number of plane waves is needed) for the low-energy, strongly localized core states of the crystal. These states are very well described by atomic orbitals. In order to overcome this problem, a conceptual advance was made by Herring (1940). He realized that the crystal states lying at higher energy (valence- and conduction-band states) must be orthogonal to the crystal core states since they are eigenfunctions of the same crystal Hamiltonian but have different energies. Consequently, Herring suggested using, as the basis set, *plane waves that have been orthogonalized (by construction) to the core states*. An expansion in terms of these *orthogonalized plane waves* (OPW) does indeed exhibit vastly improved convergence properties over simple plane waves.

A plane wave can be made orthogonal to a core state $|\varphi_c\rangle$ by a Gram–Schmidt procedure (Arfken, 1985). We construct an OPW associated with a momentum $\mathbf{k}+\mathbf{G}$ by projecting out the core states $|\varphi_c\rangle$ from a plane wave $|\mathbf{k}+\mathbf{G}\rangle$:

$$
|\text{OPW: } \mathbf{k}+\mathbf{G}\rangle = \left(1-\sum_c |\varphi_c\rangle\langle\varphi_c|\right)|\mathbf{k}+\mathbf{G}\rangle, \qquad (104)
$$

where $|\mathbf{k}+\mathbf{G}\rangle$ is the plane wave defined in Eq. (102) and $|\varphi_c\rangle \equiv |\varphi_c(\mathbf{k})\rangle$ represents a Bloch sum of core states of the form given in Eq. (94). The construction of an OPW is schematically indicated in Fig. 11. A valence- or conduction-band state $|\psi_v(\mathbf{k})\rangle$ (from now on, we will refer to the term valence states to designate noncore valence and conduction states and use the index v to label a valence- or conduction-band state) can be expanded in terms of OPWs in the following manner:

$$
\begin{aligned}
&|\psi_v(\mathbf{k})\rangle \\
&\quad = \sum_{\mathbf{G}} |\text{OPW:}\mathbf{k}+\mathbf{G}\rangle C_v(\mathbf{k}+\mathbf{G}) \\
&\quad = \sum_{\mathbf{G}} \left(1-\sum_c |\varphi_c\rangle\langle\varphi_c|\right)|\mathbf{k}+\mathbf{G}\rangle \\
&\quad\quad \times C_v(\mathbf{k}+\mathbf{G}). \qquad (105)
\end{aligned}
$$

The expansion coefficients $C_v(\mathbf{k}+\mathbf{G})$ are determined by solving the eigensystem in the OPW basis. This eigenvalue problem involves Hamiltonian matrix elements between two OPWs,

$$
\begin{aligned}
&H^{\text{OPW}}(\mathbf{k}+\mathbf{G},\mathbf{k}+\mathbf{G}') \\
&\quad = \langle \text{OPW: } \mathbf{k}+\mathbf{G}|H|\text{OPW: } \mathbf{k}+\mathbf{G}'\rangle. \qquad (106)
\end{aligned}
$$

The important point to appreciate is that the convergence is much better for the eigenvalue problem expressed in terms of the OPW basis $|\text{OPW: } \mathbf{k}+\mathbf{G}\rangle$ than it is in terms of the simple plane waves $|\mathbf{k}+\mathbf{G}\rangle$.

The OPW method is very advantageous and practical since the level of convergence of a particular calculation can be systematically tested by increasing the number of OPW basis functions in the expansion of $|\psi_v(\mathbf{k})\rangle$. The method is well suited to study the valence- and conduction-band structure of a large number of solids in which the core states do not

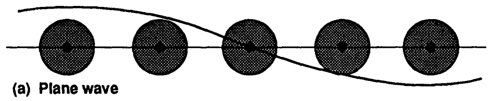

(a) Plane wave

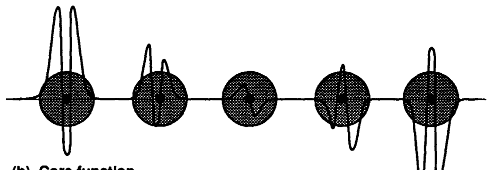

(b) Core function

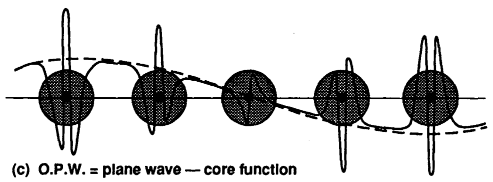

(c) O.P.W. = plane wave — core function

FIG. 11. Schematic construction of an orthogonalized plane wave (c) in which a plane wave (a) is orthogonalized to a core state (b) by construction. (Reproduced with permission from Ziman, 1984, p. 100.)

participate in the bonding and are well separated in energy from the valence states. Examples of such systems include group-IV semiconductors, III-V compound semiconductors, alkali halides, rare-gas solids, etc. One drawback of the OPW method is the assumption that the core states $|\varphi_c\rangle$ in the crystal are identical to the corresponding core states in the isolated atom. In the case in which this assumption is not realized, one must self-consistently recalculate the core states in the crystal. Moreover, the OPW method assumes a separation of the crystal eigenstates in terms of core states $|\varphi_c\rangle$ and valence states $|\psi_v(\mathbf{k})\rangle$. In some instances, such a separation is difficult to achieve in practice, as is the case for crystals containing incomplete *d* shells (Bassani and Pastori Parravicini, 1975).

We close this section by noting that, while the OPW method is not widely used nowadays, its application has been highly successful in the early development of electron band theory. In particular, pioneering utilizations of the OPW method by Herman have served as the basis for accurate band calculations of Si and Ge (Herman, 1954). Our main purpose in presenting the OPW method is its direct connection with the pseudopotential formalism.

2.2.3 Pseudopotential Theory. We now indicate how the OPW construction naturally leads to the pseudopotential concept (Heine, 1970; Harrison, 1980). It is based on the same concept that the crystal valence states $|\psi_v(\mathbf{k})\rangle$ must be orthogonal to the crystal core states $|\varphi_c\rangle$. It is possible to incorporate this idea into a form that is suitable for useful approximations.

A crystal valence state $|\psi_v(\mathbf{k})\rangle$ satisfies Schrödinger's equation

$$\left(\frac{-\hbar^2}{2m}\nabla^2+V(\mathbf{x})\right)|\psi_v(\mathbf{k})\rangle=\epsilon_v(\mathbf{k})\,|\psi_v(\mathbf{k})\rangle, \tag{107}$$

and must be orthogonal to the core states,

$$\langle\varphi_c|\psi_v(\mathbf{k})\rangle=0. \tag{108}$$

It is possible to write the valence state $|\psi_v(\mathbf{k})\rangle$ in the OPW form

$$|\psi_v(\mathbf{k})\rangle=\left(1-\sum_c|\varphi_c\rangle\langle\varphi_c|\right)|\varphi_v(\mathbf{k})\rangle, \tag{109a}$$

or, equivalently,

$$|\varphi_v(\mathbf{k})\rangle=|\psi_v(\mathbf{k})\rangle+\sum_c|\varphi_c\rangle\langle\varphi_c|\varphi_v(\mathbf{k})\rangle, \tag{109b}$$

which clearly indicates that the *pseudofunction* $|\varphi_v(\mathbf{k})\rangle$ is constructed from the valence state $|\psi_v(\mathbf{k})\rangle$, which has been smoothed in the core region by subtracting out the core orthogonality wiggles. The Schrödinger equation satisfied by the pseudofunction $|\varphi_v(\mathbf{k})\rangle$ is obtained simply by substituting Eq. (109b) in Eq. (107) and noting that the core state $|\varphi_c\rangle$ is an eigenstate of the crystal Hamiltonian with energy ϵ_c,

$$\left(\frac{-\hbar^2}{2m}\nabla^2+V(\mathbf{x})\right)|\varphi_v(\mathbf{k})\rangle+\sum_c[\epsilon_v(\mathbf{k})-\epsilon_c]\times|\varphi_c\rangle\langle\varphi_c|\varphi_v(\mathbf{k})\rangle=\epsilon_v(\mathbf{k})\,|\varphi_v(\mathbf{k})\rangle. \tag{110}$$

By comparing the OPW expansion Eq. (105) with the form of Eq. (109), we can identify the pseudofunction $|\varphi_v(\mathbf{k})\rangle$ with the plane-wave expansion (Heine, 1970)

$$|\varphi_v(\mathbf{k})\rangle = \sum_{\mathbf{G}} |\mathbf{k}+\mathbf{G}\rangle C_v(\mathbf{k}+\mathbf{G}). \tag{111}$$

Since the OPW determinant equation converges rapidly, the coefficients $C_v(\mathbf{k}+\mathbf{G})$ rapidly go to zero for large $\mathbf{k}+\mathbf{G}$, and *the pseudofunction is the smooth part of the all-electron wave function that remains after the oscillations in the core region* ($r < r_c$) *have been removed.* Consequently, the pseudofunction $|\varphi_v(\mathbf{k})\rangle$ is identically equal to the all-electron wave function $|\psi_v(\mathbf{k})\rangle$ outside the atomic core radius r_c and extends smoothly inside the core without the oscillations that $|\psi_v(\mathbf{k})\rangle$ exhibits in the core region,

$$|\varphi_v(\mathbf{k})\rangle \equiv \begin{cases} |\psi(\mathbf{k})\rangle & \text{for } r > r_c \\ \text{smooth nodeless function} & \\ \quad \text{for } r < r_c\,. & \end{cases} \tag{112}$$

Operationally, a pseudopotential can be defined as an effective potential that generates a pseudofunction $|\varphi_v(\mathbf{k})\rangle$ which is *smooth and nodeless throughout space* and whose energy eigenvalue spectrum $\epsilon_v(\mathbf{k})$ is identical to that of the corresponding all-electron wave function $|\psi_v(\mathbf{k})\rangle$. A comparison between a pseudofunction and an all-electron wave function for silicon is indicated in Fig. 12 (Yin and Cohen, 1982).

The pseudofunction $|\varphi_v(\mathbf{k})\rangle$ is a solution of the Schrödinger equation Eq. (110) associated with an effective energy-dependent *pseudopotential*

$$\begin{aligned} \hat{V}^{\text{ps}} &= V(\mathbf{x}) + \sum_c [\epsilon_v(\mathbf{k}) - \epsilon_c]\,|\varphi_c\rangle\langle\varphi_c| \\ &\equiv V(\mathbf{x}) + \hat{V}_R, \end{aligned} \tag{113}$$

where $V(\mathbf{x})$ is the crystal potential and $\hat{V}_R = \Sigma_c[\epsilon_v(\mathbf{k}) - \epsilon_c]\,|\varphi_c\rangle\langle\varphi_c|$ is a nonlocal potential operator that arises from the imposition of the orthogonalization condition, Eq. (108). The form of $\hat{V}^{\text{ps}}$ given by Eq. (113) is referred to as the Phillips–Kleinman (Phillips and Kleinman, 1959) pseudopotential.

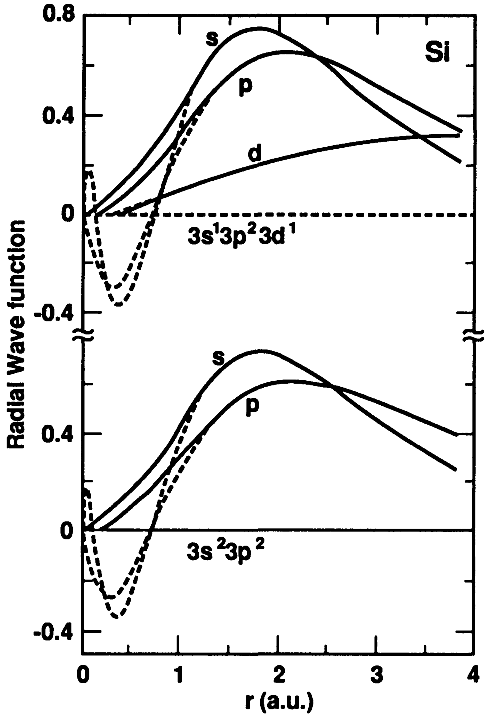

FIG. 12. Comparison between the nodeless pseudofunction (solid lines) and corresponding all-electron (dashed lines) radial wave functions of silicon for the atomic configurations $3s^1 3p^2 3d^1$ and $3s^2 3p^2$ (atomic units: 1 a.u. = 0.529 Å). (Reproduced with permission from Yin and Cohen, 1982b, p. 5671.)

Any amount of core states can be added to the pseudofunction $|\varphi_v(\mathbf{k})\rangle$ since these core states are automatically projected out by the construction in Eq. (109). This arbitrariness in the construction of the pseudofunction can be transferred to the pseudopotential itself. Consequently, there is a high degree of arbitrariness in the construction of the pseudopotential. For example, a general nonlocal pseudopotential can be defined by adding an arbitrary operator $\hat{A}$,

$$\hat{V}^{\text{ps}} = V(\mathbf{x}) + \sum_c \hat{A}\,|\varphi_c\rangle\langle\varphi_c|. \tag{114}$$

The operator $\hat{A}$ can be chosen such as to simplify the calculation of the matrix elements. The pseudopotential $\hat{V}^{\text{ps}}$ of Eq. (113) defines a pseudopotential Hamiltonian

$$\hat{H}^{\text{ps}} = \frac{-\hbar^2}{2m}\nabla^2 + \hat{V}^{\text{ps}}. \tag{115}$$

The pseudopotential concept can be formulated in the following manner: The pseudofunction equation, Eq. (110), is a transformation of the Schrödinger equation that retains the same energy eigenvalue $\epsilon_v(\mathbf{k})$ but transforms the wave function from $|\psi_v(\mathbf{k})\rangle$ to

$|\varphi_v(\mathbf{k})\rangle$. Since the core states lie at lower energy than the valence states, $\epsilon_c < \epsilon_v(\mathbf{k})$, the potential operator $\hat{V}_R$ is *repulsive* and leads to a nodeless pseudofunction $|\varphi_v(\mathbf{k})\rangle$. The popularity and usefulness of the pseudopotential concept emerge from the fact that because $\hat{V}^{\mathrm{ps}}$ is much weaker than the full potential, it leads to good convergence properties at the cost of introducing the additional complexity indicated by the form Eq. (113). Figure 13 indicates a typical pseudopotential for silicon as currently used in modern first-principles pseudopotential calculations (Yin and Cohen, 1982).

The central idea of the pseudopotential concept is the separation of the crystal eigenfunctions into core states and valence states and the solution of the resulting Hamiltonian for the energies and wave functions within the valence subspace. The resulting orthogonalization of the valence states from the core states leads to a weak pseudopotential that produces a smooth, nodeless pseudofunction $|\varphi_v(\mathbf{k})\rangle$ for which a fixed-basis-set expansion will converge rapidly. In actual calculations, the pseudofunction $|\varphi_v(\mathbf{k})\rangle$ is expanded in a convenient basis set. An extremely popular expansion set is provided by plane waves,

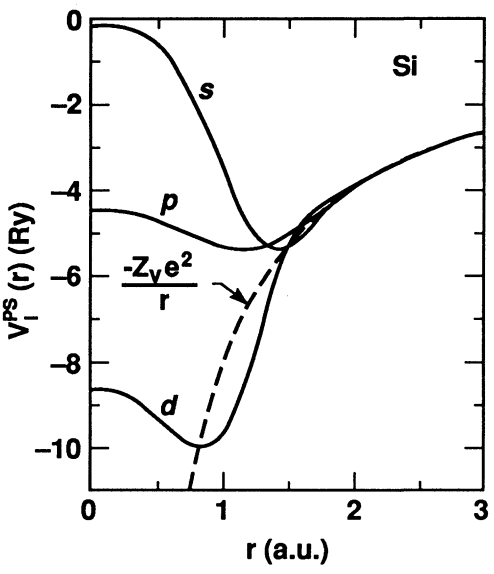

FIG. 13. *Ab initio* ionic pseudopotential of silicon. The letters *s*, *p*, and *d* denote the (nonlocal) pseudopotential for angular momenta $l=0$, 1, and 2, respectively. The dashed line indicates the bare Coulomb potential of a pointlike atomic core (atomic units: 1 Ry=13.603 eV, 1 a.u.=0.529 Å). (Reproduced with permission from Yin and Cohen, 1982b, p. 5670.)

$$|\varphi_v(\mathbf{k})\rangle = \sum_{\mathbf{G}} |\mathbf{k}+\mathbf{G}\rangle C_v(\mathbf{k}+\mathbf{G}), \tag{116}$$

where the expansion coefficients $C_v(\mathbf{k}+\mathbf{G})$ are determined by solving the resulting eigenvalue problem involving the pseudopotential Hamiltonian between two plane-wave states,

$$H^{\mathrm{ps}}(\mathbf{k}+\mathbf{G},\mathbf{k}+\mathbf{G}') = \langle \mathbf{k}+\mathbf{G}|\hat{H}^{\mathrm{ps}}|\mathbf{k}+\mathbf{G}'\rangle. \tag{117}$$

The pseudopotential method is best applicable in systems where the core states do not appreciably overlap with the valence states and are well separated in energy. Its usefulness lies in the fact that a plane-wave expansion leads to a rapid convergence. The calculation of the pseudopotential matrix elements is particularly simple in a plane-wave basis, even in the case where nonlocal pseudopotentials are used. Moreover, the solution of Poisson's equation is trivial when a plane-wave basis is used. Consequently, plane-wave basis functions are well suited for self-consistent electronic-structure calculations where Poisson's equation and Schrödinger's equation must be solved simultaneously. The pseudopotential method does not rely on any assumptions regarding the shape of the potential and affords a treatment of low-symmetry and open-structure systems. Let us mention also that the pseudopotential scheme has also been successfully implemented within a local-orbital scheme (Chelikowsky and Louie, 1984). Moreover, generalized pseudopotential schemes have been developed and applied to the treatment of solids with incomplete *d* shells (Harrison, 1969).

The pseudopotential method has been implemented within both empirical and first-principles contexts. Within the empirical pseudopotential method, the matrix elements $H^{\mathrm{ps}}(\mathbf{k}+\mathbf{G},\mathbf{k}+\mathbf{G}')$ are used as empirical parameters determined by fitting the calculated band structure $\epsilon_n(\mathbf{k})$ to experimental spectroscopic data. Within the first-principles pseudopotential method, a (generally nonlocal) pseudopotential is constructed from the all-electron potential in such a way as to reproduce the all-electron energy eigenvalue spectrum and to generate a nodeless pseudofunction that is identically equal to the all-electron wave function for $r > r_c$. We close this section by suggesting that the interested reader consult a recent comprehensive review of the utilization of pseudopotential methods in modern electronic-structure calculations (Pickett, 1989).

2.3 Scattering Methods

In electron-structure calculation schemes based on scattering methods, the crystal Bloch states $|\psi_n(\mathbf{k})\rangle$ are obtained by solving the Schrödinger equation within a Wigner–Seitz primitive cell. The solution for the entire crystal is then obtained by imposing appropriate boundary conditions at the cell boundaries. We refer to these methods as scattering methods since they often rely on basic results derived from scattering theory in the application of the boundary conditions. Scattering methods usually treat both the valence and core states and are therefore all-electron methods. The method was originally proposed by Wigner and Seitz (1933,1934) to study the cohesive energy of sodium. We indicate below extensions and refinements of this method (Bassani and Pastori Parravicini, 1975).

2.3.1 Cellular Method. Most scattering methods are based, in essence or at least in spirit, on the *cellular method.* We first present the original derivation of the cellular method and indicate modern refinements later. The fundamental principle underlying the cellular method is the expansion of the Bloch state $\psi(\mathbf{k},\mathbf{x})$ in terms of the solution of Schrödinger's equation for the atoms within one Wigner–Seitz primitive cell,

$$\psi(\mathbf{k},\mathbf{x}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} C_{lm}(\mathbf{k}) Y_l^m(\hat{\mathbf{x}}) R_l(E;r), \tag{118}$$

where $Y_l^m(\hat{\mathbf{x}})$ are spherical harmonics and the functions $R_l(E;r)$ are solutions of the atomic radial Schrödinger equation

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dR_l(E;r)}{dr}\right) + \left(\frac{2m}{\hbar^2}[E-V(r)] - \frac{l(l+1)}{r^2}\right)R_l(E;r) = 0, \tag{119}$$

in which $V(r)$ is the spherically symmetric cellular potential. The true crystal potential $V(\mathbf{x})$ is not spherically symmetric and the utilization of an atomic spherical potential $V(r)$ to generate the basis functions $Y_l^m(\hat{\mathbf{x}})R_l(E;r)$ is an approximation. A comparison between the actual crystal potential $V(\mathbf{x})$ and the spherically symmetric potential $V(r)$ within a primitive cell is indicated in Fig. 14 (Ashcroft and Mermin, 1976).

The coefficients $C_{lm}(\mathbf{k})$ appearing in the expansion Eq. (118) are determined by suitable boundary conditions imposed at the boundaries between primitive cells. For every pair of points $\mathbf{R}_j$ and $\mathbf{R}_i$ on opposite faces of the Wigner–Seitz primitive cell, the eigenfunction $\psi(\mathbf{k},\mathbf{x})$ must satisfy the boundary conditions required by Bloch's theorem,

$$\psi(\mathbf{k},\mathbf{R}_i) = \exp(i\mathbf{k}\cdot\mathbf{R})\psi(\mathbf{k},\mathbf{R}_j), \tag{120a}$$

$$\hat{\mathbf{n}}\cdot\nabla\psi(\mathbf{k},\mathbf{R}_i) = \exp(i\mathbf{k}\cdot\mathbf{R})\hat{\mathbf{n}}\cdot\nabla\psi(\mathbf{k},\mathbf{R}_j), \tag{120b}$$

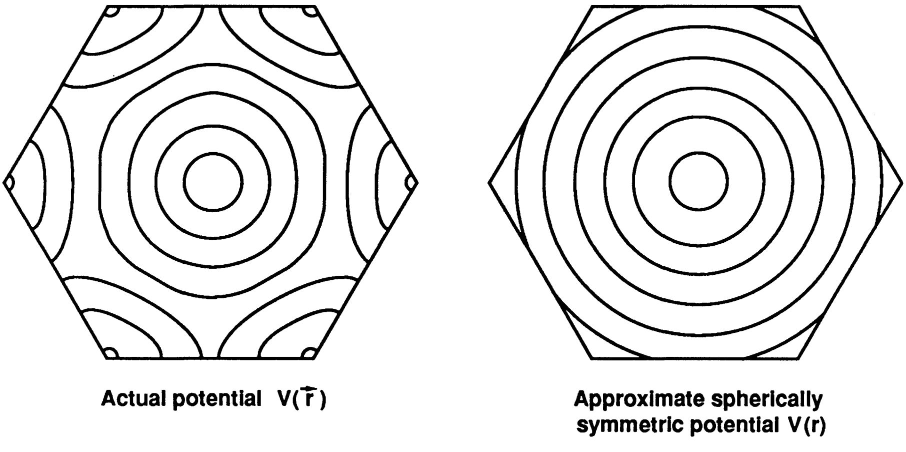
Actual potential V($\vec{r}$)
Approximate spherically symmetric potential V(r)

FIG. 14. Equipotential contours showing the true crystal potential $V(\mathbf{x})$ and the spherically symmetric potential $V(r)$ within a primitive cell. (Reproduced with permission from Ashcroft and Mermin, 1976, p. 196.)

where $\mathbf{R}=\mathbf{R}_j-\mathbf{R}_i$ is a lattice-translation vector and $\hat{\mathbf{n}}$ is a unit vector normal to the pair of surfaces to which $\mathbf{R}_i$ and $\mathbf{R}_j$ belong. The boundary conditions Eqs. (120) relate the value of $\psi(\mathbf{k},\mathbf{x})$ and of its normal derivative on opposite faces of the Wigner–Seitz primitive cell.

The prescription for calculating the crystal energy levels $\epsilon(\mathbf{k})$ and eigenfunctions $\psi(\mathbf{k},\mathbf{x})$ in the cellular method is the following: One solves the radial Schrödinger equation Eq. (119) to obtain the radial solutions $R_l(E;r)$ within the Wigner–Seitz cell. The expansion of the crystal eigenfunction $\psi(\mathbf{k},\mathbf{x})$ for a given $\mathbf{k}$ is then performed in terms of the solutions $Y_l^m(\hat{\mathbf{x}})R_l(E;r)$ as indicated by Eq. (118). The boundary conditions Eqs. (120) are then imposed for every pair of points $\mathbf{R}_i$ and $\mathbf{R}_j$ on the cell boundary. In practical implementations, however, the matching is done only at a finite number of points on the Wigner–Seitz cell. The matching points are chosen on the basis of some method. During this process, the energy E is considered as a parameter. For a given wave vector $\mathbf{k}$, the expansion Eq. (118) will satisfy the boundary conditions Eqs. (120) for every pair of points on opposite faces of the Wigner–Seitz cell only for certain values of E and $C_{lm}(\mathbf{k})$. These allowed values of E and $C_{lm}(\mathbf{k})$ for which the boundary conditions are satisfied determine the energy eigenvalue spectrum $\epsilon(\mathbf{k})=E$ and eigenfunctions $\psi(\mathbf{k},\mathbf{x})$ of the crystal (Bassani and Pastori Parravicini, 1975).

The application of the cellular method was originally performed with a spherically symmetric potential which is equal to the atomic potential $V_{\text{atom}}(r)$ inside a sphere of radius r_s, and a constant V_0 outside this sphere. Typically, the radius r_s is of the order of half the first-nearest-neighbor distance. Such a potential is said to be of the *muffin-tin* form and is written as

$$V_{\text{MT}}(r)\equiv\begin{cases}V_{\text{atom}}(r) & \text{for } r<r_s\\ V_0 & \text{for } r>r_s\,.\end{cases} \tag{121}$$

The form of the muffin-tin potential $V_{\text{MT}}(r)$ is shown in Fig. 15 (Ashcroft and Mermin, 1976).

The cellular method is an all-electron method, in contrast to the pseudopotential method in which only the valence states are calculated, and the core states are frozen at their atomic values. There are two major difficulties associated with the cellular method as originally implemented. One unsatisfactory feature concerns the application of the boundary conditions and is related to the arbitrariness in choosing the finite set of matching points on the surface of the Wigner–Seitz cell. Another difficulty concerns the *shape approximation* arising when a spherically symmetric potential is used inside the Wigner–Seitz primitive cell. In certain types of materials, such as alkali metals, the crystal potential is indeed nearly spherically symmetric and the utilization of a muffin-tin potential $V_{\text{MT}}(r)$ of the form of Eq. (121) constitutes a good approximation. In other materials, such as covalently bonded open structures, the crystal potential is poorly approximated by the spherical form of $V_{\text{MT}}(r)$. Recall that the LCAO and pseudopotential methods do not rely on any shape approximation and are more suited to treat the electronic structure of low-symmetry open structures than the cellular method is.

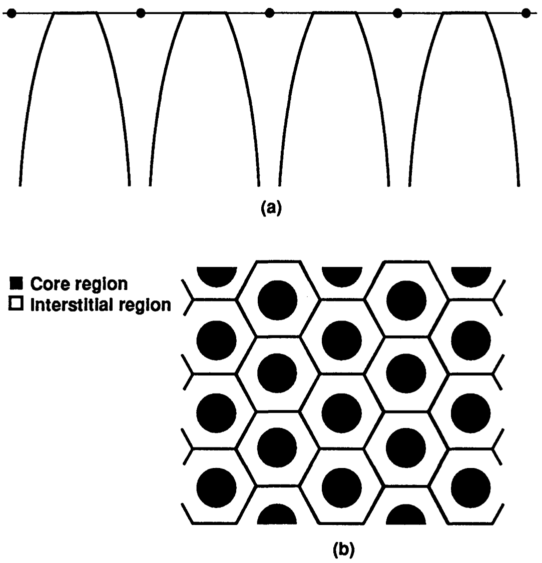

FIG. 15. Muffin-tin potential $V_{\text{MT}}(r)$. (a) Muffin-tin potential plotted along a line of ions. (b) The muffin-tin potential is constant in the interstitial regions and describes an isolated ion in the core regions. (Reproduced with permission from Ashcroft and Mermin, 1976, p. 200.)

We close this section by indicating that modern refinements of the cellular method that do not require the utilization of a muffin-tin potential $V_{\text{MT}}(r)$ have been developed (Nesbet, 1984, 1986). In these modern formulations of the cellular method, the solutions are matched over the entire surface of the Wigner–Seitz cell and can accommodate general shape-corrected as well as spherical cellular potentials.

2.3.2 Augmented Plane-Wave and Linear Muffin-Tin Methods. A solution to overcome the boundary condition problem inherent to the cellular method was originally proposed by Slater (1937). The basic principle is to augment the basis set on which $\psi(\mathbf{k},\mathbf{x})$ is expanded to include

1. the radial solutions $R_l(E;r)$ as in Eq. (118) and
2. plane waves in the interstitial regions where the potential is slowly varying.

This new basis function is called an *augmented plane wave* (APW) and is defined as

$$|\mathrm{APW}{:}\ \mathbf{k}\rangle \equiv \begin{cases} \displaystyle\sum_{l=0}^{\infty}\sum_{m=-1}^{l} a_{lm}(\mathbf{k})Y_l^m(\hat{\mathbf{x}})\ R_l(E;r) & \text{if } r<r_s \\ \exp(i\mathbf{k}\cdot\mathbf{x}) & \text{if } r>r_s\,. \end{cases} \tag{122}$$

The APW basis function $|\mathrm{APW}{:}\mathbf{k}\rangle$ contains two pieces:

1. an expansion in terms of the spherical-wave basis functions $Y_l^m(\hat{\mathbf{x}})R_l(E;r)$ within the muffin-tin sphere $r<r_s$ where the potential is atomiclike, and
2. a plane wave for $r>r_s$ where the potential is slowly varying.

The plane wave is joined to the spherical waves on a sphere whose radius r_s is usually determined by requiring that it be the largest sphere which can be inscribed within the Wigner–Seitz cell. They key point in the APW method is that the boundary conditions Eqs. (120) are already satisfied by the APW basis function $|\mathrm{APW}{:}\mathbf{k}\rangle$ at the surface of the Wigner–Seitz cell, and a much easier matching problem at the sphere surface (i.e., at $r=r_s$) provides the eigenvalues more efficiently than in the cellular method. This matching at the sphere boundary can be done by adjusting the expansion coefficients $a_{lm}(\mathbf{k})$ to assure the continuity of $|\mathrm{APW}{:}\mathbf{k}\rangle$ at $r=r_s$. By expanding the plane wave $\exp(i\mathbf{k}\cdot\mathbf{x})$ in spherical harmonics, it follows that the continuity of the function $|\mathrm{APW}{:}\mathbf{k}\rangle$ at $r=r_s$ is assured if we choose

$$a_{lm}(\mathbf{k})=4\pi Y_l^{m*}(\hat{\mathbf{k}})j_l(kr_s)\ /R_l(E;r_s)\ , \tag{123}$$

where $j_l(kr_s)$ is a spherical Bessel function (Bassani and Pastori Parravicini, 1975). In general, the slope of $|\mathrm{APW}{:}\mathbf{k}\rangle$ will be discontinuous at $r=r_s$.

The APW scheme is based on the expansion of the Bloch state $|\psi_n(\mathbf{k})\rangle$ in terms of the APW basis functions,

$$|\psi_n(\mathbf{k})\rangle=\sum_{\mathbf{G}}\ |\mathrm{APW}{:}\mathbf{k}+\mathbf{G}\rangle C_n(\mathbf{k}+\mathbf{G}), \tag{124}$$

where the expansion coefficients $C_n(\mathbf{k}+\mathbf{G})$ are obtained by solving the eigenvalue system resulting from substituting the APW expansion Eq. (122) in Schrödinger's equation. Hamiltonian matrix elements between two APWs must be computed,

$$H^{\mathrm{APW}}(\mathbf{k}+\mathbf{G},\mathbf{k}+\mathbf{G}';E) =\langle \mathrm{APW}{:}\mathbf{k}+\mathbf{G}|H|\mathrm{APW}{:}\ \mathbf{k}+\mathbf{G}'\rangle. \tag{125}$$

The evaluation of the APW matrix elements $H^{\mathrm{APW}}(\mathbf{k}+\mathbf{G},\mathbf{k}+\mathbf{G}';E)$ requires some care because of the discontinuity of slope at $r=r_s$ exhibited by the basis functions $|\mathrm{APW}{:}\mathbf{k}+\mathbf{G}\rangle$ (Bassani and Pastori Parravicini, 1975).

Contrary to the OPW method, the APW Hamiltonian matrix elements

$$H^{\mathrm{APW}}(\mathbf{k}+\mathbf{G},\mathbf{k}+\mathbf{G}';E)$$

depend on the energy and it is more efficient to perform a *linearization* about a fixed energy to cast the problem in an eigensystem form (Andersen, 1975). The resulting method, called the *linearized augmented-plane-wave (LAPW) method,* is extremely efficient and has allowed faster APW calculations by a factor of approximately 100 (Andersen *et al.*, 1987).

In the above discussion of the APW method, we have implicitly assumed that the potential was of the muffin-tin spherical form given by $V_{\mathrm{MT}}(r)$. However, it is possible to formulate the APW method without making any spherical shape approximation for the crystal potential. In this formulation, the matching sphere is immaterial and is used only in the definition of the APW basis functions $|\mathrm{APW}{:}\mathbf{k}+\mathbf{G}\rangle$. This shape-corrected formulation is known as the *full-potential linearized augmented-plane-wave (FLAPW) method.* Consequently, the FLAPW method should be regarded as being a fully shape-corrected method just as are the LCAO, OPW, and pseudopotential methods discussed above.

Another attractive feature of the APW method is that no assumptions are made about the core states and all the states (core and valence) can be treated on an equal footing. This is in contrast to the OPW and pseudopotential methods, which project out the core states from the all-electron wave function. A *panel method* is often used within the APW method to treat semicore states that are close in energy to the valence states (Skriver, 1984) and participate in the bonding in the solid. In this method, core and valence states are described by linearizing about two different energies: one near the middle of the valence band for the valence states and another one in the energy region of the core states. Consequently, the APW method can successfully be applied to the study of transition-metal elements where the pseudopotential prescription based on the separation between the core and valence states is questionable.

The form of the function in Eq. (122) can be made more general depending on the form of the function outside the sphere, $r > r_s$: If the function outside the sphere defined by the radius r_s is a plane wave, as in Eq. (122), the function is an augmented plane wave (APW); if the function outside the sphere is a Hankel function of zero kinetic energy, the function is a muffin-tin orbital (MTO); and if the function outside the sphere is a Slater orbital, the function is called a Slater-type orbital (STO) (Andersen *et al.*, 1987).

Given the importance of the MTO method in band theory and because of the impact it has had in the field of electronic structure calculations of transitions metals, we briefly discuss its major aspects. Just as in the cellular and APW methods, the crystal is divided into muffin-tin spheres of radius r_s where the atomic potential is taken to be spherically symmetric, separated by interstitial regions of constant potential [cf. the definition of $V_{\mathrm{MT}}(r)$ in Eq. (121)]. In the *atomic-sphere approximation (ASA)*, the piece of the MTO extending in the interstitial region is taken to have zero kinetic energy.

A MTO is an energy-dependent orbital constructed by a procedure similar to the APW of Eq. (122) (Skriver, 1984),

$$\chi_{lm}(E,\mathbf{x}) \equiv i^l Y_l^m(\hat{\mathbf{x}}) \times \begin{cases} R_l(E;r)+p_l(E)(r/r_s)^l & \text{if } r<r_s \\ (r_s/r)^{l+1} & \text{if } r>r_s\,. \end{cases} \tag{126}$$

The MTO defined by Eq. (126) is regular, continuous, and differentiable in all space. The *potential function* $p_l(E)$ appearing in the definition of $\chi_{lm}(E,\mathbf{x})$ is determined by the requirement of continuity and differentiability at the sphere radius r_s (Skriver, 1984),

$$p_l(E)=\frac{D_l(E)+l+1}{D_l(E)-l}, \tag{127}$$

where

$$D_l(E)\equiv\frac{r_s}{R_l(E;r_s)}\left(\frac{\partial R_l(E;r)}{\partial r}\right)_{r=r_s}. \tag{128}$$

The tail of the MTO $\chi_{lm}(E,\mathbf{x})$ is taken to be of the form $(r_s/r)^{l+1}$ and has zero kinetic energy, i.e., it is a solution of Laplace's equation, $\nabla^2\chi_{lm}(E,\mathbf{x})=0$.

The MTOs defined by Eq. (126) are combined into Bloch sums,

$$|\chi_{lm}(E,\mathbf{k})\rangle=\sum_{\mathbf{R}_j}\exp(i\mathbf{k}\cdot\mathbf{R}_j)\chi_{lm}(E,\mathbf{x}-\mathbf{R}_j). \tag{129}$$

The MTO band structure scheme involves the expansion of the Bloch state $|\psi_n(\mathbf{k})\rangle$ in terms of the MTO Bloch sums,

$$|\psi_n(\mathbf{k})\rangle=\sum_{l,m}|\chi_{lm}(E,\mathbf{k})\rangle C_{lm}(n,\mathbf{k}). \tag{130}$$

The Bloch function $|\psi_n(\mathbf{k})\rangle$ expressed by Eq. (130) will be an eigenfunction of the crystal Hamiltonian provided that the tails $(r_s/|\mathbf{x}-\mathbf{R}_j|)^{l+1}$ for all $\mathbf{R}_j\neq\mathbf{0}$ cancel the term $p_l(E)(r/r_s)^l$ appearing in the definition of $\chi_{lm}(E,\,\mathbf{x})$ inside the sphere $r<r_s$ centered at $\mathbf{R}=\mathbf{0}$. This principle of *tail cancellation* can be expressed mathematically by requiring that

$$0=\sum_{l,m}C_{lm}(n,\mathbf{k})\left[i^lY_l^m(\hat{\mathbf{x}})\frac{p_l(E)}{2(2l+1)}\left(\frac{r}{r_s}\right)^l + \sum_{\mathbf{R}_j\neq\mathbf{0}}\exp(i\mathbf{k}\cdot\mathbf{R}_j)i^lY_l^m(\hat{\mathbf{u}}_j) \times\left(\frac{r_s}{|\mathbf{x}-\mathbf{R}_j|}\right)^{l+1}\right], \tag{131}$$

where $\mathbf{u}_j=(\mathbf{x}-\mathbf{R}_j)$. A closed form for this condition can be obtained by expanding the tail centered at $\mathbf{R}_j$ in terms of the functions $i^l Y_l^m(\hat{\mathbf{x}})(r/r_s)^l$ centered at $\mathbf{R}=\mathbf{0}$:

$$\sum_{\mathbf{R}_j\neq\mathbf{0}} \exp(i\mathbf{k}\cdot\mathbf{R}_j)\, i^l Y_l^m(\hat{\mathbf{u}}_j)\left(\frac{r_s}{|\mathbf{x}-\mathbf{R}_j|}\right)^{l+1} = \sum_{l',m'} -\frac{1}{2(2l'+1)}\, S_{l'm'}^{lm}(\mathbf{k})\, i^{l'} Y_{l'}^{m'}(\hat{\mathbf{x}})\left(\frac{r}{r_s}\right)^{l'}, \tag{132}$$

where the expansion coefficients $S_{l'm'}^{lm}(\mathbf{k})$ are called *structure constants.* The tail cancellation condition, Eq. (132), can be expressed as a secular equation by collecting the coefficients of $Y_l^m(\hat{\mathbf{x}})$ $(r/r_s)^l$,

$$\sum_{l,m} [p_l(E)\delta_{l,l'}\delta_{m,m'} - S_{l'm'}^{lm}(\mathbf{k})]\, C_{lm}(n,\mathbf{k}) = 0. \tag{133}$$

Solution of Eq. (133) completely specifies the Bloch state by simultaneously satisfying the boundary conditions and Bloch's theorem.

The muffin-tin orbital method can also be linearized and, in this form, is called the linear muffin-tin orbital method (LMTO). As in the case of the APW method, the utilization of a spherically symmetric potential is not necessary within the LMTO method. The implementation of the LMTO method with a spherically symmetric potential is called the LMTO–atomic-sphere approximation (LMTO-ASA), whereas it is called the full-potential LMTO when the spherical shape approximation is relaxed (Methfessel *et al.*, 1989).

2.3.3 Green's-Function Method. Other methods based on scattering techniques have been proposed and are currently used in electron band-structure calculations. One such method is the *Green's-function* or *KKR method,* originally proposed by Korringa (1947) and Kohn and Rostoker (1954). This method uses standard Green's-function techniques to transform Schrödinger's equation into an equivalent integral equation. This method avoids the problems associated with the boundary conditions inherent to the cellular method. However, the Green's-function method usually requires the utilization of a spherical muffin-tin potential $V_{\mathrm{MT}}(r)$ as given by Eq. (121).

In the Green's-function method, the calculation of the crystal Bloch states $\psi(\mathbf{k},\mathbf{x})$ requires a knowledge of the Green's function $G(\mathbf{k},\mathbf{x}-\mathbf{x}';E)$ satisfying the free-particle Schrödinger equation

$$[(\hbar^2/2m)\nabla^2 + E]\, G(\mathbf{k},\mathbf{x}-\mathbf{x}';E) = \delta^3(\mathbf{x}-\mathbf{x}'), \tag{134}$$

describing the response to a delta-function excitation $\delta^3(\mathbf{x}-\mathbf{x}')$. The Green's function must satisfy the boundary condition

$$G(\mathbf{k},\mathbf{x}+\mathbf{R}_j;E) = \exp(i\mathbf{k}\cdot\mathbf{R}_j)\, G(\mathbf{k},\mathbf{x};E), \tag{135}$$

imposed by Bloch's theorem. The Green's function is expanded in terms of plane-wave eigenfunctions of the free-particle operator,

$$H_0 = (\hbar^2/2m)\nabla^2 + E. \tag{136}$$

Using such a plane-wave expansion, we find by direct substitution in Eq. (134)

$$G(\mathbf{k},\mathbf{x}-\mathbf{x}';E) = -\frac{1}{\Omega_{\mathbf{x}}} \sum_{\mathbf{G}} \frac{\exp[i(\mathbf{k}+\mathbf{G})\cdot(\mathbf{x}-\mathbf{x}')]}{(-\hbar^2/2m)(\mathbf{k}+\mathbf{G})^2 - E}, \tag{137}$$

where $\Omega_{\mathbf{x}}$ is the volume of the primitive cell.

The Schrödinger equation associated with a scattering potential $V(\mathbf{x})$ for a Bloch state in a crystal,

$$[-(\hbar^2/2m)\nabla^2 + V(\mathbf{x})]\,\psi(\mathbf{k},\mathbf{x}) = \epsilon(\mathbf{k})\,\psi(\mathbf{k},\mathbf{x}), \tag{138}$$

is then transformed, using standard results of scattering theory, into an integral equation

$$\psi(\mathbf{k},\mathbf{x}) = \int d^3\mathbf{x}'\, G(\mathbf{k},\mathbf{x}-\mathbf{x}';E)\, V(\mathbf{x}')\,\psi(\mathbf{k},\mathbf{x}'), \tag{139}$$

Within the muffin-tin approximation, the true crystal potential $V(\mathbf{x})$ appearing in the integral equation Eq. (139) is replaced by its spherically symmetric muffin-tin form, $V_{\mathrm{MT}}(r)$, and we obtain

$$\psi(\mathbf{k},\mathbf{x}) = \int d^3\mathbf{x}'\, G(\mathbf{k},\mathbf{x}-\mathbf{x}';E)\, V_{\mathrm{MT}}(r')\,\psi(\mathbf{k},\mathbf{x}'). \tag{140}$$

In practical implementations of the Green's-function method, the Bloch state $\psi(\mathbf{k},\mathbf{x})$ is expanded within the muffin-tin sphere as

$$\psi(\mathbf{k},\mathbf{x}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} C_{lm}(\mathbf{k})\, Y_l^m(\hat{\mathbf{x}})\, R_l(E;r). \tag{141}$$

When the expansion Eq. (141) is substituted into the integral equation, a determinant compatibility relation is obtained, and its solution allows a determination of the energy eigenvalues E and the expansion coefficients, $C_{lm}(\mathbf{k})$. Explicit expression for the Green's-function matrix elements can be found in the literature (Bassani and Pastori Parravicini, 1975).

Application of the boundary conditions and Bloch's theorem lead to a secular equation similar to that encountered in the discussion of the LMTO method [Eq. (132)]. However, in the KKR–Green's-function method, energy-dependent tail functions of the form of spherical Bessel and Neumann functions $[j_l(\kappa r)$ and $n_l(\kappa r)$ where $\kappa \sim \sqrt{E}]$ are used instead of the functions $(r/r_s)^l$ and $(rs/r)l+1$ appearing in the definition of $\chi_{lm}(E,\mathbf{x})$ [Eq. (126)]. Consequently, in the KKR–Green's-function method the structure constants $S^{lm}_{l'm'}(\mathbf{k},E)$ are energy-dependent. The resulting secular equations are known as the *KKR equations*.

As in the case of the linearized APW (LAPW) and MTO (LMTO) method, the Green's function can be made more efficient and cast in an eigensystem form by performing a linearization of the Green's-function matrix elements. The basic principles of the Green's-function method are very similar to those of the APW method, and both methods lead to the same results when the same muffin-tin potential is used.

2.4 Applications of Band Theory

The problem of calculating the electronic band structure of materials has interested a very large number of workers in the field of solid-state physics. It is clearly impossible even to summarize the outstanding achievements that have been accomplished in this field. While the quantitative features of the band structure depend on the details of the crystal potential, it is useful to emphasize general features. The purpose of this section is merely to put the somewhat abstract concepts introduced above and in the preceding section in a more practical context where the reader can appreciate the high degree of agreement between calculated band structures and experimental measurements. Moreover, we would like to give the reader a sense of the order of magnitude of important quantities: energy-band gap, bandwidth, etc.

2.4.1 Pseudopotential Band Structure of Silicon. We now present results of pseudopotential calculations of the band structure, density of states, and valence charge density for silicon (Cohen and Chelikowsky, 1982). Silicon crystallizes in the cubic diamond structure. This structure is based on a face-centered-cubic Bravais lattice (see Fig. 2) with a basis of two atoms: one at $\mathbf{d}_1=\mathbf{0}$ and one at $\mathbf{d}_2=(a/4)(\hat{\mathbf{x}}+\hat{\mathbf{y}}+\hat{\mathbf{z}})$, where a is the length of the cube side. The atoms exhibit perfect tetrahedral coordination with bond angles $\theta=\arccos\left(-\frac{1}{3}\right)=109.47°$. The cubic diamond structure is indicated in Fig. 16 (Kittel, 1957). It possesses inversion symmetry with respect to the point $(a/8)(\hat{\mathbf{x}}+\hat{\mathbf{y}}+\hat{\mathbf{z}})$ at the center of the bond joining the two basis atoms in the primitive cell.

The energy-band structure of silicon is shown in Fig. 17(a) where the energy bands $\epsilon_n(\mathbf{k})$ are plotted along high-symmetry $\mathbf{k}$ directions within the Brillouin zone (Cohen and Chelikowsky, 1982). These calculations were performed with the empirical pseudopotential method. The method is not self-consistent because the crystal potential is not recalculated due to the screening by the valence electrons, and is called empirical because the local and nonlocal pseudopotential matrix elements of Eq. (117) are fitted to reflectivity and photoemission measurements. The symbols Γ_j, X_j, and L_j label the irreducible representations of the point group at high-symmetry $\mathbf{k}$ points within the Brillouin zone (Bouckaert *et al.*, 1936). Spin-orbit interac-

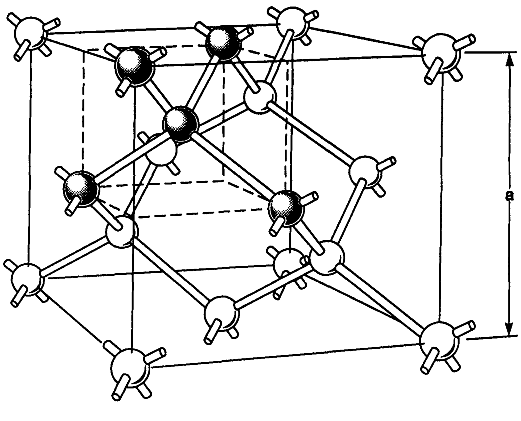

FIG. 16. Cubic diamond structure. The cubic diamond structure is based on a face-centered-cubic Bravais lattice (see Fig. 2) with a two-atom basis. (Reproduced with permission from Kittel, 1957, p. 37.)

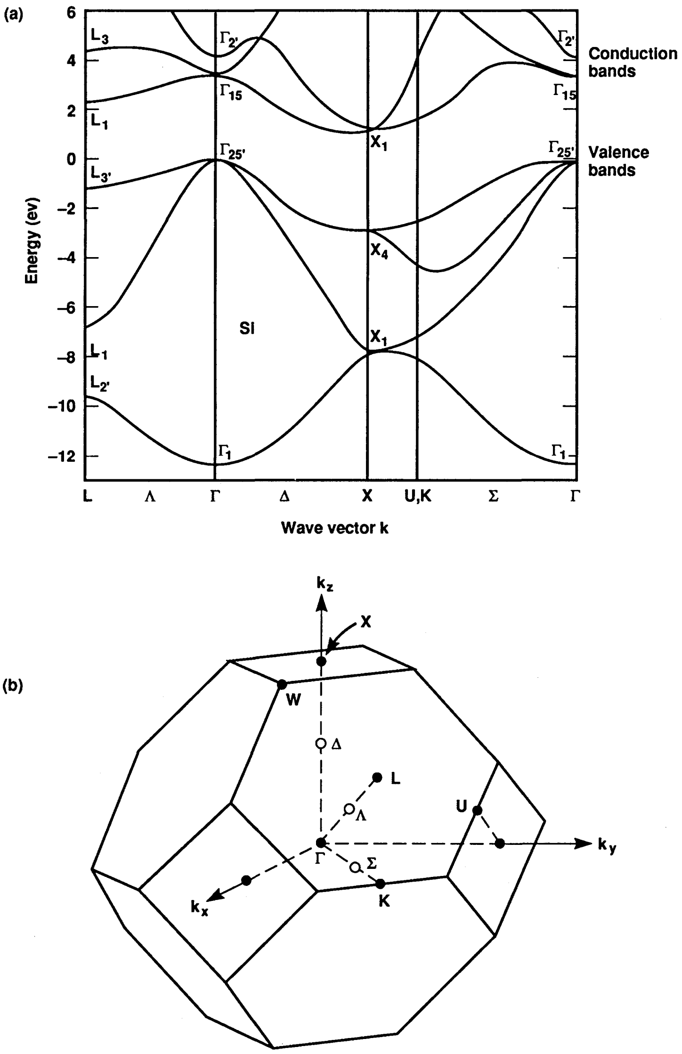

FIG. 17. (a) Energy-band structure of silicon as calculated by the empirical pseudopotential method. The calculations were performed using empirical nonlocal pseudopotentials. Spin-orbit effects are neglected. The valence-band maximum is taken as the zero of energy. The filled valence bands span the energy region $\epsilon < 0$ and the empty conduction bands span the energy region $\epsilon > 0$ as indicated. The symbols Γ_j, X_j, and L_j label the irreducible representations of the point group at high-symmetry **k** points within the Brillouin zone. Standard notation (Bouckaert *et al.*, 1936) is used to indicate the symmetry points and lines within the Brillouin zone (also shown in Fig. 2) shown in (b). (Reproduced with permission from Cohen and Chelikowsky, 1982, p. 230.) (b) Brillouin zone for the face-centered-cubic lattice. (Reproduced with permission from Kittel, 1963, p. 213.)

tions were not considered in these calculations. The Brillouin zone is shown in Fig. 17(b) with the appropriate symmetry labels [see also Fig. 2(b)].

Inspection of Fig. 17(a) reveals the existence of an *energy-band gap* separating the filled valence bands and the empty conduction bands, and, consequently, silicon is a semiconductor [see Fig. 7(a)]. Silicon exhibits an indirect band gap since the valence-band maximum and the conduction-band minimum do not occur at the same point in the Brillouin zone (**k** space): The valence-band maximum occurs at the Γ point ($\mathbf{k}=\mathbf{0}$) whereas the conduction-band minimum occurs near the X point [$\mathbf{k}=(2\pi/a)(1,0,0)$].

We now discuss the symmetry features of the energy band at the center ($\mathbf{k}=\mathbf{0}$) of the Brillouin zone. The valence-band maximum is at $\mathbf{k}=\mathbf{0}$. It is threefold degenerate (neglecting spin) and belongs to the irreducible representation $\Gamma_{25'}^{v}$, whose basis functions transform like xy,yz,zx about the center of the line joining the two basis atoms in the primitive cell. The threefold-degenerate irreducible representation $\Gamma_{25'}^{v}$ of the valence-band maximum may be formed from the ***bonding*** linear combination of p orbitals on the two silicon atoms. Similarly, the threefold-degenerate irreducible representation Γ_{15}^{c} of the conduction band may be formed from the ***antibonding*** linear combination of the p orbitals on the two silicon atoms. The ***bonding*** linear combination of the s orbitals on the silicon atoms forms the one-dimensional Γ_1^{v} irreducible rep-

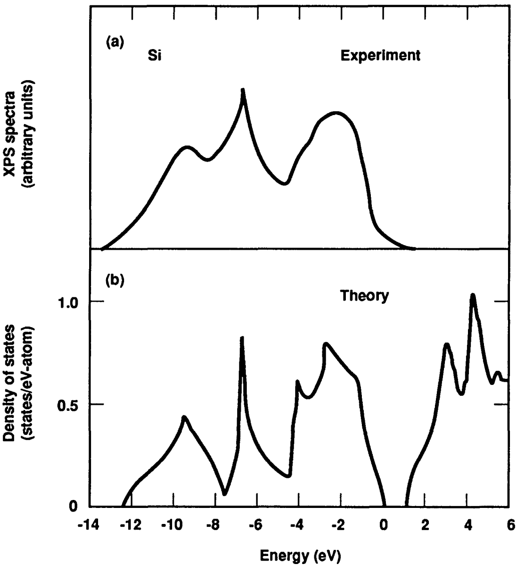

FIG. 18. Calculated electron density of states for silicon (Cohen and Chelikowsky, 1982) compared with x-ray photoemission spectroscopy measurements (Ley *et al.*, 1974). The calculations were performed using empirical nonlocal pseudopotentials. Spin-orbit effects are neglected. The van Hove singularities discussed in Sec. 1.5.3 and shown in Fig. 8 are clearly apparent in the profile of the silicon density of states. (Reproduced with permission from Cohen and Chelikowsky, 1982, p. 231.)

resentation at the bottom of the valence band, whereas the *antibonding* linear combination of the *s* orbitals on the silicon atoms forms the one-dimensional $\Gamma_{2'}^{c}$ irreducible representation of the conduction band. The inclusion of spin-orbit interactions introduces additional energy-band splittings and degeneracies due to time-reversal symmetry. The ordering and degeneracy of the bands at various points of the Brillouin zone depend sensitively on the details of the crystal potential and on the symmetry of the lattice. If spin-orbit interactions were included in the calculation, the energy bands would be doubly degenerate at each **k** point within the Brillouin zone by a combination of time-reversal and inversion symmetries [see Eq. (79)].

Figure 18 shows the calculated electron density of states for silicon (Cohen and Chelikowsky, 1982) along with x-ray photoemission spectroscopy measurements (Ley *et al.*, 1974). The general features of the density of states observed experimentally are well reproduced by the pseudopotential calculations. The van Hove singularities discussed in Sec. 1.5.3 and shown in Fig. 8 are clearly apparent in the profile of the silicon density of states. In Fig. 19, we indicate the calculated valence pseudocharge density for silicon obtained by summing the individual energy-band contributions over the occupied bands (Cohen and Chelikowsky, 1982). X-ray measurements of the charge density of silicon are also shown in Fig. 19. The agreement between the calculated and measured charge density is excellent. The pseudopotential calculation reproduces very well the essential feature of the silicon covalent bond, i.e., the accumulation of electronic charge in the bond region between nearest-neighbor silicon atoms.

2.4.2 Electron-Structure Trends among the Elements. An important goal of electron-structure theory is to establish trends and relationships between calculated energy bands and the chemical composition and crystal structure of solids. Much effort has been

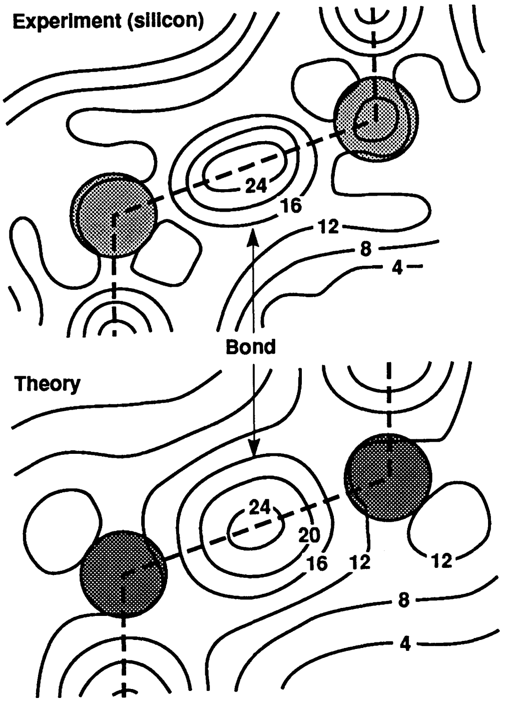

FIG. 19. Valence charge density for silicon as determined from x-ray measurements (Yang and Coppens, 1974) and as calculated using a nonlocal pseudopotential. The atomic cores are represented by the shaded regions where the pseudopotential results are not expected to be accurate. The charge contours are measured in units of e/Ω_x, where Ω_x is the volume of the unit cell. (Reproduced with permission from Cohen and Chelikowsky, 1982, p. 235.)

expended at refining the quantitative aspects of these relationships. Space limitation does not permit us to present trends of calculated quantities (energy-band gaps, bandwidths, etc.) among the various elements. Excellent contributions have appeared in the literature (Moruzzi *et al.*, 1978; Papaconstantopoulos, 1986) with the purpose of collecting calculated energy-band structures and ground-state properties of many elements. The reader interested in getting a sense of general band-structure trends among the elements is urged to consult these references.

3. FIRST-PRINCIPLES TOTAL-ENERGY METHODS

In the preceding section, our discussion has been restricted to the electron band-structure problem, i.e., the calculation of the crystal energy bands $\epsilon_n(\mathbf{k})$ and corresponding Bloch states $|\psi_n(\mathbf{k})\rangle$. We now turn to the calculation of the *total energy* of a solid. The calculation of the total energy involves the evaluation of the electronic charge density $n(\mathbf{x})$ from a knowledge of the crystal Bloch solutions $|\psi_n(\mathbf{k})\rangle$.

A tremendous conceptual advance was realized in the mid-1960s by Hohenberg, Kohn, and Sham (Hohenberg and Kohn, 1964; Kohn and Sham, 1965), who formulated the *density-functional theory.* This framework provides an extremely powerful technique for electronic-structure and total-energy calculations in solids, atoms, and molecules. In fact, it is the single most widely used formalism in first-principles electronic-structure calculations within a band-theoretical framework.

3.1 Density-Functional Theory

The basic postulate of the density-functional theory is that the ground-state energy E_N of a system of N electrons moving in a given external potential $v(\mathbf{x})$ can be obtained from a knowledge of the self-consistent charge density $n(\mathbf{x})$. This concept offers tremendous computational advantages since the electron density $n(\mathbf{x})$ is considered the basic variable rather than the complex many-electron wave function.

We now indicate the basic postulates of the density functional theory (Lang, 1973). Consider a many-electron system in its (nondegenerate) ground state and moving in a static external potential $v(\mathbf{x})$. We denote the many-electron ground state by $|\Psi\rangle$. The Hamiltonian operator of the system is

$$H=T+V+U, \tag{142a}$$

where

$$T=\frac{\hbar^2}{2m}\int d^3x \nabla\psi^\dagger(\mathbf{x})\nabla\psi(\mathbf{x}) \tag{142b}$$

is the kinetic energy,

$$V=\int d^3x\, v(\mathbf{x})\psi^\dagger(\mathbf{x})\psi(\mathbf{x}) \tag{142c}$$

is the interaction described by the external potential $v(\mathbf{x})$, and

$$U=\frac{e^2}{2}\int d^3x\, d^3y \frac{1}{|\mathbf{x}-\mathbf{y}|} \times\psi^\dagger(\mathbf{x})\psi^\dagger(\mathbf{y})\psi(\mathbf{y})\psi(\mathbf{x}) \tag{142d}$$

is the Coulomb electron–electron interaction. [In Eqs. (142), $\psi(\mathbf{x})$ refers to a fermion field operator and not to a wave function.] The equilibrium electron density $n(\mathbf{x})$ corresponding to the ground state $|\Psi\rangle$ is the expectation value

$$n(\mathbf{x})\equiv\langle\Psi|\psi^\dagger(\mathbf{x})\psi(\mathbf{x})|\Psi\rangle, \tag{143}$$

and is clearly a functional of the external potential $v(\mathbf{x})$. However, it can also be shown that $v(\mathbf{x})$ is *uniquely* determined by $n(\mathbf{x})$. Hohenberg and Kohn demonstrated that the external potential $v(\mathbf{x})$ of a nondegenerate ground state is (apart from an additive constant) a *unique functional of the ground-state electronic charge density* $n(\mathbf{x})$: Only one $v(\mathbf{x})$ can result in a given $n(\mathbf{x})$. In other words, the density distribution $n(\mathbf{x})$ associated with a ground state $|\Psi\rangle$ corresponding to an external potential $v(\mathbf{x})$ cannot be reproduced by the ground state $|\Psi'\rangle$ corresponding to another potential $v'(\mathbf{x})\neq v(\mathbf{x})$, apart from an additive constant. As a result, the many-electron ground state $|\Psi\rangle$ is also a unique functional of $n(\mathbf{x})$, since the potential $v(\mathbf{x})$ fixes the Hamiltonian H that determines $|\Psi\rangle$.

Let us define an *electron energy functional,*

$$E_v[n(\mathbf{x})]=\int d^3x\, v(\mathbf{x})n(\mathbf{x})+F[n(\mathbf{x})], \tag{144a}$$

where the quantity $F[n(\mathbf{x})]$ is a universal [i.e., it does not explicitly depend on $v(\mathbf{x})$] functional of $n(\mathbf{x})$ defined as

$$F[n(\mathbf{x})] \equiv \langle \Psi | T+U | \Psi \rangle. \qquad (144b)$$

An important theorem of the density-functional theory is that the energy functional $E_v[n(\mathbf{x})]$ obeys a variational principle in the charge density $n(\mathbf{x})$: For the correct ground-state density $n(\mathbf{x}) = \hat{n}(\mathbf{x})$ associated with the external potential $v(\mathbf{x})$ and integrating to the total number N of electrons in the system,

$$\int d^3x\, \hat{n}(\mathbf{x}) = N, \qquad (145)$$

the functional $E_v[\hat{n}(\mathbf{x})]$ evaluated at the density $n(\mathbf{x}) = \hat{n}(\mathbf{x})$ is minimum and equal to the ground-state energy E_N. Therefore, E_N is the minimum value of $E_v[n(\mathbf{x})]$ with respect to variations in densities $n(\mathbf{x})$, i.e.,

$$E_N = \min_{n(\mathbf{x})} E_v[n(\mathbf{x})], \qquad (146)$$

for densities $n(\mathbf{x})$ satisfying

$$\int d^3x\, n(\mathbf{x}) = N. \qquad (147)$$

The density $\hat{n}(\mathbf{x})$ that minimizes the energy functional $E_v[n(\mathbf{x})]$ is the correct ground-state density. The power of the Hohenberg–Kohn theorem is that useful approximations can be made, even through the exact form of the functional is not known.

It is convenient to separate the electron-electron Coulomb term from the functional $F[n(\mathbf{x})]$ and write

$$F[n(\mathbf{x})] = \frac{e^2}{2} \int d^3x\, d^3y \frac{n(\mathbf{x})n(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} + T_s[n(\mathbf{x})] + E_{\text{xc}}[n(\mathbf{x})], \qquad (148)$$

where $T_s[n(\mathbf{x})]$ is a kinetic energy functional for a noninteracting system of electrons of density $n(\mathbf{x})$ and $E_{\text{xc}}[n(\mathbf{x})]$ is a functional which contains all the many-body effects of exchange and correlation. The electron energy functional can therefore be written as

$$E_v[n(\mathbf{x})] = \int d^3x\, v(\mathbf{x})n(\mathbf{x}) + \frac{e^2}{2} \int d^3x\, d^3y \frac{n(\mathbf{x})n(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} + T_s[n(\mathbf{x})] + E_{\text{xc}}[n(\mathbf{x})]. \qquad (149)$$

The functional $E_v[n(\mathbf{x})]$ given in Eq. (149), when evaluated at the correct ground-state density $n(\mathbf{x}) = \hat{n}(\mathbf{x})$, is the expression for the ground-state energy of a many-electron system in a fixed external potential $v(\mathbf{x})$.

The exact form of $E_v[n(\mathbf{x})]$ is, of course, unknown. There are essentially three difficulties in evaluating the functional $E_v[n(\mathbf{x})]$ in actual calculations:

1. one needs to calculate the self-consistent $n(\mathbf{x})$ that minimizes $E_v[n(\mathbf{x})]$ and yields E_N;
2. the evaluation of the kinetic energy $T_s[n(\mathbf{x})]$ from the density $n(\mathbf{x})$ is not trivial since no information on the wave function is available; and
3. the exact form of the exchange- correlation functional $E_{\text{xc}}[n(\mathbf{x})]$ is unknown, except for a few simple systems, and approximations have to be made.

The first two difficulties can be overcome by a variational principle: Kohn and Sham (1965) used the variational principle for the minimization of $E_v[n(\mathbf{x})]$ to derive a set of one-electron equations,

$$\{ -(\hbar^2/2m)\nabla^2 + v_{\text{eff}}[n(\mathbf{x})] \} \psi_i(\mathbf{x}) = \epsilon_i \psi_i(\mathbf{x}), \qquad (150)$$

where $\psi_i(\mathbf{x})$ are the one-electron wave functions, ϵ_i are the one-electron energy eigenvalues, and $v_{\text{eff}}[n(\mathbf{x})]$ is an integro-differential potential operator defined as

$$v_{\text{eff}}[n(\mathbf{x})] = v(\mathbf{x}) + e^2 \int d^3y \frac{n(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} + \frac{\delta E_{\text{xc}}[n(\mathbf{x})]}{\delta n(\mathbf{x})} \equiv \varphi(\mathbf{x}) + \mu_{\text{xc}}[n(\mathbf{x})], \qquad (151)$$

where we have defined the *electrostatic potential*

$$\varphi(\mathbf{x}) \equiv v(\mathbf{x}) + e^2 \int d^3y \frac{n(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}, \qquad (152a)$$

and the *exchange-correlation potential*

$$\mu_{\text{xc}}[n(\mathbf{x})] \equiv \frac{\delta E_{\text{xc}}[n(\mathbf{x})]}{\delta n(\mathbf{x})}. \qquad (152b)$$

The self-consistent solution of the *single-particle Kohn–Sham equations* Eq. (150) forms the basis of the most self-consistent electron structure calculations performed in electron band theory.

The charge density $n(\mathbf{x})$ is calculated from a knowledge of the single-particle wave functions $\psi_i(\mathbf{x})$,

$$n(\mathbf{x}) = \sum_{i(\text{occ.})} |\psi_i(\mathbf{x})|^2, \tag{153}$$

where the summation runs over occupied states. The kinetic energy functional can be evaluated from a knowledge of the energy eigenvalues ϵ_i (Lang, 1973),

$$T_s[n(\mathbf{x})] = \sum_{i(\text{occ.})} \epsilon_i - \int d^3x v_{\text{eff}}[n(\mathbf{x})]n(\mathbf{x}). \tag{154}$$

The solution of the one-electron Kohn-Sham equations requires a knowledge of the exchange-correlation potential $\mu_{\text{xc}}[n(\mathbf{x})]$. Since the exact form of $E_{\text{xc}}[n(\mathbf{x})]$ is not known, the *local-density approximation (LDA)* is almost universally made in actual calculations. In the LDA, the exchange-correlation functional $E_{\text{xc}}[n(\mathbf{x})]$ is approximated by

$$\begin{aligned} E_{\text{xc}}[n(\mathbf{x})] &\approx E_{\text{xc}}^{\text{LDA}}[n(\mathbf{x})] \\ &= \int d^3x\, n(\mathbf{x})\epsilon_{\text{xc}}[n(\mathbf{x})], \end{aligned} \tag{155}$$

where $\epsilon_{\text{xc}}[n(\mathbf{x})]$ is the exchange-correlation energy per electron of a *uniform electron gas* of density n evaluated at the *local density* $n(\mathbf{x})$. The exchange-correlation potential in the LDA becomes

$$\begin{aligned} \mu_{\text{xc}}^{\text{LDA}}[n(\mathbf{x})] &= \frac{\delta E_{\text{xc}}^{\text{LDA}}[n(\mathbf{x})]}{\delta n(\mathbf{x})} \\ &= \frac{d}{dn(\mathbf{x})}\{n(\mathbf{x})\epsilon_{\text{xc}}[n(\mathbf{x})]\}. \end{aligned} \tag{156}$$

There exist a number of expressions for $\epsilon_{\text{xc}}[n(\mathbf{x})]$ in the literature (Lang, 1973).

The Kohn–Sham equations must be solved self-consistently: for a given atomic structure, one starts with an assumed density $n(\mathbf{x})$ and constructs the potential $v_{\text{eff}}[n(\mathbf{x})]$ from which a new density $n(\mathbf{x})$ is obtained. The procedure is repeated until self-consistency is achieved and the density $n(\mathbf{x})$ and potential $v_{\text{eff}}[n(\mathbf{x})]$ no longer change. A schematic flow diagram of the self-consistent procedure is indicated in Fig. 20 (Cohen and Chelikowsky, 1982). The solution of the one-electron Kohn-Sham equations can be obtained with the band-structure techniques discussed in Sec. 2. When the evaluation of the Hamiltonian matrix elements does not require the use of empirical parameters, the methods described in this section are commonly referred to as *first-principles* (or *ab initio*) *self-consistent electron-structure calculations*.

3.2 Total Energy

The self-consistent solution of the Kohn–Sham equations allows a calculation of the ground-state energy of the system by minimization of the energy functional

$$\begin{aligned} E_v[n(\mathbf{x})] &= \int d^3x\, v(\mathbf{x})n(\mathbf{x}) \\ &+ \frac{e^2}{2}\int d^3\mathbf{x}\, d^3\mathbf{y} \frac{n(\mathbf{x})n(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} \\ &+ T_s[n(\mathbf{x})] + E_{\text{xc}}[n(\mathbf{x})], \end{aligned} \tag{157}$$

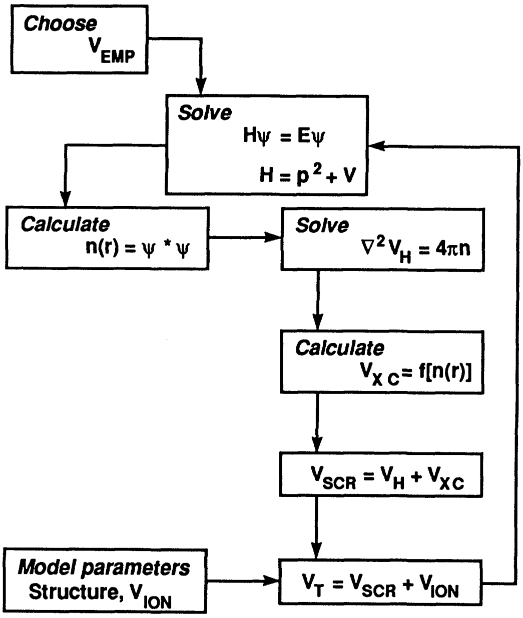

FIG. 20. Flow diagram of the self-consistent solution of Kohn–Sham one-electron equations. (Reproduced with permission from Cohen and Chelikowsky, 1982, p. 226.)

which can be written in a somewhat more useful form in terms of the energy eigenvalues ϵ_i,

$$\begin{aligned} E_v[n(\mathbf{x})] = {} & \sum_{i(\text{occ.})} \epsilon_i - \frac{e^2}{2} \int d^3x\, d^3y \\ & \times \frac{n(\mathbf{x})n(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} + E_{\text{xc}}[n(\mathbf{x})] \\ & - \int d^3x\, \mu_{\text{xc}}[n(\mathbf{x})]n(\mathbf{x}), \qquad (158) \end{aligned}$$

where we have used the expression for $T_s[n(\mathbf{x})]$ given by Eq. (154). In order to evaluate the total energy of the solid, one must add the ion-ion Coulomb interaction to the electron functional $E_v[n(\mathbf{x})]$ given by Eqs. (157) and (158):

$$\begin{aligned} E_{\text{total}} & = E_{\text{ion-ion}} + E_v[n(\mathbf{x})] \\ & = e^2 \sum_{I<J} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} + \int d^3x\, v(\mathbf{x})n(\mathbf{x}) \\ & \quad + \frac{e^2}{2} \int d^3x\, d^3y\, \frac{n(\mathbf{x})n(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} \\ & \quad + T_s[n(\mathbf{x})] + E_{\text{xc}}[n(\mathbf{x})] \\ & \equiv E_{\text{ion-ion}} + E_{\text{hartree}} + E_{\text{kinetic}} + E_{\text{xc}}, \qquad (159) \end{aligned}$$

where $\mathbf{R}_I$ ($\mathbf{R}_J$) are the positions of the ions of ionic charge Z_I (Z_J). In actual calculations, the functional $E_v[n(\mathbf{x})]$ to be minimized is evaluated by Eq. (158) using the energy eigenvalues ϵ_i of the Kohn–Sham equations. The resulting expression,

$$\begin{aligned} & E_{\text{total}}[\{\psi_i(\mathbf{x})\},\{\mathbf{R}_I\}] \\ & \quad = e^2 \sum_{I<J} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} + \sum_{i(\text{occ.})} \epsilon_i \\ & \quad - \frac{e^2}{2} \int d^3x\, d^3y \frac{n(\mathbf{x})n(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} + E_{\text{xc}}[n(\mathbf{x})] \\ & \quad - \int d^3x\, \mu_{\text{xc}}[n(\mathbf{x})]n(\mathbf{x}), \qquad (160) \end{aligned}$$

gives the total energy of the system as a function of the electronic variables $\{\psi_i(\mathbf{x})\}$ and ionic variables $\{\mathbf{R}_I\}$.

3.3 Recent Developments in Density-Functional Theory: Harris Functional

An efficient strategy to perform total-energy LDA calculations using a non-self-consistent, yet accurate, technique has recently been proposed (Harris, 1985). The starting point of this approach is to construct an initial charge density by summing the individual overlapping contributions of atomiclike densities,

$$n_f(\mathbf{x}) = \sum_k n_k(\mathbf{x}) \qquad (161)$$

where the isolated atomiclike charge densities $n_k(\mathbf{x})$ need not be the true atomic densities and the density $n_f(\mathbf{x})$ is the sum of these individual contributions or fragments. A one-electron potential $\hat{v}_{\text{eff}}[n_f(\mathbf{x})])$ is then constructed from the overlapping fragment densities $n_f(\mathbf{x})$,

$$\hat{v}_{\text{eff}}[n_f(\mathbf{x})] = \varphi_f(\mathbf{x}) + \mu_{\text{xc}}[n_f(\mathbf{x})], \qquad (162)$$

where the potential $\varphi_f(\mathbf{x})$ is given by Eq. (152a) and evaluated for the density $n_f(\mathbf{x})$,

$$\varphi_f(\mathbf{x}) \equiv v(\mathbf{x}) + e^2 \int d^3y\, \frac{n_f(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}. \qquad (163)$$

Similarly, the exchange-correlation potential $\mu_{\text{xc}}[n_f(\mathbf{x})]$ is given by Eq. (152b) and evaluated for the density $n_f(\mathbf{x})$. The energy eigenvalues $\hat{\epsilon}_i$ corresponding to the potential $\hat{v}_{\text{eff}}[n_f(\mathbf{x})]$ are determined by the solution of Eq. (150).

The difference between the true self-consistent charge density $n(\mathbf{x})$ and $n_f(\mathbf{x})$ is

$$\Delta n(\mathbf{x}) \equiv n(\mathbf{x}) - n_f(\mathbf{x}). \qquad (164)$$

Similarly, the difference between the true self-consistent potential $v_{\text{eff}}[n(\mathbf{x})]$ and the potential $\hat{v}_{\text{eff}}[n_f(\mathbf{x})]$ is

$$\begin{aligned} \Delta v_{\text{eff}}(\mathbf{x}) & \equiv v_{\text{eff}}[n(\mathbf{x})] - \hat{v}_{\text{eff}}[n_f(\mathbf{x})] \\ & = [\varphi(\mathbf{x}) - \varphi_f(\mathbf{x})] + \{\mu_{\text{xc}}[n(\mathbf{x})] \\ & \quad - \mu_{\text{xc}}[n_f(\mathbf{x})]\}, \qquad (165) \end{aligned}$$

which is assumed to be small.

It can be shown (Harris, 1985) that the use of the $n_f(\mathbf{x})$ in the total-energy functional, Eq. (160), introduces corrections which are of order $[\Delta n(\mathbf{x})]^2$ and $[\Delta v_{\text{eff}}(\mathbf{x})]^2$. The total energy can be expressed entirely in terms of the overlapping fragment density $n_f(\mathbf{x})$ and the energy eigenvalues $\hat{\epsilon}_i$,

$$\begin{aligned}E_{\text{total}} \approx e^2 \sum_{I<J} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} + \sum_{i(\text{occ.})} \hat{\epsilon}_i \\ - \frac{e^2}{2} \int d^3x\, d^3y\, \frac{n_f(\mathbf{x}) n_f(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} \\ + E_{\text{xc}}[n_f(\mathbf{x})] \\ - \int d^3x\, \mu_{\text{xc}}[n_f(\mathbf{x})] n_f(\mathbf{x}), \qquad (166)\end{aligned}$$

to within the corrections indicated above. The importance of the method derives from the fact that an accurate evaluation of the total energy E_{total} is possible through the use of Eq. (166) without the need to perform a self-consistent calculation on the charge density and the potential. Successful applications of the Harris functional have been performed on numerous solid-state systems and molecules and will be discussed in the next section.

3.4 Applications of Total-Energy Methods

Self-consistent first-principles total-energy methods have been applied to a wide variety of problems: determination of structural properties of bulk materials and solid surfaces, prediction of phase transformation between solid phases, electron structure of alloys, calculation of phonon dispersion relations in bulk and at surfaces, etc. These methods have reached a high level of sophistication and maturity. The interested reader is advised to consult the numerous recent reviews which have appeared in the field (DeVreese and Van Camp, 1985; Srivastava and Weaire, 1987; Jones and Gunnarson, 1989).

The examples given in this section will give the reader a sense of the level of accuracy and the predictive power of modern first-principles total-energy methods. As an illustration, we focus on the effect of hydrostatic pressure on the atomic and electron structure of the elements. First-principles total-energy calculations have had a tremendous impact on this area of condensed-matter physics and form the basis of our understanding of pressure-induced structural transformations among the elements (McMahan, 1986). First we examine the ability of first-principles pseudopotential methods to predict the ground-state properties and pressure-induced phase transformations of various tetrahedrally bonded materials exhibiting covalent bonding at atmospheric pressure. We then indicate similar predictions for alkali and alkaline-earth metals as afforded by the LMTO method. The first-principles pseudopotential method has had an enormous impact on our understanding of the electronic structure of simple metals and covalent materials, whereas all-electron methods such as the LMTO and the APW have had the same level of impact for transition metals. We then compare the predictions obtained by different electron-structure methods for a single material. Finally, we present recent applications of the Harris functional to a wide range of materials and compare its predictions with those afforded by self-consistent LDA calculations and with available experiments.

3.4.1 First-Principles Pseudopotentials: Covalent Materials. As a first illustration of the power of modern total-energy methods, we will examine the determination of the crystal stability and pressure-induced phase transformations in covalently bonded elemental and compound semiconductors. There is ample experimental and theoretical evidence that the application of pressure is inimical to the covalent bond. In a general argument, localization of charge in the covalent bond minimizes the potential energy at the expense of the kinetic energy. Given the respective $V^{-1/3}$ and $V^{-2/3}$ scaling of the potential and kinetic energy contributions with volume V, the latter must eventually dominate as V decreases, favoring more uniform electron density and more close-packed structures. An important aspect of pressure-induced phase transformations in materials is the general tendency for heavier members of a given group to represent high-pressure forms of lighter members (McMahan, 1986). This behavior follows from an increase in the size of the ion core radius relative to the atomic or Wigner–Seitz radius as one moves to heavier group members, thus effectively compressing the valence electrons between the two radii. The examples indicated below refine this perspective.

The first calculation which established that the determination of the static structural properties, the crystal stability, and the pressure-induced phase transformations could be achieved by first-principles total-energy methods (within the LDA) is due to Yin and Cohen (1982b). Their calculations are based on the first-principles pseudopotential method, and the input information consists only of the atomic number of the elements and the crystal structure. Numerous subsequent applications of first-principles pseudopotential total-energy methods have established that the calculated values agree with observed quantities generally to within 1% for lattice constants, 6% for bulk moduli, and 5% for cohesive energies. A summary of calculated bulk properties using first-principles total-energy methods is indicated in Table 2, where calculated values are compared with available experimental measurements for elemental and compound tetrahedrally coordinated semiconductors and insulators. Inspection of Table 2 reveals the tremendous predictive power of modern total-energy methods formulated within the LDA.

Table 2. Calculated bulk ground-state properties of elemental and compound tetrahedrally coordinated semiconductors and insulators using first-principles pseudopotential total-energy methods within the local-density approximation. The quantities shown are the equilibrium lattice constant a_0, the bulk modulus B_0, the cohesive energy E_{coh}, and the transition pressure p_{trans}. For each material, the first line gives the calculated value and the second line (in parentheses) gives the experimental result.

Material	a_0 (Å)	B_0 (GPa)	E_{coh} (eV)	p_{trans} (GPa)
C[a]	3.60	441	7.57	
	(3.57)	(443)	(7.35)	
Si[a]	5.45	98	4.67	9.9
	(5.43)	(99)	(4.63)	(12.5)
Ge[a]	5.66	73	4.02	9.6
	(5.65)	(77)	(3.85)	(10.0)
AlP[b]	5.471	845		9.3
	(5.451)	(860)		(14.0–17.0)
AlAs[b]	5.678	710		7.6
	(5.662)	(770)		(12.0)
AlSb[b]	6.153	543		5.6
	(6.135)	(580)		(8.3)
GaP[b]	5.386	868		21.7
	(5.451)	(887)		(20.0–24.0)
GaAs[b]	5.601	708		16.0
	(5.653)	(748)		(16.0–19.0)
GaSb[b]	6.032	557		6.3
	(6.118)	(570)		(6.7–8.4)
InP[b]	5.688	700		12.8
	(5.869)	(710)		(10.5–11.0)
InAs[b]	5.906	581		8.4
	(6.036)	(600)		(7.0–8.4)
InSb[b]	6.539	470		3.3
	(6.490)	(483)		(2.3)

[a] Cohen (1986).
[b] Zhang and Cohen (1987).

We now examine the equation of state and pressure-induced phase transformations in the group-IV elements C, Si, and Ge. Although these elements all have the same number of valence electrons participating in the bonding, an essential difference between C and Si or Ge is the absence of *p* states in the core of C. For Si and Ge, the *p* core states push the *p* valence electrons into the bond region between the atoms to form a single peak in the electron charge density. For C, there are no *p* core states or repulsive *p* pseudopotential and the electron charge density exhibits two peaks along the C—C bond. This difference in core structure is at the origin of many differences in the bonding properties and equations of state between C and Si or Ge.

Results of first-principles pseudopotential total-energy calculations are shown in Fig. 21 for carbon (C: Fahy and Louie, 1984), silicon (Si: Yin and Cohen, 1982b) and germanium (Ge: Yin and Cohen, 1982b) where the calculated total (pseudo)energy is calculated as a function of atomic volume. For Si and Ge, the cubic diamond structure is calculated to be the lowest-energy (i.e., most stable) phase at zero pressure and temperature. For C, the total energy of the hexagonal graphite phase [not shown in Fig. 21(a)] is experimentally known to be lower than that of cubic diamond by approximately 27 meV/atom. This result has also been theoretically confirmed (Yin and Cohen, 1984) and demonstrates that very small differences in structural energies can be accurately extracted from first-principles total energy calculations.

The external pressure that must be applied to induce a phase transformation (the transition pressure) can be extracted by a calculation of the enthalpy or graphically by drawing a common tangent joining two total-energy curves. The slope of that common tangent is the negative of the transition pressure between the two phases through the relation $p = -dE/dV$. Inspection of Fig. 21 reveals that first-principles pseudopotential calculations correctly predict that Si and Ge will transform from a cubic diamond to a β-Sn structure under the application of pressure. The pressure-induced cubic diamond $\rightarrow \beta$-Sn structural trans-

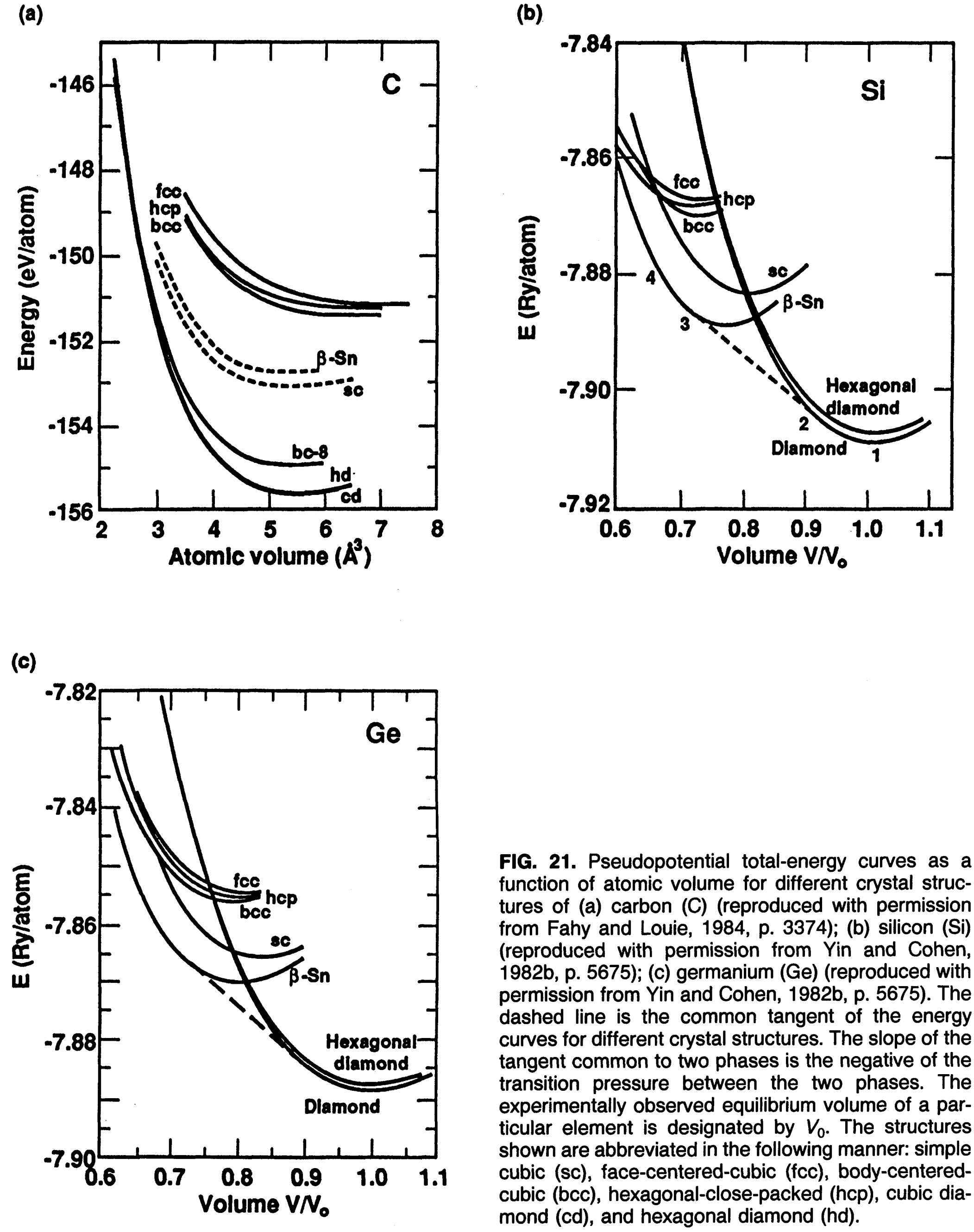

FIG. 21. Pseudopotential total-energy curves as a function of atomic volume for different crystal structures of (a) carbon (C) (reproduced with permission from Fahy and Louie, 1984, p. 3374); (b) silicon (Si) (reproduced with permission from Yin and Cohen, 1982b, p. 5675); (c) germanium (Ge) (reproduced with permission from Yin and Cohen, 1982b, p. 5675). The dashed line is the common tangent of the energy curves for different crystal structures. The slope of the tangent common to two phases is the negative of the transition pressure between the two phases. The experimentally observed equilibrium volume of a particular element is designated by V_0. The structures shown are abbreviated in the following manner: simple cubic (sc), face-centered-cubic (fcc), body-centered-cubic (bcc), hexagonal-close-packed (hcp), cubic diamond (cd), and hexagonal diamond (hd).

formation in Si and Ge does confirm the general tendency, mentioned above, for heavier members of a given group to represent high-pressure forms of lighter members. The complete structural sequence for Si is predicted to be cubic diamond→β-Sn→simple hexagonal →phase VI→hexagonal close-packed→face-centered cubic (Cohen, 1986). The cubic diamond→β-Sn transition pressure is calculated to be 9.9 GPa (12.5 GPa experimentally) for Si and 9.6 GPa (10.0 GPa experimentally) for Ge. On the other hand, no pressure-induced transformation from cubic diamond to a simple phase is predicted for C as is seen in Fig. 21. However, theoretical investigations of more complex phases (Biswas *et al.*, 1984) indicate that C would transform from the cubic diamond structure to the fourfold-coordinated bc-8 (body-centered-cubic Bravais lattice with an eight-atom basis per body-centered-cubic primitive cell) phase under the application of a pressure of approximately 1.2 TPa. It has been shown (Mailhiot and McMahan, 1991) that the phases of carbon characterized by a coordination number (number of first nearest neighbors) higher than four, such as simple-cubic, β-Sn, body-centered-cubic, hexagonal-close-packed, and face-centered-cubic, are mechanically unstable against a spontaneous transformation into the cubic diamond phase. On the basis of these results, one might speculate that all high-coordination phases of carbon would exhibit mechanical instabilities at atmospheric pressure. They also suggest the unlikely occurrence of stable or metastable local structure with coordination number higher than four in liquid and amorphous carbon as well as in carbon clusters, unlike the situation for Si. *Ab initio* molecular dynamics simulations of liquid and amorphous carbon have recently confirmed these expectations (Galli *et al.*, 1989).

3.4.2 All-Electron LMTO Method: Metals. All-electron methods have had a tremendous impact on our understanding of the electron structure of transition metals. Traditionally, the nontransition metals, e.g., alkali and alkali-earth metals, have been treated by means of pseudopotential theory, and the crystal structures predicted by this method are generally in good agreement with experiments (Heine and Weaire, 1970). It has, however, not been entirely straightforward to generalize the pseudopotential method to treat narrow *d*-band materials. All-electron methods such as the LMTO, APW, and KKR, on the other hand, afford accurate descriptions of these materials with great efficiency when implemented in linearized form. An important result emerging from first-principles all-electron calculations is that the crystal structures of elemental metals tend to occur in certain sequences when viewed as functions of atomic number or hydrostatic pressure. For example, the most striking example of this phenomenon occurs for the *d* transition metals where all the three transition series exhibit the same hexagonal-close-packed → body-centered-cubic → hexagonal-close-packed → face-centered-cubic sequence as the *d* bands become progressively filled (Skriver, 1985).

As a second illustration of the predictive power of *ab initio* total-energy methods, we consider the pressure-induced phase transformations occurring in the alkali and alkaline-earth metals. The structural transformation trend mentioned above, although adequate for the *d* transition metals, will fail in cases where states of non-*d* character are equally or more important. Examples of such materials include the alkali, the alkaline-earth, and light actinide metals. We present below results of LMTO calculations that aim at establishing correct structural trends among the alkali and alkaline-earth elements (Skriver, 1985).

According to LMTO calculations, all the alkali metals should exhibit the same crystal structure transformation, body-centered-cubic → hexagonal-close-packed → face-centered-cubic, as the atomic volume is reduced. The predicted structural energy differences for the five alkali metals as a function of the relative volume V/V_0 and LMTO pressure p are shown in Fig. 22(a). These structural transitions are driven by the pressure-induced lowering of the initially unoccupied *d* states through the Fermi level, thereby gradually transferring electrons from the *s* into the *d* band (Skriver, 1985; McMahan, 1986). The particular details of these transitions for each element depend on the relative position of the Fermi level within the density of states. Crystal structures calculated as a function of *d* occupation number indicate that this mechanism is at work only in the heavy alkali metals (K, Rb, and Cs), while the transitions in Li and Na have a different origin, at least below 35 GPa. Experimental measurements partially confirm this conclusion, but suggest that the intermediate hexagonal-close-packed structure is suppressed at room temperature (Skriver, 1985).

Calculated structural energy differences for the alkaline-earth and the two divalent rare-earth metals as a function of the relative volume V/V_0 and LMTO pressure p are indicated in Fig. 22(b). In the case of the alkaline-earth elements, there is a strong correlation between the calculated *d* occupation number and the calculated crystal structure according to which the heavy alkaline-earth metals should

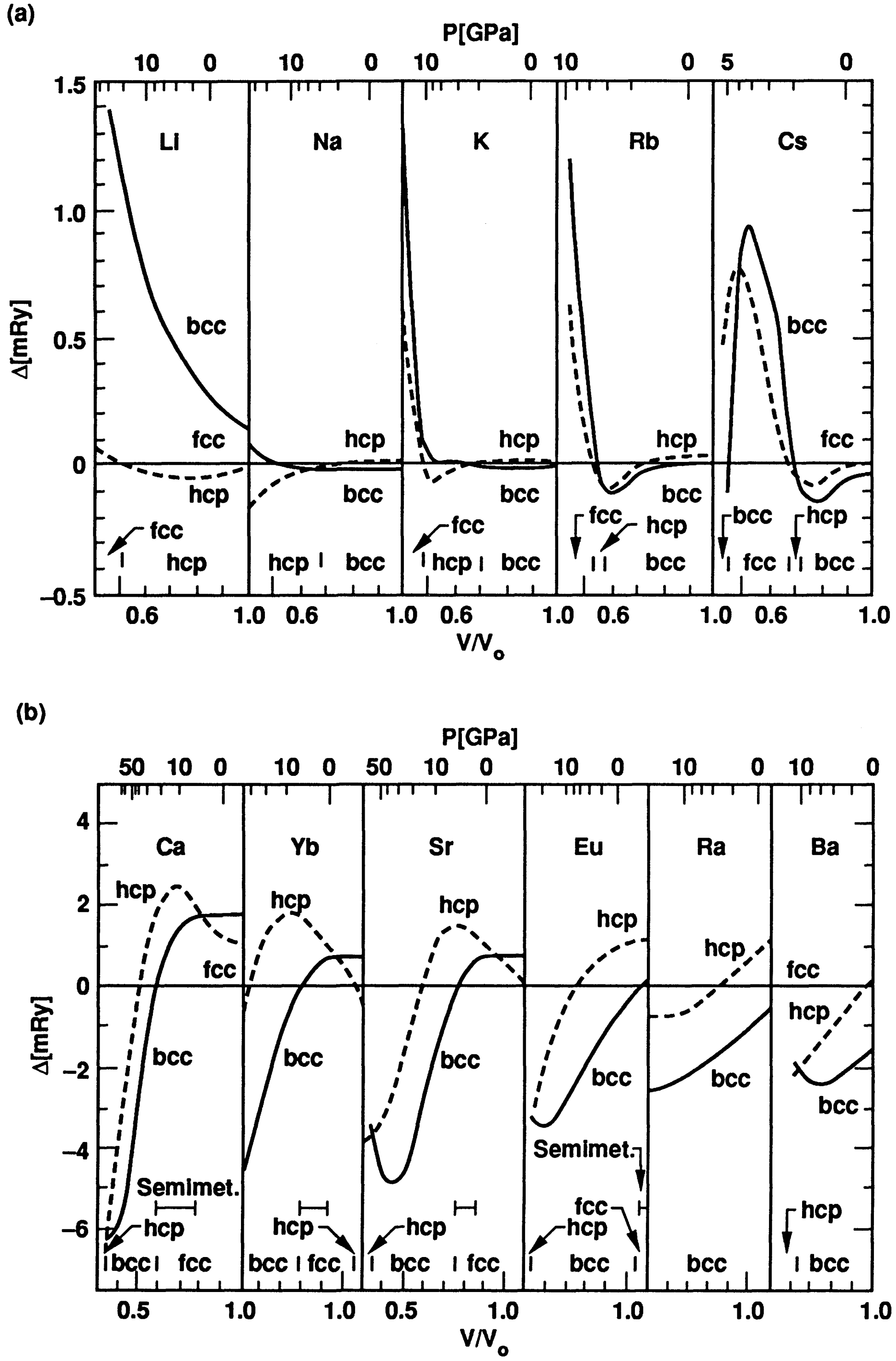

FIG. 22. (a) Structural energy differences (with respect to face-centered cubic) for the five alkali metals as a function of the relative volume V/V_0 and LMTO pressure p. (b) Structural energy differences (with respect to face-centered cubic) for the alkaline earth and the two divalent rare-earth metals as a function of the relative volume V/V_0 and LMTO pressure p. The structures shown are abbreviated in the following manner: face-centered-cubic (fcc), body-centered-cubic (bcc), and hexagonal-close-packed (hcp). (Reproduced with permission from Skriver, 1985.)

be part of the same hexagonal-close-packed→face-centered-cubic→body-centered-cubic→hexagonal-close-packed structural transformation sequence. At zero pressure, each individual metal may be characterized as being at different stages on the continuous *s*-to-*d* transition as determined by their *d* occupation number. The structural phase transitions are then driven by the pressure-induced lowering of the *d* band relative to the *s* band. The correlation is, however, not perfect and the calculated crystal structures transformations occur over a narrow range of *d* occupation numbers. Moreover, band-structure calculations indicate that Ca, Sr, and Yb in the face-centered-cubic phase should undergo a metal-semimetal-metal transition under hydrostatic pressure. The face-centered-cubic→body-centered-cubic part of the above structural sequence has been observed experimentally at room temperature, but the body-centered-cubic→hexagonal-close-packed transition is found in Ba. The experimentally determined transition pressures are also in good correspondence with the predicted values (Skriver, 1985).

3.4.3 Comparison between Different First-Principles Methods. We now compare the predictions obtained by different first-principles total-energy methods regarding the ground-state properties of a single material. The simple metals are the solids for which the most direct comparisons can be made between the different methods of calculations. The calculation of the electron structure of simple metals presents a severe test: the energy differences between various structures are small and, because the density of states is finite at the Fermi level, the calculation of the total energy requires integrals over the Brillouin zone, which necessitates the utilization of very accurate sampling techniques.

A comparison of the results obtained by different first-principles total-energy methods is illustrated in Table 3 (Martin, 1985) for aluminum. The results shown in Table 3 were obtained from various theoretical techniques: the plane-wave pseudopotential method, the Korringa–Kohn–Rostoker method, the augmented-spherical-wave method, the augmented-plane-wave method, and the linear combination of Gaussian-type orbitals method. Inspection of Table 3 clearly indicates that the results obtained by different electronic structure methods are in very good agreement between themselves. Consequently, one can conclude that careful calculations by any first-principles total-energy methods yield essentially the same results.

3.4.4 Applications of the Harris Functional. We now indicate recent applications of the Harris functional to the calculation of the ground-state properties of materials. Recall that the basis of the Harris functional (Sec. 3.3) is the construction of a charge density, $n_f(\mathbf{x})$, from the individual overlapping contributions of the fragments forming the solid or molecule, $\Sigma_k n_k(\mathbf{x})$. The total energy is then evaluated, using Eq. (166), without requiring convergence of the potential

Table 3. Calculated bulk ground-state properties of aluminum using different first-principles total-energy methods. The quantities shown are the equilibrium lattice constant a_0 (in atomic units: 1 a.u.=0.529 18 Å), the bulk modulus B_0, and the cohesive energy E_{coh}. The experimental value is also given. The theoretical techniques are the plane-wave pseudopotential, Korringa-Kohn-Rostoker (KKR), augmented-spherical-wave (ASW), augmented-plane-wave (APW), and the linear combination of Gaussian-type orbitals (LCGTO).

Calculation	a_0 (a.u.)	B_0 (GPa)	E_{coh} (Ry/atom)
Lam and Cohen (1981) pseudopotentials	7.59	71.5	−0.268
Liberman (1970,1971) KKR	7.65	78.0	−0.245
Janak (1974) KKR	7.59	80.1	−0.285
Williams *et al.* (1979) ASW	7.60	87.0	−0.298
Ross and Johnson (1970) APW	7.79	94.0	−0.300
Boettger and Trickey (1984) LCGTO	7.65	96.8	−0.235
Experiment	7.60	79.4	−0.248

and charge density to self-consistency. Figure 23 indicates results obtained for the equilibrium volume, cohesive energy, bulk modulus, and shear modulus for many materials using the Harris functional (van Schilfgaarde, 1991). The calculations were performed using the all-electron full-potential LMTO method (van Schilfgaarde, 1991). Comparisons with self-consistent total-energy LDA calculations and with experiments are also indicated. These comparisons reveal a remarkable agreement between the two methods and available exper-

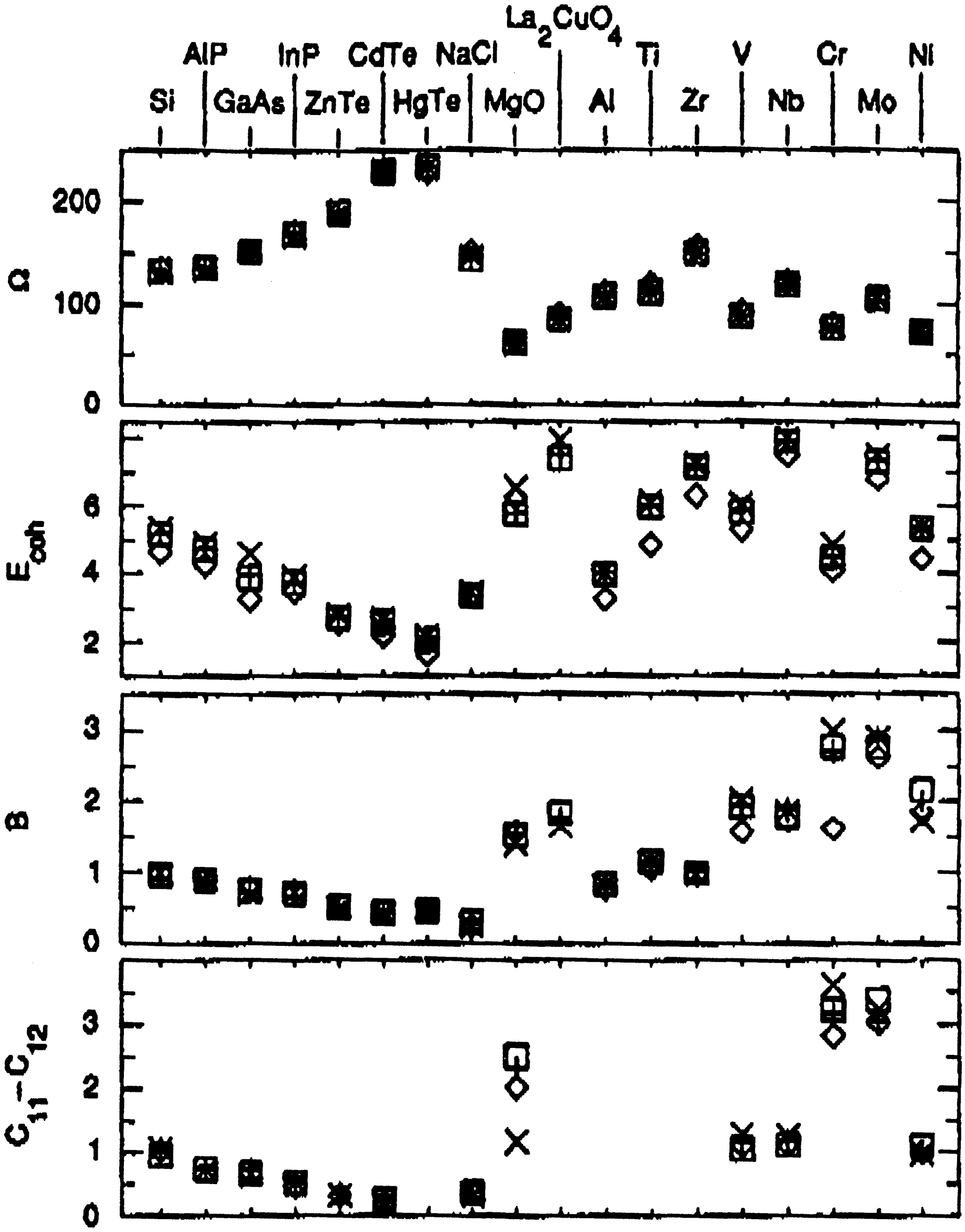

FIG. 23. Atomic volume (a.u.^{3}), cohesive energy (eV), bulk modulus (10^{11} erg cm^{-3}) and shear modulus $c_{11}-c_{12}$ (10^{11} erg cm^{-3}) for selected materials. Experiments (◇) are compared with self-consistent results (□) and with results obtained with the Harris functional using two different forms for the input densities [superposition of free-atomic densities (×) and renormalized densities (+)]. The theoretical results are obtained using the all-electron full-potential LMTO method. (Reproduced with permission from van Schilfgaarde, 1991.)

iments and suggest the substantial predictive power of the method. Also remarkable is the success of the Harris functional in describing the ground-state properties for a wide variety of materials exhibiting very different chemical bonding: covalent materials, ionic solids, oxides, simple and transition metals, etc. Since the utilization of the Harris functional consists of a *single* evaluation of the total energy without the need to perform the calculation self-consistently, the numerical work involved is substantially reduced, and, consequently, the method readily allows the treatment of large-scale materials systems. This spectacular success suggests that the utilization of non-self-consistent total-energy functionals, such as the one proposed by Harris, will have a large impact on electronic-structure calculations of complex materials.

4. *AB INITIO* MOLECULAR DYNAMICS

4.1 Introduction

In the previous sections, we have been concerned with electron-structure calculations of a *static* arrangement of atoms in a solid. In traditional total-energy minimization methods, two sets of iterations must be sequentially performed:

1. for a given set of ionic coordinates, the electronic charge density $n(\mathbf{x})$ must be iteratively recalculated until self-consistency is achieved, and
2. subsequently this procedure must be repeated by varying the ionic positions until the total energy $E_{\text{total}}[\{\psi_i(\mathbf{x})\},\{\mathbf{R}_I\}]$ given by Eq. (160) is minimized.

The present section describes recent theoretical developments aiming at simultaneously performing these two tasks within an *ab initio* framework. We now briefly indicate the essential features of static *ab initio* descriptions and empirical molecular-dynamics simulations.

Ab initio self-consistent total-energy methods involve the minimization of a total-energy functional $E_{\text{total}}[\{\psi_i(\mathbf{x})\},\{\mathbf{R}_I\}]$ with respect to electronic $\{\psi_i(\mathbf{x})\}$ and ionic $\{\mathbf{R}_I\}$ degrees of freedom. This is achieved by obtaining self-consistent solutions for the electron wave functions for a given set of ionic coordinates. The process is repeated for various ionic configurations until the total-energy functional is minimized and a minimum-energy solution has been found. For each set of ionic configurations $\{\mathbf{R}_I\}$, the electronic states of a system are obtained by solving a set of self-consistent field equations (Kohn–Sham equations) repeatedly until convergence is reached. Since perfect electronic self-consistency is achieved for each set of ionic configurations, the system follows exactly the Born–Oppenheimer surface for the ions, Φ, defined as

$$\Phi[\{\mathbf{R}_I\}] \equiv \min_{\{\psi_i(\mathbf{x})\}} E_{\text{total}}[\{\psi_i(\mathbf{x})\},\{\mathbf{R}_I\}]. \qquad (167)$$

The variations of the ionic coordinates $\{\mathbf{R}_I\}$ define the Born–Oppenheimer surface $\Phi[\{\mathbf{R}_I\}]$.

Though impressive advances in materials modeling have been achieved over the last 15 years with electron-structure methods, their application to interesting and technologically important systems that exhibit either very complex structures or large atomic motions remains limited. Moreover, it is not possible to apply these methods to finite-temperature systems where the electronic properties must be averaged over many possible configurations of the constituent ions in the system.

Molecular-dynamics (MD) simulations, on the other hand, are based on the numerical calculation of the trajectories of the atomic motion and are a very powerful tool for the study of complex condensed-matter systems. Molecular-dynamics simulations are usually based on empirical potentials that are chosen to mimic the interatomic potentials in a real material. The simplicity that follows from the use of parametrized potentials is then exploited to treat large numbers of atoms (exceeding one million) which can be in disordered, complex configurations. By varying the temperature (i.e., the mean velocity at which the constituent atoms move) during the MD simulation, technologically interesting processes can be studied: melting, crystal growth and epitaxy, ion implantation, laser annealing, defect motion, mechanical materials processing, etc. Disordered systems can be studied via time averages taken over one or many simulations starting from (usually random) initial conditions.

4.2 Lagrangian Formulation

A useful formulation for combining these two approaches into a single "*ab initio* molecular-dynamics" method has been proposed by Car and Parrinello (Car and Parrinello, 1985). Consider the problem of minimizing an energy functional $E_{\text{total}}[\{\psi_i(\mathbf{x})\},\{\mathbf{R}_I\},\{\alpha_\nu\}]$, such as Eq. (160), where $\{\psi_i(\mathbf{x})\}$ represent the electronic degrees of freedom, $\{\mathbf{R}_I\}$ represent the ionic coordinates, and $\{\alpha_\nu\}$ represent all the possible constraints imposed on the system (volume, strain, etc.). Car and Parrinello realized that the minimization of the energy functional $E_{\text{total}}[\{\psi_i(\mathbf{x})\},\{\mathbf{R}_I\},\{\alpha_\nu\}]$ can be performed using simulated annealing techniques if the concept of force is used to evolve not only the ionic degrees of freedom $\{\mathbf{R}_I\}$, as is traditionally done in MD, but also the electronic degrees of freedom $\{\psi_i(\mathbf{x})\}$ and the external constraints $\{\alpha_\nu\}$. These forces are derived from a Lagrangian for the coupled electron-ion system (Car and Parrinello, 1985, 1988a),

$$L\equiv \sum_{i(\text{occ.})} \mu_i \int d^3x \dot{\psi}_i^*(\mathbf{x},t)\dot{\psi}_i(\mathbf{x},t)+\sum_I \frac{1}{2} M_I\dot{\mathbf{R}}_I^2+\sum_\nu \frac{1}{2}\mu_\nu\dot{\alpha}_\nu^2$$
$$-E_{\text{total}}[\{\psi_i(\mathbf{x})\},\{\mathbf{R}_I\},\{\alpha_\nu\}]+\sum_{ij}\Lambda_{ij}\left(\int d^3x\ \psi_i^*(\mathbf{x},t)\psi_j(\mathbf{x},t)-\delta_{ij}\right), \tag{168}$$

where the wave functions $\psi_i(\mathbf{x})$ are subject to the holonomic constraints of orthogonality,

$$\int d^3x\ \psi_i^*(\mathbf{x},t)\psi_j(\mathbf{x},t)=\delta_{ij}. \tag{169}$$

The coefficients Λ_{ij} are the Lagrange multipliers for the orthogonality constraints and their exact form depends on the particular scheme used to orthogonalize the electronic wave functions.

The Lagrangian of Eq. (168) contains the total energy functional $E_{\text{total}}[\{\psi_i(\mathbf{x})\},\{\mathbf{R}_I\},\{\alpha_\nu\}]$ representing the potential energy, the ionic kinetic energy term, and a fictitious electronic kinetic energy term that generates a dynamics for the expansion coefficients of the electron wave functions $\{\psi_i(\mathbf{x})\}$. Similarly, a fictitious kinetic energy term is added to generate a dynamics for the constraints $\{\alpha_\nu\}$.

Application of the Euler conditions to the Lagrangian of Eq. (168) generates the equations of motion (dynamics) for the different degrees of freedom,

$$\mu_i\ddot{\psi}_i(\mathbf{x},t)=\frac{-\delta E_{\text{total}}}{\delta\psi_i^*(\mathbf{x},t)}+\sum_j\Lambda_{ij}\psi_j(\mathbf{x},t)$$
$$=-H\psi_i(\mathbf{x},t)+\sum_j\Lambda_{ij}\psi_j(\mathbf{x},t), \tag{170}$$

$$M_I\ddot{\mathbf{R}}_I=-\nabla_{\mathbf{R}_I}E_{\text{total}}, \tag{171}$$

$$\mu_\nu\ddot{\alpha}_\nu=\frac{-\partial E_{\text{total}}}{\partial\alpha_\nu}. \tag{172}$$

Equation (171) represents a physical ion dynamics for the ion trajectories. However, Eq. (170) describes a fictitious electron dynamics because the time t is not the real time but rather a construct to parametrize movement through the hyperspace defined by the electronic degrees of freedom [or volume, strain, etc. in the case of Eq. (172)]. The "equation of motion" governing the evolution of the electronic degree of freedom, Eq. (170), must therefore be interpreted in a broader sense than that governing the physical ions' dynamics, Eq. (171). For example the fictitious "mass" μ_i appearing in Eq. (170) is often used as an adjustable parameter determining the size of the time step used in the numerical integration of the equation of motion for the electronic degrees of freedom. The fictitious inertia μ_i has no real meaning, but controls the speed and character of the motion through the hyperspace defined by the electronic (or volume, strain, etc.) degrees of freedom (Remler and Madden, 1990).

The Car–Parrinello approach allows a reformulation of the electron-structure problem in which MD algorithms are used to evolve through time a coupled system of ions and electrons. We refer to this formulation as *ab initio* molecular dynamics since the forces acting on the ions are derived from an *ab* *initio* quantum-mechanical energy functional within the local-density approximation. The system is initially given a finite temperature (by assigning random velocities to the ions) and the temperature is slowly lowered and, as $T \rightarrow 0$, an equilibrium state is reached and $E_{\text{total}}[\{\psi_i(\mathbf{x})\},\{\mathbf{R}_I\},\{\alpha_\nu\}]$ is a minimum. At equilibrium, $\ddot{\psi}_i(\mathbf{x},t)=0$, and Eq. (170) becomes identical, to within a unitary transformation, to the one-electron Kohn–Sham equations [Eq. (150)] and the eigenvalues of the matrix Λ coincide with the occupied Kohn–Sham energy eigenvalues.

4.3 Nonlinear Local Optimization

The Car–Parrinello procedure described above adapts simulated annealing ideas to derive *ab initio* a dynamical model in which the electronic contributions to the atomic forces are calculated within the framework of local-density theory. Practically, the Car–Parrinello method is implemented by integrating the equations of motion Eqs. (170)–(172) and by periodically quenching the fictitious kinetic energy of the electronic system by bringing the electronic system into its ground state for a fixed configuration of the ion cores. This quenching is necessary since, because of the electron–ion interactions, thermal equilibration between the classical degrees of freedom $\{\mathbf{R}_I\}$ and the electronic wave functions $\{\psi_i(\mathbf{x})\}$ will be achieved after some time by equipartition. As a result, the electronic system heats up and moves away from the Born–Oppenheimer surface. It has been recognized that a more direct form of fictitious dynamics could be proposed that eliminates the need to quench the electronic system periodically. Such an algorithm essentially performs a local optimization of the electronic degrees of freedom (Williams and Soler, 1987) using the following "time" evolution (setting $\hbar=\mu=1$):

$$\dot{\psi}_i(\mathbf{x},t)=\frac{-\delta E_{\text{total}}}{\delta\psi_i^*(\mathbf{x},t)}+\sum_j \Lambda_{ij}\psi_j(\mathbf{x},t)$$

$$=-H\psi_i(\mathbf{x},t)+\sum_j \Lambda_{ij}\psi_j(\mathbf{x},t), \qquad (173)$$

$$M_I\ddot{\mathbf{R}}_I=-\nabla_{\mathbf{R}_I}E_{\text{total}}. \qquad (174)$$

Equation (173) can be derived from a steepest-descent condition and is the imaginary-time Schrödinger equation in which excited-state solutions decay exponentially (Woodward *et al.*, 1989). To verify that the solution of Eq. (173) monotonically approaches an energy minimum, we rewrite Eq. (173) using the orthonormality constraint $\Lambda_{ij}=H_{ji}$,

$$\dot{\psi}_i(\mathbf{x},t)=-(1-\hat{P}_N)H\psi_i(\mathbf{x},t), \qquad (175)$$

where the projection operator $\hat{P}_N$ projects on the subspace of the N occupied states $\psi_i(\mathbf{x},t)$,

$$\hat{P}_N\equiv\sum_{i=1}^{N}|\psi_i(\mathbf{x},t)\rangle\langle\psi_i(\mathbf{x},t)|. \qquad (176)$$

The time derivative of the energy eigenvalue ϵ_i associated with the state $\psi_i(\mathbf{x},t)$ is given by

$$\frac{d\epsilon_i}{dt}=-2\langle\psi_i|H(1-\hat{P}_N)H|\psi_i\rangle. \qquad (177)$$

Since $1-\hat{P}_N\geqslant 0$, it follows that $d\epsilon_i/dt\leqslant 0$ and the energy eigenvalues monotonically decrease until a minimum is attained.

The dynamical minimization method just described allows a simultaneous relaxation of the electronic and ionic degrees of freedom. At each set of fixed ionic positions, the electronic degrees of freedom $\{\psi_i(\mathbf{x})\}$ are varied according to the time evolution of Eq. (173) until a minimum on the energy surface is reached. Unlike the Car–Parrinello approach, the steepest-descent method described here ensures that perfect electronic self-consistency is achieved for each set of nuclear configurations. Consequently, the ionic trajectories follow exactly the Born–Oppenheimer surface for the ions, $\Phi[\{\mathbf{R}_I\}]$, defined by Eq. (167). However, the steepest-descent method is a *local* optimization method whereas the Car–Parrinello method is a *global* optimization method. Since the potential energy surface $\Phi[\{\mathbf{R}_I\}]$ exhibits numerous minima, a local optimization method is quite impractical, and simulated annealing techniques such as the Car–Parrinello method described above are more appropriate to identify a global energy minimum. On the other hand, local optimization methods are sufficient in many problems involving the local optimization of complex atomic structures.

Let us close this section by simply mentioning that efforts have been made to incorporate molecular-dynamics ideas within the frameworks of empirical non-self-consistent electron-structure calculation schemes. This approach, while not *ab initio* in nature, nevertheless embodies a quantum-mechanical description of the electron-ion interactions and goes beyond traditional empirical molecular-dynamics simulations using analytic two- and three-body model potentials. An example of this approach is the utilization of empirical tight-binding Hamiltonians to derive a dynamics for the electronic and ionic degrees of freedom (Broughton and Khan, 1989).

4.4 Applications

In this section, we indicate recent applications of the *ab initio* molecular-dynamics method to the study of the atomic and electronic structure of complex systems. A comprehensive description of the applications of molecular-dynamics simulations will be presented in another article and our objective here is only to demonstrate the great impact that the new approaches described above have already had on the field of electron-structure theory. These theoretical methods are in early stages of development and the applications continue to grow at an accelerated rate. Nevertheless, the results obtained so far with the method clearly indicate their predictive power and provide a sense of their usefulness.

Ab initio molecular-dynamics methods have now been applied to several systems in the amorphous or liquid state. In these calculations, the interatomic potential is constructed directly from the electronic ground state, which is treated accurately within the framework of density-functional techniques. The dynamical trajectories obtained from the equations of motion Eqs. (170)–(172) are used to generate a liquid state or an amorphous network resulting from the quenching of a liquid phase. The electron structure is simultaneously generated from the solution of the equations of motion for the electronic degrees of freedom. Examples of such applications include the study of the atomic geometry and electron structure of amorphous (Galli *et al.*, 1989a) and liquid carbon (Galli *et al.*, 1989b; Galli *et al.*, 1990), amorphous (Car and Parrinello, 1988b) and liquid silicon (Štich *et al.*, 1989), liquid sodium (Qian *et al.*, 1990), liquid arsenic (Li *et al.*, 1990), liquid GaAs (Zhang *et al.*, 1990), etc. Moreover, *ab initio* molecular-dynamics and local nonlinear optimization techniques have recently been utilized to determine minimum-energy atomic structures of complex condensed-matter systems such as semiconductor surfaces (Ancilotto *et al.*, 1990; Ihara *et al.*, 1990), clusters (Hohl *et al.*, 1990), vacancies in metals (Gillan, 1989), SiO_2 (Allan and Teter, 1987), etc. We emphasize that these calculations are the first to be performed on low-symmetry complex systems with a level of accuracy such as that afforded by density-functional methods. Previous model-potential-based molecular-dynamics studies of these systems relied on empirically derived force laws and did not offer a quantum-mechanical treatment.

5. SUMMARY

We have presented an overview of modern theoretical methods used for the calculation of the electron structure of condensed-matter systems. We have somewhat arbitrarily classified these electron-structure methods into two categories depending on how the electron states are expanded. One class of method is based on an expansion of the wave functions on a complete set of Bloch-type basis functions. The expansion coefficients are determined by solving the appropriate Schrödinger equation. We refer to these methods as *fixed-basis-set methods*. The other class of methods is based on an expansion of the wave functions on a complete set of functions that are solutions to the Schrödinger equation within a unit cell. The resulting expansion coefficients are determined by imposing appropriate boundary conditions at the unit-cell boundaries or some other more convenient surfaces. Methods based on this construction are referred to as *scattering methods*. Both classes of methods allow the calculation of the wave functions and energy eigenvalue spectrum of electron states in solids. The total energy of the system can be calculated from a knowledge of these electron wave functions and energy eigenvalues by combining the interaction energies between the ions, between the electrons, and between the electrons and the ions. We have given an overview of these total-energy methods based on local-density-functional theory and have indicated recent developments based on the utilization of non-self-consistent total-energy functionals. Examples of the predictive power of modern total-energy methods were given for a wide variety of materials including metals, semiconductors and insulators. A comparison between different implementations of local-density-functional theory was presented. We also discussed recent theoretical developments that combine first-principles electron-structure formalisms with concepts used in model-potential-based molecular-dynamics simulations. In these techniques the forces acting on the (classical) ions are calculated from first principles with an accuracy comparable to that of static calculations. These *ab initio* developments afford first-principles statistical treatments (within the confines of the local-density-functional formalism) of disordered materials (amorphous phases, liquids) at finite temperature. In this sense, they provide a natural intellectual link between *ab initio* electron-structure calculations and materials science.

GLOSSARY

Augmented Plane Wave: Piecewise-defined function consisting of the solution to Schrödinger's equation for an isolated atom within a sphere of given radius r_s and a plane wave outside this region.

Bandwidth: Energy spread between the minimum and maximum energy in a band.

Band Structure: Set of energy eigenvalue functions $\{\epsilon_n(\mathbf{k})\}$ that are continuous in the wave-vector parameter **k** and periodic with the periodicity of the reciprocal space. The band index n labels the energy eigenvalues $\epsilon_n(\mathbf{k})$ within the first Brillouin zone in reciprocal space.

Bloch Function: Eigenfunction of an electron in a crystal. A Bloch function $\psi(\mathbf{k},\mathbf{x})$ can be written as a product of a phase factor $\exp(i\mathbf{k}\cdot\mathbf{x})$ and a cell-periodic function $u(\mathbf{k},\mathbf{x}+\mathbf{R}_j)=u(\mathbf{k},\mathbf{x})$ as $\psi(\mathbf{k},\mathbf{x})=\exp(i\mathbf{k}\cdot\mathbf{x})u(\mathbf{k},\mathbf{x})$.

Bravais Lattice: Three-dimensional infinite array of discrete points with position vectors of the form $\mathbf{R}_j=n_1\mathbf{a}_1+n_2\mathbf{a}_2+n_3\mathbf{a}_3$, where the numbers n_i are integers and the vectors $\mathbf{a}_i$ are any three vectors not all in the same plane. The vectors $\mathbf{a}_i$ are called *primitive lattice vectors*.

Density-Functional Theory: Theory formulated from the realization that the total ground-state energy of a system of interacting electrons moving in an external potential is a unique functional of the electron density.

Density of States: Number of allowed electron states in a given energy range per unit volume.

Energy Band: Set of energy levels $\epsilon_n(\mathbf{k})$ describing the energy eigenvalue spectrum of a Bloch function for a given band index n.

First Brillouin Zone: Wigner–Seitz primitive cell of the reciprocal lattice.

Molecular Dynamics: Numerical simulations consisting of the calculation of dynamical trajectories describing atomic motion.

Muffin-Tin Orbital: Piecewise-defined function consisting of the solution to Schrödinger's equation for an isolated atom within a sphere of given radius r_s and a zero-kinetic-energy tail function outside this region.

Muffin-Tin Potential: Potential that represents an isolated ion within a sphere of given radius r_s about each lattice point and constant elsewhere.

Orthogonalized Plane Wave: Plane-wave function made orthogonal to the atomic core functions.

Primitive Cells: Set of nonoverlapping space-filling (when translated through all the translation vectors of a Bravais lattice) cells each containing exactly a single lattice point.

Pseudopotential: Effective potential that produces a nodeless pseudofunction without binding a core state.

Reciprocal Lattice: The reciprocal lattice of a Bravais lattice is the set of all wave vectors $\mathbf{G}_i$ satisfying the relation $\exp(i\mathbf{G}_i\cdot\mathbf{R}_j)=1$, where $\mathbf{R}_j$ defines the Bravais lattice, and generating plane waves with the periodicity of the Bravais lattice.

Space Group: Set of operations (translations, rotations, and reflections) that take a Bravais lattice into itself.

Tight-Binding Method: Electronic-structure method in which crystal Bloch functions are expanded in a set of atomic orbitals of the atoms within the primitive cell.

van Hove Singularities: Singularities present in the density of states profile of a crystalline solid which occur for energies ϵ on the constant-energy surface $\epsilon=\epsilon_n(\mathbf{k})$ for which the velocity vanishes, i.e., $\nabla_{\mathbf{k}}\epsilon_n(\mathbf{k})=\mathbf{0}$. At these singularities, the density of states $g(\epsilon)$ is continuous but the energy derivative $dg(\epsilon)/d\epsilon$ has an infinite discontinuity.

Wigner–Seitz Primitive Cell: Primitive cell with the full symmetry of the Bravais lattice.

List of Works Cited

Allan, D. C., Teter, M. (1987), *Phys. Rev. Lett.* **59**, 1136–1139.

Ancilotto, F., Andreoni, W., Selloni, A., Car, R., Parrinello, M. (1990), *Phys. Rev. Lett.* **65**, 3148–3151.

Andersen, O. K. (1975), *Phys. Rev. B* **12**, 3060–3083.

Andersen, O. K., Jepsen, O., Sob, M. (1987), in: M. Yussouff (Ed.), *Electronic Band Structure and its Applications*, Heidelberg: Springer-Verlag.

Arfken, G. (1985), *Mathematical Methods for Physicists*, 3rd ed., Orlando: Academic, pp. 516–523.

Ashcroft, N. W., Mermin, N. D. (1976), *Solid State Physics*, Philadelphia: Holt, Rinehart and Winston.

Bassani, R., Pastori Parravicini, G. (1975), *Electronic States and Optical Transitions in Solids*, Oxford: Pergamon.

Biswas, R., Martin, R. M., Needs, R. J., Nielson, O. H. (1984), *Phys. Rev. B* **30**, 3210–3213.

Bloch, F. (1928), *Z. Phys.* **52**, 555–600.

Boettger, J. C., Trickey, S. B. (1984), *Phys. Rev. B* **29**, 6434–6442.

Born, M., Oppenheimer, J. R. (1927), *Ann. Phys.* **84**, 457.

Bouckaert, L. P., Smoluchowsky, R., Wigner, E. P. (1936), *Phys. Rev.* **50**, 58–67.

Broughton, J., Khan, F. (1989), *Phys. Rev. B* **40**, 12 098–12 104.

Callaway, J. (1974), *Quantum Theory of the Solid State*, New York: Academic.

Car, R., Parrinello, M. (1985), *Phys. Rev. Lett.* **55**, 2471–2474.

Car, R., Parrinello, M. (1988a), in: A. Polian, P. Loybeyre, N. Boccara (Eds.), *NATO Advanced Study Institute Series B*, Vol. 186, New York: Plenum, 455–476.

Car, R., Parrinello, M. (1988b), *Phys. Rev. Lett.* **60**, 204–207.

Chelikowsky, J. R., Louie, S. G. (1984), *Phys. Rev. B* **29**, 3470–3481.

Cohen, M. L. (1986), *Science* **234**, 549–553.

Cohen, M. L., Chelikowsky, J. R. (1982), in: W. Paul (Ed.), *Band Theory and Transport Properties*, Vol. 1, Amsterdam: North Holland, Chap. 4B.

Devreese, J., Van Camp, P. (Eds.) (1985), *Electronic Structure, Dynamics, and Quantum Structural Properties of Condensed Matter*, New York: Plenum.

Fahy, S., Louie, S. G. (1984), *Phys. Rev. B* **36**, 3373–3385.

Galli, G., Martin, R. M., Car, R., Parrinello, M. (1989a), *Phys. Rev. Lett.* **62**, 555–558.

Galli, G., Martin, R. M., Car, R., Parrinello, M. (1989b), *Phys. Rev. Lett.* **63**, 988–991.

Galli, G., Martin, R. M., Car, R., Parrinello, M. (1990), *Phys. Rev. B* **42**, 7470–7482,

Gillan, M. J. (1989), *J. Phys.: Condens. Matter* **1**, 689–711.

Hamermesh, M. (1962), *Group Theory*, Reading, MA: Addison-Wesley.

Harris, J. (1985), *Phys. Rev. B* **31**, 1770–1779.

Harrison, W. (1969), *Phys. Rev.* **181**, 1036–1053.

Harrison, W. (1980), *Solid State Theory*, New York: Dover, pp. 97–105.

Heine, V. (1960), *Group Theory in Quantum Mechanics*, New York: Pergamon.

Heine, V. (1970), in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, Vol. 24, New York: Academic, pp. 1–36.

Heine, V., Weaire, D. (1970), in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, Vol. 24, New York: Academic, pp. 249–463.

Herman, R. (1954), *Phys. Rev.* **93**, 1214–1225.

Herring, C. (1940), *Phys. Rev.* **57**, 1169–1177.

Hohl, D., Jones, R. O., Car, R., Parrinello, M. (1990), *Chem. Phys. Lett.* **139**, 540–545.

Hohenberg, P., Kohn, W. (1964), *Phys. Rev. B* **136**, 864–871.

Ihara, S., Ho, S. L., Uda, T., Hirao, M. (1990), *Phys. Rev. Lett.* **65**, 1909–1912.

Janak, J. F. (1974), *Phys. Rev. B* **9**, 3985–3988.

Jones, R. O., Gunnarson, O. (1989), *Rev. Mod. Phys.* **61**, 689–746.

Kittel, C. (1957), *Solid State Physics*, 2nd ed., New York: Wiley.

Kittel, C. (1963), *Quantum Theory of Solids*, New York: Wiley.

Kittel, C. (1986), *Solid State Physics*, 6th ed., New York: Wiley.

Kohn, W., Rostoker, N. (1954), *Phys. Rev.* **94**, 1111–1120.

Kohn, W., Sham, L. J. (1965), *Phys. Rev. A* **140**, 1133–1138.

Korringa, J. (1947), *Physica* **13**, 392–400.

Lam, P. K., Cohen, M. L. (1981), *Phys. Rev. B* **24**, 4224–4229.

Lang, N. D. (1973), in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, Vol. 28, New York: Academic, pp. 225–300.

Ley, L., Pollak, R. A., McFeely, F. R., Kowalcyzk, S. P., Shirley, D. A. (1974), *Phys. Rev. B*, **9**, 600–621.

Li, X.-P., Allen, P. B., Car, R., Parrinello, M. (1990), *Phys. Rev. B* **41**, 3260–3263.

Liberman, D. A. (1970), *Phys. Rev. B* **2**, 244–248.

Liberman, D. A. (1971), *Phys. Rev. B* **3**, 2081–2082.

Mailhiot, C., McMahan, A. K. (1991), *Phys. Rev. B* **44**, 11578–11591.

Martin, R. M. (1985), in: DeVreese, J., Van Camp, P. (Eds.), *Electronic Structure, Dynamics, and Quantum Structural Properties of Condensed Matter*, New York: Plenum, p. 175.

McMahan, A. K. (1986), *Physica* **139 & 140B**, 31–41.

Methfessel, M., Rodriguez, C. O., Andersen, O. K. (1989), *Phys. Rev. B* **40**, 2009–2012.

Moruzzi, V. L., Janak, J. F., Williams, A. R. (1978), *Calculated Electronic Properties of Metals*, New York: Pergamon.

Nesbet, R. K. (1984), *Phys. Rev. B* **30**, 4230–4234.

Nesbet, R. K. (1986), *Phys. Rev. B* **33**, 8027–8034.

Papaconstantopoulos, D. A. (1986), *Handbook of the Band Structure of Elemental Solids*, New York: Plenum.

Phillips, J. C., Kleinman, L. (1959), *Phys. Rev.* **116**, 287–294.

Pickett, W. E. (1989), *Comput. Phys. Rep.* **9**, 115–198.

Qian, G.-X., Weinert, M., Fernando, G. W., Davenport, J. W. (1990), *Phys. Rev. Lett.* **64**, 1146–1149.

Remler, D. K., Madden, P. A. (1990), *Mol. Phys.* **70**, 921–966.

Ross, M., Johnson, K. W. (1970), *Phys. Rev. B* **2**, 4709–4714.

Sakurai, J. J. (1985), *Modern Quantum Mechanics*, Menlo Park, CA: Benjamin/Cummings.

Skriver, H. L. (1984), *The LMTO Method: Muffin-Tin Orbitals and Electronic Structure*, Berlin: Springer-Verlag.

Skriver, H. L. (1985), *Phys. Rev. B* **31**, 1909–1923.

Slater, J. C. (1937), *Phys. Rev.* **51**, 846–851.

Sommerfeld, A., Bethe, H. (1933), *Handbuch der Physik*, Vol. 24, Berlin: Springer.

Srivastava, G. P., Weaire, D. (1987), *Advances in Physics*, **36**, 463–517.

Štich, I., Car, R., Parrinello, M. (1989), *Phys. Rev. Lett.* **63**, 2240–2243.

Tinkham, M. (1964), *Group Theory and Quantum Mechanics*, New York: McGraw-Hill.

van Hove, L. (1953), *Phys. Rev.* **89**, 1189–1193.

van Schilfgaarde, M. (1991), unpublished work.

Wigner, E. P. (1959), *Group Theory and Its Applications to the Quantum Mechanics of Atomic Spectra*, New York: Academic.

Wigner, E. P., Seitz, F. (1933), *Phys. Rev.* **43**, 804–810.

Wigner, E. P., Seitz, F. (1934), *Phys. Rev.* **46**, 509–524.

Williams, A. R., Kübler, J., Gelatt, Jr., C. D. (1979), *Phys. Rev. B* **19**, 6094–6118.

Williams, A. R., Soler, J. (1987), *Bull. Am. Phys. Soc.* **32**, 562.

Woodward, C., Min, B. I., Benedek, R., Garner, J. (1989), *Phys. Rev. B* **39**, 4853–4861.

Yang, Y. W., Coppens, P. (1974), *Solid State Commun.* **15**, 1555–1559.

Yin, M. T., Cohen, M. L. (1982a), *Phys. Rev. B* **25**, 7403–7412.

Yin, M. T., Cohen, M. L. (1982b), *Phys. Rev. B* **26**, 5668–5687.

Yin, M. T., Cohen, M. L. (1984), *Phys. Rev. B* **29**, 6996–6998.

Zhang, Q. M., Chiarotti, G., Selloni, A., Car, R., Parrinello, M. (1990), *Phys. Rev. B* **42**, 5071–5081.

Zhang, S. B., Cohen, M. L. (1987), *Phys. Rev. B* **35**, 7604–7610.

Ziman, J. M. (1984), *Principles of the Theory of Solids*, London: Cambridge University Press.

ELECTRON TUBES

See TUBES, ELECTRON

ELECTRONIC AMPLIFIERS

See AMPLIFIERS, ELECTRONIC

ELECTRONIC MUSIC

See MUSIC, ELECTRONIC

ELECTRONIC CIRCUITS

EDWIN C. JONES, JR., *Electrical Engineering and Computer Engineering Department, Iowa State University, Ames, Iowa, U.S.A.*

Introduction 1
1. Basic Concepts 2
2. Types of Electronic Devices 2
3. Types of Electronic Circuits 3
4. Operating Frequencies of Electronic Devices 4
5. Power Supplies 4
6. Construction of Electronic Circuits 6
6.1 Printed Circuits 6
6.2 Wiring 7
6.2.1 Wires and Wire Gauges 7
6.2.1.1 American Wire Gauge (AWG) 7
6.2.1.2 Metric Wire Gauge 7
6.2.1.3 British Standard Wire Gauge 7
6.2.2 Cables 7
6.2.3 Coaxial Cables 7
6.2.4 Strip and Sheet Cables 7
6.3 Connectors 8
6.4 Contacts 8
6.5 Insulation 8
7. Vacuum Tubes and Vacuum-Tube Circuits 8
7.1 Historical Note 8
7.2 Tube Elements 9
7.3 Vacuum-Tube Circuit Analysis 9
8. Active Filters 11
8.1 Electrical Filters 11
8.2 Filter Design Process 11
8.2.1 Specifications 11
8.2.2 Approximation 11
8.2.3 Circuit 11
8.2.4 Element Determination 11
8.2.5 Network Conversion 11
8.3 Filter Types and Approximation Functions 12
8.3.1 Types of Filter 12
8.3.2 Modern Network Synthesis 12
8.3.3 Ideal Low-Pass Filter 12
8.3.4 Magnitude Functions 12
8.3.5 Butterworth Approximation 13
8.3.6 Legendre Approximation 13
8.3.7 Chebyshev Approximation 14
8.3.8 Elliptic Filters 15
8.3.9 Bessel Filters 15
8.4 Frequency Transformations and Scaling 15
8.4.1 Scaling 15
8.4.2 Low-Pass to High-Pass Transformation 15
8.4.3 Low-Pass to Band-Pass Transformation 15
8.4.4 Low-Pass to Band-Stop Transformation 15
8.4.5 Element Magnitude Scaling 16
8.4.6 Element Frequency Scaling 16
8.5 Network Realizations 16
8.5.1 Prototype Circuits 16
8.5.2 Cascading Second-Order Circuits 17
8.5.3 Additional Active Filters 17
Glossary 18
Works Cited 18
Further Reading 19

INTRODUCTION

Electronic circuits are electrical circuits containing one or more electronic devices in which the motion of charged particles is used and controlled in order to achieve desired effects. Amplification, switching, and control are but three of these effects. A special type of control is called filtering, which is the process of separating electrical signals on the basis of one or more important characteristics, such as frequency.

The first electronic device was the vacuum diode, but the device that started what is often called the "electronic revolution" was the vacuum triode, invented in 1906. The point-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

contact transistor was invented in 1947, and the bipolar junction transistor was invented in 1948. In 1959, the integrated circuit was invented.

Classification of electronic circuits is complex because of the wide variety of applications. One scheme is by function, including switching, analog or quasilinear circuits, and power-control circuits. Another is by operating frequency, whether audio frequency, radio frequency, microwave frequency, or optical frequency.

Electronic circuits require power supplies, nearly always in the form of a direct (dc) voltage and current. Power may come from an alternating (ac) source or from a battery. If from an ac source, rectifier circuits are required, and for good performance in most applications, filters and regulators are needed.

Most electronic circuits today are assembled on printed circuit boards, with attention paid to mounting, wires, cables, insulation, connectors, and mechanical properties. The entire process is often called packaging. Wire tables give properties of the many wires available.

The vacuum tube is still important in electronic circuits, though most of its functions are now performed by solid-state devices. Vacuum-tube instrumentation is still in use. Vacuum tubes are required in some high-power applications. There are special applications, such as a class of high-fidelity audio equipment. Analysis of circuits with one or more vacuum tubes requires the use of basic principles of circuit analysis along with the graphical "characteristic curves" that describe the tube. It is necessary to perform both operating-point and small-signal analyses.

Active filters are an important application of electronic devices. The filter design process requires a statement of specifications, development of a suitable approximation, and realization of networks that meet the requirements. Important approximations include the Butterworth, Legendre, Chebyshev, Bessel, and elliptic-function approximations. The functions that result from these approximations can be realized in a variety of active networks. The multiple-feedback, infinite-gain (MFB) filter is described in detail, and others are mentioned, with references.

1. BASIC CONCEPTS

We define an electronic circuit as an electrical circuit or system that depends for its operation on one or more electronic devices. In turn, an electronic device is one in which the motion of charged, subatomic particles (such as electrons) in a vacuum, gas, or semiconductor is used to achieve desired effects in physical systems (Fink and Christiansen, 1989, Sec. 1). Electronic devices using liquids or solids are also employed. The particles themselves are frequently called *charge carriers*. The effects of interest include amplification, switching, control, and many others. These functions make possible computers, instrumentation for research, medicine, and industry, and entertainment systems, to mention but a few examples.

The essential physical principles for electronic devices are expressed in the principles of electromagnetism, in crystal behavior, in the principles of conduction in a vacuum or in a gas, and in basic material sciences. The charge carriers are frequently electrons, but they may be positively or negatively charged ions, or holes. A hole is the vacancy in a crystal structure that remains when an electron is removed or leaves the structure. We treat it as a positive charge in analysis, and this leads us to an understanding of the device behavior.

Most electronic devices have three or more electrical connections. (The major exception is the diode, discussed in Sec. 5). In the simplest form, the device uses two terminals for the principal charge-flow terminals or connections, and the third terminal or connection is used to control the charge flow. Because of the possibility of control, the device is enabled to amplify weak signals, to switch on or off some other device such as a light, motor, or element in a computer printer, or to effect smooth control of the control surface of an airplane or of the pen in a pen recorder attached to a computer or a device such as one that records electrocardiograms.

2. TYPES OF ELECTRONIC DEVICES

Historically, the first major electronic device, as the term is used here, was the vacuum diode (Ryder and Fink, 1984). The device that is generally considered to have started the "electronic revolution" was the three-element vacuum tube called a triode (Brittain, 1977; Ryder and Fink, 1984), which is discussed in Sec. 7.2. The vacuum diode led directly to the triode and later to vacuum tubes with many elements. In many parts of the world, vacuum tubes were called *valves.*

In 1947, the invention of a solid-state electronic device was announced. The device was known as a point-contact transistor, and the co-inventors were John Bardeen, and W. H. Brattain. This invention was followed in 1948 by the invention by William Shockley of the junction transistor, which is now possibly the most commonly used discrete electronic device. These three inventors were later honored with the Nobel Prize for Physics in 1956 (Ryder and Fink, 1984). Later, the field-effect transistor was developed, and it too has become quite important in modern electronic circuits. Transistors are discussed in the articles TRANSISTORS, BIPOLAR; TRANSISTORS, FIELD-EFFECT; TRANSISTORS, MICROWAVE; and TRANSISTORS, POWER.

In 1959, two engineers, Jack Kilby and Robert Noyce, working for competing companies, developed independently and received patents for what has become known as the *integrated circuit* (Ryder and Fink, 1984, Chap. 7). The integrated circuit is a device that puts large numbers of transistors onto a single silicon (or other material) wafer. The transistors are interconnected into a system, and as this is written, the number of transistors on a single commercially available "chip" greatly exceeds 1 million (Fink and Christiansen, 1989, Sec. 8). These integrated circuits may be best known for their applications in personal computers, but in fact they have made the wide spread of useful, reliable, and inexpensive electronic systems a reality. Integrated circuits are fully discussed in the articles VLSI—INTEGRATED CIRCUIT TECHNOLOGY and MICROELECTRONICS.

Many other electronic devices exist. To name but a few, there are the Geiger counter, the photocell, the klystron, the cyclotron, the magnetron, the tunnel diode, masers, and lasers.

3. TYPES OF ELECTRONIC CIRCUITS

Electronic circuits are so widely used that a simple classification scheme is difficult, but one possibility is a classification by the type of behavior or function that is a principal description of how a circuit operates. With this in mind, we identify digital or switching circuits, analog or quasilinear circuits, and power-control circuits. Some basic ideas are given here.

Digital or switching circuits have the characteristic that the principal charge flow is either inhibited to essentially zero or permitted to a maximum value that is controlled primarily by external elements, such as resistors and power supplies. One or more of the control terminals of the electronic device is used to inhibit or control the charge flow, usually by the voltage that is applied to the terminal.

In most cases, this process leads to a "two-state" circuit, one that operates in one of two possible conditions. In turn, this idea leads to the possibility of binary arithmetic systems, the basis of computers today. A typical computer has millions of such two-state circuits, which give it the capability that it possesses. Other switching circuits are used to provide "on–off" control of lights, motors, and communications equipment. A complete discussion of computer and switching circuits is given in the article CIRCUITS, DIGITAL.

Analog or quasilinear circuits have the characteristic that the principal charge flow is controlled smoothly between minimum and maximum values. The actual values are controlled by the voltage or other control signal that is applied to the control terminal as well as by external elements such as resistors and power supplies. Examples of such devices are amplifiers that are found in entertainment, information, medical, instrumentation, and communications systems, modulation–demodulation circuits found in communications systems, oscillators and wave-form generators, and active filters, which are discussed in Sec. 8. The other circuits, when transistor or integrated-circuit based, are discussed in the article CIRCUITS, ANALOG. An important building block for analog circuits is the operational amplifier, described in the article, AMPLIFIERS, ELECTRONIC. Analog or linear circuits based on vacuum tubes are discussed in Sec. 7 of this article.

Power-control circuits, or power-electronic circuits, have some of the characteristics of analog circuits and some of the characteristics of digital circuits. The central element is, today, a solid-state, electronic switching device that either connects an electrical source to an electrical load or disconnects the load from the source. Most of these devices conduct current in only one direction, and when conditions are such that the electronic device is switched to a conducting mode, then the amount of current that flows is determined largely by the source voltage and the load that is connected. Newell (in Fink and Christiansen, 1989, Sec. 15) describes four principal types, though other classification schemes are possible. The four types are

1. two-terminal switch, provided by a diode or diodes;
2. three-terminal switch, provided by reverse-blocking triode thyristors, or silicon-controlled rectifiers (SCRs);
3. three-terminal switches with turnoff capability, provided by turnoff thyristors or gate-controlled switches;
4. power transistors, or three-terminal switches in which control of the base signal current controls the power flow.

A power-control device known as a Triac conducts current in both directions. The Triac is, effectively, an inverse or opposing parallel connection of two thyristors. It provides ac power control with a single device, such as in a dimmer switch used in lighting.

Further discussion may be found in the article POWER ELECTRONICS.

4. OPERATING FREQUENCIES OF ELECTRONIC DEVICES

Electronic devices operate over a wide range of frequency, from fractions of a hertz (Hz) to the GHz range. The terminology is not precise, but we often divide the range into the audio-frequency, radio-frequency, microwave, and optical bands (Fink and Christiansen, 1989, Sec. 1).

The audio-frequency range is often considered to extend from dc to perhaps 200 kHz, though both extremes are outside the range of audible acoustic signals. With electronic devices, very sophisticated circuits have been built for use in this range. Examples include many of the circuits in home entertainment systems, in medical instrumentation systems, in communications systems, and in educational technology.

The radio-frequency range is usually considered to be the range from about 200 kHz to about 3 GHz, though special communications systems operate between 1.0 and 200 kHz. The principal range includes the frequencies of television broadcast, radio broadcast, cellular telephones, and many other services. Electronic circuits have always been at the heart of such systems. Modern devices and circuits have led to low-cost, versatile, and reliable systems that allow point-to-point communication anywhere in the world and that permit broadcasting of information and entertainment to large numbers of people.

The microwave band may be considered as the range from 3 to 300 GHz. Signals in this range provide line-of-sight communications and data transfer between two points and are also transmitted through waveguides and coaxial cables. Other applications include radar, some television service including satellite relaying, and microwave spectroscopy.

The optical band includes the infrared (300 GHz to 400 THz), visible (400 to 800 THz), and ultraviolet (700 to as high as 30 000 THz, depending on where the line is drawn) regions of the spectrum. Electronic devices find many applications in this range. The phrase "optical transistor" refers to a transistor in which the control is provided by an optical signal. The field is called optoelectronics (*q.v.*)

Optical frequencies are now commonly employed in high-speed communications systems. Such systems usually are used for data transfer in digital format. See OPTICAL COMMUNICATIONS.

5. POWER SUPPLIES

A common component of virtually all electronic circuits is a power supply. Electronic circuits require power for operation, and nearly always this must be delivered as a direct current or voltage. As most energy is delivered from alternating sources, means to convert ac to dc are required. Such circuits are called power supplies (Fink and Christiansen,

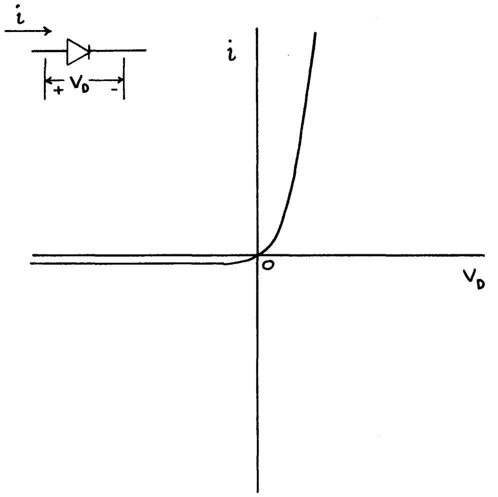

FIG. 1. Typical current-voltage characteristic for a diode.

1989, Sec. 15; Fink and Beaty, 1987, Sec. 22). Circuits that convert dc to ac are called inverters.

The essential element of a power supply is a diode, the principal two-terminal electronic device. A diode is a circuit element that allows a substantial current or charge flow in one direction and a nearly zero current flow in the opposite direction. Figure 1 shows a symbolic representation for a diode and a typical voltage–current relationship. For semiconductor diodes, the current–voltage relationship is given by the Ebers-Moll equation (Ebers and Moll, 1954).

$$i=I_0[\exp(V_D/V_T)-1], \qquad \text{(1a)}$$

where i is the diode current, I_0 is called the reverse saturation current, V_D is the diode voltage, $V_T=kT/q$, k is the Boltzmann constant, q is the electronic charge, and T is the temperature. At room temperature, kT/q is about 26 mV. In practice, it is found empirically that a correction factor is needed in the exponent, so that the Ebers-Moll equation is frequently given as

$$i=I_0[\exp(V_D/mV_T)-1], \qquad \text{(1b)}$$

where typical values of m of 1.5–2.0 are found.

When an ac source is connected in series with a resistor and a diode, as indicated symbolically in Fig. 2(a), we observe the current and voltage wave forms indicated in Fig. 2(b). Although the resulting current has a nonzero average value and thus has a dc component, it is by no means constant. This circuit is called a half-wave rectifier, and major improvements are required to make it practical for most applications.

The first improvement usually considered is to make a full-wave circuit, shown in Fig. 3(a), with the resulting wave forms in Fig. 3(b). A Fourier analysis of this wave form shows that the average value, or dc component, is doubled, and that the component at the supply or source frequency is eliminated. A figure of merit called the ripple factor is defined as the ratio of the effective value of the alternating components to the dc output (Gray, 1954). For a half-wave rectifier this figure is 1.21, while for a full-wave rectifier it is 0.48. The reduced relative magnitude and the higher frequencies involved make these components easier to separate from the desired dc component through filtering. It should be noted, however, that the zero-to-peak excursion in the output voltage is unchanged, a fact that designers of power supplies must consider in filter and rectifier design. A bridge circuit that is also a full-wave rectifier is shown in Fig 3(c). Its wave forms are similar to those of Fig. 3(b). Its main advantage is that the transformer does not require a center tap, but it requires 4 diodes.

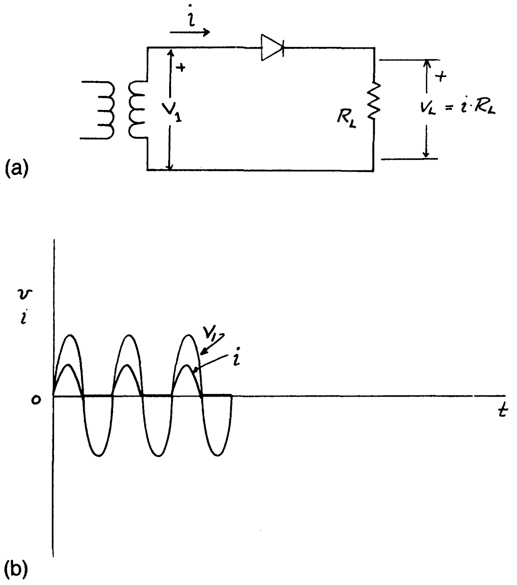

FIG. 2. (a) Half-wave rectifier circuit. (b) Current and voltage wave forms.

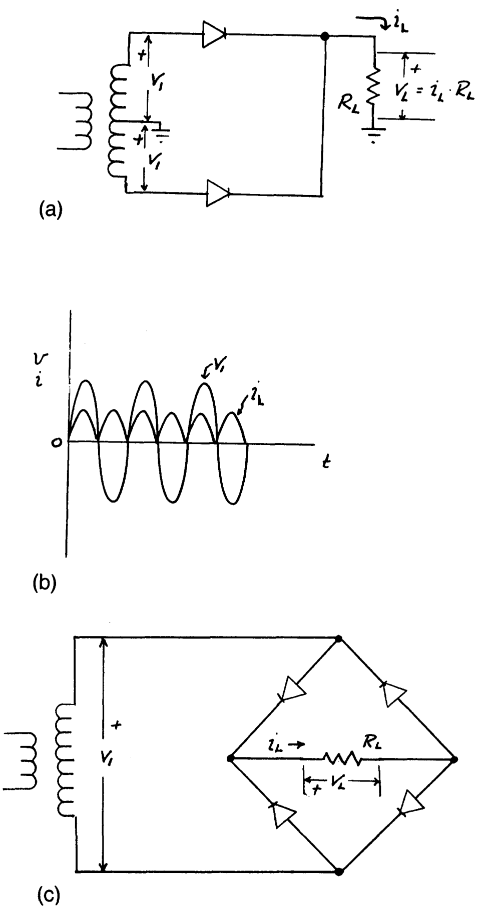

FIG. 3. (a) Full-wave rectifier circuit. (b) Current and voltage wave forms. (c) Bridge rectifier circuit.

Other improvements include the addition of filters, which are capacitors connected in parallel with the load that is using the dc power, or inductors connected in series with the load, or both. Electronic regulators have been developed over the years. These regulators reduce the ac components to very small values and maintain the dc voltage constant over a wide range of operating conditions.

Specifications for a power supply usually include the following items:

1. dc output voltage;
2. dc output current, minimum and maximum values;
3. voltage change (regulation) from no load to full load;
4. an indication of the magnitude of the ac components of the output voltage, frequently called ripple;
5. an indication of how rapidly the power supply can respond to sudden changes in the load; and
6. environmental and other operating conditions.

A wide variety of power supplies is available from many vendors. Careful selection of a power supply is crucial for operation of electronic equipment.

Many electronic devices use batteries (Fink and Christiansen, 1989, Sec. 7) because this greatly increases portability and reduces weight. Batteries are available in a wide range of sizes and shapes, voltage and current ratings, environmental operating conditions, operating and shelf lifetimes. Nickel–cadmium batteries, lead–acid cells, and nickel hydride cells are rechargeable. Most of the other commonly used batteries are not rechargeable.

6. CONSTRUCTION OF ELECTRONIC CIRCUITS

6.1 Printed Circuits

Most electronic circuits today are assembled on printed circuit boards (Fink and Christiansen, 1989, Sec. 7). This is true whether the circuits are constructed of discrete components such as resistors, transistors, and capacitors, or whether they include integrated circuits. The boards are made of laminated plastic. The material must be a good insulator. Holes are drilled in the board for mounting of components, which are then soldered to printed wiring that is plated onto the board.

The boards must be rigid and must provide protection for the circuit elements. In some cases, the components are encapsulated into resins on the board. The thickness and material choices depend on the size and on considerations of strength and stiffness. Spacing of components and of boards in an assembly is influenced by heat dissipation considerations and by voltage breakdown or insulation factors.

Frequently, a printed circuit board has a connector on one edge. The connector consists of the plated wiring brought out in a format that allows insertion of the connector into a receptacle in the overall system. An example of this is the "mother board" of a personal computer, and closely related are the expansion slots that enable the owner to customize the computer for specialized uses.

The use of printed circuit boards allows the design engineer to make changes or improvements that may be needed. It also allows the user to repair equipment by simply replacing boards and to repair the boards away from the equipment. This saves much time and improves system reliability substantially. The overall process is usually called "packaging."

6.2 Wiring

6.2.1 Wires and Wire Gauges. Wires are used to connect electrical components and systems together. The wire is made of a conducting material, usually copper, and covered with insulating materials. An assembly of two or more wires is often called a cable. Three different wire gauges for copper wire are in use. They are (Fink and Beaty, 1987, Sec. 4) the following.

6.2.1.1 American Wire Gauge (AWG). This system refers to the diameter of the wire, and divides the range of diameters from 0.0050 in. (AWG No. 36) to 0.4600 in. (AWG No. 0000, or 4/0) into 39 geometrically spaced intervals, with adjacent wire sizes having a diameter ratio of $(0.460/0.005)^{1/39} = 1.122\,932\,20$. The system has been extended down to AWG No. 50, which has a diameter of 0.0010 in. Table 1 gives an abbreviated wire table, showing sizes frequently found in electronic circuits and systems.

A useful rule of thumb is that, since the diameter ratio is very nearly equal to $2^{1/6}$, then a change of AWG numbers by 6 changes the wire diameter by a factor of 2, the area by a factor of 4, and the resistance by a factor of 4.

Table 1. Abbreviated American Wire Gauge table.

AWG size	Diameter (in.)	Area (in.^{2})	dc resistance (Ω/1000 ft) (@20 °C)
12	0.0808	0.005 13	1.64
14	0.0641	0.003 23	2.61
16	0.0508	0.002 03	4.16
18	0.0403	0.001 28	6.61
20	0.0320	0.000 804	10.5
22	0.0253	0.000 503	16.8
24	0.0201	0.000 317	26.6
26	0.0159	0.000 199	42.4
28	0.0126	0.000 125	67.6
30	0.0100	0.000 078 5	107

6.2.1.2 Metric Wire Gauge. In the systems using SI units, wire size No. 1 has a diameter of 0.1 mm, No. 2 has a diameter of 0.2 mm, etc.

6.2.1.3 British Standard Wire Gauge. This system was developed in the United Kingdom as a modification of the Birmingham Wire Gauge, which in turn was originally developed for use with soft iron wire. Other names include the New British Standard Gage, the English legal standard gage, and the Imperial wire gage. The largest diameter is 0.500 in. and is designated 7/0. The next smaller is 6/0, down to 1/0, with increasing numbers thereafter. The smallest is 0.001 in., which is designated No. 50. Thus there are 57 wire sizes. Unlike the AWG, the diameters are not geometrically spaced.

A very complete wire table may be found in Fink and Beaty (Fink and Beaty, 1987, Sec. 4, Table 4-11).

6.2.2 Cables. Cables are assemblies of 2 or more wires and usually have an outer covering in addition to the insulation on the basic wires. Often the wires are stranded, not solid, especially when there are many wires or the individual wires are large. A wide variety of cables is available to meet virtually any need. Some special-purpose cables have a mixture of wire sizes. One example is a cable that contains 2 or more relatively large wires, such as No. 14 AWG, and a large number of small wires, such as No. 22 AWG. The large wires could be used for power delivery, and the small ones for signals.

Some cables—for example, many of those used in instrumentation and audio systems—are shielded. This means that, in addition to the outer covering, there is a metallic layer enclosing all of the wires. The shield reduces problems that might arise from stray or unknown electromagnetic fields.

6.2.3 Coaxial Cables. Coaxial cables are cables made with the various conductors concentrically assembled. There is a connector in the center, one or more additional conductors separated by insulators, and, usually, a shield around the outside (though inside the final insulator). Coaxial cables are used at higher frequencies and in cases where protection from stray fields is especially important. Coaxial cable data are given by Fink and Christiansen (1989), Sec. 9 and Figs. 9-4 and 9-5.

6.2.4 Strip and Sheet Cables. Strip and sheet cables are flat cables, often used with the edge connectors found in printed circuit boards.

6.3 Connectors

Connectors are devices placed at the ends of cables or on the edges of printed circuit boards to allow electrical contact between appropriate wires or system subassemblies, with adequate mechanical properties to assure proper behavior. A wide variety of connectors for a wide variety of purposes has been developed. There are widely accepted standards that state what connectors should be used for each purpose. A look at the connectors on a personal computer will show a sample of a variety of conductors. This variety virtually eliminates any possibility of a user making incorrect connections.

Connectors must assure low-resistance electrical contacts between wires and must stay fastened regardless of mechanical problems such as vibrations. They must withstand the voltages present in a system and must perform in whatever environmental conditions are encountered.

Power-circuit connectors must handle the relatively large amounts of current involved. Signal-circuit connectors are especially vulnerable to corrosion and other problems that can increase contact resistance and lead to circuit and system failure. Special connectors have been developed for coaxial cables.

6.4 Contacts

Contacts are metallic surfaces within connectors that, when the connectors are fastened together, must touch with very low resistance to allow signals and power to be propagated from one part of a system to another without error. Contacts are susceptible to corrosion, wear, sparking in some cases, dirt, and other hazards. In critical situations, special metals are plated on the contacts to allow good performance over many years. Gold is often used for this purpose, because it does not corrode as other metals do.

6.5 Insulation

The insulation on a wire serves a number of purposes. The most important is to keep metallic wires from touching each other and causing short circuits that lead to system failure. The second is to protect the wiring itself. Insulation must meet standards of breakdown voltage and standards for vibration, bending, and other forms of mechanical motion.

7. VACUUM TUBES AND VACUUM-TUBE CIRCUITS

7.1 Historical Note

In 1884, Houston (Brittain, 1977) discussed the Edison effect, which was the observation of unilateral current flow between an incandescent element and a positively charged electrode in a vacuum. In 1897, Thomson (Ryder and Fink, 1984) identified the charge carrier as a negatively charged particle, now called an electron. This led rapidly to the development of the first vacuum tube, a two-electrode structure. The electrode that emitted the electrons was named the cathode, while the electrode that was positively charged was named the anode (a contraction of anti-cathode). The anode is also called the plate. Vacuum diodes also require a mechanism for heating the cathode. This was provided by the filament.

The vacuum triode was invented in 1906, and was the first electronic device to exhibit the control feature that has been described. de Forest (1914) discovered that by placing a meshlike control surface, called a grid, between the plate (anode) and cathode of a vacuum diode, he could control the electron flow through the mesh or grid by controlling the voltage applied to the grid. Thus he was able to achieve amplification, and his device, called the audion, was incorporated into several commercial devices. In Britain and other parts of the world, the control property gave rise to the term "valves," by which these devices are known.

Although vacuum tubes have been supplanted by transistors in most applications, many vacuum tubes are still in use. One reason is the fact that there is still a lot of useful electronic equipment built around the

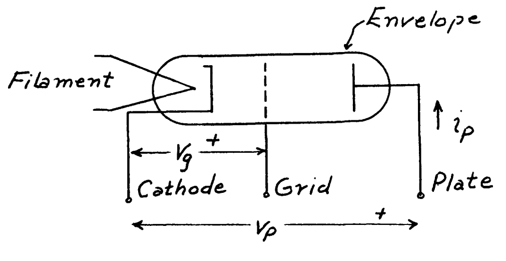

FIG. 4. Symbolic representation of a vacuum triode.

vacuum tube. A second is that in some high-power applications, the vacuum tube still has no competitor. A third reason is that the vacuum tube has some features still considered quite desirable. Some very serious high-fidelity music critics believe that the sound quality achieved in vacuum-tube amplifiers is superior to that from solid-state circuits. The distortion properties exhibited by the tube circuit when signal levels are approaching their maximum permissible limit differ substantially from the distortion properties exhibited by transistor circuits.

7.2 Tube Elements

A symbolic representation of a three-element vacuum tube, or triode, is shown in Fig. 4. The essential elements of a tube are the plate or anode, the cathode, the grid, the filament, and the (glass) envelope (Gray, 1954). The filament is heated electrically, and this heat causes the cathode to emit electrons. (In some tubes, the filament and the cathode are integrated into one unit.) When the plate voltage is positive with respect to the cathode voltage, electrons emitted by the cathode are collected by the plate. The voltage on the grid, which is usually negative with respect to the cathode, controls the electron flow from the cathode to the plate—or the current flow from the plate to the cathode. It is possible to show that, under reasonable assumptions, the voltage–current equation for a triode is given by

$$i_p = K_t[(C_{gk}/C_{pk})v_g + v_p]^{3/2}, \qquad (2)$$

where i_p is the plate current, positive charge; K_t a constant that depends on tube geometry; C_{gk} the grid–cathode capacitance; C_{pk} the plate–cathode capacitance; v_g the grid voltage; and v_p the plate voltage.

Figure 5 shows a typical appearance of these tube characteristic curves. The voltage on the plate, with respect to the cathode, is the independent variable. The voltage on the grid, with respect to the cathode, is the parameter. The plate current is the dependent variable.

As vacuum tubes developed, engineers found that the addition of additional grids to the tubes improved performance, but more importantly, led to different kinds of behavior. The pentode has very different characteristics curves. Figure 6 shows a sample set.

7.3 Vacuum-Tube Circuit Analysis

Two types of analysis are possible, and necessary, in vacuum-tube amplifier circuits. The first, called a "dc analysis," consists of finding the voltage across and the current through the tube when the signal is zero. This result is called the operating point. The sec-

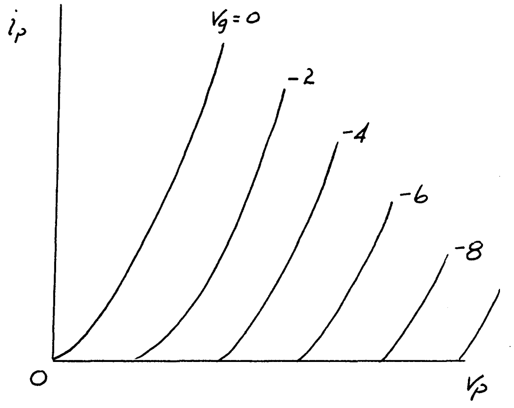

FIG. 5. Typical vacuum-triode current–voltage characteristics.

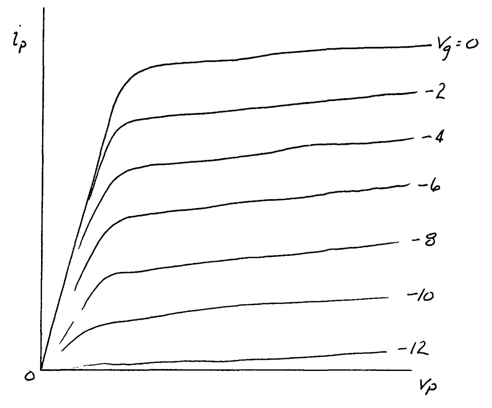

FIG. 6. Typical vacuum-pentode current–voltage characteristics.

ond, called a "small-signal analysis," consists of using equivalent-circuit parameters to find performance measures such as voltage and power gain, output resistance, and efficiency. Figure 7(a) shows a vacuum tube—a triode is chosen, but the analysis with a pentode is quite similar—with a power supply, a load resistor connected between the power supply and the plate, a second (self-bias) resistor connected between the cathode and the power supply, and a third resistor through which the grid is connected to one side of the power supply.

Figure 7(b) shows the characteristic curves with two straight lines added. Line 1 is called the "load line." Its equation is found by applying Kirchhoff's voltage law and Ohm's law to the circuit formed by the power supply, the tube, and the load and bias resistors. The second line relates the grid voltage to the plate current. Usually, the current through the grid resistor is zero, so that Eq. (4) gives the relationship between the grid voltage and the plate current. Specifically,

$$i_p = (V_{bb} - v_p)/(R_p + R_g) \qquad (3)$$

and

$$v_g = -i_p R_g \,. \qquad (4)$$

The intersection of these two lines is labeled Q, and this point tells the designer the plate current and voltage when the signal is zero. In the example shown, the plate voltage is about 150 V, the plate current is about 5.0 mA, and the grid voltage is about −5 V.

Figure 7(c) shows an equivalent circuit. It contains two elements: a voltage-controlled voltage source, μV_g, and the incremental plate resistance r_p. The voltage V_g is the small-signal voltage on the grid with respect to the cathode. The parameter μ is called the (open-circuit) voltage gain for the tube and may be determined from the characteristic curves as the ratio of a change in plate voltage caused by a change in grid voltage near the Q point. The incremental plate resistance may be determined as the slope of a line tangent to the current–voltage characteristic at the Q point.

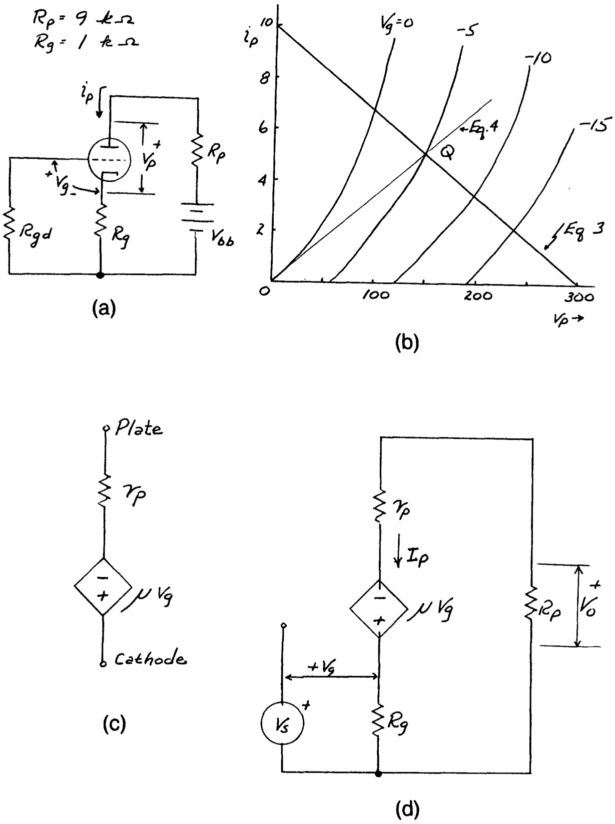

FIG. 7. Vacuum-tube analysis. (a) Circuit. (b) Load-line construction. (c) Small-signal equivalent circuit for the tube. (d) Equivalent circuit for analysis.

Figure 7(d) shows the complete equivalent circuit with external elements added. To analyze this circuit, we write the following equations:

$$I_p=\mu V_g/(R_p+r_p+R_g), \tag{5}$$

$$V_g=-I_pR_g+V_s\,, \tag{6}$$

$$V_o=-I_pR_p\,, \tag{7}$$

where V_s is the signal voltage we wish to amplify and V_o the output signal voltage. If we solve for the circuit gain, we find that

$$\frac{V_o}{V_s}=\frac{-\mu R_p}{R_p+r_p+(1+\mu)R_g}\,. \tag{8}$$

This equation, along with the operating-point analysis, shows us that the design engineer has a number of parameters to work with in the design of circuits. The negative sign means that there is a 180° phase change between input and output signals.

Many—literally thousands—of other useful circuits have been developed to use tubes over the years. A useful reference for these circuits is the book by Reich (1961).

8. ACTIVE FILTERS

8.1 Electrical Filters

One of the most important applications of electronic circuits is the process of filtering electrical and electronic signals (Huelsman and Allen, 1980; Chen, 1986; Fink and Christiansen, 1989, Sec. 12). Filtering is the process of separating desired and undesired signals based on appropriate properties. The frequency of the signal is one of the most commonly used parameters. For example, a filter in the output of a high-fidelity home entertainment system directs the low-frequency signal components to the woofer, the high-frequency signal components to the tweeter, and the remaining components to the midrange speaker. In recent years great advances have been made in this technology, with the result that the designer can now purchase a wide variety of electronic components to meet a wide variety of needs. When the filters contain electronic devices, they are usually called active filters.

8.2 Filter Design Process

In general terms, the design of a filter requires the designer to go through steps like the following set, though the details may vary depending on the situation and the experience of the designer.

8.2.1 Specifications. State the specifications that the filter must meet. These specifications would normally include some or all of amplitude and phase characteristics as functions of frequency, impedance levels at the input and output terminals, source and load characteristics, and consideration of factors such as cost, weight, reliability, sensitivity, environmental operating conditions, noise and signal levels, and other qualities deemed by the user to be important.

8.2.2 Approximation. Choose an appropriate approximation function, if one can be found, or develop a special one if necessary. The well-known approximations that follow will suffice in many situations. The approximations will be given in terms of equivalent low-pass (to be defined in the next few paragraphs) networks, and so it may be necessary to translate specifications to equivalent low-pass specifications.

8.2.3 Circuit. Choose an appropriate circuit. If a decision is being made, for whatever reason, to use an active network, then there are many circuit configurations possible, and the one that is shown to be suitable is chosen. Determine the circuit's transfer function if not already known.

8.2.4 Element Determination. From a knowledge of the transfer function required and the circuit's transfer function, determine the type and magnitude of the circuit elements. Usually they will be either resistors or capacitors, and there will not be a unique set of parameters that will meet the design criteria. Rather, there will be sets of constraint equations to be solved with additional requirements imposed by the designer, such as having two critical capacitors equal.

8.2.5 Network Conversion. If necessary, convert the network back to the required type, and scale elements in impedance level and frequency to meet the operating conditions.

8.3 Filter Types and Approximation Functions

In this section, five basic types of filter are defined, several approximations that are used in the filter design process are introduced, concepts of frequency and element scaling are developed, and one commonly used active network is developed.

8.3.1 Types of Filter. Circuit designers distinguish five basic types of electrical filter. They are as follows:

1. low-pass filter, which transmits signals whose frequency components are below some desired threshold, and attenuates signals whose frequency components are above the threshold;
2. high-pass filter, which transmits signals whose frequency components are above some desired threshold, and attenuates signals whose frequency components are below the threshold;
3. band-pass filter, which transmits signals in a range of frequencies between some required upper and lower limits, and attenuates signals lying outside this band;
4. band-stop or band-reject filter, which attenuates signals in a range of frequencies between some required upper and lower limits, and transmits frequencies lying outside this band;
5. delay filter or circuit, whose primary purpose is to correct or modify the group delay or time delay of signals from the input to the output.

8.3.2 Modern Network Synthesis. Over the years, circuit designers have developed a variety of methods for taking filter performance specifications, which are often communicated in graphical form, and approximating them analytically in a form that leads to useful network functions (see NETWORK THEORY, ELECTRICAL). The techniques to be described here are generally called "modern network synthesis." It is possible to translate high-pass, band-pass, and band-stop filter specifications to equivalent low-pass specifications, to perform the design using low-pass techniques, and to transform the resulting network back to the required type. Many methods for improving the results also exist.

8.3.3 Ideal Low-Pass Filter. An ideal low-pass filter would have a magnitude transmission characteristic of unity at frequencies from dc to the pass-band edge, and zero at frequencies above the pass-band edge [see Fig. 8(a)]. Because scaling is possible, the frequency at the pass-band edge has been scaled to 1.0 rad/s, which corresponds to a frequency of $1/2\pi$ Hz. It is shown in following sections that the specifications and the networks can be scaled and transformed in order to have the appropriate behavior at any desired freqency. Further, this scaling procedure allows computation and publication of normalized data (Zverev, 1967; Saal and Ulbrich; 1958; Weinberg, 1962; Fink and Christiansen, 1989, Sec. 12). Such tabulations are primarily for passive networks, which are not the principal focus of this article.

8.3.4 Magnitude Functions. It is not possible to achieve ideal behavior with finite networks, so it is necessary to approximate the ideal function. Two types of approximation are used—monotonic approximations and equiripple approximations. Figures 8(b) and 8(c) show these approximation types. Further, it is convenient to use magnitude-squared functions. There are two reasons for this. The first is that irrational functions are eliminated. The second is that a magnitude-squared function is easily associated with power transmission.

The process of developing a network function from the approximation requires using the mathematical process called analytic continuation to let $\omega^2 = -s^2$; finding the roots of the resulting polynomial in s, the symbol for the Laplace-transform variable usually used by electrical engineers; and finally choosing the denominator roots (function poles) in the left half of the complex plane. The reason for this choice is that it will lead to a stable network.

Although it does not arise in the first three cases to be discussed here, there are cases in which there are finite roots in the numerator also. The negative of every left–half-plane root is located in the right half-plane, and is also a root of the new function. Provided complex roots are chosen in conjugate pairs, either the left–half-plane or right–half-plane root may be chosen. The latter choice leads to nonminimum phase networks, which meet the criteria in magnitude but have additional phase

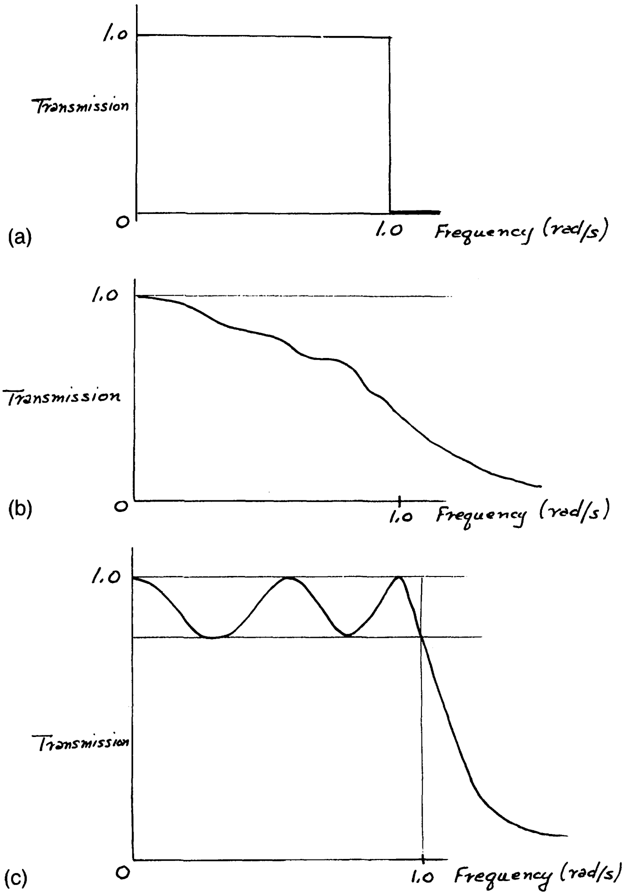

FIG. 8. (a) Ideal low-pass filter transmission characteristic. (b) Monotonic approximation. (c) Equiripple approximation.

shift, which may be an advantage in some situations and a disadvantage in others.

8.3.5 Butterworth Approximation. A useful approximation was developed by the British engineer Butterworth. Butterworth (Chen, 1986) approximated the ideal magnitude-squared function with

$$G(\omega^2)=1/(1+\omega^{2n}), \qquad (9)$$

where n is the order of the approximation. This function exhibits the monotonic behavior of Fig. 8(b), and it also has the characteristic that $2n-1$ derivatives of G are 0 at the origin. For this reason, it is also called a maximally flat monotonic approximation. All orders of this approximation take on a value of 0.5 when $\omega=1$, but higher orders of n give a better approximation to the ideal, at the expense of a more complex network.

If a designer finds that a fourth-order Butterworth response filter meets specifications, i.e., if $n=4$, then use of analytic continuation and root sorting shows that the appropriate network transfer function is

$$G(s)=\frac{1}{(s^2+0.7654s+1)(s^2+1.8478s+1)}. \qquad (10)$$

Several of the references listed at the end of this article give extensive tables for these functions (Zverev, 1967; Fink and Christiansen, 1989, Sec. 12; Weinberg, 1962).

8.3.6 Legendre Approximation. A second monotonic approximation was developed by Papoulis (1959). His goal was to find the monotonic approximation function that yields the sharpest possible discrimination at the edge of the passband. This second criterion means that the magnitude of its derivative with respect to frequency at $\omega=1$ is maximized. The approximation that meets these criteria is

$$G(\omega^2)=K/[1+L_n(\omega^2)],$$

where $L_n(\omega^2)$ is the Legendre polynomial of the first kind. Zverev (1967) and Fink and Christiansen (1989), Sec. 12, give more detail and tabulate the appropriate polynomials and roots in the form of quadratic factors.

8.3.7 Chebyshev Approximation. Chebyshev (various transliterated spellings are found in English) was a Russian engineer who studied harmonic motion, especially in steam engines, and developed the sequence of polynomials bearing his name. Cauer (1958) adapted them for use in network design. The polynomials are defined by the pair of equations

$$C_n(\omega)=\begin{cases}\cos[n\arccos(\omega)], & |\omega|\leqslant 1\\ \cosh[n\,\mathrm{arccosh}(\omega)], & |\omega|>1\end{cases}. \qquad (11)$$

It is seen that $C_0(\omega)=1$, $C_1(\omega)=\omega$. There is a recursion formula, based on trigonometric identities, that states

$$C_{n+1}(\omega)=2\omega C_n(\omega)-C_{n-1}(\omega). \qquad (12)$$

From this recursion formula, polynomials of any order can be constructed. For example, we can use the formula three times to show that

$$C_4(\omega)=8\omega^4-8\omega^2+1. \qquad (13)$$

The approximation uses the polynomials and a "ripple factor," usually denoted by the Greek letter $\epsilon\leqslant 1$, in the approximation function (Chen, 1986)

$$G(\omega^2)=1/[1+\epsilon^2C_n^2(\omega)]. \qquad (14)$$

This approximation gives a much sharper cutoff or discrimination at the edge of the passband than either the Butterworth or Legendre approximations, but the monotonic behavior is gone. Depending on whether the order n is odd or even, two differing types of equiripple approximation exist, shown in Figs. 9(a) and 9(b). It is seen that the factor ϵ controls the amount of ripple in the passband, and the order n controls the number of oscil-

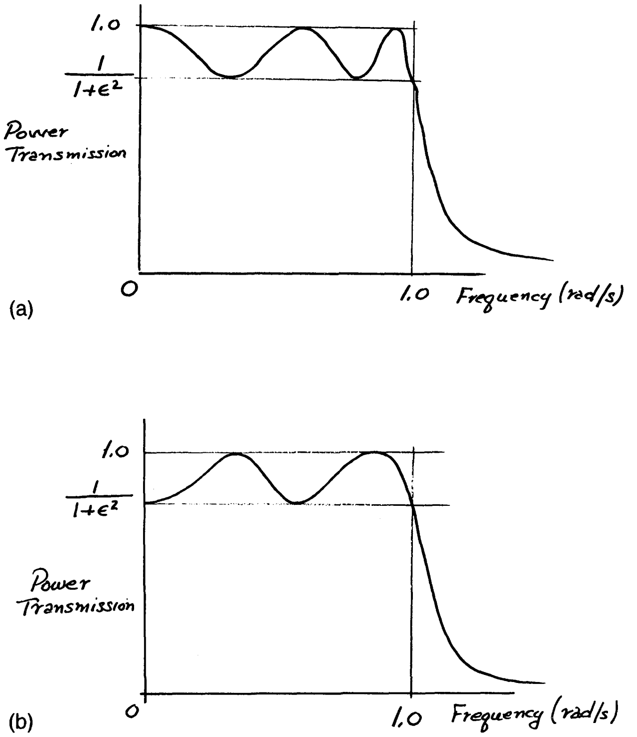

FIG. 9. (a) Chebyshev approximation for n odd. The drawing is for $n=5$. (b) Chebyshev approximation for n even. The drawing is for $n=4$.

lations. The function is monotonic in the stop band. The process of finding the quadratic polynomials in s is similar to that for Butterworth functions, but the process is more complex because of the need to consider two variables. Zverev (1967) and Fink and Christiansen (1989) Sec. 12, provide extensive lists. For example, if $\epsilon=0.3493$ and $n=4$, we find the transfer function, after using analytic continuation and root sorting, to be

$$G(s)=\frac{K}{(s^2+0.3507s+1.0635)(s^2+0.8467s+0.3564)}, \tag{15}$$

where K is a constant that can be adjusted, especially in active filters when there is a possibility of signal amplification as well as filtering.

8.3.8 Elliptic Filters. A more complex but more powerful approximation function, known variously as the Cauer-elliptic approximation, the elliptic function approximation, and the Chebyshev rational function approximation, has equiripple characteristics in both the pass band and the stop band. With elliptic functions, very sharp transitions between pass bands and stop bands are possible. These functions require three parameters for complete characterization, and as a result, the tables required are extensive. Zverev (1967) and Saal and Ulbrich (1958) give many combinations.

8.3.9 Bessel Filters. Although it is a low-pass filter, the primary use for the Bessel (also known as the Bessel-Thomson) approximation is to give a constant group or time delay for all signal frequencies below some threshold. A sequence of Bessel polynomials, not to be confused with Bessel functions, provides this approximation. The delay achieved is a maximally flat approximation to the ideal. Extensive tables of roots and other properties are provided in Zverev (1967), Weinberg (1962), and Fink and Christiansen (1989), Sec. 12.

8.4 Frequency Transformations and Scaling

The goal in this section is to show how to convert a low-pass function with a normalized band edge to a function with the values one is likely to find in practice. The first discussion will show how to scale the frequency from 1.0 rad/s to some value a. Then, transformations from low pass to high pass, band pass, and band stop will be considered. In doing this, the variable p will be substituted for s in the normalized, low-pass equations, so that s can be used in the final result.

8.4.1 Scaling. If a low-pass filter has been designed with a band edge at 1.0 rad/s, and it is desired to have the band edge at a new value of a rad/s, then the proper frequency scaling will be given by

$$p=s/a. \tag{16}$$

If we apply this to the normalized function in Eq. (10), which we rewrite as

$$G(p)=\frac{1}{(p^2+0.7654p+1)(p^2+1.8478p+1)} \tag{17}$$

and let $a=10^4$ rad/s, corresponding to a frequency of 1591.5 Hz, we get

$$G(s)=\frac{10^{16}}{(s^2+7654s+10^8)(s^2+1.8478s+10^8)}. \tag{18}$$

8.4.2 Low-Pass to High-Pass Transformation. The transformation $p=1/s$ will have the effect of inverting the frequency axis, so that the band $0\leqslant\omega\leqslant 1$ will become $1\leqslant\omega$, and the behavior that formerly existed in the low-pass region will now take place in the high-pass band. If we apply this technique to the function of Eq. (17), we get

$$G(s)=\frac{s^4}{(s^2+0.7654s+1)(s^2+18478s+1)}. \tag{19}$$

It would then be necessary to use frequency scaling on this function.

8.4.3 Low-Pass to Band-Pass Transformation. The transformation

$$p=(s^2+\omega_0^2)/Bs, \tag{20}$$

where B is the desired filter bandwidth and ω_0 is the desired frequency at the passband center, will accomplish this objective. For example, if it were desired to transform the low-pass filter of Eq. (17) to a band-pass filter having a bandwidth of 1000 Hz (6283 rad/s) and a center frequency of 8000 Hz (50 265 rad/s), then we would use the transformation

$$p=(s^2+50265^2)/6283s. \qquad (21)$$

If done without programmable calculators or computers, the arithmetic can become tedious, but it is straightforward. Alternatively, Eq. (20) can be solved for s by using the quadratic formula, and the roots of the low-pass quadratic equations can be transformed to the roots of the band-pass polynomials and the band-pass quadratic factors found without the necessity for finding roots of higher-order equations.

8.4.4 Low-Pass to Band-Stop Transformation. This transformation is the reciprocal of the low-pass to pass-band transformation. The reciprocal of Eq. (20) is used.

8.4.5 Element Magnitude Scaling. Much network design uses element value scaling, so that calculations are eased somewhat by working with a normalized load resistance of 1.0 Ω. This property also permits publication of extensive tables of element values, especially for passive networks (Zverev, 1967; Fink and Christiansen, 1989). At the end of the design, practical values are achieved by multiplying the values for inductors and resistors by the ratio of the load resistor to the normalized value, dividing the capacitance by this same value. Active network terms that have dimensions of ohms are multiplied by the same factor, terms that have dimensions of siemens (mhos) are divided by the same factor, and dimensionless terms are not changed.

8.4.6 Element Frequency Scaling. If a circuit designer has developed a network by using functions that are normalized in frequency, then the network elements can be scaled so that the desired frequency behavior occurs at a new frequency. For example, if the design has taken place at a normalized frequency of 1.0 rad/s, and the need is to translate the behavior to a frequency of 10^5 rad/s (15 915 Hz), then both capacitances and inductances are divided in magnitude by 10^5. Resistances are unaffected.

8.5 Network Realizations

8.5.1 Prototype Circuits. Once polynomials have been developed to meet the specifications, the next step is to find an electrical network. In this case it is assumed that an active or electronic circuit is desired. The number of possibilities is quite extensive, and only one will be described here. The circuit is based on the pioneering work of Sallen and Key (1955), and it uses operational amplifiers (See AMPLIFIERS, ELECTRONIC) as the central element. The circuit is usually called a multiple-feedback (MFB), infinite-gain circuit, as it depends on the very high gain achieved by the operational amplifier. Figures 10(a), 10(b), and 10(c) show the low-pass, band-pass, and

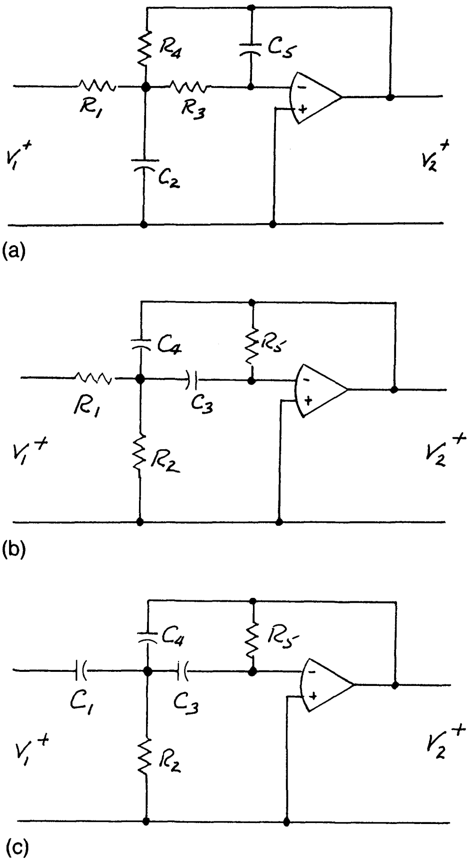

FIG. 10. Three multiple-feedback (MFB) second-order active filters. (a) Low-pass circuit. (b) Band-pass circuit. (c) High-pass circuit.

high pass configurations, respectively. A bandstop circuit is not possible with this basic circuit. The respective transfer functions for these circuits are

$$\frac{V_2}{V_1}(s) = -\frac{\dfrac{1}{R_1R_3C_2C_5}}{s^2+s\left(\dfrac{1}{C_2}\right)\left(\dfrac{1}{R_1}+\dfrac{1}{R_3}+\dfrac{1}{R_4}\right)+\dfrac{1}{R_3R_4C_2C_5}}, \tag{22}$$

$$\frac{V_2}{V_1}(s) = -\frac{\dfrac{s}{R_1C_4}}{s^2+s\left(\dfrac{1}{R_5}\right)\left(\dfrac{1}{C_3}+\dfrac{1}{C_4}\right)+\dfrac{1}{C_3C_4R_5}\left(\dfrac{1}{R_1}+\dfrac{1}{R_2}\right)}, \tag{23}$$

$$\frac{V_2}{V_1}(s) = -\frac{s^2\left(\dfrac{C_1}{C_4}\right)}{s^2+s\left(\dfrac{1}{C_3C_4R_5}\right)(C_1+C_3+C_4)+\dfrac{1}{R_2R_5C_3C_4}}. \tag{24}$$

Since the polynomials for the denominator are known in terms of quadratic factors, the designer can then choose the five element values to meet the numerical conditions. Since, in general, only three specifications must be met—two denominator coefficients and the numerator coefficient—and since five elements must be specified, the designer has two choices to make. Normally, the designer will choose these values to improve the circuit quality in some desirable sense. One possiblity, for example, is to have all capacitors equal. Another is to have a desirable ratio or equality between two resistor values. Still another is to minimize change in circuit performance caused by aging.

8.5.2 Cascading Second-Order Circuits. If a fourth-order circuit is needed, then the designer will have the option of using two of the preceding second-order circuits, connected in "cascade," as indicated in Fig. 11. With a sixth-order circuit, three such stages are needed. This works well because, under normal conditions, the stages do not interfere with or "load down" preceding stages. The ordering can be important. One good rule is to have the first or input stage realize the pair of poles farthest from the $j\omega$ axis, and then to have succeeding stages realize pole pairs progressively closer to the $j\omega$ axis.

8.5.3 Additional Active Filters. Many other techniques are available to design and build active or electronic filters. Some are similar to the circuits described earlier, in that they depend on a cascading of second-order circuits. Some do not have the negative sign, which represents a 180° phase shift that is not always acceptable.

Other techniques use combinations of operational amplifiers, resistors, and capacitors to produce devices called gyrators or frequency-dependent negative resistors (FDNRs) that replace inductors in conventional passive filter circuits (Chen, 1986).

The Huelsman-Kerwin-Newcomb (HKN) (Huelsman and Allen, 1980) filter uses operational amplifiers in a configuration that follows from state-variable descriptions of circuits. These are available commercially from a variety of sources.

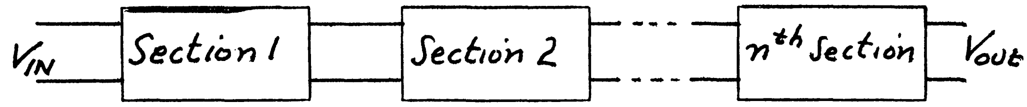

FIG. 11. Cascade of second-order active filter sections.

Two other techniques (Schauman, *et al.,* 1990; Geiger *et al.,* 1990) that are important are the switched-capacitor (SC) filter and the operational transconductance amplifier (OTA) filter. SC filters generally use the circuit topology of other active filters. However, resistors are implemented with capacitors, and the entire package is built in CMOS (complementary metal oxide semiconductor) technology. An external clock switches the capacitors between charge and discharge cycles. The circuit is tuned by adjusting the clock frequency.

While the operational amplifier is described as a voltage-controlled voltage source, the OTA is characterized as a voltage-controlled current source. The development of the OTA has led to new circuit implementation techniques, expanding the range of possibilities. Some circuits use OTAs to implement circuit elements such as resistors, though external capacitors are required in parts of the circuit.

GLOSSARY

Active Filter: An electrical filter that includes one or more electronic devices in its construction. See also **Filter**.

American Wire Gauge: A system used for designating wire sizes in the United States and elsewhere.

Bessel Approximation: A low-pass electrical filter approximation. The primary purpose of a Bessel filter is to provide a fixed time or group delay. Also known as the Bessel-Thomson approximation.

Bessel Polynomial: A sequence of polynomials, not to be confused with Bessel functions, used in the Bessel approximation.

Butterworth Approximation: A sequence of filter approximations leading to filters with maximally flat magnitude versus frequency characteristics in the pass band.

Chebyshev Approximation: A ripple type of approximation that is used in electrical filter design. It is based on the sequence of Chebyshev polynomials.

Chebyshev Polynomial: A sequence of polynomials defined by $C_n(\omega)=\cos(n \operatorname{arccos}\omega)$, $|\omega|\leqslant 1$, or by $C_n(\omega)=\cosh(n \operatorname{arccosh}\omega)$, $1\leqslant|\omega|$.

Elliptic-Function Approximation: A ripple type of filter approximation based on elliptic functions of the first kind.

Filter: An electrical circuit designed to separate components of an electrical signal based on one or more signal properties, such as frequency.

Hole: A vacancy in a crystal structure caused by the removal of an electron.

Inverter: An electrical circuit designed to convert dc to ac.

Legendre Approximation: A filter approximation that is monotonic in the pass band and that is based on Legendre polynomials.

Power Supply: In general, a component of an electrical system that provides electrical power to the system. A somewhat restricted definition is the subsystem that converts ac power to dc power.

Printed Circuit Board: A board, usually of laminated plastic, on which electrical components are mounted.

Rectifier: A circuit that converts ac to dc and that is a part of a power supply. (It also has other uses.)

Vacuum Diode: A two-terminal electrical device that has two electrodes mounted in an evacuated envelope and passes current in only one direction.

Vacuum Triode: A three-terminal electrical device that has three electrodes mounted in an evacuated envelope.

Vacuum Tube: An electrical decvice with two or more electrodes mounted in an evacuated envolope. Often the name indicates the number of electrodes. A pentode has five electrodes.

Valve: A synonym, used in Britain and elsewhere, for vacuum tube.

List of Works Cited

Brittain, J. E. (Ed.) (1977), *Turning Points in American Electrical History,* New York: IEEE, Sec. II-D.

Cauer, W. (1958), *Synthesis of Linear Communication Networks,* Vols. I and II (a translation from the German, edited by Wilhelm Klein and Franz M. Pelz), New York: McGraw Hill.

Chen, Wai-Kai (1986), *Passive and Active Filters,* New York: Wiley.

de Forest, L. (1914), "The Audion—Detector and Amplifier," *Proc. IRE* **2** (1), 15–21. (Reprinted in Britain, 1977.).

Ebers, J. J., Moll, J. L. (1954), "Large-Signal Behavior of Junction Transistors," *Proc. IRE* **42**, 1761–1772.

Fink, D., Beaty, W. (1987), *Standard Handbook for Electrical Engineers,* 12th ed. New York: McGraw-Hill, Sec. 4, Properties of Materials; Sec. 22, Electronics.

Fink, D., Christiansen, D. (1989), *Electronics Engineers Handbook,* 3rd ed. New York: McGraw-Hill, Sec. 1, Basic Phenomena of Electronics; Sec. 6, Properties of Materials Sec. 7, Discrete Circuit Components; Sec. 8, Integrated Circuits and Microprocessors; Sec. 9, UHF and Microwave Devices; Sec. 12, Filters and Attenuators; Sec. 15, Power Electronics.

Geiger, R. L., Allen, P. E., Strader, N. R. (1990), *VLSI: Design Techniques for Analog and Digital Circuits,* New York: McGraw-Hill.

Gray, T. S. (1954), *Applied Electronics,* 2nd ed., New York: Wiley, Chaps. 1 and 4.

Huelsman, L. P., Allen, P. E. (1980), *Introduction to the Theory and Design of Active Filters,* New York: McGraw-Hill.

Papoulis, A. (1959), "A New Class of Filters," *IRE Trans. Circuit Theory* **CT-6**, 277–281.

Reich, H. J. (1961), *Functional Circuits and Oscillators,* Princeton: Van Nostrand.

Ryder, J. D., Fink, D. G. (1984), *Engineers and Electrons,* New York: IEEE, Chaps. 4, 5, and 7.

Saal, R., Ulbrich, E. (1958), "On the Design of Filters by Synthesis," *IRE Trans. Circuit Theory* **CT-5**, 284–327.

Sallen, R. P., Key, E. L. (1955) "A Practical Method of Designing RC Active Filters," *IRE Trans. Circuit Theory* **CT-2**, 74–85.

Schaumann, R., Ghausi, M. S., Laker, K. R. (1990), *Design of Analog Filters,* Englewood Cliffs, NJ: Prentice-Hall.

Weinberg, L. (1962), *Network Analysis and Synthesis,* New York: McGraw-Hill.

Zverev, A. (1967), *Handbook of Filter Synthesis,* New York: Wiley.

Further Reading

Burns, S. G., Bond, P. R. (1987), *Principles of Electronic Circuits,* St. Paul, MN: West.

Coombs, Clyde (1988), *Printed Circuits Handbook,* 3rd ed. New York: McGraw-Hill.

Gray, P. E., Searle, C. L. (1969), *Electronic Principles: Physics, Models, and Circuits,* New York: Wiley.

Horowitz, P., Hill, W. (1989), *The Art of Electronics,* 2nd ed. New York: Cambridge University Press.

Humphreys, D. S. (1970), *The Analysis, Design, and Synthesis of Electrical Filters,* Englewood Cliffs, NJ: Prentice-Hall.

Reich, H. J. (1961), *Functional Circuits and Oscillators,* Princeton: Van Nostrand.

Lindquist, C. S. (1977), *Active Network Design with Signal Filtering Applications,* Long Beach, CA: Steward.

Stephenson, F. W. (Ed.) (1985), *RC Active Filter Design Handbook,* New York: Wiley.

ELECTRONIC COMPONENTS AND ASSEMBLIES: RELIABILITY AND TESTING

R. B. COMIZZOLI AND J. D. SINCLAIR, *AT&T Bell Laboratories, Murray Hill, New Jersey, U.S.A.*

Introduction 21
1. **History** 22
2. **Reliability, Failure Mechanisms, and Testing** 24
2.1 Passive Components 24
2.1.1 Capacitors 24
2.1.2 Inductors, Coils, and Transformers 24
2.1.3 Relays 24
2.1.4 Resistors 24
2.2 Active Components and Integrated Circuits 24
2.2.1 Silicon Diffusion Effects 25
2.2.2 Silicon and Silicon Dioxide Interface 25
2.2.3 Metal-Related Mechanisms 26
2.2.3.1 Electromigration 26
2.2.3.2 Mechanical Stress 27
2.2.3.3 Corrosion 27
2.2.4 Insulating-Layer Mechanisms 28
2.2.4.1 Surface-Ion Failure Mechanism 29
2.2.4.2 Charge Accumulation in Multilevel Insulators 30
2.2.4.3 Electrical Breakdown of Silicon Dioxide 30
2.2.5 Electrical Overstress 30
2.2.5.1 Electrical Latch-up 30
2.2.5.2 Electrostatic Discharge Effects 31
2.2.6 Radiation Effects 31
2.3 Enclosures for Integrated Circuits 32
2.3.1 Environmental Factors and Manufacturing Contamination 32
2.3.2 Package Types, Wire Bonding, and Chip Mounting 33
2.3.3 Plastic Packages and Flip-Chip Encapsulated Devices 33
2.3.4 Ceramic Packages 37
2.4 Assemblies 38
2.4.1 Interconnection of Components and Integrated Circuits 38
2.4.1.1 Circuit Boards 38
2.4.1.2 Failure Processes and Stress Tests 38
2.4.2 Interconnection of Circuit Boards 39
2.4.2.1 Connectors and Backplanes 39
2.4.2.2 Failure Processes and Stress Tests 40
3. **Future Requirements in Reliability and Testing** 41
3.1 Integrated Circuits—Shrinking Dimensions 41
3.1.1 Device Scaling 41
3.1.2 Particle Effects 42
3.1.3 Advanced Packaging 42
3.2 Reliability Verification and Improvement 42
Glossary 43
Works Cited 44
Further Reading 44

INTRODUCTION

The term reliability, as applied to the design and manufacture of electronic components and assemblies, is commonly defined as meeting the performance expectations of the user over the intended life of the product. The process by which reliability is achieved begins at the earliest stages of product definition and is, of necessity, a multidisciplinary evaluation of all the design, materials, manufacturing, and environmental factors that will bear on the product's performance in its intended use. The prerequisites for eliminating failures are

1. engineering evaluations that use statistical experimental methods and appropriately

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

conceived and designed accelerated-life tests to uncover potential failure mechanisms,

2. failure mode analysis and laboratory experimentation that leads to scientific understanding of degradation processes, and
3. design procedures, materials selection, and manufacturing processes that avoid deficiencies responsible for failure. Stress tests that uncover root causes are best conducted during prototype development. These tests lead to reliability models that can then be used during the development of the manufacturing process to predict product reliability.

Development of high-performance, reliable electronic components and assemblies requires

1. an in-depth knowledge of the properties and interactions of materials with the surrounding world, whether it be the ambient environment or other materials;
2. scientific understanding of the behavior of electrons, photons, and atoms in the materials;
3. a detailed understanding of the effects of processing conditions on materials properties and performance;
4. robust designs that ensure the achievement of the desired properties in spite of process variability;
5. characterization of the materials and components required to manufacture the product to ensure that tolerances are being met; and
6. a scientifically based accelerated-life–test strategy that examines all functional levels of a product including semiconductor devices, discrete components, interconnection and packaging, circuit boards, modules, and entire systems.

This article describes the critical reliability issues with electronic devices and assemblies and the processes by which failures are discovered, mechanisms are understood, and prevention technologies are developed. The kinds of approaches that can be taken to ensure that performance objectives are met will be illustrated by specific examples. Failure mechanisms, accelerated testing methods, and failure prevention strategies will be considered at each level of functionality in electronic equipment, beginning with discrete components and integrated circuits and then extending to packaging, to assemblies, and finally to large mainframe computers or other systems. References with more extensive information on certain aspects of each of these functionality levels can be consulted as needed. The statistical aspects of reliability evaluation, testing, and estimation will not be described. The pertinent statistical models and methods can be found in the references.

Following this introduction is a brief history of accelerated-life testing, failure mode analysis, and reliability methodologies. The subsequent section covers reliability assurance methods for the various levels of functionality in electronic equipment. The final section considers future requirements in reliability and testing.

1. HISTORY

Since the earliest days of telephony and telegraphy, reliable performance has been essential to the proliferation of new types of electronic equipment and systems. Achieving near 100% reliability in telecommunications is now viewed as essential to the public welfare—sometimes a life-and-death concern for the general public. The same is true for computer systems used for aircraft navigation, space vehicles, and hospital monitoring systems. Laboratory-based reliability evaluations and accelerated-life testing for electronic equipment date back at least to the 1930s. At that time all components were quite large by today's standards, a large fraction of the functionality was achieved with vacuum tubes or electromechanical devices such as relays, and interconnection between components was almost entirely by wires. Faults were frequently caused by contamination at contact interfaces and tended to be intermittent rather than permanent. Tracing their source was much more difficult than in later years with transistors. The exception was vacuum tubes, the predecessor to transistors. Vacuum tubes failed permanently at rates that were many orders of magnitude higher than those for transistors in today's integrated circuits.

The earliest computers failed to operate correctly as frequently as once every few thousand cycles. Even relatively simple computations required dual calculations by identical logic units operating in a match–compare mode to ensure accurate results.

Also in the 1930s, when pollution was a major problem in metropolitan areas, corrosion and tarnishing of metals important to electronics, such as copper, copper alloys, silver, tin, and lead, became a reliability concern. Sulfide, sulfate, and nitrate corrosion products formed on these metals from exposure to sulfur dioxide, nitrogen oxides, the acids formed from these gases, and hydrogen sulfide. Accelerated corrosion tests were conducted on metal coupons to determine the maximum tolerable exposure levels that would not compromise reliability and to find alternative metallizations that could stand up to the corrosive gases.

A landmark book (Shewhart, 1939) on the use of statistical quality-control charts and quality-control methods for the manufacture of electronic equipment appeared in 1939. Through 1950, quality control through inspection and acceptance testing was the only reliability evaluation performed, with rare exception. Testing with corrosive gases and, to a minor extent, artificial dust was part of the qualification procedure that electronics manufacturing companies used for electrical contact materials in the mid-1950s. By 1960 most products underwent at least some form of wear or environmental testing, statistical techniques were frequently applied to reliability prediction, and every supplier of electronic equipment supported a substantial failure mode analysis program. The first Physics of Failure Symposium was held in 1962 and Military Standard-883 emerged in 1968 as a broadly applied standard for reliability evaluation.

By 1970 the application of reliability programs and life testing became an essential part of the development program for equipment used by governments and their military programs. More recently Deming (1982) wrote a widely read book on quality assurance and statistical process control. With the advent of the transistor in the 1950s, and later integrated circuits that coupled hundreds and thousands of transistors together on the same silicon chip, reliability requirements soared. The fact that one failed transistor could destroy an entire chip demanded unprecedented reliability and entirely new accelerated-life test methods. By the mid-1960s, electronic switching systems were able to handle all the logical steps required to transmit thousands of phone calls per day with a down time of 2 h in 40 years. The era of high reliability had clearly been born.

Standard test methods for accelerated-life testing, including military specifications and procedures adopted by professional societies, have been used by many companies and governments as part of the purchasing criteria since the late 1960s. These test methods include salt spray tests, static and cyclic temperature and humidity tests, static and cyclic temperature-humidity-voltage tests, static and cyclic mechanical test procedures, corrosive-gas tests, and dust-particle tests. Though these early tests were successful in discovering defects and led to prevention strategies for many types of failures, they also missed many failures that emerged during use of the product because they failed to activate the defects and true mechanisms responsible for field failures. Their use also caused unnecessary changes to be made because they sometimes uncovered defects that would not lead to failures in the use environment. These kinds of misleading results remain a problem today. In modern computing and telecommunication systems, the best reliability achievable in devices is not good enough and redundancy must still be built into the system design. Complicating the problem of achieving reliability even further is the rapid integration of photonic and electronic devices, components, and interconnects into the same system. Entirely new life-testing methods will be needed to unravel potential failure processes.

Today the large number of regularly scheduled meetings like the Reliability Physics Symposium and the Components, Hybrids, and Packaging Symposium sponsored by the Institute of Electrical and Electronic Engineers, the Fault Tolerant Design Symposium, and the Reliability and Corrosion of Electronic Materials and Devices Symposium sponsored by The Electrochemical Society are clear indicators of the continuing need for improved fundamental understanding of defects and failure mechanisms, advanced stress-testing methodologies, prevention technologies with broad applicability, and reliability models that include all the relevant factors in predicting the performance of mainframe computers and computer networks.

2. RELIABILITY, FAILURE MECHANISMS, AND TESTING

2.1 Passive Components

A component is a physically distinct part that encompasses a single electrical function. Passive components provide no amplification of current or voltage and are often the simplest ones in construction and in terms of reliability phenomena (Klinger *et al.*, 1990). Passive components include such devices as capacitors, resistors, inductors, transformers, and relays.

2.1.1 Capacitors. Of all the passive components, the aluminum electrolytic capacitor has the shortest useful life because of its characteristic wear-out mechanism. This capacitor has electrodes of aluminum contained in a housing or "can" with an electrolyte. A rubberlike polymer seals the can. Loss of electrolyte through the seal results in failure, with the mean time to failure varying from about 10 to 20 years. Failure rates increase with temperature and decrease with capacitor physical size. These capacitors should not be exposed to halogenated solvents since such solvents can penetrate the seal and cause internal corrosion and failure.

Another type of capacitor, commonly referred to as an electrolyte-filled ("wet") tantalum capacitor, has sometimes failed in spectacular fashion after experiencing a power surge or reversed-bias condition. Internal short circuits form, leading to excessive current, extreme heat, and possibly fire.

Other capacitor types are not prone to wear-out failures. However, care is needed in circuit design and in selecting the capacitor. Use conditions, such as temperature and humidity, must be kept in mind to avoid short- or open-circuit failures resulting from random defects.

2.1.2 Inductors, Coils, and Transformers. The most important failure mechanism in inductors, coils, and transformers is electrical breakdown of the wire insulation. The insulating properties of polymer wire insulation degrade with time and increased temperature. Eventually, a short circuit may form between windings. Local heating will then further degrade the insulation, causing more local short circuits. In a safely designed transformer, this local heating will result in an open circuit before a fire occurs. A fuse is sometimes incorporated in the design as a safety measure.

2.1.3 Relays. Relays perform a switching function and consist of one or more pairs of movable contacts. Normally a coil is energized to move one of the contacts in a pair either to open or to close the connection. Selection of the proper relay for the conditions of voltage, current, number of expected operations, and environment is essential to reliability. Important failure mechanisms include erosion of the contact metal by high-current arcs between the contacts as they open or close, and formation of insulating films or particles on the contact surfaces. These films or particles result from chemical corrosion or arc-induced formation of adherent residues from airborne organic or silicone contaminants.

2.1.4 Resistors. A resistor is a two-terminal device in which the voltage across the terminals is directly proportional to the current. Resistors dissipate power in operation, and the main reliability issues relate to the temperature rise. Proper selection for the application avoids changes in resistance value and catastrophic failure by keeping resistor temperature within allowed bounds. In applications in which high voltages are applied, the dielectric strength of the resistor must be adequate to prevent arcs.

2.2 Active Components and Integrated Circuits

Active components amplify current or voltage and, in a circuit, can be sources of current or voltage. Active components include transistors, thyristors, rectifiers, and diodes. Some of the reliability problems and failure mechanisms found in active components also apply to integrated circuits (ICs). Historically, these mechanisms were first discovered and studied in active components such as diodes and transistors, but similar failure mechanisms can occur in integrated circuits if the knowledge gathered from components is not wisely applied. Many mechanisms were not discovered until integrated circuits were produced in large volumes. These various mechanisms are grouped here according to the types of materials and their interactions.

2.2.1 Silicon Diffusion Effects. An early and important failure mechanism of diodes and transistors was related to the presence of clusters of crystal defects during the diffusion of dopants to form *p-n* junctions (Queisser, 1963). Impurities or mechanical damage at or near the surface of the silicon locally disrupts the crystal structure. When dopants are diffused into the silicon, their diffusion rates are greater at these crystalline defects. Thus, the diffusion proceeds to a greater distance, and the underlying layer width is locally reduced, as shown in Fig. 1. This width reduction can result in reduced reverse-bias voltage capability (lower breakdown voltage) or locally increased forward current resulting in a local temperature rise. In the former case, if the breakdown voltage reduction is large enough, failures will occur when the applied voltage exceeds the degraded capability. In the latter case, local heating causes more current to pass through the region (current crowding), further increasing the temperature until a runaway instability (sometimes referred to as second breakdown) takes place, possibly causing melting and catastrophic failure.

Crystal defects and impurities were eliminated or reduced by avoiding mechanical damage or contact with metallic handling tools, by incorporating highly efficient cleaning procedures in processing, and by minimizing high-temperature excursions. Also, the effects of defects and impurities are neutralized by actually introducing lattice damage to the back of wafers to getter impurities. Current crowding can be avoided by forming an array of small emitters rather than one large emitter (Stolnitz, 1965). The small emitters are connected to each other by a resistive material such that any tendency to current crowding in one emitter is negated by the forward-bias reduction in that emitter generated by the excess current. The excess current causes a voltage drop in the resistive material, tending to turn off this part of the device.

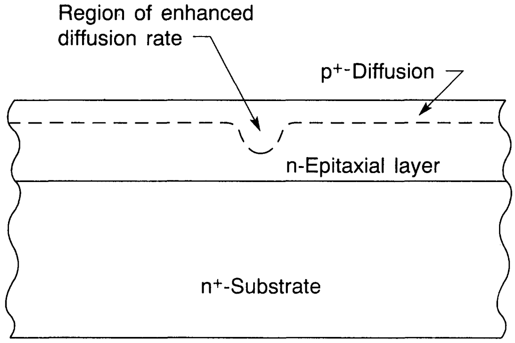

FIG. 1. Reduced breakdown voltage in diffusion-doped diodes. During diffusion of a p^+-n diode, locally enhanced diffusion rate results in a narrower *n* region. When voltage is applied, the high-field region may intersect the highly doped n^+ substrate, resulting in reduced breakdown voltage capability.

When aluminum is used as a contact material to silicon, the presence of defect clusters may cause rapid local interdiffusion of the aluminum and silicon. If the aluminum diffuses through the emitter of the transistor to the base (aluminum spiking), a short between emitter and base is formed. This short can occur during processing as a yield loss or during operating life, especially at elevated temperature. Interdiffusion can be eliminated or reduced by depositing a barrier metal, such as tungsten, between the silicon and aluminum, by doping the aluminum with silicon at deposition to reduce interdiffusion, or by minimizing processing and use temperatures.

The breakdown voltage of a reverse-biased junction may be reduced by the diffusion of copper into the silicon. The source of copper can be subtle, even ephemeral. For example, exposed copper within the cavity of a hermetically packaged transistor can react with hydrogen in the cavity atmosphere to form copper hydride vapor (Bergh, 1964). The vapor diffuses to the transistor surface where it is converted back to metallic copper. This can be avoided by sealing the device with an internal ambient that is free of hydrogen or by not having exposed copper in the package.

2.2.2 Silicon and Silicon Dioxide Interface. The strength of the electric field at the silicon surface depends on the presence of charge at the Si-SiO_2 (silicon dioxide) interface, within the SiO_2, or on the surface of the SiO_2 or other insulators. The charge in these various regions has several origins or sources (Deal, 1974). This section discusses effects related to drifting alkali ions (such as sodium) in the oxide, fixed charge at the Si–SiO_2 interface, and interface charge. Interface charge is related to traps (localized energy states that capture and release charge) and can exist in oxide traps or in traps at the Si-SiO_2 interface. The distinguishing feature of trapped charge is that the charge density depends on the number of traps filled or

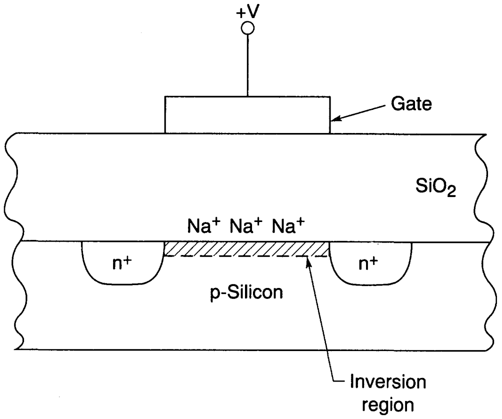

FIG. 2. Reduced threshold voltage caused by free sodium ions. During normal operation of NMOS transistors, a positive voltage on the gate turns the transistor "on." This positive voltage causes sodium ions to accumulate at the Si-SiO_2 interface. In subsequent operation, the device will turn on at lower-than-normal applied gate voltage.

empty; traps fill or empty by exchanging charge with carriers in the silicon. In contrast, the fixed charge is constant in time and location. Discussion of the effects of the charge residing on the insulator surface or at the interfaces between layered insulators is left for a later section.

The amount of charge in the oxide and the density of traps depend on impurity content, silicon crystal orientation, oxide formation conditions, annealing processes, ionizing radiation, and a number of other factors. These effects are extremely complex and are manifested by changes in a variety of electrical parameters. However, it is possible to describe briefly two of the major reliability effects related to oxide or interface charge.

Sodium ions, which have a single positive charge, are mobile in SiO_2 and move readily in the presence of an electric field, especially at elevated temperature. The numerous sources of sodium ions include dust particles; human-related contamination; impure water, solvents, acids, and other process chemicals; quartz furnace tubes used for silicon oxidation; deposited metal from contaminated sources; sodium-containing polymers that form part of the device package; and many others. Either sodium is incorporated throughout the oxide during oxide growth, as in a contaminated furnace, or it may be deposited on the surface of the oxide. In the latter case, elevated temperature causes the sodium to diffuse readily throughout the oxide.

If the sodium drift is toward the silicon, the resulting accumulation of positive charge at the Si-SiO_2 interface can affect the potential of the silicon surface. The change in potential makes the silicon less *p*-type or more *n*-type, changing the threshold voltage of MOS transistors (Fig. 2) or causing resistive short circuits between the emitter and collector of an *npn* transistor (Nicollian and Brews, 1982) (Fig. 3).

The second effect is related to the injection of electrons into SiO_2 from the silicon. In the high–electric-field region of a reverse-biased junction or a channel region in an MOS device, some conduction electrons will gain sufficient energy to leave the silicon and enter the SiO_2 (Hu *et al.*, 1985). This process, over a period of time, results in accumulation of negative charge at the interface, thereby increasing, for example, the threshold voltage of an NMOS device. This effect is minimized by altering the structure of MOS devices to decrease the peak electric fields.

2.2.3 Metal-Related Mechanism

2.2.3.1 Electromigration. Electromigration (Black, 1967) can cause failures in aluminum metal interconnection in integrated circuits and other devices. Current through the aluminum is carried by the passage of electrons. The electrons transfer a fraction of their momentum to the aluminum atoms, thus resulting in a force on the atoms. This force causes the atoms to move and to form

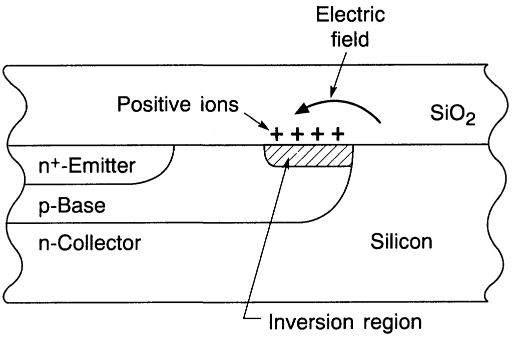

FIG. 3. Resistive shorts caused by free sodium ions. The fringing electric field at the surface of the reverse-biased collector-base junction causes sodium ions to collect at the Si-SiO_2 interface. This results in inversion of the *p* base. The fringing field moves to the left in the diagram and sweeps more sodium to the interface, resulting in an effective narrowing of the base at the interface. (From Nicollian and Brews, 1982).

voids (missing aluminum) or hillocks (accumulated aluminum). The voids can coalesce, resulting in open circuits, and the hillocks can grow until they connect to adjacent aluminum, possible resulting in a short circuit. The voids and hillocks occur principally at the intersection of three metal grains.

Electromigration is a wearout mechanism that affects many devices to some extent. The severity increases with current density in the metal line and with temperature. Factors that affect the failure rate are grain structure; constrictions in the metal arising from underlying topography, mask defects, corners, or corrosion; and hard dielectric overlayers that prevent hillock growth and, therefore, prevent void formation. For example, decreasing linewidth increases the rate of electromigration since, for a given current, the current density increases. However, for lines that are narrower than the typical grain size, the electromigration is greatly decreased. This happens since the grain structure has a "bamboo" pattern in which single grain boundaries bisect the metal line, and there are, consequently, relatively few or no junctions between three grains, as shown in Fig. 4.

Large improvements in resistance to electromigration are achieved by adding a few percent of copper to the aluminum metal (Ames *et al.,* 1970). The copper tends to segregate at the grain boundaries, inhibiting the transport of aluminum.

2.2.3.2 Mechanical Stress. Aluminum has a larger coefficient of thermal expansion than the silicon substrate and the inorganic layers deposited as interlevel dielectrics. Aluminum and dielectric deposition processes are usually performed at elevated temperature. For these two reasons, aluminum interconnections are generally in tensile stress at or near room temperature. This tensile stress results in migration of aluminum atoms from the edges of the lines to grain boundaries within the line, causing voids to appear at the edges (Yue *et al.,* 1985). For the narrow lines used in advanced device structures, these voids can cause open circuits or a reduction in metal cross section that leads to early failure by electromigration. Stress-generated voids can be minimized by using metal and dielectric deposition processes that reduce stress and by avoiding overly thick dielectric layers. Stress failures can also result from polymers used in molding device packages.

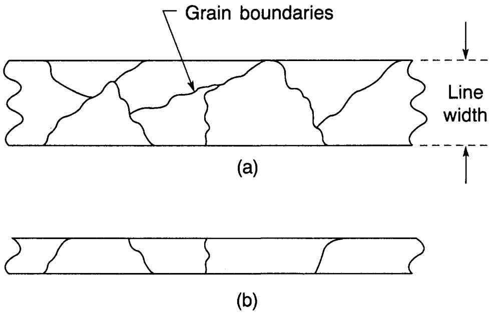

FIG. 4. Effect of grain size on electromigration. For a given aluminum grain size, a wide aluminum line (a) can have many intersections of three grains, at each of which a void can form. A narrower line (b), with the "bamboo" structure, has no intersection of three grains and electromigration is much less likely.

2.2.3.3 Corrosion. Various metals and combinations of metals are used as interconnections on monolithic and hybrid integrated circuits. The most commonly used metal is aluminum, usually alloyed with copper or silicon. Gold is sometimes used, usually in "tri-metal" combinations with a base metal, such as titanium, and a barrier metal, such as platinum or palladium. On hybrid integrated circuits, the variety of metals can be even greater and may include silver, copper, tin, lead, and others. The corrosion mechanisms depend on the metals present, contaminants, moisture, and voltage (Comizzoli *et al.,* 1986). The failure mode can be an open circuit as in the case of aluminum, or a short between interconnect lines when the soluble metal ions migrate and "plate out" at the more negative electrode to form a dendrite, as, for example, in the case of gold. In general, corrosion mechanisms are accelerated by temperature and humidity.

Dissimilar metals, e.g., gold wire bonded on aluminum pads, can corrode in the absence of an applied potential since in the presence of moisture they form an electrolytic cell with a "built-in" potential difference. This type of corrosion, called galvanic corrosion, occurs readily at bond pads in the presence of water and a contaminant such as chloride or bromide ions. Thus, plastic-packaged integrated circuits in humid environments may fail if solder fluxes containing halides penetrate to the pad area. Also, exposure of such integrated circuits to high temperature for extended time periods may decompose bromine-containing fire-retardant additives in the polymer, which subsequently results in corrosion.

Applied voltages can also cause corrosion when both electrodes are made of the same metal. Adsorbed moisture is generally required for transport of ionic species that carry current between the metal conductors. When the conductors are aluminum, both the positively biased and negatively biased conductors are susceptible to corrosion. Ionic halide contaminants destroy the protective oxide film on positively biased aluminum, exposing fresh metal that corrodes. At the negatively biased electrode, water is reduced, forming hydrogen molecules and hydroxyl ions. The latter increase the *p*H at the surface of this conductor. If the *p*H becomes sufficiently high, i.e., basic, the protective oxide and the aluminum metal will dissolve. Positively biased gold conductors are also subject to corrosion. If both adsorbed moisture and ionic halide contaminants are present, the gold conductor will corrode, forming a soluble gold halide complex. The gold complex can migrate to the negative conductor where it will be reduced to gold metal in the form of dendrites or conducting bridges that grow back toward the positive conductor. Eventually, the dendrites will short circuit the two conductors causing a failure. In the absence of halide contaminants, gold is oxidized to gold hydroxide, forming a voluminous precipitate. Failure results when this precipitate spreads across the circuit, reducing the resistance between conductors, or when the positively biased conductor has corroded through, causing an open circuit. Dendrite formation has also been observed with other metals. For example, silver dendrites have been found on ceramic chip capacitors after they were soldered using a water-soluble flux that contained halides.

Silicone polymer coatings provide excellent corrosion protection. These are applied to the chip before it is molded in epoxy. The device surface must be clean before polymer application to obtain good adhesion. If hygroscopic contamination is present at the polymer-chip interface, water absorption will occur at the contaminated interface under high humidity conditions. Corrosion of any exposed metal may then occur. Epoxy molding compounds contain some sodium and chlorine. In the late 1970s, chlorine levels of several hundred ppm were not uncommon in molding compounds. At that time, failure rates of several percent per thousand hours of 85 °C/85% relative humidity/bias were considered excellent. Now, molding compounds with chlorine contamination as low as a few tens of ppm are commercially available, and failure rates below 1% have been obtained.

Passivation layers (insulating, usually inorganic, deposited films) over the metal in nonhermetic integrated circuits greatly reduce the risk of corrosion. However, the passivation layer must be pinhole- and crack-free and adhere well in the presence of moisture and upon exposure to temperature cycling. The use of phosphorus in silicon dioxide passivation layers is beneficial on several counts, but the phosphorus content must be carefully controlled to be below 4 wt % for undensified films, to avoid leaching of phosphoric acid in the presence of moisture. This leaching has resulted in extensive corrosion of aluminum of disastrous proportions.

Corrosion rate generally increases with voltage for a given device geometry (see later discussion in Sec. 2.3.3). Power dissipation that raises chip temperature above ambient tends to reduce the incidence of corrosion, since the local relative humidity at the chip surface will be less than ambient relative humidity. Effective moisture control in hermetically packaged devices is an important factor in corrosion protection, particularly when power dissipation is low or the device operates in a reduced-temperature ambient.

2.2.4 Insulating-Layer Mechanisms. The complexity of modern semiconductor devices, hybrid integrated circuits, and advanced packaging technologies requires the use of many types of insulating layers for numerous applications such as multilevel metallization structures. For silicon integrated circuits, a thin film of silicon dioxide (SiO_2) provides the insulation layer between the gate and channel of metal oxide semiconductor (MOS) devices. An oxide layer protects the junction of bipolar circuits. Insulator films are used to isolate the different levels of metallization and also to planarize the structure for subsequent film deposition. Other insulator layers may be needed to provide stabilization or passivation, to increase breakdown voltage, or to serve as the active structure in memory devices such as metal-nitride-oxide semiconductor or floating-gate

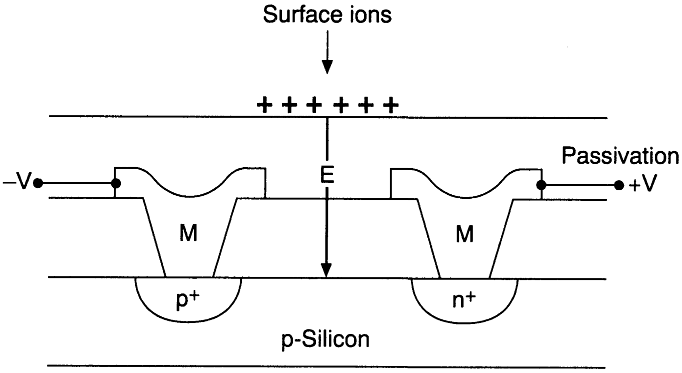

FIG. 5. Electric field in silicon caused by surface charge on insulators. Schematic cross section of an idealized diode showing the vertical electric field resulting from a net surface charge on the insulator of the device. *M*, metal; *E*, electric field vector; *V*, potential.

avalanche-effect integrated circuits. These insulating layers may include thermally grown SiO_2, SiO_2 formed by chemical vapor deposition (CVD), phosphosilicate and borophosphosilicate glass from CVD processes, silicon nitride from high-temperature CVD or low-temperature plasma-assisted CVD methods, glass layers formed by glass particle deposition and fusion, ceramics, and various polymers. Various failure mechanisms are related to these insulating layers.

2.2.4.1 Surface-Ion Failure Mechanism. Failures by this mechanism still occur, particularly in high-voltage devices, even though it was discovered and understood quite early (Atalla *et al.*, 1959).

Surface ion motion on the outermost insulating layer of a device can result in local accumulations of net ionic charge. This net charge induces a vertical electric field in the insulator and at the semiconductor surface, as shown in Fig. 5. The electric field, if large enough, can cause inversion or accumulation of the semiconductor free charge density, leading usually to increased leakage current; or it results in localized electrical breakdown in the semiconductor and reduces the device breakdown voltage below rated values.

The surface ion motion takes place on the outer layer of the device, which is normally the passivation layer in the case of an integrated circuit. A slight surface conductance, such as results when an insulator is exposed to water vapor, is sufficient to produce charge polarization. For example, at exposed metal, as exists at a bonding pad or passivation pinhole, the electric field parallel to the surface will cause ions of the same sign as that of the metal to drift away from it, as shown in Fig. 6. This ion drift tends to charge the insulator surface to the same potential as the metal, provided that the surface conductance of the insulator is much greater than its bulk conductance. By this surface-ion drift mechanism, lateral propagation of potentials to a region can result in the potential of the insulator surface being at one extreme of the potentials applied to the device, while the potential of the semiconductor (silicon) below that region is at the other extreme. In this case, the electric field in the insulator and at the silicon surface can be sufficiently large to induce the effects described above (leakage or breakdown voltage reduction). Direct deposition of charge on a device from a high-voltage source is a useful technique for studying this failure mechanism (Comizzoli and Osenbach, 1988).

A traditional method for avoiding surface-ion–induced failures is the use of a metal field plate to fix the surface potential above sensitive regions at a safe value. The distance between high-voltage nodes can also be increased within certain limits to reduce the electric field in the silicon. Oxide thickness

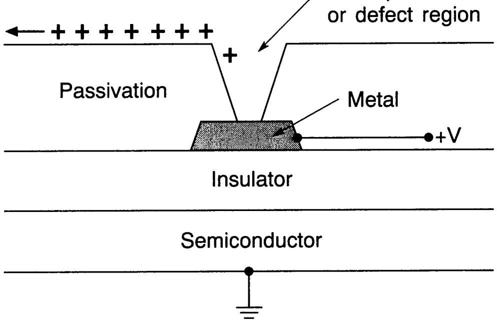

FIG. 6. Schematic cross section illustrating ion motion on the surface of an insulator near exposed metal.

can be increased, also, to reduce the electric field associated with the surface ions.

2.2.4.2 Charge Accumulation in Multilevel Insulators. Net charge can accumulate at the interfaces between insulating layers (Woods, 1974). In general, each insulator layer in a stacked structure has a different conductivity and dielectric constant. The conductivity results from ionic motion, electronic motion, or some combination of these. Initially, when voltage is applied across the insulator layers, the voltage is divided capacitively among each of the layers. Thus, the initial electric field in each layer depends on the dielectric constants and thicknesses of all of the layers. This electric field results in a current in each insulator that, in general, is not equal to the current in the other insulators. Since charge must be conserved, accumulation of net charge occurs at interfaces between insulators with unequal currents. As time progresses, the insulator system approaches a steady state in which the charge at each interface and the electric field in each insulator have attained values such that the currents resulting from the electric fields are continuous across the interfaces.

Another mechanism that can result in interface charge formation is a polarization of one of the insulators. In the applied electric field, dipoles in the insulator orient themselves parallel to the electric field. The effect is the same as that of drift of ionic or electronic charge.

If the result of the interface charge generation is an increase in the electric field at the semiconductor surface, then device operation may be degraded. The semiconductor may become inverted, resulting in excess leakage current and failure. In a high-voltage device, the excess field may result in electrical breakdown in the semiconductor at applied voltage levels below those specified. This failure mechanism can be prevented by considering the insulators' electrical properties and device design. Various structures such as field plates or shields can be used in sensitive regions.

2.2.4.3 Electrical Breakdown of Silicon Dioxide. The electrical breakdown (loss of insulation properties) of a thin insulator such as silicon dioxide (SiO_2) is one of the most difficult failure processes to understand since it is sudden and very rapid, and in general destroys the affected area by the heat generated in the breakdown event. The breakdown occurs at a highly localized region and can be induced by exceeding the breakdown electric field (field breakdown) or by maintaining a constant field for a certain time after which the insulator breaks down (time-dependent breakdown).

Much effort has been devoted to understanding breakdown effects in SiO_2. Breakdown can occur at defect sites, where the breakdown field has a wide range of values, or breakdown can be intrinsic with a value of about 1×10^7 V/cm for defect-free SiO_2. Defects that reduce breakdown voltage below the intrinsic value include thin spots, rough contours in the silicon that concentrate the electric field, and foreign material such as sodium ions, metals, and water.

Various models have been proposed to explain the time, temperature, and field dependence of SiO_2 breakdown. A model that successfully explains a wide variety of observations is the gas discharge model (Wolters and Verwey, 1986). The breakdown process in this model begins with injection of electronic charge into the oxide during the time that bias is applied. The fraction of this charge that is trapped in the oxide leads to a local increase in the electric field. When the local field is high enough, some of the trapped charge is released energetically, generating a microvoid by disruption of the SiO_2 network. In the microvoid, electrons attain much higher energies in the field so that more disruption of the structure takes place. Eventually the void structures grow and link up until a void network is formed between the electrodes. This is the breakdown event in which a surge of current passes through the SiO_2.

Various techniques are used in processing to achieve high breakdown strength. These include clean processing, growth and removal of sacrificial oxides, and complex growth and annealing processes for the oxides. Chlorine is an effective impurity getter during oxide growth and contributes to increased breakdown strength.

2.2.5 Electrical Overstress

2.2.5.1 Electrical Latch-Up. Latch-up is a purely electrical phenomenon that principally affects complementary metal oxide semiconductor (CMOS) integrated circuits. It can occur whenever four regions of alternating conductivity type, such as *npnp* or *pnpn*, exist in a structure (Schnable, 1989). In CMOS circuits, such regions occur commonly.

The series connection of *npnp* or *pnpn* forms a parasitic thyristor. This thyristor structure is not intended as part of the design, but nevertheless exists, hence the term parasitic. Thyristors switch between two states, blocking or conducting. Very little current flows across a thyristor until a certain voltage is exceeded. The device then turns on, and a large current flows at low voltage; thus the term "latch up." The parasitic thyristor functions in the same way. Thus, an overvoltage in an integrated circuit can trigger the thyristor action, and large currents will flow through the parasitic region. While the current is flowing, the device function is disrupted. If the current source has a low enough impedance, the integrated circuit may be destroyed. The thyristor action continues until the device is destroyed or the circuit is interrupted. In addition to a high voltage, latch-up can be induced by any excitation that increases the conduction carrier density in the parasitic region, such as high temperature or radiation.

2.2.5.2 Electrostatic Discharge Effects. Failures caused by electrostatic discharges (ESD) (Amerasekera and Campbell, 1987) typically occur when an electrically charged person comes into electrical contact with a device. A human being can acquire an electrical charge by contact or friction with a material such as carpeting, floor tiles, plastic seating, etc. If contact is made to a device, the charge is transferred to it. In the transfer, the device experiences a voltage surge that can cause electrical breakdown of insulators or junctions.

The approximate capacitance of a human body is 150 pF, and normal body potentials of several thousand volts can be expected from, for example, walking across a room. Under dry conditions, body potentials of more than 20,000 V are not uncommon after walking over a wool rug. If a person then touches electronic equipment, this accumulated charge is transferred to the equipment. The charge can induce large voltages on certain circuit elements, resulting in electrical breakdown of insulators or semiconductor junctions.

These electrostatic discharge effects are most severe in the production process for integrated circuits, in which the devices are exposed and can be contacted by a highly charged person. After components have been packaged, mounted on a circuit board, and assembled into an enclosure, the possibility of damage by electrostatic discharge is much smaller but can be significant if precautions are not taken.

Methods of avoiding damage include techniques to minimize charging of the human and design modifications to "harden" devices against electrostatic discharge. The former include various kinds of static-free and static-dissipative clothing and grounding straps, conductive sprays, air ionizers to discharge surfaces, and conductive floors and mats. Design modifications involve components added as part of the integrated circuit to reduce the effects of current and voltage spikes and transients. These protective circuits contain combinations of resistors and voltage-sensitive diodes or transistors. The function of the voltage-sensitive components is to switch "on," from a high-resistance state to a conducting state, when a voltage spike is applied. The electrostatic discharge voltage and current spike flow through the protective device and not through sensitive parts of the integrated circuit. Another protective device that can be fabricated as an integral part of an integrated circuit is the spark gap. This consists of a plane array of metal points with a short gap between them. When a voltage spike is applied, breakdown of the ambient gas above these points shorts out the pulse before it can damage the integrated circuit.

2.2.6 Radiation Effects. The effects of ionizing radiation are particularly important in satellite or military applications. There are both immediate and long-term effects on devices (Braunig and Fahrner, 1986). For example, in a nuclear blast, an intense radiation flux immediately passes through an integrated circuit, generating a high density of charge carriers in insulators and semiconductors alike. For a time, the device is nonfunctional since all parts of it are conducting. Long-term effects occur even for the levels of radiation occurring naturally in outer space. For example, high levels of proton or electron flux, depending on altitude, are found in the Van Allen belts. This radiation can increase the charge at the Si-SiO$_2$ interface and cause failures in metal oxide semiconductor (MOS) devices. Crystal structure damage can also occur as a result of particle flux, which can change the gain of bipolar resistors.

Radiation associated with metals and ceramics used in integrated circuit packaging can cause "soft" or "single-event upset" errors. α particles are emitted by certain nuclei in, for example, alumina device packages. An α particle generates a large number of electron-hole pairs when it passes through silicon. If the path is through a memory cell in a dynamic random access memory (DRAM), the generated carriers result in sufficient conductivity to change the state of the memory cell. These effects are eliminated by carefully selecting packaging materials and by using an α-particle shield over the surface of the integrated circuit.

2.3 Enclosures For Integrated Circuits

Once an integrated circuit has been manufactured it must be surrounded by a package or enclosure that provides some degree of protection from the environment and a means to interconnect the functional capability of the chip to the rest of the electronic system. It is essential that the materials used to achieve protection from environmental contaminants do not contribute harmful contamination at the chip surface. These subjects have been discussed in more detail by Comizzoli *et al.* (1986). Before considering the various methods for enclosure, the environmental constituents of concern will be discussed.

2.3.1 Environmental Factors and Manufacturing Contamination. Because of the importance of environmental effects, the constituents that cause corrosion and short circuits in electrical devices need to be considered when reliability tests are planned (Sinclair, 1988). These tests should take into account the severity of the environment that the product will see. The constituents of greatest concern are water-soluble ionic contaminants. These are ubiquitous components of the environment and include sea-salt particles, particles derived from fossil-fuel combustion or the incineration of municipal waste, commercial and residential fertilizers, soil dust, and salts used to melt ice and snow. The indoor concentrations of these substances typically range from 10% to 80% of their outdoor concentrations, with office environments usually at the low end of the range and factories at the high end.

Particles in the environment exist in two distinct size regions:

Coarse particles ($>2\ \mu$m diameter) consist largely of natural materials derived from soil and rocks, especially in the outdoor environment. In the manufacturing environment, a variable fraction of these particles may be process related. Air-filtration systems in office buildings and manufacturing environments are typically (30–95)% efficient in removing coarse particles from the incoming air.

Fine particles ($<2\ \mu$m diameter) typically result from fossil-fuel combustion, but they may also be process-related in manufacturing environments. Because combustion-derived particles contain larger concentrations of water-soluble ionic substances, they tend to be more corrosive than natural particles. Air-filtration systems are less effective in removing fine particles than coarse particles.

Sulfate, typically as ammonium acid sulfate, is usually the most abundant anion in fine particles indoors and outdoors. Indoor concentrations of sulfate are typically (40–75)% of the outdoor concentrations and correlate over time with the outdoor levels.

Ionic contaminants are also present in processing chemicals, which include electroplating salts, gases or liquids used for wet or dry etches, solder fluxes, acids and bases, and some cleaning solutions. These processing sources can cause contamination by direct contact. They can also raise the airborne concentrations of specific ionic compounds to levels several times higher than outdoor concentrations. These airborne contaminants may be transferred by air currents to sensitive electronic components in other parts of the surroundings.

In the presence of sufficient humidity, ionic contaminants acquire moisture and form thin films of electrically conducting solutions. The result may be enough current leakage to cause crosstalk (extraneous signals) or soft errors. The same types of moisture-related failure processes described previously for integrated circuits also apply to circuit boards and hybrid integrated circuits. When high electric fields are present, component-damaging electrical arcs can also occur.

Ionic contaminants may be deposited on sensitive surfaces during manufacture before protective coatings are applied. These contaminants, combined with exposure to atmospheric moisture, allow leakage currents to flow under the coating. Ionic contamination deposited in the field allows leakage currents to flow over the top surface of the coating between exposed bond pads or leads.

The increasing use of forced-air cooling to carry away the heat generated by high–power-density circuits can greatly accelerate the deposition of atmospheric dust particles if enhanced filtration is not provided on the incoming air stream.

2.3.2 Package Types, Wire Bonding, and Chip Mounting. The three most predominant types of packaging (plastic packages, ceramic packages, and flip-chip encapsulated devices) and the methods of connecting device leads to substrate metallizations are depicted in their simplest forms in Fig. 7–10. The metal leads extending from the plastic and ceramic (hermetic) packages in Figs. 7 and 8 are one of three common designs for connecting the package to a circuit board or hybrid integrated circuit. In this case the leads are placed in "via" holes drilled in the substrate and then soldered permanently in place, as shown in Fig. 9. In another method the leads may be plugged into a "socket," which has previously been attached to the substrate through a soldered interconnection. A third method is to "surface-mount" the package directly to the circuit board. This method of attachment is also depicted in Fig. 9. In this case the leads extending from the package are oriented parallel to the substrate surface and are connected to "bond pads" on the mounting surface, again using solder.

For plastic or ceramic packages, the silicon chip inside the package must be connected to the leads that extend out of the package. This connection is usually accomplished through gold wire bonds that are attached to the aluminum metallization of the silicon chip and to the lead metal (usually copper) by ultrasonic or thermocompression bonding. A wire bond connection is shown in Fig. 7.

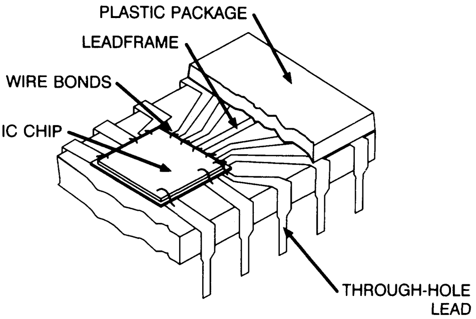

FIG. 7. Plastic package.

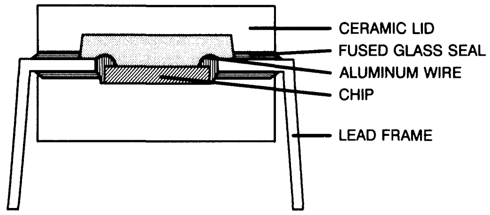

FIG. 8. Ceramic package.

In the case of flip-chip encapsulated devices, bond pads on the active side of the silicon chip are placed directly over solder bumps on bond pads on the substrate. Attachment takes place when the solder bumps are reflowed. The attached chip is then protected from the environment by encapsulation with a rubbery polymer (elastomer), which is usually a silicone. This method is illustrated in Fig. 10.

2.3.3 Plastic Packages and Flip-Chip Encapsulated Devices. Reliability issues that must be considered in the use of plastic packages and flip-chip encapsulated devices include chemical and electrochemical degradation, e.g., corrosion processes caused by moisture and other contaminants, and me-

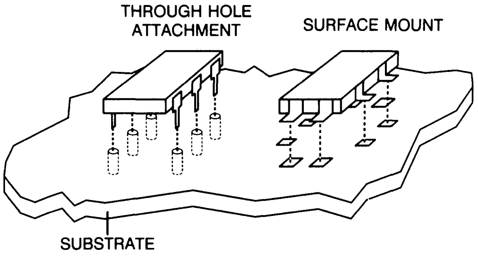

FIG. 9. Comparison of through-hole and surface-mount attachment. For through-hole attachment, the leads from the device are placed in holes that are previously drilled in the substrate. A permanent connection is made by exposing the underside of the board to molten solder. For surface-mount attachment, solder paste is applied to bond pads on the substrate, the device is then positioned on the substrate so that the bond pads and leads align, and the solder paste is flowed by hot air or infrared heating.

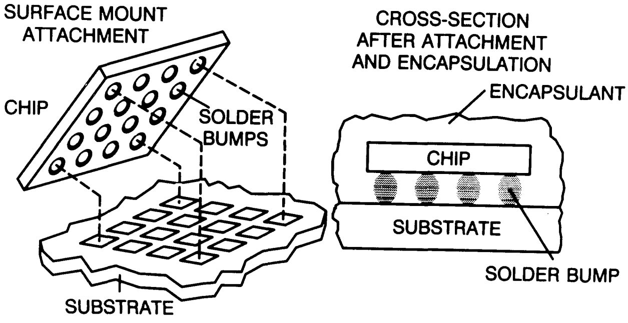
SURFACE MOUNT ATTACHMENT
CHIP
SOLDER BUMPS
SUBSTRATE
CROSS-SECTION AFTER ATTACHMENT AND ENCAPSULATION
ENCAPSULANT
CHIP
SUBSTRATE
SOLDER BUMP

FIG. 10. Flip-chip technology. The left part shows the process for attaching a chip, with bond pads on its active side facing the bond pads on the substrate. The right part gives an exaggerated view of the solder bumps connecting the chip to the substrate.

chanical stresses caused by temperature fluctuations, vibration, or shock.

Most chemical and electrochemical degradation processes involving contamination require the presence of moisture. Prevention of ionic contamination and moisture condensation inside the package is the most important function of the plastic package or encapsulation surrounding the chip. All polymeric materials that are used to protect chips are permeable to moisture to some degree. The best polymers available for electronic applications (polyimides and Novolac epoxies) typically acquire (0.1–3)% of their dry weight in moisture at saturation. While equilibration with the external humidity level is rapid (hours to days), condensation of moisture is prevented as long as the polymer tightly adheres to the active surface of the chip. Moisture-related failures are usually associated with loss of polymer adhesion to metal conductors. Ionic contamination is frequently found at the ruptured interface, though the contamination is not necessarily the cause of adhesion loss or device failure. Once delamination occurs and moisture condenses on exposed circuit elements, the electrical bias in typical integrated circuits is more than sufficient to initiate electrolytic corrosion, as illustrated in Fig. 11. The electrical malfunction associated with electrolytic corrosion can take one of several forms including leakage currents due to ion current in the condensed moisture film, electrical open circuits caused by consumption of the conductor, and electrical shorts caused by metal bridges between conductors. These are illustrated schematically in Fig. 12. The susceptibility of devices to failure by these processes increases dramatically as the minimum conductor linewidth decreases for integrated

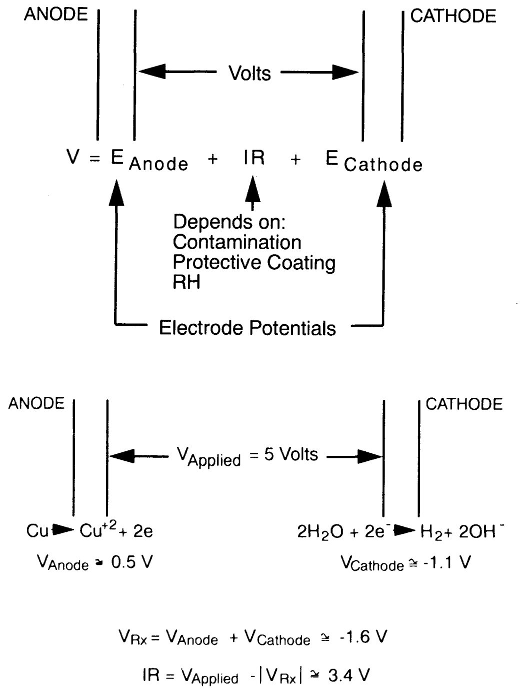
ANODE
CATHODE
Volts
V = E Anode + IR + E Cathode
Depends on:
Contamination
Protective Coating
RH
Electrode Potentials
ANODE
CATHODE
V Applied = 5 Volts
Cu ► Cu+2 + 2e
V Anode ≈ 0.5 V
2H2O + 2e- ► H2 + 2OH-
V Cathode ≅ -1.1 V
V Rx = V Anode + V Cathode ≅ -1.6 V
IR = V Applied - | V Rx | ≅ 3.4 V

FIG. 11. Voltage dissipation between conductors on a substrate. Top: illustration of how the voltage applied across conductors is dissipated when corrosion occurs. E_{anode} and $E_{cathode}$ are the electrode potentials, which take into account the concentration-mediated thermodynamic potential for the reaction at each electrode and kinetic factors. All the remaining voltage between the conductors must be dissipated in the *I-R* drop (*I* is current and *R* is resistance). Corrosion is minimized by maximizing *R*, which is accomplished by lowering relative humidity, minimizing contamination, and use of high-resistance coatings. Bottom: illustration showing that a large fraction of the voltage applied to an integrated circuit (5 V) is dissipated in the *I-R* drop, illustrating the importance of high insulation resistance. The electrode potentials used for this illustration are rough approximations.

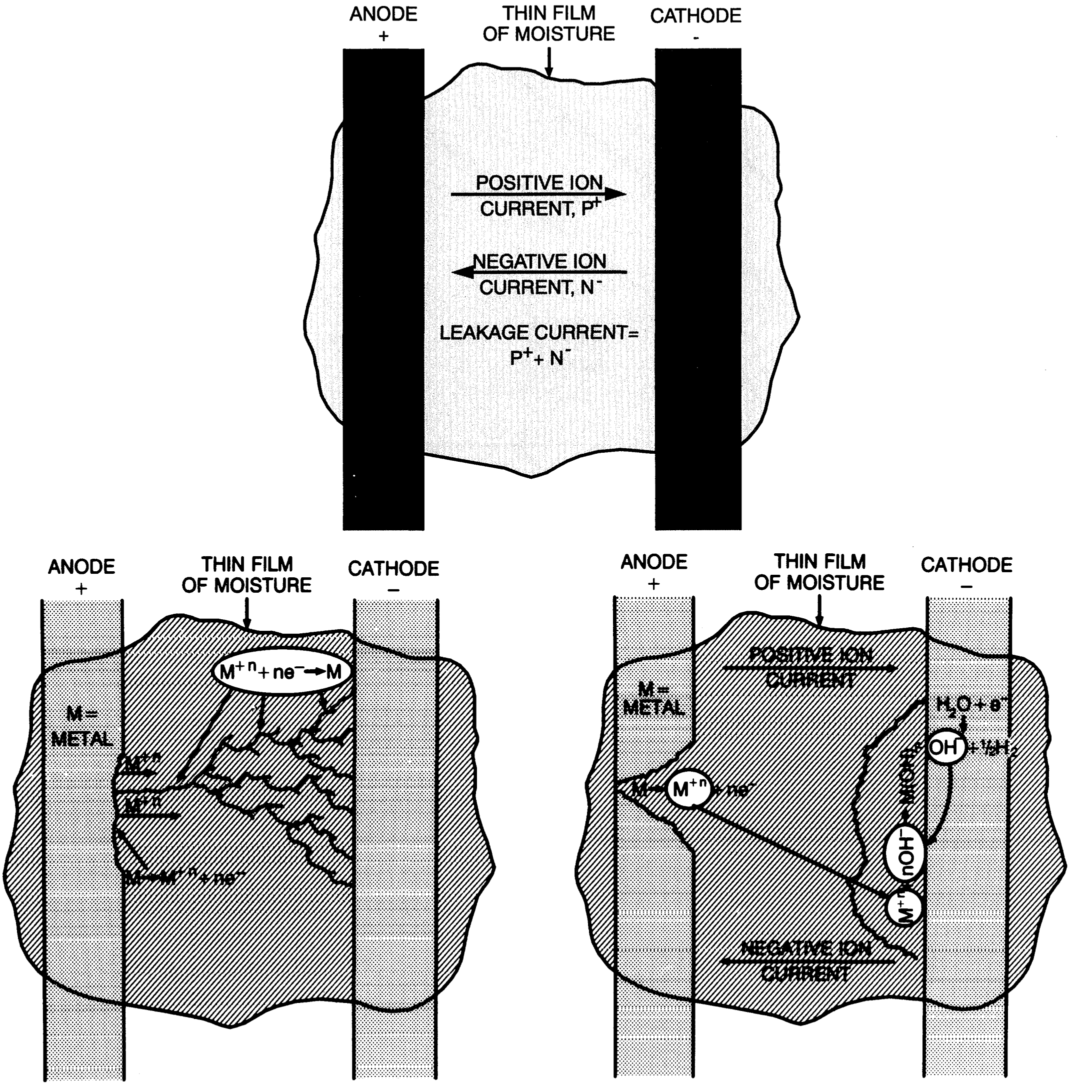
ANODE +
THIN FILM OF MOISTURE
CATHODE -
POSITIVE ION CURRENT, P+
NEGATIVE ION CURRENT, N-
LEAKAGE CURRENT= P++ N-
ANODE +
THIN FILM OF MOISTURE
CATHODE –
M= METAL
M+n+ ne-→M
M+n
M+n
M→M+n+ ne-
ANODE +
THIN FILM OF MOISTURE
CATHODE –
M= METAL
POSITIVE ION CURRENT
M→M+n+ ne-
H2O+ e-
OH-+½H2
M(OH)n
nOH-
M+n
NEGATIVE ION CURRENT

FIG. 12. Moisture/contamination failures on substrates. The three most common types of failures that can occur when contamination and moisture allow leakage currents to flow. (a) The simplest failure, a shift of the electrical parameters of a circuit out of specification. The shift is caused by the current flow associated with migration of positive and negative ions from contaminants. (b) The leakage current associated with ion migration, which may not be sufficient to cause a failure by itself, but may lead to oxidation of the anodic conductor with the formation of metal ions. These ions can migrate to the cathode. Upon reaching the cathode, the ions are reduced to the metallic state. As this process proceeds, dendritic filaments of metal grow back toward the anode. Eventually the gap between conductors is bridged, forming a short circuit. (c) Similar to (b), in that metal ions oxidized at the anode migrate to the cathode. In some cases, the important reaction at the cathode is the reduction of water to form hydroxyl ions. The hydroxyl ions and metal ions react to form an insoluble precipitate, usually near the cathode. An open circuit eventually develops at the anode when sufficient metal has oxidized to open up the conductor line. The time to failure in this case is a direct function of the cross-sectional area of the conductor and the leakage current (see Fig. 13).

circuits, the interconnecting conductors in the packages that contain them, and the printed wiring on the circuit boards on which they are mounted. This trend is illustrated in Fig. 13 for electrical open circuits caused by consumption of conductors. This figure shows that the maximum tolerable leakage current per corrosion site for a 10-year lifetime, assuming

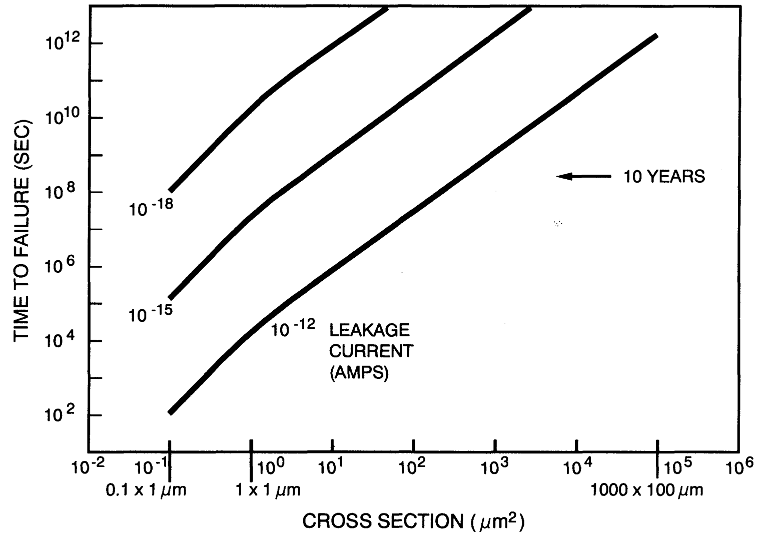

FIG. 13. Time to failure versus cross section. The time to failure, when electrical leakage currents that flow between conductors lead to corrosion of the anode or cathode, depends on the current and the cross-sectional area of the conductor. For circuit-board conductors with typical linewidths of 100–1000 μm, the time to failure is > 10 years for a leakage current of 10^{-12} A per corrosion site, which is the optimum level typically achieved for circuit boards in ordinary environments. However, for integrated circuits with linewidths of 0.1 μm, which are expected to be commercially available around the year 2000, a 10-year lifetime requires leakage currents of no more than 10^{-18} A per corrosion site.

that the current efficiency for corrosion of metal is 100%, decreases by more than 6 orders of magnitude as feature size changes from 150 to 0.5 μm (typical of circuit boards and integrated circuits, respectively, in the early 1990s).

The leakage current across insulators typically shows an exponential dependence on relative humidity. The concept of a critical relative humidity below which moisture-related failures do not occur follows from the exponential dependence. The critical relative humidity is usually in the range of (55–70)%, which roughly corresponds to the humidity at which ionic pollutants associated with fossil-fuel combustion show a large increase in hygroscopicity (moisture pickup capacity) as humidity increases. None of the failure processes ascribed to electrolytic corrosion will occur in the absence of moisture. As discussed, prevention requires highly adherent polymer coatings or packages. Controlling ionic contamination on chip surfaces well below 10^{13} ions/cm^2, and controlling bulk ionic contamination in polymers to < 5 ppm, also minimize the risk of failure.

The appropriate testing strategy is to expose the chip package to moisture stresses and ionic contamination. Three different thermal/moisture stress tests are in common use. The steam bomb test conducted at temperatures > 100 °C is the harshest but is used infrequently because it tends to produce some defects that are not observed in the field. A gentler test exposes chip packages to 85 °C/85% relative humidity with imposed electrical bias in the range of 5–30 V. The test is usually run for 1000–2000 h. Electrical testing of the devices is frequently conducted *in situ*. While the validity of this test in duplicating field experience is rarely established conclusively, years of experience have shown that devices that survive this test have a high probability of not failing in the field. In recent years, the highly accelerated stress test (HAST) has sometimes been used when rapid reliability data are needed by designers and process engineers. The typical conditions for this test are 100–160 °C and (50–100)% relative humidity, with or without bias. The duration is usually 24–48 h. The acceleration factor for a test at 120 °C/95% relative humidity is typically 10–20 times that for the 85 °C/85% relative humidity test.

Strategies to prevent moisture-related problems are not limited to providing impenetrable packages or highly adherent coatings. In some cases the humidity in the use environment can be controlled at levels below which degradation is largely eliminated. Many computer installations require that the relative humidity be kept under (50–60)% for just this reason. Other strategies include selecting metal conductors that are less susceptible to corrosion (gold is better than copper or tin-lead alloys in many situations) or designing circuitry to minimize electric fields. As feature size decreases, design engineers consider all of these options when high reliability is required.

With surface-mount plastic packages and flip-chip technology, moisture effects are not limited to electrochemical phenomena. Absorption of moisture by encapsulants can also affect the mechanical integrity of devices, depending on stresses at the encapsulant-device interface. New polymers with at least an order of magnitude lower capacity for moisture absorption will be required to eliminate the need to ensure that packages are completely dry when undergoing solder attachment. Stress concentration can cause cracking or shearing of metal conductors or wire bonds on devices or delamination of metals from insulators. Sudden temperature changes, such as those experienced during solder reflow, can cause cracking of plastics and metals due to pressure buildup at interface voids associated with rapid moisture evaporation. Solder joints can also be fractured by combined moisture/thermal effects, especially when attachment is through an array of solder joints.

Prevention of these degradation processes requires selection of plastics that are screened to be mechanically stable over a wide range of relative humidity; design of packages, solder interconnect schemes, and encapsulants to minimize stress concentration, particularly at solder joints; and environmental stress testing that combines cyclical application of thermal, mechanical, and humidity stresses to the packaged devices.

2.3.4 Ceramic Packages. When the highest level of reliability is required (military and space applications, implanted medical devices, undersea transmission systems for telecommunications), integrated circuit chips are usually enclosed in hermetic ceramic packages. These packages are typically several times more expensive than plastic packages. In recent years many reliability engineers, looking for ways to reduce cost, have questioned the need for hermeticity. As the concentration of ionic impurities in polymers continues to be lowered, as adhesion is improved, and as molding engineers continue learning how to manage stress in packages, reliability of plastic packages has improved. The need for ceramic packages may eventually be eliminated. Until then, designers need to be aware of the reliability issues for ceramic packages. The sealed character of ceramic packages ensures that there will be no influence on the device from the external environment, contaminants, or moisture, but it also ensures that any contamination or moisture contained within the device will not escape. As long as the cavity is completely dry and free of contamination and all of the materials within the cavity are free of trapped moisture that can outgas at elevated temperature, reliable operation can be expected. If moisture outgasses at elevated temperature and subsequently condenses on device surfaces at cooler temperatures, the same electrochemical phenomena described for plastic packages can occur. The appropriate screening test is a cyclical thermal test (−40 °C to 100 °C) on an operational device. Various screening tests are employed to evaluate the seal integrity, such as dye penetration, observation of escaping gas as bubbles in a liquid, and pressurization with radioactive tracer gas. Most hermetic devices are sealed in an inert-gas atmosphere. Mass spectrometry or weight change can be used to monitor the escape of the inert gas. Attention to proper manufacturing procedures eliminates most reliability problems with ceramic packages.

Surface attachment of ceramic packages with leads mounted through via holes in substrates can be accomplished with high reliability using commonly practiced assembly methods. For surface-mount attachment, mechanical stresses on solder joints and on ceramic materials can be severe. Temperature/time profiles for the substrate and package must be carefully examined to minimize residual stress when the package and substrate return to room temperature following attachment. Since solder is well known to slowly creep and recrystallize to relieve stress (even at room temperature), a surface-mount process for ceramic packages should be subjected to static and cyclic thermal stress and mechanical loads to establish the design reliability.

2.4 Assemblies

2.4.1 Interconnection of Components and Integrated Circuits

2.4.1.1 Circuit Boards. The most common type of circuit board consists of multiple layers of metal networks that are separated by a dielectric material and interconnected by metal-filled holes in the dielectric layer. The dielectric material is typically a composite of epoxy and layered sheets of woven glass fibers. Copper metallization layers are typically produced by laminating a thin foil of copper on the epoxy-glass substrate, then patterning the desired metal network by photolithographic methods, and finally thickening the metal conductor lines by electroplating copper. For many applications, ten or more of these layers are laminated together. For less demanding applications, only one epoxy/glass layer may be required, which may have the metallization printed on both sides. For single-layer boards, phenol fiber or molded plastics are also sometimes used in place of the epoxy glass. In some applications the molded plastic may be sufficiently thin to enable the circuit board to be flexed without causing damage. For most types of circuit boards, the final manufacturing step is to coat the copper with solder. An appropriately patterned solder mask covercoat is applied to the board, which is then dipped into molten solder. The solder overcoat ensures subsequent solderability of components to the board. Without this coating, the natural oxide that grows on copper would inhibit soldering. In some cases, a corrosion inhibitor that adsorbs onto the copper surface, such as benzotriazole, is used instead of solder to preserve solderability. At this point the circuit board is ready for component and connector attachment, which is usually accomplished by a soldering process, though metal-filled conductive adhesives may be used for low-cost products. In the early 1990s, the typical metal linewidth is 100–200 μm with state of the art being 25 μm for a 40-layer board. The linewidths will necessarily decrease as feature sizes for integrated circuits decrease.

Several potential failure mechanisms must be considered in evaluating the reliability of circuit boards. These are listed in Table 1. Here only multilayered epoxy/glass circuit boards will be considered, but many of the issues are common to all types. Also listed in the table are appropriate stresses, specific stress tests for bringing out these failure processes, and special considerations.

In the manufacturing process for multilayer boards, two critical issues with respect to achieving high yield and long-term reliability are the adhesion and the accuracy of registration between layers. Thermal, vibration, and shock stresses are effective in uncovering defects associated with inadequate adhesion or poor registration. Another process issue with all types of circuit boards is ionic contamination that may be derived from electroplating or solder flux residues, or from exposure to gaseous and particle contamination in the environment. During manufacturing, cleaning steps remove harmful ionic contaminants at sensitive stages of the process. A commonly used specification for cleanliness of circuit boards is that the residual water-soluble ionic residue must not exceed 1 μg/cm^2 sodium chloride or its equivalent in ionic contamination. Contamination can also occur as a discrete event or continuously during the life of the product in the field. Because of the importance of environmental effects, the environmental constituents that cause corrosion and electrical defects in devices should be considered in planning of environmental tests.

2.4.1.2 Failure Processes and Stress Tests. The typical failure process caused by contamination is corrosion of conductor lines, leading to open or short circuits. The electrical leakage currents associated with the corrosion process may also cause failure due to device parameter shifts in components or devices. In either case, the high surface conductance brought about by the combined effects of contamination and surface moisture is responsible for the failure.

Humidity coupled with electrical bias is an effective stress for uncovering these failure processes. The 85 °C/85% relatively humidity/bias test used for accelerated testing of device packages or encapsulation is also used for circuit-board and hybrid integrated-circuit evaluations. The electrical bias used to test circuit boards is usually much higher (100–300 V).

Since it is inevitable that contamination will occur over the life of the product, it is essential that the circuit board be protected with a reliable covercoat. In many cases the

Table 1. Failure processes and stress tests for circuit boards.

Degradation process	Stresses causing failure	Typical tests	Special considerations
Cracks in metal caused by differential thermal expansion of metals and insulators	Thermal	Thermal cycling (−40 °C to 100 °C), thermal shock, high temperature (> 100 °C)	
Corrosion of metallization leading to electrical open circuits in conductors and reduced insulation resistance or electrical shorts between conductors	Humidity with electrical bias, thermal, ionic contamination from particles, and corrosive gases	Exposure to elevated relative humidity (> 85%), elevated temperature (85 °C), bias, and possibly corrosive gas, e.g., $(10^{-5}$–$10^{-4})$% sulfur dioxide, nitrogen oxides, hydrogen sulfide, or hydrochloric acid, and corrosive particles, e.g., 50–100 μm/m^3 of test dust applied by forced airflow	At sufficiently high humidity, no contamination is required; contamination enhances failures by increasing the amount of surface moisture and the surface conductance
Fracture or displacement of materials caused by mechanical factors	Vibration, shock	Exposure to vibrations with variable amplitude and frequency, centrifuge (constant acceleration), thermal shock, dropping, sudden acceleration with acoustic monitoring	
Fracture of solder joints as a result of creep or fatigue	Thermal shock, mechanical stress	Thermal cycling, thermal shock, cyclic or static mechanical load	

solder mask may be a suitable covercoat. The solder mask is applied prior to component and connector attachment. For some applications it may be necessary to apply a supplemental covercoat after all solder connections are made. If the solder mask or other covercoat on a circuit board has defects such as pinholes, a leakage current may flow between the normally covered components or metal lines, which are now under a defect, and a nearby metallic conductor. Sometimes a component-damaging arc may occur (see Fig. 14). Severe arcing can pyrolyze circuit boards to form conducting carbon bridges.

Whatever method is used to provide protection from contamination, the susceptibility of the fully assembled circuit board to failure should be evaluated by a combination of corrosive-gas tests, corrosive-particle tests, and humidity/bias tests. An example is shown in Fig. 15. A synthetic dust representative of particulate pollutants is applied onto a test circuit that has been previously protected with an covercoat. A drop of water is applied, and an electrode is inserted in the water. A variable electrical bias is applied between the electrode and the conductor lines on the test structure, and the breakdown voltage is recorded. A faulty covercoat with pinholes or low dielectric strength will show low and variable breakdown voltages, while a high-quality covercoat will only break down at a voltage well in excess of any that might be encountered in the field, as illustrated in the figure.

2.4.2 Interconnection of Circuit Boards

2.4.2.1 Connectors and Backplanes. Connection of the circuit board to the rest of an electronic assembly is accomplished through plug-in and receptacle connectors. One member of the connector pair is attached to the circuit board, and its mate is part of the "backplane" network that enables tens to thousands of circuit boards to be interconnected. For small computers, the backplane is com-

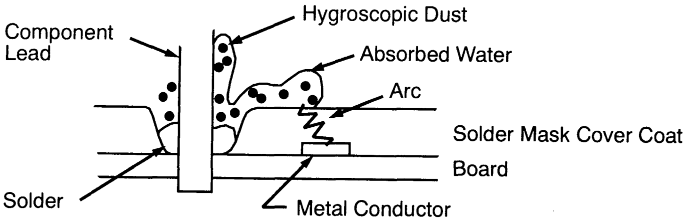

FIG. 14. Arcing failure mechanism. In this failure process, contamination that accumulates on the covercoat absorbs sufficient moisture above its critical relative humidity to extend the voltage present at the component lead over the covercoat. If a defect exists in the covercoat above a grounded metal conductor on the substrate, an arc may occur. The resulting high-voltage pulse may burn out devices anywhere in the circuit.

monly referred to as the mother board. For larger systems it can be a large array of connectors that are themselves interconnected by wires or elaborate printed wiring networks much like those on circuit boards.

2.4.2.2 Failure Processes and Stress Tests. Backplanes are susceptible to the same kinds of failure processes as circuit boards. However, the connectors require special consideration. It should also be noted that components are sometimes attached to circuit boards through standard plug-in connections rather than by solder attachment. Plug-in component attachment also requires special consideration. The critical issue with all plug-in connections is how successfully low contact resistance can be maintained over the life of the product. Corrosion films, dust particles, and wear of the contact metal can all lead to high-resistance failures through the connection. For the highest reliability applications, the contact metallization over the underlying copper is typically an electroplate of gold over nickel. The most important reliability concern is the porosity of the gold. If the gold is sufficiently porous, the underlying nickel and copper may corrode, especially in environments with high concentrations of hydrogen sulfide, e.g., near oil fields, oil refineries, or pulp mills. If corrosion occurs, the corrosion products slowly migrate across the gold surface to form a highly insulating, tenaciously adhering film. Two approaches to evaluating this potential failure process are in common use. One is to measure the porosity of the gold electroplate immediately after plating by exposing the electroplate to a corrodant and then counting the pore density revealed by corrosion spots for pores larger than a given diameter. A density greater than a set standard indicates a high susceptibility to field failure by hydrogen sulfide–induced corrosion. An alternative test is exposure of the connectors to a corrosive-gas environment, as

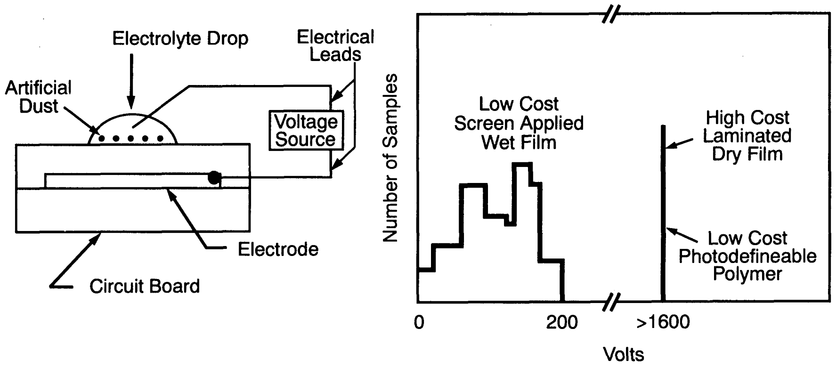

FIG. 15. Test apparatus and selected data for measurement of breakdown voltage of polymeric dielectrics. Left: The apparatus for measuring breakdown voltage. A test pattern on a circuit board is coated with a polymeric material. A hygroscopic contaminant, such as synthetic dust, is placed on the polymer followed by a drop of water. Voltage is then applied across the polymer by connecting one lead to the conductor and placing the other in the drop of water. Right: Results for three different polymeric covercoats. The screen-applied wet film shows erratic breakdown voltages well below the 1600-V design criterion for many applications. Screen-applied films typically have pinhole defects that cause low breakdown voltage.

described in Table 1, followed by evaluation of the contact resistance by electrical measurements. For most computer applications the resistance should be <20 mΩ.

Tin or a tin-lead alloy is sometimes used for plug-in connections, especially for components attached to circuit boards. The most common failure process associated with these metals, when used for connectors, is fretting corrosion leading to a high contact resistance (see Fig. 16). Fretting corrosion is typically caused by the micromotion between connector pairs associated with ordinary vibrations. Such failures have sometimes been attributed to the vibrations caused by motors for cooling fans. The vibrations disrupt the native oxide on the metal, leading to further oxidation and fretting. As the oxide debris builds up, the contact resistance increases. The appropriate stress test for evaluating susceptibility to fretting corrosion is exposure to vibrations with variable amplitude and frequency.

3. FUTURE REQUIREMENTS IN RELIABILITY AND TESTING

Several driving forces will affect not only the kinds of phenomena that are likely to be important in the future reliability of electronics, but also the way in which reliability practitioners carry out their mission (Crook, 1990). These driving forces include

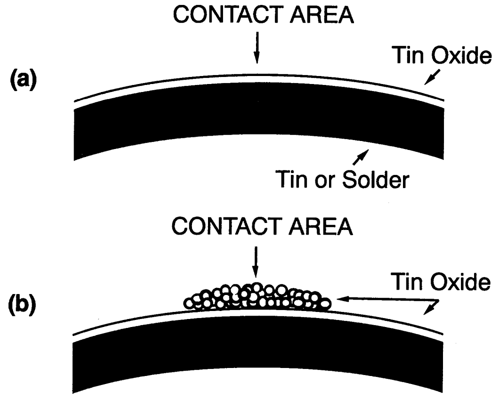

FIG. 16. Fretting corrosion of connector contacts. Motion from ordinary vibrations can produce enough oxide debris (fretting) to cause contact failure. (a) Cross section of a connector contact surface with an oxide layer over the tin or tin-lead alloy. (b) The same surface after fretting has generated tin-oxide particles that prevent intimate contact with the opposing contact surface (which is not shown).

1. the increasing density and shrinking feature sizes of integrated-circuit components;
2. the growing complexity of systems that use integrated circuits and other electronic components; and
3. the increasing critical demand for reliability from the customer.

Item **1**, as will be discussed, continually decreases the reliability margin (inherent reliability minus needed reliability) by decreasing the inherent reliability, while items **2** and **3** result in continuing pressure for an increase in the reliability level, further decreasing the available margin. Since margin cannot decrease to zero without annoying the customer, the practitioner must seek to increase the inherent reliability of the device by introducing process and design improvements as the circuit density increases.

3.1 Integrated Circuits—Shrinking Dimensions

3.1.1 Device Scaling. As feature size decreases, "scaling laws" show that the expected lifetime of a circuit may decrease drastically (Woods, 1986). These scaling laws consider the effect of shrinking dimensions on the time to failure expected for various wearout mechanisms. As dimensions decrease, wearout mechanisms are accelerated by the changes in relative dimensions. This can be illustrated for the case of electromigration failure. Assume that the thickness and width of the aluminum line decrease by the same factor K. Then the cross-sectional area of the metal line will decrease by a factor K^2. If gate oxide thickness is scaled but device voltage is not (as is often the case), the device conductance and, therefore, current increase linearly with the scale factor. Thus, the effect of scaling increases current density by K^3 for geometric factors. Additional considerations related to metal line cross sections at topological steps increase the exponent to four. Since the lifetime for electromigration failure is inversely proportional to the square of the current density, the lifetime decreases by the eighth power of the scaling factor.(Fortunately, in the case of electromigration, narrow lines benefit from the "bamboo effect" discussed in Sec. 2.2.3.1) Similar considerations apply to oxide wearout and breakdown and result in shorter mean time to failure for reduced oxide thickness. By these and other scaling effects, the ultimate reliability decreases and approaches the advertised value.

3.1.2 Particle Effects. Airborne particles that land on integrated circuits during manufacturing can affect both device yield and reliability, especially in the short term (infant mortality). The particles can interfere with photolithographic processes and cause shorts between conductors or opens in conductor lines, resulting in yield loss. Or the lithographic defects may be partial in nature, so that insulator thickness or metal width is reduced locally, increasing failures. Particle contamination can also affect breakdown voltage and leakage current.

For a given feature size, only particles that are larger than about one-tenth of the characteristic feature dimension are expected to have an effect. As feature sizes decrease with advanced device technologies, the size of particles expected to result in failures also decreases. The volume density of airborne particles increases dramatically with decreasing particle size. Thus, ever more stringent filtration requirements for manufacturing areas can be expected along with more effective protective methods and cleaning processes.

3.1.3 Advanced Packaging. Increased component density tends to result in greater power dissipation per unit area of integrated circuit. The parallel trend to greater system speed and density is driving new interconnect technologies, such as solder connection of flip chips instead of wire bonds. The requirements for denser interconnections and increased power dissipation are met with more complicated combinations of materials. Complex metal structures may be more susceptible to interdiffusion and formation of intermetallic compounds, especially at elevated operating temperatures. The reliability will depend on the quality of the selected barrier metals. Combinations of metals also increase the probability of galvanic reactions. However, the increase in device operating temperature and the correspondingly reduced humidity tends to mitigate against these and other corrosion reactions. Electromigration effects, on the other hand, will be amplified by the higher temperature.

3.2 Reliability Verification and Improvement

As the trend to higher density and decreasing feature size reduces the reliability margin of devices, the electronics industry at the same time demands greater reliability. The demand for greater reliability arises not merely because of our normal expectation that products should improve but because of the greater complexity and criticality of systems. For example, the consequences of a failure in an aircraft navigation system are much greater than those of television set failure. Even more important is the effect of a defect in a computerized aircraft landing system.

On the basis of historical trends, the expected failure rates for complex devices in the year 2000 will be approximately one failure in 10^8 device-hours (Crooks, 1990). Our present methods of assessing failure rates, by using highly accelerated stress tests and screens, will be inadequate for assessing this level of reliability. The testing requirements (number of devices and test duration) will make these kinds of tests impractical for verifying attainment of these reliability goals. Wafer-level reliability testing, a current popular method, for assessing base-line reliability of a process, will also be inadequate. This technique applies highly accelerated stresses to devices in wafer form to assess inherent reliability. Such tests have found extensive use, particularly for electromigration and oxide breakdown assessment. Wafer-level testing can give a false satisfactory result when several process parameters are out of specification in a way that the process variations cancel each other's effect. For example, a particular lot of devices may have aluminum thickness, linewidth, and steepness of step contours all out of control; but overly steep step contours, which reduce electromigration life time, may be compensated by excessive linewidth and metal thickness. If the implications of this example are extended to the full complexity of process parameters involved in electromigration reliability, it is clear that wafer-level testing on a wafer from an out-of-control process could give a satisfactory result even though actual reliability in the field will be unsatisfactory.

Since reliability screens, accelerated-life tests, and wafer-level stressing are all expected to be inadequate for the very stringent reliability demands of the future, statistical process control must become sufficiently precise and device design sufficiently robust that high reliability can be guaranteed without testing. Reliability practitioners will spend less time understanding the effect of a failure mechanism on actual field reliability and more time on understanding the effects of physical and chemical processes on material and structural variation and designing controls to monitor these processes. Improvements in reliability will be achieved by greater understanding of the degradation mechanisms of materials and their relation to processing, and the development of analytical tools and techniques to monitor the relevant parameters in a way that continuously decreases process variations.

GLOSSARY

Acceleration Factor: The inverse ratio of the time to failure of a product undergoing accelerated-life testing to its time to failure by the same mechanism in normal field service.

Active Components: Electronic devices that amplify or are sources of current or voltage. Active components include transistors, thyristors, rectifiers, and diodes.

Bonding Pads: Areas of metal on substrates, components, or integrated circuits that serve as input or output points for connecting to other substrates, components, or integrated circuits through solder joints or wire bonds.

Breakdown Voltage: Minimum voltage applied across an insulator or semiconductor that results in sudden loss of insulating properties and simultaneous current flow.

Circuit Board: A functional unit consisting of a printed wiring board with attached components and an edge-mounted connector for plugging the unit into an assembly.

Covercoat: A thin coating, usually polymeric, applied to a printed wiring board or other substrate to inhibit degradation processes caused by interaction of the printed wiring board with the environment, especially ionic contaminants and moisture.

Encapsulant: A thick coating applied to a component or device to inhibit interaction of the functional elements with the environment.

Electromigration: The degradation of aluminum or other metal conductors when subjected to high current density, due to momentum transfer from electrons to metal atoms. The momentum transfer causes atoms to move, eventually leading to open circuits.

Failure Mode Analysis: The process of characterizing a failed device or assembly by analytical techniques and electrical measurement to ascertain the cause of failure.

Field Plates: Metal plates or structures in an integrated circuit or component that are connected to a reference voltage and fix, at a specified value, the electric field below them. They stabilize the breakdown voltage of the device.

Hermetic Package: An enclosure, usually ceramic with a ceramic or metal lid, that provides electrical connections and prevents interaction of an enclosed component with the external environment.

Interconnect: A network of metal conducting pathways that connects components on or between circuit boards or other substrates or connects transistors or other elements on integrated-circuit chips.

Passivation (Passivation Layers): The surface coating on an integrated circuit or component that provides mechanical and environmental protection. Typical passivation layers are silicon nitride and silicon dioxide doped with phosphorus.

Passive Components: Electronic devices that alter a current or voltage without providing amplification. Passive components include capacitors, resistors, inductors, transformers, and relays.

Printed Wiring Board: A sheet of dielectric material with a metallization pattern on its surface, frequently packaged in multiple layers, that serves as the substrate on which components are mounted to make a circuit board.

Root Cause: The original error in judgment, process function, or product design that ultimately leads to product failure.

Sacrificial Oxides: Oxides that are grown on silicon and then removed by etching to increase the near-surface purity of the silicon.

Solder Reflow: Process of remelting solder that has been previously fused to a substrate or integrated circuit so that a component or integrated circuit can be bonded to the substrate.

Statistical Quality Control: Use of statistical techniques to pinpoint and eliminate the sources of variation in product quality.

Surface Ions: Ions on the surface of insulating materials that migrate in the presence of electric fields and may alter device function.

Undensified Film: A film, usually an insulating oxide, that is deposited at a relatively low temperature and has a density lower than that of the bulk material. Such films are mechanically weaker and less durable chemically, and provide for more rapid diffusion of impurities, than the corresponding densified material. Undensified films may be densified by heat treatments.

List of Works Cited

Amerasekera, E. A. and Campbell, D. S. (1987), *Failure Mechanisms in Semiconductor Devices,* pp. 15–20, Chichester: Wiley.

Ames, I., d'Heurle, F. M., Horstmann, R. (1970), *IBM J. Res. Devel.* **14**, 461–463.

Atalla, M. M., Bray, A. R., Lindner, R. (1959), *Proc. IEE, Pt. B,* **106**, 1130–1137.

Bergh, A. A. (1964), in: M. J. Goldberg and J. Vaccaro (Eds.), *Physics of Failure in Electronics,* New York: IEEE, Vol. 3, pp. 421–432.

Black, J. R. (1967), in: *Sixth Annual Reliability Physics Proceedings,* New York: IEEE, pp. 148–159.

Braunig, D., Fahrner, W. R. (1986), in: G. Barbottin and A. Vapaille (Eds.), *Instabilities in Silicon Devices,* Vol. 1, Amsterdam: North Holland.

Comizzoli, R. B., Frankenthal, R. P., Milner, P. C., Sinclair, J. D., (1986), "Corrosion of Electronic Materials and Devices," *Science,* **234**, 340–345.

Comizzoli, R. B., Osenbach, J. W. (1988), *J. Electrochem. Soc.* **135**, 1517–1521.

Crook, D. L. (1990), in: *Twenty-Eighth Annual Reliability Physics Proceedings,* New York: IEEE, pp. 2–11.

Deal, B. E. (1974). *J. Electrochem. Soc.* **121**, pp. 198C–205C.

Deming, W. A., (1982), *Out of Crisis,* Cambridge, MA: MIT Press.

Hu, C., Tam, S. C., Hsu, F-C., Ko, P-K., Chan, T-Y., Terrill, K. W. (1985), *IEEE Trans. Electron. Devices* **ED-32**, 375–385.

Klinger, D. J., Nakada, Y., Menendez, M. A. (1990) (Eds.), *AT&T Reliability Manual,* New York: Van Nostrand-Reinhold.

Nicollian, E. H., Brews, J. R. (1982), *MOS (Metal Oxide Semiconductor) Physics and Technology,* New York: Wiley.

Queisser, H. J. (1963), in: M. F. Goldberg and J. Vaccaro (Eds.), *Physics of Failure in Electronics,* New York: IEEE, Vol. 2, pp. 476–482.

Schnable, G. L. (1989), in: E. B. Hakim (Ed.), *Microelectronic Reliability,* Vol. 1, Norwood: Artech House.

Shewhart, W. A., (1939), *Statistical Method from the Point of View of Quality Control,* Washington: Graduate School, Dept. of Agriculture; Reprint (1986), New York: Dover.

Sinclair, J. D., (1988), "Corrosion of Electronics: The Role of Ionic Substances," *J. Electrochem. Soc.* **135**(3), C89–C95.

Stolnitz, D. (1965), in: M. F. Goldberg and J. Vaccaro (Eds.), *Physics of Failure in Electronics,* New York: IEEE, Vol. 4, pp. 227–239.

Wolters, D., Verwey, J. F. (1986), in: G. Barbottin and A. Vapaille (Eds.), *Instabilities in Silicon Devices,* Vol. 1, Amsterdam: North-Holland.

Yue, J. T., Funsten, W. P., Taylor, R. V. (1985), in: *Twenty-Third Annual Reliability Physics Proceedings,* New York: IEEE, pp. 126–137.

Woods, M. H. (1974), in: *Twelfth Annual Reliability Physics Proceedings,* New York: IEEE, pp. 259–266.

Woods, M. H. (1986), *Proc. IEEE* **74**, 1715–1729.

Further Reading

Friedlander, S. K., (1977), *Smoke, Dust and Haze: Fundamentals of Aerosol Behavior,* New York: Wiley.

Jensen, F., Petersen, N. E. (1982), *Burn-In: An Engineering Approach To The Design and Analysis of Burn-in Procedures,* New York: Wiley.

Lauger, E., Moltoft, J. (Eds.) (1982), *Reliability In Electrical and Electronic Components and Systems,* Amsterdam: North-Holland.

Marsh, L. L., Havens, R. D., Show-Chi, W., Malack, J. A., Ulsh, H. B. (1989), in: D. P. Seraphim, R. C. Lasky, L. Che-Yu (Eds.), *Principles of Electronic Packaging,* New York: McGraw-Hill, Chap. 9, "Reliability and Testing."

Pollino, E. (1989), *Microelectronic Reliability,* Vol. 2: *Integrity Assessment and Assurance,* Norwood: Artech House.

Seinfeld, J. H., (1987), *Atmospheric Chemistry and Physics of Air Pollution,* New York: Wiley.

Military Standards (1989), *Reliability Prediction of Electronic Equipment,* MIL HDBK 217E, Washington: U.S. Department of Defense.

ELECTRONIC STRUCTURE OF ATOMS AND MOLECULES

JOHN C. MORRISON, *Department of Physics, University of Louisville, Louisville, Kentucky, U.S.A.*
ANDREW W. WEISS, *Physics Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland, U.S.A.*
KATE KIRBY, *Institute for Theoretical Atomic and Molecular Physics, Harvard–Smithsonian Center for Astrophysics, Cambridge, Massachusetts, U.S.A.*
DAVID COOPER, *Department of Chemistry, University of Liverpool, Liverpool, United Kingdom*

Introduction 46
1. **The Hydrogen Atom** 46
2. **The Independent-Particle Model** 47
3. **The Hartree-Fock Approach** 49
3.1 Operator Formulation 49
3.2 Expansion Method 50
4. **Atomic Structure** 51
4.1 Central-Field Model 51
4.2 The "Building-Up" Principle 51
4.3 Russell-Saunders Coupling 51
4.4 Partial-Wave Expansion of the Coulomb Interaction 52
4.5 Antisymmetric Wave Functions for Two Electrons in *LS* Coupling 53
4.6 The Coulomb Interaction for Two-Electron Systems in *LS* Coupling 54
4.7 Coupling Schemes for Two-Electron Systems 55
4.8 Energy-Level Structure of Many-Electron Atoms 56
4.9 Hartree-Fock Model 57
4.10 Properties of Hartree-Fock Orbitals 57
4.10.1 Radial Shell Dependence 58
4.10.2 *LS* Term Dependence 58
4.10.3 Shell Collapse 60
5. **Correlation in Electronic Structure Theory** 61
5.1 Nature of Electron Correlation 61
5.2 The Hylleraas Method 62
5.3 The Configuration-Interaction Method 63
5.4 Multiconfiguration Hartree-Fock Method 64
5.5 Many-Body Perturbation Theory 65
5.6 Correlation in Atomic Spectroscopy 67
6. **Relativistic Effects in Atomic Structure** 69
7. **Fundamentals of Molecular Electronic Structure** 72
7.1 The Born-Oppenheimer Approximation 73
7.2 Qualitative Ideas 75
7.3 Classical Valence-Bond Theory 75
7.4 Molecular-Orbital Theory—The Hartree-Fock Self-Consistent Field Approach 77
7.5 Molecular Symmetry 80
7.6 Bonding in Diatomic Molecules 82
7.7 Basis Sets for Molecular Calculations 83
8. **Methods for Taking Account of Electron Correlation in Molecules** 85
8.1 Multiconfiguration Self-Consistent-Field Method 86
8.2 Configuration Interaction 88
8.3 Many-Body Perturbation Theory 89
8.4 Coupled-Cluster Methods 90
8.5 Modern Valence-Bond Theory 91
8.6 Density-Functional Theory 92
9. **Molecular Properties** 93
9.1 Geometries and Gradients 93
9.2 Energy Differences 94
9.3 Matrix Elements and Response Functions 95

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Acknowledgment 97
List of Abbreviations 97
Works Cited 97
Further Reading 98

INTRODUCTION

Much of the original insight into the quantum-mechanical nature of the electronic structure of atoms and molecules came from spectroscopy. This early experimental approach together with an increasingly complex array of scattering and resonance techniques has led to a wealth of experimental data, which can be interpreted using models of atomic and molecular structure. From the earliest years following the development of quantum mechanics, empirical models have been used that enable us to correlate and codify our knowledge and to predict the results of further experiments.

As mathematical and computational techniques have developed, electronic-structure theory has established itself as an independent branch of science. While the success of any theory depends upon its ability to reproduce the results of experiments, the nonrelativistic theory can still be formulated as a scheme for obtaining approximate solutions of the Schrödinger equation. The corresponding first-principles formulation of the relativistic theory is presently being rapidly developed. It would take many different kinds of spectroscopy and scattering experiments to gain as much knowledge about the structure and properties of an atom or molecule as can be obtained nowadays by a good *ab initio* calculation. *Ab initio* literally means "from the beginning" and implies a theoretical study based on quantum mechanics, without the input of experimental data specific to the system being studied. The power of many types of theoretical calculations comes from obtaining a wave function that can subsequently be used to calculate different physical properties. In this article, we shall develop from first principles the theoretical models that play an important role in this vast area of research.

While there are many obvious similarities between atoms and molecules, there are also important differences. For atoms, highly charged ions can be produced in the laboratory, and this makes it possible to study isoelectronic sequences over a wide range of possible ionization states. In this way, the nuclear charge of an electronic state can be turned up to quite high values. This provides a good setting for testing quantum electrodynamics and the models for applying relativistic quantum mechanics to many-electron systems. On the other hand, the various molecular phenomena associated with the relative motion of the nuclei, such as dissociation and the formation of chemical bonds, do not arise when we consider atomic structure. Apart from the effects of the hyperfine interaction, the physical and optical properties of atoms are due entirely to their electronic structure, whereas for molecules, the description of their electronic structure is only the first part of a complex and multifaceted problem.

The purpose of this article is to give an integrated account of the electronic structure of atoms and molecules. The article begins with a review of the most essential aspects of the theory of the hydrogen atom, and then gives a brief description of the independent-particle and Hartree-Fock models, which provide an intuitively satisfying framework for obtaining approximate solutions of the Schrödinger equations for atoms and molecules.

1. THE HYDROGEN ATOM

The hydrogen atom consists of a proton and an electron. Without introducing any difficulty, however, the problem can be generalized to include a nucleus of charge Ze in place of the proton. The nuclear mass will be denoted by M and the charge and mass of the electron by $-e$ and m_0. The only interaction between the nucleus and the electron that will be considered at this time is the electrostatic interaction described by the Coulomb potential $u(r) = -Ze^2/4\pi\epsilon_0 r$. As we shall see, the central nature of the potential simplifies considerably this one-electron problem. The time-independent Schrödinger equation for the relative motion may be written

$$H\Psi = E\Psi, \tag{1}$$

where Ψ is the wave function and the Hamiltonian H is given by the expression

$$H=(-\hbar^2/2m)\nabla^2+u(r). \tag{2}$$

Here m is the reduced mass of the electron–nuclear system, $m=m_0M/(m_0+M)$. The first term on the right-hand side of Eq. (2) represents the kinetic energy of the electron, and the second term corresponds to the attractive potential of the nucleus. In what follows, we shall use atomic units, for which $\hbar$, $4\pi\epsilon_0$, the mass and charge of the electron, and the Bohr radius a_0 all have the value 1. (The value of a_0 is approximately equal to 0.5292×10^{-10} m.)

The quantum-mechanical angular-momentum operator is

$$\mathbf{l}=-i\mathbf{r}\times\nabla. \tag{3}$$

The components l_k may be shown to commute with the central-field Hamiltonian and hence correspond to constants of the motion. This is the underlying reason for the importance of angular momentum to atomic physics, and it forms the basis for the shell structure of atoms.

The components of the angular momentum also commute with $\mathbf{l}^2$. It follows that H, $\mathbf{l}^2$, and l_z (for instance) are a set of mutually commuting operators and that they possess a complete set of simultaneous eigenstates. The simultaneous eigenfunctions of $\mathbf{l}^2$ and l_z are the spherical harmonics $Y_{lm}(\theta,\phi)$, where θ and ϕ are ordinary polar coordinates.

The spherical harmonic for $l=0$ is spherically symmetric, while the function $Y_{lm}(\theta,\phi)$ for $l=1$ and $m=0$ is directed along the z axis. Similarly, there are linear combinations of the spherical harmonic for $l=1$, $m=1$ and $l=1$, $m=-1$ that are directed along the x and y axes. For atoms, it is customary to use the spectroscopic notation to denote the orbital angular momentum of the electrons. In this notation, the letters *s,p,d,f,g,h,i,...* stand for the l values 0,1,2,... .

The wave function of a particle moving in a central field can be written as the product of a radial function, a spherical harmonic Y_{lm_l}, and a spin function χ_{m_s}. It is generally written

$$\Psi=\frac{P(r)}{r}\,Y_{lm_l}(\theta,\phi)\chi_{m_s}. \tag{4}$$

The two possible spin states of a single electron, for which $s=\frac{1}{2}$ and $m_s=\pm\frac{1}{2}$, are usually denoted α and β. Since the spherical harmonic Y_{lm_l}, which describes the angular dependence of the wave function Ψ, is an eigenfunction of $\mathbf{l}^2$ and l_z corresponding to the eigenvalues $l(l+1)$ and m_l, the single-particle wave function itself also has this property. Requiring that Ψ be an eigenfunction of the central-field Hamiltonian corresponding to the eigenvalue E leads to the radial equation

$$\left(-\frac{1}{2}\frac{d^2}{dr^2}+\frac{l(l+1)}{2r^2}+u(r)\right)P(r)=EP(r). \tag{5}$$

The first term in the equation corresponds to the radial part of the kinetic energy of the electron, the second term corresponds to the angular or centrifugal part of the kinetic energy, and the third term corresponds to the interaction of the electron with the central field. The centrifugal term tends to push the maxima and nodes of the higher angular momentum states away from the nucleus. For one-electron atoms, the central potential (in atomic units) is simply $u(r)=-Z/r$.

For the wave function to be finite at the origin, it is necessary for $P(r)\to0$ as $r\to0$. Furthermore, for Ψ to be normalizable, it must be true that $P(r)\to0$ as $r\to\infty$. The values of E for which there exist solutions of the radial equation satisfying the boundary conditions are the possible energies of an electron moving in a central potential. These solutions may be distinguished by the value of l to which they correspond and the number of nodes of the radial function. The principal quantum number is related to the number of nodes ν: $n=l+\nu+1$. The radial functions $P(r)$ for several states are shown in Fig. 1. The wave function of the ground state of the hydrogenic atom is

$$\psi=(Z^3\pi)^{-1/2}\exp(-Zr). \tag{6}$$

2. THE INDEPENDENT-PARTICLE MODEL

A useful first approximation of more complex atoms and molecules is obtained by assuming that each electron moves independently of the other electrons in an average field due to the nuclei and the other electrons. This assumption leads to the independent-particle model, which provides an approximate description of atomic and molecular

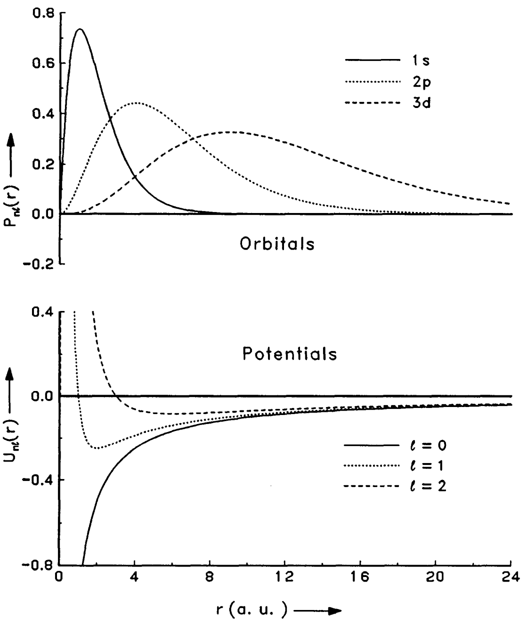

FIG. 1. Radial wave functions and potentials for the hydrogen atom. Potentials include the centrifugal repulsion term arising from the kinetic energy operator.

systems and also serves as the starting point of more accurate calculations.

The simplest wave function that describes a system of N electrons moving in an average potential is the product function

$$\Psi = \varphi_{a_1}(1)\varphi_{a_2}(2)\cdots\varphi_{a_N}(N). \tag{7}$$

Here $a_1, a_2, \ldots$ refer to the quantum numbers necessary to specify the single-electron states, and 1,2,... refer to the coordinates of the electrons. Any permutation of the electrons among the single-particle wave functions also leads to wave functions that describe this system of electrons. One may ensure that the Pauli exclusion principle is satisfied by forming an antisymmetric linear combination of the product functions:

$$\Phi = \left(\frac{1}{N!}\right)^{1/2} \sum_{P} (-1)^{p} P \times [\varphi_{a_1}(1)\varphi_{a_2}(2)\cdots\varphi_{a_N}(N)]. \tag{8}$$

The sum extends over all permutations P of the electrons among the single-particle states, and p is the parity of the permutation. The state Φ can be written in the following way:

$$\Phi = \left(\frac{1}{N!}\right)^{1/2} \begin{vmatrix} \varphi_{a_1}(1) & \varphi_{a_1}(2) & \cdots & \varphi_{a_1}(N) \\ \varphi_{a_2}(1) & \varphi_{a_2}(2) & \cdots & \varphi_{a_2}(N) \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_{a_N}(1) & \varphi_{a_N}(2) & \cdots & \varphi_{a_N}(N) \end{vmatrix}. \tag{9}$$

Such a state is called a determinantal product state or a *Slater determinant.* An interchange of the coordinates of any two of the electrons leads to the interchange of two columns of the determinant, and this will cause a sign change. If any two of the single-particle states are the same, two rows of the determinant will be equal, and the determinant will vanish. Thus the Pauli exclusion principle is automatically satisfied. Determinant states of this kind provide a useful basis for atomic and molecular calculations

Simple formulas are available for evaluating the matrix elements of one- and two-particle operators between determinantal states. The diagonal matrix elements of a general single-particle operator, $F = \Sigma_i f(i)$, are given by the formula

$$\langle \Phi | F | \Phi \rangle = \sum_{a} \langle a | f | a \rangle. \tag{10}$$

Here the letter a stands for the orbital φ_a specified by that set of quantum numbers. For this formula to apply, the orbitals must be orthonormal. The corresponding formula for the matrix element of the two-particle operator

$$G = \sum_{i<j} g(i,j) \tag{11}$$

is

$$\langle \Phi | G | \Phi \rangle = \sum_{a<b} [\langle ab | g | ab \rangle - \langle ab | g | ba \rangle]. \tag{12}$$

In the integrations implied by the matrix element $\langle ab | g | ab \rangle$, for instance, the functions φ_a on the left and right correspond to the first variable and the functions φ_b to the second. The sums over a and b include all orbitals that are occupied in the determinantal state. There are similar formulas for evaluating the off-diagonal matrix elements.

In order to construct a determinantal state Φ, it is necessary to know how the single-electron states are ordered within the determinant. This information may be provided by writing

$$\Phi=(N!)^{-1/2}|\varphi_{a_1}\varphi_{a_2}\cdots\varphi_{a_N}|. \tag{13}$$

3. THE HARTREE-FOCK APPROACH

3.1 Operator Formulation

In the previous section, the many-body wave function was approximated by determinantal product functions. For a closed-shell atom or molecule, the Hartree-Fock wave function is the variationally optimized single determinant of doubly occupied orbitals

$$\Psi=(2n!)^{-1/2}|\varphi_1\alpha(1)\varphi_1\beta(2)\varphi_2\alpha(3) \times\cdots\varphi_n\alpha(2n-1)\varphi_n\,\beta(2n)|. \tag{14}$$

One usually assumes that the orbitals are orthonormal,

$$\langle\varphi_i|\varphi_j\rangle=\delta_{ij}\,. \tag{15}$$

At this point, we will leave unspecified the symmetry and coordinate system, so that all that follows applies equally to the electronic structure of an arbitrary molecule as well as an atom. Using the formula from Sec. 2 for the diagonal matrix elements of one- and two-particle operators between determinantal states, the total-energy expression for such a wave function can be written

$$E=\langle\Psi|\mathscr{H}|\Psi\rangle =2\sum_i\langle\varphi_i|h|\varphi_i\rangle+\sum_{i,j}\left(2\left\langle\varphi_i\,\varphi_j\left|\frac{1}{r_{ij}}\right|\varphi_i\,\varphi_j\right\rangle -\left\langle\varphi_i\,\varphi_j\left|\frac{1}{r_{ij}}\right|\varphi_j\,\varphi_i\right\rangle\right), \tag{16}$$

where h is the single-particle Hamiltonian,

$$h=-(\nabla^2/2)-\sum_a(Z_a/r_a). \tag{17}$$

Here Z_a denotes the charges of the nuclei of the molecular system. The expression for the energy may be written more simply by introducing Coulomb and exchange operators, J_j and K_j, defined by the equations

$$J_j(1)\varphi(1)=\left(\int dV_2[\varphi_j(2)]^2\frac{1}{r_{12}}\right)\varphi(1), \tag{18}$$

$$K_j(1)\varphi(1)=\left(\int dV_2\varphi_j(2)\varphi(2)\frac{1}{r_{12}}\right)\varphi_j(1), \tag{19}$$

where φ is an arbitrary function. Using these equations, the expression for the total energy can be written

$$E=\langle\Psi|\mathscr{H}|\Psi\rangle =2\sum_i\langle\varphi_i|h|\varphi_i\rangle\sum_{i,j}\langle\varphi_i|2J_j-K_j|\varphi_i\rangle. \tag{20}$$

We require that the energy be stationary,

$$\delta E=0, \tag{21}$$

for small changes in the single-electron wave functions that are consistent with the orthogonality constraints. Multiplying the variation of each orthogonality constraint by a Lagrange multiplier and subtracting the result from the variation of the total energy gives

$$\delta E'=\delta E-\sum_{i,j}4\epsilon_{ij}\langle\delta\varphi_i|\varphi_j\rangle, =4\sum_i\left\langle\delta\varphi_i\left|\left(h+\sum_j(2J_j-K_j)\right)\varphi_i -\sum_j\epsilon_{ij}\,\varphi_j\right.\right\rangle. \tag{22}$$

Each integrand in Eq. (22) must vanish for any variation of the orbitals with appropriate values of the Lagrange multipliers. This condition will be satisfied only if each term vanishes identically:

$$\left(h+\sum_j(2J_j-K_j)\right)\varphi_i=\mathscr{F}\varphi_i=\sum_j\epsilon_{ij}\,\varphi_j\,. \tag{23}$$

Here $\mathscr{F}$ is called the Fock operator, and Eqs. (23) are called the Hartree-Fock equations. The matrix of Lagrange multipliers, ϵ_{ij}, is symmetric as are the orthogonality constraints themselves. For closed-shell systems, it is possible to show that the Hartree-Fock equations are covariant under unitary transformations of the orbitals and that the determinantal wave function itself is invariant under such unitary transformations. If we solve the equations for the unitary transformation that diagonalizes the matrix of Lagrange multipliers, the Hartree-Fock equations take the canonical form of single-particle eigenvalue equations,

$$\mathscr{F}\varphi_i=\left(h+\sum_j(2J_j-K_j)\right)\varphi_i=\epsilon_i\varphi_i\,. \tag{24}$$

Since the operators $2J_i-K_i$ that describe the potential field are themselves functions of the solutions, the equations must be solved iteratively until self-consistency is achieved. For this reason, the Hartree-Fock equations are often referred to as the self-consistent-field (SCF) equations.

Although the sum over j in the Hartree-Fock equations runs over all of the occupied orbitals, the definitions of the Coulomb and exchange operators may be used to show that the self-interaction term vanishes identically:

$$(J_i-K_i)\varphi_i\equiv 0. \tag{25}$$

Using this result and the definition of the exchange operator, the Hartree-Fock equations can be cast in a form in which the inhomogeneous character of these equations is apparent:

$$\left(h+2\sum_{j\neq i} J_j+J_i\right)\varphi_i = \epsilon_i\,\varphi_i+\sum_{j\neq i}\left(\int dV\,\varphi_j\varphi_i\frac{1}{r_{12}}\right)\varphi_j. \tag{26}$$

The eigenvalue of the Hartree-Fock equation may be shown to be equal to the "frozen-orbital" ionization energy, i.e., the ionization energy obtained by assuming the orbitals do not relax upon removal of one electron. These eigenvalues have therefore come to be called the orbital energies. This result, which is referred to as Koopmans's theorem, provides the final touch that enables one to interpret the variational self-consistent-field equations as a set of Schrödinger's equations for a single electron moving in the effective field of all the other electrons of the atom or molecule.

3.2 Expansion Method

The expansion method represents an alternative formulation of the Hartree-Fock scheme that is, in principle, entirely equivalent to the original methodology and particularly attractive for molecular calculations. One writes the orbitals as an expansion in a set of analytical *basis functions,*

$$\varphi_i=\sum_p c_{ip}\chi_p\,. \tag{27}$$

The central idea here is to "fit" the unknown Hartree-Fock functions with a set of analytical functions, using the variational principle to determine the fitting coefficients.

Inserting the expansion for the orbitals into the expression for the total energy gives the energy as a function of matrix elements and coefficient vectors,

$$E=2\sum_i \mathbf{c}_i^\dagger\mathbf{H}\mathbf{c}_i+\sum_{i,j}\mathbf{c}_i^\dagger\mathbf{G}_j\mathbf{c}_i\,. \tag{28}$$

The matrices **H** and **G** are given by

$$(\mathbf{H})_{pg}=\langle\chi_p|h|\chi_q\rangle, \tag{29}$$

$$(\mathbf{G}_j)_{pq}=\sum_{s,t} c_{js}c_{jt}[2(pq|st)-(ps|qt)]; \tag{30}$$

where

$$(pq|st)=\int dV_1dV_2\chi_p(1)\chi_q(1)\frac{1}{r_{12}}\chi_s(2)\chi_t(2) \tag{31}$$

are the two-electron repulsion integrals over the basis functions from which *both* the Coulomb and exchange integrals are constructed. The orthonormality constraints then take the form

$$\mathbf{c}_i^\dagger\mathbf{S}\mathbf{c}_j=\delta_{ij}\,, \tag{32}$$

where

$$(\mathbf{S})_{pq}=\langle\chi_p|\chi_q\rangle \tag{33}$$

is the basis-function overlap matrix, which is not generally diagonal.

The variational principle with the constraints included then requires that

$$\delta E'=4\sum_i \delta\mathbf{c}_i^\dagger\left(\mathbf{H}\mathbf{c}_i+\sum_j \mathbf{G}_j\,\mathbf{c}_i-\sum_j \epsilon_{ij}\,\mathbf{S}\mathbf{c}_j\right)=0 \tag{34}$$

for any arbitrary variation of the coefficient vectors $\mathbf{c}_i$. This leads to the matrix self-consistent-field equations,

$$\mathbf{F}\mathbf{c}_i=\left(\mathbf{H}+\sum_j \mathbf{G}_j\right)\mathbf{c}_i=\epsilon_i\mathbf{S}\mathbf{c}_i\,, \tag{35}$$

where we have again transformed the matrix of Lagrange multipliers to diagonal form. Because exchange has been included in the Coulomb interaction matrix $\mathbf{G}_j$, the expansion-method equations are true homogeneous equations. Just as in the operator formulation, they must be solved iteratively, the interaction matrix being a function of the expansion coefficients.

Some writers prefer to express the Fock matrix in terms of the density matrix **D**, defined by the equation

$$(\mathbf{D})_{pq}=\sum_{i}^{n} c_{ip}c_{iq}\,. \qquad (36)$$

Using the density matrix, the Fock matrix may be written

$$(\mathbf{F})_{pq}=(\mathbf{h})_{pq}+\sum_{s,t}(\mathbf{D})_{st} \times[2(pq|st)-(ps|qt)]. \qquad (37)$$

The main problem in implementing the expansion method centers around the selection of the basis set. For atomic calculations, the most commonly used basis sets consist of Slater-type orbitals,

$$\chi_{plm}^{S}(\zeta)=N_s r^p e^{-\zeta r} Y_{lm}(\theta,\varphi). \qquad (38)$$

The exponential parameter here is used as an additional variational parameter to enhance the expansion-method "fitting process."

4. ATOMIC STRUCTURE

4.1 Central-Field Model

In the previous two sections, we considered the independent-particle model in which the electrons are assumed to move independently of each other in an average field due to the nuclei and the other electrons. This assumption reduces the many-electron problem to a number of independent single-electron problems. For atoms, we can generally make the additional assumption that the potential is spherically symmetric. As we have seen previously for the hydrogen atom, the single-electron orbitals can then be written as a product of a radial function, a spherical harmonic, and a spin function.

All of the orbitals of a single electron in a central field that have the same values of n and l are said to form an atomic subshell. These orbitals are all described by the same radial function, and they all have the same energy. The state of the many-electron system can be specified by saying which of these single-electron states are occupied. This specification of an atomic state is called the electronic configuration, and it is usually denoted using the spectroscopic notation given in Sec. 1. For instance, $1s^2 2s^2 2p^3$ means that there are two electrons each in the $1s$ and $2s$ shells and three electrons in the $2p$ shell.

4.2 The "Building-Up" Principle

According to the exclusion principle, there can be at most $2(2l+1)$ electrons in each shell nl. A shell for which all of the single-electron states is occupied is said to be filled or closed. A partially filled shell is said to be open.

For a particular atom, one would expect to get the lowest energy by successively filling the lowest-lying electron shells. The configuration for the ground state would thus consist of a certain number of filled shells and at most one open shell. This leads to the historically important "building-up" or *Aufbau* principle, which was originally suggested by Bohr to explain the periodic table.

There are a number of departures from this simple description. As more electrons are added, the central field changes, and also the electron–electron interaction for particular electronic states causes deviations from the central-field description that are sufficiently large to lead to irregularities in the filling of the electron shells. There are atoms having two open shells in the ground configuration. This occurs for some of the transition elements for which a d shell is being filled, and for the rare-earth and actinide elements for which an f shell is being filled. In such cases, two open shells may have orbital energies that lie so close to each other that small departures from the central-field model can be sufficient to change the order in which the shells are filled.

4.3 Russell-Saunders Coupling

The central-field model leads to highly degenerate energy levels. For light and medium-size atoms, this degeneracy is removed mainly by the Coulomb interaction among the electrons, which splits the states of an electronic configuration into levels characterized by the quantum numbers of the total spin and orbital angular momentum. An illustration of the electrostatic splitting of an electronic configuration is given in Fig. 2, which shows the energy levels of an np^2 configuration. Each of the levels corresponds to definite values of the total angular-momentum quantum numbers S and L. The lowest level, which is denoted ${}^3P$, has a total spin quantum number $S=1$ and a quantum number of the total orbital angular momentum $L=1$. As is customary, the value of $2S+1$ is given for this level rather than the

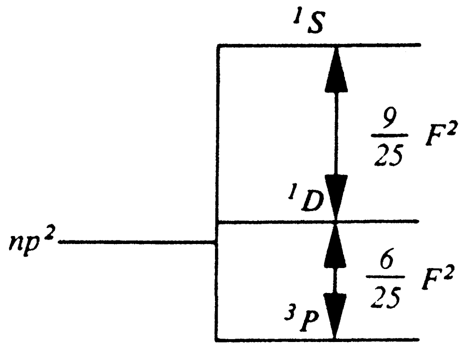

FIG. 2. Electrostatic splitting of an np^2 configuration. The quantity F^2 is defined in Sec. 4.4.

spin quantum number itself. The next higher level in Fig. 2, which is denoted ^1D, has total spin and orbital quantum numbers $S=0$ and $L=2$, while for the ^1S level, $S=0$ and $L=0$.

In order to form states that have definite values of the total angular momenta, it is generally necessary to form linear combinations of product functions or of Slater determinants. This process of forming eigenstates of the total-angular-momentum operators is referred to as addition or coupling of angular momenta. In order to discuss angular-momentum coupling in general terms, we consider now a system composed of two subsystems that are described by the states $|j_1m_1\rangle$ and $|j_2m_2\rangle$. For a composite system of this kind, the total-angular-momentum operator $\mathbf{J}$ can be written $\mathbf{J}=\mathbf{j}(1)+\mathbf{j}(2)$, and the eigenfunctions of $\mathbf{J}^2$ can be expressed as linear combinations of the product functions:

$$|(j_1j_2)JM\rangle=\sum_{m_1,m_2}|j_1m_1,j_2m_2\rangle \times\langle j_1m_1,j_2m_2|JM\rangle, \qquad (39)$$

where j_1m_1 and j_2m_2 are the angular-momentum quantum numbers of the individual states. The coefficients $\langle j_1m_1,j_2m_2|JM\rangle$ are known as the Clebsch-Gordan or vector-coupling coefficients.

The angular momenta of the electrons can be coupled together in various ways. The coupling scheme illustrated in Fig. 2, for which the spin angular momenta of the electrons are coupled together to form a total S and the orbital angular momenta are coupled to form a total L, is called LS or Russell-Saunders coupling. The total spin and orbital angular momenta are then coupled together to form the total angular momentum J.

4.4 Partial-Wave Expansion of the Coulomb Interaction

As we have seen, the spatial part of the wave function of an electron moving in a central field can be written as a product of a radial function and a function of the polar angles. In order to take advantage of this, we use the result from classical physics that $1/r_{12}$ can be expanded in partial waves as

$$\frac{1}{r_{12}}=\sum_k \frac{r_<^k}{r_>^{k+1}}P_k(\cos\omega). \qquad (40)$$

Here $r_<$ denotes the lesser and $r_>$ the greater of the two radial distances r_1 and r_2. The coordinates of a two-electron system are shown schematically in Fig. 3. In this figure, the positions of the electrons are represented by the coordinate vectors $\mathbf{r}_1$ and $\mathbf{r}_2$, which are separated by an angle ω.

The partial-wave expansion can be written in a particularly simple form using spherical tensors. For tensor operator methods the reader is referred to the book by Lindgren and Morrison (1986). Defining a tensor C with components

$$C_q^k=\left(\frac{4\pi}{(2k+1)}\right)^{1/2}Y_{kq}(\theta,\varphi) \qquad (41)$$

and using elementary properties of the spherical harmonics, the Legendre polynomial $P_k(\cos\omega)$ can be written

$$P_k(\cos\omega)=\mathbf{C}^k(1)\cdot\mathbf{C}^k(2). \qquad (42)$$

Equation (40) then becomes

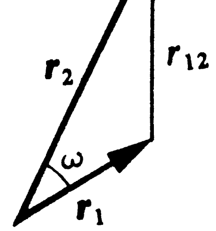

FIG. 3. The coordinate vectors of a two-electron system.

$$\frac{1}{r_{12}} = \sum_k \frac{r_<^k}{r_>^{k+1}} \mathbf{C}^k(1)\cdot\mathbf{C}^k(2). \tag{43}$$

Each term that appears in this expansion has a radial part that is represented by $r_<^k/r_>^{k+1}$ and an angular part that is represented by the scalar product of two spherical tensors. The fact that the two tensor operators are coupled together to form a scalar corresponds to the fact that the Coulomb interaction is invariant with respect to a simultaneous rotation of the coordinates of the two electrons.

As we have seen in Sec. 1, the wave functions for electrons moving in a central field also have separate radial and angular parts. For this reason, the terms in the partial-wave expansion of the Coulomb matrix elements for a two-electron system can be expressed as a product of a radial integral called a Slater integral and an angular term, which is the matrix element of the scalar product of two spherical tensor operators:

$$\begin{aligned}\left\langle n_a l_a m_s^a m_l^a, n_b l_b m_s^b m_l^b \right| \frac{1}{r_{12}} & \\ \times \left| n_c l_c m_s^c m_l^c, n_d l_d m_s^d m_l^d \right\rangle & \\ = \sum_k R^k(ab,cd)\,\delta(m_s^a,m_s^c)\,\delta(m_s^b,m_s^d) & \\ \times \langle l_a m_l^a, l_b m_l^b | \mathbf{C}^k(1)\cdot\mathbf{C}^k(2) | l_c m_l^c, l_d m_l^d \rangle. &\end{aligned} \tag{44}$$

Using our previous notation for the wave function of a particle moving in a central field, the two-electron Slater integral is defined as

$$R^k(ab,cd) = \int P_a(r_1)P_b(r_2) \times \frac{r_<^k}{r_>^{k+1}} P_c(r_1)P_d(r_2)\,dr_1dr_2. \tag{45}$$

For the special cases in which the radial wave functions of the two states on the right are the same as the two states on the left, it is convenient to define direct and exchange integrals

$$F^k(a,b) = R^k(ab,ab), \tag{46}$$

$$G^k(a,b) = R^k(ab,ba). \tag{47}$$

4.5 Antisymmetric Wave Functions for Two Electrons in *LS* Coupling

For two electrons with quantum numbers nl and $n'l'$, the antisymmetric wave function in LS coupling is

$$\begin{aligned}|\{nln'l'\}SM_SLM_L\rangle & \\ = F[\,|(nl)_1(n'l')_2,SM_SLM_L\rangle & \\ - |(nl)_2(n'l')_1,SM_SLM_L\rangle, &\end{aligned} \tag{48}$$

where F is a normalization factor. An interchange of the subscripts 1 and 2, which refer to the coordinates of the electrons, causes the state on the right to change sign. The curly brackets that appear in the state on the left constitute a special notation that we will use to specify an antisymmetric combination.

Equation (48) can be cast into another form by taking advantage of the special symmetry properties of the vector-coupling coefficients with respect to an interchange of the two constituent angular momenta,

$$\begin{aligned}\langle j_2m_2, j_1m_1 | JM\rangle & \\ = (-1)^{j_1+j_2-J}\langle j_1m_1, j_2m_2 | JM\rangle. &\end{aligned} \tag{49}$$

We may use this relation to reverse the order of the angular momenta in the second term that appears on the right. In doing this, we must reverse the coupling of the two spins as well as the two orbital angular momenta. This gives the two phase factors $(-1)^{S+1}$ and $(-1)^{l+l'+L}$. Equation (48) then becomes

$$\begin{aligned}|\{nln'l'\rangle,SM_SLM_L\rangle & \\ = F[\,|(nl)_1(n'l')_2,SM_SLM_L\rangle & \\ + (-1)^{l+l'+S+L} & \\ \times |(n'l')_1(nl)_2,SM_SLM_L\rangle]. &\end{aligned} \tag{50}$$

If $n=n'$ and $l=l'$, the two electrons are said to be equivalent. In this case, the right-hand side of Eq. (50) reduces to

$$F[1+(-1)^{S+L}]\,|(nl)_1(nl)_2,SM_SLM_L\rangle.$$

This will be nonvanishing only if $S+L$ is even, and hence only those functions for which $S+L$ is even correspond to physical states. The normalization factor F is then equal to $\frac{1}{2}$ and the normalized, antisymmetric state becomes

$$|\{nl^2\}SM_SLM_L\rangle = |(nl)_1(nl)_2,SM_SLM_L\rangle, \quad S+L \text{ even}. \tag{51}$$

Thus if $S+L$ is even, ordinary vector coupling leads in this case to proper antisymmetric wave functions.

If, on the other hand, the two electrons are not equivalent, the two states that appear in Eq. (50) are orthogonal, and F is $\sqrt{2}$.

4.6 The Coulomb Interaction for Two-Electron Systems in *LS* Coupling

We first consider the case of two equivalent electrons. Using the partial-wave expansion of the Coulomb interaction and Eq. (51), we obtain

$$\langle\{nl^2\}SL|1/r_{12}|\{nl^2\}SL\rangle = \sum_k f_kF^k(nl,nl), \tag{52}$$

where f_k is given by the formula

$$\begin{aligned} f_k &= \langle (nl)_1(nl)_2,SL|\mathbf{C}^k(1)\cdot\mathbf{C}^k(2) \\ &\quad \times |(nl)_1(nl)_2,SL\rangle \\ &= \langle l_1l_2,L|\mathbf{C}^k(1)\cdot\mathbf{C}^k(2)|l_1l_2,L\rangle, \end{aligned} \tag{53}$$

and where the two-electron Slater integral $F^k(nl,nl)$ was defined in Sec. 4.4. The Coulomb interaction is a scalar operator in both the spin and orbital spaces. For this reason, the Coulomb matrix elements are independent of M_S and M_L, and we have left out these quantum numbers.

Explicit formulas for evaluating the coefficients f_k have been given by Condon and Shortley (1959) and by Cowan (1981). At this point, we shall only cite the results for the np^2 configuration that we have considered previously. As we have seen, this configuration has three *LS* terms: ^1S, ^1D, and ^3P. The Coulomb matrix elements for these states are

$$\langle np^2\ {}^1S|1/r_{12}|np^2\ {}^1S\rangle = F^0+\tfrac{2}{5}F^2, \tag{54}$$

$$\langle np^2\ {}^1D|1/r_{12}|np^2\ {}^1D\rangle = F^0+\tfrac{1}{25}F^2, \tag{55}$$

$$\langle np^2\ {}^3P|1/r_{12}|np^2\ {}^3P\rangle = F^0-\tfrac{1}{5}F^2. \tag{56}$$

From these equations, we obtain the *LS* levels illustrated in Fig. 2. Each level has a multiplicity corresponding to the number of possible M_S and M_L values to which it corresponds. For example, the multiplicity of the ^1D state is 5 since $M_S=0$ while $M_L=-2,-1,\ldots,2$. A weighted average of the energy levels produces

$$\begin{aligned} E_{\text{av}} &= \tfrac{1}{15}[1E(^1S)+5E(^1D)+9E(^3P)] \\ &= F^0-\tfrac{2}{25}F^2. \end{aligned} \tag{57}$$

As before, we may use the explicit form of the wave function given in Sec. 4.5 and the symmetry properties of the vector-coupling coefficients to obtain the following general expression for the Coulomb matrix elements of a configuration of two nonequivalent electrons:

$$\begin{aligned} &\langle\{nln'l'\}SL|1/r_{12}|\{nln'l'\}SL\rangle \\ &\quad = \sum_k [f_kF^k(nl,n'l')+g_kG^k(nl,n'l')], \end{aligned} \tag{58}$$

where

$$\begin{aligned} f_k &= \langle l_1l_2',L|\mathbf{C}^k(1)\cdot\mathbf{C}^k(2)|l_1l_2',L\rangle, \\ g_k &= (-1)^{l+l'+S+L}\langle (nl)_1(n'l')_2,SL|\mathbf{C}^k(1) \\ &\quad \cdot\mathbf{C}^k(2)|(n'l')_1(nl)_2,SL\rangle. \end{aligned} \tag{59}$$

From these expressions, one can make some general, qualitative observations about the direct and exchange interactions. The coefficients f_k, which determine the character of the direct interaction, depend only upon the orbital quantum numbers l, l', and L. One can describe this situation by saying that the direct part of the Coulomb interaction couples the two orbital angular momenta l and l' together to form L. The two orbital momenta are also coupled by the exchange interaction, but this effect is usually quite small compared to the direct coupling. The exchange coefficients g_k also depend upon the total spin quantum number S, and so the exchange interaction couples the two spins.

A particularly simple example of a configuration of two nonequivalent electrons is provided by an ls configuration ($l'=0$). In this case, Eq. (58) becomes simply

$$\begin{aligned} &\langle\{nln's\}SL|1/r_{12}|\{nln's\}SL\rangle \\ &\quad = F^0(nl,n's)+(-1)^S\frac{G^l(nl,n's)}{2l+1}. \end{aligned} \tag{60}$$

This configuration has two *LS* terms: a triplet term with $S=1$ (^3l) and a singlet term with $S=0$ (^1l). As shown in Fig. 4, these two terms are split by an amount $2G^l(nl,n's)/(2l+1)$ by

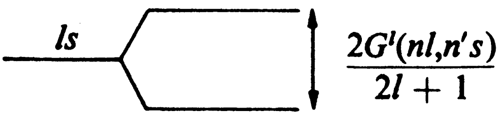

FIG. 4. Splitting of the *ls* state in *LS* coupling.

the exchange part of the Coulomb interaction. The splitting of the two spin states is a consequence of the antisymmetry of the wave function.

4.7 Coupling Schemes for Two-Electron Systems

In the previous sections, we have considered the *LS* coupling scheme for two-electron systems in some detail. We would now like to review from a general point of view the different coupling schemes that may be used to describe a configuration of two electrons. As we shall see, the coupling scheme that best describes the configuration and the general appearance of the energy levels depends upon the relative magnitude of the Coulomb and spin-orbit interactions. Our discussion here owes much to the work of Cowan and Andrew (1965).

There are four angular momenta associated with a two-electron system—$\mathbf{s}_1,\mathbf{s}_2,\mathbf{l}_1,\mathbf{l}_2$—and we begin by considering the interactions involving pairs of these angular momenta. As we have seen in Sec. 4.6, the direct part of the Coulomb interaction couples the two orbits together, and the exchange part of the Coulomb interaction couples the two spins. The magnetic interactions are responsible for coupling the spin and orbital motions together. For all but the lightest atoms, the magnetic interactions can usually be represented quite well by an effective spin-orbit operator, which for a two-electron configuration is

$$H_{\mathrm{so}}=\zeta_1\mathbf{s}_1\cdot\mathbf{l}_1+\zeta_2\mathbf{s}_2\cdot\mathbf{l}_2\,. \tag{61}$$

The one-electron spin-orbit operator $\zeta\mathbf{s}\cdot\mathbf{l}$ can be written

$$\zeta\mathbf{s}\cdot\mathbf{l}=\tfrac{1}{2}\,j(\mathbf{j}^2-\mathbf{s}^2-\mathbf{l}^2), \tag{62}$$

where $\mathbf{j}$ is the total angular momentum operator for one electron, $\mathbf{j}=\mathbf{s}+\mathbf{l}$. The effect of the spin-orbit operator upon a coupled single-electron state is

$$\langle(sl)j|\zeta\mathbf{s}\cdot\mathbf{l}|(sl)j\rangle=\tfrac{1}{2}\,j[j(j+1) -s(s+1)-l(l+1)]. \tag{63}$$

Using this relation, it is easy to show that the spin-orbit operator splits the energy levels of the electrons in the way shown in Fig. 5. This situation can be described by saying that the spin-orbit interaction couples the spin and the orbit of each individual electron.

The way in which the various angular momenta are coupled is illustrated in Fig. 6. The spin-orbit interaction couples the spin and the orbital angular momenta of each electron. The direct part of the Coulomb interaction couples the two orbits, and the exchange interaction couples the two spins.

For the low-*Z* elements the electrostatic interactions are predominant, which means that both the direct and the exchange interactions are larger than the spin-orbit interactions. In this case the coupling of the four angular momenta is usually best described by *LS* coupling,

$$|(ss')S,(ll')L\rangle. \tag{64}$$

For the heavy elements the spin-orbit interaction is comparable to or sometimes larger than the electrostatic interactions, and the coupling approaches the *j-j* limit. The states in this coupling scheme are of the form

$$|(sl)j,(s'l')j',J\rangle. \tag{65}$$

Both the *LS* and *j-j* coupling schemes are symmetric. The angular momenta of each electron are coupled together in the same way. These coupling schemes are usually best for the ground or low-lying configurations. Then the wave functions of the two elements have about the same spatial extent, and the interactions of the electrons are about the same.

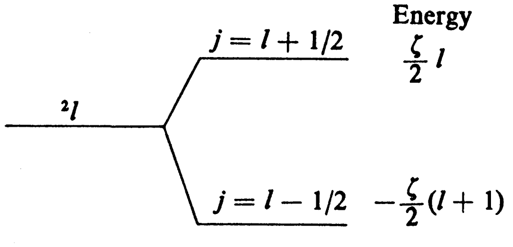

FIG. 5. Spin-orbit splitting of the single-electron energy levels.

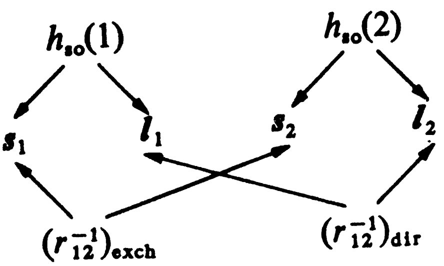

FIG. 6. Coupling of the angular momenta by the Coulomb and spin-orbit interactions.

When one of the outer electrons is more highly excited, and particularly when it has a high value of l, the configuration is highly asymmetric in the two electrons, and the symmetric couplings are usually no longer a good approximation. The highly excited electron, to which we give quantum numbers $n's'l'$, usually has a very small spin-orbit interaction. The wave functions of the two electrons are localized in different regions of space, and so the exchange interaction will also be quite small. Under these circumstances, the spin of the excited electron is weakly coupled to the other angular momenta, and the energy levels of the atom will occur in pairs with the J values of the levels of each pair differing by unity. How the angular momenta are coupled together in this case depends upon the relative magnitude of the direct interaction and the spin-orbit interaction of the inner electron. If the direct interaction is larger, then a third type of coupling is appropriate:

$$|((ll')L,s)K,s',J\rangle, \tag{66}$$

where K is the quantum number associated with an intermediate angular momentum obtained by coupling L and s together. Following Cowan and Andrews, we shall call this LK coupling, since in this case both L and K are fairly good quantum numbers.

For highly excited states, however, the direct interaction may become smaller than the spin-orbit interaction of the inner valence electron. We will then have another kind of coupling, called jK coupling:

$$|((sl)j,l')K,s'J\rangle. \tag{67}$$

In each case the angular momenta used to denote the coupling scheme are good quantum numbers to the extent that the interactions that are responsible for the coupling dominate over the other interactions of the two-electron system.

4.8 Energy-Level Structure of Many-Electron Atoms

As we have seen, the Coulomb interaction among the electrons splits a configuration up into a number of LS terms. This effect can be understood by considering the expression for the energy. The energy is given by the expectation value of H,

$$E=\langle\gamma SLM_SM_L|H|\gamma SLM_SM_L\rangle, \tag{68}$$

where

$$H=\sum_{i}^{N}\left(-\frac{1}{2}\nabla_i^2-\frac{Z}{r_i}\right)+\sum_{i<j}\frac{1}{r_{ij}} \tag{69}$$

is the Hamiltonian of the system and γ denotes the additional quantum numbers necessary to specify the many-electron state.

We can evaluate the term energies once we have formed proper LS functions. There are two general approaches that one can use to form antisymmetric LS wave functions. The first method is to express the LS wave function as a linear combination of Slater determinants. The expectation value of the energy can then be evaluated using formulas such as those given in Eqs. (10) and (12). The other general approach is to build up LS states with the proper symmetry from coupled two-electron states using the technique of fractional parentage. For the definition and use of fractional parentage the reader is referred to the books by Judd (1963) and Lindgren and Morrison (1986). The approach has the advantage that it enables one to make use of powerful tensor operator techniques for evaluating matrix elements. Whichever method one uses, the expression for the matrix elements of the Hamiltonian operator consists of a number of one-electron matrix elements involving the kinetic and nuclear-potential energy of the electrons and a number of two-electron matrix elements involving the Coulomb interaction among the electrons. As an example, we give here the expectation value of the energy for the carbon atom, which has a $1s^22s^22p^2$ configuration:

$$\begin{aligned} E &= \langle H \rangle \\ &= 2I(1s)+2I(2s)+2I(2p)+F^0(1s,1s) \\ &\quad +F^0(2s,2s)+4F^0(2s,1s)-2G^0(2s,1s) \\ &\quad +4F^0(2p,1s)-\tfrac{2}{3}G^1(2p,1s)+4F^0(2p,2s) \\ &\quad -\tfrac{2}{3}G^1(2p,2s)+F^0(2p,2p) \\ &\quad +\tfrac{1}{25}\,aF^2(2p,2p), \end{aligned} \tag{70}$$

where the single-electron integrals $I(nl)$ are defined as

$$I(nl)=\int_0^\infty P_{nl}(r)\left(-\frac{1}{2}\frac{d^2}{dr^2}+\frac{l(l+1)}{2r^2}-\frac{Z}{r}\right)P_{nl}(r)dr. \tag{71}$$

The coefficient $a=10,1,-5$ for the ^1S, ^1D, and ^3P states, respectively, and $a=-2$ for the weighted average of these states.

4.9 Hartree-Fock Model

In this section, we shall optimize the description within the general framework of the central-field model by choosing the radial part of the orbitals, $(P/r)Y_{lm_l}\chi_{m_s}$, so that the energy is minimized. This energy could either be that of a particular LS term or the average energy of a configuration. As before, the radial equations for the Hartree-Fock (HF) procedure are derived by requiring that the expectation value of the total energy be stationary for small changes in the radial part of the orbitals. The orbitals are required to be of the above form, to be orthonormal, and also to satisfy the same boundary conditions as in the one-electron case.

This leads as in Sec. 3.1 to a coupled set of differential equations known as the Hartree-Fock equations. Rather than go through this variational procedure again with central-field orbitals, we note that there is a one-to-one correspondence between the terms in the expression for the energy and the left-hand side of the resulting HF equations. The HF equation for a particular orbital can be obtained immediately by making the following replacements in the energy terms that depend on that orbital:

$$I_a \to \frac{1}{q_a}\left(-\frac{1}{2}\frac{d^2}{dr^2}+\frac{l_a(l_a+1)}{2r^2}-\frac{Z}{r}\right)P_a(r), \tag{72}$$

$$F^k(a,b) \to \frac{1+\delta(a,b)}{q_a}\frac{1}{r}\,Y^k(bb,r)P_a(r), \tag{73}$$

$$G^k(a,b) \to \frac{1+\delta(a,b)}{q_a}\frac{1}{r}\,Y^k(ab,r)P_b(r). \tag{74}$$

Here q_a is the occupation number of shell a, and the functions $Y^k(ab,r)$ are defined by

$$Y^k(ab,r)=r\int_0^\infty \frac{r_<^k}{r_>^{k+1}}P_a(r')P_b(r')dr'. \tag{75}$$

As an example, we shall obtain the HF equation for the $2p$ orbital of the carbon atom. The terms in the energy expression for carbon that depend upon the $2p$ orbital are

$$\begin{aligned} E|_{2p} &= 2I(2p)+4F^0(2p,1s)-\tfrac{2}{3}G^1(2p,1s) \\ &\quad +4F^0(2p,2s)-\tfrac{2}{3}G^1(2p,2s)+F^0(2p,2p) \\ &\quad +\tfrac{1}{25}\,aF^2(2p,2p), \end{aligned} \tag{76}$$

where as before

$$a=10,1,-5,\ \text{and}\ -2 \tag{77}$$

for the ^1S, ^1D, ^3P, and the configuration average, respectively. Making the replacements given above with $q_a=2$ and $l_a=1$, we obtain

$$\begin{aligned} &\left(-\frac{1}{2}\frac{d^2}{dr^2}+\frac{1}{r^2}-\frac{Z}{r}+\frac{2}{r}\,Y^0(1s1s,r)+\frac{2}{r}\,Y^0(2s2s,r)\right. \\ &\quad \left.+\frac{1}{r}\,Y^0(2p2p,r)+\frac{a}{25}\frac{1}{r}\,Y^2(2p2p,r)\right) \\ &\quad \times P_{2p}(r)-\frac{1}{3}\frac{1}{r}\,Y^1(2p1s,r)P_{1s}(r) \\ &\quad -\frac{1}{3}\frac{1}{r}\,Y^1(2p2s,r)P_{2s}(r) \\ &= \epsilon_{2p}P_{2p}(r). \end{aligned} \tag{78}$$

A good, systematic treatment of the Hartree-Fock method has been given by Fischer (1986).

4.10 Properties of Hartree-Fock Orbitals

The Hartree-Fock method provides a qualitative understanding of some of the general features of atoms. In the remaining part of this section, we shall use it to discuss some of the properties of atomic orbitals. In later sections of this article, we will consider correlation effects that cause departures from this single-particle description.

4.10.1 Radial Shell Dependence. The Hartree-Fock wave function for a single electron has the same angular dependence and the same nodal structure as the corresponding hydrogenic wave function. The most important difference between these two functions is due to the fact that the screening of the nucleus increases in the HF model as the electron moves away from the nucleus, while for a hydrogen atom it remains the same.

If one neglects exchange, the radial Hartree-Fock equations for a closed-shell atom can be written in the form

$$\left(-\frac{1}{2}\frac{d^2}{dr^2}+\frac{l(l+1)}{2r^2}-\frac{Z-\sigma(r)}{r}\right)P_{nl}(r)$$
$$=\epsilon_{nl}P_{nl}(r), \tag{79}$$

where the effect of all the other electrons in the atom has been incorporated into an r-dependent screening parameter, $\sigma(r)$. To the extent that the exchange terms can be approximated by a local potential, their effect can also be included in the screening. So this equation frequently provides a surprisingly good representation of the Hartree-Fock approximation.

It is the strong radial dependence of the screening that is mainly responsible for the departure from hydrogenic behavior. As an example, we consider the neutral argon atom. Near the nucleus, a $3s$ electron "sees" an effective charge of very nearly $Z=18$; however, the effective charge decreases continuously from the nucleus to the outermost reaches of the atom. At a distance of about 0.3–0.5 Bohr radii, the nucleus is almost fully screened by the two K-shell electrons and partially screened by the L-shell electrons, so that the effective charge has been reduced by about 5–6 units. Still farther out, beyond one Bohr radius, the nucleus is almost completely screened by the eight L-shell electrons as well as the K shell. As a result of this change, the orbital wave functions are far from hydrogenic.

This is illustrated in Fig. 7 in which the Hartree-Fock $3p$ orbital function for the ground state of neutral argon is shown. In the upper part of the figure, the Hartree-Fock wave function is compared with a hydrogenic $3p$ function with a value of Z chosen so that the function approximately matches the HF function in the tail region. This corresponds to a nuclear charge of 6. The bottom part of the figure compares the same Hartree-Fock function with another hydrogenic function for which the nodal points match. The outer loop of this second hydrogenic function, which corresponds to a nuclear charge of 12, completely fails to represent the Hartree-Fock function in the tail region. By comparing these two figures, one can see that as one moves inward from large r, the effective field becomes continuously stronger, drawing the wave function in towards the nucleus.

Because of the contraction of atomic orbitals as the nuclear charge increases, all atoms in the same column of the periodic table differ in size by no more than a factor of 2 or 3. The size of atoms is described by the one-electron charge-density distribution, which is obtained by integrating the total charge density over all but one of the radial coordinates:

$$\rho(r)=\int d\Omega\int dV_2\cdots$$
$$\times\int dV_N|\Psi(\mathbf{r},\mathbf{r}_2,\cdots,\mathbf{r}_N)|^2. \tag{80}$$

The charge-density distribution of three noble-gas atoms is illustrated in Fig. 8. In the figure, the horizontal scale is the same for the three atoms, but the vertical scale increases with Z because of the increase in the number of electrons. The shell structure of the atoms is clear and quite pronounced, but with the inner shells contracting closer to the nucleus with higher nuclear charge. The valence shell of krypton ($n=4$) is less than twice as far out as the valence shell ($n=2$) of neon, even though the nuclear charge increases from 10 to 36.

4.10.2 *LS* Term Dependence. As we have seen, the Hartree-Fock equations are determined by requiring that the expectation value of the total energy E be stationary with respect to small changes in the form of the radial wave functions. The HF equations should, in principle, be different for each LS term, just as the expectation value of the energy is different; however, it has become traditional since the earliest days of quantum mechanics to treat the different LS terms as merely different angular-momentum couplings of the same set of atomic orbitals. This artifice can be accommodated within the framework of the Hartree-Fock approximation by apply-

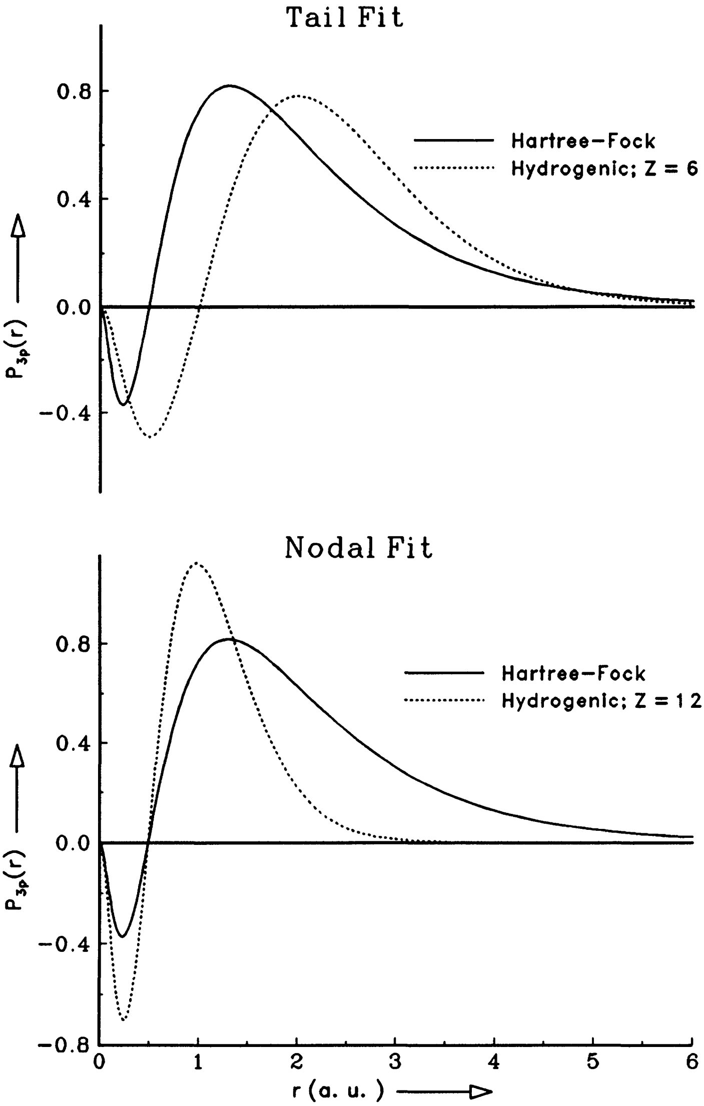

FIG. 7. Hartree-Fock and scaled hydrogenic $3p$ orbitals for the ground state of argon, $3s^2 3p^6\ {}^1S$, from the kinetic energy operator.

ing the variational principle to the weighted average of the total energies of all these states, with the degeneracies $(2S+1)(2L+1)$ as the weight factors. This leads to a set of equations for the orbitals of the center of gravity, or average energy, of the configuration and does in fact variationally define a set of orbital functions common to all states of the configuration.

The configuration-average approach frequently produces sufficiently accurate results. It is also much easier to calculate matrix elements between different states of a configuration if one has a common set of orbitals. There are cases, however, where the Hartree-Fock radial functions are quite strongly dependent on the particular LS term of the configuration. The best known example of such a term dependence was discovered by Hartree and Hartree in 1936 for the $1s^2 2s2p$ configuration of the beryllium atom, which gives rise to two terms, the triplet and singlet P states ${}^3P$ and ${}^1P$. The total energy for these states is given by

$$
\begin{aligned}
E(^{1,3}P) = {} & 2\langle 1s|h|1s\rangle + \langle 2s|h|2s\rangle + \langle 2p|h|2p\rangle + \langle 1s1s|1/r_{12}|1s1s\rangle + 2\langle 1s2s|1/r_{12}|1s2s\rangle \\
& - \langle 1s2s|1/r_{12}|2s1s\rangle + 2\langle 1s2p|1/r_{12}|1s2p\rangle - \langle 1s2p|1/r_{12}|2p1s\rangle \\
& + \langle 2s2p|1/r_{12}|2s2p\rangle \pm \tfrac{1}{2}\langle 2s2p|1/r_{12}|2p2s\rangle, \qquad (81)
\end{aligned}
$$

where h is the one-electron hydrogenic Hamiltonian. The upper sign in this expression refers to the singlet state and the lower sign to the triplet. The opposite sign of the exchange integral not only splits the term energies; it also influences the potentials of the Hartree-Fock radial equations for the $2s$ and $2p$ orbitals. The combined effect of the repulsive centrifugal and exchange potentials in the $2p$ equation of the ^1P state makes the $2p$ function for this state dramatically more diffuse than that for the triplet. The two $2p$ functions together with the $2s$ of the ^3P state are illustrated in Fig. 9. The $2s$ function is not very much different for the two states and thus serves as a convenient reference point. The mean value of r for the $2p$ orbital is 2.88 a.u. for the ^3P state and 5.02 for the ^1P state, and the maximum is twice as far out for the singlet state as the triplet.

This pronounced term dependence is closely related to the existence of a singly occupied orbital with the same principal quantum number, which gives rise to a large exchange interaction. The $3s3p$ configuration of magnesium also displays a striking singlet–triplet term dependence, as do the corresponding configurations of calcium and zinc, and the ^3P and ^1P terms of the $3p^5 3d$ configurations of singly ionized potassium. This last case is due in part to the very large angular factor for the exchange in the ^1P matrix element.

4.10.3 Shell Collapse. The case of singly ionized potassium, which we have just mentioned, provides information about another important atomic property. For light atoms, the $3d$ orbital is usually quite diffuse and hydrogenic; however, as the nuclear attraction becomes strong in comparison to the centrifugal repulsion, the d orbital is drawn into the core of the atom. In the region of the periodic table around potassium the $3d$ or-

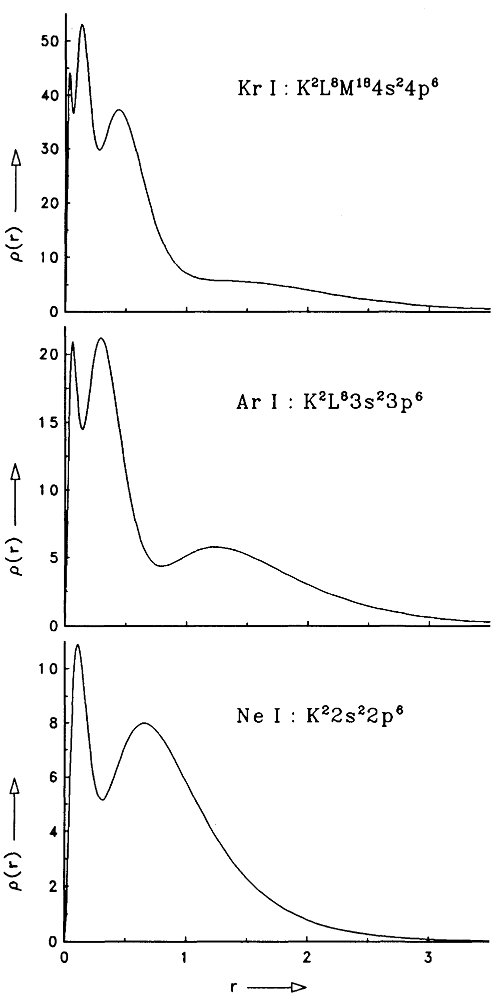

FIG. 8. Hartree-Fock radial charge-density distributions for the ground states of the noble gases Ne, Ar, and Kr. Note the change of vertical scale for the different atoms.

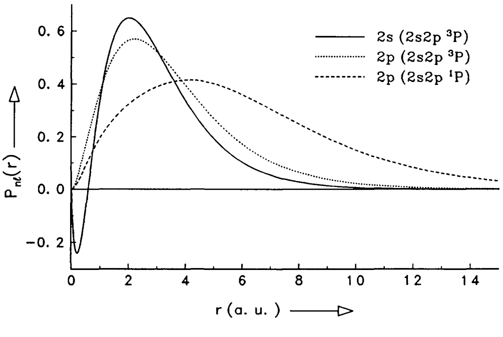

FIG. 9. Hartree-Fock orbitals for the $2s2p$ configuration of the beryllium atom.

bital "teeters" on the verge of collapse, giving rise to the sensitive term dependence just discussed. As the nuclear charge increases more the *3d* orbital collapses and becomes quite similar in spatial extent to the already filled *3s* and *3p* orbitals. In Fig. 10, we show the behavior of the mean radii of the Hartree-Fock orbitals for the elements potassium through copper. For potassium, we quote results for the two different states $3p^6 4s$ and $3p^6 3d$ and for calcium for $3p^6 3d 4s\ ^3D$. All others refer to the lowest term of the configuration $3p^6 3d^n 4s^2$, which, with the exception of chromium, represents the ground state of the atom.

In a similar manner, the *4d* shell begins to collapse around rubidium ($Z=37$), producing the second transition-metal row of the periodic table. The *5d* shell collapses shortly after cesium ($Z=55$), giving rise to the third transition-metal row.

A still more dramatic shell collapse occurs shortly beyond cesium, when the much larger centrifugal repulsion for $l=3$ is finally overcome by the attractive nuclear field, and the *4f* shell collapses. In this case, the collapse places the *4f* orbital well inside the $5s^2 5p^6$ shells. The progressive filling of this collapsed *4f* shell gives rise to the rare-earth elements, which are well known for their nearly identical chemical properties.

Some descriptive material on the properties of atomic orbitals can be found in the book by Cowan (1981) and in the review articles by Weiss (1970,1973).

5. CORRELATION IN ELECTRONIC STRUCTURE THEORY

5.1 Nature of Electron Correlation

In assigning each electron to its own single-particle state, independent-particle models implicitly treat the interaction among the electrons in an average fashion, thereby neglecting the detailed manner in which the motion of each electron is correlated with the motions of the other electrons. The error this introduces into the calculation is called the correlation error. It has become conventional to treat the Hartree-Fock model as the standard independent-particle model and to define the correlation contribution to any property as the difference between the value calculated with the exact wave function and that calculated in the Hartree-Fock approximation.

In the Hartree-Fock approximation, the wave function is written as an antisymmetrized product of single-particle functions. The product nature of the function implies that the many-electron problem is separable into a set of single-particle problems, a result that cannot be achieved for a system of interacting particles. For example, the spatial part of the Hartree-Fock function for the ground state of helium is

$$\Psi_{\mathrm{HF}}=\phi(r_1)\phi(r_2), \tag{82}$$

whereas the exact nonrelativistic Hamiltonian is given by

$$\mathscr{H}=-\frac{1}{2}\nabla_1^2-\frac{1}{2}\nabla_2^2-\frac{Z}{r_1}-\frac{Z}{r_2}+\frac{1}{r_{12}}\,. \tag{83}$$

It is the term $1/r_{12}$ that causes departures from the independent-particle description.

In order to satisfy the condition

$$(\mathscr{H}\Psi)/\Psi=E=\text{const.}, \tag{84}$$

the exact wave function must have a cusplike singularity at the nucleus:

$$\left(\frac{1}{\Psi}\frac{\partial\Psi}{\partial r_1}\right)_{r_1=0}=\left(\frac{1}{\Psi}\frac{\partial\Psi}{\partial r_2}\right)_{r_2=0}=-Z. \tag{85}$$

These conditions can be satisfied by a Hartree-Fock product function. However, the exact wave function must also satisfy the cusp condition

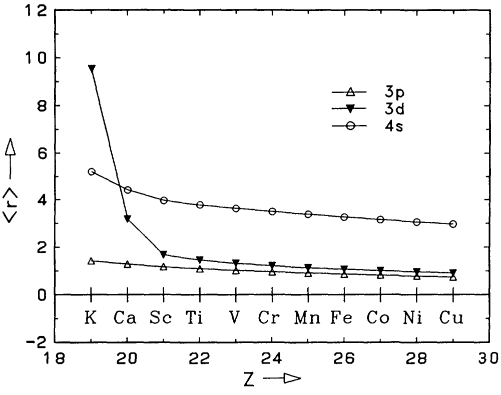

FIG. 10. Hartree-Fock orbital expectation values of *r* for the first transition elements.

$$\left(\frac{1}{\Psi}\frac{\partial\Psi}{\partial r_{12}}\right)_{r_{12}=0}=\frac{1}{2}, \tag{86}$$

which is always violated by any orbital product function, as illustrated schematically in Fig. 11. This is the root source of the correlation problem. All methods of calculating correlation corrections to the wave function can be regarded as attempts to account for this cusp condition.

The magnitude of the correlation correction to the total energy for selected first-row atoms and ions is given in Table 1. With the exception of the heliumlike systems, the energies labeled "exact" are estimates based on a combination of calculations and a careful evaluation of experimental data. The magnitude of the correlation energy is always a very small fraction of the total energy; however, it is generally comparable to the binding energy of the other electrons, which are responsible for the chemical and optical properties of atoms. Also, the changes in the wave function due to correlation may lead to more serious changes in other physical properties than it does for the energy itself.

5.2 The Hylleraas Method

The most accurate approach to the correlation problem was introduced by Hylleraas (1964). It involves explicitly introducing the interelectronic coordinate r_{12} into a variational trial function. Because of their extraordinary complexity, these calculations have been confined almost exclusively to two-electron atoms and molecules, i.e., helium and its isoelectronic ions and the hydrogen molecule. They provide benchmarks of accuracy by which all other approximations of electronic structure are judged.

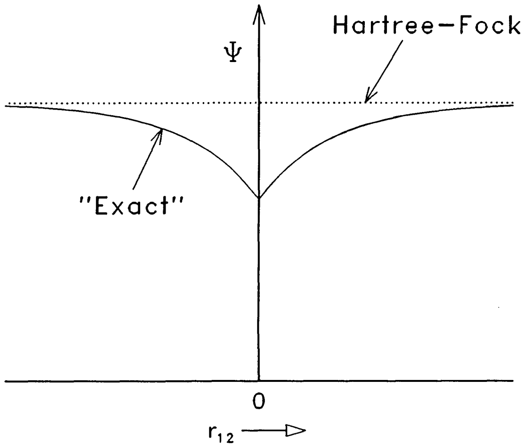

FIG. 11. Schematic behavior of Hartree-Fock and exact wave functions as $\mathbf{r}_{12}$ passes through zero, for $|r_1|=|r_2|$.

For the ground state, or any ^1S state, of a heliumlike atom, the trial function takes the form

$$\Psi=\left(\sum_{p,q,\mu} c_{pq\mu}\Phi_{pq\mu}(r_1,r_2,r_{12})\right)\times\frac{\alpha(1)\beta(2)-\beta(1)\alpha(2)}{\sqrt{2}}, \tag{87}$$

where the two-electron basis function has the general form

$$\Phi_{pq\mu}=(e^{-\zeta r_1-\zeta' r_2}r_1^p r_2^q+e^{-\zeta r_2-\zeta' r_1}r_2^p r_1^q)r_{12}^{\mu}, \tag{88}$$

ζ and ζ' being variational parameters. The total energy is then given by

$$E=\frac{\langle\Psi|\mathscr{H}|\Psi\rangle}{\langle\Psi|\Psi\rangle}=\frac{\Sigma_{i,j}c_i c_j\langle\Phi_i|\mathscr{H}|\Phi_j\rangle}{\Sigma_{i,j}c_i c_j\langle\Phi_i|\Phi_j\rangle}, \tag{89}$$

where the single indices stand for the $\{p,q,\mu\}$ triplets in Eq. (88). The variation of the energy with respect to the expansion coefficients is

$$\delta E=\frac{2}{\Sigma_{i,j}c_i\langle\Phi_i|\Phi_j\rangle c_j}\sum_i \delta c_i\left(\sum_j\langle\Phi_i|\mathscr{H}|\Phi_j\rangle c_j - E\sum_j\langle\Phi_i|\Phi_j\rangle c_j\right). \tag{90}$$

Since the variational principle requires $\delta E=0$ for any arbitrary variation of the coefficients δc_i, the quantity inside the large parentheses must vanish identically, leading to

$$\mathbf{Hc}=E\mathbf{Sc}, \tag{91}$$

where the matrix elements in this notation are defined as

$$(\mathbf{H})_{ij}=\langle\Phi_i|\mathscr{H}|\Phi_j\rangle,\quad (\mathbf{S})_{ij}=\langle\Phi_i|\Phi_j\rangle. \tag{92}$$

The calculation thus reduces to a matrix eigenvalue problem for the eigenvector of expansion coefficients. The nonlinear parameters ζ and ζ' can be chosen to ensure the correct asymptotic large-r behavior of the wave function,

Table 1. Total correlation energies (in a.u.) for selected first-row atoms and ions. [This table is taken essentially from Davis *et al.* (1991).]

Atom	Electron no.	E_{total} Exact	E_{total} Hartree-Fock	Correlation energy
He: $1s^2\ ^1S$	2	−2.903 72	−2.861 68	−0.042 04
Be^{2+}: $1s^2\ ^1S$	2	−13.655 57	−13.611 30	−0.044 26
Be: $1s^2 2s^2\ ^1S$	4	−14.667 36	−14.573 02	−0.094 34
C^{2+}: $1s^2 2s^2\ ^1S$	4	−36.534 90	−36.404 50	−0.126 40
B: $1s^2 2s^2 2p\ ^2P$	5	−24.653 93	−24.529 06	−0.124 87
C: $1s^2 2s^2 2p^2\ ^3P$	6	−37.844 99	−37.688 62	−0.156 37
N: $1s^2 2s^2 2p^3\ ^4S$	7	−54.589 25	−54.400 93	−0.188 32
O: $1s^2 2s^2 2p^4\ ^3P$	8	−75.067 2	−74.809 40	−0.257 8
F: $1s^2 2s^2 2p^5\ ^2P$	9	−99.733 8	−99.409 35	−0.324 4
Ne^{+}: $1s^2 2s^2 2p^5\ ^2P$	9	−128.144 3	−127.817 81	−0.326 5
Ne: $1s^2 2s^2 2p^6\ ^1S$	10	−128.938 8	−128.547 10	−0.391 7

$$\Psi_{r\to\infty} \to \exp(-Zr_1 - I^{1/2} r_2) + \exp(-Zr_2 - I^{1/2} r_1), \qquad (93)$$

where I is the exact ionization energy. They can also be taken to be purely variational parameters, adjusted to minimize the energy.

The Hylleraas approach, which has recently been extended to include several pairs of exponential parameters, can be applied in a straightforward fashion to excited states, to triplets, and to states of non-S symmetry. The energy spectrum of helium and heliumlike ions has been determined with a precision of 9–11 significant figures.

5.3 The Configuration-Interaction Method

Unlike the Hylleraas method, most approaches to the correlation problem make use of the general language of the independent-particle model. They are based on the multiconfiguration expansion of the wave function,

$$\Psi = c_0\Phi_0 + \sum_{a,i} c_a^i \Phi_a^i + \sum_{a,b}\sum_{i,j} c_{ab}^{ij}\Phi_{ab}^{ij} + \sum_{a,b,c}\sum_{i,j,k,} c_{abc}^{ijk}\Phi_{abc}^{ijk} + \cdots, \qquad (94)$$

where Φ_0 is normally taken to be the antisymmetric single-configuration function that most accurately describes the wave function. It is referred to as the reference state, or reference configuration, and it is usually taken to be the corresponding Hartree-Fock function. In this case, the remaining terms represent the correlation corrections. Φ_a^i denotes a symmetry-adapted antisymmetric function obtained from Φ_0 by substituting an excited orbital φ_i for orbital φ_a, occupied in Φ_0. Φ_{ab}^{ij} refers to the corresponding double-substitution configurations, and so forth. For some atomic and molecular states, the wave-function expansion converges much more rapidly if the reference function Φ_0 corresponds to a linear combination of configurations. The substitutions are then made into all of the configurations of this multiconfiguration referent.

If the Hamiltonian of the system can be represented by an approximate operator whose eigenfunctions are known, perturbation theory may be used to evaluate the coefficients in the expansion of the wave function. This approach will be discussed at some length below. The configuration-interaction (CI) method is implemented by requiring that the total energy be stationary with respect to small changes in the expansion coefficients. This leads to matrix eigenvalue equations of exactly the same form as that for the Hylleraas method. In this case, though, the matrix elements must be evaluated using the configuration-state functions described at the beginning of this section. In CI calculations, it is common to assume that the orbitals are orthonormal. The overlap matrix **S** is then a unit matrix.

A variety of strategies can be used to select the orbitals in such a way as to optimize the convergence of the expansion. The most common approach is to select a set of orthogonalized Slater-type functions or Gaussians and vary the exponential parameters to minimize the total energy. While the correction terms in the expansion of the wave function have the appearance of excited states, the correlation orbitals are generally much more compact than are the orbitals for realistic excited states. The unfettered application of the variational principle usually leads to a set of orbitals for which the correlation part of the wave function maximally overlaps the reference state.

The CI approach has been used to obtain results with an accuracy of 10 cm^{-1} or better for selected, simple atoms and molecules, utilizing expansions of 10^3–10^4 configurations. It has also been applied to the ground and excited states of a wide variety of atomic and molecular species using somewhat more truncated expansions, omitting classes of terms such as those involving correlations of the core electrons that are assumed to play a small role in valence-shell spectroscopy. Predicted excitation and ionization energies typically have an accuracy of the order of 10^3 cm^{-1}. The method has found its most extensive and intensive applications in the calculation of electronic structure of molecules, where it is not uncommon to find expansion lengths of the order of 10^5–10^6 terms. Molecular applications of CI are discussed in Sec. 8.

The principal drawback of CI arises from the rapid increase in the size of the calculation, not only as one saturates the virtual-orbital basis, but with the inclusion of higher-order substitution configurations.

5.4 Multiconfiguration Hartree-Fock Method

A natural extension of the CI method is to require that the energy be stationary with respect to simultaneous variations of *both* the CI coefficients and the orbitals. This leads to a coupled set of integro-differential equations analogous to the Hartree-Fock equations. In this case, however, the differential equation for each orbital depends upon the mixing coefficients that appear in the expansion of the wave function as well as upon the other orbitals. The conventional procedure for solving these equations is to iterate back and forth between the configuration-interaction secular equation and the integro-differential equations for the orbitals until consistency is achieved for both the orbitals and expansion coefficients.

As an example of the multiconfiguration Hartree-Fock (MCHF) procedure, we consider the ground state $1s^22s^2\ {}^1S$ of the neutral beryllium atom. It is well known that the $1s^22p^2\ {}^1S$ state mixes strongly with this state, and it is reasonable to suppose that the two-component multiconfiguration Hartree-Fock function,

$$\psi^{\text{MCHF}}=a|1s^22s^2\ {}^1S\rangle+b|1s^22p^2\ {}^1S\rangle,$$

provides a better description of the ground state than does the single-configuration wave function of the ordinary Hartree-Fock theory. Here the coefficients a and b give the weight of each configuration in the MCHF wave function. The energy is then

$$\begin{aligned}E=\,&a^2\langle 1s^22s^2\ {}^1S|H|1s^22s^2\ {}^1S\rangle\\&+2ab\langle 1s^22s^2\ {}^1S|H|1s^22p^2\ {}^1S\rangle\\&+b^2\langle 1s^22p^2\ {}^1S|H|1s^22p^2\ {}^1S\rangle.\end{aligned}\tag{95}$$

This expression assumes that

$$\int_0^\infty P_{nl}(r)P_{n'l'}(r)dr=\delta_{nn'},\quad a^2+b^2=1.\tag{96}$$

As in the single-configuration Hartree-Fock theory, we introduce a Lagrange multiplier for each constraint on the wave function. In this way, we obtain the energy functional

$$\begin{aligned}W=\,&E+\lambda_{1s,1s}\langle 1s|1s\rangle+\lambda_{1s,2s}\langle 1s|2s\rangle\\&+\lambda_{2s,2s}\langle 2s|2s\rangle+\lambda_{2p,2p}\langle 2p|2p\rangle.\end{aligned}\tag{97}$$

The requirement that this functional be stationary with respect to variations of the radial part of the wave functions leads to the MCHF equations for the $1s$, $2s$, and $2p$ orbitals,

$$\begin{aligned}LP_{1s}=\,&-\frac{2}{r}\,Y^0(1s,1s)P_{1s}+\frac{2a^2}{r}[2Y^0(2s,2s)P_{1s}\\&-Y^0(1s,2s)P_{2s}]+\frac{2b^2}{r}\bigg[2Y^0(2p,2p)P_{1s}\\&-\frac{1}{3}\,Y^1(1s,2p)P_{2p}\bigg]+\epsilon_{1s,1s}P_{1s}\\&+\epsilon_{1s,2s}P_{2s}\,,\end{aligned}\tag{98}$$

$$LP_{2s}=\frac{2}{r}[2Y^0(1s,1s)+Y^0(2s,2s)]P_{2s}$$

$$-\frac{2}{r}\,Y^0(1s,2s)P_{1s}$$

$$-\frac{2}{\sqrt{3}}\frac{b}{a}\frac{1}{r}\,Y^1(2s,2p)P_{2p}+\epsilon_{2s,2s}P_{2s}$$

$$+\epsilon_{2s,1s}P_{1s}\,, \tag{99}$$

$$LP_{2p}=\frac{2}{r}[2Y^0(1s,1s)+Y^0(2p,2p)$$

$$+\tfrac{2}{5}Y^2(2p,2p)]P_{2p}-\tfrac{2}{3}Y^1(1s,2p)P_{1s}$$

$$-\frac{2}{\sqrt{3}}\frac{a}{b}\,Y^1(2s,2p)P_{2s}+\epsilon_{2p,2p}P_{2p}\,, \tag{100}$$

where the operator L is defined as

$$L=\frac{d^2}{dr^2}+\frac{2Z}{r}-\frac{l(l+1)}{r^2}\,,$$

$Y^k(nl,n'l')$ are the functions introduced in Sec. 4.9, and the Lagrange multipliers in these equations are related to the coefficients in the energy functional.

The addition of the single-configuration $1s^22p^2$ to the wave function accounts for nearly half of the correlation energy of the beryllium atom. However, since the optimum-correlation $2p$ orbital is quite similar in radial extent to the $2s$, the added configuration has a fairly small effect upon the charge distribution of the atom.

Since the MCHF method determines the variationally optimum set of orbitals and CI coefficients simultaneously, it can provide a significantly more compact representation of correlation effects than the straightforward CI method. Relaxation of the occupied orbitals in the field of the correlation configurations is automatically taken into account. Indeed, since the orbitals seek whatever form is optimum for the multiconfiguration trial function, the distinction between occupied and excited orbitals can become somewhat blurred, which might well be exactly what is called for by the particular structure problem under consideration. For longer expansions, however, the rapid increase in the number of couplings of terms with smaller and smaller coefficients leads to convergence problems. In extensive MCHF calculations, convergence difficulties may lead one to freeze certain sets of orbitals while variationally determining the mixing coefficients and the remaining orbitals. Thus a calculation that began as a MCHF calculation may end as partially a CI calculation.

Just as with the Hartree-Fock approximation, one can expand the orbitals in a basis of analytical functions. The integro-differential equations then become matrix Hartree-Fock equations for the orbital expansion coefficients, which are, in turn, coupled with the secular equation for the CI coefficients. The expansion-method representation of the MCHF orbitals enables one to employ a radically different computational strategy. Since the total-energy expression is a function of the orbital coefficients, it is possible to develop algorithms for directly calculating numerical corrections to an initial guess of these coefficients, instead of going through the process of iteratively solving the Euler-Lagrange equations for the variational problem. In the area of atomic physics, a number of MCHF calculations have been reported recently that make use of a local basis of B splines. These calculations, which make use of the more direct iterative schemes just mentioned, also have the advantage that they are able to represent continuum and bound states together on the same footing.

5.5 Many-Body Perturbation Theory

As we have indicated previously, perturbation theory provides another way of calculating the effect of configuration interaction upon atomic and molecular properties. In this brief review of the applications of perturbation theory, we shall first describe the lowest-order corrections to the wave function and the energy of a closed-shell atom, and then indicate some of the different features that arise for higher-order treatments.

In order to implement a perturbative approach, the electrostatic Hamiltonian is separated into an approximate Hamiltonian and a perturbation, $H=H_0+V$. The approximate Hamiltonian H_0 is a sum of one-electron operators,

$$h_0(i)=-\tfrac{1}{2}\nabla_i^2-(Z/r_i)+u(r_i), \tag{101}$$

and the perturbation is

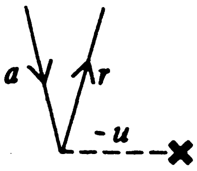

$$= \frac{|\alpha_a^r\rangle\langle r|-u|a\rangle}{\varepsilon_a-\varepsilon_r}$$

FIG. 12. Goldstone diagram that describes a process in which a single electron is excited from a core state by a potential interaction.

$$V=\sum_{i<j}^{N}\frac{1}{r_{ij}}-\sum_{i=1}^{N}u(r_i). \qquad (102)$$

A closed-shell state may be formed by constructing determinants of the eigenfunction of h_0. These states are eigenfunctions of H_0 corresponding to an eigenvalue that is the sum of the single-electron eigenvalues.

An account of the formalism of perturbation theory and its application to atomic systems has been given by Kelly (1969) and by Lindgren and Morrison (1986). In our notation, the first-order correction to the determinantal wave function $|\alpha\rangle$ is

$$\Psi^{(1)}=\sum_{\beta\neq\alpha}\frac{|\beta\rangle\langle\beta|V|\alpha\rangle}{E_0-E_0^{\beta}}. \qquad (103)$$

The perturbation V has a one-electron part that describes the interaction of the electrons with the average potential and a two-electron part that corresponds to the Coulomb interaction among the electrons. The one-body part of the perturbation gives the contribution

$$\Psi_U^{(1)}=\sum_{\beta\neq\alpha}\frac{|\beta\rangle\langle\beta|-U|\alpha\rangle}{E_0-E_0^{\beta}}, \qquad (104)$$

where

$$-U=-\sum_{i=1}^{N}u(r_i). \qquad (105)$$

Since a one-body operator can only join determinantal states that differ by a single electron, the sum over determinants only includes single-substitution states. Using this fact and equations for the matrix elements of one-body operators between determinantal states similar to those given in Sec. 2, this contribution to the first-order wave function can be written

$$\Psi_U^{(1)}=\sum_{a,r}\frac{|\alpha_a^r\rangle\langle r|-u|a\rangle}{\epsilon_a-\epsilon_r}. \qquad (106)$$

The terms in this summation are clearly described by the diagram shown in Fig. 12. This type of diagram, which is a modification of the original Feynman graph, is referred to as a Goldstone diagram. The diagram describes a process in which a single electron is excited from a core state labeled a into an excited state labeled r. As is customary, the interaction is denoted by a dotted horizontal line, the core orbital is represented by a line directed downwards, and the excited state is represented by a line directed upwards.

The contribution to the first-order wave function, which is due to the Coulomb interaction among the electrons, can be written out explicitly in terms of two-particle matrix elements:

$$\Psi_C^{(1)}=\sum_{a,r}\frac{|\alpha_a^r\rangle}{\epsilon_a-\epsilon_r}\sum_b(\langle rb|1/r_{12}|ab\rangle -\langle rb|1/r_{12}|ba\rangle) +\frac{1}{2}\sum_{a,b,r,s}\frac{|\alpha_{ab}^{rs}\rangle\langle rs|1/r_{12}|ab\rangle}{\epsilon_a+\epsilon_b-\epsilon_r-\epsilon_s}. \qquad (107)$$

The first two terms describe the excitation of a core state labeled a to an excited state r by means of a direct and an exchange interaction with a core orbital b. The last term describes a two-electron excitation. These three processes are described by the Goldstone diagrams in Fig. 13.

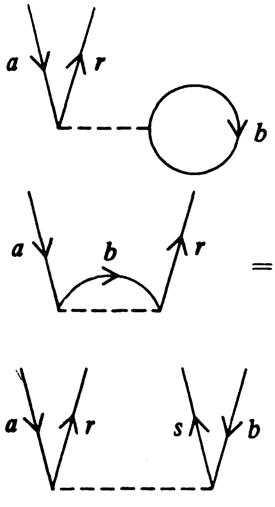

$$= \frac{|\alpha_a^r\rangle\langle rb|r_{12}^{-1}|ab\rangle}{\varepsilon_a-\varepsilon_r}$$

$$= -\frac{|\alpha_a^r\rangle\langle br|r_{12}^{-1}|ab\rangle}{\varepsilon_a-\varepsilon_r}$$

$$= \frac{1}{2}\frac{|\alpha_{ab}^{rs}\rangle\langle rs|r_{12}^{-1}|ab\rangle}{\varepsilon_a+\varepsilon_b-\varepsilon_r-\varepsilon_s}$$

FIG. 13. Goldstone diagrams that describe the excitation of a single electron from a core state by means of a Coulomb interaction.

The higher-order terms in the perturbative expansion may also be represented graphically. Some of the diagrams that correspond to second-order contributions to the wave function are shown in Fig. 14. One may notice that the lowest part of each of these diagrams is the same as one of the diagrams that appear in the first-order wave function. An additional interaction occurs in each of the diagrams, however, as one follows the diagram from bottom to top. For instance, in the first diagram in Fig. 14, there is an additional interaction between a core and an excited orbital. All of the diagrams that appear in the expansion of the wave function are said to be linked, which means that they do not have a closed, disconnected part. The linked nature of the diagram expansion can be used to show that the energy of an extended system has the proper linear dependence upon the number of particles.

5.6 Correlation in Atomic Spectroscopy

Although the Hartree-Fock approximation generally provides a good qualitative description of the atom, there are a number of areas where it is essential to include the effects of correlation. Atomic spectroscopy is one area where correlation, at least at the level of a minimal configuration interaction, is frequently important to account correctly for the experimental results. Transition strengths in particular can be quite sensitive to the corrections arising from an adequate treatment of correlation.

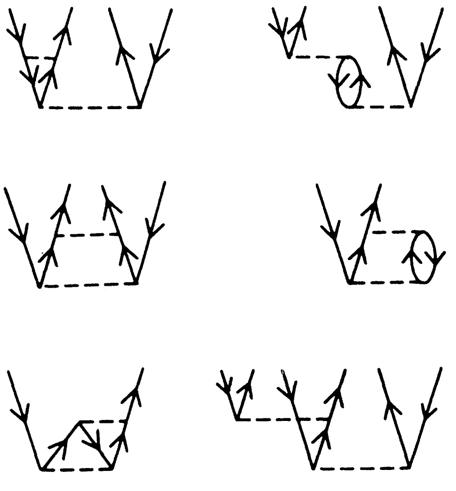

FIG. 14. Some typical second-order Goldstone diagrams.

The fundamental quantity for describing electronic transitions is the transition multiplet strength, or line strength, which for an electric dipole transition is given by

$$\mathbf{S}_{ij}=\sum_{\mu,\nu}|\langle\Phi_{i,\mu}|\mathbf{r}|\Phi_{j,\nu}\rangle|^2. \tag{108}$$

Here, Φ_i and Φ_j are the full multielectron, initial- and final-state wave functions of the transition, with μ and ν representing the degeneracies in each state. The sum over degeneracies of this scalar product, for an independent-particle model, can be carried out analytically to give the multiplet strength in the general form

$$S_{ij}=\mathscr{C}_{ij}\,\sigma_{ij}^2, \tag{109}$$

where

$$\sigma=\frac{1}{\sqrt{4l_>^2-1}}\int_0^\infty dr P_{n_i l}(r)\,r P_{n_j l'}(r), \tag{110}$$

and $\mathscr{C}_{ij}$ is a factor depending only on the orbital angular-momentum coupling in the two states. The P_{nl} functions that appear in this last formula are the radial functions of the "jumping electron." For example, in the $2s^2$–$2s2p$ transition of the beryllium atom, the relevant functions are P_{2s} and P_{2p}. It has been assumed here that all orbitals in both states are orthogonal.

Transition strengths are frequently discussed in terms of the oscillator strength, or f value, a dimensionless quantity that is given by

$$f_{ij}=\tfrac{2}{3}(\Delta E_{ij}/g_i)S_{ij}, \tag{111}$$

where ΔE is the transition energy and g_i the statistical weight of the initial state, which in LS coupling is given by $(2L+1)(2S+1)$. There are alternative, and formally equivalent, formulations of the transition strength, utilizing different transition operators, the dipole velocity and acceleration forms of the multiplet strength. All of these forms must give identical results for the exact wave functions, but need not, and usually do not, when using approximate wave functions.

The correlation corrections to the wave function can have a profound effect on the calculation of f values, as well as on the relative positions of levels in the energy level spectrum. In Fig. 15, a comparison of the Hartree-Fock and experimental binding energies for a few states of the beryllium atom are shown. All level positions are given in electron volts and are relative to the ionization limit, which is lithiumlike Be^+ $1s^22s$. One can see that there is a rather substantial shift of the $2s^2$ ground state of about 1.25 eV. Unlike most other atoms and ions, this correlation shift is accounted for largely by only a single configuration interaction, namely $2s^2+2p^2$. As discussed in the section on multiconfiguration Hartree-Fock (MCHF) methods (Sec. 5.4), the optimum $2p$ function is computed self-consistently for this state and is somewhat different from the Hartree-Fock $2p$ orbitals of the states of either $2s2p$ or $2p^2$, although not drastically so. Another feature to notice in Fig. 15 is the relatively large shift of the $2p^2$ ^1D state to a position below the ^3P of the same configuration. This energy shift may be accounted for in the MCHF theory by adding the configuration $2s3d$ to the $2p^2$ configuration.

In an MCHF calculation, the $3d$ orbital is determined self-consistently within this multiconfiguration context and it, in fact, bears very little resemblance to any independent-particle model orbital corresponding to the $2s3d$ state. This MCHF $3d$ orbital can be written as an expansion in the independent-particle model nd functions of the Rydberg series, $3d_{\mathrm{MCHF}}=\Sigma_n c_n nd$, so that the exact wave function can be written

$$\Psi(^1D)=a_0 1s^2 2p^2+\sum_n b_n 1s^2 2snd+\cdots. \qquad (112)$$

It is understood here that the sum extends over the continuum ϵd functions of the independent-particle model. This correlation effect can therefore be described as an interaction of the perturbing $2p^2$ configuration with the entire Rydberg nd series. In this case, the nd series members are displaced upwards from what the Hartree-Fock theory would predict.

Turning now to oscillator strengths, we note that the correlating effect of the $2p^2$ configuration upon the ground state will affect the f value for the transition $2s^2$ ^1S–$2s2p$ ^1P. The dominant contributions to the transition moment, taking it here to be the z component, are

$$\begin{aligned}\langle\Psi(2s^2)|z|\Psi(2s2p)\rangle &=\langle a_0 2s^2+a_1 2p^2+\cdots|z|b_0 2s2p+\cdots\rangle\\ &=a_0b_0(2s|2s)(2s|z|2p)+a_1b_0(2p|2p)\\ &\quad\times(2p|z|2s)+\cdots. \qquad (113)\end{aligned}$$

The contribution from the dominant correlating configuration in the ground state, $2p^2$, can add to, or subtract from, the HF contribution. In this case, correlation reduces the resonance-line oscillator strength by about 25%. A similar mechanism comes into play in the case of the transition to the series perturber, $2p^2$ ^1D, from $2s2p$ ^1P, for which the mixing of the $2s3d$ configuration with the $2p^2$ ^1D state reduces the oscillator strength essentially to zero. This cancellation effect occurs quite generally for transitions involving configurations that are strong series perturbers and that lie below the series that they perturb. There is a converse enhancement of the oscillator strengths throughout the perturbed series. For more complete treatments of transition probabilities, the reader is referred to the articles by Nicholls and Stewart (1962), Weiss (1973), and Fischer (1988).

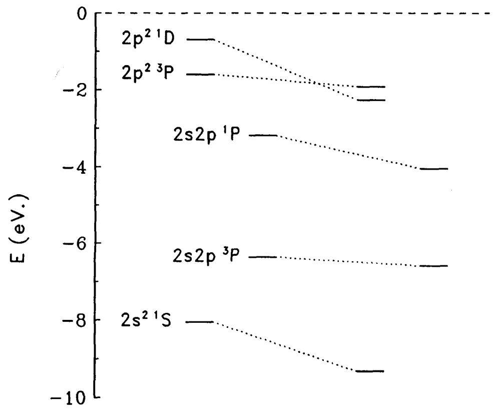

FIG. 15. Comparison of Hartree-Fock and experimental binding energies for some low-lying states of the beryllium atom. The zero of energy is the lithiumlike ion limit. Note the correlation inversion of the states of $2p^2$.

6. RELATIVISTIC EFFECTS IN ATOMIC STRUCTURE

We begin our relativistic treatment of electronic structure with a discussion of the Dirac equation. This equation can be obtained by employing the following linearized expression for the energy:

$$E = c\boldsymbol{\alpha}\cdot\mathbf{p} + \beta mc^2. \tag{114}$$

The requirement that this expression for the energy be consistent with the relativistic equation for the energy of a particle, $E^2 = c^2p^2 + m^2c^4$, leads to the condition that the components of the α and β operators have unit magnitude,

$$\alpha_1^2 = \alpha_2^2 = \alpha_3^2 = 1, \quad \beta^2 = 1, \tag{115}$$

and obey the anticommutation relations

$$\alpha_i\alpha_j + \alpha_j\alpha_i = 0, \quad \alpha_i\beta + \beta\alpha_i = 0. \tag{116}$$

Here m denotes the rest mass of the particle and c the speed of light, which in atomic units is the inverse of the fine-structure constant, 137.036. In most computational applications, the electron mass is set equal to unity just as in the nonrelativistic theory.

The time-dependent form of Dirac's equation is obtained by making the usual replacements,

$$E \to i\hbar\frac{\partial}{\partial t} + eA_0, \quad \mathbf{p} \to -i\hbar\nabla. \tag{117}$$

The customary stationary-state *Ansatz*,

$$\psi(r,t) = e^{-i(E_R/\hbar)t}\psi(r), \tag{118}$$

then yields the time-independent Dirac equation for a single particle in the external potential $U(r)$,

$$[c\boldsymbol{\alpha}\cdot\mathbf{p} + U(r) + \beta mc^2]\psi = E_R\psi. \tag{119}$$

It should be noted that the relativistic energy eigenvalue E_R includes the rest energy,

$$E_R = mc^2 + E, \tag{120}$$

where the binding energy E has the same meaning as that in the nonrelativistic theory.

The above conditions upon the α and β operators are satisfied by the matrices

$$\alpha = \begin{pmatrix} 0 & \sigma \\ \sigma & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \tag{121}$$

where the σ's are the usual 2×2 Pauli spin matrices. It is convenient to make use of this block structure and to express the four-component Dirac wave functions as two-component functions of the form

$$\psi = \begin{pmatrix} \phi^L \\ \phi^S \end{pmatrix}. \tag{122}$$

The Dirac equation reduces then to a pair of coupled first-order equations for the large and small components:

$$c\boldsymbol{\sigma}\cdot\mathbf{p}\phi^s + U\phi^L = E\phi^L, \tag{123}$$

$$c\boldsymbol{\sigma}\cdot\mathbf{p}\phi^L + (U - 2mc^2)\phi^S = E\phi^S. \tag{124}$$

For an atomic central-field problem, each of the wave-function components can be separated into a radial function and an angular factor just as in the nonrelativistic case. Because the momentum operator appears linearly in the above equations, the large and small components must have opposite parities. Following Rose (1961), we write the two components as

$$\phi^L_{nljm} = g_{nlj}(r)\chi_{\kappa,ki,j,m}\,, \tag{125}$$

$$\phi^S_{nljm} = -if_{nlj}(r)\chi_{-\kappa,j,m}\,, \tag{126}$$

where the angular factor is a coupled product of orbital and spin angular-momentum functions,

$$\chi_{\kappa,j,m} = \sum_{\mu=-1/2}^{1/2} \langle lm-\mu, \tfrac{1}{2}\mu \,|\, jm\rangle Y_{l\,m-\mu}\eta_\mu\,. \tag{127}$$

Here, the spherical harmonic $Y_{l\,m-\mu}$ and the spin eigenfunction η_μ are coupled together to form a state having the total angular-momentum quantum numbers j and m. The "relativistic quantum number" κ is defined as

$$\kappa = l, \quad j = l - \tfrac{1}{2}, \tag{128a}$$

$$\kappa = -l - 1, \quad j = l + \tfrac{1}{2}. \tag{128b}$$

The coupled equations for the radial functions can then be written

$$c\left(-\frac{d}{dr} + \frac{\kappa - 1}{r}\right)f(r) + V(r)g(r) = Eg(r), \tag{129}$$

$$c\left(\frac{d}{dr} + \frac{\kappa + 1}{r}\right)g(r) + [V(r) - 2mc^2]f(r) = Ef(r). \tag{130}$$

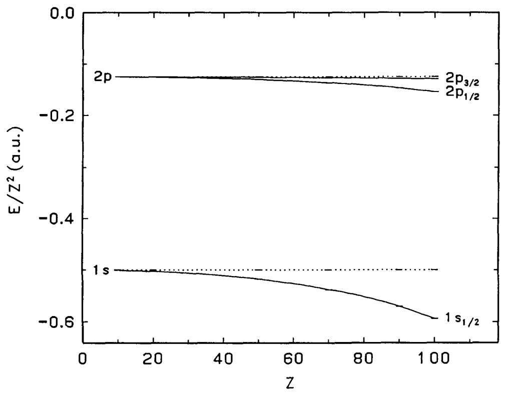

FIG. 16. Comparison of relativistic and nonrelativistic scaled binding energies for hydrogenlike ions.

For the hydrogenic atoms, these equations are exactly soluble, just as in the nonrelativistic theory.

The energy eigenvalues for the $n=1$ and 2 states of hydrogenic ions as functions of the nuclear charge Z are shown in Fig. 16. In this figure, the 2*s* state is not shown, since it is degenerate with the $2p_{1/2}$ state. The energy is scaled by Z^2, which is the exact Z dependence of the nonrelativistic theory. As indicated in the figure, relativistic effects are weaker for states with higher principal quantum numbers where the electron is farther from the nucleus and moves more slowly. It is also clear from the figure that relativistic corrections become larger with increasing nuclear charge. Relativity removes the $2p_{1/2}$-$2p_{3/2}$ degeneracy but not the $2s_{1/2}$-$2p_{1/2}$ degeneracy, which is removed by the Lamb shift. It should be clearly understood that the radial wave functions themselves are now j dependent as well as l dependent.

As in the nonrelativistic case, the radial functions are traditionally multiplied by r. P and Q functions are defined, $P_{nlj}(r)=rg_{nlj}(r)$, $Q_{nlj}(r)=rf_{nlj}(r)$. The nonrelativistic and relativistic wave functions of the $j=\frac{1}{2}$ 1*s* and 2*p* states for $Z=50$ are compared in Fig. 17. For this charge state, the large-component corrections are quite small, and for lower-Z atoms, relativistic effects, while not necessarily negligible, are generally quite small. The first and most dramatic effect of relativity with increasing Z is the development of a small component, which is entirely absent in nonrelativistic theory.

The energy eigenvalues we have discussed thus far are for the positive-energy states. The Dirac equation has a second class of solutions corresponding to a set of negative-energy states. These states, for which the large and small wave-function components reverse their roles, originally presented a puzzle as to their interpretation. Dirac resolved this by asserting that, in a vacuum, *all* these states were fully occupied, and therefore, because of the Pauli exclusion principle, completely undetectable. If, however, one of these electrons were to be excited, the remaining hole would behave as an electron but with a positive charge. The negative-energy solutions have thus become associated with the electron's antiparticle, the positron.

For an electron moving under the influence of a small effective nuclear charge, the effect of relativity may be treated as a small correction. The process of calculating a relativistic correction is initiated by formally solving the second Dirac equation for the small component,

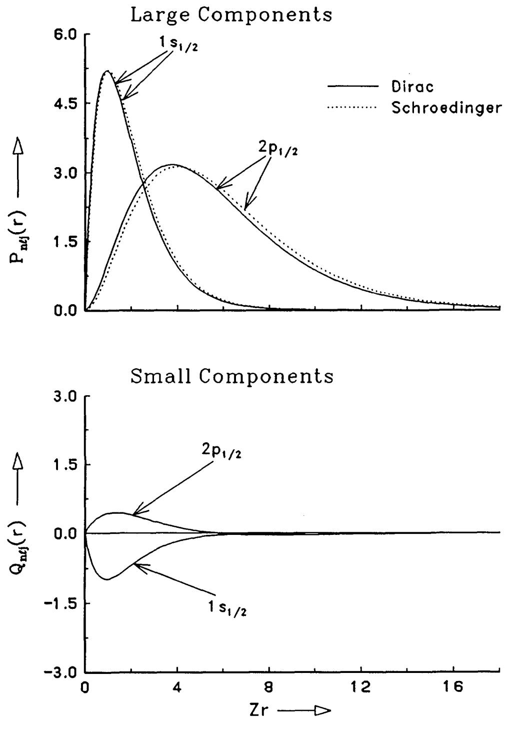

FIG. 17. Relativistic and nonrelativistic hydrogen wave functions for 1*s* and 2*p* states at $Z=50$. Both large and small components are drawn to the same scale.

$$\phi^S = \frac{c\sigma\cdot\mathbf{p}\phi^L}{2mc^2\{1+[E-V(r)]/2mc^2\}}\,. \tag{131}$$

The essential assumption of the Pauli approximation is to assume that $E-V(r)$ is much smaller than the rest energy, which allows one to expand the denominator of this equation. This leads to

$$\phi^S = \frac{1}{2mc}\,\sigma\cdot\mathbf{p}\phi^L - \frac{E-V(r)}{4m^2c^3}\,\sigma\cdot\mathbf{p}\phi^L + \cdots. \tag{132}$$

Inserting this for ϕ^S in the first Dirac equation gives

$$\begin{aligned}&\frac{1}{2m}(\sigma\cdot\mathbf{p})(\sigma\cdot\mathbf{p})\phi^L - \frac{1}{4m^2c^2}\,\sigma\cdot\mathbf{p}[E-V(r)]\sigma\cdot\mathbf{p}\phi^L\\ &\quad+\cdots+V(r)\phi^L\\ &\quad= E\phi^L.\end{aligned} \tag{133}$$

The second operator in this equation, called the Pauli operator, represents the lowest-order correction due to relativity. The leading relativistic correction to the energy is obtained by evaluating the expectation value of the Pauli operator with nonrelativistic wave functions. This leads to the energy correction

$$\begin{aligned}\delta E = &-\frac{1}{8m^3c^2}\langle p^2\phi|p^2\phi\rangle - \frac{4\pi}{8m^2c^2}\langle\phi|\delta(\mathbf{r})|\phi\rangle\\ &-\frac{1}{2m^2c^2}\left\langle\phi\left|\frac{1}{r}\frac{dV}{dr}\,\mathbf{l}\cdot\mathbf{s}\right|\phi\right\rangle.\end{aligned} \tag{134}$$

Here the first term, the mass–velocity term, can be identified as the lowest-order relativistic correction to the classical kinetic energy. The second term, the Darwin term, is purely quantum mechanical and produces an energy shift for *s* states only. The last term, the spin-orbit interaction, is nonvanishing for states with $l>0$ and produces a splitting of the *j* levels arising from such states. All of these expressions are to be evaluated using nonrelativistic wave functions. The Pauli approximation is a reasonable approach for atoms of moderate size.

The basic theoretical framework for treating multielectron atomic or molecular systems is generally the same as for the nonrelativistic theory. Orbitals are replaced, though, by Dirac-type orbitals, with large and small components, and angular-momentum couplings are based on the *j* values of the orbitals. *l* and *s* are no longer appropriate orbital quantum numbers. As of this writing, all computational schemes for the relativistic treatment of multiparticle electronic-structure problems are based on the somewhat *ad hoc* generalization of the corresponding Schrödinger equation,

$$\left(\sum_{\mu} h_{\mathrm{D}}(\mu) + \sum_{\mu>\nu}\frac{1}{r_{\mu\nu}}\right)\Psi = E\Psi, \tag{135}$$

where $h_{\mathrm{D}}(\mu)$ is the single-particle, time-independent Dirac Hamiltonian. The Dirac Hartree-Fock equations are derived in exactly the same way as for the nonrelativistic case. Since pure *j-j* coupling occurs far less often than *LS* coupling, however, the multiconfiguration HF scheme is much more important in the relativistic treatment than the nonrelativistic treatment.

While fundamental objections can be raised about the overall approach, a wealth of experience has been obtained at the Dirac-Fock and multiconfiguration Dirac-Fock stages of approximation. Some relativistic many-body calculations have been done on selected atoms, but very little work has been reported using extensive configuration interaction. The benchmark r_{12} type of study of heliumlike ions still remains to be done.

As with single-electron systems, the dominant effect of relativity is to shrink the electron density closer to the nucleus than in nonrelativistic theory. Because of changes in the screening of the inner electrons and because of the imposition of orthogonality on the orbital set, the contraction propagates to the outer reaches of the atom. This is illustrated in Fig. 18, which shows the Hartree-

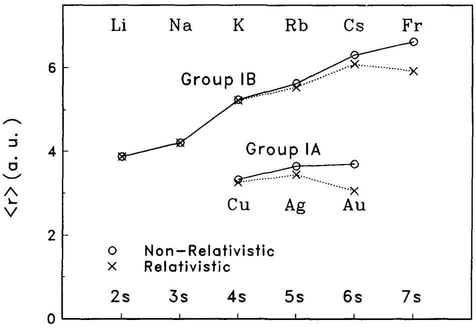

FIG. 18. Mean values of *r* for the valence orbital of the alkali-metal elements (group I of the periodic table). Comparison of nonrelativistic Hartree-Fock and Dirac-Hartree-Fock calculations.

Table 2. Ratio of relativistic to nonrelativistic HF mean values of r for the beginning and ending rare-earth elements.

Orbital	Praesodymium $4s^2 4p^6 4d^{10} 4f^3 5s^2 5p^6 6s^2$	Lutetium $4s^2 4p^6 4d^{10} 4f^{14} 5s^2 5p^6 5d 6s^2$
$4s$	0.954	0.927
$4p_{1/2}$	0.953	0.924
$4p_{3/2}$	0.992	0.983
$4d_{3/2}$	0.994	0.984
$4d_{5/2}$	1.008	1.005
$4f_{5/2}$	1.067	1.029
$4f_{7/2}$	1.080	1.050
$5s$	0.952	0.918
$5p_{1/2}$	0.950	0.912
$5p_{3/2}$	0.997	0.987
$5d_{3/2}$	...	1.084
$5d_{5/2}$	...	1.118
$6s$	0.920	0.916

Fock expectation values of r for the outermost s electron in the ground state of the alkali-metal atoms. As a result of this relativistic contraction, francium and gold are actually smaller than the corresponding lighter atoms, cesium and silver. In Table 2, the ratios of the Dirac HF and HF expectation values of r are given for the first and last rare-earth elements (Desclaux, 1972). Since the higher-angular-momentum d and f orbitals are pushed farther from the nucleus by the centrifugal repulsion, the contracted s and p functions are able to screen them more completely than in the nonrelativistic case. The end result is that the $4f$ and $5d$ orbitals expand relative to the predictions of the nonrelativistic Hartree-Fock theory.

7. FUNDAMENTALS OF MOLECULAR ELECTRONIC STRUCTURE

The increased complexity in the description of molecular electronic structure due to the presence of two or more atoms in a single entity is very significant. While many of the techniques derived for atoms in the previous sections are similar to those applied to molecules, there are a number of issues that are unique to molecules. These considerations include bonding, dissociation, nuclear motion, and the multicentered nature of the Coulombic interactions. These issues, in addition to the inevitable compromises between accuracy and computational effort, may lead to significant differences in the ways in which the various techniques are applied. The development of theoretical models of molecular electronic structure has largely taken place within the field of quantum chemistry, and so much of the vocabulary is that of chemistry rather than physics.

Molecular structure involves both nuclear and electronic coordinates. In Sec. 7.1, a widely used approximation is introduced that allows one to treat separately the motion of the nuclei and the motion of the electrons. Because the focus of this article is on "electronic" structure, very little will be said about the nuclear motion problem. However, it should be noted that the vibrational and rotational nuclear degrees of freedom complicate significantly the spectroscopy of a molecule. Transitions between electronic energy levels in a molecule are usually accompanied by transitions between vibrational and rotational levels. Electronic transitions are generally observed at ultraviolet (900–4000 Å) and optical (4000–6000 Å) wavelengths. Transitions involving only the vibrational and rotational energy levels of a molecule generally take place at much lower energies, observable from the infrared into the microwave region of the electromagnetic spectrum. In addition, large changes in the nuclear framework, such as bond stretching, leading ultimately to dissociation, have profound effects on molecular electronic structure.

There are a variety of different reasons for carrying out molecular electronic-structure calculations on a particular system. One aim might be simply to obtain a qualitative description of the bonding. A more common goal is the characterization of the key features on a potential-energy surface relevant to a chemical reaction, such as the equilibrium geometries of reactants and products, any local minima, and the positions of potential-energy barriers (transition states). Another goal, which is quite different, is to obtain a balanced, high-quality description of one or more states of a system over a wide range of nuclear geometries. In the latter case, more emphasis tends to be placed on generating a description of a consistent quality than on obtaining the highest possible accuracy at a single geometry. Such considerations generally govern which methods have become the most successful for particular applications.

7.1 The Born-Oppenheimer Approximation

Molecular calculations are usually based on a nonrelativistic Hamiltonian that involves the electronic coordinates **r** and the nuclear coordinates **R** but that contains no spin-dependent terms. The total molecular wave function $\Psi_{\text{tot}}(\mathbf{r},\mathbf{R})$ depends explicitly on **r** and **R**. This molecular Hamiltonian consists of operators relating to the kinetic energies of the nuclei and of the electrons, and of operators relating to the Coulombic interactions between all of these charged particles. In SI units,

$$H_{\text{tot}}=\sum_{A}-(\hbar^2/2M_A)\nabla_A^2-\sum_{i}(\hbar^2/2m_e)\nabla_i^2+\sum_{A<B}\frac{Z_AZ_B\,e^2}{4\pi\epsilon_0|\mathbf{R}_A-\mathbf{R}_B|}-\sum_{A,i}\frac{Z_Ae}{4\pi\epsilon_0|\mathbf{R}_A-\mathbf{r}_i|}+\sum_{i<j}\frac{e^2}{4\pi\epsilon_0|\mathbf{r}_i-\mathbf{r}_j|}, \tag{136}$$

in which A and B label nuclei, i and j label electrons, Z_A and Z_B are atomic numbers, and M_A and m_e are masses of nuclei and electrons, respectively. In order to simplify the present discussion, it is convenient to introduce symbols for the individual kinetic-energy (T) and potential-energy (V) terms in the Hamiltonian for the electrons (e) and the nuclei (N):

$$H_{\text{tot}}=T_N+T_e+V_{NN}+V_{Ne}+V_{ee}\,. \tag{137}$$

The majority of theoretical studies of molecular systems implicitly assume the Born-Oppenheimer approximation, which separates the electronic-structure problem from that of the motion of the nuclei. The electronic wave function and the corresponding energy still depend on the positions, but not on the momenta, of the nuclei. The physical justification for the Born-Oppenheimer approximation is that the nuclei are thousands of times more massive than the electrons. From the point of view of the electronic structure of the molecule, the nuclei can be regarded as instantaneously stationary, creating a fixed potential of positive charges in which the electrons move.

In the Born-Oppenheimer approximation, Ψ_{tot} is factorized into separate electronic (Ψ_{el}) and nuclear (χ_{nuc}) parts:

$$\Psi_{\text{tot}}(\mathbf{r},\mathbf{R})=\Psi_{\text{el}}(\mathbf{r}\,|\,\mathbf{R})\chi_{\text{nuc}}(\mathbf{R}) \tag{138}$$

in which $\Psi_{\text{el}}(\mathbf{r}\,|\,\mathbf{R})$ depends explicitly on **r** but only parametrically on **R**. Substitution of this product wave function into the Schrödinger equation, after neglecting certain cross terms, leads to two equations:

$$(T_e+V_{Ne}+V_{ee})\Psi_{\text{el}}=E_{\text{el}}\Psi_{\text{el}} \tag{139}$$

and

$$(T_N+V_{NN}+E_{\text{el}})\chi_{\text{nuc}}=E_{\text{tot}}\chi_{\text{nuc}}\,. \tag{140}$$

The terms that have been neglected in making this separation of nuclear and electronic coordinates have the following form:

$$-\sum_{A}\frac{\hbar^2}{2M_A}[2(\nabla_A\Psi_{\text{el}})(\nabla_A\chi_{\text{nuc}})+\chi_{\text{nuc}}\nabla_A^2\Psi_{\text{el}}] \tag{141}$$

and are generally of the order of 10^{-5} to 10^{-7} times the electronic energy.

This factorization of Ψ_{tot} into a *single* product of an electronic wave function and a nuclear-motion wave function is a very good approximation unless two or more electronic states lie very close in energy. The breakdown of the so-called adiabatic approximation is of particular importance when one considers "curve-crossing" problems (see Sec. 9.3).

A molecular electronic-structure calculation in the Born-Oppenheimer approximation involves finding approximate solutions of the time-independent Schrödinger equation for *fixed* values of the nuclear coordinates, with the "clamped-nucleus" Hamiltonian:

$$H=-\sum_{i}\frac{1}{2}\nabla_i^2-\sum_{A,i}\frac{Z_A}{r_{Ai}}+\sum_{i<j}\frac{1}{r_{ij}}\,. \tag{142}$$

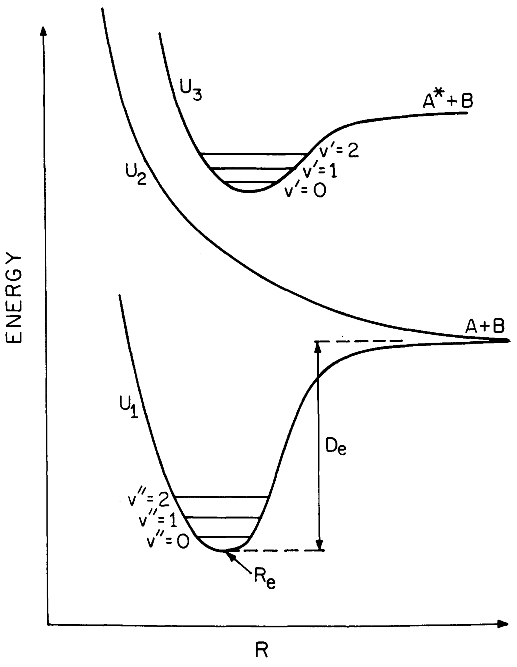

FIG. 19. Potential-energy curves of a generic molecule *AB*. U_1, a bound state, and U_2, a repulsive state, both dissociate to ground-state atoms $A+B$. Also shown is a bound-excited-state potential U_3, dissociating to the excited asymptote A^*+B. The lowest vibrational energy levels of the bound states are shown.

Here, the Hamiltonian is written in atomic units and the symbols r_{Ai} and r_{ij} are used to denote the quantities $|\mathbf{R}_A-\mathbf{r}_i|/a_0$ and $|\mathbf{r}_i-\mathbf{r}_j|$, respectively, where a_0 is the Bohr radius. The value of the energy that is obtained from such a calculation is the electronic energy E_{el}. The total electronic energy of the system is the sum of E_{el} and the fixed value of the Coulomb repulsion between the nuclei (E_{NN}). In general, the Schrödinger equation must then be solved for a number of different nuclear geometries.

A potential-energy surface $U(\mathbf{R})$ for a polyatomic system (a potential-energy curve for a diatomic molecule) consists of the total electronic energy of the system ($E_{\text{el}}+E_{NN}$) as a function of all the nuclear coordinates $\mathbf{R}$. The concept of the potential-energy surface is extremely useful in molecular physics, as it governs the theoretical interpretation of chemical reactions and molecular spectroscopy and of all atomic- and molecular-collision processes. Three idealized potential-energy curves for a diatomic molecule, *AB*, are shown in Fig. 19. The lowest potential (labeled U_1 in the figure) exhibits a minimum at a lower energy than the separated atoms and thus corresponds to a bound molecular electronic state. The minimum energy $U_1(R_e)$ occurs at the equilibrium nuclear separation R_e. The difference between $U_1(R_e)$ and the energy at the separated-atom limit is known as the binding energy, or dissociation energy D_e. The first three nuclear vibrational energy levels are shown schematically. The second potential shown in Fig. 19 (labeled U_2) represents an unbound molecular state, or "repulsive" state, for which there is no potential well and thus no binding. The top curve (U_3) represents a bound excited molecular state with the corresponding vibrational levels.

For polyatomic systems, it becomes exceedingly difficult to visualize the complete multidimensional potential-energy surfaces, and it is usual to display the surfaces in three dimensions by fixing all but two geometric variables. An example of an idealized surface for a triatomic system *ABC* is shown in Fig. 20 as contours of $U(R_{AB},R_{BC})$. Of course, provided potential-energy surfaces are available, it is possible to treat the nuclear motion problem, but this is beyond the scope of the present account, and the remainder of this article concentrates instead on the electronic-structure problem.

The clamped-nucleus Hamiltonian contains no spin-dependent terms, and so it is entirely reasonable to insist that the electronic wave functions Ψ_{el} should correspond to a

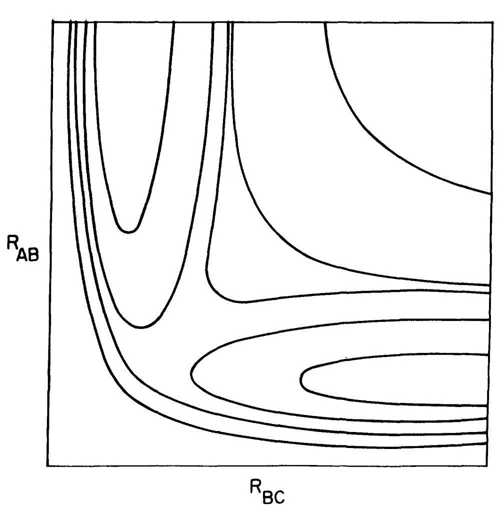

FIG. 20. Contours of $U(R_{AB},R_{BC})$ for a generic triatomic system *ABC*.

well-defined value for S, the quantum number for the total spin. This means that Ψ_{el} must be an eigenfunction of the spin operator S^2:

$$S^2\Psi_{el}=S(S+1)\Psi_{el}. \qquad (143)$$

7.2 Qualitative Ideas

One of the first successful attempts to account in a qualitative way for molecular electronic structure can be traced back to 1916, several years before the advent of the Schrödinger equation, and it is usually associated with the name of Gilbert Lewis. The basic idea behind Lewis structures is that the various atoms in a molecule share some or all of their valence electrons so as to form electron pairs. Achieving the same number of electrons as the next noble-gas element is deemed particularly favorable; for atoms such as, say, carbon or chlorine this involves completing an "octet" of electrons.

The Lewis structure of the water molecule may be represented as

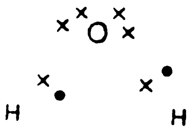

where the six valence electrons of the O atom have been denoted by crosses and the single valence electrons of each H atom by circles. By "sharing" two of its electrons with the H atoms, the oxygen atom achieves an octet. The H atoms achieve the corresponding magic number of two electrons (as in He). Each of the shared pairs of electrons may be termed a covalent bond, whereas the pairs of valence electrons on oxygen not involved in bonding are termed lone pairs. A more conventional representation of the Lewis structure for water takes the form

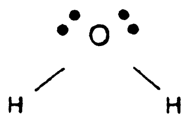

in which the notation O–H signifies a two-center, two-electron covalent bond and each pair of dots represents a lone pair.

A simple, but fairly reliable, means of predicting geometries from Lewis structures is provided by the valence-shell electron-pair repulsion or VSEPR model. The rationale is that the various classes of repulsions between pairs of electrons are ordered

$$\text{lone pair–lone pair} > \text{lone pair–bond} > \text{bond–bond}. \qquad (144)$$

In the case of H_2O, for example, the four pairs of valence electrons on oxygen would be expected to adopt an approximately tetrahedral arrangement. In the VSEPR model, the repulsions between the two lone pairs dominate the repulsions between the two covalent bonds, and so the bond angle is reduced from the tetrahedral angle $[\cos^{-1}(-\frac{1}{3})\approx 109.5°]$ to approximately 104.5°.

Many of the qualitative notions implicit in Lewis structures and in the VSEPR model were given a firm quantum-mechanical basis with the introduction of classical valence-bond theory, which is outlined in the following section.

7.3 Classical Valence-Bond Theory

The basic idea behind valence-bond (VB) theory is very appealing, and it implements in a clearcut way much of the experience of chemistry, in that the electronic wave function for a molecule is constructed directly from the wave functions of the constituent atoms or fragments. It is clear that VB concepts such as hybridization (see later), two-center two-electron bonds, and lone pairs have a strong influence on the interpretation of electronic structure and bonding. Nonetheless, VB theory has largely been eclipsed by molecular-orbital methods, which will be described later.

In constructing a molecular wave function from those of its fragments, a fundamental requirement is that it must be antisymmetric: the simultaneous interchange of the space and spin coordinates of any two electrons must induce only a change in the sign of the wave function (Pauli principle). In particular, the wave function must be antisymmetric with respect to the interchange of electrons stemming from different fragments. This leads to the exchange contribution to the energy, which is a key to understanding the formation and strength of covalent bonds. The magnitude of this exchange energy and thus the strength of covalent bonding are determined to a large extent by the degree of overlap between the wave functions of the fragments. This nonorthogonality is the hallmark of VB approaches and is also the source of most of the computational difficulties. However, for a given level of accuracy, the use of nonorthogonal orbitals can lead to much more compact wave functions than are available from methods based on orthogonal orbitals.

As a simple example of a classical VB wave function, consider the singlet ground state of the H_2 molecule, with the two nuclear centers denoted A,B and the two electrons labeled 1,2. The spin part of the wave function, which is antisymmetric, takes the form

$$\Theta=2^{-1/2}[\alpha(1)\beta(2)-\alpha(2)\beta(1)]. \tag{145}$$

In the simplest VB description, there are three "structures" to consider: the covalent structure H–H and the two ionic structures H^+H^- and H^-H^+. The spatial wave function Φ_{cov} for the covalent structure can be written

$$\Phi_{\text{cov}}=a(1)b(2)+a(2)b(1), \tag{146}$$

where a and b are atomic 1s functions on centers A and B, respectively. In a similar fashion, the spatial wave function Φ_{ion} for the two ionic structures, which must enter the total wave function with equal weights so as to preserve the molecular symmetry, can be written

$$\Phi_{\text{ion}}=a(1)a(2)+b(1)b(2). \tag{147}$$

The VB wave function Ψ_{VB} for this system then takes the form

$$\Psi_{\text{VB}}=[\Phi_{\text{cov}}+\mu\Phi_{\text{ion}}]\Theta, \tag{148}$$

where μ is a variationally optimized parameter that varies with the internuclear distance.

It is possible to rewrite the VB wave function just described in a much more compact form known as the Coulson-Fischer wave function:

$$\Psi_{\text{CF}}=\Phi_{\text{CF}}\Theta. \tag{149}$$

The H_2 molecule is described here by the spatial configuration Φ_{CF}, which has exactly the same form as Φ_{cov} above, except that the functions a and b are no longer restricted to be localized on a single center. Instead, these functions are given by

$$a=(1s_A+\lambda 1s_B),\quad b=(1s_B+\lambda 1s_A), \tag{150}$$

where $\lambda(R)$ is a (small) variationally optimized parameter, which tends to zero for large nuclear separations. The spatial function Φ_{CF} incorporates the optimal balance between the covalent and ionic structures in a *single* spatial configuration. The optimization of nonorthogonal orbitals using multicenter basis sets, as in modern VB methods (see Sec. 8.5), is crucial for the development of accurate, compact, and easily interpreted VB descriptions of molecular electronic structure.

The experimentally determined tetrahedral arrangement of the hydrogen atoms in methane (CH_4) was something of a puzzle in the early days of quantum mechanics, in that it is difficult to see why the atomic configuration of carbon ($1s^2 2s^2 2p^2$) should give rise to four equivalent C–H bonds. The solution suggested by Linus Pauling is based on taking certain linear combinations of the 2s and 2p orbitals known as hybrid orbitals. So-called sp^3 hybridization involves constructing four equivalent, orthogonal linear combinations of

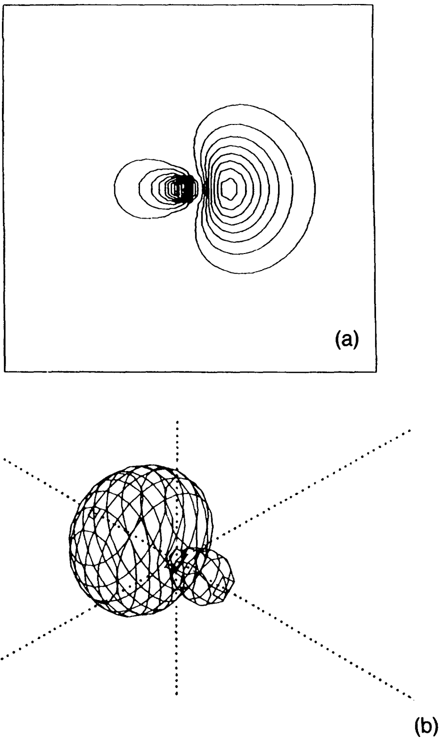

FIG. 21. An sp^3 hybrid orbital for carbon: (a) a contour plot, (b) a perspective view.

$2s$, $2p_x$, $2p_y$, and $2p_z$ functions that point towards the corners of a tetrahedron. A contour plot of an sp^3 hybrid orbital for carbon is shown in Fig. 21(a), and a perspective view in Fig. 21(b). Hybridization is a very widely used concept, particularly in organic chemistry. Hybrid orbitals of this general type arise automatically in modern VB approaches simply by minimizing the total energy, without any form of constraints on the wave function.

Some molecules cannot be realistically represented by a single Lewis structure. The mixing of different structures in valence-bond theory is called resonance, and it is usually denoted by means of a double-headed arrow. An important example in this respect is the benzene molecule C_6H_6, which features a regular hexagonal arrangement of carbon atoms. In classical VB theory, the bonding in this molecule might be represented as

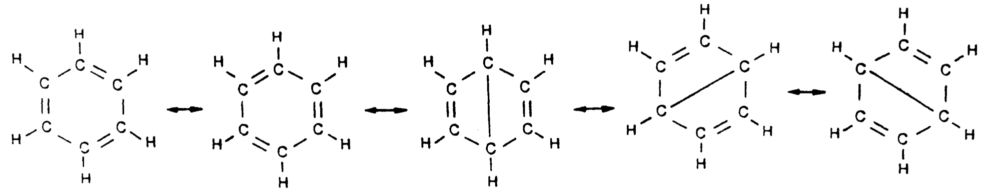

(151)

This notation signifies that the total wave function is represented as a linear combination of the various structures: it does *not* mean that the molecule is spending some of its time in one form and some of its time in another.

7.4 Molecular-Orbital Theory—The Hartree-Fock Self-Consistent Field Approach

In many respects, molecular-orbital theory is much more similar than is valence-bond theory to the model used for describing atomic structure, although there are also a number of important differences. The basic premise is that each electron is described by a function called a molecular orbital φ_i, which may have nonzero amplitude over the entire nuclear framework. A function of the spatial coordinates of one electron, $\varphi_i(\mathbf{r})$, can be, and usually is, associated with at most two electrons of opposing spins and spin functions α and β. It is usual in SCF calculations to use the same spatial functions for the electrons of α and β spins. Such a calculation is referred to as a restricted Hartree-Fock or RHF calculation.

For molecules, the simplest electronic-structure problem is provided by the hydrogen molecular ion H_2^+, with one electron bound to two protons fixed in space at a separation R. To a first approximation, the wave function of this one-electron system consists of one molecular orbital (MO) composed of the ground-state atomic orbitals ($1s$) of each constituent atom, denoted a and b:

$$\varphi_i = 1s_a + 1s_b \,. \tag{152}$$

This model is derived from the recognition that the fundamental building blocks of molecules are atoms (cf. VB theory). The concept, easily extended to larger many-electron systems, is denoted the LCAO-MO (linear combination of atomic orbitals–molecular orbital) method and takes the form

$$\varphi_i = \sum_p c_{ip} \chi_p \,, \tag{153}$$

in which the χ_p are atomic orbitals centered on each of the nuclei.

The molecular orbital for H_2^+ as expressed above is a "bonding" orbital that describes the accumulation of electron density in the region between the two nuclei. This leads to a favorable contribution to the potential energy of the molecule and the formation of a bond. An important contribution to the total energy arises from the fact that the functions $1s_a$ and $1s_b$ contract significantly upon molecule formation. An "antibonding" combination of the hydrogen atomic orbitals, $1s_a - 1s_b$, depletes the electron density in the internuclear region. It can be shown for a range of nuclear separations R that the energy corresponding to the bonding orbital of H_2^+ is lower than the energy of a hydrogen atom and a hydrogen ion at infinite separation. Conversely, the energy corresponding to the antibonding orbital at finite R is greater than that at infinite separation. The ground state of H_2^+, in which the electron occupies the lowest-energy orbital, exhibits a bound potential-energy curve, similar to U_1 in Fig. 19, whereas the first excited state of H_2^+ is repulsive with a potential curve similar to U_2.

To a large extent, the Hartree-Fock procedure for determining the optimum set of linear expansion coefficients c_{ip} for each MO takes the same form as outlined for atoms in Sec. 3. The RHF wave function for a closed-shell molecule is based on a single configuration of doubly occupied orthonormal molecular orbitals, $\varphi_1^2\varphi_2^2\cdots\varphi_n^2$, and is conveniently written as a Slater determinant. Application of the variation principle to the expectation value of the energy E_{el} results in a series of one-electron equations (as derived in Sec. 3), which can be written

$$F\varphi_i=\epsilon_i\,\varphi_i\,. \tag{154}$$

Each electron experiences the average field of all the other electrons as well as the potential defined by the clamped nuclei. Because the Fock matrix depends implicitly on the form of all the orbitals, i.e., on all c_{ip}, an iterative scheme must be employed to find a solution to this generalized eigenvalue problem. Starting from a set of atomic orbitals χ_p (or, more generally, basis functions as mentioned in Sec. 3) and an initial guess for the c_{ip}'s, the Fock matrix is constructed and the relevant matrix equation is solved to find a new set of c_{ip}'s. The process is repeated until the change in energy from one cycle to the next is less than a preselected threshold and/or the elements of **c** are essentially constant. Various sophisticated schemes are available to accelerate the convergence of this process in the minimum number of iterations. In actual practice, the scheme for optimizing the occupied orbitals may rely on repeated mixing of these orbitals with unoccupied or "virtual" orbitals.

The ground-state energy obtained in an SCF calculation cannot be lower than the ground-state energy of the exact solution of the clamped-nucleus electronic-structure problem. Although the basis functions χ_p were originally conceived of as atomic orbitals, in actual calculations the basis set is more likely to consist of a significant number of relatively simple analytic functions with no direct relationship to atomic orbitals (as mentioned in Sec. 3). The Hartree-Fock energy obtained with use of a mathematically complete set of basis functions is termed the Hartree-Fock limit energy. In practice, of course, the basis sets used in molecular calculations are not complete, and so the calculated energy is higher (less negative) than the Hartree-Fock limit energy.

The MOs that are obtained by solving the Hartree-Fock equations are termed the canonical orbitals and can be classified according to their symmetry properties. For example, the occupied (bonding) orbital of H_2^+ is, to a first approximation, an in-phase combination of 1*s* functions on the two centers. This orbital is cylindrically symmetric around the internuclear axis (σ symmetry) and it is unchanged on inversion through the center of mass (*gerade*, or g symmetry). Consequently, this orbital is denoted $1\sigma_g$, where the 1 indicates that this orbital is the first one of σ_g symmetry. The lowest unoccupied (antibonding) orbital of H_2^+ is, to a first approximation, an out-of-phase combination of 1*s* functions on the two centers. This orbital is denoted $1\sigma_u$, where the subscript u, meaning *ungerade*, indicates that this orbital changes sign on inversion through

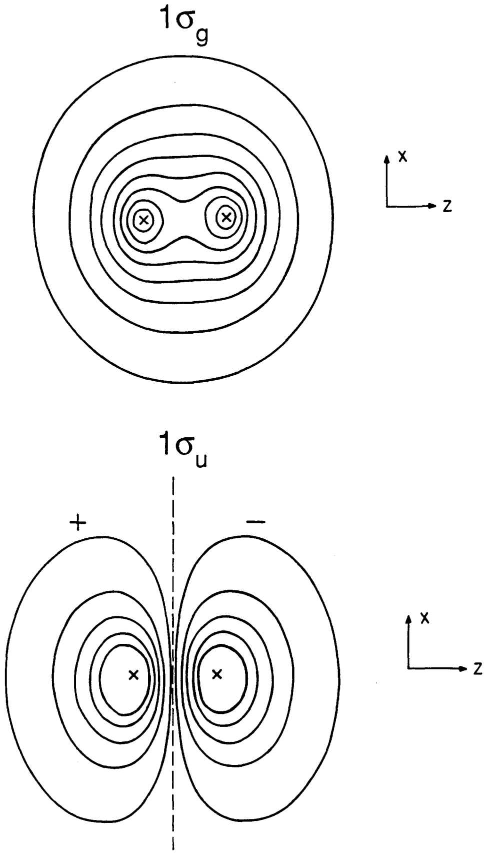

FIG. 22. Contour plots of $1\sigma_g$ bonding and $1\sigma_u$ antibonding orbitals in H_2^+.

the center of mass. These two orbitals are illustrated in Fig. 22.

For more complicated polyatomic systems it can be very useful to transform the orbitals to a more localized set, giving a clearer indication of the nature of the bonding. This relies on an important feature of the SCF wave function, namely that it is invariant against arbitrary unitary transformations of the doubly occupied orbitals. A number of schemes, of varying objectivity and computational expense, is available for generating localized MOs. In the case of methane, for example, the application of such a localization procedure results in four equivalent C–H bonding orbitals, each composed of an sp^3-like hybrid on carbon and a $1s$ function on hydrogen. A useful supplement to the use of computer graphics in analyzing canonical or localized MOs is to estimate the integrated electron density or "electron population" around different nuclear centers. There is, of course, no unique criterion for partitioning the electron density into contributions from different sites, and so several different schemes are in widespread use.

The simplest molecule beyond H_2^+ is molecular hydrogen with two electrons of opposing spins. In the ground state, each electron can be described spatially by a $1\sigma_g$ orbital of a very similar form to the bonding orbital in H_2^+. The wave function for such a system can be written

$$\Psi(1,2)=1\sigma_g(1)1\sigma_g(2)\Theta, \qquad (155)$$

where the spin function

$$\Theta=2^{-1/2}[\alpha(1)\beta(2)-\beta(1)\alpha(2)] \qquad (156)$$

is antisymmetric with respect to interchange of the two electrons 1 and 2. With a large flexible basis set of functions centered on each nucleus, the above wave function can give a Hartree-Fock energy (including the nuclear repulsion energy) of -1.134 a.u. at the equilibrium separation $R_e=1.4a_0$. The error made using the RHF approximation, or correlation energy (as defined in Sec. 5.1), is known to be 0.0409 a.u. for H_2 at R_e. The RHF energy gives over 96% of the exact total energy. However, a much more relevant quantity for molecules is the binding energy D_e, as defined in Fig. 19. D_e is computed as the difference of two energies, $E(R_e)$ and $E(\text{large } R)$, which may have very different amounts of correlation error. Thus, the dissociation energy, which is always much smaller than total energies, is very sensitive to correlation error.

In addition, there is a serious problem unique to *molecular* restricted Hartree-Fock theory that develops as the nuclei are pulled apart. In the case of H_2, multiplying out

$$\Psi(1,2)=[1s_a(1)+1s_b(1)] \times[1s_a(2)+1s_b(2)]\Theta, \qquad (157)$$

it is clear that the wave function consists of an equal mixture of terms such as $1s_a(1)1s_b(2)$, which correctly describe two neutral H atoms, and terms such as $1s_a(1)1s_a(2)$, which describe ionic asymptotes. The fact that ionic terms account for 50% of the wave function for all nuclear separations means that the RHF energy at large R becomes grossly inaccurate.

A number of molecular properties are affected by the fact that in the Hartree-Fock approximation electron–electron interactions are not accounted for exactly, but in an average sense. In general, RHF bond distances are shorter than spectroscopically derived values. For H_2, for example, the RHF minimum occurs at $R=1.38a_0$ instead of $1.40a_0$. Usually, errors in RHF bond lengths calculated with high-quality basis sets are on the order of $(0.02–0.04)a_0$. For polyatomic molecules, RHF bond angles are typically within 1°–2° of spectroscopically determined angles. With these caveats, SCF calculations have been very successful in predicting molecular geometries.

The Hartree-Fock approximation tends to exaggerate the polarity of bonds (cf. the weight given to ionic structure in the case of H_2). Thus the dipole moment, which characterizes the distribution of charge in the molecule and which is calculated as the expectation value of a one-electron operator, as indicated in Sec. 9.3, tends to be overestimated at the Hartree-Fock level of theory, often by as much as 0.5 to 1.0 D.

Much of the discussion so far has been restricted to closed-shell systems with doubly occupied MOs. RHF calculations for open-shell systems are less straightforward, although certain simplifications do arise for the special case in which all the open-shell orbitals are exactly half-filled and the corresponding electron spins are coupled to the maximum possible value. The label ROHF is often used to denote open-shell restricted Hartree-

Fock calculations. In practice, ROHF calculations usually require the input of coupling coefficients that define the energy expression to be minimized. Note that restricted Hartree-Fock wave functions for closed- and open-shell systems are eigenfunctions of the spin operator S^2.

One alternative to the use of RHF and ROHF wave functions is to allow different orbitals for different spins. Such a calculation is denoted unrestricted Hartree-Fock or UHF. The α and β spin orbitals in such a wave function are separately orthonormal, but the spatial parts of the two sets of orbitals are not orthogonal to one another. In spite of the increased number of orbitals to be optimized, it is generally much more straightforward to generate UHF wave functions than ROHF wave functions. UHF calculations yield lower energies than RHF wave functions and can often describe reasonably well the breaking of a single two-electron bond (see also Sec. 8.1). The major objection to UHF wave functions is that they are not eigenfunctions of the spin operator S^2, i.e.,

$$S^2\Psi_{\mathrm{UHF}} \neq S(S+1)\Psi_{\mathrm{UHF}} .$$

Such a situation is known as spin contamination and can lead to spurious, unphysical behavior, especially when states of different spin multiplicity are close in energy to one another. As discussed in Sec. 8.3, spin contamination can lead to very serious problems if UHF wave functions are used as the starting point for more sophisticated calculations. UHF wave functions are available from many of the standard quantum chemistry computational packages and are consequently very widely used. There has very recently been significant progress in developing open-shell SCF methods in which the wave function is effectively constrained to have the correct eigenvalue of S^2.

There are several useful theorems that apply to closed-shell RHF wave functions. According to Koopmans's theorem, previously discussed in Sec. 3.1, the orbital energies ϵ_i, obtained as solutions of the Fock equations, may be thought of as ionization potentials of the molecule. The approximations inherent in Koopmans's theorem include the neglect of the orbital relaxation that follows ionization and of the differences in correlation energy between the parent molecule and its ion (see Sec. 9.2). Nevertheless, orbital energies provide useful initial guidance for the interpretation of photoelectron spectra. Differences in the relativistic contributions to the energy can cause significant errors if Koopmans's theorem is used to interpret experiments such as x-ray photoelectron spectroscopy in which core electrons are ionized. It is important to point out that the orbital energies cannot be used on their own to provide meaningful estimates of the electronic excitation energies in a molecular system.

Brillouin's theorem states that there are no nonzero matrix elements of the Hamiltonian between the SCF wave function and singly excited wave functions that differ from the first one only by the promotion of one electron from an occupied spin orbital to a virtual spin orbital. This has consequences for the construction of more sophisticated wave functions (see Sec. 8), although singly excited configurations may be crucial for the accurate calculation of molecular properties and they do contribute, albeit indirectly, to the correlation energy.

7.5 Molecular Symmetry

The nuclear framework of the molecule establishes a spatial symmetry, termed the point-group symmetry, which must be reflected in the clamped-nucleus Hamiltonian and in the electronic wave functions of the system. Molecular states are classified according to the spin multiplicity, $2S+1$, and by the symmetry properties of the spatial part of the wave function. Symmetry considerations are very useful in molecular-structure calculations. Indeed, many qualitative properties of an electronic state, such as degeneracies and selection rules for making transitions, may be deduced from symmetry arguments alone. Furthermore, it is often possible to use the molecular point-group symmetry to reduce significantly the size of electronic-structure calculations.

As a simple example of the use of symmetry to classify molecular states, we consider the water molecule. The axis convention for this molecule is shown in Fig. 23. The nuclear framework (and thus also the Hamiltonian) is unchanged upon rotation through an angle of 180° about the z axis. This axis is denoted C_2, where the subscript $n=2$ indicates that the rotation angle is $2\pi/n$ radians. The nuclear

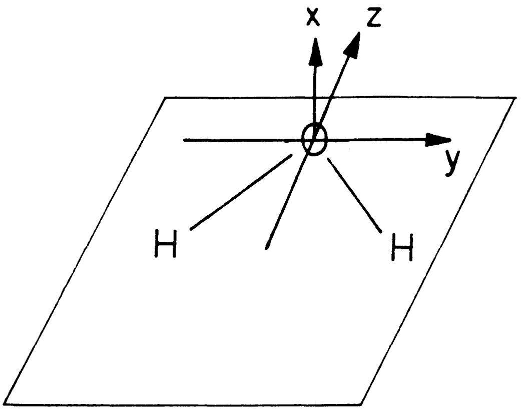

FIG. 23. Axis convention for the H_2O molecule (C_{2v} symmetry). The molecule plane is *yz*.

framework is also unchanged by reflection in the *yz* molecular plane (denoted σ_v'), reflection in the *xz* plane (denoted σ_v), or the identity operation. These four symmetry operations combine to form a group, known as the point group C_{2v}. Wave functions for a molecule possessing C_{2v} symmetry can be classified according to their symmetry properties under the various operations that form the group. Just four symmetry types or "irreducible representations" can occur, labeled A_1, A_2, B_1, and B_2. The ground state of H_2O has the term symbol $^1A_1$, where A_1 denotes the spatial symmetry and the superscript 1 indicates that the spin multiplicity is $2S+1=1$ (i.e., $S=0$). The canonical orbitals of H_2O from an RHF calculation are similarly labeled a_1, a_2, b_1, and b_2, so that the Hartree-Fock electronic configuration of the ground state of water near its equilibrium geometry is $1a_1^2 2a_1^2 3a_1^2 1b_1^2 1b_2^2$.

There are only four common labels for MOs in nonlinear molecules:

1. *a* or *b*: nondegenerate. The precise difference between *a* and *b* symmetries is linked to the behavior under rotation about the C_n axis with highest *n*. Such orbitals can accommodate a maximum of two electrons.
2. *e*: a pair of degenerate MOs. Such orbitals can accommodate a maximum of four electrons.
3. *t*: a set of three degenerate MOs. Such orbitals can accommodate a maximum of six electrons.

Each of these symbols *a,b,e,t* can be modified with various additional labels such as g and u, which we encountered in the case of H_2 (see Sec. 7.4). Other possible labels include 1, 2, and various primes, as in a''. Most chemical physicists do not concern themselves with the precise meaning of all of these symbols, but regard them instead as names that distinguish one orbital from another.

In the special case of heteronuclear diatomic molecules, or indeed any linear molecule that is not centrosymmetric (such as OCS), the point group is $C_{\infty v}$. The rotation axis, which passes through all the nuclei, is denoted C_∞, where the subscript $n=\infty$ signifies the special property of cylindrical symmetry that the nuclear framework, and thus the clamped-nucleus Hamiltonian, is invariant against *any* arbitrary rotation around the internuclear axis. The symmetry designations of the orbitals $\sigma,\pi,\delta,\phi,\ldots$, which correspond to quantized projections of the electronic orbital angular momentum along the internuclear axis of 0,1,2,3,..., are analogous to the labels used for atoms (*s,p,d,f,...*). Molecular electronic states in $C_{\infty v}$ symmetry are similarly labeled $\Sigma,\Pi,\Delta,\Phi,\ldots$ according to the value (0,1,2,3,...) of Λ, the projection of the total electronic orbital angular momentum along the symmetry axis. The full designation of term symbols for $C_{\infty v}$ molecules is $^{2S+1}\Lambda_\Omega$, in which the numerical value of Λ is replaced by the corresponding symbol ($\Sigma,\Pi,\Delta,\Phi,\ldots$), and $\Omega=|\Lambda+\Sigma|$ is the total angular momentum about the internuclear axis, where Σ (here) is the corresponding projection of the total spin *S*. For $\Lambda>0$, the molecular states are doubly degenerate, whereas for $\Lambda=0$ the states are labeled Σ^+ or Σ^- according to whether or not the electronic wave function changes sign upon reflection in any plane containing the C_∞ axis.

Of course, many states of a molecule can correspond to the same spin multiplicity and spatial symmetry, and so it is necessary to find some means of uniquely labeling each state. The ground state is designated with a capital *X* before the spin multiplicity; the ground state of the OH radical, for example, is $X\,^2\Pi_{3/2}$. Excited states with the same spin multiplicity ($2S+1$) as the ground state are labeled with capital letters (*A,B,C,...*) whereas lowercase letters (*a,b,c,...*) are used to designate states of different spin multiplicity.

Homonuclear diatomic molecules, or indeed any linear molecule that is centrosymmetric (such as CO_2), belong to the point

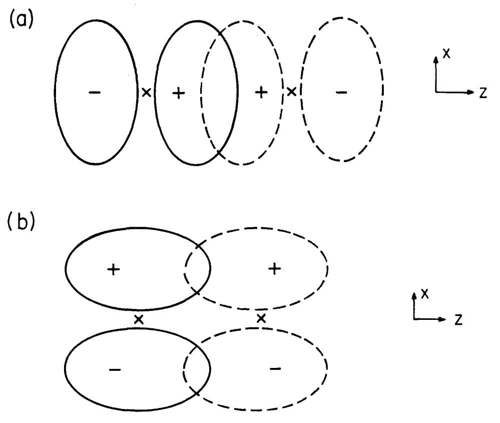

FIG. 24. Schematic representation of orbitals in a diatomic molecule: (a) bonding σ combination of $2p_z$ functions; (b) bonding π combination of $2p_x$ functions.

group $D_{\infty h}$. The labeling of orbitals and of molecular states is analogous to the designations just described for $C_{\infty v}$. The main difference, as already encountered in Sec. 7.4, is the addition of subscripts g or u to signify whether the orbital or wave function is even or odd under inversion through the center of mass. Thus the ground state of H_2, with RHF configuration $1\sigma_g^2$, is denoted $X\,^1\Sigma_g^+$.

7.6 Bonding in Diatomic Molecules

As described in Sec. 7.4, the in-phase combination of two hydrogen $1s$ orbitals in H_2 leads to a bonding orbital ($1\sigma_g$), whereas the out-of-phase combination leads to an antibonding orbital ($1\sigma_u$). Similarly, linear combinations of two hydrogen $2s$ orbitals generate the $2\sigma_g$ and $2\sigma_u$ MOs. If we bring together two $2p_z$ orbitals, where z is the internuclear axis, then the in- and out-of-phase combinations generate the $3\sigma_g$ and $3\sigma_u$ orbitals [see Fig. 24(a)]. If the two atomic centers are different, so that the molecule has $C_{\infty v}$ symmetry, then the g and u labels no longer apply and so the bonding and antibonding combinations of $2p_z$ orbitals are denoted 5σ and 6σ, respectively, with the first four σ orbitals corresponding to bonding and antibonding combinations of the $1s$ and $2s$ functions. The analogous combinations of $2p_x$ orbitals (or, equivalently, $2p_y$ orbitals) have very different symmetry properties [see Fig. 24(b)]. The two bonding combinations $2p_x+2p_x$ and $2p_y+2p_y$ constitute a degenerate pair which is labeled 1π in $C_{\infty v}$ symmetry, or $1\pi_u$ in $D_{\infty h}$ symmetry. The corresponding antibonding combinations, $2p_x-2p_x$ and $2p_y-2p_y$, constitute a degenerate pair that is labeled 2π in $C_{\infty v}$ symmetry, or $1\pi_g$ in $D_{\infty h}$ symmetry.

A schematic molecular-orbital energy-level diagram of O_2 is shown in Fig. 25, in which the arrows signify which orbitals are occupied by electrons of α and β spins. It can be seen that the electronic configuration of the ground state of this molecule is $1\sigma_g^2 1\sigma_u^2 2\sigma_g^2 2\sigma_u^2 3\sigma_g^2 1\pi_u^4 1\pi_g^2$. According to Hund's rule, electrons in partially filled degenerate orbitals, such as $1\pi_g$, preferentially couple to the maximum possible spin ($S=1$, in this case). The prediction of unpaired electrons in O_2 and thus the rationalization of the paramagnetism of this molecule was one of the early successes of MO theory.

In extending this type of qualitative argument so as to generate schematic molecular energy-level diagrams, there are five very useful principles:

1. There must be no symmetry reason why the atomic functions cannot overlap. In a diatomic molecule, a $2s$ function on one center cannot mix with $2p_x$ on another: the regions of positive and negative contributions cancel to give a net overlap of zero. For diatomic molecules, "allowed" contributions include $s\pm s$, $s\pm p_z$, $p_z\pm p_z$, $p_x\pm p_x$, and $p_y\pm p_y$.
2. Bonding MOs arise from combinations of atomic orbitals that have similar energies and that have a large value for the overlap integral.
3. For every bonding MO there is a corre-

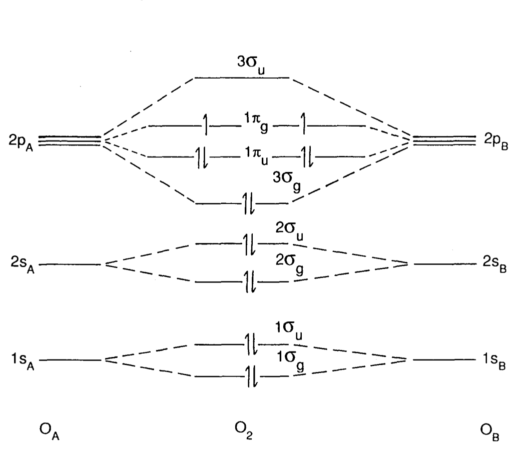

FIG. 25. A schematic molecular-orbital energy-level diagram for O_2 ($D_{\infty h}$). The arrows denote the electron configuration.

sponding antibonding MO.

4. Atomic orbitals that do not mix with functions on other centers are neither bonding nor antibonding. Such orbitals are said to be nonbonding.
5. It is usually necessary only to consider the valence electrons.

As a simple heteronuclear example, consider OH with nine electrons. This molecule dissociates to $O(1s^22s^22p^4)$ and $H(1s)$. The oxygen $1s$ and $2s$ orbitals do not enter into the bonding and are designated 1σ and 2σ, respectively. Of the four $2p$ electrons on oxygen, the one in the $2p_z$ orbital overlaps with the $1s$ function on hydrogen so as to form a σ bond (the 3σ MO). The remaining three electrons in the $2p_x$ and $2p_y$ orbitals occupy the 1π orbitals, so that the electron configuration of the ground state of OH is written $1\sigma^22\sigma^23\sigma^21\pi^3$. Bonding orbitals in heteronuclear molecules often have significantly more electron density in the vicinity of one nucleus than around the other. One such example occurs in the $1\sigma^22\sigma^2$ electronic configuration of the ground state of LiH ($X\,^1\Sigma^+$). In this case, the doubly occupied 1σ MO is effectively $1s^2$ on lithium. The 2σ bonding orbital, which formally arises from the overlap of Li$(2s)$ and H$(1s)$, has sufficient amplitude on the hydrogen center that the molecule can be said to have substantial Li^+H^- character.

Finally, we consider the ground state of HBr molecule with six valence electrons: H$(1s)$ and Br$(\cdots 4p^5)$. Significant $4s$-$4p$ mixing (or hybridization) is not likely for an element so far to the right of the periodic table. The combination H$(1s)$ + Br$(4p_z)$ leads to a σ-bonding MO. The corresponding out-of-phase combination leads to an antibonding MO of σ symmetry. It is common, particularly in organic chemistry, to denote the antibonding character by means of an asterisk, as in σ^*. For symmetry reasons, there can be no net overlap between the $1s$ orbital on hydrogen and the $4p_x$ or $4p_y$ orbitals on bromine. This means that the (degenerate) $4p_x$ and $4p_y$ orbitals give rise to a nonbonding MO of π symmetry.

A schematic energy-level diagram for HBr, based on these arguments, is presented in Fig. 26, and suggests an electronic configuration of $\sigma^2\pi^4$. We can predict a bond order of unity, i.e., a single H–Br σ bond, and four nonbonding valence electrons. Removal of an electron from the highest occupied MO (π nonbonding) should not significantly alter the bond length or the dissociation energy. Similarly, promoting one of the π electrons of the neutral molecule into the antibonding σ^* orbital should generate an excited electronic state with a weaker and longer bond. All of these various predictions have been confirmed by means of appropriate spectroscopic measurements.

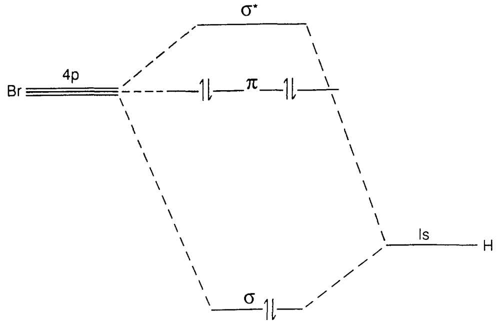

FIG. 26. A schematic molecular-orbital energy-level diagram for HBr ($C_{\infty v}$). The arrows denote the electron configuration.

7.7 Basis Sets for Molecular Calculations

A dominant factor in determining the accuracy of molecular electronic-structure calculations is the quality of the basis set $\{\chi_p\}$ in terms of which the orbitals are expanded:

$$\varphi_i = \sum_p c_{ip}\chi_p \,. \qquad (158)$$

In the most extensive contemporary calculations of correlation energy (see Sec. 8) the largest source of uncertainty is the basis-set truncation error. The calculated values of molecular properties tend to be particularly sensitive to the quality of the basis set.

In principle, molecular wave functions could be expanded in terms of a one-center expansion, as in the case for atoms. Such expansions have a number of potentially useful features, such as the ease of integral evaluation and the avoidance of linear dependence. However, one-center basis sets are only very rarely used in molecular studies, because the rate of convergence with the size of basis set is unacceptably slow. The vast majority of molecular calculations use multicenter basis sets, with basis functions centered on each of the nuclei.

Computation of the energy requires integrals of the one- and two-electron parts of the clamped-nucleus Hamiltonian introduced in Sec. 7.1. The computationally most expensive integrals are the two-electron integrals, arising from the electron–electron repulsion terms in the Hamiltonian:

$$(\chi_a\chi_c|\chi_b\chi_d) = \int \chi_a(1)^*\chi_b(2)^*(1/r_{12}) \times\chi_c(1)\chi_d(2)dr_1^3dr_2^3. \qquad (159)$$

The choice of functional form for the basis functions is a compromise between the ease of calculating accurate values of such integrals and the number of functions required to represent adequately the orbitals on which the total wave function is based.

The most commonly encountered functional forms are known as Slater-type orbitals (STOs) and Gaussian-type orbitals (GTOs). The STOs, or exponential-type functions, take the form

$$\chi = N_{n\zeta}\, r^{n-1} \exp(-\zeta r) Y_{lm}(\theta,\phi), \qquad (160)$$

in which r,θ,ϕ are polar coordinates, n,l,m are the usual atomic quantum numbers, and $N_{n\zeta}$ is a normalization constant. The value of ζ is usually referred to as the exponent. Basis functions of this type generally provide rapidly convergent representations of atomic and molecular orbitals. Note that the radial parts of these functions are nodeless, unlike hydrogen-atom orbitals.

"Minimal" STO basis sets consist of a single basis function for each occupied orbital of the constituent atoms. For a molecule containing a carbon atom, for example, this implies a 1*s* STO, a 2*s* STO, and three equivalent 2*p* STOs, each corresponding to a different value of *m*. "Double-ζ" basis sets allow two STOs with different exponents for each occupied atomic orbital, while "triple-ζ" basis sets allow three values of ζ.

In principle, all of these exponents could be taken from atomic calculations or they could be modified to take account of the molecular environment. The latter is especially important for functions centered on hydrogen nuclei, for which it is necessary to use exponents greater than the atomic value of unity. In the case of negative ions, for example, it is essential to include suitable diffuse (low-ζ) functions. In some calculations, the STO exponents have been generated by geometric series.

These various basis sets are almost always augmented with so-called "polarization functions," which correspond to the next value of *l*, i.e., *p* functions on hydrogen atoms or *d* functions on carbon atoms, for example. A double-ζ basis set augmented with polarization functions is often denoted DZP or DZ+P. Such polarization functions are essential for obtaining good descriptions of molecular systems and may have a significant effect on calculated values of molecular properties while contributing relatively little to the energies of the separated atoms.

Particular care is required in the design of large, flexible basis sets suitable for calculations that go beyond the SCF approximation. Fairly extensive tables have been published of STO basis sets suitable for molecular calculations.

STOs are still extensively used in calculations on small linear molecules, particularly diatomic molecules. However, multicenter two-electron integrals involving STOs are computationally very expensive for nonlinear systems, and it may be difficult to obtain the integrals to the required accuracy, particularly for functions with $l > 1$. This is the main reason for the dominance of GTOs as basis functions in contemporary molecular calculations.

A Cartesian GTO can be written in the form

$$\chi = N_{abc\alpha} x^a y^b z^c \exp(-\alpha r^2), \qquad (161)$$

where x,y,z are the usual Cartesian coordinates, a,b,c are integers that determine the type of GTO, α is the exponent, and $N_{abc\alpha}$ is a normalization constant. An *s* function corresponds to $a=b=c=0$—notice that there is no dependence on *n*, the principal quantum number. In the case of *d* functions $(a+b+c=2)$ there are six unique components instead of the usual five. This is equivalent to adding an extra *s* function to the basis set. Similar considerations apply to *f* functions.

As well as Cartesian GTOs, it is also common to use spherical GTOs in the form

$$\chi = N_{l\alpha} r^l \exp(-\alpha r^2) Y_{lm}(\theta,\phi). \qquad (162)$$

For *s* and *p* functions, the Cartesian and spherical GTOs are essentially equivalent, but there are only five components for a *d* function in the case of a spherical GTO.

Gaussian functions provide very poor representations of atomic and molecular orbitals in the regions of space close to the nucleus on which they are centered, since GTOs, unlike STOs, do not possess a cusp. In order to achieve a given accuracy, it is usually necessary to use significantly more GTOs than STOs, particularly for core orbitals. A factor of 3 in the number of functions required is not uncommon. Furthermore, the occurrence of $-r^2$ rather than $-r$ in the exponent means that the long-range form of a GTO is also inappropriate. This necessitates particular care in calculations of properties such as polarizabilities and in studies of intermolecular forces.

In view of the problems just outlined, it might seem surprising that GTOs are used at all. However, these disadvantages are more than offset by two further considerations: the ease with which the integrals can be computed accurately and the form of the derivative of a Gaussian function (see Sec. 9.1). The overwhelming majority of molecular electronic-structure calculations employ GTOs. It is straightforward to show that the product of two Gaussian-type orbitals on centers *A* and *B* takes the form of a Gaussian-type orbital centered on a point *P* that is located on the line between *A* and *B*. Consequently, the calculation of two-electron integrals becomes very straightforward.

The Gaussian-type functions described so far are termed "primitive" functions, because each basis function consists of a single GTO with a single exponent. A practical means of reducing the size of the basis set without significantly reducing the accuracy of molecular electronic-structure calculations is to form new basis functions, called contracted GTOs, which are fixed linear combinations of primitive GTOs:

$$g=\sum_j k_j\chi_j\,, \tag{163}$$

where the k_j are called contraction coefficients. The high-exponent primitive GTOs, which contribute mostly to the inner-shell or core orbitals, can often be contracted with little loss of accuracy. The more diffuse primitive GTOs are not usually contracted in this way.

The need for diffuse functions for negative ions and for polarization functions arises in the same way for GTOs as for STOs. Contracted Gaussian basis sets are often given acronyms such as DZP or TZVP, which give an indication of the type or the size of STO basis set that would be of similar accuracy. A TZVP basis set for the carbon atom is listed in Table 3; the letters TZVP or TZV+P denote "triple-ζ valence" (the core is less well represented) plus polarization (the single set of *d* functions). Some of the basis sets in common use with some of the standard computational packages were originally based on least-squares fits of Gaussian functions to STO basis sets and bear names such as 4-31G, which reflect this parentage. A 4-31G basis set for the carbon atom is also listed in Table 3.

The computation of the one- and two-electron integrals accounts for a significant proportion of the total computational requirements of a geometry optimization using an SCF wave function. One alternative to the *ab initio* calculation of all the integrals is to use semiempirical approximations. Such methods treat the integrals, and related quantities, as parameters to be obtained by least-squares fits to experimental data, or to the results of small *ab initio* calculations, for a large sample of molecules. A number of semiempirical techniques are in widespread use, particularly for large organic molecules, for which *ab initio* calculations might be rather expensive. Semiempirical calculations are particularly useful as a means of estimating equilibrium geometries and heats of formation and for obtaining an approximate idea of the qualitative features of the molecular orbitals. The forms of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) are particularly useful for interpreting organic reactivity, using "frontier orbital theory." However, semiempirical techniques are not in general suitable for high-quality calculations of entire potential-energy surfaces.

8. METHODS FOR TAKING ACCOUNT OF ELECTRON CORRELATION IN MOLECULES

While there has been a great deal gained in terms of qualitative descriptions of molecular electronic structure via independent-particle methods, an up-to-date account of current

Table 3. Examples of contracted GTO basis sets used in molecular electronic-structure calculations. The different contractions, which have not been normalized, are separated by extra space.

Type	Contraction coefficient	Exponent
(a) 4-31G basis set for carbon		
s	0.017 726	486.966 928
s	0.123 478	73.371 094
s	0.433 875	16.413 458
s	0.561 504	4.344 984
s	−0.121 384	8.673 525
s	−0.227 338	2.096 619
s	1.185 174	0.604 651
p	0.063 545	8.673 525
p	0.298 268	2.096 619
p	0.762 103	0.604 651
s	1.000 000	0.183 558
p	1.000 000	0.183 558
(b) TZVP basis set for carbon		
s	0.000 776	9471.0
s	0.006 218	1398.0
s	0.033 575	307.5
s	0.134 278	84.54
s	0.393 668	26.91
s	0.544 169	9.409
s	0.248 075	9.409
s	0.782 844	3.500
s	1.000 000	1.068
s	1.000 000	0.4002
s	1.000 000	0.1351
p	0.016 295	25.37
p	0.102 098	5.776
p	0.340 228	1.787
p	0.668 269	0.6577
p	1.000 000	0.2480
p	1.000 000	0.0910
d	1.000 000	0.7200

work in the field would be somewhat deficient without some mention of methods that extend this picture. As discussed in Sec. 2, independent-particle models, such as Hartree-Fock theory, include only the average electron–electron interactions, whereas the motion of the electrons is in fact strongly correlated. In molecular electronic-structure work, the impetus to move beyond the Hartree-Fock model comes not just from the desire to reduce the correlation error, defined as the difference in energy between the exact solution of the clamped-nucleus Schrödinger equation and the Hartree-Fock solution. Equally important is allowing for flexibility in the electronic wave function to describe a changing charge distribution as the nuclear framework distorts and bonds are made or broken. Particularly for systems for which excited electronic states, molecular properties (see Sec. 9), and significant portions of the potential-energy surface(s) are required, it is necessary to consider calculations which account for some degree of charge rearrangement, recoupling of electron spins, and electron correlation. Many methods have been developed for taking into account the most important electron-correlation effects. The purpose of this section is to give a flavor of the more standard and the most widely used approaches for obtaining more accurate electronic energies, wave functions, and properties for molecular systems.

8.1 Multiconfiguration Self-Consistent-Field Method

A fundamental deficiency of the MO-theory SCF approach to molecular electronic structure is that it does not provide descriptions of uniform quality as the internuclear distances are varied. In particular, the single-configuration wave function does not describe correctly the process of molecular dissociation. As was shown in Sec. 7.4, the configuration $1\sigma_g^2$ describes H_2 reasonably well near its equilibrium geometry but corresponds at large nuclear separations to a mixture of the correct ground-state asymptote, $H(^2S)+H(^2S)$, and the ionic asymptotes $H^+ + H^-(^1S)$ and $H^-(^1S) + H^+$. In order to eliminate the ionic asymptotes, a linear combination of two configurations is required, namely $1\sigma_g^2 - 1\sigma_u^2$. In this section, we describe the extension of the Hartree-Fock theory single-determinantal wave function to a superposition of Slater determinants. This treatment is analogous to the description of the multiconfiguration Hartree-Fock (MCHF) method for atoms in Sec. 5.4 but, for historical reasons, the method is usually known in the molecular literature as multiconfiguration self-consistent-field (MCSCF) theory.

The main objective of a multiconfiguration self-consistent-field calculation is usually to obtain qualitatively correct molecular wave functions that properly describe all the dissociation processes of interest and that take account of any near-degeneracy effects. Such a wave function Ψ_{MC} can be written in the form

$$\Psi_{\mathrm{MC}} = \sum_I C_I \Phi_I, \qquad (164)$$

in which the configurations Φ_I are constructed from a common set of orthonormal orbitals, which are fully optimized together with the configuration interaction or CI coefficients C_I.

The first step in selecting the configurations to be included in an MCSCF wave function is to decide which, if any, orbitals are to remain doubly occupied in all configurations. These "core" orbitals, and the electrons which occupy them, are termed "inactive." The remaining orbitals are partitioned into the "active" or "internal" orbitals, which are allowed to have different occupancies in different configurations, and the "external" or "virtual" orbitals, which are unoccupied in all configurations. During the calculation, internal and external orbitals are mixed with one another, by means of orthogonal rotations, until the optimum internal orbitals are obtained. It is also possible, by means of additional orbital rotations, to allow the doubly occupied core orbitals to relax, in a self-consistent fashion, in the optimized field defined by the active electrons.

A common choice for the internal orbitals is the set of valence orbitals. In the case of LiH, for example, the 1σ orbital corresponds to $\mathrm{Li}(1s^2)$, and it is treated as a core orbital. The valence orbitals in this case arise from $1s$ on hydrogen and from $2s$ and $2p$ on lithium. These give rise to the valence 2σ, 3σ, and 1π orbitals, of which only the 2σ orbital is occupied in the SCF description of the ${}^1\Sigma^+$ ground state. It is more usual to include *all* configurations of the correct spin multiplicity and spatial symmetry that can be generated by distributing the two active electrons among the internal orbitals, rather than to attempt to select just those configurations that are essential for correct dissociation. MCSCF wave functions in which the list of configurations consists of all allowed distributions of the active electrons in the internal orbitals are known as complete active-space self-consistent-field or CASSCF wave functions. The CASSCF wave function for our LiH example would therefore include configurations $1\sigma^2 2\sigma^2$, $1\sigma^2 3\sigma^2$, $1\sigma^2 2\sigma 3\sigma$, and $1\sigma^2 1\pi^2$. Such CASSCF calculations are typically more straightforward and computationally less expensive than smaller MCSCF calculations based on the same active space. This occurs, in part, because the CASSCF wave function is invariant to arbitrary unitary transformations of the internal orbitals.

The energy corresponding to an MCSCF wave function can be written in the form

$$E = \sum_{t,u} \gamma_{tu} h_{tu} + \sum_{t,u,v,w} \Gamma_{tuvw} (tu|vw), \qquad (165)$$

where γ and Γ are known as one- and two-electron density matrices, respectively. h_{tu} are one-electron integrals in the MO basis, and $(tu|vw)$ are the corresponding two-electron integrals defined by Eq. (159). The density matrices are composed of orbital coefficients c_{ip}, configuration-interaction coefficients C_I, and various coupling constants. Much effort has been expended to develop highly efficient algorithms that allow the construction of the density matrices without the need to precompute and store extensive lists of intermediate quantities, and to minimize both the number of iterations and the number of times that the integrals must be transformed to the MO basis.

The MCSCF method was initially developed as a natural extension of the Hartree-Fock method and was used to describe dissociation processes correctly and to take some account of electron correlation. A comparison of the LiH $X\,{}^1\Sigma^+$ RHF energy (at R_e) of $-7.987\,313$ hartrees with the corresponding energy of $-8.021\,321$ hartrees from a small, 15-configuration MCSCF calculation demonstrates an improvement of almost 1 eV. Whereas dissociation energies for LiH obtained using single-configuration wave functions were totally unreliable, often being in error by several electron volts, this small MCSCF calculation leads to a value of 2.41 eV, to be compared with an experimental value of 2.52 eV (Stwalley and Zemke, 1992). Nowadays, an MCSCF calculation is usually used as the first step in a more sophisticated configuration-interaction (CI) calculation (cf. Secs. 5.3 and 8.2): the MCSCF step is used to obtain a multiconfiguration zeroth-order wave function based on an appropriate set of valence orbitals. Molecular MCSCF or, more often, CASSCF calculations with 10^2–10^4 configurations are now essentially routine, and calculations with much longer lists of configurations are not uncommon. Many of the standard quantum chemistry computational packages now include the facility to carry out such calculations.

8.2 Configuration Interaction

A configuration-interaction (CI) wave function Ψ_{CI} consists of a linear combination of configurations Φ_I constructed from a common *fixed* set of orthonormal orbitals:

$$\Psi_{\mathrm{CI}}=\sum_I C_I\Phi_I\,. \tag{166}$$

Superficially, at least, this type of wave function looks very similar to an MCSCF expansion, but the main purposes of the two types of calculation and the classes of configurations that are usually included are somewhat different.

The correlation energy taken into account by an MCSCF wave function, so as to produce qualitatively correct features, is often referred to as "nondynamical" correlation. In general, an MCSCF calculation recovers much less than one half of the total correlation energy. The purpose of a CI calculation is to recover a large proportion of the remaining correlation, termed "dynamical" correlation, and to obtain quantitatively accurate results. It should be stressed that there is no sharp dividing line between nondynamical and dynamical correlation, although the (approximate) distinction between them remains useful.

The simplest strategy for a closed-shell molecule is to start from the single-configuration SCF wave function Φ_0, and to partition the occupied orbitals into core and internal sets, in much the same way as in the MCSCF case. The usual form of CI calculation for such a system then includes all configurations of the correct spin multiplicity and spatial symmetry that can be generated by single and double replacements of the internal orbitals in Φ_0 by the external orbitals, i.e., those that are not occupied in Φ_0. Such a calculation is typically called a "singles and doubles" CI, or SDCI, and is further labeled "single reference," because all of the excitations stem from a single configuration. The term multireference CI, or MRCI, usually denotes a configuration-interaction expression that includes all single and double excitations from more than one reference configuration, such as the configurations obtained in an MCSCF calculation. It can be particularly important for Φ_0 to be a multireference function if several points on a potential-energy surface are to be studied. In general, the number of external orbitals is very much larger than the number of internal orbitals, and so the largest class of configurations consists of those that have two electrons in external orbitals. The total number of molecular orbitals and consequently the size of a CI calculation are determined by the total number of atomic basis functions.

Whereas the MCSCF energy expression is fourth order in the orbitals, the CI problem is a linear eigenvalue problem, in that only the CI coefficients C_I are determined. The eigenvalue problem to be solved takes the form

$$(\mathbf{H}-E\mathbf{1})\mathbf{C}=\mathbf{0}, \tag{167}$$

in which $\mathbf{C}$ is a vector of the CI coefficients and $\mathbf{H}$ is the matrix of the Hamiltonian operator, with elements $\langle\Phi_I|H|\Phi_J\rangle$. The different roots of this secular problem correspond to the different electronic states. The computational effort typically scales approximately linearly with the number of roots extracted from the CI calculation. The expression for the CI energy in terms of density matrices and one- and two-electron integrals takes exactly the same form as in the MCSCF case. A typical CI calculation involves 10^4–10^6 configurations, although larger calculations are now possible, and so explicit construction and diagonalization of $\mathbf{H}$ are not a practical proposition. Instead, an iterative strategy is usually adopted.

It is usual to exploit the fact that different molecular symmetries do not couple through the Hamiltonian operator H to construct different wave functions and to solve different secular problems for each symmetry. Provided the wave function has been properly constructed from an appropriate set of orbitals, it is relatively straightforward to obtain the lowest root of each symmetry with the same accuracy as the ground state. If excited states of a given symmetry are required, successively higher roots can be extracted from the secular equation. It is important, however, to make sure that orbitals have been included that are appropriate to the *occupied* orbitals of the excited state(s) and that all necessary reference configurations are present in Φ_0. It is often difficult to describe with equal accuracy states of very different character; this can be a particular problem for molecular Rydberg states, and particular care is required in order to treat Rydberg and valence states on an equal footing. These remarks can be illustrated by looking again at the LiH molecule (Stwalley and Zemke, 1992), which in its first excited state ($A\ ^1\Sigma^+$) has the RHF configuration $1\sigma^2 2\sigma 3\sigma$. If wave functions are required for both of the $X\ ^1\Sigma^+$ and $A\ ^1\Sigma^+$ states, then it is important to make sure that the 3σ orbital takes the form of an occupied orbital of the $A\ ^1\Sigma^+$ state rather than the form of a correlating orbital for the ground state. If the $B\ ^1\Pi$ state of LiH ($1\sigma^2 2\sigma 1\pi$) were also required, then it would be much more appropriate to construct Φ_0 from orbitals obtained via an MCSCF calculation on the $B\ ^1\Pi$ state instead of the analogous calculation on the ground state. In this way, the valence 1π orbital would be much better determined. The dissociation energy of the $X\ ^1\Sigma^+$ state of LiH from a large CI calculation is 2.48 eV, which is within 0.04 eV of the experimental value and which shows an improvement of 0.07 eV over the 15-configuration MCSCF value.

A full CI, or FCI, calculation is one in which all possible replacements of internal orbitals by external orbitals are included in the wave function. It corresponds to the exact solution, within a given one-particle basis set, of the usual electronic structure problem. Usually such calculations are prohibitively expensive when using sufficiently large one-particle basis sets to obtain quantitative accuracy. Single and double excitations are, however, by far the most important for most purposes. Truncated CI expansions that start from a qualitatively incorrect reference space are not likely to give high-quality results, even with very large numbers of configurations. With this proviso, affordable FCI calculations generally give much poorer results than truncated CI expansions with larger basis sets.

A disadvantage of singles and doubles CI expansions is that they are not size consistent. As a simple example, consider two well-separated helium atoms. Clearly, the energy of this system should be exactly twice that of a single helium atom, and if this is the case, then the calculation is said to be size consistent. More formally, the wave function for this system should be equal to the product of the two helium-atom wave functions. In the case of an ordinary SDCI calculation, the list of configurations includes those that are double replacements of occupied orbitals on one center by virtual orbitals on the other, and so the total wave function is not the product of the wave functions of the two fragments. In a full CI calculation, these unwanted contributions to the energy and wave function are cancelled exactly by higher-order excitations. Various simple schemes have been suggested that allow estimates to be made of the size-consistency errors in truncated CI expansions, without the need to include further classes of excitations.

The interpretation of large CI wave functions is not very straightforward, unless it turns out that the wave function is overwhelmingly dominated by a small number of configurations. A practical alternative to examining long lists of CI coefficients is to generate natural orbitals. The natural orbitals for a given approximate CI calculation are obtained by diagonalizing the one-electron density matrix, γ_{tu}. The diagonal elements of the density matrix in this new basis are termed occupation numbers and typically take fractional values. The interpretation of the CI wave function then relies on examining the form of the natural orbitals and the values of the occupation numbers, which give a clear idea of the importance of a given orbital in the CI expansion, such as its effectiveness as a correlating function.

Configuration mixing is a very slowly convergent process for reproducing those terms in the exact wave function that are linear in the interelectronic coordinates r_{12}. One alternative to a conventional CI approach is the explicit inclusion into an approximate wave function of terms linear in r_{12}. The chief drawback of such a strategy is that it introduces additional classes of "difficult" integrals. However, significant process is currently being made by means of suitable approximations to these integrals.

8.3 Many-Body Perturbation Theory

The basic aim of perturbation theory is to account for the changes induced in a model system by a small disturbance or perturbation. It provides a scheme for obtaining successively more accurate refinements of a model problem. A good account of perturbation theory for molecular problems has been given by Szabo and Ostlund (1982).

When perturbation theory is used to treat electron correlation in molecular systems, it is often termed many-body perturbation theory or MBPT. In its simplest form, which is applicable to single-determinant closed-shell RHF wave functions, this approach is more commonly known as Møller-Plesset (MP) theory. The SCF energy is equal to the sum of the zeroth-order and the first-order energies. In the notation of Sec. 5.5, we have

$$\begin{aligned} E^{(0)}+E^{(1)} &= \langle\alpha|H_0|\alpha\rangle+\langle\alpha|V|\alpha\rangle \\ &= \langle\alpha|H|\alpha\rangle, \end{aligned} \tag{168}$$

where $|\alpha\rangle$ is the RHF wave function. The lowest-order perturbation correction to the energy is thus the second-order (MP2) correction, which can be written in the form

$$E^{(2)}=-\frac{1}{4}\sum_{i,j}\sum_{a,b}\frac{[(ia|jb)-(ib|ja)]^2}{(\epsilon_a+\epsilon_b)-(\epsilon_i+\epsilon_j)}. \tag{169}$$

MP2 calculations are relatively inexpensive and typically recover about 80% of the correlation energy for a molecular system near its equilibrium geometry. Whereas SCF (RHF) calculations typically underestimate the lengths of single bonds by ~0.02 Å and give harmonic frequencies to ~10%, the analogous MP2 calculations underestimate single bonds by ~0.003 Å and give harmonic frequencies to ~2%. MP2 is less reliable for multiple bonds.

Evaluation of the third-order and fourth-order Møller-Plesset energies necessarily involves significantly more computational effort. Many standard packages permit the calculation of MP2, MP3, and MP4 energies, although some of the MP4 terms are sufficiently expensive that the user may be given the option to compute only selected terms.

By analogy to single-reference SDCI, it is self-evident that the results from a Møller-Plesset calculation are unlikely to be very reliable if the final wave function is not dominated by the unperturbed wave function. Methods such as MP4 are not generally applicable to the calculation of entire potential surfaces, and they do not describe correctly the dissociation of closed-shell molecules. As mentioned above, MBPT calculations are not variational. On the other hand, MBPT has the significant advantage of size consistency (see Sec. 8.2), unlike most CI-based methods.

The Møller-Plesset approach can also be used with open-shell UHF wave functions, in which case the *n*th-order terms are denoted UMP*n*. Even small amounts of spin contamination can lead to very serious convergence problems. A partial solution to this problem is to project out pure spin states at each order of perturbation theory, but many of the standard packages do not offer this feature. It is not common for terms beyond fourth order to be included in MBPT calculations on molecular systems, and so hard evidence on the convergence properties of the perturbation series is scarce. Progress for Møller-Plesset theory based on open-shell RHF wave functions or for multireference MBPT has been fairly slow for molecular calculations.

8.4 Coupled-Cluster Methods

A family of methods that is currently gaining in popularity and that has undergone extensive recent development consists of the coupled-cluster approaches. Part of the motivation, as in the case of MBPT, is that coupled-cluster wave functions are usually size consistent. However, coupled-cluster calculations do not generally offer a variational upper bound to the energy. There has been a large number of variations on the basic theme of coupled-cluster theory, and the present account is necessarily restricted to a brief outline of the simplest version.

For a closed-shell RHF configuration Φ, the coupled-cluster wave function Ψ_{CC} is defined by

$$\Psi_{\mathrm{CC}}=\exp(T)\Phi, \tag{170}$$

in which T is an excitation operator, which generates single, double, triple, etc., replacements of occupied orbitals in Φ by virtual orbitals. The exponential is defined in the usual way, i.e., by a power series expansion:

$$\exp(T)=1+T+\tfrac{1}{2}T^2+\cdots. \tag{171}$$

In coupled-cluster singles and doubles (CCSD) calculations, T is restricted to the sum of single- (T_1) and double- (T_2) excitation operators, so that terms up to quadruple excitations are included in the wave function, via $\tfrac{1}{2}T_2^2$. The corresponding energy is given by

$$E=\langle\Phi|H|(1+T_2+\tfrac{1}{2}T_1^2)\Phi\rangle, \tag{172}$$

with analogous equations to determine the CCSD wave function.

Coupled-cluster calculations such as CCSD, which start from a single determinant Φ, are not suitable for calculations of entire potential surfaces—cf. SDCI or MP*n* methods (where *n* indicates the order of perturbation theory)—but they do yield excellent equilibrium properties for systems that are reasonably well described by a single-configuration closed-shell RHF wave function. It is also possible to carry out coupled-cluster calculations starting from UHF wave functions, but this raises again the problem of spin contamination. Calculations based on open-shell RHF wave functions have been reported in the literature.

There are obvious links between the coupled-cluster and MBPT methods, but progress in developing multireference methods for the coupled-cluster approach appears to be much more promising.

8.5 Modern Valence-Bond Theory

Actual *ab initio* implementations of classical valence-bond theory are somewhat disappointing. The same is true of calculations in which molecular-orbital theory SCF wave functions are projected onto classical VB structures. In order just to match the SCF energy in a given basis set, it turns out that it is usually necessary to include large numbers of structures, particularly those of ionic character, with separation of charges. All in all, the original simple intuitive model is clouded by the requirement for multistructure wave functions.

Recent years have seen a significant resurgence of interest in methods for the optimization of nonorthogonal orbitals to be used in compact and accurate VB descriptions of potential-energy surfaces of ground and excited states. Two of the key problems of classical VB theory have been largely overcome, namely, the proliferation of ionic contributions, even in intuitively covalent systems, and the difficulty of performing calculations with nonorthogonal orbitals. Sophisticated techniques from group theory lead to efficient algorithms, but the preferred method now appears to be the expansion of the wave function in terms of Slater determinants and the evaluation of matrix elements using cofactors.

The *exact* solution Ψ_{ex} of the usual electronic structure problem for an N-electron system can be written in the form

$$\Psi_{\mathrm{ex}} = \sum_k c_{Sk} (N!)^{-1/2} \mathscr{A}(\Phi \Theta^N_{SM;k}), \tag{173}$$

in which Φ is an exact N-electron spatial function (with no particular permutational symmetry) and $\mathscr{A}$ is the antisymmetrizing operator, which ensures that the simultaneous interchange of the space and spin coordinates of two electrons induces a change in the sign of the wave function (Pauli principle). Each of the allowed ways of coupling together the individual electrons spins so as to achieve a total spin of S (and projection M) is labeled by the index k, and the $\Theta^N_{SM;k}$ form a complete set of N-electron spin functions. The c_{Sk} reflect the relative importance of the different modes of spin coupling and may be termed spin-coupling coefficients.

The spin-coupled wave function corresponds to approximating the spatial function Φ by a single configuration constructed from distinct, singly occupied nonorthogonal orbitals. The spin-coupled wavefunction Φ_{SC} can be written in the form

$$\Psi_{\mathrm{SC}} = \sum_k c_{Sk} (N!)^{-1/2} \mathscr{A}(\varphi_1 \varphi_2 \cdots \varphi_N \Theta^N_{SM;k}).$$

The orbitals are expanded in a one-particle atom-centered basis set, as in MO theory, and are fully optimized simultaneously with the spin-coupling coefficients. The spin-coupled wave function is an eigenfunction of S^2 and S_z, and so questions of spin contamination never arise. The spin-coupled description accounts for most of the available nondynamical correlation and remains qualitatively correct over the entire potential-energy surface, including the various dissociation pathways. The spin-coupling coefficients are particularly sensitive indicators of the changes occurring in the total wave function as bonds are made and broken.

In many cases, the descriptions that emerge show marked similarities to those postulated by classical VB theory, except for small distortions of the orbitals. These deformations, which correspond to the use of basis functions from more than one center, are crucial to the description and have the important consequence that just one spatial configuration overwhelmingly dominates the total wave function, without the need to include configurations with double occupancy of orbitals (ionic structures), as in classical VB theory. The spin-coupled description is invariably "simpler" and much more compact than the classical VB wave function, and in some cases leads to somewhat different interpretations of the electronic structure of polyatomic systems.

A widely used alternative to the spin-coupled wave function, known as the generalized valence-bond (GVB) wave function, is usually computed with various orthogonality constraints between pairs of singlet-coupled electrons so as to reduce the computational effort.

The spin-coupled orbitals in Ψ_{SC} can be partitioned into "core" and "active" sets, in much the same way as in MCSCF calculations, and for polyatomic systems it is usual to identify the active space with the valence electrons. The closest equivalent MO-theory wave function is a CASSCF wave function with all configurations generated by distributing N electrons in N orbitals. Because it is based on a single spatial configuration, the interpretation of the spin-coupled wave function in terms of the relevant physics and chemistry remains very straightforward.

This single-configuration description can be refined by means of a nonorthogonal CI calculation, known as a spin-coupled VB or SCVB calculation, but this inclusion of dynamical correlation does not appear to affect the essential physical picture provided by the spin-coupled wave function. The virtual orbitals, into which the electrons in occupied orbitals are excited, are constructed in such a way that they provide accurate descriptions of excited states. Consequently, the SCVB wave functions for the excited states are dominated by a very small number of spatial configurations and are easy to interpret. It is usual to include single and double excitations from the occupied orbitals of the spin-coupled description of the ground state into the first few virtual orbitals and possibly also equivalent excitations from the dominant configurations for the low-lying excited states. A typical SCVB calculation providing accurate and compact descriptions of, perhaps, 20 states of each symmetry of a polyatomic species consists of 10^2–10^3 spatial configurations. SCVB calculations are variational and are well suited to the description of entire potential surfaces, but they are not usually size consistent. However, the essentially localized character of the occupied and virtual orbitals makes it straightforward in many cases to restrict the classes of excitations so as to minimize the size-consistency error.

As a simple example of the utility of SCVB calculations we consider the simple singlet methylene insertion reaction (Cooper *et al.*, 1991)

$$CH_2(^1A_1)+H_2 \rightarrow CH_4, \tag{174}$$

which involves the complex process of forming two new C–H bonds while breaking the strong bond in molecular hydrogen. Note that CH_2 belongs to the same point group, C_{2v}, as does H_2O (see Sec. 7.5). According to simple arguments based on the orbital symmetries in the RHF wave functions for the reactants, the direct approach of H_2 to $CH_2(^1A_1)$ should be very unfavorable. This is because the nonbonding pair of electrons on carbon occupies an sp^2-like hybrid orbital, pointing along the z axis away from the H atoms; the interaction of this orbital with the doubly occupied $1\sigma_g$ orbital of H_2 is energetically unfavorable. Nonetheless, this insertion reaction is known experimentally to proceed rapidly with little or no activation barrier. The six spin-coupled orbitals for the six valence electrons of $CH_2(^1A_1)$ are depicted in Fig. 27, and it is easy to discern that orbitals φ_1–φ_4 describe two equivalent C–H bonds. The two remaining orbitals, φ_5 and φ_6, take the form of nonbonding sp^x-like hybrids ($x \approx 3$) on carbon, pointing like rabbit ears in the plane perpendicular to the molecular plane. In view of this orbital description, it is easy to envisage that the initial approach of H_2 to $CH_2(^1A_1)$ could be with its bond axis pointing along the direction of one of the nonbonding hybrids. As the molecules came together, the H atom furthest from the CH_2 would then swing around towards the second nonbonding hybrid. Detailed SCVB calculations confirm that this reaction can occur along this path with *zero* barrier.

8.6 Density-Functional Theory

The term "density-functional theory" is currently used to denote a whole family of methods in which the electronic energy is computed from functional forms involving the electron density $\rho(\mathbf{r})$. Indeed, according to the Hohenberg-Kohn theorem, the ground-state energy of a nondegenerate many-electron system is a unique functional of the electron

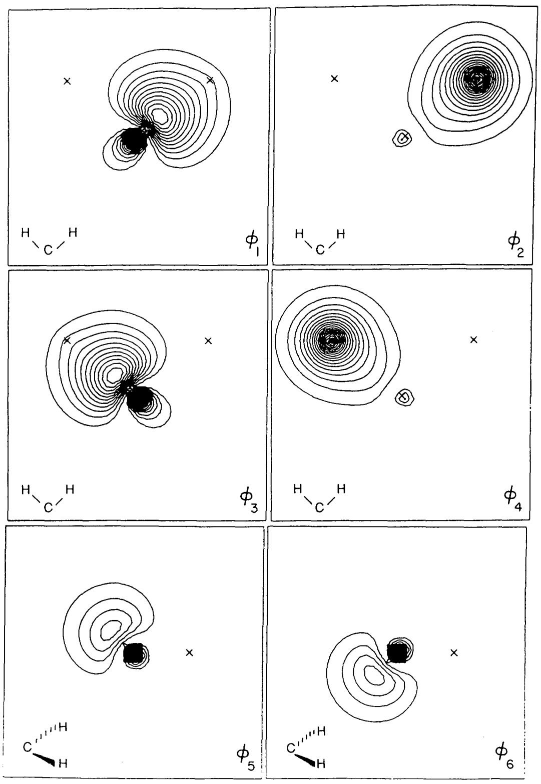

FIG. 27. Spin-coupled orbitals in CH_2 ($^1A_1$) shown as contours of $|\varphi_\mu|^2$. Spin-coupled orbitals φ_1–φ_4 are shown in the σ_v' (molecular) plane, but orbitals φ_5 and φ_6 are shown in the σ_v plane.

density. This theorem is a strong driving force behind density-functional theory but does not in itself provide a tractable computational procedure.

The most widely used practical schemes are based on the Kohn-Sham equations. A basic idea of this approach is that the exact many-electron wave function is mapped onto that of a model system of *noninteracting* electrons that has the same total electron density. In essence, the exact kinetic energy is replaced by the kinetic energy of the same number of noninteracting electrons in the lowest possible state. The electron density of this model system is constrained to be the same as that of the true system by means of an external electric field.

The total energy in the Kohn-Sham approach is written as the sum of the noninteracting kinetic energy, the electron–nuclear attraction, the "classical" electron–electron Coulomb repulsion, and a term E_{xc} that contains everything else. $E_{xc}[\rho(\mathbf{r})]$ includes the exchange and correlation energies (of the interacting system) as well as the difference in kinetic energy between the noninteracting and interacting systems. Explicitly, the Kohn-Sham equation for orbital ψ_i takes the form

$$\left(-(\tfrac{1}{2})\nabla^2+v(\mathbf{r})+\int\frac{\rho(\mathbf{r}')d^3r'}{|\mathbf{r}-\mathbf{r}'|}+v_{xc}(\rho(\mathbf{r}))\right)\psi_i =\epsilon_i\psi_i\,, \tag{175}$$

in which the first three terms correspond to the kinetic energy, the electron–nuclear attraction, and the electron–electron Coulomb repulsion, and the fourth is defined by

$$v_{xc}(\rho(\mathbf{r}))=\frac{\delta E_{xc}[\rho(\mathbf{r})]}{\delta\rho(\mathbf{r})}\,. \tag{176}$$

Various approximate exchange–correlation potentials v_{xc} have been proposed.

Provided a suitable v_{xc} has been chosen, iterative solution of the self-consistent-field Kohn-Sham equations to determine the orbitals, electron density, and total energy is relatively straightforward using either basis-set expansions or numerical methods. Approaches based on density-functional theory can provide accurate total electronic energies for molecules around the equilibrium geometry. Taken together with accurate atomic energies, these methods provide accurate dissociation energies. Density-functional theory is of particular value for molecules containing transition-metal atoms, for which it often provides results superior to those from CASSCF-CI approaches. The most efficient and accurate density-functional methods include nonlocal corrections to the exchange and correlation energies.

9. MOLECULAR PROPERTIES

9.1 Geometries and Gradients

The most obvious molecular properties to be determined are the location and character of minima on a potential-energy surface and of other stationary points, such as transition states. Significant advances have been made in recent years in the development of gradient techniques, so that the location of stationary points is now a routine task for many types of wave functions. In the present context, the term "gradient" is used to denote analytic derivatives of the total energy with respect to geometric parameters, such as bond lengths and angles.

A nuclear displacement can modify the energy of a CI wave function, for example, in three distinct ways: changes to the basis functions, because the centers on which these are located have been moved; changes to the MOs; and changes to the CI coefficients. The changes to the basis functions lead to terms in the gradient that involve two-electron integrals of the derivatives of the basis functions. The first and higher derivatives of a Gaussian-type orbital take the form of linear combinations of further GTOs, and so the integrals of basis-function derivatives are very straightforward. The same is not true of Slater-type orbitals, and this reinforces the preeminent role of GTOs in molecular calculations.

First derivatives are now available for most commonly encountered types of wave functions. Second derivatives are available in some cases, such as SCF, MP2, and CASSCF. These are useful not only for geometry optimization and the location of transition states, but also for the calculation of harmonic frequencies. Indeed, in the case of SCF wave functions, third and fourth derivatives can be calculated, making possible the calculation of anharmonic constants.

It is common, particularly for diatomic molecules, to use the calculated potential-energy curves (or surfaces) to solve the nuclear motion problem. This provides details of the vibrational and rotational energy levels of the molecular system. The wave functions $\chi_{\mathrm{nuc}}(\mathbf{R})$ and the corresponding energies can be used to generate a variety of spectroscopic constants. Comparisons of these with the equivalent experimentally derived quantities give useful information regarding the accuracy of the potential curves or surfaces. In many cases it is possible to provide reliable numerical values for constants that are very difficult to measure experimentally.

9.2 Energy Differences

Many of the quantities of interest that characterize a molecular electronic state or a manifold of states are derived from energy differences. For example, the dissociation energy D_e of a diatomic molecule is defined as the difference between the minimum energy at the equilibrium internuclear distance R_e and the energy of the component atoms at infinite separation. The excitation energy T_e of an excited molecular state is the difference between the excited-state energy and the ground-state energy, each evaluated at the respective equilibrium geometry. Other important quantities include the heights of barriers on a potential surface, as well as ionization potentials and electron affinities.

Because physical quantities are often related to energy differences, there tends to be greater emphasis in molecular calculations on obtaining potential curves or surfaces that closely parallel the exact solution, rather than attempting to generate the best possible energy at a particular geometry. However, even if the error made in an energy calculation varies as little as 0.01% as a function of nuclear coordinates or from one electronic state to another, then the errors in the corresponding energy differences can be substantial. In practice it is fairly difficult to minimize this error, but the most extensive calculations do provide estimates of important quantities that rival experimental measurements. Such calculations are of particular importance when the necessary experiments are very difficult. Electronic-structure studies are also crucial for trying to understand the magnitude of a quantity, and the way in which it varies with geometry or from state to state or from system to system.

In general, the correlation energy increases with the number of electrons in a system, but any given approximate method is likely to recover approximately the same *proportion* of the correlation energy for different systems. Difficulties arise then in the calculation of ionization potentials and electron affinities obtained as energy differences between N-electron and $(N-1)$-electron or $(N+1)$-electron systems. In addition, the correlation energy of a pair of electrons in the same spatial orbital is usually considerably greater than that of electrons in different spatial orbitals. This observation is of particular relevance to the calculation of dissociation energies, since there are usually more electron pairs in the electronic configuration of a molecule at its equilibrium geometry than there are in the constituent fragments at infinite separation. Consequently, a greater correlation error is likely in an approximate calculation at R_e

than at $R=\infty$, and so the calculated value of D_e will be too small.

Errors associated with excitation energies also usually reflect different degrees of electron pairing in different states. For example, the calculated value of T_e for the first excited state of LiH ($1\sigma^2 2\sigma 3\sigma$) is likely to be small simply because the calculated energy of the ground state ($1\sigma^2 2\sigma^2$) is subject to a larger correlation error. Assuming that the calculated dissociation energies of the two states are fairly accurate, then reliable estimates of excitation energies, ionization potentials, or electron affinities can be obtained by simple energy differences involving these dissociation energies and the experimentally determined asymptotic separations.

An additional problem, known as basis-set superposition error or BSSE, arises in studies of very weakly interacting systems and of intermolecular forces. The fundamental issue is that the individual basis sets used to describe the fragments A and B are necessarily incomplete, and so the quality of the merged basis set of the AB system, or supermolecule, will be geometry dependent. Even if there were no interaction between the two fragments, the wave function of the A fragment could in principle make some use of basis functions centered on the B fragment, thus artificially lowering the total energy of the supermolecule. Clearly, the amount of BSSE is likely to be larger for smaller separations between the two fragments. The BSSE can be minimized through the use of very extensive basis sets, but this has the potential drawback of introducing problems of linear dependence for smaller separations at which basis functions on different centers may have very large overlaps.

For very weakly bound systems, the BSSE can swamp the true interaction energy. For more strongly bound systems, BSSE is still present but it is a relatively minor source of error. An approximate correction for BSSE is provided by practical schemes collectively known as function-counterpoise or ghost-orbital methods. The basic idea is to use the same basis set for the supermolecule and for the fragments. In essence, calculations are performed (for a given geometry) for the separate A and B fragments using the *full* basis set, and these energies are subtracted from that of the AB supermolecule, i.e.,

interaction energy

$$=E_{AB}(\mathbf{R})-E_A(\mathbf{R})-E_B(\mathbf{R}), \tag{177}$$

where all the calculations use the same supermolecule basis set.

9.3 Matrix Elements and Response Functions

Molecular electronic-structure methods that produce not only an energy but also a wave function are of particular value when molecular properties are required. A number of molecular properties can be expressed as expectation values of one-electron operators: e.g., electric multipole moments, spin hyperfine constants, and field gradients at the nuclei.

The molecular property most extensively calculated is the electric dipole moment function $\mu(\mathbf{R})$, which can be expressed as

$$\mu(\mathbf{R})=\left\langle\Psi_{\mathrm{el}}(\mathbf{r}|\mathbf{R})\right|\left(-\sum_i e\mathbf{r}_i+\sum_A Z_A e\mathbf{R}_A\right) \times\left|\Psi_{\mathrm{el}}(\mathbf{r}|\mathbf{R})\right\rangle. \tag{178}$$

Because the electronic wave functions depend parametrically on the nuclear coordinates $\mathbf{R}$, the expectation value of the electric dipole operator is a function of $\mathbf{R}$. The electric dipole moment governs the absorption and emission of photons for vibration–rotation transitions within an electronic state. When $\mu(\mathbf{R})$ is identically zero, as in the case of centrosymmetric molecules such as O_2 or N_2, radiation can be emitted or absorbed only if matrix elements of higher-order moments (such as quadrupole moments) are nonzero.

In order to calculate an electric dipole moment that can be compared to an experimentally determined value for, say, a diatomic molecule, $\mu(R)$ must be obtained over a range of R values and then averaged over the nuclear wave functions:

$$\mu_{vJ}=\int \chi_{vJ}^*(R)\mu(R)\chi_{vJ}(R)dR. \tag{179}$$

Dipole moments calculated using the Hartree-Fock SCF approximation are usually in error by $(0.04\text{–}0.02)ea_0$ for closed-shell systems $[(0.2\text{–}0.4)ea_0$ for open-shell systems$]$. Using CI methods, the errors can be reduced by more than an order of magnitude. Although single-excitation terms in a CI wave function do little to improve the energy, these terms are very important for accurate calculations of the dipole-moment function.

The matrix element governing transitions from one electronic state (i) to another (f) by absorption or emission of radiation is the electric-dipole transition moment:

$$D_{if}(\mathbf{R})=\left\langle \Psi_i(\mathbf{r}|\mathbf{R})\left|\left(-\sum_i e\mathbf{r}_i\right)\right|\Psi_f(\mathbf{r}|\mathbf{R})\right\rangle . \tag{180}$$

There are no contributions from terms involving a summation over the nuclei simply because Ψ_i and Ψ_f are orthogonal. Of course, the wave functions of states i and f must be available at the same set of nuclear coordinates. In CI calculations, the computations are greatly simplified if the two wave functions are constructed from a common set of molecular orbitals. The electric-dipole transition-moment operator is a one-electron operator, and so states that differ in orbital occupation by one electron generally have the largest matrix elements.

If wave functions of the nuclear motion are available, as in the case of diatomic molecules, then $D_{if}(\mathbf{R})$ can be used to obtain the line strengths or band oscillator strengths. Electronic-structure calculations are also very important for determining lifetimes of excited states and branching ratios.

Not all molecular properties of interest are obtained as expectation values of operators. The polarizability, for instance, can be thought of either as being related to the second derivative of the energy with respect to an applied external electric field or as the change induced in the dipole moment by such a field. Properties such as the polarizability are often termed response functions because they characterize the response of a molecular system to an external perturbation. Particular care is required in the choice of sufficiently flexible basis sets for calculations of electric and magnetic response functions. The development of gradient techniques (see Sec. 9.1) has been particularly important for the efficient calculation of such quantities.

Much progress is currently being made in the development of molecular electron-correlation methods that take proper account of relativistic effects. For systems containing heavy atoms, such as second-row transition metals, such effects are crucial to achieve spectroscopic accuracy. One important consequence of relativistic effects, which appears even in molecules containing only light atoms, is the occurrence of spin-dependent terms such as spin-orbit coupling. Rather than carry out fully relativistic calculations, it is usual to include spin-orbit coupling effects as a perturbation to nonrelativistic states. The dominant contribution to fine-structure intervals in molecules arises from the Breit-Pauli Hamiltonian

$$\begin{aligned} H_{\mathrm{SO}}=&(\alpha^2/2)\sum_{i,A}(Z_A/r_{Ai}^3)(\mathbf{r}_{Ai}\times\mathbf{p}_i)\cdot\mathbf{s}_i\\ &-(\alpha^2/2)\sum_{i,j}(1/r_{ij}^3)(\mathbf{r}_{ij}\times\mathbf{p}_i)\cdot(\mathbf{s}_i+2\mathbf{s}_j), \end{aligned} \tag{181}$$

in which α is the fine-structure constant, and $\mathbf{p}$ and $\mathbf{s}$ are linear momentum and spin angular momentum operators. This Hamiltonian corresponds to an approximation valid to order α^4mc^2 (i.e., α^2 in atomic units), but it is straightforward in molecular calculations also to include the term of order $(\alpha/\pi)\alpha^4mc^2$ that arises from anomalous magnetic moment effects. Using this Hamiltonian it is possible to calculate accurate molecular spin-orbit coupling constants and to study spectroscopic perturbations such as singlet–triplet mixing, Λ doubling, and spin-splitting constants. The lowest-order contributions that are neglected in the Breit-Pauli approximation are of order α^6mc^2. These additional terms cannot easily be included as a perturbation to nonrelativistic states, but they are already fairly significant for atoms beyond halfway across the second row of the periodic table. This is clearly an area in which fully relativistic calculations are potentially very important.

Finally, we consider briefly a breakdown of the Born-Oppenheimer approximation in which terms involving changes of the electronic wave function with nuclear coordinates cannot be neglected. At particular nuclear geometries, two potential curves or surfaces of the same symmetry may interact strongly and may exhibit an "avoided crossing" (see Fig. 28) in which they appear to exchange character with one another. As indicated in Sec. 7.1, the factorization of Ψ_{tot} into a *single* product of an electronic wave

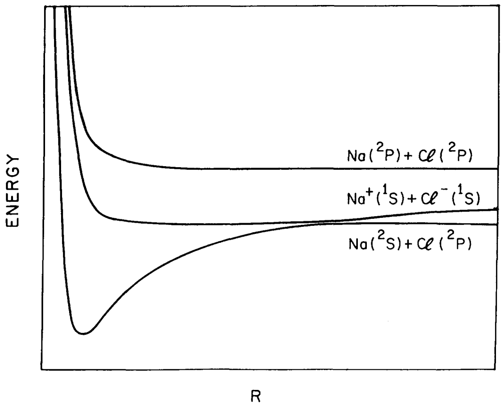

FIG. 28. Adiabatic potential-energy curves for the three lowest $^1\Sigma^+$ states of NaCl, illustrating an avoided crossing.

function and a nuclear motion wave function is dependent on potential-energy surfaces being well separated. This is clearly not the case in curve-crossing problems. In the case of diatomic systems, for example, this has the consequence that matrix elements such as $\langle\Psi_i(\mathbf{r}|R)|\partial/\partial R|\Psi_f(\mathbf{r}|R)\rangle$, arising from the nuclear kinetic energy operator, may be nonnegligible. The strong coupling of molecular states through $\partial/\partial R$ and related terms manifests itself experimentally through significant spectroscopic perturbations, leading to energy shifts, line broadening, and diffuseness of spectral lines. Such matrix elements also govern scattering processes such as charge transfer, in which an electron is transferred from one atomic center to another, e.g.,

$$O^{2+}+H\rightarrow O^{+}+H^{+}. \tag{182}$$

Matrix elements of $\partial/\partial R$ are usually calculated by numerical differentiation, although gradient techniques are to be preferred.

ACKNOWLEDGMENT

This work was partially supported by the National Science Foundation through a grant for the Institute for Theoretical Atomic and Molecular Physics at Harvard University and Smithsonian Astrophysical Observatory and by NSF Grant No. STI-9108764.

List of Abbreviations

BSSE: Basis-set superposition error.
CASSCF: Complete active-space self-consistent field.
CCSD: Coupled-cluster singles and doubles.
CI: Configuration interaction.
DZP: Double-ζ plus polarization.
FCI: Full configuration interaction.
GTO: Gaussian-type orbital.
GVB: Generalized valence bond.
HOMO: Highest occupied molecular orbital.
LCAO: Linear combination of atomic orbitals.
LUMO: Lowest unoccupied molecular orbital.
MBPT: Many-body perturbation theory.
MCSCF: Multiconfiguration self-consistent field.
MO: Molecular orbital.
MP*n*: Møller-Plesset to order *n* (starting from RHF).
MRCI: Multireference configuration interaction.
RHF: Restricted Hartree-Fock.
ROHF: Open-shell restricted Hartree-Fock.
SCF: Self-consistent field.
SCVB: Spin-coupled valence bond.
SDCI: Singles and doubles configuration interaction.
STO: Slater-type orbital.
TZVP: Triple-ζ valence plus polarization.
UHF: Unrestricted Hartree-Fock.
UMPn: Møller-Plesset to order *n* (starting from UHF).
VB: Valence bond.
VSEPR: Valence-shell electron pair repulsion.

List of Works Cited

Condon, E. U., Shortley, G. H. (1959), *The Theory of Atomic Spectra*, New York: Cambridge University Press.

Cooper, D. L., Gerratt, J., Ramondi, M. (1991), *Chem. Rev.* **91**, 929–964.

Cowan, R. D. (1981), *The Theory of Atomic Structure and Spectra*, Berkeley: University of California Press.

Cowan, R. D., Andrew, K. L. (1965), *J. Opt. Soc. Am.* **55**, 502.

Desclaux, J. P. (1972), *At. Data Nucl. Data Tables*, **12**, 311–406.

Davidson, E. R., Hagstrom, S. A., Chakravorty, S. J., Umar, V. M., Fischer, C. F. (1991), *Phys. Rev. A* **44**, 7071–7083.

Fischer, C. F. (1986), *Comp. Phys. Rep.* **3**, 273–326.

Fischer, C. F. (1988), *Nucl. Instrum Methods* **B31**, 265–272.

Hylleraas, E. A. (1964), *Adv. Quantum Chem.* **1**, 1–34.

Judd, B. R. (1963), *Operator Techniques in Atomic Spectroscopy*, New York: McGraw-Hill.

Kelly, H. P. (1969), *Adv. Chem. Phys.* **14**, 129–190.

Lindgren, I., Morrison, J. C. (1986), *Atomic Many-Body Theory*, 2nd ed., New York: Springer.

Nicholls, R. W., Stewart, A. L. (1962), in: D. R. Bates (Ed.), *Atomic and Molecular Processes*, New York: Academic.

Rose, M. E. (1961), *Relativistic Electron Theory*, New York: Wiley.

Stwalley, W. C., Zemke, W. (1992), *J. Phys. Chem. Ref. Data* (to be published).

Szabo, A., Ostlund, N. S. (1982), *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, New York: Macmillan.

Weiss, A. W. (1970), *Nucl. Instrum Methods* **90**, 121–131.

Weiss, A. W. (1973), *Adv. Atom. Molec. Phys.* **9**, 1–46.

Further Reading

Atkins, P. W. (1983), *Molecular Quantum Mechanics*, 2nd ed. Oxford: Oxford University Press. Includes a good introduction to the electronic structure of molecules.

Atkins, P. W. (1991), *Quanta: A Handbook of Concepts*, 2nd ed. Oxford: Oxford University Press. A very readable dictionary of important keywords and concepts.

Atkins, P. W., Child, M. S., Phillips, C. S. G. (1970), *Tables for Group Theory*, Oxford: Oxford University Press. Excellent reference for molecular symmetries.

Desclaux, J. P. (1982), in: R. Marrus (Ed.), *Atomic Physics of Highly Ionized Atoms*, New York: Plenum, pp. 75–112.

Desclaux, J. P. (1985), in: H. P. Kelly, Y.-K. Kim (Eds.), AIP Conference Proceedings No. 136, New York: AIP, pp. 162–175.

Froese Fischer, C. (1977), *The Hartree-Fock Method for Atoms*, New York: Wiley.

Grant, I. P. (1970), *Adv. Phys.* **19**, 747–811.

Grant, I. P., Quiney, H. M. (1988), *Adv. At. Mol. Phys.* **23**, 37–86.

Hartree, D. R. (1957), *The Calculation of Atomic Structures*, New York: Wiley.

Jucys, A. P. (1969), *Adv. Chem. Phys.* **14**, 191–213.

Kutzelnigg, W. (1989), *Z. Phys. D* **11**, 15–28.

Kutzelnigg, W. (1990a), *Z. Phys. D* **15**, 27–50.

Kutzelnigg, W. (1990b), in: Z. B. Maksic (Ed.), *The Concept of the Chemical Bond*, New York: Springer.

Lawley, K. P. (1987a), *Ab Initio Methods in Quantum Chemistry, Part I*, Advances in Chemical Physics, Vol. 67. Chichester: Wiley. Fairly up-to-date reviews of all the methods described in this article can be found in Parts I and II of this series.

Lawley, K. P. (1987b), *Ab Initio Methods in Quantum Chemistry, Part II*, Advances in Chemical Physics, Vol. 69. Chichester: Wiley.

Löwdin, P.-O. (1959), *Adv. Chem. Phys.* **2**, 207–322.

McWeeney, R. W. (1989), *Methods of Molecular Quantum Mechanics*, 2nd ed. London: Academic.

Michl, J. (Ed.) (1991), *Chem. Rev.* **91**, No. 5 (special issue), pp. 651–1108. Discussions of a wide range of molecular electronic structure methods and applications.

Roothaan, C. C. J. (1951), *Rev. Mod. Phys.* **23**, 69–97

Roothaan, C. C. J. (1960), *Rev. Mod. Phys.* **32**, 179–198

Roothaan, C. C. J., Weiss, A. W. (1960), *Rev. Mod. Phys.* **32**, 194–210

Schaefer, H. F., III, (Ed.), (1977), *Methods of Electronic Structure Theory*, Modern Theoretical Chemistry, Vol. 3, New York: Plenum.

Schaefer, H. F., III (Ed.) (1977), *Applications of Electronic Structure Theory*, Modern Theoretical Chemistry, Vol. 4. New York: Plenum.

Slater, John C. (1960), *Quantum Theory of Atomic Structure*, New York: McGraw-Hill, Vols. 1 and 2.

Sobelman, I. I. (1972), *Introduction to the Theory of Atomic Structure*, New York: Pergamon.

Szabo, A., Ostlund, N. S. (1982), *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, New York: Macmillan. Excellent presentation of methods with extensive examples.

Thorne, A. P. (1988), *Spectrophysics*, New York: Chapman and Hall.

Woodgate, G. K. (1980), *Elementary Atomic Structure*, Oxford: Clarendon.

ELECTRONIC STRUCTURE OF SURFACES

D. P. WOODRUFF, *Physics Department, University of Warwick, Coventry, United Kingdom*

Introduction 99
1. **Photoemission and Inverse Photoemission** 100
2. **Surface States at Clean Surfaces** 103
2.1 Shockley, Image-Potential, and Tamm States 103
2.2 Surface Core-Level Shifts 107
2.3 Dangling-Bond States and Scanning Tunneling Microscopy 109
2.4 Surface Reconstruction 112
2.5 Magnetic Properties of Surfaces 113
3. **Adsorbate-Induced Surface Electronic Structure** 114
3.1 Molecular Bonding and Valence Electronic Structure 115
3.2 Atomic Adsorbates and Associated Two-Dimensional Bands 116
3.3 Chemical Shifts in Core-Level Photoemission 118
Works Cited 119
Further Reading 119

INTRODUCTION

There are at least two (rather different) reasons why the electronic structure near the surface of a solid should differ from that of the underlying bulk. The first of these is entirely fundamental: a surface represents a special kind of defect in the ideal three-dimensional periodicity which characterizes a crystalline solid, and, in common with the situation for other defects, localized electronic states can occur (see ELECTRON STATES, LOCALIZED). This would be true even if the solid were ideally terminated (i.e., if all the atomic locations in the solid were unchanged by the presence of the surface), although, in general, the presence of the surface leads to a geometrical relaxation or reconstruction of the outermost atomic layers which can also modify the surface electronic structure. Of course, these geometrical and electronic structure changes are inexorably interlinked in that the stable geometrical structure is dictated by the minimization of the electronic energy. A second source of surface electronic structure is associated with compositional differences between the surface and the underlying bulk. In the case of nonelemental solids (compounds or alloys), these compositional changes may be associated with reestablishing thermodynamic equilibrium between the surface and the bulk, but another important area of interest concerns compositional differences associated with adsorbed species from the surrounding gas phase. These adsorbate-induced changes are of considerable interest because they are intimately related to the chemical reactivity of the surface, a feature which is highly relevant not only to heterogeneous catalysis and surface corrosion and passivation, but also to the nature of semiconductor interface formation and the resulting electronic properties of a fabricated device.

In discussing our present state of understanding of surface electronic structure it is necessary to have also some appreciation of the experimental methods of investigating this structure, and to understand what is actually measured. An exhaustive treatment of this topic is beyond the scope of this article and, in particular, we do not address in detail the problems associated with the fact that, as in all spectroscopies, the electronic structure cannot be investigated directly in the ground state; real measurements actually probe both the ground and some excited states in a technique-specific fashion. This article does, however, incorporate a presentation of some of the more important investigative techniques with an emphasis on their underlying physical

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

principles. These techniques are introduced as necessary to appreciate the examples of the results of experiments which support various aspects of our current understanding of surface electronic structure.

By far the most important and generally applicable experimental methods are those based on photoemission and inverse photoemission, and these techniques are described in Sec. 1. This is followed in Sec. 2 by a presentation of various aspects of the electronic structure of clean surfaces, beginning with the basic theory of surface states, particularly on metal surfaces. One of the most interesting aspects of another experimental technique, scanning tunneling microscopy, is described in the particular context of semiconductor "dangling bond" surface states. Section 2 concludes with a brief discussion of two other aspects of the electronic structure of clean surfaces, namely, their tendency in many cases to undergo reconstruction and their magnetic properties. In Sec. 3 we then discuss the electronic structure associated with adsorbed species at surfaces with particular regard to the information to be gained from electronic spectroscopy of molecular adsorbates, the nature of two-dimensional band structures formed mainly by atomic adsorbates, and the information to be gained from core-level photoemission studies.

1. PHOTOEMISSION AND INVERSE PHOTOEMISSION

The basic idea of photoemission is dictated by the demands of energy conservation as stated by Einstein in 1905, who explained the photoelectric effect by considering that the process involved a single photon giving up all its energy $h\nu$ to a bound electron in a solid. Thus, if the electron has an associated binding energy E_B, then the electron can escape with a kinetic energy given by

$$E_{\text{kin}} = h\nu - E_B. \qquad (1)$$

Evidently, if the kinetic energy of the emitted electron is measured and the photon energy is known, we have a means of determining the electronic binding energy. This process is illustrated schematically in the PES (photoelectron spectroscopy) panel of Fig. 1. Peaks in the photoelectron energy spectrum correspond directly to occupied bound states, although the actual experimental spectrum also has a rising background due to the effects of inelastic scattering. Notice that if we refer the binding and kinetic energies of Eq. (1) to the Fermi level of the solid, as is usual, there is a threshold kinetic energy needed (equal to the work function ϕ —the difference between the Fermi and vacuum levels) before photoemission can be observed in the vacuum. Typically, photoemission studies of surfaces in the laboratory use fixed photon energies in the vacuum ultraviolet (usually 21.2- or 40.8-eV lines from a He discharge) or in the soft x-ray energy range (1253.6 or 1486.6 eV from Mg or Al $K\alpha$ sources), leading to the techniques of ultraviolet and x-ray photoelectron spectroscopy (UPS and XPS). Clearly the lower energies used in UPS only give access to valence electronic states, while XPS is more extensively used for the study of core-level photoemission. Increasing use is now made of synchrotron radiation, however, which allows the use of continually variable energies throughout this range, a feature that offers important advantages for electronic structure investigations, as will be shown below, but softens the distinction between UPS and XPS defined by the two types of laboratory source.

Photoemission provides a means of probing the occupied electronic states of a solid; strictly, it also provides some information on

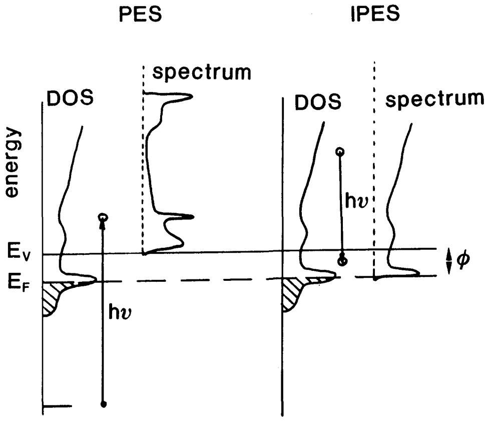

FIG. 1. Schematic diagram showing the basic principle of operation of photoelectron spectroscopy (PES) and inverse photoemission spectroscopy (IPES). Superimposed on the energy-level diagram, on which are defined the Fermi and vacuum levels, E_F and E_V, are shown the electron density of states (DOS) and the resulting PES and IPES spectra.

unoccupied states because the electron is emitted into such an unoccupied state and the transition probability must be influenced by both the initial and final states. Often, however, one is able to assume that the emitted electron has a sufficiently high kinetic energy that it can be well described by a simple free-electron state; even if this is not the case, the most interesting unoccupied states, which lie just above the Fermi level, cannot be accessed by photoemission because they lie below the vacuum level and an electron photoexcited into them is unable to escape from the solid. Complementary information on these states, however, is available from the technique of inverse photoemission, also shown schematically in the IPES (inverse photoemission spectroscopy) panel of Fig. 1.

In inverse photoemission an electron is directed at the solid and has a certain probability of losing energy by the emission of a photon, falling into a lower-energy unoccupied energy level. In this case the final electron state can lie between the Fermi and vacuum levels as well as in the continuum of higher-energy states. Although inverse photoemission is a less familiar technique than photoemission, its history actually dates from a similar period, as it is simply another name for the bremsstrahlung emission from an electron-bombarded solid which is well known in laboratory x-ray sources. The application of this method in the ultraviolet energy range is relatively new, however. In principle this technique can be exploited in either of two forms. In one, a fixed-energy incident electron beam is used and the energy spectrum of emitted photons is measured. More commonly, however, a fixed photon-energy detector is used and the incident initial-state electron energy (and thus the final-state energy *hν* below) is scanned. In this isochromat mode (also known as bremsstrahlung isochromat spectroscopy) a particularly simple version of the experiment uses a detector illustrated in Fig. 2 which comprises an iodine-filled Geiger–Müller tube fitted with a CaF_2 window; the combination of the low-energy-pass filter character of the window and the high-pass threshold ionization behavior of the iodine combine to produce a band-pass (isochromat) detector with a central energy of 9.7 eV.

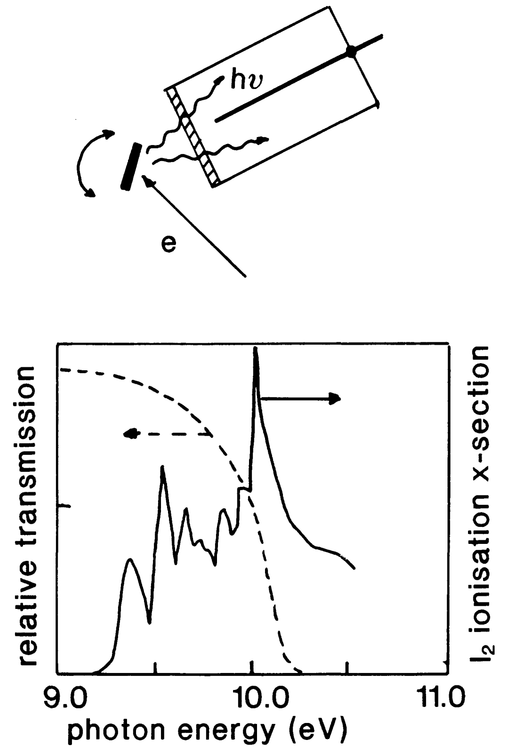

FIG. 2. Schematic diagram of the iodine-filled Geiger-Müller tube fitted with a CaF_2 window for use as a 9.7-eV isochromat detector in inverse photoemission (upper diagram), and of the low-pass window transmission and high-pass ionization cross-section threshold behavior (lower diagram) which form the basis of its operation.

Although the basis of both the photoemission and inverse photoemission experiments is energy conservation, their full power in the investigation of the electronic structure of crystalline solids and surfaces arises from the additional requirements of momentum conservation. Because a photon carries very little momentum, all optical transitions in solids are dominated by direct or vertical transitions in which the reduced electron momentum (i.e., the electron momentum within the first Brillouin zone in the reduced zone scheme) is conserved. In the case of photoemission, however, the final-state electron is detected outside the solid in the vacuum, and if both its energy and direction are measured, the final momentum is known. Energy and momentum conservation therefore mean that the initial-state energy and momentum are also known, so that it is possible to map an experimental energy-momentum (E-$\mathbf{k}$) band structure for the solid. This is the basis of angle-resolved (ultraviolet) photoelectron spectroscopy (ARUPS). In the same way, the incident electron energy and direction (and hence momentum) can be defined in inverse photoemission, allowing the unoccupied bands to be mapped by **k**-resolved inverse photoemission spectroscopy (KRIPES).

This potential is illustrated in Fig. 3, which shows the band structure of Cu along the Γ-*X* direction upon which are superimposed possible *s-p*–band direct photoemission and inverse photoemission transitions, together with the experimental spectra corresponding to these conditions, collected in normal emission or incidence at a Cu(100) surface. The arrowed peaks in the experimental spectra correspond to the *s-p*–band transitions shown on the band structure; with the photoelectron kinetic energy axis of these graphs (the abscissa) converted to a binding energy scale as shown, these arrows track the dispersion of the lower *s-p*–band through the Fermi level. This figure not only demonstrates the basic power of the techniques to map occupied and unoccupied energy bands in the solid, but also illustrates that variable photon energies are needed to achieve this along any particular direction of three-dimensional **k** space (such as the Γ-*X* symmetry direction). Strictly, however, our discussion is not exactly correct because we have assumed that the reduced electron momentum of the bulk solid is conserved. In reality, photoemission and inverse photoemission both involve the passage of relatively low-energy (tens of eV) electrons through a solid, and the mean free path for this process before inelastic scattering occurs is very short (0.5–2 nm). This means that the techniques only sample the outermost few atomic layers of the solid and therefore do not truly sample the translational symmetry (the periodicity) of the solid perpendicular to the surface. As the conservation of the reduced momentum is a direct consequence of the periodicity, this means that only the component of the electron momentum parallel to the surface, $\mathbf{k}_{\parallel}$, is strictly conserved in ARUPS and KRIPES. However, it is clear from the success of the example shown in Fig. 3 that the electron samples sufficient depth for a strong remnant of full **k** conservation to be maintained; in particular, the experimental spectra are all recorded at $\mathbf{k}_{\parallel}=0$ and the **k** varied in the Γ-*X* direction is the component *perpendicular* to the surface.

In investigations of true surface electronic structure, however, this formal absence of conservation of the perpendicular component of the momentum is not a problem. Electronic states which are localized at the surface can have no dispersion in the momentum component perpendicular to the surface, and so only the parallel component and its conservation are important. In this case the ability to vary

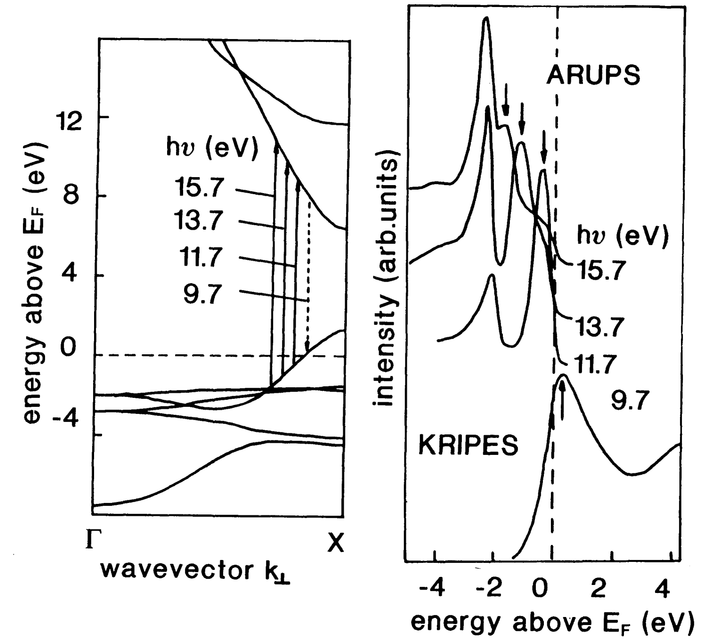

FIG. 3. Possible *s-p*–band direct transitions in photoemission and inverse photoemission for normal electron emission or incidence on Cu(100) overlaid on the appropriate part of the band structure (left), compared with the experimental spectra taken under these conditions (right). The arrowed peaks in the experimental spectra correspond to the *s-p*–band transitions indicated on the band structure. Note that the abscissa of the experimental photoemission spectra is strictly the photoelectron kinetic energy, but this is converted to the binding energy of the initial state being probed by subtracting the photon energy. The KRIPES spectrum axis is similarly converted to the final-state binding energy. After Woodruff *et al.* (1983).

the photon energy is also no longer so important because different $\mathbf{k}_{\parallel}$ can be accessed with a single photon energy simply by varying the electron emission or incidence direction. On the other hand, the absence of dispersion in the perpendicular component of the momentum for surface-localized states is a convenient way of distinguishing them from bulk states and can only be explored with variable photon energies. A more detailed description of angle-resolved photoemission and **k**-resolved inverse photoemission may be found in Kevan (1992) and Smith and Woodruff (1986), respectively.

2. SURFACE STATES AT CLEAN SURFACES

2.1 Shockley, Image-Potential, and Tamm States

At a clean surface we can identify two essentially different types of intrinsic surface state. One of these types (which includes both the so-called Shockley and image-potential states) can occur even for nearly-free-electron solids and is a direct consequence of including the effects of the crystal truncation in an otherwise perfectly periodic system in three dimensions. A rather different type of surface state is associated with more localized electronic states of the constituent atoms and arises because of the different local electronic environment of the surface atoms; these are generally known as Tamm states, and the most obvious example is the so-called dangling-bond states which occur at semiconductor surfaces. A general feature of any electronic state which is localized at the surface is that it must be unable to couple to a bulk state with the same energy and parallel component of momentum, and must therefore lie in a projected bulk band gap. (Notice that this need not be a complete bulk gap; we require only that no allowed states with the same parallel component of the wave vector exist). Its associated wave function therefore decays in an evanescent fashion into the solid (and must also decay exponentially with distance outside the surface at large distances if the energy lies below the vacuum level). In addition, however, some electronic states lying within an allowed band of the underlying solid have an enhanced electron density in the surface region relative to the bulk state to which they couple; these states are known as surface resonances.

Surface states which arise only from the termination of the solid at the surface and which are not related to semi-localized bulk states are known as Shockley states. Their possible existence and character can be derived from a simple nearly-free-electron model of the band structure of a solid by recognizing that mathematically valid solutions to Schrödinger's equation for an electron in a periodic potential do exist in the band gaps, but they have complex values of electron momentum **k** (Fig. 4). For a bulk solid these are not physically meaningful because the imaginary part of **k** leads to the wave function having a real exponential component which must become infinite in some region of space. If a surface is present, on the other hand, these solutions can correspond to states which decay into the solid but do not apply to the half-space corresponding to the vacuum. In the vacuum they must decay exponentially, and the two solutions must match in wave-function amplitude and gradient at the solid/vacuum interface. Since the phase of the evanescent wave solu-

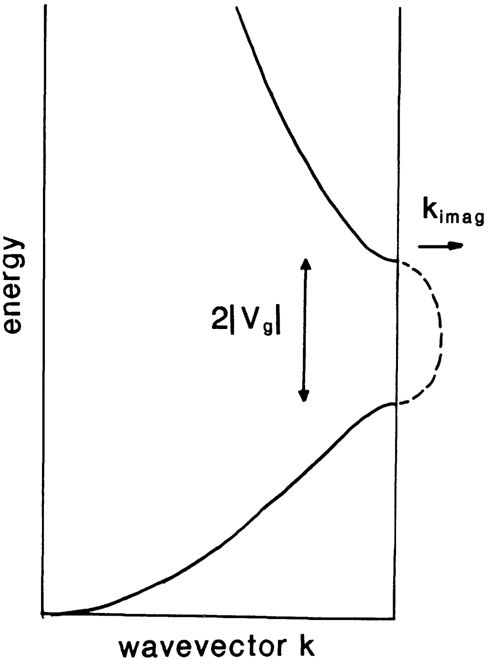

FIG. 4. Schematic diagram of a nearly-free-electron band structure showing the complex *k* solutions in the band gap.

tion at this interface changes in a systematic fashion as the energy is varied within the band gap, this matching (notably in gradient) is only possible at one particular energy in the gap, defining the Shockley-state energy. Indeed, this solution only exists for the appropriate polarity of the ion core pseudopotential amplitude, V_g (which has a magnitude of half the energy range of the band gap), and so surface states of this kind do not occur in all projected bulk band gaps.

An alternative but a particularly simple formalism which provides a description of both the Shockley and image states is as follows. Consider an electron just outside the crystal surface, at an energy below the vacuum energy and within an energy gap of the crystal. Under these conditions, the electron finds itself in a potential bounded by two rather different barriers. In particular, it is unable to penetrate the crystal because it lies in the band gap, yet it is also unable to escape into the vacuum because it is trapped by the so-called image-potential force. For a metal, the interaction of the electron with the screening charge of the substrate can be described by a Coulombic interaction with an "image" charge (of opposite sign to the electron, thus producing an attraction), which appears in the metal surface at a distance below the surface equal to the distance of the electron from the surface. For electron energies below the vacuum level, bound states can therefore be found within this potential well, defined by the repulsive solid potential and the attractive image potential.

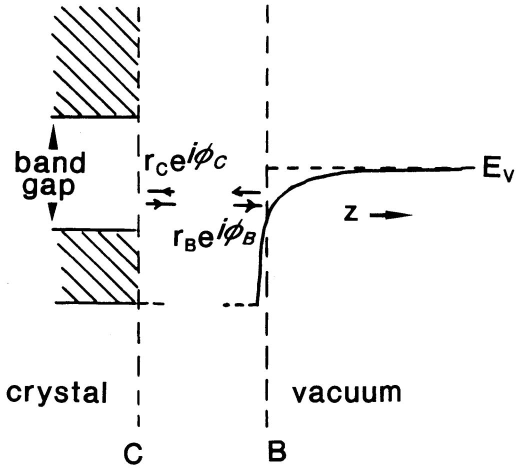

FIG. 5. Schematic diagram showing the one-dimensional potential well for an electron outside the surface of a conducting solid at a band-gap energy. The reflection amplitudes at the two barriers are shown as described in the text.

Figure 5 shows this electron potential schematically, while Fig. 6 shows the charge-density distribution associated with some of the solutions. The picture is quite similar to a one-dimensional analogue of the Coulombic potential of atomic hydrogen, and the solution involves a similar infinite series of possible solutions which converge on the vacuum level, exactly like the Rydberg states of the hydrogen atom. One way of obtaining the solutions is to consider an electron which bounces back and forth off the two walls within this potential and identifying the conditions under which the round trip involves a total phase shift which is an integral multiple (n) of 2π. This condition will generate the standing-wave solutions characteristic of a bound state. Thus, if the electron reflectances at the crystal and image potential barriers are denoted by $r_C \exp(i\phi_C)$ and $r_B \exp(i\phi_B)$, then the condition for the bound states is

$$\phi=\phi_C+2k_\perp(z_B-z_C)+\phi_B=2\pi n, \tag{2}$$

where $k_\perp$ is the component of the electron wave vector perpendicular to the solid surface, and z_C-z_B is the distance between the two barriers (typically taken to be half the atomic-layer spacing in the solid). The phase shift at the crystal barrier is determined by the energy within the gap (and hence by the size of the gap, $2V_g$) but varies by π across the gap,

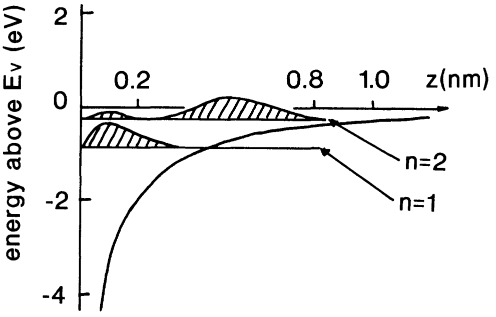

FIG. 6. Plot of the image-potential barrier and of the charge density associated with the $n=1$ and 2 bound states (Whittaker functions). The solutions are shown for the case of a quantum-defect parameter [Eq. (5)] $a=0$, which corresponds to an infinitely repulsive crystal boundary.

while the phase shift at the image barrier depends on the energy relative to the vacuum level.

Solving this problem (see, for example, Smith and Woodruff, 1986) leads to two rather different types of bound states. For $n=0$, a solution is obtained which is dominated by the crystal barrier phase shift, and corresponds to the Shockley surface state with its charge density peaked just inside the crystal surface. The higher n-value solutions, most strongly influenced by the image potential barrier, occur at energies given by

$$E=E_V-e_n+\hbar^2k_\parallel^2/2m, \tag{3}$$

$$e_n=(0.85\ \text{eV})/(n+a)^2, \quad n=1, 2, \ldots \tag{4}$$

with E_V the energy of the vacuum level. Evidently Eq. (4) defines a Rydberg-type set of states which disperse in a free-electron fashion in $k_\parallel$, but the exact energy at which these states occur is influenced by the so-called quantum-defect parameter,

$$a=\tfrac{1}{2}(1-\phi_C/\pi). \tag{5}$$

The $n=1$ and 2 solutions for $a=0$ (corresponding to an infinitely repulsive potential) are illustrated in Fig. 6.

The actual energies can be obtained from a graphical solution such as that shown in Fig. 7 for the case of Cu(111); the diagram shows the $n=0$ and 1 solutions. The $n=0$ Shockley state is predicted to lie 0.5 eV below the Fermi level (and so should be occupied at $k_\parallel=0$), while the $n=1$ lowest-energy image-potential state is 0.9 eV below the vacuum level. Both of these predictions are in rather good agreement with experimental results as shown in the KRIPES and ARUPS results of Figs. 8 and 9, summarized in the experimental band structure of Fig. 10. Notice that the occupied Shockley state is seen in the normal-emission high-resolution ARUPS data, but this state disperses up in energy as the emission angle (and thus $k_\parallel$) increases, and it crosses the Fermi level to appear in the unoccupied-state KRIPES spectra at higher incidence angles. The $n=1$ image-potential state is clearly visible in the KRIPES spectra, although the experimental resolution is too poor to resolve the higher Rydberg

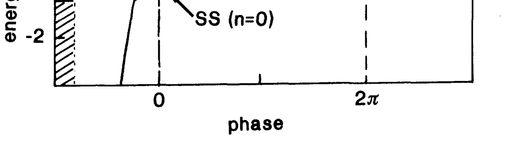

FIG. 7. Graphical solution of the multiple-reflection model of surface states for Cu(111) showing the $n=0$ (SS, Shockley state) and $n=1$ (IS, first image-potential state) solutions. The figure shows the crystal and image-potential barrier individual contributions as well as the total phase. After Hulbert *et al.* (1985).

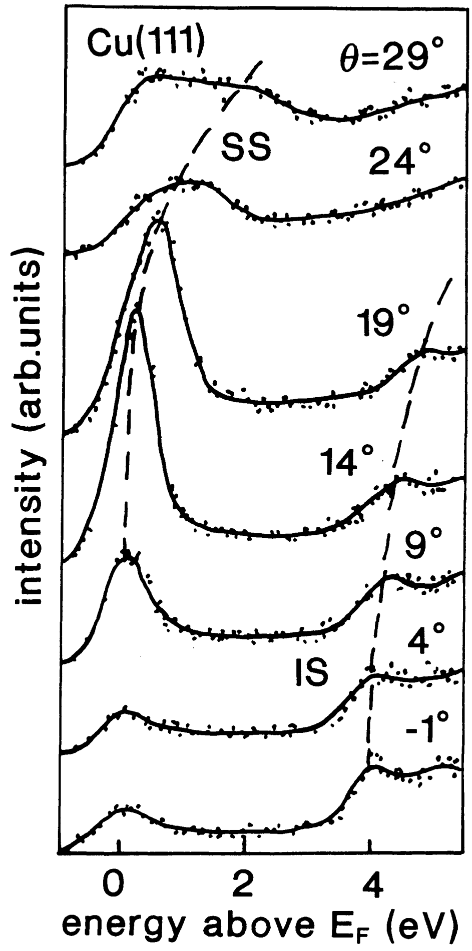

FIG. 8. KRIPES data taken from Cu(111) at a photon energy of 11.0 eV for different values of the angle of incidence Θ. The main features are labelled as IS (the $n=1$ image state) and SS (the Shockley state). After Hulbert *et al.* (1985).

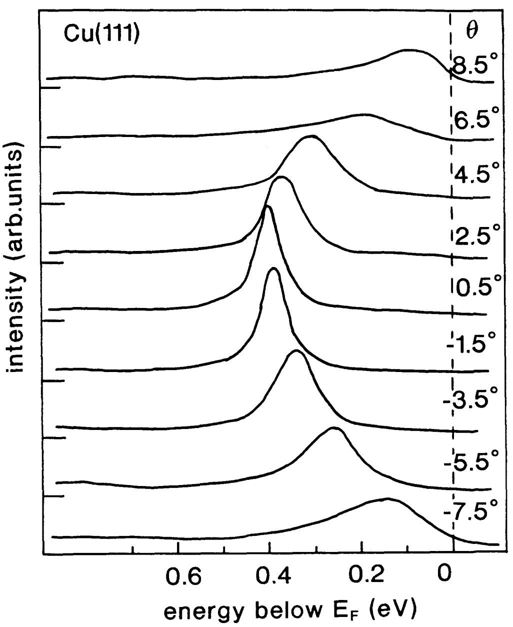

FIG. 9. High-resolution ARUPS of the Shockley state near the Brillouin zone center on Cu(111) recorded with a photon energy of 11.8 eV at different angles of emission, Θ. After Kevan (1983).

states which merge into an underlying edge. More of these higher states have been identified in separate higher-resolution two-photon photoemission experiments (in which the image-potential states are the intermediate bound states in the two-photon excitation from the occupied valence band to the vacuum continuum). Figure 10 summarizes the results of these experiments in the form of an experimental two-dimensional band structure onto which is superimposed the projection of the bulk occupied bands. Notice that both the Shockley and image-potential states which have been mapped lie in (or very close to the edge of) the bulk projected band gap.

Cu(111) also provides an example of a Tamm state which can be found at substantially larger values of $k_{\|}$. In addition to the rather free-electron-like *s-p* bands, the Cu valence bands contain the 3*d* electrons which, being far more strongly localized on the individual atoms, form very flat (weakly dispersing) fully occupied bands about 3 eV below the Fermi level (as may be seen in Fig. 3). Calculations of these *d* bands clearly require a more sophisticated treatment than the free-electron model we have discussed above, although they can be understood with a tight-binding model.

The usual way to calculate these bands is to perform a slab calculation in which the electronic structure of a small number of two-dimensionally periodic layers of the solid are stacked together. This allows one to use conventional three-dimensional band-structure methods to calculate the two-dimensional band structure of each layer, although the coupling between the layers must be dealt with differently. The slab has two surfaces, and these calculations are capable of providing a good description of the two-dimensional surface band structure, although the $k_{\perp}$ dispersion of the bulk bands is obviously not reproduced. The results of a calculation of this type are shown in Fig. 11 in a similar format to that of Fig. 10, but extended to lower energies. The dashed lines show the two-dimensional bands which are localized at the surface, and the upper such band is clearly the Shockley state which we have discussed previously. In addition, however, a rather flat band is seen some 2 eV below the Fermi level; for most values of $k_{\|}$, this surface band overlaps the bulk bands and so forms a surface resonance, but near the edge of the Brillouin zone it appears in a bulk band gap and becomes a true surface state. This band is associated with *d*-electron states in the surface atoms which experience a different local potential from those of the underlying bulk atoms and so are displaced in energy from the bulk *d* bands, a description which characterizes the Tamm surface states. These occupied surface states have also been observed (at energies a few tenths of an eV

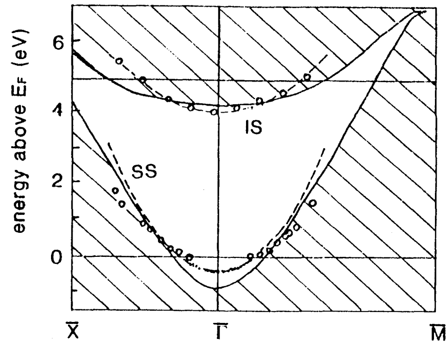

FIG. 10. Summary of the experimental surface band structure of Cu(111) based on the KRIPES (open circles) and ARUPS (small solid dots) results of Figs. 8 and 9. The hatched region is the projection of the bulk bands of copper.

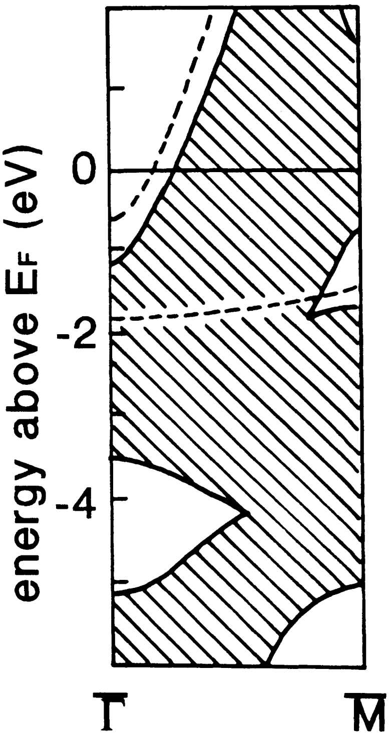

FIG. 11. Theoretical surface band structure for Cu(111) based on a slab calculation by Euceda *et al.* (1983) (after Zangwill, 1988).

below the calculated values) in ARUPS experiments performed at the appropriate higher values of $k_{\|}$.

Although this distinction between Shockley and Tamm states is rather clear in the case of Cu and quite a number of related noble and transition metals, this is not always true and evidently depends on the detailed hybridization schemes which determine the detailed electronic band structure. Nevertheless, this example serves to illustrate the main physical ideas involved in a particularly clear fashion.

We should also note here that in this section we have concentrated on a description of metal surfaces. Clearly covalently bonded semiconductors do not lend themselves to being described by a nearly-free-electron model, and a more appropriate simplification is a tight-binding approach; this highlights the fact that the electrons are far more localized in semiconductors so that we would expect mainly Tamm states. This subject is dealt with separately in Sec. 2.3.

2.2 Surface Core-Level Shifts

In the foregoing discussion we have concentrated on the use of photoemission, and particularly angle-resolved photoemission, as a direct probe of the valence electronic structure. Using higher photon energies to excite core-level photoemission, as typified by the XPS technique, provides primary information on the elemental composition of the surface region because the core-level binding energies provide a characteristic spectral signature of the emitting atoms. Small changes in these measured binding energies, however, do occur as a result of changes in the local valence electronic environment of the surface atoms, and indeed this sensitivity to the local "chemical" bonding state, and the associated "chemical" shifts in the measured binding energies, is one of the primary uses of the XPS technique in studies of surface chemistry. (Indeed, the technique was often, and is still sometimes, referred to as ESCA, or electron spectroscopy for chemical analysis.) This application of XPS is discussed further in Sec. 3.3 in regard to investigations of surface electronic structure associated with adsorption, but of interest in the context of clean surfaces is the extent to which the different electronic structure of the surface relative to the bulk is manifested in a change in the measured photoelectron binding energy even for an elemental solid. Indeed, one could view this effect as a rather special kind of Tamm surface state in that it is a binding-energy change in a localized state at the surface.

Before discussing the results of these surface core-level shift experiments, however, it is important to appreciate just what is measured in the XPS technique. In particular, photoemission from a highly localized core-level state of an atom leaves behind an ion with a localized core hole, and the time scale within which the surrounding valence electrons can respond to the appearance of this hole is sufficiently short that their effect must be considered in understanding the measured binding energy. Specifically, the loss of the core electron means that the nuclear screening effect of this electron is lost, and the remaining bound electrons will "relax" to lower binding-energy states; the effect of this relaxation is to make additional energy available to the escaping photoelectron which is measured to have a higher kinetic energy than would be expected from a simple "frozen orbital" emission process; the measured binding energy thus has a smaller magnitude than would be expected from a simple one-electron picture. This effect is important, not only because it modifies the absolute value of the measured binding energy, but because different valence electronic environments will lead to different degrees of extra-atomic relaxation, so that one component of any measured chemical or surface core-level binding-energy shift in XPS is associated with changes in the final-state ion as well as the initial-state neutral (ground-state) atom. Indeed, because removal of a core electron from an atom increases the effective nuclear charge "seen" by the valence electrons by one electronic charge, the valence electronic states in the final state can be likened to those of the next element in the periodic table (i.e., the atom having an atomic number one unit higher). As this atom may have a radically different bonding character (e.g., the difference between a halogen and an inert gas, or between an inert gas and an alkali), it is clear that this final-state effect can be substantial.

Returning now to the original question of the observability of surface core-level binding-energy shifts in photoemission, Fig. 12 shows examples of this effect in the form of W and Ta $4f_{7/2}$ photoemission energy spectra from W(111) and Ta(111) surfaces. In each case there are two clearly visible components to the spectrum, while more careful fitting of these curves reveals that three separate peaks, labeled S_1, S_2, and B, contribute to each spectrum. By applying various additional tests, such as studying the influence on the spectra of adsorbed species, these components are found to be associated with emission from top-layer atoms, second-layer atoms, and underlying bulk atoms, respectively. Note that these spectra, recorded at rather high resolution, were measured using quite low–photon-energy synchrotron radiation (70 and 66 eV, respectively) in order to maximize the surface sensitivity (by taking advantage of the lower inelastic-scattering mean free path at low kinetic energies) and to optimize energy resolution. These spectra illustrate the fact that in core-level photoemission from states which show sufficiently narrow intrinsic widths (such as these 4*f* states), the different electronic environment of surface atoms relative to bulk atoms can be distinguished, and indeed in this case even different shifts for differently coordinated near-surface atoms can be observed. One additional and rather striking feature of these data is that the surface core-level shifts of W and Ta are of opposite polarity.

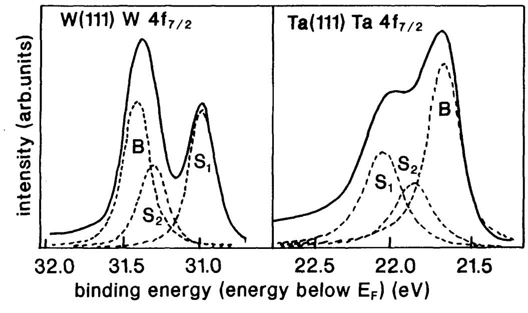

FIG. 12. W and Ta $4f_{7/2}$ photoemission spectra from clean W(111) and Ta(111) surfaces, also showing the component parts determined by peak fitting. After Eastman *et al.* (1982).

A rather simple theory of the effect of the different surface electronic environment on the one-electron binding energy (i.e., of only initial-state effects as in conventional Tamm valence states) can provide a qualitative description of this effect. It is based on the fact that the lower coordination of the surface atoms will lead to a narrowing of the associated 5*d* conduction band. If this band is other than exactly half-filled then a small amount of charge must flow into or out of the surface in order to balance the Fermi levels in the bulk and at the surface, and the resulting change in local electrostatic potential causes a change in the associated core-level binding energies of the surface atoms. This effect is illustrated schematically in Fig. 13 which shows that for a *d* band which is less than half-filled, the binding energy should increase, while for a band which is more than half-full the binding energy should decrease. The trend matches the experimental data in that W has one more (4) *d*-band electrons than Ta, so that the shift should be less positive, but the change in sign should actually occur one element to the right in the periodic table (at Re which has five 4*d* electrons) and not between Ta and W. One reason for this quantitative discrepancy is that the *d* bands are not symmetrical as assumed in the schematic of Fig. 13, and this asymmetry will change the *d*-band occupation required for the sign of the charge transfer to switch over.

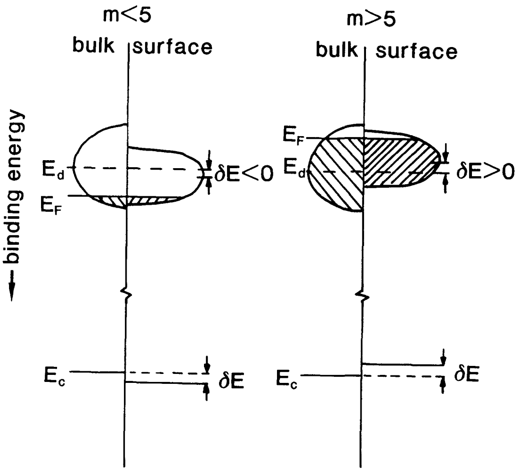

FIG. 13. Band-narrowing model of surface core-level shifts for a *d*-band metal, and an electronic occupation *m* either greater or less than 5 (half-filled). E_d is the energy of the center of the *d* band, E_c is the photoelectron binding energy of a core level of atoms in the bulk. δE is the displacement of the surface and bulk *d*-band centers and thus also the core-level shift. After Eastman *et al.* (1982).

Although this simple account of the initial-state energy change is qualitatively consistent with the results, another equally simple but more complete phenomenological theory highlights the fact that changes in the final-state energy are also likely to contribute to the observed results. In this picture we note that the key difference between photoemission of a core electron from a surface or a bulk metal atom is that in one case the final state has a core hole on a bulk atom, whereas in the other this hole is on a surface atom. We have already noted that the (relaxed) valence-electron environment around the core hole of the metal atom of atomic number Z is essentially equivalent to that around a neutral atom having an atomic number $Z+1$. The difference in energy is therefore that involved in moving a $Z+1$ atom from the bulk to the surface; i.e., the core-level shift is equal to the heat of segregation of a $Z+1$ atom in a Z-atom solid matrix.

2.3 Dangling-Bond States and Scanning Tunneling Microscopy

In introducing the idea of Tamm states one example that was mentioned was that of so-called dangling-bond states at semiconductor surfaces. Semiconductors comprise covalently bonded networks in which the valence electrons are rather highly localized in space between the atoms, in marked contrast to the highly dispersed electron density distribution of a simple free-electron-like metal. One consequence of creating a surface in such a structure is that localized electron orbitals appear at the surface which are unmatched by the atoms which were removed to create the surface. Clearly, these localized electron states have a very different electronic environment from that of a bulk atom and are likely to appear in the band gap. Another aspect to this picture makes this anticipated energetic location rather clear. If we consider, in particular, an elemental semiconductor such as Si, the bonding of two Si atoms causes the energy level of the bonding orbital (the sp^3 hybrid) to split into a bonding and antibonding orbital at higher and lower energies, respectively, and extending the Si—Si bonding to a solid network of Si atoms broadens these bonding and antibonding orbitals into the valence and conduction bands separated by the band gap, with the lower band just filled. If we now remove a Si atom from this network we can expect the energy of the unbound orbital to revert to an energy closer to its original atomic value, midway between the bonding and antibonding bands, and thus in the band gap.

Clearly a more rigorous approach is needed to establish the exact form and energy of the associated surface band, and in particular whether it should be occupied or unoccupied. Moreover, a very important complication in such a calculation is to take proper account of the reconstruction which such surfaces adopt in order to reduce the increased energy associated with these unsatisfied bonds. In the simplest case, such as the Si(100) surface, the main reconstruction is simply a "pairing" of adjacent atoms [which each have two dangling bonds on an ideally terminated Si(100) surface] with these atoms moving closer together; this movement of top-layer atoms causes distortions in several lower atomic layers because of the sensitivity of the total bonding energy to the nearest-neighbor bond lengths and relative bond angles. The atom-row pairing in the surface leads to a reconstructed surface having a surface unit mesh with dimensions equal to that of the substrate in one direction parallel to the surface, but twice as large in the direction of the other surface primitive translation vector—a so-called (2×1) structure. In the case of Si(111), the stable surface reconstruction involves atom pairing, top-layer stacking faults, and adatoms in a massive (7×7) reconstructed surface mesh, the structure of which has taken many years to solve. Nevertheless, the key conclusion concerning the surface electronic structure is that surface states are still expected to occur on this surface, and that they should be highly localized at specific surface atoms. The basic correctness of the existence of such surface states, both occupied and unoccupied, has been established using photoemission and inverse photoemission. A far more interesting experimental result in this context, however, is direct observation of the spatial localization of these surface states on the associated surface atoms.

This demonstration has been achieved using the technique of scanning tunneling microscopy (STM) (*q.v.*; see also FIELD EMISSION AND IONIZATION). The basic principle of the scanning tunneling microscope is illustrated in Fig. 14. If a fine metal tip is brought sufficiently close to a conducting surface, electrons can tunnel across the vacuum gap between them under the influence of a small voltage difference V. Clearly this quantum-mechanical tunneling can only occur when the separation is of atomic dimensions (less than about 1 nm), but under these circumstances electrons can tunnel either from occupied states in the tip to unoccupied states in the surface, or from occupied states in the surface to unoccupied states in the tip, depending on the polarity of the applied voltage. In the usual microscopic mode of the technique the tip is mounted on piezoelectric rods or a tube which allows the tip to be moved parallel and perpendicular to the surface, and as the tip is scanned parallel to the surface a feedback loop moves the tip in and out as required to maintain a constant tunneling current I_0. The voltage applied to the piezoelectric mounting to move the tip in and out can then be converted into a distance moved, so that a constant–tunneling-current contour map of the surface can be constructed. Because the quantum-mechanical tunneling current is exponentially dependent on the gap size, this technique is extremely sensitive, and the contour map can have a perpendicular spatial resolution of less than 10 pm, with a lateral resolution also well below atomic dimensions.

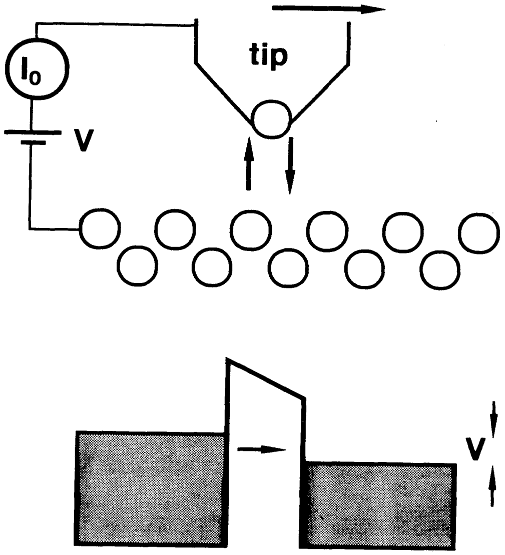

FIG. 14. Schematic of the principle of operation of a scanning tunneling microscope scanned at a bias potential difference V and a fixed tunneling current I_0.

Microscopies are generally regarded as probes of geometrical rather than electronic structure, but in the case of STM the nature of the probe means that it is extremely sensitive to the electronic structure; indeed, an approximate description of the contours actually followed in the microscope is that they correspond to constant valence-electron density contours at a rather low value of this density. The fact that the image typically appears as if it is an atomic picture of the surface is therefore potentially misleading in that there is no general theorem which connects peaks in these low-density charge contours outside the surface with underlying atomic positions (although such a correlation probably does exist for many surfaces). In this regard, the amplitude of the corrugations perpendicular to the surface in a typical STM image is found to be much larger at a semiconductor surface than at a metal surface, an effect which can be attributed to differences in the degree of localization of the valence charge in these materials. Specifically, however, in the situation in which we have dangling-bond states localized at particular atoms on the surface, we may expect that STM offers a means of spatially mapping these electronic states (with atomic resolution). This is, indeed, found to be the case, and a particularly clear example is provided by STM of the Si(111) surface [showing the (7×7) reconstruction]. Because the electronic states sampled by the tunneling experiment depend on the applied voltage difference *V*, and particularly on the polarity of this voltage, the STM images from this surface depend strongly on this parameter, as may be seen in Fig. 15.

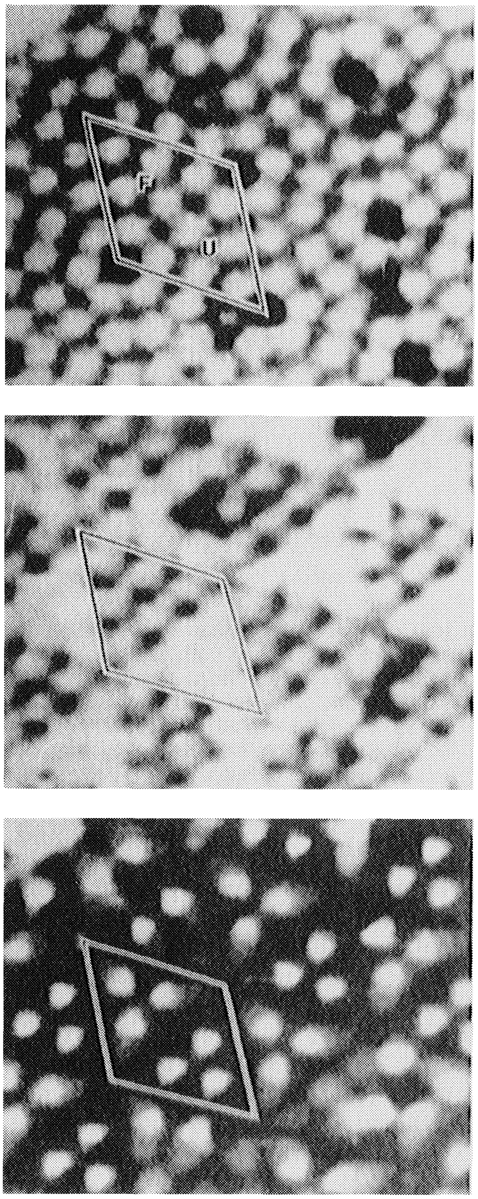

FIG. 15. STM images of Si(111) recorded with different bias voltages applied to the sample as follows: top, +2 V; center, +1.45 V; bottom, −1.45 V. The surface (7×7) unit mesh is superimposed on the micrographs for clarity. *F* and *U* refer to the two inequivalent ("faulted" and "unfaulted") halves of this unit mesh. After Hamers *et al.* (1986).

The sensitivity to electronic structure in STM can be exploited further by a modification of the usual imaging mode. In particular, if the tip is held at a fixed position above the surface, and the applied potential difference is scanned while measuring the changing tunneling current (now with the perpendicular positioning feedback loop switched off), the current voltage (*I*-*V*) characteristics should show steps associated with new tunneling channels being opened by appropriate occupied or unoccupied energy levels of the surface. By plotting a conductance versus applied voltage (*I*/*V*-*V*) graph these electronic energy levels can then appear as peaks similar to those seen in photoemission and inverse photoemission. This is confirmed by the data presented in Fig. 16, which shows conductance-voltage curves of this type recorded in both polarities and at various positions on the surface within a single surface unit mesh, as well as a spatially averaged spectrum. Also shown are (necessarily spatially averaged) photoemission and inverse photoemission spectra from this surface. It is clear that all of the main features of the occupied and unoccupied surface-state spectra produced by optical means are reproduced by the spatially averaged tunneling spectra, and also that each peak in the spectra is associated with a different part of the unit cell in real space. Indeed, direct spatial maps of these states can be obtained by building complete images based on the amplitude of specific features in the spatially resolved conductance-voltage spectra.

Although this example highlights the spatially localized character of semiconductor surface dangling-bond states (parallel as well as perpendicular to the surface), it also provides an elegant demonstration of the power of STM for electronic structure investigations of surfaces in suitable cases. Unfortunately, this restriction to "suitable cases" is an important one; although this technique has been applied to a number of other semiconductor surface problems with success, its broader applicability appears to be limited. For example, one might hope that it would be possible to map localized (Tamm) surface states on metal surfaces, but in the case of the transition and noble metals it appears that tunneling is dominated by the less spatially localized

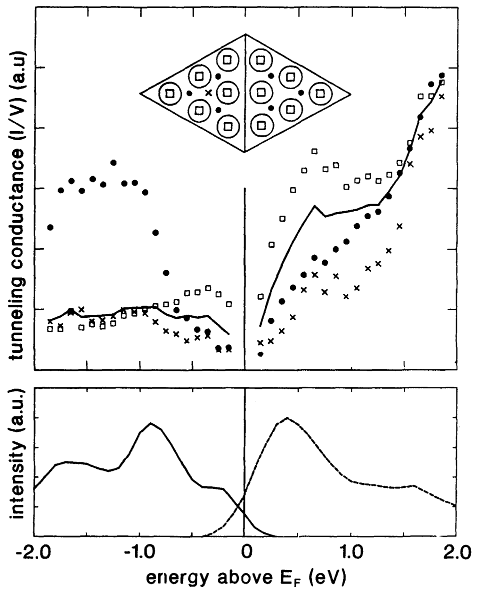

FIG. 16. Comparison of photoemission, inverse photoemission (lower part), and both spatially averaged (line) and spatially resolved (points) tunneling conductance-voltage spectra (upper part) from Si(111). The inset shows the areas sampled in the spatially resolved tunneling spectra, which are labeled by the same circles, squares, and crosses. After Hamers *et al.* (1986).

s-p electrons. Another potential limitation is that the tunneling process appears to favor predominantly transitions involving states at the surface Brillouin zone center (with $k_{\|} = 0$), although if no states exist here, strong features at other points in the zone can be observed. This selectivity in **k** clearly makes the technique complementary to the more conventional ARUPS and KRIPES methods.

Notice that one aspect of the existence of surface states on semiconductor surfaces, which often lie in the complete bulk band gap (i.e., not simply in a projected bulk gap, as on metal surfaces), is that it is then possible for the surface to be metallic, rather than semiconducting. In addition, the extent to which such surface states may be occupied can be strongly influenced by the type and extent of doping of the bulk semiconductor. The redistribution of charge between the surface and the bulk to balance the Fermi levels can then lead to interesting band-bending effects in the near-surface region. A proper quantitative description of surface states in semiconductors is far more complex than in metals, however, because of the essentially universal phenomenon of surface reconstruction (for III-V and II-VI, as well as elemental, semiconductors). In effect, the essential inevitability of dangling-bond states at ideally terminated semiconductor surfaces means that some rearrangement (both geometric and electronic) is necessary, but a quantitative modelling of the electronic structure requires an accurate geometrical structure as the starting point. This geometrical structure is not easily obtained because the covalently bonded network means that adjusting the position of one atom in the top layer causes movements of adjacent atoms in lower layers, typically down to 3 or 4 layers below the surface, and so there are many unknown structural parameters. Indeed, this intimate connection between the geometrical and electronic structure of the surface has meant that some progress in unraveling these structures has been made by measuring the electronic structure (with ARUPS and KRIPES) and comparing this with the results of model calculations based on different geometries. Further discussions of all of these effects may be found, for example, in Hansson and Uhrberg (1988), and in an article by the same authors in the book edited by Kevan (1992).

2.4 Surface Reconstruction

We have already remarked that semiconductor surfaces (almost invariably) have a local geometrical structure which is significantly different from that of an ideally terminated bulk solid, and that these reconstructions can be understood as a consequence of the rather local bonding in these materials and thus of the problem of reducing the excess energy associated with the dangling bonds present at an ideally terminated solid. It is now clear, however, that many metal surfaces also show substantial surface reconstructions, the origin of which is somewhat less clear. Early examples of these reconstructions were on W(100) and Mo(100), and it is tempting to suggest that even in these metallic systems a similar local bonding description based on the role of the localized valence *d* electrons can be used; indeed there has been some success in using local bonding models to explain the phenomenon.

An alternative picture of these reconstructions which has also found some favor, however, is related to the idea that a surface charge-density wave (CDW) may be associated with the reconstruction. This model relies on collective electronic effects rather than local bonding, and the reconstruction is seen as the freezing out of a particular surface phonon mode that has a surface wave vector spanning heavily nested portions of the two-dimensional Fermi surface. A necessary, if not sufficient, condition for this to occur is that the two-dimensional Fermi surface of the metallic surface does contain the required large parallel flat segments. In order to test this theory, therefore, ARUPS can be used to map out the two-dimensional Fermi surface by determining at which values of $k_{\|}$ surface states cross the Fermi level; each such crossing defines a point on the two-dimensional Fermi surface. This clearly requires very many ARUPS measurements to be made, ideally with good energy and angular (and thus $k_{\|}$) resolution, but this has been performed for a number of surfaces including the W and Mo(100) faces mentioned above.

The results for W(100) are summarized in Fig. 17 which shows the experimentally determined two-dimensional Fermi surface for this crystal face. Individual crossing points of the Fermi level are shown as circles, but there is also a large region which is hatched in which the ARUPS shows a surface state so close to the Fermi level that is difficult to be sure whether the Fermi surface is completely flat or not. Clearly these areas are heavily nested and provide strong support for some kind of charge-density-wave–driven reconstruction. Unfortunately, the nesting vectors are not quite in the expected range for the observed reconstruction [to a so-called $c(2\times2)$ surface periodicity]; a classic CDW reconstruction to this phase would require vectors joining nested regions along the $\bar{\Sigma}$ symmetry direction [the line joining the zone center ($\bar{\Gamma}$) and $\bar{M}$ points] with a length equal to the $\bar{\Gamma}$ to $\bar{M}$ distance. However, the largest vector in this direction joining the hatched (nested) regions is almost 10% shorter than this value. It appears, however, that these results are consistent with a further development of this basic theory in which electron-phonon effects are explicitly considered, indicating that a collective electronic effect does play a role in these reconstructions. A fuller discussion of these data, those on other surfaces, and their general significance can be found in a review by Smith and Kevan (1991).

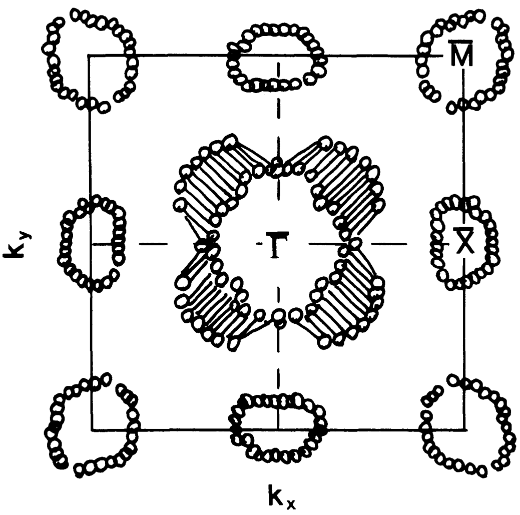

FIG. 17. Experimentally determined 2D Fermi surface of W(100) mapped in the first surface Brillouin zone. Data were recorded in one quarter of the zone only. The main crossing points of the Fermi level are marked by circles, but the hatched area also denotes regions in which the surface state is at, or very close to, the Fermi level. After Smith *et al.* (1990).

2.5 Magnetic Properties of Surfaces

Although the idea of surface magnetism in ferromagnetic materials has been discussed for many years, attempts to measure this frequently led to negative results, and the idea developed that the surfaces of such materials were magnetically "dead." It now seems clear, however, that this only true for contaminated surfaces, and that clean surfaces of ferromagnetic materials are, indeed, magnetic. Moreover, controlled studies of adsorption reveal that there is no simple universality to the destruction of the surface magnetism by adsorption, and indeed adsorbed species can themselves acquire magnetic electronic level splitting through hybridization with the underlying substrate.

The existence of magnetic surface states was inferred in quite early applications of ARUPS to the study of Ni surfaces by Plummer and Eberhardt (1979), through the observation of surface states appearing in regions of k space that only correspond to the bulk band gap for one spin state of the metal d electrons. More direct observation of the spin splitting of the d states, which produces the magnetism, has come from spin-resolved

ARUPS and KRIPES measurements. These techniques rely on the fact that not only energy and momentum, but also electron spin is conserved in the photoemission and inverse photoemission processes. This means that if the spin of photoemitted electrons is detected, or if the incident electron beam in an inverse photoemission experiment is spin-polarized, the occupied and unoccupied states of the near-surface region can be mapped in energy, momentum, and spin. Examples of such measurements on Ni(110) are shown in Fig. 18 for normal photoelectron emission and normal electron incidence inverse photoemission from Ni(110). The photoemission data clearly show the spin splitting of the occupied portion of the 3*d* bands, while the inverse photoemission shows that the unoccupied portion of the *d* bands is entirely of one (minority) spin; the unequal occupation of the two different spin states is the origin of the ferromagnetic behavior of this material.

Of course, we have already remarked in Sec. 1 that the ARUPS and KRIPES techniques are quite effective in mapping bulk band structure, in that the conservation of $k_\perp$ remains rather dominant despite the relatively small sampling depth of these methods. Although measurements such as those shown in Fig. 18 therefore provide valuable support for our understanding of bulk ferromagnetism, they do not necessarily demonstrate surface magnetism. Other measurements do, however. For example, it is possible to determine the temperature dependence of the spin polarization of electrons photoemitted from such a surface, and this is found to deviate very significantly from that to be expected for bulk magnetism, showing that at least part of the measured magnetic properties is surface specific.

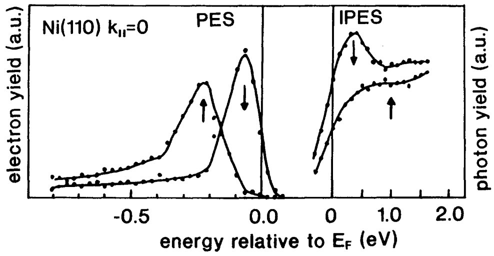

FIG. 18. Comparison of experimental results of spin- and angle- (momentum-) resolved photoemission and inverse photoemission data from Ni(110) recorded at normal emission and normal incidence, respectively. Taken from Schmitt *et al.* (1987) and Ungaris *et al.* (1982).

A small extension of studies of this type to the investigation of the magnetic properties of ultrathin metal films (i.e., of one to ten single atomic layers or so) is also proving extremely interesting, as is the study of metastable (constrained) phases of similarly thin films produced by epitaxial growth, such as of fcc Fe and Co by growth on copper surfaces. So far studies of such material by spin-polarized ARUPS and KRIPES have been limited (these experiments combining the problems of simultaneous energy, angle, and spin resolution with the small amounts of material available in monolayer films), but it is already clear that the magnetic properties of these layers can differ significantly from those of the bulk. For example, the fcc Fe films are found to be ferromagnetic, but the Curie temperature is thickness dependent (e.g., 230 K for one monolayer and 390 K for films of 3- and 5-monolayer thicknesses) (Stampanoni *et al.*, 1987).

3. ADSORBATE-INDUCED SURFACE ELECTRONIC STRUCTURE

A primary reason for studying the electronic structure of a surface in the presence of adsorbed species is to try to understand the nature of the adsorbate-substrate bonding together with any possible modification of the intrinsic electronic structure of the underlying surface and of the adsorbate itself (particularly for an adsorbed molecular species) which may help us to understand the chemical reactivity of the surface. A slightly different motivation also exists in the case of semiconductor surfaces for which the adsorbed species may be oxygen, a metallic species, or another semiconductor material; in this case one may wish to understand the early stages of semiconductor-insulator, semiconductor-metal (Schottky barrier) and semiconductor-semiconductor (heterojunction) interfaces of importance in fabricating semiconductor devices, particularly by ultrahigh-vacuum methods such as molecular beam epitaxy.

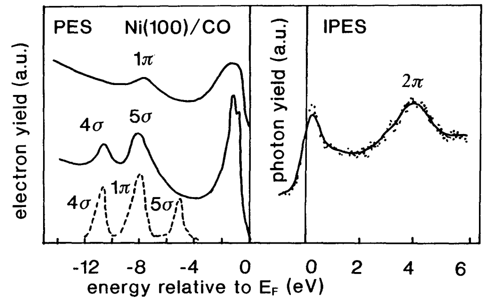

FIG. 19. Photoemission and inverse photoemission spectra recorded from CO adsorbed on Ni(100). Also shown as a dashed line is the gas-phase photoemission spectrum suitably aligned in energy. The photoemission spectra, taken from Allyn, Gustafsson, and Plummer (1977), are recorded in two different polarization conditions to separate out the different symmetry orbitals. The inverse photoemission spectrum is from Johnson and Hulbert (1987).

3.1 Molecular Bonding and Valence Electronic Structure

In the case of adsorbed molecules or molecular fragments, photoemission and inverse photoemission can provide a valuable spectral "fingerprint" of the molecular character of the adsorbate because of the characteristic energies of their valence orbitals. More careful inspection can also provide valuable information on the adsorbate bonding and molecular modification. One example of this is in studies of the adsorption of molecular carbon monoxide on transition-metal (and noble-metal) surfaces, a topic of great chemical interest in heterogeneous catalysts for processes as variable as methanol production, gasoline and methane production, and automobile exhaust-gas pollution control. Figure 19 shows valence-band photoemission and inverse photoemission spectra taken from a Ni(100) surface onto which CO has been adsorbed. The photoemission spectrum of the gas-phase molecule in this energy range contains three main features associated with the 4σ, 1π, and 5σ orbitals in order of decreasing binding energy, these orbitals being mainly derived from the C and O $2p$ atomic orbitals of the constituent atoms. In the adsorbed-molecular spectrum, however, one typically only observes two peaks (other than those associated with the underlying metal), although by the use of polarization selection rules in angle-resolved photoemission, combined with the fact that the C—O bond axis is almost always found to be perpendicular to the surface (thus preserving the symmetry of the situation), it is possible to show that the smaller-binding-energy peak actually comprises two overlapping peaks of different (π and σ) symmetries. The origin of this effect is that while the 4σ and 1π orbitals are largely undisturbed by the adsorption, the 5σ orbital is pulled down in energy by the interaction with the substrate to overlap the photoemission peak from the 1π state. This selective modification of the 5σ state is consistent with the fact that the molecule bonds with the C end against the metal, because the charge density associated with the 5σ orbital is most concentrated beyond this carbon end and so will overlap the metallic charge and hybridize with metal states of the appropriate symmetry. This hybridization is expected to form a CO-metal bonding–antibonding pair of levels, with the bonding level being lower in energy than either of the constituent levels. The displacement of the 5σ energy level can thus be assigned to this substrate bonding shift.

A more complete picture of the CO–transition-metal bonding picture in this molecular description is shown in Fig. 20 and illustrates two further important points. One is that the CO-5σ–metal d-band bonding actually only truly contributes to the adsorbate bonding if its associated antibonding level is at least partially unoccupied (it is shown above the Fermi level in Fig. 20). The second

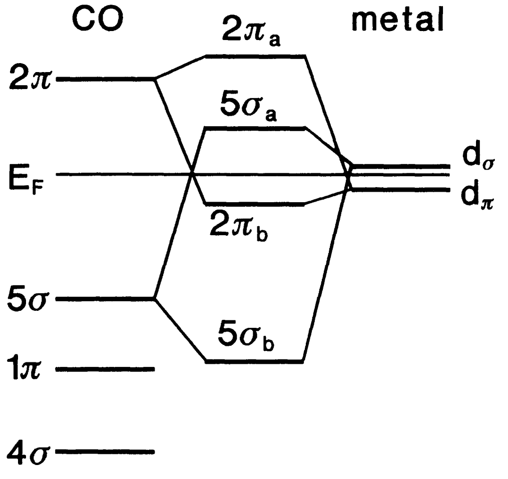

FIG. 20. Schematic energy level diagram showing the conventional (Blyholder) bonding scheme for CO adsorbed on a transition-metal surface.

feature is that a bonding interaction is also believed to occur between the unoccupied CO 2π orbital and the metal *d* band. The interaction is normally described in terms of CO charge donation to the metal from the 5σ orbital and backdonation of metal charge to the CO 2π. The involvement of the CO 2π orbital is of great importance in this picture because this orbital is antibonding with respect to the C–O bond and is completely unoccupied in the free molecule. If the surface bonding does donate charge to the CO 2π state this will therefore weaken the C–O bond making it easier to fracture, a process which is important in some of the CO chemistry which is aided by transition-metal catalysts. In this respect it is clear that the key issue is not the energetic location of the unoccupied 2π-derived level, which is also antibonding with respect to the CO-metal bond (although it is presumably this level which is responsible for the peak seen in the inverse photoemission spectrum), but rather the existence or otherwise of an occupied component of this state which is expected to be located close to the Fermi level. This is actually much more difficult to observe because it should occur in the energy range occupied by the substrate *d* band which typically dominates photoemission spectra. Nevertheless, there are some reported observations of this state in photoemission.

One rather different method that has been used to observe this feature (which is less troubled by the interfering influence of a dominant metal substrate signal) is based on the use of metastable He atoms, rather than ultraviolet photons, as the source of energy input to eject electrons from occupied states. In this technique, variously referred to as He metastable atom spectroscopy or surface Penning ionization spectroscopy, a metastable excited He-atom beam is directed at the surface and the emitted electrons are collected and energy analyzed as in photoemission. The excited He atoms used have one of the ground-state $1s$ electrons excited into a $2s$ state to form the metastable singlet 2^1S state (or the triplet 2^3S state) but can deexcite when they interact with the surface to give up their excess energy of 20.6 eV (or 19.8 eV for the triplet state, both being close to the 21.2 eV of the He I resonance source photons typically used in laboratory UPS). The form of the de-excitation can vary, but in the case of studies of adsorbed CO, appears to be by an Auger de-excitation process (see AUGER ELECTRON SPECTROSCOPY) in which an electron in the valence band of the surface tunnels into the $1s$ state of the He atom, and the He $2s$ electron is emitted with an energy characteristic of the total available energy in the metastable atom minus the binding energy of the valence electron involved. A key feature of this process is that the charge exchange is very localized, so that for a surface with adsorbed CO, the technique probes mainly the electronic states of the adsorbed molecule and not of the underlying substrate. This extremely high surface specificity makes it particularly useful for investigating the electronic states localized on adsorbed species rather than those associated with the underlying substrate, although the technique is clearly less simple and generally less applicable than UPS.

3.2 Atomic Adsorbates and Associated Two-Dimensional Bands

In studies of the electronic structure associated with adsorbed atomic species the valence electrons of the adsorbed atom no longer offer a particularly incisive fingerprint of the chemical identity of the adsorbate, but they still aid in understanding the adsorbate-substrate bonding, and for ordered overlayers the adsorbate-induced surface localized states form two-dimensional bands which disperse in $k_{\parallel}$ as a result of adsorbate-adsorbate interactions both through space (directly) and through interaction with the substrate. An example of data of this kind, in the form of experimental two-dimensional band structures obtained by ARUPS from ordered half-monolayer structures of C, N, and O on Ni(100), is shown in Fig. 21. Three main spectral regions may be seen in these data. The largest binding-energy bands [binding energies of about 12 eV (C) to 22 eV (O)] correspond to the adsorbate $2s$ electronic states and form rather flat bands because of their relatively high degree of localization; the upward dispersion away from the center of the surface Brillouin zone is as expected from a simple tight-binding model.

In the energy range around 4–8-eV binding energy are seen bands associated with the adsorbate $2p$ states. These states overlap the photoemission signal from the Ni substrate, and so the full extent of the adsorbate-induced bands cannot be followed reliably at all points in **k** space. Nevertheless, we can expect that at

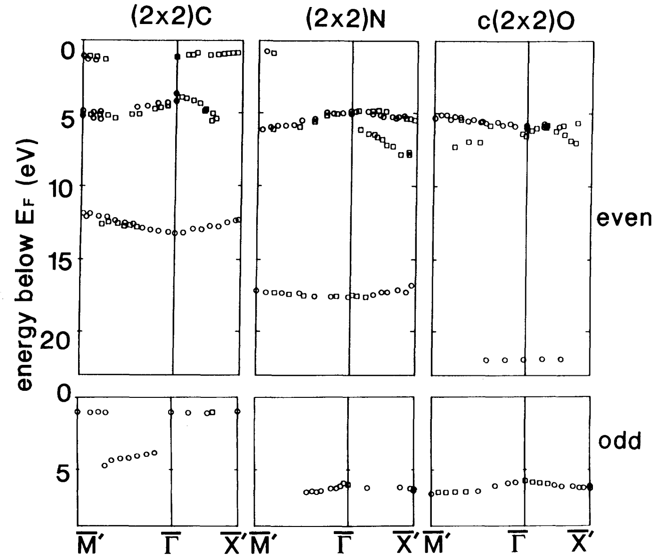

FIG. 21. Experimentally determined surface band structures of the adsorbate-induced states measured by ARUPS from half-monolayer structures of C, N, and O on Ni(100), plotted within the first surface Brillouin zone corresponding to a $c(2\times 2)$ overlayer mesh. After Kilcoyne *et al.* (1989).

the zone center the p_x and p_y states will be degenerate while the p_z-derived states (z perpendicular to the surface) will appear at a somewhat different energy, as a result of their different adsorbate-adsorbate and adsorbate-substrate interactions, even at this symmetry point. Note that the p_y states have odd parity relative to the mirror planes which contain the detector, and this difference in symmetry can be distinguished by polarization effects in photoemission; these states are plotted separately in the lower part of the figure. Different dispersion away from the zone center of the p_x-, p_y-, and p_z-derived states is expected to lead to three nondegenerate bands elsewhere in the two-dimensional Brillouin zone, although all three cannot always be discerned in the experimental data.

In addition, the C and N adsorption structures show some evidence of new surface localized states just below the Fermi level, although as these occur in an energy region overlapping the strong substrate d-band emission, the associated ARUPS peaks can only be reliably identified in some narrow regions of **k** space in which the d-band features are weak. These states close to the Fermi level are of particular interest, because although the half-monolayer coverage of O on Ni(100) appears to produce no significant distortion of the metallic substrate, both N and C do lead to a significant reconstruction. Following the arguments of Sec. 2.4 we might therefore expect to see electronic structure changes close to the Fermi level accompanying the geometrical reconstruction. The form of the reconstruction is illustrated in Fig. 22, in which the top-layer Ni atoms for an ideally terminated Ni(100) surface are shown as dashed circles, while the actual reconstructed sites are denoted by hatched circles; the adsorbed C or N atoms are shown as filled

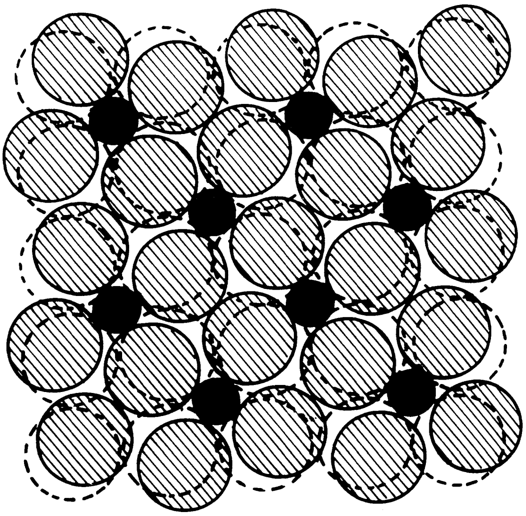

FIG. 22. Schematic diagram of the reconstruction induced by half-monolayer coverages of either C or N on a Ni(100) surface. The filled circles are the adsorbed atoms, the hatched circles are the top-layer Ni atoms, and the dashed circles show the positions of these top-layer Ni atoms for an ideally truncated surface.

circles. The origin of this adsorbate-induced substrate reconstruction is still far from clear but is of considerable academic as well as potential practical importance as a heterogeneous catalyst. The latter derives from the fact that the C adsorption structure appears to be the catalytically active surface phase in a methanation reaction conducted over this surface, implying that it is the electronic and chemical properties of this surface, rather than those of the clean metal, that are of chemical importance. In view of this, the small changes in the surface in the metal *d*-band region might be anticipated, but as yet the information available is too sparse to provide any real understanding of this problem.

3.3 Chemical Shifts in Core-Level Photoemission

As we have already remarked in Sec. 2.2, small shifts in the measured photoelectron binding energies of atomic core levels are well known to occur when the atom has its local valence electronic structure modified by different types of bonding, and these so-called chemical shifts (typically varying from a few tenths of an electron volt to several electron volts) provide an extremely valuable fingerprint of local bonding interactions at surfaces as well as in molecules and bulk compounds. The intrinsic surface sensitivity of the XPS method, which derives mainly from the short inelastic-scattering mean free path for electrons in the 10–1500-eV energy range appropriate to the method, means that it is particularly well suited to these surface studies.

A quantitative understanding of the magnitude (and even direction) of the observed chemical shifts is not at all straightforward because of the dual role of initial- and final-state effects, as described in Sec. 2.2. It is tempting to assign these chemical shifts to simple initial-state charge-transfer effects, but final-state relaxation energy shifts can be at least as large as these initial-state energy changes, and can be either of the same sign or the opposite sign as the initial state effects. For example, calculations of these two effects for O, Na, Si, and Cl, on an ideal free-electron surface (jellium) with an electron density appropriate to that of Al, yield initial-state energy shifts for adsorbate core levels of approximately +3.0, −1.0, +1.8, and +1.4 eV, respectively, but final-state shifts due to adsorbing these species of +7.8, +4.6, +5.2, and +4.8 eV (Williams and Lang, 1977). One possible quantitative route to the use, rather than interpretation, of these chemical shifts, is via the equivalent hole picture described earlier (Sec. 2.2) in which one recognizes that core ionization leaves behind a relaxed valence electronic environment equivalent to that of a neutral atom of the next element of the periodic table. It is therefore possible, by the construction of thermodynamic Born-Haber cycles (in which the energy changes in ionizing, adsorbing, and transmuting an atom of atomic number Z to one of $Z+1$ are followed around a closed loop), to relate core-level shifts to heats of adsorption (or segregation) at surfaces (Johansson and Mårtensson, 1980; Rosengren and Johansson, 1981).

Far more common, however, is the use of chemical shifts as a fingerprint of changes in local electronic environment or "chemical state." In studies of basic surface chemistry one often uses the measured photoelectron binding energies in solids or molecules of known valency to help identify the valency of adsorbed species. In other cases, one may simply concentrate on the fact that different chemical shifts signal changes in chemical state which can be correlated with changes seen in other techniques (which could be electronic, vibrational, or structural probes). An example of this kind of application, which also highlights the fact that interesting chemical shifts may be seen in the substrate as well as the adsorbate atoms, is shown in Fig. 23. It is well known that the Al $2p$ core-level photoemission peaks (spin-split doublets with a separation of approximately 0.4 eV) show a large chemical shift of some 2.6 eV when an aluminum surface is oxidized. In the case of Al(111), this broad oxide-state peak is seen at large oxygen exposures [measured in Langmuirs, where one Langmuir corresponds to an exposure at a pressure of 10^{-6} Torr (1.33×10^{-4} Pa or 1.33×10^{-6} mbar) for 1 s], but at lower exposures three different, intermediately shifted pairs of peaks (with shifts of approximately 0.5, 1.0, and 1.5 eV) appear. Figure 23 shows the rich spectrum of features resulting from this effect, but although the surface sensitivity and energy resolution have been enhanced by using synchrotron-radiation photons of energy 120 eV, proper resolu-

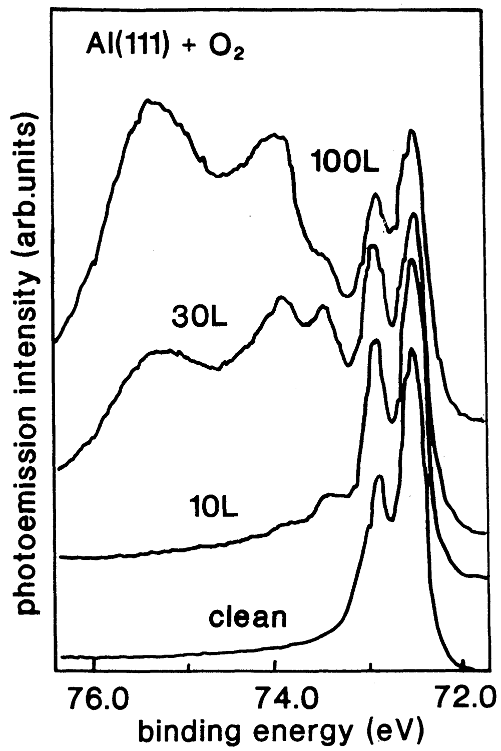

FIG. 23. Al 2*p* photoemission spectra recorded at a photon energy of 100 eV from an Al(111) surface after various exposures of oxygen, given in Langmuir units. (1 Langmuir$=10^{-6}$ Torr·s$=1.33\times10^{-4}$ Pa·s). After McConville *et al.* (1987).

tion of all these peaks requires simple quantitative curve fitting. The analysis of the data indicates that all of these extra peaks are associated with surface Al atoms bonded to a simple chemisorbed atomic oxygen overlayer, whereas the higher chemical shift oxide peak is associated with Al atoms in a thin surface oxide, probably involving both overlayer and underlayer oxygen. One surprising feature of this result is that it indicates that quite local Al–O bonding must occur even on this rather free-electron-like metal surface.

List of Works Cited

Allyn, C. L., Gustafsson, T., Plummer, E. W. (1977), *Chem. Phys. Lett.* **47**, 127–132.

Eastman, D., Himpsel, F., van der Veen, J. F. (1982), *J. Vac. Sci. Technol.* **20**, 609–616.

Euceda, A., Bylander, D. M., Kleinman, L. (1983), *Phys. Rev. B* **28**, 528–534.

Hamers, R. J., Tromp, R. M., Demuth, J. E. (1986), *Phys. Rev. Lett.* **56**, 1972–1975.

Hansson, G. V., Uhrberg, R. I. G. (1988), *Surf. Sci. Rep.* **9**, 197–292.

Hulbert, S. L., Johnson, P. D., Stoffel, N. G., Royer, W. A., Smith, N. V. (1985), *Phys. Rev. B* **31**, 6815–6817.

Johansson, B., Mårtensson, N. (1980), *Phys. Rev. B* **21**, 4427–4457.

Johnson, P. D., Hulbert, S. L. (1987), *Phys. Rev. B* **35**, 9427–9436.

Kevan, S. D. (1983), *Phys. Rev. Lett.* **50**, 526–529.

Kevan, S. D. (Ed.) (1992), *Angle-Resolved Photoemission*, Amsterdam: Elsevier.

Kilcoyne, A. L. D., Woodruff, D. P., Rowe, J. E., Gaylord, R. H. (1989), *Phys. Rev. B* **39**, 12604–12611.

McConville,C. F., Seymour, D. L., Woodruff, D. P., Bao, S. (1987), *Surf. Sci.* **188**, 1–14.

Plummer, E. W., Eberhardt, W. (1979), *Phys. Rev. B* **20**, 1444–1459.

Rosengren, A., Johansson, B. (1981) *Phys. Rev. B* **23**, 3852–3858.

Schmitt, W., Kämper, K.-P., Güntherodt, G. (1987), *Surf. Sci.* **189/190**, 741–746.

Smith, K. E., Elliott, G. S., Kevan, S. D. (1990), *Phys. Rev. B* **42**, 5385–5287.

Smith, K. E., Kevan, S. D. (1991), *Prog. Solid State Chem.* **21**, 49–131.

Smith, N. V., Woodruff, D. P. (1986), *Prog. Surf. Sci.* **21**, 295–370.

Stampanoni, M., Pescia, D., Zampieri, G., Bona, G. L., Vaterlaus, A., Meier, F. (1987), *Surf. Sci.* **189/190**, 736–740.

Ungaris, J., Seiler, A., Celotta, R. J., Pierce, D. T., Johnson, P. D., Smith, N. V. (1982), *Phys. Rev. Lett.* **49**, 1047–1050.

Williams, A. R., Lang, N. D. (1977) *Surf. Sci.* **68**, 138–148.

Woodruff, D. P., Johnson, P. D., Smith, N. V. (1983), *J. Vac. Sci. Technol. A* **1**, 1104–1110.

Zangwill, A. (1988), *The Physics of Surfaces*, Cambridge: Cambridge University Press.

Further Reading

Inglesfield, J. E., Holland, B. W. (1981), in: D. A. King, D. P. Woodruff (Eds.), *The Chemical Physics of Solid Surfaces and Heterogeneous Catalysis*, Vol. 1, Amsterdam: Elsevier, Chap. 3.

Kevan, S. D. (Ed.) (1992) *Angle-Resolved Photoemission*, Amsterdam: Elsevier.

Scheffler, M., Bradshaw, A. M. (1983), in: D. A. King, D. P. Woodruff (Eds.), *The Chemical Physics of Solid Surfaces and Heterogeneous Catalysis*, Vol. 2, Amsterdam: Elsevier, Chap. 3.

Smith, K. E., Kevan, S. D. (1991), *Prog. Solid State Chem.* **21**, 49–131.

Smith, N. V. (1988), *Rep. Prog. Phys.* **51**, 1227–1294.

Smith, N. V., Woodruff, D. P. (1986), *Prog. Surf. Sci.* **21**, 295–370.

Woodruff, D. P., Delchar, T. A. (1986), *Modern Techniques of Surface Science*, Cambridge: Cambridge University Press.

Zangwill, A. (1988), *The Physics of Surfaces*, Cambridge: Cambridge University Press.

ELECTRONIC TRANSPORT IN SOLIDS

OLE PER HANSEN, *University of Copenhagen, Copenhagen, Denmark*

Introduction 121
1. **Currents, Fields, and Relaxation Time; Symbols** ... 122
2. **The Drude Model: Ohm's Law** 122
2.1 Optics 123
2.2 dc Electric and Magnetic Fields........ 123
2.2.1 Hall Effect, Magnetoresistance, and Mobility 124
2.2.2 Multicarrier Systems: Semiconductors and Semimetals 124
2.2.3 Metals........ 126
2.3 Magneto-optics: Cyclotron Frequency 128
2.4 Thermal Conductivity, Fourier's Equation, and Thermal Diffusivity........ 129
2.5 Limitations 130
3. **Wave Features and *k* Space** .. 130
3.1 Metal-Insulator Transitions ... 130
4. **The Boltzmann Equation**..... 131
4.1 Scattering of Charge Carriers.. 132
4.1.1 Charged Impurities........ 132
4.1.2 Lattice Vibrations 132
4.1.3 Mixed Scattering 134
4.2 The Wiedemann-Franz Law... 135
5. **Linear Transport Equations: Peltier and Thermoelectric Tensors**........ 136
6. **Symmetry Restrictions**....... 137
6.1 Onsager Relations 137
6.2 Point-Group Symmetry 137
7. **Sample Geometries** 138
7.1 Hall Geometry 138
7.2 Corbino Geometry........ 138
7.3 van der Pauw Method 139
8. **Landau Quantization**........ 140
8.1 Shubnikov–de Haas Oscillations 140
9. **Transport in Low-Dimensional Systems** ... 140
9.1 The Quantum Hall Effect 141
9.2 Size-Quantized Conductivity... 142
Works Cited 143

INTRODUCTION

"Transport in solids" covers quite a large number of phenomena, but obviously, transport in liquids and gases is not a part of these. In the present context, "electronic" transport in solids will be delimited as the transport of charge and heat in solid electrical conductors, i.e., metals, semimetals, and semiconductors. Insulators, with their vanishingly small conductivities, will not be treated.

The concept of charge, i.e., the elementary charge of the electron, is supposed to require no explanation. Transport of charge means an electric current and, as the electron carries a mass, a transport of mass. A heat current, on the contrary, is defined in a somewhat subtle way: when a particle enters or leaves a material, it gives to the surroundings, or takes from them, the work function of the material. Such a current of work function is just proportional to the electric current, and consequently it is not a new concept. The new concept, the heat current, appears when the current of work function is subtracted from the energy current.

The two concepts are illustrated in Fig. 1 in terms of velocity distributions for a definite direction, say x, i.e., the probability of finding electrons with velocity v_x is displayed as function of v_x. In 1(a) there are no currents at all. In 1(b) there is an electric current, but no heat current. Finally in 1(c) there is a heat current, but no electric current.

The units of the two currents are coulomb per second and joule per second, with the shorter names ampere (A) and watt (W), respectively. The current passing through a

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

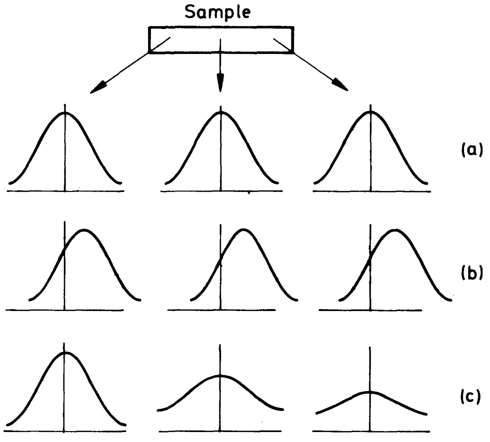

FIG. 1. Velocity distributions at three places in a sample, i.e., probabilities of finding the velocity v_x is displayed as function of v_x. In (a) the distribution is the same everywhere, and the average value of v_x is zero everywhere. In (b) the distribution is the same everywhere, but is shifted with respect to (a) so that it has a nonzero average value. Finally in (c) the average value of v_x is zero everywhere but the distributions become broader towards one end of the sample (the warm end).

unit square perpendicular to the direction of flow is denoted current density. The two current densities are thus measured in units A m^{-2} and W m^{-2}.

To produce a current, an external action is needed. This can be gradients of electrical potential, of temperature, or of electron concentration, i.e., spatial changes of these quantities. If a static magnetic field is applied to a sample in which a current flows, the nature of the flow is changed. By itself, a static magnetic field does not produce a current, and accordingly, in the equations describing the currents, such a field is treated differently from the gradients of electrical potential, temperature, or concentration. The very same remark holds for an applied static pressure.

The structure of the present article is largely historic, starting with the description used in the beginning of this century, and ending with a few of the last decades' discoveries. Quantum-mechanical features are treated, but without writing down the Schrödinger equation. The reader is assumed to be familiar with vectors and tensors, as well as with complex numbers.

For a deeper insight, the textbooks by Ziman (1960) and by Smith and Jensen (1989) are recommended, as well as cross references in the present encyclopedia, e.g., TRANSPORT PHENOMENA; HEAT TRANSFER; THERMAL CONDUCTION IN SOLIDS; SOLIDS, DIFFUSION IN; SUPERCONDUCTIVITY, LOW-TEMPERATURE; HIGH-TEMPERATURE SUPERCONDUCTIVITY.

1. CURRENTS, FIELDS, AND RELAXATION TIME; SYMBOLS

With two types of currents and three types of forces to drive the currents, plus the possibilities of an applied static magnetic field or pressure, the number of phenomena becomes large. Moreover, the actions inside the solid, the so-called scattering of electrons, need a level of description. For this, the most elementary one is in terms of a single parameter: the relaxation time τ. These concepts, with their symbols, are illustrated in Fig. 2.

2. THE DRUDE MODEL: OHM'S LAW

Already in 1900 P. Drude put forward a model that took account of relaxation effects. Drude's model consists mainly in describing the flow by the average velocity, $\mathbf{v}_a$, resulting from external forces and internal "frictional" forces, along with assuming that all carriers act independently. The Drude frictional force is

$$\mathbf{F}_{\text{Drude}} = -m\mathbf{v}_a/\tau\,; \tag{1}$$

and, in the very simplest case where the external force arises from an electric field **E**, the Newtonian equation of motion becomes

$$m\frac{d\mathbf{v}_a}{dt} = q\mathbf{E} - \frac{m\mathbf{v}_a}{\tau}\,. \tag{2}$$

The meaning of the Drude relaxation time τ is illustrated in Fig. 3, where a step function for the electric field is considered. The expression for $v_a(t)$ is the solution of Eq. (2), and from this the relaxation time appears as the time needed for the velocity to reach the fraction $1-e^{-1}$ of the steady state value $(q\tau/m)E_0$.

The electric current density in the steady state of Fig. 3 is

$$\mathbf{j} = nq\mathbf{v}_{a,\infty} = (nq^2\tau/m)\mathbf{E}_0 = \sigma_0\mathbf{E}_0, \tag{3}$$

where σ_0 is the electric conductivity in the Drude model,

$$\sigma_0 = nq^2\tau/m. \tag{4}$$

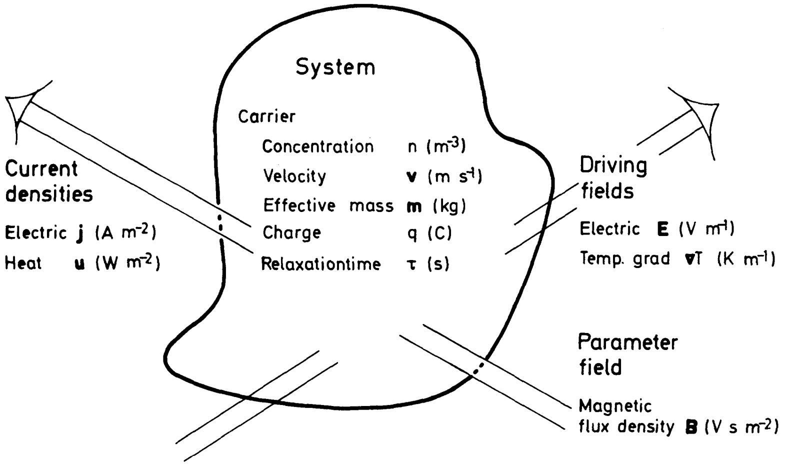
System
Carrier
Concentration n (m⁻³)
Velocity v (m s⁻¹)
Effective mass m (kg)
Charge q (C)
Relaxationtime τ (s)
Current densities
Electric j (A m⁻²)
Heat u (W m⁻²)
Driving fields
Electric E (V m⁻¹)
Temp. grad ∇T (K m⁻¹)
Parameter field
Magnetic flux density B (V s m⁻²)

FIG. 2. Symbols (units) used in the present article. The carrier charge q can be either $-e$ (electrons) or $+e$ (holes), e being the elementary charge: $e=1.602\times10^{-19}$ C. The effective mass **m** is different from the mass of a free electron; it has a value characteristic for each solid. It can be anisotropic like the relaxation time τ.

The linear relation between current density and electric field expressed by Eq. (3) displays Ohm's law.

2.1 Optics

If, instead of the step function, an alternating electric field of frequency f is applied,

$$\mathbf{E}=\mathbf{E}_0 e^{-i2\pi ft}, \tag{5}$$

then there is no steady state, $dv_a/dt=-i2\pi f v_a$, and the conductivity becomes

$$\sigma=\sigma_0(1-i2\pi f\tau)^{-1}. \tag{6}$$

For $2\pi f\tau=1$, the imaginary and the real parts of σ are identical.

The link to optical properties is provided by the Maxwell equation for the curl of the magnetic field strength:

$$\nabla\times\mathbf{H}=\mathbf{j}+\epsilon_L\frac{\partial\mathbf{E}}{\partial t}=\epsilon\frac{\partial\mathbf{E}}{\partial t}. \tag{7}$$

The quantity ϵ so defined is denoted permittivity, and using $j=\sigma E$ the following relation results:

$$\epsilon=\epsilon_L-\sigma/i2\pi f, \tag{8}$$

where ϵ_L is the permittivity related to the ion cores.

Figure 4 illustrates the application of the Drude model to gold. The high-frequency part of the data was obtained from measurements of reflectance versus frequency; the low-frequency part, from waveguide transmission.

2.2 dc Electric and Magnetic Fields

When electric as well as magnetic fields are applied, the term $q\mathbf{E}$ in Eq. (2) is replaced by the Lorentz force, i.e.,

$$\mathbf{m}\cdot\frac{d\mathbf{v}_a}{dt}=q(\mathbf{E}+\mathbf{v}_a\times\mathbf{B})-\frac{\mathbf{m}\cdot\mathbf{v}_a}{\tau}, \tag{9}$$

where **m** allows for an anisotropic mass. In the steady state the solution for $\mathbf{v}_a$ is

$$\mathbf{v}_a=[(\mathbf{m}/\tau)+q\boldsymbol{B}]^{-1}\cdot q\mathbf{E}. \tag{10}$$

Here $\boldsymbol{B}$ is the tensor representation of the magnetic flux density,

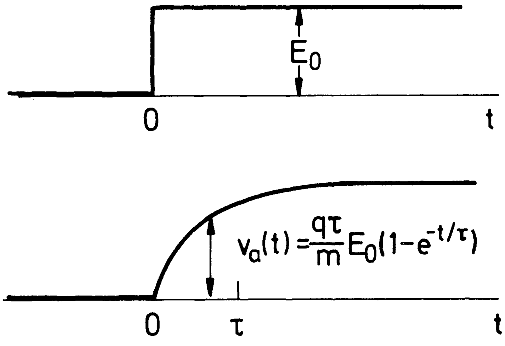

FIG. 3. Step function of the electric field and the velocity resulting from Eq. (2) as function of time.

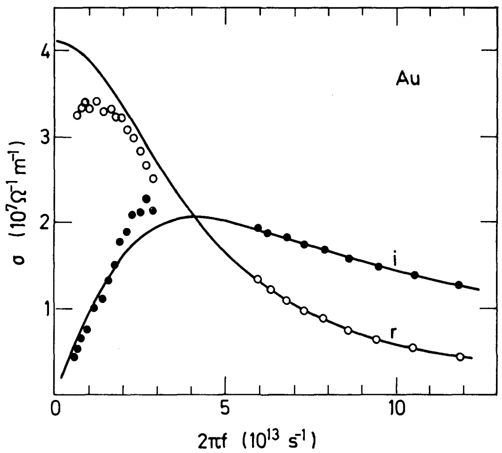

FIG. 4. Real (*r*) and imaginary (*i*) parts of the conductivity of gold in the infrared as function of the frequency. The high-frequency part represents a fit to data of reflectance vs frequency with $\sigma_0=4.09\times10^7\ \Omega^{-1}\ \mathrm{m}^{-1}$, $\tau=2.46\times10^{-4}$ s. The continuous curve is based on these values and Eq. (6). The low-frequency part was obtained from measurements of waveguide transmission (Smith and Jensen, 1989, p. 103).

$$\boldsymbol{B}=\begin{pmatrix} 0 & -B_z & B_y \\ B_z & 0 & -B_x \\ -B_y & B_x & 0 \end{pmatrix}, \tag{11}$$

and the expression (10) for $\mathbf{v}_a$ is found using the identity

$$\mathbf{v}_a\times\mathbf{B}=-\boldsymbol{B}\cdot\boldsymbol{v}_a\,. \tag{12}$$

From Eq. (10) the expression for the conductivity becomes

$$\sigma=\sigma_0\cdot\,(\mathbf{1}+q\tau\boldsymbol{B}\cdot\boldsymbol{m}^{-1})^{-1}, \tag{13}$$

$$\sigma_0=nq^2\tau\mathrm{m}^{-1}. \tag{14}$$

2.2.1 Hall Effect, Magnetoresistance, and Mobility. The concepts of Hall effect and magnetoresistance refer to the inverse of the conductivity tensor, the resistivity tensor $\boldsymbol{\rho}$. With $\boldsymbol{\sigma}$ given by Eq. (13) the resistivity tensor becomes

$$\rho=\sigma^{-1}=(\mathbf{1}+q\tau\boldsymbol{B}\cdot\boldsymbol{m}^{-1})\cdot\rho_0=\rho_0+(1/nq)\boldsymbol{B}. \tag{15}$$

The ways in which the elements of $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ can be measured are discussed in Sec. 7. At present it is just noted that Hall effect and magnetoresistance are described by, respectively, the nondiagonal and diagonal elements of $\boldsymbol{\rho}$.

If the direction of $\mathbf{B}$ is chosen as the z direction, it follows from Eqs. (11) and (15) that

$$\rho_{xy}=-\rho_{yx}=-B/nq. \tag{16}$$

The Hall coefficient is defined as

$$R_{\mathrm{H}}=\rho_{yx}/B. \tag{17}$$

The expression for R_{H} finds extremely widespread use in determination of the carrier density, and, as the charge q appears to the first power, also of the type of carriers, i.e., electrons or holes. In contrast to this, the diagonal elements—being independent of B—do not explain the experimentally well-known phenomenon of magnetoresistance. The reason for this failure can be understood from the fact that the average velocity $\mathbf{v}_a$ was assumed to be the same for each carrier. Consequently, the force $q\mathbf{v}_a\times\mathbf{B}$ is the same for each carrier and is balanced by the electric field from the surface charge, cf. Fig. 20. Hereby, the effect of $\mathbf{B}$ on $\mathbf{v}_a$ is cancelled. From this argumentation a nonzero magnetoresistance is to be expected whenever different carriers, or groups of carriers, have different $\mathbf{v}_a$'s. This is especially the case for semiconductors and semimetals with two types of carriers: negatively charged electrons ($q=-e$) and positively charged holes ($q=e$). Here Eq. (13), when used for each of the two types, gives in most cases a fair description of the Hall effect as well as of the magnetoresistance. For metals more subtle ideas have to be invoked.

In the expressions for $\boldsymbol{\sigma}$ and $\boldsymbol{\rho}$ the factor $q\tau\mathbf{m}^{-1}$ appears on the magnetic flux density. This factor is denoted as mobility, μ. Usually a positive value is meant, i.e.,

$$\mu=e\tau\mathrm{m}^{-1}. \tag{18}$$

Mobilities are measured in units of inverse magnetic flux density: $[\mu]=(\mathrm{V\ s\ m^{-2}})^{-1}$.

2.2.2 Multicarrier Systems: Semiconductors and Semimetals. In semiconductors and semimetals, electrons as well as holes can be carriers of the electric current density. To distinguish the two, subscripts e and h will be used. Thus the equation

$$\mathbf{j}=\mathbf{j}_e+\mathbf{j}_h \tag{19}$$

expresses the idea on which transport in multicarrier systems is based: the current density is found by adding the contributions from each group. In fields sufficiently weak, Ohm's law will be valid for each group, i.e.,

$$\mathbf{j}_e=\boldsymbol{\sigma}_e\cdot\mathbf{E},\quad \mathbf{j}_h=\boldsymbol{\sigma}_h\cdot\mathbf{E}. \tag{20}$$

This means that the conductivity is the sum of $\boldsymbol{\sigma}_e$ and $\boldsymbol{\sigma}_h$:

$$\boldsymbol{\sigma}=\boldsymbol{\sigma}_e+\boldsymbol{\sigma}_h\,. \tag{21}$$

Although the idea presented by these three equations is simple, it takes some algebra to carry out the calculations when a magnetic field is applied. Thus, whereas Eq. (15) could be obtained from Eq. (13) with no calculations at all, it is now necessary to evaluate $(\mathbf{1}+q\tau\boldsymbol{B}\cdot\boldsymbol{m}^{-1})^{-1}$ in Eq. (13) before the two parts $\boldsymbol{\sigma}_e$ and $\boldsymbol{\sigma}_h$ can be added together, and finally the inverse of their sum, i.e., $\boldsymbol{\rho}$, can be found. Only some results for the case of isotropic mass and z axis along **B** will be presented.

The inversion that appears on the right-hand side of Eq. (13) gives

$$\sigma_{xx}=\sigma_{yy}=\frac{\sigma_0}{1+(q\tau/m)^2B^2}\,, \tag{22}$$

$$\sigma_{zz}=\sigma_0\,, \tag{23}$$

$$\sigma_{xy}=-\sigma_{yx}=\frac{\sigma_0}{1+(q\tau/m)^2B^2}\,\frac{q\tau}{m}\,B, \tag{24}$$

$$\sigma_{xz}=\sigma_{zx}=\sigma_{yz}=\sigma_{zy}=0. \tag{25}$$

When electrons as well as holes are present, every element of $\boldsymbol{\sigma}$ becomes a sum of an electron and a hole contribution. As $\boldsymbol{\sigma}$ gets the symmetry of $\boldsymbol{\sigma}_e$ and $\boldsymbol{\sigma}_h$, its inverse can be found from the relations

$$\rho_{xx}=\rho_{yy}=\sigma_{xx}/[(\sigma_{xx})^2+(\sigma_{xy})^2], \tag{26}$$

$$\rho_{zz}=1/\sigma_{zz}\,, \tag{27}$$

$$\rho_{yx}=-\rho_{xy}=\sigma_{xy}/[(\sigma_{xx})^2+(\sigma_{xy})^2], \tag{28}$$

$$\rho_{xz}=\rho_{zx}=\rho_{yz}=\rho_{zy}=0. \tag{29}$$

The resulting resistivity tensor is most conveniently considered in very weak or very strong magnetic fields, where "weak" and "strong" refer to μB being smaller or larger than one.

To zeroth order in μB, i.e., with no magnetic flux density applied, $\boldsymbol{\rho}$ simply becomes a scalar:

$$\boldsymbol{\rho}=(\sigma_{0e}+\sigma_{0h})^{-1}\mathbf{1}\quad\text{for}\quad B=0. \tag{30}$$

To first order in μB the result is

$$\boldsymbol{\rho}=(\sigma_{0e}+\sigma_{0h})^{-1}\begin{pmatrix}1 & -\tilde{\mu}B & 0\\ \tilde{\mu}B & 1 & 0\\ 0 & 0 & 1\end{pmatrix}, \tag{31}$$

where $\tilde{\mu}$ is short for

$$\tilde{\mu}=\frac{\sigma_{0h}\,\mu_h-\sigma_{0e}\,\mu_e}{\sigma_{0e}+\sigma_{0h}}\,. \tag{32}$$

Magnetoresistance only appears when second-order terms in μB are included. In this case,

$$\rho_{xx}=\rho_{yy}=\frac{1}{\sigma_{0e}+\sigma_{0h}}\left(1+\frac{\sigma_{0e}\sigma_{0h}(\mu_e+\mu_h)^2}{(\sigma_{0e}+\sigma_{0h})^2}\,B^2\right), \tag{33}$$

$$\rho_{zz}=1/(\sigma_{0e}+\sigma_{0h}), \tag{34}$$

$$\rho_{yx}=-\rho_{xy}=\frac{\sigma_{0h}\,\mu_h-\sigma_{0e}\,\mu_e}{(\sigma_{0e}+\sigma_{0h})^2}\,B, \tag{35}$$

$$\rho_{xz}=\rho_{zx}=\rho_{yz}=\rho_{zy}=0. \tag{36}$$

For very strong magnetic fields, expansions in $(\mu B)^{-1}$ and $(\mu B)^{-2}$ are made. This gives

$$\rho_{xx}=\rho_{yy}=\frac{(n_e/\mu_e)+(n_h/\mu_h)}{e(n_h-n_e)^2}\quad\text{for}\quad \mu B\gg 1, \tag{37}$$

$$\rho_{zz}=1/(\sigma_{0e}+\sigma_{0h}), \tag{38}$$

$$\rho_{yx}=-\rho_{xy}=\frac{1}{e(n_h-n_e)}\,B, \tag{39}$$

$$\rho_{xz}=\rho_{zx}=\rho_{yz}=\rho_{zy}=0. \tag{40}$$

Thus the transverse magnetoresistance saturates while the Hall element becomes independent of the mobilities.

The appearance of n_h-n_e in the denominators of Eqs. (37) and (39) means that these equations do not apply for compensated materials, i.e., when the densities of electrons and holes are equal. In this case it can be shown that the transverse magnetoresistance increases as B^2 while the Hall element is still linear in B, but with a prefactor that depends on the mobilities.

As an example of the application of the multicarrier concept, Figs. 5(a) and 5(b) display respectively the Hall effect and the resistivity in p-type InSb. The experimental results

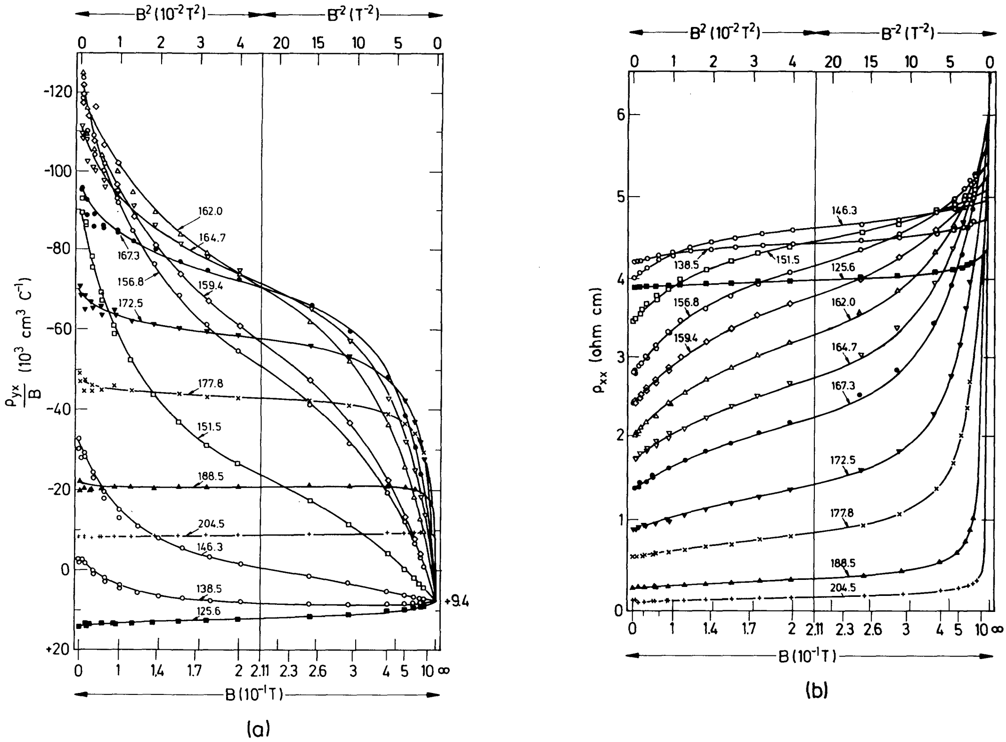

FIG. 5. (a) Hall coefficient and (b) magnetoresistance vs magnetic field for a single crystal of p-type InSb at various temperatures. The points display experimental values and the continuous curves display the results of calculations based on a three-carrier model (Blatt, 1968, p. 293).

were interpreted by a three-carrier model: electrons, heavy holes, and light holes. Extension of the two-carrier model to this case is straightforward. The six parameters, three densities and three mobilities, were determined such that the experimental points were reproduced as well as possible. Obviously the multicarrier formulas display sufficient complexity to allow for the fairly extensive field variations of Hall effect and magnetoresistance.

2.2.3 Metals. It might be expected that metals would be single-carrier systems, electrons being the only type of carrier. However, because of the complexity of the Fermi surface it turns out that transport can be mediated also by holes (See ELECTRON STRUCTURE OF SOLIDS). In such cases, the multicarrier concepts presented in the preceding section find applications also for metals.

As an example, the Fermi surface of aluminum is shown in Fig. 6. Out of the three valence electrons per Al atom, two fill up completely the first Brillouin zone. The one electron left is shared between the second and third zone. Figure 6(a) shows portions of the free-electron sphere when translated back into the first zone. The convex surface encloses holes, and, as each zone has two states per atom, the sum of electron and hole densities equals twice the density of Al atoms:

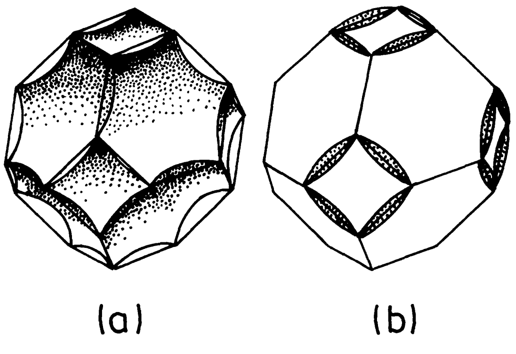

FIG. 6. (a) First Brillouin zone of Al with the parts of the free-electron sphere in the second zone translated back into the first. (b) As in (a) but now for the parts in the third zone (Ashcroft and Mermin, 1976, p. 301).

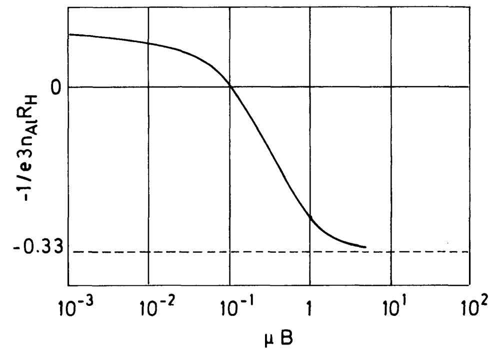

FIG. 7. Experimental values of Hall effect in polycrystalline Al (Ashcroft and Mermin, 1976, p. 15 and Lück, 1966). R_H denotes the measured Hall coefficient, and n_{Al} the density of aluminum atoms.

$$(n_e)^{II}+(n_h)^{II}=2n_{Al}\ . \qquad (41)$$

Figure 6(b) shows parts of the free-electron sphere in the third zone when translated back into the first zone. As only one electron per atom was left after filling the first zone, the electron densities in the second and third zone must sum up to the density of Al atoms:

$$(n_e)^{II}+(n_e)^{III}=n_{Al}\ . \qquad (42)$$

Thus according to Eqs. (41) and (42) the difference between hole and electron densities becomes

$$(n_h)^{II}-(n_e)^{III}=n_{Al}\ . \qquad (43)$$

Within the multicarrier concept this gives an expected strong-field limit of the Hall coefficient:

$$R_H=\frac{1}{e}\frac{1}{(n_h)^{II}-(n_e)^{III}}=\frac{1}{en_{Al}}\ . \qquad (44)$$

This is precisely what is observed; i.e., the strong-field Hall coefficient of polycrystals as well as single crystals has a value consistent with one positive carrier per Al atom; see Fig. 7.

In contrast to the strong-field limit of the Hall coefficient, the magnetoresistance has features that are not explainable in terms of the multicarrier concepts. Figure 8 displays results on single crystals of Al; the saturation in strong fields predicted by Eq. (37) is obviously not observed. Instead the magnetoresistance increases linearly, and the strength of this linear increase depends on the orientation of the magnetic field.

The horizontal variable in Fig. 8, B/ρ_0, is commonly used when magnetoresistance data are presented. With the simple expression $m/ne^2\tau$ for ρ_0 [cf. Eqs. (14) and (15)], B/ρ_0 becomes

$$B/\rho_0=\mu Bne. \qquad (45)$$

Materials for which the magnetoresistance depends on B only through B/ρ_0 are said to obey Kohler's rule.

The nonsaturation of magnetoresistance, exemplified above by the case of Al, is observed as well in other metals with closed

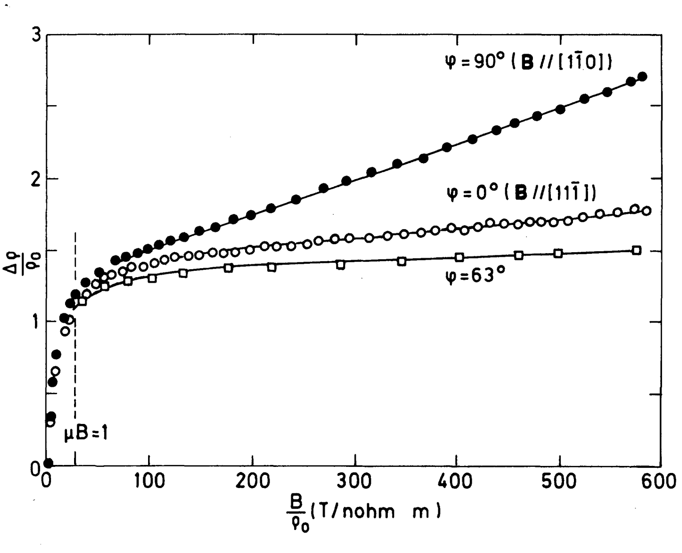

FIG. 8. Magnetoresistance of single-crystalline Al vs the Kohler variable B/ρ_0; $\rho_0=0.60$ nΩ cm (Kesternich *et al.*, 1975).

Fermi surfaces, such as In, Na, and K. Many proposals have been offered to explain this, among them magnetic breakdown (Kesternich *et al.*, 1975) and surface imperfections (Bruls *et al.*, 1981).

Aluminum belongs to a large group of metals having closed Fermi surfaces. However, metals with open Fermi surfaces exist, and in this case ranges of **B** directions can give rise to open carrier orbits. A typical example is the crimped cylinder, shown in Fig. 9. For this, open carrier orbits appear when **B** is perpendicular to the cylinder axis. The metals Cu, Ag, and Au belong to this category.

A theory based on the geometry of the Fermi surface was worked out by Lifshitz *et al.* (1957). The result was saturation of magnetoresistance for closed orbits and quadratic increase for open orbits. As the appearance of open orbits depends on the field direction, magnetoresistance for this case is expected to be strongly anisotropic.

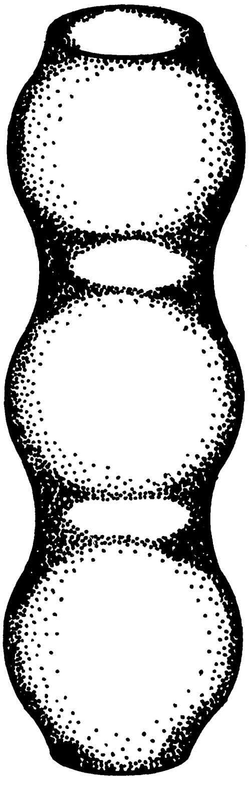

FIG. 9. Crimped cylinder (Lifshitz *et al.* 1957).

2.3 Magneto-optics: Cyclotron Frequency

The generalization of Eq. (10) to the case of an oscillating electric field, as in Eq. (5), is straightforward. The result for the conductivity becomes

$$\sigma=\sigma_0[(1-i2\pi f\tau)\mathbf{1}+(q\tau/m)\boldsymbol{B}]^{-1}. \qquad (46)$$

Free carriers, besides giving rise to the conductivity, also modify the applied field. This is especially so for metals, where the large density of free carriers screens out the electric field. Oppositely, for semiconductors with small densities of free carriers, the electric field inside the material can to a first approximation be taken as the vacuum field.

Thus for semiconductors the power absorption per unit volume can be approximated by

$$P_{\mathrm{av}}=\tfrac{1}{2}\,\mathrm{Re}(\mathbf{j}_0\cdot\mathbf{E}_0^*), \qquad (47)$$

where the subscript av means time average, Re means real part, the asterisk denotes complex conjugation, and $\mathbf{E}_0$ is the amplitude of the electric field as given by Eq. (5).

Figure 10 illustrates a common configuration: The electromagnetic wave propagates parallel to the static magnetic flux density and perpendicular to the sample surface (Faraday configuration). When the wave equation for this configuration is solved, it turns out that the solutions for **E** are in the form of a circularly polarized wave. With coordinates chosen as in Fig. 10 this means x and y components of the electric field related by

$$E_{0x}=\pm iE_{0y}, \qquad (48)$$

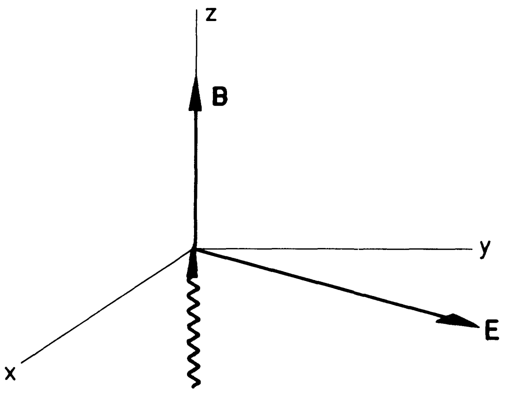

FIG. 10. Faraday configuration of a static magnetic flux density **B** and the electric field **E** of an electromagnetic wave propagating in the direction of **B**, which is perpendicular to the sample surface.

where the upper (lower) sign is for right (left) polarization. When used in Eq. (47), the power absorption of the two polarizations becomes

$$P_{\mathrm{av}}^{\pm}=(E_{0x})(E_{0x})^{*}\mathrm{Re}(\sigma_{xx}\mp i\sigma_{xy}). \tag{49}$$

The expressions (22) and (24) for σ_{xx} and σ_{xy} then give

$$P_{\mathrm{av}}^{\pm}=(E_{0x})(E_{0x})^{*}\frac{\sigma_0}{1+(2\pi f\pm 2\pi f_c)^2\tau^2}. \tag{50}$$

This has a maximum for either right or left polarization when the frequency equals the cyclotron resonance frequency f_c:

$$f_c=\pm(1/2\pi)(eB/m), \tag{51}$$

where the upper (lower) sign is for electrons (holes).

Cyclotron resonance was discovered in the semiconductor germanium in 1953 by Dresselhaus, Kip, and Kittel [see the review by Lax and Mavroides (1960)]. From measurements of the frequency at resonance, the carrier mass can be calculated using Eq. (51). The results of such measurements and calculations are displayed in Fig. 11 for the case of germanium. Three resonances were observed for each direction of the magnetic flux density. The resulting masses are seen to be strongly anisotropic.

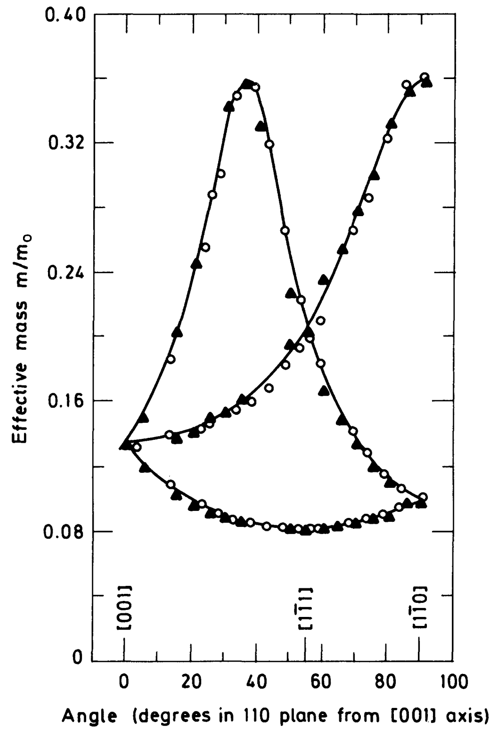

FIG. 11. Effective electron masses of germanium vs angle of magnetic flux density (Lax and Mavroides, 1960).

2.4 Thermal Conductivity, Fourier's Equation, and Thermal Diffusivity

The concept of heat current was illustrated in Fig. 1(c). The sample has a warm end and a cold end; i.e., a temperature gradient ∇T exists. The relation between ∇T and the heat current can be found by considering a particle traveling with velocity $\mathbf{v}_a$ from one region to another. To keep in thermal equilibrium with its immediate surroundings it energy has to change at a rate

$$\frac{\partial E}{\partial t}=c\mathbf{v}_a\cdot\nabla T, \tag{52}$$

c being the heat capacity per particle. As the particle moves a distance $v_a\tau$ without scattering, it contributes to the heat current by $v_a\tau\partial E/\partial t$. Taking the average over all directions of the velocity and summing over all particles, the heat-current density becomes

$$\mathbf{u}=-\tfrac{1}{3}Cv_a^2\tau\nabla T, \tag{53}$$

C being the heat capacity per volume.

The proportionality between heat-current density and the temperature gradient is often referred to as Fourier's equation, the constant of proportionality being the thermal conductivity κ:

$$\mathbf{u}=-\kappa\nabla T,\quad \kappa=\tfrac{1}{3}Cv_a^2\tau. \tag{54}$$

The diffusive nature of heat conduction is displayed when the time derivative of the energy is written as $C\partial T/\partial t$. Then energy conservation can be expressed by

$$C\frac{\partial T}{\partial t}+\nabla\cdot\mathbf{u}=0, \tag{55}$$

which together with Fourier's equation, leads to the diffusion equation for T:

$$\frac{\partial T}{\partial t}-\frac{\kappa}{C}\nabla^2 T=0. \tag{56}$$

The quantity κ/C is denoted thermal diffusivity. It is a main parameter for the description of heat pulse propagation in solids (*q.v.*).

2.5 Limitations

Although quite successful in describing a number of phenomena pertaining to electric and thermal conduction, the Drude model suffers from the lack of quantum-mechanical features. First, the particle quantum statistics is lacking, i.e., there is no way to include the Pauli principle for electrons. Also, the special quantization of electron orbits when a magnetic field is applied is absent, as is the whole concept of band structure of solids. The scattering should be given a quantum-mechanical description as well; however, a relaxation-time approximation can often be used.

3. WAVE FEATURES AND *k* SPACE

If scattering is described by a relaxation time, then the only nonclassical features that need to be introduced are the wave nature of moving particles and the Pauli exclusion principle.

The wave nature is expressed by the de Broglie relation:

$$k=2\pi/\lambda=mv/\hbar, \tag{57}$$

where k is the wave number, λ is the wavelength, and $\hbar$ is Planck's constant divided by 2π. Wave numbers are quantized according to the rule of periodic boundary conditions, i.e., along each of the three spatial dimensions, there should be an integer number of wavelengths. This rule selects discrete values of the wavelength:

$$\lambda\times\text{integer}=\text{spatial dimension}; \tag{58}$$

and according to (57) this means **k** values that are discrete and equally spaced.

This property of **k** values makes the **k** space a very useful concept. It is illustrated in Fig. 12(a). The discrete **k** values are denoted **k** states. According to Eqs. (57) and (58) they are spaced by $2\pi/L$; i.e., in the volume $(2\pi)^3/L_xL_yL_z$ there is just one **k** state, which means that the density of states in **k** space is $(2\pi)^{-3}$ per unit volume.

The Pauli principle is accounted for by the Fermi-Dirac distribution function—see Fig. 12(b)—which gives the probability of actually finding an electron in a state **k**:

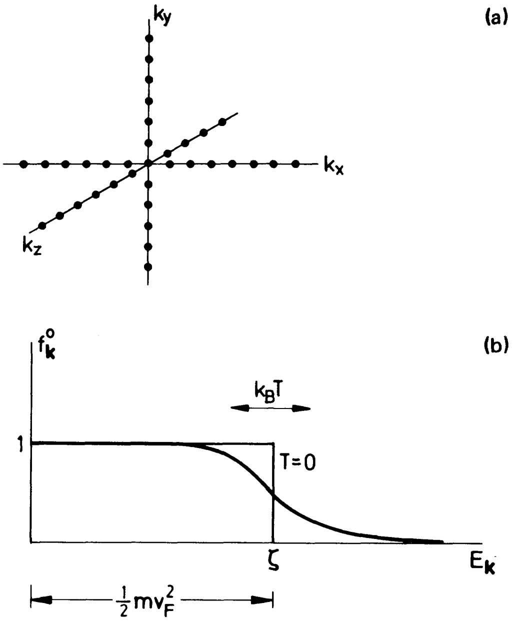

FIG. 12. Quantum-mechanical features. (a) Wave-number space, or **k** space. The spacings are $2\pi/L$, L being the (x,y,z)-edge lengths of the sample. (b) The Fermi-Dirac distribution function; ζ is the Fermi level.

$$f^0_{\mathbf{k}}=\{\exp[(E_{\mathbf{k}}-\zeta)/k_BT]+1\}^{-1}. \tag{59}$$

Here $E_{\mathbf{k}}$ is the energy of the state **k** and ζ is the energy at the Fermi level.

The wave vector **k** changes in time according to

$$\hbar\frac{d\mathbf{k}}{dt}=q(\mathbf{E}+\mathbf{v}\times\mathbf{B}). \tag{60}$$

As was the case for the Drude model, it is the Lorentz force that determines the dynamics of the system; cf. Eq. (9).

3.1 Metal-Insulator Transitions

The transition from a metallic, conducting state to an insulating state may be viewed in two ways: the starting point can be the metal or it can be the insulator.

Let the starting point be the metal, and let this be described by the mobility $\mu=e\tau/m$. As the mean free path $l=\tau v_F$ may be more easily visualized than τ, the mobility is written alternatively as $\mu=el/mv_F$. The question is then the following: Is there a lower limit of the mean free path below which the description breaks down? Ioffe and Regel (1960) have argued: Yes, if the mean free path, calculated from the ordinary mobility formulas, becomes less than the interatomic distances, then the mechanism of mobility itself must be altogether different, and the concept of a mean free path has no meaning. Thus, with symbols,

$$a \leqslant l. \tag{61}$$

which is known as the Ioffe-Regel rule.

If, on the contrary, the starting point is the insulating state, then the electrons are orbiting the atoms within a distance of the order of the Bohr radius

$$r_B = \frac{h^2 \epsilon_0}{m_0 \pi e^2} \frac{\epsilon}{m/m_0}. \tag{62}$$

In this case the question is: How close can neighboring atoms be before electrons start to become itinerant? An early estimate was made by Mott [see Mott (1982) for a review] of the critical value of the lattice constant a_c:

$$a_c \leqslant 4.5 r_B. \tag{63}$$

As an example of this, consider a semiconductor being doped with increasing concentration of donor atoms. A transition will then occur to a conducting metallic state. Experimental results for silicon doped with phosphorus is shown in Fig. 13. Here the metal-insulator transition takes place at a density $n_c = 3.74 \times 10^{18}$ cm^{-3}. If phosphorus is taken to be a monovalent donor in silicon, then a_c can be estimated as

$$a_c \sim n_c^{-1/3} = 0.64 \times 10^{-6} \text{ cm}.$$

The Bohr radius in silicon is approximately 20 Å, which gives $a_c \sim 3.2 r_B$, i.e., the order of magnitude expected from the Mott criterion.

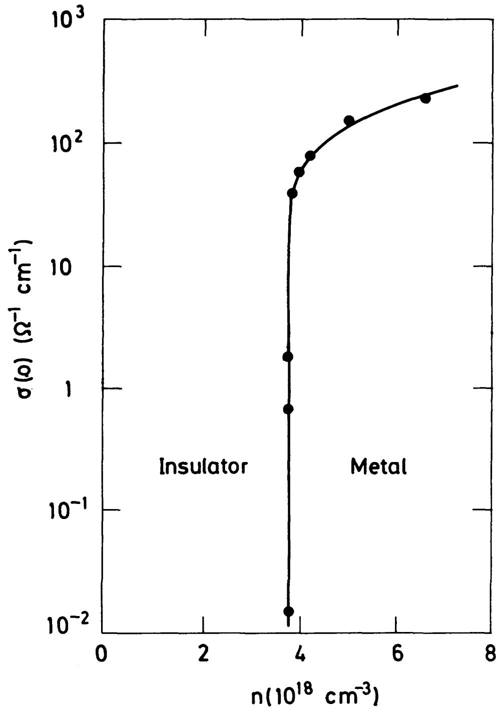

FIG. 13. Semilogarithmic plot of observed "zero temperature" conductivity $\sigma(0)$ vs donor concentration n for phosphorus donors in silicon (Kittel, 1986, p. 269).

Situations where either Eq. (61) or (63) is not fulfilled will not be treated in the present article. The reader may consult ELECTRON STATES, LOCALIZED and SEMICONDUCTORS, AMORPHOUS—ELECTRONIC PROPERTIES.

4. THE BOLTZMANN EQUATION

The Boltzmann equation describes the evolution in time of $f_\mathbf{k}$ as $\mathbf{k}$ changes according to Eq. (60) and the $\mathbf{k}$ states get scattered:

$$\frac{\partial f_\mathbf{k}}{\partial t} = \left(\frac{\partial f_\mathbf{k}}{\partial t}\right)_{\text{scatt}} - \frac{d\mathbf{k}}{dt} \cdot \nabla_\mathbf{k} f_\mathbf{k} - \frac{d\mathbf{r}}{dt} \cdot \nabla_\mathbf{r} f_\mathbf{k}, \tag{64}$$

where $(\partial f_\mathbf{k}/\partial t)_{\text{scatt}}$ is the rate of change of $f_\mathbf{k}$, due to scattering alone, and $d\mathbf{r}/dt$ is the carrier velocity. The inclusion of the last term, with the spatial gradient of $f_\mathbf{k}$, provides in one and the same equation a unified description of the effects of electric and magnetic fields and temperature gradients.

Various approximations have been devised to solve the Boltzmann equation. Let it be solved; then the current densities can be obtained as integrals over $\mathbf{k}$ states:

$$\mathbf{j} = \frac{2}{(2\pi)^3} \int q \mathbf{v}_\mathbf{k} f_\mathbf{k} d^3k, \tag{65}$$

$$\mathbf{v} = \frac{2}{(2\pi)^3} \int (E_\mathbf{k} - \zeta) \mathbf{v}_\mathbf{k} f_\mathbf{k} d^3k. \tag{66}$$

The factor of 2 in front of the integrals arises from the spin degeneracy of $\mathbf{k}$ states. The subtraction appearing in Eq. (66) accounts for the current of work function, as described in the Introduction.

4.1 Scattering of Charge Carriers

The scattering term in the Boltzmann equation describes the change in the occupation of the state **k**. This occupation increases when carriers in other states, **k**′, are scattered into **k**, and decreases with scattering out of **k**. The result can be written as an integral over the other states:

$$\left(\frac{\partial f_{\mathbf{k}}}{\partial t}\right)_{\text{scatt}}=\frac{V}{(2\pi)^3}\int [w(\mathbf{k},\mathbf{k}')f_{\mathbf{k}'}(1-f_{\mathbf{k}}) -w(\mathbf{k}',\mathbf{k})f_{\mathbf{k}}(1-f_{\mathbf{k}'})]d^3k, \qquad (67)$$

where V is the sample volume, $w(\mathbf{k},\mathbf{k}')$ is the transition probability per unit time from a filled state $\mathbf{k}'$ to an empty state $\mathbf{k}$, and the factors with $f_{\mathbf{k}'}$ and $1-f_{\mathbf{k}}$ account for the probability of finding the state $\mathbf{k}'$ occupied and the state $\mathbf{k}$ empty, respectively.

An immediate consequence of Eq. (67) is the principle of detailed balance: In equilibrium, with no fields and temperature gradients, the scattering term in the Boltzmann equation must vanish for every conceivable scattering process. This means that the integrand must vanish:

$$w(\mathbf{k},\mathbf{k}')f^0_{\mathbf{k}'}(1-f^0_{\mathbf{k}})-w(\mathbf{k}',\mathbf{k})f^0_{\mathbf{k}}(1-f^0_{\mathbf{k}'})=0, \qquad (68)$$

which by use of Eq. (59) leads to the principle of detailed balance:

$$\frac{w(\mathbf{k},\mathbf{k}')}{w(\mathbf{k}',\mathbf{k})}=\exp\left(\frac{E_{\mathbf{k}'}-E_{\mathbf{k}}}{k_BT}\right). \qquad (69)$$

For elastic scattering this means that $w(\mathbf{k},\mathbf{k}')=w(\mathbf{k}',\mathbf{k})$.

The complexity of the scattering integral can be avoided by simply writing the scattering term as

$$\left(\frac{\partial f_{\mathbf{k}}}{\partial t}\right)_{\text{scatt}}=-\frac{f_{\mathbf{k}}-f^0_{\mathbf{k}}}{\tau}. \qquad (70)$$

Whether or not this expression will follow strictly from Eq. (67) depends on the scattering process considered. When approximations have to be used, Eq. (70) is called the relaxation-time approximation.

4.1.1 Charged Impurities. When carriers are scattered against impurities, the large impurity mass makes the scattering elastic, and it turns out [Smith and Jensen (1989), p. 97] that for metals the relaxation rate can be written

$$\tau_{\text{imp}}^{-1}=\pi^3 n_{\text{imp}}v_F/k_F^2, \qquad (71)$$

where n_{imp} is the density of impurities, and where v_F and k_F are the Fermi velocity and the Fermi-sphere radius, respectively. The result in Eq. (71) is valid under the assumption that the impurity potential can be taken as a δ function. Through screening of the impurity charge by the mobile carriers—a positively charged impurity will attract a cloud of electrons—its bare Coulomb potential becomes modified by an exponential factor:

$$\phi_{\text{imp}}=-\frac{e}{4\pi\epsilon_0 r}e^{-k_s r}, \qquad (72)$$

where k_s^{-1} is the screening length, given by

$$k_s^2=\frac{e^2}{\epsilon_0}\int\left(-\frac{\partial f^0}{\partial E}\right)D(E)dE, \qquad (73)$$

$D(E)$ being the density of states.

The screening in metals is markedly different from screening in semiconductors. For metals the energy derivative of f^0 is a δ function and k_s^2 becomes

$$(k_{s,\text{TF}})^2=(e^2/\epsilon_0)D(E_F)=3e^2n/2\epsilon_0E_F\,. \qquad (74)$$

The quantity $(k_{s,\text{TF}})^{-1}$ is denoted as the Thomas-Fermi screening length. Contrary to this, for semiconductors $\partial f^0/\partial E=-f^0/kT$, and the integral of Eq. (73) leads to

$$(k_{s,\text{D}})^2=e^2n/\epsilon_0k_BT, \qquad (75)$$

where $(k_{s,\text{D}})^{-1}$ is denoted as the Debye screening length.

When, for metals, k_s is much larger than k_F, the impurity potential can be taken as a δ function, and Eq. (71) gives the resistivity due to impurity scattering:

$$\rho_{\text{imp}}=\frac{m}{ne^2\tau_{\text{imp}}}=\frac{\pi^2}{2}\frac{h}{e^2}\frac{1}{k_F}\frac{n_{\text{imp}}}{n} \qquad (76)$$

for impurities with a single charge.

4.1.2 Lattice Vibrations. Carriers are being scattered also against the vibrating ions of the solid. The energy of these vibrations is quantized in units of $\hbar\omega_{\mathbf{q}}$, so-called phonons, where $\omega_{\mathbf{q}}$ is the frequency of a vibration with wave vector $\mathbf{q}$. This type of scattering is principally inelastic, as the carrier absorbs or emits one or more phonons. For single-phonon processes energy conservation means

$$E_{\mathbf{k}'} = E_{\mathbf{k}} \pm \hbar\omega_{\mathbf{q}}. \tag{77}$$

This type of scattering dominates the transport phenomena at high temperature where phonons are many in number.

Several modes of lattice vibrations exist. In acoustic modes, known also as sound waves, the frequency is proportional to the wave number q:

$$\omega_{\mathbf{q}} = v_s q, \tag{78}$$

where v_s is the sound velocity. In optical modes the frequency is to a first approximation independent of **q**. Optical modes appear in crystals with two or more atoms per primitive basis and can be visualized as these atoms vibrating against each other. If the atoms in the basis carry a charge there will be an associated electric polarization wave, and the mode is denoted as polar optical. All these modes have transverse and longitudinal forms according to whether the motion of the atoms is at right angles or parallel to the direction of **q**. Below, only a few qualitative remarks pertinent to acoustic phonons will be made.

The density of phonons with wave vector **q** is given by

$$n_q = \frac{4\pi q^2}{\exp(\hbar v_s q/k_{\mathrm{B}}T) - 1}\,\frac{3}{(2\pi)^3}, \tag{79}$$

where the factor of 3 accounts for one longitudinal and two transverse modes. It has a maximum at $\hbar v_s q = 1.6 k_{\mathrm{B}}T$. In Debye's model an upper value of $\omega_{\mathbf{q}}$ is introduced by requiring that the total number of modes should equal the number of degrees of freedom:

$$\frac{3V}{(2\pi)^3}\int_0^{\omega_{\mathrm{D}}} 4\pi q^2 dq = 3N, \tag{80}$$

N being the number of atoms and ω_{D} the Debye frequency. For a cubic lattice $Na^3 = V$, a being the lattice constant, and Eq. (80) gives

$$\omega_{\mathrm{D}} = v_s q_{\mathrm{D}} = v_s(6\pi^2)^{1/3}/a \approx 1.2 v_s(\pi/a). \tag{81}$$

The Debye temperature T_{D} is defined by

$$k_{\mathrm{B}}T_{\mathrm{D}} = \hbar\omega_{\mathrm{D}}. \tag{82}$$

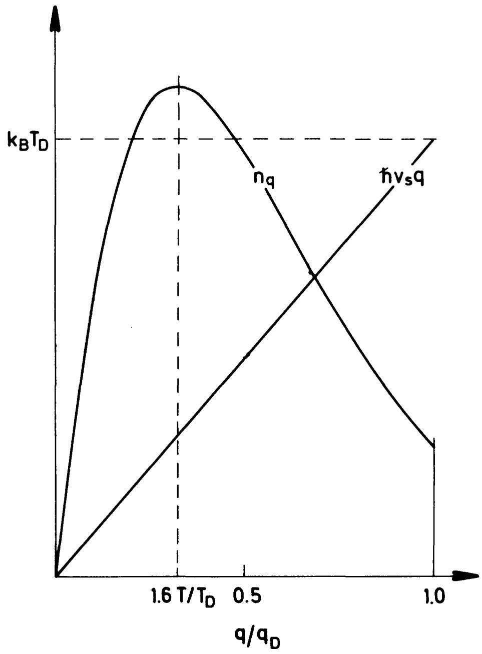

FIG. 14. Illustration of concepts in Debye model for phonons: Phonon density $n_{\mathbf{q}}$ (arbitrary units), according to Eq. (79) for $T/T_{\mathrm{D}} = 0.2$, and phonon energy $\hbar v_s q$, both vs reduced phonon wave vector q/q_{D}. The maximum in $n_{\mathbf{q}}$ occurs at a phonon energy of $1.6 k_{\mathrm{B}}T$.

The concepts introduced above are illustrated in Fig. 14. The density of phonons is

$$n_{\text{phonon}} = \int_0^{q_{\mathrm{D}}} n_q dq. \tag{83}$$

For $T \gg T_{\mathrm{D}}$, n_{phonon} increases linearly with T, while for $T \ll T_{\mathrm{D}}$ its temperature dependence is T^3.

Although electron-phonon scattering is principally inelastic, analysis of the scattering integrals shows that, provided $\hbar\omega_{\mathbf{q}} \ll |E_{\mathbf{k}} - \zeta|$, it can be treated as quasielastic.

In metals $|E_{\mathbf{k}} - \zeta| \approx k_{\mathrm{B}}T$, and as the phonon energy is less than $k_{\mathrm{B}}T_{\mathrm{D}}$, the requirement for quasielastic scattering becomes $T_{\mathrm{D}} \ll T$. Furthermore, for cubic metals with lattice constant a, and one conduction electron per atom, the Fermi-sphere radius is

$$k_{\mathrm{F}} = (3/\pi)^{1/3}(\pi/a) \approx 1.0(\pi/a) \tag{84}$$

and electron-phonon scattering can be illustrated as in Fig. 15. The geometry of momentum conservation gives

$$q \approx 2k_{\mathrm{F}} \sin \tfrac{1}{2}\theta, \tag{85}$$

and a relative change $\Delta k/k$ of the wave vector in the direction of **k**:

$$\Delta k/k \approx 1-\cos\theta = 2\sin^2(\tfrac{1}{2}\theta) = 2(q/2k_F)^2. \tag{86}$$

Above T_D the phonon wave vector $\mathbf{q}$ can be replaced by its maximum value q_D, which, along with the estimates of q_D and k_F in Eqs. (81) and (84), gives a $\Delta k/k$ of the order of 1. This means that a single electron-phonon scattering destroys completely the direction of **k**. For $T<T_D$ the situation is different. Here q may be replaced by its most probable value $1.6q_D(T/T_D)$, in which case $\Delta k/k \approx 1.8(T/T_D)^2$.

Besides being proportional to $\Delta k/k$, the scattering rate is further proportional to the phonon density and to the density of carrier states. In metals the density of carrier states will have to be taken at the Fermi level and is thus independent of temperature, whereas the phonon densities have the dependences T and T^3, as already explained. When put together, the result for the scattering rate becomes

$$\tau^{-1}_{\text{phonon,metal}} \sim \begin{cases} T & \text{for} \quad T>T_D\,, \\ T^5 & \text{for} \quad T<T_D\,. \end{cases} \tag{87}$$

Equation (87) gives the limiting behaviors of the so-called Grüneisen-Bloch law. Experimental data for a wide range of metals fit this remarkably well; cf. Fig. 16.

In semiconductors, the low carrier density means that small-angle scattering is of no relevance, and furthermore, that the electron density of states becomes temperature dependent as $T^{1/2}$. This leads to a scattering rate

$$\tau^{-1}_{\text{phonon,semicond.}} \sim \begin{cases} T^{3/2} & \text{for} \quad T>T_D\,, \\ T^{7/2} & \text{for} \quad T<T_D\,. \end{cases} \tag{88}$$

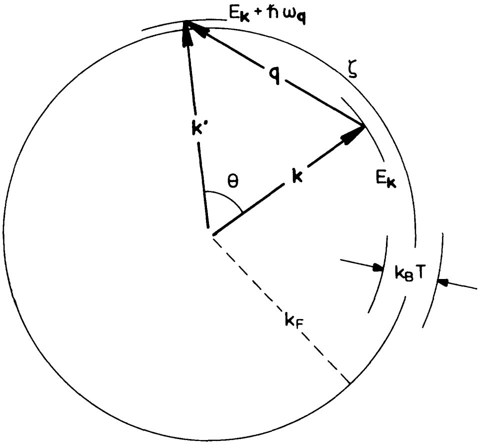

FIG. 15. Electron-phonon scattering: Electron with wave vector **k** and energy E_k absorbs a phonon with wave vector **q** and energy $\hbar\omega_q$. Fermi energy and Fermi-sphere radius are denoted ζ and k_F; θ is the scattering angle.

Figure 17 shows agreement between detailed calculations of mobility in germanium and experimental data gathered over two decades of years. Below 60 K the dependence is $T^{-3/2}$ in accordance with the high-temperature part of Eq. (88). The detailed derivation giving the $T^{-3/2}$ dependence is due to Bardeen and Shockley.

4.1.3 Mixed Scattering. When more than one scattering process is present, the combined effect is to a first approximation described by adding the scattering rates. Thus if the processes are those just considered, the total scattering rate is

$$\tau^{-1} = \tau^{-1}_{\text{imp}} + \tau^{-1}_{\text{phonon}}. \tag{89}$$

For metals, the impurity scattering rate is independent of temperature, and, as the phonon scattering vanishes at low temperatures,

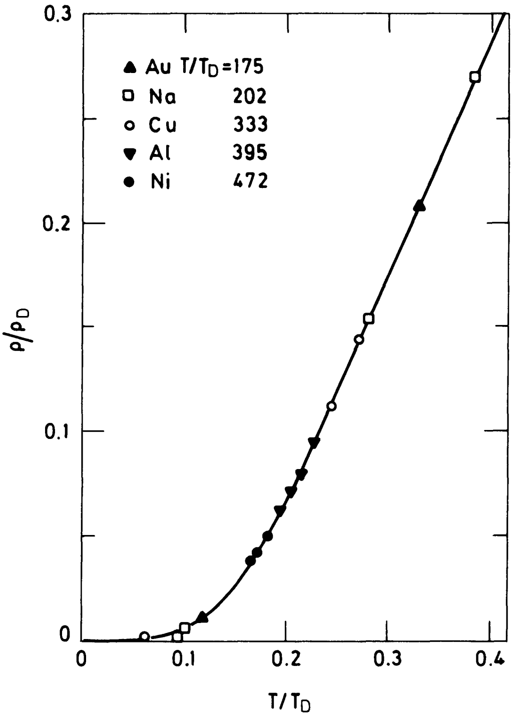

FIG. 16. The Grüneisen-Bloch law fitted to experimental data on the resistivity of metals, T_D and ρ_D being the Debye temperature and the resistivity at this temperature (Ziman, 1960, p. 366).

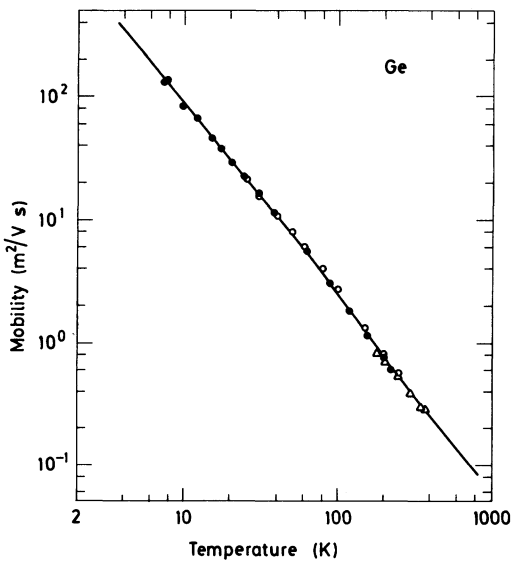

FIG. 17. Theoretical electron mobility of germanium (solid curve) and measured values: △, 1953; ○, 1954; ×, 1971; and ●, 1973 (Rode, 1975, p. 83).

Eq. (89) gives rise to a resistivity with a nonzero limiting value at low temperatures. This behavior is found experimentally for a large number of metals. It is known as Matthiessen's rule.

For the experimental data in Fig. 16 the measured resistivities were analyzed in accordance with Eq. (89), such that the resistivity ratio displayed pertains to the phonon scattering part only, and similarly for Fig. 17. But here, in addition to the scattering by acoustic phonons, scattering by nonpolar optical modes and scattering from one electron valley to another were considered. Willardson and Beer (1975) is recommended for further reading.

4.2 The Wiedemann-Franz Law

Within the relaxation-time approximation the Drude results (4) and (54) for the electric and thermal conductivities are retained—but with a different interpretation of the average velocity v_a: This should be the velocity v_F at the Fermi level. Furthermore, the expression for the heat capacity is not the one of classical particles, $C_{\text{classical}}=\frac{3}{2}k_B n$, but reflects that only the fraction $\sim k_B T/\frac{1}{2}mv_F^2$ of the electrons is taking part in thermal excitations. This gives rise to a heat capacity

$$C=\pi^2 n k_B^2 T/m v_F^2, \qquad (90)$$

and, according to Eqs. (4) and (54), a thermal conductivity

$$\kappa=(\pi^2/3)(k_B/e)^2 T\sigma. \qquad (91)$$

In the Drude model a similar result is obtained but with $\frac{3}{2}$ instead of $\pi^2/3$ as the numerical prefactor.

The independence on temperature of the ratio $\kappa/\sigma T$ is known as the Wiedemann-Franz law:

$$\kappa/\sigma T=\mathscr{L} \quad \text{independent of } T. \qquad (92)$$

From Eq. (91) the ratio is found to be

$$\begin{aligned}\kappa/\sigma T&=(\pi^2/3)(k_B/e)^2\\ &=\mathscr{L}_0\\ &=2.443\times10^{-8}\ \text{V}^2\ \text{K}^{-2},\end{aligned}$$

which is the so-called Lorenz number. Values of $\mathscr{L}$ at 0 and 100 °C for some metals are given in Table 1. Note that the numerical values are unambiguously in favor of the quantum description, the Drude value being 1.114×10^{-8} $(\text{V/K})^2$.

Table 1. Lorenz numbers for some metals (Mott and Jones, 1958, p. 307).

Metal	Lorenz number $\mathscr{L}$ [10^{-8} (V/K)^2] 0°	100°	Metal	Lorenz number $\mathscr{L}$ [10^{-8} (V/K^2)] 0°	100°
Mg	···	2.31	Pt	2.51	2.60
Al	···	2.23	Cu	2.23	2.33
Mo	2.61	2.79	Ag	2.31	2.37
W	3.04	3.20	Au	2.35	2.40
Fe	2.47	···	Zn	2.31	2.33
Rh	2.57	2.54	Cd	2.42	2.43
Ir	2.49	2.49	Sn	2.52	2.49
Ni	···	2.28	Pb	2.47	2.56
Pd	2.59	2.74	Bi	3.31	2.89

Figure 18 displays the thermal conductivity of Au at temperatures below 150 K. Three regions can be seen: a linear dependence on T at the lowest temperatures, a constant value above ~100 K, and a maximum at intermediate temperatures. This behavior can be qualitatively understood from the discussion of electron-impurity and electron-phonon scattering in Secs. 4.1.1 and 4.1.2: The Wiedemann-Franz law, Eq. (91), was obtained through cancellation of the relaxation time appearing in the expressions for σ and κ. However, when small-angle scattering is important, i.e., for $T<T_D$, these relaxation times are no longer identical. In other words, the relaxation-time approximation is not valid. Energy is more easily relaxed than wave number, and below T_D the Wiedemann-Franz law could not be expected to be valid.

At the lowest temperatures the impurity scattering becomes dominating; let it be below a temperature T_i. At these temperatures the relaxation-time approximation is again valid, the conductivity is independent of temperature, and the Wiedemann-Franz law is obeyed. The characteristic temperature dependences of σ, κ, and $\mathscr{L}$ for metals are displayed in Fig. 19.

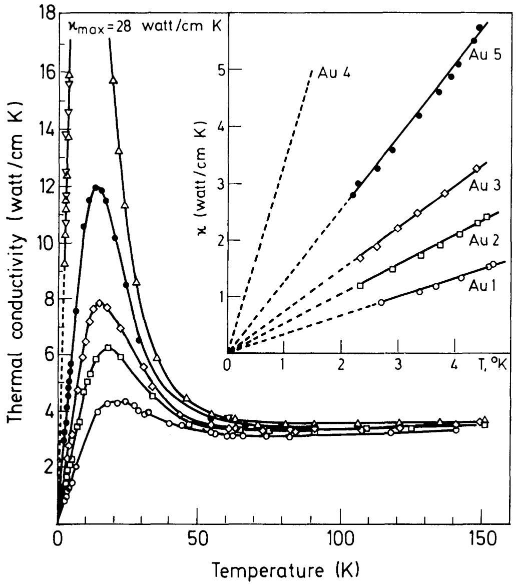

FIG. 18. Thermal conductivity of gold in varying states of physical and chemical purity, ○, Au 1, 99.9% pure, unannealed; □, Au 2, 99.9% pure, annealed; ◇, Au 3, JM3226, freshly drawn; △, Au 4, JM3226, annealed; ▽, Au 4, JM3226, annealed—second run; ●, Au 5, JM3226 redrawn (Olsen, 1962, p. 50).

5. LINEAR TRANSPORT EQUATIONS: PELTIER AND THERMOELECTRIC TENSORS

Despite the complicated scattering term in the Boltzmann equation, the expressions for the current densities will, to a first approximation, always be linear functions of the electric field and the temperature gradient:

$$\mathbf{j}=\mathbf{L}_{EE}\cdot\mathbf{E}+\mathbf{L}_{ET}\cdot\nabla T, \tag{93}$$

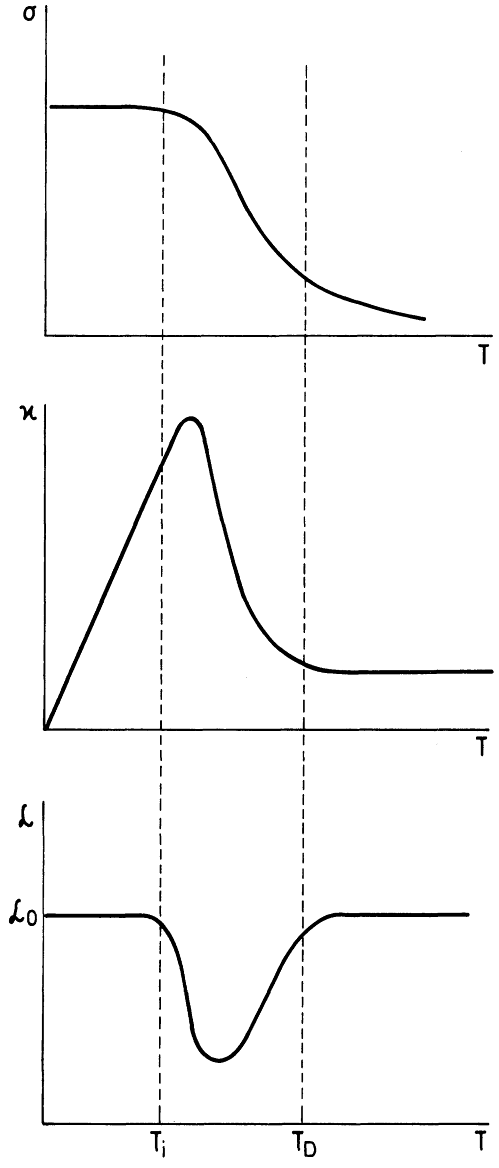

FIG. 19. Characteristic temperature dependences of the electric and thermal conductivities, and of $\mathscr{L}=\kappa/\sigma T$, for metals.

$$\mathbf{u}=\mathbf{L}_{TE}\cdot\mathbf{E}+\mathbf{L}_{ET}\cdot\nabla T. \tag{94}$$

The experimental conditions imposed in transport measurements are conveniently discussed with reference to these equations. These conditions are primarily the following two: no temperature gradient or no electric current.

When the temperature is everywhere the same, Eq. (93) is the basis for describing measurements of electric conductivity:

$$\boldsymbol{\sigma}=\mathbf{L}_{EE}\,. \tag{95}$$

Some details pertaining to the measurement of elements of $\boldsymbol{\sigma}$ and of its inverse, the resistivity $\boldsymbol{\rho}$, are given in Sec. 7. Equation (94) describes a flow of heat proportional to the electric field, or, by Eq. (93), to the electric current density:

$$\mathbf{u}=\mathbf{L}_{TE}\cdot\mathbf{L}_{EE}^{-1}\cdot\mathbf{j}=\boldsymbol{\Pi}\cdot\mathbf{j}; \tag{96}$$

The tensor in front of **j** is the Peltier tensor **Π**.

The other condition means that the sample should be electrically insulated. In this case, one end of the sample is clamped to a block which is maintained at a constant temperature, while a heater at the other end imposes a heat current through the sample. Equation (93) then tells that along with the temperature gradient, there will be an electric field:

$$\mathbf{E}=-\mathbf{L}_{EE}^{-1}\cdot\mathbf{L}_{ET}\cdot\nabla T. \tag{97}$$

The tensor in front of ∇T is the thermoelectric tensor **Q**:

$$\mathbf{Q}=-\mathbf{L}_{EE}^{-1}\cdot\mathbf{L}_{ET}\,. \tag{98}$$

The relation between the heat-current density and the temperature gradient is found by inserting the expression for **E**, Eq. (97), into Eq. (94):

$$\mathbf{u}=(\mathbf{L}_{TT}-\mathbf{L}_{TE}\cdot\mathbf{L}_{EE}^{-1}\cdot\mathbf{L}_{ET})\cdot\nabla T=-\boldsymbol{\kappa}\cdot\nabla T, \tag{99}$$

where $\boldsymbol{\kappa}$ is the thermal conductivity tensor.

6. SYMMETRY RESTRICTIONS

Two types of symmetry restrictions exist. One is linked to time reversal, the other one to spatial rotations. For a detailed understanding, the textbook by Smith *et al.* (1967), Chaps. 2 and 4, is recommended.

6.1 Onsager Relations

The Onsager relations deal with transport under time reversal. For the magnetic flux density—being produced by a circulating current—time reversal means a change of direction:

$$\mathbf{B}\rightarrow-\mathbf{B}\quad\text{for } t\rightarrow -t. \tag{100}$$

The Onsager relations tell how the transport tensors behave when the direction of **B** is changed. They are the following:

$$\mathbf{L}_{EE}(\mathbf{B})=\tilde{\mathbf{L}}_{EE}(-\mathbf{B}), \tag{101}$$

$$\mathbf{L}_{TE}(\mathbf{B})=-T\tilde{\mathbf{L}}_{ET}(-\mathbf{B}), \tag{102}$$

$$\mathbf{L}_{TT}(\mathbf{B})=\tilde{\mathbf{L}}_{TT}(-\mathbf{B}). \tag{103}$$

where the tilde means transposition. Applied to the transport tensors $\boldsymbol{\sigma}$, $\boldsymbol{\rho}$, $\boldsymbol{\kappa}$, **Π**, and **Q**, Eqs. (101)-(103) lead to

$$\boldsymbol{\sigma}(\mathbf{B})=\tilde{\boldsymbol{\sigma}}(-\mathbf{B}),\quad \boldsymbol{\rho}(\mathbf{B})=\tilde{\boldsymbol{\rho}}(-\mathbf{B}), \tag{104}$$

$$\boldsymbol{\kappa}(\mathbf{B})=\tilde{\boldsymbol{\kappa}}(-\mathbf{B}), \tag{105}$$

$$\boldsymbol{\Pi}(\mathbf{B})=T\tilde{\mathbf{Q}}(-\mathbf{B}). \tag{106}$$

6.2 Point-Group Symmetry

Suppose a crystal is placed in an apparatus that measures some macroscopic property. If a space-group symmetry operation is physically performed on the crystal but not on the apparatus, the atoms of the crystal will be brought back into their original positions with respect to the measuring apparatus, and there will be no change in the value of the quantity measured. The space-group translations associated with glide planes and screw axes involve atomic dimensions. They are therefore unobservable at the macroscopic level, and it is the point-group operations that give information about the symmetry of macroscopic crystal properties. The forms of the resistivity tensor and of the thermoelectric tensor for all 32 point groups can be found in the two papers by Akgöz and Saunders (1975). The effects of the Onsager relations are included in these papers as well.

7. SAMPLE GEOMETRIES

In the measurement of transport quantities a decisive role is played by the geometry of the sample and the way in which it is connected to the measuring circuit. As an example, consider the equation for the electric current density when the sample has everywhere the same temperature: $\mathbf{j}=\sigma\cdot\mathbf{E}$. The measurements can be performed in two ways: by control of the electric current density or by control of the electric field. Specific sample geometries are required for the two cases. They are named after their inventors: Hall (1880) and Corbino (1911). A mixed geometry, van der Pauw (1958/1959), mainly reflects technical convenience.

7.1 Hall Geometry

The Hall geometry is displayed in Fig. 20. The sample is in the form of an elongated, rectangular bar. The current enters and leaves the sample through the end faces, and its direction is along the long dimension. Let this be denoted by x. By this geometry the direction of the current is fixed:

$$\mathbf{j}=(j,0);\tag{107}$$

and the electric field is to be found from

$$\mathbf{E}=\rho\cdot\mathbf{j},\tag{108}$$

which means

$$E_x=\rho_{xx}j,\quad E_y=\rho_{yx}j.\tag{109}$$

These field components are measured between potential probes placed along the long dimension of the sample. The component E_y results from charges deflected by the field $\mathbf{B}$ and residing at the y surfaces. It is called the Hall field.

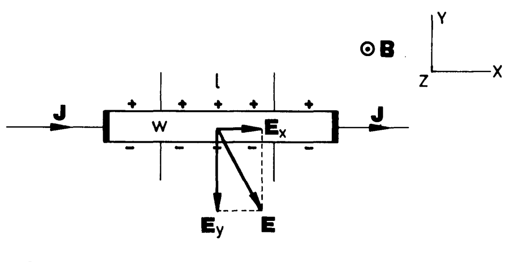

FIG. 20. Hall geometry. A magnetic flux density **B** is applied perpendicular to the paper plane. Spacings between potential probes are denoted *l* and *w*.

If only one type of charge carrier is present—let it be electrons of density n—then ρ_{yx} depends in a simple way on the density. The dependence can be found from the following considerations: In the steady state, the total force in the y direction is zero, meaning

$$\mathbf{E}_y=-\mathbf{v}\times\mathbf{B},\tag{110}$$

and the velocity $\mathbf{v}$ is related to the current density through

$$\mathbf{j}=-en\mathbf{v},\tag{111}$$

which along with (109) and (110) gives

$$\mathbf{E}_y=\frac{1}{ne}\,\mathbf{j}\times\mathbf{B}=R_{\mathrm{H}}\mathbf{B}\times\mathbf{j},\tag{112}$$

where

$$R_{\mathrm{H}}=-1/ne=\rho_{yx}/B\tag{113}$$

is called the Hall coefficient.

The components ρ_{xx} and ρ_{xy} of the resistivity tensor are found from measurements of the transverse and longitudinal voltages, E_yw and E_xl, respectively.

7.2 Corbino Geometry

Figure 21 illustrates an imaginary transformation from Hall geometry in Fig. 21(a) to Corbino geometry in Fig. 21(c): One of the current contacts is stretched until the two y faces meet, whereby this contact forms a ring, while the other one becomes a disk. The meeting of the two y faces means a short circuit of the Hall field, and this leaves the radial component of the electric field as the only component. Thus, the Corbino geometry fixes the direction of the electric field.

The current–field relation for the Corbino geometry can be expressed in cylindrical coordinates (r,θ) by

$$\mathbf{E}=(E,0),\tag{114}$$

$$\mathbf{j}=\sigma\cdot\mathbf{E},\tag{115}$$

from which

$$j_r=\sigma_{rr}E,\quad j_\theta=\sigma_{\theta r}E.\tag{116}$$

The potential difference V_{Corbino} between the disk and the ring is the radial integral

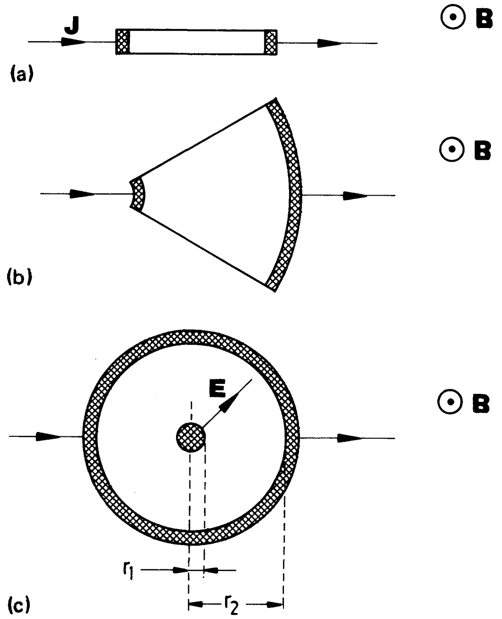

FIG. 21. Imaginary transformation from (a) Hall to (c) Corbino geometry.

$$V_{\text{Corbino}} = \int_{r_1}^{r_2} \mathbf{E}(r)\cdot d\mathbf{r}, \tag{117}$$

and the field $E(r)$ is determined by the requirement that, in the steady state, the current I entering the disk passes through every concentric circle:

$$2\pi r_1 j(r_1) = 2\pi r j(r) = I/t, \tag{118}$$

t being the thickness of the sample; or, by Eq. (116),

$$E(r) = j(r)/\sigma_{rr} = I/2\pi r t \sigma_{rr}\,. \tag{119}$$

Inserting this expression in the integral of Eq. (117) gives

$$V_{\text{Corbino}} = \frac{I}{2\pi t \sigma_{rr}} \ln \frac{r_2}{r_1} = R_{\text{Corbino}} I, \tag{120}$$

where

$$R_{\text{Corbino}} = \frac{I}{2\pi t \sigma_{rr}} \ln \frac{r_2}{r_1}\,. \tag{121}$$

Thus the Corbino geometry allows a genuine determination of the diagonal component σ_{rr}.

The current circulating in the Corbino sample, I_θ, is determined by both components σ_{rr} and $\sigma_{\theta r}$:

$$I_\theta = t \int_{r_1}^{r_2} j_\theta(r)\,dr = t \int_{r_1}^{r_2} \sigma_{\theta r} E(r)\,dr$$
$$= t\sigma_{\theta r} V_{\text{Corbino}} = \frac{I}{2\pi} \frac{\sigma_{\theta r}}{\sigma_{rr}} \ln \frac{r_2}{r_1}\,. \tag{122}$$

7.3 van der Pauw Method

The Hall and Corbino geometries are distinguished by parallel current streamlines and radial electric field, respectively. These features make the interpretation of the measured quantities easy, but are paid for by the difficulties in sample and contact preparation. In the van der Pauw method a slab of arbitrary shape, and with four point contacts along its edge, is used. This is sketched in Fig. 22. To find the resistance, two measurements have to be made: One of these is the potential difference V_{OP} between contacts P and O, using M and N as current contacts. The ratio between V_{OP} and the current is denoted $R_{MN,OP}$:

$$R_{MN,OP} = V_{OP}/I_{MN}\,. \tag{123}$$

The other one is the potential difference V_{PM}, using N and O as current contacts. This gives

$$R_{NO,PM} = V_{PM}/I_{NO}\,. \tag{124}$$

The square resistance $R_\square$ is then found from

$$R_\square = \frac{\pi}{\ln 2} \frac{R_{MN,OP} + R_{NO,PM}}{2} f, \tag{125}$$

where f is a geometrical factor of the order 1. The square resistance is the resistivity divided

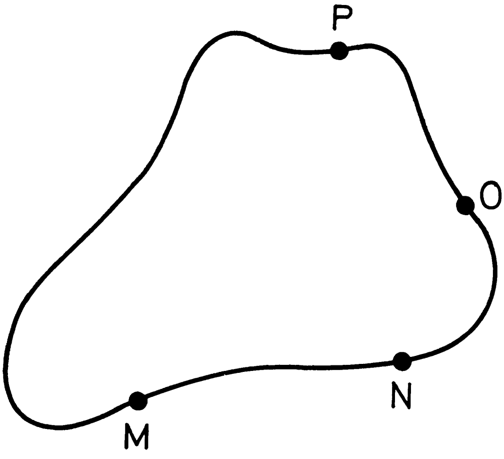

FIG. 22. van der Pauw geometry: slab of arbitrary shape with four contacts along its edge.

by the thickness of the slab. The Hall effect may be measured by this method as well. For the derivation of Eq. (125) the original paper of van der Pauw (1958/59) should be consulted.

8. LANDAU QUANTIZATION

In the previous sections the influence of quantum mechanics on electronic transport has been treated at the very simplest level: the de Broglie relation and the Pauli exclusion principle (Fermi-Dirac distribution function). This type of treatment does not account for a number of oscillatory phenomena observed when an electronic sample is placed in a strong magnetic field at low temperatures.

The theory as well as the experimental observations appeared in 1930. The Leningrad physicist L. D. Landau worked out the theory of diamagnetism in metals while visiting Cambridge University. At the same time at the University of Leiden, de Haas and van Alphen observed oscillatory susceptibility, and Shubnikov and de Haas observed oscillatory magnetoresistance.

The Landau quantization appears in the plane perpendicular to an applied magnetic flux density. It restricts the **k**-space areas to the following values:

$$A_{\mathbf{k}\perp}=(p+\tfrac{1}{2})(2\pi/\hbar)eB, \quad p=0,1,2,\ldots . \qquad (126)$$

The period of oscillatory phenomena is found from this expression by equating $A_{\mathbf{k}\perp}$ to the **k**-space area actually occupied by electrons. Thereby, a series of equidistant values of the inverse field, B^{-1}, appears, their spacing being the period $P(B^{-1})$:

$$P(B^{-1})=\frac{2\pi}{\hbar}\,e\,\frac{1}{A_{\mathbf{k}\perp,\text{occupied}}}\,. \qquad (127)$$

In the simplest case the **k**-space volume occupied by electrons has the shape of a sphere with radius k_F. The area $A_{\mathbf{k}\perp}$ and the carrier density is then expressed as

$$A_{\mathbf{k}\perp,\text{occupied}}=\pi k_\text{F}^2 \qquad (128)$$

and

$$n=2\,\frac{1}{(2\pi)^3}\,\frac{4}{3}\,\pi k_\text{F}^3=\frac{1}{3\pi^2}\,k_\text{F}^3\,. \qquad (129)$$

From measured periods $P(B^{-1})$ the carrier density n can be calculated, as from Eqs. (127)–(129):

$$P(B^{-1})=(2e/\hbar)(3\pi^2 n)^{-2/3}. \qquad (130)$$

The review of Landwehr and Rashba (1991) is recommended for further reading.

8.1 Shubnikov–de Haas Oscillations

The oscillatory period given by Eq. (127) is observed in a variety of solid-state phenomena. In electronic transport its most conspicuous effects show up in the magnetoresistance. An example is shown in Fig. 23.

9. TRANSPORT IN LOW-DIMENSIONAL SYSTEMS

The quantization of wave numbers was mentioned in Sec. 3: The wave-number spacings are $2\pi/L$, L being the spatial dimension along (x,y,z). If one of these dimensions is reduced—let it be z—then the k_z spacing increases. As **k** states are occupied in accordance with their energy, which depends on **k** as k^2, the large k_z values mean that ultimately only states in one (k_x,k_y) plane, the one having the smallest k_z, are occupied. This is sketched in Fig. 24(a). The occupied area in **k** space, which in the three-dimensional sample was a sphere, now becomes a disk. The system is said to be two-dimensional. The density of **k** states in the plane is $(2\pi)^{-2}$ per unit area.

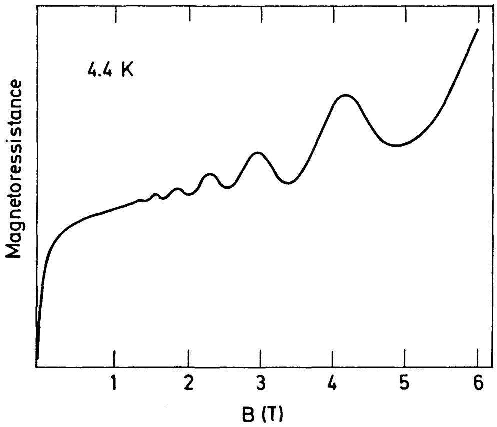

FIG. 23. Magnetoresistance of bismuth doped slightly with Sn.

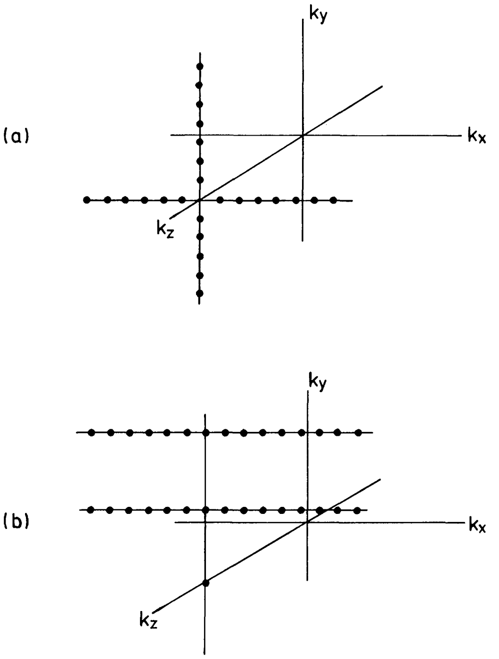

FIG. 24. **k** space for (a) a two-dimensional system and (b) a one-dimensional system.

If, in the two-dimensional system, one of the spatial dimensions is reduced—let it be y—then the k_y spacings increase, and the occupied **k** states will be on k_x lines. The system is then said to be one-dimensional. [See Fig. 24(b)].

The improvement within the last two decades of crystal-growing techniques to the limit where growth can be controlled at the level of atomic layers has opened up the possibility for studying a large variety of two-dimensional systems. Along with this, the techniques for making patterns in the two-dimensional systems have given access to one-dimensional systems as well.

9.1 The Quantum Hall Effect

Within the two-dimensional systems the discovery by von Klitzing in 1980 of the quantum Hall effect was a most remarkable event, for which the Nobel prize was awarded in 1985. An example of quantized Hall resistance is shown in Fig. 25 along with Shubnikov–de Haas oscillations. Instead of the linear dependence of Hall voltage on B, expected from Eqs. (16) and (109), a series of steps appears. The values of these transverse resistance steps can be obtained by the following considerations: With the notation used in Fig. 20 the Hall voltage V_H is written as

$$V_H = E_y w = \rho_{yx} j w = (B/ne) j w. \qquad (131)$$

The Landau levels are circles in the (k_x, k_y) plane in Fig. 24(a). The number of **k** states in each of these circles equals the area $(2\pi/\hbar)eB$ between them [cf. Eq. (126)], multiplied by the density of states $(2\pi)^{-2}$ per unit area, i.e., $(e/h)B$. Thus if ν circles are filled up with electrons, there will be $\nu(e/h)B$ electrons per unit area. On multiplying numerator and denominator in Eq. (131) by the thickness t of the two-dimensional layer, nt becomes the density per unit area and jwt becomes the current I, and so

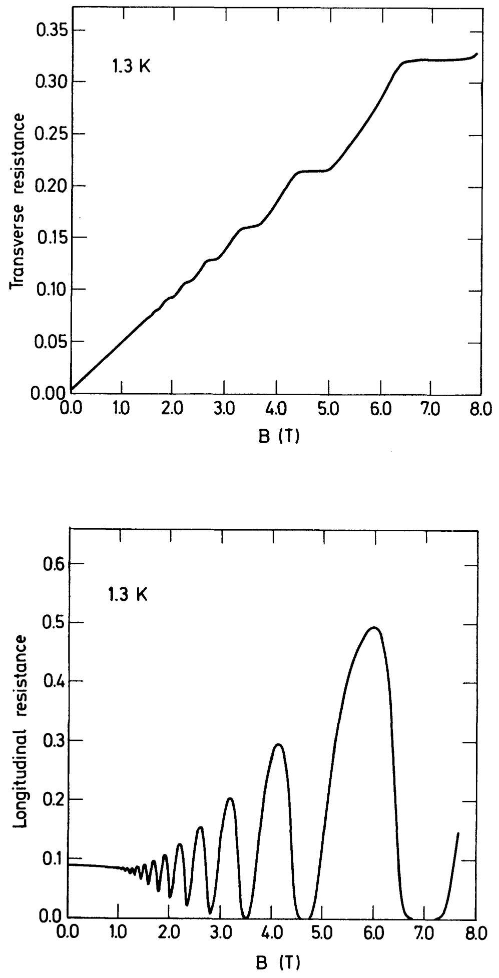

FIG. 25. The transverse and longitudinal resistances of a GaAs/GaAlAs heterostructure (two-dimensional electron system).

$$V_{\mathrm{H}}=\frac{1}{\nu}\frac{h}{e^2}I. \tag{132}$$

Measurement of the transverse resistances at the plateaus thus gives access to the important combination of fundamental constants, h/e^2. The values are so accurately given by Eq. (132) that in 1988 the Comité International des Poids et Mesures recommended that from 1 January 1990 the ohm should be represented by means of the quantum Hall effect, assigning the value 25 812.807 Ω to the von Klitzing constant, that is to say, to the quotient of the Hall voltage divided by the current for the plateau $\nu=1$ (see CONSTANTS, FUNDAMENTAL).

The number ν is called the filling factor because it tells how many Landau levels are filled. The very intense research that followed the discovery in 1980 led in 1982 to the observation of a step also at $\nu=\frac{1}{3}$, and soon after at a number of fractional fillings. After this, the names integer quantum Hall effect (IQHE) and fractional quantum Hall effect (FQHE) were used for the two phenomena. They have been and are still a challenge to theory. For further reading the following is recommended: QUANTUM HALL EFFECT, the book *The Quantum Hall Effect,* edited by Prange and Girvin (1987), and the review by MacDonald (1992).

9.2 Size-Quantized Conductivity

Figure 26 shows the conductance of a so-called point contact, which is an example of a one-dimensional constriction in an otherwise two-dimensional system. The quantization may be qualitatively understood as follows: For each of the k_x lines in Fig. 24(b) the current is written as an integral over the energy:

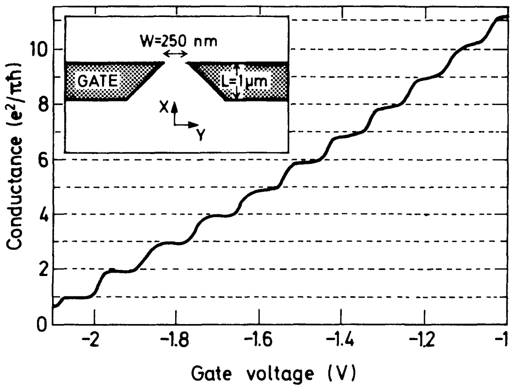

FIG. 26. Point-contact conductance as a function of gate voltage. Temperature: 0.6 K. Inset: Point-contact layout (van Wees *et al.*, 1988).

$$I=e\int_0^{eV}\nu\frac{dn}{dE}\,dE, \tag{133}$$

where V, in the upper limit of the integral, is the voltage across the point contact. The way in which one-dimensionality is distinguished is by its density of states per unit energy:

$$\frac{dn}{dE}=\frac{dn}{dk}\frac{dk}{dE}. \tag{134}$$

Only for the one-dimensional case is dn/dk a constant, $(2\pi)^{-1}$. For the velocity the general $\nu=\hbar^{-1}(dE/dk)$ is used, and the following expression for I results (the factor of two accounts for the spin degeneracy):

$$I=(2e^2/h)V. \tag{135}$$

The steps in Fig. 26 arise when the size of the constriction is changed. Such change is accomplished by the horizontal variable of Fig. 26, the gate voltage. The gate is the hatched area in the inset of Fig. 26. Applying a negative voltage to this increases the energy of the two-dimensional electron gas underneath, and thereby creates a saddle-point potential for the electrons. The gate voltage determines the width of the "saddle," and this in turn determines the number of k_x lines by which electrons can cross the constriction. Changing this number by 1 creates a step in Fig. 26.

The two phenomena considered in this last section are just examples, very important ones though, of electronic transport in solids in which the Drude assumption, that all carriers act independently, is no longer valid. This invalidation has been brought about by improved crystal-growing and lithographic techniques. In the small spatial dimensions thus created, electron waves can keep coherence, and as a result interference effects appear. The word mesoscopic, indicating a range between micro and macro, is being used to designate this whole field. (See MESOSCOPIC SYSTEMS.)

List of Works Cited

Akgöz, Y. C., Saunders, G. A. (1975), *J. Phys. C* **8**, 1387–1396; **8**, 2962–2970.

Ashcroft, N. W., Mermin, N. D. (1976), *Solid State Physics*, New York: Holt, Rinehart and Winston.

Blatt, F. J. (1968), *Physics of Electronic Conduction in Solids*, New York: McGraw-Hill.

Bruls, G. J. C. L., Bass, J., van Gelder, A. P., van Kempen, H., Wyder, P. (1981), *Phys. Rev. Lett.* **46**, 553–555.

Corbino, O. M. (1911), *Phys. Z.* **12**, 561–568.

Hall, E. H. (1880), *Am. J. Sci. Arts*, 3rd ser., **XIX**, 200–205.

Ioffe, A. F., Regel, A. R. (1960), in A. F. Gibson, F. A. Kröger, R. E. Burgess (Eds.), *Progress in Semiconductors*, Vol. 4. London: Heywood, pp. 237–291.

Kesternich, W., Ullmaier, H., Schilling, W. (1975), *Philos. Mag.* **31**, 471–488.

Kittel, C. (1986), *Introduction to Solid State Physics*, New York: Wiley.

Landwehr, G., Rashba, E. I. (Eds.) (1991), *Landau Level Spectroscopy, Modern Problems in Condensed Matter Sciences*, Vols. 27.1 and 27.2, Amsterdam: North-Holland.

Lax, B., Mavroides, J. G. (1960), in F. Seitz, D. Turnbull (Eds.), *Solid State Physics: Advances in Research and Applications*, Vol. 11, New York: Academic, pp. 201–400.

Lifshitz, I. M., Azbel', M. Ya., Kaganov, M. I. (1957), *Sov. Phys. JETP* **4**, 41–55. See also the review by Lifshitz, I. M., Kaganov, M. I., "Some Problems of the Electron Theory of Metals," *Sov. Phys. Usp.* **2**, 831–855 (1960); **5**, 878–907 (1963); **8**, 805–851 (1966).

Lück, R. (1966), *Phys. Status Solidi* **18**, 49–57.

MacDonald, A. H. (1992), *Helv. Phys. Acta.* **65**, 133–142.

Mott, N. F. (1982), *Proc. R. Soc. London Ser.* A **382**, 1–24.

Mott, N. F., Jones, H. (1958), *The Theory of the Properties of Metals and Alloys*, New York: Dover.

Olsen, J. L. (1962), *Electron Transport in Metals*, New York: Wiley-Interscience.

Prange, E. Richard, Girvin, Steven M. (Eds.) (1987), *The Quantum Hall Effect*, New York: Springer.

Rode, D. L. (1975), in: R. K. Willardson, Albert C. Beer (Eds.), *Semiconductors and Semimetals*, Vol. 10. New York: Academic, pp. 1–89.

Smith, A. C., Janak, J. F., Adler, R. B. (1967), *Electronic Conduction in Solids*, New York: McGraw-Hill.

Smith, H., Jensen, H. Høgaard (1989), *Transport Phenomena*, Oxford: Oxford University Press.

van der Pauw, L. J. (1958/59) *Philips Tech. Rev.* **20**, 220-224.

van Wees, B. J., van Houten, H., Beenakker, C. W. J., Williamsson, J. G., Kouwenhoven, L. P., van der Marel, D., Foxon, C. T. (1988), *Phys. Rev. Lett.* **60**, 848–850.

Willardson, R. K., Beer, Albert C. (Eds.) (1975), *Semiconductors and Semimetals*, Vol. 10, New York: Academic.

Ziman, J. M. (1960), *Electrons and Phonons—The Theory of Transport Phenomena in Solids*, Oxford: Oxford University Press.

ELECTRORHEOLOGY

C. F. ZUKOSKI, *Department of Chemical Engineering, University of Illinois, Urbana, Illinois, U.S.A.*

Introduction 145
1. **The Electrorheological Response** 146
2. **Origins of the ER Response** 149
3. **Applications of ER Devices** ... 152
Works Cited 153

INTRODUCTION

The electrorheological (ER) response is characterized by a large change in suspension mechanical properties upon exposure to electric fields. Viscosity changes of several orders of magnitude commonly occur. Suspensions displaying an ER response are typically composed of polar solids suspended in low–dielectric-constant oils, and, as a result, have low conductivities. As a consequence, field strengths of 0.5–5 kV/mm employed to achieve large responses do not result in prohibitive power consumption, opening the possibility for a wide range of technologies. While the ER response has been known for over 40 years, a recent resurgence of interest can be traced to the availability of control instrumentation made possible by solid state electronics. The ability to rapidly alter and continuously vary stress transfer properties plays an important role in applications of the ER response in devices such as dampers, shock absorbers, clutches, and actuators. Despite the length of time the response has been known, much remains poorly understood, and the study of ER systems is very active from both a fundamental and a technological perspective. Here a review of current (up to the beginning of 1991) understanding of the ER response will be provided, and directions that future research and development are likely to take are discussed.

The ER response was discovered by Winslow in 1947 (Winslow, 1947, 1949) and patents were issued at that time on applications involving stress-transfer devices. Winslow identified much of the underlying physics giving rise to the ER response and investigated a number of fluid formulations. Klass and Martinek (1967) and Uejima (1972) provided the most comprehensive rheological and suspension characterization of ER fluids published until late in the 1980s. Klass and Martinek questioned Winslow's contention that the ER response was due to interactions between particles produced by differences in polarizability of the solid and continuous phases and, on the basis of the short response times, suggested that the ER mechanism involved polarization of electric double layers. Uejima, working with suspensions of microcrystalline cellulose, demonstrated that dramatic increases in response can follow from the addition of small quantities of water to ER suspensions. The effects of water are still poorly understood and continue to plague commercialization of ER technologies. In the 1970s an active program investigating electrorheology was established in the Soviet Union, where a variety of formulations were explored and careful rheological characterization of ER fluids was initiated (Shulman *et al.*, 1981). From the mid 1960s to the late 1970s, Mason and co-workers at McGill University supplied some of the first fundamental data on particle interactions in the presence of shear and electric fields and demonstrated that rates of structure formation and degradation could be determined from a summation of electrically induced polarization forces developed between particles in the presence of an electric field and the viscous drag provided by the suspending solvent (Zia *et al.*, 1975).

What might be called a modern era of ER research and development was initiated in the late 1970s in the U.K., culminating in the publications of Stangroom and co-workers,

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

who developed polymer-based suspensions displaying large ER responses (Stangroom, 1978, 1983). This work stimulated automotive industry interest in developing active suspensions and engine mounts. Since the early 1980s research programs in industrial and academic laboratories have begun to elucidate the fundamental character of the ER response. These efforts have spanned a broad spectrum of interests ranging from device design to physical properties of polarized suspensions. A poor understanding of the links between field strength, shear rate, current density, and material properties contribute to the limited applications of ER technologies. In this article, the effects of large electric fields on nonaqueous suspensions will be described from a phenomenological perspective in Sec. 1 followed in Sec. 2 with a development of the polarization model of the ER response. In recent years, novel applications of the ER response have been described, and these will be discussed in Sec. 3. Reviews discussing the ER response have been written by Block and Kelly (1988), Gast and Zukoski (1989), and Jordan and Shaw (1989).

Table 1. Characteristics of electrorheological suspensions.

Suspending Fluid Properties	
Relative permittivity	$\epsilon_c=2-15$
Low-field conductivity	$\sigma_c=10^{-7}$–10^{-13} s/m
Viscosity	$\eta_c=0.01$–10 Pa s
Density	$\rho_c=0.6$–2 g/cm^3
Particle Properties	
Shape	Roughly spherical
Size	$\sigma=0.1$–100 μm
Relative permittivity	2–60
Suspension Properties	
Zero-field viscosity	$\eta_\infty=0.1$–10 Pa s
Volume fraction	$\phi=0.05$–0.50
Conductivity	$\Sigma=10^{-6}$–10^{-13} s/m
Typical field strengths	$E=0.5$–5 kV/mm

1. THE ELECTRORHEOLOGICAL RESPONSE

Suspensions displaying large electrorheological responses are typically composed of a polar solid suspended in a low–dielectric-constant oil. Continuous phases of chlorinated hydrocarbons and silicon, mineral, and vegetable oils have been reported. Properties of importance in the continuous phase are low conductivity, high dielectric breakdown voltage, and ability to disperse the solid phases. Typical physical properties are given in Table 1.

The dispersed phase is composed of particles in the 0.1 to 100-μm-diam range with polarizabilities much larger than that of the continuous phase. For example, when containing 0.5–4 wt% water (based on solid-phase mass), suspensions of microcrystalline cellulose and wheat flour produce very effective ER suspensions. On the other hand, if the solid phase is dried at 80°–100 °C for several hours, suspensions of the same materials have small ER responses. The change in activity is associated with increases in polarizability produced by hydrating the solid phase. Besides starches and cellulosics, polar solids such as titania, barium titanate, charge-transfer salts, conducting polymers, hydrated polymeric systems, and zeolites have been reported as giving rise to large ER responses. Water is often identified as an agent which activates many solid phases. Particles are often roughly spherical and few studies have systematically explored the effects of anisotropic shapes. Typical physical properties of the dispersed phase are given in Table 1.

ER suspensions are made up at 1–50 volume% solids and, in the absence of an electric field, display the properties expected of nonaqueous suspensions. The particles tend to flocculate because of long-range van der Waals interparticle forces and thus to settle out of suspension. This poses problems for devices which are called upon to respond in an intermittent manner. As a result, stabilizing agents are commonly incorporated in fluid formulation. Surfactants and grafted or adsorbed polymeric materials have been investigated, but the effects of these agents are complex and poorly understood, even in the absence of an external electric field. As discussed in more detail below, the choice of stabilizer is made difficult by the need for the electric field to hold particles at small pair separations without irreversibly trapping them in a deep potential energy minimum produced by van der Waals forces (for a discussion of colloidal interactions see Russel *et al.*, 1989). Thus a balance of stabilizer size and the softness of the resulting interparticle potential must be sought in suspension formulation.

Depending on volume fraction and degree of dispersion, in the absence of an externally applied field, ER suspensions may display a yield stress (i.e., a stress below which the materials respond as an elastic solid rather than as a fluid). Most formulations aim to reduce the yield stress in the absence of a field to as low a value as possible. As the shear rate is raised, unelectrified ER suspensions are shear thinning and, at elevated shear rates, reach a Newtonian shear viscosity of η_∞ which depends primarily on suspension volume fraction and can be estimated from

$$\eta_\infty/\eta_c=(1-\phi/\phi_m)^{-2}, \tag{1}$$

where η_c is the viscosity of the continuous phase and ϕ_m is a maximum packing fraction often taking a value between 0.63 and 0.71 (Russel *et al.*, 1989).

Devices employing the ER response hold the suspension between two electrodes. Because of the large field strengths required to achieve the desired viscosity changes (on the order of 0.5–5 kV/mm), gaps on the order of 0.5–5 mm are commonly encountered. The ratio of the gap width to particle size, l/σ, ranges from 5 to 100 in most devices. To date, no systematic studies have been carried out to explore the effects of l/σ on the resulting ER response.

When the electric field is applied to the quiescent suspension, the particles rearrange from an initially random microstructure to one where fibers or columns of particles are seen to span the electrode gap (Fig. 1). At very low volume fractions the fibers tend to be one particle wide and form slowly (over the course of tens of minutes depending on the field strength; see below for a more detailed description of the time dependencies). As the volume fraction is raised, the columns thicken and become branched.

If one electrode is moved parallel to the other so as to produce a small strain in the field-induced structures, the strands respond in an elastic manner, showing that they are capable of sustaining a stress without rupturing, and thus they display solid-like characteristics. At small steady rates of deformation, the strands break and reform in a continuous manner. The strands remain attached to each electrode and the strands thicken until a characteristic, steady-state size distribution is reached. At higher shear rates, the strands rupture completely such that there are solid-like regions attached to each electrode separated by a fluid-like zone. The fluid-like zone expands until it occupies the entire gap (Klingenberg and Zukoski, 1990) (Fig. 1) as the shear rate is still further raised. At these shear rates the electrified suspension viscosity approaches η_∞.

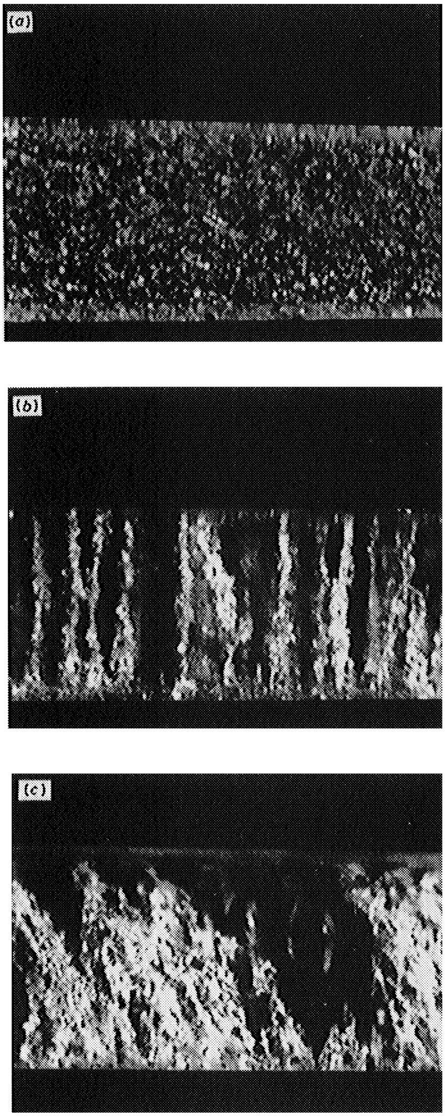

FIG. 1. 100-μm-diam hollow silica spheres suspended in corn oil between two metal electrodes spaced 3 mm apart, at (a) 0 V/mm, (b) 750 V/mm with a zero shear rate, and (c) 750 V/mm and a shear rate of 0.5 s^{-1} (reproduced from Klingenberg and Zukoski, 1989).

In slit or pipe flows at pressure drops below those required to break the strands held in place by a given field, the strands act as weakly extensible elastic solids. At low to intermediate volume fractions, the network behaves like an excellent deep bed filter with particles being captured in the fibrous structure and clear fluid being expressed at the low-pressure outlet. As the pressure drop is raised the fibers bend and then break, giving rise to plug flow at just above the yield pressure. At still larger pressure drops, the field-induced structures are completely destroyed.

The current passed by the ER suspensions typically increases faster than linearly with the electric field and depends on both the continuous- and dispersed-phase conductivities. At a constant field strength the current drops at shear rates near those required to form a fluid-like zone and increases again upon cessation of flow when the percolating strands reform. With increases in the degree of hydration of the dispersed phase, the ionic character of dispersant, or operating temperature, currents increase. Not only do increasing power losses place economic constraints on ER suspensions with elevated conductivities, but the resulting Joule heating degrades the ER response.

The shear behavior of ER fluids is characterized by a field-strength–dependent dynamic yield stress, shear thinning with increasing shear rates, and a high–shear-rate viscosity that is independent of field strength. These features are conveniently captured in a Bingham type of constitutive equation,

$$\tau=\tau_y+\eta_\infty\dot{\gamma}, \tag{2}$$

where τ is the stress, $\dot{\gamma}$ is the shear rate, and τ_y is the field-strength–dependent dynamic yield stress. In a plot of stress versus $\log\dot{\gamma}$, a plateau stress is observed at low $\dot{\gamma}$ and a slope of unity at elevated $\dot{\gamma}$ (Fig. 2). The stress extrapolated to zero shear rate defines τ_y. Here τ_y increases with field strength E, but the stress transferred by the suspension becomes independent of E at elevated $\dot{\gamma}$ [as expected from Eq. (2)] suggesting that one of the key parameters of interest in understanding ER suspension behavior and designing devices is τ_y. The shear rate where τ lifts off the low–shear-rate plateau corresponds well with the appearance of the fluid-like regions seen in flow visualization experiments, indicating that the transition to Newtonian flow behavior at high $\dot{\gamma}$ is associated with degradation of field-induced structures.

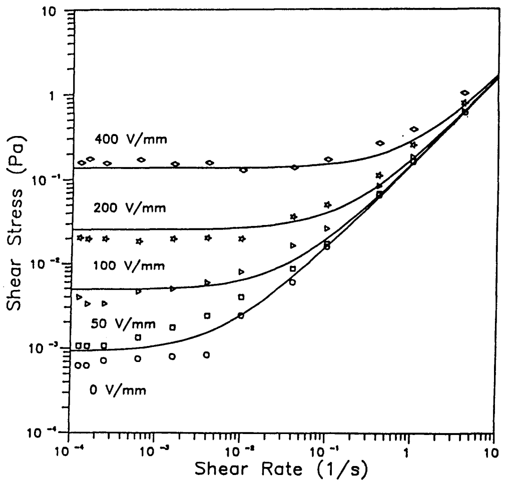

FIG. 2. Shear stress in pascals as a function of shear rate at various field strengths for an ER suspension composed of hydrated polymethylacrylate particles in a chlorinated hydrocarbon. Solid lines drawn from Eq. (2) with τ_y proportional to E^2.

For many ER suspensions, a critical field strength is required to break up the flocculated network produced by colloidal forces, and the overall yield stress is well approximated as the sum of a yield stress observed when $E=0$ and a field-dependent yield stress (Gow and Zukoski, 1990). For suspensions with small yield stress at $E=0$, τ_y is found to scale as E^n. For many suspensions up to field strengths of 1–2 kV/mm, n is found to lie near 2. At higher field strengths n often approaches unity. In addition, τ_y/E^2 is found to increase linearly at low volume fractions, but to saturate at volume fractions near 0.3. τ_y/E^2 is a function of the material properties of the dispersed and continuous phases, as discussed in more detail below.

The kinetics of stress-transfer changes with variations in field strength or shear rate are poorly understood, but appear to be dominated by two characteristic times: that required to produce electrode-spanning structures and that required for the shear to rearrange these structures. The time to produce percolating strands depends primarily on particle polarization properties, volume fraction, and the square of the field strength, while shear degrades the strands in a time characterized by $\dot{\gamma}^{-1}$. The result is that at intermediate shear rates and/or field strengths there is a large stress overshoot before the steady-state stress is reached. The stress overshoot is of importance in the design of control algorithms for situations where rapid deceleration is required.

The dynamic response of ER suspensions has been less well studied. In the presence of an electric field, the suspensions have a substantial elastic modulus and a small limiting strain above which they display nonlinear responses. As a result, examination of the linear viscoelastic properties is somewhat instrumentation limited. However, interest in applications such as acoustic dampers will result in continued studies of the small-strain dynamic response of ER materials.

2. ORIGINS OF THE ER RESPONSE

In his original publications, Winslow (1947, 1949) identified strand formation as a key process in successful ER fluids, and postulated that these structures were formed as the result of the interactions of particles polarized in the external field. Increases in viscosity were then associated with breaking polarization-induced interparticle "bonds." Klass and Martinek (1967) did not observed strands in their fluids and postulated that the increases in viscosity could arise from polarization of double layers that surround the slightly charged particles. Later, because of the apparent requirement of water to activate ER suspensions, Stangroom (1983) postulated that water migrated within particles by electroosmosis, and that interparticle bonds were formed by a water bridging mechanism.

By far the most successful approach to understanding and codifying the ER response lies in the polarization model as first suggested by Winslow. Here, particles of different dielectric constant than the continuous phase polarize when the suspension is exposed to an external electric field. To a first approximation, the spherical particles of diameter σ interact as point dipoles which feel forces attempting to align pairs along the lines of electrical force. The magnitude of the polarization force is given by $(3/16)\pi\epsilon_0\epsilon_c\sigma^2\beta^2E^2$ where $\beta[=(\epsilon_p-\epsilon_c)/(\epsilon_p+2\epsilon_c)]$ is the relative polarizability of the particles, with ϵ_p and ϵ_c being the particle and continuous-phase dielectric constants. Here, the permittivity of free space is ϵ_0.

If the particles are small and feel negligible colloidal interaction forces, and the field is adjusted such that $(3/16)\sigma^3\beta^2E^2/kT$ is of order unity, the suspensions display phase behavior of dipolar materials. Here, kT is the product of Boltzmann's constant and the absolute temperature. At low volume fractions and modest field strengths, the particles establish a cluster length distribution that can be predicted from a balance of thermal and polarization forces using classical statistical mechanics. As the volume fraction or field strength is raised, a percolation transition is approached where chains of infinite length are formed (Gast and Zukoski, 1989).

Typical ER suspensions consist of large particles where the ratio of polarization to thermal forces is large [i.e., $(3/16)\pi\sigma^3\beta^2E^2/kT \gg 1$]. Under these conditions, low-temperature approximations are valid and suspension structures are determined by kinetic rather than equilibrium considerations. Through polarization interactions, particle columns and fibers are formed, percolate across the electrode gap, and reach a steady-state width distribution. If one electrode is moved parallel to the second, stress is transferred through a combination of viscous and electromechanical mechanisms. Within the polarization-model framework, time scales for structure formation can be found by balancing the viscous drag on particles moving in the continuous phase against the characteristic polarization interaction. The characteristic time for percolating structures to form when an initially random suspension of non-Brownian particles is exposed to an electric field is predicted to vary as $t^*=16\eta_c/\epsilon_0\epsilon_c\beta^2E^2$ and to be dependent on volume fraction. Simulation studies, which have been confirmed experimentally, have demonstrated that the average time to produce percolating structures, $\langle t_p\rangle$, scales as $\langle t_p\rangle/t^*\sim\phi^{-4.5}$ for $\phi<0.5$. This result demonstrates that the polarization model can predict measured response times reported in the millisecond range. In the design of ER devices, while stress-transfer properties plateau as volume fraction is raised, response times can be easily varied over a wide range through changes in ϕ (Klingenberg *et al.*, 1989).

With growing concerns for shorter response times, the rate of particle polarization becomes an important parameter. The characteristic particle polarization time may be estimated from $\epsilon_p \epsilon_0 / K_p$, where K_p is the particle conductivity. As K_p increases, dipoles are formed more rapidly when the electric field is applied. Few studies have investigated the role of particle conductivity on ER suspension response time. As a result, the role of particle response time (and thus K_p) in the ER response is poorly understood and subject to debate (Uejima, 1972; Block *et al.*, 1990; Gow and Zukoski, 1990).

In the polarization model, the continuous shear behavior of ER suspensions results from a balance of polarization and viscous interaction forces, such that the relative viscosity of the electrified suspension will be a function of the Mason number, $Mn = \eta_c \dot{\gamma} / 2\epsilon_c \epsilon_0 \beta^2 E^2$. As exemplified by Fig. 3, this scaling works well for many ER suspensions. Essentially the model predicts that at low Mn (i.e., high field strengths or low shear rates), polarization forces dominate suspension structure and stress is transferred by electromechanical means through the percolating particle strands. At large Mn, the viscous forces dominate, and the polarization-induced structures are degraded. Thus the suspension behaves as if no field were applied (i.e., $\tau/\dot{\gamma} = \eta_\infty$). The Bingham constitutive equation can be rearranged to yield

$$\eta/\eta_\infty = (Mn^*/Mn) + 1, \tag{3}$$

where Mn^* becomes a volume-fraction- and material-dependent, dimensionless yield stress. This constitutive equation for the steady shear behavior of ER suspensions is found to be quite accurate at small and large Mn. At intermediate values, Eq. (3) under-predicts the ER response because it does not account for synergistic effects between electromechanical and viscous stress-transfer mechanisms (Marshall *et al.*, 1989).

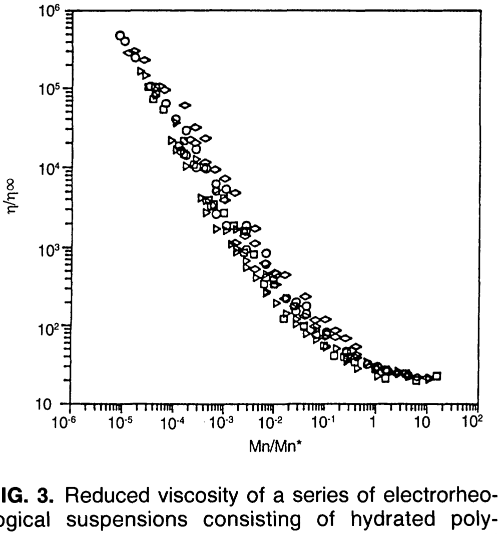

FIG. 3. Reduced viscosity of a series of electrorheological suspensions consisting of hydrated polymethacrylate particles suspended in a chlorinated hydrocarbon for volume fractions ranging from 0.07 to 0.35 for shear rates between 10^{-4} and 10 s^{-1} and field strengths of 50–400 kV/m. $M_n^* = 0.45\ \phi$.

Detailed simulation studies of hard particles suspended in a viscous continuous phase accounting for exact hydrodynamic and electrostatic interactions have shown that the polarization mechanism is capable of capturing both the qualitative features of the continuous flow behavior and the magnitude of the ER response based on particle volume fraction, continuous-phase viscosity, and the dielectric constants of the two phases (Bonnecaze and Brady, 1990; Heyes and Melrose, 1990; Klingenberg *et al.*, 1991).

As observed experimentally for $E \leqslant 1$–3 kV/mm, the polarization model predicts that yield stress scales, $\tau^* = (3/16)\pi\epsilon_0\epsilon_c\beta^2E^2$. However, quantitative prediction of the yield stress is very sensitive to particle interactions at small surface-to-surface separations. Thus details of the electrostatic interaction accounting for large ratios of ϵ_p/ϵ_c (i.e., outside the point dipole limit) and multiparticle interactions play an important role in setting the magnitude of the ER response. In addition, the nature of the short-range repulsive interactions provided by stearic stabilizers or surface roughness will determine how close particle cores may approach and thus have impact on the magnitude of the yield stress. In keeping with experimental observations, for fixed short-range repulsive interactions, the polarization model predicts a large increase in dimensionless yield stress (τ_y/τ^*) as β approaches unity.

As the volume fraction is raised, molecular-dynamics simulations predict that the dimensionless yield stress will saturate at a volume fraction near 0.4. This is due to the formation of dense clusters of particles that transfer stress less efficiently than the single-particle chains that dominate at lower volume fractions, a result well supported by experimental evidence (Klingenberg *et al.*, 1990) (Fig. 4).

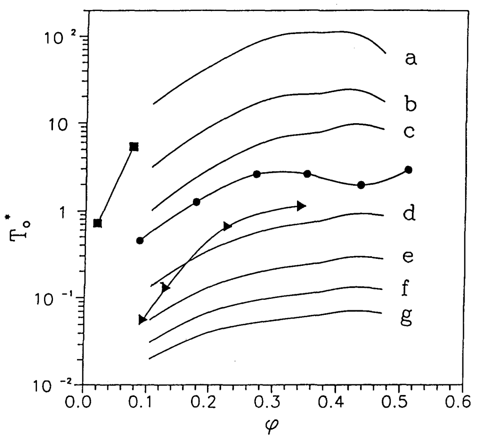

FIG. 4. Dimensionless yield stress $T_0^* = \tau_{yE}/(3/16)\pi\beta^2 E^2$ predicted from simulations of sheared ER suspensions. Solid lines are predictions for conducting particles with minimum surface-to-surface separations of (a) $10^{-5}\sigma$, (b) $10^{-4}\sigma$, and (c) $10^{-3}\sigma$, and dielectric spheres which come into contact for ϵ_p/ϵ_c of (d) 10.0, (e) 4.0, (f) 2.0 and (g) 1.0 (limit form). Experimental points are for (squares) polyaniline in silicone oil with $\epsilon_p/\epsilon_c = 36$; (circles) hollow silica spheres in corn oil with $\epsilon_p/\epsilon_c = 3.9$; and (triangles) hydrated polymethacrylate particles in a chlorinated hydrocarbon with $\alpha = 3.2$. (Reproduced from Klingenberg *et al.*, 1991).

As $\epsilon_p/\epsilon_c \to \infty$, the polarization model predicts that the yield stress will grow without bound for particles in contact. This picture is altered by chemical limitations of the systems under study. For example, conducting particles are observed to form chains when exposed to electric fields and thus a large ER response is expected. However, the chains also conduct electricity and the power requirements to maintain the field across the gap are unacceptable. Even if the electrodes are coated such that no loss current can flow, if conducting particles are held at too small a separation and the field strength is raised, the dielectric breakdown voltage of the continuous phase is reached and the ER response is severely degraded. However, the polarization model predicts that conducting particles held apart by a thin insulating layer suspended in a dielectric continuous phase represent upper bounds on the ultimate strengths of ER suspensions. Based on a model of conducting particles suspended in a continuous phase of dielectric constant 4 at a volume fraction of 0.37, and a minimum pair surface-to-surface separation of $1 \times 10^{-5}\sigma$, a yield stress of 240 kPa is predicted at 10 kV/mm (Klingenberg *et al.*, 1990). This is a response that, if available, would open a host of technologies. Unfortunately, an experimental yield stress of this magnitude has not yet been reported.

The successes of the polarization model lie in its ability to reproduce time scales, magnitudes of yield stresses, and dependence of the viscosity on field strength, shear rate, and volume fraction. However, models based solely on electrostatic mechanisms are not yet able to account for all experimental observations. First, the models make no attempt to predict the current-carrying capacity of ER suspensions. This is a problem of great technological importance as energy consumption limits ER device economic viability, and the resulting Joule heating produces excessive temperature changes and degrades the response. Second, there is debate about the requirements of a loss current for there to be an observable ER response. Third, the polarization model cannot yet explain why some materials with large bulk dielectric constants do not produce suspensions with large ER responses. Much of the uncertainty over the origin of this failure of the polarization model revolves around the details of the electrochemical system under investigation, how current is transferred through the suspension, and the dispersed-phase polarization mechanisms. Fourth, the effects of temperature on ER suspensions are not fully understood in terms of the polarization model. Some suspensions appear to respond well at high temperature and poorly at low temperature while in others, the reverse is observed. To meet the needs of the automotive industry, the effects of temperature will need to be characterized and controlled.

The polarization model is largely a linear description of the ER response. Often the field strengths used in ER devices exceed the limits of linearity, and then descriptions of the physics and chemistry of nonpolar suspensions subjected to large field strengths are poorly characterized. Understanding many of the failings of the polarization model requires more experimental study. Advances in understanding of the ER response will also be aided by studies of how current is transferred in highly nonpolar continuous phases.

In conclusion, the polarization model has been successful in establishing the bounds where linear behavior can be expected and what materials should produce high-quality ER responses. As a consequence, the way is paved for detailed and rigorous research and development activities. For example, studies of the effects of particle size, shape, and polarization time will lead to greater understanding of how material properties control the ER response. In addition the influence of continuous-phase conductivity and chemistry occurring at the electrode surfaces is likely to influence ER fluid development. Finally, the role of thermal motion in controlling the field where percolating chains are produced, details of the microstructure, and the role of free charge in limiting the ER response are subjects of fundamental and technological interest worthy of further investigation.

3. APPLICATIONS OF ER DEVICES

Devices using ER fluids typically rely on the ability of the ER suspension to change viscosity in a reversible manner upon the application of an electric field. The continuous variability of stress-transfer capability, fast switching speeds, and ease of control offer tantalizing opportunities for device design. Clutches, shock absorbers, robotic actuators, and engine mounts are commonly considered applications for ER suspensions. Here, a series of parallel plates with alternating polarity contain the ER suspension. Plates of one polarity move relative to those of opposite polarity and stress is transferred. When the field is changed, the magnitude of the stress transferred is altered. Advantages of ER devices are rapid and continuously variable stress-transfer capability and the possibility of solid-state electronic control of stress transfer. Engine mounts and active suspensions are currently under consideration in the automotive industry. Power requirements, temperature limitations, and toxicity are considered to be the main stumbling blocks to commercialization (Shulman *et al.*, 1981; Block and Kelly, 1988; Jordan and Shaw, 1989).

In an alternative series of applications, ER fluids are used in hydraulic lines. Electrical valves are then used to control flow rates and pressure drops. Problems faced here are the heterogeneous nature of the ER suspensions. Structures capable of sustaining the required stresses can be generated in the electric field. However, the strands and columns are of limited utility in controlling flow of the continuous phase (Shulman *et al.*, 1981; Block and Kelly, 1988; Jordan and Shaw, 1989).

Control of flexural stiffness is also considered a potential application. For example, a plate-like member is constructed such that it contains compartments with electrodes and ER fluid. When the member is subjected to a load, the compartments can be stiffened in such a manner as to damp out vibrations.

While many applications of the ER response involve limiting the macroscopic motion of a solid surface, new applications under consideration include acoustic dampers and transducers where the modulus of the transducer and thus frequency range to which it is sensitive can be tuned with the electric field. Here, rather than being concerned with the ultimate strength of the ER fluid at some field strength as represented, for example, by τ_y, one is concerned with field-strength and volume-fraction dependences of the storage and loss moduli of the electrified ER suspension in the linear response region.

In another intriguing application, the continuous phase of an ER suspension is made of a chemically crosslinking material such as a silicon sealant, epoxy, or urethane. A mold is constructed of electrodes configured in the shape of the ultimate piece and dipped into the reactive suspension. When the electrified mold is lifted out, the suspension clings to the mold where the electrodes have a narrow separation and runs off the mold where the field is not intense. If the field is kept on the electrode configuration until the material has set, an "electroshaped" object will result. This type of technology has applications in rapid generation of spare parts and prototypes.

The ER response has been known for almost five decades. Few applications have reached the commercial stage because of a variety of problems. The ultimate strength, erosive properties, long-term chemical and colloidal stability, current-carrying capacity, and temperature range are problems which have plagued ER fluid formulation. Recent advances in understanding of the origin of the response and thus the types of chemical and physical characteristics that give rise to large ER responses are beginning to eliminate some of these problems. In addition to formulation problems, limited visibility of ER devices has led to a lack of experience in design, and thus there is little working knowledge of long-term device and fluid performance. A strong coupling between the device design, its applications, and fluid formulation will be required before successful commercialization is achieved.

Driving the interest in ER devices is the availability of solid state power supplies and circuits to control stress transfer. Advances in power supplies and in circuits that can survive extreme environments, coupled with design of new ER suspensions, hold promise for commercial products in the next 5–10 years.

List of Works Cited

Block, H. J., Kelly, J. P., Qin, A., Watson, T. (1990), *Langmuir* **6**, 6.

Block, H., Kelly, J. P. (1988), *J. Phys. D.* **21**, 1661.

Bonnecaze, R. T., and Brady, J. F. (1990), in *Proceedings of the 2nd International Conference on Electrorheological Fluids*, edited by J. O. Carlson, A. F. Sprecher and H. Conrad, Lancaster, PA: Technomic.

Gast, A. P., Zukoski, C. F. (1989), *Adv. Colloid Interface Sci.* **30**, 153.

Gow, C. J., Zukoski, C. F. (1990), *J. Colloid Interface Sci.* **136**, 175.

Heyes, D. M., Melrose, J. M (1990), *Mol. Sim.* **5**, 293.

Jordan, T., Shaw, M. (1989), *IEEE Trans. Elec. Insul.* **24**, 849.

Klass, D. L., Martinek, T. W. (1967), *J. Appl. Physics* **38**, 67; **38**, 75.

Klingenberg, D. J., Zukoski, C. F., (1990), *Langmuir* **6**, 15.

Klingenberg, D. J., van Swol, F. B., Zukoski, C. F. (1989), *J. Chem. Phys.* **91**, 7888.

Klingenberg, D. J., van Swol, F., Zukoski, C. F. (1991), *J. Chem. Phys.* **94**, 6160 and 6170.

Marshall, L., Goodwin, J. W., Zukoski, C. F. (1989), *J. Chem. Soc. Faraday Trans. 1* **85**, 2785.

Russel, W. B., Saville, D. A., Schowalter, W. R. (1989), *Colloidal Dispersions*, Cambridge, U.K.: Cambridge Univ. Press.)

Shulman, F. P., Gorodkin, R. G., Korobku, E. V., and Gleb, V. K. (1981), *J. Non-Newtonian Fluid Mechanics* **8**, 29.

Stangroom, J. E. (1978), UK Patent Specification 1501635; (1980), UK Patent Specification 1570234.

Stangroom, J. E. (1983), *Phys. Technol.* **14**, 290.

Uejima, H. (1972), *Jpn. J. Appl. Phys.* **11**, 319.

Winslow, W. M. (1949), *J. Appl. Phys.* **20**, 1137.

Winslow, W. M. (1947), US Patent Specification 2417850.

Zia, I. Y., Cox, R. G., Mason, S. G (1975), *Proc. R. Soc. A* **113**, 516.

ELECTROSTATIC CAPACITIVE ENERGY STORAGE

D. K. HASKELL AND A. C. KOLB, *Maxwell Laboratories, Inc., San Diego, California, U.S.A.*
W. G. MCMILLAN, *Department of Chemistry and Biochemistry, University of California, Los Angeles, California, U.S.A.*

Introduction 155
1. **Capacitor Anatomy and Circuit Characteristics** 156
1.1 The Thunderhead: A Natural Geophysical Capacitor 156
1.2 Historical Sketch: The Leyden Jar "Condenser" 157
1.3 Capacitor Components and Structure 157
1.4 Capacitance 158
1.5 Capacitor Circuit Analysis 159
1.6 Dissipative Processes; The Equivalent Series Resistance 160
1.7 The Equivalent Series Inductance 161
2. **Macroscopic Dielectric Phenomenology** 161
2.1 Dielectric Polarization **P** 161
2.2 The Induced Bulk Dipole Moment **M** 162
2.3 The Displacement Field **D** 162
2.4 The Dielectric Permittivity ϵ 162
2.5 Thermodynamics of the Capacitor System 162
2.6 (Free-)Energy Storage 163
3. **Molecular Dielectric Phenomenology** 164
3.1 The Polarizability Tensor 164
3.2 The Clausius–Mossotti Equation 164
3.3 The Electronic Polarizability α_e 164
3.4 The Orientation Polarizability α_μ 165
3.5 The Distortion Polarizability α_d 166
3.6 Dielectric Saturation 167
3.7 Ferroelectric and Piezoelectric Materials 167
3.8 Dielectric Breakdown Strength 167
3.9 Breakdown Mitigation 168
4. **Types of Capacitors** 168
4.1 The Great Diversity of Capacitor Types 168
4.2 Capacitor Structure and Geometry 168
4.3 The Electrodes 168
4.4 Dielectric Materials 169
4.5 Electrolytic Capacitors 171
4.6 Chemical Double-Layer "Battery" Capacitors 171
5. **Capacitor Applications** 172
5.1 Energy-Storage Capacitors 172
5.2 dc Filters 173
5.3 Timing Circuits 173
5.4 Electromagnetic Interference (EMI) Suppression 173
6. **Whither Capacitor Technology?** 173
6.1 Industrial and Military Requirements 173
6.2 Some Critical Capacitor Technologies 174
Glossary 174
Works Cited 175
Further Reading 175

INTRODUCTION

Maxwell's equations, constituting the foundation of electromagnetic theory, deal with three primary entities: the electric field **E**, the magnetic field **B**, and the current density *j*. Corresponding to these three entities are the three primary elements of modern electronic circuits: capacitors, for the storage of energy in the electric field; inductors, for the storage of energy in the magnetic field; and resistors,

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

which convert electric current into heat energy.

This article addresses the structure, technology, applications, and development of one of these primary electronic elements: the capacitor. To cope with the near infinity of capacitor designs, sizes, electrode–dielectric combinations, and electrical characteristics, a generic approach is adopted in Sec. 1, in which the anatomy of a capacitor is first illustrated with Nature's thunderhead. A brief historical sketch of the first manmade capacitor—or "condenser"—the Leyden jar, is followed by a description of the essential components of a capacitor, together with mathematical definitions of capacitance, dielectric permittivity and relative permittivity, and free-energy storage. An analysis of a typical *R-C-L* circuit introduces the circuit characteristics, which then form the basis of the technical evaluation of capacitor quality in terms of the equivalent series resistance (ESR), equivalent series inductance (ESL), and dissipation factor.

The relation of these gross electrical characteristics to the underlying physics is examined at two levels: first, Sec. 2 offers a field description of the dielectric polarization, induced bulk dipole moment, Maxwell's displacement field, permittivity, the general thermodynamics, and the Gibbs free energy, and how these determine the capacitor electrical characteristics; and second, Sec. 3 describes dielectric phenomenology at the molecular level, including the several types of polarizability, dielectric saturation, breakdown, and breakdown mitigation.

These fundamentals provide the basis of the discussion in Sec. 4 of the various types of capacitors, their structure and geometry, electrode types, and dielectric materials.

Section 5 singles out four illustrative capacitor applications: Energy storage, dc filters, timing circuits, and electromagnetic interference suppression.

The article concludes in Sec. 6 with a projection of the technology of high–energy-density capacitors.

1. CAPACITOR ANATOMY AND CIRCUIT CHARACTERISTICS

1.1 The Thunderhead: A Natural Geophysical Capacitor

On a warm summer day a rapidly moving mass of moisture-laden air sweeps northward from the Gulf of California, forming clouds by adiabatic expansion and cooling as it rises over the high country of northern Arizona. As the afternoon wears on and insolation from the waning sun declines, the cloud water droplets become supercooled. When freezing is finally spontaneously initiated in the droplets of a local volume or "cell" of the cloud, the heat liberated creates a sudden buoyancy, resulting in violent updrafts that propel the cell into a thunderhead towering far above the cloud deck. Accompanying this development is a separation of electrical charge of ill-understood mechanism but variously ascribed to "friction", which builds an enormous difference in the electrostatic potential Φ between the positively charged top and negative base of the thunderhead (Feynman *et al.*, 1964).

Nature thus produces an awesome electrical capacitor of geophysical dimensions, in which the two regions of accumulated charge near the thunderhead top and base serve as the electrodes, with the intervening air as the dielectric medium (Fig. 1). The underlying surface of the Earth acts as a third electrode, positively charged from induction by the negative cloud base.

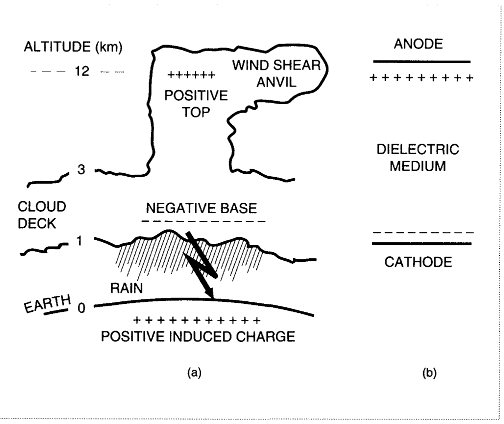

FIG. 1. (a) Charge distribution in a typical thunderhead. (b) Schematic capacitor analog.

If the electric field **E** ($\equiv -\mathrm{grad}\Phi$) reaches the breakdown strength of the air, ~3 MV/m, the capacitor discharges through a spectacular arc or lightning stroke. Lightning can occur between parts of a single cloud, between clouds, or between cloud and Earth. Because of the limited conductivity in the "electrodes," the stroke momentarily depletes the charge concentrations about the ends of the arc, reducing the potential difference, quenching the arc, and allowing the charge separation to resume its buildup.

Evidently to produce such high electric fields in thunderstorms requires a rate of charge separation far exceeding the charge neutralization rate due to conductivity. Conductivity is always present, accounting for a gradual "leakage" of electrode charge and eventual loss of the stored electrical energy.

These natural geophysical capacitors thus mirror on a grand scale not only the structure and general properties of manmade capacitors, but also their failure modes—conduction, leakage, energy loss, dielectric breakdown, and interelectrode arcing—and their recovery—arc quenching.

Collectively, the 2000 thunderstorms in progress worldwide drive a continuous power source of ~700 MW. But in the short duration (~2 ms) of a lightning stroke to Earth from a cloud at 1 km altitude, the power is much greater: ~10 kA at 3 GV produces 3×10^7 MW and an energy release ~6×10^{10} J or 15 tons of high-explosive equivalent! Fortunately, Benjamin Franklin did not tap into this enormous energy source efficiently when he performed his dangerous kite-flying experiment in June 1752. However, using his wet kite string as conductor, he was able to charge a Leyden jar—the forerunner of the modern capacitor.

1.2 Historical Sketch: The Leyden Jar "Condenser"

Electrification produced by friction on certain substances, notably amber and jet, was mentioned by Theophrastus (321 B.C.) and was said to have been familiar to Thales of Miletus (600 B.C.) (Fleming, 1910). The first electrostatic generating machine, employing a revolving ball of sulfur, was built by Otto von Guericke ca. 1663. In 1729, Stephen Gray, of the Greyfriars' School in England, discovered that static electricity is conducted by certain materials and soon was transmitting electrostatic charges around his house over "wires" made of such things as hempen string. The apparent flow of charge led to the belief that electricity was some sort of fluid, and prompted research into ways of collecting this electric fluid.

In 1745, Ewald Georg von Kleist, Bishop of Pomerania, theorized that electricity could be collected in a glass container, glass being already well recognized in the mid-1700s for its electrical insulating properties. A bottle partially filled with water was fitted with a brass nail inserted through a cork in the mouth and touching the water. Holding the bottle by hand, von Kleist presented the projecting nail head to the pole of an electrostatic friction generator. Upon removing it from the generator and touching the nail with his other hand, he received a strong electrical shock.

During the following year, Pieter van Musschenbroeck, Professor of Physics at the University of Leiden, conducted a similar experiment. He also started with water in a bottle, but soon discovered that the charge could be intensified by replacing the water with mercury. By outlining his experiment in a letter to a colleague, who presented it to the French Academy of Sciences, van Musschenbroeck received the credit, and the "Leyden jar" thus acquired its name. In later experiments the Leyden jar was improved by coating the inner and outer surfaces with tin foil (Fig. 2).

The Leyden jar was the original "condenser"—so designated for its ability to concentrate and store electric charge. Spectacular demonstrations of the condenser's power were carried out, including electrocuting birds and fish, and igniting gunpowder and ether. Sir William Watson conveyed the discharge of a Leyden jar over metallic wires for a distance of 2 miles, observing that the transmission appeared instantaneous. This gave rise to a proposal in 1753, now credited to Charles Morrison (Singer, 1958), for an electrostatic telegraph in which the successive charging of an appropriate sequence of the 26 alphabet wires connecting the transmitter and receiver could convey any message.

1.3 Capacitor Components and Structure

The essential elements of a capacitor in a circuit are schematically illustrated in Fig. 3. Upon the closing of switch S1, with S2 open, equal and opposite charges appear on the

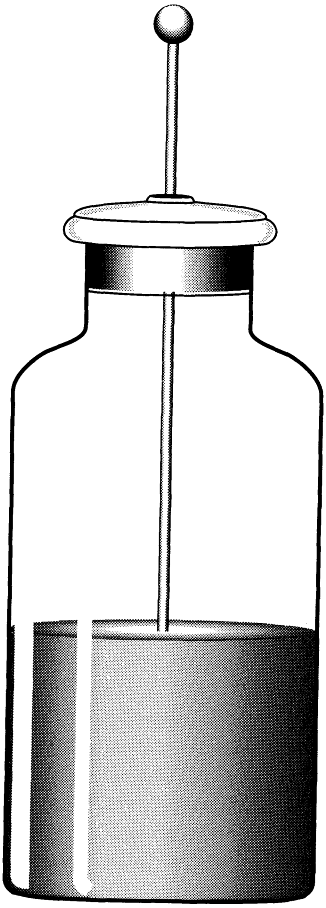

FIG. 2. Glass Leyden jar.

upper and lower plates. These charges result from the transfer of electrons, which continues until the potential difference between the electrodes is equal to the applied voltage—supplied, in the case shown, by a battery.

Modern capacitors operate on the same physical principles as the Leyden jar. As shown schematically in Fig. 1(b), a typical capacitor still utilizes some form of charged electrodes separated by a dielectric material, usually in a multiple-layer sandwichlike structure, in either planar or cylindrical geometry.

The most significant advances in capacitor technology have been made in new materials that permit the use of large areas of dielectric exposed to high electric fields, and in the design of the electrodes. Many common organic plastics, such as various polyesters, polycarbonates, polypropylene, polystyrene, etc., are currently used in film form as capacitor dielectrics. The electrodes of most capacitors are now either aluminum foils 4–12 μm thick, or extremely thin metal coatings—of the order of a few hundred angstroms thick—deposited directly on the surface of the dielectric film.

Inorganic compounds constitute the dielectric of "ceramic" capacitors (see DIELECTRIC PROPERTIES OF INSULATORS), which are today mass produced with such materials as barium titanate ($BaTiO_3$) and lanthanum-doped lead zirconate/titanate (PLZT). Aluminum and tantalum oxides are used as dielectric media in another class, the electrolytic capacitor.

1.4 Capacitance

The measure of the ability of a capacitor to store electric charge, denoted by its capacitance C, is usually expressed as the ratio of charge Q to potential difference V (see ELECTRODYNAMICS, CLASSICAL):

$$C=Q/V. \tag{1}$$

This definition is adequate for low voltages where the capacitance is essentially constant, but is inconvenient for higher voltages where the capacitance may become voltage dependent. In that case it is desirable to use the more general definition of capacitance $C(V)$ as the differential coefficient

$$C(V)=dQ/dV, \tag{2}$$

which, of course, includes a constant capacitance as a special case.

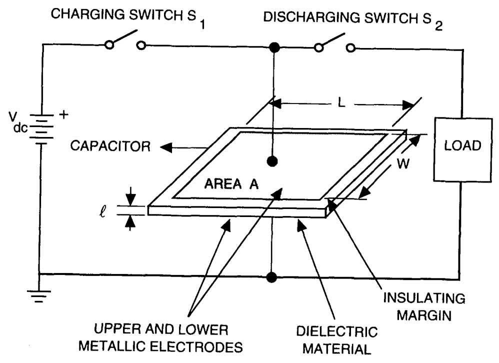

FIG. 3. Essential elements of a capacitor in a circuit.

Capacitance depends not only upon the geometric factors of electrode area A and separation distance l, but also and particularly upon the permittivity ϵ of the dielectric medium. Within the dielectric between electrodes bearing charges $\pm Q$, increasing the charge by dQ causes a corresponding increase in the electric field E of magnitude

$$d\mathsf{E} = dQ/A\epsilon, \tag{3}$$

and an increase in potential difference of

$$dV = l d\mathsf{E}. \tag{4}$$

By Eq. (2) the capacitance is thus

$$C = \epsilon A/l. \tag{5}$$

In the mks system with charge Q in coulombs and V in volts, C is in farads.

The farad, named after the nineteenth-century English physicist Michael Faraday, is generally too large a unit of capacitance for most practical applications. It is even much too large for the cloud-to-ground thunderstorm system of Fig. 1(a): an area of 10 km^2 and altitude 1 km, with air permittivity essentially the same as for the vacuum,

$$\epsilon_0 = 10^7/4\pi c^2 \text{ C/V m}, \tag{6}$$

where the speed of light $c = 3\times10^8$ m/s, has capacitance

$$C = (10^7/4\pi c^2)10^7/10^3 = 0.09\mu F. \tag{7}$$

The microfarad (μF), or 10^{-6} farad, is a more common unit of capacitance, although for the miniature capacitors used in integrated circuits the picofarad (or, in the older literature, the micromicrofarad, $\mu\mu$F) is sometimes more appropriate. However, modern capacitors for high density of energy storage and pulse-power delivery may have capacitances of the order of a farad, more than a millionfold greater than the geophysical capacitors of nature's thunderheads. The capacitance range of industrial capacitors thus spans at least 12 orders of magnitude!

1.5 Capacitor Circuit Analysis

In use, a capacitor must of course be connected to other circuit elements. The solder joints, electrode contacts, leads, and other "external" peripherals always have some series resistance R and inductance L that must be taken into account in time-dependent applications, whether ac of pulse operation.

With reference to Fig. 4(a), the applied voltage, in the complex form $V = V_0 e^{i\omega t}$, with ω $(=2\pi\nu)$ being the angular frequency, is the sum of the voltage drops across each of the series elements:

$$V = V_0 e^{i\omega t} = L dI/dt + RI + \frac{1}{C}\int I\,dt. \tag{8}$$

As shown in standard textbooks (Page, 1935), this equation is solved by the substitution of the time-dependent current,

$$I = I_0 e^{i\omega t}, \tag{9}$$

which yields

$$V = I[R + i(\omega L - 1/\omega C)] \equiv IZ. \tag{10}$$

In an ac circuit, the complex impedance Z thus occupies the role played by the resistance in a dc circuit:

$$\begin{aligned} Z &\equiv R + i(\omega L - 1/\omega C) \\ &= [R^2 + (\omega L - 1/\omega C)^2]^{1/2} \\ &\quad \times [R/Z_0 + i(\omega L - 1/\omega C)/Z_0] \\ &= Z_0 e^{i\delta}, \end{aligned} \tag{11}$$

where Z_0 stands for the modulus radical in the second line, and the "phase angle" δ is given by

$$\tan\delta = (\omega L - 1/\omega C)/R. \tag{12}$$

The relationships among V, I, and δ in the complex plane are shown in Fig. 5. The circuit "reactance" X is given by the numerator of Eq. (12):

$$X \equiv \omega L - 1/\omega C. \tag{13}$$

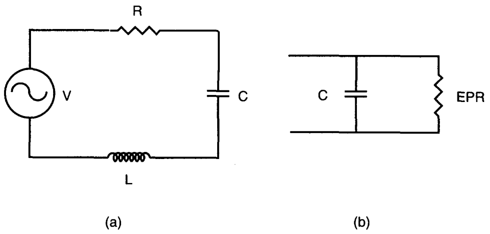

FIG. 4. (a) External series resistance R and Inductance L. (b) Equivalent parallel resistance (EPR).

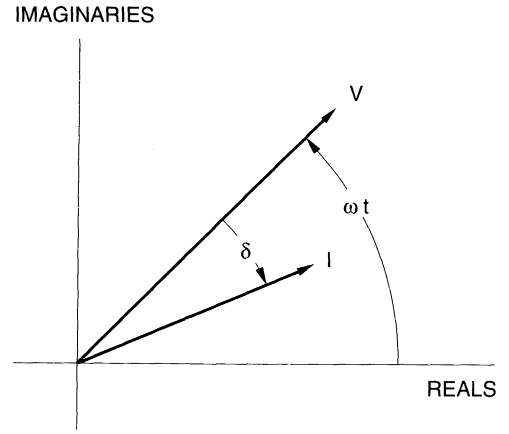

FIG. 5. Representation of V, I, and phase lag δ in the complex plane.

Maximum power is achieved under the resonant condition $\delta=0$ where, from Eq. (12), the resonance angular frequency ω_r is given by

$$\omega_r=1/\sqrt{LC}. \tag{14}$$

1.6 Dissipative Processes; The Equivalent Series Resistance

The conversion of electrical energy stored in a capacitor into heat not only reduces the energy efficiency but also may limit the number and repetition rate of cycles in pulsed operations because of intolerable temperature rise. Three principal mechanisms account for the loss of energy in capacitor operation:

1. Ohmic heating in the combined series resistance R_s of the metal conductors: the electrodes, the foils in wound capacitors, the soldered connections in extended-foil capacitors, the contact resistance in tabbed capacitors, etc.;
2. charge leakage and accompanying Ohmic heating due to the omnipresence of a nonzero conductivity of the dielectric medium, and a corresponding finite shunt or equivalent parallel resistance, R_p [Fig. 4(b)];
3. dielectric energy loss, arising from the conversion of the mechanical energy of atoms and molecules into heat as they try to follow the changes in the electric field.

In the absence of any series inductance L these three loss mechanisms can be accounted for by introducing an equivalent resistance in series with the capacitor—the equivalent series resistance (ESR) (Hall, 1978).

For each type of loss mechanism it is customary to define a dimensionless dissipation factor D as the ratio of the energy loss rate to the energy storage rate. If in Eq. (13) the series inductance L is essentially zero, the magnitude of the capacitor reactance is

$$X_c=1/\omega C. \tag{15}$$

Then, if I is the current through the capacitor, the rate of increase in the stored (free) energy G is

$$dG/dt=I^2X_c=I^2/\omega C. \tag{16}$$

The dissipation factor for a series resistance R_s, which carries the same current I as the capacitor, is evidently

$$D_s=I^2R_s/I^2X_c=\omega CR_s\,. \tag{17}$$

The dissipation factor for the parallel leakage resistance R_p that subtends the same voltage as the capacitor is

$$D_p=(I^2X_c/R_p)^2R_p/I^2X_c=1/\omega CR_p\,. \tag{18}$$

Similarly, dielectric loss can be attributed to an equivalent parallel resistance R_d that subtends the same ac voltage as that across the capacitor. This leads to the same form of dissipation factor as for D_p, Eq. (18):

$$D_d=1/\omega CR_d\,. \tag{19}$$

The total dissipation factor D is simply the sum of these three contributions:

$$D=\omega CR_s+(1/R_p+1/R_d)/\omega C. \tag{20}$$

Noting that D is essentially the ratio of energy lost to energy stored, the ESR becomes

$$R_{\text{ESR}}=D/\omega C=R_s+(1/R_p+1/R_d)/(\omega C)^2. \tag{21}$$

The reciprocal of the dissipation factor, $Q=1/D$, is known as the quality factor. Lumping the three loss mechanisms into the ESR has the virtue of allowing the familiar circuit analysis and language to be applied.

With reference to Fig. 6(a), the impedance, Eq. (11), becomes

$$Z=R+iX_c\,, \tag{22}$$

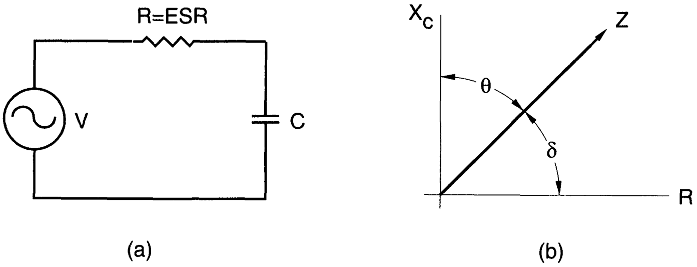

FIG. 6. (a) Capacitor circuit with equivalent series resistance. (b) Relationships in the complex plane.

The tangent of the phase angle δ is again given by the reactance/resistance ratio,

$$\tan\delta = X_c/R. \tag{23}$$

The "loss angle" θ is the complement of the phase angle δ; its tangent is the dissipation factor:

$$\tan\theta = \cot\delta = R/X_c = D. \tag{24}$$

The quality factor Q, being the reciprocal of D, is evidently

$$Q = \tan\delta = \cot\theta. \tag{25}$$

These relationships are shown in the complex plane, Fig. 6(b).

1.7 The Equivalent Series Inductance

As with the equivalent series resistance, each capacitor has a series inductance L, called the equivalent series inductance (ESL). This inductance is inherent in the geometry of the capacitive elements, the connections, the type of insulators or feedthrough used, the current return, etc.

The ESL of a capacitor can be determined by finding its resonant frequency. This test, known as the "standing-wave method", imposes a standing wave, generated by a variable-frequency oscillator, upon a transmission line terminated by the capacitor under test. The frequency at which a null voltage is detected at the capacitor by a high-impedance voltmeter is the resonant angular frequency ω_r of the capacitor, from which, according to Eq. (14):

$$L = 1/C\omega_r^2, \tag{26}$$

wherein L is the inductance in henries and C is the capacitance in farads. This method produces valid readings until the capacitor plates begin to operate as transmission lines, but this occurs only in extremely low-inductance designs where ESL < 10 nH.

2. MACROSCOPIC DIELECTRIC PHENOMENOLOGY

2.1 Dielectric Polarization **P**

An electric capacitor consists essentially of a dielectric medium sandwiched between a pair of metallic electrodes, usually in a parallel-plate configuration, that can be provided with charges of equal magnitudes but of opposite signs. Being supplied by an outside source and unattached to the dielectric medium, these electrode charges are referred to as "free" or "real." It will be shown that they constitute the source of the so-called *displacement field* **D** across the dielectric.

When the field of the electrode charges is applied to a dielectric medium, the positively charged nuclei of the constituent atoms experience a displacement along the field direction and the negatively charged electrons in the opposite direction, so that both signs of charge contribute in the same sense to the charge displacement. If we designate the amount of charge displaced—or charge "flux"—across a unit area normal to the direction of the displacement by the *polarization vector* **P**, and the resulting polarization charge density by ρ_{pol}, conservation of charge requires that

$$\text{div}\mathbf{P} = -\rho_{\text{pol}}. \tag{27}$$

As was already evident from Eq. (5), the greater the permittivity, ϵ, the greater the charge that can be stored on the electrodes for a given potential difference. The underlying physical reason is that the polarization charges induced on the surface of the dielectric material are of opposite sign to the charges on the electrodes, and thus reduce the potential difference.

2.2 The Induced Bulk Dipole Moment $\mathbf{M}$

The polarization vector $\mathbf{P}$ represents the dipole moment induced per unit volume, or the "induced dipole-moment density" (Landau and Lifshitz, 1960). Thus the total dipole moment induced in volume V of the dielectric is given by the volume integral

$$\mathbf{M} = \int_V d\tau\, \mathbf{P}. \tag{28}$$

2.3 The Displacement Field $\mathbf{D}$

From Eq. (27) the total charge density can be written as

$$\rho = \rho_{\text{free}} + \rho_{\text{pol}} = \rho_{\text{free}} - \operatorname{div}\mathbf{P}. \tag{29}$$

In the mks system, Maxwell's fundamental equation defining the electric field $\mathbf{E}$ as the field of *all* the charges takes the form

$$\operatorname{div}\mathbf{E} = \rho/\epsilon_0, \tag{30}$$

where ϵ_0 is the vacuum permittivity of Eq. (6), introduced for convenience of the mks units. Combining Eqs. (29) and (30), we obtain

$$\rho_{\text{free}} = \operatorname{div}(\epsilon_0\mathbf{E} + \mathbf{P}) \equiv \operatorname{div}\mathbf{D}. \tag{31}$$

Evidently the free charges constitute the source of the displacement field $\mathbf{D}$, defined by

$$\mathbf{D} \equiv \epsilon_0\mathbf{E} + \mathbf{P}. \tag{32}$$

In the cgs system, the corresponding relation takes the form

$$\mathbf{D} \equiv \mathbf{E} + 4\pi\mathbf{P}. \tag{33}$$

2.4 The Dielectric Permittivity ϵ

To obtain the connection between $\mathbf{P}$ and $\mathbf{E}$ for dielectric materials at low field strengths it is customary to eliminate $\mathbf{D}$ through the linear constitutive relation*

$$\mathbf{D} = \epsilon\mathbf{E}, \tag{34}$$

where ϵ is the permittivity. However, for high field strengths, the dielectric medium is no longer linear, and it is more useful to define the (nonconstant) permittivity ϵ as the differential coefficient,

$$\epsilon(\mathrm{E}) = d\mathrm{D}/d\mathrm{E} = \epsilon_0 k(\mathrm{E}), \tag{35}$$

where the *relative* permittivity k is defined as the ratio ϵ/ϵ_0 of the dielectric permittivity ϵ to the permittivity ϵ_0 of the vacuum. Although generally a function of E, in the low-field regime this ratio is independent of the field strength, and is called the dielectric constant k:

$$k \equiv \epsilon/\epsilon_0. \tag{36}$$

In the cgs–Gaussian system at low field strengths, $\mathbf{D}$ is simply proportional to $\mathbf{E}$ through the dimensionless dielectric constant k:

$$\mathbf{D} = k\mathbf{E}. \tag{37}$$

Table 1 lists the low-field relative permittivity k and approximate breakdown strength of common capacitor dielectric materials at low frequency and 25 °C.

2.5 Thermodynamics of the Capacitor System

A capacitor may be considered as a "closed" system to which the combined first and second laws of thermodynamics apply during the charging process. If we denote by w_{el} the electrical work done on the capacitor during charging, the Gibbs free energy G has the differential (cf. Guggenheim, 1950)

$$dG = V\,dP - S\,dT + \delta w_{\text{el}}, \tag{38}$$

where, as usual, V is volume, P pressure, S entropy, and T absolute temperature. Thus, for the charging process at constant pressure and temperature,

$$dG = \delta w_{\text{el}}. \tag{39}$$

For a parallel-plate capacitor containing a dielectric slab of area A and thickness l, the work done on the system in transporting charge dQ from the negative to the positive electrode across the interelectrode separation distance l against an existing field $\mathbf{E}$ (normal to the electrode plates) is

$$\delta w_{\text{el}} = \mathrm{E}l\,dQ = V\,dQ, \tag{40}$$

---

*For nonisotropic media the permittivity is a symmetrical three-dimensional tensor, for which generally $\mathbf{D}$ is parallel to $\mathbf{E}$ only along the principal axes. However, for liquids and amorphous solids, such as high polymers above their glass transition temperatures, as well as for crystalline materials of sufficiently high symmetry, the tensor ellipsoid reduces to a sphere so that the scalar relation (34) applies, at least for low field intensities where ϵ is constant.

Table 1. Typical properties of common dielectric materials (at 25 °C).

Medium	Relative permittivity, ϵ/ϵ_0	Breakdown strength (MV/m)^a
Vacuum	1	
Air (STP)	1.000 583	3.0
Polytetrafluoroethylene (Teflon)	2.0	80
Polypropylene	2.2	400
Polyethylene	2.2	140
Paraffin	2.25	
Polystyrene	2.5	180
Polycarbonate	2.7	300
Silicone oil	2.5	30
Polyester (Mylar)	3.2	300
Polyimide (Kapton)	3.3	275
Glass (Pyrex)	4.0	30
Castor oil	4.5	20
Bakelite	6.0	20
Kraft paper (impregnated)	6.6	230
Mica	7.8	120
Aluminum oxide (Al_2O_3)	8.0	
Polyvinylidenedifluoride	11.0	450
Tantalum oxide (Ta_2O_5)	27.0	
Water	80	10
Barium titanate ceramics	500–10 000	20

[a] Values of measured breakdown strengths are highly variable since these are sensitive not only to inhomogeneities and minor impurities but also to sample preparation and duration of exposure to the field.

where V is the potential difference (voltage) between the capacitor electrodes. Since $\mathbf{D}$ is the field of the real charges residing on the capacitor electrodes, its corresponding increment in magnitude is

$$d\mathsf{D} = dQ/A. \tag{41}$$

In what follows it will be convenient to regard the dielectric medium as uniform as well as isotropic. Thus any extensive thermodynamic property, such as the total energy, entropy, or free energy, can be obtained simply as the product of its *specific* (i.e., per unit volume) value—indicated by an overbar—with the total volume of the dielectric medium. Accordingly, combining Eqs. (40) and (41), the increment $d\bar{G}$ in the (Gibbs free) energy density is

$$\begin{aligned} d\bar{G} &= \mathsf{E} l A d\mathsf{D}/lA \\ &= \mathsf{E} d\mathsf{D} \\ &= [\epsilon(\mathsf{E})/\epsilon_0]\epsilon_0 \mathsf{E} d\mathsf{E} \\ &= k(\mathsf{E})\epsilon_0 \mathsf{E} d\mathsf{E}, \end{aligned} \tag{42}$$

where $k(\mathsf{E})$ is the relative permittivity of Eq. (36), generally a monotonically decreasing function of the field strength E.

The characteristics of bulk dielectric materials described above translate into the technical parameters of the manufactured capacitors. The connection between the potential-dependent capacitance C of Eq. (2) and the field-dependent relative permittivity $k(\mathsf{E})$ of Eq. (35) is evidently given by

$$C = dQ/dV = Ad\mathsf{D}/ld\mathsf{E} = (A/l)\epsilon_0 k(\mathsf{E}). \tag{43}$$

2.6 (Free-)Energy Storage

One of the principal applications of capacitors—particularly large capacitors—is the storage of electrical energy and its rapid delivery at high power. If the capacitor were ideal—i.e., a pure, voltage-dependent capacitance $C(V)$ with no conductivity and zero inductance—the Gibbs free energy G stored would be the integral of the voltage V times the charge increment dQ. But since from Eq. (2) $dQ = CdV$,

$$G = \int_0^V V\,dQ = \int_0^V C(V)V\,dV = \langle C\rangle V^2/2, \tag{44}$$

where $\langle C\rangle$ represents the average capacitance over the voltage range 0–V. Evidently if the capacitance has the constant value C_0 independent of voltage,

$$G = C_0 V^2/2. \tag{45}$$

In many applications the energy rating of a capacitor is not of primary concern but rather is determined only by the capacitance and working voltage required. In other cases, such as metal forming by high magnetic fields, the energy delivered to the work piece is critical. In low-energy applications, some capacitors may contain only a few microjoules of energy, while at the other extreme large capacitors can store up to 100 kJ or more—hence their designation as "energy-storage capacitors."

3. MOLECULAR* DIELECTRIC PHENOMENOLOGY

3.1 The Polarizability Tensor

Dielectric materials can be divided into two general classes, polar or nonpolar, depending upon whether or not they possess a permanent electric dipole moment in their ground state.

To describe the behavior of dielectric materials in capacitors, four types of fields must be distinguished:

1. the field **D** of the "real" charges on the electrodes;
2. the field due to the polarization **P**;
3. the electric intensity **E**, being the macroscopic force field averaged over the dielectric medium; and
4. the "local" (or effective) microscopic field **F**.

At the molecular level, the local field **F** that acts upon a molecular element of the dielectric medium is a self-consistent superposition of the externally applied field due to the charged capacitor electrodes, and the fields contributed by all the other charges in the medium, including those induced by the field **E**.

The response of an atom or dipolar group to the local field can be described in terms of its polarizability α, which gives rise to an induced dipole moment μ,

$$\mu=\alpha\mathbf{F}, \tag{46}$$

and polarization,

$$\mathbf{P}=n\alpha\mathbf{F}, \tag{47}$$

where n is the molecular number density. Since under some circumstances the induced moment μ is not parallel to **F**, the polarizability α is generally a (symmetrical) tensor whose operation on the vector **F** changes its direction, magnitude, and units into those of μ.

Three types of response contribute to the polarization:

1. *the electronic polarizability* α_e—measured by the dipole moment resulting from the induced displacement, under the local electric field, of the negative electrons relative to the positively charged atomic nuclei;
2. *the orientation polarizability* α_μ —resulting from the partial alignment of the permanent bond dipole moments along the local electric field; and
3. *the distortion polarizability* α_d —resulting from the bending and stretching of the chemical bonds within the molecules that tie together atoms bearing different formal charges, due to their different electronegativities.

Both the electronic and distortion polarizabilities may depend upon the orientation of the molecule or substituent group relative to the direction of the applied electric field, each being described by its respective polarizability tensor.

3.2 The Clausius–Mossotti Equation

In the cgs–Gaussian system the four fields **E, D, P,** and **F** are connected through the three equations (33), (37), and (47), so that a fourth equation is required for their complete determination. For low field strengths this is provided by the Lorentz–Lorenz field derived for a spherical cavity in the dielectric (Debye, 1945):

$$\mathbf{F}=\mathbf{E}+(4\pi/3)\mathbf{P}. \tag{48}$$

This leads to the Clausius–Mossotti equation,

$$(k-1)/(k+2)=(4\pi/3)n\alpha, \tag{49}$$

where α is the sum of the three polarizabilities:

$$\alpha=\alpha_e+\alpha_\mu+\alpha_d\,. \tag{50}$$

3.3 The Electronic Polarizability α_e

Since electrons in molecules have binding energies comparable with those of the constituent atoms but have masses thousands of times smaller, only they can respond to the high frequencies of visible light—which, of course, is itself generated by electronic oscillations. This provides a means of determining the electronic polarizability. For nonmagnetic materials the dielectric constant k reduces to the square of the refractive index n (not to be confused with the molecular number density used elsewhere):

---

*The subject of electrodynamics is plagued by a multiplicity of unit systems. The electromagnetic technology of capacitors, motors, etc., now usually employs the mks system, which has been used in the preceding section. In contrast, the physics and chemistry of molecules, atoms, and subatomic particles still generally employ the cgs–Gaussian system, which will be followed in this section. The main difference lies in the presence or absence of factors of 4π and ϵ_0 in various formulas.

$$k=n^2, \tag{51}$$

so that the Clausius–Mossotti equation (49) can be written in terms of the molar refraction R, defined by

$$R=V(n^2-1)/(n^2+2)=(4\pi/3)N_A\alpha_e, \tag{52}$$

where V is the molar volume and Avogadro's number $N_A=6.022\times10^{23}$. The electronic polarizabilities α_e^s for various atomic species s can then be expressed in terms of the corresponding atomic refractions, R_s:

$$\alpha_e^s=(3/4\pi)R_s/N_A=0.396R_s\ (\text{Å}^3), \tag{53}$$

where R_s is in cm^3.

Values of the atomic refractions are approximately additive and are found to be practically independent of molecular structure. For a polymer, the electronic polarizability α_e can thus be estimated in terms of the sum of the atomic refractions R_s of the individual atoms composing the typical monomer unit in the polymer film.

For calibrating the intuition, a useful limiting case is that of a perfectly conducting sphere of radius a, for which the electronic polarizability is simply the cube of the radius: $\alpha_e=a^3$ (Joos, 1951).

3.4 The Orientation Polarizability α_μ

The electronic polarizability is reasonably constant up to very high field strengths and is rapidly reversible with essentially no hysteresis or other dielectric loss, but in polar media contributes only a small fraction of the total polarization. By contrast, the rotation of the permanent bond dipoles towards alignment with the applied field involves relative and cooperative movement of the atoms in the dielectric medium, and thus encounters not only a kind of viscous resistance but also the opposing thermal agitation which tends toward randomized orientation. At very high field strengths these permanent dipoles approach complete alignment or "saturation," so that any further polarizability is reduced to the electronic contribution α_e.

Moreover, when the applied field is removed, the permanent dipoles do not "spring" back (as do in the electrons) to their original unaligned orientations but rather are much more gradually disaligned by the thermal agitation; as with the previous alignment, the relative atomic motions required in this disalignment encounter a viscous resistance that may give rise to some hysteresis. It is desirable to minimize hysteresis, since it leads to energy dissipation and heating, which may impose limits on the number and repetition rate of charge–discharge cycles of the capacitor.

At sufficiently low temperatures a polymer may exist in a quasicrystalline state, with a correspondingly rigid structure that strongly resists rotation of the bond dipole moments; upon heating above the "glass transition temperature," the crystallinity disappears in favor of a glasslike structure, which allows greater rotational freedom and thus increases the dielectric permittivity. It is therefore desirable to select a dielectric material whose glass transition temperature lies below the normal operating temperature of the capacitor.

In some dipolar materials the interaction between neighboring dipoles is sufficiently strong to maintain mutual alignment even in the absence of an external field (see DIELECTRIC PROPERTIES OF INSULATORS). When the preferential alignment between neighbors is parallel the material becomes ferroelectric; when antiparallel, antiferroelectric. In extreme cases, ferroelectric crystals can exhibit bulk dipole moments and are known as electrets. In capacitor technology ferroelectric materials can exhibit large hysteresis and correspondingly large dielectric loss.

The energy u of the interaction of a permanent dipole moment μ with the effective local field $\mathbf{F}$ depends upon the relative orientation according to the scalar product*

$$u=-\mu\cdot\mathbf{F}. \tag{54}$$

---

*Strictly speaking, **F** as used in this section is only that part of the local field that affects the dipole orientation. In a given orientation, the dipole induces in the surrounding medium a charge distribution whose "reaction field" acts back on the dipole. If this reaction field is parallel to the dipole and follows its reorientation, it plays no part in orienting the dipole itself.

By convention, the direction of the dipole moment vector $\boldsymbol{\mu}$ extends from the negative (tail) end to the positive (head) end, so that the orientation of lowest energy corresponds to $\boldsymbol{\mu}$ parallel to $\mathbf{F}$. The reference zero of the orientation energy evidently corresponds to $\boldsymbol{\mu}$ perpendicular to $\mathbf{F}$. If the only orientation-dependent energy of a dipolar group were its interaction, Eq. (54), with the local field $\mathbf{F}$, one could write the Hamiltonian function H_{F} for a typical dipole of moment $\boldsymbol{\mu}$ as

$$H_{\mathrm{F}} = H_0 - \boldsymbol{\mu}\cdot\mathbf{F}, \tag{55}$$

where H_0 is the Hamiltonian for the field-free case. This form is not quite equivalent to assuming that the dipoles are mutually independent, since the effective field $\mathbf{F}$ at one dipole contains contributions from the fields of all the other dipoles in its vicinity. In the absence of the external field there is no preferred direction against which to measure the orientation of the dipole, so that the relative orientation of $\boldsymbol{\mu}$ with respect to any specific direction is random. We can thus write the change in chemical potential (i.e., in the Gibbs free energy per molecule) upon applying the external field in terms of the ratio of the respective molecular partition functions q as

$$-\beta\bar{G}/n = \ln(q_{\mathrm{F}}/q_0), \tag{56}$$

where

$$\beta = 1/kT \tag{57}$$

and

$$\begin{aligned} q_{\mathrm{F}}/q_0 &= \oint d\Omega\, e^{\beta\boldsymbol{\mu}\cdot\mathbf{F}} \Big/ \oint d\Omega \\ &= \frac{1}{2}\int_0^{\pi} d\theta \sin\theta\, e^{\beta\mu\mathrm{F}\cos\theta} \\ &= (\sinh\beta\mu F)/\beta\mu\mathrm{F}. \end{aligned} \tag{58}$$

The average energy $\bar{E}/n$ per dipole is thus

$$\begin{aligned} \langle u\rangle &= -\langle\boldsymbol{\mu}\cdot\mathbf{F}\rangle \\ &= \bar{E}/n \\ &= \partial(\beta\bar{G}/n)/\partial\beta \\ &= -\mu\mathrm{F}(\coth x - 1/x) \\ &= -\mu\mathrm{F}\mathscr{L}(x), \end{aligned} \tag{59}$$

where the Langevin function (Mayer and Mayer, 1940) is defined by

$$\mathscr{L}(x) \equiv \coth x - 1/x, \tag{60}$$

with

$$x \equiv \beta\mu\mathrm{F}. \tag{61}$$

At the two extremes of small and large x, the Langevin function takes the values

$$\mathscr{L}(x) \rightarrow \begin{cases} (x/3)(1 - x^2/15 + \cdots) & (x<1) \\ 1 \text{ asymptotically} & (x \gg 1). \end{cases} \tag{62}$$

In terms of the polarizability α_μ, the average energy is given by

$$\langle u\rangle = -\alpha_\mu \mathrm{F}^2. \tag{63}$$

From Eq. (59) we thus obtain the polarizability as

$$\alpha_\mu = \langle\boldsymbol{\mu}\cdot\mathbf{F}\rangle/\mathrm{F}^2 = \mu\mathscr{L}(\beta\mu\mathrm{F})/\mathrm{F} = \beta\mu^2\mathscr{L}(x)/x. \tag{64}$$

For low field strengths ($\beta\mu\mathrm{F} \ll 1$), this becomes

$$\alpha_\mu = \beta\mu^2/3. \tag{65}$$

Dipole moments are commonly expressed in terms of the debye unit (D) of 10^{-18} esu cm (or 3.3356×10^{-30} C m). Thus the transfer of one electron charge (4.8×10^{-10} esu or 1.6×10^{-19} C) through a distance of 1 Å (10^{-8} cm or 10^{-10} m) yields a dipole moment of 4.8 D (or 1.6×10^{-29} C m).

3.5 The Distortion Polarizability α_d

The restoring forces that oppose stretching (or compressing) of chemical bonds are generally much stronger than those that oppose bond bending. Accordingly, most of the distortion of the molecular framework by an electric field comes from modifying bond angles rather than bond lengths. Because the formal charge that resides on an atom varies with its environment in the molecule, there is no general theory for the distortion polarizability of molecules akin to that for the electronic polarizability. Although the distortion polarizability is generally relatively small and therefore often neglected for molecules in which the atoms are covalently bonded, certain ionic lattices—notably of the perovskite structure—can exhibit large distortion polarizabilities and correspondingly enormous dielectric permittivities (see DIELECTRIC PROPERTIES OF INSULATORS).

3.6 Dielectric Saturation

A high permittivity can result from a high density of polar groups only if the dipoles are sufficiently mobile and mutually independent to be able to rotate into alignment with the field. That this is not always the case has been shown by the analysis of the dielectric constant of water (Kirkwood, 1939), in which the intermolecular hydrogen bonds maintain an approximate tetrahedral geometry of nearest neighbors. Lack of a fundamental theory that permits reliable estimation of the energies of various conformations, the energy barriers toward reorientation and the local field F requires resort to experimental measurements in each case.

Insofar as the torque on a dipolar group within the dielectric medium is dominated by the local field **F**, dielectric saturation is approached when the dipole orientation energy $\mu \cdot \mathbf{F}$ becomes comparable with the thermal energy kT; i.e., when

$$\mathsf{F} \approx kT/\mu. \tag{66}$$

Thus, the larger the dipole moment, the lower the internal field F at which saturation occurs.

3.7 Ferroelectric and Piezoelectric Materials

An extreme case of dielectric saturation is the phenomenon of ferroelectric behavior exhibited by several classes of capacitor dielectrics, e.g., certain high polymers and ceramics (see DIELECTRIC PROPERTIES OF INSULATORS).

Dielectric materials are generally electrostrictive. This means that under an imposed field the dielectric undergoes a mechanical deformation; in the simplest case, this deformation is proportional to the square of the applied electric field. In nonpolar organic dielectrics, such as polypropylene, kraft paper, polyester, etc., this effect is extremely small. However, highly polar organic polymers and the inorganic barium titanate have well-characterized piezoelectric properties. For obvious reasons piezoelectric behavior is best minimized in capacitors.

3.8 Dielectric Breakdown Strength

Besides a high dielectric permittivity, the other dominant requirement of the dielectric material is a high electrical breakdown strength. Although the decline in permittivity at high fields means that the energy density no longer increases quadratically with field strength, it nevertheless continues to increase monotonically with the applied electric field, so that to achieve high energy densities it is desirable to use the highest field possible. But the applied field is limited by the breakdown strength of the dielectric.

The development of a breakdown "avalanche" starts with the insertion or creation of a free electron in the medium, possibly by cold-cathode emission from a promontory on the negative electrode or by ionization of an atom in the dielectric material produced by the passage of a radioactive disintegration product (α, β, or γ particle) or a cosmic-ray muon. The free electron, accelerated in the applied field, suffers collisions with both the nuclei and other electrons of the atoms and molecules in its path. Because of the great mass disparity, collisions of electrons with nuclei are essentially elastic and serve only to deflect the path of the electron with little loss in its kinetic energy, just as a tennis ball rebounds from a wall collision with little loss in kinetic energy. Also, as long as the electron kinetic energy lies below the energy of the first excited state of the atom (or molecule) with which it collides, the free electron cannot lose energy since the bound electrons are forbidden to absorb energy by the Pauli exclusion principle; that is, the electrons of the target atom have no higher quantum state available at that energy. However, once the kinetic energy of the free electron exceeds the energy of the first excited state, it can excite an atomic electron and lose a corresponding amount of kinetic energy. Unfortunately, this excitation threshold energy is usually not far below the ionization energy of the atom, so that there is a high probability that such a collision will now produce a second free electron. Thus when the field is high enough, the process of electron multiplication continues until the resulting avalanche of electrons causes an arc—that is, dielectric breakdown—between the electrodes. This arc carries a very high current, and acts as a conducting "short circuit" between the capacitor plates.

3.9 Breakdown Mitigation

Two approaches are commonly used to avoid or mitigate avalanche formation:

1. *prevention* (or, at least, deferring to higher fields) of electron multiplication through the agency of a "getter" molecule whose high electron affinity serves to capture and bind the electron in an ion of essentially zero mobility, thus rendering it incapable of further multiplication;
2. *cure* of the breakdown arc by cutting off its source of current through local vaporization and oxidation of a small electrode area surrounding the arc termini. This severs electrical conduction to the remainder of the charged electrode and thus "quenches" the arc. For the metallized electrodes commonly used in capacitor manufacture, only a small fraction of the stored charge is short-circuited by the arc, and the capacitance is reduced only in proportion to the area of the electrode vaporized. For this reason, this process is referred to as "self-healing failure" (de Rosa *et al.*, 1968).

A kind of chemical self-healing occurs on the anodes of electrolytic capacitors (Sec. 4.5), which can reform their dielectric oxide coating when this is temporarily ruptured by cracking or arc formation.

4. TYPES OF CAPACITORS

4.1 The Great Diversity of Capacitor Types

Capacitors range in size from the microscopic elements of integrated circuits (of characteristic picofarad capacitance) to very large units (~farads) used, for example, in pulsed-power supplies for electromagnetic launchers and high-energy laser systems. Their various applications span a similarly wide range: from solid-state electronics, through sensors, computers, communication systems, radio, television, and missile guidance systems, to Marx generators, public power conditioning, etc. The current catalog of a leading electronics supplier devotes 23 pages to discrete capacitor circuit elements produced by a dozen manufacturers, listing several hundred capacitor types per page; a further 40 pages are devoted to electronic power supplies, of which capacitors constitute an integral part.

Rather than attempt to encapsulate the wide range of capacitor features that support this great diversity, the present description will focus on the relatively few generic components, structural geometries, electrical characteristics, and practical engineering considerations.

4.2 Capacitor Structure and Geometry

All capacitors share the common structure of a dielectric medium sandwiched between two electrodes. Apart from the materials used and the enormous range in sizes, the principal differences in capacitors lie in the structures employed to pack the largest capacitance in the smallest volume. To this end, capacitors of small capacitance for low-voltage operation are typically constructed in the form of stacked flat multilayers of alternating dielectric and electrode sheets. Large capacitors such as those used for high-density energy storage and rapid-discharge, pulse-power applications are almost invariably constructed in a cylindrical geometry by winding together (from separate rolls) dielectric film interleaved with metal electrode foil. Cylindrical windings made on sufficiently large-diameter mandrels can be subsequently flattened to reduce volume and allow more efficient packaging.

4.3 The Electrodes

Capacitors with foil electrodes can employ a variety of metals, but for long shelf life and operational lifetimes it is desirable to use metals that resist corrosion. Noble metals—e.g., silver and gold—would satisfy this requirement, but cost considerations favor the much less expensive aluminum. Although aluminum metal is very electropositive, it has the remarkable property of forming in air a very thin transparent oxide coat that is strongly adherent and protects the underlying metal from further oxidation (at least in low humidity). Typical aluminum foils in use range in thickness from 8 to 12 μm. Such foils are readily available commercially, are very stable, and are easy to wind in the manufacturing process. Special care is required, however, to ensure good electrical connection of the capacitor terminals to these large-area aluminum electrodes.

For some applications the metal foil is replaced by a thin metal coating, ~100 Å thick, that is laid down directly on the dielectric surface by vapor deposition or sputtering "metallization." Depending upon the capacitor application, it may be necessary to control the thickness (i.e., resistance) of such metallized electrodes within narrow limits in order, on the one hand, to avoid excessive I^2R heat generation (and energy loss) in rapid discharge, and, on the other, to limit the current and quench the arc when a local breakdown occurs (see Sec. 3.8).

The selection of the electrode system is usually based on voltage and peak current or peak power specifications. High-voltage capacitors almost always have foil electrodes to minimize internal resistance.

4.4 Dielectric Materials

Although capacitors can be made with any type of dielectric medium—vacuum, gas, liquid, or solid—the common premium on low-space requirements usually dictates the use of a solid. Thin flexible films combine large energy densities with convenience in manufacturing (e.g., winding cylindrical capacitors).

A common classification of capacitors is by the type of dielectric employed.

Paper. Kraft paper, made from the cellulose in wood pulp and produced in a very fine, homogeneous tissue of high purity, is still used in many capacitor types, although it is steadily being supplanted by synthetic plastics. Usually several layers of thin paper are used instead of a single thicker sheet to avoid possible overlap of any random defects (ion impurities, pinholes, etc.). The natural moisture content of the cellulose must be removed prior to final packaging of paper capacitors.

Mica. Minerals of the mica group have monoclinic crystal structure and are readily cleaved along the basal plane into thin sheets that are flexible and elastic and have excellent mechanical and electrical properties. The dielectric constant is about 8, with little variation in the radio-frequency range. Mica also exhibits a very low loss. For these reasons mica was one of the earliest materials used as a dielectric, and remains today a mainstay for small electronics capacitors.

Reconstituted mica, a paperlike material made from ground mica platelets, has a dielectric constant of about 6, and can be wound together with metal foil into large cylindrical capacitors using the same manufacturing processes as for paper or plastic film

High-Polymer Plastic Films. Many of the high polymers developed by the chemical industry—usually for textile fibers—have been found suitable as dielectric materials. Because of the mechanical requirements of flexibility these are generally linear polymers, as contrasted with cross-linked two- and three-dimensional structures—which tend to be more brittle. They thus fall into the class of "thermoplastics," which have a softening temperature around 135 °C. Although the polymer chains are linear in a topological sense, they are usually kinked and highly entangled in films that are either extruded or melt-cast. It is therefore customary to subject newly formed film to uniaxial or biaxial stretching prior to metallizing and winding. Stretching generally improves the strength and electrical properties, but may impart a "memory" that causes later shrinkage, especially at temperatures approaching the softening point.

Of all the commercially available dielectric films, polyvinylidene-difluoride and its copolymers have the highest permittivities (see Table 1). This is a consequence of the relatively large difference in electronegativities of the carbon atom (2.5) and the fluorine atom (4.0), which leads to a C–F bond dipole moment of nearly 2 D. An important contributing factor to the dipole-moment density is the small volume occupied by these low–atomic-number elements (6 and 9, respectively).

Ceramics. The early ceramic capacitors were generally made of low-loss, low-permittivity materials such as Steatite. Steatite has a dielectric constant of about 8 with an excellent dissipation factor (<0.001). These capacitors have good thermal and voltage stability and can be operated at 500 volts or more.

Research into other metal-oxide ceramics led to the use of titanium dioxide ($k \sim 80$) for a period, but the formulation of barium titanate mixtures having dielectric constants upwards of 5000 (Tareev, 1975) sparked the widespread use of ceramic capacitors today. Research into metal-oxide semiconductor devices has produced dielectric constants approaching 200 000.

The most common method for producing ceramic capacitors is called the tape casting process wherein the dry ceramic powders are mixed into a liquid organic binder to form a slurry the consistency of cake batter. This slurry is then metered onto a moving belt coated with a polymer film such as polypropylene. The thickness of this "green" ceramic material is controlled by an adjustable gate known as a "doctor blade." The material is then dried to a pliable condition that can be cut into the desired configurations. These cut sheets are then fired in a furnace at about 1300 °C. The firing process is critical in that too low a firing temperature yields a porous material with low capacity and high dissipation factor, while too high adversely affects the capacitor temperature coefficient and also increases the dissipation factor.

After firing, the electrode is patterned onto the ceramic. Several electrode materials are used, but a common one is silver. The ceramic chip is then refired at about half the initial firing temperature to "set" the electrode. These chips can be arranged in series or parallel to obtain higher voltage or capacitance. Leads are then attached, and the assembly is either potted in epoxy or encased in a specialized housing.

Water as Dielectric. Water capacitors are used in large pulse-power systems where rapid discharge and high voltages are involved and particular geometries are desired—usually in the final stages of pulse compression. Typically these capacitors have a coaxial geometry and are best analyzed as pulse-forming lines rather than pure capacitances.

Low-conductivity water is used as the dielectric medium, but the key requirement is a short time during which voltage is applied. For this reason, water capacitors are generally charged by the discharge of a Marx generator containing film-type capacitors. The breakdown field in a coaxial water capacitor has been empirically determined to depend on the area A of the electrodes and the pulse width t according to the equation

$$\mathrm{E}_B(\mathrm{MV/cm}) \approx b/[t(\mu\mathrm{s})]^{1/3}[A(\mathrm{cm}^2)]^{0.06}, \tag{67}$$

where b (≈ 0.23) is a constant that depends on the polarity of the center electrode and the electrode surface condition. Roughening of the electrode surface and applying a prepulse at a relatively low voltage have been found to increase breakdown strength by contributing to the formation of a space charge that makes the buildup of the field at the electrodes more gradual.

Impregnating, Potting, and Canning. Low-voltage dc capacitors are often unimpregnated or "dry," and wrapped with a heavy insulating film capped at each end with epoxy—a design known as "dry wrap and fill." For low-ac-voltage ratings the paper or film, wound in a roll, may be impregnated or partially impregnated, the excess liquid impregnant being drained off, after which the capacitor is potted or canned. Since capacitance depends sensitively upon electrode configuration and spacing, for maximum stability it is desirable to "freeze" these parameters in fixed geometry. For small capacitors this is readily accomplished with standard methods of "potting" in plastics that solidify into rigid structures.

Larger capacitors, particularly those subject to frequent high-field-stress and/or thermal cycling, may develop voids from delamination or cracking. To ameliorate any attendant degradation of capacitance or breakdown strength, most capacitors designed for high-voltage (>500 V) operation are impregnated under vacuum with an insulating dielectric liquid.

The highest voltage ratings are obtained by connecting individual capacitor windings in series within the container and/or by arranging multiple electrodes within a single winding to create series capacitors. Parallel connection of windings is frequently used for large capacitance values as well.

The final operations in capacitor manufacture involve providing the internal connections to the capacitor terminals and sealing in an appropriately configured container ("canning"). For efficiency in packaging, cylindrically wound capacitors are often flattened. Depending on the application and requisite can capacitance, several subunit capacitors can be incorporated in a single can, with either series or parallel connection.

4.5 Electrolytic Capacitors

Electrolytic capacitors are similar in structure to conventional electrostatic capacitors in that they also contain two electrodes separated by a dielectric medium. However, in contrast with capacitors in which the metallic electrodes are deposited ("metallized") onto the dielectric, electrolytic capacitors utilize metal-foil anodes—originally of aluminum, but now increasingly of tantalum—upon which a dielectric oxide film is formed by an electrolysis process. It is this electrolysis process that gives this type of capacitor the generic name "electrolytic": it is not necessary for electrolytic capacitors to contain an electrolyte, although most types do. Prior to electrolysis, the anodes are put through an etching process designed to increase their surface area by several-hundred-fold. This process is of critical importance for the selection of the capacitance–voltage (*CV*) product, described below.

The electrode metals—aluminum or tantalum—are chosen on the basis of the desirable properties of their oxide films, particularly their uniformity, adherence to the metal substrate, and dielectric characteristics. The oxide film thickness is determined by the "forming voltage" used in the electro-oxidation process. To avoid altering the oxide-film properties, the working voltage of the resulting capacitor is generally set between 25% and 90% of the forming voltage. Table 2 lists the characteristics of the aluminum and tantalum oxide films.

The anodized foils are separated by an insulating material to prevent direct mechanical contact between successive layers. This separator is typically porous and acts as a reservoir to hold the electrolyte. The electrolyte is the "real" cathode of the electrolytic capacitor, the cathode foil serving only to make contact to the terminal of the capacitor.

Tantalum has several advantages over aluminum. The higher dielectric constant of Ta_2O_5 (Table 2) imparts a volumetric improvement ratio of about 3 over equivalent aluminum electrolytics. Being much less electropositive, tantalum is relatively inert and does not release hydrogen gas by electrolysis of the electrolyte. Moreover, whereas aluminum electrolytics may require "reforming" after extended disuse, tantalum oxide layers are far more stable.

Table 2. Properties of electrolytic metals and their oxide films.

Metal	Oxidation Potential (*V*)	Oxide	ϵ/ϵ_0	Thickness (Å/V)
Aluminum	1.5	Al_2O_3	8	13.5
Tantalum	0.7	Ta_2O_5	27.6	17

As with aluminum, the thickness of the tantalum oxide layer is increased by the application of higher forming voltages, which translate to higher rated voltages in the finished capacitors. This anodizing process must be optimized, however, since as the oxide layer increases in thickness it tends to bridge and seal off pores in the anode structure, resulting in dramatic losses in capacitance.

Both aluminum and tantalum have the desirable feature of "self-healing" (cf. Sec. 3.9) in that any cracks or separation of the oxide layer from the underlying metal are immediately repaired through reaction with the electrolyte.

Leakage of the liquid electrolyte has been a common problem, especially for aluminum electrolytics in which gas may be formed by reaction of the aluminum metal with the electrolyte. To obviate this problem Bell Telephone engineers pioneered in substituting solid manganese dioxide (MnO_2) for the electrolyte solution. This material has good conductivity and is a strong oxidant, which preserves the self-healing property.

There are two main types of tantalum capacitors marketed today, solid and wet. The solid version utilizes a "solid electrolyte" such as the manganese dioxide described above, whereas the wet tantalum electrolytics employ a liquid electrolyte such as a solution of sulfuric acid, H_2SO_4. Figure 7 schematically illustrates the components of a "wet" powder anode capacitor.

4.6 Chemical Double-Layer "Battery" Capacitors

A new type of electrolytic capacitor is currently being developed by several university and industrial concerns. These are the chemical double-layer capacitors, often referred to as "battery capacitors," based on a double-layer charge-storage principle (Boland, 1985; Yoshida *et al.*, 1986; Rose, 1988; Evans, 1990; Boos and Argade, 1991). The

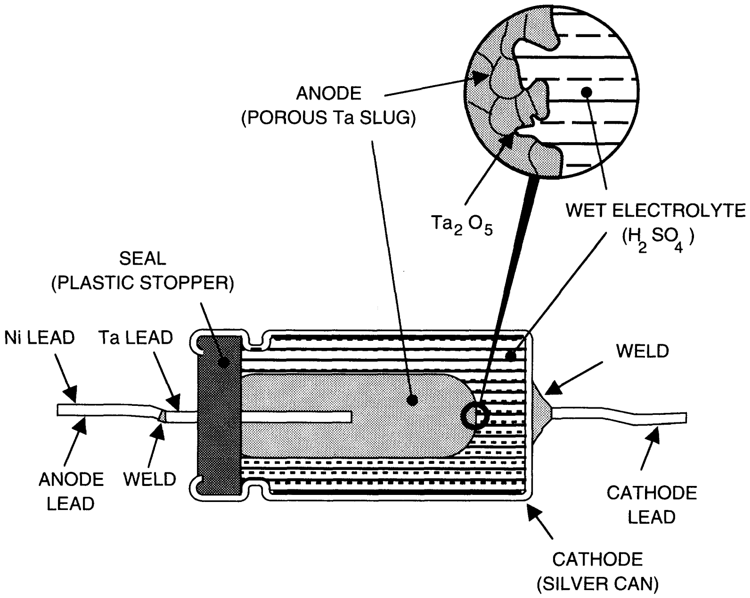

FIG. 7. Components of a "wet" powder anode capacitor.

capacitance of this device derives from the potential difference across the double layer (von Helmholtz, 1879) that is formed by polarization of the electrolyte in the field at the surface of the electrodes. When two of these double layers are brought together with interposed electrolyte, typically sulfuric acid, equal but opposite charges can be stored on the interfaces. Because of the high surface areas of these interfaces, extremely high capacitance values are obtainable, resulting in high volumetric energy densities.

Advanced constructions have been reported in the references given above. Several groups are investigating the use of these "battery" capacitors in pulsed-power applications, such as components for electric vehicles.

5. CAPACITOR APPLICATIONS

Capacitors are used for a broad range of circuit applications—filtering, blocking dc currents, coupling or decoupling components, frequency tuning, timing, energy storage, etc.; the list is virtually endless. This section will touch on a few illustrative uses of capacitors but must be understood to be by no means comprehensive.

5.1 Energy-Storage Capacitors

Energy-storage capacitors are designed primarily for use in high-power electrical applications that require delivery of high-power pulses over specific times. The amount of energy to be stored may range from a few joules to greater than 50 kJ. Typically, the larger energy-storage facilities are assembled in large capacitor banks storing multimegajoules of energy that can be discharged in very short periods of time, giving extremely high bursts of power. Examples of such systems include the pulse-power conditioning capacitor bank that drives the large solid-state lasers used in the NOVA inertial-confinement fusion experiments at the Lawrence Livermore National Laboratory (Haskell *et al.*, 1981), and high-power x-ray generators developed for the U.S. Defense Nuclear Agency and the U.S. Department of Energy.

Energy-storage capacitors must be designed to withstand high voltages; in the case of pulse-shaping or "peaking" circuits, these may be as high as several megavolts. Applications exist requiring each individual capacitor to deliver in excess of 2.5×10^5 A of discharge current over several microseconds. The magnetic forces associated with such large currents must be taken into account in capacitor design.

5.2 dc Filters

dc filter capacitors are used in power supply circuits for reducing the ac ripple voltage of the rectifier. There are many types of circuits for producing dc, ranging from single-phase, half-wave rectifiers to 12-phase, double-wave rectifiers. The ripple voltage levels allowable for typical polymer film capacitors are expressed as a percentage of the dc voltage and decrease as frequency increases. For example, at 60 Hz the allowable ripple voltage level is 25% of the dc level, whereas at 1 kHz the ripple voltage cannot exceed 5% without a significant reduction in life.

dc filter capacitors are also employed in the power supplies of hard tube modulators. To prevent a drooping voltage during a pulse, the capacitor bank should appear as a source of nearly infinite current.

5.3 Timing Circuits

Timing circuits generally utilize the *RC* characteristics of the circuit to time voltage increases or decreases. During the charging process of a capacitor in an *RC* circuit, the buildup of the transient voltage V across the capacitor and the declining circuit current I follow the exponential forms

$$V/V_0 = 1 - e^{-t/\tau}, \tag{68}$$

$$I/I_0 = e^{-t/\tau}, \tag{69}$$

where V_0 is the applied voltage, I_0 ($\equiv V_0/R$) the initial current at time $t=0$, and the characteristic time unit τ is given by

$$\tau = RC. \tag{70}$$

Practically, all capacitors and resistors have losses that affect their time characteristics. Aluminum and tantalum electrolytics have relatively high dc leakage currents that appear electrically as a shunt resistance across the capacitor. This can cause an increase in the *RC* time constant, of which designers must be aware.

5.4 Electromagnetic Interference (EMI) Suppression

Feedthrough and dual-inline-pin (DIP) mount multilayer ceramic (MLC) capacitors are designed to have very low terminal inductances. This permits bypassing to ground of any electromagnetic interference superimposed on the desired signal. In military applications this is important in overcoming adversarial jamming of communications, radars, etc., in sophisticated electronic warfare.

6. WHITHER CAPACITOR TECHNOLOGY?

6.1 Industrial and Military Requirements

Many modern industrial processes—pulsed magnetic forming or welding of metals, explosive shearing, highly penetrating x-ray imaging, high-energy lasing—require delivery of repetitive high power in short pulses. Although present technology does not yet permit achieving electrical energy storage at densities comparable with chemical energy densities, the much shorter time in which an electrical pulse can be delivered makes the *power* density of capacitors highly competitive with other power sources. Moreover, in recent years rapid advances have been made in the technologies of both high–energy-density storage and high power-delivery capabilities of electrical capacitors.

These advances are expected soon to make practical several important applications, heretofore unachievable, ranging from electrical launch of military projectiles and missiles from mobile vehicles (e.g., ships, tanks, self-propelled artillery, antiballistic missile defense) to auxiliary power for rapid acceleration of electrically driven automobiles. Since the recharging of the capacitors can be conducted over the relatively long periods between pulses, this can be carried out at the much lower power levels currently available from ordinary electric generators driven by conventional combustion engines or gas turbines.

6.2 Some Critical Capacitor Technologies

To realize the projected energy and power densities required for capacitors in the foregoing applications, advances are necessary in two key technical areas: increased dielectric permittivity and improved breakdown strength.

For orientation, to achieve in an electrical capacitor an energy density 1% that of TNT would require that

$$\langle \epsilon \rangle \mathrm{E}^2/2 = \epsilon_0 \langle k \rangle \mathrm{E}^2/2 \sim 66 \ \mathrm{MJ/m^3}, \qquad (71)$$

or equivalently,

$$\langle k \rangle \mathrm{E}^2_{\mathrm{kV/mil}} = 0.93 \times 10^4. \qquad (72)$$

This is to be compared with the presently achievable energy density of about 15-fold less.

To reduce this gap, two approaches appear promising:

1. *Increase in permittivity:* In comparison with the relative permittivity of the best commercially available dielectric films (Table 1), the relative permittivity of the various mixed titanates can exceed 10 000 (Tareev, 1975)—but currently only at much reduced breakdown strengths. One challenge, then, is to learn how to develop ceramics, or ceramic–film matrices, that can withstand high electric fields. Again for orientation, an average relative permittivity $\langle k \rangle$ of 10^4 would from Eq. (71) require a field of $\mathrm{E} \approx 1$ kV/mil or 40 MV/m.
2. *Increase in breakdown strength:* Current understanding of the propagation and multiplication of electrons through collisional ionization in solids and gases and the possibility of suppressing ultimate breakdown through electron avalanche formation is in a relatively primitive state. Sulfur hexafluoride is one of the few electron"getters" that have been used successfully for this purpose.

In conclusion, the general principles outlined in this article and illustrated particularly by high–energy-density applications are relevant to capacitors of all types, sizes, and dielectric composition.

GLOSSARY

Capacitor: A device for storing electrostatic energy, consisting of two electrodes (or "plates") separated by a layer of dielectric material. The capacitor is "charged" by transporting electric charge from one electrode to the other, usually under the action of an applied potential difference, thus establishing an electrostatic field within the dielectric (Secs. 1.2 and 1.3).

Ceramic Capacitor: A capacitor whose dielectric is a ceramic material, such as barium titanate (Secs. 1.3 and 4.4).

Dielectric: A nonconducting medium (solid, liquid or gas) that can serve as the province of an electric field (Secs. 1.3, 3.4, and 4.4).

Dielectric Loss: Electrical energy input minus electrical energy recovered from a capacitor, resulting from various combinations of dielectric conduction, leakage, electrical hysteresis, etc. (Secs. 1.1 and 1.6).

Dielectric Permittivity, ϵ: The differential coefficient expressing the rate of change of the displacement field D with the electric field E: $\epsilon = d\mathrm{D}/d\mathrm{E}$. The permittivity of a *linear* dielectric is independent of the field, so that $\mathrm{D} = \epsilon \mathrm{E}$ (Secs. 1.4 and 2.4).

Dielectric Saturation: The phenomenon approached at applied electric fields that are so strong as to cause virtually complete alignment of the permanent dipoles of the dielectric medium, thus greatly reducing the permittivity (Sec. 3.6).

Dipole Moment, μ: For a neutral molecule containing an unsymmetrical charge distribution, the product of the magnitude of the charges separated with the (vector) distance of separation (Secs. 2.2 and 3.4).

Electrolytic (Capacitor): A capacitor whose dielectric medium has been formed directly on one or both electrodes by an electrolytic (e.g., anodizing) process—for instance, as aluminum oxide on an aluminum electrode (Secs. 1.3 and 4.5; Table 2; Fig. 7).

Local (Electric) Field, F: The effective electric field at the site of a molecule (or substituent group), comprising the applied external field together with the superposed fields due to all the other charges and dipole moments of the medium (Secs. 3.1–3.4).

Metallization: The process of forming a thin adherent metallic layer, e.g., by vapor deposition, on a dielectric film to serve as a capacitor electrode (Sec. 4.3).

Polarizability, α: A measure of the ability of a molecule (or substituent group) exposed to an electric field $\mathbf{F}$ to develop a dipole moment μ along the field direction, by displacement of its electrons relative to the nuclear framework and through (partial) alignment of any permanent dipole moments present: $\mu=\alpha\mathbf{F}$ (Sec. 3.1).

Polar Medium: A medium containing permanent molecular (or substituent group) dipole moments (Sec. 3.1).

Relative Permittivity, k: The ratio of the permittivity (ϵ) of a dielectric material to that (ϵ_0) of the vacuum: $k=\epsilon/\epsilon_0$. A *linear* dielectric medium has a constant k value, independent of field strength (Sec. 2.4; Table 1).

Self-Healing (Electrode): A metallized electrode that automatically quenches an arc breakdown by local vaporization of the thin metallic film thus restoring the nonconducting *status quo ante*; also, an anodized electrode of an electrolytic capacitor that automatically reforms the dielectric oxide layer if a crack or breakdown develops (Sec. 3.9).

List of Works Cited

Boland, R. C., "Double Layer Energy Storage Device," U.S. Patent No. 4,542,444, 17 September 1985.

Boos, D. L., Argade, S. D., (1991), "Historical Background and New Perspectives for Double-Layer Capacitors," presented at the International Seminar on Double-Layer Capacitors, Deerfield Beach, FL, 9–11 October (unpublished).

Debye, P. (1929), *Polar Molecules*, New York: Chemical Catalog Co.; reprint (1945), New York: Dover, 2.

de Rosa, L. A., *et al.* (Eds.) (1968), *ITT Reference Data for Radio Engineers*, 5th ed. Indianapolis: Sams, pp. 5–17.

Evans, D. A. (1990), "High Reliability Double-Layer Capacitor," in: *Proceedings 10th Capacitor and Resistor Technology Symposium, March 28–29, 1990*, Huntsville, AL: Components Technology Institute., pp. 35–39.

Feynman, R. P., Leighton, R. B., Sands, M. (1964), *The Feynman Lectures in Physics*, Reading, MA: Addison-Wesley, Vol. II, Chap. 9.

Fleming, J. A. (1910), "Electricity," *The Encyclopaedia Britannica*, 11th ed. New York: Encyclopaedia Britannica Co., Vol. IX, pp. 179–193.

Guggenheim, E. A. (1990), *Thermodynamics*, 2nd ed., Amsterdam: North-Holland, Chap. XII.

Hall, H. P. (1978), *Electronic Defense News (EDN)*, 5 January, pp. 73–74.

Haskell, D. K., Cooper, R. A., Sevigny, J. A., Merritt, B. T., Whitham, K. (1981), "Development of a High Energy Storage Capacitor for NOVA," in: C. K. Choi (Ed), *Proccedings of the 9th Symposium on Engineering Problems of Fusion Research* (2 vols.), New York: IEEE, p. 1422.

Joos, G. (1951), *Theoretical Physics*, 2nd ed., translated by I. M. Freeman, New York: Hafner, Chap. XII.

Kirkwood, J. G. (1939), *J. Chem. Phys.* **7**, 911 (1939).

Landau, L. D., Lifshitz, E. M. (1960), *Electrodynamics of Continuous Media*, London: Pergamon, Chap. II.

Mayer, J. E., Mayer, M. G. (1940), *Statistical Mechanics*, New York: Wiley, Chap. 15.

Page, L. (1935), *Introduction to Theoretical Physics*, 2nd ed., Princeton, NJ: Van Nostrand, Sec. 150.

Rose, M. F. (1988), "Performance Characteristics of Large Surface Area Chemical Double Layer Capacitors," in: *Proceedings of the 33rd International Power Sources Symposium, 13–16 June 1988*, Pennington, NJ: The Electrochemical Society, pp. 572–592.

Singer, C. (ed.), (1958), *A History of Technology*, Vol. IV: *The Industrial Revolution, 1750 to 1850* Oxford: Oxford University Press, Chap. 22.

Tareev, B. (1975), *Physics of Dielectric Materials* Moscow: Mir, Chap. 5.

von Helmholtz, H. (1879), *Ann. Phys. (Leipzig)* **3**, 337.

Yoshida, A., *et al.*, "Electric Double Layer Capacitor and Method for Producing Same," U.S. Patent No. 4,597,028, 24 June 1986.

Further Reading

Böttcher, C. J. F. (1973) in: O. C. Van Belle, P. Bordewijk, A. Rip (Eds.): *Theory of Electric Polarization*, Vol. I, *Dielectrics in Static Fields*, revised 2nd ed., Amsterdam: Elsevier.

Böttcher, C. J. F., Bordewijk, P. (1978), *Theory of Electric Polarization*, Vol. II: *Dielectrics in Time-Dependent Fields*, 2nd ed., Amsterdam: Elsevier. This and the preceding volume contain much useful information, especially on high-polymer dielectrics, not readily available elsewhere.

Buchanan, R. C. (1986), *Ceramic Materials for Electronics*, New York: Dekker. Ceramics for capacitor applications.

Chelkowski, A. (1979), *Dielectric Physics*, translated by J. Tomaszczyk, Amsterdam: Elsevier. A very readable account, including an extensive discussion of nonlinear dielectric media.

Fröhlich, H. (1958), *Theory of Dielectrics, Dielectric Constant and Dielectric Loss*, 2nd ed., Oxford: Oxford University Press. Particularly useful for its clear presentation of the various approximations to the local field.

Hughes, L. E. C., Holland, F. W. (1967), *Handbook of Electronic Engineering*, 3rd ed. Cleveland: CRC. A useful reference work emphasizing the behavior of components in various types of circuits.

Jackson, J. D. (1962), *Classical Electrodynamics*, New York: Wiley. An excellent and comprehensive text.

McClellan, A. L. (1963), *Tables of Experimental Dipole Moments*, San Francisco: Freeman. A comprehensive and eminently useful literature survey and listing of measurements of the dipole moments of thousands of molecules.

Seanor, D. A. (1982), *Electrical Properties of Polymers*, New York: Academic. Organic insulating materials and applications.

Shugg, W. T. (1986), *Handbook of Electrical and Electronic Insulating Materials*, New York: Van Nostrand. Good practical references on common dielectric materials and applications.

Terman, F. E. (1943), *Radio Engineers' Handbook*, New York: McGraw-Hill. A venerable classic antedating solid-state electronics, but still useful.

Van Vleck, J. H. (1932), *The Theory of Electric and Magnetic Susceptibilities*, Oxford: Oxford University Press. A classic treatment of the fundamentals.

Von Hippel, A. R. (Ed.) (1954), *Dielectric Materials and Applications*, Cambridge, MA: MIT Press. A comprehensive account, including relevant results of World War II radar research.

Wilson, A. C. M. (1980), *Insulating Liquids: Their Uses, Manufacture and Properties*, New York: Peter Peregrinus. A reference on the various liquid dielectrics.

ELECTROSTATICS

R. CASANOVA ALIG, *David Sarnoff Research Center, Princeton, New Jersey, U.S.A.*

	Introduction	177
1.	**Theory**	178
1.1	Concepts and Definitions	178
1.2	Laws and Equations	179
1.3	Applications	180
2.	**Sources**	181
3.	**Polarization**	182
3.1	Fundamentals	182
3.2	Applications	183
3.3	Other Polarization Phenomena	183
4.	**Atmospheric Electrostatics**	184
4.1	Global Phenomena	184
4.2	Weather	185
4.3	Sparks and Shocks	185
5.	**Fluid Electrostatics**	186
6.	**Vacuum Electrostatics**	188
7.	**Outlook**	188
	Glossary	189
	Works Cited	190
	Further Reading	190

INTRODUCTION

Electrostatics is concerned with the effects of positive and negative charges at rest. The fundamental charges are the proton and the electron. In general, for every proton, there is also an electron. These charges are equal in magnitude; the proton is positive and the electron is negative. Like charges repel; unlike charges attract.

The basic electrostatic law is Coulomb's law: Two electric charges attract or repel each other with a force that is proportional to the product of the charges and that varies inversely with the square of the distance between them.

While Coulomb's law for the force is simple and compact for a single charge source, it becomes complicated and unwieldy for sources consisting of multiple separated charges and distributions of charge. A large body of theory exists for obtaining the force from these sources; this theory is discussed in Sec. 1. The physical origins of these sources are discussed in Sec. 2. Sources surrounded by materials interact with the material to alter the forces; these interactions are discussed in Sec. 3. Many of us are most familiar with electrostatic phenomena in the atmosphere, discussed in Sec. 4. There are many applications of electrostatics in industry, both in fluids, discussed in Sec. 5, and in vacuum, discussed in Sec. 6. The subject of electrostatic applications in fluids is customarily called "electrostatics." Vacuum and atmospheric applications, while equally significant, are generally considered as topics within the subject area. Strictly speaking, the electrostatic force is fundamental to chemistry and biology as well, as it bonds atoms and molecules together; these topics are discussed in other articles.

The earliest recorded references to the electrostatic force come from the ancient Greeks. In *De magnete*, published in 1600, William Gilbert described many observations of electrostatic phenomena that distinguish electricity from magnetism. A number of electrostatic generators appeared during the next century. Electrical repulsion, conduction, and induction were observed. Benjamin Franklin set the foundations for atmospheric electrostatics in 1752 when he demonstrated the electrical nature of lightening using a kite. He is also credited with the names "positive" and "negative" for the two kinds of charge. In 1785 C. A. Coulomb confirmed the inverse square law of force and measured its magnitude. The basic theory of electrostatics was established by Faraday and Gauss and others in the first half of the 19th century. The word "electrostatic" was introduced about a century ago by Maxwell and Thomson (Roller, 1967).

The Leyden jar, the earliest capacitor, was an early reservoir of charge. In fact, the word "charge" derives from the concept of loading a reservoir, as in loading a cannon. The Leyden

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

jar was supplanted by the voltaic cell in the early 19th century, and, subsequently, electrostatics languished. Electrostatics was rejuvenated in the first half of the 20th century, beginning with the Cottrell precipitator and culminating with the introduction of xerography by Carlson.

Vacuum electrostatic devices were invented during the first half of the 20th century, and electronic tubes rapidly took on technological importance. Image formation with an electron beam in the 1920s was followed by the development of television and the electron microscope. Since the charge motion in vacuum devices is ballistic, these devices can be fully simulated by computer. With the power to simulate, the performance of vacuum devices has been vastly improved during the last half-century.

Throughout the 20th century, atmospheric electrostatics has benefitted from more extensive and accurate observations coupled with the data-handling capacity of modern computers.

1. THEORY

1.1 Concepts and Definitions

(see also ELECTRODYNAMICS, CLASSICAL)

The unit charge, the coulomb (C), consists of 6.242×10^{18} electrons or protons. In a vacuum a charge of 1 C repels a like charge, at a distance of 1 m, with a force of about 9×10^9 newtons (N), or approximately 1 million tons. In physics and industry the charges encountered are typically vastly smaller than the coulomb.

The charge is the electrical analog of the gravitational mass in mechanics. Just as raising water away from the earth against the gravitational force results in the storage of energy as potential energy, so the separation of two unlike charges, against the Coulomb force, results in the storage of energy as potential energy. Near the surface of the earth, the gravitational potential energy is mgh, where m is the mass, g is the acceleration of gravity, and h is the separation. The analogous stored electrical potential energy is $Q\phi$, where Q is the charge, and ϕ is the unit of electrical potential; the potential is frequently expressed in volts (V), a joule per coulomb. Just as the separation h is the distance between two locations, the electrical potential is the difference in potential between two locations. Thus, it is often called the potential difference.

Since the volt describes energy storage associated with the separation of charge Q from a source charge, just as gh describes the energy storage associated with the separation of mass m from the source mass (the earth), then the volt per meter is the electrical analog of g. This quantity is called the electric field. The field at a point in space is defined by all the source charges and their arrangement about the point. Then the force on a charge at that point is the charge Q multiplied by the field $\mathbf{E}$. The energy change upon moving the charge between two nearby points separated by $\mathbf{h}$ is the scalar product of the force and the vector separation of the points, i.e., $Q\mathbf{E}\cdot\mathbf{h}$, or $Q\phi$. Near the center of a parallel-plate capacitor, the electric field is uniform, just as the gravitational field g is uniform near the surface of the earth.

The capacitance measures the capacity of a system to store electric energy. It relates the stored charge Q to the potential ϕ at which it is stored; there is no familiar mechanical analog for capacitance. The farad (F) is the unit of capacitance, measured in coulombs per volt. The farad is far greater than the amount of capacitance normally encountered, e.g., the capacitance of a parallel-plate capacitor with area 1 m^2 and separation 1 cm is less than 10^{-9} F. From this example, the capacitance is seen to characterize the geometry of the capacitor, and, indeed, in some systems of units, its unit is a length.

A charge in an electric field experiences the Coulomb force. If the charge is free, i.e., the electron or proton can move, or it is attached to a small mass, the charge is accelerated according to Newton's second law and it moves. When the charge moves, the phenomenon falls outside the domain of electrostatics. In customary usage, however, slowly moving charged particles are still considered to be "quasielectrostatic" phenomena. A moving charge is observed as an electric current. In fact, the electric current is used to define the fundamental unit of electricity; this unit is the ampere, a coulomb per second. As the ampere is the fundamental unit, the coulomb is defined as an ampere second, and, thus, the coulomb assumes the awkward size noted above.

While it is widely used, the coulomb is not the only unit of electric charge. The constant in Coulomb's law relating force to charge and distance is dependent on the charge unit and the distance and force units. The units used in this article, the coulomb and the mks system, are called SI, or rationalized mks, units. However, if the Coulomb's-law constant is set equal to unity in the cgs system, the unit of charge is the statcoulomb, equal to 2.081×10^9 electrons or protons. This unit of charge is used in the electrostatic (esu) and Gaussian systems of units. Other systems of units in use include the electromagnetic (emu) and the Heaviside–Lorentz systems. Returning to SI units, the Coulomb's-law constant, 9×10^9 N m^2 C^{-2}, is frequently expressed as $(4\pi\epsilon_0)^{-1}$, where ϵ_0 is called the permittivity of free space; $\epsilon_0 = 1/(36\pi)\times 10^{-9}$ farad/m. [More precisely, the factor 9 is $(2.997\,93)^2$ from light speed.]

1.2 Laws and Equations

Electric charges occur in fundamental units on particles that can be imperceptibly small. The forces are described by Coulomb's law. For larger particles and/or larger charges, the charge is distributed in space; the distribution is characterized quantitatively by the volume charge density $\rho(\mathbf{r})$. The forces and fields associated with extended charge densities require more extensive mathematics, especially when the densities are complicated. The densities may be complicated by the absence of symmetry, because of an unusual distribution in space, by being disjoint, etc. Most of the theory in electrostatics is developed to calculate these electric fields. By contrast, in gravitation, where the force between masses is described by Newton's law, similar to Coulomb's law for the forces between charges, fewer examples of complicated mass densities occur in daily life, and so there is less need for this extensive theory; of course, all the theory developed for electrostatics is transferable to gravitation.

Following Coulomb's work, Gauss showed that the divergence of the electric field is proportional to the charge density; symbolically, in SI units, this statement is

$$\nabla\cdot\mathbf{E} = \rho/\epsilon_0. \tag{1}$$

Upon integration of this equation over volume and applying the divergence theorem from mathematics, also credited to Gauss, Gauss's law emerges. Gauss's law states that the electric flux through a closed surface is proportional to the charge enclosed by the surface. The electric flux through a plane surface in a constant field is the product of the electric field component normal to the surface and the surface area; other surfaces, or planar surfaces in regions of variable field, may be decomposed into differential plane areas, and the electric flux is obtained by integration. Thus, for a charge Q at the center of a spherical imaginary surface of radius R, the field component E normal to the surface is the same by symmetry everywhere over this surface, so that the flux is $4\pi R^2 E$, and Gauss's law yields $4\pi R^2 E = Q/\epsilon_0$, i.e., Coulomb's law. This imaginary surface is frequently called a Gaussian surface. Gauss's law is often used to find the field from symmetric charge densities; e.g., the field E immediately outside a sheet of uniform charge density σ is $\sigma/2\epsilon_0$.

In electrostatics, Faraday's law, one of Maxwell's equations, simplifies to the statement that the curl of $\mathbf{E}$ is zero because the Coulomb force is conservative. Mathematically, this implies the existence of a scalar function $\phi(\mathbf{r})$, such that

$$\mathbf{E} = -\nabla\phi. \tag{2}$$

The function ϕ is the electric potential discussed in Sec. 1.1. When Eq. (2) is inserted in Eq. (1), the differential equation

$$\nabla^2\phi = -\rho/\epsilon_0 \tag{3}$$

is obtained; it is called Poisson's equation, and when ρ is everywhere zero, it is called Laplace's equation. In engineering applications of electrostatics, the electric field is generally needed, because it provides the forces acting on charges; however, the field is a vector function, comprised of up to three components, each of which is a distinct scalar function. To obtain these functions from Eq. (1) is frequently a difficult problem. It is often easier to solve Eq. (3) for the scalar potential function, a single function, and use Eq. (2) to obtain the electric field components.

Equation (3) is not enough to specify ϕ. Also the field obtained from ϕ with Eq. (2) is insensitive to the addition of a constant to ϕ, so the point at which $\phi=0$ is arbitrary; it is often called ground, or ground potential, as the earth is generally the zero level of potential in atmospheric electrostatics. To specify ϕ, boundary conditions are required. Specification of ϕ on a closed surface is a suitable boundary condition, frequently called the Dirichlet problem. Specification of the derivative of ϕ normal to a closed surface is also a suitable boundary condition, called the Neumann problem. The problem is said to be closed or open depending on whether ϕ is to be determined inside or outside, respectively, the boundary. Interior and exterior are often used instead of closed and open, respectively.

Materials in which charge moves freely are called conductors. ϕ is constant throughout a conductor. A region enclosed by a conductor that contains no charge is also at constant ϕ and the field is zero.

1.3 Applications

Electrostatics is governed by the theory outlined above. Its application to specific problems in technology requires skillful use of a variety of mathematical tools. Prior to computers, the principal tool was analytic solution of differential equations. Frequently a problem similar to the specified problem is solved. The solvable problem generally has symmetries that allow Eq. (3) to be reduced from a partial differential equation in three space variables to a set of up to three ordinary differential equations. Often a symmetry is associated with a constant boundary potential on a surface defined by a constant value of a coordinate, e.g., on a spherical surface where the radius is constant in the spherical coordinate system. Thus, unusual coordinate systems are frequently encountered in the analytic solutions of electrostatic problems. Other techniques, such as the method of images and Green's functions, are also used. Solutions to these solvable problems are an important guide in applications of electrostatics (Smythe, 1967).

During the last two decades, numerical computers have vastly extended the domain of electrostatic problems that can be solved. All numerical methods are based on the division of the problem into a large number of small parts. For a closed Dirichlet problem with no free charges, a simple assignment is a rectangular mesh that covers the region interior to the closed surface on which the potential is specified. The potential that fulfills Laplace's equation at a mesh point is the arithmetic mean of the potentials on the neighboring mesh points. Since the potential is specified on the boundary enclosing the mesh, it is apparent that the potential at every mesh point is defined through a set of linear algebraic equations equal in number to the number of mesh points. Sets of thousands or more equations are rapidly solved with modern computers. This method is called the finite-difference method. More difficult electromagnetic problems, mainly problems outside the domain of electrostatics, may require a similar, but more general, method called the finite-element method; of course, the finite-element method, which is widely available, may be applied to all electrostatic problems (DiStasio, 1979; McAllister, 1985; and Kashy, 1990).

The finite-difference method is adequate for most interior problems. Open, or exterior, problems, where the solution extends to infinity, however, are difficult because the computer admits only potentials defined at specified finite locations. Open problems are better done with the division of the boundary surface into a large number of surface elements, frequently called discretization elements. When a surface charge density is assigned to each of these discretization elements, the potential and field everywhere outside the boundary can be calculated with Coulomb's law, so that the difficulty at infinity is avoided. The specified potentials on the boundary surface are used to assign a charge density to each discretization element. This method is frequently called the charge density method (Cendes, 1990). The charge densities can also be recessed behind the boundary surface, resulting in another method called the charge simulation method. When the discretization elements associated with these recessed charge densities are contracted, resulting in isolated charges behind the boundary surface, this method reduces to the well-known method of image charges.

While the mathematical tools in use prior to computers generally required only dexterity with differential equations and algebra, numerical methods require access to computer hardware and, generally, a library of software. While necessary, software to simply solve the problem is frequently not sufficient for a useful solution. Most electrostatic problems require preprocessor software to compose the problem for computer solution and postprocessor software to exhibit the solution.

Composition is required because the computer requires the division of the problem into a large number of small parts, each of which is specified by a set of numbers. As every change of geometry in the problem requires a new set of numbers, preprocessors are needed to generate these numbers. Exhibition beyond listing the potentials and fields at a few points is generally needed to grasp the implications of a solution. Exhibition is done with postprocessors using data visualization techniques, such as contours of the same potential, called equipotential plots, and as electric flux plots. Flux plots contain field lines that show the field direction with arrows and field magnitude by their density.

2. SOURCES

On the largest scale, the universe is believed to be electrically neutral. On the microscopic scale, atoms and molecules are also neutral. Thus charge cannot be generated, only separated. To generate positive and negative charges, electrons must be removed from atoms and molecules, i.e., atoms and molecules must be ionized. Ions and electrons may move about by attachment to other particles or by conduction in certain materials.

Phenomena resulting in ionization include electrical, mechanical, thermal, and chemical processes. Electrical ionization generally occurs near sharp points; hence, the charge transfer it causes is frequently called corona discharge. On mechanical contact between different materials, electrons pass between them, and the electrons are carried away when the materials are separated. Thus, some materials become charged when rubbed with other materials; this is called triboelectricity. As thermal energy is added to atoms and molecules, some of them lose electrons, a process called thermionic emission. Chemically, a battery separates electrons from materials; a metallic conductor is used to direct these electrons to other locations, and eventually back to the battery. Because these electrons rarely escape the conductor, they are not generally considered sources of electrostatic phenomena. There are also other ionization processes, such as ionization by light and radiation. In the remainder of this section, electrical, mechanical, and thermal ionization processes will be discussed.

Corona discharge (see SPARKS, ARCS, AND OTHER ELECTRIC DISCHARGES) results from the ionization of gaseous atoms and molecules in large electrical fields. These fields generally occur near sharp metallic points. Electrical forces on the electrons and nucleus of an atom are in opposite directions, and when the forces are sufficiently large, an electron is removed. In a corona discharge the electric field near a sharply pointed electrode maintained at a high potential relative to its neighborhood ionizes gas atoms which subsequently stream away to charge nearby surfaces. This gas stream can be so intense that it is felt as a wind. When the gas is air, blue light accompanies the ionization, as well as the pungent odor of ozone, audible noise, and electrical static. This light is historically known as St. Elmo's Fire, a spectacular glow from the mast and rigging of large sailing vessels seen upon the approach of an electrical storm. In the dark it is awesome and foreboding. This blue light can also sometimes be seen near high-tension (-voltage) electrical transmission lines. In fact corona discharge limits the power transport on these lines, for it limits the voltage used in them. Corona discharge is also responsible for brush discharge, a discharge characterized by visible streamers. Electrical ionization near sharp metallic points also occurs in vacuum, where electrons emerge directly from the metal. This is called field emission (see FIELD EMISSION AND IONIZATION), a widely used electrical source in electron microscopes.

Triboelectricity is the exchange of charge by different uncharged materials brought into mechanical contact and then separated. It is the earliest manifestation known, dating back to Thales of Miletus (640–546 B.C.), who reported the charging of amber (called elektron in Greek, from which Gilbert introduced the word "electricity" in 1600) upon rubbing with other dry substances. The charged amber attracts other substances and so has similarity to a magnet. Indeed, the separated charges in amber render it an electret, the electrical analog of a magnet. Triboelectricity is the source of many familiar manifestations of electrostatics, such as static cling of garments made from modern synthetic materials, shocks after walking in rugs, the attraction of hair and paper to a plastic comb, etc.

The charge appears to be transferred by friction or rubbing. In fact, it is transferred at the locations of contact between the materials, and charging occurs when the contact is broken and the materials separate. Enhanced charging associated with additional rubbing occurs because more locations come in contact, not from the greater work of rubbing. At the locations of contact, the charge transfer occurs between nearby energy levels at the material surfaces. While some of these energy levels are intrinsic to the material bulk and surface, others are extrinsic to defects and adsorbed impurities. These latter levels restrict reproducibility and quantification of the triboelectric phenomenon.

Thermionic emission (*q.v.*) follows the addition of heat to solids. Some electrons accumulate sufficient thermal energy to escape from the solid. As the solid must be heated to very high temperatures, thermionic emission generally is used to obtain charged particles in enclosed systems, viz., in the absence of oxygen, such as in vacuum tubes.

The charging of an object frequently occurs by induction, historically called influence. Induction is the redistribution of charge in a conductor; charge distributions are discussed in Sec. 3. When it is followed by removal of net charge, the object becomes charged. The electrophorus uses this principle. When small particles separate from a pointed electrode, as in ink-jet printing, they are charged inductively.

3. POLARIZATION

3.1 Fundamentals

As described in the Introduction, electrostatics is concerned with the effects of positive and negative charges; sometimes these charges are locally separated, resulting in dipoles. Polarization is concerned with the effects of dipoles. The fundamental dipole $\mathbf{p}$ is a pair of positive and negative charges of equal magnitude Q separated by a vector $\mathbf{r}$ directed from the negative to the positive charge; thus, $\mathbf{p}=Q\mathbf{r}$. For a larger number of charges, or a distribution $\rho(\mathbf{r})$ of charges, $\mathbf{p}=\Sigma Q_i\mathbf{r}_i$ or $\mathbf{p}=\int\rho(\mathbf{r})\mathbf{r}\, d^3r$, the first, or dipole, moment of the charge distribution. Dipoles have units of charge-distance, and as many molecules are dipolar, the dipole moment 3.33×10^{-30} coulomb meter$=0.393qa_0$, is frequently called the Debye, where q is the fundamental charge and $a_0=0.0529$ nm is the Bohr radius of the hydrogen atom; the magnitude of the Debye is a consequence of its origin in esu units. The name dipole comes from the "two pole" source of magnetic fields, and, implicitly, the dipoles described here are electric dipoles.

The concept of the dipole depends on local separation of charges, i.e., the field is evaluated far from the dipole. As the field of a charge depends inversely on the square of the distance from the charge, so the field of a dipole depends inversely of the cube of the distance from the dipole. As a charge is a scalar quantity, the field around it is isotropic; as a dipole is a vector quantity, the field around it depends on the angle between the dipole and the vector to the field point. A second dipole a distance away may experience this field; the interaction energy depends on the angles formed by the dipole vectors with their separation vector, as well as the separation between the dipole vectors.

A distribution of dipoles generally adopts a configuration to minimize the dipole energy. An electric field $\mathbf{E}$ reconfigures the dipole distribution to produce a polarization field $\mathbf{P}=\chi\epsilon_0\mathbf{E}$, where χ is the electric susceptibility of the dipole distribution, i.e., of the material. This property of the material makes it a dielectric. For a material to be dielectric, χ must be positive. Polarization has units of dipole moment density, i.e., from the definition the units are farads/meter times volts/meter, or coulombs/meter^2, i.e., coulomb meter/meter^3.

Returning to the concept of the single dipole, generally at the atomic or molecular level, the local electric field $\mathbf{E}_l$ can modify its moment. This modified dipole moment is $\mathbf{p}=\alpha\mathbf{E}_l$, where α is the polarizability. The macroscopic susceptibility is related to the microscopic polarizability by the macroscopic relation $\mathbf{P}=\chi\epsilon_0\mathbf{E}$ from the preceding paragraph and the microscopic relation $\mathbf{P}=N\alpha\mathbf{E}_l$, where N is the volume density of dipoles. The ratio of $\mathbf{E}_l$ to $\mathbf{E}$ is a geometric characteristic of the material. As the density of the material decreases, this ratio approaches unity, so $\alpha=\chi\epsilon_0/N$. Dielectric polarization arises from electronic polarization, the distortion of the electron distribution of an atom or molecule; atomic polarization, a change in molecular geometry; orientation polarization, a change in the angles between molecules; and space-charge polarization, the movement of aggregations of charge.

The sum of the polarization **P** and the electric field is frequently defined as the electric displacement vector **D**, i.e., $\mathbf{D}=\epsilon_0\mathbf{E}+\mathbf{P}$. Then $\mathbf{D}=\epsilon\mathbf{E}$, which defines the permeability ϵ of the material (see DIELECTRIC PROPERTIES OF INSULATORS). (Sometimes **D** and **E** are not parallel and ϵ is then a tensor.) The ratio $\epsilon/\epsilon_0=1+\chi$ is the relative dielectric constant of the material. By analogy to the magnetic flux density vector **B** of magnetostatics, the electric displacement is also called the electric flux density. By the same analogy the electric field **E** is the analog of the magnetic field **H**, and the polarization **P** is the analog of the magnetization **M** (see MAGNETS AND MAGNETOSTATICS).

The polarization in materials is generally defined by the relative dielectric constant ϵ/ϵ_0. When this ratio exceeds unity, the electric field in the material is reduced. Mathematically, this reduction arises because the response to a charge distribution ρ is described by **D**; i.e., Maxwell's equation for electrostatics is $\nabla\cdot\mathbf{D}=\rho$ in a dielectric. Physically, the polarization **P** and the source charges ρ are both charge distributions, the polarization, of course, being a dipole distribution. A reduced field in a volume is very useful for storing charges, e.g., in a capacitor, and for shielding charges, i.e., for electrical insulation.

3.2 Applications

A capacitor, previously known as a condenser, is an electrical device used to store charges, generally equal amounts of both positive and negative charge. Capacitors, along with inductors, resistors, and transistors, are components of electronic circuits (see CIRCUIT ELEMENTS). On electrical-appliance circuit boards, they are frequently visible as cylinders or encapsulated packages. As mentioned in Sec. 1.1, the unit of capacitance is the farad (after Michael Faraday); this unit is larger than the capacitance of commonly used capacitors, and so the capacitance will often be indicated in picofarads [pF or (archaic) $\mu\mu$F, both of which are 10^{-12} F] or nanofarads (nF, 10^{-9} F). These capacitors generally consist of alternating layers of metal foil and dielectric film or paper. Alternate layers of the foil contain the positive and negative charge, and the film reduces the field between these charges so the capacitor can be made physically smaller. Other capacitors for circuit boards, and for more general applications, may use liquid or gas for the dielectric material. Integrated circuits contain many capacitors for which the semiconductor is the dielectric material. Capacitors are frequently used to store energy for brief pulses of high power, as in a camera flash device or an automotive ignition condenser (see ELECTROSTATIC CAPACITIVE STORAGE OF ENERGY). For sustained power, chemical batteries can store 10 to 100 times more energy per unit volume than capacitors. Capacitance properly belongs in the domain of electrostatics only when charge is stored; when the capacitor is charged or discharged, as frequently occurs in circuits, phenomena outside electrostatics occur.

Insulators are electrical materials used to maintain the separation of positive and negative charges, i.e., to shield or isolate charges. Usually these charges are moving in wires, and so are perceived as electric currents. Electrical insulation is frequently encountered as a wrapping for electrical wires or as ceramic structures for supporting electric wires and objects. Electrostatically, insulation extends the range of fields for which electrostatic conditions prevail. However, at sufficiently high fields, positive and negative charges no longer remain separated. The resultant disruption of the insulation is called electrical breakdown, and the charge transport is frequently perceived as short luminous glow, such as sparks and lightning, which are discussed in Sec. 4. Insulators emphasize another property of dielectric materials, called the dielectric breakdown strength, i.e., the electric field at which the material breaks down; a typical dielectric breakdown strength is 100 kV/cm.

3.3 Other Polarization Phenomena

In materials the polarization is generally dielectric, i.e., it is produced by an electric field. However, it may also be produced by an applied stress or pressure, called piezoelectricity. Piezoelectricity also includes the converse relation of strain to an electric field. Piezoelectric materials are used to transform, or transduce, electrical and acoustic signals, as in microphones and loudspeakers (see PIEZO-

ELECTRIC DEVICES). Quartz is a widely used piezoelectric material. For small stresses the polarization is linearly proportional to the stress, but the relation is geometrically complicated, as the stress tensor T is frequently expressed as a six-component vector, so, as the polarization is a vector, the polarization is related to the stress by the 18 piezoelectric moduli d, i.e., $P_j=\Sigma_{i=1}^{6} d_{ij}T_i$.

Polarization may also be produced thermally, called pyroelectricity. Pyroelectricity is used in thermometers and radiation detectors, as in a bolometer. Tourmaline is a common pyroelectric material.

Even when the polarization is produced by an electric field, the susceptibility may be negative, in which case the material is paraelectric.

Polarization differs from magnetization in the existence of free electric charges which eventually reconfigure to cancel the polarization; there is, of course, no analogous magnetic charge. There are some materials with polarizations that last for many years, called electrets, used in commercial applications; many of these are organic polymers. In addition, the stable state of some other materials is characterized by a polarization field; these materials are ferroelectric by analogy to ferromagnetism (see FERROELECTRIC DEVICES). These materials, e.g., Rochelle salt, have many of the characteristics of a ferromagnet, e.g., a Curie temperature, domains, hysteresis in the polarization response to an electric field, and a residual polarization. Of course, free charges in the air collect to cancel the residual polarization, so that our common experience with permanent magnets does not have an electric analog.

4. ATMOSPHERIC ELECTROSTATICS

4.1 Global Phenomena

The atmosphere is the layer of air surrounding the earth. The atmosphere and earth together are electrically nearly neutral, but the earth is charged negative while the atmosphere is positive, each with about 500 000 coulombs. The earth is a good conductor, and so this negative charge is distributed nearly uniformly over its surface. While the atmosphere is a poor conductor, the ionosphere that encloses the atmosphere is a good conductor, so that the outer boundary of the atmosphere is at an electric potential that is nearly constant, typically 300 000 V positive relative to the grounded earth. The positive charge is distributed within the atmosphere, and so the earth, atmosphere, and ionosphere form a spherical capacitor consisting of a pair of concentric conductors, i.e., the earth and ionosphere, with the earth carrying the negative charge and the positive charge distributed in the atmosphere between the conductors.

The sources of this charge, that is, its generation at the local, or even atomic, level, are not well known, and there may be several significant sources (see ATMOSPHERIC PHYSICS). The separation of these charges leading to the negative charge on the earth and positive charge in the atmosphere is intimately connected with the weather. The sources of the charge include the interaction of cosmic rays with gas molecules in the atmosphere, radioactive decay of elements in the earth, ejection of charged water droplets in the turbulence of ocean waves, and triboelectrification associated with the contact of wind-blown particles with the earth surface. Cosmic rays are very high-energy charged particles that come from beyond the solar system; their flux onto the earth is independent of time and geography. They are distinct from the solar wind, a flux of charged particles from the sun that is deflected by the ionosphere. A cosmic ray is generally absorbed by the ionization of many atmospheric gas atoms. As the atmospheric density decreases with altitude and cosmic rays impinge from above, this mechanism results in an altitude of maximum charge production, and some measurements are consistent with this maximum.

Radioactive decay of elements in the earth can lead to ionization of gases trapped within the earth that subsequently escape into the atmosphere. It can also generate short-lived radioactive decay gases, such as radon, which enter the atmosphere and decay, thereby generating charged particles in the same manner as cosmic rays. (Recently, concern has been raised about radon entering the atmosphere within homes, where it may be trapped, and decay resulting in damaged human tissue.) While this source might seem to be consistent with maximum charge production near the earth, updraft of the radioactive gases confuses this conclusion.

When a bubble breaks at the ocean surface, small drops emerge that carry positive charge. These drops diffuse into the atmosphere over the ocean and the earth. On the basis of measurements over selected areas of the ocean where the waves are turbulent, evidenced by whitecaps, and on estimates of the ocean area covered by these waves, a significant portion of the positive charge in the atmosphere could originate from these drops.

4.2 Weather

The atmosphere in the global capacitor mentioned above is not a perfect insulator, i.e., it is leaky. The atmosphere has a conductivity, and if no other conductive phenomena were present, most of the positive charge would return to the earth within one hour. A phenomenon generally thought be responsible for moving positive charge away from the earth is stormy weather, e.g., thunderstorms. In fact, the quantity of positive charge in the atmosphere is proportional to the thunderstorm activity on the earth at that moment. It peaks during the hours corresponding to afternoon over the Amazon and central Africa, regions well-known for thunderstorms.

Thunderstorms are associated with lightning (indeed, thunder is the sound associated with lightning). Lightning is electrical breakdown in the atmosphere, and it occurs where the field between the bottom of the thunderstorm and the earth exceeds the breakdown field of the atmosphere. It serves to move substantial charge (1–10 coulombs) rapidly from a local region of the atmosphere to the earth. Lightning is, of course, widely observed, familiar, and dangerous, and it is the subject of a large literature (Wahlin, 1986).

In fair weather, the electric field near the surface of the earth is about 100 V/m directed toward the earth. This field generally decreases with altitude because the density of positive charge in the atmosphere generally decreases with altitude. Thus, it is consistent with a potential of 300 000 V on the ionosphere 100 km above the earth's surface. The charge density on the earth's surface is ϵ_0 times this field, or 1 mcoulomb/km^2, resulting in about 500 000 coulombs when integrated over the earth's surface. Since capacitance is the charge stored divided by the potential, the capacitance of the earth and atmosphere is near 1 F.

This electric field is subject to large variations as the weather or locale changes. This is apparent from the knowledge that much larger fields, 100 kV/cm, are required for breakdown, and breakdown is commonly observed. Not only is large-scale breakdown in the form of lightning a common phenomenon, but smaller breakdowns in the form of shocks and sparks are even more common.

4.3 Sparks and Shocks

Sparks and lightning are the same electrical breakdown phenomenon, but sparks generally describe a lower-voltage breakdown. Electric spark describes the light associated with these smaller breakdowns (see SPARKS, ARCS, AND OTHER ELECTRIC DISCHARGES). The light emission accompanies local heating of the air (indeed, this heating is the cause of the thunder associated with a lightning discharge). Heating can lead to ignition of chemical processes. Sparks are widely used for ignition, as in the spark plugs of gasoline engines, where potentials of 10 kV discharge across a 1-mm gap thousands of times per minute to ignite the fuel mixture. Unwanted sparks and ignition also occur. In some environments, such as petroleum tanks or grain elevators, sparks are a well-known hazard. In these environments, where large amounts of materials are moved through pipes into a reservoir, triboelectric charging between the material and the pipe walls occurs. Sparks occur as this charge moves in the reservoir. If conditions such as the ignition energy and the air-to-material ratio are met, heat from the spark can ignite a massive and deadly explosion.

Electric shock describes the sensation associated with electric current passing through the human body. It is sometimes felt momentarily when the electrostatic charge on the body surface (the human body is a conductor) flows away in the breakdown associated with a spark, as often occurs when one reaches for a door knob after walking across a rug. Of course, shocks are well known, and more dangerous, when the body becomes part of an electrical circuit, and these shocks clearly reveal that the intensity of a shock is related to the current. Thus, the rapidity of the discharge in a spark, as well as the quantity of charge discharged, determines the magnitude of the sensation associated with a spark. The quantity of charge that can accumulate on the body is limited by the breakdown field in air. As this field is 10^6 V/m or more, the maximum charge on a human body surface is a few μcoulombs, assuming the body surface area is about a square meter. As the body can sense currents of a milliampere or less, the discharge time is about a millisecond. Recalling the preceding discussion of sparks igniting explosions, the ignition energy of ether can be only a millijoule, corresponding to the energy $Q\phi$ released from the body in a 1-mm-long spark. Thus, sparks can be a hazard in some hospital environments.

Sparks felt as shocks are only a fraction of the body electrical discharges that occur. Many sparks, too small to feel, have catastrophic consequences for modern electrical equipment, particularly integrated circuits. When a spark is examined in detail, as in the photographic analysis of lightning, it is found to have endpoints which are highly localized, generally to local sharp points of the electrodes between which the charge flows. Thus, in an electrostatic discharge to an integrated circuit, much of the charge may flow to a single element of the circuit. The current may be large enough to melt a conductor on the circuit, resulting in an open circuit. Because the circuit elements are so close together, the charge may cause local breakdown between two nearby conductors, fusing the insulating substrate, and leaving a new conducting path, i.e., a short circuit. A discharge is not even necessary, as large currents can be induced by nearby charged bodies. These short and open circuits may not lead to obvious malfunction of the integrated circuit; instead they may result in spurious behavior that is very difficult to characterize by subsequent testing. These short and open circuits are latent defects that can distort the operation of the integrated circuit. Special care is required to ensure that people are grounded electrically when working with or near exposed integrated circuits (Boxleitner, 1988).

Atmospheric electric potential differences are generally measured with an electrostatic voltmeter. Potential differences across circuit elements are generally measured by diverting a small portion of the current through the voltmeter. However, the currents in the atmosphere are too small to be able to divert a small portion and still have measurable current. Hence, an instrument called an electrostatic voltmeter is used. It utilizes the Coulomb attraction of the plates of a capacitor to measure the potential difference between two electrodes; a part of one of the electrodes is made flexible to obtain the measurement. Thus, only the current to charge the capacitor is needed. An early version of this voltmeter, made by Cavendish, was used by Coulomb to measure the constant ϵ_0.

5. FLUID ELECTROSTATICS

The atmosphere surrounding the earth is a fluid, and it is filled with dust particles. Dust particles are frequently charged, for they adhere to TV screens, collect inside electrical appliances, etc. Lightning is seen in dust storms. Dust drifts in air and responds slowly to external forces, e.g., gravity causes dust to collect on horizontal surfaces in homes. Electrostatic forces can also direct the flow of dust and other particles, and, because electric forces can be far stronger than gravitation, the response can be faster. The electrostatic force on charged particles has been used in the past century in many applications, such as electrostatic precipitation, separation, coating, printing, etc. Particles are generated, if they do not exist already, charged, exposed to a strong field, deposited, and, if appropriate, collected.

Electrostatic precipitation was the earliest application. A precipitator removes particles from a gas, frequently air. It is generally used to clean the air and reduce pollution, although sometimes the precipitated particles have value. It is widely used to collect fly-ash from coal-burning electric plants. In operation the particles are charged with the corona discharge from wires at high voltage. These charged particles are driven by an electric field to the walls of the precipitator where they collect; eventually gravity directs them to a container for removal. The efficiency of collection is an important characteristic of a precipitator. It depends on the type of particles, e.g., large or small, to be collected, as well as on the turbulence of the gas. Electrostatic precipitators were introduced by F. G. Cottrell before 1911 and were rapidly adopted for pollution control. They are also used for dust control in smelters, acid mist control in chemical plants, etc. (Oglesby, 1978).

Electrostatic separation is used to concentrate minerals, purify foods, and sort mixed garbage. The mixture to be separated consists of small particles, and the technique is based on the different affinities of the particles in the mixture for electric charge. Typically, the particles are charged triboelectrically or by corona discharge, passed through an electric field, and collected in a series of bins. Thus, it is analogous to the use of mass and gravity in winnowing. Major applications include iron ore concentration, and separation of zircon from beach sand and of rodent droppings from grain (Inculet, 1984).

Electrostatic coating is used to apply surface coatings on conducting objects. Nearly all objects sold today are coated and much of it is done electrostatically. The coating material, a powder or aerosol, is charged, and an electric field directs it to the object to be coated. The charging is done by corona or induction. The coating is permanently affixed to the object by chemical reaction or fusing. Sometimes the coating particles are oriented by the electric field, as in electrostatic flock coating. The atmospheric field around plants can be used to direct chemical coatings. Major concerns include the transfer efficiency of the coating material and the uniformity of the coating. Sometimes nonuniformities are planned by charging only selected areas of the object to coat in a pattern; these coatings resemble electrostatic printing (Hughes, 1984).

Electrostatic printing includes the xerographic process of photocopying. In xerography a high-resolution, detailed charge pattern, e.g., a page of alphanumerics or graphic art, is coated electrostatically with a dark powder. This charge pattern is created with light on a charged photoconductive layer. Coating is done with a powder of small, charged particles; selectively coating only the detailed pattern is dependent on the dominance of the electrostatic force over other forces present. While a wide variety of photocopying techniques can be made to work in the laboratory, many of them fail in the transition to commercial use. Chester Carlson was successful in leading this transition in the last half-century to make photocopying a part of every office. Electrostatic printing also includes ink-jet printing where a fine beam of tiny charged ink particles is accelerated and directed by electrostatic fields to write as with a fine pen.

To achieve the electrostatic force in these applications, large fields, and, hence, high voltages, are needed. In most applications transformer-rectifier assemblies are sufficient, but for some very high-voltage applications above 100 kV, like particle accelerators, electrostatic generators are needed. Electrostatic generators, such as the Van de Graaff and Wimshurst machines, transport charge from a source to a collector. Electrostatic generators, being relatively safe because of the limited current they can deliver, are widely used in classroom demonstrations. However, this current limitation restricts their use in industry.

The electrostatic force cannot overwhelm other forces for all particles. The gravitational force dominates the motion of large particles—particles larger than a few mm in diameter, as charge is generally confined to their surface while gravity acts on their bulk. Aerodynamic forces dominate the motion of very small particles, as the viscosity of air causes submicron-size particles to move generally with the air flow. The motion of much smaller charged particles in a liquid due to an electric field is called electrophoresis, while the motion of similar uncharged particles due to a nonuniform electric field is called dielectrophoresis.

Not all atmospheric electrostatic phenomena contribute beneficially in technology. Adverse effects are common in the textile and plastic industries where electrostatic forces interfere with the desired motion of the materials. Antistatic agents are used to reduce these forces. Fabrics tend to cling together or balloon out, sometimes leading to such disasters as the failure of a parachute to open, but more commonly disturb the industrial handling of dry textiles and the comfort and appearance of clothing. Textile charging occurs triboelectrically, and it is exacerbated by the speed of the machines. Charging can be reduced with increased conductivity; higher humidity will generally increase conductivity. Hygroscopic materials are often used as antistatic agents, as they collect water from the atmosphere. In fact, natural fibers are less susceptible to electrostatic charges than synthetic fibers because they are more hygroscopic. Electrostatic forces in plastics lead to dust and dirt accumulation, and conducting surface treatments are frequently used to dissipate charge (Horvath, 1982).

6. VACUUM ELECTROSTATICS

Charged particles in air drift in response to the electrostatic force; i.e., their movement is the net displacement after many collisions with air molecules. Charged particles in vacuum respond ballistically to an electrostatic force; i.e., their movement can be predicted with Newton's laws of motion. Drift phenomena were discussed in Sec. 5 and ballistic phenomena will be described in this section. Charged particles, e.g., electrons, in vacuum are frequently formed into beams that are focused, reflected, and deflected in much the same way as are light beams. Thus, the study of these beams is called electron optics (see CHARGED-PARTICLE OPTICS). The lenses, mirrors, and prisms of electron optics are frequently electrostatic fields. The television display tube is a common and well-known example using vacuum electrostatics (Szilagyi, 1988).

While television is the principal consumer application of vacuum electrostatics, as the replacement of electronic vacuum tubes by solid-state devices is virtually universal, there are many laboratory, and industrial applications. These include the electron microscope (see ELECTRON MICROSCOPY), avionics and oscilloscope display tubes (see DISPLAY TECHNOLOGY), particle accelerators (see ACCELERATORS, LINEAR; ACCELERATORS, POTENTIAL DROP; CYCLOTRONS; SYNCHROTRONS), mass spectrometers (see SPECTROMETERS, MASS), electron-beam lithography (see LITHOGRAPHY), x-ray tubes, etc. All these devices rely on a source of charge, frequently thermionic or field emission, a guidance system based on electrostatic or magnetic fields, and detectors to convert the charge to a useful signal. For example, in a television tube the electron beam is emitted and focused electrostatically by the electron gun, deflected magnetically, and converted into a visual image by phosphors. An electron microscope has condenser, objective, and projection lenses, just as in an optical microscope, and many of these lenses are electrostatic. Deflection in an oscilloscope tube is generally done electrostatically because of the high speeds needed.

Electrostatic electron-optical devices, such as lenses, are shaped conducting electrodes in the vacuum enclosure. Voltages are applied to these electrodes by wires through the enclosure wall. For example, when an ion beam passes between two flat plates at different voltages, it is bent, or deflected, by the electric field, just as light is bent by a prism. Also, when different voltages are applied to flat plates with coaxial openings, a beam of electrons passing through the openings is focused. The electric field distribution in the opening acts on the electrons in the same way as the refractive index distribution in a glass lens acts on light.

The electron wavelength is so small that electrostatic optical systems can be described by geometrical optics. The analogs of geometrical-optics theory, such as paraxial-ray theory, aberrations, etc., are used to characterize electrostatic lenses. In addition, the electrons in the beam repel each other electrostatically, leading to a beam divergence that is not found in light optics; this repulsion is called space charge.

The electrostatic theory outlined in Sec. 1 is applicable to electrostatic optical systems, and modern computers permit detailed analyses of complicated systems. Because the electron and ion motion is ballistic in vacuum, complicated motions can also be simulated with computers. Even space-charge effects can be simulated. Thus, the growth of computer technology during the past half-century has resulted in refined precision and improved efficiency in vacuum electrostatic technology. This has occurred in spite of a vacuum mechanical technology (see VACUUM TECHNIQUES AND INSTRUMENTATION) that can be cumbersome and aggravating (Herrmannsfeldt, 1988).

7. OUTLOOK

Atmospheric electrostatics will continue to benefit from the growth of observations in and above the earth's atmosphere. Electronic computers will allow assimilation of this data into models of the atmosphere. Concerns about the environment will accelerate this activity, particularly concerns about the interaction of electromagnetic fields with human life (Brodeur, 1989).

Electrostatic applications in fluids will benefit from computer simulation when concepts from fluid dynamics are applied to simulate charged-particle drift in electrostatic fields. Reductions in size and cost and improvements in efficiency will lead to consumer versions of many of the existing industrial applications.

Applications of vacuum electrostatics will mature, as substitutes for vacuum devices are always attractive. These applications have already drawn tremendous benefit from computer simulations of ballistic electron motion. New applications may emerge from microtechnology in vacuum electronics (Tech, 1990).

Electrostatic theory is very well established and unlikely to change. Computer solution of electrostatic problems is also firmly established and viewed as a tool for the simulation of particle motion.

GLOSSARY

Breakdown (Electric): Electrical rupture of a dielectric.

Capacitance: Capacity of a system to store electric potential energy.

Capacitor (Condenser): A device for storing electric charges.

Charge: (noun) Quantity of electricity. (verb) To add charge to a material.

Charge Density: Distribution of charge in space. It may be volume, surface, or linear.

Corona Discharge: Corona is the faint glow on surfaces of electrical conductors exposed to high fields. Discharge is the accompanying flow of charge.

Coulomb: Unit of charge equal to 6.242 $\times 10^{18}$ electrons or protons.

Coulomb's Law: Basic law of electrostatics: The force exerted between two electric charges is directly proportional to the product of the charges and inversely proportional to the square of the distance between them.

Dielectric: A nonconducting material which can sustain electric fields and has positive electric susceptibility.

Dipole: A pair of equal and opposite charges separated by a small distance.

Discharge: Reduction of potential difference by charge transfer.

Displacement Vector: The sum of the electric field and polarization in a material.

Earnshaw's Theorem: A charge cannot exist in stable equilibrium among a group of fixed charges.

Electret: A block of dielectric material with permanent electrostatic polarization.

Electron: The most elementary charge of negative electricity.

Electron Optics: Branch of electronics that deals with beams of electrons behaving analogously to beams of light.

Electroscope: Instrument for detecting the presence of charge.

Electrophoresis: The movement of suspended particles in a fluid with an electric field.

Electrostatics: The part of the science of electricity for which the distribution of charge is static.

Equipotentials: Surfaces on which the electric potential is constant; also, lines formed by the intersection of such surfaces with specified planes.

Farad: Unit of capacitance needed to store one coulomb with a potential difference of 1 V, i.e., one coulomb per volt.

Ferroelectric: A material with abnormally high electric permeability and appreciable residual polarization and hysteresis.

Field (Electric): A vector quantity characteristic of a point in space equal to the electric force on a small charge at that point, divided by the magnitude of the charge. The unit is force over charge, or potential over distance.

Gauss's Law (or Theorem): The charge enclosed by an imaginary surface in space is equal to the integral of the normal component of the electric field over this surface.

Ground: (noun) Potential of the earth, frequently defined as the zero of potential. (verb) To connect an electric object to the earth.

Ionosphere: The layer of the earth's atmosphere 40 to 400 km above the earth's surface, containing free charged particles.

Laplace's Equation: $\nabla^2\phi=0$ where ϕ is a function defined in space. In electricity ϕ is the electric potential.

Leyden Jar: An early capacitor, developed at the University of Leyden in the 1740s, consisting of a glass jar coated with tin foil within and without and topped with a brass knob connecting to the inner foil. It is charged by applying a source of electricity to this knob.

Paraelectric: A nonconducting material which can sustain electric fields and has negative susceptibility.

ENERGETICS OF BIOLOGICAL PROCESSES

IRWIN H. SEGEL AND LEIGH D. SEGEL *Department of Biochemistry & Biophysics, University of California, Davis, California, U.S.A.*

Introduction 207
1. **General Principles** 207
1.1 Varieties of Metabolic Lifestyles 207
1.2 The Free-Energy Change (ΔG) of Biochemical Reactions 209
1.2.1 The Standard State 210
1.2.2 ΔG of Oxidation-Reduction Reactions 211
1.2.3 Which Way and How Far, But Not How Fast 212
2. **Common Energy Metabolism Pathways** 212
2.1 Glycolysis—An Anaerobic Pathway 212
2.1.1 Mechanism of ATP Generation in Glycolysis—Substrate-Level Phosphorylation (Chemical Coupling) 214
2.1.2 Gluconeogenesis—The Reversal of Glycolysis 216
2.2 Respiration—Aerobic Energy Metabolism 217
2.2.1 The Electron Transport System and the Chemiosmotic Mechanism of Oxidative Phosphorylation 217
3. **Photosynthesis** 219
4. **How Do Cells Drive Energy-Requiring Reactions?** 221
4.1 Limitations of Coupled Reactions 221
4.2 "Energy-Rich" Compounds 222
4.3 Active Transport 223
4.4 Other Types of Energy-Rich Compounds 224
Glossary 224
Further Reading 227

INTRODUCTION

Energy is the capacity to do work. The subject of *bioenergetics* deals with the energy changes accompanying the reactions by which living cells (a) mobilize and trap chemical or light energy from their environment and (b) use that energy to drive various vital processes, i.e., to do work. "Work" in a biological sense includes the biosynthesis of specialized organic compounds from simpler precursors, the formation of complex macromolecules from low-molecular-weight building blocks (for repair, growth, and cell multiplication), the "uphill" transport of substances into cells from a dilute environment, as well as physical work performed by flagella, cilia, and muscles. A complete coverage of bioenergetics would include topics that could also be discussed under headings of biophysics, thermodynamics, photochemistry, and cellular biomechanics. This article will focus on cellular energy metabolism at the molecular level. Many key compounds will be referred to by name and standard abbreviation. Readers should consult a biochemistry text for their chemical structures.

1. GENERAL PRINCIPLES

1.1 Varieties of Metabolic Lifestyles

Organisms can be broadly classified as either *phototrophs* (those which use sunlight as their primary energy source) or *chemotrophs* (those which use chemical substances for energy). Chemotrophs can be subdivided into two groups—chemoorganotrophs (which use preformed organic compounds) and chemolithotrophs (which use inorganic substances). Chemotrophs include animals, yeasts, molds, and most species of bacteria.

In broad terms, energy is made available to chemotrophs when electrons from the "en-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ergy source" are transferred in a stepwise manner to some final acceptor (Fig. 1). (The energy source originates as "food" taken from the environment.) Thus, energy metabolism in chemotrophs is based on the progress of favorable, enzyme-catalyzed reactions which include one or more oxidation-reduction steps. The overall process results in the oxidation of the "energy source" (also called the *initial electron donor* or *reducing agent*) and the reduction of the final or *terminal electron acceptor* (also called the *oxidizing agent*). Some of the energy released in this overall process is conserved by generating adenosine triphosphate (ATP) from adenosine diphosphate (ADP) and inorganic phosphate (P_i). ATP [Fig. 2(a)] is used directly or indirectly to drive many energy-requiring cellular processes (see Sec. 4). Because cells also need reducing power for many biosynthetic processes, some of the electrons of the energy source are diverted to the nicotinamide-containing coenzymes NAD^+ and $NADP^+$ [Fig. 2(b)]. The overall oxidation-reduction-coupled formation of ATP and NAD(P)H is analogous to the charging of a battery by a water wheel placed in a flowing stream. Most chemoorganotrophs are also *heterotrophs*, meaning that they use preformed organic compounds as their primary source of carbon for biosynthetic processes. Most chemolithotrophs and phototrophs (see below) are *autotrophs*, meaning that CO_2 serves as their major carbon source.

Phototrophs (higher plants, algae, and photosynthetic bacteria) also rely on oxidation-reduction reactions to generate ATP and reduced coenzymes. However, in these organisms, the initial electron source is a pigment system or "photosystem" (chlorophyll, etc.) which, upon absorbing light energy, becomes a much stronger electron donor. ATP is generated as the electrons flow back through a series of intermediary carriers to the original (or to a different) photosystem. When electrons are retained for biosynthetic reductions, another electron source is needed to fill the "positive hole" left in the original pigment system (see Sec. 3). Green plants and algae use water for this purpose (liberating oxygen as a byproduct). Some photosynthetic bacteria use other inorganic substances such as H_2S. Phototrophs use light energy to synthesize organic compounds (carbohydrates,

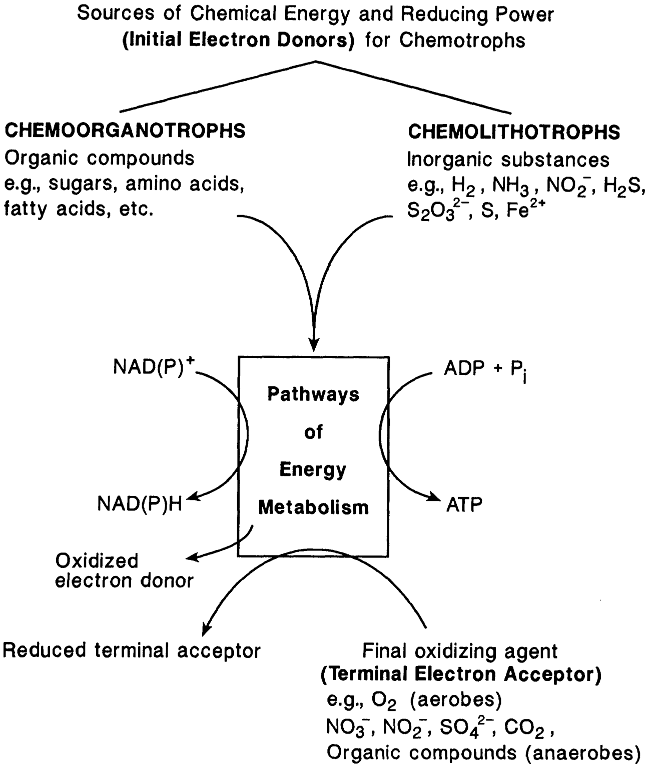
Sources of Chemical Energy and Reducing Power
(Initial Electron Donors) for Chemotrophs
CHEMOORGANOTROPHS
Organic compounds
e.g., sugars, amino acids, fatty acids, etc.
CHEMOLITHOTROPHS
Inorganic substances
e.g., H2, NH3, NO2−, H2S, S2O3 2−, S, Fe2+
NAD(P)+
NAD(P)H
Pathways of Energy Metabolism
ADP + Pi
ATP
Oxidized electron donor
Reduced terminal acceptor
Final oxidizing agent
(Terminal Electron Acceptor)
e.g., O2 (aerobes)
NO3−, NO2−, SO4 2−, CO2,
Organic compounds (anaerobes)

FIG. 1. Varieties of chemotrophic energy metabolism. Most chemolithotrophs are also autotrophs, i.e., they use CO_2 as their major source of cellular carbon. Most chemoorganotrophs are also heterotrophs, i.e., they use preformed organic compounds as their source of cellular carbon as well as for energy.

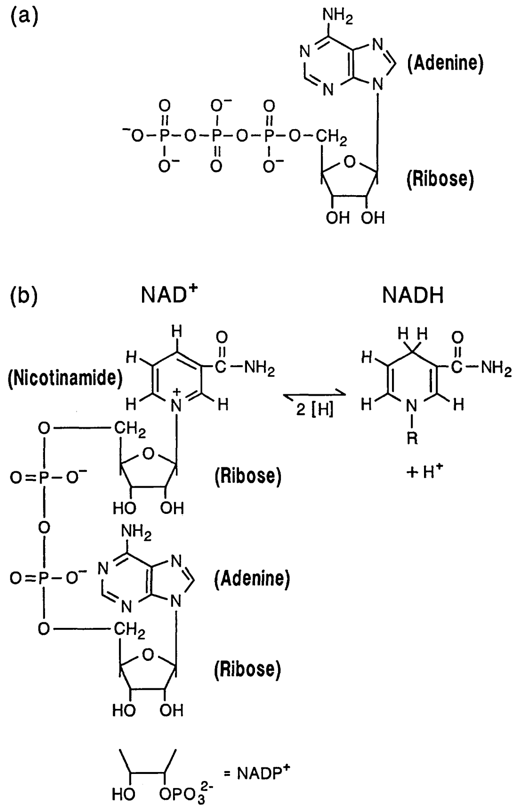

FIG. 2.(a) Adenosine triphosphate (ATP). The compound is shown in its ionized form as it exists at physiological *p*H. ATP is composed of adenine (a purine), ribose (a 5-carbon sugar), and three linearly linked phosphoryl groups. (b) Nicotinamide adenine dinucleotide (NAD^+) and its reduced form, NADH. The coenzyme (derived from the vitamin, niacin) carries two electrons, but only one complete hydrogen. Thus, for each $2e^-$ transferred to NAD^+ (as a hydride ion), an H^+ is also produced. The similar nicotinamide adenine dinucleotide phosphate ($NADP^+$) contains a phosphoryl group esterified to the OH on carbon number 2 of the adenine-linked ribose.

fats, and proteins) from CO_2. In the dark, these organisms act as chemotrophs using their stored organic compounds as energy sources and biosynthetic building blocks.

The following sections describe how biochemists deal with the above types of reactions in quantitative terms.

1.2 The Free-Energy Change (ΔG) of Biochemical Reactions

All physical and chemical processes can be described by a set of basic equations derived from the laws of thermodynamics—laws which describe the conservation and direction of energy flow (see THERMODYNAMICS). In brief, the first law of thermodynamics states that energy can neither be created nor destroyed. In any particular process, one form of energy may be converted into another (including energy into work and vice versa), but the total energy of the system plus its surroundings remains constant. The first law is simply a law of *conservation of energy*. The second law deals with the *increase in universal entropy* that accompanies all processes. One statement of this law is that any real process results in a decreased capacity of the (ever-constant) energy to perform further work. We will see the second law in action when we compare the biochemically relevant "energy yield" of certain metabolic processes with the energy required to reverse the process completely.

The thermodynamic quantity most useful for discussing biochemical reactions is the *Gibbs free energy change* ΔG which indicates the spontaneity of a process.* By convention, a thermodynamic Δ is taken as the difference between the final state and the initial state, defined as "final minus initial." Thus, a negative ΔG means that the components of the system under consideration will contain less energy after a process compared to before the process (i.e., energy will be released as a result of the process). If ΔG is negative, the process as formulated can proceed because it goes "downhill" (see below). ΔG is composed of two other thermodynamic quantities—ΔH (the enthalpy change) and ΔS (the entropy change). ΔH represents the heat liberated or absorbed by a process (a negative ΔH value means that heat is liberated). Caloric values of foods (e.g., 4 kcal g^{-1} for carbohydrates, 9 kcal g^{-1} for fats) are, in fact, the $-\Delta H$ values for the complete oxidation of these substances to CO_2 and H_2O. ΔS can be considered to be a measure of the change in randomness or disorder of reaction components as a result of a process. The relationship between the three thermodynamic quantities is $\Delta G = \Delta H - T\Delta S$. Thus, a reaction with a negative ΔG could be primarily enthalpy driven (ΔH is highly negative; heat is released), or primarily entropy driven (ΔS is highly positive), or driven by a combination of the two factors. For example, the folding of a newly synthesized linear polypeptide chain into a compact functional protein is largely entropy driven. The obvious decrease in entropy of the aminoacyl residues is more than compensated for by the large increase in entropy of water molecules that were originally highly ordered around hydrophobic (nonpolar) amino acid groups (analogous to tiny water droplets on an oily surface). Upon folding of the polypeptide chain, these water molecules return to the relatively unstructured bulk solution when the hydrophobic residues move to the interior of the protein.

---

* In classical thermodynamics, $-\Delta G$ is equivalent to (a) the maximum work that could be performed by the process and also (b) the $T\Delta S$ of the reaction system plus its surroundings.

ΔG can also be expressed in terms more useful for characterizing the progress of a chemical reaction:

$$\Delta G = RT \ln Q - RT \ln K_{eq}$$

or

$$\Delta G = RT \ln(Q/K_{eq}). \tag{1}$$

R is the gas constant (1.987 cal mole^{-1} deg^{-1} or 8.314 J mole^{-1} deg^{-1}) and T is the absolute temperature. Q is the *mass action quotient.* That is, if the reaction is $A+B \rightleftharpoons C+D$, then $Q=[C][D]/[A][B]$ at the moment, or under steady-state conditions such as exist in cells. K_{eq}, the *equilibrium constant,* is the mass-action quotient at equilibrium, i.e., at the lowest energy level of the system. At equilibrium there is no net reaction. Thus, *the sign and the magnitude of* ΔG *describe the displacement of the system from equilibrium and, consequently, the direction and extent of the reaction if it were to proceed.* For example, if $Q < K_{eq}$, (i.e., ΔG is negative), the reaction as written is not yet at equilibrium and can proceed further to the right. The more negative the ΔG, the further the reaction can proceed before attaining its equilibrium position. If $Q = K_{eq}$ (i.e., ΔG is zero), the reaction is at equilibrium and will not proceed in either direction. If $Q > K_{eq}$ (i.e., ΔG is positive), the reaction as written is beyond the equilibrium position and, if it proceeds at all, it will go in the reverse direction (which has a ΔG equal in magnitude, but negative in sign). A reaction having a negative ΔG is said to be "spontaneous," "favorable," or "exergonic" ("energy yielding" or "energy releasing"). Work (in the molecular sense) can be performed by coupling such a reaction to one having a positive ΔG (i.e., to a "nonspontaneous," "unfavorable," "energy-requiring," or "energy-absorbing" reaction).

For most biochemical applications of Eq. (1), the reaction is assumed to occur in dilute aqueous solution of constant water concentration and at a fixed *p*H (usually 7.0). Thus $[H_2O]$ and $[H^+]$ are generally not included in K_{eq} or Q expressions. For example, K_{eq} or Q for the hydrolysis of ATP at physiological *p*H ($ATP + H_2O \rightleftharpoons ADP + P_i + H^+$) is usually given as $[ADP][P_i]/[ATP]$. The numerical value of K_{eq} or Q expressed this way will vary with *p*H and water concentration.

ΔG values are more convenient than K_{eq} and Q for discussing biochemical reactions because the energy quantities are additive. For example, given two sequential reactions, $A+B \rightleftharpoons C+D$ with a ΔG_1 followed by $C+E \rightleftharpoons F+G$ with a ΔG_2, the overall reaction, $A+B+E \rightleftharpoons D+F+G$, will have a ΔG_3 equal to $\Delta G_1 + \Delta G_2$.

1.2.1 The Standard State. In order to catalog and compare different reactions, a reference or *standard state* is used. That standard is $1M$ of all reactants and products (with the exception of water and H^+). Thus, Q at the standard state is unity. The standard state ΔG is notated as $\Delta G^{0\prime}$ (the superscript zero referring to $1M$ reaction components, the prime indicating that $[H^+]$ is different from $1M$, usually $10^{-7}M$). From Eq. (1), we see that if $Q=1$,

$$\Delta G^{0\prime} = -RT \ln K_{eq}. \tag{2}$$

In effect, $\Delta G^{0\prime}$ is another way of stating the K_{eq} of the reaction;* it tells how K_{eq} compares to unity. For example, if $\Delta G^{0\prime}$ is negative, it means that $1M$ of all products is not sufficient to be at equilibrium with $1M$ of all reactants and, consequently, the reaction will proceed to the right. Combining Eq. (2) with Eq. (1) yields another form of Eq. (1):

$$\Delta G = \Delta G^{0\prime} + RT \ln Q; \tag{3}$$

at 25 °C,

---

* The minus log-linear mathematical relationship between $\Delta G^{0\prime}$ and K_{eq} is similar to that between *p*H and $[H^+]$. In both cases, the two conventions give the same information, but in different language.

$$\Delta G = \Delta G^{0\prime} + 1364 \log Q \text{ cal mole}^{-1}$$

$$= \Delta G^{0\prime} + 5706 \log Q \text{ J mole}^{-1}.$$

Discussions of biochemical reactions often use $\Delta G^{0\prime}$ rather than true ΔG values because the intracellular concentrations of reaction components (and hence, the *in vivo* Q) is not always known, but K_{eq} can be experimentally determined *in vitro.* This is not as unreasonable as it may seem. Certainly, intracellular metabolite concentrations are not $1M$ (most fall between $10^{-7}M$ and $10^{-2}M$). But note that for reactions of the type $A \rightleftharpoons P$ or $A+B \rightleftharpoons C+D$ (i.e., equimolecular in both directions), the reaction components need only to be equal in concentration for Q to equal 1. Q can also equal 1 if $[C][D]=[A][B]$ even when no single component is present at $1M$ or at a concentration equal to that of another component.

1.2.2 ΔG of Oxidation-Reduction Reactions. As mentioned earlier, many reactions of energy metabolism are oxidation-reduction reactions. Table 1 lists some compounds of biological interest and their relative tendencies to gain electrons at 25 °C and *p*H 7 under otherwise standard conditions. The numerical value of any E_0' (standard reduction potential as volts per electron) indicates the reduction potential of a half-reaction relative to that of $H^+ + 1e^- \rightarrow \frac{1}{2}H_2$, which is taken as zero under true standard conditions ($1M$ H^+, 1 atm H_2) and −0.414 V at 10^{-7} M H^+, 1 atm H_2. E_0' values may be thought of as electron pressures and, as such, they are independent of the number of electrons in the half-reaction.

Because no substance can gain electrons without another substance losing electrons, a complete oxidation-reduction reaction must be composed of two half-reactions. When any two half-reactions are coupled, e.g., $A_{ox}+2e^- \rightarrow A_{red}$ with an E_0' of −0.19 V and $B_{ox}+2e^- \rightarrow B_{red}$ with an E_0' of −0.32 V, the one with the greater tendency to gain electrons (the one with the more positive reduction potential) goes as written (as a reduction). Consequently, the other half-reaction (the one with the lesser tendency to gain electrons, as shown by the less positive reduction potential) is driven backwards (as an oxidation). Thus, the overall reaction is $A_{ox}+B_{red}$

Table 1. Standard reduction potentials of some biologically important half-reactions.[a]

Half-reaction (written as a reduction)	E_0' at *p*H 7.0 (V)
$\frac{1}{2}O_2+2H^++2e^- \rightarrow H_2O$	0.82
$Fe^{3+}+1e^- \rightarrow Fe^{2+}$	0.77
$SO_4^{2-}+2H^++2e^- \rightarrow SO_3^{2-}+H_2O$	0.48
$NO_3^-+2H^++2e^- \rightarrow NO_2^-+H_2O$	0.42
Cytochrome a_3-$Fe^{3+}+1e^- \rightarrow$ cytochrome-a_3-Fe^{2+}	0.55
Cytochrome-a-$Fe^{3+}+1\ e^- \rightarrow$ cytochrome-a-Fe^{2+}	0.29
Cytochrome-c-$Fe^{3+}+1e^- \rightarrow$ cytochrome-c-Fe^{2+}	0.25
Ubiquinone$+2H^++2e^- \rightarrow$ ubiquinone-H_2	0.10
Fumarate$+2H^++2e^- \rightarrow$ succinate	0.03
Acetaldehyde$+2H^++2e^- \rightarrow$ ethanol	−0.16
Oxalacetate$+2H^++2e^- \rightarrow$ malate	−0.18
$FAD+2H^++2e^- \rightarrow FADH_2$	−0.18[b]
Pyruvate$+2H^++2e^- \rightarrow$ lactate	−0.19
Cystine$+2H^++2e^- \rightarrow 2$ cysteine	−0.22
1,3-diphosphoglycerate$+2H^++2e^- \rightarrow$ glyceraldehyde-3P$+P_i$	−0.29
$NAD(P)^++2H^++2e^- \rightarrow NAD(P)H+H^+$	−0.32
$2H^++2e^- \rightarrow H_2$	−0.41
$CO_2+H^++2e^- \rightarrow$ formate	−0.42
Ferredoxin-$Fe^{3+}+1e^- \rightarrow$ ferredoxin-Fe^{2+}	−0.43
Acetate$+2H^++2e^- \rightarrow$ acetaldehyde	−0.60

---

[a] Standard conditions: Unit activity of all components except H^+, which is maintained at 10^{-7} M. Gases are at 1 atm pressure.

[b] The value given is for free $FAD/FADH_2$. The E_0' of the protein-bound coenzyme varies according to its environment.

$\rightleftharpoons A_{red}+B_{ox}$. The oxidized forms of those substances with highly positive reduction potentials are good oxidizing agents (and are easily reduced). The reduced forms of those substances with highly negative reduction potentials are good reducing agents (and are easily oxidized).

The relative tendency of the overall oxidation-reduction reaction to proceed may be calculated from the difference between the reduction potentials of the component half-reactions:

$$\Delta E_0' = [E_0' \text{ of the half-reaction containing the oxidizing agent}] - [E_0' \text{ of the half-reaction containing the reducing agent}]. \quad (4)$$

If the overall reaction as written is favorable, Eq. (4) yields a positive $\Delta E_0'$. $\Delta E_0'$ provides the same information as $\Delta G^{0\prime}$ except in units of volts. The relationship between $\Delta E_0'$, $\Delta G^{0\prime}$, and K_{eq} is given by

$$\Delta G^{0\prime} = -n\mathscr{F}\Delta E_0', \quad (5)$$

$$\Delta E_0' = (RT/n\mathscr{F})\ln K_{eq}, \quad (6)$$

where n is the number of electrons transferred per molecule (i.e., the number of equivalents per mole) and $\mathscr{F}$ is Faraday's constant (23 063 cal V^{-1} equivalent^{-1} or 96 496 J V^{-1} equivalent^{-1}), a factor that converts volts×equivalent^{-1} to cal×equivalent^{-1} or joules×equivalent^{-1}.

For the reaction $A_{ox} + B_{red} \rightarrow A_{red} + B_{ox}$, the nonstandard-state ΔE is given by

$$\Delta E = \Delta E_0' - (RT/n\mathscr{F}) \ln Q$$

or

$$\Delta E = (RT/n\mathscr{F}) \ln(K_{eq}/Q), \quad (7)$$

where $Q = [A]_{red}[B]_{ox}/[A]_{ox}[B]_{red}$. Thus, at any Q,

$$\Delta G = -n\mathscr{F}\Delta E. \quad (8)$$

1.2.3 Which Way and How Far, But Not How Fast. While energy considerations allow us to describe the *direction* of a biochemically relevant reaction and the *extent* to which it will proceed, nothing can be inferred from ΔG values about the *rate* of the reaction. Yet, it is the rapid rates of chemical reactions in cells that make life possible. (Clearly, a reaction can be biologically useful only if it can be completed in a time period that is very short compared to the lifetime of a cell.) For example, sucrose is a good biological energy source. The ΔG for sucrose oxidation to CO_2 and H_2O is highly negative. But a teaspoon of sucrose (~10 g) does not burst into flames before our eyes. Indeed, the oxidation rate of dry sucrose (or sucrose in sterile solution) is too slow to measure. On the other hand, a moderately active 75-kg individual can oxidize 10 g of sucrose in about 20 min. (and do so at a relatively low temperature). The agents responsible for rapid reaction rates *in vivo* are *enzymes*—protein catalysts that accelerate the progress of chemical reactions toward equilibrium. The roles of several enzymes in cellular energy metabolism are described in the following sections.

2. COMMON ENERGY METABOLISM PATHWAYS

2.1 Glycolysis—An Anaerobic Pathway

Glycolysis is a ubiquitous pathway among chemoorganotrophs for the utilization of glucose as an "energy source." Other common hexoses, e.g., fructose (derived from sucrose), galactose (derived from lactose), mannose, and the glucosyl residues of the storage polysaccharide, glycogen, feed into the pathway at appropriate points. As shown in Fig. 3, the pathway is "primed" by consuming two moles of ATP per mole of glucose. However, each glucose is split into two moles of triose phosphate, each of which yields 2 moles of ATP. Consequently, the net yield is 2 ATP per glucose. In animal cells and several bacterial species, the end product is lactate. Yeast produces ethanol and CO_2 as end products. Except for the final step(s) (which deal with the disposition of pyruvate), animal, yeast, and several other bacterial versions of glycolysis are identical. Glycolysis (as all other anaerobic fermentations) involves internal oxidation-reduction; i.e., no external oxidizing agent (terminal electron acceptor) is required. In glycolysis leading to lactate, carbon atoms 3 and 4 of glucose are oxidized (a total of 6 e^- are removed), while carbon atoms 1 and 6 are reduced (a total of $6e^-$ are gained). Note (Fig. 3) that 4 of the 6 electrons are transiently carried on NADH. The fate of each carbon atom of glucose upon conversion to 2 lactate is shown below.

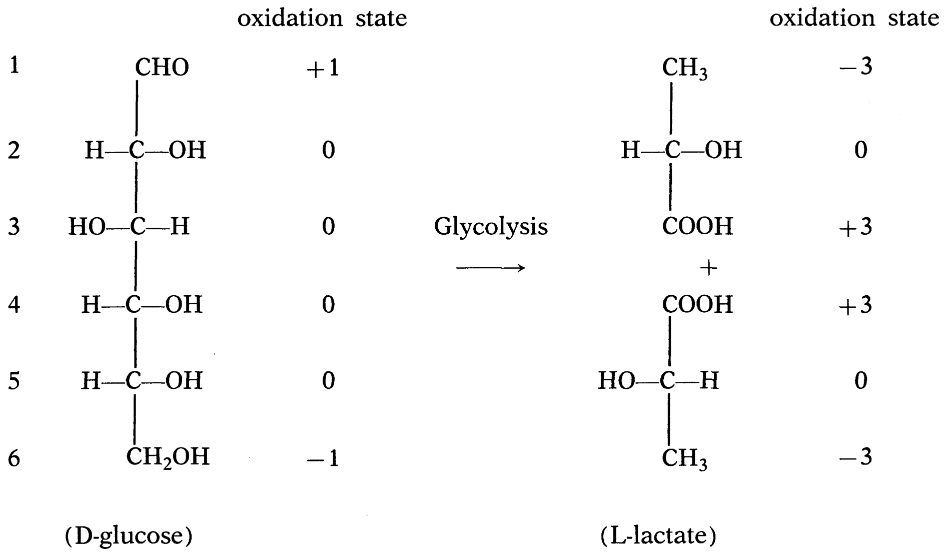

The overall conversion is highly exergonic:

$$\text{glucose} \rightleftharpoons 2\ \text{lactate}, \tag{9}$$

$$\Delta G_9^{0\prime} = -47.8\ \text{kcal mole}^{-1}$$

$$= -200\ \text{kJ mole}^{-1}.$$

The energy made available is more than enough to drive the formation of ATP:*

$$\text{ADP} + \text{P}_i \rightleftharpoons \text{ATP} + \text{H}_2\text{O}, \tag{10}$$

$$\Delta G_{10}^{0\prime} = +7.6\ \text{kcal mole}^{-1}$$

$$= +31.8\ \text{kJ mole}^{-1}.$$

In fact, two ATPs are formed for each glucose converted to 2 lactate. The overall process is

$$\text{glucose} + 2\text{ADP} + 2\text{P}_i$$

$$\rightleftharpoons 2\ \text{lactate} + 2\text{ATP} + 2\text{H}_2\text{O}. \tag{11}$$

Under standard conditions, the overall $\Delta G^{0\prime}$ is given by

$$\Delta G_{11}^{0\prime} = \Delta G_9^{0\prime} + 2\Delta G_{10}^{0\prime}$$

$$= -32.6\ \text{kcal}\ (-136.4\ \text{kJ})\ \text{mole}^{-1}.$$

Thus, even with the formation of ATP included, the overall process is highly favorable under standard (as well as under nonstandard, *in vivo*) conditions.

It should be no surprise that sustained energy (i.e., ATP) -requiring processes must be coupled to continuous energy-yielding processes (e.g., glycogen must be continuously metabolized through glycolysis to provide a continuous source of ATP for muscle contraction in an animal fleeing a predator or chasing its lunch). However, it may not be immediately obvious that the reverse is also true: Glycolysis cannot proceed for any useful period of time unless the ATP formed is used. The reason for this obligate coupling is that the intracellular levels of ADP and ATP are (of necessity) fairly low: ~0.3 and 3 μmoles g^{-1}, respectively. An actively contracting muscle may use glucosyl residues of glycogen at a rate of 60 μmoles g^{-1} min^{-1}. Thus, the original ADP pool would last less than 1 sec unless it were continuously regenerated. Similarly, NAD$^+$ (present at about 0.06 μmoles g^{-1}) must be regenerated from the NADH formed at the glyceraldehyde-3-P dehydrogenase-catalyzed step. This is accomplished by the NADH-dependent reduction of pyruvate to lactate. Figure 4 shows how the ADP/ATP and NAD$^+$/NADH cycles must mesh with the overall utilization of glucose in glycolysis.

The enzymes of glycolysis are present in the soluble fraction of cells. In general, the enzymes are considered to be free-floating residents of the cytosol. However, recent stud-

---

*The use of standard abbreviations (e.g., P_i for HPO_4^{2-} plus $H_2PO_4^-$ will make some reactions appear to be unbalanced.

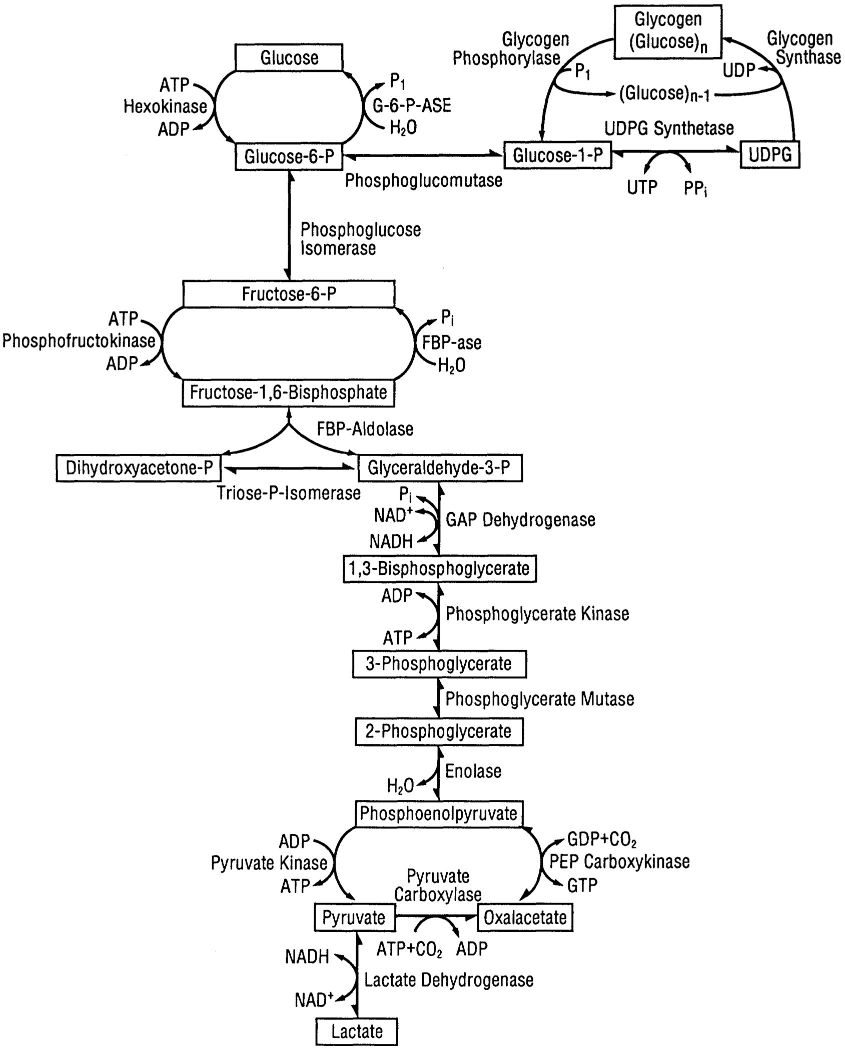
Glucose
ATP
Hexokinase
ADP
P1
G-6-P-ASE
H2O
Glucose-6-P
Phosphoglucomutase
Glucose-1-P
Glycogen
(Glucose)n
Glycogen
Phosphorylase
P1
(Glucose)n-1
UDP
Glycogen
Synthase
UDPG Synthetase
UDPG
UTP
PPi
Phosphoglucose
Isomerase
Fructose-6-P
ATP
Phosphofructokinase
ADP
Pi
FBP-ase
H2O
Fructose-1,6-Bisphosphate
FBP-Aldolase
Dihydroxyacetone-P
Glyceraldehyde-3-P
Triose-P-Isomerase
Pi
NAD+
NADH
GAP Dehydrogenase
1,3-Bisphosphoglycerate
ADP
ATP
Phosphoglycerate Kinase
3-Phosphoglycerate
Phosphoglycerate Mutase
2-Phosphoglycerate
H2O
Enolase
Phosphoenolpyruvate
ADP
Pyruvate Kinase
ATP
Pyruvate
Carboxylase
GDP+CO2
PEP Carboxykinase
GTP
Pyruvate
Oxalacetate
ATP+CO2
ADP
NADH
NAD+
Lactate Dehydrogenase
Lactate

FIG. 3. Glycolysis and gluconeogenesis. Glycolysis (the conversion of glucose to lactate) provides a net gain of two ATP per glucose. Note that gluconeogenesis (the reversal of glycolysis) requires the expenditure of six ATP per glucose.

ies suggest that some enzymes may associate with each other or with structural elements of the cell (e.g., membrane proteins in erythrocytes, contractile proteins in muscle).

2.1.1 Mechanism of ATP Generation in Glycolysis—Substrate-Level Phosphorylation (Chemical Coupling). The formation of ATP in glycolysis proceeds via *substrate-level phosphorylation,* which simply means that the overall process proceeds through discrete chemical intermediates. The reactions of the first ATP formed illustrate how cells couple exergonic and endergonic reactions through a common intermediate. In effect, the cell harnesses the energy made available by the oxidation of glyceraldehyde-3-phosphate (G-3-P) to drive the unfavorable formation of ATP:

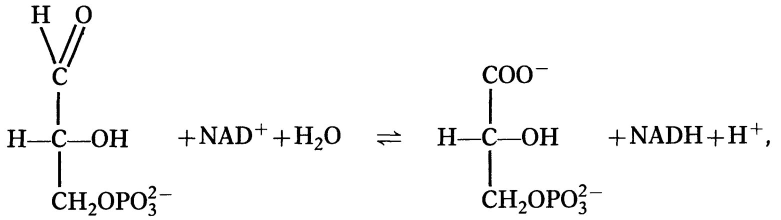

$$\mathrm{HC(=O)\text{—}CH(OH)\text{—}CH_2OPO_3^{2-}} + \mathrm{NAD^+} + \mathrm{H_2O} \rightleftharpoons \mathrm{COO^-\text{—}CH(OH)\text{—}CH_2OPO_3^{2-}} + \mathrm{NADH} + \mathrm{H^+}, \quad (12)$$

(G-3-P) [3-phosphoglyceric acid, (3-PGA)]

$$\Delta G^{0\prime}_{12} = -10.6 \text{ kcal mole}^{-1} = -44.4 \text{ kJ mole}^{-1},$$

$$\mathrm{ADP} + \mathrm{P}_i \rightleftharpoons \mathrm{ATP} + \mathrm{H_2O}, \quad (13)$$

$$\Delta G^{0\prime}_{13} = +7.6 \text{ kcal mole}^{-1} = +31.8 \text{ kJ mole}^{-1}.$$

While the sum of these two reactions is quite favorable,

$$\text{G-3-P} + \mathrm{NAD^+} + \mathrm{ADP} + \mathrm{P}_i \rightleftharpoons \text{3-PGA} + \mathrm{NADH} + \mathrm{H^+} + \mathrm{ATP}, \quad (14)$$

$$\Delta G^{0\prime}_{14} = -3 \text{ kcal mole}^{-1} = -12.5 \text{ kJ mole}^{-1},$$

cells have no mechanism for coupling two independent reactions occurring in the same cellular compartment if the reactions have no common component. Instead, cells catalyze two sequential reactions that accomplish the same result. First, G-3-P is oxidized by NAD$^+$ in a reaction in which inorganic phosphate replaces water. The product then is 1,3-bisphosphoglycerate (1,3-BPG) instead of 3-phosphoglycerate. The $\Delta G^{0\prime}$ of this reaction is not as negative as that of reaction (12) above because energy is conserved by forming the carboxyl-phosphoryl anhydride:

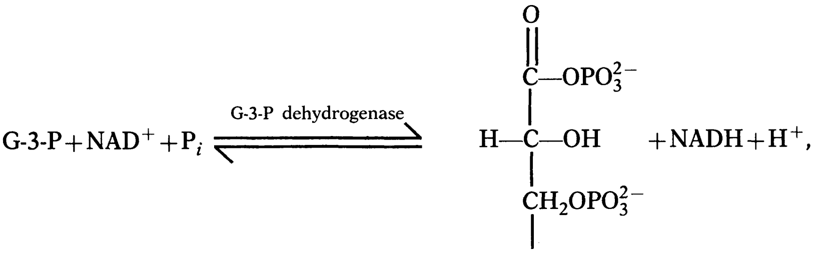

$$\text{G-3-P} + \mathrm{NAD^+} + \mathrm{P}_i \xrightleftharpoons{\text{G-3-P dehydrogenase}} \mathrm{O{=}C(OPO_3^{2-})\text{—}CH(OH)\text{—}CH_2OPO_3^{2-}} + \mathrm{NADH} + \mathrm{H^+}, \quad (15)$$

(1,3-BPG)

$$\Delta G^{0\prime}_{15} = +1.4 \text{ kcal mole}^{-1} = +5.9 \text{ kJ mole}^{-1}.$$

Then, the enzyme phosphoglyceryl kinase catalyzes the transfer of the C-1 phosphoryl group to ADP:

$$\text{1,3-BPG} + \mathrm{ADP} \overset{\text{PG kinase}}{\rightleftharpoons} \text{3-PGA} + \mathrm{ATP}, \quad (16)$$

$$\Delta G^{0\prime}_{16} = -4.4 \text{ kcal mole}^{-1} = -18.4 \text{ kJ mole}^{-1}.$$

The sum of reactions (15) and (16) equals reaction (14).

Although $\Delta G^{0\prime}$ for reaction (15) is positive, the actual *in vivo* ΔG is negative because 1,3-BPG is continuously removed by reaction (16). [Thus, Q for reaction (15) *in vivo* is always $< K_{eq}$ as long as glycolysis is proceeding.]

The formation of the second ATP per triose recovers the two phosphoryl groups used to prime glycolysis. Again, the overall sequence is driven by a favorable oxidation-reduction, although in this case, the electron movements are internal and occur both fore and aft of the ATP-yielding step. First, 2-phosphoglycerate (2-PGA) is dehydrated by the enzyme enolase yielding phosphoenolpyruvate (PEP). Then, the enzyme pyruvate kinase catalyzes a phosphoryl transfer from PEP to ADP yielding ATP and pyruvate. The latter reaction is extremely favorable because of the concurrent tautomerization of the putative product (enol-pyruvate) to the more stable keto form. (In effect, the phosphoryl group of PEP freezes pyruvate into the unstable enol form.) Overall, C-2 of 2-PGA loses two electrons which appear on C-3 of pyruvate:

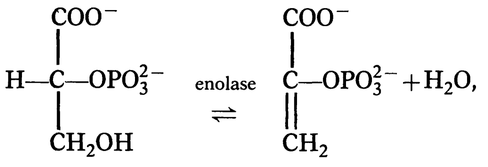

$$\begin{array}{c}\mathrm{COO^-}\\ |\\ \mathrm{H{-}C{-}OPO_3^{2-}}\\ |\\ \mathrm{CH_2OH}\end{array} \overset{\text{enolase}}{\rightleftharpoons} \begin{array}{c}\mathrm{COO^-}\\ |\\ \mathrm{C{-}OPO_3^{2-}}\\ \|\\ \mathrm{CH_2}\end{array} + \mathrm{H_2O}, \qquad (17)$$

(2-PGA) (PEP)

$$\Delta G^{0\prime}_{17} = +0.44 \text{ kcal mole}^{-1}$$

$$= +1.84 \text{ kJ mole}^{-1}$$

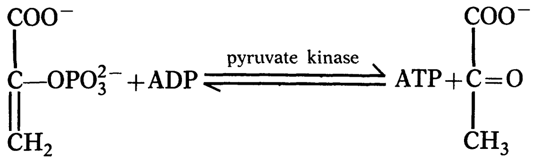

$$\begin{array}{c}\mathrm{COO^-}\\ |\\ \mathrm{C{-}OPO_3^{2-}}\\ \|\\ \mathrm{CH_2}\end{array} + \mathrm{ADP} \underset{}{\overset{\text{pyruvate kinase}}{\rightleftharpoons}} \mathrm{ATP} + \begin{array}{c}\mathrm{COO^-}\\ |\\ \mathrm{C{=}O}\\ |\\ \mathrm{CH_3}\end{array} \qquad (18)$$

(PEP) (pyruvate)

$$\Delta G^{0\prime}_{18} = -6.2 \text{ kcal mole}^{-1} = -25.9 \text{ kJ mole}^{-1}.$$

As in all reactions of glycolysis, the actual ΔG of reaction (17) will be at least slightly negative *in vivo* while glycolysis is proceeding.

2.1.2 Gluconeogenesis—The Reversal of Glycolysis. Many types of cells can synthesize glucose from noncarbohydrate precursors. The pathway, called gluconeogenesis, is for the most part, a reversal of glycolysis. This bidirectionality is possible because most of the glycolysis reactions are close to equilibrium *in vivo* and consequently, can be easily reversed by small adjustments in the concentrations of reaction components. However, three glycolysis reactions (those catalyzed by hexokinase, phosphofructokinase, and pyruvate kinase) are not reversible in the physiological sense. (Their ΔG values are highly positive in the gluconeogenesis direction—see Sec. 4.1.) As shown in Fig. 3, cells reverse the interconversions at these points by using different reactions with negative ΔG values. (It is noteworthy that the three metabolic traffic circles are the major regulatory points in glycolysis and gluconeogenesis.)

Glycolysis yields ATP. So we would expect glyconeogenesis to require ATP and this is indeed the case. But note that while glycolysis yields 2 ATP per glucose converted to 2 pyruvate, gluconeogenesis requires the equivalent of 6 ATP (4 ATP plus 2 GTP) to drive the synthesis of one glucose from 2 pyruvate. The biochemical economics of the situation is a consequence of the second law of thermodynamics. An examination of all metabolic pathways will reveal that it always costs more ATP (or its equivalent) to synthesize a more complex molecule from simpler precursors than a cell can obtain by breaking the complex molecule down to the simpler precursors.

2.2 Respiration—Aerobic Energy Metabolism

The biological oxidation of a reduced organic compound by oxygen forming CO_2 and H_2O is called *respiration.* Certain bacteria perform a variation on this theme in which a different inorganic substance (e.g., sulfate, nitrate) serves as the terminal electron acceptor. Respiration is a highly exergonic process. For example, the oxidation of glucose by O_2 ($C_6H_{12}O_6+6\ O_2 \rightleftharpoons 6\ CO_2+6\ H_2O$) has a $\Delta G^{0\prime}$ of -686 kcal mole^{-1} or -2870 kJ mole^{-1}. The overall pathway of respiration includes the participation of

1. *glycolysis* (glucose+2 ADP+2P_i+2NAD$^+$ $\rightarrow$2 pyruvate+2 ATP+2 NADH +2H$^+$),
2. an enzyme complex called *pyruvate dehydrogenase* (pyruvate+NAD$^+$+coenzyme *A* $\rightleftharpoons CO_2$+NADH+H$^+$+acetyl-coenzyme *A*), and
3. an eight-step pathway called the *tricarboxylic acid* (TCA) (or citric acid, or Krebs) *cycle.* This last pathway effects the complete oxidation of the remaining two carbon atoms of the acetyl group to CO_2 concomitant with the gathering of eight electrons on nicotinamide and flavin (FAD)-containing coenzymes:

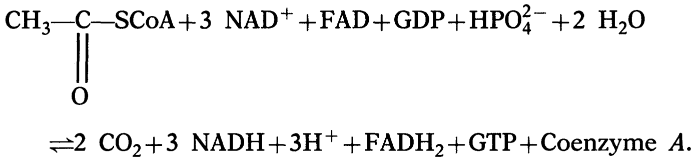

$$CH_3\text{—}\underset{\underset{O}{\|}}{C}\text{—}SCoA+3\ NAD^+ + FAD + GDP + HPO_4^{2-} + 2\ H_2O \rightleftharpoons 2\ CO_2 + 3\ NADH + 3H^+ + FADH_2 + GTP + \text{Coenzyme } A. \qquad (19)$$

Finally,

4. a membrane-embedded *electron transport system* (ETS) channels electrons from NADH and $FADH_2$ to O_2 yielding water. ATP is generated from ADP + P_i by
5. a membrane-associated ATPase (also called *ATP synthase*) (see below).

In eucaryotic cells, processes **2** through **5** all occur within the mitochondria. Acetyl-coenzyme *A* is also an intermediate in the metabolism of fatty acids. Consequently, processes **3** through **5** function in the oxidation of fats as well as carbohydrates.

2.2.1 The Electron Transport System and the Chemiosmotic Mechanism of Oxidative Phosphorylation. Although one mole of GTP (an ATP equivalent) is produced per mole of pyruvate in the TCA cycle, the majority of the ATP is formed by a process called *oxidative phosphorylation* which taps the energy produced when NADH and $FADH_2$ are oxidized by O_2. The yield is 3 ATP for each NADH plus 2 ATP for each $FADH_2$ produced intramitochondrially. NADH produced in the cytosol by glycolysis is used to reduce dihydroxyacetone phosphate to glycerol phosphate. The latter compound enters the mitochondria where it is oxidized by FAD. Thus, the total yield is 36 ATP for each glucose oxidized to 6 CO_2 and 6 H_2O.

The components of the electron transport system (which are associated with the inner mitochondrial membrane) include two types of substances which carry only electrons and two which carry one or two complete hydrogen atoms (H=electron plus proton). Most cytochromes (heme-containing proteins) and the non-heme, iron-sulfur-containing proteins belong to the first type. The second type include flavoproteins (proteins containing the coenzyme flavin mononucleotide or flavin-adenine dinucleotide), and a quinone (ubiquinone, also called coenzyme Q). Electrons are transferred from NADH or $FADH_2$ to O_2 by a series of favorable sequential oxidation-reduction reactions (Fig. 5).

Unlike the formation of ATP in glycolysis, or GTP in the TCA cycle, the formation of ATP in oxidative phosphorylation does not proceed through discrete chemical intermediates. Instead, the oxidation of NADH or $FADH_2$ generates a charge and a *p*H gradient across the proton-impermeable inner mitochondrial membrane, i.e., the "energy-rich" intermediate is physical in nature rather than chemical.

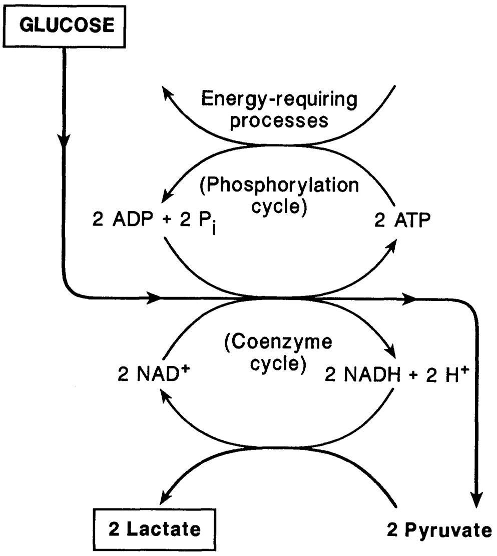
GLUCOSE
Energy-requiring processes
(Phosphorylation cycle)
2 ADP + 2 P_i
2 ATP
(Coenzyme cycle)
2 NAD+
2 NADH + 2 H+
2 Lactate
2 Pyruvate

FIG. 4. ATP and coenzyme cycling. Because of the limited supply of ADP and NAD^+ in a cell, they must be continually regenerated in order to permit a prolonged period of glycolysis.

The features of this "chemiosmotic" model (in its simplest form) are as follows.

The electron carriers of the electron-transport system alternate in order between those which carry complete hydrogen atoms and those which carry only electrons. Furthermore, the carriers are arranged asymmetrically within the membrane such that when electrons are transferred from a complete hydrogen carrier to an electron-only carrier, the protons are released on the exterior side of the membrane. When electrons are transferred from an electron-only carrier to one that carries a complete hydrogen, the needed protons are obtained from the interior side of the membrane (i.e., from the mitochondrial matrix). The final carrier, cytochrome a_3 (also called cytochrome oxidase), promotes both the transfer of electrons to O_2 and the release of protons on the exterior side of the membrane. As a result of these sequential reactions, protons are pumped out of the mitochondrial matrix into the cytosol. (The outer mitochondrial membrane is "leaky" and thus, the intermembrane space is at equilibrium with the cytosol.) Essentially, the favorable free-energy change of NADH or $FADH_2$ oxidation is used to generate an *ion electrochemical potential* ($\Delta\Psi$) and an H^+ *concentration gradient* (ΔpH) across the inner membrane. Thus, compared to the cytosol, the mitochondrial matrix becomes negatively charged and has a higher *p*H. The return of protons to the matrix (i.e., the reaction $H^+_{out} \rightleftharpoons H^+_{in}$) is an exergonic process. In general, ΔG for the movement of a charged substance from "side 1 " of a membrane to "side 2" is given by

$$\Delta G = Z\mathcal{F}\Delta\Psi_{2/1} + RT\ln([C]_2/[C]_1), \qquad (20)$$

where Z is the net charge on the substance, $\mathcal{F}$ is Faraday's constant, $\Delta\Psi_{2/1}$ is the electrochemical (or membrane) potential expressed as

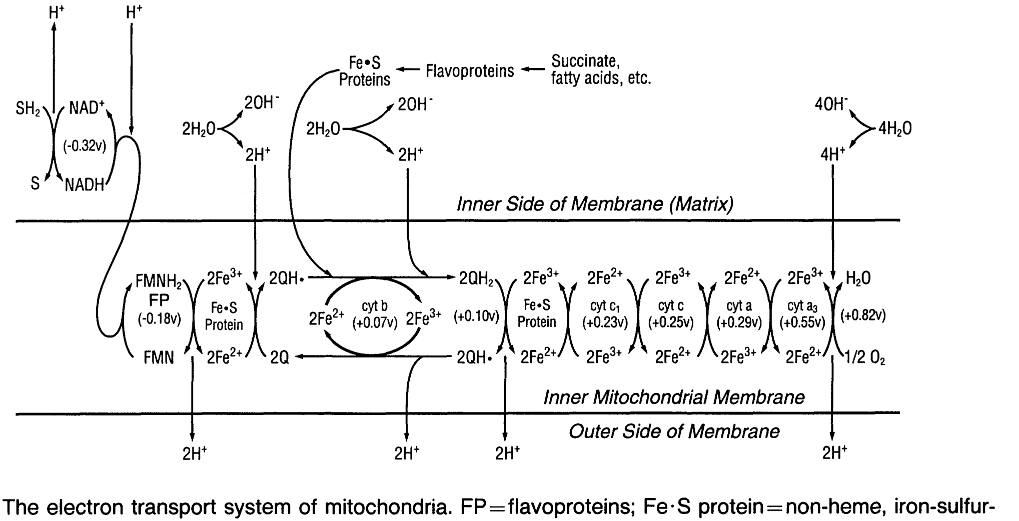
H+
H+
SH2
NAD+
(-0.32v)
S
NADH
2H2O
2OH-
2H+
Fe•S Proteins
Flavoproteins
Succinate, fatty acids, etc.
2H2O
2OH-
2H+
4OH-
4H2O
4H+
Inner Side of Membrane (Matrix)
FMNH2
FP (-0.18v)
FMN
2Fe3+
Fe•S Protein
2Fe2+
2QH•
2Q
cyt b (+0.07v)
2Fe2+
2Fe3+
2QH2
(+0.10v)
2QH•
2Fe3+
Fe•S Protein
2Fe2+
2Fe2+
cyt c1 (+0.23v)
2Fe3+
2Fe3+
cyt c (+0.25v)
2Fe2+
2Fe2+
cyt a (+0.29v)
2Fe3+
2Fe3+
cyt a3 (+0.55v)
2Fe2+
H2O
(+0.82v)
1/2 O2
Inner Mitochondrial Membrane
Outer Side of Membrane
2H+
2H+
2H+
2H+

FIG. 5. The electron transport system of mitochondria. FP = flavoproteins; Fe·S protein = non-heme, iron-sulfur-containing proteins; cyt = cytochromes (heme-containing proteins); Q, QH·, QH_2 = quinone, semiquinone, and hydroquinone forms of coenzyme Q.

side 2 compared to side 1, $[C]_2/[C]_1$ is the actual concentration ratio (side 2/side 1), i.e., Q for the reaction as written. For the movement of protons ($Z=+1$) from side 1=cytosol =outside to side 2=matrix=inside, Eq. (20) can be written as

$$\Delta G_{\text{out}\to\text{in}}=\mathscr{F}\Delta\Psi_{\text{in/out}} +2.3RT\log([\text{H}^+]_{\text{in}}/[\text{H}^+]_{\text{out}}) \tag{21}$$

or, noting that $p\text{H}=-\log[\text{H}^+]=\log(1/[\text{H}^+])$,

$$\Delta G_{\text{out}\to\text{in}}=\mathscr{F}\Delta\Psi_{\text{in/out}}-2.3RT\ \Delta p\text{H}, \tag{22}$$

where $\Delta p\text{H}=p\text{H}_{\text{in}}-p\text{H}_{\text{out}}$.

The voltage equivalent of the ΔG for proton influx, called the *protonmotive force*, $\Delta\tilde{\mu}_{\text{H}^+}$, is obtained by dividing Eq. (22) by $\mathscr{F}$:

$$\begin{aligned}\Delta\tilde{\mu}_{\text{H}^+}&=\Delta\Psi_{\text{in/out}}-(2.3RT/\mathscr{F})\Delta p\text{H}_{(\text{in-out})}\\ &=\Delta\Psi_{\text{in/out}}-0.059\ \Delta p\text{H}_{(\text{in-out})}\\ &\qquad\text{at } 25\ ^\circ\text{C}.\end{aligned} \tag{23}$$

As defined above, $\Delta\tilde{\mu}_{\text{H}^+}$ will have a negative sign under the conditions described. Readers should be aware that some sources express the $\Delta\tilde{\mu}_{\text{H}^+}$ driving force in positive terms. This is analogous to speaking of ATP ($\Delta G^{0\prime}$ of hydrolysis$=-7.6$ kcal mole^{-1} or -31.8 kJ mole^{-1}) as having a standard *phosphoryl group transfer potential* of $+7.6$ kcal mole^{-1} or $+31.8$ kJ mole^{-1}.

The formation of ATP is accomplished by a multisubunit, membrane-embedded proton-translocating enzyme called ATPase or ATP synthase. This enzyme couples the endergonic ATP synthesis half-reaction ($\Delta G^{0\prime}=+7600$ cal mole^{-1}) to the simultaneous exergonic proton influx half-reaction. For example, if the ΔpH across the inner membrane is 1.2 units (interior more alkaline) and the $\Delta\Psi$ is -0.13 V (interior negative), the ΔG for proton influx will be -4635 cal mole^{-1} or $-19\,393$ J mole^{-1} [Eq. (22)]. If at the same time, the intramitochondrial [ADP]/[ATP] ratio is maintained at 6 while $[\text{P}_i]=0.01M$, the ΔG for ATP synthesis will be $+9267$ cal mole^{-1} or $+38\,773$ J mole^{-1} [Eq. (3)]. Clearly, the influx of one H$^+$ cannot provide sufficient driving force to produce ATP. However, the coupled influx of 2H$^+$ will just be sufficient to tilt the overall reaction in the ATP formation direction:

$$\text{ADP}+\text{P}_i\rightleftharpoons\text{ATP}+\text{H}_2\text{O}, \tag{24}$$

$$\begin{aligned}\Delta G_{24}&=+9267\text{ cal per mole,}\\ &=+38\,773\text{ J per mole,}\end{aligned}$$

$$2\text{H}^+_{\text{out}}\rightleftharpoons 2H^+_{\text{in}}, \tag{25}$$

$$\begin{aligned}\Delta G_{25}&=-9270\text{ cal per 2 moles,}\\ &=-38\,786\text{ J per 2 moles,}\end{aligned}$$

Sum:

$$\text{ADP}+\text{P}_i+2\text{H}^+_{\text{out}}\rightleftharpoons\text{ATP}+\text{H}_2\text{O}+2\text{H}^+_{\text{in}}, \tag{26}$$

$$\Delta G_{26}=\text{slightly negative.}$$

In healthy cells, reaction (26) will continue indefinitely as the H$^+$ is pumped back out of the mitochondria by the continued operation of the ETS, and ATP leaves the mitochondria while ADP and P$_i$ enter.

3. PHOTOSYNTHESIS

Photosynthesis is the process whereby higher plants, algae, and certain bacteria use light energy to drive the synthesis of reduced organic compounds (e.g., glucose) from CO_2 and an electron source (H_2O in plants; other inorganic substances such as H_2S or H_2 in bacteria). Photosynthesis and respiration are the two major limbs of the solar-powered carbon cycle of nature (Fig. 6). Indeed, heterotrophic life on earth is completely dependent on photosynthesis because chemoorganotrophic food chains can not be forever circular. (The chemical elements of biomolecules—carbon, hydrogen, oxygen, nitrogen, phosphorous, and sulfur—are completely recyclable, but the second law of thermodynamics dictates that energy is not.) A complete description of the photosynthetic process is beyond the scope of this article, but can be summarized as follows:

1. The "light" reaction (generation of ATP and NADPH):

$$\begin{aligned}&12\ \text{H}_2\text{O}+18\ \text{ADP}+18\text{P}_i+12\ \text{NADP}^+\\ &\xrightarrow{60h\nu} 18\ \text{ATP}+12\ \text{NADPH}+12\ \text{H}^+\\ &\qquad+6\ \text{O}_2+18\ \text{H}_2\text{O}.\end{aligned}$$

(Water appears on both sides of the reaction because labeling experiments have estab-

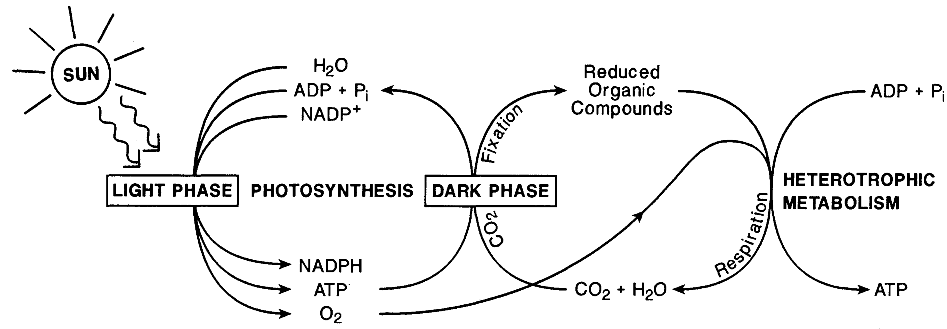

FIG. 6. The major limbs of the carbon cycle. At any time, living organisms contain about 830×10^{15} g of carbon (90% of which is present in plants), while the atmosphere contains about 670×10^{15} g of carbon as CO_2. About 10% of the total carbon is recycled per year.

lished that all 12 atoms of the oxygen produced are derived from water.)

2. The "dark" reaction (CO_2 reduction to glucose):

$$\begin{aligned}&6\ CO_2+18\ ATP+12\ NADPH+12\ H^+\\&\quad+12\ H_2O\\&\rightarrow C_6H_{12}O_6+18\ ADP+18\ P_i\\&\quad+12\ NADP^+.\end{aligned}$$

If one NADPH is considered to be energetically equivalent to 3 ATP, the overall synthesis of glucose from 6 CO_2 plus 6 H_2O is seen to cost 54 ATP. (Recall that aerobic cells obtain only 36 ATP for each glucose completely oxidized to 6 CO_2 plus 6 H_2O).

Figure 7 shows (in abbreviated form) how the capture of light energy by plants leads to the generation of ATP and NADPH. Two photons are captured by photosystem I (a chlorophyll and electron-carrier–containing center within the thylakoid membrane structures of the chloroplast). The absorption of light energy elevates two electrons to a highly negative reduction potential. (One can also envision that light excites a pigment component of photosystem I so that it becomes a much stronger reducing agent.) The two electrons are transferred to $NADP^+$ through a series of electron carriers (phylloquinone and iron-sulfur proteins, including ferredoxin) producing NADPH. A similar excitation occurs when photosystem II absorbs two photons. This time, the two electrons are transferred (via pheophytin, quinones, cytochromes, and plastocyanin) to the oxidized photosystem I, restoring that system to its original state. During the electron transport process, protons are pumped into the thylakoid lumen. (The direction of H^+ movement is the opposite to that which occurs in mitochondria.) The resulting $\Delta\tilde{\mu}_{H^+}$ is used to generate one ATP, as described earlier. Finally, two electrons from H_2O make up the electron deficiency in photosystem II yielding oxygen as a product. The overall noncyclic process (so called because the electrons that leave photosystem I do not return to that system) generates 1 NADPH + 1 ATP (and liberates $\frac{1}{2}$ O_2) per 2 photons. Twelve repeats yield sufficient NADPH for the "dark" reaction, but only 12 of the 18 ATP that are required. The remaining 6 ATP are produced by six traversals of a cyclic process in which two electrons excited from photosystem I are transferred back to the same photosystem, generating a $\Delta\tilde{\mu}_{H^+}$ which yields 1 ATP per 2 photons. The overall process requires the input of 60 einsteins (moles) of photons (36 into photosystem I plus 24 into photosystem II) in order to produce the 18 moles of ATP and 12 moles of NADPH required for the reduction of 6 moles of CO_2 to one mole of glucose. The energy of an einstein of photons is given by $E=Nh\nu=Nhc/\lambda$, where N is Avogadro's number, h is Planck's constant, ν is the light frequency, c is the speed of light, and λ is the wavelength of the light. If the wavelength of the captured light is given in cm, the energy in an einstein of photons can be calculated as

$$E=2.86/\lambda_{\text{cm}}\ \text{cal einstein}^{-1}$$

or

$$E=11.95/\lambda_{\text{cm}}\ \text{J einstein}^{-1}. \qquad (27)$$

Thus, about 2900 kcal of visible light energy (average $\lambda=600$ nm) are required to form one mole of glucose. Recall that the oxidation

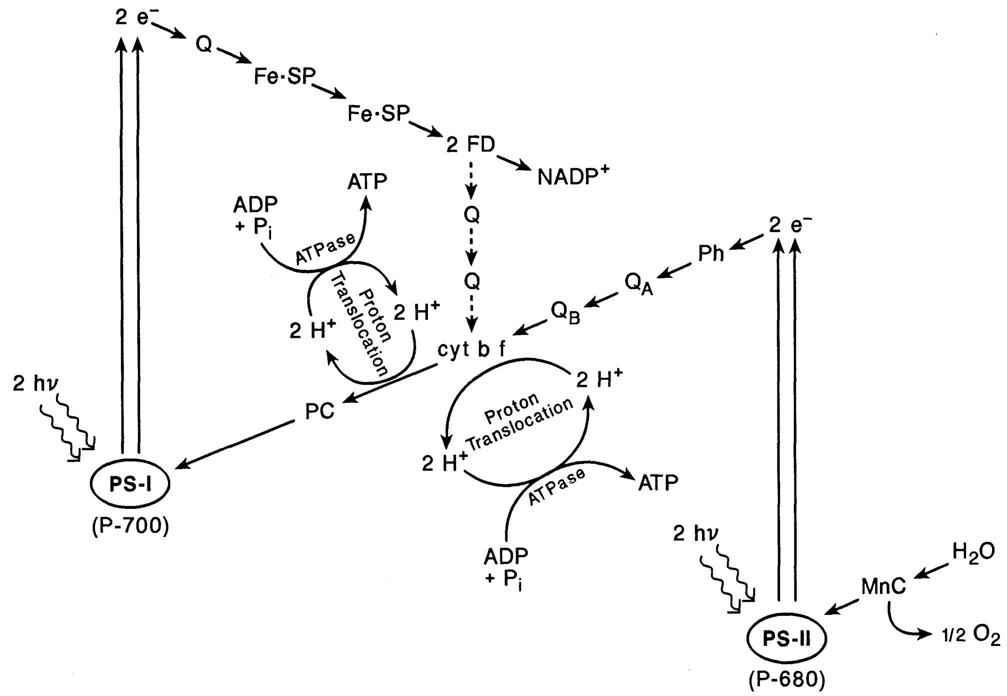

FIG. 7. The path of electrons during noncyclic (solid line) and cyclic (broken line) photophosphorylation in plants. Photosystem I (PS-I) has an absorption maximum (λ_{max}) of 700 nm (7×10^{-5} cm) and thus is sometimes indicated as P700. Photosystem II (PS-II, P680) has a λ_{max} of 680 nm. Q, Q_A, Q_B are quinones; FeSP represents non-heme, iron-sulfur proteins; Ph is pheophytin, a chlorophyll molecule in which the central Mg^{2+} is replaced by $2H^+$; PC is plastocyanin, a copper-containing protein; cytbf is a complex of two cytochromes; MnC is a manganese-containing reaction center involved in transferring electrons from H_2O to the "positive hole" of photosystem II. The photosynthetic apparatus harnesses the (literally) free energy of sunlight ultimately making all manner of energy-requiring biological processes possible, including the formation of complex, highly organized (low entropy) macromolecular and cellular structures from simple, low molecular weight (high entropy) inorganic and organic precursors.

of glucose via aerobic respiration yields only 686 kcal $mole^{-1}$. But this "inefficiency" of the carbon cycle is of no consequence because the primary biological energy source, sunlight, is (for all practical considerations) unlimited and inexhaustible.

4. HOW DO CELLS DRIVE ENERGY-REQUIRING REACTIONS?

4.1 Limitations of Coupled Reactions

In order to appreciate the role of ATP, let us examine a simple cellular task—the formation of glucose-6-phosphate (G-6-P) which is necessary to initiate glycolysis. Theoretically, G-6-P can be formed from glucose and P_i by the G-6-P phosphatase-catalyzed reaction

$$\text{glucose}+P_i \overset{\text{G-6-Ptase}}{\rightleftharpoons} \text{G-6-P}+H_2O, \qquad (28)$$

$$\Delta G^{0\prime}_{28}=+3.3\ \text{kcal}\ (+13.8\ \text{kJ})\ \text{mole}^{-1},$$

$$K_{eq}=3.81\times10^{-3}M^{-1}.$$

At the usual intracellular levels of glucose (5 mM) and P_i (10 mM), the maximum (equilibrium) level of G-6-P would be 0.19 μM. It would be necessary to maintain the [G-6-P] somewhat below 0.19 μM in order to keep the reaction proceeding to the right. But there is a practical limitation to this approach: The enzymes that use G-6-P require about 30-μM G-6-P in order to work at their half-maximum rates. At the normal cellular levels of these enzymes, G-6-P could not be metabolized at a biologically useful rate. So, while the continual formation of G-6-P via the above reaction can (in theory) be made thermodynamically feasible, it is not kinetically feasible.

Another way to drive a reaction is to increase the concentrations of reactants. In theory, an equilibrium G-6-P concentration of 30 μM could be attained if the concentrations of glucose and P_i were maintained at about 0.09 *M* each. But once again, the limitations of living cells make this an impractical option. Most cells will not accumulate solutes to internal levels much greater than $10^{-2}M$. There are many reasons for this upper limit. One is that cells must expend energy to concentrate solutes to levels higher than that in the external environment. The energy cost increases with increasing internal concentration. Another reason is that high concentrations of charged compounds, such as inorganic phosphate, inhibit the activities of many enzymes. Perhaps the simplest explanation is that cells must maintain thousands of different organic compounds (including proteins) in solution. The solvent capacity of a cell is not sufficient to maintain a large number of different compounds at high concentrations. Thus, some highly endergonic reactions can be neither pulled nor pushed by simply manipulating internal concentrations. How then do cells drive these reactions? The answer lies in the special properties of a group of compounds biochemists call *energy-rich compounds.*

4.2 "Energy-Rich" Compounds

To the biochemist, the term *energy-rich compound* generally refers to a compound that has a rather highly negative $\Delta G^{0\prime}$ of hydrolysis. One class of such compounds are the *nucleoside triphosphates,* of which ATP is the most abundant in cells. ATP plays a central role in cellular energy metabolism, linking the oxidative (or photosynthetic) pathways that supply chemical (or light) energy to the biosynthetic pathways that require energy to proceed. Consider the problem of forming G-6-P. On paper, the hydrolysis of ATP yields sufficient "energy" to drive the synthesis of G-6-P from glucose + P_i:

$$\text{glucose} + P_i \xrightleftharpoons{\text{G-6-Ptase}} \text{G-6-P} + H_2O \qquad (29)$$

$$\Delta G^{0\prime}_{29} = 3.3 \text{ kcal } (+13.8 \text{ kJ}) \text{ mole}^{-1},$$

$$\text{ATP} + H_2O \xrightleftharpoons{\text{ATPase}} \text{ADP} + P_i, \qquad (30)$$

$$\Delta G^{0\prime}_{30} = -7.6 \text{ kcal } (-31.8 \text{ kJ}) \text{ mole}^{-1}.$$

However, reactions (29) and (30) cannot be coupled in the usual sense because there is no common intermediate. So, if glucose, P_i, and ATP are mixed together in the presence of ATPase and glucose-6-phosphatase, both reactions would proceed independently toward their respective equilibrium positions. The concentration of G-6-P at equilibrium would be almost the same as in the absence of ATP and ATPase. (The uncoupled situation is analogous to burning a gallon of gasoline on the driveway. A lot of energy will be released, but your car will not move.)

If we add reactions (29) and (30) we obtain

$$\text{ATP} + \text{glucose} \xrightleftharpoons{\text{hexokinase}} \text{G-6-P} + \text{ADP}, \qquad (31)$$

$$\Delta G^{0\prime}_{31} = -4.3 \text{ kcal } (-18 \text{ kJ}) \text{ mole}^{-1},$$

$$K_{eq} = 1421.$$

Reaction (31) is more than just the sum of two uncoupled reactions. It is, in fact, the actual reaction used by cells to produce G-6-P from glucose. However, reaction (31) is not catalyzed by glucose-6-phosphatase plus ATPase, but rather by a completely different enzyme called hexokinase. Hexokinase accomplishes the "coupling" of reactions (29) and (30) by combining them into a single reaction.

The hexokinase-catalyzed reaction exemplifies a key biological role of energy-rich compounds. Although energy-rich compounds have highly negative free energies of hydrolysis, their use seldom involves simple hydrolysis. Instead, they are used as *group transfer agents* that donate parts of their structure to acceptors other than water (when the appropriate enzyme is present). In the hexokinase reaction, ATP serves as a phosphoryl-group ($-PO_3^{2-}$) donor. The advantage of ATP over inorganic phosphate as the precursor of the phosphate ester group is obvious: The hexokinase-catalyzed phosphate transfer reaction proceeds far to the right. At the steady-state levels of ATP and glucose in a cell, G-6-P can accumulate to a level that permits the subsequent G-6-P–utilizing enzymes to operate at biologically useful rates. Clearly, a very low steady-state concentration of ATP can accomplish the same result as a much higher concentration of inorganic phosphate. We might say, then, that the use of ATP as a phosphorylating agent is nature's way of conserving the limited solvent capacity of a cell.

ATP serves not only as a phosphoryl donor, but also as a pyrophosphoryl donor, as an adenylyl (AMP) donor, and as an adenosine donor. Other nucleoside triphosphates (NTP) play similar roles. The group transfer may be an end in itself (e.g., G-6-P, nucleic acid synthesis), or may be a prelude to further reactions. Indeed, most biosynthetic reactions are driven by group transfer from an NTP producing an "activated" intermediate containing an easily displaced leaving group. For example, consider the synthesis of sucrose from glucose. The $\Delta G^{0\prime}$ for the reaction 2 glucose $\rightleftharpoons$ sucrose+H_2O is about +7 kcal mole^{-1}. However, *in vivo*, a plant cell might proceed by the sequence of coupled reactions shown below (UTP stands for uridine triphosphate):

	$\Delta G^{0\prime}$ (kcal×mole^{-1})	
$2\ \text{glucose} + 2\ \text{ATP} \overset{\text{hexokinase}}{\rightleftharpoons} 2\ \text{G-6-P} + 2\ \text{ADP},$	-4.3×2	(32)
$\text{G-6-P} \overset{\text{phosphoglucose isomerase}}{\rightleftharpoons} \text{fructose-6-P (F-6-P)},$	$+0.5$	(33)
$\text{G-6-P} \overset{\text{phosphoglucomutase}}{\rightleftharpoons} \text{G-1-P},$	$+1.7$	(34)
$\text{G-1-P} + \text{UTP} \overset{\text{UDPG synthetase}}{\rightleftharpoons} \text{uridine diphosphate glucose (UDPG)} + PP_i,$	$+0.7$	(35)
$\text{UDPG} + \text{F-6-P} \overset{\text{sucrose-P synthase}}{\rightleftharpoons} \text{sucrose-P} + \text{UDP},$	-0.7	(36)
$PP_i + H_2O \overset{\text{inorganic pyrophosphatase}}{\rightleftharpoons} 2\ P_i,$	-5.0	(37)
$\text{sucrose-P} + H_2O \overset{\text{sucrose phosphatase}}{\rightleftharpoons} \text{sucrose} + P_i,$	-3.8	(38)
Sum: $2\ \text{glucose} + 2\ \text{ATP} + \text{UTP} + 2\ H_2O \rightleftharpoons \text{sucrose} + 2\ \text{ADP} + \text{UDP} + 3\ P_i.$	-15.2	(39)

Essentially, the formation of one mole of sucrose from 2 moles of glucose is driven by the free energy of hydrolysis of 3 moles of nucleoside triphosphate (NTP). The expenditure of 3 NTP converts what would be an unfavorable (endergonic) reaction into a highly favorable (exergonic) reaction sequence that proceeds far to the right. (But notice that an actual NTP hydrolysis reaction does not occur anywhere in the sequence.)

Although the above reactions were characterized by their $\Delta G^{0\prime}$ values (some of which were slightly positive), it is crucial to recognize that if sucrose is continuously formed from 2 glucose, then the actual ΔG of every reaction in the sequence must be negative. (If the overall flow of a stream is from high to low ground, there can be no regions along the course with a net uphill flow.)

Sucrose has the same overall oxidation state as glucose. Consequently, only ATP and UTP are required for the biosynthetic process. If we examine the formation of a more reduced product from glucose (e.g., fatty acids), we would find that reducing power in the form of NADPH is required along with ATP.

4.3 Active Transport

The accumulation of solutes across a membrane and against a concentration gradient is a major energy-requiring activity of cells. Many such *active transport* processes are driven by a transmembrane electrochemical potential, $\Delta\tilde{\mu}$, usually of H^+ or Na^+. This $\Delta\tilde{\mu}$ is generated either by an electron transport system within that membrane, or by a membrane-embedded ATPase that couples the "outward" translocation of H^+ (or Na^+) to ATP hydrolysis. Uptake of the solute (e.g., a neutral amino acid) is accomplished by a specific membrane transport system (sometimes called a "permease") which cataylzes the *cotransport* (*"symport"*) of the amino acid and H^+ or Na^+. In effect, the downhill flow of H^+ or Na^+ drives the simultaneous uphill flow of amino acid. The amino acid can be accumulated to an internal/external concentration ratio such that the positive ΔG for further uptake,

$$\Delta G = RT \ln \frac{[\text{amino acid}]_{\text{in}}}{[\text{amino acid}]_{\text{out}}}, \tag{40}$$

numerically equals the negative ΔG provided by the H^+ or Na^+ electrochemical potential [Eq. (20)], i.e., the sum of the two half-reactions has a zero ΔG. The protons taken up by symport are reextruded, thereby maintaining the $\Delta\tilde{\mu}$ across the plasma membrane.

4.4 Other Types of Energy-Rich Compounds

In addition to nucleoside triphosphates, cells contain several other types of compounds that have highly negative ΔG's of hydrolysis—or, in a functional sense, have high group-transfer potentials. These include several classes of anhydrides such as inorganic pyrophosphate, acyl phosphates (e.g., acetyl phosphate, 1,3-bisphosphoglycerate), acyl adenylates (particularly amino acyl adenylates), and phosphosulfates (e.g., adenosine-5′-phosphosulfate). Other types include thioesters (e.g., acetyl-coenzyme *A*), enol phosphates (PEP may be the sole example), guanidinium phosphates (e.g., creatine phosphate), and sugar nucleotides (e.g., UDPG). A complete description of each type, their metabolic roles, and the chemical basis of their high group-transfer potential is beyond the scope of this article. In general, one or more of the following factors contribute to the instability of all energy-rich compounds: electrostatic repulsion between like-charged atoms, spontaneous ionization or tautomerization of a primary reaction product, and resonance stabilization of the products of the reaction. All energy-rich compounds share an essential feature: They are *thermodynamically unstable*, but *kinetically stable*. That is, they can persist in solution for prolonged periods without decomposing but react readily in the presence of a specific group-transfer enzyme.

GLOSSARY

Active Transport: "Uphill" transport and accumulation of a solute within a cell (or cellular compartment) against a concentration gradient. Active transport is an energy-requiring process.

Alcohol Fermentation: Yeast glycolysis; the anaerobic conversion of one mole of glucose to two moles of ethanol plus two moles of CO_2.

Aldolase (FBP aldolase; aldolase A): Enzyme that cleaves the 6-carbon fructose-1,6-bisphosphate into two triose phosphates (dihydroxyacetone phosphate and glyceraldehyde-3-phosphate).

ATP: Adenosine-5′-triphosphate (adenine-ribose-P-P-P), the most common "energy-rich" compound in cells. ATP is generally used as a source of transferable phosphoryl, pyrophosphoryl, adenylyl, and adenosyl groups which "activate" the acceptor.

ATPase (ATP synthase): An enzyme that catalyzes the reversible reaction $ATP + H_2O \rightleftharpoons ADP + P_i$.

Autotroph: Organism that uses CO_2 as its sole source of carbon.

Calorie: Unit of energy, work, and heat. One calorie (with a small c) is equivalent to the heat needed to raise the temperature of 1 g of water by 1 °C. One Calorie ("large calorie," with a capital C) is 1000 calories (1 kcal). The caloric values of foods are given in terms of Calories. 1 cal = 4.184 J.

Chemical Coupling: See substrate-level phosphorylation.

Chemiosmotic Mechanism: Process whereby the endergonic synthesis of ATP from ADP and P_i by a membrane-spanning enzyme (ATPase or ATPsynthase) is driven by the coupled (simultaneous) exergonic influx of protons catalyzed by the same enzyme.

Chlorophyll: Green, magnesium-containing protein-bound porphyrin that serves as the primary light-absorbing pigment of higher plants, algae, and some bacteria.

Chloroplast: Organelle within photosynthetic organisms that contains the light-capturing pigments and the enzymes of photosynthesis as well as many other enzymes.

Coenzyme: Low-molecular-weight (compared to proteins) organic compound that serves as a group-specific carrier. Most coenzymes are derived from water-soluble vitamins.

Coenzyme A: Sulfur-containing coenzyme, derived from the vitamin pantothenic acid, that carries acyl groups in thioester linkage.

Coenzyme Q: Membrane-bound quinone that serves as a $1e^-$ or $2e^-$ carrier in the mammalian electron transport system.

Coupled Reactions (Apparent): Reactions (true or hypothetical) whose sum equals an actual reaction.

Coupled Reactions (True): Sequential reactions with a common component, e.g., $A \rightleftharpoons B$, $B \rightleftharpoons C$.

Cyclic Photophosphorylation: ATP synthesis by photosynthetic organisms driven by electrons returning via associated carriers to the same photosystem from which they were excited by light.

Cytochrome: Protein containing a heme (iron porphyrin) center that acts as an electron transfer agent.

Dehydrogenase: Common name for the class of enzymes that catalyze coenzyme-dependent oxidation-reduction reactions.

Einstein: One mole (Avogadro's number) of photons.

Electron Transport System: Electron carriers, generally membrane-bound, that transfer electrons sequentially from an initial donor (e.g., NADH, $FADH_2$) to the terminal acceptor.

Energy-Rich Compound: Biochemical jargon for a compound having a highly negative ΔG of hydrolysis.

Enolase: Common name of the enzyme that catalyzes the interconversion of 2-phosphoglycerate and phosphoenolpyruvate.

Enthalpy Change (ΔH): Heat change accompanying a reaction. ΔH is negative if heat is evolved; positive if heat is absorbed.

Entropy Change (ΔS): Change in the degree of randomness (disorder) as a result of a reaction. The ΔS of the reaction components (the "system") may increase, decrease, or remain constant, but the total ΔS of the system plus its surroundings always increases as a result of any real reaction.

Enzyme: Protein catalyst that accelerates the rate at which a reaction approaches equilibrium. Living cells contain more than 1000 different enzymes, each highly specific for a particular reaction.

E_0': Symbol for the standard reduction potential of a half-reaction at $p\mathrm{H} = 7$.

Equilibrium: Lowest energy level of a system; energy level at which the process or reaction does not proceed in a net forward or reverse direction. If the equilibrium is a dynamic one (as all chemical equilibria are), the forward and reverse reaction rates are equal.

Equilibrium Constant (K_{eq}): Mass action quotient of reaction components at equilibrium. For the common reaction type $A + B \rightleftharpoons C + D$, $K_{eq} = [C][D]/[A][B]$.

Flavoprotein: Protein containing the tightly bound, 2H-carrying coenzyme flavin mononucleotide (FMN) or flavin adenine dinucleotide (FAD).

Free-Energy Change: See Gibbs free-energy change.

Gibbs Free-Energy Change (ΔG): Maximum work that can be performed by a process; a measure of the displacement of a system from equilibrium.

Gluconeogenesis: Synthesis of glucose from noncarbohydrate organic precursors such as lactate, TCA cycle intermediates, and amino acids.

Glycolysis: Anaerobic conversion of glucose to lactate or ethanol plus CO_2.

Group-Transfer Potential: Free energy of hydrolysis of an energy-rich compound expressed as a positive value.

Heterotroph: Organism that uses preformed organic compounds as its source of carbon.

Iron-Sulfur Protein: Electron-transfer protein containing one or more iron atoms bound to acid-labile sulfur atoms or cysteine residues of the protein polypeptide (or both).

Isomerase: Common name (in carbohydrate metabolism) for an enzyme that catalyzes the interconversion of keto and aldo structural isomers.

Kinase: Common name for an enzyme that catalyzes a phosphoryl transfer reaction between a nucleosidetriphosphate and an acceptor. Specificity is given by a prefix (e.g., hexokinase, ribokinase).

Lithotroph: Organism that uses inorganic compounds as its source of energy.

Mass-Action Quotient: Ratio of concentration driving forces of opposing reactions (i.e., for the reaction $nA+mB \rightleftharpoons rC+sD$, the mass action quotient Q is $[C]^r[D]^s/[A]^n[B]^m$).

Metabolism: Sum total of all the enzyme-catalyzed chemical reactions that occur in cells.

Mitochondrion: Double-membraned organelle within eucaryotic cells that contains pyruvate dehydrogenase, the enzymes of the TCA cycle, the electron transport system, enzymes of fatty acid oxidation, and many other enzymes.

Mutase: Common name (in carbohydrate metabolism) for an enzyme that catalyzes the intramolecular transfer of a phosphoryl group from one carbon atom to another.

NAD^+: Nicotinamide adenine dinucleotide (nicotinamide-ribose-P-P-ribose-adenine), a coenzyme that transiently carries a pair of electrons. [See Fig. 2(b)].

$NADP^+$: Electron-carrying coenzyme similar in structure to NAD^+. [See Fig. 2(b)].

Noncyclic Photophosphorylation: ATP synthesis by photosynthetic organisms driven by electrons from photosystem II as they flow to photosystem I via associated carriers.

Oxidation: The loss of electrons.

Oxidative Phosphorylation: ATP formation from ADP + P_i concomitant with electron transfer from a reduced coenzyme (NADH, $FADH_2$) to O_2 via a membrane bound electron transport system.

Oxidizing Agent: Compound that accepts electrons donated by another substance (e.g., NAD^+ is an oxidizing agent in many dehydrogenase-dependent redox reactions).

***p*H:** Logarithmic convention for indicating the H^+ (hydrogen ion; proton) concentration ($p\mathrm{H} = -\log[\mathrm{H}^+]$).

Photophosphorylation: Light-driven formation of ATP.

Photosynthesis: Light-driven overall formation of reduced organic compounds from CO_2.

Photosystem: Light-absorbing center of photosynthetic organisms; contains various pigments (including chlorophyll).

Phototroph: Organism that uses light as the primary source of energy.

Protonmotive Force ($\Delta\tilde{\mu}_{H^+}$): Driving force (potential energy) for the movement of protons (H^+) down a concentration and electrochemical gradient.

Pyruvate Dehydrogenase: Complex of three distinct enzymes that catalyze the overall oxidative decarboxylation of pyruvate forming acetyl-*S*-coenzyme *A*, CO_2, and NADH.

Redox Reaction: Overall oxidation-reduction reaction in which one substance (the oxidizing agent) gains electrons (and thus, is reduced) while another substance (the reducing agent) loses electrons (and thus, is oxidized).

Reducing Agent: Compound that donates electrons to another substance (e.g., NADH is a reducing agent in many dehydrogenase-dependent redox reactions.

Reduction: The gaining of electrons.

Respiration: Overall biological oxidation (by oxygen) of reduced organic compounds to CO_2 and H_2O.

Spontaneous Reaction: A reaction that will proceed in the direction written (if it proceeds at all) because the reaction components are not at equilibrium (i.e., a reaction with a negative ΔG). The word "spontaneous" implies nothing about the reaction rate.

Standard Free-Energy Change ($\Delta G^{0\prime}$): ΔG of a reaction when all components are $1M$.

Standard Reduction Potential (E_0'): Tendency of a half-reaction to proceed as a reduction under standard conditions at $p\mathrm{H}=7$ (expressed in terms of volts) relative to that of the $2H^+ + 2e^- \rightarrow H_2$ half-reaction which is -0.41 V.

Standard State: Reference state of $1M$ of all reactants and products used for tabulating and comparing energy changes accompanying a reaction.

Steady State: Condition in which the rate of formation of a substance equals the rate at which the substance is removed, resulting in a constant level of that substance over time. That steady-state level may or may not equal the equilibrium level. The intermediates of most metabolic pathways are in a steady state *in vivo.*

Substrate-Level Phosphorylation: Formation of ATP via discrete phosphorylated chemical intermediates (also called chemical coupling).

Symport: Simultaneous transport of two substances across a membrane, one of which moves down a concentration and/or electrochemical gradient driving the accumulation of the other "uphill."

Terminal Electron Acceptor: Final electron acceptor (oxidizing agent) in an electron transport chain.

Ubiquinone: Another name for Coenzyme Q.

Further Reading

Arnon, D. I. (1987), *Trends Biochem. Sci.* **12**, 39–42 (Jan.).

Atkinson, D. E. (1977), *Cellular Energy Metabolism and its Regulation,* New York: Academic.

Harold, F. M. (1986), *The Vital Force: A Study of Bioenergetics,* New York: Freeman.

Hinkle, P. C., McCarty, R. E. (1978), *Sci. Am.* **238** (3), 104–122.

Lowe, A. G., Jones, M. N. (1984), *Trends Biochem. Sci.* 11–12 (Jan.).

Mathews, C. K., van Holde, K. F. (1990), *Biochemistry. Parts III and IV.* Menlo Park: Benjamin/ Cummings.

McPartland, A. A., Segel, I. H. (1986), *Biochem. Educ.* **14** (3), 137–141.

Nichols, D. G. (1982), *Bioenergetics: An Introduction to the Chemiosmotic Theory,* New York: Academic.

Rawn, J. D. (1989), *Biochemistry. Parts 3 and 4,* Burlington: Patterson.

Rousseau, D. L., Jelinski, L. W. (1991), *Encyclopedia of Applied Physics,* Vol. 2, New York: VCH Publishers. pp. 571–608.

Segel, I. H. (1976), *Biochemical Calculations,* Chap. 4, New York: Wiley.

Stenish, J. (1989), *Dictionary of Biochemistry and Molecular Biology,* New York: Wiley.

Voet, D., Voet, J. G. (1990), *Biochemistry. Part IV,* New York: Wiley.

ENERGY CONVERSION

PAUL BERDAHL *Lawrence Berkeley Laboratory, Berkeley, California, U.S.A.*

Introduction 229
1. **Forms of Energy** 230
2. **General Properties of Energy Converters** 230
2.1 Energy Efficiency as a Figure of Merit 230
2.2 Simple Nonthermal Energy Converters 231
2.3 Thermal Energy Converters 231
2.4 Irreversible Converters and Onsager's Theorem 232
3. **Conventional Large-Scale Electric Generation Systems** 232
3.1 Sources of Electric Energy: Mainly Combustion 232
3.2 Turbines 233
3.3 Generators 233
3.4 Environmental Aspects of Large-Scale Energy Conversion 234
4. **Alternative Approaches for Energy Conversion to Electricity** 234
4.1 Photovoltaic Conversion 235
4.2 Photoelectrochemical Conversion 236
4.3 Thermoelectric Conversion 236
4.4 Thermophotovoltaic Conversion 237
4.5 Ferroelectricity, Pyroelectricity, and Piezoelectricity 237
4.6 Thermionic Conversion 238
4.7 Sodium Heat Engine (SHE) 238
4.8 Fuel Cells 238
4.9 Magnetohydrodynamics 239
5. **Efficient Conversion of Electricity to Electromagnetic Radiation** 239
5.1 Radio-Frequency and Microwave Radiation 239
5.2 Infrared Radiation 240
5.3 Visible Light Emission 241
Works Cited 241
Further Reading 242

INTRODUCTION

Energy conversion is a vast subject, involving the transformation of energy among all its myriad forms. Virtually all processes that involve change are made possible by energy conversion. From a global perspective, available energy originates primarily with the sun and energy-rich mineral substances (e.g., uranium) and is transformed into useful forms by plants, animals, and human activity. Organisms utilize whatever forms of energy are available for survival, gaining the capacity for activity and change by transforming energy from one form into another.

Some important but diverse examples of energy-conversion processes in biology are photosynthesis in plants and muscle contraction in animals. In chemistry examples include electrolysis of water, chemical reactions that absorb or produce heat, and chemical reactions that produce electricity. In nuclear physics, examples are fission and fusion reactions and radioactive decay. In solid-state physics some important examples are light absorption by materials to produce electricity or heat, and light emission in various processes. In mechanical and electrical engineering important examples include internal combustion and steam engines, mechanical machinery (e.g., gear trains), turbines, electric generators and motors, refrigeration technology, etc. These examples illustrate the broad spectrum of energy-conversion processes.

From the perspective of the disciplines of applied physics (as opposed, for example, to biology), the phenomena of electromagnetism form a central part of energy-conversion technology. For this reason and to limit the number of topics that must be discussed, the

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

primary emphasis of this article is placed on processes in which electricity is produced and on processes in which electricity is converted to electromagnetic radiation. Chemical processes such as combustion and nuclear processes such as fission and fusion will be regarded simply as providing heat sources for subsequent conversion to electricity. For a treatment of nuclear power the reader is referred to NUCLEAR ENERGY, FISSION, and NUCLEAR ENERGY, FUSION. For a treatment of energy conversion with emphasis on mechanical technology such as engines, see Russell *et al.* (1989). After a discussion of the forms of energy in Sec. 1, and the general properties of energy converters in Sec. 2, conventional large-scale systems are discussed in Sec. 3. Subsequently in Sec. 4 alternative approaches to electricity generation are treated, followed in Sec. 5 by presentation of several techniques for efficient conversion of electricity to electromagnetic radiation. A complementary article, STORAGE OF ENERGY, covers large-scale energy storage systems.

1. FORMS OF ENERGY

Forms of energy may be classified as kinetic or potential. Kinetic energy may be classified into macroscopic forms (e.g., rotation of a flywheel) and microscopic forms (e.g., a particle beam, or motion of atoms in a gas). Potential energy may be classified in terms of the fundamental forces involved. The "strong" nuclear force mediates interactions between nuclei, nucleons, quarks, etc., while the "weak" nuclear force is responsible for radioactive decays. The electromagnetic force controls the motions of charged particles in situations in which nuclear reactions are unimportant. This is the case for ordinary matter, as it is simply composed of assemblages of electrons and nuclei. The gravitational force, shared by all matter, is likewise a fundamental force.

From a technological point of view, electromagnetic forces are particularly important, since they govern all processes that are not nuclear or gravitational. Electromagnetic forms of energy may be further subdivided into macroscopic forms, such as in the electric field of a capacitor or the magnetic field of a solenoid, and microscopic forms. Microscopic electromagnetic energy is contained in the particular arrangement of electrons and nuclei in a given material; it may be further subdivided into thermal and nonthermal. The thermal portion of the energy in a system is that associated with the fact that the absolute temperature is not zero. It can further be associated with the disorder (entropy) of the system, but not merely with motions of particles in the system, as sometimes supposed. For example, the zero-point motion required by quantum mechanics occurs even at absolute zero. The kinetic energy associated with this motion is unavailable, but not unimportant: the zero-point motion of electrons is responsible for the rigidity of solid matter.

Changes in microscopic electromagnetic energy, occurring during energy conversion, may be classified into chemical and physical changes. Chemical changes require the breaking or formation of electronic bonds between atoms, as occurs in combustion or at the electrode of a battery. Physical changes are those in which no chemical reaction occurs; for example, ionization of a simple plasma or a phase transition such as vaporization of a liquid. The distinction between chemical and physical changes is not always sharp. For example, the absorption of light by a semiconductor can be described as either a chemical or physical change. From a chemical point of view an incoming photon displaces an electron bonding two atoms. Both the broken bond and the liberated electron are then free to move in the crystal lattice. From a physical point of view, the incoming photon excites an electron from the valence band to the conduction band, leaving a hole behind in the valence band. Thus, photon absorption by semiconductors is an example of a class of "interdisciplinary" processes for which the perspectives of both chemical and physical points of view are both quite important.

2. GENERAL PROPERTIES OF ENERGY CONVERTERS

2.1 Energy Efficiency as a Figure of Merit

For any energy-conversion process the obvious and conventional figure of merit is the efficiency, the ratio of useful output energy to input energy. High-efficiency processes are important, not just because they are more effective, but also because they reduce the amount of waste heat that must be dissipated. In many cases, the maximum theoretical value is unity. However, when input thermal energy is involved, for heat engines, the maximal (Carnot) efficiency is less than unity. Likewise, for heat pumps (see Sec. 2.3) the question of maximal efficiency is more complex. For these reasons it is often useful to define the "second- law efficiency," which is the ratio of the efficiency to the maximum value permitted by the second law of thermodynamics.

In specialized applications, the best figure of merit may not be energy efficiency, but rather some related quantity. For example, in illumination engineering one uses luminous efficacy, measured in lumens of light output per watt of input. The lumen is a specialized unit of the quantity of light that accounts for the fact that the human eye is more sensitive to some wavelengths than others.

2.2 Simple Nonthermal Energy Converters

Simple nonthermal energy converters are those in which energy is transformed without using or producing (i.e., dissipating) heat. This is an idealized case since some heat generation can always be expected, but it can be closely approached by some real systems. A lever arm is one example. An efficient electric motor is another example. Because of time-reversal invariance of virtually all the basic laws of physics, nonthermal processes are reversible. Dissipative mechanisms that generate heat (and entropy) are associated with a specific direction in time and thus cause irreversibility. In the limit of zero heat dissipation the lever used to produce a large force over a small displacement can be used in reverse to produce a small force over a large displacement. Likewise, the efficient electric motor can be used in reverse to generate electricity.

The lever example will serve to illustrate the properties of simple nonthermal energy converters more explicitly. The same type of analysis can be applied directly to simple mechanical systems such as ropes and pulleys, gear trains, etc., and by extension to various other devices such as motors. The relation between input and output displacement is

$$d_{\text{out}} = C d_{\text{in}} ,$$

where C is a constant determined by the geometry of the lever. Let the input and output forces be F_{in} and F_{out}, respectively. The work done by the lever is just $d_{\text{out}} F_{\text{out}}$. The relationship between input and output forces can be determined from the geometry of the lever, but it can be equally well be obtained from energy conservation ($d_{\text{out}} F_{\text{out}} = d_{\text{in}} F_{\text{in}}$):

$$F_{\text{out}} = C^{-1} F_{\text{in}} .$$

Any simple converter that has a linear relation between input and output displacements and in which energy is neither lost nor stored must have the same relation between input and output forces. The time-reversed motion of the simple converter merely involves the change in sign of the displacements, while the forces remain the same as those for forward motion. Energy is then transferred from output to input.

Friction can now be considered to see how dissipative processes affect converter performance. Frictional forces oppose the motion of the converter and therefore reduce the output force below the value $C^{-1} F_{\text{in}}$, accompanied by the irreversible generation of heat. Time-reversal invariance is now broken, as the frictional forces change sign when the direction of motion is changed, and heat is, of course, generated irrespective of the direction of operation.

2.3 Thermal Energy Converters

Thermal energy converters accept input energy Q in the form of heat at a high temperature T_2 while producing work W and rejecting the waste heat $Q-W$ at a lower temperature T_1. Since the input entropy Q/T_2 must be greater than or equal to the output entropy $(Q-W)/T_1$, it follows immediately that the maximum efficiency is given by the Carnot value:

$$\eta_C = W/Q = 1 - T_1/T_2. \tag{1}$$

Also of interest is the fact that the same considerations can also be used to describe the maximum (reversible) performance of a refrigerator (i.e., a heat pump). In this case each energy flow is time-reversed: $Q-W$ is negative and represents heat removal at the low temperature T_1, W represents input work, and Q represents heat rejected at the high temperature T_2. The technical figure of merit, called the coefficient of performance or COP, is thus given by

$$\mathrm{COP}=(Q-W)/W=T_1/(T_2-T_1),$$

where the first equality defines COP, and the second equality gives the maximal Carnot value. For further discussion of these issues, see HEAT ENGINES AND REFRIGERATORS and THERMODYNAMICS, EQUILIBRIUM.

The approach of efficiency and COP to the ideal maximum ("second-law") values in reversible machines is associated with indefinitely large heat-transfer areas and slow changes of time-dependent variables, so that the irreversible entropy increase can approach zero. However, economic considerations dictate that any conversion device must have limited size and process energy at a finite rate, so that it is clear that economic machines will not necessarily exhibit high second-law efficiencies.

2.4 Irreversible Converters and Onsager's Theorem

Various transport processes have the potential for use in energy-conversion devices due to the interaction of one process with another. An important example is the Peltier effect observed with thermocouples in which the transport of electricity causes the transport of heat. A reciprocal process is that the transport of heat in the same system causes the transport of electricity (Seebeck effect). Lord Kelvin (1854) postulated and Onsager (1931) proved that the coefficients describing these reciprocal phenomena must be equal. From Onsager's proof it is clear that the existence of reciprocity relations is quite general; they apply to systems undergoing irreversible transport that are not too far from thermal equilibrium. Suppose F_1 and F_2 are generalized forces (such as temperature or concentration gradients) that produce the fluxes J_1 and J_2 (e.g., heat and particle flow). The forces and fluxes can be chosen such that the rate of increase in entropy, per unit volume, is simply $J_1F_1+J_2F_2$. The linear relation between generalized forces and fluxes defines the kinetic coefficients L_{ij}:

$$J_1=L_{11}F_1+L_{12}F_2,$$

$$J_2=L_{21}F_1+L_{22}F_2.$$

Here, the nondiagonal elements of L describe how the processes labeled 1 and 2 interact with one another; Onsager's theorem gives the result that $L_{12}=L_{21}$. In the specific example of the Peltier and Seebeck effects at the junction of dissimilar materials, the result assumes the form

$$\pi=VT,$$

where π is the Peltier coefficient (heat flow per unit current), V is the thermoelectric power (open-circuit voltage per unit temperature difference), and T is the absolute temperature.

3. CONVENTIONAL LARGE-SCALE ELECTRIC GENERATION SYSTEMS

3.1 Sources of Electric Energy: Mainly Combustion

Large-scale commercial power plants process enormous quantities of energy. High efficiency is vital in order to reduce the large cost of the input energy. The existing technology of such plants represents the competition to be improved or supplanted by any new large-scale energy-processing techniques that are developed. Figure 1 shows the electric energy production, broken down by energy source, for the United States. The overall U.S. energy picture also includes large contributions from gas and oil for industrial, commercial, and residential heating, and oil use for transportation (not shown in the figure). Total U.S. energy use represents about 23% of world energy use of 3.2×10^{20} J/y (United Nations, 1992). As may be seen in Fig. 1, combustion of coal provides the dominant share of electricity output. Combustion of oil was important in the 1970s but declined during the 1980s. Natural gas is of importance because it is available, burns cleanly, and can be allowed to pass through combustion turbines. Nuclear power is a very important component in the energy mix, still increasing in energy output despite the present absence of orders for new plants. Hydroelectric power is the only major source that is converted directly into electricity without first being converted into heat.

With the exception of hydroelectric power, all mainstream electrical energy production is by means of heat engines. Either steam or combustion gases are expanded through turbines, and the shaft horsepower is converted to electricity. Many alternative energy-conversion schemes also use heat-engine technology, such as solar thermal conversion, geothermal energy, fusion power, etc. In the next two sections, the major components of this technology, turbines and generators, will be discussed further.

3.2 Turbines

Turbines are devices for converting the kinetic energy of a moving fluid to the mechanical energy of a rotating shaft, using nozzles and/or vanes. They replaced the reciprocating engines of the 19th century, because of decreased wear on moving parts and reduction in vibration.

Steam turbines are ubiquitous in large power systems, and can handle power up to above 1 GW, which is the full output of a large power plant. The rotation rate is generally 1800 or 3600 rpm in the United States, and 1500 or 3000 rpm in the rest of the world, synchronized to the desired 60- or 50-Hz electrical output frequency. Typical input conditions for a modest-sized (15-MW) unit (Hall *et al.*, 1987) are 755 K and 59 bar (900 °F, 850 lb in^{-2}). Temperatures above 811 K (1000 °F) are not usually used because the combination of high temperature, high pressure, and corrosive environment reduces reliability. If heat is rejected at 300 K, the corresponding Carnot efficiency is 0.60. The actual efficiency is approximately 0.29, which is the product of the ideal turbine (Rankine) cycle efficiency (0.37) and the turbine–generator efficiency (0.79). With a larger (30-MW) turbine and a higher input temperature of 811 K (1000 °F), the efficiency can be raised to 0.33.

The terminology "combined cycle" is generally used to describe the use of a gas turbine as a topping cycle, followed by a steam turbine. Topping cycles are those which utilize very high-temperature heat, which is then passed to a "bottoming" cycle. The combustion products pass directly through the turbine, which limits fuel to natural gas or clean-burning petroleum liquids. Turbine input temperatures can be rather high (1300 K), with output temperatures (800 K) suitable as input for a steam turbine. Efficiency in plant sizes of 100–600 MW is about 0.42 (Hall *et al.*, 1987), which is very high for conversion of thermal energy to electricity.

3.3 Generators

Generation of electricity from shaft rotational power is most easily (and efficiently) performed with an alternating-current generator. The simplest arrangement possible is a static magnetic field in which a loop of conducting wire is rotated. In accordance with Faraday's law, a voltage is induced in the rotating loop, proportional to the time derivative of the magnetic flux threading the wire loop. If the loop is connected to a load resistor, current will flow and energy will be

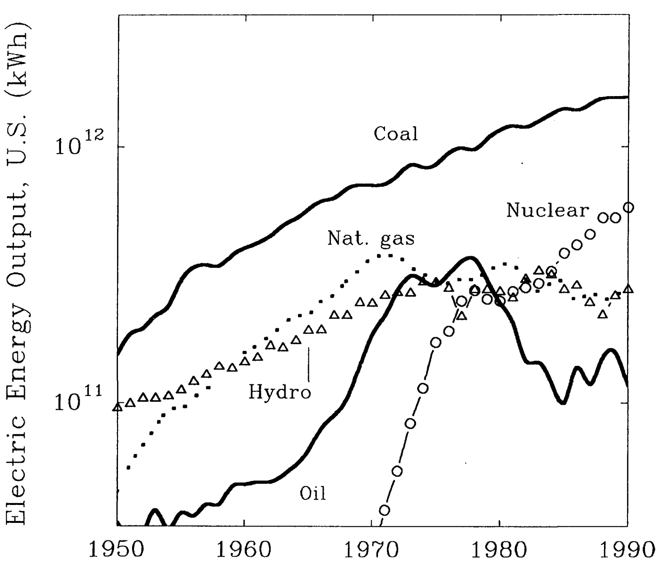

FIG. 1. Electric energy production in the United States for the period 1950–1990. Data are from the *Annual Energy Review 1990* of the U.S. Energy Information Administration (1991).

delivered to the load. If the wire resistance and bearing friction are small, such a generator has a high efficiency; it only suffers from the engineering difficulty of low output power per unit volume. The example of the wire loop can also be used to illustrate the reciprocity of simple, nonthermal, reversible energy converters. If an alternating current of suitable frequency is applied to the wire loop, it will rotate, acting as a motor. The motive force is the Lorentz force due to the magnetic field's action on the moving charges in the current-carrying wire.

For real generators (and motors), the output power per unit volume and per unit mass is nearly as important as the efficiency. For this reason, it is important to have high torque and high rate of shaft rotation. The high torque is achieved by maximizing the magnetic fields of both field and armature coils, using soft iron to enhance greatly the resulting magnetic induction. The magnetization of iron saturates at an induction of about 2 T, and so this is roughly the field that can be achieved in the gap between the iron in the armature and that of the rotor. Still higher fields are desirable and may be feasible in the future by the use of superconducting coils without iron.

Losses in a typical 10 000-kV A generator, operated at 80% power factor (8-MW output), include (Appleyard and Rosenberg, 1987) I^2R losses in the field and armature windings, losses due to magnetic hysteresis in the iron, and losses due to induced eddy currents caused by the varying magnetic field. Eddy-current losses in the iron are reduced by adding 3% silicon to increase resistivity and by the use of laminations. All of the losses can be effectively minimized, resulting in the remarkable overall generator efficiency of 0.968.

3.4 Environmental Aspects of Large-Scale Energy Conversion

The operation of large-scale energy-conversion systems has various environmental impacts. Some of these impacts are similar to any other large-scale activity such as construction of roads or buildings. More important, however, are the impacts of acquiring the energy resource and disposing of the by-products. Thus, for example, fossil-fuel plants emit oxides and sulfates that lead to acid-rain formation. These emissions must be controlled at the plant to levels that will minimize ecological damage downwind. Another problem with the combustion of carbonaceous fuels is the greenhouse warming of the earth due to CO_2 production. CO_2 hinders the earth's natural cooling process in which thermal infrared radiation is emitted into space. It is well established that the concentration of carbon dioxide in the atmosphere has been increasing substantially over the last several decades, and combustion is clearly a major source. In the future, therefore, it may be desirable to limit CO_2 production from combustion to assist in limiting global warming.

Nuclear plants have special problems related to the highly radioactive by-products that are produced. Thus it is necessary to guard against accidents that may release large amounts of radioactivity into the atmosphere, as in fact occurred at Chernobyl, U.S.S.R., in 1986 (Hohenemser, 1988). Also, some of the long-lived radioactive wastes produced by nuclear plants must be isolated from the environment for hundreds of years.

All energy-conversion systems release heat to the environment. This "thermal pollution" must be managed to avoid undue perturbation of the environment. For example, if river water is used for cooling, aquatic life is affected by the resulting temperature rise.

Certain energy-generation systems are often regarded as being particularly benign. These include hydropower, windpower, solar power, and perhaps geothermal systems. However, each system certainly does have specific environmental impacts. Perhaps the ultimate "clean" technology is energy conservation, since increases in the efficiency of energy use can actually reduce the environmental impacts of energy use.

4. ALTERNATIVE APPROACHES FOR ENERGY CONVERSION TO ELECTRICITY

Several important alternative schemes for producing electricity involve heat as an intermediate step, such as geothermal energy and solar thermal energy. These techniques will not be further discussed here, since the conversion of heat to electricity can be performed by the "conventional" techniques outlined in Sec. 3. Likewise, the use of turbines to convert the kinetic energy of a moving fluid to electricity, in hydroelectric conversion and for wind energy, will not be discussed. The reader is referred to the separate articles GEOTHERMAL ENERGY, SOLAR ENERGY, and WIND ENERGY and to the article on hydroelectric power by Wachter (1987).

4.1 Photovoltaic Conversion

Photovoltaic energy converters are heat engines in which thermal radiation from a high-temperature source is converted into electricity, using the electron-hole plasma excited in a semiconductor as the engine's working fluid. When photons with energies greater than the semiconductor's band gap are absorbed, electron-hole pairs are produced. The minority carriers thus produced (holes in *n*-type material, electrons in *p*-type material) carry most of the energy of the input photon. Each minority carrier diffuses until it falls down the built-in potential barrier at the *p-n* junction, becoming a majority carrier on the other side. In a well-designed photovoltaic cell, the short-circuit current corresponds to nearly one electron circulating in the external circuit for each absorbed photon (quantum efficiency of unity). The open-circuit voltage is comparable to the energy gap of the semiconductor (typically $\frac{1}{2}$ of this value) and increases slowly (logarithmically) with increasing input photon flux. This increase of output voltage with increasing photon flux is due to the higher thermodynamic quality of the input energy; it corresponds to higher effective radiation source temperature or, what amounts to the same thing, less entropy per unit energy in the input radiation. (The relatively smaller entropy input means that the energy efficiency is less constrained by the second law of thermodynamics. For an *ideal* photovoltaic converter, the open-circuit voltage is actually *determined* by the second law of thermodynamics.)

The "incidental" losses in photovoltaic cells, such as those due to series resistance, optical reflection, parasitic photon absorption, etc., can be reduced to levels of just a few percent, but there are two inherent losses that are particularly important. Photons with energies below the band gap cannot produce electron-hole pairs, and the corresponding energy is lost. Photons that have energies well above the band gap energy produce energetic electron-hole pairs in which the excess energy appears in the form of kinetic energy. This excess energy is thermalized by inelastic scattering with lattice vibrations (phonons) on the short time scale of roughly 10^{-13} s and becomes unavailable for conversion. The inherent losses in photovoltaic conversion clearly depend on the spectral distribution of the input radiation; a broad spectrum leads to large inherent losses, and a narrow spectrum can permit efficient operation. Thus the best conversion efficiencies may be expected in photovoltaic conversion of laser light. For example, in an experiment by D'Amato *et al.* (1992), a GaAs cell converted 826-nm (1.5-eV) radiation to electricity with an efficiency equal to 0.59.

For solar radiation input, we may regard the input spectrum as that of a blackbody with absolute temperature $T = 5800$ K:

$$b(E) = (2E^3/h^3c^2)[\exp(E/kT) - 1]^{-1}, \qquad (2)$$

where b is the radiant energy flow per unit solid angle, area, and photon energy at the surface of the sun. E is the photon energy, c is the speed of light, h is Planck's constant, and k is Boltzmann's constant. [Actually, Eq. (2) is an approximation because the solar spectrum is somewhat colder in the ultraviolet region and is modified by the earth's atmosphere; it totals about 1 kW m^{-2} at the earth's surface.] From Eq. (2), the average photon energy can be obtained:

$$\langle E \rangle = 2.701kT.$$

For $T = 5800$ K, we have $\langle E \rangle = 1.35$ eV, in the near-infrared region (920 nm). A systematic analysis of ideal photovoltaic conversion (Sze, 1981) shows that the best band gap is about 1.1 eV for highly concentrated solar radiation (efficiency, 0.37) and about 1.3 eV for unconcentrated radiation (efficiency, 0.31). Important solar-cell materials and their band gaps are Si (1.12 eV), InP (1.35 eV), GaAs (1.42 eV), CdTe (1.56 eV), and $CuInSe_2$ (1.02 eV). Materials with larger and smaller energy gaps are considered for schemes in which the solar spectrum is split into several components in order to raise efficiency.

As of this writing, the highest conversion efficiencies for silicon cells are approximately 0.27 with concentrated sunlight and 0.23 with unconcentrated light (1 sun). For GaAs cells, 0.29 has been achieved with concentrated light. Amorphous Si cells, which are less expensive to manufacture, have achieved efficiencies of 0.12. With a combined GaAs-GaSb tandem cell, in which the GaSb cell converts long-wavelength radiation passing through the GaAs cell, it appears possible to achieve efficiencies for concentrated radiation of over 0.35 (Fraas *et al.,* 1989). The total market for photovoltaic cells in 1988 was about 35 MW (Carlson, 1990), the cells being primarily for calculators and small off-grid power systems. For further information, see PHOTOVOLTAIC DEVICES.

4.2 Photoelectrochemical Conversion

The term photoelectrochemical conversion is applied to several processes in which the absorption of light causes the transport of electrical charge in an electrochemical cell. Photoelectrochemical cells (PECs) are similar to photovoltaic cells in that they utilize a semiconductor junction; however, the usual semiconductor *p-n* junction is replaced by a semiconductor-electrolyte junction. One example of a PEC is *n*-type CdSe (energy gap, 1.7 eV), in an aqueous solution of potassium ferricyanide–ferrocyanide (Licht and Peramunage, 1990). The efficiency of this device was 0.164 in sunlight. The stability of the output photocurrent was greatly enhanced (to at least hundreds of hours) by adjustment of electrolyte *p*H and cyanide concentration.

Semiconducting materials for photoelectrochemical conversion tend to exhibit corrosion, which limits cell lifetimes. Narrower band-gap materials such as Ge (0.6 eV) and even Si (1.1 eV) are particularly difficult to use; wider gap materials such as GaAs (1.4 eV) and the cadmium chalcogenides appear more promising.

An alternative approach to photoelectrochemical conversion uses dye molecules in solution to absorb the incident light, rather than a semiconductor. The excited dye molecule then injects an electron into the semiconducting electrode. In this scheme, it would appear that the function of the semiconductor is to serve as a transparent electrical conductor which, being fabricated from small particles, has a large surface area for adsorption of the active dye molecules. O'Regan and Gratzel (1991) have described a system using a charge-transfer dye (a brownish-red trimeric ruthenium complex), an iodide-triiodide redox electrolyte, and an *n*-type TiO_2 electrode composed of nanophase (10 nm) particles. The conversion efficiency was 0.079, stable over a period of 2 months.

4.3 Thermoelectric Conversion

Thermoelectric conversion is one of the oldest techniques for producing electricity. It has applications in systems in which it is important to produce electricity without moving parts. Examples include high-reliability space applications and gas-fired appliances for auxiliary electric output. One constructs a loop, called a couple, from two dissimilar materials. If one junction is heated and the other cooled, electricity flows in the loop. Useful electrical output may be obtained by cutting one side of the loop and inserting the load resistor across the gap. The open-circuit (Seebeck) voltage is given by the product of the thermoelectric power V of the couple and the input temperature difference. Two types of irreversible losses are important. First, the current circulating in the loop produces Joule I^2R heating. Also, ordinary heat conduction from the hot source to the cold sink dissipates input heat without contributing directly to the output electricity. Because of high thermal conductivities, metals exhibit low output power. On the other hand, the extraction of useful amounts of electrical power requires that the materials be electrical conductors.

In practice, narrow-band-gap semiconductors are the most useful materials. A single material can be used to form a couple if one leg is doped *p* type and the other *n* type. PbTe (energy gap, 0.32 eV) is an important material; typical dopants are Na for *p* type, PbI_2 for *n* type. Deviations in the Pb:Te ratio also affect conductivity type. Bi_2Te_3 (0.13 eV) and the alloy $(Bi_{1-x}Sb_x)_2Te_3$ are also used. The same materials are used for the reverse application, that is, to produce refrigeration (heat-pumping) effects when current is passed through the thermoelectric couple in the reverse direction. For conversion of high-temperature heat (say above 500°C, 773 K), silicon–germanium alloys are preferred (e.g., $Si_{0.8}Ge_{0.2}$), on account of good mechanical properties at elevated temperatures and somewhat larger band-gap energies (0.6–1.1 eV), which are desirable at high temperatures. Silicon and germanium have the same valence and crystal structure, but different atomic masses, a desirable situation that leads to lower (phonon-mediated) thermal conductivity than is measured for either element in pure form.

One of the most important applications for thermoelectric generators has been to provide a stable electrical power supply that lasts for years in space missions. The heat source is a radioisotope generator or a nuclear reactor. Efficiencies for operational systems working between 1073 and 344 K with outputs of a few watts are on the order of 0.08 (Dick *et al.*, 1985). Higher efficiency is difficult to obtain because of the irreversible heat conduction. However, there appears to be some promise for further improvements with advanced materials such as Ru_2Si_3 (Ohta *et al.*, 1991). For further information, see THERMOELECTRIC DEVICES.

4.4 Thermophotovoltaic Conversion

In this conversion scheme a heat source, which can be any high-temperature source such as combustion, nuclear, or solar, is used to heat a radiator into incandescence. This high-temperature radiation is transmitted to a photovoltaic cell, where it is converted to electricity. This technique is not yet used in applications, but it has thermodynamic advantages over thermoelectric conversion, due to the lack of a conduction path between the hot source and cold sink. For the best performance, the heat radiator should be a selective emitter that emits its thermal radiation only in a narrow wavelength band that is matched to the peak in the spectral conversion-efficiency curve of a suitable photovoltaic cell. It is also desirable that the temperature of the hot side be roughly matched to the band gap of the photovoltaic cell. Thus if one could use 5800-K radiation (as for solar photovoltaics), a band gap of 1–1.5 eV would be best. For 1500-K radiation, which is a more practical radiator temperature for extended material lifetimes, band gaps as low as 0.3 eV can be considered.

Summarizing and extending earlier work on thermophotovoltaic conversion, Nelson (1986) described systems with gas-fired mantles of Yb_2O_3 and Er_2O_3, which are selective emitters matched to Si and Ge photovoltaic cells, respectively. Another interesting development by Fraas *et al.* (1989) has been the GaSb photovoltaic cell (energy gap, 0.72 eV). Estimates by Morgan *et al.* (1989) of attainable conversion efficiency with the GaSb cell, without a selective emitter, are 0.14 for hot and cold temperatures of 1373 and 350 K.

4.5 Ferroelectricity, Pyroelectricity, and Piezoelectricity

Some crystalline materials of low symmetry, in the form of single crystals, have a permanent electric dipole moment. This dipole moment is due to the fact that, within each unit cell, the negative charge is displaced with respect to the positive charge. These materials are called ferroelectric. They also exhibit a temperature dependence of the electric dipole moment, a phenomenon referred to as pyroelectricity. Other crystalline materials can acquire a dipole moment proportional to an applied mechanical stress. These materials (which include the ferroelectrics) are termed piezoelectric. Examples of ferroelectric materials are KH_2PO_4, $BaTiO_3$, $LiNbO_3$, and GeTe. An example of a material that is piezoelectric but not ferroelectric is crystalline quartz. Other important piezoelectric materials are lead zirconate–lead titanate alloys, and the polymer polyvinylidene fluoride (PVF_2). See also FERROELECTRIC DEVICES.

It is possible to construct heat engines in which the pyroelectric material is the dielectric of a capacitor, which is alternately heated and cooled. For the simplest cycle, a load resistor is simply connected in parallel with the capacitor. At present, high output impedance and low efficiency remain problems. The efficiency is roughly 0.01 (12% of the Carnot efficiency in a limited temperature range), but further improvements appear possible (Olsen *et al.*, 1985). Energy densities as large as 0.9 MJ m^{-3} have been achieved using ferroelectric polymers as the dielectric (Olsen and Bruno, 1986).

Piezoelectric devices are used as transducers, such as microphones and acoustic buzzers, and for translation stages for small mechanical motions, such as those required in scanning tunneling microscopes. See PIEZOELECTRIC DEVICES. However, they have potential for electric generation as well. They convert mechanical motion (strain) to a high-voltage, low-current output. Conversion efficiencies can be high, of the order of 0.5, but output power per unit volume is small. Thus piezoelectric generators are limited to specialized applications; they certainly cannot compete with standard electrical generators fabricated with copper and iron.

4.6 Thermionic Conversion

Thermionic devices utilize input heat to eject electrons from an emitter (see THERMIONIC EMISSIONS). These electrons cross an evacuated or plasma-filled space and impinge on a collector electrode. While crossing the vacuum gap, the electrons lose some of their thermal kinetic energy and cause the buildup of electrostatic potential energy, which can be withdrawn by allowing the electrons to circulate in an external circuit.

When a vacuum is used between emitter and collector, the space charge due to the emitted electrons inhibits the emission of additional electrons. This problem can be minimized by the use of a very narrow gap (a few micrometers) or by the introduction of a plasma into a larger gap (a few hundred micrometers). The introduction of a plasma between the electrodes greatly complicates the analysis but improves the performance. Cesium vapor is often used for the plasma, a material that also coats the electrodes and facilitates electron emission by reason of its low work function. It appears that barium oxide is used as the electron emitter in the 6-kW Soviet TOPAZ system (Rasor, 1991). Sources of thermodynamic losses include collisions in the plasma, exchange of thermal radiation between emitter and collector, and heat loss from the emitter due to thermal conduction along the electrical lead.

Output power and efficiency are strongly increasing functions of the hot-side temperature. For the present technology, emitter temperatures should exceed 1600 K. Rasor (1991) quotes typical parameters for a thermionic diode converter as source temperature, 1800 K; sink temperature, about 800 K; output power, 71 W; and efficiency, 0.12.

4.7 Sodium Heat Engine (SHE)

This approach to electricity generation is also sometimes given the name alkali-metal thermoelectric converter (AMTEC). The technique uses a novel material called β''-alumina as a special solid electrolyte (fast ion conductor). It is composed of Al_2O_3 and a small percentage of sodium and potassium oxides. It is otherwise electrically insulating but permits Na^+ ions to diffuse rapidly through it.

In the SHE cycle (Weber, 1974), high-temperature (900–1300-K) liquid sodium is pressed against a β''-alumina plate with a porous metal electrode on the backside. Sodium atoms become ionized at the front surface of the plate and pass through it as Na^+ ions. This action causes an electric field to develop across the alumina plate, which is used to force the electrons to flow from the liquid sodium through the load, recombining with the sodium ions at the porous metal electrode. In order to maintain a large concentration gradient across the alumina electrolyte, the sodium is permitted to evaporate from the backside, being condensed at the heat rejection temperature of 400–800 K. After condensation, the sodium fluid is pumped back to the hot side of the device by means of a novel electromagnetic pump that has no moving parts. This pump utilizes a portion of the output electricity, passing a current through the liquid metal in the presence of a magnetic field from a permanent magnet. The Lorentz force thus created causes the small needed liquid flow.

Efficiencies as high as 0.19 have been obtained (Hunt *et al.*, 1981), working between 1073 and 473 K, and efficiencies of 0.25 are projected for 1-kW devices (McBride *et al.*, 1989).

4.8 Fuel Cells

Fuel cells are electrochemical devices in which a chemical fuel, which is usually H_2 but which may also contain CO, CH_4, etc., is oxidized to produce electric output. Thus they are similar to batteries in that they use chemical reactions to produce electricity (see ELECTROCHEMISTRY and STORAGE OF ENERGY), but they differ from batteries in that they are continuously supplied with fuel rather than relying on a finite internal supply. They are not heat engines, as all processes can take place at a uniform temperature. However, the chemical reactions taking place may cause the net evolution or absorption of heat.

The important phosphoric acid fuel cell will serve as an example. (The following information is summarized from a handbook on fuel cells by Kinoshita *et al.*, 1988). It uses concentrated phosphoric acid at about 200°C as the electrolyte. The electrolyte permits the desired mobility of ionic species between the anode (negative terminal) and the cathode (positive terminal). The operating temperature is elevated to provide adequate ionic conductivity. A Pt catalyst is used on both anode and cathode. A matrix of SiC is used to contain the electrolyte. The reaction at the anode is

$$H_2 \rightarrow 2H^+ + 2e^-,$$

and the cathode reaction is

$$\tfrac{1}{2}O_2 + 2H^+ + 2e^- \rightarrow H_2O.$$

Thus the net effect of the chemical reactions is that the hydrogen is oxidized to water, and two electrons circulate through the external circuit for each water molecule formed. The electrical circuit is completed by the diffusion of the protons (H^+ ions) through the electrolyte.

Phosphoric acid fuel cells appear near commercialization for large-scale power generation, after decades of research to reduce cost and extend operational lifetime. Cost estimates are only a factor of about 3 above that of conventional fossil-fuel-fired power plants for 11-MW units. An efficiency of 0.41 is expected.

Other important fuel cells, identified by their electrolytes and their operating temperatures, are the polymer electrolyte (80 °C), alkaline (100 °C), molten carbonate (650 °C), and solid oxide (yttria-stabilized zirconia, 1000 °C) cells. The polymer and alkaline types have been used for power generation in space. For example, a 7-kW alkaline fuel cell with KOH electrolyte has been used in the U.S. Space Shuttle.

4.9 Magnetohydrodynamics

In magnetohydrodynamic (MHD) systems a conducting fluid is forced to flow through a duct in the presence of a transverse magnetic field (see MAGNETOHYDRODYNAMICS). An electric current flows that is orthogonal to both fluid motion and magnetic field direction, due to the Lorentz force ($q\mathbf{v}\times\mathbf{B}$, the product of charge, velocity, and magnetic field) acting on the mobile charges in the conducting fluid. The effect was first demonstrated by Michael Faraday, using mercury. Various other fluids have been used, including seawater. (The reverse MHD effect is under consideration as a technique for ship propulsion.) However, most engineering investigations have focused on ionized gases as the working fluid.

The MHD generator is seen as a potential replacement for the turbogenerator. For example, heat from combustion of fossil fuel can be used to produce very high-temperature gas, which is expanded through the MHD duct, producing dc output current. Very high temperatures are required to maintain the gas partially ionized and, hence, electrically conducting. These high temperatures lead to materials lifetime problems. The plasma can be seeded with readily ionized material such as alkali metals to increase the conductivity available at a useful input temperature. Rosa (1987) describes an example of a topping cycle that operates between 2800 and 1400 K, as an addition to a fossil plant. Efficiencies in the range of 0.5–0.6 are projected for such a plant. Another MHD application is the generation of very high electric outputs for periods of a few minutes, using a rocket engine to provide the needed input plasma.

5. EFFICIENT CONVERSION OF ELECTRICITY TO ELECTROMAGNETIC RADIATION

An important class of energy-conversion processes are those in which electricity is used to generate electromagnetic radiation. A few particularly efficient processes are briefly enumerated here, ordered from low to high frequency.

5.1 Radio-Frequency and Microwave Radiation

Radio-frequency and microwave radiation is produced by connecting a source of alternating current to an antenna. Antenna efficiencies are high, and so the conversion of dc or low-frequency electrical power to high-frequency power is the key to efficient production of radiation. Modern semiconductor electronics generally makes it possible to convert electrical energy freely and efficiently from one frequency to another (see POWER ELECTRONICS). For example, even to convert from one dc voltage to another, it is convenient to use ac electrical power as an intermediate step. Such switching converters may use MOSFETs (metal oxide semiconductor field-effect transistors) to chop the input voltage at 100 kHz, an inductor to transform the voltage, and a filter to smooth the output voltage. Efficiencies above 0.9 are readily achievable. For high-power switching (thousands of amperes, thousands of volts), thyristors are available. These *p-n-p-n* devices can switch in a few microseconds. MOS transistors can switch efficiently at higher frequencies (to a few hundred megahertz), but there comes a point at which for high-power applications, vacuum tubes such as magnetrons and klystrons become more efficient than solid-state devices. This crossover is due, at least in part, to the fact that high-speed semiconductor devices are generally small, which conflicts with the need to dissipate heat at elevated power levels. Milnes (1980) gives an example of an *n*-channel Si MOSFET that can produce 16 W of power at 700 MHz with an efficiency of 0.26. Another example is work by Popovic *et al.* (1991), who used an array of GaAs FETs to produce 21 W of continuous power at 5 GHz with an efficiency of 0.2.

Magnetrons are microwave-frequency vacuum tubes that utilize the cyclotron motion of electrons in the field of a permanent magnet to convert dc electrical power to microwave frequencies. Electrons emitted from the negative cathode are subjected to a magnetic field orthogonal to their motion, which causes them to move in helical orbits. This motion is used to couple electromagnetic energy into a resonant microwave cavity. Pulsed magnetrons are used in radar applications as sources of high peak power (see RADAR TECHNOLOGY). Fixed-frequency magnetrons are used for heating applications in the familiar microwave oven and for various industrial processes. Farney (1982) cites an example of 25-kW output at 915 MHz, with an efficiency above 0.7. Weaver (1989) cites similar values for the maximum available continuous power available from commercial tubes at 1 GHz, declining to about 300 W at 30 GHz.

5.2 Infrared Radiation

In the infrared portion of the electromagnetic spectrum, thermal sources, such as blackbody cavities, become useful. The spectrum of radiation is given by Eq. (2). The efficiency of thermal emission can be high (unity), but there is a general difficulty in limiting this broad-band emission to a desired specific wavelength range.

One notable source of infrared radiation is the CO_2 laser (see LASERS, GAS). Among gas lasers, its efficiency is unusually high; high output powers are available as well. The radiative transition involves changes in the vibrational (and rotational) modes of the CO_2 molecule, unlike the electronic transitions used at shorter wavelengths. The most efficient transition is at 9.6 μm wavelength (31 THz), where the energy efficiency approaches 0.27 (Verdeyen, 1981). Hundreds of kilowatts of continuous output have been produced by large devices. And, as is the case with many lasers, *very* high instantaneous power outputs ($>10^{12}$ W) are achievable in very short pulses.

Another remarkable class of high-efficiency infrared laser is the semiconductor diode laser (see LASERS, SEMICONDUCTOR). Considerable progress has been made to extend operation into the visible portion of the spectrum, but high-efficiency devices operate primarily in the near-infrared region (near 1 μm). The mid-infrared region (near 10 μm) is also accessible but requires cryogenic operating conditions. Diode lasers are most often fabricated from the semiconductor GaAs, alloyed with AlAs for shorter wavelengths (larger band gap) or with InAs for longer wavelengths. Reported electrical-to-optical energy-conversion efficiencies now exceed 0.50 (Endriz *et al.*, 1992: 60 W output, 200-μs pulses, 50 Hz, 25 °C). These remarkable efficiencies indicate that each of the steps involved in producing light from a GaAs laser is highly efficient: low parasitic losses in getting the electrical energy in (series resistance), efficient production of minority carriers by *p-n* junction injection, efficient recombination of minority carriers with majority carriers to produce infrared light, and efficient coupling of the infrared energy out of the device aperture.

One important application of semiconductor laser diodes is for pumping of solid-state lasers. In one recent solid-state laser-pumping experiment a GaAs laser-diode bar emitted continuous 10-W output of 810-nm light through a 10-mm by 1-μm aperture (Shannon and Wallace, 1991). This light was used to pump a solid-state Nd:YAG laser (yttrium aluminum garnet crystal doped with Nd), which emits at the slightly longer wavelength of 1.06 μm. The output of the Nd:YAG laser was 1.9 W (optical-to-optical efficiency, 0.19).

5.3 Visible Light Emission

Visible electromagnetic radiation (i.e., light) is particularly important because it corresponds with the human visual range. The sensitivity of the human eye ranges from 400 to 700 nm, and peaks at 555 nm in the green. The use of light for illumination consumes roughly 25% of electricity production, and so the efficiency of lighting is economically very important. The quantity of visual light is measured in lumens (lm), a unit devised to account for the variation of the eye's sensitivity with wavelength. By definition, the *luminous efficacy* of light of 555 nm wavelength is 683 lm/W.

The simplest lighting device is the incandescent bulb. The tungsten filament is heated electrically to about 3000 K. Higher temperatures are not permitted by filament lifetime considerations. The spectral emission peaks in the near-infrared region and thus the efficiency for visible light emission is low. The luminous efficacy is roughly 15–20 lm/W and depends on filament temperature.

Fortunately, gas-discharge lamps have been designed that have much higher efficacy. The ubiquitous conventional fluorescent lamp, together with the ballast used to stabilize the discharge current, has a luminous efficacy of roughly 100 lm/W (Dakin, 1991). It employs an Ar and Hg plasma discharge in which the electron temperature rises well above that of the ions to about 11 000 K. Typical concentrations of species in the plasma are (Dakin, 1991), in units of cm^{-3}, Ar, 10^{17}; Hg, 2×10^{14}; free electrons, 2×10^{11}. The mercury in the plasma produces 254-nm radiation in the ultraviolet part of the spectrum. This radiant energy is absorbed by a phosphor coating on the inside of the glass tube that encloses the plasma, causing fluorescence in the visible spectrum. Different phosphors are used to achieve the color rendering desired in a particular lamp. The function of the Ar buffer gas is to prevent diffusion of the energetic electrons to the cool glass enclosure.

Low-pressure sodium lamps contain mixtures of Na and Ne and operate in a regime similar to fluorescent lamps (Dakin, 1991). The partial pressure of Na is about 1 Pa (10^{-5} atm). In this case the dominant emission is from the sodium transitions at 589 nm, in the yellow portion of the visible spectrum. There is no need for a fluorescent phosphor with its attendant losses, and so the luminous efficacy is very high, nearly 200 lm/W. However, the nearly monochromatic emission has the consequence that color perception of illuminated scenes is not possible.

One other important class of lamps is called "high-intensity discharge" lamps. These utilize high-temperature, high-pressure (near 10^5 Pa) electrical discharges in which the ions, atoms, and electrons are nearly in thermal equilibrium at temperatures of 4000–6000 K (Waymouth, 1991). At these temperatures the peak of the blackbody spectrum is in the visible range, which aids in the avoidance of copious infrared emission. Constituents for the plasma are chosen for strong optical emission in the visible part of the spectrum and include Hg, various metal halides, and sodium, usually in combination. The spectral content of the light varies with elemental composition of the plasma; efficacies range up to 100 lm/W.

List of Works Cited

Appleyard, R. E., Rosenberg, L. T. (1987), "Alternating-Current Generators," in: D. G. Fink, H. W. Beaty (Eds.), *Standard Handbook for Electrical Engineers*, New York: McGraw-Hill, Sec. 7-92.

Carlson, D. E. (1990), *Annu. Rev. Energy* **15**, 85–98.

Dakin, J. T. (1991), *IEEE Trans. Plasma Sci.* **19**, 991–1002.

D'Amato, F. X., Berak, J. M., Shuskus, A. J. (1992), *IEEE Photonics Technol. Lett.* **4**, 258–260.

Dick, P. J., Brittain, W. M., Brouns, R. C. (1985), in: *Proceedings of the 20th Intersociety Energy Conversion Engineering Conference*, Warrendale, PA: Society of Automotive Engineers, Vol. 3, pp. 536–541.

Endriz, J. G., Vakili, M., Browder, G. S., DeVito, M., Haden, J. M., Harnagel, G. L., Plano, W. E., Sakamoto, M., Welch, D. F., Willing, S., Worland, D. P., Yao, H. C. (1992), *IEEE J. Quantum Electron.* **28**, 952–965.

Farney, G. K. (1982), "Crossed-Field Tubes," in: D. G. Fink, D. Christiansen (Eds.), *Electronics Engineers' Handbook*, New York: McGraw-Hill, Sec. 9-40.

Fraas, L. M., Girad, G. R., Avery, J. E., Arau, B. A., Sundaram, V. S., Thompson, A. G. (1989), *J. Appl. Phys.* **66**, 3866.

Hall, D. H., Mizen, L. R., Allen, R. P. (1987), "Prime Movers," in: D. G. Fink, H. W. Beaty (Eds.), *Standard Handbook for Electrical Engineers*, New York: McGraw-Hill, Secs. 6-18 and 6-23.

Hohenemser, C. (1988), *Annu. Rev. Energy* **13**, 383–428.

Hunt, T. K., Weber, N., Cole, T. (1981), in: J. B. Bates and G. C. Farrington (Eds.), *Fast Ionic Transport in Solids*, Amsterdam: North-Holland, p. 263.

Kinoshita, K., McLarnon, F. R., Cairns, E. J. (1988), *Fuel Cells, A Handbook*, available from the U.S. NTIS Energy Distribution Center, P.O. Box 1300, Oak Ridge, TN 37831, pp. 1-156.

Lord Kelvin (1854), *Proc. R. Soc. Edinburgh*, **123**, Collected Papers I, 237-241.

Licht, S., Peramunage, D. (1990), *Nature* **345**, 330–334.

McBride, J. R., Novak, R. F., Schmatz, D. J., Copple, W. B., Brockway, J. T., Aron, N., Grab, G. A., Hunt, T. K. (1989), in: W. D. Jackson, D. A. Hull (Eds.), *Proceedings of the 24th Intersociety Energy Conversion Engineering Conference*, New York: IEEE, Vol. 2, pp. 683–688.

Milnes, A. G. (1980), *Semiconductor Devices and Integrated Electronics*, New York: Van Nostrand Reinhold, Sec. 6.4.

Morgan, M. D., Horne, W. E., Day, A. C. (1989), "Radioisotope Thermal Photovoltaic Application of the GaSb Cell," in: *Space Photovoltaic Research and Technology*, NASA Conf. Pub. 3107, pp. 349-358.

Nelson, R. E. (1986), "Thermophotovoltaic Technology," U.S. Patent No. 4 584 426.

Ohta, T., Vining, C. B., Allevato, C. E. (1991), in: *Proceedings of the 26th Intersociety Energy Conversion Engineering Society*, La Grange Park, IL: American Nuclear Society, Vol. 3, pp. 196–201.

Olsen, R. B., Bruno, D. A. (1986), "Pyroelectric Conversion Materials," in: *Proceedings of the 21st Intersociety Energy Conversion Engineering Conference*, Washington, DC: American Chemical Society, Vol. 1, pp. 89–93.

Olsen, R. B., Bruno, D. A., Briscoe, J. M. (1985), *J. Appl. Phys.* **58**, 4709–4716.

Onsager, L. (1931), *Phys. Rev.* **37**, 405–426.

O'Regan, B., Gratzel, M. (1991), *Nature* **353**, 737–740.

Popovic, Z. B., Weikle, R. M., Kim, M., Rutledge, D. B. (1991), *IEEE Trans. Microwave Theory Tech.* **39**, 193–199.

Rasor, N. S. (1991), *IEEE Trans. Plasma Sci.* **19**, 1191–1208. See also Rasor, N. S. (1983), "Thermionic Energy Converter," in: S. S. L. Chang (Ed.), *Fundamentals Handbook of Electrical and Computer Engineering*, New York: Wiley, Vol. II, p. 668.

Rosa, R. J. (1987), *Magnetohydrodynamic Energy Conversion*, Washington, DC: Hemisphere, Preface and Sec. 8.2.

Russell, C. R., Wailes, R., Budenholzer, R. A., Woodruff, E. B., Armstrong, L. V., Cromer, O. C., Baxter, A. D., Durant, F. C., III, Yamada, T., Saunders, R. M., Gregory, D. P., Heywood, J. B., Wolfe, R., Wernick, J. H., Soo, S. L., Lozinsky, M. G., Powers, P. N. (1989), in: P. W. Goetz (Ed.), *Encyclopaedia Britannica*, 15th ed. Chicago: Encyclopaedia Britannica, Vol. 18, pp. 448–548.

Shannon, D. C., Wallace, R. W. (1991), *Opt. Lett.* **16**, 318–320.

Sze, S. M. (1981), *Physics of Semiconductor Devices*, New York: Wiley, Sec. 14.2, and Ref. 13 therein.

United Nations (1992), *1990 Energy Statistics Yearbook*, United Nations Publication No. E./F.92.XVII.3, New York: United Nations, pp. 122 and 130.

U.S. Energy Information Administration (1991), *Annual Energy Review 1990*, available from National Energy Information Center, Energy Information Administration, Forrestal Bldg., Rm. 1F-048, Washington, DC 20585.

Verdeyen, J. T. (1981), *Laser Electronics*, Englewood Cliffs, NJ: Prentice-Hall, Sec. 10.4.4.

Wachter, G. F. (1987), "Hydroelectric Power Generation," in: D.G. Fink, H.W. Beaty (Eds.), *Standard Handbook for Electrical Engineers*, New York: McGraw-Hill, Sec. 9.

Waymouth, J. F. (1991), *IEEE Trans. Plasma Sci.* **19**, 1003-1012.

Weaver, F. J. (1989), "Magnetrons," in: F. F. Mazda (Ed.), *Electronics Engineer's Reference Book*, 6th ed. London: Butterworths, Sec. 32.4.

Weber, N. (1974), "A Thermoelectric Device Based on Beta-Alumina," *Energy Conversion* **14**, 1–8.

Further Reading

The extensive proceedings of the annual conferences entitled *Intersociety Energy Conversion Engineering Conference* contain a comprehensive view of ongoing research on energy conversion.

Large-Scale Electric Energy Systems

EPRI Journal, published by the Electric Power Research Institute, Inc., Palo Alto CA.

Fink, D. G., Beaty, H. W. (Eds.) (1987), *Standard Handbook for Electrical Engineers*, New York: McGraw-Hill.

The New Encyclopaedia Britannica (1991), "Energy Conversion," 15th ed. Chicago: Encyclopaedia Britannica, Vol. 18, pp. 332–413. A good historical treatment of practical energy conversion technology.

Williams, R. H., Larson, E. D. (1988), "Aeroderivative Turbines for Stationary Power," *Annu. Rev. Energy* **13**, 429–489.

General Properties of Energy Converters

Callen, H. B. (1985), *Thermodynamics and an Introduction to Thermostatistics*, New York: Wiley.

Kittel, C. Kroemer, H. (1980), *Thermal Physics*, New York: Freeman.

Landau, L. D., Lifshitz, E. M. (1980), *Statistical Physics*, 3rd ed. Oxford: Pergamon, Part 1, Sec. 120.

Yourgrau, W., Van Der Merwe, A., Raw, G. (1966), *Treatise on Irreversible and Statistical Thermophysics*, New York: Macmillan.

Alternative Approaches for Energy Conversion to Electricity

Angrist, S. W. (1982), *Direct Energy Conversion,* Boston: Allyn and Bacon.

Chandra, S. (1985), *Photoelectrochemical Solar Cells,* New York: Gordon and Breach.

Fahrenbruch, A. L., Bube, R. H. (1983), *Fundamentals of Solar Cells: Photovoltaic Solar Energy Conversion,* New York: Academic.

Kettani, M. A. (1970), *Direct Energy Conversion,* London: Addison-Wesley.

Sze, S.M. (1981), *Physics of Semiconductor Devices,* 2nd ed. New York: Wiley, Chap. 14.

ENGINEERING ACOUSTICS

RICHARD K. COOK, *4111 Bel Pre Road, Rockville, Maryland, U.S.A.*
STANLEY L. EHRLICH, *Stan Ehrlich Associates, Newport, Rhode Island, U.S.A.*

Introduction 245
1. **Transducers and Electroacoustics** 246
1.1 Condenser Microphones (Capacitive Transducers) 247
1.2 Calibration and Reciprocity 248
2. **Measurement of Physical Properties** 249
2.1 Sound in Solids, Liquids, and Gases 250
2.1.1 Sound Speed in Natural Waters 252
2.2 Piezoelectric Constants 252
2.2.1 α- and β-Quartz and Tourmaline 253
2.2.2 Polarized Ferroelectrics, Barium Titanate, Lead Zirconate–Titanate, and Polymers 254
2.3 Piezomagnetic Constants 256
2.3.1 Nickel, 2-V Permendur, and the Rare-Earth–Iron Alloys 257
2.4 Surface Tension, Surface Waves in Liquids, and Surface Acoustic Waves in Solids 258
2.5 Interior of the Earth 258
3. **Hearing and Physiology** 260
4. **Telephony** 261
5. **Sonar Methods** 262
6. **Sonics** 265
6.1 Flaw Detection 266
6.2 Thickness Metrology 267
Glossary 267
Works Cited 267
Further Reading 270

INTRODUCTION

The physical basis of sound and the properties of sound waves in solids, liquids, and gases are set forth elsewhere in this Encyclopedia (see ACOUSTICS, LINEAR; ACOUSTICS, NONLINEAR; ACOUSTIC PROPERTIES OF LIQUIDS; SOLIDS, ACOUSTIC PROPERTIES OF; ATMOSPHERIC ACOUSTICS; ACOUSTICAL INSTRUMENTATION). Included there are the theory of and mathematical physics for propagation and the methods and instruments for measurement.

This article describes the areas in the physical sciences and engineering to which acoustical methods have been applied. In the physical sciences these include measurements of elastic constants, piezoelectric and piezomagnetic constants, electric permittivity and magnetic permeability constants, and internal friction for solids. In liquids and gases, acoustical methods measure bulk modulus of elasticity, internal relaxation and friction, and surface tension. Areas in engineering include the control of the sounds of music and speech in rooms and auditoria (see ACOUSTICS, ARCHITECTURAL), the control of excessive noise in the environment (see NOISE, ACOUSTICAL), the detection of flaws in engineering materials, and the monitoring of progress toward collapse when such materials are periodically stressed. Military applications have included the location of enemy artillery by acoustical ranging. In underseas warfare, the location and tracking of enemy submarines are accomplished by means of sonar techniques. The locations of surface vessels and submarines under way are also determined by means of detecting the ship noise radiated and propagated through the water (see ACOUSTICS, UNDERWATER).

The audiometer is an electroacoustical instrument used by otologists and by audiologists (see ACOUSTICS, PHYSIOLOGICAL). It measures the amount of hearing loss experienced by persons with hearing varying from normal to moderately impaired to profoundly deaf. The instrument has been developed in parallel with prosthetic electroacoustical devices; e.g.,

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

hearing aids, which amplify speech, music, and incidental sounds of the environment for the benefit of persons with impaired hearing. Other bioacoustical applications are mainly for diagnosis, by means of ultrasound, of such human ailments as tumors, kidney stones, and cardiovascular irregularities, and for treatment of gallstones and kidney stones by lithotripsy (see BIOLOGICAL EFFECTS OF SOUND AND ULTRASOUND).

The application of acoustical methods in modern applied physics and technology is legion. This article includes a relatively small number of such methods. However, they have been applied also to other areas of acoustics: physiology and psychology of speech and hearing; music and musical instruments (*q.v.*); the acoustics of buildings; and the control of noise in the environment. Most of the latter applications have been to the development of acoustic instruments and to standard methods for measuring sound in the atmosphere.

In closing this introduction we note that the first theoretical research into the effect of a moving atmosphere on the propagation of sound waves was carried out by W. Voigt in the latter part of the 19th century. A few years later H. A. Lorentz applied this acoustical (theoretical) analysis to electromagnetic waves, and thereby arrived at what is called today the "Lorentz transformation." The work of Lord Rayleigh (1894), who some say covered everything one would ever want to know about acoustics, is worthy of note here.

1. TRANSDUCERS AND ELECTROACOUSTICS

Virtually all applications require at some stage the controlled production of vibration and/or radiation of sound waves. This is most often achieved by applying energy as electrical wave forms to an electroacoustical transducer. The article ACOUSTICAL INSTRUMENTATION has a general description of the various transducers used in acoustics.

Conversely, measurement of vibrations and radiated sound waves is done in the great majority of cases by means of calibrated transducers—microphones, hydrophones, seismometers, accelerometers, etc. The study of transducers has given rise to the science of electroacoustics, whose main objectives are the study of the conversion of wave forms of electrical energy into vibration and sound waves and, conversely, the study of the conversion of wave forms of sound energy into electrical wave forms in electronic circuits. Analysis of the electrical signal output of transducers, by means of sophisticated computerized electronic circuits, is included in the science of electroacoustics.

This is not to say, however, that the radiation of sound into various media is engineered in all cases by means of electroacoustic transducers. Powerful explosive sources, which may be called chemoacoustic transducers, are still used to study the propagation of sound to large distances in the atmosphere, in the ocean, and into the earth. The parameters of such sources are not normally as well controlled as in electroacoustics.

Most transducers are reversible; the physical mechanism that allows a transducer to receive sound as a microphone may also permit it to send sound as a loudspeaker. In water, the corresponding transducers are called a hydrophone and a projector. Furthermore, many reversible transducers have the same generalized transfer impedance for the two directions of energy flow, and these are called reciprocal transducers. When a transducer is known to be reciprocal, one can calculate its performance in transmitting from measurements in receiving, and vice versa. Similarly, one may test a reversible transducer to determine if it is reciprocal. A familiar analogous example in electrical engineering is the electric motor. It is an electromechanical transducer that converts electric currents into mechanical forces. When it is used reversibly as an electric generator (dynamo), it converts mechanical forces into electric currents.

Electrical phenomena are accompanied by two physically distinct fields—the electric (electrostatic) fields of electric charges and the magnetic fields of moving charges. Reversible acoustic transducers fall into two classes—one in which the physical action depends entirely on electric fields, the other in which it depends on magnetic fields.

The moving-coil loudspeaker is a generally useful transducer for radiation of sound into the atmosphere (Rice and Kellogg, 1925). The transducer is reversible, and it can be designed as a microphone. The basic form is a helical coil of conducting wire elastically restrained to move in its axial direction. The coil is in a fixed radial magnetic field, on which the entire transducer action depends. See Fig. 1.

Piezoelectric (solid) materials are useful for sound radiation into solids and liquids (Cady, 1964; Mason, 1950; Mason, 1981; National Defense Research Committee, 1946c; see also PIEZOELECTRIC DEVICES). The action is electrostatic, and the transducers are reversible. The action is due to electric dipoles distributed throughout the solid material. The dipole moment changes under applied mechanical stress. Quartz and tourmaline are naturally occurring piezoelectric crystalline minerals. Small single crystals can be used as seeds for artificially growing larger single crystals. Synthetic piezoelectrics can be produced by application of electrostatic fields to several types of plastic or polymer materials and ceramic materials. Such materials are called ferroelectric, and some of them are called electrets when polarized. Polyvinylidene fluoride (PVDF) is an example of a polymer material (Sessler, 1981), while barium titanate and lead zirconate–titanate are examples of the ceramics (Berlincourt, 1981, 1992).

Piezomagnetic (solid, basically magnetostrictive) materials may be considered to be analogous to the piezoelectric materials, except that the action is electromagnetic (National Defense Research Committee, 1946d; see also MAGNETOSTRICTIVE AND MAGNETOMECHANICAL DEVICES). There are no known purely piezomagnetic minerals, even single crystals (Cady, 1964). However, synthetic piezomagnetics can be produced by the application of electromagnetic fields to several types of magnetostrictive metals and metal-oxide ceramics. Such materials are ferromagnetic, and their initial discovery clearly preceded anything ferroelectric. Nickel and the cobalt–iron alloy 2-V Permendur are examples of the metals that found considerable application in the past, while rare-earth–iron alloys such as Terfenol D, a terbium–dysprosium–iron ternary alloy (Clark, 1988), have become important because of their extremely large strain response (a few parts per thousand) to magnetic field excitation.

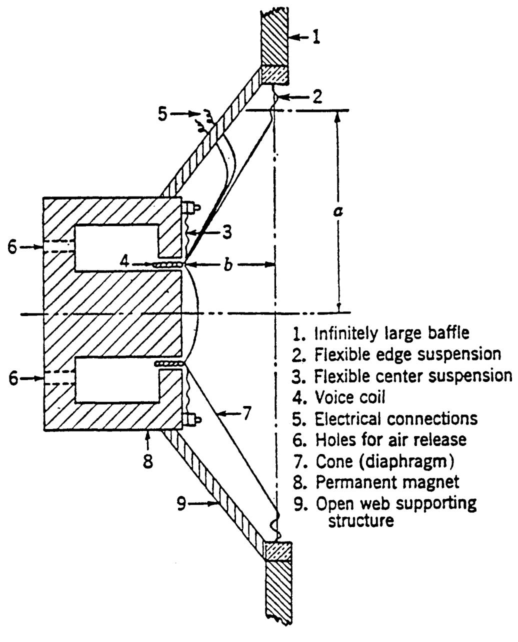
1
2
5
a
3
6
4
b
6
7
8
9
1. Infinitely large baffle
2. Flexible edge suspension
3. Flexible center suspension
4. Voice coil
5. Electrical connections
6. Holes for air release
7. Cone (diaphragm)
8. Permanent magnet
9. Open web supporting structure

FIG. 1. Construction of a moving-coil loudspeaker.

An important transducer for making accurate measurements on sound waves in air is the condenser microphone (Wente, 1917; Cook, 1941; Hilliard, 1950). The basic action is electrostatic. A detailed description of the microphone is given in Sec. 1.1. In the article TRANSDUCERS several other types of transducers are described. See also ACOUSTICAL INSTRUMENTATION.

1.1 Condenser Microphones (Capacitive Transducers)

We consider here in detail only the condenser microphone (Fig. 2). It is particularly important for making accurate measurements of sound waves in air. Conversely, as a result of its reciprocal action, it can be used to produce accurately known sound waves when a known voltage is applied to its electrical terminals. Some technical aspects of the condenser microphone are described in the article ACOUSTICAL INSTRUMENTATION.

The condenser microphone consists basically of a flexible diaphragm placed very close to a rigid metallic back plate, the two forming the plates of an electrical condenser. It responds to the *sound pressure* associated with the molecular motions in a sound wave in air, as does the ear of man. It is used with a nonfluctuating polarizing voltage, large in comparison with the smaller voltages produced by the acoustical action. The response of the microphone is the fundamental quantity required for its use as an accurate acoustical measuring instrument: at any particular fre-

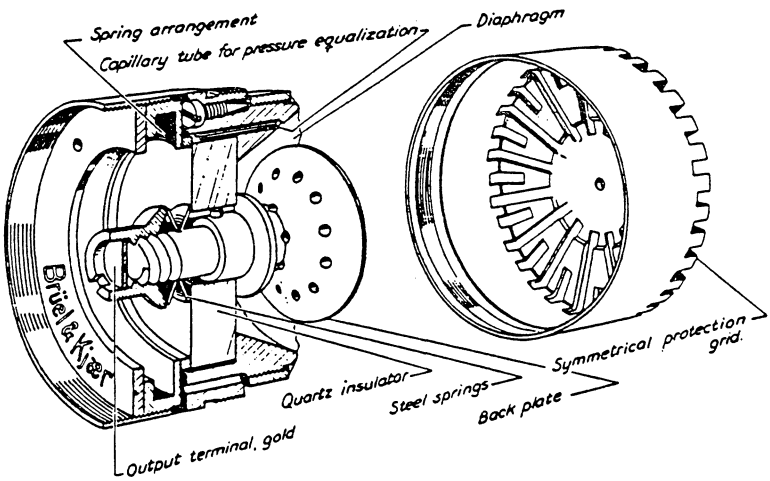

FIG. 2. Exploded view of conventional condenser microphone.

quency of sinusoidal oscillation, it is the complex-number ratio of the microphone's sinusoidal electrical output to the (sinusoidal) pressure acting on the diaphragm.

Practical requirements on standard condenser microphones are freedom from nonlinear distortion, uniformity of response of both amplitude and phase as functions of frequency, stability of response (including freedom from hysteresis) with temperature and humidity changes, and physical size small enough to minimize the amount of scattered sound power.

The condenser microphone appears to meet these requirements more satisfactorily than any other type of microphone. A knowledge of its construction is not sufficient to yield with requisite accuracy an absolute value of its response, which is best determined by means of the reciprocity method. Nevertheless, simple lumped parameters, which describe very well the microphone's response and electrical admittance as a function of frequency, can be determined empirically from the reciprocity calibration.

1.2 Calibration and Reciprocity

It is often important to have an accurate measure of the response, or output, of an acoustical instrument to a known stimulus or input. For convenience, let us assume that the acoustical instrument of interest is the condenser microphone, discussed above. The voltage or current response to a sound pressure or sound particle-velocity stimulus, determined quantitatively, represents a calibration of the instrument (Cook, 1940; Beranek, 1988; Bobber, 1970; National Defense Research Committee, 1946a, 1946b). Although there are many details to be concerned with, such as whether the input is a pure-tone sinusoid, a transient impulse, or a band-limited random noise, and whether the instrument is operating in a free field or in a different specified environment, we shall consider here only the simple case of pure tones in a free field. The general symbol for response of a transducer in a sound field (ANSI/ASME, 1984) is M_{yx}, where y is related to the output and x to the input. Thus the open-circuit voltage response to pressure in a sound field would employ the quantity symbol M_{Up}, and have the unit, based on the International System (SI), of volt per pascal, for which the unit symbol is V/Pa.

If one has a microphone with a known receiving sensitivity as a function of frequency, its output as a function of frequency may first be measured in a sound field created by a loudspeaker. Then another microphone, for which a calibration is desired, may be substituted for the reference microphone and the loudspeaker driven by the same electrical signal as a function of frequency. By comparing the outputs of the two microphones, one may deduce the receiving sensitivity of the second microphone. This procedure is called a secondary calibration, since it is dependent on knowledge of the calibration of a standard reference instrument. Most engineering and almost all manufacturing procedures employ this approach to perform necessary tasks speedily and at low cost.

When it is necessary to calibrate a primary standard, without any previous assumed knowledge of its response, there is a more detailed procedure available called a reciprocity calibration. To perform a reciprocity calibration it is necessary to have three transducers, one of which is a loudspeaker, one a microphone, and the third a reversible and reciprocal unit capable of serving as a loudspeaker and as a microphone. The concept of reciprocity (Rayleigh, 1894) was developed by several investigators, including Cook (1940 and 1941), and the basic mathematics can be found in Beranek (1988). The magnitude of the absolute receiving response of the reversible transducer in a free field is given by

$$|M_{Up}| = \{[E^{m}_{\mathrm{oc}}/i][e_{\mathrm{oc}}/e^{m}_{\mathrm{oc}}][2d\lambda/\rho c]\}^{1/2}, \quad (1)$$

where the first ratio in brackets is the open-circuit voltage at the microphone when the reversible transducer is driven with a constant current, the second ratio is the open-circuit voltage of the reversible transducer divided by the open-circuit voltage of the microphone when each is driven by the loudspeaker under identical conditions, and the third ratio is the magnitude of the reciprocity factor, $|J|$, for a small transducer:

$$|J| = 2d\lambda/\rho c. \quad (2)$$

In Eq. (2), d is the spacing between the loudspeaker and the reversible transducer (operating as a microphone) or the microphone, and also the spacing between the reversible transducer (operating as a loudspeaker) and the microphone; λ is the wavelength of the sound wave; ρ is the density of the medium, in this case air; and c is the speed of sound.

The calibration of the microphone, $|M^{m}_{Up}|$, is simply

$$|M^{m}_{Up}| = |M_{Up}|(e^{m}_{\mathrm{oc}}/e_{\mathrm{oc}}). \quad (3)$$

By the principle of reciprocity, the transmitting current response, $|M_{pI}|$, of the reversible transducer is given by

$$|M_{pI}| = |M_{Up}|/|J|. \quad (4)$$

The same technique is applicable with hydrophones and projectors in water and also in other liquids and gases.

2. MEASUREMENT OF PHYSICAL PROPERTIES

Acoustical phenomena and the propagation of sound are studied by measuring physical quantities in the medium of sound propagation (Beranek, 1986; Pierce, 1989). Measured quantities are the sound pressure at a point, the particle velocity, the optical density, deflection at a surface or interface, and many others. The passage of a sound wave is manifest as a continuous time series of limited duration. The wave form of the time series represents the variation in amplitude of the acoustical quantity. Problems such as losses by dispersion (scattering of a wave by turbulence or temperature gradients, for example) are treated in Sec. 2.1. The basic function of a receiving transducer (microphone, hydrophone, or seismometer) is to produce an electrical quantity with a wave form that is a copy of the wave form of its corresponding acoustical quantity. In gases and liquids the acoustical quantity usually observed by a receiving transducer (microphone, hydrophone) is the fluctuating sound pressure at a point within the medium. For solids the quantity observed may be the fluctuating particle velocity (vibration) at a point on the surface; seismic vibrations in the earth are measured in terms of surface motion by means of geophones and seismometers placed on its surface.

The electrical wave form of the transducer is received by an electrical network that performs the entire task of signal processing and display/readout. In many automated or computerized operations the display/readout is replaced by a feedback circuit to control the desired operation. For purposes of circuit analysis the transducer is represented in the network by an electroacoustical equivalent circuit (Mason, 1950) and the establishment of analogies between acoustical and electrical quantities (see Fig. 3). Some typical wave forms are continuous sinusoidal with harmonic overtones, sharp pulses separated by times long in comparison with the pulse width

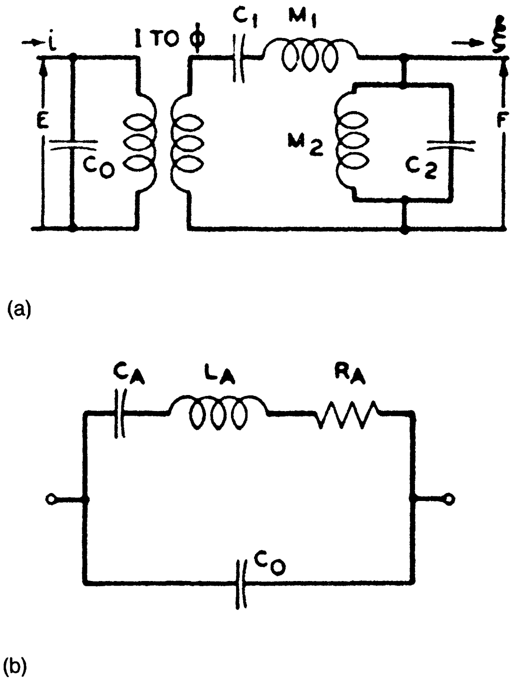

FIG. 3. Equivalent electromechanical and electrical circuits for a piezoelectric crystal (Mason, 1950).

(sometimes called low duty cycle), filtered random noise, amplitude- and/or frequency-modulated wave forms, etc. Important functions of the network are amplification, phase shifting or time delaying of sinusoids, calculation of the mean square amplitude over various time intervals, measurement of elapsed time between two pulses, and correlation of the wave form of the received pulse with a known or estimated replica of the transmitted pulse.

Some acoustical investigations and applications require measurement with two or more microphones. Then additional useful network functions are, for instance, multiplication of two voltages and averaging over a time interval (cross-correlation coefficient); and multiplication of two voltages, the Fourier components of one being phase-shifted by $\pi/2$ radians relative to the other (calculation of intensity in a sound field).

2.1 Sound in Solids, Liquids, and Gases

The passage of sound waves through a material is basically the transit of elastic waves obeying Hooke's law. In solids, liquids, and gases, the waves are composed of very small oscillations of pressure (stress), displacement (density), and temperature. The oscillations of audible sound waves in the atmosphere are typically of the order of 10^{-3} to 10^{-6} times the amounts of the ambient pressure, density, and temperature that enter into the atmosphere's equation of state.

In liquids and solids, surface waves also can be generated. Acoustical methods applied to such phenomena yield measurements of additional physical quantities. For example, the surface tension of liquids is readily measurable by means of surface waves. In piezoelectric crystals, ceramics, and polymers and in piezomagnetic metals and ceramics, electroacoustical methods of measurement are normally chosen to yield accurate values of piezoelectric and piezomagnetic moduli, respectively, which are very difficult to obtain by means of static mechanical and electrical stimuli.

We can see from the following considerations the advantages of using an acoustical method to determine the elastic constants and moduli of metals used commonly in engineering and physics. Many years ago the determination required the use of large static loads on large specimens in order to measure accurately enough the very small deflections produced by the application of the load. The development of electroacoustics has made it possible to measure the velocity of sound in solids. The phase velocity of sound waves (produced by suitable transmitting transducers) is measured by receiving transducers which pick up the sound. The measurements yield results on relatively small samples, of considerably greater accuracy than those obtainable from static measurements of the elastic moduli.

Simple measurements of sound velocity are also possible in crystalline materials, plastics, liquids, and gases. Where the sound velocity is independent of the direction of the sound waves in the material, as in gases and in most liquids, it is commonly called sound speed. The basic methods of measurement are very similar in all media. A sound wave is described by small oscillations of the medium, sinusoidal in time, and satisfying a linearized differential equation for propagation through the bulk volume of the medium. The sound can be idealized as plane waves having longitudinal oscillations, parallel to the direction of propagation. The physical quantity that is readily measurable (in liquids and gases) is the sound pressure, p, whose representation is

$$p=p_0e^{-\alpha x}e^{j(\omega t-\omega x/c)}=p_0e^{j(\omega t-kx)}, \tag{5}$$

where $k=\omega/c-j\alpha$ is the propagation constant. The wave is not dispersive. The phase velocity c is independent of the generalized frequency ω. The attenuation of the pressure amplitude takes place exponentially with distance along the propagation path and is denoted by α.

The differential propagation equations for isotropic homogeneous media show that

$$c^2=\gamma Y/\rho, \tag{6}$$

where γ is the ratio of the specific heat of the material at constant ambient pressure to the specific heat at constant volume; Y is the isothermal modulus of elasticity for the propagation mode; ρ is the density of the material. The acoustical measurements on longitudinal waves at practically all frequencies, in any solid or liquid material, yield adiabatic moduli of elasticity, γY. The basic physical mechanism is an oscillatory increase of temperature, as can be seen from the equation of state, when a material is compressed by passage of a sound wave and its density is increased.

If shear waves are the mode of propagation, then there is no density variation, hence no temperature variation, and so γ is replaced by 1. The acoustical measurement yields directly the shear modulus of elasticity.

The phase velocity in gases is in fact a function of frequency, as well as of temperature and ambient pressure as indicated above. In a monatomic gas (helium, argon, etc) sound propagation is essentially a transport-of-momentum process. There is variation of the sound speed—dispersion—and increasing absorption of sound energy at high frequencies. These latter effects arise from the internal viscosity and thermal conductivity of the gas and are collectively called classical phenomena. Thus at low frequencies the gaseous propagation takes place adiabatically, and at very high frequencies the oscillations occur almost isothermally.

In a polyatomic gas a principal mechanism during the passage of a sound wave is the time τ required for the internal rotational and vibrational degrees of freedom of the individual molecules to acquire energy from intermolecular collisions. At oscillation periods greater than τ, the sound speed is less than that at short oscillation periods (less than τ), This variation in sound speed is called "intrinsic dispersion." The entire phenomenon is usually called a "relaxation" effect. In the neighborhood of a relaxation frequency the sound absorption is increased to an amount two or three orders of magnitude greater than the absorption due to viscosity and thermal conduction. For carbon dioxide at 20 °C and at 1.0 atm pressure, the relaxation frequency is at 20 kHz; see Fig. 4. The time τ is determined from the acoustical relaxation frequency.

The attenuation factor α is, in liquids, due mainly to viscous losses arising from shearing motions in longitudinal waves. The propagation equations then show that α is directly proportional to ω^2. Therefore c and α/ω^2 characterize a liquid at any given temperature and pressure.

In most liquids, however, there is an additional attenuation. In natural waters this arises from the presence of dissolved salts, principally $MgSO_4$ in sea water. The excess absorption is due to relaxation mechanisms. The product $\alpha\lambda$ has maxima at various acoustical frequencies dependent upon the relaxation times of the transient chemical effects accompanying the transit of a sound wave. Acoustical measurements of absorption can therefore yield the relaxation times of the molecular interactions taking place in the liquid with

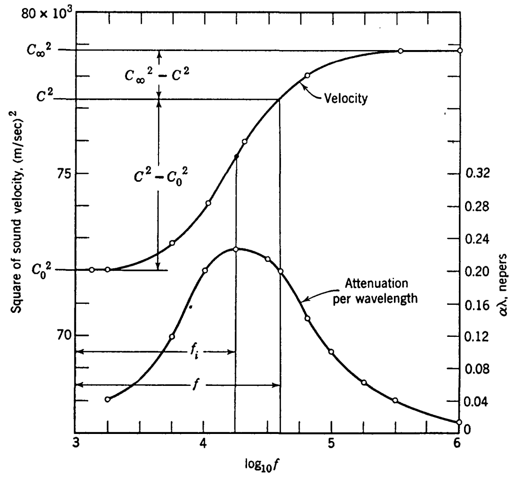

FIG. 4. Velocity dispersion and concomitant absorption in carbon dioxide (R. W. Leonard in Hueter and Bolt, 1955).

dissolved salts. In other liquids, the cause of the absorption may be a molecular-structure rearrangement and relaxation.

In solids the attenuation mechanisms are not yet all known. A conspicuous possible mechanism in polycrystalline materials is the scattering of sound energy (out of the main wave) by grains of the material. Acoustical measurements of α then yield in effect the cross-sectional area (for scattering of sound) by the polycrystalline grains.

2.1.1 Sound Speed in Natural Waters. The local speed of sound, c, must be known in detail for various maritime and military operations in many natural waters. In the ocean the variation of c with depth is particularly important. The gravity force gives rise to a stratified structure in which the variation of c with vertical depth is much greater than that in horizontal directions. Laboratory measurements over many years have resulted in an empirical formula that yields c as a function of the temperature T, depth d, and salinity S. These three quantities can be conveniently measured at sea, and the empirical formula for c (Wilson, 1960a, 1960b, 1962) that is simplified here,

$$c=1449.3+4.572T-0.0445T^2+0.0165d$$
$$+1.398(S-35), \qquad (7)$$

will then yield an "indirect" value for it. The ranges in which the above formula is applicable are $T=-3$ to $+30$ °C, $d=0$ to 10 000 m, and $S=33$ to 37 parts per thousand.

Measurements made in the ocean show that c has a minimum value at depths of about 1000 m in equatorial and temperature latitudes. The minima give rise to acoustical "SOFAR" (SOund Fixing And Ranging) channels in the oceans, centered at about 1000 m depth, in which sound is propagated in horizontal directions to large distances with only small loss. There are also sound channels that can develop at or near the surface as a consequence of seasonal changes in temperature.

The need for "direct" measurements of the variations of c with depth has led to the development of a velocimeter (Greenspan and Tschiegg, 1963). This instrument uses an electric "sing-around" principle. A short pulse is radiated by a piezoelectric crystal source; it travels a few centimeters through the water to a crystal hydrophone and is immediately reshaped as a fresh pulse by the electroacoustical circuitry. The repetition frequency of the pulse transit is a function only of the sound speed c. A calibration of the instrument is necessary; the velocimeter is immersed in water in a tank in which both T and d (or the hydrostatic pressure) can be varied. The upshot is that either by the direct velocimeter, or by the indirect empirical formula, sound speed c can be measured in natural water with an uncertainty no greater than ±0.02%.

2.2 Piezoelectric Constants

Piezoelectric crystals have unusually convenient physical properties. The crystal itself can be used as a transducer. A dynamical determination by acoustical methods can be made of piezoelectric, elastic, and dielectric constants. After the discovery of the piezoelectric effect in crystalline quartz by the Curie brothers, Jacques and Pierre (Curie and Curie, 1880), they made early measurements of the effect in several other crystals by static methods that they reported in their initial paper. The basic physical action in a piezoelectric crystal is the production of electric charges on the external surfaces of the crystal, when internal stresses are applied; transient mechanical loads were applied by the Curies, and the electric charges produced were measured by ballistic galvanometers and electrometers. The following year they reported the discovery of the inverse effect (Curie and Curie, 1881): electrostatic fields applied to a piezoelectric material generate stresses and/or strains. As indicated in the article TRANSDUCERS, the action is electrostatic and reversible.

The first applications of piezoelectricity were to an electrometer with increased sensitivity (c. 1896) and to an ultrasonic quartz–steel sandwich transducer (now called a *Tonpilz,* which means "sound mushroom") for use in detecting German submarines during World War I (c. 1915).

Beginning about 1920, electric circuits were used to create sinusoidal (in time) mechanical forces and motions in crystals at audio (i.e., sonic) frequencies. The electric charges and currents produced as a result of the piezoelectric effect were measured at the same audio frequencies by means of electronic circuits. Measurements on piezoelectric resonators were made particularly at and near the acoustical natural frequencies of vibration.

Such measurements were rapidly extended through ultrasonic frequencies and into the radio-frequency range. The frequency of oscillation of electric circuits may be accurately controlled and stabilized by the "high-Q" (small internal friction losses compared to stored piezoelectric energy) acoustical oscillations of crystals such as quartz (Cady, 1922; Nicolson, 1917). They are also the basis for narrowband filters designed for electronic circuits. The piezoelectric effect serves as the coupling between acoustical oscillations of the crystal and the electronic circuitry.

The basic constitutive equations for piezoelectricity (ANSI/IEEE, 1987) are given in tensor notation as

$$T_{ij}=c^{E}_{ijkl}S_{kl}-e_{kij}E_k\,, \tag{8}$$

$$D_i=e_{ikl}S_{kl}+\epsilon^{S}_{ik}E_k\,, \tag{9}$$

and in compressed matrix notation as

$$T_p=c^{E}_{pq}S_q-e_{kp}E_k\,, \tag{10}$$

$$D_i=e_{iq}S_q+\epsilon^{S}_{ik}E_k\,, \tag{11}$$

where $S_{ij}=S_p$ when $i=j$, $p=1, 2, 3$, and $2S_{ij}=S_p$ when $i\neq j$, $p=4, 5, 6$, and the summation convention for all repeated indices is understood. Other forms of the constitutive equation pairs, with the subscripts omitted, may be written as

$$S=s^{E}T+d_tE, \quad D=dT+\epsilon^{T}E, \tag{12}$$

$$S=s^{D}T+g_tD, \quad E=-gT+\beta^{T}D, \tag{13}$$

$$T=c^{D}S-h_tD, \quad E=-hS+\beta^{S}D. \tag{14}$$

In the above equations, T is the stress in pascals, S is the strain and is dimensionless, D is the electric displacement in coulombs per square meter, and E is the electric field in volts per meter; c is the elastic stiffness in pascals, s is the elastic compliance in reciprocal pascals, ϵ is the permittivity in farads per meter, and β is the impermittivity in meters per farad; the piezoelectric constants are d in meters per volt (= coulombs per newton), e in coulombs per square meter [= newtons per (meter volt)], g in meter volts per newton (= square meters per coulomb), and h in volts per meter (= newtons per coulomb). The subscript t indicates a transposed matrix and the superscripts denote variables that are held constant in the determination of the associated parameter.

Piezoelectric crystals are now used extensively throughout the communications industries: as transducers, as frequency-control elements, and as filters in electronic circuits. Ferroelectric ceramics and polymers have taken over some of these applications in recent years—for example, transducers for underwater sound applications (sonar), accelerometers, and microphones.

2.2.1 α- and β-Quartz and Tourmaline. The following summary describes application of an acoustical method to determine the piezoelectric constants of quartz as a function of temperature.

α-quartz is the stable crystalline form at temperatures below 573 °C. It has symmetry designated 32 in the international notation and D_3 in the Schoenflies notation. When heated above 573 °C, the crystal structure changes suddenly to symmetry 622 and D_6, respectively. This form, stable above 573 °C, is called β-quartz.

α-quartz has two nonzero piezoelectric constants (tensor components), d_{11} and d_{14}. d_{11} is the electric polarization per unit normal stress on a plane perpendicular to the electric axis. d_{14} is the electric polarization per unit shearing stress on a plane perpendicular to the optic axis. For more precise definitions of piezoelectric constants see ANSI/IEEE (1987), and for a summary of the measurements of the d constants for α-quartz see Cady (1964).

For β-quartz, its crystal symmetry requires that $d_{11}=0$. However, d_{14} is not zero. Up to the time in 1950 when the acoustical method being described here was applied, no quantitive measurements of d_{14} for α-quartz at high temperatures nor for β-quartz at temperatures above 573 °C had been reported in the technical literature.

Another technique of measurement was a determination of the equivalent electroacoustical circuits (Fig. 3) (Van Dyke, 1928) of long, relatively thin bars (Fig. 5) driven electrically. The voltages and currents were sinusoidal at frequencies near resonance for the fundamental longitudinal modes of the bars. The equivalent circuits were determined from measurements using an electrical Q meter (Figs. 6 and 7). The only other physical quantities needed were the physical dimensions and orientation angles of the bars and the density of quartz. The resonance frequencies of the bars were in the range from 35 to 60 kHz. The piezoelectric constants for α-quartz are given in Fig. 8. Since quartz belongs to one of the 11 crystal classes (point groups) having no center of inversion or plane of symmetry, two different types of the same species may exist. Each type is the mirror image of the other; neither type can be made to look exactly like the other by a simple rotation. If one type appears right-handed with right-handed rectangular axes, the other will appear left-handed. The left-handed type is regarded as the crystallographic inversion of the right-handed type, rather than its mirror image.

Polar crystals, such as tourmaline, have a polar axis along which an electric polarization develops when the crystal is subjected to hydrostatic pressure. This can also be produced by a sound pressure at audio frequencies. Tourmaline is in the trigonal–ditrigonal polar class, which is No. 19 of the 32 crystal classes and has the symmetry $3m$ or C_{3v}. A determination of the piezoelectric d constant under hydrostatic pressure d_h, where

$$d_h = 2d_{31} + d_{33}, \tag{15}$$

was made on a black California tourmaline crystal. The hydrostatic pressure was applied by means of sound pressures at several frequencies. The source of sound was a condenser microphone operating as a loudspeaker to produce an accurately known sound pressure. The charge on the crystal produced by the sound pressure was measured electronically. The result for the measurement of d_h was 2.23×10^{-12} C/N. The crystal motions took place adiabatically. The accompanying thermal oscillations could produce pyroelectric charges in polar tourmaline. However, it appears that such charges would have a negligible effect on the piezoelectric measurement, and the above result is the same as that which would be obtained statically. Applying measurements of d_{31} and d_{33} measured by Voigt (1910) (also given in Mason, 1950) in Eq. (15) agrees with the above value of d_h within 10%.

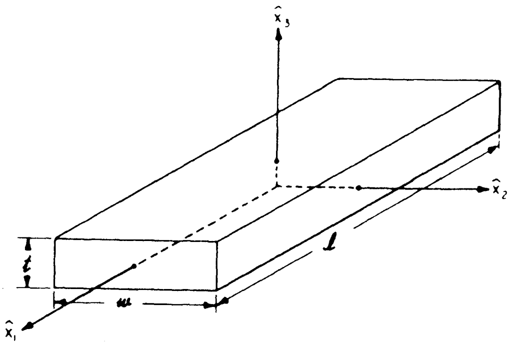

FIG. 5. A rectangular bar positioned in a Cartesian coordinate system (ANSI/IEEE, 1987).

2.2.2 Polarized Ferroelectrics, Barium Titanate, Lead Zirconate–Titanate, and Polymers. Ferroelectric crystals and polycrystalline materials are analogous to ferromagnetic materials in that they have polarization-field and strain-field curves in the form of hysteresis loops. The principal prop-

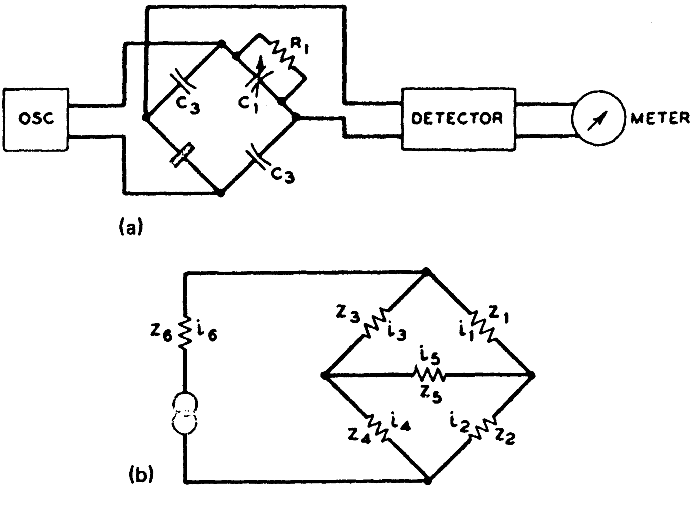

FIG. 6. Bridge circuit for obtaining quantitative measurements of piezoelectricity (Mason, 1950).

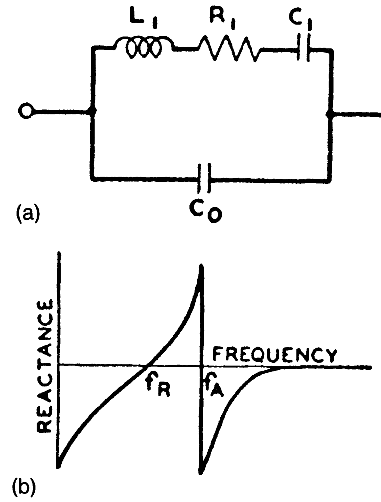

FIG. 7. Equivalent electrical circuit of a piezoelectric crystal and a plot of its reactance vs frequency (Mason, 1950).

erties of these materials tend to vary considerably with temperature. The earliest known ferroelectric crystal is Rochelle salt ($NaKC_4H_4O_6 \cdot 4H_2O$), which is ferroelectric in the temperature range from −18 to +24 °C (the two Curie temperatures). While all piezoelectric crystals, in addition to the first-order (linear) piezoelectric effect, also have a second-order electrostrictive effect in which a strain occurs proportional to the square of the applied electric displacement, it is small except in ferroelectric materials. About 1942 it was discovered that barium titanate, which occurs as a polycrystalline ceramic as well as a single crystal, is ferroelectric at temperatures up to 120 °C, its Curie temperature (Fig. 9). Mason (1950) made the excellent prediction: "The amount of motion generated in barium titanate ceramic is larger than that in magnetostrictive materials, and it appears that barium titanate may be an important electromechan-

General form of the matrices

$$\left(\begin{array}{c|c} c^E & e_t \\ \hline e & \epsilon^S \end{array}\right)$$

c^E in 10^9 Pa

e in 10^{-2} C/m^2

ϵ^S in 10^{-12} F/m

Right-handed quartz

$$\left(\begin{array}{cccccc|ccc}
86.74 & 6.99 & 11.91 & 17.91 & 0 & 0 & 17.1 & 0 & 0 \\
6.99 & 86.74 & 11.91 & -17.91 & 0 & 0 & -17.1 & 0 & 0 \\
11.91 & 11.91 & 107.2 & 0 & 0 & 0 & 0 & 0 & 0 \\
17.91 & -17.91 & 0 & 57.94 & 0 & 0 & 4.06 & 4.06 & 0 \\
0 & 0 & 0 & 0 & 57.94 & 35.82 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 35.82 & 39.88 & 0 & -34.2 & 0 \\
\hline
17.1 & -17.1 & 0 & 4.06 & 0 & 0 & 39.21 & 0 & 0 \\
0 & 0 & 0 & 4.06 & 0 & -34.2 & 0 & 39.21 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 41.03
\end{array}\right)$$

Left-handed quartz

$$\left(\begin{array}{cccccc|ccc}
86.74 & 6.99 & 11.91 & 17.91 & 0 & 0 & -17.1 & 0 & 0 \\
6.99 & 86.74 & 11.91 & -17.91 & 0 & 0 & 17.1 & 0 & 0 \\
11.91 & 11.91 & 107.2 & 0 & 0 & 0 & 0 & 0 & 0 \\
17.91 & -17.91 & 0 & 57.94 & 0 & 0 & -4.06 & -4.06 & 0 \\
0 & 0 & 0 & 0 & 57.94 & 35.82 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 35.82 & 39.88 & 0 & 34.2 & 0 \\
\hline
-17.1 & 17.1 & 0 & -4.06 & 0 & 0 & 39.21 & 0 & 0 \\
0 & 0 & 0 & -4.06 & 0 & 34.2 & 0 & 39.21 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 41.03
\end{array}\right)$$

NOTE: The matrix e_t is the transpose of the piezoelectric matrix e.

FIG. 8. Elasto-piezo-dielectric matrices for right- and left-handed quartz (ANSI/IEEE, 1987).

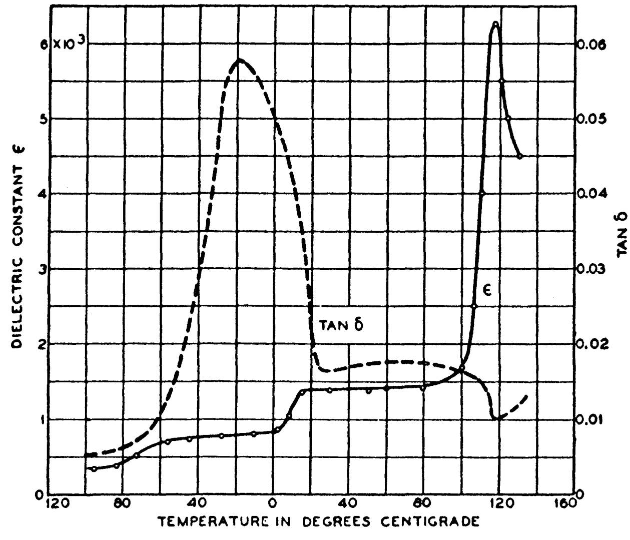

FIG. 9. Dielectric constant and loss tangent (ratio of the imaginary component to the real component of the dielectric constant) of barium titanate as a function of temperature (Mason, 1950).

ical transducing element." That did, in fact, prove to be true for several years until the discovery in the 1960s that ternary ceramic alloys, the lead zirconate–titanates, generated even larger strains (Berlincourt, 1981). Small amounts of additive substances are employed to modify the properties of ceramics by vendors, who often give their specific formulations trade names. All of the ceramic materials are normally polarized by applying a high electric field, usually at a temperature near the Curie temperature, and allowing the ceramic pieces to cool slowly to room temperature before removing the applied field. After the removal of the field, a ceramic piece is left in a remanent state of polarization and it is effectively piezoelectric with a linear strain-to-field ratio as long as the applied field is not excessive. For practical engineering application, it is not common to apply a dc polarizing voltage or to drive an element beyond its linear range.

Piezoelectrically polarized polymer materials, such as polyvinylidene fluoride (PVDF or PVF_2) discovered by Kawai (1969), are very flexible and may be polarized either electrically or by mechanically stretching in a direction normal to the polarization direction, or both (Sessler, 1981). When polarized electrically only, the two normal directions are effectively identical and $d_{31}=d_{32}$. However, when all or part of the poling is the result of mechanical stretching, these constants are unequal (Wilson, 1985). Since much greater polarization can be obtained by the latter method with a resultant much higher output voltage to applied mechanical strain capability, it is usually the polarization method of choice. For some applications this can lead to undesirable nonuniformity in response in the two primary directions normal to the polarization, that is, the stretch direction and the direction normal to it. By mounting a pair of polymer elements with their stretch directions orthogonal, Ricketts (1992) has shown that at least the two lowest modes of a plate may be neutralized to provide an effectively uniform response that is useful to increase receiving frequency capability.

Among recent developments are flexible ceramic materials, in which nonactive flexible materials are combined into composite ceramics for purposes such as to increase the hydrostatic piezoelectric constant by decreasing d_{31} and, sometimes, increasing d_{33}. Piezoelectric rubber compounds with conductive rubber electrodes are examples of commercially available flexible ceramic materials.

2.3 Piezomagnetic Constants

The discovery of the phenomenon of magnetostriction is attributed to James Prescott Joule in 1842, but it was not until 1865 that the inverse effect was discovered by Villary. Magnetostrictive materials are analogous to the ferroelectric electrostrictive materials discussed above. Because the earlier known materials were all ferromagnetic metals, they were generally more rugged than the piezoelectric crystals. As better fabrication techniques were developed, the advantage of ruggedness has all but disappeared. By polarizing such materials to make them effectively linear, one obtains the analog of the polarized piezoelectric materials; these are called piezomagnetic (Cady, 1964). They differ in that nonmagnetic phenomena were characterized by only 32 crystallographic classes, whereas magnetic phenomena are characterized by 90 crystallographic classes (Birss, 1964). The additional 58 crystal classes represent conditions for which changes in directions of magnetic moments are a significant symmetry feature. Many of these additional classes may be simulated with piezoelectrics when moving charges are involved; however, that is beyond the scope of this article.

For most practical applications of piezomagnetic materials, the dimensional considerations are much simplified, sometimes involving only one dimension when the direction of polarization and the principal direction of mechanical strain are identical. Shortly following the publication of the first IEEE standard on piezoelectricity in 1949, the latest revision being ANSI/IEEE Standard 176 (ANSI/IEEE, 1987), it became apparent that one could more easily compare transducers employing the analogous mechanisms of piezoelectricity and piezomagnetism if the same mathematical formulation is used (Ehrlich, 1952). The result was the first IEEE standard on piezomagnetism in 1971, the latest revision being IEEE Standard 319 (IEEE, 1990).

The basic forms of the constitutive equations for piezomagnetism follow the same format as Eqs. (12)–(14):

$$S=s^{H}T+d_{t}H, \quad B=dT+\mu^{T}H, \tag{16}$$

$$T=c^{H}S-e_{t}H, \quad B=eS+\mu^{S}H, \tag{17}$$

$$S=s^{B}T+g_{t}B, \quad H=-gT+\nu^{T}B, \tag{18}$$

$$T=c^{B}S-h_{t}B, \quad H=-hS+\nu^{S}B, \tag{19}$$

where Eq. (16) is the analog of Eq. (12), Eq. (17) that of the simplified Eqs. (10) and (11), Eq. (18) that of Eq. (13), and Eq. (19) that of Eq. (14). The magnetic variables, the magnetic field strength H in amperes per meter and the magnetic flux density B in teslas, are treated as the analogs of the respective electric variables E and D; the magnetic constants are the permeability μ in henrys per meter and the reluctivity ν in meters per henry; and the piezomagnetic constants are d in webers per newton (=meters per ampere), e in teslas [=newtons per (ampere meter)], g in reciprocal teslas (=ampere meters per newton), and h in newtons per weber (=amperes per meter). Where necessary, a superscript m should be used above the piezomagnetic constants to distinguish them from their piezoelectric analogs.

Applications of piezomagnetic materials have included underwater transducers, oil-well logging, drills, soldering tools, and situations where heat conduction away from a relatively closed volume is of major concern. Because eddy-current losses in magnetostrictive metals, which are much better electrical conductors than most piezoelectrics, tend to increase with the square of the frequency, it is usually necessary to use relatively thin laminated materials, similar to those in electromagnetic transformers. Although many designs for use at frequencies up to 100 kHz have been produced, the modern trend is to apply magnetostrictive materials to transducers at low sonic frequencies.

2.3.1 Nickel, 2-V Permendur, and the Rare-Earth–Iron Alloys. Of the three ferromagnetic elements iron, cobalt, and nickel, the more favorable magnetic and piezomagnetic properties of nickel made it the material of choice in the early 1940s. Permanent magnets were most often used with soft annealed transducers to provide the magnetic flux-density bias necessary for linear operation (National Defense Research Committee, 1946d). Development of iron–cobalt alloys was at first found to result in excessively brittle materials; then, by adding small amounts of materials such as vanadium and chromium to the alloy composition, the brittleness problem was effectively overcome. One such alloy that found wide use in the 1940s and early 1950s was 2-V Permendur, an Allegheny Ludlum Co. trade name for an alloy of 49% iron, 49% cobalt, and 2% vanadium. Advantages of 2-V Permendur over nickel included its greater piezomagnetic strain capability and its high remanent polarization after magnetization, which avoided the need for biasing magnets (National Defense Research Committee, 1946d). Then, on the development of the ferroelectric ceramics noted above with their much higher strain capability than then-known magnetostrictive materials, they became preferred for most applications.

In 1963 measurements at low temperatures of the magnetostrictive strain capabilities of terbium and dysprosium of approximately 1% leapfrogged the capabilities of the ceramics. This has generated renewed interest in the piezomagnetic properties of all the rare-earth–iron compounds, which is tempered by the costs of the rare-earth materials and the fabrication of the alloys. A terbium–dysprosium–iron ternary alloy developed at the Naval Surface Weapons Center in White Oak, MD (formerly the Naval Ordnance Laboratory) and called Terfenol-D has been receiving the greatest attention recently (Clark, 1988). It appears to have reasonably favorable mechanical, magnetic, and piezomagnetic properties in combination.

2.4 Surface Tension, Surface Waves in Liquids, and Surface Acoustic Waves in Solids

The free surface of a liquid can support waves, in addition to the elastic waves within the bulk volume discussed above. The surface waves—ripples—are supported by both gravity and surface tension. The nominal speed of sound (elastic waves) in pure water is 1400 m/s. Ripple speed is substantially less and so is independent of the bulk modulus of elasticity of water. In the analysis for ripple motions the water may be regarded as an incompressible fluid. The differential equation for ripple propagation shows that the square of the wave speed c is given by

$$c^2 = g\lambda/2\pi + 2\pi T/\rho\lambda \qquad (20)$$

for sinusoidal waves on the surface of a liquid of depth much greater than the wavelength λ. Here the acceleration of gravity is $g=9.8$ m/s^2; the surface tension T has units of N/m; and the density of pure water is $\rho = 10^3$ kg/m^3.

The speed c varies with the wavelength λ. For a water surface there is a minimum in c near $\lambda = 1.7$ cm, and c is about 23 cm/s. For longer wavelengths c is determined mainly by the acceleration of gravity g, and the waves are deep-water gravity waves. For shorter waves c is determined mainly by the surface tension T, and the waves are surface-tension ripples.

Gravity waves on a liquid surface offer an acoustical method for measurement of g. Apparently no data on such measurement have ever been published. Lord Rayleigh, in 1890, measured T by means of ripples, in spite of the unavailability then of convenient modern-day transducers. The ripple method assures that T is determined free of any assumption respecting angles of contact between water and a solid, as in a capillary tube measurement of T. The acoustical ripple method admits "of application to surfaces purified to the utmost from grease." Some of his results are given in Table I.

Surface acoustic waves in solids were described as early as 1885 by Rayleigh to explain the propagation of seismic disturbances from earthquakes. More recently, the applications to piezoelectric materials of surface acoustic waves (SAW) has resulted in new devices, such as an "interdigital transducer" (Fig. 10), which may be operated as a filter (Apfel, 1974). Apfel, in turn, refers to three earlier review articles on SAW applications in microwave ultrasonics (Kino and Shaw, 1972; deKlerk, 1972; White, 1970). In addition, the reader is referred to Campbell (1990) and the article RESONATORS AND FILTERS, ACOUSTIC.

2.5 Interior of the Earth

Acoustical studies of the interior of the Earth may be directed toward the geological structure of the Earth based on detecting naturally occurring phenomena, or toward the detection, mapping, and exploitation of oil fields based on seismic signals returned following transmitted sound pulses (White, 1965).

The global structure of the solid Earth is spherically layered (see EARTH, INTERIOR STRUCTURE OF). Earthquakes serve as sources of elastic waves in the earth. Both longitudinal and shear elastic waves are radiated from earthquake sources. In addition, Rayleigh (elastic) waves on the Earth's surface, as well as tsunami (gravity) waves on the surfaces of the oceans, are generated by earthquakes. The

Table 1. Surface tension T of water (from Rayleigh, 1894).

Surface condition	T (N/m)
Clean	0.074
Greasy	0.053
Saturated with olive oil	0.041
Saturated with oleate of soda	0.025

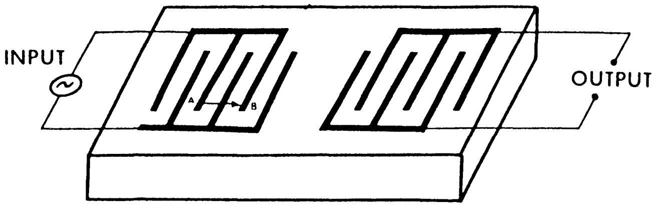

FIG. 10. A schematic of a simple surface acoustic wave (SAW) "interdigital" transducer (Apfel, 1974).

basic theory and mathematical physics for elastic waves on and in the Earth are essentially acoustical. The waves are detected and measured by seismometers that are perforce placed on the surface at various locations around the world. The seismometer transducers measure vibratory displacements at infrasonic (very low) frequencies, usually less than 1.0 Hz. The times of incidence and durations of the several wave types are also measured. These data are basic for a determination of the internal structure of the Earth.

Additional basic data are the observed frequencies and amplitudes of oscillation of the normal modes of oscillation of the whole Earth following a great earthquake. These may be likened to the reverberation of sound in an empty room after a sharp transient sound. The greatest observed period of oscillation may be of the order of an hour or so.

Acoustical methods of analysis are then applied to the basic data on elastic waves transmitted through the interior of the Earth, combined with laboratory data on the speed of longitudinal and shear waves in geological rock. By these means a spherically shaped surface of discontinuity has been found at the boundary between the Earth's mantle and the outer core. This is at a depth below the surface of 2900 ± 10 km. This accuracy comes about from waves originating at earthquakes deep in the mantle.

Both longitudinal and shear waves of sound travel through the mantle. However, only longitudinal waves pass through the outer core beneath the mantle–core boundary. This can mean either that the outer core of the Earth is liquid or that shear waves in it are rapidly attenuated. The acoustical analysis leads also to an estimated distribution of temperature and density as a function of depth below the surface and therefore to an estimate of the hydrostatic pressure at various depths.

Explorations for oil and minerals beneath the surface of the Earth are conducted within the mantle. At the operating depths from the surface (zero depth) up to several kilometers, transmitted acoustic signals may be longitudinal or shear waves. The shear waves may be polarized, analogous to optical waves, and the polarization directions are commonly divided into horizontal shear (SH) and vertical shear (VH) components. When an oil field is discovered and producing wells are operating, the monitoring of a well is called oil-well logging. Bore holes are drilled in the vicinity of the well and transmitting and receiving transducers, an example of which is shown in Fig. 11, may be lowered into the bore holes in a pattern designed for efficient data gathering and signal processing (Leslie and Randall, 1992).

3. HEARING AND PHYSIOLOGY

In the practice of medicine, physicians are called upon to treat difficulties and defects in hearing. A common defect is loss of the ability to hear well such everyday sounds (Stevens and Davis, 1983) as the speech of other persons, music, traffic, etc. Until about fifty years ago, ear-nose-throat (ENT) physicians had used tuning forks, producing sounds at various sinusoidal frequencies, to assess the ability of a patient to hear.

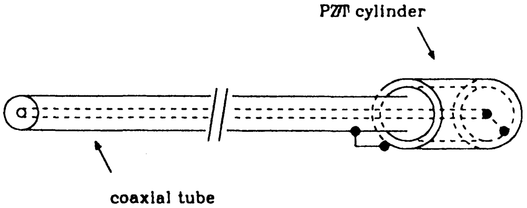

FIG. 11. Seismic micromonopole transducer construction details. PZT is a trade name for a lead zirconate–titanate piezoelectric ceramic material (Leslie and Randall, 1992.)

During the 1930s audiometers were invented, developed, and manufactured to produce sound at various audible frequencies. The audiometer is now used in physicians' offices to measure quantitatively the hearing level (minimum audible sound pressure) of a person under diagnosis. They are used as well in routine general physical examinations. The instrument is basically electroacoustical; voltages at various sinusoidal frequencies in the range 64–8000 Hz and of adjustable amplitude are generated electronically and applied to an earphone. The latter is held against the patient's ear, with guidance by the audiologist. The voltage is then adjusted until the earphone sound is just barely audible to the patient (Fig. 12).

The acoustical (physical) action is production of sound pressure in the ear canal, which vibrates the tympanic membrane. The latter, in turn, vibrates the bones of the middle ear, which finally causes waves on the basilar membrane of the inner ear (cochlea) sensed by the nerves there (von Békésy, 1935; Stevens and Davis, 1983; see also ACOUSTICS, PHYSIOLOGICAL).

The audiometer earphone is essentially the same (electroacoustically) as the receiver of a telephone. However, it is used as a measurement transducer, and so the engineering requirements on its temporal stability of response, freedom from harmonic distortion, etc., are more stringent than they are for telephones. The important point is that the use of the audiometer yields a relatively accurate physical measure of a person's ability to hear.

Vibrations of the skull can be transmitted to the cochlea and are then also perceived as sound. Multielectrode cochlear implants are among the more recent types of hearing prostheses, and new methods are being developed for testing them (McDermott *et al.*, 1992). The transmission is thought to take place directly to the cochlea encapsulated in the temporal bone of the skull and is referred to as "bone conduction" of sound. Audiometers have a "bone vibrator," which like the earphone is a measurement transducer. The vibrator can be applied to the frontal bone of the skull. Sinusoidal vibrations are produced at the same frequencies as those of the sound pressures produced by the earphone, and the threshold of hearing by bone-conduction vibrations is measured. From a comparison of air-conduction and bone-conduction measurements, the otologist can make an improved diagnosis of impaired hearing.

Other physiological acoustics tests can now be employed by internal-medicine physicians in such procedures as the diagnosis of tumors (Romijn *et al.*, 1991), *in utero* examination of a fetus, and measurement of blood flow (Embree and O'Brien, 1990). Sonar methods are used, in which ultrasound is introduced at the skin surface by means of a piezoelectric transducer. Variations in density and elasticity of the internal tissues scatter the ultrasound, which is reflected back to the same transducer. Extensive computer-aided tomography (CAT) signal processing techniques (Muller *et al.*, 1979; Devaney, 1985) are applied to the received sound, which is finally displayed

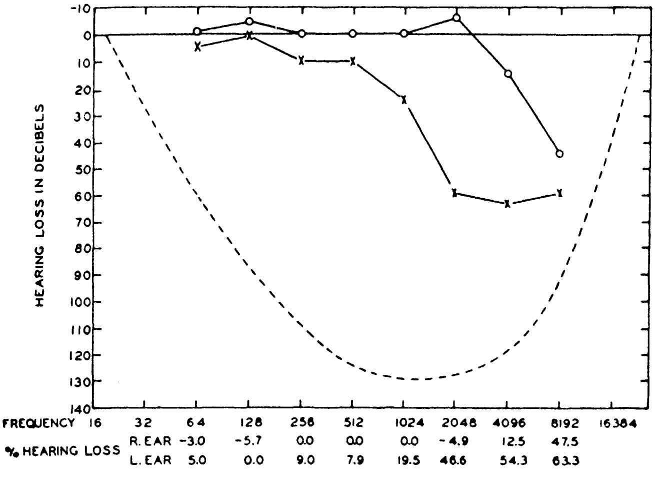

FIG. 12. An audiogram for a person suffering from high-tone loss of hearing. The circles are for the left ear, the crosses for the right. The zero ordinate represents normal hearing; the dotted curve represents 100% loss of hearing (Stevens and Davis, 1983).

continuously as a two-dimensional visual image, or sonogram, as in a motion picture. Synthetic receive aperture (SRA) imaging is an important technique applicable to the study of tissues in that it offers a significant reduction in the number of transducer elements required (Nock and Trahey, 1992). The advantages of ultrasonic tomography are that it is nonintrusive medically and that it causes no structural atomic–molecular changes of body tissue. Further discussion of sonar methods for other applications is set forth in Sec. 5.

4. TELEPHONY

The earliest development of electroacoustical transducers was in the telephone industry, in the 1870s (Bell, 1876; Martin, 1951; Hunt, 1982). The sounds of speech were transduced into electrical signals by "transmitters" (microphones). Then after transmission over wire lines the electrical signals were transduced into sound waves in air, at the listener's ear, by "receivers" (quadratic earphones). Modern telephone transmitters are usually electret capacitive microphones (Sessler and West, 1962), and receivers are modified loudspeakers. Until recently, the receiver has been basically a small piston, driven electromagnetically, which produces sound in the enclosed volume of air formed by the receiver and the ear to which it is applied. Newer receivers are more likely to be electret capacitive loudspeakers, primarily because they can be produced at lower cost. The electromagnetic receivers had the advantage of being compatible with electromagnetic pickup devices in hearing aids before conversion to sound in the hearing aid. With the newer electret receivers it is necessary to employ a sound-to-electromagnetic converter between the telephone receiver and a hearing aid, and these devices are manufactured to meet this newer requirement.

The action of a telephone system has, however, very little to do with the physics of sound in the atmosphere. The principal dependence of the telephone industry has been on applications of psychoacoustics to the intelligibility of speech sounds in the presence of possible electric-circuit noise and of excessive distortion in the wave form of the voice signal.

Also, the industry applies extensively the principles of electroacoustics. The audio-frequency bandwidth of the entire system, from transmitter through the electrical circuits to the sound radiated into the ear of the listener, is limited to the range for acceptable speech intelligibility. The voice currents may travel over wire lines thousands of miles long, over land and under water, having numerous "repeaters" (amplifiers of the voice currents) serving also as electrical filters. The voice currents may be "multiplexed" by using each current to modulate its own ultrasonic carrier frequency, well separated from all the other carriers. Then a coaxial conductor can have hundreds of conversations on the line simultaneously, with no noticeable interference between individual voice currents.

Some or all of the transmission may be by means of modulated radio waves. This may involve reflection by or retransmission by satellites as well as the direct transmission between radiating and receiving antennas. The latter have become especially popular in the development of cellular telephones and "cordless" telephones, making it convenient to move about in a house, to telephone from automobiles and other moving vehicles, and for emergency communication such as is often necessary following earthquakes, hurricanes, etc.

The development of long optical fibers with low transmission loss is a newer alternative for the transmission of multiplexed telephone signals. They have the advantage of much greater bandwidth capability for a given diameter, permitting greater signal-handling capacity. Furthermore, optical fibers are much less sensitive to electromagnetic interference and produce much lower output levels of electromagnetic radiation. The development of optoacoustic and acousto-optic transducers for direct conversion of acoustic signals from and to optical signals has been a natural consequence (Tsai, 1992). See OPTICAL COMMUNICATION.

Among the by-products of telephone voice signal transmission, an important one is digital data transmission. This may be done either by assignment of separate telephone channels or by piggybacking on existing channels in the frequency band below 300 Hz. Telephone lines are readily adapted to facsimile (FAX) transmission of documents and to electronic mail (e-mail), for example.

Telephone transducers are not used as measuring instruments. Nonetheless, they are measured acoustically during manufacture to ensure that they meet specification requirements. Such test measurements are made with condenser microphones calibrated to yield, from the voltage output, the amount of sound pressure that produces the voltage. Receivers are placed on an artificial ear that has a condenser microphone to measure the amount of sound produced by the receiver when test audio-frequency voltages are applied to it. Similarly, a transmitter has test audio-frequency sound pressures, measured by a condenser microphone, applied to the transmitter's membrane diaphragm. Its voltage output is then measured with electronic circuitry. In brief, telephone transducers are not required to have precisely controlled responses. Their phases and amplitudes, for signals sinusoidal in time, may vary from transducer to transducer over small ranges.

5. SONAR METHODS

Sonar is a word coined by F. V. Hunt as the acoustical analog of radar. Where radar was devised as an acronym for "radio detection and ranging," an attempt to find suitable words to make sonar also an acronym resulted in "sound navigation and ranging." The classical uses of sonar are detection, classification, tracking (and localization), navigation, communication, and countermeasures (see Horton, 1957; Urick, 1983; Bartram *et al.*, 1982).

Table 2. Military operational uses of sonar (underwater transducer applications in naval sonar).

Missions	Platforms
Area (volume) search	Submarines
Detect, classify, and track	Surface ships
Hold, harass, and/or attack	Fixed-wing aircraft with sonobuoys
Ballistic missile submarine defense	Rotary-wing aircraft
Detect and evade	Torpedoes
Counterattack	Mines
Convoy defense	Fixed-position systems
Screen	
Intelligence	
Strategic–Turnstile	
Tactical–Technical	

Table 3. Some equipment for detection and location of underwater objects (sonar systems and transducer arrays).

- Submarine sonars
  - Spherical array
  - Wide aperture array
  - Towed array
- Surface ship sonars
  - Cylindrical array
  - Towed array
  - Intercept receiver systems
- Fixed- and rotary-wing airborne sonars
  - Sonobuoy systems
  - Seaplane-dipped systems
  - Helicopter-dipped and-towed sonars
- Other sonar systems
  - Minehunting systems
  - On-board trainers

Sonar methods are those in which steady-state, transient, and pulsed sound waves, including their reflections and scattered secondary waves, are applied to determine internal physical features of materials and gases. Some typical applications are

1. the determination of the locations of internal flaws and the gross structural features of materials used in engineering, particularly during industrial processing;
2. in the natural waters of the seas and oceans, the depth of the waters beneath a ship under way and the locations of icebergs (historically, the latter was the earlier application, following the sinking of the Titanic in 1912);
3. detection and location of underwater objects and vehicles such as submarines (Tables 2 and 3), and the location for commercial fishing and environmental studies of schools of fish and crustacea;
4. for geophysical prospecting in the earth, the determination of subsurface structures, such as salt domes having petroleum and natural gas, and for oil-well logging;
5. in the laboratory, convenient measurement of sound speed and attenuation in liquids by means of pulsed ultrasound, and also for the measurement of elastic and transducer properties of solids by resonance techniques, including vector and tensor quantities;
6. in natural waters, the direct measurement

of local sound speed *in situ* by means of velocimeters, which record continuously in a sing-around electronic circuit the travel time of an ultrasonic pulse between two piezoelectric transducers;

7. in the human body, examination of internal structures by ultrasound introduced at the skin surface, with the scattered sound being received and displayed continuously in a visual image.

The applications employ a fairly wide variety of transducers, depending on factors such as spatial-resolution requirements, size limitations, attenuation in the transmitting medium as a function of frequency, ease of access to the object of interest, and the normal engineering cost problems. For many of the applications ultrasonic frequencies are used. The transducers are usually either polarized piezoelectric ceramics, such as one of the lead zirconate–titanate compositions, or piezoelectric crystals, such as quartz. One transducer serves as a source for an ultrasonic pulse of sound and another to receive the reflected waves from the object insonified; an alternative is for the same transducer to perform both functions with electronic switching to convert between the source and the receiver configurations. The electronic methods for the generation and detection of pulsed ultrasound are similar to those used with electromagnetic waves in radar systems. For most modern long-range sonar system applications, sonic frequencies are used to take advantage of the lower sound attenuation in seawater at lower frequencies. The transducers tend to be larger and the piezoelectric materials get some competition from polarized piezomagnetic, or magnetostrictive, materials as the transducing mechanism of choice. Flextensional transducers (Royster, 1970; Rolt, 1990; Decarpigny *et al.* 1991; Boucher, 1992; Ring, 1992) that use longitudinal active elements to drive shells in bending modes are often employed to limit transducer size (Fig. 13); also, arrays of large numbers of elements of longitudinal transducers are commonly used (Fig. 14). In geophysical prospecting for minerals and oil in the near-surface layers of the earth, the frequencies used tend toward the very low sonic and infrasonic. The transducers used tend to be impulsive, such as

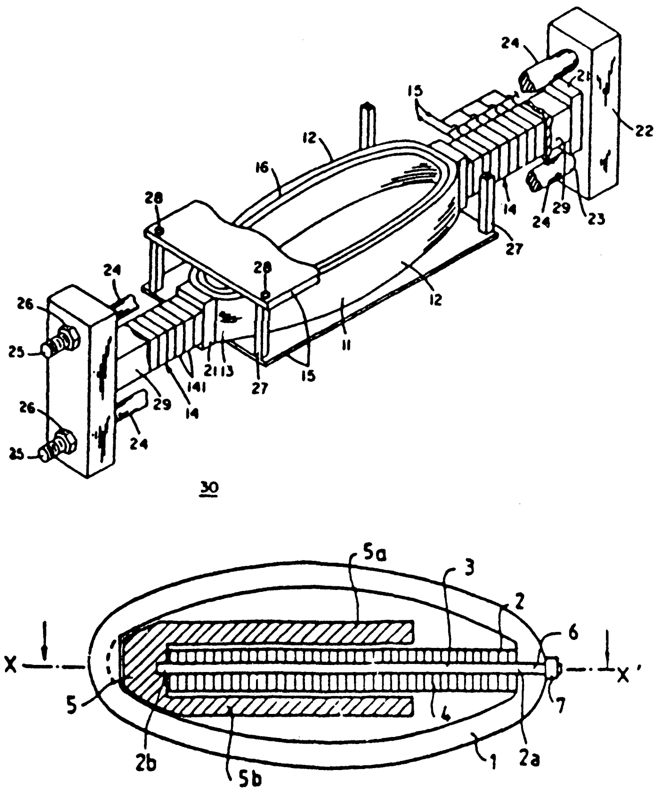

FIG. 13. Two recent flextensional transducer patents (Royster, 1970; Rolt, 1990; Decarpigny *et al.*, 1991; Boucher, 1992; Ring, 1992).

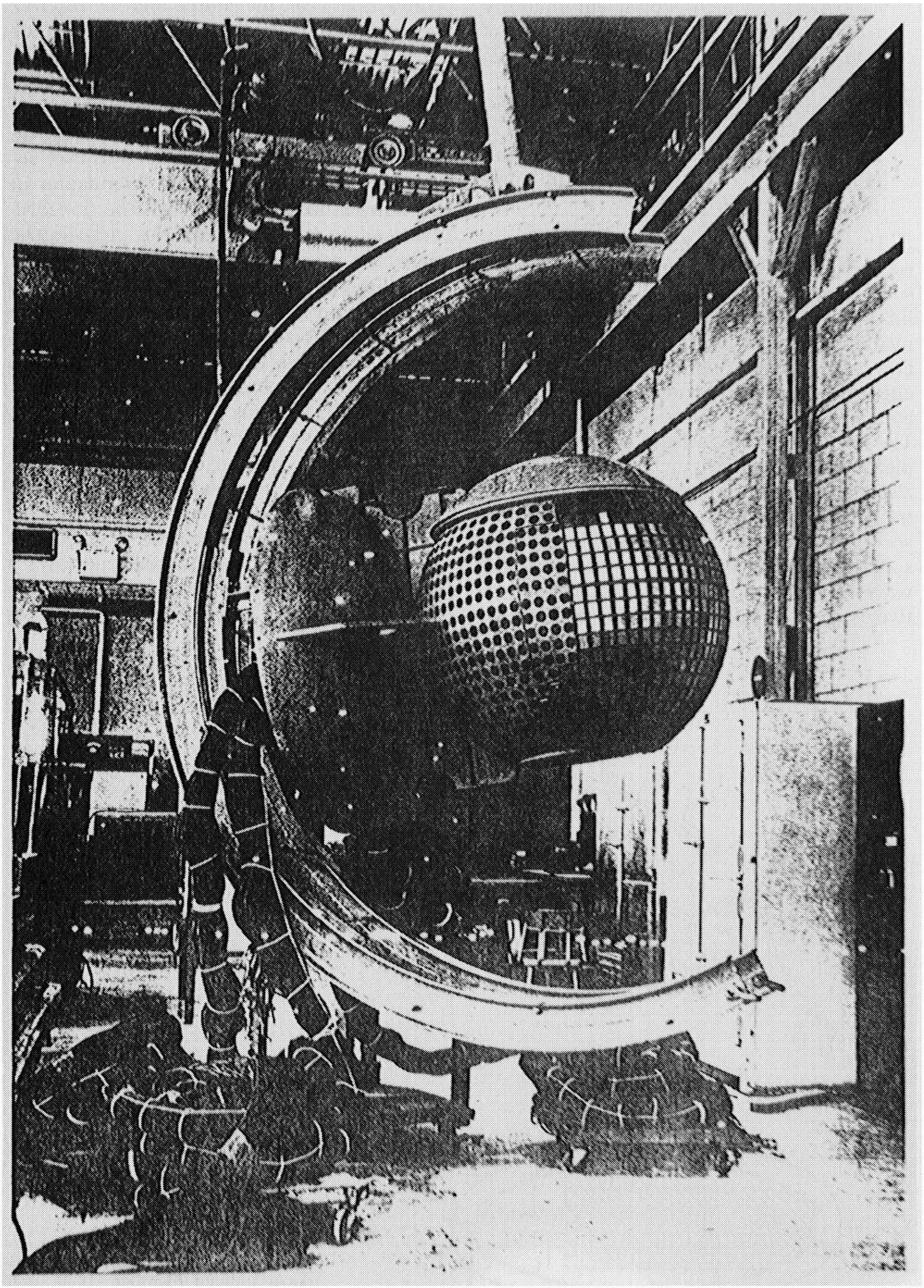

FIG. 14. Partially filled large spherical array mounted in a test fixture. A Raytheon Company development; see Hueter (1972).

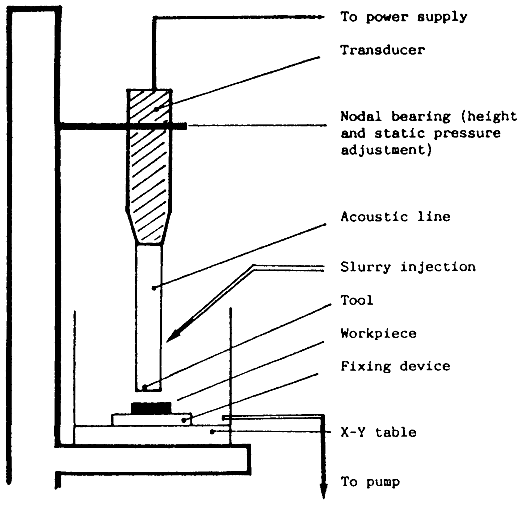

FIG. 15. Ultrasonic machining apparatus: schematic description of the main parts (Tierce and Decarpigny, 1988).

produced by hydraulic mechanisms or by explosives, to produce elastic waves. Detection of the waves generated and the waves scattered from underground structures often employ geophone transducers at the surface of the earth in the vicinity. Where the location of an oil well is known, the process of oil-well logging employs higher frequencies than those used in prospecting, permitting the use of more conventional or smaller transducers, especially where the use of explosives is not tolerable. Natural phenomena, such as earthquakes, are detected by seismographs, often at frequencies below 1 Hz; that is, the number of seconds, or even hours, for each cycle is measured.

6. SONICS

Sonics is the application of acoustical technology to industrial measurement, testing, analysis, and processing of materials (see Beranek, 1988; Hueter and Bolt, 1955). Many standards for sonics equipment have been developed under the auspices of the International Standards Organization (ISO), the International Electrotechnical Commission (IEC), a variety of national agencies, such as the American National Standards Institute (ANSI), and scientific and engineering societies, such as the Acoustical Society of America (ASA) and the Institute of Electrical and Electronics Engineers (IEEE); some of the specific standards are cited in Beranek (1988).

Aerosols of fine particles in the atmosphere are agglomerated by intense sound waves. In liquids used for cleaning small manufactured parts, intense sound causes cavitation, which accelerates the cleaning process. Sonic cavitation also accelerates chemical reactions, homogenization of liquid mixtures, and dispersion of solids into liquids. Internal flaws in the materials of engineering are detected by irradiation with ultrasound (sound waves at ultrasonic frequencies). The continuous monitoring of the dimensions of manufactured objects is accomplished. An example is thickness measurement and detection of defect laminations in strip steel as the strip leaves the rolling mill. Controlled intense sound is used to drill very hard materials; one example is the use of a conventional piezoelectric transducer with an inverted horn (Fig. 15) to produce a large amplitude over a small area (Tierce and Decarpigny, 1988). In other applications, large shaker transducers are at-

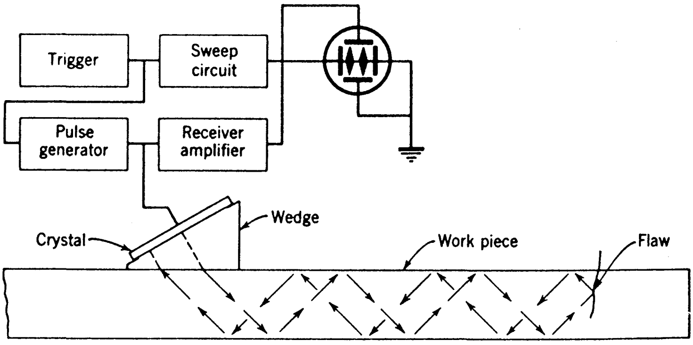

FIG. 16. Beam-refraction flaw-detection system for inspection of pipes (Hueter and Bolt, 1955).

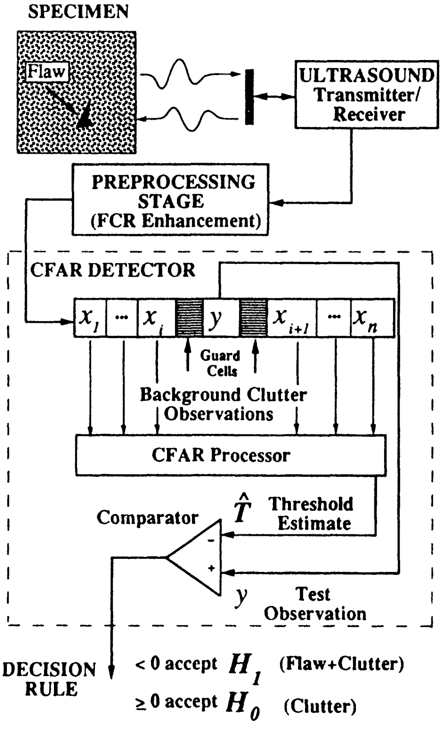
SPECIMEN
Flaw
ULTRASOUND Transmitter/ Receiver
PREPROCESSING STAGE (FCR Enhancement)
CFAR DETECTOR
x_1 ··· x_i y x_{i+1} ··· x_n
Guard Cells
Background Clutter Observations
CFAR Processor
Comparator
$\hat{T}$ Threshold Estimate
−
+
y Test Observation
DECISION RULE
< 0 accept H_1 (Flaw+Clutter)
≥ 0 accept H_0 (Clutter)

FIG. 17. Block diagram of ultrasonic constant false-alarm rate (CFAR) flaw detection system (Saniie and Nagle, 1992).

tached to conventional drilling equipment to speed up the drilling process by orders of magnitude. Only two of the many applications of sonics noted above are discussed further in this section.

6.1 Flaw Detection

Acoustical methods for detection of flaws in manufactured articles, particularly those made of metal that are required to carry large loads and stresses, are extensively used in the manufacturing industry. The methods are referred to as nondestructive evaluation, often abbreviated to as NDE, or NDT, where the T stands for testing (Firestone, 1945). A commonly used technique is the generation of ultrasonic pulses and measurement of the sound scattered by flaws. The basic concept is virtually the same as the sonar concept for some of the underwater applications of acoustics; see Sec. 5. Typical articles to be tested for internal flaws are metallic load-bearing plates, axles in rotating machinery, laminated strips and plates, the walls of tanks and pipelines subjected to substantial internal hydrostatic pressures, etc.

We describe here a hypothetical system for flaw detection (Fig. 16). The piezoelectric crystal transducer radiates longitudinal (or shear or surface, depending on the angle of

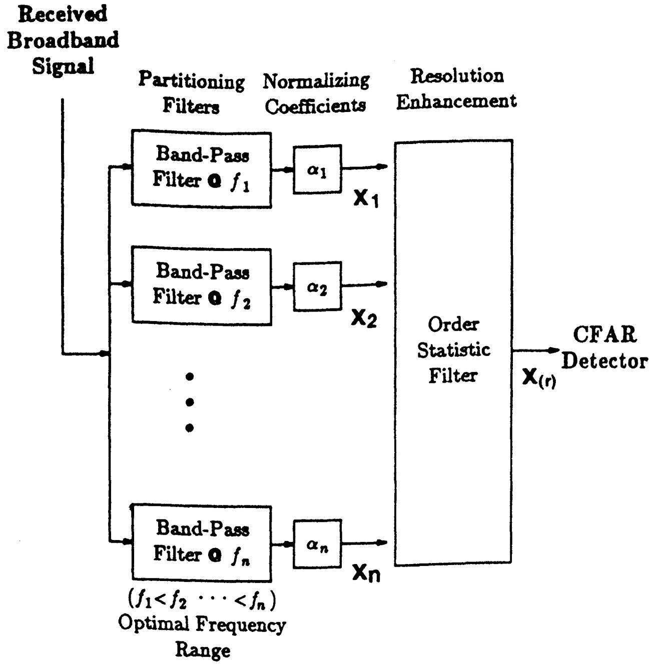
Received Broadband Signal
Partitioning Filters
Normalizing Coefficients
Resolution Enhancement
Band-Pass Filter @ f_1
α_1
X_1
Band-Pass Filter @ f_2
α_2
X_2
Band-Pass Filter @ f_n
α_n
X_n
Order Statistic Filter
$X_{(r)}$
CFAR Detector
$(f_1 < f_2 \cdots < f_n)$ Optimal Frequency Range

FIG. 18. Split-spectrum signal processing schematic (Saniie and Nagle, 1992).

the wedge) waves of ultrasound into the relatively long cylindrical steel pipe (Hueter and Bolt, 1955). Sound is scattered by the flaw, a thin disklike cavity in the metal, and there is sound reflected from the end of the pipe. The same crystal transducer serves as a microphone for detection and measurement of scattered and reflected ultrasound. The results of the electronic processing of the microphone signal are displayed on an oscilloscope-video screen.

The electroacoustic source for this hypothetical system produces (a) a sinusoidal wave at 1.0 MHz, which is the resonance frequency of the piezoelectric crystal source; (b) a pulsed output 5 periods ($=5$ μs) long; (c) a time delay (between successive pulses) long enough to allow the crystal source to be used in the interim as a microphone and to allow the ultrasound in the metal to die away to zero; (d) a display, sequential in time, of the signals received by the microphone.

When the process of flaw detection is difficult because of the presence of interfering and random echoes, called "clutter," one method of overcoming the problem is the application of constant false-alarm rate signal processing (Saniie and Nagle, 1992). A system block diagram is shown in Fig. 17 and the signal-processing schematic in Fig. 18.

6.2 Thickness Metrology

The measurement of the thickness of metal parts is often required during a manufacturing process. The needs are (a) for rapid and accurate measurement of nominally identical parts coming off an assembly line; (b) for measurement with access limited to only one side of the part; and (c) for measurement of curved-surface as well as flat-surfaced pieces.

A basic acoustical technique is detection of the resonance frequencies and/or transit times of ultrasound radiated into the metal part. This is accomplished by means of a piezoelectric transducer, usually quartz or lead zirconate–titanate, coupled to the metal by a thin film of oil to assure acoustical contact. Typical thicknesses tend to be from 1.0 to 40 mm. The transducer transmits longitudinal waves of frequency-modulated ultrasound at an accessible face of the metal part. Reflection of sound at the opposite face creates a standing elastic (acoustic) wave between the two faces of the piece. At a resonance frequency there is a whole number m of half wavelengths in the thickness of the piece, and the effect in the electronic circuit connected to the transducer is maximum. The ultrasonic signal is frequency modulated, so that m varies repetitively from m to $m+a$. The velocity of sound c in the metal must be known. The repetitive standing-wave pattern in the electronics is displayed on an oscilloscope-video screen showing the resonance frequencies produced by the oscillator. An example is a stainless-steel plate $h=3$ mm thick, sound velocity $c=5790$ m/s; the difference between successive resonance frequencies is $\Delta f = c/2h = 965$ kHz. The electronic display is designed and calibrated so that the indicated (or calculated) thickness is uncertain by less than 1.0%.

GLOSSARY

Acoustic-Gravity Wave: A wave in the atmosphere, whose phase velocity is dependent on both the bulk modulus of elasticity of the atmosphere and the acceleration of gravity.

Acoustic Impedance: Ratio of sound pressure to volume velocity.

Electroacoustics: The science whose main objective is the study of the conversion of sound energy into electrical waveforms, and conversely the conversion of electrical waveforms into vibration and sound waves.

Geophone: A transducer, from motions of the surface of the earth to an electrical waveform.

Hydrophone: A transducer, from underwater sound to an electrical waveform.

Non-Destructive Testing (NDT) and Evaluation (NDE): The use of sound waves, particularly ultrasound, to examine the interiors of optically opaque solids. Voids and inclusions, which can impair the strength of the material, scatter ultrasound which can be detected by transducers on the surface.

Particle Velocity: The velocity of small displacements in the medium that mark the passage of sound waves.

Projector: A transducer, from an electrical waveform to underwater sound.

Rayleigh Wave: A SAW on the surface of the earth.

Reciprocity Method: A protocol for calibration of microphones and hydropones, for which the reciprocal theorem is valid in the transducer, and from which an absolute measure of the transducer response to sound waves can be obtained.

Sing-Around Circuit: A closed electroacoustic loop, in which one link is a short fixed-length path underwater for an ultrasonic pulse which travels recurrently around the loop. The greater the speed of sound in the water, the higher is the frequency of recurrence of the pulse around the loop.

SOFAR: An acronym for "sound fixing and ranging," by means of the "sound channel" of minimum sound speed, at a depth of about 1000 meters in many of the natural seas and oceans.

SONAR: An acronym for "sound navigation and ranging" by means of underwater sound, in the seas and oceans.

Sound Pressure: The small fluctuations in pressure that mark the passage of sound waves through gases and liquids

Specific Acoustic Impedance: Ratio of sound pressure to particle velocity.

Surface Acoustic Wave (SAW): An elastic wave on the surface of a solid, whose particle motions are in ellipses normal to the solid surface, and whose phase velocity is slightly less than the shear-wave velocity.

Volume Velocity: The arithmetical product of a plane area in the medium, and the component of particle velocity normal to the plane.

List of Works Cited

ANSI/ASME (1984), Standard Y10.11, *Letter Symbols and Abbreviations for Quantities Used in Acoustics,* New York: American Society of Mechanical Engineers.

ANSI/IEEE (1987), Standard 176, *Piezoelectricity,* New York: Institute of Electrical and Electronics Engineers.

Apfel, R. E. (1974), *Phys. Teacher* **12**, 109–111.

Bartram, J. F., Ehrlich, S. L., Fredenburg, D. A., Heimann, J. H., Kuzneski, J. A., Skitzki, P. (1982), in: D. G. Fink, D. Christiansen (Eds.), *Electronics Engineers' Handbook,* 2nd ed. New York: McGraw-Hill.

Bell, A. G. (1876), *Am. Acad. Arts Sci. Proc.* **12**, 1–10. Reprinted (1981) in: R. B. Lindsay (Ed.), *Acoustic Transducers,* Stroudsburg, PA: Hutchinson and Ross, pp. 12–21. Also U.S. Patent No. 174 465, issued 1876 March 07.

Beranek, L. L. (1986), *Acoustics,* New York: Acoustical Society of America.

Beranek, L. L. (1988), *Acoustical Measurements,* New York: Acoustical Society of America.

Berlincourt, D. (1981), *J. Acoust. Soc. Am.* **70**, 1586–1595.

Berlincourt, D. (1992), *J. Acoust. Soc. Am.* **91**, 3034–3040.

Birss, R. R. (1964), *Symmetry and Magnetism,* Amsterdam: North-Holland.

Bobber, R. J. (1970), *Underwater Electroacoustic Measurements,* Washington, DC: U.S. GPO.

Boucher, D. (1992), in: *Proceedings of the Third International Workshop on Transducers for Sonics and Ultrasonics,* Orlando, FL: U.S. Navy Research Laboratory, Underwater Sound Reference Div.

Cady, W. G. (1922), *Proc. Inst. Radio Eng.* **10**, 83–95, 106–109. Reprinted (1974) in: R. B. Lindsay (Ed.), *Physical Acoustics,* Stroudsburg, PA: Dowden, Hutchinson and Ross, pp. 223–238.

Cady, W. G. (1964), *Piezoelectricity* (reprint), New York: Dover.

Campbell, C. (1990), *Surface Acoustic Wave Devices and Their Signal Processing Applications,* New York: Academic.

Clark, A. E. (1988), in: B. Hamonic, J. N. Decarpigny (Eds.), *Power Sonic and Ultrasonic Transducers Design* (Proceedings of the International Workshop in Lille, France, 1987), Berlin: Springer, pp. 41–99.

Cook, R. K. (1940), *J. Res. Natl. Bur. Std.* **25**, 489–505.

Cook, R. K. (1941), *J. Acoust. Soc. Am.* **12**, 415–420.

Curie, P. J., Curie, P. (1880), *C. R. Hebd. Séances* **91**, 294–296. Translated from French and reprinted (1972) in: R. B. Lindsay (Ed.), *Acoustics,* Stroudsburg, PA: Dowden, Hutchinson and Ross, pp. 373–374.

Curie, P. J., Curie, P. (1881), *C. R. Hebd. Séances* **92**, 186.

Decarpigny, J. N., Hamonic, B., Wilson, O. B. (1991), *IEEE J. Oceanic Eng.* **16**, 107–122.

deKlerk, J. (1972), *Phys. Today* **25** (11) 32–39.

Devaney, A. J. (1985), *J. Acoust. Soc. Am.* **78**, 120–130.

Ehrlich, S. L. (1952), *Proc. Inst. Radio Eng.* **40**, 992.

Embree, P. M., O'Brien, W. D. (1990), *IEEE Trans. Ultrason. Ferroelec. Freq. Contr.* **37**, 176–189.

Firestone, F. A. (1945), *J. Acoust. Soc. Am.* **17**, 287–299.

Greenspan, M., Tschiegg, C. E. (1963), in: *Underwater Acoustics,* New York: Plenum Press.

Hilliard, J. K. (1950), *Soc. Motion Pict. Telev. Eng. J.* **54**, 303–314. Reprinted (1981) in: R. B. Lindsay (Ed.), *Acoustic Transducers,* Stroudsburg, PA: Dowden, Hutchinson and Ross, pp. 142–149.

Horton, J. W. (1957), *Fundamentals of Sonar,* Annapolis, MD: U.S. Naval Institute.

Hueter, T. F. (1972), *J. Acoust. Soc. Am.* **51**, 1025–1040.

Hueter, T. F., Bolt, R. H. (1955), *Sonics*, New York: Wiley.

Hunt, F. V. (1982), *Electroacoustics*, New York: Acoustical Society of America.

IEEE (1990), Standard 319, *Magnetostrictive Materials: Piezomagnetic Nomenclature*, New York: Institute of Electrical and Electronics Engineers.

Kawai, H. (1969), *Jpn. J. Appl. Phys.* **8**, 975.

Kino, G. S., Shaw, J. (1972), *Sci. Am.* **227** (4), 50–68.

Leslie, H. D., Randall, C. J. (1992), *J. Acoust. Soc. Am.* **91**, 12–27.

Martin, W. H. (1951), *Bell Syst. Tech. J.* **10**, 46–62.

Mason, W. P. (1950), *Piezoelectric Crystals and Their Application to Ultrasonics*, New York: Van Nostrand.

Mason, W. P. (1981), *J. Acoust. Soc. Am.* **70**, 1561–1566.

McDermott, H. J., McKay, C. M., Vandali, A. E. (1992), *J. Acoust. Soc. Am.* **91**, 3367–3371.

Muller, R. K., Kaveh, M., Wade, G. (1979), *Proc. Inst. Electr. Electronic Eng.* **67**, 567–587.

National Defense Research Committee (1946a), Summary Technical Report of Div. 6, Vol. 10, *Basic Methods for the Calibration of Sonar Equipment*, Washington, DC: U.S. GPO.

National Defense Research Committee (1946b), Summary Technical Report of Div. 6, Vol. 11, *A Manual of Calibration Measurements of Sonar Equipment*, Washington, DC: U.S. GPO.

National Defense Research Committee (1946c), Summary Technical Report of Div. 6, Vol. 12, *Design and Construction of Crystal Transducers*, Washington, DC: U.S. GPO.

National Defense Research Committee (1946d), Summary Technical Report of Div. 6, Vol. 13, *Design and Construction of Magnetostriction Transducers*, Washington, DC: U. S. GPO.

Nicolson, A. M. (1917):, U. S. patent (number not available), based on work with Rochelle salt at Bell Telephone Laboratories.

Nock, L. F., Trahey, G. E. (1992), *IEEE Trans. Ultrason. Ferroelec. Freq. Contr.* **39**, 489–495.

Pierce, A. D. (1989), *Acoustics*, New York: Acoustical Society of America.

Lord Rayleigh (1894), *Theory of Sound*, New York: Macmillan; reprint (1945), New York: Dover.

Rice, C. W., Kellogg, E. W. (1925), *Trans. Am. Inst. Electr. Eng.* **44**, 461–475. Reprinted (1981) in: R. B. Lindsay (Ed.), *Acoustic Transducers*, Stroudsburg, PA: Dowden, Hutchinson and Ross, pp. 188–202.

Ricketts, D. (1992), in: *Proceedings of the Third International Workshop on Transducers for Sonics and Ultrasonics*, Orlando, FL: U.S. Navy Research Laboratory, Underwater Sound Reference Div.

Ring, M. D. (1992), in: *Proceedings of the Third International Workshop on Transducers for Sonics and Ultrasonics*, Orlando, FL: U.S. Navy Research Laboratory, Underwater Sound Reference Div.

Rolt, K. D. (1990), *J. Acoust. Soc. Am.* **87**, 1340–1349.

Romijn, X., *et al.* (1991), *Ultrasonic Imaging* **13**, 27–55.

Royster, L. H. (1970), *Appl. Acoust.* **3**, 117–126.

Saniie, J., Nagle, D. T. (1992), *IEEE Trans. Ultrason. Ferroelec. Freq. Contr.* **39**, 618–630.

Sessler, G. M. (1981), *J. Acoust. Soc. Am.* **70**, 1596–1608.

Sessler, G. M., West, J. E. (1962), *J. Acoust. Soc. Am.* **34**, 1787–1788.

Stevens, S. S., Davis, H. (1983), *Hearing*, New York: Acoustical Society of America.

Tierce, P., Decarpigny, J. N. (1988), in: B. Hamonic and J. N. Decarpigny (Eds.), *Power Sonic and Ultrasonic Transducers Design* (Proceedings of the International Workshop in Lille, France, 1987), Berlin: Springer, pp. 185–207.

Tsai, C. S. (1992), *IEEE Trans. Ultrason. Ferroelec. Freq. Contr.* **39**, 529–554.

Urick, R. J. (1983), *Principles of Underwater Sound for Engineers*, 3rd ed., New York: McGraw-Hill.

Van Dyke, K. S. (1928), *Proc. Inst. Radio Eng.* **16**, 742–764.

Voigt, W. (1910), *Lehrbuch der Kristallphysik*, Leipzig: Teubner.

von Békésy, G. (1935), *Elek. Nachr. Techn.* **12**, 71–83.

Wente, E. C. (1917), *Phys. Rev.* **10**, 39–51.

White, J. E. (1965), *Seismic Waves, Radiation, Transmission, and Attenuation*, New York: McGraw-Hill.

White, R. M. (1970), *Proc. Inst. Electron. Electronic Eng.* **58**, 1238–1276.

Wilson, O. B. (1985), *An Introduction to the Theory and Design of Sonar Transducers*, Washington, DC: U.S. GPO.

Wilson, W. D. (1960a), *J. Acoust. Soc. Am.* **32**, 1357.

Wilson, W. D. (1960b), *J. Acoust. Soc. Am.* **32**, 641–644.

Wilson, W. D. (1962), *J. Acoust. Soc. Am.* **34**, 866.

Further Reading

Cady, Walter G. (1946) *Piezoelectricity*, New York: McGraw-Hill. The applications of piezoelectricity have been dominated by the usage of piezoelectric-acoustical resonators. The usages include physical standards of time and frequency (quartz), the accurate control of radio-frequency oscillators (quartz), narrow-band electroacoustical filters used

in multiplexed telephone lines (ethylene diamine tartrate), hydrophones and projectors for under water sound (barium titanate), microphones and loudspeakers for audio frequencies in air (Rochelle salt, ammonium dihydrogen phosphate, electrets), and medical ultrasonic applications. Cady's monumental work covers the basic physical principles for these applications.

Hueter, Theodor F., Bolt, Richard H. (1955) *Sonics*, New York: Wiley. The authors describe techniques for the applications of axoustics in engineering and physics. They state that "sonics encompasses the analysis, testing, and processing of materials and products by the use of mechanical vibratory energy." About half of the text is necessarily devoted to explanations of the fundamental principles of electroacoustics. They describe apparatus and instrumention for the industrial application of substantial acoustical power.

Mason, Warren P. (1948) *Electromechanical Transducers and Wave Filters*, 2nd ed. New York: D. Van Nostrand Co. Efficient electromechanical transducers, to transform acoustical energy into electrical waveforms and vice versa, are at the heart of almost all applications of acoustics. Mason's comprehensive treatise summarizes most of the known technology of transducers.

Rayleigh, Lord (1894, 1896), *Theory of Sound*, London: Macmillan and Co. 2nd ed., Vols. I and II. The two volumes were reprinted by Macmillan in 1937 and 1940 repectively, with a few small corrections. This treatise was written before the development and application of modern electronics to acoustical measurements. However, it remains to this day the primary treatise on the theoretical and mathematical aspects of the physics of sound.

Tolstoy, Ivan, Clay C. S., (1966) *Ocean Acoustics*, New York: McGraw-Hill. The book deals with the theory of sound propagation in the oceans, and with the comparison of theory with experiment. The authors use, as the essence of underwater acoustical theory, the simple scalar wave equation in which the propagation of sound pressure depends on the single parameter *c*, the local speed of sound at the point (x,y,z).

ENGINES, DIESEL

See DIESEL ENGINES

ENVIRONMENTAL HEALTH AND SAFETY

B. H. STROHM AND P. J. BEATTIE, *Toxic Materials Control Activity, General Motors Corporation, Detroit, Michigan, U.S.A.*

Introduction 271
1. **Toxicology and the Identification of Environmental Health Hazards** 272
1.1 Hazard vs Safety
1.2 Descriptive Toxicology 273
1.3 Regulatory Toxicology 273
2. **Defining Chemical Toxicity** 273
2.1 Severity of Effect 273
2.2 Systemic Toxicity 273
2.3 Dose 273
3. **Dose-Response Relationships** 274
3.1 The Dose-Response Curve 274
3.1.1 LD_{50} 274
3.1.2 Thresholds 274
3.1.3 Potency 275
3.1.4 Multiple Chemical Responses 275
4. **Physiological and Biochemical Determinants of Toxicity** 276
4.1 Absorption 276
4.2 Distribution 277
4.3 Metabolism 277
4.4 Excretion 277
5. **The Assessment of Risks Due to Chemicals in the Environment** 277
5.1 Hazard Identification 277
5.2 Dose-Response Analysis 8
5.3 Exposure Assessment 278
5.4 Risk Characterization 278
5.5 Uncertainty in Risk Assessments 279
6. **Environmental Chemicals and Cancer** 279
6.1 Environmental Factors Associated With Cancer 280
6.2 Carcinogenic Mechanisms
6.2.1 Genetic and Epigenetic Carcinogens 280
6.3 Cancer Risk Assessment 281
6.3.1 Acceptable Risk 281
6.3.2 Relative Risk 282
7. **Some Environmental Agents of Concern** 283
7.1 Heavy Metals 283
7.1.1 Lead 283
7.1.1.1 Fate 283
7.1.1.2 Health Effects 284
7.1.2 Mercury 284
7.1.2.1 Fate 284
7.1.2.2 Health Effects 284
7.1.3 Cadmium 285
7.1.3.1 Fate 285
7.1.3.2 Health Effects 286
7.2 PCBs 286
7.2.1 Fate 286
7.2.2 Health Effects 286
7.3 Ozone 287
7.4 Extremely Low Frequency (ELF) Magnetic Fields 287
8. **Conclusion** 288
Acknowledgment 288
Glossary 288
Works Cited 288
Further Reading 288

INTRODUCTION

Environmental health is a very broad area of study encompassing biological, chemical, physical, sociological, and psychological factors that may affect man's health and well-being. Historically, communicable disease has been at the center of most environmental health and safety concerns. Identification of causative organisms, routes of transmission, populations at risk, and the subsequent development of strategies to prevent transmission

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

or treatments and cures for communicable diseases have dominated public health activities. Today, while biological hazards such as acquired immunodeficiency syndrome (AIDS) continue to be enormous public health issues, other hazards have emerged and are demanding greater attention and an increasing share of society's environmental health resources.

As countless headlines, special reports, legislative acts, and regulations attest, concerns over the presence of chemicals in the environment and their possible effects upon human health have attained preeminence. It is imperative that we understand the true nature of these chemical hazards and associated risks. It is not enough for us merely to be able to take necessary actions to eliminate unacceptable environmental health hazards. We must also be capable of separating real chemical health hazards from perceived or insignificant ones so that limited health resources are not misappropriated.

Mankind has evolved in continual contact with chemical and physical hazards. Radioactive compounds such as uranium and radon gas, the sun's ultraviolet radiation, extremely low-frequency magnetic fields, toxic organic chemicals such as benzene and benzo(a)pyrene, heavy metals such as lead and mercury, and chemical oxidants such as ozone, nitrogen oxides, and sulfur oxides, have been present in man's environment from our beginnings. As Ames *et al.* (1987) have recently observed, "Nature is not benign." The human diet is laden with a tremendous diversity of naturally occurring toxins, mutagens, and carcinogens. Human activities which accompany technological, industrial, and economic advancement have, in some cases, increased the possibility of exposure to both new and existing chemical or physical hazards. Total elimination of these hazards is not feasible in that we will always be left with exposure to naturally occurring toxins or physical agents.

Effective environmental health and safety management is dependent upon our ability to identify significant health hazards and to take action to protect the public from excessive exposure to the many potentially toxic chemicals present in our environment. This article will focus upon the identification and evaluation of chemical hazards and the assessment of the risks they present through their release to the environment.

The methodologies employed when assessing environmental risks are essentially the same for both chemical and physical health hazards. In particular, the relationship between dose and toxicity was first observed and characterized in studies of the biological effects of radiation. Specific treatment of physical environmental hazards such as UV and ionizing radiation and electromagnetic fields are covered in BIOLOGICAL EFFECTS OF ELECTROMAGNETIC AND PARTICLE RADIATION.

1. TOXICOLOGY AND THE IDENTIFICATION OF ENVIRONMENTAL HEALTH HAZARDS

Toxicology is the study of adverse chemical effects upon biological systems. It has progressed from ancient observations on plant and animal poisons to an understanding at the molecular level of some of the mechanisms involved in chemically induced disease processes, such as cancer. Toxicology is a cross-disciplinary field of study encompassing a number of sciences including chemistry, biology, biochemistry, and pharmacology.

Toxicology has evolved to fulfill a critical need in modern society. With the ever-increasing role that chemicals and chemical technologies play in our daily lives, there has been a commensurate realization of the need to test these materials for possible consequences upon human health and the environment. Industrial chemicals, drugs, food additives, consumer products, cosmetics, etc., must all be evaluated in a manner which enables us to predict possible adverse effects due to their intended use.

The final application of toxicology is in the identification and characterization of chemical hazards. This task is complicated by the fact that the toxicity or effect produced by any one chemical is dependent upon a multitude of factors. The same chemical can elicit beneficial effects under one set of conditions, and harmful or lethal effects under another. Therapeutic drugs, or, in some cases, essential minerals are the most common examples of this phenomenon.

1.1 Hazard vs. Safety

In attempting to characterize chemical hazards, toxicology studies also provide the information necessary for assessing chemical safety. Chemical toxicity is a relative measure defined by both the severity of effect and the dose or exposure necessary to produce the effect. At sufficient dose, all chemicals can produce adverse effects. Oxygen, at high concentrations, can damage the lung, and approximately 100 g of table salt can produce life-threatening electrolyte imbalances.

Although this implies that no chemical is entirely safe, it has also been observed that no chemical should be considered entirely harmful. Even the most toxic chemicals exhibit minimally effective doses, below which adverse effects can be avoided. In fact, use of a very toxic chemical may be less hazardous than use of a relatively nontoxic one, depending upon differences in the resulting exposures.

1.2 Descriptive Toxicology

Descriptive animal toxicity testing under well-defined conditions of exposure forms the foundation of chemical hazard evaluation. These studies are the starting point for evaluating the health and safety ramifications of chemicals in the environment. Most descriptive toxicity testing follows guidelines prescribed by regulation and is designed to predict possible effects in humans (Environmental Protection Agency, 1985a; Hayes, 1989). In general, animal models have been found to be very good predictors of possible human responses.

1.3 Regulatory Toxicology

From these experimental results, regulatory toxicologists must extrapolate findings, most often from homogeneous animal colonies, to a heterogeneous human population. Regulatory toxicologists provide agencies, legislators, and consumers with information on the hazards or risks associated with the use and occurrence of chemicals in our daily lives. This activity usually culminates in the setting of standards which limit the quantity of a chemical to which an individual may be exposed through air, food, water, or the use of chemical materials at work or home.

2. DEFINING CHEMICAL TOXICITY

Chemicals can affect virtually every biological system of the body. From a biochemist's perspective, life itself can be viewed as a sophisticated and complex series of interrelated chemical reactions and equilibria. Perturbation of any of these reactions by a foreign chemical could be considered an effect. Whether this effect is considered a toxic or adverse effect or merely an adaptive response in order to maintain homeostasis depends upon the severity of the effect.

2.1 Severity of Effect

Severe toxic effects are most often thought of as life-threatening or debilitating. They can manifest themselves rapidly after short-term single exposures, or can be delayed and appear after some passage of time. Certain chemical carcinogens have exhibited latency periods of 20–30 y from time of initial exposure before onset of clinically observable symptoms or tumors.

Severity is also related to whether the effect is reversible or irreversible. The capacity of a tissue to repair itself or of an organ's ability to regain function influences severity. The liver possesses significant regenerative capacity and, thus, toxic effects which manifest in the liver are often reversible. Cells of the central nervous system do not replicate and tend to be the site of more severe effects.

2.2 Systemic Toxicity

With the exception of the lung, the most severe chemically induced environmental health effects appear at sites distant from the point of entry. These are termed "systemic effects" and are usually somewhat selective in that they are produced at exposures below those concentrations which would normally elicit a local effect. Chemical selectivity is a product of both the specific target or site of action for the exogenous compound and the importance of this reactive site to the viability and function of the tissue or organ.

2.3 Dose

The single most important factor determining the toxicity of a chemical is dose. As first expressed by Paracelsus in the 1500s, all chemical substances are potentially harmful to living organisms. The level of exposure or amount of a chemical to which an individual is exposed will decide the effect.

The total dose received by an individual is a function of the exposure concentration, duration, and, for intermittent exposures, frequency. Animal toxicity testing is classified into acute, subacute, subchronic, or chronic on the basis of duration of exposure. Toxicity is generally a function of the total dose. At lower concentrations of exposure, toxicity is more likely to be manifested by exposures of longer duration or greater frequency.

Acute effects, produced by short-term exposure to high doses of a chemical, often produce different effects than long-term, low-level exposure to the same chemical. Ethanol is a familiar example of a chemical which produces primary acute effects upon the central nervous system, and long-term effects in the liver. Toxicity, or the severity of response, tends to decrease as the dose administered is dispersed over time. The effect produced by consumption of six 1-oz servings of alcohol in 6 h is less than that produced by six 1-oz servings in 1 h. If the dose is spread out over a great enough length of time, there may be no observable effects (e. g., one-tenth of an ounce of alcohol per day for 60 days).

This decrease in severity of effect as the dose is fractionated over time occurs in large part through detoxification and excretion of the chemical between successive doses. Diminished effects may also be observed when the cumulative effect or injury produced by each administered dose is partially or fully reversed before the subsequent dose is administered.

3. DOSE-RESPONSE RELATIONSHIPS

The dose-response relationship is the most fundamental and important concept in toxicology. It is derived from the premise that biological responses to a chemical will increase in magnitude or frequency as dose is increased. Characterization of the effects produced by a chemical at varying doses or concentrations of exposure, via different routes, and with changing exposure duration and frequency define chemical hazards. Chemical safety then becomes an exercise in determining where on the dose-response curve a specific individual's potential dose or exposure may lie.

3.1 The Dose-Response Curve

The generalized sigmoid curves depicted in Fig. 1 are classically observed in toxicity testing, provided sufficient numbers of animals and doses are employed over a broad enough range of concentrations. The *y* axis corresponds to the magnitude or frequency of the response produced. It can be representative of the severity of an effect produced in an individual (i.e., low, moderate, or severe irritation) or be a measure of the percentage of individuals exhibiting a response in an exposed population (e.g., one in one million persons exposed may develop cancer).

The *x* axis corresponds to the dose or level of exposure. Doses are most often described as the weight of test substance per weight of animal (i.e., milligram per kilogram). Airborne exposures are expressed as parts of toxicant per parts of air (parts per million—ppm) or, for suspended particulate matter, quantity per volume of air (mg/m^3).

The sigmoid dose-response curve is representative of a normally distributed test population. All biological responses to chemicals exhibit some degree of individual variability, which accounts for the shape of this curve. At low doses only sensitive individuals in a population will respond to a chemical insult. As dose increases, the percentage of individuals responding will increase, and at the highest doses only resistant individuals will not exhibit a response.

3.1.1 LD_{50}. One of the most frequently monitored toxicological endpoints is acute lethality. Lethal-dose studies act as a starting point for the design of subsequent long-term exposure investigations. The acute dose of a substance producing death in 50% of the animals exposed is termed the lethal dose–50, denoted LD_{50}, and is shown graphically in Fig. 1 for two different compounds, A and B.

3.1.2 Thresholds. Figure 1 also depicts a point for both chemicals A and B where the percentage of animals responding is essentially zero. The corresponding dose is termed the threshold, or no-observed-effect level (NOEL). Estimation of chemical no-observed-

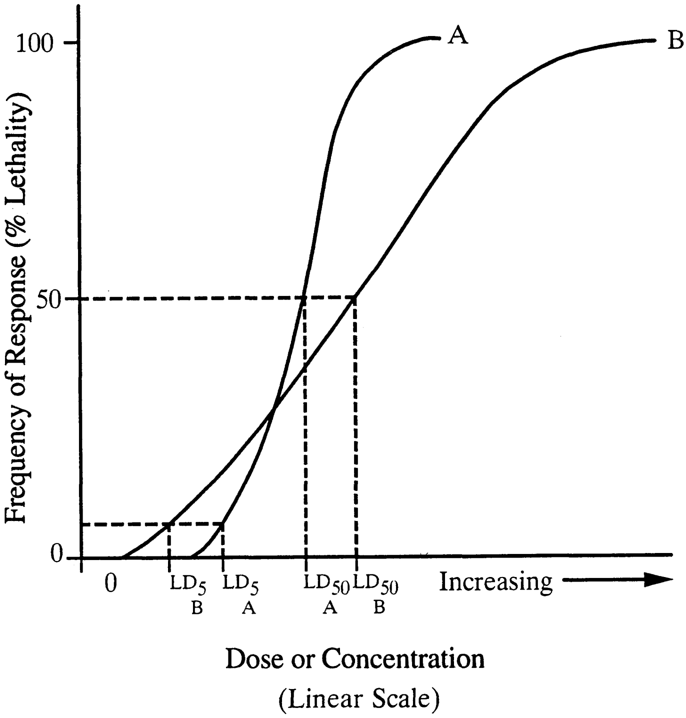

FIG. 1. Hypothetical dose/response curves for two compounds, A and B, depicting lethal dose-50 and lethal dose-5 intercepts.

effect levels from animal studies is fundamental to the evaluation of chemical hazards and risk assessment.

3.1.3 Potency. Common endpoints, such as lethality, provide a means for comparing the toxicity of chemical substances. Figure 1 indicates that the dose of compound A needed to produce lethality in 50% of the animals is less than that for compound B. Compound A would be considered more toxic or potent in producing death in 50% of the animals exposed than Compound B.

In comparing potency of chemicals, it is important to consider the anticipated level of exposure. As Fig. 1 demonstrates, the shapes of the dose-response curves for two chemicals may reveal that differences in potency can arise depending upon the dose levels being compared. While Compound A appears more potent with respect to LD_{50}'s, the doses required to produce death in 5% of the animals indicate that compound B is more potent. Moreover, compound B appears to have a lower threshold dose, indicating a greater hazard than A at lower exposure levels.

3.1.4 Multiple Chemical Responses. An added dimension of chemical dose-response evaluation is the fact that almost all chemicals are capable of producing more than one effect. Each of these effects has its own characteristic dose-response relationship. As Fig. 2 indicates, the no-observed-effect level for one effect may be observed at a dose producing nearly maximal response for other effects. One of the objectives of toxicological testing is to identify the most sensitive adverse effect and determine the lowest level of exposure necessary to produce it.

In some cases, the spectrum of effects produced by a chemical with increasing dose may reflect an increase in severity of response. For example, at very high doses a chemical may kill a large number of liver cells, destroying liver function and causing death in some or all of the animals exposed. At lower doses, fewer cells may be killed and the animal survives, but the affected cells produce other types of functional detriment, which may be observed as jaundice. At still lower doses, no cells are killed but alterations in cell function or structure may occur that can be observed only through specialized biochemical tests or histological examination. Finally, a dose may be achieved at which no effects are

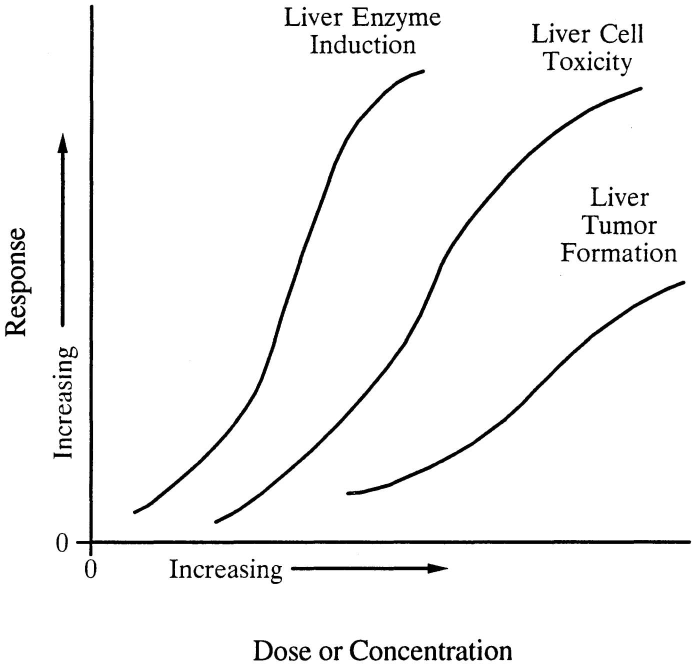

FIG. 2. Hypothetical dose/response curves for three effects produced by administration of the same compound.

observed or only changes that have no adverse impact upon the animal.

4. PHYSIOLOGICAL AND BIOCHEMICAL DETERMINANTS OF TOXICITY

The physiological and biochemical determinants of chemical toxicity are intimately related to the dose-response relationship. While administered dose is what is measured in classic toxicity testing, it is the target-site concentration that ultimately determines the response. The factors that establish and modify the relationship between the administered dose and the actual chemical concentration produced at the target site are absorption, distribution, metabolism, and, excretion.

4.1 Absorption

The physical chemical properties of a substance, the dose coupled with the route of exposure, determine how readily a chemical is absorbed across the body's epithelial surfaces. The surface areas exposed and, thus, the potential for chemical absorption vary with route of exposure. While the average individual has a skin surface area of approximately 1.8 m^2, chemicals are rarely absorbed via whole-body immersion. In contrast, the lung, with an average alveolar surface area of 50–100 m^2, and the small intestine, with a microvilli-enhanced surface area of approximately 185 m^2, present very large continuous surfaces for exposure. With every inspiration of chemical-contaminated air and for ingested chemicals along the entire length of the gastrointestinal tract, chemical absorption takes place. In general, and with decreasing effectiveness, toxic chemicals produce their most severe and most rapid effects upon inhalation, ingestion, and, lastly, skin absorption.

Absorption of a chemical into the body, as well as distribution of the chemical to its ultimate target site, is dependent upon passage of the compound across cellular membranes. Membranes represent barriers to chemical absorption only for compounds possessing specific chemical properties. For example, ionized chemicals do not cross membranes to the same degree as their nonionized forms. Nonpolar and lipid-soluble chemicals cross biological membranes more readily than water-soluble compounds. Mechanisms of transport across membranes can involve passive diffusion, active or facilitated transport, as well as endocytosis.

4.2 Distribution

Upon absorption of a chemical into the body fluids, distribution determines specific target-tissue concentrations. The circulatory system is the principal pathway for chemical distribution. While some toxicants are sufficiently soluble to be transported in solution, the majority move through the circulatory system in association with plasma proteins. Transport through the blood can result in accumulation of a chemical at its site of action, storage in tissues which do not manifest toxicity, or passage to tissues which metabolize or eliminate the chemical.

Chemicals can bind plasma proteins through covalent, ionic, hydrogen, van der Waals, or hydrophobic interactions. Dissociation of the ligand- (toxicant-)protein complex occurs when the affinity of a chemical for another biomolecule or tissue component is greater than that of the plasma protein to which the toxicant was originally bound. If binding is assumed to be reversible, redistribution of a chemical will occur in order to restore equilibrium whenever the concentration in one tissue is diminished.

4.3 Metabolism

Metabolism of chemicals is a major factor which changes parent-compound tissue concentration and, thus, shifts chemical equilibrium and distribution. Xenobiotic (foreign chemicals) metabolism is also a major determinant of chemical toxicity, either increasing the toxicity through bioactivation of a compound to a more reactive form, or detoxification, which decreases the toxicity of a chemical, usually rendering it in a more readily excretable form (Anders, 1985; Caldwell and Jakoby, 1983).

4.4 Excretion

Chemical excretion decreases whole-body concentration and causes chemical redistribution as equilibrium between body compartments is reestablished. Chemical removal from the body is expressed in terms of biological half-lives or the time required to remove one-half of the chemical from the body. Chemical elimination rates may have several phases depending upon the specific biological mechanisms involved. In general, nonpolar chemicals, which are readily absorbed, will not be excreted until in a more polar, water-soluble form. Metabolic conversion critical to removal of these compounds may often be the rate limiting step for chemical excretion.

5. THE ASSESSMENT OF RISKS DUE TO CHEMICALS IN THE ENVIRONMENT

The previous discussion was intended to highlight the complexities involved in attempting to define chemical toxicity, the first step in identifying potential environmental health hazards. Hopefully, it is evident that chemical toxicity is not a single value or determination, but rather a dependent variable which is particularly sensitive to dose. This complexity and dose dependency is the result of underlying biochemical mechanisms that determine whether a given exposure can be detoxified and eliminated without consequence or results in target-site concentrations of an active compound high enough to produce adverse effects. The greater our understanding of the mechanisms involved in chemical toxicity and detoxification, the more precisely we can identify potentially hazardous exposures, which will ultimately lead to more accurate environmental health hazard assessments.

The formalized process in which experimentally derived health hazard information is compared with estimates of actual human exposure to determine the likelihood of a toxic effect is called risk assessment. Risk assessment can be divided into four steps: hazard identification; dose-response assessment; exposure assessment; and risk characterization.

While the following discussion will highlight chemical risk assessment, the process and the four steps involved are equally applicable to the assessment of environmental risks due to physical hazards. Acceptable levels of exposure to naturally occurring radon gas, as well as allowable levels of exposure to ionizing radiation during medical x-rays, are determined by this process.

5.1 Hazard Identification

Hazard identification involves the evaluation of all available toxicological information on a chemical. Toxic effects most often observed in animals are identified, characterized, and extrapolated to man. Species, sex, mechanism of action, and exposure characteristics utilized in generating the experimental data must be interpreted and their applicability to potential human exposures determined.

An example of the complexity of this type of consideration is the toxicity of *d*-limonene, a naturally occurring monoterpene extracted from orange peels. The use of *d*-limonene as a degreasing solvent has escalated in an attempt to find safer alternatives to the halogenated solvents traditionally used in these applications. *d*-limonene was recently tested in a two-year carcinogenicity study in male and female rats and mice (National Toxicology Program, 1988). Results indicated that *d*-limonene was carcinogenic in male rats only, producing tumors of the kidney. Further mechanistic investigation identified a protein (α-2μ-globulin), found only in male rats, that was structurally altered through binding with a bioactivated form of *d*-limonene. This altered protein then accumulates in the male rat kidney, apparently initiating inflammatory tissue reactions which eventually lead to tumor formation. There is no indication that *d*-limonene binds DNA directly. The hazard identification step of risk assessment includes determination of whether this type of data is relevant to potential human exposures.

5.2 Dose-Response Analysis

The second step in risk assessment arises from the data reviewed in the hazard identification step and involves analysis of dose-response relationships. For each chemical, all potential adverse effects must be examined and the individual dose-response curves evaluated. The effects produced at the lowest level of exposure are usually identified through this procedure. Since environmental exposures are most often well below the concentrations or doses employed in experimental studies, responses must be projected or extrapolated to levels of exposure beyond those actually tested.

5.3 Exposure Assessment

The steps of hazard identification and dose-response assessment provide the risk assessor with a characterization of the potential effects that a chemical might produce and the types of exposures necessary to elicit these effects. Determination of the risk or likelihood that an individual may experience an effect due to the presence of a chemical in the environment requires an estimate of dose or exposure. Estimating environmental exposures to chemicals can be as complex as the evaluation of toxicological information.

Chemicals can enter the environment from point sources, such as the outfall of a waste water treatment plant, or non-point sources, such as the tailpipe of an automobile. Computer modeling is often employed to predict down-range concentrations of chemicals released to the atmosphere or entering surface or ground water.

Human exposures to environmental chemicals can occur via the air, from consumption of food and beverages, and, in some cases, through skin contact with contaminated soil or water. Upon release of a chemical to the environment, numerous factors will determine which of these pathways of exposure is most probable and whether the resulting exposure concentrations are significant.

The environmental fate of a chemical is dependent upon physical chemical characteristics, the point of origin, the medium to which the chemical is released, transport between environmental compartments, and chemical transformation. With the exception of persistent chemicals and chemicals which bioaccumulate, most chemicals from the time of release are either diluted or degraded, decreasing in concentration from their point of origin. Transport between environmental compartments can take place via numerous pathways including evaporation of liquids into vapor or gaseous phases, precipitation of chemicals out of solution, absorption to particulates, particulate fallout from the atmosphere, or atmospheric resuspension of particulates. Chemical transformation or degradation can take place through chemical, photochemical, and biological mechanisms.

Exposure assessment must estimate the resulting human uptake of chemicals that may be widely and unevenly distributed in the environment. Assumptions and estimates are made regarding contaminant concentrations in food, air, water, and soil, as well as quantities of various foods typically consumed, volumes of air inhaled, and amount of skin contact with soil.

5.4 Risk Characterization

The final step in risk assessment is risk characterization. In this process, information from the previous steps is summarized. Risk characterization should not focus upon one estimate of risk resulting from one exposure scenario or measure of toxicity, but rather should compare a range of estimates, identifying the underlying assumptions and uncertainties. "Weight of evidence" is often the phrase used to describe the most accepted approach to risk characterization and includes a maximum likelihood or "best" estimate based upon all available data.

Risk characterization for noncarcinogenic effects centers upon determination of no-observed-effect levels (NOELs) for all adverse effects identified. Human acceptable daily intakes (ADIs) are derived from the experimental NOEL. ADIs usually incorporate safety factors which attempt to compensate for uncertainties involved in extrapolating animal findings to man.

Risk characterization for carcinogens is complicated by the theoretical plausibility that all exposures carry with them some finite risk. This risk may be so extraordinarily small that, in fact, a practical threshold does exist. Cancer risk assessments attempt to provide decision makers with information which allows for the determination of tolerable levels of risk.

The risk-assessment product represents the scientific treatment of all relevant data. This product is presented to a risk manager who must then make decisions as to the necessary course of action. Risk management decisions result in banning of a chemical, limiting use, or imposing no restrictions.

5.5 Uncertainty in Risk Assessments

There are two principal sources of uncertainty when one is assessing the environmental risks of chemicals. The first is that most chemical toxicity data come from animal studies which must be extrapolated to man. The second is the need to estimate possible human exposures to environmentally disseminated chemicals.

Unlike the controlled conditions employed in animal experimentation, real-world exposures to chemicals are often unknown. They can occur via multiple routes, can be of inconsistent duration and frequency, and usually involve very low concentrations that can fluctuate over time. The mere presence of a chemical in the environment does not necessarily represent a hazard or risk. As analytical methods improve, chemicals will be detected more often and at lower concentrations than ever before. The health significance of these findings will be determined by an evaluation of potential exposure.

Precise estimation of dose would require a tremendous amount of data, specific to each individual—unique to his/her environment, personal habits, and preferences. Food consumption and personal activity patterns, such as occupation and amount of time spent indoors or outdoors, are all determinants of potential environmental chemical exposure. In the absence of this type of data, risk assessments utilize estimates of exposure. Given the importance of dose in determining whether a chemical will produce a toxic effect and thus truly represent a hazard, environmental exposure estimates represent a critical underpinning of any environmental risk assessment.

While the many uncertainties involved in the risk assessment process have led some to think that it is of little value to the risk manager, when performed consistently risk assessments do offer a means for comparing risks and separating significant hazards from trivial ones.

6. ENVIRONMENTAL CHEMICALS AND CANCER

The threat of cancer due to chemicals in the environment is the most publicized and possibly most significant chemical health and safety issue confronting society today. Public concern over cancer is justifiable. Approximately 22% of all deaths in the U.S. are now due to cancer. But upon further analysis of age-adjusted mortality rates for all cancers combined, it becomes apparent that overall cancer rates have remained the same or declined since 1950, with the notable exception of smoking-related lung cancer (American Cancer Society, 1988). This trend is contrary to what one might expect after examining the production and use of synthetic organic chemicals since 1945, which has increased from about 5×10^8 kg/y to 1.5×10^{11} kg/y (Davis and Magee, 1979), and is certainly in conflict with the numerous reports of increased cancer risks due to the presence of chemicals such as Alar, PCBs, and formaldehyde in our environment. The reasons for this apparent discrepancy between the perceptions of risks due to chemicals and actual cancer rates are evident after investigation of what we know about the causes of cancer and also how cancer risk assessments are often performed and reported.

6.1 Environmental Factors Associated With Cancer

The causes of cancer are multifactorial and include genetic, biological, and environmental factors. It is generally accepted that 80% to 90% of all cancers are causally associated to varying degrees with numerous types of environmental factors. Environmental factors are those over which we have some control, and for this reason have been the focus of many cancer prevention strategies. One manner in which these factors have been identified is through comparison of the incidences of specific types of cancer in different parts of the world. For example, colon and breast cancer, which are among the major types of cancer in the U.S., are quite rare among the Japanese in Japan, but not among Japanese Americans. Smoking, diet, ionizing radiation, drugs, sunlight, and chemicals are all known to be capable of producing cancer.

Precise determination of the percentage of cancers caused by any one factor is not possible. Estimates have been proposed based upon the limited data available, but with the exception of cancers associated with tobacco, there remains considerable debate over the proportionate causes of cancer. In perhaps the most often cited study, Doll and Peto (1981) have estimated that 30% of U.S. cancers are associated with tobacco, 35% with nutritional factors, 3% with alcohol, 7% with reproductive and sexual behavior, 3% with physical factors such as UV radiation, 10% with infection, 1% with food additives, 2% with pollution, 1% with industrial products, 1% with medicines, and 4% with occupation. While some controversy remains as to whether environmental chemicals may play more or less of a role in the induction of cancer, other studies seem to indicate that the relative hazards presented by environmental chemicals when compared to other sources of exposure to carcinogens may be very low.

6.2 Carcinogenic Mechanisms

Cancer is characterized by the abnormal, uncontrolled, and seemingly unrestricted growth of the body's own cells, giving rise to cellular masses called tumors. Tumors can be benign, remaining at the site of origin, or malignant, spreading or metastasizing from the primary site to give rise to secondary tumors in other areas of the body. The symptoms of cancer are usually the result of tumor growth invading surrounding tissues and interfering with normal body function.

Chemical carcinogens include both naturally occurring and man-made chemicals. While chemical carcinogens exhibit a spectrum of toxic effects and properties similar to other chemical toxicants, they differ through their ability to produce heritable change at the cellular level. The transformation of a normal cell to a cancerous, or neoplastic, cell with subsequent tumor formation proceeds through a complex series of events. The initiating event for many chemical carcinogens is thought to be interaction with DNA. This interaction results in change in the DNA sequence, which ultimately alters cellular growth characteristics giving rise to the transformed cell. Most chemical carcinogens require bioactivation to reactive forms which then bind DNA.

The entire process involved in chemically induced cancer is under intricate control and subject to numerous modifiers. Species, strain, sex, age, and nutritional status have all been observed to alter significantly an animal's response to chemical carcinogens.

6.2.1 Genetic and Epigenetic Carcinogens. Chemical carcinogens have generally been classified into two types based upon mechanisms of action: the genetic and epigenetic carcinogens.

Genetic, or genotoxic, carcinogens are DNA reactive and thought to be directly responsible for DNA alteration and initiation of tumor development. Epigenetic carcinogens do not bind DNA directly and, apparently, facilitate tumor development through other mechanisms.

In general, genotoxic carcinogens are more potent than chemicals working through epigenetic mechanisms. Additionally, there is mounting support for the existence of thresholds or levels of exposure to epigenetic carcinogens which do not induce tumor formation.

6.3 Cancer Risk Assessment

Results from animal experiments have been the major source of data for estimating human risk from exposure to potentially carcinogenic chemicals. Animal tests for chemical carcinogenicity generally employ rodents (mice or rats) exposed to a chemical for a major portion of their life span. Rodents have been found to be very similar to man anatomically and in the way they metabolize chemicals. Generally, the tumors produced by chemicals in rodents have similar characteristics to those found in man.

A limitation in these studies, due to the small number of animals tested, is that weak carcinogens would probably not be identified as cancer causing. In place of increasing the size of the study, which would be costly, large doses (termed the maximum tolerated dose) of chemicals are typically used in order to detect even weak carcinogens. These doses, in virtually all cases, are much larger than those to which humans would be subjected. The rationale for this is that, in general, the number of tumors will increase as the dose of a carcinogen increases. While the use of the maximum tolerated dose in animal studies is controversial and some difficult questions remain to be answered, regulatory agencies, in order to maintain a conservative and prudent public health policy, must rely upon the results of these animal tests.

In general, it appears that animal testing is a very good indicator of a chemical's possible carcinogenicity in humans. Of the 32 chemicals thought to be carcinogenic in humans, 26 have been shown to produce cancer in experimental animals. In these comparisons, the target organ of carcinogenicity was the same 80% of the time.

Beyond the determination of whether a chemical may or may not produce cancer in humans, risk managers are concerned with the probability or risk to humans upon exposure to a suspect carcinogen. From these animal bioassays, regulatory toxicologists attempt to determine the likelihood of humans developing cancer at a particular exposure level. Currently, regulatory agencies, such as the U. S. Environmental Protection Agency (EPA), assume that there is no threshold or dose at which a cancer-causing chemical would not induce tumors. From these animal studies, which are conducted at relatively high doses, regulators attempt to predict the response in humans due to exposures to these chemicals at much lower doses.

These types of extrapolations, while questionable with respect to their relevance to actual human carcinogenic risks, do allow for the establishment of exposure levels to suspect carcinogens which would not be anticipated to produce significant numbers of cancers in exposed individuals. In addition, this type of dose-response analysis allows comparison of the potency of chemicals to induce cancer when tested in similar animal bioassays. Chemicals with a high potency induce higher numbers of tumors at lower levels of exposure or induce tumors in a much shorter period of time. Low-potency carcinogens induce lower numbers of tumors and require greater doses or a greater length of time to produce tumors in animals. As Table 1 illustrates, comparison of carcinogenic potencies indicates that all carcinogens are not alike. There do exist real and measurable differences in the capacity of chemicals to produce tumors in animals; in this example bis(chloromethyl)ether is approximately 100,000 times as potent as chloroform.

6.3.1 Acceptable Risk. While the analyses in Table 1 are useful when comparing carcinogenic potencies of chemicals within the dose ranges studied, the actual tumor incidence rates produced in these studies are beyond acceptable limits for potentially ex-

Table 1. Comparison of carcinogenic potencies.[a]

Chemical	Potency[b]
Chloroform	6.1×10^{-3}
Styrene	3.0×10^{-2}
Vinyl chloride	2.3
n-Nitroso di-n-propylamine	7.0
Bis(chloromethyl)ether	2.2×10^{2}

---

[a] Based upon values from Environmental Protection Agency (1990).

[b] Relative risk per unit dose

Table 2. Estimates of lifetime risk to humans from exposure to methylene chloride[a] (95% upper confidence limit of additional risks)

Air concentration $\mu g/m^3$	Multi-stage model	One-hit model	Weibull model	Log-probit model
1	1.8×10^{-7}	2.0×10^{-7}	4.8×10^{-10}	3.5×10^{-31}
10	1.8×10^{-6}	2.0×10^{-6}	1.7×10^{-8}	1.6×10^{-22}
100	1.8×10^{-5}	2.0×10^{-5}	6.1×10^{-6}	2.5×10^{-15}
1 000	1.8×10^{-4}	2.0×10^{-4}	2.0×10^{-5}	1.3×10^{-9}
10 000	1.8×10^{-3}	2.0×10^{-3}	6.1×10^{-4}	2.4×10^{-1}

[a] Environmental Protection Agency (1985b).

posed human populations. Animal carcinogenicity testing is limited in its power of detection by a combination of the numbers of animals typically employed and the spontaneous background rates of tumor formation observed in control animals. Most animal studies identify tumor incidence rates of 10% or greater, which would translate into 1 cancer per 10 individuals exposed.

EPA policy has established an acceptable level of risk for suspect carcinogens to be one cancer per one million individuals exposed. This necessitates the extrapolation of experimentally derived tumor incident rates in the range of one cancer per ten animals exposed to doses which might induce tumors in one individual per one million exposed.

Current EPA policy requires the use of a linearized multistage extrapolation model to derive 95% upper confidence limits for these doses. Other models for extrapolation have been proposed, which utilize equally valid but less conservative mechanistic assumptions.

Table 2 compares estimates generated by use of the EPA multistage model with those produced by three other models. These data were derived by the EPA in assessing health risks due to airborne methylene chloride exposure (Environmental Protection Agency, 1985b). At varying air concentrations ($\mu g/m^3$) the corresponding estimates of numbers of cancers produced per population exposed are presented for the four models. Depending upon the model employed, a lifetime (70-y exposure) risk to humans exposed at 1 $\mu g/m^3$ can range from two cancers per 10^7 individuals exposed (one-hit model) down to 3.5 cancers per 10^{31} persons exposed (log-probit model).

It is important to understand that these incidence rates are hypothetical extrapolations and do not represent actual cancer rates. Misrepresentation of these estimates without explanation of the numerous assumptions which go into their generation, and with no indication of the full statistical range of possible outcomes, has led to much of the public's confusion and perception of cancer risks due to chemicals in the environment.

6.3.2 Relative Risk. In the absence of accurate quantitative methods to evaluate potential carcinogenic hazards, attempts have been made to provide risk managers, policy makers, and the public with qualitative measures of the relative hazards or risks attributable to environmental chemicals. Utilizing carcinogenic potencies and estimates of human exposure, relative risks can be used to assist in prioritizing environmental health needs.

Ames *et al.* (1987) have used a hazard ranking system which computes a HERP index (human exposure dose/rodent potency dose) for comparing carcinogenic risks. Table 3 presents rankings from these data for several chemicals with known human exposures. The greater the HERP percentage, the greater the relative hazard.

This type of comparison provides risk managers and the public with an intuitive aid to understanding the health significance of chemicals in the environment. Hazards from environmental pollutants can be compared to naturally occurring carcinogens in the diet, such as aflatoxin and nitrosamines, or to chloroform in tap water (which is formed as a byproduct during drinking-water chlorination to protect us from pathogenic viruses and bacteria).

Table 3. Relative risk based upon human exposure dose/rodent potency dose (HERP) (Ames *et al.*, 1987).

Daily human exposure	Carcinogen dose per 70-kg person	HERP%
Tap water (1 L)	Chloroform, 8.3 μg (U.S. average)	0.001
Well water (1 L, Woburn, MA)	Chloroform, 12 μg	0.0002
Conventional home air (14 h/day)	Formaldehyde, 598 μg	0.6
Mobil home air (14 h/day)	Formaldehyde, 2.2 μg	2.1
PCBs: daily dietary intake	PCBs, 0.2 μg(U.S. average)	0.0002
Bacon, cooked (100 g)	Diethylnitrosamine,0.1 μg	0.006
Peanut butter (32 g, one sandwich)	Aflatoxin, 64 ng	0.03
Ethylene dibromide: daily dietary (from grains and grain products)	Ethylene dibromide, 0.42 μg (U.S. average)	0.0004

7. SOME ENVIRONMENTAL AGENTS OF CONCERN

Environmental health questions today and in the future will ultimately be resolved through the use of the risk assessment process. As detailed in the previous discussions, toxicity is determined by dose or level of exposure. This precept is equally applicable when considering both chemical and physical agents. Environmental exposures of concern will most often arise as the result of either persistent chemicals which bioconcentrate or bioaccumulate through the food chain to produce elevated body burdens, or widespread emission sources which may result in exposure of large populations. Heavy metals and polychlorinated biphenyls (PCBs) represent persistent chemicals which have been the subject of numerous investigations and public health initiatives. Ozone and extremely low-frequency magnetic fields represent environmental agents with widespread sources of exposure.

7.1 Heavy Metals

Heavy metals are naturally occurring elements which are present throughout the Earth's crust. They are unlike many organic chemicals in that they are neither created nor destroyed by human activities. Environmental health hazards due to heavy metals often arise from their redistribution in the environment giving rise to an increased likelihood of exposure. Industrial, agricultural, and mining operations may enhance the bioavailability of these metals. The toxicity and resulting hazard of a specific heavy metal can vary dramatically depending upon form and route of exposure. In the general population, food sources represent the greatest source of exposure to metals.

Several heavy metals, because of a combination of their persistence in the environment as well as their ability to bioaccumulate through the food chain or bioconcentrate in human tissues, are of particular environmental concern.

7.1.1 Lead. Environmental lead is the result of both natural and anthropogenic sources. In the recent past, combustion of leaded gasoline was the primary source of lead from human activities. Today, with the removal of tetraethyl lead from gasolines, lead emissions have been dramatically reduced. Other sources of concern have included lead pigments used in paints and lead solder used in plumbing. Now, nonferrous smelters, battery plants, and chemical plants remain the principal sources of anthropogenic lead, although these sources have decreased significantly due to improvements in control technologies.

7.1.1.1 Fate. In air, lead exists as a particulate. Atmospheric transport depends in large part upon the aerodynamic diameter of this particulate. Large particulates (those generally greater than 2 microns in diameter) are removed through settling relatively rapidly, and deposited close to the source. Smaller particulates may be transported up to 1000 miles.

Concentration of lead in surface waters is largely dependent upon solubility of the specific lead compound, *p*H, and dissolved salt content. Solubility tends to increase as salt content decreases. A significant proportion of lead in surface waters can be carried as suspended particulates of large, undissolved lead precipitates.

Lead in soil is affected by a large number of factors including absorption and chelation with organic substrates. Most lead is retained very strongly in soil and little is tranported into surface waters or ground water. Lead leaching in soil is normally very slow. Factors which may increase lead migration include the presence of water-soluble chelators and decreased soil *p*H.

Plants and animals may bioconcentrate lead, but biomagnification through the food chain has not been observed. Lead exposure in the general population is most likely through ingestion of food and drinking water. Airborne particulates containing lead can result in inhalation exposure and are important in the transport and deposition of lead-containing dusts and soils onto vegetation and into surface waters. On the basis of 1982 data, the U. S. Food and Drug Administration (FDA) has estimated that the average daily intake of lead by consumption of food, water, and beverages is 25.1 μg/d for two-year-old children, 32 μg/d for adult females, and 45.2 μg/d for adult males (Environmental Protection Agency, 1986). More recent analysis indicates lead intakes have continued to decrease through 1988.

7.1.1.2 Health Effects. Health concerns regarding lead in the environment have focused upon effects upon children. Young children are more susceptible to the effects of lead exposure because of greater intakes (principally through ingestion) on a body-weight basis compared to adults, greater absorption and retention rates, and incomplete development of the blood-brain barrier, which tends to limit lead distribution to the brain in adults. Lead concerns have also arisen from the use of lead-containing solders in water distribution systems and lead-containing paints in older homes. The health effects of lead exposure have most often been correlated with total body burden. The distribution of lead in the body is ultimately dependent upon the relative affinity of tissues for the element. Steady-state intercompartmental distribution is achieved upon consistent and extended exposure, and is reflected in blood lead concentrations.

Lead is capable of producing a wide spectrum of human health effects depending upon exposure, dose, and duration. At high levels of exposure lead can produce cardiovascular effects resulting in high blood pressure; hematological effects, including anemia due to the inhibition of heme synthesis; and toxic effects upon the kidney.

Most environmental concerns have focused upon potential central nervous system effects due to lead overexposure. Symptomatic encephalopathy is most often observed in overexposed children, producing effects ranging from ataxia to stupor, coma, and convulsions. This form of lead poisoning is uncommon in North America and is seen most often at blood lead levels of greater than 100 μg/dL. Low-level lead exposure has recently become a concern because of possible subclinical effects upon the central nervous system. Neurobehavioral effects, such as hyperactivity, poor classroom learning behavior, and small decrements in I.Q. scores have been studied in association with various bloodlead levels of generally 50 μg/dL and below. A major problem with these studies is that the possible effects upon I.Q. and other neurobehavioral endpoints are very small in comparison to other competing factors such as parents' I.Q. or vocabulary, and thus become very difficult to identify (Agency for Toxic Substances and Disease Registry, 1988).

7.1.2 Mercury. Mercury presents a wide spectrum of potential biological effects depending upon whether it is in an elemental, inorganic, or organic form. The major source of mercury in the environment is natural degassing of the Earth. Anthropogenic sources include mining, smelting, and burning of fossil fuels.

7.1.2.1 Fate. Once released to the environment, mercury may undergo significant transformation. Metallic mercury is readily oxidized to divalent mercury in aquatic environments. There, it may be methylated by anaerobic bacteria. Methyl mercury may bioaccumulate, eventually finding its way into the human food chain.

7.1.2.2 Health Effects. Elemental mercury vapor is lipid soluble and readily crosses the membranes of the lung. Upon entry into the bloodstream, it has a high affinity for red blood cells as well as the high lipid content of the central nervous system. In contrast, metallic mercury when ingested is very slowly absorbed in the gastrointestinal tract and is generally thought to be of little toxicological concern via this route of exposure. Gastrointestinal absorption of inorganic salts of mercury is approximately 7% and that of methyl mercury, 90–95% (Goyer, 1986).

All forms of mercury cross the placenta, presenting a potential risk to fetal development. Concentrations of mercury in the fetus after maternal exposure to alkyl mercury are twice as high as those found in maternal tissues. The developing fetus apparently acts as a sink following maternal exposures and thus is particularly sensitive to the effects of mercury.

Mercury vapor is most often encountered in industrial settings, although some studies have focused upon the use of mercury in dental fillings and potential exposures in dental-health care workers.

Acute high-level exposure may result in corrosive effects upon the lung or gastrointestinal tract and central nervous system excitability resulting in tremors. Chronic exposure to mercury vapor affects the central nervous system. Symptoms include tremors of muscles involved in fine motor functions such as the fingers and eyelids, progressing to generalized trembling of the whole body. There may also be changes in personality and behavior, loss of memory, depression, and hallucination. Gingivitis and severe salivation are two distinctive characteristics of mercury toxicity.

Acute ingestion of inorganic salts of mercury causes ulcerations and necrosis of the gastrointestinal tract. Within 24 h, effects upon the kidney proximal tubules are observed resulting in oliguria, anuria, and uremia. Chronic low-level exposure to mercury salts may induce an immunological glomerular nephritis, again resulting in renal dysfunction.

Organic mercury is the most readily absorbed form of mercury and represents the form of greatest toxicological significance. Studies of human populations exposed to methyl mercury have produced most of the data regarding potential toxicological effects. Several significant environmental exposures have resulted from industrial and agricultural incidents. In Japan, industrial release of mercury compounds into Minamata Bay resulted in bacterial methylation and subsequent bioaccumulation of mercury in fish. Contaminated fish were caught and ingested by local populations. Fish were found to contain approximately 11 mg/kg mercury. A second incident, in Iraq, involved the ingestion of bread made from seed grain which had been treated with a methyl mercury fungicide. During the period of exposure, mercury intake was estimated to range from 80 μg/kg d to 250 μg/kg d. Principal effects observed were upon the central nervous system including degeneration of neurons of the cerebral cortex, particularly in the visual areas. Progressive clinical features included paresthesia, ataxia, dysarthria, and deafness.

Prenatal exposure from the Minamata incident produced effects upon psychomotor development, including delays in the time required to learn to walk and talk, as well as seizures. Severe *in utero* exposures may be observed to produce symptoms similar to those observed in cerebral palsy.

7.1.3. Cadmium. Cadmium is produced as a by-product of zinc and lead in mining and smelting. Its principal use is in electroplating or galvanizing. In addition, it is used as a color pigment in paints and plastics, and as cathode material in nickel-cadmium batteries.

7.1.3.1 Fate. Cadmium is found in water in the 2^+ oxidation state. Chemically, cadmium is very similar to zinc, and many of its biological activities are the result of competition for zinc binding sites in biologically active proteins.

Depending upon chemical form and solubility, cadmium can be found in water and sediment. In one study of a marine estuary system, it was observed that water cadmium concentrations varied significantly with depth during the summer months. It was determined that during stagnant periods the aerobic surface water contained relatively high concentrations of dissolved cadmium, predominantly as a soluble ion, $CdCl^+$. The anaerobic bottom layers of water contained relatively low cadmium levels, as a result of microbial reduction, to produce insoluble cadmium sulfides (Manahan, 1991).

Cadmium is known to bioconcentrate in shellfish and plants. Muscles, scallops, and oysters may contain 100 to 1000 μg/kg of cadmium. Cadmium is more readily taken up by plants than other heavy metals such as lead. Wheat and rice have been shown to concentrate cadmium when grown in highly contaminated soils. In nonpolluted areas, content may range from 0.01 to 0.1 $\mu g/g$ and in polluted areas, 0.1 to 1 $\mu g/g$. A major source of human exposure in the general population is through cigarette smoke. One cigarette contains 1 to 2 μg of cadmium. Smoking one pack or more per day can double an individual's total body burden of cadmium (International Agency for Research on Cancer, 1976).

7.1.3.2 Health Effects. The principal human health effects observed due to cadmium have resulted from occupational exposure to relatively high airborne levels of cadmium fumes. Acute exposure can produce acute chemical pneumonitis and pulmonary edema. Long-term occupational exposure has elicited chronic obstructive pulmonary disease and emphysema, and chronic renal tubular disease. Animal studies indicate that cadmium ingestion can produce nephropathy when cadmium levels in the kidney reach approximately 200 $\mu g/g$. Other evidence suggests that effects of cadmium include hypertension and effects upon both male and female reproductive systems. More recently cadmium chloride aerosols have been found to produce lung cancer in rats.

7.2 PCBs

PCBs were commercially produced in the U.S. from 1929 through 1977. By 1970 it was estimated that approximately one billion pounds of PCBs had been sold in North America. PCBs were used principally in transformers and capacitors because of dielectric properties and fire resistance. The environmental persistence of PCBs, and potential toxicity, vary with the specific isomer. The resistance to biodegradation and relative toxicity appear to increase with increasing chlorine content.

7.2.1 Fate. Current sources of PCB release are relatively minor since their manufacture has been discontinued. While potential emissions from landfills containing PCB-contaminated waste have generated public concern, the low water solubility of PCBs and relatively low vapor pressure, in addition to their affinity for organic substrates, minimize migration into the air and water. The higher chlorinated PCBs (Aroclor 1248, 1254, and 1260) are resistant to biodegradation and have been found to bioconcentrate in fish, shrimp, and oysters by factors ranging from 26 000 to 660 000.

Air concentrations for PCBs have been found to be very low. Means for remote locations have been measured at approximately 0.1 fg/m^3 ranging to 5–10 fg/m^3 in urban environments. The low water solubility of PCBs limits the significance of this potential route of exposure. A survey of finished water supplies from 113 cities, conducted by the EPA between 1975 and 1977, identified PCBs in only 4 water supplies. PCB concentrations were all detected at levels near or below the limits of detection, which are 0.1–0.2 ppb (Agency for Toxic Substances and Disease Registry, 1990).

7.2.2 Health Effects. Exposure to PCBs in the general population occurs principally through the consumption of fish, with dairy products ranking second. Average daily intake via food for adult males has been estimated at 0.56 μg.

The toxicity of PCBs has been studied extensively in animals. Acute exposures via inhalation, ingestion, and dermal contact indicate that PCBs are low in toxicity.

PCBs are highy lipophilic and tend to accumulate in fat, skin, and other high-fat-containing organs. Upon chronic low-level exposure they can be detected in blood and lipid biopsies, and attempts have been made in human studies to correlate these levels with possible effects. Long-term exposure to low levels of PCBs have produced a number of effects upon the liver, ranging from enzyme induction to hepatocellular necrosis, depending upon dose. Other studies in animals have suggested subtle effects upon the immune system and possibly fetal toxicity. Studies of human populations exposed to PCBs have indicated that high level of exposure can produce changes in liver function and effects upon cutaneous tissue exhibited as an acneiform skin lesion termed "chloracne". Aroclor 1260 has been shown to produce liver tumors in animals. Environmental and occupational epidemiological studies have failed to detect any similar activity in human populations (Silberhorn *et al.*, 1990).

7.3 Ozone

The generation of ozone in the troposphere is discussed in AIR POLLUTION: COMPONENTS, CAUSES, AND CURES. Ozone is the highly reactive triatomic form of oxygen and is a major constituent of photochemical smog. It is formed primarily via atmospheric reactions between nitrogen oxides and organic hydrocarbons. Highly reactive oxidants, such as ozone and other similarly generated smog constituents (i.e., organic peroxides, aldehydes, and free radicals), usually produce acute and immediate effects upon biological tissues.

The lung is the most important target organ impacted by reactive airborne contaminants such as ozone, and has been the subject of numerous toxicological investigations. In animals, at very high levels of exposure, ozone produces deep lung tissue destruction leading to pulmonary edema and death. Three-hour LD_{50}'s range from 20 to 50 ppm depending upon the species tested (Lippmann, 1989).

Atmospheric concentrations of ozone range from 40 ppb in rural areas to 300 ppb in heavy-smog urban areas. Animals exposed to ozone in the 300-ppb range have exhibited subtle changes in pulmonary function, including increased respiratory frequency and a decrease in tidal volume. In addition, mild changes in the morphology of lung cells have been noted.

Pulmonary-function investigations of human populations exposed to ozone in the 300-ppb range have produced both subtle positive effects and negative findings. It appears that certain individuals of the population may be more susceptible to the effects of ozone upon lung function. In addition, it is possible that over time, low-level exposures may lead to a greater ozone tolerance in exposed individuals (Folinsbee *et al.*, 1988).

7.4 Extremely Low Frequency (ELF) Magnetic Fields

The potential biological effects of extremely low-frequency magnetic fields have been the subject of a number of *in vitro* and epidemiological investigations. Extremely low-frequency magnetic fields (frequencies below 300 Hz) are generated both naturally and through man-made sources. Naturally occurring time-varying fields include those arising from solar activity and thunderstorms, and can reach intensities of 0.5×10^{-3} tesla (0.5 mT). Two manmade ELF magnetic field types of recent environmental concern are those generated by high-voltage transmission lines (intensities of approximately 15 μT) and land-based naval communication systems (approximately 14 μT) (World Health Organization, 1987). In addition, household appliances operating on a 60-Hz line voltage source can produce immediate field intensities as high as 2.5 μT. These field strengths decrease rapidly with distance, the ambient field strength in most households being less than 0.3 μT.

Time-varying magnetic fields induce electronic fields and currents in biological tissues. *In vitro* studies have shown that these fields, when of sufficient intensity, can alter cell surface binding properties, an effect which leads to a number of secondary events in biological tissues. Biological effects include altered cell surface binding of calcium, suppression of T-cell lymphocyte cytotoxicity, and inhibition of human lymphocyte activation (Tenforde, 1990).

Most biological effects have been studied utilizing peak field intensities greater than 0.5 mT with current densities induced in the exposed samples exceeding 10 mA/m^2. These levels of exposure are greater than expected from most man-made environmental sources. ELF magnetic fields have not been reported to induce mutations or cytogenetic changes.

The greatest concern regarding ELF magnetic fields has resulted from possible associations with human cancer. Several environmental and occupational epidemiology studies have suggested weak associations between the presence of an ELF magnetic field source and incidence of cancer. Other studies, similarly conducted, have failed to observe this relationship. Measurements of the magnetic field dose were not available in these studies and analysis of confounding factors, such as smoking habits, diet, etc., were not considered. It is likely that the general population is not at risk from environmental sources of ELF magnetic fields. Ground-level power-line field strengths and household appliance ambient levels of exposure are all generally less than those occurring naturally. To date, no convincing association between any human exposure to ELF magnetic fields and cancer has been established. This lack of significant effects is supported by the relatively high ELF magnetic field intensities required to produce effects *in vitro*. Because of the lack of a direct effect of ELF magnetic fields upon the genetic makeup of the cell, it is likely that if any association with cancer is observed it will be the result of an indirect effect exhibiting a practical threshold.

8. CONCLUSION

Environmental health and safety expectations will continue to evolve as the relative significance of factors potentially affecting man's health and well being change. Certain risks will continue to be reduced with medical, technological, and engineering advances, while other new risks may appear. As the major historical causes of mortality and morbidity, such as communicable disease and famine, are brought under control, we will be left to deal with less obvious environmental health and safety issues. The success of efforts to improve health and safety further will become increasingly dependent upon distinguishing between potential risks. Risk assessment will play a greater role than ever before in directing future environmental health and safety initiatives. True health and safety benefits will only be realized if the science involved in this process is separated from science policy and perceptions of risk.

ACKNOWLEDGMENT

The authors would like to thank Mrs. Claudia Kinman for her inestimable assistance in the preparation of this manuscript.

GLOSSARY

Bioactivation: Biological metabolism of a substance to a more reactive form.

Carcinogen: Agent that has been found to induce abnormal uncontrolled growth of cells (tumors) in laboratory animals or in humans.

Detoxification: Biological metabolism of a substance to a less reactive, more readily excretable form.

DNA (Deoxyribonucleic Acid): Biomolecule found in the nucleus of cells; thought to be the repository of hereditary charcteristics.

Endocytosis: Process whereby materials are taken into a cell by invagination of the plasma membrane.

Toxicology: The study of adverse effects of chemicals on biological systems.

Xenobiotic: Compounds foreign to the normal metabolism of an organism.

List of Works Cited

American Cancer Society (1988), *Cancer Facts and Figures—1988*, New York: American Cancer Society.

Ames, B. N., Magaw, R., Gold, L. S. (1987), *Science*, **236**, 271–280.

Anders, M. W. (Ed.) (1985), *Bioactivation of Foreign Compounds*, New York: Academic.

Agency for Toxic Substances and Disease Registry (1988), *The Nature and Extent of Lead Poisoning in Children in the United States: A Report to Congress*, Washington: U.S. GPO.

Agency for Toxic Substances and Disease Registry (1990), *Toxicological Profile for Selected PCB's*, Washington: U.S. GPO.

Caldwell, J., Jakoby, W.B. (Eds.) (1983), *Biological Basis of Detoxication*, New York: Academic .

Davis, D.L., Magee, B.H. (1979), *Science*, **206**, 1356–1358.

Doll, R., Peto, R. (1981), *The Causes of Cancer*, Oxford: Oxford Univ. Press.

Environmental Protection Agency (1985a), *Health Effects Testing Guidelines*, Report No. CFR 798, 50 FR 39252. Springfield, VA: National Technical Information Service.

Environmental Protection Agency (1985b), *Health Assessment Document for Dichloromethane (Methylene Chloride/Final Report) February 1985*, Report No. EPA/660/8-82/004F. Springfield, VA: National Technical Information Service.

Environmental Protection Agency (1986), *Air Quality Criteria for Lead*, Report No. EPA/600/8-83/028F. Springfield, VA: National Technical Information Service.

Environmental Protection Agency (1990), *Health Effects Assessment Summary Tables*, Report No. OERR 9200.6-303 (90-1/2). Springfield, VA: National Technical Information Service.

Folinsbee, L. J., McDonnell, W. F., Horstman, D. H. (1988), "Pulmonary Function and Symptom Responses After 6.6-Hour Exposure to 0.12 ppm Ozone With Moderate Exercise," *J. Air Waste Management Assoc.*, **38**, 28–35.

Goyer, R. A. (1986), in C. D. Klassen, M. O. Amdur, J. Doull (Eds.), *Toxicology: The Basic Science of Poisons*, 3rd Ed., New York: Macmillan.

Hayes, A. W. (Ed.) (1989), *Principles and Meth-*

ods of Toxicology, 2nd ed. New York: Raven.

International Agency for Research on Cancer (1976), *IARC Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Man*, Vol. 11, Geneva: World Health Organization.

Lippmann, M. (1989), "Health Effects of Ozone: A Critical Review," *J. Air Waste Management Assoc.*, **39**, 676–695.

Manahan, S. E. (1991), *Environmental Chemistry*, 5th ed. New York: Lewis, p. 145.

National Toxicology Program (1988), *National Toxicology Program Draft Technical Report on the Toxicology and Carcinogensis Studies of d-Limonene*, Springfield, VA: National Technical Information Service.

Silberhorn, E. M., Glauert, H. P., Robertson, L. W. (1990), "Carcinogenicity of Polyhalogenated Biphenyls: PCB's and PBB's," *CRC Rev. Toxicol.*, **20**, 440–476.

Tenforde, T. S., (1990), in B. W. Wilson, R. G. Stevens, L. E. Anderson, (Eds.), *Extremely Low Frequency Electromagnetic Fields: The Question of Cancer*, Columbus: Battelle, pp. 291–315.

World Health Organization (1987), *Magnetic Fields*, Environmental Health Criteria 69. Geneva: World Health Organization.

Further Reading

Blau, G. E. (1985), in: Neely, W. B. (Ed.), *Environmental Exposure from Chemicals*, Vols. 1 and 2. Boca Raton, FL: CRC.

Clayton, D. B., Drewski, D., Munro, I. (Eds.), (1985), *Toxicological Risk Assessment*, Vols. I and II, Boca Raton, FL: CRC.

Graham, J. D., Green, L. C., Roberts, M. J. (Eds.) (1988), In Search of Safety: Chemicals and Cancer Risk, Cambridge, MA: Harvard University Press.

Klaassen, C. D., Amdur, M. O., Doull, J. (Eds.) (1986), *Toxicology: The Basic Science of Poisons*, 3rd ed. New York: Macmillan.

Paustenbach, D. J. (Ed.) (1989), *The Risk Assessment of Environmental and Human Health Hazards: A Textbook of Case Studies*, New York: Wiley.

EQUATIONS OF STATE

MARVIN ROSS, *Lawrence Livermore National Laboratory, University of California, Livermore, California, U.S.A.*

Introduction 291
1. **Fundamental Principles** 292
1.1 Intermolecular Forces 292
1.2 Statistical Thermodynamics 293
2. **Overview of the Phase Diagram** 293
2.1 Critical Point 293
2.2 Law of Corresponding States 294
2.3 Melting of Solids 294
2.4 Superfluids 296
2.5 Plasmas—The Saha-Boltzmann Ionization Equation 296
3. **Statistical Mechanics of Fluids** 297
3.1 Virial Expansion for a One-Component Gas 297
3.2 van der Waals Equation 298
3.3 Computer Simulations—Exact Results 298
3.4 Perturbation Theory 299
3.5 Integral-Equation Theories 299
4. **The Theory of Solids** 300
4.1 The Local-Density Approximation 300
4.2 The Mie-Grüneisen Equation of State 301
4.3 Thermal Electron Properties 301
4.4 Empirical Equations of State 302
5. **Matter at Extreme Conditions** 302
5.1 Static Experiments 303
5.2 Shock-Wave Experiments 303
5.3 Pressure-Induced Transitions 304
5.4 Metallic State at Ultrahigh Density 305
5.5 Statistical Atom Models 305
5.6 Shell-Structure Effects in Compressed Solids 306
Acknowledgment 307
Glossary 307
Works Cited 308
Further Reading 309

INTRODUCTION

The equation of state (EOS) of a system describes the relationship between its thermodynamic variables. The simplest example is a gas of noninteracting particles, an ideal gas, in which the relationships between pressure P, volume V, temperature T, the number of molecules N, Boltzmann's constant k, and the energy E are given by

$$PV=NkT, \quad E=\tfrac{3}{2}\,NkT. \tag{1}$$

Real systems involve interactions between particles, and a knowledge of the EOS requires an understanding of the relationship between interatomic forces and the observed macroscopic thermodynamic properties.

In principle, the existing theoretical formalism allows the equation of state and the properties of matter to be determined once these forces are known. In practice, the barriers to *ab initio* calculations are formidable because of the difficulty of solving the many-body problem. Consequently, it has proven necessary to introduce simplifying assumptions into the governing equations. The price of simplification, however, is a loss of generality, and the resulting theory is adequate over only a limited region of the phase diagram. Liquids, solids, and gases have distinctive patterns of electronic structure and atomic ordering, leading to different theoretical approximations. Some well-known approximations are the Thomas-Fermi model of dense matter, the van der Waals model of the liquid, and the Debye model of the solid. Each of these contains certain features essential for characterizing the material properties over a limited range of conditions or of a region of the phase diagram (Eliezer *et al.*, 1986; Godwal *et al.*, 1983; Ross 1985).

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. FUNDAMENTAL PRINCIPLES

On the most fundamental level the forces between molecules are manifestations of the electromagnetic interactions between electrons and nuclei of the entire system (Torrens, 1972). For many systems the sum of these interactions may be expressed as effective intermolecular potentials. The connection between interatomic forces and EOS is provided by thermodynamics and statistical mechanics (McQuarrie, 1976; Goodstein, 1975).

1.1 Intermolecular Forces

Some typical rare-gas pair potentials are shown in Fig. 1. One of the best-known functional forms, the Lennard-Jones (LJ) potential, for a pair of atoms separated by a distance r can be written as

$$\phi(r)_{\mathrm{LJ}}=\epsilon\left[\left(\frac{r^*}{r}\right)^{12}-2\left(\frac{r^*}{r}\right)^{6}\right], \qquad (2)$$

where ϵ and r^* are the depth and position of the potential well. The attractive r^{-6} term represents the long-range dipole-dipole interaction. In particular, the r^{-12} repulsion has been found to be too stiff, and theory tells us that an exponential repulsion should be the proper form. A more useful function, particularly for examining high-pressure data, has been the exponential-six (exp-6) potential,

$$\phi(r)_{\text{exp-6}}=\frac{\epsilon}{\alpha-6}\left\{6\exp\left[\alpha\left(1-\frac{r}{r}\right)^{*}\right]-\alpha\left(\frac{r^*}{r}\right)^{6}\right\}. \qquad (3)$$

The parameter α regulates the stiffness of the interatomic repulsion. For most systems $\alpha \sim 13$–14. The LJ and exp-6 potentials are referred to as van der Waals–type potentials. Higher accuracy can be achieved by introducing additional parameters, higher multipole attractive interactions, or three-body terms (Barker, 1986).

The potential in an ionic crystal may be expressed as the sum of a Coulomb and a van der Waals potential. For example, the potential between ions i and j with charges Z_i and Z_j may be written as

$$\phi(r_{i,j})=\frac{Z_iZ_je^2}{r_{i,j}}+\phi(r_{i,j})_{\text{exp-6}}. \qquad (4)$$

Metals constitute a particular difficulty because of the inherently many-body nature introduced by itinerant electrons. For the case of simple polyvalent metals, practical pair potentials have been derived from theory (Shimoji, 1977).

Consider a system of ions placed in a neutralizing background electron gas. The electrons will tend to accumulate near the ions, screening the ion-ion interaction. A formal treatment of this problem leads to an effective pairwise interaction potential for a pair of ions,

$$\phi(r)=\frac{Z^2e^2}{r}+\frac{1}{2\pi^2}\int_0^\infty\frac{\sin qr}{qr}\,\phi_{\mathrm{BS}}(q)q^2dq. \qquad (5)$$

The first term is the bare Coulomb repulsion between ions, and the second is the screening correction due to the modulations in electron density near the ions. $\phi_{\mathrm{BS}}(q)$, referred to as the band-structure term, is a function of the electron-ion potential and dielectric constant and is in general density dependent. A typical metal ion-ion potential is shown in Fig. 2. The electrons piling up near the ion lead to a diminished charge elsewhere and to the oscillations in the potential.

Apart from the density dependence, the simple metal pair potentials are similar to those of rare gases. The potentials of both

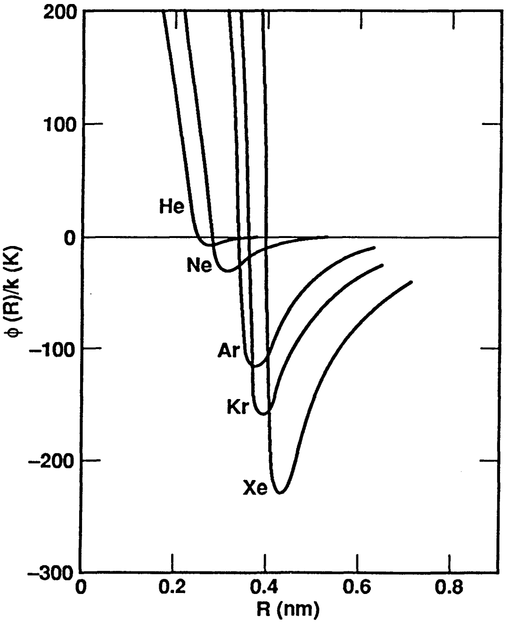

FIG. 1. Comparison of rare-gas pair potentials.

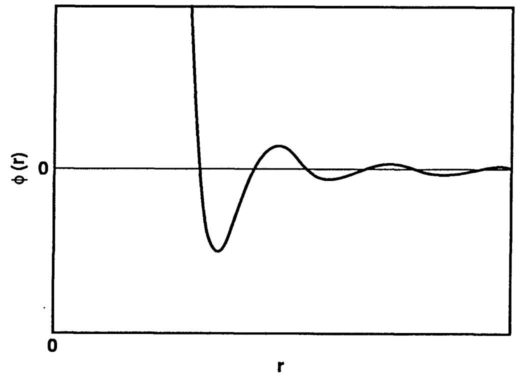

FIG. 2. A typical metallic pair potential.

have minima roughly near the average nearest-neighbor separation, and a steeply repulsive behavior at small interatomic spacings, although the metal repulsion tends to be less steep. For covalently bonded systems such as carbon or quartz, three-body potentials are required (Tersoff, 1989).

1.2 Statistical Thermodynamics

The pressure and energy for a closed isothermal system, with independent variables N,V,T, are linked to the interatomic forces through the Helmholtz free energy,

$$A(N,V,T) = -kT \ln Z(N,V,T), \tag{6}$$

where Z is the partition function

$$Z = \sum_{c} \exp[-E_c(N,V,T)/kT]. \tag{7}$$

The sum is over all possible configurations of the system in which E_c is equal to the total kinetic and potential energy for all particles of the system in that configuration. For the case of simple atoms interacting via pairwise additive potentials, we may write

$$E_c = \sum_{i} t_i + \sum_{i<j} \phi_{ij}. \tag{8}$$

Pressure and energy are obtained from the partial derivatives of Eq. (6),

$$P = -\left(\frac{\partial A}{\partial V}\right)_{T,N} = kT\left(\frac{\partial \ln Z}{\partial V}\right)_{T,N} \tag{9}$$

and

$$E = -T^2\left(\frac{\partial A/T}{\partial T}\right)_{V,N} = kT^2\left(\frac{\partial \ln Z}{\partial T}\right)_{V,N}. \tag{10}$$

The Gibbs free energy is $G = A + PV$.

2. OVERVIEW OF THE PHASE DIAGRAM

Figure 3 is a schematic phase diagram of a typical simple substance (Young, 1991). In the region of high pressures or densities and low temperatures, the substance is a solid; at higher temperature and lower pressures it is a gas, and there is a liquid region at intermediate values of P and T below the critical point. For $T > T_c$ and $P > P_c$, the substance is generally referred to as a fluid. The lines that separate these phases are called *coexistence curves*. At very high temperatures gases dissociate and ionize to become plasmas, sometimes referred to as the fourth state of matter.

2.1 Critical Point

Two points of special interest on the phase diagram are the triple point and the critical point (CP). At the triple point, solid, liquid, and gas all exist in equilibrium. The critical point is the end point of the vapor-pressure curve. A liquid cannot exist at any temperature higher than T_c or pressure higher than P_c. At the critical point, the volumes of a specific amount of liquid and gas become equal to each other, and above the critical point there is only a homogeneous fluid phase with no distinction between liquid and gas. By following a path around the critical point, a homogeneous liquid can always be transformed into a homogeneous gas without pass-

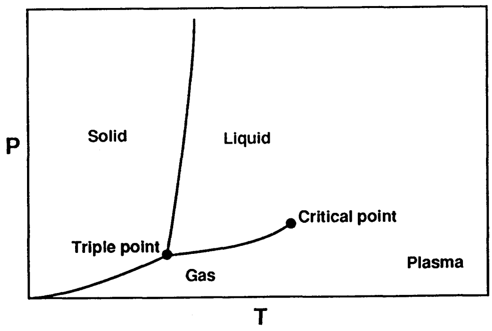

FIG. 3. A typical phase diagram.

ing through a phase transition. Figure 4 represents the isotherms of xenon near the critical temperature (Hapgood and Schneider, 1954). From the curves, we see that, at the critical temperature,

$$\left(\frac{\partial P}{\partial \rho}\right)_{T_c}=0, \quad \left(\frac{\partial^2 P}{\partial \rho^2}\right)_{T_c}=0.$$

In contrast to simple insulating fluids such as Xe, in which atoms interact through weak van der Waals forces, the interactions in liquid metals are more complicated. The change of state from liquid to gas across the vapor-pressure curve is often accompanied by a transition from metallic to nonmetallic behavior. At temperatures below the critical point, metallic properties, such as electrical conductivity, change discontinuously from vapor to liquid and continuously for temperatures above. This behavior seems to be typical for liquid metals and has been observed for Hg, Cs, Rb, and K (Freyland and Hensel, 1985). Some typical critical temperatures are, for argon, cesium, and molybdenum, respectively, 151, 1924, and (estimated) 14 300 K.

According to experiments, when the critical temperature T_c is approached, several physical quantities such as density, compressibility, and heat capacity have been observed to vary in a manner proportional to powers of $|T-T_c|$. For example, in the case of density $|\rho-\rho_c| \sim |T-T_c|^{\beta}$; or the compressibility $-(1/V)(\partial V/\partial P)_T \sim (T-T_c)^{-\gamma}$. The exponents characterizing the relationship are called critical indices (Baker, 1990).

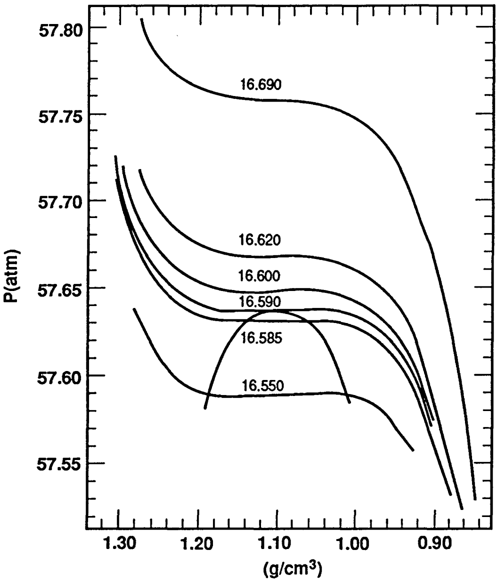

FIG. 4. Isotherms of xenon near the critical temperature. The numbers on the curves represent the temperature in degrees Celsius. The inverted parabola is the liquid-gas critical curve.

2.2 Law of Corresponding States

The ratios of P, V, and T to the critical values P_c, V_c, and T_c are called the reduced pressure, volume, and temperature. These may be written

$$P_r=P/P_c\,, \quad V_r=V/V_c\,, \quad P_r=T/T_c\,.$$

To a good approximation, especially at moderate pressures, all gases obey the same equation of state when described in terms of the reduced variables, P_r, V_r, T_r. This approximation is called the *law of corresponding states*. This is illustrated in Fig. 5 for a number of different gases, where the compressibility factor $Z_r=P_rV_r/RT_r$ is plotted at various reduced temperatures against the reduced pressure. The law of corresponding states stems from the fact that most molecules have similar functional forms for their intermolecular potentials. For example, although N_2 and n-heptane are very different in size, the spatial dependence of their interaction energies is similar and can be approximately described by Eq. (1) or (2). Thus the potential parameters ϵ and r^* can be rewritten in terms of reduced variables such as ϵ/kT_c, etc., leading to reduced expressions for thermodynamic properties.

In contrast, a comparison of the coexistence curves of the rare gas argon and the metals cesium and mercury (Fig. 6) shows a lack of correspondence when the interatomic forces are fundamentally different as in Eqs. (3) and (5). Note that even the metals cesium and mercury do not correspond, suggesting that the force laws operating in these fluids differ fundamentally. Melting data for rare gases also show that these systems obey corresponding states (Fig. 7).

2.3 Melting of Solids

From purely thermodynamic considerations, the fundamental theory of melting can be summarized as

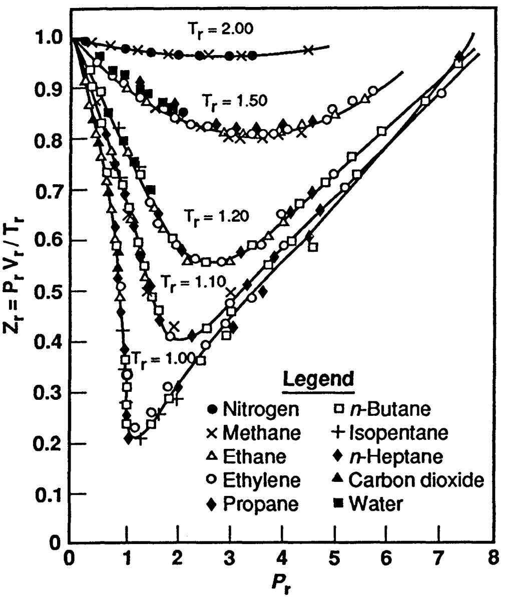

FIG. 5. Reduced compressibility factor Z_r, as a function of reduced temperature and pressure for a variety of gases (Su, 1946).

$$T_s = T_f, \quad P_s = P_f, \quad G_s(P,T) = G_f(P,T),$$

where s denotes solid and f fluid. The three conditions correspond to thermal, mechanical, and chemical equilibrium. Each of the quantities must take the same value in the solid as in the fluid phases. The traditional difficulty in calculating melting curves of real substances arises from uncertainties in statistical mechanics, particularly of the fluid free energy, and a lack of knowledge of the intermolecular forces. With an improved understanding of the forces and better statistical and quantum theories, accurate calculations of phase boundaries are now possible (Vos *et al.*, 1991).

However, for many applications where the forces are poorly known or the systems are beyond the scope of present methods, it remains necessary to rely on simpler and more empirically based methods such as the Lindemann rule.

Lindemann assumed that a solid melts when the mean square amplitude of vibrations of atoms about their equilibrium positions reaches a certain fraction of the lattice spacing, found empirically to be about 10%. He then arrived at a relationship between the melting temperature T_m, the Debye temperature Θ_D, the atomic mass M, the melting volume V_m, and a constant C that depends on the material,

$$T_m = C\Theta_D^2 (M V_m^{2/3}). \tag{11}$$

A liquid analog of the Lindemann law states that the hard-sphere liquid packing fraction (η), which is the ratio of the effective atomic hard-sphere volume v^* to the volume per atom for the system, v, has a constant value at freezing equal to about 0.46.

The Simon equation,

$$P_m = P_0 + a(T_m/T_0)^n, \tag{12}$$

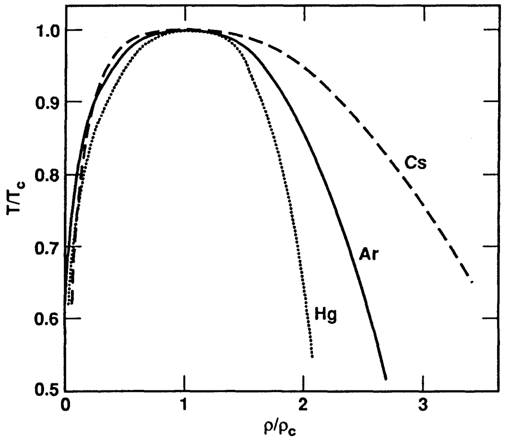

FIG. 6. Comparison of T/T_c vs ρ/ρ_c coexistence curves for argon, cesium, and mercury.

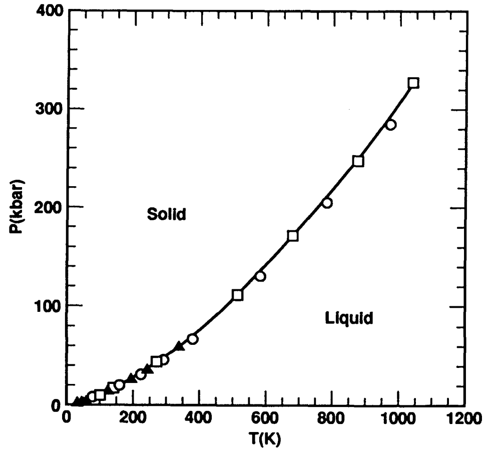

FIG. 7. Neon melting data (▲) compared with scaled helium (○) and argon (□) data (Vos *et al.*, 1992).

is a widely used equation for fitting melting data. P_0 and T_0 refer to some initial set of conditions, and a and n are constants.

Solid–liquid coexistence curves are occasionally found to have negative slopes in the P-T plane. For example, in the case of water, the liquid is denser than the solid and the melting curve of water slopes backward.

2.4 Superfluids

Helium is a simple substance with a unique phase diagram. The combination of weak attractive forces and a large zero-point motion prevents the liquid from freezing below 25 bar. Figure 8 shows the phase diagram for ^{4}He (Wilks and Betts, 1987). There are two liquid phases (I and II). Near 2.17 K the specific heat rises in a λ-like shape.

The properties of helium I above the λ temperature are not very different from those of normal classical liquid, but helium II below the λ line exhibits superfluid properties. A model to understand this behavior postulates a mixture of a superfluid and a normal fluid with thermal excitations obeying Bose statistics. The fraction of superfluid decreases with increasing temperature and only normal fluid exists above the λ line. The model also predicts that thermal excitations increase with increasing pressure, leading to the negative slope of the λ line. The phase diagram for ^{3}He shows a similar solid-liquid-vapor scheme but a superfluid does not appear until cooled below 2.6 mK, where it also exhibits a λ transition. Normal ^{3}He is governed by Fermi statistics and the superfluid properties are associated with the formation of ^{3}He pairs behaving as bosons much as in the BCS theory of electron superconductivity.

2.5 Plasmas—The Saha-Boltzmann Ionization Equation

At very high temperatures gas molecules dissociate and atoms are found in multiply ionized states (Ebeling *et al.*, 1977; Zel'dovich and Raizer, 1966). Under conditions of local thermodynamic equilibrium and negligible interatomic interactions, the ratio of the densities of the different states of ionization is given by the Saha equation. N_r, the number density of atoms in the rth stage of ionization, is related to that of the $(r+1)$th stage, N_{r+1}, by

$$\frac{N_{r+1}}{N_r}N_e=\frac{u_{r+1}}{u_r}\frac{2(2\pi mkT)^{3/2}}{h^3}\exp\left(\frac{-I_r}{kT}\right), \tag{13}$$

where N_e is the free-electron density, u_r is the internal partition function of the rth stage, I_r is the energy required to remove an electron from the ground state of the r-times ionized atom (its ionization potential), m is the electron mass, and h is Planck's constant. The Saha equation can be derived by minimizing the Gibbs free energy of a reacting mixture.

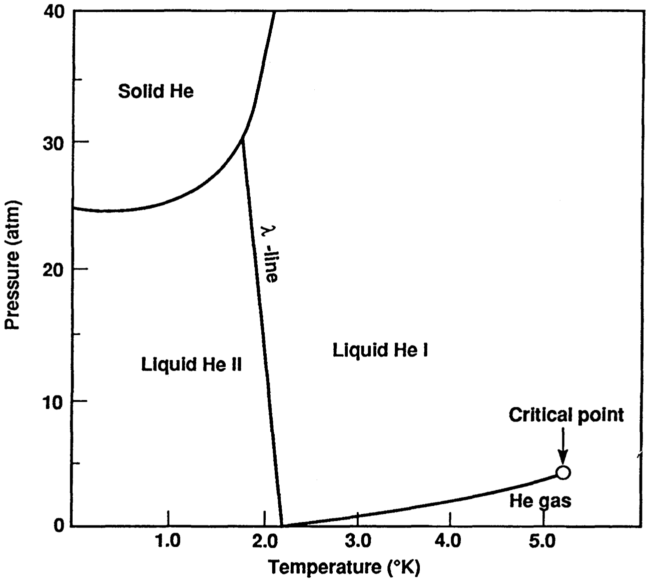

FIG. 8. The four states of ^{4}He in the *PT* plane.

The pressure, for negligible interactions, is that of a perfect gas [Eq. (1)], where N is the number of free atoms, ions, and electrons.

Ionization of atoms begins at values of kT much lower than the ionization potential. The reason here is that the statistical weight of the free-electron state is very large. With the exception of the alkali metals, which have very low ionization potentials, the first ionization potentials of the majority of atoms and molecules vary between 7 and 15 eV ($I/k \sim 80\,000$–$170\,000$ K). Ionization usually begins at thermal energies of the order of a few tenths of the ionization potential. The degree of ionization increases with temperature. When the temperature is of the order of several tens of thousands of degrees practically all of the atoms are singly ionized and the gas contains multiply ionized atoms.

The effect of small nonideal contributions from interparticle interactions is to lower the ionization potential and therefore enhance ionization. The corrections are collectively responsible for the phenomenon of "pressure ionization." For the conditions under which the Saha equation has been derived, nonideal effects only produce a slight enhancement of the degree of ionization. The Saha equation is widely used for astrophysical applications where the properties of stellar and interstellar matter are determined by the chemical composition and the degree of ionization and state of excitation (Clayton, 1983).

3. STATISTICAL MECHANICS OF FLUIDS

The equation of state of a fluid can be expressed concisely in terms of the intermolecular potential and the radial distribution function $g(r)$ (Hansen and McDonald, 1986; Allen and Tildesley, 1987). The radial distribution function measures the probability of finding a particle at a given distance from a fixed particle in the fluid. A typical form for $g(r)$ is shown in Fig. 9. In the case of a solid, $g(r)$ would be seen as a series of narrow isolated peaks corresponding to shells of neighbors. For a spherically symmetric potential the energy and pressure can be written as

$$E=\frac{3}{2}NkT+2\pi N\rho\int_0^\infty \phi(r)g(r)r^2dr \qquad (14)$$

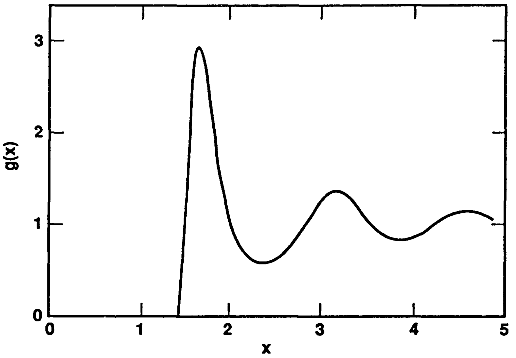

FIG. 9. A typical $g(r)$ function for a liquid.

and

$$PV=NkT-\frac{2\pi N\rho}{3}\int_0^\infty r^3\frac{d\phi(r)}{dr}g(r)dr. \qquad (15)$$

The first term corresponds to the kinetic energy of an ideal gas and the second part arises from interactions between the particles, where $\phi(r)$ is the interatomic potential and ρ is the number of particles per unit volume. When three-body and higher-order interactions are present, additional integral terms include the many-body distribution functions, $g^{(3)}(r_{12},r_{13},r_{23})$, etc. $g(r)$ can be obtained experimentally by neutron diffraction, or by theoretical methods using computer simulations or solving integral equations.

3.1 Virial Expansion for a One-Component Gas

The equation of state of a sufficiently dilute gas is $P=\rho kT$. At higher densities the equation of state can be put in the form

$$\frac{P}{kT}=\rho+B_2(T)\rho^2+B_3(T)\rho^3+\cdots, \qquad (16)$$

known as a virial expansion. The $B_n(T)$ are the virial coefficients and depend on intermolecular forces. The second virial coefficient of a classical monatomic gas is

$$B_2(T)=-\frac{1}{2}\int_0^\infty (e^{-\phi(r)/kT}-1)4\pi r^2dr, \qquad (17)$$

B_3 involves the interaction between three molecules, etc. With increasing density higher coefficients become significant. For a (hypothetical) gas of hard spheres,

$$\phi=\begin{cases}+\infty, & r<d\\ 0, & r\geqslant d\end{cases}$$

and

$$B_2=2\pi d^3/3\ ,$$

where d is the diameter or distance of closest approach of the centers of two spheres. The virial expansion for a hard-sphere fluid has been determined by Ree and Hoover (1964).

3.2 Van der Waals Equation

Many empirical equations of state are useful for representing fluid *PVT* data. One of the best-known examples is the equation of van der Waals,

$$P=\frac{NkT}{V-b}-\frac{aN^2}{V^2}\ . \tag{18}$$

An expansion in powers of $\rho=N/V$ leads to a virial series. The constants a and b are evaluated by fitting the equation to experimental *PVT* measurements, or more usually from the critical constants of the gas. At the critical point both the first and second derivatives vanish, $(\partial P/\partial V)=0=(\partial^2P/\partial V^2)$. Solving these equations for the critical constants yields

$$T_c=\frac{8a}{27bR}\ ,\quad V_c=3b,\quad P_c=\frac{a}{27b^2}\ .$$

The van der Waals equation can be understood in terms of the thermodynamic EOS,

$$P=\left(\frac{\partial P}{\partial T}\right)_V T-\left(\frac{\partial U}{\partial V}\right)_T\ . \tag{19}$$

The first term is the contribution from the kinetic pressure and the second relates to the volume dependence of the internal energy. The van der Waals equation can be derived from simple kinetic theory, viz., b is the volume occupied by the effective hard-sphere repulsions and $-a/V^2$ is the attractive pressure due to intermolecular attraction. The van der Waals equation provides a reasonably good representation of the *PVT* data of gases in the range of moderate deviations from ideality. It contains most of the essential features of a fluid and is the starting point for more complicated empirical models. The basic realism of the model has been confirmed by computer simulations and fluid perturbation theory.

3.3 Computer Simulations—Exact Results

The most important recent advances in the theory of fluids have come from computer simulations carried out by Monte Carlo (MC) and molecular dynamics (MD) methods (Ciccotti and Hoover, 1986; Allen and Tildesley, 1987) (see MOLECULAR DYNAMICS; MONTE CARLO METHODS) and the development of approximate but computationally fast and accurate models. In both methods, N molecules are confined to a box of volume V at temperature T. Computer size and speed limit N to several thousand particles. But these practical limitations are continually decreasing. To minimize surface effects, periodic boundary conditions are employed. This consists of filling three-dimensional space by replications of the original cell in which the molecules in each cell occupy the same relative positions. In the MD method, properties are simulated using numerical methods to solve Newton's equations for N particles. The MC method employs probabilistic or stochastic methods and is limited to equilibrium properties. The MD method is the more versatile and is applicable to the study of nonequilibrium systems (Hoover, 1991).

Computer simulations should be viewed as experiments producing exact data that complement laboratory experiments. For the case of strongly coupled degenerate plasmas with densities comparable to those found in high-energy astrophysics and laser-fusion-compression experiments, computer simulations are the only "experimental" data. Simulations have been used to determine the pair distribution and thermodynamic and transport properties over a wide range of conditions. One of the most useful models has been the hard-sphere system. Its isotherm is shown in Fig. 10. The pressure data for the fluid may be expressed as

$$P_{\mathrm{HS}}/kT\rho=(1+\eta+\eta^2-\eta^3)/(1-\eta)^3. \tag{20}$$

Here η is the hard-sphere packing fraction, $\eta=(\pi/6)\rho d^3$. The Helmholtz free energy A_{HS} in excess of that for a perfect gas is then

$$A_{\mathrm{HS}}/NkT=(4\eta-3\eta^2)/(1-\eta)^2\ . \tag{21}$$

Simple analytic expressions exist for exact thermodynamic results on model systems such as the Lennard-Jones (Ree, 1981) and inverse-power potentials (Young and Rogers, 1984) and one-component plasmas (Stringfellow *et*

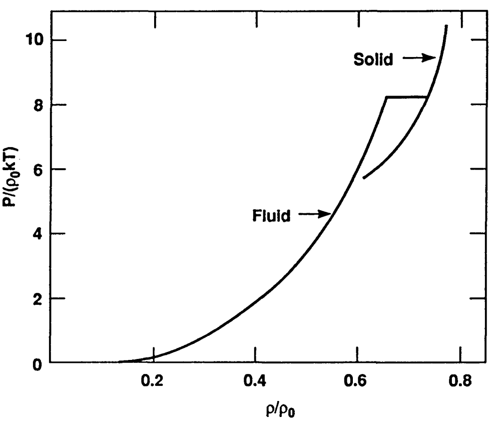

FIG. 10. Isotherm of a hard-sphere system shows the solid and fluid branches and the freezing line. The quantity ρ_0 is the close-packed density. Systems with purely repulsive forces do not have gas-liquid coexistence curves.

al., 1990). These results have been extremely valuable in the development and testing of computationally simpler approximate methods such as perturbation theories and integral equations. The books by Hafner (1987) and Shimoji (1977) contain many useful applications to real liquids.

3.4 Perturbation Theory

The idea of representing a fluid as hard spheres or charged ions moving in a background of interatomic interactions forms the basis of the van der Waals model and perturbation theory, the rationale being that if the free energy and pair distribution function of a reference system are well known, as, for example, from computer simulations, then the properties for a model system can be computed as corrections to those of the reference system. For example, if we divide the pair potential into a hard-sphere repulsion for $r<d$ and an attraction $u_1(r)$ for $r>d$, then the free energy of the system can be written

$$A=A_0-\frac{N}{2}\int_d^\infty d^3r\, g(r)u_1(r)+\cdots,$$

where A_0 is the free energy of the reference system and the second term is the correction. The "closeness" of reference and model fluids largely determines the accuracy of theory. For the case of a hard-sphere reference fluid, there is no unique way to choose the value of a diameter.

In the variational formulation, the Helmholtz free energy of an interacting system is approximated by the use of the Gibbs-Bogolyubov inequality

$$A\leqslant A_0+(\phi-\phi_0)_0. \tag{23}$$

This equation states that the free energy of the interacting system is bounded above by the free energy of a reference system (A_0) plus the difference in energy between the actual system and the reference, averaged over all configurations of the reference. For a hard-sphere reference system this equation can then be simplified to

$$\frac{A}{NkT}<\frac{A_{\rm HS}(\eta)}{NkT}+\frac{\rho}{2kT}\int_d^\infty \phi(r)g_{\rm HS}(r,\eta)4\pi r^2dr. \tag{24}$$

The correct value of A (and d) is obtained by minimizing A with respect to d. The pressure and energy are obtained from partial derivatives of A [Eqs. (9) and (10)]. Since the hard-sphere potential is unrealistically stiff in comparison with realistic intermolecular potentials, the pressures obtained are typically too high. This deficiency has been remedied by the use of inverse-power (soft-sphere) reference systems (Young and Rogers, 1984). For many simple liquid metals, the one-component plasma has provided a useful reference system (Ross *et al.*, 1981).

3.5 Integral-Equation Theories

The equations for thermodynamic properties of fluids are valid at all densities but require accurate values for the radial distribution function. Although computer simulations generate the most accurate results, integral equations are the best of the approximate methods.

Integral equations result from the combination of two relations. The first one, which is exact, is the Ornstein-Zernike relation between the pair correlation function $h(r)=g(r)-1$ and the direct correlation function $c(r)$,

$$h(r)-c(r)=\rho\int h(r')c(|r-r'|)dr'. \tag{25}$$

The second is a closure relation between these functions, examples of which are given by the

Percus-Yevick (PY) and hypernetted-chain approximations (HNC), respectively:

$$g(r)=\begin{cases}\exp[-\beta\phi(r)][1+\gamma(r)] & \text{(PY)}, \quad (26)\\ \exp[-\beta\phi(r)]+\gamma(r) & \text{(HNC)}, \quad (27)\end{cases}$$

where $\gamma(r)=h(r)-c(r)$. For purely repulsive potentials, the PY and HNC equations of state bracket the "exact" computer simulation results. This suggests that some interpolation of the two equations would provide improvements. In the scheme introduced by Rogers and Young (1984) the PY and HNC closures are interpolated by a function that depends on the interparticle distance,

$$g(r)=\exp[-\beta\phi(r)]\left(1+\frac{\exp[f(r)\gamma(r)]-1}{f(r)}\right), \tag{28}$$

where $f(r)$ is the "mixing function" ($0\leqslant f\leqslant 1$). A convenient form is the mixing function

$$f(r)=1-\exp(-\alpha r), \tag{29}$$

where α is the adjustable parameter used to achieve thermodynamic consistency. Equations (25), (28), and (29) together constitute the thermodynamically consistent approximation. In this scheme the integral equation is reduced to the PY closure at $r=0$ and HNC closure at $r=\infty$. This feature is intuitively satisfying since it is known that PY is better for treating short-range repulsive forces and HNC for long-range forces. Pressure and energy may be calculated with Eqs. (14) and (15).

4. THE THEORY OF SOLIDS

The most fundamental property of a solid is its ground-state electronic energy E_0, to which must be added a vibrational contribution (E_{vib}) arising from atomic motion and thermally excited electrons (E_{el}). Thus

$$E(V,T)=E_0(V)+E_{\text{vib}}(V,T)+E_{\text{el}}(V,T), \tag{30}$$

$$P(V,T)=P_0(V)+P_{\text{vib}}(V,T)+P_{\text{el}}(V,T). \tag{31}$$

4.1 The Local-Density Approximation

Current theoretical methods for determining E_0 use the local-density approximation (LDA) (Lundquist and March, 1983). This method simplifies the many-body fermion problem into a mean-field approximation in which each electron moves in the self-consistent average field of all the electrons and the nuclei (V_{en}). The complicated many-body exchange and correlations included in the average field are conveniently approximated by a functional $\epsilon_{\text{xc}}(\bar{r},[\rho])$ of the local electron density. The total ground-state energy is

$$E=T[\rho]+\int d\bar{r}\; V_{en}(\bar{r})\rho(\bar{r}) -\frac{e^2}{2}\int d\bar{r}'\,\frac{\rho(\bar{r})\rho(\bar{r}')}{|\bar{r}-\bar{r}'|} +\int d\bar{r}'\rho(\bar{r})\epsilon_{\text{xc}}(\bar{r},[\rho])+U_{\text{ion-ion}} \tag{32}$$

and

$$\rho(r)=\sum_j |\phi_j(r)|^2. \tag{33}$$

In simpler terms, these are, respectively,

$$E=T+U_{en}+U_{ee}+U_{\text{xc}}+U_{\text{ion-ion}}, \tag{34}$$

where T is the total electron kinetic energy and the remaining terms are the ion-electron (U_{en}), electron-electron (U_{ee}), and electron exchange and correlation (U_{xc}) energies, and $U_{\text{ion-ion}}$ is the ion-ion Coulomb energy. The energy is minimized using a wave-function gradient method ($\partial E/\partial\phi_j=0$), which leads to one-electron Schrödinger equations in which the potential field is the same for each electron:

$$\left(-\frac{\hbar^2}{2m}\nabla^2+V_{en}(r)+e^2\int dr'\,\frac{\rho(r')}{|r-r'|} +\frac{\partial U_{\text{xc}}}{\partial\rho(r)}\right)\phi_j(r)=\epsilon_j\phi_j(r), \tag{35}$$

ϕ_j and ϵ_j are the single-particle wave functions and eigenvalues. The methods for solving these equations are referred to as "electron-band theory" and are described in detail in the article ELECTRON STRUCTURE OF SOLIDS. The pressure can be obtained from the virial theorem,

$$PV=2T/3+U/3, \tag{36}$$

where U, the total potential energy, is the sum of the U terms in Eq. (34).

In general, the local-density approximation works surprisingly well for solids and even for free atoms and small molecules, using only a fraction of the computer time required by the Hartree-Fock approximation. For example, extensive sets of self-consistent field calculations (Moruzzi *et al.*, 1978; Skriver, 1984) demonstrated that electron-band calculations can accurately describe the lattice constant, cohesive energy, and compressibility of the full range of metals, given only the atomic number. It has also been capable of providing accurate predictions of many high-pressure phase transitions (McMahan, 1986; Yin and Cohen, 1982).

4.2 The Mie-Grüneisen Equation of State

The total zero-point and thermal energy of the crystal is given by

$$E_{\text{vib}} = \sum_{\alpha=1}^{3N} \left(\frac{1}{2} h\nu_\alpha + \frac{h\nu_\alpha}{e^{h\nu_\alpha/kT} - 1} \right). \qquad (37)$$

The summation is over the $3N$ normal modes (ν_α) of the crystal, N being the total number of atoms. If the interatomic potential is known, then ν_α can be calculated from lattice dynamics (Maradudin *et al.*, 1971). The Mie-Grüneisen equation of state (Girafalco, 1973)

$$E = E_0 + E_{\text{vib}}, \qquad (38)$$

$$P = -\frac{dE_0}{dV} + \frac{\gamma}{V} E_{\text{vib}}, \qquad (39)$$

is obtained by assuming that the dimensionless variable γ_α for each mode

$$\gamma_\alpha \equiv -\frac{d \ln \nu_\alpha}{d \ln V} \qquad (40)$$

can be replaced by its average over all the normal modes, $\gamma = \bar{\gamma}_\alpha$. γ is referred to as the Grüneisen constant. In the classical limit, for a system of harmonically vibrating atoms $E_{\text{vib}} = 3NkT$, and the pressure becomes

$$P = P_0 + 3\gamma NkT/V. \qquad (41)$$

In the Debye model the sum in Eq. (37) is replaced by an integral over a continuum distribution of frequencies,

$$\sum_{\alpha=1}^{3N} \rightarrow \frac{9N}{\nu_m^3} \int_0^{\nu_m} \nu^2 d\nu.$$

This leads to a modified expression for E_{vib}; however, the remaining formalism [Eqs. (36)–(39)] is unchanged. For either model γ must be determined as a function of volume. With the exception of simple metals and small molecules, this procedure is generally not feasible. Instead, approximations are introduced that bypass a knowledge of the microscopic physics and rely on macroscopic variables—the pressure-volume curve—to obtain useful approximations for γ. Several models for γ can be expressed in a single equation (Ross, 1985):

$$\gamma = -\frac{V}{2} \frac{\partial^2 (P_0 V^{2m/3})/\partial V^2}{\partial (P_0 V^{2m/3})/\partial V} + \frac{1}{3}(m-2), \qquad (42)$$

where $m=0$ is for the elastic continuum model of Slater, $m=1$ is obtained from a harmonic expansion of the static lattice energy of Dugdale and McDonald (Rice *et al.*, 1958), and $m=2$ is based on the harmonic vibration of an atom in the spherically symmetric field of neighbors. A number of useful empirical relations have been offered that bypass these models. These assume $\gamma/V = \text{const}$ or $\gamma = CV + \frac{1}{2}$, where C is a constant fitted at normal density. $\gamma = \frac{1}{2}$ is the limiting case for lattice of positive ions in a background degenerate electron gas in which both T and $V \rightarrow 0$. In the perfect-gas limit of $T \rightarrow \infty$, $\gamma = \frac{2}{3}$. A particularly large number of applications of the Mie-Grüneisen equations are to be found in Zharkov and Kalinin (1971).

4.3 Thermal Electron Properties

Contributions from thermally excited electrons may be obtained from finite-temperature self-consistent band calculations for electrons in a rigid lattice, using Fermi statistics to determine the energy-level occupation numbers (McMahan and Ross, 1977). Accordingly,

$$E_{\text{el}}(V,T) = \sum_i \epsilon_i(V) [n_i(T) - n_i(0)], \qquad (43)$$

$$P_{\text{el}}(V,T) = -\sum_i \frac{d\epsilon_i(V)}{dV} [n_i(T) - n_i(0)], \qquad (44)$$

where $\epsilon_i(V)$ are the ground-state eigenvalues obtained from the same electron-band calculations used to generate $E_0(V)$ and $P_0(V)$, and $n_i(T)$ is given by the Fermi-Dirac distribution function. For $T < 0.1T_F$, the electron thermal components E_{el} and P_{el} are obtained from the exact low-temperature expansions for the free-electron gas,

$$E_{\text{el}}(V,T)=\frac{\pi^2}{6}N(E_F)(kT)^2+\cdots \tag{45}$$

and

$$P_{\text{el}}(V,T)=\frac{\pi^2}{6}\frac{\partial N(E_F)}{\partial V}(kT)^2+\cdots, \tag{46}$$

where $N(E_F)$ is the density of electron states at the Fermi energy E_F.

4.4 Empirical Equations of State

The finite-strain model developed by Birch (1978) provides a convenient analytic fit to data that makes it possible to determine the initial bulk modulus and its pressure derivatives. In the absence of phase transitions, it has also provided a reasonably reliable method for extrapolating EOS data.

An expression for the pressure is written as an expansion in terms of a strain function f,

$$P=3f(1+2f)^{5/2}K_0(1-2af+4bf^2+\cdots), \tag{47}$$

where

$$f=\frac{1}{2}\left[\left(\frac{\rho}{\rho_0}\right)^{2/3}-1\right],$$

and with ρ and K_0 being the density and bulk modulus, respectively (subscript 0 indicates zero pressure). The higher-order terms are related to the first and second pressure derivatives of the bulk modulus (denoted by primes):

$$a=\frac{3}{4}(4-K_0'),$$

$$b=\frac{3}{8}\left[K_0K_0''+K_0'(K_0'-7)+\frac{143}{9}\right].$$

The bulk modulus K_0 and its pressure derivatives K_0' and K_0'' can be obtained from a fit of $F=P/3f(1+2f)^{5/2}$ versus f.

The Murnaghan equation is a simpler, but less accurate, function often useful for the same purposes. It is written as

$$P=\frac{K_0}{B_1}\left[\left(\frac{\rho}{\rho_0}\right)^{B_1}-1\right], \tag{48}$$

which leads to the expression for the bulk modulus

$$K=K_0\ (\rho/\rho_0)^{B_1};$$

here B_1 is a fitting parameter found empirically to be about 4.

Vinet *et al.* (1989) have proposed an EOS of the form

$$P=3K_0(V/V_0)^{-2/3}[1-(V/V_0)^{1/3}] \times\exp\{\eta[1-(V/V_0)^{1/3}]\}, \tag{49}$$

where $\eta=3/2(K_0'-1)$. They have shown that this function is capable of fitting the EOS of a wide variety of materials to high compression.

5. MATTER AT EXTREME CONDITIONS

Systematic investigations of high-pressure properties date back to the 18th century. By the latter part of the 19th century, Amagat had extended the pressure measurements on gases to about 3 kbar. P. W. Bridgman, of Harvard, ushered in the modern age of high-pressure technology. Bridgman was actively engaged in high-pressure research from 1906 to his retirement in 1954 and received a Nobel prize in 1946. During and immediately after World War II, shock-wave research appears to have developed independently in the United States and the former USSR. References to the early history can be found in the reviews by Rice *et al.* (1958) and Al'tshuler (1965). A more detailed account of shock compression experiments is to be found in this encyclopedia (see HIGH-PRESSURE EFFECTS IN SOLIDS: SHOCK WAVES).

The Universe is a natural high-pressure laboratory where existing conditions may dwarf any conditions attainable in terrestrial laboratories. The center of the Earth has pressures of about 3.5 Mbar and estimated temperatures of 6000 K. The conditions in Jupiter's center are believed to be about 100 Mbar and 24 000 K. The centers of stars and white dwarfs have pressures in excess of thousands of megabars and with temperatures of tens of millions of kelvins. In neutron stars, pressure of millions of megabars are believed to exist. In an indirect way physicists have access to a laboratory in which densities go beyond 10^{15} g/cm^3 (Eliezer and Ricci, 1991).

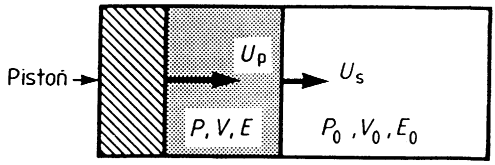

FIG. 11. Schematic of a shock wave traveling at a velocity U_s, driven by a piston moving with velocity U_p. In an experiment the piston is typically a metal plate accelerated to a constant velocity by means of a high explosive or a laser.

5.1 Static Experiments

In static experiments, the sample is squeezed between a set of opposing pistons or anvils (Jayaraman, 1983; see also HIGH-PRESSURE TECHNIQUES). Temperatures seldom exceed several hundred degrees Celsius. The highest pressures are obtained using opposed diamonds as anvils holding a sample between them. These are called diamond-anvil cells (DAC) and are currently capable of achieving maximum working pressures up around 400 GPa (4.0 Mbar) for transition metals and 200 GPa (2.0 Mbar) for rare-gas solids and hydrogen. The pressure is determined from a secondary standard such as the shift in the fluorescence frequency of ruby or the measured lattice parameter of a standard material such as platinum obtained from shock-wave studies (see Sec. 5.2).

5.2 Shock-Wave Experiments

In shock-wave experiments, a supersonic wave is passed through the material under investigation (Rice *et al.*, 1958; Al'tshuler, 1965; Zel'dovich and Raizer, 1966). A measurement of the shock wave and mass velocity are sufficient to determine the *P*, *V*, and *E* of the compressed state. Figure 11 illustrates a one-dimensional, steady, planar shock wave moving at a velocity U_s, generated by accelerating a piston from rest to some final constant velocity U_p, referred to as the mass or particle velocity. The width of the discontinuity is generally a few molecular mean-free-path lengths. The thermodynamic pressure, volume, and energy (P,V,E) behind the shock front are related to the initial properties (P_0,V_0,E_0) in front of the shock wave by the dynamic variables U_s and U_p through the Rankine-Hugoniot equations: conservation of mass,

$$V/V_0=1-U_p/U_s, \tag{50a}$$

conservation of momentum,

$$P-P_0=U_sU_p/V_0, \tag{50b}$$

and conservation of energy,

$$E-E_0=\frac{1}{2}(P+P_0)(V_0-V). \tag{50c}$$

Thus a measurement of U_s and U_p, with a knowledge of initial conditions, is sufficient to determine the *P*, *V*, and *E* of the compressed state. The resulting *P*-*V* curve is known as a Hugoniot and represents the locus of points that can be reached from an initial condition. Shock compression is typically accompanied by a large rise in temperature. Figure 12 shows some sample Hugoniots measured for various metals. Measurements have been carried out on most elements and a very large number of compounds (Marsh, 1980). Chemical explosives have been used to obtain pressures up to 1 Mbar in liquids and up to 13

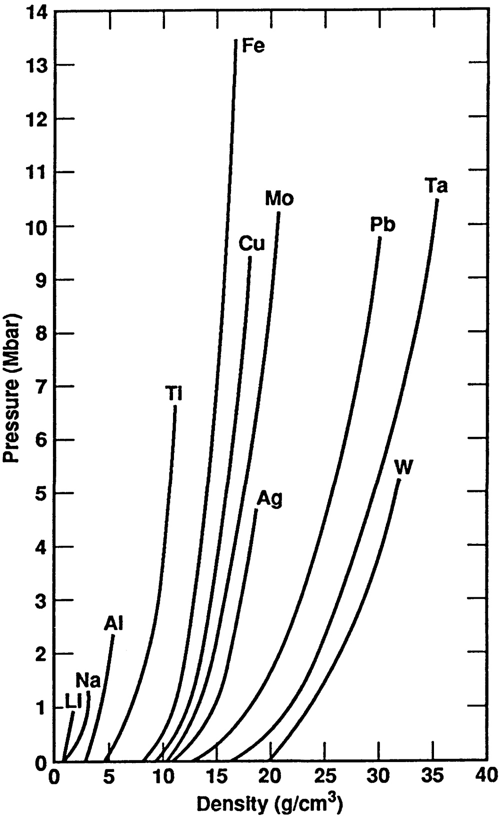

FIG. 12. Experimental Hugoniots for metals.

Mbar in metals, with accompanying temperatures of tens of thousands of kelvins. Final pressures ranging from 20 to 1000 Mbar have been reported using underground nuclear explosions (Avronin *et al.*, 1987). In inertial-confinement studies, pellets of liquid deuterium are subjected briefly to laser-driven dynamic pressures of about 1000 Mbar and temperatures in excess of 10^7 K.

Theoretical Hugoniots have been calculated for metals (Sec. 4), combining the zero-temperature pressure-volume isotherm calculated by electron-band theory with electron and ion thermal pressures as functions of volume [Eqs. (30)–(31)]. Figure 13 shows results for platinum (Holmes *et al.*, 1989). The isotherms obtained from reductions of Hugoniots are accurate to about 10% in pressure at a given density and can be used as pressure EOS standards for diamond-anvil-cell static compression experiments.

5.3 Pressure-Induced Transitions

With increasing pressure many solids undergo a series of structural transitions to more densely packed systems characterized by a greater number of nearest-neighbor atoms. At sufficiently high pressure all materials are expected to become metallic. Metallization may occur by the delocalization of valence electrons, as in the case of diatomic molecules and covalently bonded semiconductors, or with the crossing of an empty conduction band with a filled valence band as in closed-shell insulators such as the case of xenon (Reichlin *et al.*, 1988). The driving force in all insulator-to-metal transitions is electron delocalization, which lowers the kinetic energy by increasing the volume available to an electron.

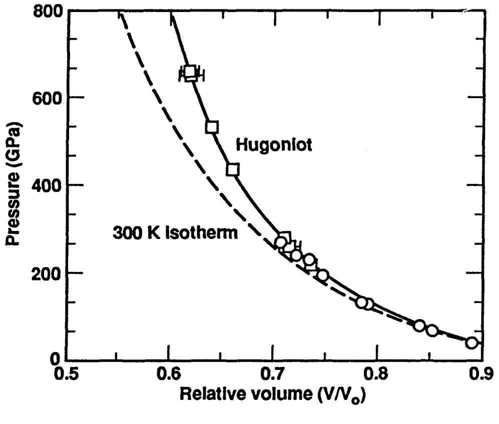

FIG. 13. Theoretical Hugoniot (solid line) for platinum and 300-K isotherm (dashed line), as compared against the Hugoniot data (squares and circles) (Holmes *et al.*, 1989).

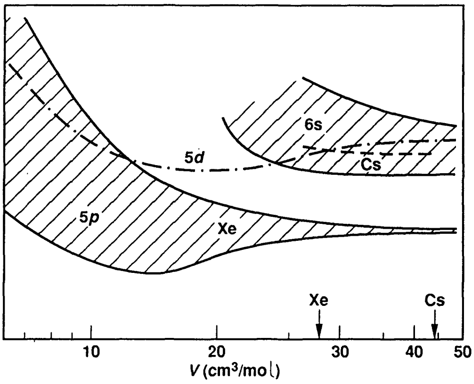

FIG. 14. 5*d* band effects in Cs and Xe. The location of the 5*p*, 5*d*, and 6*s* bands are shown schematically, plotted against atomic volume (in cm^3/mol). The Fermi energy of Cs is shown as a dashed line; the zero-pressure volumes by arrows. In Xe the Fermi energy is at the top of the 5*p* band.

Many of the interesting high-pressure structural phase transitions and irregularities in the equation of state can be understood in terms of systematic changes with compression in the electronic structure (McMahan, 1986). In most metals, valence *d* electrons are relatively localized and are closer to the nucleus than *s* electrons. Thus compression raises the energy of the *s*-state electrons, which transfer into *d* states, causing changes in the compressibility and structural stabilities. This is shown schematically in Fig. 14. Cesium, with a single conduction electron that is transferred from an *s* to *d* state, presents the most striking illustration of the rich pressure-induced effects in the phase diagram, Fig. 15. At higher pressures the *d* state, which is empty in xenon, crosses the filled 5*p* band, leading to metallization.

A feature of the trivalent rare earths is the sequence of structures hcp→Sm types →dhcp→fcc that are observed for decreasing atomic number from right to left across the series. Duthie and Pettifor (1977) have shown that it is the degree of *d*-band occupancy or the *s*-*d* transition that is responsible the for the observed crystal structure sequence without any significant participation of *f* electrons.

The high-pressure sequence of structures for the individual elements also follows this pattern. With a change in the lattice parameter, s electrons are transferred to d states and, in effect, change the properties of an element to resemble those of a neighbor.

5.4 Metallic State at Ultrahigh Density

The simplest model of a solid at extremely high density is the system of N ions in a volume V immersed in a uniform neutralizing background of electrons with density $\rho = N_e/V$ (Lundquist and March, 1983). A convenient parameter r_s is defined by $4/3\pi\, r_s^3 = 1/\rho$. Then the total energy is given by

$$E = \frac{2.21}{r_s^2} - \frac{0.916}{r_s} - 0.115 + 0.031 \ln r_s - \frac{A_M}{r_0} \text{ Ry per electron,} \tag{51}$$

where r_0 is the nearest-neighbor distance.

The first term is the electron kinetic energy. The second is the electron exchange, and the third and fourth are the Wigner approximation to the electron-gas correlation energy. Therefore the first four terms add up to the uniform electron-gas energy E_g. The final term, the Madelung energy, comes from the energy of point lattice charges in the uniform negative background. For the close-packed and body-centered lattices $A_M = 1.792$ and for the simple-cubic case $A_M = 1.760$. Structural stability calculations require more precise values. The Madelung term favors the bcc structure as the most stable. This implies the existence of a bcc phase at some very high density where electron band effects have been leveled out.

The one-component plasma (OCP) is a generalization of the electron-gas model to include the thermal motion of classical ions. The total energy is written as

$$E = E_g + \frac{3}{2} NkT + U_{\text{th}}, \tag{52}$$

where E_g is the uniform electron-gas energy, the second term is the ion kinetic energy, and third term is the ion thermal energy. Monte Carlo simulations have been used extensively to obtain the OCP solid and fluid equations of state and the location of the melting transition. DeWitt and co-workers (Stringfellow *et al.*, 1990) have reported accurate fits to U_{th} for the solid and fluid as a function of temperature and density. These fits are of the form

$$U_{\text{th}} = a\Gamma + b\Gamma^{1/4} + c + d\Gamma^{-1/4}, \tag{53}$$

where Γ, the plasma parameter, is defined as $\Gamma = (Ze)^2/a_s kT$, where $4/3\pi\, a_s^3 = N/V$. These data have been widely used to model liquid metals, dense plasmas, and ion mixtures.

5.5 Statistical Atom Models

Thomas-Fermi statistical atom models are probably the best known extreme-density theories and are often synonymous with the subject (Lundquist and March, 1983; More, 1979). These may be derived from the LDA to the total energy (Sec. 4.1) by introducing a local-density free-electron approximation for the kinetic energy as well as for the exchange and correlation. This further simplifies the numerical solution, but removes any vestige of the quantum shell structure. The description of the kinetic energy by a local approximation is only valid for sufficiently slowly varying electron densities. Accordingly, the replacement of the quantum-mechanical ki-

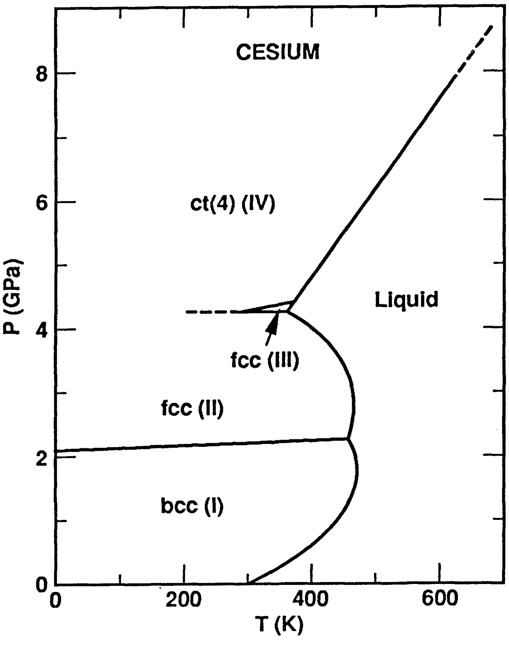

FIG. 15. Cesium phase diagram.

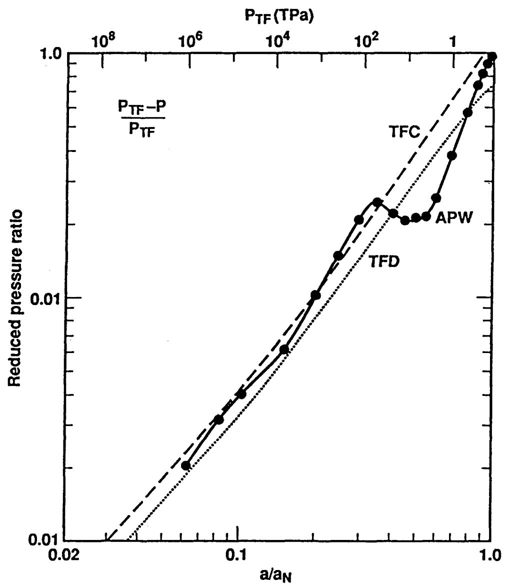

FIG. 16. Comparison of augmented-plane-wave (APW), TFC, and TFD pressures for aluminum. Pressures are given relative to P_{TF} as shown, and are plotted against reduced lattice constant. The data points indicate the actual APW calculations.

netic energy by its statistical analog should be carried out as an expansion in $\nabla\rho(r)$ and leads to

$$E=C_k \int \rho^{5/3}dV+\frac{1}{72}\int \left|\frac{\nabla\rho}{\rho}\right|^2 \rho\, dV+U_{ei}$$
$$+U_{en}+U_{xc}+U_{\text{ion-ion}}, \tag{54}$$

where

$$C_k=(3\hbar^2/10m)(3/8\pi)^{2/3}\ .$$

The gradient term has the virtue of preventing a large electron density near the nucleus. Statistical theories are generalized to finite temperatures by replacing $\rho(r)$ with the Fermi function $\rho(r,T)$. In this case the thermal contribution from $U_{\text{ion-ion}}$ can be approximated by U_{th} [Eq. (53)] using an appropriate value of Z. The requirement that E be minimized with respect to variations in the electron density is subject to the condition that the number of electrons N remains fixed and leads to an integral equation from which the local charge density may be obtained:

$$\frac{5}{3}C_k\,\rho(r)^{2/3}+\frac{1}{72}\left|\frac{\nabla\rho}{\rho}\right|^2\ \frac{1}{36}\frac{\nabla\rho}{\rho}+V_{en}(\bar{r})$$
$$+e^2\int d\bar{r}'\,\frac{\rho(\bar{r}')}{|\bar{r}-\bar{r}'|}+\frac{d(U_{xc})}{d\rho}=\mu. \tag{55}$$

This is the statistical-model analog of the one-electron Schrödinger equation [Eq. (35)] where μ, the chemical potential, replaces the one-electron eigenvalue. The pressure may be determined from the virial theorem [Eq. (36)].

The neglect of the gradient term and correlation energy leads to the Thomas-Fermi-Dirac (TFD) theory, and the additional neglect of exchange leads to the Thomas-Fermi (TF) model. The great virtue of TF is that all its properties scale with Z, and a single set of calculations at all densities for one Z suffices for all Z. This feature is not present in the other statistical models. The TFC theory is an attempt to include exchange in TF while retaining its advantageous scaling features by using the TF charge densities (ρ_{TF}) to approximate ρ in TFD.

5.6 Shell-Structure Effects in Compressed Solids

The major disadvantage of the statistical models is that although they often predict average properties reasonably well at very high pressure, they fail on details because they ignore the electron shell structure of the atoms. A more satisfactory treatment of electrons in solids rigorously including shell structure is the self-consistent LDA method.

In Fig. 16, the LDA pressure-volume isotherm as calculated by LDA methods is compared with predictions of the TFD and TFC statistical models. All results are shown relative to the larger Thomas-Fermi pressures according to the reduced pressure ratio ($P_{TF}-P)/P_{TF}$. The calculations show two well-defined deviations in pressure from predictions of the statistical models for the aluminum isotherms. These occur as first the L- and then the K-electron shells are pressure ionized. The pressure deviation due to shell effects in the first case corresponds to an oscillation of about 13% amplitude about the statistical-model pressures, and occurs over the range 5 to 75 TPa.

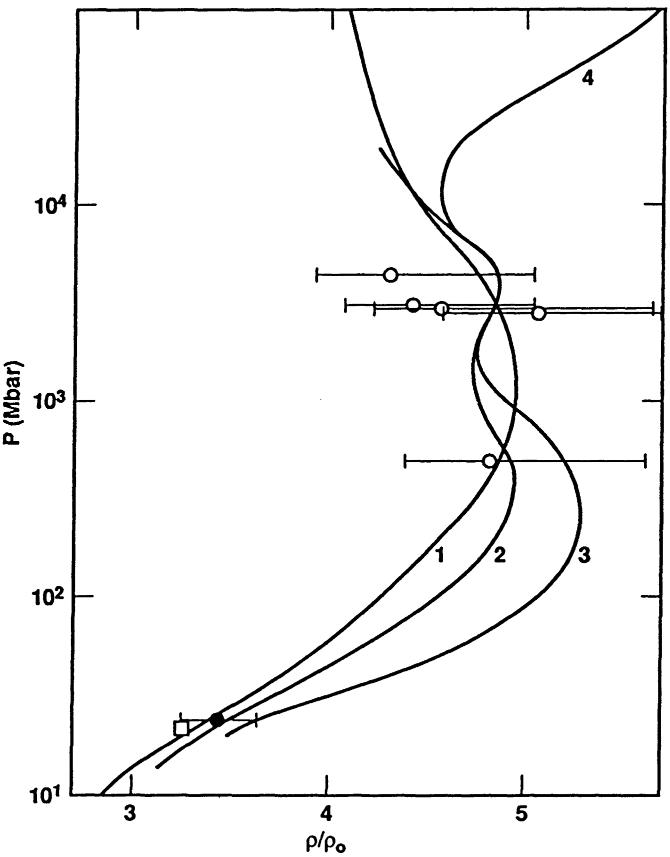

FIG. 17. Experimental data (of pressure versus compression ρ/ρ_0) for shock-compressed Al compared to a Thomas-Fermi calculation (curve 1) and two LDA-type models (curves 2 and 3). Note the oscillations in the latter. Thermal radiation effects (curve 4) appear above 10^4 Mbar. (Vladimirov *et al.*, 1984.)

Underground nuclear explosions have been used to generate pressures of up to 10^3 Mbar in shock experiments with metals an attempt to discover possible shell effects (Avrorin *et al.*, 1987). Figure 17 shows some experimental data for aluminum. They are compared to theoretical calculations made with a temperature-dependent Thomas-Fermi and local-density model that includes electron structure. Although the experiments are not sufficiently accurate to distinguish between theories, the predicted effect of shell structure on the calculated Hugoniot is seen clearly.

ACKNOWLEDGMENT

This work was performed under the auspices of the U. S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

GLOSSARY

Bose-Einstein Statistics: A counting method used in quantum statistical mechanics that allows an unlimited number of particles (bosons) to occupy a single state. This applies to phonons and particles with zero and integer-valued spins.

Corresponding States: The scaling of thermodynamic variables of liquids by taking the ratio of their values in the liquid to those at their critical point.

Diamond-Anvil Cell (DAC): A technique that employs a set of opposing diamonds to obtain measurements of the EOS and optical properties of materials at high pressure.

Electron-Band Theory: A generic term for the collection of methods used in solving the quantum-mechanical (Schrödinger) equations for electrons in a lattice.

Fermi-Dirac Statistics: A counting method used in quantum statistical mechanics that allows no more than one particle (fermion) to occupy a single state. This applies to electrons and particles with half-odd-integer–valued spins.

Grüneisen Constant: Ratio of the logarithmic derivative of the Debye temperature to that of the volume. Used to calculate the thermal properties of a solid.

Hugoniot: Locus of pressure-volume points that can be reached by a shock compression experiment starting from a given initial condition.

Lattice Dynamics: Refers to the vibrations of atoms in a solid about their equilibrium positions and the effect of these vibrations on the properties of the solid.

Local-Density Approximation: The simplification made in atomic and molecular physics for replacing the complicated electron exchange and correlation energies by simple free-electron-gas expressions.

Mie-Grüneisen Equation: An equation of state of the solid that employs the Grüneisen constant to calculate the thermal contribution to the pressure.

Molecular Dynamics: Computer simulation of molecular properties using numerical integration methods to solve Newton's equations for N interacting particles.

Monte Carlo Method: Computer simulation of molecular properties using stochastic methods.

Phase Diagram: A convenient method of representing the stability regions of solid, liquid, gas, and plasma under the various conditions of temperature and pressure.

Plasma: An electrically conducting gas of ionized particles.

Pressure Units: Primary units are force/area or dyn/cm. Literature values are commonly quoted in terms of MPa (10^5 dyn), bar (10^6 dyn), kbar (10^3 bar), GPa (10^{11} dyn or 10 kbar).

van der Waals Equation: The first quantitatively successful equation of state that described the liquid-vapor branches of the phase diagram.

van der Waals Forces: Weak long-range interactions between neutral molecules.

List of Works Cited

Al'tshuler, L. V. (1965) *Sov. Phys.—Usp.***8**, 52–91.

Allen, M. P., Tildesley, D. J. (1987), *Computer Simulation of Liquids*, Oxford: Clarendon.

Avrorin, E. N., Vodolaga, B. K., Voloshin, N. P., Kovalenko, G. V., Kuropatenko, V. F., Simonenko, V. A., Chernovolyuk, B. T. (1987), *Sov. Phys.—JETP* **66**, 347–354.

Baker, G. A., Jr. (1990), *Quantitative Theory of Critical Phenomena*, New York: Academic.

Barker, J. A. (1986), *Phys. Rev. Lett.* **57**, 230–233.

Birch, F. (1978), *J. Geophys. Res.* **83**, 1257–1268.

Ciccotti, G., Hoover, W. G. (1986), *Molecular-Dynamics Simulation of Statistical-Mechanical Systems*, Amsterdam: North-Holland.

Clayton, D. D. (1983), *Principles of Stellar Evolution and Nucleosynthesis*, Chicago: Univ. of Chicago Press.

Duthie, J. C., Pettifor, D. G. (1977), *Phys. Rev. Lett.* **10**, 564–567.

Ebeling, W., Kraeft, W. D., Kremp, D. (1977), *Theory of Bound States and Ionization Equilibrium in Plasmas and Solids*, Berlin: Akademie.

Eliezer, S., Ricci, R. A. (1991), *High Pressure Equations of State: Theory and Applications*, Amsterdam: North-Holland.

Eliezer, S., Ghatak, A., Hora, H. (1986), *An Introduction to Equations of State: Theory and Applications*, Cambridge: Cambridge University Press.

Freyland, W., Hensel, F. (1985), in: P. P. Edwards, C. N. R. Rao (Eds.), *The Metallic and Nonmetallic States of Matter*, London: Taylor and Francis, pp. 93–117.

Girafalco, L. A. (1973), *Statistical Physics of Materials*, New York: Wiley.

Godwal, B. K., Sikka, S. K., Chidambaram, R. (1983), *Phys. Rep.* **102**, 121–197.

Goodstein, D. L. (1975), *States of Matter*, Englewood Cliffs, NJ: Prentice-Hall.

Hafner, J. (1987), *From Hamiltonians to Phase Diagrams*, Berlin: Springer.

Hansen, J. P., McDonald, I. R. (1986), *Theory of Simple Liquids*, London: Academic.

Hapgood, H. W., Schneider, G. W. (1954), *Can. J. Chem.* **32**, 98–112.

Holmes, N. C., Moriarty, J. A., Gathers, G. R., Nellis, W. J. (1989), *J. Appl. Phys.* **66**, 2962–2967.

Hoover, W. G. (1991), *Computational Statistical Mechanics*, New York: Elsevier.

Jayaraman, A. (1983), *Rev. Mod. Phys.* **55**, 65–108.

Lundquist, S., March, N. H. (1983), *Theory of the Inhomogeneous Electron Gas*, New York: Plenum.

Maradudin, A. A., Montroll, E. W., Weiss, G. H., Ipatova, I. P. (1971), *Theory of Lattice Dynamics in the Harmonic Approximation*, New York: Academic.

Marsh, S. P. (1980), *LASL Shock Data*, Berkeley: University of California Press.

McMahan, A. K. (1986), *Physica* **139&140B**, 31–41.

McMahan, A. K., Ross, M. (1977), *Phys. Rev. B* **15**, 718–725.

McQuarrie, D. A. (1976), *Statistical Mechanics*, New York: Harper & Row.

More, R. M. (1979), *Phys. Rev. A* **19**, 1234–1246.

Moruzzi, V. L., Janak, J. F., M. Williams, A. R., (1978), *Calculated Electronic Properties of Metals*, Oxford: Pergamon.

Ree, F. H. (1981), *J. Chem. Phys.* **73**, 5401–5403.

Ree, F. H., Hoover, W. G. (1964), *J. Chem. Phys.* **40**, 939–950.

Reichlin, R., Brister, K. E., McMahan, A. K., Ross, M., Martin, S., Vohra, Y. K., Ruoff A. L., (1988), *Phys. Rev. Lett.* **62**, 669–672.

Rice, M. H., McQueen, R. G., Walsh, J. M. (1958), *Solid State Phys.* **6**, 1–63.

Rogers, F. J., Young, D. A. (1984), *Phys. Rev. A* **30**, 999–1007 (1984).

Ross, M. (1985), *Rep. Prog. Phys.* **48**, 1–52.

Ross, M., DeWitt, H. R., Hubbard, W. B. (1981), *Phys. Rev. B.* **24**, 1016–1020.

Shimoji, M. (1977), *Liquid Metals*, London: Academic.

Skriver, H. L. (1984), *The LMTO Method*, Berlin: Springer.

Stringfellow, G. S., DeWitt, H. E., Slattery, W. L. (1990), *Phys. Rev. A* **41**, 1105–1111.

Su, Gouq-Jen (1946), *Ind. Eng. Chem.* **38**, 803–806.

Tersoff, J. (1989), *Phys. Rev. B* **39**, 5566–5568.

Torrens, I. M. (1972), *Interatomic Potentials*, New York: Academic.

Vinet, P., Rose, J., Smith, J. R., Ferrante, J., (1989), *J. Phys. C* **1**, 1941–1963.

Vladimirov, A. S., Voloshin, N. P., Nogin, V. N., Petrovtsev, A. V., Simonenko, V. A. (1984), *Sov. Phys.—JETP Lett.* **39**, 82–85.

Vos, W. L., Schouten, J. A., Young, D. A., Ross, M. (1991), *J. Chem. Phys.* **94**, 3835–3838.

Wilks, J., Betts, D. S. (1987), *An Introduction to Liquid Helium*, Oxford: Clarendon.

Yin, M. T., Cohen, M. L. (1982), *Phys. Rev. B* **26**, 5668–5687.

Young, D. A. (1991), *Phase Diagrams of the Elements*, Berkeley: Univ. of California Press.

Young, D. A., Rogers, F. J. (1984), *J. Chem. Phys.* **81**, 2789–2793.

Zel'dovich, Ya. B., Raizer, Yu. P. (1966), *Physics of Shock Waves and High Temperature Hydrodynamic Phenomena*, Vols. 1 and 2, New York: Academic.

Zharkov, V. N., Kalinin, V. A. (1971), translated by A. Tybulewicz, *Equations of State for Solids at High Pressures and Temperatures*, New York: Consultants Bureau.

Further Reading

Ebeling, W., Kraeft, W. D., Kremp, D. (1977), *Theory of Bound States and Ionization Equilibrium in Plasmas and Solids*, Berlin: Akademie.

Eliezer, S., Ghatak, A., Hora, H. (1986), *An Introduction to Equations of State: Theory and Applications*, Cambridge: Cambridge University Press.

Eliezer, S., Ricci, R. A. (1991), *High Pressure Equations of State: Theory and Applications*, Amsterdam: North-Holland.

Girafalco, L. A. (1973), *Statistical Physics of Materials*, New York: Wiley.

Godwal, B. K., Sikka, S. K., Chidambaram, R. (1983), *Phys. Rep.* **102**, 121–197.

Goodstein, D. L. (1975), *States of Matter*, Englewood Cliffs, NJ: Prentice-Hall.

Hansen, J. P., McDonald, I. R. (1986), *Theory of Simple Liquids*, London: Academic.

Hafner, J. (1987), *From Hamiltonians to Phase Diagrams*, Berlin: Springer.

Jayaraman, A. (1983), *Rev. Mod. Phys.* **55**, 65–108.

Lundquist, S., March, N. H. (1983), *Theory of the Inhomogeneous Electron Gas*, New York: Plenum.

Torrens, I. M. (1972), *Interatomic Potentials*, New York: Academic.

Zel'dovich, Ya. B., Raizer, Yu. P. (1996), *Physics of Shock Waves and High Temperature Hydrodynamic Phenomena*, Vols. 1 and 2, New York: Academic.

EXCITONS

ROBERT S. KNOX, *Department of Physics and Astronomy, University of Rochester, Rochester, New York, U.S.A.*
WAYNE H. KNOX, *AT&T Bell Laboratories, Holmdel, New Jersey, U.S.A.*

Introduction 311
1. **General Principles** 312
1.1 Excitons as Excited States 312
1.1.1 Frenkel and Wannier-Mott Excitons 312
1.1.2 Intermediate Cases 313
1.2 Excitons as Bound Charge Carriers 313
1.3 Role in Determining Material Properties 313
1.3.1 Optical Properties 314
1.3.1.1 Absorption 314
1.3.1.2 Emission and Nonlinear Effects 315
1.3.2 Other Properties 315
1.4 Other Topics 315
1.4.1 Exciton Trapping 315
1.4.2 Exciton-Exciton Interactions 315
2. **Excitons in Semiconductors** 316
2.1 Crystals: General Remarks 316
2.2 Quantum Wells, Wires, and Dots 316
2.2.1 Electron and Hole States in Two and Three Dimensions 316
2.2.2 Modification of Exciton Properties in Two Dimensions. 317
2.2.3 Effects of Applied Electric Fields 317
2.2.4 Device Applications of Quantum-Well Modulators 318
2.2.5 Lower-Dimensional Systems 318
2.2.6 High-Density Effects 318
2.3 Amorphous Semiconductors 318
3. **Excitons in Other Materials** 319
3.1 Molecular Crystals 319
3.1.1 General Characteristics 319
3.1.2 Triplet States 319
3.1.3 Photon Channeling 319
3.2 Other Large–Band-Gap Crystals 320
3.2.1 Rare-Gas Crystals 320
3.2.2 Ionic Crystals 320
3.3 Biomaterials 320
3.3.1 Photosynthesis 320
3.3.1.1 Chlorophyll Aggregates 321
3.3.1.2 A Dimer State as a "Miniexciton" 322
3.3.2 DNA Damage 322
4. **Summary** 322
Glossary 322
Works Cited 324
Further Reading 325

INTRODUCTION

An electronic excitation capable of motion in an atomically dense medium is called an exciton. Frenkel (1931) introduced the concept during an investigation into how absorbed light is turned into heat in a crystal. Because of the translational symmetry of a crystal, the primary excitation resulting from absorption of a photon has no reason to reside at a particular site. Since the eigenstates of a crystal are wavelike, Frenkel argued, the excited state produced should act as a traveling wave packet whose energy is essentially that of the photon absorbed, until it decays into vibrational quanta (heat).

Early research on excitons concentrated on their role in determining the optical properties of crystals. It was concluded that strong exciton-lattice coupling led to broad Gaussian absorption lines; weak coupling, to narrow Lorentzian lines. Absorption features quite distinct from those associated with the production of free electrons and holes were predicted and observed. During the 1950s and 1960s, motion of the exciton was firmly established, either as diffusive, which is the case at virtually any coupling strength, or as ballistic

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

(describable by a velocity), which occurs in certain ideal cases involving weak coupling. Exciton diffusion is observed largely by experiments on doped crystals, where the dopant acts as a trap and detector (Simpson, 1956). In the ballistic case, the exciton velocity is measured by exploiting a subtle magnetic effect in very narrow spectral lines, as first demonstrated in cadmium sulfide at low temperature (Thomas and Hopfield, 1961).

The decay of the exciton usually results in some product other than pure heat, although a certain amount of heat is inevitably produced. Some possible outcomes are light emitted at a longer wavelength, free electron-hole pairs, bound electrons and holes in the form of stored chemical energy, and fission into two or more excitons. The exciton, whether mobile or not, is thus a very important first intermediate in many of the standard optically induced processes in condensed matter.

Currently the exciton plays many useful roles. The optical properties it contributes are a source of much detail about energy-band structures of solids. Its relatively enhanced absorption strength and nonlinear behavior in quantum wells are important for engineering the optoelectronic properties of materials. Exciton transport is used to transport light effectively through submicron passages. There is a good understanding of how the exciton participates in the initial processes of photosynthesis.

1. GENERAL PRINCIPLES

1.1 Excitons as Excited States

Atoms, molecules, and ordinary condensed materials consist of very well characterized arrangements of nuclei and electrons. When a system is stably bound in its lowest electronic energy state, the electrons occupy certain orbitals. Electronic excitation of the system is well described as the placement of one electron into a different, usually more extensive, orbital.

1.1.1 Frenkel and Wannier-Mott Excitons. Figure 1 shows two limiting cases of the excitation process in condensed material, which for the moment we will assume to be a regular solid. In Fig. 1(a) the erstwhile bound electron, whose new orbital is shown as a stippled region, closely circles the core of an atom from which it has been lifted by the excitation energy. The "+" is meant to indicate two things: that the remaining atomic core is positively charged, having started out to be neutral, and that the electron previously resided in a well-localized position. Absent the remaining atoms, one would have merely an excited atom. It was this limiting case of excitation that Frenkel considered. Let this state of the system be denoted $\Phi_{\mathbf{R}}^{ex}$ and let **R**, the position vector of the excitation, represent any site of the system. In a regular lattice one can show, in analogy with the case of electron and vibrational states, that a good state of the system is $\Phi_{\mathbf{K}}=\Sigma_{\mathbf{R}}\exp(i\mathbf{K}\cdot\mathbf{R})\Phi_{\mathbf{R}}^{ex}$. Here **K** is a wave vector. (Because the lattice is regular, with each site being essentially identical to any other site, the different $\Phi_{\mathbf{R}}^{ex}$ are also essentially identical, except for the shift between cells.) Such a state is more nearly an energy eigenstate and is suited to describe particles or quasiparticles that move. In Frenkel's case the motion is that of a packet of energy, approximately the amount required to produce the related atomic excited state. No net charge is transported.

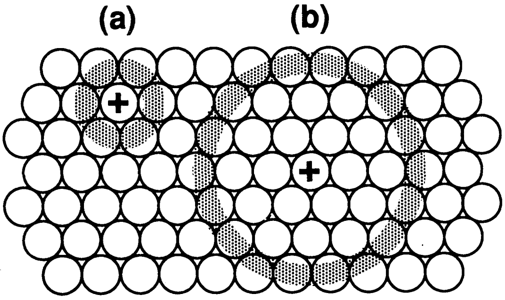

FIG. 1. Sketch of (a) the tightly bound (Frenkel) limiting case of the exciton and (b) the weakly bound (Wannier-Mott) case. In each case the hole, represented by "+," and the electron, shown in terms of a probability cloud around the hole, may move through the lattice as a unit.

Wannier (1937) and Mott (1938) observed that similar motion and energy transport would occur in the limit of a much more extended state, as in Fig. 1(b). There we see the electron quite distant from its associated positive charge. The pair form a quasi-hydrogen atom and Wannier showed that a hydrogenic series of energy levels would be obtained. Their excitation energies E_n would be given by

$$E_n = E_G - R_{ex}/n^2,$$

where E_G is the gap between the bands from which the hole and electron are taken, n is a positive integer, and R_{ex} is the "exciton Rydberg." It is precisely analogous to the hydrogenic Rydberg $R(=13.6\ \text{eV})$, differing from it only by factors resulting from the presence of a medium and the differing masses of the particles involved. The Coulomb attraction between the electron and hole is greatly reduced from that found in atomic hydrogen, because of the polarizable material intervening between them. The Wannier-Mott result was

$$R_{ex} = (13.6\ \text{eV})(1/\epsilon^2)(m^*/m),$$

where ϵ is the dielectric constant, m is the free-electron mass (the mass of the proton in the hydrogen atom is neglected), and $m^* = m_e^* m_h^*/(m_e^* + m_h^*)$ is the reduced mass made up from the effective masses of the electron and hole.

The Wannier-Mott exciton is observed extensively in semiconductors where the dielectric constant is indeed quite large; binding energies (R_{ex}) of tens of meV are common. Frenkel excitons are observed in molecular crystals and some ionic crystals, where binding forces are such that the molecules and ions retain fairly strong identities.

1.1.2 Intermediate Cases. Although we refer to Frenkel or Wannier-Mott excitons, the two are fundamentally the same and should be considered limiting cases of small and large radii, respectively (Knox, 1963, Sec. 5). In many materials, as illustrated especially well in the alkali halide crystals, excitons whose radii are of the order of one or two lattice constants occur. In this case the hole resides on the halide$^-$ sublattice and the electron mainly on the alkali$^+$ sublattice. Since halide$^-$ as a free ion has no bound excited state, there is no "atomic limit" or true Frenkel exciton, but in these materials the radius of the lowest-energy exciton can be relatively small nonetheless. In rare-gas crystals the two limiting cases of exciton clearly coexist. The lowest Wannier-Mott state ($n=1$) has such a small radius that it acts essentially as an excited state of a rare-gas atom; the higher states ($n=2,3,\ldots$) are so extended that they have a hydrogenlike series of energy levels.

1.2 Excitons as Bound Charge Carriers

Many crystal properties, especially the electrical, are best discussed in terms of their energy-band structure. In a typical semiconductor, electrons move relatively freely in conduction bands, holes in valence bands (Fig. 2). The band picture assumes completely independent particles and admits no excitons, which depend for their very existence on the Coulomb interaction between electrons and holes. In the absence of this interaction the lowest electronic excited state of the crystal occurs when an electron is lifted from a filled valence band to the conduction band. The energy required is E_G and there may or may not be a change in the wave vector of the electron [Figs. 2(a), 2(b)].

The band picture must be used with great care when electron-hole interactions become significant—for example, when optical properties at photon energies near E_G are involved or when electron-hole recombination occurs in a stepwise process, the first step being creation of an exciton by mutual electron-hole capture. The exciton is not a one-particle state, but it has a wave-vector–dependent energy that can be superimposed on a one-particle band diagram. In Fig. 2 exciton bands are shown as light dashed lines, and indeed excitonic optical features are observed at energies just below band-to-band optical features in these cases. One problem with generalizing this picture is that exciton bands are not necessarily closely associated with a nearby conduction band. For example, if the valence-band maximum were at the right edge of the zone in each of the two cases in Fig. 2, instead of at the center, the exciton would be indirect in (a) and direct in (b).

1.3 Role in Determining Material Properties

Excitons play a primary role in optical properties and a secondary role in others. One can appreciate the former by reference to the atom: if there were no "exciton effect," every atom would have only a ground state and a continuum of states, and therefore no line spectrum. However, merely taking account of the exciton effect, i.e., not ignoring the interaction between electrons and holes, does not guarantee the appearance of line spectra in solids. First, a band gap must exist in order to

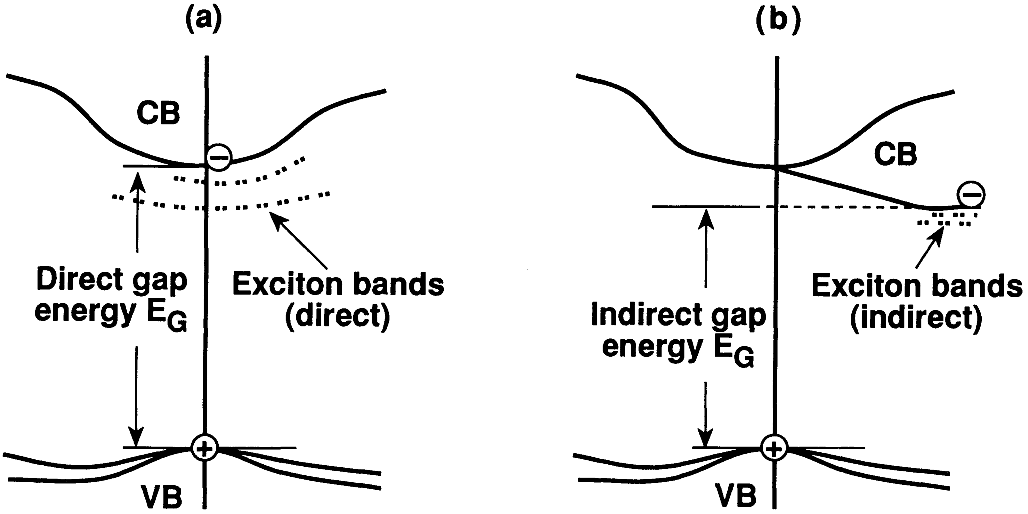

FIG. 2. Band structures in materials typical of (a) direct gaps and (b) indirect gaps. Energy is plotted vertically, wave-vector components (*k*) horizontally. Positive and negative *k* correspond to different directions in *k* space to convey the complexity of band structures. Dotted lines beneath the conduction bands (CB) show energies of excitons, relative to a zero of energy taken at the top of the valence bands (VB). The electron and hole symbols indicate the regions of one-particle space from which the exciton bands are constructed.

permit observation of bound states. Then selection rules, which are atomiclike only for direct-gap excitations, may greatly restrict the formation of lines or change them into continua because of the involvement of phonons.

1.3.1 Optical Properties. The spectral range of interest here spans the visible spectrum and parts of the infrared and ultraviolet. This corresponds to band gaps of 0.5 to 5 eV, typical of semiconductors and insulators, although excitonic phenomena also influence optical properties at even smaller bandgaps.

1.3.1.1 Absorption Figures 3(a)–3(d) show four typical and widely divergent cases that are useful for discussing general principles.

As in the case of any absorption process, the cross section (σ) at photon energy E may be separated into two factors,

$$\sigma(E) = |M|^2 f(E),$$

where $|M|^2$ is an overall strength factor and $f(E)$ is a shape factor. $|M|^2$ depends on the exciton size, on the symmetry of the electronic band states involved, and on exciton-phonon coupling strengths when phonons are required to allow the transition. $f(E)$ is essentially a density of states, weighted by factors depending on the "directness" and "allowedness" of the transitions involved. For excitons constructed from bands resembling those in Fig. 2(a), $f(E)$ can be dominated by strong distinct lines, such as those appearing in the absorption spectra of solids in the Frenkel limit [RbBr, Fig. 3(a)]. In the case of Cu_2O, Fig. 3(b), directness is seen in the sharp lines but a phonon-assisted continuum is seen because the lines have relatively weak transition probabilities compared to the phonon-assisted background. Indeed, the $n=1$ line is so weak as to be virtually invisible in this spectrum. Cu_2O is an example of a solid in which the conduction and valence bands have the same parity, so that the envelope must be *p*-like [not spherical, as in Fig. 1(b)], drastically reducing the factor $|M|^2$.

In GaAs [Fig. 3(c)], an exciton series exists but suffers considerable phonon broadening and blends into the electron-hole continuum. Finally, in AgBr, we see an "indirect" exciton spectrum in which $f(E)$ consists of a series of steps. No lines are seen because exciton-lattice interaction is required to allow any transition at the edge, and the lattice vibrational continuum produces a range of allowed energies. The lowest shoulder, due to simultaneous photon and phonon absorption, disappears at low temperatures because few thermal phonons exist.

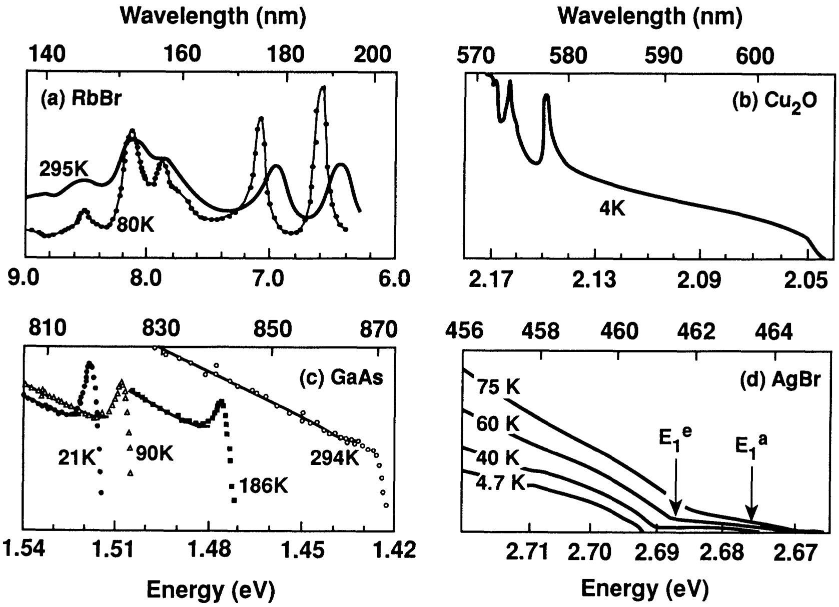

FIG. 3. Optical absorption spectra at temperatures indicated. (a) RbBr, Frenkel/intermediate case with strong lines (after Eby *et al.*, 1959); (b) Cu_2O, clear Wannier-Mott case (after Baumeister, 1961); (c) GaAs, Wannier-Mott case with line broadening greater than the spacing (after Sturge, 1962); (d) AgBr, indirect-gap mixed case showing no spectral lines, but two shoulders, one due to phonon absorption, beginning at E_1^a, and one to phonon emission, beginning at E_1^e (after Joesten and Brown, 1966). Vertical axes represent absorption coefficient in cm^{-1}, with maximum scales of (a) 7×10^5; (b) 320; (c) 1.2×10^5; (d) 4.

1.3.1.2 Emission and Nonlinear Effects. Exciton states may decay radiatively, but this process competes weakly with nonradiative decay modes. Emission is seen at low temperatures as lines near the band-gap energy ("edge") and as a Stokes-shifted band when strong coupling to phonons exists.

At high exciton densities optical properties of a material may be affected significantly. This topic, which is of great practical interest, is discussed in Sec. 2.

1.3.2 Other Properties. As neutral quasiparticles, excitons cannot participate in dc conductivity. However, they act to change the photoionization spectrum from what might be expected on a simple band model and they act as intermediate states in the electron-hole recombination process.

1.4 Other Topics

1.4.1 Exciton Trapping. The interaction of an exciton with crystal imperfections can greatly alter its state during its lifetime. As a moving packet, it is scattered by phonons and point defects. Strong interaction with phonons leads to self-trapping (Toyozawa, 1980), a state in which there is slow diffusive motion or no motion at all. Similarly, the exciton's energy may be transferred to an excited state of an impurity having energy lower than that of the exciton. Then being out of resonance with states of neighbors, the excitation is again immobile or trapped. On the band picture, when a hole is trapped at an acceptor, it may capture an electron into a stable orbit before recombination occurs. The result is again a trapped exciton. In the semiconductor case this state will generally have a large radius, like that of the free exciton in Fig. 1(b), and be fairly easy to ionize.

1.4.2 Exciton-Exciton Interactions. Under high-intensity excitation a material may become populated so densely with excitons that collisions between them are likely during their lifetime. As densities increase, this phenomenon is revealed in one of two ways: the formation of biexcitons (Knox *et al.*, 1969), whose energy is slightly less than twice that of one exciton because of the binding; or annihilation (Northrup and Simpson, 1958), in which a localized Auger process takes place resulting in heat, carriers, and either one exciton remaining or none. Whether biexciton formation or annihilation occurs is a question of competing rate processes. Biexcitons are generally observed in semiconductors and annihilation in molecular crystals.

At fairly high excitation densities in Cu_2O, behavior corresponding closely to Bose-Einstein condensation occurs; the exciton provides perhaps the only example of this phenomenon other than the case of liquid helium (Snoke *et al.*, 1990). Under very high excitation, a new phase can appear in which excitons have coalesced into electron-hole droplets (e.g., in silicon; see Gourley and Wolfe, 1978). These new phases are particularly interesting in the case of Cu_2O because conditions are very favorable for the observation of condensation. Cu_2O has a forbidden direct gap that provides excitons with long lifetimes, a small exciton Bohr radius and large exciton binding energy, and certain characteristics that minimize the likelihood of the presence of stable exciton pairs and the electron-hole liquid (Wolfe, 1992).

2. EXCITONS IN SEMICONDUCTORS

2.1 Crystals: General Remarks

Semiconductors have relatively large dielectric constants and relatively small electron and hole effective masses near the band extrema. Therefore, excitons in bulk semiconductors are spatially extended and are well-described by the effective-mass approximation in the Wannier picture. Excitonic effects appear at the fundamental absorption edge in III–V materials, and also at the direct edge of group IV materials such as Si and Ge with approximately 1–10-meV binding energies. In bulk GaAs material, the exciton Bohr radius is 14 nm, corresponding to a 4.2-meV binding energy. A large body of literature on excitons in bulk semiconductors exists. See, for example, Dimmock (1967), Pankove (1975), and Rashba and Sturge (1982).

The detailed optical characteristics of excitons are excellent probes of the microscopic crystalline quality since they display great sensitivity to crystal growth conditions, defect structures, temperature, strain, and applied electric and magnetic fields. In addition, excitonic absorption properties display dramatic density-dependent effects, with large changes in absorption observed at either high doping densities or high optical excitation levels. These properties bring important new considerations to optical device applications.

2.2 Quantum Wells, Wires, and Dots

2.2.1 Electron and Hole States in Two and Three Dimensions. As a result of recent advances in materials science, it has become possible to fabricate semiconductors one of whose dimensions is of the order of the exciton diameter in the bulk. Layered semiconductor structures have been fabricated by a number of different techniques in different material systems, such as GaAs-AlGaAs and GaAs-InGaAs, with monolayer accuracy. As a result, the densities of states for electrons and holes are significantly altered, and the effects are apparent in the optical properties (Dingle *et al.*, 1974). Figure 4(a) shows the electron density of states for three- and two-dimensional (3D,2D) semiconductors. Whereas the 3D semiconductor exhibits a square-root den-

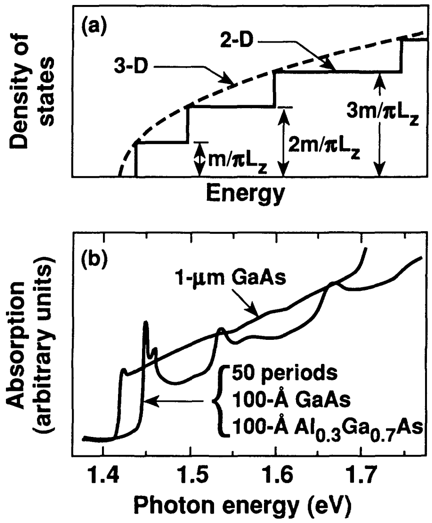

FIG. 4. Effect of dimensionality on properties of GaAs. (a) Densities of conduction electron states in bulk (3D) and in quantum wells (2D). The step heights involve *m*, the electron effective mass, and L_z, the confining well width. (b) Optical absorption in 3D and 2D samples of GaAs, showing the effect of the underlying density of states (after Schmitt-Rink *et al.*, 1989). The horizontal scales are slightly different; the major peaks in (b) correspond to steps in (a).

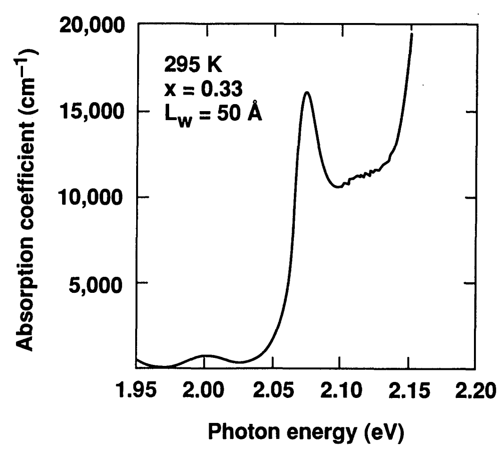

FIG. 5. Excitonic features in quantum wells of CdZnTe/ZnTe at room temperature (after Lee *et al.*, 1991).

sity of states, the 2D, or quantum-well, system exhibits a number of steps, as a result of the quantization of the electron into "particle-in-a-box" states. The effects of this quantization are clearly seen in experiments, as shown in Fig. 4(b) [see Schmitt-Rink *et al.* (1989)], which compares the absorption of bulk and quantum wells of GaAs. The valence bands are also significantly modified in quantum wells. One clear effect of quantum wells is the splitting of the heavy- and light-hole degeneracy at $k=0$ which occurs in bulk GaAs. As a result, two exciton lines are clearly observed near the band gap.

2.2.2 Modification of Exciton Properties in Two Dimensions. Figure 4(b) shows that at room temperature the quantum wells exhibit sharp excitons near each step in the density of states. In fact, each quantum subband has an excitonic feature that appears at a slightly lower energy (the binding energy) than the subband continuum. These sharp excitonic features in quantum wells are important for device applications, since excitons in quantum wells exhibit large sensitivity to applied fields and optical saturation effects. Figure 5 shows that excitonic features have also been obtained with room-temperature quantum wells in II–IV compounds as well: CdZnTe/ZnTe (Lee *et al.*, 1991). These systems are interesting for short-wavelength (<600 nm) applications.

2.2.3 Effects of Applied Electric Fields. Electric fields can be applied to quantum wells in two different field configurations: electric field parallel to and perpendicular to the quantum-well walls. The results are different: in the parallel field case, the exciton is readily field-ionized because there is no bound state in a parallel field. The exciton then broadens as the field is applied (Miller *et al.*, 1985) but does not substantially shift. The parallel-field effect has been used in femtosecond excitonic optoelectronics applications (Knox *et al.*, 1989b).

When the electric field is applied perpendicular to the quantum wells, the exciton shifts to lower energies but only slowly broadens, because the field ionization is significantly suppressed by the confinement of the electrons and holes. This effect (dubbed the "quantum-confined Stark effect") (Miller *et al.*, 1985,1986) forms the basis of a number of new and important optical modulators and related devices. Figure 6 shows the absorption in a quantum well waveguide for optical fields in different modes and for various perpendicular fields (Weiner *et al.*, 1985). In polariza-

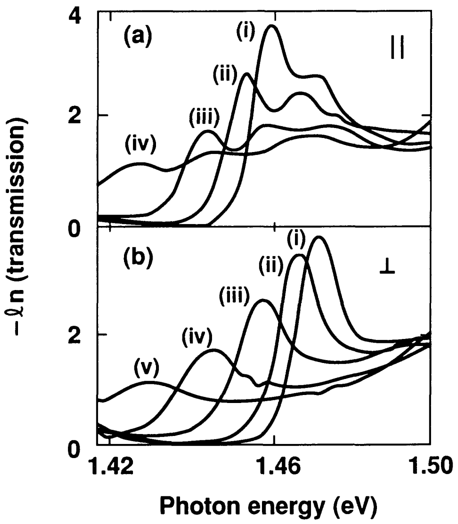

FIG. 6. Effect of electric field on properties of quantum-well excitons (after Weiner *et al.*, 1985). The applied dc field is perpendicular to the quantum-well walls in each case. The polarization of the incident light is (a) parallel and (b) perpendicular to well walls. Applied electric field strengths, in units of 10^5 V/cm and in order of increasing roman numerals: (a) 0, 0.6, 1.0, 1.5; (b) 0, 0.6, 1.1, 1.5, 2.0. Vertical axis: negative natural logarithm of the transmission coefficient.

tion perpendicular to the well walls [Fig. 6(b)] only the light-hole exciton is observed.

2.2.4 Device Applications of Quantum-Well Modulators. The achievement of a material system that offers large, fast, and sensitive light modulation by applied electric fields at room temperature opens the way for optical device applications. One such implementation is the "SEED" or self-electrooptic-effect device, which consists of a quantum-well modulator that contains a feedback loop (Fig. 7). Such a device operates as a light-intensity modulator that sets its own operating voltage, depending on the intensity of the incident light. This device exhibits bistable behavior that has been used to demonstrate a primitive optical computer (Miller, 1990).

2.2.5 Lower-Dimensional Systems. The steplike density of states obtained with quantum wells is further modified when the width of the sample also becomes comparable to the bulk exciton diameter, in a system called a "quantum wire." In a perfect wire, a one-dimensional density of states would result (Sercel and Valhala, 1990).

Although perfect zero-dimensional systems should exhibit a density of states reminiscent of sharp atomlike energy levels, at present inhomogeneous broadening and other effects prevent such an absorption spectrum from being observed. In the case of an applied magnetic field perpendicular to the quantum-well layers, the excitons are confined to essentially zero-dimensional states and a nearly perfect Landau-ladder absorption spectrum results (Akimoto and Hasegawa, 1967). By application of a strong magnetic field, the *ns* states of the excitons can be separated and studied. In CdSe, quantum dots only 3.0 nm in diameter have been prepared by a chemical technique (Brus, 1986). They exhibit a strong exciton peak, but not a characteristic density of states. Inhomogeneous broadening effects can easily mask such structure, since the energy levels depend critically on the particle size, which can vary by 20% or more.

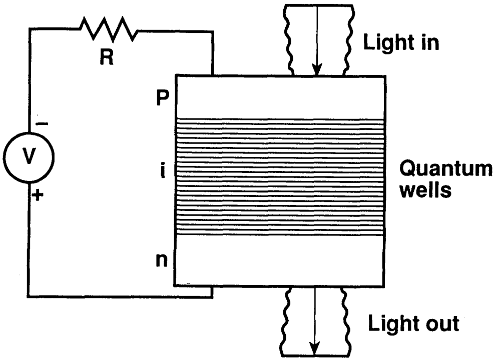

FIG. 7. Schematic of the self-electrooptic-effect device (after Miller *et al.*, 1985).

2.2.6 High-Density Effects. Excitons in semiconductors exhibit interesting and complex effects at high densities, by which we generally mean a density of one exciton per exciton Bohr radius cubed, or higher. Such densities may be achieved either by doping or by high-intensity optical excitations. In the case of quantum wells, excess electrons or holes can be produced by doping the barriers only ("modulation doping"; Dingle *et al.*, 1978). The excess electrons or holes then fill the quantum wells, leaving the ionized impurities in the barriers, and resulting in high in-plane transport mobilities. In such a case, excitonic features near the band edge are severely modified because of many-body interactions, and a correlation singularity called the "Mahan exciton" or "Fermi-edge singularity" results (Mahan, 1967a,b; Skolnick *et al.*, 1987; Livescu *et al.*, 1988).

At high excitation intensities, the excitonic nonlinear optical response is complicated. Excitonic absorption is affected in several ways by the injection of additional carriers. The exciton-exciton interaction contains screening terms, renormalization terms, and exchange terms. In addition, excitons have an ultrafast (300-fs) nonlinear response at room temperature due to exciton-phonon interactions (Knox *et al.*,1985) that has been used in various applications. AC Stark effects are exhibited by excitons when sub-band-gap excitation is used (Mysyrowicz *et al.*, 1986; Knox *et al.*, 1989a). In the case of high densities of optical excitations in modulation-doped quantum wells, carrier-carrier scattering times as short as 10 fs have been measured (Knox *et al.*, 1988).

2.3 Amorphous Semiconductors

Amorphous semiconductors are of great practical importance, but the line broadening resulting from their intrinsic disorder prevents observation of any ordinary exciton effects (see Zallen, 1983).

3. EXCITONS IN OTHER MATERIALS

3.1 Molecular Crystals

Of the many possible examples of molecular crystals, we will discuss a large subset that typifies them, namely, organic crystals. Low-lying electronic excited states of organic molecules usually involve rearrangement of the electron distribution among the same molecular π orbitals without a pronounced change in "size" of the molecule. Therefore, in organic molecular crystals, excitons are close to the Frenkel limit, and, to a first approximation, spectra are essentially perturbed molecular spectra retaining all their structural complexity.

3.1.1 General Characteristics. At normal temperatures, molecular-crystal spectra generally appear simply to be broadened and somewhat shifted spectra of the same molecules in vapor phase. The broadening is due partly to vibrations, partly to disorder, and only slightly to exciton band effects.

At low temperatures, where thermal broadening is greatly reduced, molecular crystal spectra can become exquisitely structured. The primary feature added to the molecular spectrum is "Davydov splitting," which is due to the interaction between like molecules having different orientations in the crystal's unit cell (Davydov, 1948). This effect is seen in Fig. 8. Light polarized in different directions is absorbed maximally at slightly different wavelengths, corresponding to the two Davydov exciton components. The broadened solution spectrum is also shown in Fig. 8 (Wolf, 1958).

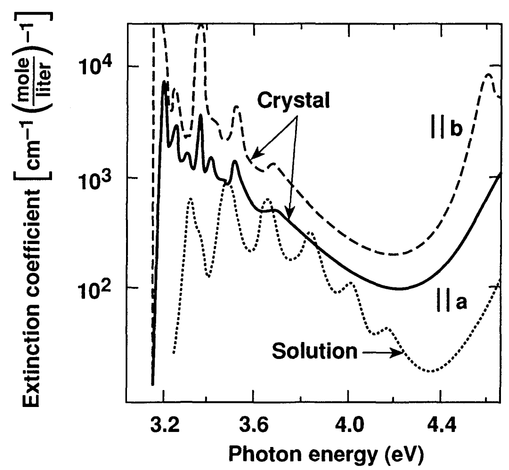

FIG. 8. Absorption spectra of anthracene in solution (lower curve) and as a crystal (two upper curves) (after Wolf, 1958). The crystal curves differ in the polarization of the absorbed light relative to the crystal axes, which are denoted "a" and "b."

3.1.2 Triplet States. The long lifetime of triplet excited states, which are well characterized in organics, affords the chance to examine transport and coherence of excitation over considerable spatial extents.

In most systems triplet states are difficult to observe optically because of the spin selection rule. However, they may be produced as decay products of higher excited states, and their spectral profile can then be found by an up-conversion process: an additional photon absorbed by the triplet itself leads to a highly excited state that (in anthracene, for example) fluoresces in the blue. The excitation spectrum for this process can be converted into what is essentially an absorption spectrum for the triplet (Avakian *et al.*, 1963). Another application of the blue fluorescence is its use as a detector of triplet excitons through their mutual annihilation, an effect that has been used to determine the diffusion constant of triplets (Avakian and Merrifield, 1964).

Exciton coherence is an important but complex topic. An excellent introduction and review, with special reference to organic crystals, is given by Burland and Zewail (1979).

3.1.3 Photon Channeling. The exciton carries the energy of the photon that created it, but need not occupy as large a volume. This fact, along with the ability of the exciton to migrate, has been put to practical use in recent proposals relating to designs of subwavelength-resolution microscopes (Lieberman *et al.*, 1989). Light channeled into a pipette with a gold-coated tip forms a subwavelength source, provided that the light is used sufficiently near the tip that diffraction effects do not wash out the channeling effect. However, the dimensions of the effective light source produced by this method are limited to about > 50 nm by evanescence, which reduces the contrast between tip and channel. The pipette is transformed into an active source by filling its tip with anthracene. The excitons formed in this crystal by the primary light source diffuse to localized emitters in the tip that then act as a secondary source of smaller dimension (Kopelman *et al.*, 1991).

3.2 Other Large-Band-Gap Crystals

3.2.1 Rare-Gas Crystals. In their ground states, crystals of the rare gases (Ne,Ar,Kr, Xe,Rn) resemble molecular crystals, because of their weak interatomic van der Waals attraction. However, exciton states in these crystals are much different because the electron must occupy an orbital of considerably larger spatial extent than the hole it leaves behind. This then places the electron in a region where the dielectric is effective in reducing the Coulomb interaction, and a Wannier-Mott series of exciton states is observed.

In a typical rare-gas atom, the ground state has a filled outer shell of p electrons (p^6), and the first excited state configuration is p^5s. Remarkably, the radius of the lowest-energy exciton is roughly the same as that of the excited s electron of the free atom. This coincidence has made the rare gases an ideal, though not entirely practical, class of solids in which to study the electronic structure of the exciton and its relation to band structures. Recent work has relied almost exclusively upon synchrotron radiation to produce the vacuum-ultraviolet photon energies necessary in these crystals, which exist at normal pressures only at very low temperatures (Schwentner *et al.*, 1985).

3.2.2 Ionic Crystals. Excited states that can now be recognized as excitons were first observed in the ultraviolet absorption spectra of alkali halide crystals. The typical strongly ionic crystal consists of closed-shell ions and is electronically very similar to the rare-gas crystal. For example, in NaCl, Cl^- is isoelectronic to Ar and the Cl^- sublattice has the solid Ar structure. See Knox and Teegarden (1968) for a general review.

Early workers interpreted strong absorption lines in the vacuum ultraviolet to an electron transfer process, in which one of the Cl^- electrons was moved to a neighboring Na^+. The energy required was related to the electron affinity of the halide, the ionization energy of the alkali, and the net difference in the Coulomb interaction felt by the electron. The state created is indeed just the lowest Wannier-Mott state built from hole states in the uppermost valence band (based almost exclusively on Cl^- orbitals) and electron states in the lowest conduction band (based largely on Na^+ orbitals) in which the radius of the exciton is such that is spends much time on the nearest-neighboring Na^+ ions.

Wannier-Mott absorption series exist in the alkali halides, but the existence of polar lattice modes prevents a complete analogy with the rare gases. At higher orbital radii, the low-frequency dielectric constant begins to dominate the effective Coulomb interaction and weakens the binding, so that by the time the electron and hole have enough energy to dissociate, they have become polarons. In addition, the hole usually becomes self-trapped on a pair of neighboring halide ions. This trapping process is so strong that the exciton itself is subject to it. This results in probably the largest known luminescence Stokes shifts (typically 3 to 4 eV) (e.g., Kabler, 1964).

3.3 Biomaterials

The bulk of biological material is transparent in the visible region because it consists of protein, nucleic acids, and lipids, all of which have few π-electron systems that would support transitions in the 2–4-eV range. The exciton is of known importance in certain pigmented proteins (Sec. 3.3.1), and as a radiation by-product may have some role in protein and DNA damage (Sec. 3.3.2).

3.3.1 Photosynthesis. The exciton is prominent in models of "light-harvesting," the first step in photosynthesis (see, e.g., Sauer, 1986). Hundreds of chlorophyll (Chl) molecules are associated with each photochemical reaction center in green plants. It has been found possible to model the Chls as a loosely packed molecular crystal in which excitons diffuse toward the "trap" where photochemistry is initiated.

An absorption spectrum typical of green plant chloroplasts is shown in Fig. 9. The strong bands in the vicinity of 1.88 and 3.1 eV (670 and 430 nm) are directly attributable to Chl; the remaining background absorption is due to "accessory pigments" that transfer energy to the Chl. Energy of light absorbed

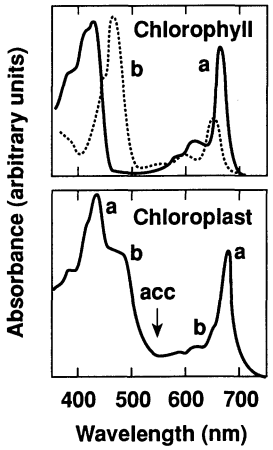

FIG. 9. Typical absorption spectrum of a chloroplast, here isolated from a spinach leaf (lower figure). The "a" peaks are due to chlorophyll-*a*, "b" to chlorophyll-*b*, whose individual absorption profiles are shown in the upper figure. The region "acc" involves accessory pigments that transfer energy to the chlorophylls. Excitons represented by the 670-nm "a" band act to transport absorbed solar energy to chemical reaction centers. (After Lin, 1991.)

anywhere in the spectrum quickly ends up in the 670-nm state of Chl somewhere on the lattice, and then migrates toward a reaction center by exciton diffusion. The several hundred chlorophylls are thus said to act as an "antenna." This colorful terminology refers only to the contribution of the Chls to the effective cross section for absorption.

Photosynthetic bacteria behave similarly, but rely on bacteriochlorophyll (Bchl), a form of Chl with its lowest singlet excited states in the 1.55-eV (800-nm) region.

3.3.1.1 Chlorophyll Aggregates. X-ray crystallography of proteins (see CRYSTALLOGRAPHY) has confirmed the basic hypotheses about the Chl antenna: the chromophores are packed closely enough for efficient energy transfer but loosely enough so as not to form photochemical quenching centers. One of the first antenna protein structures obtained was that of the photosynthetic bacterium *Prosthecocloris aestuarii* (Matthews *et al.*, 1979); see Fig. 10. The protein illustrated is one part of a trimer, and the trimers containing 21 Bchl are arrayed in the membrane to form the "loosely coupled molecular crystal." The excited eigenstates of the monomer are in fact delocalized, being linear combinations of localized excitations on the individual Bchls, and are close indeed to Frenkel's original exciton but on a microscopic to mesoscopic scale (Pearlstein, 1992).

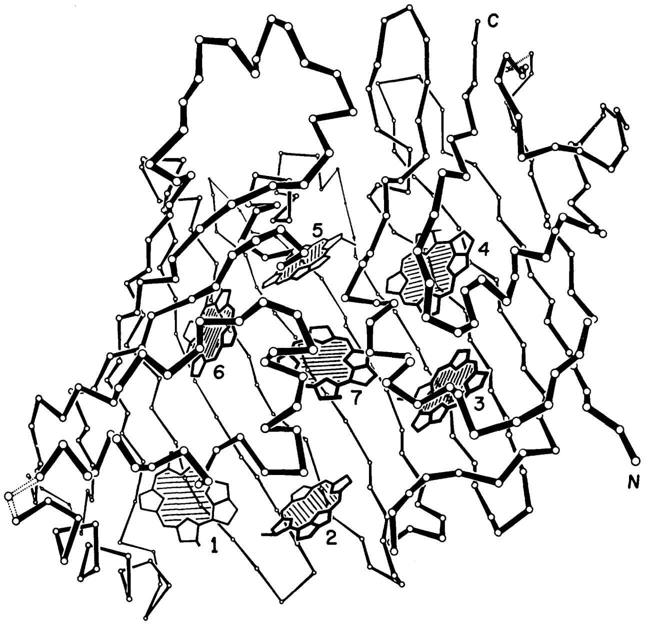

FIG. 10. Organization of a chlorophyll protein, in this case bacteriochlorophyll. The connected line segments represent the sequence of amino acids "packaging" the small numbered rings that represent the chromophores of the chlorophyll molecules. (Matthews *et al.*, 1979. Reproduced by permission of Academic Press, Inc.)

Transport of excitation energy into the trap (see next section) is very efficient. Chl fluorescence in green plants has an efficiency of about 2%, compared with 33% in solution. This corresponds to a trapping efficiency of 94%.

3.3.1.2 A Dimer State as a "Miniexciton." The lowest excited states of two strongly interacting identical molecules are a complete mixture of states of the individual molecules: $\frac{1}{2}(\Phi_A+\Phi_B)$, where Φ_A and Φ_B are states of localized excitation on the two molecules A and B, respectively. These two states are Frenkel's states for the case $N=2$ and have been termed "miniexcitons" in the organic crystal literature. The energy levels of the miniexcitons are above and below those of the monomer (Fig. 11). In photosynthetic organisms, evolution appears to have led to the construction of antenna-trap systems, exemplified by the sketch of Fig. 11. Anywhere from 50 to 300 chlorophyll molecules act as the antenna, supplying energy by exciton diffusion to the dimers, which act as photochemical traps and are the principal component of the reaction center. The traps mediate excitation-induced charge separation and the charges are directed to chemical systems capable of utilizing the energy.

3.3.2 DNA Damage. While their role in propagating destructive packets of energy along the bases in the helix structure of deoxyribonucleic acid (DNA) is not completely established, exciton diffusion is prominent in theoretical models of the process (Gueron and Shulman, 1968).

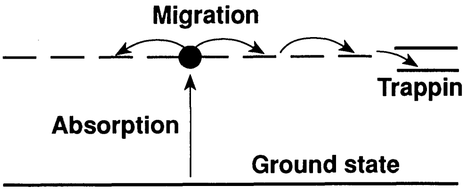

FIG. 11. Illustrating the principle of the photosynthetic antenna-trap system. Excited states of two strongly coupled "trap" chromophores, represented by the thick lines, lie above and below those of five identical weakly coupled "antenna" chromophores. An excitation packet, or localized exciton, makes a random walk over the antenna molecules, eventually falling into the trap. In practice, there are dozens of antenna chromophores arranged in three dimensions about the trap.

4. SUMMARY

The exciton is the first product of light absorption, and its characteristics are therefore frequently revealed by a material's absorption spectrum. The exciton may also be produced by an incident particle or by a quantum of ionizing radiation. It is a system state whose lifetime is usually in the subnanosecond range, but during that lifetime important energy transduction can take place. The exciton has an acknowledged practical importance in natural photosynthesis and it has recently become an increasingly important factor in the design of semiconductor quantum well devices.

GLOSSARY

Absorption Line: A sharp feature in an absorption spectrum characterized by A, an overall strength; ν_0, a center frequency; and Γ, a full width at half maximum. See Gaussian and Lorentzian.

Auger Process: A transition in which two electron-hole pairs recombine, with their mutual energy appearing as one highly energetic pair.

Bands: See energy bands.

Chloroplast: A micrometer-size unit in a cell of a photosynthetic organism. It contains light-harvesting chlorophyll proteins in membranes and much of the primary photochemical apparatus.

Chromophore: That part of a complex molecule containing the orbitals that are active in making optical transitions in the infrared and visible regions of the spectrum.

Coherence: A persistence of a definite phase relation in a state of a system. For the exciton it may be a relation to the phase of the excitation light or relationships between phases of localized wave functions making up the exciton.

Conduction Band: An energy band, normally almost empty in a nonmetal and lying higher than the valence-band maximum energy by an amount E_G, the band gap.

Crystal Imperfections: Any of several standardized departures from ideality in a crystal, important for determining properties of real crystals. Examples are vacancies, dislocations, excess electrons, surfaces, and phonons.

Density of States: The number of energy eigenstates per unit energy range, usually referring to states having a virtually continuous distribution of energies.

Effective Mass: The ratio of force to acceleration of a quasiparticle. In typical semiconductors the electron effective mass m_e^* is often much smaller than the ordinary electron mass, and the hole effective mass m_h^* usually comparable to or somewhat smaller than the ordinary electron mass. Even though no actual mass is transported, the exciton has an effective mass of $m_e^* + m_h^*$.

Eigenstate: The general quantum-mechanical equivalent of Bohr's stationary state. An energy eigenstate can be considered one of any system's "natural" states, one which would not change if the system had no external interaction.

Energy Bands, Energy-band Structure: The energies of orbitals of electrons in solids. The energies of the eigenstates fall into well-defined near-continua known as "bands." Each eigenstate is characterized by a wave vector k.

Exciton-Lattice Coupling: A contribution to the Hamiltonian (energy) of an exciton induced by vibrations of the crystal lattice about its average configuration (phonons).

Gaussian: A function of ν given by $A \times \exp[-(\nu-\nu_0)^2/2\sigma^2]$. See Absorption Line for symbol definitions. Here σ is related to the full width at half maximum by $\sigma = \Gamma\sqrt{8} \ln 2 = 2.355\Gamma$.

Heavy (Light) Holes: Holes in a valence band at whose maximum the effective mass is relatively larger(smaller); applies to cases in which the valence-band energies themselves are equal (degenerate), usually at $\mathbf{k}=0$.

Hole: A missing electron, in particular reference to an almost-filled atomic shell or energy band in a solid. Frequently likened to an empty space in an otherwise filled parking lot. It can move as a result of motion of the other electrons (autos). See also Effective Mass, Quasiparticle.

Landau Ladder: A sequence of quantum states of electrons in a magnetic field, so called because of its equal spacing of energy levels that is in turn due to the isomorphism of the problem with that of the simple harmonic oscillator.

Light Holes: See Heavy Holes.

Lorentzian: A function of ν given by $A(\Gamma/2)^2/[(\nu-\nu_0)^2+(\Gamma/2)^2]$. See Absorption Line for symbol definitions.

Phonon: A quantum of vibrational excitation in an extended system, usually possessing a wave vector $\mathbf{k}$, polarization (transverse or longitudinal), and type (acoustic or optical). The acoustic phonon has vanishing energy as $|\mathbf{k}| \rightarrow 0$.

Quantum Dot: One of an ensemble of small clusters of semiconductor material in a matrix of larger band gap, the ensemble having a very narrow size distribution with dimensions smaller than the radius of an exciton in the corresponding bulk material.

Quantum Well: A physical sandwich structure consisting of very thin (~ 10 nm) layers, usually of semiconductor material, designed such that electrons, holes, and excitons move essentially in two dimensions, confined by the "bread."

Quantum Wire: Similar to quantum dots and quantum wells but confined in two dimensions instead of in three and one, respectively.

Quasiparticle: A term applied to any localizable wavelike state of a condensed system whose motion can be approximated by simple particlelike dynamics. Typically the term is applied when the effective mass is nonzero: electrons (when characterized by an effective mass), holes, and excitons are examples.

Reaction Center: A molecular unit, usually a dimer composed of two identical chromophores, that accepts electronic excitation energy and mediates separation of charge in the first chemical step of photosynthesis.

Selection Rules: Restrictions on the allowedness of transitions between quantum-mechanical states, usually dictated by symmetry of the system.

Stokes Shift: The amount (in energy or wavelength units) by which the peak position of an emission band differs from that of the associated absorption band. It is a measure of the strength of the interaction between the excited state and the lattice.

Valence Band: An energy band normally mostly filled with electrons, often those participating in the covalent bonding of the solid.

List of Works Cited

Akimoto, O., Hasegawa, H (1967), *J. Phys. Soc. Jpn.* **22** 181–191.

Avakian, P., Abrahamson, E., Kepler, R. G., Caris, J. C. (1963), *J. Chem. Phys.* **39**, 1127–1128.

Avakian, P., Merrifield, R. E. (1964), *Phys. Rev. Lett.* **13**, 541–543.

Baumeister, P. W. (1961), *Phys. Rev.* **121**, 359–362.

Brus, L. (1986), *J. Phys. Chem.* **90**, 2555–2560.

Burland, D. M., Zewail, A. H. (1979), *Adv. Chem. Phys.* **40**, 369–484.

Davydov, A. S. (1948), *Zh. Eksperim. Teor. Fiz.* **18**, 210–216.

Dimmock, J. O. (1967), in: R. K. Willardson and A. C. Beer (Eds.), *Semiconductors and Semimetals,* New York: Academic, Vol. 3, pp. 259–319.

Dingle, R., Wiegmann, W., Henry, C. H (1974), *Phys. Rev. Lett.* **33**, 827–830.

Dingle, R. C., Stormer, H., Gossard, A. C., Wiegmann, W. (1978), *Appl. Phys. Lett.* **33**, 665–667.

Eby, J. E., Teegarden, K. J., Dutton, D. B. (1959), *Phys. Rev.* **116**, 1099–1105.

Frenkel, J. (1931), *Phys. Rev.* **37**, 17–44, 1276–1294.

Gourley, P. L., Wolfe, J. P. (1978), *Phys. Rev. Lett.* **40**, 526–530.

Gueron, M., Shulman, R. G. (1968), *Annu. Rev. Biochem.* **37**, 571–596.

Joesten, B. L., Brown, F. C. (1966), *Phys. Rev.* **148**, 919–927.

Kabler, M. N. (1964), *Phys. Rev.* **136**, A1296–A1302.

Knox, R. S. (1963), *Theory of Excitons,* New York: Academic.

Knox, R. S., Teegarden, K. J. (1968), in: W. B. Fowler (Ed.), *Physics of Color Centers,* New York: Academic, pp. 1–15, 625–626.

Knox, R. S., Nikitine, S., Mysyrowicz, A. (1969), *Optics Commun.* **1**, 19–21.

Knox, W. H., Fork, R. L., Downer, M. C., Miller, D. A. B., Chemla, D. S., Shank, C. V., Gossard, A. C., Wiegmann, W. (1985), *Phys. Rev. Lett.* **54**, 1306–1309.

Knox, W. H., Chemla, D. S., Livescu, G., Cunningham, J. E., Henry, J. E. (1988), *Phys. Rev. Lett.* **61**, 1290–1293.

Knox, W. H., Chemla, D. S., Miller, D. A. B., Stark, J. B., Schmitt-Rink, S. (1989a), *Phys. Rev. Lett.* **62**, 1189–1192.

Knox, W. H., Henry, J. E., Goossen, K. W., Li, K. D., Tell, B., Miller, D. A. B., Chemla, D. S., Gossard, A. C., English, J., Schmitt-Rink, S. (1989b), *IEEE J. Quant. Electron.* **25**, 2586–2595.

Kopelman, R., Lieberman, K., Lewis, A., Tan, W. (1991), *J. Lumin.* **48/49**, 871–875.

Lee, D., Johnson, A. M., Zucker, J. E., Feldman, R. D., Austin, R. F. (1991), *J. Appl. Phys.* **69**, 6722–6724.

Lieberman, K., Harush, S., Lewis, A., Kopelman, R. (1989), *Science* **247**, 49–51.

Lin, S. (1991), Ph.D. thesis, University of Rochester.

Livescu, G., Miller, D. A. B., Chemla, D. S., Ramaswamy, M., Chang, T. Y., Sauer, N., Gossard, A. C., and English, J. H. (1988). *IEEE J. Quant. Electron.* **24**, 1677–1689.

Mahan, G. D. (1967a), *Phys. Rev.* **153**, 882–889.

Mahan, G. D. (1967b), *Phys. Rev.* **163**, 612–617.

Matthews, B. W., Fenna, R. E., Bolognesi, M. C., Schmid, M. F., Olson, J. M. (1979), *J. Molec. Biol.* **131**, 259–286.

Miller, D. A. B. (1990), *Optics Quant. Electron.* **22**, S61–S98.

Miller, D. A. B., Chemla, D. S., Schmitt-Rink S. (1986), *Phys. Rev. B* **33**, 6976–6982.

Miller, D. A. B., Chemla, D. S., Damen, T. C., Gossard, A. C., Wiegmann, W., Wood, T. H., Burrus, C. A. (1985), *Phys. Rev. B* **32**, 1043–1060.

Mott, N. F. (1938), *Trans. Faraday Soc.* **34**, 500–508.

Mysyrowicz, A., Hulin, D., Antonetti, A., Migus, A., Masselink, W. T., Morkoc, H. (1986), *Phys. Rev. Lett.* **56**, 2748–2751.

Northrup, D. C., Simpson, O. (1958), *Proc. R. Soc. London Ser. A* **244**, 377–386.

Pankove, J. I. (1975), *Optical Properties in Semiconductors,* New York: Dover.

Pearlstein, R. M. (1992), Photosynthesis Research, **31**, 213–226.

Rashba, E. I., Sturge, M. D. (1992), *Excitons,* Amsterdam: North-Holland.

Sauer, K. (1986), in: L. A. Staehelin, C. J. Arntzen (Eds.), *Photosynthesis III,* in *Encyclopedia of Plant Physiology,* N.S., Vol. 19. Berlin: Springer, Chap. 2, 85–97.

Schmitt-Rink, S., Chemla, D. S., Miller, D. A. B. (1989), *Adv. Phys.* **38**, 89–188.

Schwentner, N. G., Koch, E. E., Jortner, J. (1985), *Electronic Excitation in Condensed Rare Gases,* Berlin: Springer-Verlag.

Sercel, P. C., Valhala, K. J. (1990), *Phys. Rev. B* **42**, 3690–3710.

Simpson, O. (1956), *Proc. R. Soc. London Ser. A* **238**, 402–411.

Skolnick, M. S., Rorison, J. M., Nash, K. J., Mowbray, D. J., Tapster, P. R., Bass, S. J., Pitt, A. D. (1987), Phys. Rev. Lett. **58**, 2130–2133.

Snoke, D. W., Wolfe, J. P., Mysyrowicz, A. (1990), *Phys. Rev. B* **41**, 11171-11184.

Sturge, M. D. (1962), Phys. Rev. **127**, 768–776.

Thomas, D. G., Hopfield, J. J. (1961), Phys. Rev. **124**, 657–665.

Toyozawa, Y. (1980), in: R. Kubo, E. Hanamura (Eds), *Relaxation of Elementary Excitations,* Berlin: Springer-Verlag, pp. 3–18.

Wannier, G. H. (1937), *Phys. Rev.* **52**, 191–200.

Weiner, J. S., Chemla, D. S., Miller, D. A. B., Damen, T. C., Burns, C. A., Wood, T. H., Gossard, A. C., Wiegmann, W. (1985). *Appl. Phys. Lett.* **47**, 1148–1150.

Wolf, H. C. (1958), *Z. Naturforsch.* **A13**, 414-419.

Wolfe, J. P. (1992), *J. Lumin.* **53**, 327-334.

Zallen, R. (1983), *The Physics of Amorphous Solids,* New York: Wiley, p. 267.

Further Reading

Chemla, D. S., Miller, D. A. B. (1985), Room-Temperature Excitonic Nonlinear-Optical Effects in Semiconductor Quantum-Well Structures, *J. Opt. Soc. Am.* **B2**, 1155–1173.

Cho, K. (Ed.) (1979), *Excitons,* Berlin: Springer-Verlag.

Grassano, U. M., Terzi, N. (Eds.) (1987), *Excited-State Spectroscopy in Solids,* Amsterdam: North-Holland.

Knox, R. S. (1963), *Theory of Excitons,* New York: Academic.

Pope, M., Swenberg, C. E. (1982), *Electronic Processes in Organic Crystals,* Oxford: Oxford University Press, esp. Chap. 1.

EXPLOSIVES

RAY ENGELKE AND STEPHEN A. SHEFFIELD, *Detonation Systems Group, Los Alamos National Laboratory, Los Alamos, New Mexico, U.S.A.*

Introduction 327
1. **History of Condensed-Phase Explosives** 330
1.1 Explosive Materials Development and Use 330
1.2 Appearance of a Quantitative Picture of Detonation 331
2. **Planar Steady Detonation Theory** 332
2.1 Hydrodynamic Jump Conditions for Laminar Steady Flow and Across Discontinuities 332
2.2 Rayleigh Line and Hugoniot Curve 333
2.3 Chapman-Jouguet Theory 333
2.4 Zel'dovich–von Neumann–Doering Theory 335
3. **Equations Governing Reactive Flow** 336
3.1 Fluid-Mechanical Equations 336
3.2 Chemical-Heat–Release Relations 337
3.3 The Material Equation of State 337
3.4 Qualitative Features of Flow with Reaction 338
3.5 Taylor Wave 338
3.6 Solutions of Reactive Flow Equations for Complex Systems 339
4. **Experimental Aspects of Initiation and Steady Detonation** 340
4.1 Initiation of Detonation 340
4.1.1 1D Homogeneous Initiation 341
4.1.2 1D Heterogeneous Initiation 342
4.1.3 Deflagration-to-Detonation Transition 343
4.2 2D Steady Detonation in Homogeneous and Heterogeneous Materials 343
5. **Properties of Chemical Explosives** 346
5.1 Explosive Classifications 347
5.2 Chemical and Detonation Properties for Selected Explosives 348
5.3 Sensitivity Properties 349
5.4 Initiation and Detonation Measurement Techniques 351
6. **Summary** 353
Glossary 355
Works Cited 356
Further Reading 357

INTRODUCTION

When the term "explosion" is mentioned, most people picture a violent, uncontrolled, chaotic event such as shown in Fig. 1. It is a principal purpose of this article to give a qualitatively different view of explosive phenomena. We will present a picture where explosive events are governed and rigorously describable in terms of the conservation of mass, momentum, and energy and certain material specific properties of the explosive material. Furthermore, we present evidence that when viewed experimentally, on the proper space and time scales, the observed phenomena are, in fact, orderly and, when variables are well controlled in the experiment, have some simplicity. The multiple-exposure photograph shown in Fig. 2 shows such a case.

Explosives have the property that they exist in an energetically metastable chemical state that can be controllably destabilized. The result of such destabilization can be extreme levels of power generation and extreme physical states in the explosive and its surroundings (e.g., very high pressure). Energy is released in a reactive wave process in which the passage of a shock wave triggers the stored energy release and, in turn, some of this

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

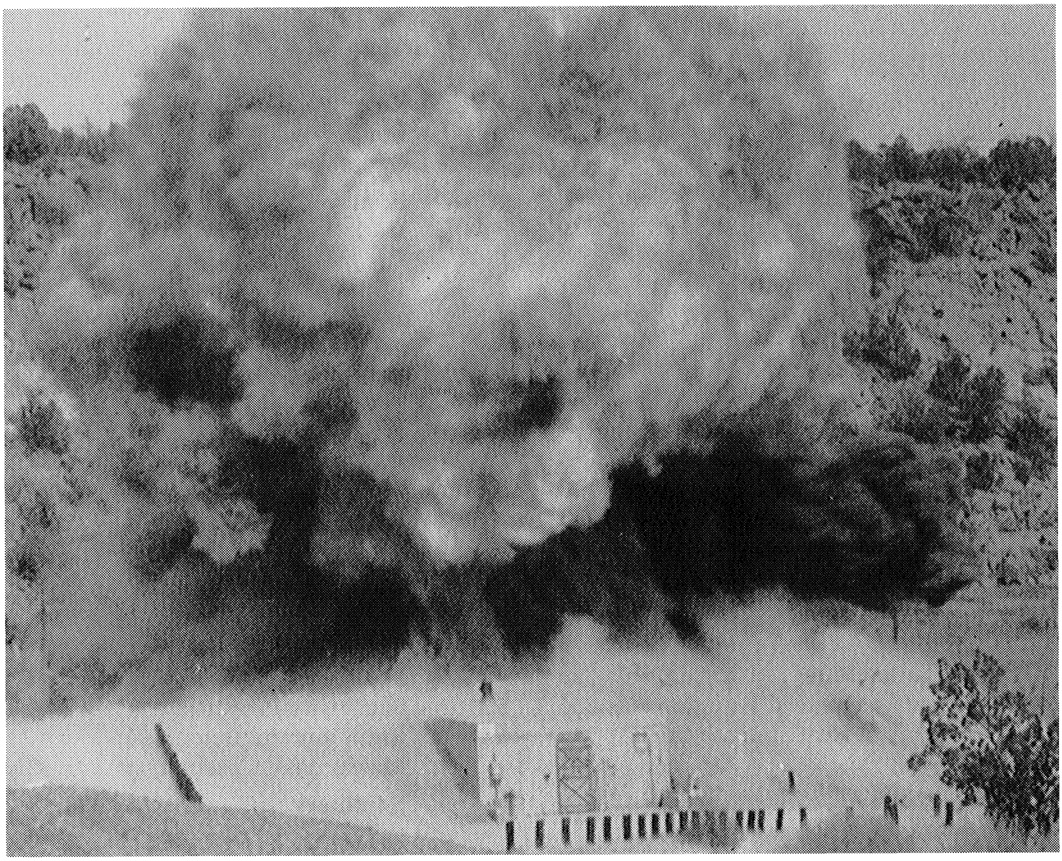

FIG. 1. Photograph of a large explosion at a firing bunker for explosive experimentation.

energy sustains the shock-wave motion: this is called a "detonation wave." Pressures produced in high-performance explosive detonations are typically in the 3×10^5-atmosphere range. The high detonation pressures mean that the explosive and any material (e.g., steel) in contact with it must be treated as compressible fluids using the principles of shock-wave physics.

Explosions and explosive action are not always due to high-explosive detonations; in fact, a wide variety of types and rates of reactions exist (depending on the particular metastable components involved). Types of explosive reactions range from nuclear explosions (the result of a nuclear chain reaction), to rapid first-order phase transitions (for example, vapor explosions such as occurred at Mt. St. Helens in 1980), to rapid burning of gunpowder, pyrotechnics, and propellants (such as occur in guns), to dust/air explosions (that sometimes occur in grain elevators). Detonating gas and condensed-phase explosives represent a significant part of the explosive reactions that occur. Rates of reaction range over six orders of magnitude (from a few millimeters per second to a few kilometers per second) depending on the particular phenomena involved.

Because of space limitations, we concentrate on condensed-phase materials that do work by detonation. Primarily because of their initial high mass density and the associated high energy density, liquid and solid materials have the most practical importance. However, most of the processes to be described are thought to occur both in condensed phases and in gases. The space scales are usually much smaller in condensed-phase materials; this makes it more difficult to measure and understand the processes in detail. Because of this, much of our understanding is

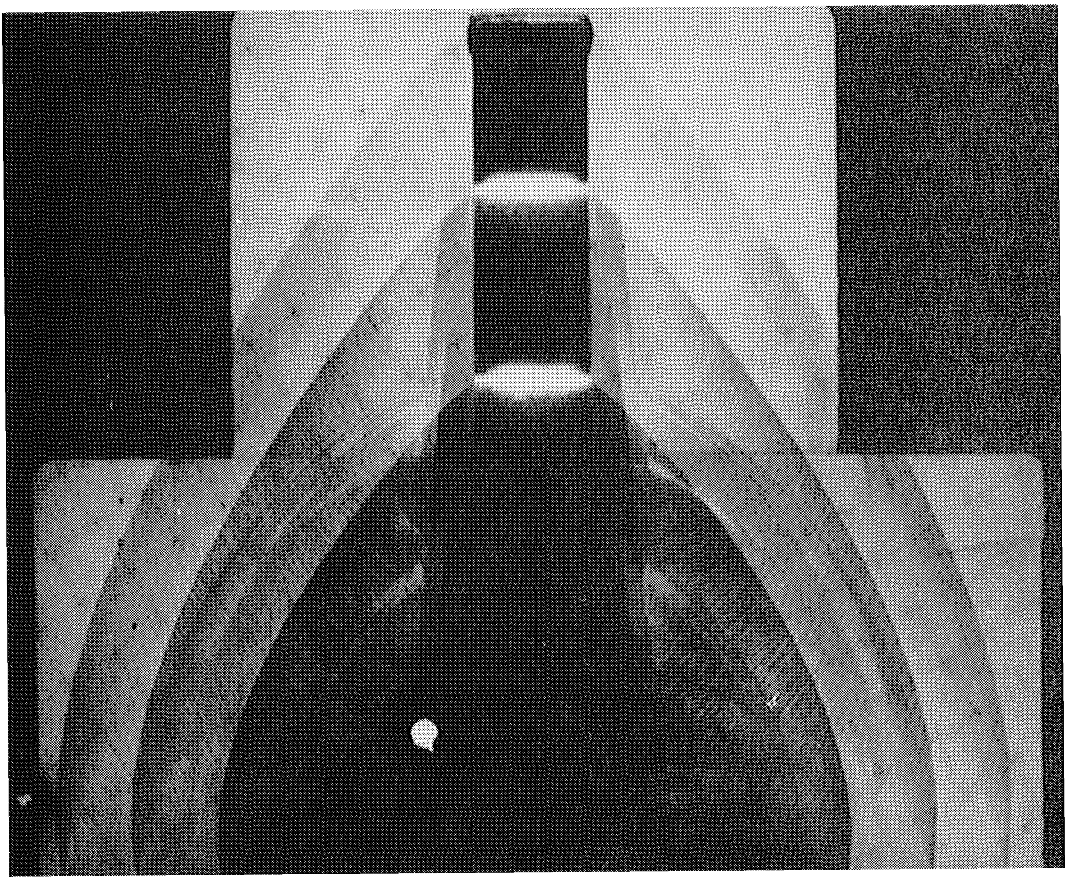

FIG. 2. Triple-exposure image-intensifier camera photograph of a detonation wave propagating in a cylindrical tube (ca. 10 cm diam) loaded with explosive. The detonation front at three locations can be clearly seen, along with the three associated shock waves driven into the surrounding liquid. The image converter tube was electronically shuttered to give an exposure time of 20 ns. This was done at three times about 40 μs apart.

phenomenological and is based on studies of one-dimensional (1D) waves and their interaction with the surrounding "confinement" material.

In chemical explosives, the energy storage is usually via the proximity of a fuel and an oxidizer. In high-performance chemical explosives the fuel and oxidizer are usually present in the same molecule (molecular explosives), but isolated by chemical bonds—e.g., TNT (2,4,6-trinitrotoluene). In lower- or special-performance chemical explosives, the fuel and oxidizers may be contained in components that are mixed together—e.g., mining slurry explosives.

In this article we begin by giving a short history of the theory of detonation and of explosive materials development (Sec. 1); only a few important details are covered to give some idea of the long and interesting history in this area. This is followed by two sections dealing with some of the important theoretical developments that have been made. Section 2 deals with planar steady detonation theory, including the conservation relations, simple shock-wave theory, and the two simple models that describe planar steady detonation. Section 3 discusses the general reactive fluid-flow equations, heat-release functions, equations of state, and various qualitative features of reactive flow. Theoretical solutions relating to complex systems are only briefly mentioned and most of our discussion of such systems is experimentally based. The experimental aspects of initiation and steady detonation are discussed in Sec. 4 including the differences between "homogeneous" and "heterogeneous" explosives. Section 5 contains chemical, detonation, and sensitivity information about certain selected high-explosive materials and discusses how some of the measurements in detonating materials are made.

By necessity, many details relating to explosive materials have been only briefly covered or have been left out altogether. For instance, deflagration, propellants, and pyrotechnics, even though they are important topics in some explosives circles, have only been mentioned. Further information relating to these and other areas can be obtained in the cited references and the articles indicated for further reading at the end of the paper.

1. HISTORY OF CONDENSED-PHASE EXPLOSIVES

Explosive technology has progressed from the use of rapid-burning materials to very sophisticated molecular explosives and explosive mixtures tailored to particular applications. This process has taken place over a period of more than a thousand years; major materials and theory advances have been made in the last 150 years. Our understanding of the physics of initiating and detonating materials is considerable, but much remains to be done to understand the microscopic chemistry. In this section we briefly discuss the history of explosive materials development and in what areas explosive materials are used today. This is followed by a short discussion of the development of theoretical concepts relating to detonation and the accompanying reactive flow.

1.1 Explosive Materials Development and Use

Explosive technology began when the oxidizer potassium nitrate (KNO_3, saltpeter) was discovered, probably in China or India. Fireworks were reported in China in the twelfth century. In England, Roger Bacon made the first pure KNO_3 and mixed it with charcoal and sulfur to make a form of black powder in about 1242. Berthold Schwarz, a German Monk, is credited with using black powder as a gun propellant in about 1300. A formula for black powder (in the weight ratio 75.0/15.62/9.38 wt %KNO_3/charcoal/sulfur, virtually the same mixture used today) was published in Brussels in 1560. This material began to be used as a blasting agent in mines in the 1600s and its use dramatically increased during the next 300 years.

Development of molecular explosives (materials in which the fuel and oxidizer are contained in a single molecule) started in 1846 when Sobrero (a professor at the University of Turin in Italy) discovered nitroglycerin (NG). It was promisingly powerful but dangerously unpredictable. When Immanuel Nobel and his son Alfred built a NG plant in Sweden in 1862 and invented an igniter for initiation in 1863, NG began to be used for blasting applications; this use was accompanied by severe accidents. By 1875 Nobel had learned that safety could be enhanced by absorbing the NG in *kieselguhr* (a porous siliceous earth) or by making a gelatin with a nitrated cotton material (nitrocellulose). These materials were called "dynamites" and gained almost immediate success in blasting operations.

TNT was developed in the late 1800s and became the main explosive used in World War I; it still is widely used today. TNT is relatively insensitive and melts at 79 °C so that it can be easily melted, mixed with other materials (such as aluminum), and then cast to a shape. More powerful materials, such as HMX (cyclotetramethylene tetranitramine), are now of great importance for weapons applications. Today, hundreds of molecular explosives are known, but only a relatively few have wide application.

After a disaster in 1947 in which two ships loaded with fertilizer-grade ammonium nitrate (AN) blew up in Texas City harbor, it became apparent that AN was an explosive. Shortly after, ANFO (a mixture of 94/6 wt % AN and fuel oil) was found to be a commercially valuable explosive. Through its cheapness and ease of preparation, this material rapidly became a very important blasting agent and remains so today. A blasting agents industry has developed, based on mixtures of nitrates, fuels, and other materials, in which explosives (slurries, emulsions, etc.) are tailor made for particular applications. These materials account for the bulk of explosives used today.

The blasting industry worldwide uses billions of kilograms of explosive products every year in mining and construction. Some products are fabricated by the action of explosives on them, e.g., explosive forming of complex metal shapes, production of small industrial diamonds from graphite, dissimilar-metal welding, and fabrication of intermetallics. Explosive materials are used in the space industry as components of the propellant in rocket engines, as well as to control and time events on rocket systems. Explosive devices and materials are used to enhance safety, e.g., auto air bags, aircraft ejector systems, safety flares, and the common safety match. The oil industry utilizes explosives to perforate well casings deep in the ground and to create seismic disturbances so that underground oil reservoirs can be mapped. Explosive materials (pyrotechnics) are used to create brilliant fireworks displays.

The weapons industry relies almost completely on the action of explosives for destructive purposes. Even nuclear weapons rely on specialized chemical high explosives and explosive components to create the proper conditions so that a nuclear explosion can occur. A variety of materials are used for igniters, rocket propellants, gun propellants, shaped charges, bomb fills, flares, etc.

Researchers, particularly physicists, use special explosive devices to create shock waves in inert materials so that states of matter (such as the conditions of pressure and temperature that occur deep in the earth) can be studied. Explosive action on specially designed electrical inductor systems can lead to extremely high electrical power outputs and magnetic fields that can be used to develop one-shot–device plasma guns, x-ray heads, or railgun particle acceleration systems. Transient light sources (useful for illuminating short time-scale events) a hundred times brighter than the sun can be produced by explosively generated shock waves in inert gas.

1.2 Appearance of a Quantitative Picture of Detonation

The first experimental observations (ca. 1880) of detonation waves (in gases) were made by French workers (Berthelot, Vielle, Mallard, and Le Chatelier). Later Chapman and Jouguet, independently, gave a theoretical treatment of detonation in gases (ca. 1900). In the Chapman-Jouguet (CJ) treatment, a detonation is idealized as a planar mathematical pressure discontinuity (shock wave) that propagates steadily. In the CJ picture, the passage of the shock causes complete release of the stored chemical energy; i.e., there is no spatially resolved zone of chemical reaction. The fluid mechanics required to devise the CJ picture comprises primarily the mass, momentum, and energy conservation-law jump conditions across a shock wave (the Rankine-Hugoniot conditions). An important result of the CJ treatment is that, given an energetic material, there is a minimum velocity at which a steady plane detonation wave can be propagated through it. The CJ theory is discussed in Sec. 2.3.

Later (ca. 1940), Zel'dovich, von Neumann, and Doering (ZND), independently, refined the CJ model by assuming that the conservation conditions of mass, momentum, and energy for inviscid flow (the Euler equations) apply, and by relaxing the assumption of an instantaneous heat release triggered by the shock. They modeled the chemical heat release as a single forward exothermic rate process. With the assumption that the chemical energy is released at a finite rate, one finds that a steady planar solution of the flow and shock-jump equations exists. This (ZND) detonation consists of a shock wave followed by a steady zone of chemical reaction. The hydrodynamic state at the end of the steady reaction zone is the CJ state, and the flow variables there are independent of the rate form. The fluid-mechanical state within the ZND reaction zone is *dependent* on the rate form. The ZND theory is discussed in Sec. 2.4.

Quite early on, it was experimentally recognized that detonations exist that show behavior significantly more complex than that described by either the CJ or ZND models. For instance, real detonation waves have (e.g.)

1. time dependence in the way they approach steadiness (see Sec. 4.1),
2. dependence on the lateral size of the explosive charge, even when the detonation is traveling steadily (see Sec. 4.2), and
3. inherent longitudinal and transverse instabilities (see Fickett and Davis, 1979, and Dremin *et al.*, 1970).

All of these complications have been studied theoretically and experimentally.

2. PLANAR STEADY DETONATION THEORY

The purpose of this section is to help the reader form an initial picture of what a detonation is. We give a quantitative treatment of a detonation as a shock-induced exothermic chemical reaction in which the chemical reaction, in turn, drives the shock wave. The reader should note that our treatment is for laminar steady detonation. The picture so derived is a useful reference point for interpreting the more complex space- and time-dependent cases discussed in Secs. 3–5. The laminar steady theories are particularly valuable because the mathematics required to obtain them is straightforward and because of the rarity of analytical solutions in more complex situations.

The simplicity of the theory of laminar steady detonation results from the simple form that the conservation relations for mass, momentum, and energy take in this case. These conservation relations are obtained in Secs. 2.1 and 2.2 and then used to obtain the CJ and ZND theories in Secs. 2.3 and 2.4.

2.1 Hydrodynamic Jump Conditions for Steady Laminar Flow and Across Discontinuities

The fluid flows characteristic of detonation are constrained by the conservation of mass, momentum, and energy. Therefore to discuss such processes quantitatively, we define the relations between mass, momentum, and energy at two locations (stations) in a *steady* planar compressible flow. We will use the term one dimensional (1D) to mean *planar* one dimensional. The fluid flow is viewed from a Galilean reference frame in which matter flows from left to right (see Fig. 3). The pertinent flow quantities at station 1 are the velocity at which matter crosses the boundary (U_s), the mass density (ρ_0), the pressure (P_0), and the internal energy per unit mass (E_0); the analogous quantities at station 2 are $U_s - u_p$, ρ, P, and E. The quantity u_p is the change in the mass velocity of the fluid as it undergoes the fluid-mechanical processes that occur between stations 1 and 2. Note that if the process causing the flow is a detonation wave moving from right to left (at speed U_s) in the laboratory frame, then the Galilean frame used to define stations 1 and 2 and to derive Eqs. (1)–(3) below is the one in which the wave is stationary; i.e., we have transformed from the laboratory frame to one in which the wave is motionless and matter streams through the plane defined by the wave. As drawn in Fig. 3, moving from station 1 to station 2 corresponds to a pressure increase and a mass-velocity decrease. As will be seen, this corresponds to stations 1 and 2 being in the unreacted and detonating material, respectively.

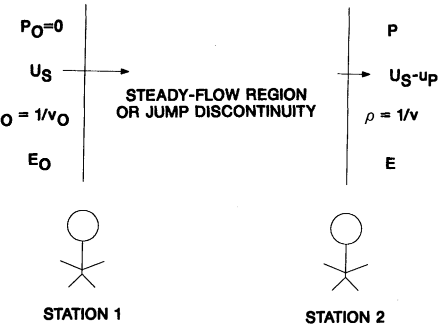

FIG. 3. Locations (stations) and quantities used in obtaining the conservation conditions in steady flow or across discontinuities.

Below we assume a special case where the pressure at station 1 is negligible relative to other pressures in the flow (i.e., $P_0 = 0$). As will be shown, typically in condensed-phase detonation $P_0/P \approx 10^{-5}$ when $P_0 = 1$ atm, and so this is a very good approximation.

The mass conservation condition under the stated conditions is that the mass flux at stations 1 and 2 must be equal, i.e.,

$$\rho_0 U_s = \rho(U_s - u_p). \qquad (1)$$

The momentum flux at a station is given by the mass flux there times the velocity at which matter crosses the boundary. This velocity is called the particle velocity, u_p. For stations 1 and 2, the values of momentum flux are $\rho_0 U_s^2$ and $\rho(U_s - u_p)^2$, respectively. The change in momentum per unit time per unit area between the two stations must be balanced by the pressure difference between the two stations, i.e.,

$$-P = \rho(U_s - u_p)^2 - \rho_0 U_s^2. \qquad (2)$$

Similarly, the change in flux of internal energy plus kinetic energy (per unit area per unit mass) between the two stations must be

12balanced by the rate at which the pressure difference between the two stations is doing work, i.e.,

$$-P(U_s-u_p)=\rho(U_s-u_p)[E+\tfrac{1}{2}(U_s-u_p)^2] -\rho_0U_s[E_0+\tfrac{1}{2}U_s^2]. \qquad (3)$$

Note that the validity of Eqs. (1)–(3) does not depend on the station spatial separation. Thus, they can be applied to a flow discontinuity by allowing the separation between the two stations to approach zero. Used in this way they define the jump conditions (called the Rankine-Hugoniot conditions) across a detonation or shock wave.

2.2 Rayleigh Line and Hugoniot Curve

Relationships that a steady laminar detonation flow must obey can be obtained from Eqs. (1) to (3) by elimination of u_p and U_s. For example, eliminating u_p from Eqs. (1) and (2) gives

$$P=K(v_0-v), \qquad (4)$$

where $v\equiv 1/\rho$ is the volume per unit mass (specific volume) and $K\equiv\rho_0^2U_s^2$ is a parameter. Equation (4) is a straight line in P,v space. Two such lines, with different K values (slopes), are shown in Fig. 4; note that at fixed ρ_0, a larger slope corresponds to a larger wave speed. All 1D steady flow or flow across jump discontinuities [with initial state ($P_0=0$, v_0)] must have P, v values that satisfy Eq. (4) to conserve mass and momentum. Such a set of P,v points is called a Rayleigh line. The Rayleigh-line concept will prove useful below in defining what detonation processes are possible.

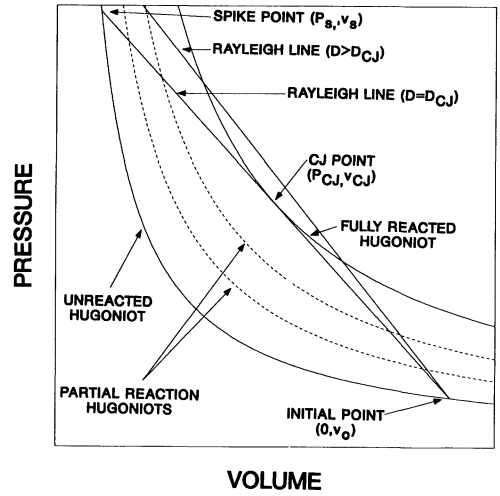
SPIKE POINT (P_s,v_s)
RAYLEIGH LINE ($D>D_{CJ}$)
RAYLEIGH LINE ($D=D_{CJ}$)
CJ POINT (P_{CJ},v_{CJ})
FULLY REACTED HUGONIOT
PRESSURE
UNREACTED HUGONIOT
PARTIAL REACTION HUGONIOTS
INITIAL POINT (0,v_0)
VOLUME

FIG. 4. Pressure-volume plane quantities useful in the discussion of the CJ and ZND models of detonation.

If both U_s and u_p are eliminated from Eq. (3) by use of Eqs. (1) and (2), one finds that

$$(E-E_0)-\tfrac{1}{2}P(v_0-v)=0. \qquad (5)$$

Thus, for 1D steady processes or for flow across shocks that start from the initial point ($P_0=0$, v_0, E_0), the final state must lie on the curve defined by Eq. (5), in order to conserve mass, momentum, and energy. If the process connecting the initial and final states is a single shock process, this E,P,v, relationship is called the Hugoniot curve of the shocked material. The Hugoniot curve defines all possible final shock states reachable from an initial state by a single shock process. The curve labeled "unreacted Hugoniot" in Fig. 4 is such a locus.

2.3 Chapman-Jouguet Theory

The Rayleigh line and Hugoniot curve are two equations in the three variables P, v, and E. A third independent equation $E=E(P,v)$ (i.e., an equation of state) closes this system and allows one to solve for permissible E, P, and v values in terms of the other parameters in the equations. This is how one obtains the Chapman-Jouguet theory of detonation.

In order to make the discussion of the CJ theory concrete we will specialize to a material in which the internal energy per unit mass is given by

$$E(P,v,\lambda)=Pv/(\gamma-1)-q\lambda, \qquad (6)$$

where γ is a constant characteristic of the material, q is the chemical heat stored in the material per unit mass, and λ is a parameter that measures the proportion of the chemical energy that has been released into the flow; $\lambda=0$ and 1 correspond to no and complete chemical reaction, respectively. The $Pv/(\gamma-1)$ part of Eq. (6) is called a "γ-law" equation of state (EOS). For condensed-phase explosives, a γ value of about 3 gives a reasonable description of many observed condensed-phase detonation phenomena. It is important to note that a γ-law EOS is not, necessarily, a "perfect gas" EOS. This can be seen by noting that for a perfect gas, γ (the ratio of specific heats, $C_p/C_v) = 1+2/n$, where n is the number of degrees of freedom available to the particles composing the gas. Thus, since $3 \leqslant n < \infty$, γ is constrained by $\frac{5}{3} \geqslant \gamma > 1$ for a perfect gas. The values $\gamma \approx 3$ useful in condensed-phase detonation allow the gas to be "stiffer" than is possible for a perfect gas.

Now suppose that a steady planar detonation wave moves at velocity D through a material defined by Eq. (6) and, in so doing, instantaneously releases the chemical heat present. The material before shock passage is at zero pressure and specific volume v_0, and is completely unreacted ($\lambda = 0$), i.e., $E_0 = E(0, v_0, 0) = 0$. After shock passage, the material is at pressure P, specific volume v, and completely reacted ($\lambda = 1$), i.e., $E = E(P, v, 1) = Pv/(\gamma - 1) - q$. The question can now be asked: What values of P and v do the conservation laws allow? Using the values of E and E_0 corresponding to before and after detonation wave passage in Eq. (5) gives

$$Pv/(\gamma-1) - q - \tfrac{1}{2}P(v_0 - v) = 0. \tag{7}$$

Equation (7) and the Rayleigh line [Eq. (4)] must both be satisfied if the conservation laws are obeyed.

Elimination of P from Eq. (7) by use of the Rayleigh-line condition yields a quadratic equation for the specific volume of the possible detonation states, i.e.,

$$v^2 - Bv + C = 0, \tag{8}$$

where $B = 2\gamma v_0/(\gamma+1)$ and $C = (\gamma-1)[1+2q/D^2]v_0^2/(\gamma+1)$. The value of $B^2 - 4C$ determines the number of real roots of Eq. (8). At this point, the value of D (the detonation wave velocity) is a free parameter in C, related to the slope of the Rayleigh line. When $D = \sqrt{2q(\gamma^2-1)}$, $B^2 - 4C = 0$ and Eq. (8) has one real root. For D larger (smaller) than this value, Eq. (8) has two (zero) real roots. The CJ detonation velocity is obtained if one assumes there must be a unique solution to Eq. (8). Geometrically this uniqueness of detonation velocity corresponds to tangency of the Rayleigh line and the completely reacted Hugoniot of the material (see Sec. 2.4).

At this point, the conservation conditions have given the important result that for $D < \sqrt{2q(\gamma^2-1)}$, a steady 1D detonation wave is impossible. When $D = \sqrt{2q(\gamma^2-1)}$, there is a unique value of the detonation wave velocity—this is the CJ value (D_{CJ}). It will be shown below that for a wave traveling at D_{CJ}, the flow at the CJ point is sonic relative to the following flow. Consequently, such a wave is immune from hydrodynamic disturbances from the rear (other than shocks).

Next consider the case where $D > D_{\mathrm{CJ}}$; in this situation there are two solutions. The solution with $P > P_{\mathrm{CJ}}$ (see Fig. 4) corresponds to a detonation wave that moves subsonically relative to its following flow and therefore can be attenuated by rarefactions from the rear. This solution is rejected as being uninteresting for this reason. The solution with $P < P_{\mathrm{CJ}}$ travels supersonically relative to the following flow. More will be said about this type of solution in the next section.

The solution of Eq. (8) for the specific volume at the CJ state is $B/2$; i.e., $v_{\mathrm{CJ}} = \gamma v_0/(\gamma+1)$. Substitution of v_{CJ} into the Rayleigh-line equation yields the CJ pressure $P_{\mathrm{CJ}} = \rho_0 D_{\mathrm{CJ}}^2/(\gamma+1)$. For $\gamma = 3$, one sees that $v_{\mathrm{CJ}}/v_0 = \frac{3}{4}$; i.e., the volume per unit mass is reduced by 25% for a γ-law gas at the CJ state relative to the uncompressed state. Typical measured values of ρ_0 and D_{CJ} for condensed-phase explosives are 1.80 g/cm^3 and 8.0 km/s. Use of these values (and $\gamma = 3$) in the P_{CJ} relation yields a CJ pressure of about 29 GPa (290 kbar); as a comparison, the yield strength of common steels is about 0.5 GPa. For these detonation parameters, *1 MW of power* is generated by a 1.9-cm^2 area on the detonation front.

Using results of the last paragraph, one can prove that the flow at the CJ point is sonic. A few other facts are needed for the proof. Firstly, note that if U_s is eliminated between Eqs. (1) and (2), one finds that $u_p = \sqrt{P(v_0 - v)}$. Use of P_{CJ} and v_{CJ} in this equation shows that $u_{\mathrm{CJ}} = D/(\gamma+1)$. Next note that the sound speed in a fluid is given by $c = \sqrt{(\partial P/\partial \rho)_S}$ (the S subscript denotes constant entropy) and that for a γ-law EOS the isentropes are given by $Pv^\gamma = \text{constant}$. Therefore $c_{\mathrm{CJ}} = \sqrt{\gamma P_{\mathrm{CJ}} v_{\mathrm{CJ}}} = \gamma D/(\gamma + 1)$, so that $u_{\mathrm{CJ}} + c_{\mathrm{CJ}} = D$; i.e., a sound signal at the CJ plane travels exactly at the detonation wave speed. We have proved the sonic characteristic of the flow at the CJ point for a constant-γ EOS. It is important to note that the sonic character of the flow at the CJ point is a general result and is true for any reasonably behaved EOS (Fickett and Davis, 1979).

2.4 Zel'dovich–von Neumann–Doering Theory

It is self evident that the CJ model assumption of an instantaneous chemical heat release is, at best, an approximation. Zel'dovich, von Neumann, and Doering (ZND) advanced the theory of steady 1D detonation beyond the CJ theory by finding a solution of the flow equations with a *resolved* chemical reaction zone (see, e.g., von Neumann, 1963). In this section, we derive the ZND model.

In the ZND model one assumes that the detonation process consists of a shock wave that takes the unreacted material from its initial state $(P,v,\lambda)=(0,v_0,0)$ to a "von Neumann spike" point $(P_s,v_s,0)$ that is still on the unreacted Hugoniot (see Fig. 4). Since the chemistry is now resolved, there is a continuum of partial reaction Hugoniots between the unreacted and fully reacted Hugoniots corresponding to different values of λ. For the γ-law form [Eq. (6)] these Hugoniots are given by

$$[Pv/(\gamma-1)]-\lambda q-\tfrac{1}{2}P(v_0-v)=0, \qquad (9)$$

where the value of the variable λ determines which partial reaction Hugoniot is being considered. In traversing the ZND reaction zone, one proceeds down the Rayleigh line from the spike point $(P_s,v_s,0)$ to the CJ point $(P_{\mathrm{CJ}},v_{\mathrm{CJ}},1)$—crossing the full set of partial reaction Hugoniots in the process (see Fig. 4). The q in Eq. (9) is the difference in enthalpy between reactants and products at the CJ state and is called the "heat of detonation" (see Sec. 5.2). Note that the chemical heat release in Eq. (9) is assumed to be governed by a single chemical progress variable (λ), and increased reaction always corresponds to increased exothermicity.

The fluid mechanics already discussed is sufficient to describe this process in detail because the flow in the reaction zone is steady. Thus, a space point in the unshocked material and any space point in the steady reaction zone constitute satisfactory stations 1 and 2 of Fig. 3—so the Rankine-Hugoniot conditions obtained in Sec. 2.1 are applicable. The (P,v,λ) state in the reaction zone corresponding to a particular λ is found by solving Eq. (8), but now with C defined as

$$C(\lambda)=\frac{(\gamma-1)}{(\gamma+1)}\left(1+\frac{2q\lambda}{D^2}\right)v_0^2. \qquad (10)$$

The detonation-wave velocity is determined from the tangency of the Rayleigh line and the *fully reacted* Hugoniot, i.e., again $D=\sqrt{2q(\gamma^2-1)}$, as in the CJ theory. The specific volume as a function of λ [$v=v(\lambda)$] can be obtained from Eqs. (8) and (10). The pressure through the reaction zone as a function of λ is obtained by use of $v=v(\lambda)$ in the Rayleigh-line equation [Eq. (4)].

We still do not know the reaction-zone structure in space or time [e.g. $P=P(x)$]. To proceed further one needs to know how the chemical heat is released as a function of the thermodynamic variables, i.e., a chemical rate relation of the form

$$\frac{d\lambda}{dt}=r(P,v,\lambda). \qquad (11)$$

By use of what has been found above, Eq. (11) can be written as $d\lambda=r(P(\lambda),v(\lambda),\lambda)dt$; integrating gives $\lambda=\lambda(t)$. Given $\lambda=\lambda(t)$, $P=P(t)$ and $v=v(t)$ are known, since $P(\lambda)$ and $v(\lambda)$ are known. The dependence of P and v on the spatial coordinate x is determined as follows. The spatial position of a fluid particle relative to the detonation shock is given by

$$\frac{dx}{dt}=D-u_p. \qquad (12)$$

We know from Sec. 2.3 that $u_p=\sqrt{P(v_0-v)}$, and we know $P(t)$ and $v(t)$ from the above work. Therefore Eq. (12) can be integrated to give a function $t=t(x)$, which, in turn, can be used in $P(t)$ and $v(t)$ to yield the functions $P(x)$ and $v(x)$, i.e., the space dependence of P and v through the ZND reaction zone. A typical particle-velocity snapshot of the ZND reaction-zone profile obtained through this process is shown in Fig. 5.

A striking physical result of the ZND theory shown on Fig. 5 is that the pressure *falls* through the ZND reaction zone. This result is implicit in Fig. 4, i.e., as one moves across the partial-reaction Hugoniots from $\lambda=0$ to 1, the pressure falls.

Consider the ZND picture for a Rayleigh line with $D>D_{\mathrm{CJ}}$; the solution with $P>P_{\mathrm{CJ}}$ is again rejected because the flow at the point of complete reaction is subsonic relative to the following flow. The solution with $P<P_{\mathrm{CJ}}$ cannot be reached from the spike point by moving down the Rayleigh line because of the absence of partial-reaction Hugoniots at some

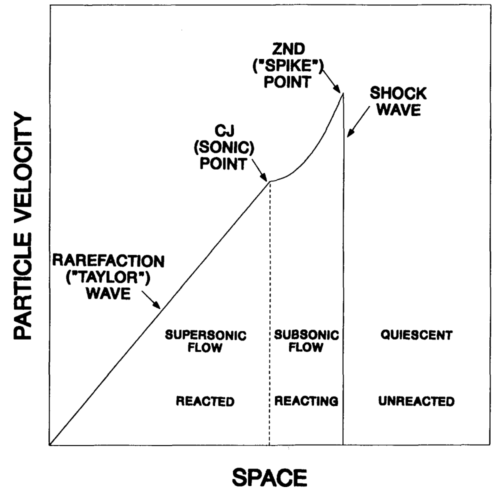

FIG. 5. Schema of the flow in a ZND detonation in the particle-velocity versus space plane; the Taylor wave is also shown (see Sec. 3.5).

points on the path and is rejected for this reason. More general types of the heat-release function (e.g., with reversibility of reaction or reactions with both exothermic and endothermic parts) can make the point with $P < P_{\mathrm{CJ}}$ accessible (see Fickett and Davis, 1979).

The main features of the ZND model are borne out by experiment; e.g., direct measurement of the particle velocity versus time in detonating explosives shows a particle-velocity discontinuity (shock front) followed by a zone of decreasing particle velocity (see Fig. 6). X-radiography of the explosive products have been used to show that the flow behind the shock does become sonic. This latter feature was done by tracking strong rarefaction waves introduced into the reaction products.

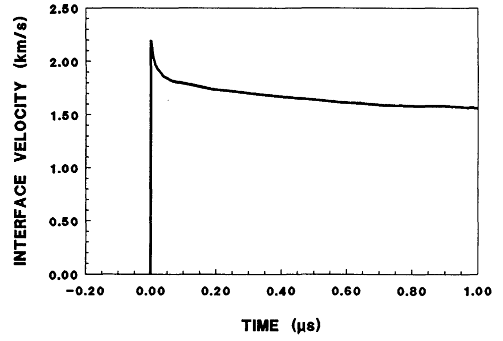

FIG. 6. Experimentally measured interface velocity history showing the ZND reaction-zone structure. See Sec. 5.4 for details of the experiment.

Various other detonation properties can also be measured, e.g., spike pressure, detonation velocity, etc. Methods of making such measurements are discussed in Sec. 5.4.

3. EQUATIONS GOVERNING REACTIVE FLOW

In Sec. 2 a description was given of 1D steady detonation theories where the chemical-heat addition either is unresolved in time or is especially simple in form; i.e., one forward exothermic rate process. The essentially algebraic form of the CJ and ZND theories must be generalized when more complex detonation phenomena are considered. Here, "complexity" can mean, for example, multidimensionality in the flow (e.g., steady two-dimensional detonation—see Sec. 4.2) or flow in which the shock wave moves unsteadily (e.g., the shock initiation of detonation—see Sec. 4.1). These more involved situations are what are met most often in the experimental situation.

The more general situation is governed by the applicable partial differential equations of fluid mechanics (with initial and boundary conditions) and material constitutive relations. The fluid-mechanical equations governing condensed-phase detonation are usually taken to be the Euler equations of inviscid compressible flow with chemical reaction—augmented by the suitably generalized shock-jump conditions if the flow is multidimensional. The Euler equations govern compressible fluid flow in which transport processes are neglected. The equations are statements of the conservation of mass and energy together with a dynamical (momentum) equation that governs the particle velocity ($\mathbf{u}$) field.

In this section, we describe these equations and then apply them to a few situations to illustrate their usefulness.

3.1 Fluid-Mechanical Equations

As a preliminary to consideration of the Euler equations, it is important to note that in a compressible flow, the time derivative of a flow-field variable [say $F \equiv F(x,y,z,t)$] on a fluid particle is given by the total time (or material) derivative of that quantity evaluated on the particle's path, i.e.,

$$\frac{dF}{dt}=\left[\frac{\partial}{\partial t}+(\mathbf{u}\cdot\nabla)\right]F, \tag{13}$$

where **u** is the velocity vector of the particle. Equation (13) is derived by taking the total derivative of *F* (via the chain rule) and then evaluating the result on a particle path. This result will be needed below where we are interested in defining the dynamical behavior of any particular fluid particle.

When the notation of Eq. (13) is used, the three applicable fluid-mechanical conditions, written in the laboratory reference frame, are as follows. The mass conservation condition for a particle of the fluid is

$$\frac{d\rho}{dt}+\rho\nabla\cdot\mathbf{u}=0. \tag{14}$$

In Eq. (14), the second term results from the mass density being dependent on the fluid particle's volume; the $\rho\nabla\cdot\mathbf{u}$ term accounts for dilation of the fluid particle as the flow evolves. The momentum and energy equations, given next, are dependent on the fluid particle mass and, consequently, do not contain dilation terms. The next equation is the expression of Newton's second law for a continuous inviscid compressible fluid; i.e.,

$$\rho\frac{d\mathbf{u}}{dt}=-\nabla P, \tag{15}$$

where *P* is pressure at the particle's location. When $\nabla P=0$, Eq. (15) expresses the conservation of linear momentum. Note that Eq. (15) is a nonlinear partial differential equation, since the operator d/dt depends on **u** [see Eq. (13)].

The energy conservation condition is

$$\frac{dE}{dt}=-P\frac{dv}{dt}, \tag{16}$$

where *E* and *v* are the internal energy per unit mass and the volume per unit mass, respectively. Equation (16) mirrors the adiabatic nature of the flows being treated. It shows that energy transport occurs only via mechanical work done by or on a particle.

The chemical-heat release in a fluid particle is specified in *E* via dependence on chemical "progress" variables.

3.2 Chemical-Heat-Release Relations

The progress-variable (λ) evolution is specified by "rate laws" that govern their time dependence on the state variables of the system. Such a form is

$$\frac{d\overline{\lambda}}{dt}=r(P,v,\overline{\lambda}), \tag{17}$$

where, in general, $\overline{\lambda}$ is an *n*-dimensional vector.

The energy function that drives the fluid mechanics is the global heat released (Q) into the flow. Thus, when a number of chemical processes are occurring simultaneously, the pertinent mechanical energy is the sum of the individual energy releases q_i, i.e.,

$$Q=\sum_{i=1}^{n}\lambda_i q_i. \tag{18}$$

The difficulty in attempting to probe microscopic energetic events in a detonation via mechanical measurements is inherent in the ambiguities associated with Eq. (18). To probe the microscopic processes that produce the global heat-release function *Q*, one needs to examine events at the microscopic level—e.g., with spectroscopic techniques. This type of investigation is a current area of research in detonation physics and chemistry.

3.3 The Material Equation of State

Equations (14)-(17) are $5+n$ differential equations in $6+n$ dependent variables. Thus, this system of equations is underdetermined. One further constraint involving the $6+n$ dependent variables is required.

The constraint that closes the system is the equation of state (EOS) of the reactive fluid, i.e., a relationship of the form

$$E=E(P,v,\overline{\lambda}); \tag{19}$$

see, e.g., Eq. (6).

We note that since one is dealing with mechanical equations, it is not required that the system temperature be a variable in the EOS; i.e., a full $P(v,T,\overline{\lambda})$ EOS is not required. However, if the chemical-heat-release relations of Sec. 3.2 depend on temperature, this variable must also be known in order to define fully the reactive flow problem. Defining the temperature field in detonations is a current area of research. It is implicit in the above discussion that there is local thermodynamic equilibrium in the system, except for the chemical progress variables. The rate laws [Eq. (17)] define their time evolution.

Shock-jump conditions and Eqs. (14)–(19) (with boundary and initial conditions) define the reactive flow problems found in detonating condensed systems, within the stated approximations. These governing equations are a hyperbolic system; hence a salient feature of their solutions is transport of energy by wave motion (Wood and Kirkwood, 1957).

The usefulness of these equations in the study of detonation phenomena will be illustrated in Secs. 3.4 and 3.5.

3.4 Qualitative Features of Flow with Reaction

Some insight into the effect of chemical-heat addition on a compressible flow can be obtained by eliminating dE/dt between the energy conservation condition and the EOS. From the EOS, we know that

$$\dot{E} \equiv \frac{dE}{dt} = \left[\frac{\partial E}{\partial P}\right]_{v,\lambda} \dot{P} + \left[\frac{\partial E}{\partial v}\right]_{P,\lambda} \dot{v} + \left[\frac{\partial E}{\partial \lambda}\right]_{v,P} \dot{\lambda}.$$

Eliminating $\dot{E}$ from this expression via $\dot{E} = -P\dot{v}$ and rearrangement gives

$$\dot{P} = \left[\frac{[\partial E/\partial v]_{P,\lambda} + P}{[\partial E/\partial P]_{v,\lambda}}\right] \frac{\dot{\rho}}{\rho^2} - \frac{[\partial E/\partial \lambda]_{P,v}}{[\partial E/\partial P]_{\lambda,v}} \dot{\lambda}. \qquad (20)$$

By writing the 1D Euler equations in characteristic form, Wood and Kirkwood (1957) showed that the coefficient of $\dot{\rho}$ in Eq. (20) is the local sound speed ($\equiv c$) squared—at fixed chemical composition, i.e., fixed λ. This speed is called the "frozen" (composition) sound speed. The second term in Eq. (20) depends on the chemical-heat-release rate $\dot{\lambda}$ ($\equiv r$). Use of standard partial derivative relations allows it to be rewritten as

$$-\left(\frac{\partial P}{\partial \lambda}\right)_{E,v} \dot{\lambda} \equiv -P_\lambda \dot{\lambda}.$$

With these definitions, Eq. (20) can be written as

$$\dot{P} = c^2 \dot{\rho} + P_\lambda \dot{\lambda}. \qquad (21)$$

The quantity $P_\lambda \dot{\lambda}$ is a measure of the local rate of chemical-heat addition. For example, for the energy relationship defined by Eq. (6), $P_\lambda \dot{\lambda} = [q(\gamma-1)/v]\dot{\lambda}$. When $P_\lambda \dot{\lambda}$ is positive, Eq. (21) shows that (at fixed $c^2\dot{\rho}$) $\dot{P}$ is larger than it would be without chemical heat addition.

One can use Eq. (21) to obtain an important relationship between the heat-addition rate and the sonic character of steady 1D chemically driven flow. In this case, Eq. (21) specializes to $uP_x - c^2 u\rho_x = P_\lambda \dot{\lambda}$. For steady 1D flow, Eqs. (14) and (15) simplify to $u\rho_x = -\rho u_x$ and $\rho u u_x = -P_x$, respectively; these equations imply that $u\rho_x = P_x/u$. Substitution of this result into the specialized form of Eq. (21) gives

$$\dot{P} = P_\lambda \dot{\lambda} / [1 - (c/u)^2], \qquad (22)$$

where the sign of the denominator of Eq. (22) determines the sonic character of the flow.

Interesting insights into the character of the flow behind a steady 1D detonation shock can be obtained from Eq. (22). For subsonic regions (e.g., in the ZND reaction zone) $c^2 > u^2$; and for $P_\lambda \dot{\lambda} > 0$, $\dot{P}$ is negative; that is, the pressure falls through the ZND reaction zone. An important feature of Eq. (22) is that at points where the flow is sonic ($u = c$), either P_λ or $\dot{\lambda}$ must be zero—otherwise $\dot{P}$ is infinite. For the simple EOS defined by Eq. (6), $P_\lambda [= q(\gamma - 1)/v]$ cannot be zero and so the only possible sonic point corresponds to complete reaction (i.e., $\dot{\lambda} = 0$ at $\lambda = 1$). This is the CJ point defined earlier. For more complex rate laws, P_λ can be 0 (e.g., for a rate containing both exothermic and endothermic parts), and sonic points other than the CJ point are possible. Such a situation can give rise to "weak" detonations in which $P < P_{\mathrm{CJ}}$ (see Fickett and Davis, 1979).

3.5 Taylor Wave

In Secs. 2.3 and 2.4, we derived the fluid-mechanical state at the point of complete chemical reaction in the CJ and ZND models—i.e., the "CJ point" values of the flow field. A natural question to ask is, "What fluid flow connects the CJ point to the final low-density and -pressure state of the expanded reacted gases?" Here we answer this question by use of the compressible-flow equations of Sec. 3.1. The fluid flow that connects these two states is called the "Taylor wave" after G. I. Taylor as explained below.

First note that in a fluid describable by the Euler equations (and in which there are no chemical reactions occurring or shock waves present), the thermodynamic states experienced by an initially isentropic fluid must lie on an isentrope (because shock waves or chemical reactions are the only possible sources of entropy production). The fluid that traverses the CJ plane of a detonation is fully reacted, and we assume that the subsequent flow has no discontinuities; therefore the state of this material lies on an isentrope. We assume that the reacted fluid is describable by a γ-law EOS [see Eq. (6)]; then the isentropes are given by $P=C\rho^{\gamma}$, where C is a constant. In this case, the Euler equations governing the density [Eq. (14)] and momentum [Eq. (15)] are

$$\rho_t+u\rho_x+\rho u_x=0 \quad \text{and} \quad u_t+uu_x+\gamma C\rho^{\gamma-2}\rho_x=0,$$

respectively.

Since there are no pertinent space or time scales in the flow following the CJ point, Taylor (1950) assumed that the flow can be described via the self-similarity variable $z\equiv x/t$. When the density and momentum partial differential equations are expressed in terms of z, they become the ordinary differential equations (ode) $(u-z)\rho'=\rho u'$ and $(u-z)\times u'=\gamma C\rho^{\gamma-2}\rho'$, where the prime denotes the ordinary derivative with respect to z. Division of one of these ode's by the other results in the differential relation $u'=(\gamma C\rho^{\gamma-3})^{1/2}\rho'$. This can be integrated to yield

$$u=\frac{D}{\gamma-1}\left[\frac{2\gamma}{\gamma+1}\left(\frac{\rho}{\rho_{\mathrm{CJ}}}\right)^{(\gamma-1)/2}-1\right], \tag{23}$$

where the known values at the CJ point of $u_{\mathrm{CJ}}=D/(\gamma+1)$, $c_{\mathrm{CJ}}=\sqrt{\gamma P_{\mathrm{CJ}}v_{\mathrm{CJ}}}=\gamma D/(\gamma+1)$, and $\rho_{\mathrm{CJ}}=\rho_0(\gamma+1)/\gamma$ have been used to evaluate the integration constant and to simplify the expression. Equation (23) can be used in the equation $(u-z)u'=\gamma C\rho^{(\gamma-2)}\rho'$ to find the dependence of ρ on z. One finds two cases: a trivial solution $\rho'=0$ and

$$\rho(x,t)=\rho_{\mathrm{CJ}}\left[\frac{(\gamma-1)x/t+D}{\gamma D}\right]^{2/(\gamma-1)}. \tag{24}$$

Use of Eq. (24) in Eq. (23) gives the particle velocity flow field

$$u(x,t)=\frac{1}{\gamma+1}(2x/t-D). \tag{25}$$

Figure 5 shows a schema of the $u(x,t)$ Taylor wave solution. Note that if the back boundary is free, Eq. (24) shows that the density and pressure are zero along the $x-t$ line $x/t=-D/(\gamma-1)$. Thus, for a $\gamma=3$ material, the speed of free expansion of the reacted gas is one-half the detonation speed.

We now have obtained a quantitative picture of a 1D steady detonation as a shock wave followed by a steady space-dependent reaction zone that ends at a sonic point. The sonic point is, in turn, followed by a self-similar flow that takes the flow field variables to ambient values (see Fig. 5).

3.6 Solutions of Reactive Flow Equations for Complex Systems

There are a number of difficulties associated with obtaining solutions of Eqs. (14)–(19) for real physical systems. A mathematical obstacle is that the flow equations are nonlinear (in general, partial) differential equations, and so the solution techniques available for linear systems are not available. This difficulty is usually avoided by solving the set of equations numerically (see Mader, 1979).

A more fundamental physical problem involves Eqs. (17) and (19). These equations require specification of constitutive properties of dense systems consisting of mixtures of many molecular species at extreme conditions of pressure, density, and temperature. There have been various attempts at obtaining the EOS of explosive-reaction products by use of statistical-mechanical techniques. Such an approach is to be contrasted with the use of an experimentally calibrated form such as Eq. (6). A recent example of the statistical-mechanical attack on the EOS problem is that given by Ree and van Thiel (1985); also see references therein.

Specification of the chemical-heat–release function(s) [Eq. (17)] in condensed-phase explosives is in a primitive state. Usually a mathematically simple rate form containing adjustable parameters is postulated and its parameters calibrated to the results of experiments on real systems. This calibrated rate can then be used with the other governing equations to make predictions of behavior of systems “closely” related to the calibration. Some experimental spectroscopic work in probing the microscopic chemical events in detonations has been done recently. Such studies are in their embryonic phase.

Much theoretical work has been done on numerical solution of the flow equations for multidimensional systems in an effort to understand the initiation and detonation regimes. The interested reader can find discussions of such work in Mader (1979), Nunziato *et al.* (1982), and the proceedings of the various Detonation Symposia.

4. EXPERIMENTAL ASPECTS OF INITIATION AND STEADY DETONATION

Initiation and detonation of explosives depend on the physical nature of the materials, their density, the dimensions of the explosive, and the surrounding confinement materials. In terms of their physical nature, explosives are generally grouped into homogeneous and heterogeneous materials. Homogeneous materials are typically liquids or single crystals of molecular explosives with no physical imperfections (e.g., bubbles, voids, etc.) that can cause perturbations in the input shock and the flow behind it. Homogeneous materials viewed with the macroscopic probes characteristic of detonation physics experiments appear uniform.

Heterogeneous explosives are generally all other types; these are usually pressed, cast, or extruded into the shapes desired. Here, heterogeneous means a material that contains any kind of imperfections that can cause fluid-mechanical irregularities ("hot spots") when a shock or detonation wave passes over them. Such hot spots cause associated space-time fluctuations in the thermodynamic fields (e.g., the pressure or temperature fields). These thermodynamic variations affect the local chemical heat-release rate. When averaged over a sufficiently large space scale, such variations convoluted with the underlying chemical rate(s) produce an average heat-release rate that is a combination of chemistry and mechanics. Examples of some conditions that cause hot spots are

1. voids collapsing by impact of one side on the other,
2. shock-wave propagation through irregular particles causing complex shock interactions,
3. shock-wave interactions between particles and voids leading to jetting,
4. plastic flow involving crystal breakage and shearing, and
5. shocking impedance mismatch between components of the explosive that cause stagnation points and shock reflections and interactions.

Some of these have been modeled using wave propagation computer codes (e.g., see Mader, 1979, and Nunziato *et al.*, 1982). Because of these hot spots, heterogeneous initiation and detonation phenomena are qualitatively different from that observed in homogeneous materials.

In Sec. 4.1, we discuss 1D shock initiation of detonation in homogeneous and heterogeneous explosives. Growth to detonation is qualitatively different in these two situations. Detonation can also develop from a burning or deflagration process in some explosive materials with quite particular material and confinement conditions; this initiation process will also be addressed. Steady 2D detonations in homogeneous and heterogeneous explosives have qualitatively different detonation velocity versus charge diameter relationships (diameter-effect curves); the reasons for this behavior will be discussed in Sec. 4.2.

4.1 Initiation of Detonation

As was shown in Sec. 2.4, detonation is a self-sustaining, shock-generated chemical reaction, and so any successful initiation process must eventually lead to shock formation. The most direct initiation method is to introduce a shock mechanically, e.g., with a thrown plate. If the resultant pressure input is sufficiently large, the shock will grow to detonation. This process has been labeled a shock-to-detonation transition (SDT). Shock inputs obviously have a particular duration, depending on how they are generated, and the duration may or may not be relevant to the initiation process. If the shock duration is long compared to the initiation time, it has no relevance. Durations of the same order as the initiation time or shorter can become important because pressure relief from the rear may quench the initiation reactions. Qualitative differences occur in SDT of homogeneous and heterogeneous materials, due to the hot-spot phenomena generated in the latter case. Both the homogeneous and heterogeneous SDT processes will be discussed in some detail.

When the pressure input is less than that required for direct SDT, two other notable processes can take place: a burning or deflagration-to-detonation transition (DDT), or what has been called a low-velocity detonation (LVD). DDT will be discussed in some detail but LVD will only be briefly described here.

LVD is a process completely different from steady laminar detonation. LVD is by its nature multidimensional, depending strongly on the material conditions and the surrounding confinement to create a condition conducive to low-level "hot-spot" reaction (e.g., a mechanical wave in a container of a liquid explosive can "precondition" the explosive by cavitating it; the cavitation bubbles can then become hot-spot loci activated by a low-pressure shock in the liquid explosive). LVD waves are in a sense constant velocity, but long propagation distances (several tens of centimeters) are seldom observed because the wave either dies out or makes the transition to ordinary detonation when slightly perturbed. LVD has been observed in liquid, cast, and pressed explosive charges but is usually associated with liquids or dynamite. LVD phenomena are a safety concern because initiation inputs are on the order of 0.1 GPa. Propagation velocities are about 2 km/s (one-third to one-fourth those of high-order detonation) with only a small portion of the available chemical energy consumed. LVD pressures are about 1.5 GPa or lower but can lead to fatal accidents (many accidents with nitroglycerin have been attributed to this phenomenon). See Johansson and Persson (1970) for additional information concerning this interesting phenomenon.

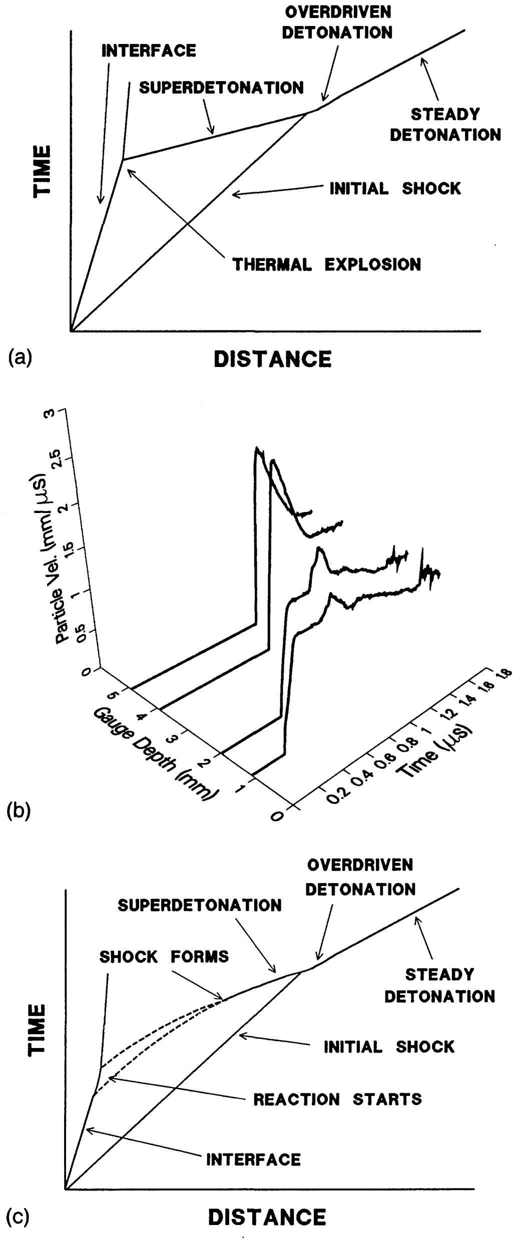

OVERDRIVEN DETONATION
INTERFACE
SUPERDETONATION
STEADY DETONATION
TIME
INITIAL SHOCK
THERMAL EXPLOSION
(a)
DISTANCE
Particle Vel. (mm/µs)
0
0.5
1
1.5
2
2.5
3
Gauge Depth (mm)
5
4
3
2
1
0
Time (µs)
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
(b)
OVERDRIVEN DETONATION
SUPERDETONATION
SHOCK FORMS
STEADY DETONATION
TIME
INITIAL SHOCK
REACTION STARTS
INTERFACE
(c)
DISTANCE

FIG. 7. (a) A time-distance diagram for the homogeneous initiation process as described by Campbell *et al.* (1961a). (b) The *in situ* particle-velocity measurements of the homogeneous initiation process that led to the diagram in (c) are shown. (c) A time-distance diagram for the homogeneous initiation process as described by Sheffield *et al.* (1989).

4.1.1 1D Homogeneous Initiation. Homogeneous SDT has been studied in detail by Campbell *et al.* (1961a), leading to the classical homogeneous initiation model shown in the time-distance diagram of Fig. 7(a). The explosive is shocked and, after an induction period τ (which depends on the initial shock pressure), a thermal explosion (runaway chemical reaction) occurs at the explosive-driver interface. This explosion produces a superdetonation (i.e., detonation in the explosive precompressed by the initial shock; this detonation has a higher pressure and velocity than a detonation in uncompressed material would have), which runs forward and eventually overtakes the initial shock. This produces an overdriven detonation condition which then decays to a steady detonation.

The simplest theory predicts the induction time (time to thermal explosion) using Arrhenius kinetics as

$$\tau = (C_v R T^2 / AQE) e^{E/RT}, \tag{26}$$

where C_v is the constant-volume specific heat, R is the gas constant, A is a collision factor, Q is the heat release, T is the shock temperature, and E is the Arrhenius activation energy. A number of assumptions, including no heat transfer, lead to this simplified form which comes from a more general theory developed by Frank-Kamenetskii (1955). Because of the exponential dependence of the Arrhenius rate on E, Eq. (26) indicates that the induction time is a very sensitive function of the temperature (provided $E \gg RT$); i.e., a small increase in temperature yields a large increase in reactivity.

It seems reasonable to suspect that a superdetonation does not form immediately after the thermal explosion as depicted in Fig. 7(a) but that rather some type of extended reaction-wave buildup process occurs. Recent *in situ* particle-velocity measurements of this growth process in chemically sensitized nitromethane have shown a developing reactive wave over relatively long time and distance scales [see Fig. 7(b)]. When these measurements are transformed into the time-distance plane, the trajectories of the various waves become those shown in Fig. 7(c). This modified homogeneous initiation model consists of the following processes:

1. The explosive is initially shocked, raising the pressure and temperature to the point that a chemical reaction starts;
2. this reaction produces mechanical waves that coalesce and strengthen;
3. a reactive compression or shock wave is formed within the material;
4. this reactive wave builds in strength (possibly to a steady superdetonation); and
5. this reactive shock overtakes the original shock wave, producing an overdriven condition that eventually settles down to a steady detonation (Sheffield *et al.*, 1989).

This picture is qualitatively different from SDT in heterogeneous materials that will be discussed next.

4.1.2 1D Heterogeneous Initiation. Heterogeneous SDT was also studied in detail by Campbell *et al.* (1961b), leading to much of the basic understanding that exists today. They showed that wave growth occurs at the front as well as behind it. This leads to a relatively smooth growth of the initiating shock to a detonation, in contrast to the abrupt changes observed in the homogeneous case. In this case, no overdriven detonation is observed, leading to the idea that the transition to detonation occurs at or near the shock front. Experiments on heterogenized nitromethane in which *in situ* particle-velocity histories were measured are shown in Fig. 8. These profiles agree with the heterogeneous initiation model of Campbell *et al.* (1961b).

The difference in behavior from that observed in homogeneous explosives is attributed to shock-induced hot spots forming at or near the shock front as a result of physical imperfections in the explosive and their consequent effect on the chemical-energy release into the flow (see Sec. 4). These hot spots can result from any phenomena that disrupts the smooth nature of the flow. On this basis, one could expect differences in initiation behavior to occur because of, for example, the particle size of material used to fabricate the explosive, and this is the case.

Little is known about the exact mechanisms that produce the shock-wave growth other than that the hot spots develop at the inhomogeneities, stimulating chemical-energy release locally and making the explosive initiate at much lower pressure inputs than would otherwise be the case. For example, pure liquid nitromethane can be made to initiate with a sustained shock input of about 8 to 9 GPa, while the same material with a large number of carborundum heterogeneities will initiate at 2 to 3 GPa (Campbell *et al.* 1961b). Many experimental studies have looked at hot-spot phenomena, but the size and nature of the inhomogeneities have not been sufficiently experimentally controlled to allow detailed understanding.

Because the hot spots produce reaction locally, heterogeneous explosives have a much less sensitive, state-dependent, average chemical-heat–release rate than homogeneous materials. The dependence on the bulk material pressure and temperature (all-important in the homogeneous case) is secondary.

4.1.3 Deflagration-to-Detonation Transition. Under some circumstances, transition to detonation can take place from much smaller stimuli than those required in a SDT process. When the initiation and confinement conditions are appropriate, relatively low thermal or pressure inputs start the explosive burning and this can eventually lead to detonation. This process is called deflagration-to-detonation transition (DDT). While SDT takes place on a time scale of a few microseconds, DDT takes place on a time scale 10 to 1000 times slower. DDT has been observed in both cast and pressed solid explosives. Experimental and theoretical research relating to DDT phenomena is active because of the safety concerns associated with small stimuli leading to accidental detonations. Heavy confinement is usually required for DDT.

DDT in low-density pressed materials (e.g., HMX) has been experimentally shown to involve the following phenomena (McAfee *et al.*, 1989). After the explosive starts to burn, the heavy confinement causes the pressure to increase with a consequent increase in the burning rate. The hot product gases move out into the pores of the unburned explosive and convective burning develops. This also drives a compaction wave ahead of the burning making it more difficult for the gases to penetrate the porous bed of material. This compaction wave compresses the material to about 90% of theoretical maximum density (TMD). As pressure builds up, stress waves are formed and coalesce, compacting the explosive to nearly 100% TMD, and a plug is formed in the material. As the combustion process gets near the plug, pressures increase and a strong shock forms near the front of the plug. When the shock amplitude increases sufficiently, a SDT process takes place in the 90% TMD material ahead of the plug. DDT in cast materials is thought to involve similar phenomena with the addition that the burning and compaction processes cause crushing of the material creating pores with resultant enhanced convective burning.

Baer and Nunziato (1989) have had considerable success in modeling DDT processes in a number of materials by using a two-phase flow model. Although many of the phenomena they calculate are similar to the experimental work in low-density HMX, careful experiments are needed to validate the physical and chemical assumptions of their model.

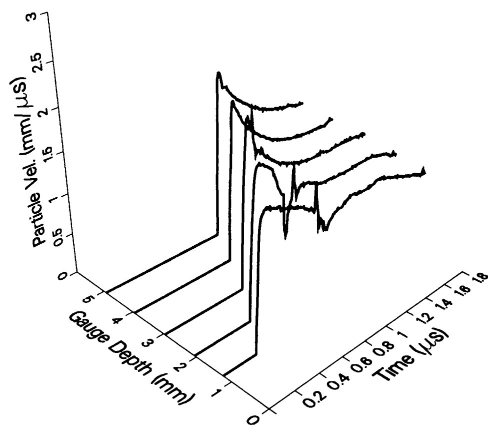

FIG. 8. *In situ* particle-velocity measurements of a heterogeneous initiation process in heterogenized nitromethane described by Sheffield *et al.* (1989).

In addition to heavy confinement, many explosive properties are important (e.g., density, porosity, compressibility, material strength, shock sensitivity, reaction rates, heats of combustion and detonation) in the DDT process. Very little work relating to the detailed chemistry involved in the DDT process has been attempted.

4.2 2D Steady Detonation in Homogeneous and Heterogeneous Materials

An experimental situation with some conceptual simplicity is steady two-dimensional detonation. Such detonations are observed in explosive charges that have an axis of symmetry. The most common experimental configuration is a right circular cylinder of explosive detonated in air (see Fig. 9). A detonation wave propagating in such a charge reaches a steady state after an initiation transient. In high-density explosive charges, experiment shows that a steady two-dimensional shock followed by a steady chemical reaction zone is closely approached after the detonation wave has propagated about six charge diameters. Thereafter, the steady 2D flow in the chemical reaction zone is determined by the interaction of the chemical-energy–release rate and the fluid mechanics of the divergent flow (see Fig. 10). At the steady state, the mode of initiation has been forgotten.

The hydrodynamic equations governing steady axisymmetric 2D flow in a detonation reaction zone are functions of two space

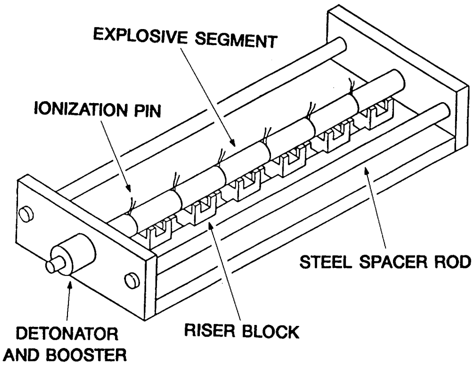
EXPLOSIVE SEGMENT
IONIZATION PIN
STEEL SPACER ROD
DETONATOR AND BOOSTER
RISER BLOCK

FIG. 9. Schema of the "rate-stick" assembly with which accurate measurements of steady 2D detonation velocity are made.

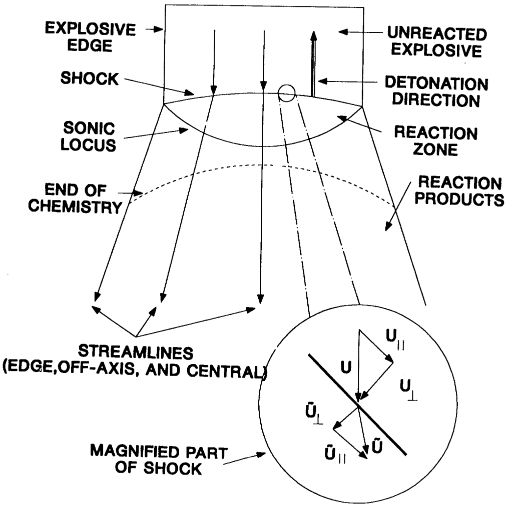
EXPLOSIVE EDGE
SHOCK
SONIC LOCUS
END OF CHEMISTRY
UNREACTED EXPLOSIVE
DETONATION DIRECTION
REACTION ZONE
REACTION PRODUCTS
STREAMLINES (EDGE,OFF-AXIS, AND CENTRAL)
MAGNIFIED PART OF SHOCK
$\mathbf{u}_{\parallel}$
$\mathbf{u}$
$\mathbf{u}_{\perp}$
$\tilde{\mathbf{u}}_{\perp}$
$\tilde{\mathbf{u}}_{\parallel}$
$\tilde{\mathbf{u}}$

FIG. 10. Quantities of interest in steady 2D detonation. The enlarged section of the shock (inset) shows how shock curvature deflects the streamlines in the flow (represented by $\mathbf{u}$ and $\tilde{\mathbf{u}}$).

variables and independent of time (in the Galilean frame traveling with the detonation wave).

It is known experimentally that a steadily propagating 2D detonation front, as discussed above, is curved. The shock on the charge symmetry axis leads the remainder of the shock wave; the qualitative form of such a detonation is shown in Fig. 10. A number of features of steady 2D detonation can be obtained from the fact that the detonation shock is curved in this way.

To deduce these features, we need the shock-jump conditions for fluid motion that is oblique (i.e., nonperpendicular) to the shock. Suppose that one rides in the shock reference frame and that the fluid velocity observed from this frame is **u**. As shown in the inset in Fig. 10, the velocity vector **u** can be resolved into components that are perpendicular ($u_\perp$) and parallel ($u_\parallel$) to the local shock surface. The conditions that relate the values of the components of **u** in front of and in back of the shock wave are as follows: The component $u_\parallel$ is continuous across the shock, while $u_\perp$ is governed by Eqs. (1)–(3) with U_s and $U_s - u_p$ replaced by the values of $u_\perp$ and $\tilde{u}_\perp$ in front of and behind the shock wave (see Fig. 10).

One sees from these considerations that more oblique shocks result in lower mass density, pressure, and internal energy jumps across the shock wave than occurs in normal incidence. Further, as can be seen from the inset in Fig. 10, the curvature of the steady 2D detonation shock causes divergence of the flow stream lines away from the central streamline as material crosses the shock. This divergence has a number of effects. First, it introduces mechanically based terms into the 2D Euler equations that have the appearance of an endothermic chemical reaction (see Fickett and Davis, 1979). This "mechanical" endothermicity is due to some of the chemical heat being diverted into producing lateral fluid motion. Because of this, the detonation speed is reduced from that observed in steady 1D detonation. A second more subtle effect is that this "mechanical" endothermicity produces a zero in the total heat-release rate before the full chemical-heat release is completed. As shown in Sec. 3.4, a zero in the heat-release rate is accompanied by a sonic point in the flow [see Eq. (22) and the following discussion]. The chemical heat released behind this sonic point cannot affect the shock and so it is lost in this sense. This effect also contributes to a reduction in the speed of the steady 2D detonation shock wave relative to the planar case.

A question of interest is, "How does the steady 2D detonation speed depend on the lateral charge size?" This is answered experimentally by measuring the detonation speed as a function of the lateral charge dimension. Such measurements produce an explosive's "diameter-effect" curve as shown in Fig. 11.

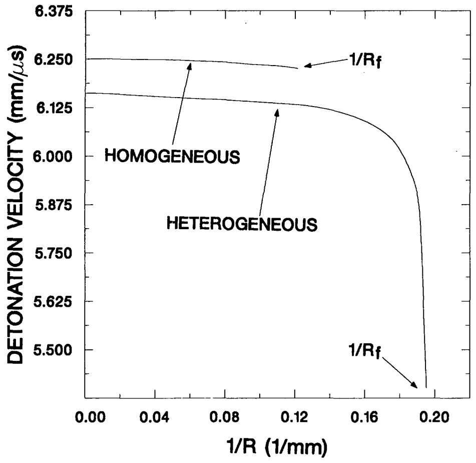

FIG. 11. Typical "diameter-effect" curves for homogeneous and heterogeneous explosives. Steady 2D detonation velocity is plotted versus the reciprocal of charge radius *R*.

Diameter-effect curves are usually plotted in the detonation velocity vs 1/(lateral dimension) plane; this choice allows one to estimate the steady 1D detonation speed by an extrapolation to zero. For cylindrical charges, the pertinent lateral dimension is the charge radius, *R*.

A striking feature of diameter-effect curves (see Fig. 11) is that they end at a finite lateral size. In cylindrical charges, this lateral dimension is called the "failure radius" and denoted by R_f. It is an experimental fact that it is impossible to propagate a steady 2D detonation in a cylindrical charge whose radius is less than R_f. R_f is dependent on the molecular structure of the explosive, its density, the material surrounding the explosive (confinement), the initial temperature, and other variables. Cylindrical charges fired with lateral dimension smaller than R_f result in shock waves that attenuate rapidly and completely. It is thought that the failure phenomenon is a result of detonation waves in charges near failure being catastrophically hydrodynamically unstable to local perturbations.

The 2D steady detonation of homogeneous and heterogeneous materials (see Sec. 4) exhibit strong differences from each other. This parallels the behavior, already discussed, seen in the initiation behavior of such materials and can be traced to the same origin—that is, the absence or presence of hot-spot phenomena (see Sec. 4.1.1 and 4.1.2).

We discuss the homogeneous case first, because it is simpler. The diameter-effect curves of homogeneous explosives are well represented by a line in the D vs $1/R$ plane (see Fig. 11). Second, the detonation-velocity decrease from the steady one-dimensional value $D(\infty)$ to the speed at failure, $D(R_f)$, is small; a value of the velocity deficit $\equiv [D(\infty) - D(R_f)]/D(\infty) \approx +0.01$ is typical.

Heterogeneous explosives have diameter-effect curves that are qualitatively different from those of homogeneous materials; they are strongly downward concave and have a velocity deficit $\approx +0.10$ (see Fig. 11).

It has been experimentally shown in a nitromethane system that the downward concave portion of a diameter-effect curve is caused by flow-field fluctuations (hot spots). Experimental diameter-effect curves for nitromethane exist for the pure liquid and for the liquid heterogenized by the addition of inert silica particles (see Fig. 11). The addition of the heterogeneities produces the expected downward concave region and the roughly ten times larger value of the velocity deficit at failure of detonation relative to that of the homogeneous material. A detailed analysis (Engelke and Bdzil, 1983) shows that the chemical-heat–release rate for the homogeneous material is much more sensitive to the thermodynamic state than is the rate for the heterogenized material.

These results for nitromethane offer a more complete view of the heat-release rates needed to produce the qualitatively different detonation phenomena seen in homogeneous and heterogeneous materials than is allowed by the hot-spot picture alone. A more encompassing view is that homogeneous (heterogeneous) behavior is observed in explosives that have strong (weak) dependence of the heat-release rate on the local average thermodynamic state. Within this picture, e.g., a physically homogeneous material could exhibit heterogeneous detonation behavior provided the thermodynamic-state dependence of the underlying microscopic chemical rates is weak enough. Such behavior is rare enough that the partitioning of explosives as homogeneous and heterogeneous as discussed in Sec. 4 is a useful viewpoint.

5. PROPERTIES OF CHEMICAL EXPLOSIVES

A large number of different materials fall under the category of condensed-phase explosives. These range from rapid-burning compositions to molecular explosives that detonate. Condensed explosives can be categorized into propellants (gun and rocket propellants consisting of mixtures of fuels and oxidizers), pyrotechnics (mixtures of metal fuel and oxidizers), molecular explosives (nitroglycerin, TNT, etc.), and slurry or blasting explosives (mixtures of fuels and oxidizers). Materials in the first two categories are designed to burn rapidly (deflagrate) and serious consequences may develop if they detonate. Materials in the last two categories are designed to detonate, and failure of detonation can lead to difficulties. All except the molecular explosives (which contain in each molecule the fuel and oxidizer necessary to react) are mixtures of fuels and oxidizers. The particular properties desired can be controlled by proper choice of the ingredients. Since there are many choices for fuels and oxidizers, a very large number of combinations are possible. Because of this, developing a particular propellant, pyrotechnic, or blasting material with certain desired properties may involve considerable art as well as science. In order to eliminate the difficulties of describing a large number of mixtures, we have chosen to include only molecular explosives in this discussion. The other three categories of materials will be touched on briefly below to provide only basic information.

Propellants fall into two different classes: gun propellants and rocket propellants. "Single-based" propellants consist of nitrocellulose (NC) only. "Double-based" propellants consists of NC and nitroglycerin (NG) in various compositions (NG is soluble in NC). Gun propellants can be single based, double based, or composite; rocket propellants are either double based or composites. Composite propellants are typically mixtures of NC and/or NG plus other fuel, oxidizer, and binder materials chosen to give the propellant the desired properties. Propellant burn rates are on the order of cm/s with the rate being proportional to the ambient pressure to a power less than unity (typically 0.4 to 0.9). The burning process consists of several steps: heatup, gasification or vaporization, mixing, and the final combustion chemistry. Shapes of propellant grains are designed to control the surface burning area in a rocket so that the ambient pressure and output thrust are controlled. Burning rates can be enhanced by addition of molecular explosives, such as HMX, with some formulations containing up to 50% HMX. Obviously, very careful designs are required to control the burning in these materials since a detonation would be disastrous.

Pyrotechnics are mixtures of metal-based oxidizers (for example, Fe_3O_4) and fuels that are usually metals (for example, Al, Mg, etc.). They are used in ordinary matches, safety flares, gunpowder, fireworks, auto air bags, etc. Pyrotechnic materials can be designed to produce heat, light, noise, or gas depending on the particular formulation. To produce this variety of outputs there are obviously a great many different combinations of fuels and oxidizers. Chemical reactions are generally solid state, and so the reactivity depends strongly on the physical characteristics of the components, how they are mixed, and special preparations of the mixtures. Because these materials are sensitive to very small inputs (so that they will react easily), great care must be exercised in their manufacture to keep from having unwanted ignitions. Manufacturing plants must be carefully designed to eliminate the possibility of introducing small inputs of static discharge, friction, heat, or impact. Even excess amounts of moisture can sometimes cause difficulties. While not detonable, when large quantities of pyrotechnics burn rapidly, considerable damage can result.

Blasting agents or slurry explosives are mixtures of ammonium nitrate and other oxidizers, various fuels including metals (usually aluminum), gelling agents, gassing agents, etc. Since these materials are usually formulated for a particular application, finding a composition with widespread use is difficult. These materials have the advantage that the various components are not explosives, so that an explosive material exists only after mixing. A delivery unit can be loaded with nonexplosive components and driven to the blasting area, and the materials mixed to form a detonable composition just prior to loading it into the blasting configuration, considerably enhancing safety. Blasting agents are thought to have the same fundamental detonation properties as molecular explosives.

The chemical properties of many molecular explosives have been measured; however, only a few materials have been studied to the extent that a complete set of initiation and detonation information is available. Some parameter measurements are intrinsically difficult to make (e.g., reaction-zone length).

On the basis of the ZND detonation theory (Sec. 2), if the unreacted explosive and the reaction-product Hugoniots are known, the von Neumann spike and CJ point can be determined. In practice the unreacted Hugoniot is not well known at higher pressures. Accurate measurements are not possible in this regime, since high-pressure input shocks cause prompt chemical reaction. This makes it necessary to extrapolate to the higher-pressure unreacted Hugoniot values. Similar measurement problems are encountered in determining the reaction-product Hugoniot; estimates are usually developed by using a 2D reactive-wave computer code and varying adjustable parameters to calibrate a particular EOS form to experimental data (for example, a JWL equation of state, see Kury *et al.*, 1965). Because of this and the difficulties in making direct measurements, reaction-zone parameters (e.g., spike and CJ pressures and reaction zone length) are not well defined for any explosive. In contrast, detonation velocity and shock sensitivity are relatively easily measured. Nearly all explosive properties depend on the initial mass density, and so this quality must be accurately known to specify an experiment properly. In addition, some properties, such as shock sensitivity, are dependent on the particle-size distribution of the material used to fabricate the explosive (see Sec. 4.1.2).

In this section we discuss the explosive classifications of primary, secondary, and insensitive materials; these are rather loosely defined groupings according to sensitivity. The chemical and detonation properties for a few representative molecular explosives are given to provide an idea of the magnitude of the various parameters (e.g., detonation velocity). Sensitivity of explosives to shock, impact, thermal, and spark inputs are discussed in order to provide the information necessary to understand the explosive classifications discussed in Sec. 5.1. Finally a brief discussion of some of the experimental methods used in obtaining detonation and initiation information is given and illustrated with figures.

5.1 Explosive Classifications

Explosives are usually classified according to the broad sensitivity classifications: primary, secondary, and insensitive. Primary explosives can be initiated with rather small inputs of mechanical shock, heat, flame, or spark and are used to start an explosive train. Examples of primary explosives are lead azide and mercury fulminate. Small amounts of these materials are located in initiators so that small electrical or mechanical inputs will reliably initiate them. They, in turn, reliably initiate the output charges, e.g., booster pellets and gunpowder. Extreme care is required when handling these materials to prevent accidental initiation.

Secondary explosives are considerably less sensitive to the same inputs and can be handled with less concern about accidental initiation. These materials are reliably initiated by moderate- to high-pressure shock inputs. They are used in some high-input initiating devices (e.g., exploding bridge-wire detonators or exploding-foil–driven flyer-plate detonators), output pellets and boosters, and main charges in explosive assemblies. Examples are 2,4,6-trinitrotoluene (TNT), cyclotetramethylene tetranitramine (HMX), and pentaerythritol tetranitrate (PETN). Hundreds of molecular explosives fit in this category.

Insensitive explosives are difficult to initiate and shock inputs for reliable initiation must be quite large, i.e., careful planning is usually required to eliminate failures, particularly when there can be environmental changes, e.g., low ambient temperature. Examples of materials in this category are 1,3,5-triamino-2,4,6-trinitrobenzene (TATB) and nitroguanidine (NQ).

Some materials are borderline and could be placed in either of two categories, depending on the particular sensitivity property being used as the criterion. For instance, nitroglycerin might be put into the secondary explosive category because it requires a rather high-level shock input to initiate it in the absence of heterogeneities (e.g., bubbles); in this case, it initiates homogeneously as discussed in Sec. 4.1.1. With heterogeneities present, heterogeneous initiation can take place (see Sec. 4.1.2) at much lower input pressures. Under these circumstances it is classified as a primary explosive.

Table 1. Chemical properties of selected molecular explosives.[a] [Data from Dobratz and Crawford (1985) and Meyer (1987).]

Common name	Chemical name[a]	Chemical formula	Molecular weight	Theoretical density (g/cm^3)	Melting point (°C)	ΔH_f Heat of formation (kcal/mol)	$-\Delta H_{det}$ Heat of detonation[b] (kcal/g)
Lead azide	Lead azide	$Pb(N_3)_2$	291.3	4.8	dec.[c]	+112	0.37
NG	Nitroglycerin or glycerol trinitrate	$C_3H_5N_3O_9$	227.1	1.60	13.2	−88.6	1.48
PETN	Pentaerythritol tetranitrate	$C_5H_8N_4O_{12}$	316.2	1.78	140	−128.7	1.51
Tetryl	Trinitro-2,4,6-phenylmethylnitramine	$C_7H_5N_5O_8$	287.0	1.73	130	+4.67	1.45
RDX	Cyclo-1,3,5-trimethy-lene-2,4,6-trinitramine	$C_3H_6N_6O_6$	222.1	1.81	205[d]	+14.7	1.48
HMX	Cyclotetramethylene tetranitramine	$C_4H_8N_8O_8$	296.2	1.90	285	+17.9	1.48
HNS	Hexanitrostilbene	$C_{14}H_6N_6O_{12}$	450.3	1.74	316[d]	+18.7	1.36
TNT	2,4,6-trinitrotoluene	$C_7H_5N_3O_6$	227.1	1.65	80.9	−16.0	1.29
NM	Nitromethane	CH_3NO_2	61.0	1.13	−29	−27	1.36
AN	Ammonium nitrate	NH_4NO_3	80.7	1.73	169	−87.3	0.38
TATB	1,3,5-triamino-2,4,6-trinitrobenzene	$C_6H_6N_6O_6$	258.2	1.94	dec.[c]	−36.9	1.08
NQ	Nitroguanidine	$CH_4N_4N_4O_2$	104.1	1.78	257	−22.1	0.88

---

[a] Common chemical names are used rather than the more descriptive names used in the Chemical Abstract Index Guide (see Dobratz and Crawford, 1985).

[b] Heats of detonation are calculated (rather than experimental) and assume that H_2O is a gas.

[c] Decomposes before melting.

[d] Melts with some decomposition.

5.2 Chemical and Detonation Properties for Selected Explosives

Chemical properties of a selected group of molecular explosives (some from each sensitivity class), listed in descending order of sensitivity, are presented in Table 1. Most of these are organic molecules containing nitro groups (NO_2). Detonation-reaction product gases are typically N_2, CO, CO_2, and H_2O, along with small quantities of other chemical species. Very little is known about the chemical reaction mechanisms by which the reactant is transformed into products.

Heats of formation given in Table 1 are the difference in enthalpy between the explosive molecule and its elements (in their reference state). The heat of detonation is the enthalpy difference per gram between the explosive molecule and the product molecules that result from the reaction. Comparison of the heat of detonation values gives a partial measure of the expected output; higher heats of detonation usually correspond to more powerful explosives, as is the case for HMX, RDX, and PETN.

Detonation properties for the explosives of Table 1 are given in Table 2. Densities are given for each entry (except those that are liquids) because the values are density dependent. Detonation velocity is the most accurately measured detonation parameter; consequently, there is a considerable amount of such data available for these and other explosives. In some cases, these data include the dependence of detonation velocity on initial density and temperature. High detonation velocities are associated with high performance; this is apparent in Table 2. The CJ pressure is more difficult to measure and, hence, these values are less reliable. The von Neumann spike and reaction-zone length data are only given for a few materials because of the measurement difficulties. Improved measurements are now being obtained by laser-velocity interferometry and *in situ* gauging

Table 2. Detonation properties of selected molecular explosives. [Densities are given in parentheses after each property value. Data from Dobratz and Crawford (1985) and Gibbs and Popolato (1980). (Most reaction-zone estimates were made by the authors, based on the literature.)]

Material	Theoretical density (g/cm^3)	Detonation velocity (mm/μs)	Estimated CJ pressure (GPa)	Estimated spike pressure (Gpa)	Estimated reaction zone length (mm)	Failure radius (mm)	Plate dent depth (mm)[a]
Lead azide	4.80	5.5(3.80)	23 (4.0)	—	—	—	—
NG	1.60[b]	7.7	25	—	0.2	1.1[c]	—
PETN	1.78	8.26 (1.76)	30 (1.67)	—	—	—	9.8 (1.66)
Tetryl	1.73	7.85 (1.71)	26 (1.71)	—	—	—	8.1 (1.68)
RDX	1.81	8.7 (1.77)	34 (1.77)	—	—	—	10.4 (1.75)
HMX	1.90	9.11 (1.89)	39 (1.89)	—	—	—	10.1 (1.73)
HNS	1.74	7.00 (1.70)	23 (1.74)	—	—	—	—
TNT	1.65	6.93 (1.64)	21 (1.63)	28 (1.64)	0.6 (1.64)	7.5 (1.62)[d]	7.0 (1.63)
NM	1.13[b]	6.35	13	20	$\approx$0.1[e]	7.6[c]	4.1
AN	1.73	5.27 (1.30)	22 (1.73)	—	—	>50 (1.4)[c]	—
TATB	1.94	7.76 (1.88)	29 (1.88)[f]	38 (1.88)[f]	2. (1.88)[f]	4–5 (1.88)[d,f]	8.3 (1.87)
NQ	1.78	7.93 (1.62)	25 (1.63)	—	—	—	—

---

[a] Charge diameter 41.3 mm; cold-rolled steel dent plates.

[b] Liquid.

[c] Confined in glass.

[d] Unconfined or paper tube confinement.

[e] Experiments are underway to define this more accurately.

[f] PBX 9502, which is 95/5 wt. % TATB/Kel-F.

techniques (see Sec. 5.4), so that better estimates are forthcoming. TATB has one of the longer reaction zones at about 2 mm and NM has one of the shorter ones at about 0.1 mm.

Explosives are often mixed with a few percent of polymeric materials (to make them easier to process—press, mold, machine, etc.). Such mixtures are called plastic bonded explosives (PBX).

Many explosive properties depend on initial temperature; e.g., NG has a failure radius of 2 mm at an initial temperature of −20 °C and 0.6 mm at 70 °C. Properties given in Table 2 are at ambient initial temperature.

5.3 Sensitivity Properties

Explosive materials vary widely in their sensitivity to different inputs. Because experience has shown it necessary to understand explosive hazards, a rather large number of sensitivity tests have been developed. Some of these are peculiar to a particular operation or hazard. Methods have been developed to measure response to various inputs, including shock, friction and crushing, heat, and spark or static electricity.

"Gap" tests, of which there are several different types, are widely used in measuring shock sensitivity. In this test, a standard explosive driver system produces an output shock that is attenuated to different pressure levels by varying a thickness of inert attenuator material. The attenuator thickness at which the explosive sample is just initiated is the "gap" thickness reported. Data from two different gap tests are given in Table 3 for some of the materials of Table 1. A large gap indicates a more sensitive material since, with a larger amount of the inert material present, the pressure is attenuated to a lower value before it enters the explosive. These data indicate that the materials are listed in Table 3 in descending order of sensitivity (with minor exceptions).

Another shock sensitivity test is obtained by impacting a flyer plate on the explosive; in this test, a critical energy input for initiation is the recorded index. Flyer material, thickness, and velocity can all be varied to change the characteristics of the input pressure pulse. The required critical energy has been found to be relatively constant over a range of input pressures and pulse widths. Values obtained

Table 3. Sensitivity properties on selected molecular explosives. Densities are given in parentheses after each sensitivity value where appropriate. [Data from Dobratz and Crawford (1985), Gibbs and Popolato (1980), and Smith (1987).]

Material	Theoretical density (g/cm^3)	Gap test[a] (mm)	Gap test[b] (mm)	Drop weight[c] (m)	Spark input[d] (J)	DTA exotherm (°C)	Critical energy (cal/cm^2)
			Primary Explosive				
Lead azide	4.80	27.2 (3.66)	—	—	0.007	—	0.03 (1.26)
			Secondary Explosives				
PETN	1.78	14.4 (1.58)	5.21 (1.76)	0.11	0.19	190	4. (1.6)
RDX	1.81	7.9 (1.72)	5.18 (1.74)	0.28	0.21	200	16? (1.55)
HMX	1.90	8.7 (1.81)	4.27 (1.79)	0.33	0.23	260	—
TNT	1.65	6.25 (1.56)	0.33 (1.63)	0.80	0.46	250	34 (1.64)
NM	1.13	—	0.18–0.43 (1.13)	>3.2[e]	—	—	405 (1.13)
			Insensitive Explosives				
TATB	1.94	1.12 (1.88)	0.13 (1.87)	>1.77	4.25	340	72–88 (1.76)
NQ	1.78	2.72 (1.27)	no-go (1.58)	>1.77	—	240	—

[a] NSWC (Naval Surface Warfare Center) Gap Test.
[b] Los Alamos (LANL) Small-Scale Gap Test.
[c] 5-kg weight using Type 12 tooling (sandpaper surfaces).
[d] Highest electrostatic-discharge energy at 5000 volts without ignition; sample confined.
[e] 2.5-kg weight using Type 12 tooling (sandpaper surfaces), modified tooling for liquids.

from this test are given in the last column of Table 3. These data indicate that the critical energy to initiate an explosive follows the same order as the gap test sensitivity. A notable exception is liquid nitromethane, which initiates homogeneously in contrast to the other materials, which undergo heterogeneous initiation (see Sec. 4.1).

"Impact" sensitivity is important because this type of test mimics inputs in many accident situations. Drop weight setups (there are several with different drop heights, weights, anvils, and anvil surfaces) are used to determine impact sensitivity. Samples are usually in the form of powders or pellets. The samples are placed in an anvil setup and then the drop weight is allowed to undergo free fall onto the sample. The height of free fall that causes reaction is the value reported. Evidence of reaction is determined by both active measurement and postmortem examination of the debris from the test. Data from one type of impact test setup (Type 12 tooling) are given in Table 3. Sensitivity as measured by this test generally agrees with that of the other tests.

A differential thermal analysis (DTA) test is often used to determine sensitivity to thermal input. This test involves a continuous comparison of an explosive sample's temperature with that of an inert standard, when both are heated at a constant rate. Energy excursions are determined in this test, i.e., the temperatures at which endotherms or exotherms occur. The temperature at which large exotherms are observed are given in Table 3. Another thermal input test involves the immersion of an explosive sample in a constant high-temperature metal bath and measuring the time to reaction or explosion. By use of the thermal explosion analysis developed by Frank-Kamenetskii (1955), Arrhenius rate parameters can be determined from this experiment.

Sensitivity to spark input (such as might be produced by the human body) is of interest because of the possibility of discharging through an explosive material when working with it. Critical spark-energy values from this test are test specific (i.e., there are several different test setups). Data given in Table 3 are from one such setup. Note that this test gives the same sensitivity order as the other tests.

Because threshold reaction conditions are being measured, probabilistic data comes from the above tests. Therefore, it is necessary to run several tests at each input condition to establish statistically the lowest input reaction condition.

The data of Table 3 show large differences in sensitivity among the various explosive materials. TATB and NQ are obviously very insensitive explosives. The differences between lead azide (a primary explosive) and PETN (which is considered one of the more sensitive secondary explosives) are considerable. Differences between the secondary and insensitive explosives are less apparent. The order of the materials in Table 3 (from top to bottom) is from more to less sensitive, but there are order reversals for some tests. This indicates that one cannot reliably determine the sensitivity of an explosive by conducting only one type of test.

5.4 Initiation and Detonation Measurement Techniques

Because of the harsh environment produced by the detonation and initiation processes, measurements of certain phenomena are extremely difficult, e.g., measurement of the detailed chemistry occurring in an initiation or detonation. However, considerable progress has been made in measuring the mechanical aspects of initiation and detonation phenomena. A wide variety of experimental techniques are employed to make such measurements including various gauging techniques, laser velocity interferometry, streak cameras, framing cameras, flash x-ray, "time of arrival" ionization and piezoelectric pins, and integrated output tests such as plate dent testing. We will only discuss a few of these. Articles in the various Detonation Symposia contain more detailed discussion of experimental methods.

In situ gauging (i.e., low-perturbation gauges located in the flow field) is becoming important. These gauges give direct information on wave profiles in the flow and contain implicit information on how the chemical heat release is occurring. The particle-velocity waveforms shown in Figs. 7(b) and 8 are examples of *in-situ* gauging measurements in which the initiation process has been measured in nitromethane-based materials. These measurements provided the insight necessary to modify the long-held homogeneous-initiation model as discussed in Secs. 4.1.1 and 4.1.2. *In situ* Manganin pressure measurements are also providing important insights into these processes.

Laser-velocity-interferometry measurements also provide wave-profile information as an initiation or detonation wave accelerates an explosive/window-material interface from which a laser beam is reflected. An important advantage (over *in situ* gauging) is that extremely high–time-resolution measurements are possible even at very high pressure. A disadvantage is that the wave profiles are perturbed by the mechanical impedance mismatch at the window through which the observations are made. Figure 12 shows information regarding ORVIS (Optically Recorded Velocity Interferometer System) velocity-interferometer measurements of a detonation wave in a TATB-based explosive interacting with a 12-μm-thick copper foil backed by a water window. Fig. 12(a) shows the experimental setup, Fig. 12(b) the electronic streak-camera record, and Fig. 12(c) the particle-velocity history obtained from reducing the data. Wave reverberations in the foil (with subnanosecond time resolution) are readily apparent in the reduced data. This method provided the reaction-zone-length estimate given in Table 2 for TATB (PBX 9502).

Figures 6 and 13 contain information from a Fabry-Pérot velocity-interferometer measurement of a detonation wave in nitromethane (NM) interacting with a polymethylmethacrylate (PMMA) window. In this case, the reflecting film was a submicron thick layer of aluminum. Figure 13(a) shows the experimental setup, Fig. 13(b) the experimental streak-camera record; and the interface particle-velocity history obtained from reducing the record in Fig. 13(b) is presented in Fig. 6. PMMA is a good mechanical impedance match with NM, and so perturbations of the detonation-wave profile were small. These data provided the reaction-zone-length estimate given in Table 2 for NM.

Detonation velocity is precisely measured in rate-stick experiments as shown in Fig. 9. Ionization switches (in some cases consisting of 50-μm-diam copper wire) are installed between precisely measured pieces of the explosive and provide wave arrival time information from which the detonation velocity is obtained. Diameter-effect curves (see Sec. 4.2 and Fig. 11) are generated when the charge diameter is varied in a set of such rate-stick experiments.

Wedge experiments yield run distance- or time-to-detonation data from sustained shock input experiments. A precisely machined explosive wedge is mounted on a plane-wave explosive driver/attenuator system as shown in Fig. 14(a). The wedge is illuminated by a light source (usually explosively driven) and

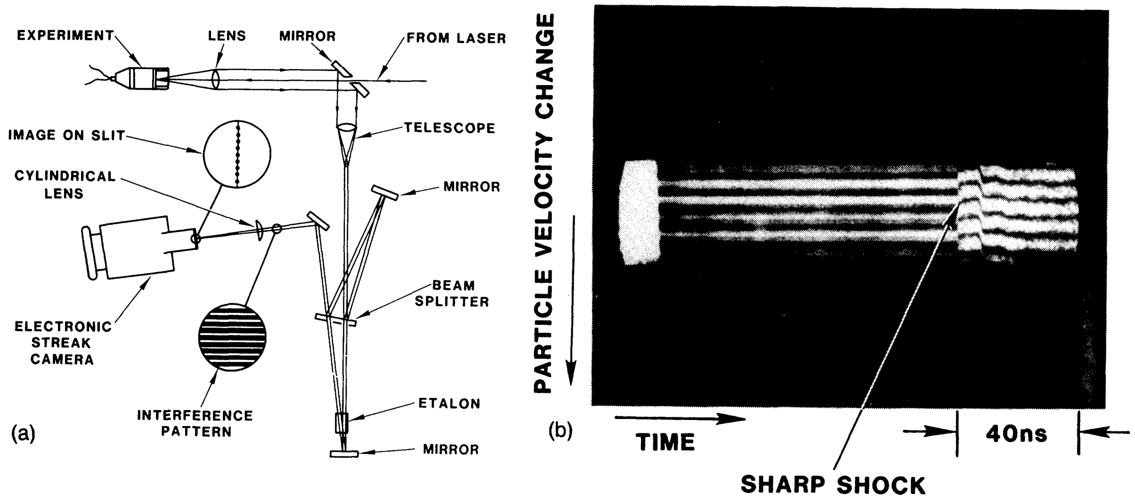

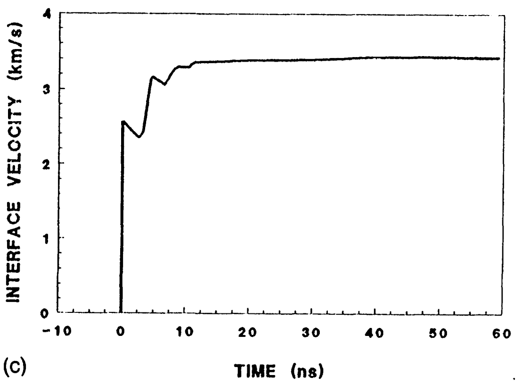

FIG. 12. (a) The experimental setup for an ORVIS. (b) An electronic-streak-camera recording of the fringe changes with time. (c) The interface velocity history obtained by reduction of the record in (b). This experiment consisted of a detonation wave in TATB (PBX 9502) interacting with a 12-μm-thick copper foil backed by a water window.

a streak camera is used to monitor the light changes as a function of time along the wedge surface as the shock interacts with the surface; such a trace is shown in Fig. 14(b). The point at which the distance-time streak record has a large change in slope is the distance or time to detonation for that particular pressure input. This is clearly shown in Fig. 14(c) in which the slope (shock velocity) is plotted versus distance of shock travel into the wedge. By doing several experiments, in which the input shock pressure is varied, a set of pressure versus distance- or time-to-detonation data are compiled. A plot of these data in the log(pressure) vs log(run distance) plane is typically a straight line (Ramsay and Popolato, 1965) as shown in Fig. 15. Note that for the explosives on this plot, PETN is the most-sensitive and TATB and NQ are the least-sensitive materials.

Rotating-mirror framing cameras and electronic image-converter cameras are used to make stop-action pictures of detonation events. Exposure times range from microseconds to nanoseconds, depending on the camera used. For example, Fig. 2 was taken using an electronic image-converter camera that was triply shuttered so that there are three stop-action pictures at different times superimposed in the photograph. This picture shows the progress of the detonation wave up the charge and also the form of the shock waves driven into the liquid surrounding the charge. Estimates of the reaction-product Hugoniot of the explosive can be obtained from this type of data by using it to calibrate computer models of the event. A related test involves the measurement of the wall expansion of a precisely machined and explosively loaded metal cylinder when the explosive is detonated. This experiment is conducted with a streak camera and is called a "cylinder test." Information obtained from this test is also used for calibrating reaction-product Hugoniots (see Sec. 5).

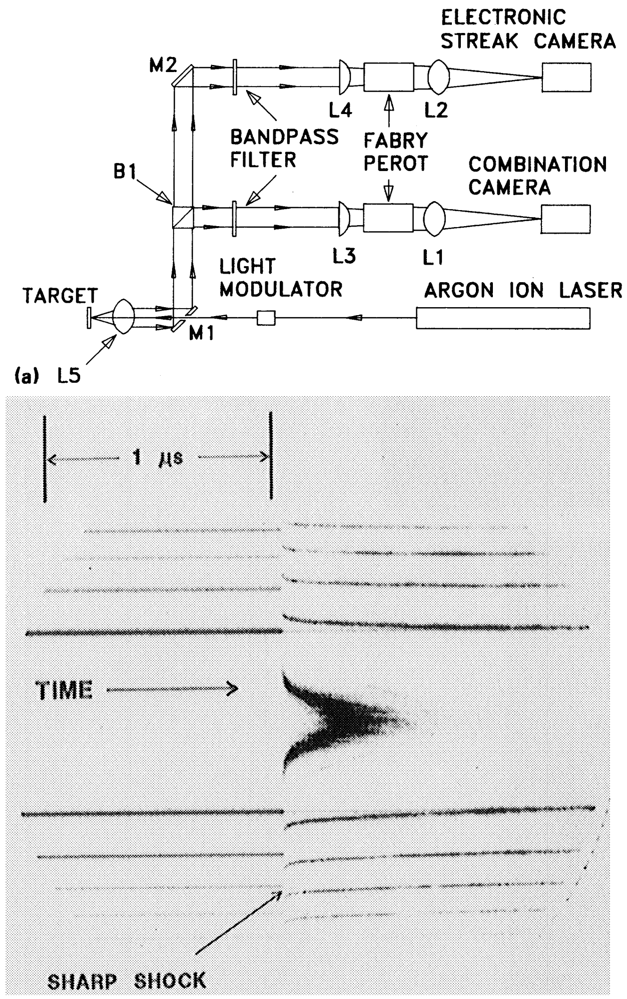

FIG. 13. (a) The experimental setup for a Fabry–Perot velocity interferometer system. (b) An electronic-streak-camera recording of the fringe changes with time. The interface velocity history obtained by reduction of the record in (b) is plotted in Fig. 6. This wave form was obtained from an experiment in which a detonation wave in liquid nitromethane interacted with a polymethylmethacrylate (PMMA) window.

Failure radii for propagation of steady detonation waves are determined by making a series of shots at progressively smaller radii until detonation failure occurs. Failure is determined by the condition of dent blocks or post-shot examination of the debris. Failure radius is reduced by heavier confinement, e.g., NM has a failure diameter of $\approx$16 mm in Pyrex and $\approx$2.5 mm in brass.

Other measurement methods are discussed in several of the references cited in the "Further Reading" list.

6. SUMMARY

Developments in condensed-phase explosive technology in the last 150 years have led to substantial increases in understanding of the initiation and detonation process. A planar steady detonation is a chemical-reaction–supported shock. The shock takes the unreacted material to a high pressure (the von Neumann spike); this is followed by decreasing pressure in the chemical-reaction zone in which the explosive is transformed into gaseous reaction products. The chemical-reaction zone terminates at a sonic point (the CJ state). *In situ* and interface velocity measurements of the reaction zone support the 1D steady detonation theory. Spectroscopic measurements that attempt to interrogate the reaction-zone chemistry are a current area of research.

Explosives are classified according to sensitivity as primary (the most sensitive), secondary (which have intermediate sensitivity), and insensitive explosives (the least sensi-

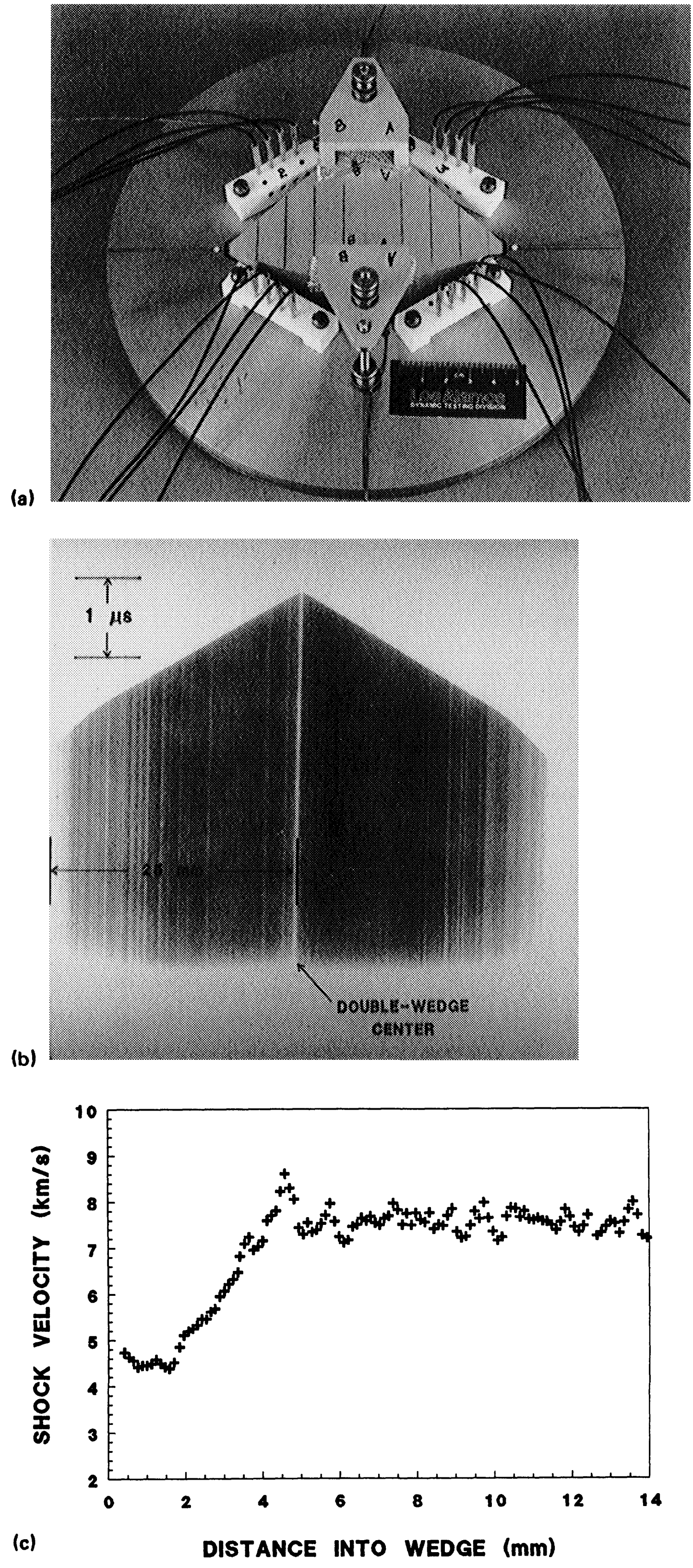

FIG. 14. (a) The experimental setup for an explosively driven wedge experiment. The 16 wires are leads for electrical pins to measure the input conditions into the explosive wedge. (b) A streak-camera record showing shock contact with the wedge surface. Shock contact with the wedge surface reduces the reflectivity and therefore the light intensity. The dark area in the streak record photograph is the region of high reflectivity. (c) The shock velocity plotted versus distance into the wedge obtained from the space-time trajectory of the shock contact with the wedge surface. The distance-to-detonation point is about 4 mm.

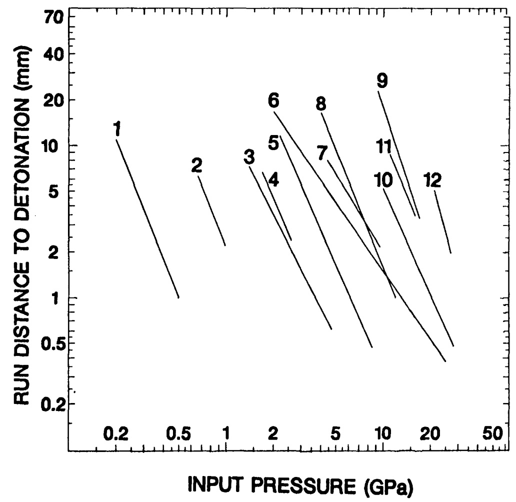

FIG. 15. Plot of sustained input shock pressure vs run distance to detonation (in the log-log plane), as obtained from wedge tests, for a number of explosives. Explosives are as follows (numbers correspond to numbers on graph): (1) PETN ($\rho_0 = 1.0$ g/cm^3); (2) PETN ($\rho_0 = 1.4$ g/cm^3); (3) PBX 9407 (94/6 wt. % RDX/Exon, $\rho_0 = 1.60$ g/cm^3); (4) PETN ($\rho_0 = 1.75$ g/cm^3); (5) Tetryl ($\rho_0 = 1.70$ g/cm^3); (6) PBX 9404 (94/3/3 wt. % HMX/nitrocellulose/chloroethylphosphate, $\rho_0 = 1.84$ g/cm^3); (7) HMX ($\rho_0 = 1.891$ g/cm^3); (8) pressed TNT ($\rho_0 = 1.63$ g/cm^3); (9) cast TNT ($\rho_0 = 1.635$ g/cm^3); (10) superfine particle TATB ($\rho_0 = 1.806$ g/cm^3); (11) TATB ($\rho_0 = 1.876$ g/cm^3); (12) NQ ($\rho_0 = 1.688$ g/cm^3). Data taken from Gibbs and Popolato (1980).

tive). Sensitivity is measured by a number of well-defined test methods in which sensitivity to shock, impact, heat, and electrical inputs are measured.

The initiation and 2D steady detonation processes depend qualitatively on whether the explosive is homogeneous or heterogeneous. Homogeneous explosives are typically liquids or single crystals; their initiation and detonation behavior depend on the bulk temperature and pressure generated by an input shock. Heterogeneous explosives have voids, grain boundaries, etc., which perturb the flow so that spatially localized hot spots are produced; these hot spots are energy concentrations where the chemical reaction rate is greatly increased. These regions control the initiation and detonation behavior of heterogeneous explosives.

Explosives are used in gun propellants, rocket propellants, and pyrotechnics; such materials are designed to burn (deflagrate) rather than detonate. Under some accident conditions, these explosives can burn too rapidly or even detonate, with great consequent damage to the surroundings.

Important areas of explosive research have been neglected in our exposition. Examples of this are

1. the study of the hydrodynamic stability of detonations to perturbations,
2. studies concerning numerical solution of the partial differential equations governing multidimensional time-dependent initiation and detonation, and
3. another type of steady 1D detonation solution possible when more complex chemical-heat–release functions are used ("weak detonations").

Discussion of these and other important topics can be found in the Further Reading list.

We have attempted to show that, in contrast to the intuitive feeling that explosives are disorderly, irreproducible, and chaotic processes, they can be quite orderly and are governed by the conservation equations of compressible fluid mechanics along with material constitutive relations.

Evidence has been discussed showing that there are explosives with widely varying characteristics. These include, e.g., explosives with different sensitivities to accidental insult, levels of power and pressure generation during detonation, and cost. Because of this variation, one can choose an explosive appropriate to the application. Further work is being done to increase this variety of choice. An important example of this is the study of insensitive explosives for use in nuclear weapons, the goal being to produce a situation where it is virtually impossible to produce detonation of the explosive, except by design.

GLOSSARY

CJ (Chapman-Jouguet) Model: The simplest model of planar steady detonation, in which the exothermic chemical reaction is assumed to be instantaneous.

DDT (Deflagration-to-Detonation Transition): The mechanical compaction and reaction buildup processes that an explosive goes through to make the transition from deflagration (burning) to detonation.

Diameter-Effect Curve: The relation between steady 2D detonation velocity and the lateral charge size.

EOS (Equation of State): The functional relationship between internal energy, pressure, and volume for a material in thermodynamic equilibrium.

Failure Radius: The radius of a cylindrical explosive charge below which it is impossible to propagate a steady detonation.

Heterogeneous Explosive: A material in which the release of chemical heat is controlled by local high-temperature and -pressure regions at flow irregularities ("hot spots").

Homogeneous Explosive: A material in which the release of chemical heat is controlled by the bulk temperature and pressure.

Hot Spots: Localized regions in a shocked explosive where energy is concentrated by fluid-flow irregularities in the vicinity of physical imperfections.

Hugoniot: A curve, obtained from the conservation laws, that restricts the locus of end states achievable behind a shock wave or in steady laminar flow.

LVD (Low-Velocity Detonation): A low-level reactive wave in an explosive that is dependent on the existence of nonreactive precursor waves. LVD is not a true detonation.

Rankine-Hugoniot Conditions: The three mathematical conditions that guarantee conservation of mass, momentum, and energy across flow discontinuities (e.g., shocks); they also apply between any two locations in laminar steady flow.

Rayleigh Line: A line in the P,v plane obtained from the conservation laws that restricts the states possible behind a shock wave or in steady laminar flow.

SDT (Shock-to-Detonation Transition): The reaction buildup process that an explosive goes through to make the transition from a shock input to detonation.

Superdetonation: A detonation that occurs in an explosive precompressed by an initial shock.

Taylor Wave: The pressure-relief wave (rarefaction) following a planar steady detonation.

Thermal Explosion: A rapid chemical reaction that occurs when the temperature in an explosive gets above a threshold value.

von Neumann Spike: The leading (high-pressure) point in the ZND model of planar steady detonations (sometimes referred to as the chemical peak).

ZND (Zel'dovich, von Neumann, Doering) Model: The simplest model of planar steady detonation with a resolved chemical-reaction zone.

List of Works Cited

Baer, M. R., Nunziato, J. W. (1989), in: *Proceedings of the Ninth Symposium (Intl.) on Detonation*, Office of Naval Research OCNR 113291-7, Washington: U. S. GPO, pp. 293–305.

Campbell, A. W., Davis, W. C., Travis, J. R. (1961a), *Phys. Fluids* **4**, 498–510.

Campbell, A. W., Davis, W. C., Ramsay, J. B., Travis, J. R., (1961b), *Phys. Fluids*, **4**, 511–521.

Dobratz, B. M., Crawford, P. C. (1985), *LLNL Explosives Handbook*, UCRL-52997, Livermore, CA: Lawrence Livermore National Laboratory.

Dremin, A. N., Savrov, S. D., Trofimov, V. S., Shvedov, K. K. (1970), *Detonation Waves in Condensed Media* (transl. from Russian), Springfield, VA: National Technical Information Service, AD-751417.

Engelke, R., Bdzil, J. B. (1983), *Phys. Fluids*, **26**, 1210–1221.

Fickett, W., Davis, W. C. (1979), *Detonation*, Berkeley: Univ. of California Press.

Frank-Kamenetskii, D. A. (1955), *Diffusion and Heat Exchange in Chemical Kinetics*, Princeton: Princeton Univ. Press.

Gibbs, T. R., Popolato, A. (1980), *LASL Explosive Property Data*, Berkeley: Univ. of California Press.

Johansson, C. H., Persson, P. A. (1970), *Detonics of High Explosives*, New York: Academic.

Kury, J. W., Hornig, H. C., Lee, E. L., McDonnel, J. L., Ornellas, D. L., Finger, M., Strange, F. M., Wilkins, M. L. (1965), in: *Proceedings of the Fourth Symposium (Intl.) on Detonation*, Office of Naval Research ACR-126, Washington: U. S. GPO, pp. 3–13.

Mader, C. L. (1979), *Numerical Modeling of Detonation*, Berkeley: Univ. of California Press.

McAfee, J. M., Asay, B. W., Campbell, A. W., Ramsay, J. B. (1989), in: *Proceedings of the Ninth Symposium (Intl.) on Detonation*, Office of Naval Research OCNR 113291-7, Washington: U. S. GPO, pp. 265–279.

Meyer, R. (1987), *Explosives*, Weinheim and New York: VCH.

Nunziato, J. W., Kipp, M. E., Setchell, R. E., Walsh, E. K. (1982), *Shock Initiation in Heterogeneous Explosives*, SAND81-2173, Albuquerque: Sandia National Laboratories.

Ramsay, J. B., Popolato, A. (1965), in: *Proceedings of the Fourth Symposium (Intl.) on Detonation*, Office of Naval Research ACR-126, Washington: U. S. GPO, pp. 233–238.

Ree, F. H., van Thiel, M. (1985), in: *Proceedings* *of the Eighth Symposium (Intl.) on Detonation,* NSWC MP, Washington: U. S. GPO, pp. 86–194.

Sheffield, S. A., Engelke, R., Alcon, R. R. (1989), in: *Proceedings of the Ninth Symposium (Intl.) on Detonation,* Office of Naval Research OCNR 113291-7, Washington: U. S. GPO, pp. 39–49.

Smith, L. C. (1987), *LANL Explosives Orientation Course: Sensitivity and Sensitivity Tests,* LA-11010-MS, Los Alamos: Los Alamos National Laboratory.

Taylor, G. I. (1950), *Proc. Royal Soc. A* **200**, 235–247.

von Neumann, J. (1963), *Collected Works,* Vol. 6, New York: Pergamon, pp. 203–218.

Wood, W. W., Kirkwood, J. G. (1957), *J. Appl. Phys.* **25**, 395–398.

Further Reading

Belyaev, A. F., Bobolev, V. K., Korotkov, A. I., Sulimov, A. A., and Chuiko, S. V. (1973), *Transition From Deflagration to Detonation in Condensed Phases* (transl. from Russian), Springfield, VA: National Technical Information Service.

Dobratz, B. M., Crawford, P. C. (1985), *LLNL Explosives Handbook,* UCRL-52997, Livermore: Lawrence Livermore National Laboratory.

Dremin, A. N., Savrov, S. D., Trofimov, V. S., Shvedov, K. K. (1970), *Detonation Waves in Condensed Media* (transl. from Russian), Springfield, VA: National Technical Information Service AD-751417.

Fickett, W., Davis, W. C. (1979), *Detonation,* Berkeley: Univ. of California Press.

Gibbs, T. R., Popolato, A. (1980), *LASL Explosive Property Data,* Berkeley: Univ. of California Press.

Johansson, C. H., Persson, P. A. (1970), *Detonics of High Explosives,* New York: Academic.

Lewis, B., von Elbe, G. (1987), *Combustion, Flames, and Explosions of Gases,* Orlando: Academic.

Mader, C. L. (1979), *Numerical Modeling of Detonation,* Berkeley: Univ. of California Press.

McLain, J. H. (1980), *Pyrotechnics,* Philadelphia: The Franklin Institute Press.

Picatinny Arsenal/Large Caliber Weapons Systems Laboratory (1960–1983), *Encyclopedia of Explosives and Related Items* (in ten volumes), New Jersey: Dover.

Proceedings of the Third to Ninth (Intl.) Symposia on Detonation, sponsored by the Office of Naval Research, Washington: U. S. GPO.

Strehlow, R. A. (1984), *Combustion Fundamentals,* New York: McGraw-Hill.

Zel'dovich, Ia. B., Raizer, Iu. P. (1966–1967), *Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena,* New York: Academic.

FIELD EMISSION AND FIELD IONIZATION

ANTHONY E. BELL, *Oregon Graduate Institute of Science and Technology, Beaverton, Oregon, U.S.A.*
DOUGLAS F. BAROFSKI, *Oregon State University, Corvallis, Oregon, U.S.A.*

Introduction 379
1. **Principles** 379
1.1 Field Emission 379
1.2 Field Ionization 382
1.2.1 Ion Formation 383
1.2.2 Ion Current Generation 384
1.3 Field Evaporation and Field Desorption 384
1.3.1 Image-Force Model of Desorption 384
1.4 Liquid-Metal Ion Sources (LMIS) 385
2. **Devices and Instruments** 386
2.1 Emitters 386
2.2 Field-Emission Microscope 386
2.3 Field-Ion Microscope 387
2.4 Scanning Tunneling Microscope 387
2.5 Flash X-Ray Technology 388
2.6 Electron-Beam Lithography 389
2.7 Flat-Panel Field-Emission Display (FED) Technology 389
2.8 Field-Emission Microwave Generation 390
Glossary 390
Works Cited 391
Further Reading 391

INTRODUCTION

Ions and electrons may be generated by the application of very high electric fields, of the order of 3–5 GV/m in the case of electrons and 10–50 GV/m in the case of ion generation. Fields of this order are conveniently generated by applying a voltage of several kV to sharp needles with an apex radius of ~ 100 nm. In the case of liquid-metal ion sources, the liquid necks down to form a spatially stable jet of radius ~ 2 nm, which is sufficiently small to generate the required high electric fields in spite of the shielding by the nearby liquid cone from which the jet issues. The basic mechanism in all of the above cases involves tunneling of electrons through a potential energy barrier which has been altered by the high electric field.

The above phenomena have been applied to the creation of three very important instruments: the field-emission microscope which has a resolution of 1 nm, the field-ion microscope, and the scanning tunneling microscope. The latter two instruments have atomic-level resolution.

Several important industrial applications of both field-electron emission and field ionization have been developed. They include flash x-ray emission, electron-beam lithography, and, very recently, flat-panel display technology.

1. PRINCIPLES

1.1 Field Emission

It is well known that thermally excited electrons can escape from a heated conductor by passing *over* the potential-energy barrier that confines electrons to the conductor. This is called thermionic emission. Electrons may also escape from a conductor by the application of a sufficiently high negative electric field. In this case, the electrons escape from the conductor by tunneling *through* the field-deformed potential-energy barrier, thereby generating a current of field-emitted electrons. Tunneling of an electron from a conductor into vacuum was first investigated by Fowler and Nordheim (1928), who employed the new wave mechanics of Schrödinger to calculate the tunneling current. This calculation was based on the fact that small particles, like electrons, could exhibit wavelike proper-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ties under some circumstances. The wavelength λ of the particle, in this case an electron, is given by

$$\lambda = h/p = h/(2mE)^{1/2}, \qquad (1)$$

where h is Planck's constant, p is the electron momentum, m is the electron mass, and E is the electron energy.

The distinction between thermionic emission and field emission may be understood by examining the potential-energy diagrams shown in Figs. 1 and 2 for electrons confined inside a conductor. According to the principles of classical mechanics, a particle confined, as shown in Fig. 1, cannot escape from within the potential-energy well unless it is given enough energy to reach the top of the potential-energy barrier and hence pass *over* it. This is what happens in thermionic emission when the electron acquires enough thermal energy when the emitting conductor is heated. However, the wavelike properties of the electron also allow the electron to pass *through* the barrier, at an appreciable rate, if the barrier is lowered below the level E_F in a distance comparable with λ. Solid-particle tunneling is analogous to the transmission of light through a transparent medium of varying refractive index.

The electron field-emission tunneling current density J is exponentially dependent on the ratio $\phi^{3/2}/F$ and was calculated by Fowler and Nordheim to be

$$J(0) = (CF^2/\phi t^2)\ \exp\{-Bv\phi^{3/2}/F\}, \qquad (2)$$

where F is the electric field strength, ϕ is the conductor work function, and C and B are constants, given by

$$C = e^2/8\pi h\phi t^2, \qquad (3)$$

$$B = 4(2m)^{1/2}/(3eh/2\pi). \qquad (4)$$

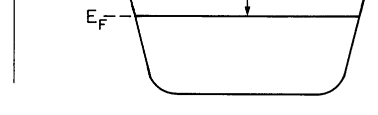

FIG. 1. Plot of potential energy for an electron confined within a structureless metal crystal. In order for the electron to escape it must receive an energy of ϕ, the metal work function, either from a photon or by heating of the metal.

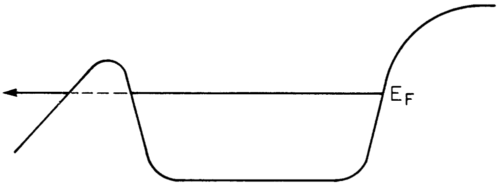

FIG. 2. The electron potential-energy curve of the metal crystal has been bent over by the application of a high electric field, $F = -dV/dx$, so that the slope of the downward sloping line is equal to the magnitude of the field. The deformation of the potential energy by the field now makes it possible for the electron at the Fermi level to "leak" out of the potential-energy well by quantum mechanical tunneling even at 0° K.

If J is expressed in A/m^2, F is expressed in V/m, and ϕ is given in units of eV, then $C = 1.54 \times 10^{-10}$ and $B = 6.83 \times 10^{-9}$. The terms t and v represent slowly varying dimensionless functions of $F^{1/2}/\phi$. Typically, $v \sim 0.5$. Fowler and Nordheim assumed that the electrons in the metal were confined by a box-shaped potential energy distribution, in which details of the electron energy band structure were ignored. The temperature dependence of electron emission, $J(T)$, is given by

$$J(T) = J(0)\ \sin(\pi p/d), \qquad (5)$$

where d and p are given by

$$d = ehFt/4\pi[2m\phi]^{1/2} = 9.75 \times 10^{-11} F/t\phi^{1/2}, \qquad (6)$$

$$p = kT/d, \qquad (7)$$

where k is Boltzmann's constant, d is in eV, and p is dimensionless.

Electric fields of the order of 3–5 GV/m are needed in order to thin the potential energy barrier sufficiently to allow appreciable amounts of tunneling. Fields of this magnitude are conveniently created by applying a negative potential V of a few thousand volts to a needle of small radius r separated from a grounded anode electrode a few cm away; then the electric field at the emitter apex is given by

$$F = V/kr, \qquad (8)$$

where k depends on the needle shape and, to a small extent, on the details of the grounded-electrode geometry. Typically, $4 < k < 8$, vs 1 for the case of a spherical conductor of the same radius.

Converting current density J to current I through $I = JA$ and using $F = V/kr$, the Fowler-Nordheim equation is converted to

$$I/V^2 = (AC/r^2k^2)\ \exp\{-Bkr\phi^{3/2}v/V\}. \qquad (9)$$

Hence a plot of $\ln(I/V^2)$ vs $1/V$ will give a nearly straight line of slope s given by

$$s = -Bkrv\phi^{3/2}. \qquad (10)$$

Here we have assumed $t \sim 1$. If r and k are known, then ϕ may be obtained by an iterative procedure in which it is initially assumed that $v = 1$. From this first approximation to the work function ϕ_1, a better value of v may be obtained by computing $v(F^{1/2}/\phi_1)$. This can be used to obtain a better value of ϕ by substituting the improved value of v in Eq. (10). Alternatively, work functions of individual crystal planes of the needle crystal may be obtained if I/V measurements are made for the planes and the work function of one plane is known. Work functions are then given by

$$\phi^{3/2}/\phi_{\text{ref}}^{3/2} = s/s_{\text{ref}}. \qquad (11)$$

Since I is exponentially dependent on $-\phi^{3/2}$, regions of the emitter needle of high work function will emit considerably fewer electrons than other regions of the needle. For clean field emitters, the close-packed planes—e.g., (110), (100) for tungsten—have higher work functions than the rest of the needle. Hence these planes emit considerably fewer electrons than other regions.

If the needle is brought very close to the counter electrode, the potential-energy barrier is thinned even when the potential difference between the two electrodes is only a few millivolts. This is the mechanism by which electron tunneling is effected in the scanning tunneling microscope (STM), first introduced by Young (1972) and later developed to achieve atomic resolution by IBM workers: Binnig, Rohrer, Gerber, and Weibel (1982). In this technique, a sharply pointed needle of a conductor is scanned across a surface at very close distances from it in order to develop a scanning image of the surface at very high magnifications and at atomic resolution. At close enough separations, a few millivolts difference in potential is sufficient to cause appreciable electron tunneling. Unlike the field emission case, tunneling at close distances can occur in either direction, depending on the relative potentials of the two electrodes. A distinction can be made between field emission, which involves electron tunneling from metal to vacuum (MV), and scanning tunneling microscope (STM) tunneling, which involves tunneling from metal to vacuum to metal (MVM). Semiconductors may, of course, be substituted for conductors.

The energy distribution of field-emitted electrons is of interest from both a theoretical and a practical point of view. The energy distribution is affected by surface conditions and, to a lesser extent, by bulk band-structure properties of the emitting needle.

For many metals, the electronic energy levels in the metals may be regarded as being distributed according to the free-electron theory of metals in which the density of energy levels per unit of energy increases with energy so that more electrons have energies at the Fermi level than at energy levels below it. The thickness of the high-field–reduced potential-energy barrier is smaller at the Fermi level than at energy levels below it. Both the reduced barrier thickness and the higher energy-level density favor electron emission at energies equal to the Fermi level so that more electron emission will occur at this energy level than any other, as shown in Fig. 3.

At low temperatures, few electrons populate energy levels in the metal at energies above the Fermi level. Hence, this is the cutoff level above which few electrons are emitted at low temperatures. As emitter temperatures are increased, more electrons occupy energy levels above the Fermi level and some tunneling begins to occur from these levels. This leads to a rounding off of the peak of the energy distribution, as shown in Fig. 4. Neglecting band-structure effects (see Swanson and Bell, 1973), the energy distribution of field-emitted electrons is given by

$$J(\epsilon) = J(0)\ \exp(\epsilon/d)\{1 + \exp(\epsilon/pd)\}/d. \qquad (12)$$

Plots of $J(\epsilon)$ vs ϵ for various values of T are shown in Fig. 4. At a temperature corresponding to $p = \frac{1}{2}$, a symmetric distribution of the emitted electrons occurs about $\epsilon = 0$, which corresponds to emission from the Fermi level. The half-width $\Delta E_{1/2}$ of the energy distribution at low temperatures is given approximately by

$$\Delta E_{1/2} = d\ \log_e 2. \qquad (13)$$

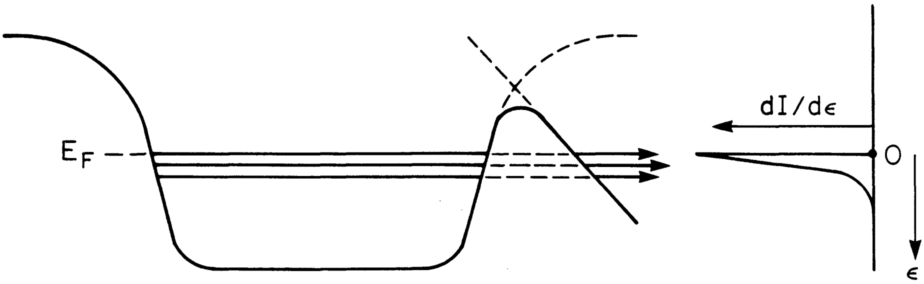

FIG. 3. An explanation of why the energy distribution curve of field-emitted electrons, shown on the right, is peaked at the energy corresponding to the Fermi energy level. There are no electrons above the Fermi level to supply electrons for tunneling, and below the Fermi level the tunneling probability drops off exponentially.

As shown by Gadzuk and Plummer (1971), at high temperatures (>1000 K) the energy distribution widens considerably beyond the amount predicted by Eq. (13).

There is considerable interest in $\Delta E_{1/2}$ because of the role it plays in limiting performance when high-intensity, small-diameter field-emitted electron beams are used in applications such as electron-beam lithography, a crucial technology needed for fabricating integrated circuit lithography masks.

At very high temperatures, a large proportion of the electron emission comes from electrons with enough thermal energy to surmount the potential-energy barrier. In this case there is a considerable further increase in the value of $\Delta E_{1/2}$ as emission comes both from tunneling through and emission over the potential-energy barrier.

Another interesting effect that arises as a property of the energy distribution is the possibility of both positive- and negative-energy exchanges of the field-emitted electrons with the emitter. If the average energy of the emitted electrons is less than that of the electrons arriving at the emitter surface via conductors of the circuit, heating of the emitter will occur. This will occur for values of $p<0.5$. For values of $p>0.5$, cooling of the emitter will occur because electrons will leave the emitter with more energy than provided by replacement electrons which arrive at the emitter at the Fermi energy level. The temperature at which the transition from cooling to heating occurs was investigated by Swanson *et al.* (1966) for tungsten field emitters.

1.2 Field Ionization

As suggested first by Oppenheimer (1928), the application of electric fields of the order of 1–50 GV/m to free molecules or atoms enables bound electrons to escape by tunneling in a process called field ionization (FI) depicted in Fig. 5. On the basis of their theoretical studies, Inghram and Gomer (1954) showed that experiments of Mueller and Bahadur (1956), in which they were able to obtain hydrogen-ion images of a tungsten needle cooled to −195 °C, were the first direct experimental evidence of field ionization.

Field-ion currents at conductor surfaces can be calculated by considering the issues of ion formation and gas supply to the conductor surface and the subsequent generation of ion current separately.

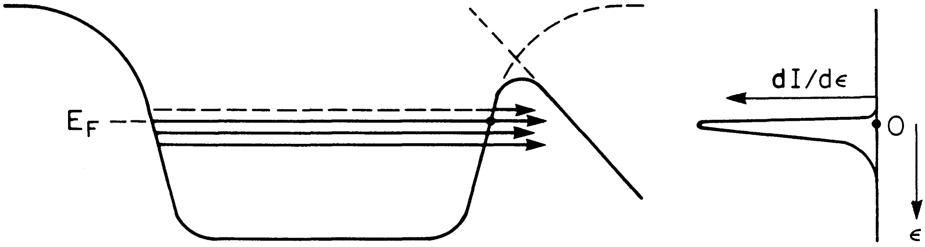

FIG. 4. The combined effect of temperature and electric field is to allow some electrons to tunnel from above the Fermi level, as shown by the dotted line.

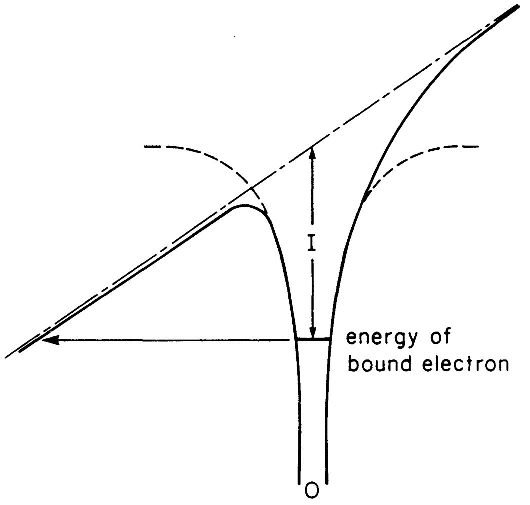

FIG. 5. The effect of the electric field on the potential energy of an electron for a field-free atom (dotted line) and for the potential energy in the presence of an electric field. Tunneling of the bound electron can now occur from the atom.

1.2.1 Ion Formation. The process of field ionization of an atom or molecule is directly comparable to field emission, except that the potential-energy box of the metal is replaced by the narrower, Coulombic well for the single atom or molecule.

The mean lifetime for the ground-state electron in the molecule is given by

$$\tau = 1/\nu D, \tag{14}$$

where ν is the frequency with which the bound valence electron arrives at the barrier and D is the tunneling probability of the electron through the electric-field reduced potential-energy barrier. Typical values of ν range from $4.1 \times 10^{16}\ \mathrm{s}^{-1}$ for H to $2.4 \times 10^{16}\ \mathrm{s}^{-1}$ for Ar. The expression for D, derived by the method of Wentzel, Kramers, and Brillouin, is a useful approximation:

$$D(E, V(x)) = \exp\{-(8m/h^2)^{1/2} \times [V(x) - E]^{1/2}\} dx, \tag{15}$$

where $V(x)$ is the electron potential energy and x_1 and x_2 are, respectively, the positions where the tunneling electron enters and exits from the potential energy barrier at the energy E.

Field emitters are also used to generate the high fields required for FI, except that the tips are sharper for FI, because FI requires higher field strengths than field electron emission from conductors. The higher fields for FI are a consequence of the fact that the ionization potentials of molecules are greater than the work functions of conductors.

Close up to a conductor surface, the required fields for FI are reduced because the tunneling electron energy is reduced by its image charge in the nearby metal. This is shown in the following expression, given by Mueller and Tsong (1969) for the potential energy of the tunneling electron at a distance x from the metal:

$$V(x) = -e^2/|x_i - x| + eFx - e^2/4x + e^2/|x_i + x|. \tag{16}$$

The first term is the Coulomb potential energy of the electron due to the positively charged ion, the second is due to the field, while the last two terms are due to the images in the metal of the electron and the ion.

Tunneling can only occur if there is an empty energy level available in the metal; hence, molecules closer than a critical distance x_c will not field ionize because the energy level of their ground-state electron is below the conductor Fermi energy level. An approximate value for x_c is given by

$$eFx_c = I - \phi. \tag{17}$$

Estimates of τ, based on the Wentzel–Kramers–Brillouin (WKB) model for He near W, range from 10^3 s at an electric field of 10 GV/m to 10^{-13} s at a field of 50 GV/m.

The validity of the above mechanism of ion formation has been confirmed by kinetic energy distribution measurements of field ionization carried out by Tsong and Mueller (1964). One consequence of the mechanism is a very narrow energy spread which is explained as follows. The kinetic energy of an ion will depend on how far away from the metal point field ionization occurs, because the potential energy of an ion will be $\sim eV(x)$ and decreases from a maximum close to the metal surface to a minimum at the collector. The maximum energy occurs for ions at the critical distance from the metal, x_c. The tunneling probability D declines rapidly with distance away from the emitter because it is a highly sensitive function of field and inverse potential-energy barrier width, both of which decrease the further away from the surface an atom is.

Tsong and Mueller (1964) found that the very narrow energy half-widths for the energy distributions obtained for He/W corresponded to an effective ionization zone of only 0.02 nm wide, a value that is about 30% less than the value predicted by the approximate one-dimensional theory discussed above.

Jason (1967) discovered additional weak, periodic peaks in the low-energy tails of the kinetic energy distributions of hydrogen and neon ions. These peaks were shown to correspond to quantized states (resonances) of a tunneling electron in the triangular well between the emitter's surface and the potential-energy line of the electric field.

1.2.2 Ion Current Generation. Ion current generation is a complex function of F, T, and emitter geometry, but may be discussed in terms of two limiting cases: high and low electric fields.

The ion current produced in a small volume near the emitter is given by

$$I=eS[1-\exp(-\Delta t/\tau)], \tag{18}$$

where S, the gas supply, is the number of particles entering the volume in question and Δt is the residence time a nonionized particle spends in the small volume. In the high-field limit τ is very small and $\Delta t>\tau$ so that

$$I\sim eS.$$

The quantity S was found to be considerably greater than the normal gas-phase supply function S_0 at zero field; S_0 is given by

$$S_0=A_0P(2\pi mkT)^{-1/2}. \tag{19}$$

where P is gas pressure, and A_0 is the area enclosing the ionizing volume.

The field enhancement of gas supply results in about an order of magnitude increase in supply and is a result of the reduced potential energy of a particle in the electric field due to polarization and dipole forces:

$$-V(F)=\mu E+\alpha F^2/2, \tag{20}$$

where μ is the permanent dipole of the particle and α is the polarizability of the particle. S is related to S_0 by

$$S=S_0\exp[-\pi V(F)/kT]. \tag{21}$$

At the low electric field limit, τ is long relative to the residence time of a gas particle in the very thin ionization zone above the emitter ($\Delta t<\tau$), and Eq. (18) simplifies to

$$I\sim eS_0\Delta t/\tau=enV_{\mathrm{FI}}/\tau, \tag{22}$$

where V_{FI} is the volume in which ionization occurs and n is the equilibrium gas density in that volume given by

$$n/n_0=(T_G/T_T)^{-1/2}\exp[-V(F)/kT_0], \tag{23}$$

where n_0 is the gas density at zero field, T_G is the temperature in the gas, and T_T is the temperature of the emitter surface, which is usually cooled well below room temperature. Equation (22) predicts a steep rise of current with field and this is observed.

The electric-field enhancement of gas density near the emitter surface for molecules with large dipole moments such as water or acetone can amount to several orders of magnitude. Beckey (1977) found that at fields of the order of 7 GV/m, multilayer adsorption occurred in accord with the theory of Brunauer, Emmett, and Teller (1938) for multilayer adsorption. Ion clusters, such as $[(H_2O)_n{*}H_3O]^+$, consistent with these ideas have been observed in field ionization mass spectrometry.

1.3 Field Evaporation and Field Desorption

As discovered by Mueller (1941), adsorbed atoms or molecules may be removed from a surface by field desorption as ions if the electric field is high enough. Field evaporation, a special case of field desorption, is the thermally activated evaporation of ions over the image-force (Schottky) barrier. This image-force barrier will be described in the following section.

1.3.1 Image-Force Model of Desorption. The potential-energy curves for an atom and for an ion, both in the absence of and in the presence of an electric field, are shown respectively in Figs. 6 and 7. If for an atom the desorption energy is Λ, then by using the Born–Haber thermochemical method, the barrier Q_0 to desorption of an ion of charge ne is given by

$$Q_0=\Lambda+\Sigma I_n-n\phi, \tag{24}$$

where ΣI_n is the sum of the ionization potentials up to the nth ionized state. This equation explains why the potential-energy curve for the ion is above that for the metal since

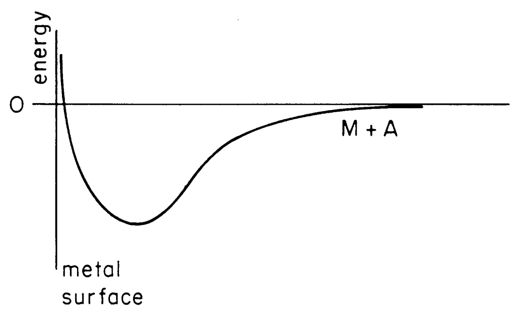

FIG. 6. Plot of the potential energy of an atom interacting with a metal surface.

$\Sigma I_n \gg n\phi$. The electric field reduces the potential energy of an ion at a distance x from the surface by eFx. There is a further lowering because of the interaction between the ion and its negatively charged image in the metal surface, located a distance x behind the metal surface. Hence, the total potential energy of an ion at x, relative to that of a free atom at infinity is

$$Q(F,x) = Q_0 - eFx - (ne)^2/4x. \tag{25}$$

The maximum in the potential-energy curve, obtained by differentiation, is at

$$x_m = (ne/4F)^{1/2} \tag{26}$$

and the energy maximum at this location is

$$Q(F,x_m) = Q = Q_0 - (n^3e^3F)^{1/2}. \tag{27}$$

Thermal evaporation, at the rate κ, can now occur over the energy maximum, known as the Schottky hump, according to

$$\kappa = \nu_p e^{-Q/kT}, \tag{28}$$

where ν_p is the vibration frequency of the bound particle. Substituting Eq. (27) into (28) and rearranging gives

$$F = (ne)^{-3}[\Lambda + I_n - n\phi - kT\ln(\kappa/\nu_p)]^2. \tag{29}$$

If F is minimized with respect to n, $F_{\min}$ will be the predicted evaporation field, and the nearest integral value of n that minimizes F will be the predicted charge state of the evaporating ion. Two polarization terms, $\mu_{\text{eff}}F$ and $\frac{1}{2}(\alpha_a - \alpha_i)F^2$, should also be added to Q_0. The first arises from the effective polarization of the double layer of electric charge that exists at the surface of a metal due to spillover of electrons from the metal, and the second is the interaction energy difference between that of an atom and that of the corresponding ion in an electric field. The α terms are the polarizabilities of an atom and an ion, and μ_{eff} is the effective dipole moment of the double electric charge layer. The image-force model, as embodied in this value of Q, yields quantitative predictions of desorption rate as a function of electric field that are in agreement with experimental results by Brandon (1965), and by Barofski and Mueller (1968).

1.4 Liquid-Metal Ion Sources (LMIS)

If a sufficiently high positive voltage is applied to a needle coated with a thin film of a liquid metal, the liquid at the needle apex will distort into a cone- or cusp-shaped form from the end of which issues a stream of positive ions. The onset of ion emission comes when the surface tension forces of the liquid metal are overcome by the stresses generated by the electric field. The end of the stressed liquid elongates into an extremely thin nipple of about 2–4 nm in diameter and about 10 nm long. This has been predicted by calculation carried out by Kingham and Swanson (1984) and verified by high-voltage electron microscopy work carried out by Ben Assayag and Sudraud (1984). The electric field at the end of the nipple is high enough for field evaporation of the liquid metal to occur at a rate that limits further growth of the nipple.

Such ion sources of liquid metals have been created from a variety of metals such as gallium, indium, aluminum, uranium, gold, and alloys of palladium and boron, palladium and arsenic, gold and silicon, etc. These alloy

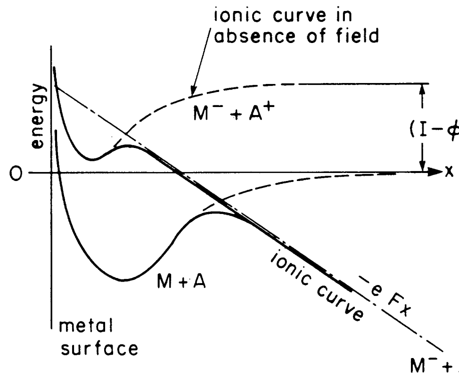

FIG. 7. Effect of an electric field on the potential-energy curve for an ionized atom A^+ and on that of a neutral atom A. At close distances from the metal, the neutral atom is favored, while at larger distances, the ionic species is clearly the lower-energy form.

ion sources can, in mass-separated ion-beam columns, provide ions of B, As, or Si. Separation is usually performed by use of a Wien filter, which consists of crossed (perpendicular) magnetic and electrostatic fields. The needles are typically 1–5 μm in diameter and can be fabricated from any conductor which will not be corroded by the liquid metal. The liquid metals are restricted to those that have vapor pressures low enough (Swanson and Bell, 1989) for the ion source to have a reasonably long life.

Emitted ions are characterized by energy spreads that are low enough ($\sim$5 eV) (Swanson, 1982) for LMIS to be of use in focused ion beam applications.

2. DEVICES AND INSTRUMENTS

2.1 Emitters

Points suitable for field-emission experiments with radii in the range 20–200 nm are easily fabricated by electrochemical etching of small-diameter (0.25 mm) metal wires. Usually, the process of drawing wires causes the wire to form crystals of the metal with a preferred orientation along the wire axis. For tungsten this is the (110) direction. Other orientations may be produced by zone melting of the wire in contact with a suitable seed crystal. For tungsten, electrochemical etching can be performed either with a positive potential of a few volts dc applied to the metal or with a few volts ac using sodium hydroxide as the electrolyte.

Room-temperature operation of field emitters requires ultrahigh vacuum because ions formed by electron impact return to the emitter and cause formation of rough, atomic-scale features over which enhanced electric fields exist. Electron emission increases at these sites and can run away leading to a vacuum arc that can destroy the emitter.

High-temperature operation of the emitter greatly enhances emitter life by causing increased surface self-diffusion of the emitter material. This causes the rough surface features to smooth away by surface diffusion. This healing mechanism is utilized in the 1800-K operation of (100)-oriented field emitters of W coated with a highly stable compound of zirconium and oxygen. Such emitters have lowered sticking coefficients for gases which would also otherwise change the I/V characteristics of the needle by adsorption-induced work-function changes. Because of the lowered work function of the W(100) plane due to the adsorbed ZrO overlayer at the needle apex of such emitters, there is an additional advantage: electron emission is confined to a cone of half angle $\sim$13 vs 30° for a clean emitter. Beam confinement is especially useful for electron optical applications.

Points suitable for field ionization experiments typically have radii in the range 10–100 nm. FI emitters are initially shaped by the same electrochemical etching techniques used for field-emission emitters. The final shaping of an FI emitter is performed in an FI microscope (see Sec. 2.3). As the voltage applied to an electrochemically etched emitter is gradually raised, the field strength over local protuberances increases until they are stripped away by field evaporation. The process is continued until the apex of the emitter attains a so-called field-evaporation end form in which the surface is atomically smooth and representative of the metal's crystallographic structure.

2.2 Field-Emission Microscope

In 1937, Mueller introduced the field-emission projection microscope (FEM) in which an image of the emitter surface was projected onto a fluorescent screen with a magnification M given approximately by the ratio of screen to emitter radii: R/r.

Actual magnification is reduced from this value by about 30% because of compression of the trajectories of the electrons due to the difference between the electric-field distribution for a sphere, for which the above expression is exact, and the actual field distribution for a needle-shaped emitter. The resolution of the microscope is about 1 nm and is limited by the fact that the electrons emerge from the emitter surface with a variation in velocities in the plane perpendicular to the emitter surface's normal. This results in the production of a blurred spot at the fluorescent screen that limits the ability of a viewer to distinguish between the projected images of adjacent surface features on the emitter.

A FEM image of a clean tungsten surface is shown in Fig. 8. When the needle is annealed, the point develops an end form that minimizes the surface energy of the needle. This

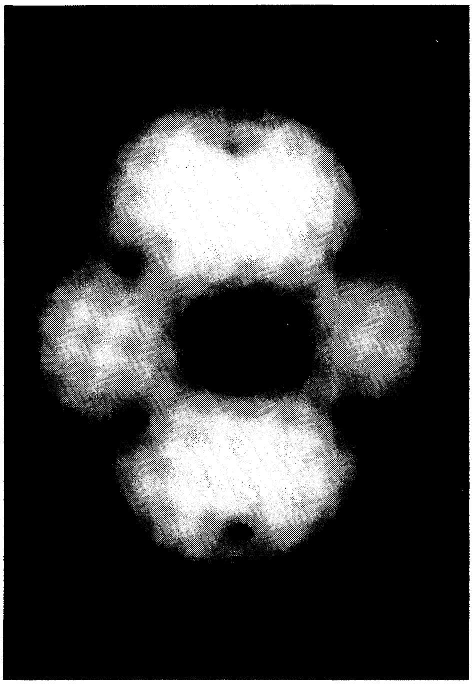

FIG. 8. Image of a field-emission pattern of a tungsten (110)-oriented needle. (Courtesy of L. W. Swanson, FEI Corp., Beaverton, Oregon.)

results in a mostly hemispherical shape on which low-energy, low-index planes of the metal exist as flats joined together by curved regions of the emitter surface.

The dark areas of the projected field-emission pattern correspond to the low-index planes, which have high work functions. These planes are distributed on the needle surface in a way that is determined by the crystal symmetry of the needle material. Hence the symmetry of the field-emission image is a reflection of the symmetry of the crystal structure of the tungsten needle. The FEM has been used extensively for the study of adsorption phenomena on clean, well-defined crystal surfaces.

2.3 Field-Ion Microscope

Mueller showed that, if a needle suitable for field-emission experiments is subjected to a reverse electric field, i.e., the needle was positive with respect to the extractor electrode, and a gas at low pressure was present, then ions are formed from the gas by field ionization. If, in addition, the needle is cooled to temperatures less than ~100 K, a projection image of the surface is formed on a fluorescent screen, as in the field-emission microscope case, with a resolution, however, that is sufficiently high to allow individual atoms of the surface to be distinguished. This can be seen in Fig. 9 which shows an image of a W needle, cooled to 20 K, in the presence of helium at 10-μm pressure.

The contrast mechanism is due to the fact that higher electric fields exist above the atoms compared to that between them so that the bright spots on the photograph correspond to individual atoms. The interiors of the single-crystal planes are not visible because the electric fields are too low.

Cooling of the needle causes an increase in ion current because gas atoms that hit the needle and are trapped there by polarization forces make jumps on the surface of the needle to a height that decreases with gas temperature. Consequently, they spend a longer time in the critical ionization zone, just above the needle surface, and so are more likely to be ionized at lower needle temperatures and thereby produce an increased current.

The resolution of the field-ion microscope, δ, is dependent on the needle apex radius r_t, needle temperature T, electric field at the emitter apex, F_0, gas atom atomic weight M, and gas atom size δ_0:

$$\delta = A(r_t/MF_0)^{1/2} + B(r_tT/F_0) + \delta_0, \qquad (30)$$

where A and B are constants. The above formula was derived in the book by Mueller and Tsong (1969). It can be seen that resolution is favored by low temperatures, high gas ionization fields F_0, small needle radii, and small gas molecule size. This is the reason that resolution is highest for helium, which has the highest value for F_0 and the smallest diameter. Helium also permits operation at temperatures <20 K where most other gases would freeze out as solids.

2.4 Scanning Tunneling Microscope

The atomic-resolution scanning tunneling microscope (STM), developed in 1981 by Binnig and Rohrer at IBM Zurich, Switzerland (see Binnig *et al.*, 1982), represents a revolutionary advance in microscope technology. The resolution in the plane of the sample is ~0.2 nm and much better in the vertical direction. This is good enough for imaging of

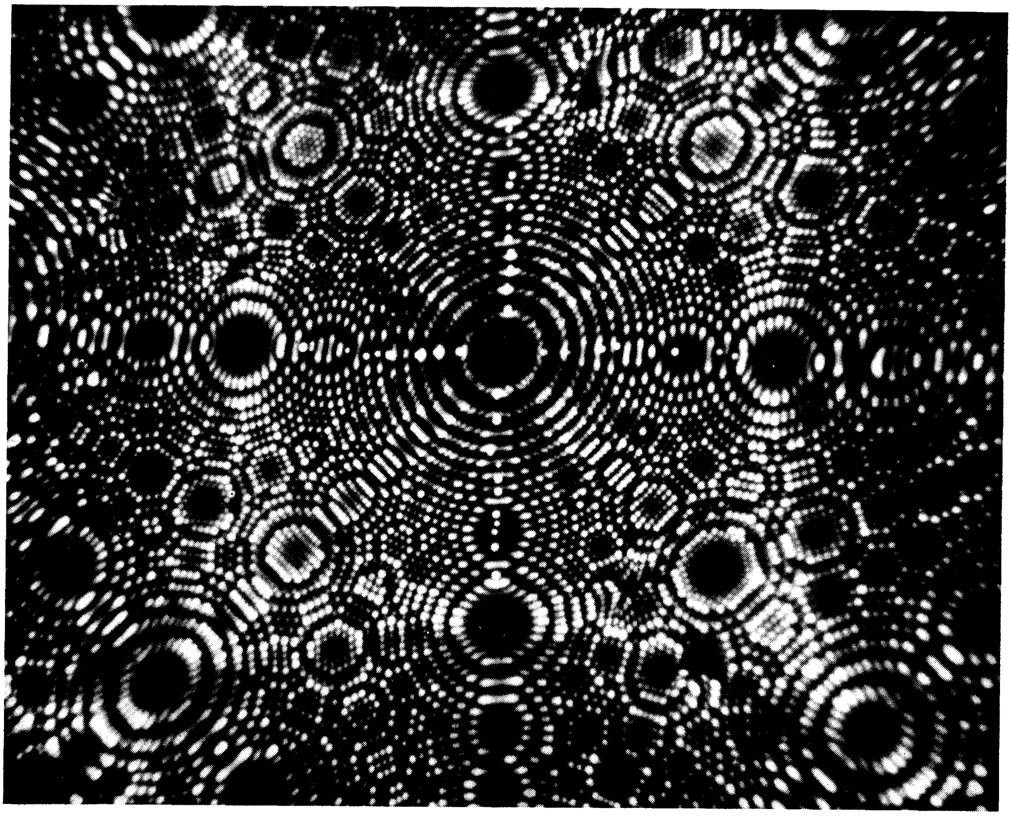

FIG. 9. Image of a field-ion microscope pattern of an iridium single crystal in which individual atoms at the edges of the single-crystal planes can be seen. In some cases, the atoms of small-diameter planes can be seen even in the interior of the planes.

individual atoms, as can be seen for the image in Fig. 10 of the surface of a Si(111)7×7 crystal. Because of the closeness of the electron source—the emitter tip—to the sample, the presence of gases at atmospheric pressure does not prevent operation of the STM, although it does present a problem for surface science studies involving clean surfaces which would rapidly become contaminated in the presence of atmospheric gases. The contrast in the STM image arises from the overlap of the wave function of an atom at the apex of the STM needle with that of a surface atom. This is often greater over a surface atom than that between atoms.

The success of the STM and its rapid development during the last decade are due in large part to the fact that STM workers have found techniques to overcome the following problems:

1. vibrations that tend to cause tip-damaging collisions between the tip and surface,
2. difficult to control tip-surface distances, and
3. conversion of STM data into useful surface images.

As discussed by Pohl (1986), vibrational problems have been greatly reduced by reducing the dimensions of the STM and making compact rigid devices. Tip-surface distances are being controlled to less than 0.1 nm by means of piezoelectric drivers. The development of superb graphics imaging software, mostly by IBM workers, has contributed greatly to the acceptance of STM technology.

2.5 Flash X-Ray Technology

If current densities are too high at the field emitter surface, the emitter end vaporizes and turns into a plasma that, in a period of ~10 ns, expands to fill the space between anode and

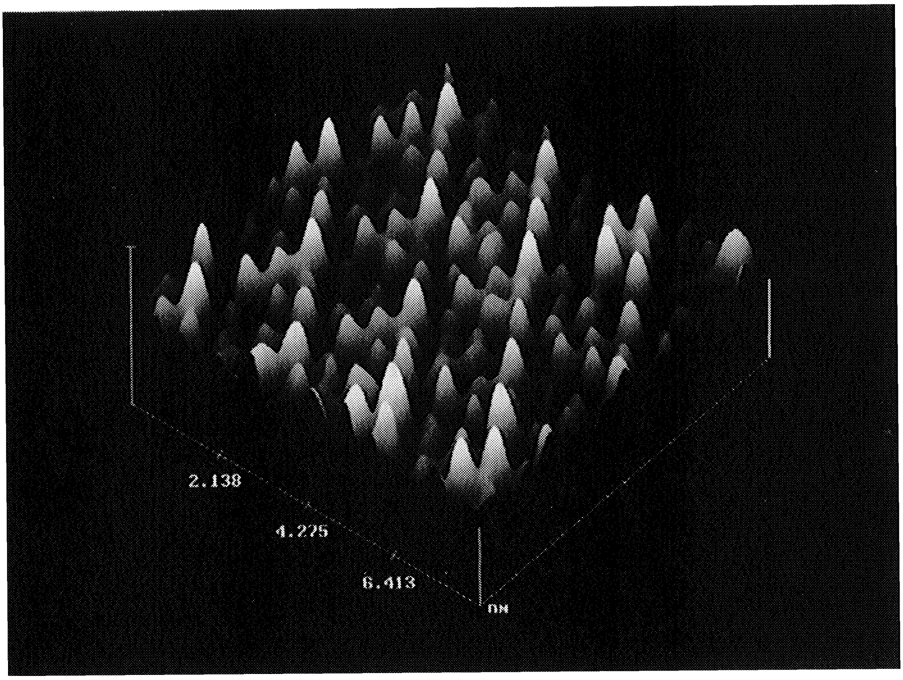

FIG. 10. A scanning tunneling microscope image of silicon showing the 7×7 reconstruction of the Si (111) surface. (Courtesy of Jeff Tetzlaff, Perkin Elmer Corp., Electronics Division.)

extractor electrode. This causes a tip-destroying arc. If, on the other hand, the field emission extraction voltage is turned off in time, the expanding plasma does not have time to reach the extractor and short out the electrodes. This is the basis of a flash x-ray technology for producing ultrahigh-power, short pulse-length, x-ray sources. By using extractor voltage pulses of ~10 ns or less, x-ray systems using electron bursts of up to 10 000 A and 2 MeV were developed by Dyke and co-workers (see Charbonnier *et al.*, 1963) at Femcor, now Hewlett Packard in McMinnville, Oregon.

2.6 Electron-Beam Lithography

The printing of patterns on resist-coated silicon wafers by passing light through lithographic masks is a key process in integrated-circuit manufacture. The mask layout, on the other hand, can also be created by using computer-directed, finely focused electron beams. Recently, field-emission sources using the zirconiated tungsten cathodes, discussed above in Sec. 2.1, have been used for high-performance electron-beam lithography systems devoted to the production of application-specific integrated circuits.

2.7 Flat-Panel Field-Emission Display (FED) Technology

A very promising recent application of field emission has emerged: the use of individually addressable field-emitter arrays for flat panel displays. A high-resolution color display using field-emission sources was recently demonstrated by Meyer and co-workers (see Meyer, 1991) at Leti Ceng, Grenoble, France. In a 256×256-line, 2-mm-thick, 6-in. diagonal monochrome display, lifetimes of >5000 h have been demonstrated. Intrinsic contrasts of >100:1 were attained with a brightness of 150–300 cd/m^2. The average power consumption for the screen was 1 W/dm^2.

The individual emitters of, presumably, molybdenum, deposited on a glass substrate, were 1.2 μm high and arranged 10 000/mm^2, allowing about 1000 emitters per pixel, thus providing a very high level of redundancy should individual emitters fail. A problem with field-emitter arrays is that only a small fraction contribute to emission because the emission current, as seen earlier, is exponentially dependent on emitter radius. Thus sharper emitters dominate emission because they emit at lower voltage than duller ones. An attempt to increase voltage to allow emission from duller tips leads to catastrophically high emission from the sharp emitters which fail through generation of a vacuum arc which can destroy a large fraction of all array emitters simultaneously. Leti workers overcame this problem by incorporating a resistive strip in the base of each emitter so that at high currents the voltage on an emitter was substantially less than that across low-emitting blunt emitters. This solution also greatly reduced the degree of flicker in the display.

According to Kuo (1992), the field-emission flat-panel display may play a prominent part in new flat-panel display developments because of its apparent advantage over competing display technologies of requiring fewer manufacturing steps.

2.8 Field-Emission Microwave Generation

Another promising application of the field-emission array technology, discussed in the previous section, is for generation of high-power microwaves in the frequency range 10–100 GHz. The first unequivocal demonstration of the possibility of microwave modulation of field-emission tungsten emitter arrays was carried out by Charbonnier *et al.* (1963) using a two-cavity device which these authors called the femitron: field-emission microwave generator. The analysis and experimental results of this work indicate that field-emission microwave amplifiers promise higher power than solid state devices, offer instant turn-on, unlike thermionic electron devices, and hold out the possibilities of much higher efficiencies than thermionic devices.

More recent discussions of the potentialities of field emission for microwave generation and amplification have been described at meetings of the recently instituted international conferences on vacuum microelectronics in which new developments of miniaturized field-emission devices are discussed.

GLOSSARY

Electron-Beam Lithography: The use of a focused beam of electrons to create patterns in resist materials in order to facilitate the fabrication of fine features for integrated-circuit fabrication.

Fermi Level: The maximum energy level of electrons in a metal at absolute zero or the energy for which the electron occupancy probability is 0.5. It is also the electrochemical potential of an electron in a solid.

Field Desorption: Removal of an adsorbed atom from a surface by application of a high electric field.

Field Emission: Extraction of electrons from a solid by tunneling.

Field-Emission Microscope: A high-resolution (~ 1 nm) point-projection microscope utilizing the emission of electrons by tunneling of the electrons through a field-deformed potential energy barrier.

Field Evaporation: Removal of an atom from a surface by application of a high electric field. This process was described by Mueller as a special case of field desorption.

Field-Ion Microscope: A point-projection microscope of atomic resolution utilizing atom ionization by the high electric field at the surface of a conductive needle.

Field Ionization: The process of ionization of an atom by tunneling of an electron from it through the electric-field–deformed potential energy barrier.

Image Force: The force between an electric charge and its induced opposite (image) charge in a conductor or semiconductor.

Liquid-Metal Ion Source: A novel high-intensity ion source generated by the application of a high electric potential to a liquid conductor coating a needle.

Scanning Tunneling Miscroscope: A recently developed microscope of atomic resolution which is based on the principle of tunneling of electrons from atom(s) of an atomically sharp conductive needle to adjacent surface atoms on the surface of a conductor or semiconductor.

Thermionic Emission: Emission of electrons brought about by heating.

Tunneling: Removal of an atomic-sized particle, in this case an electron, from an atom or a solid by utilizing the wavelike properties of very small particles which can then pass through a potential energy barrier if it is sufficiently thin.

Work Function: Energy required to remove an electron from a metal to a position just outside it.

List of Works Cited

Barofski, D. F., Mueller, E. W. (1968), *Surf. Sci.* **10**, 177–196.

Beckey, H. D., (1977), *Principles of Field Ionization and Field Desorption Mass Spectrometer,* New York: Pergamon.

Ben Assayag, G., Sudraud, P. (1984), *J. Phys. (Paris) Colloq.* **C9**, Suppl. 223–226.

Binnig, G., Rohrer, H., Gerber, C., Weibel, E., (1982), *Appl. Phys. Lett.* **40**, 178–180.

Brandon, D. G., (1965), *Science* **3**, 1–18.

Brunauer, S., Emmett, P. H., Teller, E., (1938), *J. Am. Chem. Soc.* **60**, 309–319.

Charbonnier, F. M., Barbour, J. P., Garrett, L. E., Dyke, W. P., (1963), *Proc. IEEE* **51**, 991–1004.

Fowler, R. H., Nordheim, L. W. (1928), *Proc. Roy. Soc. A* **119**, 173–181.

Gadzuk, J. W., Plummer, E. W., (1971), *Phys. Rev. B* **3**, 2125–2129.

Inghram, M. G., Gomer, R., (1954), *J. Chem. Phys.* **22**, 1279–1280.

Jason, A. J., (1967), *Phys. Rev.* **156**, 266–285.

Kingham, D. R., Swanson, L. W., (1984), Appl. Phys. A **34**, 123–132.

Kuo, H., panel discussion at 5th International Vacuum Microelectronics Conference, Hotel Chateau Wilhelminenberg, Vienna, Austria, 13–17 July 1992 (unpublished).

Meyer, R., (1991), in: S. Namba, Y. Nannichi, T. Utsumi (Eds.), *Technical Digest of the IV International Conference of Vacuum Microelectronics, August 22–24, 1991, Nagahama, Japan,* Osaka, Japan: Program Committee of the Fourth International Vacuum Microelectronics Conference.

Mueller, E. W., (1937), *Z. Phys.* **106**, 541–550.

Mueller, E. W., (1941), *Naturwiss* **29**, 533–537.

Mueller, E. W., Bahadur, K. (1956), *Phys. Rev.* **102**, 624–631.

Mueller, E. W., Tsong, T. T., (1969), *Field Ion Microscopy,* New York: American Elsevier, 1–312.

Oppenheimer, J. R., (1928), *Phys. Rev.* **31**, 66–81.

Pohl, D. W., (1986), *IBM J. Res. Dev.* **30**, 417–427.

Swanson, L. W., Bell, A. E. (1973), *Adv. Electron. Electron Phys.* **32**, 194–304.

Swanson, L. W., Bell, A. E., (1989), in: I. G. Brown (Ed.), *The Physics and Technology of Ion Sources,* New York: Wiley, 313–330.

Swanson, L. W., Crouser, L. C., (1969), *J. Appl. Phys.* **40**, 4741–4749.

Swanson, L. W., Crouser, L. C., Charbonnier, F. M., (1966), *Phys. Rev.* **151**, 327–340.

Tsong, T. T., Mueller, E. W., (1964), *J. Chem. Phys.* **41**, 3279–3284.

Young, R. D., (1972), *Rev. Sci. Instrum.* **43**, 999–1011.

Further Reading

Gomer, R., (1961), *Field Emission and Field Ionization,* Cambridge: Harvard University Press.

Swanson, L. W., Bell, A. E., (1973), *Adv. Electron. Electron Phys.*, **32**, 194–304.

Mueller, E. W., Tsong, T. T., (1969), *Field Ion Microscopy,* New York: American Elsevier.

Beckey, H. D., (1977), *Principles of Field Ionization and Field Desorption Mass Spectrometer,* New York: Pergamon.

FIELD-ION MICROSCOPY

GARY L. KELLOGG, *Sandia National Laboratories, Albuquerque, New Mexico, U.S.A.*

Introduction 393
1. **Physical Processes** 394
1.1 Field Ionization 394
1.2 Field Evaporation and Field Desorption 395
2. **Instrumentation and Principles of Operation** 396
2.1 Field-Ion Microscope 396
2.2 Atom-Probe Mass Spectrometer 399
3. **Applications of the Field-Ion Microscope and Atom-Probe** 402
3.1 Surface Science 402
3.2 Metallurgy and Materials Science 404
3.3 Novel Materials and Special Applications 405
4. **Summary** 405
Glossary 406
Works Cited 406
Further Reading 407

INTRODUCTION

The field-ion microscope was invented by the late Erwin W. Müller (Müller, 1951). Four years later, the first direct images of individual atoms on a solid surface were obtained with this instrument (Müller, 1956a). For nearly 20 years thereafter, the field-ion microscope had the distinction of being the only microscope capable of atomic resolution. Although it is now possible to resolve individual atoms in the high-resolution electron microscope (Crewe *et al.*, 1970) and scanning tunneling microscope (Binnig *et al.*, 1983), the field-ion microscope remains a powerful basic research tool for the investigation of atomic processes at solid surfaces.

The physical process that underlies the operation of the field-ion microscope is field ionization. In a high electric field, electrons may tunnel from gas-phase atoms or molecules to the surface of a solid. The loss of electrons results in the formation of positive ions. It is these "field ions" that produce direct images of the surface atoms in the field-ion microscope. In a related process known as field desorption, the surface atoms themselves are removed from the solid as positive ions. The discovery of field desorption led to the development of the atom-probe (Müller *et al.*, 1968), an instrument that is used to identify selected atoms observed in the field-ion microscope. Together, the field-ion microscope and atom-probe have been used to investigate a wide variety of problems in materials science and solid-state physics.

In this article the principles and applications of the field-ion microscope and atom-probe are reviewed. The processes that underlie the operation of the instruments, namely field ionization and field desorption, are discussed first. The emphasis in this section is placed on physical concepts rather than mathematical detail. A more extensive mathematical description can be found in the article FIELD EMISSION AND FIELD IONIZATION. Next, the design and operation of the instruments are reviewed. Included are discussions of the basic design of a field-ion microscope, methods of sample preparation, and the interpretation of field-ion microscope images. The different types of atom-probes currently in operation and the methods used to interpret atom-probe data are also discussed. Finally, selected applications of the instruments in the areas of surface science, metallurgy, and materials science are reviewed.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. PHYSICAL PROCESSES

1.1. Field Ionization

Field ionization is a process in which an atom or molecule loses one or more electrons by the application of a high electric field. The strength of the electric field required for field ionization is of the order of tens of V/nm. In the field-ion microscope the field-ionization process takes place near the surface of a solid sample (either a metal or a semiconductor). The high electric field is achieved by forming the sample into a sharply pointed needle (called a tip) and applying a positive voltage. The radius of curvature at the tip apex is of the order of tens of nm and the applied voltage is typically 3–20 kV.

Field ionization relies on the quantum-mechanical process of tunneling, for which there is no classical analog. The potential-energy diagram shown in Fig. 1 indicates how field ionization of an atom occurs at a metal surface. The metal is represented by a free-electron model with electronic states filled to the Fermi level. The uppermost electronic level of an atom near the metal surface is raised above the Fermi level of the metal by the application of a high electric field. If the width of the potential barrier between the atom and metal is sufficiently narrow (the width decreases as the height of the atomic level is increased above the Fermi level), an electron can tunnel from the gas atom to the metal. This tunneling process produces a positive ion next to the positively charged metal surface. Electrostatic repulsion accelerates the ion away from the surface to the fluorescent screen of the field-ion microscope.

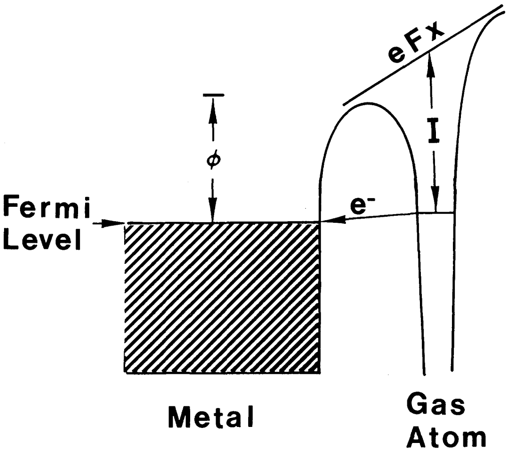

FIG. 1. Electron potential-energy diagram for an atom near a metal surface in the presence of an electric field *F*. In field ionization an electron tunnels from the gas atom to the metal.

In general, the tunneling probability increases as the atom moves closer to the surface and the barrier narrows. However, if the highest filled energy level of the atom falls below the Fermi level of the metal, tunneling cannot occur because there are no unoccupied states to receive the electron. This condition defines the critical distance of field ionization, x_c, given by

$$eFx_c = I - \Phi, \qquad (1)$$

where e is the electron charge, F is the applied field, I is the ionization potential of the atom, and Φ is the work function of the metal. At distances smaller than x_c field ionization is forbidden. Beyond the critical distance the field falls off very rapidly, resulting in a narrow zone in which field ionization can occur. It should be noted that the above relationship neglects contributions of the image potential and polarization effects, which are assumed to be small in comparison to I and Φ. For the typical case of helium ionized above a tungsten surface, Eq. (1) predicts that the critical distance of field ionization is approximately 0.4 nm.

The most widely accepted theoretical treatment of field ionization at a metal surface was originated by Inghram and Gomer (1954) and involves the application of the Wentzel-Kramers-Brillouin (WKB) method to determine the tunneling probability. The barrier penetration is given by an expression that depends on the potential and total energies of the tunneling electron. Various approximations for the electron potential energy have led to both numerical and analytic solutions for the tunneling probability. The total ion current, which is a product of this probability and the gas supply to the emitter apex, can be calculated and compared to experiment. Measurements of the energy distributions of field-ionized gases and the total ion current as a function of electric field have, in general, established the validity of this approach. Detailed discussions of field-ionization theory can be found in books by Gomer (1961), Müller and Tsong (1969), and Tsong (1990).

Atomic resolution with the field-ion microscope was achieved following the discovery that the lateral resolution increases dramatically at lower temperatures. It is observed experimentally that a temperature of approximately 20 K produces the best resolution (see Sec. 3.1). This observation, along with the fact that field ionization occurs away from the surface at the critical distance of field ionization, indicates that the gas atoms are not field ionized upon a single impact with the surface, but are thermally accommodated, i.e., their kinetic energy is reduced to that of the thermal energy of the surface atoms. Without thermal accommodation, there would be no temperature dependence of the resolution. The accommodation process involves a large number of hops from one impact site to another for each gas atom. Polarization attraction causes the atoms to return to the surface after each impact and concomitant loss of kinetic energy. Derivations of the theoretical resolution of the field-ion microscope assume that the gas atoms are fully accommodated to the sample temperature. The hopping-gas model has been used successfully to explain the measured ion current-voltage relationship.

It is now generally accepted that, in addition to the hopping-gas atoms above the sample surface, there is also a layer of adsorbed atoms on the surface. Inert gases such as He or Ne, which are typically used for imaging in the field-ion microscope, are weakly bound to the surface in the presence of a high electric field at low temperatures. The term used to describe this phenomenon is field adsorption. The binding has been described as an attraction between the field-induced dipole of the gas atom and its image dipole (Tsong and Müller, 1970). More recently, it has been suggested that the applied electric field raises the electronic levels of the gas atoms to a point where chemisorption may occur (Nath *et al.*, 1986). In either case, the binding is relatively weak compared to ordinary chemisorption, and the atoms are released at temperatures above 100–150 K. Atomic-resolution images of metal surfaces by the thermal-induced desorption of field-adsorbed gas atoms has been reported (Kellogg and Tsong, 1981).

1.2 Field Evaporation and Field Desorption

In addition to ionizing gas-phase atoms above a surface, the high electric field used in the field-ion microscope may also induce the removal of substrate surface atoms and adsorbates. The process is known as field evaporation when referring to removal of substrate atoms or, more generally, field desorption when referring to removal of either substrate surface atoms or adsorbates. In field-ion microscopy the imaging gas is typically chosen such that the field required for field ionization is somewhat less than the substrate evaporation field. Under these conditions, it is possible to raise the field to remove atoms for sample preparation (cleaning) purposes and then reduce the field for imaging. As a result, field desorption (evaporation) is used routinely in field-ion microscopy. As discussed later, field desorption is also the process that underlies the operation of the atom-probe field-ion microscope.

The process of field desorption is understood qualitatively in terms of one-dimensional potential-energy diagrams as shown in Fig. 2. The curve shown in Fig. 2(a) represents an atom at a surface in the absence of an applied electric field. The quantity λ is either the sublimation energy for field evaporation of a protruding atom at the edge of a plane or the binding energy for field desorption of an adsorbate. To remove the atom as an n-fold charged ion requires an energy given by

$$Q_0=\lambda+\sum_n I_n-n\Phi, \tag{2}$$

where I_n is the nth ionization potential of the atom and Φ is the work function of the surface. The situation is illustrated in Fig. 2(b). In an applied electric field, the ionic potential-energy curve is given by the superposition of the curve shown in Fig. 2(b) and the potential due to the field, $-neFx$. The result of this superposition is shown in Fig. 2(c). Field desorption is viewed as a thermal activation of the surface atom or adsorbate over the barrier indicated in Fig. 2(c). The activation energy for field desorption is therefore given by $Q_n(F)$, the height of the barrier.

In Müller's "image hump" model of field evaporation (Müller, 1956b), the ionic potential is approximated by a simple image potential, $-ne^2/4x$. The superposition of this curve with the field potential $-neFx$ results in a

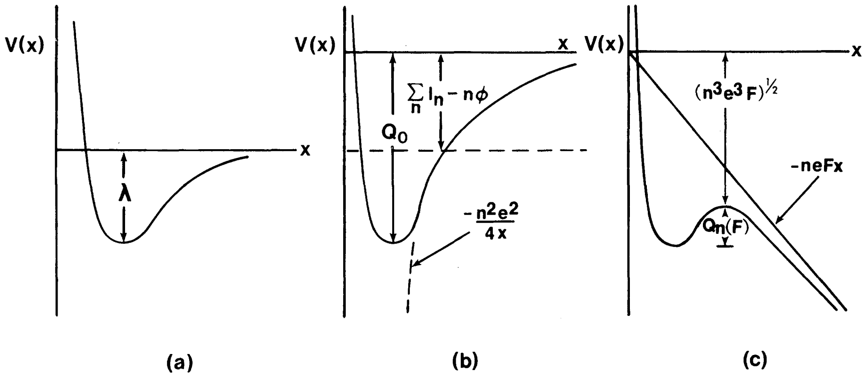

FIG. 2. Atomic and ionic potential-energy curves related to field desorption. (a) Atomic potential-energy curve for an atom held on the surface with a binding (or sublimation) energy λ. (b) Ionic potential-energy curve for an n-fold charged ion approximated by a simple image potential. (c) Ionic potential energy for the same ion in an applied field F. The energy required to remove the atom as an n-fold charged ion is $Q_n(F)$.

"Schottky saddle" with a maximum at $(n^3e^3F)^{1/2}$ below the zero-field maximum. The activation energy $Q(F)$ is given by the zero-field energy (the energy to remove the ion in the absence of a field) reduced by the energy at the Schottky maximum, i.e.,

$$Q_n(F) = Q_0 - (n^3e^3F)^{1/2}. \qquad (3)$$

In more refined models the difference in polarizability between the atom and the ion is included in the expression for $Q_n(F)$. Despite the simplicity of this model, it predicts the low-temperature evaporation fields and the ionic charge states of many metals with surprising accuracy.

In Gomer's "charge-exchange" model (Gomer, 1959), it is the crossing point of the atomic and ionic potential-energy curves that defines the maximum of the activation barrier. The activation energy is given by the difference between this intersection point and the minimum of the atomic potential-energy curve. Although the charge-exchange model is considered to be more realistic than the image-hump model (the image potential is not well defined at such small distances), it is less convenient for the calculation of measurable parameters because the exact forms of the curves are generally not known. However, by making suitable approximations, investigators have used this model successfully to analyze data on the field-evaporation rates of several metals.

More recently, quantum-mechanical models of field desorption have begun to emerge (McMullen and Perdew, 1987; Kreuzer and Nath, 1987). In the model of Kreuzer and Nath, the electronic properties of the metal are calculated within a tight-binding cluster model and the local electric field is taken from self-consistent jellium calculations (see COLLECTIVE PHENOMENA IN SOLIDS). A scaling law is derived from this approach which predicts experimentally measured evaporation fields to within 10%–20%, which is really no better than the above phenomenological models. However, the quantum-mechanical calculations are found to predict the field dependence of the activation energy for field evaporation of W and Rh much better than the image-hump model.

2. INSTRUMENTATION AND PRINCIPLES OF OPERATION

2.1 Field-Ion Microscope

The basic design of the field-ion microscope is surprisingly simple, given its ability to resolve individual atoms at a solid surface. The essential requirements are a vacuum system that can be charged with a low pressure of an inert gas (the imaging gas), a sharply pointed needle (the tip) cooled to 77 K or lower, a fluorescent screen, and a means to apply a high positive voltage to the tip. These requirements can be met by the simple design shown in Fig. 3. Although stainless-steel vacuum systems are standard nowadays, glass systems similar to this design were common prior to the 1970s.

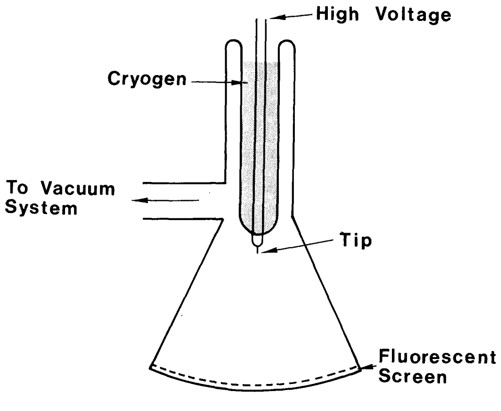

FIG. 3. Schematic diagram showing the essential components of a field-ion microscope. The voltage is referenced to the fluorescent screen, which is deposited on a conductive film.

The standard method for the preparation of sample tips used in field-ion microscopy is electrochemical etching. The starting material for metals is typically a piece of wire of 0.1–0.2 mm diameter. The wire is dipped (by hand or manipulator) into an appropriate chemical solution and a current is passed through it. Semiconductor samples are prepared from thin whiskers or slices of wafers and are sharpened in chemical solutions. Over the years, researchers have developed etching recipes for a variety of semiconductor and metal samples. Many of the polishing solutions are taken from those developed for the etching of thin-field transmission electron-microscope samples. Several tables of tip-polishing recipes are published in the literature (Müller and Tsong, 1969; Bowkett and Smith, 1970; Miller and Smith, 1989; Tsong, 1990).

The field-ion microscope was developed as an extension of the earlier field- (electron-) emission microscope (Müller, 1936). In fact, the same system design used for field-ion microscopy (e.g., Fig. 3) may also be used for field-emission microscopy. In field-emission microscopy a negative voltage is applied to the tip in ultrahigh vacuum. The required voltage for the field-emission microscope is about a factor of 10 lower than that needed for the field-ion microscope (see below). The negative electric field at the tip apex ejects electrons in a process known as field emission (the electrons tunnel out from the solid) and these electrons form an image of the tip surface on the fluorescent screen. Although simpler to operate than the field-ion microscope, the field-emission microscope has a resolution of 3–4 nm, not sufficient to resolve individual atoms.

To operate a field-ion microscope, the chamber is evacuated to a pressure in the 10^{-6} Pa range or lower and several tenths of a Pa of the imaging gas is introduced (He and Ne are common). The tip is cooled with a cryogen and a positive voltage in the range of 3–20 kV is applied to the emitter. The high electric field at the tip apex (20–50 V/nm) ionizes imaging gas atoms as discussed above and projects them to the fluorescent screen. The pattern of spots that forms on the screen is a direct image of the protruding surface atoms.

Contrast variations in the field-ion microscope result from the fact that field ionization occurs preferentially above the protruding surface atoms. The enhanced electric field above the protruding atoms causes the hopping-gas atoms (see above) to have a higher probability of ionization in these localized regions. Ions formed above the atoms follow the nearly radial electric field lines to the fluorescent screen and produce a highly magnified image of the surface.

An example of a field-ion image from a rhodium tip is shown in Fig. 4. The small spots in this micrograph correspond to individual surface atoms. It is clear from Fig. 4 that there are regions on the surface where the atoms do not protrude sufficiently to be imaged (the dark circular areas). These regions correspond to the low-index (smooth) single-crystal planes of the multifaceted surface. The atoms within these planes do not protrude out far enough to cause local field variations and

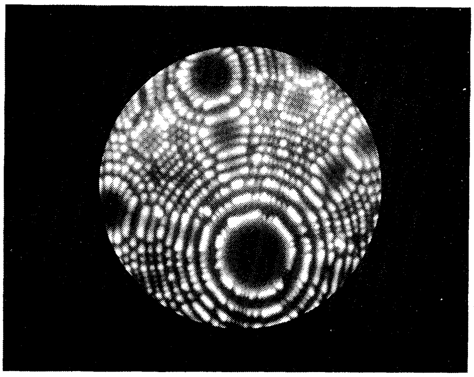

FIG. 4. An example of a field-ion microscope image from a sample of rhodium.

differences in contrast. The concentric rings of spots that surround the dark regions correspond to the edge atoms of successively deeper atomic layers. The edge atoms of the different low-index planes join each other to form the higher-index planes. The field-ion image is a stereographic projection of the hemispherical tip surface. From the symmetry of the spot patterns and the known crystallographic structure of the sample (e.g., pure metals), it is relatively easy to determine the Miller indices of the different crystal faces observed in the image. For alloy samples and more complicated structures, computer simulations of field-ion images are very useful for obtaining the crystal structure.

The field-ion microscope is a point-projection microscope and therefore has a magnification given by

$$M=D/\beta r, \tag{4}$$

where D is the tip-to-screen distance, r is the tip radius, and β is an image compression factor that accounts for the fact that the electric field lines do not follow exactly the radial direction from the tip to the screen (the field lines are compressed by the presence of the tip shank). The value of β typically is in the range from 1.5 to 1.8. Thus, a magnification of several million is obtained from a tip with a radius of curvature of a few tens of nm. Because the ions are generated very close to the tip surface, there is no relative motion of the surface with respect to the beam of field ions. As a result, external vibrations are not magnified as in conventional electron microscopes or scanning tunneling microscopes and means to isolate the system from such vibrations are not required.

The resolution of the field-ion microscope is determined by three factors: the spatial extent of the region above the protruding surface atom where ionization occurs, the lateral velocity component of the emerging ions, and diffraction effects due to the finite de Broglie wavelengths of the field ions. The spatial extent of the ionization zone is defined by the overlap region of the protruding surface atom and imaging gas atom. The lateral velocity component of the field ions and the diffraction effects have been calculated, and an expression for the overall resolution has been derived (see, for example, Müller and Tsong, 1969; Miller and Smith, 1989; Tsong, 1990). From this expression it has been found that the resolution of the field-ion microscope improves with increasing field strength, decreasing temperature, and decreasing tip radius. On the basis of these criteria, He is the best choice for an imaging gas because it has the highest ionization field and can be used at low temperatures without condensation onto the sample surface. However, as discussed below, many materials cannot withstand the high fields required for imaging in He.

The processes of field evaporation and field desorption are used routinely in field-ion microscopy to obtain the smooth, nearly hemispherical end forms required to obtain images like the one in Fig. 4. When one first views a sample in the field-ion microscope, there are usually asperities, crystallographic irregularities, oxides, and other surface contaminants left from the electrochemical etching process. With increasing field, the surface contaminants are field desorbed and the asperities and irregularities are removed by field evaporation. Once a smooth end form is obtained, the field-evaporation process proceeds in a uniform fashion. Atoms in the vicinity of the low-index planes are removed from the edge of the plane inward, one atomic layer at a time. Thus, it is possible both to clean the surface and to probe into the bulk of the sample with the use of field desorption and evaporation.

The field strength at which the sharpest image is produced in the field-ion microscope is known as the best image field. In order to obtain stable images, the evaporation field of the material under investigation must be higher than the best image field. Although somewhat subjective, the best image fields for the common imaging gases are generally agreed to be 44 V/nm for He, 37 V/nm for Ne, and 22 V/nm for H_2 and Ar. It has been found that only the most refractory metals have evaporation fields higher than the 44 V/nm required for imaging in He. Many other materials are suitable for imaging in Ne or Ar, but the heavier ions have a much lower efficiency for exciting the phosphor of the fluorescent screen.

A major breakthrough in field-ion microscopy came with the development of channel-plate image intensifiers. A channel plate, placed just in front of the fluorescent screen, converts the relatively weak ion image into an intense electron image by an electron multiplication process. The electron image is proximity focused onto the fluorescent screen without loss of resolution. With channel plates it is now possible to apply the technique to almost all metals and alloys, most semiconductors, and even a few exotic materials such as the recently discovered ceramic superconductors.

2.2 Atom-Probe Mass Spectrometer

Except for very special conditions, the field-ion microscope provides an image of the atomic structure only. However, the chemical identity of atoms observed in the field-ion microscope can be determined with a related instrument known as the atom-probe. The atom-probe is a time-of-flight mass spectrometer that analyzes ions generated at the surface of a field-ion tip by field evaporation or desorption. With the atom-probe, field-ion microscope combination, it is possible to image individual atoms on a solid surface, to remove selected atoms in a controlled layer-by-layer fashion, and to determine the chemical identity of the atoms as they are removed. Given these capabilities, it is clear that the atom-probe field-ion microscope is a microanalytical probe unsurpassed in sensitivity.

The first atom-probe was introduced in 1968 (Müller *et al.*, 1968) and is now known as the probe-hole atom-probe or conventional atom-probe. Shown schematically in Fig. 5, the probe-hole atom-probe is a field-ion microscope that has been modified to permit time-of-flight mass analysis of field-desorbed surface atoms. The tip is mounted on a manipulator allowing preselected atoms to be aligned with a small hole in the viewing screen. Surface atoms are removed as positive ions by field desorption. The desorption takes place over a very short interval of time (tens of ns) with the application of a high-voltage (0.5–3.0 kV) pulse to the tip. The tip voltage is added to the dc imaging voltage such that the sum of the two voltages removes surface atoms at a reasonable rate. The ions of interest travel through the probe hole and along a drift tube to a detector sensitive to the impact of individual ions. From the known drift length (l) and measured flight time (t), the mass-to-charge ratios (m/n) of the desorbed ions are determined from the simple expression

$$m/n = (2e/l^2)(V_t + V_p)t^2, \tag{5}$$

where e is the elementary charge and V_t and V_p are the applied dc and pulse voltage, respectively. The mass-to-charge ratio specifies (although sometimes not uniquely) the chemical identity of the desorbed ions.

In most applications the atom-probe is not used to identify individual surface atoms, but rather to probe the composition of a selected volume of the sample material. In this case the probe hole covers an area extending a few nm over the tip surface and the tip is pulsed repetitively. As atoms are removed, the analysis extends into the bulk of the sample. Data obtained in this manner (the ion flight times) may be expressed in several forms. The most common procedure is to plot a simple mass spectrum, i.e., a histogram of the ion signal intensity (the number of ions of a given m/n ratio) versus the m/n ratio. The intensities may be plotted either linearly or logarithmically, the latter having the advantage of emphasizing the low-concentration species. An example of an atom-probe mass spectrum from a $YBa_2Cu_3O_{7-x}$ sample is shown in Fig. 6.

Ion signals may also be plotted as "ladder diagrams" or "cumulative profiles." In these forms of data representation the number of ions of a given mass-to-charge ratio is plotted against the number of ions of a different species (ladder diagram) or the total number

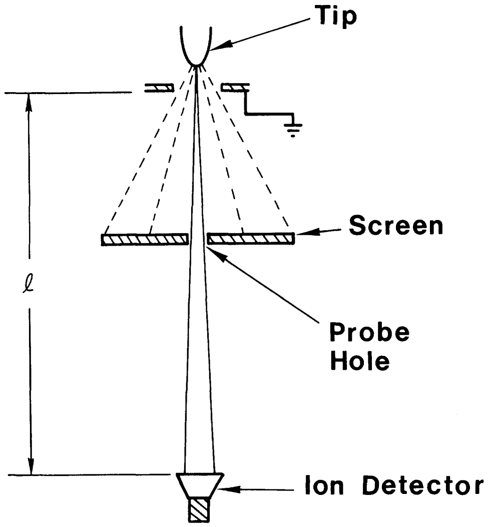

FIG. 5. Schematic diagram of a probe-hole time-of-flight atom-probe field-ion microscope.

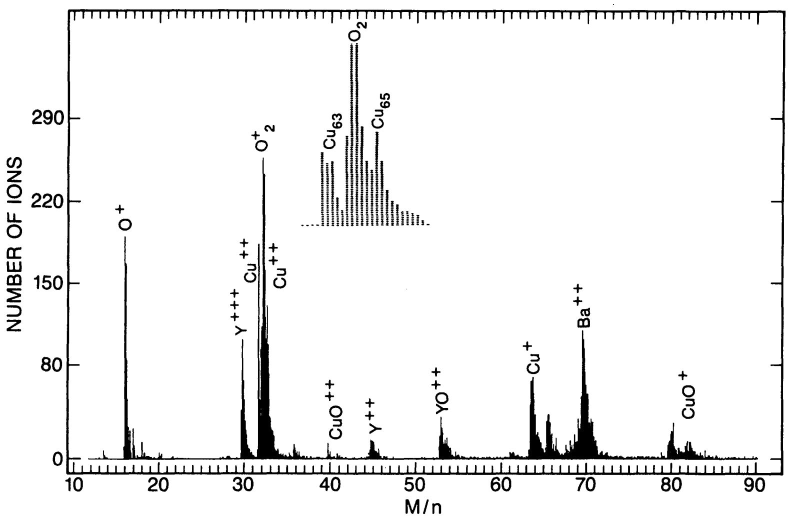

FIG. 6. An example of a time-of-flight mass spectrum taken from a sample of $YBa_2Cu_3O_{7-x}$. [From Kellogg and Brenner (1987).]

of ions detected (cumulative profile). The slopes of these plots can be used as a quick visual aid to determine how the concentration of a given species varies with respect to another species or with all species through the probed region. An example of a ladder diagram from atom-probe analysis of an oxidized Rh sample is shown in Fig. 7.

The probe-hole atom-probe described above has two major limitations. The first is its mass resolution, which is limited primarily by the energy spread of the ions as they enter the drift region. The energy spread arises from premature field desorption (ions that are desorbed before the voltage pulse reaches its maximum value). The resolution of the conventional atom probe is typically $M/\Delta M \sim 250$. The second limitation is the small extent of the probe area over which ions are collected. These restrictions have led researchers to develop a number of variations of the original atom-probe design.

One way to improve the mass resolution of the atom-probe is to use a magnetic-sector (instead of a time-of-flight) mass spectrometer to analyze the desorbed ions. An instru-

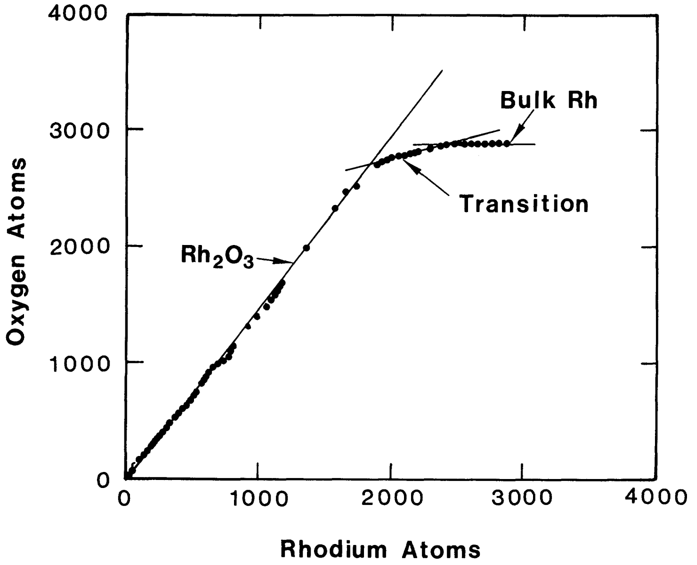

FIG. 7. An example of a ladder diagram resulting from atom-probe analysis of an oxidized rhodium tip. [From Kellogg (1987b).]

ment based on this concept was developed by Barofski and Müller (1968) and improved by Müller and Sakurai (1974). Although the resolution ($M/\Delta M \sim 2000$) is much better than that of the conventional probe-hole atom-probe, the magnetic-sector instrument suffers from a narrow mass range and low transmission coefficient. As a result, it has seen only limited application.

A more common way to improve the mass resolution of the conventional atom-probe is to replace the straight drift tube with an energy-compensating lens (Müller and Krishnaswami, 1974). The lens system used in most atom-probes is based on a 163° sector design developed for other applications by Poschenrieder (1971). The mass resolution of the energy-compensated atom-probe is better than $M/\Delta M = 2000$. This system also has the advantage that noise signals from the background imaging gas are filtered out and accurate elemental analysis can be carried out while viewing the field-ion microscope image. The energy-compensated atom-probe is the state-of-the-art system for applications in metallurgy and materials science.

A third method to improve the mass resolution of the atom-probe is to eliminate the high-voltage pulse entirely and replace it with a dc voltage and a short-duration laser pulse. In the pulsed-laser atom-probe, the dc voltage is set below the evaporation field of the material, and the thermal stimulation provided by the laser pulse initiates the field-evaporation process. Time-of-flight mass analysis is carried out the same as in the conventional atom-probe. With the use of laser pulses, the kinetic energy spread of the desorbed ions is negligible. Although the first pulsed-laser atom-probe systems, introduced independently by Kellogg and Tsong (1980) and Drachsel *et al.* (1980), were not designed for high mass resolution, subsequent versions with a straight drift-tube configuration were able to achieve mass resolution better than even the energy-compensated atom-probe (Tsong *et al.*, 1982). The pulsed-laser atom-probe has the additional advantage of being able to analyze materials whose electrical resistivity is too high to transmit the nanosecond pulses in the voltage-pulsed atom-probe. This capability has greatly facilitated atom-probe investigations of semiconductor materials. The ability to induce desorption at lower field strengths has also stimulated investigations of surface chemical reactions with the pulsed-laser technique. A general review of the pulsed-laser atom probe technique and its applications has been given by Kellogg (1987a).

In all of the atom-probe designs mentioned above, the diameter of the probed area on the surface is just a few nm. This design is particularly well suited for the study of impurities at lattice defects, etc., but poses a limitation for adsorption studies because the total number of ions collected is very small. This limitation can be addressed with the use of the imaging atom-probe (Panitz, 1973). In this design the viewing screen of the field-ion microscope is replaced with a set of channel-plate ion detectors. By stacking two plates in series, the impact of individual ions may be detected. The drift region in this system is the tip-to-screen region, i.e., there is no separate drift tube. Ions desorbed from anywhere within the imaged area of the surface produce both a current signal that can be used for time-of-flight mass analysis and an image spot on the screen. The imaging feature of this atom-probe makes it possible to obtain elemental maps of selected surface species. The association of a given chemical species with a specific structural feature or defect is easily seen in these elemental maps. However, the shorter flight path imposes an inherent mass resolution of $M/\Delta M < 100$.

Elemental maps may also be obtained with the position-sensitive atom-probe developed by Cerezo *et al.* (1988). As with the imaging atom-probe, ions are collected from the entire imaged area of the surface (no probe hole). However, instead of a fluorescent screen behind the image intensifier, this instrument uses a position-sensitive anode (wedge-and-strip) to record the arrival of ions. If the desorption rate is maintained at less than one ion per pulse, the chemical identity and position of each ion arriving at the detector can be determined and stored in a computer. After field evaporation through the near-surface region, a three-dimensional map of each element in the analyzed volume can be reconstructed from the stored data.

In general, the choice of a particular atom-probe design is based on its intended application. The conventional (probe-hole) and energy-compensated designs are used most often in metallurgical applications. The ability to carry out chemical analysis in a highly localized region of the sample makes the atom-probe an ideal instrument to investigate impurity segregation to defects, precipitate formation in alloys, and other atomic-level phenomena. The imaging atom-probe is better suited to investigations of adsorption and surface reaction processes than the probe-hole instruments, but has seen applications in both areas. The pulsed-laser atom-probe is used for semiconductor studies and also adsorption studies where the ability to vary the electric field over a wide range is important. The position-sensitive atom-probe is a relatively new instrument, but should develop into a useful tool for investigations of a wide range of problems in metallurgy and materials science.

3. APPLICATIONS OF THE FIELD-ION MICROSCOPE AND ATOM-PROBE

3.1 Surface Science

The most impressive feature of the field-ion microscope is its ability to obtain beautiful, atomically resolved images of metal surfaces. It would seem therefore, that an obvious application of the instrument would be the determination of surface atomic structure. Since the bulk structure for the common metals examined in field-ion microscopy is known from x-ray diffraction, the field-ion microscope should be able to determine if the surface atomic structure is a simple termination of the bulk structure or if the surface atoms are displaced from their bulk positions. A surface whose atomic structure differs from the bulk-terminated structure is said to be "reconstructed." Unfortunately, for many of the surface planes of interest (primarily atomically smooth planes having low Miller indices), the interior atoms are not resolved and no information on atomic structure is available. Even for those crystal planes that are fully resolved, there is a concern that the applied electric field may have an influence on the atomic structure, and the imaged positions of the atoms may not reflect the true positions of the surface atoms under equilibrium conditions. As a result, the field-ion microscope has not seen widespread application over the years as a probe of surface atomic structure.

Within the past several years, however, researchers have begun to investigate certain types of surface reconstructions with the field-ion microscope. The uncertainty with respect to the electric field mentioned above is alleviated by studying systems for which the reconstructed surface is energetically very stable and transformations from the bulk-terminated to the reconstructed atomic structure can be observed. An example is the (110) surface of Pt. Field evaporation at low temperatures (77 K or less) produces the bulk-terminated structure, which, in the case of an fcc (110) plane, is a periodic array of close-packed atomic rows. The bulk-terminated structure is stable only at these very low temperatures. Upon heating to room temperature (or above) in the absence of any applied fields, the surface atoms rearrange to form the equilibrium structure. This structure can be "frozen in" by quenching the surface back to 77 K or less. For the Pt(110) plane, imaging with the field-ion microscope reveals that the equilibrium surface structure is the so-called missing-row reconstruction, in which every other close-packed row of atoms is absent (Kellogg, 1985a). This observation confirms atomic models of the reconstruction of macroscopic Pt(110) surfaces based on low-energy electron-diffraction and ion-scattering studies. Moreover, the results demonstrate that such reconstructions exist even for very small clusters of atoms (as few as five) and involve very short-ranged atomic interactions. Field-ion microscopy has been used subsequently for similar studies of reconstructions on other surfaces. The results of these investigations are reviewed in detail in Chap. 4 of the recent book by Tsong (1990).

An area of surface science where the field-ion microscope has seen much more extensive application is the investigation of the diffusion and clustering of individual atoms on top of perfectly ordered solid surfaces. In surface diffusion studies, the movement of a single atom on a perfectly ordered surface is tracked as a function of time and temperature. As an example, Fig. 8 shows snapshots of a nickel atom as it moves around on the (110) surface of tungsten. In studies of atomic interactions, the motion of two or more atoms is examined. Pioneered by Ehrlich in the mid-1960s (Ehrlich and Hudda, 1966), field-ion microscope studies have provided diffusion parameters and interaction energies for a large number of atom-surface combinations. Although atomic resolution has been achieved with other techniques, the field-ion micro-

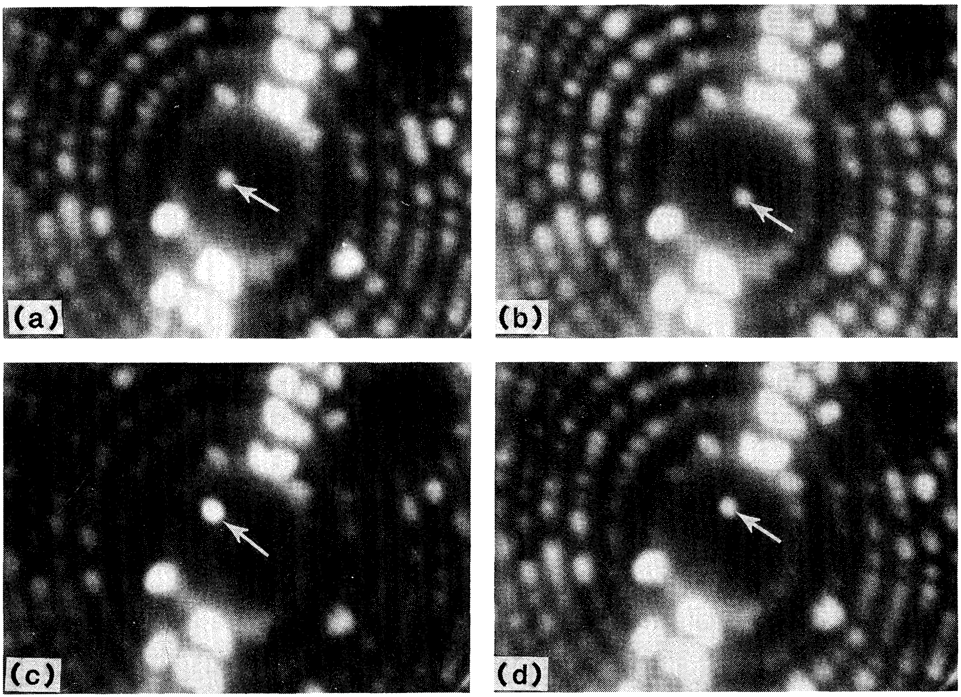

FIG. 8. Field-ion micrographs showing the migration of a nickel atom on the (110) plane of tungsten. The images were recorded with the tip at 77 K. Between each micrograph the tip was heated to a temperature of 200 K for 1 min. The arrow points to the nickel atom, which performs a two-dimensional random walk. [From Kellogg (1987c).]

paralleled in its ability to obtain quantitative data on the diffusion and interaction of individual atoms on surfaces. Field-ion microscope studies of surface diffusion and atomic interactions have been reviewed in articles by Kellogg *et al.* (1978), Ehrlich and Stolt (1980), Bassett (1983), and Tsong (1988, 1990).

A surprising discovery that has resulted from field-ion microscope investigations of single-atom surface diffusion is that the migration of single atoms on certain surfaces does not take place by the site-to-site hopping process one would expect intuitively. Instead, the atom migrates by a series of exchange displacements with atoms in the top layer of the substrate. First observed on the grooved (110) surfaces of Pt (Bassett and Webber, 1978) and Ir (Wrigley and Ehrlich, 1980), exchange displacements have recently been shown to occur on the smooth (100) surfaces of these same metals (Kellogg and Feibelman, 1990; Chen and Tsong, 1990). The exchange mechanism on the fcc (100) surfaces is deduced from maps of the sites visited by the migrating atom. Diffusion by the exchange process produces different maps than does diffusion by ordinary hopping. The exchange-diffusion mechanism has important implications in various crystal and epitaxial growth processes.

The investigation of small clusters of atoms with the field-ion microscope has also provided fundamental insight concerning the interaction of atoms at surfaces. Especially noteworthy is the finding that atomic interactions cannot be modeled by simple pairwise interactions. This discovery is clearly demonstrated by the observation that Re atoms on the W(110) surface do not form stable two-atom clusters, but readily form stable three-atom clusters (Fink and Ehrlich, 1984). Three-body or higher-order interactions are required to explain this observation. Many-body interactions are also needed to explain the stability of atomic configurations observed in cluster nuclei on the (100) surfaces of Ir and Pt (Schwoebel and Kellogg, 1988; Schwoebel *et al.*, 1989). On Pt(100), for example, field-ion microscopic observations show that the stable configuration oscillates between a linear chain and a two-dimensional island as the number of atoms in the cluster is increased of atoms in the cluster is increased from three to six. Theoretical calculations indicate that up to four-body interactions are needed to explain this observation (Schwoebel *et al.*, 1989).

The chemical analytical capabilities of the atom-probe mass spectrometer have extended applications of the field-ion microscope to other areas of surface science. One example is surface segregation in metal alloy systems. Often it is found that under equilibrium conditions the surface and near-surface layers of an alloy are enriched or depleted in one of the alloy components. With the atom-probe field-ion microscope and controlled field evaporation, it is possible to obtain composition profiles with single-atomic-layer depth resolution. Investigations by Tsong and co-workers have shown that in certain alloy systems the concentration of the segregant changes from enrichment to depletion upon removal of successive layers of surface atoms (Tsong *et al.*, 1978; Ng *et al.*, 1979). The atomic-layer depth resolution of the atom-probe is essential to observe this behavior experimentally.

The atom-probe and field-ion microscope have also been used to investigate the reaction of gas-phase adsorbates on metal surfaces. Information on surface morphological and compositional changes induced by various reaction processes including surface oxidation have been obtained in studies with the field-ion microscope and atom-probe, respectively. For example, in one such study the inhibition of the CO oxidation reaction on Rh by the formation of a surface oxide was clearly demonstrated (Kellogg, 1985b). Detailed studies of the kinetics of several reaction processes have been carried out with a pulsed field-desorption technique developed by Block and co-workers (Block *et al.*, 1986). Field-ion–related techniques are particularly attractive for the study of surface reactions related to catalysis because the analyzed area of a field-emitter tip is quite similar to the surface of commercial catalyst particles.

3.2 Metallurgy and Materials Science

Although the field-ion microscope and atom-probe are considered surface-sensitive probes, the ability to field evaporate into the bulk of the sample has led to numerous investigations of problems in physical metallurgy. A few examples of applications in this area are mentioned in the following paragraphs. For more extensive discussions of these applications, the reader is referred to a review by Brenner (1978) and books by Hren and Ranganathan (1968), Müller and Tsong (1969), Bowkett and Smith (1970), Miller and Smith (1980), and Tsong (1990).

One obvious metallurgical application of the field-ion microscope is the investigation of structural defects in pure metals. Point defects (vacancies and interstitials), which are beyond the resolution of conventional electron microscopes, are easily detected in the field-ion microscope. As early as 1959, the field-ion microscope was used to examine the thermal vacancy concentration in an annealed Pt wire (Müller, 1959). Line and planar defects (dislocations, stacking faults, twin boundaries, etc.) have also been investigated extensively with the field-ion microscope (Bowkett and Smith, 1970). In addition, the field-ion microscope has made significant contributions to our understanding of interface structures in metals. A particularly noteworthy accomplishment in this area was the clarification of the structure of high-angle grain boundaries (Müller, 1963; Brandon *et al.*, 1964).

A considerable amount of field-ion microscope research has been carried out related to the crystal and defect structure of ordered alloys. The early work in this field was directed towards image interpretation. In certain alloy materials it was discovered that only one of the constituent elements produced image spots in the field-ion microscope. Although it caused some confusion at first, this phenomenon proved to be useful to examine defect structures. Later, field-ion microscope studies provided valuable information concerning disorder-order transitions in platinum-cobalt (Tsong and Müller, 1966), nickel-molybdenum (Yamamoto *et al.*, 1974) and nickel-iron (Taunt and Ralph, 1975) alloys.

The field-ion microscope and atom-probe have made significant contributions to our understanding of decomposition processes in metal alloys. Precipitate formation in various Fe-, Al-, and Ni-based alloys has been a fruitful area of investigation. The techniques are particularly suited to the study of precipitates smaller than ~10 nm. In some alloy systems small precipitates image very clearly in the field-ion microscope and their number and size distributions are easily obtained. In other systems this information must be inferred from statistical analysis of atom-probe data. Composition modulations in spinodally decomposed alloys have also been investigated in atomic detail with the field-ion microscope and atom-probe. Extensive references to the original work in precipitation and other alloy decomposition processes can be found in the reviews cited above.

Another important application of field-ion microscopy has been the investigation of radiation damage induced by the bombardment of various forms of energetic particles. For example, studies by Seidman and co-workers (Seidman, 1978) have provided detailed information concerning the range profile of the impinging particles, the extent and structure of depleted zones, the type and size distributions of defects, and the removal of these defects by thermal processing.

In recent years materials science applications of the field-ion microscope and atom-probe have gone beyond studies of metals and alloys and now include semiconductor materials. Detailed investigations of metal-silicide interfaces were first reported by Nishikawa *et al.* (1983), and many similar studies of silicon-metal and compound semiconductor-metal interfaces have followed. The atom-probe is shown to be ideally suited for the investigation of compositional variations across the interface with single-atom-layer depth resolution. Studies of semiconductors are not confined to metal-semiconductor interfaces. For example, three-dimensional composition maps, compiled from analysis with the position-sensitive atom-probe, have been obtained from GaInGa/InP quantum-well heterostructures (Liddle *et al.*, 1989).

3.3 Novel Materials and Special Applications

The field-ion microscope and atom-probe have been used in the analysis of some very interesting and "novel" materials. For example, the atomic structure of the icosohedral alloy Al–12% Mn was investigated by Melmed and Klein (1986). The results indicate that there is long-range icosohedral order and that the fivefold symmetry is not due to microtwinning. Attempts to examine amorphous alloys (metallic glasses) have also been reported (Jacobaeus *et al.*, 1980; Piller and Haasen, 1982). As in the case of ordered alloys, it is found that some of the component atoms do not produce image spots, which makes quantitative structural determinations difficult. More recently, field-ion microscopy has been used to investigate the structure of high-T_c superconducting ceramic oxides (Kellogg and Brenner, 1987; Melmed *et al.*, 1988). Interesting patterns consisting of parallel rows of atoms are observed in the field-ion images, and atom-probe analysis yields a composition fairly consistent with the known stoichiometry of the material. Field-ion–related techniques have been developed to examine the morphology of biological molecules deposited on a metal surface (Panitz, 1982). Quasi three-dimensional images of several biological molecules (e.g., ferritin) deposited on a tungsten surface have been obtained with benzene as the imaging gas. The field-ion microscope and atom-probe have also been involved in the development of field emission, field-ion, and liquid-metal ion sources. Although these developments have considerable technological interest, a discussion of field-electron and ion sources is beyond the scope of this article.

4. SUMMARY

Over the past 40 years, the field-ion microscope has grown from a curiosity instrument capable of resolving the surface atoms of a few refractory metals under very special conditions to a basic research tool with widespread applications in the areas of surface and materials science. This growth is due, in part, to numerous advancements in the instrumentation associated with the field-ion microscope. Certainly, the ability to conduct chemical analysis with the atom-probe and its many variations has contributed to the widespread materials applications of the technique. In recent years, the development of new techniques with atomic resolution such as the high-resolution electron microscope and scanning tunneling microscope have rekindled an interest in viewing matter at the atomic level. This interest combined with an ever-increasing drive to fabricate new materials with unique and tailorable properties by manipulating matter at the atomic level should provide a long-range incentive for continued fundamental and applied research with the field-ion microscope and related instruments.

GLOSSARY

Atom-Probe: A field-ion microscope combined with a time-of-flight mass spectrometer, used to determine the chemical identity of ions generated at the surface of an emitter tip.

Channel Plate: A device placed in front of the fluorescent screen in a field-ion microscope used to convert the relatively weak ion image into an intense electron image by a process of electron multiplication.

Critical Distance of Field Ionization: The shortest distance from a surface for which field ionization may occur. Within this distance, the uppermost ionization level of the gas atom or molecule falls below the Fermi energy of the surface and tunneling is forbidden.

Energy-Compensated Atom-Probe: A variation of the time-of-flight atom-probe in which a electrostatic lens is used to improve mass resolution by compensating for the distribution in the initial kinetic energies of field desorbed ions.

Field Adsorption: The binding of gas atoms or molecules (usually inert-gas atoms) to a solid surface induced by the presence of a high electric field.

Field Desorption: The removal of atoms or molecules from a surface as ions by the application of a high electric field.

Field-Electron Microscope: An instrument in which a negative voltage is applied to an emitter tip to eject electrons and form an image on a fluorescent screen. Unable to achieve atomic resolution, the field-electron microscope was the predecessor to the field-ion microscope.

Field Evaporation: A special case of field desorption in which atoms of the emitter material itself are removed by a high electric field.

Field-Ion Microscope: An instrument in which the field-ionization process is used to obtain atomic-resolution images of solid surfaces.

Field Ionization: The ionization of gas atoms or molecules in a high electric field by an electron tunneling process.

Imaging Gas: A gas (usually inert) admitted to a field-ion microscope to produce the image. The gas atoms are ionized above the tip surface and accelerated to an imaging detector.

Imaging Atom-Probe: A variation of the time-of-flight atom-probe in which a large-area, imaging, single-ion detector is used to obtain elemental maps of preselected atomic or molecular species.

Magnetic-Sector Atom-Probe: An atom-probe that uses a magnetic-sector mass spectrometer to determine the identity of field desorbed ions or molecules.

Position-Sensitive Atom-Probe: A variation of the time-of-flight atom-probe in which the identity and spatial location of each atom removed is recorded. Full three-dimensional reconstructions of the removed material are obtained with this instrument.

Pulsed-Laser Atom-Probe: A variation of the time-of-flight atom-probe in which ions are generated by the combination of a dc electric field and a short-duration laser pulse. It is used to examine low-conductivity materials and to investigate field-desorption phenomena as a function of electric field.

Thermal Accommodation: The loss of kinetic energy of impinging gas-phase atoms by successive hops across a cryogenically cooled sample surface. The spatial resolution of the field-ion microscope is greatly improved by this process.

Tip: A sharply pointed wire sample or whisker whose apex surface is examined in the field-ion microscope and atom-probe.

List of Works Cited

Bassett, D. W. (1983), in: V. T. Binh (Ed.), *Surface Mobilities on Solid Materials,* New York: Plenum, pp. 63–108.

Barofski, D. F., Müller, E. W. (1968), *Surf. Sci.* **10**, 177–196.

Bassett, D. W., Webber, P. R. (1978), *Surf. Sci.* **70**, 520–531.

Binnig, G., Rohrer, H., Gerber, Ch., Weibel, E. (1983), *Phys. Rev. Lett.* **50**, 120–123.

Block, J. H., Kruse, N., Abend, G. (1986), in: A. D. Romig, and W. F. Chambers (Eds.), *Microbeam—1986*, San Francisco: San Francisco Press, pp. 389–398.

Bowkett, K. M., Smith, D. A. (1970), *Field Ion Microscopy,* New York: American Elsevier.

Brandon, D. G., Ralph, B., Ranganathan, S., Wald, M. (1964), *Acta Met.* **12**, 813–821.

Brenner, S. S. (1978), *Surf. Sci.* **70**, 427–451.

Cerezo, A., Godfrey, T. J., Smith, G. D. W. (1988), *Rev. Sci. Instrum.* **59**, 862–866.

Chen, C., Tsong, T. T. (1990), *Phys. Rev. Lett.* **64**, 3147–3150.

Crewe, A. V., Wall, J., Langmore, J. (1970), *Science* **168**, 1338–1340.

Drachsel, W., Nishigaki, S., Block, J. H. (1980), *Int. J. Mass Spectrom. Ion Phys.* **32**, 333–343.

Ehrlich, G., Hudda, F. (1966), *J. Chem. Phys.* **44**, 1039–1049.

Ehrlich, G., Stolt, K. (1980), *Annu. Rev. Phys. Chem.* **31**, 603–637.

Fink, H.-W., Ehrlich, G. (1984), *Phys. Rev. Lett.* **52**, 1532–1534.

Gomer, R. (1959), *J. Chem. Phys.* **31**, 341–345.

Gomer, R. (1961), *Field Emission and Field Ionization*, Cambridge, MA: Harvard University Press, Chap. 3.

Hren, J. J., Ranganathan, S. (1968), *Field Ion Microscopy*, New York: Plenum, pp. 1–244.

Inghram, M. G., Gomer, R. (1954), *J. Chem. Phys.* **22**, 1279–1280.

Jacobaeus, P., Madsen, J. U., Kragh, F., Cottrill, R. M. J. (1980), *Phil. Mag.* **B41**, 11–20.

Kellogg, G. L. (1985a), *Phys. Rev. Lett.* **55**, 2168–2171.

Kellogg, G. L. (1985b), *J. Catal.* **92**, 167–172.

Kellogg, G. L. (1987a), *J. Phys. E: Sci. Instrum.* **20**, 125–136.

Kellogg, G. L. (1987b), *Appl. Phys. Lett.* **51**, 100–102.

Kellogg, G. L. (1987c), *Surf. Sci.* **187**, 153–164.

Kellogg, G. L., Brenner, S. S. (1987), *Appl. Phys. Lett.* **51**, 1851–1853.

Kellogg, G. L., Feibelman, P. J. (1990), *Phys. Rev. Lett.* **64**, 3143–3146.

Kellogg, G. L., Tsong, T. T. (1980), *J. Appl. Phys.* **51**, 1184–1193.

Kellogg, G. L., Tsong, T. T. (1981), *Surf. Sci.* **110**, L599–L605.

Kellogg, G. L., Tsong, T. T., Cowan, P. (1978), *Surf. Sci.* **70**, 485–518.

Kreuzer, H. J., Nath. K. (1987), *Surf. Sci.* **183**, 591–608.

Liddle, J. A., Cerezo, A., Grovenor, C. R. M. (1989), *J. Phys. (Paris)* **50** (Coll. C8), 437–442.

McMullen, E. R., Perdew, J. R. (1987), *Phys. Rev. B* **36**, 2598–2606.

Melmed, A. J., Klein, R. (1986), *Phys. Rev. Lett.* **56**, 1478–1481.

Melmed, A. J., Shull, R. D., Chiang, C. K., Fowler, H. A. (1988), *Science* **239**, 176–178.

Miller, M. K., Smith, G. D. W. (1989), *Atom-Probe Microanalysis, Principles and Applications to Materials Problems*, Pittsburgh: Materials Research Society.

Müller, E. W. (1936), *Phys. Z.* **37**, 838–845.

Müller, E. W., (1951), *Z. Phys.* **131**, 136–142.

Müller, E. W., (1956a), *J. Appl. Phys.* **27**, 474–476.

Müller, E. W. (1956b), *Phys. Rev.* **102**, 618–624.

Müller, E. W. (1959), *Z. Phys.* **156**, 399–410.

Müller, E. W. (1963), *J. Phys. Soc. Jpn.* **18** (Suppl. 2), 1–17.

Müller, E. W., Krishnaswami, V. (1974), *Rev. Sci. Instrum.* **45**, 1053–1059.

Müller, E. W., Sakurai, T. (1974), *J. Vac. Sci. Technol.* **11**, 878–882.

Müller, E. W., Tsong, T. T. (1969), *Field Ion Microscopy, Principles and Applications*, New York: American Elsevier.

Müller, E. W., Panitz, J. A., McLane, S. B. (1968), *Rev. Sci. Instrum.* 39, 83–86.

Nath, K., Kreuzer, H. J., Anderson, A. B. (1986), *Surf. Sci.* **176**, 261–283.

Nishikawa, O., Tsunashima, Y., Nomura, E., Wada, M., Horie, S., Shibata, M., Yoshimura, T., Uemori, R. (1983), *Surf. Sci.* **126**, 529–533.

Ng, Y. S., Tsong, T. T., McLane, S. B. (1979), *Phys. Rev. Lett.* **42**, 588–591.

Panitz, J. A. (1973), *Rev. Sci. Instrum.* **44**, 1034–1038.

Panitz, J. A. (1982), *Ultramicroscopy* **7**, 241–247.

Piller, J., Haasen, P. (1982), *Acta Met.* **30**, 1–8.

Poschenrieder, W. P. (1971), *Int. J. Mass Spectrom. Ion Phys.* **6**, 413–426.

Seidman, D. N. (1978), *Surf. Sci.* **70**, 532–565.

Schwoebel, P. R., Kellogg, G. L. (1988), *Phys. Rev. Lett.***61**, 578–580.

Schwoebel, P. R., Foiles, S. M., Bisson, C. L., Kellogg, G. L. (1989), *Phys. Rev. B* **40**, 10639–10642.

Taunt, R. J., Ralph, B. (1975), *Surf. Sci.* **47**, 569–582.

Tsong, T. T. (1988), *Surf. Sci. Rep.* **8**, 127–209.

Tsong, T. T. (1990), *Atom-Probe Field Ion Microscopy*, Cambridge, UK: Cambridge University Press.

Tsong, T. T., Müller, E. W. (1966), *Appl. Phys. Lett.* **9**, 7–10.

Tsong, T. T., Müller, E. W. (1970), *Phys. Rev. Lett.* **25**, 911–913.

Tsong, T. T., McLane, S. B., Kinkus, T. J. (1982), *Rev. Sci. Instrum.* **53**, 1442–1448.

Tsong, T. T., Ng, Y. S., Krishnaswami, S. (1978), *Appl. Phys. Lett.* **32**, 778–780.

Wrigley, J. D., Ehrlich, G. (1980), *Phys. Rev. Lett.* **44**, 661–663.

Yamamoto, M., Futamoto, S., Nenno, S., Nakamura, S. (1974), *Jpn. J. Appl. Phys.* **13**, 1461–1462.

Further Reading

Bowkett, K. M., Smith, D. A. (1970), *Field Ion Microscopy*, New York: American Elsevier.

Gomer, R. (1961), *Field Emission and Field Ionization*, Cambridge, MA: Harvard University Press.

Hren, J. J., Ranganathan, S. (1968), *Field Ion Microscopy*, New York: Plenum.

Kellogg, G. L. (1987), "Pulsed-Laser Atom-Probe Mass Spectroscopy", *J. Phys. E: Sci. Instrum.* **20**, 125–136.

Miller, M. K., Smith, G. D. W. (1989), *Atom-Probe Microanalysis, Principles and Applications to Materials Problems,* Pittsburgh: Materials Research Society.

Müller, E. W., Tsong, T. T. (1969), *Field Ion Microscopy, Principles and Applications,* New York: American Elsevier.

Panitz, J. A. (1985), "High-Field Techniques," *Methods of Experimental Physics* 22, Chap. 7, pp. 349–423.

Tsong, T. T. (1990), *Atom-Probe Field Ion Microscopy,* Cambridge , UK: Cambridge University Press.

FIELD THEORIES, UNIFIED

See UNIFIED FIELD THEORIES

FILMS, MAGNETIC

See MAGNETIC FILMS

FISSIONS, NUCLEAR

See NUCLEAR ENERGY, FISSION

FLOWS, TURBULENT

See TURBULENT FLOWS

FLUID DYNAMICS, EQUILIBRIUM

Ronald L. Panton, *Department of Mechanical Engineering, University of Texas, Austin, Texas, U.S.A.*

Introduction and Scope 409
1. **Simple Compressible Substance** 410
2. **Kinematics** 410
3. **Local Conservation Laws** 411
3.1 Continuity, Momentum, and Energy 411
3.2 Angular Momentum, Kinetic Energy, and Thermal Energy 411
3.3 Temperature and Entropy 412
4. **Global Conservation Laws** 412
4.1 Continuity, Momentum, and Energy 412
5. **Interfaces** 413
5.1 No-Slip Condition 413
5.2 Jump Conditions 413
5.3 Surface Tension and Continuum Properties 413
6. **Constitutive Equations** 413
6.1 Thermal Conductivity 414
6.2 Viscosity 414
7. **Navier-Stokes Equations** 414
8. **Scales and Dimensional Considerations** 414
9. **Compressible Flows** 414
10. **Incompressible Flows** 415
10.1 Adiabatic Walls 415
10.2 Isothermal Walls 415
10.3 Characteristics of Incompressible Flows 416
10.4 Pressure and Vorticity 416
10.5 Mechanical Energy (Engineering Bernoulli Equation) 417
11. **Simple Laminar Flows** 417
11.1 Parallel Flows 417
11.2 Viscous Diffusion 418
11.3 Stagnation Point 418
12. **Wakes, Jets, and Drag** 419
13. **Swirling Flows** 422
13.1 Various Vortex Profiles 422
13.2 von Kármán Viscous Pump 423
13.3 Vortex Breakdown 424
14. **High–Reynolds-Number Flows** 425
14.1 Inviscid Flows 426
14.1.1 Internal Flows 427
14.1.2 External Flows 428
14.2 Boundary Layers 430
14.2.1 Physical Characteristics 431
14.2.2 Three-Dimensional Effects 432
15. **Low–Reynolds-Number Flows** 433
15.1 Internal Flows—Lubrication Theory 434
15.2 External Flows 434
16. **Stability and Turbulence** 435
Glossary 435
Works Cited 435
Further Reading 435

INTRODUCTION AND SCOPE

Fluid dynamics is a continuum science that deals with a sufficient number of molecules so that meaningful averages of density, mass-averaged velocity, internal energy, pressure, temperature, and so forth may be formed. A fluid dynamic event must have length and time scales large enough to conceptually form these averages.

It is a tacit assumption that if we subtract the average velocity from each molecular velocity, the remaining random components will still have a Maxwellian distribution. This

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

means that equilibrium thermodynamics prevails locally. In cases where rarefied gases are involved this assumption is sometimes modified.

1. SIMPLE COMPRESSIBLE SUBSTANCE

A simple compressible substance is a fluid that receives reversible work only from pressure forces. This article will deal only with fluids of this type. The thermodynamics of a simple compressible substance is completely described by two equations of state: a pressure-density-temperature relation and a specific-heat function of density and temperature (see THERMODYNAMICS, EQUILIBRIUM). Thus fluids with anisotropic structure or electrical and magnetic properties are excluded. Occasionally liquids are characterized as incompressible substances. Then density and pressure are no longer thermodynamic variables, and one equation of state for the specific heat as a function of temperature describes the liquid. Transport laws for the molecular migration of chemical species, momentum, or internal energy contain coefficients that depend on the thermodynamic state. These coefficients, namely the diffusion coefficients, viscosity, and thermal conductivity, are in this sense thermodynamic properties of the fluid.

If a shear stress is applied to a fluid it continues to deform as long as the stress is applied. On the other hand, application of a shear stress to a solid produces only a small strain that generates internal forces which resist the force and produce an equilibrium. Removal of the force in a perfectly elastic solid recovers the energy used to deform the solid. On the other hand, the shearing work in a fluid cannot be recovered and becomes random molecular motion in the form of internal energy. In fluid mechanics, solids are frequently idealized as rigid objects.

2. KINEMATICS

The *Eulerian* viewpoint is to analyze fluid flows from a fixed coordinate system where properties are functions of position $\mathbf{x}$ and time t. The *substantial* or *material derivative*, first introduced by Stokes, is defined as

$$\frac{D(\ )}{Dt}=\frac{\partial(\ )}{\partial t}+\mathbf{v}\cdot\nabla(\ ). \tag{1}$$

As applied to an Eulerian function this derivative has the interpretation of the rate of change as one moves along with a fluid particle. If $\mathbf{r}$ is the position vector of a material particle, $D\mathbf{r}/Dt=\mathbf{v}$ is the velocity and $D\mathbf{v}/Dt=\mathbf{a}$ is the acceleration. Consider two neighboring material particles P and P', separated by an increment $\delta\mathbf{r}$, and imagine following the motion of P' relative to P. Then the material derivative of $\delta\mathbf{r}$ is the difference in their velocities,

$$\frac{D(\delta\mathbf{r})}{Dt}=d\mathbf{v}=\delta\mathbf{r}\cdot\nabla\mathbf{v}. \tag{2}$$

Using definitions below and tensor identities, Eq. (2) can be changed into

$$\frac{D(\delta\mathbf{r})}{Dt}=\delta\mathbf{r}\cdot\mathbf{S}+\tfrac{1}{2}\boldsymbol{\omega}\times\delta\mathbf{r}. \tag{3}$$

Here, the *strain rate* tensor is defined as

$$\mathbf{S}=\tfrac{1}{2}\nabla\mathbf{v}+\tfrac{1}{2}(\nabla\mathbf{v})^{\mathrm{T}}, \tag{4}$$

and the *vorticity* vector is defined as

$$\boldsymbol{\omega}=\nabla\times\mathbf{v}. \tag{5}$$

The vorticity is equal to twice the solid-body rotation of P' about P while all the deformation motions (both angular and extensional) are produced by $\mathbf{S}$. The trace of $\mathbf{S}$, that is $\nabla\cdot\mathrm{v}$, is the *rate of expansion* of a unit volume. Flow regions where $\boldsymbol{\omega}=\mathbf{0}$ are called *irrotational.*

At any given instant in time *streamlines* are defined as lines that are everywhere parallel to the velocity vectors. Similarly *vortex lines* are at any instant everywhere parallel to the vorticity vectors (the term "line vortex" has a distinctly different use). *Particle path* is the term for the trajectory of a fluid particle of fixed identity. In *steady flow,* a field that is independent of time, the streamlines and the particle paths coincide; otherwise they are distinct.

3. LOCAL CONSERVATION LAWS

3.1 Continuity, Momentum, and Energy

Any flow field is governed by three independent conservation laws. The conservation of mass, usually called the *continuity equation,* governs the density ρ:

$$\frac{1}{\rho}\frac{D\rho}{Dt}=-\nabla\cdot v. \tag{6}$$

The fractional rate of change of density is equal to the rate of expansion.

The *momentum equation* attributes acceleration to pressure forces, viscous forces, and gravity:

$$\rho\frac{D\mathbf{v}}{Dt}=-\nabla p+\nabla\cdot\tau+\rho\mathbf{g}. \tag{7}$$

Here p is the pressure, the viscous stress tensor is τ, and the gravity vector is $\mathbf{g}$.

The total energy equation governs the sum of internal energy e and macroscopic kinetic energy $\frac{1}{2}v^2$:

$$\rho\frac{D}{Dt}(e+\tfrac{1}{2}v^2)$$
$$=-\nabla\cdot q-\nabla\cdot(p\mathbf{v})+\nabla\cdot(\tau\cdot v)+\rho\mathbf{v}\cdot g. \tag{8}$$

Microscopic energy transfer mechanisms that have no discernible force are called heat. The symbol $\mathbf{q}$ is the *heat flux.* Conduction, radiation, and diffusion (in a mixture of chemical species) are in this category. Macroscopic energy transfers are the total work rate (force times velocity) of surface forces and gravity forces. An alternative perspective is to introduce a gravitational potential energy (*gh,* where h is the vertical distance above a reference level and g the magnitude of $\mathbf{g}$) and delete the work of the gravity force $\rho\mathbf{v}\cdot\mathbf{g}$.

3.2 Angular Momentum, Kinetic Energy, and Thermal Energy

If $\mathbf{r}$ is the particle position from an arbitrary origin, the macroscopic *angular momentum* about that origin is $\mathbf{r}\times\mathbf{v}$. An equation governing this quantity can be derived by taking the vector product of $\mathbf{r}$ and the linear momentum equation, Eq. (7). This relation is called the moment-of-momentum equation.

When angular momentum is transported and stored at the molecular level (a preferred orientation of spinning microscopic particles), one must also propose an angular momentum conservation law and provide an angular momentum transport tensor. For such fluids Dahler and Scriven (1961) summarize the results, which include a stress tensor that is not symmetric. This approach is needed only for exotic fluids such as ferromagnetic fluids (iron filings suspended in oil).

Homogeneous isotropic fluids have no net molecular angular momentum; that is, the molecular angular momentum is randomly oriented. In this case comparing the angular momentum conservation law and the moment-of-momentum equation shows that the stress tensor is symmetric and the equations become identical, namely

$$\rho\frac{D(\mathbf{r}\times\mathbf{v})}{Dt}=-\mathbf{r}\times\nabla p+\nabla\cdot(\mathbf{r}\times\tau)+\rho\mathbf{r}\times\mathbf{g}. \tag{9}$$

Thus in fluid dynamics of simple fluids the angular momentum conservation law does not supply an equation that is independent of the linear momentum law.

An equation governing macroscopic kinetic energy is found from the inner produce of $\mathbf{v}$ and the momentum equation. It is

$$\rho\frac{D\frac{1}{2}v^2}{Dt}=-\mathbf{v}\cdot\nabla p+\mathbf{v}\cdot(\nabla\cdot\tau)+\rho\mathbf{v}\cdot g. \tag{10}$$

Here, only a portion of the surface work, the velocity times gradients of surface forces (unbalanced forces), accelerates the fluid to increase the kinetic energy. All of the gravity work appears as kinetic energy while heat transfer never directly produces kinetic energy. Subtracting Eq. (10) from the total energy equation, Eq. (8), yields an equation governing the internal energy:

$$\rho\frac{De}{Dt}=-p\nabla\cdot v+\tau:\nabla\mathbf{v}-\nabla\cdot q. \tag{11}$$

The remaining portion of the surface work, "balanced" forces times deformations, increases the random molecular motions and appears as a source of internal energy. *Compression or expansion work,* $p\nabla\cdot$v, is reversible and *viscous dissipation work,* $\tau:\nabla\mathbf{v}$, is irreversible (because τ is symmetric $\tau:\nabla\mathbf{v}=\tau:\mathbf{S}$). Any net heat flux produces internal energy.

3.3 Temperature and Entropy

A very useful form of the thermal energy equation is derived by employing various thermodynamic relations to arrive at a relation for the temperature T,

$$\rho c_p \frac{DT}{Dt} = \beta T \frac{Dp}{Dt} + \boldsymbol{\tau}:\nabla \mathbf{v} - \nabla \cdot q. \qquad (12)$$

It should be noted that the thermodynamic functions for *specific heat at constant pressure, $c_p(p,T)$*, and *thermal expansion coefficient, $\beta(p,T)$*, are unrestricted.

The second law of thermodynamics for the *entropy s* takes the form

$$\rho \frac{Ds}{Dt} = \frac{1}{T} \boldsymbol{\tau}:\nabla \mathbf{v} - \nabla \cdot \frac{\mathbf{q}}{T} - \frac{1}{T^2} \mathbf{q} \cdot \nabla T. \qquad (13)$$

Any viscosity law relating $\boldsymbol{\tau}$ and $\nabla \mathbf{v}$ must have a form that always gives a positive contribution to increase the entropy. The heat-transport terms have been written in a form that displays a reversible effect and an irreversible effect. If the reversible effect is integrated over a volume of interest and Gauss's divergence theorem applied, it becomes a boundary effect of heat flux divided by temperature. Without local viscous dissipation and heat transport, a flow becomes isentropic.

4. GLOBAL CONSERVATION LAWS

Any local law or relation for an arbitrary quantity f that contains a rate of change $\partial f/\partial t$ may be expressed in a global (or integral) form. Consider a finite region of volume V and surface S with outward unit normal $\mathbf{n}$ and an arbitrarily specified surface velocity $\mathbf{w}$. For continuous f, Leibnitz's theorem of calculus gives a relation for the rate of change of the total integrated value of f,

$$\frac{d}{dt} \int f \, dV = \int \frac{\partial f}{\partial t} dV + \int \mathbf{n} \cdot w \, f \, dS. \qquad (14)$$

In order to derive a global law one substitutes the local law for $\partial f/\partial t$; then Gauss's divergence theorem is used to convert as many volume integrals as possible to surface integrals.

4.1 Continuity, Momentum, and Energy

Taking f as the density and employing the continuity equation [Eq. (6)] yields

$$\frac{d}{dt} \int \rho \, dV = - \int \mathbf{n} \cdot (\mathbf{v} - \mathbf{w}) \rho \, dS. \qquad (15)$$

The mass within a region increases because of the mass flux from the relative velocity $\mathbf{n} \cdot (\mathbf{v} - \mathbf{w})$ normal to the moving surface.

The global momentum equation is found by taking f as $\rho \mathbf{v}$ and using Eq. (7):

$$\frac{d}{dt} \int \rho \mathbf{v} \, dV = - \int [\mathbf{n} \cdot (\mathbf{v} - \mathbf{w}) \rho \mathbf{v} + \mathbf{n} p - \mathbf{n} \cdot \boldsymbol{\tau}] dS + \int \rho \mathbf{g} \, dV. \qquad (16)$$

The body force of gravity may be converted to a surface integral, $\rho \mathbf{g} = -\nabla (g \rho h)$, if the flow is incompressible.

The total energy equation, with $E = e + \frac{1}{2} v^2 + gh$, for a region in arbitrary motion is similarly found employing Eq. (8):

$$\frac{d}{dt} \int \rho E \, dV = - \int [\mathbf{n} \cdot (\mathbf{v} - \mathbf{w}) \rho E - \mathbf{n} \cdot \boldsymbol{\tau} \cdot v + \mathbf{n} \cdot v p + \mathbf{n} \cdot \mathbf{q}] dS. \qquad (17)$$

In practical applications, the work of the pressure forces is sometimes represented by two effects as follows:

$$\int \mathbf{n} \cdot v p \, dS = \int [\mathbf{n} \cdot (\mathbf{v} - \mathbf{w}) \rho (p/\rho) + \mathbf{n} \cdot w p] dS. \qquad (18)$$

The first term is called *flow work* and the second the *moving boundary work*. The advantage of this form is that the flow work can be combined with convection of internal energy by introducing the enthalpy $h = e + p/\rho$.

It is also possible to write a global mechanical energy equation. It is

$$\frac{d}{dt} \int \rho \tfrac{1}{2} v^2 dV = - \int [\mathbf{n} \cdot (\mathbf{v} - \mathbf{w}) \rho \tfrac{1}{2} v^2 - \mathbf{n} \cdot \boldsymbol{\tau} \cdot v + \mathbf{n} \cdot v p] dS + \int [\mathbf{n} \cdot g \rho - \boldsymbol{\tau}:\nabla \mathbf{v} + p \nabla \cdot v] dV. \qquad (19)$$

Several practical results derive from this equation [for example, see Eqs. (46) and (89)].

5. INTERFACES

Surfaces where the density is discontinuous characterize not only fluid-solid interfaces, liquid-liquid interfaces, and liquid-gas interfaces, but also shock and combustion waves. In situations with several chemical species it may be necessary to envision that an interface contains absorbed mass; however, pure substances can usually be dealt with assuming that no mass, momentum, or energy is associated with the surface.

5.1 No-Slip Condition

The translational mobility of fluid molecules allows them to interact with neighboring fluid elements. This is the microscopic origin of viscous action, and because of this molecular migration we do not allow discontinuities in tangential velocity at an interface within a fluid. Similarly, at an interface between a fluid and a solid wall, molecules enter the interface and interact with molecules from the wall. After "collision" they return to the bulk fluid with a tangential velocity nearly that of the wall. A kinetic theory model for a gas and a solid wall shows that the tangential gas velocity at the interface differs from the wall velocity by a slip velocity u_s given by

$$\frac{u_s}{U}=\frac{c\lambda}{L}\frac{d(u/U)}{d(n/L)}. \tag{20}$$

Here $u(n)$ is the velocity profile, n is the normal direction, U and L are the characteristic scales of the continuum motion that normalize $u(n)$, c is a constant of order one, and λ is the mean free path. The approximation $u_s=0$ is thus very accurate when the mean free path is small compared to the continuum length, that is, λ/L is small. Indeed, the first approximation in the field of *rarefied gas dynamics* is to use continuum equations with the slip boundary condition above.

The extension of this argument to a fluid-fluid interface produces the result that the tangential component of the velocity is continuous across the interface.

5.2 Jump Conditions

By postulating that mass, momentum, and energy cannot be stored within an interface, it is possible to derive jump conditions that relate variables on either side of a plane interface. Brackets indicate that the quantity is evaluated on either side, and the interface is regarded as stationary. From the mass equation we get

$$[\rho\mathbf{n}\cdot\nu]_{\mathrm{I}}-[\rho\mathbf{n}\cdot\nu]_{\mathrm{II}}=0. \tag{21}$$

Without mass flow across the interface the normal velocity at the interface is zero. The normal momentum relation gives

$$[\rho(\mathbf{n}\cdot\nu)^2-p+\mathbf{n}\cdot(\mathbf{n}\cdot\tau)]_{\mathrm{I}}-[\rho(\mathbf{n}\cdot\nu)^2-p+\mathbf{n}\cdot(\mathbf{n}\cdot\tau)]_{\mathrm{II}}=0. \tag{22}$$

Tangential momentum and the assumption of continuity of the tangential velocity component (no-slip) show that the tangential viscous stress is continuous,

$$[\mathbf{n}\times(\mathbf{n}\cdot\tau)]_{\mathrm{I}}-[\mathbf{n}\times(\mathbf{n}\cdot\tau)]_{\mathrm{II}}=0. \tag{23}$$

Energy conservation results in

$$[\rho\mathbf{n}\cdot\nu(h+\tfrac{1}{2}\nu^2)-\mathbf{n}\cdot q]_{\mathrm{I}}-[\rho\mathbf{n}\cdot\nu(h+\tfrac{1}{2}\nu^2)-\mathbf{n}\cdot q]_{\mathrm{II}}=0. \tag{24}$$

Continuity of the heat flux across the wall is required if there is no mass transfer.

5.3 Surface Tension and Continuum Properties

If a surface is highly curved the simple model above is invalid, and it is necessary to propose a *surface tension* (surface energy). The surface tension coefficient σ relates the pressure change across the surface to the principal radii of curvature, R_1 and R_2:

$$p_{\mathrm{I}}-p_{\mathrm{II}}=\sigma(1/R_1+1/R_2). \tag{25}$$

Although σ is conceptually like a pressure in a two-dimensional world, it is treated as a thermodynamic property of the interface.

6. CONSTITUTIVE EQUATIONS

The microscopic mechanisms to create fluxes of mass, momentum, or energy from a nonuniformity of the appropriate flow property give rise to constitutive relations. Such relationships are not universally valid; however, many chemically simple substances obey linear relations.

6.1 Thermal Conductivity

Fourier's law for heat conduction is linear in the temperature gradient,

$$\mathbf{q} = -\kappa \nabla T. \tag{26}$$

The *thermal conductivity* κ is a thermodynamic property of the material.

6.2 Viscosity

Newton's viscosity law is linear in the rate of expansion and deformation,

$$\tau = \lambda \nabla \cdot v I + 2\mu \mathbf{S}. \tag{27}$$

The coefficients λ (*second viscosity*) and μ (*viscosity*) are thermodynamic properties. The thermodynamic concept of pressure and the mechanical concept of a normal stress are related by Stokes's assumption that the average normal stress is equal to the thermodynamic pressure. This implies that $\lambda = -\frac{2}{3}\mu$. In liquids the viscosity and conductivities decrease with increasing temperature while in gases they increase. Pressure does not have a large effect in either case.

7. NAVIER-STOKES EQUATIONS

This term *Navier-Stokes equations* is applied to the mass, momentum, and thermal energy equations written for a substance obeying Fourier's heat conduction law and Newton's viscosity relation. They are

$$\frac{1}{\rho}\frac{D\rho}{Dt} = -\nabla \cdot v, \tag{28}$$

$$\rho \frac{D\mathbf{v}}{Dt} = -\nabla p - \tfrac{2}{3}\nabla(\mu \nabla \cdot \mathbf{v}) + 2\nabla \cdot (\mu \mathbf{S}) + \rho \mathbf{g}, \tag{29}$$

$$\rho c_p \frac{DT}{Dt} = \beta T \frac{Dp}{Dt} + \tfrac{2}{3}\mu(\nabla \cdot v)^2 + 2\mu \mathbf{S}{:}\nabla \mathbf{v} + \nabla \cdot (\kappa \nabla T). \tag{30}$$

Another use of the name Navier-Stokes is for the simpler equations that govern an incompressible flow. They will be given below.

8. SCALES AND DIMENSIONAL CONSIDERATIONS

The simplest characterization of a flow occurs when variables are made nondimensional using characteristic scales derived from the specific flow situation of interest. Let the size of the region of interest be characterized by L, the flow speed by U_0, and the fluid reference state by density ρ_0 and temperature T_0 with speed of sound a_0. The proper temperature scale depends on the heat-transfer boundary conditions. In any event, several nondimensional parameters will appear in the equations. The Mach number, Reynolds number, and Froude number refer to dynamics:

$$M = U_0/a_0, \quad Re = \rho U_0 L/\mu_0, \quad Fr = U_0^2/gL. \tag{31}$$

Other nondimensional thermodynamic parameters express the ratio of specific heats ($\gamma = c_{p0}/c_{v0}$, where the specific heat at constant volume is c_v), the isothermal compressibility ($\alpha^* = \alpha_0 \rho_0 c_{p0} T_0$), the thermal expansion coefficient ($\beta^* = \beta_0 T_0$), and the Prandtl number ($Pr = \mu_0 c_{p0}/\kappa_0$). As a first approach to classifying flows it is appropriate to let the thermodynamic parameters be arbitrary and consider different extremes in the dynamic parameters.

9. COMPRESSIBLE FLOWS

Variations in the density must be accounted for in several flow situations. Consider a flow where the fluid undergoes a velocity change U_0 (say from U_0 to complete rest). If the fluid density is ρ_0, the pressure difference needed to cause this velocity change is approximately $\rho_0 U_0^2$. Recognizing that $\Delta p/\Delta\rho = a_0^2$ for a reversible process, we can estimate the percentage of density change as

$$\frac{\Delta\rho}{\rho_0} = \frac{\Delta\rho}{\Delta p}\frac{\Delta p}{\rho_0} = \frac{1}{a_0^2}\frac{\Delta p}{\rho_0} = \frac{1}{a_0^2}\frac{\rho_0 U_0^2}{\rho_0} = \frac{U_0^2}{a_0^2} = M^2. \tag{32}$$

Hence the Mach number is an indicator of compressibility effects from fluid accelerations. To make this point quantitative consider the isentropic (frictionless and without heat conduction) flow of a perfect gas. If the flow is brought to rest at a stagnation point, the pressure change is

$$p_0-p_\infty=\tfrac{1}{2}\rho U_\infty^2[1-\tfrac{1}{4}M^2+O(M^4)]. \qquad (33)$$

This equation shows a 2–3% correction to the incompressible Bernoulli equation for $M \approx 0.3$. The value of $M=0.3$ is often quoted as the limit for incompressible flow; however, the shape of the body can also have an influence. Although the equation above limits the ability to convert inertia to pressure, there is no limit to decreases in pressure by expansion, that is, the conversion of pressure to inertia. The next generation of high-lift wings will have strong expansions of the flow and compressible effects at lower values of M. High-speed compressible flows (see COMPRESSIBLE FLOWS) occur when velocity changes are brought about by strong pressure changes that are accompanied by density changes.

Another type of compressible flow occurs when there is significant energy entering the flow from a wall. The subsequent changes in thermodynamic state cause associated density variations. For example, flow in a natural gas pipeline is compressible even if the initial Mach number is 0.05. This is a "low-speed" compressible flow. "Natural" or "free" convection is another such situation. Here, the characteristic velocity is small and is related to the heat transfer from a hot or cold surface. Acoustics (see ACOUSTICS, LINEAR; ACOUSTICS, NONLINEAR) is another type of compressible flow caused by the rapid normal oscillation of a surface. It is a linear subject for small-amplitude motions.

10. INCOMPRESSIBLE FLOWS

The *incompressible flow* of a general compressible fluid can occur under two major situations. Both are thought of as low–Mach-number limits but with different heat-transfer boundary conditions.

10.1 Adiabatic Walls

When the walls confining the flow are essentially adiabatic, the only mechanism to produce thermal energy is viscous dissipation in regions where velocity gradients exist. Once generated, this energy can be convected by the flow and conducted across temperature gradients. The appropriate temperature scale to measure changes from the reference value is $\mu_0 V_0^2/\kappa_0$. These thermal effects are so weak at low Mach numbers that the flow is nearly isothermal and the density is constant. For $M^2 \to 0$ the appropriate Navier-Stokes equations, Eqs. (28)–(30), simplify to

$$0=\nabla\cdot v, \qquad (34)$$

$$\rho\frac{D\mathbf{v}}{Dt}=-\nabla p+\mu\nabla\cdot\nabla\mathbf{v}+\rho\mathbf{g}, \qquad (35)$$

$$\rho c_p\frac{DT}{Dt}=\beta T\frac{Dp}{Dt}+2\mu\mathbf{S}:\nabla\mathbf{v}+\kappa\nabla^2 T. \qquad (36)$$

Furthermore, as $M^2 \to 0$ the coefficients μ, k, β, and c_p are approximately constant. The continuity and momentum equations are now uncoupled from the energy equation, and the fluid flow can be solved without regard to thermal effects; hence fluid dynamics and heat transfer become separate subjects.

10.2 Isothermal Walls

When the walls containing the flow are purposefully held at a distinctly different temperature than the reference, the thermal energy distribution is dominated by conduction from the walls. Now the proper scale is T_1-T_0, where T_1 is the characteristic wall temperature. In the limit of low Mach number, the continuity and momentum equations, Eqs. (28) and (29), remain the same but the energy equation, Eq. (30), becomes

$$\rho c_p\frac{DT}{Dt}=\nabla\cdot(\kappa\nabla T). \qquad (37)$$

These equations are the "low-speed" compressible equations. The density, specific heat, and transport properties are not constant, and the thermal and mechanical events are coupled. Further simplifications depend upon assumptions about the fluid, such as an incompressible liquid, or flow situation, such as free convection.

Another special case is of major interest and it is an incompressible flow: the case of isothermal walls with a "small" temperature difference, that is $(T_1-T_0)/T_0 \to 0$, and simultaneously a low Mach number. The Navier-Stokes equations [Eqs. (28), (29), and (30)] for this case reduce to

$$0=\nabla\cdot v, \qquad (38)$$

$$\rho\frac{D\mathbf{v}}{Dt}=-\nabla p+\mu\nabla\cdot\nabla\mathbf{v}+\rho\mathbf{g}, \qquad (39)$$

$$\rho c_p \frac{DT}{Dt} = \kappa \nabla^2 T. \tag{40}$$

These equations are the same as those for the adiabatic case except that the temperature equation is simpler. Again the thermodynamic coefficients are constant and the continuity and momentum equations uncoupled from the energy equation. The fluid mechanics of this case and the adiabatic case are the same. The only difference is in the temperature field due to heat-transfer events.

10.3 Characteristics of Incompressible Flows

With appropriate boundary conditions, the continuity and momentum equations yield solutions for the velocity vector and the pressure. The pressure is a mechanical item, as it causes no appreciable change in the thermodynamic state. Indeed, thermodynamics is relevant only in determining values of the coefficients in Eqs. (39) and (40).

Consider the incompressible flow in a region of fixed shape with specified boundary values. In nondimensional variables the flow pattern depends on only two parameters, *Re* and *Fr*. The Reynolds number, $Re = \rho U_0 L/\mu_0$, indicates the inertia effects compared to viscous effects. The Froude number, $Fr = U_0^2/gL$, indicates the importance of inertia and gravity. It is only dynamically significant when the flow is a liquid with a free surface. Then the pressure change across the interface is a boundary condition in the problem.

For a single-phase flow one may introduce h, the distance above a horizontal reference level, as a potential for the gravity force, $\rho \mathbf{g} = \nabla(-\rho g h)$. Let a *modified pressure* be defined by

$$p' \equiv p + \rho g h - p_{\text{ref}}. \tag{41}$$

Now the pressure and gravity forces in the momentum equation, Eq. (35), can be combined as follows:

$$-\nabla p + \rho \mathbf{g} = -\nabla p - \nabla(\rho g h) = \nabla p'. \tag{42}$$

Thus the pressure in a flow without a free surface has a dynamically important part p' and a hydrostatic part with the universal solution $p_{\text{hydro}} = \rho g h - p_{\text{ref}}$. Because pressure is not in a boundary condition, the effect of gravity is to cause a hydrostatic pressure that has no dynamical importance. This also means that the total force on a submerged object consists of a dynamic part (lift and drag) and a buoyancy part. The fact that a fluid is a liquid or a gas is indicated by the value of ρ but nowhere else.

The typical way to express the pressure on a surface is as a *pressure coefficient* that includes the arbitrary reference,

$$C_p \equiv (p - p_{\text{ref}})/\tfrac{1}{2}\rho U_\infty^2. \tag{43}$$

The scale quantity $\frac{1}{2}\rho U_\infty^2$ is called the *dynamic pressure*, and it is the appropriate scale unless the Reynolds number is very low. Similarly the lift or drag force on a body is nondimensionalized by some typical cross-section area times the dynamic pressure $\frac{1}{2}\rho U_\infty^2$.

The equations governing an incompressible flow are invariant under an "unsteady" translational transformation of the coordinate axes: $\mathbf{X} = \mathbf{x} + \int \mathbf{V}(t)\,dt$. The invariance is not complete because the flow, viewed in the transformed coordinates, must be considered to have an additional time-dependent body (gravity) force $\rho d\mathbf{V}/dt$ or, in an alternative viewpoint, a pseudopressure $\pi = p + \rho \mathbf{x} \cdot d\mathbf{V}/dt$ where p is the true pressure in the inertial system.

10.4 Pressure and Vorticity

The pressure field associated with a given flow is governed by an elliptic equation

$$\nabla^2 p = -\rho \nabla \mathbf{v} : \nabla \mathbf{v}. \tag{44}$$

This equation is derived from the inner product of ∇ and Eq. (39). Viscous effects enter indirectly in determining the velocity field.

The vector product of ∇ and Eq. (39) produces an equation governing the vorticity. It is

$$\frac{D\boldsymbol{\omega}}{Dt} = \boldsymbol{\omega} \cdot \nabla \mathbf{v} + \nu \nabla \cdot \nabla \boldsymbol{\omega}. \tag{45}$$

Here the kinematic viscosity $\nu \equiv \mu/\rho$ is employed as a convenient notation. The first two terms in this relation have an analogy with Eq. (2) that describes the lengthening and turning of a material line element $\delta \mathbf{r}$. The analogy is with $\boldsymbol{\omega} \rightarrow \delta \mathbf{r}$. If viscous effects of the last term in Eq. (45) are negligible, then a vortex line, a line parallel to $\boldsymbol{\omega}$ at all points, may be considered as a moving with the fluid particles, and the magnitude of vorticity increases as the line is stretched. The viscous term is a diffusion effect with the same coefficient as the diffusion of momentum. Pressure does not appear because a pressure gradient will always act through the center of mass and cannot engender any solid-body rotation in a particle.

10.5 Mechanical Energy (Engineering Bernoulli Equation)

In many practical situations there is a single fluid stream that enters a device, goes through a process, and then exits. The device might be as simple as a pipe with bends and valves, or as complex as the turbine of a hydroelectric dam. If one considers a uniform flow at both entrance and exit and at least a periodic flow within the device, then the global mechanical energy equation for the device is very simple (for an incompressible flow). It is

$$\left[\frac{p}{\rho}+\tfrac{1}{2}v^2+gh\right]_1-\left[\frac{p}{\rho}+\tfrac{1}{2}v^2+gh\right]_2 =-\frac{dW/dt}{dm/dt}+\frac{\Phi}{dm/dt}. \tag{46}$$

In this equation dW/dt is the work rate of any shaft that crosses the region, dm/dt is the mass flow rate, and Φ is the viscous dissipation integrated over the interior of the region:

$$\Phi\equiv\int\tau\mathbf{:}\nabla\mathbf{v}\;dV=\frac{dm}{dt}\,gh_L. \tag{47}$$

An ideal frictionless flow has $\Phi=0$. The quotient of Φ by the product of the mass flow rate and the acceleration of gravity is called the *head loss* between stations 1 and 2. It has units of length. Sometimes g is included in the definition, that is, $gh_L\rightarrow h_L$ (units of energy per unit mass). The nondimensional form is obtained by dividing by the square of some characteristic velocity, U, and is called the *loss coefficient:*

$$K=gh_L/\tfrac{1}{2}U^2. \tag{48}$$

In pipe flow the loss increases in direct proportion to the length L divided by the diameter D, and this is accounted for by defining a *friction factor* f, so that

$$K_{\text{pipe}}=gh_L/\tfrac{1}{2}U^2=(L/D)f. \tag{49}$$

Loss coefficients are a function of the device geometry, the Reynolds number, and possibly the Froude number.

In many applications the internal flow is at a high Reynolds number and turbulence exists. Turbulence may be thought of as large-scale disorganized velocity fluctuations imposed on a mean flow. Most turbulent energy is in the large-scale fluctuations. However, turbulence flows have processes that reduce the length scales of the turbulent motion to smaller and smaller values until viscosity limits the process. These small-wavelength motions are where the viscous dissipation for the head loss takes place. At very high Reynolds numbers the amount of turbulence produced is independent of the Reynolds number and depends only on the geometry. Likewise the amount of energy proceeding to the small scales becomes independent of the viscosity (Reynolds number). For example, in pipe flow the friction factor at high Reynolds numbers depends only on the ratio of wall roughness height to diameter.

11. SIMPLE LAMINAR FLOWS

A convenient classification of flow situations is internal flows and external flows. In internal flows entrance and exit conditions are specified and walls confine the flow. For external flows a closed body is imagined to be in an infinite fluid that has a uniform velocity at infinity. In many instances we may also isolate a certain part of a flow to discuss a local flow pattern. A brief but thorough review of exact laminar flow solutions may be found in Wang (1989 and 1991).

11.1 Parallel Flows

This situation has one nonzero velocity component that does not vary in the flow direction. The flow is driven by a gradient of the (modified) pressure or by motion of the confining walls. Of course, the confining walls must be parallel. The continuity equation is satisfied identically and the momentum equation, Eq. (35), becomes a linear Poisson equation for $u(y,z)$ with a constant source term,

$$\nabla^2 u = -\frac{1}{\mu}\frac{dp'}{dx}. \tag{50}$$

Since the right-hand side is a constant it becomes the scale for the velocity and the flow is independent of the Reynolds number. As an example the *Poiseuille flow* between stationary parallel plates a distance H apart has the velocity profile

$$u = \frac{H^2}{2\mu}\frac{dp'}{dx}\left[\left(\frac{y}{H}\right) - \left(\frac{y}{H}\right)^2\right]. \tag{51}$$

Solutions for parallel flow in tubes of many cross-section shapes have exact closed-form answers (however, flow in a rectangular tube has only a series-expansion solution).

Couette flow is the name for flow between parallel plates, one stationary and one moving in its own plane at velocity U. This flow is driven by the wall motion, which scales the velocity,

$$u/U = y/H. \tag{52}$$

The linear nature of Eq. (50) means that flows with appropriate boundary conditions, such as the ones above, may be added.

Although these flow solutions are independent of the Reynolds number, they are only observed at low Reynolds numbers. For higher Reynolds numbers they undergo a transition to another flow that has a secondary, nonparallel motion, and/or turbulence.

11.2 Viscous Diffusion

A flow that illustrates the diffusive character of viscous effects is the *Rayleigh problem*. Stokes originally solved this problem; however, Rayleigh used it to estimate the friction drag on a flat plate. Consider a horizontal flat plate with a semi-infinite fluid above it in the y direction. The plate initially at rest is impulsively moved in its own plane, direction x, with a velocity U. This sliding motion generates a net viscous force that is diffused outward to accelerate other fluid particles, Fig. 1. The momentum equation, Eq. (35), for the velocity $u(y,t)$ takes on a parabolic form,

$$\frac{\partial u}{\partial t} \equiv \nu \frac{\partial^2 u}{\partial y^2}. \tag{53}$$

The velocity profile is a similarity solution in terms of the single variable

$$\eta = y/2(\nu t)^{1/2}. \tag{54}$$

All points with combinations of y and t that give the same η have similar velocities. The profile is

$$u = U[1 - \mathrm{erf}(\eta)]. \tag{55}$$

The associated problem where the wall is stationary but the fluid is impulsively moved with velocity U is described by the profile

$$u = U\,\mathrm{erf}(\eta). \tag{56}$$

The important result in either problem is the depth of penetration, δ, of the viscous effects after a time t. It is

$$\delta \sim (\nu t)^{1/2}. \tag{57}$$

The growth of a shear layer (a thin layer with a large velocity gradient) can be estimated using this result. Consider a flow with characteristic width D and flow direction length L. Set t equal to the time it takes to convect the flow through the region; $t = L/U$. By the time the flow reaches the position L, viscous effects will diffuse across the region a depth δ:

$$\frac{\delta}{D} \sim \left(\frac{\nu}{DU}\frac{L}{D}\right)^{1/2} = \left(\frac{1}{Re_D}\frac{L}{D}\right)^{1/2}. \tag{58}$$

This offers another interpretation of the Reynolds number as an indicator of the length scale of viscous diffusion compared to a cross-stream dimension.

11.3 Stagnation Point

When a flow approaches a body and must divide to go around on either side there is an *attachment* or *stagnation point* on the surface where the streamline direction is multivalued, Fig. 2. For high Reynolds numbers the viscous effects are confined to a region near the wall. In some small region near the stagnation point the body can be assumed equivalent to a flat wall (normal direction y, origin at the stagnation point, x and z in the plane of the wall, velocity components u,v,w), and the flow far away, $y \to \infty$, is decelerating in a linear fashion; that is $v \sim -(\alpha+\beta)y$. This flow eventually proceeds along the wall in the x and z directions. As $y \to \infty$, $u \sim \alpha x$ and $w \sim \beta z$ in various proportions according to the shape of the body. For equal flows in x and z the point is an axisymmetric stagnation point. For no flow in the z direction the flow is a plane two-dimensional flow, and the stagnation point is actually a line perpendicular to the flow direction.

The Navier-Stokes equations for these flows can be solved to give a velocity profile that rises from zero at the wall to match smoothly the flow described above. Thus the velocity profiles u and w at any location scale with αx and βz. The nondimensional coordinate normal to the wall is $\eta \equiv (\alpha/\nu)^{1/2}y$. This implies that the thickness of the profile is proportional to $(\nu/\alpha)^{1/2}$. From dimensional considerations one can show that α must have the form $\alpha = CU/L$, where C is a constant determined by the shape of the body, L a length characterizing the size of the body, and U the free-stream velocity. Although the solution satisfies the Navier-Stokes equations, the formulation of the boundary conditions is only realistic for a high Reynolds number UL/ν. Another significant result is that the pressure is constant in the y direction across the viscous layer.

12. WAKES, JETS, AND DRAG

Consider a fixed body with a characteristic dimension d in an infinite fluid. The flow at infinity is a uniform streaming motion in the x direction with a velocity U. For any Reynolds number greater than zero the effect of the body is confined to a region downstream of the body called the *wake*. For definiteness consider a two-dimensional body and a flow

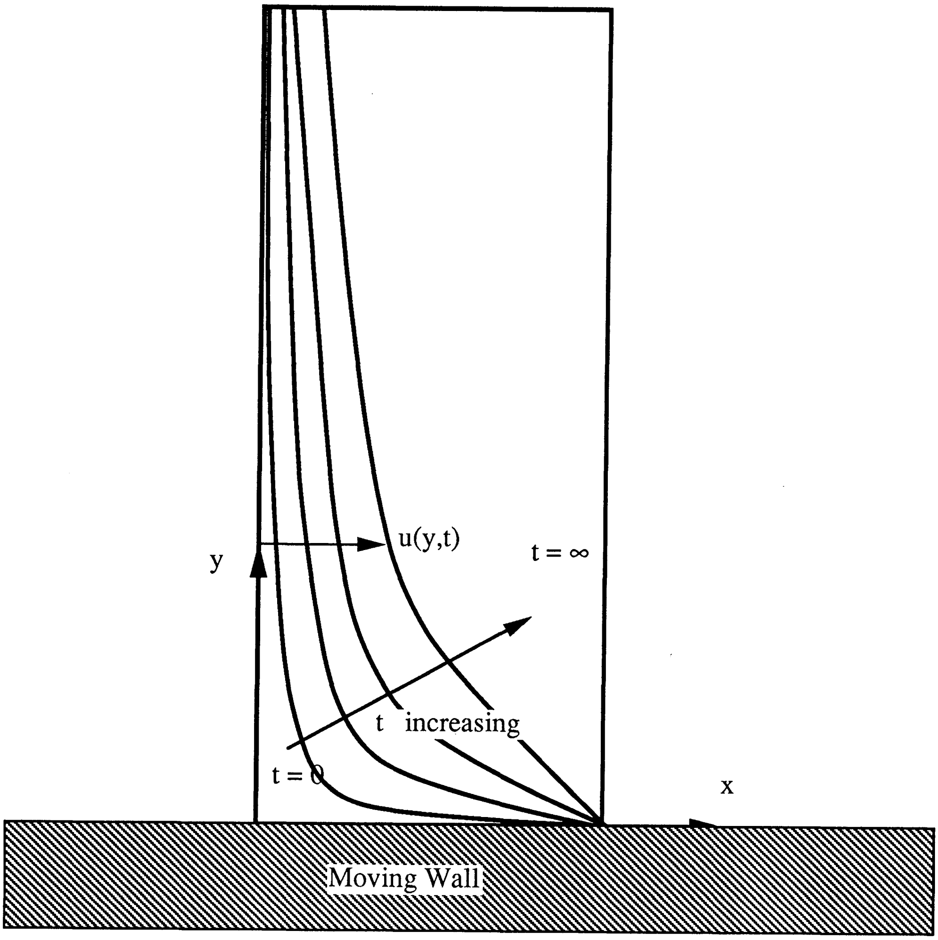

FIG. 1. Viscous diffusion away from a wall that is impulsively set in sliding motion. Acceleration of the fluid is entirely the result of viscous forces.

of a unit depth. A rectangular control region is shown in Fig. 3. It has a uniform flow in at the left with a value U. Because of the body, the flow on the right has a wake velocity profile $u_w(y)$. The upper and lower surfaces are so far removed that the velocity is U; however, in order to conserve mass there is also a small velocity perpendicular to these surfaces. The surfaces are far enough removed so that the pressure is the same on all surfaces. By applying the momentum equation to this region one finds that the drag force is related to an integral property of the wake profile,

$$C_D=\frac{F_D}{\frac{1}{2}\rho U d}=\int_{-\infty}^{\infty}\left[\frac{u_w}{U}-\left(\frac{u_w}{U}\right)^2\right]d\left(\frac{y}{d}\right)=2\frac{\Theta}{d}. \qquad (59)$$

The nondimensional drag coefficient above is the customary form for any bluff body. The drag force F_D is compared to the dynamic pressure $\frac{1}{2}\rho U$ times a cross-section area ld. The dimensional form of the integral, defined as Θ, is called the *momentum thickness*. As one goes downstream the wake decays; $u_w \rightarrow U$ and the thickness becomes larger. This can only happen in such a way that the integral above is the same constant for any downstream position. A survey of results of many drag experiments is given in Hoerner (1965).

In practical situations the wake flow is turbulent; nevertheless, the theory above is still valid with the time-averaged profile used for u_w (see TURBULENT FLOWS). The term *far wake* refers to the downstream region where the wake velocity profiles have established a similarity form. Let $L(x)$ be a characteristic width dimension of the wake and $W(x)$ be the velocity defect on the centerline, that is,

$$W(x)=U-u_w(x,y=0). \qquad (60)$$

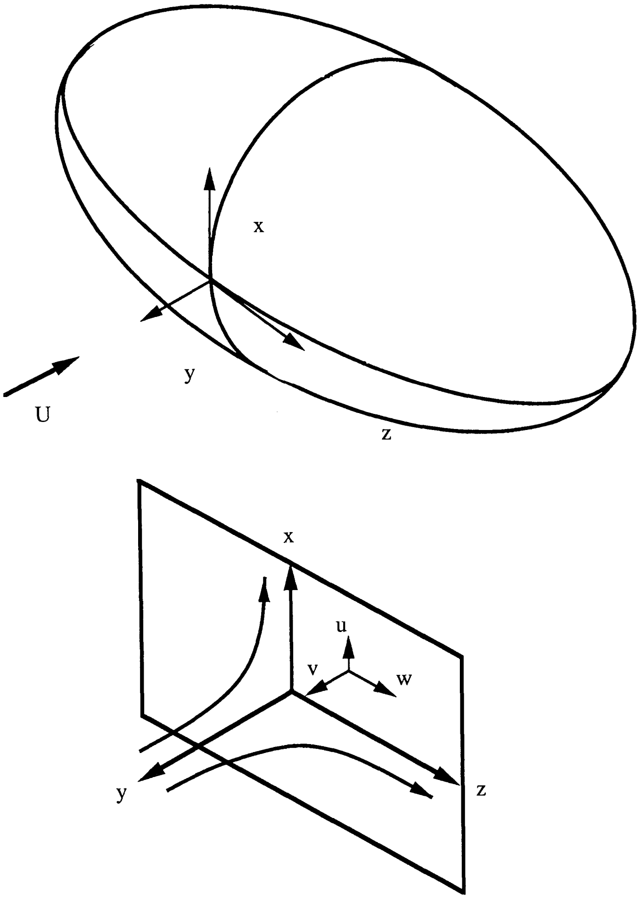

FIG. 2. Stagnation point in a three-dimensional flow. The nose of an ellipsoid has a flow that is an example of a generic stagnation point on a flat wall.

Sufficiently far downstream the exact shape of the body has been "forgotten" and the wake *defect velocity profile* is given by

$$\frac{U-u_w(x,y)}{W(x)}=F(\eta), \tag{61}$$

where

$$\eta=y/L(x).$$

In the far wake the second term in the momentum thickness definition becomes negligible and one can deduce that the growth of the wake is $L\sim x^{1/2}$ and the decay of the centerline velocity defect is $W\sim x^{-1/2}$. A detailed analysis of the momentum equation including assumptions about the turbulence is required to determine $F(\eta)$. Experiments show that the plane two-dimensional wake $u_w(x,y)$ behaves as

$$\frac{W}{U}=1.58\left(\frac{x-x_0}{\Theta}\right)^{-1/2},$$
$$\frac{L}{\Theta}=0.252\left(\frac{x-x_0}{\Theta}\right)^{1/2}. \tag{62}$$

Here x_0 is an effective origin to account for details of the flow near the body. A similar analysis for an axisymmetric wake $u_w(r,z)$ shows that the decay equations are

$$\frac{W}{U}\sim\left(\frac{z-z_0}{\Theta}\right)^{-2/3},$$
$$\frac{L}{\Theta}\sim\left(\frac{z-z_0}{\Theta}\right)^{1/3}. \tag{63}$$

Precise coefficients for these decay equations are hard to obtain. Part of the reason is that the axisymmetric wake does not sustain turbulence in its latter stages of decay.

Jets issuing into an infinite reservoir have many of the same characteristics as wakes. They are unstable at very low Reynolds numbers and commonly turbulent. Instead of a drag force, the jet is a source of positive momentum in a stagnant reservoir. The momentum crossing any downstream plane is constant,

$$U^2D=W^2L\int_{-\infty}^{\infty}\left(\frac{u_j}{W}\right)^2 d\left(\frac{y}{L}\right). \tag{64}$$

Here U and D are constants characterizing the velocity and width of the jet, $W=u_j(x,y=0)$ is the centerline jet velocity, and $L(x)$ is the local jet width. Again the exact similarity profile of the jet, $u_j/W=F(\eta)$ with $\eta=y/L$, depends on the details of the turbulence.

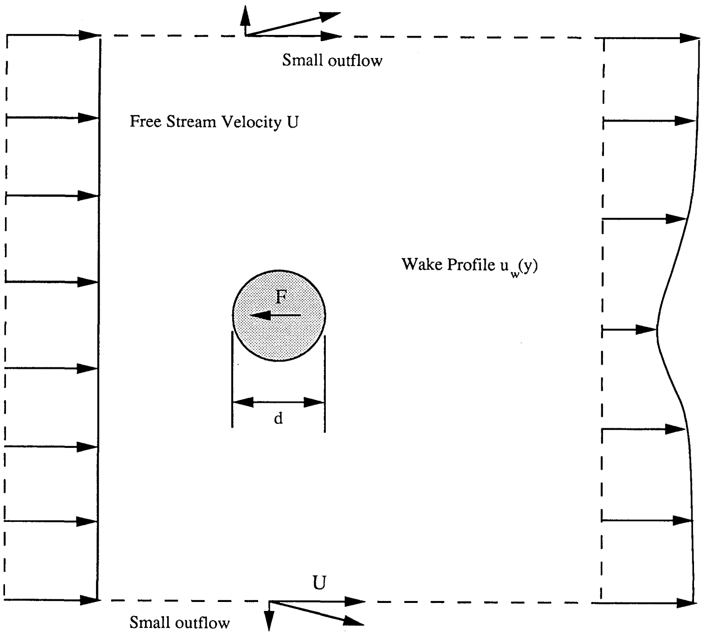

FIG. 3. Wake of a body in an infinite fluid. A control volume analysis shows that the drag of the body is related to the momentum defect of the wake velocity profile.

The development laws for jet width and centerline velocity for plane two-dimensional jets are

$$\frac{W}{U}=2.7\left(\frac{x-x_0}{D}\right)^{-1/2},$$
$$\frac{L}{D}=0.078\left(\frac{x-x_0}{D}\right). \qquad (65)$$

The corresponding relations for axisymmetric jets are

$$\frac{W}{U}=6.4\left(\frac{x-x_0}{D}\right)^{-1},$$
$$\frac{L}{D}=0.067\left(\frac{x-x_0}{D}\right). \qquad (66)$$

In both cases the linear growth of the jet width implies an increase in the mass flow across downstream stations by entrainment of reservoir fluid.

13. SWIRLING FLOWS

The term *vortex* does not have a rigorous definition. It is applied to any flow pattern with spiral or circular streamlines. A measure of vortex strength is the *circulation.* Circulation is defined as the tangential component of velocity integrated around a specified closed loop, a line integral.

$$\Gamma=\int_C \mathbf{t}\cdot v\, ds. \qquad (67)$$

Stokes's theorem allows conversion of the line integral into an integral over any surface with C as its edge. By noting that $\omega=\nabla\times\mathbf{v}$, we get

$$\Gamma=\int_C \mathbf{n}\cdot\omega\, dA. \qquad (68)$$

A flow with circulation must have vorticity somewhere within the circuit C.

13.1 Various Vortex Profiles

If the vorticity is concentrated into a Dirac δ function along the z axis we get a pure circular flow called an *ideal vortex.* The swirling flow velocity in cylindrical coordinates is

$$v_\theta=\Gamma/2\pi r. \qquad (69)$$

This is also known as a *potential vortex* because it has $\omega=0$ (the necessary and sufficient condition for a velocity potential to exist) everywhere except the origin. The profile satisfies the Navier-Stokes equations exactly, except for the singularity at the origin.

The *Rankine vortex* has uniform vorticity in a core that exist out to a distance r_0. This gives a solid-body rotation in the core, $r<r_0$:

$$v_\theta=(\Gamma_0/2\pi r_0)(r/r_0). \qquad (70)$$

Outside the core the vorticity is zero and the ideal profile above exists. This vortex model has a discontinuity in slope (or shear stress) at the core radius.

A vorticity distribution that is Gaussian in the radius will not be steady but will diffuse outward because of viscosity. The vortex associated with this distribution is called a *Lamb* or *Oseen* vortex. The velocity profile at a certain time t is

$$v_\theta=\frac{\Gamma}{2\pi r}\left[1-\exp\left(-\frac{r^2}{4\nu t}\right)\right]. \qquad (71)$$

The governing equation in cylindrical coordinates is

$$\frac{\partial(rv_\theta)}{\partial t}=\nu\left(\frac{\partial^2(rv_\theta)}{\partial r^2}-\frac{1}{r}\frac{\partial(rv_\theta)}{\partial r}\right). \qquad (72)$$

This vortex can be viewed as the decay of an ideal vortex and Eq. (71) is an exact solution to the Navier-Stokes equations. The circulation at infinity, or total vorticity, is constant. The Oseen vortex contains an infinite amount of angular momentum and kinetic energy.

A vortex that has a finite amount of angular momentum is the profile discovered by G. I. Taylor. If the differential equation above is differentiated with respect to t the form governing $\partial v_\theta/\partial t$ is the same as that governing v_θ. The profile found by differentiating the Oseen profile is

$$v_\theta=\frac{Hr}{8\pi\nu t^2}\exp\left(-\frac{r^2}{4\nu t}\right). \qquad (73)$$

Here H is the strength constant. There are an infinite number of vortex solutions like the *Taylor vortex,* each with a sharper profile and a faster time decay. In fact, these vortex solutions form a complete set of basis functions of Eq. (72).

When a vortex is "pulled" or stretched along its axis the vorticity is intensified. Hence, one may reach an equilibrium between intensification and radial diffusion to obtain a steady-state profile. This is known as *Burgers's vortex.* The straining is given by radial and axial profiles

$$v_r = -ar, \quad v_z = 2az. \tag{74}$$

The swirling profile is a steady Oseen profile:

$$v_\theta = \frac{\Gamma}{2\pi r}\left[1 - \exp\left(-\frac{ar^2}{2\nu}\right)\right]. \tag{75}$$

The parameter a defines the straining.

An important illustration that the longitudinal and radial flows in a vortex can be quite complex is *Sullivan's vortex.* At $r \to \infty$ the flow is like Burgers's vortex. Near the axis, however the flow has two cells as shown in Fig. 4. The velocities are

$$v_r = -ar + \frac{6\nu}{r}\left[1 - \exp\left(-\frac{ar^2}{2\nu}\right)\right],$$

$$v_z = 2az\left[1 - 3\exp\left(-\frac{ar^2}{2\nu}\right)\right], \tag{76}$$

$$v_\theta = \frac{\Gamma}{2\pi r} H\left(-\frac{ar^2}{2\nu}\right),$$

where the function $H(\zeta)$ is defined by

$$H(\zeta) = \frac{1}{37.905}\int_0^\zeta \exp\left(-z + 3\int_0^z [1 - \exp(-s)]s^{-1}ds\right)dz. \tag{77}$$

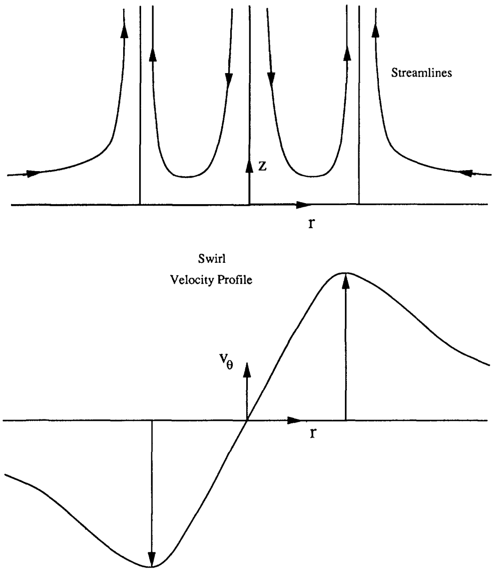

FIG. 4. Sullivan's vortex. Velocity profile and cross-section view of streamlines show a complex pattern with flow going upward in one part of the vortex and downward in another part.

13.2 von Kármán Viscous Pump

The interaction of a swirling flow and a solid wall occurs in several situations. In one of the most general problems, a solid wall (disk) in the x-y plane is imagined to rotate about the z axis with a specified angular velocity. A second parallel disk at $z = z_0$ has another specified angular velocity. One of the simplest cases is a rotating disk in a semi-infinite fluid, the *von Kármán viscous pump.* Here, the second disk is stationary and z_0 is at infinity. Another interesting case is when the disk at infinity rotates and the disk at the origin is at rest, that is the flow of a fluid with a solid-body rotation above a stationary wall. This is known as the *Bödewadt problem.*

In both of these problems the proper scaling for the z directions is

$$\zeta = z(\Omega/\nu)^{1/2}. \tag{78}$$

The velocity of swirl has the form

$$v_\theta = r\Omega G(\zeta). \tag{79}$$

In the von Kármán problem G goes monotonically from $G = 1$ at $\zeta = 0$ (the rotating wall) to $G = 0$ at $\zeta = \infty$. Actually, the fluid motion is confined to a region near the moving wall; $G(\zeta = 4)$ is nearly zero. Viscous forces have diffused from the rotating disk to accelerate the fluid near the disk. Note that the magnitudes of the velocities are proportional to $r\Omega$ and the depth of rotating fluid proportional to $(\nu/\Omega)^{1/2}$. For high rotation speeds the layer is very thin.

The first effect is that the rotating disk accelerates nearby particles by viscous forces. Since the fluid has no pressure gradient to keep the particles in a circular path they are flung outward along the disk. The radial velocity is given by

$$v_r = r\Omega F(\zeta), \tag{80}$$

At the rotating disk $F(\zeta = 0) = 0$ and at infinity $F(\zeta = \infty) = 0$; however, a peak occurs at about $\zeta = 1$. This outward flow near the disk must be replaced by a flow from infinity toward the disk, hence the term *viscous pump*. The inward flow is scaled by

$$v_z = -(\nu\Omega)^{1/2} H(\zeta). \tag{81}$$

Since the flow must turn when it approaches the disk, $H(\zeta=0)=0$, and the constant flow from infinity toward the disk is determined from the solution as $H(\zeta=\infty)=0.886$.

The Bödewadt problem is similar except that the disk, and therefore the flow, at infinity is rotating as a solid body while the disk at $z=0$ is stationary. At infinity there must be a pressure gradient in the radial direction that turns the particles in a circular pattern. At the stationary disk (wall) the viscous forces destroy this motion in a thin layer near the wall. Now, the equilibrium of the centrifugal motion and the pressure gradient is upset. The pressure gradient forces the particles toward the center as they slow down. This flow is accompanied by a uniform outflow along the z axis.

To a great extent these problems are complementary. The absence of a pressure gradient in the von Kármán problem yields a spiral outward flow along the rotating disk; the presence of a pressure gradient yields a spiral inward flow along the stationary disk when the rotating fluid is slowed in the Bödewadt problem.

13.3 Vortex Breakdown

Intense vortices undergo a transition called vortex *breakdown* or vortex *bursting*. Viewed in a coordinate frame where the breakdown is stationary, the flow has a large axial velocity which is usually jetlike in the core. A typical empirical axial profile is

$$v_z = V\left[1 + Q\exp\left(-\frac{ar^2}{2\nu}\right)\right]. \tag{82}$$

Here Q and a are parameters. The swirl profile is gives by the steady Oseen or Burger vortex discussed above. The transition begins as flow on the vortex axis slows down and the vortex expands. Two major types of breakdown are observed. Figures 5 and 6 from Faler and Leibovich (1977) give examples. Both types have a stagnation point on the axis and a region of reverse flow. If a dye filament is introduced in the vortex core the *bubble-type breakdown* will show a large bubble of collected dye in a separated region with reverse flow. There is a slow exchange of fluid in the bubble with the surrounding flow. Dye going around the bubble will form a tail that spirals into a region of turbulent flow. In the *spiral-type breakdown* the dye filament has a kink at the stagnation point and spirals out into a turbulent flow. The final flow in either

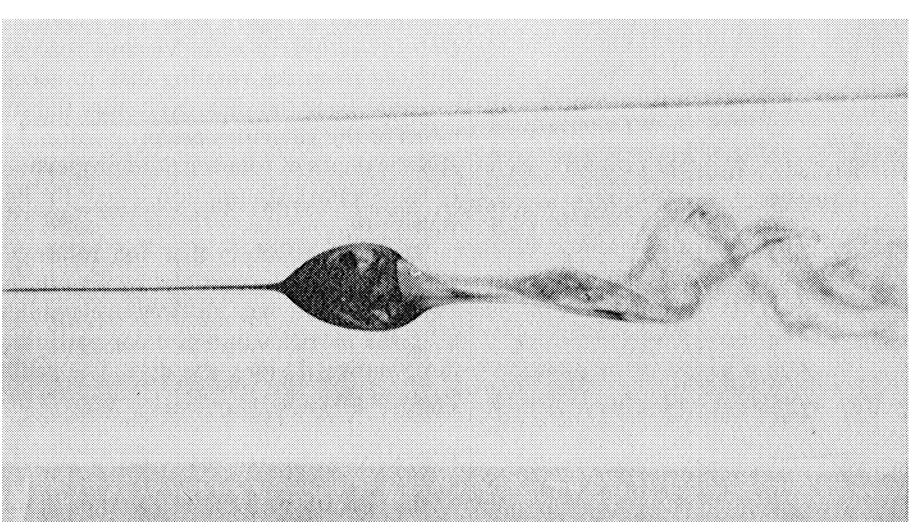

FIG. 5. Bubble-type vortex breakdown. The recirculation bubble is simultaneously filled and emptied from the rea . (From Faler and Leibovich, 1977).

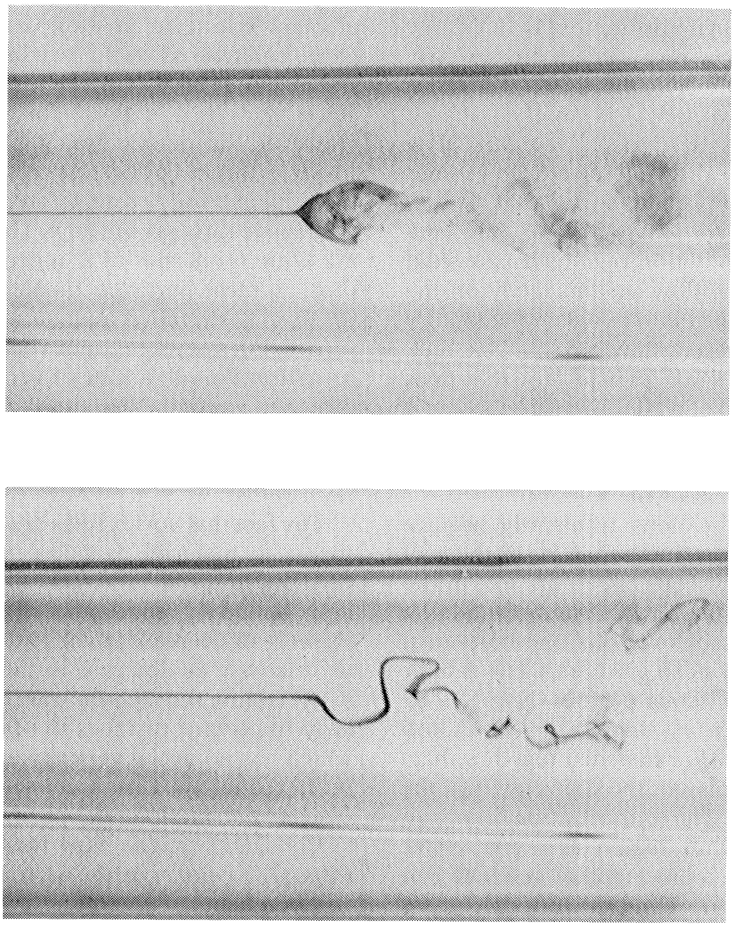

FIG. 6. Spiral-type vortex breakdown. The common mode that occurs on aircraft delta wings. The kink always breaks down into turbulence after a turn or two. (From Faler and Leibovich, 1977.)

case will be a turbulent vortex with about twice the radius of the original vortex.

An axisymmetric theory of an inviscid flow shows that there are two distinct velocity profiles, conjugate flow states, such that each fluid particle has the same angular momentum and the same total pressure. Thus the initial and final states are connected by jump conditions analogous to a shock wave or a hydraulic jump. It is further found that vortices can support axisymmetric traveling waves. The vortex is called supercritical if waves can only propagate downstream and subcritical if waves can propagate upstream. One proposition is that vortex breakdown occurs when disturbance waves from downstream collect at a location where they can no longer move against the flow. The fact that vortex breakdown is a finite-amplitude flow structure makes this an oversimplification; however, it is observed experimentally that the upstream vortex flow is supercritical and the downstream flow subcritical. Review articles on vortex breakdown are Escudier (1988) and Leibovich (1984).

14. HIGH-REYNOLDS-NUMBER FLOWS

The viscosity of common liquids and gases is small, and it does not take a very high velocity to yield a high Reynolds number. Consider a characteristic flow dimension of 10 cm and a velocity of 1 m/s; then for air (ν=0.150 cm^2/s) the Reynolds number is

$UL/\nu=6670$. Water in this sense is even less viscous than air ($\nu=0.010$); for water the corresponding Reynolds number is 10^5. Large Reynolds numbers mean, according to Sec. 10.3, that the diffusion distance for viscous shear stress (or vorticity) is small compared to the characteristic dimension of the flow field. Because the diffusion distance becomes small, the first approximation is that the viscous effects (or vorticity gradients) are confined to vanishingly thin regions on the solid boundaries of the flow. In this situation the flow may be divided into two complementary regions with different flow physics. The bulk of the flow field is an *inviscid flow* that does not satisfy the no-slip condition. The inviscid flow slips past the wall on the thin viscous region. The magnitude of the slip velocity over the walls is a result of the solution.

Practical applications frequently require knowledge of the wall shear stress or the wall heat flux. These viscous effects are lost in the inviscid approximation. To understand the events in the viscous regions one must deal directly with the steep gradients. The proper physics in the viscous regions is given by *boundary-layer theory.* Boundary-layer theory and inviscid flow are matched together, having as the common part the slip velocity of the inviscid solution. The concentration of certain physics into a thin region as a parameter becomes large is a prototype of singular perturbation problems. Thus, the term boundary layer is applied to other events of a similar physical structure.

14.1 Inviscid Flows

At high Reynolds numbers, the first approximation is that an inviscid flow covers the entire domain. If the governing equations are nondimensionalized with a characteristic velocity U, length L, and fluid density ρ and the Reynolds number is set equal to infinity, all terms originally preceded by the viscosity coefficient drop out of the equations. Euphemistically one says that the flow is that of an inviscid fluid.

The inviscid vorticity equation is

$$\frac{D\boldsymbol{\omega}}{Dt}=\boldsymbol{\omega}\cdot\nabla\mathbf{v}. \tag{83}$$

From this equation some important facts about vorticity may be deduced. The similar structure of Eq. (83) governing $\boldsymbol{\omega}$ and Eq. (2) governing the length vector of a material line element $\delta\mathbf{r}$ leads to an intuitive understanding of the behavior of vorticity in inviscid flows. A vortex line is an instantaneous concept in the sense that it is defined for a given time. However, for inviscid flow it is permissible to imagine that a vortex line is attached to the material and follows along on the particles as they move through the flow. The vorticity has the same properties as a material line drawn through these particles. Vorticity increases if the line is stretched and turns if the line is turned. Furthermore, since vorticity obeys the condition $\nabla\cdot\boldsymbol{\omega}=\mathbf{0}$, a tube of vortex lines has a constant strength, the strength being measured by the circulation or the integral of the vorticity across any cross section of the tube, Eq. (67).

The fact that vortex lines are attached, so to speak, to material particles means that the vorticity is specified as part of the incoming flow boundary conditions. The subsequent increase or decrease in vorticity is the same as the increase or decrease in the length of the material line through the particles. A particularly important result is the fact that a flow which has no initial vorticity will continue to be irrotational. This special class is called *ideal* or *potential flows.*

Consider a steady flow with a vortex line through a certain set of particles. These particles also have streamlines, which are also particle paths, and together with the vortex line they form a surface called the *Lamb surface.* The normal to the surface will be perpendicular to both $\boldsymbol{\omega}$ and $\mathbf{v}$, that is, in the direction $\mathbf{v}\times\boldsymbol{\omega}$. On a Lamb surface it is possible to integrate the momentum equation to find the famous *Bernoulli equation,*

$$p/\rho+gh+\tfrac{1}{2}v^2=\text{const.} \tag{84}$$

The value of the constant may be different for each Lamb surface. If the flow is irrotational, Lamb surfaces can no longer be identified and the same constant applies to the whole flow.

The momentum equation normal to the streamlines cannot be explicitly integrated; however, it is informative. At any point a streamline has a radius of curvature R, whose direction defines a coordinate n normal to the streamline. The inviscid momentum equation for steady flow may be written as

$$\frac{d}{dn}\left[\frac{p}{\rho}+gh\right]=-\frac{v^2}{R}. \tag{85}$$

The bracketed term is actually the modified pressure p'/ρ. Any change in this pressure across streamlines is balanced by the centrifugal force v^2/R. In particular, if the streamline is straight, $R\to\infty$, there is no change in the pressure in the normal direction. For example, a jet flowing into a large reservoir has the same pressure in the core of the jet exit as that of the reservoir.

Ideal flows are dominated by the shape of the boundaries. To see why this is true first note that in an ideal flow the particles do not rotate,

$$\omega=\nabla\times\mathbf{v}=\mathbf{0}. \tag{86}$$

This is the necessary and sufficient mathematical condition for the existence of a *velocity potential* ϕ that obeys

$$\mathbf{v}=\nabla\phi. \tag{87}$$

The incompressible continuity equation demands that particles do not expand and furnishes a relation that governs ϕ,

$$\nabla\cdot v=\nabla^2\phi=0. \tag{88}$$

Thus the solution for the velocity and streamlines is governed by an elliptic equation. Specifying the boundary shape and the normal component of velocity across the boundary (which for a solid wall is zero) completely determines the flow on the interior. The entire flow is sensitive to a change in the boundary; in particular, there is always an upstream influence. These flows are dominated by geometry. Once the streamlines and velocities are set by the geometry and kinematics (no rotation or expansion), the pressure force required to produce these motions may be found from the Bernoulli equation. It is only at this stage that a fluid property, the density, enters as a scaling factor for the level of the pressure, that is, p/ρ.

14.1.1 Internal Flows. The ideal flow into a channel contraction is shown in Fig. 7. The geometry forces the streamlines to curve and crowd together as they enter the smaller section. Inviscid flow slips over a wall; hence, any interior streamline might be considered as a wall containing a flow. One might also imagine that the flow is one half of a symmetric contraction.

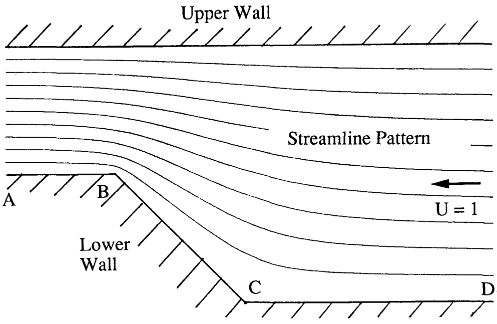

FIG. 7. Ideal flow into a contraction. Streamlines for equal increments of volume flow. Wider distance between streamlines indicates a slower velocity, and a smaller distance a faster velocity.

Streamlines have been drawn in equal increments, that is the volume flow rate AV is constant (A is the cross-sectional area between streamlines and V the average velocity). This means that where the streamlines are far apart the velocity is low and vice versa. By Bernoulli's equation the pressure at any location decreases if the velocity is high and increases if the velocity is low. The centers of curvature are first above the streamlines and then below the streamlines. Correspondingly the pressure gradient across streamlines changes sign in these different regions.

First, consider the flow along the upper flat wall. The streamtube smoothly contracts to a smaller area. Correspondingly in Fig. 8 the velocity monotonically increases to a higher value. The pressure will also show a monotonic decrease to the final value in the small channel. The situation is quite different for the bottom wall. As the flow approaches the concave corner, the stream tube broadens. Theory shows that there is always a stagnation point at a concave corner. Thus the pressure here is the maximum value, the stagnation pressure. The high pressure not only slows the flow approaching the corner but it gives a normal pressure gradient across the streamlines required by the radius of curvature. The rising pressure as the flow enters the corner (even one with a smooth contour) is a practical concern, because it is a possible place for the boundary layer to separate (see Boundary Layers, Sec. 14.2). As the flow proceeds up the sloping wall and approaches the convex corner the streamlines converge. The velocity at a

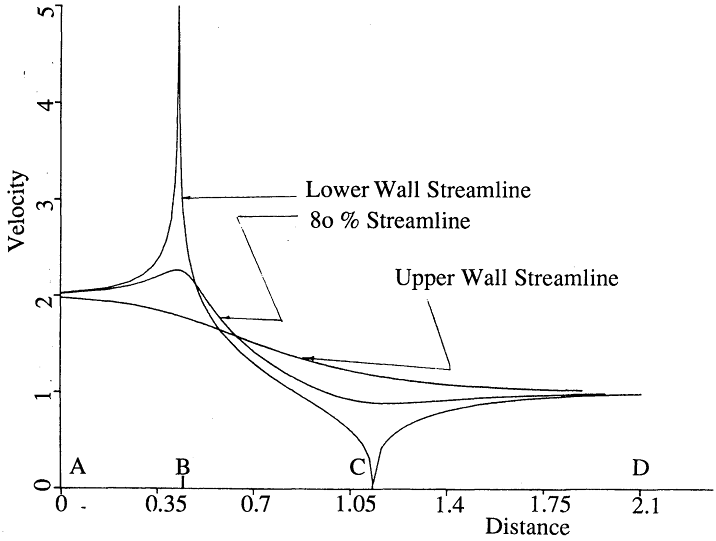

FIG. 8. Ideal flow into a contraction. Velocity as a function of distance along the streamlines of Fig. 7. A slower velocity indicates a higher pressure and vice versa. For streamlines in the lower half of the contraction the pressure first rises, then decreases to a very low value, and subsequently rises again to the final value.

convex corner is theoretically infinite. The corresponding pressure is unreasonably negative. Downstream of the corner, the flow encounters another adverse pressure gradient and again the likelihood of separation.

Mathematically one can show that the flow in the lower half of the channel first slows down approaching the concave corner, and then accelerates to overshoot the final velocity near the convex corner. On the other hand, the flow in the upper half of the channel monotonically accelerates to the higher value. It has no regions of adverse pressure gradients.

14.1.2 External Flows. A solid body is frequently so far removed from other objects that it can be considered to be in an infinite fluid. First, consider the nose region of a blunt but smoothly rounded body, either two dimensional or axisymmetric. The oncoming flow slows down and stops at the stagnation point. Here the static pressure is now equal to the free-stream stagnation or total pressure (free-stream static plus free-stream dynamic pressure). The pressure on the surface then drops as the flow accelerates away to either side. In the process the velocity always becomes greater than that of the approaching free stream, and the pressure is correspondingly lower than the free-stream pressure.

Downstream from the point where the maximum surface velocity occurs, the flow slows down and the associated pressure rises. If one looks at the object head on, there is a high-pressure region surrounding the stagnation point (or stagnation line in the case of a two-dimensional object) followed by a low-pressure region that continues out to the edge of the object. A somewhat surprising fact is that net integrated drag force on the frontal area is zero regardless of the shape of the nose.

After the flow reaches its maximum velocity the pressure increases and an adverse pressure gradient exists. Thus on the shoulders of an object there is a possibility of boundary-layer separation. This is called leading-edge separation in the case of wings. If it does occur it is usually a very thin region that soon reattaches to the body. In that case it does not materially affect the drag. On bodies with sharp corners, a square-fronted object, for example, the flow separates in a large region along the sides. Now the low-pressure region is lost (the proper object for an ideal flow calculation is the body plus an accounting for the separated wake region), and a substantial nose drag is encountered.

The design of wings or blades to produce a large normal force, lift force, is the subject of aerodynamics (*q.v.*). The exact shape of a wing results from an effort to control separation and decrease drag.

Many effects of ideal flow can be understood with reference to the energy of the flow. In order to have a finite kinetic energy we consider a body that is moving through a fluid that is at rest at infinity. Let the body have a unidirectional velocity $U(t)$. Consider a region having one surface at infinity, containing the fluid on the inside, and a second surface at the wall of the body. The global mechanical energy, Eq. (19), is especially simple for inviscid, incompressible flow. It is

$$\frac{d}{dt}\int_{\text{flow}} \rho\tfrac{1}{2}v^2 dV = -\int_{\text{body}} \mathbf{n}\cdot vp\, dS = UF_D. \qquad (89)$$

This equation shows that the work done by the moving body, UF_D, must appear as a increase in the global kinetic energy of the fluid.

For a steadily moving closed body *without a lift force,* the ideal flow in the wake is a potential flow that returns to a still state in a short distance behind the body. The rate of increase in the energy of the flow field is zero and so the drag is also zero, $C_D=0$. In many cases this is unrealistic because flow separates from the body. The turbulent separated flow produces a wake that has a finite amount of kinetic energy. In this case we interpret the drag force work as that required to increase the length of the turbulent wake. This is the energy viewpoint that is complementary to the momentum analysis of Sec. 12. The idea that a body moving through a fluid can leave a permanent mark at infinity also applies to ship waves. The waves propagate to infinity with undiminished energy. Hence, by the equation above there is a wave drag that is equal to the rate at which energy is supplied to the wave system. A similar event happens in the flow over a lifting body. The lift force can only be generated by a net circulation. Therefore the vortex lines associated with this circulation must ultimately, for a three-dimensional body, trail off in a singular surface in the wake. These vortex lines produce a flow of finite energy in the plane perpendicular to the body motion; see Fig. 9. Thus the drag, called *induced drag* or *drag due to lift,* is related to the increase in the length of the wake. For a more extensive discussion see AERODYNAMICS.

If a body is accelerated, the energy of the flow surrounding it increases, and an extra pressure force at the body surface is needed to accomplish this. In the case of a potential flow all velocities scale with U and the average extent of the flow is a constant fraction K of the body volume V. The drag force that the acceleration must overcome is

$$UF_D = \frac{d}{dt}\int \rho\tfrac{1}{2}v^2 dV = \frac{d}{dt}(\rho\tfrac{1}{2}U^2 K\mathrm{V})$$
$$= K\rho\mathrm{V}U\frac{dU}{dt}. \qquad (90)$$

The force is equivalent to a certain mass of fluid, $K\rho\mathrm{V}$, added to the inertia of the body, hence the term *added mass.* The added-mass constant for a sphere is $K=\frac{1}{2}$. The theory still holds for surfaces also. A disk moving normal to its plane has an equivalent $K\rho\mathrm{V}=\rho 8r_0^3/3$. If the fluid is a liquid, the added-mass forces can be large because ρ is large. In gases they are negligible unless the motion is rapid, for

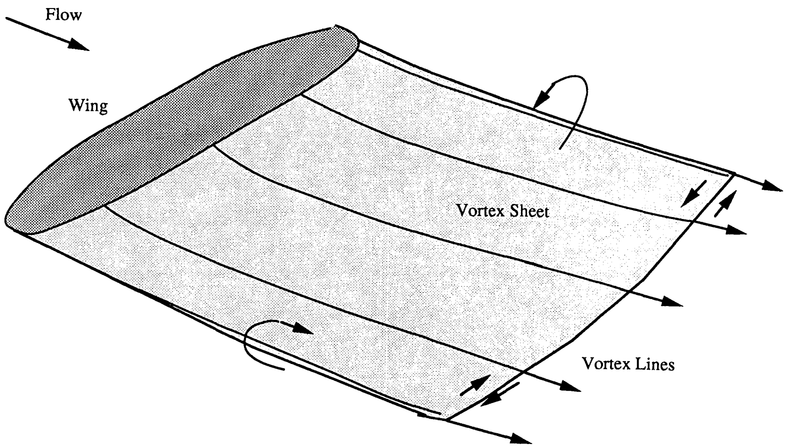

FIG. 9. Ideal potential-flow vortex sheet behind a finite wing. This sheet is a singular surface in the solution of the Laplace equation. Spanwise velocity is discontinuous across the sheet even though the longitudinal velocity is continuous. A finite vortex may exist at the edges of the sheet.

example, a loud-speaker cone, or the surface is large, such as the opening of a parachute.

14.2 Boundary Layers

Inviscid flow theory fails where the fluid slips over the walls. It also fails where one stream slips over another at a higher speed (this would be a sheet of vorticity). These apparent discontinuities in velocity are smoothed out by viscosity in thin regions called *boundary layers.* They exist at high Reynolds numbers because the diffusion length scale for viscous effects is much smaller that the characteristic dimension of the flow field.

Elementary boundary-layer theory is developed for plane bodies with surfaces having a smoothly varying radius of curvature, R. A set of coordinates is laid out with the x axis along the surface and y normal to the surface (Fig. 10). If R is large compared to the thickness of the boundary layer, the coordinates may be considered as a Cartesian system, that is, no curvature terms appear in the equations. Flow in a steady two-dimensional boundary layer, with velocity components $u(x,y)$ and $v(x,y)$, is governed by the following equations:

$$\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}=0, \tag{91}$$

$$\frac{\partial p}{\partial y}=0 \Rightarrow p=p(x \text{ only}), \tag{92}$$

$$u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}=-\frac{1}{\rho}\frac{dp}{dx}+\nu\frac{\partial^2 u}{\partial y^2}. \tag{93}$$

The first equation is the continuity equation and the second is the y-direction momentum equation. Equation (92) reduces to a statement that the pressure is constant across the boundary layer. This means that the known pressure of the inviscid flow is impressed on the boundary layer. The last equation is the x-direction momentum equation. One viscous term, $\nu\partial^2 u/\partial x^2$, has dropped out of the equation as negligible. The boundary-layer equations are parabolic in mathematical character. The terms $u\partial u/\partial x$ and $\nu\partial^2 u/\partial y^2$ dominate the x and y behavior, respectively.

Appropriate boundary conditions are the no-slip condition at the wall, the condition that inviscid flow cannot satisfy,

$$u(x,0)=0, \quad v(x,0)=0, \tag{94}$$

and initial u velocity profile within the layer,

$$u(x_0,y)=u_i(y). \tag{95}$$

The final boundary condition is that the u velocity (at large y) match the inviscid velocity (at the wall). Specifically,

$$u(x,y\to\infty)\equiv u_e(x)=u_{\text{inviscid}}(x,y=0). \tag{96}$$

The velocity at infinity in the boundary layer $u_e(x)$ matches the (slip) velocity of the inviscid flow on the wall. The velocity $u_e(x)$ is called the *external flow*, also sometimes called the free stream since it occurs as $y\to\infty$. No initial nor matching conditions on the vertical velocity v can be prescribed without overdetermining the problem. The matching condi-

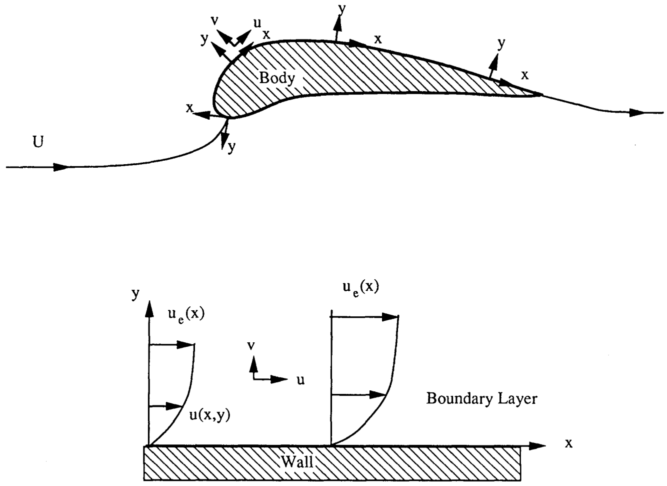

FIG. 10. Boundary-layer coordinates. Coordinates are tangent to the surface and normal to the surface. The boundary layer is so thin and the curvature of the surface so small that the coordinates are essentially Cartesian.

tion also allows the pressure to be given in terms of u_e. Evaluating Eq. (93) as $y \to \infty$ yields

$$u_e \frac{du_e}{dx} = -\frac{1}{\rho}\frac{dp}{dx}. \tag{97}$$

This is a differential form of the Bernoulli equation.

The simplest boundary-layer flow is the flow over a flat plate with zero pressure gradient; $u_e = u_0$ and $u_i = u_0$. It is known as the *Blasius flow.* This is actually a member of a class of solutions called *Falkner-Skan flows.* These are *similarity* solutions in the sense that there are trajectories in the x-y plane with similar velocities. That is,

$$\frac{u(x,y)}{u_e(x)} = f'(\eta), \tag{98}$$

where

$$\eta = y/\delta(x). \tag{99}$$

The notation to call the velocity f' is because f has a physical interpretation. Specifically $f = \text{const}$ is a streamline. The boundary-layer equations admit similarity solutions if u_e and δ vary as power functions

$$u_e = u_0(x/L)^m, \tag{100}$$

$$\delta = [\nu x/u_e(x)]^{1/2}. \tag{101}$$

The solution $m=1$ is the plane stagnation point discussed in Sec. 11.3. Solutions $0 < m < 1$ are associated with inviscid flows over a plane two-dimensional wedge of included angle $\Theta_0 = 2\pi m/(m+1)$. The angle is zero for $m=0$ and Blasius's solution is obtained. All of the flows above are accelerating for $m > 0$. Unique solutions for decelerating flows exist only for the range $-0.0904 \leqslant m < 0$; see Fig. 11. The $m = -0.0904$ solution has zero shear stress and is on the verge of separating for all x.

The similarity solutions are also used as initial profiles for bodies that begin with the appropriate local geometry. Any body with a blunt leading edge begins with the stagnation-point boundary layer, a body beginning with a sharp edge of finite angle begins with a wedge flow boundary layer, and a body beginning with a sharp cusp has the Blasius flow as the starting solution.

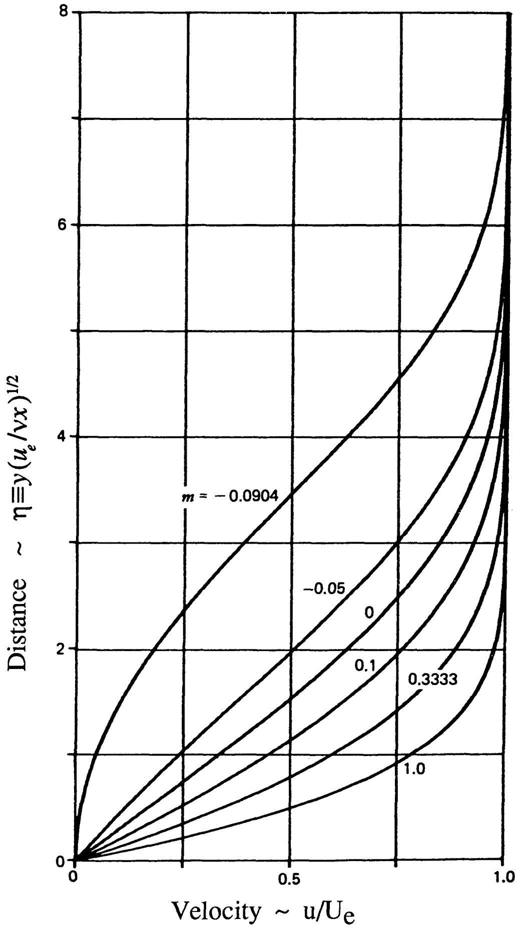

FIG. 11. Velocity profiles in Falkner-Skan boundary layers. The parameter m indicates the external velocity by $u_e(x) = ax^m$. Positive m indicates an accelerating flow with a favorable pressure gradient. Adverse gradients exist only up to $m = -0.0904$, where the flow is on the verge of separating.

14.2.1 Physical Characteristics.

Boundary layers have the same pressure across the layer at any x location. This pressure is imposed by the inviscid flow and largely determines the development of the layer. Without any pressure gradient, the Blasius solution, the boundary-layer thickness grows as $\delta \sim x^{1/2}$ and the wall shear stress decreases as $\tau \sim x^{-1/2}$. An accelerating flow, $dp/dx < 0$, is called a *favorable pressure gradient.* It tends to thin the boundary layer, increase the shear stress, and broaden the velocity profile. An *adverse pressure gradient,* $dp/dx > 0$, tends to thicken the layer rapidly and decrease the wall shear stress. Since the same pressure force acts on all particles in the layer, those next to the wall with the least momentum are slowed most rapidly. It takes only a slight application of adverse pressure to turn the particles next to the wall around and cause them to flow upstream. When this happens *separation* is said to occur. A thin separation bubble that reattaches quickly can be calculated with boundary-layer equations.

However, if the separation region is large, then the assumption that the boundary-layer theory is thin is violated and the calculation is stopped. A large separated region modifies the inviscid flow upstream. Hence a proper boundary-layer calculation needs to have the modified external velocity as its boundary condition. If the actual pressure distribution is used in the boundary-layer calculation, then the prediction of the separation point is reasonably accurate.

The parabolic nature of the boundary-layer equations means that effects are carried downstream (the timelike direction), but that there is no upstream influence. A change in the external velocity $u_e(x)$ after some point in the flow will not be felt upstream of that point. The dominant convective term $u(\partial u/\partial x)$ gives the timelike behavior. In an adverse pressure gradient, when particles near the wall are turned around u becomes negative and the direction of signal characteristics changes. Thus information is carried backward by the reverse flow near the wall in the region of separated flow. In very thin separation bubbles the reverse convection can be ignored; however, in large separation regions it cannot.

14.2.2 Three-Dimensional Effects.

Boundary layers on axisymmetric bodies are only slightly more complicated than two-dimensional layers. The only new physical feature is the divergence or convergence of the streamlines as they move to places of higher or lower radius. There is a transformation, Mangler's transformation, that maps the axisymmetric problem back into an equivalent two-dimensional boundary layer. Divergence thins the layer and increases the shear stress while convergence thickens the layer and decreases the stress. Pressure gradients do not change sign in the transformation. An adverse gradient is still adverse in the equivalent two-dimensional problem.

A boundary layer on a general cylindrical object placed at a sweep angle with respect to the free stream is three dimensional in the sense that all three velocity components are variables. However, the flow in a plane perpendicular to the leading edge is exactly like a two-dimensional boundary layer. This is known as the sweep independence principle. Once this flow is known the spanwise flow parallel to the cylinder axis can be calculated.

Truly three-dimensional boundary layers can have much more complicated behavior. Consider a set of inviscid streamlines that are curving over a surface boundary layer. The inviscid flow curvature is produced by a transverse pressure gradient (the pressure gradient in the flow direction slows or speeds the velocity):

$$\frac{d}{ds}\left(\frac{p}{\rho}\right) = -\frac{V^2}{R}\,. \qquad (102)$$

This gradient acts undiminished through the boundary layer, and, in particular, it acts on the particles deep within the layer that have only a small velocity. It causes, through an essentially inviscid mechanism, a much smaller radius of curvature. Thus a general rule is that fluid within the boundary follows more highly curved streamlines than does the external flow. The particle paths within the boundary layer tend to separate. At any position the particles through a boundary-layer profile have come from a variety of upstream locations.

Likewise the particles at this location will spread to different regions as they go further along. This transverse spreading is another distinct action that happens in a three-dimensional boundary layer. Since particle paths are subcharacteristics of the boundary-layer equations, transverse spreading must be accounted for in formulating boundary conditions and solving problems.

Secondary flow is the flow in the direction perpendicular to the external streamline and parallel to the wall. It is a jetlike profile, zero at the wall and zero again at $y \to \infty$. This means that the net viscous force in the outer part of the jet is in the direction of the jet. Therefore it can add to the pressure force that originally produced the jet and generate velocities that are higher in magnitude those in the inviscid flow. It is not uncommon to find velocities in the upper parts of boundary layers that are significantly higher than those in the external inviscid flow.

15. LOW-REYNOLDS-NUMBER FLOWS

When the Reynolds number of a flow is very low, the inertia effects are negligible compared to viscous and pressure effects. These flows are also known as *Stokes flows* or *creeping flows;* however, "low–Reynolds-number flows" is the most descriptive name.

At moderate or high Reynolds numbers the proper scales are ρ, U, and L [except for the normal coordinate in boundary-layer theory where $(\nu/U)^{1/2}$ is needed]. For low Reynolds numbers the density (inertia) is irrelevant and the appropriate scales are μ, U, and L. When these scales are used, the momentum equation in dimensionless variables is

$$Re\,\frac{D\mathbf{v}^*}{Dt}=-\nabla p^{**}+\nabla\cdot\nabla\mathbf{v}^*. \tag{103}$$

In the low–Reynolds-number limit the inertia term drops out and the pressure and viscous forces must balance everywhere in the flow,

$$\mathbf{0}=-\nabla p^{**}+\nabla\cdot\nabla\mathbf{v}^*. \tag{104}$$

An equivalent form in terms of the stress tensor is

$$\mathbf{0}=-\nabla p^{**}+\nabla\cdot\boldsymbol{\tau}^*. \tag{105}$$

The appropriate form of the vorticity equation is

$$\begin{aligned}\mathbf{0}&=\nabla\cdot\nabla\boldsymbol{\omega},\\ \mathbf{0}&=\nabla\cdot\nabla(\nabla\times\mathbf{v}).\end{aligned} \tag{106}$$

This equation states that diffusion is the only mechanism to govern vorticity in low-*Re* flows. It is important to notice that all low-*Re* equations are linear and independent of time.

As the viscous forces become stronger and stronger, the pressure rises to counteract them. For this reason the proper nondimensional pressure is

$$p^{**}=\frac{p-p_{\text{ref}}}{\mu U/L}. \tag{107}$$

Note that the low–Reynolds-number pressure scale compared to the high–Reynolds-number pressure, the dynamic pressure, scale is the inverse Reynolds number:

$$\frac{\mu U/L}{\frac{1}{2}\rho U^2}=\frac{1}{2\,Re}. \tag{108}$$

In low–Reynolds-number flows the pressures are quite large compared to the dynamic pressure.

Any flow property in a low-*Re* flow is independent of the fluid density. This means that the lift, drag, and pressure coefficients that were previously defined using the dynamic pressure are actually inappropriate. Nevertheless, in practice the high–Reynolds-number forms are retained, and as a result they always have a Re^{-1} dependence as *Re* approaches zero. For example, as $Re\to 0$ the drag coefficient for a sphere is

$$C_D=\frac{F}{\frac{1}{2}\rho U^2\pi r_0^2}=\frac{24}{Re}. \tag{109}$$

The accuracy of this equation deteriorates if *Re* is greater than 0.5. (Aside: intuitively we would not expect a theory for $\epsilon\to\infty$ to be good unless $\epsilon>$ a few hundred, say 100. A completely equivalent theory in terms of a new parameter $1/\epsilon\equiv\eta\to 0$ would mean $\eta=1/\epsilon<0.01$ or so.)

The fact that linear equations govern low-*Re* flows leads to many theoretical consequences. Consider a problem with $\mathbf{v}$ as a solution. The reverse flow, that is $\mathbf{v}\to-\mathbf{v}$, is also a solution. Furthermore, the stress and pressure change sign; $\boldsymbol{\tau}\to-\boldsymbol{\tau}$ and $p-p_{\text{ref}}\to-(p-p_{\text{ref}})$. Hence, one may argue that a symmetric object must have symmetric streamlines and an antisymmetric pressure distribution. Several conclusions stem from these facts. The requirement for symmetric streamlines means that external flow over an object cannot have a wake. It also requires the flow over a rectangle on a wall to separate both in front of and behind the rectangle; see Fig. 12.

The ability of Stokes equations to produce flow into a corner has been investigated by Dean and Montagnon (1949) and Moffatt (1964). In plane flow smooth solutions are

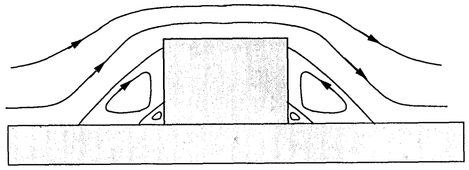

FIG. 12. Low-Reynolds-number flow over a block. Streamlines must be symmetric and so separation takes place on both sides of the block. Low-Reynolds-number flow cannot smoothly turn a 90° corner and this leads to separation. The separation vortex is actually an infinite sequence of vortices of smaller and smaller size.

not possible if the corner angle is less than 146.3°. For smaller angles a vortex must occur in the corner. Hence the pattern of separation in front of and behind the rectangular block is expected. Actually the vortex itself generates a smaller vortex nearer the corner. Theoretically there is an infinite sequence of smaller and smaller vortices (called Moffatt vortices) that are nested in the corner.

15.1 Internal Flows—Lubrication Theory

Internal flows in porous media and in lubrication situations have channels that are very narrow compared to their length. Consider a channel with rigid walls and a slowly varying cross section and curvature. Flow in such a channel is locally like a viscous parallel (Poiseuille) flow driven by a pressure gradient (see flows of Sec. 10.1). The total pressure drop is the integrated value of the local values. When one wall moves and the other wall is stationary, a situation common to bearings, the local velocity profile is a linear combination of a Couette profile and a Poiseuille profile. The local pressure gradient is established by the local channel width and the requirement of a constant flow rate.

The essential hydrodynamic action in lubrication is illustrated by flow in a converging passage with one wall moving. The flow in a slider block, Fig. 13, typifies the physics. The moving wall drags fluid into the passage by viscous forces. Pressure builds up in the center and tends to retard the flow in the front half and aid the flow out the second half. The force that can be supported is indicated by the maximum pressure. It is related to the other parameters of the problem by

$$p_{\max}-p_0=\frac{3\mu UL}{4h_0^2}\,\frac{h_0-h_1}{h_0}\,. \qquad (110)$$

This equation is an approximation for small $(h_0-h_1)/h_0$. In a journal bearing the rotating shaft is eccentric to the journal. This creates a converging passage where high pressures are generated that support the shaft. The fluid is returned to the other side in the diverging section. It has very low pressures and frequently cavitation occurs.

15.2 External Flows

The streaming flow over a solid body at low Reynolds number turns out to be a singular perturbation problem. Stokes's equation, Eq. (105), neglects convective terms like $u\partial u/\partial x$, which must become significant at some large distance from the body. Thus the singularity in the theory is at infinity. In three-dimensional flow it is possible to solve Stokes's equations near the body and to have the solutions match to U_∞ far away. In a certain sense the first-order outer flow is the constant

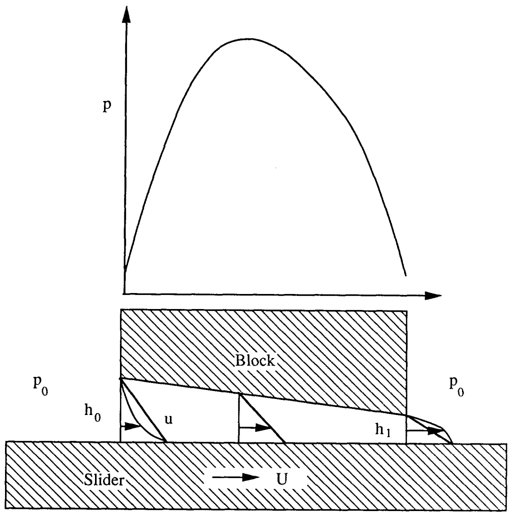

FIG. 13. Slider-block lubrication flow. Viscous forces drag the fluid into the narrowing channel. A pressure gradient is established to retard or accelerate the flow and maintain a constant flow rate. The velocity profiles are a linear Couette component plus the Poiseuille component from the pressure gradient.

U_∞. Thus in three-dimensional problems the singular behavior is not even noticed. For any three-dimensional object it is possible to get a good solution and a drag coefficient. In two dimensions the situation is quite different. Stokes's equation has no solutions that will tend to U_∞, and the far field must be solved and matched to the near-field Stokes flow. An alternative approach is to linearize the problem about U_∞ as formulated by Oseen. He approximated the convective term as $U_\infty \partial u/\partial x$ and could find solutions that cover both regions. The drag on a circular cylinder from an Oseen solution is

$$C_D = \frac{8\pi}{Re[0.5 - \Gamma + \ln(8/Re)]}. \qquad (111)$$

Here Re is based on the diameter and Γ is the Euler constant. From a practical standpoint it is very difficult to produce a low–Reynolds-number flow without end or sidewall effects.

16. STABILITY AND TURBULENCE

The flows discussed above are not necessarily unique solutions. Sometimes solutions are known to exist only for restricted values of the Reynolds number. The question of stability and nonuniqueness is addressed in the article FLUID DYNAMICS, NONEQUILIBRIUM. Many times the flow that exists is a time-dependent flow with a mean profile and fluctuations of 10%–20% called turbulence. For this subject the reader should consult the article TURBULENT FLOWS.

GLOSSARY

Conservation Laws, Local: Differential equations in space and time that govern the properties of matter at each and every point in space.

Conservation Laws, Global: Integral equations that relate the average properties of a region to the events at the surface of the region.

Constitutive Equations: Relations for the molecular fluxes of mass, momentum, or energy in terms of thermodynamic properties and the state of motion of the fluid.

Interfaces: Surfaces where the density is discontinuous because of a change of material or shock or combustion wave.

Jump Conditions: Relations that govern discontinuities of fluid properties at interfaces.

Stability: The growth or decay of a small perturbation that is imagined to exist within the flow. Contrast with "mathematical stability": stability against small changes in boundary conditions.

Turbulence: Flows that have an unsteady fluctuation in the local velocity. The fluctuations must be sustained and have a continuous spectrum.

Viscous Dissipation: The portion of the work of viscous stresses that heats the fluid to increase its internal energy.

Vorticity: Solid-like rotation of a fluid particle.

List of Works Cited

Dahler, J. S., Scriven, L. E. (1961), *Nature* **192**, 36–43.

Dean, W. R., Montagnon, P. E. (1949), *Proc. Cambridge Phil. Soc.* **45**, 389–394.

Escudier, M. (1988), *Prog. Aero. Sci.* **25**, 189–229.

Faler, J. H., Leibovich, S. (1977), *Phys. Fluids* **20**, 1385–1400.

Hoerner, S. F. (1965), *Fluid-Dynamic Drag* (published by the author).

Leibovich, S. (1984), *AIAA J.* **22**, 1192–1206.

Moffatt, H. K. (1964), *J. Fluid Mech.* **18**, 1–18.

Wang, C. Y. (1989), *Appl. Mech. Rev.* **42**, S269.

Wang, C. Y. (1991), *Annu. Rev. Fluid Mech.* **23**, 159–178.

Further Reading

Batchelor, G. K. (1967), *An Introduction to Fluid Dynamics*, Cambridge: Cambridge University Press. (Bible of serious researchers.)

Landau, L. D., Liftshitz, E. M. (1975 reprint) *Fluid Mechanics*, New York: Pergamon. (The physicists' classic.)

Lumley, J. L., Van Dyke, M., Reed, H. L. (eds.)(1991), *Annu. Rev. Fluid Mech.* **23**. (This volume contains an index to Vols. 1–23. Articles on a wide variety of subjects are of a uniformly high quality.)

Panton, R. L. (1984), *Incompressible Flow*, New York: Wiley. (Student text on the fundamental concepts, theoretical structure, and examples of flow phenomena.)

Schlichting, H. (1979), *Boundary Layer Theory,* New York: McGraw-Hill. (A valuable source of experimental data from older references.)

Tritton, D. J. (1977), *Physical Fluid Dynamics,* New York: Van Nostrand-Reinhold. (Fluid dynamics described with a minimum of mathematics.)

White, F. M. (1991), *Viscous Fluid Flow,* New York: McGraw-Hill. (Contains engineering data on boundary layers including heat transfer.)

FLUID DYNAMICS, NONEQUILIBRIUM

RANDALL TAGG, *Department of Physics, University of Colorado at Denver, Denver, Colorado, U.S.A.*

Introduction 437
1. **Mathematical Tools** 439
1.1 Characteristic Scales and Nondimensional Form of the Equations 440
1.2 Analytical Approaches 441
1.2.1 Energy Methods and Global Stability 441
1.2.2 Operator Form of the Disturbance Field Equations 441
1.2.3 Linear Stability 442
1.2.4 Evolution of Arbitrary Disturbances 444
1.2.5 Convective versus Absolute Instabilities 444
1.2.6 Temporal or Spatial Modulation: Floquet Theory 445
1.2.7 Weakly Nonlinear Theory and Amplitude Equations 446
1.2.8 Bifurcation Theory 449
2. **Instabilities** 451
2.1 Waves and Instabilities at Flat Fluid Interfaces 451
2.1.1 Heavy Fluid Above Light Fluid: Rayleigh-Taylor Instability 452
2.1.2 Shocks at Fluid Interfaces: Richtmeyer-Meshkov Instability 453
2.1.3 Parametrically Excited Surface Waves 453
2.1.4 Surface Waves 454
2.2 Capillary Phenomena 455
2.3 Propagating Fronts 456
2.4 Thermal- or Buoyancy-Driven Instabilities 457
2.5 Electrohydrodynamic Instabilities in Liquid Crystals 461
2.6 Centrifugal Instabilities 461
2.6.1 Flow between Rotating Cylinders: The Taylor-Couette System 462
2.6.2 Other Types of Centrifugal Instability 464
2.7 Shear Flow 464
2.7.1 Kelvin-Helmholtz Instability 465
2.7.2 Parallel and Nearly Parallel Flows 466
3. **Dynamical Systems and Chaos Applied to Fluid Dynamics** 469
3.1 Scenarios for Evolution to Chaos 471
3.2 Empirical Identification of Fluid Flows as Dynamical Systems 472
3.3 Measurements of Low-Dimensional Chaotic Systems 473
4. **Summary** 473
Glossary 474
Works Cited 477
Further Reading 479

INTRODUCTION

Fluid matter, under applied stresses, flows and dissipates energy through viscous friction within the fluid and at bounding walls. Thermodynamically, this is already a nonequilibrium system needing external fluxes of energy, momentum, or mass to sustain the fluid motion. However, if the stresses are small enough, the state of the fluid is dominated by dissipative processes and remains close to the equilibrium conditions of the fluid at rest. The velocity field has the same symmetry as the externally imposed conditions. This state of fluid motion is called the thermodynamic branch or the primary flow.

Experience shows that a point is reached where the primary flow is not stable, even though it is in dynamic equilibrium with energy input balancing energy dissipation and

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

applied stresses balancing viscous drag. At the point of instability, internal fluctuations or small external disturbances are selectively amplified and the fluid motion bifurcates (branches) to a new type of motion with different symmetry. As the stresses are increased from zero, the fluid flow may undergo a sequence of such symmetry-breaking bifurcations, or it may suddenly make a transition directly to turbulence. A smooth change in an externally imposed control parameter—for example, mass flux—can be used to bring fluid flows to the point of instability. However, in some flows there is also a spatial evolution from a simple uniform structure to turbulence as the fluid passes downstream from the source. Thus there is a wide range of possible responses of fluid flows to increasing applied stresses.

The hydrodynamic instability of the basic flow, the resulting secondary flows, and the process of transition to turbulence are the subjects of this article. This is an extremely broad field, since instability is the rule rather than the exception in fluid dynamics. It is also a highly interdisciplinary field, involving mathematicians, physicists, engineers, and any other applied scientists working with processes involving fluid motion.

There are many ways in which transitional flows are important to science and technology. First, an important new scientific viewpoint comes from the observation that nonequilibrium systems may spontaneously order themselves into cooperative motion with well-defined patterns and length scales. This process is referred to as self-organization, and the resulting patterns are called dissipative structures. These structures can be a key element of technological devices; an example from optics is the spontaneous amplification and saturation of a single mode in a laser cavity. Perhaps there are hydrodynamic analogs that will also be put to practical use. An important and unifying idea is the slaving principle: certain critical modes arise from fluctuations and entrain all the remaining modes. As a result, the original continuum equations of motion may be reduced in many problems to a finite set of nonlinear ordinary differential equations. The study of such equations—dynamical systems—has been under intense development in recent decades.

Understanding hydrodynamic instabilities also helps clarify the origins and mechanics of turbulence. In many turbulent flows, large-scale structures resemble the secondary flows that arose during the instabilities leading to the turbulent state. Also, turbulence may sustain itself through processes very similar to instabilities, such as through inertial effects at local inflection points within the turbulent velocity field or through instability of a region of vorticity to stretching and reorientation. Finally, turbulent flows are governed by the same deterministic equations that describe transitional flows; thus there may be some relation between the occurrence of extreme disorder in turbulent flows and the observation of chaotic dynamics in some of the simpler flows arising from instability.

Technological applications often seek either to induce or to delay instabilities and the onset of turbulence. In heat-transfer or mixing processes, secondary flows and turbulence are desirable because they improve transport. Similarly, the onset of turbulence in airfoil boundary layers may prevent boundary-layer separation that could adversely affect the airfoil lift. On the other hand, if both separation and turbulent boundary layers could be prevented, the airfoil would have improved lift-to-drag characteristics; hence there has been considerable research into the design of laminar airfoils in which transition is suppressed by geometry, suction, etc. In other applications (such as ventilation) it is also desirable to maintain steady uniform flows because of the higher energy costs associated with the more dissipative turbulent flows; such flows are also likely to cause harmful vibrations. Still other applications seek a compromise whereby secondary flows are created to increase transport but further transition is arrested in order to minimize energy consumption. Finally, flows on the verge of transition might be useful as amplifiers of weak disturbances.

An essential feature of nonequilibrium flows is the nonlinearity of the equations of motion. An important consequence of nonlinearity is the nonuniqueness of states of motion satisfying the equations and boundary conditions. Some, but not all, such states may be stable; which one of a multiplicity of stable solutions is selected depends on the history by which the flow state was prepared. This is in contrast to linear problems, which respond to imposed conditions in a unique way. It is now also known that a system of only three coupled nonlinear ordinary differential equations can produce motion that is chaotic; that is, the motion is so sensitive to initial conditions that predictability decays exponentially with time.

The study of nonequilibrium systems advances through close interaction of precise experiments (often under computer control), mathematical analysis, and large-scale numerical simulations. While many important theoretical models have been developed to understand transitional flows, experiments identify both the successes and limitations of these models in describing real processes. Often the bridge between experiments and the mathematical models is made using accurate numerical simulations. Simulations, in turn, serve as exploratory tools when either theory or experiment is intractable. Unfortunately, many problems in transitional flows involve ranges of space and time scales that are still beyond the capabilities of computers. Over all, the subject remains full of unanswered questions, and the theory is undergoing rapid growth. Thus it is helpful to approach this subject with a broad overview of the basic instability mechanisms and of the observational facts.

In the following, we start with a mathematical primer in order to create a point of view for examining the variety of instability phenomena. This point of view is tentative; for example, we are still far from a complete understanding of transition in shear flows and in wakes. The next part of the article then discusses several important and physically distinct types of instabilities, including a description of the geometries and relevant parameters, an outline of the basic theory, and a summary of experimental observations. Notation conforms to the usage found in the literature and may vary somewhat from one type of flow to another. The article continues with a discussion of the application of the theory of chaotic dynamical systems to nonequilibrium flows. It concludes with a guide to further reading. Instabilities in compressible, magnetohydrodynamic, and non-Newtonian fluids have been excluded (see COMPRESSIBLE FLOWS; MAGNETOHYDRODYNAMICS). Also, the important subjects of experimental and numerical methods are beyond the scope of this article, although some references are given in the guide to further reading.

1. MATHEMATICAL TOOLS

The starting point for analysis of nonequilibrium flows is the set of mass, momentum (Navier-Stokes), and energy conservation equations for an incompressible fluid. The fluid is assumed to be a pure fluid in local thermodynamic equilibrium obeying Newton's law relating viscous stress to rate of strain and Fourier's law of heat conduction. For the origin of these equations and the generalization to a wider class of fluids (for example, compressible fluids), see FLUID DYNAMICS, EQUILIBRIUM.

The independent variables are position

$$\mathbf{r}=(x_1,x_2,x_3)=(x,y,z) \text{ with Cartesian}$$

$$\text{unit vectors } \hat{\mathbf{x}},\hat{\mathbf{y}},\hat{\mathbf{z}}$$

$$=(r,\theta,z) \text{ with cylindrical polar}$$

$$\text{unit vectors } \hat{\mathbf{r}},\hat{\boldsymbol{\theta}},\hat{\mathbf{z}},$$

and time t. The fields are velocity

$$\mathbf{u}=(u_1,u_2,u_3)=(u,v,w),$$

vorticity

$$\omega\equiv\nabla\times\mathbf{u},$$

pressure p, temperature T, density ρ, concentration c. The thermodynamic coefficients for various fluid properties are kinematic viscosity ν ($\mathrm{m^2\,s^{-1}}$), specific heat c_p ($\mathrm{J\,kg^{-1}\,K^{-1}}$) (at constant pressure), thermal conductivity k_{th} ($\mathrm{W\,m^{-1}\,K^{-1}}$), thermal diffusivity $\kappa\equiv k_{\mathrm{th}}/\rho c_p$ ($\mathrm{m^2\,s^{-1}}$), diffusion coefficient κ_c ($\mathrm{m^2\,s^{-1}}$), surface tension γ ($\mathrm{N\,m^{-1}}$), and coefficient of expansion

$$\alpha\equiv-\frac{1}{\rho}\left(\frac{\partial\rho}{\partial T}\right)_p .$$

A body force per unit mass $-g\hat{\mathbf{z}}$ acts due to a gravitational field with local acceleration g in the downward vertical direction $-\hat{\mathbf{z}}$. With these definitions (but omitting concentration-dependent terms), the equations are

$$\nabla\cdot\mathbf{u}=0, \tag{1a}$$

$$\frac{\partial\mathbf{u}}{\partial t}+(\mathbf{u}\cdot\nabla)\mathbf{u}=-\frac{1}{\rho}\nabla p+\nu\nabla^2\mathbf{u}-g\hat{\mathbf{z}}, \tag{1b}$$

$$\frac{\partial T}{\partial t}+(\mathbf{u}\cdot\nabla)T=\kappa\nabla^2T+\boldsymbol{\tau}':\nabla\mathbf{u}+\frac{\alpha}{\rho c_p}\left(\frac{\partial p}{\partial t}+(\mathbf{u}\cdot\nabla)p\right), \tag{1c}$$

$$\rho=\rho(T), \tag{1d}$$

$$\Delta p=\gamma\nabla\cdot\hat{\mathbf{n}}, \tag{1e}$$

where $\tau'_{ij}\equiv\nu(\partial u_i/\partial x_j+\partial u_j/\partial x_i)$ is the viscous part of the stress tensor for a Newtonian fluid and Δp is the jump in pressure entering a fluid whose interface has outward normal $\hat{\mathbf{n}}$. The nonlinearity of these equations is evident in the advective terms $(\mathbf{u}\cdot\nabla)\mathbf{u}$, $(\mathbf{u}\cdot\nabla)T$, and $(\mathbf{u}\cdot\nabla)p$ and the dissipative term (often ignored) $\boldsymbol{\tau}':\nabla\mathbf{u}$. Further nonlinearity may arise from the boundary conditions, typically in the case of deformable surfaces. Indeed, the problem is not completely posed until initial conditions and boundary conditions are set; these will be discussed in Sec. 2 below for each particular type of instability.

1.1 Characteristic Scales and Nondimensional Form of the Equations

It is very useful, though sometimes tricky, to identify the intrinsic scales of the problem. There is typically a length scale d over which the gradients are originally imposed (e.g., the gap between two parallel plates held at different temperatures in a thermal convection experiment, or the boundary-layer thickness). The other length scales of the system may be referred to this length scale; for example, experiments are confined to a dimension L in a direction where the theory idealizes an infinite extent. The dimensionless length $\Gamma\equiv L/d$ is called the aspect ratio for that direction. Systems with large aspect ratio are almost translationally invariant in the long direction and allow (approximately) a separation of spatial variables. Small–aspect-ratio systems usually preclude a separation of the spatial variables; however, they may be dynamically simpler in the sense that they permit only a small finite number of active modes.

One of the most important dimensionless constants applied to the study of transition flows is the Reynolds number Re. Assuming a characteristic velocity of magnitude U (for example, the free-stream velocity in boundary layer flow or the center-line velocity in flow through ducts), we have

$$Re\equiv Ud/\nu. \tag{2}$$

This is often interpreted as the ratio of "inertial forces" to "viscous forces." An alternative interpretation is that $Re=\tau_a/\tau_\nu$. Here τ_a is the advective time scale U/d, the time for a fluid element with characteristic speed U to transport momentum across the characteristic length d. The viscous time scale $\tau_\nu\equiv d^2/\nu$ measures the time scale for momentum to diffuse across the same length d.

New space and time scales may enter the problem after flow transition occurs. This applies especially to secondary-flow patterns, in which there is a length $2\pi/k$ associated with a mode with wave vector k (usually this scale is set by d) and a correlation length ξ that describes the distance over which the organized motion remains correlated. There is also a time scale, often quite long, for disruptions of a pattern to diffuse through the system. In flows that evolve spatially into fully turbulent motion, there are empirical space and time scales associated with the streamwise development; however, these scales await a systematic theoretical description.

Typically, the first step of analysis is to convert the governing equations to nondimensional form by expressing dependent and independent variables in terms of the intrinsic scales. Results may then be compared from experiments with similar geometries but different sizes and fluids. Also, the relative importance of terms may be inferred; this must be done with caution, however, since solutions can have singular dependence on small parameters over subdomains of the flow. An example is boundary layers, in which there are large gradients and viscous terms remain important even when the viscosity is very small.

To illustrate a nondimensional form of the Navier-Stokes equation, let $\mathbf{r}'=\mathbf{r}/d$, $t'=t/\tau_a=tU/d$, $\mathbf{u}'=\mathbf{u}/U$, and $p'=p/\rho U^2$. Then

$$\frac{\partial \mathbf{u}'}{\partial t'}+(\mathbf{u}\cdot\nabla)\mathbf{u}'=-\nabla p'+\frac{1}{Re}\nabla^2\mathbf{u}'-\frac{1}{Fr^2}\hat{\mathbf{z}}. \tag{3}$$

Here the additional dimensionless parameter $Fr\equiv U/(gd)^{1/2}$ (Froude number) measures the relative importance of inertial and gravitational terms and is important in free-surface problems. Note that since g is fixed and the range of kinematic viscosities ν for common working fluids is limited, it is usually not possible to match both Reynolds number and Froude number in comparing the motion of real ships to that of ship models. Such a match might be achieved, however, in some modeling experiments by using liquid helium.

There is an ambiguity in the choice of time scale: one might have chosen $t''=t/\tau_\nu=t\nu/d^2=(1/Re)t'$. Indeed, sometimes this is more appropriate since the growth and spreading of disturbances near the first instability is characterized by a viscous time scale. The choice of scales is seldom unique; furthermore, many problems (for example, boundary-layer problems) require a separation into spatial regions or time domains governed by completely different scales.

Other dimensionless parameters will arise in the specific instability problems discussed in Sec. 2 below.

1.2 Analytical Approaches

This section is rather formal and the reader may wish to look briefly first at particular examples of instabilities in Sec. 2. The following material is, however, an important primer for much of the current research literature on instability phenomena.

1.2.1 Energy Methods and Global Stability. A general argument shows that the solution corresponding to the thermodynamic branch is unique when the external stresses are sufficiently small. In this case, if a fluid system is started from an arbitrary initial state consistent with the governing equations (for example, the initial velocity field is divergence-free), then there will be a transient motion that ultimately settles into the unique primary flow. Similar arguments, based on energy considerations or on various integral forms of the governing equations, can even set bounds on the applied stresses beneath which the primary flow remains stable. Little is assumed about the nature of the stresses, so that the theory shows when the system is globally stable to all perturbations. Sometimes these limits correspond to the observed stability limits, but often they are too conservative: the flow may remain stable well beyond the established limits. Nonetheless, these techniques yield valuable insights into the physical processes involved in flow stability; see Joseph (1976a,b, 1985) for further discussion.

1.2.2 Operator Form of the Disturbance Field Equations. Consider the total velocity and pressure fields as the sum of the base fields $(\mathbf{U},P)$ and disturbance fields $(\mathbf{u},p)$. The base fields* separately satisfy the equations of motion and boundary conditions. In the following, we assume fixed boundaries (see Secs. 2.1–2.3 for problems with moving boundaries); also, we ignore thermal effects (see Sec. 2.4). The resulting equations for the disturbances are

$$\nabla\cdot\mathbf{u}=0, \tag{4a}$$

$$\frac{\partial \mathbf{u}}{\partial t}+(\mathbf{U}\cdot\nabla)\mathbf{u}+(\mathbf{u}\cdot\nabla)\mathbf{U}+(\mathbf{u}\cdot\nabla)\mathbf{u}$$
$$=-\nabla p+\frac{1}{Re}\nabla^2\mathbf{u}, \tag{4b}$$

with homogeneous conditions on the boundary δV,

$$\mathbf{u}|_{\delta V}=\mathbf{0},\quad p|_{\delta V}=0. \tag{4c}$$

This system may be rewritten in operator form as

$$\mathscr{L}\Psi+\mathscr{N}(\Psi,\Psi)=\mathscr{M}\frac{\partial\Psi}{\partial t}, \tag{5a}$$

where

$$\Psi\equiv\begin{pmatrix}u_x\\u_y\\u_z\\p\end{pmatrix}, \tag{5b}$$

---

*The base field need not be the primary flow (thermodynamic branch). We could be linearizing about a secondary flow to find *its* stability properties.

$$\mathscr{L}\equiv\begin{pmatrix} \frac{1}{Re}\nabla^2-\mathbf{U}\cdot\nabla-\frac{\partial U_x}{\partial x} & -\frac{\partial U_x}{\partial y} & -\frac{\partial U_x}{\partial z} & -\frac{\partial}{\partial x} \\ -\frac{\partial U_y}{\partial x} & \frac{1}{Re}\nabla^2-\mathbf{U}\cdot\nabla-\frac{\partial U_y}{\partial y} & -\frac{\partial U_y}{\partial z} & -\frac{\partial}{\partial y} \\ \frac{\partial U_z}{\partial x} & \frac{\partial U_z}{\partial y} & \frac{1}{Re}\nabla^2-\mathbf{U}\cdot\nabla-\frac{\partial U_z}{\partial z} & -\frac{\partial}{\partial z} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} & 0 \end{pmatrix}. \qquad (6a)$$

$$\mathscr{M}\equiv\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad (6b)$$

and the nonlinear term is

$$\mathscr{N}(\Psi^{(1)},\Psi^{(2)})=-\begin{pmatrix} \mathbf{u}^{(1)}\cdot\nabla u_x^{(2)} \\ \mathbf{u}^{(1)}\cdot\nabla u_y^{(2)} \\ \mathbf{u}^{(1)}\cdot\nabla u_z^{(2)} \\ 0 \end{pmatrix}. \qquad (6c)$$

Note further that $\mathscr{N}$ is a bilinear operator:

$$\begin{aligned}&\mathscr{N}(\Psi^{(1)}+\Psi^{(2)},\Psi^{(3)}+\Psi^{(4)})\\ &\quad=\mathscr{N}(\Psi^{(1)},\Psi^{(3)})+\mathscr{N}(\Psi^{(1)},\Psi^{(4)})\\ &\qquad+\mathscr{N}(\Psi^{(2)},\Psi^{(3)})+\mathscr{N}(\Psi^{(2)},\Psi^{(4)}), \qquad (7a)\end{aligned}$$

$$\mathscr{N}(a\Psi^{(1)},b\Psi^{(2)})=ab\mathscr{N}(\Psi^{(1)},\psi^{(2)}). \qquad (7b)$$

(Nonlinear terms arising from the kinematical conditions of free-surface problems do not have this property.)

1.2.3 Linear Stability. An important first step in analyzing the system equation (5a) is to neglect the nonlinear term $\mathscr{N}(\Psi,\Psi)$ and solve the linear problem

$$\mathscr{L}\Psi=\mathscr{M}\frac{\partial\Psi}{\partial t}, \qquad (8a)$$

subject to the same boundary conditions

$$\mathbf{u}|_{\delta V}=\mathbf{0},\quad p|_{\delta V}=0. \qquad (8b)$$

Typically, systems are idealized as unbounded in one or two directions, say x and y. For example, fluid channels are presumed infinitely long and infinitely wide (but they have finite depth). There are two choices in completing the statement of the problem:

1. Temporal instability: the linear system is considered an initial-value problem with a specified initial disturbance $\Psi(x,y,z,t=0)$. The initial disturbance and subsequent solution are constrained to be bounded functions of the spatial variables, but the solution may, in the event of instability, become unbounded in time.*
2. Spatial instability: there is a prescribed harmonic forcing at some surface, say $x=0$: $\Psi(x=0,y,z,t)=\Psi(y,z)e^{i\omega_0 t}$. (Some specification of the spatial derivatives of Ψ at $x=0$ is also necessary.) The solution is presumed to also be harmonic in time but may become unbounded in space, say as $x\to\infty$.

According to the method of normal modes, one ignores particular initial or forcing conditions and looks for a complete set of linearly independent solutions to the problem posed by Eq. (8) and its boundary conditions. Arbitrary disturbances can then be represented by series expansions in the normal modes (see Sec. 1.2.4). If any of these normal modes is found to grow exponentially in time or space (depending on how the problem is posed), the base flow is deemed to be unstable.

If the base flow $\mathbf{U}$ is steady, then the operator $\mathscr{L}$ (which contains $\mathbf{U}$) is time independent, and so we can expect a separation of the normal-mode solutions into the form

$$\Psi(x,y,z,t)=\Psi_n(x,y,z)\exp(s_n t), \qquad (9)$$

where Ψ_n then satisfies the generalized eigenvalue equation

$$\mathscr{L}\Psi_n=s_n\mathscr{M}\Psi_n. \qquad (10)$$

---

*In inviscid flow theory, the possibility has arisen for there to be divergence, in finite time, of the velocity in some local region of space.

The index n shows that there are many modes consistent with the boundary conditions, as in the analogous example of vibrations of a stretched string. In most problems studied so far, the base flow only depends on one coordinate, say z, so that the operator $\mathscr{L}$ is also independent of x and y. In this case we may further expect the normal-mode solutions to be of the form

$$\Psi_n(x,y,z)=\Psi_{nk_xk_y}(z)\exp(ik_xx+ik_yy). \tag{11a}$$

In systems with cylindrical symmetry, $\mathscr{L}$ is independent of z and θ and the solutions must be strictly periodic in θ. We then expect solutions of form

$$\Psi_n(r,\theta,z)=\Psi_{nmk_z}(r)\exp(ik_zz+im\theta). \tag{11b}$$

Clearly, there are many practical problems (for example, swept wings, complex turbine geometries, etc) that are intrinsically three dimensional, so that the above separations are not possible. These are much less tractable for analysis and numerical methods must be employed; however, it is often best to first consider two-dimensional limiting cases.

Thus the linear stability problem often reduces to the form

$$\mathscr{L}(k_x,k_y,z;Re)\Psi_{nk_xk_y}=s_n(k_x,k_y;Re)\mathscr{M}\Psi_{nk_xk_y}, \tag{12}$$

where the dependence of the linear operator on (k_x,k_y,z,Re) has been made explicit and the k_y,k_y dependence of the eigenvalues $s=s_n(k_x,k_y;Re)$ constitutes a dispersion relation. Note that we may think of the index n as describing different "bands" of dispersion relations $s=s_n(k_x,k_y;Re)$, because there are infinitely many solutions $\Psi_{nk_xk_y}(z)$ even when $(k_x,k_y;Re)$ are fixed.

If the problem is posed as a temporal instability problem, the wave numbers k_x and k_y must be real so that the solution remains a bounded function of the spatial variables. However, the eigenvalue s may be taken to be complex:

$$s_n(k_x,k_y;Re)=\sigma_n(k_x,k_y;Re)+i\omega_n(k_x,k_y;Re). \tag{13}$$

If as the Reynolds number Re is increased, $\sigma_n(k_x,k_y;Re)$ passes from a negative value to a positive value, we have instability to a temporally growing mode, that is, a mode whose magnitude grows as $\exp(\sigma_nt)$. The value of Re where this crossover occurs is given by the marginal stability condition:

$$\sigma_n(k_x,k_y;Re)=0. \tag{14}$$

This implicitly defines a relationship $Re=f_n(k_x,k_y)$, and the minimum Re over all choices of n,k_x,k_y establishes a critical Reynolds number Re_c and critical wave numbers k_{xc},k_{yc}. Usually, the index n is ordered so that $n=1$ corresponds to the least stable band. Often, the symmetry is such that only the magnitude $k\equiv(k_x^2+k_y^2)^{1/2}$ is important; in other cases [especially parallel shear flows with a base flow $U(y)$ in the x direction], it can be shown that instability occurs soonest for two-dimensional modes so that we can let $k_y=0$ and define $k\equiv k_x$. Assuming $n=1$ for the most unstable band, the marginal stability condition then reduces to a marginal curve $Re=f(k)$ whose minimum defines the critical values (k_c,Re_c)—see Fig. 1(a). Evaluating the imaginary part of s at these critical values gives a critical frequency $\omega_c\equiv\omega_1(Re_c,k_c)$. If this frequency $\omega_c=0$, the instability has historically been called an exchange of stabilities: one steady flow is replaced by another steady flow. If instead $\omega_c\neq0$, this is usually called a Hopf bifurcation: a steady flow changes to a time-periodic flow.

When the problem is posed as a spatial stability problem, the eigenvalue s is con-

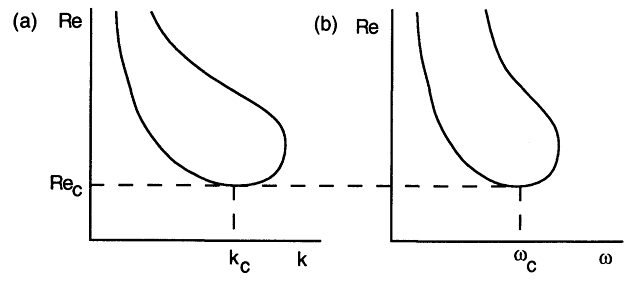

FIG. 1. Marginal stability curves (a) for temporal instability (b) for spatial instability.

strained to be purely imaginary and equal to the driving frequency:

$$s_n(k_x,k_y;Re)=i\omega_0. \tag{15}$$

This implicitly defines a functional dependence of a wave number, now complex, on the other parameters:

$$k_x(\omega_0,Re,k_y)=k_{xr}(\omega_0,Re,k_y)+ik_{xi}(\omega_0,Re,k_y). \tag{16}$$

The wave number k_y is usually set equal to zero or to some real value: the time-harmonic forcing at $x=0$ is either uniform or spatially periodic in the y direction. If, as the Reynolds number is increased, the imaginary part $k_{xi}(\omega_0,Re)$ goes from positive to negative, then we have instability to a spatially growing mode whose magnitude grows as $\exp(k_{xi}x)$. The marginal stability condition is now written as

$$k_{xi}(\omega_0,Re,k_y)=0. \tag{17}$$

Assuming that $k_y=0$, this gives a marginal curve

$$Re=Re(\omega_0), \tag{18}$$

whose minimum defines the critical values (ω_c,Re_c): see Fig. 1(b). The real part of the critical wave number is then determined by $k_c\equiv k_{xr}(\omega_c,Re_c)$.

Usually the temporal stability problem is much simpler, since the linear problem, Eq. (12), has a much more complicated dependence on k than on s. It is important to note that, regardless of whether the problem is posed as one of temporal instability or spatial instability, there is a unique critical Reynolds number Re_c where linear stability theory predicts transition to occur. However, above the critical Re_c one may be interested in knowing how the temporal and spatial growth rates are related. An approximate relation is given by Gaster's transformation (Gaster, 1962):

$$k_i\approx-\sigma/v_{gc}, \tag{19a}$$

where v_{gc} is the group velocity of the critical eigenmode:

$$v_{gc}\equiv\frac{\partial\omega}{\partial k}\quad\text{evaluated at } Re_c,k_c. \tag{19b}$$

It is an intriguing aspect of study of hydrodynamic instabilities that for some flows (for example, flow between rotating cylinders) the first transition occurs very close to the predicted critical Reynolds number, but in other flows (for example, boundary-layer instability) linear stability theory overestimates the threshold for transition. The latter situation is understood to result from the susceptibility of the flow to finite, rather than infinitesimal, disturbances; however, the process of transition in this case remains incompletely understood.

1.2.4 Evolution of Arbitrary Disturbances. Of course, a real experiment cannot just launch a single normal mode into the flow to see if it grows. Thus the normal-mode approach must be related to the practical situation of more general initial disturbances or more general forcing conditions. Suppose the domain is bounded in the transverse coordinate z, as in channel flow or in thermal convection in horizontal layers. (In other problems, like pipe flow or flow between cylinders, replace z with r). Then the spectrum of eigenvalues $s_n(k_x,k_y;Re)$ for given k_x and k_y is discrete and the corresponding set of eigenfunctions is complete. An arbitrary disturbance then evolves as a superposition of normal modes:

$$\Psi(x,y,z,t)=\int_{-\infty}^{\infty}dk_x\int_{-\infty}^{\infty}dk_y\sum_{n=1}^{\infty}A_n(k_x,k_y)\times\Psi_{nk_xk_y}(z)e^{i(k_xx+k_yy)}e^{s_n(k_x,k_y;Re)t}, \tag{20}$$

where the mode amplitudes $A_n(k_x,k_y)$ are determined by the initial condition $\Psi(x,y,z,t=0)$.

For flows with at least one boundary at infinity, such as boundary layers and free shear layers, the spectrum will contain, for fixed k_x and k_y, a *finite* number of discrete eigenvalues $s_n(k_xk_y;Re)$, and the corresponding set of eigenfunctions cannot be complete. There are various ways of dealing with this situation, such as by solving the linearized equations with Fourier-Laplace transforms (Huerre and Monkewitz, 1990) or by admitting a wider class of basis functions subject to the generalized boundary conditions that disturbances $\mathbf{u},p$ remain bounded (rather than vanish) as $z\to\infty$ (Grosch and Salwen, 1978; Salwen and Grosch, 1981).

1.2.5 Convective versus Absolute Instabilities. What happens, within the scope of the linearized theory, to a local initial disturbance as it propagates in space and time? This question leads to an important distinction between types of instabilities. Neglecting nonlinear effects, a local disturbance will generate a wave packet that travels with the group velocity of the fastest-growing mode. If the disturbance travels away from the origin fast enough that the flow at the origin returns to the unperturbed state, even though the disturbance itself continues to grow as it moves [Fig. 2(a)], then the flow is said to undergo a convective instability. If the growth and spreading are sufficiently rapid that the flow at the origin is permanently affected, even though the center of the disturbance moves off in the direction of the group velocity [Fig. 2(b)], then the flow is said to undergo an absolute instability.

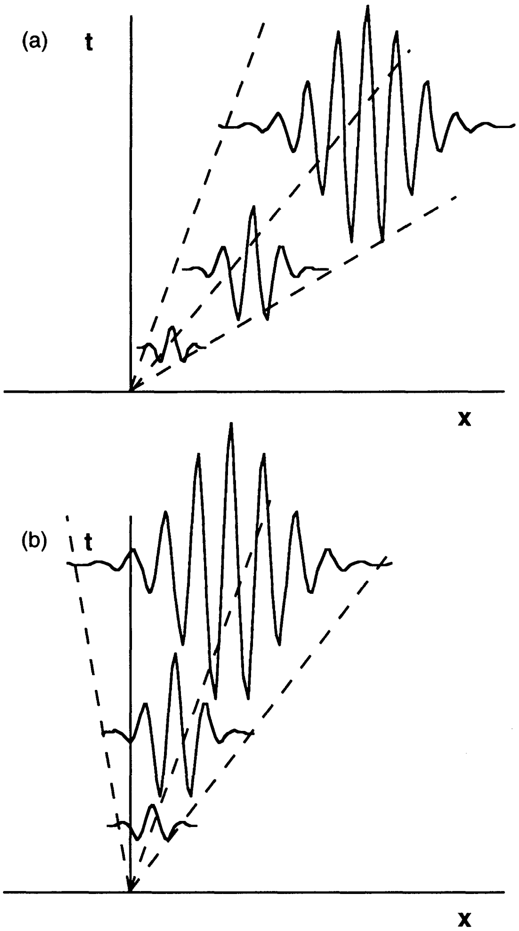

FIG. 2. Space–time diagram showing propagation of a local disturbance in (a) convectively unstable flow; (b) absolutely unstable flow.

Many open flows, in which there is a steady unidirectional mass flow through the system, are likely to be convectively unstable. If the length of the experimental system is too short, the instability might be overlooked because disturbances are washed out of the system before they are detected. On the other hand, even if the disturbances are detected, the resulting secondary flow will be erratic in phase and amplitude because it is driven by random fluctuations upstream of the detector. Power spectra of the velocity fluctuations will show broad peaks even though the linear-stability theory indicates a well-defined frequency at onset of the instability. In contrast, the secondary mode in an absolutely unstable flow becomes coherent (assuming no further instabilities) and exhibits a well-defined frequency (sharp peaks in a power spectrum). Note that one way to overcome the randomness in a convectively unstable flow is to force the flow with a controlled disturbance stronger than any fluctuations or noise in the system.

While convective instabilities are associated with open flows, they can occur in closed systems that produce traveling-wave modes with nonzero group velocity. Examples are binary-fluid convection and flow between counterrotating cylinders. If some means of feedback is provided—even in an open flow, for example, by having the flow impinge on a downstream obstacle such that pressure disturbances are radiated back upstream—the entire flow domain can still become filled with a coherent secondary flow. This is sometimes referred to as a global instability. Feedback provided by the end walls of convectively unstable closed systems can produce very interesting large length- and time-scale dynamics. An example is the "blinking state" in binary-fluid convection, where the waves slowly switch back and forth in time from left-going to right-going.

1.2.6 Temporal or Spatial Modulation: Floquet Theory. In some cases the base flow may be periodic in space and/or time. This may be due either to a physical modulation of stresses that drive the flow (Davis, 1976) or to the fact that a well-established secondary flow is considered as the base flow and we wish to consider *its* stability relative to further changes in the flow patterns.

If the base flow is time periodic with period T, then the Floquet theory of linear differential equations with periodic coefficients (Iooss and Joseph, 1990) leads to a modified eigenfunction of the form

$$\Psi=\Psi_n(t)\exp(s_n t), \tag{21a}$$

where

$$\Psi_n(t+T)=\Psi_n(t), \tag{21b}$$

and s_n, called the Floquet exponent, has an imaginary part within the bounds

$$-\pi/T\leqslant\omega_n<\pi/T. \tag{21c}$$

If the base flow is *spatially* periodic in the x direction with period Λ (but presumed homogeneous in the y direction), then the eigenfunction is modified to

$$\Psi_n=\Psi_{nk_xk_y}(x,z)\exp(ik_x x+ik_y y) \tag{22a}$$

with a spatially periodic part

$$\Psi_{nk_xk_y}(x+\Lambda,z)=\Psi_{nk_xk_y}(x,z), \tag{22b}$$

and the real part of the spatial Floquet exponent is within the bounds

$$-\pi/\Lambda\leqslant k_{xr}<\pi/\Lambda. \tag{22c}$$

(By analogy with solid-state physics, we call this range the "first Brillouin zone" and the Floquet theory for the form of Ψ_n given above is also known as Bloch's theorem.)

A special case of Floquet theory, applied to the parametric instability of surface waves, is discussed in Sec. 2.1.3.

1.2.7 Weakly Nonlinear Theory and Amplitude Equations. The linear-stability calculations, though difficult enough, do not determine the form of the flow after transition occurs. Disturbances cannot continue to grow exponentially; because of the nonlinearity of the equations, the base flow will be altered and higher harmonics of the disturbing mode will be produced through self-interaction. Perhaps this will lead to a saturation of the growing mode into a state that resembles the critical eigenfunction. If this new saturated secondary flow is itself stable and grows continuously from zero amplitude as the Reynolds number is increased, we say there is a supercritical bifurcation, and perturbation theory may be used to find approximations to the secondary flow.

The theory requires identification of a small parameter, typically $\epsilon\equiv(Re-Re_c)/Re_c$. We seek an equation for the complex amplitude $A(t;\epsilon)$, presumed small, of the eigenmode $\Psi_{1k_xk_y}(z)\exp(ik_x x+ik_y y-i\omega_c t)$ that has marginally destabilized the base flow at $Re=Re_c$. Landau (1944), with his usual heuristic brilliance, proposed an equation of the form

$$\frac{dA}{dt}=\sigma_0(1+ic_0)\epsilon A-(g_r+ig_i)|A|^2A, \tag{23}$$

the Landau equation. (Landau's formulation has been adapted to the notation of this article; also, many authors use τ_0^{-1} in place of σ_0.)

Several comments are in order. First, the linearized equation

$$\frac{dA}{dt}=\sigma_0(1+ic_0)\epsilon A \tag{24}$$

has eigenvalue $\sigma_0(1+ic_0)\epsilon$, which is just the first-order term in an expansion of the eigenvalue s in the parameter ϵ:

$$s=i\omega_c+\sigma_0(1+ic_0)\epsilon+O(\epsilon^2). \tag{25}$$

Considering this as a Taylor-series expansion, we identify

$$\sigma_0(1+ic_0)=\frac{\partial s}{\partial\epsilon}=Re_c\frac{\partial s}{\partial Re}\bigg|_{Re=Re_c}. \tag{26}$$

A second and very important comment is that this is the only form of the equation up to cubic order in A that is consistent with translational invariance in x and y. In another coordinate frame with coordinates $(x',y')=(x-x_0,y-y_0)$, the amplitude is $A'=A\exp(ik_x x_0+ik_y y_0)$. The physics must be the same, and so the same amplitude equation must be obtained for A'. Thus any phase shifts in A must cancel from the amplitude equation, as they do for the Landau form. We say that the amplitude equation is equivariant under translations in the (x,y) plane.

Finally, the complex Landau constant $g=g_r+ig_i$ describes the nonlinear saturation of the amplitude and the finite-amplitude corrections to the frequency ω. Before showing this, we note that computation of the Landau constant depends on the particular flow and requires a detailed perturbation analysis. The numerical value of g depends on the normalization chosen for the eigenfunction Ψ; unfortunately, authors sometimes overlook the important requirement to state this normalization when quoting a value for g.

In order to solve the Landau equation, write $A=\rho e^{i\phi}$. Then two equations result:

$$\frac{d\rho}{dt}=\sigma_0\epsilon\rho-g_r\rho^3, \tag{27a}$$

$$\frac{d\phi}{dt}=\sigma_0 c_0\epsilon-g_i\rho^2. \tag{27b}$$

Looking for solutions with ρ steady in time, we get, for $g_r>0$,

$\epsilon<0$: $\rho=0$, $\phi=\sigma_0 c_0\epsilon t$, stable;

$\epsilon>0$: $\rho=0$, $\phi=\sigma_0 c_0\epsilon t$, unstable,

$\rho=(\sigma_0\epsilon/g_r)^{1/2}$, $\phi=\sigma_0(c_0-g_i/g_r)\epsilon t$, stable;

for $g_r<0$,

$\epsilon<0$: $\rho=0$, $\phi=\sigma_0 c_0\epsilon t$, stable,

$\rho=(-\sigma_0\epsilon/g_r)^{1/2}$, $\phi=\sigma_0(c_0-g_i/g_r)\epsilon t$, unstable;

$\epsilon>0$: $\rho=0$, $\phi=\sigma_0 c_0\epsilon t$, unstable.

Note that Eq. (27a) for the magnitude ρ is decoupled from the phase equation. However, Eq. (27b) for the phase ϕ retains a dependence on ρ through the nonlinear term $-g_i\rho^2$. The solutions for ϕ may be interpreted as corrections to the frequency: $\omega=\omega_c+d\phi/dt$. Thus for $\epsilon>0$ in the supercritical case ($g_r>0$), the frequency is altered by a linear correction $\sigma_0 c_0\epsilon$ and a leading-order nonlinear correction $-g_i\rho^2=-\sigma_0(g_i/g_r)\epsilon$. The solutions for ρ and their stability are shown in the bifurcation diagrams of Fig. 3. Note the characteristic $\epsilon^{1/2}$ dependence of the bifurcating solutions. We see that the sign of the real part of the Landau constant is crucial. For positive g_r there is a stable supercritical mode bifurcating from the base state. For negative g_r, there is subcritical bifurcation to a solution that is unstable, and higher-order terms in the amplitude equation are needed to resolve what happened for $\epsilon>0$.

In fact, an exact time-dependent solution may be found for Eq. (27a) given an initial nonzero disturbance ρ_0:

$$\rho^2=\frac{\sigma_0\epsilon}{g_r}\,\frac{1}{1+(\sigma_0\epsilon/g_r\rho_0^2-1)e^{-2\sigma_0\epsilon t}}. \tag{29}$$

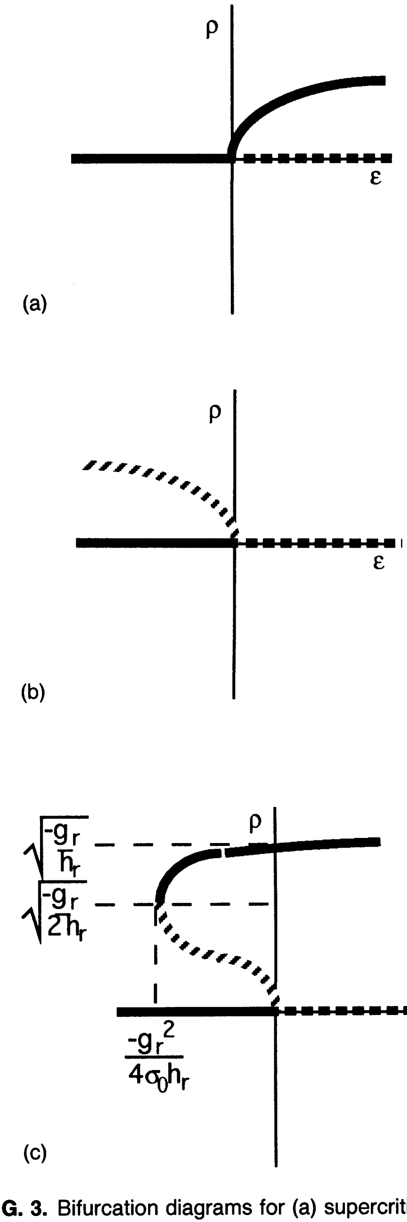
ρ
ε
(a)
ρ
ε
(b)
$\sqrt{\frac{-g_r}{h_r}}$
$\sqrt{\frac{-g_r}{2h_r}}$
ρ
$\frac{-g_r^2}{4\sigma_0 h_r}$
ε
(c)

FIG. 3. Bifurcation diagrams for (a) supercritical bifurcation, (b) subcritical bifurcation, (c) subcritical bifurcation combined with turning point, resulting from a fifth-order term in the amplitude equation.

There have been beautiful experiments in several systems showing how the time-dependent growth ($\epsilon>0$) or decay ($\epsilon<0$) conforms to this equation. As an example from wake flow, see Provansal *et al.* (1987). Furthermore, the $\epsilon^{1/2}$ growth predicted by the Landau equation has been shown in precise experiments for several systems, including thermal convection and Taylor-Couette flow between rotating cylinders. For critical discussion of some of these experiments, see Ahlers (1989).

An amplitude equation with a fifth-order term $(h_r+ih_i)|A|^4A$ with $g_r<0$ but $h_r>0$ gives the bifurcation diagram in Fig. 3(c). The overlap of two stable solutions for $-g_r^2/4\sigma_0 h_r<\epsilon<0$ is experimentally observed as hysteresis: a carefully done experiment will observe the base flow while increasing ϵ up to $\epsilon=0$ but, upon decreasing ϵ, not recover the base flow until the lower limit $\epsilon=-g_r^2/4\sigma_0 h_r$. Even when the base flow is stable in the overlap range, a finite disturbance of sufficient size can produce the secondary-flow solution instead. Thus it may be very difficult to approach the linear-stability threshold $\epsilon=0$ in experiments involving subcritical bifurcations. This problem is exacerbated by the possibility that the upper-branch secondary flow, typically a two-dimensional flow, may be unstable to a wider class of perturbations (for example, three-dimensional disturbances). Thus one never sees a saturated secondary mode: only, perhaps, a transient that resembles it. Typically (in shear flows) there is a cascade of instabilities to turbulence, a situation sometimes referred to as snap-through transition to turbulence.

The systematic derivation of Landau's equation for specific flow problems was first accomplished for convection by Malkus and Veronis (1958), for channel flow by Stuart (1958, 1960) and Watson (1960), and for Taylor-Couette flow between rotating cylinders by Davey (1962). An important method for obtaining amplitude equations is the technique of center-manifold reduction (see Iooss and Adelmeyer, 1992; Haken, 1983; Richtmeyer, 1981). After the amplitude equations are obtained up to some order of approximation by such a reduction process, it may be necessary to convert them to normal form: a nonlinear change of variables eliminates as many nonlinear terms as possible (see Guckenheimer and Holmes, 1983). A variety of different flow problems may be reduced to the same normal form, which can often be predicted simply with such general considerations as the dimensionality of the problem (the number of interacting marginally stable modes) and the governing symmetries. Thus there is a universality among widely different physical systems in the sense that they are described by the same equation. In order to get the numerical values of the *coefficients* of the normal form, however, one must carry out the reduction process in detail for each particular fluid flow.

The existence of modes with a continuous band of wave vectors (k_x,k_y) centered around the critical wave vector $(k_c,0)$ is accounted for by allowing the amplitude to vary slowly in space and time. This leads to an extension of the Landau equation called an envelope equation (Newell, 1974, 1989). For steady-state transitions ($\omega=0$), as in thermal convection (see below), the envelope equation is

$$\frac{dA}{dt}=\sigma_0\epsilon A+\sigma_0\xi_0^2\left(\frac{\partial}{\partial x}-\frac{i}{2k_c}\frac{\partial^2}{\partial y^2}\right)^2 A-g|A|^2A, \tag{30}$$

where $\sigma_0\equiv\partial\sigma/\partial\epsilon$ and $\sigma_0\xi_0^2\equiv\partial^2\sigma/\partial k^2$ (evaluated at $\epsilon=0$ and $k=k_c$).

If there is a transition to oscillatory modes in quasi one-dimensional systems with reflection symmetry, one obtains either traveling waves or standing waves (depending on the nonlinear parameters); this leads to a coupled pair of complex Ginzburg-Landau equations:

$$\begin{aligned}\frac{dA}{dt}&=\sigma_0(1+ic_0)\epsilon A+v_g\frac{\partial A}{\partial x}\\&\quad+\sigma_0\xi_0^2(1+ic_1)\frac{\partial^2 A}{\partial x^2}-(g_r+ig_i)|A|^2A\\&\quad-(h_r+ih_i)|B|^2A,\end{aligned} \tag{31a}$$

$$\begin{aligned}\frac{dB}{dt}&=\sigma_0(1+ic_0)\epsilon B-v_g\frac{\partial B}{\partial x}\\&\quad+\sigma_0\xi_0^2(1+ic_1)\frac{\partial^2 B}{\partial x^2}-(g_r+ig_i)|B|^2B\\&\quad-(h_r+ih_i)|A|^2B.\end{aligned} \tag{31b}$$

Recalling that $s\equiv\sigma+i\omega$, the linear parameters are defined by derivatives evaluated at $\epsilon=0$ and $k=k_c$:

$$\sigma_0(1+ic_0)\equiv\frac{\partial s}{\partial\epsilon}$$

and

$$\sigma_0\xi_0^2(1+ic_1)\equiv\frac{\partial^2\sigma}{\partial k^2}.$$

Also, $v_g\equiv\partial\omega/\partial k$ is the group velocity, which must be small for the above equations to be a valid approximation. These equations can lead to some surprising dynamics that have been confirmed in experiment, including (for appropriate boundary conditions) spontaneous reversal of traveling-wave direction or to localized states. This discussion falls under the general topic of nonlinear pattern formation, the subject of reviews by Cross and Hohenberg (1993) and Newell *et al.* (1993).

1.2.8 Bifurcation Theory. Linear-stability theory predicts when a particular solution is unstable to a class of disturbances. The flow that results may or may not be described in terms of small perturbations away from the original flow, with dynamics described by amplitude equations. However, a more general question concerns the overall "universe" of possible solutions given a particular level of imposed stresses. Recall that for small stresses, there is a unique solution. As the stresses are increased, say in response to some control parameter μ, new solutions become possible. The classification of such solutions, their stability, and their relation to each other is the subject of bifurcation theory. Usually the theory is organized according to the systematic variation of one or more control parameters $(\mu_1,\mu_2,\ldots,\mu_m)$. We refer to the space of all possible values of these parameters as an m-dimensional parameter space with elements $\mu \equiv (\mu_1,\mu_2,\ldots,\mu_m)$.

Often, the first experiments with a fluid dynamical system will map the domains occupied in parameter space by various observed flows. Such a map is often called a phase diagram,* by analogy to thermodynamic response of substances to changes in temperature and pressure. As the parameters $\mu_1,\mu_2,\ldots,\mu_m$ are explored, the observed flows are not necessarily all possible solutions; they are just some of the stable ones. Furthermore, the observed flows may depend on the history of preparation. Often special efforts must be made to produce a flow that is a solution but is not usually obtained under normal experimental procedures. For example, control parameters may be quickly jumped to new values. There will be a strong transient, after which the final state may be quite different from that obtained by slow variation of the parameters to the same final values. Bifurcation theory serves as a guide to this search process.

For some flows (for example, boundary layers and free shear layers), the parameters evolve downstream and are thus not completely under the experimenter's control. In this case, bifurcation theory may be of more limited use but can still serve as a conceptual guide to understanding the transition process. Furthermore, the use of bifurcation theory is much simpler for small finite-sized systems or for systems hypothetically constrained to periodic boundary conditions. For such systems, there is a denumerable set of solutions, whereas infinite, translationally invariant systems possess continuous families of solutions. Moreover, a precise description of real experiments rests on a proper accounting of the finite size and of inevitable perturbations to assumed symmetries (Benjamin, 1978a,b).

A more modest goal than finding all possible fluid-flow solutions for a given μ is to find only those solutions that lie in a neighborhood in parameter space of the critical value μ_c where a particular solution becomes unstable. Under such circumstances, we say that the original solution "bifurcates" to the new solutions. Such bifurcation phenomena are most easily understood by first considering a single control parameter μ. Three complementary views of bifurcations are obtained by examining the response to changes in μ of

1. the spectrum of eigenvalues s of the normal modes of the problem linearized about the original solution;
2. some measure X that distinguishes different flow states from one another—for example, a component of the flow velocity at a particular point in space;
3. the dynamics in an abstract phase space whose coordinates are the amplitudes of relevant degrees of freedom of the system.

Discussion of phase-space dynamics is postponed to Sec. 3.

Consider a solution steady in time. We know that instability occurs when the eigenvalue s of some perturbation crosses the imaginary axis. If the eigenvalue remains real as it crosses [Fig. 4(a)], we speak of a steady-state bifurcation. If the eigenvalue is complex, then it crosses the imaginary axis at two complex-

---

*This is distinct from the term "phase portrait," discussed in Sec. 3.

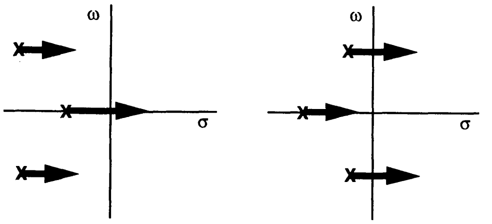

FIG. 4. Trajectories of eigenvalues in the complex s plane as a parameter μ is varied: (a) steady-state bifurcation; (b) Hopf bifurcation.

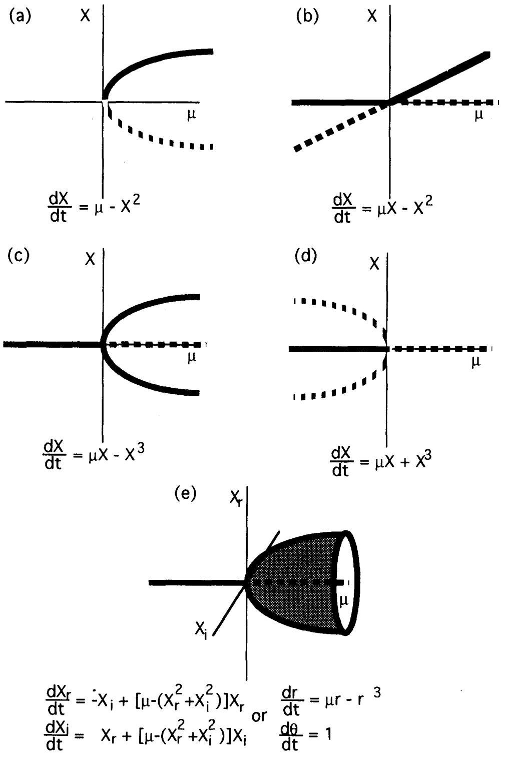
(a)
X
μ
$\frac{dX}{dt} = \mu - X^2$
(b)
X
μ
$\frac{dX}{dt} = \mu X - X^2$
(c)
X
μ
$\frac{dX}{dt} = \mu X - X^3$
(d)
X
μ
$\frac{dX}{dt} = \mu X + X^3$
(e)
X_r
X_i
μ
$\frac{dX_r}{dt} = -X_i + [\mu-(X_r^2+X_i^2)]X_r$
$\frac{dX_i}{dt} = X_r + [\mu-(X_r^2+X_i^2)]X_i$
or
$\frac{dr}{dt} = \mu r - r^3$
$\frac{d\theta}{dt} = 1$

FIG. 5. One-parameter bifurcations of steady solutions, with simplified versions of normal-form equations. Solid curves are stable solutions and dotted curves are unstable solutions. (a) Saddle node; (b) transcritical; (c) supercritical pitchfork; (d) subcritical pitchfork; (e) supercritical Hopf.

conjugate points $\pm i\omega_c$ [Fig. 4(b)] and we speak of a Hopf bifurcation.*

Next consider the variation of the measure X with μ. Plots of the curves $X(\mu)$ for all the possible solutions near μ_c are called bifurcation diagrams. The types of single-parameter ("codimension one") bifurcations from steady-state solutions are shown in Fig. 5 (Crawford, 1991). The measures X may be considered to be solutions of normal-form equations of the type

$$\frac{dX}{dt} = f(X,\mu), \tag{32}$$

where the appropriate $f(X,\mu)$ are shown in the figure. For the steady-state bifurcations, X may be considered real; for the Hopf bifurcation, the measure X for the bifurcating solution is complex with both an amplitude and a phase.

The pitchfork bifurcations in Fig. 5 assumed the symmetry that solutions with both X and $-X$ exist. Imperfections in the system will break this symmetry, changing these bifurcations into a composition of transcritical and saddle-node bifurcations as in Fig. 6(a). Furthermore, the idealized base state may be slightly wrong, again because of imperfections, and an inhomogeneous term must be put into the normal forms. The $X=0$ solutions of the transcritical and pitchfork bifurcations no longer solve the perturbed normal form and, as shown in Figs. 6(b) and 6(c), there is no true bifurcation of the base state into a qualitatively different solution, just a smooth deformation. This is called an imperfect bifurcation, and the generalization of the normal-form equations to account for perturbations is called an unfolding of the bifurcation. Note that there may be a nearby solution of different form, but it is no longer connected at some μ_c to the base state. Only the saddle-node bifurcation survives as generic, which means that it is not altered qualitatively by general perturbations of the normal form.

Suppose two parameters can be varied. Special cases of the bifurcations produced by varying, say, μ_1 can be obtained by tuning the other parameter, μ_2. These are called codimension-two bifurcations and include, for example,

---

*A complication arises in interpreting the bifurcation to traveling waves in systems with a preferred direction of flow, say channel flow. At any fixed value of μ, the traveling wave can appear as a stationary solution by transformation to a reference frame moving with the phase velocity of the traveling wave. However, the amplitude equation or "normal form" describing the bifurcation will be different from the normal form for a steady-state bifurcation because it will contain coefficients that cause the frequency of the wave to vary with amplitude [such as g_i in Eqs. (23) and (31)].

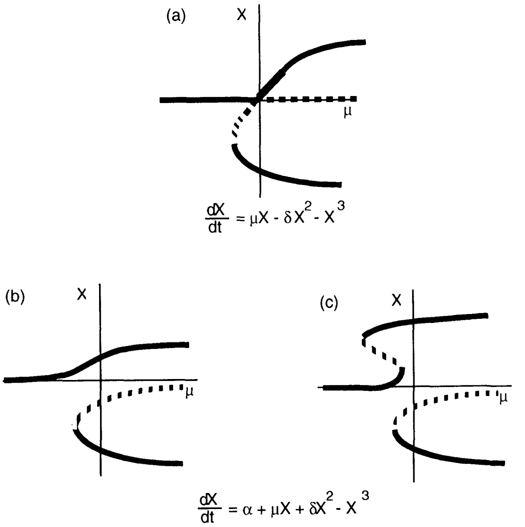

FIG. 6. The effect of system imperfections on a pitchfork bifurcation. (a) Reflection symmetry is no longer valid, and so a small term δX^2 is added to the normal form. (b), (c) Two cases resulting from a more general perturbation that causes the basic state $X=0$ to deform smoothly rather than reach a bifurcation: a small term α is added to the normal form.

1. switching of a bifurcation from supercritical to subcritical;
2. simultaneous bifurcation (bicriticality) of two independent modes that are not degenerate as a result of symmetry;
3. collision of two real eigenvalues to produce a complex-conjugate pair just when the eigenvalues cross the imaginary axis so that the frequency grows from zero (soft-mode instability).

Symmetry plays a crucial role in many fluid-dynamics bifurcation problems; the theory of bifurcations with symmetry has been under active development in recent years. For recent reviews, see Golubitsky *et al.* (1988) and Crawford and Knobloch (1991).

2. INSTABILITIES

The range of possible instabilities in fluid problems is enormous, and so we concentrate here on a few fundamentally different physical processes.

2.1 Waves and Instabilities at Flat Fluid Interfaces

Consider a flat interface $z=0$ between a fluid layer of density ρ_1 and a second one of density ρ_2 lying above the first (Fig. 7). The bottom layer terminates at $z=-h_1$ and the top layer terminates at $z=h_2$. There is an interfacial tension γ between the two fluids, and a body force $-\rho_i g\hat{\mathbf{z}}$ acts on each fluid (here and below, $i=1$ or 2). The base state is fluids at rest: $\mathbf{U}_1=\mathbf{0}$ and $\mathbf{U}_2=\mathbf{0}$. As shown in Fig. 7, let the fluid layer be deformed to a surface $z=\zeta(x,y,t)$, with consequent velocity fields $\mathbf{u}_1$ and $\mathbf{u}_2$. Our objective is to see how this deformation and these velocity fields evolve in time.

This is one case where the inviscid approximation works well. Furthermore, since the flow starts out irrotational ($\nabla\times\mathbf{u}_i=0$), we express the velocity fields in terms of potentials: $\mathbf{u}_i=\nabla\phi_i$. Incompressibility leads to Laplace's equation:

$$\nabla^2\phi_i=0, \tag{33}$$

with boundary conditions

$$u_z=\frac{\partial\phi_i}{\partial z}=0 \text{ at } z=\mp h_i \text{ for finite-width layers,} \tag{34}$$

or

$$\lim_{|z|\to\infty}\phi_i=0 \quad \text{for semi-infinite layers.} \tag{34b}$$

Given ϕ, the pressure p_i of each fluid is determined from Bernoulli's equation for unsteady motions:

$$\frac{\partial\phi_i}{\partial t}+\frac{1}{2}(\nabla\phi_i)^2+gz+\frac{p_i}{\rho_i}=C_i(t), \tag{35}$$

where $C_i(t)$ is an arbitrary function of time and can be absorbed into ϕ_i.

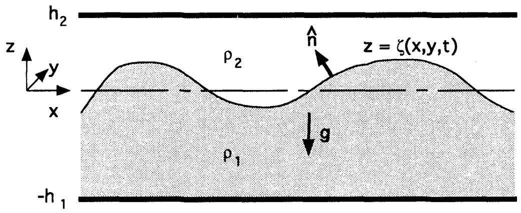

FIG. 7. Deformed interface $z=\zeta(x,y,t)$ between two fluids; gravity $\mathbf{g}$ acts downwards. The unit vector $\hat{\mathbf{n}}$ is normal to the surface.

Two linear conditions at the interface must also be satisfied. First, the pressure jump across the interface whose unit normal vector is $\hat{\mathbf{n}}$ is given by

$$p_2 - p_1 = -\gamma \nabla \cdot \hat{\mathbf{n}} \tag{36a}$$

$$= -\gamma \nabla \cdot \frac{(-\zeta_x, -\zeta_y, 1)}{[1+(\zeta_x)^2+(\zeta_y)^2]^{1/2}}, \tag{36b}$$

where the subscripts denote partial derivatives and all terms are evaluated at $\mathbf{r}=(x,y,\zeta(x,y,t))$. Combined with Bernoulli's equation above, the condition becomes

$$\rho_2\left(\frac{\partial \phi_2}{\partial t}+\frac{1}{2}(\nabla\phi_2)^2+g\zeta\right)-\rho_1\left(\frac{\partial \phi_1}{\partial t}+\frac{1}{2}(\nabla\phi_1)^2+g\zeta\right)$$

$$= -\gamma \nabla \cdot \frac{(-\zeta_x, -\zeta_y, 1)}{[1+(\zeta_x)^2+(\zeta_y)^2]^{1/2}}. \tag{37}$$

Second, the interface must follow the fluid motions. A scalar field $F(x,y,z) \equiv z-\zeta(x,y,t)$ has the constant value 0 on the interface, and so

$$\left(\frac{\partial}{\partial t}+\mathbf{u}_i \cdot \nabla\right)[z-\zeta(x,y,t)]=0, \tag{38}$$

which implies that the potentials satisfy

$$\frac{\partial \phi_i}{\partial z}=\frac{\partial \zeta}{\partial t}+\frac{\partial \phi_i}{\partial x}\frac{\partial \zeta}{\partial x}+\frac{\partial \phi_i}{\partial y}\frac{\partial \zeta}{\partial y}. \tag{39}$$

Again, all terms are evaluated at $\mathbf{r}=(x, y,\zeta(x,y,t))$.

Linearization keeps terms only to first order in the perturbations ϕ_1, ϕ_2, and ζ (and their derivatives). Furthermore, derivatives evaluated at the interface are approximated by derivatives evaluated at $z=0$. We expect solutions of the form

$$\Psi=\begin{pmatrix}\phi_1\\ \phi_2\\ \zeta\end{pmatrix}$$

$$=\begin{pmatrix}\hat{\phi}_1(z)\\ \hat{\phi}_2(z)\\ \zeta_0\end{pmatrix}\exp(ik_x x+ik_y y)\exp(st)+\text{c.c.} \tag{40}$$

(c.c. stands for "complex conjugate"). Insertion into the linearized problem yields the dispersion relation [here $k \equiv (k_x^2+k_y^2)^{1/2}$]

$$s^2=\frac{(\rho_2-\rho_1)gk-\gamma k^3}{\rho_1 \coth kh_1+\rho_2 \coth kh_2} \quad \text{for finite layers;} \tag{41a}$$

$$s^2=\frac{(\rho_2-\rho_1)gk-\gamma k^3}{\rho_1+\rho_2} \quad \text{for semi-infinite layers.} \tag{41b}$$

If $\rho_2 < \rho_1$ (light fluid over heavy fluid), $s=i\omega$ is purely imaginary, and so the modes are neutrally stable traveling waves:

$$\zeta(x,y,t)=\zeta_0 \exp[i(k_x x+k_y y-\omega t)]. \tag{42}$$

When the densities of both fluids are significant, the modes are called internal waves. For the case $\rho_2=0$, $\rho_1 \equiv \rho$, and semi-infinite layers, we recover the linear dispersion relation for "water" waves,

$$\omega^2=gk-(\gamma/\rho)k^3, \quad \text{capillary gravity waves.} \tag{43}$$

The two limiting cases are

$$\omega \approx \begin{cases}(gk)^{1/2}, & \text{long wavelength: gravity waves}\\ (\gamma k^3/\rho)^{1/2}, & \text{short wavelength: capillary waves.}\end{cases} \tag{44}$$

Both terms in the dispersion relation have equal size when

$$\lambda \equiv 2\pi/k$$

$$=2\pi(\gamma/\rho g)^{1/2}$$

$$\approx 1.7 \text{ cm (for pure water).} \tag{45}$$

$(\gamma/\rho g)^{1/2}$ is called the capillary length. A full treatment of this problem, including damping due to viscosity, is found in Chandrasekhar (1961).

2.1.1 Heavy Fluid Above Light Fluid: Rayleigh-Taylor Instability. Perhaps one of the most intuitive instabilities is that of a heavy fluid lying on top of a light fluid. Anyone who paints a ceiling thinks of this situation. A flat layer of heavy fluid can certainly be supported by the hydrostatic pressure of a lighter fluid, but any deformation of the interface quickly grows so that the fluids overturn (unless the heavy layer is thin enough to be supported, as it divides into droplets, by adhesion to the ceiling). The classic analysis of this problem is due to Lord Rayleigh.

A small perturbation grows if the eigenvalue s of the mode given in Eq. (40) has a positive real part. If the fluid is initially at rest but has an initial deformation $\zeta(x,y;t=0) =\zeta_0 \cos kx$, the interface will evolve in the linear approximation as $\zeta(t)=\zeta_0 \cos kx \cosh st$. From Eq. (41), growth will occur if $\rho_2>\rho_1$ and $k<k_{\text{cutoff}}=[(\rho_2-\rho_1)g/\gamma]^{1/2}$. Without surface tension all wave numbers are unstable. Thus the effect of surface tension is to place a small-wavelength cutoff: only modes with wavelengths longer than $2\pi/k_{\text{cutoff}}$ will grow. For pure water above air, this wavelength is again 1.7 cm. The cutoff wavelength is simply demonstrated by showing that water will remain in an inverted container if the opening is covered by a wire mesh. There is also a wave number of maximum growth rate. For semi-infinite layers $k_{\max}=k_{\text{cutoff}}/\sqrt{3}$.

Taylor (1950) recognized that this instability would also take place if the system had a uniform acceleration **a** in the direction towards the more dense layer, that is, if the light fluid pushes the heavy fluid. If there is both a uniform body force per unit mass **g** and a uniform acceleration **a**, one obtains an equivalent system by transforming to a noninertial frame at rest relative to the accelerating fluids and assuming a body force per unit mass $\mathbf{g}'=\mathbf{g}-\mathbf{a}$. Thus the surface of liquid in an upright container will be unstable if it is accelerated downwards faster than free fall.

An important application of this idea occurs in the study of inertial-confinement fusion (see FUSION, INERTIAL CONFINEMENT). A spherical pellet of deuterium is rapidly compressed by an radial inward thrust (ablative acceleration) of vapor produced by an intense laser pulse heating the pellet exterior. As the less dense vapor accelerates into the more dense interior of the pellet, an instability develops that disrupts the desired uniform compression. Another important application comes from plasma fusion: curved magnetic fields can act like an effective gravitational force. Inverted situations can arise so that confined plasmas are subject to Rayleigh-Taylor instabilities (see Infeld and Rowlands, 1990).

Considering again a planar geometry, the nonlinear behavior of the unstable interface is quite complex. The dense fluid may form falling spikes while the lighter fluid rises as bubbles with rounded fronts. The deforming interface is subject to shear instabilities, and there is a complicated process of breakup, coalescence, and mixing. A review of this process and its study may be found in Sharp (1984). Recently, well-controlled experiments (Fermigier *et al.*, 1992) have applied an amplitude-equation formalism to explain the weakly nonlinear pattern formation in unstable thin layers.

2.1.2 Shocks at Fluid Interfaces: Richtmeyer-Meshkov Instability. Treating a shock as an instantaneous acceleration, one expects instabilities in shock waves passing through a fluid interface (Richtmeyer, 1960; Meshkov, 1969). In fact, instability occurs regardless of the direction of the shock. Strong mixing is induced by the instability. An astrophysical application is the study of supernova explosions. Shock-tube experiments by R. F. Benjamin are compared to numerical simulations in Cloutman and Wehner (1992).

2.1.3 Parametrically Excited Surface Waves. In 1831 Michael Faraday observed pronounced deformations of the surface of liquid in a dish that oscillated up and down. Furthermore, the frequency of the surface patterns appeared to be *half* that of the driving frequency. This was an early record of the phenomenon of parametric instability, which has since been found and exploited in a wide variety of systems (mechanical, electrical, optical, etc.).

Benjamin and Ursell (1954) found that the linearized analysis of Faraday's experiment with a vibrating layer of fluid leads to Mathieu's equation [Eq. (50) below]. Let the undisturbed surface lie at $z=0$ and assume that the vertical position of the bottom of the dish vibrates in time as $z_0(t)=-h+\epsilon \cos\Omega t$. As with Taylor's analysis of accelerating fluid layers, we can transform to a noninertial frame by replacing gravity g with an effective gravity $g+\epsilon\Omega^2 \cos\Omega t$. The nomenclature "parametric instability" arises because the forcing of this system arises through one of the *parameters* (gravity). The surface is described in the noninertial frame by

$$z=\zeta(x,y,t)=\zeta_0(t)S_{lm}(x,y). \tag{46}$$

The indices l and m are integers and arise as follows. For rectangular containers of dimensions a and b,

$$\begin{aligned}S_{lm}(x,y)&=\cos(k_x x)\cos(k_y y)\\&=\cos(l\pi x/a)\cos(l\pi x/b),\end{aligned} \tag{47a}$$

where the second part of the equation shows that the wave number $k^2 \equiv (k_x^2 + k_y^2)^{1/2}$ must be quantized according to $k^2 = k_{lm}^2 = (l\pi/a)^2 + (m\pi/b)^2$ in order to satisfy the boundary condition that at each wall the surface is flat in the direction normal to the wall (a kinematic consequence of the condition that the normal component of the velocity must vanish). For cylindrical containers of radius r_0

$$S_{lm}(r,\theta) = J_l(k_{lm}r)\cos(l\theta+\phi), \tag{47b}$$

where the same boundary condition requires that k_{lm} is the mth zero of the derivative $J_l'(k_{lm}r_0)$.

The amplitude $\zeta_0(t)$ satisfies the equation

$$\frac{d^2\zeta_0}{dt^2} + \beta\frac{d\zeta_0}{dt} + (\omega_{lm}^2 + \epsilon\Omega^2 k_{lm}\tanh k_{lm}h\ \cos\Omega t)\zeta_0 = 0, \tag{48}$$

where β is the phenomenological damping coefficient and, from Eq. (41a),

$$\omega_{lm}^2 = k_{lm}(g+\gamma k_{lm}^2/\rho)\tanh k_{lm}h. \tag{49}$$

Defining $\tilde{\zeta}_0 \equiv e^{\beta t}\zeta_0$ and $\tau \equiv \Omega t/2$ we have Mathieu's equation:

$$\frac{d^2\tilde{\zeta}_0}{d\tau^2} + (p - 2q\cos 2\tau)\tilde{\zeta}_0 = 0, \tag{50}$$

where

$$p \equiv 4(\omega_{lm}^2 - \beta^2)/\Omega^2, \tag{51a}$$

$$q \equiv 2\epsilon k_{lm}\tanh k_{lm}h. \tag{51b}$$

Mathieu's equation has the remarkable property (McLachlan, 1947) that for a given level of forcing q there are bands $b_n(q) < p < a_n(q)$ ($n = 0,1,2,3,\ldots$ with $b_0 = -\infty$) within which a solution $\tilde{\zeta}_0$ diverges exponentially with time and has the Floquet form

$$\tilde{\zeta}_0 = e^{(\mu+i\pi)\Omega t/2}\phi_n(t) \quad \text{for } n \text{ odd:}$$

$$\text{subharmonic relative to } \cos\Omega t$$

$$[\text{Floquet exponent } (\mu+i\pi)\Omega/2], \tag{52a}$$

$$\tilde{\zeta}_0 = e^{\mu\Omega t/2}\phi_n(t) \quad \text{for } n \text{ even:}$$

$$\text{harmonic relative to } \cos\Omega t$$

$$(\text{Floquet exponent } \mu\Omega/2), \tag{52b}$$

with $\phi_n(t+2\pi/\Omega) = \phi_n(t)$ and $\mu(p,q) > 0$ (real). Thus the actual surface deformation is given by

$$\zeta_0(t) = e^{(\mu\Omega/2-\beta)t}e^{i\pi\Omega t/2}\phi_n(t)$$

$$\text{or} \tag{53}$$

$$\zeta_0(t) = e^{(\mu\Omega/2-\beta)t}\phi_n(t)$$

and is unstable provided the frequency ratio $4(\omega_{lm}^2 - \beta^2)/\Omega^2$ lies inside one of the unstable bands and $\mu(p,q) > 2\beta/\Omega$. This requires a threshold value of the forcing amplitude ϵ that will depend on frequency (recall that q is proportional to ϵ); note that in the limit of zero damping ($\beta = 0$) there are resonant frequencies $\omega_{lm} = n\Omega/2$ at which there is an unstable solution for vanishingly small forcing $\epsilon = 0$. In a real system with small damping, a small modulation signal adjusted to resonance can cause a large response (assuming that nonlinearities saturate the unstable solution), and we have a parametric amplifier.

If the Faraday experiment is done in relatively small containers, the mode frequencies will be distinct so that only one mode might become parametrically unstable for carefully selected amplitude ϵ and frequency Ω. However, as ϵ is increased, bands of instability for different modes overlap and there is competition. Since the full equations are nonlinear and since symmetry-caused degeneracy ensures that three or more modes interact when bands overlap, one has the ingredients for chaotic dynamics. This is exactly what has been observed by Ciliberto and Gollub (1985). In their experiment, a cylinder of radius $r_0 = 6.35$ cm held a water layer of depth $h \approx 1$ cm. The (4,3) mode was excited for a frequency $\Omega/2\pi = 15.77$ Hz at a drive amplitude $\epsilon \approx 37$ μm (quite small). Nearby, at 16.21 Hz, a (7,2) mode could be excited with nearly the same drive amplitude. At a small range of intermediate frequencies and drive amplitudes greater than 100 μm, the bands overlapped and chaotic dynamics was found. For such systems, the nonlinear perturbation analysis, though difficult, is tractable thanks to the clearly separated modes. Thus considerable progress has been made in a first-principles quantitative understanding of the chaotic dynamics of this fluid system. Recent reviews may be found in Miles and Henderson (1990) and Gollub (1991).

2.1.4 Surface Waves.

Consider a small-amplitude surface wave as the base state:

$$\Psi_1 = A_1\Psi_{10}e^{i(\mathbf{k}_1\cdot\mathbf{r}-\omega_1 t)}. \tag{54a}$$

What is its stability? An important mechanism for instability is through resonant wave interactions with other wave modes. The simplest such interaction is the resonant triad, arising from the nonlinear mixing of two waves,

$$\begin{aligned}\Psi_2&=A_2\Psi_{20}e^{i(\mathbf{k}_2\cdot\mathbf{r}-\omega_2 t)}\\ \Psi_3&=A_3\Psi_{30}e^{i(\mathbf{k}_3\cdot\mathbf{r}-\omega_3 t)},\end{aligned}\tag{54b}$$

to form a wave varying like

$$\exp\{i[(\mathbf{k}_2+\mathbf{k}_3)\cdot\mathbf{r}-(\omega_2+\omega_3)t]\}.$$

This will be seen by the original wave $(\mathbf{k}_1,\omega_1)$ as a forcing term that is resonant if

$$\mathbf{k}_1=\mathbf{k}_2+\mathbf{k}_3,\tag{55a}$$

$$\omega_1(\mathbf{k}_1)=\omega_2(\mathbf{k}_2)+\omega_3(\mathbf{k}_3).\tag{55b}$$

This is a stringent condition on the dispersion relations and may not be satisfied for any choice of $\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3$. Indeed, for gravity waves $\omega=(gk)^{1/2}$ the resonant conditions cannot be satisfied at this order, and a higher-order four-wave resonance must be sought instead (Phillips, 1974). For gravity capillary waves, however, the resonant-triad conditions can be met, and the consequences have been explored in some beautiful experiments by Henderson and Hammack (1987). For reviews of wave interactions, with discussion of amplitude equations, see Phillips (1974), Craik (1985), and Hammack and Henderson (1993).

2.2 Capillary Phenomena

In the above examples, surface tension played a stabilizing role, as was the case when it caused a cutoff of short-wavelength modes in the Rayleigh-Taylor instability. In base states that start out with curved surfaces, however, surface tension may be destabilizing. If we ignore damping, this may be understood by energy considerations: a base state will be unstable if a new surface of lower energy can be found. This is indeed the case, for example, in the breakup of a cylindrical liquid column into drops (Fig. 8).

The analysis of surface deformations is similar to that for surface waves, except that we now have a cylindrical interface defined by $r=a+\zeta(z,\theta)$. For perturbations $\zeta(z)=\zeta_0\, e^{i(kz+m\theta)+st}$, a linear analysis gives

$$s^2=\frac{\gamma}{\rho a^3}\frac{kaI_n'(ka)}{I_n(ka)}(1-k^2a^2-m^2),\tag{56}$$

where $I_n(ka)$ is a modified Bessel function of order n. Only the term $1-k^2a^2-m^2$ determines the sign of s^2 and we see, then, that s is real and instability occurs if $m=0$ (axisymmetric disturbances) and $k<1/a$. In other words, the column is unstable for all wavelengths $\lambda>2\pi a$, the circumference of the column. The fastest growing mode occurs for $ka=0.6970$, or $\lambda=1.435(2\pi a)$, with a growth rate $s=0.3433(\gamma/\rho a^3)^{1/2}\approx(0.12\ \mathrm{s})^{-1}$ for a water column with radius $a=0.5$ cm and surface tension $\gamma=73$ dyn/cm. In experiments, the column is produced as a jet exiting from a nozzle with velocity U. Forcing the column with a speaker oscillating at frequency ω_0 rad/s gives a well defined wave number $k=U/\omega_0$. The growth rate may be inferred by fitting an exponential envelope to the jet profile; good agreement is found (Donnelly and Glaberson, 1966) on comparing the initial growth with the inviscid theory. See Drazin and Reid (1981) and Chandrasekhar (1961) for further discussion. For experimental investigation of the nonlinear regime, see Goedde and Yuen (1970).

Capillarity is also important for studying the shapes of rotating and oscillating liquid drops. Indeed, the problem of equilibrium shapes and their stability for rotating fluid masses held together either by gravity or by

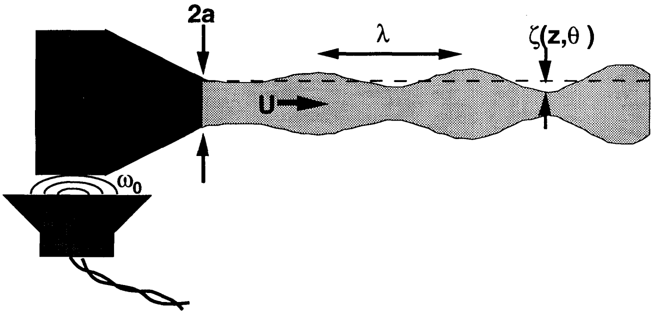

FIG. 8. Liquid column or "jet" emanating from a nozzle. The column is deformed by a disturbance of wavelength $\lambda=2\pi U/\omega_0$, where ω_0 is the angular frequency of a loudspeaker exciting the nozzle. The disturbance grows if $\lambda>2\pi a$, the circumference of the column.

surface tension is a classic one (Chandrasekhar, 1969). In 1834 Jacobi found a family of ellipsoidal figures of equilibrium that broke the axisymmetry of the rotating spheroids derived by Maclaurin a century earlier. In 1885, Poincaré also found a pear-shaped family of solutions and coined the term "bifurcation" to describe the branching of the solution families. A resurgence of interest in the problem for the case where the drops are held together by surface tension has occurred thanks to the ability to conduct extended experiments in zero gravity. This has also lead to extensive studies of the modes of oscillation of liquid drops. These studies are motivated, in part, by the development of zero-gravity materials processing technology. For a review, see Wang (1988).

2.3 Propagating Fronts

An important class of nonequilibrium phenomena concerns the shape of a propagating front separating one medium from another. Examples of this situation include

1. propagation of one fluid into another, where the two fluids have different viscosities (viscous fingering); this arises, for example, in secondary oil recovery when water is used to drive oil through the porous beds towards extraction wells;
2. advance of a region undergoing the chemical reactions of combustion into an unignited phase (flame fronts); this is of obvious importance for the efficiency of combustion processes;
3. expansion of a domain undergoing crystallization (directional solidification); the resulting structures may strongly influence the metallurgical properties of commercial alloys;
4. the motion of an interface of a fluid that is wetting a solid substrate; this concerns technological problems of coating, draining, and drying.

The phenomenology of these problems is very rich, often producing intricate structures as a result of fingering instabilities. The front deforms into fingers that penetrate the second medium; these fingers may themselves experience fingering instabilities, etc. However, in some circumstances (see below) the front shape is ultimately governed by a selection process that can be quite subtle.

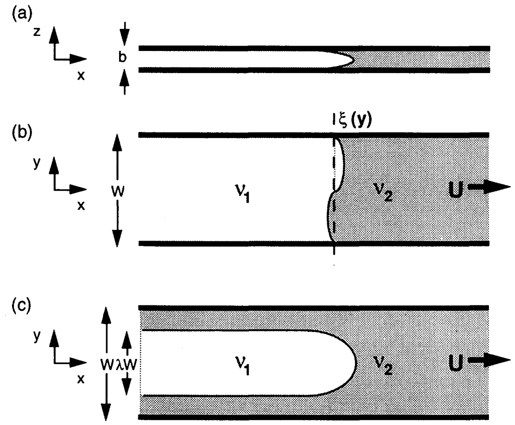

FIG. 9. Hele-Shaw cell. (a) Cross section of narrow direction; (b) cross section of wide direction, showing hypothetical interface undergoing deformation $x-Ut=\xi(y)$; (c) a viscous finger.

One problem that has received considerable attention is the Saffman-Taylor problem concerning the shape of the interface between a less viscous fluid (for example, air) pushing a more viscous fluid (for example, oil or glycerine) through a very narrow channel called a Hele-Shaw cell (Fig. 9). Because the gap b between the top and bottom walls of the cell is so small, the flow is two dimensional and is dominated by viscosity, obeying Darcy's law that the velocity—defined here to be an average across the narrow gap—is proportional to the pressure gradient:

$$\mathbf{u}_i=-(b^2/12\mu_i)\nabla p_i, \qquad (57)$$

where $i=1$ or 2 distinguishes the two fluids and μ_i is the dynamic viscosity. Gravity is neglected on the assumption that the cell is horizontal. Incompressibility then implies that the pressure will satisfy Laplace's equation in each fluid:

$$\nabla^2 p_i=0. \qquad (58)$$

The base flow is simply the case where both fluids and the interface move with velocity $\mathbf{U}=U_0\hat{\mathbf{x}}$. Observing from a frame of reference moving with the undisturbed interface, we want to known what happens if there is a deformation

$$x=\xi(y)=\xi_0\cos(ky)e^{\sigma t}. \qquad (59)$$

We omit discussion of the kinematic and pressure-jump conditions at the interface (which has interfacial tension γ), noting that in simplest form* they are similar to those given in Sec. 2.1. Linearized analysis gives the result

$$\sigma=[(\mu_2-\mu_1)U-(\gamma b^2/12)k^2]k/(\mu_1+\mu_2). \qquad (60)$$

This implies that $\sigma>0$ and instability occurs for wave numbers

$$k^2<[12(\mu_2-\mu_1)/\gamma b^2]U. \qquad (61)$$

If the flow is confined by side walls a distance W apart (where free-slip boundary conditions are usually assumed as a good approximation), the wave numbers k are restricted to values $k=2\pi n/W$. Taking $n=1$ for the smallest wave number and defining the dimensionless parameter

$$d_0\equiv\frac{b^2}{W^2}\frac{\gamma}{(\mu_2-\mu_1)U}, \qquad (62)$$

we see that instability occurs if the velocity is large enough that $d_0<3/\pi^2$.

The nonlinear problem turns out to be one of very great subtlety. This is because surface tension is a singular perturbation, so that solutions for small surface tension do not asymptotically approach those for zero tension. A one-parameter family of steady-state solutions was found for zero surface tension by Saffman and Taylor (1958):

$$\frac{\xi(y)}{W}=\frac{1-\lambda}{\pi}\ln\cos\frac{\pi y}{\lambda W}, \qquad (63)$$

where the parameter λ is the width of the finger relative to the width of the cell [as $y\rightarrow\pm\lambda W/2$, $\xi(y)\rightarrow-\infty$]. However, there is no indication which value of λ will be chosen by an experiment. Furthermore, Eq. (61) may be interpreted to say that as surface tension vanishes, even very-short-wavelength perturbations destabilize a front. Relative to such short-wavelength perturbations, the Saffman-Taylor front looks locally flat and ought to be subject to the same instability. Experiments show, however, that viscous fingers are stable and that a width parameter λ is uniquely determined by the velocity. Considerable progress has been made recently in understanding the intricacies of this problem; see Saffman (1986) and Bensimon *et al.* (1986).

2.4. Thermal- or Buoyancy-Driven Instabilities

We know intuitively that fluid layers that are warmer at the bottom will have a tendency to overturn. Gradients in concentration of a solute can have a similar effect. The overturning is driven by buoyancy with respect to some body force, usually gravity; the overturning is resisted by viscosity and diffusion. If a sufficient source of heat or solute is steadily applied, the overturning can be sustained. The transport of heat or solute will be greatly enhanced over the rate that would occur through diffusion alone. This enhanced means of transport is called convection; if the fluid motion responsible for the convection occurs spontaneously, this is called free convection. If the motion is imposed (for example, by external pumps) then this is forced convection, a case not necessarily associated with instabilities. There are also cases of mixed convection in which externally imposed motion is influenced by buoyancy forces. The more general concept of transport arising through fluid motion is also termed advection, with the term "convection" usually restricted to heat transport or buoyancy-related solute transport.

The situation producing free convection can be generalized to cases in which gradients are imposed horizontally; for example, one side of a vertical chamber may be kept warmer than the other. Such cases are of immense practical importance but, even though the basic mechanism for instability is the same, the analysis is substantially different from the case of vertically imposed gradients. Finally, rotation may be imposed on a differentially heated annulus, producing models important to geophysical and astrophysical studies.

Early quantitative experiments on convection in horizontal fluid layers were reported by Bénard in 1900. Heating the layers from below, he often obtained arrays of nearly hexagonal cells of overturning fluid. In 1916, Rayleigh published his theory for buoyancy-

---

*As discussed in Bensimon *et al.* (1986), there are more appropriate pressure-drop conditions that include the effect of the velocity of the interface as it slips along the bounding plates.

induced instability in fluid layers heated from below. Ironically, it was realized much later (Block, 1956; Pearson, 1958) that Bénard's experiments corresponded to what is now known as Marangoni convection, in which the source of instability is temperature variation of the surface tension of the fluid. However, the original idea of fluid layers heated from below (still referred to as Bénard convection) has led to some of the most refined and precise experiments in fluid mechanics.

Figure 10 shows the basic configuration of experiments with horizontal fluid layers. (For a general review of the physics of convection, see Normand *et al.*, 1977.) In the ideal case, the layers have infinite horizontal extent while the top and bottom plates have infinite thermal conductivity (and thus completely uniform temperature). In practice, materials are chosen so that the top and bottom plates have much larger conductivity than the fluid, while the side walls are chosen to be good insulators. It should be noted that transparent top plates are often used in order for the fluid motion to be visible, while the bottom plate is often made of polished metal that serves as a mirror in the shadowgraph method for visualizing fluid-density differences. Heat is usually supplied by heaters embedded in the bottom plate, with servo control to establish either constant temperature difference or constant heat flux. The top-plate temperature may be regulated by carefully directed fluid currents from a temperature-controlled bath. The whole system is isolated from room-temperature fluctuations by an insulating container.

A rectangular container with horizontal dimension L_x and L_y and gap d between the plates has aspect ratios L_x/d and L_y/d; a cylindrical container of diameter D has an aspect ratio D/d. Experiments are often designed with either large aspect ratio (say greater than 30) or small aspect ratio (1–3). Sometimes one aspect ratio (say in the y direction) is intentionally made smaller than the other in order to establish a preferred alignment of the convective rolls, rendering the horizontal variation of the observed patterns nearly one dimensional. The large–aspect-ratio systems conform better to the ideal of infinite horizontal extent; however, after instability occurs the situation can be very complex because the (near) isotropy and translational invariance of the system produce a quasicontinuum of interacting modes. The small–aspect-ratio systems pose a substantially different stability problem, but one that may be solvable by numerical methods. In this case, a small number of modes can dominate the behavior of the system, and the dynamics is, in principle, reducible to a system of ordinary differential equations. An interesting variation on the design of convection apparatus has been the construction of annular containers.

Consider an infinite incompressible fluid layer whose top ($z=d/2$) and bottom ($z=-d/2$) layers are held at temperatures T_t and T_b, respectively. For small enough temperature differences, heat is simply conducted through the fluid, which remains motionless. This base state is

$$\mathbf{U}=\mathbf{0} \quad \text{(fluid is at rest)}, \tag{64a}$$

$$T=T_b+\frac{T_t-T_b}{d}\left(z+\frac{d}{2}\right) \quad \text{(linear temperature profile)}, \tag{64b}$$

$$\rho=\rho_b[1-\alpha(T-T_b)] \quad \left(\text{neglecting } \frac{d\alpha}{dT}\right), \tag{64c}$$

$$P=P_b-\int_0^z \rho(z)g\,dz. \tag{64d}$$

Here $\alpha\equiv-(1/\rho)(\partial\rho/\partial T)$ is the thermal expansion coefficient of the fluid.

If the bottom plate is cooler than the top, the system is stably stratified and fluid remains at rest. If the temperature gradient is reversed by heating the fluid layer from below, instability sets in when the thermally induced buoyancy is strong enough to overcome the dissipative effects of viscosity ν and thermal diffusivity κ. Imagine a small parcel of fluid of size a that is slightly warmer (by an

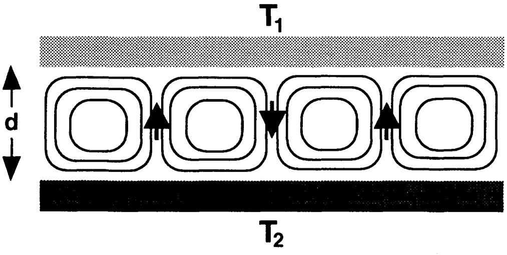

FIG. 10. Convective cells forming in fluid heated from below. When temperature T_2 is sufficiently larger than temperature T_1, fluid rises and falls in cellular patterns.

amount θ) than the surrounding fluid. The buoyancy force is $\rho_b\alpha\theta a^3 g$; the parcel accelerates until the buoyancy force is opposed by a viscous drag approximately equal to $\rho\nu u a$ (compare to Stokes's formula, $6\pi\rho\nu u a$, for the drag on a sphere of radius a moving with velocity u). This gives $u \sim g\alpha\theta a^2/\nu$. Thus the time for the parcel to move a distance comparable to its size is $\tau \sim a/u \sim \nu/g\alpha\theta a$. However, heat diffuses out of the parcel with a time scale $\tau_\theta \sim a^2/\kappa$. This time must be longer than the time scale of motion of the parcel; otherwise the cause of the buoyancy (the temperature difference) is removed. Thus we require $\tau < \tau_\theta$, meaning $\nu/g\alpha\theta a < a^2\kappa$ or $\alpha g\theta a^3/\nu\kappa > 1$. Now the characteristic scale for θ is $T_b - T_t$, while the characteristic scale for the size of the parcel is d. Thus we expect instability to occur when the Rayleigh number

$$Ra \equiv \alpha g(T_2 - T_1)d^3/\nu\kappa \tag{65}$$

exceeds some critical value Ra_c. The numerical value of Ra_c must be obtained from the detailed stability analysis using the equations of motion and appropriate boundary conditions.

Let the deviation from the base state be given by velocity $\mathbf{u}(u,v,w)$, temperature θ, and pressure p. In the Oberbeck-Boussinesq approximation the temperature variation is coupled to the dynamics only through the buoyancy force per unit volume $\rho_b\alpha\theta g$; otherwise, the density and other material properties of the fluid are assumed constant. These can be well approximated in the laboratory provided the temperature differences remain moderate (in practice, of order several kelvins). The physical quantities can be scaled by d (distance), $\tau_\theta \equiv d^2/\kappa$ (time), κ/d (velocity), and $\kappa\nu/\alpha g d^3$ (temperature). With the scalings the full nonlinear equations (Boussinesq equations) for the disturbance fields are

$$\nabla\cdot\mathbf{u} = 0, \tag{66a}$$

$$\frac{1}{Pr}\left(\frac{\partial\mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p\right) = \nabla^2\mathbf{u} + \theta\hat{\mathbf{z}}, \tag{66b}$$

$$\frac{\partial\theta}{\partial t} + (\mathbf{u}\cdot\nabla)\theta = \nabla^2\theta + Ra\, w. \tag{66c}$$

The Prandtl number $Pr \equiv \nu/\kappa$, a quantity solely based on the fluid physical properties, expresses the relative importance of the two nonlinear terms $(\mathbf{u}\cdot\nabla)\mathbf{u}$ versus $(\mathbf{u}\cdot\nabla)\theta$, and thus may be expected to strongly affect the nonlinear behavior. The term $Ra\, w$ comes from the advective transport of thermal energy in the base-state profile by the z component w of the disturbance velocity. There are two types of boundary conditions at $z = \pm\frac{1}{2}$ (note that z now refers to the dimensionless vertical coordinate): free (stress-free), $\theta = \partial u/\partial z = \partial v/\partial z = w = 0$, appropriate (approximately) if the convecting fluid is bounded by another fluid of much lower viscosity and the interface is assumed to remain flat; and rigid (no slip), $\theta = u = v = w = 0$, appropriate if the convecting fluid is bounded by rigid walls.

The linearized problem is obtained by neglecting the advective terms $(\mathbf{u}\cdot\nabla)\mathbf{u}$ and $(\mathbf{u}\cdot\nabla)\theta$. Assuming a system of infinite horizontal extent, solutions may be expected of the following form:

$$(u,v,w,\theta,p) = (\hat{u}(z),\hat{v}(z),\hat{w}(z),\hat{\theta}(z),\hat{p}(z)) \times \exp[st + i(k_x x + k_y y)]. \tag{67}$$

Some manipulation yields equations coupling $\hat{w}$ and $\hat{\theta}$:

$$Pr[(D^2 - k^2)^2\hat{w} - k^2\hat{\theta}] = s(D^2 - k^2)\hat{w}, \tag{68a}$$

$$(D^2 - k^2)\hat{\theta} + Ra\hat{w} = s\hat{\theta}, \tag{68b}$$

where $D \equiv d/dz$. Two types of boundary conditions at $z = \pm\frac{1}{2}$ are

$$\hat{\theta} = \hat{w} = D^2\hat{w} = 0 \quad \text{(stress-free or "free")}, \tag{69a}$$

$$\hat{\theta} = \hat{w} = D\hat{w} = 0 \quad \text{(no slip or "rigid")}. \tag{69b}$$

The eigenvalue s for this problem, with either type of boundary conditions, is real, and so the secondary flow that results from instability is a steady flow. Although the case of stress-free boundary conditions at both walls is unrealistic, it has the advantage that closed-form solutions for the fields and dispersion relation $s = s(k,Ra,Pr)$ can be obtained. They are

$$(\hat{w},\hat{\theta}) = (W_n,\Theta_n)\sin(n\pi z), \tag{70}$$

$$s^2 + (1+Pr)(k^2 + n^2\pi^2)s + Pr\frac{(k^2+n^2\pi^2)^3 - k^2 Ra}{k^2 + n^2\pi^2} = 0. \tag{71}$$

The condition $s = 0$ gives a family of marginal stability curves indexed by the integer n:

$$Ra(k,n) = (n^2\pi^2 + k^2)^3/k^2. \qquad (72)$$

The minimum Rayleigh number is obtained for $n = 1$; minimization must also be done over k. The resulting critical Rayleigh number Ra_c and critical wave number k_c are summarized in Table 1 for all three possible pairs of boundary conditions. These critical values do not depend on Prandtl number. We note that the stability problem with rigid boundaries requires numerical solution of a transcendental equation to obtain the marginal curve $Ra(k)$.

A nonlinear analysis must determine if the resulting eigenmodes saturate into a stable secondary flow. Furthermore, for an infinite system, there is no preferred direction for the wave vector **k**. Thus it is necessary to investigate the nonlinear selection of a stable pattern out of all possible combinations of modes with wave numbers of the same magnitude but different directions. Most study has been given to three possible cases:

1. two wave numbers, $\pm k_c\hat{\mathbf{x}}$ (parallel rolls);
2. four wave numbers, $\pm k_c\hat{\mathbf{x}} \pm k_c\hat{\mathbf{y}}$ (cross rolls);
3. six wave numbers, $\pm k_c\hat{\mathbf{x}}$ and $\pm\frac{1}{2}k_c\hat{\mathbf{x}} \pm \frac{1}{2}\sqrt{3}k_c\hat{\mathbf{y}}$ (hexagons).

Perturbation theory establishes a nonlinear state for each of these types and then examines its instability to perturbations of the form of the other two types. For fluids that satisfy the Oberbeck-Boussinesq approximations, parallel rolls are the only stable choice for Rayleigh numbers just above threshold. These rolls have an amplitude that satisfies a form of the Landau equation discussed in Sec. 1.2.7 with real coefficients:

$$\frac{dA}{dt} = \sigma_0\epsilon A - g|A|^2 A. \qquad (73)$$

Here $\epsilon \equiv (Ra - Ra_c)/Ra_c$ and g is a real Landau constant. One measure of the roll amplitude accessible to experiment is the ratio of the heat transfer with convection to the heat transfer in the conducting state. This is called the Nusselt number. Precise experiments show that $Nu \propto \epsilon^{1/2}$, in accordance with the $\epsilon^{1/2}$ growth of the amplitude A predicted by the Landau equation.

As the Rayleigh number is raised above the critical value Ra_c, the convection rolls may select a wave number k from a continuous finite range of wave numbers bounded by the marginal stability curve. Thus nonlinear theory must also address, for each wave number k, the question of stability of rolls with that wave number to arbitrary disturbances when Ra exceeds the marginal stability value $Ra_m(k)$. Rolls with wave numbers close to the marginal stability boundary turn out to be unstable to large–length-scale modulations (Eckhaus instability) that end up replacing the roll pattern with a parallel set of rolls with a stable wavelength. Alternatively, a cross-roll instability changes wavelength by replacing the unstable rolls with a perpendicular set of rolls. A final mechanism for wave-number adjustment is through local reorientation of the rolls in what is called the zigzag instability. For the rolls that fall within the wave-number band stable to these perturbations, however, there will eventually be a secondary instability to a qualitatively different type of flow when the Rayleigh number is increased to a large enough value. The "zoo" of possibilities includes "bimodal instability," which produces oscillations traveling along the roll boundaries, and "skew-varicose" deformation of the rolls. The Prandtl number Pr becomes very important in determining the instability mode for finite-amplitude rolls because, as remarked above, Pr measures the relative importance of the nonlinear terms $(\mathbf{u}\cdot\nabla)\mathbf{u}$ and $(\mathbf{u}\cdot\nabla)\theta$. The result of this analysis is a stability balloon (in this case called the Busse balloon) in a parameter space whose coordinates are (k,Pr,Ra). For coordinates outside this balloon, steady convection rolls become unstable to the variety of perturbations discussed above. A discussion of the nonlinear pattern selection, the stability balloon, and the various

Table 1. Critical parameters for various conditions applied to convection in horizontal fluid layers.

Boundary condition	Ra_c	k_c
Free-free	$27\pi^4/4 \approx$ 657.51	$\pi/\pi/\sqrt{2} \approx 2.221\,4$
Rigid-free	1100.6	2.68
Rigid-rigid	1707.76	3.116 32

types of modes that destabilize convection rolls is found in the review by Busse (1985).

When the Rayleigh number is increased further (beyond secondary instability), the flow eventually becomes time dependent. Depending on the Prandtl number, this might start out as periodic in time, but for further increases of *Ra* the flow becomes irregular or chaotic. Eventually, at large Rayleigh number, the motion is turbulent, characterized by strong three dimensionality and a wide range of spatial scales.

There are some important experimental wrinkles in the theory for convection near threshold. The first of these is that the side walls joining the top and bottom plates can exert an overriding influence on the pattern chosen by the convecting rolls (Koschmieder, 1974, 1975). This is called side-wall forcing and leads, for example, to concentric rolls in cylindrical containers. It also leads to a rounding of the transition into an imperfect bifurcation, with rolls first appearing near the walls. However, by careful thermal bonding of the side walls to the end plates, horizontal temperature inhomogeneities induced by the side walls may largely be eliminated. There is still an effect due to the tendency of convection rolls to minimize dissipation by aligning perpendicular to the walls; rolls in rectangular containers tend to align parallel to the short side of the container. However, in large–aspect-ratio systems, the patterns are no longer exclusively determined by the side-wall geometry. The resulting flows can have astonishing behavior: for Rayleigh numbers not far above threshold, irregular time dependence can appear. This is due to the nucleation, drifting, and annihilation of dislocations in the roll patterns, particularly evident in cylindrical containers in which the alignment of rolls perpendicular to walls is frustrated. The complex, slow irregular motion is an example of spatiotemporal chaos and is discussed further in Newell (1989) and Manneville (1990).

There are many variations on convection between horizontal plates. Extremely precise measurements are done on liquid helium (Behringer, 1985). The effects of rotation and of magnetic fields (for conducting fluids) are discussed in Chandrasekhar (1961) and continue to be investigated. Another important variation is convection in binary fluid mixtures. As a result of the possibility of thermally induced concentration gradients (the Soret effect), the instability mechanism can be significantly altered. Consequently, it becomes possible for the conducting state to make a transition to time-dependent flow in the form of traveling waves or standing waves; see Cross and Hohenberg (1993).

2.5 Electrohydrodynamic Instabilities in Liquid Crystals

Electric fields can excite convective motion in nematic liquid crystals. From an experimental point of view, this is an attractive system because

1. the control parameters (voltage and frequency) can be set very precisely;
2. the motion is easy to visualize thanks to the optical anisotropy of the liquid-crystal medium;
3. the gap between bounding planes can be quite small (20–100 μm), meaning that extremely large aspect ratios are possible and effects having to do with large-scale modulations and/or defects in the secondary-flow patterns are readily studied (Lowe and Gollub, 1985);
4. a preferred direction may be established for the secondary-mode wave vector, so that the system is nearly one dimensional;
5. fluctuations in the system couple strongly to visible light and may therefore be studied very sensitively by light-scattering experiments.

Theoretically, however, the dynamic equations are very complex, with several physical parameters describing the fluid and with strong interaction of several fields: electric, velocity, director, space charge, and current density. The material constants are not all precisely known. The medium is anisotropic, with dielectric constants $\epsilon_\perp$ and $\epsilon_\parallel$ and conductivities $\sigma_\perp$ and $\sigma_\parallel$ depending on direction relative to the applied electric field **E**.

Electrohydrodynamic instabilities are discussed in de Gennes (1974), Stephen and Straley (1974), and Dubois-Violette *et al.* (1978).

2.6 Centrifugal Instabilities

A situation that mimics a density stratification is an angular-momentum stratification in differentially rotating fluids. Imagine a set of curved streamlines, each with a radius of curvature r about some (perhaps approximate) common center—see Fig. 11(a). We adopt cylindrical coordinates and assume only an azimuthal velocity component $V(r)\hat{\theta}$ and a corresponding angular momentum per unit volume $L(r)=\rho V(r)r$ (here the density ρ is assumed constant). This curved flow will be sustained by a pressure gradient that provides the centrifugal acceleration of the fluid:

$$\frac{dp}{dr}=\frac{\rho V^2(r)}{r}. \tag{74}$$

Let the streamline of fluid at radius r_1 be displaced to a larger radius r_2, acquiring a new velocity V'. Ignoring viscosity, conservation of angular momentum requires that the angular momentum of the streamline at its new position equal its original angular momentum,

$$V'r_2=V(r_1)r_1. \tag{75}$$

Thus the new velocity $V'=v(r_1)r_1/r_2$. The flow is unstable if the corresponding centrifugal acceleration exceeds the local pressure gradient,

$$\frac{\rho V'^2}{r_2}>\left.\frac{dp}{dr}\right|_{r_2}. \tag{76}$$

Substitution for V' and for dp/dr at r_2 leads to the conclusion that instability will occur if

$$V^2(r_1)r_1^2>V^2(r_2)r_2^2 \tag{77}$$

or, more generally,

$$\frac{d}{dr}L^2(r)<0. \tag{78}$$

Thus curved flow of an inviscid fluid is unstable if the square of the angular momentum decreases with radius; this is Rayleigh's criterion for centrifugal instability. If instability occurs but there is a bounding wall that confines the flow within an outer radius, the flow is forced to overturn and fluid flows in helical paths along vortices whose axes are aligned in the azimuthal direction [Fig. 11(b)].

2.6.1 Flow between Rotating Cylinders: The Taylor-Couette System. The flow between two coaxial independently rotating cylinders (Fig. 12) has been studied in great detail because experiments can be done very precisely and theory can take advantage of the simple geometry (DiPrima and Swinney, 1985). Suppose that the inner-cylinder radius and angular velocity are r_1 and Ω_1 and the outer-cylinder radius and angular velocity are r_2 and Ω_2. (By convention we usually choose Ω_1 to be positive.) Then, assuming the length L is infinite, an analytic solution exists for the primary velocity, vorticity, and pressure fields:

$$\mathbf{U}=V(r)\hat{\theta}=(Ar+B/r)\hat{\theta}, \tag{79a}$$

$$\nabla\times\mathbf{U}=2A\hat{\mathbf{z}}, \tag{79b}$$

$$p=p_0+\int_{r_1}^{r}\rho\frac{V^2(r)}{r}\,dr. \tag{79c}$$

The no-slip boundary conditions $V(r_1)=\Omega_1 r_1$ and $V(r_2)=\Omega_2 r_2$ determine A and B:

$$A=-\frac{\Omega_1 r_1^2-\Omega_2 r_2^2}{r_2^2-r_1^2}=-\frac{\eta^2-\mu}{1-\eta^2}\Omega_1, \tag{80a}$$

$$B=\frac{(\Omega_1-\Omega_2)r_1^2 r_2^2}{r_2^2-r_1^2}=\frac{(1-\mu)\eta^2}{1-\eta^2}\Omega_1 r_2^2. \tag{80b}$$

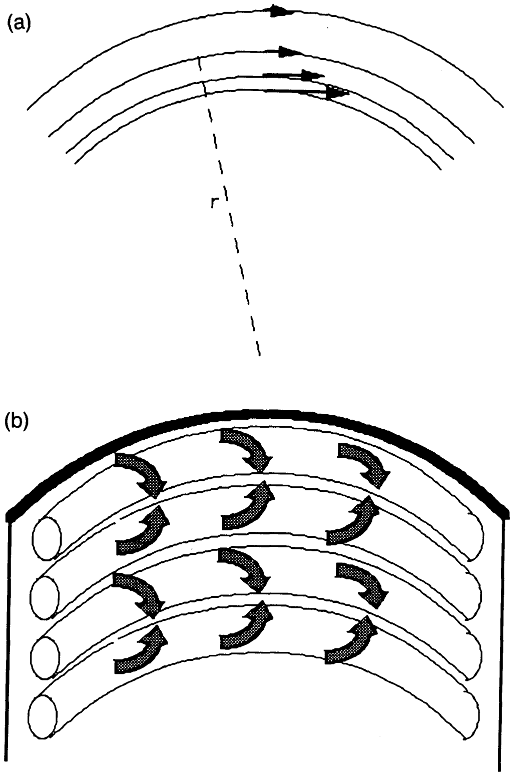

FIG. 11. Curved flows. (a) View in the r-θ plane of curved streamlines of the base flow. (b) Three-dimensional view of vortices formed by centrifugal instability when the flow is bounded by an outer wall.

The dimensionless constant $\mu \equiv \Omega_2/\Omega_1$ is the angular velocity ratio and $\eta \equiv r_1/r_2$ is the radius ratio.

Originally this system was used for measuring the dynamic viscosity $\rho\nu$ of a fluid. When the outer cylinder is rotated, the torque on the stationary inner cylinder is measured by suspending the inner cylinder from a torsion fiber. The torque G per unit length transmitted across a layer at any radius r is independent of r and is given by

$$G = \rho\nu r \frac{d(V/r)}{dr} \times 2\pi r \times r$$
$$= -4\pi\rho\nu B$$
$$= -4\pi\rho\nu \frac{\eta^2}{1-\eta^2} r_2^2(\Omega_1 - \Omega_2). \tag{81}$$

However, the linear relation between torque and differential angular velocity $\Omega_1 - \Omega_2$ is not observed for all angular velocities, suggesting that instability occurs. In fact, the Rayleigh criterion (see Sec. 2.7.2) applied to this flow says that the flow is always unstable if $\Omega_1\Omega_2 < 0$, that is, if the cylinders are oppositely rotating. If the cylinders are rotating in the same direction (with the convention that Ω_1 is positive), the Rayleigh criterion says that instability should occur for $\Omega_1 r_1^2 > \Omega_2 r_2^2$ or equivalently $\Omega_2/\Omega_1 < (r_1/r_2)^2$. If the outer cylinder is at rest, this condition is simply $\Omega_1 > 0$. These criteria are, however, for inviscid fluids. It is expected that viscosity stabilizes the flow, which is indeed the case.

In a landmark work, Taylor (1923) theoretically calculated the stability threshold for flow of a viscous fluid between independently rotating cylinders and found excellent agreement with his own experiments. Taylor assumed that the gap $d \equiv r_2 - r_1$ was small: $d/r_2 \ll 1$. He also confined his analysis to axisymmetric disturbances. Today, computers have made the generalization to arbitrary gap and nonaxisymmetric disturbances fairly straightforward. The base flow is perturbed by modes

$$\mathbf{u} = [u(r)\hat{\mathbf{r}} + v(r)\hat{\theta} + w(r)e^{i\pi/2}\hat{\mathbf{z}}]e^{i(kz+m\theta)}e^{st}, \tag{82a}$$

$$p = p(r)e^{i(kz+m\theta)}e^{st}. \tag{82b}$$

The resulting linear system leads to a dispersion relation $s = s(Re,k,m,\mu,\eta)$, and the marginality condition (real part of s vanishes) gives a family of curves $Re_{\text{marg}} = Re(k,m,\mu,\eta)$. Assuming that the speed ratio μ and radius ratio η are fixed at externally imposed values, the critical Reynolds number and wave numbers are found by minimizing over k and m. For co-rotating cylinders, stationary outer cylinder, and weakly counter-rotating cylinders ($\mu_{01} < \mu < \infty$, where $\mu_{01} < 0$) the critical mode has $m = 0$, is steady in time ($\omega = 0$), and takes the form of axisymmetric Taylor vortices. For counter-rotating cylinders with $\mu < \mu_{01}$, the critical modes are nonaxisymmetric and time periodic ($\omega \neq 0$); the resulting flows are either spiral vortices (waves traveling in the axial and azimuthal direction) or "ribbons" (standing waves in the axial direction, but traveling in the azimuthal direction). The point μ_{01} is a bicritical point where axisymmetric and nonaxisymmetric modes compete. (There are other bicritical points, $\mu_{m,m+1}$, where nonaxisymmetric modes m and $m+1$ compete.)

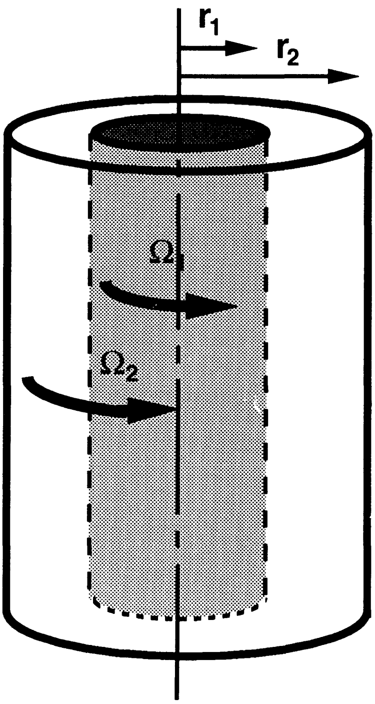

FIG. 12. Couette-Taylor geometry. Fluid flows in the annulus between two coaxial, independently rotating cylinders.

The Taylor-Couette system is remarkable for the fact that an incredible variety of qualitatively different flow patterns may be established by changing the two cylinder speeds and the radius ratio. Furthermore,

Table 2. Sequence of flow states between coaxial cylinders with the inner cylinder rotating, for radius ratio $\eta=0.88$. The imposed rotation is expressed in terms of the ratio of the Reynolds number *Re* to the critical value Re_c at which Taylor vortices first appear.

Re/Re_c	Type
0–1	Base flow (Couette flow)
1–1.2	Taylor vortices
1.2–7.5	Wavy vortices–there may be transitions within this range to a different number *N* of wavy vortices within the finite axial length, as well as to a different number *m* of waves in the azimuthal direction.
7.5–9.5	Modulated wavy vortices with small-scale modulation
9.5–12.5	Modulated wavy vortices with large-scale modulation
12.5–19	Weakly turbulent vortices, still showing modulated waves
19–40	Turbulent vortices with fast regular oscillations embedded in them
40–700	Strongly turbulent Taylor vortices
>700	Strongly turbulent flow with no remnant of Taylor vortices.

nonuniqueness of solutions to the Navier-Stokes equations under identical imposed conditions is amply demonstrated in this system (Coles, 1965): many different states are obtained by arranging different histories of the flow development. The system has been most intensively studied for the case in which the outer cylinder is at rest, usually with radius ratio $\eta\approx0.88$. Table 2 shows the sequence of flow states that results for this radius ratio as the Reynolds number is steadily increased. The ranges of values of Reynolds numbers are only rough estimates, since the precise values depend on the exact radius ratio and on the history of the flow state. Note that five qualitatively different flows are found *before* turbulence occurs, in contrast to most open flows like pipe flow and boundary-layer flow. The weakly turbulent vortices have been particularly interesting, because experiments (Fenstermacher *et al.*, 1979; Brandstater and Swinney, 1987) have shown that the flow has the characteristics of temporal chaos (see Sec. 3 below).

When the outer cylinder is rotated as well, an even greater variety of flows is found. One of the most amusing is the "barber pole" or spiral turbulence found for counter-rotating cylinders. Turbulent spots similar to those in boundary layers or channel flow may also be found. Finally, for co-rotating cylinders, a variety of time-periodic flows results from the instability of Taylor-vortex flow. For a survey with many excellent photographs, see Andereck *et al.* (1986).

2.6.2 Other Types of Centrifugal Instability. From a practical point of view, one of the most important types of centrifugal instabilities occurs in boundary layers along concave walls. Instability leads to rolls aligned with the direction of flows. These are called Görtler vortices. If, instead, one has flow forced through a curved channel by a pressure gradient, the rolls that form (again along the direction of flow) are called Dean vortices. Finally, flow impinging on a stagnation point must be diverted through curved streamlines, from which centrifugal instabilities may arise. For these variations, see Drazin and Reid (1981).

2.7 Shear Flow

Many flows of practical interest, such as flow through pipes or flow around airfoils, are mainly in one direction (say $\hat{\mathbf{x}}$) but have gradients mainly in a perpendicular direction (say $\hat{\mathbf{y}}$), with the result that vorticity is mainly in the mutually perpendicular direction ($\hat{\mathbf{z}}$); see Fig. 13. Such flows are called shear flows and the bulk of the stability literature concerns them.

Four facts make the study of transition in shear flow very challenging:

1. Viscosity plays a dual role. On one hand, it damps disturbances and thus restricts the family of unstable modes. However, it may

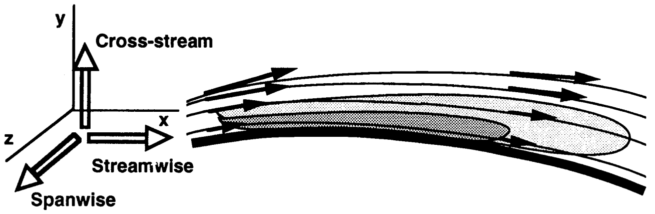

FIG. 13. Two-dimensional shear flow. The flow is mainly in the streamwise direction, the gradients mainly in the cross-stream direction, and the vorticity mainly in the spanwise direction. The streamlines shown are assumed to lie in the *xy* plane. Shaded regions show the gradation of the magnitude of the vorticity, which to first approximation is homogeneous in the streamwise direction.

also transport momentum into some modes and thereby become the *cause* of instability.

2. Transition often occurs well before linear-stability theory predicts the growth of infinitesimal disturbances. This is partly explained by the finding that many of the instabilities are subcritical bifurcations, so that there is a range of Reynolds numbers for which the flow is stable to infinitesimal disturbances but unstable to disturbances above some threshold (possibly quite small in practice.) Furthermore, it has been conjectured that disturbances, even though damped, may persist long enough to provide an altered base flow that can support instability even though the original base flow is stable. Thus environmental disturbances play a heightened role in bringing about instability. The term *receptivity* denotes the mechanisms by which normal modes of the flow are coupled to these external disturbances; examples include fluctuations in the free-stream vorticity, pressure disturbances transmitted through the walls or back up the flow from downstream obstacles, and wall-surface roughness.
3. Even when external disturbances are sufficiently controlled to observe linear-instability modes, these are usually convectively unstable: growing disturbances are "washed" downstream and possibly out of the system before they are big enough to be observed. If unstable modes *are* observed, they have irregular phase (and consequent broad peaks in power spectra) unless the launching disturbance is externally imposed—e.g., by exciting the flow with a vibrating ribbon. Indeed, successful comparison with theory for shear flow instabilities relied on two important experimental tactics: strong suppression of background disturbances through the use of settling chambers, contraction regions, etc, and imposition of an overriding external disturbance with vibrating ribbons, loudspeakers, etc.
4. There is seldom a well-differentiated sequence of instabilities. The effective control parameters (e.g., Reynolds number) evolve spatially. One band of normal modes becomes "detuned" as another is excited, so that there may be a spatially dependent competition of instability modes. A further complication is the possibility that growing secondary flows become unstable even before they equilibrate (Bayly *et al.*, 1988).

Nonetheless, careful control and interpretation of experiments shows some success for the basic stability analysis. In the following, we will first discuss a basic mechanism for instability in inviscid shear flow. Then we show how the linear stability of a wide variety of base flows can be approached within the common formalism of the parallel-flow approximation, first assuming zero viscosity and then accounting for the subtle effects that viscosity can have in flows that are inviscidly "stable." Finally, there will be a brief catalog of specific flows and some discussion of the transition to turbulence, which is largely inevitable once the initial instability sets in.

2.7.1 Kelvin-Helmholtz Instability. Consider the external case [Fig. 14(a)] of a region of inviscid fluid moving with uniform velocity $U_1\hat{\mathbf{x}}$ beneath a region of the same fluid moving with uniform velocity $U_2\hat{\mathbf{x}}$. This is an idealized, infinitesimally thin shear layer. The growth of disturbances $y=\zeta_0\, e^{i(k_x x+k_z z)+st}$ may be treated in the same way as waves on flat fluid interfaces, except that there is no density difference and no surface tension. (Also, the y

and z axes have been exchanged, in keeping with convention for shear flows.) Linear analysis gives the dispersion relation

$$s = \pm k_x \frac{U_2 - U_1}{2} - i k_x \frac{U_1 + U_2}{2}. \tag{83}$$

The result is a traveling wave with an x component of the phase velocity equal to the mean velocity of the fluid layers; for *any* velocity difference, one of the modes is unstable, the wave grows, and the shear layer deforms. The streamlines of the deformed layer seen from the frame moving with the phase velocity (which can be the laboratory frame if $U_1 = -U_2$) form a Kelvin cat's-eye pattern [Fig. 14(c)].

A useful point of view of the instability (Sherman, 1990) is to consider the shear layer, which is a vortex sheet, as a linear array of vortices [Fig. 14(b)] with circulation $\Gamma = (U_1 - U_2)a$ and undistorted positions $(x_{n0}, y_{n0}) = (na, 0)$, where $-\infty < n < \infty$ and a is the uniform spacing of the vortices along the x axis. The vortices interact through the motion they induce in the surrounding fluid:

$$\begin{pmatrix} \dfrac{dx_m}{dt} \\ \dfrac{dy_m}{dt} \end{pmatrix} = \frac{\Gamma}{2\pi} \sum_{\substack{n=-\infty \\ n \neq m}}^{\infty} \frac{1}{(x_m - x_n)^2 + (y_m - y_n)^2} \times \begin{pmatrix} -(y_m - y_n) \\ x_m - x_n \end{pmatrix}. \tag{84}$$

There is no motion of the vortices as long as they remain exactly on the x axis and are uniformly spaced. If their initial positions are perturbed by $(\delta x_{n0}, \delta y_{n0}) = (0, \epsilon \cos kx)$, then it is found that the array curls up into an intricate pattern of billows very much like that seen in real shear layers—see Sherman (1990), pp. 441–444, and references therein for results and important practical details of such computations.

For completeness, we include the dispersion relation when the two regions have different densities ρ_1 and ρ_2 and there is an interfacial tension γ:

$$s = \pm \left(\frac{k_x^2 \rho_1 \rho_2 (U_1 - U_2)^2}{(\rho_1 + \rho_2)^2} - \frac{kg(\rho_1 - \rho_2) + \gamma k^3}{\rho_1 + \rho_2} \right)^{1/2} - ik \frac{\rho_1 U_1 + \rho_2 U_2}{\rho_1 + \rho_2}, \tag{85}$$

where $k \equiv (k_x^2 + k_y^2)^{1/2}$. There can be instability for some k_x and k_y if and only if

$$|U_1 - U_2| > U_c \equiv \left(\frac{2(\rho_1 + \rho_2)}{\rho_1 \rho_2} \right)^{1/2} \times [(\rho_1 - \rho_2) g \gamma]^{1/4}, \tag{86}$$

with the first instability occurring (as the velocity difference is increased) for a wave number equal to

$$k_x = k_c \equiv [(\rho_1 - \rho_2) g / \gamma]^{1/2}, \quad k_y = 0. \tag{87}$$

Note that $1/k_c$ is just the capillary length defined in Sec. 2.1 [following Eq. (45)].

2.7.2 Parallel and Nearly Parallel Flows.

Many flows may be approximated as being parallel, with the fields given by

$$\mathbf{U} \approx U(y)\hat{\mathbf{x}}, \quad P \approx P(x) \quad \text{(two-dimensional base flow)} \tag{88a}$$

or

$$\mathbf{U} \approx U(r)\hat{\mathbf{z}}, \quad P \approx P(z) \quad \text{(axisymmetric base flow).} \tag{88b}$$

Let us then restrict our attention to two-dimensional flows. If viscosity is ignored and the geometry does not vary with x, it turns out that any profile $U(y)\hat{\mathbf{x}}$ and a constant pressure

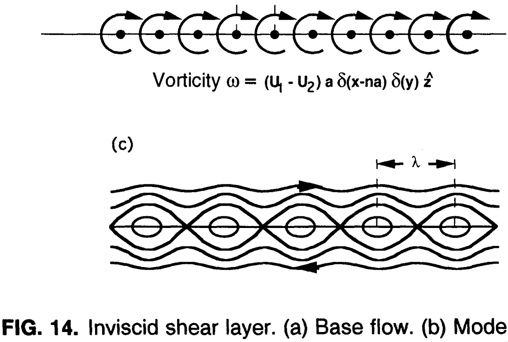

FIG. 14. Inviscid shear layer. (a) Base flow. (b) Model of base flow as an array of equally spaced vortices. (c) Kelvin cat's-eye pattern of streamlines for a disturbance of wavelength λ to the base flow, as seen in frame moving with phase velocity of the wave.

P_0 satisfy the resulting flow equations. In more general situations (like Fig. 13) and with viscosity included, we will assume that the streamwise variation and curvature effects may first be ignored: the base flow is considered to be "locally" parallel. Of course, for some flows (flows through uniform ducts), the base flow is exactly parallel. A classification of parallel and nearly parallel flows is shown in Fig. 15.

Consider a two-dimensional base flow $\mathbf{U}=(U(x,y),V(x,y),0)$. In the following, lengths are scaled by some characteristic length d—for example, the half-width $h/2$ of the channel for flow between parallel plates; velocity is scaled by a characteristic velocity U_0, e.g., the maximum velocity $U_{\max}$; the time scale is d/U_0 and the pressure scale is ρU_0^2. Assuming $V \ll U$ and $\partial U/\partial x \ll \partial U/\partial y$, the terms V and $\partial U/\partial x$ are neglected. Linear-stability analysis proceeds by considering the perturbed velocity and pressure fields: $U(y)\hat{\mathbf{x}}+\mathbf{u}(x,y,z,t)$, $P(x)+p(x,y,z,t)$. For infinitesimal disturbances $\mathbf{u}=(u,v,w)$ and p, the linearized equations are

$$\frac{\partial \mathbf{u}}{\partial t}+U(y)\frac{\partial \mathbf{u}}{\partial x}+v\frac{dU}{dy}\hat{\mathbf{x}}=-\nabla p+\frac{1}{Re}\nabla^2\mathbf{u}, \qquad (89a)$$

$$\nabla\cdot\mathbf{u}=0. \qquad (89b)$$

$\mathbf{u}(x,y,z)$ must vanish at the boundaries $y=y_1$ and y_2 (one or both of which may be at infinity). Homogeneity of the equations in x,z, and t suggests solutions of the form

$$(\mathbf{u},p)=(\hat{\mathbf{u}}(y),\hat{p}(y))e^{i(\alpha x+\beta z-\alpha ct)}. \qquad (90)$$

with four unknown functions of y: $\hat{\mathbf{u}}\equiv(\hat{u},\hat{v},\hat{w})$ and $\hat{p}$. Note that $-i\alpha c=\alpha c_i-i\alpha c_r$ is a new name for the eigenvalues $s=\sigma+i\omega$ used in previous problems (this new notation is customary in shear flow literature). Squire's transformation,

$$\tilde{\alpha}\equiv(\alpha^2+\beta^2)^{1/2},\quad \tilde{c}\equiv c,\quad \tilde{\alpha}\tilde{u}\equiv\alpha\hat{u}+\beta\hat{w},$$
$$\tilde{v}\equiv\hat{v},\quad \frac{\tilde{p}}{\tilde{\alpha}}\equiv\frac{\hat{p}}{\alpha},\quad \tilde{\alpha}\widetilde{Re}\equiv\alpha Re, \qquad (91)$$

permits the system of equations to be reduced to an equivalent two-dimensional problem solving for $\tilde{\mathbf{u}}\equiv(\tilde{u},\tilde{v},0)$ and $\tilde{p}$. For every three-dimensional disturbance satisfying the eigenvalue relation $f(Re,\alpha,\beta,c)=0$, there is a two-dimensional disturbance satisfying $f(\alpha Re/\tilde{\alpha},\tilde{\alpha},0,c)=0$. If the three-dimensional disturbance becomes marginally stable ($c_1=0$) for α and Re, then the corresponding two-dimensional disturbance is marginally unstable for $\tilde{\alpha}$ and $\widetilde{Re}=\alpha Re/\tilde{\alpha}$. Since $\alpha<\tilde{\alpha}$, we have $\widetilde{Re}<Re$, that is, the two-dimensional disturbance becomes unstable at lower Reynolds number. This important result, that the earliest instability occurs for two-dimensional modes, is known as Squire's theorem.

Having reduced the problem back to two dimensions (and dropping the tilde notation), we write u and v in terms of a stream function ψ:

$$u=\frac{\partial\psi}{\partial y},\quad v=-\frac{\partial\psi}{\partial x}, \qquad (92)$$

where

$$\psi=\phi(y)e^{i\alpha(x-ct)}. \qquad (93)$$

Then, defining $D\equiv d/dy$, some manipulation yields the Orr-Sommerfeld equation:

$$(U-c)(D^2-\alpha^2)\phi-U''\phi+\frac{i}{\alpha Re}(D^2-\alpha^2)^2\phi=0. \qquad (94)$$

The boundary conditions are

channel, $\alpha\phi=D\phi=0$ at $y=\pm 1$; (95a)

boundary layer, $\alpha\phi=D\phi=0$ at $y=0,\infty$; (95b)

mixing layer, jets, wakes,

$\alpha\phi=D\phi=0$ at $y=\pm\infty$. (95c)

The pressure $p(x,y,t)$ satisfies Poisson's equation:

$$\nabla^2 p=-2U'\frac{\partial v}{\partial x}. \qquad (96)$$

Letting $p(x,y,t)=\hat{p}(y)e^{i(\alpha x+\beta z-\alpha ct)}$ and using $v=-(\partial/\partial x)[\phi(y)e^{i\alpha(x-ct)}]$ gives

$$(D^2-\alpha^2)\hat{p}=-2\alpha^2U'\phi. \qquad (97)$$

The Orr-Sommerfeld equation [Eq. (94)], though linear, has been "notoriously difficult to solve" (Panton, 1984). Before discussing this further, we consider the inviscid limit $Re\to\infty$. The result is the Rayleigh equation:

$$(U-c)(D^2-\alpha^2)\phi-U''\phi=0. \qquad (98)$$

It must be noted, however, that viscosity can act as a singular perturbation. By taking the limit $Re\to\infty$, the highest-order derivatives have been dropped. Thus there must be fewer

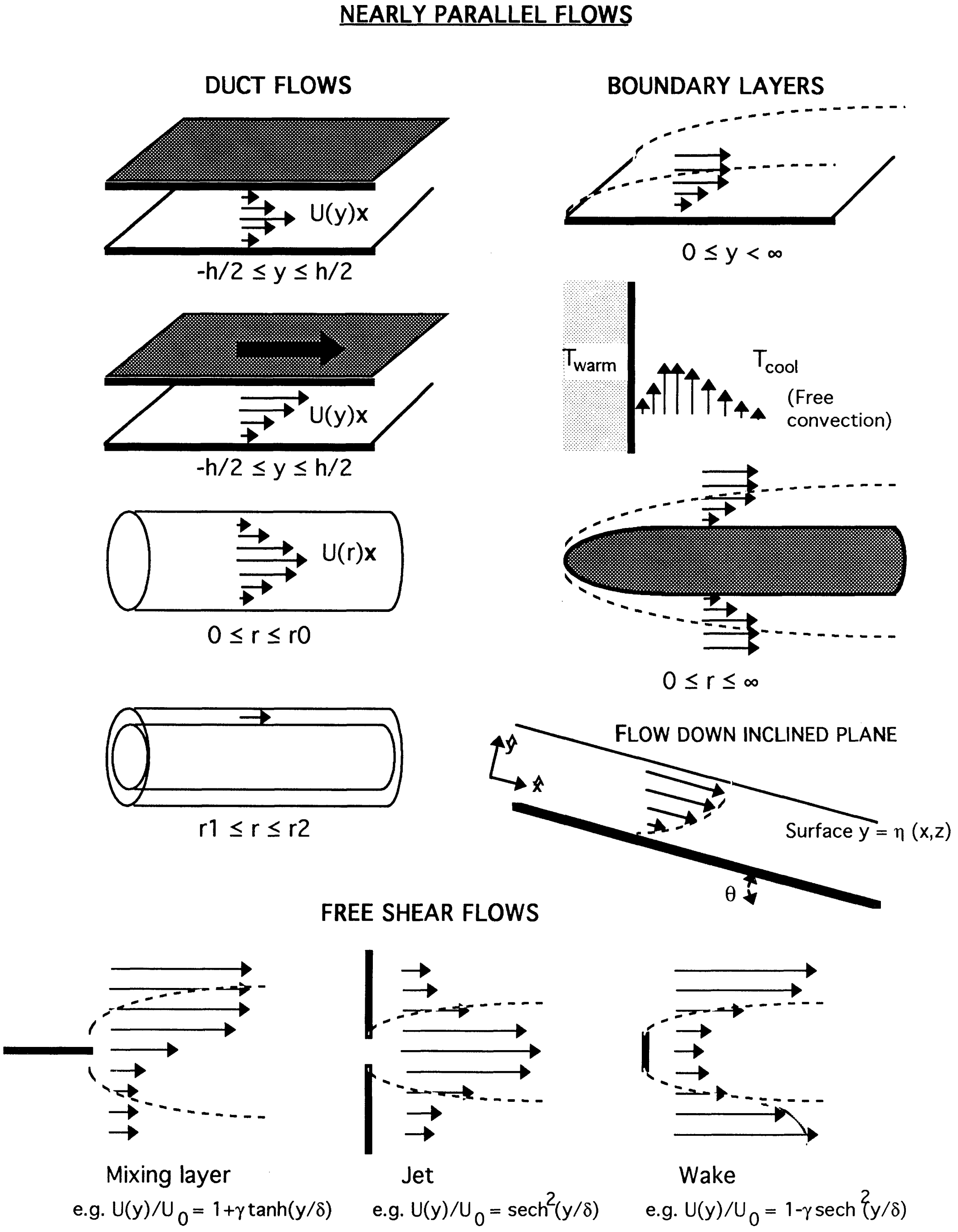
NEARLY PARALLEL FLOWS
DUCT FLOWS
U(y)x
-h/2 ≤ y ≤ h/2
U(y)x
-h/2 ≤ y ≤ h/2
U(r)x
0 ≤ r ≤ r0
r1 ≤ r ≤ r2
BOUNDARY LAYERS
0 ≤ y < ∞
Twarm
Tcool
(Free convection)
0 ≤ r ≤ ∞
FLOW DOWN INCLINED PLANE
y
x
Surface y = η (x,z)
θ
FREE SHEAR FLOWS
Mixing layer
e.g. U(y)/U0 = 1+γtanh(y/δ)
Jet
e.g. U(y)/U0 = sech²(y/δ)
Wake
e.g. U(y)/U0 = 1-γsech²(y/δ)

FIG. 15. The wide variety of nearly parallel flows whose instability may be treated with the Orr-Sommerfeld equation (or its axisymmetric counterpart).

boundary conditions for the Rayleigh equation: $\phi = 0$ at the boundaries (corresponding to vanishing normal velocity component). More subtle consequences of ignoring the viscous term will be discussed below. However, the inviscid equation itself yields important insights. The most important of these is Rayleigh's stability criterion: a necessary condi-

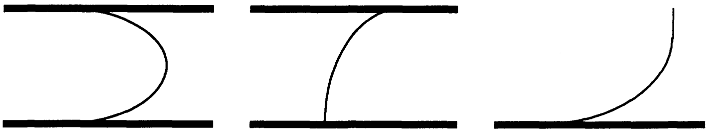

Inviscidly stable: no inflection point

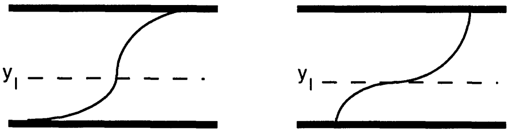

Inviscidly stable: $U''(y)[U(y)-U(y_I)] > 0$

Possibly unstable: $U''(y)[U(y)-U(y_I)] < 0$

FIG. 16. Classification of flow profiles according to Rayleigh's and Fjørtoft's inviscid stability criteria.

tion for a flow to be inviscidly unstable ($c_i > 0$) is that there be an inflection point in the base flow profile: $U''(y)$ must change sign for some y within the flow. Further necessary conditions, Fjørtoft criteria, are

1. $U''(y_I) = 0$ at the inflection point y_I must correspond to a local *maximum* in the magnitude of the vorticity $|U'(y_I)|$ rather than a minimum;
2. $U''(y)[U(y) - U(y_I)] < 0$ somewhere in the flow.

An evaluation of various flow profiles in terms of these conditions is shown in Fig. 16. It should be stressed that these are *necessary* conditions: if the conditions *are* met, this is no guarantee that the profile is unstable.

Some actual flow profiles (for example, mixing layers, wakes, jets, and boundary layers with adverse pressure gradients) satisfy Rayleigh's and Fjørtoft's criteria as candidates for being unstable in the inviscid limit. It is often the case that such flows do in fact become unstable over wave-number ranges in accord with results from Rayleigh's equation, and that the actual viscous eigenmodes with positive growth rate may be determined as a regular perturbation of the inviscid eigenmodes with positive growth rate (expanding in powers of $1/Re$). In such flows, viscosity acts mainly to delay instability.

Other flows, particularly duct flows and boundary layers, are predicted to be stable in the inviscid limit. Observation shows, however, that instability does occur, and for such flows, viscosity can play a subtle role in extracting energy from the mean flow to feed a growing eigenmode. The modes brought about by this mechanism are usually referred to as Tollmien-Schlichting waves, after the scientists who computed approximate solutions for the case of the Blasius boundary layer (the boundary layer over a flat plate placed parallel to a free stream with zero streamwise pressure gradient—see FLUID DYNAMICS, EQUILIBRIUM). The accurate solution of the Orr-Sommerfeld equation in such cases has been an extremely delicate task (both analytically and numerically) that has gone on for more than 70 years. The linear-stability results for various inviscidly "stable" flows are summarized in Table 3.

In reality, however, all of these systems undergo transition much sooner, especially when no extra precautions are taken to eliminate disturbances from the flow. In his classic experiments on transition in pipe flow, Osborne Reynolds found laminar flow up to $Re = 13\,000$. Typical transition Reynolds numbers for pipe flow are 2000, although careful experiments have maintained laminar flow up to $Re = 50\,000$.

Table 3. Critical Reynolds numbers for flows that are inviscidly stable. For the boundary layer, δ^* is the displacement thickness (Panton, 1984). The planes bounding Poiseuille flow are separated by distance h while the pipe has radius r_0. The characteristic velocity U is the maximum velocity of the base flow in each case.

Type	Re	Re_c
Blasius boundary layer	$U\delta^*/\nu$	520
Channel (plane Poiseuille) flow	$Uh/2\nu$	5772
Plane Couette flow	Uh/ν	∞
Pipe flow	Ur_0/ν	∞

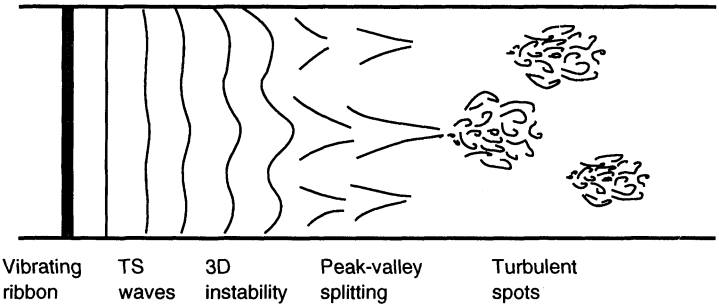

FIG. 17. Schematic of transition in boundary layers or plane channel flow.

Several classic experiments have been done on shear-flow transition. For boundary-layer transition, see Schubauer and Skramstad (1947) and Klebanoff *et al.* (1962). For plane channel flow, see Nishioka *et al.* (1975). In these experiments, free-stream disturbances were kept less than 0.1%, and the Tollmien-Schlichting (TS) waves are excited by vibrating ribbons placed in the flow. The general picture of transition that has emerged from these and other experiments is shown schematically in Fig. 17. After they are excited from the base flow, two-dimensional TS waves become unstable to three-dimensional disturbances generating streamwise vorticity; measurements show a "peak–valley" spanwise (z) variation of the intensity of the velocity fluctuations about the base flow. Associated with this are pronounced Λ-like structures in the visualized flow. Eventually, near the peaks, rapid fluctuations in velocity are observed, and there is a local burst of turbulence (Emmons spot). These turbulent spots or bursts eventually fill up the whole region and the flow completes its transition to turbulence.

For further discussion of shear-flow transition, see the textbooks and monographs in the reading list. Also, see the reviews by Bayly *et al.* (1988), Herbert (1988), Maslowe (1985), and Morkovin (1985).

3. DYNAMICAL SYSTEMS AND CHAOS APPLIED TO FLUID DYNAMICS

Fluid flows involve an infinite number of degrees of freedom, but it appears that there are cases where a dynamical description is accurately collapsed into the nonlinear interaction of a small finite number of degrees of freedom. This suggests the application of dynamical systems theory, with its powerful geometric classification of the solutions and bifurcations of evolution equations like those of the type

$$\frac{d\mathbf{x}}{dt}=\mathbf{f}(\mathbf{x},\mu). \tag{99}$$

In this context, $\mathbf{x}$ denotes a phase-space vector of n state variables $(x_1,\ldots,x_n)$ (for example, the amplitudes of critical modes near an instability), while μ is an m-component parameter-space vector $(\mu_1,\ldots,\mu_m)$, as in Sec. 1.2.8. A plot of a representative sample of possible trajectories $\mathbf{x}(t)$ in phase space is called a phase portrait; this is often more useful than an analytical solution since the qualitative behavior of whole families of solutions may be understood at once. A key feature of Eq. (99) is that it is deterministic: an initial condition $\mathbf{x}_0$ uniquely specifies the solution $\mathbf{x}(t)$ at future times.* However, some types of solutions are found to be effectively unpredictable: two arbitrarily close initial conditions will diverge exponentially fast in time, so that an inevitable small uncertainty in the initial conditions is quickly magnified to become a large uncertainty in the resulting motion. Solutions that have this sensitive dependence on initial conditions are called chaotic.

The existence of chaotic solutions to Eq. (99) has been known since Poincaré's studies in the 19th century. However, the possible application to fluid dynamics (and to physical systems in general) is relatively recent, dating from Lorenz's (1963) discovery of chaotic solutions in a severely truncated model of convection and the conjecture by Ruelle and Takens (1971) that such solutions could be related to turbulence. Since these works, a large amount of experimental evidence has supported the conclusion that some fluid flows may accurately be described as chaotic dynamical systems. Typically, though, these flows are in strongly confined geometries (or they are confined by a persistent structural feature of the flow, such as Taylor vortices). Since spatial variables are not present in Eq. (99), the irregular motion is further said to be temporal chaos. There may be a relation between such examples and fully turbulent flow, in the sense that turbulence also results from a deterministic set of equations (Navier-Stokes and continuity), but these equations depend on both space and time. Thus it is an unwarranted extrapolation to say that the origin and nature of turbulent flows are *explained* by chaos.

---

*Technically, up to some possibly finite limit T—for example, in cases where a solution diverges to infinity in finite time.

With this caveat, we proceed to describe how fluid flows may be characterized, when appropriate, in terms of dynamical systems theory. First, Eq. (99) is said to be a dissipative dynamical system if an arbitrary volume δV of points in phase space is mapped by the dynamics onto volumes steadily decreasing towards zero with time. Arbitrary initial conditions are thereby swept towards zero-volume subsets of phase space called attractors. The simplest type of attractor is a fixed point, for which $d\mathbf{x}/dt$ vanishes, that is, $\mathbf{f}(\mathbf{x},\mu)=0$. The primary fluid flow in systems subjected to steady forces is an example of a fixed-point solution. Fixed-point solutions may become unstable as μ is varied, possibly being replaced by other fixed points as attractors. Alternatively, the new attracting set could be a limit cycle, with the solution oscillating periodically in time. These are the result of a Hopf bifurcation. Limit cycles can in turn become unstable to a torus, characterized by quasiperiodic motions having two incommensurate frequencies. There can be a further succession of solutions that are "N-tori," with N incommensurate frequencies. Landau conjectured that turbulence arises from just this succession of bifurcations to modes with increasing numbers of independent frequencies (and phases). However, it has been shown (Newhouse *et al.* 1978) that arbitrarily near to a three-torus there is a much different type of solution upon which nearby trajectories exponentially diverge from one another. This is called a strange attractor, and the dynamics it represents is chaotic. Although the trajectories are stretched apart from one another, they are also folded back in a manner analogous to the stretching and folding of bread dough. For this to be possible, the phase space must be at least three dimensional. As a result, the attractor has the intricate geometric structure known as a fractal: it is a set of points in phase space that, unlike the tori, has a noninteger dimension. Lorenz's convection model had three equations that generated just such a solution for suitable values of the parameters; this is the well-known Lorenz attractor.

3.1 Scenarios for Evolution to Chaos

The above discussion outlines only one of the possible routes to chaos. This is the sequence stationary–periodic–quasiperiodic–chaotic. Experimental evidence of such a sequence includes the transitions in flow between rotating cylinders: Couette flow→Taylor vortex flow→periodic wavy vortex flow →modulated wavy vortex flow→turbulent Taylor vortices (Fenstermacher *et al.* 1979).

Another sequence is the period-doubling sequence of bifurcations of limit cycles, ending again in a chaotic phase portrait. An astonishing result is that such behavior may be understood, even quantitatively, by studying the bifurcation of attracting sets of arguments x in one-dimensional maps:

$$x(n+1)=f(x(n),\mu). \tag{100}$$

Here time has been replaced by a discrete index n. For a three-dimensional phase space, such maps can be extracted from phase portraits by means of Poincaré sections, which are cross sections of the phase portraits obtained by finding the intersections $(x(n),y(n),0)$ of a three-dimensional trajectory $(x(t),y(t),z(t))$ with some two-dimensional plane $z=0$ (only the intersections going in one direction across the plane are used). It may be that, to a good approximation, one of the coordinates x of the intersection will be a single-valued (but noninvertable) function of the value of the same coordinate for the previous intersection. (More generally, some combination of the coordinates may be found to be a single-valued function of the same combination for the previous intersection.) For small μ, all possible starting values of x will be attracted to a single stable fixed point x (which can be taken to be the origin by a coordinate transformation). Eventually, however, this can become unstable at a value

$u=\mu_1$ to a period-2 solution represented by the pair of points x_1^*,x_2^* that alternately map onto one another; these eventually become unstable to a mapping that sequentially maps four points $x_1^{**},x_2^{**},x_3^{**},x_4^{**}$ onto one another. This period-doubling cascade or subharmonic cascade proceeds over smaller and smaller intervals of μ up to some limit point μ_c where the map becomes chaotic. Furthermore, the sequence μ_1, μ_2, μ_3, etc. satisfies, for a large family of maps $f(x,\mu)$, the universal relation

$$\lim_{i\to\infty}\frac{\mu_i-\mu_{i-1}}{\mu_{i+1}-\mu_i}=\delta, \qquad (101)$$

with $\delta=4.669\,201\,609\,102\,990\,9$, known as the Feigenbaum constant. Experimental observation of such a cascade includes the study of small–aspect-ratio convection (Libchaber and Maurer, 1980; Libchaber *et al.* 1982). The cascade has been observed in other fluid experiments as well as many other nonlinear systems. It is usually only possible to follow a few period doublings because the required precision of μ increases very rapidly. Thus the experimental verification of Eq. (101) is necessarily imprecise.

Another route to chaos observed in experiments is called temporal intermittency, in which the motion remains very nearly periodic for an interval of time, then behaves irregularly for another interval, then becomes regular again, and so on. The fraction of time occupied by the irregular motion increases as some parameter μ is increased. This route to chaos was first explained by Pomeau and Manneville (1980), and an experimental example is the convection experiment of Bergé *et al.* (1980).

Unfortunately, it is seldom possible to predict *a priori* which route is followed for a given small–aspect-ratio experiment. Slightly different experimental conditions can lead to a different evolution of the dynamics.

3.2 Empirical Identification of Fluid Flows as Dynamical Systems

How can experimental data from real fluid flows be fitted into the highly abstract picture of dynamical systems theory? First, a time series of data may be recorded of, say, the velocity at a single spatial location. The square of the discrete Fourier transform of this time series yields a power spectrum that shows the dominant frequencies in the flow. Of course, a steady flow will yield data that are unchanging in time, and the corresponding spectrum will only show the level of external noise in the experiment and its instrumentation. This is useful to establish a noise "floor" for comparison to spectra obtained for time-varying flows. If the flow is periodic, there will be a dominant frequency ω_1 and its harmonics $2\omega_1$, $3\omega_1$, etc. If the flow is quasiperiodic, there will be two independent frequencies ω_1 and ω_2 along with their sums and differences $n_1\omega_1 \pm n_2\omega_2$. If the flow becomes chaotic, however, the spectrum will include broadband features (there may still be peaks in the spectrum). Though not definitive of chaos, the appearance of broadband power spectra as a parameter varies is a clue that the underlying dynamics has made a transition to chaos. A practical concern in producing power spectra is the prevention of aliasing: data must be sampled at least twice as fast as the highest frequency in the flow or else spurious components will appear in the spectra.

Of course, it is desirable to find the phase-space coordinates describing the instantaneous state of the fluid and to map trajectories of the fluid motion in phase space. For some flows, such as parametrically forced surface waves, the normal modes are readily identified and their amplitudes may be measured directly (albeit with considerable experimental ingenuity). In other flows, the identification of normal modes is theoretically intractable. Nevertheless, important features of geometry of the motion in phase space are preserved if coordinates are chosen that depend, in principle, on the mode amplitudes. One technique is the time-delay method (Takens, 1981): a phase space is constructed out of sequences of measurements of a single dynamical variable, such as one component of the velocity at a particular spatial location: $\mathbf{x}(t)=(u(t),u(t+\tau_1),u(t+\tau_2),\ldots,u(t+\tau_{n-1}))$. The choice of time delays τ_i is done either empirically or according to a variety of optimization procedures. The embedding theorem of Takens shows that the geometry of a d-dimensional attractor is faithfully represented in such a construction provided $n>2d+1$. For this procedure to be useful for further studies, the external noise level of the measurements must typically be less than 1%, and 2^{15} (32K) to 2^{18} (256K) data samples are desired.

Another method to identify a basis of modes and their amplitudes is through an adaptation of pattern-recognition techniques to fluid problems. Although developed by Lumley in the 1960s, this has only recently been applied to an understanding of fluid motions as dynamical systems. The approach is known by several names, including Karhunen-Love expansion or proper orthogonal decomposition. An ensemble of realizations of the velocity fields $\mathbf{u}(\mathbf{r})$ is obtained from experiment or from numerical simulations. (Explicit reference to time is left out to emphasize that a statistical ensemble of the spatial fields is sought; if the flow statistics are "stationary," these realizations can be samples of the flow field at different times.) An orthogonal basis $\phi_1(\mathbf{r}),\dots,\phi_n(\mathbf{r})$ is found empirically such that the projection of $\mathbf{u}(\mathbf{r})$ onto $\phi_1(\mathbf{r})$ captures, when averaged over the ensemble, the most energy; the projection of the remaining field onto $\phi_2(\mathbf{r})$ captures the next most energy, etc. Other criteria may be used, since the modes containing the most energy may not be the modes that govern the dynamics. In any case, a *given* time-dependent realization of the field $\mathbf{u}(\mathbf{r},t)$ may be projected onto this basis so that

$$\mathbf{u}(\mathbf{r},t)=\sum_{i=1}^{n} A_i(t)\phi_i(\mathbf{r}). \tag{102}$$

The sum is truncated when the energy in the remaining modes is comparable to the noise in the data. This procedure gives a phase-space trajectory of amplitudes $\mathbf{x}(t)=(A_1,A_2,\dots,A_n)$. This has been applied, for example, to the interpretation of moderately turbulent boundary layer in terms of dynamical systems (Aubry *et al.*, 1988). One drawback of the approach is that it requires huge amounts of data to generate phase portraits of sufficient length to be useful for quantitative studies of the attractors. However, it is a powerful technique and is reviewed in Berkooz *et al.* (1993).

3.3 Measurements of Low-Dimensional Chaotic Systems

Chaos and strange attractors are reviewed in CHAOTIC PHENOMENA. Further discussion, with reference to fluid dynamics problems, may be found in Bergé *et al.* (1986) and Manneville (1990). Here we simply comment on some practical matters.

Computation of fractal dimension requires sufficient experimental precision that an adequate "scaling region" may be found for the density of points on the attractor. That is, a range of radii ϵ about representative points on the attractor is sought so that, for example, the number of points within an n-dimensional hypersphere of radius ϵ about such points will accurately scale as ϵ^d across the range, where d is called the correlation dimension. The upper limit is set by the size of the attractor and the lower limit by the noise in the data. Typically, a resolution in the data greater than 10 bits is desired.

On the other hand, as the attractor dimension becomes larger, the space in which it is embedded likewise must be enlarged and it becomes more and more sparsely populated for a fixed number of data points. The number of points required to sample adequately the structure of the attractor blows up and there currently is a practical limit of $d \approx 4$ for reliable dimension computations.

Perhaps the most decisive indicator that the data come from a chaotic dynamical system is the existence of at least one positive Lyapunov exponent. This exponent measures the average rate at which nearby trajectories on a chaotic attractor diverge from one another. For example, phase-space trajectories from data for weakly turbulent Taylor vortex flow were found to have a positive Lyapunov exponent using the algorithm of Wolf *et al.* (1985).

4. SUMMARY

There are many other important classes of instabilities. For example, instabilities in wake flows have received a lot of attention since they seem to mix the properties of convective and absolute instability, depending on spatial location. See the review by Oertel (1990). For geophysical problems, instabilities affected by stratification and by Coriolis forces become very important. These are extensively discussed in the book by Pedlosky (1979); further discussion of stability problems in rotating flows is found in Greenspan (1968). It is likely that any identifiable fluid state (for example, intense vortex flow) has a retinue of instability phenomena and journals such as *Physics of Fluids*, *Journal of Fluid Mechanics*, *Physical Review E,* and *European Journal of Mechanics B/Fluids* continue to have a large share of articles on instability phenomena. The past decade has seen surges of interest in the application of dynamical systems theory to fluid problems and in nonlinear pattern formation; the extension of these areas into open flows remains a considerable challenge. Meanwhile, engineering concerns cause research in shear-flow transition to proceed unabated. Finally, the ultimate problem—turbulence—draws the attention of all researchers in nonequilibrium fluid flows. See TURBULENT FLOWS.

GLOSSARY

Absolute Instability: An instability in which a local disturbance results in a permanently altered flow at the origin of the disturbance even though the effects of the disturbance may also propagate away from the origin.

Amplitude Equations: A simplified description of the evolution of a new state of flow in terms of amplitudes of a small number of characteristic modes of instability.

Aspect Ratio: The ratio of the length of an apparatus in a direction perpendicular to the gradient in applied stresses to the width of the apparatus in the direction parallel to the gradient.

Attractor: A set in phase space that has the properties that (a) if an initial condition lies on the attractor, then the resulting trajectory remains on the attractor as time increases; (b) there is a neighborhood of the attractor such that initial conditions lying in the neighborhood have trajectories that converge on the attractor as time increases.

Basic Flow: An established state of flow that is examined for its stability with respect to perturbations.

Bifurcation: A branching, as a control parameter is varied, of the number of states of a system satisfying the governing equations; not all of the states are necessarily stable.

Boundary Layer: A thin layer in a flow near a solid wall, in which there are large gradients in flow velocity, temperature, etc.

Capillary Waves: Waves on a free surface of sufficiently short wavelength such that surface tension is the dominant restoring force.

Centrifugal Instability: Instability in flow with curved streamlines, due to the inability of the pressure gradient to provide sufficient centrifugal acceleration to fluid flowing in a streamline perturbed from its equilibrium position.

Chaos: Seemingly random motion of a system governed by deterministic equations, where the state quickly becomes unpredictable because of sensitive dependence on initial conditions.

Control Parameter: A parameter (for example, flow rate) that may be independently varied to control the state of a system.

Convective Instability: An instability in which a local disturbance results in a growing perturbation to the flow that travels away from the origin sufficiently fast that the flow at the origin returns to its original undisturbed state.

Critical Reynolds Number: The Reynolds number at which infinitesimal disturbances will render a basic flow unstable. (The same definition applies to other control parameters—for example, critical Rayleigh number.)

Critical Wave Number: The wave number [2π/wavelength] of the Fourier component of a disturbance that is the first to destabilize a basic flow state.

Dispersion Relation: The dependence of the growth rate and frequency of a mode of motion on the wave number of that mode.

Dissipative Dynamical System: A dynamical system in which arbitrary volumes of initial conditions in phase space reduce to sets of zero volume as time increases. A physical system in which there are frictional forces such as viscosity may also be called a dissipative dynamical system.

Dissipative Structures: Spontaneously ordered patterns in a nonequilibrium system that are sustained through a steady input of energy.

Dynamical System: A system of equations that describes the evolution of a state with time (usually according to a system of ordinary differential equations) or with increasing number of iterations of a mapping rule. The physical system modeled by such equations may also be considered a dynamical system.

Envelope Equations: A generalization of amplitude equations to allow variation of the amplitudes of the characteristic modes of instability over length scales long compared to the wavelength of the modes.

Equivariant: The property that equations of motion look the same after a change of coordinates arising from a symmetry operation such as translation, rotation, or reflection.

Fingering Instabilities: Instability in an interface or advancing front that causes the interface or front to deform into fingers.

Görtler Vortices: Vortices arising from the instability of a boundary layer near a curved wall; the axes of the vortices are in the streamwise direction.

Gravity Waves: Waves on a free surface, of sufficiently long wavelength such that gravity is the dominant restoring force.

Hele-Shaw Cell: An experiment configuration in which fluid is constrained to flow between two very close parallel plates.

Hopf Bifurcation: A bifurcation in which the new state of a system is characterized by an additional frequency of oscillation with time and the amplitude of the new state grows continuously from zero as a parameter is varied.

Hysteresis: The difference between the path along which a system evolves as a control parameter is increased and the path along which it evolves as the control parameter is decreased between the same two limiting values.

Imperfect Bifurcation: A smooth change in solution to the equations of motion (or in the state of a physical system) as a parameter is varied that is not a true branching from some original solution but would be a new branch if the corresponding physical system had perfect symmetry, was infinitely long, etc.

Inflectional Instability: The basic means by which many shear flows are unstable, in situations where the Rayleigh criterion is met: there is an inflection point in the plot of the streamwise velocity component versus a coordinate perpendicular to the flow direction.

Laminar Flow: A relatively simple state of fluid motion in which velocities remain either steady, periodic, or quasiperiodic in time. The flow is pictured as layers (laminae) of fluid sliding past one another, in contrast to turbulent flow in which there is a complex random eddying motion that is constantly stretching and bending elements of fluid.

Limit Cycle: An attracting set in phase space that represents a periodic solution to the equations of motion.

Linear-Stability Analysis: The solution of the equations of motion for small perturbations of a basic flow when all nonlinear terms in the perturbations are ignored. The solutions depend exponentially on either time or space: exponential growth indicates that the basic flow is unstable to the perturbations, while exponential decay indicates that the basic flow is stable.

Marangoni Convection: Motion induced in a fluid layer due to thermally induced spatial variation of the surface tension of the layer.

Marginal Stability Curve: Plot, as a function of wave number, of the Reynolds (or Rayleigh) number at which a Fourier component of a disturbance is at the threshold of growth.

Nonlinearity: Terms in the equations of motion that involve products of two or more of the dependent variables (for example, velocity components) or their derivatives; this includes products of a given dependent variable with itself.

Nusselt Number: Ratio of heat transport by a convective flow to the heat transport due to thermal conduction alone.

Open Flow: A flow in which mass is continuously passed into and then out of a system.

Parameter Space: The space of parameters selected to control or describe the flow; each coordinate represents the value of a different parameter.

Parametric Instability: Instability arising from the periodic variation of a parameter in the equations of motion that multiplies a term containing one of the state variables.

Phase Diagram: A map of regions in parameter space corresponding to various observed states of motion.

Phase Portrait: The trajectory in phase space traced by a dynamical system as time evolves, starting with one or more initial conditions. Such portraits give useful geometric meaning to the dynamics: steady states are points, periodic oscillations are closed curves, etc.

Phase Space: A space whose coordinates are the amplitudes of the various degrees of freedom of a dynamical system.

Prandtl Number: Ratio of the kinematic viscosity to the thermal diffusivity of a fluid.

Primary Flow: The unique state of flow produced when applied stresses are sufficiently small.

Quasiperiodic: Time dependence characterized by two or more independent frequencies that are not rationally related to one another (the ratio of the frequencies is not the ratio of two integers).

Rayleigh-Bénard Convection: Motion induced in a fluid layer heated from below due to the buoyancy of the warmer fluid relative to the cooler fluid.

Rayleigh Number: A dimensionless number describing the ratio of the effects of buoyancy to the effects of dissipation due to viscosity and thermal diffusion [see Eq. (2)]. $Ra \equiv \alpha g(T_2 - T_1)d^3/\nu\kappa$, where α is the thermal expansion coefficient, g is the acceleration due to gravity, $T_2 - T_1$ is the temperature difference that occurs across a gap d, ν is the kinematic viscosity, and κ is the thermal diffusivity.

Resonant Wave Interactions: Strong coupling of wave modes that satisfy the condition that the wave number of one of the modes is the vector sum of the wave numbers of the other modes and, correspondingly, the frequency of the one mode is the sum of the frequencies of the other modes.

Reynolds Number: A dimensionless number describing the ratio of inertia forces to viscous forces in a system [see Eq. (65)], $Re \equiv Ud/\nu$, where U is a characteristic velocity, d is a characteristic dimension across which changes in velocity of order U occur, and ν is the kinematic viscosity of the fluid.

Secondary Flow: The new, more complex part of a flow field that results after an instability—for example, Taylor vortices in flow between independently rotating cylinders.

Shear Flow: Flow whose dominant characteristic is a gradient in flow velocity in a direction perpendicular to the flow direction.

Side-Wall Forcing: Influence of bounding walls on the selection of patterns in secondary flows.

Slaving Principle: The idea that a few critical modes (those that have just become unstable) determine the dynamics of all the remaining degrees of freedom in a system.

Spatial Instability: An instability in which time-dependent forcing at one spatial location (or at one surface in the flow) leads to a disturbance that grows with spatial distance away from the origin.

Spatiotemporal Chaos: Chaotic motion that becomes uncorrelated in both space and time.

Stability Balloon: A confined region of stability in the space of parameters describing a flow (most readily visualized when there are three parameters, such as Rayleigh number, Prandtl number, and wave number in the case of convecting rolls).

Strange Attractor: An attractor whose geometry is fractal (has noninteger dimension) as a consequence of dynamics that stretches and folds the trajectories starting from nearby points on the attractor. Chaotic dynamics yields phase portraits that are strange attractors.

Stream Function: A scalar function of position and time from which the two velocity components of a two-dimensional flow field may be obtained as partial derivatives with respect to the two space variables: $u = \partial\psi/\partial y$ and $v = -\partial\psi/\partial x$. This formulation automatically guarantees the continuity equation $\partial u/\partial x + \partial v/\partial y = 0$. The stream function is constant along streamlines.

Streamline: A curve whose tangent is everywhere parallel to the fluid velocity.

Subcritical Bifurcation: A branching to a new solution whose amplitude increases from zero as a control parameter decreases from the point of bifurcation.

Supercritical Bifurcation: A branching to a new solution whose amplitude increases from zero as a control parameter increases from the point of bifurcation.

Symmetry-Breaking Bifurcation: A branching to a new solution that has less symmetry than the original solution.

Taylor-Couette System: The experimental system for studying flow between independently rotating coaxial cylinders. This system is characterized by a clearly distinguished sequence of bifurcations leading from the primary flow to turbulence. It is a much-studied system because of the high degree of success with which it can be modeled by mathematical analysis or numerical simulation.

Temporal Chaos: Chaotic motion that remains highly correlated in space although it quickly becomes unpredictable in time.

Temporal Instability: An instability in which an initial [time = 0] disturbance throughout the spatial domain of flow grows with time.

Tollmien-Schlichting Waves: Traveling-wave disturbances to a steady shear flow; the waves are characterized by well-defined wave number and frequency and are, at infinitesimal amplitude, the eigenmodes of the linear-stability problem for the steady shear flow.

Turbulence: A highly disordered state of fluid motion in which the velocity, pressure, and (when applicable) temperature appear to vary randomly in space and time. The flow is intrinsically three-dimensional, with local vortexlike motions that are continually being altered by stretching and reorientation of the vortices. Turbulent flows transport mass and momentum orders of magnitude more readily than nonturbulent (laminar) flows.

Turbulent Spot: A local burst of turbulence appearing in a boundary layer; analogous phenomena occur in channel flow and flow between counter-rotating cylinders. Sometimes these are referred to as "Emmons spots," after the person who observed such bursts in a shallow water channel.

Vortex Sheet: A hypothetical plane of discontinuity in the velocity of a parallel two-dimensional flow; the discontinuity may be viewed as the limit of an infinitely dense array of parallel vortices all lying in the same plane (see Fig. 14).

List of Works Cited

Ahlers, G. (1989), in: D. L. Stein (Ed.), *Lectures in the Sciences of Complexity,* Redwood City, CA: Addison-Wesley, p. 175.

Andereck, C. D., Liu, S. S., Swinney, H. L. (1986), *J. Fluid Mech.* **164**, 155–183.

Aubry, N., Holmes, P., Lumley, J. L., Stone, E. (1988), *J. Fluid Mech.* **192**, 115–173.

Bayly, B. J., Orszag, S. A., Herbert, T. (1988), *Annu. Rev. Fluid Mech.* **20**, 359–391.

Behringer, R. P. (1985), *Rev. Mod. Phys.* **57**, 657–688.

Benjamin, T. B. (1978a), *Proc. R. Soc. London Ser. A* **359**, 1–26.

Benjamin, T. B. (1978b), *Proc. R. Soc. London Ser. A* **359**, 27–43.

Benjamin, T. B., Ursell, F. (1954), *Proc. R. Soc. London Ser. A* **225**, 505–515.

Bensimon, D., Kadanoff, L. P., Liang, S., Shraiman, B., Tang, C. (1986), *Rev. Mod. Phys.* **58**, 977–1000.

Bergé, P., Dubois, M., Manneville, P., Pomeau, T. (1980), *J. Phys. (Paris) Lett.* **41**, L341–L345.

Bergé, P., Pomeau, Y., Vidal, C. (1986), *Order within Chaos,* New York: Wiley.

Berkooz, G., Holmes, P., Lumley, J. L. (1993), *Annu. Rev. Fluid Mech.* **25**, 539–575.

Block, M. J. (1956), *Nature* **178**, 650–651.

Brandstater, A., Swinney, H. L. (1987), *Phys. Rev. A* **35**, 2207–2220.

Busse, F. H. (1985), in: H. L. Swinney, J. P. Gollub (Eds.), *Hydrodynamic Instabilities and the Transition to Turbulence,* 2nd ed. Berlin: Springer, p. 97.

Chandrasekhar, S. (1961), *Hydrodynamic and Hydromagnetic Instability,* Oxford: Clarendon.

Chandrasekhar, S. (1969), *Ellipsoidal Figures of Equilibrium,* New Haven: Yale University Press.

Ciliberto, S., Gollub, J. P. (1985), *J. Fluid Mech.* **158**, 381–398.

Cloutman, L. D., Wehner, M. F. (1992), *Phys. Fluids A* **4**, 1821–1830.

Coles, D. (1965), *J. Fluid Mech.* **21**, 385–425.

Craik, A. D. D. (1985), *Wave Interactions and Fluid Flows,* Cambridge, England: Cambridge University Press.

Crawford, J. D. (1991), *Rev. Mod. Phys.* **63**, 991–1037.

Crawford, J. D., Knobloch, E. (1991), *Annu. Rev. Fluid Mech.* **23**, 341–387.

Cross, M. C., Hohenberg, P. C. (1993), *Rev. Mod. Phys.* **65** (to be published).

Davey, A. (1962), *J. Fluid Mech.* **14**, 336–368.

Davis, S. H. (1976), *Annu. Rev. Fluid Mech.* **8**, 57–73.

de Gennes, P. G. (1974) *The Physics of Liquid Crystals,* Oxford: Clarendon.

DiPrima, R. C., Swinney, H. L. (1985), in H. L. Swinney, J. P. Gollub (Eds.), *Hydrodynamic Instabilities and the Transition to Turbulence,* 2nd ed. Berlin: Springer, p. 139.

Donnelly, R. J., Glaberson, W. (1966), *Proc. R. Soc. London Ser. A* **290**, 547–556.

Dubois-Violette, E., Durand, G., Guyon, E., Manneville, P., Pieranski, P. (1978), in L. Liebert (Ed.), *Liquid Crystals,* Solid State Physics, Supplement 14, New York: Academic, pp. 147–208.

Drazin, P. G., Reid, W. H. (1981), *Hydrodynamic Stability,* Cambridge, England: Cambridge University Press.

Fenstermacher, P. R., Swinney, H. L., Gollub, J. P. (1979), *J. Fluid Mech.* **94**, 103–128.

Fermigier, M., Limat, L., Wesfried, J. E., Boudinet, P., Quilliet, C. (1992), *J. Fluid Mech.* **236**, 349–383.

Gaster, M. (1962), *J. Fluid Mech.* **14**, 222–224.

Goedde, E. F., Yuen, M. C. (1970), *J. Fluid Mech.* **40**, 495–511.

Gollub, J. P. (1991), *Physica D* **51**, 501–511.

Golubitsky, M., Stewart, I., Schaeffer, D. G., Langford, W. F. (1988), *Singularities and Groups in Bifurcation Theory*, Vol. 2, Berlin: Springer, p. 485.

Greenspan, H. P. (1968), *The Theory of Rotating Fluids*, Cambridge, England: Cambridge University Press.

Grosch, C. E., Salwen, H. (1978), *J. Fluid Mech.* **87**, 33–54.

Guckenheimer, J., Holmes, P. (1983), *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Berlin: Springer.

Haken, H. (1983), *Advanced Synergetics: Instability Hierarchies of Self-Organizing Systems and Devices*, Berlin: Springer.

Hammack, J. L., Henderson, D. M. (1993), *Annu. Rev. Fluid Mech.* **25**, 55–97.

Henderson, D. M., Hammack, J. L. (1987), *J. Fluid Mech.* **184**, 15–41.

Herbert, T. (1988), *Annu. Rev. Fluid Mech.* **20**, 487–526.

Huerre, P., Monkewitz, P. A. (1990), *Annu. Rev. Fluid Mech.* **22**, 473–537.

Iooss, G., Adelmeyer, M. (1992), *Topics in Bifurcation Theory and Applications*, Singapore: World Scientific.

Iooss, G., Joseph, D. D. (1990), *Elementary Stability and Bifurcation Theory*, 2nd ed. Berlin: Springer, Sec. VII.6.

Infeld, E., Rowlands, G. (1990), *Nonlinear Waves, Solitons and Chaos*, Cambridge, England: Cambridge University Press.

Joseph, D. D. (1976a), *Stability of Fluid Motions I*, Berlin: Springer.

Joseph, D. D. (1976b), *Stability of Fluid Motions II*, Berlin: Springer.

Joseph, D. D. (1985), in: H. L. Swinney, J. P. Gollub (Eds.), *Hydrodynamic Instabilities and the Transition to Turbulence*, 2nd ed. Berlin: Springer, p. 27.

Klebanoff, P. S., Tidstrom, K. D., Sargent, L. M. (1962), *J. Fluid Mech.* **12**, 1–34.

Koschmieder, E. L. (1974), *Adv. Chem. Phys.* **26**, 177–212.

Koschmieder, E. L. (1975), *Adv. Chem. Phys.* **34**, 109–133.

Landau, L. D. (1944), *C. R. Acad. Sci. U.R.S.S.* **44**, 311–314.

Libchaber, A., Maurer, J. (1980), *J. Phys. (Paris) Colloq.* **41-C3**, 51.

Libchaber, A., Laroche, C., Fauve, S. (1982), *J. Phys. (Paris) Lett.* **43**, L211–L216.

Lorenz, E. (1963), *J. Atmos. Sci.* **20**, 130–141.

Lowe, M., Gollub, J. P. (1985), *Phys. Rev. Lett.* **55**, 2575–2578.

Malkus, W. V. R., Veronis, G. (1958), *J. Fluid Mech.* **4**, 225–260.

Manneville, P. (1990), *Dissipative Structures and Weak Turbulence*, San Diego: Academic.

Maslowe, S. A. (1985), in: H. L. Swinney, J. P. Gollub (Eds.), *Hydrodynamic Instabilities and the Transition to Turbulence*, 2nd ed. Berlin: Springer, pp. 181 and 295.

McLachlan, N. W. (1947), *Theory and Application of Mathieu Functions*, Oxford: Clarendon.

Meshkov, E. E. (1969), *Izv. Akad. Nauk SSSR, Mekh. Zhidk. Gaza* **5**, 151 ([NASA Tech. Trans. F-13074 (1969)].

Miles, J. W., Henderson, D. (1990), *Annu. Rev. Fluid Mech.* **22**, 143–165.

Morkovin, M. V. (1985), *Guide to Experiments on Instability and Laminar-Turbulent Transition in Shear Flows* (Note for AIAA Professional Study Series: Instabilities and Transition to Turbulence), Cincinnati, Ohio: AIAA.

Newell, A. C. (1974), *Lectures in Applied Mechanics*, Vol. 15, Providence, RI: American Mathematical Society, p. 157.

Newell, A. C. (1989), in: D. L. Stein (Ed.), *Lectures in the Sciences of Complexity*, Redwood City, CA: Addison-Wesley, p. 107.

Newell, A. C., Passot, T., Lega, J. (1993), *Annu. Rev. Fluid Mech.* **25**, 399–453.

Newhouse, S., Ruelle, D., Takens, F. (1978), *Commun. Math. Phys.* **64**, 35–40.

Nishioka, M., Iida, S., Ichikawa, Y. (1975), *J. Fluid Mech.* **72**, 731–751.

Normand, C., Pomeau, Y., Velarde, M. G. (1977), *Rev. Mod. Phys.* **49**, 581–624.

Oertel, H. Jr. (1990), *Annu. Rev. Fluid Mech.* **22**, 539–564.

Panton, R. L. (1984), *Incompressible Flow*, New York: Wiley.

Pearson, J. R. A. (1958), *J. Fluid Mech.* **4**, 489–500.

Pedlosky, J. (1979), *Geophysical Fluid Dynamics*, Berlin: Springer.

Phillips, O. M. (1974), *Annu. Rev. Fluid Mech.* **6**, 93–110.

Pomeau, Y., Manneville, P. (1980), *Commun. Math. Phys.* **74**, 189–197.

Provansal, M., Mathis, C., Boyer, L. (1987), *J. Fluid Mech.* **182**, 1–22.

Richtmeyer, R. D. (1960), *Commun. Pure Appl. Math.* **13**, 297–319.

Richtmeyer, R. D. (1981), *Principles of Advanced Mathematical Physics*, Vol. II, New York: Springer, Chaps. 29–31.

Ruelle, D., Takens, F. (1971), *Commun. Math. Phys.* **20**, 167–192; **23**, 343–344.

Saffman, P. G. (1986), *J. Fluid Mech.* **173**, 73–94.

Saffman, P. G., Taylor, G. I. (1958), *Proc. R. Soc. London Ser. A.* **245**, 312–329.

Salwen, H., Grosch, C. E. (1981), *J. Fluid Mech.*

104, 445–465.

Schubauer, G. B., Skramstad, H. K. (1947), *J. Res. Natl. Bur. Stand.* **38**, 251–292.

Sharp, D. H. (1984), *Physica D* **12**, 3–18.

Sherman, F. S. (1990), *Viscous Flow*, New York: McGraw-Hill.

Stephen, M. J., Straley, J. P. (1974), *Rev. Mod. Phys.* **46**, 617–704.

Stuart, J. T. (1958), *J. Fluid Mech.* **4**, 1–21.

Stuart, J. T. (1960), *J. Fluid Mech.* **9**, 353–370.

Takens, F. (1981), in: *Lecture Notes in Mathematics*, Vol. 898, Berlin: Springer, p. 366.

Taylor, G. I. (1923), *Philos. Trans. R. Soc. London Ser. A* **223**, 289–343.

Taylor, G. I. (1950), *Proc. R. Soc. London Ser. A* **201**, 192–196.

Wang, T. G. (1988), *Adv. Appl. Mech.* **26**, 1–62.

Watson, J. (1960), *J. Fluid Mech.* **9**, 371–389.

Wolf, A., Swift, J. B., Swinney, H. L., Vastano, J. A. (1985), *Physica D* **16**, 285–317.

B. Further Reading

Acheson, D. J. (1990), *Elementary Fluid Dynamics*, Oxford: Clarendon Press, Chap. 9. A very readable introductory text.

Bergé, P., Pomeau, Y., Vidal, C. (1986), *Order within Chaos: Towards a Deterministic Approach to Turbulence*, New York: Wiley. Many insights on the practical details of interpreting experiments in terms of chaotic dynamical systems.

Betchov, R., Criminale, W. O. (1967), *Stability of Parallel Flow*, New York: Academic. Somewhat dated but useful monograph with emphasis on numerical approaches.

Chandrasekhar, S. (1961), *Hydrodynamic and Hydromagnetic Instability*, Oxford: Clarendon. A classical monograph emphasizing linear analysis of a wide variety of problems (including variations due to magnetic fields or rotation).

Collet, P., Eckmann, J. (1990), *Instabilities and Fronts in Extended Systems*, Princeton, NJ: Princeton University Press. Addresses many subtleties involved in spatially extended systems.

Craik, A. D. D. (1985), *Wave Interactions and Fluid Flows*, Cambridge, England: Cambridge University Press. A good complement to the text by Drazin and Reid, with emphasis on many applications of wave-interaction theory.

Cvitanovich, P. (1984), *Universality in Chaos*, Bristol, England: Adam Hilger. A volume of reprints of classic papers on chaos theory and experiment.

Drazin, P. G., Reid, W. H. (1981), *Hydrodynamic Stability*, Cambridge, England: Cambridge University Press. A comprehensive modern text on stability theory.

Haken, H. (1983), *Advanced Synergetics: Instability Hierarchies of Self-Organizing Systems and Devices*, Berlin: Springer. A wide-ranging treatment of the theory of nonequilibrium systems (not just fluids).

Hussaini, M. Y., R. G. Voigt (1990), *Instability and Transition*, Vols. 1 and 2. Berlin: Springer. A collection of papers stemming from a workshop on numerical modeling of transition phenomena.

Infeld, E., Rowlands, G. (1990), *Nonlinear Waves, Solitons and Chaos*, Cambridge, England: Cambridge University Press. Connects hydrodynamic stability theory to similar problems in plasma physics.

Iooss, G., Adelmeyer, M. (1992), *Topics in Bifurcation Theory and Applications*, Singapore: World Scientific. A recent summary and set of examples of the methods of bifurcation theory, including center manifold reduction and the use of symmetry.

Joseph, D. D. (1976), *Stability of Fluid Motions I and II*, Berlin: Springer. Discusses energy methods for establishing stability limits.

Koschmieder, L. (1993), *Bénard Cells and Taylor Vortices*, Cambridge, England: Cambridge University Press. Two important types of instability are discussed by a prominent experimentalist.

Kundu, P. K. (1990), *Fluid Mechanics*, San Diego: Academic, Chap. 11. A very complete intermediate fluid mechanics text, with detailed explanations of basic ideas and phenomena.

Lin, C. C. (1955), *The Theory of Hydrodynamic Stability*, Cambridge, England: Cambridge University Press. A classic monograph.

Manneville, P. (1990), *Dissipative Structures and Weak Turbulence*, San Diego: Academic. A thorough combination of physical insight, mathematics, and interpretation of experiments; provides a modern view of the physics of weakly turbulent flows.

Panton, R. L. (1984), *Incompressible Flow*, New York: Wiley, Chap. 22. An intermediate text with many useful insights on instabilities and turbulence, with emphasis on shear flows.

Saffman, P. G. (1992), *Vortex Dynamics*, Cambridge, England: Cambridge University Press. A detailed mathematical treatment of important classes of flow phenomena understood through the motion and interaction of finite regions of vorticity; discusses vortex instabilities (a topic omitted from this article).

Sherman, F. S. (1990), *Viscous Flow*, New York: McGraw-Hill, Chap. 13. Another intermediate text with many useful insights on instabilities and turbulence, with emphasis on shear flows and some discussion of numerical methods.

Swinney, H. L., Gollub, J. P. (Eds.) (1985), *Hydrodynamic Instabilities and the Transition to Turbulence*, 2nd ed. Berlin: Springer. A collection of articles on a wide range of stability phenomena and applications of dynamical systems theory to stability problems.

Tritton, D. J. (1988), *Physical Fluid Dynamics*, 2nd ed. Oxford: Clarendon, Chaps. 14, 17, and 18. An introductory text with considerable physical insight.

Van Dyke, M. (1982), *An Album of Fluid Motion*, Stanford, CA: Parabolic. A collection of some of the best fluid flow illustrations.

White, F. M. (1990), *Viscous Fluid Flow,* 2nd ed. New York: McGraw-Hill, Chap. 5. Another intermediate text with many useful insights on instabilities and turbulence, with emphasis on shear flows.

Whitham, G. B. (1974), *Linear and Nonlinear Waves,* New York: Wiley. A classic text on the mathematics of waves, with very general scope.

Yih, C.-S. (1969), *Fluid Mechanics,* New York: McGraw-Hill, Ch. 9. An intermediate text with discussion of some types of instabilities not found elsewhere.

32
Fourier and Other Mathematical Transforms

Ronald N. Bracewell†

32.1 Introduction 1151
32.2 The Fourier Transform 1151
32.3 Continuous versus Discrete Transforms 1152
32.4 Some Common Transforms 1155
32.5 The Laplace Transform 1155
32.6 Convergence Conditions 1158
32.7 Why Transforms are Useful 1159
32.8 Fields of Application 1160
32.9 The Hartley Transform 1162
32.10 The Fast Fourier Transform 1163
32.11 The Fast Hartley Algorithm 1164
32.12 The Mellin Transform 1165
32.13 The Hilbert Transform 1165
32.14 Multidimensional Transforms 1166
32.15 The Hankel Transform 1167
32.16 The Abel Transform 1167
32.17 Tomography and the Radon Transform 1168
32.18 The Walsh Transform 1170
32.19 The *Z* Transform 1171
32.20 Convolution 1172
32.21 Summary 1173
Glossary 1173
References 1174
Further Reading 1175

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

32.1 Introduction

Fourier analysis, which gained prominence from the work of J. B. J. Fourier in the early 1800s, led immediately to applications in mechanics and heat conduction. It also contributed to the advance of pure mathematics as regards the basic notions of limit, convergence, and integrability; the impact on mathematics and applied physics has continued to this day. Applications of transform methods were developed in connection with differential and integral equations and became very powerful; more recently, numerical analysis, aided by electronic computing, has added an extra dimension to the applied relevance of mathematical transforms and especially of the Fourier transform. The analytical and computational aspects are dealt with first; among applied examples, heat conduction, Fourier-transform spectroscopy, diffraction, sampled data, and tomography are mentioned.

When one looks for antecedents from which Fourier analysis might have evolved, they are not hard to find. Euler had published trigonometric series, and the sum to infinity, in such statements as

$$\sin x - \tfrac{1}{2}\sin 2x + \tfrac{1}{3}\sin 3x + \cdots = \tfrac{1}{2}x \tag{32.1}$$

Gauss analyzed motion in astronomical orbits into harmonics and indeed utilized the fast algorithm now favored for computing. Much earlier in Roman times, Claudius Ptolemy expressed motion in planetary orbits by the geometrical equivalent of trigonometric series and, according to Neugebauer (1983), the idea of epicycles has roots in Mesopotamian astronomy where the solar motion was matched by zigzag functions, rough approximations of the sinusoids to come.

32.2 The Fourier Transform

There are many transforms, each characterized by its own explicit operator, which we may call as **T**. The operand, or entity operated on, is a function such as $f(x)$, where x is a real variable ranging from $-\infty$ to ∞. The notation $\mathbf{T}\{f(x)\}$ signifies the outcome of applying the operator **T** to the function $f(x)$. To illustrate, the operation that converts a given function $f(x)$ to its Fourier transform, which is a different function $F(s)$, is as follows: ''Multiply the function $f(x)$ by $\exp(-i2\pi sx)$ and integrate with

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

respect to x from $-\infty$ to ∞". Applying this operation to $f(x) = \exp(-|x|)$, we find that $\mathbf{T}\{f(x)\} = F(s) = 2/[1 + (2\pi s)^2]$, which is the Fourier transform of $\exp(-|x|)$. The symbolic expression of the Fourier transform operation is

$$F(s) = \int_{-\infty}^{\infty} f(x) e^{-i2\pi sx} \mathrm{d}x \qquad (32.2)$$

It is apparent that any particular value of $F(s)$ (for example, $F(2)$, which equals 0.0126) takes into account the whole range of x; that is, the value depends on the shape of $f()$ as a whole, not on any single point. Thus the Fourier operation is quite unlike the operation that converts $f(x) = \exp(-|x|)$ to $\sin[\exp(-|x|)]$; the outcome of this latter operation is referred to as a *function of a function*, and the resulting values each depend only on the single value of x. When the result depends on the shape of $f(x)$ over part or all of the range of x, an entity such as $F(s)$ is called a *functional* of $f()$. The variable s is called the *transform variable* and may have a physical meaning; if so, its units will be cycles per unit of x. A short list of Fourier transforms for illustration is shown in Table 32.1.

In this list, rect x is the unit rectangle function (equal to unity where $|x| < 0.5$, elsewhere zero) and sinc $x = (\sin \pi s)/\pi s$. The last five lines are representative theorems of the form, "If $f(x)$ has Fourier transform $F(s)$, then [modification of $f(x)$] has transform [modification of $F(s)$]". Extensive lists of such transform pairs and theorems are available from the reference texts; the short list given would cover a sizable fraction of the analytic forms encountered in the literature.

With some transforms – the Abel transform is an example – each transform value depends on only a part of, not all of, $f()$; and with other transforms, the transform variable does not necessarily have a different identity (as s is different from x) but may have the same identity (Hilbert transform). The integral from $-\infty$ to x is a transform with both of the above restrictive properties.

Tab. 32.1 Selected Fourier transforms. The quantity a is a constant.

F(x)	F(s)
$e^{-\|x\|}$	$2/[1 + (2\pi s)^2]$
$\delta(x)$	1
$\cos(2\pi x/a)$	$\frac{1}{2}\delta(s + a^{-1}) + \frac{1}{2}\delta(s - a^{-1})$
rect x	sinc s
$e^{-\pi x^2}$	$e^{-\pi s^2}$
$e^{-\pi(x/a)^2}$	$\|a\| e^{-\pi(as)^2}$
$f(x/a)$	$\|a\|\ F(as)$
$f(x + a)$	$e^{i2\pi as} F(s)$
$f'(x)$	$i2\pi s F(s)$
Autocorrelation of $f(x)$	$\|F(s)\|^2$
$\int_{-\infty}^{\infty} f(x - u) g(u) \mathrm{d}u$	$F(s)G(s)$

All the transforms dealt with here are linear transforms, which are the commonest type; they all obey the superposition rule that $\mathbf{T}\{f_1(x) + f_2(x)\} = \mathbf{T}\{f_1(x)\} + \mathbf{T}\{f_2(x)\}$ for any choice of the given functions $f_1(x)$ and $f_2(x)$. An example of a nonlinear transformation is provided by $\mathbf{T}\{f(x)\} = a + bf(x)$, as may be tested by reference to the superposition definition; clearly the term *linear* in "linear transform" does not have the same meaning as in Cartesian geometry.

32.3
Continuous versus Discrete Transforms

Before defining the main transforms succinctly by their operations $\mathbf{T}$, all of which involve integration over some range, it is worth commenting on a numerical aspect. One could take the point of view, as is customary with numerical integration, that the desired integral is an entity in its own right; that the integral may on occasion be subject to precise evaluation in analytic terms, as with $F(s) = 2/[1 + (2\pi s)^2]$; and that if numerical methods are required, a sum will be evaluated that is an approximation to the desired integral. One would then discuss the desired degree of approximation and how to reach it. Now this is quite unlike the customary way of thinking about the discrete Fourier transform. What we evaluate is indeed a sum, but we consider the sum as precise and not as an approximation to an integral. There are excellent reasons for this. Meanwhile, the important thing to realize is that there are both a Fourier transform and a discrete Fourier transform, each with its own definition. The discrete Fourier transform operation is

$$F(\nu) = \frac{1}{N}\sum_{\tau=0}^{N-1} f(\tau)e^{-i2\pi\nu\tau/N} \qquad (32.3)$$

The word "discrete" is used in antithesis to "continuous," and in the cases discussed here, it means that an independent variable assumes integer values. To understand the discrete Fourier transform, which is exclusively what we compute when in numerical mode, it is best to forget the Fourier integral and to start a fresh. Instead of starting with a complex function $f(x)$ that depends on the continuous real variable x, we start with N data (complex in general, but often real) indexed by an integer serial number τ (like time) that runs from 0 to $N - 1$. In the days when FORTRAN did not accept zero as a subscript, summation from $\tau = 0$ caused much schizophrenia, but the mathematical tradition of counting from zero prevailed and is now unanimous. In cases where $f(\,)$ is a wave form, as it often is, the quantity τ can be considered as time that is counted in units starting from time zero. Clearly, N samples can never fully represent $\exp(-|x|)$, for two reasons: the samples take no account of the function where x exceeds some finite value, and no account is taken of fine detail between the samples. Nevertheless, one may judge that, for a given particular purpose, 100 samples will suffice, and the confidence to judge may be bolstered by trying whether acquisition of 200 samples significantly affects the purpose in hand. Numerical intuition as developed by hand calculation has always been a feature of mathematical work but was regarded as weak compared with physical intuition. Nowadays, however, numerical intuition is so readily acquired that it has become a matter of choice whether to deal with questions about the size of N by traditional analytic approaches. A new mix of tools from analysis, finite mathematics, and numerical analysis is evolving.

The discrete transform variable ν reminds us of frequency. If τ is considered as time measured in integral numbers of seconds, then ν is measured in cycles per second, and is indeed like frequency (cycles per second or Hertz), but not exactly. It is ν/N that gives correct frequencies in Hertz, and then only for $\nu \leq N/2$. Where ν exceeds $N/2$, we encounter a domain where the discrete approach conflicts with the continuous. When the Fourier transform is evaluated as an integral, it is quite ordinary to contemplate negative values of s, and a graph of $F(s)$ will ordinarily have the vertical $s = 0$ axis in the middle, giving equal weight to positive and negative "frequencies." (The unit of s is always cycles per unit of x; if x is in meters, s will be a spatial frequency in cycles per meter; if x is in seconds, s will be a temporal frequency in cycles per second, or hertz.) However, the discrete Fourier transform, as conventionally defined, explicitly requires the transform variable ν to range from 0 to $N-1$, not exhibiting negative values at all. There is nothing wrong with that, but persons coming from continuous mathematics or from physics may like to know that, when ν is in the range from $N/2$ to $N-1$, the quantities $N-\nu$ correspond to the negative frequencies familiar to them as residing to the left of the origin on the frequency axis. This is because the discrete transform is periodic in ν, with period N.

In the familiar Fourier series

$$p(x) = a_0 + \sum_{1}^{\infty}(a_\nu \cos 2\pi\nu x + b_\nu \sin 2\pi\nu x) \tag{32.4}$$

for a periodic function $p(x)$ of period 2π, the first term a_0 represents the direct current, zero frequency, or mean value over one period as calculated from

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi} p(x)\,\mathrm{d}x \tag{32.5}$$

Therefore, the first term $F(0)$ of the discrete Fourier transform is the average of the N data values. This is the reason for the factor $1/N$ in front of the summation sign in Eq. (32.3), a factor that must be remembered when checking. In practical computing, it is efficient to combine the factor $1/N$ with other factors such as calibration factors and graphical scale factors that are applied later at the display stage. The remaining Fourier coefficients, given by

$$a_\nu = \frac{1}{\pi}\int_0^{2\pi} p(x)\cos 2\pi\nu x\,\mathrm{d}x \tag{32.6}$$

$$b_\nu = \frac{1}{\pi}\int_0^{2\pi} p(x)\sin 2\pi\nu x\,\mathrm{d}x \tag{32.7}$$

are related to the discrete Fourier transform by $a_\nu - ib_\nu = F(\nu)$. The minus sign arises from the negative exponent in the Fourier kernel $e^{-i2\pi sx}$. The reason for the choice of the negative exponent is to preserve the convention that d/dt be replaceable by $+i\omega$ in the solution of linear differential equations, as when the impedance of an inductance L to alternating voltage of angular frequency ω is written as $+i\omega L$ (more usually as $j\omega L$).

How to decide whether the discrete Fourier transform is an adequate approximation to the Fourier transform is a very interesting question. But the question itself is open to challenge. If I am studying cyclicity in animal populations, perhaps seasonal influence on bird migration, I may start with 365 reports of how many birds were seen each day of the year. In such a case, and in many other cases, discrete data mean that the integrals, even though convenient, are themselves the approximations; the discrete Fourier transform, given N equispaced data, is a valid entity in its own right. However, unexpected discrepancies may arise over the choice of N, which may be very large or very small. Slow computing (N very large), unwanted sensitivity to measurement error (N very small), and aliasing are its negative consequences. Aliasing is the word for the following phenomenon. Measurements are made of some time-varying phenomenon at regularly spaced time intervals, perhaps temperature is recorded twice a day or perhaps speech samples are taken at a 10-kHz rate. Such data can represent harmonic components with period longer than one day or longer than 2×10^{-4} s, but cannot faithfully follow faster harmonic variation. The samples will not ignore the presence of such high frequencies, because the high-frequency variations will indeed be sampled, but the samples will be consistent with, and indistinguishable from, a long-period sinusoidal component that is not actually present. The imperfectly sampled component emerges under the alias of a lower, counterfeit frequency.

32.4
Some Common Transforms

As a convenient reference source, definitions of several transforms (Laplace, Fourier, Hartley, Mellin, Hilbert, Abel, Hankel, Radon) are presented in Table 32.2. When one has the transform, there is a way of returning to the original function in all the cases chosen. In some cases, the inverse operation $\mathbf{T}^{-1}$ is the same as the defining operation $\mathbf{T}$ (e.g., Hartley and Hilbert, which are reciprocal transforms), but the majority differ, as shown. In addition, examples of each transform are presented. These will be found to convey various general properties at a glance and may be helpful for numerical checking.

32.5
The Laplace Transform

A long and diverse history (Deakin, 1985) characterizes the Laplace transform, which was in use long before Laplace, but became known to current generations mainly through its pertinence to the linear differential equations of transient behavior in electricity and heat conduction. Many tough technological problems of electric circuits that arose in connection with telegraphy, submarine cables, and wireless, and related industrial-process problems of thermal diffusion, were cracked around the turn of the century, sometimes by novel methods such as those of Heaviside (1970), which were to be justified subsequently (Nahin, 1987) to the satisfaction of academic mathematics by systematic application of the Laplace transform. Heaviside is remembered for stimulating the application of the Laplace transform to convergence of series and for Maxwell's equations, the delta function, the Heaviside layer, impedance, nonconvergent series that are useful for computing, fractional order derivatives and integrals, and operational calculus.

Table 32.2 gives, as an example, the Laplace transform of $f(x) = \exp(-x - 1.5)\mathbf{H}(x + 1.5)$. The Heaviside unit step function $\mathbf{H}(x)$ jumps, as x increases from 0 to 1, the jump being where $x = 0$; one of its uses is as a multiplying factor to allow algebraic expression of functions that switch on. The transform of $f(x)$, which is easy to verify by integration, is $(\exp\ \ 1.5s)/(1 + s)$; the transform variable s may be complex but must lie among those numbers whose real parts are greater than -1 (otherwise the integral does not exist). It is rather cumbersome to exhibit the complex transform graphically on the complex plane, and so an illustration is omitted. To invert the transform requires integration on the complex plane along a semicircular contour with indentations if necessary to circumvent points where the integrand goes to infinity (poles). The

Tab. 32.2 Transform definitions, inverses, and examples. (Ticks are at unit spacing.

Name of transform	Nature of function domain variable	Example of function	Nature of transform variable	Nature of transform	Example of transform	Defining formula and the inverse
Laplace	Continuous real	$f(x) = e^{-x-1.5}H(x+1.5)$ 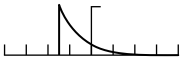	Continuous complex	Complex	$\dfrac{e^{1.5s}}{1+s}, -1 < \mathrm{Re}\, s$	$F_L(s) = \int_{-\infty}^{\infty} f(x)e^{-sx}\mathrm{d}x$ $\dfrac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} F_L(s)e^{xs}\mathrm{d}s$
Fourier	Continuous real	$f(x) = e^{-x-1.5}H(x+1.5)$ 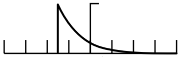	Continuous real	Complex	$F(s) = \dfrac{e^{i3\pi s}}{1+i2\pi s}$	$F(s) = \int_{-\infty}^{\infty} f(x)e^{i2\pi sx}\mathrm{d}x$ $f(x) = \int_{-\infty}^{\infty} F(s)e^{i2\pi xs}\mathrm{d}s$
Discrete Fourier	Discrete, real	$f(\tau) = e^{-\tau-1.5}H(\tau+1.5)$ 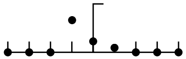	Discrete, real	Complex	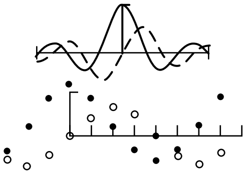	$F(v) = N^{-1}\sum_{\tau=0}^{N-1} f(\tau)e^{-i2\pi v\tau}$ $f(\tau) = \sum_{v=0}^{N-1} F(v)e^{i2\pi\tau v}$
Hartley[a]	Continuous real	$f(x) = e^{-x-1.5}H(x+1.5)$ 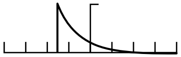	Continuous real	Real	$H(s) = \dfrac{\mathrm{cas}(-3\pi s) + 2\pi s\,\mathrm{cas}3\pi s}{1+4\pi^2 s^2}$	$H(s) = \int_{-\infty}^{\infty} f(x)\mathrm{cas}2\pi sx\mathrm{d}x$ $f(x) = \int_{-\infty}^{\infty} H(s)\mathrm{cas}2\pi xs\mathrm{d}s$
Discrete Hartley	Discrete, real	$f(\tau) = e^{-\tau-1.5}H(\tau+1.5)$ 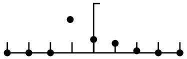	Discrete, real	Real	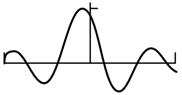	$H(v) = N^{-1}\sum_{0}^{N-1} f(\tau)\mathrm{cas}2\pi v\tau$ $f(\tau) = \sum_{v=0}^{N-1} H(v)\mathrm{cas}2\pi v\tau$
Mellin	Continuous real	$f(x) = e^{-x}H(x)$	Discrete, real	Real	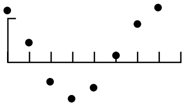	$F_M(s) = \int_{0}^{\infty} f(x)x^{s-1}\mathrm{d}x$

		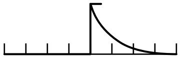			$\begin{array}{c\|cccc} s & 0 & 1 & 2 & 3 \\ \hline F_M(s) & 1 & 1 & 2 & 6 \end{array}$	$f(x) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+\infty} F_M(s)x^{-s}\mathrm{d}s$
Hilbert	Continuous real	$f(x) = e^{-\pi x^2}\cos 4\pi x$	Continuous complex	Real	$F_{\mathrm{Hi}}(x) = -e^{-\pi x^2}\sin 4\pi x$	$F_{\mathrm{Hi}}(x) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{f(x')\mathrm{d}x'}{x'-x}$
		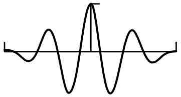			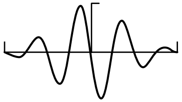	$f(x) = -\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{F_{\mathrm{Hi}}(x')\mathrm{d}x'}{x'-x}$
Abel	Continuous real	$f(r) = 1, r < 0.5$	Continuous complex	Real	$F_A(s) = \sqrt{1-s^2}, s < 1,$	$F_A(s) = 2\int_s^{\infty}\frac{f(r)r\mathrm{d}r}{\sqrt{r^2-s^2}}$
		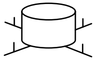			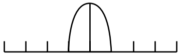	$f(r) = \frac{1}{\pi}\int_r^{\infty}\frac{f_A'(x)\mathrm{d}x}{\sqrt{x^2-r^2}}$
Hankel	Continuous real	$f(r) = 1, r < 0.5$	Continuous complex	Real	$F_{\mathrm{Ha}}(s) = jincs\frac{J_1(\pi s)}{2s}$	$F_{\mathrm{Ha}}(s) = 2\pi\int_0^{\infty} f(r)J_0(2\pi sr)r\mathrm{d}r$
		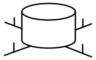			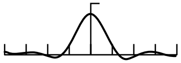	$f(r) = 2\pi\int_0^{\infty} F_{\mathrm{Ha}}(s)J_0(2\pi rs)s\mathrm{d}s$
Radon	2-D Continuous real	$f(x, y) = \delta(x-a)\delta(y)$	2-D, Continuous real	Real	$F_{\mathrm{Ra}}(x', \theta) = \delta(x' - a\cos\theta)$	$F_{\mathrm{Ra}}(x', \theta) = \int_{-\infty}^{\infty} f(x, y)\mathrm{d}y'$
		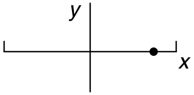			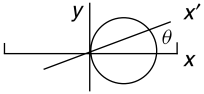	Complicated inverse

---

[a] The function cas$x = \cos x + \sin x$.

constant c in the inversion formula is to be chosen to the right of all poles.

To some extent, Laplace transforms were computed numerically, but more typically, development led to compilations of analytic transforms resembling the tables of integrals (Campbell and Foster, 1948; Erdélyi *et al.*, 1954). Programs for deriving the Laplace transform of the impulse response from electrical networks given diagrammatically are also available. Consequently, it is hardly ever necessary to analytically derive Laplace transforms today. The analytic solution of transients in electric circuits, a subject traditionally used for sharpening the minds of electrical engineers, is obsolescent because impulse responses and transfer functions have been concisely published (McCollum and Brown, 1965). Furthermore, the advent of integrated circuits has meant that inductance is seldom included in new designs, and those circuits containing more than two or three elements have become less common. Mature programs are also available for step-by-step integration of circuit differential equations.

On the numerical side, the Laplace transform has also been largely eroded by use of the Fourier transform. This is because angular frequency ω is a mathematically real quantity, and it ought to be possible to compute the behavior of an electrical, acoustical, or mechanical system without reference to a complex frequency $\omega - i\sigma$. Certainly, the Laplace transform is computable over its strip of convergence from any single slice therein. Nevertheless, practitioners of control theory find it convenient to assume the complex plane of s in terms of poles and zeros that are off the real frequency axis, and theirs is one tradition that keeps the complex plane alive; the convenience stems from the fact that the Laplace transform is analytic, and thus specifiable by its poles and isolated zeroes. There are problems that used to be handled by the Laplace transform, paying strict attention to the strip of convergence, because the Fourier integral did not converge; but these situations are now universally handled by Fourier methodology with the aid of delta-function notation for impulses and their derivatives, and no longer call for special treatment. When it comes to discrete computing, the impulse, and its associated spectrum reaching to indefinitely large frequencies, may in any case be forgotten. Thus, it has been wondered (Körner, 1988) ''whether the Laplace transform will keep its place in the standard mathematical methods course for very much longer,'' but it will never die out; a new balance between curricular segments will be struck.

32.6
Convergence Conditions

Much attention used to be given to the existence of the Fourier integral because of paradoxes with such wanted entities as $f(x) = 1, f(x) = \cos x$, or $f(x) = \delta(x)$, where $\delta(x)$ is the unit impulse at $x = 0$, none of which possessed a Fourier integral. Today, we reason as a physicist would, recognizing that a voltage waveform cannot have a value of 1 V forever, but must have turned on at some time in the past and will turn off at some time in the future. The finite-duration function does have a Fourier transform. We then consider a sequence of waveforms of longer and longer duration and the corresponding sequence of transforms, arriving at the concept of "transforms in the limit." This attitude has received mathematical respectability under the rubric of generalized functions (Lighthill, 1958) and is the basis for saying that the Fourier transform of $\delta(x)$ is 1 (while conversely the Fourier transform of 1 is $\delta(s)$). The elaborate conditions for the existence of a transform when generalized functions were excluded have thus lost interest. Even $\delta'(x)$ now has the indispensable transform $i2\pi s$; under the rules of analysis, $\delta'(x)$ was an unimaginable entity, certainly not qualifying as a function of x; it was considered as a commonplace dipole by physicists, and as a local load such as a moment applied at a point on a beam in mechanics.

The fact that the Laplace integral converged when the Fourier transform did not give the Laplace transform a certain prestige, even though convergence was achieved at the cost of tapering the given function by a real, exponentially decaying factor. In addition, the strip of convergence had to be specified for the complex transform variable s. The convenience of dealing with the real and physically intuitive frequency as the transform variable has shifted preference in favor of the Fourier and Hartley transforms. The only effective condition for the existence of a Fourier or Hartley transform today is that the given function should have a physical interpretation, or be representable by a sequence of physically interpretable functions whose individual transforms approach a limit. Consequently, it is no longer necessary to require that $f(x)$ be absolutely integrable ($\int_{-\infty}^{\infty} |f(x)| \mathrm{d}x$ exists) or that any discontinuities be finite; on the contrary, the "shah function" $\mathrm{III}(x) = \Sigma_{n=-\infty}^{n=\infty} \delta(x-n)$, which could be said to possess an infinite number of infinite discontinuities, now has a Fourier transform thanks to the theory of generalized functions (Bracewell, 1956). Interestingly, the Fourier transform of $\mathrm{III}(x)$ is $\mathrm{III}(s)$.

The function $\sin(x^{-1})$ raises a convergence question as a result of possessing an infinite number of maxima in any interval containing $x = 0$; this sort of behavior is without interest in the world of numerical computing but of considerable interest to the theory of integration. Possession of an infinite number of maxima does not in itself define the convergence condition because the Fourier integral may converge if the amplitude of the oscillation dies down so that the function exhibits bounded variation. Nor does bounded variation define the convergence condition because Lipschitz has demonstrated functions of unbounded variation whose Fourier integrals converge. However, the Lipschitz condition is not the ultimate convergence condition, as has been shown by Dini (Bracewell, 1986a). This style of analysis has lost practitioners as activity has moved in the direction of finite, or discrete, mathematics.

32.7
Why Transforms are Useful

Many problems can be posed in the form of a differential equation (or a difference equation, or an integral equation, or an integro-differential equation) that has to be solved for some wanted function subject to stated boundary conditions or initial conditions. Laplace's equation in three dimensions describes the potential distribution set up by an array of electric charges, and the diffusion equation describes the heat flow distribution set up by a given distribution of heat. By applying a transformation such as the Laplace or Fourier to each term of such an equation, we arrive at a new equation that describes the transform rather than the original wanted function. The interesting thing about this is that the new equation may be simpler, sometimes solvable just by algebra. We solve that equation for the transform of the solution and then invert. Not all differential equations are simplified in this way; those that do are characterized by linearity and coordinate invariance (such as time invariance), and the presence of these characteristics in nature is responsible for a good deal of the numerical activity with transforms. Transfer functions, such as the frequency response curves of amplifiers, are corresponding manifestations of these same characteristics. The passage of a speech waveform through an amplifier is described by a differential equation that may be hard to solve; but having used a Fourier transform to go to the frequency domain, we apply the transfer function, frequency by frequency, by complex multiplication to get the transform of the output. Then retransforming gives the output waveform.

There is also a differential equation, describing the bending of a beam under the influence of a load distribution that may be considered as a spatial input analogous to an input waveform, whereas the curve of deflection is analogous to the output waveform. Although Hooke's law, the first of the linear laws, may apply, we do not use transform methods. If we analyze the load distribution into spatially sinusoidal components and find the bending response to each component and linearly sum the responses, we will get the desired shape of the bent beam, but there is no transfer function to facilitate getting the individual responses by simple algebra. The reason is that we have linearity but not space invariance – if we shift the load, the response does not shift correspondingly without change of shape; a sinusoidal load does not produce sinusoidal deflection. If, on the contrary, we delay the input to an amplifier or a vibratory mechanical system, the response is correspondingly delayed but is unchanged as to shape; furthermore, a sinusoidal input produces a sinusoidal output.

32.8
Fields of Application

Fourier (Grattan-Guinness, 1972) originally thought of representing the temperature on a heat-conducting bar as the sum of sinusoids. To avoid a problem of integration, he considered the bar to be bent around on itself in a large circle, a distortion that is not harmful to the discussion of any given finite straight bar because the arc of interest can be made as straight as you wish by taking the circle large enough. Because the temperature distribution on the ring is of necessity now periodic in space, only a fundamental and harmonics need to be considered, plus the constant temperature a_0 representing the mean temperature. As time elapses, the temperature distribution varies as the heat flows under the influence of the temperature gradients, ultimately approaching the uniform value a_0 in the limit. Fourier found that the component sinusoids decay exponentially with a time constant proportional to the spatial period, or wavelength, the nodes of each sinusoid remaining fixed. By attenuating each component in accordance with the elapsed time and summing one gets the same result as if the spatially variable heat flow were followed in real time. This is an example of the duality of the function domain (space domain in this instance) and the transform domain (spatial frequency domain) that permeates Fourier applications.

Music can be considered in terms of the wave form of the wave that conveys the sound through the air (function domain), or in terms of the harmonic constituents (spectral domain) that are separately discernible by the ear and are treated separately by an amplifier. In crystallography, there is the arrangement of the atoms in space (crystal lattice domain) and the spatial Fourier components (reciprocal lattice domain) that under illumination by X-rays or neutron beams, evidence themselves by diffraction at defined angles. Image formation with cameras and radio telescopes can be conceived as operating on the object domain, or "sky plane," or we can think in terms of complex coherence measurements in the transform domain. All these dual modes of thought, under their respective terminologies, are fully equivalent; it helps to be familiar with both and to be able to translate from one domain to the other. In addition, it is most helpful to be able to translate between fields, converting a problem in one subject into the analogous problem in another subject where the solution may be intuitively obvious. As an example, persons who know very well that the diffraction from a pair of side-by-side pinholes is sinusoidal in space may not know that the spectrum of a pair of audible clicks in succession is sinusoidal in frequency. How much richer this knowledge becomes when they are able to translate from acoustics to optics and vice versa!

As a formal illustration of the methodology, let us calculate the response of an electric circuit consisting of an inductance L in series with a resistance R to which a voltage impulse of strength A is applied at $t = -1.5$. Equating the sum of the voltages in the circuit to zero, as taught by Kirchhoff, gives the differential equation

$$A\delta(t) = L\frac{\mathrm{d}i}{\mathrm{d}t} + Ri \tag{32.8}$$

where $i(t)$ is the current flow in response to the applied voltage. Taking the Fourier transforms term by term (Table 32.1), we find that

$$Ae^{i2\pi\times 1.5s} = i2\pi sLI(s) + RI(s) \tag{32.9}$$

where $I(s)$ is the transform of the wanted current. Solving this algebraic equation gives

$$I(s) = \frac{Ae^{i2\pi\times 1.5s}}{R + i2\pi Ls} \tag{32.10}$$

and taking the inverse Fourier transforms of both sides give

$$i(t) = \frac{A}{L}e^{-R(t+1.5)/L}\mathbf{H}(t+1.5) \tag{32.11}$$

The transform involved is illustrated in Table 32.2 for the Fourier transform. The method for solving the same problem by the Laplace transform is similar but involves reference to convergence of the integral, a complication that is circumvented when generalized function theory is combined with the Fourier integral.

Newton showed how to split sunlight into its constituent colors with a prism, where we think in the spatial domain, but there is another way that we learned from Michelson that is explicable in the time domain. We split a beam of light, and then recombine the two beams on a photodetector, but not before a controlled delay is introduced into one of the beams, for example, by retroreflection from a movable plane mirror. The detector output reveals the autocorrelation of the light beam from which, by using the autocorrelation theorem (Table 32.1) and numerical Fourier transformation, we get the spectral distribution of power.

32.9
The Hartley Transform

Table 32.2 illustrates by example that the Fourier transform in general is a complex function of the real transform variable s; consequently, two transform curves must be drawn, one for the real part and one (broken) for the imaginary part. The example $f(\tau)$ for the discrete Fourier transform is based on samples of the previous $f(x)$. Imaginary values of the discrete transform $F(\nu)$ are shown as hollow circles. Three features may be noted: no matter how closely samples are spaced, some detail can be missed; no outlying parts beyond a finite range are represented; the indexing convention 0 to $N-1$ has the effect of cutting off the left side of $F(s)$, translating it to the right, and reconnecting it. To convey the nature of this third comment, the points for $\tau > N/2$ have been copied back on the left.

The Hartley transform differs from the Fourier transform in that the kernel is the real function cas $2\pi sx$ instead of $\exp(-i2\pi sx)$. The cas function, which was introduced by Hartley (1942), is defined by cas $x = \cos x + \sin x$ and is simply a sinusoid of amplitude $\sqrt{2}$ shifted one-eight of a period to the left. The consequences of the change are that the Hartley transform is real rather than complex and that the transformation is identical to the inverse transformation. As may be obvious from the graphical example, the Hartley transform contains all the information that is in the Fourier transform and one may move freely from one to the other using the relations

$$H(s) = \text{Re}\, F(s) - \text{Im}\, F(s) \tag{32.12}$$

and

$$2F(s) = H(s) + H(N-s) - iH(s) + iH(N-s) \tag{32.13}$$

The convenience that arises from familiarity with complex algebra when one is thinking about transforms loses its value in computing. What one thinks of compactly as one complex product still means four real multiplications to computer hardware, which must be instructed accordingly.

The Hartley transform is fully equivalent to the Fourier transform and can be used for any purpose for which the Fourier transform is used, such as spectral analysis. To get the power spectrum from the complex-valued Fourier transform, one forms $[\text{Re}\, f(s)]^2 + [\text{Im}\ F(s)]^2$; starting from the real-valued Hartley transform, one forms $[H(s)]^2 + [H(-s)]^2$. The phase is obtained from

$$\begin{aligned}\tan\phi(s) &= \text{Im}(s)/\text{Re}(s)\\ &= [H(-s)/H(s)] - \pi/4\end{aligned} \tag{32.14}$$

We see that for purposes of spectral analysis by the Hartley transform, it is not necessary to work with complex quantities, because power spectrum is an intrinsic property independent of choice of kernel; the phase depends on the x origin which is locked to the peak of the cosine function in one case and the peak of the cas function in the other, hence the term $\pi/4$.

32.10
The Fast Fourier Transform

Around 1805, C.F. Gauss, who was then 28, was computing orbits by a technique of trigonometric sums equivalent to today's discrete Fourier synthesis. To get the coefficients from a set of a dozen regularly spaced data he could explicitly implement the formula that we recognize as the discrete Fourier transform. To do this, he would multiply the N data values $f(\tau)$ by the weighting factors $\exp(-i2\pi\nu\tau)$, sum the products, and repeat these N multiplications N times, once for each value of ν, making a total of N^2 multiplications. But he found that, in the case where N is a composite number with factors such that $N = n_1 n_2$, the number of multiplications was reduced when the data were partitioned into n_2 sets of n_1 terms. Where N was composed of three or more factors, a further advantage could be obtained. Gauss (1876) wrote, "illam vero methodum calculi mechanici taedium magis minuere, praxis tentatem docebit." He refers to diminishing the tedium of mechanical calculation, as practice will teach him who tries. This factoring procedure, usually into factors of 2, is the basis of the fast Fourier transform (FFT) algorithm, which is explained in many textbooks (Rabiner and Gold, 1975; IEEE, 1979; Elliott and Rao, 1982; Nussbaumer, 1982; Bracewell, 1986a; Press *et al.*, 1986) and is available in software packages. The fast method (Cooley and Tukey, 1965) burst on the world of signal analysis in 1965 and was for a time known as the *Cooley-Tukey algorithm* (IEEE, 1967), but as the interesting history (Heideman *et al.*, 1985) of prior usage in computing circles became known, the term FFT became universal.

Most FFT programs in use take advantage of factors by adopting a choice of N that is some power P of 2, that is, $N = 2^P$. The user may then design the data collection to gather, for example, $256 = 2^8$ readings. Alternatively, when such a choice does not offer, a user with 365 data points can simply append sufficient zeros to reach $512 = 2^9$ values. This might seem wasteful, but an attendant feature is the closer spacing of the resulting transform samples, which is advantageous for visual presentation. Perhaps one could do the job faster, say by factoring into 5×73. There are fast algorithms for 5 points and for many other small primes, but not for 73, as far as I know; it is simply not practical to store and select from lots of special programs for peculiar values of N. On the other hand, a significant speed advantage is gained if one elects more rigidity rather than more flexibility, tailors one's data collection to a total of 4^P values, and uses what is referred to as a *radix-4 program*. Because $1024 = 4^5$, the radix-4 approach is applicable to $N = 1024$ data samples (or to 256 for example), but not to 512 unless one appends 512 zeros. Packing with just as many zeros as there are data is commonly practiced because twice as many transform values result from the computation, and when the power spectrum is presented graphically as a polygon connecting the computed values, the appearance to the eye is much smoother.

Much practical technique is involved. If the sound level of an aircraft passing over a residential area is to be recorded as a set of measurements equispaced in time, the quantity under study begins and ends at zero value. But in other cases, such as a record of freeway noise, the noise is present when measurements begin and is still there when they cease; if the N values recorded are then packed with zeros, a discontinuity is introduced whose effects on the transform, such as overshoot and negative-going oscillation, may be undesirable. Packing with plausible (but unobserved) data can eliminate the undesired artifacts and is probably practiced in more cases than are admitted to. Authors often mitigate the effects of implied discontinuities in the data by multiplying by a tapering function, such as a set of binomial coefficients, that approaches zero at both the beginning and end of the data taken; they should then explain that they value freedom from negatives more than accuracy of amplitude values of spectral peaks or than resolution of adjacent peaks.

The FFT is carried out in P successive stages, each entailing N multiplications, for a total of NP. When NP is compared with N^2 (as for direct implementation of the defining formula), the savings are substantial for large N and make operations feasible, especially on large digital images, that would otherwise be unreasonably time consuming.

32.11
The Fast Hartley Algorithm

When data values are real, which is very commonly the case, the Fourier transform is nevertheless complex. The N transform values are also redundant (if you have the results for $0 \leq \nu \leq N/2$ you can deduce the rest). This inefficiency was originally dealt with by the introduction of a variety of efficient but unilateral algorithms that transformed in half of the time of the FFT, albeit in one direction only; now we have the Hartley transform, which for real data is itself real, is not redundant, and is bidirectional. The Hartley transform is elegant and simple and takes you to the other domain, regardless of which one you are in currently (Bracewell, 1986b; Buneman, 1989).

When a Hartley transform is obtained, there may be a further step required to get the more familiar complex Fourier transform. The time taken is always negligible, but even so the step is usually unnecessary. The reason is that although we are accustomed to thinking in terms of complex quantities for convenience, it is never obligatory to do so. As a common example, suppose we want the power spectrum, which is defined in terms of the real and imaginary parts of the Fourier transform by $P(\nu) = [\mathrm{Re}\, F(\nu)]^2 + [\mathrm{Im} F(\nu)]^2$. If we already have the Hartley transform $H(\nu)$, then it is not necessary to move first to the complex plane and then to get the power spectrum; the desired result is obtained directly as $\{[H(\nu)]^2 + [H(N - \nu)]^2\}/2$. Likewise phase $\varphi(\nu)$, which is required much less often than $P(\nu)$, is defined by $\tan\,[\varphi(\nu)] = \mathrm{Im}\ F(\nu)/\mathrm{Re}\ F(\nu)$; alternatively, one can get phase directly from $\tan\,[\varphi(\nu) + \pi/4] = H(N - \nu)/H(\nu)$, thus circumventing the further step that would be necessary to go via the well-beaten path of real and imaginary parts.

To illustrate the application to power spectra, take as a short example the data set {1 2 3 4 5 6 7 8}, whose discrete Hartley transform is

$$H(\nu) = \{4.5 - 1.707 - 1 - 0.707 \\ - 0.5 - 0.293\ 0\ 0.707\} \tag{32.15}$$

The first term, 4.5, is the mean value of the data set. The power spectrum for zero frequency is 4.5^2, for frequency $1/8(\nu = 1)$, $P(1) = (-1.707)^2 + (0.707)^2$, for frequency $2/8(\nu = 2)$, $P(2) = (-1)^2 + 0^2$. Similarly $P(3) = (-0.707)^2 + (-0.293)^2$ and $P(4) = (-0.5)^2 + (0.5)^2$. The highest frequency reached is 4/8, corresponding to a period of 2, which is the shortest period countenanced by data at unit interval.

The encoding of phase by a real transform has added a physical dimension to the interest of the Hartley transform, which has been constructed in the laboratory with light and microwaves (Villasenor and Bracewell, 1987, 1988, 1990; Bracewell, 1989; Bracewell and Villasenor, 1990) and has suggested a new sort of hologram.

32.12
The Mellin Transform

The vast majority of transform calculations that are done every day fall into categories that have already been dealt with, and much of what has been said is applicable to the special transforms that remain to be mentioned. The Mellin transform has the property that $F_M(n+1)$ is the nth moment of $f(x)$ when n assumes a finite number of integer values 1, 2, 3, ... The special value $F_M(1)$ is the zeroth moment of, or area under, $f(x)$. But the transform variable does not have to be integral or even real, so one can think of the Mellin transform as a sort of interpolate passing through the moment values. When the scale of x is stretched or compressed, for example, when $f(x)$ is changed to $f(ax)$, the Mellin transform becomes $a^{-2}F_M(s)$, a modification that leaves the position of features on the s-axis unchanged and is useful in some pattern-recognition problems.

If we plot $f(x)$ on a logarithmic scale of x, a familiar type of distortion results, and we have a new function $f(e^{-x})$ whose Laplace transform is exactly the same as the Mellin transform of $f(x)$. An equally intimate relation exists with the Fourier transform. Consequently, the FFT may be applicable in numerical situations. Because of the intimate relationship with moments and with spectral analysis, Mellin transforms have very wide application. A specific example is given by the solution of the two-dimensional Laplace equation expressed in polar coordinates, namely $\partial^2 V/\partial r^2 + r^{-1}\partial V/\partial r + r^{-2}\partial^2 V/\partial\theta^2 = 0$. Multiply each term by r^{s-1} and integrate with respect to r from 0 to ∞. We get $d^2F_M/d\theta^2 + s^2F_M = 0$. Solve this for $F_M()$ and invert the transform to get the solution. In this example, a partial differential equation is converted to a simple differential equation by the transform technique.

32.13
The Hilbert Transform

As the example in Table 32.2 shows, the Hilbert transform, or quadrature function, of a cosinusoidal wave packet is a similar, but odd, waveform sharing the same envelope. But what do we mean by the envelope of an oscillation that only touches the intuitively conceived envelope at discrete points? The Hilbert transform provides an answer in the form $\sqrt{[f(x)]^2 + [f_{Hi}(x)]^2}$. Likewise, the original wave packet reveals its phase at its zero crossings. But what is the phase at intermediate points? The Hilbert transform supplies an instantaneous phase ϕ in the form $\tan \phi = f_{\mathrm{Hi}}(x)/f(x)$. The operation **T** for the Hilbert transform is simply convolution with $-1/\pi x$. It is known that the Fourier transform of $-1/\pi x$ is $i \operatorname{sgn} s$, where $\operatorname{sgn} s$ is 1 for $s > 0$ and -1 for $s < 0$. Therefore, by the convolution theorem (last line of Table 32.1), according to which the Fourier transform of a convolution is the product of the separate Fourier transforms, it would seem that a fast Hilbert transform of $f(x)$ could be calculated as follows. Take the FFT of $f(x)$, multiply by i for $0 < \nu < N/2$ and by $-i$ for $N/2 < \nu < N$, set $F(0)$ and $F(N/2)$ equal to zero, and invert the FFT to obtain the Hilbert transform. This sounds straightforward, but the procedure is fraught with peril, for two reasons. We propose to multiply a given function $f(x)$ by $-1/\pi[(x+ \text{const})]$ and to integrate from $-\infty$ to ∞, but we are only given N samples. The extremities of $-1/\pi x$ approach zero and have opposite signs, but there is infinite area under these tails no matter how far out we start. Consequently, we ask two oppositely signed large numbers to cancel acceptably. How can we expect satisfaction when the convolving function $-1/\pi x$ is not symmetrically situated about the extremes of the data range? The second reason is that we ask for similar cancellation in the vicinity of the pole of $1/x$. Experience shows that satisfactory envelopes and phases only result when $f(x)$ is a rather narrow-band function. Under other circumstances, an N-point discrete Hilbert transform can be defined and will give valid results free from worries about the infinities of analysis, but the outcome may not suit expectation.

An optical wave packet $\exp(-\pi t^2/T^2) \sin 2\pi \nu t$ of equivalent duration T easily meets the narrow-band condition when the duration T is much greater than the wave period $1/\nu$; it has a Hilbert transform $\exp(-\pi t^2/T^2)\cos 2\pi\nu t$. The square root of the sum of the squares yields $\exp(-\pi t^2/T^2)$ for the envelope, in full accord with expectation.

32.14
Multidimensional Transforms

The two-dimensional Fourier and Hartley transforms are defined respectively by

$$F(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \times e^{-i2\pi(ux+vy)} \mathrm{d}x\mathrm{d}y \qquad (32.16)$$

$$F(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \times \operatorname{cas}[2\pi(ux + vy)]\mathrm{d}x\mathrm{d}y \qquad (32.17)$$

where the transform variables u and v mean spatial frequency components in the x and y directions. Work with images involves two dimensions, electrostatics and X-ray crystallography involve three, and fluid dynamics involves four. Multidimensional transforms can be handled numerically with a one-dimensional FFT subprogram or a fast Hartley as follows. Consider an $N \times N$ data array. Take the 1-D (one-dimensional) transform of each row and write the N transform values in over the data values. Now take the 1-D transform of each resulting column

(Bracewell, 1984; Bracewell *et al.*, 1986). In three and four dimensions, the procedure is analogous (Hao and Bracewell, 1987; Buneman, 1989). Further simple steps lead to the Hartley transform and to the real and imaginary parts of the Fourier transform if they are required, but usually they are not; more often the quadratic content (power spectrum) suffices.

When a 2-D function has circular symmetry, as commonly arises with the response functions of optical instruments, not so much work is required, as explained below in connection with the Hankel transform. Cylindrical symmetry in 3-D is essentially the same, while spherical symmetry in 3-D is also referred to below.

32.15
The Hankel Transform

In two dimensions, where there is circular symmetry as expressed by a given function $f(r)$, the two-dimensional Fourier transform is also circularly symmetrical; call it $F_{\mathrm{Ha}}(s)$. It can be arrived at by taking the full 2-D transform as described earlier, or it can be obtained from a single 1-D Hankel transform as defined in Table 32.2. The inverse transform is identical. There is apparently no opening for the Hartley transform because in the presence of circular symmetry the 2-D Fourier transform of real data contains no imaginary part. The kernel for the Hankel transform is a zero-order Bessel function, which is a complication that hampers the FFT factoring approach, but there is an elegant sidestep around this, which is explained below in connection with the Abel transform. Under spherical symmetry, the 3-D Fourier transform reduces to a different one-dimensional transform

$$4\pi \int_0^\infty f(r)\mathrm{sinc}(2sr)r^2\mathrm{d}r \qquad (32.18)$$

The inverse transform is identical.

To illustrate by a well-known result from optical diffraction, we consider a telescope aperture $f(r)$ representable as $\mathrm{rect}(r/D)$, a two-dimensional function that is equal to unity over a circle of diameter D. The Hankel transform is $D^2\mathrm{jinc}Ds$, the familiar Fraunhofer diffraction field of a circular aperture. The jinc function [$\mathrm{jinc}x = J_1(\pi x)/2x$], which is the Hankel transform of the unit rectangle function of unit height within a radius of 0.5, has the property that jinc $1.22 = 0$; this is the source of the constant in the expression $1.22\lambda/D$ for the angular resolution of a telescope.

32.16
The Abel Transform

Most commonly, although not always, the Abel transform arises when a 2-D function $g(x, y)$ has circular symmetry, as given by $f(r)$. The Abel transform (Table 32.2) then simplifies to $F_{\mathrm{A}}(x) = \int_{-\infty}^{\infty} g(x, y)\mathrm{d}y$. In other words, if the given $f(r)$ is represented by a square matrix of suitably spaced samples, then the Abel transform results when the columns are summed. There might not seem to be any future in trying to speed up such a basic operation, apart from the obvious step of summing only half way and doubling. However, when it is remembered that for each of $N^2/8$ matrix elements, we have to calculate $\sqrt{x^2 + y^2}$ to find r, and thence $f(r)$, it gives pause. The alternative is to proceed by equal steps in r rather than in y; then the oversampling near the x-axis is mitigated. But the variable radial spacing of elements stacked in a column needs correction by a factor $r/\sqrt{r^2 - s^2}$, which takes more time to compute than $\sqrt{x^2 + y^2}$. This is an excellent case for decision by using the millisecond timer found on personal computers. Of course, if many runs are to be made, the factors $r/\sqrt{r^2 - s^2}$ can be precomputed and the preparation time can be amortized over the successive runs.

Figure 32.1 shows a given function $g(x, y)$ and its one-dimensional projection (labeled P) which is derived by integrating along the y-axis in the (x, y) plane. Integrating along the y'-axis of a rotated coordinate system gives the projection P'. Now if $g(x, y)$ were circularly symmetrical, being a function $f(r)$ of r only, then the projections P and P' would be identical and equal to the Abel transform of $f(r)$. This is the graphical interpretation of the Abel transform.

Applications of the Abel transform arise wherever circular or spherical symmetry exists. As an example of the latter, consider a photograph of a globular cluster of stars in the outer reaches of the galaxy. The number of stars per unit area can be counted as a function of distance from the center of the cluster; this is the projected density. To find the true volume density as a function of radius requires taking the inverse Abel transform (Table 32.2) of the projected density.

With the Abel transform under control, we can now see a way of doing the Hankel transform without having to call up Bessel functions. The Abel, Fourier, and Hankel transforms form a cycle known as the *FHA cycle* (Bracewell, 1956), so that if we take the Abel transform and then take the FFT, we get the Hankel transform; the theorem is

$$\int_0^\infty \mathrm{d}rJ_0(2\pi\xi r)r \int_{-\infty}^{\infty} \mathrm{d}se^{i2\pi rs} \int_s^\infty \frac{\mathrm{d}x2xf(x)}{\sqrt{x^2 - s^2}} = f(\xi) \qquad (32.19)$$

The FFT required will not be complex, except in the extraordinary case of complex 2-D data; consequently, it will in fact be appropriate to use the fast Hartley to get the Hankel transform. Because of symmetry, the result will also be exactly the same as obtained with the FFT, if after taking the FFT we pay no attention to the imaginary parts that have been computed, which all should be zero or close to zero.

The FHA cycle of transforms is a special case of the projection slice theorem, a theorem which refers to the more general situation where $g(x, y)$ is not circularly symmetrical. Circular symmetry characterizes instruments, especially optical instruments, which are artifacts. Lack of symmetry characterizes data; tomographic data will be taken as the illustration for the projection-slice theorem.

32.17
Tomography and the Radon Transform

Consider a set of rotated coordinates (x', y') centered on the (x, y) plane, but rotated through θ. The expression $\int_{-\infty}^{\infty} g(x, y)\mathrm{d}y$ given for the Abel transform, representing a line integral in the y direction at a given value of x, would equal the line integral $\int_{-\infty}^{\infty} g(x, y)\mathrm{d}y'$ in the rotated direction y' provided $g(x, y)$ had circular symmetry as specified for the Abel transform. But when $g(x, y)$

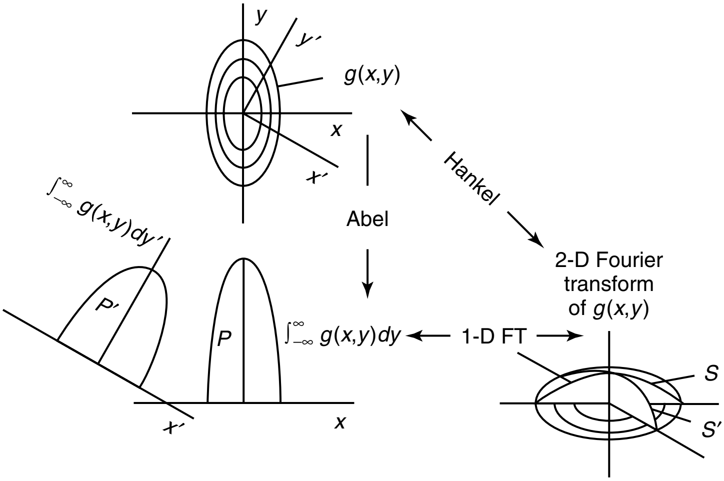
y
y′
g(x,y)
x
x′
$\int_{-\infty}^{\infty} g(x,y)dy'$
Abel
Hankel
2-D Fourier transform of g(x,y)
P′
P
$\int_{-\infty}^{\infty} g(x,y)dy$
1-D FT
S
S′
x′
x

Fig. 32.1 Illustrating the projection-slice theorem, which states that if a distribution $g(x, y)$ has a projection P', in the y' direction, its 1-D Fourier transform is the slice S' through the 2-D Fourier transform of $g(x, y)$. The set of projections P' for all inclination angles of the (x', y') coordinates constitutes the Radon transform. In the presence of circular symmetry where $g(x, y) = f(r)$, the projection P in any direction is the Abel transform of $f(r)$. The 1-D Fourier transform of P is the slice S in any direction; this slice S is then the Hankel transform of $f(r)$. Thus the Abel, Fourier, and Hankel transforms form a cycle of transforms.

does not have symmetry, then the line-integral values depend both on x' and on the angle θ (Figure 32.1). The set of integrals with respect to dy' is the Radon transform of $g(x, y)$, named after Johann Radon (1917). Such integrals arise in computed X-ray tomography, where a needle beam of X-rays scans within a thin plane section of an organ such as the brain with a view to determining the distribution of absorption coefficient in that plane. If there are N^2 pixels for which values have to be determined, and because one scan will give N data, at least N different directions of scan spaced $180°/N$ apart will be needed to acquire enough data to solve for the N^2 unknowns. In practice, more than $2N$ directions are helpful to compensate for diminished sample density at the periphery. To compute a Radon transform is easy; the only tricky part is summing a given matrix along inclined directions. One approach is to rotate all the matrices and interpolate onto a rotated grid, for each direction of scan; but this may be too costly. At the other extreme one sums, without weighting, the matrix values lying within inclined strips that, independently of inclination, preserve unit width in the direction parallel to the nearer coordinate direction. How coarse the increment inclination angle may be dependent on acceptability as judged by the user in the presence of actual data.

The harder problem is to invert the line-integral data to retrieve the required absorption coefficient distribution. A solution was given by Radon

(1917). Later Cormack (1963, 1964, 1980), working in the context of X-ray scanning of a solid object, gave a solution in terms of sums of transcendental functions. Other solutions include the modified back-projection algorithm (Bracewell, 1956; Bracewell and Riddle, 1967) used in CAT scanners (Brooks and Di Chiro, 1976; Rosenfeld and Kac, 1982; Deans, 1983). The algorithm depends on the projection-slice theorem (see Figure 32.1). According to this theorem, (Bracewell, 1956) the 1-D Fourier transform of the projection P' (or scan) of $g(x, y)$ in any one direction is the corresponding central cross section or slice S' through the 2-D Fourier transform of the wanted distribution $g(x, y)$. The proof is as follows. Let the 2-D Fourier transform of $g(x, y)$ be $G(u, v)$ as defined by

$$G(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) \times e^{-i2\pi(ux+vy)} \mathrm{d}x\mathrm{d}y \quad (32.20)$$

Setting $v = 0$, so as to have the representation $G(u,0)$ for the slice S, we get

$$\begin{aligned} G(u, 0) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) e^{-i2\pi ux} \mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} g(x, y) \mathrm{d}y \right] \times e^{-i2\pi ux} \mathrm{d}x \\ &= \int_{-\infty}^{\infty} P(x) e^{-i2\pi ux} \mathrm{d}x \quad (32.21) \end{aligned}$$

where $P(x)$ is the projection of $g(x, y)$ onto the x-axis. Thus, the 1-D transform of the projection $P(x)$ is the slice $G(u, 0)$ through the 2-D transform of $g(x, y)$. If we rotate the coordinate axes to any other orientation (x', y'), we see that the same proof applies.

Because the density of polar coordinate samples is inversely proportional to the radius of the Fourier transform plane, a simple correction factor followed by an inverse 2-D Fourier transform will yield the solution. However, a solution was found (Bracewell and Riddle, 1967; Brooks and Di Chiro, 1976), based on this theoretical reasoning, to avoid numerical Fourier transforms entirely. An equivalent correction term, arrived at by convolving each projection P' with a few coefficients, can be directly applied to each P', after which the modified projections are accumulated on the (x, y) plane by back projection to reconstitute $g(x, y)$. Back projection means assigning the projected value at x' to all points of the (x, y) plane which, in the rotated coordinate system, have the abscissa x'. Accumulation means summing the back-projected distributions for all inclination angles.

32.18
The Walsh Transform

A function defined on the interval (0, 1) can be expressed as a sum of sines and cosines of frequency 1,2,3,. . ., but can also be expressed as a sum of many other sets of basis functions. Among the alternatives, Walsh functions (Walsh, 1923; Elliott and Rao, 1982; Hsu and Wu, 1987) are particularly interesting because they oscillate between values of +1, 0, and −1, a property that is most appropriate to digital circuits, telecommunications, and radar. Furthermore, multiplication by a Walsh function value takes much less time than multiplication by a trigonometric function. Walsh functions, not being periodic, are not to be confused with the periodic square cosine and sine functions $C(x) = \mathrm{sgn}(\cos x)$ and $S(x) = \mathrm{sgn}(\sin x)$; but on a finite support they do form a complete set from which any given function can be composed. They are also orthonormal (mutually orthogonal and with fixed quadratic content, as with Fourier components), which leads to simple relations for both analysis and synthesis. The Walsh (or Walsh-Hadamard) transform has found use in digital signal and image processing and for fast spectral analysis. Fast algorithms are available that use only addition and subtraction and have been implemented in hardware. A vast, enthusiastic literature sprang into existence in the 1970s, a guide to which can be found in the text by Elliott and Rao (1982).

32.19
The *Z* Transform

In control theory, in dealing with signals of the form

$$f(t) = \sum_{-\infty}^{\infty} a_n \delta(t-n) \qquad (32.22)$$

and systems whose response to $\delta(t)$ is

$$h(t) = \sum_{0}^{\infty} h_n \delta(t-n) \qquad (32.23)$$

the response $g(t)$ is the convolution integral

$$g(t) = \int_{-\infty}^{\infty} f(t')h(t-t')\mathrm{d}t' \qquad (32.24)$$

This response is a series of equispaced impulses whose strengths are given by $\Sigma_i a_i h_{n-i}$, an expression representable in asterisk notation for convolution by $\{g_n\} = \{a_n\} * \{h_n\}$ (in this notation the sequence $\{a_n\}$ sufficiently represents $f(t)$). For example, a signal $\{1\ 1\ 1\ 1\ 1\ 1\ \ldots\}$ applied to a system whose impulse response is $\{8\ 4\ 2\ 1\}$ produces a response

$$\{1\,1\,1\,1\,1\,1\ldots\} * \{8\,4\,2\,1\}$$
$$= \{8\ 12\ 14\ 15\ 15\ 15\ldots\} \qquad (32.25)$$

This is the same rule as that that produces the coefficients of the polynomial that is the product of the two polynomials $\Sigma a_n z^n$ and $\Sigma h_n z^n$, as may be verified by multiplying $1 + z + z^2 + z^3 + z^4 + z^5 + \ldots$ by $8 + 4z + 2z^2 + z^3$. The z transform of the sequence $\{8\ 4\ 2\ 1\}$ is, by one definition, just the polynomial $8 + 4z + 2z^2 + z^3$; more often one sees $8 + 4z^{-1} + 2z^{-2} + z^{-3}$. If, conversely, we ask what applied signal would produce the response $\{8\ 12\ 14\ 15\ 15\ 15\ldots\}$, we get the answer by long division:

$$(8 + 12z + 14z^2 + 15z^3 + 15z^4 + 15z^5$$
$$+ \cdots)/(8 + 4z + 2z^2 + z^3) \qquad (32.26)$$

Occasionally, one of the polynomials may factor, or simplify, allowing cancellation of factors in the numerator and denominator. For example, the z transform of the infinite impulse response $\{8\ 4\ 2\ 1\ 0.5\ldots\}$, where successive elements are halved, simplifies to $8/(1 - z/2)$. But with measured data, or measured system responses, or both, this never happens and the z notation for a polynomial quotient is then just a waste of ink compared with straightforward sequence notation such as $\{8\ 12\ 14\ 15\ 15\ 15\ldots\} * \{8\ 4\ 2\ldots\}^{-1}$. Whenever sampled data are operated on by a convolution operator (examples would be finite differences, finite sums, weighted running means, and finite-impulse-response filters), the z transform of the outcome is expressible as a product of z transforms. Thus to take the finite difference of a data sequence, one could multiply its z transform by $1 - z$ and the resulting polynomial would be the z transform of the desired answer; in a numerical environment one would simply convolve the data with $\{1 - 1\}$. In control theory and filter design, the complex plane of z is valued as a tool for studying the topology of the poles and zeroes of transfer functions.

32.20
Convolution

Sequences to be convolved may be handled directly with available subprograms for convolution and inverse convolution that operate by complex multiplication in the Fourier transform domain. When two real sequences are to be convolved, you can do it conveniently by calling the two Hartley transforms, multiplying term by term, and calling the same Hartley transform again to get the answer. Some subtleties are involved when the sequences are of unequal length or in the unusual event that neither of the factors has symmetry (even or odd) (Bracewell, 1986b). If one of the sequences is short, having less than about 32 elements, depending on the machine, then slow convolution by direct evaluation of the convolution sum may be faster, and a shorter program will suffice. When the Fourier transform is used, the multiplications are complex but half of them may be avoided because of Hermitian symmetry. Software packages such as CNVLV (Press *et al.*, 1986) are available that handle these technicalities by calling two unilateral transforms, each faster than the FFT, or two equivalent subprograms; one fast Hartley transform, which is bilateral and, conveniently for the computer, real valued, now replaces such packages. Fast convolution using prime-factor algorithms is also available if general-purpose use is not a requisite.

As an example, suppose that $\{1\ 2\ 1\}$ is to be convolved with $\{1\ 4\ 6\ 4\ 1\}$, a simple situation where we know that the answer is the binomial sequence

$$\{1\ 6\ 15\ 20\ 15\ 6\ 1\} \tag{32.27}$$

If we select $N = 8$ for the discrete transform calculation, the given factors become in effect

$$f_1(\tau) = \{1\ 2\ 1\ 0\ 0\ 0\ 0\ 0\} \tag{32.28}$$

and

$$f_2(\tau) = \{1\ 4\ 6\ 4\ 1\ 0\ 0\ 0\} \tag{32.29}$$

respectively, where the boldface emphasizes the zeroth elements $f_1(0)$ and $f_2(0)$. The sequence $\{1\ \ 2\ \ 1\}$ is commonly used to apply some smoothing to a data sequence, but because the center of symmetry at the element 2 is offset from the origin at $\tau = 0$, a shift will be introduced in addition to the smoothing. Therefore it makes sense to permute the sequence cyclically and use

$$f_1(\tau) = \{2\ 1\ 0\ 0\ 0\ 0\ 0\ 1\} \tag{32.30}$$

To compute the convolution

$$\begin{aligned} f_3 &= f_1 * f_2 \\ &= \{2\ 1\ 0\ 0\ 0\ 0\ 0\ 1\} \\ &\quad * \{1\ 4\ 6\ 4\ 1\ 0\ 0\ 0\} \end{aligned} \tag{32.31}$$

we take the two 8-element Hartley transforms to get the values H_1 and H_2 tabulated in Table 32.3. Multiply the corresponding values as shown under H_1H_2 and take the Hartley transform again. The result is as expected; notice

Tab. 32.3 Performing convolution by multiplying Hartley transforms.

F_1	F_2	H_1	H_2	H_1H_2	F_3
2	1	0.5	2	1	6
1	4	0.427	1.457	0.622	15.008
0	6	0.25	−0.5	−0.125	20
0	4	0.073	−0.043	−0.003	15.008
0	1	0	0	0	6
0	0	0.073	0.043	0.003	0.992
0	0	0.25	−0.5	−0.125	0
1	0	0.427	−1.457	−0.622	0.992

that the peak value 20 occurs where the peak value 6 of $f_2(\tau)$ occurs; this is a result of the precaution of centering {1 2 1} appropriately. The noninteger results are a consequence of rounding to three decimals for demonstration purposes, and these errors will be present, though smaller, if more decimals are retained.

32.21
Summary

A great analytic tradition of mathematical transform theory has gained far-ranging everyday importance by virtue of new numerical possibilities opened up by automatic computing machines.

Glossary

Alias: A sinusoid of low frequency spuriously introduced by insufficient sampling in the presence of a sinusoidal component of semiperiod shorter than the sampling interval.

Convolution of Two Functions: A third function composed by spreading each element of one given function out into the form of the second function and superimposing the spread components.

Discrete Transform: One suited to functions, such as those constituted by equispaced data samples, where the function values occur at discrete intervals and are usually finite in number.

Fast Fourier Transform (FFT): An algorithm for computing the discrete Fourier transform in less time than would be required to evaluate the sum of the products indicated in the defining formula.

Frequency, Negative: A convenient fiction arising from the representation of real sinusoids by complex quantities. The representation of the real function cos $2\pi ft$ in the form $\frac{1}{2}\exp[i2\pi ft] + \frac{1}{2}\exp[i2\pi(-f)t]$ involves clockwise rotation at frequency f and counterclockwise rotation at frequency $-f$.

Frequency, Spatial: The reciprocal of the period of a periodic function of space. Values of spatial frequency are expressed in cycles per meter, or in cycles per radian, according to whether the spatial variable is distance or angle.

Frequency, Temporal: The reciprocal of the period of a periodic function of time. Values are expressed in cycles per second, or hertz.

Heaviside Unit Step Function: A function $\mathbf{H}(x)$ that is equal to zero to the left of the origin and equal to unity to the right. The value $\mathbf{H}(0)$ at the origin has no effect on the value of integrals but may conventionally be taken as 0.5.

Inverse Transformation: An operation that, when applied to the transform of a function, effectively recovers the function.

Linear Transformation: A transformation with the property that the transform of the sum of any two functions is the sum of the separate transforms.

Tomography: Originally, a photographic technique for obtaining an X-ray image of a slice of tissue within the body; now applied in many fields to a technique of combining projections in many orientations to reconstruct an image.

Transform: A mathematical function, each value of which is derived from a set of values of a given function by an explicit operation.

References

Bracewell, R.N. (1956) *Aust. J. Phys.*, **9**, 198–217.

Bracewell, R.N. (1984) *Proc. IEEE*, **72**, 1010–1018.

Bracewell, R.N. (1986a) *The Fourier Transform and Its Applications*, 2nd edn rev., McGraw-Hill, New York.

Bracewell, R.N. (1986b) *The Hartley Transform*, Oxford University Press, New York.

Bracewell, R.N. (1989) *J. Atmos. Terrest. Phys.*, **51**, 791–795.

Bracewell, R.N., Buneman, O., Hao, H. and Villasenor, J. (1986) *Proc. IEEE*, **74**, 1283–1284.

Bracewell, R.N. and Riddle, A.C. (1967) *Astrophys. J.*, **50**, 427–434.

Brooks, R.A. and Di Chiro, G. (1976) *Phys. Med. Biol.*, **21**, 689–732.

Buneman, O. (1989) *IEEE Trans. ASSP*, **37**, 577–580, (Copyright held by the Board of Trustees of Leland Stanford Junior University.).

Campbell, G.A. and Foster, R.M. (1948) *Fourier Integrals for Practical Applications*, Van Nostrand, New York.

Cooley, J.W. and Tukey, J.W. (1965) *Math. Comput.*, **19**, 297–301.

Cormack, A.M. (1963) *J. Appl. Phys.*, **34**, 2722–2727.

Cormack, A.M. (1964) *J. Appl. Phys.*, **35**, 2908–2913.

Cormack, A.M. (1980) *Med. Phys.*, **7**, 277–282.

Deakin, M.A.B. (1985) *Math. Educ.*, **1**, 24–28.

Deans, S.R. (1983) *The Radon Transform and Some of its Applications*, Wiley, New York.

Elliott, D.F. and Rao, K.R. (1982) *Fast Transforms: Algorithms, Analyses, Applications*, Academic, New York.

Erdélyi, A., Oberhettinger, F., Magnus, W. and Tricomi, F.G. (1954) *Tables of Integral Transforms*, McGraw-Hill, New York.

Gauss, C.F. (1876) *Werke*, Vol. 3, Königliche Gesellschaft der Wissenschaffen, Göttingen.

Grattan-Guinness, I. (1972) *Joseph Fourier, 1768–1830*, The MIT Press, Cambridge, MA.

Hao, H. and Bracewell, R.N. (1987) *Proc. IEEE*, **75**, 264–266.

Hartley, R.V.L. (1942) *Proc. IRE*, **30**, 144–150.

Heaviside, O. (1970) *Electromagnetic Theory*, Vols 1–3, Chelsea, New York.

Heideman, M.T., Johnson, D.H. and Burrus, C.S. (1985) *Arch. Hist. Exact Sci.*, **34**, 265–277.

Hsu, C.-Y. and Wu, J.-L. (1987) *Electron. Lett.*, **23**, 466–468.

IEEE (1967) Special issue on Fast Fourier Transform, *IEEE Trans. Audio Electroacoustics* **AU-2**, 43–98.

IEEE (1979) Digital signal processing committee, in *Programs of Digital Signal Processing*, (eds IEEE ASSP Soc.), IEEE Press, New York.

Körner, T.W. (1988) *Fourier Analysis*, Cambridge University Press, Cambridge.

Lighthill, M.J. (1958) *An Introduction to Fourier Analysis and Generalized Functions*, Cambridge University Press, Cambridge.

McCollum, P.A. and Brown, B.F. (1965) *Laplace Transforms Tables and Theorems*, Holt, Rinehart and Winston, New York.

Nahin, P.J. (1987) *Oliver Heaviside, Sage in Solitude*, IEEE Press, New York.

Neugebauer, O. (1983) *Astronomy and History, Selected Essays*, Springer, New York.

Nussbaumer, H.J. (1982) *Fast Fourier Transform and Convolution Algorithms*, Springer, New York.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986) *Numerical Recipes*, Cambridge University Press, Cambridge.

Rabiner, L.R. and Gold, B.A. (1975) *Theory and Application of Digital Signal Processing*, Prentice Hall, Englewood Cliffs, NJ.

Radon, J. (1917) *Ber. Sächs. Akad. Wiss. Leipzig, Math.-Phys. Kl.*, **69**, 262–277.

Rosenfeld, A. and Kac, A.C. (1982) *Digital Picture Processing*, Vol. 1, Academic, New York.
Villasenor, J.D. and Bracewell, R.N. (1987) *Nature*, **330**, 735–737.
Villasenor, J.D. and Bracewell, R.N. (1988) *Nature*, **335**, 617–619.
Villasenor, J.D. and Bracewell, R.N. (1990) *J. Opt. Soc. Am.*, **7**, 21–26.
Walsh, J.L. (1923) *Am. J. Math.*, **45**, 5–24.

Further Reading

Texts treating the various transforms and computational methods are identifiable from their titles in the list of works cited. An indispensable source for locating recent material on any of the special branches mentioned in this article is *Mathematical Abstracts*.

FRACTAL GEOMETRY

PAUL MEAKIN, *Department of Physics, University of Oslo, Oslo, Norway*

Introduction 501
1. **Self-Similar Fractals** 501
1.1 The Cantor Set, A Simple Example 501
1.2 Statistically Self-Similar Fractals 502
1.3 The Characterization of Self-Similar Fractals 503
1.4 Simple Rules for Self-Similar Fractals 505
2. **Self-Affine Fractals** 505
2.1 The Brownian Process, A Simple Example 505
2.2 The Characterization of Self-Affine Fractals 507
3. **Fractal Surfaces and Interfaces** 507
3.1 Some Applications 507
3.2 The Growth of Rough Surfaces 508
3.2.1 Self-Similar Rough Surfaces 508
3.2.2 Self-Affine Rough Surfaces 510
4. **Practical Considerations** 511
Glossary 512
Works Cited 513
Further Reading 513

INTRODUCTION

A quantitative description of the structure of physical objects plays an important role in our understanding of a wide range of phenomena. In many areas such as spectroscopy, solid-state physics, and engineering physics, the symmetry properties associated with this geometric description lead to important insights that would be difficult to obtain in other ways. Until recently structures have been described in terms of Euclidean geometry (straight lines, planar surfaces, spherical particles, etc.) and the associated symmetries of invariance to rotation, reflection, translation, and inversion. However, many systems of practical importance (colloids, rough surfaces and interfaces, polymer molecules, etc.) cannot be described satisfactorily in such terms. In the decade or so following the development and popularization of fractal geometry by Mandelbrot (1977, 1982) it has been shown that fractal geometry and the associated symmetry of scale invariance can be used to describe a wide variety of disorderly structures.

1. SELF-SIMILAR FRACTALS

1.1 The Cantor Set, A Simple Example

The first and perhaps the most simple example of a fractal (the Cantor set) is illustrated in Fig. 1. The Cantor set can be constructed by first taking a line and removing the middle third. In the next stage of this process the middle third from each of the remaining line segments is removed, etc. After n generations the number of line segments has grown to 2^n but their total length has decreased to $(\frac{2}{3})^n$. In the limit $n \rightarrow \infty$ a self-similar fractal has been constructed. If this fractal is dilated by a factor of 3, it can be covered by two replicas of itself. Such self-similar fractals can be characterized in terms of their fractal dimensionality D given by $D = \log n / \log \lambda$, where n is the number of replicas required to cover the fractal after dilation by a factor of λ. For the Cantor set illustrated in Fig. 1 the fractal dimensionality is $\log 2 / \log 3$, or about 0.6309. This is intuitively reasonable; the Cantor set is clearly more than a point ($D=0$), since it contains an infinite number of them, but less than a line

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

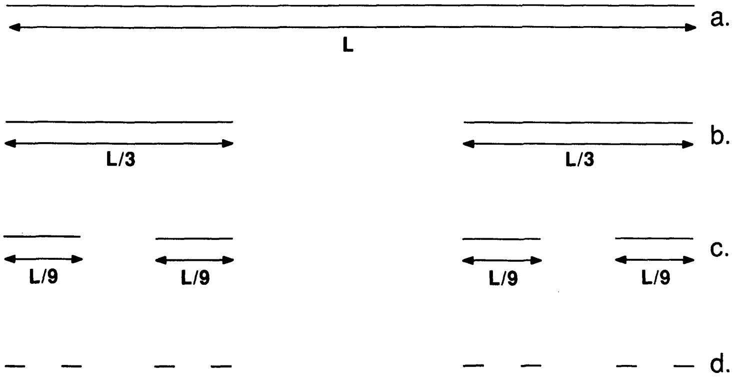

FIG. 1. Three stages in the construction of a Cantor set with a fractal dimensionality of log 2/log 3. The bottom line shows the third-generation prefractal.

$(D=1)$, since its total length is zero. In many applications the fractal dimension can be thought of as the exponent that relates mass M to length L,

$$M \sim L^D. \tag{1}$$

Here, L is a characteristic length, such as the radius of gyration R_g or maximum diameter, that describes the overall spatial extent. Equation (1) is also appropriate for Euclidean shapes where D is now the ordinary, Euclidean, dimensionality d.

After they were first introduced, fractals such as the Cantor set were considered to be very unusual objects with no possible applications in the physical sciences. In some areas fractals are still referred to by terms such as "strange sets"; this no longer seems to be appropriate. There is no precise definition of a fractal, but in general terms a fractal is an object that has the same degree of complexity on different length scales.

In real systems the geometric scaling relationships that characterize fractal objects do not extend over an infinite range of length scales. There are in general both lower (ϵ) and upper (L) cutoff lengths that limit the range of fractal scaling. For example, in the case of a flexible polymer molecule ϵ might correspond to the monomer size and L to the radius of gyration. If the ratio L/ϵ is large, then fractal geometry can be a very important asset in our attempts to understand complex, disorderly systems. If L/ϵ is small (say less than one order of magnitude), then fractal geometry is not likely to be of much practical importance. However, it may still be of considerable conceptual value if the structure was assembled by a mechanism that would lead to a fractal structure if it were not perturbed by other processes.

1.2 Statistically Self-Similar Fractals

Highly organized regular fractals such as the Cantor set (Fig. 1) that can be mapped exactly onto themselves after a change of length scales do not provide realistic models for describing most natural structures. Such natural fractals have a more complex disorderly structure that is self-similar only in a statistical sense. Statistically self-similar fractals can be described in terms of the scaling relationships such as Eq. (1) that describe regular, hierarchical fractals, but these equations must now be interpreted statistically (for example L might be the average radius of gyration for a large ensemble of structures of mass M).

Figure 2 shows an example of a statistically self-similar fractal structure. It is a projection, onto a plane, of a cluster of spherical particles generated by a three-dimensional model for the diffusion-limited cluster-cluster aggregation (colloidal flocculation: Meakin, 1988; Jullien and Botet, 1986). The fractal dimensionality of this structure is about 1.8. Since $D<2$, the projection also has a fractal dimensionality of 1.8 (see Sec. 1.4).

The use of correlation functions has, for a long time, been a valuable approach toward the quantitative characterization of disorderly systems. For example, density correlation functions such as $C^n(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ defined as

$$C^n(\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_n) = \langle \rho(\mathbf{r}_0)\rho(\mathbf{r}_0+\mathbf{r}_1)\cdots\rho(\mathbf{r}_0+\mathbf{r}_n)\rangle \tag{2}$$

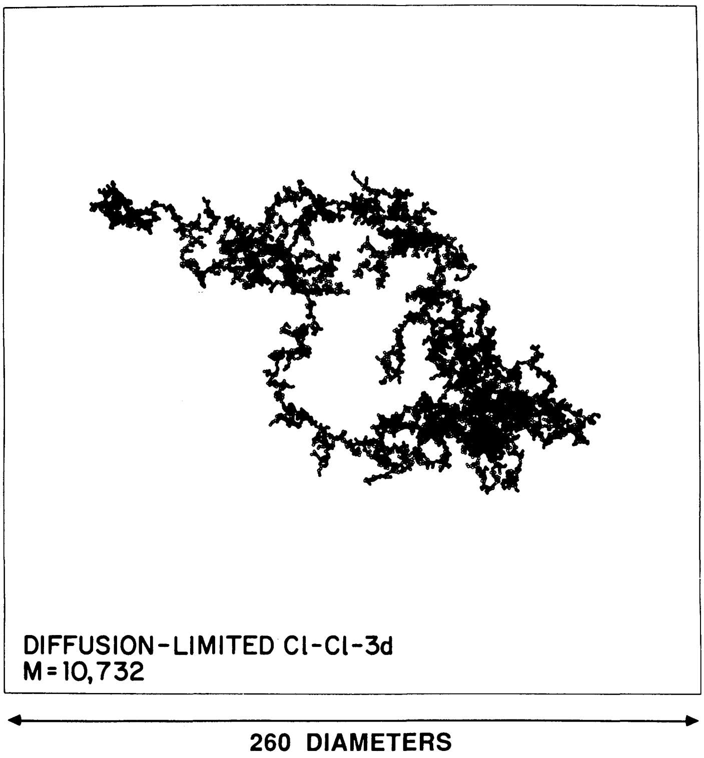

FIG. 2. A cluster of 10,732 particles generated using a three-dimensional off-lattice diffusion-limited cluster-cluster aggregation model. In this model the cluster size distribution is allowed to evolve in a natural way.

can be used to describe both fractal and nonfractal structure. Here $\rho(\mathbf{r})$ is the density at position $\mathbf{r}$ and the averaging is over all possible origins ($\mathbf{r}_0$). For self-similar fractals these correlation functions have a homogeneous power-law form,

$$C^n(\lambda\mathbf{r}_1,\lambda\mathbf{r}_2,\ldots,\lambda\mathbf{r}_n)=\lambda^{-n\alpha}C^n(\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_n). \quad (3)$$

The exponent α (called the co-dimensionality) in Eq. (3) is $d-D$, where d is the Euclidean dimensionality of the embedding space.

By far the most frequently used correlation function is the two-point density-density correlation function $C(r)$ given by

$$C(r)=\langle\langle\rho(\mathbf{r}_0)\rho(\mathbf{r}_0+\mathbf{r})\rangle\rangle_{|\mathbf{r}|=r}\,. \quad (4)$$

Here $\langle\langle\ \rangle\rangle$ implies averaging over all origins ($\mathbf{r}_0$) and orientations. In addition $C(r)$ may be averaged over an ensemble of samples. For a simple self-similar fractal $C(r)$ has the power-law form

$$C(r)\sim r^{-\alpha} \quad (5)$$

and the fractal dimensionality D_α is equal to $d-\alpha$.

1.3 The Characterization of Self-Similar Fractals

Correlation functions such as those described above can be used to measure the fractal dimensionality. In practice only the two-point density-density correlation function has been used extensively for this purpose. Figure 3(a) shows the two-point density-density correlation functions for clusters of different sizes generated using the three-dimensional diffusion-limited cluster-cluster aggregation model illustrated in Fig. 2. These correlation functions have the form

$$C(r)=r^{-\alpha}f(r/L), \quad (6)$$

where L is the cutoff length. The cutoff function $f(x)$ has the form $f(x)=\text{const}$ for $x\ll 1$ and $f(x)$ decays faster than any power of x with increasing x for $x\gg 1$. Figure 3(b) shows that the density-density correlation function can be represented by the scaling form

$$C(r)=N^{(D-d)/D}g(r/N^{1/D}), \quad (7)$$

where N is the number of particles in the cluster. Since $L\sim N^{1/D}$ and $\alpha=d-D$, the scaling forms in Eqs. (6) and (7) are equivalent.

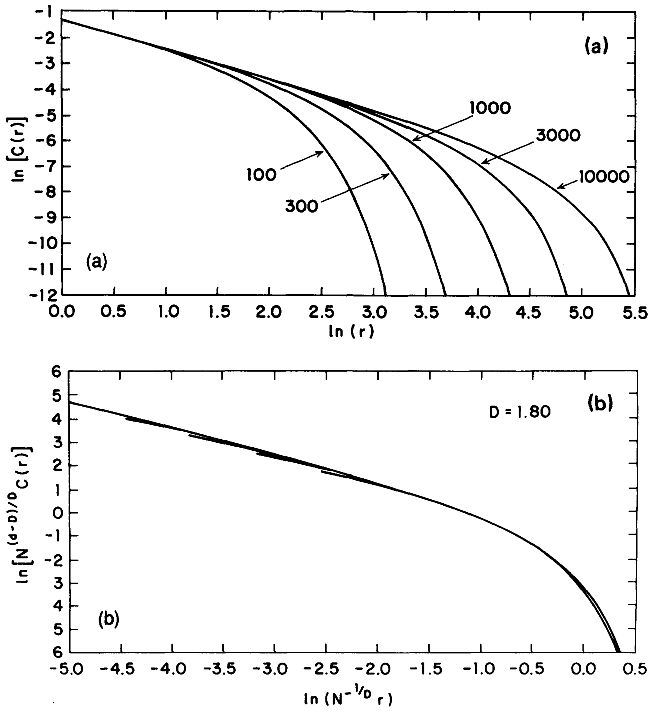

FIG. 3. Scaling of the two-point density-density correlation function for clusters generated using an off-lattice diffusion-limited cluster-cluster aggregation model illustrated in Fig. 2. (a) The correlation functions obtained from 10,000 100-particle clusters, 1000 300-particle clusters, 100 1000-particle clusters, 39 3000-particle clusters, and 13 10,000-particle clusters. (b) How these correlation functions can be scaled onto a common curve using the scaling form given in Eq. (7).

[The functions $f(x)$ and $g(x)$ are related by $g(x) \sim x^{-\alpha} f(x)$]. The results shown in Fig. 3 demonstrate that the internal structure of the clusters and their global mass-length scaling can be described in terms of the same fractal dimensionality ($D \simeq 1.8$).

Most approaches to the characterization of self-similar fractals are based on Eq. (1). For example, if we are concerned with structures formed by growth processes or systems in which a large number of objects of different sizes are present, then the fractal dimensionality can be obtained from the dependence of the radius of gyration on the mass. For an ensemble of statistically self-similar fractals we have

$$\langle R_g \rangle \sim M^{\beta}, \tag{8}$$

where $\langle R_g \rangle$ is the mean radius of gyration for structures of mass M. The corresponding fractal dimensionality D_β is then given by $D_\beta = 1/\beta$. In practice D_β is obtained by fitting the dependence of $\log R_g$ on $\log M$ by a straight line and taking the slope of the straight line as the exponent β. If data are available from many realizations over a broad range of length scales, the exponent β can be measured over a number of mass intervals to assess how accurately Eq. (8) represents the dependence of R_g on M.

Another popular approach to the measurement of the fractal dimension of self-similar fractals is to cover the fractal by a series of grids with elements having sides of length ϵ. The number of elements completely or partially occupied by the fractal is then given by

$$N(\epsilon) \sim \epsilon^{-D_\epsilon} \tag{9}$$

so that the fractal dimensionality (D_ϵ) can be obtained by plotting $\log N(\epsilon)$ against $\log \epsilon$. In practice this method appears to be quite sensitive to corrections associated with a limited range of accessible length scales, and it is difficult to obtain reliable results.

In many cases structures grow from a unique "seed" or nucleation site. In this case the fractal dimensionality can be obtained by measuring the mass $M(l)$ contained within a distance l measured from this unique point. For a self-similar fractal $M(l)$ is given by

$$M(l) \sim l^{D_\gamma} \tag{10}$$

so that D_γ can be obtained by fitting a straight line to the dependence of $\log M(l)$ on $\log l$ over a suitable range of length scales.

In principle all of the methods described above (and many other methods) should lead to the same value for the fractal dimensionality ($D_\alpha = D_\beta = D_\gamma = D_\epsilon = D$, where D is the "all purpose" fractal dimensionality) for simple self-similar fractals. In practice it is a good idea to use several approaches to measure D. This provides an assessment of the uncertainty in D and some indication of whether or not the structure is indeed a self-similar fractal. These methods can be applied equally well to fractal structures generated by physical phenomena or computer simulations.

1.4 Simple Rules for Self-Similar Fractals

The ability to describe complex, disorderly structures in quantitative terms (via fractal geometry) has stimulated scientific interest in problems that only a decade or so ago were thought to lie outside of the realm of quantitative scientific investigation. For a variety of model systems we now have quantitative (frequently exact but rarely rigorous) results and at least the beginnings of a sound theoretical understanding. In attempting to apply the concepts of fractal geometry to self-similar systems the following simple rules or ideas have been found to be useful (Vicsek, 1989; Meakin, 1990).

1. Two fractals with dimensionalities D_1 and D_2 can be placed together in the same region of a d-dimensional embedding space or lattice without contacting each other if $d > D_1 + D_2$. If the two fractals are following a relative trajectory with a fractal dimensionality of D_t, then they will not contact each other (except by accident) if $d > D_1 + D_2 + D_t$. An important implication of this rule is that fractals with $D < 2$ will be asymptotically transparent in three-dimensional space since they will not be contacted by "point" objects ($D = 0$) such as photons or electrons following linear ($D = 1$) trajectories. For such fractals one part is (in general) not hidden by another, and the fractal dimensionality of a projection onto a plane is the same as that of the fractal itself. If $D > 2$, then the structure is asymptotically opaque and the fractal dimensionality cannot be determined by analyzing a projection onto a plane. It follows from this that the projection of a D-dimensional fractal onto a d-dimensional space will have a fractal dimension of D if $D < d$. In this event the area (measure) of the projection will be proportional to the mass (measure of the fractal in the d-dimensional space).
2. A d_1-dimensional cross section of a D-dimensional fractal in a d_2-dimensional space will have a fractal dimensionality of $D + d_1 - d_2$.
3. The (set theoretical) intersection of two fractals with dimensionalities D_1 and D_2 in a d-dimensional space is given by $D_1 + D_2 - d$. This rule can be applied repeatedly to obtain the dimensionality of the intersection of three ($D_1 + D_2 + D_3 - 2d$) or more fractals.
4. The union of two fractals with dimensionalities D_1 and D_2 has a fractal dimensionality of max (D_1, D_2).
5. The product of two fractals with dimensionalities D_1 and D_2 has a dimensionality of $D_1 + D_2$. For example, the region swept out by a fractal of dimensionality D following a trajectory of dimensionality D_t is $D + D_t$ (if $D + D_t < d$).
6. Many random fractals can be described in terms of a power-law distribution of unoccupied "holes,"

$$N_s \sim s^{-\tau}, \tag{11}$$

where N_s is the number of holes of size s (s would be the number of sites contained in the hole for a lattice model). For such fractals the size distribution exponent τ is given by $\tau = (d + D)/d$.

2. SELF-AFFINE FRACTALS

2.1 The Brownian Process, A Simple Example

Fractals that have different scaling structures in different directions are said to be self-affine. Perhaps the most simple and most important example of a self-affine fractal is the Brownian process $B(t)$ that describes the distance moved by a Brownian particle in time t. It is well known that (on average) the distance moved by a Brownian particle in time t is proportional to $t^{1/2}$ so that the Brownian process can be rescaled by simultaneously changing the time scale by a factor of b and the distance scale by a factor of $b^{1/2}$.

More formally this symmetry property of the Brownian process can be written as

$$B(t) \equiv b^{-1/2} B(bt). \tag{12}$$

In this equation the symbol "≡" should be interpreted as meaning "statistically equivalent to."

Figure 4 shows different "lengths" of the same discretized Brownian process in which the distance is increased randomly by $+1$ or -1 each time the time is incremented by 1. In each part of the figure the horizontal (time) scale is proportional to the total time T and the vertical (distance) scale is proportional to $T^{1/2}$. The observation that the four rescaled curves in Fig. 4 look "similar" illustrates the self-affine scaling of the Brownian process.

The Brownian process can be generalized to give the "fractal" Brownian process $B_H(t)$, for which the self-affine scaling properties can be represented as

$$B_H(t) \equiv b^{-H} B(bt), \tag{13}$$

where the exponent H is referred to as the Hurst exponent (roughness exponent or wandering exponent). Values of H larger than $\frac{1}{2}$ correspond to persistent processes and $H<\frac{1}{2}$ implies antipersistent fluctuations. This is il-

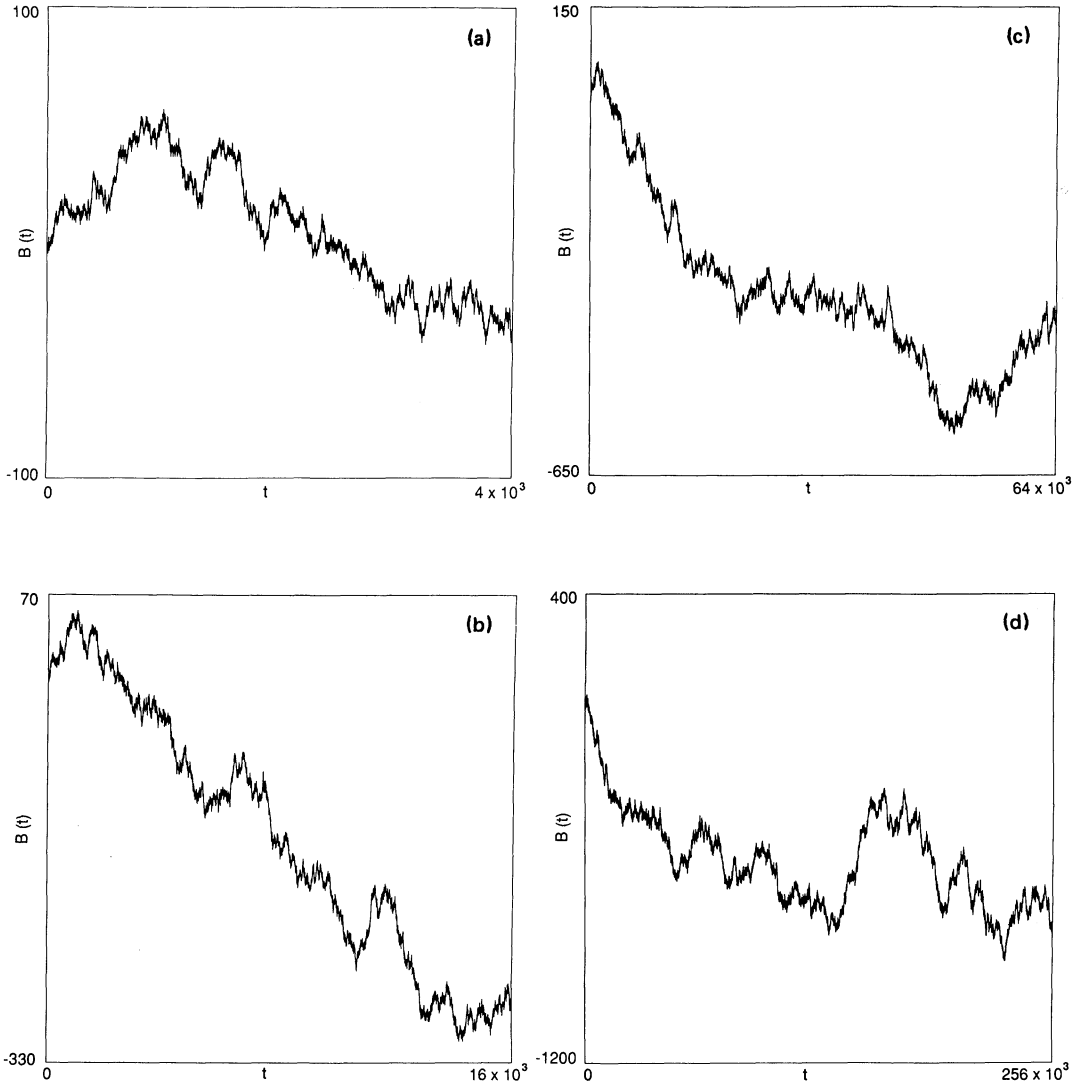

FIG. 4. Four sections of the same (discretized) Brownian process starting at the point [$t=0$, $B(t)=0$]. In each successive stage the time scale is increased by a factor of 4 and the vertical scale by 2 ($4^{1/2}$).

lustrated in Fig. 5 where fractal Brownian curves with Hurst exponents of 0.1, 0.5, and 0.9 are shown.

2.2 The Characterization of Self-Affine Fractals

In many cases (such as the Brownian process described above) self-affine fractals can be represented as single-valued functions $z(\mathbf{x})$ of the coordinates $x_1, x_2, \ldots, x_n$. For this class of self-affine fractals the scaling properties can be described in terms of the correlation functions $C_q(x)$ defined as

$$(C_q(x))^q = \langle |z(\mathbf{x}_0) - z(\mathbf{x}_0+\mathbf{x})|^q \rangle_{|\mathbf{x}|=x} . \tag{14}$$

In this case it has been assumed that all of the coordinates $(x_1, \ldots, x_n)$ are equivalent. In general the self-affine fractal $z(\mathbf{x})$ can be characterized by taking cross sections through the function $z(\mathbf{x})$ in the direction x_m, $y(x_m)$, and measuring the correlation function

$$(C_q(x_m))^q = \langle |y(x_m^0) - y(x_m^0 - x_m)|^q \rangle . \tag{15}$$

For self-affine fractals the correlation functions $C_q(x)$ or $C_q(x_m)$ have the form

$$C_q(x) \sim x^H \tag{16a}$$

or

$$C_q(x_m) \sim x_m^{H_m} . \tag{16b}$$

3. FRACTAL SURFACES AND INTERFACES

3.1 Some Applications

One of the most important applications of fractal geometry is the quantitative description of irregular surfaces and the development of a better understanding of their behavior. During the past few decades the technological importance of rough surfaces has motivated the development of a large

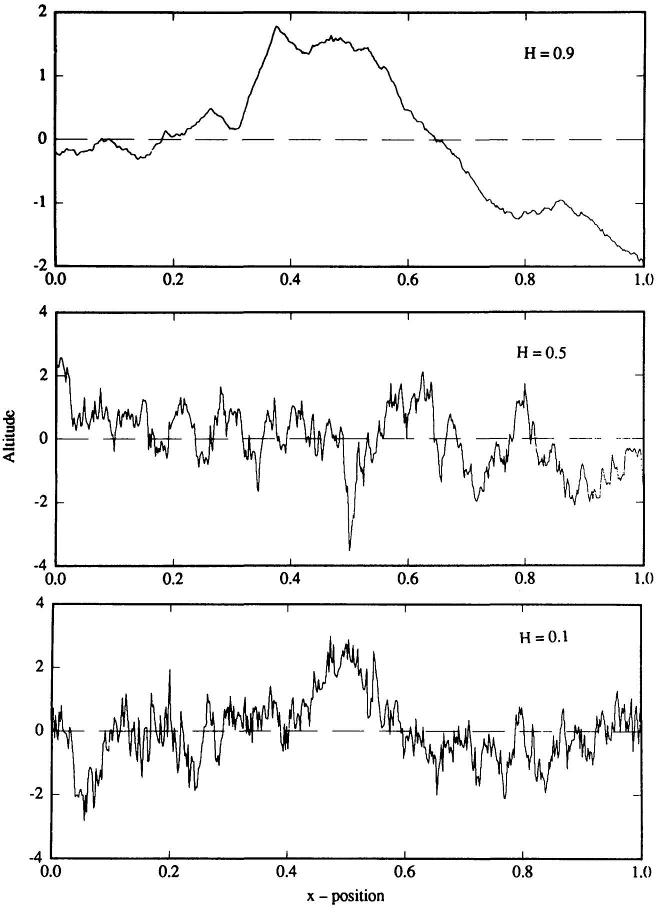

FIG. 5. Fractal Brownian curves generated using three different values for the Hurst exponent [$H=0.9$ (top), $H=0.5$, and $H=0.1$]. This figure is taken from Feder (1988) and was provided by J. Feder.

number of procedures for characterizing their structure. Many of these approaches involve a large number of parameters that are of little fundamental importance. In a recent review (Nowicki, 1985) 32 parameters and/or functions that have been used to characterize rough surfaces are described. More recently it has been shown that a wide variety of rough surfaces generated by processes such as fracture (Mandelbrot *et al.*, 1984; Charmet *et al.*, 1989; Herrmann and Roux, 1990), corrosion (Oppenheim *et al.*, 1991; Holland-Moritz *et al.*, 1991), deposition (Meakin, 1987; Family and Vicsek, 1991; Krug and Spohn, 1991), or fluid-fluid displacement (Feder, 1988) can be described in terms of fractal geometry. Both self-similar and self-affine fractal surfaces are common but most surfaces appear to be self-affine.

3.2 The Growth of Rough Surfaces

In many processes of practical importance rough surfaces are generated from a more or less smooth surface. Under these circumstances the surface roughness can often be described in terms of the correlation lengths $\xi_\perp$ and $\xi_\parallel$ and the manner in which they grow. The correlation length $\xi_\perp$ describes the amplitude of the surface roughness in a direction perpendicular to the general direction of the surface and $\xi_\parallel$ is the distance over which fluctuations in the surface height persist in a direction parallel to the coarse-grained surface. The length $\xi_\perp$ can be defined as

$$\xi_\perp^2=\langle h^2\rangle-\langle h\rangle^2, \qquad (17)$$

where $h(\mathbf{x})$ is the height of the surface above (or below) position $\mathbf{x}$ on the initially smooth surface.

For many simple processes the correlation length $\xi_\perp$ grows algebraically with increasing time

$$\xi_\perp\sim t^\omega. \qquad (18)$$

For self-affine surfaces the correlation lengths $\xi_\perp$ and $\xi_\parallel$ are related by

$$\xi_\perp\sim\xi_\parallel^H, \qquad (19)$$

where H is the Hurst exponent. In some cases surface properties in the x and y directions parallel to the surface may be quite distinct so that Eq. (17) may be replaced by

$$\xi_\perp\sim\xi_x^{H_x}\sim\xi_y^{H_y}. \qquad (20)$$

In some simple cases the amplitude of the surface roughness ($\xi_\perp$) may grow indefinitely according to Eq. (18); but in many cases $\xi_\perp$ is limited by other physical processes, and this limits the range of length scales over which fractal scaling can be observed.

Simple model systems that are quite well understood are used to illustrate how rough surfaces can be characterized using fractal geometry and scaling ideas in Secs. 3.3 and 3.4.

3.2.1 Self-Similar Rough Surfaces. The invasion percolation model (Lenormand and Bories, 1980; Wilkinson and Willemsen, 1983; Feder, 1988) provides a simple description of the slow displacement of a wetting fluid by a nonwetting fluid in a porous medium. In the site-invasion percolation model the sites on a lattice are assigned random "threshold" values at the start of a simulation. At each step in the simulation the unoccupied perimeter site with the lowest threshold value is filled to represent the fluid-fluid displacement process (unoccupied perimeter sites are unoccupied sites with one or more occupied nearest neighbors). In the two-dimensional version of this model the displacement pattern is self-similar with a fractal dimensionality of $\frac{91}{48}$ (about 1.89), or about 1.82 if growth is not allowed to take place in regions that have been surrounded by the growing cluster. The invasion front (outer perimeter) has a fractal dimensionality of $\frac{4}{3}$ (Grossman and Aharony, 1986; Saleur and Duplantier, 1987; Coniglio *et al.*, 1987) for both versions of the model (with and without "trapping").

If the fluid-fluid displacement processes take place in a vertical or inclined cell and the two fluids have different densities, the process may be either stabilized (Birovljev *et al.*, 1991) or destabilized by gravity. In this situation the invasion percolation process can be simulated using thresholds given by

$$t_i=x_i+gh_i, \qquad (21)$$

where t_i is the threshold associated with the ith site, x_i is a random number (in the most simple model, used here, the random numbers are uniformly distributed over the range $0<x_i<1$), and h_i is the height of the ith site. Figure 6 shows the displacement fronts generated during simulations carried out using the gravity-stabilized invasion percolation model. In this model the invasion front evolves

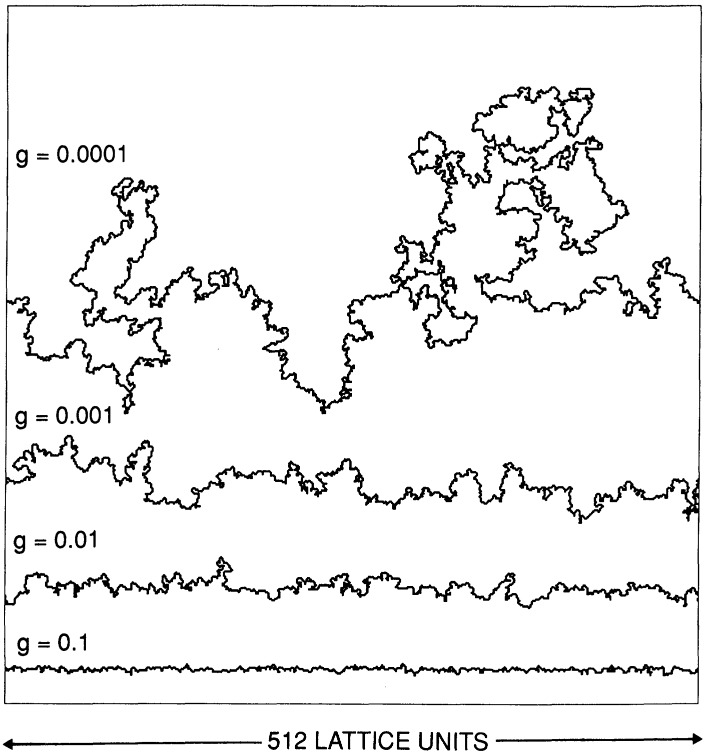

FIG. 6. Invasion fronts (unoccupied external perimeter sites) obtained from two-dimensional gradient-stabilized invasion percolation simulations. Fronts generated using four values of the stabilizing gradient g (10^{-1}, 10^{-2}, 10^{-3}, and 10^{-4}) are shown.

toward an asymptotic regime in which the statistical properties become stationary. The fronts shown in Fig. 6 were recorded in this asymptotic regime.

Self-similar fractal surfaces can be described in terms of the correlation functions defined in Eqs. (2) and (4). For structures generated by physical processes, Eqs. (3) and (5) are accurate over a limited range of length scales ($\epsilon \ll r \ll \xi$). For the invasion percolation simulations described in this section, the inner cutoff length ϵ is one lattice unit and the outer cutoff length ξ is related to the stabilizing gradient g by

$$\xi \sim g^{-\gamma}, \tag{22}$$

where the exponent γ is given by (Wilkinson, 1984)

$$\gamma = \nu/(\nu+1), \tag{23}$$

where ν is the correlation length exponent for ordinary percolation. The exponent ν has a value of exactly $\frac{4}{3}$ for two-dimensional percolation (Stauffer, 1985).

Figure 7(a) shows the density-density correlation functions $C(r)$ for the invasion fronts obtained from two-dimensional invasion percolation simulations with five values for the stabilizing gradient (g): $g=0.1$, 0.01, 0.001, 0.0001, 0.00001. In these plots a crossover from a slope of $-\frac{2}{3}$ on short length scales (corresponding to a fractal dimensionality of $\frac{4}{3}$) to a slope of -1 on longer length scales ($D=1$) can be seen. The results shown in Figs. 6 and 7 were obtained by starting with a flat surface and allowing the simulations to proceed until the vertical correlation length $\xi_\perp$ has grown to a stationary value $\xi_\perp(\infty)$ given by Eq. (22). In this stationary regime the growth of the invasion front exhibits large fluctuations, but the mean statistical properties such as $\xi_\perp$ and $C(r)$ do not change.

Figure 7(b) shows how the correlation function $C(r)$ shown in Fig. 7(a) can be collapsed onto a single curve. This figure illustrates that $C(r)$ (in the stationary regime) can be represented by the scaling form

$$C(r) = g^{(8/21)} f(g^{4/7} r) \tag{24}$$

or

$$C(r) = g^{\alpha\gamma} f(g^{\gamma} r), \tag{25}$$

where the scaling function $f(x)$ has the form $f(x) \sim x^{-2/3}$ for $x \ll 1$ and $f(x) \sim x^{-1}$ for $x \gg 1$.

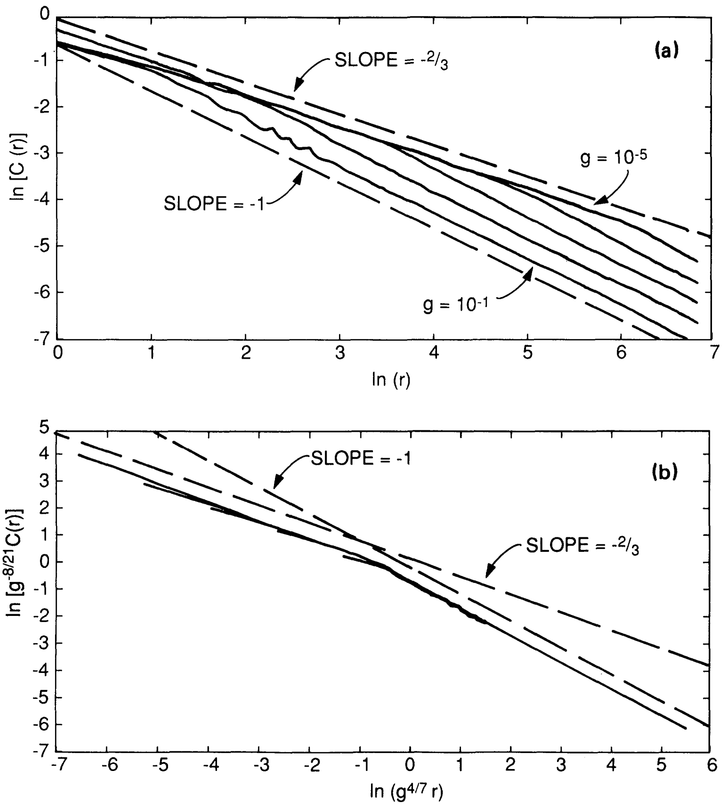

FIG. 7. Two-point density-density correlation functions for invasion fronts similar to those shown in Fig. 6. Each correlation function was obtained from 20 simulations carried out using a strip width L of 2048 lattice units. (b) How these correlation functions can be scaled using the scaling form given in Eq. (24).

This means that the surface appears to be self-similar ($D=\frac{4}{3}$) for $x \ll 1$ and flat ($D=1$) for $x \gg 1$.

3.2.2 Self-Affine Rough Surfaces. The Brownian process described above provides a valuable paradigm for the geometry of rough surfaces. The correlation functions $C_q(x)$ [Eq. (14)] can be used to measure the Hurst exponent (by least-squares fitting of straight lines to the dependence of $\log C_q(x)$ on $\log x$. For $q=1$ and 2 the value for the Hurst exponent is well within the 1% of the expected value ($H=\frac{1}{2}$).

In many real systems the correlation length $\xi_\perp$ may be finite because $\xi_\perp$ has not had enough time to grow or because it is limited by physical processes or finite size effects. Figure 8 shows "surfaces" generated using a simple modification of the discrete Brownian process model in which the probability of moving toward the origin is $0.5+k|x|$ and the probability of moving away from the origin is $0.5-k|x|$, where x is the displacement from the origin. This may be regarded as a model for the motion of a Brownian particle in a harmonic potential. Figure 9(a) shows the height-difference correlation functions $C_2(t)$ [Eq. (13)] obtained from simulations carried out using this model for six values of the parameter k (and $k=0$). Figure 9(b) shows that these correlation functions can be represented quite well by the scaling form

$$C_2(t)=k^{-1/2}h(kt), \tag{26}$$

where the scaling function $h(x)$ has the form $h(x) \sim x^{-1/2}$ for $x \ll 1$ and $h(x)=\text{const}$ for $x \gg 1$. This means that the surface appears to be self-affine ($H=\frac{1}{2}$) for $x \ll 1$ and flat for $x \gg 1$.

If a cross section is taken through a self-affine curve or surface parallel to the coarse-grained direction of the surface, then the intersection between the self-affine surface and the d-dimensional intersecting plane is a self-similar fractal with a dimensionality D_s given by

$$D_s=d-H. \tag{27}$$

Consequently, the relatively reliable methods that have been used to analyze self-similar fractals can be used [via Eq. (27)] to measure

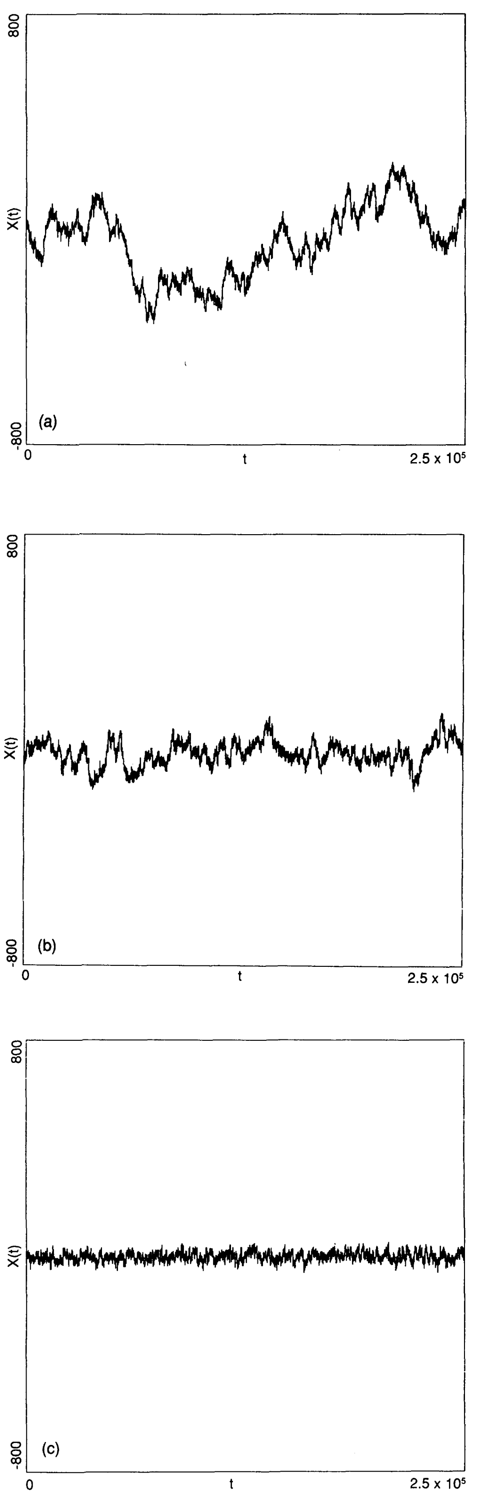

FIG. 8. Displacement curves $x(t)$ obtained from a simple model for Brownian motion in a harmonic potential. (a)–(c) Results for $k=10^{-2}$, 10^{-3}, and 10^{-4}, respectively.

the Hurst exponent of self-affine fractal surfaces. This is the basis of the slit island method (Mandelbrot *et al.*, 1984) that has been applied successfully to a variety of rough surfaces.

4. PRACTICAL CONSIDERATIONS

In Sec. 3 it was shown that quite accurate values for the fractal dimensionality can be obtained for self-similar and self-affine surfaces using simple procedures. Very similar approaches can be used for other types of fractal structures. However, large quantities of data spanning a large range of length scales were available from the simple models used in Sec. 3. In practice the statistical uncertainties may be larger and the range of length scales smaller for experiments or more realistic models. In most cases statistical uncertainties can be reduced to quite tolerable levels by averaging, but systematic uncertainties due to correlations to the asymptotic scaling behavior and other finite-size effects may be much more difficult to detect and control. In addition, real systems may have a much more complex scaling structure than that of the examples used here.

An account of the geometric properties of systems that must be described in terms of multifractal geometry and related concepts such as multiscaling and multiaffinity is beyond the scope of this survey. Information on those topics may be found in recent books (Feder, 1988), reviews (Paladin and Vulpiani, 1987; Stanley and Meakin, 1988; Meakin, 1990), and conference proceedings (Pietronero, 1988). In general, there is no well-established general procedure for characterizing the scaling structure of these more complex systems. In some cases it appears that the corrections to the asymptotic scaling behavior are large and convergence is slow.

One of the main difficulties in the past has been a failure to distinguish carefully between self-similar and self-affine fractals. This is particularly true in the case of surfaces and interfaces. As a consequence much of the literature in this area is subject to reinterpretation and/or is ambiguous.

Finally, there are no simple standards or criteria for determining whether a structure is fractal or not. To some extent this depends on

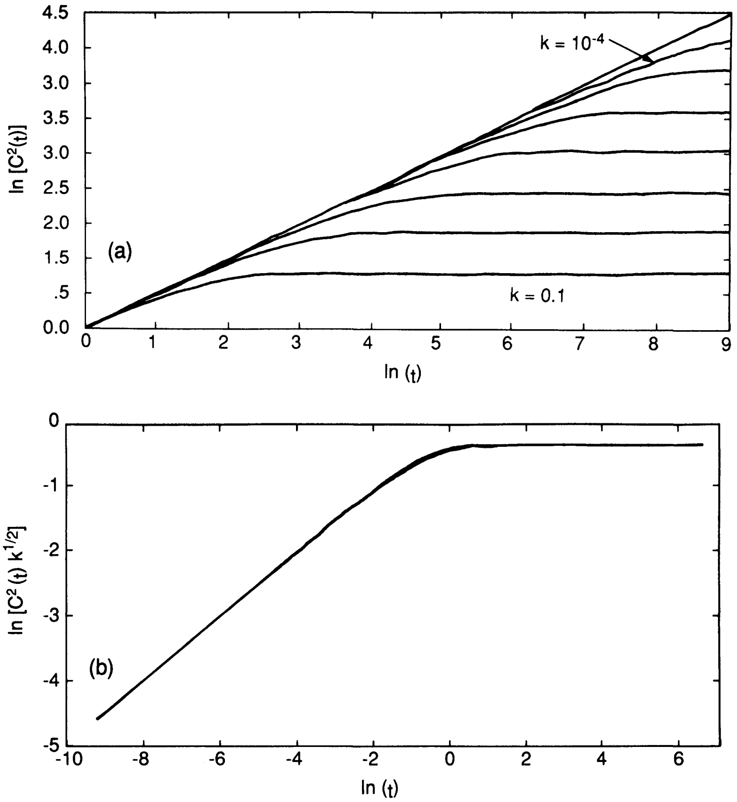

FIG. 9. (a) Correlation functions $C_2(t)$, obtained from simulations carried out using the model illustrated in Fig. 8 for eight values of k (0, 10^{-4}, 3×10^{-4}, 10^{-3}, 3×10^{-3}, 10^{-2}, 3×10^{-2}, and 10^{-1}). (*b*) How these curves can be collapsed onto a common curve using the scaling form given in Eq. (24).

the application and on theoretical considerations. However, geometric scaling (power-law behavior) over at least a decade of length scales combined with some understanding of the deviations from scaling outside of this range is probably a minimum requirement for fractal analysis to be a useful practical tool. In practice, scaling over more than two orders of magnitude of the length scale is rare.

GLOSSARY

Dimensionality: Dimensionality can be defined in many ways, but in practical terms it is the exponent relating mass (or measure) to length.

Embedding Space: The Euclidean space in which a structure resides. In most practical cases this is ordinary three-dimensional space ($\mathbb{R}^3$), but two-dimensional embedding spaces are common in computer simulations and for processes occurring at smooth interfaces.

Fractal: A structure that exhibits geometric scaling. In general terms a fractal is a structure that has similar complexity on all length scales; it "looks the same" on different length scales or at different magnifications.

Percolation: The transition associated with the formation of a continuous path spanning an arbitrarily large ("infinite") range. Site percolation on a lattice is a simple model for percolation. In this model the sites on a lattice are selected randomly and filled. For an infinitely large lattice an infinite cluster will be formed when a fraction p_c (the site percolation threshold probability) of the lattice sites have been filled. This cluster is a self-similar fractal. However, the entire system (including all the smaller clusters) has a finite density (p_c) and is uniform on all but small length scales.

Prefractal: An intermediate (nonasymptotic) stage in the construction of a regular fractal.

Radius of Gyration: The root mean square (rms) radius measured from the center of mass.

Self-Affine Fractal: A fractal that can be rescaled by a transformation that requires different changes of length scale (with different exponents) in different directions.

Self-Similar Fractal: A fractal that can be rescaled by an isotropic change of length scales (by the same amount in all directions).

List of Works Cited

Birovljev, A., Furuberg, L., Feder, J., Jøssang, T., Måløy, K. J., Aharony, A. (1991), *Phys. Rev. Lett.* **67**, 584–587.

Charmet, J. C., Roux, S., Guyon, E. (Eds.) (1989), *Disorder and Fracture,* NATO ASI Series B, Vol. 235, New York: Plenum.

Coniglio, A., Jan, N., Majid, I., Stanley, H. E. (1987), *Phys. Rev.* B **35**, 3617–3620.

Family, F., Vicsek, T. (Eds.) (1991), *Dynamics of Fractal Surfaces,* Singapore: World Scientific.

Feder, J. (1988), *Fractals,* New York: Plenum.

Grossman, T., Aharony, A. (1986), *J. Phys.* A **19**, L745–L751.

Hermann, H. J., Roux, S. (Eds.) (1990), *Statistical Models for the Fracture of Disordered Materials,* Amsterdam: North Holland.

Holland-Moritz, E., Gordon, J., Kanazawa, K., Sonnenfeld, R. (1991), *Langmuir,* **1**, 1981–1987.

Jullien, R., Botet, R. (1986), *Aggregation and Fractal Aggregates,* Singapore: World Scientific.

Krug, J., Spohn, H. (1991) in: C. Godreche (Ed.), *Solids Far from Equilibrium: Growth Morphology and Defects,* Cambridge, U.K.: Cambridge Univ. Press.

Lenormand, R., Bories, S. (1980), *C. R. Acad. Sci. (Paris)* **B291**, 279–281.

Mandelbrot, B. B. (1977), *Fractals: Form, Chance and Dimension,* San Francisco: Freeman.

Mandelbrot, B. B. (1982), *Fractal Geometry of Nature,* New York: Freeman.

Mandelbrot, B. B., Passoja, D. E., Paullay, A. J. (1984), *Nature* **308**, 721–722.

Meakin, P. (1987), *CRC Crit. Rev. Solid State Mater. Sci.* **13**, 143.

Meakin, P. (1988), C. Domb and J. L. Lebowitz (Eds.), in: *Phase Transitions and Critical Phenomena,* New York: Academic, Vol. 12, p. 353–498.

Meakin, P. (1990), *Prog. Solid State Chem.* **20**, 135–233.

Nowicki, B. (1985), *Wear* **102**, 161–176.

Oppenheim, C. I., Trevor, D. J., Chidsey, C. E. D., Trevor, P. L., Sieradski, K. (1991), *Science* **254**, 687–689.

Paladin, G. and Vulpiani, A. (1987), *Phys. Rep.* **156**, 147–225.

Pietronero, L. (1988) (Ed.), *Fractals: Physical Origin and Properties,* New York: Plenum.

Saleur, H., Duplantier, B. (1987), *Phys. Rev. Lett.* **58**, 2325–2328.

Stanley, H. E., Meakin, P. (1988) *Nature* **335**, 405–409.

Stauffer, D. (1985), *Introduction to Percolation Theory,* London: Taylor and Francis.

Vicsek, T. (1989), *Fractal Growth Phenomena,* Singapore: World Scientific.

Wilkinson, D. (1984), *Phys. Rev. A* **30**, 520–531.

Wilkinson, D., Willemsen, J. (1983), *J. Phys. A* **16**, 3365–3376.

Further Reading

At the present time several dozen books concerned with fractal geometry and its applications in the physical sciences have appeared. The Mandelbrot (1977, 1982) classics (particularly *The Fractal Geometry of Nature,* Mandelbrot, 1982) are still a primary source of information. For those interested in the applications of fractal geometry to physical processes the books by Feder (1988) and Vicsek (1989) are highly recommended. A recent reprint collection (with useful commentaries) assembled by Family and Vicsek (1991) provides an up-to-date account of the rapidly developing surface growth area. Surveys of the applications to growth phenomena may be found in the books by Feder and Vicsek and recent reviews (Meakin, 1988, 1990). A collection of reviews concerned with application in chemistry has been edited by Avnir [D. Avnir (Ed.) (1989), *The Fractal Approach to Heterogeneous Chemistry: Surfaces, Colloids, Polymers,* Chichester: Wiley]. Many conference proceedings have appeared during the past decade. A selection of those most relevant to applied physics include A. Aharony and J. Feder (Eds.) (1989), *Fractals in Physics, Essays in Honour of Benoit B. Mandelbrot,* Amsterdam: North Holland; M. Fleischmann, D. J. Tildesley, and R. C. Ball (Eds.), (1990), *Fractals in the Natural Sciences,* Princeton: Princeton Univ. Press; and the proceedings edited by Pietronero (1988).

FULLERENES

M. S. DRESSELHAUS, *Department of Electrical Engineering and Computer Science and Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*

G. DRESSELHAUS, *Francis Bitter National Magnet Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*

P. C. EKLUND, *Department of Physics and Astronomy and Center for Applied Energy Research, University of Kentucky, Lexington, Kentucky, U.S.A.*

Introduction 515
1. **Structure of Fullerenes** 517
2. **Synthesis of Fullerene Solids**. 519
3. **Doping Methods and New Compounds** 519
4. **Crystalline Structure** 521
5. **Vibrational Modes** 524
6. **Electronic Structure** 527
7. **Optical Properties** 529
8. **Electrical and Thermal Transport** 532
9. **Superconductivity** 534
10. **C_{60}-Related Tubules** 536
11. **Applications** 540
Glossary 541
Works Cited 542
Further Reading 543

INTRODUCTION

Fullerenes denote large closed-cage polyhedral molecules of carbon atoms, the most common fullerene being C_{60}. In an early Japanese-language publication it was suggested that an icosahedral C_{60} molecule might be stable chemically (Osawa, 1970). Independent of this Japanese work, early Russian workers showed theoretically that C_{60} should have a large electronic gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) (Bochvar and Gal'pern, 1973; Stankevich *et al.*, 1984). These early theoretical suggestions for icosahedral C_{60} were not widely recognized and were only rediscovered after the experimental work of Kroto, Smalley, and co-workers in the middle 1980s (Kroto *et al.*, 1985) established the stability of the C_{60} molecule in the gas phase.

In the early gas-phase work, carbon clusters were synthesized by the laser ablation of a graphite surface, using a pulsed laser to generate a hot plasma of carbon atoms and radicals with temperatures over 10 000 °C. Pulses of helium gas were used to cool the plasma rapidly to form carbon clusters. The first to report clusters containing 40–100 carbon atoms was the EXXON group (Smalley, 1992), who showed mass spectra with prominent peaks at C_{60} and C_{70} and only clusters with even numbers of carbon atoms (see Fig. 1). Soon thereafter, Kroto, Smalley, and co-workers in their landmark paper (Kroto *et al.*, 1985) identified C_{60} with a molecule containing 60 carbon atoms located on equivalent sites at the vertices of a regular truncated icosahedron (Fig. 2), and identified other C_n species with polyhedral closed-cage clusters (Smalley, 1992). They gave the name buckminsterfullerene to C_{60} (Kroto *et al.*, 1985) because of its resemblance to the geodesic domes designed and built by R. Buckminster Fuller (Fuller, 1984). With time and usage, the nomenclature for this family of carbon-based molecules has been shortened to fullerenes. This early gas-phase work kindled considerable experimental and theoretical interest in fullerenes, though the activity worldwide was limited initially by the low availability of suitable samples to carry out the experiments suggested by the early work.

Starting in the 1970s astrophysicists were trying to identify some unusual infrared (ir) emissions from interstellar dust and from

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

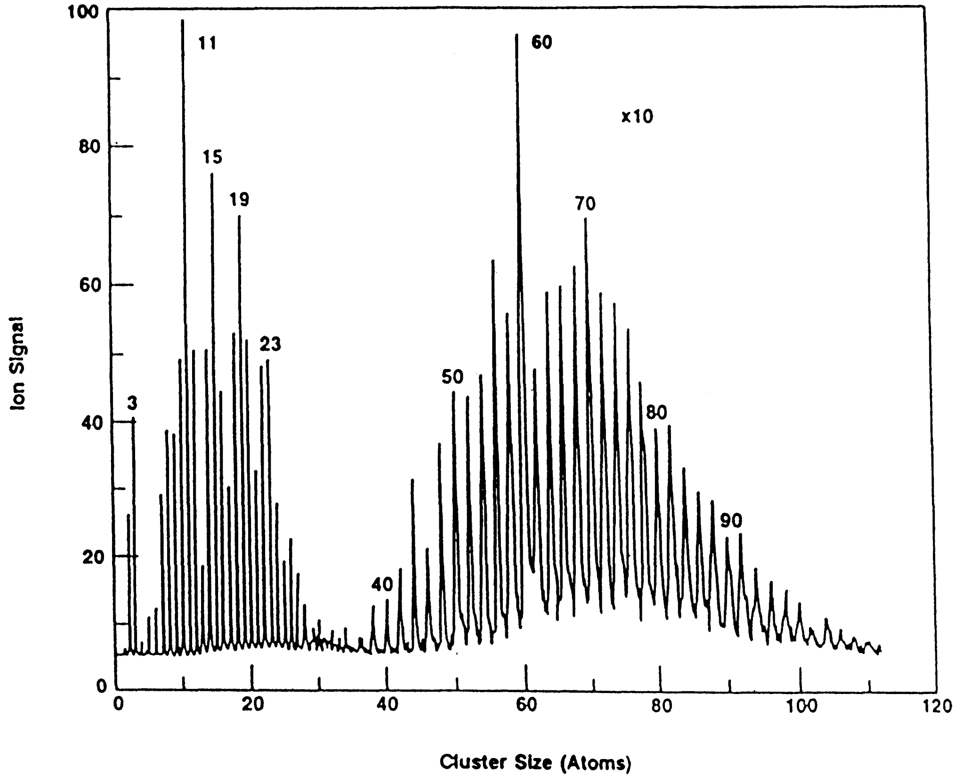

FIG. 1. Time-of-flight mass spectrum of carbon clusters produced in a supersonic beam by laser vaporization of graphite (from Rohlfing *et al.*, 1984).

carbon-rich stars. This work also suggested the possibility of larger carbon clusters (Hare and Kroto, 1992) and the linkage of these astronomical carbon clusters with carbon soot and with concentric polyhedral shell structures observed in carbon microparticles by Iijima using transmission electron microscopy (Iijima, 1991). Hare and Kroto (1992) have argued that carbon clusters similar to fullerenes play a role in the galactic carbon cycle.

In the fall of 1990 a new type of condensed matter, based on C_{60}, was synthesized for the first time. This was made possible by the breakthrough of Krätschmer, Huffman, and co-workers (Krätschmer *et al.*, 1990) in finding a simple method using a carbon arc discharge and liquid chromatography for preparing gram quantities of C_{60}, which had previously only been available in trace quantities in the gas phase (Kroto *et al.*, 1985; Smalley, 1992).

It was soon found by Haddon and co-workers at AT&T Bell Laboratories (Haddon *et al.*, 1991) that the intercalation of alkali metals into C_{60} to a stoichiometry M_3C_{60} (where M=K,Rb,Cs) greatly modifies the electronic properties of the host fullerene solid and yields relatively high-transition-temperature ($18 \leqslant T_c \leqslant 33$ K) superconductors. The discovery of superconductivity (Haddon, 1992) in these compounds clearly spurred a great deal of interest in C_{60}-related materials.

This review discusses the structure of fullerenes in Sec. 1, the techniques of synthesis in Sec. 2, the various doping techniques in Sec. 3, the crystalline structure of fullerene solids in Sec. 4, the spectroscopic results on vibrational modes in fullerenes in Sec. 5, the electronic structure in Sec. 6, the optical properties in Sec. 7, the results of carrier transport experiments in Sec. 8, and the superconducting properties of the alkali-metal–doped fullerenes in Sec. 9, while in Sec. 10 we review the structure and electronic properties of nanometer-diameter graphene tubules terminated with hemispherical fullerene caps, and finally in Sec. 11 we discuss some possible applications of fullerenes.

1. STRUCTURE OF FULLERENES

Many independent experiments discussed in this review show that crystalline materials formed from the fullerenes are molecular solids. Therefore the structure and properties of these solids are strongly dependent on the structure and properties of the constituent fullerene molecules.

From a structural standpoint, the 60 carbon atoms in C_{60} are at the vertices of a regular truncated icosahedron, where every atomic site is equivalent to every other site [see Fig. 2(a)]. Since the average nearest-neighbor carbon-carbon (C–C) distance a_{C-C} in C_{60} (1.44 Å) is almost identical to that in graphite (1.42 Å), and since most of the faces on the regular truncated icosahedron are hexagons, we can think of the C_{60} molecule as a "rolled up" graphene sheet (a single layer of crystalline graphite) in which the pentagonal faces are substituted for some of the hexagons to form a closed surface. A regular truncated icosahedron has in fact 20 hexagonal faces and 12 additional pentagonal faces to form a closed surface, in keeping with Euler's theorem, discussed below. The diameter of the C_{60} molecule is 6.83 Å, treating the carbon atoms as points (Fischer *et al.*, 1992). When one takes account of the size of the π-electron cloud associated with the carbon atoms, the outer diameter of the C_{60} molecule becomes 10.18 Å (see Table 1), consistent with a π-electron cloud of thickness 3.35 Å, which is the interlayer separation in graphite. The symmetry operations of the C_{60} molecule include 12 fivefold axes passing through the centers of a pair of pentagonal faces on opposite sides of the molecule, 20 threefold axes likewise passing through centers of a pair of hexagonal faces, 15 twofold axes going through the centers of the edges connecting two hexagons, and invariance of the entire molecule under inversion.

In the Huffman-Krätschmer arc method larger fullerenes are also formed. The most

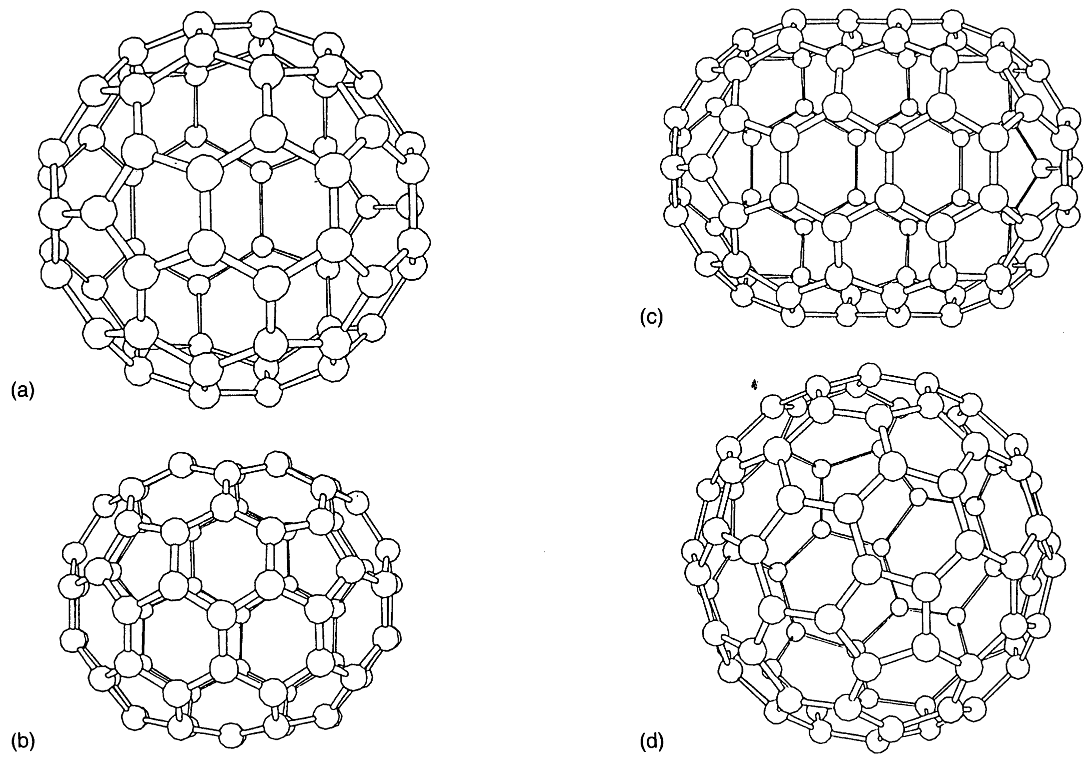

FIG. 2. (a) In the C_{60} molecule each carbon atom is at an equivalent position at the vertices of a regular truncated icosahedron. (b) The C_{70} molecule as a rugby ball. (c) The C_{80} molecule as an extended rugby ball. (d) The C_{80} molecule as an icosahedron.

Table 1. Physical constants for solid C_{60}.

Quantity	Value
Average C—C distance	1.44 Å
fcc lattice constant	14.17 Å
C_{60} mean ball diameter	6.83 Å
C_{60} ball outer diameter	10.18 Å
C_{60} ball inner diameter	3.48 Å
Tetrahedral site radius	1.12 Å
Octahedral site radius	2.07 Å
Mass density	1.72 g/cm^3
Molecular density	1.44×10^{21}/cm^3
Compressibility ($-d\ln V/dP$)	6.9×10^{-12} cm^2/dyn
Bulk modulus	14 GPa
Structural phase transitions	255 K, 165 K
Binding energy per atom	7.40 eV
Electron affinity	2.65 eV
Ionization potential (first)	7.58 eV
Ionization potential (second)	11.5 eV
Heat of formation	40 kcal/mol
Band gap (HOMO-LUMO)	1.75 eV
Spin-orbit splitting of C($2p$)	0.00022 eV
Velocity of sound v_t	2.1×10^5 cm/s
Velocity of sound v_l	3.6×10^5 cm/s
Debye temperature	165 K
Thermal conductivity (300 K)	0.4 W/mK
Phonon mean free path	50 Å
Static dielectric constant	3.9–4.2

abundant of the higher fullerenes is C_{70} [Fig. 2(b)], but significant quantities of C_{76}, C_{78}, and C_{84} have also been isolated and studied in some detail (Diederich and Whetten, 1992). Groups of large-mass carbon clusters as large as C_{240}, C_{330}, and clusters in the range C_{600}–C_{700} have been identified by mass spectroscopy (Smalley, 1992), but these larger fullerenes have not yet been isolated or studied in detail.

All of these closed-cage carbon clusters contain 12 pentagonal faces necessary to close the molecular cage and follow Euler's theorem for polyhedra,

$$f+v=e+2 \tag{1}$$

where f, v, and e are respectively the number of faces, vertices, and edges of a polyhedron. If we consider polyhedra formed by h hexagonal faces and p pentagonal faces, then

$$f=p+h,$$

$$2e=5p+6h,$$

$$3v=5p+6h, \tag{2}$$

so that

$$6(f+v-e)=p=12, \tag{3}$$

from which we conclude that all fullerenes must have 12 pentagonal faces, and the number of hexagonal faces is arbitrary. These arguments further show that the smallest possible fullerene is C_{20}, which would form a regular dodecahedron with 12 pentagonal and no hexagonal faces. Furthermore, since the addition of a single hexagon adds two carbon atoms, all fullerenes must have an even number of carbon atoms, in agreement with the observed mass spectra for fullerenes shown in Fig. 1 (Smalley, 1992).

Of the larger fullerenes, C_{70} is the easiest to make and to isolate. C_{70} is formed by adding a ring of five hexagons around the equatorial plane of the C_{60} molecule, normal to one of the fivefold axes [see Fig. 2(b)]. In contrast to the C_{60} molecule with I_h symmetry, the C_{70} molecule has the lower symmetry D_{5h}, which lacks inversion symmetry but rather has a mirror plane at the equator. Fullerenes have been found to form isomers, that is, polyhedra with a given number n of carbon atoms C_n arranged in different geometrical structures (Diederich and Whetten, 1992). For example, C_{80} might be formed in the shape of an elongated rugby ball prepared by adding two rows of five hexagons normal to a fivefold axis of C_{60} at the equator [see Fig. 2(c)]; an icosahedral form of C_{80} can also be specified as shown in Fig. 2(d). As another example of fullerene isomerism, Diederich and Whetten have reported that C_{78} has five distinct isomers.

Generally speaking, each carbon atom has four valence electrons that can participate in bonding to other atoms. In C_{60} each carbon atom has three nearest-neighbor carbon atoms and forms bonds to each of these atoms. Two of the bonds are located at the fusion of a hexagon and a pentagon [see Fig. 2(a)] and are single bonds. The third bond is located at the fusion between two hexagons and must be a double bond occupied by two (rather than one) valence electrons [see Fig. 2(a)]. The molecular structure is further stabilized by introducing a small distortion of the bond lengths to form the Kekulé structure of alternating longer single and shorter double bonds around the hexagonal face. The full icosahedral I_h symmetry is, however, preserved under these distortions. The single bonds that define the pentagonal faces are increased from the average bond length of 1.44 to 1.46 Å, while the double bonds are decreased to 1.40 Å (Johnson *et al.*, 1992). Since each carbon atom has its valence requirements fully satisfied, a crystalline solid of C_{60} is expected to be an insulator or a semiconductor. The semiconducting behavior is further stabilized by the above-mentioned distortions in the intramolecular C–C bond lengths.

2. SYNTHESIS OF FULLERENE SOLIDS

C_{60} molecules can be prepared by several methods, though many workers use a method first developed by Krätschmer and Huffman based on an alternating-current (ac) discharge between graphite electrodes in approximately 200 Torr of He (Krätschmer *et al.*, 1990). This discharge produces a carbon soot, which can contain up to ~15% of the fullerenes C_{60} and C_{70}. The smaller fullerenes are separated from the soot using Soxhlet extraction with a toluene solvent. Separation of C_{60} from C_{70} and the higher-molecular-weight fullerenes is typically accomplished by liquid chromatography (LC) (Diederich *et al.*, 1991; Chen *et al.*, 1991). The identity of the separated fractions from the LC column is verified, for example (Diederich and Whetten, 1992), by the comparison of the observed optical spectra (Sec. 7) and NMR data with published results. Further characterization by vibrational spectroscopy (including both infrared transmission and Raman spectroscopy; see Sec. 5) allows an assessment to be made of the compositional purity of a fullerene sample. Liquid chromatography using organic solvents as the elutant generally allows separation of the fullerenes according to their molecular weights, but can also separate isomers, e.g., C_{78} with C_{2v} symmetry from C_{78} with D_3 symmetry (Diederich *et al.*, 1991).

For experiments where rigid exclusion of solvent impurities is necessary, it is possible to separate C_{60} and C_{70} from the arc-derived soot without the use of solvents by heating the soot in vacuum and in a temperature gradient. In this method, the C_{60} and C_{70} both are sublimed from the soot placed in the hottest part of the furnace, and these molecules drift toward the cold end of a quartz tube connected to a vacuum line. The C_{60} and C_{70} condense in separate bands on the tube wall, and the positions of these bands are related to the different sublimation temperatures for C_{60} and C_{70}.

Actual property measurements of fullerenes are made on either microcrystalline powder samples, films, or single crystals. Microcrystalline C_{60} powder is obtained by vacuum evaporation of the solvent from the solution obtained from the LC column. These powders are further heated to ~300 °C for several hours in vacuum to remove residual solvent. Pristine C_{60} films can be deposited onto a substrate such as clean silicon (100) by sublimation of the C_{60} powder in a vacuum of $\sim 10^{-6}$ Torr. Single crystals can be grown either from solution or by vacuum sublimation (Meng *et al.*, 1992). The solution growth method tends to yield larger single crystals, but such crystals tend to contain trapped solvent. The smaller crystals (~0.5 mm in size) grown by sublimation (in vacuum or in an inert gas such as argon) tend to be more perfect structurally.

3. DOPING METHODS AND NEW COMPOUNDS

There are three distinct ways to introduce foreign atoms into a C_{60}-based solid (Smalley, 1991). One method is "endohedral" doping, which involves the addition of a rare earth, an alkaline earth, or an alkali-metal ion into the interior of the fullerene ball. Smalley (1992) has proposed that this endohedral configuration be denoted, for example, by $La@C_{60}$ to represent one endohedral lanthanum in C_{60}, or $Y_2@C_{82}$ for two Y atoms inside a C_{82} fullerene. For the large fullerenes, the endohedral addition of up to three metal ions has been demonstrated by mass-spectroscopy techniques (Smalley, 1991). It has been shown that when La is inside the fullerene ball, it has a valence of +3 (Smalley, 1991), indicating the presence of three delocalized electrons on the C_{60} ball in antibonding states that could be available for conduction in a solid of endohedrally doped C_{60}.

A second method is the substitutional doping of an impurity atom with a different valence state for a carbon atom on the surface of a fullerene ball. Since a carbon atom is so small, and since the average nearest-neighbor C–C distance $a_{C\text{–}C}$ on the C_{60} surface is only 1.44 Å (Johnson *et al.*, 1992), the only species that can be substituted for a carbon atom on the C_{60} ball surface is boron, making the charged ball *p* type. Smalley and co-workers have demonstrated that it is possible to replace more than one carbon atom by boron on a given ball (Smalley, 1991). Also for graphite, the only substitutional dopant is boron, and for the same reasons as for C_{60}. However, for diamond, which has a larger C–C nearest-neighbor distance of $a_{C\text{–}C}=1.544$ Å, both boron and nitrogen can enter the lattice substitutionally (Field, 1979). It has also been reported that it is possible to place a potassium atom endohedrally inside the C_{60} ball while at the same time substituting a boron for a carbon atom on the surface of the ball (Smalley, 1991).

In the third method of doping fullerene solids (also called intercalation), the dopant (e.g., an alkali metal or an alkaline earth, *M*) is introduced into the interstitial positions between adjacent balls (exohedral locations). In this doping method, charge transfer can take place between the *M* atoms and the balls. This method of doping forms exohedral fullerene solids and closely parallels the process of intercalating an alkali-metal ion between carbon layers in graphite (Dresselhaus and Dresselhaus, 1981).

Among the alkali metals, Li, Na, K, Rb, and Cs have all been used as exohedral dopants for C_{60}, as well as intermetallic mixtures of these alkali metals such as Na_2K and Rb_xCs_y (Haddon, 1992). These guest atoms are called "donor-type intercalants" in analogy to the graphite intercalation compound nomenclature (Dresselhaus and Dresselhaus, 1981).The doping of C_{60} with alkali metals can be achieved in a two-zone furnace, similar to the apparatus used to prepare alkali-metal graphite intercalation compounds (Dresselhaus and Dresselhaus, 1981). The intercalation of alkali metals works well in the preparation of powdered samples, whereby the stoichiometry is controlled by the number of moles of alkali metal introduced into the sealed ampoule (quartz or Pyrex) relative to the number of moles of C_{60}, the temperatures of the reagents, and the duration of the reactions. Alkali-metal doping of thin (~1000 Å) C_{60} films can be carried out similarly, using a sealed quartz (or Pyrex) tube (with approximate dimensions 25 cm long × 1 cm diameter) which contains the C_{60} film/Si substrate sample and the alkali metal at opposite ends. The films are maintained at a higher temperature (200 °C) than the alkali metal (*M*) (100–180 °C) to avoid condensation of the alkali metal on the film surface. The progress of the reactions can be monitored qualitatively by the color change observed in the films (see Sec. 7) and more quantitatively by the doping-induced downshift of the strongest C_{60} Raman line (see Sec. 5) or by the magnitude of the electrical conductivity (see Sec. 8). The doping reaction in C_{60} single crystals is carried out in a two-zone furnace for long times (~1 to 2 weeks) because of the larger distances that the dopant must travel (Xiang *et al.*, 1992).

Some success has also been achieved with the intercalation of alkaline-earth metal dopants, e.g., Ca, for which charge transfer also takes place (Kortan *et al.*, 1992; Chen *et al.*, 1992). The intercalation of alkali-metal and rare-earth metal atoms into C_{60} leads to the transfer of their valence electrons to the C_{60} ball and the formation of positive ions (i.e., K^+ or Ca^{2+}).

Doping fullerenes with acceptor-type dopants, which become negatively charged ions inside the host material, has been more difficult because of their high electron affinities (see Table 1). A stable crystal structure has nevertheless been identified upon doping C_{60} with iodine, where the iodine atoms form a sheet between layers of C_{60} balls; no measurable conductivity has, however, been observed in the case of iodine doping of fullerenes, indicating that charge transfer between the iodine and C_{60} probably does not occur (Fischer *et al.*, 1992).

In a totally different approach to the doping problem, varying numbers of fluorine atoms have been added via radial F—C bonds to the outside of the C_{60} ball, where the attachment of a F atom converts a double bond in C_{60} into a single bond (Fischer *et al.*, 1992). A report has been given of adding up to 60 fluorine atoms per ball exohedrally to form the compound $C_{60}F_{60}$ (Fischer *et al.*, 1992). Similar bonding arrangements of C_{60} to hydrogen have also been observed up to a maximum hydrogen concentration, $C_{60}H_{60}$ (Fischer *et al.*, 1992). However, it is not known whether all the H bonds are directed radially outward or whether some may perhaps be endohedrally bonded. Several stable bromine-doped fullerenes have also been reported including $C_{60}Br_6$, $C_{60}Br_8$, and $C_{60}Br_{24}$, and proposals for their structures have also been made (Birkett *et al.*, 1992).

Synthetic fullerene chemistry has become a very active research field, largely because of the novelty of the C_{60} molecule and the variety of chemical modifications to the C_{60} shell that appear to be possible. Since many of the fullerene reactions appear to be reversible, and the amount of the fullerene material available to most chemists is small, research in this field has been especially challenging. Some of the simplest of the fullerene compounds, such as $C_{60}H_x$, $C_{60}F_x$, and $C_{60}Br_x$, have already been mentioned. More complex compounds have also been synthesized, and many of these are based on forming covalent bonds to each of the two carbon atoms joining two hexagons [Fig. 2(a)]. As various moieties bond to these two carbons, the latter are pulled out of their equilibrium positions, and the local structure of the fullerene is modified.

One of the simplest compounds that distorts carbon bonds on the C_{60} molecule is the epoxide $C_{60}O$ (Fischer *et al.*, 1992), which crystallizes in the same simple-cubic space group as pristine C_{60} as discussed in Sec. 4. Many more complicated moieties can also be attached at the double-bond site. The addition of phenyl groups through a carbon bridge atom has formed the basis of a large number of new and interesting compounds (Wudl *et al.*, 1992; Fagan *et al.*, 1992). In some cases the bonding at the double-bond sites is done through a metal atom such as Pt, Os, or Ir to which other groups attach. The first structurally characterized metal complex of C_{60} was $(t\text{-BuC}_5H_5N)_2OsO_4$. More recently, many metal-containing derivatives of C_{60} have been synthesized (Fagan *et al.*, 1992; Hammond and Kuck, 1992). In preparing many of these compounds, specific groups are added for special chemical reasons, while other chemical species are used to achieve a chemically stable compound. From these synthesis studies, it is concluded that C_{60} behaves as an electron-deficient species. Since the double bonds of C_{60} are strained by the curvature of the C_{60} ball, it is expected that C_{60} will be reasonably reactive. The effect of forming a metal derivative complex is to remove one double C=C bond, lengthen that bond (e.g., to 1.50 Å for Pt attachment), modify the charge distribution locally, raise the energy of the lowest unoccupied molecular orbital, and lower the electron affinity (Fagan *et al.*, 1992). The attachment of aromatic groups has also been shown to link fullerenes (Wudl *et al.*, 1992). Most of the novel fullerene synthetic chemistry has been done in solution. As new fullerene compounds are prepared, physical chemists are measuring their properties. Thus, the preparation of new chemical compounds based on fullerenes and the study of their physical properties are now important topics in chemical research.

One especially interesting direction that this synthetic research has taken is the preparation of an organic ferromagnet with a Curie temperature of 16.1 K by Wudl and co-workers (Allemand *et al.*, 1991). This material, exhibiting almost no magnetic hysteresis, was synthesized by the reaction of the organic donor TDAE [tetrakis(dimethyl-amino)ethylene] complex with C_{60} (Wudl, 1992). The "soft ferromagnetism" for this material showed no Curie-Weiss behavior of the susceptibility for the temperature range above the Curie temperature. The significance of this work is the demonstration of magnetism in systems having only *s* and *p* orbitals. Typically magnetic materials are identified with systems having unfilled *d* and *f* orbitals, such as transition metals (e.g., Fe, Ni, Co) and rare-earth elements (e.g., Sm, Eu, Ce).

4. CRYSTALLINE STRUCTURE

In the solid state, C_{60} molecules crystallize into a cubic structure with a lattice constant of 14.17 Å (Fischer *et al.*, 1992) and a density of 1.72 g/cm^3 (corresponding to 1.44×10^{21} C_{60} molecules/cm^3). At room temperature the molecules are rotating, and the crystal structure is face-centered cubic (fcc) with four equivalent spinning C_{60} balls per unit cell. The pertinent space group is denoted by O_h^5 or $Fm\bar{3}m$. This structure has been established directly by x-ray diffraction. The equivalence of each carbon site on a C_{60} ball is confirmed by a single sharp line in the NMR spectrum

(Johnson *et al.*, 1992). Relative to the other allotropic forms of carbon, solid C_{60} is relatively more compressible, with an isothermal volume compressibility (see Table 1) of 6.9 $\times 10^{-12}$ cm^2/dyn (Fischer *et al.*, 1992), approximately three times larger than graphite because the van der Waals charge cloud around the C_{60} balls can easily be compressed in three dimensions, rather than in one dimension as in the case of graphite perpendicular to the carbon layers.

Below about 250 K, the C_{60} balls become oriented, and the structure becomes simple cubic (space group T_h^6 or $Pa\bar{3}$), with four C_{60} molecules per unit cell as the four balls within the fcc unit cell become inequivalent (Fischer *et al.*, 1992). We can understand the crystal structure of the idealized ordered phase by referring to Fig. 3(a), where the standard orientation of a C_{60} molecule is shown with respect to a cubic coordinate system. Here [100] axes pass through three mutually orthogonal twofold molecular axes (the centers of the hexagon–hexagon edges) and four [111] axes pass through the centers of the hexagonal faces. The alternative standard orientation is shown in Fig. 3(b), which is obtained from Fig. 3(a) through rotation by $\pi/2$ about the [100] axes. In the low-temperature idealized ordered phase, the centers of four C_{60} molecules are located at coordinate locations (0,0,0), $(\frac{1}{2},0,\frac{1}{2})$, $(\frac{1}{2},\frac{1}{2},0)$and $(0,\frac{1}{2},\frac{1}{2})$, and each molecule is rotated from the standard orientation shown in Fig. 3(a) by 22° about the threefold [111], $[\bar{1}\,\bar{1}\,1]$, $[\bar{1}1\bar{1}]$, and $[1\,\bar{1}\,\bar{1}]$ axes, respectively. This idealized orientation is stabilized by bringing the electron-rich double-bond hexagon-hexagon edge of one C_{60} molecule adjacent to the electron-poor pentagonal faces of the adjacent C_{60} molecule. Supporting evidence for a phase transition of the C_{60} molecules associated with orientational ordering is provided by many property measurements, such as differential scanning calorimetry, resistivity, NMR, ultrasonic attenuation, Raman spectroscopy, and thermal conductivity (Fischer *et al.*, 1992). An additional phase transition has also been reported to occur at still lower temperatures ($\sim$160 K), and some authors have proposed additional phase transitions. As-prepared powdered C_{70}, on the other hand, has been reported to form a hexagonal structure at room temperature with lattice constants $a=10.63$ Å and $c=17.39$ Å (Fischer *et al.*, 1992), though other structures are stable in different temperature ranges.

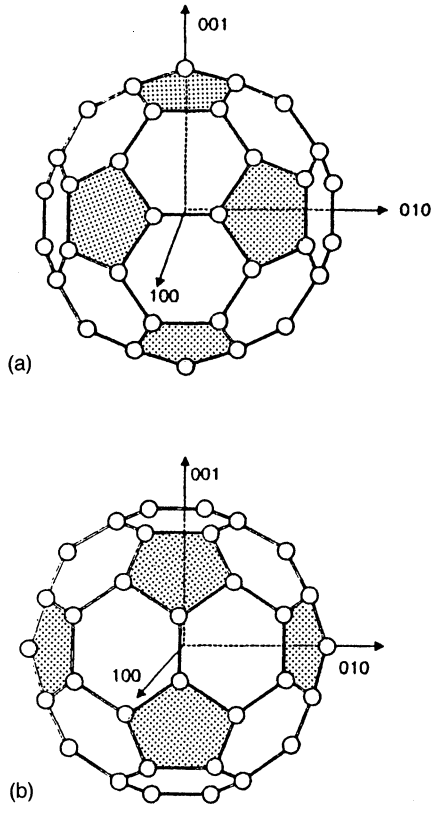

FIG. 3. (a) The C_{60} molecule shown in one of two "standard" orientations such that the Cartesian axes in a cubic crystal pass through three orthogonal twofold axes. (b) The other "standard" orientation, related to that in (a) by a rotation of $\pi/2$ about the (100) Cartesian axis.

Several stable crystalline phases for doped C_{60} have been identified. Most widely studied are the crystalline phases formed with alkali metals, though some structural work has also been done, for example, with alkaline-earth metals (Kortan *et al.*, 1992) and halogen intercalants (Zhu *et al.*, 1992). Fleming *et al.* at AT&T Bell Laboratories first reported that when C_{60} is doped with the alkali metals (M=K,Rb,Cs), stable crystalline phases are formed for the compositions M_3C_{60}, M_4C_{60}, and M_6C_{60} (Haddon, 1992). One electron is transferred to the C_{60} ball per M-atom dopant, resulting in M^+ ions at the tetrahedral and/or octahedral interstices of the C_{60} host structure (see Fig. 4). For the composition M_3C_{60}, the resulting metallic crystal exhibits a fcc arrangement for the C_{60} balls (see Fig. 4). Within this fcc host structure the alkali-metal ions can be

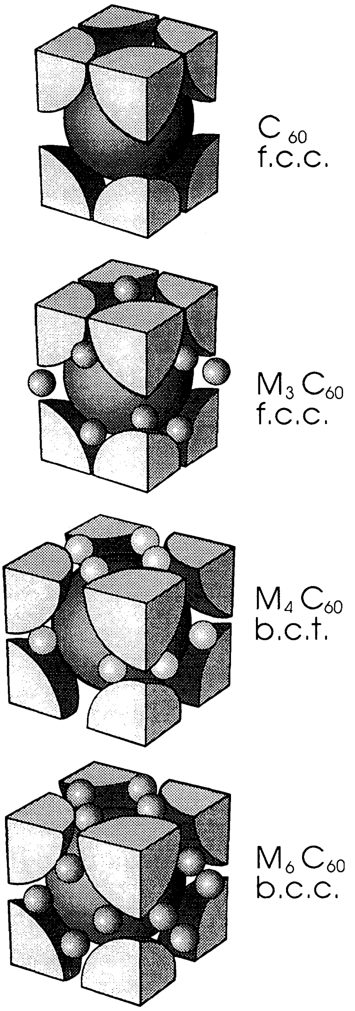

FIG. 4. Structures for solid C_{60}, M_3C_{60}, M_4C_{60}, and M_6C_{60} (M=K,Rb,Cs). The large balls denote C_{60} molecules and the small balls are alkali-metal atoms. For fcc M_3C_{60}, which has four C_{60} balls per cubic unit cell, the M atoms can be on either octahedral or tetrahedral sites. Undoped solid C_{60} above 250 K also exhibits an fcc crystal structure, but in this case all tetrahedral and octahedral sites are unoccupied. For bcc M_6C_{60} all the M atoms are on distorted tetrahedral sites. For bct M_4C_{60} the dopant is also found on distorted tetrahedral sites (from Fleming *et al.*, 1991b).

located either at tetrahedral $(\frac{1}{4},\frac{1}{4},\frac{1}{4})$ sites or at octahedral $(\frac{1}{2},0,0)$ interstitial sites (referenced to a simple cubic coordinate system). In the fcc structure, x-ray diffraction data indicate that the positively charged alkali-metal ions lie adjacent to an electron-rich C=C double bond. For example, from the standpoint of the near neighbors, the K guest species in K_3C_{60} are located in the cavity between adjacent C_{60} balls, on any one of four tetrahedral sites or two octahedral sites. For the tetrahedral sites, the radius of the cavity available to a K^+ ion is 1.12 Å, while for an octahedral site it is 2.07 Å (Stephens *et al.*, 1991). The shortest distance between a K^+ ion at a tetrahedral site and the carbon atom on a freely rotating C_{60} ball is 2.66 Å (Stephens *et al.*, 1991). It is of interest to note that the Cs^+ ion is apparently too large to form a stable Cs_3C_{60} compound with the fcc structure, and this is believed to be the reason why Cs_3C_{60} [which crystallizes in a body-centered cubic (bcc) structure] is not superconducting (see Sec. 9).

The two other stable phases for M_xC_{60} (namely, M_4C_{60} and M_6C_{60}) each have different crystalline structures. The M_4C_{60} phase, which is more difficult to prepare than M_3C_{60} or M_6C_{60}, has a body-centered tetragonal structure (bct) with a tentatively assigned space group *I4/mmm* (Fleming *et al.*, 1991b). The compound M_6C_{60} has a bcc structure (see Fig. 4) and corresponds to the space group T_h^5 or *Im*3. The compounds K_6C_{60}, Rb_6C_{60}, and Cs_6C_{60} all correspond to alkali-metal–saturated compounds, which are all semiconductors. A list of lattice constants for the various stable C_{60}-related structures is given in Table 2 for pristine C_{60} and alkali-metal–doped M_xC_{60} compounds (Fleming *et al.*, 1991b).

Since the ionic radii of the alkali-metal ions Na^+ and Li^+ are much smaller than those of the K^+, Rb^+, and Cs^+ ions, a number of unusual structural phenomena are observed. First, it is possible to fit more than one Na^+ ion in an octahedral site (Rosseinsky *et al.*, 1992; Tanigaki *et al.*, 1992), so that the fcc structure can be preserved upon adding up to six Na atoms per C_{60} with two Na ions going into tetrahedral sites and up to four on octahedral sites to form Na_6C_{60}. There is no stable Na_3C_{60} crystal, and materials prepared with this stoichiometry tend to phase separate (disproportionate) into Na_2C_{60} and Na_6C_{60} (Rosseinsky *et al.*, 1992). Also because of the small ionic radius of Na, the Na ions tend to fill the smaller tetrahedral sites when alloys of Na with heavier alkali metals are used as dopants, while the heavier alkali metals tend to be on octahedral sites. One unusual effect (Stephens, 1992) is that alkali-metal doping with Na_2Cs, Na_2Rb, Na_2K, and Li_2Cs results in a crystal with closer ball–ball separation than is present in undoped C_{60} (see Table 2), presumably due to the attractive forces of the

Table 2. Lattice constants and superconducting T_c values for stable C_{60}-related materials.

M_xC_{60}	Phase	Site Sym.[a]	Lattice constants (Å)	T_c (K)
C_{60}	fcc	-	14.11	-
Ca_5C_{60}	fcc	O and T	14.01	8.4
Na_2KC_{60}	fcc	O and T	14.025	2.5
Na_2RbC_{60}	fcc	O and T	14.028	2.5
Li_2CsC_{60}	fcc	O and T	14.120	12
Na_2CsC_{60}	fcc	O and T	14.134	12
K_3C_{60}	fcc	O and T	14.253	19.28
K_2RbC_{60}	fcc	O and T	14.253	21.80
K_2CsC_{60}	fcc	O and T	14.253	24
$K_{1.5}Rb_{1.5}C_{60}$	fcc	O and T	14.253	22.15
Rb_2KC_{60}	fcc	O and T	14.336	24
Rb_3C_{60}	fcc	O and T	14.436	22.9
Rb_2CsC_{60}	fcc	O and T	14.493	31.3
K_4C_{60}	bct	T[b]	11.886(*a*), 10.774(*c*)	
Rb_4C_{60}	bct	T[b]	11.962(*a*), 11.022(*c*)	-
Cs_4C_{60}	bct	T[b]	12.057(*a*), 11.443(*c*)	-
K_6C_{60}	bcc	T	11.385[c]	-
Rb_6C_{60}	bcc	T	11.548[c]	-

[a] The octahedral and tetrahedral site symmetries are denoted by O and T, respectively.

[b] Distorted tetragonal site. For the body-centered-tetragonal structure (bct), the in-plane and *c*-axis lattice constants are denoted by *a* and *c*, respectively.

[c] To compare a_{fcc} to a_{bcc}, use the relation $a_{\text{bcc}} = a_{\text{fcc}}/\sqrt{2}$.

electrostatic interactions between the alkali-metal ions and the charged C_{60} molecules.

The structure of the alkaline-earth compound Ca_yC_{60} (for $y \leqslant 5$) has the same fcc space-group symmetry as the heavy alkali-metal M_xC_{60} compounds ($x \leqslant 3$) (Kortan *et al.*, 1992), and the Ca atoms occupy both tetrahedral and octahedral sites. Because of the smaller size of the Ca^{2+} ion [as is also the case for the Na^+ ion (Rosseinsky *et al.*, 1992)], the octahedral sites can accommodate multiple Ca ions, and it is believed that up to three Ca ions can be accommodated in a single octahedral site (Kortan *et al.*, 1992). Also for Ca_5C_{60}, the lattice constant $a_0 = 14.01$ Å is smaller than that for C_{60} prior to doping (see Table 2). Unlike alkali-metal–doped C_{60}, where each *M* atom transfers one electron to C_{60}, the alkaline-earth metal dopants (such as Ca) have been reported to transfer two electrons per metal dopant for low metal concentrations, but less than two electrons per metal dopant for the highest dopant concentrations (Kortan *et al.*, 1992).

The high electronegativity of pristine C_{60} hinders acceptor doping. Consequently, relatively few C_{60}-based structural studies have thus far been made with acceptors. Solid-state C_{60} structure studies by Fischer and co-workers (1992) show an uptake of iodine to a composition $C_{60}I_4$, and in this compound the iodine molecules lie in planes between layers of C_{60} balls. Calculations of the electronic structure of other proposed halogenated C_{60} materials (Oshiyama *et al.*, 1992) suggest that such phases would be thermodynamically stable. The prototypical acceptor dopant would appear to be oxygen. The epoxide $C_{60}O$ crystallizes in the same crystal structure as C_{60} (i.e., space group $Pa\overline{3}$) and undergoes a first-order phase transition at ~280 K from a phase above, in which the $C_{60}O$ balls are spinning, to a phase where the molecules are ordered. We note that the transition temperature in $C_{60}O$ is about 30 K higher than the corresponding transition temperature in C_{60} (Heiney *et al.*, 1992). Heiney and co-workers have argued that the physical basis for the similarity in crystal structure between C_{60} and $C_{60}O$ is related to the high probability that the attached oxygen is directed to an octahedral or tetrahedral cavity in the cubic structure, leaving an essentially unaltered C_{60} ball at each lattice site.

5. VIBRATIONAL MODES

Because of the high symmetry of the C_{60} molecule (point group I_h), there are only 46 distinct mode frequencies corresponding to the $180-6=174$ degrees of freedom for the isolated C_{60} molecule, and of these, only 4 are infrared active (all with T_{1u} symmetry) and 10 are Raman active (2 with A_g symmetry and 8 with H_g symmetry). Raman and infrared spectroscopies provide sensitive methods for distinguishing C_{60} from higher fullerenes with lower symmetry (e.g., C_{70} has D_{5h} symmetry). Since most of the higher-molecular-weight fullerenes have lower symmetry as well as more degrees of freedom, they have many more infrared- and Raman-active modes.

The simplicity of the infrared (ir) spectrum of solid C_{60} [see Fig. 5(a)], which shows four prominent spectral lines, provides a convenient method for characterizing the compositional purity of C_{60} samples, especially with regard to their contamination with C_{70} [see Fig. 5(b)]. The ir spectrum of solid C_{60} remains almost unchanged relative to that of the isolated C_{60} molecule, with only the addition of a few weak features, such as the peak at 1539 cm^{-1} in Fig. 5(a). This is indicative of the molecular nature of solid C_{60}. The ir spectra shown in Fig. 5 are for thin C_{60} and C_{70} films on alkali halide substrates. The positions of the toluene bands indicate how ir spectroscopy might be used to monitor residual sol-

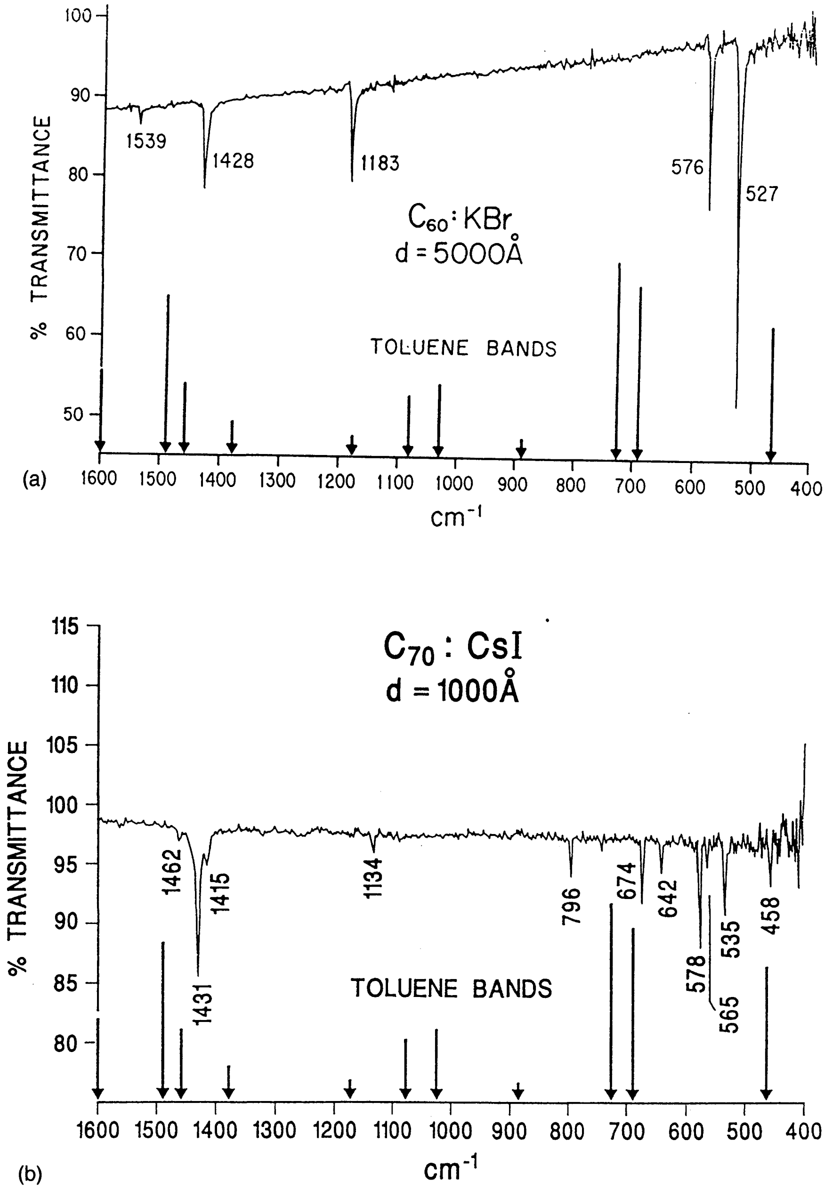

FIG. 5. The infrared spectra of (a) C_{60} (on a KBr substrate) and (b) C_{70} (on a CsI substrate) for the indicated fullerene film thicknesses. Also shown schematically are the ir bands for toluene, a common solvent for fullerenes.

vent in fullerene-based solid-state samples at the 1%–2% level.

Likewise, Raman spectroscopy (see Fig. 6) provides valuable information about intramolecular and intermolecular bonding in solid C_{60} and C_{60}-related compounds (K. A. Wang *et al.*, 1992). The Raman spectrum in Fig. 6 for solid C_{60} shows 10 strong Raman lines, the number of Raman-allowed vibrational modes expected for the free molecule. These spectral lines are therefore assigned to intramolecular modes. The normal modes in molecular C_{60} above about 1000 cm^{-1} involve carbon-atom displacements that are predominantly tangential to the C_{60} ball surface, while the modes below $\sim$800 cm^{-1} involve predominantly radial motions. The A_g breathing mode (1469 cm^{-1}) corresponds to breathing-type tangential displacements of the five carbon atoms around each of the 12 pentagons. Under high laser flux from an Ar-ion laser, this mode frequency shifts down irreversibly to 1458 cm^{-1}, which has been interpreted as a signature of a photoinduced transformation of undoped solid C_{60} (Zhou *et al.*, 1992a).

In the solid phase, cubic crystal-field interactions lower the symmetry of the molecule, causing a splitting of the fivefold-degenerate H_g-derived molecular modes into threefold (T_g) and twofold (E_g) lattice modes (Dresselhaus *et al.*, 1992a). The study of spectral line splittings and their symmetry in solid C_{60} requires the use of well-annealed single crystals, polarization studies, and studies of the temperature dependence of the Raman spectra to uncover the nature of the solid state interaction. These techniques are also needed to monitor the effect of the 250-K phase transition on the spectral line shifts, linewidths, line splittings, and their symmetries.

The addition of alkali-metal dopants to the metal-saturated stoichiometry M_6C_{60} (where M=K,Rb,Cs) perturbs the Raman spectra only slightly (Eklund *et al.*, 1992), with very little change from one alkali-metal dopant to another. For metallic K_3C_{60} (see Fig. 6) and Rb_3C_{60}, the coupling between the phonons and a low-energy electronic continuum in these metallic systems strongly broadens the

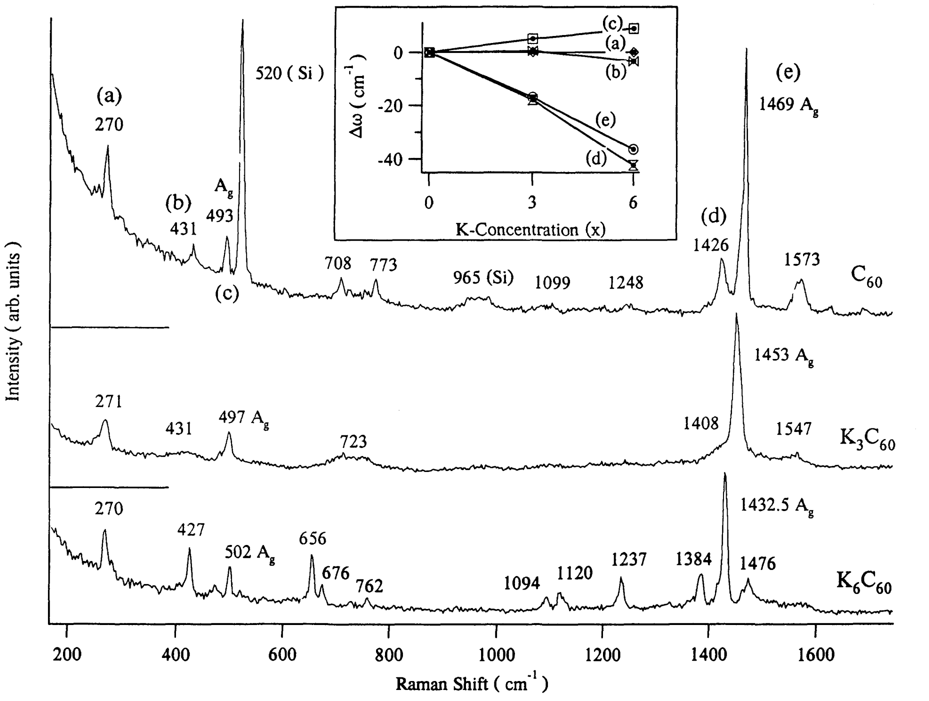

FIG. 6. Room-temperature Raman spectra (resolution $\sim$6 cm^{-1}) from films of undoped C_{60}, K_3C_{60}, and K_6C_{60} (from K. A. Wang *et al.*, 1992). Si denotes the silicon substrate. The inset shows the potassium concentration (x) dependence of five representative Raman modes in K_xC_{60} ($0 \leqslant x \leqslant 6$). Several of the peaks for K_6C_{60} (and M_6C_{60}, M=Rb,Cs) films are doublets, which are not resolved in these data.

H_g-derived modes (K. A. Wang *et al.*, 1992) and gives rise to modifications in the Raman line shape.

As a result of alkali-metal doping, electrons are transferred to the π-electron orbitals on the surface of the C_{60} molecules, elongating the C–C bonds and downshifting the intramolecular tangential modes. In the inset in Fig. 6, the dependence of the Raman frequencies on K concentration x for five dominant modes in K_xC_{60} is presented (Zhou *et al.*, 1992b). The softening of the 1469-cm^{-1} tangential A_g mode by alkali-metal doping is used as a convenient method to characterize the stoichiometry x of K_xC_{60} samples. The radial A_g mode, on the other hand, stiffens slightly because of competing effects associated with a mode softening arising from charge-transfer effects (similar to the situation for the tangential modes) and a larger mode-stiffening effect due to electrostatic interactions between the negatively charged C_{60} ion and the surrounding positively charged alkali-metal ions as their atomic separations change during a normal-mode vibrational displacement (Eklund *et al.*, 1992).

The low-frequency acoustic and rotational-librational modes associated with the displacements and rotations of the C_{60} balls in a cubic lattice have been calculated (Y. Wang *et al.*, 1991), and experimental low-frequency Raman scattering measurements on single crystals of C_{60} have been performed.

Closely related to these spectroscopic studies of the lattice properties are measurements of the bulk modulus, the pressure-dependent Raman spectra, the ultrasonic propagation and attenuation, and the specific heat (see Table 1). Models for the intramolecular modes have been used to study the electron–phonon interaction, with particular emphasis on its connection to the observed superconductivity in the M_3C_{60} compounds.

6. ELECTRONIC STRUCTURE

On the basis of the spectroscopic studies discussed in Sec. 5, we conclude that the fullerenes form molecular solids. Thus we expect their electronic energy band structures to be closely related to the electronic levels of the isolated molecules. Each carbon atom in C_{60} has two single bonds along adjacent sides of a pentagon and one double bond between two adjoining hexagons. If these bonds were coplanar, they would be very similar to sp^2 trigonal bonding in graphite. The curvature of C_{60} causes the planar-derived trigonal orbitals to hybridize with the remaining radial p orbitals, thereby giving rise to different electronic states than in graphite. The shortening of the double bonds and lengthening of the carbon atoms in the single bonds in the Kekulé arrangement of the carbon atoms in the C_{60} molecule also strongly influence the electronic structure.

The most extensive calculations of the electronic structure of fullerenes thus far have been done for C_{60}. To obtain the molecular orbitals for C_{60}, the multielectron problem must be solved (Herzberg, 1945). Results for C_{60} are shown in Fig. 7(a) for the free molecule (Saito and Oshiyama, 1991). Because of the molecular nature of solid C_{60}, the electronic structure for the solid phase would be expected to be closely related to that of the free molecule, and some authors have followed this approach. More attention has, until recently, been given to a band-structure approach. An example of a band-structure model is shown in Fig. 7(b) where we see results for the fcc solid, based on total-energy calculations using the local-density approximation (Saito and Oshiyama, 1991). Various band calculations for C_{60} yield a narrow-band ($\sim$0.4–0.6 eV bandwidth) solid, with a HOMO-LUMO–derived direct band gap of $E_{HL} \sim 1.5$–1.7 eV separating the lowest unoccupied molecular orbital (LUMO) from the highest occupied molecular orbital (HOMO). Since the HOMO and LUMO levels have the same odd parity, electric dipole transitions between these levels are symmetry forbidden in the free C_{60} molecule. In the solid, transitions between the direct band gap states at the Γ and X points in the fcc Brillouin zone are also forbidden, although they may be allowed at many other lower symmetry points in the Brillouin zone. The allowed electric dipole transitions for the free C_{60} molecule are indicated by arrows in Fig. 7(a). Experiments, such as luminescence emission and optical absorption, show different photon threshold energies, since both processes require assistance by vibrational excitations.

Photoemission and inverse photoemission experiments reported by Weaver and coworkers on M_xC_{60} for M=K, Rb, and Na and

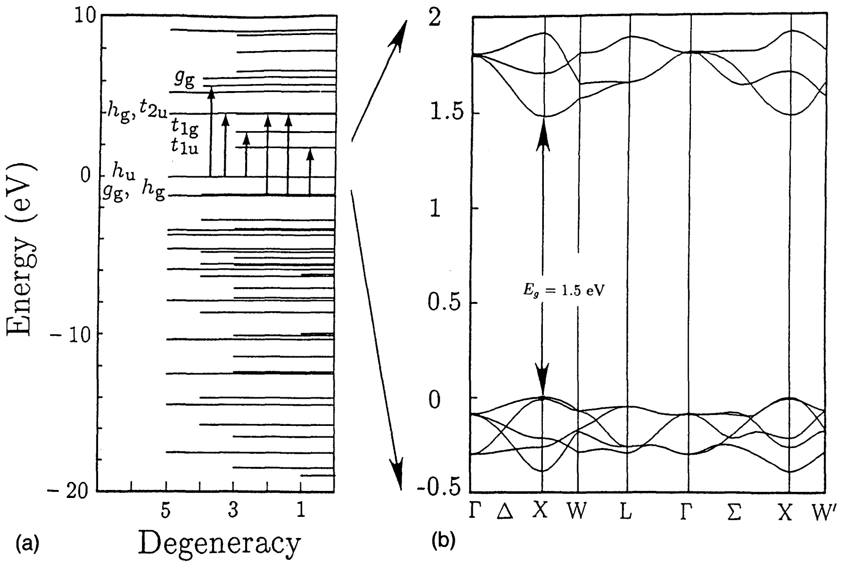

FIG. 7. Calculated electronic structure for (a) an isolated C_{60} molecule and (b) the fcc solid C_{60} where the direct band gap at the X point is 1.5 eV (from Saito and Oshiyama, 1991).

their alloys (Weaver, 1992) show direct evidence for filling of the LUMO levels upon alkali-metal doping, consistent with optical and transport measurements. The implied energy gap in the density of states for these experiments is, however, sensitive to screening and correlation effects.

Doping C_{60} with an alkali metal transfers electrons to the LUMO levels, which, because of their T_{1u} symmetry (Fig. 7), can accommodate three spin-up and three spin-down electrons. Total-energy calculations on K_xC_{60} show that the tetrahedral sites fill up before the octahedral sites (Saito and Oshiyama, 1991). Assuming one electron to be transferred to the C_{60} ball per alkali-metal atom dopant, the LUMO (T_{1u}-derived) levels are expected to be half occupied at the alkali-metal stoichiometry M_3C_{60} and totally full at M_6C_{60}, leading to a filled-shell configuration with A_g symmetry. Thus M_6C_{60} would be expected to be semiconducting with a band gap between the T_{1u} and T_{2g} levels (see Fig. 7), while M_3C_{60} should be metallic provided that no band gap is introduced at the Fermi level by a Peierls distortion or a Jahn-Teller distortion of the molecule.

Fermi-surface calculations for K_3C_{60} indicate a small hole Fermi surface around the Γ point and a larger and more complicated Fermi surface consisting of both electron and hole orbits (see Table 3) (Erwin and Pederson, 1991). No results from direct measurements of the Fermi-surface topology for any of the M_3C_{60} materials are yet available. Because of the weak interaction of the balls with each other and with the alkali-metal dopants, solid M_3C_{60} should also be viewed as a molecular metallic solid having energy bands with little dispersion. This gives rise to a very high density of states near the Fermi level, which is important in understanding superconductivity in M_3C_{60} (see Sec. 9).

Table 3. Physical constants for M_3C_{60} (M=K,Rb).

Quantity	K_3C_{60}	Rb_3C_{60}
T_c	19.3 K	29.6 K
fcc lattice constant	14.253 Å	14.436 Å
$-d \ln a/dP$	1.20×10^{-2} GPa^{-1}	1.52×10^{-2} GPa^{-1}
Bulk modulus	28 GPa	22 GPa
Cohesive energy	24.2 eV	-
Heat of formation	4.9 eV	-
Density of states	25 states/ (eV/C_{60})	35 states/ (eV/C_{60})
Electron effective mass	$1.3m_e$	-
Hole effective mass	$1.5m_e$,$3.4m_e$	-
Resistivity	2.5×10^{-3} Ω cm	-
Radial A_g mode frequency	493.0 cm^{-1}	495.0 cm^{-1}
Tangential A_g mode frequency	1468.5 cm^{-1}	1449 cm^{-1}

7. OPTICAL PROPERTIES

During early research on the separation of C_{60} from impurities and higher fullerenes using liquid chromatography (Ajie *et al.*, 1990), it became apparent that fullerenes in solution (organic solvents) exhibited characteristic colors that would be useful for identifying and purifying these fullerenes. For example, C_{60} and C_{70} appear magenta and reddish-orange, respectively, in toluene and benzene solutions.

More detailed studies show that the optical properties of fullerenes are complex and provide important insights into the electronic structure of the fullerenes. Since the HOMO–LUMO optical transitions in C_{60} and C_{70} are symmetry forbidden, a great deal of attention has focused upon understanding the optical behavior near the fundamental absorption edge, which is very sensitively studied by optical transmission spectra. Such spectra are shown in Figs. 8(a) and 8(b) versus wavelength in the ultraviolet (uv)–visible range (200–700 nm) for C_{60} and C_{70}, respectively, in hexane solution (Ajie *et al.*, 1990). The strong absorption bands in the solution spectra are identified with electric-dipole-allowed transitions between occupied (bonding) and empty (antibonding) sp^2-hybridized molecular orbitals above the LUMO level. For a C_{60} molecule, electric dipole transitions between the one-electron HOMO level with h_u symmetry and the one-electron LUMO level with t_{1u} symmetry are forbidden.

Using many-electron molecular states (Herzberg, 1945), we identify the fully occupied ground-state HOMO level with a singlet S_0 state with A_g symmetry. Allowed electric

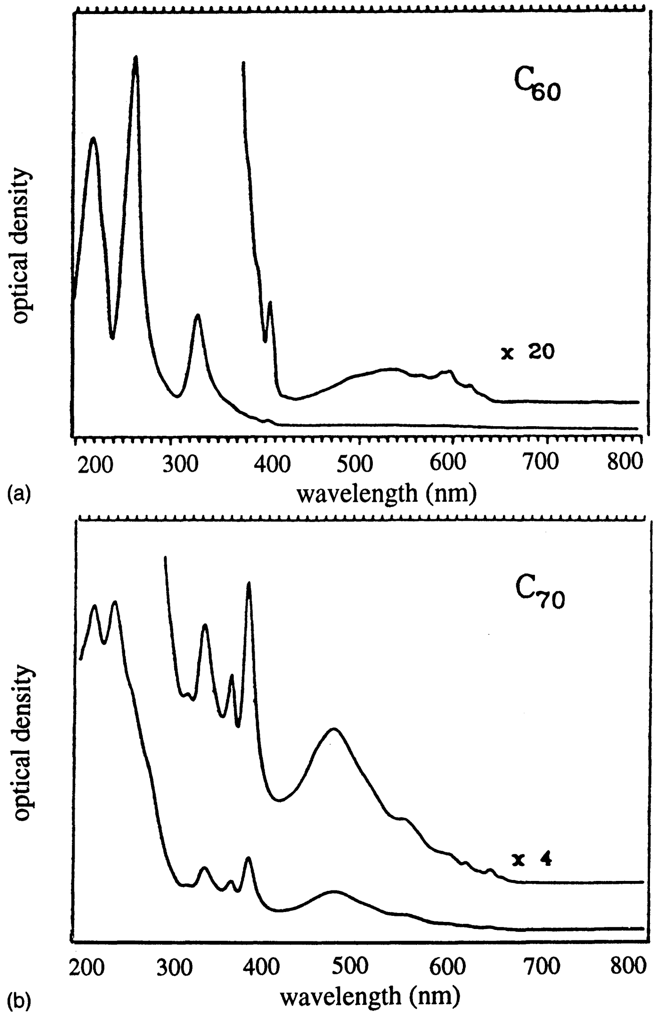

FIG. 8. UV-visible absorption spectra shown as optical density vs wavelength for (a) C_{60} and (b) C_{70} in hexane solutions. For both of these molecular solution spectra, the data near the absorption edge are shown more clearly by using solutions containing higher concentrations of fullerenes as indicated (from Ajie *et al.*, 1990; Ebbesen *et al.*, 1991).

dipole transitions can likewise be described in terms of either a one-electron picture or a many-electron (exciton) picture. In a one-electron picture, the lowest allowed electric dipole transition occurs by exciting an electron from the filled h_u orbital to the unoccupied t_{1g} level (see Fig. 7), or in terms of a many-electron picture from a singlet S_0 ground state to a singlet exciton state denoted by S_1. In terms of the many-electron molecular description, the weak absorption observed between 490 and 640 nm for C_{60} in solution [Fig. 8(a)] is assigned to transitions between the singlet ground state S_0 and the first excited singlet state S_1 (which, for the one-electron molecular orbital description, corresponds to the t_{1u} orbital, activated by vibronic coupling). From the long-wavelength limit (~ 640 nm) of this weak absorption, the energy of the S_1 state can be estimated to be ~ 1.9 eV (Haddon, 1992). Referring to Fig. 8(a), the broad weak absorption band is identified with the $h_u \rightarrow t_{1u}$ forbidden one-electron transition between multiplets associated with the S_0 and S_1 singlet excitonic states. The strong transition starting at ~ 2.9 eV relates to the dipole-allowed transition between the h_u and the t_{1g} one-electron states (see Fig. 7). An analysis of the C_{70} absorption spectrum (plotted as the optical density vs. wavelength) [see Fig. 8(b)] is more difficult than for C_{60} since molecular-orbital calculations for C_{70} reveal a large number of closely spaced orbitals both above and below the HOMO–LUMO gap $E_{HL} = 1.65$ eV (Saito and Oshiyama, 1991).

Further insight into the electronic structure is provided by pulsed-laser studies of C_{60} and C_{70}, which probe the photodynamics of the optical excitation spectra. The importance of these dynamic studies in C_{60} is to show that photoexcitation in the long-wavelength portion of the UV–visible spectrum leads to the promotion of the system from the singlet S_0 state into a singlet S_1 state, which decays quickly with nearly 100% efficiency via an intersystem crossing to the lowest excited triplet state T_1, which has the same orbital quantum numbers as the S_1 state, and therefore the S_1 and T_1 states are strongly coupled. Once in the lower T_1 triplet excitonic state, a luminescence recombination emission to the ground state may be possible, but has not yet been observed directly. This suggests that nonradiative decay from the $T_1 \rightarrow S_0$ levels is much more efficient in returning the system to the ground state. Luminescence emission at 1.7 eV has, howver, been observed and has been identified with $S_1 \rightarrow S_0$ singlet fluorescence. The luminescence contains vibrational sidebands at 0.18 and 0.09 eV associated with the two vibrational A_g modes of C_{60}. In the presence of oxygen, the lifetime of the triplet state becomes very short, and the decay to the ground state becomes very rapid. Furthermore, once C_{60} reaches the T_1 triplet state, efficient electronic dipole excitation to other excited triplet states can occur with high efficiency, thereby greatly increasing the absorption coefficient and providing a mechanism for achieving efficient optical limiting (see Sec. 11).

To date, most of the optical studies on pristine fullerene solids have been carried out in transmission on thin solid films deposited on various substrates (such as quartz, Si, or KBr). The uv–visible transmission spectra for C_{60} and C_{70} solid films are observed to be remarkably similar to the respective solution spectra, thus providing further evidence for the molecular nature of fullerene solids (Ren *et al.*, 1991). The optical properties of solid C_{60} and C_{70} have been studied over a wide frequency range using the variable-angle spectroscopic ellipsometry (VASE) technique. From the VASE measurements the complex refractive index $\tilde{n}(\omega) = n(\omega) + ik(\omega)$ is determined, which is related to the optical dielectric function $\epsilon(\omega) = \epsilon_1(\omega) + i\epsilon_2(\omega) = \tilde{n}^2(\omega)$. In Fig. 9 we show the results for $\epsilon_1(\omega)$ and $\epsilon_2(\omega)$ obtained from VASE and transmission–Fourier-transform ir studies on thin solid films of C_{60} at $T = 300$ K (Rao *et al.*, 1992). The strong, sharp structure at low energy is identified with infrared-active optic phonons, and at higher energies the structure is due to electronic transitions, as discussed below.

Vibronic coupling of the vibrational molecular modes to the LUMO states provides a mechanism by which optical absorption across the HOMO–LUMO gap can be observed. No absorption threshold was detected down to 0.5 eV in the M_6C_{60} materials, establishing an experimental upper bound for the band gap between the t_{1u}- and t_{1g}-derived one-electron bands.

Photoemission and inverse photoemission experiments have been especially useful (Haddon, 1992; Weaver *et al.*, 1991) in providing information on the density of states within a

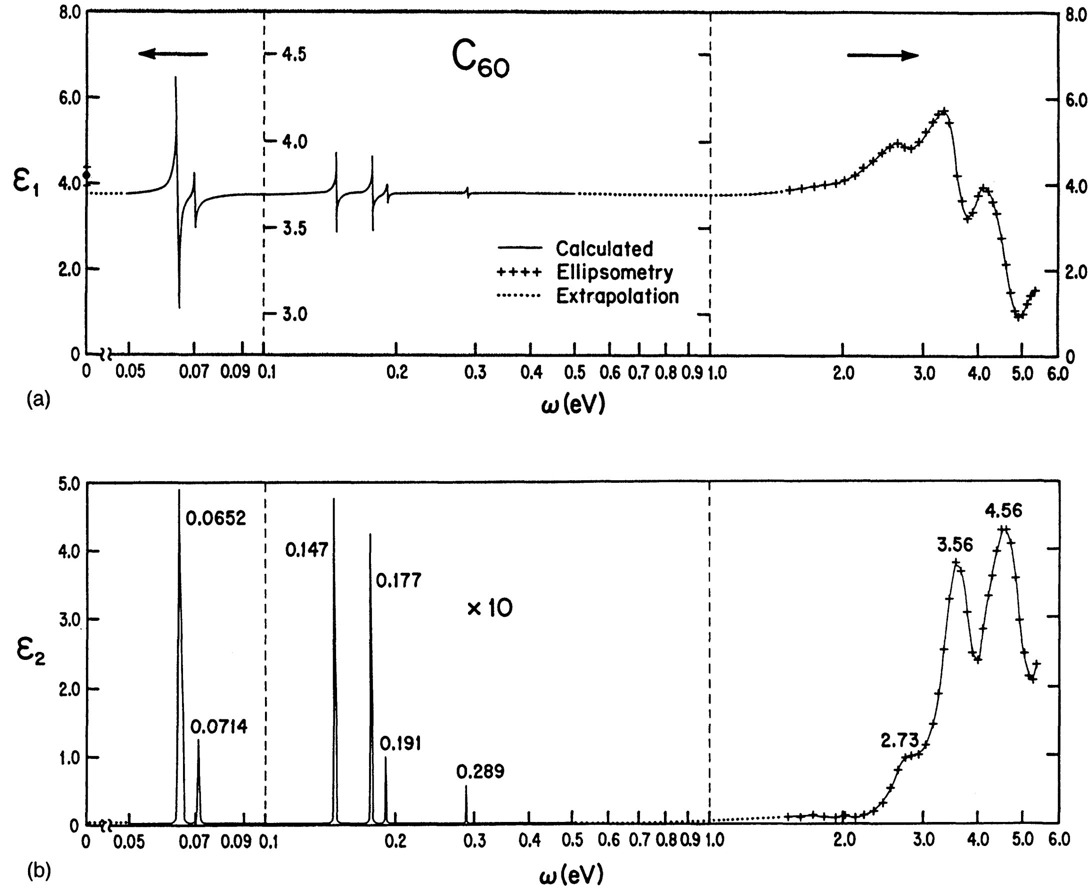

FIG. 9. (a) Real and (b) imaginary parts of the dielectric function $\epsilon(\omega)$ determined from variable-angle spectroscopic ellipsometry (VASE) and Fourier-transform infrared (FTIR) measurements for C_{60} (from Rao *et al.*, 1992).

few electron volts of the Fermi level for both undoped and doped fullerenes. Typical photoemission spectra ($E<E_F$, where E_F is the Fermi energy) and inverse-photoemission spectra ($E>E_F$) are shown in Fig. 10 for C_{60} and $K_xC_{60}(0\leqslant x\leqslant 6)$ (Weaver, 1992), where the intensity maxima correspond to peaks in the density of states, which is here compared to calculations based on a one-electron picture and a constant-optical-matrix element approximation. Photoemission and inverse-photoemission spectra (PES and IPES) provide convincing evidence for charge transfer and for band filling as x in K_xC_{60} increases. Referring to Fig. 10, we can see the peak in the density of states, associated with the t_{1u}-derived band just above the Fermi level E_F in the C_{60} trace, move closer to E_F as K is added and some K_3C_{60} phase is introduced. Upon further addition of potassium, the t_{1u}-related peak crosses the Fermi level, and eventually, for $5.8\leqslant x\leqslant 6.0$, the t_{1u}-derived peak falls below E_F, indicating complete filling of the t_{1u}-derived level; the data for this composition further show a small band gap to the next higher-lying t_{1g}-derived level. Photoemission and inverse-photoemission (Jost *et al.*, 1991) studies have confirmed strong similarities between electronic structure of K_xC_{60} and both Rb_xC_{60} and Cs_xC_{60}. However, differences have also been clearly demonstrated between the density of states for K_xC_{60} and for both Na_xC_{60} and Ca_xC_{60}, for which multiple metal ions can be accommodated in the octahedral sites (see Sec. 4).

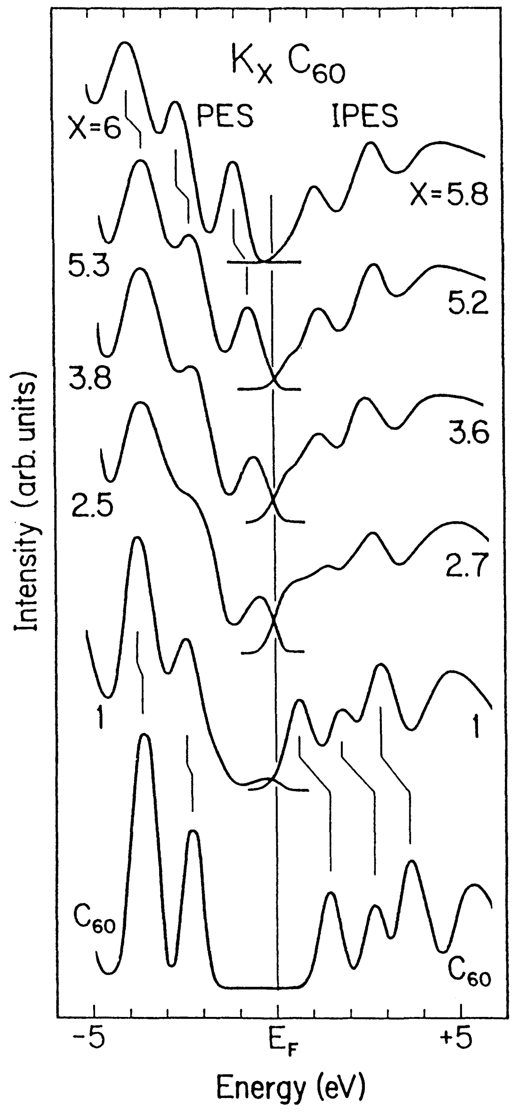

FIG. 10. Photoemission (PES) and inverse photoemission (IPES) (from Weaver, 1992). The spectra are given for K_xC_{60} and show the effects of K incorporation in C_{60} films. The $x=1$ spectra show emission at E_F from grains of K_3C_{60}. Adding more K results in an increase of emission below E_F and a decrease of intensity above E_F. The top spectra show that E_F shifts into the gap when the LUMO bands are filled.

8. ELECTRICAL AND THERMAL TRANSPORT

Because of the high resistivity of undoped C_{60}, the doping of C_{60} with alkali metals creates carriers at the Fermi level and decreases the electrical resistivity ρ of C_{60} by several orders of magnitude. As x in M_xC_{60} increases, the resistivity ρ approaches a minimum at $x=3.0\pm0.05$ (Haddon, 1992), corresponding to a half-filled t_{1u}-derived conduction band. Then upon further increase in x from 3 to 6, ρ again increases, as is shown in Fig. 11 (Kochanski *et al.*, 1992). It should be noted that stable M_xC_{60} compounds (M =K,Rb,Cs) only occur for $x=0$, 3, 4, and 6. (For Na_xC_{60}, stable compounds have been reported for $x=2$ and $x=6$.) Since $x=4$ for M_xC_{60} (M=K,Rb,Cs) is less stable than the others, samples with $x\neq3$ often phase separate into the more stable compounds with $x=0$, 3, and 6. The compounds corresponding to filled molecular levels (C_{60} and M_6C_{60}) are the most stable structurally and correspond to maxima in the resistivity. Furthermore, even at the minimum resistivity in M_xC_{60}, a high value of ρ is found for K_3C_{60} ($\rho_{\min}\cong2.5\times10^{-3}$ Ω cm), typical of a high-resistivity metal (Haddon, 1992). Furthermore, no structure is observed (Fig. 11) in the $\rho(x)$ data near $x=4$, indicating that K_4C_{60} may not be a metallic phase.

Studies of the temperature dependence of the resistivity of polycrystalline M_xC_{60} samples in the normal state show that conduction is by a thermally activated hopping process except for a small range of x near 3 where the conduction is metallic (Haddon, 1992). The magnitude of the arrows in Fig. 11, corresponding to the magnitude of the activation energy E_a, indicate that E_a increases as x deviates further and further from the resistivity minimum at $x=3$, and E_a has, for example, a magnitude of 0.12 eV for $x=1$ (Kochanski *et al.*, 1992). In the metallic regime, $\rho(T)$ results for a superconducting single-crystal K_3C_{60} sample (see Fig. 12) show a linear T dependence of $\rho(T)$ above T_c (Maruyama *et al.*, 1991; Xiang *et al.*, 1992), although the linear T dependence is much weaker than that for the high-T_c cuprates (Ginsberg, 1990). In contrast, many *thin-film* superconducting M_3C_{60} samples have been shown to have a negative slope in $\rho(T)$ vs T for $T\geqslant T_c$. The observed temperature dependence of the resistivity for films of M_xC_{60} (for $0<x\leqslant6$) has been interpreted in terms of a granular conductor with grain sizes in the 60–80-Å range

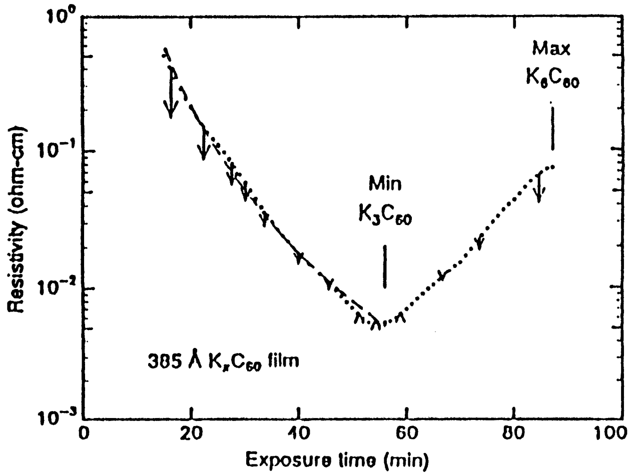

FIG. 11. Dependence of the resistivity of a K_xC_{60} film on exposure time to a K molecular beam. The experiment is carried out in ultrahigh vacuum at an ambient temperature near 74 °C. The arrows show changes in the resistivity of a similar sample as it was heated from 60 to 134 °C (from Kochanski *et al.*, 1992).

(Palstra *et al.*, 1992). This granularity strongly affects the superconducting properties of the M_3C_{60} films, as well as their transport properties in the normal state.

The small value of the observed Hall coefficient suggests the existence of multiple carrier types in K_3C_{60} including both electrons and holes (Palstra *et al.*, 1992). This interpretation is corroborated by the observed sign change in R_H from negative below 220 K to positive above, and by Fermi-surface calculations (Erwin and Pederson, 1991; Oshiyama *et al.*, 1992).

The high electrical resistivity and the magnitude of the optical band gap of C_{60} can be reduced by the application of high pressure, with decreases in resistivity of about one order of magnitude observed per 10 GPa of pressure (Nuñez-Regueiro *et al.*, 1991). At a pressure of ~20 GPa, a phase transition to a more insulating phase has been reported. As the interball C–C distance decreases upon application of pressure and becomes comparable to the intraball C–C distance, an electronic transition might be expected to occur.

Thermal conductivity measurements from 30 to 300 K on single-crystal C_{60} by Yu *et al.* (1992) show a low value of the thermal conductivity κ, and this small magnitude of κ is attributed to both a high density of defect states and a low Debye temperature in C_{60}. In addition, a discontinuity is observed at 260 K, associated with a first-order structural phase transition (see Sec. 4).

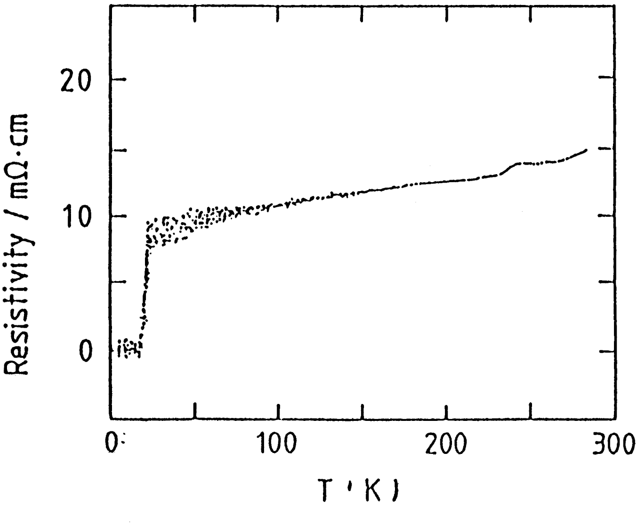

FIG. 12. Temperature dependence of the resistivity ρ of single-crystal K_3C_{60}, showing a linear increase in ρ above the T_c (from Maruyama *et al.*, 1991).

9. SUPERCONDUCTIVITY

Perhaps the most striking property of the C_{60}-related materials is the observation of high-temperature superconductivity. The first observation of superconductivity in an alkali-metal–doped carbon material dates back to 1965 when superconductivity was observed in the first-stage alkali-metal graphite intercalation compound (GIC) C_8K (Hannay *et al.*, 1965). Except for the novelty of observing superconductivity in a compound having no superconducting constituents, this observation did not attract a great deal of attention, since the T_c was very low ($\sim$140 mK). Later, higher T_c's were observed in other GICs [e.g., $KHgC_8$, $T_c = 1.9$ K (Dresselhaus and Dresselhaus, 1981)].

The early observation of superconductivity at 18 K in K_3C_{60} (Haddon *et al.*, 1991) was soon followed by observations of superconductivity at even higher temperatures; in Rb_3C_{60} ($T_c = 30$ K) and $Rb_xCs_yC_{60}$ ($T_c = 33$ K) (Tanigaki *et al.*, 1992). A large increase in T_c is achieved on going to compounds with larger intercalate atoms, resulting in unit cells of larger size and with larger lattice constants as seen in Tables 2 and 4. As the lattice constant increases, the coupling between adjacent C_{60} molecules decreases, narrowing the width of the LUMO level, and thereby increasing the corresponding density of states. Several experiments and calculations (Oshiyama *et al.*, 1992) provide supporting evidence for an increased density of states at the Fermi level resulting from an increase in lattice constant. Figure 13 shows an empirical, nearly linear relation between T_c and the lattice constant a_0 for superconducting alkali-metal–doped samples with a_0 greater than that for undoped C_{60} (see Table 2) (Tanigaki *et al.*, 1992). This correlation includes compounds derived from alkali-metal dopants, alloys of different alkali metals (Fleming *et al.*, 1991a), and samples under pressure (Sparn *et al.*, 1992). Because of the connection between the electronic density of states at the Fermi level $N(E_F)$ and the lattice constant, similar plots of T_c vs $N(E_F)$ have been made (Oshiyama *et al.*, 1992). The increase in T_c is consistent with the BCS expression relating the transition temperature to the density of states $N(E_F)$ and to the electron–phonon coupling V,

$$T_c \sim \omega_{ph} \exp[-1/VN(E_F)]. \qquad (4)$$

However, for the donor C_{60} compounds intercalated with Li, Na, and Ca where a_0 falls below that for C_{60} itself (see Table 2), the linear correlation between T_c and a_0 breaks down.

The reason why T_c is so much higher in the M_3C_{60} materials relative to other carbon-based superconductors such as C_8K appears to be closely related to the higher density of states expected in a molecular solid with little dispersion in the electronic energy-band diagram. This is achieved in the M_3C_{60} compounds at the Fermi level when the T_{1u} LUMO molecular level is half filled with carriers. In addition, several authors have found some

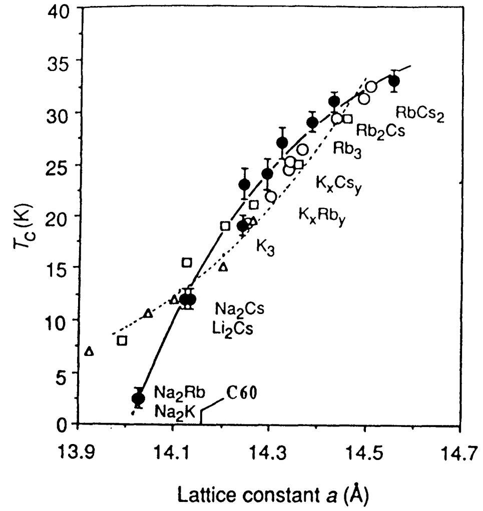

FIG. 13. Dependence of T_c for various M_3C_{60} compounds on the lattice constant (from Tanigaki *et al.*, 1992). The open triangles and squares relate to pressure-dependent studies. The solid line is a fit to the experimental data and the dashed curve is predicted from density-of-states calculations.

Table 4. Superconductivity parameters in M_3C_{60} (M=K,Rb).

Parameter	K_3C_{60}	Rb_3C_{60}
fcc a_0 (Å)	14.253	14.436
T_c (K)	19.3	29.6
$2\Delta(0)/kT_c$	5.2	5.3
$(dT_c/dP)_{p=0}$ (K/GPa)	−7.8	−9.7
$H_{c1}(0)$ (mT)	13	12
$H_{c2}(0)$ (T)	49	78
$H_c(0)$ (T)	0.38	0.44
ξ_0 (nm)	2.6±0.3	2.0±0.25
λ_L (nm)	240±15	247±15
$\kappa_s=\lambda_L/\xi_0$	92	124

enhancement in the electron–phonon coupling in C_{60}-related materials. If the electron–phonon interaction provides the pairing mechanism for superconductivity in C_{60}-based materials, it is believed that it is the H_g-derived phonons that play the dominant role in the coupling. The observation of broad H_g-derived Raman lines in M_3C_{60} is consistent with a strong electron–phonon coupling. Calculations of the contributions of the various phonon modes to the electron–phonon interaction strength (Schlüter *et al.*, 1992) show that the high-frequency intramolecular phonons can account for the high T_c values that are observed experimentally.

Using scanning tunneling microscopy (STM), the temperature dependence of the superconducting band gap has been investigated for both K_3C_{60} and Rb_3C_{60} (Zhang *et al.*, 1991). The results show a good fit for the temperature dependence of the superconducting band gap 2Δ by BCS theory (see Fig. 14), although the zero-temperature band-gap value $[(2\Delta(0)/kT_c)]$ was found to be 5.3 for K_3C_{60} and 5.2 for Rb_3C_{60}, well above 3.5 for the case of a BCS superconductor. These results imply that the M_3C_{60} compounds are strong-coupling superconductors (Zhang *et al.*, 1991), indicating that the paired electrons are more strongly attracted to each other.

There have been several measurements of the isotope effect in superconducting M_3C_{60} to provide insight into whether or not the electron pairing mechanism responsible for superconductivity in the doped fullerenes involves the coupling between electrons and phonons as in BCS superconductors. In the experiments reported to date, the isotopic enrichment has involved ^{13}C. The experiments all suggest that $T_c \propto M^{-\alpha}$, consistent with phonon-mediated electron pairing, where the observed α values are compared with the BCS prediction of $\alpha = 0.5$. The experimental values of α for superconducting fullerenes at present differ from one group to another, though all groups suggest that intramolecular C_{60} vibrations are involved in the pairing mechanism for superconductivity and that the electron–phonon coupling constant is quite large.

Of interest also is the dependence of the superconducting parameters on pressure and magnetic field. Closely related to the high compressibility of C_{60} and M_3C_{60} (M=K,Rb) is the large linear decrease in T_c with pressure observed in K_3C_{60} and Rb_3C_{60} (see Table 4) (Sparn *et al.*, 1992). By attributing the smaller size of the K^+ ion relative to Rb^+ (0.186 Å) to an "effective chemical pressure" of 1.06 GPa, the pressure dependence of T_c for both K_3C_{60}

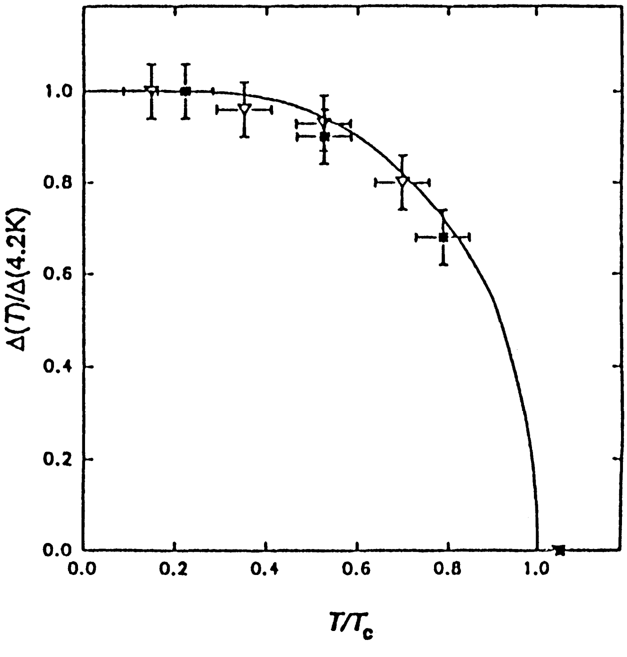

FIG. 14. Summary of the temperature-dependent superconducting energy gap results for K_3C_{60} (dark squares) and Rb_3C_{60} (open triangles). For comparison the values of the gap 2Δ have been normalized by the respective low-temperature (4.2 K) values and the temperature has been scaled by T_c. The solid line corresponds to the temperature dependence of $\Delta(T)/\Delta(0)$ calculated by the BCS theory (from Zhang *et al.*, 1991).

and Rb_3C_{60} can be explained. Listed in Table 4 are values for the characteristic superconducting parameters relevant to their magnetic field behavior, including estimates of the upper critical field (in tesla) at $T=0$, $H_{c2}(0)$, the lower critical field $H_{c1}(0)$, the coherence length ζ_0 at $T=0$, and the London penetration depth ξ_L. The large values of $\kappa_s=\xi_L/\zeta_0$ listed in Table 4 confirm that the M_3C_{60} compounds are strongly type-II superconductors.

10. C_{60}-RELATED TUBULES

In addition to ball-like fullerene molecules, it is possible to synthesize tubular fullerenes with very high aspect ratios (length/diameter). In analogy to the C_{60} ball, we can define C_{60}-derived tubules by bisecting a C_{60} molecule at the equator and joining the two resulting hemispheres with a cylindrical tube of carbon atoms in a honeycomb lattice one monolayer thick, and with the same diameter as C_{60}. If the C_{60} is bisected normal to a fivefold axis, the armchair tubule shown in Fig. 15(a) is formed, and if the C_{60} is bisected normal to a threefold axis, the zigzag tubule in Fig. 15(b) is formed (Dresselhaus *et al.*, 1992b).

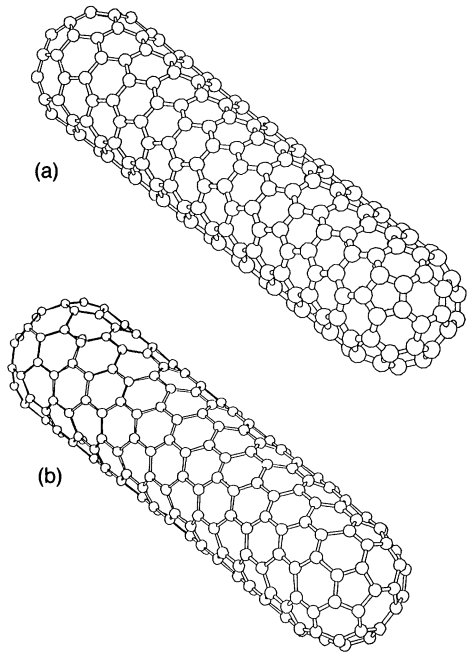

FIG. 15. By rolling up a graphene sheet (a single layer from a 3D graphite crystal) as a cylinder and capping each end of the cylinder with half of a C_{60} molecule, a "C_{60}-related tubule" one layer in thickness is formed. Shown here is a schematic model for a tubule arising from (a) an armchair cap with a fivefold axis, and (b) a zigzag cap with a threefold axis (from Dresselhaus *et al.*, 1992b).

Two reports have been made on the observation of carbon fibers of very small (nanometer) diameters (Iijima, 1991; Endo *et al.*, 1991, 1992) based on high-resolution transmission electron microscopy (TEM) measurements (see Fig. 16). These images suggest the existence of micron-long tubules that have cross sections consisting of several concentric coaxial tubes and a hollow core [see Fig. 16(a)]. In Fig. 16(b), we see one fiber with only two concentric tubules, and another in Fig. 16(c) with an inner diameter of only 23 Å, about three times larger than the diameter of C_{60} (Iijima, 1991). Such carbon tubules have been nucleated on vapor-grown carbon fibers (Endo *et al.*, 1991, 1992) or on the surface of the negative carbon electrode in a dc arc discharge (Iijima, 1991). These fullerene tubules differ in a fundamental way from the scroll-like graphite whiskers synthesized by Bacon many years ago in a dc carbon-arc discharge at much higher pressures of the rare gas than are used to synthesize fullerene tubules (Bacon, 1960). Highly elongated fullerenes with shapes between the traditional icosahedral fullerenes and the graphene tubules have also been observed by high-resolution transmission electron microscopy (Wang and Buseck, 1991).

The growth mechanism for cylindrical fullerene-related tubules is especially interesting. According to one model, the nucleation of a C_{60} tubule instead of a C_{60} ball is believed to be stabilized by a defect in the cap region. In the growth process, the pentagonal faces in the cap region act like catalytic particles to promote tubule growth (Endo and Kroto, 1992). Saito *et al.* (1992a) have suggested a series of steps that incorporate a C_2 unit through the motion of two pentagonal faces on a hemispherical fullerene-derived cap and promote tubule growth.

Iijima (1991) reports that the majority of the carbon tubules that he has observed by high-resolution TEM have screw axes and chirality. Using Fig. 17, we can specify the chirality and the tubule diameter of any graphene tubule in terms of the chiral vector

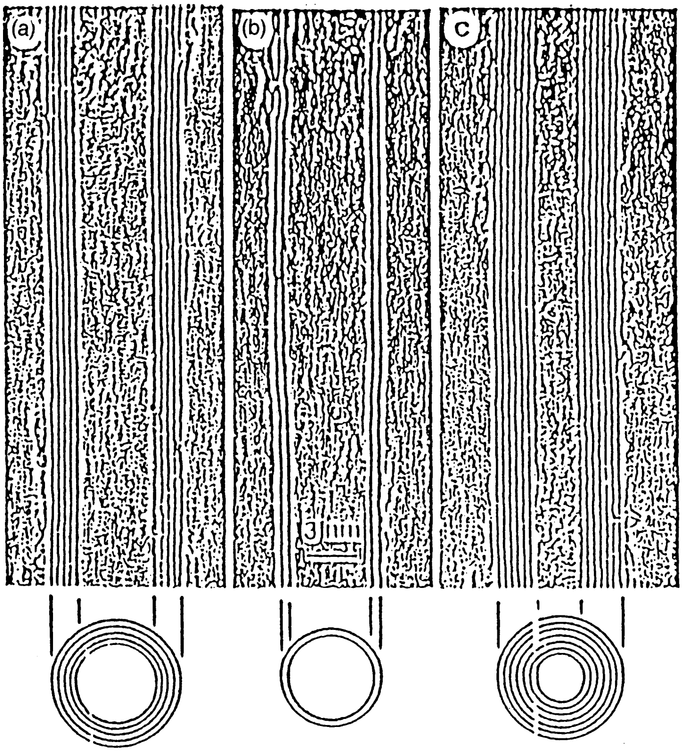

FIG. 16. The observation of N concentric carbon tubules with various inner diameters d_i and outer diameters d_0 was reported using TEM (from Iijima, 1991). (a) $N=5$, $d_0=67$ Å, (b) $N=2$, $d_0=55$ Å, and (c) $N=7$, $d_i=23$ Å, $d_0=65$ Å.

$\mathbf{c}_h=n\hat{\mathbf{a}}_1+m\hat{\mathbf{a}}_2$ that connects two crystallographically equivalent sites on a two-dimensional (2D) graphene sheet [Fig. 17(a)] (Dresselhaus *et al.*, 1992b), where a graphene sheet denotes a single layer of the 3D graphite lattice. The construction in Fig. 17 shows the chiral angle θ with respect to the zigzag direction ($\theta=0°$) and the unit vectors $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$ of the hexagonal honeycomb lattice. The armchair tubule [Fig. 15(a)] corresponds to $\theta=30°$ on this construction. An ensemble of possible vectors specified by pairs of integers (n,m) denoting the vector $\mathbf{c}_h=n\hat{\mathbf{a}}_1+m\hat{\mathbf{a}}_2$ is given by Fig. 17(b) (Saito *et al.*, 1992b).

The cylinder connecting the two hemispherical caps is formed by superimposing the two ends of the vector $\mathbf{c}_h$, and the cylinder joint is made along the two lightly dotted lines in Fig. 17, which are perpendicular to the vector $\mathbf{c}_h$ at each end (Dresselhaus *et al.*, 1992b). The chiral tube thus generated has no distortion of bond angles other than distortions caused by the cylindrical curvature of the tubule. Differences in chiral angle θ and in the tubule diameter d give rise to differences in the properties of the various graphene tubules, as discussed below. In the (n,m) notation for $\mathbf{c}_h=n\hat{\mathbf{a}}_1+m\hat{\mathbf{a}}_2$, the vectors $(n,0)$ denote zigzag tubules and the vectors (n,n) denote armchair tubules. All other vectors (n,m) correspond to chiral tubules where the tubule axis has the symmetry properties of a screw axis (Saito *et al.*, 1992b). Of course, both right- and left-handed chiralities are possible for chiral tubules. Therefore, it is expected that chiral tubules are optically active to either right- or left-circularly polarized light, depending on their chirality. In terms of the integers (n,m), the tubule diameter d is given by

$$d=\sqrt{3}a_{\text{C-C}}(m^2+mn+n^2)^{1/2}/\pi, \tag{5}$$

where $a_{\text{C-C}}$ is the nearest-neighbor carbon–carbon distance and the chiral angle θ is given by

$$\theta=\tan^{-1}[\sqrt{3}n/(2m+n)]. \tag{6}$$

For example, a zigzag tubule ($\theta=0°$) with $(m,0)=(9,0)$ has a tubule diameter of 7.05 Å, while an armchair tubule (5,5) has $d=6.83$ Å, which exactly matches the diameter of 6.83 Å for a C_{60} ball.

Because of the special atomic arrangement of the carbon atoms in a C_{60} tubule, substitutional impurities other than boron are inhibited by the small size of the carbon atoms.

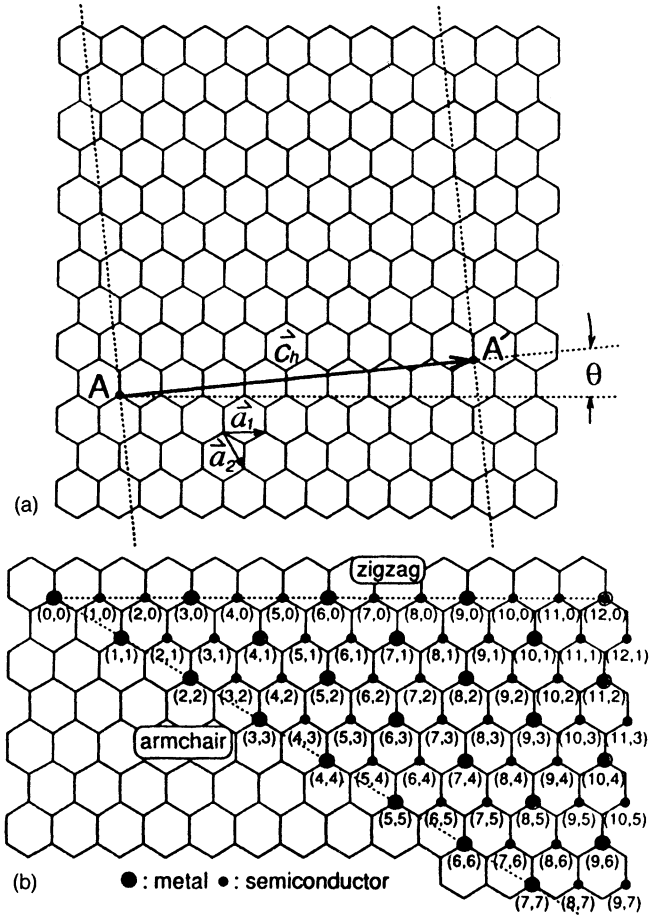

FIG. 17. (a) The vector $\overrightarrow{AA'}$ or $\mathbf{c}_h = n\hat{\mathbf{a}}_1 + m\hat{\mathbf{a}}_2$ on the honeycomb lattice defined by unit vectors $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$ and the chiral angle θ with respect to the zigzag axis. (b) Possible vectors for general tubules, including zigzag, armchair, and chiral tubules. The large dots and small dots respectively denote metallic and semiconducting behavior for each of the possible tubules (from Dresselhaus *et al.*, 1992b).

Furthermore, the screw-axis dislocation, the most common defect found in bulk graphite, is inhibited by the monolayer structure of the C_{60} tubule. The special geometry of the C_{60}-related tubules should make these tubules relatively incompressible along the tubule axis, when compared to other materials (Overney *et al.*, 1992).

From a theoretical standpoint, fullerene tubules are interesting as examples of a one-dimensional periodic structure along the tubule axis. Geometrical confinement in the radial direction is provided by the monolayer thickness of the tubule. In the circumferential direction, periodic boundary conditions apply to the electronic and vibrational states associated with the enlarged unit cell that is formed in coordinate space and the subsequent zone folding that occurs for these states in reciprocal space. We can then expect to observe one-dimensional (1D) dispersion relations for electrons and phonons in C_{60}-derived tubules.

A number of methods have been used to calculate the 1D electronic bands for fullerene tubules, and these energy bands are closely related to the 2D graphene sheet used to form the tubule. Their results for the 1D electronic energy bands show that for small-diameter graphene tubules, about $\frac{1}{3}$ of the tubules are metallic and $\frac{2}{3}$ are semiconducting, depending on the tubule diameter d and chiral angle θ. Metallic conduction in a fullerene tubule is achieved when (Saito *et al.*, 1992c)

$$2n+m=3q, \tag{7}$$

where n and m are integers specifying the tubule diameter and chiral angle and q is an integer. Tubules satisfying Eq. (7) are indicated in Fig. 17(b) as large dots, and these are the metallic tubules. The small dots in this figure correspond to semiconducting tubules. Figure 17(b) shows that all armchair tubules are metallic, but only $\frac{1}{3}$ of the possible zigzag tubules are metallic (Saito *et al.*, 1992b). This effect is illustrated in the density of states for metallic and semiconducting tubules, shown in Fig. 18 for the (9,0) and (10,0) tubules, respectively (Saito *et al.*, 1992b). Here we see a finite density of states at the Fermi level for the (9,0) tubules, and a vanishing density of states for the (10,0) tubules, consistent with the metallic nature of the (9,0) tubules and the semiconducting nature of the (10,0) zigzag tubules. It is surprising that the calculated electronic structure can be either metallic or semiconducting, depending on the choice of (n,m), although there is no difference in the local chemical bonding between the carbon atoms in the tubules, and no doping impurities are present (Saito *et al.*, 1992b). These surprising results can be understood on the basis of the electronic structure of 2D graphite, which is a zero-gap semiconductor (Painter and Ellis, 1970) with bonding and antibonding π bands degenerate at the K point (zone corner) of the hexagonal Brillouin zone. The periodic boundary conditions permit only a few wave vectors to exist in the circumferential direction. If one of these wave vectors along the tubule axis direction passes through the K point in the graphite Brillouin zone (satisfying Eq. 7), then metallic conduction results.

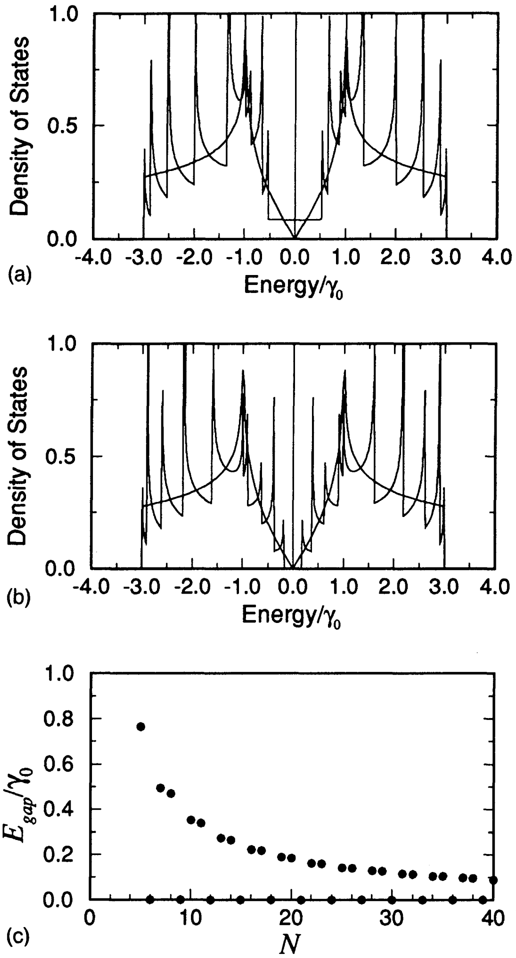

FIG. 18. Electronic 1D density of states for two (n,0) zigzag tubules: (a) the (10,0) tubule, which has semiconducting behavior, and (b) the (9,0) tubule, which has metallic behavior (from Saito *et al.*, 1992b). The zero of energy is chosen to coincide with the Fermi level. (c) Plot of the semiconducting gap E_{gap} normalized to the nearest-neighbor C–C overlap energy γ_0 for graphite vs the number of carbon atoms in the zigzag ring.

Metallic 1D energy bands are generally unstable under a Peierls distortion. However, the Peierls energy gap obtained for the metallic tubules is found to be greatly suppressed by increasing the tubule diameter, so that the Peierls gap quickly approaches the zero-energy gap of 2D graphite. Thus if we consider finite temperatures or fluctuation effects, such a small Peierls gap can be neglected. It is also of interest to note that as the tubule diameter increases, more wave vectors become allowed for the circumferential direction, the tubules become more two dimensional, and the semiconducting band gap disappears, as is illustrated in Fig. 18(c), which shows the diameter dependence of the semiconducting band gap for graphene tubules. Calculation of the electronic structure for two concentric tubules shows that pairs of concentric metal–semiconductor or semiconductor–metal tubules are stable.

From these results one could imagine designing a minimum-size conductive wire consisting of two concentric graphene tubules with a metallic inner tubule covered by a semiconducting (or insulating) outer tubule. This concept could be extended to the design of tubular metal–semiconductor devices without introducing any doping impurities (Saito *et al.*, 1992b). There are many possibilities for arranging arrays of metallic and semiconducting graphene tubules from a conceptual standpoint. It will be interesting to see what kinds of tubules and tubule arrays can be synthesized from a practical standpoint.

The elastic properties of fullerene tubules have also been considered by carrying out lattice-mode calculations of long fullerene tubules based on 30-carbon-atom hemispherical caps from C_{60} and consisting of 100, 200, and 400 carbon atoms (Overney *et al.*, 1992). These calculations show that the lowest bending vibrational mode frequency decreases as the length of the fullerene tube increases, saturating at a cluster size of about 200 carbon atoms. The theoretical magnitude of the beam rigidity of these tubules is found to exceed that of presently available materials (Overney *et al.*, 1992), so that graphene tubules are expected to offer outstanding mechanical properties.

11. APPLICATIONS

Because graphite is a material of extreme properties (modulus, strength, thermal conductivity, anisotropy, melting point, etc.), and because of the similarities between the bonding in graphite and C_{60}, we might also expect fullerenes to exhibit extreme properties. As in graphite where the planes are weakly van der Waals bonded, solid C_{60} has weak intermolecular bonding and strong intramolecular bonds. Although research on solid C_{60} and related materials is still at an early stage, these materials are already beginning to show many exceptional properties, some of which may lead to practical applications.

One possible application for C_{60} is as an optical limiter. Optical limiters are used to protect materials from damage by high light intensities via a rapid saturation of the transmitted light intensity with increasing incident intensity. Outstanding performance for C_{60} relative to presently used optical limiting materials has been observed at an optical wavelength of 5320 Å for 8-ns pulses using solutions of C_{60} in toluene and in CH_3Cl (Tutt and Kost, 1992). Although C_{70} in similar solutions also showed optical limiting action, the performance of C_{60} was superior. The proposed mechanism for the optical limiting is that C_{60} and C_{70} are more absorptive once the triplet excited state is populated. The initial absorption of a photon leads to electronic excitation to an excitonic state, with a subsequent intersystem crossing to a triplet state, followed by further stronger absorption of photons as excitations from the triplet state to higher-lying triplet states take place (see Sec. 7) (Tutt and Kost, 1992).

The possible use of C_{60} for the fabrication of industrial diamonds offers another area for possible applications. It is found that when rapid nonhydrostatic pressures (in the range of 20 GPa) are applied at room temperature to C_{60}, the material is transformed instantaneously into bulk polycrystalline diamond at high efficiency (Nuñez-Regueiro *et al.*, 1992). It is believed that the presence of pentagons in the C_{60} structure promotes the formation of sp^3 bonds during the application of high anisotropic stress. Enhanced nucleation of a high density of diamond crystallites on a Si substrate has been achieved by Meilunas *et al.* (1991) through the initial deposition of a $\sim$1000-Å C_{70} film, followed by positive-ion bombardment, and then by chemical vapor deposition growth of diamond in a microwave plasma reactor. A base layer of ion-activated C_{70} was found to be more effective in promoting sp^3 bonding (Meilunas *et al.*, 1991) than a similarly treated C_{60} buffer layer. Normally, for diamond films to grow from a mixture of gaseous CH_4 and H_2, the substrate surface (usually Si) must be pretreated with diamond grit polish or must contain small diamond seeds.

A photoconducting device can be fabricated using fullerene-doped polymers. Polyvinylcarbazole (PVK) doped with a mixture of C_{60} and C_{70} has been reported to exhibit exceptionally good photoconductive properties (Wang, 1992), which may lead to the development of future polymeric photoconductive materials. The effects of the fullerenes ($\sim$2.7% by weight) on the charge generation and transport still need to be understood.

Ever since C_{60} was discovered (Kroto *et al.*, 1985), the possibility of using fluorinated derivatives such as $C_{60}F_{60}$ as lubricants was considered. It was recently shown that highly fluorinated C_{60} is unstable in the presence of trace amounts of H_2O (Taylor *et al.*, 1992) and therefore is probably unsuitable for application as a lubricant.

As further research on these materials is carried out, it is expected, because of the extreme properties of carbon-based materials generally, that other interesting physics and chemistry will be discovered, and promising applications will be found for fullerenes and fullerene-related materials.

ACKNOWLEDGMENTS

The authors acknowledge fruitful discussions with Professor M. Endo, Professor R. Saito, Professor M. Fujita, and Professor R. A. Jishi. The authors gratefully acknowledge support from NSF Grant No. DMR-88-19896 and from AFOSR Grant No. 90-0123 (M. S. D. and G. D.) and from the University of Kentuckey Center for Applied Energy Research (P. E.).

GLOSSARY

Antibonding States: Electronic states for which the electron charge distribution is decreased through interactions with neighboring atoms.

Armchair Tubule: Graphene tubule with orientation such that a cut normal to the tubule axis gives an up-up-down-down edge for the carbon-atom locations.

Bonding States: Electronic states where the energy is lowered by the excess charge distribution coupling to adjacent atoms.

Dopant: Foreign atom or molecule intentionally added to a host material.

Doping: The addition of foreign atoms or molecules to the host material.

Endohedral Doping: Placement of a foreign atom inside the fullerene cage.

Exohedral Doping: Placement of a foreign atom outside the fullerene cage.

Fullerene: An abbreviation of "buckminsterfullerene," the name given to a closed-cage polyhedral molecule of carbon atoms.

Functional Groups: A chemical group serving a particular purpose in the synthesis or in the end product of a reaction.

Glassy Carbon: A highly disordered, weakly conducting form of carbon containing graphitic ribbons, having high porosity and good mechanical properties.

Graphene Sheet: An isolated single layer of hexagonally arranged carbon atoms such as that found in graphite.

HOMO: Highest occupied molecular orbital.

Hybridization: The coupling and admixing of two electronic orbitals.

Icosahedron: A regular polyhedron having 20 faces and 12 vertices.

Infrared-Active Modes: Vibrational modes that are excited by infrared radiation.

Intercalation: The insertion of guest species in layers or in specially designated sites of the host material.

Kekulé Structure: A chemical structure where carbon atoms in benzene or analogous rings show alternating single and double bonds.

LUMO: Lowest unoccupied molecular orbital.

Octahedral Site: A lattice site with nearest neighbors at the centers of the faces of a cube.

Peierls Distortion: A lattice distortion resulting in a lowering of the total energy and of the symmetry.

Peierls Gap: A gap in the electronic density of states found at the Fermi surface as a result of the Peierls distortion.

π Electron: An electron whose wave function extends normal to the plane where the nuclei are located.

Pristine Fullerene: An undoped fullerene.

Raman-Active Modes: Vibrational modes that are symmetry allowed to participate in the inelastic (Raman) scattering of light.

Regular Truncated Icosahedron: A polyhedron formed by truncating each vertex of an icosahedron so that all of the 90 edges are of equal lengths.

Screw Axis: Symmetry axis that requires both a rotation and a translation.

Soot: A very fine (submicron) powder of carbon or carbonaceous material.

Soxhlet Extraction: A method for continuous extraction of a solute from a solid in a solvent near its boiling point.

sp^2 Trigonal Bonding: Three planar bonds derived from one s and two p atomic carbon orbitals making angles of 120° with each other.

sp^3 Tetrahedral Bonding: Four bonds derived from one s and three p atomic carbon orbitals making tetrahedral angles.

Tetrahedral Site: A lattice site with near neighbors located at the corners of a regular tetrahedron.

Two-Zone Furnace: An oven with two different constant-temperature zones. Vapor transport of the reactant (colder zone) and subsequent intercalation of the reactant into the host crystal or film takes place in the hotter zone.

van der Waals Bond: Weak dipolar bond between neutral atoms or molecules.

van der Waals Charge Cloud: An electron charge distribution that weakly binds near-neighbor atoms or molecules.

Vibronic Coupling: Coupling between electrons and vibrations of a molecule that turns on a symmetry-forbidden optical transition between two molecular states.

Zigzag Tubule: Graphene tubule with orientation such that a cut normal to the tubule axis gives a zigzag edge (up-down) of carbon atoms.

List of Works Cited

Ajie, H., Alvarez, M. M., Anz, S. J., Beck, R. D., Diederich, F., Fostiropoulos, K., Huffman, D. R., Krätschmer, W., Rubin, Y., Schriver, K. E., Sensharma, D., Whetten, R. L. (1990), *J. Phys. Chem.* **94**, 8630–8633.

Allemand, P.-M., Khemani, K. C., Koch, A., Wudl, F., Holczer, K., Donovan, S., Grüner, G., Thompson, J. D. (1991), *Science* **253**, 301–303.

Bacon, R. (1960), *J. Appl. Phys.* **31**, 283–290.

Birkett, P. R., Hitchcock, P. B., Kroto, H. W., Taylor, R., Walton, D. R. M. (1992), *Nature* **357**, 479–481.

Bochvar, D. A., Gal'pern, E. G. (1973), *Dokl. Akad. Nauk SSSR* **209**, 610–615 (*Proc. Acad. Sci. USSR* **209**, 239–244).

Chen, C.-C., Kelty, S. P., Lieber, C. M. (1991), *Science* **253**, 886–888.

Chen, Y., Stepniak, F., Weaver, J. H, Chibante, L. P. F., Smalley, R. E. (1992), *Phys. Rev. B* **45**, 8845–8848.

Diederich, F., Whetten, R. L. (1992), *Accts. Chem. Res.* **25**, 119–126.

Diederich, F., Ettl, R., Rubin, Y., Whetten, R. L., Beck, R., Alvarez, M., Anz, S., Sensharma, D., Wudl, F., Khemani, K. C., Koch, A. (1991), *Science* **252**, 548–551.

Dresselhaus, M. S., Dresselhaus, G. (1981), *Adv. Phys.* **30**, 139–326.

Dresselhaus, G., Dresselhaus, M. S., Eklund, P. C. (1992a), *Phys. Rev. B* **45**, 6923–6930.

Dresselhaus, M. S., Dresselhaus, G., Saito, R. (1992b), *Phys. Rev. B* **45**, 6234–6242.

Ebbesen, T. W., Tanigaki, K., Kuroshima, S. (1991), *Chem. Phys. Lett.* **181**, 501–504.

Eklund, P. C., Zhou, Ping, Wang, Kai-An, Dresselhaus, G., Dresselhaus, M. S. (1992), *J. Phys. Chem. Solids* **53**, 1391–1413.

Endo, M., Kroto, H. W. (1992), *J. Phys. Chem.* **96**, 6941–6944.

Endo, M., Fujiwara, H., Fukunaga, E. (1991), *Meeting of the Japanese Carbon Society* (unpublished), pp. 34–35.

Endo, M., Fujiwara, H., Fukunaga, E. (1992), *Second C_{60} Symposium in Japan* (unpublished), pp. 101–104.

Erwin, S. C., Pederson, M. R. (1991), *Phys. Rev. Lett.* **67**, 1610–1613.

Fagan, P. J., Calabrese, J., Malone, B. (1992), *Accts. Chem. Res.* **25**, 134–142.

Field, J. E. (1979), in: J. E. Field (Ed.), *Properties of Diamond*, London: Academic, p. 281.

Fischer, J. E., Heiney, P., Smith, Amos B., III (1992), *Accts. Chem. Res.* **25**, 112–118.

Fleming, R. M., Rosseinsky, M. J., Ramirez, A. P., Murphy, D. W., Tully, J. C., Haddon, R. C., Siegrist, T., Tycko, R., Glarum, S. H., Marsh, P., Dabbagh, G., Zahurak, S. M., Makhija, A. V., Hampton, C. (1991a), *Nature* **352**, 701–703.

Fleming, R. M., Ramirez, A. P., Rosseinsky, M. J., Murphy, D. W., Haddon, R. C., Zahurak, S. M., Makhija, A. V. (1991b), *Nature* **352**, 787–788.

Fuller, R. Buckminster (1984), in: W. Marlin (Ed.), *The Artifacts of R. Buckminster Fuller: A Comprehensive Collection of His Designs and Drawings*, New York: Garland.

Ginsberg, D. M. (Ed.) (1990), *Physical Properties of High Temperature Superconductors II*, Singapore: World Scientific.

Haddon, R. C. (1992), *Accts. Chem. Res.* **25**, 127–133.

Haddon, R. C., Hebard, A. F., Rosseinsky, M. J., Murphy, D. W., Duclos, S. J., Lyons, K. B., Miller, B., Rosamilia, J. M., Fleming, R. H., Kortan, A. R., Glarum, S. H., Makhija, A. V., Muller, A. J., Eick, R. H., Zahurak, S. M., Tycko, R., Dabbah, G., Thiel, F. A. (1991), *Nature* **350**, 320–322.

Hammond, G. S., Kuck, V. J. (Eds.) (1992), *Fullerenes*, ACS Symposium Series No. 481, Washington, DC: American Chemical Society, Chap. 6, p. 91.

Hannay, N. B., Geballe, T. H., Matthias, B. T., Andres, K., Schmidt, P., MacNair, D. (1965), *Phys. Rev. Lett.* **14**, 225–226.

Hare, J. P., Kroto, H. W. (1992), *Accts. Chem. Res.* **25**, 106–111.

Heiney, P. A., Vaughan, G. B. M., Fischer, J. E., Coustel, N., Cox, D. E., Copley, J. R. D., Neumann, D. A., Kamitakahara, W. A., Creegan, K. M., Cox, D. M., McCauley, J. P., Jr., Smith A. B., III, (1992), *Phys. Rev. B* **45**, 4544–4547.

Herzberg, G. (1945), *Molecular Spectra and Molecular Structure*, Princeton, NJ: Van Nostrand. Icosahedral symmetry is not treated explicitly.

Iijima, S. (1980), *J. Cryst. Growth* **50**, 675–683.

Iijima, S. (1991), *Nature* **354**, 56–58.

Johnson, R. D., Bethune, D. S., Yannoni, C. S. (1992), *Accts. Chem. Res.* **25**, 169–176.

Jost, M. B., Troullier, N., Poirier, D. M., Martins, J. L., Weaver, J. H., Chibante, L. P. F., Smalley, R. E. (1991), *Phys. Rev. B.* **44**, 1966–1969.

Kochanski, G. P., Hebard, A. F., Haddon, R. C., Fiory, A. T. (1992), *Science* **255**, 184–186.

Kortan, A. R., Kopylov, N., Glarum, S., Gyorgy, E. M., Ramirez, A. P., Fleming, R. M., Theil, F. A., Haddon, R. C. (1992), *Nature* **355**, 529–532.

Krätschmer, W., Lamb, L. D., Fostiropoulos, K., Huffman, D. R. (1990), *Nature* **347**, 354–358.

Kroto, H. W., Heath, J. R., O'Brien, S. C., Curl, R. F., Smalley, R. E. (1985), *Nature* **318**, 162–163.

Maruyama, Y., Inabe, T., Ogata, H., Achiba, Y., Suzuki, S., Kikuchi, K., Ikemoto, I. (1991), *Chem. Lett.* 1849.

Meilunas, R., Chang, R. P. H., Lu, S., Kappes, M. M. (1991), *Appl. Phys. Lett.* **59**, 3461–3463.

Meng, R. L., Ramirez, D., Jiang, X., Chow, P. C., Diaz, C., Matsuishi, K., Moss, S. C., Hor, P. H., Chu, C. W. (1992), unpublished.

Nuñez-Regueiro, M., Monceau, P., Rassat, A., Bernier, P., Zahab, A. (1991), *Nature* **354**, 289–291.

Nuñez-Regueiro, M., Monceau, P., Hodeau, J.-L. (1992), *Nature* **355**, 237–239.

Osawa, E. (1970), *Kagaku (Kyoto)* **25**, 854–863, in Japanese.

Oshiyama, A., Saito, S., Hamada, N., Miyamoto, Y. (1992), *J. Phys. Chem. Solids* **53**, 1457–1471.

Overney, G., Zhong, W., Tománek, D. (1992), *Z. Phys. D* (in press).

Painter, G. S., Ellis, D. E. (1970), *Phys. Rev. B* **1**, 4747–4752.

Palstra, T. T. M., Haddon, R. C., Hebard, A. F., Zaanen, J. (1992), *Phys. Rev. Lett.* **68**, 1054–1057.

Rao, A. M., Wang, Y., Ren, S. L., Lehman, G. W., Zhou, P., Eklund, P. C. (1992), *Phys. Rev. B* (in press).

Ren, S. L., Wang, Y., Rao, A. M., McRae, E., Hager, G. T., Wang, K. A., Lee, W. T., Ni, H. F., Selegue, J., Eklund, P. C., (1991) *Appl. Phys. Lett.* **59**, 2678–2680.

Rohlfing, E. A., Cox, D. M., Kaldor, A. (1984), *J. Chem. Phys.* **81**, 3322–3330.

Rosseinsky, M. J., Murphy, D. W., Fleming, R. M., Tycko, R., Ramirez, A. P., Siegrist, T., Dabbagh, G., Barrett, S. E. (1992), *Nature* **356**, 416–418.

Saito, R., Dresselhaus, G., Dresselhaus, M. S. (1992a), *Chem. Phys. Lett.* **195**, 537–542.

Saito, R., Fujita, M., Dresselhaus, G., Dresselhaus, M. S. (1992b), *Appl. Phys. Lett.* **60**, 2204–2206.

Saito, R., Fujita, M., Dresselhaus, G., Dresselhaus, M. S. (1992c), *Phys. Rev. B* **46**, 1804–1811.

Saito, S., Oshiyama, A. (1991), *Phys. Rev. Lett.* **66**, 2637–2640.

Schlüter, M. A., Lannoo, M., Needels, M., Baraff, G. A., Tománek, D. (1992), *Phys. Rev. Lett.* **68**, 526–529.

Smalley, R. E. (1991), in: G. S. Hammond and V. J. Kuck (Eds.), *Fullerenes: Synthesis, Properties, and Chemistry of Large Carbon Clusters*, ACS Symposium Series, Washington, DC: American Chemical Society.

Smalley, R. E. (1992), *Accts. Chem. Res.* **25**, 98–105.

Sparn, G., Thompson, J. D., Whetten, R. L.,Huang, S.-M., Kaner, R. B., Diederich, F., Grüner, G., Holczer, K. (1992), *Phys. Rev. Lett.* **68**, 1228–1231.

Stankevich, I. V., Nikerov, M. V., Bochvar, D. A. (1984), *Russ. Chem. Rev.* **53**, 640–655.

Stephens, P. W. (1992), *Nature* **356**, 383–385.

Stephens, P. W., Mihaly, L., Lee, P. L., Whetten, R. L., Huang, S. M., Kaner, R., Diederich, F., Holczer, K. (1991), *Nature* **351**, 632–634.

Tanigaki, K., Hirosawa, I., Ebbesen, T. W., Mizuki, J., Shimakawa, Y., Kubo, Y., Tsai, J. S., Kuroshima, S. (1992), *Nature* **356**, 419–421.

Taylor, R., Avent, A. G., Dennis, T. J., Hare, J. P., Kroto, H. W., Walton, D. R. M., Holloway, J. H., Hope, E. G., Langley, G. J. (1992), *Nature* **355**, 27–28.

Tutt, L. W., Kost, A. (1992), *Nature* **356**, 225–226.

Wang, K. A., Wang, Y., Zhou, Ping, Holden, J. M., Ren, S. L., Hager, G. T., Ni, H. F., Eklund, P. C., Dresselhaus, G., Dresselhaus, M. S. (1992), *Phys. Rev. B* **45**, 1955–1958.

Wang, S., Buseck, P. R. (1991), *Chem. Phys. Lett.* **182**, 1–4.

Wang, Y. (1992), *Nature* **356**, 585–587.

Wang, Y., Tomanek, D., Bertsch, G. F. (1991), *Phys. Rev. B* **44**, 6562–6565.

Weaver J. H., (1992), *Acct. Chem. Res.* **124**, 143–149.

Weaver, J. H., Martins, J. L., Komeda, T., Chen, Y., Ohno, T. R., Kroll, G. H., Troullier, N., Haufler, R. E., Smalley, R. E. (1991), *Phys. Rev. Lett.* **66**, 1741–1744.

Wudl, F. (1992), *Accts. Chem. Res.* **25**, 157–161.

Xiang, X.-D., Hou, J. G., Briceño, G., Vareka, W. A., Mostovoy, R., Zettl, A., Crespi, V. H., Cohen, M. L. (1992), *Science* **256**, 1190–1191.

Yu, R. C., Tea, N., Salamon, M. B., Lorens, D., Malhotra, R. (1992), *Phys. Rev. Lett.* **68**, 2050–2053.

Zhang, Z., Chen, C. C., Lieber, C. M. (1991), *Science* **254**, 1619–1621.

Zhou, Ping, Rao, A. M., Wang, Kai-An, Robertson, J. D., Eloi, C. C., Meier, M. S., Ren, S. L., Bi, X. X., Eklund, P. C., Dresselhaus, M. S. (1992a), *Appl. Phys. Lett.* **60**, 2871–2873.

Zhou, Ping, Wang, K. A., Rao, A. M., Eklund, P. C., Dresselhaus, G., Dresselhaus, M. S. (1992b), *Phys. Rev. B* **45**, 10 838–10 840.

Zhu, Q., Cox, D. E., Fischer, J. E., Kniaz, K., McGhie, A. R., Zhou, O. (1992), *Nature* **355**, 712–714.

Further Reading

McLafferty, F. W., Smalley, R. E. (Eds.), (1992) *Accts. Chem. Res.* (1992) **25**, No. 3. A compilation of reviews with numerous references.

Curl, R. F., Smalley, R. E. (1988), "Probing C_{60}," *Science* **242**, 1017–1022.

Curl, R. F., Smalley, R. E. (1991), "Fullerenes," *Sci. Am.* **265**(4), 54–63.

Hammond, G. S., Kuck, V. J. (Eds.), (1992), *Fullerenes: Synthesis, Properties, and Chemistry of*

Large Carbon Clusters, ACS Symposium Series 481, Washington, DC: American Chemical Society.

Kroto, H. (1988), "Space, Stars, C_{60} and Soot," *Science* **242**, 1139–1145.

Smalley, R. E. (1991), "Great Balls of Carbon: The Story of Buckminsterfullerene," *The Sciences* **31**, March-April, 22–26.

FUNDAMENTAL CONSTANTS

See CONSTANTS, FUNDAMENTAL

FUSION, INERTIAL CONFINEMENT

BARRETT H. RIPIN, *Space Plasma Branch, Plasma Physics Division, Naval Research Laboratory, Washington, D.C., U.S.A.*

Introduction 545
1. **History** 547
2. **Principles of Controlled Fusion** 548
2.1 Nuclear Fusion Process 548
2.2 Fusion Fuels: Temperature Requirement 549
2.3 The Lawson Criterion: Fuel Density-Confinement Requirement 549
2.4 Power-Reactor System Requirements 550
3. **Elementary Description of ICF Physics** 550
3.1 The Need for High Pellet Compression 550
3.1.1 Difficulty of Attaining High Pellet Compression 551
3.2 Pellet Gain 551
3.3 Ignition 552
4. **Approaches to Achievement of Inertial Fusion** 552
4.1 Critical Issues of High-Gain Inertial Fusion 553
4.1.1 Good Driver-Target Interaction Physics 553
4.1.2 High Symmetry of Driver-Beam Illumination 557
4.1.3 Hydrodynamic Implosion Stability 557
4.1.4 Robust Pellet Design 558
4.1.5 Reactor-Compatible System 558
5. **ICF Drivers** 559
5.1 Lasers 559
5.1.1 Nd-Glass Lasers 559
5.1.2 KrF Lasers 562
5.1.3 Laser-Driver Summary 562
5.2 Ion-Beam Accelerators 562
5.2.1 Heavy-Ion Accelerators 563
5.2.2 Light-Ion Accelerators 564
5.3 Comparison of ICF Drivers 565
6. **ICF Power Reactors** 567
7. **Recent Progress in the ICF Program** 569
8. **Prognosis and Conclusion** 571
Glossary 572
Works Cited 573
Further Reading 573

INTRODUCTION

Inertial confinement fusion, frequently referred to as ICF, inertial fusion, or laser fusion, is a means of producing energy by imploding small pellets containing thermonuclear fusion fuel. Inertial fusion is the younger of the two principal approaches to the development of controlled nuclear fusion for energy production; the other approach, magnetic fusion, is described more fully under FUSION, MAGNETIC CONFINEMENT. Inertial fusion also has defense applications in addition to energy production.

The inertial fusion concept is simple: A small hollow pellet containing fusionable fuel material is crushed so rapidly and forcefully that it is compressed into a small, dense, hot core that undergoes rapid, explosivelike, nuclear fusion reactions before it flies apart. Figure 1 illustrates this process. Powerful beams of energy, such as light from a laser or particles from an accelerator, are directed at the surface of the pellet with enough intensity to ablate surface material at high speed. The outward-going material acts like a powerful rocket exhaust and accelerates the remaining shell and fuel inward. The fuel is compressed to over 1000 times liquid density by the implosion. If this is done properly, a small region in the center of the compressed fuel becomes hot enough to initiate rapid thermonuclear fusion reactions. Fusion energy is

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

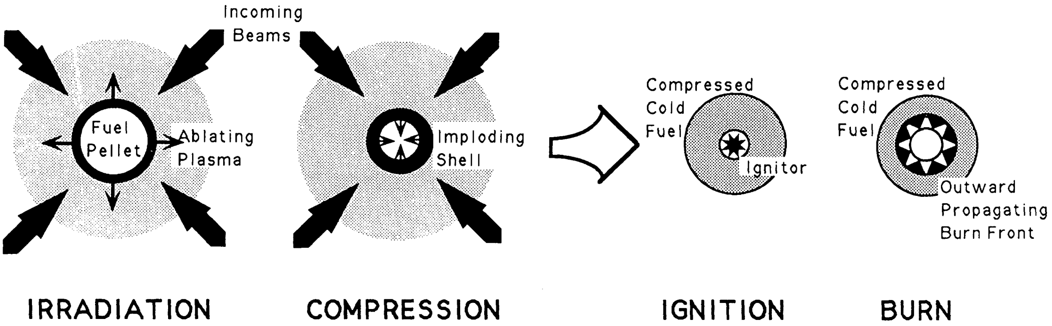

FIG. 1. Inertial confinement fusion concept. The four basic phases of an inertial fusion pellet implosion are (1) *irradiation* of the pellet surface and formation of a plasma with an intense energy beam, such as a laser; (2) *compression* of the pellet and fuel driven by the rocketlike blowoff of the surface material; (3) *ignition* of a central fuel core to about 1000–10 000 times liquid D-T density and 100 000 000 °C; (4) *burn* of the remaining compressed fuel while it is "inertially" confined.

produced when the fuel nuclei collide and bind (fuse) into heavier species, as depicted in Fig. 2 for the two heavy-hydrogen isotopes, deuterium (D) and tritium (T). Fusion energy is the result of the final mass being less than that of the sum of the lighter fuel components; mass is converted to energy. "Thermonuclear ignition" occurs when the energy released from the fusion reactions supplies enough heat to initiate fusion in the surrounding cold compressed fuel. This forms a thermonuclear "burn front" that rapidly propagates outward throughout the remaining fuel. In a small fraction of a nanosecond, before the pellet has time to disassemble, much of the pellet fuel undergoes fusion releasing a large energy burst, a small explosion, with an energy yield over 100 times higher than was initially applied to the pellet. In most applications, this process is repeated over and over again.

Although the ICF concept is simple, the execution has proved to be difficult because of the extreme conditions that need to be satisfied. The driver beams must be very powerful, between 100 and 1000 TW. For reference, the total world's instantaneous electric generation capacity is about 1 TW. At these intensities pressures over 1000 times that of the space shuttle's boosters are applied to the ICF shell. The shell undergoes several trillion *g*'s acceleration to speeds over 350 km/s (about 1000 times faster than a supersonic transport plane) in just a few tenths of billionths of a second. The compressed pellet core will reach about one trillion atmospheres pressure. As incredible as these numbers seem, many have been attained already and others are within sight.

If inertial fusion is perfected, then it may provide a virtually limitless source of energy in the mid-21st century. Energy from the fusion microexplosions is converted into heat for electric power and process heat production in a power plant reactor. Energetic fusion neutrons can break chemical bonds for synthetic fuel production as well as being used to breed tritium fusion fuel, fissile fuel, and radioisotopes. ICF has important defense applications to test hardness of equipment. Some people have even suggested that ICF could someday be used for space-craft propulsion. Spinoff applications and scientific advances have already emerged in such diverse fields as

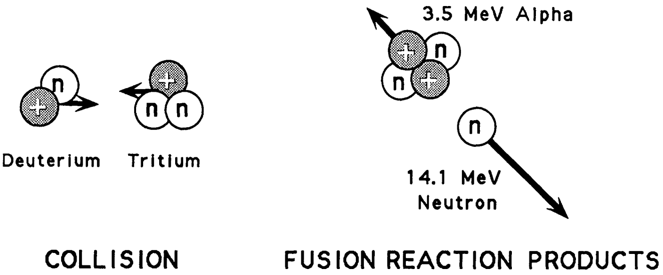

FIG. 2. Fusion reaction of deuterium and tritium. The 17.6-MeV energy release in the D-T fusion reaction is split between an alpha particle (helium nucleus) and a neutron.

x-ray lasers, x-ray lithography, astrophysics, space physics, innovative particle accelerator concepts, materials research, atomic physics, and ultrahigh-speed diagnostic instrumentation.

Controlled fusion is an attractive option for electric power production for a multitude of reasons. First, one of the best thermonuclear fuels, deuterium, is a plentiful constituent of ordinary water. Although common water is predominantly H_2O, the natural isotopic abundance of deuterium to hydrogen is 1 part in 6700. There is enough deuterium fuel in sea water to last several million years. Second, the energy concentration of thermonuclear fuel is high (about 3.45×10^{11} joules per gram for D-D reactions). There is no need for vast arrays of fuel storage tanks. A gallon of water contains about 1/8 gram of deuterium, the equivalent energy of about 300 gallons of gasoline. The total annual fuel for a 1000-MWe power plant, about 15 000 kg, could be carried in one truck. Third, fusion, being a nuclear process, emits no carbon dioxide to contribute to global warming or atmospheric pollutants. Fourth, universally available fuel reduces the impact of flukes of geographic fuel deposits on world politics and economics and acts as a stabilizing effect. Fifth, fusion is relatively safe. Even though fusion is a nuclear process and uses radioactive materials, its waste and accident hazard potential are much lower than those of fission plants.

Any future energy source must be economically competitive to be adopted. Economic analyses of fusion plant designs indicate that power costs are comparable to conventional and fission sources. Capital costs for fusion plants are likely to be higher, because of the complexity of fusion systems, but fuel and environmental costs should be lower. Much of the "hidden costs" of energy, such as distributed environmental effects (e.g., acid rain and CO_2 accumulation), individual health costs (e.g., cancers from radon and other emissions from coal burning), and other energy use costs ignored in the usual accounting, estimated to be about $\$500 \times 10^9$ a year (Hubbard, 1991), would be eliminated with fusion.

If fusion is so wonderful, why do we not have any operating fusion plants yet? The answer, unfortunately, is that the scientific and technical challenges encountered in the quest for controlled fusion have turned out to be among the most difficult undertaken by mankind, and it will be several decades before fusion can be commercialized.

Two vastly different approaches are currently being pursued to harness fusion for power production: magnetic fusion energy (MFE) and inertial confinement fusion (ICF). MFE, the older approach, began in the 1950s. In the MFE approach, a low-density thermonuclear fuel is contained and heated to thermonuclear temperatures in large magnetic "bottles" for long periods of time. The ICF approach was conceived in the 1960s. In contrast to MFE, ICF attempts to fuse very dense fuel during brief periods.

1. HISTORY

Fusion research began with the realization that it was the process that generated the sun's energy. Before this, there was a long standing dilemma: If the solar energy flux is produced by chemical reactions, such as combustion, then the sun could produce energy for only a few thousand years; thus, a much more efficient process had to be responsible. In the 1920s, Arthur Eddington conjectured that mass was being converted into energy, probably by a fusion process. This was not accepted initially because the solar interior was thought to have too low a temperature for sufficient nuclear fusion reactions to occur. Then, in 1928, George Gamow realized that nuclei can "tunnel through" the potential barrier caused by their mutual electrostatic repulsion at low temperatures. In 1931 Hans Bethe worked out the complex cycle of fusion reactions of the inertially (gravitationally) held material in stellar interiors. Solar "inertial" fusion is the Earth's oldest and most benign source of energy. We only intercept a minute fraction of the sun's huge fusion energy production (3.9×10^{26} Watts, from a consumption rate of 6.2×10^{11} kg/s of hydrogen) on Earth. Even so, sunlight, one way or another, generates all the Earth's renewable energy reserves: it drives the wind and the ocean currents, and fosters the growth of plants and trees that, in turn, are burned directly or in the form of petroleum and coal. However, it is difficult to replicate solar inertial fusion conditions on earth, since we do not have the advantage of having the sun's huge gravitational pull to contain the hot thermonuclear fuel indefinitely.

Man has succeeded in producing substantial fusion energy in only one way to date, by the explosion of a thermonuclear (hydrogen) bomb. A key principle of an H-bomb's operation is the use of the inertia of the device's mass to contain the thermonuclear material and delay it from flying apart before the bulk of it can undergo fusion. This is a formidable task since the bomb's thermonuclear fuel must be simultaneously heated to thermonuclear temperature (above a few kilovolts) and compressed to high density (well above that of normal solids). The inertia must resist pressures comparable to those found in stellar cores. The energy source used to drive the hydrogen bomb's fuel to thermonuclear conditions is an atomic (fission) bomb. The ICF concept also relies on inertially held fuel, similar to a thermonuclear bomb, but on a much smaller scale, without the use of a fission driver, and in a safe, controlled manner.

Research in controlled inertial fusion began in earnest in the early 1960s, shortly after the invention of the laser. Researchers (Nuckolls *et al.*, 1972) realized in about 1961 that immense pressures could be applied by using intense laser light to compress small pellets containing thermonuclear fuel analogously to a fusion bomb. Nuckolls introduced the idea of temporally shaping the laser pulse to drive implosions ablatively to densities over 1000 times liquid density. The ICF program was totally classified at this time because of its connection with the fusion bomb program and, perhaps, because achievement seemed imminent. These calculations led to a classified experimental laser fusion program at Lawrence Livermore National Laboratory (LLNL) in 1963. There were also a number of independent theory and experimental programs that sprang up throughout the world in the mid to late 1960s that were on a similar track. Readers interested in a more complete history should see the references.

Although ICF has similarities to its larger weapon cousin, it differs in important regards. ICF pellet explosions must be small enough to be contained safely in a reactor chamber; one milligram of D-T fuel burned completely yields fusion energy with about 340 MJ, the equivalent of 160 pounds of high explosive. The practical limit confinable in a power reactor is a 500–1000 MJ yield. It turns out that achieving fusion yield on a miniaturized scale is much more difficult a task than making a nuclear bomb, and the two are not as closely related as initially thought. The first major declassification of the program occurred in 1972 with the paper of Nuckolls *et al.*

However, impressive progress is being made and scientists are optimistic about achieving ICF. The remainder of this article will describe these current research trends.

2. PRINCIPLES OF CONTROLLED FUSION

2.1 Nuclear Fusion Process

Nuclear fusion is the result of the collision and binding of two or more lighter nuclei. The D-T fusion process, illustrated in Fig. 2, is exothermic because the final total mass is less than that of the original deuterium and tritium, even though a heavier alpha particle is formed in the process. The energy released, E_r, is related to the lost mass Δm and the speed of light by Einstein's well-known relation $E_r = \Delta mc^2$. The 17.59-MeV excess energy from the D-T mass loss appears as kinetic energy of the products, a 3.5-MeV alpha particle and a 14.1-MeV neutron. Fusion can be thought of as the inverse of the more familiar nuclear fission process, which is the *splitting* of a heavy nucleus into two or more lighter parts. Fusion is exothermic for light elements; fission releases energy when heavy elements, such as uranium, split. The colliding nuclei must be very energetic to overcome the "Coulomb barrier" between them, caused by their mutual electrostatic repulsion, to get close enough for their nuclear forces to act and make them fuse. Singly charged hydrogen isotopes require about 200-keV directed energy to overcome the Coulomb barrier; more highly charged species require even greater directed energy. Overcoming the Coulomb barrier is a very stiff requirement. The good news is that a deuterium-tritium fuel thermal temperature of only $(50\text{–}100)\times 10^6$ °C (i.e., 5–10 keV) has a sufficient fraction of nuclei in the high-energy tail of its distribution function to be useful for fusion energy production. The adjective "thermonuclear" modifying fusion refers to the use of such a "thermal" fuel distribution. The bad news, of course, is that these temperatures are still exceedingly high and no material can contain them. At the temperatures required for thermonuclear fusion, all matter is in the plasma state, that is, ionized. Plasmas are made up of charged particles and respond in a complex manner to electric and magnetic fields in addition to the usual hydrodynamic forces. In fact, a large fraction of the obstacles to perfecting fusion power lies in the difficulty of controlling the response of the plasma fuel to the strong long-range collective electrical forces introduced by its own charged-particle nature.

2.2 Fusion Fuels: Temperature Requirement

D-T is the most frequently considered fuel mixture for controlled fusion energy production because it has the lowest temperature threshold requirement, high fusion reaction rate, and high Q value, the energy yield per reaction. The D-T fusion cross section has a resonance near zero energy and a peak of about 5 barns (5×10^{-24} cm^2) at 125 keV. Although tritium is not naturally plentiful, since it is unstable with about a 13-year half-life, it can be bred in the lithium thermal blanket of a fusion reactor. Other fusion reactions require higher temperatures and have lower reaction rates. Table 1 lists several fusion reactions that are candidates for energy production. Doubtless, the first fusion reactors will operate on D-T fuel, but so-called advanced fuels, such as the ones using deuterium (D-D and D-^{3}He) and proton reactions, have reactor advantages. D-D reactions do not require tritium breeding. The proton-fusion reaction products are neutronless. D-^{3}He is attractive because there are no radioactive reaction products and the specific yield is high. Unfortunately, the supply of ^{3}He on earth may be extremely small. However, ^{3}He is plentiful on the moon's surface, where it might be mined, and recently it was found in high concentration in the sea water near Pacific Ocean bottom thermal vents.

2.3 The Lawson Criterion: Fuel Density-Confinement Requirement

Even when the plasma fuel is hot enough to overcome the Coulomb barrier, the probability of two nuclei being on a collision path is quite small. The fuel must therefore be held together long enough for most of it to fuse. The containment requirement is given by Lawson's criterion, which for equal mixtures of D-T fuel above 10 keV temperature is

$$n\tau \geqslant 10^{20}\ \mathrm{s\,m}^{-3}. \tag{1}$$

The fuel number-density confinement-time product $n\tau$ must exceed the value of the right-hand side of Eq. (1), the Lawson number; this is basically a trade-off between density and confinement time. MFE and ICF attempt to satisfy Lawson's criterion in diametrically opposite ways. Namely, for MFE, low-density fuel ($n > 10^{20}$ m^{-3}, about 10^{-5} of atmospheric density) is heated and contained for long periods ($\tau > 1$ s) in magnetic bottles;

Table 1. Thermonuclear reactions considered for controlled fusion. D-T will be the fusion fuel in the earliest reactors since it has the lowest threshold temperature and ρR requirements, and high energy release Q. In general, "advanced fuels" have stringent requirements and lower yield, although they offer other advantages. ^{11}B-p fusion is non-neutronic, and has no radioactive byproducts, and it is conceivable that electricity could be generated directly from its charged reaction products without a thermal conversion stage. D-^{3}He offers similar advantages to ^{11}B-p, and has a lower threshold and high Q, but ^{3}He has a high temperature threshold for fusion and may be extremely rare on Earth. The energy released to each product from the fusion reactions shown in column 2 is in parentheses in MeV; n stands for neutron, p stands for proton, and notation ^{A}He refers to the isotope of He with atomic number A.

Fuel	Fusion reaction	Q (MeV)	T (keV)	ρR (g/cm^2)
D, T	^{4}He(3.5)+n(14.1)	17.6	5–10	2–5
D, D	^{3}He(0.82)+n(2.45) T(1.01)+p(3.02)	3.65	20–50	10–20
^{11}B, p	3 ^{4}He(8.68) ^{4}He+Be(8.59)	8.68	100–400	50–500
D, ^{3}He	^{4}He(3.6)+p(14.7)	18.3	~50	

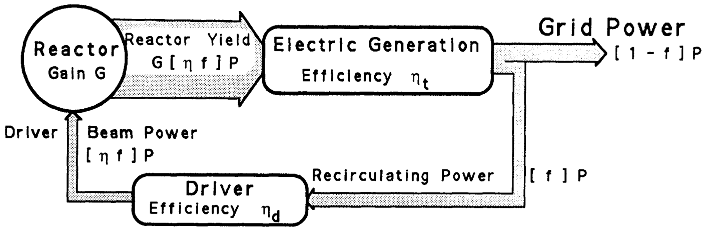

FIG. 3. Power flow in an ICF power station. Power supplied to the grid is the reactor output less the portion wasted in the thermal portion of the power plant, heat exchanges, turbines, generators, etc. (typically 30%–40%) and the fraction *f* needed to run the driver and other power station systems (<25%).

for ICF, high-density fuel ($n > 10^{31}$ m^{-3}) is produced for ultrashort "inertial" times ($\tau < 10^{-10}$ s).

The Lawson criterion has a more convenient form for the inertial fusion case, i.e., in terms of the fuel mass-density radius product, ρR; it is

$$\rho R > 3 \text{ g/cm}^2. \qquad (1')$$

Equation (1′) follows from Eq. (1) by noting that number and mass density are related by $n = \rho / m_{\text{ion}}$ and that the dissembly time of a sphere of radius R, the equivalent confinement time for ICF, is $\tau \approx R/4c_s$, where c_s is the speed of sound in the fuel. Actually, the requirement for ICF, Eqs. (1′) and (1), is about an order of magnitude higher than the MFE equivalent because a much higher fraction (~30%) of the available fuel needs to be burned in a pellet explosion compared to a marginal MFE reactor. The ρR criterion dictates that the D-T fuel must be compressed to over 1000 times its normal liquid density of $\rho_{\text{DT}} = 0.2$ gm/cc. This requirement will be discussed in more detail later.

2.4 Power-Reactor System Requirements

For a fusion power plant to be economically practical, its output must exceed all of its own operating needs, including inefficiencies, by a large margin. The recirculating power fraction needed to run the plant should remain below 25% for economic viability. Figure 3 shows a block diagram of the power flow in an ICF power reactor. Assuming the power plant has 35% thermal efficiency, it is easy to estimate the minimum required pellet gain, G, i.e., pellet energy yield divided by its absorbed driver energy. In terms of the driver efficiency (η_d) of converting its operating power into a useful beam, the criterion is

$$\eta_d G > 10. \qquad (2)$$

Equation (2) dictates a pellet gain requirement, for example, of $G > 100$ for a 10% efficient driver. Laser drivers under consideration have efficiencies of <10% and would require gains over 100, whereas ion-beam accelerators with projected efficiencies of 25%–35% could get by with pellet gains as low as 40 or 30. A great deal of current ICF research effort focuses on determining whether this gain-efficiency product can be met, and what driver-pellet combination is best. It should be pointed out that most other ICF applications have less stringent requirements than power production.

3. ELEMENTARY DESCRIPTION OF ICF PHYSICS

A good description of the basic physics involved in inertial fusion can be made from elementary considerations. We go through the basic ICF principles below.

3.1 The Need for High Pellet Compression

As noted earlier, limiting the maximum yield of the pellet explosion to 500–1000 MJ or so tells us that we are dealing with at most milligrams of D-T and pellets of millimeter dimensions. A millimeter-radius pellet has an initial D-T fuel $\rho_0 R_0$ of about 0.02 gm/cm^2 assuming, for simplicity, that it is a uniform drop at liquid density ($\rho_{\text{DT}} = 0.2$ gm/cc). The fuel must reach $\rho R > 3$ to satisfy the Lawson criterion, Eq. (1′), a factor of 150 higher than its initial state. Fuel compression is the only solution. An estimate of the final compressed fuel radius needed is made by noting that

$$\frac{\rho R}{\rho_0 R_0} \approx \frac{[m/\frac{4}{3}\pi R^3]R}{[m/\frac{4}{3}\pi R_0^3]R_0} \approx \left(\frac{R_0}{R}\right)^2 \sim 150 \Rightarrow \frac{R_0}{R} \sim 12.$$

Since $\rho/\rho_0 \sim (R_0/R)^3$, a convergence ratio of $R_0/R \sim 12$ corresponds to a final compressed fuel density of order 2000. Detailed computer calculations call for compression between 1000 and 10 000. It is advantageous to strive for compression levels above the minimum. The higher the fuel density, the lower the required driver energy; the higher the fraction of fuel that undergoes fusion reaction, the higher the pellet gain.

3.1.1 Difficulty of Attaining High Pellet Compression. Anyone who has attempted to squeeze and compress solid- or liquid-phase material knows that it requires tremendous pressures because of its lack of compressibility. The minimum specific energy needed to compress cold D-T is approximately the same as for a Fermi-degenerate electron gas, $\mathscr{E}_{\text{Fermi}}$, assuming that the implosion is isentropic, i.e., the fuel is heated only by compressive work. If excess heat, termed "preheat," gets into the fuel during the implosion process then more driver energy is needed. Preheat is caused by energetic particles, radiation, or shock-wave deposition from the driver-pellet interaction. Compression energy $\mathscr{E}_c$ more than a factor of two over the minimum $\mathscr{E}_{\text{Fermi}}$ level begins to be problematic. We can estimate the energy needed to compress an ICF pellet by referring to Table 2. A minimum energy of about 11 kJ per milligram D-T (10^{-3} gm $\times\, \mathscr{E}_{\text{Fermi}}$) is needed to compress the fuel to 1000 times liquid density. Thus, the minimum beam pulse energy is over 100 kJ since less than 10% of the driver energy gets converted into compressive energy. The maximum acceptable driver output is in the range of 5–10 MJ, assuming a pellet gain of 100 and a reactive explosive yield limit of 500—1000 MJ.

Other phenomena can ruin pellet implosion besides preheat. Fuel can squirt out of the fuel core before reaching high density if nonuniformities or pressure imbalances are present, just as Jell-O will ooze between your fingers if squeezed. It is critical that the pressure driving the implosion be applied very symmetrically and uniformly. Any latent nonuniformities are exacerbated by inevitable hydrodynamic instabilities that occur during pellet compression. One particularly troublesome hydrodynamic instability, Rayleigh–Taylor (RT), occurs when high-density fuel is pushed (accelerated) by a much lower density ablation plasma. Uniformity requirements put stringent constraints on the driver, geometry, and implosion approach and will be discussed further later.

3.2 Pellet Gain

Pellet gain G is the ratio of thermonuclear energy yield to the driver energy incident on the pellet assembly (also known as the target capsule), i.e.,

$$G=\frac{E_{\text{out}}}{E_d}=\eta\phi\left(\frac{\mathscr{E}_b}{\mathscr{E}_c+(m_h/m)\mathscr{E}_h}\right). \qquad (3)$$

The expression on the right-hand side of Eq. (3) results from recognizing that the energy yield is a product of the total fuel mass m, the thermonuclear energy released per unit mass $\mathscr{E}_b$, and the fraction of the fuel that undergoes fusion ϕ, i.e., $E_{\text{out}}=\phi m\mathscr{E}_b$. Only a small fraction η of the driver energy couples usefully to the fuel to compress it and to heat a small mass m_h in the ignitor region to thermonuclear burn temperature, i.e., $\eta E_d=m\mathscr{E}_c+m_h\mathscr{E}_h$. The beam-pellet coupling efficiency η includes both reflective beam losses and inefficiency in the rocket ablation process; as mentioned earlier, η is typically about 0.1. It is difficult to achieve a fuel burn-up fraction ϕ greater than 0.3, which limits the coefficient in front of the bracket to only about 0.03. To obtain gains of about 100 then requires fuel gain [the quantity in parentheses in Eq. (3)] over 3000.

The tabulation of $\mathscr{E}_b$, $\mathscr{E}_h$, and $\mathscr{E}_c$ for D-T in Table 2 allows us to estimate the maximum gain expected from an ICF pellet. If *all* the D-T fuel in the pellet is uniformly heated to 5 keV, the best fuel gain that can be expected is less than 600, resulting in an overall pellet gain of only about 20, which is far too small for power plant use. The critical trick, therefore, is to heat only a central core, the ignitor region. The ignitor region should contain the least amount of fuel that can initiate a prop-

Table 2. Characteristic energy densities involved in pellet fusion using D-T fuel.

Burning D-T yield	$\mathscr{E}_b=3.4\times10^{11}$ J/g
Heat to burn-temperature	$\mathscr{E}_h=5.8\times10^{8}$ J/g (@ 5 keV)
Energy to compress fuel	$\mathscr{E}_c>\mathscr{E}_{\text{Fermi}}=1.1\times10^{7}$ J/g (@$1000\times n_{\text{liq}}$) (note $\mathscr{E}_{\text{Fermi}}\propto n^{2/3}$).

agating fusion burn throughout the remaining cold fuel. This technique of heating only a fuel core to ignition is called "hot-spot ignition."

3.3 Ignition

Ignition occurs when the hot-core (ignitor) radius R_c is greater than an α-particle mean free path R_α, where R_α is given by

$$\rho R_\alpha \approx 0.3 \text{ g/cm}^2 \quad \text{for} \quad T_c = 10 \text{ keV}. \qquad (4)$$

About 20% of the ignitor fusion yield is in α particles that, in turn, can heat about six times the original ignitor mass of cold fuel up to fusion temperature. This process cascades and a nuclear fusion "burn front" propagates outward into the dense cold fuel. The ρR Lawson criterion, Eq. (1′), for the ignitor is an order of magnitude smaller than for the pellet as a whole. There is a great premium on achieving as high an ignitor density as feasible because its mass, and hence energy requirement, decreases dramatically, viz., $m_h = \rho R_\alpha^3 \sim (\rho R_\alpha)^3/\rho^2$. The upper limit fuel gain of 30 000 ($\sim \mathcal{E}_b/\mathcal{E}_{\text{Fermi}}$), corresponding to an overall simple pellet-gain limit of 1000, is not really attainable. This is because fuel is heated in the ignitor, the ignitor region will be larger than the theoretical minimum because of asymmetries and mixing of cold and hot fuel at its boundary, and the actual energy required to compress the fuel, $\mathcal{E}_c$, is higher than the minimum Fermi-degenerate level, $\mathcal{E}_{\text{Fermi}}$.

The pellet gain needed for power production lies between 30 and 1000. The lower limit assumes the driver has a high efficiency ($\sim$30%), such as might be attainable from an ion-beam accelerator; the upper limit is near the theoretical limit for a simple ICF pellet. A gain of 100–150 would be sufficient with a 10% efficient driver and is the generally accepted program goal. Success is not assured even though the pellet gain needed is below the ideal by an order of magnitude. The pellet implosion process is extraordinarily complex because of the extremes of density and temperature conditions created and highly nonlinear interactions of the driver beam with the pellet. Close attention must be paid to optimizing the driver coupling efficiency and implosion physics compatibility. Although this is a very challenging task, steady and impressive progress is being made.

Figure 4 shows a generic ICF gain curve. For driver energies below that required for hot-spot ignition, the yield is from volumetric fusion of fuel nuclei in the tail of the thermal distribution. A driver energy in the ignition region satisfies Eq. (4) and propagating burn begins; the thermonuclear yield rapidly increases forming an "ignition cliff." Finally, for driver energies above the minimum ignition level, the gain increase becomes proportional to the additional fuel mass (and driver energy). Energy application requires a driver-efficiency–gain product $\eta_d G$ over 10 to overcome the inefficiencies of the implosion process and power-plant thermal cycle, as noted earlier; this requirement is shown as a hatched area on the diagram. In addition, 500–1000 MJ is the maximum yield that can be reasonably confined in a reactor; hence, the practical driver pulse limit is between 5 and 10 MJ.

4. APPROACHES TO ACHIEVEMENT OF INERTIAL FUSION

Two approaches to achieve high-gain implosions are being pursued, namely,

1. direct drive, in which the driver beam strikes the pellet surface directly, and
2. indirect drive (also known as the "classified approach" or x-ray drive), in which the beam energy is first converted to x rays in a *Hohlraum*; it is the x rays that then strike and implode the pellet.

The direct-drive and indirect-drive ideas are sketched in Fig. 5.

Each approach has advantages and limitations, and neither has been proven as yet to be suitable (or unsuitable) for achieving high-gain controlled fusion. Direct drive has advantages of simplicity and potentially higher efficiency, but has stringent driver illumination symmetry requirements. Driver-beam uniformity requirements are reduced in the indirect-drive approach, but its efficiency is likely to be lower as a result of the extra x-ray conversion step. Excellent progress is being made in developing both approaches and it is premature to choose one over the other at this time. Both approaches must ultimately achieve similar pellet conditions.

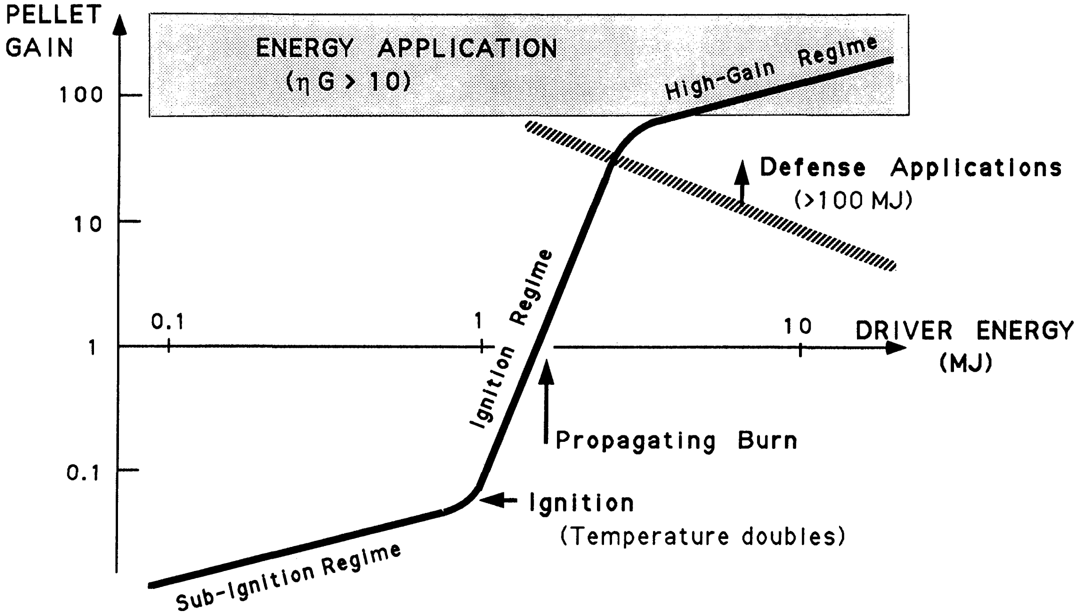

FIG. 4. Gain curve for an ICF pellet. Fusion yield gradually increases with driver energy until ignition conditions are met in the central core. Above ignition, fusion products from the core ignite surrounding fuel and the gain rapidly increased; this is the ignition "cliff." At still higher driver energy, the gain increases in proportion to the additional fuel mass.

4.1 Critical Issues of High-Gain Inertial Fusion

High-gain pellet requirements for energy production are listed in Table 3. The same basic conditions apply to both direct and indirect drives, but sometimes in different manifestations. The first three requirements are common to defense and energy implementations, while the last two requirements are mainly energy application concerns. Each requirement is briefly discussed in turn.

4.1.1 Good Driver-Target Interaction Physics. Good driver-beam absorption in the plasma formed outside the pellet surface is important. The size (and therefore cost) of these sophisticated drivers increases rapidly with energy. Recall that the driver efficiency needed to meet the $\eta_d G > 10$ criterion is about 10%, and so there is little leeway for additional losses. Additionally, beam energy must be absorbed properly to generate the high ablation pressures needed to implode the pellet efficiently. Ablation plasma with a thermal energy distribution is the best. Energetic (also called "fast" or "suprathermal") electrons generated in some beam-target interaction processes can penetrate and deposit energy into the dense fuel. If too many suprathermal electrons are present they can heat the fuel

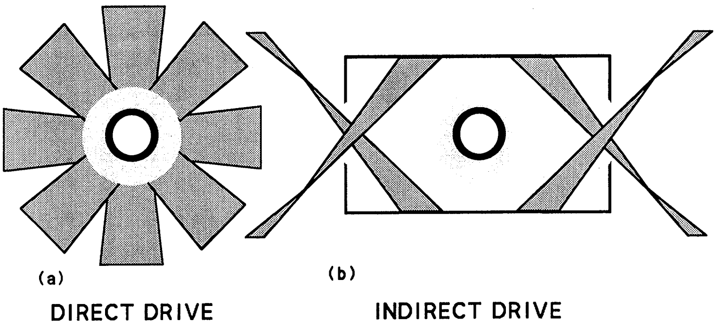

FIG. 5. Comparison of direct- and indirect-drive ICF approaches. (a) In the direct-drive method, the drive beams focus onto the pellet surface to drive the implosion. (b) In contrast, in the indirect-drive approach, the driver beams create x rays in a *Hohlraum* which, in turn, strike the pellet surface to drive the implosion.

Table 3. Critical issues of inertial fusion: approximate conditions that need to be satisfied for successful inertial-fusion energy production.

Good driver-target interaction physics	
High deposition of driver energy onto the pellet	$75\% < \eta_d < 100\%$
Minimal preheating of the cold thermonuclear fuel	$E_{\text{preheat}} < 1\%\ E_d$
High ablation pressure and hydrodynamic efficiency	$P > 10$ Mbar; $\eta > 10\%$
High symmetry of driver-beam illumination	
Uniform pressure for pellet implosion	$\Delta P/P < 1\%$
Hydrodynamic Implosion Stability	
Achieve high pellet compression (densities)	$n_c > 1000 n_{\text{liquid}}$
Minimal mix of hot and cold fuel or ablator	$m_{\text{mix}} < (\text{few } \%)\ m_h$
Robust pellet design	
Low driver energy for ignition and gain	$E_d < 10$ MJ
High-gain potential	$G \geqslant 100$
Reactor compatible system	
Efficient driver	$\eta_d > 10\%$–50%
Repeatable driver	$f_{\text{rep}} \sim 5$–10 Hz
Reliable and reactor compatible driver	Millions of shots
Economically competitive	cents/kWh

prematurely (preheat). The higher back pressures caused by preheated fuel prevent the high compression needed for high-gain explosion. Just a percent or so of the driver energy in suprathermal electrons can spoil the process. Energetic ions can be accelerated in the absorption process, too; these are undesirable because they transmit momentum to the pellet poorly (lower the pressure) and they waste driver energy.

Virtually all ICF pellet implosion research in the past, and in the foreseeable future, uses laser drivers. This is because they alone can presently generate the high irradiances needed for pellet implosion.

Laser light cannot penetrate beyond the so-called critical density, n_c, in the plasma surrounding the pellet shell, as illustrated in Fig. 6(a). The critical density is well below solid density; light is absorbed in this "underdense" ablation plasma by inverse bremsstrahlung and other processes. The critical density is the density at which the incoming light frequency equals the local plasma frequency. The plasma frequency is the natural oscillation rate of electron waves in a plasma and is given numerically by $f_{\text{pe}} = 9 \times 10^3 n_e^{1/2}$ Hz, where n_e is the electron density in electrons/cm^3. For 1-μm wavelength laser light the critical density is 10^{21} electrons/cm^3. The critical density scales with light wavelength like λ^{-2}; hence, the shorter the laser wavelength, the higher the density to which light can penetrate. Energy absorbed in the underdense plasma flows up the density profile through heat conduction to the pellet surface, where it ablates shell material that in turn applies pressure to drive the shell inward. The closer the absorption region is to the ablation surface, the higher the hydrodynamic efficiency, the fraction of the absorbed energy converted into kinetic energy of the imploding shell and fuel.

Poor laser absorption and suprathermal particle generation have been major problem until recently. Normally, light travels up the plasma density gradient towards the target, where it is absorbed by a process called inverse bremsstrahlung. In the inverse bremsstrahlung process, electrons absorb photons as they are jiggled by the light wave's electric field near ions; the energy taken up by the electrons collisionally couples to the ions. It is the inverse process to ordinary bremsstrahlung, photon generation due to electrons scattering from ions. The higher the critical density (shorter wavelength light), the more efficient is the absorption process. Unabsorbed light at the critical density reflects back towards the laser.

At the very high light intensities needed for ICF, the interaction of the light with the target plasma can also exhibit some fascinating, but usually deleterious, processes. For laser irra-

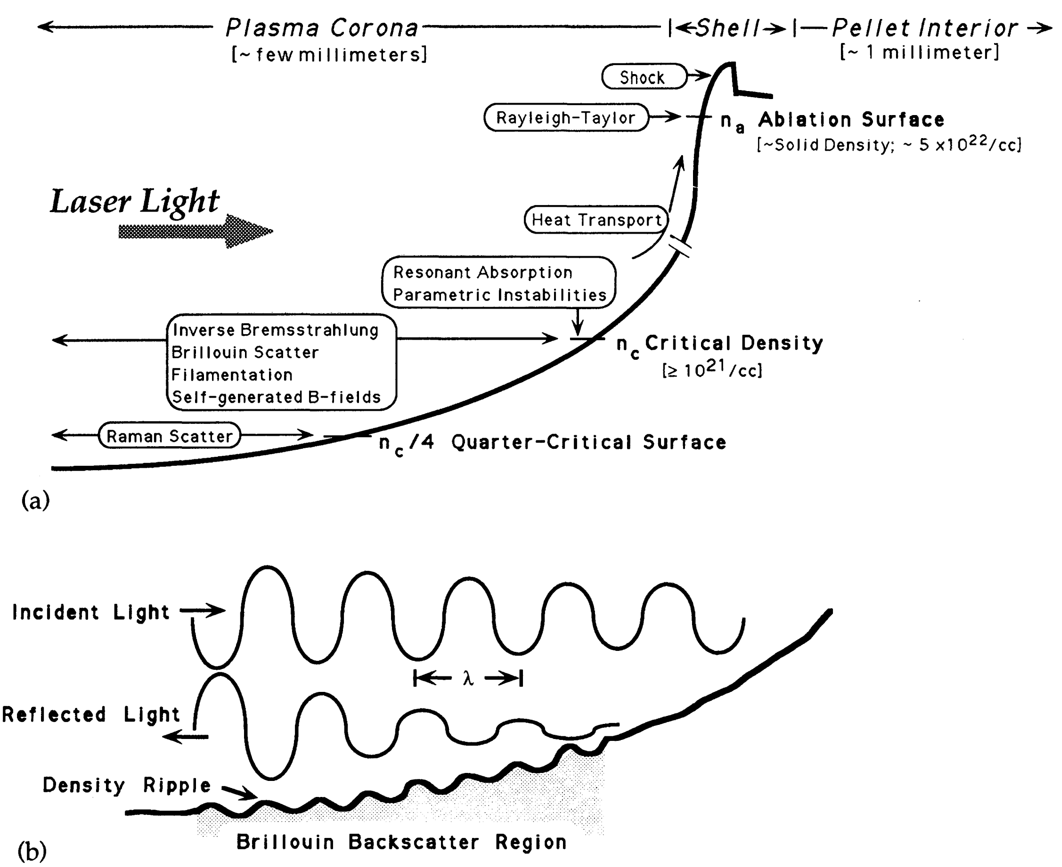
Plasma Corona
[~ few millimeters]
Shell
Pellet Interior
[~ 1 millimeter]
Shock
Rayleigh-Taylor
n_a Ablation Surface
[~Solid Density; ~ 5 x10^{22}/cc]
Laser Light
Heat Transport
Resonant Absorption
Parametric Instabilities
Inverse Bremsstrahlung
Brillouin Scatter
Filamentation
Self-generated B-fields
n_c Critical Density
[≥ 10^{21}/cc]
Raman Scatter
$n_c/4$ Quarter-Critical Surface
(a)
Incident Light
λ
Reflected Light
Density Ripple
Brillouin Backscatter Region
(b)

FIG. 6. Interaction and absorption of laser light. (a) Regions in which laser light interaction processes occur in the underdense ablation plasma. The underdense plasma extends up to the critical density, n_c (10^{21} electrons/cm^3 for 1.06-μm Nd-laser light), about a factor of 50 times less dense than solid D-T. (b) Mechanism responsible for the stimulated Brillouin scatter (SBS) instability. SBS is capable of reflecting almost all the incident light before it is absorbed, a fascinating but very undesirable effect.

diance-(wavelength)^2 products $I\lambda^2$ above a few times 10^{13} W μm^2/cm^2, the light can modify the plasma density profile by pushing plasma out of its high-intensity electric field regions through a process called ponderomotive force. Ponderomotive force in a plasma is analogous to vacuum radiation pressure, but is much stronger for a given wave intensity. There are hosts of plasma instabilities that are driven by ponderomotive-force–induced density changes. Ponderomotive force effects tend to be self-aggravating and to grow to large amplitudes, and can become very disruptive. For instance, a fascinating process known as stimulated Brillouin scatter (SBS) can reflect over 80% of the incident 1-μm (Nd-laser) light before it even reaches the absorption layer!

Stimulated Brillouin scatter instability is in a class of nonlinear phenomena called parametric instabilities. Parametric processes involve the interaction of a large-amplitude wave, often called the pump, with two or more other waves in the system. Typically, density ripples form in the plasma from the interference of the incident and reflected waves, as illustrated in Fig. 6(b). It is the ponderomotive force that pushes plasma away at the peaks of the interference pattern to create the ripple. The density ripple acts much like a Bragg reflection grating with just the right periodicity (half the incident laser wavelength) to backscatter more of the incident wave; this in turn further deepens the density ripple, causing more scatter, and so on. The three waves involved are the incident (i), the reflected (r), and the wave caused by the density ripple, commonly called the beat (b) wave. The three waves must have frequency and wave numbers that conserve energy and momentum; this is equivalent to satisfying the following relations:

$\omega_i = \omega_r + \omega_b$, energy conservation

$k_i = k_r + k_b$, momentum conservation.

The parametric processes are particularly strong when all three waves are normal modes (resonant, natural waves) of the plasma. Since plasma can support many types of waves, there are several possible parametric instabilities. Parametric processes generally require the incident intensity to be above some "threshold" value to become unstable. The threshold intensity occurs at the point where the parametric growth of the secondary waves just exceeds their natural damping rates; for 1-μm light, threshold values are in the same range as intensities needed for pellet implosion.

Underdense parametric instabilities important in the laser fusion context include the stimulated Brillouin scatter instability (SBS), in which the incident and reflected electromagnetic waves interfere to form a plasma ion acoustic (sound) wave; the stimulated Raman instability (SRS), in which the beat wave is an electron plasma wave; and the two-plasmon decay (TPD) instability, in which incident light decays into two plasma electron waves at one quarter of the critical density. These underdense instabilities tend to reflect the light and generate very energetic, "suprathermal," electrons.

Light not reflected in the underdense plasma can still cause trouble. Intense laser light reaching the critical density region can steepen it by the same ponderomotive force phenomena. The critical-surface steepening in turn suppresses inverse bremsstrahlung and promotes undesirable absorption processes, such as resonant absorption and parametric-decay instability. Resonant absorption and parametric-decay absorption are less efficient than inverse bremsstrahlung and generate undesirable suprathermal populations of electrons and ions. Figure 6(a) schematically indicates where many of these processes occur in the density profile leading up to the target surface. Most parametric phenomena get worse in larger (more reactorlike) plasmas. Since laser irradiance above 10^{14} W/cm^2 is needed to implode pellets, parametric instabilities could be very strong and disruptive for the multimillimeter-sized plasmas characteristic of ICF reactor pellets.

The laser-light interaction and absorption problems discussed above are now largely resolved as a result of extensive theoretical and experimental investigation by the laser fusion community. One way to alleviate problems is to avoid pellet designs that require highly shaped pulses with intense peaks ($>10^{16}$ W/cm^2) that were prevalent in the early phases of the ICF program. However, it is now clear that the best solution is to use short-wavelength laser light. Figure 7, a plot of absorption fraction versus laser irradiance for 1-μm laser light and its second, third, and fourth harmonics, shows the benefits of short-wavelength light clearly. This finding rules out the use of CO_2 lasers ($\lambda \approx 10$ μm), HF or DF lasers ($\lambda \approx 2$–3 μm), and iodine lasers ($\lambda \approx 1.3$ μm), which were once candidate drivers. It also rules out using the fundamental wavelength of Nd-glass lasers ($\lambda \approx 1$ μm) and probably even its second harmonic ($\lambda \approx 0.5$

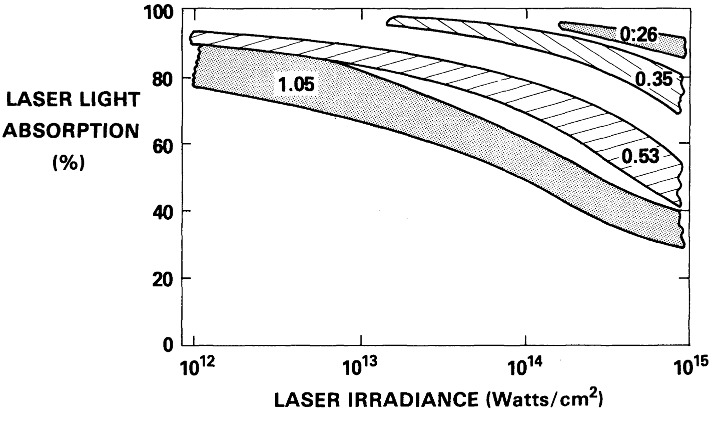

FIG. 7. Light absorption decreases with increasing laser intensity. Low absorption is due, in large part, to parametric instabilities. Note that absorption gets better with shorter-wavelength (higher-frequency) laser light.

μm), which have been the most common research drivers through the 1980s. Fortunately, Nd-laser light can be converted to its third or even fourth harmonic efficiently ($\sim$70%). Short-wavelength light is used almost exclusively in most current research programs. The KrF gas laser, which operates at $\lambda \approx 0.25$ μm and has both a favorable wavelength and the potential of reasonable efficiency ($\sim$10%), is currently being developed for ICF. Finally, even though it appears that the major laser-target interaction issues are settled, researchers have, more frequently than not, been confronted with unexpected results whenever new parameter regimes are explored. Therefore, laser interaction issues could well resurface when the next generation of experiments produce more reactorlike conditions.

Absorption processes for light- and heavy-ion beams are thought to be more straightforward and inherently benign compared to those of laser light. Initial experiments on stopping distances of ions in solids and plasma support this expectation. However, no particle-beam experiments have been performed at parameters even close to those required for ICF. One open question is whether particle beams are compatible with the direct-drive approach. This is because particle beams tend to deposit energy deeper into solid-density target material than laser ablation fronts and it is not yet clear whether high-gain direct-drive targets can accommodate this.

4.1.2 High Symmetry of Driver-Beam Illumination. Pellets need to implode to about $\frac{1}{15}$ of their initial radius to compress their fuel to several thousand times liquid density. Before irradiation, most pellet shell designs have wall thicknesses, ΔR_0, between about $\frac{1}{10}$ and $\frac{1}{50}$ of their radius R_0; i.e., have "initial aspect ratios," $R_0/\Delta R_0$, between 10 and 50. The shells become several times thinner than their initial values when they are subjected to the pressure and ablation of the intense driver irradiation. This "in-flight aspect ratio," $R/\Delta R$, which is typically in the range of 30–100, is the relevant quantity for determining implosion uniformity requirements and Rayleigh–Taylor instability susceptibility. During implosion, the shells must traverse distances of about 100 times their own thickness uniformly without breaking. Simple kinematics dictates that pressure nonuniformities over the pellet surface be no worse than about 1%. This exceptional pressure uniformity requirement translates into almost the same degree of uniformity for the incident-beam irradiance. This is especially true in the direct-drive approach with short-wavelength lasers where absorption occurs close to the ablation surface. Transmittal of beam nonuniformities to the pellet is the one disadvantage of short-wavelength light. Typical high-powered lasers tend to be rather nonuniform because of the very high amplification they undergo, even if multiple beams overlap on the target. Fortunately, there has been spectacular progress made in solving the driver uniformity problems. A variety of beam smoothing and balancing techniques were developed which can give pellet illuminations of the required quality. Smoothing methods are described later in this article in Sec. 5.1.

Breakthroughs in beam smoothing have largely removed the principal disadvantage of the direct-drive compared to the indirect-drive approach. Since the driver beams do not strike the pellet surface directly in the indirect approach, but rather their generated x rays, beam uniformity requirements are reduced and more driver flexibility is allowed.

4.1.3 Hydrodynamic Implosion Stability. Hydrodynamic stability of the pellet during implosion is closely related to the beam-symmetry issue discussed above. Both phenomena affect the ability to maintain a spherical pellet implosion. As discussed above, the beam needs to have about 1% or better uniformity. However, even with perfectly smooth illumination small unavoidable imperfections in the pellet itself will grow in amplitude through hydrodynamic instability.

Hydrodynamic instability, specifically the Rayleigh–Taylor (RT) instability, occurs because of a system's tendency to go to its minimum energy state. For instance, if a glass of water is turned upside down and the water-air surface is *perfectly* smooth, the 14-lb/in.^{2} air pressure is sufficient to hold the water in the glass. However, the slightest ripple on the water surface grows exponentially through Rayleigh–Taylor instability and the water sloshes out onto the floor, the lower-energy state. Rayleigh–Taylor instability can occur when a low-density fluid supports or accelerates a higher-density fluid.

Pellet implosions have two regions where hydrodynamic instability is important. Pressure from the low-density ablation plasma accelerates the dense pellet shell and fuel inwards and, hence, this interface is unstable. Hydrodynamic instability occurring near the ablation surface can ruin the implosion by breaking up the pellet shell before achievement of maximum compression. The pellet implosion is also susceptible to RT instability at the interface between the hot ignitor and the surrounding cold denser fuel towards the end of the compression process. The amplitude of RT growth during the linear phase is

$$A(t)=A_0 \exp(\gamma_{\mathrm{RT}} t), \qquad (5)$$

where the Rayleigh–Taylor growth rate, γ_{RT}, is

$$\gamma_{\mathrm{RT}}=\alpha(k_p g)^{1/2} \qquad (6)$$

for two fluids separated by a sharp interface undergoing an acceleration of magnitude g. The wave number of the perturbation, k_p, is related to the perturbation wavelength λ_p by $k_p=2\pi/\lambda_p$. The Attwood number α is unity for a large density jump between these two fluids.

The most dangerous RT wavelength for ICF pellet implosion near the ablation surface is $\lambda_p \sim \Delta R$, the thickness of the pellet shell. The acceleration g is approximated by using $R=gt^2/2$, and, hence, the number of growth e-foldings of this most dangerous mode, $N_{m\mathrm{RT}} = \gamma_{\mathrm{RT}} t$, is approximately

$$N_{m\mathrm{RT}} \approx (2\pi R/\Delta R)^{1/2}. \qquad (7)$$

Rayleigh–Taylor instability is a very serious problem, as can be seen by noting, for example, that ripples on a shell with an in-flight aspect ratio $R/\Delta R$ of 100 will grow 25 e-folds (or 7.7×10^{10} times) in amplitude, clearly unacceptably large.

The pellet design must keep RT growth at an acceptable level. This is a difficult, but not impossible, task. Fortunately, there are clever ways to reduce the instability growth—for instance, use of thicker shells to increase ΔR, accelerating the shell for the first half of its implosion phase and letting it coast the rest of the way to cut R in half, shaping the beam pulse power in time, or reducing the steepness of the ablation region, as well as other techniques. Most ICF researchers think that by careful design $N_{m\mathrm{RT}}$ can be kept below 7 or 8; this is an acceptable level provided that pellet surface imperfections are 100 Å or less. Mass producing pellets with close to perfect smoothness is a challenge, but current state of the art is close to achieving this. Understanding hydrodynamic growth in ICF targets and devising techniques to reduce it to acceptable levels are, perhaps, the most challenging topics in current ICF research.

The interface between the hot fuel ignitor core and the surrounding dense cold fuel is also susceptible to Rayleigh–Taylor instability because the hot ignitor core is typically less dense than the cold fuel. A deleterious effect that can result is that some of the cold fuel gets mixed into the hot ignitor region, thereby cooling it below threshold for efficient thermonuclear burn. If this happens then a much bigger and more expensive driver is required to heat the ignitor and the gain of the pellet decreases.

Researchers think that there are high-gain pellet designs that overcome the hydrodynamic instabilities described here. But the uncertainties are large enough that the Rayleigh–Taylor physics underpinnings and pellet designs need thorough testing. The next major ICF experimental facility will explore these ignition-regime physics issues.

4.1.4 Robust Pellet Design. Pellets for an ICF commercial power plant must have gain of order 100. In addition, they must work with driver pulses below 10-MJ energy, and they need to be robust so that they work under less than ideal conditions. Pellets must work one after another, several times each second, millions of times per year even though being injected several meters into position in a reactor chamber, where the reactor environment is harsh, where there is some variance of driver performance from shot to shot, and where alignment may not be perfect.

4.1.5 Reactor-Compatible System. Finally, for power generation, the driver and pellet design need to be compatible with the reactor. As noted earlier, the driver must be efficient ($\geqslant$10%) and be capable of sustaining a high repetition rate (several Hz) for sustained periods between mean failures. Beam entry and focusing needs to be consistent with the reactor geometry and environment. The first wall of the reactor needs to survive millions of microexplosions and have facilities to purge the system of pellet debris. This latter requirement tends to favor the simple minimal-mass direct-drive approach. Last, but not least, whatever system eventually satisfies all the physical requirements must also be economically viable. Pellet capsules have to be mass produced at low per-unit cost and the driver and reactor capital and operational costs must be competitive. The capital cost of a fusion power plant can be higher than other alternatives, but its lower fuel, environmental-impact, and hidden costs must compensate.

Most of the ICF critical issues discussed above are under active investigation. They are being addressed through computer simulation and laboratory experiments. In addition, it is noted that important aspects of the feasibility of high-gain inertial fusion are claimed to have been demonstrated at the Nevada test site in an underground nuclear explosion program called Halite/Centurion.

5. ICF DRIVERS

The driver is intimately connected with the implosion physics requirements. All joules are not created equal. First, the driver must satisfy all the critical pellet-coupling requirements, have adjustable pulse shape, and be efficient enough for a power reactor or be affordable for a defense facility. For energy application, the driver must also be capable of high pulse-repetition rate, be reliable, and be reactor compatible.

Two lasers that potentially satisfy these requirements, at least for the less demanding defense need, are the solid-state neodymium (Nd) glass laser and the krypton-fluoride (KrF) gas excimer laser. Other laser systems were considered in the course of the ICF program but were discarded, primarily because of failure to meet the ICF-pellet physics requirements; their wavelengths were too long.

Particle-beam accelerators are attractive ICF driver contenders because of their high intrinsic efficiency (~30%) and repetitive pulse capability. Two ion-beam approaches are under investigation, i.e., low-energy light ions at very high currents and high-energy heavy ions at low current. In either case, the resulting ion beams need to be focusable to power densities above 10^{14} W/cm^2 and to have suitable stopping ranges. The light-ion approach, also called pulsed power because of the type of accelerator technology involved, is attractive because of its relatively low cost and compactness. But pulse power has major technological hurdles to overcome in focusing its beams to intensities sufficient for ICF and in demonstrating reactor compatibility. The heavy-ion approach, on the other hand, uses extensions of conventional high-energy accelerator technology. The heavy-ion approach is thought by many researchers capable of satisfying all requirements; but there is only a small program developing this option at present.

It should be pointed out that there is no clear consensus as to which, if any, of these drivers is the most suitable for inertial fusion applications. Driver development is a very expensive and long-term enterprise. Technology is advancing rapidly. Breakthroughs occur occasionally that transform a seemingly dead-end approach into a contender; there is no reason not to expect more to come. In addition, all the driver-beam/target interaction issues have not been resolved. Many critical interaction issues need actual fusion reactor conditions to be settled. Some drivers are more compatible with direct drive than indirect drive, and vice versa. Finally, the ultimate economics of drivers is very uncertain; a prohibitively expensive technology today may be inexpensive in the 21st century, when fusion commercialization would occur. For these reasons a multipronged driver approach is favored.

5.1 Lasers

5.1.1 Nd-Glass Lasers. The development of high-power lasers in 1961 was the stimulus that spurred serious contemplation of controlled inertial fusion. The glass laser in particular was the first laboratory device able to concentrate energy to large enough intensities to develop the tremendous pressures needed to implode ICF pellets. A few percent of neodymium ions embedded in optical glass or crystal host, when pumped using flashlamp light, produces strong lasing action near 1 μm (~1060 nm) wavelength, which is in the near infrared. Energy can be stored in the inverted population over a long time interval (360 μs) and then stimulated to lase in less than 1 ns. The large ratio of storage to lasing times allows tremendous power concentration.

Even though the glass laser is presently too inefficient (less than 1%) for an energy reactor driver, it is the driver of choice for most inertial fusion research. This is because the glass laser can develop the needed power densities, is adaptable, and can test a wide range of ideas. Since inception, glass lasers have grown over a million-fold in energy output. Figure 8 shows photographs of three glass lasers used in ICF research. The Pharos III laser, located in the Space Plasma Branch of the Naval Research Laboratory (NRL), is typical of about a half-dozen medium-sized lasers in the world that have been used to elucidate much of the ICF-relevant laser-target interaction physics. The next Nd-glass laser shown is the 24-beam Omega laser located at the Laboratory for Laser Energetics of the University of Rochester that, along with the twelve-beam Gekko XXII facility at Osaka University in Japan, are the principal direct-drive implosion facilities. Finally, the large ten-beam Nova laser located at the Lawrence Livermore National Laboratory (LLNL) is the most powerful laser in the world; Nova primarily tests indirect-drive issues. Nova may soon be upgraded from about 100-kJ pulse level to the 1- to 2-MJ range to explore critical pellet ignition issues around the turn of the century.

Glass lasers used in the ICF program have undergone considerable evolution over the years. They are very flexible in beam pulse length and wavelength capability. Second-, third-, and fourth-harmonic conversion of the fundamental 1-μm wavelength into green, blue, and ultraviolet light with efficiencies of 70%–80% is routinely achieved by passing the light through optically active crystals, such as potassium di-hydrogen phosphate (KDP). It is fortunate that Nd-glass laser beams can operate at short wavelengths because that is what the great preponderance of research data has indicated is required for ICF.

Pressure applied to the pellet surface needs to be ultra-uniform ($\sim$1%), yet even good-quality high-power glass lasers have focal intensity variations more than an order-of-magnitude larger. Laser beam uniformity requirements are somewhat relaxed for the indirect-drive approach, but even so there are deleterious consequences of nonuniform beams. Fortunately, beam uniformity problems are largely solved. The group at Osaka University, Japan, introduced the idea of breaking the beam up into a multitude of beamlets with a random phase distribution. This has the effect of averaging nonuniformities by roughly a factor of $N^{1/2}$, where N is the number of overlapped beamlets in the focal region, and transforming long-wavelength distortions into shorter-wavelength interference patterns. Pressure nonuniformities over small distances tend to be smoothed by thermal diffusion between the absorption and ablation regions of the target. A further significant development was the invention of the induced spatial incoherence (ISI) method at NRL (Lehmberg and Obenschain, 1983). ISI accomplishes simultaneous spatial and temporal smoothing of the beam's focus by breaking up the beam into N beamlets, like the Osaka approach, and introducing a time delay into each beamlet path greater than the coherence time of the laser beam. Thus, interference patterns from all the overlapped beams move rapidly around and average, because of the statistically independent nature of the ISI beamlets. Nonuniformities are not "frozen" into the focal distribution, as would be the case for very coherent lasers. The coherence length of laser light is shortened from meters to less than a millimeter by introducing bandwidth, a slight frequency spread, onto the beam. Beam patterns that are uniform within the 1% level required for direct-drive were demonstrated using ISI with both 1-μm Nd-glass lasers and doubled (0.5-μm) Nd-glass lasers. A side benefit of using a uniform beam is that there are no high-intensity spikes to exceed threshold and to excite problematic parametric instabilities and filamentation, as discussed earlier. Beam-smoothing advances are still being made. The Laboratory for Laser Energetics, at the University of Rochester, has invented a variant of ISI that works with a frequency-tripled Nd-glass beam, which, by itself, does not have enough inherent bandwidth available. NRL is testing an "echelon-free" version of ISI for use in a developmental krypton-fluoride laser, Nike.

The poor (< 1%) efficiency of glass lasers and their high cost (> \$1000/J at present) are not suitable for an energy reactor. Nonetheless, a glass laser is the leading driver candidate for the first defense application facility, where high efficiency is not essential. In addition, there are active programs to develop higher efficiency, less expensive, and high-repetition-rate solid-state lasers. One approach is to use arrays of solid-state laser diodes instead of flash lamps to pump Nd-glass in a

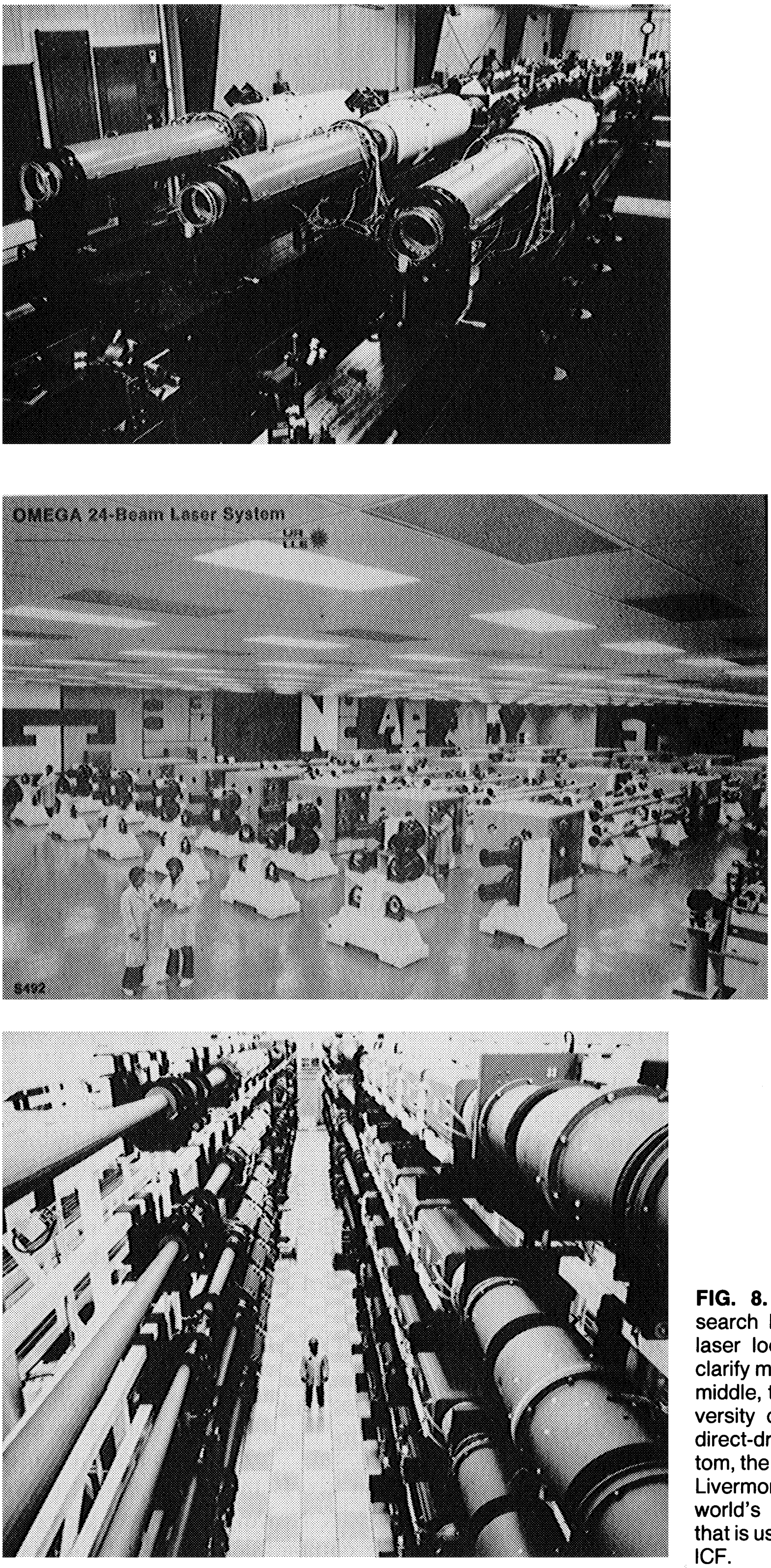

FIG. 8. Three Nd-glass ICF research lasers. Top, the Pharos III laser located at NRL that helped clarify many ICF physics processes; middle, the Omega laser at the University of Rochester used to test direct-drive pellet implosions; bottom, the Nova laser at the Lawrence Livermore National Laboratory, the world's most powerful glass laser, that is used to develop indirect-drive ICF.

narrow, efficient, frequency interval. Diode pumping works on a small scale but it is an open question whether the scheme scales economically to the MJ level. Most research lasers have used the master oscillator/power amplifier (MOPA) architecture. The laser pulse starts in an oscillator and is amplified by many mirrorless amplifiers staged in a long linear chain. New laser configurations are now emerging, such as the LLNL Athena design, which may lead to lower per-joule cost in large systems.

5.1.2 KrF Lasers. The most attractive alternative laser driver to the glass lasers for ICF is the krypton-fluoride (KrF) laser, which lases at about 249-nm wavelength. An excimer is a bound molecule formed with the ground state of a rare-gas atom, such as Kr, and an excited state of another component, fluorine in this case, i.e., KrF*. Excitation is usually accomplished by shooting energetic electron beams through the laser gas. The KrF laser is an ICF candidate because it has both short wavelength, for desirable target coupling, and, perhaps, sufficient (~10%) intrinsic efficiency. Gas lasers can also be designed to operate at high repetition rates and beam-smoothing techniques can be employed. Thus, KrF could satisfy all the requirements for a defense-facility driver and perhaps even an energy-plant driver. However, KrF proponents have many technical obstacles to overcome before it becomes the driver of choice.

KrF laser development is about one decade behind Nd-glass lasers. Development has been slow and fraught with technical problems. The first KrF system designed for ICF was the 2-kJ Aurora laser at the Los Alamos National Laboratory. KrF lasers do not store pump energy over relatively long times and then dump it into a short intensified laser pulse as do glass lasers. Rather, KrF lasers need to be continuously pumped throughout the laser pulse. To get the high intensities and efficiencies needed for ICF, a compression scheme is used to concentrate the energy of a half-microsecond or so pulse into about 10 ns. There are two pulse compression methods, optical multiplexing and Raman compression. Each technique has its own set of advantages and drawbacks. Optical multiplexing, the current method of choice, works by starting with an oscillator pulse with the correct pulse duration, splitting it into multiple beamlets, and delaying each beamlet sequentially in time so that they enter the KrF laser amplifiers in a long train of pulses that match the laser pump duration. After amplification the beamlets are recombined, one on top of another, before focusing onto the target. It remains to be seen if such a complex system, with hundreds of mirrors, can be made robust and efficient enough for practical use.

5.1.3 Laser-Driver Summary. All lasers have a reactor compatibility problem. Each beam has some optical element, a mirror or a lens, exposed to the harsh reactor interior environment. These components must tolerate large pressure impulses, radiation, neutrons, debris, and other reaction products from the pellet explosion. While this is not a big problem for low repetition rates, such as for a defense system, it is of serious concern for energy reactors that operate at 5–10 microexplosions each second, day after day, month after month. This is not necessarily a fatal flaw, because the last optical elements could be placed far away, may be protected with coatings or gas flows, or may be easily replaceable. Drivers in energy plants need to be reliable and put out millions of pulses between mean failures. Laser systems need to undergo significant engineering advances to meet this reliability requirement.

In the near term, however, there is no driver better than the Nd-glass laser for sorting out the myriad of physics issues to demonstrate the scientific feasibility of ICF for either defense or energy application. These physics investigations, carried out in many laboratories throughout the world, can test the viability of pellet designs and uncover and help overcome technical obstacles and determine whether the direct-drive or indirect-drive approach is best. The current strategy is to do just that. The proposed 1–2-MJ Nova-Upgrade Nd-glass laser at LLNL is designed to explore the pellet ignition regime and to test high-gain system designs.

5.2 Ion-Beam Accelerators

Particle beams can drive pellet implosions as long as they satisfy all the implosion physics requirements. Particle beams, like laser drivers, can work in either the direct or indirect mode. However, indirect drive appears more compatible with known ion stop ping distances than direct drive. In many respects, particle beams are preferable to laser beams. Ion energy deposition in targets seems more straight forward and benign than that with lasers. Particle accelerators have high efficiency, typically between 25% and 50% from wall plug to beam. Particle accelerators do not have delicate optics exposed to the reactor. Accelerators are amenable to reliable high-repetition-rate operation. Three particle-beam approaches have been considered for ICF: electron beams, light-ion beams, and heavy-ion beams. The U.S., Russia, and others pursued the electron-beam approach until the 1980s when it was demonstrated that electrons do not couple properly to pellets. Electrons cause excessive target preheat due to their long range. Ion beams are still candidates.

ICF target designs require ion deposition ranges between 40 and 200 mg/cm^2. The deposition mechanism for ions passing through cold material is primarily excitation and ionization of bound electrons the ions, described by the Bethe formula. The range of ions in cold material scales approximately as E^2/Z^2A, where E is the energy, Z is the effective charge, and A is the atomic number of the projectile ion. Thus, low-energy light ions (low atomic number A) and high-energy heavy ions can have similar ranges. Figure 9 lists a comparison of light- and heavy-ion ranges in cold aluminum. For example, a 100-MeV lithium ion ($A=7$) has about the same range as a 10-GeV lead ion ($A=207$). Of course, more low-energy ions are required than high-energy ions for the same deposited energy density in the target. As ICF targets become heated and ionized by the beam, stopping ranges decrease, an advantage for the ICF application.

Two completely different accelerator technologies produce light-ion and heavy-ion beams, viz., pulsed-power devices and high-energy accelerators. The former efficiently accelerates high currents (tens of MA) of light ions to moderate energies (< 100 MeV) and the latter accelerates low currents (tens of kA) of heavy ions to high energy (~ 10 GeV). Each approach has its own strengths and weaknesses.

5.2.1 Heavy-Ion Accelerators. Heavy-ion drivers are the leading contenders for use in ICF energy production plants. The technology is well adapted for high efficiency, high repetition rate, high reliability, and good target energy deposition. Heavy-ion accelerators are extensions of conventional linear accelerators (linacs) used for high-energy physics research (see ACCELERATORS, LINEAR). Ions from a source are injected into a long train of linac acceleration modules. The accelerating electric fields in the linac are applied by using rf fields in cavities or by using induction coils. The former is an rf linac, the latter an induction linac. Both approaches are being pursued, rf in Germany and induction in the U.S. Figure 10 shows schematics of rf and induction heavy-ion accelerators.

Heavy-ion accelerators differ from conventional accelerators in important respects. They accelerate heavy ions ($A > 200$) rather than the protons or electrons usually used in most elementary-particle research accelerators. They need to accelerate higher beam currents in highly compressed pulses to get the very high energy densities needed for ICF. Significant engineering hurdles need to be overcome to build a suitable heavy-ion driver.

There has been progress in developing heavy-ion driver technology. In the U.S., the

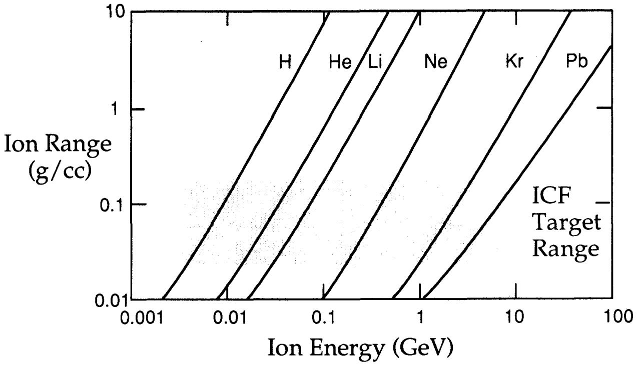

FIG. 9. Ion stopping ranges in solid material as a function of ion kinetic energy. Low-energy light ions and high-energy heavy ions have stopping ranges suitable for imploding ICF pellets.

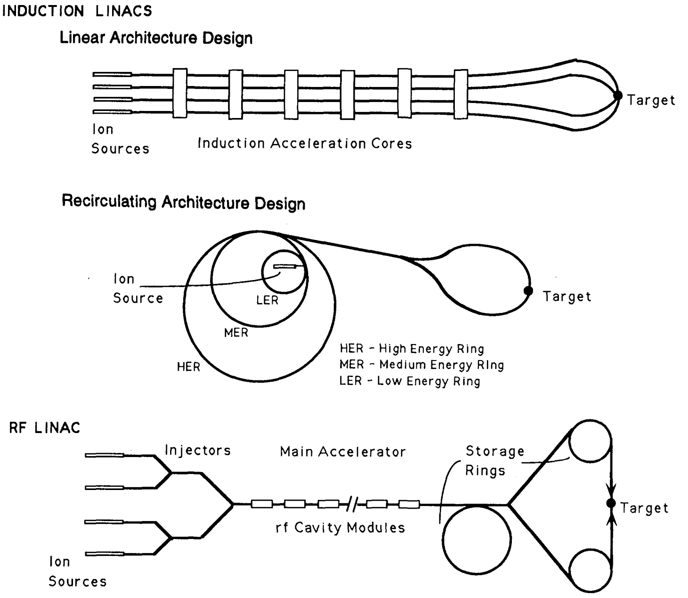

FIG. 10. Schematics of a rf linac (German approach) and two induction (U.S. approach) heavy-ion accelerators for ICF.

Lawrence Berkeley Laboratory (LBL) has built a small test module under its Heavy Ion Fusion Accelerator Research (HIFAR) program. Multiple ion sources, multiple beam transport, and pulse length compression were successfully demonstrated using the HIFAR four-beam accelerator. The next step, currently in its formative phase, is the Induction Linac Systems Experiment (ILSE) project. ILSE is, in many ways, a one-tenth scale model of a full heavy-ion system. ILSE will test beam-combining techniques, space-charge effects, emittance growth, and other beam-transport issues of a full-sized driver. A rf-linac heavy-ion accelerator program called SIS/ESR at GSI, Darmstadt, Germany, intends to study space effects in storage rings, bunching techniques, and ion ranges in hot matter (to study range shortening in plasma). The next step planned for the Darmstadt program is to build a heavy-ion fusion (HIF) demonstration accelerator, first at 1 GeV, 100 kJ, to be upgraded later to 1 MJ energy content per pulse.

Heavy-ion accelerators are not without potential problems. These machines need excellent beam quality (low-emittance beams) so that they can focus to high irradiance onto small ICF capsules in a fusion energy reactor. At the high currents needed, the beam's space-charge fields introduce emittance growth during acceleration that aberrates the beam. Additionally, a disruptive longitudinal resistive instability can develop between the beam and the accelerator structure that limits the maximum current in the induction-linac approach. There is little experience with accelerators in this regime and these issues need to be resolved. Nonetheless, the consensus is that heavy-ion beam drivers are suitable for ICF.

The current strategy is to develop heavy-ion technology in increasingly larger modules at the same time that high-gain ICF implosion physics is tested with glass lasers. There is an implicit assumption, yet untested, that a high-gain system that works with a laser driver will be compatible with a heavy-ion driver. Heavy-ion drivers presently appear best matched to the indirect-drive approach. One reason for this is that it may be difficult to irradiate a pellet uniformly directly with a small number of ion beams. Another reason indirect-drive is more compatible with particle beams is that ion ranges are too long for the most efficient rocket ablation. You may ask why ICF pellet physics cannot be done entirely with heavy-ion drivers instead of lasers. The answer is that it could, but most interaction and implosion experiments require a heavy-ion system capable of depositing the high energy densities needed for ICF. The entry fee for such a system is over 10^9 dollars and is at least a decade away.

5.2.2 Light-Ion Accelerators. Pulsed-power devices accelerate high-current, light-

Table 4. Driver requirements summary for defense and energy applications.

Requirement	Defense	Energy
Driver efficiency	Any (Yield >100 MJ)	$\eta_d G>10$ (e.g., $G>100$, $\eta_d>10\%$)
Pulse repetition rate	Single shot	5–10 Hz
Driver size	Cost limited	$E_d<10$ MJ (reactor limit)
Other	Experimental access	Reliable, economic, reactor compatible,...

ion beams. Their operation is very different from conventional and heavy-ion particle accelerators. Pulsed-power electron accelerators arose in the 1960s, primarily as an intense x-ray source. Pulse power works by arranging a series of electrical energy transfer stages, using capacitors and transmission lines, with successively shorter pulse duration. The temporal pulse compression progressively bunches the stored energy to higher and higher instantaneous powers, i.e., eventually into the multiterawatt regime. For example, a capacitor bank charges up over several minutes in the first stage, the stored energy is switched into a capacitive storage line in several hundred microseconds, and then transferred to a pulse-forming network in about 100 ns, and, finally, bunched and applied across the beam-emitting diode in about 10 ns. In the light-ion mode of operation, the polarity of the diode is arranged so that ions emitted from the anode can leave the machine. Figure 11(a) is a photograph of a research pulse-power machine at NRL called Gamble II, where much pulse-power technology was pioneered. Diode operation is complicated and the preponderance of pulse-power research is devoted to engineering diodes to produce the desired beam. For example, the ideal diode must optimize the ion emission fraction, produce a beam with acceptable emittance, focus the beams to small (pellet) dimensions, produce the ion species desired, and suppress prepulses, among other properties. An advantage of pulse-power accelerators is that they may be relatively inexpensive and compact compared to other drivers.

A large light-ion accelerator designed for the ICF program, the Particle Beam Fusion Accelerator II (PBFA-II) device at Sandia National Laboratories (SNL), is shown in Fig. 11(b). The goal of PBFA II is to deliver a 10-ns pulse of 1 MJ of lithium ions onto a 3-mm radius target at 100 TW/cm^2. PBFA II has cylindrical geometry to cover a pellet surface better. If this can be done, then PBFA II may meet the minimum requirements for ignition and some defense applications. The maximum focused lithium beam power from PBFA-II was about 7 TW/cm^2 in 1990. The light-ion–beam research thrust is to develop diodes that emit much higher quality (more focusable) beams needed to reach the required 100-TW/cm^2 level. Unfortunately, as with heavy ions, it is difficult to do pellet implosions until high intensities are in hand. Nonetheless, relevant experiments have been done with pulse power. Absorption of light ions in cold material and plasma as well as concepts to transport a focused beam along a long (e.g., 2 m) channel have been demonstrated at NRL and Sandia. NRL also pioneered the plasma opening switch and inductive storage developments, which makes possible suppression of prepulse energy and higher power compression, and makes for a more compact driver. Sandia has developed a modular inductive-based accelerator called HERMES and is testing high-repetition-rate accelerators that may ultimately satisfy all the requirements of an energy reactor driver. However, light-ion drivers are far from meeting reactor requirements and considerable development is needed to test their viability.

5.3 Comparison of ICF Drivers

The only driver technology that can currently test the pellet-implosion critical issues of high-gain ICF is the Nd-glass laser. The largest existing glass laser, Nova, which currently delivers about 100 kJ per pulse, may be upgraded in the late 1990s to deliver blue (third-harmonic) pulses in the 1–2-MJ range to test pellet ignition processes. Experiments with such a laser can explore the pellet ignition regime and provide the information needed to design a pellet/driver system for a defense facility. Further, if the ignition experiments go favorably, then most of the information will be in hand to establish the driver requirements and limitations of a high-gain

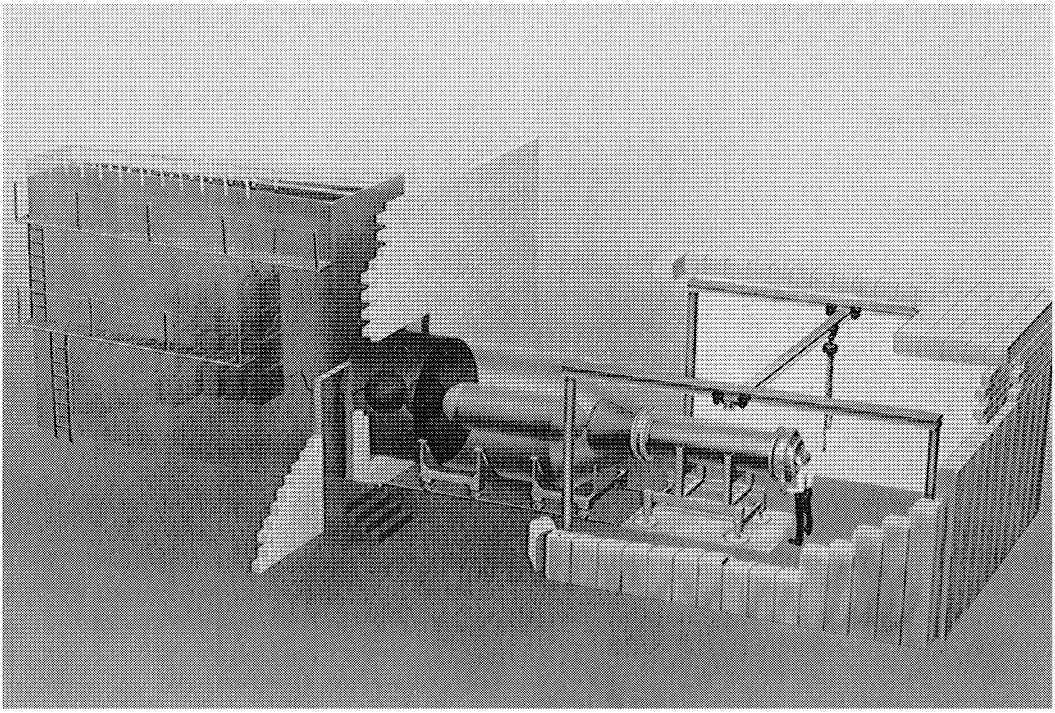

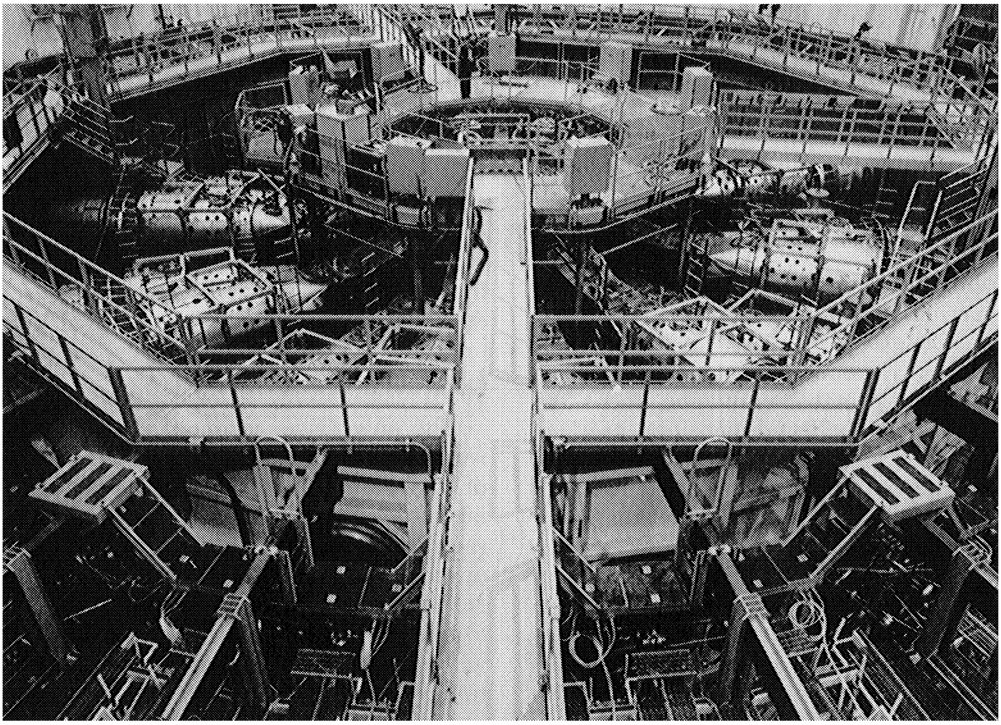

FIG. 11. Two light-ion pulse-power accelerators used in ICF research. (Top) A drawing of the Gamble II pulse-power research accelerator located at NRL. (Bottom) Particle-beam fusion accelerator (PBFA) II located at Sandia National Laboratory that can deliver almost 1 MJ of light-ion beam energy in a single pulse.

pellet system for power production. But, as noted before, it is unlikely that glass lasers are suitable for a power reactor. A comparison of driver requirements for power production and military applications is indicated in Table 4. Of all the driver candidates for energy application, heavy-ion accelerators have the most promise, but are also the least mature technology.

6. ICF POWER REACTORS

An ICF power plant consists of the driver, reactor and associated turbine, generators, and support systems. The reactor is the chamber in which the pellet microexplosions occur and the released fusion energy is converted into the heat used to drive the turbines and generators. ICF reactors will be very different from fission reactors and share the same desirable radiological-hazard and environmental benefits of the magnetic fusion reactors. The reactor will have a first wall to protect and contain the "blanket." The blanket converts the fusion products, primarily neutrons, into heat and provides for breeding of tritium. Breeding is needed to replenish the used-up tritium fuel. A distinct advantage of ICF is that its reactor can be simpler and more innovative than those envisioned for a MFE plant. This is because the driver, the most complex and costly portion of the system, is located outside the reactor area; only the beams enter the reactor itself. This separation of driver from reactor vastly simplifies dealing with such practical concerns as maintenance and reliability.

On the other hand, an ICF reactor has some unique demanding requirements, not the least of which is standing up to several million microexplosions per year (each releasing the energy equivalent of several hundred pounds of TNT) for several years at a time. Protection of the first surface in the reactor and any beam-steering or -focusing elements in the line of sight exposed to the bursts and debris is a challenging engineering feat. Additionally, the beam must propagate through the reactor to the target while maintaining adequate focus and minimal losses. Each driver type has its own unique set of constraints connected with beam transport. Means must be provided to inject target capsules (ICF pellets including any associated *Hohlraum* structures) into proper position in the center of the reactor several times each second. Target capsules have to be insensitive to errors in alignment or beam irradiation parameters. They also must be inexpensive. The reactor and power plant should also have provisions for dealing with the target-capsule debris accumulation. Although each burst contributes a small amount of debris, accumulations over millions of shots can be appreciable, particularly for targets involving *Hohlräume*. Direct-drive–based systems have reduced target debris accumulation since they use only bare pellets, and these can be made up largely from their own cryogenic D-T fuel. On the other hand, direct drive requires extraordinary pellet illumination symmetry of the incoming beams, which can complicate the reactor geometry.

There are not yet any ICF reactor designs with enough engineering detail actually to build a prototype. Several reactor concepts appear compatible with drivers and may be economically viable. Figure 12 shows sketches of several ICF reactor ideas. The first wall protects the thermal blanket and frequently maintains vacuum integrity of the reactor. As such it bears the brunt of the energy, radiation, and debris from each microexplosion. It must have a reasonable service life since it is located deep inside the reactor and is not easily repaired or replaced. The first wall also must efficiently transmit the raw fusion energy, mostly in the form of neutrons, to heat the surrounding thermal blanket and to breed additional tritium fuel. Reactors fall into three types characterized by the kind of first-wall material they incorporate, viz., dry wall, wet wall, and gas filled.

Dry walls are solid material barriers. Low-atomic-number materials are preferred, such as graphite, silicon carbide, etc., to reduce activation levels and radiation damage. Most metals, like steel, are unacceptable since they typically have long-lived activation and would require an enormous reactor diameter (> 100 m) to limit radiation damage to acceptable levels. Some dry-wall concepts are quite innovative. For instance, one idea is deliberately to use a material in which a thin (few-micron) layer of the first wall ablates away on each pulse as it absorbs the brunt of the explosion products; eventually, a fresh wall would have to be installed. Another dry-wall concept called

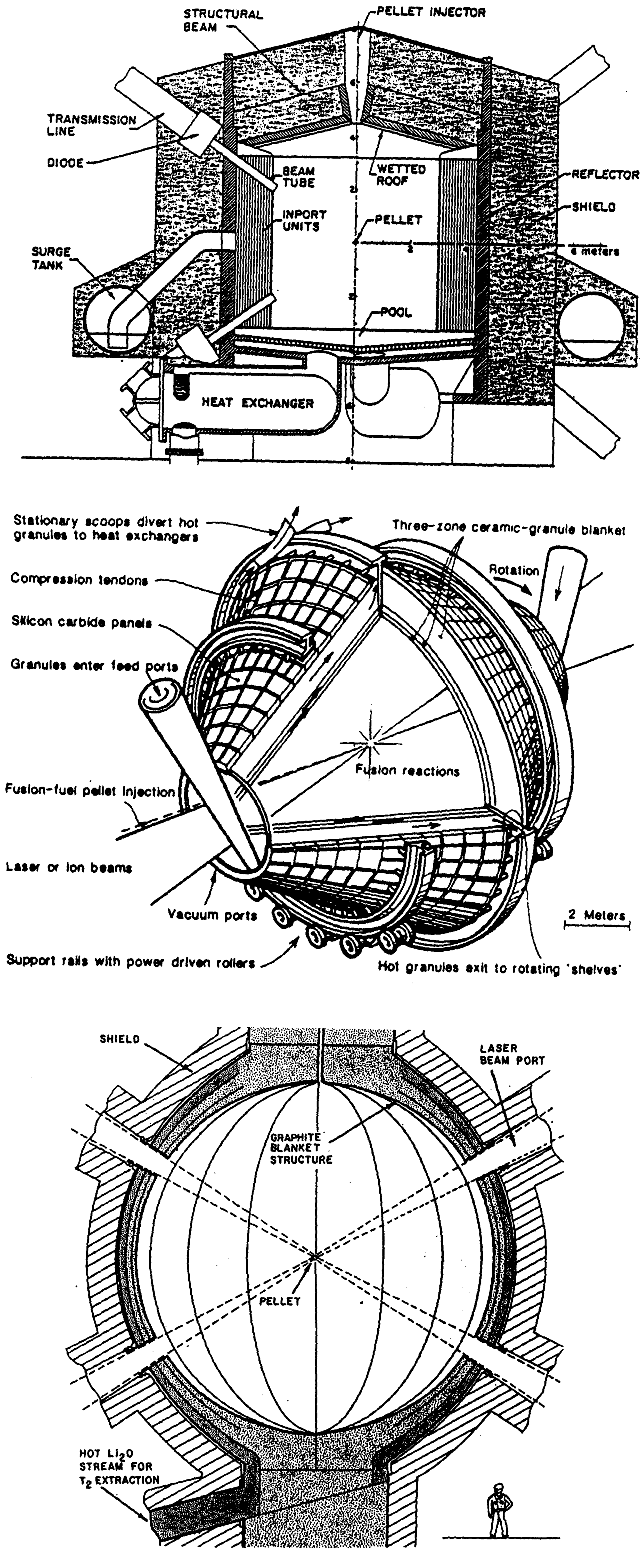
STRUCTURAL BEAM
PELLET INJECTOR
TRANSMISSION LINE
DIODE
BEAM TUBE
WETTED ROOF
REFLECTOR
SHIELD
INPORT UNITS
PELLET
SURGE TANK
6 meters
POOL
HEAT EXCHANGER
Stationary scoops divert hot granules to heat exchangers
Three-zone ceramic-granule blanket
Compression tendons
Rotation
Silicon carbide panels
Granules enter feed ports
Fusion reactions
Fusion-fuel pellet injection
Laser or ion beams
Vacuum ports
2 Meters
Support rails with power driven rollers
Hot granules exit to rotating 'shelves'
SHIELD
LASER BEAM PORT
GRAPHITE BLANKET STRUCTURE
PELLET
HOT Li_2O STREAM FOR T_2 EXTRACTION

FIG. 12. Three ICF power reactor concepts. (Top) Liquid first-wall reactor called LIBRE. (Middle) Dry first-wall reactor consisting of lithium ceramic granules, called CASCADE. (Bottom) Buffer-gas first-wall reactor design called SOLASE.

CASCADE would use a thick flowing layer of small lithium-ceramic granules as both the blanket material and the first wall; heat and contaminants can be removed from the granules outside the reactor chamber.

Wet-wall reactors employ a layer of flowing liquid metal, such as lithium, as the first surface. Structural damage to liquids is self-healing and contaminant extraction occurs outside the reactor cavity. The liquid lithium may be just a wet layer flowing over the surface of a dry wall, or, innovatively, a thick flowing "waterfall," such as in the HYLIFE concept, which acts as a combined blanket and first wall. One property of wet walls that needs further attention is whether their vapors cause interference to the incoming driver beams.

If incoming beams can tolerate ambient pressures in the mTorr range, then deliberate introduction of gas into the reactor vessel can be protective. Gas "first walls," such as incorporated in University of Wisconsin's SOLASE reactor, work by absorbing the prompt x rays and target debris before they reach the first solid surface. This buffers the blanket wall from much of the prompt impact of the explosions. However, SOLASE requires gas pressures in the reactor chamber that can interfere with the transport of some types of incoming beams. Also, a shock wave is set up in the buffer gas that eventually does strike the wall.

It is likely that the ICF reactor that eventually ends up in a power plant will look quite different from present versions since serious development can only take place when a specific driver and pellet implosion is selected, and a prototype plant is within sight.

7. RECENT PROGRESS IN THE ICF PROGRAM

There has been steady progress in the worldwide inertial fusion effort since its inception. The United States has by far the largest ICF programs and driver facilities in the world. The program has been reviewed numerous times since its inception and has consistently been found to be of high quality and has made steady progress on all fronts. Groups in Europe, Japan, Russia, Canada, China, and elsewhere have all made important contributions. Table 5 lists the major ICF experimental facilities in the world as of 1992. There are many smaller experiments and theory efforts worldwide in addition to those listed in Table 5. Most of the physics phenomena that affect the interaction of beam energy with the target and the pellet response are now known and understood well enough to design a large driver facility to test pellet ignition and gain.

Among the non-U.S. efforts, the Japanese program at Osaka University is particularly broad and vigorous. Collaborations among ICF researchers cross many national boundaries, even though the U.S. indirect-drive classification policy through 1992 has been inhibiting.

Road blocks have been met from time to time while trying to resolve inertial-fusion critical issues. But most of these have been overcome through the interplay of good experiments, theory, computer simulation, and inventiveness. Some significant ICF advances include the following:

1. Interaction physics has been largely solved and is no longer the dominant critical issue.
2. Rayleigh–Taylor instability and mix understanding and control has made strides.
3. Effective laser-beam smoothing and balancing methods are invented.
4. Agreement of calculations with experiments is now quite good.
5. Advances have been made in laser and ion-beam driver technologies.
6. Fabrication of high-surface-quality and cryogenic targets has been demonstrated.
7. Low-ignition-threshold indirect-drive concept was invented.
8. Lastly, Halite/Centurion underground nuclear tests have "...excellent performance, putting to rest fundamental questions about basic feasibility to achieve high gain."

Progress is also evident in the march toward satisfying the Lawson criterion, Eq. (1′). The $Tn\tau$ plot shown in Fig. 13 shows a steady advance towards ignition, break-even, and high gain. The next major step to be taken in the program is the demonstration of pellet ignition. Although the indirect-drive approach is the largest component of the ICF program, it is important to note that the smaller programs have made major progress in solving

Table 5. Major ICF research facilities (1992). Laboratory is usually denoted by its acronym: LLNL is Lawrence Livermore National Laboratory; SIOFM is the Shanghai Institute Of Fine Mechanics; CEA is Commissariat à l'Energie Atomique; AWE is Atomic Weapons Establishment; UR-LLE is University of Rochester—Laboratory for Laser Electronics; RAL is Rutherford-Appleton Laboratory; LANL is Los Alamos National Laboratory; SNL is Sandia National Laboratory; LBL is Lawrence Berkeley Laboratory; and GSI is Gesellschaft für Schwerionenforschung. Beam energy and wavelengths or particle types are typical of pulses useful for ICF research. The primary ICF research area is denoted by I-D (indirect drive), D-D (direct drive), Driver dev. (driver technology), and ICF feasibility, which refers to classified experiments done to establish the feasibility of some ICF issues in underground nuclear explosion tests (UGT) in the late 1980's.

Laboratory	Location	Facility	Beam (E_d; λ)	Beams	ICF
Nd-laser					
LLNL	U.S.A.	Nova	80 kJ; 340 nm	10	I-D
Osaka	Japan	Gekko XII	15 kJ; 530 nm	12;sphere	D-D, I-D
CEA	France	Phebus	10 kJ; 1053 nm	2	I-D
AWE	U.K.	Helen	10 kJ; 1053 nm	2	I-D
Kurchatov	U.S.S.R.	Del'fin 2	5 kJ; 1054 nm	64;sphere	D-D
UR-LLE	U.S.A.	Omega	2.5 kJ; 340 nm	24;sphere	D-D
SIOFM	China	Jiquang 12	2 kJ; 1054 nm	6	...
Ecole Poly.	France	LULI	1 kJ; 1053 nm	6	D-D
RAL	U.K.	Vulcan	1 kJ; 530 nm	6	D-D
NRL	U.S.A.	Pharos III	1.5 kJ; 530 nm	3	D-D
KrF-laser					
LANL	U.S.A.	Aurora	2 kJ; 250 nm	5	I-D
NRL	U.S.A.	Nike^{uc}	2 kJ; 250 nm	1	D-D
RAL	U.K.	Sprite	0.1 kJ; 250 nm	1	Driver dev.
Light-ion					
SNL	U.S.A.	PBFA-II	1 MJ; p, Li$^+$	1; cylinder	I-D
NRL	U.S.A.	Gamble-II	100 kJ; p	1	Driver dev.
Heavy-ion					
LBL	U.S.A.	HIFAR	...	4	Driver dev.
GSI	Germany	SIS	...	1	Driver dev.
UGTs					
LLNL	U.S.A.	Centurion	NA	...	ICF feasibility
LANL	U.S.A.	Halite	NA	...	ICF feasibility

the beam symmetry and smoothness requirements and the beam-target interaction issues, which are critical for both direct drive and x-ray drive. The significant improvements in beam quality and balance demonstrated at the University of Rochester and beam-smoothing techniques from NRL have substantially met the requirements for ICF. Some of these beam improvements are incorporated in the LLNL Nova system (called "Precision-Nova"). At this juncture, neither the direct- nor indirect-drive approach is assured of either success or failure. However, the positive claims of the classified Halite/Centurion program together with the laboratory results have been very encouraging.

There is steady progress in driver development, too. LLNL's Nova laser is working well, University of Rochester's Omega laser has demonstrated impressive beam-quality improvement at short wavelength, and Osaka's Gekko XII system has demonstrated good reliability and versatility, as noted above. KrF laser development at NRL includes the incorporation of a form of the induced spatial incoherence beam-smoothing technique. Sandia has made progress meeting its near-term lithium-beam milestones in the PBFA II light ion program. There has also been progress in the HIFAR heavy-ion accelerator program, located at Lawrence Berkeley Laboratory (LBL), as well as the European heavy-ion efforts at Darmstadt, despite low funding levels. At present, the glass laser is the only driver technology advanced enough to perform a laboratory ignition demonstration in the near

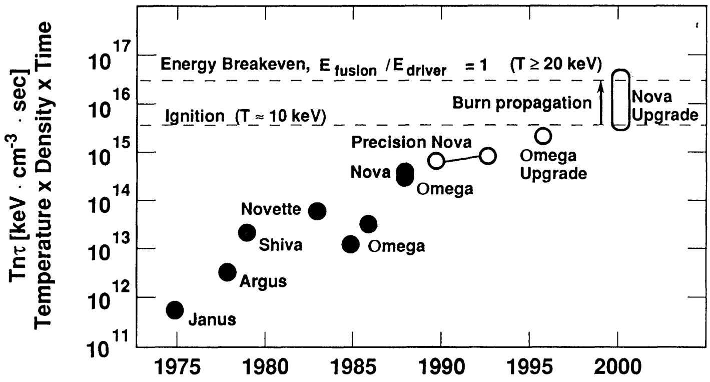

FIG. 13. ICF progress on the Lawson $Tn\tau$ diagram. The temperature–density–confinement-time product has progressively increased over the years to within a factor of ten of scientific feasibility. The next generation research facility, Nova-Upgrade, should achieve ICF ignition. (From the FPAC Final Report).

term. Barring a technology breakthrough, glass lasers are ruled out as an energy driver because of their low efficiency (< 1%). However, other driver technologies have the potential of yielding an inertial energy driver. No driver yet meets the repeatability, efficiency, and reliability requirements and high-gain target compatibility ultimately needed.

8. PROGNOSIS AND CONCLUSION

There were two major reviews of the U.S. fusion program in 1990. The U.S. Department of Energy formed the Fusion Policy Advisory Committee (FPAC) to recommend a broad plan to bring fusion energy to commercialization for both the magnetic and inertial fusion programs. Europe has decided not to establish a major ICF program at this time, but rather is watching the U.S. effort. A schematic of the FPAC plan for inertial fusion development is shown in Fig. 14. The U.S. National Academy of Sciences (NAS) also completed an in-depth review of the inertial fusion program, primarily for defense application. These studies noted the immense difficulty and long lead times involved in fusion research and recommended that research efforts be immediately accelerated so that commercialization could begin in the early 21st century. Both reviews concluded that inertial-fusion defense applications were important and likely to be successful. They also concluded that prospects for the more stringent energy-production goal were encouraging. In particular, FPAC recommended that both the ICF and MFE programs proceed with experiments to explore burning fusion conditions, so-called "ignition" facilities. These fusion facilities would form the basis required to build an ICF engineering test reactor. An ignition facility would also lead to the Laboratory Microfusion Facility (LMF) for defense applications.

The next major step in the ICF program will be the construction of an ignition facility. The ignition facility envisioned consists of a major upgrade of the present 10-beam, 100-kJ Nova facility at LLNL to a 20-beam Nd-glass laser that delivers 1–2 MJ of third-harmonic (blue) light pulses onto ICF targets. The plan calls for this "Nova-Upgrade" to reach the conditions necessary to explore pellet ignition in about the year 2000. The Japanese program at Osaka is also discussing the possibility of a large facility of its own. If these ignition-level experiments are successful, then fusion researchers would have most of the information

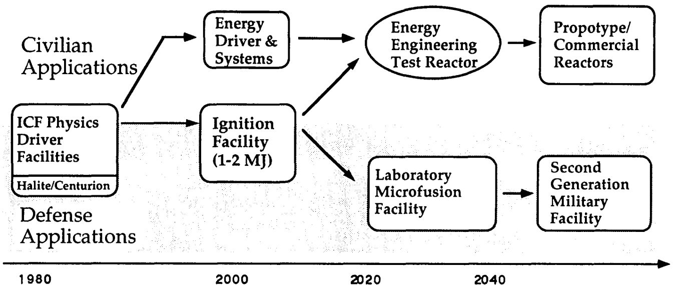
Civilian Applications
Energy Driver & Systems
Energy Engineering Test Reactor
Propotype/ Commercial Reactors
ICF Physics Driver Facilities
Halite/Centurion
Ignition Facility (1-2 MJ)
Laboratory Microfusion Facility
Second Generation Military Facility
Defense Applications
1980
2000
2020
2040

FIG. 14. The U.S. Department of Energy plan for ICF development. Energy and defense applications have a common development path until high gain is shown to be feasible. Development of a high-efficiency driver, such as heavy-ion accelerators, proceeds in parallel for the energy application. (From the FPAC Final Report).

needed to build a high-yield ICF implosion facility. As noted earlier, commercialization of ICF for energy production requires more than successful high-yield pellet implosions. Energy production also requires a highly reliable repetitively pulsed system, driver and reactor, with sufficient efficiency and economic viability to be competitive. FPAC urged the Department of Energy to undertake a program specifically to examine prospects for achieving these additional requirements and to pursue the most attractive path. For this reason, the development of a heavy-ion driver has high priority.

The next decade will be an exciting period in ICF development. If all goes well, the feasibility of high-yield pellet explosions will be verified, a selection of a driver approach compatible with power reactors made, and an engineering test facility initiated leading to commercialization in the next century. The development of fusion for power production is a formidable task, taking much more time, resources, and ingenuity than initially envisioned. Commercialization is far from being assured, even now after three decades of development. But, if it is ultimately successful, then the payoff economically, environmentally, and geopolitically will be immense.

GLOSSARY

Ablation: The process of removing material from the surface of a pellet as a result of intense radiation from the driver beam.

Aspect Ratio: Ratio of the radius of a pellet to the thickness of its shell wall.

Break-Even: The point at which the amount of fusion energy produced is equal to the energy used in a fusion device.

Burn Front: A self-sustaining thermonuclear fusion reaction region that propagates outward throughout the remaining fuel.

Capsule: The target of ICF driver beams. In the direct-drive approach the capsule is the pellet. In the indirect approach the capsule is the pellet-*Hohlraum* combination.

Critical Density: The density at which the incoming light frequency equals the local plasmas frequency.

Deuterium (D): The heavy isotope of hydrogen with one neutron; it is a component of fusion fuel that is abundantly found in natural water.

Direct-Drive Approach: The ICF method in which the driver beams directly irradiate the pellet surface.

Driver: The device that produces the intense beams that are focused on the target to implode the pellet. Drivers include lasers and particle beams.

Fusion Policy Advisory Committee (FPAC): The committee set up in 1990 by the U.S. Department of Energy that laid out magnetic and inertial fusion energy development plans through commercialization.

Gain: The ratio of the fusion yield of an ICF pellet to the input driver energy.

Heavy-Ion Accelerator: Instrument for accelerating high–atomic-number particle beams for ICF.

HIFAR: Heavy-ion fusion accelerator research program located at Lawrence Berkeley Laboratory which is exploring heavy-ion driver technology.

Hohlraum: An enclosure in which driver-beam energy is converted into x rays; these x rays are contained in the *Hohlraum* and irradiate the pellet surface.

Hot-Spot Ignition: Ignition of small amount of thermonuclear fuel in the center of the bulk of the cold compressed fuel.

Hydrogen Bomb: Also called H bomb, an explosive that derives its energy release from thermonuclear fusion.

Ignition: The process in which the fusion energy released supplies enough heat to propagate a self-sustaining fusion burn front into the surrounding thermonuclear fuel.

Indirect-Drive Approach: Also known as the "classified approach," the ICF method in which the beam energy is first converted into x rays in a *Hohlraum*; the x rays then irradiate and implode the pellet.

Induced Spatial Incoherence (ISI): A method to produce an ultrasmooth laser beam profile.

Inertial Confinement Fusion (ICF): Also frequently referred to as inertial fusion, inertial-fusion energy (IFE), or laser fusion, a means of producing bursts of fusion energy by imploding small pellets containing thermonuclear fuel. IFE refers specifically to the civilian energy-production application.

Isentropic Compression: Heating of a material by compressive work.

Lawson Number: The minimum product of density and confinement time ($\sim 10^{20}$ s m^3) of a fusion fuel plasma needed to produce net energy in a fusion reactor.

Magnetic Fusion Energy (MFE): An approach toward controlled thermonuclear fusion in which low-density thermonuclear fuel is contained and heated to thermonuclear temperatures in large magnetic "bottles" for long periods of time.

Nova: The largest neodymium-glass laser in the world, located at Lawrence Livermore National Laboratory; it is used primarily for indirect-drive ICF research. Upgrades to Nova include "Precision Nova," which improved the laser's beam uniformity and balance, and the proposed "Nova Upgrade" which, if approved, will increase the current 100-kJ beam energy up to 1–2 MJ.

Omega: A large neodymium-glass laser located at the University of Rochester that is used for direct-drive ICF research.

Parametric Instabilities: Parametric processes involve the interaction of a large-amplitude wave, often called the pump, with two or more other waves in the system. These processes are often unstable, grow to large amplitude, and cause interesting, but deleterious, effects for ICF applications.

PBFA II: A large light-ion accelerator at Sandia National Laboratory being tested as an ICF driver.

Pellet: A spherical shell containing the fusion fuel in the ICF approach.

Plasma: Ionized matter that responds to electric and magnetic forces as well as mechanical forces. Plasma is often called the fourth state of matter.

Plasma Frequency: A natural oscillation frequency of electron waves in a plasma.

Ponderomotive Force: A force caused by intense electromagnetic wave gradients that is important in many laser plasma interaction processes.

Preheat: Excess heating of the fuel before final compression.

Pulse Power: Type of driver used to accelerate intense light-ion (low atomic number) beams for ICF.

Rayleigh–Taylor Instability: An instability caused when a dense fluid is accelerated by a less dense fluid. This instability occurs during the pellet implosion and can break up the pellet before maximum compression.

Thermonuclear Fuel: Low-atomic-number nuclei that can undergo fusion reaction and give off net energy. Deuterium and tritium are the thermonuclear fuels most considered for controlled fusion applications.

Tritium (T): The unstable heavy isotope of hydrogen with two neutrons; it is a fusion fuel component that would be bred in the blanket of a fusion reactor to replenish supplies.

Underground Tests (UGT): Experiments performed in underground nuclear detonation tests at the DOE nuclear test site in Nevada.

List of Works Cited

Fusion Policy Advisory Committee (FPAC) Final Report (1990), DOE/S-0081; available from the National Technical Information Service, U. S. Department of Commerce, Springfield, VA 22161.

Hubbard, H. M. (1991), *Sci. Am.* **264** (4), 36–42.

Lehmberg, R. H., Obenschain, S. P. (1983), *Opt. Commun.* **46**, 27–31.

NAS Review of the Department of Energy's Inertial Confinement Fusion Program Final Report (1990), National Academy Press; available from the Naval Studies Board, National Research Council, 2101 Constitution Avenue, Washington, DC 20418.

Nuckolls, J., Wood, L., Thiessen, A., Zimmerman, G. (1972), *Nature* **239**, 139—142.

Further Reading

Duderstadt, J. J., Moses, G. A. (1982), *Inertial Confinement Fusion*, New York: Wiley. A text book on inertial confinement fusion.

Johnson, T. (1984), "Inertial Confinement Fusion: Review and Perspective," *Proc. IEEE* **72**, 548–594. A well-written perspective on the status of inertial confinement research.

Kruer, W. L. (1988), *The Physics of Laser Plasma Interactions*, Redwood City, CA: Addison Wesley.

Ripin, B. H., Kruer, W. L. (1986) in: *Physics Through the 1990s: Plasma and Fluids*, Washington, DC: National Academy Press, 221–242. Part of a comprehensive survey of the frontiers of physics and recommendations for research emphasis into the 1990s.

Schwartz, H. J., Hora, H., Eds. (1971–1980), *Laser Interaction and Related Plasma Phenomena*, Vols. 1–5, New York: Plenum; Hora, H., Miley, G. H., Eds. (1984–1991), *ibid.*, Vols. 6–9. A series published every two years containing progress in inertial controlled fusion and laser plasma interactions.

FUSION, MAGNETIC CONFINEMENT

H. L. BERK, *Institute for Fusion Studies, The University of Texas at Austin, Austin, Texas, U.S.A.*

Introduction 576
1. **Energy from the Fusion Reaction** 576
1.1 Fusion Reactions 576
1.2 Fuel Availability 577
1.3 Magnetic Confinement 578
2. **Intrinsic Energy-Loss Mechanisms** 578
2.1 Lawson Criterion 578
2.2 Loss Mechanisms 579
2.2.1 Bremsstrahlung 579
2.2.2 Synchrotron Radiation 579
2.2.3 Diffusion 579
2.2.4 Turbulence Losses 580
3. **Magnetically Confined Equilibrium** 580
3.1 Magnetohydrodynamic Equilibrium Equations 580
3.2 Two-Dimensional Pinch Solutions 581
3.3 End Stoppering 582
3.4 Toroidal Configurations 582
3.4.1 General Considerations 582
3.4.2 Vertical Field 583
3.4.3 Compact Toroids 583
3.4.4 Inductive Voltage Drive 585
3.4.5 Stellarators 586
4. **Particle Orbits and Classical Transport in Magnetically Confined Plasmas** 588
4.1 Particle Orbits in Magnetic and Electric Fields 588
4.2 Spatially Varying Fields 589
4.2.1 *Q* Value in a Mirror Machine 590
4.2.2 Electrostatic Potentials in Tandem Mirrors 590
4.3 Orbits in a Toroidal Field 591
4.4 Neoclassical Diffusion 592
4.4.1 Bootstrap Current 593
4.4.2 Ware Pinch 593
4.5 Radial Electric Field 594
4.6 Orbits in Stellarators 594
5. **Plasma Instability and Anomalous Transport** 595
5.1 Ideal MHD Theory 595
5.1.1 Simplified Dynamical Model 595
5.1.2 Generalization to Plasmas 596
5.1.3 Role of Field-Line Curvature 596
5.2 Deviation from MHD Stability Guidelines 596
5.2.1 Passive Current Carriers 596
5.2.2 Finite-Larmor-Radius Stabilization 597
5.2.3 Migma Concept 597
5.3 Stability in Tori 597
5.3.1 Line-Average Stabilization 597
5.3.2 Beta Stability Limit 597
5.3.3 RFP Stability 597
5.4 Resistive Stability 598
5.5 Microinstability 598
5.5.1 Velocity-Space Instability 599
5.5.2 Drift-Wave Instability 599
5.6 Turbulent Diffusion 599
6. **Status of Magnetic Fusion Research** 600
6.1 Historical Overview 600
6.2 Status of Non-Tokamak Concepts 600
6.2.1 Mirror Machines 600
6.2.2 Reversed-Field Pinch 601
6.2.3 Compact Toroids 601
6.2.4 Stellarators 602
6.3 Status of Tokamak Research 602
6.3.1 Impurity Control and Fueling 602
6.3.2 Auxiliary Heating 603
6.3.3 Plasma Shaping 603
6.3.4 Current Drive 603
6.3.5 Achieved Confinement Parameters 604
6.4 Conclusion 604
Glossary 605
Works Cited 607
Further Reading 607

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

Nuclear fusion is the combining of light nuclei to form more massive ones with the conversion of nuclear binding energy to kinetic energy. This energy conversion process is the source of heat in the stars and in the Sun. If this reaction can be replicated on Earth in a controlled manner, there would exist an energy source to supply mankind in perpetuity.

In order to achieve the fusion reaction, charged nuclei have to approach to within distances of 10^{-14} m of each other for nuclear transitions to occur. However, this close approach is impeded by the Coulomb repulsive force of the positively charged nuclei. Only if particles have enough kinetic energy, so that they can penetrate this Coulomb barrier, can the fusion reaction take place. Hence, to achieve fusion it is necessary to heat the fuel to a sufficiently high temperature that particles can penetrate close to each other's nuclear core. Then a nuclear fusion environment can be created where the kinetic energy released by the fusion reaction products maintains the temperature of the fuel.

The fusion environment is at a temperature even greater than that found in the Sun. In order to achieve this environment technologically, special conditions are required to establish insulation of the fuel from the surrounding "cold" Earth temperature. Otherwise the energy losses to the neighboring environment will exceed the energy produced by the fusion reaction. The task of producing adequate heat confinement is the crucial aspect of the controlled-fusion program. As will be described here, great progress has occurred since the program's initiation in the 1950s, yet much more is needed in order to achieve one of the great challenges to mankind: Earth-based energy from controlled fusion.

1. ENERGY FROM THE FUSION REACTION

Only a few fusion reactions have sufficiently high rates and energy gain to be exploitable in a practical fusion reactor. As a result the issue of what fuels are available for use on Earth when controlled fusion is feasible is significant and is discussed in this section.

1.1 Fusion Reactions

The three most important fusion reactions for energy applications are

$$d+d\rightarrow\begin{cases}{}^3\mathrm{He}+n+3.3\ \mathrm{MeV}\\ t+p+4.0\ \mathrm{MeV}\end{cases},$$

$$d+t\rightarrow{}^4\mathrm{He}+n+17.6\ \mathrm{MeV},$$

$$d+{}^3\mathrm{He}\rightarrow{}^4\mathrm{He}+p+18.3\ \mathrm{MeV}.$$

Figure 1 shows the cross sections and reaction rates of these three reactions. These curves show that the energy dependence of the rate of fusion reactions decreases rapidly below the energy of maximum reactivity, a manifestation that the Coulomb barrier can quench the fusion reaction. A fusion environment must achieve temperatures in excess of the ideal ignition temperature, T_{ideal}, which is the lowest temperature for which a self-sustaining fusion reaction can be achieved when energy losses due only to bremsstrahlung radiation are accounted for. For the *d-t* reaction T_{ideal} is 5×10^7 K, or, in units more commonly used by research workers, 4 kiloelectronvolts (keV). For the *d-d* and *d*-^{3}He

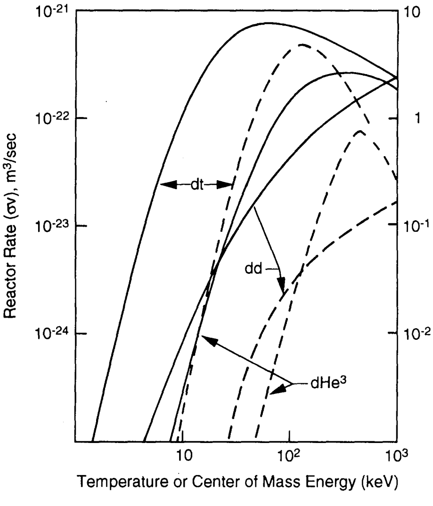

FIG. 1. Solid curves: reaction rates averaged over a Maxwellian distribution as functions of temperature (in keV units); dashed curves: cross sections as functions of center-of-mass energy for the *d-d*, *d-t*, and *d*-^{3}He fusion reactions.

reactions the ideal temperatures are 40 and 30 keV, respectively.

1.2 Fuel Availability

One of the principal motivations for using controlled fusion for energy is the availability of fuel reserves. Among the reactions listed in Fig. 1, the *d-d* reaction clearly is a fuel resource with virtually inexhaustible supply. Deuterium is a stable isotope of hydrogen, which is stored in sea water with a natural abundance of 1 part in 6500. The energy content of 1 L of water from the *d-d* reaction (2×10^9 joules) is equivalent to the energy content of about 50 L of oil. The effective energy content can even be raised substantially if the tritium and ^{3}He products produce further fusion reactions. With deuterium extracted from sea water, if energy is consumed at a rate of 10^{21} J/y, the *d-d* reaction can provide an energy source for $\sim 10^9$ y.

The main drawback of the *d-d* reaction is a relatively low reaction rate and yield, which makes its implementation more difficult as compared to the *d-t* and *d*-^{3}He reactions. Of the two, *d-t* has a clear advantage in achieving a fusion environment, as the ignition temperature is appreciably lower and reaction rates are higher than for the *d*-^{3}He reaction. For this reason the first planned demonstration of controlled fusion conditions is based on the *d-t* reaction.

As a long-term energy source, the *d-t* reaction is feasible, although not as directly as with the *d-d* reaction. Tritium is an unstable element with a half-life of 12 years. Hence tritium must be constantly bred during reactor operation. Breeding is possible via the nuclear reactions

$$n+{}^6\mathrm{Li}\rightarrow t+{}^4\mathrm{He}+4.8\ \mathrm{MeV},$$
$$n+{}^7\mathrm{Li}\rightarrow t+{}^4\mathrm{He}+n-2.5\ \mathrm{MeV}.$$

Both reactions can be used to produce tritium. The *n*-^{6}Li reaction is exothermic, and if the released energy is exploited, the effective yield per *d-t* reaction would be 22 MeV. Though the ^{7}Li reaction is endothermic, it does not appreciably degrade the energy output of the entire *d-t* cycle. ^{6}Li has a natural abundance of 7.5% while ^{7}Li has a natural abundance of 92.5%. Thus as tritium will ultimately be bred from lithium, it is the availability of lithium that governs the energy potential of the *d-t* reaction. The officially estimated "economically" extractable lithium reserves are equivalent to about one thousand years of energy supply. In addition the lithium found in sea water, if economically extractable, would imply an energy reserve on the scale of millions of years.

Despite its long-term potential, the implementation of fusion energy based on *d-t* has several difficulties. The need to breed tritium complicates reactor design; and, since it is radioactive, there will be siting restrictions. There is controversy about whether the hazards associated with the long-lived radioactive material generated from the neutrons produced by the *d-t* reaction are a significant environmental concern (though it is generally agreed that such problems are far less severe than with fission reactors). Perhaps the most significant problem is whether the long-term integrity of containment vessels can be maintained under the continuous flux of 14-MeV neutrons (advanced materials research is needed to demonstrate that walls can be fabricated from material that anneals with minimum deterioration from defects caused by neutron bombardment).

These concerns give impetus to the desire to exploit the *d*-^{3}He fusion reaction, which by itself produces only charged fusion products (though some neutrons are produced from the *d-d* side reaction). Another potential benefit of the *d*-^{3}He reaction is the possibility of directly converting charged fusion products to electrical energy with an efficiency of $\sim 80\%$, whereas energy derived from energetic neutrons needs first to be converted to heat, so that the ultimate conversion to electricity is limited to a thermodynamic energy efficiency of $\lesssim 40\%$.

Unfortunately, there is a crucial uncertainty in the *d*-^{3}He cycle, that of fuel availability. Though ^{3}He is a stable element, it is light enough to escape the Earth's gravitational field. Hence only trace amounts of ^{3}He are found on Earth. Paradoxically, ^{3}He is present on the moon (Wittenberg *et al.*, 1986). The ^{3}He, which originates in the flux of the solar wind, is embedded in lunar rock. It has been estimated that this ^{3}He source can supply mankind with energy for hundreds of years. On a longer time scale, one can envision space travel to the larger planets, Jupiter and Saturn, where the gravitational field is strong enough to retain ^{3}He. An alternative scenario is that Earth-based sources of ^{3}He can come from other fusion reactors operating on *d-d* or *d-t* reactions. Neutron-producing reactors could be sited in remote areas and supply ^{3}He for electricity-producing ^{3}He reactors close to populated areas.

The above discussion indicates that with appropriate technological progress, the fusion reaction from several fuels can provide civilization with a long-term energy source. The choice of fuel cycle will depend on the quality of the energy confinement scheme that is devised, the environmental standards that are deemed acceptable, and the technologies that are developed in the future.

1.3 Magnetic Confinement

Still, the first major step remains to be proved; i.e., can a controlled-fusion Earth-based environment be established? There has been considerable progress toward answering this question, but much more information is needed from future research. This article is devoted to explaining the principles of magnetic confinement, where an ionized gas (commonly called a plasma) is isolated from the surrounding "cold" environment by magnetic pressure. For an alternative approach, see FUSION, INERTIAL CONFINEMENT.

In the subsequent sections we present the following material: In Sec. 2 we discuss the confinement parameters that must be achieved if the energy produced from the fusion reaction is to exceed the energy required to produce the hot plasma conditions necessary for copious fusion reactions to occur. In addition we will present an overview of the principal energy-loss mechanisms associated with a confined hot plasma. In Sec. 3 we discuss magnetically confined equilibrium and give examples of several confinement schemes. In Sec. 4 we examine in more detail the mechanisms for particle confinement and transport for two of the most important confinement concepts, the mirror machine and the tokamak. The viability of a plasma confinement concept usually depends on the stability of the configuration. An overview of stability considerations is given in Sec. 5. Finally, in Sec. 6, past progress and the present state of magnetic confinement research is discussed.

2. INTRINSIC ENERGY-LOSS MECHANISMS

2.1 Lawson Criterion

All fusion confinement schemes rely on heating a plasma in a configuration that is sufficiently well insulated from its surroundings. The Lawson criterion, the product of the particle density n and energy lifetime τ arising from all possible loss mechanisms, is a measure of how close a given configuration is to creating a fusion environment. It is obtained by observing that a plasma (with, say, equal numbers of electrons and hydrogenic ions) with a temperature T (throughout Sec. 2 we assume that electron and ion temperatures are equal), electron density n, and lifetime τ loses power density at a rate $P_l = 3nkT/\tau$. In order to have a self-sustaining fusion reaction, this power loss must be compensated by the energy produced by the fusion reaction. It is envisioned that charged fusion products will be contained in the system with their kinetic energy being absorbed by the background plasma, while fusion-produced neutrons escape from the plasma region. A self-sustaining reaction (a burn) occurs when the power production of the charged-particle fusion products compensates the thermal energy losses of the plasma. Thus the criterion for a fusion burn is

$$3nkT/\tau \lesssim \gamma n^2 \langle \sigma v \rangle W_F \,, \qquad (1)$$

where $\langle \sigma v \rangle$ is the fusion reaction rate averaged over the background ion distribution functions (σ is the fusion cross section, v $\equiv$ relative ion velocities), γ is a numerical factor that depends on the fuel mixture (for *d-d*, $\gamma = \frac{1}{2}$, for *d-t*, $\gamma = n_d n_t / n^2 \leqslant \frac{1}{4}$, etc.; the subscripts d and t refer to deuterium and tritium), and W_F is the energy of the charged fusion products. From Eq. (1) the Lawson criterion follows:

$$n\tau > \frac{3kT}{\gamma \langle \sigma v \rangle W_F} \equiv (n\tau)_{\mathrm{L}} \,, \qquad (2)$$

where the right-hand side depends on temperature and fuel mix. Figures of $n\tau$ versus temperature, for the optimum fuel mix, are given in Fig. 2. One observes that $n\tau$ values in excess of 10^{20} m^{-3} s are needed to achieve self-sustaining fusion.

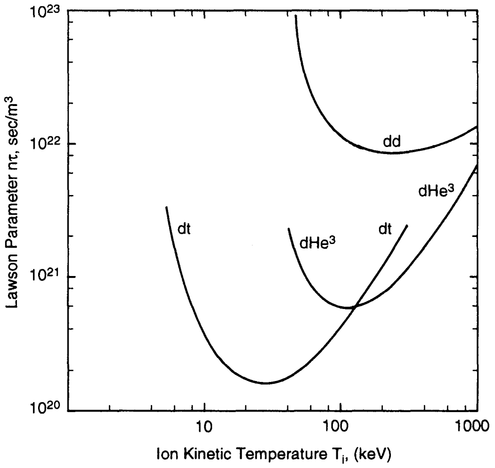

FIG. 2. Lawson diagram for sustaining fusion burn for the *d-d*, *d-t*, and *d*-^{3}He reactions.

A less stringent, modified Lawson criterion applies to the condition for obtaining net energy when energetic neutrons are produced. For the *d-t* fuel cycle with ^{6}Li used for the breeding of tritium, 22 MeV per reaction is released, with only 3.5 MeV going into charged products inside the plasma. With a 40% heat conversion efficiency, the critical $n\tau$ that is needed to achieve fusion energy is reduced by a factor of 3. However, external power must then be continuously supplied to this so-called "wet-wood burner" mode of operation.

2.2 Loss Mechanisms

The feasibility of attaining fusion energy depends on devising thermal insulation so that Eq. (2) can be satisfied. There are several intrinsic power-loss mechanisms that place upper bounds on $n\tau$, including radiation, diffusive losses due to collisions, and diffusion due to small-scale turbulence. The composite loss rate can be written as $1/\tau = \Sigma_i 1/\tau_i$, with τ_i^{-1} the effective energy-loss rate per electron-ion pair for each individual mechanism.

2.2.1 Bremsstrahlung. Radiation due to bremsstrahlung is a basic loss mechanism and is due to the radiation emitted by electrons colliding with each other and with ions. The power radiated P_{brem} is given by

$$P_{\text{brem}} = 3nkT/\tau_{\text{brem}} \tag{3}$$

with $n\tau_{\text{brem}} = 2.8\times10^{21} T_{10k}^{1/2}/Z_{\text{eff}}$ m^{-3} s, where T_{10k} is the plasma temperature in units of 10 keV, and $Z_{\text{eff}} = \Sigma_i Z_i^2 n_i/n_e$ with Z_i the atomic number of species i. The subscript e refers to electrons (we note that the $n\tau$ curves in Fig. 2 take into account the bremsstrahlung power loss as well as the fusion power gain). As bremsstrahlung is not significantly reabsorbed by the plasma in configurations appropriate to controlled fusion, the losses are not readily "recycled" back into the system (except perhaps with some thermodynamic efficiency ~35%–40%). Impurities sputtered from the wall can accumulate in plasmas and can cause significant increase in the bremsstrahlung loss rates even if the hydrogen component is not significantly diminished. Hence the suppression of impurity accumulation is critical in plasma containment. Because of bremsstrahlung losses many possible neutron-free fusion reactions cannot be used for fusion power production. Examples include $p + {}^{11}B \rightarrow 3\alpha + 8.7$ MeV, $p + {}^6\text{Li} \rightarrow {}^4\text{He} + {}^3\text{He} + 4$ MeV, etc.

2.2.2 Synchrotron Radiation. The other important radiation process with significant energy loss is synchrotron radiation. This is single-particle radiated power due to cyclotron motion of electrons in a magnetic field, which is given by

$$P_{\text{sync}} = 3nT/\tau_{\text{sync}} \tag{4}$$

with

$$n\tau_{\text{sync}} = 1.9\times10^{21}\beta/T_{10k} \text{ m}^{-3}\text{ s}$$

with $\beta = 4\mu_0 kT/B^2$ the ratio of kinetic pressure to magnetic pressure, a fundamental plasma parameter in magnetic fusion. If the radiated power were to propagate without further absorption, this parasitic loss mechanism would severely affect the viability of nearly every magnetic fusion concept. Fortunately, most of the radiation is reabsorbed by the plasma, and even escaping radiation can be reflected by external conducting walls and reabsorbed by the plasma. Calculations show that synchrotron losses do not constitute a constraining loss mechanism in a *d-t* reactor, whereas they do constrain *d-d* or *d*-^{3}He reactors to operate at moderately high β values (typically $\beta \gtrsim 15\%$).

2.2.3 Diffusion. The other important mechanism that determines plasma lifetime is particle diffusion. In subsequent sections we shall describe magnetic containment configurations in which the particle orbits are isolated from the surrounding walls. However, binary collisions and fluctuating fields cause particles to deviate from their orbits, and eventually these particles diffuse to the walls. Energy losses due to binary collisions can be reliably calculated. Such losses, together with radiation, establish the intrinsic containment properties of a given configuration. In toroidally shaped devices collisional losses are generally compatible with reactor operation. In open-ended devices, these losses are a significant constraint for power production.

2.2.4 Turbulence Losses. Losses due to fluctuations from instabilities appear to govern the lifetime of almost every magnetic fusion experiment that has been developed. Roughly, these losses can be categorized in two forms: those due to macroinstability and those due to microinstability. In macroinstability, the unstable mode affects the entire plasma column and the loss of integrity of the confinement scheme. In such cases the losses are usually incompatible with achieving energy production. However, in some cases macroinstability is localized to the plasma interior, and the integrity of the thermal insulation can still be maintained.

In microinstability the instability structure is on a scale small compared to the plasma size. These fluctuations cause stochastic motion of particles, much like that produced by collisional diffusion. Typically it is this diffusive process that appears to determine thermal containment time of even the best-contained plasmas. The detailed mechanisms for such losses are not fully understood, and the theoretical and empirical descriptions of these losses are still being established (Callen, 1992).

The control of these self-induced fluctuations is a key to obtaining the plasma parameters needed to fulfill the Lawson condition. Such fluctuations can have a beneficial effect. For example they can prevent the buildup of unwarranted impurity in the plasma center, or they can allow dynamo action to maintain steadily the magnetic structure of a configuration that would not be possible to sustain in absence of fluctuations. It now appears that, at least in tokamak experiments, the natural fluctuation level is compatible with the achievement of controlled fusion conditions.

3. MAGNETICALLY CONFINED EQUILIBRIUM

In order to prevent heat from escaping from a confined hot plasma to the cold surroundings, a means of providing both confinement forces and thermal insulation is needed. Magnetic fields can in principle provide both properties.

Many different magnetic field configurations have been proposed. Roughly the proposals can be divided into two classes, toroidal and open ended. In the former case the trajectories of magnetic field lines lie on a surface of a torus, as shown in the spiral lines in Fig. 3. Charged particles, which constitute the plasma, move nearly parallel to these field lines, and as a result under many conditions the particles remain inside the torus. In open-ended machines the trajectories of magnetic field lines inside the plasma reach the surrounding wall as shown in the two configurations of Fig. 4. However, particles are inhibited from reaching the wall if the magnetic field strength increases toward the wall.

In this section we will discuss some basic ideas in magnetic plasma containment and show how they have been specifically applied to various containment concepts.

3.1 Magnetohydrodynamic Equilibrium Equations

To understand containment one can examine solutions of the magnetohydrodynamic (MHD) equilibrium equations (see MAGNETOHYDRODYNAMICS), which take the form

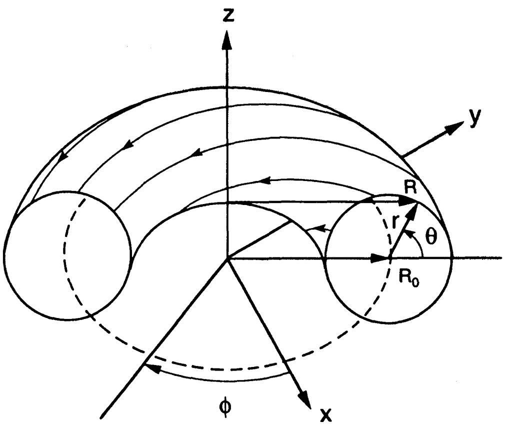

FIG. 3. Typical toroidal configuration showing coordinates (Z,R,ϕ) and (r,θ,ϕ).

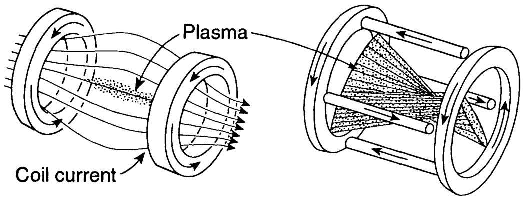

(a) Simple mirror with Helmholtz Coils

(b) Min-B mirror with quadrupole coils

FIG. 4. Two mirror configurations which are examples of open-ended machines. The symmetric mirror (a) tends to be MHD unstable. With a quadrupole field, shown in (b), stabilizing concave-shaped field lines are produced.

$$\mathbf{j}\times\mathbf{B}=\nabla\cdot\mathbf{P},\qquad \nabla\times\mathbf{B}=\mu_0\mathbf{J}, \tag{5}$$

where $\mathbf{B}$ is the magnetic field, $\mathbf{j}$ the current density, and $\mathbf{P}$ the generalized momentum flux tensor. It is frequently appropriate to take $\mathbf{P}$ as isotropic, whereupon $\mathbf{P} = p\mathbf{I}$ with p the scalar pressure and $\mathbf{I}$ the unit dyadic (pressure isotropy is established if the particle lifetime is longer than the collisional relaxation time to a Maxwellian distribution). These equations can be written in stress conservation form,

$$\nabla\cdot\mathbf{\Pi}=0, \tag{6}$$

with the stress tensor given by $\mathbf{\Pi}=\mathbf{BB}/\mu_0-(B^2/2\mu_0)\mathbf{I}-\mathbf{P}$. One can show from the virial theorem that Eq. (5) only allows solutions if on a surface S bounding the containment volume, the "force" $\int\mathbf{\Pi}\cdot\mathbf{n}\,dS$ is nonzero, with $\mathbf{n}$ the unit normal vector to the bounding surface S. This means that in order to obtain equilibrium one must supply an external anchor. One type would simply be plasma leaning against external solid material walls. This method is not generally satisfactory since thermal heat losses would usually be too large [nonetheless there are serious "wall containment" concepts, such as originally proposed by Budker, being developed that depend upon the relatively low heat conduction across magnetic fields to obtain the needed thermal insulation (Ryutov, 1975)]. The more standard technique is to have magnetic coils, which establish external magnetic fields that can isolate the plasma from wall contact. In turn these coils need to be anchored by external forces.

3.2 Two-Dimensional Pinch Solutions

To begin to understand the solutions of Eq. (5), it is informative to consider a system that is unbounded in the z direction, symmetric in the θ direction (where θ is the angle about the z axis), and contained radially in r. If we take the momentum flux tensor $\mathbf{P}$ as isotropic, $\mathbf{P}=p\mathbf{I}$ (note that p is the particle pressure) and the magnetic field of the form $\mathbf{B}=\hat{\mathbf{z}}B_z(r)+\hat{\theta}\,B_\theta(r)$, the general solution to Eq. (5) is

$$p(r)+\frac{B_z^2(r)}{2\mu_0}-\left\langle p+\frac{B_z^2}{2\mu_0}\right\rangle=-\frac{B_\theta^2(r)}{2\mu_0}, \tag{7}$$

where

$$\langle a(r)\rangle=2r^{-2}\int_0^r dr'r'a(r').$$

There are several issues that this form illustrates. First of all, at a wall $r = a$, the pressure $p(a)$ can vanish, and the averaged contained pressure is then

$$\langle p\rangle=[B_z^2(a)+B_\theta^2(a)-\langle B_z^2\rangle]/2\mu_0. \tag{8}$$

We see that the particle pressure can be expressed in terms of a containment parameter,

$$\langle\beta\rangle=2\mu_0\langle p\rangle/B_z^2(a), \tag{9}$$

where $B_z(a)$ is also the applied external magnetic field if there are no embedded external current carriers inside the plasma. Equation (7) then becomes

$$\langle\beta\rangle=1-\langle b_z^2\rangle+b_\theta^2(a) \tag{10}$$

with $b_j(r)=B_j(r)/B_z(a)$.

We now note that there is freedom in specifying b_z or b_θ. The case where $b_\theta=0$ is the "theta pinch" solution, where the nomenclature arises because in such a configuration the plasma current is in the θ direction. In this case one can show that $1>b_z^2(r)$, and hence $\langle\beta\rangle\leqslant 1$. This is why β is frequently used as a measure of effectiveness of stored plasma pressure with respect to an applied magnetic field.

In principle the greatest effectiveness in confining plasma is obtained with the z-pinch solution where plasma currents are in the z direction and $B_z=0$. This case does not require any external magnetic field to achieve radial containment (note that the θ pinch is unbounded in the z direction, which is why the virial theorem is not violated). Thus from just its equilibrium properties, the z pinch is extremely attractive for fusion application. Unfortunately, in its most basic form the z pinch is highly unstable, and drastic modification of the configuration is usually considered necessary to have a tractable concept. Nonetheless, interest in the pure z pinch continues today. Novel ideas for stabilization are being investigated by Lehnert in the ex-trap experiment (Drake *et al.*, 1989). There is also considerable interest by several groups in the high-density fast z pinch produced from thin fibers, e.g., Haines *et al.* (1991), where the time scale for producing fusion energy might be more rapid than the time scale for instability to develop.

The most direct way to stabilize a z pinch is to have simultaneously both $\hat{\mathbf{z}}$-directed and $\hat{\boldsymbol{\theta}}$-directed currents. In Eq. (7) the kinetic pressure and the magnetic pressure due to B_z are completely interchangeable. In the limit of zero kinetic pressure, the solution is a "force-free solution," where the plasma currents are parallel to the magnetic field so that the magnetic force $\mathbf{j}\times\mathbf{B}=0$. There is a class of solutions where the direction of B_z changes inside the discharge, and such solutions are characteristic of the reversed-field pinch (RFP), one of the important concepts in controlled fusion research.

Another prototype solution contained in the structure of Eq. (7) is one where B_z is nearly constant within the discharge and the plasma beta is low. Typically, $\beta_z=2\mu_0 p/B_z^2 \ll 1$ while $\beta_\theta=2\mu_0 p/B_\theta^2 \approx 1$. Such discharges are stable for a wide range of parameters and are the basis of the so-called tokamak, the lead concept for achieving fusion by magnetic containment. Many issues in tokamak containment will be discussed later.

Both these "low-β" solutions sacrifice high-β effectiveness in exchange for achieving stability. Beta values ~10% have been achieved in these concepts and the determination of the upper permissible beta is one of the important questions under investigation in controlled fusion.

3.3 End Stoppering

The magnetohydrodynamic (MHD) solutions presented thus far do not exhibit axial containment. For axial containment either we have to "stopper" the ends or we have to wrap the straight solutions into a torus. First we discuss end stoppering.

Axial stoppering can be achieved by varying the magnitude of the magnetic field along the z axis if the pressure tensor is anisotropic. As particle orbits in a magnetic field cause the pressure tensor to be symmetric about the magnetic field direction, while the parallel component is unaffected by the magnetic field, the pressure tensor $\mathbf{P}$ takes the form

$$\mathbf{P}=p_\perp(\mathbf{I}-\mathbf{bb})+p_\parallel\mathbf{bb}, \tag{11}$$

where $\mathbf{b}$ is in the direction of the magnetic field, $p_\perp$ is the perpendicular pressure, and $p_\parallel$ is the parallel pressure. The component of Eq. (11) along the magnetic field gives the relation

$$\frac{\partial p_\parallel}{\partial s}=\frac{-(p_\perp-p_\parallel)}{B}\frac{\partial B}{\partial s}, \tag{12}$$

where s is the distance along a field line. This equation shows that the pressure $p_\parallel$ can be confined along the magnetic field if B increases toward the ends. This axial containment is an example of the mirror principle. In Sec. 4 we will study in more detail how electric and magnetic forces act on particles to achieve containment.

3.4 Toroidal Configurations

The two-dimensional pinch solution is frequently considered as a mathematical limit of a large-aspect-ratio toroidal system. However, this "attachment" of ends of a nearly straight system onto itself brings in many new physical features, some of which are discussed in this section and others in Sec. 4.

3.4.1 General Considerations. To study containment in toroidal configurations, Eq. (5) has to be solved in three-dimensional geometry. A typical three-dimensional configuration is shown in Fig. 3. In such a configuration there is a toroidal axis of radius R_0 (the dashed curve in Fig. 3) on which a magnetic field closes on itself. Frequently R_0 is constant, but in principle R_0 can vary as it does in some concepts. The other field lines wrap around the toroidal axis as indicated by the lightly drawn spirals in Fig. 3. The winding number of a field line is the ratio of the number of times it wraps around poloidally (the short way around) to the number of times it wraps around toroidally (the long way around). The winding number is commonly called the rotational transform ι; its inverse, $1/\iota$, is denoted as q, the safety factor, which is a parameter closely related to plasma stability. A convenient set of coordinates, shown in Fig. 3, is the toroidal angle ϕ, the poloidal angle θ, and the radial distance to the magnetic axis r. For a toroidally symmetric configuration one can readily show that if the rotational transform is irrational, a single field line is constrained to lie on a surface. Such "magnetic flux" surfaces can also be attained in configurations that are not axisymmetric. This property leads to the stellarator concept, to be discussed below, where the flux structure can be formed directly from external magnetic coils, rather than the internal plasma current.

3.4.2 Vertical Field. In a toroidal pinch, there will be a toroidal plasma current $I_p\hat{\boldsymbol{\phi}}$ that produces B_θ fields. The average self-magnetic force, $\langle \mathbf{I}_p \times \mathbf{B}_s \rangle$ with $\mathbf{B}_s$ the magnetic field produced by $\mathbf{I}_p$, produces the so-called hoop force directed outward in the $\hat{\mathbf{R}}$ direction. This self-force tends to produce expansion of the current ring. To obtain confinement, this force must be compensated by an external vertical magnetic $\mathbf{B}_v$, directed such that $I_p\hat{\boldsymbol{\phi}} \times \mathbf{B}_v$ is in the $-\hat{\mathbf{R}}$ direction. For a torus of large aspect the magnitude of $\mathbf{B}_v$ is found to be

$$\mathbf{B}_v = \frac{aB_\theta(a)}{2R_0}\left(\ln\frac{8R_0}{a} - \frac{3}{2} + \frac{l_i}{2} + \langle \beta_\theta \rangle\right), \tag{13}$$

where $B_\theta(a)$ is the poloidal component of the magnetic field that is generated by the plasma current $I_p \approx 2\pi a B_\theta(a)/\mu_0$,

$$\langle \beta_\theta \rangle = 2\mu_0 \int_0^a dr\, rp(r)/a^2 B_\theta^2(a)$$

is the average beta with respect to the magnetic field, and l_i (the internal inductance) is a quantity of order unity that depends on the detailed current profile. The need of such an external vertical field can be viewed as a consequence of the virial theorem.

From Eq. (13) one infers that the intrinsic β value with respect to the vertical field is large, with $\beta \approx R/a$. Such high beta values can arise if the toroidal magnetic field B_ϕ is zero. However, stability considerations demand the presence of an external toroidal magnetic field B_ϕ, which considerably reduces the β that can be achieved. In the RFP the toroidal B_ϕ and poloidal B_θ magnetic fields are comparable in magnitude, while in the tokamak B_ϕ is considerably larger than B_θ—typically $B_\phi/B_\theta \approx R/a$.

We also note that toroidal magnetic fields are typically produced by magnetic field coils where the coils are wound through the hole in the plasma as shown schematically in Fig. 5(a). In Fig. 5(b) the plasma region and toroidal field coil are shown as they exist in the Text tokamak. The inner coil structure gives rise to complications that would be worth while to avoid if an alternative configuration proved suitable. This is a motivation for the study of compact toroids where there are no external coils through the plasma hole.

3.4.3 Compact Toroids. As $\beta \approx a/R$ appears to be a limitation of confinement, increased β values are achieved by increasing a/R. The extreme is to have $a/R \approx 1$. If in this case the external toroidal field coils are eliminated entirely, one need not have hardware through the inside of the torus. Instead, the plasma can reach the axis of the machine, and one has a configuration known as a compact toroid.

To discuss the structure of the magnetic fields in a compact toroid, we introduce a magnetic flux function ψ_p where

$$\mathbf{B}_\theta = \nabla\psi_p \times \hat{\boldsymbol{\phi}}/2\pi R.$$

The magnetic field then lies along contours of constant ψ_p. A schematic figure of the flux surface of a compact toroid is shown in Fig. 6. The z axis is an axis of revolution. The poloidal magnetic field is closed and reverses direction inside the separatrix (the separatrix is the curve that intersects at the points marked by $\times$, which separate the closed and open field lines). The closed field lines constitute the region of containment. Outside the separatrix the field lines are open, and the plasma that reaches these outer lines, as well as plasma reaching the z axis, flows toward the walls. By and large, the flux surfaces are produced by the toroidal plasma current. However, the hoop force still needs to be compensated for by an external vertical magnetic field, just as in the large–aspect-ratio case.

As an example, a prototype analytic solution for the flux ψ_p without toroidal field, which satisfies Eq. (5), is

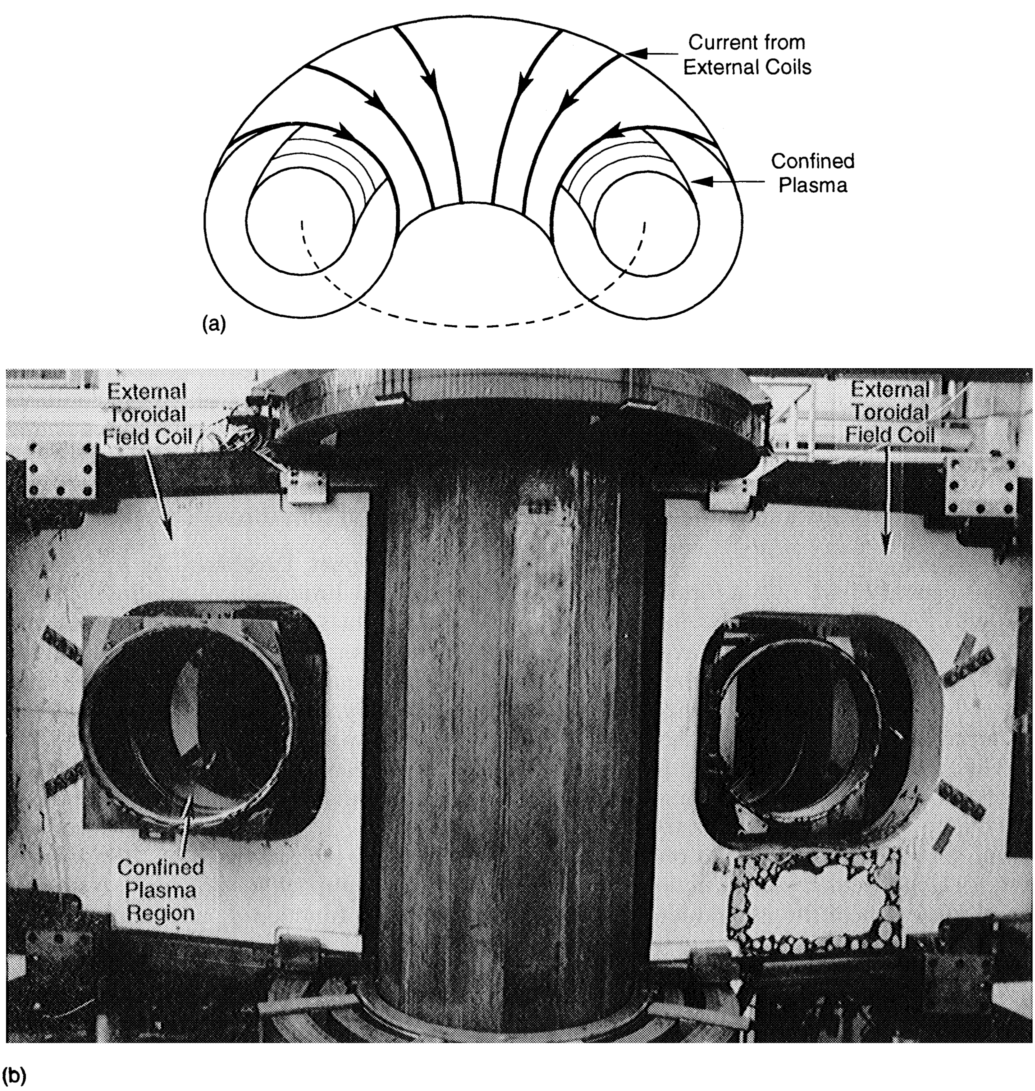
Current from External Coils
Confined Plasma
(a)
External Toroidal Field Coil
External Toroidal Field Coil
Confined Plasma Region
(b)

FIG. 5. (a) In the usual toroidal configuration the plasma is surrounded by external toroidal field coils as indicated schematically by the heavy lines with arrows. (b) The plasma region and field coils are shown for the Text tokamak, at the University of Texas, Austin. The central region is the structural support as well as the magnetic flux column that is used as a transformer to drive plasma current.

$$\psi_p = 4\psi_{\max}(R^2/R_0^4)(R_0^2 - R^2 - \alpha^2 z^2), \qquad (14)$$

where

$$\frac{\psi_{\max}}{2\pi} = \frac{R_0^2}{4}\left(\frac{\mu_0 p_0}{2(1+\alpha^2/4)}\right)^{1/2},$$

$$p = p_0 \psi_p / \psi_{\max} .$$

R and z are radial and axial cylindrical coordinates indicated in Fig. 6, and R_0 is a constant. The plasma pressure is confined to the region $\psi_p > 0$, and the curve $\psi_p = 0$ defines the separatrix. For this solution the separatrix forms an ellipsoid of revolution with major axis R_0 in the radial direction, and R_0/α in the z direction. The required external magnetic field $\mathbf{B}_{\text{ext}}$ is directed in the z direction. The needed external field B_{ext} can be expressed in terms of a beta condition $\beta_0 = 2\mu_0 p_0 / B_{\text{ext}}^2$, with p_0 the pressure at the 0 point of Fig. 6 where the poloidal magnetic field vanishes. For a spherically shaped separatrix ($\alpha = 1$), $\beta_0 = \frac{45}{16}$, while for an axially elongated separatrix ($\alpha \to 0$), $\beta_0 = 1$.

Another type of compact toroid, known as the spheromak, is one where the plasma also

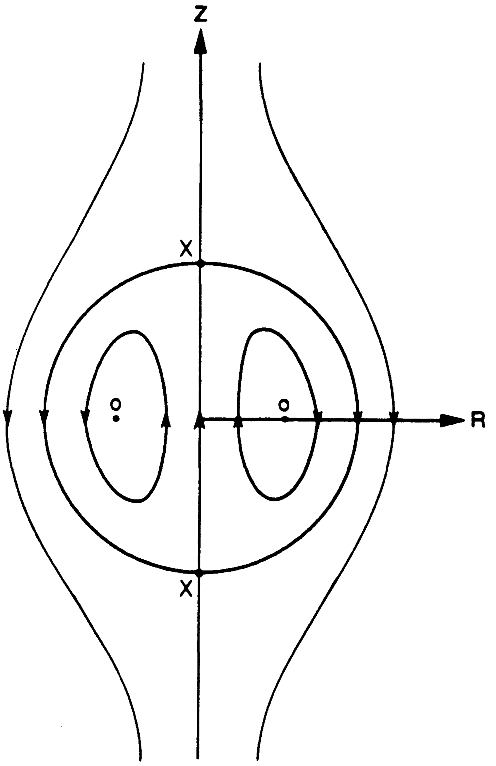

FIG. 6. Schematic flux surface of a compact toroid. The three-dimensional configuration is envisaged by revolving this plot around the z axis. Closed poloidal field lines are closed inside the separatrix curve passing through the points labeled ×. Outside the separatrix, field lines are open. The poloidal flux maximizes at the poloidal field null inside the toroid where $B_\theta=0$ and which is indicated by the points labeled 0. There are no external current carriers penetrating the torus. A toroidal field inside the toroid can arise from plasma currents flowing along poloidal field lines.

carries poloidal current, so that toroidal magnetic field exists inside the containment region, but not outside the separatrix. In these configurations the roles of kinetic pressure and magnetic field pressure from the toroidal field are nearly interchangeable. Now, both the kinetic pressure and toroidal magnetic field pressure maximize at the 0 point and vanish at the separatrix.

A third type of compact toroid is Astron (Christofilos, 1962). This is a configuration where current carriers are energetic particles. In this case the external magnetic field B_{ext} confines the energetic particles in orbits that are essentially the betatron orbits that appear in the theory of particle accelerators (see BETATRONS). The dominant momentum flux associated with the orbits of this configuration is

$$\mathbf{P}=m_b\langle n_b v_\theta^2\rangle\hat{\theta}\hat{\theta} \tag{15}$$

with m_b and n_b the mass and density of the energetic particles and v_θ the θ-directed velocity component. Now Eq. (5) needs to be solved using Eq. (15) for $\mathbf{P}$. If the current formed by this beam of energetic particles is intense enough, the shape of the poloidal magnetic fields is similar to that obtained with the other compact toroid solutions.

3.4.4 Inductive Voltage Drive. The diverse toroidal concepts discussed will last only a magnetic diffusion time unless magnetic flux is "pumped" into the configuration. The diffusive nature of the discharge can be inferred from combining Faraday's induction law, Ohm's law, and Ampère's law, which leads to a diffusionlike equation for the flux ψ_p that is essentially of the form

$$\frac{\partial\psi_p}{\partial t}=\frac{1}{\sigma\mu_0}\,\Delta\psi_p \tag{16}$$

with σ the electrical conductivity. Unless a source of flux is present, Eq. (16) predicts that the magnetic configuration decays with a characteristic time τ_M given by

$$\tau_M\approx\mu_0\sigma a^2/4\approx 4\times10^{-4}a^2T^{3/2}(\text{eV})/Z_{\text{eff}}\ \text{s},$$

where a is the minor radius in meters and T_e(eV) the electron temperature in electron volts, and the electrical conductivity is $\sigma = 1.3\times10^3T_e^{3/2}$ (eV)$/Z_{\text{eff}}$ siemens/m with Z_{eff} the mean atomic number of the discharge. Sample diffusion times are $\tau_M=6\times10^{-3}$ s for $a=0.2$ m, $T_e=50$ eV, which are typical of startup parameters; $\tau_M=1.0$ min for $a = 1.0$ m, $T_e=3$ keV, which are typical parameters for present-day tokamaks; and $\tau_M=0.8$ h for $a=2$ m, $T=15$ keV, which are typical of fusion reactor parameters. We observe that stronger magnetic flux pumping is required for small devices at low temperatures than for large devices at high temperatures.

The most obvious flux pump mechanism is to link magnetic flux, just as one does in a transformer, as shown in Fig. 7. The flux ψ_F in the solenoid core is $\psi_F = B_FA_F$, where B_F is the magnetic field in the solenoid and A_F the cross-sectional area of the solenoid. If B_F

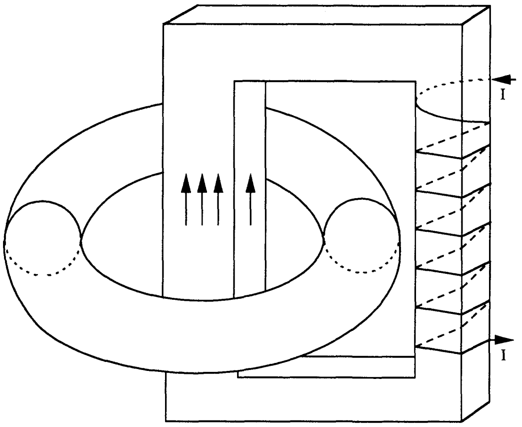

FIG. 7. Transformer link to a toroidal plasma. The change of magnetic flux inside the central transformer column supplies toroidal voltage to the plasma.

changes from B_{FM} to $-B_{FM}$, it follows from Faraday's law that a voltage V_F is imposed on the core plasma given by

$$V_F = \frac{-\partial\psi_F}{\partial t} = \frac{-\partial B_F}{\partial t} A_F .$$

In turn, the voltage is Ohmically dissipated by the plasma, so that $V_F = -A_F\, \partial B_F/\partial t = I_p\mathcal{R}$ with I_p the plasma current and $\mathcal{R}$ the plasma resistance ($\mathcal{R} \approx 2R/\sigma a^2$). Because the total flux that can be pumped is limited to $2A_F B_F$, then for a fixed current and resistance the solenoid can supply voltage for a time t_F given by $t_F = 2B_{FM}A_F/I_p\mathcal{R}$. This Ohmic method of magnetic flux pumping will allow a reactor to operate on the scale of hours. Though quite long by today's standards, Ohmic magnetic flux pumping implies periodic shut-down and start-up of the reactor, which is not desirable from several engineering considerations (one of the more significant reasons is that the many start-up phases lead to large changes in the wall temperature, which cause continual thermal stress that eventually causes wall fracture). Hence, alternative flux pumping techniques, frequently called current drive, are desirable (Fisch, 1987). For example, radio-frequency waves have been used to drive steady plasma currents through a variety of mechanisms. Though less efficient than Ohmic drive, they do offer the possibility of achieving truly steady-state operation in a toroidal machine where plasma current is needed to form nested flux surfaces.

3.4.5 Stellarators. A way to avoid the need for driving current in a toroidal plasma is to design a stellarator configuration where flux surfaces form solely from the current in the external coils. The demonstration that "vacuum magnetic fields" form flux surfaces can be inferred from the properties of a periodic stellarator with helical symmetry and arbitrarily large aspect ratio. Consider a cylindrical coordinate system (r,θ,z), with the system extended in z, and define a phase variable $\chi = \theta - \alpha z$ with α a constant. The vacuum magnetic field can be expressed as $\mathbf{B} = \nabla\Phi$, which, with $\nabla \cdot \mathbf{B} = 0$, requires Laplace's equation to be satisfied, $\Delta\Phi = 0$. A helical stellarator solution has the form

$$\Phi = B_0\, z + \frac{1}{\alpha} \sum_{n=1}^{\infty} b_l I_l(l\alpha r) \sin(l\chi), \tag{17}$$

where $I_l(x)$ is the Bessel function of imaginary argument. The field lines generated by these solutions lie tangent to a helical flux function ψ_H,

$$\psi_H = B_0 \frac{\alpha r^2}{2} - r \sum_{l=1}^{\infty} b_l I_l'(l\alpha r) \cos l\chi. \tag{18}$$

These solutions are the first approximations to the fields that can be generated in a torus with a large aspect ratio ($R_0/a \gg 1$, where R_0 and a are typical major and minor radii). Helical solutions are obtained by winding the external current carrier in a helix M/l times around poloidally for each toroidal turn, with M an integer and $\alpha = M/lR_0$. The pitch of the coil is M/l. Examples of the shapes of constant–helical-flux contours at fixed z generated from various coil configurations are given in Fig. 8(a). At a given poloidal cross section there are l current paths flowing in one direction [shown by the symbol ⊕ in Fig. 8(a)] and perhaps l current paths in the opposite direction (shown by the symbol ⊖). If the latter set of currents is absent the configuration is called a heliotron or torsatron; with a torsatron a single coil forms toroidal and helical fields. The magnetic field lines inside the torus wind around the magnetic axis with a rotational transform that is less than the pitch of the current loops. The rotational transform increases with increasing radial distance until a separatrix is reached (in these figures the separatrices are the contours that cross on themselves), where the

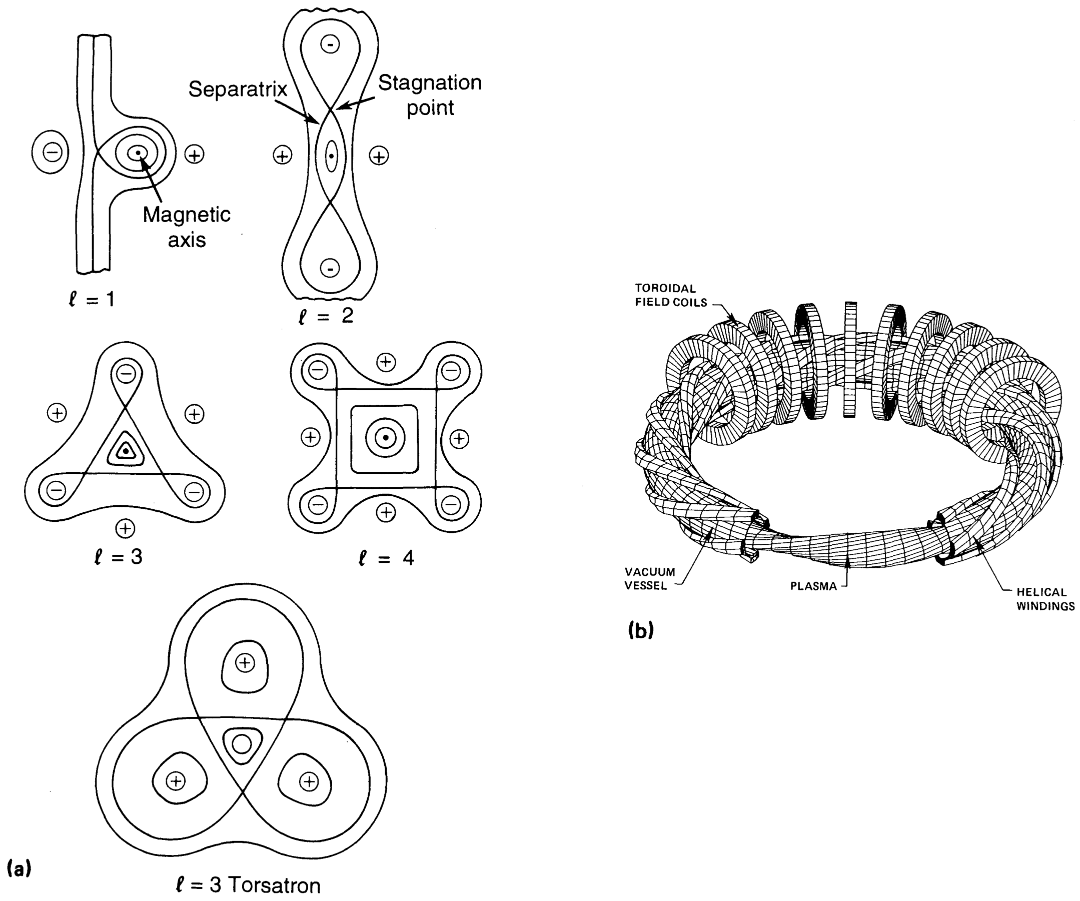

FIG. 8. (a) Magnetic surfaces showing separatrix points, and separatrices for $l=1$, 2, 3, and 4 stellarators and for the $l=3$ torsatron. (b) Three-dimensional view of coil set of an $l=2$ classical stellarator (Wendelstein VII-A) showing the twisting magnetic flux surface (from Carreras *et al.*, 1988).

rotational transform equals the pitch. At radii outside the separatrix, field lines are not contained but intersect the wall. Unlike in a typical tokamak, $q = 1/\iota$ decreases away from the magnetic axis.

In practice a straight system must be wrapped into a large–aspect-ratio torus as in Fig. 8(b). Observe that the magnetic flux surfaces twist to follow the external helical currents. However, strict helical symmetry is broken by the toroidal shape. With strict helical symmetry it follows from Eq. (18) that the magnetic fields inside the separatrix have nested flux surfaces. With helical symmetry broken (as it must be in a toroidal configuration), it is not at all obvious that a field line generates a magnetic flux surface rather than filling a volume of space. Fortunately, the flux surface property is retained over a large volume surrounding the magnetic axis. However, flux surfaces disappear near the separatrix (there field lines meander over a finite volume) and additional structure in the field lines (so-called magnetic islands) arises near surfaces with rational rotational transform. With proper coils, an acceptable volume of mostly nested magnetic surfaces can be designed. Within these good surfaces, plasmas can be retained to beta values comparable to what is possible with tokamaks.

Historically, the stellarator concept was one of the pioneering confinement concepts in controlled fusion. The three-dimensional aspect of the configuration made design and analysis difficult, and progress in the field (in part because of poor design choices) lagged behind tokamak research. In more recent years, as a result of clearer conceptual ideas of the stellarator concept, very promising experimental results have emerged in stellarator research. Today, the stellarator concept, with its multiple design options and its intrinsic truly steady-state nature, is a leading alternative to tokamaks as a means for achieving fusion-grade plasmas (Carreras *et al.*, 1988).

4. PARTICLE ORBITS AND CLASSICAL TRANSPORT IN MAGNETICALLY CONFINED PLASMAS

4.1 Particle Orbits in Magnetic and Electric Fields

Without magnetic and electric fields or collisions, a particle in an enclosed volume moves in a straight line until it hits the surrounding wall. Magnetic and electric fields can be designed to prevent the particles from hitting bounding surfaces. The orbits are determined from the equations of motion,

$$\mathbf{v}=\frac{d\mathbf{r}}{dt}, \quad M_j\frac{d\mathbf{v}}{dt}=e_j(\mathbf{v}\times\mathbf{B}+\mathbf{E}), \tag{19}$$

with M_j and e_j the mass and charge, respectively, of species j. When $\mathbf{B}$ is spatially uniform and $\mathbf{E}=0$, the particle motion perpendicular to the magnetic field is a circle with a radius $\rho_j=v_\perp/\omega_{cj}$ (known as the Larmor radius) about a point called the guiding center, where $v_\perp$ is the velocity component perpendicular to the magnetic field and $\omega_c = e_jB/M_j$ is the cyclotron frequency of species j. However, the parallel velocity, $v_\parallel\equiv\mathbf{v}\cdot\mathbf{B}/B$, remains constant and the particle moves without containment parallel to the magnetic field in a manner shown in Fig. 9(a). With a homogeneous static electric field present perpendicular to the magnetic field, a particle takes on an additional drift motion, $\mathbf{v}_E=\mathbf{E}\times\mathbf{B}/|B|^2$, giving rise to a motion shown in Fig. 9(b), with the motion of the guiding center being perpendicular to the direction of $\mathbf{E}$.

The motion we have described has not taken into account collisions between particles due to the binary Coulomb interaction. Because of the long-range nature of the Coulomb potential, the many small-angle deflections arising from collisions with relatively large impact parameters are more important than the large-angle deflections arising from collisions with small impact parameters. The small-angle deflections cause the particle pitch angle $\theta=\cos^{-1}(v_\parallel/|\mathbf{v}|)$ of species j to change by 90° in a characteristic scattering time τ_j, which is given by

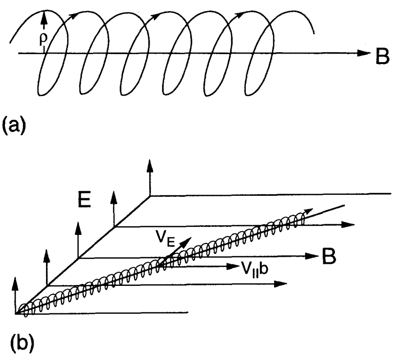

FIG. 9. (a) Motion of a charged particle in a uniform magnetic field. (b) Motion of a charged particle in uniform electric and magnetic fields that are perpendicular to each other. The particle guiding center moves in a plane perpendicular to **E**.

$$\begin{aligned} n_e\tau_i&\approx 1.5\times10^{10}A^{1/2}E_i^{3/2}(\text{keV})/Z^3,\\ n_e\tau_e&\approx 1.5\times10^{10}(M_e/M_i)^{1/2}E_e^{3/2}(\text{keV})/\\ &\quad(1+1.5Z), \end{aligned} \tag{20}$$

where E_j(keV) is the mean energy of species j in keV, $j\equiv i$ refers to ions, $j\equiv e$ refers to electrons, A and Z refer to the ions' atomic mass and number, respectively, and we assume only one ion species for simplicity.

As a particle scatters, the guiding center of a particle likewise deflects, and in a scattering time $t = \tau_j$, the guiding center is deflected by roughly one Larmor radius. For time scales much longer than τ_j the evolution across field lines is described by random walk theory. The diffusion coefficient D_j for species j is given by $D_j=\Delta x_j^2/\Delta t$, with Δx_j the random step that occurs after a time Δt. In our case, $\Delta x=\rho_j$ and $\Delta t=\tau_j$, which gives the "classical" diffusion coefficient of a particle in a straight magnetic field,

$$D_j\approx\rho_j^2/\tau_j. \tag{21}$$

When collisions are taken into account in fields where **E** and **B** are mutually perpendicular, particle diffusion across field lines gives rise to a thermal conduction coefficient $\chi\sim D_i$. As a result, for a slab model of a plasma, with the magnetic field in the z direction and an electric field in the x direction, and with the plasma confined in x within a characteristic spatial scale $L_p \gg \bar{\rho}_i$ ($\bar{\rho}_i = \sqrt{2T_i/m_i}$), the characteristic lifetime for energy confinement, τ_{th}, is

$$\tau_{\text{th}} \approx L_p^2/\chi_i \approx L_p^2 \bar{\tau}_i/\bar{\rho}_i^2 \, , \tag{22}$$

where for $\bar{\tau}_i$ we use Eq. (20) with $E_i = T_i$. As $L_p \gg \bar{\rho}_i$, the energy lifetime is many 90° deflection times.

In addition collisions prevent the unbounded acceleration of particles when a parallel electric field is present. With collisions the parallel equation of motion gives roughly

$$\frac{dv_\parallel}{dt} = e_j \frac{E_\parallel}{M_j} - \frac{v_\parallel}{\tau_j}$$

with a steady-state solution $v_\parallel = e_j E_\parallel \tau_j / M_j$ which produces a plasma current

$$j_\parallel = -en_e(v_{\parallel e} - v_{\parallel i}) \approx \tau_e n_e e^2 E_\parallel / M_j \equiv \sigma E_\parallel$$

with e the magnitude of the electron charge. Within a numerical coefficient this value of σ reproduces the electrical conductivity used in Eq. (16).

4.2 Spatially Varying Fields

By considering magnetic and electric fields that vary slowly in space compared to a Larmor radius, and slowly in time compared to a cyclotron period $\sim 1/\omega_c$, one can demonstrate particle containment along the magnetic field. Under these conditions the quantity $\mu \equiv m(\mathbf{v}_\perp - \mathbf{v}_E)^2/B$ is an adiabatic invariant, i.e., it remains conserved to high accuracy during a particle's motion. As a consequence of this invariance, the parallel velocity of the guiding center, $v_\parallel$, is given by

$$v_\parallel = \pm \{ (2/M_j) [\mathscr{E} - \mu B(s) - e_j \, \Phi(s)] \}^{1/2}, \tag{23}$$

where $-\nabla\Phi = \mathbf{E}$, $\mathscr{E} = \frac{1}{2}mv^2 + e_j \, \Phi$ is the total conserved particle energy, $B(s) = |\mathbf{B}|$, and Φ is the electrostatic potential. From Eq. (23) it is clear that if the magnetic field is increasing, a particle can be contained along the field line. Specifically, when $\Phi = 0$, the particle containment condition is

$$\frac{v_{\perp 0}^2}{v^2} > \frac{B_0}{B_{\max}} \equiv \frac{1}{R} \, , \tag{24}$$

where $B_{\max}$ and B_0 are respectively the maximum and mid-plane values of the magnitude of the magnetic field, $v_{\perp 0}$ is the velocity component perpendicular to the magnetic field, the subscript 0 refers to the value at the mid-plane, and R is called the mirror ratio. This type of containment is referred to as mirror containment, where particles that satisfy Eq. (24) oscillate between two turning points inside the mirror peaks [for an axially symmetric machine where $B(-s) = B(s)$ the turning points are at $s = \pm s_T$, where $B(s_T)/B_0 = v^2/v_{\perp 0}^2$]. Particles for which $v_{\perp 0}^2/v^2 < 1/R$ are in the so-called loss cone, and they leave the mirror machine in one transit time.

The particles that satisfy Eq. (24) are axially trapped and remain there for many axial bounce periods. The limiting lifetime, τ_{sc}, is determined by collisions causing a particle to pitch-angle scatter into the loss cone. This time is roughly $\tau_{sc} \sim \tau_j(1 - 1/R)$ for $R - 1 \ll 1$ and $\tau_{sc} \sim \tau_j \ln R$ for $R \gg 1$, with τ_j given by Eq. (20).

In the absence of parallel electric fields, electrons are lost more rapidly than ions, as $\tau_{sce}/\tau_{sci} \approx (M_e/M_i)^{1/2}(T_e/\bar{E}_i)^{3/2}$. Hence the plasma will charge up positively, producing an axial electric field pointing outward toward the mirror peaks. This field reinforces the containment of electrons and weakens the containment of ions. The magnitude of the negative potential adjusts so that the axial losses of electrons and ions will be equal. Thus the electrostatic potential is of intrinsic importance in the containment of plasma in a mirror machine. It is called an ambipolar potential, as it regulates the charged-particle loss to occur with nearly zero current flow.

The containment condition for ions and electrons with an ambipolar potential is obtained using Eq. (23) by demanding that the turning point s_T lie inside the mirror peak position at $s = s_{\max}$. This modifies the containment condition given by Eq. (24), to

$$v_{\perp 0}^2 > \frac{v_0^2}{R} + \frac{2e_j\Delta\Phi}{M_j R}, \tag{25}$$

where $\Delta\Phi = \Phi(0) - \Phi(s_{\max})$. For ions, where the charge $e_j = Ze$ is positive, we see that all ions with kinetic energy in the midplane less than $e_j\Delta\Phi/(R-1) \equiv E_{\text{crit}}$ are lost in one transit.

On the other hand, for electrons, where $e_j = -e$ is negative, the containment region is enhanced. All low-energy particles for which $\frac{1}{2}M_e v_0^2 < e\Delta\Phi/(R-1) = E_{\text{crit}}$ are trapped independent of pitch angle. Such particles can only be lost if during their diffusion process they gain energy. Electrons with midplane kinetic energies greater than E_{crit} are lost by pitch-angle scattering caused by collisions.

In mirror machine operation the ambipolar potential $e\Delta\Phi$ is generally several times the electron temperature T_e, which, as we shall see, is an order of magnitude less than the ion mean thermal energy. Hence, most electrons are trapped independent of pitch angle, where they remain for many collision times. Thus the bulk of electrons relax to nearly a Maxwell distribution. Only a high-energy tail of electrons, whose density is roughly $n_e\exp(-e\Delta\Phi/T_e)$, do not thermalize since these electrons are lost by pitch-angle scattering. Hence, the effective rate of loss $\nu_e(\Delta\Phi)$ of electrons with an ambipolar potential present is roughly

$$\nu_e(\Delta\Phi) \approx \frac{1}{\tau_e} \exp\left(-\frac{e\Delta\Phi}{T_e}\right).$$

If $q\Delta\Phi/(R-1) \lesssim \bar{E}_i$, the ion loss rate is somewhat reduced, but basically comparable to $1/\tau_i$ given by Eq. (20). Thus equating the electron and ion loss rates, we find that the ambipolar potential is roughly given by

$$\frac{\Delta\Phi}{kT_e} \approx \ln\left[\left(\frac{M_i}{M_e}\right)^{1/2}\left(\frac{\bar{E}_i}{T_e}\right)^{3/2}\right]. \qquad (26)$$

Typically this factor is ~5 while $\bar{E}_i/T_e \sim 10$. The reason why such a large ratio of $\bar{E}_i/T_e$ can be maintained in steady state is because a typical ion is lost with an energy somewhat less than its mean value $\bar{E}_i$, while a typical electron is lost with a mean energy $e\Delta\Phi + T_e \gg T_e$.

4.2.1 Q Value in a Mirror Machine. The ratio of the power produced by fusion to the power lost by classical collisions is frequently referred to as Q. Detailed solutions of the Fokker–Planck equation (the equation that describes the diffusive evolution of particle distribution functions due to multiple small-angle scattering) in mirror geometry show that Q can only slightly exceed unity for the *d-t* reaction. This result severely limits using a simple mirror machine for net energy production from the fusion reaction. Therefore, various methods to enhance the efficiency of energy production have been proposed. One simple procedure is to convert the energy flux of outgoing ions directly to electrical potential energy so that the effective Q value then becomes $Q_0/(1-\eta)$ where Q_0 is the original Q value and η is the ultimate fraction of the outgoing energy that can be reintroduced into the plasma ($\eta \sim 0.7$ is a realistic expectation).

4.2.2 Electrostatic Potentials in Tandem Mirrors. An alternative idea is the tandem mirror concept, where parallel electric fields are actively used to enhance the containment. A schematic of a basic tandem mirror configuration is shown in Fig. 10. There is a central cell mirror containing the bulk of the energy, where the potential is negative with respect to the surrounding end cells so that ions can be contained, while the end cells are separate mirror machines whose electrostatic potential with respect to the wall is positive so that electrons are contained as they are in a conventional mirror machine. This potential structure can be established if the mean ion energy in the end plugs is larger than the central-cell ion temperature. However, because of the much larger volume of the central cell, most of the energy is stored there. The role of end plugs is to serve as a containment dam that comes from the electric field structure. Pitch-angle scattering of ions in the central cell does not cause particle ion loss, while the potential between the plasma and the end wall prevents pitch-angle scattering of electrons from causing particle loss. The bulk losses arise only from pitch-angle scattering of ions and electrons in the high-energy tail. The ions in the

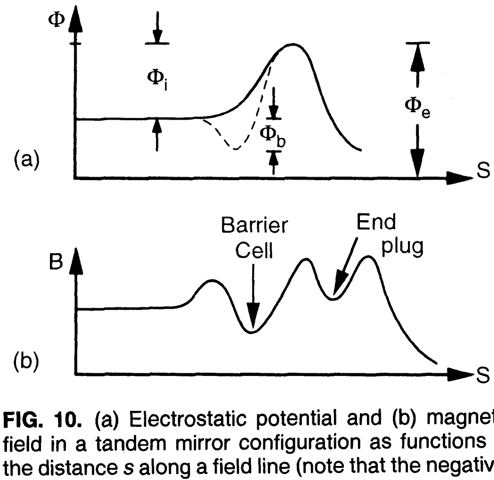

FIG. 10. (a) Electrostatic potential and (b) magnetic field in a tandem mirror configuration as functions of the distance *s* along a field line (note that the negative-*s* region is a reflection of the positive-*s* region). Φ_i and Φ_e are the containing potentials for ions and electrons. With a thermal barrier present, Φ_b is the potential that isolates electrons in the central cell from the end plug.

plugs are susceptible to the usual mirror losses, but their low inventory of total stored energy diminishes the seriousness of these losses. As a result the Q value of a tandem mirror can be increased substantially above that predicted for a simple mirror machine. Even further improvement in Q can be designed by actively controlling the ambipolar potential and forming so-called thermal barriers. The dashed curve in Fig. 10 indicates how the potential would look with a thermal barrier. With a thermal barrier present, the temperatures of electrons in the plug and central cell can be decoupled, which ultimately allows the ions in the plug to have a mean energy comparable to that in the central cell. Hence the parasitic losses due to collisions in the plug are not as severe as in the more conventional tandem mirror, and the Q value can be improved still further.

4.3 Orbits in a Toroidal Field

The alternative to an open-ended system is to have a toroidally shaped plasma. Then to the extent that particles follow field lines, one would expect containment inside the torus. However, toroidal containment is more subtle, as particle trajectories differ somewhat from the trajectory of a magnetic field even in the absence of electric fields. For example, consider the simplest magnetic field to return on itself, which arises from a straight wire, and has the form

$$\mathbf{B}=B_\phi(R_0/R)\hat{\boldsymbol{\phi}}\,, \tag{27}$$

where ϕ is the toroidal angle about the z axis, $\hat{\boldsymbol{\phi}}$ is a unit vector in the ϕ direction, R is the major radius, and R_0 is a convenient constant denoting the typical major radius. A magnetic field line in such a field is a circle of radius R. However, such a field does not allow charged particles to be contained in a closed volume. The magnetic field given by Eq. (27) has a finite magnetic field gradient $(1/|B|)\nabla|B| = -(1/R)\hat{\mathbf{R}}$ and field-line curvature $(\mathbf{b}\cdot\nabla)\mathbf{b} = -\hat{\mathbf{R}}/R$, where $\mathbf{b}=\mathbf{B}/|B|$ is the unit vector along a magnetic field line [note that in vacuum magnetic fields where $\nabla\times\mathbf{B}=0$, $\mathbf{b}\times(\mathbf{b}\cdot\nabla)\mathbf{b} = \mathbf{b}\times\nabla|B|/|B|$]. From Eq. (19) one can show that magnetic gradients and field-line curvature produce an orbit drift velocity transverse to the magnetic field, $\mathbf{v}_D$, given by

$$\mathbf{v}_D=\frac{v_\parallel^2}{\omega_c}\,\mathbf{b}\times(\mathbf{b}\cdot\nabla)\mathbf{b}+\frac{v_\perp^2}{2\omega_c}\,\frac{\mathbf{b}\times\nabla|B|}{B}\,. \tag{28}$$

In the field given by Eq. (28), particle trajectories primarily follow the ϕ-directed field, but in addition the drift velocity in Eq. (28) causes the orbits to drift indefinitely in the negative z direction transverse to the magnetic field with a speed

$$v_D=(\tfrac{1}{2}\,v_\perp^2+v_\parallel^2)/\omega_c R. \tag{29}$$

To prevent unbounded motion a poloidal magnetic field needs to be present. One simple way of establishing such a field is with a current ring carrying a current $I_{\phi 0}$ at a major radius $R=R_0$, which gives a magnetic field $\mathbf{B}_\theta=B_{\theta 0}\hat{\theta}r_0/r$ if $r\ll R$, with $B_{\theta 0}=\mu_0 I_{\phi 0}/2\pi r_0$, where r is the minor radius and r_0 is a constant. Such a configuration is referred to as a hard-core pinch or levitron. Alternatively, the toroidal current can be produced by the various toroidal diffuse pinches where the plasma carries toroidal current, as described in the previous section. The magnetic field then has the generic form

$$\mathbf{B}=B_\phi(R_0/R)\,[\hat{\boldsymbol{\phi}}+q(r)\hat{\boldsymbol{\theta}}], \tag{30}$$

with $q(r)$ the inverse rotational transform. In Fig. 11 we show typical drift orbits of positive charged particles in this field projected onto a plane $\phi=\text{const}$ when $v_\parallel/v \gg (r/R)^{1/2}$. In addition, not shown in the figure, there is rapid oscillatory motion about the drift trajectory due to the Larmor motion. Figure 11(a) is for passing particles, a terminology that applies to those particles where $v_\parallel/v>(2r/R)^{1/2}$. These particles always circulate in a fixed θ direction. The dotted line denotes the magnetic flux surfaces formed from the magnetic field given by Eq. (30). One should note that co-moving particles—i.e., those particles whose sense of motion produces a current along the magnetic field that is in the same sense as the prevailing current [in Fig. 11(a) the prevailing current points toward the reader]—are shifted inward, while counter-moving particles (i.e., those particles that move in the opposite direction) are shifted outward. This asymmetry produces a bias that helps sustain current in a toroidal discharge. This current bias also exists in the electron motion. The trajectory's structure can be understood by noting that the particles' velocity is primarily tangential to

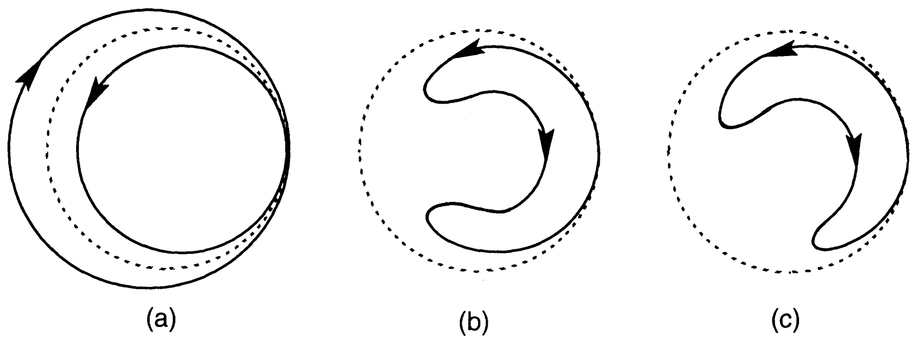

FIG. 11. Particle motion in a tokamak projected onto a fixed-ϕ cross section. In (a), particles moving counter to the toroidal current are shifted to orbits that are outside the dotted flux surface, while particles moving with the toroidal current have orbits interior to the flux surface. The orbit of a trapped particle is shown in (b) and (c). In (c) the orbit is displaced by a toroidal electric field, giving rise to an overall inward drift motion.

the flux surface. However, the additional drift velocity causes a positively charged particle moving in the clockwise direction to move radially inward from the flux surface when it is above the midplane. This inward displacement is compensated when the particle is below the midplane where the drift velocity takes the particle back to the flux surface. A similar argument holds for the counterclockwise rotating particle. The overall excursion from the flux surface

$$\delta r \approx q(r)\rho(r)(v_\parallel^2+v_\perp^2/2)/v_\parallel v_\perp$$

(recall that ρ is the Larmor radius $v_\perp/\omega_c$) is inward for "co-moving" particles (particles moving parallel to the current) and outward for "counter-moving" particles as indicated in Fig. 11(a).

On a flux surface the magnetic field varies, and the mirror ratio on the flux surface is approximately

$$R=\frac{B_\phi(R=R_0-r)}{B_\phi(R=R+r)}=\frac{1+\epsilon}{1-\epsilon}\approx 1+2\epsilon \qquad (31)$$

with $\epsilon=r/R$ the inverse aspect ratio. If $v_\parallel^2/v_\perp^2<2\epsilon$, particles are mirror trapped on the flux surface and they are constrained to remain on the outside of the torus. They are referred to as "trapped" particles. The orbit of a trapped particle, projected onto a cross-sectional plane where ϕ is constant, is shown in Fig. 11(b). Because of the shape of the trajectory, the orbit of a trapped particle is referred to as a "banana" orbit. These orbits are still contained in a finite volume, but their orbit width, given by

$$\delta r \approx q\rho/\epsilon^{1/2} \equiv r_b\,,$$

is somewhat larger than that of a passing particle.

4.4 Neoclassical Diffusion

The complications of the orbit structure in a toroidal magnetic field cause an enhancement of the previously estimated classical diffusion. The appropriate theory is called neoclassical diffusion (Galeev and Sagdeev, 1979) and we now present a heuristic discussion of its implications.

Using random-walk theory, we estimate the diffusion coefficient of a particle by noting that a particle will diffuse an orbit width of its trajectory, Δr, in an effective collision time τ_{eff}. The diffusion coefficient D can be estimated as

$$D=\Delta r^2/\tau_{\text{eff}}\,, \qquad (32)$$

while the overall energy containment time is then given by

$$\tau_E \approx a^2/D=(a^2/\Delta r^2)\tau_{\text{eff}}\,.$$

The majority of particles are passing and their effective relaxation time is a 90° scattering time τ_j [τ_j is either τ_i or τ_e, given in Eq. (20), depending on the type of particle], while their mean displacement is $\Delta r \approx q\rho$. Hence, the diffusion coefficient for passing particles, D_{pass}, is given by $D_{\text{pass}} \approx q^2\rho_j^2/\tau_j$, with ρ_j the Larmor radius of species j. However, this value for the diffusion coefficient turns out to be too small. Surprisingly, it is the diffusion of the trapped-particle minority that primarily determines the containment characteristics. This is because Δr is larger for a trapped particle than for a passing particle and τ_{eff} is shorter. For a trapped particle $\Delta r \approx q\rho/\epsilon^{1/2}$, which is the characteristic deviation of a trapped particle from a flux tube, and $\tau_{\mathrm{eff}} \approx \tau_j\, \epsilon$, which is the characteristic time a trapped particle remains trapped. The shorter τ_{eff} arises because of the diffusive nature of the Coulomb collisional process. Hence if τ_j is the time for a particle to diffuse 90° in velocity space, $\tau_j \sin^2 \Delta\theta$ would be the characteristic time to diffuse a pitch-angle width $\Delta\theta$; for trapped particles $\sin \Delta\theta \approx R-1 \sim \epsilon^{1/2}$, where R is the mirror ratio on the flux tube of a tokamak. Hence the diffusion coefficient of a trapped particle is

$$D_{\mathrm{trap}} \simeq \Delta r^2/\tau_{\mathrm{eff}} \simeq q^2\rho^2/\epsilon^2\tau_j\,.$$

The fraction of phase space that is occupied by trapped particles is approximately $\sin \Delta\theta \simeq \epsilon^{1/2}$, and hence the global diffusion coefficient due to both trapped and passing particles is

$$D \approx \epsilon^{1/2} D_{\mathrm{trap}} + D_{\mathrm{pass}} \approx \frac{q^2\rho^2}{\tau_j}\left(\frac{1}{\epsilon^{3/2}}+1\right).$$

Clearly the contribution to D from the trapped particles is more important than from the passing particles. The global energy lifetime due to thermal conduction is then determined by

$$n\tau_E \approx \frac{a^2 n}{D} \approx \frac{a^2}{\rho_i^2 q^2}\,\frac{n\tau_i}{(a/R_0)^{3/2}}$$

with ρ_i the ion Larmor radius. The large factor a^2/ρ_i^2 ensures that the intrinsic thermal conduction losses in a tokamak can satisfy the Lawson criterion for ignition by a large margin.

4.4.1 Bootstrap Current. The so-called bootstrap current is another important collisional effect. It arises because

1. co-rotating particles tend to be shifted inward while counter-rotating particles are shifted outward;
2. the primary diffusion process is significant only when particles are trapped or nearly trapped.

As a consequence the statistical weight of a particle, even when it is passing, is determined by its weight when it was trapped. At a given radius r the co-circulating passing particles then have a weight of trapped particles that were centered around an inward-shifted radius $r - q\rho/\epsilon^{1/2}$, while a counter-circulating particle has an equal weight shifted to the outward radius $r + q\rho/\epsilon^{1/2}$. The current density at the flux surface r is then given by

$$j_{\|} = e_j[n^+(r) - n^-(r)]\,v_{\mathrm{th}}\,,$$

where n^+ is the density (proportional to the statistical weight) of co-moving particles and n^- the density of counter-moving particles. Using $n^+(r + 2q\rho/\epsilon^{1/2}) \approx n^-(r)$ we find the bootstrap current density

$$j_{\|} \approx -2e_j\,\frac{\partial n}{\partial r}\,\frac{q\rho}{\epsilon^{1/2}}\,v_{\mathrm{th}}\,.$$

This can be a substantial fraction of the current required to maintain a discharge. It allows for a substantial decrease in the current drive needed to maintain a steady-state tokamak.

4.4.2 Ware Pinch. As the bootstrap current results from a spatial gradient that "drives" current, it follows from an Onsager relation that a toroidal electric field will induce a radial spatial gradient. The phenomenon is called the Ware pinch.

In this phenomenon the $\hat{\phi}$-directed electric field that is being used to drive current is not quite parallel to the magnetic field. For passing particles there is an inward flow speed $\mathbf{E}\times\mathbf{B}/|B|^2 \approx E_\phi B_\theta/|B|^2$. However, the flow speed is too slow for trapped particles, which actually move inward at the speed E_ϕ/B_θ. This enhanced speed arises for the following physical reason.

When Eq. (23) is used to determine the turning point of a trapped particle with the electric potential given by $\Phi = \int^s ds' E(s')$, one finds that the turning points are displaced. As a result, a banana orbit is shifted as shown in Fig. 11(c). In the figure shown, the particle spends more time above the mid-plane than below it. The downward curvature drift then directs the particle towards the flux surface more frequently inward than it is directed outward. As a result, after averaging over a particle's bounce orbit, a trapped particle will move inward at an inward radial speed given by

$$\dot{r} = -E_\phi/B_\theta\,,$$

where E_ϕ is taken positive if it is in the direction of the prevailing toroidal current. As only trapped particles with a density $n_T \approx (r/R)^{1/2} n$ move at this rate, the actual inward flux is

$$n_T \bar{v}_r \approx n(r/R)^{1/2}(E_\phi/B_\theta).$$

4.5 Radial Electric Field

The effect of a radial electric field can also cause a profound change in particle containment. Because electrons move rapidly along the field lines forming flux surfaces, there is a strong tendency for a magnetic flux surface to be an electric potential surface. Consequently, the strongest electric fields in a tokamak tend to be normal to the flux surface. Such an electric field produces the standard **E×B** drift, whose flow is tangential to a flux tube, nearly in the θ direction. This motion alters the characteristic displacement of a particle from a flux surface. Without an electric field, the magnetic drift motion of particles causes a typical radial flow speed $v_r \sim (v_\parallel^2 + \frac{1}{2}v_\perp^2)/\omega_c R$ that alternates towards and away from the flux surface, with a radian frequency $\omega_b \approx v_\parallel / qR$ in the poloidal direction, giving rise to a displacement

$$\delta r_b = v_r/\omega_b \approx q(v_\parallel^2 + \tfrac{1}{2}v_\perp^2)/\omega_c v_\parallel$$

from the flux surface. With a radial electric field present, the particle has an additional poloidal velocity component $v_{E\theta} \approx cE_r/B$, which should be added to the θ component of the parallel flow, giving a total θ-component flow speed

$$v_\theta = (\epsilon/q)(v_\parallel + qv_{E\theta}/\epsilon).$$

Hence the circulation frequency is

$$\omega_b \approx v_\theta/r = (v_\parallel + qv_{E\theta}/\epsilon)/qR,$$

and the radial displacement is

$$\delta r_b \approx q(v_\parallel^2 + v_\perp^2/2)\{\omega_c[v_\parallel + q(v_{E\theta}/\epsilon)]\}.$$

Without a radial electric field, the largest particle displacement occurs for small $v_\parallel/v$ where the distribution of particles peaks. The radial electric field changes the region of phase space for which the displacement from a flux tube is maximum. Instead of $v_\parallel \approx 0$, the displacement maximizes at $v_\parallel \simeq v_{E\theta}/\epsilon$. It is then possible for the large particle displacement to affect only particles in the tail of the distribution function where there are relatively few particles, while all the particles in the bulk remain close to the magnetic surface.

Recent tokamak experiments, e.g. (Asdex Team, 1989), have found that there is a profound improvement of plasma containment from more usual discharges in the so-called *H*-mode regime. The radial electric field appears to be an essential element in controlling particle transport.

4.6 Orbits in Stellarators

In stellarators, where toroidal symmetry is broken, particle containment can differ profoundly from systems with toroidal symmetry. In the latter case with a finite rotational transform, the particle displacement from a flux tube is guaranteed to be some factor of a Larmor radius. Hence with toroidal symmetry present, as B becomes larger the particle displacement from the flux surface becomes smaller. The lack of toroidal symmetry allows special classes of particles to deviate substantially from the flux surfaces, to distances that can be comparable to the machine size itself, independent of the magnitude of the magnetic field. In the absence of electric fields, the particles most susceptible to this large deviation are those that are mirror trapped by the magnetic-field modulations due to the combined toroidal and helical field structure. They form so-called super-banana orbits, where in one bounce time the orbit appears like a trapped banana orbit. As the magnetic moment is conserved, the banana orbit moves with the magnetic field at the banana tip, remaining constant. This constraint in itself does not guarantee orbit containment. Thus, many stellarator configurations in the absence of radial electric fields have inherent loss regions. With a radial electric field present, orbits will remain closer to flux surfaces, and containment can be improved. However, as the electrostatic potential tends to be comparable to the temperature of the plasma, it is very difficult to modify the containment properties of the energetic particles. This particularly applies to the charged fusion products that need to be retained in a reactor during ignition. Hence the large orbit displacement from a flux surface will give rise to a direct particle loss of a relatively large fraction of the fusion products in a stellarator (perhaps ~25%). Such a factor does not quench ignition scenarios, but does introduce additional problems. For example, losses tend to be concentrated in specific places of the wall, which induces very special heat load and sputtering problems that need to be addressed in reactor designs.

5. PLASMA INSTABILITY AND ANOMALOUS TRANSPORT

A plasma is characterized by several intrinsic and disparate spatial and temporal scales. Hence, even in the same system, several different types of instabilities can arise that cause different responses. Long-wavelength instability can cause rapid disintegration of a configuration, or can establish a threshold which prevents operation above a critical beta value. Short-wavelength instability tends to saturate at relatively low perturbed field values, so that while the integrity of a system can still be maintained, the plasma transport is frequently enhanced above intrinsic losses.

5.1 Ideal MHD Theory

A first approach to the description of plasma stability is to use ideal magnetohydrodynamic (MHD) theory (see MAGNETOHYDRODYNAMICS). In this model, viscosity, resistivity, and heat conduction are ignored, and the plasma pressure satisfies the ideal gas law. In Sec. 3 we discussed equilibria that arise from the MHD equations. If these equilibria are perturbed, the time-dependent response determines whether small perturbations remain small or grow to finite values. The former case implies stability, while the latter situation implies instability.

5.1.1 Simplified Dynamical Model. When a plasma is described by the ideal MHD model, the method of determining stability is a generalization of the principle of virtual work, that is readily illustrated in a one-dimensional system with two degrees of freedom. A potential $V(\xi)$ is given, say, in the shape of Fig. 12 with ξ the position and V the potential. The equation of motion is given by

$$\frac{d^2\xi}{dt^2}+\frac{\partial V(\xi)}{\partial \xi}=0. \tag{33}$$

In Fig. 12, equilibrium points exist at positions *a*, *b*, *c*, and *d*, where $\partial V/\partial \xi = 0$.

From the principle of virtual work the stability of the equilibrium points are determined from the sign of the perturbed potential energy,

$$\delta W \equiv \delta\xi^2 \frac{\partial^2 V(\xi_i)}{\partial \xi^2},$$

where ξ_i is the equilibrium point where $\partial V(\xi_i)/\partial\xi=0$, $\delta\xi \equiv \xi-\xi_i$ is an arbitrarily small displacement from ξ_i, and δW is the virtual work that must be applied to displace the system. If $\delta W>0$ for every arbitrarily small displacement, the system is locally stable and an initially small perturbation will remain small, while if $\delta W<0$, an initially small perturbation will grow to finite and perhaps even infinite amplitude. For this one-dimensional system determining the sign of δW is equivalent to determining the sign of $\partial^2 V/\partial \xi^2$.

In the simple system shown in Fig. 12 it is clear that initially small deviations about points *a* and *c* lead to finite deviations from these points, while small deviations from points *b* and *d* remain small. The system illustrated here is simple enough that we can ascertain that a small deviation about point *c* leads to a

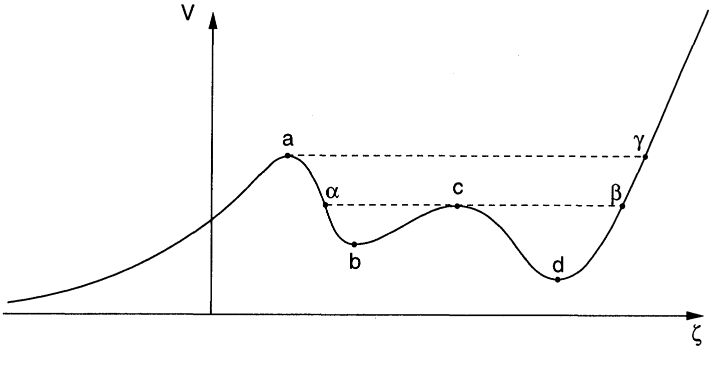

FIG. 12. Potential diagram for one-dimensional motion.

bounded oscillation between points α and β. With a small amount of dissipation present, one expects a state initially at c to relax to either of the points b or d. However, a small deviation about point a can quickly lead to an unbounded excursion. Hence this simple example illustrates how instability can lead to rapid disruptions or to a nonlinear bounded state that can eventually relax to a stable state.

5.1.2 Generalization to Plasmas. These relatively simple considerations can be applied to more complicated systems with many degrees of freedom and even to systems where the degrees of freedom form a continuum, as it does with the MHD equations; see for example Friedberg (1982) for a detailed derivation and detailed application of the method. If one can show that the virtual work (or, as it is frequently called, the MHD energy) is positive for every possible small displacement, stability of the ideal MHD equations is guaranteed. The MHD energy is a very physical quantity which describes the change caused by distortion of the configurational energy stored in magnetic and thermodynamic energies. An isolated system can only evolve with its total energy (including additional positive kinetic energy) conserved. Hence, if the MHD energy of the configuration increases for every possible distortion, it is clear that such perturbations are prohibited from spontaneously evolving to large amplitudes since the system will have more energy than in the equilibrium state. However, if MHD energy is negative upon distortion, it is energetically possible for the system to evolve from small amplitudes, with the excess energy taken up by the kinetic energy of the flow.

5.1.3 Role of Field-Line Curvature. From the detailed expression for the MHD energy, one can show that the field-line curvature plays a major role in determining MHD stability. Consider, for example, the plasma contained in the magnetic configurations shown in Fig. 4. In Fig. 4(a) the field lines are convex with respect to the contained plasma, while in Fig. 4(b) the field lines are concave. The MHD energy principle demonstrates that the configuration shown in Fig. 4(a) is unfavorable for stability, while the configuration shown in Fig. 4(b) is highly favorable to stability. For a low-beta system this can be shown to be equivalent to saying that systems where $|B|$ decreases outwardly are unfavorable to stability, while systems where $|B|$ increases outwardly are favorable to stability. Intuitively, one can argue that a plasma with a pressure p and mass density nM has a mean conserved magnetic moment $\mu \approx p/nMB$, so that the force density on the plasma is $\mathbf{F} = -n\mu\nabla|\mathrm{B}|$. For increasing $|B|$, as one moves away from the confined region, there is an inward force that restores plasma perturbations, while if $|B|$ decreases outwardly the force is outward and causes perturbations to convect outwardly.

Magnetic mirror machines offer a particularly insightful example of the importance of field-line curvature. In a symmetric mirror machine the field-line curvature is always unfavorable for stability at the mid-plane. This almost always causes symmetric mirror machines to be MHD unstable except under very restricted conditions.

To obtain a concave-outward field-line curvature system, as shown in Fig. 4(b), one needs to forsake azimuthal symmetry. For example, two axially aligned Helmholtz coils with currents flowing in the same direction can establish a simple mirror machine with unfavorable field-line curvature in the midplane. Placing four current-carrying rods parallel to the coils' axis but radially displaced and with the currents in the rods alternating in direction, as shown in Fig. 4(b), leads to a magnetic field with favorable concave field-line curvature properties nearly everywhere.

Experimentally, the entire character of plasma containment in mirror machines was decisively improved once minimum-B fields were used that prevented the extremely disruptive curvature-driven mode from being excited.

5.2 Deviation from MHD Stability Guidelines

The need to break azimuthal symmetry leads to several technological side problems such as stress problems at high fields, difficulty of access (e.g., for the injection of neutral beams), increased single-particle orbit loss, etc. Hence, considerable interest has remained in devising systems that are basically simple mirrors that maintain azimuthal symmetry. It is conjectured that instability can be attained by creating plasmas that do not obey the dynamics of an ideal MHD system.

5.2.1 Passive Current Carriers. The most common idea is to place an *in situ* current carrier within the plasma where the current carrier does not respond to the displacements of the plasma. The energetic particle beam in the Astron concept is a proposed example of one such current carrier. A passive current carrier can create a minimum-B field in a symmetric configuration, which can then produce stability in the MHD response of the background plasma. The crucial question is whether the current carrier indeed acts as a *passive* current carrier, and this question needs further study. There are numerous experimental studies in mirror machines and tokamaks that show stability enhancement with energetic particles in the system (e.g., Perkins and Post, 1963; Dandl *et al.*, 1971; JET Team, 1991). There are also regimes where new instabilities appear to be caused by the energetic particles.

5.2.2 Finite–Larmor-Radius Stabilization. Another idea to achieve stability in a symmetric mirror machine is to exploit the finite orbit size of the bulk plasma ions, which can lead to a plasma response that differs from MHD predictions and to so-called finite–Larmor-radius stabilization (Rosenbluth *et al.*, 1960). This mechanism readily stabilizes the shorter-wavelength modes, but has difficulty in stabilizing longer-wavelength modes. In fact, a mode that corresponds to a nearly rigid displacement of the plasma column is extremely difficult to stabilize. This mode can be stabilized either at extremely low densities (far below densities of interest in fusion) or at relatively high β where currents induced in neighboring conducting walls can serve as a passive feedback mechanism to restore the rigid displacements of the column. At intermediate densities active feedback mechanisms or ponderomotive force from radio-frequency fields is needed to achieve stability.

5.2.3 Migma Concept. An example of a finite–Larmor-radius concept is the Migma (Maglich, 1988) experiment. Here the machine size is roughly two ion orbits and all the ion orbits pass close to the axis of the machine. In this device, stability at low density has been achieved, but it is below the threshold density predicted for the onset of the unstable displacement mode. Further experiments are needed to test stability for the predictions of finite orbit theory.

5.3 Stability in Tori

In toroidal devices MHD stability based on the constraints imposed by field-line curvature is very important. The stability criteria are somewhat subtle to apply. The curvature direction alternates in sign, being unfavorable to stability at major radii R greater than that of the magnetic axis R_0, while it is favorable to stability at major radii R less than R_0. The net result depends on the average of the two effects.

5.3.1 Line-Average Stabilization. For a low-β circular tokamak, average favorable curvature arises if the safety factor satisfies $q > 1$, while unfavorable average curvature arises if $q < 1$. Tokamaks generally operate with most of the flux tubes having $q > 1$. In fact, in the center where $q \lesssim 1$, periodic bursts of instability are usually observed, so-called sawtooth oscillations. Though these instabilities are driven by the unfavorable curvature, ideal MHD theory usually does not properly predict them. They are usually a manifestation of resistive instability, which we shall shortly explain.

5.3.2 Beta Stability Limit. When both good and bad curvatures exist on a field line, unstable perturbations try to localize in the bad-curvature regions. This causes the field lines to bend, which according to the energy principle induces a stabilizing "bending energy." At sufficiently low β, ideal MHD theory predicts stability for a toroidal system, since the destabilizing energy perturbation driven by the unfavorable curvature cannot compensate the bending energy. Only above a critical beta value β_c can a perturbation be localized to the bad-curvature region and overcome the stabilizing bending energy. The result is that overall stable operation is limited to beta values below β_c, commonly called the Troyon limit (Troyon *et al.*, 1984), which is typically $\beta_c \approx 2I_n$ (with $I_n = \mu_0 I_p / aB_\phi$, I_p the plasma current, a the minor radius, and B_ϕ the toroidal field).

5.3.3 RFP Stability. If a perturbation has a given pitch along the torus, this pitch can be identical to the rotational transform on one magnetic surface. On that surface there is no bending energy produced by the perturbation. However, if there is magnetic shear, i.e., q varies with the flux surface ψ (i.e., $\partial q/\partial\psi\neq 0$), bending energy is produced on neighboring field lines which allows for a stable MHD response for beta values below the appropriate critical β_c. Only if there is a flux tube where $\partial q/\partial\psi=0$ can MHD instability occur at extremely low beta. In pinches, where $q\ll 1$, there will be a surface where $\partial q/\partial\psi=0$ if the toroidal magnetic field does not reverse, and this region is a source of virulent instability. When toroidal field reversal is experimentally achieved, there is a dramatic decrease in the magnetic fluctuation levels. As a result the reversed-field pinch (RFP), where toroidal field reverses direction inside the plasma, is considered the most interesting low-q toroidal configuration.

5.4 Resistive Stability

If plasma displacements can be localized to bad-curvature regions without inducing the stabilizing bending energy, then curvature-driven modes can be excited at low-β values. In ideal MHD theory, the ideal Ohm's law, $\mathbf{v}\times\mathbf{B}+\mathbf{E}=0$, with $\mathbf{v}$ the plasma flow speed, forces stabilizing bending energy to be induced when a displacement is localized on a flux tube. With finite resistivity η Ohm's law is altered to

$$\mathbf{v}\times\mathbf{B}+\mathbf{E}=\eta\mathbf{J},$$

with $\mathbf{J}$ the plasma current. The finite resistivity then allows plasma to slip through magnetic field lines. As a result the plasma displacement can be localized to the unfavorable-curvature region, without inducing stabilizing bending energy (Furth *et al.*, 1963). Thus, because of plasma resistivity, instability can arise from the unfavorable curvature even at low β.

When this mechanism applies, instability growth rates are much slower than the growth rates typical of MHD instability. This mechanism gives rise to the sawtooth magnetic signals inside the $q=1$ surface of tokamaks. The instability causes rapid thermal mixing of the plasma inside the $q=1$ surface. However, as the plasma is stable outside the $q=1$ surface, the integrity of the thermal insulation is maintained on the outside of the plasma. In fact the sawtooth oscillations have the favorable feature of preventing high-Z impurities from collecting in the center of these discharges, which would cause severe energy losses from radiation.

The RFP is also prone to resistive instability, and the level of magnetic fluctuations is much larger in the RFP than in the tokamak. Here instability tends to be caused throughout the plasma as q is much less than unity everywhere. There is a concern whether the thermal insulation properties needed for fusion can ultimately be achieved in an RFP with this instability drive present.

In addition to the curvature instability drive, the current profile itself is a source of instability. Here again resistivity allows the current profile drive to manifest itself at low beta, whereas MHD theory would predict stability. In RFP plasmas these current-drive instabilities force the configuration to relax close to the lowest energy state consistent with the amount of magnetic flux (or, as it is frequently called, magnetic helicity) pumped into the system. This state is one with toroidal field reversed in the inside of the plasma relative to the outside, and it is maintained indefinitely with a fixed amount of input voltage. Simple diffusion theory for a symmetric equilibrium predicts that a reversed-field state cannot be continuously maintained. However, a three-dimensional dynamolike effect that is driven by magnetic turbulence is what allows the maintenance of a reversed toroidal field (Taylor, 1986).

5.5 Microinstability

Another class of instabilities of importance, quite different from MHD-driven instabilities, is referred to as microinstabilities. They generally are of short wavelength (compared to the system size) and their fluctuations do not generally lead to catastrophic disassembly, but rather to an enhancement of transport. Even in MHD-favorable systems, it is the microinstabilities that often determine the intrinsic containment time of a plasma system.

A homogeneous plasma at a fixed global temperature T is, of course, in thermodynamic equilibrium and is free of instability. Any departure from thermodynamic equilibrium leads to intrinsic "free energy" drives that have the potential to cause instability and thereby to cause relaxation to the thermodynamic equilibrium state at a rate more rapid than predicted by binary collisions.

5.5.1 Velocity-Space Instability. The farther away a system is from thermodynamic equilibrium, the more virulent the drive of the free energy. In simple mirror machines where the ion kinetic distribution is not Maxwellian, there is a velocity-space containment condition that must be satisfied, which is given by Eq. (24). As a consequence the distribution function will have an "inverted" population of particles such that the phase-space density of low-energy particles is less than that of higher-energy particles. Such a situation gives rise to a laserlike instability (generally called loss-cone instability), in which the inverted distribution tries to relax through coherent oscillations (generally at harmonics of the ion cyclotron frequency) to a "monotonic" state where the phase-space density is a decreasing function of energy. However, as the low-energy particles are not contained, rapid loss arises on the scale of the particle bounce time.

In order to sustain a simple mirror plasma it was found necessary to supply low-energy particles, either through a gas feed or through a low-energy particle gun. In the 2X-IIB experiment it was found that with these techniques it was possible to achieve a high-β steady-state plasma. Particles that were injected at high energy were contained for a classical slowing-down time. The low-energy phase-space region was filled primarily with plasma arising from the external source, which either prevented instability or caused a low enough fluctuation level that it heated slightly the low-energy source but did not greatly perturb the energy-containment properties of the high-energy particles. The limitation of this method was that the electron temperature was constrained to remain low (typically $T_e/T_i \approx 0.01$ with $T_e \lesssim 100$ eV) and the high-ion-energy lifetime could not exceed several milliseconds.

One of the pleasing features of the tandem mirror configuration is that the loss-cone instability can be prevented. In the center of a tandem mirror the ions do not have a loss cone because these particles are contained by the electric field. In the end cells the loss-cone drive is quenched by the outflow of particles from the center. As a result tandem mirrors routinely operate without loss-cone instability.

5.5.2 Drift-Wave Instability. In toroidal plasmas one is generally not concerned with velocity-space instabilities. However, as intrinsic temperature and density gradients must exist to isolate the system thermally from the walls, the system cannot be in global thermodynamic equilibrium. The intrinsic gradients give rise to a diamagnetic flow rate, which from the equilibrium equations is

$$\mathbf{J}_j = \mathbf{B} \times \nabla p_j / B^2 \equiv e_j n_j \mathbf{v}_{Dj},$$

where the species j can refer to electrons or ions and v_{Dj} is the diamagnetic drift frequency of species j. In a frame moving at the equilibrium $\mathbf{E}\times\mathbf{B}$ speed, $\mathbf{v}_{De}$ and $\mathbf{v}_{Di}$ would be the mean flow speed of electrons and ions, respectively. These flows are in opposite directions and it is impossible to find a reference frame where the overall flow of every species can be ignored. These equilibrium flows are the free-energy source for instability. If we consider a perturbation with a spatial dependence proportional to $\exp(i\mathbf{k}\cdot\mathbf{r})$, a low-frequency response will arise near the angular frequency $\omega \sim \mathbf{k}\cdot\mathbf{v}_{Dj}$. It is found that when instability arises at this frequency, it causes anomalous relaxation of the plasma. Often the magnitude of the growth rate γ approaches that of the real frequency.

5.6 Turbulent Diffusion

The rate of diffusion caused by drift instabilities can be estimated by the following mixing-length argument. A wave with an electric field amplitude $\delta\mathbf{E}$ will grow until the $\delta\mathbf{E}\times\mathbf{B}$ drift velocity of the wave forces the plasma to displace a wavelength in a growth time, i.e.,

$$\delta\mathbf{r} \simeq \delta\mathbf{E}\times\mathbf{b}/\gamma B \approx 1/k,$$

where $\delta\mathbf{r}$ is the plasma displacement by the wave. The resulting wave breaking then forces saturation. Thus, we have a random motion $\delta r \sim 1/k$ in a correlation time $\tau = 1/\gamma \approx 1/kv_{Dj}$. Taking the diffusion coefficient to be $D_{\text{anom}} \approx \delta r^2/\tau$ gives

$$D_{\text{anom}} \approx \frac{\gamma}{k^2} \approx \frac{1}{e_j kBn} \frac{\partial(nT_j)}{\partial r}.$$

If $k \sim 1/r$, this diffusion coefficient yields so-called Bohm diffusion (taking $r\,\partial/\partial r \sim 1$)

$$D_{\mathrm{Bohm}} \simeq T_j/eB.$$

Such diffusion produced by drift waves is possible in a shearless system. In tokamaks and other toroidal devices, the magnetic shear often forces the wave number to scale as the ion Larmor radius ρ_i, giving rise to a modified Bohm scaling

$$D_{\mathrm{MB}} \approx (T_j/eB)(\rho_i/L_p)$$

with $L_p^{-1} = \nabla|nT_j|/nT_j$.

The interplay among the number of degrees of freedom is sufficiently complicated that a definitive description of turbulent transport still needs to be developed, and it is currently an extremely active research area (Callen, 1990).

6. STATUS OF MAGNETIC FUSION RESEARCH

6.1 Historical Overview

The quest to achieve controlled fusion developed soon after the successful development of the hydrogen bomb in the 1950s. The rapid success of new scientific technology, including fission power, produced great optimism that controlled fusion could be achieved relatively quickly. However, the instability problems in plasma containment were not fully appreciated. Indeed, many research groups initially inferred the achievement of thermonuclear conditions when neutrons were produced in experiments. It usually turned out that these neutrons resulted from runaway ions produced by the large electric fields associated with the turbulent disruption of the plasma.

By the early 1960s a general pessimism pervaded the program, as plasma confinement in nearly all devices was poor, with confinement properties often being attributed to Bohm diffusion. During this difficult period the subtleties needed to achieve good confinement began to be appreciated.

Major breakthroughs developed, often due to the understanding of "kitchen" physics, i.e., the solving of problems associated with wall conditioning, impurity control, vacuum conditions, etc., and to the understanding of how to achieve configurations that would be stable to gross instabilities. Experimental results demonstrated relatively long-lived containment in a variety of concepts. The experimental concept that received the greatest attention was the tokamak, which has become the leading magnetic fusion concept and is at present the most developed. There have been important breakthroughs in other concepts as well, including mirror machines, RFP, compact toroids, and stellarators.

By the mid 1970s the success of multiple approaches to achieving fusion and the appreciation of the need to develop a long-term energy source led to a broad support by the U.S. government of the many approaches in magnetic fusion. However, budgetary constraints became important in the U.S. fusion program in the 1980s. As a result, most of the research programs not connected to tokamaks were terminated, often before scientific issues were clarified. Thus, crucial questions in ascertaining the potential usefulness of many of the approaches remain unanswered. In some cases they can be addressed in fusion programs outside the U.S., where significant research facilities exist in stellarators, RFP, and mirror machines.

To present an overview of progress in magnetic fusion in somewhat more detail, we shall give a brief status report of the approaches that are alternative to tokamaks and then present a more detailed status report of the tokamak concept.

6.2 Status of Non-Tokamak Concepts

6.2.1 Mirror Machines. Initial mirror-machine experiments were in symmetric mirror fields, and if the density was not too low, they demonstrated poor confinement because of the destabilizing curvature-driven instability. A major breakthrough occurred (Ioffe and Sobolev, 1964) when it was demonstrated that the placement of quadrupole coils to form an MHD-stable minimum-B configuration resulted in an order of magnitude increase in the containment time.

The next major breakthrough in mirror research was the demonstration that a high-beta plasma can be sustained by neutral-beam injection (Coensgen *et al.*, 1976). This experiment also demonstrated the importance of neutral-beam technology, which was rapidly applied to tokamak and stellarator research.

An appreciation that the energy gain of a mirror machine under fusion conditions needed to be improved then led to the invention of the tandem mirror concept (Dimov *et al.*, 1976; Fowler and Logan, 1977). The experimental group at Lawrence Livermore Laboratory demonstrated feasibility of the tandem mirror and, in particular, showed that loss-cone instability can be completely suppressed as first suggested by theory (Post and Rosenbluth, 1966). The attempt to improve the tandem mirror concept by creating thermal barriers proved to be illusive. At present the GAMMA-10 tandem mirror, the only tandem mirror which is in operation in 1992, in Tsukuba, Japan, has shown that at relatively low density, quiescent long-lived tandem mirror plasmas can be obtained with thermal-barrier operation (Miyoshi *et al.*, 1991). The major improvement needed is to increase the density and the lifetime in thermal barrier operation.

Though simple mirror machines have marginal applicability as an energy source, they may have a special uniqueness for application as a 14-MeV neutron source which will be needed for testing the properties of materials that surround a fusion reactor (Berk and Ryutov, 1990). A *d-t* plasma in a mirror machine can supply neutron intensity to achieve a needed fluence in the shortest possible time and at relatively low cost, because a high-β (which allows a high intensity) and small-volume (which reduces overall cost) plasma can be achieved. Currently, studies of this application are ongoing in several laboratories, and may lead to a resurrection of interest in mirror-machine research.

6.2.2 Reversed-Field Pinch. Many different high-β pinch concepts were examined in the 1960s. The most successful ultimately turned out to be the Zeta experiment (which originally was designed as a stabilized *z* pinch), where a quiescent operational region was discovered (Robinson and King, 1969). In this regime the toroidal magnetic field in the center was reversed from its direction at the edge. Surprisingly, the reversal could be sustained by toroidal voltage, despite the prediction of two-dimensional diffusion theory that it would not be sustained.

Though the Zeta experiment was terminated, perhaps prematurely in light of its discovery, the reversed-field pinch configuration has now been studied in many facilities worldwide. These experiments have demonstrated reduced fluctuation level and improved containment with increasing temperature. These results are consistent with the theoretical understanding that the plasma will try to relax towards its lowest-energy state and that the relaxation process allows a steady plasma to be maintained through turbulent dynamo processes. In principle this mechanism can be extended to achieve steady operation with non-Ohmic helicity injection.

Though magnetic fluctuations in an RFP are low compared to most other high-β pinches, the fluctuations are appreciably higher than in tokamaks, and are present close to the plasma edge. Such fluctuations lead to impurity buildup, heat loading problems, and the penetration of magnetic fields into the walls, which in turn enhances fluctuations. It is foreseen that by increasing the current and temperature of an RFP, the fluctuation level will decrease, as the rate of relaxation needed to remain close to the lowest-energy state is reduced. The verification of this conjecture is the principal goal of the RSX experiment in Padua, Italy, which will have a plasma current of 4 MA (Antoni *et al.*, 1989).

6.2.3 Compact Toroids. Another offshoot of pinch research was the study of compact toroids. Early on it was found that by programming of the wall currents in a theta-pinch experiment, the *z*-directed magnetic fields could be reversed into an elongated compact toroid without any apparent toroidal magnetic field. However, these experiments would typically disrupt because of the onset of a rotational instability. By appropriate preparation it was demonstrated (Es'kov *et al.*, 1979) that elongated field-reversed theta pinches can be made free of the rotational instability. After a great deal of study these results were confirmed by other experimental teams. In the 1980s it was shown in Japan that minimum-*B* quadrupole fields could be used to ensure MHD stability (Ishimura *et al.*, 1989).

An alternative approach to compact toroids started in the late 1970s with the study of the spheromak, a compact toroid which has both toroidal and poloidal magnetic fields. Several novel ideas were developed to form plasmas using pulsed techniques. Both the poloidal magnetic flux and toroidal magnetic flux need to be embedded in the plasma during its formation phase, and then the plasma has to be guided into a storage region where it is contained. In the more successful experiments a relatively high electron temperature was attained, $T_e > 100$ eV (Yamada *et al.*, 1989), and external magnetic helicity injection was used to sustain the configuration (Jarboe *et al.*, 1983).

Perhaps the first compact toroid conceived was the Astron concept (Christofilos, 1962). Here field reversal is to be achieved with the current produced by energetic particles whose Larmor radius is comparable to the radius of the plasma. The demonstration of field reversal has been achieved (Andrews *et al.*, 1971), but there is still difficulty in achieving hot plasma containment in an Astron configuration.

One possible future direction in compact toroid research is to merge the Astron and field-reversed theta-pinch concepts. The joint system should stabilize the so-called tilt instability (Rosenbluth and Bussac, 1979), an MHD instability that might otherwise prevent the achievement of reversed-field theta-pinch devices with relatively small Larmor radii.

Finally, it should be noted that the reversed-field theta pinch is an intrinsically high-β device. If suitable plasma confinement can be successfully achieved, the configuration would be especially suitable for advanced fuel cycles, such as d-^{3}He, since synchrotron radiation losses would be insignificant and the fusion power production density would be high, while the total power production would be moderate because of the system's compactness (Kernbichler *et al.*, 1991).

6.2.4 Stellarators. Progress in stellarators was initially very slow, in part because of the complications in designing three-dimensional magnetic fields. For example, it was found that stellarators were frequently built improperly without the existence of magnetic flux surfaces. Careful field design by the research team at the Max-Planck-Institut für Plasmaphysik led to more favorable stellarator containment by the late 1960s (Berkl *et al.*, 1969). Further innovations by the Kyoto group in Japan were successful (Uo *et al.*, 1977). With these studies research teams found that by heating with rf fields or with high-energy neutral beams, confinement times could be achieved that were similar to those in comparable size tokamaks. Further, it was shown that the radial electric field could be controlled to improve particle containment significantly, even that of the injected high-energy particles.

The stellarator has considerable flexibility of design, and indeed new interesting designs emerge every few years. Some of the more interesting magnetic-field design concepts that have emerged are (Grieger *et al.*, 1992) modular magnetic coils that allow for ready disassembly for maintenance, and magnetic field design that can minimize the particle excursion from a flux tube, produce large stabilizing magnetic shear, produce favorable minimum-B properties, etc.

Stellarators are now the principal backup to tokamaks. The often-cited advantages of stellarators over tokamaks are that stellarators have been free from the catastrophic disruptive loss to which tokamaks are susceptible and that there is no need for an elaborate current drive mechanism. The complexity of a stellarator is also its virtue as the optimum stellarator is probably yet to be designed. A large stellarator (LHD) is now under construction in Nagoya, Japan, and plans for a large stellarator exist in Garching, Germany.

6.3 Status of Tokamak Research

Tokamak research achieved a major breakthrough in 1968 when a favorable regime of operation was reported (Artsimovich *et al.*, 1969). The favorable regime was rather narrow, with too strong a current producing violent MHD instabilities since $q < 1$ on axis, whereas too weak a current could not produce enough power to sustain the discharge. The discovery of a window of favorable operation truly changed the direction of controlled-fusion research. Many groups throughout the world converted to tokamak research.

A great deal of progress has occurred since the initial breakthrough was reported in 1968. Some of the crucial developments are described below.

6.3.1 Impurity Control and Fueling. The development of wall preparation and fueling techniques has mitigated the impurity problem in tokamaks. Initially tokamaks operated with a high Z_{eff} ($\sim$5–10), whereas $Z_{\mathrm{eff}} < 2$ is now standard in tokamak operation. Recently, technology has been developed for injecting pellets that allows for central fueling which often leads to substantially improved confinement characteristics at low Z_{eff}. However, the issue of whether impurity buildup can be prevented over extremely long pulse times (> 10 sec) and under high heating power remains crucial and will be investigated in future experiments.

6.3.2 Auxiliary Heating. Auxiliary heating is generally necessary at high temperature as Ohmic heating loses its effectiveness because, at fixed current, the heating rate is proportional to $T_e^{-3/2}$. Only if large plasma current densities are present, which can occur in high-field machines with $B_\phi > 10$ T, is it conceivable that Ohmic heating can bring a plasma to ignition conditions. The advances in neutral-beam and radio-frequency technologies have made possible the achievement of the plasma temperature needed for producing fusion. For example, the TFTR experiment at Princeton, New Jersey, and the JET experiment in Culham, England, have produced plasmas with ion temperatures in excess of 20 keV and electron temperatures of ~10 keV. In the TFTR experiment the best confinement characteristics are obtained in the "supershot" regime (Bell *et al.*, 1989), where neutral beams directly fuel the center of the discharge. Neutral beams have also been used to simulate collective alpha-particle effects of a fusion-producing plasma (Wong *et al.*, 1991; Heidbrink *et al.*, 1991). In the JET experiment, high-energy tails have been generated with ion cyclotron heating that simulate the effect of charged energetic fusion products in the plasma (JET Team, 1991). In this experiment sawtooth instability is eliminated, perhaps as a result of hot-particle stabilization of MHD instability.

6.3.3 Plasma Shaping. The shaping of a tokamak plasma has become a crucial feature in tokamak performance. The initial tokamak experiments were with circular-shaped poloidal cross sections which lead to relatively low critical beta values for the onset of MHD stability. By means of plasma shaping, MHD stability properties can be substantially improved. Recently the D-III-D tokamak, a shaped tokamak at General Atomic in San Diego, California, has achieved peak beta values of 40% and an average beta value of ~10% (Lazarus *et al.*, 1991). It is believed that this plasma can test the so-called second stability regimes (Coppi *et al.*, 1979), where stability can be attained in a tokamak even at high β. If such a high-β stability regime can indeed be achieved it makes d-^{3}He operation feasible in a tokamak.

In addition, with shaping it is easier to cause field lines near the edge to be deliberately diverted to the wall (such field-line design is called a divertor). The purpose is to decouple the properties of the plasma–wall interactions from the properties of the confined plasma and thereby prevent impurity buildup in the plasma. In addition, it is found that divertor operation facilitates H-mode operation, where the plasma develops a sharp pressure gradient at the edge and the overall plasma confinement increases by a factor of ~2–3 from the usual operating scenario (Jackson *et al.*, 1992).

Another type of shaping is for the current profile. It has recently been observed (Zarnstorff *et al.*, 1991) that there is better confinement if the current profile is altered from the natural profile reached in steady operation. Techniques to develop favorable current profiles in steady state need to be developed.

6.3.4 Current Drive. Noninductive current-drive techniques have made considerable progress. Tokamak currents have been successfully driven by rf sources (at lower hybrid and electron cyclotron frequencies) and by neutral beams. At low plasma currents, these current-drive mechanisms allow long-time plasma operation in large and small machines (Jacquinot *et al.*, 1992). However, these current-drive schemes are intrinsically less efficient than the inductive (Ohmic) drive methods, and the efficiency decreases at higher densities. In reactor operation at relatively high β, the requirements of current drive are lowered by the large bootstrap current induced, and it is conceivable that an acceptable current-drive efficiency can be achieved (Cordey, 1988).

In addition to the above-mentioned current-drive mechanisms there are alternative ideas being investigated, such as magnetic helicity injection (Jarboe *et al.*, 1983), which was first studied in RFP research.

Without current drive the duty cycle of a tokamak reactor needs to be high, and thermal stresses, due to wall cooling and heating as the tokamak is turned on and off, need to be minimized. The JET facility has recently shown that the plasma current can be reversed by ac cycling of the Ohmic drive (Tubbing *et al.*, 1992). They demonstrated that the time interval in which no plasma equilibrium exists can be short. This mode of operation can conceivably mitigate the heat-stress problem.

6.3.5 Achieved Confinement Parameters. The historical progress in tokamak containment can be observed in Fig. 13, which presents values of the parameter $n\tau T$ achieved in various tokamaks during the last two decades. Note that the quality of plasma confinement now being achieved in tokamaks is nearly equivalent to achieving energy breakeven with the *d-t* fusion reaction. In addition, the JET facility (JET Team, 1992) has now actually operated with *d-t*, producing $Q{\sim}0.15$ in a lean *d-t* mixture ($Q{\sim}0.45$ would have been produced in an optimum *d-t* mixture in the same experiment; in addition, plasma confinement conditions that would have produced $Q > 1$ in an optimum *d-t* mixture were produced in a *d-d* discharge). Similar runs are being planned for TFTR.

6.4 Conclusion

The advances in tokamak research have reached the point where the next major step is to build an ignition experiment and to study the system during its burning phase. It is proposed that this experiment be run by an international collaboration that is now called ITER (Tomabechi *et al.*, 1991). In addition, work on smaller tokamaks is necessary in order to understand physical phenomena better and to find ways to improve the intrinsic operation of tokamaks. As the advances in tokamaks have often occurred in unplanned areas, substantial improvements can certainly be expected.

Research in alternative confinement schemes has not had support comparable to tokamaks, and it is not surprising that their achievements lag behind the tokamak. Some of these schemes may evolve to reactor-grade concepts, while others might be highly favorable for nonenergy-producing applications such as a compact neutron source for materials testing or a compact source for producing medically desired radioactive isotopes. Re-

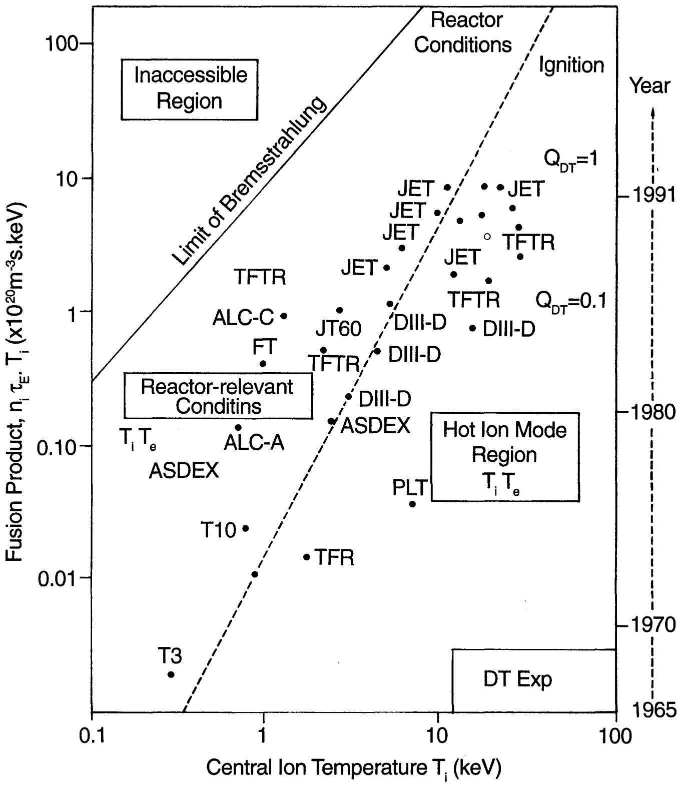

FIG. 13. Fusion parameter $n\tau_E T_i$ vs central ion temperature $T_i(0)$ for a number of machines worldwide in the period 1965–1991 (figure published with permission from JET Joint Undertaking, which retains copyright of the material).

search in these alternative concepts has frequently contributed significantly to the overall progress in the physics of plasma containment and to the generation of technological tools. Hence, the controlled fusion program will continue to profit by continuing its support of alternative directions of research.

GLOSSARY

Ambipolar Potential: The electrostatic potential that a plasma establishes for maintaining overall charge neutrality.

Anomalous Transport: The transport in a plasma due to self-generated electromagnetic fluctuations.

Banana Orbit: The shape of an orbit of particles in a tokamak that are trapped by the increasing magnetic field.

Beta: The ratio of kinetic pressure of a plasma to the magnetic field pressure.

Classical Transport: A theoretical description of transport of a plasma in a uniform magnetic field.

Compact Toroid: A confinement concept of a toroidal plasma without magnetic field coils through its center.

Lawson Criterion: The product of density and plasma lifetime, which is a parameter that measures the quality of plasma confinement with respect to what is needed to achieve fusion conditions.

Magnetic Flux Surfaces: The surface on which magnetic field lines lie in a toroidal confinement concept.

Magnetohydrodynamics (MHD): The fluid description of ionized gases.

MHD Energy: The perturbed energy of a system described by the MHD equations.

Mirror Machine: A confinement concept consisting of open-ended magnetic field lines with the magnetic field intensity increasing towards the ends.

Neoclassical Transport: A collisional transport theory that is based on the finite orbit size due to drift orbit excursions in tokamaks and other plasma confinement machines.

Pinch: A generic plasma confinement configuration with large plasma currents.

Plasma: Ionized gas.

Reversed-Field Pinch: A pinch device where the toroidally directed field reverses between the plasma center and the edge.

Rotational Transform: The number of times a field line winds around the poloidal cross section of a torus for each time it winds around the toroidal cross section.

Safety Factor (q): The inverse of the rotational transform; $q > 1$ generally indicates stability in tokamaks, while $q < 1$ indicates a likelihood of instability.

Stellarator: A three-dimensional magnetic configuration that does not require large plasma currents to achieve plasma containment.

Tokamak: A pinchlike configuration with a strong toroidal magnetic field and a strong toroidal current carried by the plasma.

Listing of Works Cited

Andrews, M. L., Davitian, H., Fleischmann, H., Kusse, B., Kribel, R. E., Nation, J. A. (1971), *Phys. Rev. Lett.* **27**, 1428–1431.

Antoni, V., Bagatin, M., Baseggio, E., Bassan, M., Buffa, A., Cappello, S., Carraro, L., Costa, S., Giudicotti, L., Innocente, P., Martin, P., Martini, S., Merlin, D., Ortolani, S., Paccagnella, R., Pasqualotto, R., Puiatti, M. E., Scapin, M., Scarin, P., Valisa, M., Villoresi, P., Zago, S., Zilli, E. (1989), *Plasma Physics and Controlled Nuclear Fusion Research 1988*, Vienna: IAEA, Vol. 1, p. 661–669.

Artsimovich, L. A., Bobrovskii, G. A., Gorbunov, E. P., Ivanov, D. L., Kirillov, V. D., Kuznetzov, E. I., Mirnov, C. V., Petrov, M. P., Razumova, K. A., Strelkov, V. C., Sheglov, D. A. (1969), *Plasma Physics and Controlled Nuclear Fusion Research,* Vienna: IAEA, Vol. 1, pp. 157–173.

Asdex Team (1989), *Nucl. Fusion* **29**, 1959–2040.

Bell, M. G., Arunasalam, V., Barnes, C. W., Bitter, M., Bosch, H.-S., Bretz, N. L., Budny, R., Bush, C. E., Cavallo, A., Chu, T. K., Cohen, S. A., Colestock, P. L., Davis, S. L., Dimock, D. L., Dylla, H. F. Efthimion, P. C., Ehrhardt, A. B., Fonck, R. J., Fredrickson, E. D., Furth, H. P., Gammel, G., Goldston, R. J., Greene, G. J., Grek, B., Grisham, L. R., Hammett, G. W., Hawryluk, R. J., Hendel, H. W., Hill, K. W., Hinnov, E., Hosea, J. C., Howell, R. B., Hsuan, H., Hulse, R. A. Jaehnig, K. P., Janos, A. C., Jassby, D. L., Jobes, F. C., Johnson, D. W., Johnson, L. C., Kaita, R., Kieras-Phillips, C., Kilpatrick, S. J., Krupin, V. A., LaMarche, P. H., Langer, W. D., LeBlanc, B., Little, R., Lysojvan, A. I., Manos, D. M., Mansfield, D. K., Mazzucato, E., McCann, R. T., McCarthy, M. P., McCune, D. C., McGuire, K. M., McNeill, D. H., Meade, D. M., Medley, S. S., Mikkelsen, D. R., Motley, R. W., Mueller, D., Murakami, Y., Murphy, J. A., Nieschmidt, E. B., Owens, D. K., Park, H. K., Ramsey, A. T., Redi, M. H., Roquemore, A. L., Rutherford, P. H., Saito, T., Sauthoff, N. R., Schilling, G., Schivell, J., Schmidt, G. L., Scott, S. D., Sinnis, J. C., Stevens, J. E., Stodiek, W., Stratton, B. C., Tait, G. D., Taylor, G., Timberlake, J. R., Towner, H. H., Ulrickson, M., Von Goeler, S., Wieland, R. M., Williams, M. D., Wilson, J. R., Wong, K.-L., Yoshikawa, S., Young, K. M., Zarnstorff, M. C., Zweben, S. J. (1989), *Plasma Physics and Controlled Nuclear Fusion Research 1988*, Vienna: IAEA, Vol. 1, pp. 27–40.

Berk, H. L., Ryutov, D. D. (1990), *Comments Plasma Phys. Controlled Fusion* **13**, 173–190.

Berkl, E., Eckhartt, D., Gierke, G. v., Grieger, G., Hinnov, E., Hagenow, K. U., Ohlendorf, W. (1969), *Plasma Physics and Controlled Nuclear Fusion Research 1968*, Vienna: IAEA, Vol. 1, p. 513–528.

Callen, J. D. (1990), *Phys. Fluids B* **2**, 2869–2960 and related articles in this issue.

Callen, J. D. (1992), *Phys. Fluids B* **4**, 2142–2154.

Carreras, B. A., Greiger, G., Harris, J. H., Johnson, J. L., Lyon, J. F., Motojima, O., Rau, F., Renner, H., Rome, J. A., Uo, K., Wakatani, M., Wobig, H. (1988), *Nucl. Fusion* **28**, 1613–1627.

Christofilos, N. C. (1962), *Nucl. Fusion Suppl., Part 1*, 159–168.

Coensgen, F. H., Cummins, W. F., Gormezano, C., Logan, B. G., Molvik, A. W., Nexsen, W. E., Simonen, T. C., Stallard, B. W., and Turner, W. C. (1976), *Phys. Rev. Lett.* **37**, 143–146.

Coppi, B., Ferreira, A., Mark, M. W. K., Ramos, J. J. (1979), *Nucl. Fusion* **19**, 715–725.

Cordey, J. G. (1988), *Plasma Phys. Controlled Fusion* **30**, 1625–1635.

Dandl, R. A., Eason, H. O., Edmonds, P. H., England, A. C. (1971), *Nucl. Fusion* **11**, 411–423.

Dimov, G. I., Zakaidakov, V. V., Kishinevskii, M. E. (1976), *Fiz. Plasmy* **2**, 597–610.

Drake, J. R., Brunsell, P., Brzozowski, J., Eninger, J. E., Hedin, E. R., Karlsson, P., Lehnert, B., Li, Jin, Scheffel, J., Sätherblom, H. E., Tennfors, E., Wilner, B. (1989), *Plasma Physics and Controlled Nuclear Fusion Research 1988*, Vienna: IAEA, Vol. 2, pp. 751–756.

Es'kov, A. G., Kurtmullaev, R. K., Kreshchuk, A. P., Laukhin, Ya. N., Malyutin, A. I., Markin, A. L., Martyushov, Yu. S., Mironov, B. N., Orlov, M. M., Proshletsov, A. P., Semenov, V. N., Sosunov, Yu. B. (1979), *Plasma Physics and Controlled Nuclear Fusion Research 1978*, Vienna: IAEA, Vol. 2, pp. 187–204.

Fisch, N. J. (1987), *Rev. Mod. Phys.* **59**, 175–341.

Fowler, T. K., Logan, B. G. (1977), *Comments Plasma Phys. Controlled Nucl. Fusion Res.* **2**, 167–172.

Freidberg, J. P. (1982), *Rev. Mod. Phys.* **54**, 801–902.

Furth, H. P., Killeen, J., Rosenbluth, M. N. (1963), *Phys. Fluids* **6**, 459–484.

Galeev, A. A., Sagdeev, R. Z. (1979), in: M. A. Leontovich (Ed.), *Reviews of Modern Physics 7*, New York: Consultants Bureau, pp. 1–180.

Grieger, G., Lotz, W., Merkel, P., Nührenberg, J., Sapper, J., Strumberger, E., Wobig, H., the W7-X Team, Burhenn, R., Erckmann, V., Gasparino, U., Giannone, L., Hartfuss, H. J., Jaenicke, R., Kühner, G., Ringler, H., Weller, A., Wagner, F., W7-AS Team (1992), *Phys. Fluids B* **4**, 2081–2091.

Haines, M. G., Dangor, A. E., Choi, P., Coppins, M., Bayley, J. M., Arber, T., Chittenden, J. P., Mitchell, I. H., Nifkeer, S., McCall, G. H., Scheffel, J. (1991), *Plasma Physics and Controlled Nuclear Fusion Research 1990*, Vienna: IAEA, Vol. 2, pp. 769–792.

Heidbrink, W. W., Strait, E. J., Doyle, E., Sager, G., Snider, R. T. (1991), *Nucl. Fusion* **31**, 1635–1648.

Ioffe, M. C., Sobolev, R. I. (1964), *At. Energy* **17**, 366–372.

Ishimura, T., Ueda, Y., Sugimoto, S., Okada, S., Ito, Y., Ohi, S., Goto, S., Takahashi, T., Shimamura, S., Nogi, Y., Hamada, S. (1989), *Plasma Physics and Controlled Nuclear Fusion Research 1988*, Vienna: IAEA, Vol. 2, pp. 705–711.

Jackson, G. L., Winter, J., Taylor, T. S., Greenfield, C. M., Burrell, K. H., Carlstrom, T. N., DeBoo, J. C., Doyle, E. J., Groebner, R. J., Lao, L. L., Rettig, C., Schissel, D. P., Strait, E. J., DIII-D Research Team (1982), *Phys. Fluids B* **4**, 2181–2188.

Jacquinot, J., Bures, M., JET Team (1992), *Phys. Fluids B* **4**, 2111–2121.

Jarboe, T. R., Henins, I., Sherwood, A. R. (1983), *Phys. Rev. Lett.* **51**, 39–42.

JET Team (1991), *Plasma Physics and Controlled Nuclear Fusion Research 1990*, Vol. 1, Vienna: IAEA, pp. 41–65.

JET Team (1992), *Nucl. Fusion* **32**, 187–203.

Kernbichler, W., Hendler, M., Momota, H., Tomita, Y., Ishida, A., Ohi, S., Ohnishi, M., Sato, K., Miley, G. H., Berk, H. L., Dove, W. (1991), *Plasma Physics and Controlled Nuclear Fusion Research 1990*, Vienna: IAEA, Vol. 3, pp. 555–566.

Lazarus, E. A., Chu, M. S., Ferron, J. R., Helton, F. J., Hogan, J. T., Kellman, A. G., Lao, L. L., Lister, J. B., Osborne, T. H., Snider, R., Strait, E. J., Taylor, T. S., Turnbull, A. G. (1991), *Phys. Fluids B* **3**, 2220–2229.

Maglich, B. (1988), *Nucl. Instrum. Meth. Phys. Res.* **A271**, 13–36.

Miyoshi, S., Cho, T., Hojo, H., Ichimura, M., Inutake, M., Ishii, K., Itakura, A., Katanuma, I., Kiwamoto, Y., Mase, A., Nakashima, Y., Saito, T., Tsubouchi, D., Yamaguchi, N., Yatsu, K. (1991), *Plasma Physics and Controlled Nuclear Fusion Research 1990*, Vienna: IAEA, Vol. 2, pp. 539–546.

Perkins, W. A., Post, R. F. (1963), *Phys. Fluids* **6**, 1537–1558.

Post, R. F., Rosenbluth, M. N. (1966), *Phys. Fluids* **9**, 730–749.

Robinson, D. C., King, R. E. (1969), *Plasma Physics and Controlled Nuclear Fusion Research*, Vol. I, Vienna: IAEA, p. 263–276.

Rosenbluth, M. N., Bussac, M. N. (1979), *Nucl. Fusion* **19**, 489–498.

Rosenbluth, M. N., Krall, N. A., Rostoker, N. (1960), *Nucl. Fusion Suppl., Part 1,* 143–150.

Ryutov, D. D. (1975), *Usp. Fiz. Nauk* **116**, 341–342.

Taylor, J. B. (1986), *Rev. Mod. Phys.* **58**, 741–763.

Tomabechi, K., Gilleland, J. R., Sokolov, Yu. A., Toschi, R., ITER Team (1991), *Nucl. Fusion* **31**, 1135–1217.

Troyan, F., Gruber, R., Saurenmann, H., Semenzato, S., Succi, S. (1984), *Plasma Phys. Controlled Nucl. Fusion* **26**, 209–215.

Tubbing, B. J., Gottardi, N. A. C., Green, B. J., How, J. A., Huart, M., Konig, R., Lowry, C. G., Lomas, P. J., Noll, P., O'Rourke, J. J., Rebut, P. H., Stork, D., Tanga, A., Taroni, A., Ward, D. J. (1992), *Nucl. Fusion* **32**, 967–972.

Uo, K., Iiyoshi, A., Obiki, T., Morimoto, S., Sasaki, A., Kondo, K., Motojima, O., Yoshioka, S., Ohtake, I., Nakasuga, M., Sato, M., Hanatani, K., Mutoh, T. (1977), *Plasma Physics and Controlled Nuclear Fusion Research 1976,* Vienna: IAEA, Vol. 2, pp. 103–111.

Wittenberg, L. J., Santarius, J. E., Kulcinski, G. L. (1986), *Fusion Technol.* **10**, 167–178.

Wong, K. L., Fonck, R. J., Paul, S. F., Roberts, D. R., Fredrickson, E. D., Nazikian, R., Park, H. K., Bell, M., Bretz, N. L., Budny, R., Cohen, S., Hammett, G. W., Jobes, F. C., Meade, D. M., Medley, S. S., Mueller, D., Nagayama, Y., Owens, D. K., Synakowski, E. J. (1991), *Phys. Rev. Lett.* **66**, 1874–1877.

Yamada, M., Ellis, R. A., Jr., Janos, A. C., Levinton, F. M., Mayo, R. M., Motley, R. W., Nagata, M., Ono, Y., Satomi, N., Ueda, Y., Jarboe, T. R., Barnes, C. W., Fernandez, J. C., Henins, I., Marklin, G. J., and Wysocki, F. J. (1989), *Plasma Physics and Controlled Nuclear Fusion Research,* Vienna: IAEA, Vol. 2, pp. 539–555.

Zarnstorff, M., Barnes, C. W., Efthimion, P. C., Hammett, G. W., Horton, W., Hulse, R. A., Mansfield, D. K., Marmar, E. S., McGuire, K. M., Rewoldt, G., Stratton, B. C., Synakowski, E. J., Tang, W. M., Terry, J. L., Xu, X. Q., Bell, M. G., Bitter, M., Bretz, N. L., Budny, R., Bush, C. E., Diamond, P. H., Fonck, R. J., Fredrickson, E. D., Furth, H. P., Goldston, R. J., Grek, B., Hawryluk, R. J., Hill, K. W., Hsuan, H., Johnson, D. W., McCune, D. C., Meade, D. M., Mueller, D., Owens, D. K., Park, H. K., Ramsey, A. T., Rosenbluth, M. N., Schivell, J., Schmidt, G. L., Scott, S. D., Taylor, G., Wieland, R. M. (1991), *Plasma Physics and Controlled Nuclear Fusion Research 1990,* Vienna: IAEA, Vol. 1, pp. 109–122.

Further Reading

Galeev, A. A., Sudan, R. N. (Eds.) (1983), *Handbook of Plasma Physics; Basic Plasma Physics I, II,* Amsterdam: North-Holland.

Teller, E. (Ed.) (1981), *Fusion, Volume I: Magnetic Confinement,* New York: Academic.

Post, R. F. (1987), "The Magnetic Mirror Approach to Fusion," *Nucl. Fusion* **27**, 1579–1739 (1987).

Nucl. Fusion (1990), Vol. 30, p. 1637–1996. (This issue contains review papers on all the principal approaches to magnetic fusion.)

FUSION, NUCLEAR

See NUCLEAR ENERGY, FUSION

FUSION TECHNOLOGIES

FARROKH NAJMABADI, *Institute of Plasma and Fusion Research, University of California at Los Angeles, Los Angeles, California, U.S.A.*

Introduction 1
1. **Fusion-Reactor Overview** 3
2. **Fusion-Plasma Technologies** 4
2.1 Magnetic Fusion-Energy Reactors 4
2.1.1 Plasma Generation and Heating 4
2.1.2 Noninductive Current Drive 6
2.1.3 Magnets 6
2.1.4 Particle Exhaust and Fueling 8
2.1.5 Alternative Magnetic-Confinement Concepts 9
2.2 Inertial Fusion-Energy Reactors 10
3. **Fusion-Reactor Technologies** 11
3.1 First Wall, Blanket, and Shield 11
3.1.1 Function 11
3.1.2 Tritium Breeding 12
3.1.3 Structural Material 12
3.1.4 Conceptual Designs 13
3.1.5 Tritium-Recovery Systems 14
3.2 Power-Conversion Systems 15
3.3 Inertial Fusion-Energy Reactors 16
4. **Advanced Fuels** 16
5. **Conclusions** 17
Glossary 17
Works Cited 19
Further Reading 19

INTRODUCTION

Fusion is the primary source of energy in the universe—it is the phenomenon that powers the sun and the stars. In a fusion reaction, two light nuclei combine to produce a heavier nuclide, thus liberating a large amount of energy. Because the reacting nuclei have positive charges, their relative speed must be high enough to overcome the electrostatic repulsive forces. Relative kinetic energies of 100–200 keV are required to fuse hydrogen-isotope nuclei, which have the smallest Coulomb barrier because they are singly charged. For a thermalized mixture of the fusing nuclei (i.e., with Maxwellian velocity distributions), production of an appreciable amount of fusion energy requires a temperature of more than 5 keV (6×10^7 K) and peaks at tens of keV. At these high temperatures, the fusing fuel is in the plasma state, i.e., a completely ionized gas in which electrons are not bound to the nuclei. Obviously, no material wall can contain the fusion plasma, nor is the gravitational confinement of the fusion plasma (as in the sun) practical. The two principal methods for isolating the fusion plasma from the surroundings are magnetic- and inertial-confinement schemes (see FUSION, MAGNETIC CONFINEMENT; FUSION, INERTIAL CONFINEMENT).

The fusion reaction that involves two isotopes of hydrogen, deuterium (D) and tritium (T),

$$D+T\rightarrow {}^4He(3.52\ \text{MeV})+n(14.06\ \text{MeV}),$$

is the least difficult one to achieve in terrestrial conditions. This is because it has the largest cross section and occurs at the lowest plasma temperature. Other fusion reactions of hydrogen isotopes, which are usually referred to as advanced-fuel reactions, have also been considered for terrestrial application and are discussed in Sec. 4.

Examination of the above-mentioned D-T fusion reaction has underlined the potential desirable features of fusion and caused a worldwide quest for a practical fusion reactor. It has also highlighted the challenges in achieving this goal. First, fusion reactions produce a large amount of energy from a small amount of fuel. For example, a 1000-MWe power station requires about 1 kg/day

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

of D-T fuel, while a similar coal-fired plant would require 2000 tonne/day of coal. The fuel for fusion reactions is readily available. Since for every 6700 hydrogen atoms there is a deuterium nucleus, the water in the Earth's oceans contains a virtually inexhaustible supply of fusion fuel. Because tritium is radioactive with a half-life of 12.3 years, it is not found in significant quantities in nature. It can, however, be produced from lithium, which is an abundant natural resource. If the fusion plasma chamber is surrounded with a blanket containing lithium or lithium compounds, the neutrons produced in a D-T fusion reaction interact with the lithium to produce tritium, which is subsequently burned in the reactor.

Second, fusion reactors have the potential to be environmentally benign. The amount of fusion fuel in the reactor is limited. Therefore, even if an uncontrolled fusion burn occurs, it quickly consumes the available fuel and is extinguished. The "ashes" from fusion reactions are α particles, the nuclei of helium, which are not radioactive. There are two sources of radioactivity in a fusion reactor. The first is the tritium fuel itself. Because the tritium is generated in the reactor and is burned *in situ*, the amount of tritium in the system is much smaller than the amount of nuclear fuel in a fission reactor. The second source is from the neutrons produced by the D-T reactions, which can transmute the materials in the reactor and make them radioactive. Fortunately, research has shown that by the proper choice of materials and by careful design, the level of induced radioactivity can be kept low (many orders of magnitude smaller than a comparable fission reactor).

Achieving and maintaining the high temperatures, T, required for the fusion burn are the most fundamental challenges for fusion energy research. Basically, the fusion fuel should be heated from ambient temperature to about 10^8 K for fusion reactions to start. At these temperatures, the plasma loses its energy through radiation (bremsstrahlung, synchrotron, and line radiation), heat conduction, turbulent convection of plasma particles, etc. The average time it takes before the energy escapes from the plasma is referred to as the energy-confinement time τ_E. To maintain the fusion temperature, energy must be supplied. Although one of the by-products of fusion reactions (the energetic neutron) leaves the plasma, the other (charged α particle) slows down by colliding with the plasma particles and deposits its energy into the plasma, which sustains the plasma temperature. The necessary condition for a sustained D-T thermonuclear burn is that the "triple product" $nT\tau_E$ exceed 2×10^{21} s keV/m^3 (n is the plasma density). If the triple product is smaller than this value, the α-particle power is not sufficient to sustain the burn and the temperature decreases.

Most of the fusion research in the past 40 years has focused on achieving the necessary plasma temperature and triple product for fusion burn. The required fusion temperature has been attained and several experiments are approaching the required triple product. To date, almost all of the experiments have been performed in hydrogen or deuterium plasmas in order to avoid difficulties associated with handling radioactive tritium. Thus fusion-power production has been small and from D-D reactions (see Sec. 4). Operation with D-T plasmas is planned for two major experiments in the world. The Joint European Torus (JET) at Culham Laboratories has already produced 2 MW of D-T fusion power for 2 s with a small amount of T fuel, in preparation for extensive and powerful (several tens of MW) tritium operation planned for 1995. The Tokamak Fusion Test Reactor (TFTR) at the Princeton Plasma Physics Laboratory is expected to start D-T operation in 1993 with a fusion power of about 10 MW. The fusion ignition and long-pulse operation remain to be demonstrated in a larger experiment, the International Thermonuclear Experimental Reactor (ITER), proposed as a joint venture by the European Community, Japan, the Russian Federation, and the United States. The ITER project is now in the engineering design activity phase with construction planned to begin before the year 2000.

Conceptual fusion-reactor design studies have been carried out since the early 1970s. In addition to assessing whether a confinement system is attractive as a fusion-power reactor, these studies have uncovered potential problems, proposed solutions that are consistent with other reactor subsystems, and identified critical areas that should receive priority in experimental programs. It is a characteristic of this field that many of the research papers appear as laboratory reports or proceedings of conferences. The most up-to-date assessment of fusion reactors can be found in the ARIES Tokamak Reactor Study reports (Najmabadi *et al.*, 1991, 1992, 1993) the SSTR report (Seki *et al.*, 1991), and the overviews of Sombrero, Osiris, and Prometheus IFE reactors (Abdou and Waganer, 1993; Bourque *et al.*, 1992; Sviatoslavsky *et al.*, 1992), all of which have been national efforts involving scientists from major national laboratories, universities, and industry involved in fusion research.

1. FUSION-REACTOR OVERVIEW

Although several experiments are approaching the triple product that is required for fusion ignition, many other technological issues must be resolved before an economical fusion-power reactor can be realized. These issues can be loosely divided into two sets, the first of which relates to the plasma itself (Sec. 2). It is necessary to be able to create a plasma reliably, provide auxiliary heating to achieve fusion temperature and burn, and provide control of the major plasma parameters (e.g., n and T) for the duration of the burn. For steady-state or long-pulse reactors (e.g., the majority of the magnetic-confinement schemes), the plasma must be replenished with new fuel and the fusion ash (α particles produced by the D-T reaction) removed from the plasma. Obviously, the plasma technologies are confinement-method specific (see FUSION, MAGNETIC CONFINEMENT; FUSION, INERTIAL CONFINEMENT).

The second set of fusion-reactor technology issues relates to the systems surrounding the plasma (Sec. 3). The functions of these systems are mainly to convert the energy produced by the plasma into heat and then into electricity, to attenuate the neutron flux produced by the D-T plasma, and to produce sufficient tritium to fuel the reactor. Not only must these systems operate reliably in a harsh radiation environment, but they should also convert the plasma energy into electricity efficiently so that the reactor is economically viable. Furthermore, these systems must be built of special materials and designed in such a way that the attractive safety and environmental features of fusion are realized.

The technologies of a fusion reactor encompass many fields of science and engineering such as plasma physics, plasma–material interaction physics, materials science and engineering, heat transfer and fluid dynamics, chemical engineering, electrical engineering, electronics, power-systems engineering, vacuum-system engineering, all aspects of nuclear engineering, and (for magnetic-confinement concepts) superconductivity or (for inertial-confinement concepts) lasers, optics, accelerators, and pulse-power technologies. In some areas, new fields of science (e.g., plasma physics) had to be established, and in others, new frontiers had to be sought.

2. FUSION PLASMA TECHNOLOGIES

2.1 Magnetic Fusion-Energy Reactors

The most advanced magnetic-confinement approach is that of the tokamak (see FUSION, MAGNETIC CONFINEMENT). Therefore the plasma technologies for a tokamak reactor are emphasized here. A conceptual tokamak reactor design is shown in Fig. 1. (Alternative magnetic-confinement concepts are briefly described in Sec. 2.1.5.)

In the tokamak, which is similar in shape to a doughnut, the confining magnetic field is the result of combining a "toroidal field" (the long way around the doughnut), produced by coils outside of the plasma, and a "poloidal field," which is produced by the toroidal current flowing in the plasma. This combination causes the formation of a magnetic configuration that consists of nested surfaces. Each field line wraps helically and endlessly around one of these nested surfaces. The plasma heat and particles that flow quickly along the magnetic field lines are, therefore, well confined in this magnetic configuration. The plasma then diffuses slowly across the flux surface through cross-field transport processes.

The most efficient way to produce the toroidal current flow in the plasma is by magnetic induction (i.e., transformer action). A solenoid coil is placed in the center of the tokamak doughnut. Then, as the current flows in the solenoid coil, it induces a toroidal current in the plasma. Unfortunately, the toroidal current is induced only when the field produced by the solenoid is increasing (i.e., its current is rising)—a state that is practically impossible to maintain. Therefore, in most of the present-day experiments, when the cur-

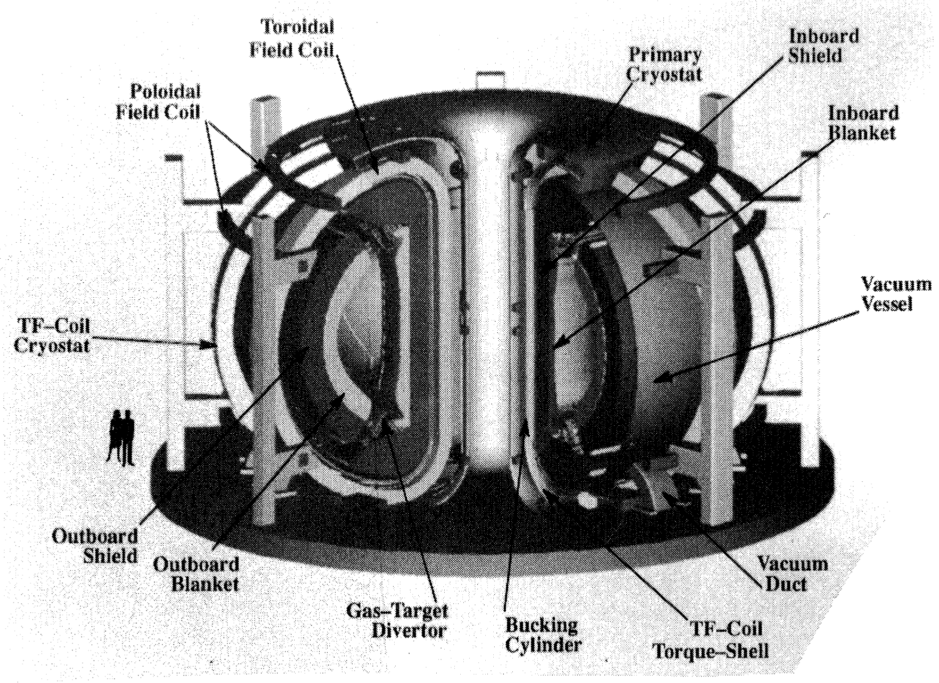

FIG. 1. The ARIES-IV conceptual tokamak fusion-reactor design (Najmabadi, 1993). The D-T plasma is located in the vacuum chamber inside the torus and is surrounded by the tritium-breeding blanket, the magnetic shield, the toroidal magnetic-field coils, and the vacuum vessel. The poloidal-field coils (the ohmic-heating, equilibrium-field, and shaping coils) are also shown. Divertor plates are also visible at the top and the bottom of the plasma region.

rent in the solenoid has reached its maximum, the plasma is shut down, the solenoid current is reset, and a new discharge is initiated. Theoretical and experimental research, however, has shown that the toroidal current can be produced by noninductive means (Sec. 2.1.2).

2.1.1 Plasma Generation and Heating. The plasma is usually generated by breaking down the deuterium gas that fills the plasma chamber. This is accomplished by initiating a current ramp in the solenoid coil set. The resulting induced voltage in the chamber then breaks down the gas. To facilitate plasma generation in most tokamaks, microwaves are also employed to ionize the initial gas partially.

As the plasma current is raised to its final value, the plasma heats Ohmically. However, since plasma electrical resistivity is reduced drastically as plasma temperature increases, an Ohmically heated tokamak plasma can only reach temperatures of about 2–3 keV. Auxiliary plasma heating is, therefore, required to achieve fusion ignition temperatures (~ 10 keV).

One of the most common methods of auxiliary heating employs an intense beam of high-energy neutral particles (neutral beam), since neutral particles can move through the magnetic field unaffected and penetrate the plasma. The beam particles are then ionized and trapped in the plasma as a result of collisions with plasma ions and electrons. The resulting high-energy ions then "slow down" by Coulomb collisions and deposit their energy in ions and electrons. The mean distance that a neutral ion travels in a given plasma before it is ionized scales directly with its energy. Since it is most efficient to heat the center of the plasma, the beam energy is usually determined by the plasma density, temperature, and size. For a fusion-reactor plasma, beam energies of 100–200 keV are required for perpendicular injection (perpendicular to the toroidal circumference of the plasma) and 500–3000 keV for tangential injection (e.g., for current-drive purposes, Sec. 2.1.2).

The schematic of a neutral-beam injector is shown in Fig. 2. The neutral beam begins as a hydrogen plasma produced by a radio-frequency (rf) plasma generator or by a high-current discharge. Then the ions leaving the source are accelerated, in one or several stages by a series of electrostatic grids, to their final energy. The resulting ion beam is then neutralized by passing through a neutralizer chamber containing cold hydrogen gas. The beam ions pick up electrons from the neutral gas via the charge-exchange process and travel out of the neutralizer chamber as high-energy neutral particles. The neutral gas, which is ionized by the ion beam, drifts slowly to the walls of the chamber. Meanwhile, the neutralized beam passes through a drift tube before entering the reactor plasma. The small fraction of the original ion beam that is not neutralized in the neutralizer chamber is deflected by a small magnet in the drift tube. At present, the neutral-beam injectors on the TFTR experiments are capable of operating at 120 keV and delivering ~30 MW of heating power to the plasma for a pulse length of ~2 s (easily upgradable to steady-state operation).

Because the efficiency of the gas neutralizer chamber to neutralize the positive-ion beams degrades rapidly for a beam energies larger than ~150 keV, negative-ion beams should be utilized for higher energies. A completely different technology is required for this scheme. To produce a beam of negative ions, either the plasma in the ion source is seeded with cesium atoms, or the ions extracted from the source are reacted with specially coated surfaces. The negative ions are then accelerated by either electrostatic or rf quadrupoles. The high-energy negative-ion beam is then neutralized either through the charge-exchange process in a xenon plasma or by using the laser photodetachment process. Negative-ion beam accelerators are under development. At present, negative-ion sources are capable of 50-keV H^--ion beams with a current of ~0.5 A at steady state. Stable H^- beams at 300 keV for 0.5 s have also been achieved. The first heating experiments with negative-ion beams, which are planned for JT-60U (a large toka-

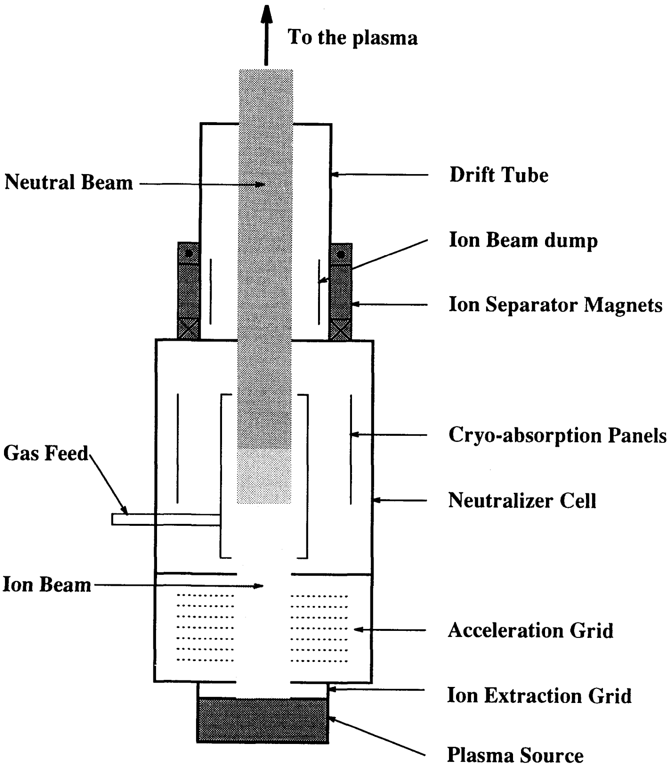

FIG. 2. A schematic of a positive-ion neutral-beam injector system.

mak experiment at the Japan Atomic Energy Research Institute) around 1995, will employ 500-keV deuterium beams (22 A for 10 s).

Another common method of auxiliary heating is by rf waves. In principle, the electromagnetic wave energy can be converted into plasma-particle kinetic energy through resonant wave–particle interaction. Because the plasma consists of charged particles, a variety of wave modes exist in a magnetized plasma and several rf heating schemes are possible and have been successfully demonstrated on present experiments (e.g., Alfvén, lower-hybrid, and electron-cyclotron resonance heatings, etc). One of the most advanced rf-heating techniques is ion-cyclotron range-of-frequencies (ICRF) heating. Depending on the wave frequency and the launcher design, ICRF waves can heat electrons, ions, or a minority ion species. In the latter scheme, a deuterium plasma is doped with a small amount of hydrogen or helium, and the ICRF wave frequency is matched to the cyclotron frequency of this minority species (~ 100 MHz, similar to television broadcast frequencies). Then the rf wave heats these minority species to the MeV range of energies, and their energy is transferred to plasma electrons and ions by Coulomb collisions.

All rf heating systems are, in essence, similar to a radio transmitter. They include an oscillator, an amplifier, a transmission line, and a launching structure inside or at the plasma chamber wall. The oscillator and amplifier can be a tetrode, a klystrode (a linear beam device having features of both klystron and tetrode), or a solid-state device based on field-effect transistors. The transmission line is a coaxial cable or waveguide, and the launching structure can be either an antenna or a waveguide. The major fusion-specific component of the rf heating system is the launcher.

In order for the launched rf power to couple well to the plasma, the wave number and polarization of the launched wave must be strictly controlled. Then the wave is able to penetrate the plasma and be absorbed at the desired location. To protect the launcher from electromagnetic radiation from the plasma, a "Faraday shield" is usually placed in front of it. The launcher should also be hardened against damage from the neutrons and charged particles from the plasma (Sec. 3.1).

2.1.2 Noninductive Current Drive. As mentioned before, plasma current driven by inductive means cannot be sustained indefinitely. Even though a pulsed tokamak reactor with a burn time of a few hours is a possibility, extensive research has also been directed towards noninductive means of driving the plasma current.

Theoretical research in the early 1970s predicted that in a sufficiently hot plasma, the radial pressure gradient and the dynamics of plasma flow on the flux surfaces combine to produce a self-driven current, usually referred to as the bootstrap current. Existence of the bootstrap current was experimentally confirmed only recently. Major tokamak experiments have operated with up to 80% of the plasma current driven by the bootstrap effect. Unfortunately, no tokamak plasma equilibrium has yet been found (either theoretically or experimentally) in which all of the plasma current is driven by the bootstrap effect. The difference between the plasma and bootstrap currents must still be driven by some other means.

Research has shown that both neutral-beam and rf-wave heating can be used to drive the plasma current. For example, by tuning the ICRF launching structure, most of the energy of rf waves can be used to accelerate the electrons parallel to the field line, which generates toroidal currents. The efficiency of these noninductive current-drive schemes, however, is low, and most steady-state tokamak reactor designs operate in regimes that maximize the bootstrap current, reducing the necessary current-drive power to a manageable level (~ 100 MW of current-drive power in a 1000-MWe power station).

2.1.3 Magnets. In addition to the toroidal-field magnets and the "Ohmic-heating" solenoid discussed above, other magnets are required for a tokamak. In principle, since a current-carrying loop can move along its axis freely, a tokamak plasma can move up or down and hit the plasma chamber walls. To keep the plasma at its location, a pair of "equilibrium-field" coils with currents flowing in the toroidal direction (but opposite to that of the tokamak plasma) are required. These coils are usually located symmetrically at about the toroidal midplane of the tokamak torus (one above and one below) and have radii that are usually larger than that of the plasma.

An important measure of the efficiency of a magnetic-confinement concept is the plasma β, the ratio of the plasma pressure to the pressure exerted by the external magnetic field that confines the plasma. The higher the β value, the smaller and more economical the external magnets can be. A major accomplishment of plasma physics research during the last decade has been achieving an understanding of the relationship of the plasma β to plasma shape, current, toroidal field, etc. It has been shown, both theoretically and experimentally, that elliptical or D-shaped plasma cross sections (with an elongation of about 2) can lead to improved plasma β. Plasma shaping, however, requires additional coils, all having toroidal currents, strategically placed around the plasma.

Most of the magnetic fusion-reactor concepts use superconducting magnets because the power consumed by normal-conducting magnets would be very large. Basically, superconductivity occurs in certain materials when their temperatures are reduced below some critical temperature. Most of the superconducting magnets, at present, use niobium–titanium (NbTi) or niobium–tin (Nb_3Sn) alloys as the superconducting material cooled by liquid helium to temperatures of about 4 K (some designs use either of these superconductors at lower temperatures). The commercial superconducting magnets available today mainly use NbTi and are of medium size (e.g., those used for magnetic-resonance imaging). Nearly all experience with large-size superconducting magnets to date is in high-energy physics (e.g., accelerators, bubble chambers) or in fusion. Compared to other superconducting magnets available today, fusion magnets are usually larger (many meters in diameter), may have noncircular cross sections (e.g., D-shaped toroidal-field coils), and have a high magnetic stored energy (tens of GJ). As a result, the design of these magnets is dominated by the structure, which is embedded in the coil to support the massive electromagnetic forces. Furthermore, the design of fusion magnets is also severely constrained by fusion-specific issues, such as allowable location for the coils and the structure, ability to operate in a radiation environment, and capability to produce and withstand high rates of field change (i.e., several T/s for poloidal-field coils).

Because of the engineering constraints, the peak field produced by a practical superconducting magnet is usually well below the critical field limit of the superconducting material. The maximum field produced by a Nb_3Sn superconducting magnet (a small-bore solenoid) is 21 T. The maximum field produced by a fusion magnet with a noncircular cross section is $\sim$9 T. It should be noted that the magnetic field strength on the plasma axis in a tokamak is considerably lower than the peak field on the toroidal-field magnets because of geometrical effects. For the ITER design, research and development are directed towards superconducting magnets with peak fields of 11–14 T, while conceptual tokamak-reactor designs have envisioned magnets with a peak field ranging from 14 to 21 T (depending on the value of plasma β).

A cross section of a toroidal-field magnet is shown in Fig. 3. In this type of design, filaments of niobium and tin are twisted together in groups and embedded in a copper sleeve to form a cable. Several cables are then placed together in a steel conduit, which also has passages for liquid-helium coolant. This conduit is heat treated so that the niobium and tin filaments interact, forming the superconducting alloy. A layer of insulating material is placed on the outside of the conduit, and layers of the superconductor conduits are interleaved with steel supports. The coil is then placed in a stainless Dewar or cooling vessel.

The superconducting magnets must be kept insulated from the nearby plasma because neutrons produced by fusion reactions can damage the superconducting material. Furthermore, so much heat would be generated by the interaction of neutrons with the magnet material that refrigeration of the magnets would be impractical. In order to attenuate the neutron flux on the magnets, they are usually placed behind the blanket and shield components (Sec. 3.2).

2.1.4 Particle Exhaust and Fueling. In most modern tokamaks, the plasma-shaping coils are energized such that a magnetic separatrix is formed inside the plasma chamber (Fig. 4). In the region inside this separatrix,

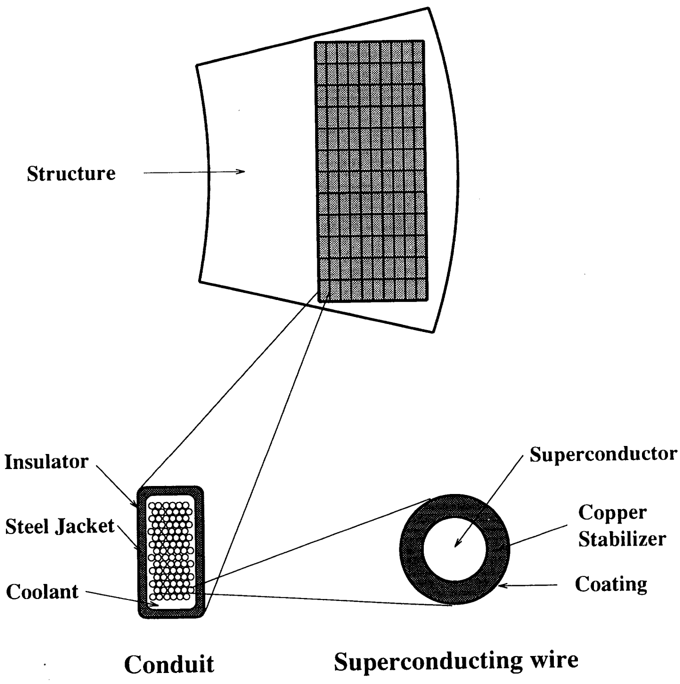

FIG. 3. A cross section and internals of the inner leg of a superconducting toroidal-field magnet.

called core plasma, the nested flux-surface structure of the tokamak is preserved. The field lines outside the separatrix, however, leave the vicinity of the plasma within a single poloidal transit (but many toroidal transits) and intersect material walls (divertor plates) in certain locations. The heat and particles from the core plasma diffuse outward through the separatrix into the scrapeoff-layer region and then flow rapidly along the field lines to the divertor plates. Clearly, very high heat and particle fluxes appear at the intersection of the separatrix with the divertor plate (strike point). Note that there is a large body of experimental evidence that shows that controlling the scrapeoff-layer plasma properties has profound influence on the quality of core-plasma confinement. In fact, the best tokamak discharges have been obtained in a divertor configuration.

The plasma that intersects the divertor plates is neutralized through surface recombination processes. Some of the neutral atoms are ionized, follow the field lines, and flow back to the divertor plate and, in effect, are recycled in the divertor region. If the separatrix plasma density is high enough, a high-recycling divertor regime is formed in which more than 90% of the neutral particles from the divertor plate are re-ionized in the plasma and are recycled. This mode of operation spreads the heat and particle load on the divertor plate and reduces the peak fluxes at the strike point. Further reduction in heat and particle fluxes can be achieved by slanting the divertor plate with respect to the field lines (Fig. 4).

At present, the design of the divertor represents one of the most fundamental challenges in designing fusion reactors, and it is not yet clear whether the high-recycling divertors will be able to reduce the particle and heat fluxes to manageable values. Many other concepts are, therefore, being pursued. In one, the plasma near the divertor is doped with impurities that cool the divertor plasma through line radiation. Another involves terminating the divertor plasma in a dense, but much cooler, neutral gas. Both of these concepts would spread the heat and particle load over a much larger area in the divertor.

Most of the neutral atoms that are produced on the divertor plate are ionized and flow back to the divertor plate. The rest, which are a mixture of fusion fuel (D and T) and α particles, is pumped, which removes the fusion ash from the system. Cryosorption pumps (see CRYOGENICS) are a leading technology for fusion reactors. Cryopumps include an array of sieves that are cooled to liquid-helium temperature. The sieves are surrounded by chevrons that are cooled to liquid-nitrogen temperatures. Impurity gases such as oxygen and nitrogen are adsorbed by the chevrons while hydrogen isotopes condense on the sieves. Helium can also be trapped by the sieves if liquid argon is sprayed onto the surfaces to overlay the helium atoms. Cryopumps are simple and attain very high pumping speeds at low pressures (10^{-7} atm). Their drawbacks are that they require periodic warming up to remove the gases trapped on the surface, and that the trapping of tritium onto the cryopump surfaces will result in a high inventory of tritium in the reactor. An alternative approach is to design the divertor system such that the neutral pressure at the front of the pumping duct is relatively high ($\sim 10^{-5}$ atm). In this case, nonadsorbing pumps, such as turbomolecular pumps, can be used. The exhaust gas from the reactor is processed to remove the helium ash and impurities.

For the fusion burn to continue, the reactor must be refueled with the D and T atoms that have been exhausted from the reactor, together with a supply of fuel to replenish that which was burned. In most current experiments, fueling is performed primarily by puffing the fuel gas at the edge of the plasma. This fueling technique is not very effective because most of the fuel does not reach the plasma center. An alternative method, also currently used in major tokamak experiments, is to inject high-speed pellets of frozen fuel. A typical pellet is several millimeters in diameter and is accelerated to speeds of ~2 km/s by high-pressure gas propellant in a gas gun (Fig. 5).

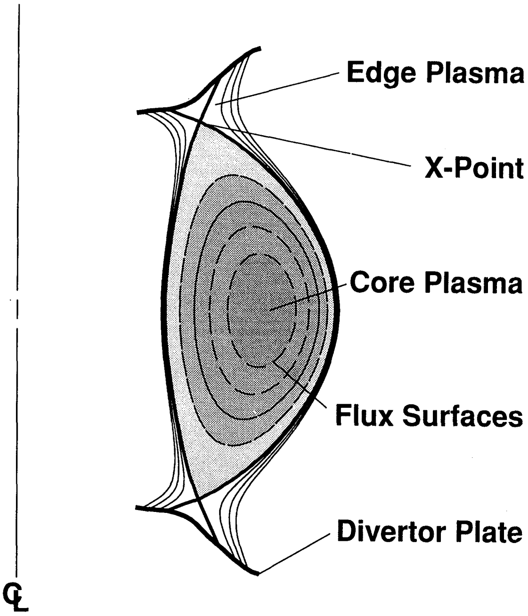

FIG. 4. A cross section of a tokamak plasma. The lines indicate the cross sections of flux surfaces (magnetic field lines are located on these surfaces). The "separatrix" surface marks the boundary between closed magnetic surfaces of the core plasma and the field lines in the scrapeoff-layer plasma that leave the vicinity of the plasma and intersect the divertor plates.

2.1.5 Alternative Magnetic-Confinement Concepts. Of the many magnetic-confinement concepts that have been investigated in the past 30 years, the tokamak is the most advanced. Other promising concepts include the stellarator, reversed-field pinch, and field-reversed configurations (see FUSION, MAGNETIC CONFINEMENT). Each has desirable features that simplify or eliminate plasma support technologies. These features are highlighted below.

The confining magnetic field structure in a stellarator is similar to that of a tokamak, although it is produced by complicated coils external to the plasma. The major advantage of the stellarator is that because it does not require a toroidal current in the plasma, it can operate in steady-state mode without a current-drive system.

The reversed-field pinches also have nested flux surfaces similar to those in the tokamak. However, most of the toroidal field is produced by current flowing in the plasma. The direction of the toroidal field is reversed near the plasma edge and only a weak external toroidal field is required to maintain the plasma configuration. Ohmic heating to ignition may also be possible. Elimination of the auxiliary heating system and major simplification of the external magnets are major features of this concept. Steady-state operation, however, requires a noninductive current-drive technique.

The field structure in a field-reversed configuration is very different from the above-noted concepts; it resembles that of a current loop (plasma) in an external solenoidal field. The experimentally achieved plasma β is very high, the external magnets are simple, and it is more amenable to direct conversion of charged-particle energy to electricity. At this

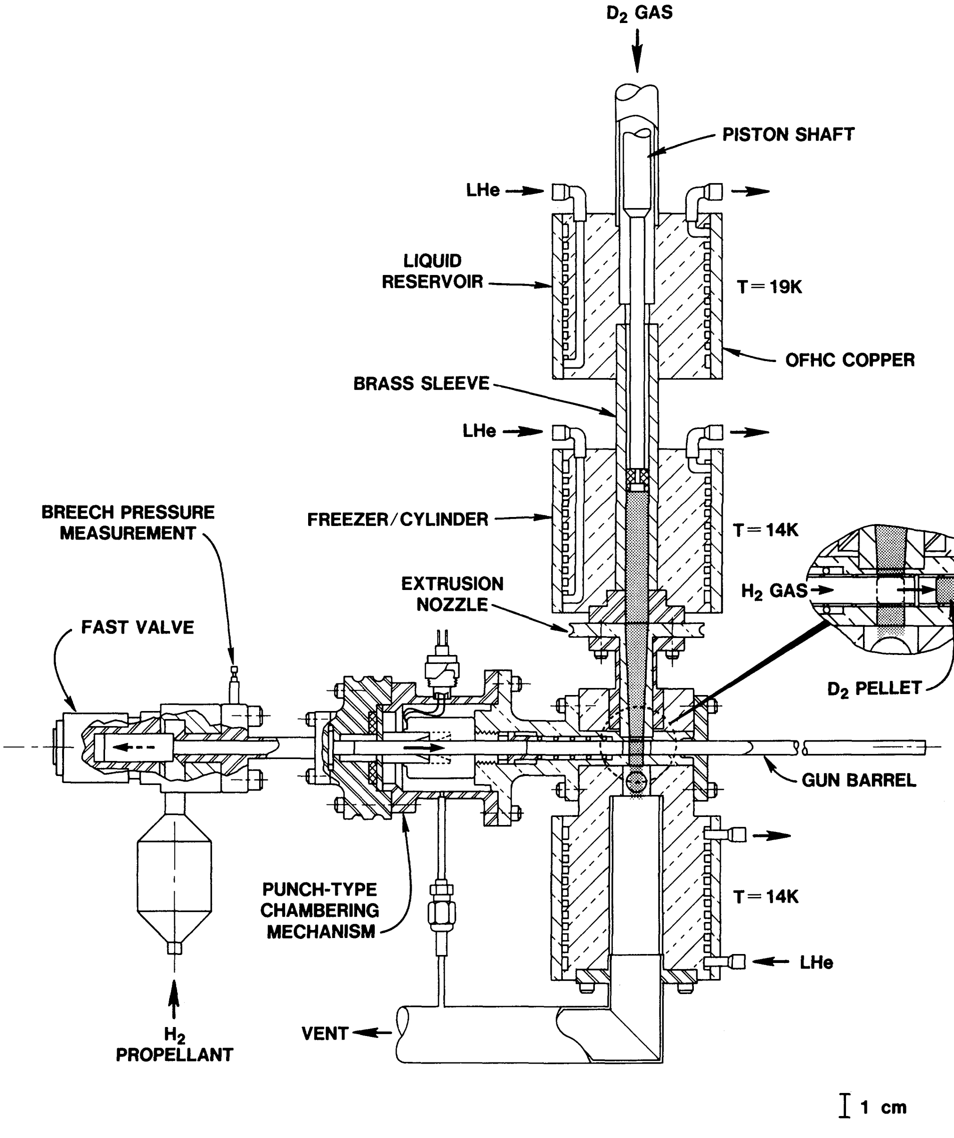
D_2 GAS
PISTON SHAFT
LHe
LIQUID RESERVOIR
T = 19K
OFHC COPPER
BRASS SLEEVE
LHe
FREEZER/CYLINDER
T = 14K
BREECH PRESSURE MEASUREMENT
EXTRUSION NOZZLE
H_2 GAS
FAST VALVE
D_2 PELLET
GUN BARREL
PUNCH-TYPE CHAMBERING MECHANISM
T = 14K
LHe
H_2 PROPELLANT
VENT
1 cm

FIG. 5. Schematic of cryogenic extruder and gun assembly for a repeating pneumatic pellet injector (figure courtesy of S. Milora, Oak Ridge National Laboratory).

time, however, the experimental data base for this concept is limited.

2.2 Inertial Fusion-Energy Reactors

Because of its inertia, material at fusion temperature stays together for a finite, albeit short, period of time (nanosecond time scale). Since fusion ignition and burn require attainment of the triple product $nT\tau_E$, inertial-confinement fusion can be achieved if the fusion fuel is compressed to a very high density (1000 times liquid density). Such a compression can also adiabatically heat the fuel to fusion temperatures.

In the inertial-confinement scheme (see FUSION, INERTIAL CONFINEMENT), a small pellet containing the fusion fuel is bombarded by powerful lasers or particle beams (typically 100–1000 TW of power is needed). Ablation of the material on the pellet surface, therefore, causes an implosion that compresses the center of the pellet to high density and high temperature. A small region in the center of the pellet can become hot enough to initiate a fusion reaction. The energy released from the fusion reactions can heat the surrounding colder fuel to fusion temperature, resulting in a fusion-burn front that rapidly propagates outward and consumes the rest of the fusion fuel before the pellet flies apart.

Candidate drivers include lasers and particle-beam accelerators. Solid-state neodymium (Nd) glass lasers are presently used in inertial-confinement experiments. The efficiency of this laser (~1%) is too low for reactor applications, however. Krypton fluoride (KrF) gas excimer lasers are predicted to have a higher efficiency (~10%) and are under development as a potential driver. Particle beams are more attractive since they have higher intrinsic efficiency (~25%). Low-energy light-ion beams with a high current are used in the present experiments, but high-energy heavy-ion beams with a low current, which are under development, are thought to be the best driver for a power reactor. The technologies of the above drivers are covered in detail in FUSION, INERTIAL CONFINEMENT, Sec. 5.

The ratio of the energy released by pellet explosion to the driver energy is usually referred to as the pellet gain G. A high pellet gain (typically 100 for a D-T reaction) is required to compensate for the inefficiencies in the driver and thermal-power-conversion system. Achieving ignition and high pellet gain imposes severe constraints on the driver and pellet designs. First, the driver pulse should be shaped properly in order to maximize the fraction of the driver energy that is used in compressing the pellet. Second, and more importantly, the pellet implosion should be highly uniform (pressure nonuniformity should be less than 1%). This requires that the pellet surface should be very smooth (less than 100-Å imperfections). Also, the pellet should be heated uniformly, either by ensuring that the driver beam strikes the pellet uniformly (direct drive) or by converting the driver energy into x rays in a *Hohlraum* (indirect drive).

In an inertial-fusion reactor, the pellet microexplosion occurs in the reactor chamber. Practical constraints on the size and lifetime of the reactor chamber limit the yield of pellet explosions to a maximum of 500–1000 MJ (a pound of high explosive releases about 2 MJ of energy). Therefore, a typical power reactor (1000 MWe) requires a pulse rate of a few Hz. During each pulse, a pellet is injected several meters into the reactor chamber and should be tracked until it is in the precise location and alignment for the driver. The driver should deliver its energy with the precise pulse shape and focus on the target. After the implosion, the debris of the pellet (and the *Hohlraum*, if used) should be evacuated from the reactor chamber so that the beam can propagate through the reactor chamber with minimum loss while maintaining adequate focus. These pulses should be repeated reliably millions of times per year, and the wall of the reactor chamber should be able to withstand the many microexplosions (see Sec. 3.6). Mass production of millions of highly smooth pellets at an acceptable cost is also a challenging technological issue. Research is ongoing on most of these issues.

3. FUSION-REACTOR TECHNOLOGIES

3.1 First Wall, Blanket, and Shield

3.1.1 Function. The plasma reaction chamber is the first material boundary that faces the plasma. Usually referred to as the first wall, this region is subjected to bombardment by electromagnetic radiation from the plasma (mainly x rays), the flux of charged particles diffusing from the plasma, fast neutral atoms produced by charge-exchange processes in the plasma, and energetic neutrons produced by the D-T reaction. Each component of this harsh environment imposes severe constraints on the first-wall design (Sec. 3.1.2). The energy deposited in the plasma by the auxiliary heating system and by the fusion α particles must eventually leave the plasma. This energy appears mostly as surface heat flux on the first wall (radiation and charged-particle flux) and should be removed as useful heat. Sputtering of the first wall by plasma and neutral particles would cause erosion of the first-wall material, and the first wall may have to be coated with a low-Z material so that eroded material will not contaminate the plasma dramatically.

The fusion-produced neutrons have a long mean free path between collisions with solid and liquid materials (typically ~10 cm for 14-MeV neutrons). Most of these neutrons pass through the first wall and are absorbed in the blanket. The main function of the blanket is to moderate and capture these neutrons, extract their energy, and simultaneously breed enough tritium to replace that which is burned in the fusion process.

Typically, blankets are 0.8 to 1 m thick and attenuate the neutron flux by about 2 orders of magnitude. In MFE reactors, a shield is usually placed behind the blanket in order to attenuate the neutron flux by about 4 orders of magnitude so that a superconducting magnet can be located behind the shield. Such shields are typically 0.5 to 1 m thick.

3.1.2 Tritium Breeding. The primary reactions leading to tritium production are

$$^{6}\mathrm{Li}+n\rightarrow \mathrm{T}+{}^{4}\mathrm{He}$$

and

$$^{7}\mathrm{Li}+n\rightarrow \mathrm{T}+{}^{4}\mathrm{He}+n.$$

The first reaction leads to energy production (4.8 MeV released per reaction). The second consumes energy (2.87 MeV) but it gives rise to a second neutron that can subsequently react with ^{6}Li. The natural abundances of these lithium isotopes are 7.4% ^{6}Li and 92.6% ^{7}Li. The technology for enriching the natural lithium in the ^{6}Li isotope is available and is not very expensive. In some conceptual blanket designs, therefore, enrichment of up to 20% in ^{6}Li is utilized in order to optimize the nuclear performance of the blanket.

The number of tritium atoms produced in the blanket for each D-T neutron is referred to as the tritium breeding ratio (TBR). Since each fusion reaction uses one tritium atom to produce one neutron, a TBR of 1 or greater is required for tritium self-sufficiency. In a fusion reactor, not all of the fusion neutrons are absorbed in Li; some are absorbed in the structural material or coolant of the blanket. Therefore, in certain designs, a neutron-multiplying material is added to the blanket. Of particular interest is beryllium, which interacts with one neutron and produces two neutrons. These extra neutrons then compensate for neutron absorption in the blanket structural material.

The tritium-breeding materials can be lithium or lithium compounds. The choice is usually based on high-temperature capability; chemical compatibility with the coolant and the structural material; the desire for a high tritium-release rate (the tritium produced leaves the material rapidly, leading to a low tritium inventory in the blanket); the availability of, or the need for, a neutron-multiplier material; the tritium-recovery method; simplicity of design; etc. The first and most obvious choice is lithium itself in liquid form, to be used as both the breeder and coolant. Lithium eutectic compounds such as lithium–lead ($\mathrm{Li_{62}Pb_{38}}$), aqueous solutions of lithium salts, and fused lithium salts such as FLiBe can also be used as the coolant and breeder. Lastly, solid lithium compounds, such as lithium oxide ($\mathrm{Li_2O}$) and lithium orthosilicate ($\mathrm{Li_4SiO_4}$), have also been considered.

3.1.3 Structural Material. Material choices are critical in designing the fusion first wall and blanket. Structural material choices are usually based on their ability to operate at high temperatures, to withstand radiation damage by neutrons, and to have only a minimum amount of induced radioactivity caused by the fusion neutrons. Radiation damage is the most important criterion because it has the greatest influence on material performance and lifetime. During neutron irradiation, the absorption of a neutron by the material through scattering and absorption reactions can induce two major nuclear effects: atomic displacement and production of impurities and gas (hydrogen and helium).

During the scattering events, the energy transferred from the neutron to the nuclei can cause an atom in the solid structure to be displaced from its equilibrium lattice site, leaving behind a vacancy in the lattice. A scattered neutron, as well as the displaced atom, can interact with other atoms in the lattice and cause displacement cascades. This process continues until the energy transferred to the lattice atom becomes low enough so that it can be absorbed as vibrational energy by the lattice. The displaced atoms themselves usually come to rest at locations that are not regular lattice sites and, therefore, are referred to as interstitials. Even though each neutron produces many vacancy-interstitial pairs, most recombine rapidly and leave quite small net levels of point defects in the lattice structure. Voids in the material form if the vacancies are mobile enough (if the material temperature is high enough) to combine and if some agent, such as a small amount of gas, is present to stabilize the embryonic void. The void will then subsequently grow and cause dimensional changes of the structural material, referred to as "swelling," which can easily reach several tens of percent. Because the swelling rate depends on the material temperature and neutron flux, which are not uniform in a fusion blanket, the swelling rate will not be uniform, which can lead to large internal stresses in the structure.

The strength (or hardness) of annealed metals usually increases under irradiation and typically both the yield strength and the ultimate tensile strength increase. Unfortunately, this increase in hardness is generally accompanied by a loss of ductility, or embrittlement (material fails under small levels of strain). Embrittlement can have several causes; the most important is related to the large amount of helium gas that is produced in the structural material through (n,α) reactions. At high temperatures, the helium atoms become mobile and diffuse to the boundaries between grains in the metal. Subsequently, helium bubbles form and intergranular failure occurs.

In addition to swelling and embrittlement, physical, mechanical, and thermal properties of the material change by neutron irradiation (e.g., through the production of point defects or by the change in the composition of the alloying elements). As mentioned before, the irradiation effects are very sensitive to both neutron flux and fluence and also to the operating temperature of the material. Chemical interaction among the structural and breeding materials and the coolant, as well as gradients in the temperature, flux, and stress, can also have a dramatic impact on the response of the material to radiation and on its lifetime.

During the 1970s, most of the research effort on structural materials was focused on stainless steels because of the experience in the fission industry. During the 1980s, research focused on low-activation structural material (nuclear waste produced in a reactor built with these materials does not require geological disposal as is needed for fission plants). Primary materials in this category are low-activation ferritic steels (in which induced activation is reduced by substituting for certain alloying elements) and vanadium alloys. Recently, ceramic composites such as silicon carbide (SiC) and fiber-reinforced SiC-matrix composites have received considerable attention. Ceramic composites are of particular interest because the level of induced radioactivity is very low, as shown in Fig. 6 (6 orders of magnitude lower than a metallic structure one day after shutdown).

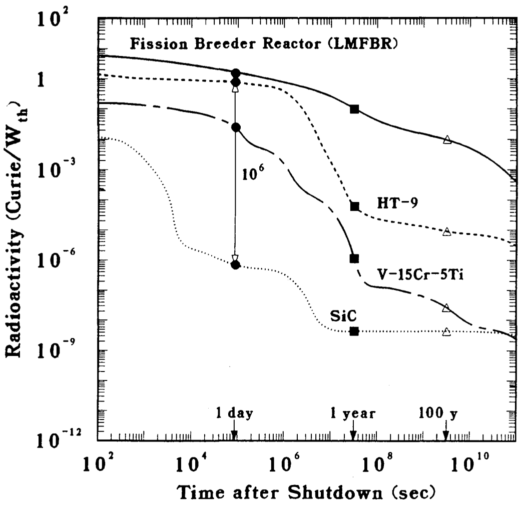

FIG. 6. Normalized radioactivity inventories after shutdown in D-T fusion reactors made of various structural material: HT-9 ferritic steel, V–15Cr–5Ti vanadium alloy, and SiC composites. For comparison, radioactivity levels of a liquid-metal fast-breeder reactor are also shown.

Because of the radiation damage, the lifetime of the first wall and blanket of a D-T reactor is limited and they should be replaced periodically (about every three to six years would be acceptable). Unfortunately, experimental radiation-damage data are limited to a small number of materials and to a low value of neutron fluence. Furthermore, experiments are usually carried out in fission reactors (14-MeV fusion neutrons cause a much higher production of gases, which drastically impacts the material behavior). Fusion-reactor–material test facilities, therefore, have been proposed that incorporate an intense 14-MeV-neutron source.

3.1.4 Conceptual Designs. In addition to liquid lithium or lithium eutectic compounds, water and helium gas have been considered as the coolant. In principle, engineering issues such as chemical compatibility and operating temperature determine the coolant once the structural and breeder material are chosen. Three design types are favored at present. The first is a water-cooled system with steel struc-

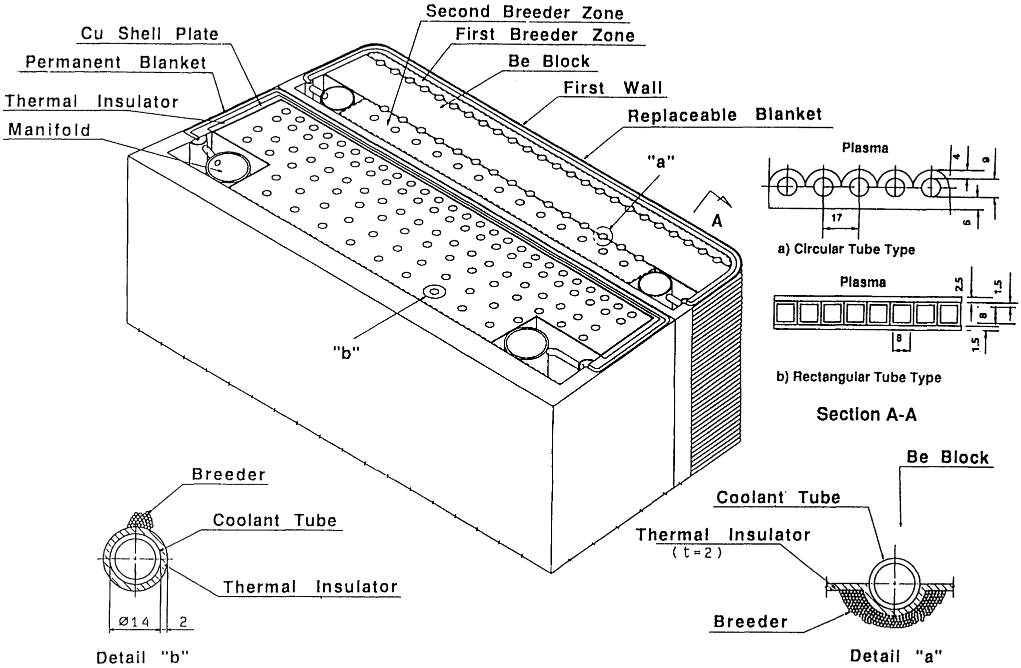
Second Breeder Zone
First Breeder Zone
Cu Shell Plate
Permanent Blanket
Be Block
Thermal Insulator
First Wall
Manifold
Replaceable Blanket
"a"
A
"b"
Plasma
4
9
17
6
a) Circular Tube Type
Plasma
2.5
1.5
8
8
1.5
b) Rectangular Tube Type
Section A-A
Be Block
Breeder
Coolant Tube
Coolant Tube
Thermal Insulator
(t = 2)
Thermal Insulator
Ø14
2
Breeder
Detail "b"
Detail "a"

FIG. 7. Internal structure of one of the blanket modules of the steady-state tokamak reactor, SSTR-1 (Seki *et al.*, 1991). This design uses stainless steel as the structural material, pressurized water as the coolant, beryllium as neutron multiplier, and solid tritium breeding material (figure courtesy of Y. Seki of Japan Atomic Energy Research Institute).

ture and a solid breeder material such as lithium oxide (Fig. 7). This combination requires a neutron multiplier (e.g., Be) in order to achieve a TBR > 1. This design builds upon pressurized-water fission-reactor experience. The main drawbacks to this design are that the outlet temperature of the water coolant is low, which results in a low thermal-to-electrical conversion efficiency, and that the radioactivity induced in the steel structure, even though lower than a fission reactor, is still large by fusion-reactor standards. Tritium leakage into the water coolant also requires expensive tritium-separation systems.

The second type of design (Ehst *et al.*, 1987) uses vanadium alloy as the structural material and liquid lithium as the coolant (no multiplier). Actually, because most of the energy is deposited in the lithium itself, the lithium is a heat-transport fluid rather than a coolant—a major design simplification. Because lithium is electrically conductive, the magnetohydrodynamic (MHD) pressure associated with a conductive fluid moving perpendicular to a magnetic field can lead to high pressure and high pumping power. Reducing the pressure and pumping power to a manageable level will require either a low-velocity, large-area flow channel or nonconducting coolant-channel walls. The thermal-to-electrical conversion efficiency is high. From safety and environmental viewpoints, it ranks better than the water–steel design discussed above; however, lithium fires, which may be caused by an accident, are safety concerns. Lastly, vanadium manufacturing is not as mature as that of steel.

The third class of designs uses a ceramic composite such as SiC composite as structural material, inert helium gas as the coolant, and a solid breeder material such as lithium oxide (Fig. 8). This type of design can achieve the full safety and environmental potential of fusion because the radioactivity inventory is several orders of magnitude lower than fission reactors, and the level of decay afterheat is similarly smaller. Although ceramic composite manufacturing is still in its infancy, it is being pursued vigorously by the aerospace industry.

3.1.5 Tritium-Recovery Systems. Tritium that is bred in the blanket should be recovered rapidly to minimize the T inventory

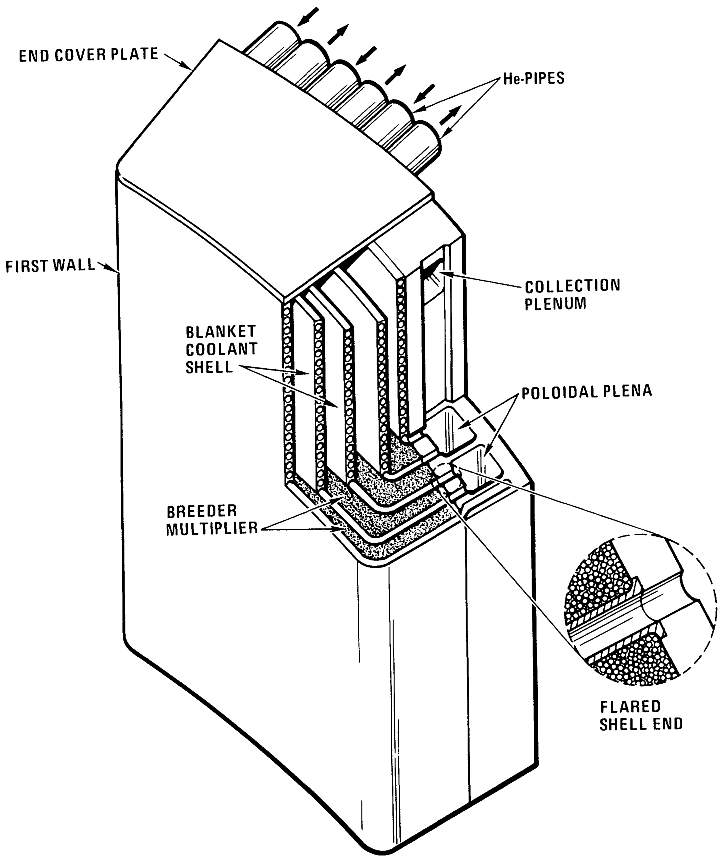

FIG. 8. Internal structure of one of the ARIES-I blanket modules. This design uses SiC composite as the structural material, helium gas as the coolant, Be as the neutron multiplier, and lithium zirconate tritium breeders (Najmabadi *et al.*, 1991).

in the reactor. Designs with solid breeder material rely on diffusion of T out of the solid. As such, the breeder is usually fabricated in small pellets to increase the surface area of the solid. The temperature of the breeder material is also chosen to maximize T diffusion. The T diffusing from the breeder is transported out of the blanket by a purge stream of helium gas. In designs with liquid-lithium breeder, T is in the form of lithium tritide. A small side stream of liquid lithium is brought into contact with an equal volume of a molten lithium salt. Lithium tritide tends to pass to the salt stream and can be readily separated because it is several times lighter. Tritium is then recovered by electrolysis. The molten-salt process is similar to that developed to reprocess fission-reactor fuel.

3.2 Power-Conversion Systems

Most conceptual fusion-reactor designs utilize thermal-to-electrical power-conversion systems similar to those in existing fission or coal-burning power plants. The energy produced in the reactor is used to heat a gas (usually steam) to high temperatures. The hot gas, in turn, expands and cools down in a turbine, producing mechanical energy that is converted to electricity. Of the three types of blanket designs discussed above, the water-cooled system probably needs an intermediate coolant loop and has a low conversion efficiency. Liquid-metal–cooled designs also require an intermediate liquid-metal loop because tritium is present in the primary coolant. Gas-cooled designs can either use a Brayton cycle (coolant sent directly to the turbine) or advanced Rankine steam cycles similar to those proposed for near-term coal-fired power plants. Intermediate loops can be eliminated from these designs if care is taken to keep the tritium separated from the primary coolant.

The power produced in the fusion plasma is carried away by radiation (microwaves and x rays) and by charged particles with energies from several tens of eV (scrapeoff-layer plasma) to several keV (bulk plasma) to the MeV range (fusion by-products) depending on the magnetic-confinement concept. Because the "effective" temperatures of the charged particles or radiation photons are very high, the Carnot bound for the efficiency of converting their energy into electricity is virtually 100%. Furthermore, such direct-conversion schemes can avoid expensive and complicated "conventional" thermal-power-conversion systems.

Several schemes have been proposed for converting charged-particle energy directly into electricity. Such a system would operate as an inverse of a particle accelerator (i.e., the charged particles are slowed down and absorbed by a series of electrodes at different voltages). Since the plasma particles move readily along the magnetic field lines, direct conversion of charged-particle energy is more amenable to confinement schemes in which some field lines leave the main plasma (a field-reversed configuration as opposed to a tokamak). Furthermore, each electrode only absorbs particles with energies close to its biasing voltage. Therefore if the plasma exhaust has a Maxwellian energy distribution, the conversion efficiency for a practical system (limited number of electrodes) is not very high. In certain magnetic-confinement concepts (such as mirrors), the plasma has a positive voltage with respect to the ground and, therefore, the energy distribution of exhausted plasma particles is narrow. In this case, conceptual designs predict a conversion efficiency of ~80% (Logan *et al.*, 1984).

Several schemes have also been proposed for converting plasma radiation into electricity. In one proposal, the plasma radiation is utilized to superheat the coolant for use in an advanced thermal-conversion scheme such as MHD generators. In another, solid-state rectennas (a rectenna is combination of an antenna and a rectifier) have been used to convert synchrotron radiation from the plasma into electricity. These devices have been studied for conversion of microwave energy into electricity (for use with solar-power satellites, which transmit their energy to Earth stations through microwaves). One of the major difficulties, which is similar to that of the charged-particle conversion scheme, is that each rectenna is tuned to a certain frequency range while the synchrotron radiation from the plasma has a broadband spectrum.

3.3 Inertial Fusion-Energy Reactors

The major difference between IFE and MFE reactors is the first-wall design. While the first-wall design is a challenge in an MFE reactor, the design constraints are even more severe for an IFE reactor. This is because of the thermal and structural fatigue caused by millions of microexplosions as well as erosion caused by debris. Any beam-steering or -focusing element that is in direct sight of the explosion should also be protected.

Several innovative designs have been suggested for an IFE reactor. They include

1. inclusion of sacrificial coating on the first wall that would be reapplied *in situ* occasionally,
2. a thick flowing layer of small lithium ceramic granulates,
3. a wetted first wall in which liquid metals such as lithium or lead diffuse through the wall onto its surface, and
4. a thick flowing waterfall (or jet) of lithium.

These concepts are reviewed in detail in FUSION, INERTIAL CONFINEMENT, Sec. 6 and Fig. 12.

In an MFE reactor, the external magnets are usually placed as close as possible to the plasma in order to minimize their costs. The first wall, blanket, and shield are, therefore, squeezed into the space between the plasma and the magnet, making maintenance operations difficult. In an IFE reactor, there are no magnets. The first wall and blanket have a more regular geometry (cylindrical or spherical), which would facilitate maintenance.

4. ADVANCED FUELS

In addition to the D-T fusion reaction, the following have cross sections that are large enough to be of interest in terrestrial applications (as expected, all involve hydrogen isotopes):

$$\mathrm{D}+\mathrm{D}\rightarrow{}^{3}\mathrm{He}(0.92\ \mathrm{MeV})+n(2.45\ \mathrm{MeV}),$$

$$\mathrm{D}+\mathrm{D}\rightarrow\mathrm{T}(1.01\ \mathrm{MeV})+p(3.03\ \mathrm{MeV}),$$

$$\mathrm{D}+{}^{3}\mathrm{He}\rightarrow{}^{4}\mathrm{He}(3.67\ \mathrm{MeV})+p(14.67\ \mathrm{MeV}).$$

The two D-D reactions occur with equal probability. The resulting T and ^{3}He produced by these reactions can subsequently undergo D-T and D-^{3}He reactions. Roughly 40% of the fusion energy is released in neutron power in a D-D reactor. Even though the D-^{3}He reaction does not produce neutrons, D-D reactions occur in a D-^{3}He plasma, which lead to neutron production.

These advanced-fuel reactions require a triple product that is ~ 10 times larger than that required for D-T fusion. They also require a higher plasma temperature (40–60 keV) and, obviously, require more efficient confinement concepts. It also appears that the required pellet gain for an IFE reactor is so large that an advanced-fuel IFE reactor may be impractical.

The major advantage of advanced fuels is that no T breeding is required, making blanket design less difficult. Tritium-handling systems, however, are still needed because T is produced in the plasma, albeit in smaller quantities. Because the neutron flux in a D-^{3}He reactor is 10 times less than in a D-T reactor, the first-wall and blanket component may last the reactor lifetime and need not be replaced. In should be noted that the choice of structural material has a more drastic impact than advanced fuels on achieving the safety and environmental features of fusion (the amount of induced radioactivity in a D-^{3}He reactor is only a factor of ~ 5 lower than that in a D-T reactor with similar structural material, while the induced radioactivity can be reduced by up to 6 orders of magnitude by choosing the appropriate structural material). Lastly, ^{3}He is not available in appreciable quantities on Earth. The proposed eventual source of ^{3}He is the regolith of the lunar surface, which contains about 10 parts per billion of ^{3}He, deposited by the solar wind.

5. CONCLUSIONS

Long-term energy needs require energy sources that are inexhaustible and environmentally benign. Fusion is one of the very few long-term energy options.

Achieving and maintaining the high temperatures required for the fusion burn are the most fundamental challenges for fusion-energy research. Technical progress towards achieving a fusion reactor has been dramatic in the past decade. The required fusion temperature has been attained, and several experiments are approaching the required triple product for maintaining fusion burn. These results, together with the D-T experiments planned for JET and TFTR, serve to demonstrate the scientific feasibility of fusion power.

These accomplishments have initiated an internationally coordinated strategy to address the remaining technological issues that will demonstrate the economic feasibility of fusion power. The first step in this direction is the International Thermonuclear Experimental Reactor (ITER), proposed as a joint venture by the European Community, Japan, the Russian Federation, and the United States.

Conceptual fusion-reactor design studies have been carried out since the early 1970s in order to uncover potential problems, to propose solutions that are consistent with other reactor subsystems, and to identify critical areas that should receive priority in experimental programs. These studies have also shown that potential safety and environmental attributes of fusion can be realized at least at the level of a self-consistent conceptual design study.

At present, the national energy strategy of the United States calls for a fusion demonstration power reactor to begin operation, following the ITER program, in 2025. For this goal to be realized, several other major facilities are required in order to develop the necessary technologies. Steady-state tokamak experiments, a high-fluence 14-MeV-neutron source for material testing, and a technology device for testing blanket concepts (usually referred to as the volumetric neutron source) are necessary and are in various stages of planning.

GLOSSARY

Auxiliary Heating System: The system which heats the plasma to fusion temperatures ($\sim 10^8$ K).

Blanket: The components that surround the fusion chamber and convert the fusion energy into heat. In a D-T fusion reactor, the blanket contains lithium or lithium-containing compounds to breed the tritium fuel.

Breakeven: Refers to a fusion plasma condition in which the fusion energy released is equal to auxiliary-heating energy needed to sustain the plasma temperature.

Cryopump: A high-vacuum pump that uses plates cooled to cryogenic temperatures to adsorb gas atoms and molecules.

Current Drive: A system to drive an electric current in a magnetically confined plasma.

Direct Drive: The inertial-confinement fusion method in which the driver directly irradiates the pellet surface.

Driver: A very high-energy laser or a particle beam used for implosion of the pellets (inertial-confinement fusion).

Embrittlement: A process caused by neutron irradiation which leads to a decrease in the ductility and fracture toughness of the material.

Energy-Confinement Time: The average time it takes before the energy escapes from the plasma.

First Wall: First material wall facing the plasma in a fusion device.

Gain: The ratio of fusion energy released during pellet implosion to the driver energy.

Hohlraum: An enclosure in which the energy of the driver is converted into x rays. It is used in the indirect-drive approach to the inertial-confinement fusion.

ICRF Heating: An efficient method of auxiliary heating of the magnetically confined plasmas by using electromagnetic waves. ICRF is an acronym for ion cyclotron range of frequency.

ITER: The acronym for International Thermonuclear Experimental Reactor, a large tokamak experiment proposed as a joint venture by the European Community, Japan, the Russian Federation, and the United States.

Ignition: A fusion plasma condition in which the fusion energy released supplies enough energy to the fusion plasma to sustain the fusion reactions.

Indirect Drive: The inertial-confinement fusion method in which the driver energy is first converted into x rays in a *hohlraum.* The x rays then irradiate the pellet surface.

Inertial-Confinement Fusion: A method for isolating the fusion plasma from the material wall. In this scheme, bursts of fusion energy are produced by using lasers or particle beams to implode small pellets containing deuterium and tritium. The fusion reaction takes place in the short time before the pellet flies apart.

JET: An acronym for Joint European Torus. It is the largest tokamak experiment in the world today and is located at Culham Laboratories in the United Kingdom.

JT60: A large tokamak experiment located at Japan Atomic Energy Research Institute at Naka, Japan.

Low-Activation Material: Certain material for which the radioactivity induced by neutron irradiation is low. In particular, disposal of such material does not require geological storage as is required for nuclear-fission waste. It is proposed that fusion-power reactors will be built of this sort of material.

Magnetic Confinement Fusion: A method for isolating the fusion plasma from the material wall. In this scheme, the plasma is confined in a magnetic bottle.

Neutral Beam Injector: An efficient method of auxiliary heating of magnetically confined plasmas. It employs an intense beam of high-energy electrically neutral particles, since neutral particles can move through the magnetic field unaffected.

Ohmic Heating: Heating of the plasma by the electric current flowing in the plasma.

Pellet Injector: A method for fueling magnetically confined plasmas. A frozen pellet of deuterium and tritium is injected at high velocity (~ 2 km/s) into the plasma.

Plasma: A state of matter; an ionized gas in which electrons are not bound to the nuclei.

Radiation Damage: Processes leading to degradation in the properties of materials, caused by neutron irradiation. Examples includes atomic displacement, swelling, and embrittlement.

Scrapeoff Layer: The region of the plasma in a tokamak in which the field lines intersect a material wall (i.e., limiter or divertor plates).

Shield: A component located behind the blanket of a magnetic-confinement fusion reactor to protect the superconducting magnets.

Superconductivity: A state achieved by certain alloys and compounds in which electrical current can flow without any resistance. It usually occurs at cryogenic temperatures.

Swelling: A process caused by neutron irradiation that leads to an increase in the size of the material.

TFTR: An acronym for Tokamak Fusion Test Reactor. It is a large tokamak experiment located at the Princeton Plasma Physics Laboratory.

Tokamak: The most advanced magnetic-confinement concept.

Triple Product: The product of plasma density, n, temperature, T, and energy-confinement time, τ_E. The necessary condition for a sustained D-T fusion burn is that the triple product exceed 2×10^{21} s keV/m^3.

Tritium Breeding Material: Lithium or a lithium-containing compound included in the blanket of a D-T fusion reactor to produce tritium to fuel the reactor.

Tritium Breeding Ratio: The number of tritium atoms produced in the blanket for each D-T neutron. Since each D-T fusion reaction uses one tritium atom to produce a neutron, a TBR of 1 or greater is required for tritium self-sufficiency.

List of Works Cited

Abdou, M. A., Waganer, L. (1993), *Fusion Technologies, 1992,* New York: Elsevier Science (to be published).

Bourque, R. F., Meier, W. R., Monsler, M. J. (1992), *Fusion Technol.* **21**, 1465–1469.

Ehst, D., Baker, C., *et al.* (1987), "Tokamak Power System Studies: A Second Stability Power Reactor," Argonne National Laboratory Report No. ANL/FPP/86-1 (unpublished).

Logan, B. G., Henning, C. D., *et al.* (1984), "MARS Mirror Advanced Reactor Study Final Report," Lawrence Livermore National Laboratory Report No. UCRL-53480 (unpublished).

Najmabadi, F., Conn, R. W., *et al.* (1991), "The ARIES-I Tokamak Reactor Study—Final Report," University of California, Los Angeles, Report No. UCLA-PPG-1323 (unpublished); also, Conn, R. W., Najmabadi, F. (1991), in: *Plasma Physics and Controlled Nuclear Fusion Research 1990,* Vol. 3. Vienna: International Atomic Energy Agency, p. 569.

Najmabadi, F., Conn, R. W., *et al.* (1992), "The ARIES-III D-^{3}He Tokamak Reactor Study—Final Report," University of California, Los Angeles, Report No. UCLA-PPG-1384 (unpublished); also, Najmabadi, F., Conn, R. W., *et al.* (1993), "The ARIES-II D-^{3}He Tokamak Reactor Study," in *Proceedings of the IEEE 14th Symposium on Fusion Engineering,* San Diego, CA, 1991, IEEE Publication No. 91CH3035-3. Piscataway, NJ: IEEE, p. 213.

Najmabadi, F., Conn, R. W., *et al.* (1992), "The ARIES-II and ARIES-IV Second Stability Tokamak Reactor Studies—Final Report," University of California, Los Angeles, Report No. UCLA-PPG-1461 (unpublished); also, Najmabadi, F., Conn, R. W., *et al.* (1993), in: *Plasma Physics and Controlled Nuclear Fusion Research 1992,* Vienna; International Atomic Energy Agency (to be published).

Seki, Y. *et al.* (1991), "Concept Study of the Steady State Tokamak Reactor (SSTR)," Japan Atomic Energy Research Institute Report No. JAERI-M 91-081 (unpublished); also Kikuchi, M., Conn, R. W., Najmabadi, F., Seki, Y. (1991), *Fusion Eng. Design* **16**, 253–270.

Sviatoslavsky, I. N., Sawan, M. E., Peterson, R. R., *et al.* (1992), *Fusion Technol.* **21**, 1470–1474.

Further Reading

Conn, R. W. (1981), in: E. Teller (Ed.), *Fusion,* Vol 1. New York: Academic, p. 193. A thorough overview of magnetic fusion reactors.

Conn, R. W., Chuyanov, V. A., Inoue, N., Sweetman, D. R. (1992), *Sci. Am.* **266** (4), 102–111. An overview of the proposed International Thermonuclear Experimental Reactor (ITER).

Duderstadt, J. J., Moses, G. A. (1982), *Inertial Confinement Fusion,* New York: Wiley. A widely quoted text book on inertial fusion.

Holdren, J. P., *et al.* (1989), "Report of the Senior Committee on Environmental, Safety, and Economic Aspects of Magnetic Fusion Energy," Lawrence Livermore National Laboratory Report No. UCRL-53766 (unpublished); also (1988) *Fusion Technol.* **13**, 7–56. Thorough and up-to-date reviews of safety and environmental aspects of fusion reactors.

Rose, D., Clarke, M. (1961), *Plasma and Controlled Fusion,* New York: Wiley. A widely quoted early work on magnetic fusion.

GALAXIES AND COSMOLOGY

DAVID LAYZER, *Department of Astronomy, Harvard University, Cambridge, Massachusetts, U.S.A.*

Introduction 21
1. **Ancient and Classical Cosmology** 22
1.1 Greek Cosmology 22
1.2 The Revival of Greek Cosmology 22
1.3 Newton's Cosmology 22
1.4 Kant and Laplace 23
1.5 Cosmology in the 19th Century 24
2. **The Milky Way and Other Galaxies** 24
2.1 Cepheids as Distance Indicators 24
2.2 Shapley's Galaxy 25
2.3 The Role of Dust 25
2.4 Galactic Rotation 26
2.5 The Nature of Spiral Nebulae 26
2.6 Composition of Galaxies 27
2.7 Galaxy Types 27
2.8 Chemical Composition and Age 28
3. **"The Realm of the Nebulae"** 29
3.1 The Cosmological Principle 29
3.2 The Cosmic Expansion 30
3.3 The Spatial Distribution of Galaxies 11
3.4 Prevalence of Dark Matter 31
3.5 The Cosmic Microwave Background 31
3.6 The Age of the Universe 33
4. **Modern Cosmology** 34
4.1 Relativistic Cosmology 34
4.2 The Primordial Fireball 37
4.3 Origin of Helium and Other Light Elements 37
4.4 The Primordial Fireball and Particle Physics 38
4.5 Phase Transitions, Domain Walls, and Magnetic Monopoles 38
4.6 Inflationary Cosmology 39
4.7 Origin and Growth of Density Fluctuations in the Fireball 40
4.8 The Vacuum–Energy Problem 40
4.9 A Cold Universe? 40
4.10 The Cold Universe and Particle Physics 41
Works Cited 42

INTRODUCTION

Cosmology is the branch of physical science that treats the structure, origin, and evolution of the physical universe as a whole. Although people have speculated about these questions for millennia, cosmology became a full-fledged science, in the modern sense, only in the 20th century, when telescopes powerful enough to yield abundant data about stellar systems outside the Milky Way became available. The history of scientific cosmology falls naturally into three main periods. The first period begins with the ancient Greeks and goes right through the 19th century, at the end of which the prevailing cosmological notions were actually farther from contemporary notions than were those of the Greek atomists. The second period begins in 1912 with Henrietta Swan Leavitt's discovery of the period–brightness relation of Cepheid variables—still the foundation stone of cosmological distance estimates—and ends in 1965. In that year, Arno Penzias and Robert Wilson announced the discovery of the cosmic microwave background, a uniform sea of blackbody radiation at a temperature of 2.7 K. Contemporary cosmology deploys Earth-, balloon-, rocket-, and satellite-based telescopes that explore deep space over virtually the entire electromagnetic spectrum, from radio wavelengths to γ rays. Three questions are central to modern observational cosmology: How are galaxies distributed in space on all scales? What is the rate of cosmic expansion, and how does it vary with distance? And what is the mean

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

cosmic mass density? The answers to these questions are crucial to an understanding of how the universe began and how it has evolved.

Modern theoretical cosmology is dominated by Einstein's theory of gravitation. Relativistic cosmology rests on that theory and on the assumption that, apart from "local" irregularities, the universe looks the same everywhere and in all directions. Relativistic cosmology, together with modern estimates of Hubble's constant (the mean rate of cosmic expansion), predicts that the universe began to expand from an initial state of exceedingly high density between 10 and 20 thousand million years ago. It says nothing, however, about the temperature of that initial state. As discussed below, most workers during the past quarter century have assumed that the cosmic microwave background is the remnant of a primordial fireball. But that scenario, despite its popularity, is encountering serious observational and theoretical difficulties. It is not yet clear whether these difficulties are growing pains or symptoms of a fatal disease.

1. ANCIENT AND CLASSICAL COSMOLOGY

1.1 Greek Cosmology

In the sixth century B.C., Greek natural philosophers constructed the first nonmythical accounts of the structure, origin, and evolution of the physical universe. These early cosmogonies postulated an infinitely extended, uniform, static, and structureless primordial cosmic medium that gradually differentiated under the action of opposed, indwelling tendencies.

The fifth-century atomists Leukippos and Demokritos invented a radically different picture of the world. Unlike their predecessors, they thought of reality not as a continuum but as fundamentally discrete and discontinuous. Behind the world of appearances they imagined an infinite variety of indivisible atoms moving in otherwise empty space. They suggested that the Sun, the Moon, the planets, and the stars are hot, glowing rocks. And they argued that the order of the cosmos was not purposeful but merely a consequence of motions governed by "blind Necessity."

An opposing tradition of Greek scientific cosmology put the Earth at the center of a system of nested concentric spheres containing, respectively, the Moon, the Sun, the planets, and the stars. Eudoxos and other Greek mathematicians tried to account for the observed motions of celestial bodies by attributing to these spheres various combinations of uniform circular motions.

1.2 The Revival of Greek Cosmology

Greek cosmology played an important role in the revival of Greek science that took place in Europe in the 16th and 17th centuries. Copernicus rejected the received cosmology of his day, a geocentric model of the solar system constructed by Ptolemy (second century A.D.) on lines laid down by Aristotle, in favor of the mathematically simpler heliocentric cosmology of Aristarchus (third century B.C.), which he reconstructed from hints that he found in newly translated Greek texts. Copernicus, in the first half of the 16th century, and Kepler, at the end of the century, considered the Sun to be at the center of the universe. The opposing view, that the Sun is just a nearby star and that the stars are uniformly sprinkled throughout an infinite space, was advocated by Giordano Bruno, who was executed by the Church in 1600 for his vigorous advocacy of it. The hypothesis nevertheless grew in popularity, along with the other major tenets of Greek atomism. Huygens and Newton not only accepted the picture of an infinite universe more or less uniformly populated by stars but devised scientifically sound methods for estimating the distances of nearby stars on the assumption that they are intrinsically as bright as the Sun (described in Layzer, 1984). Huygens wrote a popular book, *Cosmotheoros*, expounding the notion that each star has its own family of planets, many or all of which are abodes for life.

1.3 Newton's Cosmology

Newton's theory of gravitation, published in 1687, opened up the possibility of constructing a physical theory of the universe as a whole. The second of Newton's "axioms or laws of motion" states that the acceleration of any object is proportional to the net force acting on it (and inversely proportional to the object's mass). Newton posited that accelerations are to be measured relative to a cosmic reference frame that he called "absolute space." He conjectured that the "fixed stars" are on average at rest in this frame, but his theory did not explain why. Newton's universal law of gravitation states that the gravitational force acting on any particle of matter is the (vector) sum of the gravitational attractions exerted by all the other particles in the universe, each such attraction being proportional to the mass of the attracting particle and inversely proportional to the square of its distance. In a letter to the theologian Richard Bentley written in 1692, Newton discussed the application of his theory of gravitation to the universe as a whole. He argued that if matter was initially distributed evenly within a finite volume of space, it would,

> by its gravity ... fall down into the middle of the whole space and there compose one great spherical mass. But if the matter was evenly disposed throughout an infinite space, it could never convene into one mass; but some of it would convene into one mass and some into another, so as to make an infinite number of great masses, scattered at great distances from one to another throughout all that infinite space. And thus might the sun and the fixed stars be formed, supposing the matter were of a lucid nature.

Newton argued, correctly, that the laws of motion and gravitation cannot by themselves explain why the planets all move around the Sun in the same sense in nearly circular, nearly coplanar orbits. These regularities, he said, are consequences of *initial conditions,* which are not prescribed by his theory and must therefore have been laid down by the Creator. Newton also suggested that the Creator intervenes from time to time to restore the orderly arrangement of planetary orbits, which is constantly being degraded by the planets' mutual gravitational attractions.

1.4 Kant and Laplace

During the second half of the 18th century, Newton's view of the Creator's cosmological role was challenged by Immanuel Kant and by Pierre Simon de Laplace. In his *General History of Nature and Theory of the Heavens* (1755; discussed in Layzer, 1984) Kant proposed a new picture of the cosmos and a new theory, based on Newtonian physics, of its evolution from a simple initial state. Kant's ideas are important not only because they anticipated modern findings but also because they rested for the most part on sound physical inferences from observational evidence.

The Solar Nebula. Kant believed that a diffuse gas cloud would collapse into a spinning, flattened disk. He suggested that the planetary system evolved from such a cloud.

The Galaxy. Developing an earlier suggestion by Thomas Wright, an English amateur astronomer and theologian, Kant inferred from the observed distribution of stars and diffuse starlight on the celestial sphere that the Sun and its neighbors belong to a huge, highly flattened, rotating, self-gravitating swarm of stars whose structure and dynamics are similar to those of the planetary system. Just as the Earth orbits the Sun, so, Kant argued, the Sun orbits the distant center of our stellar system, the Galaxy. Moreover, Kant argued, the Galaxy, like the solar system, evolved from a huge, initially spherical gas cloud that flattened as it shrank.

External Galaxies. Extending Huygens's and Newton's hypothesis that the stars are suns, Kant conjectured that the Galaxy is comparable in size, shape, and intrinsic brightness to innumerable similar systems (external galaxies) sprinkled throughout an infinite space. Kant reasoned that external galaxies would appear in the telescope as faint, extended patches of starlight, because their individual stars would be too faint to be resolved by the telescope. Astronomers had observed such faint, fuzzy patches, and Kant boldly speculated that some of these, at least, were not nearby clouds of diffuse, glowing gas but distant galaxies. Pushing his analogy between the Galaxy and the solar system still further, Kant speculated that galaxies, including our own, are themselves members of huge, flattened, spinning, self-gravitating systems.

Late in the 19th century, Laplace revived and developed Kant's hypothesis that the Sun and planets formed from an initially structureless gas cloud whose dimensions were comparable with those of the present solar system. Laplace argued that as the cloud radiates away energy, its angular momentum remains constant. Most of the mass collapses into a central star, but the matter that carries the indestructible angular momentum remains behind to form a thin, differentially rotating disk. Later, according to Laplace, the planets form from condensations in this gaseous disk. Although Laplace's hypothesis was not—and still has not been developed into—a complete theory, it accounted in a natural way for the most conspicuous regularities of the planetary system and the satellite systems of Saturn and Jupiter. (See ASTRONOMY.)

Laplace also showed by ingenious and elaborate calculations that the orbits of the planets are stable over long periods, thereby demolishing Newton's argument that divine intervention is needed to maintain order in the planetary system. Although Laplace's calculations represented a major advance, they did not completely resolve the issue of the planetary system's long-term stability. Even today important questions remain. Numerical calculations have shown that at least in the inner solar system the planetary orbits are unstable in the sense that infinitesimal deviations from any given orbit increase exponentially with time. But numerical calculations also indicate that the characteristic times for the development of finite deviations (large eccentricities and inclinations) are many orders of magnitudes greater than the characteristic times associated with the exponential growth of infinitesimal deviations from the unperturbed orbit. It has been suggested that a planet's representative point in a "parameter space" whose points represent all possible Keplerian orbits performs a random walk. Finite deviations from the initial orbit would then grow only as the square root of the elapsed time. But a rigorous mathematical theory is still lacking.

1.5 Cosmology in the 19th Century

The 19th century saw the development of powerful mathematical theories of the dynamics and stability of gravitating systems (the work of Henri Poincaré is especially noteworthy) and an accumulation of useful data, but little progress in observational or theoretical cosmology. At the end of the century, most astronomers believed, on the basis of star counts, that the Sun lay close to the center of the Galaxy, and Kant's hypothesis that the "spiral nebulae" are stellar systems comparable in size and composition to the Milky Way was still highly controversial.

2. THE MILKY WAY AND OTHER GALAXIES

The main outlines of our present picture of the astronomical universe emerged between 1912 and 1929. During this brief period astronomers recognized and solved two major problems. They developed a strategy and a set of effective techniques for estimating the distances of exceedingly remote stars and stellar systems, and they recognized the importance of the dimming and reddening of starlight by intervening interstellar dust and learned how to allow for these effects.

2.1 Cepheids as Distance Indicators

The initial breakthrough came in 1912, when Henrietta Swan Leavitt, a research assistant at the Harvard College Observatory, discovered a remarkable property of certain exceptionally bright stars called Cepheids (Leavitt, 1912). These stars vary periodically in brightness, with periods ranging from 1 to 100 days. Leavitt found a number of Cepheids in the Small Magellanic Cloud, a huge cloudlike aggregation of stars in the southern sky, and she measured their periods and mean apparent brightnesses. She discovered that the period of a Cepheid is tightly correlated with its mean brightness. So the period of a Cepheid, which is unaffected by distance or intervening obscuration, is a reliable index of its apparent brightness, and hence—because the Cepheids in the Cloud are all at nearly the same distance—of its intrinsic brightness. A modern version of this plot is shown in Fig. 1.

To calibrate Leavitt's period–brightness relation, Hertzsprung (1914) used a statistical technique to estimate the distances of 13 Cepheids in the Milky Way. This enabled him to convert Leavitt's relation between period and apparent brightness into a relation between period and intrinsic brightness. Now astronomers could deduce the true brightness of a Cepheid from its (easily measured) period, and, by comparing its true and apparent brightnesses, estimate its distance. A new era of astronomical exploration had begun.

Hertzsprung himself made two important applications of the period–brightness relation. He found that the Cepheids and hot blue stars of the Milky Way were, as expected, strongly concentrated toward a plane passing through the Sun, but that within this plane they were

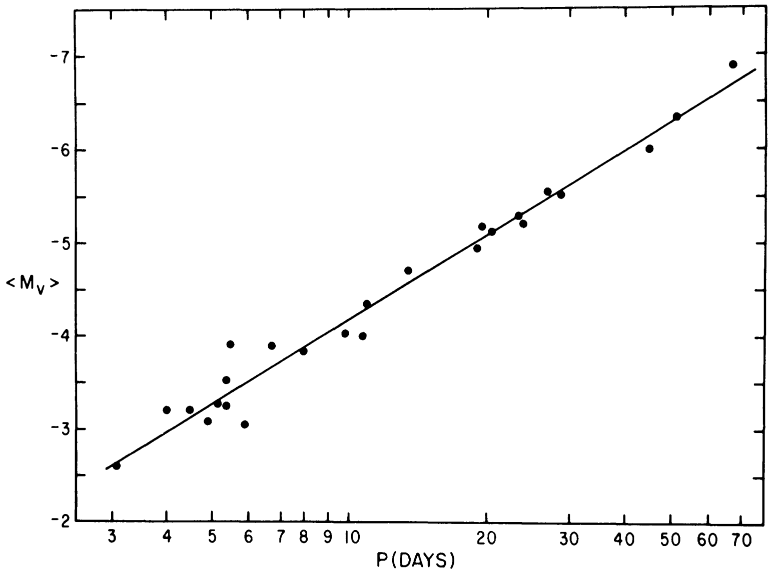

FIG. 1. Period-brightness relation for classical Cepheids belonging to galactic star groups. From Stothers (1983).

not evenly distributed in direction. Their space density increased in the direction of the constellation Sagittarius. Hertzsprung concluded that, contrary to contemporary astronomical opinion but in agreement with Kant, the Sun is not at the center of the Galaxy. Hertzsprung was also able to estimate the distance of the Small Magellanic Cloud. Although his estimate was too small by a factor of 7, it supported the conclusion that this system and its companion, the Large Magellanic Cloud, are external galaxies rather than (as many astronomers believed) star clouds within the Galaxy.

2.2 Shapley's Galaxy

In 1918 Shapley (1918; see also Shapley, 1961, and Hubble, 1936) published the results of an ambitious study of globular star clusters based on data taken with the 60-in. reflector on Mount Wilson, then the world's largest telescope. (Globular clusters are compact, nearly spherical systems, containing from a hundred thousand to ten million stars.) By extrapolating the Leavitt-Hertzsprung period–brightness relation to RR Lyrae stars (Cepheid variables with periods less than 1 day, found in globular clusters), Shapley was able to estimate the distances of 69 clusters. He found that the globular clusters outline a vast spheroidal system whose central plane coincided with that of the Milky Way (the disk-shaped system defined by Galactic Cepheids and hot blue stars). The center of the spheroidal system lay in the direction found by Hertzsprung, but at a much greater distance, and the diameter of the spheroidal system greatly exceeded that of the disk-shaped system.

Shapley's Galaxy was much larger than any astronomer had ever suggested, and the Sun's position was much more peripheral. Two features of Shapley's model, however, were not merely surprising but incomprehensible on the basis of gravitational theory: the precise coincidence between the central plane of the spheroidal system and the central plane of the Milky Way, and the absence of globular clusters in a slab centered on the Galaxy's central plane; see Fig. 2 (left).

2.3 The Role of Dust

These incomprehensible features of Shapley's model of the Galaxy disappear if one postulates a thin layer of obscuring matter, in the form of dust, strongly concentrated toward the common plane of the system of globular clusters and the Milky Way; see Fig. 2 (right). The Milky Way then expands in the direction of the Galactic center, while the system of globular clusters shrinks. Indeed the postulate of a dust layer allows the centers

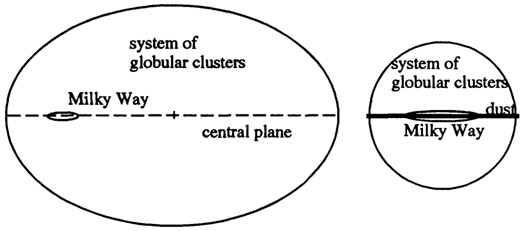

FIG. 2. Left: Shapley's system of globular clusters (large ellipse) had the same central plane as the Milky Way but its diameter was ten times greater. Right: The hypothesis of a dust layer concentrated toward the central plane brings the centers of the two systems together, makes the system of globular clusters more nearly circular, and reduces the discrepancy in size. Both figures are schematic.

of the two systems to coincide and causes the globular clusters to be concentrated toward this new common center. The absence of globular clusters in a slab centered on the central plane becomes a straightforward consequence of obscuration in this plane. Supporters of this idea pointed out that many spiral galaxies, viewed nearly edge-on, show evidence of heavy obscuration in a thin central layer.

Definitive evidence for such an obscuring layer in the Milky Way was presented by Trumpler (1925). Trumpler estimated the distances of 100 so-called Galactic star clusters. Unlike globular clusters, Galactic clusters are found mainly within the flattened disk-shaped system defined by hot blue stars and Cepheids, and they contain, typically, a few hundred stars. Trumpler found that the colors of stars in these clusters vary systematically with the cluster's distance from the Sun: the more distant the cluster, the redder its stars. In addition, the diameters of the clusters, inferred from their distances and angular diameters, increased systematically with distance. Both systematic effects were explained by the postulate of a uniformly dense layer of dust in the Galactic plane, which would both redden and dim the light of distant stars.

In 1932, Stebbins (1933) succeeded in inferring the thickness of the obscuring layer from photoelectric measurements of the colors of globular clusters. This enabled him to correct Shapley's estimates of the distances of globular clusters, which were then seen to form a nearly spherical system, densest at its center, with the Sun at a distance of 30 000 light-years. This value is close to current estimates.

2.4 Galactic Rotation

Kant's hypothesis of 1755 that the stars of the Milky Way revolve about a distant center under the action of gravity, like planets around the Sun, was revived in 1925 by Lindblad (1925), who showed that certain striking statistical properties of the observed motions of nearby stars relative to the Sun are explained by the hypothesis that stars in the Sun's neighborhood are moving in nearly circular orbits about a distant center. Oort (1927) looked for and found evidence of differential rotation in the motions of distant stars in the Galactic plane. He showed that the velocity of a star in a circular orbit about the Galactic center varies with the radius of the orbit just as the orbital velocity of a planet varies with its distance from the Sun. Oort's analysis yielded the first well-founded estimates of the period of the Sun's orbit around the center of the Galaxy (about 10^8 years), the radius of the orbit (about 30 000 light-years), and the mass inside the orbit (about 10^{11} solar masses).

2.5 The Nature of Spiral Nebulae

While the Galaxy was taking shape, astronomers were assembling evidence and arguments bearing on Kant's hypothesis that "spiral nebulae" are galaxies comparable in size and composition with the Milky Way. In 1917 G. W. Ritchey photographed a nova in a spiral nebula, using the 60-in. telescope on Mount Wilson (see Hubble, 1936). Ritchey and Heber D. Curtis of the Lick Observatory immediately began to search the photographic-plate collections of the Mount Wilson and Lick Observatories for evidence of other novae in directions well away from the plane of the Galaxy. They soon found five more, all in spiral galaxies. Curtis argued that because these novae were 10 000 times fainter than the faintest novae in the Milky Way, they had to be 100 $(=\sqrt{10000})$ times as far away; he showed that the linear diameters of spiral galaxies, inferred from their angular diameters on the assumption that their novae have the same intrinsic brightness as Galactic novae, were indeed comparable to the diameter of the Milky Way.

Six years later, Hubble (1936) published the results of a detailed study of three bright nebulae, including the spiral nebula in Andromeda. Plates made with the newly completed 100-in. reflector on Mount Wilson resolved many individual stars in each of these objects, including a number of Cepheids, from which Hubble could estimate the distances of the three systems by using the Leavitt-Hertzsprung relation between period and intrinsic brightness. The derived distances agreed well with those estimated earlier by Curtis; and the statistical properties of the stellar populations in all three systems resembled those of stars in the Galaxy.

2.6 Composition of Galaxies

In calibrating his distance indicators, Hubble assumed that samples of stars drawn from different galaxies or from different regions in the same galaxy have the same statistical properties—that they are drawn, in effect, from a single, homogeneous population. Distance estimates based on this assumption implied that globular clusters in the Andromeda galaxy are intrinsically four times fainter than their counterparts in the Milky Way. This conclusion was troubling, because it cast doubt on its own premise: if globular clusters differ systematically from one galaxy to another, why shouldn't the stars of which they are composed also differ systematically from one galaxy to another?

In 1944 Walter Baade proposed that there are two qualitatively distinct stellar populations; see Baade (1956, 1963). Population I inhabits the disks of spiral galaxies while Population II inhabits the spheroidal component of spiral galaxies and nearly all of elliptical galaxies. Thus spirals are mixtures of the two populations, with population I dominant in the disk and population II in the nucleus or central bulge and in the halo, while elliptical galaxies are nearly pure population II. Population I includes the Sun and its neighbors. Its most distinctive constituents are hot blue stars (supergiants), long-period Cepheids, galactic clusters like the Hyades and the Pleiades, and interstellar dust and gas. In a first approximation, these constituents have the same statistical properties in all spiral and irregular galaxies. Typical members of population II include globular clusters, RR Lyrae stars, and stars with very low abundances of elements heavier than helium. Although astronomers now subdivide both populations into subpopulations, Baade's original distinction is still meaningful and important.

Baade suspected that Cepheids in the two stellar populations would turn out to obey different period–brightness relations. A decisive test of this conjecture had to await the development of telescopes and photographic techniques capable of producing photometric images of population II stars in an external galaxy, where their apparent brightness could be compared directly with the apparent brightnesses of bright population I Cepheids. (In the Galaxy, this comparison depends on a large and uncertain correction for interstellar dimming.) In 1950 the 200-in. reflector on Mount Palomar became functional. The first long-exposure photographs of the Andromeda galaxy made with this instrument contained excellent images of the red giants that are the brightest stars in population II. Knowing the brightness ratio between such stars and RR Lyrae stars in the Milky Way, Baade could infer the apparent brightness of RR Lyrae stars in Andromeda, although these stars were too faint to show up on the 200-in. plates. The inferred brightnesses of Andromeda's RR Lyrae stars were smaller by a factor of 4 than they would have been if they conformed to Shapley's extrapolation of the Leavitt-Hertzsprung period–brightness relation for population I Cepheids. Confirming evidence followed quickly, when RR Lyrae stars were recorded in globular clusters of the Large and Small Magellanic Clouds: their apparent brightnesses agreed perfectly with Baade's prediction. The inferred brightness of Andromeda's globular clusters—and indeed of the Andromeda Galaxy—could now be increased by a factor of 4, making them and it comparable, respectively, to our own globular clusters and the Milky Way.

2.7 Galaxy Types

From their appearance on photographic plates, Hubble assigned galaxies to two great classes: spirals and ellipticals. The photographic images of ellipticals are smooth and regular, and the contours of constant surface brightness are concentric ellipses with an axial ratio greater than 0.3. The photographic images of spirals suggest a more or less foreshortened disk marked by bright knots and dark lanes arranged in a more or less regular spiral pattern. Seen edge-on, spirals show dark equatorial lanes and often, though not always, a central bulge.

These differences between the photographic appearances of ellipticals and spirals reflect basic structural differences. The ellipticals are spheroidal (or perhaps ellipsoidal) stellar systems in which the space density of stars increases strongly and smoothly toward the center. The disks of spirals are more highly flattened and less centrally condensed than ellipticals. The spiral markers—hot blue stars, often in compact groups, glowing ionized gas clouds, dark clouds, dark lanes—are all characteristic of Baade's population I. But most spirals also contain a spheroidal component of population II objects. This component contributes the central bulge (if there is one) and a spheroidal halo, as in our Galaxy.

The photographic images of spirals differ in several conspicuous ways, including the relative size of the spheroidal nucleus, the pitch or openness of the spiral arms, the conspicuousness of spiral markers, and the conspicuous of a central bar. Hubble showed that the first three characteristics vary together and thus define a linear sequence. Figure 3 shows Hubble's original "tuning-fork" diagram. Figure 4 shows exemplars of some of the galaxy types. The sequence begins with galaxies that have large spheroidal nuclei and disks but no spiral markers; these are designated by the symbol S0. As we proceed along the sequence, the relative size of the spheroidal nucleus decreases, and spiral markers appear, defining tightly wound, nearly circular arms (type Sa). Between types Sb and Sd, the markers become more conspicuous, the arms open out, and the spheroidal nucleus becomes less conspicuous, eventually disappearing entirely. The sequence ends with galaxies like the Magellanic Clouds (Sm), which are disks with extremely open arms and exceptionally bright spiral markers. The fourth characteristic, conspicuousness of a central bar, adds a second dimension to the classification scheme.

More recent classification schemes use additional characteristics, such as the relative thickness of the spiral arms and the conspicuousness of a ring enclosing the central bar.

2.8 Chemical Composition and Age

Studies of stellar evolution have shown that the stars that make up spheroidal galaxies and the spheroidal components of spirals—population II—are very old. The ages of population II stars do not differ significantly from the age of the universe (Sec. 3.6). Population I is younger than population II, but the ages of population I stars span a wide range, from a million years or less to a few thousand million years. The bright blue stars and clumps of bright blue stars that largely define spiral arms have ages of only a million to a few million years.

The two populations also differ markedly in chemical composition. Spectroscopic analyses and studies of stellar evolution have shown that elements heavier than helium are

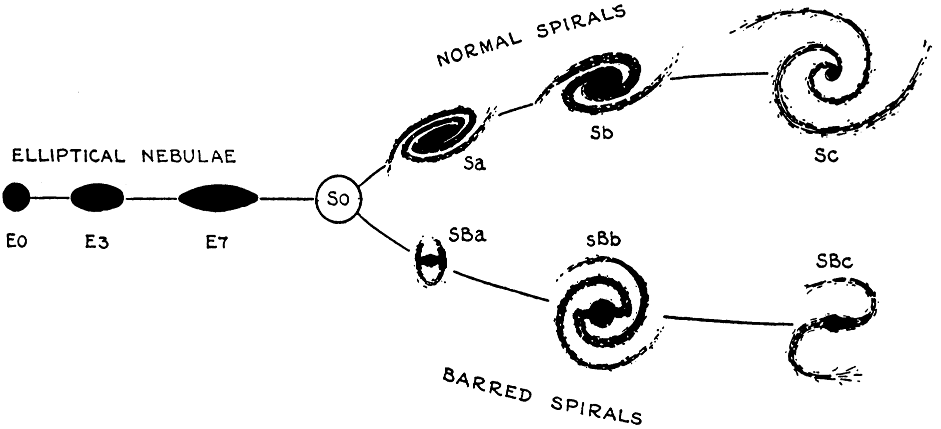

FIG. 3. Hubble's "tuning-fork" diagram. From Hubble (1936).

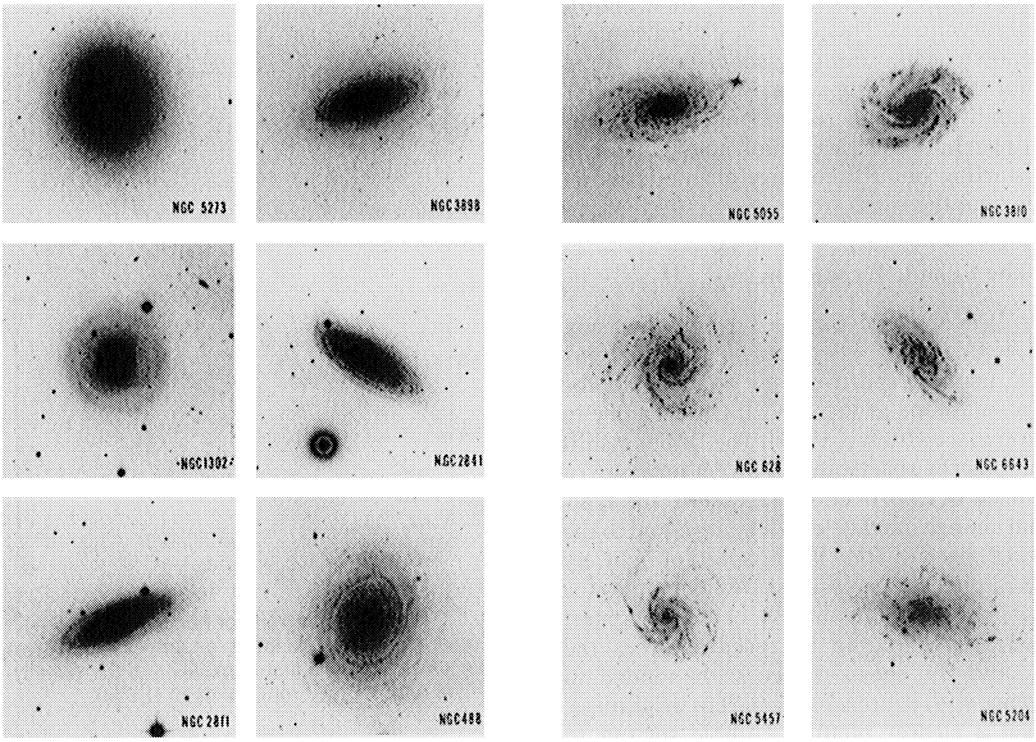

FIG. 4. Illustrating the progressive change of appearance along the classification sequence, from type S0 (upper left) to Sm or Irr (at the bottom right). From Sandage (1975a).

less abundant in population II stars than in population I stars by factors of 10 to 100 or more.

These observational findings suggest the following evolutionary hypotheses.

1. Population II stars, globular star clusters, and the spheroidal components of galaxies formed early in the history of the universe from "primordial" material consisting almost entirely of hydrogen and helium.
2. Chemical elements heavier than helium were synthesized in the deep interiors of short-lived massive stars, which eventually exploded, ejecting heavy-element–rich gases into interstellar space. As the enriched gas cooled, it settled into a thin disk, along with dust formed from the gas.
3. Population I stars formed from gas enriched (or contaminated) by this heavy-element–rich gas.

More detailed evolutionary scenarios have been proposed but are still highly speculative.

3. "THE REALM OF THE NEBULAE"

3.1 The Cosmological Principle

Having confirmed that the "spiral nebulae" are galaxies comparable to our own stellar system, Hubble set out in 1923 to chart their distribution in space, using data from the 100-in. telescope. Allowing for the effects of obscuration by dust in the Galaxy's central plane, Hubble concluded from counts of galaxies in small patches randomly distributed over the whole sky that the spatial distribution of galaxies is statistically uniform—that its statistical properties are the same in all directions and at all depths. In *The Realm of the Nebulae* (Hubble, 1936) he summed up his findings in these words: "There is no evidence of a thinning-out, no trace of a physical boundary. There is not the slightest suggestion of a supersystem of nebulae isolated in a larger world." Hubble's study provided the first solid evidence to support the ancient hypothesis that the astronomical universe is unbounded and that it looks much the same from every vantage point and, from a given vantage point, in all directions. This hypothesis has come to be known as the *cosmological principle.* Since Hubble's pioneering work, the quantity and the quality of the astronomical evidence supporting it have increased enormously. It forms the starting point for nearly all modern cosmological theories.

3.2 The Cosmic Expansion

In 1912 Vesta M. Slipher began to measure Doppler shifts in the spectra of external galaxies, using a spectrograph attached to the 24-in. telescope at the Lowell Observatory in Arizona (see Hubble, 1936). The Doppler shift in an object's spectrum is a direct measure of its speed of approach or recession, the fractional change in wavelength being equal to the object's speed divided by the speed of light. By 1925 Slipher had measured Doppler shifts for 45 galaxies. All but a few of the brightest galaxies appeared to be receding from the Sun, and the speed of recession tended to increase with decreasing apparent brightness. To get a clearer picture, Hubble set out to estimate the distances of the galaxies in Slipher's sample, using a strategy similar to the one Shapley had used to estimate the distances of globular clusters: He used Cepheids in nearby galaxies to calibrate brighter distance indicators, which he then used to calibrate still brighter distance indicators, and so on. In 1929 Hubble published a plot of Doppler shift versus distance for 22 galaxies on Slipher's list. The data suggested a linear relation between distance and redshift, with considerable scatter in redshift, especially among nearby galaxies. Later Hubble and Milton Humason were able to extend the plot to much fainter galaxies, strongly confirming the suggestion of a linear relation between redshift and distance. Figure 5 shows a modern version of the empirical redshift–apparent-brightness relation. Deviations from this relation can be attributed to clumpiness in the spatial distribution of galaxies, which gives rise to local gravitational fields and local gravitationally induced motions.

The ratio between velocity and distance is called the Hubble constant. Once its value is known, the measured redshifts of galaxies can be used as distance indicators. Our present maps of the universe of galaxies rest mainly on measured redshifts.

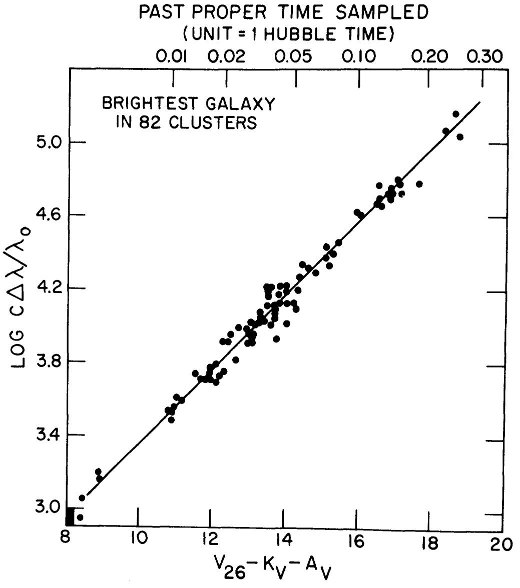

FIG. 5. The Hubble diagram for 82 first-ranked cluster galaxies. The box in the lower left corner is the range within which Hubble's first formulation of the velocity-distance relation lay. From Sandage (1975b).

Present estimates of the Hubble constant range from 45 to 90 (km/s)/Mpc (1MPc=1 megaparsec=3.26 million light-years). The estimates cluster around two widely separated values, 50 and 85 (km/s)/Mpc. The wide range of estimates reflects in part the variety of objects that have been used as distance indicators beyond the range of Cepheids (the original, and still the primary, extragalactic distance indicators). Such objects must have observable characteristics tightly linked to intrinsic brightness, in the way that the period of light variation of a Cepheid is tightly linked to its intrinsic brightness; and they must be bright enough to be visible at distances well beyond the scale of "local" structure. The most popular current indicators are spiral galaxies with measured rotation curves (Tully and Fisher, 1977; see also Sec. 3.4), giant elliptical galaxies with measured internal velocity dispersions, novae, supernovae of a particular type (Ia), and planetary nebulae. Estimates of the Hubble constant also differ because astronomers disagree about how to correct the data for the effects of large-scale clustering on the measured velocities. Finally, there is disagreement about how to correct in specific cases for the so-called Malmquist effect: in any sample limited by apparent brightness, the intrinsically brightest members of a given class of objects—supernovae of a particular type, say—are overrepresented. At the time of this writing (August 1992), excellent cases have been made for both the high and the low values of the Hubble constant; see Huchra (1992) and Tammann (1992).

As discussed in Sec. 3.6, the Hubble constant is one of two measurable quantities that determine the age of the universe; the other is the mean cosmic density.

3.3 The Spatial Distribution of Galaxies

Studies of the spatial distribution of nearby galaxies have revealed a pronounced tendency toward hierarchic clustering.

Many galaxies belong to binary and small multiple systems. Many, perhaps most, large- and medium-sized spirals have smaller satellites. For example, the Magellanic Clouds are satellites of the Milky Way.

Single galaxies and small multiple systems often belong to groups with diameters ranging from 3 to 12 million light-years. The Local Group contains two main subgroups, one dominated by the Galaxy, the other by Andromeda, as well as a dozen or so dwarf galaxies not attached to either subgroup.

The Local Group and many other nearby groups belong to less distinct aggregates called clouds, and some groups and clouds are clumped in clusters with as many as a thousand or more visible members and diameters of up to about 30 million light-years. Some clusters contain subclusters; others have a smooth, spherically symmetric structure analogous to the structure of globular star clusters.

Galaxy clusters are not uniformly distributed in space; there is a higher order of clustering—superclustering, first detected in the late 1930s. But in the late 1970s evidence began to accumulate that the character of the spatial distribution of galaxies changes qualitatively beyond the level of clusters. By the mid-1980s it had become clear that galaxies and galaxy clusters are concentrated in sheetlike structures that form the interstitial regions of a vast network of "bubbles"—convex, significantly underdense regions with diameters of the order of 100 to 300 million light-years (for a review, see Rood, 1988). The large-scale distribution of galaxies and galaxy clusters resembles the distribution of soap in soap suds; see Fig. 6. Thus a qualitative change in the character of clustering seems to take place at the very largest scales on which the distribution is nonuniform. On scales up to the scale of galaxy clusters, the overdense regions are convex. At the scale of superclustering, it is the underdense regions that are convex.

3.4 Prevalence of Dark Matter

Newton's theory of gravitation allows us to infer the mass of a self-gravitating system from the velocities of its members and its diameter. For example, we can infer the mass of the Sun from the Earth's orbital velocity and orbital radius, and the mass of the Earth from the Moon's orbital velocity and orbital radius. Application of this principle has led to the conclusion that *a large fraction—indeed the bulk—of the mass gravitationally bound in galaxies and larger systems is nonluminous, not only in visible light but also at shorter (ultraviolet, x-ray, γ-ray) and longer (infrared, millimeter and submillimeter, radio) wavelengths.*

Several different kinds of evidence, including those listed below, support this conclusion.

1. Neutral atomic hydrogen emits and absorbs radiation at a rest wavelength of 21 cm. In spiral galaxies, atomic hydrogen is concentrated in a thin sheet that coincides with the galaxy's central plane. Parcels of atomic hydrogen circle the galactic nucleus like planets around a central sun. When we look at a spiral galaxy edge-on, we see, on one side of the center of the disk, parcels of atomic hydrogen that are approaching us; on the other side, parcels that are receding. Measurements of the Doppler shift of the 21-cm line at different angular distances from the center of the disk, corrected for foreshortening, yield the orbital velocity of hydrogen in the galaxy's central plane as a function of radius (provided the distance of the galaxy is known). From this "velocity profile" one easily calculates the mass within each radius for which the velocity has been measured. Using this technique, Rubin and colleagues (for a summary, see Rubin, 1979) have shown that the mass density in a typical bright spiral decreases

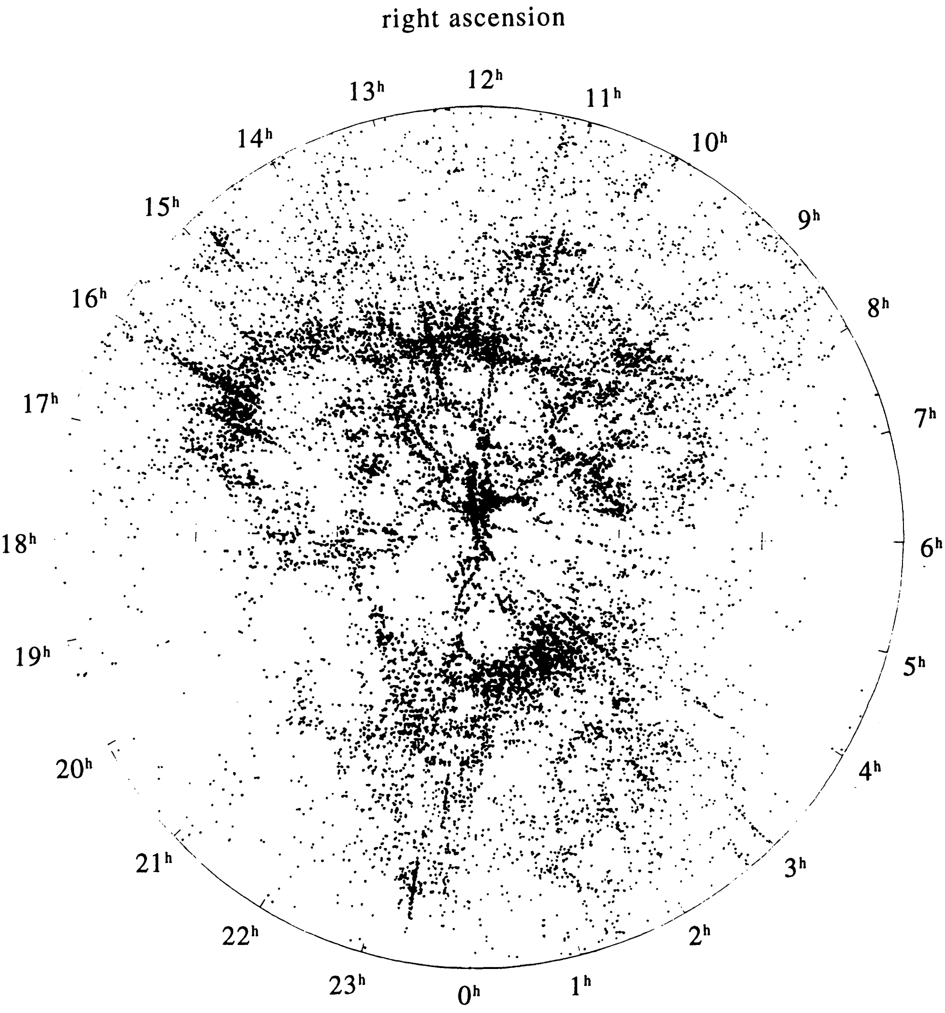

FIG. 6. Large-scale distribution of galaxies. A "pie slice" of the Universe about 500 million light years in radius, with the Earth at the center. The map represents a wedge bounded by celestial longitudes 8.5 degrees and 44.5 degrees. Convex underdense structures some 100 million million light years in diameter are clearly visible. Diagram courtesy of John Huchra, Smithsonian Astrophysical Observatory.

with radius much more slowly than the brightness. The actual radius of a typical bright spiral probably lies between 10^5 and 3×10^5 light-years, as much as 10 times the radius of the visible galaxy. The masses of bright spirals estimated in this way are typically of the order of 10^{12} solar masses, about 10 times the mass associated with visible stars.

2. Estimates of the mass of the Galaxy based on the velocities of distant globular clusters and small satellite galaxies also yield values of order 10^{12} solar masses.
3. Doppler-shift measurements for galaxies in small multiple systems, groups, and clusters yield estimates of the total mass and the mass:light ratio.

These and other kinds of evidence consistently indicate that 90% to 95% of the mass bound in self-gravitating systems is not in the form of visible stars. Observations at radio, ultraviolet, and x-ray frequencies show that only a small part of the invisible mass is in the form of interstellar and intergalactic gas (though in galaxy clusters intergalactic gas does contribute significantly). In what form does the dark mass reside?

The most conservative candidates are compact stellar remnants (white dwarfs, neutron stars, and more massive objects that are in the process of collapsing into black holes); objects of planetary or subplanetary mass; and very faint stars of low mass. All these possibilities, however, conflict with inferences that have been drawn from a widely held cosmological hypothesis. This hypothesis, discussed in Sec. 4.2 below, posits that the cosmic microwave background (Sec. 3.5) is the remnant of a primordial fireball. As discussed in Sec. 4.3, the results of calculations of helium production in the fireball can be reconciled with astronomical estimates of the primordial abundance of helium only if much (or perhaps all) of the nonluminous mass in the universe is in some form other than ordinary matter.

3.5 The Cosmic Microwave Background

In 1964–1965 Arno Penzias and Robert Wilson of Bell Laboratories, Holmdel, N.J., stumbled on an important new constituent of the universe (Penzias and Wilson, 1965). While calibrating a radiometer tuned to a wavelength of 7.3 cm, they discovered a signal that could not be attributed either to their instruments or to external sources of known kinds. The signal was equally strong in all directions and showed no diurnal or seasonal variation. Meanwhile, in Princeton, N.J., a few miles from the observing site where Penzias and Wilson had set up their radiometer, another group, headed by Robert Dicke, was preparing an experiment to detect radiation, at a wavelength of 3 cm, with precisely these properties. A mutual colleague put the two groups in touch.

Dicke's group was looking for evidence of an early, extremely hot, radiation-dominated phase in the history of the universe. In 1948, George Gamow, Ralph Alpher, and Robert Herman had posited such a "primordial fireball" on theoretical grounds, and had predicted that it would leave a relic in the form of a uniform, isotropic sea of thermal (blackbody) radiation with a temperature of a few degrees above absolute zero; see Peebles (1971). The isotropic radiation field discovered by Penzias and Wilson fitted this prediction. Its measured temperature was close to 3 K. A few months later Dicke's group detected the radiation at 3 cm and derived the same value of the temperature. Measurements carried out with the COBE (Cosmic Background Explorer) satellite in 1990 (Mather *et al.*, 1990) showed that, over the whole of the relevant wavelength range, the spectrum of the radiation deviates by less than one part in a hundred from the spectrum of a blackbody at a temperature close to 2.7 K.

Measurements carried out from the ground, from balloons and aircraft, and from the COBE satellite have yielded important information about angular variations in the temperature of the radiation on angular scales from arc minutes to tens of degrees. Because of the Doppler effect, the radiation should appear brighter in the direction of the Earth's motion relative to the local standard of rest defined by the radiation field, fainter in the opposite direction—and it does. The observed deviations yield an estimate of the velocity of the Galaxy relative to the local standard of rest that is consistent with estimates based on measurements of galactic redshifts; see Jerjen and Tammann(1992).

On arc-minute scales, variations in the temperature of cosmic microwave background are less than 2 or 3 parts in 100 000. As discussed in Sec. 4.4, this upper limit rules out the simplest theories for the origin of structure in a hot universe.

In 1992 measurements made with a differential radiometer on board COBE gave the first strong evidence for deviations from isotropy not due to the Earth's motion relative to the radiation background (Smoot *et al.*, 1992). On scales larger than the radiometer's beamwidth of 7° the temperature variations have a noiselike character. Their amplitude is one part in 10^5 and is independent of the angular scale. The significance of these findings is discussed below (Sec. 4.8).

3.6 The Age of the Universe

Several independent arguments indicate that the oldest astronomical systems have comparable ages and that the universe as a whole began to expand from a highly compressed initial state in the finite past.

The Age of the Solar System. Unstable atomic nuclei (radioactive elements) decay exponentially. From comparisons between the measured abundances in rocks of a radioactive element with a measured half-life, its decay product, and a stable element chemically similar to the decay product, planetary scientists have been able to estimate the ages of the Earth's solid crust, meteorites, and the Moon. These estimates indicate that the Moon and the meteorites solidified 4.6×10^9 years ago and that the Earth's solid crust formed about 0.8×10^9 years later.

Ages of the Elements (for a review, see Cowan *et al.*, 1991). Theoretical studies of nucleosynthesis in supernovae predict relative abundances of stable heavy elements that agree reasonably well with measured abundances in meteorites. These studies provide estimates of the initial abundances of radioactive elements. A comparison between the estimated initial abundance of a radioactive element with a known half-life and its measured present abundance, relative to the initial and present abundances of a chemically similar stable element, yields an estimate of the time elapsed since the radioactive element was synthesized. Such estimates yield "ages" in excess of 7×10^9 years. An "age" estimated in this way is a weighted average of the ages of the supernovae that produced the heavy elements present in meteorites and other solid objects in the solar system.

Ages of Star Clusters (for a review, see Cowan *et al.*, 1991). The brightness and color (or effective temperature) of a star depend on its initial mass, its initial chemical composition, and its age. Theoretical calculations show how the brightness and color of a star with given initial mass and chemical composition vary with age. In a plot of brightness versus color, a star's representative point traces out a certain evolutionary track (Fig. 7). Stars of different mass trace out different tracks. If all stars in a cluster had the same age and chemical composition and hence differed only in mass, their representative points in a brightness–color diagram would lie on a curve whose shape would change as the cluster evolved; see Fig. 7. The measured brightness–color diagrams of clusters are in fact consistent with the hypothesis that stars in a cluster form at the same time and have the same initial chemical composition. In principle, then, the ages of clusters can be deduced from their brightness–color diagrams, interpreted in the light of theories and calculations of stellar evolution.

The ages of Galactic clusters (clusters that lie close to the Milky Way's central plane) range from a few million years to $(7 \text{ or } 8) \times 10^9$ years. The ages of globular star clusters are more difficult to estimate for a variety of technical reasons. Current estimates of the ages of the *oldest* globular clusters range from 14 to 18 thousand million years.

The Age of the Universe. The measured redshifts of distant galaxies imply that the system of galaxies is uniformly expanding. That is, the distance between any two sufficiently widely separated galaxies doubles in a fixed period of time. Relativistic cosmology (Sec. 4.1) allows us to calculate the time that has elapsed since the beginning of the cosmic expansion, given the present value of velocity–distance ratio (Hubble constant) and the present value of the cosmic mass density. The uncertainties in estimates of both values are still large. The resulting estimates of the time elapsed since the beginning of the cosmic expansion are between 10 and 20 thousand million (10^9) years.

4. MODERN COSMOLOGY

4.1 Relativistic Cosmology

In constructing his theory of gravitation, general relativity, Einstein set out to remedy two shortcomings of Newton's theory of gravitation. The first was its failure to explain the phenomenon, noted by Galileo and later confirmed by accurate experiments, that gravitational acceleration is independent of the physical properties of the accelerated object. In Einstein's theory the presence of mass distorts the structure of the four-dimensional space-time manifold. A gravitationally accelerated object moves along a geodesic (straightest line) in curved space-time. Since geodesics are determined by the structure of the manifold itself, Einstein's theory explains why grav-

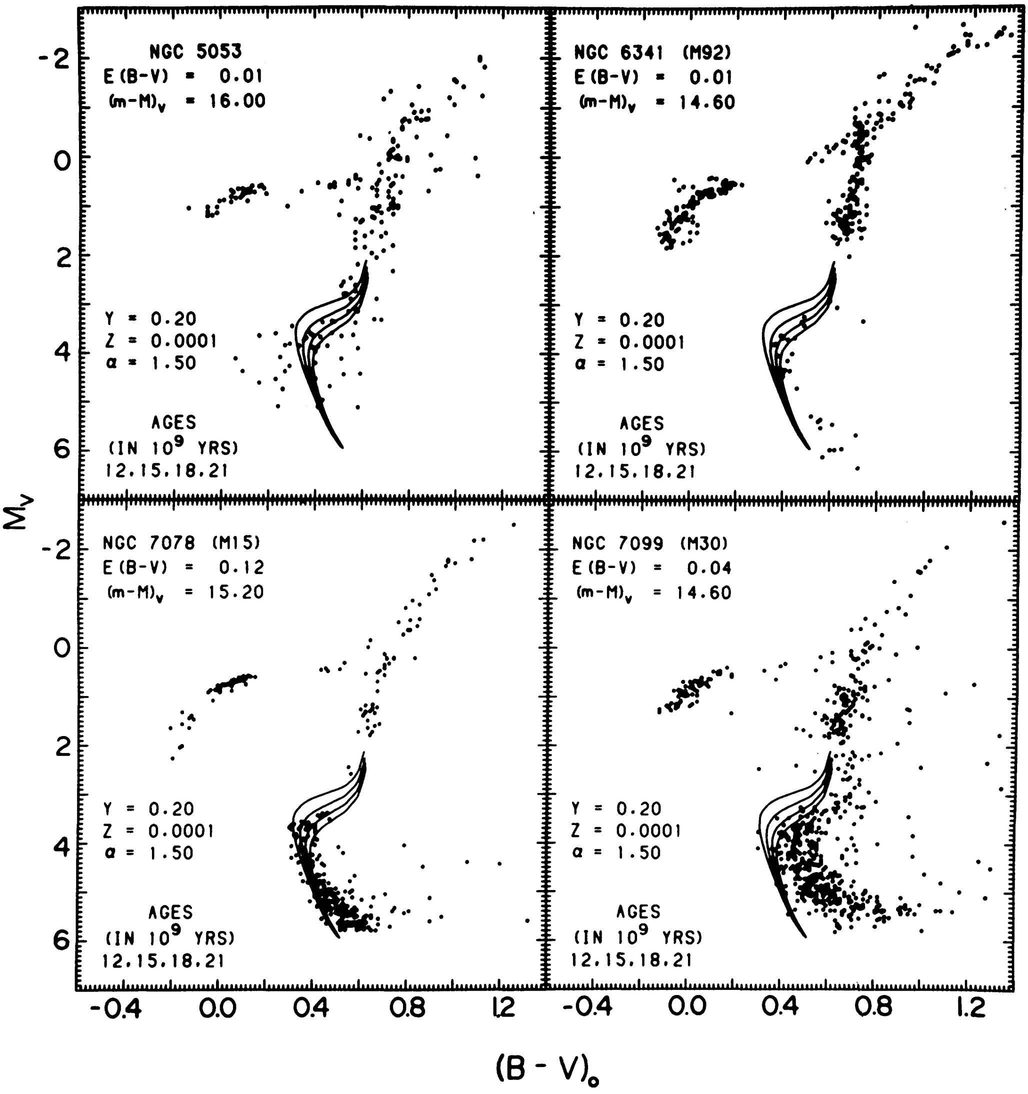

FIG. 7. Comparison on the observational plane of isochrones computed in this study with the published observations of NGC 5053, M92, M15, and M30. In each case, the cluster name and the adopted reddening and apparent distance modulus are specified in the top left-hand corner. The set of theoretical isochrones is characterized by the noted values of *Y*, *Z*, and *a*, and ages in the order of decreasing turnoff luminosity are given the lower left-hand corner of each plot. From Vandenberg (1976).

itational acceleration does not depend on any property of the accelerated object (and predicts that no such dependence will ever be found).

The second defect in Newton's theory was its failure to explain "Mach's principle": the coincidence between the local inertial reference frame (in which Newton's laws of motion are valid) and the cosmological reference frame (defined by the large-scale distribution of matter and motion). As mentioned earlier, Newton himself defined absolute acceleration as acceleration relative to absolute space, and he asserted that the "fixed stars" are at rest relative to absolute space. Ernst Mach, at the end of the 19th century, argued that absolute space is a scientifically empty notion. He suggested that the distribution of matter and motion on a cosmic scale somehow determines the local inertial reference frame. Newton's theory, however, does not explain the connection. Indeed, Newton's theory cannot even be applied in a consistent way to a uniform, unbounded distribution of mass filling all space.

In 1916 Einstein applied his recently published theory of gravitation (see GRAVITATION AND GENERAL RELATIVITY) to a uniform, static cosmic distribution of mass. He found that such a distribution of mass was inconsistent with his law of gravitation. At this point he made what he later described as the greatest scientific mistake of his career: instead of modifying his cosmological assumptions he modified the law of gravitation, adding a term whose coefficient is the so-called cosmological constant. The resulting equations do have a static solution, provided space is finite and has positive curvature. (A three-dimensional space of constant positive curvature is the three-dimensional "surface" of a sphere in four-dimensional Euclidean space; but its properties can also be specified without reference to a fourth spatial dimension, just as the properties of the two-dimensional surface of a sphere in three dimensions can be specified without reference to a third dimension.)

Alexander Friedman (1922), a dynamical meteorologist, discovered that Einstein's law of gravitation in its original form predicts that a uniform cosmic medium expands uniformly from a singular state of infinite density in the finite past (or contracts toward such a state in the finite future). Hubble's relation between distance and redshift, announced in 1929, confirmed this prediction. Because it is a consequence of Einstein's theory of gravitation and the cosmological principle, both of which enjoy strong observational support and neither of which has been seriously challenged, Friedman's theory is now accepted by nearly all contemporary cosmologists.

In Friedman's theory, space itself—not just the material medium filling space—is expanding (or contracting). This feature of the theory has a variety of observable consequences. For example, the travel time for light from a distant source is greater in an expanding space than it would be in a static space because the distance between the source and the observer is continually increasing. Because of the expansion of space, there are, at any given moment, objects so distant that light emitted by them at that moment never reaches the observer. Thus the theoretically observable universe is finite. Its (spherical) surface is called the event horizon. In a Friedman universe the mass within the event horizon increases with time; the radius of the event horizon increases at a greater rate than the radius of a sphere that contains a fixed quantity of mass.

The cosmological principle allows space to be curved but requires the curvature to be the same at every point. The curvature may be positive, negative, or zero. In space of zero curvature, Euclidean geometry prevails. For example, the angle sum of every triangle is exactly equal to 180° and the ratio between the circumference of a circle and its diameter is exactly equal to π. If space has negative curvature, the angle sum of every triangle is less than 180° and the ratio between the circumference of a circle and its diameter is greater than π. If space has positive curvature, the angle sum of every triangle is greater than 180° and the ratio between the circumference of a circle and its diameter is less than π. Friedman's theory specifies the temporal variation of the curvature, and so the curvature of a Friedman universe is characterized by the value of a single parameter.

Whether the curvature is positive, negative, or zero, the cosmic medium expands from a singular state of infinite density in the finite past. This "cosmological singularity" marks the beginning of time. If the curvature is negative or zero, the expansion continues forever and the mass density decreases steadily toward zero. In both cases, the medium is infinite. If the curvature is positive, space has finite volume (just as the surface of a sphere in three dimensions has finite area). In this case the universe expands to a state of minimum but finite mass density, then contracts toward a final cosmological singularity.

Friedman's theory predicts a simple relation between the mean cosmic mass density, the velocity–distance ratio (the Hubble constant), and the curvature of space. In a flat (zero-curvature) Friedman universe the cosmic mass density is a fixed multiple of the square of the Hubble constant. Depending on whether the cosmic mass density is less than or greater than this value, the spatial curvature is negative or positive.

Observational estimates of the cosmic mass density rest mainly on measurements of the velocities of galaxies, in groups, clusters, and larger aggregates. These estimates strongly suggest that the cosmic mass density is somewhere between 10% and 30% of the value that corresponds to zero spatial curvature. Inflationary cosmology (Sec. 4.6) predicts that the cosmic mass density is equal to the value that corresponds to zero spatial curvature. There is no direct observational support for such a high value.

Given the Hubble constant and the spatial curvature, Friedman's theory predicts the age of the universe—the time elapsed since the initial singularity. With a Hubble constant of 50 (km/s)/Mpc, the predicted age of the universe is about 17 thousand million years if the cosmic density is one-fifth the critical value. With the same Hubble constant, the predicted age is only 13 thousand million years if the cosmic density is equal to its critical value, as inflationary cosmology (Sec. 4.6) predicts. The second estimate is *smaller* than current estimates of the ages of the oldest globular star clusters. Doubling the Hubble constant halves the estimated ages.

4.2 The Primordial Fireball

The background radiation discovered by Penzias and Wilson has all the characteristics of thermal (blackbody) radiation at a temperature of 2.7 K. As mentioned above, observations from the COBE satellite have shown that its spectrum closely resembles the Planck spectrum. This observation implies that the background radiation must have been produced (or have already been in existence) at a time when the universe was opaque in the relevant range of wavelengths. This in turn requires the mass density of the universe to have been much greater in the past than it is now; no other way of accounting for the accurate blackbody spectrum of the background radiation is known. In the opinion of most cosmologists, these considerations rule out alternatives to relativistic cosmology that postulate a nonexpanding universe or a universe whose average mass density remains constant in time.

Theory predicts that after the expanding universe became transparent, the radiation, once thermalized, would have kept its thermal character. Its temperature, however, would have decreased at a rate proportional to the cube root of the number density of nucleons (protons and neutrons) or of photons. Since the average energy per photon in blackbody radiation is proportional to the temperature, the ratio between the energy density of the radiation and the energy density of matter, including rest energy, decreases continually as the universe expands. Extrapolating backward in time to the very early universe, one finds that nearly all the energy was in the form of thermal radiation at exceedingly high temperature. This is the primordial fireball. The fireball hypothesis accounts simply and accurately for the blackbody spectrum and the nearly perfect isotropy of the cosmic microwave background.

4.3 Origin of Helium and Other Light Elements

At very early epochs, the temperature of the universe (according to the fireball hypothesis) was so high that atomic nuclei heavier than the proton could not exist (Peebles, 1971; Weinberg, 1972). As the universe expanded, it cooled. At a temperature around 10^9 K deuterium began to form. Nearly all the deuterium subsequently burned to helium. As the temperature and density fell still further, nucleosynthetic reactions ceased and the relative abundances of the various nuclei were frozen in. These are the so-called primordial abundances, characteristic of the gas from which the first self-gravitating systems later formed, according to the fireball hypothesis.

The abundances of helium, deuterium, and other light elements produced in the fireball have been calculated with great accuracy. They depend on a single cosmological parameter, the cosmic mass density of ordinary matter (that is, matter composed of neutrons, protons, and electrons). If the fireball hypothesis is correct, it should be possible to choose this parameter to satisfy observational estimates of the cosmic mass density, the primordial abundance of helium, and the primordial abundance of deuterium and other light elements.

The most widely accepted current estimate of the primordial helium abundance is 23% by mass. The corresponding predicted value of the contribution of ordinary matter to the cosmic mass density is close to current estimates of the average mass density of *luminous* matter. Since, as discussed in Sec. 3.4, luminous matter accounts for at most 5% to 10% of the mass bound in self-gravitating systems (as deduced from their internal motions), the fireball hypothesis requires that at least 90% to 95% of the mass in the universe—essentially all the mass we cannot actually see—resides in "nonordinary" matter. What form does this "nonordinary" matter take?

During the 1970s many cosmologists believed that massive neutrinos, whose existence is predicted by grand unified theories of the strong, weak, and electromagnetic interactions (Sec. 4.5), could account for the nonluminous mass, but this possibility fell from favor in the 1980s because it failed to account for the distribution of nonluminous mass over the observed wide range of linear scales. "Cold dark matter," consisting of relatively massive particles of an unspecified kind, subsequently became the favored candidate, but this hypothesis fails to account for the spatial distribution of nonluminous mass on the largest scales (Saunders *et al.*, 1991).

As for deuterium, its predicted abundance is about 10 times greater than estimates based on observations of deuterium in interstellar space. The interstellar medium, however, has been formed in part from gas ejected by stars, where deuterium is burned into helium. Calculations that take this into account reduce but have not yet succeeded in eliminating the discrepancy between the observational evidence and the theoretical prediction.

The predicted abundance of lithium agrees well with astronomical estimates, but recent work (Gilmore *et al.*, 1991) suggests that beryllium is far more abundant than it should be according to the fireball hypothesis.

4.4 The Primordial Fireball and Particle Physics

The present-day universe is thought to contain no appreciable quantity of antimatter. But it is a consequence of the fireball hypothesis that proton–antiproton pairs and neutron–antineutron pairs would have been present in great numbers at very early times, when the cosmic temperature (measured in energy units) exceeded the rest energy of a proton–antiproton pair. At very early times the number densities of baryons would have been very nearly equal to the number densities of the corresponding antibaryons. If the number densities of baryons and antibaryons had been exactly equal, all the baryon–antibaryon pairs would have annihilated as the universe expanded and cooled. Only radiation would have remained, and the present ratio between the number densities of photons and baryons would have been infinite. Its actual value, 10^9 give or take a factor of 10, is indeed very large.

In 1967 André Sakharov outlined a scheme by which a tiny imbalance between baryon and antibaryon number densities might have evolved from a symmetric initial state with equal numbers of particles and antiparticles. Sakharov suggested that unstable superheavy particles produced in the very early universe might have decayed in ways that produced slightly more baryons than antibaryons as the universe expanded and cooled. This hypothesis required that the laws governing the decay of unstable particles discriminate between (a) particles and antiparticles, (b) left and right, and (c) the two directions of time. The first two "broken symmetries" were established experimentally in 1957 and the third in 1964. Sakharov's hypothesis also presupposed a theory of strong, weak, and electromagnetic interactions that would allow decay processes in which the net baryon number (the difference between the number of baryons and the number of antibaryons) changes. This is a feature of grand unified theories, developed in the 1960s and 1970s.

Grand unified theories allow neutrinos to have a finite rest mass. As mentioned above, massive neutrinos were once (but are no longer) a favorite candidate for the nonluminous mass of the universe.

Grand unified theories give protons a finite lifetime, because they allow particle-decay processes that violate baryon conservation. The simplest grand unified theory, however, predicts a proton lifetime that is shorter than the present experimental lower limit of about 10^{31} years.

4.5 Phase Transitions, Domain Walls, and Magnetic Monopoles

The now-standard unified theory of electromagnetic and weak interactions, developed by Sheldon Glashow, Steven Weinberg, and Abdus Salam, posits a uniform field of massive spinless particles called Higgs bosons. This field, which is part of the physical vacuum, is in its lowest energy. But in this state, the field itself has a nonzero value. This fact is ultimately responsible for the huge differences in strength and range between the weak and electromagnetic interactions.

In 1972, D. A. Kizhnits and A. D. Linde pointed out that at temperatures greater than a certain critical value (about 3×10^{15} K) the Higgs field vanishes in its state of lowest energy. At these temperatures, symmetry between the weak and electromagnetic interactions is "restored." It follows that when the temperature of the early hot universe fell below the critical value, the physical vacuum underwent a phase transition, somewhat analogous to the phase transition that occurs in a piece of iron when its temperature drops below the Curie point. At temperatures higher than the Curie point, the iron is unmagnetized; its physical state defines no preferred direction. At temperatures below the Curie point, spontaneous magnetization occurs, defining a preferred direction (or directions) and thus breaking the directional symmetry that prevailed at high temperatures. Similarly, when the cosmic temperature fell below its critical value, the physical vacuum underwent a phase transition, resulting in a breaking of the symmetry between the weak and electromagnetic interactions.

In a large piece of iron that has undergone spontaneous magnetization, the magnetic field may have different directions in contiguous "domains." Analogously, when the vacuum undergoes a phase transition, the Higgs field assumes different values in adjacent domains. The energy density in the walls separating adjacent domains turns out to be very high. Indeed, the contribution of domain walls to the energy (or mass) density of the universe has been estimated to be several powers of 10 higher than the upper limit set by cosmological observations.

An analogous phase transition is associated with the physical vacuum posited in grand unified theories. (The critical temperature for this transition is around 10^{27} K). Besides domain walls, this phase transition also produces, in great profusion, extremely energetic structures called magnetic monopoles (or grand monopoles). The estimated contribution of these structures to the cosmic energy density is many powers of 10 higher than the limit allowed by observation.

So the promising line of development initiated by the union of particle physics and the fireball hypothesis encountered seemingly insuperable difficulties.

4.6 Inflationary Cosmology

A way out of these difficulties was found in 1981 by Alan Guth. Guth proposed the following "scenario" for the very early universe. An initially hot universe, whose temperature exceeds the critical grand-unification temperature (10^{27} K) expands and cools. When the temperature falls below the critical temperature, the vacuum remains in a metastable supercooled state; it does not immediately make the symmetry-breaking phase transition discussed above. During this stage of supercooled expansion, the energy density of the vacuum comes to dominate the energy density of radiation and matter. Friedman's theory predicts that the universe now expands at an exponential rate (because the expansion does not reduce the energy density of the vacuum). The exponential expansion ("inflation") enormously reduces both the average spatial curvature and the amplitude of any local irregularities in mass density that may have been present. It generates a universe whose average mass density is indistinguishable from the critical value mentioned at the end of Sec. 4.1.

The phase transition to the stable vacuum state corresponding to the now greatly reduced cosmic temperature occurs through the formation of bubbles of stable vacuum. Collisions between the walls of these bubbles release the energy stored in the initial, metastable vacuum state, causing the universe to reheat. These collisions also create magnetic monopoles, but, because of inflation, the space density of these monopoles is now acceptably low.

Unfortunately, the mechanism for "reheating" in this scenario—collisions between the walls of the stable-vacuum bubbles—creates an observationally unacceptable level of nonuniformity and anisotropy in the cosmic distribution of matter and energy (including the cosmic microwave background).

An alternative theory, called the "new" inflationary scenario, was put forward in 1982 by Andrei Linde and, independently, by A. Albrecht and P. J. Steinhardt. It envisions a gradual phase transition and a more uniform mode of release of the energy stored in the initial vacuum state. This scenario, however, can be made to work only with an implausible degree of fine-tuning of properties of the Higgs field; even then it encounters severe technical difficulties.

To overcome these difficulties, Linde (1990) has proposed a still more speculative scenario, "chaotic inflation," in which quantum fluctuations in the energy of the Higgs field play a key role. Such fluctuations become important (as we look backward in time) at the so-called Planck epoch, about 10^{-43} s after the nominal beginning of the cosmic expansion, when quantum fluctuations cause Einstein's theory of gravitation to break down. In the "chaotic inflation" scenario, these fluctuations in the energy of the Higgs field give rise to an infinite collection of distinct "miniuniverses" with a wide range of initial conditions and hence of fundamental properties, including perhaps the dimensionality of space-time and the values of basic physical constants. We happen to inhabit a miniuniverse whose initial conditions gave rise to a set of physical constants and a scenario of cosmic evolution hospitable to the emergence of life. Some writers offer the last statement as an answer to the question of why certain constants and parameters that characterize the universe (such as the ratio between the number of photons and the number of nucleons) have the values they do. This answer is known as the *anthropic principle.*

For a review of inflationary cosmology, see Linde (1990).

4.7 Origin and Growth of Density Fluctuations in the Fireball

The new inflationary scenario predicts that large-scale fluctuations in the mass density are generated during the inflationary phase. These fluctuations obey the cube-root-of-N law: The fluctuation in the number of particles in a region containing N particles on average is approximately proportional to the cube root of N. Eventually these fluctuations are supposed to evolve into the precursors of galaxies, galaxy clusters, and the large-scale sheet-and-bubble structure. But two difficulties have arisen with this theory. First, numerical simulations using either hot dark matter (massive neutrinos) or cold dark matter have so far failed to replicate the observed structure. Second, a coefficient in the theory that determines the amplitude of the fluctuations must be arbitrarily assigned a value 10^{12} times smaller than the value predicted by particle physics.

The theory has, however, had one spectacular success. It predicts that large-scale fluctuations in mass density are accompanied by fluctuations or ripples in the temperature of the cosmic microwave background (see Peebles, 1980). If the amplitude of the primordial density fluctuations were just large enough for the fluctuations to evolve into the large-scale structure we now observe, the fractional amplitude of the temperature ripples would be about 10^{-5}. The COBE measurements referred to in Sec. 3.7 support this prediction (Wright *et al.*, 1992). Moreover, the measured spectrum of the ripples, on angular scales greater than 7°, has the form predicted for density fluctuations that obey a cube-root-of-N law.

4.8 The Vacuum–Energy Problem

Quite apart from difficulties with the various inflationary scenarios, the enterprise of uniting particle physics with the fireball hypothesis encounters a formidable difficulty (Linde, 1990). If present ideas about particle physics are correct in their main outlines, the fireball hypothesis implies that the vacuum has undergone at least three phase transitions: the first broke the grand-unified symmetry of strong and electroweak interactions; the second broke the symmetry of the electromagnetic and weak interaction; in the third, the quark sea condensed into baryons and mesons. Each of these phase transitions caused the mass density of the vacuum to decrease. In units of grams per cubic centimeter, the estimated changes are respectively 10^{80}, 10^{25}, 10^{14}. The present mass density of the vacuum is at most about 10^{-29} g cm^{-3}; it could well be precisely 0 (apart from quantum fluctuations). Despite many strenuous and ingenious attempts, no convincing explanation has yet been offered for the present low value of the cosmic mass density, given the magnitude of its past changes.

4.9 A Cold Universe?

It was suggested by Layzer in 1952 that the hierarchy of self-gravitating systems, from planetary systems to galaxy clusters, might have been formed by a single repetitive process: gravitational clustering in an expanding universe. At a given stage in the process, clusters S_n are the "molecules" of an expanding cosmic gas. As the gas expands, it becomes less uniform. The clusters S_n congregate in clusters S_{n+1} that become the new "molecules." Because gravitation, the "intermolecular" interaction responsible for clustering, has no preferred scale, the process occurs repeatedly on successively larger scales. Thus small systems are formed earlier than large systems—protostars before protoclusters, protoclusters before protogalaxies, and so on.

In 1962, Ya. B. Zel'dovich argued that an initially cold universe composed mainly of hydrogen would solidify when its density fell to about one-tenth the density of water, and that the solid expanding medium would subsequently break up into large chunks. Zel'dovich abandoned this hypothesis in favor of the fireball hypothesis after the discovery of the cosmic microwave background in 1965, but it was incorporated by Layzer into the theory of gravitational clustering (Layzer, 1984). There does not yet exist a rigorous statistical theory of an expanding "gas" of gravitationally interacting "molecules," but in its present heuristic form the theory of gravitational clustering accounts semiquantitatively for the major features of the clustering hierarchy. For example, it predicts that the average binding energy per unit mass of clusters of mass M is proportional to $M^{1/3}$. Thus binary stars, rich globular clusters, and galaxies, whose masses are roughly in the ratios $1{:}10^6{:}10^{12}$, are predicted to have binding energies per unit mass in the ratios $1{:}10^2{:}10^4$, in satisfactory agreement with observation. The same theory also predicts the approximate value of the coefficient in this relation.

Gravitational clustering does not account, however, for cosmic structure on the very large scales, the sheet–bubble structure described in Sec. 3.3. It has recently been suggested (Layzer, 1992a) that density fluctuations present at the Planck epoch, the initial state of the cold universe, evolve into structures qualitatively similar to the observed sheet–bubble structure. As in the "new" inflationary theory, but for a different reason, the primordial large-scale density fluctuations are predicted to have the cube-root-of-N form mentioned in Sec. 4.8.

To account for the cosmic microwave background, Layzer and Ray Hively postulated in 1973 that 90% to 95% of the stars that formed early in the cosmic expansion had masses of order 10 to 15 solar masses. The light emitted by these stars during their 10-million-year lifetimes would have been transformed into approximately blackbody radiation by its interaction with solid grains formed from heavy elements synthesized in the massive stars and ejected during the explosive phase of their evolution. The observed energy density of the microwave background and the fact that its spectrum very accurately approximates that of a blackbody impose stringent constraints on this hypothesis. Wright (1992) has shown that these conditions can easily be met if a reasonable fraction of the iron manufactured in the massive stars condenses into "whiskers," elongated structures that thermalize starlight very effectively.

Because the stars that produce the cosmic microwave background form in a medium of fluctuating large-scale density, the resulting background radiation bears the imprint of these fluctuations. The amplitude and the spectrum of the predicted temperature ripples agree at large scales with the COBE measurements (Layzer, 1992b); as mentioned in Sec. 4.7, so do the amplitude and spectrum of the temperature ripples predicted by the "new" inflationary scenario. The two theories differ, however, in their predictions about the temperature-fluctuation spectrum at angular scales between a few degrees and a few arc minutes. As the sensitivity of observations in this region improves, they will refute one (or both) of the theories.

The cold scenario predicts that the nonluminous mass resides mainly in the collapsed cores and ejecta of supernovae. If 90% to 95% of the mass of a primordial galaxy resided in relatively short-lived massive stars, the remnants of these stars would now populate a spheroidal system comparable in extent with the primordial galaxy, in agreement with observation.

4.10 The Cold Universe and Particle Physics

The implications of the cold scenario for particle physics differ from those of hot scenarios in two major ways. First, the vacuum-energy problem does not arise in the cold model because a vacuum phase transition does not occur in a universe initially at zero temperature. Second, the cold scenario requires the existence of several levels of quark and lepton substructure between the present experimental limit, which corresponds to energies of a few hundred giga-electron-volts, and the Planck energy of 10^{19} GeV. This requirement is needed to give the coefficient in the cube-root-of-N law the "right" value—the value for which predictions about large-scale structure agree with observation.

List of Works Cited

Baade, W. (1956), *Publ. Astron. Soc. Pacific* **68**, 5–16.

Baade, W. (1963), *Evolution of Stars and Galaxies*, Cambridge, MA: Harvard Univ. Press.

Cowan, J. J., Thielemann, F.-K, Truran, J. W. (1991), *Annu. Rev. Astron. Astrophys.* **29**, 447–497.

Friedman, A. (1922), *Z. Phys.* **10**, 377–386.

Gilmore, G., Edvardsson, B., and Nissen, P. E. (1991), *Astrophys. J.* **378**, 17–21.

Hertzsprung, E. (1914), *Astron. Nachr.* **196**, 203.

Hubble, E. P. (1936), *The Realm of the Nebulae*, New Haven: Yale Univ. Press.

Huchra, J. (1992), *Science* **256**, 321–325.

Jerjen, H., Tammann, G. A. (1992), to be published.

Kant, I. (1755), *Allgemeine Naturgeschichte und Theorie des Himmels*, Königsberg and Leipzig, translated by W. Hastie, in *Kant's Cosmogony* (Glasgow, 1900).

Layzer, D. (1984), *Constructing the Universe*, New York: Freeman (Scientific American Library).

Layzer, D. (1992), *Astrophys. J.* **392**, L5–L8.

Layzer, D. (1993), *J. Chem. Phys.*, to be published.

Leavitt, H. S. (1912), *Harvard College Observatory Circular* No. 173.

Lindblad, B. (1925), *Ark. Mat. Astron. Fys.* **19A**, Nos. 21, 27, 35; **19B**, No. 7.

Linde, A. (1990), *Particle Physics and Inflationary Cosmology*, Chur, Switzerland: Harwood.

Mather, J. C., Cheng, E. S., Eplee, R. E., Jr., Isaacman, R. B., Meyer, S. S., Shafer, R. A., Weiss, R., Wright, E. L., Bennett, C. L., Boggess, N. W., Dwek, E., Gulkis, S., Hauser, M. G., Janssen, M., Kelsall, T., Lubin, P. M., Moseley, S. H., Jr., Murdock, T. L., Silverberg, R. F., Smoot, G. F., Wilkinson, D. T. (1990) *Astrophys. J.* **354**, L37–L40.

Oort, J. (1927), *Bull. Astron. Inst. Netherlands* **3**, 275–282.

Peebles, P. J. E. (1971), *Physical Cosmology*, Princeton, NJ: Princeton Univ. Press.

Peebles, P. J. E. (1980), *The Large-Scale Structure of the Universe*, Princeton, NJ: Princeton Univ. Press.

Penzias, A. A., Wilson, R. W. (1965), *Astrophys. J.* **142**, 419–421.

Rood, H. J. (1988), *Annu. Rev. Astron. Astrophys.* **26**, 245–294.

Rubin, V. C. (1979), *Comments Astrophys.* **8**, 79–87.

Sandage, A. (1975a), in: A. Sandage, M. Sandage, J. Kristian (Eds.), *Galaxies and the Universe*, Chicago: Univ. of Chicago Press, Chap. 1.

Sandage, A. (1975b), in: A. Sandage, M. Sandage, J. Kristian (Eds.), *Galaxies and the Universe*, Chicago: Univ. of Chicago Press, Chap. 19.

Saunders, W. *et al.* (1991), *Nature* **349**, 32–38.

Shapley, H. (1918), *Astrophys. J.* **48**, 89.

Shapley, H. (1961), *Galaxies*, revised ed., Cambridge, MA: Harvard Univ. Press.

Smoot, G., Bennett, C. L., Kogut, A., Wright, E. L., Aymon, J., Boggess, N. W., Cheng, E. S., De Amici, G., Gulkis, S., Hauser, M. G., Hinshaw, G., Jackson, P. D., Janssen, M., Kaita, E., Kelsall, T., Keegstra, B., Lineweaver, G., Lowenstein, K., Lubin, P. M., Mather, J., Meyer, S. S., Moseley, S. H., Murdock, T., Rokke, L., Silverberg, R. F., Tenorio, L., Weiss, R., Wilkinson, D. T. (1992), *Astrophys. J. Lett.* **396**, L13–L18.

Stebbins, J. (1933), *Proc. Nat. Acad. Sci.* **19**, 222.

Tammann, G. A. (1992), in *Extragalactic Astronomy including Observational Cosmology* (to appear in *Phys. Scr.*).

Trumper, R. J. (1925), *Proc. Astron. Soc. Pacific* **37**, 307–318.

Tully, R. B., Fisher, J. R. (1977), *Astron. Astrophys.* **54**, 661–678.

Weinberg, S. (1972), *Gravitation and Cosmology*, New York: Wiley.

Wright, E. L., Meyer, S. S., Bennett, C. L., Bogess, N. W., Cheng, E. S., Hauser, M. G., Kogut, A., Lineweaver, C., Mather, J. C., Smoot, G. F., Weiss, R., Gulkis, S., Hinshaw, G., Janssen, M., Kelsall, T., Lubin, P. M., Moseley, S. H., Jr., Murdock, T. L., Shafer, R. A., Silverberg, R. F., Wilkinson, D. T. (1992) *Astrophys. J. Lett.*, **396**, L13–L18.

GALLIUM ARSENIDE

DAVID C. LOOK, *Wright State University, Dayton, Ohio, U.S.A.*

Introduction 43
1. **Basic Properties** 45
1.1 Lattice Structure 45
1.2 Defects and Impurities 45
1.3 Lattice Motion 46
1.4 Electron States 47
2. **Crystal Growth** 48
2.1 Bulk Growth 48
2.1.1 Liquid-Encapsulated Czochralski Growth 48
2.1.2 Bridgman and Gradient-Freeze Growth 49
2.2 Epitaxial Growth 50
2.2.1 Molecular-Beam Epitaxy 50
2.2.2 Metal-Organic Vapor-Phase Epitaxy 51
2.2.3 Vapor-Phase and Liquid-Phase Epitaxy 51
3. **Optical Properties** 52
3.1 Light Absorption 52
3.1.1 Band-to-Band, Exciton Absorption 52
3.1.2 Band-to-Impurity or -Defect Absorption 53
3.1.3 Acceptor-to-Donor Absorption 53
3.1.4 Free-Carrier Absorption 53
3.1.5 Phonon Absorption 53
3.1.6 Local Vibrational-Mode Absorption 53
3.2 Light Emission: Photoluminescence 54
3.3 Modulated Reflectance: Photoreflectance 55
4. **Electrical Properties** 56
4.1 Hall-Effect Measurements 56
4.2 Electron and Hole Concentrations 57
4.3 Mobilities 57
5. **Applications** 59
5.1 Discrete Electronic Devices 60
5.1.1 Metal-Semiconductor Field-Effect Transistors 60
5.1.2 Heterojunction Field-Effect Transistors 61
5.1.3 Heterojunction Bipolar Transistors 62
5.2 Integrated Circuits (ICs) 62
5.2.1 Digital ICs 63
5.2.2 Microwave ICs (MMICs) 64
5.3 Optical Devices 65
5.3.1 Sources 65
5.3.2 Detectors 65
5.3.3 Integrated Optics 66
Glossary 66
Acronyms 67
Works Cited 67
Further Reading 68

INTRODUCTION

Gallium arsenide (GaAs) is the second most popular semiconductor material in use today [behind silicon (Si)] and is the material of choice for the highest-speed digital and microwave applications. The reasons are basically twofold:

1. The electrons move very fast in response to an applied electric field; and
2. the resistivity can be made either very low, permitting strong current conduction in devices, or very high, allowing "inert" platforms for integrated circuits.

These two properties have fostered a mature microwave industry, exploiting applications such as microwave communications and phased-array radar, and a budding digital industry, with a 16-Gflops supercomputer having all-GaAs logic due in the market in 1992 or 1993 and a smaller hybrid (GaAs/Si) machine already introduced in July 1991. However, it is more difficult to grow and process GaAs than Si, so that Si devices are cheaper

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

and thus completely control most of the present consumer markets (such as television). A developing market is optoelectronics, i.e., the joining of electronic and optical devices on a single chip. Here GaAs has a distinct advantage over Si because of inherent properties that make it a much more efficient emitter and absorber of light. Already, in fact, one of the chief uses of GaAs is as a substrate for light-emitting diodes (LEDs) and other discrete optical devices. But one can imagine more complicated applications in the future, such as transmitters and receivers with laser sources, detectors, amplifiers, modulators, demodulators, and logic elements all fabricated at the same time on one small piece of GaAs.

Because the commercial interest in GaAs depends upon its basic optical and electrical properties, it is necessary to consider these subjects in some detail. As with most useful semiconductors, these properties can be controlled by impurities and defects, which themselves are controlled by the growth conditions. For example, GaAs can be given a nearly metallic conductivity by heavily doping with Si, and a nearly insulating quality by doping with Cr (chromium) or by "doping" with a high concentration of a particular defect, As misplaced on a Ga site. To develop these concepts, we will first discuss the basic lattice properties and the electrical, optical, and thermal consequences of these properties. We will follow that discussion with a summary of some of the crystal-growth techniques, then a detailed development of the optical and electrical properties, and finally an enumeration of present and future applications. For the convenience of the reader, a summary of some of the more useful measured parameters associated with GaAs is given in Table 1. Much more detailed and critical summaries of these and other values can be found in the excellent review given by Blakemore (1982) and also in the volume *Properties of Gallium Arsenide*, involving a number of authors from the EMIS (Electronic Materials Information Service) group (1990).

Table 1. Properties of GaAs at 300 K.

Length of side of unit cube	5.65325 Å
Nearest-neighbor distance	2.44793 Å
Atomic density	4.4279×10^{22} cm^{-3}
Molecular weight	144.642 amu
Crystal density	5.3174 g/cm^3
Speed of sound	
Longitudinal wave, [100] direction	4.731×10^5 cm/s
Transverse wave, [100] direction	3.345×10^5 cm/s
Young's modulus	85.5×10^{10} dyn/cm^2
Poisson ratio ([100] direction)	0.31
Phonon energy (at the Γ point): LO phonon	36.1 meV
TO phonon	33.2 meV
Melting point	1513 K
Thermal expansion coefficient (linear)	5.73×10^{-6} K^{-1}
Debye temperature (thermal)	360 K
Specific heat (at constant pressure)	0.327 J/g·K
Dielectric constant: low frequency	12.85
high frequency	10.88
Refractive index (infrared)	3.299
Energy gap (direct)	1.423 eV
Effective mass (at Γ point): electron	$0.063m_0$
heavy hole	$0.50m_0$
light hole	$0.076m_0$
split-off hole	$0.145m_0$
Density of states (effective): conduction band	4.21×10^{17} cm^{-3}
valence band	9.51×10^{18} cm^{-3}
Carrier concentration (intrinsic)	2.25×10^6 cm^{-3}
Electrical conductivity (intrinsic)	3.0×10^{-9} Ω^{-1} cm^{-1}
Hall mobility (intrinsic): electrons	8500 cm^2/V·s
holes	450 cm^2/V·s

1. BASIC PROPERTIES

1.1 Lattice Structure

GaAs crystals do not appear in nature but can be grown by many different techniques, to be discussed later. The lattice is of the cubic sphalerite, or zinc-blende, type, which involves two face-centered cubic lattices, one consisting of Ga and the other of As, displaced from each other by one-quarter of a body diagonal. A unit cube is illustrated in Fig. 1. Note that each atom is surrounded by four atoms of the opposite kind sitting at the corners of a tetrahedron. Such an arrangement is very common for semiconductor materials. The size a of the unit cube, the nearest-neighbor distance, and several other relevant properties are listed in Table 1. An important feature of the GaAs lattice is relatively easy cleavage along the {110} family of planes. For ingots grown along the [100] direction, with wafers sliced perpendicular to that direction, the {110} cleavage planes are at right angles to each other, which facilitates cutting the wafer into smaller pieces. A weaker cleavage occurs along the {111} planes, but this property is normally of little consequence.

1.2 Defects and Impurities

Intrinsic (defect- and impurity-free) GaAs would have a room-temperature resistivity ρ of about 3×10^8 Ω cm. Such a material, by itself, would be of little use to the electronics industry. However, by doping with Si or other elements, ρ can be reduced to 10^{-3} Ω cm, which makes a very useful material indeed. Thus, impurities in semiconducting materials are extremely important because they allow modification and control of the electrical and optical properties. Defects are also important for the same reasons. Group II elements, such as Be, Mg, Zn, and Cd, generally substitute for a Ga atom, and act as "acceptors" (hole producers) because they bring one less electron than needed for bonding purposes and thus attempt to "accept" an electron from another source (e.g., the valence band or a donor impurity), leaving a hole in that source. Group VI elements, such as S, Se, and Te, generally substitute for As atoms, and act as "donors" since they bring an extra electron over that needed for bonding. Group IV elements, such as C, Si, Ge, and Sn, can go on either Ga or As sites, and thus can act as donors or acceptors, respectively. This behavior is termed "amphoteric." However, under the typical conditions being used for present-day GaAs crystal growth, C usually goes on the As site as an acceptor, and Si, Ge, and Sn on the Ga site as donors.

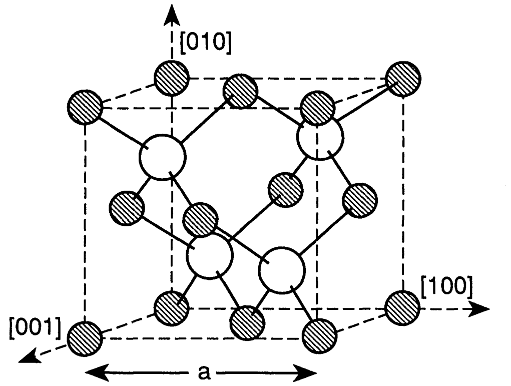

FIG. 1. Unit cube for GaAs, where a is the lattice constant. [From Blakemore (1982). Reproduced by permission of the American Institute of Physics.]

Typical background impurity concentrations range from 10^{14} cm^{-3} (2 parts per billion atomic, or 2 ppba) in epitaxial GaAs, to 10^{16} cm^{-3} (200 ppba) in bulk GaAs. The most commonly observed background elements are probably C as an acceptor and Si as a donor. The most common elements used for doping are C, Be, Zn, and Cd as acceptors, and Si, Sn, and Te as donors. Some of the principal considerations in doping are the solubility and the diffusion propensity under high-temperature processing conditions.

Defects can consist of missing atoms, called vacancies; atoms wedged in between other atoms, called interstitials; and atoms on the wrong sites, called antisites. By far the most important defect, and the only one identified so far in GaAs with any degree of certainty, is the arsenic antisite, i.e., As on a Ga site, or $\mathrm{As_{Ga}}$. It is said that this defect is the most studied of any in the history of semiconductors, and the reason for that is its importance in the production of semi-insulating GaAs, which is used as a substrate material for GaAs microwave devices as well as analog and digital integrated circuits (see Sec. 5). It is much more difficult to study defects than impurities, because defects are not responsive to any of the usual analytical techniques, such as mass spectroscopy. However, simple defects, such as vacancies, can be created by high-energy electron irradiation and thus fingerprinted with a fair degree of accuracy. For example, the arsenic vacancy V_{As} seems to be reasonably well identified by certain electrical characterization techniques. Optical and electrical signatures of several other defects have also been found, but identifications are uncertain. What is clear, however, is that defects can have either adverse or favorable effects on GaAs materials and devices, depending on the circumstances. The concentrations of electrically active defects typically range from 10^{13} cm^{-3} or less in epitaxial materials to 10^{16} cm^{-3} in bulk materials.

Besides the simple "point" defects described above, larger defects such as dislocations, precipitates, and voids can also exist. For example, bulk GaAs grown by the liquid-encapsulated Czochralski (LEC) method typically contains about 50 000 dislocations/cm^2 (counting only those which intersect the surface). Epitaxial GaAs has a much smaller dislocation density, but may have other imperfections. For example, molecular-beam epitaxial (MBE) layers can have from 10 to 1000 "oval" defects (small defects resulting from liquid Ga droplets) per cm^2 on the surface, depending on thickness and growth conditions. There is still a degree of controversy over how much these large defects may affect a device in close proximity to one of them. As an example, some studies have suggested that a metal-semiconductor field-effect transistor (MESFET) lying within about 30 μm of a dislocation will suffer a significant change in its threshold voltage. Such questions have great importance for integrated-circuit applications.

1.3 Lattice Motion

As with any solid, the atoms in GaAs can move and vibrate in directions and modes constrained by the lattice symmetry. The dynamics of the deformation and motion with respect to an external mechanical force are described by a tensor **c** with components c_{ij}, where this tensor is sometimes called the "elastic response tensor," and the components are variously called the "moduli of elasticity," the "elastic stiffness constants," or simply the "elastic constants." As an example, we would write the stress component X_x, which represents a force applied in the x direction (the capital letter) to a unit area of a plane whose normal also lies in the x direction (the subscript), as

$$X_x = c_{11}e_{xx} + c_{12}e_{yy} + c_{13}e_{zz} + c_{14}e_{yz} + c_{15}e_{zx} + c_{16}e_{xy}, \tag{1}$$

where **e** is the strain tensor. Similar equations could be written for Y_y, Z_z, Y_z, Z_x, and X_y. For the cubic lattice of GaAs several of the components vanish, and of the remaining ones only three, c_{11}, c_{12}, and c_{44}, are independent. These three components govern the speed of sound in any direction, since a sound wave results from the elastic response to an external force. For example, a wave propagating in a [100] direction with particle motion also in the same direction has a velocity $v_L = (c_{11}/\rho)^{1/2}$, where ρ is the crystal density. If the particle motion is in a (100) plane perpendicular to the direction of the motion, then $v_T = (c_{44}/\rho)^{1/2}$. Numerical values of v_L and v_T in this case are given in Table 1. For wave and particle motion in other directions, the equations are slightly more complicated.

Normally when we consider wave motion in a solid, we think of low-frequency sound waves, with frequencies $\nu \simeq 1\text{–}10^4$ Hz. In fact, however, in an equivalent one-dimensional GaAs lattice, the calculated ν can extend all the way up to approximately $[2ac_{11} \times (m_{As}^{-1} + m_{Ga}^{-1})/4\pi^2]^{1/2} \simeq 7.5 \times 10^{12}$ Hz, where a is the lattice constant. The density of states of the phonons (quantized lattice vibrations) as a function of frequency for GaAs is given in Fig. 2, and it is seen that there is a very strong peak near $\nu = 8 \times 10^{12}$ Hz, which is mainly due to optical phonons. For many experiments, the longitudinal-optical (LO) phonon at $\mathbf{k}=0$ in the Brillouin zone (Γ point), which has a frequency $\nu = 8.55 \times 10^{12}$ Hz (35.4 meV), is an important factor to be considered. For example, an electron or hole that attains a very high kinetic energy (becomes "hot"), by absorbing a light photon or being accelerated by an electric field, often gives off much of this excess energy by emitting optical phonons.

The lattice vibrations can also store energy. In the simplest ("classical") picture, each atom will have $\frac{1}{2}kT$ kinetic energy and $\frac{1}{2}kT$ of potential energy for each of its 3 degrees of freedom. Thus, for N atoms, the total energy U should be $3NkT$, and the specific heat $C = \partial U/\partial T = 3Nk$. Per unit mass we then would expect $C = 3k(2N_a/M) = 0.345$

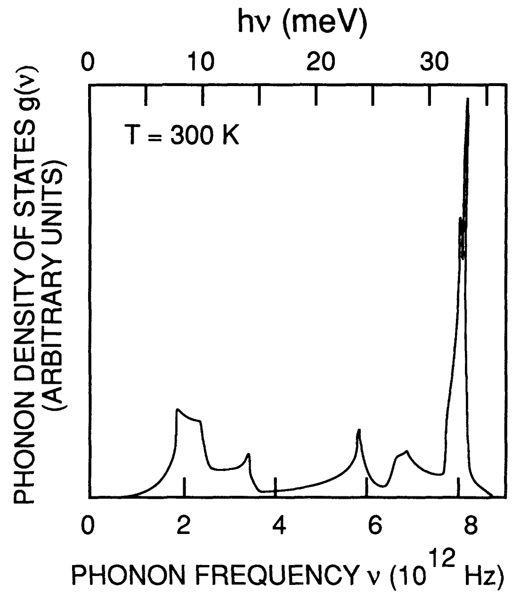

FIG. 2. Frequency dependence of the total spectral density of all lattice vibrational modes. [From Dolling and Cowley (1966). Reproduced by permission of the IOP Publishing Co.]

J/g·K, where N_a is Avogadro's number and M is the molecular weight. In fact, the specific heat at constant pressure C_p is about 0.327 J/g·K at 300 K and fairly well follows a Debye model, with a thermal Debye temperature $\Theta_D \simeq 360$ K, above about 100 K. It is interesting that this Θ_D corresponds to a frequency $k\Theta_D/h \simeq 8\times 10^{12}$ Hz, which is the most prominent frequency in the phonon density-of-states spectrum, Fig. 2.

If parts of a sample are at different temperatures, then heat will be transferred by the phonons from the hot areas to the cooler areas. The heat transferred across a unit area per unit time in a sample of length L is given by $Q=\kappa(T_1-T_2)/L$, where κ is the thermal conductivity and T_1 and T_2 are the temperatures at the respective ends of the sample. In classical theory, κ can be related to C by $\kappa=\rho Cvl/3$, where ρ is the density, v an average speed of sound, and l a phonon mean free path. Experimentally, $\kappa \simeq 0.55$ W/cm·K, in pure material at 300 K, which suggests a mean free path $l=200$–300 Å, depending on v. At low temperature, say 20 K, κ can approach 50 W/cm·K, and since C falls to about 0.02 J/g·K, l becomes very large, about 0.01 cm. Doping, even up to 10^{19} cm^{-3}, affects κ very little at 300 K, but begins to have a large effect below 50 K.

Anharmonic terms in the potential energy of the vibrating atoms are responsible for thermal expansion of the lattice with temperature. The linear expansion coefficient of GaAs is about 5.7×10^{-6} K^{-1} at 300 K and increases to about 7.2×10^{-6} K^{-1} at 1000 K.

1.4 Electron States

Lattice vibrations and the associated phonon modes are concerned with collective motions of the nuclear cores of the atoms, and the surrounding electrons are simply considered to move along with the cores. However, just as the collective motions of the nuclear cores (the lattice vibrations) can be described in terms of energy bands, so the interactions of the surrounding electrons with each other also lead to bands of energy separated by gaps, as shown in Fig. 3. The total density of states for all of the bands shown in Fig. 3 is plotted in Fig. 4. Here, the bands below $\mathscr{E}=0$ are called the valence bands, and are filled

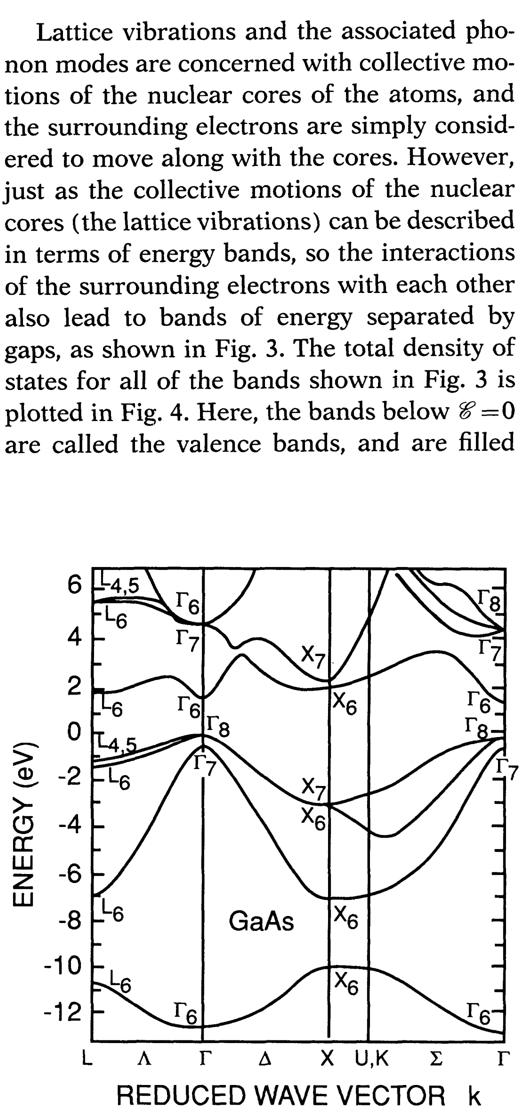

FIG. 3. Electron energy vs reduced wave vector for the four GaAs valence bands and the first four conduction bands. [From Chelikowsky and Cohen (1976). Reproduced by permission of the American Institute of Physics.]

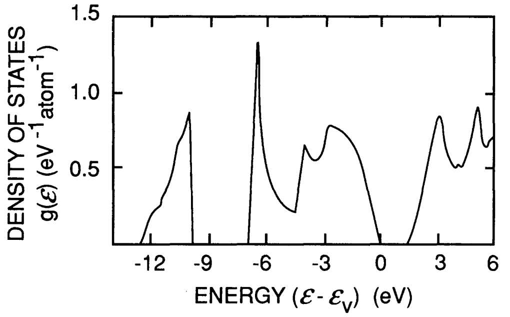

FIG. 4. Density of states for the lower part of the conduction band and for the four valence bands. [From Chelikowsky and Cohen (1976). Reproduced by permission of the American Institute of Physics.]

with electrons, while the bands above $\mathscr{E}=0$ are empty. (In this article, we will denote *general* energies by $\mathscr{E}$, in order to distinguish from electric field *E*, but *specific* energies by *E* with a subscript, which is a more common notation.) The lowest energy gap between filled and empty states occurs at the Γ point (k_x,k_y,k_z=0,0,0) in the Brillouin zone, and is "direct," i.e., the maxima of the Γ_8 valence band (Γ_{8V}) and the Γ_6 conduction band (Γ_{6C}) both occur at the same **k** value. The resultant band gap, 1.519 eV at $T=0$ K and 1.423 eV at $T=300$ K, is extremely important to the optical and electrical properties of GaAs, as will be shown later. The higher conduction bands, L_{6C} and X_{6C}, are of importance only at higher temperatures or for other situations in which the electrons are excited to high energies. Similarly, the Γ_7 valence band (the "split-off" band) normally has little influence on the optical and electrical properties.

The curvature in $\mathscr{E}$ vs **k** of the various bands is also very important because it influences both the density of states (Fig. 4) and the effective mass, which is given by $m^*=\hbar^2/(d^2\mathscr{E}/dk^2)$. Indeed, it is the strong curvature of the Γ_{6C} band which leads to a low effective mass ($m^*_{6C}=0.067m_0$ at $T=0$) and thus to high electron acceleration in an electric field. It is, in large part, the low effective electron mass, the relatively high band gap, and the direct nature of this gap that give GaAs many advantages over its chief rival, Si. However, there are also disadvantages, which will be discussed later.

2. CRYSTAL GROWTH

2.1 Bulk Growth

Since GaAs crystals are not found in nature, they must be grown by artificial means. The sizes of such crystals range from thin layers of a few angstroms (10^{-8} cm) thickness to large ingots more than 10 cm in length and 10 cm in diameter. We term the production of such ingots "bulk" growth and describe it in some detail below. [Also, it has been reviewed recently by AuCoin and Savage (1985).] The efficient, inexpensive growth of long, large-diameter GaAs ingots is crucial to the emerging GaAs integrated-circuit industry.

There are six major methods used to grow bulk GaAs:

1. horizontal Bridgman (HB);
2. vertical Bridgman (VB);
3. horizontal gradient-freeze (HGF);
4. vertical GF (VGF);
5. low-pressure liquid-encapsulated Czochralski (LP-LEC); and
6. high-pressure LEC (HP-LEC).

A large problem with GaAs crystal growth is that the As vapor pressure at elevated temperatures is much higher than that of Ga, so that As tends to evaporate from the growing crystal. The Bridgman and gradient-freeze techniques solve that problem by putting excess As into a closed tube containing the melt and growing crystal, whereas the LEC method encapsulates the melt with liquid B_2O_3, and further adds a high pressure of inert gas in order to hold in the As. It should be pointed out that this partial-pressure difference of the constituent elements is a problem not encountered in the growth of an elemental semiconductor such as Si. Thus, simpler techniques are available for Si growth, and, e.g., inexpensive 20-cm-diam Si wafers are readily available today.

2.1.1 Liquid-Encapsulated Czochralski Growth. The high-pressure LEC technique is presently the dominant method for producing semi-insulating (SI) GaAs, which is used as the basic substrate material for many GaAs devices. An advantage of the HP-LEC process is that polycrystalline GaAs can first be synthesized from elemental Ga and As in the same reactor that is later used for the single-crystal growth. The synthesis takes place at about 700 °C, with a liquid B_2O_3 layer and 60 atm of inert gas to keep the As from escaping. Then the temperature is increased to the GaAs melting point, the pressure reduced to 2–4 atm, and crystal growth begun by inserting a GaAs seed through the B_2O_3 layer and into the melt. The temperature is lowered until solidification begins to occur on the seed, and the seed is then pulled, along with the attached crystal, through the B_2O_3 encapsulant at a rate of 1–10 mm/h. A typical product of this process, an ingot somewhat larger than 76 mm in diameter and weighing about 4 kg, is shown in Fig. 5. This ingot would next be ground to obtain a nearly uniform 76-mm diameter and then sliced into wafers, which, after polishing, would have a thickness of about 650 μm.

As can be seen in Fig. 6, the only limit to crystal size in the LEC technique is the diameter of the crucible holding the melt. In fact, 125-mm-diam crystals have already been grown, and one manufacturer states that 125-mm, 25-kg crystals will be the norm in a few years. A disadvantage of the HP-LEC method is the size and cost of the reactor, which must withstand 60 atm pressure. The LP-LEC method helps somewhat in this regard, but has the disadvantage that As must be injected from an exterior source to maintain the proper stoichiometry.

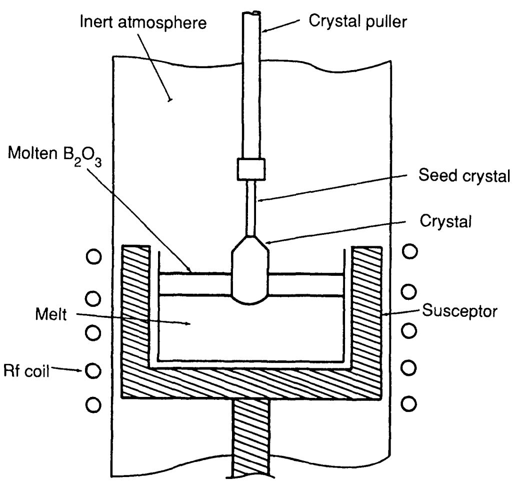

FIG. 6. Schematic of a liquid-encapsulated Czochralski crystal puller. [From AuCoin and Savage, (1985). Reproduced by permission of Howard W. Sams and Co.]

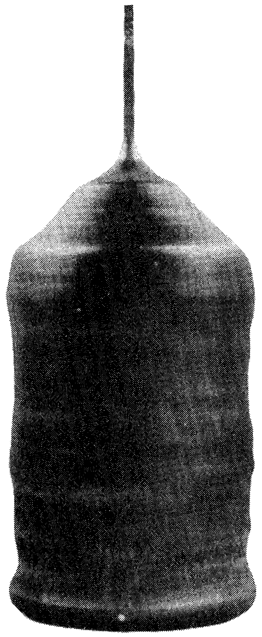

FIG. 5. A 4-kg undoped semi-insulating crystal of GaAs grown by the LEC method. (Courtesy D. E. Holmes.)

2.1.2 Bridgman and Gradient-Freeze Growth. Horizontal Bridgman and gradient-freeze reactors are much simpler and less expensive to set up than their LEC counterparts. As shown in Fig. 7, the seed and melt typically sit in a quartz boat, which itself is placed at one end of a sealed quartz ampoule with excess As at the other end to maintain stoichiometry. In the HGF variation, two (or more) furnaces produce a temperature profile (note that the right side is hotter, in Fig. 7), and the profile is slowly changed (cool zone moved to the right) to freeze the entire charge. In the HB variation, the furnace is moved, rather than the profile being electronically adjusted. In spite of their simplicity and the advantage of low dislocation density, these techniques have several disadvantages:

1. The crystals grow best in a ⟨111⟩ direction and so must be sliced at an angle to get the preferred (100) wafers;
2. because of that, and also because the tops of the crystals are flat to begin with, the wafers cannot be made round without wasting a significant amount of material; and
3. since the growth takes place in a quartz environment, Si is a common impurity and

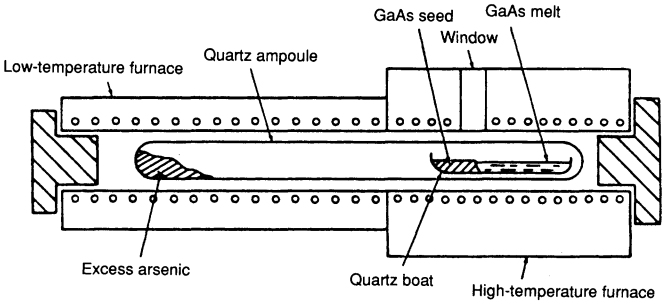
GaAs seed
GaAs melt
Window
Low-temperature furnace
Quartz ampoule
Excess arsenic
Quartz boat
High-temperature furnace

FIG. 7. Schematic of a gradient-freeze furnace. [From AuCoin and Savage, (1985). Reproduced by permission of Howard W. Sams and Co.]

makes the crystals *n*-type. To get semi-insulating material, Cr must be added, and that introduces a whole new set of problems for the subsequent device fabrication. Thus, HB and HGF wafers are usually used for applications such as solar cells, which do not need SI substrates.

All of the above problems can be circumvented by going to vertical growth in pyrolitic boron nitride (PBN) crucibles. The resulting crystals are round, have a ⟨100⟩ growth axis, are semi-insulating, and maintain the low dislocation density which is characteristic of the Bridgman and GF methods (because of low thermal gradients during growth). One obvious problem of the vertical growth techniques is that now the melt is touching all sides of the crucible, and it may wet or break the crucible (or crystal) because of expansion upon freezing. However, it appears that several American vendors have overcome these problems and are now marketing 76-mm-diam SI GaAs wafers produced by the VGF and VB methods. It will be interesting to see whether the VGF and VB methods will prove to be superior to the LEC technique for SI growth.

2.2 Epitaxial Growth

Although the starting material for GaAs devices is the substrate, typically 650 μm thick, the active (conductive) layer is usually much thinner, from 0.01 to 1 μm. Such thin conductive layers can be produced either by ion implantation or by epitaxial growth. Ion implantation is simple and cost-effective because the dopant (usually Si for *n*-type material, Be for *p*-type material) is implanted directly into the substrate. However, the dopant then has a nonflat profile (nearly Gaussian), and it may become further distorted during the high-temperature (~750–850 °C) anneal necessary to reduce lattice damage and move the ions onto substitutional lattice sites. Epitaxial growth of the active layer, on the other hand, can lead to very abrupt doping profiles, which are necessary for some devices.

2.2.1 Molecular-Beam Epitaxy. A schematic drawing of a molecular beam epitaxial (MBE) growth chamber is given in Fig. 8. Effusion cells, containing solid sources of Ga, As, and the various dopants, are heated until they produce the desired evaporation rate of

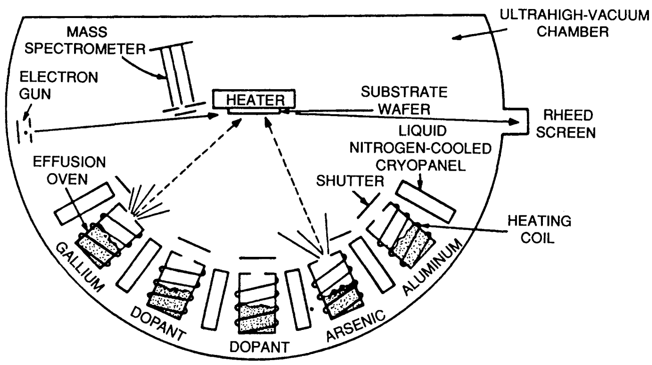
ULTRAHIGH-VACUUM CHAMBER
MASS SPECTROMETER
ELECTRON GUN
HEATER
SUBSTRATE WAFER
RHEED SCREEN
LIQUID NITROGEN-COOLED CRYOPANEL
EFFUSION OVEN
SHUTTER
HEATING COIL
GALLIUM
DOPANT
DOPANT
ARSENIC
ALUMINUM

FIG. 8. Schematic of an MBE growth chamber. (Courtesy of M. B. Panish.)

the atomic or molecular species. Because of the ultrahigh vacuum (10^{-11} Torr before growth, 10^{-6} Torr during growth), the various beams move almost without interaction before reaching the substrate surface. Gaseous sources, such as arsine, are also available and can be replenished much more easily. For characterization, built-in tools such as mass spectroscopy and reflection high-energy electron diffraction (RHEED) have proved very useful in setting the growth parameters precisely. Excellent single-crystal layers can be grown at substrate temperatures 200–600 °C, although the 200 °C material has an extremely high (and often very useful!) concentration (10^{20} cm^{-3}) of As antisites. Background impurity concentrations in the 10^{14}-cm^{-3} range can be routinely achieved, as can Si doping as high as 10^{19} cm^{-3}, and Be or C doping up to 10^{20} cm^{-3}. Coupled with the very precise thickness control (1 monolayer!), MBE has a versatility perhaps unmatched by any other technique. However, it also has several disadvantages:

1. high initial expense;
2. low growth rate (~ 1 μm/h);
3. long down time for source loading; and
4. difficulty in obtaining doping and thickness uniformity on large wafers.

Thus, MBE at present finds its primary application in growing very thin, complicated structures for discrete devices, although problems **3** and **4** above are being successfully addressed.

MBE growth and characterization have been recently reviewed by Sakaki (1989).

2.2.2 Metal-Organic Vapor-Phase Epitaxy. The major competing technology to MBE for thin layers is MOVPE (metal-organic vapor-phase epitaxy), also known as MOCVD (metal-organic chemical vapor deposition) and OMVPE (organometallic VPE). The apparatus is very simple, as illustrated in Fig. 9, and easy to set up, compared to an MBE system. In Fig. 9, the source gasses, typically trimethylgallium (TMGa) or triethylgallium (TEGa) and arsine (AsH_3), enter the chamber, are dispersed by a baffle, and then react over the substrate which is sitting on a susceptor held at 500–800 °C. The reaction products can be fairly complex and are not fully understood, but it is known that diffusion through a boundary gas layer at the susceptor surface is an important part of the process. This is in contrast to the MBE case in which the primary species essentially do not react until they reach the substrate surface and find each other through hopping processes. As compared with MBE, some of the advantages of MOVPE for GaAs growth are lower initial cost; higher growth rate (~ 10 μm/h); better thickness and doping uniformity on large wafers; and less down time for source replenishment. Some of the disadvantages are

1. the toxicity of arsine, which effectively raises the apparatus costs because of the necessity for safety devices; and
2. less precision in the control of very thin layers.

Attempts are being made to address problem **1** by finding new, safer gaseous sources of As, and problem **2** by working at low pressures, which allows the gas flow to be turned on and off quickly. In fact, by working at very low pressures, in a vacuum chamber, some of the advantages of MOVPE and MBE can be combined and a whole new discipline has developed, MOMBE (metal-organic MBE). It is likely that each of these three techniques, MBE, MOVPE, and MOMBE, will find particular applications in the future.

MOVPE growth of III-V compounds has been reviewed recently by Dapkus and Coleman (1989).

2.2.3 Vapor-Phase and Liquid-Phase Epitaxy. Two of the earliest and still viable techniques for growing GaAs layers are vapor-phase epitaxy (VPE) and liquid-phase epitaxy (LPE). The most common VPE system passes

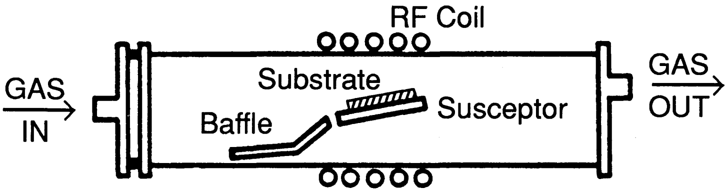

FIG. 9. Schematic of an MOVPE growth chamber. [From Coleman and Dapkus, (1985). Reproduced by permission of Howard W. Sams and Co.]

$AsCl_3$, carried by H_2, over a boat containing liquid Ga, and the reaction products, which are As_4 and GaCl, then flow further down the line and react to form GaAs on a substrate that is held at about 750 °C. In LPE, a supersaturated solution of As in Ga is placed in contact with a substrate and then slowly cooled, releasing the As and forming a layer of GaAs on top of the substrate. The advantages of VPE and LPE are simplicity of design; very high purity of the resulting layers (especially for VPE); and high growth rates (10–25 μm/h for VPE). The disadvantages are poor control of thin layers and poor uniformity over large areas. Although some clever solutions to these problems have been proposed and implemented, it is doubtful that VPE and LPE will be able to compete in producing thin, complex structures. However, VPE will continue to be used for applications requiring thick, high-quality layers, and LPE is still widely employed for optoelectronics applications involving ternary and quaternary alloys.

3. OPTICAL PROPERTIES

As stated earlier, GaAs has a direct energy gap that has a strong influence on its optical properties and in fact gives it a major advantage over Si for devices, such as lasers or photodetectors, which require absorption or emission of light. Furthermore, a whole range of wavelengths is available by forming ternaries such as $Al_xGa_{1-x}As$, and quaternaries such as $In_xGa_{1-x}As_yP_{1-y}$. Thus, the emerging optoelectronics industry is largely based on GaAs and other III-V compounds. In this section we will give a cursory review of the basic optical properties of GaAs. A good general account of optical properties in semiconductors can be found in the book by Pankove (1975).

3.1 Light Absorption

When light of intensity I_0 impinges upon a solid, some of it will be reflected, some absorbed, and the rest transmitted. The absorption should be an exponential process, and for a very thick sample (no light transmitted) we would expect that at a depth z into the material

$$I(z)=I_0(1-R)e^{-\alpha z}, \tag{2}$$

where R is the fractional reflectance and α is defined as the absorption coefficient. For thinner samples of thickness d, a fraction T will be transmitted, and after accounting for multiple internal reflections, T is given by

$$T=\frac{(1-R)^2e^{-\alpha d}}{1-R^2e^{-2\alpha d}}\,. \tag{3}$$

Transmission experiments based on Eq. (3) are widely used to determine α in semiconductor specimens. The absorption can arise from various processes, as enumerated below.

3.1.1 Band-to-Band, Exciton Absorption. A strong absorption takes place when the light energy $h\nu$ is sufficient to excite electrons from the valence band (VB) to the conduction band (CB) at the Γ point ($\mathbf{k}=0$). The absorption continues to increase above $h\nu=E_g$, as shown in Fig. 10, because the GaAs band structure (Fig. 3) allows many higher-energy transitions. The sharp peaks in Fig. 10 are due to exciton formation, i.e., bound states between the electrons and holes formed by the light. The excitons can either remain free or themselves become bound to donors and acceptors.

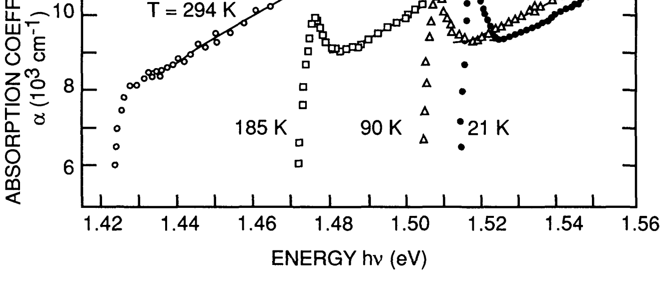

FIG. 10. Absorption coefficient vs photon energy for interband electronic transitions in GaAs at various temperatures. [From Sturge (1962). Reproduced by permission of the American Institute of Physics.]

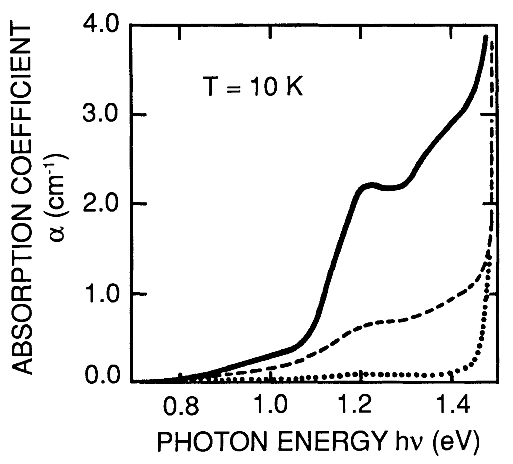

FIG. 11. Absorption coefficient vs photon energy for EL2/conduction-band electronic transitions in semi-insulating LEC GaAs. The dashed and dotted lines show photoquenching after white-light illumination for 1 and 10 min, respectively. [From Martin (1981). Reproduced by permission of the American Institute of Physics.]

Note that α quickly approaches $10^4\ \mathrm{cm}^{-1}$ for $h\nu > E_g$, which means that most of the light is absorbed within about 1 μm.

3.1.2 Band-to-Impurity or -Defect Absorption. An electron can be excited from a bound state of an impurity or defect to the conduction band if $h\nu > E_C - E_B$, where E_B is the energy of the bound electron. A good example of this process involves the famous defect EL2 (related to $\mathrm{As_{Ga}}$), as shown in Fig. 11. Since in SI GaAs, EL2 is only partially occupied, two excitation processes are possible: electrons to the CB from occupied states, and holes to the VB from unoccupied states. Then α is given by

$$\alpha(h\nu) = \sigma_n(h\nu)N^0_{\mathrm{EL2}} + \sigma_p(h\nu)N^+_{\mathrm{EL2}}, \qquad (4)$$

where σ_n and σ_p are the respective photoionization cross sections. Since σ_n and σ_p are known as functions of $h\nu$, both N^0_{EL2} and N^+_{EL2} can be determined by measuring α at two values of $h\nu$. Also, the Fermi energy can be determined from the fractional occupation.

3.1.3 Acceptor-to-Donor Absorption. An electron on an acceptor can be excited to a nearby donor, separated by distance r, if

$$h\nu = E_g - E_D - E_A + q^2/\epsilon r, \qquad (5)$$

where the last term is the Coulomb energy. Such transitions are commonly seen in GaAs if impurity concentrations are low, but usually in emission rather than absorption.

3.1.4 Free-Carrier Absorption. An electron or hole can make a transition to a higher state within the same band. Theoretically, it is expected that $\alpha \propto n\lambda^2$, where $\lambda = c/\nu$, but one set of data in GaAs suggests that $\alpha \simeq 7.5 \times 10^{-8}$ $n\lambda^3$ (if $\lambda > 4\ \mu\mathrm{m}$), for n and λ in the same unit system (Spitzer and Whelan, 1959). For example, if $n \simeq 10^{17}\ \mathrm{cm}^{-3}$ and $\lambda = 4 \times 10^{-4}$ cm, then $\alpha \simeq 0.5\ \mathrm{cm}^{-1}$, a rather small number. Thus, free-carrier absorption is usually not a problem in GaAs unless $n\lambda^3 > 10^7$ (e.g., n or $p > 10^{19}$ if $\lambda = 1\ \mu\mathrm{m}$).

3.1.5 Phonon Absorption. The phonon (lattice-vibration) bands can absorb light irradiation because the unequal charges on dissimilar atoms form dipoles. However, two or more phonons need to be involved in order to conserve momentum, since the photon momentum is very small. A portion of the lattice absorption spectrum is given in Fig. 12. For example, the strong peak at 64.8 meV is thought to be due to two TO (transverse optical) phonons, whereas much of the spectrum beyond that point is evidently due to three-phonon absorption.

3.1.6 Local Vibrational-Mode Absorption. Atoms that are light with respect to the host, such as Be, C, B, and Si in GaAs, can vibrate at high frequencies and be observed as relatively sharp peaks in the long-wavelength spectrum.

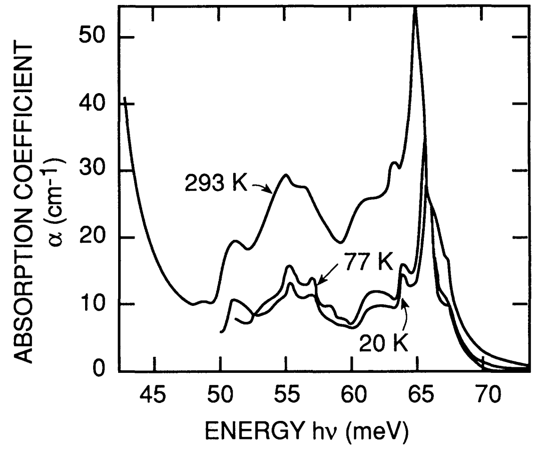

FIG. 12. Absorption coefficient vs photon energy for some of the lattice vibrational modes in GaAs. [From Cochran *et al.* (1961). Reproduced by permission of the American Institute of Physics.]

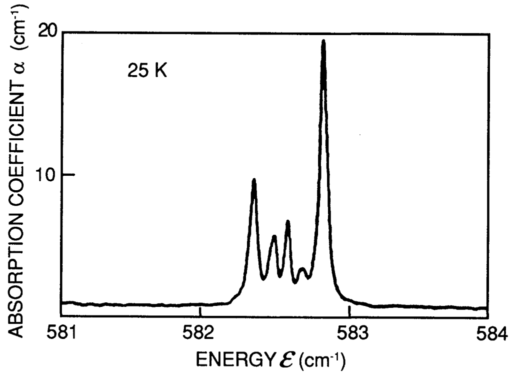

FIG. 13. Absorption coefficient vs photon energy for the carbon local-vibrational modes in GaAs. [From Theis *et al.* (1982). Reproduced by permission of the American Institute of Physics.]

From simple considerations, we would expect that $2\pi\nu \simeq (2k/m)^{1/2}$, where k is the force constant and m the mass of the impurity. Thus, for example, $^{11}B_{Ga}$ has an energy of 517 cm^{-1} (19.3 μm), and $^{10}B_{Ga}$, 540 cm^{-1} (18.5 μm), about the predicted ratio for the isotopes. Oxygen (interstitial) has lines near 730 cm^{-1}, $^{9}Be_{Ga}$ near 482 cm^{-1}, $^{28}Si_{Ga}$ near 384 cm^{-1}, and $^{12}C_{As}$ near 582 cm^{-1}. The local-vibrational-mode (LVM) spectrum of C, shown in Fig. 13, is especially interesting since it displays a nearest-neighbor isotope effect because the Ga atoms in the surrounding tetrahedron can be of either mass 69 or mass 71 (Theis *et al.*, 1982). The energy splittings, which can be fairly well predicted, are quite small, and require a high-resolution spectrometer to be observed. Note that the appearance of these splittings proves that C is on an As site, since if it were on a Ga site, the surrounding As atoms would all be of mass 75 and would produce no splitting. The LVM intensities of the above impurities can be used for concentration determinations, since the absorption cross sections are known. For this purpose, fortunately, a high-resolution instrument is not necessary.

3.2 Light Emission: Photoluminescence

The transitions involved in light absorption, described in Secs. 3.1.1–3.1.3, can also be responsible for light emission, if electrons or holes are in the excited states. However, for transitions involving band states the absorption process will have a much broader spectrum than the emission process, because absorption transitions can occur from any filled state to any empty state, whereas emission will occur primarily from a narrow, thermalized distribution of the excited carriers. Thus, sharp photoluminescence (PL) signals may be obtained from impurity concentrations as low as 10^{10} cm^{-3}, as long as extraneous broadening mechanisms are not present, and this capability makes PL extremely useful for semiconductor characterization.

Photoluminescence spectra representing four different samples (GaAs 1, 2, 3, and 4) are shown in Figs. 14 and 15 (Steiner *et al.*, 1989). Here X denotes either an exciton or an exciton-photon bound state known as a polariton; D denotes one of the hydrogenic donors, which are difficult to identify in GaAs because of very small energy separations between the various donor species; and the other symbols are obvious except for U which denotes an unknown band. Thus, the energy for the peak designated X is that resulting from the annihilation of a free exciton (or polariton), while the energy from the annihilation of an exciton bound to a neutral donor (D^0,X) is lower because of the binding energy. Conduction-band–to–acceptor transitions (Sec. 3.1.2) are represented by (e,C), and donor-acceptor-pair (DAP) transitions (Sec. 3.1.3) by (D^0,C),

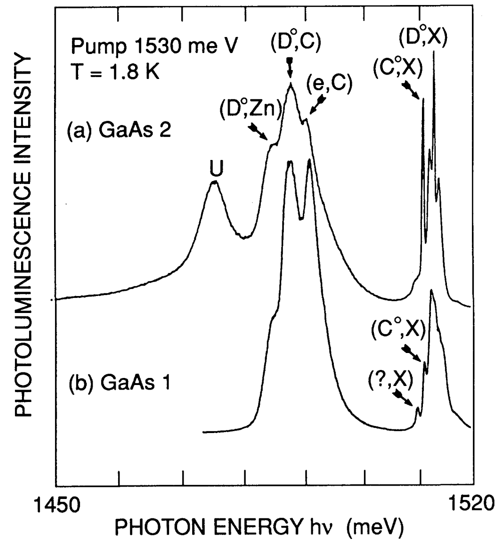

FIG. 14. Photoluminescence spectra in GaAs. [From Steiner *et al.* (1989). Reproduced by permission of the National Research Council of Canada.]

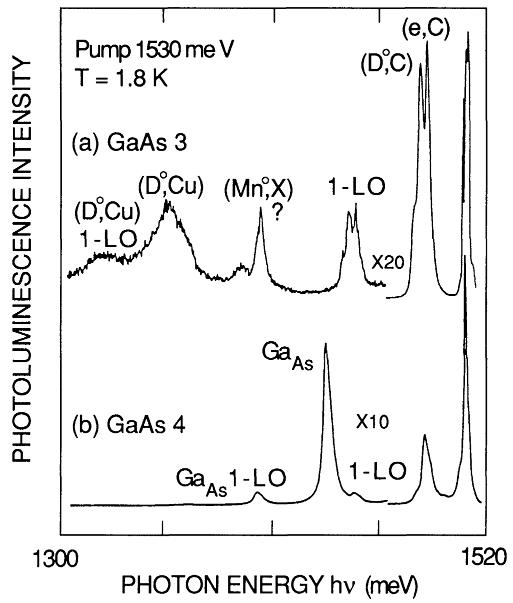

FIG. 15. Photoluminescence spectra in GaAs. [From Steiner *et al.* (1989). Reproduced by permission of the National Research Council of Canada.]

(D^0,Zn), and (D^0,Cu). The transition involving the defect Ga_{As} is also possibly of DAP nature. Furthermore, note that in several cases LO phonons are involved in the transitions. In general, PL spectra in GaAs nearly always show C, evidently a ubiquitous impurity, and in undoped SI GaAs a line at 0.68 eV due to the defect As_{Ga} is usually present.

Excited electrons and holes may also recombine through nonradiative (NR) channels which then diminish the overall PL intensity. Such channels are most effective if they involve a continuum or quasicontinuum of energy states between the initial excited state and final ground state, because then low-energy phonons can efficiently support the transitions through the chain of states. In some cases, higher-energy transitions can occur by the emission of several phonons at the same time, but this is a more unlikely process. Surfaces, or large defects, that involve a high density of dangling bonds can be efficient supporters of NR recombination. Thus, a high PL intensity often is more a result of good surface and bulk crystal quality rather than a high number of radiative recombination centers.

3.3 Modulated Reflectance: Photoreflectance

A disadvantage of PL measurements is that they must be carried out at very low temperatures, $T \lesssim 10$ K, in order to avoid kT broadening of the lines which involve a band, or even total loss of the lines for which the initial state is a shallow donor or acceptor. Furthermore, excitons are stable only at low temperatures in bulk material. Absorption (α) measurements at higher temperatures (see Fig. 10) still give a reasonably sharp response at the threshold energy (even though broadened by kT), but have little structure above the threshold, since the excitons are absent. The reflectance R, on the other hand, can give a sharper response because of its dependence on the index of refraction n, as shown below. For normal incidence of the light

$$R = \frac{(n-1)^2 + k^2}{(n+1)^2 + k^2}, \tag{6}$$

where $k = c\alpha/4\pi\nu$ and c is the speed of light. Near and below the band gap, $k \ll n$ so that $R \simeq (n-1)^2/(n+1)^2$. From the Kramers-Kronig relationship,

$$\begin{aligned} n(\mathcal{E}) - 1 &= \frac{ch}{2\pi^2}\, \mathrm{P} \int_0^\infty \frac{\alpha(\mathcal{E}')}{\mathcal{E}'^2 - \mathcal{E}^2}\, d\mathcal{E}' \\ &= \frac{ch}{4\pi^2}\, \mathrm{P} \int_0^\infty \ln\left(\frac{1}{\mathcal{E}'^2 - \mathcal{E}^2}\right) \\ &\quad \times \frac{d\alpha(\mathcal{E}')}{d\mathcal{E}'}\, d\mathcal{E}', \end{aligned} \tag{7}$$

where P denotes the Cauchy principal value of the integral. Since $\alpha(\mathcal{E})$ depends on the joint density of states, which is the highest near critical points—i.e., points at which $d[E_C(k) - E_V(k)]/dk = 0$—the derivative $d\alpha(\mathcal{E})/d\mathcal{E}$ will be a sharply varying function of $\mathcal{E}$. In the presence of an electric field E, α will become

$$\alpha(\mathcal{E},E) = \alpha(\mathcal{E},0) + \Delta\alpha(\mathcal{E},E) \tag{8}$$

and thus

$$\Delta n(\mathcal{E},E) = \frac{ch}{2\pi^2} \int_0^\infty \frac{\Delta\alpha(\mathcal{E}',E)}{\mathcal{E}'^2 - \mathcal{E}^2}\, d\mathcal{E}'. \tag{9}$$

When the field E is modulated, then Δn and thus R will also be modulated. An obvious way to modulate the field is to put a thin, transparent Schottky barrier on the sample and

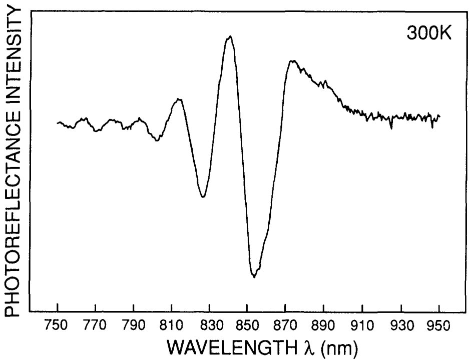

FIG. 16. Photoreflectance spectrum in a GaAs MBE layer.

apply an ac voltage. However, a far more convenient way is simply to change the built-in surface field by shining a modulated laser beam on the sample, since the free carriers produced will neutralize surface states and decrease the electric field. This phenomenon is called photoreflectance (PR). Besides obtaining information on the surface states, it is also possible to determine the bulk carrier density, since the initial electric field depends on this density. An example of PR spectra for a 1000-Å, undoped layer of MBE GaAs grown on a 2500-Å, *n*-type layer doped with 2×10^{17} cm^{-3} Si is shown in Fig. 16. The oscillations are a manifestation of the Franz-Keldysh effect, which is concerned with the behavior of the electronic wave function at a band edge in the presence of an electric field. The oscillations in this case can be easily analyzed to give a field $E=5\times10^4$ V/cm, very close to what was predicted from the expected surface potential, the doping in the active layer, and the thickness of the undoped layer. The photoreflectance technique has been reviewed by Pollak and Shen (1990) and Bottka *et al.* (1988).

4. ELECTRICAL PROPERTIES

4.1 Hall-Effect Measurements

The operation of virtually all GaAs devices involves the flow of charge carriers; therefore, the electrical properties of the material are very important. Ohm's law states simply that the current density $\mathbf{J}=\sigma\cdot\mathbf{E}$, where σ is the conductivity tensor. If the electric field **E** is applied in the x direction, and magnetic field **B** in the z direction, then for a large class of carrier scattering mechanisms (to be discussed later) it can be shown that (for electrons)

$$J_x=\frac{\sigma_{xx}^2+\sigma_{xy}^2}{\sigma_{xx}}E_x\bigg|_{B\to0}=\sigma_{xx}E_x=\frac{ne^2\langle\tau\rangle}{m^*}=ne\mu E_x, \tag{10}$$

$$R_{\mathrm{H}}=\frac{1}{B}\frac{\sigma_{xy}}{\sigma_{xx}^2+\sigma_{xy}^2}\bigg|_{B\to0}=-\frac{1}{ne}\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2}=-\frac{r}{ne}, \tag{11}$$

where $\langle\tau\rangle$ is an average (over energy) of the momentum relaxation time, and the conductivity mobility μ is defined as $\mu=e\langle\tau\rangle/m^*$. (Note that the symbols σ and n should not be confused with those used in the previous section.) The quantity R_{H}, the Hall coefficient, is quite useful in all of semiconductor physics because it is very simply related to the carrier concentration n, and because it is easily measured: $R_{\mathrm{H}}=E_y/J_xB$. Another very useful quantity is $\sigma R_{\mathrm{H}}=-r\mu\equiv\mu_{\mathrm{H}}$, the Hall mobility, where we have redefined $\sigma_{xx}=\sigma$. A calculation of r vs T for electrons in relatively pure GaAs is shown in Fig. 17. It is clear that r deviates very little from unity in this case. In fact, at 296 K, for n ranging from 1×10^{12} to 1×10^{18} cm^{-3}, and compensation ratio N_I/n ranging from 1 to 10, r varies only between 1.02 and 1.26 (Look, 1989). For holes, r is more difficult to calculate and can be somewhat larger, although it is usually less than 2. In any case,

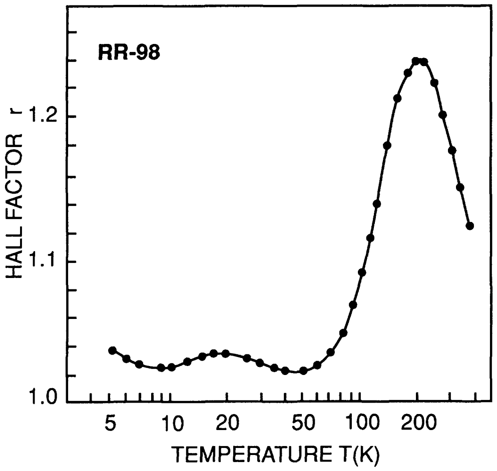

FIG. 17. Temperature dependence of the electron Hall factor *r* for a relatively pure, VPE GaAs layer. [From Look (1989). Reproduced by permission of John Wiley and Sons.]

Hall measurements give very useful approximations to n and μ and thus are widely used for their determinations. Below, we relate the quantities n and μ to materials parameters.

4.2 Electron and Hole Concentrations

A useful relationship in semiconductor analysis is the condition of charge balance:

$$n+\sum_i N_{Ai}^- = p+\sum_j N_{Dj}^+ , \tag{12}$$

where it is assumed here, for simplicity, that all acceptors either are neutral or have a single negative charge, and the same for donors except that the charge would be positive. More complex situations can be handled in a straightforward manner, although the equations can become tedious (Look, 1989). Typically, only one impurity or defect in the sample will be partially ionized, and all the others will be either neutral or fully ionized. If the partially ionized center is a donor, and if the sample is strongly n-type, then $n \ll p$ and Eq. (12) becomes

$$n+N_A^{\mathrm{net}} = N_D^+ = \frac{N_D}{1+(g_1/g_0)(n/N_C)e^{-\alpha/k}\exp(E_{D0}/kT)} , \tag{13}$$

where N_A^{net} is the total concentration of acceptors having energies below that of the donor N_D minus the total concentration of other donors having energies above that of N_D. Also, g_1/g_0 is a degeneracy factor, $E_D = E_{D0} - \alpha T$ is the temperature-dependent donor energy with respect to the conduction band (here the symbol α differs from that used in the previous section), and $N_C = 2(2\pi m^* kT)^{3/2}/h^3$ is an effective conduction-band density of states, which holds as long as the Fermi energy E_F is more than a few times kT below the conduction band. An example of fitting n vs T data by Eq. (13) is shown in Fig. 18. For example, the fit to sample RR-98, a VPE layer, has the parameters $N_D = 3.1\times10^{14}$ cm^{-3}, $N_A^{\mathrm{net}} = 2.5\times10^{13}$ cm^{-3}, $E_{D0} = 4.3$ meV, and $(g_0/g_1)\times\exp(\alpha/k) \simeq 1$. Here N_D and N_A are known to represent mainly Si and Zn, respectively, by comparison with PL experiments and other data. It is interesting that this particular N_A is the lowest acceptor concentration ever measured in GaAs, although numbers this low are more common in Si and Ge. Many other donors and acceptors in GaAs have been characterized in this way, with E_{D0} being the prime "fingerprinting" parameter. Some results are given in Table 2.

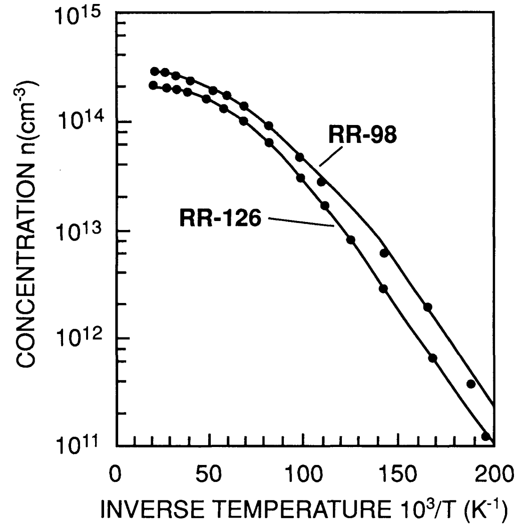

FIG. 18. Electron concentration vs inverse temperature for two VPE GaAs layers. Sample RR-98 has a very low compensation ratio. [From Look and Colter (1983). Reproduced by permission of the American Institute of Physics.]

4.3 Mobilities

Electrons or holes suddenly subjected to an electric field do not accelerate indefinitely, but suffer collisions, which eventually lead to a steady-state velocity (analogous to that of a skydiver freefalling through the air). The time required for the velocity (or momentum) to relax was designated τ in Sec. 4.1, and the mobility was shown to be $\mu = e\langle\tau\rangle/m^*$, where the brackets denote averaging over energy. An implicit assumption in the derivation leading to that form of μ is that the collisions are elastic, i.e., that the energy does not change as a result of the collision, even though the momentum does. If that is true, then at a given electron energy the total collision *rate* (τ^{-1}) should just be equal to the sum of the individual collision rates, i.e., $\tau^{-1} = \tau_1^{-1} + \tau_2^{-1} + \tau_3^{-1} + \cdots$. For many of the collision or scat-

Table 2. E_{D0} (donor) or E_{A0} (acceptor) values determined by Hall-effect measurements for various centers in GaAs

Species	Type	Energy (eV)
Si_{Ga}, Ge_{Ga}, Sn_{Ga}, S_{As}, Se_{As}, Te_{As}	Donor	$E_C-0.006$
Defects, V_{As}-related	Donor	$E_C-0.13$ to 0.20
Defect?	Donor	$E_C-0.43$
As_{Ga}	Donor, first level	$E_C-0.75$
	Donor, second level	$E_V+0.52$
Cr_{Ga}	Donor	$E_V+0.32$
Cr_{Ga}	Acceptor	$E_C-0.78$
V_{Ga} (vanadium)	Acceptor	$E_C-0.2$
Fe_{Ga}	Acceptor	$E_V+0.52$
Ni_{Ga}	Acceptor	$E_V+0.46$
Cu_{Ga}	Acceptor, first level	$E_V+0.15$
	Acceptor, second level	$E_V+0.44$
Ni_{Ga}	Acceptor	$E_V+0.20$
Co_{Ga}	Acceptor	$E_V+0.16$
Mn_{Ga}	Acceptor	$E_V+0.11$
C_{As}, Be_{Ga}, Mg_{Ga}, Zn_{Ga}, Cd_{Ga}	Acceptor	$E_V+0.026$ to 0.035

tering mechanisms in GaAs, such as Coulomb scattering from ionized impurities or scattering involving low-energy acoustic phonons, that description is correct. However, at 296 K in GaAs the strongest scattering mechanism results from the absorption or emission of optical phonons, and the most important of these (the LO phonon) has an energy of 0.036 eV, greater than kT (0.025 eV). Thus, for optical-phonon scattering, τ would depend on both initial- and final-state energies, and thus would not be a useful description of the scattering process. Because of this problem, an accurate analysis of mobility in GaAs must involve a numerical solution of the Boltzmann transport equation. However, the *measurement* of mobility, as shown earlier, is easily accomplished by using the Hall effect: $\mu_H = R_H\sigma$.

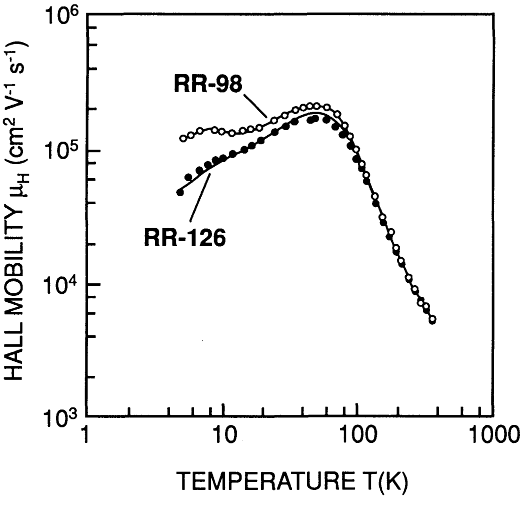

FIG. 19. Mobility vs temperature for the same two GaAs layers as shown in Fig. 18. [From Look and Colter (1983). Reproduced by permission of the American Institute of Physics.]

In Fig. 19 are shown μ vs T fits for the same two samples displayed in Fig. 18. Here the fit of N_A for sample RR-98 is 1.5×10^{13} cm^{-3}, rather than the 2.5×10^{13} cm^{-3} deduced (more accurately) from the n vs T data. Discrepancies of this magnitude are not unusual because of the complexity of mobility analysis.

Although sample RR-98 is quite pure, at low temperatures the ionized donors and acceptors still make their presences known, as shown in Fig. 20. The solid line is the theoretical μ vs T for a *completely* pure sample, and it is seen that μ should approach 7.3×10^6 cm^2/V·s at liquid-helium temperature, 4.2 K. In fact, a value of 1.2×10^7 cm^2/V·s has been measured at 0.35 K by separating the electron-producing ionized donors from the electrons themselves in an $Al_{0.35}Ga_{0.65}As$/GaAs heterostructure (Pfeiffer *et al.*, 1989). Such a value is in good agreement with the theory shown in Fig. 20.

As a figure of merit, μ(77 K) is much better than μ(296 K), because ionized impurities and defects are much stronger scattering centers at low temperatures. An empirical curve relating $\mu(77)$ to N_D+N_A in bulk material is shown in Fig. 21 (Wolfe *et al.*, 1970). Thus,

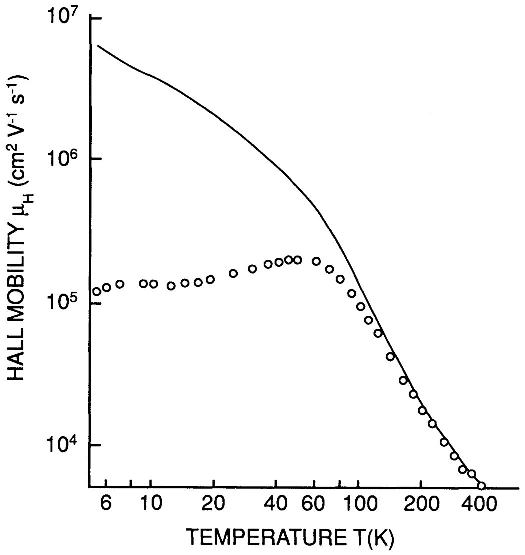

FIG. 20. The experimental Hall mobility (circles) vs temperature for a relatively pure, VPE GaAs layer. Also shown is the theoretical curve (solid line) for GaAs with no defect or impurity scattering. [From Look (1989). Reproduced by permission of John Wiley and Sons.]

any differentiation at this level of quality must be made on the basis of the *peak* mobility, which will occur at a lower temperature, the better the sample. As of this writing, the record peak mobilities, the temperatures at which they occur, and the references are as follows: VPE, 3.05×10^5 cm^2/V·s at 43 K, (Wolfe *et al.*, 1970); MOVPE, 3.35×10^5 cm^2/V·s at 38 K (Razeghi *et al.*, 1989); and MBE, 4.0×10^5 cm^2/V·s at $\simeq$35 K (Stanley *et al.*, 1991). For the VPE result, n vs T fitting gives $N_D = 4.6 \times 10^{13}$, $N_A = 3.0 \times 10^{13}$ cm^{-3}, and $E_{D0} = 5.57$ meV. Unfortunately, n vs T data were evidently not analyzed for the other two cases, but $n(77) \simeq N_D - N_A$ was about 3×10^{13} cm^{-3} in both cases. (None of these results have been corrected for Hall r factor.) Clearly, all of these techniques are presently capable of producing very pure GaAs layers although that was not the case for MBE and MOVPE until recently. It also should be realized that such high purities would almost certainly not be obtained in a production environment.

Before leaving the discussion of mobility it is important to note that the linear (Ohmic) relationship $v = \mu E$ does not hold to infinitely high field E, but in fact the electron velocity v goes through a peak at about $E \simeq 4$ kV/cm, and then settles to a relatively constant value near 1×10^7 cm/s. The reason for the peak is that electrons gain enough energy at $E \simeq 4$ kV/cm to transfer from the lowest $\mathscr{E}$ vs k valley (Γ) to the next higher valley (L), and the L valley has a lower mobility (because of a higher effective mass). This effect can be very important in the operation of high-speed devices (see Sec. 5) and in fact is the basis of one of them, the Gunn diode. Further discussion of the v vs E relationship can be found in Blakemore (1982).

5. APPLICATIONS

The general interest in GaAs began with the successful development of bulk crystal growth in the early 1950s, but greatly accelerated as novel devices emerged in the early 1960s. Today many different types of electronic, optical, and optoelectronic devices are based on GaAs, or at least use GaAs in some aspect. As yet, there is no large-scale consumer market, such as computers or television, to demand any of these devices in great numbers, but at least three such markets loom on the horizon: electronically scanned (phased-array) radar, direct satellite-to-home TV broadcasting, and collision-avoidance radar for automobiles. Other, smaller markets will keep the industry going, and perhaps slowly expanding, so that there is reason to develop existing devices further and design new ones. In this section,

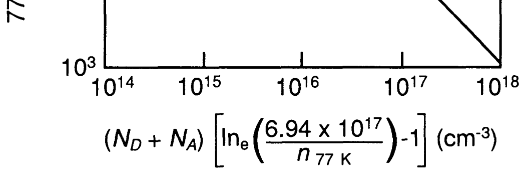

FIG. 21. The 77-K electron Hall mobility in GaAs vs an expression involving N_D, N_A, and n. [From Wolfe *et al.* (1970). Reproduced by permission of the American Institute of Physics.]

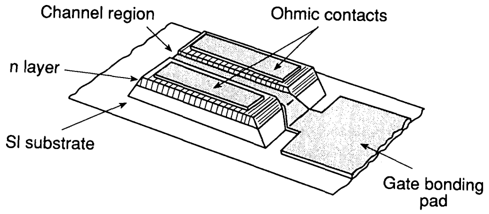

FIG. 22. Schematic of a typical MESFET. [From Arnold *et al.* (1989). Reproduced by permission of the IEE Publishing Dept.]

we will consider devices that are already in use or have the potential to be utilized in the near future.

5.1 Discrete Electronic Devices

5.1.1 Metal-Semiconductor Field-Effect Transistors. Without a doubt, the "workhorse" of present-day GaAs technology is the metal-semiconductor field-effect transistor (MESFET). A generic MESFET is shown in Fig. 22, although high-frequency versions usually have a somewhat different design. The *n* layer ("active" layer) may be produced directly by implantation of Si into the SI substrate, or by epitaxial growth with Si doping. The gate is recessed, in this design, to obtain the desired thickness of the channel under the gate, and thereby control the pinch-off or threshold voltage. Before application of the Ohmic and gate metal, the active layer is isolated from the active layers of other devices by a "mesa" etch that removes the conductive layer between devices. In this way, each device can operate independently, since in theory, currents should not be able to flow through the high-resistivity substrate. In actuality, the presence of "cross-talk" between closely spaced devices remains a major hindrance to the realization of high-density integrated circuits. One solution has been to insert a thin (~ 1 μm) layer of MBE GaAs, grown at 200 °C, between the substrate and the active layer.

A cross section of a typical MESFET is shown in Fig. 23. The active-layer regions with cross-hatching are depleted of free electrons, which are lost to surface states, substrate-interface states, or the gate metal. The device works by applying a negative voltage to the gate, thus drawing more free electrons to the metal and further reducing the thickness of the channel under the gate. The change in channel charge dQ resulting from the change of gate voltage dV_g depends on the gate-to-source capacitance C_{gs}; i.e., $dQ=C_{gs}dV_g=(\epsilon l_g Z/w)dV_g$ where ϵ is the dielectric constant, w is the thickness of the depletion region under the gate, l_g is the length of the gate (parallel to the current flow), and Z is the width of the gate (perpendicular to w and l_g). Also, the drain current I_D will change since $dI_D=nev(dw)Z=vdQ/l_g$, where v is the velocity of the electrons. For high drain voltages, the field near the drain will be very high and most of the electrons will be moving at the saturation velocity v_s (cf. Sec. 4.3). The primary operational characteristic of the MESFET is the intrinsic transconductance g_{m0} normalized to unit gate width. If $v \simeq v_s$, then

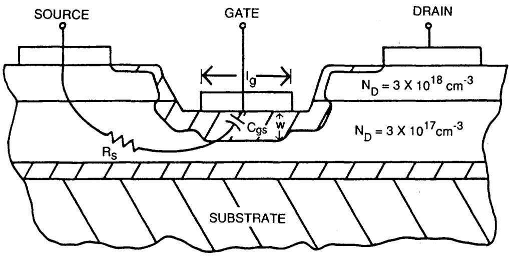
SOURCE
GATE
DRAIN
l_g
$N_D = 3 \times 10^{18}$ cm^{-3}
$N_D = 3 \times 10^{17}$ cm^{-3}
C_{gs}
w
R_s
SUBSTRATE

FIG. 23. Cross-sectional schematic of a typical MESFET. [From Goronkin *et al.*, (1985). Reproduced by permission of Howard W. Sams and Co.]

$$g_{m0}=\frac{dI_D}{dV_g}=C_{gs}\frac{v_s}{l_g}=\epsilon v_s\frac{Z}{w}\,. \tag{14}$$

Thus, it is very important to have a high saturation velocity v_s, large device width Z (perpendicular to plane of Fig. 23), and small depletion depth w. Unfortunately, too large a width Z limits the integration density, and too small a depth w can lead to gate leakage. Parasitic elements in the device will also limit its performance. One of the most harmful is the source resistance R_s, which comes from the source contact resistance and also from the "access" resistance due to the material between the source and gate. Because of R_s, the transconductance is reduced as follows:

$$g_m=\frac{g_{m0}}{1+g_{m0}R_s}\,. \tag{15}$$

Thus, for a normalized $g_{m0}=500$ mS/mm, a normalized R_s of 2 Ω mm will drop g_m in half. Therefore, much effort has been expended on reducing both the contact and access resistances. Of course, high mobility leads to low resistance, and, as discussed in Sec. 4.3, GaAs excels in this respect.

Another important operational parameter of MESFETs is the noise figure, defined as $NF=10\log_{10}F$ where

$$F=1+K_v\nu l_g[g_{m0}(R_g+R_s)]^{1/2}. \tag{16}$$

Here ν is the frequency of operation, R_g is the gate-metal resistance, and K_v is an undetermined constant on the order of unity. Although Eq. (16) is empirical, it has been shown to be representative of devices having a wide range of operating parameters. Thus, besides decreasing l_g and R_s, it is also of benefit to decrease R_g, usually by using a thick gate metal or *T*-shaped gate. Unfortunately, however, there is a trade-off between F and the transconductance g_{m0}, so that it is difficult to optimize gain and noise figure at the same time. Present technology allows l_g of 0.5 μm in production MESFETs, and down to 0.1 μm or even lower in laboratory devices. One of the best results (as of 1992) for an 0.5-μm gate-length MESFET is $NF=0.8$ dB along with an associated gain G_a of 6.3 dB at 16 GHz (Lau *et al.*, 1989). The K_v factor is 1.4 for this device.

The frequency response of MESFETs and other devices is usually measured in terms of f_T, the frequency at which the current gain descends to unity, and $f_{\max}$, the frequency at which the unilateral power gain is unity. Roughly speaking, f_T is more appropriate for low-noise applications, such as the front ends of amplifiers, and $f_{\max}$ for power applications. For a MESFET,

$$f_T=\frac{g_{m0}}{2\pi C_{gs}}=\frac{v_s}{2\pi l_g}, \tag{17}$$

$$f_{\max}=\frac{f_T}{2[(R_s+R_g+R_i)/R_{sd}+2\pi R_g C_{gd} f_T]^{1/2}}, \tag{18}$$

where R_i is a resistance in series with R_s and R_g in the usual equivalent circuit. Thus, the importance of using a material with high saturation velocity, such as GaAs, is obvious. Also important is the need to make the gate length as small as possible, but here there are trade-offs, since reproducibility, yield, and uniformity over the wafer will probably suffer. The best values of f_T reported (as of 1992) for a GaAs MESFET are $f_T=55$ GHz for $l_g=0.5$ μm, and $f_T=126$ GHz for $l_g=0.25$ μm (Feng *et al.*, 1990). By adding 18% In to the GaAs ($In_{0.18}Ga_{0.82}As$), values of $f_T=126$ GHz and $f_{\max}=232$ GHz are achieved for ion-implanted devices with $l_g=0.20$ μm.

5.1.2 Heterojunction Field-Effect Transistors. A limitation of MESFETs is that the free electrons providing the conduction will be scattered by the donors that supplied them. Such scattering diminishes the mobility μ, which increases R_s, and to a lesser extent reduces v_s. A glance at Eqs. (14)–(18) shows that v_s and R_s are prominent components of the figures of merit g_{m0}, g_m, NF, and f_T, so that it would be very helpful to reduce the donor scattering. This can be accomplished by a clever device pictured in Fig. 24, and variously known as the heterojunction FET (HFET), modulation-doped FET (MODFET), two-dimensional electron-gas FET (TEGFET), selectively doped heterojunction transistor (SDHT), or high-electron-mobility transistor (HEMT). The principal feature of this device is a heavily doped $Al_xGa_{1-x}As$ ($x\simeq0.2$–0.4) layer on top of an undoped GaAs layer (see review by Tu *et al.*, 1985). Since the lowest conduction band

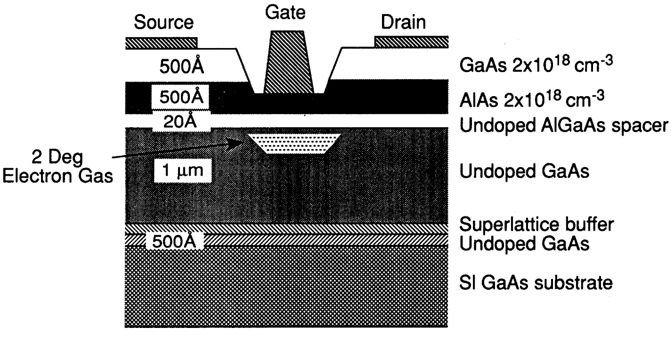

FIG. 24. Cross-sectional schematic of a heterojunction structure. [From Arnold *et al.* (1989). Reproduced by permission of the IEE Publishing Dept.]

in $Al_xGa_{1-x}As$ is higher in energy than that in GaAs, the electrons in the $Al_xGa_{1-x}As$ will transfer to the undoped GaAs, which should result in a higher μ (lower R_s) and higher v_s. Besides the "standard" $Al_xGa_{1-x}As/GaAs/$ GaAs substrate design, newer versions include $Al_xGa_{1-x}As/In_yGa_{1-y}As/GaAs$ substrate, and $Al_xIn_{1-x}As/In_yGa_{1-y}As/InP$ substrate. State-of-the-art results are $f_T=150$ GHz and $f_{max}=405$ GHz (corrected for gate-pad capacitance) for an 0.15-μm gate-length $In_{0.52}Al_{0.48}As/In_{0.53}Al_{0.47}As/InP$ device (Chao *et al.*, 1990); $NF=1.4$ dB with $G_a=6.6$ dB at 93 GHz for the same device; $f_T=120$ GHz, $f_{max}=290$ GHz, and $NF=2.1$ dB with $G_a=6.3$ dB at 94 GHz for an 0.1-μm gate-length $Al_{0.25}Ga_{0.75}As/In_{0.15}Ga_{0.85}As/GaAs$ device (Tan *et al.*, 1990).

An advantage of HFETs over MESFETs seems to be a substantially lower noise figure and higher gain at a given frequency. For example, the device mentioned above with $NF=1.4$ dB at 93 GHz had $NF=0.3$ dB and $G_a=17.2$ dB at 18 GHz, which are much better results than those reported for MESFETs in the previous section. Thus, HFETs should be superior for applications in which noise is a dominant consideration such as in radio astronomy. Also, the advantage of HFETs over MESFETs increases rapidly as temperature is lowered (e.g., the low-temperature mobility of HFETs is much higher), and so HFETs will be more likely used if cooling is practical. However, the economy of MESFET production will prevail in many of the less demanding applications.

5.1.3 Heterojunction Bipolar Transistors. Another candidate for certain digital and microwave applications is the heterojunction bipolar transistor (HBT), pictured in Fig. 25. A problem with regular bipolar designs is hole injection from the base into the emitter. This limits the doping in the base region, which leads to an unacceptably high base resistance R_b and low f_{max}, since

$$f_{max}=(f_T/8\pi R_b C_{bc})^{1/2}. \qquad (19)$$

Because $Al_xGa_{1-x}As$ has a lower valence band edge than GaAs, hole injection from a GaAs base to an $Al_xGa_{1-x}As$ emitter is reduced, and the GaAs base doping can be greatly increased, thus reducing R_b. State-of-the-art results for HBTs are $f_T=165$ GHz and $f_{max}=100$ GHz for a device with a 3.6×3.6-μm^2 InP emitter on an InGaAs base (Chen *et al.*, 1989); $f_T=65$ GHz and $f_{max}=218$ GHz for a 1×1-μm^2 $Al_{0.3}Ga_{0.7}As$ emitter on a GaAs base (Asbeck *et al.*, 1989). Clearly, there is a trade-off between f_T and f_{max}. Also, the noise figures, while better than those in Si bipolar designs, are much poorer than those in MESFETs or HFETs. For example, published results for a device having a 2×2-μm^2 emitter are $NF\simeq 4.6$ dB with $G_a=10.4$ dB at 16 GHz (Asbeck *et al.*, 1989). Thus, the HBT would probably not be used in low-noise microwave applications. Even for power microwave applications it has been suggested that the smaller emitter sizes necessary for high-frequency capabilities would make the thermal resistance unacceptably high; however, some useful power devices have been fabricated. Overall, the very high transconductances available (as high as 10,000 mS/mm reported) and correspondingly high current-drive capabilities, along with good threshold voltage control, make the HBT very attractive for digital applications.

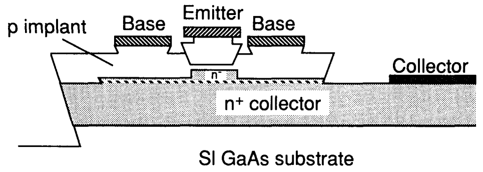

FIG. 25. Cross-sectional schematic of an HBT structure. [From Arnold *et al.* (1989). Reproduced by permission of the IEE Publishing Dept.]

5.2 Integrated Circuits (ICs)

For digital logic and memory circuits there has long been a drive to put increasingly larger numbers of devices on the same chip in order to satisfy size, cost, and speed requirements. Although later in coming, the microwave field is now experiencing a similar drive, due mainly to size, cost, and reliability constraints. A microwave circuit consisting of transistors, capacitors, inductors, and resistors on the same chip is known as a monolithic microwave integrated circuit (MMIC). We will discuss GaAs-based digital ICs and MMICs, and make comparisons with the competitive Si technology.

5.2.1 Digital ICs. Because of its inherent speed advantages (cf. Sec. 4.3), it has long been forecast that GaAs would soon overtake Si as the material of choice for high-speed digital applications. The reason that this has not yet happened is perhaps best illustrated by considering the most ideal logic design. When thousands and even millions of devices are placed very close to each other on a small chip, the basic logic units should be as simple as possible, and should dissipate very little power. One such design is called direct-coupled FET logic (DCFL), and a simple DCFL inverter circuit is illustrated in Fig. 26. The lower ("switching") transistor is an enhancement ("E") mode MESFET, and the upper ("pull-up") transistor is a depletion ("D") mode device. For a typical logic "0" input voltage (less than the threshold voltage of ~0.2 V), the switching (E-mode) transistor will not conduct, and V_{out} will be held at $V_{DD}=1.7$ V since the pull-up D-mode transistor will conduct. However, in real life, the output will be connected to the gate of another switching transistor [Fig. 26(b)] which will then conduct and drop V_{out} to the Schottky barrier potential of a GaAs MESFET, about 0.7 *V* (logic "1" in this case). Further connections to other gates (more "fanout") will drop V_{out} even more, perhaps to about 0.6 *V* for a fanout of four. Thus, in this case the actual logic "1" is ~0.6 *V*, rather than 1.7 *V*. This example points out the first problem with GaAs MESFETs, the low maximum gate voltage (~0.7 V), which is an inherent feature of the GaAs Schottky barrier. In Si, a native oxide (SiO_2) is inserted between the semiconductor and metal [metal/oxide/semiconductor (MOS) device], which allows much higher voltages on the gate before conduction. Then logic "1" can be kept at a high value, and noise spikes or variations in device processing are not as critical, even for much higher fanouts; i.e., the "noise margin" is higher. Unfortunately, the GaAs native oxide is poor and introduces a large surface-state density, as do most other insulators on GaAs. Another problem is that a non-negligible current is drawn by the E-mode device, even in the off state (~0.1 *V*), so that not too many input transistors can be added in parallel (low "fanin"). Because of the small difference between the off voltage (~0.1 *V*) and the threshold voltage (0.2–0.3 *V*), the device uniformity must be controlled very precisely, and this has been a problem in the past. Recently, however, HFET threshold voltages have exhibited standard deviations of only 11 mV for E-HFETs and 14 mV for D-HFETs over a full 3-in. wafer.

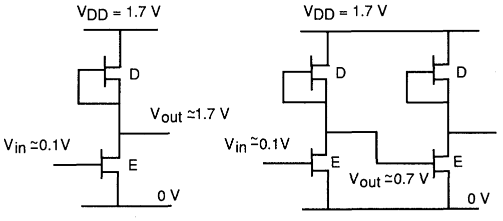

(a) (b)

FIG. 26. Schematic of a logic circuit. [From Flynn *et al.* (1989). Reproduced by permission of the IEE Publishing Dept.]

A way to achieve higher allowable voltage swings between turnoff and turnon is to use depletion-mode technology throughout the design, and hold the gates at a high negative voltage in the off state. Two such logic designs are presently popular: buffered FET logic (BFL) and Schottky-diode FET logic (SDFL). Unfortunately, both of these schemes require a second power supply as well as level-shifting diodes to provide the proper voltages on the gates of the fanout transistors. Although increasing the noise margin, as well as fanin and fanout capabilities, these designs also greatly increase the total number of circuit elements required, and consequently the total power dissipation. Many digital MESFET, HFET, and HBT logic circuits are now being produced, and some of the more conservative and reliable designs have found their way into the computer marketplace. Convex Computer Corporation has introduced a new series of real-time computers that use both BiCMOS (bipolar complementary MOS) and GaAs logic circuits and have performance at the 800-Mflops level. Also, Cray Computer Corporation will soon introduce the Cray-3 Supercomputer, with up to 16 processors, giving a 16-Gflops capability. All of the logic functions in the Cray machine are performed by GaAs devices, which are designed around a conservative 1-μm gate-length, depletion-mode MESFET technology. The basic devices have 100-ps gate delays, while the circuits are clocked at 2 ns (500 MHz). Gate complexities range from 500 to 2000 gates, and more than 450 different logic types are employed.

Circuits with VLSI (very large scale integration) complexity are also being developed (see VLSI TECHNOLOGY). Honeywell has reported a 16×16 complex multiplier chip for digital signal processing applications. This unit uses 1-μm HFET technology implemented in a DCFL scheme, and consists of 4500 gates on an 7.8×8.5-mm^2 chip. Operated at 500 MHz, the unit can perform a full 16-bit complex multiplication every 8 ns.

For memory circuits, 4-K SRAMs (static random access memories) have been developed in both MESFET and HFET technology, with the latter having an 0.5-ns access time. Also, a 16-K HFET SRAM with an access time of 3.4 ns at 77 K has been produced. This unit has a total of 107 519 devices and is implemented in E/D DCFL technology, as are most of the HFET SRAM designs. With switching delays of less than 10 ps now available from HFET devices, it seems likely that GaAs memory devices will find use in some specialized applications in the future, even though their integration levels are still well below those of Si. A review of HFET LSI (large scale integration) technology has been given by Abe *et al.* (1989).

5.2.2 Microwave ICs (MMICs) One area in which GaAs already dominates Si is in microwave applications at frequencies above 1 GHz. As discussed in Sec. 5.1, f_T of greater than 100 GHz has already been demonstrated in all the basic technologies discussed here: MESFET, HFET, and HBT. Not only are these devices faster than any Si devices, but GaAs also has the enormous advantage of having readily available semi-insulating substrates, with $\rho > 10^7$ Ω cm at 296 K. Thus, MESFETs, MIM (metal/insulator/metal) capacitors, spiral inductors, transmission lines, resistors, air bridges, and via holes can all be fabricated on a single chip without having to worry about parasitic elements due to substrate conductivity. The resulting circuits are called MMICs (monolithic microwave integrated circuits) and are illustrated in Fig. 27. MMIC technology is impractical in Si, because stable Si substrates with $\rho > 10^4$ Ω cm, the minimum usable resistivity, are not available. The main advantages of MMIC circuits over discrete and hybrid circuits are in cost, size, and reliability.

Many MMIC circuits, covering a frequency range of dc to 115 GHz, have been designed and fabricated from single-function circuits such as small-signal and power amplifiers, switches, phase shifters, oscillators, and mixers, to multifunction circuits such as direct-broadcast-satellite receivers [with low-noise rf (radio-frequency) amplifiers, mixers, oscillators, and IF (intermediate frequency) amplifiers] and transmit/receive modules for phased-array radar (with low-noise and power amplifiers, phase shifters, and reciprocal switches). Other applications include electronic warfare instruments, decoy seekers, and smart munitions, in the military area; satellite transponders, mobile communication units, and position indicators, in the communications area; and cellular telephones, automotive anticollision devices, and high-definition TV, in the consumer area.

It is likely that MMIC will be the biggest driver of GaAs technology in the near future. At least 25 companies around the world have made significant contributions to MMIC designs since the first *X*-band amplifier was introduced by Plessey in 1974, and other companies are becoming involved each year. A major boost to MMIC technology in the United States is resulting from a 6-year (1988–1994), >$500M program, funded by the Defense Department. Although specifically intended for military applications, it will undoubtedly have an impact in the consumer

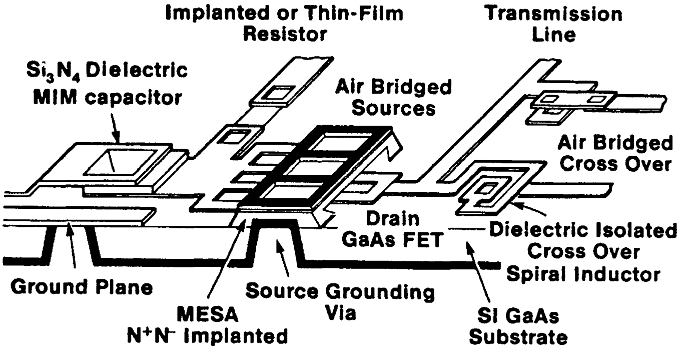

FIG. 27. Schematic of a MMIC circuit. [From Arnold *et al.* (1989). Reproduced by permission of the IEE Publishing Dept.]

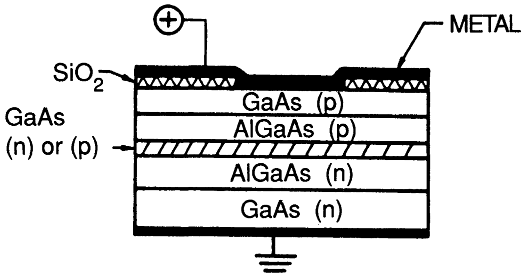

FIG. 28. Schematic of a double heterojunction injection laser. [From Shen and Hartman (1985). Reproduced by permission of Howard W. Sams and Co.]

arena, also. Reviews of MMIC technology have been given by Arnold *et al.* (1989) and Pucel (1985).

5.3 Optical Devices

Because of its direct band gap, GaAs is an efficient emitter and detector of light, an advantage not enjoyed by Si (cf. Sec. 3.1.1). This property has led to many GaAs-based optical devices, including light-emitting diodes (LEDs), lasers, and various kinds of detectors. Although there are many present applications for such devices, such as bar-code readers, optical disks, and surveying instruments, and other potential applications such as all-optical computers and solar cells, the biggest impact, both now and in the future, will probably be in the communications and digital processing areas. Thus, we will concentrate on these areas in our examples below.

5.3.1 Sources. Light sources involving GaAs include LEDs and lasers, both emitting at a wavelength of about 0.85 μm (i.e., close to the band gap; cf. Sec. 1.4). This wavelength was ideal for minimum signal loss and wavelength dispersion of some of the early silica fibers, but present fiber technology has moved the loss minimum to 1.3 or 1.5 μm, and the dispersion minima to nearly the same wavelengths. At any of these wavelengths, a choice must be made between low-cost multimode fiber or the more expensive single-mode fiber, which have bandwidth-length products of about 200 MHz km and 100 GHz km, respectively. Thus, for short link lengths or low bit rates, the multimode, more lossy fiber can be used with an 0.85-μm GaAs laser or LED, whereas for longer links and higher bit rates, single-mode fiber with minimum loss at 1.3 or 1.55 μm is used with InGaAs or InGaAsP sources lattice-matched to InP. The material for most of these devices is grown by epitaxial techniques, usually LPE or MOVPE, since defects leading to nonradiative recombination must be kept at a minimum.

A simple implementation of a double-heterostructure injection laser is shown in Fig. 28. Although the carriers recombine and emit light in the cross-hatched GaAs region, the AlGaAs (actually $Al_{0.3}Ga_{0.7}As$) regions serve two important purposes: confinement of the injected carriers due to the band-gap discontinuities, and confinement of the optical field due to the index-of-refraction discontinuity. The result of the AlGaAs cladding layers is a greatly reduced current threshold for lasing.

5.3.2 Detectors. Optoelectronic detectors are of two basic types: photoconductors in which the light increases the majority-carrier concentration (usually electrons) and thus the drift current; and photodiodes, in which the electrons and holes which are created in (or diffuse into) the junction region are swept by the junction electron field. A very common detector for both signal processing and the optical communications areas is the reverse-biased *p-i-n* photodiode, in which a high-resistivity drift region of width *w* is placed between the *p* and *n* regions. This scheme effectively increases the junction width and thus the amount of light absorbed. It also decreases the junction capacitance, which decreases the R_LC time constant due to the load; however, it increases the transit time across the junction, which will be the speed-limiting feature if *w* is made too large. A schematic of a planar photodiode is shown in Fig. 29. Here the *p* region (AlGaAs) has a higher conduction-

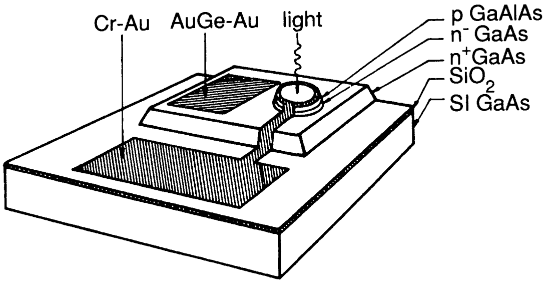

FIG. 29. Schematic of a planar photodiode. [From Bar-Chaim *et al.* (1983). Reproduced by permission of the American Institute of Physics.]

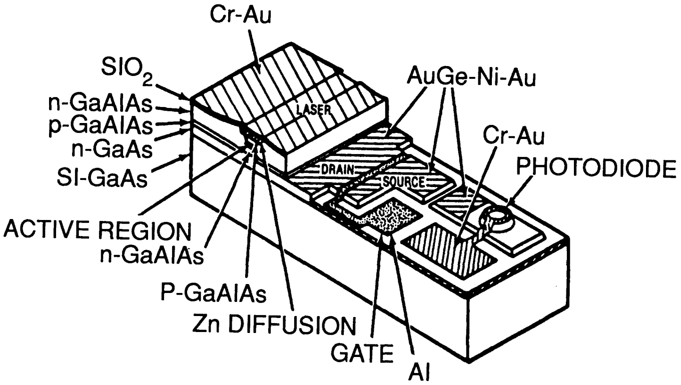

FIG. 30. Schematic of a monolithically integrated repeater circuit. (Courtesy A. Yariv).

band energy than the *n* region (GaAs) in order to "launch" the electrons into the drift space.

At very high reverse biases, the carriers in the drift region will pick up enough energy to create further electron-hole pairs by impact ionization. This process will "avalanche" and produce a very high optical gain. Avalanche detectors are used for applications in which noise is not a major consideration. A useful version of this detector has been designed around $Al_xGa_{1-x}As/GaAs$ superlattices.

For digital signal-processing applications, the fiber links are relatively short, so that the 0.85-μm minimum-loss fibers are not a limiting factor. Then, GaAs sources and detectors are ideally suited for most purposes. For optical communications, on the other hand, longer fibers are needed so that the higher-quality 1.3- and 1.55-μm minimum-loss fibers are normally used. In that case, the sources and detectors must be optimized for longer wavelengths, and are usually based on ternary materials, such as InGaAs, or quaternary materials, such as InGaAsP, lattice matched to InP. However, it must be noted that InP-based technology is not as mature, and it is also more difficult to grow ternaries and quaternaries than binaries.

5.3.3 Integrated Optics. Overall performance of a circuit can be greatly enhanced if several functions can be integrated on the same chip. In Fig. 30, we show a monolithically integrated repeater, consisting of a *p-i-n* detector, a MESFET, and a heterojunction laser diode. Here the *p*-AlGaAs layer serves as the window layer for the photodiode and as the current-confinement layer in the laser. Although this is a rather simple circuit, many OEICs (optoelectronic integrated circuits) nowadays have more than ten components, and it is likely that complexity will continue to increase in order to realize greater cost and speed advantages over hybrid circuits.

A good review of OEIC technology and applications has been given by Hirano (1987).

GLOSSARY

Acceptor: An impurity or defect in the crystal lattice that has the capability of taking an electron from the conduction band, the valence band, or a donor center (thus decreasing, increasing, or not changing the conductivity, respectively).

Donor: An impurity or defect in the crystal lattice that has the capability of donating an electron to the conduction band, the valence band, or an acceptor center.

Effective Mass: The mass that a free electron or hole in a crystal lattice "appears" to have when subjected to electric or magnetic forces.

Energy Gap: The energy difference between the highest point (in $\mathscr{E}$ vs **k** space) of the valence (filled) bands and the lowest point of the conduction (empty) bands.

Epitaxial Growth: Single-crystal growth on a substrate, usually of the same material, or material with a similar lattice constant.

Heterostructure: A crystal formed by the joining of two dissimilar materials, usually by epitaxial growth of one on the other.

Microwave: Referring to frequencies in the range 1 to 30 GHz.

MMIC (Monolithic Microwave Integrated Circuit): A circuit consisting of several microwave components, such as transistors, capacitors, and inductors, fabricated monolithically on the same chip.

Optoelectronic: Referring to the combining of optical components, such as lasers and detectors, and electronic components, such as oscillators and amplifiers, in the same circuit.

Phonons: Quantized lattice vibrations, having discrete, "particlelike" behavior.

Photons: Quantized light waves, having particlelike behavior.

Relaxation Time: The time required for randomization of electron or hole momentum through collisions (scattering). A different time is involved in energy relaxation.

List of Acronyms

BiCMOS	bipolar complementary metal/oxide/semiconductor
BFL	buffered FET logic
CB	conduction band
CMOS	complementary metal/oxide/semiconductor
DCFL	direct-coupled FET logic
GF	gradient freeze
HB	horizontal Bridgman
HBT	heterojunction bipolar transistor
HEMT	high electron mobility transistor
HFET	heterostructure field-effect transistor
HGF	horizontal gradient freeze
HP-LEC	high-pressure liquid-encapsulated Czochralski
IC	integrated circuit
IF	intermediate frequency
LEC	liquid-encapsulated Czochralski
LED	light-emitting diode
LO	longitudinal optical
LPE	liquid-phase epitaxy
LP-LEC	low-pressure liquid-encapsulated Czochralski
LSI	large-scale integration
LVM	local vibrational mode
MBE	molecular-beam epitaxy
MESFET	metal/semiconductor field-effect transistor
MIM	metal/insulator/metal
MMIC	monolothic microwave integrated circuit
MOCVD	metal-organic chemical vapor deposition
MODFET	modulation-doped field-effect transistor
MOMBE	metal-organic molecular beam epitaxy
MOS	metal/oxide/semiconductor
MOVPE	metal-organic vapor-phase epitaxy
NF	noise figure
NR	nonradiative
OEIC	optoelectronic integrated circuit
OMVPE	organometallic vapor-phase epitaxy
PBN	pyrolytic boron nitride
PL	photoluminescence
PR	photoreflectance
rf	radio frequency
RHEED	reflection high-energy electron diffraction
SDFL	Schottky diode FET logic
SDHT	selectively-doped heterostructure transistor
SI	semi-insulating
SRAM	static random access memory
TEGa	triethylgallium
TEGFET	two-dimensional electron gas field-effect transistor
TMGa	trimethylgallium
TO	transverse optical
VB	valence band
VB	vertical Bridgman
VGF	vertical gradient freeze
VLSI	very large scale integration
VPE	vapor-phase epitaxy

List of Works Cited

Abe, M., Mimura, T., Kobayashi, N., Suzuki, M., Kosugi, M., Nakayama, M., Odani, K., Hanyu, I. (1989), *IEEE Trans. Electron Devices* **36**, 2021–2031.

Arnold, J., Charlton, R. W. W., Turner, J. A. (1989), in: D. Haigh and J. Everard (Eds.), *GaAs Technology and Its Impact on Circuits and Systems*, London: Peter Peregrinus, Chap. 2.

Asbeck, P. M., Chang, M-C. F., Higgens, J. A., Sheng, N. H., Sullivan, G. J., Wang, K-C., *IEEE Trans. Electron Devices* **36**, 2032–2042.

AuCoin, T. R., Savage, R. O. (1985), in: D. K. Ferry (Ed.), *Gallium Arsenide Technology*, Indianapolis: Howard W. Sams, Chap. 2.

Bar-Chaim, N., Lau, K. Y., Ury, I., Yariv, A. (1983), *Appl. Phys. Lett.* **43**, 261–262.

Blakemore, J. S. (1982), *J. Appl. Phys.* **53**, R123-R181.

Bottka, N., Gaskill, D. K., Sillmon, R. S., Henry, R., Glosser, R. (1988), *J. Electron. Mater.* **17**, 161–170.

Chao, P. C., Tessmer, A. J., Duh, K.-H. G., Ho, P., Kao, M.-Y., Smith, P. M., Ballingall, J. M., Liu, S.-M. J., Jabra, A. A. (1990), *IEEE Electron Device Lett.* **11**, 59–62.

Chelikowsky, J. R., Cohen, M. L. (1976), *Phys. Rev. B* **14**, 556–582.

Chen, Y.-K., Nottenburg, R. N., Panish, M. B., Hamm, R. A., Humphrey, D. A. (1989), *IEEE Electron Device Lett.* **10**, 267–269.

Cochran, W., Fray, S. J., Johnson, F. A., Quarrington, J. E., Williams, N. (1961), *J. Appl. Phys.* **32**, 2102–2106.

Coleman, J. J., Dapkus, P. D. (1985), in: D. K. Ferry (Ed.), *Gallium Arsenide Technology*, Indianapolis: Howard W. Sams, Chap. 3.

Dapkus, P. D., Coleman, J. J. (1989), in: R. J. Malik (Ed.), *III-V Semiconductor Materials and Devices*, Amsterdam: North-Holland, Chap. 4.

Dolling, G., Cowley, R. A. (1966), *Proc. Phys. Soc. London* **88**, 463–494.

EMIS Group (1990), *Properties of Gallium Arsenide*, Bristol: Adam Hilger.

Feng, M., Lau, C. L., Eu, V., Ito, C. (1990), *Appl. Phys. Lett.* **57**, 1233–1235.

Flynn, B. W., Reekie, H. M., Mavor, J. (1989), in: D. Haigh and J. Everard (Eds.), *GaAs Technology and Its Impact on Circuits and Systems*, London: Peter Peregrinus, Chap. 6.

Goronkin, H., Grondin, R. O., Ferry, D. K. (1985), in: D. K. Ferry (Ed.), *Gallium Arsenide Technology*, Indianapolis: Howard W. Sams, Chap. 5.

Hirano, M. (1987), *Optoelectronics—Devices and Technology* **2**, 137–175.

Lau, C. L., Feng, M., Lepowski, T. R., Wang, G. W., Chang, Y., and Ito, C. (1989), *IEEE Electron Device Lett.* **10**, 409–411.

Look, D. C. (1989), *Electrical Characterization of GaAs Materials and Devices*, New York: Wiley.

Look, D. C., Colter, P. C. (1983), *Phys. Rev. B.* **28**, 1151–1153.

Martin, G. M. (1981), *Appl. Phys. Lett.* **39**, 747–748.

Pankove, J. I. (1975), *Optical Processes in Semiconductors*, New York: Dover.

Pfeiffer, L., West, K. W., Stormer, H. L., Baldwin, K. W. (1989), *Appl. Phys. Lett.* **55**, 1888–1890.

Pollak, F. H., Shen, H. (1990), *J. Electron. Mater.* **19**, 399–406.

Pucel, R. A. (1985), in: D. K. Ferry (Ed.), *Gallium Arsenide Technology*, Indianapolis: Howard W. Sams, Chap. 6.

Razeghi, M., Omnes, F., Nagle, J., Defour, M., Acher, O., Bove, P. (1989), *Appl. Phys. Lett.* **55**, 1677–1679.

Sakaki, H. (1989), in: R. J. Malik (Ed.), *III-V Semiconductor Materials and Devices*, Amsterdam: North-Holland, Chap. 5.

Shen, C. C., Hartman, D. H. (1985), in: D. K. Ferry (Ed.), *Gallium Arsenide Technology*, Indianapolis: Howard W. Sams, Chap.11.

Spitzer, W. G., Whelan, J. M. (1959), *Phys. Rev.* **114**, 59–63.

Stanley, C. R., Holland, M. C., Kean, A. H., Stanaway, M. B., Grimes, R. T., Chamberlain, J. M. (1991), *Appl. Phys. Lett.* **58**, 478–480.

Steiner, T., Zhang, Yu, Charbonneau, S., Villemaire, A., Thewalt, M. L. W., Maciaszek, M., Bult, R. P. (1989), *Can. J. Phys.* **67**, 242–250.

Sturge, M. D. (1962), *Phys. Rev.* **127**, 768–773.

Tan, K. L., Dia, R. M., Streit, D. C., Lin, T., Trinh, T. Q., Han, A. C., Liu, P. H., Chow, P.-M. D., Yen, H. C. (1990), *IEEE Electron Device Lett.* **11**, 585–587.

Theis, W. M., Bajaj, K. K., Litton, C. W., Spitzer, W. G. (1982), *Appl. Phys. Lett.* **41**, 70–72.

Tu, C. W., Hendel, R. H., Dingle, R. (1985), in: D. K. Ferry (Ed.), *Gallium Arsenide Technology*, Indianapolis: Howard W. Sams, Chap. 4.

Wolfe, C. M., Stillman, G. E., Lindley, W. T. (1970), *J. Appl. Phys.* **41**, 3088–3091.

Further Reading

Blakemore, J. S. (1982), "Semiconducting and Other Major Properties of Gallium Arsenide," *J. Appl. Phys.* **53**, R123–R181.

EMIS Group (1990), *Properties of Gallium Arsenide*, Bristol: Adam Hilger.

Ferry, D. K. (Ed.) (1985), *Gallium Arsenide Technology*, Indianapolis: Howard W. Sams.

Haigh, D., Everard, J. (Eds.) (1989), *GaAs Technology and Its Impact on Circuits and Systems*, London: Peter Peregrinus.

Look, D. C. (1989), *Electrical Characterization of GaAs Materials and Devices*, New York: Wiley.

Malik, R. J. (Ed.) (1989), *III-V Semiconductor Materials and Devices*, Amsterdam: North-Holland.

Milnes, A., Miner, C. J. (Eds.) (1990), *Semi-Insulating III-V Materials, Toronto, 1990*, Bristol: Adam Hilger.

Pankove, J. I. (1975), *Optical Properties in Semiconductors*, New York: Dover.

GAME THEORY

MARTIN SHUBIK, *Cowles Foundation for Research in Economics, Yale University, New Haven, Connecticut, U.S.A.*

Introduction 69
What is Game Theory? 69
A Historical Sketch 70
1. **The Languages of Game Theory** 70
1.1 The Extensive Form 71
1.2 The Strategic Form 72
1.3 The Coalitional Form 72
1.4 Other Structures 73
1.5 A Continuum of Agents 73
1.6 Incomplete Information 73
1.7 Common Knowledge 74
1.8 Utility Theory 74
2. **Solution Concepts** 74
2.1 The Zero-Sum Game and Maxmin 74
2.2 The Nonconstant Game and Noncooperative Equilibrium 75
2.3 Noncooperative Equilibrium and the Extensive Form 76
2.4 Solutions to Games with Incomplete Information 77
2.5 Cooperative Games 77
2.5.1 The Core 78
2.5.2 The Value 78
2.5.3 The Nucleolus 78
2.5.4 The Bargaining Set 79
2.5.5 The Stable Set 79
3. **Applications** 79
3.1 Economics 79
3.1.1 Competition among Few 80
3.1.2 Fair Division; Bargaining 80
3.1.3 Auctions 80
3.1.4 Market Games and General Equilibrium 81
3.1.5 Strategic Market Games and Money 81
3.1.6 Accounting 81
3.2 Political Science 81
3.2.1 Voting 81
3.2.2 Deterrence Theory 82
3.3 Biology 82
3.4 Other Applications 83
4. **Modeling, Gaming, and Empirical Validation** 83
4.1 Experimental Gaming 83
4.2 Towards Dynamics and Bounded Rationality 84
Glossary 84
Works Cited 85
Further Reading 88

INTRODUCTION

What is Game Theory?

The theory of games is devoted to the analysis of individual and group decision making in situations involving a diversity of individual goals and differing levels of strategic control over the environment. It provides a mathematical theory of strategy as an extension of individual rational behavior. It is complementary to social psychology and other behavioral theories whose models of man presume limited rationality, constraints to calculation, and nonconscious behavior. Game theory examines the implications of the interactions of individually rational agents in situations involving more than one agent. It illustrates the relationship between individual and group rationality. Several simple intuitively plausible concepts concerning the extension of individual rationality to group rationality are shown to be false. The theory of games offers explanations and insights as to why this is so.

By limiting the model of the individual to rational man, the theory of games helps to sort out those aspects of behavior that require the introduction of psychological, sociopsychological, or sociological factors from those that can be explained by pure economic factors.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

A Historical Sketch

Although several of the basic ideas concerning formal strategic analysis could be attributed as far back as the great classic of Sun Tzu (ca. 500 B.C.), modern game theory can be dated from the 1920s with the work of von Neumann (1928) on the two-person zero-sum game or game of pure opposition. Games such as chess, Go, or checkers are in this class. The gain to one player is attained only at the cost of a loss to the other. Although Borel (1924) had considered two-person zero-sum games, and considerably earlier Waldgrave had examined a simple game called "Le Her," the proof of the basic solution for a two-person zero-sum game—the minimax theorem (discussed in Sec. 2.1)—was von Neumann's seminal contribution.

Prior to von Neumann, Zermelo (1913) had proved (by backward induction) the first formal theorem in what is now known as game theory. However, this was an isolated result rather than the formal start of a new theory.

The result of Zermelo establishes that the outcome in a game of chess is strictly determined. Chess is a game of *perfect information*, which means that at every point in the game each individual knows precisely what has transpired (this contrasts with a game such as bridge or poker where the deal is a random move not disclosed to all). Starting at any end point of the game, as it is played sequentially the player who is called upon to move faces a one-person maximization problem if he selects the option to terminate the game with a win or draw. Thus we can replace the original game with a smaller game by removing the last set of choices and replacing them by the value of the optimal selection of the player who is called upon to move. This procedure can be repeated all the way back until the outcome of the game is determined at the start. To this day it has not been determined if chess is a win, loss, or draw for white, but we know that if people or machines could compute fast enough it would not be worth playing, like tic-tac-toe which is also a two-person zero-sum game with perfect information.

The work that launched game theory as a major discipline was the book *The Theory of Games and Economic Behavior* by von Neumann and Morgenstern (1944). Cournot (1897) and Edgeworth (1881) had recognized the importance of economic interaction among few agents, but Morgenstern was the first modern economist to realize the possibility of the construction of a general theory for the study of cross-purposes maximization. Morgenstern's collaboration with von Neumann resulted in the publication of the large and impressive book that introduced game theory as a scientific discipline of its own.

This book contained four logically separate contributions that the authors linked together. They were utility theory (see Sec. 1.8); the extensive form of a game (see Sec. 1.1); the theory of two-person zero-sum games (see Sec. 2.1); and a theory of cooperative games (see Sec. 1.3) based on a solution called the *stable set* (see Sec. 2.5.5), which reflects a form of social stability.

The 1950s saw the takeoff in the development of the mathematical aspects of theory of games and some applications. Nash (1951) defined the noncooperative equilibrium (see Secs. 2.2 and 2.3) and proved its existence for a broad class of games. Gillies (1953) and Shapley considered the core (see Sec. 2.5.1); Shapley (1953a) developed the value. Nash (1950,1953) wrote two seminal papers on two-person bargaining and fair division (see Sec. 3.1.2). Shapley (1953b) wrote a fundamental paper on stochastic games (see Sec. 1.4). Among the early applications Mayberry *et al.* (1953) applied Nash's results to duopoly and Shapley and Shubik (1954) applied the value to measure power in committee voting. As early as 1951 Nyblen applied game theory to macroeconomics and Harsanyi (1956) connected the work of Zeuthen (1930) with that of Nash in his early work on the development of a no-side-payment value (see Sec. 2.5.2). By the end of the 1950s the roots of game theory had been well established.

1. THE LANGUAGES OF GAME THEORY

A key contribution of the theory of games has been to provide several basically different abstractions together with special mathematical notation to study various aspects of multiperson optimization at various levels of detail. Each abstraction together with the mathematical notation used provides a language of discourse for a broad set of problems.

The three major representations of a game have been the extensive, the strategic (or normal), and the coalitional forms. The first provides microscopic detail concerning the moves, information conditions, and payoffs of the game; the strategic form suppresses much of this detail and describes a game only in terms of strategies and payoffs (see Sec. 1.2 below); the coalitional form goes to an even higher level of aggregation. The concern is with the joint gains of coalitions; thus there is no direct interest in details even about strategies. In this representation the game is portrayed by coalitions and the amounts they can achieve.

1.1 The Extensive Form

The basic description of a game in full process detail is provided by the *extensive form*. First developed by von Neumann and generalized by Kuhn (1953) for *n*-person games, the extensive form provides a precise language to describe choices, moves, information, and payoffs for any finite game.

Two simple examples of a game tree serve to illustrate the extensive form. A game in extensive form is represented by a rooted tree, which means that the tree has a distinguished node indicating where the game begins. In Figs. 1(a) and 1(b) the node is identified with the letter O. Each node of the game tree, except the terminal nodes, represents a choice point for one of the players (or Nature or chance, considered as an extra player, who when called upon to choose selects a move according to some given probability distribution). A move is represented by a branch of the tree. At each choice point the moves available to the player are indexed. In Figs. 1(a) and 1(b) player 1 selects his move first. His choice is to go left or right. These are indexed by 1 or 2. All choice points for each player are partitioned into information sets. These sets describe the state of insurance under which a player operates.

The game in Fig. 1(a) has perfect information. This means that at any point in the game the player whose choice it is knows exactly where he is. All his information sets contain only one element apiece. The game in Fig. 1(b) does not have perfect information. The information set for player 2 contains two nodes. This means that when player 2 selects his move, he does not know whether player 1 has selected left or right. Once the sequence of moves has reached a terminal point in the game tree the game is over and the numbers in parentheses at the bottom of each terminal node specify the payoffs to each player. The pair of numbers to the far left in Fig. 1 indicate that for the play of the game where each selects his first move the payoff is 1 to player 1 and −1 to player 2.

A (pure) strategy for a game in extensive form is a complete plan that tells the player how to select his move as a function of his current information at every point in the game. Given a strategy for each player these uniquely determine a path down the whole game tree and hence a play and an outcome.

A game of any complexity, such as chess, has myriads of strategies, many of which are redundant. But although it is not practical to enumerate all of the strategies in games involving more than three or four moves, the concept of a strategy as a complete contingent plan enables us to define precisely how contingent planning depends on information. This can be illustrated with the two simple examples. The game shown in Fig. 1(a) has two strategies for player 1. His moves and strategies coincide. He can select 1 or 2. Player 2, however, has four strategies because his choice

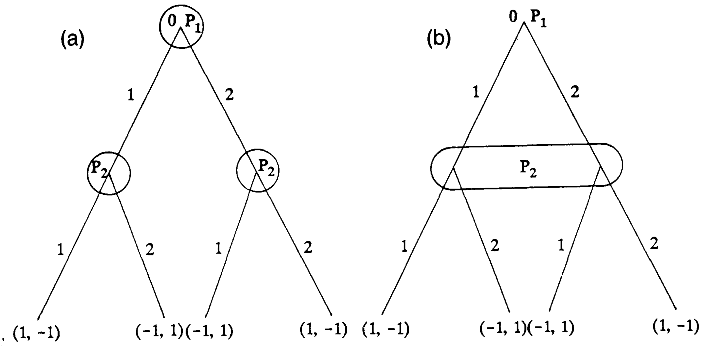

FIG. 1. Representations of a game in extensive form. (a) The case of perfect information. (b) The case of imperfect information.

of move is based upon contingent planning, which tells him what to do if player 1 selects 1 and what to do if player 1 selects 2. In each instance he can select one of two moves. In contrast in the game shown in Fig 1(b), when player 2 is called upon to select his move, he does not know what player 1's choice has been, and hence his choice is limited to selecting move 1 or 2 without conditioning the selection on what player 1 has done.

A mixed strategy for a player is a selection among his pure strategies according to some random distribution. We may regard each pure strategy as a complete book of instructions that one could give to an agent acting on one's behalf. A mixed strategy involves randomizing to decide which pure strategy to use.

1.2 The Strategic Form

The strategic (sometimes called the normal) form of a game is defined in terms of the players, their strategy sets, and their payoffs. Suppose that each player i has a set of strategies S_i, where we denote a specific strategy by s_i. Associated with each player i is a payoff function $P_i(s_1,s_2,\ldots,s_n)$. This describes the game in strategic form.

A two-player game can be illustrated simply in matrix form, or, more precisely, in bimatrix notation, i.e., in each cell there are two numbers, the first indicating the payoff to player 1 and the second to player 2. The two games illustrated in extensive form in Figs. 1(a) and 1(b) can be illustrated in strategic form in Tables 1(b) and 1(a).

The game with perfect information is displayed as a 2×4 matrix, as player 1 has two strategies (which coincide with his moves) but player 2 has four strategies as he has information which he can utilize for contingent planning. The numbers above the matrix are merely the names of the strategies and can be regarded as standing for the labels to four books of instructions.

Table 1. (a) Matching pennies. (b) As in (a), but with complete information.

(a)

	1	2
1	1	−1
2	−1	1

(b)

	1	2	3	4
1	1	1	−1	−1
2	−1	1	−1	1

1=(1,1;2,1), 2=(1,1;2,2), 3=(1,2;2,1), 4=(1,2;2,2)

1.3 The Coalitional Form

The coalitional form of a game is represented by its *characteristic function* if the game has transferable utility (referred to as TU). This means that when any group of players collaborate, whatever they earn, they are able to make transfers among themselves so that without loss of value they can transfer their joint gain among themselves in any way they choose. The characteristic function of an n-person game is a superadditive set function $v(S)$ defined on all of the 2^n-coalitions that can be formed from n players, where n is the number of players in the set N of all players. An additional economic assumption is that the set n is superadditive. This can be stated formally as

$$v(S \cup T) \geqslant v(S) + v(T). \tag{1}$$

The economic intuition behind this condition is that a coalition will not form unless there is the possibility of joint gain to the parties involved. A simple example of the characteristic function of a three-person game in coalitional form is shown below:

$$\begin{gathered} v(\{1\})=0,\quad v(\{2\})=0,\quad v(\{3\})=0, \\ v(\{1,2\})=1,\quad v(\{1,3\})=2,\quad v(\{2,3\})=3, \\ v(\{1,2,3\})=4. \end{gathered} \tag{2}$$

This is interpreted as follows: $v(\{i,j\})$ stands for the amount that can be obtained by the set of players whose names are i and j. Here we observe that any player by acting alone obtains nothing and that all three together obtain 4. An examination of the two-person coalitions shows that player 3 brings the most to any coalition of two.

Implicit in representation of the payoffs to any coalition by a single number is the idea that the coalition can divide the proceeds among themselves in any way the members see fit. In many situations the assumption that there is some perfect means of side payment among all players is not justified. The players may obtain payoffs as a result of the play of the game and may not be in a position to make side payments or adjustments to what each achieves. A game where there are no side payments is referred to as an NTU game (no transferable utility). It is represented by a characterizing function $V(S)$, where it is no longer possible to assign a single number to indicate the amount that a coalition S obtains, but instead a set of feasible efficient outcomes specifying what the coalition can achieve must be given. The analysis of NTU games is considerably more complex than TU games, but the NTU games have been used extensively in economic theory.

1.4 Other Structures

The notations developed in Secs. 1.1, 1.2, and 1.3 all involve a finite number of agents playing in a game with a given start and a finite end with the payoffs at the end. Many aspects of human cooperation and competition are not well modeled in this form. In general, for many situations there has been a yesterday, and history may matter. Furthermore, there may be a tomorrow, and the time at which the game ends may be indefinite.

In other instances the game under consideration might best be modeled with a continuous set of strategies, such as in a duel where two individuals are walking towards each other and each has to decide when to fire.

Another important class of games that are modeled naturally with continuous moves consists of games of pursuit, leading to differential games. The first book on differential games was by Isaacs (1965). Case (1979) applied differential games to some problems in economic competition. There is also a growing literature on other dynamic games; these include repeated games, stochastic games, games of survival, games of economic survival, and recursive games.

1.5 A Continuum of Agents

The original development of the theory of games by von Neumann and Morgenstern dealt with games with a finite set N of players. In applications it became desirable to consider both games with an infinite but countable set of players and games with a continuum of players. When studying mass economic and mass voting behavior, the modeling of these games is called for to examine the various forms of limiting behavior that might be encountered, as action in the market or at the ballot box involves millions of more or less anonymous individuals. One can consider an exchange economy with a baker, a butcher, and a candlestick maker, and then compare it with another economy where there are k individuals of each type, where we allow k to become arbitrarily large. A natural question to ask is, if we assume some form of behavior for the individual agent, does the behavior of the mass of agents approach a limit as k becomes large? As long as k is finite, although the power or possession of an individual in a replicated game shrinks relative to the game as a whole, nevertheless the player remains of measurable size. Shubik (1959a) considered the replications of a game, basing the idea on an interpretation of earlier work of Cournot and Edgeworth; Debreu and Scarf (1963) used replication in the study of the behavior of the core (Sec. 2.5.1).

The way we can capture the intuitive idea of an individual who is so small with respect to the market that he has no influence whatsoever on price is to consider a continuum of traders where each is of measure zero. Milnor and Shapley (1957) first considered *oceanic games* in the study of voting with a continuum of voters. Aumann (1964) considered cooperative markets with a continuum of traders. Dubey and Shapley (1980) considered games in strategic form and the noncooperative equilibrium solution with a continuum of traders. The idea of a continuum of economic agents plays a central role in understanding the concept of competition in an enterprise economy. It has enabled economists to formalize the remarks of Adam Smith on the "invisible hand."

1.6 Incomplete Information

The game-theory models described in Secs. 1.1–1.5 all have as an implicit assumption that the players know each other's preferences and strategy sets as part of the rules of the game. For example, although in chess it is reasonable to assume that each player knows the other's preferences concerning the outcome of the game, it is not likely that in a many-person trade each knows the valuations placed on all goods by the others. Harsanyi (1967,1968a,1968b) provided a means to take this lack of knowledge into account.

An n-player game with incomplete information in strategic form is described as follows. There are n players and each player i may be one of t_i types. There are n probability distributions $(p_1, p_2, \ldots, p_n)$ that can be used to determine each player type. At the start of the game all player types are determined and each player is informed privately of his type. The game is then played simultaneously by all the player types. Each player obtains a payoff that is determined by both the choice of all strategies and the selection of his type.

The Harsanyi formulation presents a way to model both incomplete information and beliefs about beliefs. Mertens and Zamir (1985) established that there is an equivalence between the Harsanyi formulation and consistent beliefs.

1.7 Common Knowledge

Although Harsanyi succeeded in modeling incomplete information in a situation in which all agents share a common prior, a "prior" being the *a priori* probability for the occurrence of an expected event, it is possible that knowledge about who knows what is not shared. Aumann (1976) has phrased it that a fact is common knowledge among the players if every player knows it, every player knows that every player knows it, and so on.

The importance of the role of common knowledge is illustrated in an example noted by Myerson (1991). A village has 100 couples. The men gather each day to praise their wives if they believe them to be faithful or to curse them if they believe otherwise. If a wife is unfaithful she and her lover tell all men in the village except her husband. In this village every wife has been unfaithful to her husband but each husband lives in the bliss of having no evidence about his wife. A visiting guru announces to all that there is an unfaithful wife in the village. One hundred days after the guru's visit all husbands curse their wives.

1.8 Utility Theory

Utility theory deals with the nature of the description of human preferences. A common error made in the discussion of game theory is that somehow the development of the theory of games depends upon a precise and unrealistic specification of measurable and comparable utility among all individuals—as though some soma pill or exchangeable utility existed. In fact, all that is needed for the logical basis of a theory of games is that each individual has some rule of choice when confronted with a set of alternatives. There is an enormous literature on preferences and utility [see, for example, Fishburn (1970)] that is complementary but incidental to many of the basic ideas of the theory of games.

2. SOLUTION CONCEPTS

2.1 The Zero-Sum Game and Maxmin

The seminal contribution of von Neumann was to show that for the class of two-person zero-sum games under reasonable conditions, the concept of rational behavior can be extended by means of the concept of individual rationality. The *security level* of player i is the greatest amount that he can guarantee for himself under all circumstances. This is the amount maxmin $P_i(s_1, s_2)$, where P_i is the payoff for player i, the s_i are the strategies of the players, and the term *maxmin* refers to the maximum that individual i can obtain on the assumption that j will select a strategy that minimizes i's payoff. An outcome to a game is called individually rational if each player obtains at least his security level. A game in which there is a single individually rational outcome is called *strictly determined.* An example in Table 2(a) shows a strictly determined two-person zero-sum game. The example in Table 2(b) shows the simple matching pennies game, which is not strictly determined.

Table 2. A game (a) with and (b) without a saddle point.

(a)

	1	2	min	
1	2	0	0	
2	3	1	1	max
max	3	1		
		min		

(b)

	1	2	min	
1	1	−1	−1	
2	−1	1	−1	max
max	1	1		
		min		

In Table 2(a), as the game is zero sum we do not need to place two entries in each cell; given the payoff to player 1 we know that the payoff to player 2 is the negative of this value. The matrix is bordered on the right by the values of the row minima, and it is bordered below by the column maxima. The row minima indicate the payoffs that player 1 can attain by playing strategies 1 or 2, assuming that player 2 minimizes player 1's payoff. The maximum entry in this column represents the best outcome that player 1 can guarantee for himself. There is a similar interpretation for the column maxima. The security level for player 1 is the maximum of the row minima, and the security level for player 2 is the minimum of the column maxima. In this instance we see that they coincide. A game that has this property is said to have a *saddle point* in pure strategies. A three-dimensional graph of the payoff surface to player 1 as a function of the two strategies would show the saddle point as the lowest mountain pass, i.e., it is shaped like a saddle. If both individuals are individually rational, they will each obtain the saddle-point value with no need to communicate except through their actions or selection of strategies.

The game shown in Table 2(b) is the strategic form of the game shown in Fig. 1(b); it can be interpreted as the familiar game of matching pennies. If player 1 matches player 2 he wins 1; if he fails to match he loses 1. When we border this matrix with the row minima and the column maxima, there is no coincidence of the maxmin with the minmax. von Neumann's theorem established that if we permit mixed strategies, i.e., the players are permitted to randomize over their pure strategies, then there will always be a coincidence of the maxmin of the expected values with the minmax of the expected values of the payoffs. In terms of this simple example, the way the game of matching pennies is usually played is that both sides toss their coins, i.e., employ random strategies simultaneously. Each player obtains an expected value of $\frac{1}{2}(1)+\frac{1}{2}(-1)=0$. This is the value to each in this symmetric, fair game.

The minimax theorem has had considerable influence outside of game theory alone. There is an intimate relationship between it and the duality theory in linear programming, where the mixed strategies have an interpretation in terms of shadow prices.

2.2 The Nonconstant Game and Noncooperative Equilibrium

The dominant solution concept for an n-person game in strategic form is the noncooperative equilibrium solution presented by Nash (1950,1951) but given over 100 years earlier in a more specialized form by Cournot.

Consider a set N of players with index $i=1,\ldots,n$. Player i selects a strategy s_i from his set S_i of pure strategies. $s=(s_1,s_2,\ldots,s_n)$ is a vector of strategies with one strategy for each player. We denote by s/s_i^* the set of strategies s with the ith element replaced by s_i^*.

The payoff to an individual i is a function of his strategy and the strategies of all others. We denote this as $P_i(s)$. A pure strategy noncooperative equilibrium is a vector of strategies $s^*=(s_1^*,s_2^*,\ldots,s_n^*)$ such that for each $i \in N$ and $s_i \in S_i$, s_i^* is an optimal response. Given the strategies of all other players no individual can improve his payoff by selecting a strategy s_i differing from s_i^*. This definition can be extended to include the use of mixed strategies by the players, where each selects a probability distribution over his set of pure strategies.

The basic idea behind the Nash equilibrium is one of self-fulfilling expectations. If each believes that the others are going to behave in certain ways, there will exist a mutually consistent set of expectations such that if each optimizes given his expectations, then the predictions of all will be fulfilled. Unfortunately, the equilibrium is a static concept, and little of any generality is known of the dynamics that might lead to a noncooperative equilibrium.

From the viewpoint of prediction it would be desirable if a game had a unique equilibrium point. Unfortunately this is often not so. Furthermore, in general, noncooperative equilibria do not lead to efficient outcomes [see Dubey (1986) and Dubey and Rogawsi (1990)]. Tables 1(a) and 1(b) provide examples of two simple 2×2 bimatrix games, where the first number in each cell is the payoff to the first player and the second number the payoff to the second.

In Table 3(a) there are two equilibrium points, one at (1,1) and another at (2,2). The first favors the first player and the second the second. They both face a problem in coordination: if each tries for the advantage they could easily both obtain nothing. Aumann

Table 3. (a) Battle of the sexes; (b) Prisoner's Dilemma.

(a)

	1	2
1	2,1	0,0
2	0,0	1,2

(b)

	1	2
1	5,5	−5,10
2	10,−5	0,0

(1987) has extended the idea of a noncooperative equilibrium to include the concept of a correlated equilibrium, where the individuals are willing to randomize jointly. Thus in this game they will flip to choose between strategies (1,1) and (2,2). There is a mixed-strategy noncooperative equilibrium point where each individual randomizes independently with probabilities for player 1 of $(\frac{2}{3},\frac{1}{3})$ and for player 2 $(\frac{2}{3},\frac{1}{3})$. The independent randomization results in an expected payoff to both of $(\frac{2}{3},\frac{2}{3})$, whereas the joint randomization gives them an expected payoff of $(\frac{3}{2},\frac{3}{2})$.

Table 3(b) illustrates a well-known game called the Prisoner's Dilemma. It provides a striking counterexample to the proposition of vector domination as a decision rule for the single individual. In one-person decision making, if an individual has several strategies and he faces several different states of Nature, then if one strategy has payoffs that dominate the payoffs obtained from another strategy under every state of Nature, it is both intuitively obvious and true that the first strategy will be selected over the second and will yield the player a higher payoff. This attractive rule does not generalize to two-person nonzero-sum games. We observe, if we examine just the payoffs to player 1 alone, that his second strategy dominates the first. He can observe that if player 2 selects his first (second) strategy player 1 obtains 10 (0); if instead player 1 uses his first strategy then he would obtain 5 (−5). But 10 is larger than 5 and 0 is larger than −5. If both players use this reasoning they both select their second strategies and end up with payoffs of (0,0), yet if they were able to collaborate they both could use their first strategy and obtain (5,5), which is strictly better for both. There has been considerable experimentation with this game and in the one-shot versions individuals do tend to select their second strategy. This particular game is utilized to show that individually rational behavior does not necessarily coincide with group rational behavior.

Another desirable property of a solution might be stability. Unfortunately, even games with unique equilibrium points do not necessarily have stability. Consider the game shown in Table 4. This 3×3 game has a unique mixed-strategy equilibrium point with a mix of $(\frac{1}{3},\frac{1}{3},\frac{1}{3})$, which yields an expected value of payoff to each player of 3. But by observing the matrix we see that if the players just blundered about the matrix sequentially using optimum response their expected value would be 4.5. Starting at any of the cells with a positive-entry optimum response leads to a six-cycle.

2.3 Noncooperative Equilibrium and the Extensive Form

Many of the applications of the noncooperative equilibrium solution have been to games in strategic form. One can also consider using this solution to a game in extensive form. When the extra detail of the extensive form is considered, several new and important problems that are obscured by the shorthand of the strategic form appear. In particular, Selten (1975) introduced the concept of a perfect equilibrium point. A perfect equilibrium has the property that it is an equilibrium not only in the game as a whole, but in every subgame. This can nontechnically be best illustrated with a simple example. Figure 2 shows a game in extensive form. It is a game with perfect information. At all points in the game the players know exactly where they are. There is no concealment of information. After player 1 has made his first choice, if he selects his second move then player 2 will be

Table 4. A game with a unique unstable equilibrium.

	1	2	3
1	4,5	5,4	0,0
2	5,4	0,0	4,5
3	0,0	4,5	5,4

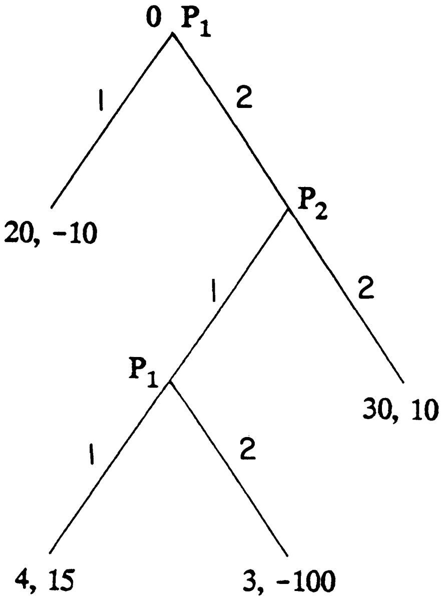

FIG. 2. A game with a perfect equilibrium and a threat.

called upon to choose his move. We could consider the rest of the tree as a separate game with its origin at player 2's choice point. This idea amounts to making an assumption that history does not matter. At any point in the game it is only the future that counts. This game has two equilibrium points. One is a perfect equilibrium and the other involves a threat by player 1 in which the history of the previous selection of moves matters. The equilibrium point in which player 1 selects move 1 and player 2 is prepared to select his move 1 is the perfect equilibrium and has a payoffs of (20, − 10). This can be calculated by backward induction. Suppose that the game had reached the second choice point for player 1; regarding this as a separate game, player 1 should choose move 1 and obtain 4. We can replace his choice point by the pair of payoffs (4,15). Player 2 is now faced with his choice. He will select his move 1 rather than move 2 as his payoff of 15 is better than a payoff of 10. We can replace his choice point by the payoff pair (4,15). Player 1 can now consider the game where by selecting 1 he obtains 20, or by selecting 2 he obtains 4. He selects 1 and the payoffs at this perfect equilibrium point are (10, − 10). A glance at the game tree shows that there is an outcome at which both players can improve. If both select their move 2 they obtain (30,10). This is also an equilibrium outcome, but its stability is maintained only by a threat strategy of the following variety. Player 1 announces "I will choose 2; if player 2 chooses 1, then I will choose 2 as my second choice." This can be interpreted as a threat in the sense that if player 1 were at his second choice point, his pure maximization of payoff, forgetting history, is to select move 1 and obtain 4; by selecting move 2 he obtains 3 but causes a loss of 100 to player 2. There is one viewpoint that individuals will not carry out unenforceable threats unless it is in one's immediate self-interest. Another viewpoint is that it is natural even in a one-play situation for player 1 to offer cooperation and to be prepared to back his words with his deeds if player 2 fails to cooperate.

Van Damme (1987) offers an excellent coverage of the many modifications to and variations of equilibrium-point theory for games in extensive form. Harsanyi and Selten (1987) present a monumental but controversial work arguing for a one-point equilibrium solution for all games.

2.4 Solutions to Games with Incomplete Information

The Harsanyi formulation of games with incomplete information permits an extension of the noncooperative solution to these games with the understanding of Bayesian rationality among the players. All players have personal probabilities about the types of player each other player might be, but their beliefs are all *consistent* in the sense that all personal probability distributions can be derived from a common *a priori* distribution, i.e., it is as though Nature randomizes from a known given set of *n* random variables and discloses what type each player is only to that player by revealing the value of his random variable.

Myerson (1991), in a recent book, provides a coverage of many of the more recent investigations of games with incomplete information.

2.5 Cooperative Games

The solutions to games in cooperative or coalitional form deal with concepts relevant to dividing joint gains among individuals who have agreed to act in concert to achieve an optimal product. There are several well-known solutions, each of which approaches this problem from a different point of view. Briefly, the core illustrates countervailing power; the value, fair division; the nucleolus, the minimization of the maximal amount that any coalition can claim; the bargaining set, a form of countervailing power between players; and the stable set, a form of social stability.

2.5.1 The Core. An *imputation* in an n-person cooperative (TU) game is a vector x of dimension n where each component x_i is the amount obtained by individual i. The sum of all x's adds up to $v(N)$, i.e., it exhausts the joint product. An imputation x dominates an imputation y via a coalition S if $x_i > y_i$ for all i in S. The *core* of an n-person game is the set of all undominated imputations. The core for the example of a three-person game in characteristic-function form shown in Sec. 1.3 is illustrated in Fig. 3. The corners of the triangle have coordinates (4,0,0), (0,4,0), and (0,0,4). Any point (a,a_2,a_3) in the triangle has coordinates that add to 4. Any point in the trapezoid delineated by $ACDE$ has the property that the sum of any two coordinates yields to a coalition of any two players more than they can obtain by themselves. This area is the core of the three-person game represented. The core need not always exist. For example, if the coalitions {1,2} and {1,3} each were able to obtain 3, the new game would have no imputation, giving all coalitions more than they could get by independent action. An interpretation of the core is that it is a set of imputations for which both societal and subgroup optimality is simultaneously satisfied.

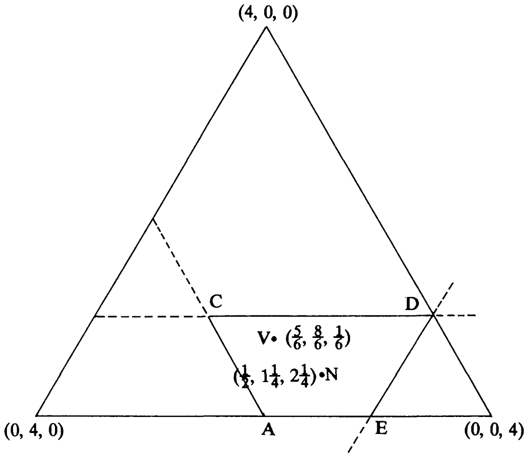
(4, 0, 0)
C
D
V• $(\frac{5}{6}, \frac{8}{6}, \frac{1}{6})$
$(\frac{1}{2}, 1\frac{1}{4}, 2\frac{1}{4})$•N
(0, 4, 0)
A
E
(0, 0, 4)

FIG. 3. A three-person game in characteristic-function form. The trapezoid *ACDE* is the core; *V* is the value; *N* the nucleolus.

There has been considerable application of core theory to economics [see Hildenbrand, (1982)]. The connection between the core and superadditivity, which is at the basis of economic cooperation, has been studied by Wooders (1983).

2.5.2 The Value. The value of an n-person game in characteristic function form is a single-point solution arrived at by considering the combinatorics of all the ways that an individual could join a coalition of any size, calculating the marginal worth of his addition to the coalition, and then averaging over all of the contributions. It attaches a value or expected worth to every player.

Remarkably, Shapley (1953b) was able to derive the value from essentially four axioms:

1. Symmetry: If two players are identical (i.e., they make the same contribution to any coalition one or the other joins), they each obtain the same value.
2. Dummy player: If a player contributes nothing to the worth of every coalition he joins, his value is zero.
3. Efficiency: The sum of the values assigned to all players exhausts the available gain achieved by the coalition of all, i.e., it equals $v(N)$.
4. Additivity: If two separate games are considered jointly as though they were a single game, the sum of the values of the two games will equal the values of the new single game.

The specific formula derived by Shapley is as follows:

$$\varphi = \sum_{S \subset N} \sum_{i \in s} \frac{(n-s)!(s-1)!}{n!} [v(S) - v(S/i)], \tag{3}$$

where the term in the square brackets is the marginal contribution of player i to coalition S, and s is the number of players in S.

Aumann and Shapley (1974) in an important book have extended the analysis of the value from games with a finite number of players to games with a continuum of players.

2.5.3 The Nucleolus. The *excess* of a coalition S at imputation a is $V(S) - a(S)$, i.e., it is a measure of how much the coalition is giving up by accepting the imputation a rather than acting independently. The *nucleolus*, formulated by Schmeidler (1969), is that point which minimizes the maximum excess among all of the 2^n excesses for all coalitions. Intuitively, the excess is a measure of the dissatisfaction that a coalition might have for a specific imputation. The selection of the nucleolus minimizes the maximum dissatisfaction among all coalitions. Sobolev (1975) has provided axioms for the nucleolus. When the core is nonempty the nucleolus lies within it. In Fig. 3 it is denoted by N.

2.5.4 The Bargaining Set. A bargaining point is an imputation with the property that for each pair of players i,j any objection that might be raised by i against j can be met by a counter objection by j against i. An objection consists of a coalition S containing i but not j and an imputation feasible for S that is preferred to the given imputation by every member of S. A counter objection consists of another coalition T containing j but not i and an imputation that is (weakly) preferred to the objection of every member of T also in S and is (weakly) preferred to the original imputation by every member of T not in S. The bargaining set consists of those imputations such that to every objection there is a counter objection [see Aumann and Maschler (1964)]. The bargaining set exists in many minor variants all of which include the core. Peleg (1967) has proved that it is never empty (for side-payment games).

2.5.5 The Stable Set. The original solution for a game in coalitional form proposed by von Neumann and Morgenstern was the stable set. Formally, an imputation a dominates an imputation b if there is a coalition S such that all members of S are better off with b than with a and the coalition can enforce b. The core consists of all imputations none of which is dominated by another imputation. It has internal stability; no point in the core dominates any other. The stable set has this property, but it also has the extra property that together all points in a stable set are able to dominate any point outside. This is the property of external stability. Consider the three-person simple-majority voting game where any coalition of two or more obtains the prize of say, 1, while others obtain 0. The unique symmetric stable set consists of the three imputations $(\frac{1}{2},\frac{1}{2},0)$, $(\frac{1}{2},0,\frac{1}{2})$, and $(0,\frac{1}{2},\frac{1}{2})$. There are also other nonsymmetric stable sets to this game that appear to illustrate discriminatory behavior by two of the players against the third. von Neumann and Morgenstern stressed that the nonuniqueness of the stable sets and the complexity of the domination relationship reflected the many possible social structures of a society. The case for the stable set would be much stronger if it always was nonempty, as von Neumann conjectured. Unfortunately Lucas (1969) was able to show by means of a 10-person example that this was false. Shapley and Shubik (1969a) showed that the Lucas counterexample was by no means pathological, but could occur in an exchange economy.

There are several other solution concepts such as the kernel [Davis and Maschler (1965); Peleg (1987)] and variants of the core [see Shubik (1982)].

3. APPLICATIONS

The nature of the applications of game theory has ranged from qualitative advice giving and the construction of simple illustrative models (such as the Prisoners' Dilemma) in strategic analysis, through serving as an aid in theory construction in economics, political science, biology, and other disciplines, to the applied mathematical calculations of specific models in dueling theory and weapons evaluation [see, for example, Drescher (1961); Shubik (1983a)]. The last 30 years have seen considerable growth in the applications of game theory to economics, political science, biology, law, accounting, military science, ethics, and sociology, but little to engineering or physics, although this may change. Much attention has been given to mass behavior, but a key distinction between mass-people behavior and mass-particle behavior appears to be the different type of communication and information flows between individual particles or people. Neutrons do not appear to talk to each other; people do. However, polarization might be regarded as an analog to correlation of behavior caused by communication.

3.1 Economics

The major applications of game theory have been to economics. The monumental book of von Neumann and Morgenstern stressed the investigation of economic behavior. Furthermore, as the subject developed, the relationship between the early mathematical economics of Cournot and Edgeworth was recognized.

3.1.1 Competition among Few. One of the earliest problems in mathematical economics was the treatment of competition among the few. The extension of analysis from monopoly with one maximizing agent to duopoly with competition between two maximizing agents whose fates were intertwined represents a considerable advance. Cournot's analysis was clearly a special case of the noncooperative equilibrium solution put forward in full generality by Nash over a century later.

From the 1950s onward there has been a steady explosion in microeconomic applications to the study of oligopolistic competition and industrial organization. Unfortunately, economic dynamics gives no indication of being any easier than a three-, four-, or five-body problem in physics. Even limiting concern to competition between two firms, the dynamics depends delicately on the relative sizes of the firms, the choice of strategic variables, the sequencing of moves, and the state of information. Cournot and Bertrand (1883) both discussed simultaneous-move games, but one used quantity and the other price as the strategic variable. In contrast, Stackelberg (1952) considered sequential moves.

Even with the difficulties in the study of dynamics, considerable strides have been made in the study of competition among the few as is evinced by the books by Shubik (1959b), Friedman (1990), Tirole (1988), and others.

3.1.2 Fair Division; Bargaining. Underlying much economic thought is the concept of fair division. Is there such a thing as a just and equitable way to slice the economic pie? If there is, what is its relationship to the price system and a division arrived at competitively?

There is an extensive literature on bargaining and fair division between two parties. Zeuthen (1930) was an early precursor of the game-theoretic approach. The two basic papers by Nash (1950, 1953) were followed by the contributions of Shapley (1953b) and Harsanyi (1956,1963), extending the investigations of bargaining from two individuals to any number, covering both TU and NTU instances, and considering the possibility and the plausibility of the use of threats in bargaining. The axiomatic method has been used extensively, and Roth (1979) provides a detailed discussion of many of the variations in the axioms proposed as criteria for fair division. Other solutions proposed include those of Raiffa (1953), Kalai and Smorodinsky (1975), and Maschler and Perles (1981). Raiffa (1983), in a perceptive book dealing with teaching and experimentation in negotiation, provides valuable insights into the relationship between axioms and experiments in bargaining, as does Roth (1987).

3.1.3 Auctions. One of the earliest ways of disposing of goods at a price is some form of auction. Herodotus provides a description of an auction of brides in Babylonian times. A historical sketch of auctions has been provided by Shubik (1983b); Engelbreth-Wiggans (1980) and Milgrom and Weber (1982) have excellent surveys. The seminal paper is that of Vickery (1961). A simple example (that is related to both bargaining and auctions) serves to illustrate the nature of the game with incomplete information.

Two partners wish to break up a partnership. Each believes that the other values the business at somewhere between 0 and 1000 with all values equiprobable. They agree on a buy–sell agreement. They toss to see who makes the offer. Partner 1 knows that his value is C. He must make the offer to partner 2. If he bids aC his profit will be composed of two terms, the profit he makes if he buys or the profit if he sells:

$$(C-aC)\frac{ac}{1000}+\frac{1}{1000}\int_{y=aC}^{1000}(y-C)dy. \tag{4}$$

This amounts to

$$\frac{1}{1000}\left(\frac{3}{2}a(1-a)C^2+\frac{1000^2}{2}-1000C\right). \tag{5}$$

It can be shown that the optimal or expected most profitable bid is $a=\frac{1}{2}$. The partner will bid half of his true value. Suppose that $C=500$, then his profit is approximately 48.9. If he bids 500, his expected profit is zero.

3.1.4 Market Games and General Equilibrium. When we consider all games that can be represented in coalitional form by a characteristic function, only a small part will represent games arising from economies with a price system. Studying TU games, Shapley and Shubik (1969a,1976) defined a class of games called *totally balanced games* and showed that all exchange economies give rise to totally balanced games and *vice versa*. They also studied [Shapley and Shubik (1969b)] the relationship among the core, value, and competitive equilibrium for large market games. This was part of a program to show that many conceptually different solutions approached the same outcomes for large market games. The outcomes are the competitive equilibria. For market games the relationship of the limiting behavior of the core, nucleolus, value, and bargaining set to the competitive equilibria of an economy has been examined. A remarkable coincidence of solutions based on strikingly different assumptions is obtained for mass economic exchange.

3.1.5 Strategic Market Games and Money. One application of games in strategic and extensive form is to the theory of money and financial institutions. A Strategic Market Game [Shubik (1973); Shapley and Shubik (1977); and Dubey and Shubik (1978)] is a playable game whose rules provide a complete and consistent description of markets, price formation mechanisms, credit arrangements, and default procedures. A modern market may be regarded as an efficient aggregating–disaggregating device that operates on the actions of masses of anonymous individuals, accepting as inputs goods and money and producing as outputs prices and the redistribution of the goods and money. An example of an elementary price formation mechanism is as follows. Suppose that there are m sellers and n buyers of some good k. If all bid and offer simultaneously and if b^i_k is the amount of money offered by individual i for k and q^j_k is the amount of good k offered by j, then the price is given by

$$p_k = \sum_{i=1}^{m} b^i_k \left(\sum_{j=1}^{n} q^j_k \right)^{-1} . \tag{6}$$

There are many mechanisms for price formation. The one illustrated here is among the simplest. Others call for consideration of asymmetric information and sequential moves. By considering games with a continuum of traders, mass markets can be investigated. The properties of a mass market have been axiomatized [see Dubey *et al.* (1980)]. A key question to be asked is, under what conditions do efficient prices emerge in the economy? Market games are described in coalitional form; strategic market games are in strategic or extensive form. The former are preinstitutional; the latter, by the need to include a process description, provide elemental descriptions of market and financial institutions.

3.1.6 Accounting. The Shapley value is in essence a combinatoric averaging of the marginal contributions of each individual to the whole. As such it has direct application to some classical problems in accounting concerning the allocation of joint costs and incentive implications of such allocations. Shubik (1962) suggested this application of the value in order to examine the compatibility of accounting schemes that assign overheads to individual subunits with incentive schemes for the individual units. In other words, does the scheme of assigning profits and costs encourage individual units to behave in a manner compatible with the overall goals of the firm? The literature on applications of game theory to joint allocation has grown considerably. The books edited by Moriarty (1981) and Young (1985) provide many examples of application. Littlechild and Thompson (1977) apply this approach to aircraft landing fees; Young *et al.* (1982), to cost allocation in water resources development.

3.2 Political Science

Possibly the most direct and important application to political science has been in the theory of voting. The earliest application was the Shapley–Shubik (1954) power index, which is derived from the value in application to a class of games called *simple games*, which can be used to represent voting situations. In a simple game, the characteristic function takes on only two values, zero or one, which represent a motion lost or won.

3.2.1 Voting. It is reasonable to attribute the same *a priori* level of power or influence to each individual with one vote in a situation with a simple majority vote. If weighted voting is permitted, a natural question to ask is, how does an individual's power increase as his vote increases? If he has more than 50% of the vote he has complete control over the outcome. Thus we may ask, what is the nonlinear function that portrays the increase in power of the individual as his vote is increased? The index of power can be illustrated by two simple examples. Suppose first there is a simple majority game with five voters, each with a single vote. It can be denoted by (3:1,1,1,1,1), where the first number is the quota, i.e., the number of votes required for victory, and the other five numbers indicate the number of votes for each individual. As all individuals are identical their *a priori* power will be the same and can be denoted as $(\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5})$. We now consider a slightly more complicated game. Suppose that there is the same number of votes with one fewer voter, and hence one of the voters has two votes. The game can be denoted by (3:2,1,1,1). The power of these voters can be calculated by considering each one's contribution to the 24 permutations of the order of voting which can take place. The tabulation below shows a way to represent all lineups. A dot placed over a vote indicates that that vote will be pivotal, i.e., by bringing this member into the coalition previously formed, he provides the margin that enables that coalition to control the outcome.

$2,\dot{1},1,1$	six ways
$1,\dot{2},1,1$	six ways
$1,1,\dot{2},1$	six ways
$1,1,\dot{1},2$	six ways

If the voter with two votes enters the coalition first there are 3! or six ways in which the three one-vote voters can enter. Similarly, there are six ways that each of the other patterns can form. The two-vote voter is pivotal in 12 of the 24 possibilities. The voting power of all four can be described by $(\frac{1}{2},\frac{1}{6},\frac{1}{6},\frac{1}{6})$. The index has been applied among others to the United Nation's security council and the electoral college [Mann and Shapley (1964)]. It is possible to design voting schemes where an individual with some votes may have no power whatsoever. In the weighted voting game (8:4,4,4,3) the voter with three votes is never pivotal in a winning coalition.

An interesting connection can be made with the power index and the Chow factor used in electrical engineering to measure redundancy and the probability of failure in networks. The concept of a winning coalition can be viewed as equivalent to asking when there are enough elements in a redundant system to keep it functioning [Dubey and Shapley (1980)]. There is also an alternative index suggested by Banzhaff (1965), which is discussed in Dubey and Shapley. The important concept of minimal winning coalition has been imaginatively applied to voting theory by Riker (1962).

The fully strategic aspects of voting require the study of games in extensive form. The basic work of Farquharson (1969) analyzed the logic of voting for candidates other than one's most preferred candidate. There is also a growing literature on special voting structures such as preference voting and proportional representation.

3.2.2 Deterrence Theory. Starting with the provocative work of Schelling (1960) and Ellsberg (1961), there has been a steady growth in game-theoretic application to international relations in general [see the exhaustive bibliography of O'Neill (1992)], and problems in the analysis of threats and nuclear warfare. Two recent examples of game-theoretic models concerned with nuclear threat and stability are Powell (1990) on nuclear deterrence theory and Bracken and Shubik (1991) on problems of stability in a multipolar nuclear world.

3.3 Biology

Since the early 1970s there has been a considerable growth in the application of game-theoretic models to biology. This work has as its major proponent Maynard Smith (1982). The central concept is the evolutionary stable strategy (ESS), which may be interpreted as a relative of the noncooperative equilibrium, where, however, the choice of strategy is interpreted genetically rather than as a conscious strategic choice. We may consider a population of animals with pairwise random encounters. Each animal is "coded" to be either aggressive or peaceful. The popu-

Table 5. A game with an evolutionary stable equilibrium.

	1	2
1	0,0	9,4
2	4,9	5,5

lation is split 50–50 between the two types. The payoffs shown in Table 5 may be regarded as the reproduction success of the two individuals after encounter. Thus, if an aggressive individual meets another aggressor, neither has offspring. If an aggressor meets a peaceful individual he will have 9 offspring and the other will have only 4. We can check that the 50–50 split of the population is in equilibrium. The population is stable. We assume that reproduction involves only the single animal and that the next generation is coded to follow the strategy of its parent. The fitness of each individual to reproduce is determined by its strategy and the strategy of the individual it encounters. The ESS is equivalent to a mixed strategy in the game, but is based on a frequency interpretation. Selten (1980), Crawford (1989), and others have considerably extended the original idea of Maynard Smith to include asymmetric conflicts, multispecies involvement, and the possibility of learning. It appears to be a safe bet that there will continue be a considerable growth in the application of game theory to biology in the next few years.

3.4 Other Applications

There is a growing literature of game-theoretic applications in the law as is evidenced by a forthcoming special issue of *The Cincinnati Law Review* devoted to game theory [Shubik (1992b)].

There is also work in agency theory. The simple nongame-theoretic view of organization has individuals arranged in various forms of hierarchial structure with the power of decision making flowing down from high to the lower ranks. The game-theoretic view takes into account that individuals often employ others whose goals and information differ from their own. Thus even though *A* may be superior to *B*, *B* may be in a position to withhold information from *A* in a manner that gives him considerable control over *A*. The agency problem takes this into account specifically. Papers in anthropology [Barth (1959)], sociology and philosophy [Rawls (1971)], and even biblical interpretation [Brams (1980)] have appeared.

4. MODELING, GAMING, AND EMPIRICAL VALIDATION

4.1 Experimental Gaming

Roth (1987), in a provocative discussion of laboratory experimentation in economics, has suggested three kinds of activities for those engaged in economic experimental gaming. They are "speaking to theorists," "searching for facts," and "whispering in the ears of princes." The third of these activities, from the viewpoint of society, is possibly the most immediately useful. In the United States the princes tend to be corporate executives, politicians, and generals or admirals. However, the topic of operational gaming merits a separate treatment.

Perhaps the most important item to keep in mind is the fundamental distinction between game theory and the approach of social psychology to the same problems. Underlying a considerable amount of game theory is the concept of *external symmetry*. Unless stated otherwise all players in game theory are treated as though they are intrinsically symmetric in all nonspecified attributes. They have the same perceptions and abilities to calculate and they have the same personalities. Much of game theory is devoted to deriving normative or behavioral solution concepts to games played by personality-free players in context-free environments. In particular, one of the impressive achievements of formal game theory has been to produce a menu of different solution concepts that suggest ingeniously reasoned sets of outcomes to nonsymmetric games played by externally symmetric players.

In contrast, the approach of the social psychologists has for the most part been to take symmetric games and observe the broad differences in play that can be attributed to nonsymmetric players. Personality, passion, and perception count and the social psychologist is interested to see how they count.

A detailed survey of the role of gaming relevant to the study of game theory is given by Shubik (1992a). One of the earliest experiments was that of Mosteller and Nogee (1951), devoted to measuring individual utility scales. The book by Siegal and Fouraker (1960) launched work on the study of bilateral monopoly and bargaining. Raiffa's (1983) imaginative work on the art and science of bargaining and negotiation and the experimental work of Roth (1987,1988) on bargaining indicate how this line of investigation has developed. Early work on competition among few players in a market setting was done by Hoggatt (1969), Friedman (1967,1969), Shubik *et al.* (1972), V. L. Smith (1965), and others. More recently V. L. Smith (1986), in collaboration with many others, has performed many experiments on auction markets, but with the emphasis more on the economics of many competitors than on the game-theoretic aspects of competition among the few.

There is now a vast literature on experimentation with matrix games where one of the major questions involves the validity of various forms of noncooperative equilibrium as a predictor of limiting behavior. The book by Rapaport *et al.* (1976) provides an excellent introduction to experimentation with matrix games.

4.2 Towards Dynamics and Bounded Rationality

The evidence from human experience and experimental gaming is overwhelming that in many contexts the personality-free rational calculator model of the human is a gross oversimplification. Work has begun on models of bounded rationality, where the individual's computational and perceptual abilities are limited. Formal theoretical work by Abreu and Rubinstein (1988) has considered games played by players modeled as simple automata with costly computation. Important works by Geanakoplos and Gray (1991) and by Baum (1991) have shown that for players with limited ability to compute, looking ahead further in a game such as chess is not the best strategy if the players do not have a perfect way of evaluating intermediate positions. These results may eventually have deep impact on the concept of long-range planning.

The work in gaming and in bounded rationality points towards a merging of the approaches towards the game-theoretic study of mass human economic and other behavior, not only by means of experimental games, but by mass simulations where each individual is represented as a behavioral particle in a mass stochastic environment. Thus it may well be that in the future at least part of the trend in game-theoretic research will be towards a statistical mechanics of human behavior, blending individual optimization with the many possibilities provided by human communication for correlated behavior.

GLOSSARY

Continuum of Players: A concept capturing the idea that a single player or a finite subset of players has negligible strategic influence. This enables the analyst to use measure-theory techniques, which are of considerable value in studying mass social behavior.

Equilibrium, Noncooperative: A solution concept for games that is based on individual optimization under the assumption of mutually consistent expectations. If all expect a certain outcome and each optimizes individually, based on this expectation, then the actions of all will result in their expectations being realized. Although equilibria exist for many games, little is known about the stability of these equilibria. The dynamics of adjustment calls for a theory of learning that is not covered by game theory.

Game, Market: A specialization of a game in coalitional form so that it reflects the basic economic model of trade and bargaining among individual owners where all coalitions are permitted and can form costlessly.

Game, Strategic Market: A specialization of a game in strategic form that reflects the basic economic model for anonymous trade utilizing a market mechanism.

Game, TU or NTU: A game is a TU (transferable utility) game if there exists a commodity for which all individuals have constant marginal utility regardless of the quantity they possess. It is a perfectly transferable store of wealth. The TU assumption is a good approximation for some economic situations where money more or less serves this purpose. For many situations the TU assumption is unreasonable and NTU (no transferable utility) games do not use this assumption. TU games are mathematically far more tractable than NTU games.

Information Set: The subset of alternatives that are distinguished by the player when he is called upon to move. If the player can always distinguish all of his alternatives he is said to be perfectly informed.

Prior: Term used in economics and statistics to denote the *a priori* probability distribution for the occurrence of an expected event.

Rooted Tree: A connected graph with no loops and a distinguished node (the root) which indicates the start of the game.

Strategy: A complete set of instructions as to how a player will select his moves under all contingencies. In general, even for most board games the number of strategies is hyperastronomical. Even with the largest computers an exhaustive investigation of all strategies is not feasible.

Value: A solution to a game in coalitional form that awards each player his expected marginal contribution to all coalitions on the assumption that all the ways in which he can be added to any coalition are equiprobable.

List of Works Cited

Abreu, D., Rubinstein, A. (1988), "The Structure of Nash Equilibria in Repeated Games with Finite Automata," *Econometrica* **56**, 1259–1281.

Aumann, R. J. (1964), "Markets with a Continuum of Traders," *Econometrica* **32**, 39–50.

Aumann, R. J. (1976), "Agreeing to Disagree," *Ann. Statistics* **4**, 1236–1239.

Aumann, R. J. (1987), "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica* **55**, 1–18.

Aumann, R. J., Maschler, M. (1964), "The Bargaining Set for Cooperative Games," in: M. Drescher, L. S. Shapley, A. W. Tucker (Eds.), *Advances in Game Theory*, Princeton, NJ: Princeton University Press, pp. 443–447.

Aumann, R. J., Shapley, L. S. (1974), *Values of Non-Atomic Games*, Princeton, NJ: Princeton University Press.

Banzhaff, J. F. (1965), "Weighted Voting Doesn't Work: A Mathematical Analysis," *Rutgers Law Rev.* **19**, 317–343.

Barth, F. (1959), "Segmentary Opposition and the Theory of Games: A Study of Pathan Organizations," *J. R. Anthropol. Soc. (London)* **89**, 5–21.

Baum, E. B. (1991), "Minimax Is Not Optimal," NEC Research Institute Report, Princeton, NJ.

Bertrand, J. (1883), "Théorie Mathématique de la Richesse Sociale" (review), *J. Savants (Paris)*, 499–508.

Borel, E. (1924), "Sur les Jeux ou Interviennent l'Hasard et l'Habilité des Joueurs," in: J. Hermann (Ed.), *Elements de la Thèorie des Probabilitès*, Paris: Librairie Scientifique, pp. 204–224.

Bracken, J., Shubik, M. (1991), "Strategic Offensive and Defensive Forces and Worldwide Coalition Games," in: R. Avenhaus, H. Karkar, M. Rudnianski (Eds.), *Defense Decision Making*, Berlin: Springer, pp. 169–185.

Brams, S. J. (1980), *Biblical Games: A Strategic Analysis of Stories in the Old Testament*, Cambridge, MA: MIT Press.

Case, J. H. (1979), *Economics and the Competitive Process*, New York: New York University Press.

Cournot, A. A. (1897), *Researches into the Mathematical Principles of the Theory of Wealth*, New York: Macmillan (Bacon translation of original 1838 French).

Crawford, V. P. (1989), "Learning and Mixed Strategy Equilibria in Evolutionary Games," *J. Theoret. Biology* **140**, 537–550.

Davis, M., Maschler, M. (1965), "The Kernel of a Cooperative Game," *Naval Res. Logistics Quart.* **12**, 223–259.

Debreu, G. Scarf, H. (1963), "A Limit Theorem on the Core of an Economy," *Intl. Econ. Rev.* **4**, 236–246.

Drescher, M. (1961), *Games of Strategy: Theory and Applications*, Englewood Cliffs, NJ: Prentice-Hall.

Dubey, P. (1986), "Inefficiency of Nash Equilibria," *Math. Operations Res.* **11**, 1–8.

Dubey, P., Rogawski, J. D. (1990), "Inefficiency of Smooth Market Mechanisms," *J. Math. Econ.* **19**, 285–304.

Dubey, P., Shapley, L. S. (1980), "Non Cooperative Exchange with a Continuum of Traders: Two Models," *Technical Report of the Institute for Advanced Studies*, Hebrew University of Jerusalem.

Dubey, P., Mas-Colell, A., Shubik, M. (1980), "Efficiency Properties of Strategic Market Games: An Axiomatic Approach, *J. Econ. Theor.* **22**, 339–362.

Edgeworth, F. Y. (1881), *Mathematical Psychics*, London: Kegan Paul.

Ellsberg, D. (1961), "The Crude Analysis of Strategic Choice," *Am. Econ. Rev.* **51**, 472–478.

Engelbrecht-Wiggins, R. (1980), "Auctions and Bidding Models: A Survey," *Management Sci.* **26**, 119–142.

Farquharson, R. (1969), *Theory of Voting*, New Haven, CT: Yale University Press.

Fishburn, P. (1970), *Utility Theory*, New York: Wiley.

Friedman, J. (1967), "An Experimental Study of Cooperative Duopoly," *Econometrica*, **35**: 379–397.

Friedman, J. (1969), "On Experimental Research on Oligopoly," *Rev. Econ. Studies* **36**, 399–415.

Friedman, J. W. (1990), *Game Theory*, New York: Oxford University Press.

Geanakoplos, J., Gray, L. (1991), "When Seeing Further Is Not Seeing Better," *Bull. Santa Fe Institute*, Fall-Winter, **6** (2), 4–9.

Gillies, D. B. (1953), "Some Theorems on *n*-Person Games," Ph.D. dissertation, Department of Mathematics, Princeton University.

Harsanyi, J. C. (1956), "Approaches to the Bargaining Problem Before and After the Theory of Games: A Critical Discussion of Zeuthen's, Hicks' and Nash's Theories," *Econometrica* **24**, 144–157.

Harsanyi, J. (1963), "A Simplified Bargaining Model for the *n*-Person Cooperative Game," *Intl. Econ. Rev.* **4**, 194–220.

Harsanyi, J. C. (1967), "Games with Incomplete Information Played by 'Bayesian' Players, Part I," *Management Sci.* **14**, 159–182.

Harsanyi, J. C. (1968a), "Games with Incomplete Information Played by 'Bayesian' Players, Part II," *Management Sci.* **14**, 320–334.

Harsanyi, J. C. (1968b), "Games with Incomplete Information Played by 'Bayesian' Players, Part III," *Management Sci.* **14**, 486–502.

Harsanyi, J. C., Selten, R. (1987), *A General Theory of Equilibrium Selection in Games*, Cambridge, MA: MIT Press.

Hildenbrand, W. (1982), "Core of an Economy," in: K. J. Arrow, M. D. Intrilligator (Eds.), *Handbook of Mathematical Economics*, Amsterdam: North-Holland, Chap. 18.

Hoggatt, A. C. (1969), "Response of Paid Subjects to Differential Behavior of Robots in Bifurcated Duopoly Games," *Rev. Econ. Studies* **36**, 417–432.

Isaacs, R. (1965), *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control, and Optimization*, New York: Wiley.

Kalai, E., Smorodinsky, M. (1975), "Other Solutions to Nash's Bargaining Problem," *Econometrica* **43**, 513–518.

Kuhn, H. W. (1953), "Extensive Games and the Problem of Information," in: H. Kuhn, A. W. Tucker (Eds.), *Contributions to the Theory of Games*, Princeton, NJ: Princeton University Press, pp. 193–216.

Littlechild, S. C., Thompson, G. F. (1977), "Aircraft Landing Fees and Game Theory Approach," *Bell J. Econ.* **8**, 186–204.

Lucas, W. F. (1969), "The Proof that a Game May Not Have a Solution," *Trans. Am. Math. Soc.*, **137**, 219–229.

Mann, I., Shapley, L. S. (1964), "The a priori Voting Strength of the Electoral College," in: M. Shubik (Ed.), *Game Theory and Related Approaches to Social Behavior*, New York: Wiley, pp. 151–164.

Maschler, M., Perles, M. (1981), "The Superadditive Solution for the Nash Bargaining Game," *Intl. J. Game Theory* **10**, 163–193.

Mayberry, J., Nash, J. F., Jr., Shubik, M. (1953), "A Comparison of Treatments of a Duopoly Situation," *Econometrica* **21**, 141–155.

Mertens J. F, Zamir, S. (1985), "Formulation of Bayesian Analysis for Games with Incomplete Information," *Intl. J. Game Theory* **14**, 1–29.

Milgrom, P., Weber, R. J. (1982), "A Theory of Auctions and Competitive Bidding," *Econometrica* **50**, 1089–1122.

Milnor, J. W., Shapley, L. S. (1957), "On Games of Survival," in: M. Drescher, A. W. Tucker, P. Wolfe (Eds.), *Contributions to the Theory of Games*, Vol. III, Princeton, NJ: Princeton University Press, pp. 15–45.

Moriarity, S. (Ed.) (1981), *Joint Cost Allocations*, Tulsa: University of Oklahoma Press.

Mosteller, F., Nogee, P. (1951), "An Experimental Measurement of Utility," *J. Pol. Econ.* **59**, 371–404.

Myerson, R. B. (1991), *Game Theory*, Cambridge, MA: Harvard University Press.

Nash, J. F., Jr. (1950), "The Bargaining Problem," *Econometrica* **18**, 155–162.

Nash, J. F., Jr. (1951), "Non-Cooperative Games," *Ann. Math.* **54**, 289–295.

Nash, J. F., Jr. (1953), "Two Person Cooperative Games," *Econometrica* **21**, 128–162.

Nyblen, G. (1951), *The Problem of Summation in Economic Science*, Lund: Gleerup.

O'Neill, B. (1992), "Sources in Game Theory for International Relations Specialists," in: R. J. Aumann, S. Hart (Eds.), *Handbook of Game Theory*, Vol. 2, Amsterdam: North-Holland.

Peleg, B. (1967), "Existence Theorem for the Bargaining Set M 1(i)," in: M. Shubik (Ed.), *Essays in Mathematical Economics in Honor of Oskar Morgenstern*, Princeton, NJ: Princeton University Press, pp. 53–56.

Peleg, B. (1987), "An Axiomatization of the Core of Cooperative Games without Side Payments," *J. Math. Econ.* **14**, 203–214.

Powell, R. (1990), *Nuclear Deterrence Theory: The Search for Credibility*, Cambridge, MA: Cambridge University Press.

Raiffa, H. (1953), "Arbitration Schemes for Generalized Two-Person Games," in: H. Kuhn, A. W. Tucker (Eds.), *Contributions to the Theory of Games*, Princeton, NJ: Princeton University Press.

Raiffa, H. (1983), *The Art and Science of Negotiation*, Cambridge, MA: Harvard University Press.

Rapaport, A., Guyer, M. J., Gordon, D. G. (1976), *The* 2×2 *Game*, Ann Arbor, MI: University of Michigan Press.

Rawls, J. (1971), *A Theory of Justice*, Cambridge, MA: Harvard University Press.

Riker, W. H. (1962), *The Theory of Political Coalitions*, New Haven, CT: Yale University Press.

Roth, A. E. (1979), *Axiomatic Models of Bargaining*, Lecture Notes in Economics and Mathematical Systems No. 170, Berlin: Springer.

Roth, A. E. (Ed.) (1987), *Laboratory Experimentation in Economics: Six Points of View*, Cambridge, MA: Cambridge University Press.

Roth, A. E. (1988), "Laboratory Experimentation in Economics: A Methodological Overview," *Econ. J.* **98**, 974–1031.

Schelling, T. C. (1960), *Strategy and Conflict*, Cambridge, MA: Harvard University Press.

Schmeidler, D. (1969), "The Nucleolus of a Characteristic Function Game," *SIAM J. Appl. Math.* **17**, 1163–1170.

Selten, R. C. (1975), "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *Intl. J. Game Theory* **4**, 25–55.

Selten, R. C. (1980), "A Note on Evolutionary Stable Strategies in Asymmetric Animal Conflicts," *J. Theoret. Biology* **84**, 101.

Shapley, L. S. (1953a), "Stochastic Games," *Proc. Natl. Acad. Sci. U.S.A.* **39**, 1095–1100.

Shapley, L. S. (1953b), "A Value for *n*-Person Games," in: H. Kuhn, A. W. Tucker (Eds.), *Contributions to the Theory of Games*, Princeton, NJ: Princeton University Press, pp. 305–317.

Shapley, L. S., Shubik, M. (1954), "A Method for Evaluating the Distribution of Power in a Committee System," *Am. Pol. Sci. Rev.* **48**, 787–792.

Shapley, L. S., Shubik, M. (1969a), "On Market Games," *J. Econ. Theor.* **1**, 9–25.

Shapley, L. S., Shubik, M. (1969b), "Pure Competition, Coalitional Power, and Fair Division," *Intl. Econ. Rev.* **10**, 337–362.

Shapley, L. S., Shubik, M. (1976), "Competitive Outcomes in the Cores of Market Games," *Intl. J. Game Theory* **4**, 229–237.

Shapley, L. S., Shubik, M. (1977), "Trade Using One Commodity as a Means of Payment," *J. Polit. Econ.* **85**(5), 937–968.

Shubik, M. (1959a), "Edgeworth Market Games," in: R. D. Luce, A. W. Tucker (Eds.), *Contributions to the Theory of Games*, Vol. III, Princeton, NJ: Princeton University Press, pp. 267–278.

Shubik, M. (1959b), *Strategy and Market Structure*, New York: Wiley.

Shubik, M. (1962), "Incentives, Decentralized Control, the Assignment of Joint Costs and Internal Pricing," *Management Sci.* **8**, 325–343.

Shubik, M. (1973), "Commodity, Money, Oligopoly, Credit, and Bankruptcy," *West. Econ. J.* **11** (1), 24–38.

Shubik, M. (1982), *Game Theory in the Social Sciences, Concepts, and Solutions*, Cambridge, MA: MIT Press.

Shubik, M. (Ed.) (1983a), *Mathematics of Conflict*, Amsterdam: North-Holland.

Shubik, M. (1983b), "A History of Auctions" in: Engelbrecht-Wiggins, R., Shubik, M., Stark, R. (Eds.), (1983), *Auctions, Bidding and Contracting: Uses and Theory*, New York: New York University Press.

Shubik, M. (1992a), "Game Theory and Experimental Gaming," in: R. J. Aumann, S. Hart (Eds.), *Handbook of Game Theory*, Vol. 2, Amsterdam: North-Holland.

Shubik, M. (1992b), "Game Theory, Law, and the Concept of Competition," *Cincinnati Law Rev.* **60**, 285–303.

Shubik, M., Wolf, G., Eisenberg, H. B. (1972), "Some Experiences with an Experimental Oligopoly Business Game," *General Systems* **17**, 61–75.

Siegal, S., Fouraker, L. (1960), *Bargaining and Group Decision Making: Experiments in Bilateral Monopoly*, New York: McGraw-Hill.

Smith, J. Maynard (1982), *Evolution and the Theory of Games*, Cambridge, England: Cambridge University Press.

Smith, V. L. (1965), "Experimental Auction Markets and the Walrasion Hypothesis," *J. Pol. Econ.* **73**, 387–393.

Smith, V. L. (1986), "Experimental Methods in the Political Economy of Exchange," *Science* **234**, 167–172.

Sobolev, I. V. (1975), "The Characterization of Optimality Principles in Cooperative Games by Functional Equations," in: N. N. Vorobiev (Ed.), Proceedings of Seminar on Mathematical Methods in the Social Sciences, Vilnius, Lithuania (unpublished), Vol. 6, pp. 150–165 (in Russian with English summary).

Stackelberg, H. F. von (1952), *The Theory of the Market Economy*, London: Hoage (translated from the German original, 1934).

Tirole, J. (1988), *The Theory of Industrial Organization*, Cambridge, MA: MIT Press.

Van Damme, E. (1987), *Stability and Perfection of Nash Equilibria*, Berlin: Springer.

Vickery, W. (1961), "Counterspeculation, Auctions, and Competitive Sealed Tenders," *J. Finance* **16**, 8–37.

von Neumann, J. (1928), "Zur Theorie der Gesellschaftsspiele," *Mathematische Annalen* **100**, 295–320.

von Neumann, J., Morgenstern, O. (1944), *The Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press.

Wooders, M. H. (1983), "The Epsilon Core of a Large Replica Game," *J. Math. Econ.* **11**, 277–300.

Young, H. P. (Ed.) (1985), *Cost Allocation: Methods, Principles, Applications*, Amsterdam: North-Holland.

Young, H. P., Okada, N. Hashimoto, T. (1982), "Cost Allocation in Water Resources Development," *Water Resources Res.* **18** (3), 463–475.

Zermelo, E. (1913), "Uber eine Anwendung der Mengenlehre auf die Theorie des Schachspiels," in: E. W. Hobson, A. E. H. Love (Eds.), *Proceedings of the Fifth International Congress of Mathematicians*, Cambridge, England: Cambridge University Press, Vol. II, pp. 501–504.

Zeuthen, F. (1930), *Problems of Monopoly and Economic Warfare*, London: Routledge.

Further Reading

Aumann, R. J., Shapley, L. S. (1974), *Values of Non-Atomic Games*, Princeton, NJ: Princeton University Press.

Drescher, M. (1961), *Games of Strategy: Theory and Applications*, Englewood Cliffs, NJ: Prentice-Hall.

Farquharson, R. (1969), *Theory of Voting*, New Haven, CT: Yale University Press.

Harsanyi, J. C., Selten R. (1987), *A General Theory of Equilibrium Selection in Games*, Cambridge, MA: MIT Press.

Raiffa, H. (1983), *The Art and Science of Negotiation*, Cambridge, MA: Harvard University Press.

Rapaport, A., Guyer, M. J., Gordon, D. G. (1976), *The* 2×2 *Game*, Ann Arbor, MI: University of Michigan Press.

Riker, W. H. (1962), *The Theory of Political Coalitions*, New Haven, CT: Yale University Press.

Shubik, M. (1982), *Game Theory in the Social Sciences, Concepts, and Solutions*, Cambridge, MA: MIT Press.

Shubik, M. (1984), *A Game Theoretic Approach to Political Economy*, Cambridge, MA: MIT Press.

Siegal, S., Fouraker, L. (1960), *Bargaining and Group Decision Making: Experiments in Bilateral Monopoly*, New York: McGraw-Hill.

Smith, J. Maynard (1982), *Evolution and the Theory of Games*, Cambridge, MA: Cambridge University Press.

Van Damme, E. (1987), *Stability and Perfection of Nash Equilibria*, Berlin: Springer.

von Neumann, J., Morgenstern, O. (1944), *The Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press.

GAMMA-RAY SPECTROMETERS

See SPECTROMETERS, GAMMA

GAS LASERS

See LASERS, GAS

GASDYNAMICS

See FLUID DYNAMICS, EQUILIBRIUM; RAREFIED GAS DYNAMICS

GENERAL RELATIVITY

See GRAVITATION AND GENERAL RELATIVITY

GEODESY

CLYDE C. GOAD AND IVAN I. MUELLER, *Department of Geodetic Science and Surveying, The Ohio State University, Columbus, Ohio, U.S.A.*

Introduction 89
1. **Geometric Geodesy** 90
1.1 Preview 90
1.2 Computations on the Ellipsoid 92
1.2.1 Geodetic Coordinates and the Geodetic Datums 92
1.2.2 Plane Coordinates 93
2. **Determination of the Reference Surfaces and Gravimetric Geodesy** 93
2.1 Geometric Methods 93
2.1.1 Geometric Determination of the Reference Ellipsoid 93
2.2 Gravimetric Geodesy 94
2.2.1 Equipotential (Level) Surfaces and Gravity 94
2.2.2 Normal Gravity and Gravity Anomaly 95
2.2.3 Determination of the Geoid from Gravity Measurements... 95
2.2.4 Leveling, Orthometric Heights, and the Geoid 96
3. **Modern Cartesian Three-Dimensional Geodesy** .. 97
3.1 Preview 97
3.2 Satellite Geodesy 98
3.2.1 Introduction 98
3.2.2 Principal Methods 98
3.2.2.1 General Dynamic Method 98
3.2.2.2 Partial Dynamic Method ......
3.2.2.3 Short-Arc, Point-Positioning, and Translocation Methods ... 100
3.2.3 Applications of Satellite-Derived Geodetic Parameters 100
3.2.4 Special Satellite Geodetic Techniques 102
3.2.4.1 Global Positioning System (GPS) 102
3.2.4.2 Satellite Altimetry 103
3.2.5 Lunar Laser Ranging (LLR) .. 103
3.3 Very Long-Baseline Interferometry 103
3.3.1 Introduction 103
3.3.2 Basic Instrumentation 103
3.3.3 Applications 104
4. **Concluding Remarks** 104
Glossary 105
Works Cited 106
Further Reading 106

INTRODUCTION

Geodesy is traditionally defined as the science of studying the figure of the earth and its gravity field. Now that geodetic measurements have become very accurate and easily repeatable, the traditional definition must be extended to include temporal variations as well.

Although inquiry into the figure of the earth can be traced to antiquity, the recognized founder of scientific geodesy is Eratosthenes (276–195 B.C.) of Alexandria (Torge, 1980). Eratosthenes studied the rays of the sun at Alexandria and Syene (Assuan, today) at summer solstice noting that at Syene the rays entered a well in the vertical direction and at Alexandria (roughly on the same meridian) the rays were off the local vertical direction by about 1/50 of a circle. Using the estimated distance between Syene and Alexandria of 5000 stadia (1 Egyptian stade $=157.5$ m), he calculated the radius of the earth to be 6267 km. This is less than 2% lower than the actual mean radius of 6371 km.

Historically, other estimates differ by 3%–10% from the value known today. In 1669–1670, at the urging of the newly formed Academy of Sciences in Paris (1666), J. Picard carried out arc measurements on the meridian through Paris between Malvoisine and Amiens with the aid of a triangulation network. He was the first to use a telescope with

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

cross hairs. Picard's value for the Earth radius was good to +0.01%, which aided Newton in the verification of the law of gravitation which he formulated in 1665–1666.

From observations of pendulum motion (the period), one can infer an increase of gravity from the equator to the poles. Using such observations and their own works, Isaac Newton (1643–1727) and Christian Huygens (1629–1695) suggested Earth models flattened at the poles.

If we let the letter a stand for the equatorial radius and b the polar radius, the normalized deviation or flattening is defined as $f=(a-b)/a$. Today we know that $1/f=298.257$. Historically this value also was attempted to be found through meridian arc measurements. Depending on where the arc measurements were obtained, the determination of the flattening could have even the wrong sign (polar radius greater than equatorial). In 1743, A.C. Clairaut (1713–1765) synthesized the physical and geodetic concerns in a theorem named for him that permits the computation of flattening from two gravity measurements at different latitudes. However, the practical implementation of this theorem suffered until the twentieth century from the lack of precise and well-distributed gravity measurements and the inability to reduce these data to the Earth ellipsoid.

Historically, the task of determining the figure of the Earth was, for simplicity, subdivided into two almost independent tasks: horizontal surveys for determination of latitudes and longitudes and vertical surveys for determination of heights. The reference surface for horizontal surveys is a suitable but otherwise arbitrary rotational ellipsoid. The heights, however, are referred to an equipotential surface or mean sea level (which later will be called the geoid). The use of different reference surfaces for horizontal and for vertical coordinates makes the accurate computations of three-dimensional Cartesian coordinates impossible, unless the separation between the mean sea level and the reference ellipsoid is known accurately. The ability to compute Cartesian coordinates is of importance because modern survey methods, such as those using artificial satellites, provide such coordinates and their comparison and compatibility with older surveys is essential.

Geodesy can be subdivided into the following groups: geometric geodesy, gravimetric geodesy, and space geodesy. Geometric geodesy deals with the measuring and computational methods that are applicable in the surveys of large areas, thus beyond the scope of plane surveying, and with the geometric determination of the size and shape of the reference ellipsoid. The geometric effort results in horizontal coordinates (geodetic latitude and longitude) referred to a suitable reference ellipsoid, and in heights above sea level. These results are mainly based on distance, direction measurements, and height differences. Gravimetric geodesy deals with the determination of the gravity equipotential surface most closely approximating mean sea level (known as the geoid) as a graphical depiction of the gravity field of the Earth, and with related problems. Geodetic astronomy provides the link between the "arbitrary" geometric reference system and the real world (nature) and thus serves to control the distorting effects of arbitrariness. This is achieved by attempting to keep the directions that are referred to the arbitrarily chosen reference ellipsoid close to "natural" directions determined from astronomical observations. Space geodesy provides local, regional, and global geodetic control through three-dimensional coordinates and vectors determined mostly from observations on artificial satellites and other space techniques. Since the orbits of satellites are affected by, among other things, the gravity field of the Earth, these observations also provide a means to improve our knowledge of this field.

In the following, these areas of study along with geophysical implications are described in varying detail.

1. GEOMETRIC GEODESY

1.1 Preview

If a whole country is to be surveyed, a network of control points evenly distributed over the entire country is set up and the mutual positions of the stations are determined on an ellipsoid that fits well the Earth (mean sea level) over the area in question. The mutual positions of the control points are given by their coordinates in a reference system especially set up for the area (geodetic datum). *A* datum may cover the area of only a given country or of several ones. For exam-

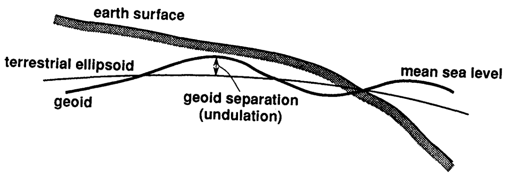

FIG. 1. Geoid and terrestrial ellipsoid generally do not deviate from each other by more than about 100 m.

ple, the North American Datum of 1983 covers Canada, Greenland, Mexico, and the United States. In order to determine the coordinates of control points, the procedure has been as follows.

Initially, points of the first-order network, separated by distances of generally 20 to 50 km, are selected. Their connecting lines with adjacent points should form, as nearly as possible, equilateral triangles. Then all angles (triangulation) or all distances (trilateration) or their combinations in the individual triangles are measured. The inevitable measuring discrepancies are eliminated in a least-squares adjustment. In order to determine the length of the triangle sides, i.e., the scale, the length of at least one side *must* be measured.

In order to understand how the position of the control points on the ellipsoid can be derived from triangulation or trilateration (or ultra-accurate traversing), the ellipsoid will have to be described in some detail. The actual shape of the Earth approximated by the geoid is defined as the surface that is everywhere perpendicular to the plumb line (direction of gravity) and that, over the oceans, nearly coincides with the undisturbed mean surface of the oceans. The geoid is a complicated surface that cannot be described by a closed mathematical expression (see also Sec. 2.2). Therefore a region or area such as a nation or continent is approximated by a best-fitting rotational ellipsoid (terrestrial ellipsoid or spheroid) from which the mean sea level seldom deviates by more than 100 m (see Figs. 1 and 2). The position of a terrain point on the ellipsoid may be thought of as its projection (along its plumb line to the geoid and then) along the ellipsoidal normal to the ellipsoid.

To orient the network on the ellipsoid, one needs the astronomic (sometimes called the geographic) latitude and longitude of at least one control point, and the angle subtended by one triangle side and the astronomic meridian (*astronomic azimuth*). The determination of these astronomic quantities is the task of geodetic astronomy. Since the angle between the true plumb line and the ellipsoidal normal (deflection of the vertical) is always significant (Fig. 3), astronomic latitude, longitude, and azimuth are measured at several evenly

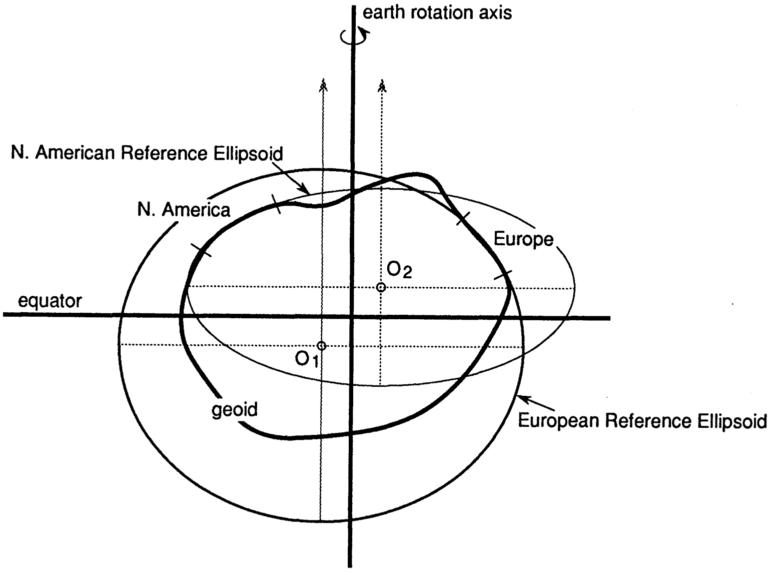

FIG. 2. The reference ellipsoids fit the geoid best over the area where they are intended to be used.

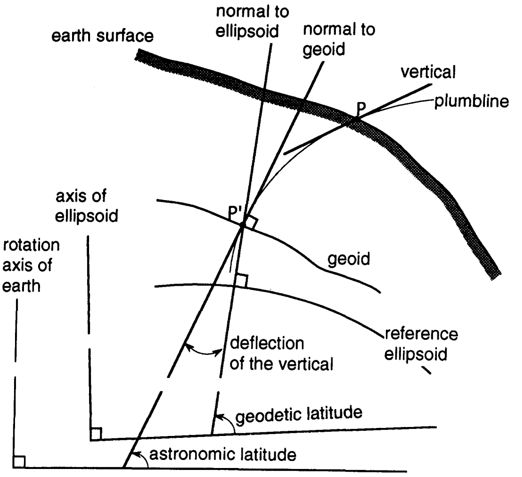
earth surface
normal to ellipsoid
normal to geoid
vertical
plumbline
P
axis of ellipsoid
P'
rotation axis of earth
geoid
reference ellipsoid
deflection of the vertical
geodetic latitude
astronomic latitude

FIG. 3. Astronomic and geodetic coordinates and the deflection of the vertical.

distributed locations. Because the deflection of the vertical is dependent on the position and orientation of the ellipsoid, the best-fitting reference ellipsoid for the area can be determined from the condition that the sum of squares of the deflections of the vertical attains a minimum.

If the network is oriented on the ellipsoid, then all triangle sides (angles) can be computed from the base line and the adjusted angles (side lengths). From the triangles the geodetic coordinates of all control points can then be calculated.

Since the geodetic coordinates (latitude, longitude) are not desirable for computations in subsequent lower-order surveying, they are usually transformed into two-dimensional plane coordinates. The transformation to the plane is necessary because the end result of the survey, the map, is a plane figure, and also because calculations in the plane are much simpler than on the ellipsoid. This transformation procedure (map projection) causes some distortions that can be mathematically expressed and thus accommodated if necessary (see CARTOGRAPHY AND GEOGRAPHIC INFORMATION SYSTEMS).

1.2 Computations on the Ellipsoid

1.2.1 Geodetic Coordinates and the Geodetic Datums. The points projected on the reference ellipsoid are defined either by their geodetic latitude and longitude or by their orthogonal ellipsoidal coordinates, but mostly by the former (Fig. 4). The geodetic latitude of a point P is the angle φ subtended by the ellipsoidal normal at P and the equator. The geodetic longitude of P is the angle λ subtended by the meridian plane through P (the plane through P and the rotation axis) and a zero or prime meridian plane of a selected point Q (e.g., Greenwich). The orthogonal ellipsoidal coordinates x,y are geodesic arcs (shortest distances between points) that define the position of P with respect to an arbitrary origin P_0. The meridian through P_0 at the geodetic longitude λ_0 serves as the y axis. The abscissa x is the PF portion of the geodesic through P, perpendicular to the meridian λ_0. The ordinate is the length of the meridian arc P_0F.

A plane containing the ellipsoidal normal at P_0 and point P intersects the ellipsoid along

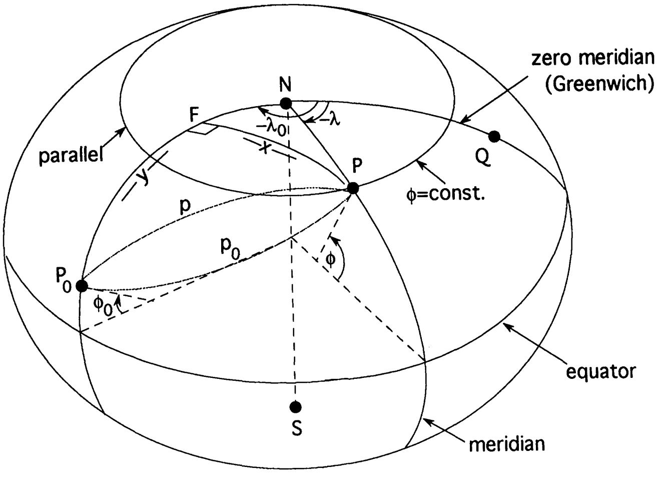
zero meridian (Greenwich)
N
F
$-\lambda_0$
$-\lambda$
x
parallel
y
P
Q
φ=const.
p
p_0
P_0
φ
ϕ_0
equator
S
meridian

FIG. 4. Coordinates on the ellipsoid.

the normal section p_0. The plane containing the normal at P and the point P_0 intersects the ellipsoid along the normal section p. Since the ellipsoidal normals at P and P_0 usually do not intersect each other (exceptions: P and P_0 are on the same meridian or parallel), the two normal sections, p and p_0, are not identical. Thus, either one has to decide which normal section should be used for the calculations, or the uniquely defined geodesic between P and P_0, usually located between the normal sections, must be used.

1.2.2 Plane Coordinates. The geodetic coordinates also must be mapped onto a plane to provide plane coordinates for lower-order surveys and mapping. The most commonly used conformal projections of the ellipsoid to the plane are the transverse Mercator and the Lambert conic projections (see CARTOGRAPHY AND GEOGRAPHIC INFORMATION SYSTEMS).

In the transverse Mercator projection, points on the ellipsoid are projected onto a cylinder of elliptic base whose axis in the ellipsoidal equator passes through its center. The orientation of the cylinder is such that it is secant or tangent to a certain central meridian. Sectors of 1.5°–3.0° on both sides of the central meridian can be mapped onto the cylinder without exceedingly large scale distortions. The cylinder then is turned to the next zone (3°–6°) and the zone around the new central meridian is projected. Since the cylinder fits the ellipsoid along the central meridians, this projection is advantageous in the mapping of areas of long north–south extensions.

In the Lambert conic projection, points on the ellipsoid are projected onto a cone of circular base whose axis coincides with the rotation axis of the ellipsoid. The cone is set either tangent to the ellipsoid along a standard parallel or in a secant position with two standard parallels. Since the cone fits the area well along a single parallel or between the two standard parallels, it is advantageous in the mapping of areas of long east–west extensions.

The transverse Mercator is used in some 22 states in the United States, and also by many in Europe, Africa, and Australia. The Lambert projection is used by 30 states and by a number of Central and South American countries. Some states in the United States use both projections. Other projections could also be used in certain cases. For example, the stereographic projection could be employed to advantage for a small circular area (e.g., the polar regions).

2. DETERMINATION OF THE REFERENCE SURFACES AND GRAVIMETRIC GEODESY

As mentioned earlier, height measurements are referred to the geoid and planimetric measurements to a rotational ellipsoid, which in the area of the survey fits the geoid best. The three interrelated problems of geodesy treated in this section are

1. determination of the best-fitting ellipsoid,
2. determination of the geoid, and
3. relationships between spirit levelling and the geoid.

2.1 Geometric Methods

2.1.1 Geometric Determination of the Reference Ellipsoid. Figure 5 shows a meridian section of the reference ellipsoid. Two meridian arcs m and m' are shown with terminal points at latitudes φ_1, φ_2, φ'_1, and φ'_2. There exists a known functional relationship between these quantities and the semiaxes a and b of the meridian ellipse. If the lengths of two meridian arcs are measured and the geodetic latitudes at the ends are determined

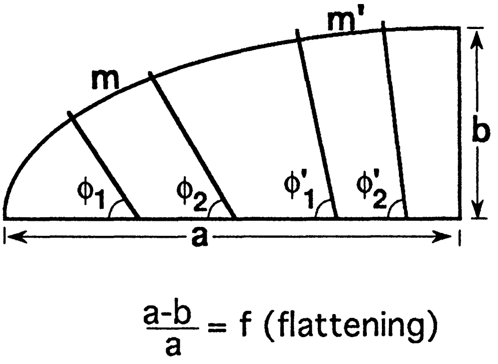

FIG. 5. Determination of the reference ellipsoid from two arc measurements.

(e.g., from astronomic latitude observations, assuming the first time around that there are no deflections of the vertical), one can obtain two equations from which the two unknowns *a* and *b* can be computed. The larger the difference in latitude, the more accurate will be *a* and *b*. In practice, the lengths of the several hundreds of kilometers long meridian arcs are measured indirectly by a chain of triangles and a base line. The shape of the triangulation chain is determined by measuring all angles, and the scale is derived from the base line and the base extension network.

A number of arc measurements were carried out in the eighteenth and nineteenth centuries according to the above principle. Several ellipsoids were derived. One of the best known ellipsoid determinations was made by Bessel in 1841. He used a total of 10 arc measurements: seven in Europe, two in eastern India, and one in Peru. In the adjustment he assumed the measured distances to be free of errors and derived the ellipsoid parameters in such a way that, along the arc, the differences between the astronomic and geodetic latitudes became minimum. The Bessel ellipsoid is still used although it is known today that the semimajor axis is too short by almost 1 km (6,377,397 m versus the current value of 6,378,137 m; and $1/f=299.15$ vs 298.257).

In the above-described method to determine the ellipsoid, the various arc measurements are considered independent, and the resulting ellipsoid(s) will fit well only the arc(s) used in the measurements. Better results can be obtained by the so-called area method. This method uses connected triangle chains or large networks where, in many points, the astronomical latitude, longitude, and azimuth are measured. The parameters of the best-fitting ellipsoid are then determined as follows: First measure and adjust all triangles of the network using the technique of least-squares adjustments. Then compute preliminary geodetic latitudes and longitudes on an approximate geodetic datum (an approximate ellipsoid, on which geodetic latitude and longitude of the initial point, as well as an initial azimuth, are defined arbitrarily, but within certain limits). The thus determined geodetic latitudes and longitudes, and the azimuths obtained from the adjustment, will deviate from the measured astronomical latitudes, longitudes (deflection of the vertical), and azimuths by small amounts, but systematically. Now change the approximate datum by changing the dimensions of the ellipsoid and shift and rotate the network by changing the initial point's geodetic latitude, longitude, and azimuth such that the new deflections of the vertical should become smaller and randomly distributed over the area. Here the area in consideration could be as small as an individual nation or as large as a continent. This gives an ellipsoid whose normals will coincide with the directions of the plumb lines as well as possible, and thus one which fits the geoid better. The deflections of the vertical are not observation errors, but are physical quantities. They are the result of the deviation of the level surface (geoid) from a rotational ellipsoid. The next section deals with how one can determine the deflection values.

2.2 Gravimetric Geodesy

2.2.1 Equipotential (Level) Surfaces and Gravity. Figure 6 shows some level surfaces of the Earth. They are by definition orthogonal to the plumb lines. Shifting masses along such a surface requires no work; thus the potential of the Earth's gravity field is constant on each of them. In contrast, moving a mass from one surface to another requires work that is independent of the path taken. If the potential of a certain equipotential surface is denoted by *W* (the work required to bring a unit mass from infinity to the surface in the gravity field), and one moves the unit mass from this surface to a somewhat lower equi-

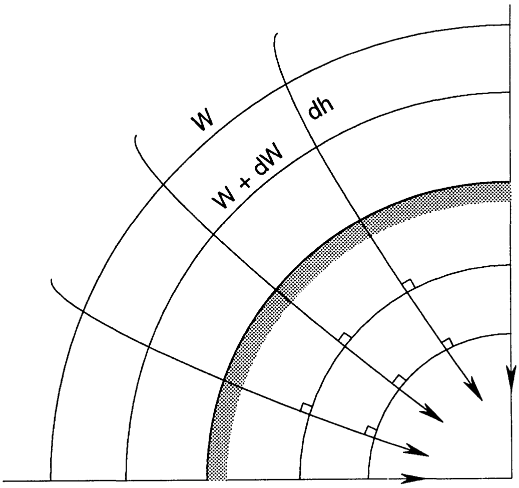

FIG. 6. Equipotential (level) surfaces in the Earth's gravity field and plumb lines.

potential surface $W+dW$, then $dW=-gdh$, where g denotes the gravity and dh the normal distance between the two equipotential surfaces. Thus, the gravity is the derivative of the potential with respect to the normal (plumb line). In an arbitrary orthogonal coordinate system the potential at point $P(x,y,z)$ is given by $W=W(x,y,z)$. The equation of the equipotential surface through P is $W(x,y,z)=\text{const.}$

These are the fundamental relations between gravity, its potential, and the equipotential surfaces. Since the gravity is measurable, it can be assumed that measurement of gravity permits the determination of the shape of equipotential surfaces and, since the geoid is an equipotential surface, also the determination of the shape of the geoid.

2.2.2 Normal Gravity and Gravity Anomaly. In an orthogonal coordinate system with its origin at the center of mass of the Earth, and the z axis coinciding with the rotation axis of the Earth, the potential at an external point P (x,y,z) is

$$W=G\int \frac{dm}{r}+\frac{\omega^2}{2}\,(x^2+y^2),$$

where G is the Newtonian gravitational constant, dm is the element of the Earth's mass, r is the distance of the mass element from P, and ω is the rotational speed of the Earth. If this equation is equal to a constant, one has the equation of the equipotential surface through P, or, with x,y,z as variables, the equation of any equipotential surface in the gravity field of the Earth. Unfortunately, the equation is not solvable in this form since we know neither the density distribution nor the shape of the surface over which the integration has to be carried out. Alternatively, the gravity potential at P is decomposed to two parts: U, the normal potential, and T, the disturbing potential. The normal potential is defined as the potential at P due to the attraction of a rotational equipotential ellipsoid (spheroid) that (ideally) has the same mass as the Earth and the same potential as the geoid. The center of this level ellipsoid is located at the center of mass of the Earth. Thus $W=U+T$.

The above equation is valid only for level surfaces located entirely outside the physical surface of the Earth. However, experience has shown that it is still usable in the proximity of the ocean surface, *i.e.*, for the geoid ($W=W_g$) and for its corresponding level ellipsoid, the terrestrial ellipsoid or spheroid ($U=W_g$), which approximates the geoid very well.

The normal gravity on the surface of the terrestrial ellipsoid (currently adopted parameters $a=6\,378\,137$ m, $1/f=298.257$, $GM=3\,986\,005\times10^8$ m^3/s^2, $\omega=7.292\,115\times10^{-5}$ rad/s) can be calculated rigorously from closed-form formulas if the geodetic latitude of the station is known.

The disturbing potential $T=W-U$ can be represented by the gravity anomaly $\Delta g=g-\gamma$, where g is the gravity reduced to the geoid, and γ the normal gravity on the terrestrial ellipsoid, both located on the plumb line of the observing station. The reduction of the measured gravity to the geoid is done using assumptions regarding the density of the masses out of the geoid.

2.2.3 Determination of the Geoid from Gravity Measurements. With the help of the gravity anomalies the distance between the geoid and the terrestrial ellipsoid (geoid undulations) can be determined. According to the formula of Stokes (spherical approximation) derived more than 100 years ago,

$$N=\frac{1}{4\pi}\frac{R}{\gamma}\int_\sigma \Delta g S(\psi)d\sigma,$$

where N=geoidal undulation, R=mean Earth radius, Δg=average gravity anomaly in the area element $d\sigma$, γ=average gravity of the Earth, $d\sigma$=differential area element, ψ=angular distance on the sphere between the area element and the computation point, and

$$S(\psi)=\frac{1}{\sin(\psi/2)}-6\sin\frac{\psi}{2}+1-5\cos\psi$$
$$-3\cos\ln\left(\sin\frac{\psi}{2}+\sin^2\frac{\psi}{2}\right)$$

(Stokes's function).

The integration must be carried out over the entire surface of the Earth. However, for political and practical reasons, gravity anomalies are not available all over the Earth. This fact complicates the calculations of the geoid enormously. In order to obtain gravity infor-

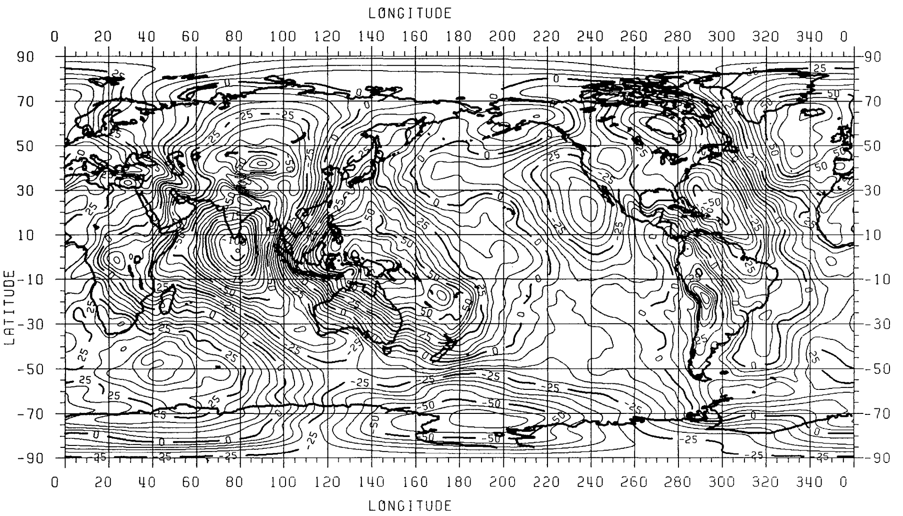

FIG. 7. Ohio State University 91A gravity model geoid heights (meters) relative to the GRS 80 reference ellipsoid (courtesy of Richard H. Rapp, Dept. of Geodetic Science and Surveying, Ohio State Univ.)

mation over unsurveyed areas, such as the polar regions, the orbital perturbations of a large number of satellites are analyzed and isostatic theories are used. Over the oceans, gravity anomalies are predicted or deduced from sea surface heights observed by altimetry satellites.

The geoid determination is, at present, possible to an accuracy of about 20 cm to 1 m. A global gravimetric geoid is shown in Fig. 7.

Differentiation of Stokes's equation with respect to a meridional and a parallel length increment yields the north-south and east-west components of the deflections of the vertical respectively.

2.2.4 Leveling, Orthometric Heights, and the Geoid. Once the geoid is determined, geometric levelling, starting from the geoid, is used to give the height of points located on the Earth's surface. For example, to determine the height of location *B* in Fig. 8, by geometric levelling starting at location *A*, the height differences dh_1, dh_2 between the level surfaces at subsequent rod positions are measured. Since the height of *B* is defined as the length of its plumb line down to the geoid (orthometric height), one actually should measure the quantities dh_1', dh_2',.... Because the work to proceed from one level surface to the other is the same for either location, one has the following relation between *dh* and the corresponding dh': $g\,dh = g'\,dh'$, where *g* is the mean gravity along the distance *dh*, and *g'* the mean along dh'. It follows that Σdh, the leveled height (difference) of *B*, is slightly different from $\Sigma dh'$, the correct orthometric height (difference) of *B*. Hence, a correction to the measured height, called the orthometric correction, should be applied in order to get the correct orthometric height. This correction is a function of the gravity values (*g*) along the path on the Earth's surface from location *A* to location *B* (see Fig. 8) and along the plumb line (*g'*) at location *B*. The latter can only be estimated on the basis of models of the structure of the Earth's crust. Therefore, in precision levelling, the gravity along the level lines is measured by means of a gravimeter. If a gravimeter is not available, as an approximation gravity values based on the reference ellipsoid (normal gravity) are used.

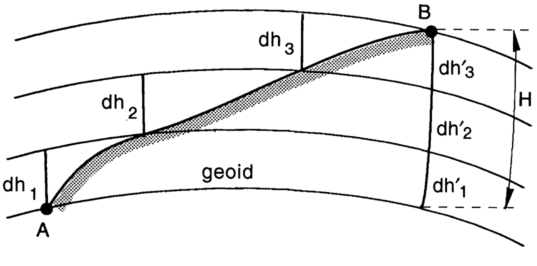

FIG. 8. The effect of the convergence of the level surfaces on geometric levelling.

3. MODERN CARTESIAN THREE-DIMENSIONAL GEODESY

3.1 Preview

The techniques described in the previous sections were developed either for horizontal (latitude, longitude) or for vertical positioning (height above sea level). These coordinates cannot be combined into three-dimensional ones of the same accuracy, because the geoid-ellipsoid separation component of the vertical coordinate (the geoid undulation) is not known with the same accuracy as either the horizontal coordinates or the height above sea level. This, together with the fact that the direct applicability of most traditional techniques is limited by the line of sight, i.e., by the curvature of the Earth, made the development of other technologies, not hampered by these problems, desirable for a number of traditional and heretofore not attempted geodetic applications. Satellite, radio astronomic, and inertial survey techniques are the ones that seem to be most suitable for such purposes. In the following these techniques together with some applications are described.

As the rapidly emerging industrial technology develops, more precise geodetic information can be achieved. For some areas of the world there still is little or no observational material. For these areas, additional or new surveys are required to relate one datum to another more precisely and to develop a standard world reference system that will satisfy future needs. Space technology is providing useful techniques for such surveys.

The theoretical use of artificial Earth satellites for obtaining geodetic information has been studied for some time. Scientific papers advocating the use of satellites for geodetic purposes have been published since 1956, and geodetic purposes were outlined for data obtained from Project Vanguard, the first U.S. satellite of geodetic significance, during the 1958–59 International Geophysical Year (Kozai, 1959; Kaula, 1966).

Many techniques and a great deal of knowledge were ultimately derived from this project. With this information, the constant growth of space technology, and the development of electronic measuring devices, it has become possible to plan and develop other satellites specifically equipped for geodetic purposes.

The feasibility of using pulsed lasers to range to artificial Earth satellites was first demonstrated in 1964 with the returns from the first satellites with laser retroreflectors. Since that time many reflector-equipped satellites have been launched and tracked. Two satellites, Starlette (France) launched in February of 1975 and LAGEOS (U.S.) in May of 1976, were specifically designed for precise laser tracking and Earth dynamic studies. Present-day lasers have demonstrated accuracies of 2–50 cm for a single range and about 5–10 parts in 10^9 in distance precision over baselines of 900–3000 km.

Our natural satellite, the moon, has also become a target for precise distance measurements, after lunar laser reflectors had been deposited on its surface.

Interferometric observations of signals from extragalactic radio sources (quasars) are also useful geodetic tools. The most widely used technique, very long-baseline interferometry (VLBI), was first developed in the 1960s for studies of source structure and size. The technique has been applied since to several other fields such as radio astrometry and relativity studies as well as to geodesy and geophysics. With interferometry, angular resolution is achieved by comparing signals from two widely separated antennas. The limiting factor on the resolution is the distance between the antennas and the size of the antenna. Earlier interferometers needed a physical communication link (e.g., cable) to transmit signals from one antenna to the other; therefore, the extent of antenna separation was naturally limited. This problem was solved by the development of VLBI, where the real communication link is bypassed by the use of highly stable atomic clocks at each radio telescope. The clocks are used to control the recordings of the radio signals on digital type tape recorders. These tapes then are transported to a common location where in essence the event is replayed and signals are compared by a computer. Because of the volume of the signals received and the complexity of the operation, the computer requirements are extensive.

In addition to the techniques mentioned above, three-dimensional coordinates may also be obtained by inertial survey systems. The principle of these systems, first developed for navigation, is that changes in position can be determined by the second integral of acceleration. Accelerometers measure specific force, which is the combined effect of acceleration and gravity: therefore gravity must be known (measured or calculated) to obtain acceleration. The orientation of the specific force is determined by gyroscopes. A computer keeps track of the orientation with respect to the reference frame of interest and performs the integration in the coordinates. The end product is two- or three-dimensional relative coordinates, depending on the number and direction of accelerations measured. The systems designed for geodetic use are mounted either in ground vehicles or in helicopters. Because of the cost of the observing equipment, vehicle, and computer operations, the use of these techniques so far has been mostly limited to government agencies and a few commercial organizations. The introduction of ring-laser gyros and fiber optic technology will aid in making available more precise, less expensive inertial systems.

3.2 Satellite Geodesy

3.2.1 Introduction. The scientific purpose of satellite observations falls into either dynamic or geometric categories. The dynamic purpose of satellite tracking is to observe positions and motions of satellites as a function of time, with sufficient accuracy for developing a theory of motion capable of predicting future positions at least as accurately as they can be observed. In order to make these predictions, an extensive theory of the motion, the precise knowledge of the physical parameters defining the force fields in which the satellite moves, and the accurate geocentric position of the observer are needed. Since, at least initially, the physical parameters and the tracking station positions are not known adequately, even when the theory of motion is sufficiently extensive, the observed positions and motions of the satellite will differ from the predicted ones. From these differences, corrections to the assumed parameters and station coordinates may be calculated, generally by means of a least-squares adjustment procedure.

In the area of celestial mechanics, a branch of classical mechanics, many improvements on the theory of motion have been developed during the centuries. The best mathematical minds of humanity devoted their energy to this problem, but mostly in connection with planetary motions. The circumstances of an artificial Earth satellite are different from those of an ordinary planet in two main respects: its orbit is much closer to the Earth than the orbit of any planet with respect to its primary, and it moves, at least partially, in an atmosphere while the natural satellites move practically in a vacuum. The first fact implies that the motion of an artificial satellite is significantly affected by the asymmetries of the Earth's gravitational field and by its temporal variations (e.g., tides), and therefore it could be used to determine the parameters defining this field. These parameters in turn may yield information on the shape, the mass distribution, and the dynamic behavior of the Earth. The second fact opens up possibilities to determine the structure of the atmosphere and to provide an insight to its behavior. There are also other phenomena that will influence the motion, such as the attraction of the sun and the moon, radiation pressure emanating from the sun, lunar and solar tidal distortions, and effects of the magnetic field of the Earth.

For principal geodetic results, the observables are the difference between transmitted and received frequencies [called the Doppler frequency after the Austrian Christian Doppler (1803–1853) who first demonstrated the principle using sound waves], range, or direction components (e.g., right ascension and declination), either observed individually or in certain combinations. Sometimes the instrumentation will collect integrated Doppler values that can be related to range changes. Receivers designed to track signals from the Global Positioning System (GPS) satellites can measure Doppler, phase delay, or advance of the carrier when compared to the phase of the receiver's oscillator, and pseudorange, which is the speed of light multiplied by the difference between transmit time based on the satellite's clock and received time based on the receiver's clock.

3.2.2 Principal Methods

3.2.2.1 General Dynamic Method. In the simplest case, when a spherical satellite moves around its spherical primary, affected only by the latter's attraction, the resulting normal orbit may be defined by six parameters, E_i ($i=1, 2, \ldots, 6$). These parameters may be selected according to various theoretical and computational criteria. From the didactic point of view, the simplest set is the classical Keplerian elements (Fig. 9), which define the orientation of the plane of the orbit in space (two parameters), the size, shape, and orientation of the Keplerian ellipse in that plane (three parameters), and finally, the position of the satellite at some given epoch (one parameter). In the gravitational force field of the spherical primary, five of these elements are constant and the satellite moves in its defined orbit in accordance with the laws of Kepler. As mentioned, the circumstances of a near-Earth artificial satellite are different from the above, the main consequence being that the orbital elements will no longer be time invariant, or, in other words, their derivatives with respect to time, $\dot{E}_i$, will not be zero. The variation of a given element from some reference epoch T_0 to the epoch of utilization T can symbolically be described by the equation

$$E_i = E_i^0 + \int_{T_0}^{T} \dot{E}_i dT,$$

where E_i^0 is the element in question at the epoch T_0 and E_i is that at T. The integral represents the perturbation of the element E_i. The function $\dot{E}_i$ is the rate of change of the element E_i due to all perturbing forces (nonspherical part of the Earth's attraction, atmospheric drag, etc.) and as such is a function of several hundred parameters, P_j ($j=1,2,\ldots,n$), defining these forces. For example, an adequate description of the Earth's gravitational field may require as many as 500, the atmospheric drag 10, and the radiation pressure 5 force parameters. The integral may be solved analytically (method of general perturbations) or numerically (method of special perturbations).

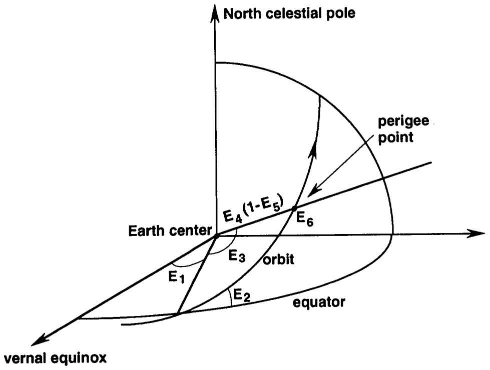

FIG. 9. Keplerian orbital elements: Right ascension of the node (E_1), inclination (E_2), argument of perigee (E_3), semi-diameter of the elliptic orbit (E_4), eccentricity of the ellipse (E_5), time of perigee passage (E_6).

Once the orbital elements are thus computed at the epoch T, they can be readily converted into positional and velocity components of the satellite and referenced to some well-defined coordinate system. If the parameters defining the observer's position in this coordinate system, S_k ($k=1,2,3$), are also known, then the observables can be calculated, i.e., predicted. To refer the positions of the satellite and the observer to the same (usually Earth-fixed) coordinate system, the precise knowledge of the Earth's rotation (precession, nutation, polar motion, and time) is also required.

Provided that the theory of motion is correct and that the observations have been freed of systematic errors, the differences between the computed, C_l ($l=1,2,\ldots,m$), and the observed, O_l, values of the observables will be due to the erroneous geocentric coordinates of the observer S_k and the erroneous orbital elements at epoch T_0 that are included along with the parameters describing the force field in the vector components P_j. Assuming that both the differences $O_l - C_l$ and the errors dP_j, dS_k are differentially small, then the following type of relations may be established:

$$O_l - C_l = \sum_j \frac{\partial C_l}{\partial P_j} dP_j + \sum_k \frac{\partial C_l}{\partial S_k} dS_k ,$$

where

$$\frac{\partial C_l}{\partial P_j} = \frac{\partial C_l}{\partial E_i} \frac{\partial E_i}{\partial P_j} ,$$

E_i being an orbital element at the instant of observation.

If the above equation is regarded as an observational equation, then the quantities dP_j and dS_k are the unknown corrections to the assumed parameters and station coordinates, respectively, and $\partial C_l/\partial P_j$ and $\partial C_l/\partial S_k$ are coefficients, representing the variations of the observables with respect to P_j and S_k, computed from the prediction mathematics. The left side of the equation, $O_l - C_l$, is the known difference between the observation and its predicted value. In order to obtain a satisfactory solution, the number of observations m must greatly exceed the number of unknown parameters n plus the station unknowns $p \times k$ (p is the number of stations), and a least-squares adjustment is performed in the traditional sense. To perform such a calculation is a formidable task, considering that recent solutions, for example, included data from over 30 satellites observed over two- to four-week periods from as many as 50–100 stations. Thus, the unknowns may include 150–300 components of positions in addition to about 2000 harmonic coefficients of the Earth's gravity field, about 90 tidal coefficients, thousands of orbital constants (e.g., E_i^0), possibly Earth rotation parameters, etc. Such general solutions, because of the high cost of solving the large number of equations, are performed infrequently, and only by a few organizations having access to large computers.

3.2.2.2 Partial Dynamic Method. Once the results of a general solution, such as the one outlined above, are available, additional stations can be added on at a much smaller cost. In such partial solutions, positions of observing stations are obtained from least-squares adjustments, where the Earth's gravity field and the positions of many of the stations are held at values determined in the preceding general solution. Usually a separate computer program is used for this purpose because program efficiency is greater when the objectives are more limited. In these solutions a shorter time span of data (two to five days) may be used, which will also reduce the effect of some errors in the force field. The unknowns in such a solution, in addition to the coordinates of the new stations, include the six initial orbital constants E_i^0, usually a drag (scaling) parameter, and maybe components of the pole position. In other words, in such solutions most perturbations are treated as known phenomena that can be calculated from the parameters determined in the general solution.

3.2.2.3 Short-Arc, Point-Positioning, and Translocation Methods. Two special cases of the partial solution are the so-called short-arc and the point-positioning methods. In the former, the data are limited generally to satellite passes of 10–30-min lengths. These relatively inexpensive solutions contain as unknowns only six initial orbital parameters per pass and the coordinates of the participating stations. The method is limited to relative positioning with respect to a reference station whose coordinates are held to their estimated (not necessarily geocentric) values. Because of the shortness of the arc, the perturbation models may be simpler than in the partial solutions; for example, the gravity field may be satisfactorily described by as few as 50–75 parameters, depending mainly on the length of the arc and the altitude of the satellite. The relative station positions obtained from short-arc solutions are considered generally free of orbital (or other) biases equally affecting stations observing the same arcs.

In the point-positioning method the data span is several days long for Doppler satellites and less than that for GPS satellites, and the orbital elements are held to the values obtained from a precise satellite trajectory (ephemeris). Thus, the only unknowns in the solution are the coordinates of the observing stations and their clock offsets or frequency drifts. The satellite ephemeris is generated and made available by some organization that keeps continuous track of the satellites in question. A prime example of this method is positioning by use of instruments that measure the range difference between a ground station and two satellite positions by means of integrating the Doppler shift of radio transmissions from the Navy Navigational Satellites (NNS). Precise ephemerides of these satellites are generated, from which satisfactory orbital elements may be obtained for the instants of the observations. Predicted, and therefore less accurate, elements are also generated and injected into the satellite's memory, which in turn are retransmitted and can be used with certain types of receivers. In this latter case it is advisable to observe the satellite passes from at least two stations simultaneously and to solve only for relative positions that, similarly to the short-arc case, will be less affected by the biases in the orbital elements than the positions themselves. This mode of operation is termed translocation.

3.2.3 Applications of Satellite-Derived Geodetic Parameters. For the geodesist it is important to know the positions of a large number of well-distributed stations, especially if these positions are determined accurately in some homogeneous way and are referred to a well-defined and accessible coordinate system. Even with the present modest accuracies, several immediate applications come to mind. In addition to the most obvious one, namely, to aid in the accurate determination of orbits, some of the most important geodetic applications are as follows:

1. The gravity harmonic coefficients obtained from general dynamic solutions provide an accurate long-wavelength representation of the Earth's gravity field, and hence the geoid. This long-wavelength information is combined with gravity anomaly data, obtained from terrestrial gravity observations (over the continents) and satellite altimetry data (over the oceans), to give the most accurate global geoid known at present.
2. Three parameters—*GM*, the geocentric gravitational constant, a, the equatorial radius of the Earth, and J_2, the second zonal harmonic of the gravity field—together with ω, the rotation rate of the Earth, contain enough information to define the size and shape of a best-fitting geocentric equipotential Earth ellipsoid (spheroid). The orientation of this ellipsoid is also defined through satellite observations, as well as astrogeodetic ones, and maintained by some international agencies, such as the International Earth Rotation Service (IERS). This ellipsoid serves as a global Geodetic Reference System and is modified from time to time (e.g., GRS67 and GRS80) by the International Association of Geodesy. The values are published each quadrennium in its *Geodesist's Handbook*, as a special edition of the journal *Bulletin Géodésique*.
3. An independent geodetic datum can be related to others by determining transformation parameters between them and a common satellite system. The problem of estimating datum transformation parameters is very often an ill-conditioned one. Therefore a careful analysis is needed in order to minimize the correlations between the estimated transformation parameters.
4. Possible distortions within a geodetic network can be detected by inspecting the residuals after systematic components arising from the above transformation are removed. These distortions may be due to various errors that are partly observational and partly computational in nature. This application presumes that the satellite-determined positions are more accurate and homogeneous than those established by ground surveys.
5. A primary geodetic network can be strengthened by incorporating accurate satellite (Doppler, laser, or GPS) positions in the adjustment process.
6. Mapping control can be established in unsurveyed areas.
7. Geodynamic activity can be monitored. Some of the geodetic space techniques have the fundamental capability of accurately monitoring the motion in space of a point on the Earth's surface or the capability of measuring the changes in the distance between two or more points. The present capabilities make it possible to measure with unprecedented accuracies certain important geodynamic phenomena, such as tectonic plate motions, deformations around fault zones, Earth tide displacements, temporal changes in the Earth's gravity due to the post-glacial rebound, sea surface variability and apparent positional changes due to polar motion, and variations in the rotational rate of the Earth. Some of these phenomena have been measured and monitored in the past by means of traditional instrumentation. Prime examples are the polar motion and Earth rotation efforts coordinated by the International Polar Motion Service (IPMS) and the Bureau International de l'Heure (BIH), based on optical astronomic techniques such as use of zenith telescopes and astrolabes. Doppler satellite techniques also provided information on the motion of the pole. These techniques allow the determination of polar motion to an accuracy of 50–100 cm and of Earth rotation to 0.7–1 ms over two to five days averaging time (Mueller and Zerbini, 1989).

   The most accurate measurements of geodynamic phenomena are obtained by the very long-baseline interferometers, the lunar laser rangers, the satellite laser rangers, and, recently, the GPS satellites. These systems have already demonstrated the following accuracies: plate and fault motions, 1–2 cm/yr or better; polar motion, 2–3 cm over 12 h averaging; and Earth rotational rate, 0.05 ms over 12 h averaging.
8. A Terrestrial Reference Frame for Earth Dynamics may be established. The expected increase in observational accuracy and the resulting increase in the resolution of geophysical phenomena requires a redefinition of many of the concepts used in the more traditional analysis described ear-

lier, in which, with the possible exception of the Earth tides phenomena, the Earth was assumed to be a rigid body. It is the departures from rigidity that are interesting and are to be studied in suitable reference systems needing definition. The realization of such a reference system must necessarily use observations from the Earth, and it could be defined through the adopted coordinates of super control stations, determined with high precision from laser observations to satellites or to the moon, VLBI observations, and GPS phase observations. The following criteria seem desirable for such a coordinate system:

- **a.** The system should be as nearly invariant as possible with respect to changes in the number, distribution, and data acquisition rates of observing stations in different parts of the world.
- **b.** The system should facilitate the rapid determination of sudden changes in the position or motion of individual stations, as well as sudden changes in polar motion and the rotation of the Earth.
- **c.** The system should be approximately fixed with respect to the mantle (should rotate with the Earth) in some average sense.

These criteria lead to the choice of a coordinate system in which our previous knowledge of the motion of the stations with respect to each other and to the mantle is used to model the station motions due to plate movements, etc. Such a coordinate system conceptually would be similar to that used by astronomers in the fundamental star catalogs, where the adopted coordinates of stars together with their relative motions define the frame of reference. To maintain such a system the International Earth Rotation Service (IERS) was established in 1988 in Paris, France; its responsibilities include coordinating the continuing observations at the fundamental stations and issuing directives on changes in the adopted coordinates and motion parameters or models, when the need for such changes arises from the data analyses. An international terrestrial reference frame is then realized through the coordinates of the fundamental stations at some reference epoch.

3.2.4 Special Satellite Geodetic Techniques

3.2.4.1 Global Positioning System (GPS). The GPS has been designed to provide for passive, all-weather, precise navigation from any location on or near the Earth at any time. The GPS satellites will be operated by the U.S. Air Force for the U.S. Department of Defense. When fully operational, the GPS constellation of satellites will consist of 21 (and three spare) satellites, equally spaced in three orbit planes, with 12-h periods. This will allow visibility of at least four satellites in good geometry for any point on Earth. The GPS satellites will employ precise atomic frequency standards to assure high oscillator stability and precise time information. Each satellite will broadcast on two L-band frequencies, L_1 (1575.42 MHz) and L_2 (1227.60 MHz), which will allow elimination of first-order ionospheric refraction. Additional information will be modulated onto these L-band frequencies for the following purposes: to identify each satellite uniquely; to provide for measurement of the signal travel time to the receiver by measuring the phase shift required to correlate the codes to receiver-generated codes; and to allow restriction of the use of the system if desired. The master control station is located at Colorado Springs, Colorado, where all operational tracking data are collected, with precise orbital information (predictions) then being made. This information is then loaded into the satellite memory at least once per day as the broadcast message for the satellite. The three major observables of the GPS satellites are pseudorange, phase difference, and phase difference rate (Doppler). The pseudorange is basically a range observation from the receiver to the GPS satellite, but since the receiver's clock and the satellite's clock are not synchronized, this range will be contaminated with clock errors. Resection of ranges to four different satellites will provide the three-dimensional position of the receiver as well as an estimate of the receiver clock offset to GPS system time. This will be the primary method for navigation. The differences between phase observables from different stations to different satellites will be free of most of the systematic errors affecting the original observables, such as the clock errors of the satellites and receivers, and hence these differences are the primary geodetic observables. The Doppler observable is the weakest and the least used in geodesy but will be useful if velocity estimates are required.

The GPS observations have already demonstrated accuracies approaching one part in 10^9 in measuring distances of a few thousand kilometers, and are used heavily in monitoring crustal deformations.

3.2.4.2 Satellite Altimetry. A short burst of electromagnetic waves is directed downward from the satellite, and the return signal, reflected from the sea surface, is monitored by a fast-response instrument. The resulting signal's delay time and shape are then used to determine the vertical distance from the satellite to the ocean surface and back to the detector. If the satellite's position is known from other tracking data, the sea surface height can be determined. If the geoid is also known, the separation of the sea surface and the geoid can be determined, a phenomenon called the sea surface topography, which is of great interest to oceanographers. Altimeters with a precision of approximately 50–100 cm have flown on Skylab and GEOS3. An improved altimeter with 10-cm precision operated successfully on the SEASAT 1 satellite for a few months until the spacecraft failed in October 1978. The most ambitious TOPEX ocean sensing satellite, with sub-10-cm altimeter, was launched in summer, 1992.

3.2.5 Lunar Laser Ranging (LLR). The goal of LLR is to use ranges to the moon observed from some stations on the Earth in order to improve the knowledge of the lunar orbit and of the positions of the lunar reflectors as well as the observing stations. Also, it is desired to study gravitational theory and the rotation of the moon by analysis of position deviations of the moon remaining after modeling known parameters. Retroreflectors were placed on the moon by the crews of Apollo 11, 14, and 15. Two other reflectors were placed by the Soviet Lunar 17 and 21 missions. Laser ranging to most of these reflectors began shortly after each of these missions, and is still continuing. The arrangement of the four working reflectors on the moon is a quadrilateral with sides of about 1000 km, a good length for the study of lunar rotation.

Although the great majority of lunar ranging observations to date have been made at only one observatory (McDonald Observatory, at Fort Davis, Texas), a remarkable number of accomplishments have been made by the lunar laser ranging program, which are not elaborated here.

3.3 Very Long-Baseline Interferometry

3.3.1 Introduction. The primary very long-baseline interferometer (VLBI) observable for geodetic and astrometric studies is the time interval between the arrival of a radio signal at one end of the interferometer and its arrival at the other end. This interval is called the delay, and its time derivative is the delay rate observable. From a sufficient set of these data as functions of time, the geometry of the interferometer baseline and the position of the observed radio sources can be determined.

3.3.2 Basic Instrumentation. The basic geometry of a VLBI interferometer is shown in Fig. 10. Each observing station receives and records independently the signals from a radio source. The time of arrival of the signals in

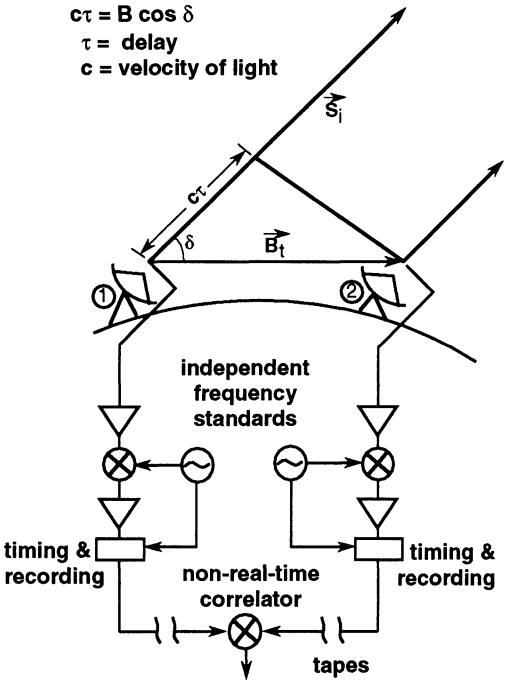

very long baseline interferometry (VLBI)

FIG. 10. Basic geometry of radio source interferometry.

terms of the local clocks is also recorded. In the simplest case the signal from the source is a single plane wave and the observing sites are fixed in space. The difference in recorded arrival time (τ) is then a direct measure, in light-time ($c\tau$), of the component of the interferometer baseline in the direction of the source ($B\cos\delta$). Measuring delays from sources located on orthogonal planes in space completely determines the baseline length and orientation. Two additional factors must now be taken into account. As the Earth rotates in space, the projected baseline in the direction of the source changes with time, introducing the delay rate observable. As the stations move, the signals recorded by each station are Doppler shifted. These effects must be removed before the recordings can be correlated to determine the delay observable.

VLBI instrumentation can be divided into the following main components: antenna, receiver, frequency standard, and recorder.

Antennas used for VLBI have varied in size from the 305-m bowl at Arecibo, Puerto Rico, to a 4-m dish of the Jet Propulsion Laboratory. The sites have been distributed geographically as widely as Australia and the Crimea.

Commercial rubidium and cesium beam clocks and hydrogen masers have been used as VLBI frequency standards. Rubidium clocks have fractional frequency stabilities of parts in 10^{11} while hydrogen masers can be two or more orders of magnitude better. Two time scales are important. Over periods of a few minutes frequency stability must be adequate to provide sufficient integration time for extracting the VLBI observables. Over periods from several hours to a day, the frequency standard must function as a stable clock for recording signal arrival time.

3.3.3 Applications. VLBI delay observations are composed of three components. The first component is due to the geometry of the station locations and source location, as illustrated in Fig. 10. This component contains all of the information about the geodetic and astrometric parameters that are of interest. The second component is due to instrumental effects, principally clock errors. The third component is due to effects of the propagation medium (e.g., the atmosphere and the ionosphere). Only the geometric component is considered here.

From Fig. 10 we see that the geometric delay τ can be written in vector form as

$$\tau c = -\mathbf{B}_t \cdot \mathbf{R}\mathbf{S}_i ,$$

where $\mathbf{S}_i$ is the unit vector in the direction of the radio source in an inertial coordinate system, $\mathbf{B}_t$ is the vector separation of the receiving stations, usually modeled in an Earth-fixed coordinate system, $\mathbf{R}$ is a transformation matrix that accounts for the rotations of the Earth in space, c is the speed of light, and τ is the time delay observable. Hence, $\mathbf{R}$ will contain parameters of the precision, nutation, and polar motion of the Earth. The vector $\mathbf{S}_i$ is a function of the right ascension and declination of the radio source, while the vector $\mathbf{B}_t$ has Δx, Δy, and Δz as components.

If sufficient observations are taken, one can solve for all the above-mentioned parameters in a well-defined Earth-fixed and celestial coordinate systems. The estimated parameters may include baseline components, nutation parameters and polar motion, radio source right ascensions and declinations, and clock drifts and drift rates.

If the observations made are delay rates rather than delays, a slightly different situation emerges. It is not possible with delay rates alone to estimate all of the parameters of geodetic and astrometric interest. The polar component of the baseline cannot be estimated, and, in addition, at least one of the baseline coordinates or source coordinates must be determined by some other means.

4. CONCLUDING REMARKS

With the introduction of artificial satellites in the late 1950s, the geodetic sciences have been under continuous and rapid change. Today the impacts of very high-quality measurement systems such as lasers, very long-baseline interferometry, precise absolute gravimeters, satellite-based altimetry, and the Global Positioning System (GPS) are bringing allied sciences such as geophysics, oceanography, remote sensing and photogrammetry, and navigation into much closer contact. For example, one would not consider using traditional measurements such as angles and distances from theodolites and electronic distance-measuring devices to establish a modern national geodetic network as was done as recently as 1987. Today the increased precision with greatly reduced costs of GPS, strengthened by satellite laser ranging and/or VLBI, would be almost the exclusive techniques employed. Real-time millimeter measurements are being collected in search of earthquake and volcanic precursors. Similar measures of ice-sheet thicknesses and sea-level monitoring along with precise satellite altimetry are opening up new vistas for studying energy transfers between oceans and the atmosphere, ocean circulation, climate, and internal Earth structure.

GLOSSARY

Anomaly, Gravity: The difference between the gravity on a geop (defined below) and the normal gravity on the corresponding spherop (defined below). The gravity anomaly is generated by masses deviating from those of the normal Earth. The effects of these disturbing masses on the gravity anomaly are given by the principal equation of gravimetric geodesy. The sign of the anomaly is in the sense observed minus normal. In the conventional theories the geop used is the geoid (defined below).

Clairaut's Theorem: An expression for the variation of the normal gravity on the Earth spherop, which is the basis of standard gravity formulas such as the international gravity formula. It also establishes the relationship between normal gravity and flattening of the Earth spherop that makes it possible to compute the flattening from surface gravity observations.

Deflection of the Vertical: *Absolute deflection* is the angle at a point on the geoid between the normal to the geoid at this point and the normal to the Earth spherop at the corresponding point. It may be determined from gravity values evenly distributed over the whole of the Earth surface.

Doppler Effect: The change of frequency of sound, light, or radio waves with the relative velocity of the source and the observer: if the source and observer are moving closer together, the observed frequency is higher than the transmitted frequency.

Ellipsoid, Best-Fitting: A reference ellipsoid whose size, shape, and orientation have been determined from the condition that it should fit best to the geoid or to parts of it. The criteria for best fit vary. Usually it is required that the square sum of the astrogeodetic deflections (or their deviations from the absolute deflections) or of the geoid undulations, extended for the area in question, should be a minimum.

Ellipsoid, Level: An equipotential surface of ellipsoidal shape. If the Earth were a rotating fluid of uniform density, its shape would be a rotational ellipsoid.

Ellipsoid, Reference: Any one of arbitrarily selected ellipsoids used in geodetic calculations to represent the Earth (International ellipsoid, Clarke's ellipsoid of 1866, GRS 80, and so forth).

Ellipsoid, Rotational: The figure generated by rotating an ellipse about its minor axis. It is defined by two parameters (two semidiameters or one semidiameter and the flattening, and so forth). Also called *ellipsoid of revolution.*

Flattening (Rotational Ellipsoid or Ellipse): The ratio of the difference between the major and minor semidiameters to the major semidiameter.

Flattening, Gravitational: The ratio of the difference between the polar and equatorial normal gravities to the equatorial normal gravity.

Geographic Position: The geographic position of a terrestrial point is its position relative to a geodetic reference surface, which is used in mapping as a substitute for the actual shape of the Earth. The most often used reference surfaces are an ellipsoid and a geopotential surface. The position usually is given in curvilinear coordinates, such as latitude, longitude, and height above the reference surface. It is customary to distinguish between ellipsoidal (geodetic) and gravitational (astronomic) coordinates depending upon whether the ellipsoid or the geopotential surface, respectively, is used as a reference surface.

Geoid: That equipotential surface of the Earth's gravity potential (geop) that most nearly coincides with the undisturbed mean sea level.

Geop: An equipotential surface in the gravity field of the Earth. On such a surface the geopotential is a constant and the gravity vector is perpendicular to it at every point. Also called *geopotential surface.*

Geopotential: The gravity potential of the actual Earth. The sum of the gravitational (attraction) potential and the potential of the centrifugal force. A function describing the variation of the geopotential in space. The function whose partial derivative in any direction gives the gravity component in that direction.

Gravity: The resultant of gravitation and the centrifugal force caused by the rotation of the Earth. The vertical gradient of the geopotential. It may be determined by methods such as reversing pendulum, long pendulum, upward projected ball, free fall, free and nonfree fall, and so forth. Also called *absolute gravity.*

Gravity, Normal: The gravity caused by the mass attraction of the normal Earth (defined below) combined with the centrifugal force due to its rotation. The normal gradient of the spheropotential.

Height, Orthometric: The distance between the geop through the station in question and the geoid, measured along the plumb line of the station.

Latitude, Geodetic: The angle between the ellipsoidal normal to the point in question and the equatorial plane of a reference ellipsoid. Positive in the Northern Hemisphere.

Longitude, Geodetic: The angle between the geodetic meridian plane of the observer and that of an arbitrary point near Greenwich. It is measured in the equatorial plane of the reference ellipsoid and numbered usually from the point near Greenwich. It is determined, however, by means of triangulation, trilateration, or similar methods from the geodetic meridian plane of the initial point of the geodetic datum to which it refers. Also called *geocentric longitude.*

Meridian, Geodetic: A curve (ellipse) on the reference ellipsoid that contains points of equal geodetic longitude.

Normal Earth: A mass such that its external bounding equipotential surface is the spherop (defined below), and its gravity is given by the prescribed gravity formula.

Spheroid: The spherop that possesses, in the normal gravity field of the Earth, the same constant of potential as the geoid in the true gravity field.

Spherop: An equipotential surface in the normal gravity field of the Earth. On such a surface the spheropotential (defined below) is constant. The normal gravity is perpendicular to it at every point. It may be shown that, neglecting terms smaller than the second power of the flattening, the surface of a spherop is identical to that of a rotational ellipsoid. Also called *spheropotential surface.*

Spheropotential: The potential function of the normal gravity. It may be defined either as the external gravity potential of a level ellipsoid approximating the Earth, or as a function consisting of the first few terms of some expansion of the geopotential.

List of Works Cited

Kaula, W. M. (1966), "Tests and combination of satellite determinations of the gravity field with gravimetry," *J. Geophys. Res.* **71**, 5303–5314.

Kozai, Y. (1959), "The motion of a close earth satellite," Astron. J. **64**, 367–377.

Mueller, I., Zerbini, S. (Eds.) (1989), *Interdisciplinary Role of Space Geodesy,* Heidelberg: Springer-Verlag.

Torge, Wolfgang (1980), *Geodesy: An Introduction,* New York: de Gruyter.

Further Reading

Only cursory discussions could be given in this encyclopedia article. So additional texts are given to aid the reader in searching for more detailed discussion of the geodetic sciences.

General Overview

Mueller, I. I., Ramsayer, K. H. (1979), *Introduction to Surveying,* New York: Ungar.

Plane Surveying

Moffit, F. H., Bouchard, H. (1992), *Surveying,* 8th ed. New York: Harper.

Photogrammetry

Wolf, P. R. (1974), *Elements of Photogrammetry,* New York: McGraw-Hill.

Geodesy

Bomford, G. (1971), *Geodesy,* 3d ed. Oxford: Clarendon.

Geodetic Astronomy

Mueller, I. I. (1969), *Spherical and Practical Astronomy As Applied to Geodesy,* New York: Ungar.

Gravimetric Geodesy

Heiskanen, W. A., Moritz, H. (1961), *Physical Geodesy,* San Francisco: Freeman. (1979 reprints available from Department of Geodetic Science and Surveying, The Ohio State University, Columbus, Ohio 43210-1247.)

Satellite Geodesy

Kaula, William M. (1966), *Theory of Satellite Geodesy,* Waltham, MA: Blaisdell.

Geophysical Geodesy

Lambeck, K. (1988), *Geophysical Geodesy: The*

Slow Deformations of the Earth, Oxford: Clarendon.

Map Projections

Pearson, F. (1990), *Map Projections: Theory and Applications*, Boca Raton: CRC Press.

Plane Rectangular Coordinate Systems

Adams, O., Claire, C. N. (1948), *Manual of Plane-Coordinate Computations*, U.S. Coast and Geodetic Survey, Spec. Pub. No. 193, Washington, D.C.: U.S. GPO.

Cartography

Robinson, A., Sale, R., Morrison, J. (1978), *Elements of Cartography*, 4th ed., New York: Wiley.

GEOELECTROMAGNETISM

L. J. LANZEROTTI, *AT&T Bell Laboratories, Murray Hill, New Jersey, U.S.A.*

R. A. LANGEL, *NASA Goddard Space Flight Center, Greenbelt, Maryland, U.S.A.*

A. D. CHAVE, *Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, U.S.A.**

Introduction 109
1. **Earth's Internal Magnetic Field** 109
1.1 Introduction 109
1.2 Gauss's Representation of the Field 110
1.3 Length Scales 110
1.4 Time Scales 111
1.5 Origin of the Main (Core) Field 112
1.6 Origin of the Crustal Field 113
2. **Earth's Variable Magnetic Field** 114
2.1 Introduction 114
2.2 Magnetosphere and Ionosphere 114
2.3 Variations in Magnetosphere–Ionosphere Current Systems .. 115
3. **Electric Currents at Earth's Surface** 116
4. **Electromagnetic Induction** ... 118
4.1 Introduction 118
4.2 Theoretical Considerations.... 119
4.3 Data Analysis 119
4.4 Conductivity Structures 120
4.5 Oceanic Studies 121
Glossary 121
Works Cited 122
Further Reading 123

INTRODUCTION

Four sources contribute to the electromagnetic field observed near Earth and at its surface: electric currents in the core, permanent magnetization in the crust or upper lithosphere, the dynamo action of moving, conducting seawater in Earth's main field, and "external" electric currents above Earth (in the ionosphere and magnetosphere) and their induced components inside Earth. The core, or main, field is by far the strongest of the sources. Because the core field must diffuse through Earth's mantle, this field yields information both on the region of its generation and on the electrical conductivity of the mantle. Studies of the external currents provide information on the vast plasma-filled space that surrounds the planet. Analyses of the induced electric current components yield further information on the conductivity structure of Earth's crust and mantle or on ocean currents.

1. EARTH'S INTERNAL MAGNETIC FIELD

1.1 Introduction

The science of geomagnetism began in 1600 with publication of the book *De Magnete* by a British physician, William Gilbert, in which he concluded that Earth itself is a great magnet. The fact that the geomagnetic field causes a compass to point in a particular direction was known to the early Chinese, an observation that illustrates that the magnetic field is a vector. If a magnetized needle is perfectly balanced on a pivot so that it is free to rotate about a horizontal axis, it will be

---

*This is WHOI Contribution No. 8216.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

nearly horizontal near Earth's equator, be nearly vertical near Earth's poles, and be inclined relative to the horizontal elsewhere, consistent with a dipolar form.

Except during magnetic storms (see Sec. 2) and in highly anomalous areas of the Earth's surface, about 99% of the magnetic field at the Earth's surface arises from electric currents in the conducting liquid core. About 90% of that field from the core takes the form of a magnetic dipole that can be modeled as lying at the Earth's center. In 1980 the model dipole was inclined at 11.2° to the axis of Earth's rotation in the direction of 289.3° E longitude. (A slightly more accurate dipole representation could be obtained by translating the inclined dipole from the center of the Earth in the general direction of the western Pacific. This is called an "eccentric dipole field," and in 1980 the dipole would have been displaced about 450 km from the Earth's center.) The position of the north pole of the model dipole (latitude 78.8° N, longitude 289.3° E) is called the geomagnetic north pole. The two locations in the northern and southern hemispheres where the field is exactly vertical are called the dip poles. The dip poles must be determined by *in situ* measurements, and in 1984 the north dip pole was measured at 77.0° N, 258.7° E. The dip poles and the dipole poles are not colocated, and each has a corresponding equator. Both sets of poles change position with time. There may also be local dip poles found near strongly magnetized mineral deposits in Earth's crust.

The temporal change of the field from the core is often referred to as a secular variation, though that term is most often used to refer only to the first time derivative of the change. Changes through geologic time are studied using measured magnetic properties of rocks, which also provide evidence that the field polarity occasionally reverses.

The main geomagnetic field (described by **B**, the magnetic induction or flux density, rather than **H**, the magnetic intensity) is described by Maxwell's equations. Within volumes where the current density is negligible **B** is curl-free and can be represented as the gradient of a scalar potential:

$$\mathbf{B}=-\nabla V, \tag{1}$$

$$\nabla^2 V=0. \tag{2}$$

1.2 Gauss's Representation of the Field

A solution to Eq. (2) for the geomagnetic field was first found by Gauss in the form of a spherical-harmonic expansion,

$$\begin{aligned} V=a \sum_{n=1}^{\infty} \sum_{m=0}^{n} \{ & (a/r)^n [g_n^m(t)\cos m\phi \\ & + h_n^m(t)\sin m\phi] \\ & + (r/a)^{n+1}[q_n^m(t)\cos m\phi \\ & + s_n^m(t)\sin m\phi]\} P_n^m(\cos\theta). \end{aligned} \tag{3}$$

Here a is the radius of the Earth, r is the radial distance from the center of the Earth, θ is the colatitude, ϕ is east longitude, $P_n^m(\theta,\phi)$ are the associated Legendre functions of degree n and order m, and g_n^m, h_n^m, q_n^m, and s_n^m are functions of time alone. P_n^m are Schmidt normalized so that

$$\int_0^\pi P_n^m P_n^{m'} \sin\theta \, d\theta = \delta_{nm'\delta_{mm'}} 4/(2n+1).$$

Terms in a/r describe fields originating within the Earth, and terms in r/a describe fields originating outside the Earth—i.e., in the ionosphere, magnetosphere, etc. If **B** is known everywhere on the spherical surface $r=a$ at the time t, then the Gauss coefficients, the g_n^m, h_n^m, q_n^m, and s_n^m, can be determined exactly.

1.3 Length Scales

At Earth's surface the strength of the dominant dipole field is about 6×10^4 nT at the poles and 3×10^4 nT at the equator. The remainder of the field has a complicated spatial pattern. A quantitative summary of the importance of the various horizontal wavelengths at Earth's surface can be obtained by considering separately the fields for various spherical harmonic degrees n. The strength of the "signal" in **B** at the wavelengths corresponding to a given degree n in Eq. (3) can be measured by averaging $|\mathbf{B}_n|^2$ over the Earth's surface:

$$\langle |\mathbf{B}_n|^2 \rangle = (n+1) \sum_{m=0}^{n} [(g_n^m)^2 + (h_n^m)^2]. \tag{4}$$

Figure 1, from Cain *et al.* (1989), shows the observed $\langle |\mathbf{B}_n|^2 \rangle$ as a function of n for low n.

The usual interpretation of Fig. 1 is that at the surface of Earth $\mathbf{B}_n$ comes principally from the core for $1 \leqslant n \leqslant 12$, principally from

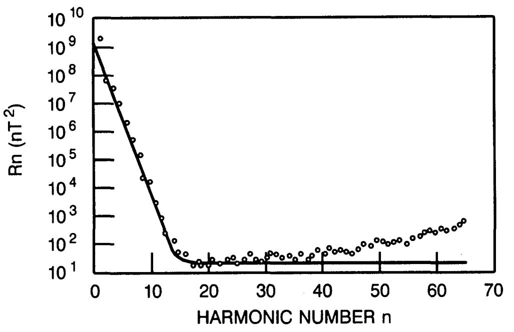

FIG. 1. Total mean square contributions to the vector field by all harmonics of degree n. Units are nT^2. Solid line is a fit to the data omitting a random noise component. Random noise is the cause of the rise in the data for $n > 40$. From Cain *et al.* (1989).

the crust (called "anomaly fields") for $n > 16$, and with both sources contributing for $13 \leqslant n \leqslant 15$. With existing satellite data, the crustal portion of the spectrum can be extended to an n of at least 50, which corresponds to a spatial wavelength of about 800 km. It is more convenient to use a characteristic size in kilometers for the anomaly fields, where the characteristic sizes are taken to be the distance between the half-amplitude contours of an anomaly. Aeromagnetic and shipborne magnetic surveys can typically map anomalies of characteristic size ranging from <1 to >100 km (see GEOPHYSICS).

1.4 Time Scales

Measurements of the remanent magnetization of ancient rocks suggest that the geomagnetic field strength has not changed by more than a factor of 3 to 4 over the last 3 billion years, except that it drops by a factor of 5 or more during geomagnetic reversals. Existing data indicate that the field may not become 0 during a reversal and that these reversals have occurred with increasing frequency during the past 70 million years, with a current average rate of about six per million years. The occurrence pattern of reversals appears to be random rather than periodic. For example, the typical occurrence period between the last few reversals has been of the order of 170 000 years, but the last complete reversal occurred over 700 000 years ago. During an individual reversal, the intensity of **B** may become anomalously low for about 10 000 years or less, and the reversal of its direction is typically completed in about 4000 years. Archeological evidence (magnetization of baked hearths and pots) indicates that the field intensity 2000 years ago at some sites was 50% higher than it is at present. In the last century the model dipole field intensity has been decreasing at a variable rate, averaging about 0.05% per year.

The secular variation $\partial_t \mathbf{B}$ is of the order of 60 nT/y. It varies greatly with position and with time. About 25% of the energy in $\partial_t \mathbf{B}$ at Earth's surface can be removed by subtracting from $\partial_t \mathbf{B}$ the field obtained by rotating the **B** pattern rigidly and steadily westward at about 0.2 deg/y. It is now believed that the secular variation is produced by fluid motions in the upper core of the Earth and that these motions may include a general westward rotation plus large spatially variable components. Figure 2 shows the variable rate of change of the main dipole term, g_1^0, in Eq. (3). Also shown in Fig. 2 is the change in length of day in milliseconds. There is a strong correlation between the two quantities with the length of day change leading the dipole change by about 13

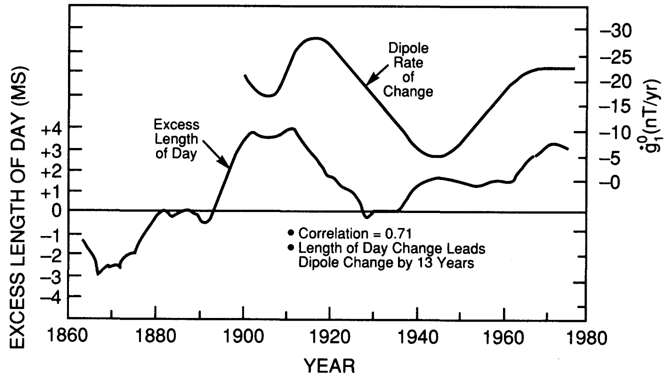

FIG. 2. Rate of change of axial dipole, $\dot{g}_1^0$, in nT/y and change of length of day in milliseconds from 1860 to 1980.

years. This correlation and lead–lag relationship is most likely related to electromagnetic coupling between the core and the mantle.

1.5 Origin of the Main (Core) Field

That portion of the magnetic field **B** at the Earth's surface that originates in the core is produced by electric currents driven by a self-sustaining dynamo process; fluid flowing across magnetic lines of force generates emf's that drive the electric currents that maintain **B**. The relationship between the fluid velocity and the magnetic field is expressed by the dynamo equation

$$\partial_t \mathbf{B} = (1/\sigma\mu_0)\nabla^2\mathbf{B} + \nabla\times(\mathbf{v}\times\mathbf{B}). \tag{5}$$

where σ is the conductivity, μ_0 the permeability, and **v** the fluid velocity. The two terms on the right-hand side of the equation describe diffusion and convection, respectively. When σ is very high, as in the liquid core, convection dominates diffusion. For most purposes the diffusion term can be ignored, and the magnetic field can be considered "frozen in" to the fluid; i.e., the field moves with the fluid. In a dynamo, energy from some source causes fluid flow, **v**, in some initial magnetic field such that, through the processes embodied in Eq. (5), the magnetic field grows or is sustained. **B** can be conveniently decomposed into toroidal ($\mathbf{B}_T$) and poloidal ($\mathbf{B}_P$) components, where $\mathbf{B}_T$ is any field of the form $\mathbf{B}_T = -\mathbf{r}\times\nabla t$ and $\mathbf{B}_P$ is any field of the form $-\nabla\times(\mathbf{r}\times\nabla p)$, with t and p any scalar functions of position. Toroidal fields have only an azimuthal component. Poloidal fields have components in meridional planes; e.g., dipole fields. Following Parker (1979), the fundamental physical processes that are responsible for generating toroidal fields from poloidal fields and vice versa are assumed to be differential rotation of the outer-core fluid and the cyclonic motion of convective upwelling in the outer core.

The angular velocity of the core is presumed to decrease with distance from the spin axis, as evidenced, for example, by the geomagnetic westward drift. This nonuniform rotation causes a shear in the lines of force of any poloidal (e.g., dipole) magnetic field, stretching the field in the azimuthal direction. Thus an azimuthal, or toroidal, field will be generated from a poloidal field. One suggestion is that this process reaches its limit when the Lorentz force balances the Coriolis forces of the convective motions, in which case the toroidal field could easily exceed 10^{-3} T. If the Coriolis forces are balanced by the fluid pressure, then the toroidal field could be much smaller. The dipole field today has a strength of only about 10^{-5} T at the core–mantle boundary. Because toroidal fields require meridional electrical currents, these magnetic fields cannot extend outside the core into the nonconducting mantle and are therefore unobservable at the surface of Earth. What may be the largest portion of the magnetic field of our planet is therefore undetectable to us.

Similarly, cyclonic motion of localized convective upwelling provides a mechanism for generating poloidal fields from toroidal fields. In this case, as a local packet of fluid upwells to form a convection cell, it carries with it the lines of force of the azimuthal field, forming a bulge in the field. Since the Earth is spinning, such fluid upwelling is cyclonic. This means that the bulge rotates as it rises. The net result is a small patch of field with a longitudinal component as well as an azimuthal component. The averaging together of many such cells is thought to be the principal means of generating toroidal fields from poloidal fields.

The core is not perfectly conducting and such a process, left to itself, would damp out in a few tens of thousands of years. An energy source is required to drive the convection of the core fluid. Some have suggested that radiogenic sources in the core could furnish the necessary energy, but this is disputed by others. The inner core (see EARTH, INTERIOR STRUCTURE OF THE) is formed by freezing of part or all of the liquid in the outer core, a process that could also provide the necessary heat source. The most popular current notion is that, at the inner-core boundary, the fluid of the outer core separates into a heavy fraction, which freezes to form the solid inner core, and a lighter fraction, which remains in the outer core. This lighter fraction is buoyant and drives the outer-core convection. In this scheme, the ultimate source of energy is the gravitational energy of the Earth, hence the term "gravitationally powered dynamo" for this generation mechanism.

1.6 Origin of the Crustal Field

As it forms, a rock generally acquires a magnetization along the direction of the ambient field. The strength and permanence of the magnetization depend upon the mineralogy of the rock and upon the strength of the ambient field. Examination of this magnetization in rock samples in the laboratory provides information on the geomagnetic field at the time the rock was formed. Ferrimagnetic materials lose their magnetization when heated above a critical temperature, called the Curie temperature, typically about 600 °C for magnetite. Some ferrimagnetic rocks when cooled to below their Curie points acquire a relatively strong and stable field. This is the dominant process (thermal remanant magnetization—TRM) that gives the lineated magnetic anomalies found on the ocean floor. These lineated magnetization reversals resulted as the ocean floor cooled at and near midocean spreading centers, and thus are records of the process of sea-floor spreading. Rocks can also acquire a viscous magnetization (VM) in which the acquired induced magnetization intensity is proportional to the length of time a rock spends in the ambient field. VM has been thought to be of particular importance in the lower crust of the Earth.

Spatial variations in the anomaly field reflect differences in near-surface magnetization of the rocks. Magnetization variations occur not only from variations in the amount of magnetic minerals but also from variations in topography and in the depth to the Curie isotherm. Variations in magnetization content reflect the past history of the rock: the chemistry of the rock source and the processes by which it was formed, the thermal and mechanical history of the rock formation, and alteration due to diagenesis and metamorphism. Figure 3 shows a relative magnetization map of the United States derived from satellite magnetic anomaly data, with some of the geologic and tectonic features indicated for comparison. (The units in Fig. 3 have been

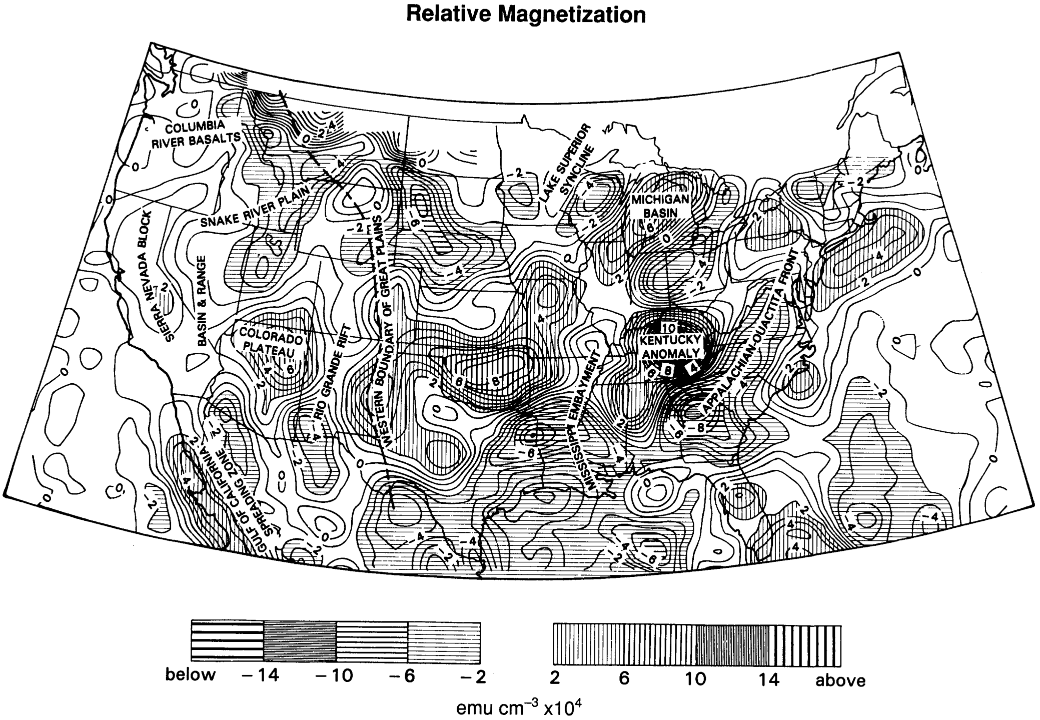
Relative Magnetization
COLUMBIA RIVER BASALTS
SNAKE RIVER PLAIN
SIERRA NEVADA BLOCK
BASIN & RANGE
COLORADO PLATEAU
RIO GRANDE RIFT
WESTERN BOUNDARY OF GREAT PLAINS
GULF OF CALIFORNIA SPREADING ZONE
LAKE SUPERIOR SYNCLINE
MICHIGAN BASIN
KENTUCKY ANOMALY
MISSISSIPPI EMBAYMENT
APPALACHIAN-OUACHITA FRONT
below
−14
−10
−6
−2
2
6
10
14
above
emu cm^{-3} x10^4

FIG. 3. Relative magnetization anomaly map of the United States derived from POGO satellite data. Units are 10^4 emu/cm^{-3}. Some geologic features are indicated. (Adapted from Mayhew, 1982).

scaled upward by 10^4; i.e., the contour level 2 means 2×10^{-4} emu/cm^3.)

2. EARTH'S VARIABLE MAGNETIC FIELD

2.1 Introduction

The magnetic field of the Earth, if measured with sensitive instruments on the surface of or in the space around the planet, is found to vary on time scales from months to days to fractions of a second. The amplitudes of these variations can range typically from a percent or so to as little as one part in a million or less of the background field magnitude. These changes arise from electrical currents flowing in the space environment around the Earth, particularly in the ionosphere (*q.v.*) (altitudes of around 100 to 150 km) and in the magnetosphere (*q.v.*).

2.2 Magnetosphere and Ionosphere

In a steady-state perspective, the dominant current systems in the overall magnetosphere can be illustrated as shown in Fig. 4. Currents with densities ~25–150 mA/m^2 flow along the Earth's day-side magnetopause (typically located at ~$10R_E$ from Earth's center) and provide for the separation of the Earth magnetic-field–dominated regime from that of the interplanetary medium (consisting of the solar wind and interplanetary magnetic field). On the night side, in the plasma sheet of Earth's magnetotail (which begins at distances of the order of ~$10R_E$ behind the Earth), currents of density typically ~10 to 80 nA/m^2 (total magnetotail current ~10^8 A) flow dawn to dusk through the center of the magnetotail and close around its boundary. During geomagnetic disturbances, a ring current of order 10^6–10^7 A (~1–10 nA/m^2) grows from near the Earthward edge of the magnetotail current sheet, approaches closer to Earth, and, in large enough geomagnetic storms, completely encircles the planet. This ring

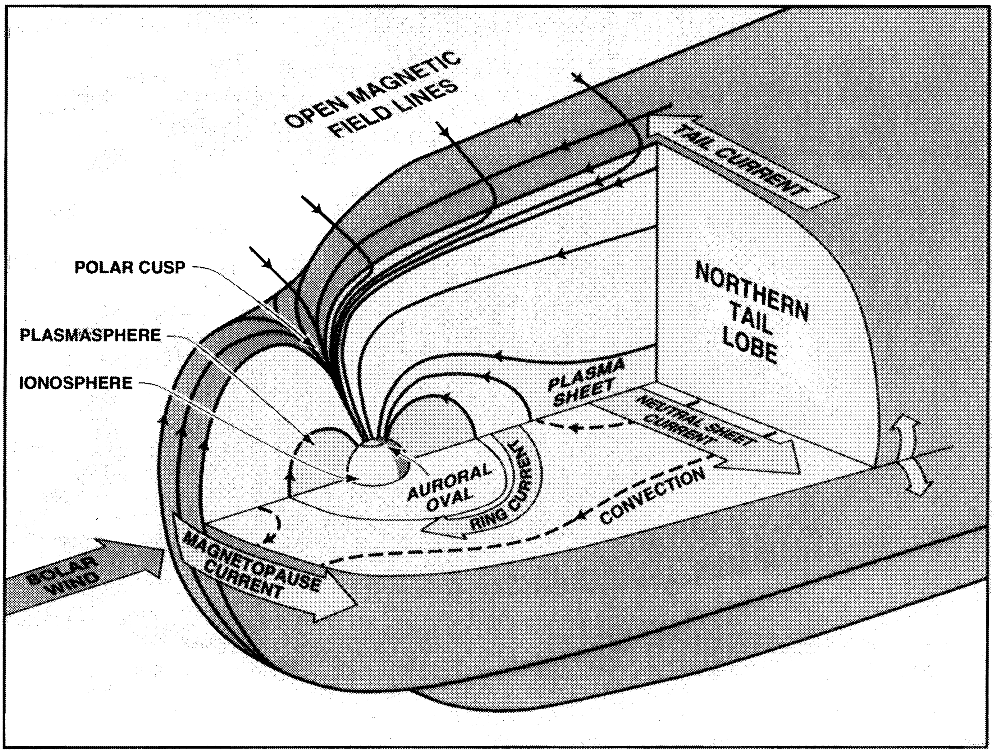
OPEN MAGNETIC FIELD LINES
TAIL CURRENT
POLAR CUSP
PLASMASPHERE
IONOSPHERE
NORTHERN TAIL LOBE
PLASMA SHEET
NEUTRAL SHEET CURRENT
AURORAL OVAL
RING CURRENT
CONVECTION
SOLAR WIND
MAGNETOPAUSE CURRENT

FIG. 4. Dominant space plasma current systems shown in a "cutaway" view of the Earth's magnetosphere.

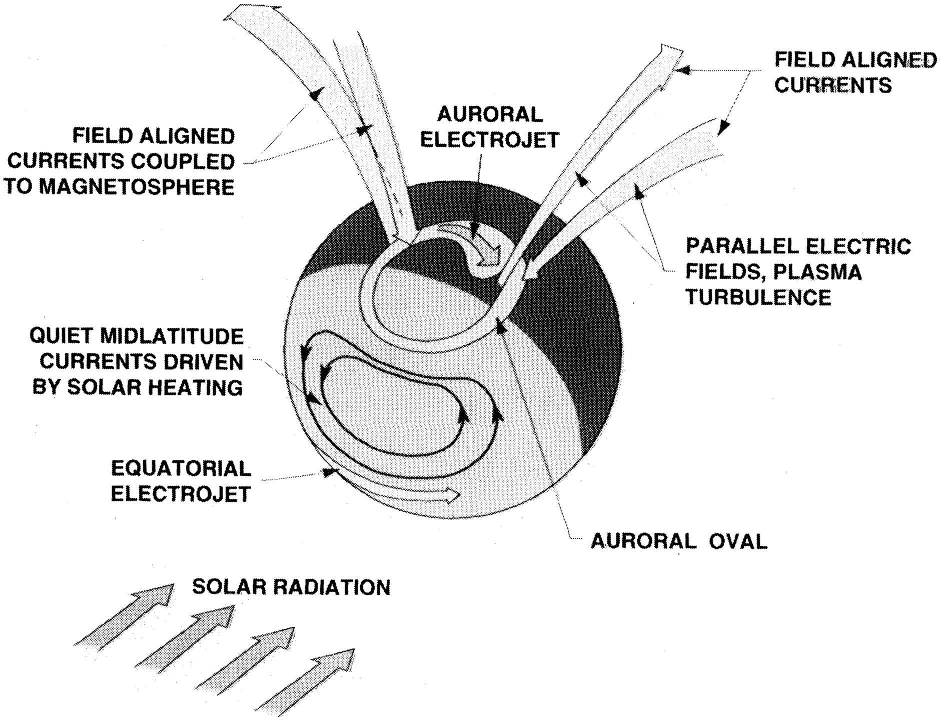
FIELD ALIGNED CURRENTS COUPLED TO MAGNETOSPHERE
AURORAL ELECTROJET
FIELD ALIGNED CURRENTS
PARALLEL ELECTRIC FIELDS, PLASMA TURBULENCE
QUIET MIDLATITUDE CURRENTS DRIVEN BY SOLAR HEATING
EQUATORIAL ELECTROJET
AURORAL OVAL
SOLAR RADIATION

FIG. 5. Dominant current systems associated with the Earth's ionosphere.

current can be connected to the ionosphere, particularly the auroral current system, through magnetic-field–aligned currents (a feature that indicates that geomagnetic field lines are not always equipotentials, as might be thought from simple considerations).

The dominant current systems in Earth's ionosphere are illustrated in Fig. 5. Day-side ionosphere currents are produced primarily by enhanced ionization caused by the fluxes of ultraviolet radiation from the sun. These currents, as well as those on the day-side magnetopause, produce variations in day-side magnetic fields. These day-side ionospheric currents have seasonal dependences (because of the tilt of the Earth's rotation axis) and a dependence on the phase of the approximately 11-year solar sunspot cycle. The most intense of these daytime currents—the equatorial electrojets—flow within a few degrees of the magnetic equator. Night-side auroral currents, with intensities of the order of 10^6 A or more and latitudinal extents of 100 km or more, flow in the ionosphere at latitudes near 65°. These ionospheric auroral currents are connected, through magnetic-field–aligned currents, to the Earthward edge of the magnetotail current sheet. The auroral current system appears as an essentially continuous oval band of illumination in the high upper atmosphere—the ionosphere—encircling each geomagnetic pole. The day-side portions of these auroral currents are connected, again via field-aligned currents, to the currents flowing on the day-side magnetopause. Occasionally single or multiple, approximately sun-aligned, auroral arcs stretch across the polar caps. The relationships of such arcs to magnetosphere current systems remain somewhat uncertain.

2.3 Variations in Magnetosphere–Ionosphere Current Systems

The solar wind, the continued outflow into the interplanetary medium of the 10^6-K solar corona, provides the energy for the complex plasma environment surrounding the Earth. Variations in the solar-wind flow speed, plasma number density, and interplanetary magnetic field strength and direction all can perturb, often severely, the geospace environment. These occurrences are called geomagnetic storms. For example, a large enough increase

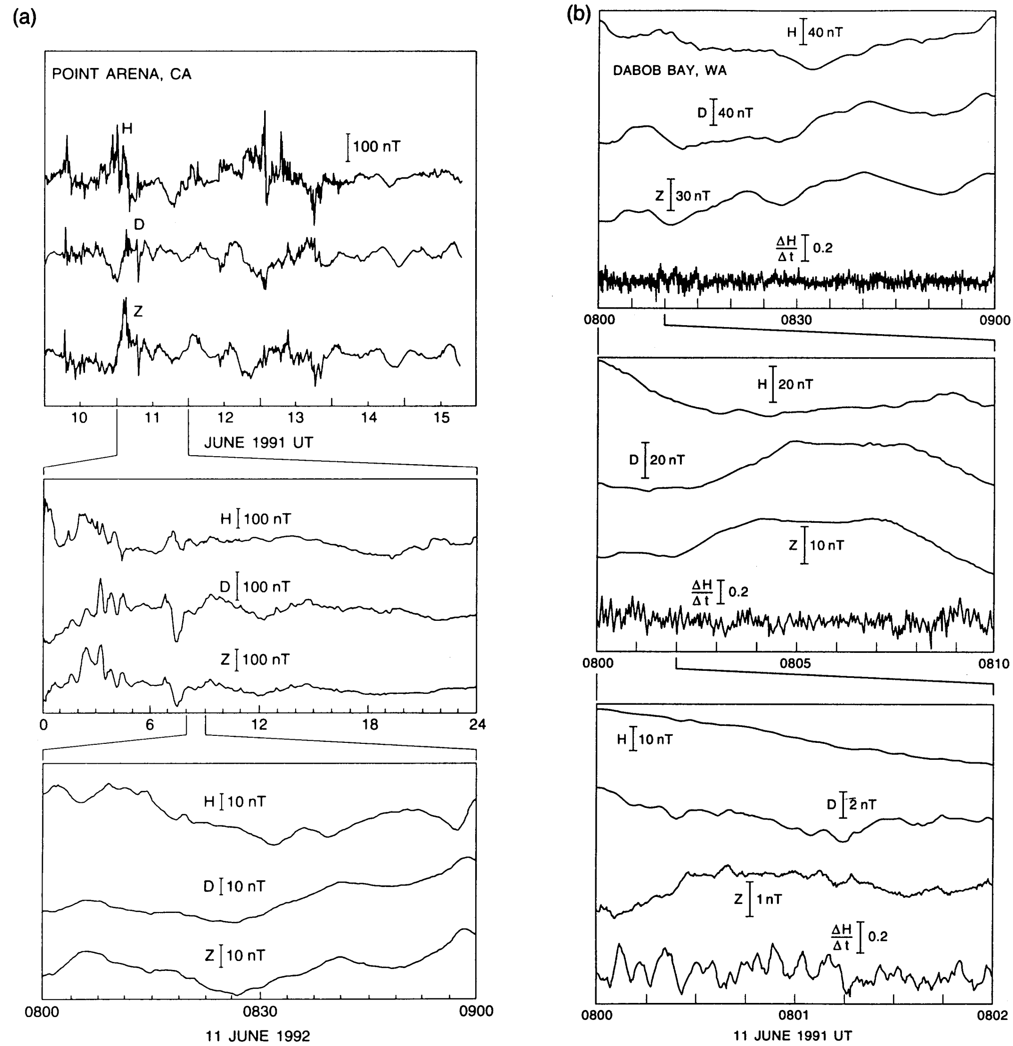

FIG. 6. (a) Time–intensity traces of geomagnetic fluctuations measured at increasing sensitivity and time resolution, days to hours, at Point Arena, CA, a mid-geomagnetic-latitude location on the west coast of the United States. The magnetic field is measured in three orthogonal geomagnetic components: *H* (south–north), *D* (west–east), *Z* (vertical). (b) Time-intensity traces of geomagnetic fluctuations measured at increasing sensitivity and time resolution, hours to seconds, at Dabob Bay, WA, a mid-geomagnetic-latitude location on the west coast of the United States.

in the pressure of the solar-wind flow can push the magnetopause substantially closer to Earth than normal, and not infrequently within the orbit of a geosynchronous satellite ($\sim 6.6R_E$). When this happens, the effective field at this location will have a direction and magnitude approximately opposite to that of the geomagnetic dipole. At the time of such solar-wind disturbances, the ionospheric auroral currents are considerably enhanced and shift equatorward from their usual locations; in some extreme cases, aurorae have been reported as far south as Hawaii. During such very large geomagnetic storms, the ring current becomes greatly enhanced and produces a decrease of as much as a fraction of a percent in the intensity of the main magnetic field as measured at a near-equatorial location on the Earth's surface.

In more usual circumstances, less intense variations in the solar-wind pressure and changes in the interplanetary field direction can produce geomagnetic disturbances (substorms) that last on the order of 1 to 3 h. One

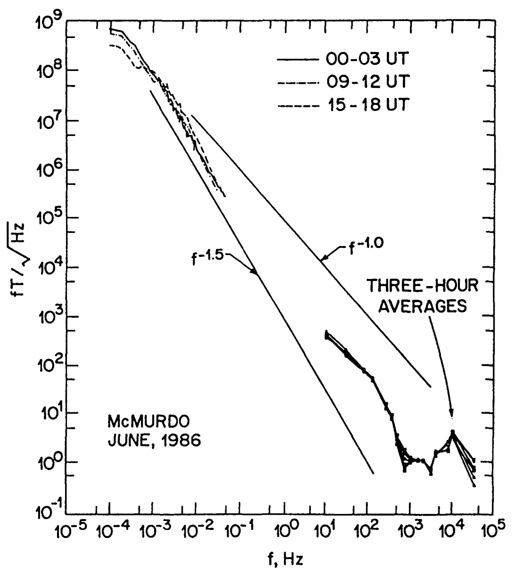

FIG. 7. Daily average amplitude spectra for magnetic field fluctuations for three geomagnetically quiet and three geomagnetically disturbed days in June 1986. From Lanzerotti *et al.* (1990).

or several such events can occur during a 24-h interval depending upon the state of the solar wind. In particular, if the interplanetary magnetic field has a vector component opposite in direction to that of the main dipole field at the magnetopause, the interplanetary and magnetosphere fields can become connected. Such a "reconnection" will drive field-aligned currents into and out of the polar-region ionospheres. These geomagnetic substorms are principally manifest in the ionosphere on the night side of the Earth at latitudes near that of the nominal auroral zone, although some effects can occasionally be seen more globally. For example, at the onset of a geomagnetic substorm, a characteristic magnetic perturbation—a damped wavelike variation with a period of the order of 60 s (termed a Pi2 pulsation)—is observed both on the Earth's surface over a wide range of geomagnetic latitudes and longitudes and in the magnetosphere itself.

Externally produced fluctuations in the geomagnetic field can be investigated and characterized in several ways. These include making time plots of the intensity of the field fluctuations and computing the power spectra (or amplitude spectra) of the fluctuations for a given measurement interval. Plotted in Figs. 6(a) and 6(b) are time–intensity plots of geomagnetic data acquired at ground-based measuring locations in California and in the Pacific Northwest. The data traces show that as the time resolution is increased, higher frequency variations with decreased amplitudes become evident. Further, a comparison of the 1-h traces of the individual components between the two locations on the geomagnetically disturbed day of 11 June shows that the more northern location (Dabob Bay) sees the larger range of variations.

The decreasing amplitudes with increasing frequency of the fluctuations means that the spectra of the geomagnetic variations can be characterized as red. Indeed, power spectra computed for fluctuation periods ranging from as long as a day (24 h) to as short as a fraction of a second have overall frequency dependences that vary as f^{-n}, where n has values in the range ~1 to ~3 to 5, depending upon the geomagnetic location and the level of geomagnetic activity. Superimposed on such mean spectra are often peaks that correspond to hydromagnetic wave resonances in the magnetosphere (typically in the period band ~1 to 600 s). The daily variation of the geomagnetic field due to the 24-h solar illumination cycle, and its harmonics at 12, 8, and 6 h, are also found. Daily average amplitude spectra for geomagnetic field fluctuations measured at McMurdo Station, Antarctica, for quiet and disturbed magnetic conditions in the austral winter are shown in Fig. 7. The daily variation and its harmonics are greatly reduced in magnitude at such a high-latitude station during the winter months.

3. ELECTRIC CURRENTS AT EARTH'S SURFACE

The principal sources of electric currents at the surface of the Earth from sources other than those described in Sec. 2 arise from lightning in Earth's atmosphere and from the atmospheric clear-weather electric field. Some 2000 thunderstorms are, on average, in progress around the world at any given time. These storms produce on the order of 100 cloud-to-ground discharges per second. The lightning occurrence regions are heavily concentrated in the equatorial regions of South America, Africa, and southeast Asia (see ATMOSPHERIC STRUCTURE). Cloud-to-ground lightning discharges typically transfer some 25 C of negative charge from the cloud to Earth's surface. Locally at Earth's surface, the peak current in a lightning discharge flowing upward can be a few hundred amperes for a few tenths of a second. These currents can be deadly to people struck directly or who might find themselves in close proximity to a lightning discharge channel. A hundred or more deaths in the United States can be attributed directly to lightning each year. Lightning also damages or destroys trees and manmade structures.

There is a vertical potential drop at the surface of the Earth of the order of ~200 V/m under clear-weather conditions. The typical clear-weather current density resulting from this is of the order of 3×10^{-12} A/m^2. Essentially, the Earth and the ionosphere can be characterized as forming the two plates of a giant leaky capacitor, with a total voltage drop of perhaps 300 kV. The clear-weather potential has a universal time (UT) variation in amplitude of about 20%, with a broad maximum in the 12–20 UT interval corresponding to the increased daytime thunderstorm activity in South America and Africa. The recharging of the Earth to maintain a net negative charge of ~10^6 C occurs through the transfer of negative charge to the surface by the thunderstorm activity.

4. ELECTROMAGNETIC INDUCTION

4.1 Introduction

Electromagnetic fields are induced in the conducting Earth by the external, ionospheric and magnetospheric, electric current systems. An additional internal source due to dynamo interaction of water currents with Earth's steady magnetic field is important in the oceans and near coastlines, generally at periods of a day or longer. As discussed in Sec. 2, the spectrum of field fluctuations produced by external sources covers a period range from a fraction of a second to more than a year. These fluctuations can interact with the conductive rocks in the Earth to depths of hundred of kilometers through the process of electromagnetic induction. Using measurements of the resulting electric and magnetic fields collected at Earth's surface, it is possible to infer the subsurface electrical conductivity.

The most widely used such electromagnetic method is magnetotellurics (MT) (see GEOPHYSICS), which is based on the principle that the electromagnetic fields induced by a fluctuating external magnetic field will attenuate with depth in the earth through the well known skin effect. From Ampère's law, the time-variable surface magnetic field is a measure of the total electric current flowing inside the Earth. If the surface electric field is also known, then the distribution of current, and hence the subsurface conductivity, can be determined. Since longer-period disturbances penetrate a conductor to greater depths, surface electric and magnetic field data covering a range of frequencies can provide information on the vertical profile of conductivity. Making such measurements with a line or a two-dimensional grid of instruments can in principle give a two- or three-dimensional picture of the subsurface electrical structure. Magnetotelluric investigations are usually applied to specific geophysical targets on a local to regional scale covering a few to several hundred kilometers.

A comparison method is geomagnetic depth sounding (GDS), which is based solely on the analysis of fluctuating magnetic field data collected at Earth's surface. Since GDS requires either measurements of, or assumptions about, the spatial scales of the inducing fields, it is generally applied on a global basis at long periods (hours to days) where a low-order spherical harmonic source is likely to be applicable. The analysis gives valuable information about the deep electrical structure of the Earth, down to as much as 1000 km depth.

A closely related topic occurs in applied or exploration geophysics where the objective is to obtain information about shallow subsurface structures, often for some economic purpose. Electromagnetic exploration equipment typically consists of an artificial source (e.g., electric or magnetic dipole) at the Earth's surface and one or more electric or magnetic receivers in either a fixed or airborne configuration. While the physical principles are identical to those for electromagnetic induction studies using the natural source, exploration practice and interpretation are often distinct, being driven by specific needs.

4.2 Theoretical Considerations

Electromagnetic induction is governed by the pre-Maxwell equations in which the magnetic effects of displacement currents are neglected. In the absence of an explicit source current density, these are

$$\nabla\cdot\mathbf{B}=0, \tag{6}$$

$$\nabla\times\mathbf{E}+\partial_t\mathbf{B}=\mathbf{0}, \tag{7}$$

$$\nabla\times\mathbf{B}-\mu_0\mathbf{J}=\mathbf{0}, \tag{8}$$

$$\mathbf{J}=\sigma\mathbf{E}, \tag{9}$$

where the symbols have their usual meaning and the magnetic permeability is assumed to have its free-space value everywhere (an excellent approximation for most Earth materials). As a consequence of the pre-Maxwell approximation, the electric current density is divergence-free.

Equations (7)–(9) can be combined to yield separate differential equations for the electric field and magnetic induction vectors:

$$\nabla\times\nabla\times\mathbf{E}+\mu_0\sigma\partial_t\mathbf{E}=\mathbf{0}, \tag{10}$$

$$\nabla\times\nabla\times\mathbf{B}-\frac{\nabla\sigma}{\sigma}\times\nabla\times\mathbf{B}+\mu_0\sigma\partial_t\mathbf{B}=\mathbf{0}. \tag{11}$$

These are the equations of electromagnetic induction and are diffusion equations analogous to those for the flow of heat.

In general, analytic solutions of Eqs. (10) and (11) are available only for one-dimensional (1D) conductivity structures, most often by considering a stack of horizontal layers and solving recursively for the fields in each in terms of exponential basis functions. For the simplest instance where the horizontal dimensions of the external, driving magnetic field are infinite (plane-wave approximation), as appears to be reasonable for magnetotelluric sounding at periods shorter than about 10^4 s or so, and when the structure is 1D, induced electric currents will only flow in the horizontal plane. This means that the vertical electric and magnetic fields must be zero everywhere.

Numerical solutions for two-dimensional (2D) structures are widely available using either finite-element, finite-difference, or integral-equation approaches. In this instance, the Cartesian form of Eqs. (6)–(9) may be separated into two independent modes that are described by distinct ordinary differential equations. The transverse electric (TE) mode is described by an electric field component in the invariant conductivity direction and the two orthogonal magnetic induction terms, so that electric currents circulate in horizontal planes. The transverse magnetic (TM) mode is given by the magnetic field component in the invariant conductivity direction and the two orthogonal electric field terms, so that electric currents may flow across the conductivity gradient and hence involve electric charges. Current magnetotelluric practice is centered on studies of 2D structures using the TM mode, which is relatively insensitive to weak three-dimensional (3D) effects. Numerical solutions for fully 3D structures are still in the experimental stage, being limited largely by computing technology (Hohmann, 1987).

Derivation of the electrical conductivity structure given a set of response functions and their associated errors is performed using either trial-and-error forward modeling techniques or a more formal inversion scheme. The latter yields both a model that fits best in some well-defined sense and, in many instances, an estimate of the resolution of the model features. Inversion is now routine for 1D data and is rapidly becoming standard for 2D situations as well.

4.3 Data Analysis

Typical magnetotelluric data consist of time series of the two horizontal electric field and the three magnetic field components collected at a number of sites. The electric field variations are usually measured with a sensitive voltmeter placed across long (~100 m) wires coupled to the Earth with a set of electrodes. The magnetic field variations may be sensed with either induction coils or fluxgate magnetometers.

Magnetotelluric data analysis is accomplished by estimating the elements of the response tensor relating the horizontal electric and magnetic field components in the frequency domain,

$$\begin{bmatrix}E_x\\E_y\end{bmatrix}=\begin{bmatrix}Z_{xx} & Z_{xy}\\Z_{yx} & Z_{yy}\end{bmatrix}\begin{bmatrix}B_x\\B_y\end{bmatrix}. \tag{12}$$

A similar relation between the vertical and two horizontal magnetic field components is also of interest. The response tensor is typically computed using time-series analysis methods covered in standard texts. However, considerable recent progress has been made with robust processing algorithms (Chave and Thomson, 1989) that are relatively insensitive to the dramatically non-Gaussian form of natural electromagnetic variations, yielding much more reliable results than more conventional approaches. In addition, uncorrelated noise on the magnetometer channels in Eq. (12) will inevitably bias the response functions downwards, and can be dealt with using a remote reference method with a second, auxiliary magnetometer.

The response tensor in Eq. (12) possesses symmetry properties that depend on the dimensionality of the underlying conductivity structure. For a 1D earth, the diagonal elements will be zero to within the experimental errors, while the off-diagonal elements will be identical. For a 2D earth, the tensor can be rotated so that the diagonal elements vanish and the off-diagonal terms are unequal; the rotation angle gives the strike direction (the direction of the projection onto Earth's surface) of any subsurface anomaly, with a 90° ambiguity that can be removed using the vertical field measurement. For a 3D earth, the tensor has no special symmetries. In many instances, a good model for the Earth may consist of a regional 1D or 2D structure distorted by near-surface 3D galvanic effects, and special tensor decompositions exist to remove the surface distortion (Groom and Bailey, 1989).

4.4 Conductivity Structures

At low temperatures, the electrical conductivity of rocks is usually controlled by the mobility of ions in pore fluids and hence depends on the amount and connectedness of interstitial fluids. At higher temperatures (> 600 °C), electronic conduction mechanisms become important, and at temperatures above the solidus, partial melting may occur, resulting in a dramatic rise in conductivity. In general, electrical conductivity mechanisms are exponentially thermally activated, so that even if a number of mechanisms are possible, one will typically be dominant for a given set of physicochemical conditions. The range of electrical conductivity in Earth materials covers at least eight decades, more than any other

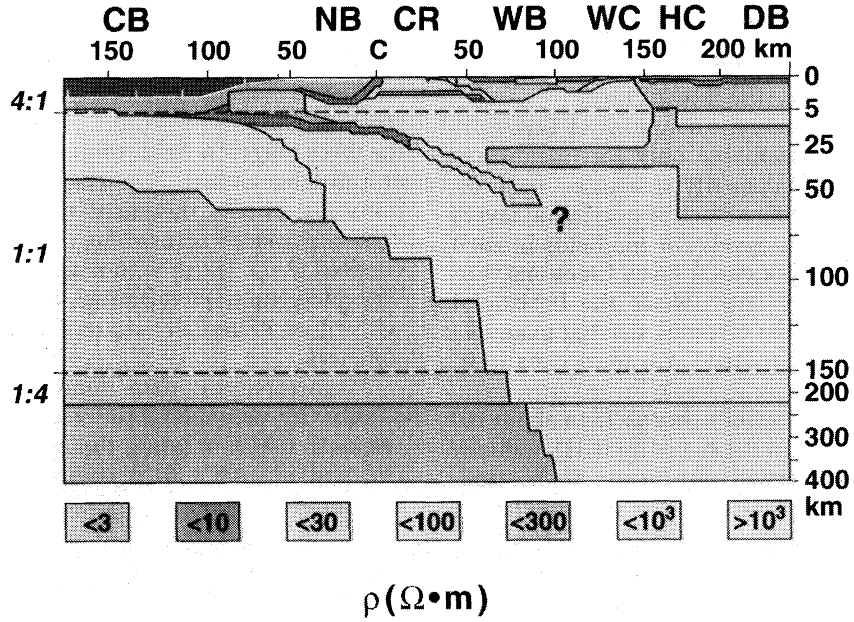

FIG. 8. East–west electrical resistivity cross section derived by forward modeling of magnetotelluric data taken from the Pacific Ocean to east of the Cascade Mountains in central Oregon during the EMSLAB Project. The seafloor portion of the model is shown only to 175 km offshore, beyond which changes are minor (note vertical exaggerations at 5 and 150 km depth). The geologic regions that the transect crossed include the Cascade Basin (CB), Newport Basin (NB), Coast Range (CR), Willamette Basin (WB), Western Cascades (WC), High Cascades (HC), and Deschutes Basin (DB). From Wannamaker *et al.* (1989).

physical parameter used to define such natural materials. Typical values range from 3 S/m, typical of seawater or moderately molten rock, down to 10^{-6} S/m at the top of crystalline pre-Cambrian shields.

The electrical conductivity structure of the Earth is quite heterogeneous in the uppermost 100 km or so on land, reflecting its complex geologic history. The largest coordinated magnetotelluric experiment ever performed is the EMSLAB effort in the Pacific Northwest, where a transect starting 400 km offshore at the Juan de Fuca Ridge was continued across the shoreline and to the eastern side of the Cascade Mountains. A 2D conductivity structure down to 200 km depth constructed from the TM mode data from this investigation is shown in Fig. 8. A key feature is a moderately conductive layer dipping at $\approx 20°$ under the Coast Range, which can be traced to about 100 km inland. This is due to subduction of the Juan de Fuca plate at the coast, and the enhanced conductivity is caused by dewatering of seafloor sediments and crust at the detachment fault. Another high-conductivity zone occurs under the Cascade Mountains, presumably due to enhanced temperature and partial melting associated with orogeny.

Global geomagnetic studies have yielded information about the electrical conductivity down to more than 1000 km depth. While details are elusive, there are suggestions for abrupt increases in conductivity associated with phase-transition–induced seismic velocity changes that have been detected at 350 and 650 km. Recent emphasis has been on detecting lateral heterogeneity of the deep electrical structure of the Earth associated with mantle dynamics.

4.5 Oceanic Studies

Oceanic electromagnetic induction differs from its continental counterpart in two important ways. First, the overlying, conductive seawater layer markedly attenuates the incident ionosphere-produced magnetic field changes at periods shorter than a fraction of an hour; thus the high-frequency magnetotelluric fields that are usable on land cannot be employed for studies at sea. Second, electric fields are induced by the motion of the conductive seawater through Earth's magnetic field at periods longer than a day. While the former means that stand-alone oceanic magnetotelluric studies will have limited resolution, the inclusion of offshore data is essential for land studies near coastlines. The ocean serves as an efficient collector of electromagnetic fields, and the extent to which electric currents in the ocean leak into the continent and down to the conductive deep mantle has a dominant effect on magnetotelluric data to distances of hundreds of kilometers inland.

There is increasing interest in physical oceanographic applications of electromagnetic methods since the electric field at periods longer than a day yields a spatially averaged measure of the water velocity. It can be shown that the horizontal electric field is proportional to the seawater-conductivity–weighted velocity integrated from seafloor to sea surface and averaged over a radius of a few times the water depth. This is effectively the depth-averaged (barotropic) velocity in most of the oceans, a quantity that is difficult to measure with conventional instrumentation. A recent study of the oceanic response to wind forcing (which is barotropic at periods shorter than ~100 days) based on electric field measurements is described in Chave *et al.* (1992). When a seafloor cable is employed, the velocity field is averaged horizontally as well, effectively giving the transport through an ocean section (Larsen, 1992).

GLOSSARY

Aurora: Lights in the upper atmosphere (from about 90 to 300 km altitude) produced by the excitation of atmospheric gases by energetic particles; the localized areas in the two polar regions where aurorae are typically observed are determined by the detailed topology of the magnetosphere.

Core: The center portion of the Earth. Probably composed of an iron–nickel alloy. Divided into two parts: the inner core is a solid (with mushy surface) with radius approximately 1220 km; the outer core is liquid with radius of approximately 3480 km.

Crust: The outermost, solid, layer of the Earth. Its thickness is about 5 km in oceanic areas and between 15 and 70 km in continental areas. The lower boundary of the crust is the Mohorovicic seismic discontinuity (or Moho).

Ferrimagnetic: A substance, generally an iron oxide, that has a strong magnetization because the strength of its antiparallel magnetic lattices is stronger in one direction than in others. The magnetization may be either permanent (remanent) or induced or both.

Ionosphere: Ionized region of the upper atmosphere, generally from about 90 to about 1000 km in altitude; the area beyond this is defined as the magnetosphere.

Magnetic Induction: The diffusionlike process by which electromagnetic fields are produced by a time-varying external magnetic field.

Magnetic Storm: A worldwide variation in the geomagnetic field. Near the Earth a storm typically has the following character in the horizontal field component: (1) sudden commencement or sudden sharp increase in magnitude (not all storms have this), (2) main phase or decrease in magnitude over a period of a few hours, (3) recovery phase or increase of the field to its prestorm or quiet level, typically taking one to four days.

Magnetization: Magnetic moment per unit volume. Rock magnetization may be either permanent (remanent; that is, "frozen" into the rock as it last cooled) or induced (a response of the rock to its ambient magnetic field).

Magnetopause: Boundary between the flowing solar-wind plasma and the magnetosphere.

Magnetosphere: Region of space around the Earth in which the terrestrial magnetic and electric fields usually dominate the transport and motions of charged particles.

Magnetotail: Region of the magnetosphere in the antisunward direction: if visible, the magnetotail would have characteristics like the tail of a comet.

Magnetotellurics: A method for inferring the electrical structure of the Earth using measurements of the time variations of electric and magnetic fields at Earth's surface.

Mantle: The region of the Earth between the core and crust; it has a thickness of about 2300 km.

Motional Induction: The diffusionlike process by which electromagnetic fields are induced by the motion of moving, conducting seawater through Earth's magnetic field. This differs from the core dynamo process in that the water motions are not affected by the induced fields (i.e., the Lorentz force is negligible).

Plasma: Gas (atomic and/or molecular species) that is ionized. Naturally occurring plasmas usually contain magnetic fields.

Solar Wind: Expansion of the solar corona, primarily hydrogen ions, into the interplanetary medium.

Substorm: Interval of one to three hours of auroral, geomagnetic, and magnetospheric activity, usually followed by a several-hour interval of relative quiescence.

Sunspot Cycles: Variation with time of the appearance of the number of sunspot groups on the visible solar photosphere.

Lists of Works Cited

Cain, J. C., Wang, Z., Schmitz, D. R., Meyer, J. (1989), "The geomagnetic spectrum for 1980 and core-crustal separation," *Geophys. J.* **97**, 443–447.

Chave, A. D., Thomson, D. J. (1989), "Some comments on magnetotelluric response function estimation," *J. Geophys. Res.* **94**, 14 225–14 226.

Chave, A. D., Luther, D. S., Filloux, J. H (1992), "The barotropic electromagnetic and pressure experiment, 1. Barotropic current response to atmospheric forcing," *J. Geophys. Res.* **97**, 9565–9593.

Groom, R. W., Bailey, R. C. (1989), "Decomposition of magnetotelluric impedance tensors in the presence of local three-dimensional galvanic distortion," *J. Geophys. Res.* **94**, 1913–1925.

Hohmann, G. W. (1987), "Numerical modeling for electromagnetic methods of geophysics," in: M. N. Nabighian (Ed.), *Electromagnetic Methods in Applied Geophysics*, Tulsa, OK: Society of Exploration Geophysicists, Vol. 1.

Lanzerotti, L. J., Maclennan, C. G., Fraser-Smith, A. C. (1990), "Background magnetic spectra: $\sim 10^{-5}$ to $\sim 10^{5}$ Hz," *Geophys. Res. Lett.* **17**, 1593–1596.

Larsen, J. C. (1992), "Transport and heat flux of the Florida Current at 27 °N derived from cross-section voltages and profiling data: theory and observations," *Philos. Trans. R. Soc. London A* **338**, 169–236.

Mayhew, M. A. (1982), "An equivalent layer magnetization model for the United States derived from satellite altitude magnetic anomalies," *J. Geophys. Res.* **87**, 4487–4845.

Parker, E. N. (1979), *Cosmical Magnetic Fields*, Oxford: Clarendon.

Wannamaker, P. E., Booker, J. R., Jones, A. G., Chave, A. D., Filloux, J. H., Waff, H. S., Law, L. K. (1989), "Resistivity cross section through the Juan de Fuca subduction system and its tectonic implications," *J. Geophys. Res.*, **94**, 14 127–14 144.

Further Reading

Cain, J. C. (1987), "The earth as a magnet," in: S.-I. Akasofu, Y. Kamide (Eds.), *The Solar Wind and the Earth,* Tokyo: Terra Sci, pp. 56–69.

Chave, A. D., Booker, J. R. (1987), "Electromagnetic induction studies," *Rev. Geophys.* **25**, 989–1003.

Hargraves, J. K. (1979), *The Upper Atmosphere and Solar-Terrestrial Relations,* New York: Van Nostrand Reinhold.

Jacobs, J. A. (1987–1991), *Geomagnetism,* London: Academic, Vols. 1–4.

Kennel, C. F., Lanzerotti, L. J. Parker, E. N. (1979), *Solar System Plasma Physics,* Amsterdam: North Holland, Vols. 1–3.

Langel, R. A. (1991), "Recent progress in estimating uncertainty in geomagnetic field modeling," *Geophys. Astrophys. Fluid Dyn.* **60**, 37–88.

Lanzerotti, L. J. (1987), "Solar-Terrestrial Physics," in: *Encyclopedia of Physical Science and Technology,* Orlando, FL: Academic, Vol. 12, p. 834.

Merrill, R. T., McElhiny, M. W. (1983), *The Earth's Magnetic Field,* London: Academic.

Moffatt, H. K. (1978), *Magnetic Field Generation in Electrically Conducting Fluids,* Cambridge, UK: Cambridge University Press, 343 pp.

Nabighian, M. N. (Ed.) (1987), *Electromagnetic Methods in Applied Geophysics,* Tulsa, OK: Society of Exploration Geophysicists, Vol. 1.

Park, G. K. (1991), *Physics of Space Plasmas,* Redwood City, CA: Addison-Wesley.

Parkinson, W. D. (1983), *Introduction to Geomagnetism,* New York: Elsevier.

Rokityansky, I. I. (1982), *Geoelectromagnetic Investigation of the Earth's Crust and Mantle,* Heidelberg: Springer.

Uman, M. A. (1987), *The Lightning Discharge,* Orlando, FL: Academic.

GEOMETRICAL METHODS

V. ALAN KOSTELECKÝ, *Physics Department, Indiana University, Bloomington, Indiana, U.S.A.*

	Introduction	125
1.	**Analytic Geometry**	126
1.1	Plane Analytic Geometry	126
1.2	Conic Sections	127
1.3	Plane Trigonometry	128
1.4	Curvilinear Coordinates	128
1.5	Solid Analytic Geometry	129
1.6	Example: The Kepler Problem	131
2.	**Differential Geometry**	132
2.1	Manifolds	132
2.2	Vectors and One-Forms	132
2.3	Tensors	134
2.4	Differential Forms	135
2.5	Fiber Bundles	136
2.6	Connection and Curvature	137
2.7	Example: Electrodynamics	138
2.8	Complex Manifolds	140
2.9	Global Considerations	141
2.10	Further Examples	142
3.	**Projective Geometry**	143
3.1	Some Theorems	143
3.2	Homogeneous Coordinates	144
3.3	Group of Projective Transformations	145
4.	**Algebraic Geometry**	145
4.1	Affine Varieties	145
4.2	Projective Varieties	146
4.3	Classification	146
	Glossary	146
	Further Reading	148

INTRODUCTION

The word "geometry" derives from Greek, meaning "earth measurement." Geometry was originally the mathematics describing the shapes of objects and their spatial relationships. Simple geometrical notions and ideas were known to ancient Babylonians and Egyptians 4000 years ago. Starting approximately 2500 years ago, the ancient Greeks developed fundamental geometrical ideas, including some relatively rigorous proofs based on logical reasoning. Dating from this era is Euclid's *Elements*, which introduced the basis for the axiomatic method and summarizes the knowledge at that time.

Prior to the sixteenth century, geometry and algebra were treated as independent subjects. The notion of combining the two was introduced in 1631 by René Descartes (1596–1650). This led to the field of analytic geometry, which permits the investigation of geometric questions using analytical methods. This area was extensively investigated in the eighteenth century, in particular by Leonhard Euler (1707–1783) and Gaspard Monge (1746–1818). Toward the end of the eighteenth century the use of calculus resulted in the beginnings of differential geometry, studied by Christian Gauss (1777–1855) and others. The introduction by Bernhard Riemann (1826–1866) of the theory of algebraic functions initiated the field of algebraic geometry. In parallel with these developments, the synthetic approach to geometry was extended by Victor Poncelet (1788–1867), who formulated postulates for projective geometry. In the past century and a half, the work of David Hilbert (1862–1943) and others has led to an extension of the scope of geometry to include the study of geometrical relationships between abstract quantities.

This article presents material concerning analytical, differential, projective, and algebraic geometry. The choice of topics and their depth of coverage were dictated primarily by consideration of their importance in applied physics and by limitations of space. In particular, the reader is warned that the weighting assigned to the topics discussed is uncorrelated with their present importance as mathematical fields of research. The treatment is not mathematically rigorous, but introduces sufficient mathematical terminology to make basic textbooks in the subject accessible. Some

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

of these are listed in the references at the end of the article.

1. ANALYTIC GEOMETRY

The underlying concepts of analytic geometry are the simple geometric elements: points, lines and curves, planes and surfaces, and extensions to higher dimensions. The fundamental method is the use of coordinates to convert geometrical questions into algebraic ones. This is called the "method of coordinates."

To illustrate the basic notion, consider a straight line l. Following the method of coordinates, select one point O on l as the origin. This separates l into two halves. Call one half positive, the other negative. Any point P on the line can then be labeled by a real number, given by the distance OP for the positive half and by the negative of the distance OP for the negative half. There is thus a unique real number x assigned to every point P on l, called the Cartesian coordinate of P. Geometrical questions about the line can now be transcribed into analytical ones involving x.

1.1 Plane Analytic Geometry

In two dimensions, basic geometric entities include points, lines, and planes. For a plane π the method of coordinates provides to each point P an assignment of two real numbers, obtained as follows. Take two straight lines in the plane, and attribute Cartesian coordinates to each line as described above. For simplicity, the lines will be assumed perpendicular and intersecting at their origins. These lines are called rectangular coordinate axes, and the Cartesian coordinates of the first are called abscissae while those of the second are called ordinates. The lines themselves are also referred to as the abscissa and the ordinate. The location of a point P on π is then specified uniquely by two real numbers, written (x,y). The number x is defined as the perpendicular distance to the first coordinate axis, while y is the distance to the second. Using these Cartesian coordinates, geometrical questions about points can be expressed in analytical terms. For example, a formula for the distance d between two points P and Q specified by the coordinates (x_1,y_1) and (x_2,y_2) is

$$d=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}. \tag{1}$$

Given an assignment of Cartesian coordinates on a plane π, a curve segment s in the plane may be analytically specified by providing a set of paired real numbers (x,y) assigned to all points on the curve. In many useful cases, s can be specified by an equation $f(x,y)=0$ between x and y that is satisfied by all points P on s but not by any other points on π. For example, the equation $x=0$ describes the straight line consisting of all points having coordinates of the form $(0,y)$, i.e., the ordinate. The method of coordinates thus permits geometrical questions about s to be transcribed into analytical ones concerning $f(x,y)$. For example, the set of points lying both on a curve $f(x,y)=0$ and on another curve $g(x,y)=0$ is specified by values (x,y) satisfying both equations, which can in principle be found by analytical methods.

A simple example of a curve in the plane is a straight line l. The slope m of l can be defined in terms of the coordinates (x_1,y_1) and (x_2,y_2) of any two distinct points on l. Provided $x_1 \neq x_2$, the slope is given by

$$m=\frac{y_1-y_2}{x_1-x_2}. \tag{2}$$

The slope is zero for lines parallel to the abscissa, and is undefined (infinite) for lines parallel to the ordinate. A line l with given finite slope m is uniquely specified by its intersection point $(x,y)=(0,c)$ with the ordinate. The equation for l is

$$y=mx+c. \tag{3}$$

If l is parallel to the ordinate instead, it is determined by its intersection point $(x,y)=(a,0)$ with the abscissa, and its equation is simply $x=a$.

The equation of a straight line l is also determined entirely by the coordinates (x_1,y_1) and (x_2,y_2) of any two distinct points on l. It can be written

$$\frac{y-y_1}{y_2-y_1}=\frac{x-x_1}{x_2-x_1}. \tag{4}$$

Alternatively, a straight line can be viewed as the curve given by the most general equation linear in the coordinates x and y:

$$Ax+By+C=0, \tag{5}$$

where at least one of A and B is nonzero.

Analytical solutions to geometrical problems involving straight lines and points can be obtained using the above results. For example, the equation of the line l_P that is perpendicular to a given line l with equation $y=mx+c$ and that passes through the point P on l with abscissa x_1 is

$$y=-\frac{1}{m}x+\frac{m^2+1}{m}x_1+c. \tag{6}$$

Another example is the expression for the perpendicular distance d_P between a line l with equation $y=mx+c$ and a point P at (a,b), which is

$$d_P=|b-ma-c|/\sqrt{m^2+1}. \tag{7}$$

1.2 Conic Sections

An important curve is the circle, denoted by S^1, which can be viewed as the set of points in the plane that are equidistant from a specified fixed point. The fixed point C is called the center of the circle, and the distance r between the center and the points on the circle is called the radius. If the Cartesian coordinates of C are (h,k), then the equation of the circle is

$$(x-h)^2+(y-k)^2=r^2. \tag{8}$$

The circle is a special case of a set of curves called conic sections or conics. These curves include ellipses, parabolas, and hyperbolas. Geometrically, the conics can be introduced as the curves obtained by slicing a right circular cone with a plane. Analytically, they can be viewed as the curves given by the most general expression quadratic in the coordinates x and y:

$$Ax^2+Bxy+Cy^2+Dx+Ey+F=0, \tag{9}$$

where at least one of the coefficients A, B, C is nonzero. These are called second-order curves. From this equation it follows that any five points lying on the conic specify it completely. The quantity B^2-4AC is called the discriminant of the conic. If the discriminant is positive, the conic is a hyperbola; if negative, an ellipse; and if zero, a parabola.

A third definition, combining geometrical and analytical notions, is often useful. Consider a straight line l, a fixed point F not on l, and another point P. Denote the distance between P and F by d_F and the perpendicular distance between P and l by d_l. Then the conic sections are given by the set of points P that obey the equation

$$d_F=ed_l, \tag{10}$$

where $e \geqslant 0$ is a constant real number called the eccentricity. The line l is called the directrix and the point F is called the focus. If $e>1$, the conic is a hyperbola. If $e=1$, the conic is a parabola. If $0<e<1$, the conic is an ellipse. The degenerate case $e=0$ gives a circle; in this case, the directrix is at infinity.

The equation determining a conic has a particularly simple form, called the canonical form, if the focus F is chosen to lie on the abscissa and the directrix l is chosen parallel to the ordinate. The canonical form depends on at most two real positive parameters a and b, where $a \geqslant b$ is taken for convenience.

For a hyperbola, the canonical form is

$$(x^2/a^2)-(y^2/b^2)=1. \tag{11}$$

The eccentricity of the hyperbola is $e=\sqrt{a^2+b^2}/a$. One focus is the point $(ae,0)$, and the corresponding directrix is the line $x=a/e$. There is a second focus at $(-ae,0)$ and a second directrix at $x=-a/e$. The hyperbola has two branches, each of which asymptotically approaches the lines $y=\pm bx/a$ as $|x|$ becomes large. The distance between the points where the hyperbola intersects the abscissa is $2a$. This is also the difference between the distances from the two foci to any given point on the hyperbola. If $a=b$, the hyperbola is called rectangular.

For a parabola, the canonical form is

$$y^2=4ax. \tag{12}$$

The eccentricity is $e=1$, the focus is at $(a,0)$, and the directrix is the line $x=-a$.

For an ellipse, the canonical form is

$$(x^2/a^2)+(y^2/b^2)=1. \tag{13}$$

The ellipse has eccentricity $e=\sqrt{a^2-b^2}/a$. There are again two foci, at $(\pm ae,0)$, and two directrices $x=\pm a/e$. The sum of the distances from the two foci to any given point on the ellipse is a constant, $2a$. The line between the points of intersection of the ellipse with the abscissa is called the major axis of the ellipse, and it has length $2a$. Similarly, the minor axis of the ellipse is given by the intersection points with the ordinate and has length $2b$. If

$a=b$ the equation reduces to that of a circle of radius a centered at the origin.

1.3 Plane Trigonometry

Consider a point P with coordinates (x,y) lying on a circle of radius r centered at the origin O. Denote by X the point $(x,0)$. Call θ the angle XOP between the line segments OX and OP. The choice of a unit of measure for angles permits the assignment of a numerical value to θ. One widely used unit is the degree, defined by the statement that there are 360 degrees in a circle. The SI unit is the radian, of which there are 2π in a circle.

Certain functions of the angle θ, called trigonometric or circular functions, are of particular use in plane analytic geometry. The ratio $\sin\theta=y/r$ is called the sine of θ while $\cos\theta=x/r$ is called the cosine of θ. The sine is odd in θ while the cosine is even, and both functions have period π radians. They obey the relations

$$\sin^2\theta+\cos^2\theta=1 \tag{14}$$

following from the Pythagorean theorem, and

$$\sin(\theta\pm\phi)=\sin\theta\,\cos\phi\pm\sin\phi\,\cos\theta, \tag{15}$$

$$\cos(\theta\pm\phi)=\cos\theta\,\cos\phi\mp\sin\theta\,\sin\phi. \tag{16}$$

The latter two equations are called addition formulas. Other, related functions of θ include the tangent $\tan\theta=y/x=\sin\theta/\cos\theta$, the cosecant $\csc\theta=r/y=1/\sin\theta$, the secant $\sec\theta=r/x=1/\cos\theta$, and the cotangent $\cot\theta=x/y=\cos\theta/\sin\theta$. From these definitions and the above equations many identities can be obtained. Inverse trigonometric functions can also be introduced; for example, if $x=\sin\theta$ then $\sin^{-1}x=\theta$.

Consider a triangle with angles A, B, C and sides of length a, b, c, where by convention the side labeled by a is opposite the vertex with angle A and there are similar conventions for the other sides. A basic problem in plane trigonometry is to determine one of a, b, c, A, B, C in terms of the others. This is called solving a triangle. The following relations hold: the law of sines,

$$\frac{\sin A}{a}=\frac{\sin B}{b}=\frac{\sin C}{c}; \tag{17}$$

the first law of cosines,

$$a=b\cos C+c\cos B; \tag{18}$$

and the second law of cosines,

$$a^2=b^2+c^2-2bc\cos A. \tag{19}$$

1.4 Curvilinear Coordinates

For certain geometrical problems, the analytical details of a calculation may be simplified if a non-Cartesian coordinate system is used. Consider two functions $u=u(x,y)$ and $v=v(x,y)$ of the Cartesian coordinates x and y on a plane π. Take the functions to be continuous and invertible, except perhaps at certain special points that require separate treatment. Any curve $u=c$ for some constant c is called a coordinate curve, as is any curve $v=c$. A point P on π is uniquely specified by two real numbers (u_1,v_1) that are the values of the constants determining the two coordinate curves passing through P. This construction generalizes the method of coordinates, and the functions u and v are called curvilinear coordinates. If the coordinate curves meet at right angles, the curvilinear coordinates are called orthogonal. All the analytical geometry described above using Cartesian coordinates can be rephrased using orthogonal curvilinear coordinates.

An important set of orthogonal curvilinear coordinates is generated by the equations

$$x=r\cos\theta,\quad y=r\sin\theta;$$

$$r=\sqrt{x^2+y^2},\quad \theta=\tan^{-1}(y/x), \tag{20}$$

where $r\geqslant 0$ and $0\leqslant\theta<2\pi$. In this system, the coordinate curves consist of circles of varying radii centered at the origin and straight lines through the origin at varying angles with respect to the abscissa. The coordinates (r,θ) of a point P are called plane polar coordinates. As an illustration of their use, consider the conic sections expressed in polar coordinates. In canonical form, with the origin of the polar coordinates placed at the focus at $(ae,0)$, the equation for a conic section is

$$r=l/(1+e\cos\theta), \tag{21}$$

where l is called the latus rectum. It is given by $l=b^2/a$ for hyperbolas and ellipses and by $l=2a$ for parabolas, and it represents the distance from the focus to the curve as measured along a straight line parallel to the ordinate. The quantity l/e is the distance from the focus to the associated directrix.

The conic sections themselves can be used to generate systems of orthogonal curvilinear coordinates. For example, parabolic coordinates can be defined by

$$x=\tfrac{1}{2}(u^2-v^2), \quad y=uv, \tag{22}$$

where $v \geqslant 0$. The coordinate curves are parabolas. Similarly, elliptic coordinates can be defined by

$$x=a \cosh u \cos v, \quad y=a \sinh u \sin v, \tag{23}$$

where $u \geqslant 0$ and $0 \leqslant v < 2\pi$. Here, the so-called hyperbolic functions $\sinh u$ and $\cosh u$ are defined by

$$\sinh u=\tfrac{1}{2}(e^u-e^{-u}), \quad \cosh u=\tfrac{1}{2}(e^u+e^{-u}). \tag{24}$$

The coordinate curves are ellipses and hyperbolas. Another common set is the system of bipolar coordinates, defined by

$$x=\frac{a \sinh v}{\cosh v-\cos u}, \quad y=\frac{a \sin u}{\cosh v-\cos u}, \tag{25}$$

with $0 \leqslant u < 2\pi$. The coordinate curves are sets of intersecting circles.

1.5 Solid Analytic Geometry

Solid analytic geometry involves the study of geometry in three dimensions rather than two. Many of the ideas of plane analytic geometry extend to three dimensions. For instance, the method of coordinates now provides an assignment of three real numbers (x,y,z) to each point P. A three-dimensional rectangular coordinate system can be introduced by taking three mutually perpendicular straight lines, each given Cartesian coordinates, to form the coordinate axes. The axes are called the abscissa, the ordinate, and the applicate. Each of the values (x,y,z) is defined as the perpendicular distance to the corresponding axis.

A two-dimensional surface σ can now be specified by providing an equation $f(x,y,z)=0$ that is satisfied only by points on the surface. The method of coordinates thus converts geometrical questions about σ to analytical questions about $f(x,y,z)$. Similarly, a curve s can be viewed as the intersection set of two surfaces. If the surfaces are specified by the equations $f(x,y,z)=0$ and $g(x,y,z)=0$, s is given analytically by the set of points (x,y,z) obeying both equations simultaneously.

By definition, a surface of the first order is given by the most general equation linear in x, y, z:

$$Ax+By+Cz+D=0. \tag{26}$$

If at least one of A, B, C is nonzero, this equation describes a plane. A straight line can be viewed as the intersection of two nonparallel planes and is therefore given analytically by two equations of this form. Just as in the two-dimensional case, the analytical formulation allows solutions to geometrical questions involving planes, lines, and points to be obtained. For example, the perpendicular distance d_P between a plane given by the above equation and a point P located at (a,b,c) can be shown to be

$$d_P=\frac{|Aa+Bb+Cc+D|}{\sqrt{A^2+B^2+C^2}}. \tag{27}$$

As another example, two planes given by

$$A_1x+B_1y+C_1z+D_1=0,$$

$$A_2x+B_2y+C_2z+D_2=0 \tag{28}$$

are parallel if and only if

$$(A_1,B_1,C_1)=(cA_2,cB_2,cC_2) \tag{29}$$

for some constant c.

In analogy to the two-dimensional introduction of conics as curves obeying a quadratic expression in x and y, a surface of the second order is defined to consist of points satisfying a quadratic expression in x, y, and z:

$$Ax^2+By^2+Cz^2+Dxy+Exz+Fyz+Gx +Hy+Iz+J=0. \tag{30}$$

Such surfaces are also called quadrics. An important example is the sphere, denoted by S^2, which can be viewed as the set of points equidistant from a fixed point called the center. The distance from the center to any point on the sphere is called the radius. If the Cartesian coordinates of the center are (h,k,l), the equation of a sphere of radius r is

$$(x-h)^2+(y-k)^2+(z-l)^2=r^2. \tag{31}$$

The quadrics can be classified. Among the surfaces described are ellipsoids, hyperboloids, paraboloids, cylinders, and cones. Canonical forms of these surfaces are

$$(x^2/a^2)+(y^2/b^2)+(z^2/c^2)=1 \tag{32}$$

for an ellipsoid;

$$(x^2/a^2)+(y^2/b^2)-(z^2/c^2)=1 \tag{33}$$

for a hyperboloid of one sheet;

$$(x^2/a^2)-(y^2/b^2)-(z^2/c^2)=1 \tag{34}$$

for a hyperboloid of two sheets;

$$(x^2/a^2)+(y^2/b^2)=2z \tag{35}$$

for an elliptic paraboloid;

$$(x^2/a^2)-(y^2/b^2)=2z \tag{36}$$

for a hyperbolic paraboloid;

$$(x^2/a^2)+(y^2/b^2)=1 \tag{37}$$

for an elliptic cylinder;

$$(x^2/a^2)-(y^2/b^2)=1 \tag{38}$$

for a hyperbolic cylinder;

$$(x^2/a^2)=2z \tag{39}$$

for a parabolic cylinder; and

$$(x^2/a^2)\pm(y^2/b^2)-(z^2/c^2)=0 \tag{40}$$

for a cone. The parameters a, b, c are called the lengths of the principal axes of the quadric.

The notions of plane trigonometry also extend to three dimensions. A spherical triangle is defined as a portion of a spherical surface that is bounded by three arcs of great circles. Denote by A, B, C the angles generated by straight lines tangent to the great circles intersecting at the vertices, and call the lengths of the opposite sides a, b, c as for the planar case. The angles now add up to more than π radians, by an amount called the spherical excess E:

$$A+B+C=\pi+E. \tag{41}$$

The following relations hold for a spherical triangle: the law of sines,

$$\frac{\sin A}{\sin a}=\frac{\sin B}{\sin b}=\frac{\sin C}{\sin c}; \tag{42}$$

the first law of cosines,

$$\cos a=\cos b\cos c+\sin b\sin c\cos A; \tag{43}$$

and the second law of cosines,

$$\cos A=-\cos B\cos C+\sin B\sin C\cos a. \tag{44}$$

Curvilinear coordinates can be introduced via three locally continuous invertible functions $u(x,y,z)$, $v(x,y,z)$, $w(x,y,z)$, following the two-dimensional case. A coordinate surface is specified by setting any curvilinear coordinate u, v, or w equal to a constant. The coordinate curves are generated by the intersection of the coordinate surfaces, and the system is said to be orthogonal if the surfaces intersect at right angles. Many useful three-dimensional orthogonal curvilinear coordinate systems can be generated from families of quadrics. One particularly useful set is the system of spherical polar coordinates, defined by

$$\begin{gathered} x=r\sin\theta\cos\phi,\\ y=r\sin\theta\sin\phi,\\ z=r\cos\theta, \end{gathered} \tag{45}$$

where $r\geqslant 0$, $0\leqslant\theta\leqslant\pi$, and $0\leqslant\phi<2\pi$. The coordinate surfaces are spheres centered at the origin, right circular cones with axes along the applicate and vertices at the origin, and half-planes with the applicate as one edge. Other common coordinates are the cylindrical coordinates, given by

$$x=r\cos\theta,\quad y=r\sin\theta,\quad z=z, \tag{46}$$

where $r\geqslant 0$ and $0\leqslant\theta<2\pi$. This system is generated from plane polar coordinates by translation along the applicate. The coordinate surfaces are right circular cylinders centered at the origin, half-planes with the applicate as one edge, and planes parallel to the plane of the abscissa and ordinate.

The notions of plane and solid analytic geometry can be extended to higher dimensions, too. A space can be defined in which the method of coordinates specifies a point by n real numbers $(x^1,x^2,\dots,x^n)$. This n-dimensional space, called Euclidean space, is denoted by the symbol R^n. Using coordinates, geometrical questions in n dimensions can be converted to analytical ones involving functions of n variables. Surfaces of the first order are $(n-1)$-dimensional hyperplanes, and surfaces of the second order, or quadric hypersurfaces, can be introduced. An example is the hypersphere of radius r in n dimensions, denoted by S^{n-1}, which when centered at the origin satisfies the equation

$$(x^1)^2+(x^2)^2+\cdots+(x^n)^2=r^2. \tag{47}$$

The notion of curvilinear coordinates also extends to higher dimensions.

A one-dimensional curve s in n dimensions can be specified by $n-1$ equations among the n coordinates $(x^1,...,x^n)$. If s is continuous, its points can be labeled by a parameter t that is a real number. Any particular point can be specified by giving the values of the n coordinates $(x^1,...,x^n)$. As t varies, so do the coordinates. This means that an alternative specification of s can be given in terms of the n expressions

$$x^j=x^j(t), \quad j=1,...,n, \tag{48}$$

determining the n coordinates $(x^1,...,x^n)$ as functions of t. This is called the parametric representation of a curve. Similarly, the points of a continuous two-dimensional surface can be labeled by two real numbers (t^1,t^2). The surface can be specified either in terms of $n-2$ equations among the n coordinates $(x^1,...,x^n)$ or in parametric form by the n equations

$$x^j=x^j(t^1,t^2), \quad j=1,...,n. \tag{49}$$

A parametric representation can also be given for continuous surfaces of more than two dimensions.

1.6 Example: The Kepler Problem

An example of the appearance of analytic geometry in a physical problem occurs in the study of the classical motion of two bodies under a mutual inverse-square attractive force. Consider for definiteness two bodies of masses m_1 and m_2, each acted on by the gravitational field of the other and free to move in three dimensions. This is called the Kepler problem.

The first step is to introduce a convenient coordinate system. For simplicity, the origin can be placed on one mass. The problem can then be reduced to determining the relative position of the second mass and the uniform motion of the center of mass. The latter is neglected here for simplicity. It is natural to select a system of spherical polar coordinates with the applicate along the direction of the angular momentum. Since angular momentum is conserved, the motion of the second mass about the origin must lie in a plane. This means that plane polar coordinates (r,θ) suffice to describe the position of the second mass relative to the first.

It can be shown that the resulting equations governing the motion of the second mass are precisely those obtained for the behavior of a reduced mass $m=m_1m_2/(m_1+m_2)$ orbiting a fixed center of force. In polar coordinates, the kinetic energy T of the reduced mass is

$$T=\tfrac{1}{2}m(\dot{r}^2+r^2\dot{\theta}^2), \tag{50}$$

where a dot over a letter signifies a derivative with respect to time. The potential energy is

$$V=-k/r \tag{51}$$

with $k=Gm_1m_2$, where G is Newton's gravitational constant.

The equations of motion are

$$\frac{d}{dt}(mr^2\dot{\theta})=0 \tag{52}$$

and

$$m\ddot{r}-mr\dot{\theta}^2+(k/r^2)=0. \tag{53}$$

The first of these integrates immediately in terms of the constant magnitude L of the angular momentum:

$$mr^2\dot{\theta}=L. \tag{54}$$

This equation can be used to eliminate $\dot{\theta}$ from Eq. (53) by direct substitution. Also, since

$$\frac{d}{dt}=\dot{\theta}\frac{d}{d\theta}, \tag{55}$$

the independent variable in Eq. (53) can be converted from time t to angle θ. An additional change of dependent variable from r to

$$s=(1/r)-(mk/L^2) \tag{56}$$

converts Eq. (53) into the simple form

$$\frac{d^2s}{d\theta^2}=-s. \tag{57}$$

The solution is readily found. Reconverting s to the variable r yields the equation for the orbit as

$$r=l/(1+e\cos\theta), \tag{58}$$

where a particular choice for the location of the abscissa relative to the orbit has been made for simplicity. In this equation,

$$e=\sqrt{1+2EL^2/mk^2},\quad l=L^2/mk, \tag{59}$$

and E can be identified with the energy of the two bodies in the orbit. This demonstrates that the motion of two masses under gravity is described by a conic section; cf. Eq. (21). The energy E determines the shape of the orbit. If $E>0$, $e>1$ and the orbit is a hyperbola. If $E=0$, $e=1$ and the orbit is a parabola. If $E<0$, $e<1$ and the orbit is an ellipse. Finally, if $E=-mk^2/2L^2$, $e=0$ and the orbit is a circle.

2. DIFFERENTIAL GEOMETRY

The requirement of differentiability provides a restriction on geometrical objects that is sufficiently tight for new and useful results to be obtained and sufficiently loose to include plenty of interesting cases. Differential geometry is of vital importance in physics because many physical problems involve variables that are both continuous and differentiable throughout their range.

2.1 Manifolds

A manifold is an extension of the usual notions of curves and surfaces to arbitrary dimensions. The basic idea is to introduce an n-dimensional manifold as a space that is like Euclidean space R^n locally, i.e., near each point. Globally, i.e., taken as a whole, a manifold may be very different from R^n. An example of a one-dimensional manifold is a straight line. This is both locally and globally like R^1. Another one-dimensional example is a circle S^1. The neighborhood of each point on a circle looks like the neighborhood of a point in R^1, but globally the two are different. The circle can be constructed by taking two pieces of R^1, bending them, and attaching them smoothly at each end. Generalized to n dimensions, this notion of taking pieces of R^n and attaching them smoothly forms the basis for the definition of a manifold.

To define a manifold more rigorously, first introduce the concept of a topological space T. This is a set S and a collection t of (open) subsets of S satisfying the following criteria:

1. Both the null set and S itself are in t.
2. The intersection of any two subsets of t is in t.
3. The union of any collection of subsets of t is in t.

Suppose in addition there is a criterion of separability: For any two elements of S there exist two disjoint subsets of S, each containing one of the elements. Then T is called a Hausdorff space. The elements of S for a manifold are its points.

Next, define a chart C of the set S as a subset U of S, called a neighborhood, together with a continuous invertible map $f:U\to R^n$ called the coordinate function. For a manifold, the subset U plays the role of a region locally like R^n, and the function f represents the introduction of local coordinates in that region. Two charts C_1, C_2 with overlapping neighborhoods and coordinate functions f_1, f_2 are called compatible if the composition map $f_1\circ f_2^{-1}$ is differentiable. The requirement of compatibility basically ensures that the transition from one coordinate patch to another is smooth. A set of compatible charts covering S is called an atlas.

A differentiable manifold M can now be defined as a Hausdorff topological space with an atlas. Given that the range of the coordinate functions is R^n, the dimension of M is defined as n and M is sometimes denoted by M^n. An example of an n-dimensional manifold is the hypersphere S^n. An example of an object that is *not* a manifold is a figure-eight curve, since the neighborhood of the intersection point is not locally like R^n for any n.

2.2 Vectors and One-Forms

The usual definition of a vector in a Euclidean space as a directed straight-line segment does not immediately extend to a general manifold. For instance, the circle S^1 does not contain any straight-line segments. Instead, vectors at a point of a manifold can be introduced using the notion of the tangents to curves passing through the point.

Consider a curve s through a point P. In a neighborhood of P, local coordinates $(x^1,\dots,x^n)$ can be used to specify s in the parametric representation $x^j=x^j(t)$, $j=1,\dots,n$. A vector tangent to s at P can be specified by the n quantities dx^j/dt forming its components. A familiar example in mechanics is the velocity vector of a moving particle, obtained by differentiation with respect to time of the particle's position vector. If the tangent vectors to all possible curves in the manifold through P are considered, an n-dimensional vector space (see ALGEBRAIC METHODS, Sec. 3) is generated. This is called the tangent space T_PM to M at P.

In differential geometry, it is desirable to introduce basic concepts in a manner that is independent of any coordinate choice. For this reason, the differential-geometric definition of a tangent vector is different from the more intuitive one above. Given a curve s, introduce an arbitrary differentiable function f assigning a real number to every point t on s. The derivative df/dt of f along s is called the directional derivative. In a local coordinate patch,

$$\frac{df}{dt}=\sum_{j=1}^{n}\frac{dx^j}{dt}\partial_j f, \tag{60}$$

where $\partial_j f=\partial f/\partial x^j$. This shows that the operator d/dt acting on the space of real functions on M contains all components of the tangent vector, each associated with the corresponding partial derivative ∂_j. A tangent vector at P can therefore be defined as the directional-derivative operator d/dt, with a natural coordinate basis of vectors for the vector space being the set of partial-derivative operators $\{\partial_j\}$. However, this definition has the disadvantage that it still explicitly includes the parameter t.

The formal definition of a tangent vector is therefore slightly more abstract. Given the space $F(M)$ of all differentiable real functions on a manifold M, a tangent vector at P is defined as an object v acting on elements of $F(M)$ to produce real numbers,

$$v:F(M)\rightarrow R, \tag{61}$$

that satisfies two criteria:

$$v(af+bg)=av(f)+bf(g),$$

$$v(f\circ g)=g(P)v(f)+f(P)v(g), \tag{62}$$

where $f,g\in F(M)$ and $a,b\in R$. This definition extracts the important properties of the tangent vector without explicit reference to a coordinate system or parametrization. Note that the coordinate realization of a tangent vector at P along x_j as ∂_j acting at P satisfies this definition. The set of all linearly independent tangent vectors at P spans the tangent space T_PM to M at P, and the set $\{\partial_j\}$ forms a basis for T_PM called the coordinate basis. An arbitrary vector v can be expanded in this basis as $v=\Sigma_j v^j\partial_j$. Physicists sometimes say that the components v^j are the contravariant components of a vector. Although in a coordinate basis the intuitive physics notion of a vector and the differential-geometric one contain the same information about components, the latter also contains information about the coordinate basis itself. In the remainder of this article except where noted, the word vector refers to the differential-geometric object.

Since T_PM is a vector space, there exists a dual vector space $\mathrm{Hom}(T_PM,R)$ consisting of linear maps

$$\omega:T_PM\rightarrow R \tag{63}$$

(see ALGEBRAIC METHODS, Sec. 3.7). This space is called the cotangent space at P and is denoted by T^*_PM. Notice that duality also implies $T_PM=\mathrm{Hom}(T^*_PM,R)$. Elements of T^*_PM are called one-forms. An important example of a one-form is the total differential df of a function $f\in F(M)$, defined as the element of T^*_PM satisfying

$$df(v)=v(f) \tag{64}$$

for any $v\in T_PM$.

In a chart around P, the set $\{dx^j\}$ of total differentials of the coordinates forms a natural coordinate basis for the cotangent space T^*_PM. It is a dual basis to $\{\partial_j\}$, since

$$dx^j(\partial_k)=\partial_k x^j=\delta^j_k. \tag{65}$$

An arbitrary one-form ω can be expanded in the dual basis as $\omega=\Sigma_j\ \omega_j dx^j$. Note that for an arbitrary vector $v=\Sigma_j v^j\partial_j$ the action of ω on v is then

$$\omega(v)=\omega_j v^k dx^j(\partial_k)=\omega_j v^j. \tag{66}$$

In this equation and subsequent ones, the Einstein summation convention is introduced to simplify notation: Repeated indices in the same term are understood to be summed. The vector v is said to be contracted with the one-form ω. Physicists sometimes say the components ω^j form the covariant components of a vector. As an example, the definitions above can be used to show that

$$df=\partial_j f dx^j, \tag{67}$$

a standard result.

2.3 Tensors

The generalization of vectors and one-forms to tensors is straightforward. A tensor T of type (a,b) can be defined at a point P of a manifold M as a multilinear mapping of a one-forms and b vectors giving a real number:

$$T\colon T^*_P M \otimes \cdots \otimes T^*_P M \otimes T_P M \otimes \cdots \otimes T_P M \to R, \quad (68)$$

where there are a factors of $T^*_P M$ and b factors of $T_P M$. The space of tensors of type (a,b) at P is denoted $T^a_b(P)$. Examples introduced above include $T^1_0(P) = T_P M$ and $T^0_1(P) = T^*_P M$.

A tensor T of type (a,b) can be expanded using a coordinate basis. In the natural basis introduced above,

$$T = T^{j_1 j_2 \cdots j_a}_{k_1 k_2 \cdots k_b} \partial_{j_1} \cdots \partial_{j_a} dx^{k_1} \cdots dx^{k_b}. \quad (69)$$

Almost all physicists and the older mathematics literature call the quantities $T^{j_1 j_2 \cdots j_a}_{k_1 k_2 \cdots k_b}$ the components of an ath-rank contravariant and bth-rank covariant tensor. Most modern mathematicians by convention interchange the usage of contravariant and covariant. This article uses the physicists' convention.

A tensor is called symmetric with respect to two contravariant or two covariant indices if its components are unaltered when the indices are interchanged. A tensor with indices of only one type is said to be totally symmetric if it is symmetric with respect to all pairs of indices. Similarly, a tensor is antisymmetric with respect to two contravariant or two covariant indices if its components change sign when the indices are interchanged, and a totally antisymmetric tensor is one with pairwise-antisymmetric indices of only one type. The sum and difference of two tensors of the same type is another tensor of the same type. The tensor product $T_1 \otimes T_2$ of two tensors T_1 and T_2 of types (a_1,b_1) and (a_2,b_2), respectively, is a tensor of type (a_1+a_2,b_1+b_1) with components given by the product of components of T_1 and T_2 (see ALGEBRAIC METHODS, Sec. 3.8). Various contractions of two tensors can be introduced that generalize the contraction of a vector with a one-form.

A useful concept in physical applications is that of a tensor field of type (a,b), defined as a particular choice of tensor of type (a,b) at each point of M. The field is called smooth if the components $T^{j_1 j_2 \cdots j_a}_{k_1 k_2 \cdots k_b}$ of a tensor field are differentiable. Special cases are vector fields and one-form fields. By convention, if $a=b=0$ the field is called a scalar field and is just an element of $F(M)$, the real-valued functions on M.

An example of a tensor that plays a crucial role in physics is the metric tensor g. On the manifold M, it is a symmetric tensor field of type $(0,2)$ such that if $g(v_1,v_2)=0$ for any $v_1 \in T_P M$, then $v_2=0$. In component form in a coordinate basis near a point P

$$g = g_{jk} dx^j dx^k, \quad (70)$$

where g_{jk} form the components of a symmetric, invertible matrix. The metric tensor g associates any two vectors with a real number. For instance, in the usual geometry in a Euclidean space R^n the matrix $g_{jk} = \delta_{jk}$ and the real number is the scalar or dot product of the two vectors. In other applications different metrics may be required. For example, in special relativity space-time is taken as a four-dimensional manifold with a Minkowskian metric. If the number $g(v_1,v_2)$ has the same sign for all v_1, v_2 at all P on M, i.e., if the eigenvalues of the matrix g_{jk} are all of the same sign, the metric is called Riemannian. Manifolds admitting such metrics are called Riemannian manifolds. Other metrics are called pseudo-Riemannian. The special case of a metric with one eigenvalue of different sign is called Lorentzian. By diagonalization and normalization, it is always possible to choose a basis at any given P such that $g_{jk}(P)$ is a diagonal matrix with entries that are ± 1. If the entries are all of the same sign, the metric in this form is called Euclidean. If one entry has a different sign, it is called Minkowskian.

Since g is a map $T_P M \otimes T_P M \to R$, any given vector v defines a linear map $g(v, \ )$ from $T_P M$ to R. This map is evidently a one-form, by definition. The components v_j of this one-form are given by

$$v_j = g_{jk} v^k. \quad (71)$$

The map is said to lower the index of the vector, and the result is called the associated one-form. An inverse map can be defined that uses the matrix inverse g^{jk} of g_{jk} to raise the index of a form, yielding a vector.

A significant part of the classical literature on differential geometry is concerned with the relationships between different manifolds, in particular in manifolds endowed with metrics. Consider two manifolds M_1 and M_2 of dimensions n_1 and n_2. If there exists a smooth and regular map $f:M_1 \rightarrow M_2$, then M_1 is said to be a submanifold of M_2. The map f is called an embedding. The notion of a regular map is readily understood in coordinate patches $\{x^j\}$ on a chart U in M_1 and $\{y^k\}$ on a chart V in M_2: the matrix with components $\partial y^k/\partial x^j$ must have maximal rank n_1 at each point. Intuitively, the requirements for an embedding can be viewed as ensuring for the submanifold its differentiability, the absence of self-intersections, and that curves through a point in M_1 look locally like their images in M_2. The references at the end of this article provide details of the methods and results of this subject. A simple example of a question involving the notion of embedding is the determination of equations, called the Frenet-Serret formulas, for a curve in R^n. A more complicated example is the description of the embedding of a hypersurface M into R^n, which, according to Bonnet's theorem, is determined by the metric tensor g on M (which in this context is called the first fundamental form), by another symmetric tensor of type (0,2) called the second fundamental form, and by a set of partial differential equations called the Gauss-Codazzi equations. General results on the possibility of embedding an m-dimensional manifold into R^n are also available. An example is Whitney's theorem, which may be viewed as the statement that for compact manifolds such an embedding is possible for $n=2m+1$.

2.4 Differential Forms

A particularly important class of tensors is the set of totally antisymmetric tensors of type $(0,p)$ at a point of M^n. These span a vector space denoted by $\wedge^p T^*_P M$ or just $\wedge^p T^*$, and they are called p-forms. The number $p \leqslant n$ is called the degree of the form. For the case $p=0$, $\wedge^0 T^*_P M$ is chosen as $F(M)$, the space of real smooth functions on M. The dimension of $\wedge^p T^*$ as a vector space is given by the binomial coefficient ${}^nC_p$. Note that this implies that $\wedge^p T^*$ and $\wedge^{(n-p)} T^*$ have the same dimension.

Introduce the wedge product $\omega_1 \wedge \omega_2$ of two one-forms by the definition

$$\omega_1 \wedge \omega_2 = \omega_1 \otimes \omega_2 - \omega_2 \otimes \omega_1. \tag{72}$$

By construction, this is an antisymmetric tensor of type (0,2), i.e., a two-form. It can be shown that a coordinate basis for the two-forms is the set $\{dx^j \wedge dx^k\}$. In general, antisymmetric tensor products of one-forms can be used to generate p-forms, and an element $\omega \in \wedge^p T^*$ can be expanded in a coordinate basis as

$$\omega_p = \frac{1}{p!} \omega_{j_1 \cdots j_p} dx^{j_1} \wedge \cdots \wedge dx^{j_p}. \tag{73}$$

A natural induced wedge product exists that combines a p-form ω_1 with a q-form ω_2 to give a $(p+q)$-form. This product obeys

$$\omega_1 \wedge \omega_2 = (-1)^{pq} \omega_2 \wedge \omega_1. \tag{74}$$

A larger vector space $\wedge T^*$ consisting of the direct sum of all the spaces $\wedge^p T^*$ can also be considered. Its dimension is 2^n, and it is called the Cartan exterior algebra of $T^*_P M$.

Analogous constructions can be introduced for the case of antisymmetric tensors of type $(p,0)$, called p-vectors. The totality of these spans a space denoted $\wedge^p T$. The p-forms, $(n-p)$-forms, p-vectors, and $(n-p)$-vectors thus all form vector spaces of dimension ${}^nC_p$ at a point P of M^n. Various relations can be constructed between these spaces. An important example is the Hodge star map $*$, defined for manifolds M that have a metric g. This is a linear map $*: \wedge^p T^*_P M \rightarrow \wedge^{(n-p)} T^*_P M$ that is most easily understood by its action on coordinate components. Define the totally antisymmetric symbol by

$$e_{j_1 \cdots j_n} = \begin{cases} +1 & \text{if } (j_1 \ldots j_n) \text{ is an even permutation of } (1,\ldots,n) \\ -1 & \text{if } (j_1 \ldots j_n) \text{ is an odd permutation of } (1,\ldots,n) \\ 0 & \text{otherwise.} \end{cases} \tag{75}$$

If a p-form ω is given in a coordinate basis by Eq. (73), then

$$*\omega=\frac{\sqrt{|g|}}{p!(n-p)!}\, g^{j_1k_1}\cdots g^{j_pk_p}\omega_{k_1\ldots k_p}e_{j_1\ldots j_n}dx^{j_{p+1}} \wedge\cdots\wedge dx^{j_n}, \tag{76}$$

where g^{jk} is the inverse metric matrix introduced in Sec. 2.3 and g is the determinant of the matrix g_{jk}.

From the definition (64), the total differential of a zero-form is a one-form. An extension of the notion of differential can be introduced to obtain a $(p+1)$-form via a p-form. Formally, a map $d: \wedge^p T^* \rightarrow \wedge^{(p+1)} T^*$ called the exterior derivative can be defined by the following requirements:

1. $d(\omega_1+\omega_2)=d\omega_1+d\omega_2$ for $\omega_1,\omega_2 \in \wedge^p T$;
2. $d(\omega_1 \wedge \omega_2)=(d\omega_1 \wedge \omega_2)+(-1)^p(\omega_1 \wedge d\omega_2)$ for $\omega_1 \in \wedge^p T$ and $\omega_2 \in \wedge^q T$; and
3. $d(d\omega)=0$ for $\omega \in \wedge^p T$.

It can be shown that the exterior derivative is unique. In a coordinate basis, the exterior derivative of a p-form given by Eq. (73) is

$$d\omega_p=(1/p!)\partial_k\omega_{j_1\ldots j_p}dx^k \wedge dx^{j_1} \wedge\cdots\wedge dx^{j_p}. \tag{77}$$

A p-form field with vanishing exterior derivative is said to be closed, while one that is obtained as the exterior derivative of a $(p-1)$-form is called exact. The definition of d implies that an exact form is necessarily closed.

The exterior derivative combines in a single notation valid for manifolds M^n extensions of the gradient, divergence, and curl operations of usual three-dimensional vector calculus. For instance, the gradient of a function f is a covariant vector with components $\partial_j f$. These are precisely the components of the one-form in Eq. (67). The components of the curl make their appearance in the exterior derivative of a one-form $\omega=\omega_x dx+\omega_y dy+\omega_z dz$:

$$d\omega=(\partial_x\omega_y-\partial_y\omega_x)dx \wedge dy+(\partial_y\omega_z-\partial_z\omega_y)dy \wedge dz + (\partial_z\omega_x-\partial_x\omega_z)dz \wedge dx. \tag{78}$$

The divergence enters the expression for the exterior derivative of a two-form $\omega=\omega_{xy}dx \wedge dy + \omega_{yz}dy \wedge dz+\omega_{zx}dz \wedge dx$:

$$d\omega=(\partial_x\omega_{yz}+\partial_y\omega_{zx}+\partial_z\omega_{xy})dx \wedge dy \wedge dz. \tag{79}$$

The statement $dd=0$ contains the usual identities div(curl v) = curl(grad f) = 0 for a vector v and a function f.

The existence of the Hodge star map $*$ makes it possible to define a map from p-forms to $(p-1)$-forms by applying first $*$ [producing an $(n-p)$-form], then d [giving an $(n-p+1)$-form], and finally $*$ again. This map is called the adjoint exterior derivative and denoted δ. For Riemannian metrics it is defined as

$$\delta=(-1)^{np+n+1}*d*, \tag{80}$$

while for Lorentzian metrics there is an additional factor of -1. The adjoint exterior derivative satisfies $\delta\delta\omega=0$. A p-form field with vanishing adjoint exterior derivative is said to be coclosed, while one that is obtained as the adjoint exterior derivative of a $(p+1)$-form is called coexact.

It is possible to express the Laplacian Δ on a manifold M^n in terms of the maps d and δ:

$$\Delta=(d+\delta)^2=d\delta+\delta d. \tag{81}$$

For example, acting on a function f in three dimensions, this definition reproduces the standard expression of vector calculus,

$$\Delta f=(1/\sqrt{|g|})\partial_j(\sqrt{|g|}g^{jk}\partial_k f). \tag{82}$$

A p-form ω is said to be harmonic if $\Delta\omega=0$. This generalizes the usual notion of harmonic functions.

2.5 Fiber Bundles

In addition to involving a manifold of variables, many physical situations also exhibit symmetry of some kind. The natural geometrical framework in which to formulate such problems is the language of fiber bundles. Here, attention is restricted to a special type of bundle, appearing widely in physics, that involves continuous symmetries. The latter are described mathematically via the theory of Lie groups.

This paragraph presents a few essential definitions involving Lie groups. More details may be found in the articles GROUP THEORY and ALGEBRAIC METHODS. For the present purposes, a Lie group G may be viewed as a group that is also an r-dimensional manifold such that for two group elements $g,h \in G$ the map $gh^{-1}: G \times G \rightarrow G$ exists and is continuous. Denote coordinates in a chart near some point P of G by $\{a^A\}$, $A=1,\ldots,r$. Then the group composition function $f{:}G\times G\to G$ defined for $g(a),h(b),k(c)\in G$ by $f(h,g)=k=hg$ can be written in terms of r functions ϕ^A acting on the coordinates as

$$c^A=\phi^A(b,a). \tag{83}$$

The generators D_A of infinitesimal group transformations on G span the tangent space T_0G at the group identity and are given by

$$D_A=U_A^B\partial_B,\quad U_A^B=\frac{\partial\phi^B}{\partial b^A}\bigg|_{b=0}. \tag{84}$$

This space is called the Lie algebra associated with the group. The dual basis is spanned by the one-forms

$$\Omega^A=da^B(U^{-1})_B^A. \tag{85}$$

As a simple example, consider the group $U(1)$. The group manifold is a circle S^1; if the coordinate is denoted by θ, the group composition function is $\theta_3=\theta_2+\theta_1$. The generator D_θ is just ∂_θ and the dual basis is $d\theta$.

A fiber bundle is basically a manifold acted on by a symmetry. One important type of bundle, called a principal bundle, looks locally (but not necessarily globally) like a product of a continuous symmetry group with a manifold. The action of the symmetry provides a natural means of moving around in each local piece of bundle. The idea is to patch together these local pieces in a smooth way to get the whole principal bundle. Globally, the patching can introduce various twists into the overall structure, in which case the bundle is called nontrivial. A trivial bundle is one where no twists arise: the global and local structure are similar.

Here is a more formal definition. Given a manifold B and a Lie group G, a principal fiber bundle $E(B,G)$ is a manifold such that

1. G acts differentiably and without fixed points on E;
2. B is the quotient space of E by equivalence under G, and there exists a differentiable map $\pi{:}E\to B$; and
3. for each chart U_j in an atlas for B, there exists a differential and invertible mapping $\phi_j{:}\ \pi^{-1}(U_j)\to U_j\times G$ given by $E\to(\pi(P),f(P))$ for any point $P\in E$, where $f{:}\pi^{-1}(U_j)\to G$ satisfies $f(gP)=gf(P)$ for any $g\in G$.

The group G is called the structure group and the manifold B is called the base manifold. The map π is called the projection. The inverse image of π is the fiber; in effect, each fiber is like a copy of G. A (global) cross section or section s of a bundle is defined as a smooth map s: $B\to E$ such that $\pi\circ s$ is the identity on B. Local sections, i.e., sections defined only on $\pi^{-1}(U_j)$, always exist. If the bundle admits a global section, it is called trivial.

2.6 Connection and Curvature

Since $\{\partial_j\}$ is a basis for the tangent space of the base manifold and $\{D_A\}$ is one for the tangent space of the group, a basis for the tangent space to a point in the bundle is the set $\{\partial_j,D_A\}$. It has dual basis $\{dx^j,\Omega^A\}$. However, linear combinations could also be taken. The existence of this freedom permits the definition of a natural one-form called the connection that contains essential information about the structure of the bundle. The connection is basically a separation of the tangent space of E into two pieces, one along the group.

Formally, a connection is a choice of a subspace T_PH of T_PE at each point P of E such that

1. $T_PE=T_PG\otimes T_PH$, where T_PG is the space of vectors tangent to the fiber at P;
2. T_PH is invariant under action by G; and
3. the components in T_PG and T_PH of a smooth vector field in T_PE are also smooth. The spaces T_PG and T_PH are called the vertical and horizontal subspaces, respectively.

Some of the implications of this definition are most easily seen in a coordinate basis on the bundle. Let a basis for T_PH be defined as the linear combination

$$D_j=\partial_j-h_j^AD_A, \tag{86}$$

and require that D_j commute with D_A (among other consequences, this implies that h_j^A transforms under a particular representation of the Lie algebra of G, called the adjoint representation). Then the coefficients h_j^A are called connection coefficients and the basis elements $\{D_j\}$ are called the horizontal lifts or the covariant derivatives of the basis elements $\{\partial_j\}$. The dual to the basis $\{D_j,D_A\}$ for T_PE is the set $\{dx^j,\omega^A\}$, where the ω^A are given by

$$\omega^A=\Omega^A+h_j^Adx^j. \tag{87}$$

They form the components of a composite one-form $\omega=\omega^AD_A$ called the connection form.

The connection form ω encodes many of the interesting properties of the bundle in a concise notation. Its exterior derivative is also an important quantity in physical applications. Introduce a two-form R called the curvature form of the bundle by the definition

$$R=d\omega+\omega\wedge\omega. \tag{88}$$

The curvature is said to be a horizontal form because its action on any vertical vector vanishes. Its nonzero components are given by the expressions

$$R=R^A D_A, \quad R^A_{jk}=R^A(D_j,D_k), \tag{89}$$

and it follows that

$$[D_j,D_k]=R^A_{jk}D_A. \tag{90}$$

Applying another exterior derivative gives an identity

$$dR\equiv R\wedge\omega-\omega\wedge R=0 \tag{91}$$

called the Bianchi identity, with components

$$\sum_{jkl} D_j R^A_{kl}=0, \tag{92}$$

where the sum is over cyclic permutations of the indices j, k, l.

2.7 Example: Electromagnetism

An illustration of the role of some of these ideas in physics is provided by the formulation of the theory of electromagnetism in differential-geometric language. First, here is a summary of a few of the key equations of electromagnetism. In this section, a boldfaced symbol denotes a vector viewed as a collection of components. The symbol ∇ is the usual vector gradient operator, while $\cdot$ indicates the vector dot product and $\times$ represents the vector cross product. The Maxwell equations in SI units include Gauss's law,

$$\nabla\cdot\mathbf{E}=\rho/\epsilon_0; \tag{93}$$

Faraday's law,

$$\nabla\times\mathbf{E}+\partial_t\mathbf{B}=0; \tag{94}$$

the equation expressing the absence of magnetic monopoles,

$$\nabla\cdot\mathbf{B}=0; \tag{95}$$

and the Ampère-Maxwell law,

$$\nabla\times\mathbf{B}=\mu_0\mathbf{J}+c^{-2}\partial_t\mathbf{E}, \tag{96}$$

where ϵ_0 is the absolute permittivity, μ_0 is the absolute permeability, and $c=1/\sqrt{\epsilon_0\mu_0}$ is the speed of light *in vacuo*. Although these equations can be solved directly in simple cases, it is often useful to introduce new variables, called potentials, in terms of which the four first-order Maxwell equations are replaced with two second-order equations. The scalar and vector potentials ϕ and $\mathbf{A}$ are defined by

$$\mathbf{E}=-\nabla\phi-\partial_t\mathbf{A}, \quad \mathbf{B}=\nabla\times\mathbf{A}. \tag{97}$$

With these definitions, the homogeneous equations (94) and (95) are automatically satisfied. The two inhomogeneous Maxwell equations become coupled second-order equations for the potentials:

$$\nabla^2\phi+\partial_t\nabla\cdot\mathbf{A}=-\rho/\epsilon_0 \tag{98}$$

and

$$\nabla^2\mathbf{A}-c^{-2}\partial_t^2\mathbf{A}-\nabla(\nabla\cdot\mathbf{A}+c^{-1}\partial_t\phi)=-\mu_0\mathbf{J}. \tag{99}$$

There exists a freedom in the definition of ϕ and $\mathbf{A}$. The electric field $\mathbf{E}$ and the magnetic induction $\mathbf{B}$ are unchanged by the replacements

$$\phi\rightarrow\phi'=\phi-\partial_t\Lambda \tag{100}$$

and

$$\mathbf{A}\rightarrow\mathbf{A}'=\mathbf{A}-\nabla\Lambda, \tag{101}$$

where Λ is a function of $\mathbf{x}$ and t. These replacements are called gauge transformations. Their existence provides sufficient freedom to decouple Eqs. (98) and (99). For more details on the subjects of this paragraph, see ELECTRODYNAMICS, CLASSICAL.

It is easiest to approach the differential-geometric formulation of electromagnetism in stages, each incorporating more aspects of the theory. Here, the Maxwell equations for $\mathbf{E}$ and $\mathbf{B}$ are first expressed using the language of differential forms. The structure of the theory as a fiber bundle is then described, thereby incorporating the potentials ϕ and $\mathbf{A}$ and the notion of gauge transformations. To obtain consistent physical dimensionalities within expressions, it is convenient to work with a coordinate $x^0=ct$ with dimensions of length rather than with the time coordinate t. In what follows, the spatial coordinates (x,y,z) are denoted (x^1,x^2,x^3).

Begin with the identification of the space and time dimensions as a four-dimensional smooth manifold M. The manifold is often taken to be R^4 but this is not essential. The tangent space to M at a point P is also four-dimensional, and a basis for this space is the set $\{\partial_\mu\}$, $\mu=0,1,2,3$, of derivatives with respect to the four coordinates (x^0,x^1,x^2,x^3). An arbitrary vector can be expanded with respect to this basis. One vector, denoted by j and called the four-vector current, has components j^μ formed from the charge and current densities ρ, $\mathbf{J}$:

$$j=j^\mu\partial_\mu=(\mu_0\rho/c)\partial_0+\mu_0\mathbf{J}\cdot\boldsymbol{\nabla}. \tag{102}$$

An important tensor field on M is the Minkowskian metric g, defined to have components $g_{\mu\nu}$ in a coordinate basis forming a matrix given by

$$g_{\mu\nu}=\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{103}$$

This incorporates the essential elements of special relativity.

The Maxwell equations can be expressed in terms of a two-form field F defined on M. This antisymmetric tensor of type (0,2) is called the field strength. The components $F_{\mu\nu}$ of F are related to the components of the electric field and magnetic induction, and are given by

$$F_{\mu\nu}=\begin{pmatrix} 0 & +E^1/c & +E^2/c & +E^3/c \\ -E^1/c & 0 & -B^3 & +B^2 \\ -E^2/c & +B^3 & 0 & -B^1 \\ -E^3/c & -B^2 & +B^1 & 0 \end{pmatrix}. \tag{104}$$

This assignment of **E** and **B** is not *a priori* mathematically unique but establishes agreement of the resulting theory with experiment. In terms of the two-form F, the inhomogeneous Maxwell equations can be rewritten as

$$dF=j, \tag{105}$$

and the homogeneous ones become

$$d{*}F=0. \tag{106}$$

The two-form $*F$ is called the dual field strength. In component form in a coordinate basis, these equations read

$$\partial_\mu F^{\mu\nu}=j^\nu \tag{107}$$

and

$$e_{\mu\nu\rho\sigma}\partial^\rho F^{\mu\nu}=0. \tag{108}$$

Each of these represents four equations, and an inspection shows they reduce to the usual form of the Maxwell equations upon substitution in F and j of **E**, **B**, ρ, and **J**.

The discussion so far has excluded the potentials ϕ and **A**. These can be combined to form the components A^μ of a vector, called the gauge potential:

$$A^\mu\partial_\mu=(\phi/c)\partial_0+\mathbf{A}\cdot\boldsymbol{\nabla}. \tag{109}$$

The factor of c is introduced to maintain dimensional consistency. The metric g provides the associated one-form

$$A=A_\mu dx^\mu=g_{\mu\nu}A^\nu dx^\mu, \tag{110}$$

with components obtained by lowering the index. A complete description of the differential-geometric role of the gauge potential in electromagnetism requires a framework in which to place its nonuniqueness under gauge transformations. This freedom can be interpreted as a symmetry of Eqs. (98) and (99) expressing electromagnetism in terms of the potentials. It can be shown that this symmetry is a Lie group, called $U(1)$. A natural geometrical framework to express this is a fiber bundle, as is discussed next. For simplicity in what follows, the charge and current densities are taken to vanish. Nonzero distributions can be incorporated consistently with the addition of some extra structure.

The bundle of interest is a principal fiber bundle with the four-dimensional space-time manifold taken as the base manifold B and the symmetry group $U(1)$ of gauge transformations taken for the structure group G. Since the manifold of the group $U(1)$ is a circle S^1, the principal bundle is five-dimensional. Denote the coordinate on S^1 by θ. The introduction of a connection separates the tangent space to a point P in the bundle into a four-dimensional horizontal subspace spanned by the basis $\{D_\mu=\partial_\mu\}$ and a one-dimensional vertical subspace spanned by the generator $D_\theta=\partial_\theta$ of the Lie algebra of $U(1)$. The dual basis is the set $\{dx^\mu,\Omega^\theta=d\theta\}$. The composite connection form ω is $\omega=\Omega^\theta D_\theta=d\theta\partial_\theta$.

The gauge potential A can be identified with the value of the one-form Ω^θ on a section s of the bundle. Suppose that the surface s through the bundle E is specified in a chart U by choosing the group coordinate θ as a function of the coordinates $\{x^\mu\}$ provided by U. Then the dual form becomes

$$\Omega^\theta \equiv d\theta = \partial_\mu \theta(x) dx^\mu \equiv A_\mu(x) dx^\mu, \tag{111}$$

where the identification of the components of the one-form Ω^θ with the components of the gauge-potential one-form has been made. Under a change of cross section, which is equivalent to the action of a group element with a parameter Λ, say, the potentials A_μ change by an amount $\partial_\mu \Lambda$. This provides the geometrical interpretation for the gauge transformations (100) and (101).

The curvature two-form $d\omega + \omega \wedge \omega$ derived from the connection form ω is denoted by F. Evaluated on a local section, it has components

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{112}$$

In terms of the scalar and vector potentials, this equation reproduces the definitions of Eq. (97). The Bianchi identity in component form in this case can be written

$$d{*}F = 0, \tag{113}$$

thereby reproducing the homogeneous Maxwell equations. To complete the specification of the bundle, additional equations are needed that explicitly determine in each section the connection and the curvature. These are called equations of motion. Requiring these to transform as usual under Lorentz transformations and to be second-order differential equations for the connection or first-order equations for the curvature significantly restricts the options. An inspection of the Lorentz representation content of the general first-order term $\partial_\lambda F_{\mu\nu}$ shows that the simplest choice is $\partial_\mu F_{\mu\nu} = 0$ or its form equivalent

$$dF = 0. \tag{114}$$

This reveals the geometrical role of the remaining two equations in Maxwell's theory.

In the presence of monopoles, the homogeneous Maxwell equations are modified by the introduction of sources. A geometrical setting for the equations describing the fields of a monopole is provided by a nontrivial principal bundle. It can be shown that the essential physics is contained in a bundle with base space S^2 and structure group $U(1)$. The bundle space E looks like S^3 and the projection map π is called the Hopf map.

2.8 Complex Manifolds

Just as the requirement of differentiability for manifolds introduces many useful structures, a further restriction imposing complex analyticity is of considerable interest. The resulting manifolds, called complex manifolds, look locally like the complex plane. Some of their main features are outlined in this section. Basic methods of complex analysis are assumed here. See ANALYTICAL METHODS, Sec. 1, for more details.

The formal definition of a complex manifold M parallels that for a real differentiable manifold presented in Sec. 2.1. The key difference is that the local charts now contain maps f taking neighborhoods U into C^n, the product of n complex planes C, and that the composition map $f_1 \circ f_2^{-1}$ is required to be holomorphic rather than differentiable. This ensures that the methods of complex analysis can be used on M independently of any choice of chart. The number n is called the complex dimension of M; the real dimension is $2n$. An important feature is that a complex manifold may have two or more incompatible atlases, i.e., the union of two atlases may not satisfy the requirements for an atlas. In this case the atlases are said to define different complex structures. An example is the torus T^2 with two real dimensions; it can be shown that the complex structures on the torus are distinguished by a complex number called the modular parameter.

Denote the n complex coordinates on M in a chart U by $z^j = x^j + iy^j$, $j = 1,\ldots,n$, with $\bar{z}^j = x^j - iy^j$. The tangent space T_PM at a point P of M^n is spanned by a $2n$-dimensional coordinate basis $\{\partial/\partial x^j, \partial/\partial y^j\}$. It is useful to define

$$\partial_j \equiv \frac{\partial}{\partial z^j} = \frac{1}{2}\left(\frac{\partial}{\partial x^j} - i\frac{\partial}{\partial y^j}\right),$$

$$\partial_{\bar{j}} \equiv \frac{\partial}{\partial \bar{z}^j} = \frac{1}{2}\left(\frac{\partial}{\partial x^j} + i\frac{\partial}{\partial y^j}\right). \tag{115}$$

The cotangent space is spanned by the dual basis $\{dx^j, dy^j\}$, or equivalently by

$$\{dz^j = dx^j + i dy^j,\quad d\bar{z}^j = dx^j - i dy^j\}. \tag{116}$$

Define the linear map $J: T_PM \to T_PM$ by

$$J\partial_j = i\partial_j,\quad J\partial_{\bar{j}} = -i\partial_{\bar{j}}. \tag{117}$$

Note that $J \circ J = -I$. This map is smooth and globally defined on any complex manifold M.

It is called the almost complex structure of M. The action of J separates T_PM into two separate vector spaces, one spanned by vectors v such that $Jv=iv$ and the other by vectors such that $Jv=-iv$. It follows that a vector in T_PM can be uniquely decomposed into two pieces, called the holomorphic and antiholomorphic parts. The cotangent space T_P^*M can be separated into two corresponding pieces.

Complex differential forms of degree (p,q) can also be introduced. These are elements of a vector space denoted by $\wedge^{(p,q)}T$. In local coordinates, $\wedge^{(p,q)}T$ is spanned by a coordinate basis containing p factors of dz^j and q factors of $d\bar{z}^j$. The exterior derivative d naturally separates into the sum of two pieces,

$$d=\partial+\bar{\partial}, \tag{118}$$

called the Dolbeault operators. They satisfy

$$\partial\partial=\bar{\partial}\bar{\partial}=\partial\bar{\partial}+\bar{\partial}\partial=0. \tag{119}$$

All complex manifolds admit a Hermitian metric. A Riemannian metric g on M is said to be Hermitian if

$$g(Jv_1,Jv_2)=g(v_1,v_2) \tag{120}$$

for all vectors $v_1,v_2 \in T_PM$ at all points P. In a coordinate basis, g can be shown to have the form

$$g=g_{j\bar{k}}dz^j\wedge d\bar{z}^k+g_{\bar{j}k}d\bar{z}^j\wedge dz^k. \tag{121}$$

One can also define a two-form Ω called the Kähler form by

$$\Omega(v_1,v_2)=g(Jv_1,v_2). \tag{122}$$

If the Kähler form is closed, $d\Omega=0$, the manifold is called a Kähler manifold and the metric g is said to be a Kähler metric. In a chart, the components of a Kähler metric can be written as

$$g_{j\bar{k}}=\partial_j\partial_{\bar{k}}K, \tag{123}$$

where K is a scalar function called the Kähler potential. Compact Kähler manifolds in one complex dimension are called Riemann surfaces and are of great importance in certain branches of physics, notably string theory. Examples of Riemann surfaces are the two-sphere S^2 and the two-torus T^2.

2.9 Global Considerations

Essentially all the differential geometry considered above has involved local concepts. It is also of interest to address the issue of the extent to which the local properties of a manifold determine its global ones. The study of global properties of a manifold forms part of the branch of mathematics called topology (*q.v.*) and as such is tangential to the scope of this article. This section provides a sketch of some connections between the two subjects. Details may be found in the references provided at the end of the article.

One link between the geometry and topology of a differentiable manifold M can be introduced by considering the space of all closed p-forms on M. This space can be separated into classes, each containing closed forms differing from one another only by exact ones. The set of all classes is a vector space called the pth de Rham cohomology group of M and denoted $H^p(M)$. This vector space contains topological information about M. For example, the dimension of H^p, called the pth Betti number, is a topological invariant of M that contains information about the holes in M. The Betti numbers also determine the number of harmonic forms on M.

There are relationships between the number of critical points of functions on a manifold M and the topology of M. This is the subject of the calculus of variations in the large, or Morse theory. Among the results obtained are the Morse inequalities, which relate the number of certain types of critical points of a function to combinations of the Betti numbers on M.

The presence of a metric on M permits other types of global information to be obtained. An example is the Hodge decomposition theorem. This can be viewed as the statement that on a compact orientable Riemannian manifold M without boundary, any p-form can be uniquely decomposed into the sum of an exact form, a coexact form, and a harmonic form.

The issue of describing the global structure of a bundle (not necessarily principal) provides another link between geometry and topology. It is possible to develop measures of the ways in which a given bundle differs from the trivial bundle. The relevant mathematical objects are called characteristic classes. They are elements of the cohomology classes of the base manifold, and are given different names depending on the type of bundle being considered. Among these are Pontrjagin, Euler, and Chern classes, corresponding to orthogonal, special orthogonal, and unitary structure groups. Elements in these classes can be expressed in terms of the curvature two-form of the bundle. Another set of characteristic classes, the Steifel-Whitney classes, determines the orientability of a manifold and whether a spinor field can be consistently defined on it.

There are also relations between certain aspects of differential operators on bundles and the topology of the bundles. These are given by index theorems. An important example is the Gauss-Bonnet theorem, which connects the number of harmonic forms on a manifold (this is a property of the exterior derivative operator) to an integral over the Euler class (this is a topological quantity). Another important example is the Riemann-Roch theorem for complex manifolds. These are special cases of the Atiyah-Singer index theorem.

2.10 Further Examples

Many sets of smooth physical variables can be viewed as differentiable manifolds, and so differential-geometric concepts such as vectors, tensors, forms, and bundles play key roles in much of theoretical physics. Examples can be found in every major branch of physics. For instance, the modern formulation of the Hamiltonian dynamics of a system proceeds via the investigation of a manifold M called the phase space, with local coordinates corresponding to the generalized coordinates and momenta of the system. A closed nondegenerate two-form called the symplectic form is defined on M, making the phase space a symplectic manifold. The study of the properties of the phase space using the methods of differential geometry provides information about the behavior of the system. An extension of this example occurs in quantum mechanics (see QUANTUM MECHANICS). Quantization of a system involves the introduction of complex structure on the symplectic manifold. The study of this procedure is called geometric quantization.

Differential geometry is particularly crucial in the development of theories of fundamental interactions and particles. The geometrical constructions presented above for electromagnetism can readily be extended to other theories of fundamental forces. For example, the equations believed to describe the underlying physics of the strong interactions (see the article STRONG INTERACTIONS) form a theory called chromodynamics. This theory can be expressed geometrically using a principal bundle over space-time but where the structure group is the eight-dimensional Lie group called SU(3) rather than U(1). The presence of a multidimensional group manifold with a nontrivial group composition law means that, unlike the electrodynamic case, the horizontal lifts are inequivalent to the basis for the tangent space to the base manifold. As a result, the structure of the Bianchi identities and the equations of motion are somewhat more complicated in detail. The essential construction, however, remains the same.

Another important physical theory is general relativity, which provides a good description of the gravitational interactions at the classical level (see the article GRAVITATION AND GENERAL RELATIVITY). This theory can also be given a geometrical interpretation as a fiber bundle, but it is of a somewhat different kind, called a bundle of frames. Each point on a fiber of this bundle consists of a choice of basis vectors for the tangent space to the space-time manifold, and the symmetry group that plays the role of the structure group of a principal bundle now acts to rotate these bases into one another. A connection form and an associated curvature still exist, and they are closely related to the Christoffel symbols and the Riemann space-time curvature tensor of general relativity. In addition, there exists new freedom arising from the choice of basis vector on the base manifold, which leads to the existence of a second natural one-form on the bundle called the solder form or vierbein. This also has an associated two-form, called the torsion. In Einstein's general relativity the torsion form is specified to be zero, although other possibilities can be envisaged.

Attempts to unify the known fundamental forces and particles (see UNIFIED FIELD THEORIES) make wide use of geometrical constructions. Examples of such theories in four dimensions are the grand unified theories, describing the strong, weak, and electromagnetic forces in a single framework. The geometrical structures discussed above can be extended to more complicated symmetry groups large enough so that the connection forms include all the force fields needed for these theories. Certain elementary particles (see PARTICLES, ELEMENTARY) play the role of sources for these fields and can also be incorporated in bundles called associated bundles. Many unified theories involve higher-dimensional manifolds, in which physical space-time is a submanifold. These include the so-called Kaluza-Klein theories. Often, the symmetries of the extra dimensions permit them to play the role of the structure group in a principal bundle.

Generalizations of the geometrical framework of gravitation are also possible. For example, if the base manifold for a bundle of frames is generalized in a certain way, it is possible to specify bundles describing extensions of general relativity that include fundamental particles and forces other than gravity and that incorporate enlarged symmetries called supersymmetries. The resulting theories are called supergravities.

String theories are candidate unified theories including gravity that are believed to be consistent with quantum mechanics. In these theories, the fundamental forces and particles are interpreted as objects that are extended in one dimension (hence the name string). As a string propagates in space-time, it sweeps out a two-dimensional surface called the world sheet. A description of the world sheet involves the study of complex manifolds, in particular Riemann surfaces, as well as the notions of global differential geometry.

3. PROJECTIVE GEOMETRY

In its basic form, projective geometry is essentially the theory of perspective, i.e., the study of those features of geometrical objects that remain the same when the objects are projected from a point onto a line or plane. The elements of projective geometry are implicitly used by artistic painters, designers, and other people who represent three-dimensional objects on a two-dimensional medium. In its generalized form, the subject is fundamental in axiomatic geometry. It can be viewed as subsuming the classical Euclidean and non-Euclidean geometries.

There are two approaches to projective geometry. Synthetic projective geometry seeks to develop the subject as a series of deductions starting from certain axioms, in the Euclidean tradition. Analytical projective geometry introduces homogeneous coordinates and uses analytical techniques to obtain results. The two approaches are complementary, although projective geometries exist for which coordinates cannot be introduced.

A key feature of projective geometry is that parallel lines are assumed to meet in a single point, called the point at infinity, and that parallel planes meet in a single line, called the line at infinity. One advantage of these assumptions is that geometrical statements do not require exceptions for parallelism. For example, it is now true that any two lines in the plane determine a point, and any two planes in three dimensions determine a line.

In a plane, the statement that two lines determine a point is strikingly similar to the statement that two points determine a line. In general, projective-geometric statements involving points and lines in the plane remain valid when the roles of the points and lines are interchanged. In the plane, points are said to be dual to lines. In three dimensions the notion of duality applies between points and planes, or between lines and lines. A similar concept exists in higher dimensions.

With these ideas, a set of axioms for synthetic projective geometry can be formulated in terms of three basic notions: point, line, and incidence. The latter is meant in the sense of intersection: for example, a point is incident to a line if it lies on the line. The axioms can be expressed in dual pairs, so that propositions deduced necessarily have valid duals.

3.1 Some Theorems

There are several theorems that play a central role both in the development of the basic theory and in its extension to more abstract situations. A key result is Desargues's theorem: Given six distinct points in two sets, $\{A_1,A_2,A_3\}$ and $\{B_1,B_2,B_3\}$ (i.e., the vertices of two triangles), if the lines A_1B_1, A_2B_2, A_3B_3 meet at a point, then the three points C_1, C_2, C_3 given respectively by the pairwise line intersections A_1B_2 and A_2B_1, A_2B_3 and A_3B_2, A_3B_1 and A_1B_3 are collinear. This theorem holds in all projective geometries in three dimensions or more and in certain two-dimensional cases, including the usual plane projective geometry. However, in two dimensions non-Desarguesian geometries also exist.

Another important result that holds for a large class of projective geometries including the usual plane and solid ones is Pappus's theorem: Given two lines a and b lying in a plane and two sets of three distinct points $\{A_1, A_2, A_3\}$ incident to a and $\{B_1, B_2, B_3\}$ incident to b, then the three points C_1, C_2, C_3 given respectively by the pairwise line intersections A_1B_2 and A_2B_1, A_2B_3 and A_3B_2, A_3B_1 and A_1B_3 are collinear. Non-Pappian geometries also exist.

A pencil of lines about a point P is defined as the set of all lines lying in a plane and incident with P. A line s in the plane not incident with P is called a section of the pencil, and the pencil is said to project the section from P. Two pencils can be projectively related through a common section. Two distinct sections are said to be related by a projective transformation from the point P. The fundamental theorem of projective geometry states that a projective transformation is specified when three collinear points and their images are given. The theorem generalizes to projective transformations of higher-dimensional figures.

Conic sections (see Sec. 1.2) have a natural construction in projective geometry, and their theory can be developed entirely within this subject. Since all conics can be generated by projection of a circle from a point onto a plane, the projective approach gives them a unified treatment and consequently several results of analytical geometry can follow from a single projective theorem. Plane-projective definitions also play an important role. For example, the locus of intersections of corresponding lines in two projectively related pencils is a conic. A well-known result in this branch of the subject is Pascal's theorem: Given six points $\{A_1,A_2,A_3,A_4,A_5,A_6\}$ incident to a conic, then the three points B_1, B_2, B_3 given respectively by the pairwise line intersections A_1A_2 and A_4A_5, A_2A_3 and A_5A_6, A_3A_4 and A_6A_1 are collinear. The dual to Pascal's theorem is sometimes called Brianchon's theorem. These methods of projective geometry can also be extended to the study of quadrics and higher-dimensional hypersurfaces.

3.2 Homogeneous Coordinates

In analytical projective geometry, a set of coordinates called homogeneous coordinates is introduced. Consider first homogeneous coordinates on the line. A Cartesian coordinate system assigns a single real number x to each point P. In contrast, a homogeneous coordinate system assigns *two* real numbers (x_0,x_1) to each point, where $x=x_1/x_0$ and at least one of (x_0,x_1) is nonzero. Evidently, the homogeneous coordinates (x_0,x_1) and (cx_0,cx_1), where c is a constant, both represent P. The advantage of homogeneous coordinates is that the point $(0,1)$ at infinity is treated on the same footing as, say, the origin $(1,0)$. It also makes any polynomial equation $f(x)=0$ homogeneous in (x_0,x_1) without affecting the degree of the equation.

In the plane, the homogeneous coordinates of a point P specified in Cartesian coordinates by (x,y) are three real numbers (x_0,x_1,x_2), not all zero, for which $x=x_1/x_0$, $y=x_2/x_0$. A line in Cartesian coordinates is given by the linear equation $Ax+By+C=0$. In homogeneous coordinates this becomes the homogeneous linear equation

$$Ax_1+Bx_2+Cx_0=0. \tag{124}$$

The line at infinity has equation $x_0=0$ and is thereby treated on a similar footing to other lines; for example, the x and y coordinate axes have equations $x_2=0$ and $x_1=0$, respectively. All these ideas generalize to higher dimensions.

In addition to providing a framework in which analytical calculations can be developed, the homogeneous coordinate system offers a simple setting for duality. For example, given Eq. (124), the three numbers (A,B,C) can be viewed as homogeneous coordinates for a line in the plane. Then, coordinate statements about a point are expressed in terms of three numbers (x_0,x_1,x_2), while statements about a line are expressed in terms of a dual set of three numbers (A,B,C). A single equation thus represents a line or a point depending on which three numbers are considered variables.

Any set of three coordinates $(\hat{x}_0,\hat{x}_1,\hat{x}_2)$, obtained from the homogeneous coordinate system (x_0,x_1,x_2) in the plane by an invertible linear transformation

$$\hat{x}_j=A_{jk}x_k \tag{125}$$

(see ALGEBRAIC METHODS, Sec. 3.3), also leaves unchanged the degree of any polynomial function of the coordinates. The set $(\hat{x}_0,\hat{x}_1,\hat{x}_2)$ can be taken as alternative homogeneous coordinates.

3.3 Group of Projective Transformations

Instead of being taken as a change of coordinates for a fixed point P, the linear transformation (125) can be interpreted as a mapping from a point P at (x_0,x_1,x_2) to another point $\hat{P}$ at $(\hat{x}_0,\hat{x}_1,\hat{x}_2)$. This provides a mapping of the projective plane onto itself. Such mappings form a group G called the group of projective transformations for the plane. Similarly, groups of projective transformations can be introduced for higher-dimensional cases.

According to the so-called erlangen program, projective geometry can be viewed as the study of properties of figures that are invariant under the action of G. Various other geometries can be obtained by requiring invariance under a subgroup of G. They include the regular Euclidean geometry, as well as affine geometry and the non-Euclidean elliptic and hyperbolic geometries.

Extensions of projective geometry to fields other than the real numbers exist. For example, complex projective geometry is defined over the complex numbers. The field may be finite or even noncommutative (see ALGEBRAIC METHODS, Sec. 2). For example, a finite geometry in the plane called $PG(2,5)$ can be constructed using 31 points and 31 lines, with six points on each line and six lines through each point. Details of these generalized projective geometries may be found in the references at the end of this article.

4. ALGEBRAIC GEOMETRY

Algebraic geometry involves the study of mathematical objects called varieties, which are generalized curves, surfaces, and hypersurfaces. The subject has several levels of abstraction, in each of which the precise meaning of the word variety is different. For the purposes of this article a relatively simple level of sophistication suffices, in which a variety can roughly be viewed as the solution to a set of polynomial equations for variables in a space. Note, however, that the modern definition of variety is considerably more abstract. It uses a branch of mathematics called the theory of schemes, about which more can be found in the references at the end of this article.

This section presents a few simple notions of algebraic geometry in the framework of polynomial equations. The discussion refers to several concepts (e.g., field, polynomial ring, rational functions) that are defined and described in the article ALGEBRAIC METHODS.

4.1 Affine Varieties

Here is a more precise definition of one important type of variety. Consider an algebraically closed field F. An n-dimensional affine space A^n over F is defined as the set of points specified by the coordinates $(f_1,...,f_n)$ with $f_j \in F$. Denote by $F\,[f_1,...,f_n]$ the polynomial ring in n variables over F. An affine variety V is a subset of A^n given by the common zeros of a set S of polynomials in $F\,[f_1,...,f_n]$. If S contains only one polynomial, V is called an affine curve for $n=2$, an affine surface for $n=3$, and an affine hypersurface for $n>3$.

A subset of V satisfying the definition of a variety is called a subvariety. If V is the union of two subvarieties, it is called reducible; otherwise, it is irreducible. For example, an irreducible affine curve is one for which the defining polynomial is irreducible (i.e., cannot be factored). An irreducible component of V is defined as a maximal irreducible subvariety of V. One result in this subject is that any variety V can be written uniquely as the union of finitely many distinct irreducible components.

Starting with a variety V, a sequence of irreducible varieties can be constructed such that each member of the sequence is a subvariety of the preceding one. This sequence is of finite length, and the number of subvarieties in it is called the dimension of V.

The unions and finite intersections of subvarieties of V are also subvarieties. This means that the complements of the subvarieties of V can be used as the collection t of subsets for a topological space (see Sec. 2.1) Therefore, A^n and hence also V can be endowed with a topology, called the Zariski topology. This topology is not Hausdorff but, unlike the usual Hausdorff topology on C^n, it is defined for all affine varieties over F.

4.2 Projective Varieties

Several extensions of the notion of affine variety to more general varieties exist. One generalization uses an approach similar to that taken in the construction of differentiable manifolds: The meaning of variety is extended to include objects constructed by patching together affine varieties. This generalization then looks locally like an affine variety but globally is different. An important result in algebraic geometry is that certain subsets of projective spaces form varieties of this sort, called projective varieties.

An n-dimensional projective space P^n over F can be introduced as the set of points specified by the homogeneous coordinates $(f_0,f_1,...,f_n)$ with $f_j \in F$ not all zero, subject to the restriction that two such sets of homogeneous coordinates related via a single nonzero constant $c \in F$ as

$$(f_0,f_1,...,f_n)=(cf_0,cf_1,...,cf_n) \tag{126}$$

specify the same point (cf. Sec. 3.2). Denote by $H[f_1,...,f_n]$ the ring of homogeneous polynomials in n variables over F. A projective variety V is a subset of P^n given by the common zeros of a set S of polynomials in $H[f_1,...,f_n]$. If S contains only one polynomial, V is called a projective curve for $n=2$, a projective surface for $n=3$, and a projective hypersurface for $n>3$.

4.3 Classification

The ultimate aims of algebraic geometry are the classification and characterization of varieties. These are difficult and unsolved problems in the generic case. To attack the classification problem, a means of relating varieties to one another is needed. This is provided by the notion of a rational map.

A rational map $f:V \to A^n$ from an affine variety to n-dimensional affine space is basically a set of n rational functions f_j. The domain of f is by definition taken as the union of the domains of the n functions f_j. A rational map $f:V_1 \to V_2$ between two affine varieties $V_1 \subset A^{n_1}$ and $V_2 \subset A^{n_2}$ is defined to be a rational map $f:V_1 \to A^{n_1}$ such that the range of f lies in V_2. If the map f also has a rational inverse, it is called a birational equivalence.

The classification problem is approached by seeking a classification up to birational equivalence. Ideally, this means providing discrete and/or continuous numerical quantities that are invariant under birational equivalence and that characterize inequivalent varieties. Then, given a birationally equivalent set of varieties, a standard subset with desirable features (e.g., no singularities) can be sought and a classification attempted. Finally, one can seek some means of measuring the deviation from this standard subset of the remaining members of the birational-equivalence class.

An example is provided by the special case of the algebraic curves over F. For these varieties, a discrete quantity called the genus g can be introduced, which is a nonnegative real number that is invariant under birational equivalence. Curves with $g=1$ are sometimes called elliptic curves. For each nonzero g the birational-equivalence classes can be labeled by a one-dimensional continuous variable if $g=1$ and by a $(3g-3)$-dimensional set of continuous variables otherwise. The continuous variables are called moduli. They also form an irreducible variety, called moduli space, that can in turn be studied with the methods of algebraic geometry.

If the field F is the field C of complex numbers, the resulting algebraic curves are the Riemann surfaces. A curve with $g=0$ is topologically a sphere, while one with $g=1$ is topologically a torus. The set of Riemann surfaces plays an important role in string theories (see Sec. 2.10). For example, at a particular order in perturbation theory for a scattering process the string world sheet is topologically a Riemann surface with punctures for the incoming and outgoing strings. The methods of differential and algebraic geometry play a significant role in the evaluation of such contributions to the scattering amplitudes.

GLOSSARY

Considerations of space prevent an extensive glossary being provided for this article. Instead, the following is a list incorporating important concepts together with the number of the section in which the concept appears.

Abscissa: See Sec. 1.1.
Adjoint Exterior Derivative: See Sec. 2.4.
Affine Space: See Sec. 4.1.

Affine Variety: See Sec. 4.1.
Almost Complex Structure: See Sec. 2.8.
Antisymmetric Symbol: See Sec. 2.4.
Antisymmetric Tensor: See Sec. 2.3.
Applicate: See Sec. 1.5.
Atlas: See Sec. 2.1.
Base Manifold: See Sec. 2.5.
Betti Number: See Sec. 2.9.
Bianchi Identity: See Sec. 2.6.
Birational Equivalence: See Sec. 4.3.
Brianchon's Theorem: See Sec. 3.1.
Bundle of Frames: See Sec. 2.10.
Cartan Exterior Algebra: See Sec. 2.4.
Cartesian Coordinates: See Sec. 1.1.
Characteristic Class: See Sec. 2.9.
Chart: See Sec. 2.1.
Circle: See Sec. 1.2.
Closed Form: See Sec. 2.4.
Coclosed Form: See Sec. 2.4.
Coexact Form: See Sec. 2.4.
Complex Manifold: See Sec. 2.8.
Complex Structure: See Sec. 2.8.
Cone: See Sec. 1.5.
Conic Section: See Sec. 1.2.
Connection: See Sec. 2.6.
Contraction: See Sec. 2.2.
Contravariant Components: See Sec. 2.2.
Coordinate Basis: See Sec. 2.2.
Cosine: See Sec. 1.3.
Cotangent Space: See Sec. 2.2.
Covariant Components: See Sec. 2.2.
Covariant Derivative: See Sec. 2.6.
Cross Section: See Sec. 2.5.
Curvature Form: See Sec. 2.6.
Curvilinear Coordinates: See Sec. 1.4.
Cylinder: See Sec. 1.5.
Cylindrical Coordinates: See Sec. 1.5.
De Rham Cohomology: See Sec. 2.9.
Desargues's Theorem: See Sec. 3.1.
Differential Forms: See Sec. 2.4.
Directrix: See Sec. 1.2.
Discriminant: See Sec. 1.2.
Dolbeault Operator: See Sec. 2.8.
Dual Basis: See Sec. 2.2.
Dual Vector Space: See Sec. 2.2.
Duality, Projective: See Sec. 3.
Eccentricity: See Sec. 1.2.
Einstein Summation Convention: See Sec. 2.2.
Ellipse: See Sec. 1.2.
Ellipsoid: See Sec. 1.5.
Embedding: See Sec. 2.3.
Erlangen Program: See Sec. 3.3.
Euclidean Space: See Sec. 1.5.
Exact Form: See Sec. 2.4.
Exterior Derivative: See Sec. 2.4.
Fiber: See Sec. 2.5.
Fiber Bundle: See Sec. 2.5.
Finite Geometry: See Sec. 3.3.
Focus: See Sec. 1.2.
Fundamental Theorem of Projective Geometry: See Sec. 3.1.
Genus: See Sec. 4.3.
Geometric Quantization: See Sec. 2.10.
Group of Projective Transformations: See Sec. 3.3.
Harmonic Form: See Sec. 2.4.
Hausdorff: See Sec. 2.1.
Hermitian Metric: See Sec. 2.8.
Hodge Decomposition Theorem: See Sec. 2.9.
Hodge Star: See Sec. 2.4.
Homogeneous Coordinates: See Sec. 3.2.
Hopf Map: See Sec. 2.7.
Horizontal Lift: See Sec. 2.6.
Horizontal Subspace: See Sec. 2.6.
Hyperbola: See Sec. 1.2.
Hyperboloid: See Sec. 1.5.
Hypersphere: See Sec. 1.5.
Hypersurface: See Sec. 1.5.
Incidence: See Sec. 3.
Index Theorem: See Sec. 2.9.
Irreducible Variety: See Sec. 4.1.
Kähler Metric: See Sec. 2.8.
Kepler Problem: See Sec. 1.6.
Lie Algebra: See Sec. 2.5.
Lie Group: See Sec. 2.5.
Line At Infinity: See Sec. 3.
Manifold: See Sec. 2.1.
Maxwell Equations: See Sec. 2.7.
Method of Coordinates: See Sec. 1.
Metric: See Sec. 2.3.
Modular Parameter: See Sec. 2.8.
Moduli: See Sec. 4.3.
Morse Theory: See Sec. 2.9.
Neighborhood: See Sec. 2.1.
Ordinate: See Sec. 1.1.
Pappus's Theorem: See Sec. 3.1.
Parabola: See Sec. 1.2.
Paraboloid: See Sec. 1.5.
Parametric Representation: See Sec. 1.5.
Pascal's Theorem: See Sec. 3.1.
Pencil: See Sec. 3.1.
Plane Analytic Geometry: See Sec. 1.5.
Plane Polar Coordinates: See Sec. 1.4.
Point at Infinity: See Sec. 3.
Potentials: See Sec. 2.7.
Principal Bundle: See Sec. 2.5.
Projective Geometry: See Sec. 3.
Projection Map: See Sec. 2.5.

Projective Transformation: See Sec. 3.1.
Projective Variety: See Sec. 4.2.
Quadric: See Sec. 1.5.
Riemann Surface: See Secs. 2.8, 2.10, 4.3.
Riemannian Manifold: See Sec. 2.3.
Scalar Field: See Sec. 2.3.
Sine: See Sec. 1.3.
Slope: See Sec. 1.1.
Solid Analytic Geometry: See Sec. 1.5.
Sphere: See Sec. 1.5.
Spherical Polar Coordinates: See Sec. 1.5.
Spherical Triangle: See Sec. 1.5.
Structure Group: See Sec. 2.5.
Submanifold: See Sec. 2.3.
Symmetric Tensor: See Sec. 2.3.
Symplectic Form: See Sec. 2.10.
Tangent: See Sec. 1.3.
Tangent Space: See Sec. 2.2.
Tensor: See Sec. 2.3.
Tensor Field: See Sec. 2.3.
Topological Space: See Sec. 2.1.
Variety: See Sec. 4.
Vector: See Sec. 2.2.
Vector Field: See Sec. 2.3.
Vertical Subspace: See Sec. 2.6.
Wedge Product: See Sec. 2.4.
Zariski Topology: See Sec. 4.1.

Further Reading

Abraham, R., Marsden, J. (1985), *Foundations of Mechanics*, Reading, MA: Addison-Wesley.

Borsuk, K. (1969), *Multidimensional Analytic Geometry*, Warsaw: PWN–Polish Scientific.

Coxeter, H. (1974), *Projective Geometry*, Toronto: Univ. of Toronto Press.

Eisenhart, L. (1909), *Differential Geometry*, New York: Atheneum.

Eisenhart, L. (1926), *Riemannian Geometry*, Princeton, NJ: Princeton Univ. Press.

Flanders, H. (1963), *Differential Forms*, New York: Academic.

Frampton, P. (1987), *Gauge Field Theories*, Reading, MA: Benjamin Cummings.

Fulton, W. (1974), *Algebraic Curves*, New York: Benjamin.

Goldstein, H. (1980), *Classical Mechanics*, Reading, MA: Addison-Wesley.

Goodman, A. W. (1963), *Analytic Geometry and the Calculus*, New York: Macmillan.

Green, M., Schwarz, J., Witten, E. (1987), *Superstring Theory*, 2 Vols., Cambridge, U.K.: Cambridge Univ. Press.

Hartshorne, R. (1977), *Algebraic Geometry*, Berlin: Springer-Verlag.

Kobayashi, S., Nomizu, K. (1965), *Foundations of Differential Geometry*, New York: Wiley.

Milnor, J., Stasheff, J. (1974), *Characteristic Classes*, Princeton, NJ: Princeton Univ. Press.

Nash, C., Sen, S. (1983), *Topology and Geometry for Physicists*, New York: Academic.

Ramond, P. (1981), *Field Theory*, Reading, MA: Benjamin Cummings.

Schutz, B. (1980), *Geometrical Methods of Mathematical Physics*, Cambridge, U.K.: Cambridge Univ. Press.

Semple, J., Roth, L. (1985), *Algebraic Geometry*, Oxford: Clarendon.

Spivak, M. (1979), *Differential Geometry*, Vols. 1–5, Wilmington: Publish or Perish.

Springer, G. (1981), *Riemann Surfaces*, New York: Chelsea.

Steenrod, N. (1974), *The Topology of Fibre Bundles*, Princeton, NJ: Princeton Univ. Press.

Struik, D. (1948), *A Concise History of Mathematics*, New York: Dover.

Veblen, O., Young, J. (1938), *Projective Geometry*, Vols. I and II, New York: Blaisdell.

Winger, R. (1962), *Projective Geometry*, New York: Dover.

Woodhouse, N. (1992), *Geometric Quantization*, Oxford: Oxford Univ. Press.

Zariski, O. (1971), *Algebraic Surfaces*, Berlin: Springer-Verlag.

GEOMETRICAL OPTICS

See OPTICS, GEOMETRICAL

GEOPHYSICS

R. R. B. von Frese, *Department of Geological Sciences, The Ohio State University, Columbus, Ohio, U.S.A.*
W. J. Hinze, *Department of Earth and Atmospheric Sciences, Purdue University, West Lafayette, Indiana, U.S.A.*

Introduction: The World of Geophysics 149
1. **Seismic Exploration** 150
1.1 Principles 151
1.2 Velocity 153
1.3 Instrumentation 153
1.4 Interpretation 154
1.4.1 Seismic Refraction Surveying 155
1.4.2 Seismic Reflection Surveying 156
2. **Gravity Exploration** 159
2.1 Principles 159
2.2 Density 160
2.3 Instrumentation 160
2.4 Surveying Procedures 161
2.5 Gravity-Data Reductions 161
2.6 Gravity Anomalies 162
2.7 Anomaly Isolation and Enhancement 163
2.8 Interpretation 164
3. **Magnetic Exploration** 166
3.1 Principles 166
3.2 Magnetic Properties 169
3.3 Instrumentation 169
3.4 Surveying Procedures 170
3.5 Magnetic-Data Reductions 171
3.6 Anomaly Isolation and Enhancement 172
3.7 Interpretation 173
4. **Electrical Exploration** 174
4.1 Self-Potential Method 175
4.2 Resistivity Methods 175
4.3 Induced-Polarization Method 179
4.4 Electromagnetic Methods 181
4.5 Telluric and Magnetotelluric Methods 183
5. **Applications of Geophysics to Lithospheric Investigations** 184
5.1 Mantle and Crustal Processes 184
5.2 Crustal Studies 185
5.3 Hydrocarbon Exploration 187
5.4 Mineral Exploration 189
5.5 Ground-Water Investigations 191
5.6 Environmental and Engineering Investigations 192
6. **Summary** 194
Glossary 194
Works Cited 195
Further Reading 195

INTRODUCTION: THE WORLD OF GEOPHYSICS

Geophysics is an interdisciplinary science that integrates the observations, hypotheses, and laws of geology with the techniques and principles of physics to understand the composition, nature, structure, and processes of the Earth. The world of geophysics extends from the magnetosphere—the outermost limits of the universe surrounding the Earth that are affected by the magnetic field of the Earth—to the center of the Earth some 6372 km from the Earth's surface. The term *solid-Earth geophysics* is applied to the branch of the science of geophysics that relates to the solid Earth including the liquid outer core of the Earth.

Geophysics is the "remote sensing" science that permits us to investigate the Earth's interior indirectly. Just as a physician uses our pulse, heart rate, temperature, x rays, and other physical evidence to investigate the composition, structure, and nature of the interior of our bodies, the solid-earth geophysicist uses a range of measurements on a variety of force fields that traverse the Earth to obtain information on the "internal condition" of the Earth. This approach is necessary because we have neither the resources nor technology to

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

sample adequately the Earth's third dimension in a more direct fashion.

The geophysical methods are based on measuring force fields, which in turn respond to the physical properties of Earth materials that they encounter. For example, seismic methods utilize elastic properties and the density of the Earth and thus the velocity of propagation of elastic waves, while electric methods are based on a variety of electrical properties of the Earth that affect its carrying of electrical current. In a similar manner, gravity and magnetic methods are dependent on the mass (density) and magnetization, respectively, of the Earth. Thus the measurements of force fields such as seismic (mechanical), gravitational, magnetic, and electric fields provide significant information about the properties of the Earth. This is especially true when the observations are taken over the surface of the Earth. Spatial variations in the measurements are ample proof that the Earth is not homogeneous, although to a first degree it is spherically symmetric in structure and composition.

A major effort of solid-Earth geophysicists is the mapping of spatial variations or the anomalies in the various force fields. Once an anomaly is adequately detailed, the geophysicist translates the effect into its source within the Earth by a procedure termed *inversion*. Unfortunately, this process does not provide unique answers, but by integration of the inversion of the results from several geophysical force fields and use of available surface geologic information, an appropriate geologic interpretation is made of the location, structure, and composition of the anomalous Earth materials.

The variety of force fields that are available to geophysicists for studying the Earth is advantageous in decreasing the ambiguity of geologic interpretation. In addition, the methods contrast significantly in depth of investigation, resolution of discrete anomalous Earth materials, efficiency, and ease of observation. The common characteristics of each of the methods is the strong base on physical principles, the need for geologic information in all phases of conducting the methods, and the intensity of computational aspects necessary to make force-field observations geologically significant.

In the following discussion, we will limit our presentation to the principal geophysical methods that are used to investigate the Earth's lithosphere. The lithosphere is the semirigid outermost shell of the Earth that varies in thickness over the surface of the Earth. It normally has a thickness of the order of 150 km, but in limited regions, such as the mid-oceanic ridge, it is almost absent, with the underlying plastic mantle (i.e., asthenosphere) nearly reaching the surface. The lithosphere is composed of two shells of variable thickness: a lower unit composed of dense, high–seismic-velocity mantle and the crustal surface shell. The crust, which has an average thickness of less than 10 km in the ocean basins and is generally 35 km thick in the continents, is highly complex, reflecting the erosion, sedimentation, metamorphism, igneous activity, and plastic movement of the mobile asthenosphere underlying the lithosphere that have occurred over the Earth's 4600-Ma history. These processes have led to the differentiation of chemical elements, vertical and horizontal movements, zones of crustal weakness, and the focusing of geological processes, such as volcanism, in limited areas of the Earth's surface. In this manner, the variations in the structure, composition, and nature of the lithosphere and specifically the crust that are mapped and investigated by solid-Earth geophysics control man's Earth resources and such prominent societal concerns as volcanic activity and earthquakes.

The following discussion considers the principles of the major methods (i.e., seismic, gravity, magnetic, and electrical) that are used to explore the Earth's interior. Applications of these methods range from global or regional scale surveys designed to image the crustal properties of an ocean or continent to essentially microscale surveys conducted down a borehole to establish the physical properties of the adjacent rocks. The discussion is concluded with a few examples of how geophysics affects our understanding of the Earth and related societal problems.

1. SEISMIC EXPLORATION

Seismology involves the temporal and spatial analysis of ground motion to deduce information concerning the structural and compositional properties of the subsurface. Measured phenomena include signal wavelengths, travel times, amplitudes, and phases. Seismic survey results present a relatively explicit and highly resolved picture of the subsurface that appears very analogous to a cross section of the subsurface. As a result the seismic method has found extensive and preferred use for many subsurface investigations. The tradeoff, however, is in the extreme cost of the method. Next to drilling, the seismic method is the most expensive exploration tool to use because it involves sophisticated field procedures and data-processing capabilities.

1.1 Principles

In seismic exploration, a displacement of the ground surface is introduced at the source point. Common mechanisms for producing the disturbance include explosions, hammer blows, ground vibrators, or strain release due to earthquake activity. The passage of the disturbance at a distant point of observation is measured by an instrument called a seismometer or geophone. The ground motion as observed by the geophone is recorded by a seismograph as a seismogram. Seismograms commonly exhibit a number of simple harmonic wave forms that describe the disturbance in time at the geophone position. Because the Earth is elastic to a first order, these wave forms can be interpreted in terms of vertical and lateral elastic property variations to help identify the compositional and structural characteristics of the media through which the disturbance has traveled.

Most of the materials of the Earth are elastic and therefore resist an applied stress by a force that is approximately proportional to the displacement. These materials follow Hooke's law that relates the elastic moduli to the ratio of stress to strain. Stress and strain in elastic materials are both tensors described by nine components, of which six are independent. Therefore, in the most general case, six equations, having a total of 36 unknown coefficients, are required to relate stress to strain. These coefficients may be simplified to the two Lamé constants λ and μ (or the shear modulus) for isotropic and homogeneous materials.

A material is said to be isotropic if its elastic properties are the same in all directions, and it is called homogeneous if strain is the same from point to point throughout the body. Applying Hooke's law to evaluate time-dependent stresses on a unit cube of isotropic, homogeneous material yields an equation of motion for the ground particles in each of the three orthogonal (j) directions, given by

$$\sigma\frac{\partial^2 u_j}{\partial t^2}=(\lambda+\mu)\frac{\partial\theta}{\partial x_j}+\mu\nabla^2 u_j, \tag{1}$$

where u_j is the displacement in the jth direction, σ is the density, θ is the cubical dilation (i.e., volume strain), x_j is the unit spatial coordinate in the jth direction, and ∇ is the gradient operator.

The equation of motion (1) predicts the existence of body waves called *P* and *S* waves that travel through the interior of elastic media. The *P* wave is also called the pressure, primary, compressional, longitudinal, or dilatational wave. In this wave, the direction for particle motion is back and forth along the direction in which the wave is traveling. The wave front advances by alternating compressions and rarefactions and travels through solid materials at a rate that is about 1.7 times the velocity V_S of the *S* wave.

The *S* wave is also called the shear, secondary, transverse, or rotational wave. In this wave, the direction of particle motion is transverse or perpendicular to the direction in which the wave is traveling. *S* waves are said to be vertically polarized in the sense that particles move in parallel lines in the vertical direction, or horizontally polarized in the sense that the motion moves in parallel lines in the horizontal plane.

Other types of waves that may be derived from the equation of motion (1) are the surface waves. The most important of these in seismic exploration are the Rayleigh waves, which exist only along a free surface such as the air–ground boundary. They represent the principal component of ground roll, which often severely distorts the occurrence of other wave forms on the seismogram. When generated by strong earthquakes or explosions, this ground roll can be very damaging to buildings, bridges, dams, and other structures. Rayleigh-wave particle motion is vertical, elliptical, and retrograde with respect to the direction of propagation. Amplitudes decrease exponentially with depth. Rayleigh-wave velocity is less than the velocities of any of the body waves, being approximately $0.9V_S$ in a homogeneous isotropic medium.

Rayleigh-wave velocity also varies with frequency, and hence it is said to be *dispersive*. Seismic velocity basically increases with depth in the Earth, so that the longer-period components tend to appear on the seismogram ahead of the shorter-period components that have propagated through shallower zones of lower velocity. Accordingly, the observed patterns of Rayleigh-wave dispersion are useful for analyzing the broad-scale velocity structure of the outer part of the Earth.

The propagation of seismic waves is governed by Huygens's principle, Fermat's principle, and Snell's laws. By Huygens's principle, each point on a wave front acts as the source of new wavelets, whose envelope constitutes the wave front at later time. For a homogeneous isotropic material, wave propagation can be described in terms of rays that are normal to the wave fronts because constructive interference of wavelets occurs only at the envelope. Seismic disturbances propagating through a first-order Earth of homogeneous isotropic strata follow straight ray paths in regions of constant velocity. However, in regions where velocities increase or decrease in the direction of propagation, curved ray paths are followed that bend respectively away from or toward the directions normal to the strata interfaces.

Basic to propagation theory is Fermat's principle, which states that the ray reaching a given point from a given source has reached the point by a minimum-time path between the two points. If the intervening medium contains parts having different velocities, the ray path will not be straight, but it will be minimum in time. In seismic prospecting, we are particularly interested in minimum-time paths.

Three things occur when an advancing wave front in one medium characterized by a velocity V_1 encounters a boundary with another medium with a different velocity V_2. First, waves are reflected according to Snell's law of reflection as given by

$$\text{angle of incidence } (i) = \text{angle of reflection.} \tag{2}$$

Second, waves enter the second medium as refracted waves at angles i_r according to Snell's law of refraction, which states that

$$\sin i / \sin i_r = V_1/V_2. \tag{3}$$

When $V_2 > V_1$, Eq. (3) predicts a "critical angle" i_c of incidence where reflected and refracted wave paths are equal and the wave travels along the interface between the two media so that

$$\sin i_c = V_1/V_2. \tag{4}$$

Finally, each incident wave type produces refracted and reflected waves of the opposite wave type across the boundary, having incidence and transmission angles that are governed by the velocity properties of the media in a fashion comparable to the relationships described in Eqs. (2) and (3).

In seismic prospecting the strength of signals returned to the detector or geophone is commonly quite weak, and hence we seek to minimize the conversion of incident energy into other wave types at media boundaries in the subsurface. The energy partitioning of the incident wave at the boundary between media is governed by a set of relations known as the Zoepperitz equations. These equations indicate that if we set up our experiment such that incident angles of 20° or less are maintained, the conversion of incident wave energy into other wave types is minimized. Under this circumstance, the fraction of incident energy reflected at an interface is given by

$$R = (\sigma_2 V_2 - \sigma_1 V_1)/(\sigma_2 V_2 + \sigma_1 V_1), \tag{5}$$

where $\sigma_1 V_1$ and $\sigma_2 V_2$ are the acoustic impedances of media 1 and 2, respectively. The sign of the reflection coefficient R for any boundary between media gives the polarity of the reflections from that boundary that are received at the geophone. Accordingly, reflections identified on a seismogram provide a straightforward indication of how acoustic impedances vary with depth.

Other factors affect the propagation of seismic waves through the subsurface including diffraction effects, spherical divergence or geometric spreading effects, and nonelastic wave-energy absorption effects. Factors such as these can severely distort seismic signals and hence greatly complicate the identification and interpretation of events recorded on the seismogram. Analytical techniques that migrate diffraction events back onto the diffracting point in the seismic section are commonly implemented to minimize the effects of diffractions, whereas amplitude corrections based on a simplistic exponential attenuation curve of amplitude with time or distance may be applied to minimize the effects of spherical divergence, absorption, and other transmission losses.

1.2 Velocity

Seismic data are interpreted in terms of the elastic-wave velocity variations of the subsurface formations. These velocities are dependent on a number of factors including depth of burial of the rocks, lithology, geologic age, fluid saturation, and the anisotropic properties of the subsurface rocks. In general, elastic velocities increase with depth, but this trend can be modified by the effects of temperature, pressure, compaction, cementation, and density of the subsurface materials. Velocities generally increase as these parameters increase, except for temperature, where velocity decreases with increasing temperature at the rate of about 2% per 100 °C.

Elastic-wave velocity variations due to lithology involve the effects of porosity, lithification, chemical composition, and texture of the rocks. Porosity is a major factor particularly in sedimentary rocks and unconsolidated materials. In general, velocity decreases with increasing porosity. Velocities increase for igneous and metamorphic rocks with decreasing silica content and metamorphic grade. In general, igneous and metamorphic rocks are higher-velocity materials than sedimentary rocks and unconsolidated materials. Unconsolidated materials generally exhibit lower velocities than consolidated materials. Velocity variations related to geologic age involve the effects of depth of burial and lithology variations within the subsurface rocks. In general, velocity values are greater for older rocks than younger rocks, although this generalization has many exceptions.

Regions of sedimentary rocks or unconsolidated materials that are water saturated tend to be characterized by increased *P*-wave velocities over those in the same formations that are unsaturated. However, because the shear strength of water is zero, *S*-wave velocities for fluid-saturated regions are markedly less than the propagation velocities of the *P* wave, which is the wave type used most commonly in exploration.

Many rocks tend to show stratification in the organization of their constituent materials, which leads to velocity anisotropy. As a result, velocities for sediments are generally greater parallel to bedding by 10%–15% than across the beds. In some igneous and metamorphic rocks, velocities parallel to banding may be 100% greater than velocities across the banding. Other sources of velocity anisotropy include crystal layering and fracturing at microscopic to macroscopic scales.

1.3 Instrumentation

The instrumentation of seismic surveying has become extremely portable with the recent advances in solid-state microcircuitry and processing. Seismic instrumentation basically involves geophones connected to a clock that provides timing to ground motion for estimating the time of travel of a disturbance between the source point and the geophone position. A processing unit is used to make a number of adjustments to the output from the geophones. These adjustments include gain control to compensate for variations in input geophone signal strength. Filters and other signal enhancement processing are also commonly applied. Ultimately the processed output from the geophones is electronically digitized for video display and archiving on magnetic tape or optical disk.

Seismic sources for exploration applications must transfer energy efficiently into the ground, producing a wave form with a sharp rise time to facilitate interpretation. Chemical explosives are traditionally used in the search for hydrocarbons and mineral resources. For engineering applications, the sledgehammer, weight drop, shotgun, studgun, and other mechanical impactors find widespread application. In regions where the use of explosives is prohibited, such as urban areas, mechanical vibrators (e.g., VIBROSEIS, Mini-Sosie) are commonly implemented. In this application, a sinusoidal vibration of continuously varying frequency is applied to the ground during a sweep period, typically lasting 7 s or longer. The field record, which consists of a superposition of many such long wave trains reflected from multiple interfaces, is correlated with the sweep frequency. The correlated record resembles the conventional seismic record that results from an impulsive source. For marine applications, sparkers, pingers, boomers, gas exploders, and air guns are commonly used to input mechanical energy to the water and ultimately the solid Earth.

Modern geophones are constructed as small as golf balls and operate along electromagnetic principles. An effective geophone design uses a wire coil rigidly attached to the geophone housing, which in turn is coupled to the ground. Surrounding the wire coil is an inertial mass that is magnetic and suspended by springs from the geophone housing. When the seismic event or ground disturbance passes the geophone position, the housing moves with the ground and an electromagnetic current is induced in the wire coil, thereby providing a temporal history of ground motion. Generally, the fundamental frequency of the geophone is tuned to the ground vibrations and thus senses the time derivative of the ground displacement.

1.4 Interpretation

Inversions of seismic data are predominantly implemented in the context of stratified Earth models. The Earth is basically a stratified system as a consequence of the action of its gravity field since it was formed. Indeed, seismic data provide the principal line of evidence for our view of the Earth as a stratified planet consisting of a layered core, overlain by a layered mantle, which in turn is overlain by a predominantly layered crust.

A seismogram records ground motion at the geophone position as a function of time. It may show several distinct seismic arrivals produced not only by the different wave types, but also by the different wave paths that exist through the stratified Earth. As an example, consider the simple horizontal two-layered system shown in Fig. 1. Seismograms collected at various source-to-receiver spacings or offsets would show a systematic pattern of arriving seismic events that may be identified to produce the travel-time curve given in Fig. 1. For example, out to the distance *A*, the earliest arrivals at the shorter distances are *P* waves following a direct path from the source to the geophone. For the direct wave paths, travel times increase linearly with impact distance so that the slope of this line is inversely proportional to the velocity V_1 of the first layer.

Later-arriving events will include the slower shear waves, surface waves, and reflected *P* waves. The arrival times for *P*-wave reflections increase hyperbolically with increasing offset. At offsets greater than or equal to the critical reflection point (see Fig. 1), *P* waves

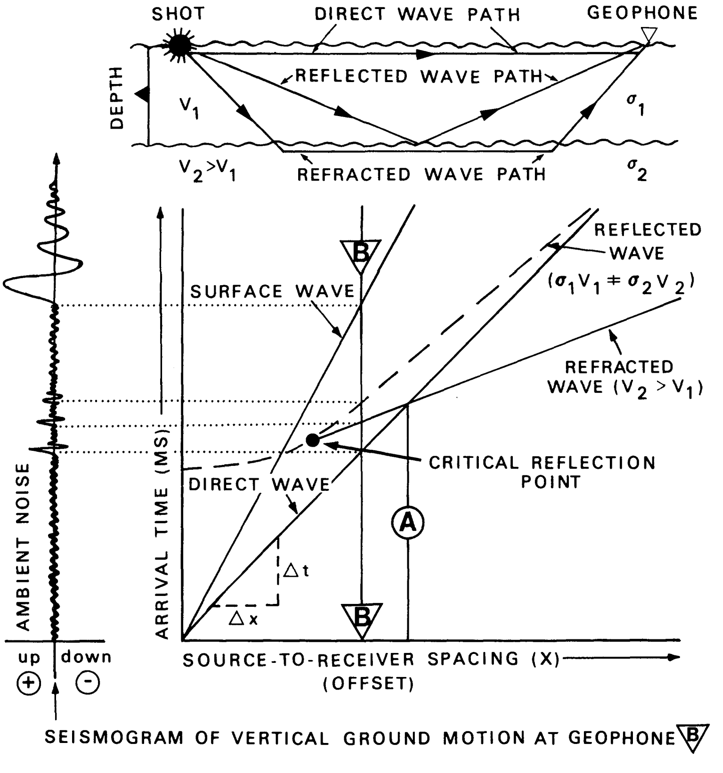
SHOT
DIRECT WAVE PATH
GEOPHONE
DEPTH
REFLECTED WAVE PATH
V_1
σ_1
$V_2>V_1$
REFRACTED WAVE PATH
σ_2
B
REFLECTED WAVE
$(\sigma_1 V_1 \neq \sigma_2 V_2)$
SURFACE WAVE
REFRACTED WAVE $(V_2>V_1)$
CRITICAL REFLECTION POINT
DIRECT WAVE
A
Δt
Δx
B
ARRIVAL TIME (MS)
AMBIENT NOISE
up
down
+
−
SOURCE-TO-RECEIVER SPACING (X)
(OFFSET)
SEISMOGRAM OF VERTICAL GROUND MOTION AT GEOPHONE B

FIG. 1. Idealized response of geophones to *P* waves and Rayleigh surface waves propagating through a horizontal two-layer subsurface. *S* waves (etc.) have not been included to simplify the travel-time graph. However, *S* waves propagate at a rate of about $0.6V_P$ so that they follow the *P*-wave arrival times by a delayed time which is proportional to offset.

refracted from the base of the V_1 layer will begin to appear on the seismogram. The critical reflection is defined at the offset where the reflected and refracted wave paths are equal. Refracted waves lag behind the direct waves out to the crossover distance A. At larger offsets, the refracted P waves arrive ahead of the direct waves. Arrival times for the refracted P waves vary linearly with distance, where the slope of the travel-time branch is inversely proportional to the P-wave velocity of the second layer, V_2. The projection of the refracted-wave branch to zero offset (i.e., $x=0$) is defined as the intercept time for this refraction. The extension of this analysis for obtaining travel-time graphs for systems consisting of three or more horizontal layers follows directly from the preceding discussion.

Analyses of travel-time graphs focus basically on the use of refracted or reflected waves for computing the depth to the underlying layers. Accordingly, two independent seismic surveying methods exist for locating and mapping subsurface interfaces. The method of seismic refraction surveying uses seismic energy that returns to the surface after traveling through the ground along refracted ray paths. This method is normally used to locate refracting interfaces (refractors) that separate layers of increasing seismic velocity with depth, but the method is also applicable where velocities vary laterally. In seismic reflection surveying, the travel times are measured for arrivals reflected to the surface from subsurface interfaces (reflectors) between zones of varying reflectivity which is a function of acoustic impedance differentials.

1.4.1 Seismic Refraction Surveying.

Analyses of refraction travel-time graphs involve ambiguities that must be considered in any interpretation. For example, a refractor branch of decreasing slope or gradient with respect to the previous branch can signal the existence of a zone of velocity increasing either laterally or vertically away from the shot point. Zones of decreasing velocity at depth will not be detected. A complexity in analyzing refraction travel-time graphs is that refractors in the most general case involve significant components of dip. For the more general case, the reciprocal of the gradient of the refractor branch is an apparent velocity dependent not only on the velocity of the layer, but the dip as well. Apparent velocity is higher than the refractor velocity when recording along a survey line up dip from the shot point, and lower when recording down dip. To deal with the possible presence of refractor dip, we reverse the refraction experiment by recording signals from energy sources positioned at each end of the survey line. The forward and reverse travel-time curves for refracted rays will differ in their gradients and intercept times in the presence of components of refractor dip. In general, the travel-time graph provides all parameters necessary for estimating the attributes of subsurface refractors such as velocity, depth, thickness, and dip.

A further complexity in the interpretation of refraction travel-time graphs is that the refracting interfaces more commonly involve irregular, nonplanar variations, which in turn, cause travel-time variations to occur along the refractor branches. However, analyzing the differences in travel time (i.e., the delay times) that occur at common offsets between the up-dip and down-dip refractor branches permits us to map the refracting interface irregularities.

In the reduction of seismic refraction data, two corrections are commonly applied. A topographic or elevation correction may be needed to remove travel-time variations due to differences in elevation of individual shots and geophones. Thickness variations in the low-velocity zone of weathering at the Earth's surface will also distort the arrival times on a refraction travel-time branch. Hence, elevation and weathering corrections may be applied to reduce travel times to a common elevation datum plane. In shallow marine refraction surveying, the water layer is commonly treated as a weathered zone and the correction is applied to replace the water layer by material of velocity equal to the velocity of the seabed.

The refraction method is applied at regional to global scales using earthquake or high-energy explosion-generated seismic data to provide basic information on the distribution and nature of the Earth's layering from the core to the surface. The refraction method using smaller energy sources is also a principal tool for investigating more local attributes of the near-surface Earth, such as the distribution of refractor interfaces, thicknesses, dips, and velocities to distinguish layer composition, and for finding vertical discontinuities (e.g., faults, areas of steep dip, salt domes). In application, however, the refraction method may suffer from several limitations. For example, to obtain complete travel-time graphs, profile lengths of up to 3 to 5 times the depth of investigation are required. This causes rapidly increasing demands on the energy source for carrying out the refraction survey. The refraction method also suffers from the velocity-inversion problem, where it cannot detect low-velocity zones beneath higher-velocity layers. In addition, refraction surveys can suffer from the blind-zone problem, where a layer is encountered that is so thin that it produces a critical refraction at only a single geophone position. To pick a refractor branch reliably, the layer must be thick enough to produce refractions on several seismograms. The occurrence of blind zones and velocity inversions within the subsurface can significantly affect the accuracy of the interpretation of refraction data. Finally, the complexities of picking refraction travel times and constructing the travel-time graph essentially limit the interpretation to a relatively small number of layers (up to about 5 or 6 layers) in most applications.

1.4.2 Seismic Reflection Surveying.

Reflection surveying offers several significant advantages over the refraction method for subsurface exploration. For example, reflection surveying can provide much greater subsurface detail in terms of the number of layers that can be mapped, as well as greater detail on the relief of the beds. Velocity inversions and blind zones also are not a concern to reflection surveys. In addition, geophone array lengths or spread lengths are generally much smaller than those required for refraction surveying, so that the source energy requirements for conducting reflection surveys can be much less.

A major problem with the application of reflection surveying is to distinguish reflections from all the other seismic events on the seismogram. Reflected pulses are commonly weak and do not appear in isolation on a seismogram, but rather they are distorted by interfering effects of numerous other seismic events arriving along different wave paths and from different wave types. As a result, sophisticated and expensive equipment, field techniques, and processing technologies are required to facilitate the recognition of reflections in the survey data. A further complication in the interpretation of the reflection data is that independent determinations of the subsurface velocities are required for effective and accurate depth analysis of the reflectors.

The interpretation of reflections on the seismogram involves the highly idealized assumption of normal-incidence reflection through a stratified subsurface, which invokes a problematic representation of reality. For example, seismograms in practice are commonly collected at significant offsets with respect to the position of the source, so that the condition of true normal incidence is rarely obtained. In practice, we find that the arrival times for reflections on adjacent seismograms increase hyperbolically with increasing offset, and hence these reflections as they were obtained do not conform to the condition of normal incidence. Furthermore, interpretations based on this condition ignore refractions, which must occur across any layer interface separating different velocities. As a result, in addition to maximizing the signal-to-noise ratio of reflections in any application, a primary objective in reflection survey design, data reduction, and processing is to adjust the observed data to the normal-incidence condition for accurate interpretation.

Seismic reflection survey data are commonly presented as a suite of seismograms called a record section. Figure 2 presents a seismic section obtained from a split-spread, multichannel reflection survey. In this case, the shot point is roughly centered within an array of 48 channels recording the output from geophones. The portions of the seismograms, or seismic traces, corresponding to upward motion of the ground surface have been shaded to facilitate the recognition of reflections. Shortly after shot detonation the first arrival of seismic energy produced by direct and refracted waves reaches the innermost geophones, and this energy passes out symmetrically through the two arms of the split spread. The first arrivals are followed by a series of reflection events revealed by their hyperbolic moveout or increasing arrival times.

In conventional reflection surveying, each detector in a profile consists of an array of geophones where the geophone outputs are simply summed or integrated to produce a single channel of output for the seismic sec-

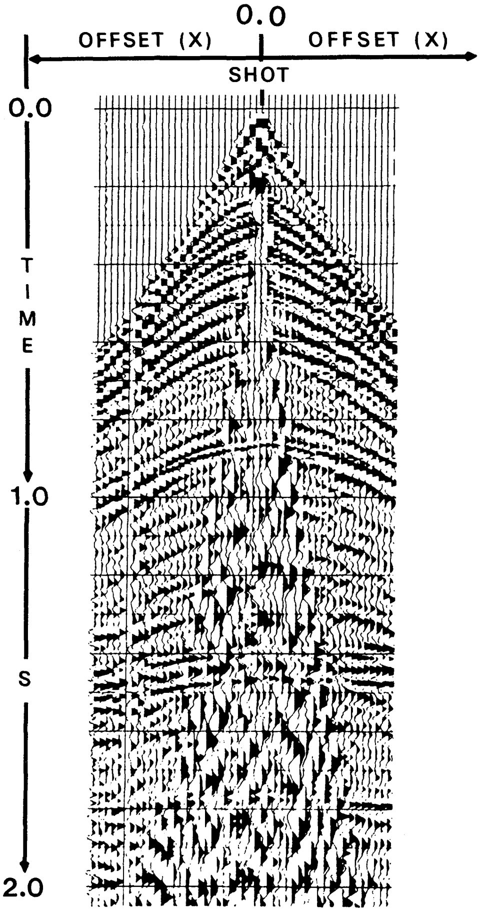

FIG. 2. A draped seismic section from a split spread array (from Prakla-Seismos GMBH by Keary and Brooks, 1984). Sets of reflected arrivals from individual interfaces are recognizable by their characteristic hyperbolic form. Reflected arrivals are masked within a triangular-shaped central zone by the arrivals of slow, large-amplitude and low-frequency surface waves (ground roll).

tion. The effective offset of a detector array is given by the distance from the shot to the center of the array. Geophone arrays provide a directional response to enhance the detection of near-vertically traveling reflected pulses, while breaking up or suppressing nonvertically traveling seismic events such as horizontally traveling Rayleigh surface waves.

In conventional reflection surveying, two reductions known as the static and dynamic corrections are commonly applied to facilitate geologic interpretation. The static correction is a combined weathering and elevation correction that removes the effects of the low-velocity surface layer and reduces all reflection times to a common elevation datum. This correction is static in the sense that it represents a constant time adjustment for the entire seismogram. Calculation of the static correction requires knowledge of the velocity and thickness of the weathered layer as well as the velocity of the material beneath the weathered layer. Shallow refraction surveys may be used to obtain this information. The velocity of the weathered layer may also be obtained directly by lowering a geophone into the layer through a borehole and measuring the arrival times for small shots fired from the surface or vice versa.

The dynamic correction is applied to adjust reflections from an interface for their hyperbolic increase in arrival times with offset. This hyperbolic effect is commonly called the *normal moveout,* defined as the difference between the arrival times of reflections from an interface observed at different offset distances. This correction is dynamic in the sense that the adjustment is not constant, but rather it must be uniquely determined for each reflection on the seismogram.

Reflections from horizontal interfaces that are below the first layer do not have purely hyperbolic moveouts because the travel paths involve refractions and other path complexities. The moveout for deeper reflections tends to decrease to zero with increasing reflector depth. The dynamic correction for these deeper reflection events may be effectively estimated by calculating the root mean square velocity of the layers overlying the reflector.

The signal-to-noise ratio of reflections may be greatly improved where the survey is performed according to the common-depth-point (CDP) method. This method is designed to enhance reflected signals and to attenuate nearly all other seismic arrivals. As illustrated in Fig. 3, this is achieved by the summation or stacking of successive seismograms taken at various offsets that are all centered about a common point on the reflector. For such symmetric wave paths, the reflected arrival-times are all nearly equal when corrected for moveout, so that the reflection will be enhanced when the CDP seismograms are stacked. Nearly all other seismic arrivals, by contrast, have wave paths that depend strongly on the offsets, so that their effects are minimized and less coherent in the stack.

The seismograms generated around a common depth point are known as a CDP gather.

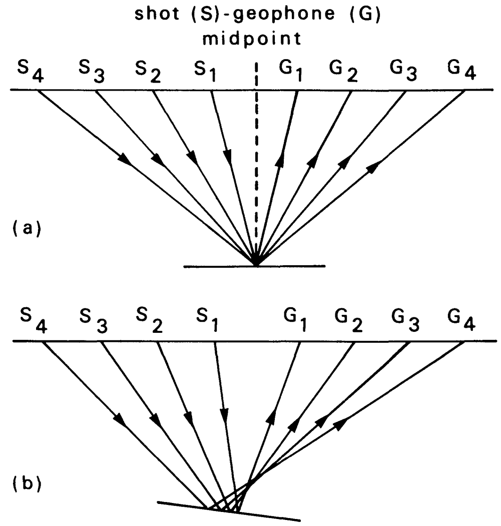

FIG. 3. Common depth point (CDP) reflection profiling (after Keary and Brooks, 1984) showing how (a) a set of rays from different shots to geophones are all reflected off a common point on a horizontal reflector, and (b) the common depth point is not achieved for a dipping reflector.

The four seismograms obtained from the wave paths illustrated in Fig. 3 provide a fourfold stack or alternatively represent 400% coverage. In theory, the CDP method is only applicable to horizontal reflectors. However, a statistical procedure has been developed that, when applied to a CDP gather, yields a stacking velocity for calculating the dynamic correction for almost any dip geometry of the reflector. The stacking velocities for a gather are computed over narrow time increments along the stack to produce a velocity spectrum. At each time increment, a range of velocity values is used to determine the dynamic corrections and the corrected traces are stacked. The stacking velocity is defined as the velocity that produces the maximum reflection amplitude in the stack.

In general, the CDP method can be applied to almost any geometry of field operation, and the performance of stacking improves with increasing multiplicity of coverage, allowing almost any desired improvement in the signal-to-noise ratio of reflections at the cost of recording more and more redundant data. Furthermore, better subsurface velocity information can be derived as a by-product of CDP stacking. Accordingly, this approach has become the standard method of multichannel reflection surveying.

Additional processing procedures are commonly implemented to facilitate the interpretation of reflection data. For example, rays that reverberate between the reflector interfaces may return to the surface to distort or masquerade as primary reflections. The correct identification of these multiple reflections is essential to avoid serious interpretation errors. In general, properties of multiple reflections such as their arrival times, amplitudes, and polarities are predictable from the attributes of the primary reflections so that procedures for identifying and removing multiple reflections are readily implemented. In addition, the differences in moveout between multiple and primary reflections increase with increasing depth so that CDP stacking will attenuate the occurrence of deeper multiple reflections in the reflection section.

A further adjustment of the reflection data is commonly warranted in the presence of diffracting sources in the subsurface and reflector undulations that are more convex than the incident wave front. For example, in the presence of a concave or synclinal undulation of the reflector surface, reflections are produced on the seismogram at vertical incidence only where the interface is horizontal. Where components of reflector dip exist, such as in the syncline, reflections will arrive on the seismogram along paths normal to reflector dip and not at normal incidence. Accordingly, such reflector undulations produce characteristic "bow-tie" signatures in the reflection section that cannot be interpreted in terms of vertically incident refections. Point diffractors such as the edges of thin faults also produce a cusp signature similar to the cusp illustrated in Fig. 4, but at maximum convexity, which distorts and may be confused with the signatures of primary reflections. Migration procedures are implemented to adjust the diffracted pulses and reflections from dipping interfaces into equivalent normal-incidence reflections. The migrated sections may be interpreted according to the idealized procedure for normal-incidence reflection using appropriate estimates of the layer velocities.

In addition to static and dynamic corrections, reflection seismograms also may undergo an amplitude correction to increase the continuity of reflection events across a section for enhanced interpretation, and to study the relative amplitude properties of the reflections. Hydrocarbon accumulations, for exam-

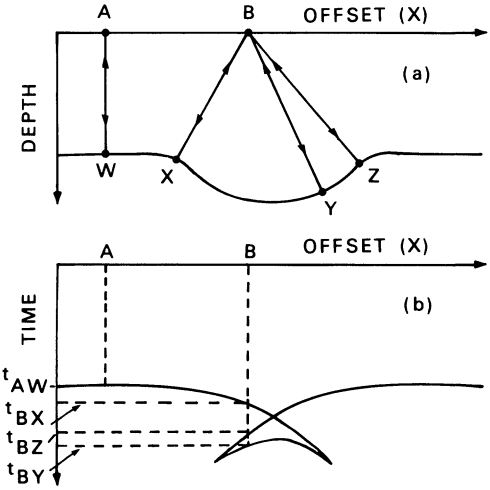

FIG. 4. (a) Reflections from a sharp synclinical feature in a reflecting interface, and (b) the resultant "bow-tie" shape of the reflection event on the non-migrated seismic section (after Keary and Brooks, 1984).

ple, may be directly revealed on amplitude-corrected sections by localized zones of anomalously strong reflections known as *bright spots*. These relatively prominent events are related to large-amplitude reflections from the top and bottom of gas zones such as gas-filled sands in the subsurface. In addition, fluid accumulations along a reflector interface may be directly signaled by *flat spots*, which are local horizontal or nearly horizontal reflection events splaying off the dipping reflector.

2. GRAVITY EXPLORATION

Gravity exploration focuses on measuring perturbations (anomalies) of the Earth's gravitational field caused by horizontal variations or contrasts in the mass and hence density properties of the subsurface. The process of obtaining models of mass distributions that fit the gravity anomalies can help to solve subsurface problems related to engineering and archaeological applications, hazard assessment, resource evaluation, studies of the Earth's tectonic features and processes, and other subsurface problems of considerable societal and scientific concern.

The multiple-stage process for applying the gravity method to the solution of subsurface problems involves data acquisition, data reduction, identification and isolation of significant anomalies, inversion of the selected anomalies for attributes of the subsurface, and finally the translation of these attributes into the context of a geological solution. Each stage of this process must be handled with appropriate care to ensure useful and accurate interpretations.

2.1 Principles

Gravity anomaly fields are potential fields, and thus they obey Laplace's equation. Their magnitudes are functions of the change in potential energy (work) so that they depend upon the relative position of the anomalous source and the point of measurement. In gravity exploration, we predominantly map the vertical component of the gravity field, because the gravity anomalies of geological mass variations are so small that perturbations of the Earth's gravity field in any direction other than the normal (i.e., vertical or radial) are negligible. For a unit point mass, the radial component of its gravity anomaly is given by

$$g(x_1,x_2,x_3) = -G \frac{\partial}{\partial R}\left(\frac{1}{R}\right)\frac{\partial R}{\partial x_1} m, \qquad (6)$$

where G is the universal gravitational constant, which has a measured value of 6.672×10^{-8} cm^3/g s^2; R is the distance from the observation point with generalized coordinates (x_1,x_2,x_3) to source point with generalized coordinates (x'_1,x'_2,x'_3); x_1 and x'_1 are the radial distances from the Earth's center to the observation point and source point, respectively; and m is the mass of the point source; with distances in cm and mass in g, the anomaly g is in milligals (1 mGal $= 10^{-5}$ m/s^2), which is the unit of measure most commonly used in gravity exploration.

Gravity anomaly fields are the summation of the effects of all anomalous sources regardless of quantity and distance. The gravity anomaly for an arbitrarily shaped body of uniform density σ, or mass per unit volume $m/dx'_1dx'_2dx'_3$, is obtained at the observation point by integrating Eq. (6) over the volume of the body according to

$$\int_{x'_{3a}}^{x'_{3b}} \int_{x'_{2a}}^{x'_{2b}} \int_{x'_{1a}}^{x'_{1b}} g(x_1,x_2,x_3)\,dx'_1dx'_2dx'_3\,, \qquad (7)$$

where the subscripts a and b refer, respectively, to the geometric lower and upper limits of the body in the orthogonal system of generalized coordinates (x'_1, x'_2, x'_3).

Conventional implementation of the above gravity equations expresses the effect of a subsurface body directly in terms of its density σ. However, in gravity exploration the density contrast, $\Delta\sigma$, is more commonly used in these equations. This is defined as the density of the subsurface source minus the density of the surrounding "country" material. Similarly, the observed gravity values are compared with the response of a homogeneous Earth; that is, the anomaly value is the observed value minus the theoretical or the normal value where there are no horizontal variations in density.

2.2 Density

Densities of crustal Earth materials vary roughly from 1.0 to 3.0 g/cm^3 (1 g/cm^3=10^3 kg/m^3). Individual mineral densities may exceed the upper limit of this range, because in addition to mineral composition, void space and void filling also control the density of Earth materials. The density of igneous rocks is primarily controlled by the relative mineral proportions. Felsic rocks contain a higher percentage of quartz and potassium- and sodium-rich feldspars than do mafic rocks such as gabbros, and thus they have a lower density. In contrast, the densities of clastic sediments and sedimentary rocks are controlled by the void space and void filling. Chemical sediments and sedimentary rocks vary considerably in density, but are primarily determined by mineral composition. In sediments and sedimentary rocks, voids will primarily be intergranular and will decrease with depth in the Earth due to compaction and, to a lesser degree, cementation. However, voids may occur in all rock types. For example, vuggy porosity can exist in chemical sedimentary rocks and in volcanic rocks, while fracture porosity is present wherever there are joints, faults, bedding planes, etc.

In geologic exploration with the gravity method, we are only concerned with horizontal variations in density or density contrasts. Vertical variations in density will not be recognized as a local crustal anomaly until they are horizontally truncated. Horizontal density contrasts in the subsurface may arise from inherent or imposed density variations. Inherent density variations may be caused, for example, by facies changes in sedimentary rocks and differentiation in plutonic bodies. Imposed effects can also lead to horizontal density contrasts. For example, deformed sedimentary strata of different densities will cause horizontal density variations, and fracturing or variable solutioning of rocks will also lead to horizontal mass disturbances.

The void space in rocks, which ranges from less than 2% in most igneous and metamorphic rocks to greater than 50% in some unconsolidated sediments at the Earth's surface, necessitates distinguishing between grain density σ_g and bulk density. Dry bulk density σ_d assumes moisture-free voids, and wet bulk density σ_w assumes complete saturation. The relationship between void space and density is

$$\sigma_w = \sigma_g - V_c\sigma_g/100 + F_c\sigma_f/100, \tag{8}$$

where V_c and F_c are the void content and fluid content, respectively, in percent, and σ_f is the fluid density.

2.3 Instrumentation

The utility of the gravity method in subsurface studies is dependent upon the efficient acquisition of precise data. Instrumentation for the measurement of gravitational acceleration has continued to improve, especially with the development of microprocessors and other electronic advances. Over the years numerous schemes have been developed and tested for measuring either the absolute or relative changes in gravity acceleration. However, today essentially all gravity measurements for geophysical exploration are relative measuring devices, gravimeters, which are based on the zero-length spring principle (LaCoste, 1934). Changes in the acceleration of gravity cause a variation in the gravitational force on a constant mass attached to a lever arm suspended from a zero-length spring. The relative position of the lever arm changes with gravity acceleration variations. Observations are made by determining the force that must be applied to the lever arm to bring it to a null position. This force is calibrated into gravity acceleration.

Standard gravimeters used in geologic exploration have a sensitivity of 0.01 mGal, and range without reset from a hundred mGal to several thousand mGal, depending on the specific instrument. However, gravimeters having a sensitivity of 0.001 mGal or 1 μGal are readily available for microgravity studies used in engineering and environmental investigations.

2.4 Surveying Procedures

Gravity surveying is usually conducted on a grid or along convenient transportation routes, with the station interval determined by the depth of geologic sources under investigation. The interval should be no greater than one-half the depth anticipated for target sources for the purpose of accurately mapping the gradients of the anomalies that are very important in interpretation. However, commonly the station interval needs to be decreased to map high-amplitude noise so that it can be identified and eliminated from the observed data. Elevations of the stations and their horizontal positions must be determined to an accuracy of approximately 4 cm (0.13 ft) and 12 m (40 ft), respectively, where we want gravimeter observations precise to 0.01 mGal, and 0.1 of those distances for μGal surveying. For this level of survey precision, accurate topographic surveying, which may have a significant impact on surveying costs and number and distribution of stations, is demanded.

2.5 Gravity-Data Reductions

Raw or field gravity data are affected by a wide variety of sources, which generally mask out the effect of the geologic sources that are the objective of the survey. As a result, field observations are subject to corrections to eliminate the extraneous effects. These corrections or reductions are derived theoretically, geologically, and empirically. Some reductions are applied universally to gravity data, while others are applied only for specific objectives and for certain geologic, topographic, and spatial size conditions.

Gravimeter observations are only relative. Therefore, to adjust measurements made at different times and with different meters relative to each other, all daily observations are referred to a single base station. Wherever possible, this base should be tied through observations to a national gravity base station. Information on base stations is available from appropriate federal agencies and many state geologic surveys.

Time variations in gravity or "drift" of the gravity measurements must be eliminated for essentially all exploration surveys. The two principal sources of drift are tidal effects and inherent fluctuations in gravimeter observations that reflect the ambient environment of the meter. Tidal effects are calculated by taking into account variations in the mass effect due to the changing relative positions of the Sun and the Moon with respect to positions on the Earth's surface. The maximum tidal variations of the Moon and Sun are respectively ±0.11 and ±0.05 mGal, and the maximum rate of change is 0.05 mGal/h. Inherent drift of gravimeters is caused by temperature and air-pressure changes, as well as fatigue of the sensing element and the mechanical shocks to which the instrument may be exposed. Under normal operating conditions the effect of inherent drift is less than the tidal variations, except for temperature-compensated meters that are subject to extreme temperature variations. Drift rates may exceed 0.1 mGal/h under these extreme conditions. Observed gravity values are generally corrected for temporal variations by monitoring the change in gravity by reoccupation of gravity stations assuming a linear change with time.

The gravity acceleration on the surface of the Earth increases as a function of latitude by approximately 5 Gal from the equator to the poles because of decrease in the Earth's radius and centrifugal force. The most recent internationally accepted equation for computing this latitude variation at sea level is the Geodetic Reference System of 1967 (GRS-1967) given by

$$g_\phi = 978.0318(1 + 0.005\,302\,4\sin^2\phi - 0.000\,005\,9\sin^2\phi), \qquad (9)$$

where g_ϕ is the predicted gravity in Gal at sea level at a north or south latitude of ϕ.

The elevation (or free-air) correction takes into account the decrease in gravity with distance from the center of the Earth. The gradient is given by

$$dg/dh = 0.308\,55 + 0.000\,219\cos 2\phi - 7.2\times10^{-8}h, \qquad (10)$$

where dg/dh is the vertical gradient of gravity in air in units of mGal/m at a station located an elevation h in meters above sea level and at latitude ϕ. The gradient changes only by 0.2% from the equator to the pole and by 0.3% from sea level to 10 km. Therefore the gradient is usually calculated assuming

$$dg/dh = 0.3086 \text{ mGal/m} = 0.094\,06 \text{ mGal/ft.} \tag{11}$$

The mass or Bouguer correction takes into account the gravitational attraction of the Earth material between the station elevation and the elevation datum. The correction is assumed to be equal to the gravity effect of a horizontal slab of infinite radius and thickness equal to the elevation difference, which can be shown to be

$$\begin{aligned} g_m &= 2\pi G\sigma h \\ &= 0.041\,85\sigma h \text{ mGal/m} \\ &= 0.012\,76\sigma h \text{ mGal/ft} \end{aligned} \tag{12}$$

for σ in g/cm^3 and h in meters or feet, respectively. The assumption of a horizontal slab made in the mass correction is commonly violated in practice where sufficient topographic relief exists between gravity stations. Therefore, a terrain correction is used to account for topographic deviation from a slab that has a thickness equal to the height of the gravity station. The effect of topography, whether above or below the station, is always to decrease the mass correction, because the attractive force of the topography above the observation site is away from the Earth, and the effect of topography below the station is also negative because it involves lack of Earth material. The precision with which terrain corrections can be made for surveys conducted in areas of rugged topography is often the controlling factor in the success of a gravity survey.

Additional specialized reductions are used for marine and airborne gravimetry. Gravity observations made in moving platforms such as ships or aircraft are corrupted by extraneous accelerations including horizontal and vertical accelerations produced by three-dimensional platform motions, the Eötvös effect of modified centrifugal acceleration that every moving gravimeter experiences, and other effects. Accelerations from local motions such as pitch, roll, and heave of the measurement platform are readily minimized by using an effectively damped gravimeter on a gyroscopically stabilized platform that can maintain the horizontal position to within a few seconds of arc. Removal of extraneous Eötvös effects from gravity observations requires relatively accurate knowledge of the vessel's navigation parameters such as its speed, heading, and latitude coordinates.

A number of other corrections are applied to gravity data for specialized purposes, particularly in regional surveys, which are becoming increasingly available as a result of technological advances in adapting shipborne, airborne, and spacecraft platforms for gravity mapping. They include

1. geologic correction, which takes into account the gravitational effect of geology known from direct geologic information or supplemental geophysical studies;
2. isostatic correction, which is the gravity effect of the Earth's topography and related isostatic compensating masses according to an isostatic theory;
3. curvature correction, which considers the gravity attraction of the actual curvature of the Earth's surface in contrast to the flat slab of the mass correction; and
4. indirect reduction, which accounts for the difference in elevation between the Earth's geoid, the reference surface for elevations, and the spheroid, the reference surface for gravity measurements.

2.6 Gravity Anomalies

When corrected for a wide variety of known variations as described above, the observed or raw gravity data yield a gravity anomaly that isolates variations due to mass (geologic) variations in the Earth. Several types of gravity anomalies have been devised on the basis of the corrections actually considered in calculating the predicted gravity. They all have their specific uses, but the most useful for exploration surveys is the complete Bouguer gravity anomaly, which takes into account the latitude, elevation, mass, and terrain corrections. By contrast, the simple Bouguer gravity anomaly omits the terrain effect.

Conventional Bouguer gravity-anomaly maps of regional areas (states, nations, etc.) are based on a density of 2.67 g/cm^3 (2670 kg/m^3) in the mass correction. This density has been assumed to be the average density of upper continental rocks. Use of the standard value facilitates compositing and comparison of gravity-anomaly maps. However, where the surface material is made up of rock units of a different density, errors will result that are proportional to the station elevation.

A gravity anomaly used commonly in marine, satellite, and other regional studies is the free-air gravity anomaly, which omits both the mass and terrain effects. The free-air anomaly is useful in mapping mass imbalances of the Earth that are of interest in long-term stability studies and in investigating stress variations in the Earth that may trigger or even cause local earthquakes (Hinze and Hildenbrand, 1988). The free-air anomaly is especially useful in regions of limited local terrain. Another anomaly that is also occasionally used for a similar purpose to that of the free-air anomaly is the isostatic anomaly. This anomaly eliminates the effect of topographic compensating masses by assuming that the Earth is in isostatic equilibrium according to a specified hypothesis. Thus mass imbalances from the assumed ideal condition are reflected in the anomalies.

2.7 Anomaly Isolation and Enhancement

Gravity anomaly fields combine the effects of all horizontal variations in density within the Earth, and thus the anomalies of interest in a particular geologic application are commonly distorted by overlapping anomalies. The gravity anomaly fields include the effects of deeper, broader sources called regional anomalies, local sources, which are considered noise, as well as the anomalies of interest which are commonly called residual anomalies. The definition of the residual anomaly is not very precise, but perhaps can best be defined as the anomaly of geologic interest in a particular study. Broader anomalies than the residual are often called regional anomalies and the smaller anomalies are generally referred to as noise. Thus the specification of an anomaly as a residual is problem dependent. The separation of the residual anomaly from the gravity field is a critical problem that controls the accuracy of the interpretational process. The subjectivity of the residual determination process makes it potentially one of the major limitations of applying the gravity method to many geological problems.

Over the years numerous graphical and numerical techniques, which are in reality filters, have been devised to focus on particular anomaly types, of specified characteristics, within a field of superimposed anomalies of varying properties. These techniques can be divided into two broad categories as *isolation* and *enhancement* methods. Isolation techniques are used to eliminate from the anomaly field all anomalies that do not have a certain set of qualitatively or quantitatively specified characteristics. Thus, the remaining anomaly, the residual anomaly, is the isolated gravity expression of the geologically significant source. Enhancement techniques, in contrast, accentuate a particular characteristic or group of characteristics of anomalies that are qualitatively or quantitatively definitive of anomalies that have special geologic significance. In the process of enhancing these characteristics, the anomalies are accentuated to increase their perceptibility. The "enhanced" potential fields are primarily used in qualitative visual inspection analysis and interpretation.

The most widely used method currently of isolating and enhancing anomaly data is digital filtering in the frequency (wavelength) domain of the data. In this procedure, the input data are Fourier transformed into the frequency domain (i.e., their spectrum) by the direct transform, multiplied by an appropriate filtering function, and the modified spectrum is transformed back into the space domain by the inverse Fourier transform. This is an efficient process on digital computers, and filters can be simply and accurately designed to carry out a wide variety of anomaly isolation and enhancement processes.

Fourier transforms depend only on the amplitudes and positions of the observations and require no information on the subsurface sources of the observations for their implementation. The transforms, however, may be directly processed or filtered for breaking up the data set into wavelength components (wavelength filtering) to remove, for example,

1. the higher-frequency noise due to small, near-surface sources with a low-pass filter,
2. the longer-wavelength component of the regional anomalies with a high-pass filter, or
3. both long- and short-wavelength components with a band-pass filter.

This approach also is used for separating anomaly components on the basis of their directional attributes (directional filtering), and for extracting the horizontal gradients. The vertical gradients of gravity anomalies are related to the horizontal gradients by Laplace's equation, so that the vertical gradients too may be extracted from the wavelength transform. Horizontal and vertical derivatives depict the gradient variations in the anomaly field and emphasize higher gradients, which are related to anomalies of near-surface sources and are associated with the positions of the boundaries of gravity anomaly sources. Consequently, derivative filter data will provide approximate shape outlines for modeling anomaly sources.

In addition, because the vertical gradients of gravity anomalies can be estimated, it is a straightforward matter to design upward and downward continuation filters that process the Fourier transform for accurate models of the observations over parallel surfaces respectively above and below the observation surface. Upward continuation minimizes the effect of higher-frequency components associated with more local, shallower source anomalies. In contrast, downward continuation emphasizes these components and increases the resolution of the anomaly field to individual sources. Downward continuation and anomaly gradient filters are particularly powerful tools in the study of near-surface sources. However, they are especially susceptible to distortion from noise, measurement errors, and other data inaccuracies, so that care must be taken in the interpretation of their results.

2.8 Interpretation

The level of interpretation achieved in gravity exploration is highly variable depending on the survey objectives, the quality of the data, and the available time and resources. Interpretation can range from simple identification and location of anomalous density contrasts in the subsurface to three-dimensional modeling, leading to complete characterization of anomaly sources. A key element in gravity interpretation is the availability and quality of ancillary data regarding the subsurface. Interpretation is ambiguous because the observed anomaly can be reproduced by an infinite number of source distributions shallower than the actual source of the anomaly and because of imprecision in the data. As a result, ancillary data, especially direct geologic data, provide important constraints to interpretation. These data can be used to develop conceptual subsurface models (hypotheses) to be tested by the data interpretation. Interpretation is commonly performed in an iterative manner utilizing previous interpretations as well as other geologic and geophysical data to develop improved models for testing until an interpretation is achieved that is compatible with all data sets.

The amplitude and character of the anomaly signatures of gravity sources in the subsurface are dependent on several key factors or variables. Anomalies vary with source geometry, including volume and shape, and with source depth. The amplitude and sign (or polarity) of the gravity anomaly are direct functions of the magnitude and sign, respectively, of the density contrast. Some examples of typical gravity anomaly profiles over idealized geological mass variations are presented in Fig. 5.

The general problem of estimating quantitative parameters of the source from the anomaly data is known as anomaly inversion. The inversion of residual anomalies is used to determine shape attributes, volume, depth, depth extent, densities, and other features of the geologic sources. For example, the maximum depth to sources of gravity anomalies can be easily derived from simple horizontal distance measurements on observed anomalies as indicated in Table 1. Solving the inverse problem requires postulating a mathematical model of the source with parameters that can be related to variations in the anomaly data.

Table 1. Maximum depth to the center (Z_c) or top (Z_t) of ideal gravity sources. Here $X_{1/2}$ is one-half the width of the anomaly at one-half the amplitude; $X^*_{1/2}$ is the horizontal distance from one-half the amplitude to either one-fourth or three-fourths of the amplitude; and DE is the depth extent of the source.

Geometric form	Depth
Sphere	$Z_c \leqslant 1.3X_{1/2}$
Horizontal cylinder	$Z_c \leqslant 1.0X_{1/2}$
Vertical cylinder	$Z_t \leqslant 0.58X_{1/2}$
Vertical fault	$Z_c \leqslant X^*_{1/2}$
Vertical thin dike	$Z_t \approx 0.7X_{1/2}$, $Z_t = \mathrm{DE}$ $Z_t \approx 1.0X_{1/2}$, $Z_t \gg \mathrm{DE}$

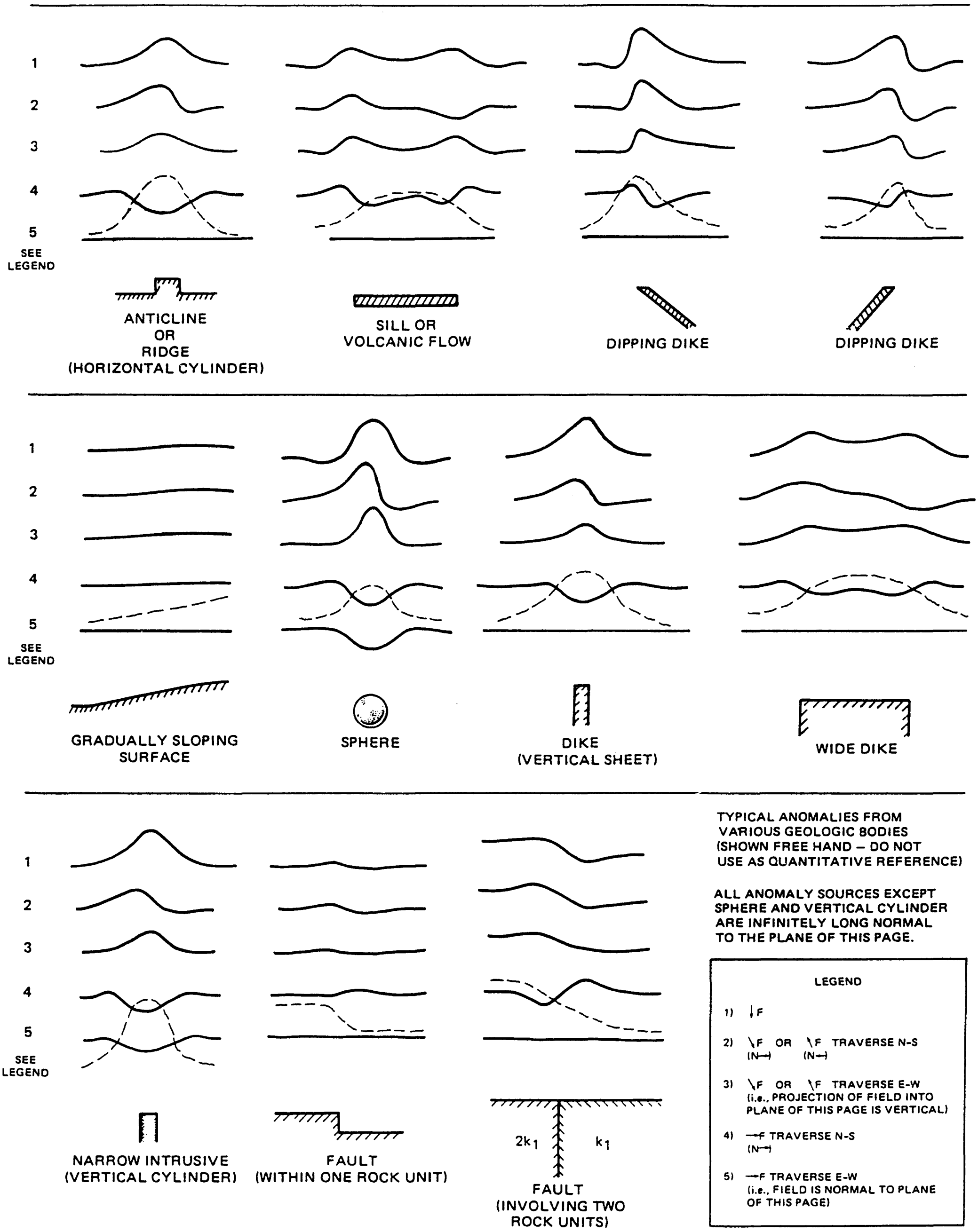
1
2
3
4
5
SEE LEGEND
ANTICLINE OR RIDGE (HORIZONTAL CYLINDER)
SILL OR VOLCANIC FLOW
DIPPING DIKE
DIPPING DIKE
GRADUALLY SLOPING SURFACE
SPHERE
DIKE (VERTICAL SHEET)
WIDE DIKE
NARROW INTRUSIVE (VERTICAL CYLINDER)
FAULT (WITHIN ONE ROCK UNIT)
2k1
k1
FAULT (INVOLVING TWO ROCK UNITS)
LEGEND
1) ↓F
2) \F OR \F TRAVERSE N-S (N→) (N→)
3) \F OR \F TRAVERSE E-W (i.e., PROJECTION OF FIELD INTO PLANE OF THIS PAGE IS VERTICAL)
4) →F TRAVERSE N-S (N→)
5) →F TRAVERSE E-W (i.e., FIELD IS NORMAL TO PLANE OF THIS PAGE)

TYPICAL ANOMALIES FROM VARIOUS GEOLOGIC BODIES (SHOWN FREE HAND – DO NOT USE AS QUANTITATIVE REFERENCE)

ALL ANOMALY SOURCES EXCEPT SPHERE AND VERTICAL CYLINDER ARE INFINITELY LONG NORMAL TO THE PLANE OF THIS PAGE.

FIG. 5. Schematic gravity (dashed) anomalies, and total-intensity magnetic (solid) anomalies (after Breiner, 1973), over idealized geologic sources with positive physical property contrasts (symbolized by k_1 for the fault between two rock units). Legend gives geomagnetic field (*F*) and anomaly profile orientations.

Evaluating anomaly values from the mathematical model of the source is called *forward modeling*. The solution of the inverse problem is based on how well anomaly predictions derived from forward modeling compare with the residual anomaly. However, a close match between predicted and residual anomalies does not guarantee a unique solution, but rather the solution is only one of a family of possible interpretations. This is because of the source ambiguity of potential fields, solution uncertainties related to errors in the residual anomaly values, and errors in postulating the mathematical model, which in every instance represents a simplified approximation of the source.

Numerous equations and algorithms exist in the literature for computing gravity anomalies of geologic sources that are used in the forward-modeling component of anomaly inversion. Relatively complex computations of gravity anomalies derived from arbitrarily shaped two- and three-dimensional sources are routinely performed by established procedures (e.g., Hinze, 1990). Simpler computations that idealize the anomaly source in terms of symmetric shapes are presented in Table 2.

Upon completion of the modeling procedures and the consequent definition of the anomaly source characteristics, there remains the problem of putting the physical model and the density contrast into a geologic context. This calls for integration of geologic principles and laws and rock properties with the gravity model.

Table 2. Equations for computing gravity anomalies of ideal sources. Here g is the gravity anomaly in Gals (m/s^2) due to a density contrast σ in g/cm^3 (kg/m^3); G is the gravitational constant, 6.672×10^{-8} cm^3/g s^2; V is volume; R is the distance between observation point and center (sphere) or axis (cylinder) or center line; x is the horizontal distance from observation point to center (sphere) or axis (cylinder) or center line (sheet) or edge (fault); d is depth to center (sphere) or axis (cylinder) or top (d_1) or bottom (d_2) of sheet or center line (fault); t is thickness; r is radius; θ is angle from horizontal. All distances are measured in centimeters.

Source	Gravity anomaly
Basic	$g=\frac{G\sigma Vd}{R^3}$
Sphere	$g=\frac{4\pi G\sigma r^3}{3d^2}\frac{1}{[1+(x^2/d^2)]^{3/2}}$
Vertical cylinder (infinite)	$g=\frac{\pi G\sigma r^2}{(x^2+d^2)^{1/2}}$
Horizontal cylinder (infinite)	$g=\frac{2\pi G\sigma dr^2}{(x^2+d^2)}$
Narow vertical sheet	$g=\frac{2G\sigma t\ \ln(x^2+d_2^2)^{1/2}}{\ln(x^2+d_1^2)^{1/2}}$
Vertical fault	$g=2G\sigma t\theta$
Finite slab	$g=2G\sigma t(\theta_1-\theta_2)$

3. MAGNETIC EXPLORATION

Magnetic exploration is based on mapping perturbations or anomalies of the geomagnetic field related to lateral magnetic property variations of the subsurface. Magnetic anomalies detected by conventional magnetometers range from small-scale features derived from magnetic property differences within the uppermost meter of soil to very-long-wavelength anomalies related to broad-scale variations of the magnetic properties near the base of the lithosphere. Hence the magnetic method is used in the solution of geologic problems over a considerable number of applications ranging from archaeological and engineering site investigations, to exploration for economic resources, to studies of crustal tectonic features and processes.

The geomagnetic field is a potential field so that any set of magnetic observations can be related to an infinite number of magnetic source distributions. This inherent ambiguity of the magnetic method plus additional uncertainties resulting from the observation, reduction, and processing of these data limit the accuracy of geologic interpretation. This ambiguity can be significantly diminished, however, with the use of constraining geologic and geophysical data.

The application of the magnetic method to the solution of geologic problems requires a multiple-step process that is analogous to that previously described for the implementation of the gravity method.

3.1 Principles

Magnetic surveys involve measuring perturbations of the geomagnetic field caused by anomalous subsurface sources at or near the Earth's surface, extracting the geological component from these measurements, and translating the anomaly data into the anomalous geology. Perturbations of the geomagnetic field are caused by horizontal variations in the magnetic polarization of the Earth, and hence by horizontal variations of the magnetic properties of the subsurface. Magnetic anomaly fields are potential fields and thus obey Laplace's equation.

In magnetic exploration, measurements are predominantly restricted to the scalar intensity (magnitude) of the total component of the Earth's magnetic field because of instrumental limitations. This total-field component represents the vector sum of the normal geomagnetic field, which originates largely in the outer core of the Earth, and the anomalous effects of magnetic sources in the subsurface. As long as the normal geomagnetic field is large compared to the anomalous field, we can assume to good approximation that the total field anomaly represents a perturbation of the geomagnetic field along its principal direction. This assumption seems to hold almost always, except in the near fields of large steel objects, or in the vicinity of iron ore deposits, or various mafic rocks and other features which produce anomalies with absolute magnitudes equal to or greater than about 10,000 nT. In the older literature, another unit was commonly employed for magnetic field strength, known as the gamma (γ), which is equivalent to the nanotesla.

For a unit point dipole, the total-field magnetic anomaly at an observation point with generalized coordinates (x_1,x_2,x_3) is

$$T(x_1,x_2,x_3) = -\mathbf{c}\cdot\nabla[\mathbf{c}'\cdot\nabla'(1/R)]M, \qquad (13)$$

where R is the distance between the observation point and source point with generalized coordinates (x'_1,x'_2,x'_3) ∇ and ∇' are the vector gradient operators in observation-point and source-point coordinates, respectively; M is the magnitude of the dipole's magnetic moment; $\mathbf{c}$ is the unit geomagnetic field vector at the observation point; and $\mathbf{c}'$ is the unit vector in the direction of the source magnetic moment. To obtain the total field magnetic anomaly of an arbitrarily shaped body of uniform magnetization intensity J, or magnetic moment intensity per unit volume $M/dx'_1dx'_2dx'_3$, at the observation point, we integrate Eq. (13) over the volume of the body:

$$\int_{x'_{3a}}^{x'_{3b}}\int_{x'_{2a}}^{x'_{2b}}\int_{x'_{1a}}^{x'_{1b}} T(x_1,x_2,x_3)\,dx'_1dx'_2dx'_3, \qquad (14)$$

where the primed variables refer to the generalized coordinates of the anomalous body as described for Eq. (7) above.

In the application of Eqs. (13) and (14), the unit vectors $\mathbf{c}$ and $\mathbf{c}'$ are expressed in terms of geomagnetic field inclination (I,I') and declination (D,D') at the observation and source points, respectively. Also, the magnetization is a function of the geomagnetic field strength at the source point. These parameters vary markedly about the Earth to complicate the analysis and interpretation of total-field anomalies. For example, the inclination and intensity of the normal field increase, respectively from 0° and about 22 000 nT at the geomagnetic equator to 90° and nearly 70 000 nT at the geomagnetic pole. The normal geomagnetic field is shifted approximately 11° from the axis of rotation of the Earth as illustrated in Fig. 6. The components of the Earth's normal magnetic field in mid-latitudes in the northern hemisphere are illustrated in Fig. 7.

To see how variations of the normal geomagnetic field can affect interpretation of total-field anomalies, consider Fig. 8, which shows

1. the vectorial representation of the Earth's field,
2. the source, which is magnetized by induction in the Earth's field, and
3. the resultant of the two fields

at a series of positions along the magnetic meridian. It is the scalar of the resultant field that is measured by most magnetometers and that, when the Earth's normal field (scalar) is subtracted, results in the anomalous field. The anomalous field is plotted above the cross section of the Earth and shows that a source of positive magnetization (higher magnetization than the country rock), which produces a

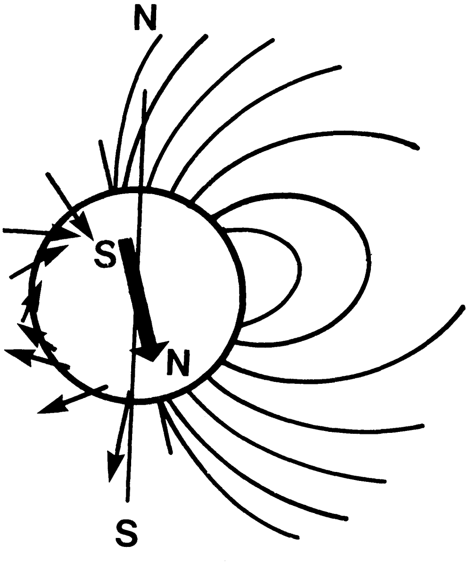

FIG. 6. Schematic diagram showing the lines of magnetic force and vectorial representation of the geomagnetic field.

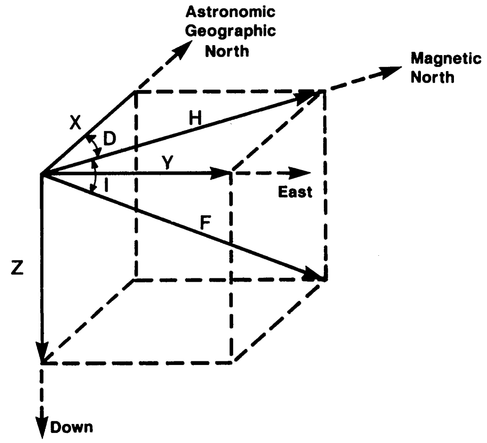

FIG. 7. Schematic diagram of components of the geomagnetic field at mid-latitudes in the northern hemisphere. In geomagnetism, declination (*D*) and inclination (*I*) are also called magnetic strike and dip, respectively.

dipolar charge with a negative pole near the upper surface in the northern geomagnetic hemisphere, will actually result in a negative anomaly on the north side of the positive anomaly. Also, the peak of the positive value is shifted to the south from the magnetic source. This shift and the intensity of the negative anomaly will increase with decreasing magnetic inclination to 45° where the positive and negative portions of the anomaly are antisymmetrical and the crossover from positive to negative values occurs directly over the anomalous source. The variation in the anomaly becomes even greater at lower inclinations, resulting in a symmetrical magnetic minimum over a positive magnetic source at the magnetic equator. In the southern magnetic hemisphere, the anomalies are reversed, with the minimum to the south of a positively magnetized source and the peak of the positive anomaly shifted to the north. The anomalies will be inverted where the anomalous source is less magnetic than the country rock. Over the United States, the magnetic anomaly from the same source magnetized by the Earth's field will increase in symmetry from the southern to the northern limits and will increase in intensity because of the increasing inclination and geomagnetic field strength.

For a point mass that is also magnetized, the anomaly-field calculations involve the same observation-to-source point geometries, so that the magnetic and gravitational potentials of the common source can be related. We can use this result, which is a manifestation of Poisson's theorem, to combine, for example, Eqs. (6) and (13) at the geomagnetic pole to obtain

$$T = (J/\sigma G)(\partial g/\partial x_1). \qquad (15)$$

This equation has many uses in comparing gravity and magnetic fields, but it shows in very simple terms that the magnetic field

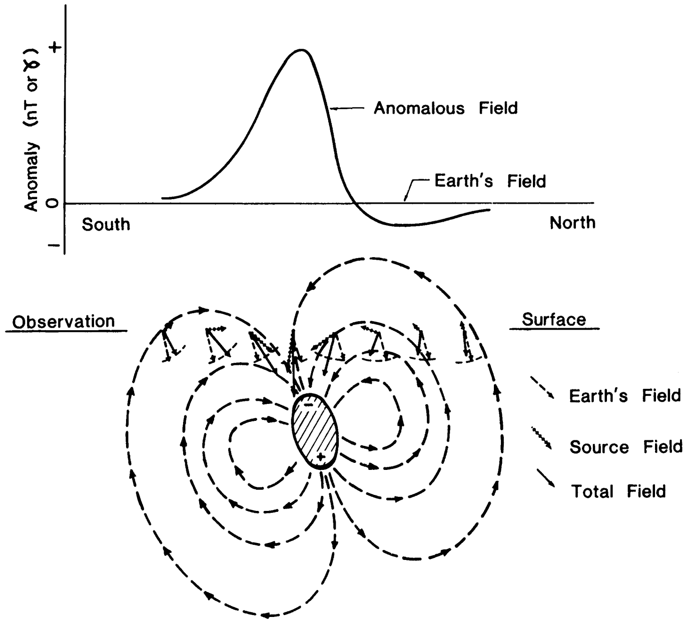

FIG. 8. Schematic illustration of the summing of the Earth's magnetic field and the anomalous source field to produce the observed total magnetic intensity plotted at the top of the figure.

varies one power faster with respect to distance than does gravity. Thus the magnetic anomaly of a source will be more sensitive to depth and will better define the location of the source than the gravity anomaly from the same source.

Conventional implementation of the above magnetic-anomaly equations expresses the effect of the subsurface body directly in terms of its magnetization intensity, J. However, in magnetic exploration, the magnetization intensity contrast ΔJ is more commonly used, defined as the magnetization intensity of the subsurface source minus the magnetization intensity of the surrounding country material. Similarly, the observed magnetic values are compared with the response of a homogeneous Earth; that is, the anomaly value is the observed value minus the theoretical or the normal value where there are no horizontal variations in magnetic polarization.

3.2 Magnetic Properties

The intensity of magnetization or the magnetic polarization J of a magnetized body is defined as the magnetic moment per unit volume, and thus it is a function of the number and orientation of elementary dipoles. The magnetization is the vectorial summation of the induced and remanent (permanent) magnetization. Induced magnetization is the magnetic polarization caused by Earth's normal (main) field and is the product of the intensity of the Earth's field and the magnetic susceptibility of the substance.

The magnetic susceptibility of rocks is primarily dependent on the volume percent content of magnetite. This ferrimagnetic mineral is a common accessory mineral in igneous and metamorphic rocks and occurs in trace amounts in sediments and sedimentary rocks. In general, the magnetic susceptibility of rocks increases with the mafic content, but the notable exceptions suggest caution in the use of this "rule" as anything other than a broad generality.

Remanent magnetization is another component of the magnetic polarization of rocks that may originate in different ways and at different times in the history of the rock. The net remanent magnetization of a rock in place is the natural remanent magnetism. This is the magnetization measured by the field magnetometer. In paleomagnetism studies, the less stable components are stripped from the rock so that particular remanent components, such as thermal remanent magnetism imposed upon an igneous rock as it cooled through the Curie point, can be investigated. Sedimentary rocks typically have low values while unmetamorphosed basalts have high natural remanent magnetizations.

In engineering, environmental, and archaeological studies with the magnetic method, we are interested not only in the magnetic properties of the common rock types, but also in those of soils and iron objects. The magnetic susceptibilities of soils reflect their parentage. Soils derived from igneous and metamorphic rocks, particularly the more mafic components of these lithologies, commonly will be relatively rich in magnetite. This is emphasized by the resistance of magnetite to alteration. Furthermore, magnetite tends to concentrate in sediments where the streams carrying magnetite and other heavy minerals lose velocity and thus deposit these mineral grains. As a result, magnetite tends to concentrate in the distributary streams in alluvial fans at the base of mountains, and in terranes covered by glacial deposits, magnetite will be concentrated in eskers, outwash deposits, and beaches of ancient glacial lakes. In addition, maghemite, a relatively magnetic form of hematite, is produced in high-organic soils. Surface soils often acquire a remanent magnetization with twice the intensity of induced magnetization by thermal effects due to fires or simply by isothermal "viscous" remanent magnetization acquired over a long period of time. Disruption of the soils by man will result in local anomalies as will natural or man-made fires. Relative to most rocks and soils, typical iron or steel objects such as pipes, drums, etc. have very large magnetizations, which often include remanent magnetizations that reach values of 10 or more times the induced components.

3.3 Instrumentation

Essentially all magnetic measurements made, whether in a stationary mode or from a moving platform, use a fluxgate, proton-precession, or alkali-vapor magnetometer. These have individual attributes that need to be considered in selecting an instrument for a particular magnetic survey.

The fluxgate magnetometer measures only the relative change in the Earth's magnetic field and is directionally dependent. The latter quality is a disadvantage because it must be oriented for measurements, and an advantage because not only can the total field be measured as is the case for the other magnetometer types, but by proper orientation the instrument can be used to measure the vertical and horizontal components of the field, which may provide an interpretational advantage under specific circumstances. The fluxgate magnetometer normally has a sensitivity of 1 nT, although with special care a sensitivity of 0.1 nT has been achieved in airborne sensors. Problems encountered in orientation of the fluxgate magnetometer on a moving platform require that the sensor be oriented in the direction of the geomagnetic field; thus the total field is measured.

The most commonly employed magnetometer in both stationary and mobile modes is that of the proton-precession magnetometer, which measures the absolute total intensity of the geomagnetic field, without the need for orienting devices, to a precision of the order of 0.1 nT. The proton-precession magnetometer measures the precession frequency of protons in a hydrogen-rich fluid that have been oriented at a large angle to the Earth's field by the strong magnetic field from a current passed through a coil of wire wound around the fluid container. On termination of the current through the coil and relaxation of the imposed magnetic field, the protons return to their original alignment about the Earth's normal field direction with a precessional frequency related to the absolute Earth's field through a well-known constant, the gyromagnetic ratio. The proton-precession magnetometer has many advantages, but it has two principal disadvantages. The observations of the magnetic field are not continuous; rather they require a finite period of time for orienting the protons with the current passing through the coil and for measuring the precessional frequency. Decreasing the reading interval will decrease the sensitivity of the instrument. This may cause a problem in making observations from a mobile platform where a choice must be made between data interval and data sensitivity. However, sampling rates can be increased with an Overhauser-effect proton magnetometer. Furthermore, the proton-precession magnetometer is subject to erroneous observations resulting from ac power interference and large magnetic gradients. Gradients of the order of 300 to 1000 nT/m, which may occur close to iron or steel objects and magnetic iron formations, may cause the instrument to malfunction.

The alkali-vapor magnetometer, which is capable of an order of magnitude or greater sensitivity over the proton-precession magnetometer, is primarily used for airborne observation of the total magnetic field. This magnetometer is based on the principle of optical pumping used in radio-frequency spectroscopy. The sampling rate of the alkali-vapor magnetometer is much higher than that of the proton-precession magnetometer and it will operate in much higher magnetic gradients. Unfortunately, the optical pumping principle will not operate when the axis of the instrument is parallel or perpendicular to the magnetic field. This problem has been eliminated in some units by including multicelled sensors.

Measurement of the gradients of magnetic fields, both vertical and horizontal, offers several attractive advantages over simple fields. Gradients minimize regional anomalies, increase resolution, and eliminate temporal magnetic variations. As a result, magnetic gradiometers, normally employed in the vertical mode and occasionally in a horizontal configuration, are regularly used, especially in the study of near-surface anomalies. Magnetic gradiometers are simply two magnetic sensors that are separated by a fixed distance, with measurements made simultaneously in both and compared for recording. Hood *et al.* (1979) present a comprehensive discussion of magnetic gradiometers as well as other magnetic instrumentation.

3.4 Surveying Procedures

Magnetic surveying for geoexploration purposes can be land, airborne, or shipborne. In land measurements, observations can be made at extremely close intervals, which permits high resolution of near-surface sources. However, they are time consuming and expensive if the survey is measured in kilometers. Furthermore, surface observations are susceptible to local magnetic anomalies originating from cultural features and magnetic polarization variations in soils and near-surface bedrock. The effects of near-surface magnetic sources are minimized by placing the magnetic sensor on a staff of several meters height and by making several observations within a limited area (a few to several meters) to detect any local source that would invalidate the observation site. However, for purposes of mapping magnetic variations in the soil, the sensor may be placed at only about 0.5 m above the ground surface. In any event, several measurements are generally made at each site to improve the precision of the observation by using an average value and to detect instrument malfunction and temporal variations in the magnetic field. Stations are laid out either on a grid or, in mapping linear anomalies, on profiles perpendicular to the magnetic strike, that may be spaced at a distance of several times the station interval.

Temporal variations in the geomagnetic field must be monitored and removed from most land surveys. These variations can be monitored by reoccupying a base station at intervals not to exceed one hour in magnetic "quiet" periods or by directly monitoring the temporal variations with a recording magnetometer established at a base within the survey area. For surveys covering limited regions, the base magnetometer can be electronically tied to the moving magnetometer and only the difference in the measurements made at the same time needs to be recorded.

Airborne magnetic surveying is preferred for most applications except for high-resolution studies of limited areas. Great care is required in planning airborne surveys to achieve the study objective in a timely and economical manner. Reid (1980) showed that to avoid aliasing errors in the data, the survey flight lines should have a maximum spacing of twice the combined altitude and depth to the expected source of anomalies, and that this should be decreased to the combined altitude and depth in cases where the data will be processed for computation of gradients, etc. The sampling interval along the flight line should be much less than this spacing. An interval of 50 m is commonly used. Flight lines should be flown perpendicular to the prevailing strike of the anomalies, with additional control lines at right angles to the principal flight lines at a spacing of roughly five times the basic flight-line interval. Control lines are used to tie the data of the principal flight lines together, minimizing temporal variations. Temporal variations are monitored by a ground-based continuous recording magnetometer, and commonly these variations are directly removed from the data before tying together the control and flight lines.

Flight elevation is a critical specification, which is generally minimized commensurate with safe operation of the aircraft and navigational requirements to maximize anomaly resolution. Surveys are flown either at a constant barometric altitude or at a constant terrain clearance, depending on the survey objectives and local geology and topography. For example, in magnetic surveying for geological mapping purposes, the principal concern is for anomaly resolution, thus these surveys are normally flown at a constant terrain clearance of 150 m or less. Survey altitude is monitored by radar and barometric altimeters, and positions are fixed by a combination of photography or videotape, radar altimeter, and electronic navigational aids. Active navigational systems deployed to increase the precision of positioning can significantly increase the cost of magnetic surveying.

3.5 Magnetic-Data Reductions

Magnetic observations can be affected by a number of extraneous sources in addition to the signature of subsurface magnetic polarization effects that we are interested in mapping. For example, the total magnetic intensity is subject to vertical changes of the geomagnetic field at the surface of the Earth of approximately 0.025 nT/m over the conterminous U.S.A. Most magnetic observations, unlike gravity measurements, are made with absolute-reading instrumentation, eliminating the need for reference stations.

Only two corrections are generally applied to observed magnetic measurements to determine the anomaly value. These account for temporal and spatial variations of the magnetic field. Temporal variations are primarily caused by particle and electromagnetic radiation from the Sun acting upon the external magnetic field, of the Earth. Lesser effects are caused by tidal effects of the Moon perturbing the ionosphere of the Earth and thus the geomagnetic field, and magnetic fields from electrical currents induced in the Earth. These temporal variations have a wide range of period and amplitude such that significant effects are observed in magnetic surveying and thus must be eliminated in most applications. Temporal variations in the geomagnetic field are monitored by reoccupation of a base at prescribed time intervals or by a recording base magnetometer, or, in the case of aeromagnetic surveying, by leveling survey observations to minimize differences in cross-line data.

Spatial variations in the geomagnetic field are caused not only by subsurface sources of magnetic polarization, but also by the geomagnetic field derived from the outer core of the Earth. To a first order, this field can be represented by a small, intensely magnetized dipole near the Earth's center and oriented approximately along the axis of rotation of the Earth (Fig. 6). An internationally accepted equation defines the field. This field, the International Geomagnetic Reference Field (IGRF), is derived from global magnetic observations and is updated every half decade. It includes time-variable terms that attempt to predict the secular (longterm) changes in the main field of the Earth from previous fluctuations, and it takes into account the latitude and longitude of the position on the Earth's surface as well as its height relative to sea level. In most local surveys, the station locations are relative only and are not defined in latitude and longitude. In these surveys the north–south and east–west gradients of the IGRF are determined and assumed to be linear over the survey area. Typically, east–west gradients range from zero near the center of the United States to 2.5 nT/km on the coasts, and in the north–south direction gradients change from 6 nT/km in the south to 2.5 nT/km along the Canadian border.

3.6 Anomaly Isolation and Enhancement

Magnetic anomaly fields combine the effects of all horizontal variations in magnetic polarization within the Earth, and thus the anomalies of interest in a particular geologic application are commonly distorted by overlapping anomalies as is the case for gravity anomalies. The requirements and procedures for isolating and enhancing residual magnetic anomalies are completely analogous to those implemented in the gravity residual anomaly separation problem described in Sec. 2.7.

Magnetic anomalies, like gravity anomalies, are readily represented by Fourier transforms to facilitate isolating and enhancing residual anomalies. Accordingly, anomaly attributes such as their wavelength and directional components, gradients, and continuations may be efficiently and accurately estimated by filtering the transform. Additional processing specific to the analysis of magnetic anomalies includes the reduction-to-pole (RTP) filter that is used to remove the distorting effect of magnetic polarization at non-normal inclination from the inductive components of the magnetic anomaly.

In conventional exploration applications, the geomagnetic-field attributes (i.e., intensity, declination, and inclination) can be assumed to be constant and equal over the subsurface sources and at the observation positions. Accordingly, total-field magnetic anomalies observed where the geomagnetic-field inclination is less than normal (vertical) are related by their directional derivatives to their equivalent anomalies at normal polarization. These directional derivatives, in turn, are nothing more than the horizontal and vertical gradients of the total magnetic anomalies weighted by the known values of geomagnetic field inclination and declination. Because the horizontal and vertical gradients are readily available from the Fourier transform of the anomalies, the design and implementation of the RTP filter are straightforward.

The RTP filter adjusts the total-field magnetic anomalies to the idealized condition of vertical polarization, which is achieved in practice only at the geomagnetic poles. As a result of this operation, the source is centered within its inductive anomaly components in the same manner that a gravity source is centered within its anomaly field. Hence, where remanent magnetic anomaly components are negligible, the RTP filter facilitates the correlation of magnetic sources with their anomalies and anomaly gradients to enhance subsurface modeling efforts. In addition, comparing total-field magnetic anomalies with their equivalent RTP anomalies facilitates identifying anomalies related to sources with significant remanent magnetization. RTP magnetic anomalies also are required to undertake quantitative analyses of the correlations between magnetic and gravity anomaly fields described in Eq. (15).

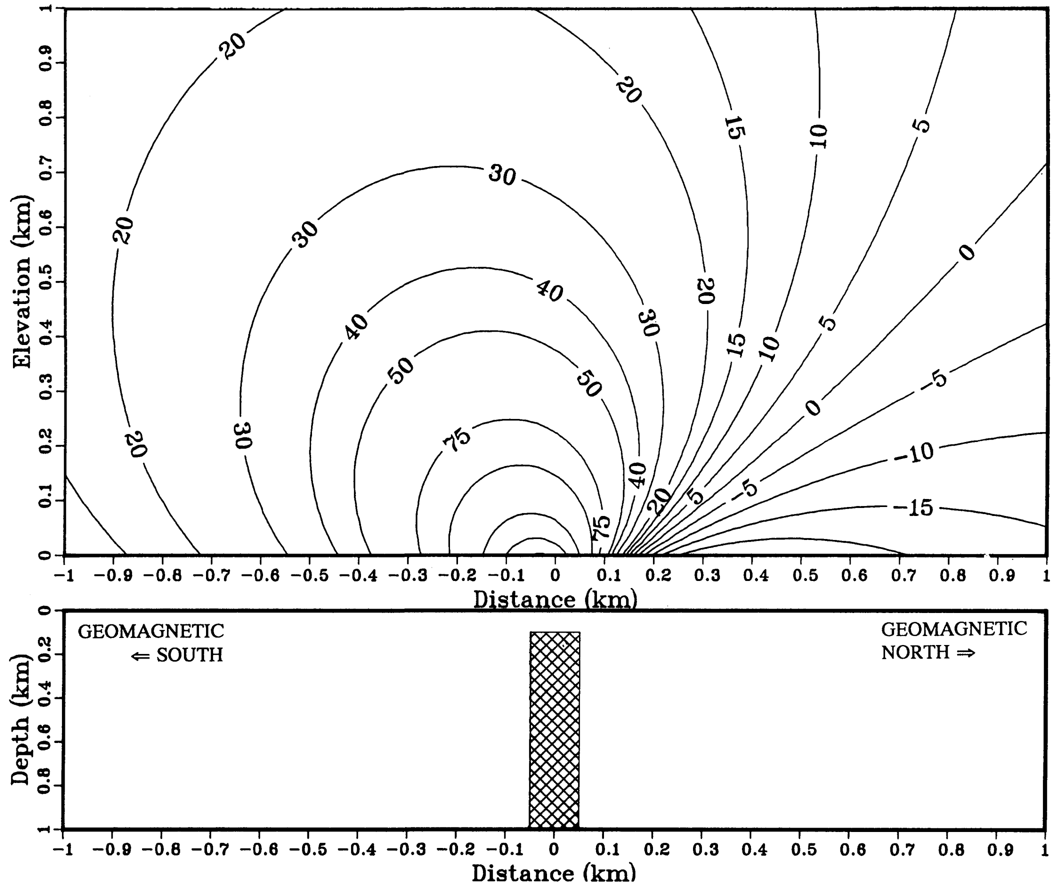

FIG. 9. Vertical contour map of total intensity magnetic anomaly in gammas (nT) for the vertical dike source with cross-section as shown extending infinitely perpendicular to the south-to-north (S-N) section in the northern geomagnetic hemisphere. The magnetization is induced due to a magnetic susceptibility contrast of 0.0025 emu/cm^3 in a field (*F*) of 60,000 nT oriented with 75° inclination (*I*) from the horizontal and 0° declination (*D*) from geographic north. Note variable contour interval.

3.7 Interpretation

The generalized approach to magnetic-anomaly interpretation is analogous to the approach for gravity-anomaly interpretation, and the reader is referred to Sec. 2.8 for more detailed discussion.

The amplitude and character of the anomaly signatures of magnetic sources in the subsurface are dependent on several key variables. Anomalies vary with source geometry, including volume and shape, and with source depth. The amplitude of the magnetic anomaly is a function of the magnetic property contrast of the source, which varies in a complex manner because of the combination of magnetization induced by the geomagnetic field and the intensity and direction of the remanent magnetization. Thus, in considering typical magnetic anomaly signatures, we must consider the relative orientation of the source in the geomagnetic field, the directional and intensity attributes of the geomagnetic field, and the total (induced plus remanent) magnetization of the source.

Typical total magnetic intensity anomaly profiles are presented in Fig. 5 in a schematic fashion for a variety of geometric shapes under five different geomagnetic field directions. The rate of decay in amplitude with distance R between sensor and source is an important variable, which changes with source shape and magnetization direction. Anomaly amplitude is proportional to $1/R^n$, where n is a real number decreasing from 3 to 0 as the

Table 3. Maximum depth to the center (Z_c) or top (Z_t) of ideal magnetic sources based on one-half the width of the anomaly ($X_{1/2}$) at one-half its maximum amplitude.

Geometric form	Depth
Sphere (single dipole)	$Z_c \leqslant 2.0 X_{1/2}$
Horizontal cylinder (line of dipoles)	$Z_c \leqslant 2.0 X_{1/2}$
Vertical cylinder (single monopole)	$Z_t \leqslant 1.3 X_{1/2}$
Edge of narrow dike (line of monopoles)	$Z_t \leqslant 1.0 X_{1/2}$

magnetic source geometries increase, respectively, from a concentrated spherical body to large tabular features of great depth extent.

The effect of observation elevation on magnetic anomalies is particularly well illustrated in Fig. 9. The plot shows vertical contours of the total intensity magnetic anomaly in gammas (nT) of an infinite vertical prism (dike) with a magnetic susceptibility contrast of 0.0025 emu/cm^3 and oriented perpendicular to the magnetic meridian at 75° magnetic inclination. The change in magnetic anomaly with elevation above the source can be readily visualized by considering the anomaly profiles indicated by the intersection of the contour lines with horizontal lines at a specified elevation.

Numerous techniques have been developed for approximating the depth to sources from a measure of the gradient of observed anomalies. These techniques provide maximum depth to the top or center of the source and are based upon recognizing the geometric form of the anomaly source from the anomaly characteristics. In general, the magnetic method provides greater depth resolution than the gravity method and thus is preferred wherever possible. The greater resolution is a result of the higher exponent in the decay function and the relative insensitivity of the magnetic method to depth extent. One of the useful depth techniques is the half-width measurement where one-half the width of the anomaly at one-half the amplitude is related to maximum depth (Nettleton, 1940). The relationships in the magnetic case are given in Table 3, based upon an assumption of vertical magnetic polarization.

The general inversion problem of estimating quantitative parameters of the source from the anomaly data is completely analogous to the gravity case, and the reader is referred to Section 2.8 for a more detailed discussion of the process. The quantitative development of forward modeling in magnetics is slightly more complex than that for gravity applications because of the dipolar formalism. This is exemplified in Table 4, which gives the equations for computing magnetic anomalies of simplified sources assuming vertical magnetic polarization. Tables 2 and 4 can be used together in joint inversions of gravity and magnetic anomalies involving common sources.

Table 4. Equations for computing magnetic anomalies of ideal sources. Here T is the magnetic anomaly in oersteds (teslas) in a vertical polarizing magnetic field caused by a magnetic susceptibility contrast k in emu/cm^3; k_m is the susceptibility of magnetite ($\simeq 0.25$); H is the vertical magnetic field in oersteds; V is volume; R is the distance between observation point and center (sphere) or axis (cylinder) or center line; x is the horizontal distance from observation point to center (sphere) or axis (cylinder) or center line (sheet) or edge (fault); d is depth to center (sphere) or axis (cylinder) or top (d_1) or bottom (d_2) of sheet or center line (fault); t is thickness; r is radius; E is depth extent; θ is angle from horizontal. All distances are measured in common linear units (e.g., centimeters, meters, kilometers, inches, feet, yards, kilofeet, miles, etc.)

Source	Magnetic anomaly
Basic	$T = kHV/R^n$, where $k = k_m \times$ (percentage magnetite)
Sphere	$T = (8\pi k H r^3 / 3d^3)(1 - x^2/2d^2)(1 + x^2/d^2)^{-5/2}$
Vertical cylinder (infinite)	$T = \pi d k H r^2 (x^2 + d^2)^{-3/2}, \quad E \gg d$
Horizontal cylinder (infinite)	$T = 2\pi k H r^2 (d^2 - x^2)(x^2 + d^2)^{-2}$
Narrow vertical sheet	$T = (2kHt)[d_1/(x^2 + d_1^2) - d_2/(x^2 + d_2^2)]$
Vertical fault	$T = 2kHtx/(x^2 + d^2)$
Finite slab	$T = 2kH(\theta_1 - \theta_2), \quad E \ll d$

4. ELECTRICAL EXPLORATION

As a major branch of geophysics, electrical exploration invokes the principles of geoelectricity for mapping subsurface structures of significance to prospecting for mineral and energy resources, and for solving hydrogeological, engineering, and archaeological problems. A considerable diversity of methods is available for geological exploration. These methods may be broadly classified according to whether they measure the ground's electrostatic (i.e., time-invariant) or electrodynamic response.

Steady-state techniques include the spontaneous-polarization (SP), resistivity, and induced-polarization (IP) methods. The results of SP and IP surveys are interpreted in terms of electrochemical potential differences in the subsurface that depend on the mineralogy of the rocks and on the composition and concentration of electrolytes dissolved in the ground water with which the rocks are in contact. Resistivity surveys map vertical and horizontal variations of subsurface electrical resistivity, or its reciprocal property, conductivity, which are related mostly to the distribution and salinity of the ground water and the resistivities of the subsurface materials.

Electrodynamic techniques include inductive electromagnetic (em) surveys and the telluric and magnetotelluric methods. The results of these methods are interpreted primarily in terms of the subsurface resistivity and dielectric properties, although magnetic permeability can also become a factor when significant concentrations of highly magnetic material such as magnetite occur within the subsurface.

4.1 Self-Potential Method

The self-polarization (or spontaneous polarization) method is a passive technique that involves measuring naturally occurring potential differences at the Earth's surface due to subsurface electrochemical activities that govern the magnitude and sign of the voltage developed when the subsurface materials are in equilibrium with an electrolyte. SP anomalies are observed over electrically conductive bodies that cut across the water table below the surface. Such conditions of the subsurface region tend to be characterized by SP anomalies that are minima with magnitudes that can reach up to several hundred millivolts.

A widely accepted model for the generation of SP anomalies at the Earth's surface is shown in Fig. 10. Electrolytes in the pore fluids below the water table release electrons through oxidation, which migrate to the top of the conductive body. These electrons reduce the electrolytes at the top of the body to produce negative ions, which may be electrolytically conducted down the system in the pore fluids. This model of electronic and electrolytic currents can polarize the body as shown in Fig. 10 for the sign and general magnitude of observed SP anomalies. The model is also consistent with the general temporal stability of SP anomalies because the body is fundamentally unaltered, serving only to transport electrons from depth.

Field procedures involve the use of nonpolarizing electrodes connected to a high-impedance millivoltmeter. Surveys commonly involve profiling at a fixed station spacing (often less than 30 m), or by maintaining one electrode in a fixed position and maneuvering the roving electrode over the survey area.

SP anomalies most often are interpreted on a qualitative basis, where the anomalous body is presumed to occur directly beneath the anomaly minimum. However, quantitative modeling can be readily implemented to estimate the depth, shape, and other parameters of the body in the subsurface using fundamental electrostatic theory.

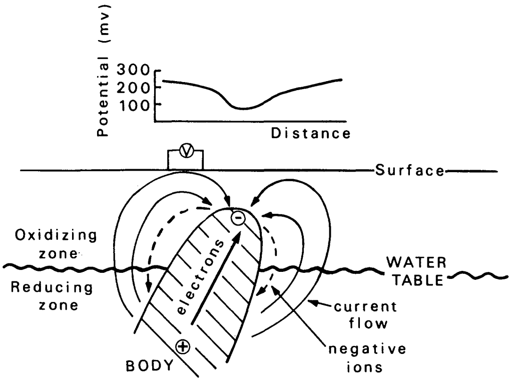

FIG. 10. The mechanism of self-potential anomalies (modified from Sato and Mooney, 1960 by Keary and Brooks, 1984). Natural potential profile and current pattern where sulfide body penetrates oxidizing zone above the water table are also illustrated.

4.2 Resistivity Methods

The resistivity method is routinely implemented in engineering, hydrogeological, and archaeological studies to investigate the shallow subsurface region in terms of vertical and horizontal variations of its resistivity properties. This is an active method whereby electric currents are introduced into the ground and the resulting potential differences are measured at the surface. By Ohm's law for zero frequency or direct current (dc), the potential difference ΔV (volts) between any two points in a circuit is related to the resistance R (ohms) of the medium between the two points and the current I (amperes) flowing in the conductive medium between the two points by

$$\Delta V = IR. \tag{16}$$

Experimentally, however, we also observe that the resistance (R) of a three-dimensional unit element of a material increases with its length L and decreases with its cross-sectional area A, according to

$$R = \rho(L/A), \tag{17}$$

where the proportionality constant ρ is defined as the resistivity (Ω m) of the material. Resistivity is the resistance of a unit element of the material. For some problems it is convenient to consider the reciprocal of resistivity which is called conductivity (S/m, where S denotes siemens). Resistivity surveying is grounded on Eqs. (16) and (17), where Eq. (16) is the basis of the field measurements, and Eq. (17) is the basis for interpreting the results of the survey.

Earth resistivities are usually measured by use of four-terminal electrode arrays positioned at the Earth's surface, where two of the electrodes galvanically drive dc current into the ground and the remaining two electrodes measure the potential difference over the surface. Electrode arrays are used to eliminate the effect of contact resistance in the measurements, which are obtained by averaging field observations made from normal and reversed-polarity input currents. Modern field systems also use very low-frequency ac current to cancel contact resistances around the electrodes. Laboratory studies have shown that resistivities observed at frequencies less than about 10 Hz are essentially the same as those obtained with dc measuring currents.

In resistivity surveying, several arrays are commonly implemented including the Wenner, Schlumberger, dipole-dipole, and pole-dipole arrays, which are illustrated in Fig. 11. The quantities measured in a field determination of resistivity are the current I flowing between the two current electrodes, the difference in potential ΔV between the potential-measuring electrodes, and the distances between the various electrodes. Accordingly, Earth resistivity as measured by four-terminal electrode arrays is given by

$$\rho = (\Delta V/I)K \equiv \rho_A, \tag{18}$$

where K is called the geometric factor of the array (Fig. 11). This equation may be used to compute the resistivity of the Earth only where the Earth is completely homogeneous. In a heterogeneous Earth, this equation is used as a definition for the apparent resistivity ρ_A of the subsurface. Values of the apparent resistivity may bear no relation to any actual value of resistivity in a heterogeneous Earth. However, the variations of apparent resistivity are indicative of how subsurface resistivities vary with depth or horizontally, so that considerable subsurface insight is revealed by obtaining and analyzing apparent resistivity.

The manner in which subsurface information is obtained from resistivity surveys can be shown using Fig. 12. The upper panel illustrates current flow and associated equipotentials for the Wenner electrode array on the

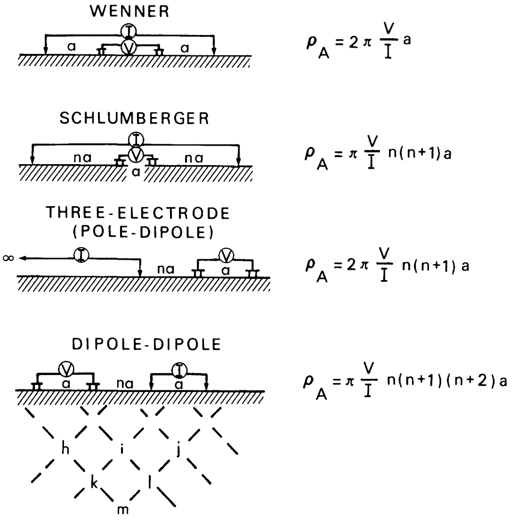
WENNER
I
V
a
a
$\rho_A = 2\pi \frac{V}{I} a$
SCHLUMBERGER
I
V
na
na
a
$\rho_A = \pi \frac{V}{I} n(n+1)a$
THREE-ELECTRODE (POLE-DIPOLE)
∞
I
V
na
a
$\rho_A = 2\pi \frac{V}{I} n(n+1)a$
DIPOLE-DIPOLE
V
I
a
na
a
$\rho_A = \pi \frac{V}{I} n(n+1)(n+2)a$
h
i
j
k
l
m

FIG. 11. Common resistivity arrays.

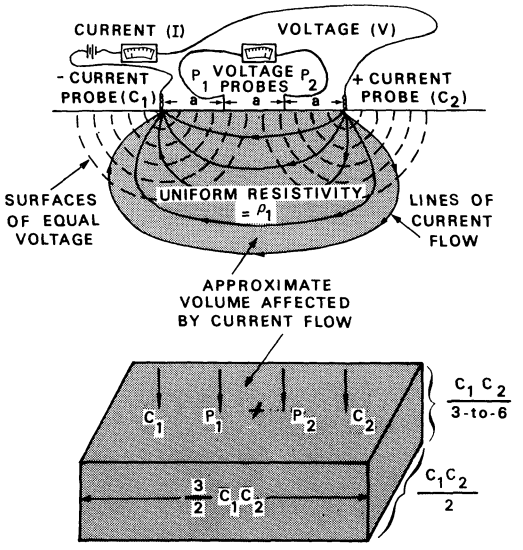
CURRENT (I)
VOLTAGE (V)
- CURRENT PROBE (C_1)
P_1 VOLTAGE PROBES P_2
+ CURRENT PROBE (C_2)
a
a
a
SURFACES OF EQUAL VOLTAGE
UNIFORM RESISTIVITY = ρ_1
LINES OF CURRENT FLOW
APPROXIMATE VOLUME AFFECTED BY CURRENT FLOW
C_1
P_1
P_2
C_2
$\frac{C_1C_2}{3\text{-to-}6}$
$\frac{3}{2}C_1C_2$
$\frac{C_1C_2}{2}$

FIG. 12. Resistivity surveying (upper panel) including a highly generalized approximation of current distribution in the subsurface region from current electrodes spaced at distance C_1C_2 (lower panel).

surface of a homogeneous material with resistivity given by ρ_1. The distribution of current within the three-dimensional subsurface may be grossly generalized in terms of the distance between the current electrodes (C_1C_2) as shown in the lower panel of Fig. 12. For the case considered here, the apparent resistivity obtained by the array will equal the resistivity of the subsurface to within the precision of the survey measurements. However, by making the array size larger, we can investigate a larger and deeper volume of the subsurface, so that if the subsurface were characterized by zones of increasing or decreasing resistivity, the apparent resistivity would increase or decrease, respectively. This mode of surveying is called *vertical electrical sounding* (VES), where we map vertical resistivity variations by measuring ρ_A as a function of changing the spacing a of the electrodes while holding the center of the array fixed.

Interpretations are based on plots of ρ_A versus the electrode spacing a as illustrated in Fig. 13. Inversion of VES data is most commonly quantified in terms of multiple horizontal resistivity layering of the subsurface. First-order estimates of layer interface depths are obtained by curve matching, where the VES data are normalized for comparison with a master set of curves to identify possible layered models of the subsurface. These estimates may be further refined by trial-and-error inversion using computer modeling.

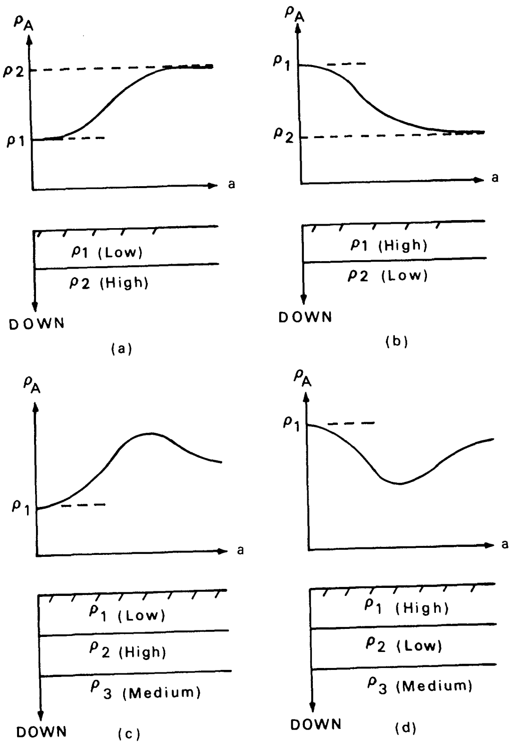

FIG. 13. Vertical electrical soundings (VES) over various layered subsurface structures.

The VES profiles are most commonly obtained using the Wenner and Schlumberger arrays. In surveying with the Wenner array, all four electrodes need to be moved between successive measurements of the apparent resistivity. The field effort to move the electrodes can be minimized with the use of Schlumberger array, where the inner potential electrodes may remain fixed and the current electrodes are expanded symmetrically about the center of the spread to obtain successive measurements of the apparent resistivity. However, when the distance between the current electrodes becomes very large, it may be necessary to increase the separation distance between the potential electrodes to obtain a measurable potential difference.

A second major mode of resistivity surveying is called *horizontal profiling*, where apparent resistivity is mapped as a function of changing offset distance of the center of the array while holding the spacing a of the electrodes constant. Lateral resistivity variations occurring within the depth range of the array will be reflected by similar variations in the apparent resistivity values. Horizontal profiling interpretations are based on plots of the apparent resistivity versus offset distance (x) such as illustrated by the examples given in Fig. 14. In these applications, horizontal profiles for different electrode spacings may be differenced to resolve further details on the lateral resistivity structure of the subsurface.

Horizontal profiling known as the *mise à la masse* method can be employed when a portion of the target ore body has been located either at outcrop or by drilling. For this application, the pole-dipole array is used where a single current electrode is placed within the body while the other current electrode is placed outside the body and far away, and the potential electrodes are moved over the Earth's surface mapping equipotential lines. In contrast to horizontal profiling, this method can provide much more information on the extent, dip, strike, and continuity of the body in the subsurface.

The interpretation of apparent resistivity in terms of the resistivity properties of the subsurface is very complex because of the extreme variability and sensitivity of resistivity to the composition and fluids of the subsurface. Electrical resistivity shows the greatest variability of the physical properties of rocks. There is broad overlap between rock types, and hence identifying a rock type solely on the basis of resistivity is generally impossible. Certain minerals such as native metals and graphite conduct currents by the movement of electrons. However, most rock-forming minerals are insulators, and electrical current is carried through a rock mainly by electrolytic conduction involving the movement of ions in the pore waters. Thus, for most rock materials the porosity and salinity of the water in the pores are the most important factors governing subsurface resistivity. An empirical relationship for the dependence of resistivity on these factors is given by Archie's law,

$$\rho = aV_c^{-m}S^{-n}\rho_w, \tag{19}$$

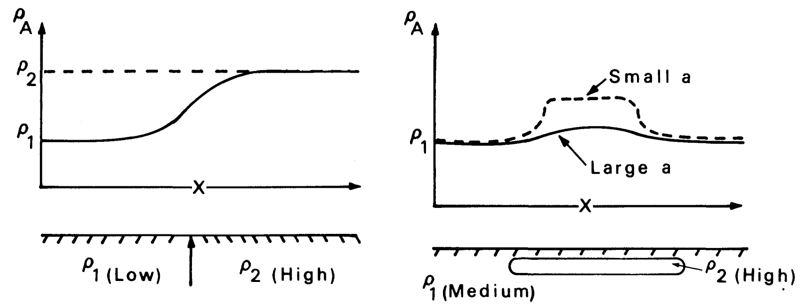

(a) Profiling across a vertical contact (b) Profiling across a shallow thin gravel deposit

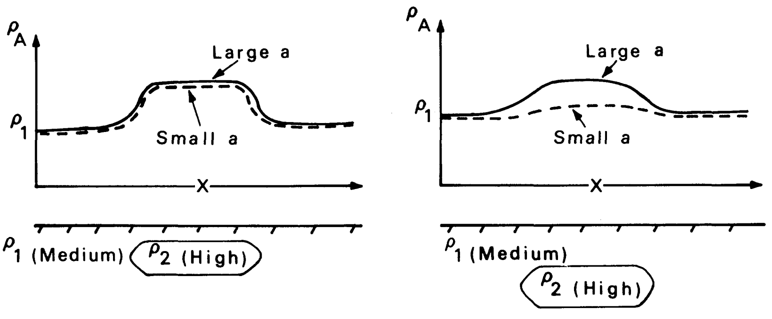

(c) Profiling across a shallow thick gravel deposit (d) Profiling across a deep thick gravel deposit

FIG. 14. Horizontal electrical profiling across various subsurface structures.

where ρ_w is the resistivity of the pore water, V_c is the fractional void content (porosity), S is the fraction of pores that are fluid filled, n is approximately 2, and $0.6 < a < 2.5$ and $1.3 < m < 2.5$, generally.

When porous rocks lie above the water table at shallow depths or when they occur at sufficiently great depths that all pore spaces are closed by the ambient pressure, the current flowing through them is maintained by electronic conduction in the mineral grains with relatively high apparent resistivities compared to the intervening zone of saturation. However, even igneous and metamorphic crystalline rock with negligible intergranular porosity and conductivity can become highly conductive in zones where they are cracked and fissured and partially filled by fluids or clay.

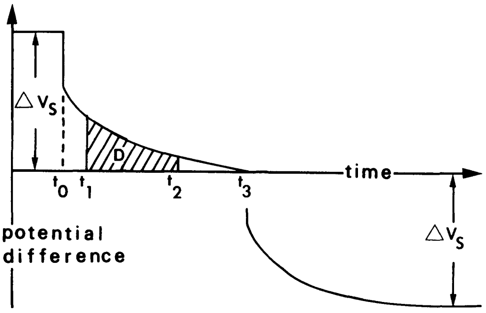

FIG. 15. The induced-polarization effect (after Kearey and Brooks, 1984). At time t_0 the current is turned off and the measured potential difference, after an initial drop from the steady state value ΔV_S, decays gradually to zero. A similar sequence occurs when the current is turned on again, but with opposite polarity to cancel electrode polarization effects. D represents the area under the decay curve for the time interval t_1–t_2.

4.3 Induced-Polarization Method

The induced-polarization (IP) method is principally used for locating disseminated sulfide ores of copper, silver, and other base metals. The method makes use of the capacitive action of the subsurface to locate zones where conductive minerals are disseminated within their host rocks.

During the course of standard resistivity surveying, it is commonly observed that the potential difference may take several seconds to several minutes to reach a stable peak value when the current is turned on, and when the current is turned off it takes a comparable amount of time for the potential difference to decay to zero, as shown in Fig. 15. This effect is called induced polarization and is particularly pronounced where electronically conducting minerals (e.g., metallic sulfides and oxides, graphite) or clay minerals are present in the subsurface. At the time scales involved, this phenomenon reflects the integrated effects of various mechanisms of interfacial polarization that occur whenever the subsurface involves electrically heterogeneous materials where charge concentrations build up at the boundaries of the heterogeneities. Mechanisms of interfacial polarization that are of particular concern to induced-polarization studies include electrode polarization or overvoltage and membrane polarization or electrolytic polarization.

Electrode polarization refers to the well-known effect of charge buildup that occurs at the interface of a conductor immersed in an electrolyte where a current is being driven through the system. This effect inhibits current flow and hence increases the apparent resistivity by an amount termed the *overvoltage* of the system. It can occur within the subsurface wherever an electronically conducting mineral grain boundary is in contact with an electrolytic pore fluid. When the current is applied, charge builds up slowly at these interfaces until saturation, where the potential difference measured by the array is maximum. When the current is turned off, the charge buildup at the interfaces slowly diffuses to cause a transitory decay of the potential difference measurements. This effect is particularly strong where the metallic minerals are individually dispersed within the subsurface so that the area of these interfaces is maximum.

Electrode polarization is the principal effect of interest in IP prospecting for metallic ores. However, other polarization effects may contribute to complicate the interpretation of IP surveys. In particular, membrane polarization effects can occur and may contribute overvoltages equivalent to those observed for a rock with up to 2% metallic minerals. Most subsurface geologic materials have a net negative charge at their interface with the pore fluid, and hence attract positive ions onto the surface. Where the diameters of the interconnecting pore throats are of the same order as the thickness of these charge buildups, negative and positive ions may build up around the blockages to inhibit current flow in the system. Membrane polarization is particularly pronounced in the presence of clay or silt particles that can block the flow of electrolyte-bearing fluid through pore spaces. In situations where membrane polarization is the primary effect, IP surveys can be interpreted for the position of the water table in the subsurface.

IP measurements are obtained as by-products of resistivity surveys, where the resistivity equipment has been upgraded to monitor the transitory behavior of potential differences over time intervals of 1–20 s or more. In practice, Schlumberger, dipole–dipole, and pole–dipole arrays are used, although in principle any standard configuration of electrodes may be employed to obtain IP measurements. IP observations are broadly classified into time-domain and frequency-domain measurements.

Time-domain IP observations involve measuring the transient behavior of potential differences as dc current is being galvanically driven into the ground from metal electrodes. The most common parameter determined is the area D beneath the decay curve over a time interval t_1–t_2 as shown in Fig. 15. Normalizing this area by the steady-state potential difference ΔV_s yields the chargeability M which is defined by

$$M=\frac{D}{\Delta V_s}=\frac{1}{\Delta V_s}\int_{t_1}^{t_2} V(t)\,dt \qquad (20)$$

in mV s/V or ms. Chargeability is commonly expressed in terms of the time interval t_1–t_2 and can be diagnostic of the presence of different minerals. For example, the chargeabilities over an interval of 1 s for pyrite and magnetite are 13.4 and 2.2 ms, respectively (Kearey and Brooks, 1984).

Frequency-domain IP observations are obtained by measuring apparent resistivities using ac current at two or more frequencies. To avoid electromagnetic induction effects, IP measurements are normally made at frequencies of 10 Hz or less where laboratory studies indicate that resistivity is essentially independent of the measuring-current frequency. They are obtained by measuring the apparent resistivity of the ground at two frequencies F and $f<F$, and expressed as the percentage frequency effect (PFE), which is defined by

$$\mathrm{PFE}=100[\rho_A(f)-\rho_A(F)]/\rho_A(F). \qquad (21)$$

Another measure of frequency-domain IP normalizes the PFE with respect to the dc or lower-frequency apparent resistivity. This measure is called the metal factor (MF), which is defined by

$$\mathrm{MF}=2\pi\times 10^5[\rho_A(f)-\rho_A(F)]/\rho_A(f)\rho_A(F), \qquad (22)$$

with units S/m.

The IP method is by far the most effective geophysical tool for locating low-grade ore deposits such as disseminated sulfides. These base-metal deposits have a strong IP effect, but they are nonconducting and hence are difficult to map by other electrical methods. IP data for the most part are interpreted on a qualitative basis, although analytical IP responses have been computed for idealized bodies such as spheres, ellipsoids, dikes, vertical contacts, and horizontal layers for more quantative interpretation.

An example of the results for a time-domain IP survey using a pole–dipole array across the Gortdrum copper–silver ore body in Ireland is shown in Fig. 16. A well-defined chargeability anomaly is observed over the ore body, although the deposit is low grade with less than 2% metallic conducting minerals. The corresponding apparent resistivity profile gives no indication of the presence of mineralization, reflecting rather the large resistivity contrast between the sandstone and dolomitic limestone with a minimum directly over the fault.

4.4 Electromagnetic Methods

With the exception of the magnetic method, the electromagnetic-induction techniques are the most widely used geophysical methods for mineral exploration. These methods are based on the induction of electric currents in buried conductive bodies by the magnetic components of electromagnetic waves generated at or near the Earth's surface. The source or primary waves originate from alternating currents generated at frequencies ranging from a few hertz to a few megahertz through loops of wire on the ground or in the air. The primary electromagnetic field attenuates as it is propagated into the Earth at a rate that depends on its frequency and the electrical properties of

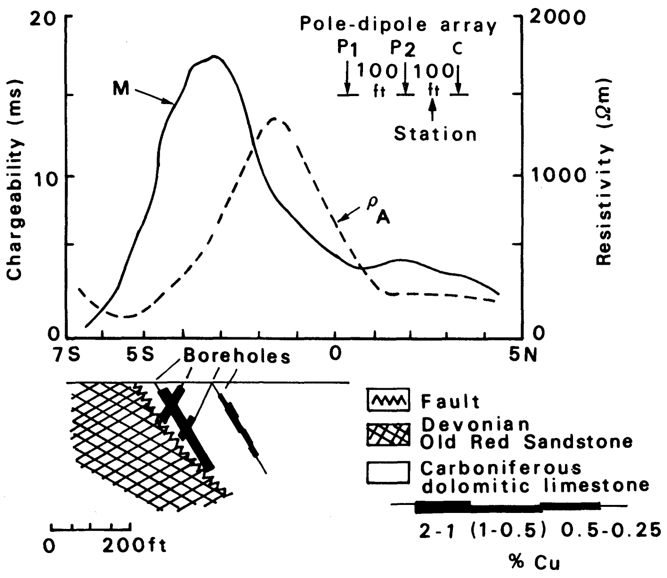

FIG. 16. Time-domain IP over Gortdrum copper-silver body (modified from Seigel, 1967, by Keary and Brooks, 1984).

the subsurface materials. The depth of penetration of the primary electromagnetic field is given by

$$Z_m = 503.8(f/\rho)^{-1/2}, \qquad (23)$$

where Z_m is in meters, the resistivity (ρ) of the Earth is in Ω m, and the frequency (f) of the primary field is in Hz. When the primary electromagnetic field waves pass through a conducting body, they induce eddy currents in the conductive materials (Fig. 17) These currents become the source of secondary electromagnetic waves that can be detected by suitable receiver coils at the Earth's surface again.

A variety of procedures are used to map anomalous electromagnetic-fields. A conventional and straightforward technique involves the use of a small receiver coil to map spatial variations of electromagnetic field inclination or dip. The current induced in the search coil by an electromagnetic wave is maximum or minimum when the plane of the coil is oriented respectively perpendicular or parallel to the direction of the electromagnetic wave. Over barren (nonconductive) ground, the plane of the search coil will be oriented perpendicular or parallel to the transmitter coil to indicate the dip properties of the primary electromagnetic wave. However, where a conductor exists within the subsurface, secondary electromagnetic signals are produced, which add vectorially to the primary electromagnetic signal to yield a resultant electromagnetic signal at the search-coil position. For good conductors the resultant electromagnetic signal can have inclination properties that are significantly different from the dip of the primary electromagnetic signal from the transmitter.

Commonly applied tilt-angle procedures include VLF and AFMAG methods, which are widely used because equipment requirements are simple and their implementation in the field is cost effective and rapid. For the very low-frequency (VLF) method, low-frequency electromagnetic waves generated by powerful radio transmitters used in long-range communications and navigational systems are implemented. Several stations around the world broadcast in the frequency range of 15–25 kHz signals that may be used for surveying

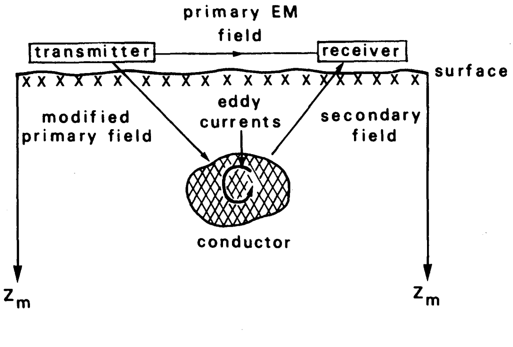

FIG. 17. General principle of electromagnetic surveying for subsurface conductors within the skin depth Z_m.

regions at distances of up to several thousand kilometers from the transmitter.

The primary electromagnetic field generated from the transmitter is essentially planar and horizontal at large distances. The magnetic component *H* lies at right angles to the direction of propagation in the horizontal plane with the electric component *E* lying in the vertical plane. A conductor that strikes in the direction of the transmitter optimally interferes with the magnetic field vector so that the induced eddy currents produce a maximum secondary electromagnetic field, whereas conductors striking at right angles to the direction of primary field propagation yield marginal to undetectable secondary electromagnetic field responses.

The audio-frequency magnetic-field (AFMAG) method utilizes the natural electromagnetic fields generated by thunderstorms, called "sferics," which propagate around the Earth between the ground surface and the ionosphere. This space constitutes an efficient electromagnetic waveguide with low attenuation so that thunderstorms anywhere in the world make an effective contribution to the field at any given point on the Earth. The field also penetrates the subsurface region, where, in the absence of electrically conducting bodies, it is essentially horizontal. Sferic sources are random so that the signal is broadband between 1 and 1000 Hz. This frequency range extends an order of magnitude lower than can be produced with conventional transmitters so that depths of investigation of several hundred meters are feasible.

Random variations in the direction and intensity of the primary field preclude the use of a single coil for determining the direction of the arriving field. Hence, the AFMAG receiver uses two orthogonal coils each inclined at 45° to the horizontal. In the absence of a secondary field, the components of the horizontal primary field perpendicular to the coils are equal so that their subtracted output will be zero. When a conductor is present, a secondary field that causes deflection of the resultant field from the horizontal occurs so that the combined output is no longer zero and is indicative of the tilt.

Tilt-angle methods such as VLF and AFMAG provide semiquantitative information on the horizontal distribution, depth, and dip orientation of conductive bodies within the subsurface. More quantitative information on the conductors can be obtained with em prospecting systems, which measure the phase and amplitude relationships between primary, secondary, and resultant electromagnetic fields. These sophisticated exploration systems commonly involve precisely located and oriented transmitters and receivers that are normally separated by distances of between 30 and 100 m. Over these distances, the primary field suffers minimal modification as it travels within both the air and the subsurface. Hence, a good estimate of the secondary wave may be obtained by subtracting the primary alternating wave from the resultant alternating wave field observed by the receiver. Modern instruments can separate the secondary electromagnetic wave into its real and imaginary components. Figure 18 generalizes the real and imaginary properties of the secondary signal induced in the conductor by primary sinusoidal fields.

The qualitative results presented in Fig. 18 suggest simple rules for characterizing conductors. A good conductor, for example, yields a large real component relative to a smaller imaginary component, whereas a poor conductor produces a relatively large imaginary component in contrast to a small real component. The ratio of the magnitudes of the real to imaginary components is commonly used to quantify the conductor, where a good conductor yields a ratio greater than unity and a poor conductor yields a ratio smaller than unity.

An example showing the results of a mobile transmitter–receiver em profile collected across a sheetlike conductor is given in Fig. 19. The conducting bodies produce negative anomalies in both the real and imaginary components, with maximum amplitudes immediately above the conductor as a consequence of

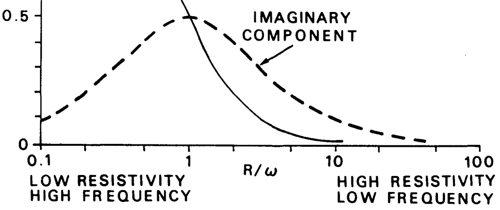

FIG. 18. Response of a single-turn conductor loop of resistance *R* to a primary sinusoidal field with angular frequency ω (after Parasnis, 1979).

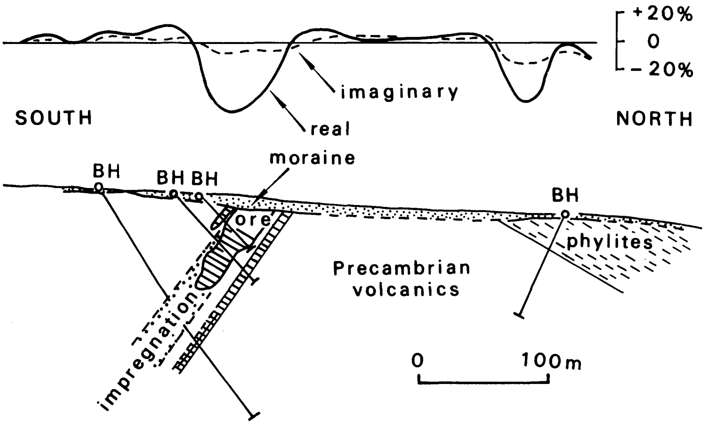

FIG. 19. Mobile transmitter-receiver profile using horizontal coplanar coils separated by 60 m and an operating frequency of 3.6 kHz in the Kankberg area of northern Sweden. Real and imaginary components are given in a percentage of the primary field. BH indicates borehole locations (modified from Parasnis, 1973, by Keary and Brooks, 1984).

the coplanar or horizontal coil system that was employed in this survey. The inclination of the body is suggested by the asymmetry of the anomalies, where the maximum anomaly gradient lies on the down-dip side. The presence of a very good conductor is indicated by the large ratio of the real to imaginary components over the ore body, whereas a smaller ratio is observed over the more resistant sequence of graphite-bearing phyllites to the north.

Mobile transmitter–receiver em surveys are conducted both on the ground and from airborne platforms such as fixed-wing aircraft and helicopters. Maximum penetration in ground surveys is limited to roughly 500 m, whereas for airborne work only about 50 m or less is common because of the weak electromagnetic coupling of the primary signal with the ground from a distant and moving airborne transmitter. Electromagnetic prospecting data are generally presented in the form of cross sections showing the amplitudes of the signals recorded by the receiver versus the horizontal position. Offset is expressed in terms of distance in ground surveys and in terms of time for airborne surveys. In general, the components of the secondary signal that are in phase (real) and out of phase (imaginary) with the primary signal are plotted separately and compared for interpretation.

Quantitative interpretation of electromagnetic data is commonly very complex because of the many factors that can influence the inductive response of the subsurface. Because of the complexity of theoretical calculations, laboratory modeling and simulations are often used to interpret em survey results. Simplified interpretation of mobile transmitter–receiver data may be accomplished for simplistic conductor geometries using master curves. However, much electromagnetic interpretation is only qualitative, especially for airborne data.

4.5 Telluric and Magnetotelluric Methods

Telluric currents flow in vast sheets parallel to the Earth's surface to depths of several kilometers. These currents flow over the whole Earth in a pattern of circular cells where each cell covers millions of square kilometers and remains essentially fixed with respect to the Sun. The source of telluric currents is ascribed to the inductive effects of magnetotelluric fields, which exist as standing waves in the space between the ionosphere and the Earth's surface. Magnetotelluric fields at the Earth's surface are essentially plane waves, with frequencies from about 10^{-5} Hz to at least the audio range, which penetrate the Earth's surface to produce the telluric currents. Ionospheric current systems are believed to be the source of magnetotelluric and telluric fields because periodic and transient fluctuations of the fields can be correlated with diurnal variations in the Earth's magnetic field resulting from solar emissions.

Telluric electric fields cause a mean potential gradient at the Earth's surface of about 10 mV/km and are constantly fluctuating in direction and magnitude at any point. However, systematic horizontal perturbations of telluric potential differences will be found at the Earth's surface when geologic structures disturb the horizontal flow of the telluric current sheets. These geological structures will represent lateral or horizontal resistivity contrasts within the subsurface such as, for example, high-resistivity salt domes or folded strata.

Telluric potential differences are measured using orthogonal pairs of electrodes made of a chemically inert substance like lead, which is nonpolarizing. The position of each station is marked by the coordinates where the line segments between the two electrode pairs bisect each other. The spacing of measurements varies from a few tens of meters to up to several hundred meters depending upon the depth of interest. The survey involves measurements collected with a mobile array, which are referenced to simultaneous readings at a spatially fixed base station. Base-station observations are required to account for temporal variations of the telluric currents.

Magnetotelluric surveying is an extension of the telluric method where one of the potential electrodes is replaced by a highly sensitive magnetometer to measure the associated magnetic field, which is in the picotesla range. The technique involves only mobile arrays (no base station is required) consisting of two electrodes to measure telluric potential differences in one direction and a sensitive magnetometer (e.g., fluxgate) to measure the magnetic field component in the orthogonal direction. In general, it can be shown that the amplitudes of orthogonal electric and magnetic field components are related according to

$$\rho_A = (0.2/f)(E/H)^2, \tag{24}$$

where f is in Hz, E is in mV/km, and H is in nT. This result shows that the apparent resistivity ρ_A varies inversely with frequency, which governs the depth of penetration of the magnetotelluric wave according to relations such as that described in Eq. (23). Hence, by breaking up the observations into their various frequency components and computing the apparent resistivities, it is possible to obtain insight on the vertical variations of resistivity within the subsurface region. Furthermore, these results can be compared at different points on the Earth's surface to establish large-scale horizontal variations of subsurface resistivity.

Magnetotelluric data are used mostly to investigate the large-scale vertical and horizontal resistivity variations of the crust and upper mantle. Interpretation of magnetotelluric data is facilitated by master curves of apparent resistivity versus frequency that are available for simplified features such as two and three horizontal layers, vertical contacts, and dikes. Numerical modeling and inversion of more detailed two-dimensional structures is also possible.

5. APPLICATIONS OF GEOPHYSICS TO LITHOSPHERIC INVESTIGATIONS

Solid-Earth geophysical methods were largely developed from basic physics principles and the investigations of Earth materials. Although the methods were used on an occasional basis for a century or more, the growth of the applications of geophysics really began in the years following World War I and grew rapidly, as did all of science, following World War II. We have seen an exponential growth in recent decades because of societal pressure to solve Earth-related problems such as mineral-resource depletion, natural-hazard mitigation, and environmental restoration, as well as the availability of high-speed computers to control instrumental observations, process numerical data, and present results of computations in graphic form. The following examples of the application of geophysics to lithospheric investigations are selected to illustrate the range of impact of geophysical methods on science and society over a range of scales.

5.1 Mantle and Crustal Processes

Problems dealing with the origin of major Earth land forms—such as continents, oceans, and mountains—have been the subject of intensive inquiry and speculation by geoscientists for many years. Unfortunately, these studies were based on a minimum of information derived primarily from surface geological observations. As a result, no hypothesis survived close scrutiny. This situation changed dramatically and quite quickly as a result of the confluence of geophysical observations following World War II.

Measurement of the spatial variation of the geomagnetic field over oceanic regions showed a pattern of linear magnetic anomalies (Raff and Mason, 1961) that were soon recognized to be symmetrical to midocean ridges that were identified by bathymetric observations. As subsequent studies have proven, Vine and Matthews (1963) correctly identified these lineations as the result of seafloor spreading from the oceanic ridges that cause magma to upwell into the crack or rift along the crest of the ridge (Fig. 20). As the spreading continues over millions of years, the cooling magma acquires a permanent magnetism in the direction of the geomagnetic field. Independent geophysical research has shown that the geomagnetic field reverses polarity occasionally. The reversals are frozen into the newly formed crust, causing the highs and lows in the magnetic field that are observed as magnetic lineations symmetric with oceanic ridges. By dating the reversals from continental rocks, it is possible to date the lineations in the oceans and thus to determine that the velocity of seafloor spreading is millimeters to centimeters per year.

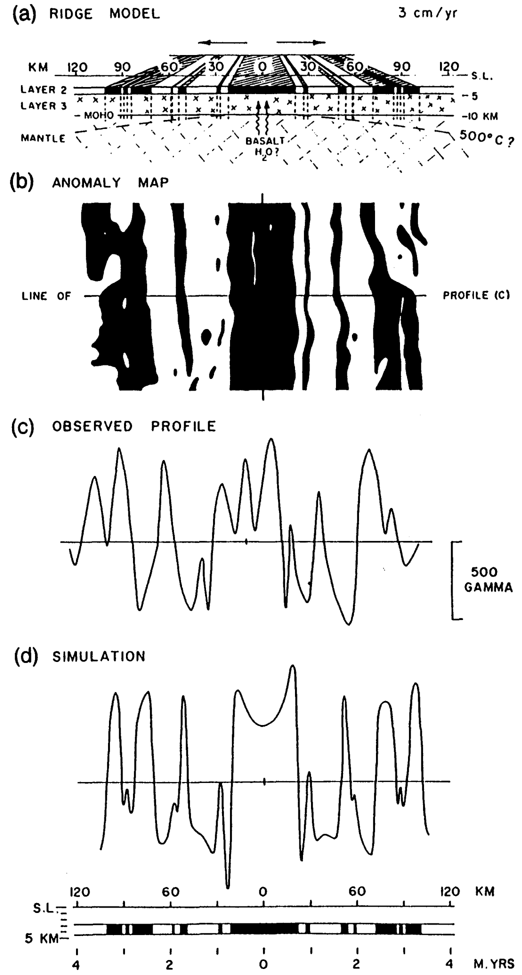

FIG. 20. The principle of seafloor spreading summarized by Vine (1968). (a) Schematic representation of the crustal model, applied to the Juan de Fuca Ridge, southwest of Vancouver Island. Shaded material in layer 2, normally magnetized; unshaded, reversely magnetized. (b) Part of the summary map of magnetic anomalies recorded over the Juan de Fuca Ridge. Black, areas of positive anomalies; white, areas of negative anomalies. (c) A total-field magnetic anomaly profile along the line indicated in (b). (d) A computed magnetic profile assuming the model and reversal time scale.

Seafloor spreading, as derived from magnetic observations in the oceans, explained the youthful nature of the oceans (< 200 Ma) as compared to the continents ($\simeq$ 4000 Ma), but conservation of space required destruction or compression of older crust if new crust was being generated in the oceans. This led to the concept of plate tectonics wherein the surface of the Earth is divided into about 20 plates that are defined by lithospheric earthquake epicenters (Morgan, 1968). Along boundaries between the plates, one of the following processes is going on:

1. crust is produced at oceanic ridges;
2. crust is destroyed by being dragged into the Earth at island arc trenches, and/or crumpled into mountain ranges on the edges of continents causing the ranges to grow; or
3. crustal plates are simply moved past each other with minimal modification and thus the continents drift over the Earth's surface.

The concepts of seafloor spreading and plate tectonics, which are accepted in the geoscience world as facts today, have revolutionized the entire geoscience discipline by providing a unifying concept for most geological processes. Its effect upon science is likened to the impact of the Copernican and atomic theories upon our understanding of the universe. Sophisticated geophysical observations and interpretations over the past quarter of a century have continued to support the initial theories and have done much to refine them, but their origin is in simple geophysical observations of the oceanic geomagnetic field, the distribution of earthquakes over the Earth, and the recognition of the reversal of the polarity of the geomagnetic field through geological time.

5.2 Crustal Studies

The continental crust is a complex mosaic of rock units of varying composition that reflect the multitude of processes that make up the geologic history of the Earth over roughly the past 4600 Ma. Surface geologic studies in the Lake Superior region of mid-North America have identified a complex geologic structure consisting of rock units of contrasting physical properties that originated 1100 Ma ago when the Earth's crust was split over a length in excess of 2000 km (Van Schmus and Hinze, 1985). This structure, the Midcontinent Rift System, is poorly known because over most of its length it is covered by younger rock strata, and even in the Lake Superior region where it crops out, the vertical range of our direct observations is extremely limited. Recent seismic reflection studies on the south shore of Lake Superior in western Michigan have mapped the strata of the rift system to depths of the order of 18 km. Hinze *et al.* (1990) have used a seismic reflection profile together with other geophysical data and surface geologic information to map the distribution of the rock formations to midcrustal levels. Knowledge of this structure makes it possible to suggest the sequence of geologic events involved in the rift and thus to interpret the geologic history of this region over a billion years ago.

Figure 21 shows a small segment of the seismic reflection profile with time of arrival of seismic waves at the surface plotted on the vertical scale as a function of distance along the profile. The coherent (continuous) events are identified as reflections from interfaces between rock strata of differing physical properties. The lack of clearly identifiable reflections within the first few seconds in the central part of this figure where the Portage Lake volcanic rocks crop out is interpreted not as a lack of reflectors, but rather as a result of their steep dip that prevents the reflections from being imaged at the surface. The reflectors of this figure and similar reflections along the profile have been used to chart the distribution of the layered strata as shown in Fig. 22. This figure also presents profiles of the anomalous geomagnetic and gravity fields. Variations in these profiles reflect, respectively, differing magnetization and mass (density) of the underlying rock formations. These are useful in identifying the rock strata imaged by the seismic method and assist in removing ambiguity in the interpretation.

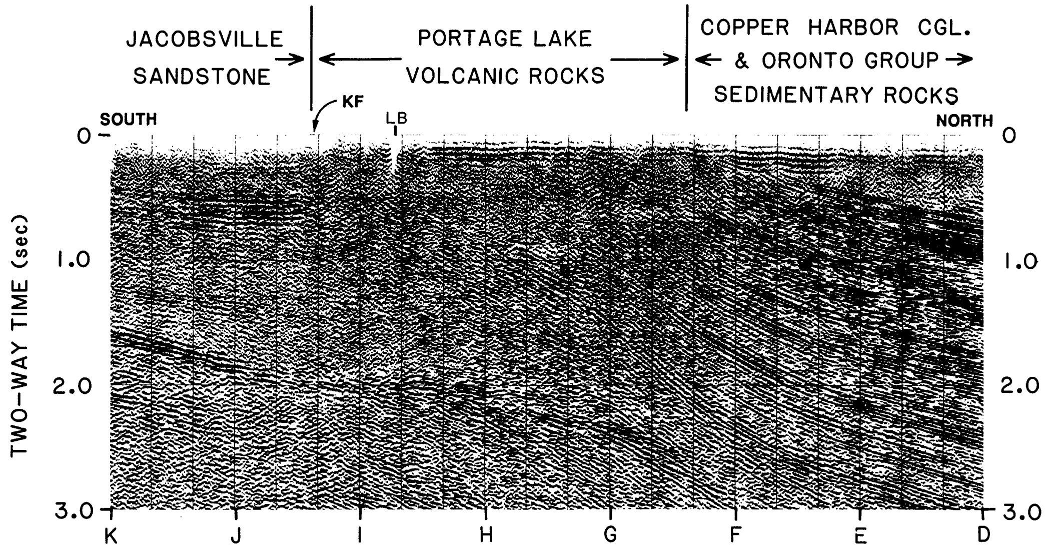

FIG. 21. Portion of migrated seismic reflection record centered on outcrop of Portage Lake volcanic rocks on the south shore of Lake Superior in western Michigan. Interpretation is presented in Fig. 22. No vertical exaggeration; 1 s of two-way time is about 3 km. KF is the observed position of the Keweenaw fault (Fig. 22). LB is position of line break in seismic profile (after Hinze et al., 1990).

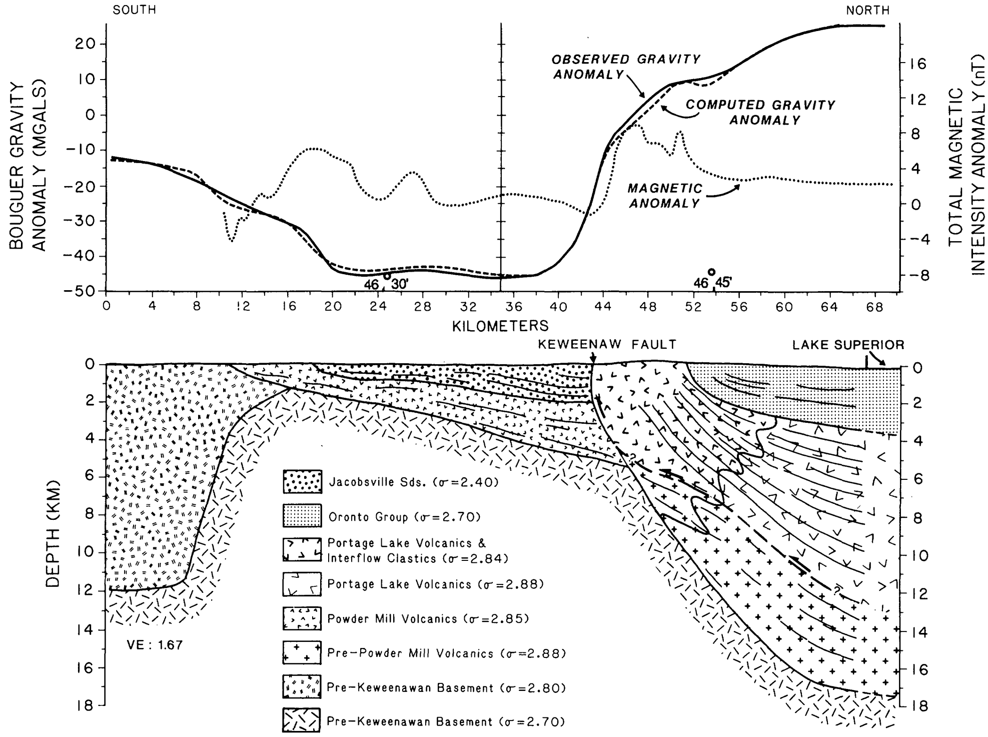
SOUTH
NORTH
BOUGUER GRAVITY ANOMALY (MGALS)
20
10
0
-10
-20
-30
-40
-50
OBSERVED GRAVITY ANOMALY
COMPUTED GRAVITY ANOMALY
MAGNETIC ANOMALY
46°30'
46°45'
TOTAL MAGNETIC INTENSITY ANOMALY (nT)
16
12
8
4
0
-4
-8
0
4
8
12
16
20
24
28
32
36
40
44
48
52
56
60
64
68
KILOMETERS
KEWEENAW FAULT
LAKE SUPERIOR
DEPTH (KM)
0
2
4
6
8
10
12
14
16
18
VE : 1.67
Jacobsville Sds. (σ=2.40)
Oronto Group (σ=2.70)
Portage Lake Volcanics & Interflow Clastics (σ=2.84)
Portage Lake Volcanics (σ=2.88)
Powder Mill Volcanics (σ=2.85)
Pre-Powder Mill Volcanics (σ=2.88)
Pre-Keweenawan Basement (σ=2.80)
Pre-Keweenawan Basement (σ=2.70)

FIG. 22. Interpretive cross section across the southern margin of the Lake Superior basin in western Michigan. Thin lines are schematic representation of principal seismic reflections. Densities of geologic units are in g/cm^3. Geologic column is from oldest to youngest from bottom to top (after Hinze *et al.*, 1990).

In Fig. 22, the positive anomalous areas in the vicinity of 16 and 48 km reflect the strong magnetization of the underlying volcanic rocks of the Midcontinent Rift System that were originally laid down horizontally, but subsequently, during the rifting process, have been structurally disturbed. The positive magnetic anomalies are critical to locating the magnetic volcanic rock. The relative negative gravity anomaly from 20 to 40 km and the strong positive gravity anomaly to the north of the 40-km position results from variations in the density and volume of the underlying rock materials. The negative anomaly from 20 to 40 km is primarily caused by the presence of low-density Jacobsville sandstone that reaches depths of up to 2 km. In contrast, the positive gravity anomaly is caused by an immense volume of relatively high-density volcanic rocks extending to depths of 18 km on this section. Confirmation of the feasibility of this interpretation is provided by the forward modeling of the gravity effect of the profile that is shown on the figure to correspond closely with the observed gravity anomaly. Furthermore, the abruptness of the change in rock type between the negative and positive gravity anomalies and the attitude of the change confirm the presence of a fault. Evaluation of the fault shows that it was caused by compressional forces, quite in contrast to the extensional stress regime required for the rifting process. By considering the types of rocks and their ages, it is thus possible to deduce that the rifting event was terminated by a change in the stress field of the Earth that caused compression of the rift structure.

This is just one small segment of a very large crustal structure that continues to be under intensive geophysical investigation. The results of these studies are not only important in deciphering geologic history, but also are useful in identifying potential earthquake hazards due to reactivation of the associated ancient zones of weakness and in localizing areas for exploration for the copper minerals that are found within the rift rocks.

5.3 Hydrocarbon Exploration

The major expenditures in solid-Earth geophysics are in the search for petroleum—oil and gas—that occur within geologic traps or reservoirs in sedimentary rocks of the continents and their shelves. Numerous requirements are necessary for the accumulation of commercial petroleum deposits, but primary among these are rocks that contain interconnected pore space and include a geological condition that will entrap the lighter-than-water hydrocarbons to prevent them from escaping to the surface and dispersing into the environment. All geophysical methods have a role in exploring sedimentary basins for these conditions. However, in excess of 95% of all petroleum geophysical expenditures are directed toward reflection seismic investigations. This method is capable of providing a subsurface image of sedimentary basins that reveals structural features such as folds, faults, domes, and unconformities. Furthermore, in some cases the seismic method can give insight into lithology, porosity, and hydrocarbon content. The resolution of this method far surpasses other geophysical techniques and thus is particularly important in pinpointing locations that are favorable for drilling in the search for hydrocarbons. The use of the seismic exploration method together with geological information is the key to decreasing the risk of wildcat drilling and providing the energy resources for the world's economy.

The success of seismic reflection in petroleum exploration is a function of increasing the signal-to-noise ratio of the received seismic pulses so that reflections from interfaces can be identified and mapped. This requires favorable geologic conditions and the application of advanced signal processing techniques. Under these conditions, it is possible not only to image the lithologic interfaces, but also to identify contacts between elements of an oil–gas pool. This is illustrated in Fig. 23. The upper portion of the figure shows an idealized anticlinal hydrocarbon deposit with a model seismic profile above it. The water (brine)–gas, the gas–oil, and the oil–water contacts are all apparent in the model. The lower portion of Fig. 23 is an actual seismic profile of a similar geological situation. The darkened coherent events in the actual seismic profile, referred to as "bright spots" by petroleum seismologists, reflect the strong acoustic impedance contrast between the gas-charged rocks and their enclosing materials. Thus, under proper conditions, it is feasible to use the seismic method as a direct hydrocarbon indicator.

Much of the world's petroleum exploration today occurs in the continental shelves. Figure 24 is a seismic reflection profile off the Grand Banks of Newfoundland with a geological interpretation of the data superimposed. The vertical line P-15 is the location of the discovery well of the Hibernia oilfield. The productive capability of this well was estimated to be in excess of 20 000 barrels per day, and the oilfield is estimated to contain in excess of 1 billion barrels of oil (Arthur *et al.*, 1982).

5.4 Mineral Exploration

As is the case for hydrocarbon exploration, mineral exploration may involve any of the solid-earth geophysical exploration methods. However, there are, in general, major differences between hydrocarbon and mineral exploration—in mineral exploration the targets are smaller and shallower, the geology is more complex and varied, and the financial resources are usually considerably less. As a result, the exploration methods based on potential-field theory—magnetic, electrical, gravitational—are most widely employed, and they tend to focus on direct detection of the mineral deposits on the basis of the physical property difference between the deposits and the barren enclosing rock formations.

In mineral exploration, it is often necessary to cover extensive regions of the continents quickly and inexpensively. As a result, many of the mineral exploration methods have been adapted to airborne measurements. The most widely used and versatile of the methods are airborne magnetic and electromagnetic methods. These methods are respectively most appropriately used for identifying and mapping strongly magnetic deposits that contain iron oxide and magnetite, and ores of high electrical conductivity consisting prima-

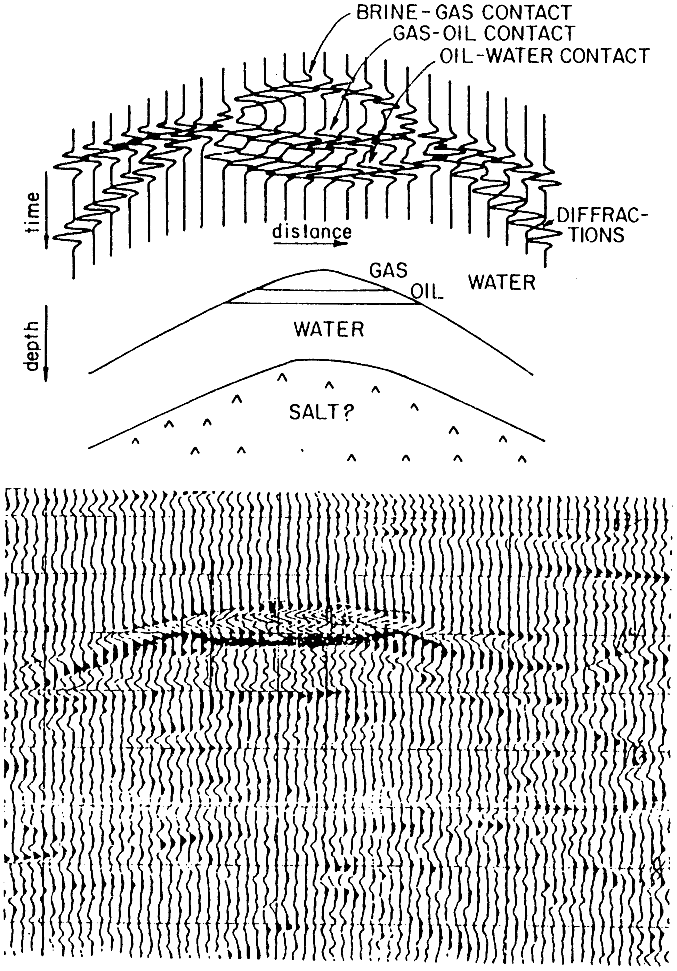

FIG. 23. (Top) Geologic and seismic reflection model of a hydrocarbon reservoir. (Bottom) Seismic reflection profile of a hydrocarbon reservoir that is an analog of model in top figure (modified from Wood, 1982).

rily of metallic sulfides. An illustration of the use of these methods in mapping diamond-bearing kimberlite pipes in Tanzania is shown in Fig. 25. On the left are shown the contours of equal magnetic intensity. Two of the known kimberlites in the region are located by intense magnetic anomalies. On the right of the figure the apparent conductivity of the surface material, as determined by an airborne electromagnetic survey, is shown for the same region as the magnetic map. All three kimberlite pipes are associated with high electrical conductivity. In this case the anomalous conductivity is not due to metallic sulfides, but rather to a clay-rich layer that forms above the kimberlite from weathering of the pipes (Palacky, 1990). The clay layer is more conducting than the weathered material from the adjacent barren rock.

5.5 Ground-Water Investigations

Ground water is a critical resource for mankind in many areas of the Earth. Although this is true in many humid regions that do not have ready access to surface waters for human consumption, irrigation, and industrial uses, it is particularly true in arid regions where ground water is the only source of water for much of the year. Thus the location of water in the subsurface that is readily available and potable is a major source of concern. In order to enhance the efficiency of the ground-water exploration process, wherein the number of wells is minimized, geophysical exploration techniques are commonly applied to the problem. Essentially all geophysical methods adapted to the shallow subsurface are used in ground-water investigations. Some methods, like the commonly used electrical

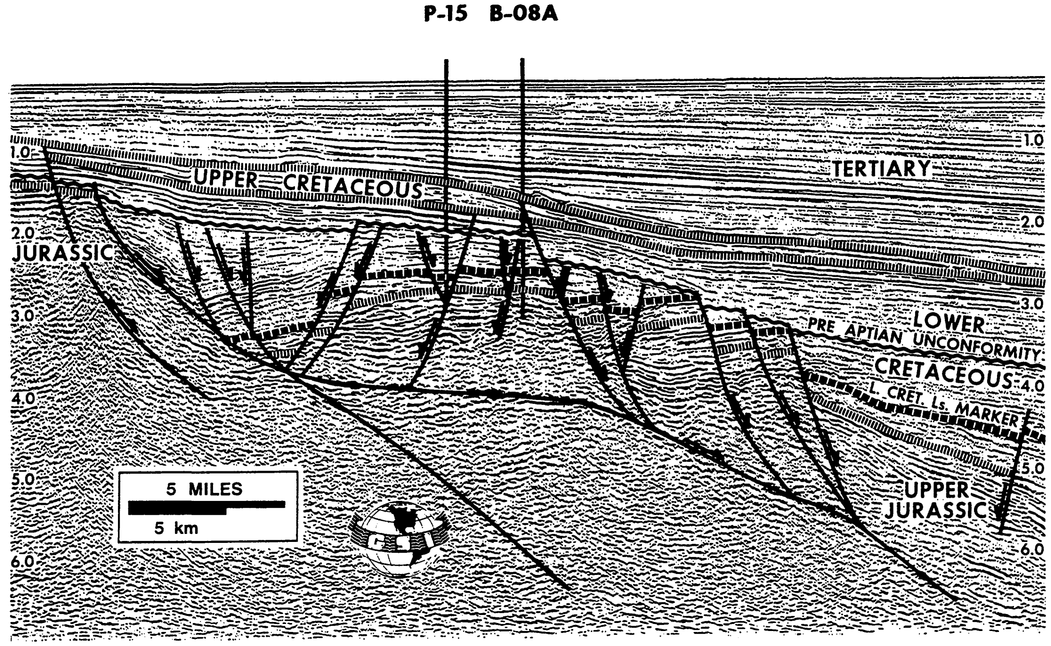

FIG. 24. Migrated seismic profile off the Grand Banks of Newfoundland with interpreted geology. P-15 and B-08A are nearby offshore drill holes. (Line provided courtesy of Geophysical Service, Inc., after Arthur *et al.*, 1982).

resistivity and electromagnetic techniques, are generally used in a direct manner, searching out the modification of the electrical properties caused by the presence of water and water of varying quality. In contrast, other methods, such as the seismic and gravity methods, are used to determine the stratigraphy and structure of the water-saturated Earth materials and thus indirectly the most favorable subsurface regions for ground-water production. These methods provide no direct information on water quality.

In recent years, geophysical methods have had an increasing role in delimiting the extent of pollution of ground-water resources. Generally, this pollution resulting from human activities is detected by electrical methods that measure the decrease in electrical resistivity brought about by the increasing ion content of polluted waters. Such is the case of salt-water intrusion near the Cross-Florida Barge Canal that was initiated to provide a shortened, navigable waterway across northern Florida between the Atlantic Ocean and the Gulf of Mexico. The sea-level canal was cut several meters into the surface carbonate rocks that comprise the Floridan aquifer (Hagemeyer and Stewart, 1990). Surface resistivity measurements (Fig. 26) indicate that low resistivity values are found in the subsurface along the course of the canal, suggesting that sea-water intrusion by the waters of the canal is polluting the aquifer. The degree and extent of the salt-water intrusion can be estimated from the values of resistivity if related geological and ground-water chemical data are available.

5.6 Environmental and Engineering Investigations

Geophysical methods have for several decades had an important role in determining various physical characteristics of the subsurface that are required for engineering designs. In recent years, with the accelerating concern for the environment and especially for maintaining the quality of ground water, there has been an increasing use of geophysical techniques. Geophysical techniques are usually much more rapid and less expensive than direct measurements in drill holes, but they have neither the sensitivity nor the resolution of direct subsurface measurements. Nonetheless, geophysical methods are often employed

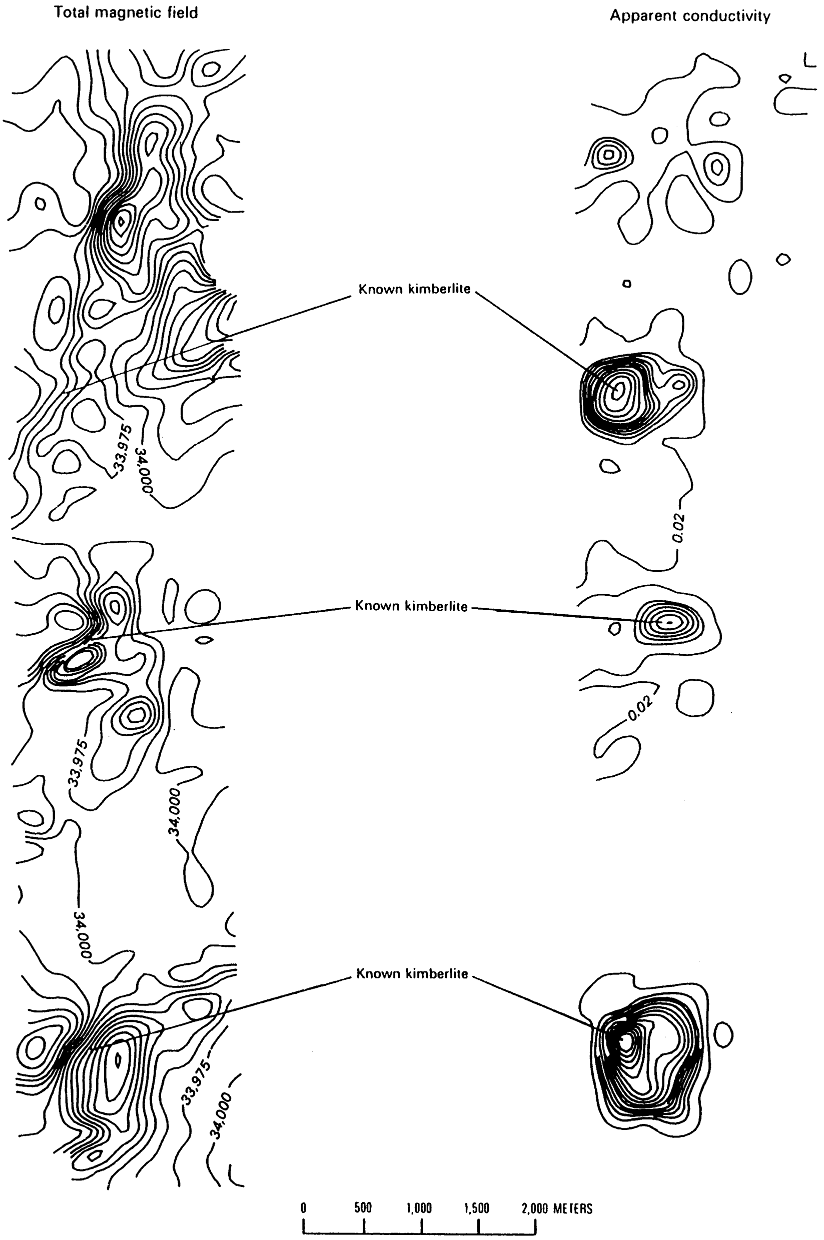
Total magnetic field
Apparent conductivity
Known kimberlite
Known kimberlite
Known kimberlite
33,975
34,000
33,975
34,000
34,000
33,975
34,000
0.02
0.02
0 500 1,000 1,500 2,000 METERS

FIG. 25. Total magnetic field and apparent electrical conductivity showing the location of three known kimberlites in Tanzania. Contour interval 25 nT and 0.02 S/m, respectively (modified from Johnson and Siegel, 1986, by Palacky, 1990).

in environmental and engineering investigations to aid in selecting optimum sites for drilling and in many cases to avoid the potential pollution hazards related to direct investigations. As in ground-water investigations, a wide range of methods is employed in environmental and engineering studies; the best method for a particular problem is often decided upon by relative costs rather than by a desire to maximize information from the study.

An illustration of the efficient use of a geophysical method to solve an environmental problem is shown in Fig. 27. A landfill

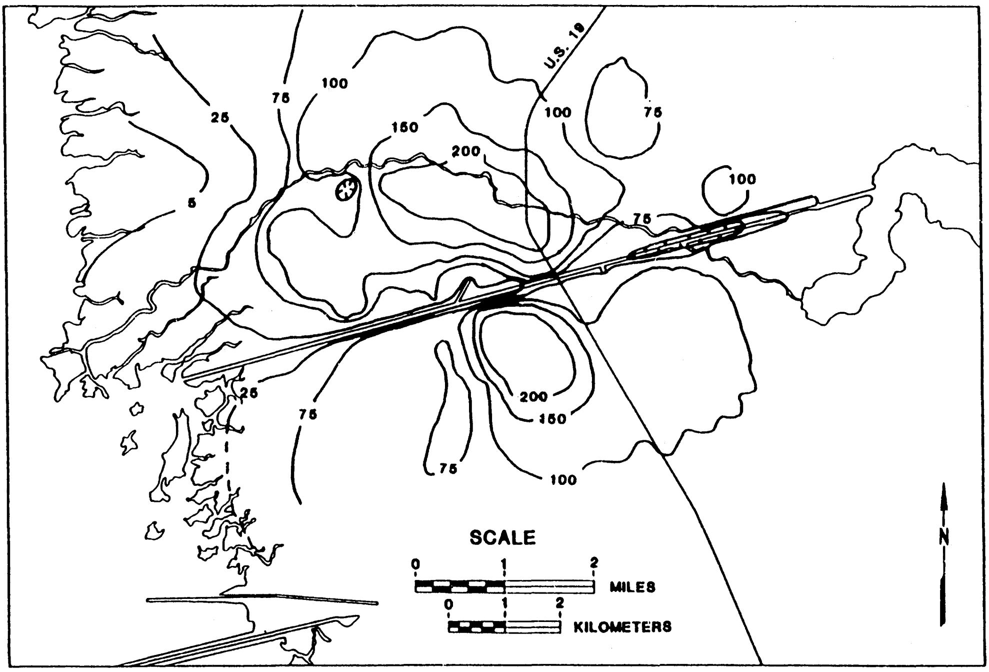

FIG. 26. Contour map of interpreted resistivity in Ω m at 30 m below land surface from vertical resistivity soundings in the vicinity of the Cross-Florida Barge Canal, northern Florida (after Hagemeyer and Stewart, 1990).

approximately 60×150 m^2 located in northwest Indiana (Roberts *et al.*, 1990) was the subject of an intensive investigation to determine its potential impact upon the groundwater resources of the region. One of the required site characteristics is the configuration of the bottom of the landfill, which was the original land surface of a ravine that was filled during the dumping process. Figure 27 shows the residual gravity-anomaly contour map of the landfill. The negative anomaly supports the assumption that the landfill is less dense than the enclosing glacial till. Figure 28 is an isopach map of the landfill derived by subtracting the prelandfill surface from the present surface landform. Comparison of these two figures illustrates how well the gravity method can be used to predict the thickness of the landfill. Furthermore, with constraints on the dimensions of the landfill from collateral data, interpreted density variations within the landfill can be used to characterize the nature of the solid waste material. In the case of the mapped landfill, Roberts *et al.* (1990) discriminated regions within the landfill composed of primarily domestic-type trash from lower-density construction and vegetation refuse.

6. SUMMARY

Solid-Earth geophysics has grown tremendously in acceptance, utilization, and effectiveness over the past half century. This science, which is thoroughly grounded in classical physics, has thrived because of improvements in instrumentation and numerical processing, but has been successful because of the effective integration of geology and physics. The problems solved by solid-Earth geophysics are largely geological; thus geological information and principles are necessary constraints in the application of the science. This is especially true because of the ambiguity inherent in the interpretation of solid-Earth geophysical data. Numerous geophysical techniques have been developed to sense remotely the variety of physical properties of Earth materials. They all have their unique uses, but solid-Earth geophysics is most successful where

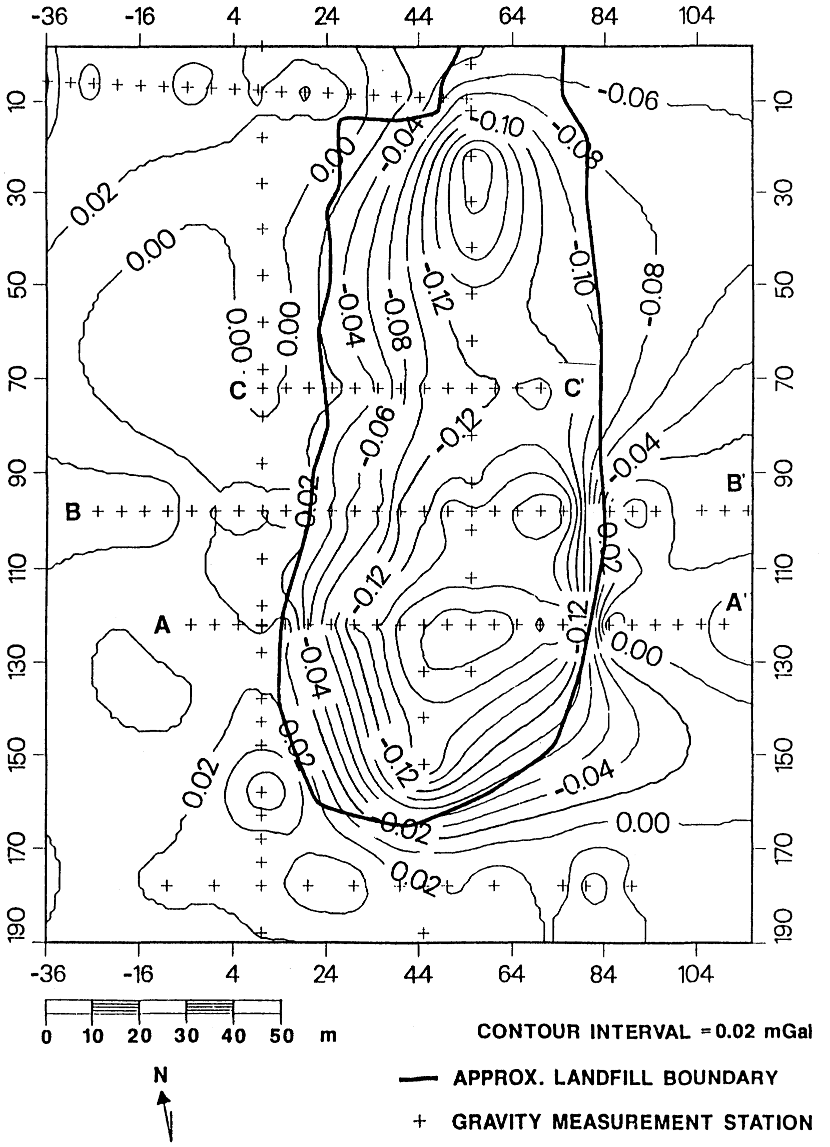
-36
-16
4
24
44
64
84
104
10
30
50
70
90
110
130
150
170
190
-0.06
-0.10
-0.08
0.02
0.00
-0.04
-0.12
C
C'
B
B'
A
A'
0 10 20 30 40 50 m
N
CONTOUR INTERVAL = 0.02 mGal
APPROX. LANDFILL BOUNDARY
+ GRAVITY MEASUREMENT STATION

FIG. 27. Residual gravity anomaly contour map of a landfill site in northwest Indiana (after Roberts *et al.*, 1990).

multiple methods, as well as geology, are synergistically combined.

Solid-Earth geophysical studies have been unusually successful in that they provided the basic building blocks for the new paradigm of geoscience, plate tectonics and seafloor spreading, which have made it possible to establish a unifying concept for most Earth processes. However, solid-Earth geophysics as applied to studies of the lithosphere have far-ranging applications beyond that of understanding the basic nature, composition, and structure of this Earth shell. Geophysics has an important role in solving numerous societal problems such as contamination of the subsurface environment, maintaining adequate Earth resources to sustain the world's quality of life, and predicting earthquakes and volcanic activity. With the stress placed upon the Earth by a booming world population and the desire

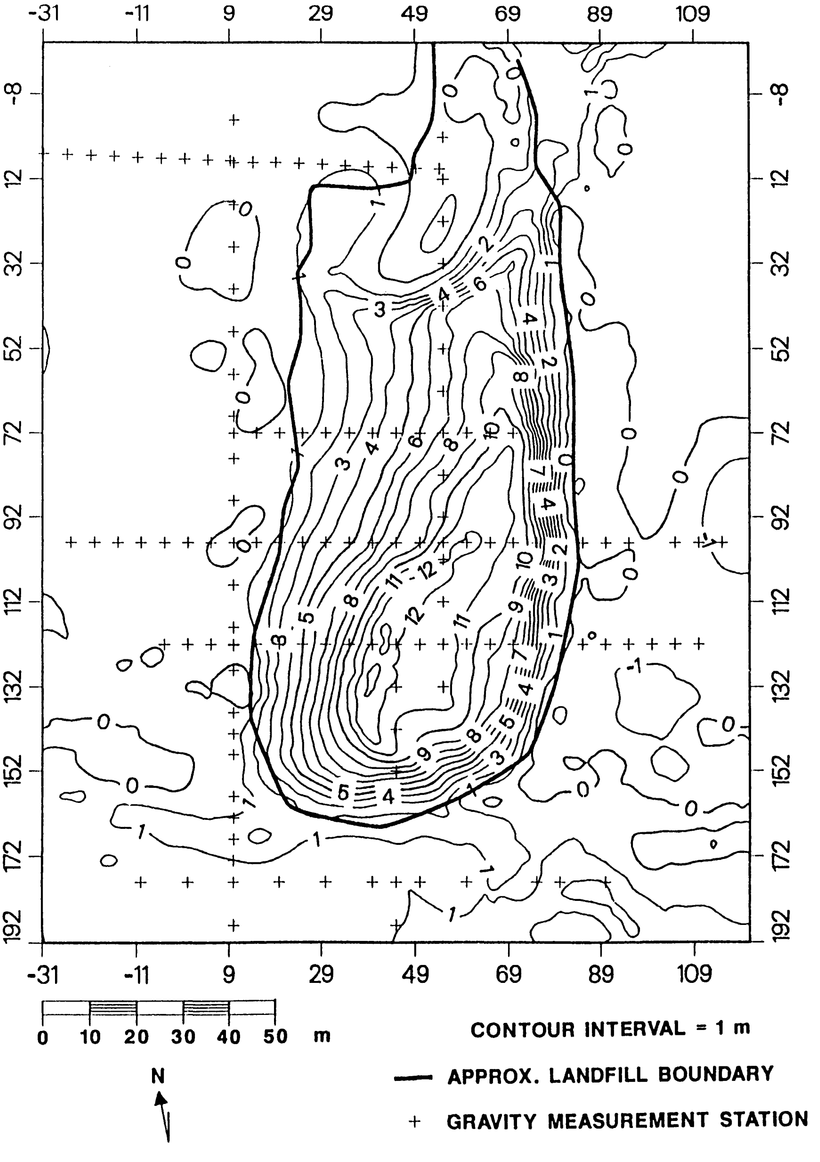
-31
-11
9
29
49
69
89
109
-8
12
32
52
72
92
112
132
152
172
192
0 10 20 30 40 50 m
CONTOUR INTERVAL = 1 m
N
APPROX. LANDFILL BOUNDARY
+ GRAVITY MEASUREMENT STATION

FIG. 28. Isopach map of the landfill site in northwest Indiana shown in Fig. 27 (after Roberts *et al.*, 1990).

to enhance the quality of life in all regions of the Earth, we can anticipate that solid-Earth geophysics will need to expand its horizons and improve its effectiveness.

GLOSSARY

Aliasing: A frequency ambiguity introduced in the representation of digital data where the sampling rate involves fewer than two samples per cycle. In this case, an input signal at one frequency yields the same sample values as a lower frequency at the output of the system. The lower frequency is an alias of the input frequency and is completely erroneous.

Curie Point: The temperature at which a material loses its magnetization.

Ma: Million years (of geologic age).

Offset: The distance between the point of disturbance (source point) and the position of a detector or the center of an array of detectors.

resonances, so that the precession signal output changes continuously with the ambient magnetic field. In addition to continuous precession signal, the method increases the signal strength and signal-to-noise ratio.

Spherical Divergence: The decrease in signal strength (energy per unit area of wave front) with distance as the wave spreads out.

Zero-Length Spring: A spring whose effective length, as referred to a fixed point of support, is zero when the external forces acting upon it are zero.

List of Works Cited

Arthur, K. R., Cole, D. R., Henderson, G. G. L., Kushnir, D. W. (1982), in: M. T. Halbouty (Ed.), *The Deliberate Search for the Subtle Trap*, American Association of Petroleum Geologists Memoirs Vol. 32, Tulsa: Society of Exploration Geophyicists, pp. 181–196.

Breiner, S. (1973), *Applications Manual for Portable Magnetometers*, Sunnyvale, CA: Geometrics.

Hagemeyer, R. T., Stewart, M. (1990), in: S. H. Ward (Ed.), *Geotechnical and Environmental Geophysics*, Vol. 2, Tulsa: Society of Exploration Geophysicists, pp. 67–77.

Hinze, W. J. (1990), in: S. H. Ward (Ed.), *Geotechnical and Environmental Geophysics*, Vol. 2, Tulsa: Society of Exploration Geophysicists, pp. 75–126.

Hinze, W. J., Hildenbrand, T. G. (1988), *Seis. Res. Lett.* **59**, 289–297.

Hinze, W. J., Braile, L. W., Chandler, V. W. (1990), *Tectonics* **9**, 303–310.

Hood, P. J., Holyroyd, M. T., McGrath, P. H. (1979), in: P. H. Hood (Ed.), *Geophysics and Geochemistry in the Search for Metallic Ores*, Economic Geology Report No. 31, Ottawa: Geological Survey of Canada, pp. 77–104.

Johnson, T. M., Seigel, H. O. (1986), in G. J. Palacky (Ed.), *Airborne Resistivity Mapping*, Ottawa: Geological Survey of Canada, Paper No. 86-22, pp. 125–129.

Keary, P., Brooks, M. (1984), *An Introduction to Geophysical Exploration*, Oxford: Blackwell.

LaCoste, L. J. B., Jr. (1934), *Physics* **5**, 178–180.

Morgan, W. J. (1968), *J. Geophys. Res.* **73**, 1959–1982.

Nettleton, L. L. (1940), *Geophysical Prospecting for Oil*, New York: McGraw-Hill.

Palacky, G. J. (1990), in: D. V. Fitterman (Ed.), *Developments and Applications of Modern Airborne Electromagnetic Surveys*, U.S. Geological Survey Bulletin No. 1925, Denver: U. S. Geological Survey, pp. 89–101.

Parasnis, D. S. (1973), *Mining Geophysics*, Amsterdam: Elsevier.

Parasnis, D. S. (1979), *Principles of Applied Geophysics*, London: Chapman & Hall.

Raff, A. D., Mason, R. G. (1961), *Bull. Geol. Soc. Am.* **72**, 1267–1270.

Reid, A. B. (1980), *Geophysics* **45**, 973–976.

Roberts, R. L., Hinze, W. L., Leap, D. I. (1990), in: S. H. Ward (Ed.), *Geotechnical and Environmental Geophysics*, Vol. 2, Tulsa: Society of Exploration Geophysicists, pp. 253–259.

Sato, M., Mooney, H. H. (1960), *Geophysics* **25**, 226–249.

Seigel, H. O. (1967), in: L. W. Morley (Ed.), *Mining and Groundwater Geophysics*, Economic Geology Report No. 26, Ottawa: Geological Survey of Canada, pp. 123–137.

Van Schmus, W. R., Hinze, W. J. (1985), *Annu. Rev. Earth Planet. Sci.* **13**, 345–383.

Vine, F. J. (1968), in R. A. Phinney (Ed.), *History of the Earth's Crust*, Princeton, NJ: Princeton Univ. Press, pp. 73–89.

Vine, F. J., Matthews, D. H. (1963), *Nature* **199**, 947–949.

Wood, L. C. (1982), in: K. C. Jain, R. J. P. deFigueiredo (Eds), *Concepts and Techniques in Oil and Gas Exploration*, Tulsa: Society of Exploration Geophysicists, pp. 45–90.

Further Reading

Dobrin, M. B., Savit, C. H. (1988), *Introduction to Geophysical Prospecting*, New York: McGraw-Hill. An introductory text on geophysical methods.

Keary, P., Brooks, M. (1984), *An Introduction to Geophysical Exploration*, Amsterdam: Elsevier. A very readable introductory account of the geophysical methods.

Keary, P., Vine, F. J. (1990), *Global Tectonics*, Oxford: Blackwell. A readable introduction to crustal and upper-mantle geophysics.

Parasnis, D. S. (1979), *Principles of Applied Geophysics*, London: Chapman & Hall. Introduction to geophysics that tends to emphasize mining applications.

Telford, W. M., Geldard, L. P., Sheriff, R. E. (1990), *Applied Geophysics*, Cambridge, U.K.: Cambridge University Press. A more detailed handbook of geophysics.

GEOTHERMAL ENERGY

JOHN C. ROWLEY, *Los Alamos National Laboratory, Los Alamos, New Mexico, U.S.A.*

Introduction 197
1. **History** 198
2. **Resource Types and Assessments** 201
2.1 Geothermal Resource Types 201
2.1.1 Earth Energy 203
2.1.2 Normal-Gradient Aquifers 203
2.1.3 Hot and Warm Aquifers 203
2.1.4 Hydrothermal Convection Resources 204
2.1.5 Geopressured Resources 204
2.1.6 Hot Dry Rock Resources 205
2.1.7 Magma Resource 205
2.2 Resource Assessments 205
3. **Geology and Hydrology of Geothermal Resources** 206
3.1 Nature and Occurrence 207
3.2 Geologic Setting 207
3.3 Permeability and Hydrology 207
4. **Technologies** 208
4.1 Exploration 208
4.2 Drilling 209
4.3 Reservoir Engineering and Production Management 211
4.4 Utilization and Conversion 212
4.4.1 Direct Use 212
4.4.2 Electric-Power Generation 213
4.5 Environmental Protection 214
5. **Economics** 214
Glossary 215
Works Cited 216
Further Reading 218

INTRODUCTION

In the widest sense, geothermal energy refers to all heat energy stored in the rocks and fluids of the Earth. It has been well known for over a century that the temperature of the Earth's crust increases with depth. It had been observed by ancient peoples that in certain places the Earth's heat is near the surface; volcanos, lava flows, hot springs, geysers, and hot mud pools are manifestations of the heat. These surface manifestations were used by early mankind for domestic chores such as cooking. These geothermal regions were often treated as sacred places and served as sanctuaries after battles. History tells us that hot baths were very popular as social gathering places, and for relaxation and recreation in early Greek, Turkish, and Roman times. This use is still widespread in the world, and hot-spring spas are especially well developed and patronized in several European and Asian countries; Japan is a notable example.

Industrial uses of steam vents, hot mineral-laden waters, and surface heat occurred in ancient times. One of the most significant of these early developments was beneficiation of salt and chemicals. The concentration of metal ores and chemicals such as borates and borax is well known. In what is now Tuscany, Italy, the Etruscans early engaged in trade in borax products from hot springs. Most of the subsequent development of geothermal energy worldwide has focused on such areas of surface manifestations and has involved simply drilling into reservoirs of naturally occurring fluids. Initially these developments focused on enhancement of the steam flows for chemical concentration and for expansion of hot-spring spas. The earliest exploitations of the so-called hydrothermal resources were similar to mankind's use of coal outcrops and the seeps of hydrocarbons. However, geothermal heat, unlike fossil fuels, cannot be transported great distances and must be used near the resource or converted to electric power for transmission to the users. At the turn of the century, the geothermal area at Larderello in Italy was the location of one of the first applications of natural steam to drive prime movers and produce electricity.

This article first provides a brief review of the history of geothermal energy uses and development. Then follows a section that de-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

fines the classifications and uses of geothermal energy and provides an overview of resource assessment. Basic sections on the geology and hydrology of hydrothermal systems set the stage for a description of nonhydrothermal resources and some advanced research projects on what is termed "heat mining." Finally, sections on technologies, environmental matters, and economic issues are presented.

1. HISTORY

Early peoples often settled or camped around the thermal areas of our planet. In these settlements the hot springs, steam and gas vents, and mud pools were used for domestic chores such as cooking, bathing, and heating. An example of users of these sources of geothermal energy are the Maori people of New Zealand. Some of these people settled in the volcanic regions of the North Island beginning in the 14th century. Their uses of the energy are recorded in legends and are practiced today, especially within the town of Rotorua (Cooke, 1970) and the extensive Taupo volcanic zone. This area subsequently was the location of the first geothermal electric-power production in New Zealand, from a 160-MW(e) power station at Wairaki in 1958 (New Zealand Electricity Department, 1970). The industrial development in that area was typical of how geothermal energy has developed: considerable drilling of wells for domestic uses and especially for resort hotels, baths, health spas, and other commercial ventures preceded the power development and thus provided a continuity of beneficial uses.

Italian thermal areas are the earliest and most widely reported. Invasion and occupation of Italy by the Etruscans occurred about 3500 years ago. Tuscany, where they settled, has numerous thermal areas. These early peoples used the borates found in the thermal waters to produce beautiful pottery glazes. The Etruscans developed trade in medicinal muds from thermal pools and developed therapeutic baths. Both enterprises continued when the Romans later came to dominate the regions. The use of steam vents to concentrate and beneficiate ores was mentioned early in the first definitive text on mining technology (Agricola, 1556). Agricola specifically mentions the production of borax and other borates from the hot pools and vents. The uses of borates as soldering fluxes for silver and gold work, as medicinal applications for eye and skin diseases, and as agents to process wool are recorded in old Greek and Roman writings. The Conte de Lederel used steam from the vents to evaporate vats of boric acid in 1818. Initially, heat for concentration was provided by wood fires under the vats, but destruction of the forests and a need to increase production led to the use of geothermal steam for borate processing. Invention of masonry domes that covered the pools (*lagone coperio*) and a system of pipes resulted in a unique gathering system to concentrate the steam. The town and region were named Larderello in honor of this developer of geothermal energy. The Larderello firm first introduced water-well drilling technology. Then the firm adopted drilling technology from the German mining industry to expand steam supply and chemical production vastly. This profitable firm and its use of the abundant geothermal energy stimulated invention of tools that increased productivity even more. Effective tools included shell-and-tube heat exchangers for production of clean steam, steam piston pumps, and motors for drill-rig winches. Near the beginning of the 20th century experiments were conducted on additional applications of steam-driven motors. In 1904, P. G. Conti, then director at Larderello, put into use the world's first geothermal steam-driven 10-kW electric dynamo to provide lighting in the plant. By 1913, the world's first commercial electric power was being generated to light all the facilities at Larderello, and a network of transmission lines had been established to provide electricity to the towns in the area. This first geothermal electric utility used a 250-kW turbogenerator (Mazzoni, 1954; Villa, 1975; Burgassi, 1987).

Japan has an abundance of heat sources provided by the young volcanism that dominates its geologic setting. Early inhabitants used the hot springs intensively; these sites were the centers of many villages, shrines, and spas. Today there are thousands of resort and health spas located at nearly all the thermal areas (Japan Spa Association, 1983). The development of the many high-grade (that is, high-temperature) hydrothermal fluid resources long recognized to exist in Japan was paced by expansion of resort and therapeutic spa areas. This required drilling of wells to expand production, and development and advancement of the expertise and technologies in exploration and drilling. Table 1 records the early development of electric power (Japan Geothermal Energy Association, 1974), which began in 1919 in the Beppu tourist thermal area of the southern island of Kyushu. Development in that region (see Table 1) was hampered because the reservoirs produced a mixture of hot water and steam, not dry steam as in Italy.

Table 1. Early history of geothermal electric-power development in Japan.[a]

1919	Vice-Admiral Y. Masuji drilled a steam well in Beppu, Kyushu.
1925	Dr. T. Heiji commenced research on geothermal development by building an experimental 1-kW(e) steam generator in Beppu.
1927	Dr. Heiji drilled a steam well at Otake, Kyushu.
1947	Tone Boring Co. drilled steam wells at Yunosawa, Shizuoka Prefecture, and Narugo, Miyagi Prefecture.
1948	Tone Boring Co. built an experimental 3-kW(e) power plant at Yunosawa near Atagawa, Isu Peninsula.
1951	The Industrial Technology Agency of Japan built an experimental 30-kW(e) power plant at Beppu. The Geological Survey of Japan initiated basic surveys of various geothermal areas in Kyushu.
1953	Kyushu Electric Power Co. commenced research work on the Otake Geothermal Field, Oita Prefecture, Kyushu.
1957	Azaumakako Co. and the Geological Survey of Japan commenced cooperative research work at Matsukawa, Iwate Prefecture.
1960	Fuijita Kanko (tourist) Co. operated a 30-kW(e) nonutility geothermal power plant at Kowakudani, Hakone, Kanagawa Prefecture.
1965	Mitsubishi Metal Mining Co. started research work in the Onuma area of Hachimantai, Iwate Prefecture.
1966	A 22-MW(e) geothermal power plant was completed at the Matsukawa field by the Japan Metals & Chemical Co.
1967	A 13-MW(e) geothermal power plant was completed at the Otake Field by the Kyushu Electric Power Co.
1974	A 10-MW(e) geothermal power plant was completed at the Onuma Field by the Mitsubishi Metal Mining Co.

[a] Japan Geothermal Energy Association (1974).

The Japanese did not continue developing the liquid-dominated reservoirs until they heard of successes in New Zealand and the discovery in 1959 of a dry steam field at Matsukawa in northern Japan. The 1959 discovery well at Matsukawa had a 4-MW(e) potential steam production. The first geothermal electric power plant was established there in 1966. The Japanese resumed development of liquid-dominated reservoirs with the Otake plant in 1967. Following these early efforts, electric-power development grew steadily. However, in Japan today more geothermal energy is used directly for such things as bathing, melting snow from roads, washing clothes, and making sake than is converted to electricity, with only 214.6 MW(e) installed electric power capacity (Huttrer, 1990) and 2686 MW(t) consumed in direct use (Freeston, 1990). Following the first "oil shock" in 1973, the Japanese government expanded its support of research and development for geothermal energy and has encouraged industry to expand development of the country's vast high-grade hydrothermal resources. It has also supported projects in exploration and drilling technology, advanced electric conversion cycles for turbines using organic fluid, and experimental programs to develop direct heat mining of subsurface hot rocks (NEDO, 1991).

The pattern of development of geothermal energy in the United States has been much like that in other parts of the world. Use by aboriginal peoples was widespread. When European colonists and settlers displaced these people, the initial uses for public resorts, therapeutic baths, and private use by individual land owners remained. The use of water-well drilling rigs to augment naturally occurring hot springs was practiced from the 19th century. Thus, the warm springs of the eastern regions and hot springs and volcanic zones in the western United States were rapidly exploited, much as they had been in Europe. The thermal areas in the state of Wyoming, now Yellowstone Park, and the thermal areas of California and the volcanic zone along the west coast were noted by early explorers and settlers. The Big Geysers area, about 135 km north of San Francisco, was discovered by a hunter in 1847 (Koenig, 1992), and a hot-spring spa hotel was developed on the site soon thereafter. Exploratory coring and drilling commenced in 1921, and by 1923 the first two wells drilled near the hotel had been completed and a 35-kW turbogenerator was supplying electric power for the resort and for subsequent drilling operations. Drilling equipment was adapted from water-well rigs, but as this equipment proved inadequate, subsequent drilling used a rotary rig with a large-capacity, steam-driven fluid pump. The next four wells were completed by 1925. Analysis of the flow, temperature, and production data (Geiger, 1926) demonstrated the potential for the Big Geysers area; however, there was no market for electric power at that time. Finally, in 1955 drilling was initiated for a power plant of 125 MW(e), which went on line as utility power in 1960 (Rintoul, 1990). This plant used production from a four-well field of about 240 000 lb/h of dry steam. Subsequent exploration and development proceeded apace (Koenig, 1992). Today this mature steam reservoir produces over 2000 MW(e) from some 30 separate power plants. The Big Geysers field development was thus the second commercial development of geothermal energy for electric power generation after Italy. The dry-steam hydrothermal reservoirs at Larderello, Italy, Matsukawa, Japan, and the Big Geysers, United States, had firmly established the viability of this type of energy source by the middle of the 20th century.

The successful development of the liquid-dominated, high-grade, hydrothermal reservoirs at Wairaki, New Zealand (1959), Otake, Japan (1967), and Cerro Prieto, Mexico (1973), had extended the range of case histories to the mixed water and steam reservoirs found worldwide (DiPippo, 1987). These three case histories formed the basis for a much wider development of geothermal energy. Figure 1 illustrates the historical progress of that development, and Table 2 summarizes the total installed capacity of 5828 MW(e) by country for 1990.

The use of natural geothermal fluids in direct domestic and industrial applications has increased rapidly worldwide. In 1992, use of geothermal energy is estimated to be at a rate of 11 385 MW(t). Advances in direct-use technology are outpacing improvements in large power plants. Recent experience shows that development of small power plants probably will increase more rapidly than development of large, conventional steam-turbine plants.

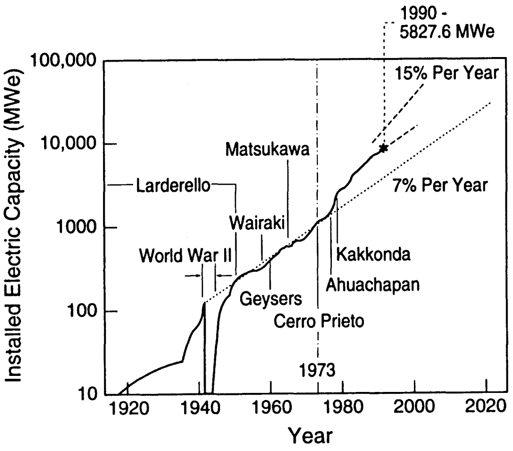

FIG. 1. Worldwide growth trend of geothermal power plants. A relatively steady growth rate of 7% per year of growth can be observed in this figure except for two noticeable perturbations. First was the interruption of growth (in Italy) during World War II, and the second was the surge of growth caused by the oil supply restrictions in the early 1970s.

Table 2. Geothermal installed capacity in megawatts (electric).[a]

Country	1985[b]	1990	Planned 1995[b]
China	0	20.8	50
Costa Rica	0	0	110
Dominica	0	0	10
El Salvador	95	95	125
France	4.2	4.2	4.2
Greece	0	0	?
Guatemala	0	2	15
Iceland	41	44.6	110
Indonesia	32.25	142.25	379
Italy	459	545	885
Japan	214.6	214.6	456.5
Kenya	45	45	105
Mexico	425	700	950
New Zealand	167.2	283.2	300
Nicaragua	35	35	?
Philippines	891	891	2164
Portugal (Azores)	3	3	?
St.Lucia	0	0	10
Thailand	0	0.3	3
Turkey	20.6	20.6	40
U.S.	1444	2770	3170
USSR	11	11	81
Totals	3887.85	5827.55	8967.7

---

[a] Source: Geothermal Resources Council questionnaire; Huttrer (1990).

[b] Capacities for 1985 and 1995 are shown for comparison.

2. RESOURCE TYPES AND ASSESSMENTS

Descriptions and definitions of geothermal energy types evolve continuously. Elder (1981) states that exploitation of the Earth's internal heat is still in its infancy. But acceleration of resource development has been aided by the efforts of the United Nations, which has focused on education of scientists and technicians from developing nations. The development of local energy sources is especially valuable and beneficial as it reduces problems of foreign exchange and offers some immunity to fossil-fuel supply crises. This has been true for Mexico (Lippman and Manon, 1987) and the Philippines (Otte, 1991). Introduction of new concepts and methods of heat mining (Armstead and Tester, 1987) and accumulation of case histories of successful exploration and exploitation have led to two important results. First, the increasing number of successful projects has accelerated development. Second, introduction of new resource concepts and methods of heat mining has greatly expanded our perception of the extent of geothermal resources.

2.1 Geothermal Resource Types

Temperature increases with depth in the Earth, as shown by the current model of the Earth's interior [Fig. 2(a)] (see EARTH, INTERIOR STRUCTURE OF). The average or normal temperature distribution nearer the surface is shown in Fig. 2(b). Sufficient global data have revealed that the average temperature gradient at the surface, that is, the rate at which the temperature increases, is about 25 °C/km, and the average surface heat flux is about 50 mW/m^2. This heat flux, q_s, is related to a nominal thermal conductivity according to the equation

$$q_s = -\bar{\lambda}\frac{dT}{dz}, \tag{1}$$

where $\bar{\lambda}$ is an average value of the thermal conductivity of the material near the surface. This heat flow from the Earth's interior is small compared to the incoming average solar heat flux on the illuminated side of the Earth, which is 175 W/m^2, a factor some 3500 times greater than the surface heat flux. The solar flux is absorbed in the atmosphere and surface features, distributed around the globe, and reradiated. The average surface temperature is set by this radiative balance at about 15 °C (Elder, 1981).

The nominal heated global volume contains an enormous amount of stored heat energy. The heat is due to the energy from

1. its original gravitational accumulation;
2. the differentiation of the core, mantle, and crustal components; and

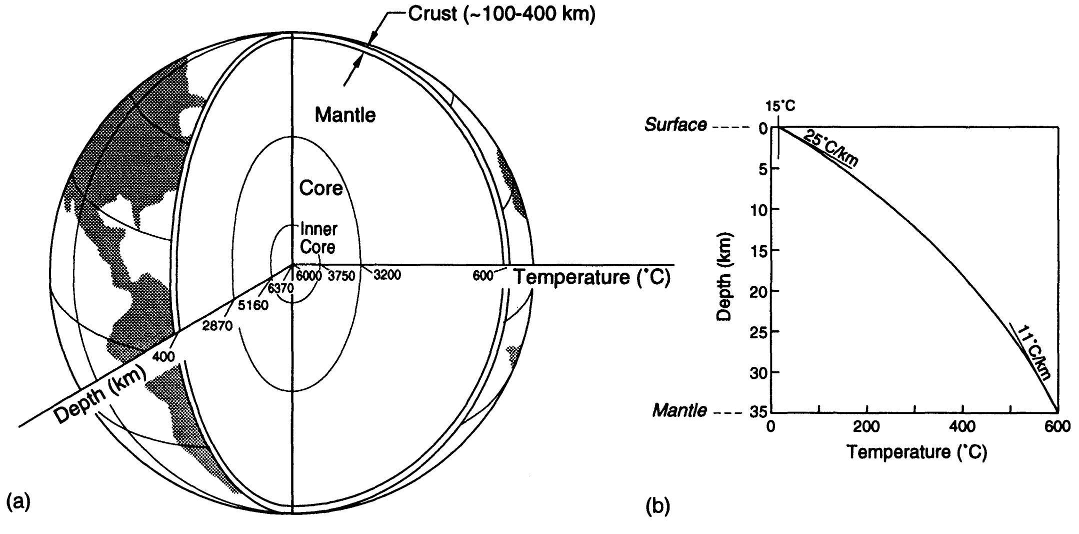

FIG. 2. Temperature distribution within the Earth (Mustoe, 1984). (a) Nominal thermal structure of the Earth. (b) Nominal temperature profile near the Earth's surface.

3. radioactive heating from natural uranium, thorium, and potassium (Bowen, 1989).

There is estimated to be some 1.3×10^{27} J stored in the crust to a depth of 10 km (Armstead and Tester, 1987). This amount is derived by integrating the temperature curve in Fig. 2(b) to a depth of 10 km and using average values of crustal rock properties. The heat content $Q(z_f)$ contained in the volume beneath the surface area A is given by

$$Q(Z_f) = A \int_0^{Z_f} \bar{\rho}(z) C(z) T(z) dz, \qquad (2)$$

where Z_f is the depth, $\bar{\rho}(z)$ is the density, $C(z)$ is the volumetric specific heat content, and $T(z)$ is the temperature as a function of depth. The average density is 2500 kg/m^3 and the average specific heat is 770 J/kg °C (Rowley, 1982). This energy, if recovered, would be equivalent to the combustion of about 2.3×10^{17} barrels of oil. This can be compared with the annual U.S. consumption of oil (6.2×10^9 barrels) or the estimated world reserves of oil (1×10^{12} barrels).

However, the easily accessible, shallow, high-grade geothermal energy resources are located in the geothermal belts or zones of the Earth as shown in Fig. 3. These areas make up only 1% or so of the total land area. They are primarily located at the crustal plate boundaries (Rybach and Muffler, 1981), and the occurrences of natural geothermal systems in these zones are often referred to as thermal anomalies. The descriptions and definitions of energy resources set forth in this section are not universal and are introduced or adopted here for convenience of discussion.

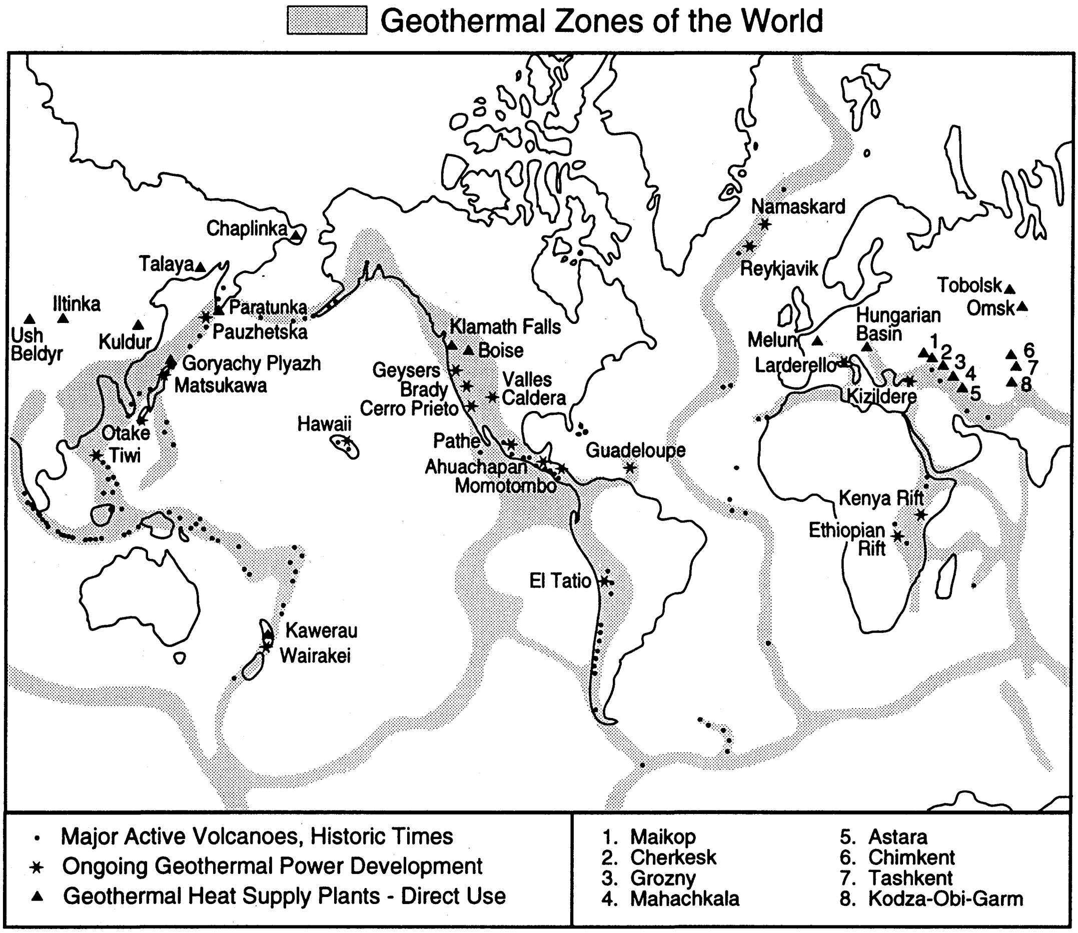
▨ Geothermal Zones of the World
Chaplinka
Talaya
Iltinka
Ush Beldyr
Kuldur
Paratunka
Pauzhetska
Goryachy Plyazh
Matsukawa
Otake
Tiwi
Hawaii
Kawerau
Wairakei
Klamath Falls
Boise
Geysers
Brady
Valles Caldera
Cerro Prieto
Pathe
Ahuachapan
Momotombo
Guadeloupe
El Tatio
Namaskard
Reykjavik
Tobolsk
Omsk
Hungarian Basin
Melun
Larderello
Kizildere
Kenya Rift
Ethiopian Rift
1
2
3
4
5
6
7
8
• Major Active Volcanoes, Historic Times
✶ Ongoing Geothermal Power Development
▲ Geothermal Heat Supply Plants - Direct Use
1. Maikop
2. Cherkesk
3. Grozny
4. Mahachkala
5. Astara
6. Chimkent
7. Tashkent
8. Kodza-Obi-Garm

FIG. 3. Geothermal zones of the world. Most of these zones are located at active tectonic plate margins.

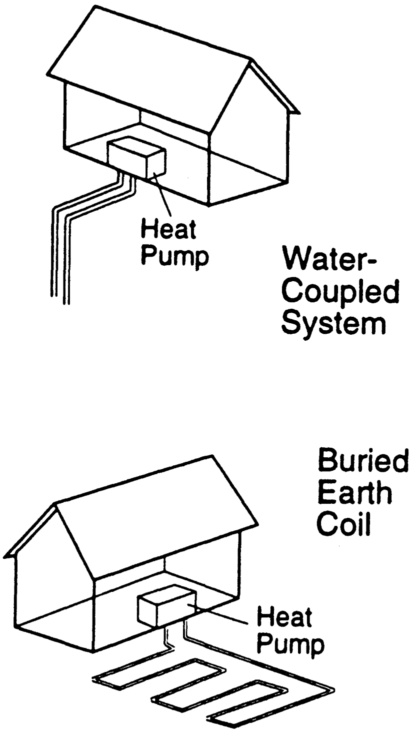

FIG. 4. Use of Earth energy resources for space conditioning.

2.1.1 Earth Energy. This resource type is applied in the United States to the near-surface use of direct soil contact and normal-gradient shallow aquifers for heat-pump installations. The heat-pump sink is either a pumped flow of shallow aquifer water, a heat-exchanger loop inserted in a water well, or a coil of tubing buried in the earth. Figure 4 illustrates these systems (Rafferty, 1989). Perhaps 100 000 units have been installed in the United States and Canada (Pratsch, 1990). These installations have mainly been in the middle west and eastern parts of these countries, but they can be installed essentially anywhere. Heat pumps coupled to the Earth or to water wells have been promoted by electric utilities to relieve peak winter heating and summer cooling loads. This shallow geothermal Earth energy resource is expected to find much wider use worldwide; currently its use is increasing at about 25% per year.

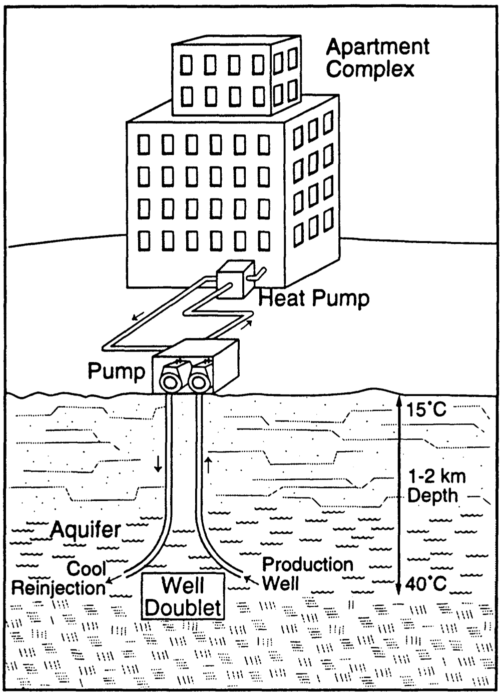

FIG. 5. Hot water from deep aquifers in normal-gradient (25 °C/km) basins is used for space heating. Two-well (doublet) systems depend on sufficient spacing between production and reinjection wells to prevent aquifer cooling. These conditions occur beneath Louisville, Kentucky, U.S.A., and Paris, France.

2.1.2 Normal-Gradient Aquifers. There is an increasing use of deep aquifers that are located in sedimentary basins. The technology consists of a doublet, or two-well system (Fig. 5), that is pumped at large flow rates and uses a heat pump to heat and cool large apartment and industrial complexes (Boisdet *et al.*, 1990). Doublet-source configurations use sufficient separation of the production and reinjection wells to avoid significant cooling of the production well.

2.1.3 Hot and Warm Aquifers. This is a widespread resource and supports direct-use application. It is found especially in thermal zones where hot and warm springs occur naturally. Water-well drilling in these areas produces hot to warm water at relatively shallow depths. Many domestic and industrial heating systems using this resource have been developed; they are rapidly expanding worldwide. A notable example is in Iceland, where Reykjavik, the capital city, has more than 90% of its thermal energy needs serviced by such a system, with a total of 705 hot-water wells supplying some 4600 kg/s. Another high-gradient region occurs in the Carpathian Basin of Hungary, where the heat flux and thermal gradient are about double the world average. This country has long used deep wells drilled into warm-water aquifers for agricultural, public bath, and industrial uses. Nearly 1000 flowing wells produce over 1200 MW(t) with a total flow rate of 12 000 kg/s at temperatures between 50 and 90 °C. Over a year this usage is equivalent to combustion of 24×10^6 barrels of oil. Agricultural use, mainly for greenhouses (29%), is significant to the economy of Hungary. In the United States (Lienau and Lunis, 1989) this resource type is distributed widely throughout the eastern states. Nearly 68 000 individual applications are based on warm or hot aquifers. These displace the combustion of perhaps 3×10^6 barrels of oil a year. District heating in Boise, Idaho, Klamath Falls, Oregon, and Mammoth Lake, California, and enhanced oil recovery and gold leaching enhancement projects have expanded the potential for using this type of resource.

2.1.4 Hydrothermal Convection Resources. In thermal areas, there exist heat sources in the form of magma that intrudes into the shallow crust or extrudes as in volcanic activity. These heat sources are cooled by heat convection to naturally circulating water. These are the resource types that have been given the most intense study and have been most widely and effectively exploited. The high-grade (> 150 °C) fluids in these convection systems are very valuable and form the basis for the conversion of thermal energy into electric power. These resources are characterized by two major subtypes of reservoirs. One is vapor dominated, with dry steam circulating in the reservoir; the other is liquid dominated, with a two-phase mixture of water and steam circulating in the reservoir. There are two essential requirements for reservoir properties: sufficient porosity and permeability, and a convective flow of water to bring the energy toward the surface. Such reservoirs of convecting fluid are tapped by drilling and will produce hot fluids, either as steam or, more commonly, as a mixture of water and steam. Figure 6 is a schematic of a typical hydrothermal system (Muffler, 1985). Exploration and exploitation of this resource type has been rapid in the past half-decade (Huttrer, 1990). Within this period a total of 166 electric power plants were installed worldwide, with 143 in the United States alone. Most hydrothermal resources are liquid dominated (DiPippo, 1987). Most are found in volcanic zones, but some are located in large sedimentary basins. One example is the Imperial Valley, California, U.S.A., which is on the active boundary of the Pacific and North American plates. Here several thermal anomalies that had no surface manifestations were discovered and drilled to develop productive reservoirs. The large development in Mexico south of the Imperial Valley, at the Cerro Prieto field (Lippman and Manon, 1987) is another example of hydrothermal resources in such a geological setting.

2.1.5 Geopressured Resources. Geopressured reservoirs containing hot brine and quantities of dissolved methane (natural gas) are often encountered when drilling in deep basins for hydrocarbon accumulations. These reservoirs are usually associated with undercompaction of sediments that have an imper-

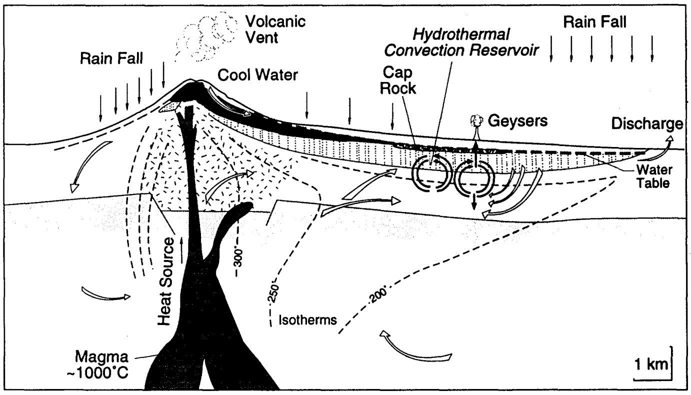

FIG. 6. Hydrothermal system. This illustrates a typical system associated with a volcanic heat source. The system consists of a hot magma, at about 1000 °C, cooled by convectively circulating water that is continually supplied by rainfall. The system contains a reservoir that is a local volume of permeable rock.

meable cap rock (shale). An internal pressure higher than hydrostatic pressure at depth is generated because of the inability of the undercompacted sediments to support the overlying rock mass. The most widely known geopressured zones are found in the Gulf of Mexico regions of Mississippi, Louisiana, and Texas, U.S.A. (Dorfman and Morton, 1985); however, many other deep basins worldwide are known to contain such resources.

2.1.6 Hot Dry Rock Resources. This resource type consists of the vast, pervasive mass of crustal rocks that exist beneath all continents (Smith, 1988). This resource type is therefore not necessarily restricted to the thermal zones or geothermal belts. The access and exploitation of this resource depend on a developing technology that drills a pair, or doublet, of wells into the hot rock, where no significant porosity or permeability exists, creates an artificial fractured reservoir, injects water into and through the artificial fracture system, and in a closed-loop process extracts the energy by the usual heat-exchanger or turbogenerator electric conversion technologies.

This technology has active research and development projects in the U.S., Japan, and Europe (Anderson, 1992). Based upon rather modest extensions of present drilling and reservoir technologies, it may be expected that commercial prototype hot dry rock resources will be in place by the end of the 20th century. Warm dry rock systems, used in conjunction with geothermal heat pumps in Sweden, have displaced nearly 100 MW(e) of needed new electrical capacity.

2.1.7 Magma Resource. Discrete intrusions of molten rock (magma) are being studied in the United States (Hardee, 1988). This concept uses discrete magma bodies close enough to the surface to be drilled into and a heat exchanger installed in the holes. The initial targets are thought to be the large, recent (silicic) calderas (volcanic collapse craters). The magma body that is believed to underlie the Long Valley caldera in California, U.S.A., has been selected for initial drilling and experiments. The magma energy concept requires considerable advances in drilling and materials technology before it can be brought to a demonstration stage. The energy of a large, high-temperature magma body is very great. For example, consider that a 4.0-km^3 volume at about 1000 °C would release about 6.5×10^4 TJ (6.5 Quads) when cooled to 150 °C. If this heat were captured and converted to electricity at a 10% conversion efficiency, it could produce 200 MW(e) for a century.

2.2 Resource Assessments

In present geothermal resource assessments and reserve estimates, it is important that the objectives of the assessments and the basis for the results are understood. At this early stage in geothermal energy development, there is a relatively limited amount of subsurface information and few published well and production data available. Therefore this account serves best as a status report.

A global assessment developed for the 1977 World Energy Conference (Roberts, 1978) calculated and estimated the geothermal energy storage to a depth of 3 km [Table 3 and

Table 3. Global geothermal resource base estimate in 10^{12} TJ, based upon geographic distribution of geothermal zones (3 km integration depth).

Region	< 100 °C	100–150 °C	150–250 °C	> 250 °C	Totals
North America	7.10	1.20	0.32	0.02	8.64
Central America	0.15	0.09	0.03	0.0	0.27
South America	4.6	0.83	0.23	0.01	5.67
Western Europe	1.5	0.07	0.02	0.0	1.59
Eastern Europe	6.7	0.25	0.06	0.0	7.01
Asia	7.5	0.79	0.21	0.01	8.51
Africa	5.2	0.33	0.07	0.01	5.61
Pacific Islands	3.3	0.26	0.17	0.01	3.74
Totals	36.05	3.82	1.11	0.06	41.04

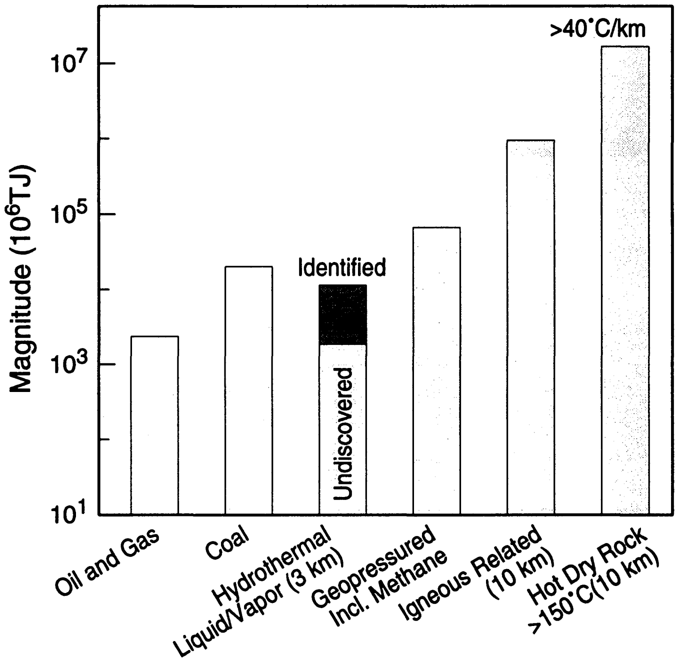

FIG. 7. Comparison of geothermal, coal, oil, and gas resources for the United States (Rowley, 1982). This figure shows that the total of all geothermal resource types is approximately 14 orders of magnitude greater than the total fossil resource base. For reference, 10^6 terajoules (TJ) equals 10^{15} British thermal units (Btu) equals 10^6 tonnes (t) of oil equals 1.4×10^6 tonnes of coal equals 1 megawatt thermal [MW(t)] of power use for one century.

Eq. (2)]. Such assessments are usually restricted to estimates of long-term potentials, priorities for alternative-energy research and development, and comparisons with the resource bases or measured quantities of other energy resources.

A few national and region geothermal resource assessments have been developed, and those for Europe (Haenel and Staroste, 1988), Italy (Muffler and Cataldi,1978), Japan (Yamaguchi, 1992), New Zealand (Donaldson and Grant, 1978), and, most notably, for United States (White and Williams, 1975; Muffler, 1978; Reed, 1983) illustrate the value of such assessments. Figure 7 summarizes these results for the United States and compares geothermal resources with coal and oil resources.

The methods used for regional assessments have been restricted generally to estimates or calculations of the geothermal energy stored in the subsurface rocks and fluids, and are usually restricted to depths of less than 3 km as a judgment of the economically accessible drilling depth. The definitions and philosophy of deriving energy assessments have been very different across the various energy resource types. However, the most definitive and effective methodology for the geothermal types has been developed by Muffler (1978).

The resources assessments of shallow aquifers rely on long-established water resource evaluation techniques. Major sedimentary basins can therefore be relatively easily evaluated for the energy in place. Another assessment method that has been advocated by McNitt (1978) and applied in Central America (Einarsson, 1975) is based on statistical analyses of already explored and drilled resources and producing, high-grade hydrothermal reservoirs. This methodology was used (Einarsson, 1975) with a very preliminary survey of hot-spring areas to project a minimum electric power production for Central America of 8000 MW(e). The value and usefulness of such methodology can only be determined by long-term comparisons with the results of actual geothermal development.

3. GEOLOGY AND HYDROLOGY OF GEOTHERMAL RESOURCES

This section describes only the resource type introduced in Sec. 2.1.4 above, hydrothermal convection resources.

3.1 Nature and Occurrence

Near-surface hydrothermal systems occur as a result of the geologic transport of above-average amounts of heat into the shallow crust. Therefore the occurrences of hydrothermal systems are not random but are determined by geological processes on local, regional, and global scales. Useful hydrothermal reservoirs contain the following elements: heat sources, permeability in the subsurface rocks that provides pathways for the fluid-transport system, water in the interconnected porosity, and usually a cap rock or blanketing overlying structure that limits leakage through natural surface manifestations.

3.2 Geologic Setting

The higher-than-average temperatures that result in so-called geothermal zones are caused by the following five geologic processes:

1. intrusion of magma into shallow levels of the crust,
2. occurrence of thinned crust and resulting higher-than-average heat flow,
3. circulation of heated groundwater along faults to depths of 1 to 5 km,
4. insulating blankets or rock layers of low thermal conductivity above hot aquifers, or
5. high volumetric radioactive heating rates in shallow, buried plutons.

The majority of high-grade resources have been found to be a result of the first process, although reservoirs based on heat sources of types **2** and **3** have been exploited. The blanketing process is typically found in sedimentary basins, which have lower temperatures and are suitable for direct use. The fifth process, radioactive heating, has been explored but as yet not exploited (Costain *et al.*, 1980).

Intrusion of magma is most commonly found within the global volcanic belts, Fig. 3, and associated with the margins of crustal plates. Intraplate heat sources, usually associated with mantle plumes, also produce near-surface hot spots. These latter geologic processes often produce large magma bodies and associated silicic calderas. In the United States, such calderas are found in the Yellowstone complex in Wyoming, Long Valley in California, and the Valles in New Mexico. The relationships of geothermal systems to volcano heat sources have been discussed extensively by Wohletz and Heiken (1992). The schematic geologic cross section in Fig. 6 typifies hydrothermal systems generated by volcanic heat sources.

3.3 Permeability and Hydrology

In most geothermal systems, porosity and permeability result from fracture networks in the brittle, crystalline rocks associated with thermal zones, tectonic stress caused by crustal plate interactions, and the forces induced by intrusion of hot magma. Even in layered sedimentary basins, the intergranular porosity is filled in by deposition of secondary minerals from the circulating geothermal fluids; the result is densification and the production of secondary, brittle fracture networks. The subsurface fracture networks are frequently associated with major fault zones and form roughly orthogonal fracture systems. Such fracture systems often lead to an elongated zone of permeability and porosity along the fault zone. Such fault and fracture zones also tend to be quite deep relative to the lateral extent of the reservoir.

The subsurface fluids are usually a result of groundwater derived from precipitation and runoff that penetrates into and then circulates within the hydrothermal system as illustrated in Fig. 6. A few reservoirs have a significant fluid component from seawater, as in the Fushime geothermal field at the southern tip of the island of Kyushu, Japan (Yoshimura *et al.*, 1985). However, systems do occur without surface recharge and discharge areas, such as the reservoirs in the Salton Trough in California, U.S., where several reservoirs without any surface manifestations of discharge such as hot springs or gas vents have been found (Elders and Pena, 1979).The hydrological regimes of the hydrothermal systems studied thus far are very diverse, and the location and nature of the reservoir are strongly influenced by the locations of the recharge and discharge areas, as well as by the size, temperature, depth, and age of the heat source. These factors have combined to produce reservoirs with areas ranging from about 0.5 to 200 km^2 as defined by the 180 °C isotherm (McNitt, 1978). The nature of the fracture networks and the hydrologic variables also cause two other commonly encountered situations, namely, the production of a condensing zone at the top of the convection cells within the reservoir, and the often resulting chemical deposition and sealing of the fractures, with subsequent formation of a cap rock.

The nature, spacing, and aperture of the subsurface fracture systems are determined by past tectonic activity and hydrothermal activity (Norton, 1988; Norton and Knight, 1977), lithology (Morse and Thorson, 1978), and local and regional crustal stresses (Nakamura, 1977). Because crystalline rocks have very low matrix permeabilities and porosities, the fracture systems control the volume and circulation of the fluids that provide the removal of heat from the reservoir and the convection of heat toward the Earth's surface. These fractured reservoir conditions are very different from those of the commonly encountered oil and gas reservoirs [although a few fractured petroleum reservoirs have been studied (Aguilera, 1980)]. The fracture porosity found in geothermal systems is generally less than 10%, but the permeability can vary by six orders of magnitude.

4. TECHNOLOGIES

The technologies applied to the development and production of geothermal energy are largely adaptations of those used in the water-well, mineral, and petroleum industries. The fractured nature of geothermal reservoirs influences every aspect of applying the most commonly used technologies to geothermal development and production. The technologies are most conveniently discussed in five categories: exploration, drilling, reservoir engineering and production management, utilization and conversion, and environmental protection. These categories are discussed in the order in which the work is normally conducted in a geothermal project.

4.1 Exploration

Exploration for geothermal resources is a relatively new and still developing area of applied science. Exploration activity and advances in the technology have expanded rapidly in the past decade and a half. Before 1965 the pioneering efforts were carried out primarily by Italy, New Zealand, Japan, Iceland, and France. The methods applied were adapted from those used for water, mineral, and petroleum resource development and the scientific study of volcanos.

Very early exploration programs concentrated on geothermal areas of large extent and abundant surface manifestations. These early efforts were often merely the drilling of step-out wells to enhance the fluid flows as at Larderello, Italy (Burgassi, 1987), and many other locations, but by the early 1970s the exploration activities had expanded until systems with little or no surface indications were being successfully located. At present there is no generally accepted approach to planning and conducting exploration for geothermal resources. However, it is usually conducted by an exploration team of geologists, geochemists, hydrologists, geophysicists, and drilling engineers. The effort usually starts at the regional level (Wright *et al.*, 1989) and as existing field work, principally geologic mapping and geochemical sampling, progresses, the area of investigation is narrowed (Ward *et al.*, 1981). Early in the assessment of the local area, a conceptual model (Elders, 1991) is developed, containing the four basic elements given in Sec. 3.1. The process, diagrammed in Fig. 8, is refined as detailed data are accumulated. The prospecting methods have as their prime objective the location of a hot subsurface volume of fluid contained in rock of sufficient porosity and permeability to sustain circulation. Table 4 lists the techniques typically employed. Very basic are the geothermometers (Henley *et al.*, 1984), data obtained from the hot springs, gas vents (fumaroles), and shallow drill holes that can often determine the base temperature of the system (Truesdell, 1984). These techniques can often determine the base temperature that the fluid reaches at depth and thus set the grade of the resource, i.e., whether it is useful for electric-power generation or, at lower temperatures, direct-use applications. The methods used must be capable of distinguishing between hot dry rocks and hot permeable rocks containing fluids.

As detailed data accumulate, temperature-gradient data from shallow core holes are coordinated with surface geological and geophysical survey results. The initial surface geophysical surveys are usually done with

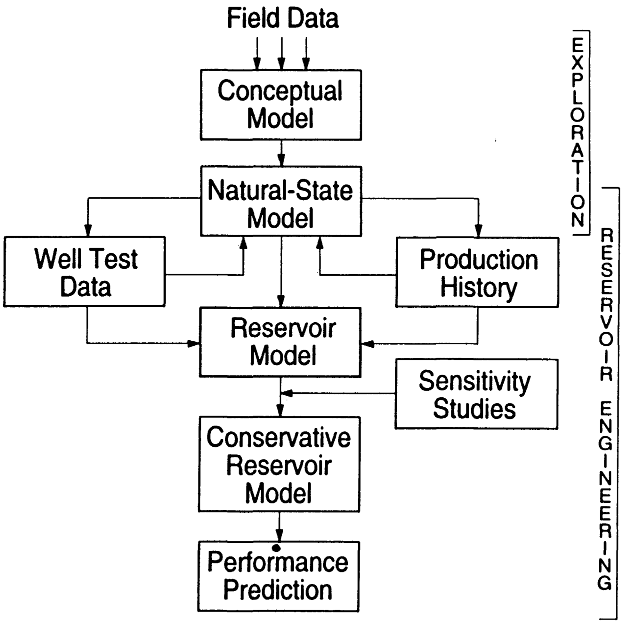
Field Data
Conceptual Model
Natural-State Model
Well Test Data
Production History
Reservoir Model
Sensitivity Studies
Conservative Reservoir Model
Performance Prediction
EXPLORATION
RESERVOIR ENGINEERING

FIG. 8. Reservoir evaluation of model development. This diagram suggests the process in the development of a conceptual reservoir model. A critical step in this process is the establishment of a valid natural-state model, the simulation of reservoir characteristics prior to start of production.

gravity and dc resistivity measurements (see GEOPHYSICS). These surveys are calibrated by the use of resistivity and density logging in shallow and intermediate-depth holes. These data can provide a sufficient refinement of the subsurface reservoir model to allow a decision to be made: whether to locate and drill a first, deep exploratory well or not to proceed with the exploration (Hill, 1985). Typical results of commonly used geophysical survey of subsurface resistance are illustrated in Fig. 9 and often are prospective for a quantity of hot fluids, because hot brine contained in pores has a considerably lower resistance than the surrounding rock (Goldstein, 1988).

4.2 Drilling

Geothermal drilling equipment and practices have developed from construction, water-well, mineral, and petroleum technologies. Early drilling projects in Italy (Mazzoni, 1954) and the Big Geysers field in California, U.S.A. (Rintoul, 1990), illustrate the progression of the early stages of drilling that used lightweight equipment. Frequently, equipment selection focused on modifications of mining drills. The selection was due to the hard, abrasive, and fractured nature of the coarsely crystalline and volcanic rocks. Equipment must be modified to control blowouts. Even though blowouts are rare (Zvulun, 1989), they do occur when drilling encounters an energetic steam-production zone and the surface valves and other flow-preventor equipment fail to contain the flow. Geothermal blowouts are similar to high-pressure gas blowouts in

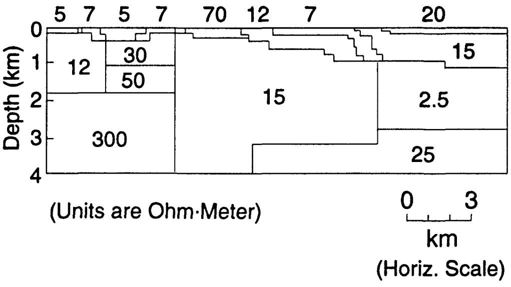
5 7 5 7 70 12 7 20
0
1
2
3
4
Depth (km)
12
30
50
300
15
15
2.5
25
(Units are Ohm·Meter)
0 3
km
(Horiz. Scale)

FIG. 9. Two-dimensional dc-resistivity reservoir model derived by numerical inversion of surface geophysical survey data. Low resistivity values indicate the presence of fluid-filled porosity.

Table 4. Geothermal exploration techniques and activities.

Discipline	Techniques
	Basic Techniques—Development of Initial Conceptual Model
Geology	Surface and photogeologic mapping, volcanic stratigraphy, structure, hydrothermal alteration Initial conceptual model
Geochemistry	Sampling and analysis of surface discharges Analysis of superficial deposits Field measurements of temperature, pH, and flow rates of discharges Geochemical geothermometry
Geophysics	Shallow to moderate-depth drill holes for temperature surveys Magnetic, gravity, and dc-resistivity surface surveys Resistance, density, and sonic-velocity logging of shallow holes
Drilling	Shallow heat flow and core holes Water and observation wells
	Supplemental Techniques and Activities—Refinement of Conceptual Model
Geology	Volcanic and plutonic geochronology Paleomagnetism Refined conceptual model
Geochemistry	Stable-isotope studies Subsurface alteration and trace-metal distribution Mixing and boiling models
Geophysics	In-field heat-flux measurements Magnetotelluric soundings Controlled-source electromagnetic soundings Self-potential Passive and active seismic surface surveys Logging and deep exploration wells
Drilling	Intermediate-depth core holes and monitoring wells Deep exploration wells

petroleum drilling, and the geothermal methods and equipment are adaptations from that industry (Goins and Sheffield, 1983). Other modifications are needed to control the wells and to drill large-diameter holes to accommodate large production flows. Drilling for production from direct-use reservoirs (Culver, 1989) requires only adaptations to accommodate the elevated temperatures.

Production drilling usually commences at a geothermal prospect once the deep exploration phase of drilling has been completed and well tests have demonstrated the reservoir potential of the field. As in all subsurface excavation processes, geothermal drilling usually follows a learning curve of improvement as knowledge of subsurface conditions increases through time. Achieving an increasing level of drilling effectiveness is very desirable, as drilling costs can be 40% to 60% of the total project budget. Geothermal drilling is more costly than wells drilled to comparable depths for oil and gas; this cost factor can be 2 to 4 times greater (Kelsey and Carson, 1987). Increased costs are directly related to the severity of subsurface conditions. Also, the average geothermal well is about twice as deep as the average U.S. oil well. These relative cost factors are summarized in Fig. 10 from the U.S. geothermal data base (Rowley, 1988).

Data on the history of geothermal drilling in the United States are presented in Fig. 11, where it is shown that the majority of production wells (over 500) have been drilled at the Big Geysers Field (steam) in California. This drilling is done using air as the drilling fluid,

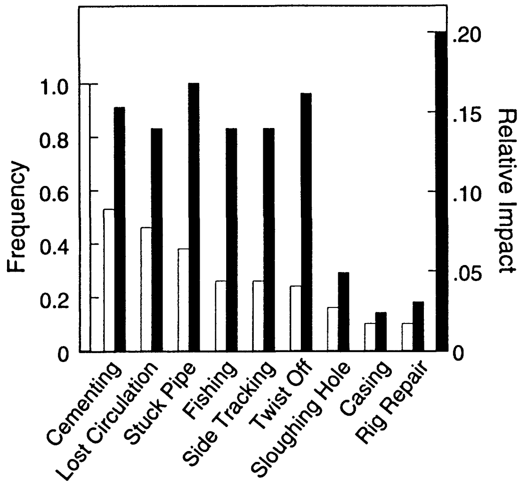

FIG. 10. Summary of frequency (open bar) and relative impact (solid bar) of geothermal drilling problems. Relative impact is a problem's frequency times the number of days' delay it causes.

as is usual for vapor-dominated reservoirs. Air drilling prevents reservoir fracturing, loss of circulating fluids, and plugging of fractures caused by the hydrostatic pressure of a liquid column (Wolke *et al.*, 1990). Directional drilling is used primarily to steer a borehole toward a production-zone target (Burkman and Nickels, 1990). Half of the geothermal wells in the United States are drilled in this way. It can also solve tight topographical constraints, lower the costs of drill pad construction, and allow for alternate drilling of reinjection wells. Directional drilling equipment needs upgrading to high temperatures (Saito, 1992) to incline and steer the wellbore so that it intersects near-vertical and azimuthal fracture trends (Shale and Nickels, 1992). Especially important are high-temperature–rated downhole drilling motors, for both air and liquid, and the steering instruments needed to steer the drilling assembly out of vertical and in a particular azimuthal direction.

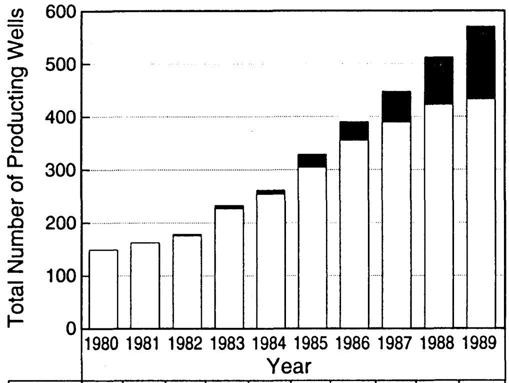

	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Liquid	0	0	5	11	12	24	35	56	82	128
Steam	151	163	175	224	252	306	354	390	428	439

☐ Steam Field ■ Liquid-Dominated Fields

FIG. 11. The total number of wells has increased; the percentage of those wells coming from liquid-dominated fields has also increased. The only steam field in the United States is at the Big Geysers field in California.

4.3 Reservoir Engineering and Production Management

Over the past decade, the ability to simulate and provide predictions of hydrothermal reservoirs has progressed rapidly (Bodvarsson and Witherspoon, 1989). The reservoir engineer's tasks are

- prediction of the production potential of the reservoir,
- determination of the number and locations of production wells,
- prediction of declines in total field and individual well production rates,
- determination of the number and locations of reinjection wells to minimize the cooling of production wells by cold injection fluids, and
- projection of the reserves of energy and a production strategy that will sustain production at the needed levels for the lifetime of the project, usually 20 to 30 years.

Simulation methods have been developed from those previously perfected for petroleum- and water-reservoir analyses. The process starts with collection of the properties of both reservoir rocks and fluid. The reservoir engineer then directs

- conduct of temperature- and pressure-logging surveys and other production measurements as the wells are drilled,
- interpretation of pressure transient test results on production wells,
- construction of reservoir computer simulation designed to model the reservoir in its undisturbed natural state, and
- extension of that model to one that simulates, over time, production from the reservoir.

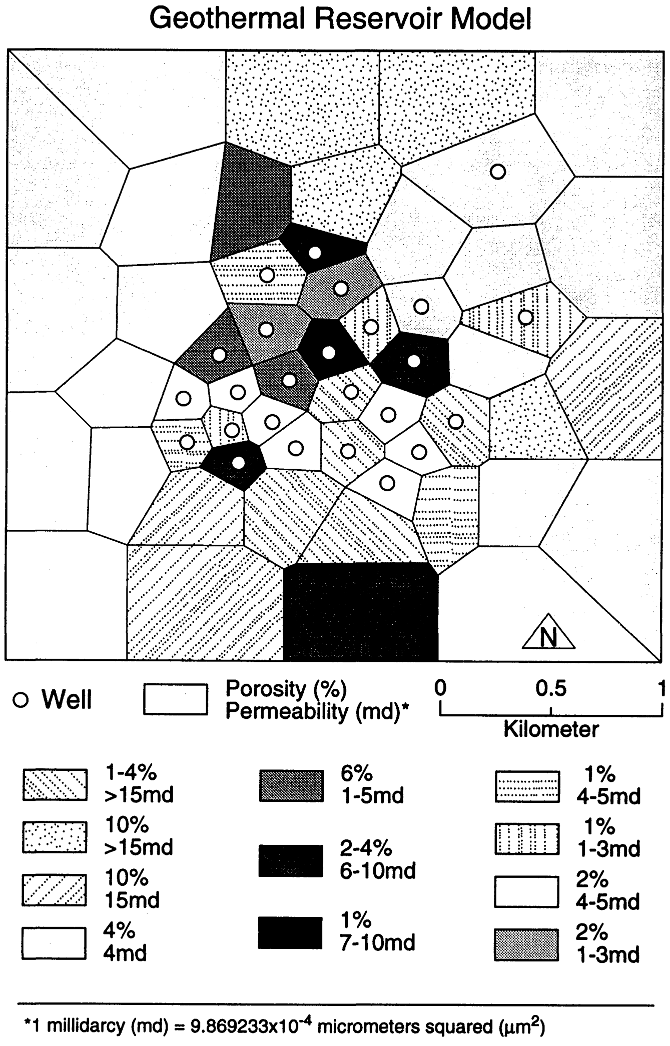
Geothermal Reservoir Model
N
○ Well
Porosity (%)
Permeability (md)*
0
0.5
1
Kilometer
1-4% >15md
10% >15md
10% 15md
4% 4md
6% 1-5md
2-4% 6-10md
1% 7-10md
1% 4-5md
1% 1-3md
2% 4-5md
2% 1-3md
*1 millidarcy (md) = 9.869233x10-4 micrometers squared (μm2)

FIG. 12. Cross section of geothermal reservoir model showing rock porosities and permeabilities within the producing zones at Olkaria Field, Kenya (Bodvarsson and Witherspoon, 1989).

The process of reservoir simulator development is illustrated in Fig. 8. Early reservoir models were first established in New Zealand (Grant *et al.*, 1982). Subsequently, they have advanced, and models now include such essential complexities as fracture flow of two-phase fluids, three-dimensional conduction of heat between fractures, and the sweep flows of cooler fluid from both natural recharge and reinjection water. Figure 12 illustrates an example of a typical computer simulation on a well-by-well basis.

Such reservoir engineering activities can provide the basis for production management. This management requires continual updating of the estimated reserves, adjustment of individual well production, and use and decisions about drilling of new wells to maintain the total needed power production.

4.4 Utilization and Conversion

The technologies used for utilization of the geothermal heat of the Earth are primarily adaptations and applications of existing industrial systems for heat exchange or conversion to electric power.

4.4.1 Direct Use. Geothermal heat is used directly for many things (Lienau and Lunis, 1989); its direct use is greater than its use for conversion to electricity (Freeston, 1990). In the United States there are 452 sites that produced 5,466 GWh of energy in 1990. It is reported that 44% is used for enhanced oil

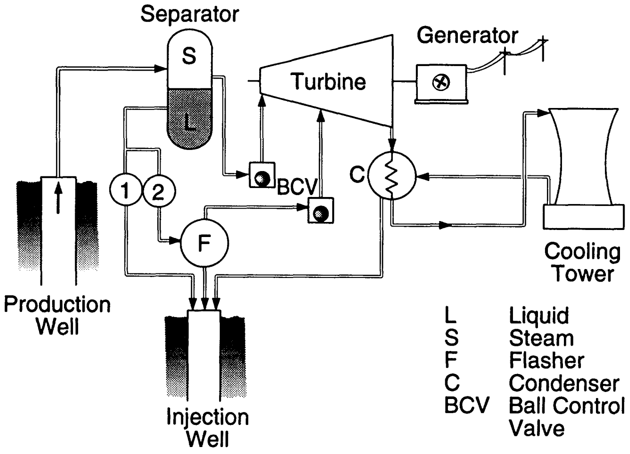

FIG. 13. Schematic of geothermal power plant showing two types. In both types, separated steam is fed to the turbine. Then in the first type (1), separated liquid is reinjected directly into the injection well. In the second type (2), the separated liquid is flashed again in flasher F, and the separated steam enters the turbine's lower pressure stage. The second type increases power plant efficiency.

recovery and 30% for earth-coupled heat-pump installations. Twenty-seven new district heating systems that have been developed in the past five years, heap-leaching for gold ore extraction, and various greenhouse, aquaculture, and building heating installations accounted for about 100 MW(t). Very significant is the rapid growth of installed heat pump units for domestic space conditioning growth. About 30 000 of them use groundwater wells and another 50 000 use direct earth-coupled loops. Together they produce 400 MW(t) annually. Japan (Higo, 1990) makes by far the most use of geothermal water with a total estimated thermal power of 3,321 MW(t), with about 90% used for bathing and hot spring resorts. The remaining 10% (Sekioka, 1990) is used for diverse applications; use for greenhouses predominates, and snow melting applications have been recorded.

4.4.2 Electric Power Generation. The technologies and facilities for steam utilization from vapor- or liquid-dominated reservoirs (DiPippo, 1987) can readily be modified from conventional thermal power plants (Kestin, 1980). Major modifications include scrubbing the steam of impurities, especially the noncondensable gases CO, CO_2, and H_2S, and reducing scaling of the pipe lines, valves, and turbines. Also, the cooled condensate fluids and the by-products of the steam-scrubbing process must be disposed of. A simplified schematic of a geothermal power plant with single-flash and double-flash options, Fig. 13, illustrates the other major difference from fossil or nuclear plants. The geothermal system is a steady flow-through process with the hot geothermal fluids being fed directly through the turbine and then reinjected after they cool. In contrast, conventional thermal plants use an external heat source in a boiler and run the fluid in a closed cycle through the turbine.

The optimum size of geothermal power plants is 50 to 100 MW(e) capacity. This has been determined from experience and is related to a tradeoff between the cost of extending the well-field piping, the energy losses in long piping runs, and the incremental costs of locating and constructing another power plant.

An advanced power-generation concept is the organic Rankine cycle, usually termed binary-cycle plants (Milora and Tester, 1976). This system removes the heat from the wellhead fluid by a heat exchanger, and, using an organic fluid tailored to match the temperature of the geothermal fluid production, expands the organic fluid in a turbine that operates on a closed cycle (DiPippo, 1987). The advantages of the new binary plants are that the spent fluids are immediately reinjected back into the subsurface, the turbine generator units are modularized, and the units can be used on resources that are at intermediate temperatures. The binary plant also keeps mineral-laden source fluid separated from the turbine working fluid. A number of binary plant designs have been developed. Several have been installed in the state of Nevada, U.S.A., with considerable commercial success. This development means that

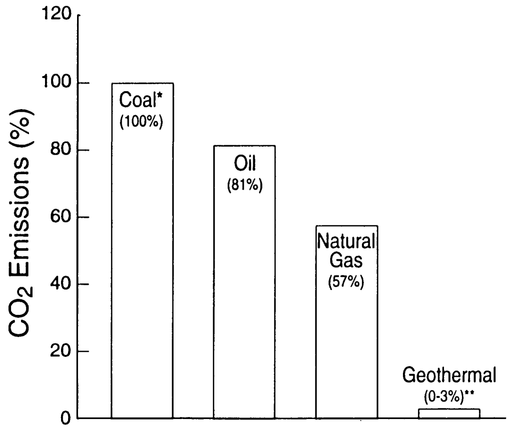

*Coal-fired plant ~500 lb/hr·MW(e) expressed as carbon.

** Newest geothermal plants use reinjection of gases and produce no CO_2.

FIG. 14. Relative carbon dioxide (CO_2) emissions from fossil-fuel and geothermal power plants, with coal taken as 100%.

electric power can be generated from small geothermal reservoirs having relatively low temperatures.

4.5 Environmental Protection

Early consideration of pollution abatement and environmental impacts has led to effective development of abatement practices, equipment, and monitoring and control methods (Hartley, 1980). The environmental sensitivity of the public and the regulatory mechanisms and environmental laws both in California, U.S.A., and in Japan have dictated that the geothermal projects now used as case histories for environmental aspects of geothermal development be given appropriate consideration. The impacts on land, water, and air have been reviewed by Weres (1988), and the nature and details of the control technologies are in place. Major problems were the handling of the large mass-flow rates of the fluids and the mitigation of the noncondensable gases, especially H_2S. The identification of potentially hazardous metals such as mercury, lead, arsenic, vanadium, and boron in the residuals from the power plants has aided the handling and disposal of these wastes. The early development and deployment of H_2S scrubbing technologies has allowed geothermal plants to meet stringent air-quality regulations. Drilling activity has been under strict regulation, and the use of H_2S abatement measures and noise-control mufflers has mitigated significant environmental impacts.

Geothermal development on a per-megawatt basis does not contribute as much to greenhouse gases as conventional thermal plants do (Fig. 14). A new reinjection technology (Hibara *et al.*, 1990) for the noncondensable gases has greatly reduced the total emissions from geothermal power plants. This technology, applied to conventional geothermal power plants, reduces their emissions to nearly zero, as in the case of binary plants.

5. ECONOMICS

The heat energy and electric power generated from geothermal resources are competitive with fossil- and nuclear-fueled heat and electricity in the United States and Japan where the private sector is conducting devel-

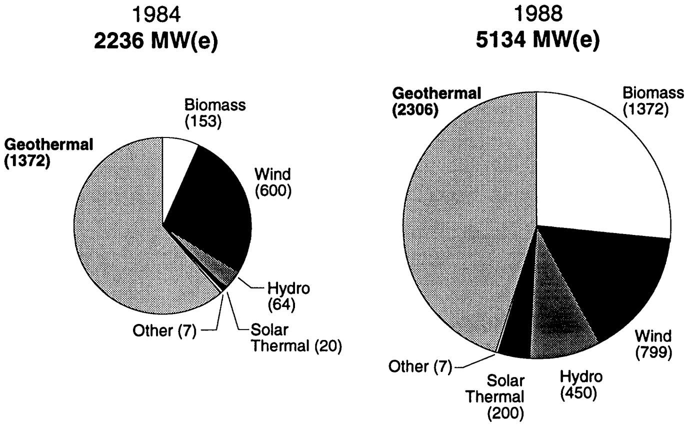

FIG. 15. Total U.S. utility power production from renewable resources.

opments. Outside these countries the national utility organizations usually have responsibility for development of electric generation, and it is difficult to judge whether the projects are competitive. However, in a country where fossil energy resources are scarce or absent, the development of geothermal energy is very attractive. In geothermal energy projects, the required technological level is lower, availability is higher, and the utility infrastructure is less complicated than that for alternative power-generation sources (Otte, 1991). Such geothermal energy projects, because they depend on local energy resources, are especially attractive from a balance-of-payments standpoint in many developing countries.

The economic cost analysis and evaluation of geothermal projects is straightforward and follows from the engineering cost estimating and financial considerations for more familiar fossil energy facilities (Tester, 1982; Blair *et al.*, 1982; Denno, 1990). In the United States the utility industry (Nola and Sioshansi, 1990) consistently produces more energy from geothermal resources than from other renewable alternatives (Fig. 15).

GLOSSARY

Aquifer: A subsurface, porous zone that contains groundwater and from which flow can be produced by a well.

Base Temperature: The maximum rock temperature reached at an unknown depth by the circulation of fluid, usually estimated by geochemical methods.

Binary Geothermal Power Plant: A two-fluid system that transfers heat from the geothermal fluid through a heat exchanger to an organic fluid that vaporizes to drive anorganic Rankine cycle turbine to produce electricity.

Direct Use: The use of geothermal (heat) energy for domestic and commercial applications that do not need conversion to electricity.

Fractured Reservoir: A reservoir of hot fluid with a circulating fluid system driven by geothermal energy occurring in fault and fissure zones that are porous and permeable.

Geothermal Anomaly: A significant deviation from the global average values in the temperature gradient or heat flow.

Geothermal Heat Pump: A heat pump that transfers energy to or from the Earth.

Geothermal Power: The rate at which geothermal energy is delivered; expressed as watts, W; in large amounts as megawatts; in direct use, MW(t); and converted to electricity, MW(e).

Geothermometers: Chemical analysis techniques based on chemical species in geothermal fluids that allow calculation of equilibrium temperatures as a function of their solubility; they indicate the base temperatures of a reservoir.

Geophysical Survey: A survey of such properties as resistivity, gravity, magnetism, and self-potential that probes the subsurface for differences in the rock properties and the presence of hot fluid.

Heat Flow: Geothermal heat flux at the Earth's surface, resulting from heat flowing from the interior.

Hydrothermal Convection System: A geothermal zone where heat is transferred to the near-surface by convective circulation of water or steam, not by the more common conductive mechanism.

Liquid-Dominated Reservoir: A hydrothermal reservoir where the subsurface conditions have caused the fluid to be present as a two-phase mixture of steam and liquid.

Magma: Mass of molten rock within the Earth's crust.

Magma Chamber: A body of magma (molten rock) in the Earth's crust.

Noncondensable Gas: A gas contained in a geothermal fluid that escapes when the fluid is flashed at the Earth's surface; examples are carbon monoxide, carbon dioxide, hydrogen sulfide, hydrogen, and helium.

Pluton: A body of rock that has solidified from magma below the Earth's surface.

Reservoir: A natural underground zone containing a significant accumulation of thermal energy.

Thermal Anomaly: Departure of a local subsurface temperature from the average surrounding temperatures.

Vapor-Dominated Reservoir: A hydrothermal reservoir where subsurface conditions have caused the fluids to be present as steam.

Works Cited

Agricola, G. (1556), *De Re Metallica*, translated by H. C. Hoover and L. H. Hoover (1950), New York: Dover, pp. 558–560.

Aguilera, R. (1980), *Naturally Fractured Reservoirs*, Tulsa, OK: Petroleum.

Anderson, D. (Ed.) (1992), *Geotherm. Resources Council Bull.* **21**, 137–173, 175, 176.

Armstead, H. C. H., Tester, J. W. (1987), *Heat Mining*, London: E. & F. F. Spon, Chap. 4.

Blair, P. D., Cassel, T. A. V., Edelstein, R. H. (1982), *Geothermal Energy: Investment Decisions and Commercial Development*, New York: Wiley.

Bodvarsson, G. S., Witherspoon, P. A. (1989), *Geotherm. Sci. Technol.* **2**, 1–68.

Boisdet, A., R. Ferrandes, C. Fouillac, F. Jaudin, J. Lemale, A. Menjoz, J. Rojas (1990), in *1990 International Symposium of Geothermal Energy, Transactions*, Vol. 14, Part I, Davis, CA: Geothermal Resources Council, pp. 55-61.

Bowen, R. (1989), *Geothermal Resources*, 2nd ed., London: Elsevier Applied Science.

Burgassi, P. D. (1987), *Geotherm. Resources Council Bull.* **16**, 3–18.

Burkman, J., Nickels, N. (1990), *Geotherm. Resources Council Bull.* **19**, 106–112.

Cooke, W. L. (1970), *Geothermics*, Special Issue No. 2, Vol. 2, Part 2, 1670–1675.

Costain, J. K., Glover, L., III, Sinha, A. K. (1980), *Trans. Am. Geophys. Union* **61**, 1–13.

Culver, G. (1989), in: P. J. Lienau, B. C. Lunis (Eds.), *Geothermal Direct Use Engineering and Design Guidebook*, Klamath Falls, OR: Oregon Institute of Technology, Geo-Heat Center, pp. 95–130.

Denno, K. (1990), *Engineering Economics of Alternative Energy Sources*, Boca Raton, FL: CRC, pp. 307–327.

DiPippo, R. (1987), *Geotherm. Sci. Technol.* **1**, 63–124.

Donaldson, I. G., Grant, M. A. (1978), *Geothermics* **7**, (2–4), 243–252.

Dorfman, M. H., Morton, R. A. (Eds.) (1985), *Geopressured-Geothermal Energy*, New York: Pergamon.

Einarsson, S. S. (1975), in: *Proceedings of the 2nd UN Symposium on the Development and Use of Geothermal Resources, San Francisco, May, 1975*, Vol. 3, Berkeley: LBL, Univ. of California, pp. 2364–2368.

Elder, J. (1981), *Geothermal Systems*, London: Academic.

Elders, W. A. (1991), *Geotherm. Resources Council Bull.* **21**, 194–203.

Elders, W. A., Pena, A. (Eds.) (1979), *Geology and Geothermics of the Salton Trough*, Univ. of California, Riverside, Museum Contribution No. 5.

Freeston, D. H. (1990), *Geotherm. Resources Council Bull.* **9**, 188–198.

Geiger, C. (1926), in: D. Anderson, B. Hall (Eds.) *Geothermal Exploration in the First Quarter Century*, Davis, CA: Geothermal Resources Council.

Goins, W. C., Jr., Sheffield, R. (1983), *Blowout Prevention*, 2nd ed., Houston, TX: Gulf.

Goldstein, N. E. (1988), *Geotherm. Sci. Technol.* **1**, 303–431.

Grant, M. A., Donaldson, I. G., Bixley, P. F. (1982), *Geothermal Reservoir Engineering*, New York: Academic.

Haenel, R., Staroste, E. (1988), *Atlas of Geothermal Resources in the European Community: Austria and Switzerland*, Luxembourg: Commission of European Communities.

Hardee, H. C. (1988), *Geotherm. Sci. Techol.* **1**, 165–224.

Hartley, R. P. (1980) in: J. Kestin (Ed.), *Sourcebook on the Production of Electricity from Geothermal Energy*, U.S. Dept. of Energy, Washington, DC: U.S. GPO, pp. 789–809.

Henley, R. W., Barton, P. B., Jr., Truesdell, A. H., Whitney, J. A. (1984), in: *Reviews in Economic Geology*, Vol. 1, El Paso, TX: Society of Economic Geologists.

Hibara, Y., Tazaki, S., Kuragasaki, M. (1990), *Geotherm. Sci. Technol.* **2**, 161–171.

Higo, M. (1990), in: *1990 International Symposium of Geothermal Energy, Transactions*, Vol. 14, Part 1, Davis, CA: Geothermal Resources Council, pp. 161–170.

Hill, D. G. (1985), *Geotherm. Resources Council Bull.* **14**, 19–22.

Huttrer, G. W. (1990), *Geotherm. Resources Council Bull.* **9**, 175–187.

Japan Geothermal Energy Association (1974), *Geothermal Energy Utilization in Japan*, Tokyo: Japan Geothermal Energy Association.

Japan Spa Association (1983), *Spas in Japan Guidebook*, Tokyo: Japan Spa Association.

Kelsey, J. R., Carson, C. C. (1987), *Geotherm. Sci. Technol.* **1**, 39–61.

Kestin, J. (Ed.) (1980), *Sourcebook on the Production of Electricity from Geothermal Energy*, U.S. Dept. of Energy, Washington, D.C.: U.S. GPO.

Koenig, J. B. (1992), in: C. Stone (Ed.), *Monograph on the Geysers Geothermal Field*, Geothermal Resources Council Special Report No. 17, pp. 7–18.

Lienau, P. J., Lunis, B. C. (Eds.) (1989), *Geothermal Direct Use Engineering and Design Guidebook*, Klamath Falls, OR: Oregon Institute of Technology, Geo-Heat Center.

Lippman, M., Manon, A. (1987), *Geotherm. Sci. Technol.* **1**, 1–38.

Mazzoni, A. (1954), *The Steam Vents of Tuscany and the Larderello Plant*, Bologna: Arti Grafiche Calderini.

McNitt, J. R. (1978), *Geothermics* **7**, 231–242.

Milora, S. L., Tester, J. W. (1976), *Geothermal Energy as a Source of Electric Power*, Cambridge, MA; MIT Press.

Morse, J. G., Thorson, L. D. (1978), *Geotherm. Resources Council. Trans.* **2**, 471–474.

Muffler, L. J. P. (Ed.) (1978), *Assessment of Geothermal Resources of the United States—1978,* U.S. Geological Survey Circular 790, Washington, DC: U.S. GPO.

Muffler, L. J. P. (1985), in: R. G. Perrine, W. G. Ernst (Eds.), *Energy: For Ourselves and Our Posterity,* Englewood Cliffs, NJ: Prentice-Hall, pp. 189–207.

Muffler, L. J. P., Cataldi, R. (1978), *Geothermics* **7**, (2–4), 53–9.

Mustoe, J. E. H. (1984), *An Atlas of Renewable (1991) Energy Resources: In the United Kingdom and North America,* Chichester, U.K.: Wiley, pp. 129–195.

Nakamura, K. (1977), *J. Volcanol. Geotherm. Res.* **2**, 1–16.

NEDO: New Energy Development Organization (1991), *Geotherm. Resources Council Bull.* **20**, 103–111.

New Zealand Electricity Department (1970), *Power from the Earth: The Story of the Wairaki Power Station,* 2nd ed., Wellington, New Zealand: New Zealand Government Printing Office.

Nola, S. J., Sioshansi, F. P. (1990), in: J. J. Hollander (Ed.), *Annual Review of Energy,* Vol. 15, Palo Alto, CA: Annual Reviews, pp. 99–119.

Norton, D. L. (1988), *Am. J. Sci.* **288**, 604–618.

Norton, D. L., Knight, J. (1977), *Am. J. Sci.* **277**, 937–981.

Otte, C. (1991), in: J. W. Tester, D. O. Wood, and N. A. Ferrari (Eds.), *Energy and the Environment in the 21st Century,* Cambridge, MA: MIT Press, pp. 755–761.

Pratsch, L. W. (1990), *Geotherm. Resources Council Bull.* **19**, 115–120.

Rafferty, K. (1989), in: P. J. Lienau, C. B. Lunis (Eds.), *Geothermal Direct Use Engineering and Design Guidebook,*Klamath Falls, OR: Oregon Institute of Technology, Geo-Heat Center, pp. 249–260.

Reed, M. J. (1983), *Assessment of Low-Temperature Geothermal Resources of the United States—1982,* U.S. Geological Survey Circular C. 0892, Washington, DC: U.S. GPO.

Rintoul, W. (1990), *Drilling Through Time,* Sacramento, CA: California Department of Conservation, Chap. 10.

Roberts, V. (Ed.) (1978), *Geothermal Energy for the Next 50 Years,* Electric Power Researach Institute Report No. ER-611-SR, pp. 4-1 to 4-100.

Rowley, J. C. (1982), in: L. M. Edwards, G. V. Chilingar, H. H. Rieke, III, W. H. Fertl (Eds.), *Handbook of Geothermal Energy,* Houston, TX: Gulf, pp. 44–176.

Rowley, J. (1988), in: *Proceedings of the International Symposium on Geothermal Energy, Beppu & Kumamoto, Japan,* Tsukuba, Japan: Geothermal Research Society of Japan, pp. 388–391.

Rybach, L., Muffler, L. J. P. (1981), *Geothermal Systems: Principles and Case Histories,* New York: Wiley.

Saito, S. (1992), in: M. V. Rao, J. Rowley (Eds.), *Proceedings of the 1992 Drilling Technology Symposium, Houston, Texas,* PD-40, New York: American Society of Mechanical Engineers, pp. 151–158.

Sekioka, M. (1990), in: *International Symposium on Geothermal Energy, Transactions,* Vol. 14, Pt. II, Davis, CA: Geothermal Resources Council, pp. 178–182.

Shale, L., Nickels, N. (1992), in: M. V. Rao, J. Rowley (Eds.), *Proceedings of the 1992 Drilling Technology Symposium, Houston, Texas,* PD-40, New York: American Society of Mechanical Engineers, pp. 117–122.

Smith, M. C. (1988), *Geotherm. Sci. Technol.* **1**, 125–163.

Tester, J. W. (1982), in: L. M. Edwards, G. V. Chilingar, H. H. Rieke, III, W. H. Fert (Eds.), *Handbook of Geothermal Energy,* Houston, TX: Gulf, pp. 471–586.

Truesdell, A. H. (1984) in: R. W. Henley, A. H. Truesdell, P. B. Barton, Jr. (Eds.), *Fluid-Mineral Equilibria in Hydrothermal Systems,* El Paso, TX: Society of Economic Geologists, Chap. 3.

Villa, F. P. (1975), in: *Proceedings of the 2nd UN Symposium on the Development and Use of Geothermal Resources, San Francisco, May, 1975,* Vol. 3, Berkeley: LBL, Univ. of California, pp. 2061–2064.

Ward, S. H., Ross, H. P., Nielson, D. L. (1981), *AAPG Bull.* **65**, 86–102.

Weres, O. (1988), *Geotherm. Sci. Technol.* **1**, 253–302.

White, D. E., Williams, D. L. (Eds.) (1975), *Assessment of Geothermal Resources of the United States—1975,* U.S. Geological Survey Circular 726, Washington, DC: U.S. GPO.

Wohletz, K., Heiken, G. (1992), *Volcanology and Geothermal Energy,* Berkeley, CA: Univ. of California Press.

Wolke, R. M., Deakins, K., Suter, R. L. (1990), *Geotherm. Resources Council Bull.* **19**, 138–143.

Wright, P. M., Nielson, D. L., Ross, H. P., Moore, J. N. Adams, M. C., Ward, S. H. (1989), *Geotherm. Sci. Technol.* **2**, 69–123.

Yamaguchi, F. (1992), *Geotherm. Resources Council Bull.* **21**, 237–240.

Yoshimura, Y., Yanagimoto, Y., Nakagome, O. (1985), *J. Jpn. Geotherm. Energy Assoc.* **25**, 167–194.

Zvulun, Y. (1989), *Geotherm. Resources Council Bull.* **18**, 5–8.

Further Reading

Edwards, L. M., Chilingar, G. V., Rieke, H. H., III, Fertl, W. H. (Eds.) (1982), *Handbook of Geothermal Energy,* Houston, TX: Gulf.

Goodman, L. J., Love, R. N. (Eds.) (1980), *Geothermal Energy Projects: Planning and Management,* New York: Pergamon.

Hayakawa, M. (1988), *Chikyu Netsugaku (Geothermy),* Tokyo: Tokai Univ. Press.

Kamata, M., Ozawa, T., Murakami, Y., Yoshida, M. (Eds.) (1985), *Chinetsu Ryutai no Kagaku (Chemistry of Geothermal Fluids*), Tokyo: Tokyo Univ. Press.

Mori, Y., Suyama, J. (1980), *Chinetsu Dokuhon (Geothermal Energy Primer),* Tokyo: Ohmu-sha.

Nasini, R. (1930), *I Sofione e i Lagoni della Toscana e la Industria Boracifera; Storia, Studi, Richerche Chimiche e Chimico-Fisiche Eseguite Principalmente nell' Ultimo Venticinquennio,* Rome: Tipografia Editrica Italia.

Yuhara, K., Seno, K. (1973), *Onsen Gaku (Research on Hot Springs; Balneology),* Tokyo: Chijinshokan.

GERMANIUM

W. L. HANSEN AND E. E. HALLER, *Lawrence Berkeley Laboratory, Berkeley, California, U.S.A.*

Introduction 219
1. **Physical and Chemical Properties** 220
2. **Electronic Properties** 220
2.1 Band Structure 220
2.2 Impurity and Defect Levels 221
2.3 Electronic Transport 223
2.3.1 Intrinsic Conduction 223
2.3.2 Extrinsic Conduction 223
3. **Optical Properties** 226
3.1 Linear Absorption Coefficient and Refractive Index 226
3.2 Photoconductivity 227
3.3 Photoluminescence 229
4. **Technology** 229
4.1 Production 229
4.2 Crystal Growth 230
4.3 Impurities and Defects 230
4.3.1 Solubility 230
4.3.2 Segregation Coefficient 231
4.3.3 Impurity Diffusion 232
4.3.4 Defects 233
5. **Applications** 233
5.1 Photodetectors 233
5.2 Nuclear-Radiation Detectors 233
5.3 Low-Temperature Ge Thermistors 234
5.4 Optics 234
5.5 Miscellaneous Uses 234
Glossary 235
Works Cited 235
Further Reading 237

INTRODUCTION

The properties of elemental germanium have been more thoroughly studied than those of any other member of the periodic table, with the possible exception of silicon. In order to review this extensive body of knowledge, various topics have been selected to illustrate the contribution of the study of germanium to semiconductor theory and technology.

The rise and fall of germanium as the commercially most important semiconductor are reviewed, followed by a description of its chemical and physical properties. Figures that summarize experimental and theoretical data have been selected, and they cover electronic and optical properties, technology, and applications. Germanium purification and crystal growth are described, and the more important current uses are listed.

Germanium was first isolated by C. Winkler in 1886 (Winkler, 1886) from the mineral argyrodite but had no commercial use until the second World War (WW II). Efforts to find a suitable material for radar detectors led to an intensive study of the semiconducting properties of germanium and silicon. These studies verified the importance of trace impurities in controlling electrical properties but did not emphasize the importance of crystalline perfection. Because of the different chemical natures of the semiconductors, silicon devices evolved into "cat's whisker" detectors and germanium in "high-voltage" rectifiers, both being made with polycrystalline material (Torrey and Whitmar, 1948).

Immediately after WW II Bell Telephone Laboratories established a solid-state division to begin a coordinated research effort by physicists, chemists, and engineers with the goal of producing a solid-state amplifier. This work led by 1948 to the invention of the point-contact transistor (Bardeen and Brattain, 1948), the field-effect transistor (Shockley and Pearson, 1948), and the junction transistor (Shockley, 1949). These devices were made from germanium single crystals.

Germanium devices remained dominant during the first decade after the invention of the transistor, mainly because of the ease of purification and crystal growth and low-temperature processing when compared to silicon (Biondi, 1958). However, once high-temperature synthesis and processing of sili-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

con were developed, germanium devices were rapidly replaced by silicon in almost all applications. This immediate dominance of silicon was due mainly to the outstanding surface-passivating and impurity-masking properties of thermally grown SiO_2 films on silicon and to the larger band gap allowing higher-temperature device operation. The thermal oxide GeO_2 on germanium is water soluble and unstable in air.

The crust of the earth contains about 8×10^{-4} wt % germanium, which is 100 times more abundant than silver, but as there are no rich ores, the supply is very price sensitive. A typical price for the past 30 years has been about US $600 per kg of pure polycrystalline Ge (1992 dollars).

1. PHYSICAL AND CHEMICAL PROPERTIES

Elemental germanium is brittle and harder than glass, a consequence of its purely covalent tetrahedral bonding. The metallic luster is due to the high refractive index and the large free-carrier concentration generated by visible light. The electrical conductivity of pure germanium increases with temperature as a result of the increase in the concentration of thermally generated electrons and holes. Germanium crystallizes from the melt ($T_{melt}=936$ °C) in the diamond lattice structure with a lattice constant of 5.65 Å. The diamond lattice is a very open structure, resulting in the expansion of germanium by 4.9% upon freezing.

Chemically, germanium reacts as would be predicted by its location in the periodic table: a column IV element located between silicon and tin. The absence of clearly distinguishing chemical properties, together with its high cost, accounts for the lack of interest by chemists in studying germanium before its unusual electrical properties were discovered.

Germanium is stable in air up to about 500 °C. It is not attacked by HCl, HF, H_2SO_4, or strong alkalis. Germanium is dissolved slowly by HNO_3, H_2O_2, and sodium hypochlorite (laundry bleach); dissolves rapidly in mixtures containing HF and a strong oxidizer, e.g., HF-HNO_3, HF-H_2O_2; and, when hot, reacts exothermally with Cl_2 to produce $GeCl_4$. Elemental germanium can be melted with negligible reaction in graphite or fused-silica containers in a controlled atmosphere (e.g., N_2, H_2, Ar, He, vacuum). The purity of Ge can be found by measuring its resistivity either at room temperature or at liquid nitrogen temperature (77 K). This simplicity of analysis led to rapid advances in purification (Tannenbaum, 1959). The use of highly purified containers of graphite or silica, together with a controlled atmosphere of H_2 that has been passed through a palladium purifier, has led to germanium crystals of a purity unattained by any other elemental substance. Large, highly perfect germanium single crystals can be grown with a total electrically active impurity concentration of less than 10^{10} cm^{-3}, or, in chemists' terminology, less than 10^{-7} ppm.

A summary of the major properties of germanium is given in Table 1. Extensive data tabulations have been published that attempt a complete survey of the germanium literature (Madelung, 1982; Madelung and Schulz, 1987, 1989).

2. ELECTRONIC PROPERTIES

2.1 Band Structure

The band structure of germanium is shown in Fig. 1. The maximum of the valence band is located at the center of the Brillouin zone, while the four conduction-band minima are located at the L points along the four [111] directions. Semiconductors with the maximum of the valence band located at a different value of k than the minimum (minima) of the conduction band are called indirect band-gap semiconductors. They are characterized by long minority-carrier lifetimes and very inefficient photon emission upon band-to-band electron-hole recombination.

The valence-band edge is derived from the $p_{3/2}$ and $p_{1/2}$ states of the free atom with the $p_{3/2}$ states at the higher energy. The top of the valence band consists of a light- and a heavy-hole band that are degenerate at $k=0$ (the center of the Brillouin zone, called the Γ point), and a third band that is split off by 0.29 eV. The direct and indirect gaps decrease as the temperature is increased, as shown in Fig. 2.

Table 1. Properties of germanium.

Property	Value
Atomic number	32
Atomic weight	72.59
Natural isotopic composition (%)	
70	20.4
72	27.4
73	7.8
74	36.6
76	7.8
Density (25 °C, g cm^{-3})	5.3234
Atomic density (atoms cm^{-3})	4.416×10^{22}
Crystal structure	Diamond cubic
Lattice parameter at 300 K (Å)	5.64613
Melting point (mp) (°C)	937.4
Boiling point (°C)	2830
Vapor pressure at mp (atm)	2×10^{-13}
Volume contraction on melting (%)	4.7
Heat of fusion (kJ mol^{-1})	36.945
Heat of sublimation (kJ mol^{-1})	374.5
Thermal expansion coefficient at 300 K (10^{-6} K^{-1})	5.9
Thermal conductivity at 300 K (W cm^{-1} K^{-1})	0.6
Energy gap at 300 K (eV)	0.664
Energy gap at 0 K (eV)	0.744
Elastic stiffness constants at 25 °C (10^{11} dyn/cm^{2})	
c_{11}	12.40
c_{12}	4.13
c_{44}	6.83
Intrinsic resistivity at 25 °C (Ω cm)	50
Intrinsic electron concentration at 300 K (cm^{-3})	1.38×10^{13}
Temperature dependence of intrinsic concentration (cm^{-3})	$1.76\times10^{16}T^{3/2}\times\exp(-0.785/2kT)$
Intrinsic electron mobility at 300 K (cm^{2} V^{-1} s^{-1})	3900
Intrinsic hole mobility at 300 K (cm^{2} V^{-1} s^{-1})	1800
Effective mass, m/m_0	
electrons: longitudinal	1.64
transverse	0.082
holes: light	0.044
heavy	0.28
Refractive index, 2.5–150 μm	4.0
Dielectric constant	16.0

2.2 Impurity and Defect Levels

A large number of impurities, defects, and their combinations (also called complexes) are known to produce energy levels in the forbidden gap of germanium (Fig. 3). These gap states are conveniently characterized as being donor or acceptor type, shallow (near the corresponding band edges) or deep (far from band edges). As in other semiconductors, there exist single, double, and multivalent centers (e.g., double acceptor Be). The energies of the ground states and the excited states of shallow donors [Fig. 4(a)] and shallow acceptors [Fig. 4(b)] can be calculated accurately by advanced effective-mass theory (Faulkner, 1969; Baldereschi and Lipari, 1973, 1974).

The shallow levels that are created by the column III and V impurities and lithium have been studied with a wide variety of techniques. In the temperature range where the thermal energy is sufficient to ionize shallow impurities but is too small to create electrons and holes across the band gap, the free-carrier

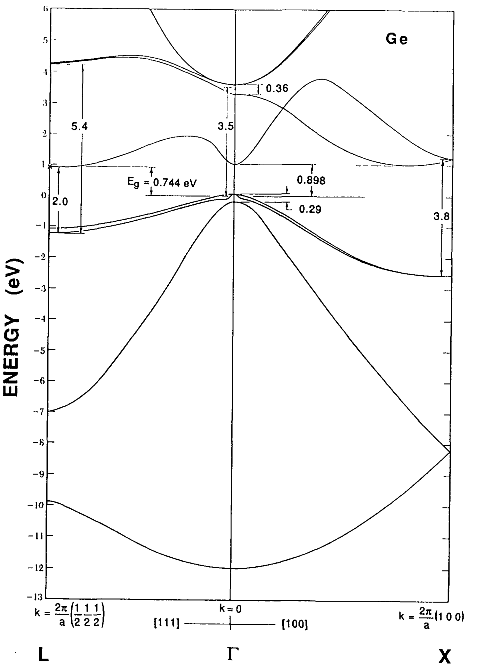
Ge
ENERGY (eV)
0.36
5.4
3.5
Eg = 0.744 eV
0.898
0.29
2.0
3.8
k = 2π/a (1/2 1/2 1/2)
[111]
k = 0
[100]
k = 2π/a (1 0 0)
L
Γ
X

FIG. 1. Band structure of germanium. The low-temperature indirect gap is 0.744 eV, and the direct band gap is 0.898 eV (after Kittel, 1986).

concentration corresponds to the difference of majority and minority dopant concentrations, the so-called net dopant concentration. The difference is the consequence of the compensation of donors by acceptors and vice versa. Compensation results from the minimum-energy principle: an electron of a donor impurity occupies an acceptor level that lies much lower in energy rather than being raised into the conduction band. Variable-temperature Hall-effect measurements (Yee *et al.*, 1974) give accurate free-carrier concentration values (Fig. 5). The free-carrier concentration is an accurate measure of the net dopant concentration in the temperature range defined above (also called exhaustion or saturation range).

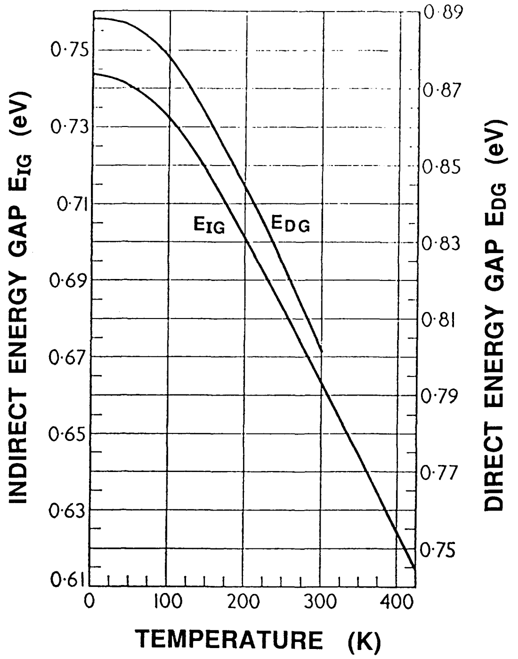

FIG. 2. Indirect (E_{IG}) and (E_{DG}) energy gaps vs temperature (after McLean, 1960).

Low-temperature optical absorption spectroscopy of lightly doped crystals reveals the hydrogenic bound excited-state spectrum of shallow donors and acceptors (Ramdas and Rodriguez, 1981). The development of ultrapure crystals required a highly sensitive and high-resolution characterization technique. Photothermal-ionization spectroscopy (PTIS), a two-step ionization process, can generate photoconductivity spectra with linewidths of the order of 10 μeV in crystal specimens containing less than 10^6 centers/cm^3. PTIS has led to the discovery of a large number of

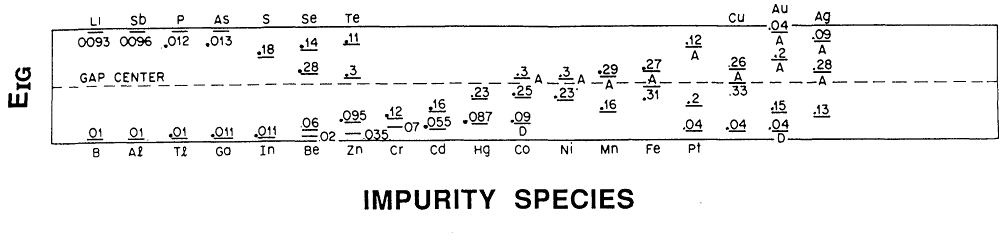
EIG
GAP CENTER
Li
Sb
P
As
S
Se
Te
Cu
Au
Ag
0093
0096
.012
.013
.18
.14
.28
.11
.3
B
Al
Tl
Ga
In
Be
Zn
Cr
Cd
Hg
Co
Ni
Mn
Fe
Pt
IMPURITY SPECIES

FIG. 3. Impurity energy levels in the forbidden gap of germanium (after Sze, 1981).

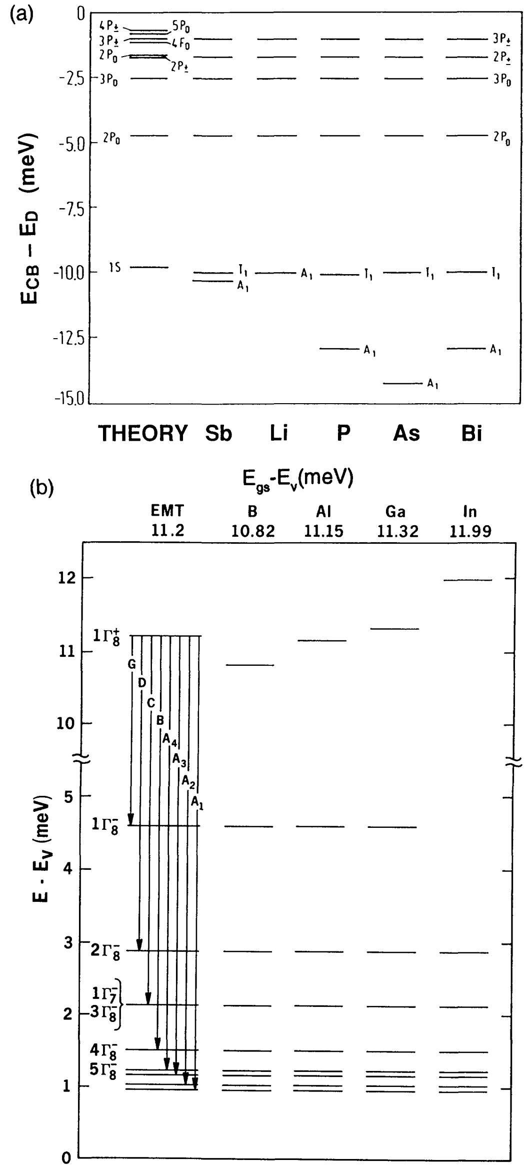
(a)
0
-2.5
-5.0
-7.5
-10.0
-12.5
-15.0
$E_{CB} - E_D$ (meV)
$4P_\pm$
$5P_0$
$3P_\pm$
$4F_0$
$2P_0$
$2P_\pm$
$3P_0$
$2P_0$
1S
T_1
A_1
THEORY
Sb
Li
P
As
Bi
(b)
E_{gs}-E_v(meV)
EMT 11.2
B 10.82
Al 11.15
Ga 11.32
In 11.99
$E - E_V$ (meV)
$1\Gamma_8^+$
G
D
C
B
A_4
A_3
A_2
A_1
$1\Gamma_8^-$
$2\Gamma_8^-$
$1\Gamma_7^-$
$3\Gamma_8^-$
$4\Gamma_8^-$
$5\Gamma_8^-$
12
11
10
5
4
3
2
1
0

FIG. 4. (a) Energy levels of shallow donors with the theoretical and experimental $2P_\pm$ levels aligned (after Faulkner, 1969). (b) Energy levels of shallow acceptors with experimental and theoretical levels aligned for the $2\Gamma_8$ levels (after Baldereschi and Lipari, 1973, 1974, 1977).

shallow-level centers that contain hydrogen bound to a neutral or a multivalent impurity (Haller *et al.*, 1981; Haller, 1986).

The very deep impurities, e.g., the transition metals, have small optical cross sections and very low solubility and are more conveniently, but less accurately, studied by deep-level transient spectroscopy (DLTS) (Pearton, 1982) (Fig. 6).

2.3 Electronic Transport

2.3.1 Intrinsic Conduction. The electrical conductivity is proportional to the product of the number of charge carriers and their mobility. In an intrinsic semiconductor one observes bipolar conduction by electrons and holes. The conductivity is given by

$$\sigma = ne\mu_e + pe\mu_p, \qquad (1)$$

where n and p are the electron and hole concentrations and μ_e and μ_p are their respective mobilities. The concentration of charge carriers obeys a mass action law so that at a given temperature $np = n_i^2 = \text{constant}$:

$$np = 4(2\pi kT/h^2)^3(m_e m_h)^{3/2}e^{-E_g/kT}, \qquad (2)$$

where m_e and m_h are the electron and hole effective masses and n_i is the intrinsic carrier concentration. This relation holds over a wide temperature range, as is shown experimentally for germanium in Fig. 7. The strong temperature dependence of n_i dominates σ, and the weaker power dependences of the mobilities do not play a significant role.

2.3.2 Extrinsic Conduction. When germanium is doped with shallow donors or acceptors (column III or V elements), n or p becomes much larger than n_i over a wide temperature range, and the conductivity will be dominated by the temperature dependence of the mobility. Three major mobility regimes

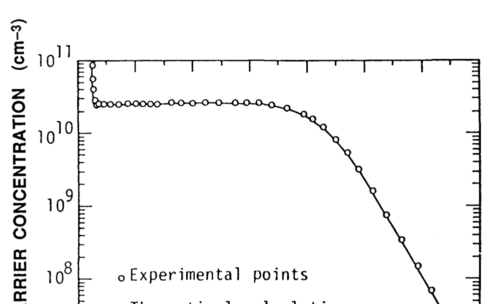

FIG. 5. Free hole concentration in an ultrapure Ge sample as measured by Hall effect (after Yee *et al.*, 1974).

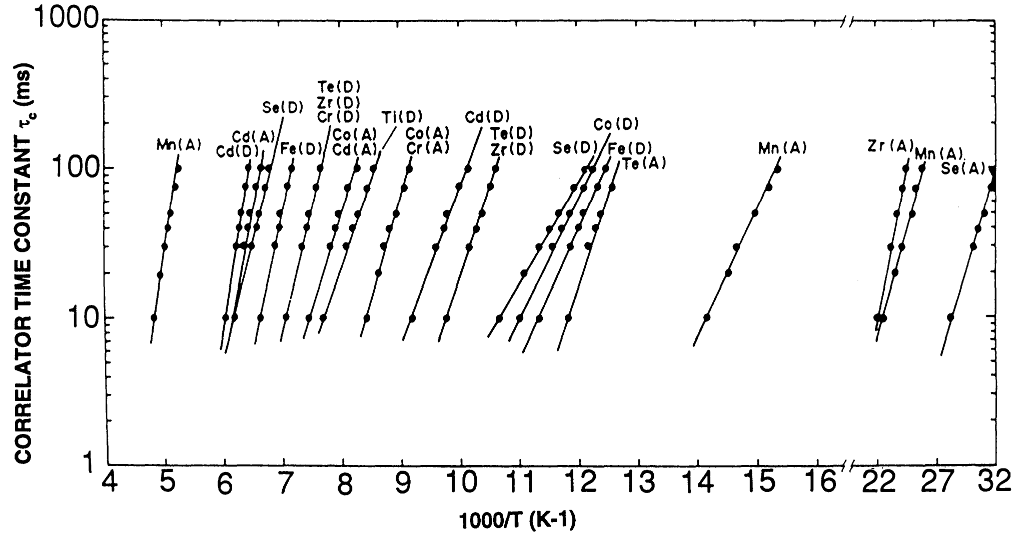

FIG. 6. Detrapping time constants of deep-level impurities thermally diffused into germanium, measured by deep-level transient spectroscopy (after Pearton, 1982).

are observed. At high temperature, lattice phonon scattering dominates:

$$\mu_{\text{phonon}} \propto T^{-3/2}. \tag{3}$$

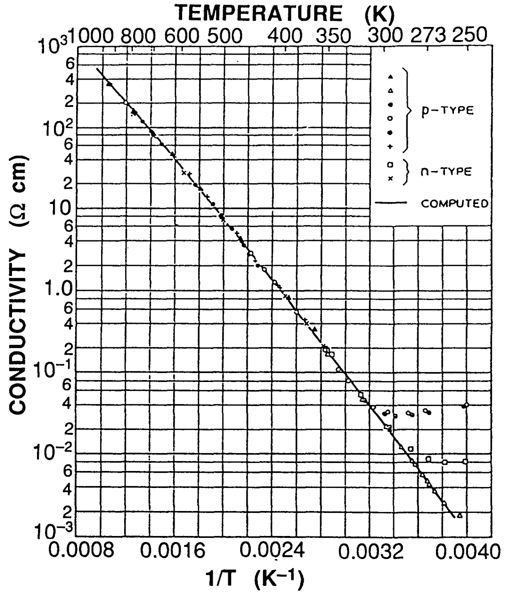

FIG. 7. The conductivity of intrinsic Ge vs temperature (after Morin and Maita, 1954).

At lower temperature, ionized-impurity scattering limits the mobility:

$$\mu_{\text{ionized imp.}} \propto T^{3/2}/Z^2N, \tag{4}$$

with Z=number of charges on the ionized impurity and N=concentration of ionized centers. In crystals with very low compensation (several orders of magnitude larger majority than minority dopant concentration) one observes a low-temperature regime in which neutral-impurity scattering dominates:

$$\mu_{\text{ionized imp.}} \propto 1/N, \tag{5}$$

with N=concentration of neutral centers. The total mobility is obtained by adding the inverse mobilities (Matthiessen's rule):

$$\mu_{\text{total}} = \left(\frac{1}{\mu_{\text{phonon}}} + \frac{1}{\mu_{\text{ionized imp.}}} + \frac{1}{\mu_{\text{neutral imp.}}} \right)^{-1}. \tag{6}$$

Figure 8 shows the hole mobilities in two low-compensation, Ga-doped Ge crystals over a wide temperature range. The onset of ionized-impurity scattering is clearly seen as an inflection at 60 K. The reason why the mobility continues to rise at lower temperatures lies in the fact that the Ga acceptors freeze out, i.e.,

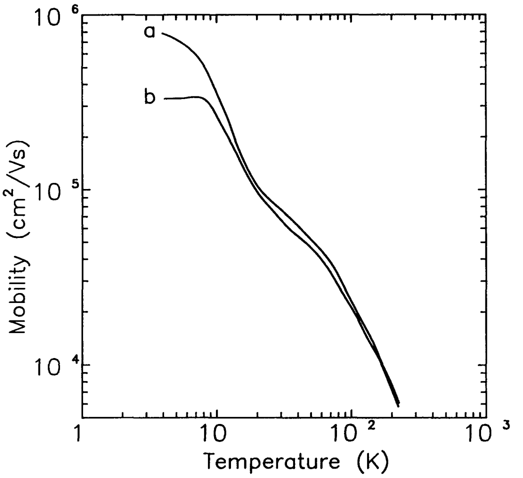

FIG. 8. Hall mobility in two moderately doped Ge:Ga crystals.

they become neutral when they bind a hole at temperatures <20 K. Below ~7 K the mobility becomes constant and is dominated by neutral impurity scattering. In moderately doped crystals that are highly compensated, a large fraction of the acceptors and donors remain ionized, and the mobility does not rise because of freezeout. In such material one observes the lattice-scattering regime at high temperatures and the ionized-impurity scattering regime at low temperatures.

At large impurity concentrations ($>10^{15}$ cm^{-3}) the conductivity is dominated by carrier hopping at low temperatures and at high doping ($>2\times10^{17}$ cm^{-3}), an impurity band forms, and the germanium exhibits metallic conduction (i.e., conduction persists down to $T=0$ K). These various conduction regimes are illustrated by experimental data in Fig. 9 for 16 compensated Ge samples with varying acceptor concentrations. In the most lightly doped sample (No. 1, $N_A=1.7\times10^{14}$ cm^{-3}) hopping conduction is observed below 5 K, and it is rather weak ($\rho\simeq10^9$ Ω cm). In the extreme of high doping concentration (Nos. 15, 16) the resistivity stays practically constant at $\sim1.3\times10^{-2}$ Ω cm over the full temperature range. These samples conduct like a metal. The effect of impurity concentration on minority electron mobility in *p*-type Ge at 300 K is shown in Fig. 10.

The discussion of electrical transport has so far been limited to low electric fields that do not heat up the free carriers to energies larger than kT. At high electric fields the carriers become hot and give up their kinetic energy by generating optical phonons. The carrier drift velocity is observed to saturate at

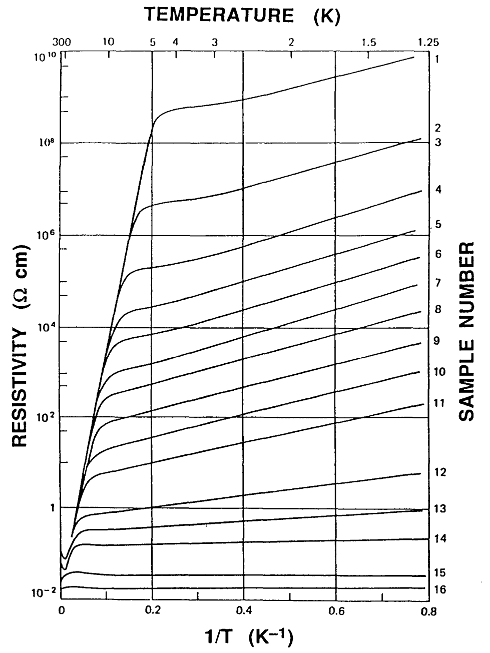

FIG. 9. Resistivity of *p*-type germanium with compensation K ($=N_D/N_A$)$=0.4$ (after Fritzsche and Cuevas, 1960). The concentrations (in cm^{-3}) of acceptors for the various curves are as follows: 1, 7.5×10^{14}; 2, 1.4×10^{15}; 3, 1.5×10^{15}; 4, 2.7×10^{15}; 5, 3.6×10^{15}; 6, 4.9×10^{15}; 7, 7.2×10^{15}; 8, 9.0×10^{15}; 9, 1.4×10^{16}; 10, 2.4×10^{16}; 11, 3.5×10^{16}; 12, 7.3×10^{16}; 13, 1.0×10^{17}; 14, 1.5×10^{17}; 15, 8.8×10^{17}; 16, 1.35×10^{18}.

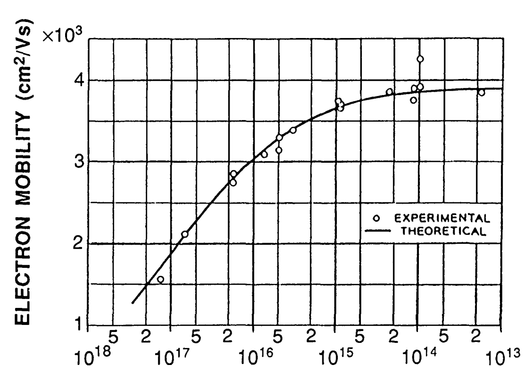

FIG. 10. Drift mobility of minority electrons in *p*-type Ge at 300 K as a function of acceptor concentration. The solid line is calculated (after Prince, 1953).

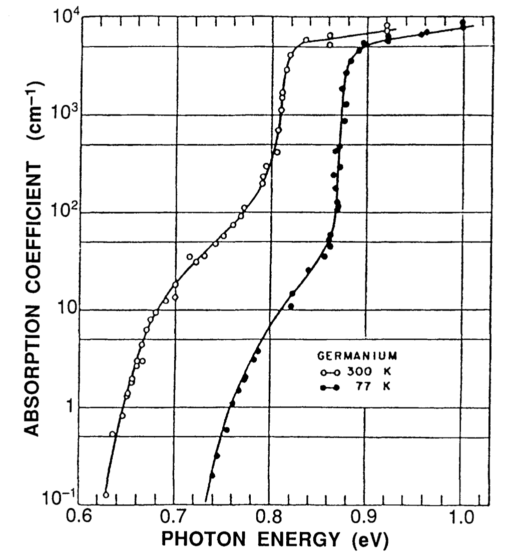

FIG. 11. Linear absorption coefficient of Ge near the fundamental absorption edge vs photon energy at 77 and 300 K.

values close to 10^7 cm s^{-1} for electric fields larger than 10^4 V cm^{-1}.

Transport phenomena become very complicated when electric and magnetic fields are present. It has been demonstrated that under appropriate conditions a population inversion of free holes between the light- and the heavy-hole bands can be achieved with crossed electric and magnetic fields at low temperatures. Such Ge crystals can be induced to lase in the far-infrared range of the electromagnetic spectrum (Gornik, 1991).

3. OPTICAL PROPERTIES

3.1 Linear Absorption Coefficient and Refractive Index

All semiconductors exhibit a very rich variety of optical properties, and Ge is no exception. Pure crystals cooled to low temperatures contain such small concentrations of free carriers that their effects can be ignored. The most striking feature for these conditions is the fundamental absorption edge, the most important feature in a plot of the linear absorption coefficient as a function of photon energy. The absorption edge occurs at an energy corresponding to the minimum band gap (Fig. 11). Because Ge has an indirect band structure, optical transitions of electrons from the top of the valence band at the Γ point of the first Brillouin zone to one of the four conduction-band minima at the *L* points require assistance of a phonon. This three-particle process is clearly less probable than a direct band-gap transition. At photon energies > 0.8 eV, direct transitions can occur, and the linear absorption coefficient rapidly rises to large values.

The linear absorption coefficient is defined by the equation

$$I = I_0 \exp(-\alpha x); \qquad (7)$$

I_0 and I are the photon beam intensities before and after passage of the beam through a semiconductor slab of thickness x. The linear absorption coefficient is typically given in units of cm^{-1}. This simple equation does not take into account external or internal reflections caused by a refractive index different from unity.

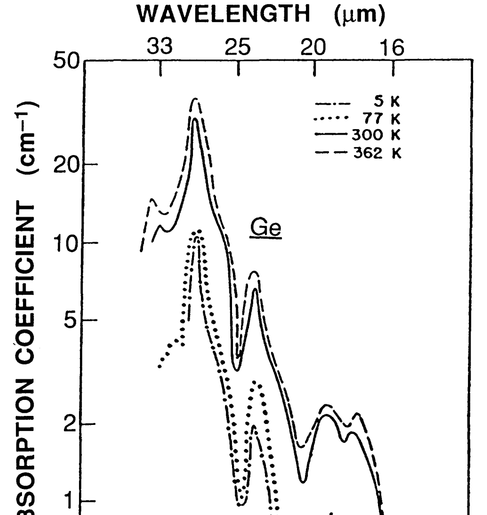

FIG. 12. Optical absorption coefficient α in the far infrared as a function of photon wave number (after Collins and Fan, 1954).

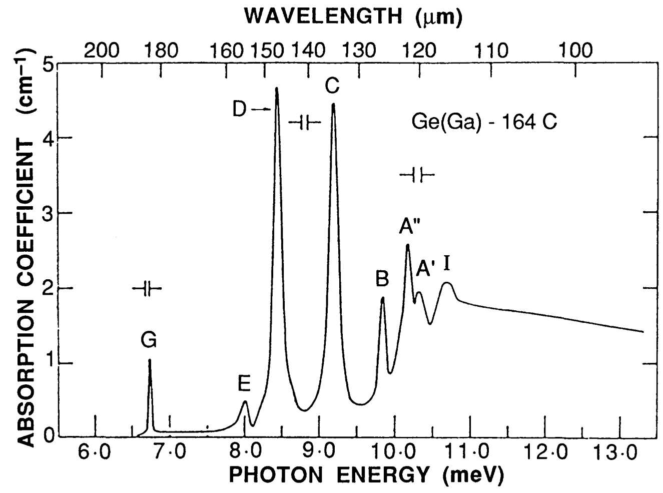

FIG. 13. The absorption spectrum of Ge doped with 2×10^{14} cm^{-3} Ga (after Jones and Fisher, 1965).

Below the fundamental absorption edge pure Ge is transparent and is used for a large number of infrared optical components. As a covalently bonded material, Ge is not infrared active, though at low photon energies lattice (multiphonon) absorption gives rise to a broad absorption feature reaching from approximately 25 to 90 meV (wave number $\tilde{\nu}=200$ to 700 cm^{-1}, $\lambda=50$ to 14 μm), with α reaching a peak value of 10 cm^{-1} at $h\nu=44$ meV ($\tilde{\nu}=350$ cm^{-1}, $\lambda=28.6$ μm) at 77 K (Fig. 12).

At even lower photon energies one encounters effects of shallow dopants. The absorption spectrum of a Ge crystal moderately doped with Ga acceptors is shown in Fig. 13. The narrow peaks correspond to hole transitions from acceptor ground states to hydrogenic bound excited states.

Table 2. Refractive index and extinction coefficient k as functions of photon energy E (after Potter, 1985).

E (eV)	n	k
525.3	0.99741	9.8×10^{-4}
100	—	2.57×10^{-2}
50	—	4.52×10^{-2}
10	0.93	0.86
5	1.394	3.197
4	3.905	3.336
3	4.082	2.145
2	5.588	0.933
1	4.325	8.09×10^{-2}
0.8	4.275	5.67×10^{-3}
0.66	—	6.73×10^{-5}
0.60	—	6.58×10^{-7}
0.59	4.104	—
0.20	4.011	—
0.099	4.003	—
0.062	3.999	3.98×10^{-4}
0.031	3.983	8.5×10^{-4}
0.0099	4.007	—
0.0025	4.004	4.77×10^{-3}

For optical applications of Ge, it is most useful to discuss the refractive index n and the extinction coefficient k. They are related to the real (ϵ') and the imaginary (ϵ'') parts of the dielectric constant:

$$\epsilon=\epsilon'+i\epsilon'', \tag{8}$$

$$\epsilon'(\nu)=n^2(\nu)-k^2(\nu), \tag{9}$$

$$\epsilon''(\nu)=2n(\nu)k(\nu). \tag{10}$$

The linear absorption coefficient and the wavelength λ are related to k:

$$k=\alpha\lambda/4\pi. \tag{11}$$

An extensive tabulation of n and k over a very wide photon energy range can be found in the literature (Potter, 1985). A reduced set of values is reproduced in Table 2.

3.2 Photoconductivity

In a semiconductor one distinguishes between two classes of optical absorption processes, both leading to an increase in the free-carrier concentration. Intrinsic photoconductivity is due to the excitation of an electron across the band gap, while extrinsic photoconductivity is due to the ionization of a neutral donor or acceptor in a cold crystal.

Diode structures are typically used for intrinsic photocarrier generation and collec-

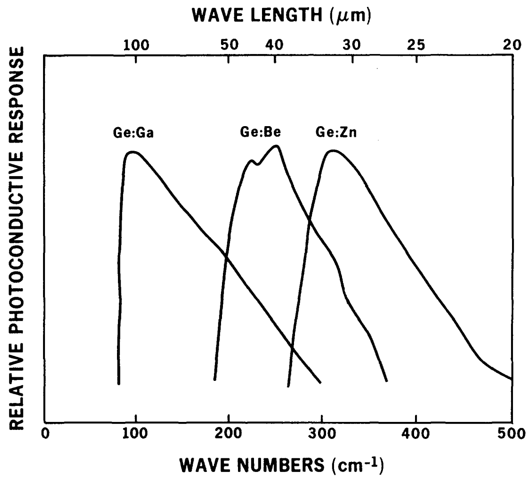

FIG. 14. Responsivity of extrinsic Ge photoconductors doped with various impurities (Haller, 1985).

tion. Ge photodiodes are widely used for infrared photoluminescence studies. They are sensitive to photons with energies larger than the indirect band gap ($E_{\mathrm{IG}}=744$ meV). Ge photodiodes are cooled, typically to liquid-nitrogen temperature (77 K), in cases where the large dark current caused by room-temperature thermal generation interferes with photodetection.

Photons with very large energies such as x rays and γ rays produce large numbers of photocarriers per photon. Through the photoeffect the energy of the incoming photons is transferred to a Ge core electron. One electron-hole pair is generated per 3 eV of the energy of the x-ray or γ quantum (Haller and Goulding, 1981).

Extrinsic photoconductivity in doped Ge has been applied to far-infrared detection. Moderately doped Ge crystals ($[\mathrm{Ga}]\sim 10^{14}$ cm^{-3}) with two Ohmic contacts are operated in the photoconductor mode. Such devices have to be cooled to liquid helium temperature in order to suppress thermal ionization of the shallow dopants. Typical dopants are Ga, Be, and Cu. Extrinsic photoconductors doped with these impurities can detect photons in an energy range from 10 meV ($\tilde{\nu}=80$ cm^{-1}, $\lambda=125$ μm) to several hundred meV ($\tilde{\nu}=1000$ cm^{-1}, $\lambda=10$ μm) (Haller, 1985; Haegel and Haller, 1986). Application of uniaxial stress along the [100] axis reduces the hole binding energy to half its unstressed value and shifts the detection limit to 5 meV (40 cm^{-1}, 250 μm) (Hall, 1962). Extrinsic Ge photoconductors are the most sensitive photon detectors beyond $\lambda=40$ μm. Figure 14 shows the individually normalized response curves of several extrinsic Ge photoconductors.

A two-step ionization process consisting of excitation of an impurity-bound carrier from the ground state to a bound excited state followed by thermal ionization from the excited state to the nearest band was discovered by Lifshits and Nad (1965). The resulting free carriers can be detected with very high sensitivity. This method has been extensively applied to the study of ultrapure Ge (Haller, 1978, 1986; Haller *et al.*, 1981). It combines high energy resolution with superb sensitivity. In ultrapure Ge, linewidths as small as 10 μeV have been obtained from hydrogenic spectra of shallow-donor complexes *D*(H,O) and *D*(Li,O) (Haller and Falicov, 1978; Haller *et al.*, 1987; Navarro *et al.*, 1988). Figure 15 displays the photothermal ionization spectroscopy results obtained with an ultrapure Ge crystal containing the shallow-acceptor complex *A*(H,C) (Kahn *et al.*, 1987). The four strongest lines are the *D* and *C* lines of the acceptors Al and $A(\mathrm{D,C})_2$. Each acceptor has a series of lines. The line spacings between

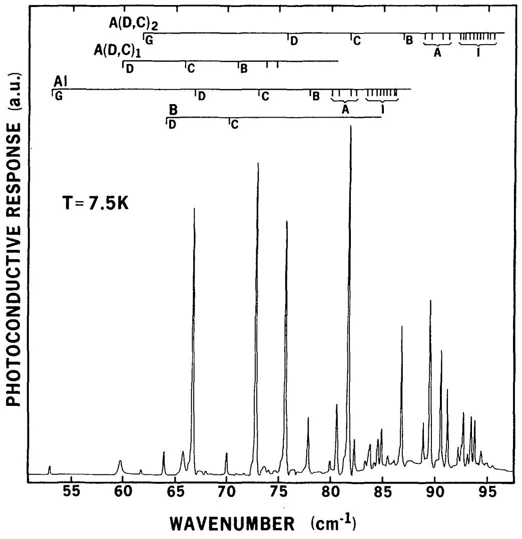

FIG. 15. Photothermal-ionization line spectrum of the shallow deuterium-carbon acceptor complex *A*(D,C), of Al, and of traces of B. The fact that two line series $A(\mathrm{D,C})_1$ and $A(\mathrm{D,C})_2$ belong to one acceptor is most unusual and is due to a splitting in the ground state of this acceptor complex. The sample contains approximately 10^9 acceptors (after Kahn *et al.*, 1987).

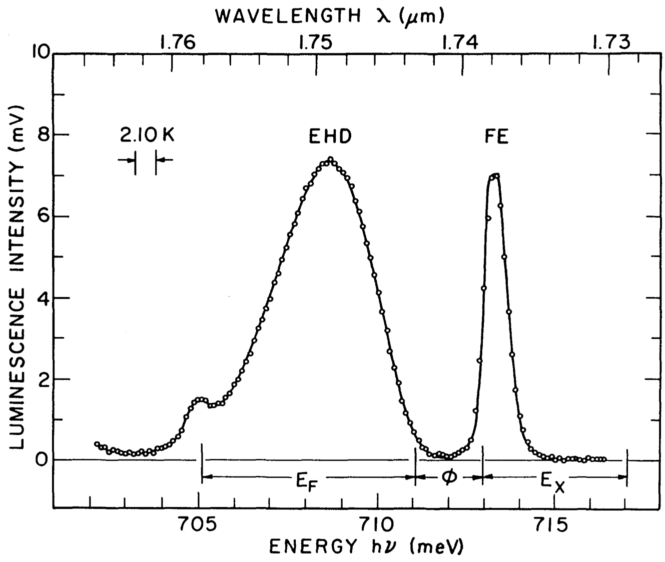

FIG. 16. Photoluminescence spectrum of optically pumped pure Ge at low temperature. The free-exciton (LA-assisted) and electron-hole-drop peaks are clearly visible (after Kelso, 1979).

specific pairs of lines are identical for all acceptors because the long-range Coulomb potential of the negatively charged acceptor core is the same for all shallow acceptors. The differences in the ground-state energies shift the individual line series to different energies. For ease of identification, the positions of the peaks in each acceptor series have been connected with a horizontal line that is labeled with the acceptor symbol.

3.3. Photoluminescence

Absorption of a photon followed by the emission of a photon is called photoluminescence (PL). The indirect-gap band structure of Ge makes PL a rare process because lattice phonons or impurities are required for momentum conservation. The PL photon originates from the decay of a free or bound exciton, an electron and a hole forming a hydrogenlike bound system. A large number of impurity-bound excitons have been studied in pure and doped Ge (Skolnick, 1989; Thewalt *et al.*, 1987).

A new state of matter, the electron-hole liquid, was predicted (Keldysh, 1968) and experimentally verified (Wolfe *et al.*, 1975; Hensel *et al.*, 1977; Jeffries and Keldysh, 1983) in pure Ge. This quantum liquid consists of condensed excitons. The electron-hole liquid is unstable because of electron-hole recombination, and it has to be regenerated continuously by a laser pumping electrons across the band gap. Figure 16 shows the PL spectrum of a pure Ge crystal at low temperature optically pumped by a laser. The free-exciton peak (FE) corresponds to the band-gap energy (744 meV) reduced by the exciton binding energy of 4 meV and the energy of the phonon (LA) assisting the transition (E_{LA}=27 meV). The electron-hole drop peak (EHD) appears at 709 meV, lower than the FE peak by the EHD binding energy.

Recently, isotopically separated Ge has become available (see Table 1). PL studies of such crystals allow the determination of the dependence of the band gap and the phonon energies on isotope mass (Fuchs *et al.*, 1991; Davis *et al.*, 1992). It can be expected that phonon transport and imaging as well as a number of special devices (see Sec. 6) will benefit from isotopically controlled crystals.

4. TECHNOLOGY

4.1 Production

Germanium concentrates are recovered as a by-product of zinc, copper, and lead refining. The concentrates are converted to oxides or germanates and treated with HCl to produce $GeCl_4$. The simple and selective chemistry involved in converting $GeCl_4$ (bp 83 °C) into high-purity germanium contributed to the early success of germanium electronics. To produce polycrystals with less than 1 ppm total impurities, the $GeCl_4$ is distilled in the

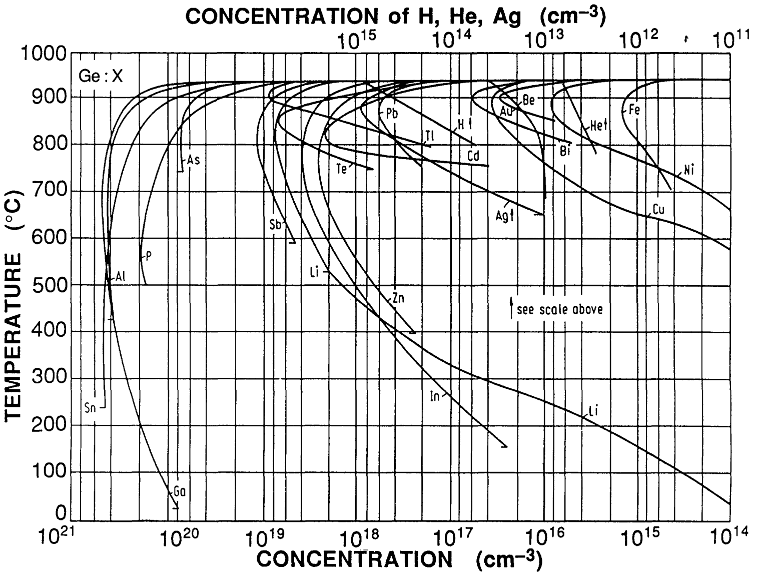

FIG. 17. Solid solubility of impurities in germanium illustrating retrograde solubility (after Trumbore, 1960).

presence of Cl_2, fractionally distilled in silica vessels, mixed with H_2O to form GeO_2, and finally reduced in H_2 and melted in graphite boats (Bridgers *et al.*, 1958). Any processing scrap, saw dust, etc. can be easily converted to $GeCl_4$ to reenter the process stream.

4.2 Crystal Growth

All electronic applications require germanium in single-crystal form and often demand ultrahigh purity. Germanium can be purified for certain impurities to better than 10^{-6} ppm by zone refining (Pfann, 1966) in silica or graphite boats, a process that was developed originally to produce high-purity germanium and now has general application in the fine chemicals industry. The zone-refined polycrystal is used as a starting charge for single-crystal growth.

Single-crystal germanium is produced by the Czochralski method (Czochralski, 1918; Teal and Little, 1950; see also CRYSTAL GROWTH), which has replaced the earlier zone-leveling process (Pfann, 1952). In the Czochralski method (CZ) an oriented seed crystal is dipped into a melt and withdrawn with temperature control to produce a cylindrical crystal of the desired diameter. This method produced the first high-quality germanium crystals (Bridgers *et al.*, 1958) and is used today for the growth of doped and ultrapure crystals. The CZ growth of dislocation-free silicon crystals has developed into the large-scale industrial production of over 10,000 tons of material per year.

4.3 Impurities and Defects

4.3.1 Solubility. The solubility is defined as the maximum concentration of an impurity in a solid phase at a given temperature that can be attained without introducing a phase

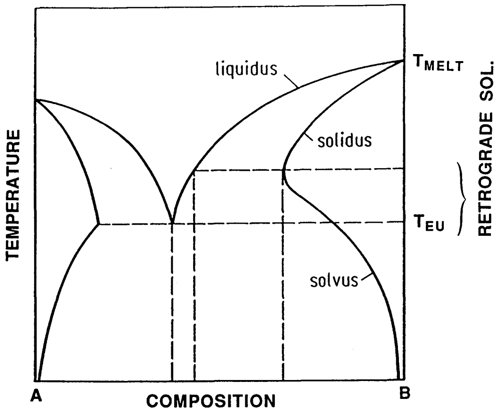

FIG. 18. Phase diagram illustrating the phenomenon of retrograde solubility (right-hand side). The left side of the figure illustrates the usual case with the maximum solubility at the eutectic temperature (after Weber, 1983).

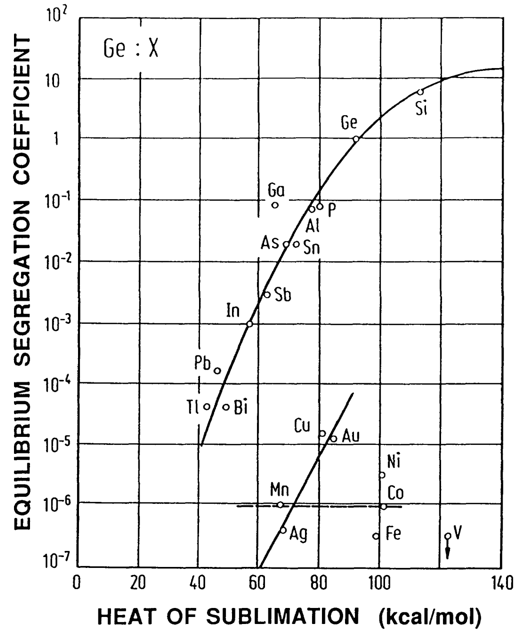

FIG. 19. Segregation coefficient of impurities vs heat of sublimation (after Trumbore, 1960).

change. Most impurities in germanium have such low solubilities that they are normally given as atoms/cm^3 instead of areas in phase diagrams (Fig. 17). As seen in the figure, most of the impurities display retrograde solubility in that the maximum solid solubility occurs below the melting temperature. This phenomenon is illustrated in Fig. 18. The left side of the figure shows the normal case where the maximum solubility occurs at the impurity(A)-germanium(B) eutectic (T_{Eu}), while the right side shows the retrograde case in which the maximum solubility is at a temperature higher than that of the eutectic.

4.3.2 Segregation Coefficient. When a solid is in contact with its liquid phase at the melting point, impurities typically will exhibit segregation effects between the two phases which at equilibrium is described by the segregation coefficient, k = (impurity concentration in solid)/(concentration in liquid). The segregation coefficient is related to several physical properties of the impurity and is shown in Fig. 19 as a function of the heat of sublimation and in Fig. 20 as a function of the maximum solubility $C^{\mathrm{eq}}_{\mathrm{max}}$. Zone purification relies on the phenomenon of impurity segregation in a narrow liquid zone that is heated by a radio-frequency heater coil. Multiple passage along a bar of germanium leads to a redistribution of the impurities along the bar. Impurities with $k < 1$ will stay preferably in the liquid and will be carried to one end of the bar, while impurities with $k > 1$ will accumulate with much less effect on the opposite side. Removal of the impure ends leads to a purified material (Pfann, 1966).

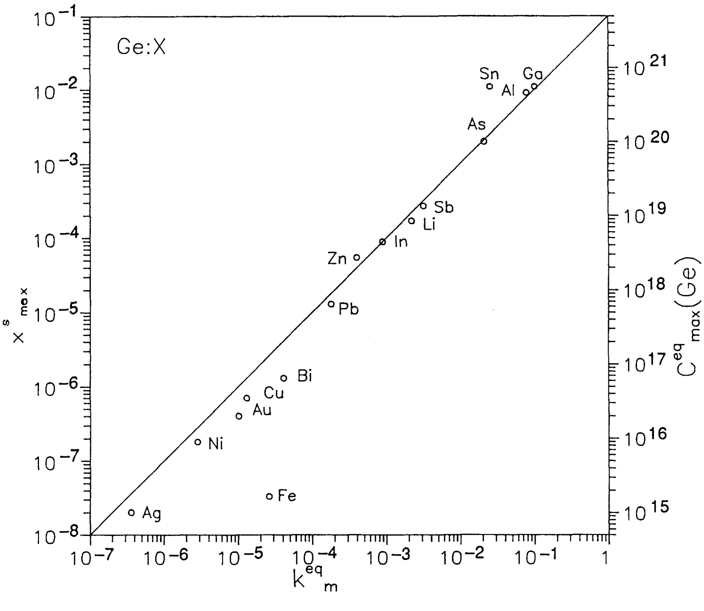

FIG. 20. Segregation coefficient $k^{\mathrm{eq}}_{\mathrm{m}}$ vs maximum solubility $C^{\mathrm{eq}}_{\mathrm{max}}$ for impurities in germanium (after Fishler, 1962).

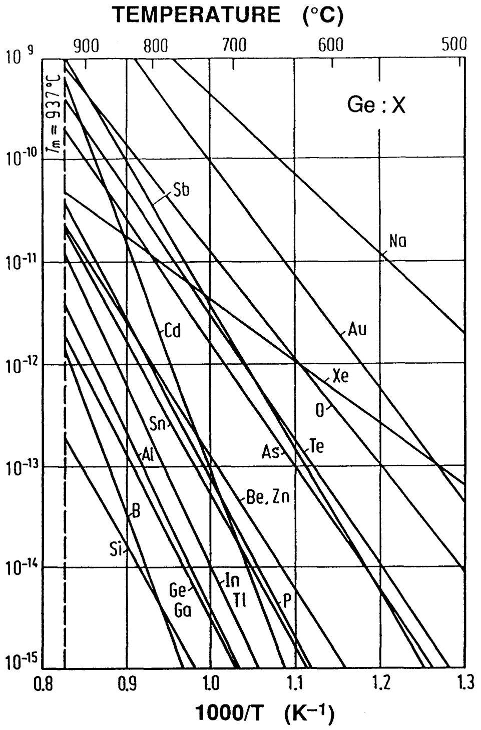

FIG. 21. Diffusion coefficient of impurities that diffuse by the vacancy mechanism. These are all slowly or moderately migrating elements (after Stolwijk, 1989).

4.3.3 Impurity Diffusion. The majority of impurities diffuse in germanium by the vacancy mechanism. This mechanism involves diffusion by impurity exchange with vacant lattice sites (vacancies). Since the concentration of vacancies is very small, the diffusion coefficients for impurities moving via vacancy assistance are small (Fig. 21) (Stolwijk, 1989). In the very open lattice structure of Ge a number of impurities can occupy interstitial spaces between the host atoms. Li, an interstitial donor, is the prototypical fast interstitial diffusor. Because Li does not form any bonds with the Ge lattice atoms, little energy (~0.4 eV) is required to jump from one tetrahedral interstitial space to the next. An Arrhenius plot of the diffusion coefficients of several interstitial impurities is given in Fig. 22.

Impurities that have both finite interstitial and substitutional solubilities usually display rapid interstitial diffusion behavior, while their electronic properties may be dominated by the substitutional component. Cu is the classical case for this dissociative diffusion mechanism (Stolwijk *et al.*, 1985). In its substitutional position it acts as a triple acceptor, as one would expect from its valence of one. In the interstitial form Cu has been characterized as a donor, though no energy level has ever been reported.

Impurities in germanium can be divided into groups that have similar properties and are studied by similar techniques. The column III and V impurities are most nearly germanium-like and thus are the slowest diffusers and have the highest solubility. The dopant atom being most like the host, the extra or missing valence electron (hole) is most weakly bound, and has a very large Bohr radius (~100 Å) and a consequent large optical cross section for ionization. These elements always diffuse by the simple vacancy method and are most accurately studied by photothermal-ionization spectroscopy.

The multivalent donor and acceptor impurities are less Ge-like, diffuse faster, and have lower solubilities than groups III-V. The multiple-charge states of these impurities have

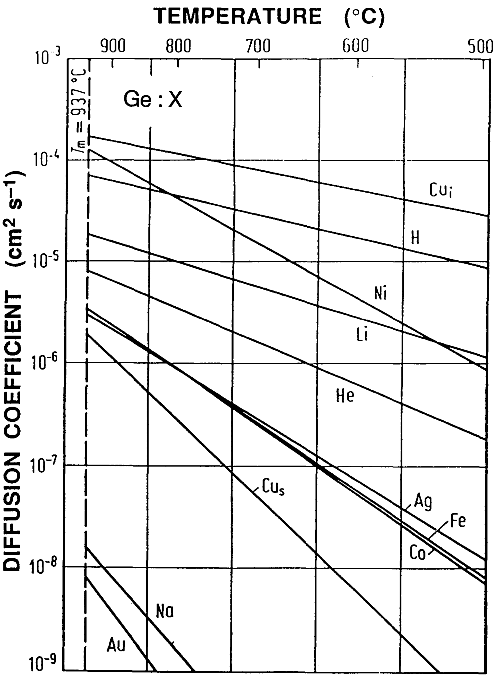

FIG. 22. Diffusion coefficient of impurities that diffuse by the interstitial mechanism. These are all fast migrating elements (after Stolwijk, 1989).

been studied by photoabsorption and photoconductivity. Diffusion is by the vacancy mechanism.

The transition metals exhibit still lower solubilities and faster diffusion. Their valence electrons are more localized, resulting in very small optical ionization cross sections. These impurities are usually studied by Hall effect or deep-level transient spectroscopy. Diffusion occurs via the vacancy mechanism.

A few impurities, e.g., H, Li, Ag, Ni, are fast-diffusing interstitial impurities that do not follow the above trends of solubility and ionization energy.

4.3.4 Defects. There exists a rich literature on point and line defects in Ge. Point defects have been generated by various kinds of energetic radiation. γ rays and electrons of ~ 1 MeV are ideally suited to form single point defects. Mashovets (1977) has reviewed the understanding of these vacancylike defects. The outstanding features of these defects are their strong localization and the resulting depth of the energy levels in the band gap. Several low-temperature annealing stages have been observed. Higher-energy radiation produces extended defects that are not understood in detail. It is known, though, that these defects act as strong trapping and recombination centers and that they compensate shallow dopants turning the crystal highly resistive.

A rich spectrum of dislocations with various Burgers vectors has been studied. The structure of these dislocations closely resembles those found in Si single crystals.

5. APPLICATIONS

The role of germanium as the dominant semiconductor material for electronic devices such as transistors ended by 1960, and its use has been relegated to those applications where some of its unique properties are of importance. Today's bulk of germanium production is consumed by detectors and optics.

5.1 Photodetectors

The long-wavelength cutoff λ_c (in μm) of a semiconductor photodetector is given by

$$\lambda_c = hc/E = (1.24\ \text{eV}\,\mu\text{m})/E, \qquad (12)$$

where E (in eV) is the energy needed to produce a free charge carrier. For a band-to-band transition in room-temperature germanium (intrinsic detector), $E=0.66$ eV, yielding $\lambda_c = 1.8\ \mu$m, which includes the technically important 1.3- and 1.5-μm low-loss windows in optical fiber communications. Intrinsic germanium photovoltaic detectors are normally reverse-biased *p-n* junctions minimizing both response time (10^{-11} s) and dark current (Sze, 1981).

Detectors can also be made using the ionization of impurities to produce charge carriers (extrinsic photodetectors). Group III and V impurities in germanium have ionization energies E_i of about 0.01 eV so that $\lambda_e = 120$ μm. Deeper-level impurities such as Be, Zn, Hg, Cu, and Au can be added to maximize the sensitivity at shorter wavelengths. Also, E_i may be reduced by uniaxial stress along [100] to extend the long-wavelength cutoff to 250 μm.

Extrinsic photodetectors must be operated at cryogenic temperatures so that the net dopants are neutral, i.e., they bind a carrier. These detectors are used as low–bias-voltage photoconductors to prevent electric field ionization of the impurities (avalanche breakdown), and, as a result of this low electric field, are much slower than intrinsic *p-n* diode detectors. For photoconductors, gain is the ratio of carrier lifetime to carrier transit time so that gain may be traded for speed. In the 30–250-μm wavelength range germanium photoconductors have the highest detectivity of any device (Haller *et al.*, 1979; Haller, 1985).

5.2 Nuclear-Radiation Detectors

The development of hyperpure germanium (Hall and Soltys, 1971; Hansen, 1971) was driven by a demand for a material to make nuclear-radiation detectors (Haller and Goulding, 1981). These detectors are giant-volume *p-i-n* diodes (up to 400 cm^3) where the *i* region forms the bulk of the device and thin ion-implanted *p* and *n* layers form blocking contacts. The *i* layer is lightly doped ($|N_A - N_D| \leqslant 2 \times 10^{10}$ cm^{-3}) and can be totally depleted at operating bias voltages of a few thousand volts. The relation between the net impurity concentration $|N_A - N_D|$, the bias voltage V, and the thickness of the region depleted of free carriers W is given by

$$W^2 = 2\epsilon\epsilon_0 V/e(N_A - N_D),$$

where ϵ is the dielectric constant, ϵ_0 is the permittivity of vacuum, and e is the charge of the electron. For example, for a depletion layer thickness of 1 cm at 1000 V bias, $|N_A - N_D|$ must be less than 2×10^{10} cm^{-3}.

These detectors are used as single-particle or photon energy spectrometers. On average, 3 eV of incident energy are required to produce one electron-hole pair. The ultimate energy resolution is dominated by the statistical nature of the charge-production processes. Because these processes are not independent, they do not follow normal statistics. The fluctuation in the number of electron-hole pairs ΔN is given by

$$\Delta N = (FN)^{1/2},$$

with $F=0.08$ (Fano factor). The full width of a 1-MeV γ-ray peak at half maximum (FWHM) can be as small as 1.2 keV, a factor of 30 smaller than a line produced by a scintillation detector. That the fluctuation in the number of electron-hole pairs is smaller than the standard statistical fluctuation is related to the fact that the ionization events are not completely random and unrelated.

5.3 Low-Temperature Ge Thermistors

Strong interest in reliable and reproducible temperature sensors by the astrophysical, radio astronomy, and "dark-matter" search communities in the subkelvin range has led to a strong development effort of neutron-transmutation doped (NTD) Ge material and devices. NTD offers the most homogeneous and controlled doping technique (Larrabee, 1984). NTD Ge is unique among semiconductors because it contains both acceptors (Ga) and donors (As) in a fixed ratio determined by the natural abundance of ^{70}Ge and ^{74}Ge and their thermal-neutron capture cross sections (Haller *et al.*, 1984). NTD Ge can be produced with resistivities ranging from 1 Ω cm to 10^9 Ω cm in a temperature range from 1 K to 15 mK. The NTD Ge thermistors are heat sensitive and do not exhibit any specific wavelength cutoff limits. The typical device has submillimeter dimensions to decrease thermal mass and is equipped with two boron-ion–implanted, degenerately doped *p* contacts (Haller, 1985).

Nonlinear conduction phenomena at very low temperature have led to interesting transport studies (Kenny *et al.*, 1989, Wang *et al.*, 1990).

5.4 Optics

The consumption of germanium for infrared optics now exceeds all other uses combined. The attraction of germanium for this application arises from its environmental inertness, good mechanical properties, moderate cost, very low dispersion, and transparency from 2 to 15 μm. Germanium's only drawback has been its high refractive index (4.0) and the difficulty of making tough antireflection coatings. This antireflection problem has now been solved by an ideal coating, diamondlike carbon, that is supertough and has a refractive index of 2.0.

Germanium exhibits significant absorption of infrared radiation in the 15- to 40-μm range due to intrinsic-crystal lattice-phonon resonances (Fig. 12). Unless specially purified, germanium will also show an absorption line at 11.7 μm because of the local vibrational mode of oxygen impurities.

5.5 Miscellaneous Uses

Graded-index optical fibers are made using a high index of refraction germanium-doped silicon core. Germanium has an almost perfect lattice match to GaAs so that high-quality epitaxial layers can be grown; these Ge-GaAs bilayers have produced the highest-efficiency solar cells. Impatt diodes, tunnel diodes (Sze, 1981), and phonon-drag detectors are made from germanium. A population inversion between the heavy-hole and light-hole bands can be induced in *p*-type germanium in crossed electric and magnetic fields at low temperature to produce a tunable, long-wavelength laser (Gornik, 1991).

A research area of great current interest is the prospect of integrating quantum-well structures on silicon using germanium. Considerable progress has been made in producing Ge-Si strained-layer superlattices for photodetectors and transistors (Bean *et al.*, 1991).

GLOSSARY

Band Structure: Electrons in solids can only assume energies in certain ranges called allowed bands. Forbidden energy ranges are called band gaps. Depending on the degree to which the highest allowed energy band is filled with electrons, and depending on the width of the band gap, a solid is an electrical conductor, an insulator, or a semiconductor.

Bound Excited State: Electrons (holes) are bound to a donor (acceptor) by the Coulomb field of the positive (negative) charge at the donor (acceptor) core. As in a hydrogen atom, there exists a ladder of discrete energy states in which the electron (hole) can exist. The lowest energy state is the ground state; all other states below the conduction-band edge (above the valence-band edge) are bound excited states. Because of the strong analogy between donors (acceptors) and the hydrogen atom, bound excited states are often called hydrogenic states or hydrogenic bound excited states.

Brillouin Zone: The regular spacing of atoms in a crystal lattice can be represented by points in an inverse or reciprocal space, also called momentum space. The Brillouin zone occupies part of this space. Every point on the Brillouin zone surface is the end point of a vector signifying a diffraction condition of the real-space crystal lattice.

Conduction Band: The first allowed energy band above the band gap in semiconductors or insulators. Electrons in this band are free to move and to conduct electricity.

Diamond Lattice: A regular array of points in space. The corners and the centers of each face of two intersecting cubes describe the basic building block of the diamond lattice. The cubes are displaced along a body diagonal by one quarter of its length. Putting real atoms on all of these lattice points leads to a crystal. These atoms form covalent bonds with each of the four nearest neighbors. Carbon, silicon, and germanium crystallize in the diamond structure.

Effective Mass: The ratio of the acceleration of an electron or hole in a crystal to the force producing the acceleration. Because of the wave nature of electrons (holes), motion in a crystal with regular atom spacings is altered and the effective mass differs from the free-electron mass. It may be different for electrons and holes, and may depend on the direction of motion. The curvatures of the conduction-band minima (valence-band maxima), i.e., the second derivatives of these curves (see Fig. 1), are inversely proportional to the effective mass.

Fano Factor: The factor by which the statistical fluctuations in the production of electrons and ions in a gas or electrons and holes in a semiconductor deviate from normal statistics.

Holes: See Valence Band.

Hydrogenic States: Bound excited states.

Interstitial: An impurity or host atom located off a lattice site. Interstitials can be very mobile in the wide-open diamond lattice and play an important role in impurity and defect motion (see Vacancy).

Neutron-Transmutation Doping (NTD): Doping resulting from transmutation of atoms of the original material that have been made radioactive by absorption (capture) of incident neutrons. Most often used for silicon, NTD produces phosphorus donors in this semiconductor. In Ge both Ga acceptors and As donors are formed through NTD.

Photothermal Ionization Spectroscopy (PTIS): Excitation of an electron (hole) at low temperatures from the ground state to one of the bound excited states in a donor (acceptor) can be followed by further excitation through thermal energy (a phonon) into the conduction (valence) band. The resulting free carrier will change the electrical conductivity of the crystal. The intense conductivity peaks at photon energies corresponding precisely to transitions from ground state to bound excited states have become the most sensitive and precise tool in the study of impurities and defect in numerous semiconductor crystals.

Vacancy: A crystal lattice site without an atom occupying it. Vacancies play a very important role in the motion of impurities and defects through a crystal.

Valence Band: The highest energy band filled with electrons. Missing electrons in the band, holes, can conduct electricity in the valence band.

List of Works Cited

Baldereschi, A., Lipari, N. O. (1973), *Phys. Rev. B* 8, 2697–2709.

Baldereschi, A., Lipari, N. O. (1974), *Phys. Rev. B* **9**, 1525–1539.

Baldereschi, A., Lipari, N. O. (1977), in: F. G. Fumi (Ed.), *Proceedings of the XIII International Conference on the Physics of Semiconductors, Rome,* Amsterdam: North-Holland, pp. 595–598.

Bardeen, J., Brattain, W. (1948), *Phys. Rev.* **74**, 230–231.

Bean, J. C., Iyer, S. S., Wang, K. L. (1991), *Silicon Molecular Beam Epitaxy,* Proc. Mat. Res. Soc., Vol. 220, Pittsburgh: Materials Research Society.

Biondi, F. J. (1958), *Transistor Technology,* Vols. 2 and 3, Princeton, NJ: Van Nostrand.

Bridgers, H. E., Scaff, J. H., Shive, J. N. (Eds.) (1958), *Transistor Technology,* Vol. 1, Princeton, NJ: Van Nostrand.

Collins, R. J., Fan, H. Y. (1954), *Phys. Rev.* **93**, 674–678.

Czochralski, J. (1918), *Z. Phys.* **92**, 219–221.

Davies, G., Lightowlers, E. C., Ozhogin, V., Itoh, K., Hansen, W. L., Haller, E. E. (1992), in: Proceedings of the Fifth International Conference on Shallow Impurities in Semiconductors, Kobe (in press).

Faulkner, R. A. (1969), *Phys. Rev.* **184**, 713–721.

Fishler, S. (1962), *J. Appl. Phys.* **33**, 1615.

Fritzsche, H., Cuevas, M. (1960), *Phys. Rev.* **119**, 1238–1245.

Fuchs, H. D., Grein, C. H., Thomson, C., Cardona, M., Hansen, W. L., Itoh, K., Haller, E. E. (1991), *Phys. Rev. B* **43**, 4835–4842.

Gornik, E. (1991), *Opt. Quan. Electron.* **23**, no. 2 (special issue on far-infrared semiconductor lasers).

Haegel, N. M., Haller, E. E. (1986), in: J. Besson (Ed.), *Materials Technology for IR Detectors,* SPIE Proceedings 659, Bellingham, WA: SPIE, pp. 188–194; see also *Infrared Phys.* **26**, 247–261.

Hall, J. J. (1962), *Phys. Rev.* **128**, 68–75.

Hall, R. N., Soltys, T. J. (1971), *IEEE Trans. Nucl. Sci.* **18**, 160–169.

Haller, E. E. (1978), *Phys. Rev. Lett.* **40**, 584–586.

Haller, E. E. (1985), *Infrared Phys.* **25**, 257–266.

Haller, E. E. (1986), in: P. Grosse (Ed.), *Advances in Solid State Physics,* Vol. 26, Braunschweig: Vieweg, pp. 203–229.

Haller, E. E., Falicov, L. M. (1978), *Phys. Rev. Lett.* **41**, 1192–1194.

Haller, E. E., Goulding, F. S. (1981), in: C. Hilsum (Ed.), *Handbook on Semiconductors,* Vol. 4, Chap. 6C, Amsterdam: North-Holland, pp. 799–827.

Haller, E. E., Hansen, W. L., Goulding, F. S. (1981), *Adv. Phys.* **30**, 93–138.

Haller, E. E. Heuschen, M. R., Richards, P. L. (1979), *Appl. Phys. Lett.* **34**, 495–497.

Haller, E. E., Navarro, H., Keilmann, F. (1987), in: O. Engstrom (Ed.), *Proceedings of the Eighteenth International Conference on the Physics of Semiconductors,* Singapore: World Scientific, pp. 837–840.

Haller, E. E., Palaio, N. P., Rodder, M., Hansen, W. L., Kreysa, E. (1984), in: R. D. Larrabee (Ed.), *Proceedings of the Fourth International Conference on Neutron Transmutation Doping of Semiconductor Materials,* New York: Plenum, pp. 21–36.

Hansen, W. L. (1971), *Nucl. Instrum. Methods,* **94**, 377–380.

Hensel, J. C., Thomas, G. A., Phillips, T. G., Rice, T. M. (1977), in: H. Ehrenreich, F. Seitz, and D. Turnbull (Eds.), *Solid State Physics,* Vol. 32, New York: Academic.

Jeffries, C. D., Keldysh, L. V. (1983), *Electron Hole Droplets in Semiconductors,* Modern Problems in Condensed Matter, Vol. 6, Amsterdam: North-Holland.

Jones, R. L., Fisher, P. (1965), *J. Phys. Chem. Solids* **26**, 1125–1131.

Kahn, J. M., McMurray, R. E., Jr., Haller, E. E., Falicov, L. M. (1987), *Phys. Rev. B* **36**, 8001–8014.

Keldysh, L. V. (1968), in: S. M. Ryvkin (Ed.), *Proceedings of the Ninth International Conference on Physics of Semiconductors, Moscow,* Leningrad, Nauka, pp. 1303–1312.

Kelso, S. (1979), PhD thesis (Physics Dept., University of California at Berkeley and Lawrence Berkeley Laboratory, Rep. LBL-8786).

Kenny, T. W., Richards, P. L., Park, I. S., Haller, E. E., Beeman, J. (1989), *Phys. Rev. B* **39**, 8476–8482.

Kittel, C. (1986), *Introduction to Solid State Physics,* 6th ed., New York: Wiley, p. 199.

Larrabee, R. D. (Ed.) (1984), *Proceedings of the Fourth International Conference on Neutron Transmutation Doping of Semiconductor Materials,* New York: Plenum.

Lifshits, T. M., Nad, F. Ya. (1965), *Sov. Phys. Dokl.* **10**, 532–533.

Madelung, P. (Ed.) (1982), *Semiconductors: Physics of Group IV Elements and III-V Compounds,* Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology, Group 3, Vol. 17a, Berlin: Springer-Verlag.

Madelung, O., Schulz, M. (Eds.) (1987), *Semiconductors: Intrinsic Properties of Group IV Elements and III-V, II-VI and I-VII Compounds,* Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology, Group 3, Vol. 22a, Berlin: Springer-Verlag.

Madelung, O., Schulz, M. (Eds.) (1989), *Semiconductors: Impurities and Defects in Group IV Elements and III-V Compounds,* Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology, Group 3, Vol. 22b, Berlin: Springer-Verlag.

Mashovets, T. V. (1977), in: N. B. Urli (Ed.), *Radiation Effects in Semiconductors 1976,* IOP Conference Series No. 31, London: Institute of Physics, pp. 30–44.

McLean, T. P. (1960), in: A. F. Gibson (Ed.), *Progress in Semiconductors,* Vol. 5, London: Heywood.

Morin, F. J., Maita, J. P. (1954), *Phys. Rev.* **94**, 1525–1529.

Navarro, H., Haller, E. E., Keilmann, I. (1988), *Phys. Rev. B* **37**, 10822–10828.

Pearton, S. J. (1982), *Solid State Electron.* **25**, 499–503.

Pfann, W. G. (1952), *Trans. Am. Inst. Min. Met. Pet. Eng.* **194**, 747–753.

Pfann, W. G. (1966), *Zone Melting*, New York: Wiley.

Potter, R. F. (1985), in: E. D. Palik (Ed.), *Handbook of Optical Constants of Solids*, New York: Academic, pp. 465–478.

Prince, M. B. (1953), *Phys. Rev.* **92**, 681–687.

Ramdas, A. K., Rodriguez, S. (1981), *Rep. Prog. Phys.* **44**, 1297–1387.

Shockley, W. (1949), *Bell Sys. Tech. J.* **28**, 435–489.

Shockley, W., Pearson, G. (1948), *Phys. Rev.* **74**, 232–233.

Skolnick, M. S. (1989), in: O. Madelung, M. Schulz (Eds.), *Semiconductors: Impurities and Defects in Group IV Elements and III-V Compounds*, Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology, Group 3, Vol. 22b, Berlin: Springer Verlag, pp. 466–489.

Stolwijk, N. A. (1989), in: O. Madelung, M. Schulz (Eds.), *Semiconductors: Impurities and Defects in Group IV Elements and III-V Compounds*, Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology, Group 3, Vol. 22b, Berlin: pp. 439–465.

Stolwijk, N. A., Frank, W., Hölzl, J., Pearton, S. J., Haller, E. E. (1985), *J. Appl. Phys.* **57**, 5211–5219.

Sze, S. M. (1981), *Physics of Semiconductor Devices*, 2nd ed., New York: Wiley.

Tannenbaum, M. (1959), in: B. Hannay (Ed.), *Semiconductors*, New York: Reinhold, p. 87.

Teal, G. K., Little, J. B. (1950), *Phys. Rev.* **78**, 647.

Thewalt, M. L. W., Labrie, D., Booth, I. J., Clayman, B. P., Lightowlers, E. C., Haller, E. E. (1987), *Physica* **146B**, 47–64.

Torrey, H. C., Whitmar, C. A. (1948), *Crystal Rectifiers*, MIT Rad. Lab. Series Vol. 15, New York: McGraw-Hill.

Trumbore, F. A. (1960), *Bell Sys. Tech. J.* **39**, 205–234.

Wang, N., Wellstood, F. C., Sadoulet, B., Haller, E. E., Beeman, J. (1990), *Phys. Rev. B* **41**, 3761–3768.

Weber, E. R. (1983), *Appl. Phys. A* **30**, 1–22.

Winkler, C. (1886), *J. Prak. Chem.* **34**, 177–229.

Wolfe, J. P., Hansen, W. L., Haller, E. E., Markiewicz, R. S., Kittel, C., Jeffries, C. D. (1975), *Phys. Rev. Lett.* **34**, 1292–1293.

Yee, J. H., Swierkowski, S. P., Armantrout, G. A., Wichner, R. (1974), *J. Appl. Phys.* **45**, 3949–3954.

Further Reading

Haller, E. E. (1985), "Advanced Far Infrared Photoconductors and Bolometers," *Infrared Phys.* **25**, 257–266.

Haller, E. E. (1986), "Physics of Ultra-Pure Germanium," in: P. Grosse (Ed.), *Advances in Solid State Physics*, Vol. 26, Braunschweig: Vieweg, pp. 203–229.

Haller, E. E. (1992), "Hydrogen-Related Phenomena in Crystalline Germanium," in: J. Pankove and N. Johnson (Eds.), *Hydrogen in Semiconductors, Semiconductors and Semimetals*, Vol. 34, New York: Academic, Chap. 11, pp. 351–380.

Haller, E. E., Goulding, F. S. (1981), in: C. Hilsum (Ed.), *Device Physics*, Handbook on Semiconductors Vol. 4, Amsterdam: North-Holland, Chap. 6C, pp. 799–827 (new edition in preparation for 1993).

Hansen, W. L., Haller, E. E. (1983), "Ultra-Pure Germanium Crystal Growth," in: E. E. Haller, H. W. Kraner, and W. A. Higinbotham (Eds.), *Nuclear Radiation Detector Materials*, New York: Elsevier (Proceedings of the Materials Research Society, Vol. 16, pp. 1–16.

Kittel, C. (1986), *Introduction to Solid State Physics*, 6th ed., New York: Wiley.

Smith, R. A. (1979), *Semiconductors*, 2nd ed., Cambridge, U.K.: Cambridge Univ. Press.

Sze, S. M. (1981), *Physics of Semiconductor Devices*, 2nd ed., New York: Wiley.

GLACIERS AND ICE SHEETS

CHARLES RAYMOND, *Geophysics Program, University of Washington, Seattle, Washington, U.S.A.*

Introduction 239
1. **Distribution and Characteristics** 241
2. **Exchange with the Environment** 242
2.1 Atmosphere 242
2.2 Land Surface 243
2.3 Ocean 243
3. **Glaciers as Flow Systems** 243
3.1 Mass Balance 244
3.2 Glacier Motion 245
3.3 Advance and Retreat 246
4. **Glacier Ice in a Changing Climate** 248
Glossary 249
Works Cited 249
Further Reading 249

INTRODUCTION

Glaciology concerns natural ice in the environment including glaciers, sea ice, seasonal snow, and frozen ground. This article discusses only glaciers. A glacier is a perennial ice body resting at least partly on land with a relatively high-altitude source area of net ice income, a relatively low-altitude sink area of net ice loss, and transfer of mass from the source to sink by flow under the action of gravity. Under the term glaciers we will include both alpine glaciers that exist on mountain slopes and the large ice sheets that completely cover the terrain (Fig. 1).

Although glaciers are restricted to high-altitude and polar environments, they form an important part of the Earth's environment. Erosion and deposition from repeated prehistoric glaciations have shaped large areas of north temperate latitudes that are now densely populated. The geomorphologic effects have therefore received considerable attention both to understand the landscape and to reconstruct the history of glacier and ice-sheet changes and the related climatological causes (Drewry, 1986). Glaciers and ice sheets store water in solid form and are fundamental to hydrology through their effects on the timing and quality of water discharges into surrounding rivers. By storing water on the land, glaciers affect global sea level (National Research Council, 1985). Changes in ice load cause time- and space- dependent depression and rebound of the Earth's surface that affect the pattern of local sea-level change and provide the principal data about the viscosity of the Earth's mantle (Peltier, 1987). Glaciers themselves are tectonic systems that experience deformation and structural changes analogous to those occurring in the solid earth. Because the physical and chemical properties of ice contain information about conditions in the atmosphere at the time of deposition, coring into ice sheets has provided a powerful tool for investigating past climate (Delmas, 1992).

The hydrological and paleoclimatological aspects of glaciers are especially important with regard to global environmental change and impacts from human activities. Two important questions may be enunciated. First, what are the potential effects of glaciers on global sea-level change? A rise in sea level would have major environmental consequences and impacts on human activities. Second, what quantitative information can glacier ice provide about climate change in the past? This is essential information for validating predictive models of climate change. In the context of these questions, our theme will be to review understanding of physical processes acting within and at the boundaries of glaciers.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

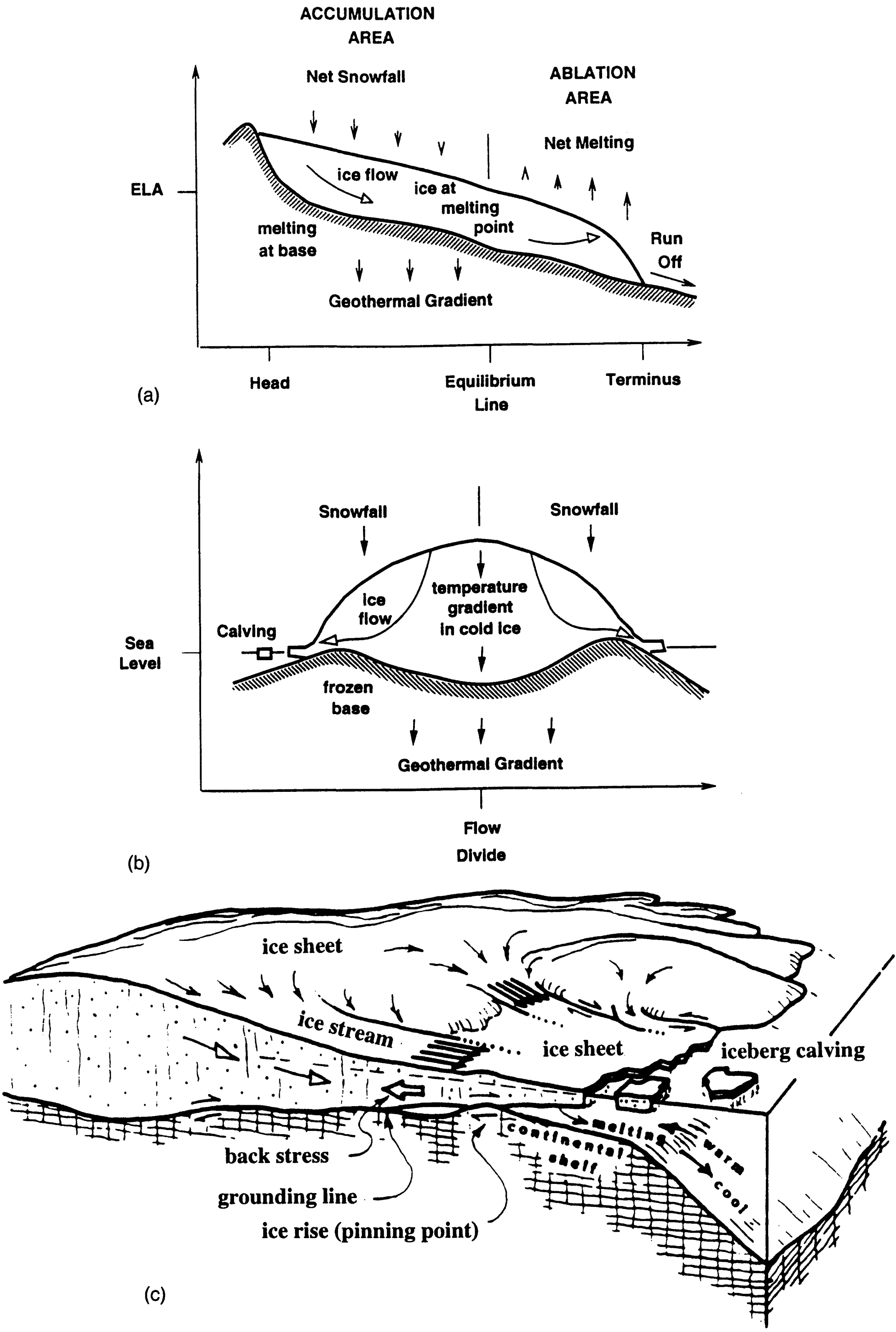
ACCUMULATION AREA
Net Snowfall
ABLATION AREA
Net Melting
ELA
ice flow
ice at melting point
melting at base
Run Off
Geothermal Gradient
Head
Equilibrium Line
Terminus
(a)
Snowfall
Snowfall
Ice flow
temperature gradient in cold ice
Calving
Sea Level
frozen base
Geothermal Gradient
Flow Divide
(b)
ice sheet
ice stream
ice sheet
iceberg calving
back stress
grounding line
ice rise (pinning point)
continental shelf
melting
warm
cool
(c)

FIG. 1. Schematic illustration of contrasting geometry and thermal conditions in (a) an idealized temperate mountain glacier and (b) a polar ice sheet shown as vertical sections along the direction of flow. Arrows indicate ice mass transfer or temperature gradients related to transfer of heat. In (a), ELA refers to altitude of the equilibrium line separating zones of net income (accumulation area) and net loss (ablation area). (c) Components of a marine ice sheet [provided courtesy of the National Research Council, U.S. Academy of Sciences; from National Research Council (1985)].

1. DISTRIBUTION AND CHARACTERISTICS

At present the largest volumes of glacier ice exist in Antarctica and Greenland. These land masses are almost entirely covered by ice, and contain the only continental scale ice sheets at present. There are many much smaller glaciers in the mountain regions of the world. Table 1 summarizes geometrical characteristics of the ice sheets and glaciers.

The volume of ice can be expressed in terms of the equivalent sea-level rise that would occur if the water-equivalent volume were spread evenly over the present area of the oceans. To assess the actual sea-level change that would occur if an ice sheet were to melt, it is necessary to account for the ice volume that exists below sea level and displaces sea water [Fig 1(c)], such as in portions of ice sheets that are grounded below sea level (marine ice sheet) or are floating (ice shelf). The West Antarctic Ice Sheet is the only large marine ice sheet at present. Consequently there is a large fractional difference between its total volume in sea-level equivalent and the actual sea-level change that it could cause (Table 1). The volumes of the Antarctic and Greenland ice sheets correspond to large sea-level equivalents and their stability under changing climate is important for this reason alone.

The total volume of ice contained in other glaciers is small in comparison to the ice sheets, but is not negligible. These glaciers tend to exist in environments with relatively high precipitation and melting, so that the exchange rate of water is quite large and comparable to the exchange rate in the big ice sheets (Table 2). Glaciers experienced a period of expansion starting about 800–600 years ago called the "Little Ice Age." Since about 1890, the glaciers have retreated on average in all mountain ranges of the world (Haeberli *et al.,* 1989). Available information about the rates of volume changes suggest that the cumulative volume loss of small glaciers over the last century has made a major contribution to the rise in sea level over this time (Meier, 1984). Information about the rates of change of the ice sheets is sparse and available estimates have large uncertainties (Table 2).

Averaged over the past several million years, land ice has been much more extensive than at present (Crowley and North, 1991). Following a thermal maximum about 50 million years ago, the Earth has experienced an irregular cooling trend. Glaciation first began in Antarctica perhaps as early as 40 million years ago with extensive volumes of ice being present by 10 million years ago. About 2.4 million years ago large periodic oscillations between glacial and interglacial conditions began to appear in the northern hemisphere at mid-latitudes. These were characterized by gradual expansion of ice sheets terminated by rapid recessions that appear to be correlated with high summer insolation on the southern limits caused by cyclic changes in the Earth's orbital geometry. The most recent continental-scale glacial expansion culminated about 22–14 thousand years ago. The total ice volume at that time was more than twice the present volume, and caused a sea-level depression of about 120 m. Recession occurred rapidly and interglacial conditions similar to the present were reached by about 10 thousand years ago. The large and sometimes rapid changes of the past point to the potential importance of changes in the future.

Table 1. Characteristics of global ice. See Oerlemans (1989) and van der Veen (1991) for sources. Uncertainties are generally within 10% except as noted by * which have larger uncertainties.

	Total area (10^6 km^2)	Mean thickness (km)	Total volume (10^6 km^3)	Sea-level Equivalent (m)	Sea-level change (m)
East Antarctica	9.9	2.6	25.9	72	60*
West Antarctica	2.1	1.6	3.4	9.4	6.0*
Greenland	1.8	1.5	2.7	8.3	7.5*
Rest	0.6*	0.3*	0.2*	0.5*	0.5*

Table 2. Mass exchange of global ice. See Oerlemans (1989) and van der Veen (1991) for sources. Uncertainties are generally less than 15% except for * where uncertainties may be larger.

	Annual Accumulation		Annual Melt						Velocity Range (m a^{-1})		
	Total (10^3 km^3)	Per unit accumulation area (m)	Total (10^3 km^3)	Per unit ablation area (m)	Annual calving (10^3 km^3)	Sea-level change (mm a^{-1})	ELA (km above sea level)	Accumulation area ratio (AAR)	Normal	Ice streams, tidewater, surges	Mean residence time (10^3 a)
Antarctica	2.2		< 0.01		2*	0–1.2	Negative	1.0	10^0–10^2	~10^3	13.3
Greenland	0.5	0.35	0.3	1.2	0.2*		1.3	0.84	10^0–10^2	~10^3	5.4
Rest	0.5	~1.3	0.5	~2.4	~0.0	0.2–0.8	0–6	0.6–0.7	10^0–10^2	10^3–10^4	0.4

2. EXCHANGE WITH THE ENVIRONMENT

Glaciers exist because of accumulation of snow from the atmosphere onto the land surface. Some glaciers (tidewater glaciers) and both of the present large ice sheets reach the ocean. The exchange of mass, momentum, and energy across these boundary surfaces (land, atmosphere, and ocean) determines boundary conditions affecting the properties of the ice and the behavior of the glacier. These exchanges also determine how a glacier affects its surroundings.

2.1 Atmosphere

The upper surface of a glacier where the ice contacts the atmosphere is most importantly a surface of mass exchange. Snowfall is the primary ice mass input to a glacier. Deposition of rime (supercooled water droplets) or hoar frost (condensation of water vapor) is normally a much smaller contribution. Heat reaches the upper surface through radiative and turbulent exchanges. This heat may warm the subsurface or, if the surface is at the melting point, cause melting and mass loss. Net outward transfer of heat can cool the subsurface or freeze any water present and add mass.

In addition to the ice mass, trace amounts of impurities are added from the atmosphere. Salts of marine and terrestrial origin, dust, pollen, volcanic ash and acids, soot, and radioactive reaction products from cosmic rays occur with varying concentration and proportion depending on production processes on the ocean and land surfaces, solar activity, air mass history, and precipitation processes. Isotopic ratios (O^{18}/O^{16} and D/H) are affected by fractionation during evaporation and condensation and are related to the temperature at which the snowfall formed. As the snow is compressed to ice, some of the gas present in the pore space is trapped, thus preserving a sample of the atmosphere. In polar conditions these constituents remain locked in the ice and constitute storehouses of information about past environments that can be accessed by coring into the ice sheets. Impurities also affect the physical properties of the ice.

Exchange of momentum with the atmosphere is negligible, the upper surface being essentially stress free except for a small normal stress from atmospheric pressure.

2.2 Land Surface

The lower surface (bed) of a glacier in contact with the land supports the weight of the glacier. When the temperature of the bed is below freezing, the momentum balance can usually be adequately described by a no-slip condition. However, when the bed is at melting temperature and water is present, then the ice may move over the bed. Several theories relate the resulting basal velocity to stress parallel (shear) and normal to the bed, water pressure in the basal zone, mechanics of the ice and subglacial material, and roughness of the bed (Lliboutry, 1987). However, none of the proposed relationships have been verified by observation. The lack of proven formulations of the basal boundary mechanics is the major gap in understanding glacier motion.

A closely related problem concerns the erosion, transport, and deposition of rock debris. These processes are fundamental to glacial action on the landscape. They also determine the subglacial environment and the mechanisms that will control basal velocity. In some circumstances, a relatively clean contact between ice and solid rock exists (Kamb and LaChapelle, 1964); then roughness and the opening of cavities between the ice and rock are important. In other circumstances the ice is underlain by soft deformable debris (Blankenship *et al.*, 1987) in which shearing makes an important contribution to the basal motion. Controls on the production and export of basal debris are poorly understood, which represents one of the barriers to predicting basal velocity.

Whether the bed is hard rock or soft unlithified debris, the condition of water at the bed has a strong influence on the basal velocity. Very high basal velocities one or more orders of magnitude above typical glacier velocities occur when the water pressure approaches flotation, either by extensive separation of the ice from the bed or by pore-pressure–induced loss of strength in the subglacial material. There is a growing understanding of the drainage along glacier beds through tunnels melted in the ice, through linked cavities at the bed, and in the pore space of debris (Rothlisberger and Lang, 1987), but accurate predictions of the morphology of the drainage system and the water pressure are not yet possible.

Geothermal heat flowing upward from the Earth must either be transported up through the ice mass or, if the ice is at the melting point, cause melting. Melting from geothermal heat is usually negligible for mass balance, but the melt water may affect the subglacial hydrology and basal velocity.

2.3 Ocean

Where the ice floats as an ice shelf, stress acting on the base is essentially the hydrostatically determined ocean water pressure, which provides a simple, well-defined momentum boundary condition at the base. Water motions in the cavity between the ice and sea floor can transport large amounts of heat. The interactions with the ice are affected by the coupled diffusion of heat and salts and depression of the melting point by pressure. These thermodynamic processes can cause significant melting or freezing onto the base of an ice shelf. Water cooled by contact with ice at high pressure under thick ice shelves in Antarctica contributes to production of cold Antarctic bottom water that resides in the deep ocean basins.

The very edge of a floating ice shelf or grounded ice in contact with the ocean is susceptible to mass loss by calving. Calving refers to fracture behind the ice front that leads to formation of a free-floating iceberg.

3. GLACIERS AS FLOW SYSTEMS

Two principal mass-transfer processes interact to determine the size and shape of a glacier. They are the distribution of mass transfer across the external boundaries that determines addition and loss to the system (mass balance), and the pattern of ice transport through internal boundaries that redistributes mass inside (ice motion). These processes act in parallel to determine the ice thickness distribution and its changes with time. Both are coupled to the external environment and internally to each other in various ways that are schematically outlined in Fig. 2. The most important interactions are shown as solid arrows. Ice mass balance is affected by the meteorological environment (atmosphere) and altitude (land surface + ice thickness). Ice motion is affected by the slope of the landscape and ice thickness on it (land surface and ice thickness) and temperature of the ice (atmosphere, solid earth, and ice thickness). The ocean has direct effects on tidewater glaciers and ice sheets. Dashed arrows indicate some feedback linkages to the environment. Examples are from ice thickness to the atmosphere caused by the elevation, temperature, and reflectivity of the surface; from ice thickness to the land surface associated with isostasy; from the ice motion to the land surface through erosion and deposition; and from the ice mass balance to the land and ocean through melting and runoff.

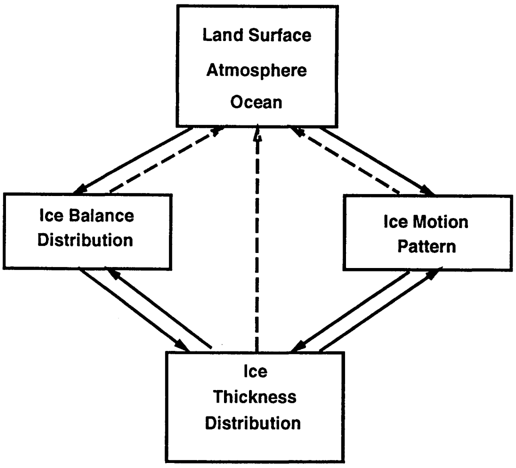

FIG. 2. Linkages between the surrounding environment (top) and the geometry of a glacier (bottom) through exchanges of ice mass across its external (left) and internal (right) surfaces.

3.1 Mass Balance

Addition (removal) of ice mass to (from) a glacier is called accumulation (ablation). The difference between the rates of accumulation and ablation on unit horizontal area is called the mass balance rate (b). It describes the spatial distribution of mass gain and loss. The integral of b over the full area of the glacier gives the total rate of mass change.

On all glaciers there is an accumulation (source) area where, averaged over time, accumulation exceeds ablation and b is positive. On glaciers terminating on land there is an ablation (sink) area where ablation exceeds accumulation and b is negative. The boundary between these zones defines the equilibrium line ($b=0$) that is described by its mean altitude (ELA). The ratio of accumulation area to total area (AAR) is commonly in the range from 0.6 to 0.7. Glaciers that calve into the ocean (or lakes) may have larger AAR values approaching 1 because mass is lost to calving at the edge as well as melting on the surface. In extreme polar conditions such as in Antarctica, there is essentially no melting on the upper surface. Then ablation is restricted to melting off the bottom of ice shelves and calving. Table 2 describes some balance-rate characteristics of glaciers and ice sheets.

Modeling of mass balance is complicated by the multiplicity of parameters that affect it. Precipitation depends on the details of microphysical processes in clouds and details of the vertical motions associated with differential heating and interaction with topography. Radiative (solar and thermal infrared) and turbulent (sensible and latent) heat fluxes depend on weather (temperature, humidity, wind, and cloud cover and type), condition of the surface (temperature, albedo, and roughness) and location (latitude, altitude, continentality, exposure angle, and horizon).

On some glaciers, the variation of altitude, exposure angle, and horizon produce a spatial pattern of b that is consistent from year to year except for nearly uniform shift Δb reflecting the conditions for a particular year. The Δb can alternatively be described in terms of the corresponding shift in ELA. This behavior provides one means for simplification.

On the basis of mass-balance observation series in several regions, it appears possible to predict ELA using winter precipitation and summer temperature. Interannual variations in several regions indicate that ELA rises about 10^2 m for each kelvin rise in summer temperature. However, on many glaciers the spatial patterns of b vary from year to year. Furthermore, in all cases it is important to recognize that phenomenological relations based on short-term records during which the atmospheric composition is nearly constant may not work for longer time scales over which concentrations of radiatively active "greenhouse" gases and volcanic aerosols could change. Such changes in the atmosphere could alter relationships between radiation and temperature (Oerlemans, 1989).

A warming climate would tend to increase accumulation at high altitudes because of availability of precipitable water vapor. It would also increase melting at low altitudes. Together these effects would increase the elevation gradient of b and alter the trends of its spatial variation. It is also important to distinguish between winter and summer temperature changes. In temperate conditions, increases in winter temperature may reduce accumulation by increasing the proportion of rain to snow with negative effect on b. In colder circumstances at high latitude or altitude, the opposite effect on b may happen because of increased total precipitation, which in any case occurs as snowfall.

A method for predicting calving is crucial for modeling the mass balance of tidewater glaciers and both of the present ice sheets. In temperate conditions floating ice shelves do not form and calving releases relatively small bergs with roughly equal vertical and horizontal dimensions. Calving rate is correlated with water depth and height of the ice cliff and is apparently related to local stress near the ice front. In polar conditions, where ice shelves form, calving occurs episodically by release of large tabular bergs with horizontal dimensions reaching tens of kilometers indicating that a large-scale stress regime is responsible. Quantitative predictive models are still not available for either case.

3.2 Glacier Motion

Glaciers, especially those flowing down valleys, are commonly described as "rivers of ice." Physically, the motion of a glacier differs from the flow of a river (a linearly viscous fluid flowing with no slip in a channel). The glacier speeds are so low (Table 2) that inertial forces are negligible. As a result the dynamic nonlinearity leading to turbulence in rivers is absent, and in this way glacier flow is simpler. Other features make it more complex. As discussed above, motion over the base introduces a relatively complex boundary condition. In addition, the fluidlike (creep) deformation of ice is described by a nonlinear relationship (flow law) between rate of deformation (ϵ) and stress (τ), which together with the basal motion introduces a rheological nonlinearity.

The foundation for modeling of glacier motion is an accurate description of a flow law for the ice. Experiments indicate that ice creep can be approximated as being like that of a "quasiviscous" fluid with an effective viscosity (η) that depends inversely on ϵ such that $\eta \propto \epsilon^{-(1-1/n)}$ with $n \cong 3$. The behavior may also be equivalently described by a power-law relationship $\epsilon \propto \tau^n$. Effective viscosity η is affected by other parameters. Most important is the thermal state. As temperature increases or as water content rises at the melting point, η decreases strongly. It is also sensitive to impurities and texture. Although many details have yet to be fully explored, a practical ice-flow law has been relatively well defined (Budd and Jacka, 1989).

The flow law of ice, field equations describing conservation of mass and momentum, and boundary conditions based on processes described in Sec. 2.2 define a fluid mechanical problem for the calculation of the instantaneous velocity field from the current geometry, thermal state of the ice, and density distribution acted on by gravity. Because of the nonlinearity and the geometrical complexity of real glaciers, this problem has been solved only with various simplified geometrical or rheological assumptions. (See Paterson, 1981; Hutter, 1983.)

The simplest model is a slab of constant thickness (h) and slope (α) with no gradients sideways across the flow or longitudinally in the down-slope flow direction. Then the stress distribution is statically determined, and the corresponding vertical variation of deformation rate is easily integrated for any specified depth profile of parameters affecting the ice-flow law (e.g., temperature). When the parameters are independent of depth one finds that speed $u \propto \alpha^n h^{n+1}$, where n is the power in the flow law.

This slab model is reasonably applicable over wide areas of the interior of ice sheets where horizontal gradients are small. These "inland" portions have smooth upper surfaces and flow lines running down slope with local and usually small perturbations from topographic features in the underlying terrain. However, the slab model is not applicable at many locations where horizontal gradients are large. This condition exists at most locations in valley glaciers as a result of drag from the sides and changes along the length in the valley profile or the basal velocity. Toward the edge of ice sheets, where mean thickness decreases, the flow commonly concentrates either into deep outlet glaciers confined between peaks reaching close to or to the surface (nunataks) or into ice streams that sometimes have no obvious relation to basal topography, but move rapidly over their bases because of special basal conditions (Bentley, 1987) [see Fig. 1(c)]. The outlet glaciers and ice streams are characterized by "fast" flow rates much higher than the more normal "slow" speeds of the surrounding ice (Table 2). The boundary zones between these fast ice currents and adjacent slowly moving ice are examples of locations where horizontal gradients are clearly important. Several analytical theories allow approximate corrections for side drag and unbalanced longitudinal forces, but numerical methods are often needed for accurate calculation.

To find the time-dependent evolution of geometry, density, and velocity, it is necessary in principle to track the parameters affecting the flow law by introducing conservation equations for energy, liquid water mass, and masses of impurities with associated boundary conditions. It is also necessary to track textural changes. While observational effects arise from all of these parameters, temperature is most important. Recent numerical models address the full coupling between the mass and heat flow caused by the strong temperature dependence in the flow law and heat advection (Huybrechts, 1990). Modeling of water, impurities, and texture is still not well developed.

For some purposes a detailed knowledge of the distribution of density (ρ) and velocity ($\mathbf{u}$) is important. One example is to determine the age, accumulated finite strain, and origin location for ice along the length of ice cores to aid in their interpretation. However, generally, to calculate the evolution of the large-scale geometry, quantities integrated over the ice thickness h can be used, and the details of the horizontal gradients are unimportant. The total mass (m) in a vertical column through the ice is given by the integral of density. The ice transport is conveniently described by the horizontal ice flux density $\mathbf{q}$ given by integrated density times horizontal velocity. The isothermal slab model gives $q \propto \alpha^n h^{n+2}$; high sensitivity of q to h and α is a feature of all theories as long as averages are taken over horizontal dimensions substantially larger than h. This approach leads to a tractable mathematical theory for advance and retreat of glaciers discussed below. The major lack is the incomplete understanding of basal motion and absence of a dependable mathematical description of the process. Predictions when basal velocity is the major contribution to the motion are especially problematical.

3.3 Advance and Retreat

Mass conservation applied to a full thickness column can be expressed as

$$dm/dt = b - \nabla \cdot \mathbf{q}. \qquad (1)$$

This equation describes quantitatively the parallel action of mass balance b and ice flow $\mathbf{q}$ on the mass distribution and corresponding thickness pattern (Fig. 2). The evolution of a glacier can be calculated from an initial condition and subsequent climate history by integration of Eq. (1) using models to predict b and $\mathbf{q}$.

When the balance-rate distribution changes on a glacier, the information about the change is spread over the full area of the glacier. The effects of the change appear at the margin spread out in time in a way that can be described by two fundamental time scales. The first is a memory length (T_m) that describes the time it takes a glacier to complete its response to the change. On this time scale or longer, the length and volume are closely related. The second is a shorter lag time (T_l) that describes the time interval between the climate change and the resulting maximum rate of change at the margin. On this time scale, the length and volume changes are not necessarily related and can even have opposite signs. Simple models of glaciers show that T_m and T_l are proportional to the length divided by velocity; see Paterson (1981) and Hutter (1983) for reviews. T_m is approximately equal to the mean residence time (Table 2) and can be expressed alternatively in terms of thickness and balance rate (Johannesson *et al.*, 1989). For typical valley glaciers with lengths in the range of a few kilometers to tens of kilometers and velocity in the range of a few tens to a few hundreds of meters per year, T_l is a few decades and T_m is decades to a few centuries. Because of the large range in lengths and velocities, there is a corresponding range in time scales that is in part responsible for diverse response histories of glaciers

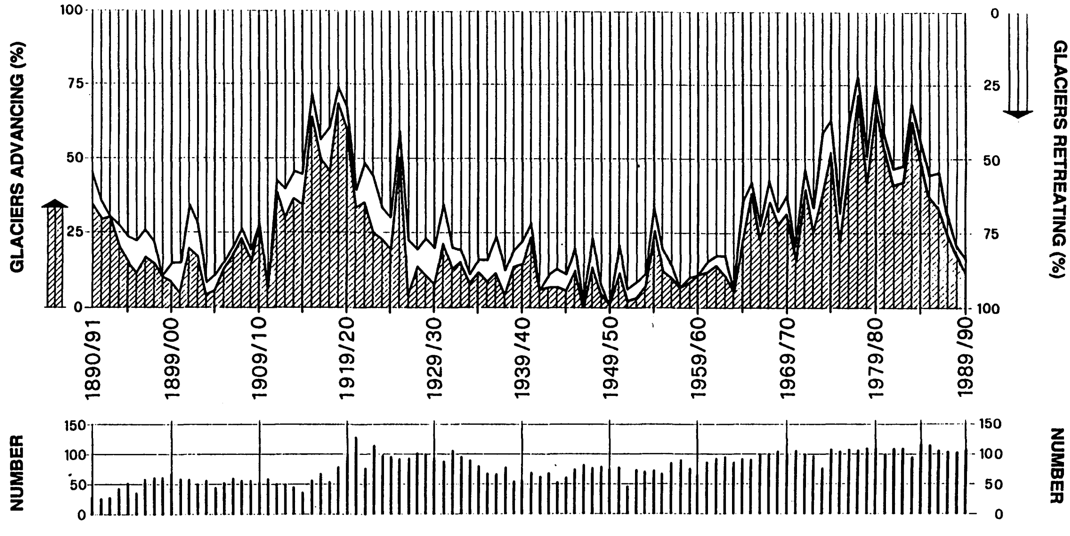

FIG. 3. Upper panel shows percentages of glaciers advancing (lower curve, scale on left), retreating (upper curve, scale on right) and stationary (space between curves) on an annual basis in the Swiss Alps over the last century. Lower panel: Number of glaciers that were observed in each year. (Adapted from Kasser *et al.*, 1983).

to climate change. As climate changes on different time scales, some glaciers may be advancing while others are retreating (Fig. 3).

As a glacier advances or retreats the elevation of its surface changes. This causes a change in balance rate on the surface in addition to those caused directly by climate. This geometry–mass-balance feedback tends to amplify changes, increase T_m, and allow nonunique steady states. For example, under the present climate a large, thick ice sheet could theoretically exist in steady state over much of northern North America. However, such an ice sheet cannot grow with the present climate because the ground surface is too low and warm to develop perennial ice.

Another example of geometry–mass-balance feedback occurs in tidewater glaciers (Meier and Post, 1987). As a tidewater glacier retreats from a moraine shoal into deeper water, the calving rate rises. This increase in loss with reduced size causes a rapid unstable retreat through deep water to a location where the bed approaches sea level. Readvance apparently occurs only very slowly because it requires the construction of a proglacial shoal by deposition in front of the glacier.

As the geometry of a glacier changes, the stress in it and the resulting velocity change. This is fundamental to the normal adjustment to climate discussed above. However, it is also possible that an incremental change in geometry can cause a dramatic shift in flow state from "slow"' to "fast" or visa versa. The primary example of this phenomenon is surge behavior. Although glaciers showing this behavior constitute only a small fraction of all glaciers, they are common in certain mountain areas and their rapid changes cause various hazards. These glaciers cannot achieve steady state in either the "slow" or "fast" states and are forced to switch between intervals of slow quiescence and fast surge (Raymond, 1987). The transitions in speed and the resulting time-dependent geometry are independent of any immediate climatological triggering. At least in some cases, the switching between quiescence and surge has to do with transformation of the drainage system along the bed that traps water and initiates the "fast" flow state as a result of very high basal velocity. Although the same hydrological processes probably do not operate beneath ice sheets and the ice streams that drain them, the possibility of dramatic auto-oscillations in these systems cannot be ruled out. Indeed, evidence has been found for changes in the speed of ice streams (Alley and Whillans, 1991), and the causes have not been identified.

4. GLACIER ICE IN A CHANGING CLIMATE

The Earth has been warming over the last century, and increasing atmospheric concentrations of radiatively active gases like carbon dioxide and methane suggest that heightened "greenhouse" warming will occur in the future. How might glaciers respond to such a scenario of "global warming"?

It is relatively certain that most glaciers in temperate latitudes would continue to experience rising ELAs, lose volume, and retreat because of increased summer melting and in some cases increased proportion of rain to snow in winter. In some exceptional cases, especially on glaciers with high-altitude accumulation areas, increased snowfall could cause growth.

In Greenland a warmer climate would cause increased melting near the margins, higher ELA, and larger ablation area. There could also be an increase in accumulation in the high interior. Although these competing effects cannot be predicted precisely, reasonable parametrizations are possible in terms of temperature change. The resulting model indicates that the Greenland Ice Sheet would retreat if the climate were to warm (Letreguilly *et al.*, 1991). The changes are predicted to be modest for temperature rises less than ~ 2 K. Temperature rise above ~ 6 K would cause disappearance of the ice sheet leaving only small ice caps and glaciers. The time scale to change the volume of the ice sheet to a new steady-state geometry is $\sim 10^4$ years.

The Antarctic Ice Sheet surface is so cold that even large warming would not produce melting with runoff from the upper surface. The principal direct meteorological effect would be increased accumulation and growth of the ice sheet. However, there are dynamic effects that could act to reduce the ice volume. Increased temperature would soften the ice, increase flow rates, and allow a thinning. This process has little effect on a short time scale when temperature changes affect only the near surface. For time scales longer than $\sim 10^4$ years thermal effects are advected and diffused to sufficient depth to speed the motion. It turns out that increased accumulation and temperature are approximately counterbalancing (Huybrechts, 1990).

The location of the ice-sheet margin is determined by calving at the coast line (except in embayments where ice shelves form). With a constant sea level and fixed land base, even large changes in accumulation rate produce only modest changes in volume. This behavior arises in part from the high sensitivity of the ice flow to changes in thickness and associated stress. Therefore, sea level fluctuations have had a dominant influence on the ice sheet dimensions in the past. During the Pleistocene ice ages, as northern hemisphere ice sheets grew and sea level dropped, the Antarctic Ice Sheet could expand on the resulting larger land base. This is one possible contribution to synchronous oscillations of northern and southern hemisphere ice sheets, even though the radiation forcing from changes in orbital parameters is quite different in the two hemispheres. Some changes could still be occurring as a result of delayed local sea level effects around the ice sheet margin because of continuing isostatic rebound arising from earlier ice volume reduction. With climate warming from present conditions, changes in the Antarctic Ice Sheet are not likely to be forced by sea-level rise from outside, although feedback from significant shrinkage of the Greenland Ice Sheet cannot be ruled out on a long time scale.

Of greater concern on the time scale of 10^2–10^3 years is the possible alteration of the drainage of the ice to the ocean through ice streams. This is of primary concern for the West Antarctic Ice Sheet because of its marine character and possible collapse by retreat of its grounding line (Mercer, 1978). Retreat could be instigated by thinning of the ice shelves through interaction with a warmer ocean, reduction of back stress on the ice streams feeding the ice shelves with a resulting increase in discharge, and thinning of inland ice. Major unknowns concern the amount of heat that could be carried by ocean currents beneath the ice shelves and lost to melting. Nor is present understanding of controls on the velocity, width, and length of ice streams sufficient to predict how the ice streams would respond to such a change in boundary conditions. The response of the West Antarctic Ice Sheet remains as one of the important unknowns of global change, especially future sea level.

GLOSSARY

Accumulation: Mass of ice added per unit area to a glacier or ice sheet.

Ablation: Mass of ice removed per unit area from a glacier or ice sheet.

Calving: Breaking off of icebergs at a glacier or ice-sheet margin with the ocean or a lake.

ELA: Elevation of the boundary between the area of accumulation and the area of ablation at the end of the summer season.

Geothermal Gradient: Temperature gradient in the Earth conducting the geothermal heat flux upward from the interior of the Earth.

Ice Shelf: Glacier ice that has flowed from the land surface onto the sea or a lake and is floating.

Ice Stream: Fast current of ice through surrounding relatively slowly moving ice.

Marine Ice Sheet: An ice sheet grounded on a bed that is below sea level even after isostatic rebound of the bed once the ice is removed.

Mass Balance Rate: The difference between the rates of accumulation and ablation.

Surge-Type Glacier: A glacier that oscillates between extended quiescent intervals of slow flow and relatively brief surge intervals of relatively very fast flow.

Tidewater Glacier: A glacier reaching sea level that calves into the ocean.

List of Works Cited

Alley, R. B., Whillans, I. M. (1991), *Science* **254**, 959–963.

Bentley, C. R. (1987), *J. Geophys. Res.* **92** (B9), 8843–8858.

Blankenship, D. D., Bentley, C. R., Rooney, S. T., Alley, R. B. (1987), *J. Geophys. Res.* **92** (B9), 8903–8911.

Budd, W. F., Jacka, T. H. (1989), *Cold Regions Sci. Technol.* **16**, 107–144.

Crowley, T. J., North, G. R. (1991), *Paleoclimatology*, Oxford Monographs on Geology and Geophysics No. 18, New York: Oxford University Press.

Delmas, R. J. (1992), *Rev. Geophys.* **30** (1), 1–21.

Drewry, D. (1986), *Glacial Geologic Processes*, London: Edward Arnold.

Haeberli, W., Muller, P., Allan, P., Bosch, H. (1989), in: J. Oerlemans (Ed.), *Glacier Fluctuations and Climate Change*, Dordrecht: Kluwer Academic, pp. 77–101.

Hutter, K. (1983), *Theoretical Glaciology, Material Science of Ice and the Mechanics of Glaciers and Ice Sheets*, Dordrecht: Reidel.

Huybrechts, P. (1990), *Ann. Glaciology* **14**, 115–119.

Johannesson, T., Raymond, C., Waddington, E. (1989), *J. Glaciology* **35** (121), 355–369.

Kamb, W. B., LaChapelle, E. R. (1964), *J. Glaciology* **5** (38), 159–172.

Kasser, P., Aellen, M., Siegenthaler, H. (1983), *97. und 98. Bericht, Glaziologisches Jahrbuch der Gletscherkommission der SNG*, Zurich: Schweizerische Akademie der Naturwissenschaften.

Letreguilly, A., Huybrechts, P., Reeh, N. (1991), *J. Glaciology* **37** (125), 149–157.

Lliboutry, L. (1987), *J. Geophys. Res.* **92** (B9), 9101–9109.

Meier, M. F. (1984), *Science* **226**, 1418–1421.

Meier, M. F., Post, A. (1987), *J. Geophys. Res.* **92** (B9), 9051–9058.

Mercer, J. H. (1978), *Nature* **271**, 321–325.

National Research Council (NRC) (1985) in: M. F. Meier (Ed.), *Glaciers, Ice Sheets, and Sea Level: Effects of a CO_2-Induced Climate Change*, Washington, D.C.: National Academy Press.

Oerlemans, J. (1989), *Climate Change* **15**, 151–174.

Paterson, W. S. B. (1981), *The Physics of Glaciers*, 2nd ed., New York: Pergamon.

Peltier, W. R. (1987) in: W. F. Ruddiman, H. E. Wright (Eds.), *North America and Adjacent Oceans During the Last Deglaciation. The Geology of North America*, Vol. K-3, Boulder: Geological Society of America, pp. 155–182.

Raymond, C. F. (1987), *J. Geophys. Res.* **92** (B9), 9121–9134.

Rothlisberger, H., Lang, H. (1987), in: A. M. Gurnell, M. J. Clark (Eds.), *Glacio-fluvial Sediment Transfer*, New York: Wiley, Chap. 10, 207–254.

van der Veen, C. J. (1991), *Rev. Geophys.* **29** (3), 433–455.

Further Reading

Delmas, R. J. (1992), *Rev. Geophys.* **30** (1), 1–21. Reviews constitution of glacier ice and implications for past climate.

Oerlemans, J. (1989), *Climate Change* **15**, 151–174. Reviews problems for modeling the future response of land ice to global warming.

van der Veen, C. J. (1991), *Rev. Geophys.* **29** (3), 433–455. Most up to date review of the mass of land ice and its changes.

Many authors (1987), *J. Geophys. Res.* **92** (B9). A collection of papers about fast flow of glaciers and ice streams.

GLASSES

Robert H. Doremus, *Materials Engineering Department, Rensselaer Polytechnic Institute, Troy, New York, U.S.A.*

Introduction 251
1. **Types and Uses of Glasses** 251
2. **Melting and Processing** 252
2.1 Melting 252
2.2 Forming 253
2.3 Fiberglass 254
2.4 Sintering and Sol-Gel Methods 254
3. **Structure** 255
4. **Phase Separation and Crystallization** 255
4.1 Phase Separation 256
4.2 Kinetics of Phase Transitions 256
4.3 Crystallization 256
4.4 Glass Ceramics 257
5. **Viscosity and the Glass Transition** 257
5.1 Viscosity 257
5.2 Glass Transition 259
5.3 Stress Relaxation 259
6. **Strength of Glass: Fracture, Fatigue, and Strengthening** 260
6.1 Strength Distributions and Surface Treatments 260
6.2 Observations of Fracture Surfaces 260
6.3 Fracture Criteria 261
6.4 Faulting 262
6.5 Fatigue 262
6.6 Strengthening 263
7. **Electrical Properties and Ionic Transport** 263
7.1 Ionic Conductivity 263
7.2 Electronic Conductivity 265
7.3 Dielectric Properties 265
8. **Chemical Properties** 265
8.1 Surfaces 265
8.2 Diffusion and Reactions of Gases in Glasses 265
8.3 Reaction of Glass with Water 266
9. **Optical Properties** 266
9.1 Vitreous Silica 266
9.2 Other Network-Forming Oxides 267
9.3 Alkali Silica Glasses 267
9.4 Transition-Metal Ions and Color 267
9.5 Applications 267
9.6 Nonlinear Optical Properties 268
Glossary 268
Works Cited 268
Further Reading 269

INTRODUCTION

Glass is a vital material in modern technology. In addition to the traditional uses in containers, windows, and for decorative objects, new applications in lamps, optical components, and electronics are continuously being developed. Special uses such as fiber-optic waveguides, telescope mirrors, laser optics, radioactive waste encapsulation, and biomedical implants testify to the great value and diversity of glass in contemporary society.

The emphasis in the present article is on silicate glasses because of their overwhelming commercial importance. Other glasses, such as fluorides, chalcogenides, and glassy alloys are described in appropriate sections because of recent research interest in them. This article has sections on melting and processing of glass, glass structure and microstructure, viscosity, and glass properties. A bibliography of books on glass is at the end of the article, as well as a list of highly selected review and research articles.

1. TYPES AND USES OF GLASSES

By far the greatest use of glass is of the soda-lime silicate composition, which has been known for at least ten millennia. This composition is relatively easy to melt from readily

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Table 1. Properties of commercial silicate glasses.

Glass type	Density $(gm/cm)^3$	Softening temp. (°C) (see text)	Annealing temp. (°C) (see text)	Young's modulus (GPa)	Coeff. thermal expansion 0–300 °C (10^{-6})	Log electrical resistance (Ω cm) 250 °C	At 25 °C and 10 Hz	
							Dielectric constant	Loss factor (%)	Refractive index @0.589 μm
Vitreous silica	2.20	1670	1190	73	0.5	12	4	0.01	1.46
Soda-lime	2.50	700	500	72	9.2	6.6	7.3	5.1	1.52
Pyrex Borosilicate	2.23	820	565	64	3.3	7.9	4.6	2.2	1.47
Alkali lead	2.86	630	435	61	8.9	8.9	6.8	1.0	1.54
Lime-magnesia aluminosilicate	2.52	915	710	88	4.2	10.5	6.3	2.3	1.53
Sodium borosilicate	2.24	710	500	60	4.6	8.7	4.9	1.4	1.48
E fiberglass	2.61	840	670	72	6.0		6.4		1.55
High lead	5.42	440	363	55	8.4	10.6	15	3.3	1.86
Barium lead	2.66	646	446	70	8.9	8.9	6.3	1.1	1.51
Barium optical	3.21	639		73	8.0				1.57

available raw materials and is highly resistant to chemical attack, including by water. A large number of different oxides can be incorporated into a silicate glass, leading to a wide variety of commercial glasses; however, only a few glasses make up the large majority of glass production. A summary of properties and uses of important commercial glasses is in Table 1, and their compositions are in Table 2.

Glass was first used for decorative objects, and exhibits such as in the Steuben store in New York City and the Corning Museum in Corning, New York, demonstrate the recent and continuing activity in glass as an artistic medium. The second use of glass, for containers for food, perfume, and a vast number of commercial liquids continues as the largest consumer of glass, in spite of strong competition from polymers and aluminum. Flat glass, mainly for windows in buildings and vehicles, is the second largest consumer of glass. Many special applications of smaller amount but high value include lamp envelopes and seals, many optical uses, fibers in polymer matrices ("fiberglass"), and consolidation of radioactive waste. Vitreous silica, a highly pure SiO_2 composition, is especially valuable because of high optical transparency, low sensitivity to thermal shock, and high chemical durability; applications include crucibles for melting semiconductors, lamp envelopes, fiber-optic waveguides, and telescope mirrors.

Glass can be defined as an amorphous solid. A material is amorphous when it has no long-range order—that is, no regular repeating structure on the scale of many atomic distances (see AMORPHOUS MATERIALS, STRUCTURE OF). A solid is a rigid material that does not flow when subjected to moderate forces; more precisely, a solid can be defined as a material with a viscosity greater than about 10^{14} Pa s (10^{15} poises).

2. MELTING AND PROCESSING

2.1 Melting

Raw materials for making silicate glasses are quartz sand, sodium carbonate or nitrate, calcium and magnesium carbonate, and oxides such as alumina, lead oxide, sodium borate, zinc oxide, and arsenic and antimony

Table 2. Compositions (wt %) of glasses in Table 1.

Glass	SiO_2	B_2O_3	Al_2O_3	CaO	MgO	PbO	Na_2O	K_2O
Vitreous silica	100							
Soda lime	72	1	2	5	4		15	
Pyrex borosilicate	81	13	2				4	
Alkali lead	77					8	9	5
Lime-magnesia aluminosilicate	64	5	10	9	10		1	1
Sodium borosilicate	68	24	1				7	
"E" fiberglass silicate	55	7	15	21			1	1
High lead	4	11	11			74		
Barium lead	66		4	(12 BaO)		4	6	7
Barium optical	48	5	1	(29 BaO, 9 ZnO)			1	7

oxides. These materials have traditionally been mined directly, with little or no purification, and so were cheap but of variable impurity content. Some deposits of quartz crystal and quartz sand are probably the purest of naturally occurring materials, having total impurity contents of less than about 100 parts per million by weight.

Glass has traditionally been melted in a "batch" in a pot or crucible of refractory ceramic material. Raw materials are mixed to the desired composition and either heated in the crucible or added to already molten glass. The low–melting-point constituents (alkali oxides) melt and dissolve the higher–melting-point constituents such as quartz and alumina.

Quantity production of glass, especially soda-lime and borosilicate compositions, is continuous in a large "tank" or furnace made of special refractory brick; a characteristic tank is 15×25 m in size, with glass 1.5 m deep, giving a daily production of as much as 700 metric tons of glass. The tank is heated by combustion of oil or natural gas across the glass surface. Melting temperatures are about 1200–1300 °C. The furnace is separated into two positions by a refractory "wall" over which glass flows. The mixed raw materials are fed onto the top of the molten glass in the melting chamber; the molten glass flows across the wall into the refining chamber, in which the glass becomes more homogeneous and bubbles are removed.

In the refining chamber the glass is "fined" by the removal of bubbles. The bubbles contain trapped air (oxygen and nitrogen) and decomposition products of the raw materials, mainly carbon dioxide. Large bubbles rise to the top of the melt, but most bubbles are too small to be removed in this way; the gases in them must dissolve in the melt to collapse the bubbles. This dissolution is aided by fining agents added to the glass batch. Arsenic and antimony oxides are common fining agents. They have two oxidation states, and increase dissolution rates of gases by changing the oxidation state of the melt. Sulfates also are common fining agents; they also function by changing the oxidation state of the melt.

2.2 Forming

Glass containers are formed from glass that flows out of the continuous furnace. A "gob" or portion of glass is placed in a mold where a first shape is formed by pressing or blowing. This initial shape is removed from the mold, reheated, and then placed in a second mold in which the final shape is blown. The continuous machine (IS or individual section machine) on which these steps are done rapidly produces a very large number of units; a usual cycle time for one container is 10–15 s.

Lamp bulbs are made by blowing into a ribbon of glass emitting from the furnace on a "ribbon machine." Again a high rate of units is produced, up to 2000 per minute. A special nozzle blows glass in the ribbon into a mold.

Glass objects such as plates, tumblers, and vases can be made by pressing molten glass in a mold with a metallic plunger. Patterns on the mold surface are imprinted into the glass and make it look more expensive than it is.

More than 100 years ago window glass was made by hand by either the crown or cylinder process. In the crown process a gob of glass was blown into a sphere and one side flattened by pressing it against a plate. An iron rod was attached to the flat part and the blow pipe removed. The globe was then spun into a flat disk from which panes could be cut. In the cylinder process the glass was formed into a large cylinder, which was split open and flattened.

Later sheet glass was drawn directly from the melting surface, flowed through rollers, and then ground and polished. These processes are expensive. A new method of making more inexpensive flat glass of high quality was developed by Alistair Pilkington of the Pilkington Glass Company in England. In this float process, a continuous strip of glass from the melting furnace floats onto the surface of a molten metal, usually tin, at a carefully controlled temperature. The flat surface of the molten metal gives a smooth, undistorted surface to the glass as it cools. This process is continuous, and high production speeds are possible. Today most flat glass is made by this method.

2.3 Fiberglass

Glass fibers are finding increasing application in reinforcing polymer-matrix composites ("fiber glass"), which are used in automobile and boat bodies, aircraft, electronic circuit boards, and a variety of other ways. Fiberglass fabrics serve as protective apparel, industrial and decorative fabrics, and tire cord. Short fibers are used in thermal insulation. A typical fiberglass composition is given in Table 2.

In making continuous fiberglass, molten glass from a melting furnace is fed to a drawing apparatus that consists of clusters of many holes called bushings. Each bushing has hundreds of holes or nozzles through which the glass is drawn into fibers of diameter about 10–100 μm in diameter. The fibers are cooled by a water spray, and a polymeric coating or "sizing" is applied to protect and lubricate the fibers. Instead of drawing glass directly from the furnace, glass can be cast into spherical "marbles" one or a few cm in diameter, which are then fed into the heated bushings.

Short fibers can be formed either directly from glass from the melting furnace or from marbles. The molten glass is fed into a rotating cylinder through a large number of small holes in the walls of the cylinder, and inside the cylinder the glass streams are broken up by an axial high-velocity flow of gas. The short fibers formed in this way are sprayed with binder as they fall onto a collecting belt which carries them through an oven that polymerizes the binder.

2.4 Sintering and Sol-Gel Methods

Several methods of melting and forming glass have been developed for special applications or properties. Glass powder can be pressed and sintered similarly to powders of crystalline ceramics. This is expensive compared to more traditional methods of melting, but can be useful for making high–melting-point glasses and adding small amounts of desired impurities to them.

Sol-gel methods of glass making are finding increasing application in making highly pure and homogeneous glasses or special shapes such as microshells. In one method a mixture of liquid metal alkoxides is hydrolyzed, polymerized, and then dehydrated to form a homogeneous glass at temperatures much below those ordinarily used for glass melting. The hydration and polymerization are carried out in an acidic alcohol-water solution at temperatures up to 60 °C; then the mixture can be formed or made to gel in bulk by heating to somewhat higher temperatures; finally the gel is pyrolyzed and dehydrated by heating at temperatures from 500 to 1000 °C to form glass. Some superior products of this method are homogeneous multicomponent glasses, compositions hard to form into glasses by direct melting because of incomplete melting and crystallization on cooling, very pure glasses, and composites of glasses with other materials. Low forming temperatures save energy, allow less corrosion-resistant container materials, and allow formation of composites with materials not resistant to higher temperatures.

Special shapes such as fibers and thin films can be made by the sol-gel method. Glass microshells for laser targets have been made by dropping gel particles through a temperature gradient in a furnace tower; the organic compounds and water help to blow the glass into thin-walled shells.

3. STRUCTURE

The structure of amorphous materials is treated in a separate article in the Encyclopedia (AMORPHOUS MATERIALS, STRUCTURE OF), and so only a short discussion of glass structure is included here. The structure of glassy polymers is treated in articles on these materials (see POLYMERS, STRUCTURE OF).

The most useful experimental techniques for studying the structure of glasses are x-ray and neutron diffraction and infrared and Raman spectroscopy. Techniques such as nuclear magnetic resonance (NMR), x-ray photoelectron spectroscopy [XPS or electron spectroscopy for chemical analysis (ESCA)] and x-ray absorption fine structure (EXAFS) with longer-wavelength x rays are useful for special structure studies, especially to determine the coordination numbers of atoms in glasses.

The building block of all silicates, both crystalline and amorphous, is the rigid silicon-oxygen tetrahedron. In silicate glasses these tetrahedra are linked by silicon–oxygen bonds into a three-dimensional network. The network can be partially broken up by other oxides in the glass, especially alkali and alkaline-earth oxides (see the review by Wright *et al.*, 1991).

From a detailed x-ray diffraction study of pure amorphous silica, Warren and later Mozzi and Warren (1969) concluded that its structure could best be described as a random network. In this model the rigid silicon-oxygen tetrahedra are linked into a three-dimensional network; randomness results from a distribution of silicon-oxygen-silicon bond angles around an average value of 144°. The distribution of angles around the maximum is rather narrow compared to a completely random distribution of bond angles from 90° to 180°. Thus the structure of amorphous silica is quite uniform at atomic distances, although there is no order beyond several layers of tetrahedra.

In silicate glasses containing alkali and alkaline-earth oxides, these cations apparently are not randomly distributed throughout the glass; their average separation is considerably less than for a uniform distribution. It is possible that alkali ions form in sheets, as they do in certain crystalline alkali silicates. This "intermediate"-range structure has been emphasized in recent studies of multicomponent silicate glasses and other compositions such as sulfides and selenides (Gaskell *et al.*, 1991).

Metals and alloys have been made into glasses by different methods of rapid cooling (see METALLIC GLASSES). The structure of glassy metals with no more than a few percent of impurities can be described as a dense random packing of hard spheres (Bernal, 1960; Finney, 1970). The packing fraction is 0.637, compared with 0.74 for close packing of spheres of the same size in a crystalline structure, such as face-centered cubic and close-packed hexagonal. This model fits the radial distribution function derived from scattering experiments.

The structures of amorphous alloys of compounds M_4X, in which M is a transition or noble metal and X a nonmetal, have been extensively studied. The best model seems to be a slightly distorted dense random packing of hard spheres for the M atoms, with the X atoms located interstitially in polyhedra in the structure. Other structures of alloys are described by Cargil and Spaepen (1981).

The structure of zirconium fluoride glasses is of special interest because it is quite different from that of silicate glasses, especially in that the coordination number of the glass-forming zirconium atoms is high. Coordination numbers of 7 and 8 for zirconium atoms have been deduced from x-ray radial distribution functions and molecular-dynamics calculations, and several different structures have been proposed for these glasses (previous work is reviewed in Ko and Doremus, 1991). From a comparison of infrared and Raman spectra of zirconium fluoride glasses and crystals, a structure of these glasses based on eight-coordinated zirconium-fluoride units bonded into a three-dimensional network was proposed (Ko and Doremus, 1991). This structure is similar to that of β-ZrF_4, which has the Archimedean antiprism structure in which zirconium is coordinated with eight fluorines.

4. PHASE SEPARATION AND CRYSTALLIZATION

Glasses normally consist of one amorphous phase. However, in certain glass compositions, especially borosilicates, the glass can separate into two amorphous or liquid phases. Crystals also grow in glasses under different conditions, sometimes enhancing properties, as in glass ceramics, but more often causing cracking and deterioration of properties (devitrification).

4.1 Phase Separation

Binary silicate mixtures that contain alkaline-earth oxides, except for barium, and many transition-metal oxides show equilibrium phase separation at high temperatures, typically above about 1700 °C. In binary alkali silicate glasses there is no equilibrium phase separation, but these mixtures are easily cooled to glasses without crystallization, and some show liquid-liquid phase separation at temperatures below the liquidus temperature of crystallization. The maximum temperature (or "consolute" temperature) for this metastable phase separation is about 850 °C at 8% Na_2O–92% SiO_2 in sodium silicate glasses and about 1000 °C at somewhat higher Li_2O concentration in lithium silicate glasses.

The commercially important soda-lime silicate glasses can also show phase separation, which can degrade properties such as chemical durability and resistance to crystallization. Addition of a small amount of aluminum oxide to these glasses strongly reduces the tendency to phase separation.

The sodium borosilicate glasses show phase separation in a wide variety of compositions. This property is the basis for making 96% silica glass ("Vycor"). The starting material is an alkali borosilicate glass that separates into two interconnected phases. The high-silicate phase is resistant to hydrochloric acid, whereas the sodium borosilicate phase dissolves in this acid leaving a porous glass with a silica matrix. This porous glass can be consolidated at about 1000 °C into a silica-rich dense glass with properties similar to vitreous silica, which must be melted at temperatures above 1800 °C. The porous glass itself is useful for filters and absorption, and as a matrix for fillers. The pore size can be changed by changing the initial composition and by annealing at temperatures in the range 600–700 °C.

4.2 Kinetics of Phase Transitions

The usual mechanism of phase transformations is nucleation of a small region of the new phase composition followed by its growth. The rates of nucleation and of growth in a glass both reach a maximum as the temperature is decreased below the phase-boundary temperature. The driving forces for nucleation and for growth increase as the temperature decreases below the phase boundary, causing an increase in the rate of nucleation or growth. As the temperature decreases still further, transport rates (diffusion or interfacial processes) become controlling, and the rates decrease as the temperature decreases. Nucleation and growth of phases is summarized in more detail in Doremus (1985).

Another mechanism of phase transformation called spinodal decomposition can give a morphology of interconnected phases. In a nucleation and growth mechanism, small fluctuations in composition of the initial phase increase the total free energy of the system, and so only regions of the equilibrium composition of the new phase form. In spinodal decomposition small composition fluctuations reduce the total free energy, and so these fluctuations occur throughout the matrix phase with a certain extent (wavelength), and as transformation proceeds the initial concentration variations increase, and the wavelength of the spatial variations of composition remains the same. Spinodal decomposition in glasses was modeled theoretically by Cahn and Charles (1965), and agreed well with experimental results.

The experimental measurements of Hammel (1967) on nucleation rates of phase separation in a soda-lime silicate glass provide one of the very few tests of nucleation theory in condensed systems and agree well with this theory.

4.3 Crystallization

Glass is metastable with respect to crystalline phase at temperatures below its equilibrium liquidus temperature. The liquidus temperatures for commercial soda-lime glasses are usually between 900 and 1000 °C for crystallization of the compound "divitrite" or $Na_2O{\cdot}3CaO{\cdot}6SiO_2$. In a zirconium fluoride glass, the liquidus temperature can be as low as 500°. Glasses can form because the rate of crystallization is low.

Crystallization or "devitrification" of glass can lead to serious problems in glass manufacture because of the resultant changes in glass properties, such as viscosity and coefficient of thermal expansion. High stresses resulting from nonuniform contraction on cooling can cause fracture of the piece. Thus, glass technologists have always been careful to avoid crystallization by not holding their glasses at temperatures where crystals grow rapidly in them.

In surface crystallization, impurities are the most important source of nucleation. On cooling from the melting temperature, dust particles adhere to the glass surface, giving preferred nucleation. Other impurities such as alkalis can corrode the surface, leading to lower glass-air surface tension and enhanced nucleation.

Experimental nucleation rates have been measured in lithium disilicate, $BaO{\cdot}2SiO_2$, and $Na_2O{\cdot}2CaO{\cdot}3SiO_2$ glasses (James, 1985). The preexponential factor in the nucleation rate equation as calculated from these was many orders of magnitude higher than expected from the theoretical value. James showed that this discrepancy could be explained by a temperature dependence of the interfacial free energy resulting from a small negative interfacial entropy. It is also possible that the simple geometrical factors assumed for interfacial rearrangement during nucleation are not valid for this system.

Nucleation rates in sodium disilicate and anorthite ($CaO{\cdot}Al_2O_3{\cdot}2SiO_2$) showed the temperature dependence expected from the theory, and interfacial energies of 0.055 and 0.19 J/m^2 were calculated (Cranmer *et al.*, 1981). The preexponential factor was found to be close to that expected theoretically.

The kinetics of crystallization can be followed in a differential scanning calorimeter (DSC), which measures the temperature difference between a standard and a sample. Either isothermal or constant-heating-rate experiments are possible. The activation energy of crystallization and the time exponent can be found from the DSC crystallization curves (Bansal *et al.*, 1983).

4.4 Glass Ceramics

Some of the few really new materials developed recently are uniformly crystallized glass or "glass ceramics." These materials were first developed at the Corning Glass Works, and are made by special heat treatment of particular glass compositions to give a fine, uniform dispersion of crystals in the glass. These partially crystallized glasses can have many superior properties, such as high strength and impact resistance, low bulk and surface electrical conductivity, low dielectric loss, low chemical reactivity, low coefficient of thermal expansion, and a range of optical properties from clear to completely opaque or white resulting from light scattering from the crystals. Important uses of glass ceramics are radomes, cookware, insulators, telescope mirrors, bearings, heat exchangers, reactor control rods, sealing, vacuum tube envelopes, substrates for electronic circuits, and capacitors. Some examples of crystalline phases in glass ceramics are β-spodumene ($Li_2O{\cdot}Al_2O_3{\cdot}4SiO_2$) and cordierite ($2MgO{\cdot}2Al_2O_3{\cdot}5SiO_2$).

To make a glass ceramic a nucleating agent is usually added to the glass composition to insure uniform crystallization. Typical nucleation agents are titanium dioxide, phosphorous pentoxide, zirconium dioxide, fluorides, and noble and platinum metals.

5. VISCOSITY AND THE GLASS TRANSITION

5.1 Viscosity

The viscosity of a glass is one of its most important technological properties. It determines the melting conditions, the temperatures of working and annealing, fining behavior (removal of bubbles from the melt), upper temperatures of use, and devitrification rate. The viscosities of different glasses vary enormously with composition and are strong functions of temperature. The most common viscosity unit is the poise (P), which is a dyne-second per square centimeter or a gram per centimeter per second. In most literature viscosity is reported in poises. The SI unit of viscosity is the pascal-second or Pa s; one of these units is equal to 10 P. The viscosity of most common fluids such as water and organic liquids is about 0.01 P at room temperature.

The working point of a glass is defined as the temperature at which it has a viscosity of 10^4 P. At this temperature the glass can be readily formed or sealed. The softening point of a glass is the temperature at which it has a viscosity of $10^{7.6}$ P. At this viscosity a rod about 24 cm long and 0.7 mm in diameter elongates 1 mm/min under its own weight. The annealing point is the temperature at which the viscosity is $10^{13.4}$ P, and the strain point, the temperature at which the viscosity of $10^{14.6}$ P. The thermal expansion curve of an annealed glass begins to deviate considerably from linearity at a viscosity of about 10^{14} or 10^{15} P. At the annealing temperature a rapidly cooled silicate glass becomes reasonably strain-free in about 15 min.

The viscosities of several commercially important silicate glasses are compared in Fig. 1, as functions of temperature. The working, softening, and annealing points are marked on the plot.

As with many transport properties the viscosity η fits an Arrhenius-type equation over some ranges of temperature:

$$\eta=\eta_0 \exp[(Q)/(RT)], \qquad (1)$$

where Q and η_0 are temperature-induced coefficients called the activation energy and the preexponential factor, respectively.

The activation energies for viscous flow of various simple glass-forming oxides are given in Table 3. Vitreous germania shows a constant activation energy over a factor of 10^{10} change in viscosity. However, vitreous silica definitely shows a change in activation energy with temperature, and B_2O_3 shows a large change in activation energy with temperature. Also, multicomponent silicate glasses for which viscosity data are plotted in Fig. 1, and most organic glasses, have changing activation energies for viscous flow as functions of temperature. From calorimetric measurements of the glass transition temperature of P_2O_5, Martin and Angell (1986) concluded that the activation energy for its viscous flow, at least from 317 to 636 °C, was constant.

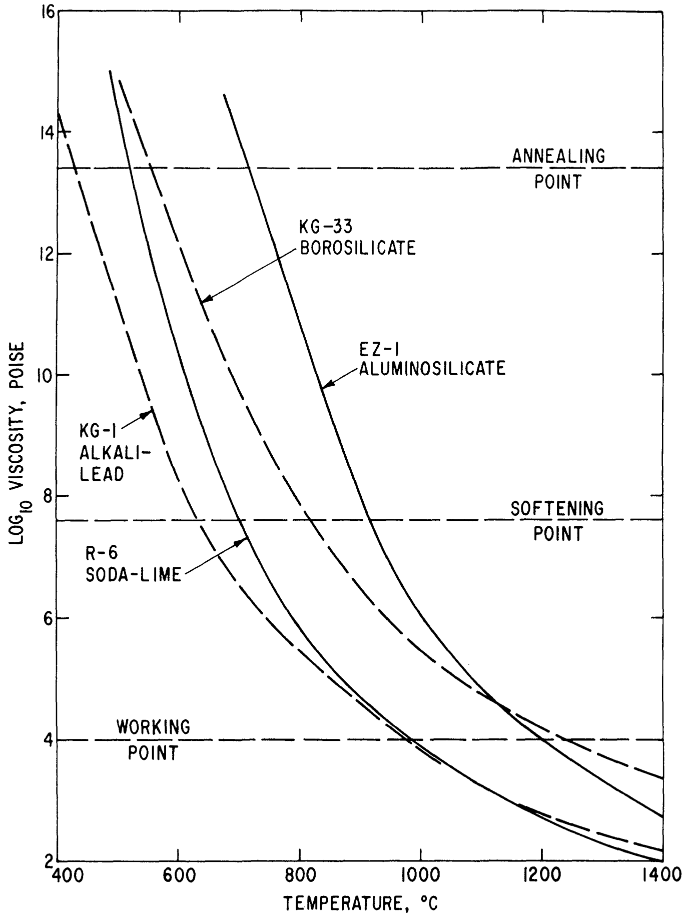

FIG. 1. Viscosities and special temperatures for some commercial silicate glasses.

Table 3. Activation energies for viscous flow of simple oxide glasses.

Glass	Activation energy (kJ/mole)	Temperature range (°C)
SiO_2	710	1100–1400
	515	1600–2500
GeO_2	310	540–1500
P_2O_5	174	545–655
B_2O_3	347–50	26–1300

The strengths of oxygen–central-atom bonds in the simple oxides such as SiO_2, GeO_2, P_2O_5, and B_2O_3 are about 100 kcal/mole. The very great difference in viscosity of these oxides is, therefore, quite surprising, since the mechanism of flow undoubtedly involves breaking of oxide bonds. The melting points of the crystalline oxides parallel the variations in viscosity: SiO_2 (cristobalite), 1710 °C; GeO_2 (hexagonal), 1116 °C; P_2O_5, 580 °C; B_2O_3, 450 °C. The differences indicate that the detailed structure of the oxide network is also important in determining the ease of viscous flow and bond breaking.

The viscosities of the simple oxides decrease sharply with the addition of most impurity ions. Such impurities as water, in the form of —OH groups, and alkali oxides are particularly effective in lowering the viscosities. Zinc and boric oxides among the multivalent oxides are particularly effective in lowering the viscosity of silicates.

The activation energy for viscous flow in multicomponent silicate glasses decreases as the temperature is increased. At low temperatures where the viscosity is about 10^{12} P the activation energy for flow of sodium silicate glasses is about 100 kcal/mole, and it decreases to 50 kcal/mol or less at high temperatures where the viscosity is about 100 P.

The viscosities of a number of non-oxide glasses drop very sharply at temperatures above the glass transition temperature; examples are zirconium fluorides and nitrates, and many molecular organic glasses, such as butyl benzene. These materials have bonds that are easily exchanged between the first and second coordination shell, as compared with the network glasses such as silicates, perhaps explaining this rapid decrease of viscosity with increasing temperature.

The empirical equation

$$\eta=\eta_0\exp[B/(T-T_0)], \qquad (2)$$

where η_0 and B are temperature independent, has been found to fit viscosity data for glasses over certain temperature ranges. This equation is the basis for two theories of viscosity: one based on "free volume" and the other on entropy. However, Eq. (2) is inconsistent with experimental data on glass-forming liquids at high viscosities. In a variety of different glasses the activation energy for viscous flow becomes temperature independent at high viscosities just where Eq. (2) would predict it should vary most sharply. In at least two glasses, P_2O_5 and GeO_2, T_0 is zero, since Eq. (2) is valid over the whole viscosity range in which measurements were made.

Two more recent theories emphasize the importance of defects in the melt structure as they affect viscosity. Mott (1987) suggested that the concentration of broken Si—O bonds in vitreous silica increases as the temperature increases, and provides for easier flow. Brawer (1985) has proposed a change in coordination number of a glass constituent as the defect likely to result in flow, or at least in structural relaxation.

5.2 Glass Transition

As a glass-forming liquid is cooled, certain of its properties such as specific volume, enthalpy, heat capacity, and electrical conductivity change abruptly in a narrow temperature range. This temperature is called the glass transition or transformation temperature.

It is significant that the viscosity does not show an abrupt change at the glass transition temperature. This result suggests that the viscosity determines the rate at which properties relax, but may not be related to the magnitude of property changes under different conditions.

The glass transition temperature can be associated with the slowing down of rearrangements in the glass structure. The rate of these molecular rearrangements is closely associated with viscous flow, and indeed the glass transition temperature occurs at a viscosity of about 10^{13} P.

5.3 Stress Relaxation

When a glass is cooled rapidly from a temperature above the transition region into this region, it retains some properties of the higher temperature, and these properties "relax" to those characteristic of the lower temperature as a function of time. To study stress relaxation, Kurkjian (1963) rapidly twisted an annealed soda-lime glass silicate rod a certain amount with a known stress. The stress in the rod at this fixed twist and at a constant temperature was then measured as a function of time. The "instantaneous" initial stress S_0 was used to normalize the stresses at different temperatures. When this was done, the stress relaxation at different temperatures in the transition range fitted on the same master curve. Kurkjian fitted his stress S as a function of time by a sum of six exponential terms; the spectrum of relaxation times was the same for the different temperatures of the experiment. An empirical equation that is often used to describe relaxation processes in glass is

$$S=S_0 \exp[(-t/\tau)^b]. \tag{3}$$

In this equation there are only two independent parameters, the relaxation time τ and the exponent b. However, the data of Kurkjian cannot be described by a single value of b.

Two recent books by Scherer (1986) and Brawer (1985) give much detailed information on relaxation processes in glass.

6. STRENGTH OF GLASS: FRACTURE, FATIGUE, AND STRENGTHENING

6.1 Strength Distributions and Surface Treatments

When the failure stress of a number of glass samples is measured under identical conditions of sample preparation and loading, the failure stresses vary over a wide range of a factor of 2 or more. The strengths are often distributed symmetrically, and so they can be described by a normal (Gaussian) distribution,

$$P=(1/d\sqrt{2\pi})\exp[-(S-S_m)^2/2d^2]. \tag{4}$$

P is the probability of finding a sample of strength (failure stress) S, S_m is the strength of greatest probability, and d is a measure of the spread of the distribution, called the standard deviation, and is equal to the root mean square of deviations from the mean strength. The integral of Eq. (4) gives the fraction F of samples that break below the stress S. Another distribution function called the Weibull distribution has become quite popular in describing strength distributions:

$$F=1-\exp[-(S/S_0)^m], \tag{5}$$

where S_0 is a scaling factor and m is a measure of the spread of the distribution—the smaller m, the broader the distribution. For m larger than about 3 the Weibull distribution is nearly symmetrical and closely resembles the normal distribution.

When a glass surface is abraded or bombarded with particles, the strength of the glass is reduced. Alternatively, if glass samples are carefully prepared and kept "pristine," they have very high strengths. Many results up to 10^{10} Pa have been reported, and Hillig (1962) measured strengths of up to 1.5×10^{10} Pa at -196 °C for fused-silica rods 1 mm in diameter. These samples were flame-drawn from larger rods and carefully protected from contact with solids. It is spectacular to bend a 1-mm rod of glass into a closed loop. If one so much as touches the fiber lightly with a finger tip, it breaks at much lower stress.

The strength of ordinary or abraded glass can be substantially increased by etching with hydrofluoric acid. Etching leads to some samples with very high strengths, because flaws in their surfaces have been removed. However, samples etched in HF are highly sensitive to damage by handling, so that many etched samples have strengths lower than before etching because new and larger flaws have been introduced into their surfaces.

6.2 Observations of Fracture Surfaces

A fracture surface on glass shows several characteristic regions—an origin usually on the glass surface, as shown in Fig. 2, a smooth surface called the mirror, a misty region outside the mirror, and finally a region of coarse surface roughness, called hackle. The distance R from the fracture origin to the mirror-mist boundary (or the distance to the mist-hackle boundary) is related to the fracture strength S by the empirical equation

$$S\sqrt{R}=A, \tag{6}$$

where A is a constant for a particular brittle material and temperature. Values of about 2×10^6 N/m$^{3/2}$ were found for A for glass at

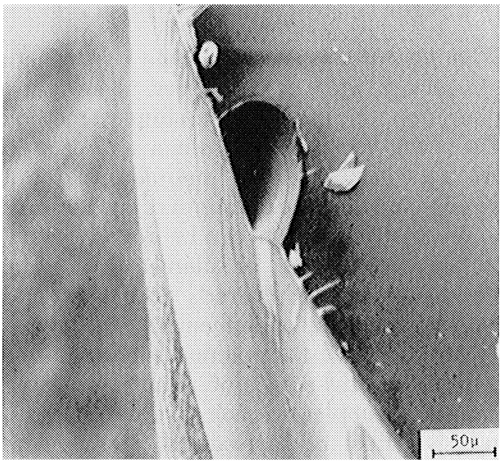

FIG. 2. Micrograph of a fracture origin in a soda-lime silicate glass. Region on the left is the glass surface; region on the right is the fracture surface.

room temperature. Apparently, A is not significantly dependent on temperature or glass composition.

Equation (6) is of considerable practical value because it allows one to make an accurate estimate of fracture stress without having to measure it directly. Thus for pieces with complex shapes where a stress analysis is difficult, or where load measurements are not possible, Eq. (6) is invaluable.

If a hard sphere is pressed against a glass surface, a crack around the contact region first develops and then grows into the glass in a conical shape. These cracks are called Hertzian cracks because the elastic stress distribution for contact of a hard sphere on a plate was calculated by Hertz.

If the indenter is sharp rather than spherical, the crack pattern is entirely different. A "median vent" crack propagates into the glass directly under the indenter during loading. When the load is released, winglike cracks called "lateral vents" propagate radially from the contact region. A region of the glass directly under the indenter is deformed in a particular way that is discussed in more detail in the section on plastic deformation in glass.

The impact of a hard spherical projectile gives damage with features of both spherical and sharp indenters. At low impact velocities Hertzian cone cracks are formed, but at higher velocities the cone cracks turn up at their ends and finally resemble the radial winglike cracks; at high velocities median cracks also propagate down into the glass.

When a hard object such as a diamond indenter is dragged across a glass surface, median and lateral cracks form along its track much as under an indenter. Abrasion of a hard surface against the glass also gives this kind of linear damage.

6.3 Fracture Criteria

The experimental results described above provide strong evidence that the surface condition of a glass determines its strength, and that flaws in the surface reduce the strength. These results therefore led to a model of surface cracks as the origin of brittle fracture. It is difficult to observe these cracks directly, and there still is uncertainty about their size, depth, and shape. Inglis (1913) solved the elastic problem of an elliptical crack subjected to a tensile strength S. He showed that the stress σ along the inside crack surface right at the tip is

$$\sigma/S = 1 + 2(c/\rho)^{1/2} \approx 2(c/\rho)^{1/2}, \tag{7}$$

in which c is the depth of the crack (semimajor axis of the ellipse) and ρ is the radius of curvature of the crack tip. If the crack is not elliptical in shape, Eq. (7) is still the proper form. Thus, if the crack is deep and has a sharp tip, the stress at the tip can be much higher than the applied stress. Various types of surface damage on glass give rise to cracks with different depths and tip radii, resulting in the influence of surface treatment on strength found experimentally. The stress drops off very sharply away from the tip along the crack surface. The stress also drops off sharply into the glass away from the crack tip. Thus, the stress is near the maximum value given by Eq. (7) in only a very small volume of the glass.

Measured strengths of glass samples are usually in the range of about 50–200 MPa, leading to design strengths less than 30 MPa. Strengths of up to 1.5×10^{10} Pa have been measured for vitreous silica at -196 °C, and theoretical strengths for silica calculated from various models are about 2.4×10^{10} Pa. The ultimate strength for soda-lime glass is about 1.6×10^{10} Pa, reflecting the lower Si—O bond density in this glass.

Another fracture criterion is based on the strain energy change in a stressed material caused by a crack. Griffith (1920) derived the equation

$$S_f^2 = 2E\gamma/\pi c \qquad (8)$$

for the fracture stress S_f in a material with a crack depth c, fracture surface energy γ, and Young's modulus E. This criterion is different from that of Inglis. In a treatment of crack energies and fracture criteria in terms of thermodynamic laws, it is shown that the Griffith equation provides a necessary but not sufficient condition for fracture, and that the Inglis equations provide the correct criterion for fracture (Doremus, 1976). The second law of thermodynamics shows that the fracture stress cannot be less than that given by Eq. (8), but can be greater.

6.4 Faulting

In contrast to tensile and bend tests of glasses, indentation of glass with a sharp point leads to some regions of deformation under the indenter that do not recover when the load is removed. This deformation in fused silica is apparently a "densification" that recovers completely on heating. In glasses with modifiers (alkali and alkaline earth oxides) this deformation can be partly a densification, but only a small part of the deformation recovers upon heating. More detailed examination of this deformed region under a sharp indenter shows that is made up of flow or fault lines between which there are only elastic strains (Hagan, 1979, 1980). This shear deformation under a concentrated compressive load is distinguished from other types of deformation and is called faulting, in analogy to geological faulting. The fault lines resemble geological faults and are quite different from slip in plastic deformation of crystals. Therefore, it seems important to distinguish this faulting deformation from plastic deformation in ductile crystals.

Viscous flow is much too slow in glass to cause appreciable deformation at ambient temperatures. The viscosity of commercial soda-lime glass at 490 °C is about 10^{15} P, with an effective activation energy of greater than 400 kJ/mole; the activation energy increases with decreasing temperature. Thus, the viscosity at room temperature is greater than 10^{59} P.

6.5 Fatigue

Fatigue in glass results when its strength deteriorates while it is stressed in an atmosphere containing water. The simplest method of studying fatigue in glass is a static test in which a constant load is imposed on a set of samples and the time of failure of each sample is recorded. Such tests can be done in bending, simple tension, or diametral loading; bending is easiest for rods and tubes and requires simple test equipment. Although static tests are the closest to practical situations they have the disadvantages of requiring a large number of samples for each of several stresses, and also must be carried out over a large spread of times, leading to long delays in obtaining results.

Fatigue can also be measured with samples tested at different loading rates; the dependence of fracture strength on loading rate is related to the stress-failure time relation in static loading. This method requires fewer samples and less time, but a testing machine with a linear load application is needed.

Failure times of Pyrex borosilicate glass rods in bending are shown in Fig. 3. Each point represents the mean of many measurements, usually forty or more. The results are plotted as reduced stress S/S_N against logarithm of fracture time, where S is the maximum stress in four-point bending and S_N is the mean fracture stress in the same test at -196 °C, where there is no fatigue. Humidity strongly influences the failure time. Rods of soda-lime glass all treated in the same way and all having a mean S_N of 86 MPa had mean failure times of 8.7, 200, and 3470 s in water, 43% r.h. (relative humidity), and 5% r.h. respectively, all at a stress $S/S_N=0.5$. Log$t_{1/2}$ is about linear with relative humidity over this range of humidities.

Surface treatment and sample history strongly influence the failure times of glass tested at the same relative humidity and relative stress S/S_N. Values of $t_{1/2}$, the time of failure at $S/S_N=0.5$, for soda-lime glass after different treatments range over five orders of magnitude, and there is no simple connection with inert strength S_N.

Experimental results on strength and fatigue of very strong silica (pure SiO_2) seem to be influenced by the chemical treatment of the silica surface. Just touching lightly the surface of a pristine silica rod greatly lowers it strength.

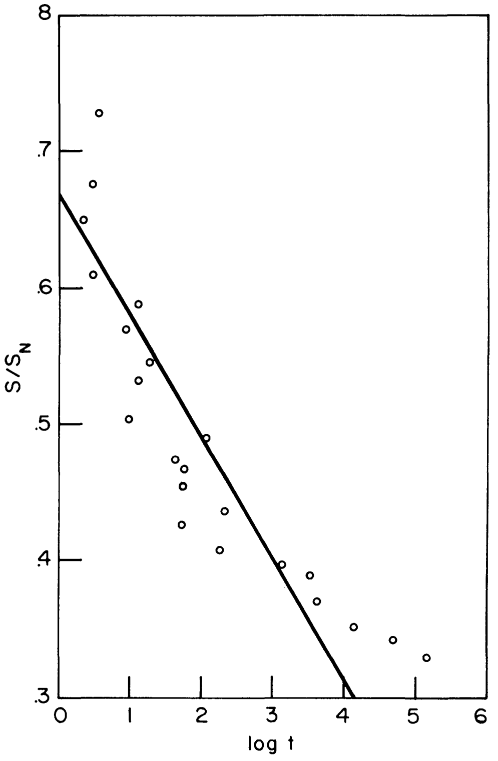

FIG. 3. Mean log(failure time) of Pyrex borosilicate glass as a function of normalized applied stress.

Silica fibers for use as waveguides are usually coated with an organic polymer to reduce their susceptibility to fatigue. These coatings reduce fatigue at short times, but at longer times, at higher temperatures, and in alkaline media the coatings are attacked, and the fatigue becomes more like that of uncoated silica.

The failure times of glass samples that are tested under identical conditions and that have identical treatments still show a large spread. This spread is related to the distribution in fracture strengths discussed above. These spreads can also be treated in statistical terms, and usually fit normal or Weibull statistical distributions.

6.6 Strengthening

Removal of surface flaws in glass by fire polishing or chemical etching leads to very strong material, but the surface can be easily damaged again, so that these methods are not practical unless the surface is immediately protected by a coating, as it is in drawing fiber glass. If the glass surface is treated to increase compressive stress in it, a higher tensile stress is required for crack propagation. Surface compressive stress can be introduced by rapid quenching, by ion exchange, and by surface crystallization. During rapid quenching, first the surface cools to a rigid state, and then the interior cools and tries to shrink, but is restrained by the already rigid surface layer. In ion exchange a large ion such as potassium is exchanged from a fused salt for sodium ions in the glass; the profile and level of stress can be controlled more easily than for quenching.

7. ELECTRICAL PROPERTIES AND IONIC TRANSPORT

7.1 Ionic Conductivity

The electrical conductivities of most oxide glasses result from the transport of monovalent cations such as sodium in them. Lithium ions are also quite mobile in glasses. Potassium and hydrogen ions sometimes carry current, but their mobilities are usually much less than those of sodium and lithium. Even in oxide glasses with no nominal addition of alkali, the conductivity results from transport of monovalent cations. In vitreous silica, electrolysis experiments show that sodium and lithium ions are the conducting species, even though they are present at concentrations of a few parts per million.

If the electrical current in a glass is carried by a single ionic species of concentration c and diffusion D, the electrical conductivity σ is given by the Einstein equation:

$$\sigma = Z^2F^2Dc/RT, \tag{9}$$

in which Z is the ionic valence, F is the Faraday, R the gas constant, and T the absolute temperature. Equation (9) agrees well with conductivity and diffusion coefficient measurements in vitreous silica, Pyrex borosilicate, and a commercial soda-lime silicate glass (see Doremus, 1973, for a review). However, many other experimental measurements have found that the conductivity must be multiplied by a factor f less than 1, often from about 0.3 to 0.7. Different workers have found different f values for the same glass. The reasons for these f values less than one are uncertain.

Below the glass transition temperature the conductivity fits an Arrhenius equation:

$$\sigma=\sigma_0 \exp(-Q/RT) \qquad (10)$$

with activation energy Q and preexponential σ_0 independent of temperature. Typical Q values for sodium and lithium transport are from 60 to 100 kJ/mole. At temperatures above the glass transition the conductivity no longer follows Eq. (10), as shown in Fig. 4. At T_g there is a discontinuous increase in activation energy, and then a continuous decrease as the temperature is increased further.

Stress or pressure influences electrical conductivity of glasses. Quenched glass has a lower density and higher conductivity than annealed glass, and compaction by hydrostatic pressure leads to lower conductivity.

At a particular temperature the electrical conductivity and alkali-ion mobility in a silicate glass increase with increasing alkali concentration. This increase corresponds with the decrease in glass viscosity as alkali is added, caused by breaking up of the silicon-oxygen network by the alkali oxide.

The addition of an oxide of a higher-valent metal ion to an alkali silicate glass tends to decrease the mobility of the alkali silicate glass and thus the electrical conductivity. For the alkaline-earth ions the decrease is related to ion size; the larger the ion, the greater its blocking action. Addition of aluminum oxide to the glass increases ionic mobility and conduction and decreases the activation energy for ionic transport.

When a second alkali oxide is added to an alkali silicate glass, the conductivity decreases sharply. This "mixed-alkali" effect is caused by a decrease in the mobility of the most mobile ion in the glass.

Anions and cations of valence greater than 1 diffuse much more slowly than monovalent cations in silicate glasses.

In zirconium fluoride glasses the conductivity is ionic, and the fluoride ion is the most mobile.

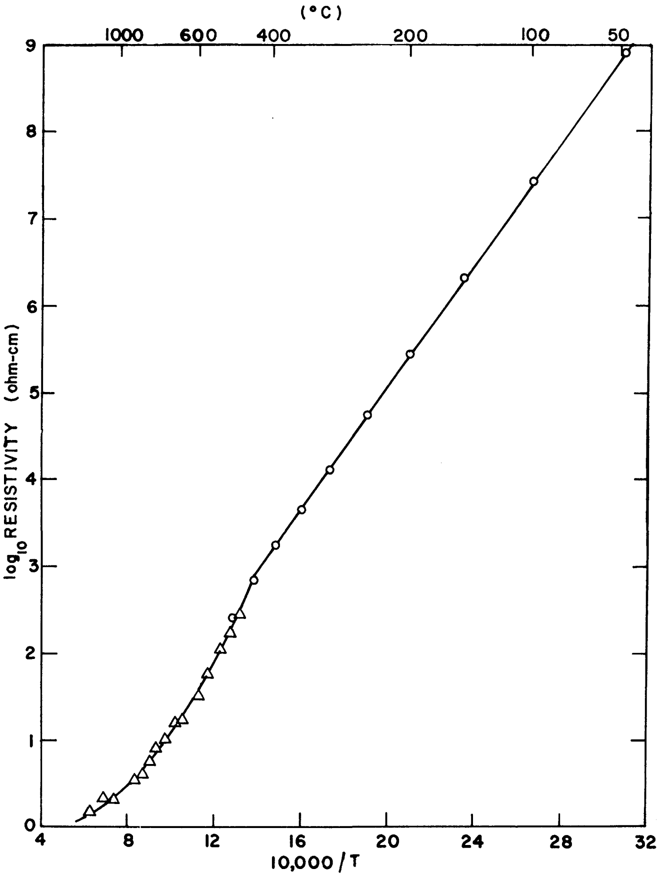

FIG. 4. Log (resistivity) of a 26.5 wt% Na_2O, 73.5% SiO_2 glass as a function of reciprocal temperature. From Doremus (1973), p. 155. T_g is 475 °C. Symbols show two different sets of measurements.

7.2 Electronic Conductivity

Certain oxide glasses containing multivalent transition metal ions are electronic conductors. Examples are barium vanadate ($BaO{\cdot}2V_2O_5$), 0.6 $V_2O_5{\cdot}0.4\ Te_2O$, and various vanadium phosphates. Also vanadium, iron, cobalt, and manganese oxides in silicate, borate, or phosphate glasses show electronic conductivity.

Chalcogenide glasses containing the Group VI elements sulfur, selenium, and tellurium alone or in combination with Group V elements phosphorus, arsenic, antimony, and bismuth, as well as other elements such as thallium and germanium, are electronic conductors. These glasses have Arrhenius temperature dependence of their conductivities, as do intrinsic semiconductors, but their conductivity is insensitive to impurities. These glasses also show a variety of switching and memory effects, as summarized in Mott and Davies (1971).

7.3 Dielectric Properties

One kind of dielectric loss in glass is closely related to electrical conductivity in the glass. The peak frequency has the same activation energy as the electrical conductivity. This loss is present in both ionic and electronic conducting glasses, and is similar to peaks found in a wide variety of dielectric materials. There are also special low-temperature loss peaks in oxide glass.

8. CHEMICAL PROPERTIES

8.1 Surfaces

Many properties of glass, such as strength, fatigue, chemical durability, and reflectivity, are strongly influenced by surface processes. The metal-oxide groups at the surface of an oxide glass are invariably hydrated, for example, to form SiOH groups. These groups can be isolated or linked with hydrogen bonds; water molecules can also absorb physically on them. Other gases such as halogens, organics, nitrogen, and even inert gases can absorb physically on glass surfaces (Dunken, 1985).

Chemical reactions with the hydroxyl groups on oxide glass surfaces can replace the hydrogen atom with a halogen or organic group. Si—O bonds in the surface can be split by reactions with water or ammonia. Silanes react with silicate surfaces to form adsorbed organic molecules, and can condense to siloxane polymers.

8.2 Diffusion and Reactions of Gases in Glasses

Gases can dissolve and diffuse molecularly in many oxide glasses because of their open structure. Table 4 gives solubilities and diffusion properties of a large range of molecular sizes and chemical types in vitreous silica. The molecular solubility is almost constant with temperature, and is reduced somewhat as the molecule becomes larger. These properties are consistent with a model of lodging of gas molecules in interstices in the glass. Diffusion is strongly influenced by molecular size. The square root of the activation energy of diffu-

Table 4. Solubility and diffusion coefficients of gases in amorphous silica.

Gas	Diameter (nm)	Solubility	Diffusion coefficient @ 1000 °C (cm^2/s)	Activation energy for diffusion (kJ/mole)
Helium	2.0	0.025	5.5×10^{-5}	20
Neon	2.4	0.020	2.5×10^{-6}	37
Hydrogen	2.5	~0.03	7.3×10^{-6}	36
Water	~3.0		$3\ \times10^{-7}$	71
Argon	3.2	0.01	1.4×10^{-9}	113
Oxygen	3.2	0.01	6.6×10^{-9}	108
Nitrogen	3.4			109
Krypton	4.2			~190
Xenon	4.9			~300

sion in vitreous silica is directly proportional to molecular size for a wide variety of gas molecules. This dependence is found in a model of jumping of molecules from one interstice to another in the glass through connecting "doorways."

Some gases can react with the silicon-oxygen network when they are dissolved in it. Examples are water, hydrogen, oxygen, and nitrogen, roughly in order of their reactivity with the Si—O bonds. Carbon dioxide and sulfur dioxide can react with constituents of the glass such as alkalis, forming, for example, sodium sulfate or sodium carbonate in the glass. Reactions of water are discussed in the next subsection. Hydrogen reacts with the silicon-oxygen network to form SiOH and SiH groups.

The concentration of dissolved molecular oxygen in a glass determines its state of oxidation. After holding for hours at glass melting temperatures the glass equilibrates with the surrounding atmosphere, and has a uniform concentration of dissolved oxygen. When the glass is cooled, the state of oxidation is determined by this oxygen concentration developed during glass melting, because the diffusion of molecular oxygen in the glass is too slow to reach equilibrium. A variety of multivalent ions change their oxidation states in response to a changing oxygen concentration in glass. A list of these ions in their oxidized state is Cr^{+6}, Mn^{+3} , Ce^{+4}, As^{+5}, Sb^{+5}, Fe^{+3}, Sn^{+4}. This list is in the order of difficulty of oxidation, chromium being the easiest to reduce and tin the easiest to oxidize. Often the ratio of amounts of ions in different oxidation states can be measured from optical absorption bands characteristic of these states, giving different colors to the glass, as described in the section on optical properties.

8.3 Reaction of Glass with Water

The surfaces of all oxides are hydrated to groups such as SiOH by reaction with water, as mentioned above. Water reacts readily with silicon-oxygen-silicon bonds to form two SiOH groups. Thus measurements of dissolved molecular water, for example by infrared absorption, give results much greater than the amount of dissolved molecular water because of this reaction.

Liquid water and water vapor exchange and interdiffuse with alkali ions in the water:

$$Na^+(\text{glass}) + H_2O = H_3O^+(\text{glass}) + NaOH. \tag{11}$$

In water vapor the sodium hydroxide deposits on the glass surface and reacts with atmospheric carbon dioxide to form sodium carbonate crystals (Husung and Doremus, 1990). The evidence for hydronium ion exchange comes from hydrogen profiles from nuclear resonant reactions on glasses rapidly chilled from the reaction temperature to prevent outgassing of water (Schnatter *et al.*, 1988). The sodium and hydronium ions interdiffuse in the glass, leading to depletion of sodium in the outer layers of the glass and its deposition on the surface. The interdiffusion can be described by the Nernst–Planck equations.

In liquid water or aqueous solutions oxide glasses also dissolve congruently—that is, the glass constituents dissolve in the solution in the same proportion as they exist in the glass. The rate of dissolution depends strongly on glass composition. A sequence of increasing dissolution rates at *p*H 7 (slowest dissolving first) is obsidian glass, crystalline quartz, and aluminosilicate minerals < vitreous silica, soda-lime glass containing alumina < soda-lime glass without alumina ≪ binary alkali silicates (Perera *et al.*, 1991). The rate of dissolution of silicate glasses increases in alkaline solution, corresponding to the increase in silica solubility as H_4SiO_4 (silicic acid) ionizes more at higher *p*H.

9. OPTICAL PROPERTIES

9.1 Vitreous Silica

Amorphous silica is highly useful as an optical material and for fiber-optic waveguides because of its high transmission of ultraviolet, visible, and near-infrared light. The transmission edge in the ultraviolet depends on the purity and oxidation state (stoichiometry) of the silica. Transition metals absorb intensely in the ultraviolet, and so they must be absent for high transmission. In silicas melted under neutral or reducing conditions, an absorption band at 0.24 μm has been associated with reduced impurities (aluminum or germanium) or an intrinsic center of a trapped electron or hole near a vacancy.

An absorption band at 11.5 eV in the ultraviolet is apparently intrinsic to the silicon-oxygen network, and therefore may be characteristic of the band gap of the vitreous silica.

In the infrared, silica has an intense absorption band at about 1100 cm^{-1} present in all silicates and assigned to the stretching vibration of the Si—O—Si bond. A band at wave number 465 cm^{-1} has been assigned to the "rocking" of the Si—O—Si bond. Lower-intensity bands at intermediate wave numbers involve various mixed vibrational modes.

9.2 Other Network-Forming Oxides

The appropriate ultraviolet absorption edges of simple oxide glasses are SiO_2, 0.16 μm; B_2O_3, 0.17 μm; P_2O_3, 0.27 μm; GeO_2, 0.22 μm.

The infrared spectrum of vitreous germania is similar to that of silica but shifted to lower frequencies, with an intense absorption band at about 880 cm^{-1} resulting from Ge—O—Ge stretching vibrations.

Boron oxide, B_2O_3, has strong infrared absorption bands at 1260 and 718 cm^{-1}, with a broad band at lower frequencies and a shoulder on the 1260-cm^{-1} band at about 1370 cm^{-1}. The Raman spectrum shows an intense band at 808 cm^{-1} and a strong band at 1260 cm^{-1}. These bands can be interpreted as resulting from a boroxyl network structure of B_2O_3. In this structure planar boroxyl (six-member B—O) rings are linked by other B—O bonds. The band at 1260 cm^{-1} is assigned to the stretching vibration of the B—O bonds linking the rings. The 808-cm^{-1} Raman band results from a "breathing" motion of the oxygen atoms in the boroxyl ring that is inactive in the infrared.

9.3 Alkali Silica Glasses

As alkali oxide is added to silica the infrared absorption band at 1100 cm^{-1} decreases and new bands at 1075 and 940 cm^{-1} develop. The band at 940 cm^{-1} has been assigned to the Si–O stretching vibration when the oxygen is not bonded to another silicon (nonbridging oxygen), and the 1075 cm^{-1} band is the 1100 cm^{-1} Si–O stretching band shifted by the presence of SiO Na^+ groups.

The optical density in the ultraviolet at wavelengths longer than the absorption edge at 0.16 μm for vitreous silica is directly proportional to the amount of alkali in the glass. This absorption is interpreted as resulting from the nonbridging oxygen ions introduced by the alkali additions.

9.4 Transition-Metal Ions and Color

Transition-metal ions color glass a variety of hues, as shown in Table 5. Thus, the green color of thick sections of most commercial glasses is caused by ferric iron ions. Other familiar colors are chromium for green, manganese for purple, and copper and cobalt for blue. The visible absorption bands listed in Table 5 result from electronic transitions of the 3*d* electrons. When the transition-metal ions are coordinated with other ions such as oxygen in the glasses, the energy levels of the *d* electrons are split by the electric field or "ligand field" of the coordinating ions, instead of being degenerate as in the free ion. The *d*-electron orbitals are strongly directional, and so the splittings are sensitive to the arrangements of the surrounding atoms.

9.5 Applications

Glass has long been an important optical material because of its transparency, stability, and chemical durability. The refractive index of glass can be varied over a wide range. A recent application of glass as a fiber-optic waveguide for communications is replacing metal cables for telephone lines. The development of very pure, strong vitreous silica fibers for this application by Bell Labs and Corning Glass was a *tour de force* of technology. The fibers, as small as 10 μm in diameter, have a different surface composition with a higher refractive index to improve the wave guiding

Table 5. Colors and absorption bands for transition-metal ions in silicate glass.

Ion	No. of *D* electrons	Absorption maxima (μm)	Color of glass
Cr^{3+}	1	0.66, 0.45	Green
Ti^{3+}	1	0.57	Purple
V^{4+}	1	1.12	Red
V^{3+}	2	0.645, 0.425	Green
Mn^{3+}	4	0.50	Purple
Mn^{2+}	5	0.435	Brown
Fe^{3+}	5	0.41	Green
Fe^{2+}	6	1.10	Blue
Co^{2+}	7	0.6	Blue
Ni^{2+}	8	0.68–0.49	Purple, grey
Cu^{2+}	9	0.79	Blue

properties of the fiber. The fibers are drawn in very long lengths (up to a kilometer) from a preform of larger diameter and controlled composition and impurity level. Intrinsic scattering and absorption are a minimum at a wavelength of about 1 μm, at which the signals are transmitted.

Zirconium fluoride and some chalcogenide glasses are much more transparent in the infrared than silicate glasses; silicates have an absorption edge at about 4 μm, whereas the fluorides are transparent to about 8 μm. This property makes the zirconium fluoride glasses attractive candidates for fiber-optic waveguides, because they have lower minimum absorption. This potential is difficult to realize because the glasses must be highly pure and contain no scattering centers such as crystallites or pores.

Glass can be used in many applications related to lasers: as a host for laser ions that lead to active devices such as oscillators, amplifiers, and Faraday rotators, and as an optical material for fibers, lenses, windows, and thin-film waveguides. Glass can dissolve a wide variety of laser ions, and is useful for high energy storage and high peak power. The spectral distribution of laser light from a glass is usually broader than from a crystal, because of the disordered glass structure. The low thermal conductivity of glass limits its use to pulsed operation with average power.

9.6 Nonlinear Optical Properties

Glasses with certain optically absorbing compounds show nonlinear optical properties at high light intensities. An example is glasses containing cadmium sulfide or selenide; apparently small (~few nm) crystals of these compounds impart nonlinear properties to the glass.

GLOSSARY

Annealing: Removal of stress in a glass by heating it at a temperature just above the glass transition temperature.

Chalcogenide: Compound of sulfur, selenium, or tellurium.

Devitrification: Nonuniform crystallization of a glass.

Faulting: Shear deformation in which planes of broken bonds form along particular directions.

Fiber-Optic Waveguide: Fine fiber of glass for conducting optical signals.

Fining: Removal of bubbles in a glass melt.

Float Glass: Flat glass made by floating a sheet of molten glass onto a bath of molten metal, usually tin.

Glass Ceramics: Uniformly crystallized glasses; can be partially glassy; crystal grains usually small (microns).

Glass Transition Temperature: Temperature at which glass properties change discontinuously. Can be taken as the temperature at which the viscosity is 10^{12} Pa s.

Hertzian Cracks: Cone cracks made by pressing a spherical indenter against a brittle solid.

Liquidus: The temperature above which a mixture consists of a single liquid phase. In a phase diagram, the temperature-composition line above which temperature there is a single phase, and below which crystallization starts.

Nonbridging Oxygen: An oxygen atom bonded to only one silicon atom.

Random Network: The amorphous structure of atoms bonded into a network that shows no order beyond a few atomic dimensions.

Soda-Lime Glass: A glass made from silica sand, sodium oxide, and calcium oxide, with other oxides possible.

Sol-Gel Fabrication: Making a glass from organic compounds that form a gel from a liquid solution.

Spinodal Decomposition: A phase transition in which a single multicomponent phase forms two interconnected phases. The spatial extension of the phases is constant with time, but their compositions change.

Tank: A large container for melting glass continuously.

List of Works Cited

Bansal, N. P., Doremus, R. H., Bruce, A. J., Moynihan, C. T. (1983), "Kinetics of crystallization of ZrF-BaF-LaF glass by differential scanning calorimetry," *J. Am. Ceram. Soc.* **66**, 233-237.

Bernal, J. D. (1960), "Geometry of the structure of monatomic liquids," *Nature* **185**, 68–70.

Brawer, S. (1985), *Relaxation in Viscous Liquidus and Glasses*, Columbus, OH: American Ceramics Society.

Cahn, J. W., Charles, R. W. (1965), "Spinodal decomposition in glasses," *Phys. Chem. Glasses* **6**, 181–190.

Cargil, G. S., Spaepen, F. (1981), "Description of chemical ordering in amorphous metal alloys," *J. Noncryst. Solids* **43**, 91–97.

Cranmer, D., Saloman, R., Yinnon, H., Uhlmann, D. R. (1981), "Nucleation rates of crystals in anorthite glass," *J. Noncryst Solids* **45**, 127–140.

Doremus, R. H. (1973), *Glass Science*, New York: Wiley and Sons.

Doremus, R. H. (1976), "Cracks and energy," *J. Appl. Phys.* **47**, 1833–1836.

Doremus, R. H. (1985), *Rates of Phase Transformations*, San Diego: Academic.

Dunken, H. H. (1985), "Surfaces," in: M. Tomezawa, R. H. Doremus (Eds.), *Treatise on Materials Science and Technolgy*, Vol. 22, San Diego: Academic.

Finney, J. L. (1970), "Random packings and the structure of simple liquids: I. The geometry of close packing," *Proc. R. Soc. London* **A319**, 479–93.

Gaskell, P. H., Eckersley, M. C., Barnes, A. C., Chieux, P. (1991), "Intermediate structure of calcium silicate, glass with neutron diffraction," *Nature* **350**, 675–677.

Griffith, A. A. (1920), "The Phenomena of Rupture and Flow in Solids", *Phil. Trans. R. Soc.* **221A**, 163-198.

Hagan, J. T. (1979), "Shear faulting under an indenter in glass," *J. Mater. Sci.* **14**, 462–475.

Hagan, J. T. (1980), *J. Mater. Sci.* **15**, 1417–1424.

Hammel, J. J. (1967), "Nucleation rate of phase separation in a soda-lime silicate glass," *J. Chem. Phys.* **46**, 2234–2244.

Hillig, W. B. (1962), "Sources of Weakness and the Ultimate Strength of Brittle Amorphous Solids", in: J. D. Mackenzie (Ed.), *Modern Aspects of the Vitreous State*, Vol. 2, London: Butterworths, pp. 152-194.

Husung, R. D., Doremus, R. H. (1990), "Infrared study of reaction of water with glass," *J. Mater. Res.* **5**, 2209–2217.

Inglis, C.E. (1913), "Stresses in a Plate Due to the Presence of Cracks and Sharp Corners", *Trans. Inst. Naval Arch.* **55**, 219-240.

James, P. F. (1985), "Nucleation rates of crystals in lithium disilicate glass," *J. Noncryst. Solids* **73**, 517–530. References to previous work.

Ko, S. H., Doremus, R. H. (1991), "Structure of zirconium fluoride glass," *Phys. Chem. Glasses.* **32**, 196–201.

Kurkjian, C. K. (1963), "Stress relaxation in soda-lime silicate glass," *Phys. Chem. Glasses* **4**, 128–132.

Martin, S. W., Angell, C. A. (1986), "Viscosity of P_2O_5," *J. Phys. Chem.* **90**, 6736–6740.

Mott, N. F. (1987), "Theory of viscosity of network liquids," *Philos. Mag.* **B56**, 257.

Mott, N. F., Davies, E. A. (1971), *Electronic Processes in Noncrystalline Materials*, Oxford; Clarendon.

Mozzi, R. L., Warren, B. E. (1969), "Structure of vitreous silica," *J. Appl. Cryst.* **2**, 164–174.

Perera, H., Doremus, R. H., Lanford, W. A., (1991), "Dissolution rates of silicate glasses in water," *J. Am. Ceram. Soc.* **74**, 1269–1274.

Scherer, G. W. (1986), *Relaxation in Glass and Composites*, New York: Wiley.

Schnatter, K. H., Doremus, R. H., Lanford, W. A. (1988), *J. Noncryst. Solids* **102**, 11–18.

Wright, A. C., Hulme, R. A., Grimley, D. J., Sinclair, R. N., Martin, S. W., Price, D. L., Galeener, F. L. (1991), "Review of research on structure of glass," *J. Noncryst. Solids* **129**, 213–232.

Further Reading

Bansal, N., Doremus, R. H. (1986), *Handbook of Glass Properties*, San Diego: Academic.

Brinker, C. J., Scherer, G. W. (1990), *Sol-Gel Science*, San Diego: Academic.

Doremus, R. H. (1973), *Glass Science*, New York: Wiley and Sons.

Hillig, W. B., Charles, R. J. (1965) in: V. F. Zackey (Ed.), *High Strength Materials*, New York: Wiley, p. 682.

Izumitani, T. S. (1984), *Optical Glass*, Tokyo: Kyoritsu Shuppan.

Mackenzie, J. D. (Ed.) (1960, 1962, 1964), *Modern Aspects of the Vitreous State*, Vols. 1–3, London: Butterworths.

McMillan, P. W. (1979), *Glass Ceramics*, San Diego: Academic.

Tomozawa, M., Doremus, R. H. (Eds.) (1977, 1979, 1982, 1985), *Treatise on Materials Science and Technology*, Vols. 12, 17, 22, 26, San Diego: Academic. Chapters of special interest for the present article:

Doremus, R. H., Vol. 22, "Fracture and Fatigue."

Drexhage, M. G., Vol. 26, "Heavy-Metal Fluoride Glass."

Dunken, H. H., Vol. 22, "Surfaces."

Friebele, Ed. J., Griscom, D. L., Vol. 17, "Radiation Effects."

Sigel, G. H., Vol. 12, "Optical Absorption."

Shelby, J. E., Vol. 17, "Molecular Solubility and Diffusion."

Tomozawa, M., Vol. 17, "Phase Separation."

Uhlmann, D. R., Kreidl, N. J. (Eds.) (1980-) *Glass: Science and Technology*, Vols. 1–5, San Diego: Academic.

Wong, J., Angell, C. A. (1976), *Glass Structure by Spectroscopy*, New York: Dekker.

Zallen, R. (1983), *The Physics of Amorphous Solids*, New York: Wiley.

GRAIN BOUNDARIES

Alexander H. King, *Department of Materials Science & Engineering, State University of New York at Stony Brook, Stony Brook, New York, U.S.A.*

Introduction 271

1. **Grain-Boundary Parameters** 272
   - 1.1 Crystal Misorientation 272
   - 1.2 Grain-Boundary Plane 273
   - 1.3 Macroscopic Degrees of Freedom 274
2. **Grain-Boundary Properties and Behavior** 274
   - 2.1 Energy 274
   - 2.2 Mobility 274
   - 2.3 Diffusion 274
   - 2.4 Strength of Materials 275
   - 2.5 Electronic Properties of Materials 275
3. **Small-Angle Grain-Boundary Structure** 277
   - 3.1 The O-Lattice 278
   - 3.2 Grain-Boundary Energy 279
4. **Large-Angle Grain-Boundary Structure** 279
   - 4.1 Geometrical Characteristics of Special Large-Angle Grain Boundaries 279
   - 4.2 Structure of Coincidence-Related Boundaries 281
   - 4.3 Off-Coincidence Grain Boundaries 282
   - 4.4 Topology of Secondary Dislocations 283
   - 4.5 Criteria of Specialness 284
   - 4.6 Beyond CSL Geometry: Structural Unit Models 284
   - 4.7 Noncubic Materials 285
5. **Techniques for the Study of Grain Boundaries** 285
   - 5.1 Electron Microscopy 285
   - 5.2 Field-Ion Microscopy (FIM) 285
   - 5.3 Computer Simulation 285
6. **Summary** 286

Glossary 286

Works Cited 287

Further Reading 287

INTRODUCTION

Grain boundaries are interfaces between identical crystals (or grains) that have different orientations. An example is shown in Fig. 1. They are a subset of solid-solid interfaces and share many features (most notably in terms of geometry) with phase boundaries, which are interfaces between unlike crystals. All such interfaces are commonly regarded as defects in crystals, reflecting a tradition of referring all structures to some ideal case or "reference structure," which is usually that of the perfect crystal. There is an important thermodynamic distinction between grain boundaries and phase boundaries, their geometric similarity notwithstanding: phase boundaries may exist in thermodynamic equilibrium, where a crystalline solid has two coexisting phases, but grain boundaries never exist in a single-phase solid in its thermodynamic ground state. In this sense, at least, grain boundaries truly are defects in crystalline solids. Grain boundaries usually exist in crystalline solids for kinetic reasons, and it is only through the application of elaborate precautions that single-crystal materials can be produced. Likewise, phase boundaries may never be able to reach a stable state and their structures are likely to be controlled by the kinetics of the phase transformation that creates them.

Although commonly regarded as planar or two-dimensional defects, grain boundaries are often not flat but may be curved or faceted, so that they frequently do not lie in a mathematical plane. They also may have finite width ranging between a few atom diameters and several hundred nanometers, depending upon

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

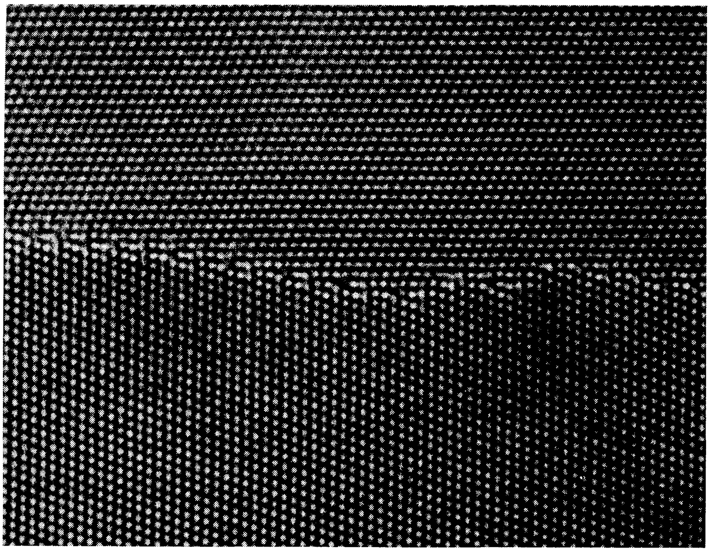

FIG. 1. Atomic-resolution electron micrograph of a tilt grain boundary in aluminum. Each white dot is a column of aluminum atoms, and the crystals are viewed along a common ⟨110⟩ direction, which is also the rotation axis. The rotation angle in this case is 90°. (Micrograph courtesy of Dr. Ulrich Dahmen of the National Center for Electron Microscopy at the Lawrence Berkeley Laboratory.)

the definition of "width" and measuring technique. These quibbles aside, however, the terms "planar" or "two-dimensional" do broadly describe the morphology of these defects. Grain boundaries form three-dimensional networks in polycrystalline materials, very much akin to the networks of liquid films that constitute foams.

The importance of grain boundaries lies in their effect upon a variety of material properties including strength, diffusion, and electrical conductivity. The net impact of these effects is great, because of the natural abundance of grain boundaries, and it was such effects that were first studied by Rosenhain, nearly a century ago (see, e.g., Ewing and Rosenhain, 1901). A typical engineering metal or ceramic is polycrystalline with a mean grain diameter of a few tens of microns. (The grain size is usually controlled in processing to optimize at least one property of the material.) At a grain size of 20 μm, a wedding ring of half a cubic centimeter contains approximately 500 cm^2, or about half a square foot, of grain boundary. As grain size is reduced, the area of grain boundary contained in any piece of solid crystalline material increases, and a subject of current importance is the preparation and properties of "nanocrystalline" materials with grain sizes in the range of nanometers and grain-boundary contents of the order of thousands of square meters per cubic centimeter.

In this brief review we will concentrate upon the structure and properties of grain boundaries rather than their control for engineering purposes. The latter subject remains at a very rudimentary stage of development, though with a great deal of potential at the time of writing (in the early 1990s).

1. GRAIN-BOUNDARY PARAMETERS

1.1 Crystal Misorientation

Grain boundaries are characterized in terms of the relative orientations of the crystals that they separate and the orientation of the grain-boundary plane. In most cases only bicrystals are addressed in present-day studies; one of the two crystals is designated as "crystal 1," while the other is "crystal 2." The default choice of reference frame for geometrical purposes is usually the natural axis system of crystal 1. It is always possible to find a simple rotation operation that would give the orientation of crystal 2 from the starting point of crystal 1 in the case of grain boundaries. Such rotation operations describe the misorientation of crystal 2 with respect to crystal 1, and may be described either as an axis-angle pair (where the axis is defined as a crystallographic direction in crystal 1, and the angle is the right-handed rotation about that axis that would bring crystal 1 into the orientation of crystal 2) or equivalently as a rotation matrix

R. Thus, for example, two equivalent descriptions of the same misorientation in cubic crystals might be

$$70.53°;[110]\equiv R=\begin{bmatrix} 0.67 & 0.33 & 0.67 \\ 0.33 & 0.67 & -0.67 \\ -0.67 & 0.67 & 0.33 \end{bmatrix} \tag{1}$$

The matrix product of any vector in crystal 1 with *R* gives the equivalent vector in crystal 2, but still referred to the axis system of crystal 1:

$$RX(1)=X(2). \tag{2}$$

Because of crystal symmetry, there may be more than one description of the misorientation of any grain boundary. For example, in the case of cubic crystals there may be 24 equivalent choices of axis systems in each crystal, meaning that there are as many as 1152 ($=2\times24^2$) possible descriptions of the misorientation between two cubic crystals. It is convenient to choose one particular description of the misorientation between two crystals, and this description has been dubbed the "disorientation." The disorientation is that description of the misorientation that embodies the smallest angle of rotation of all 1152 choices, and has the axis of rotation described as [*uvw*], where $u \geqslant v \geqslant w \geqslant 0$. The disorientation corresponding to the example given above is 60°; [111]. A distinction is often drawn between small-angle grain boundaries and large-angle grain boundaries, the latter having disorientation angles larger than approximately 15°.

1.2 Grain-Boundary Plane

Once the misorientation of two crystals is fixed, the orientation of the grain-boundary plane must be determined. The boundary plane is conventionally expressed as a crystallographic plane, relative to crystal 1. Grain-boundary planes are described as symmetrical or asymmetrical, depending on whether they have the same crystallographic form $\{hkl\}$ in both crystals. "Symmetrical" boundary planes need not contain crystal symmetry elements, and "asymmetrical" boundary planes need not be entirely devoid of them. Grain boundaries are commonly described as "tilt boundaries" when the boundary plane contains the axis of misorientation, and "twist boundaries" when the misorientation axis is perpendicular to the boundary plane, as illustrated in Fig. 2. Boundaries fitting neither description are sometimes described as having mixed character. While the distinction between tilt and twist grain boundaries is frequently useful, it must be remembered that such a description refers only to one particular description of the misorientation: a boundary that can be described as "twist" with respect to the disorientation can also have a "tilt" description with respect to another fully equivalent misorientation and a "mixed" description with respect to a third.

It is often convenient to define the grain-boundary plane normal **n**, which is a unit vector perpendicular to the boundary plane, pointing in the direction from crystal 1 into crystal 2.

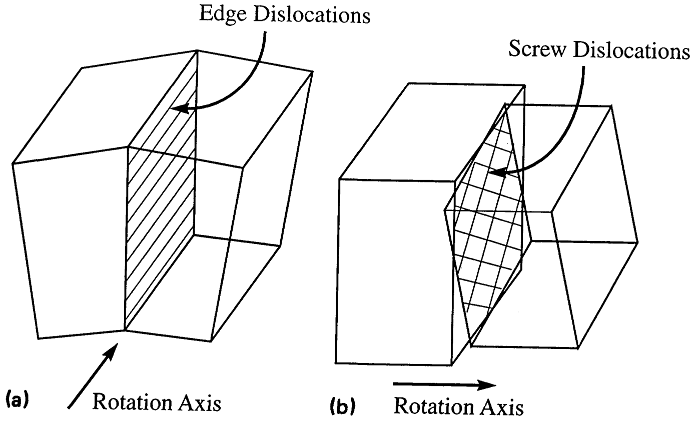

FIG. 2. Schematic diagrams showing (a) a symmetrical tilt grain boundary and (b) a twist grain boundary, and the dislocation arrays that they are normally expected to contain.

1.3 Macroscopic Degrees of Freedom

A grain boundary has five degrees of freedom with respect to crystallographic orientations, and these are now referred to as the macroscopic degrees of freedom, in contrast with the four additional microscopic degrees of freedom described below. Three of the macroscopic degrees of freedom relate to the relative orientation of the two crystals: two for the axis and one for the angle. The remaining two degrees of freedom relate to the orientation of the grain-boundary plane, which can be described by a single unit vector. It is possible through a variety of means to grow bicrystals in which all of the five macroscopic degrees of freedom of the grain are controlled.

2. GRAIN-BOUNDARY PROPERTIES AND BEHAVIOR

Grain boundaries have profound effects upon many aspects of material behavior and performance. In certain circumstances it is desirable to increase the density of boundaries in a material, while in others the opposite is true. We consider first the intrinsic properties of grain boundaries and then their effects upon material properties.

2.1 Energy

There is a free energy per unit area associated with every grain boundary, and it is usually denoted as γ. The grain-boundary energy varies with material, purity, grain-boundary parameters, and temperature, but for most metals it typically takes values in the range of about 0.5 $\mathrm{J\,m^{-2}}$. This energy is somewhat akin to a surface tension in that it always tends to minimize the grain-boundary area. Where three grain boundaries meet (a "triple line") the angles formed by the three boundary planes are determined by the relative energies of the grain boundaries. The variation of γ with misorientation is sufficient to cause the rotation of small single-crystal balls welded to single-crystal substrates as the energy of the intervening grain boundary is minimized (Herrmann *et al.*, 1975).

Because grain-boundary energy is a function of purity, it may be reduced through the segregation of impurity elements to the grain boundary. It is commonly the case that the composition of an alloy varies between the grain boundaries and the grain centers, and in many cases an impurity may be located exclusively in the grain boundaries of a specimen.

2.2 Mobility

Grain boundaries are able to move in response to various driving forces that might be applied to them. Atoms jump across the boundary from the shrinking crystal to the growing one, moving the boundary in the opposite direction. This is a thermally activated process and is typically observed at temperatures greater than half of the absolute melting temperature of the material. The mobility of grain boundaries depends upon the boundary parameters and on the purity of the material. Grain-boundary mobility is particularly sensitive to small amounts of impurity, which give rise to the "solute drag effect."

Mobility enables grain boundaries to move through a deformed material in response to the driving force provided by the strain energy stored in the crystals during the deformation process. This is called recrystallization. Grain-boundary migration leaves in its wake relatively defect-free crystals, which therefore contain lower strain energy, and the difference in energy between the deformed and recrystallized materials provides the driving force for grain-boundary migration. In contrast, the driving force for the process known as grain growth derives from the grain-boundary energy itself. As the grain size increases, the grain-boundary area and hence the total grain boundary energy are reduced. As with recrystallization, though, grain growth is effected by grain-boundary migration.

2.3 Diffusion

The transport of solute through a solid solvent, or the transport of atoms of the solid itself, occurs through the process of diffusion. At high temperatures, above two-thirds of the absolute melting temperature of the solvent, diffusion occurs predominantly through the crystal lattice, but below approximately half of the melting temperature, atom transport via the grain boundaries predominates, and the process is sometimes referred to as "short-circuit" diffusion.

Grain-boundary diffusion follows a modified form of Fick's law:

$$J = -D_{gb}\delta \frac{\partial C}{\partial x}, \tag{3}$$

where J is the diffusive flux, D_{gb} is the grain-boundary diffusivity, δ is a term with units of length, C is the solute concentration, and x is a distance measure. The value of δ is often identified with a "grain-boundary width" but in fact it is simply a term included in the equation to maintain dimensional consistency. It is usually only possible to determine the value of the product $D_{gb}\delta$ from an experiment, and the two terms cannot be disaggregated. Grain-boundary diffusion predominates at low temperature because of the small activation energy relative to that of bulk diffusion. It is mechanistically related to grain-boundary migration, although atomic jumps take place predominantly across the boundary in the case of the latter but along the boundary in the former. $D_{gb}\delta$ is found to depend upon the grain-boundary parameters, in some cases quite sensitively (Herbeuval *et al.*, 1973).

In a curious phenomenon known as diffusion-induced grain-boundary migration (or DIGM), grain-boundary diffusion is found to produce grain-boundary migration as a by-product (King, 1987). This results in the solute borne by the boundary being deposited in its wake, creating locally alloyed (or de-alloyed) regions.

Grain-boundary diffusion is capable of transporting atoms through significant distances at moderate temperatures, and it is a matter of some concern in the design of integrated circuits, where length scales are very small. In certain cases grain-boundary diffusion is important in the formation of circuit components, while in others it can be a mechanism for their failure. With ever-shrinking device dimensions, current densities in integrated circuits have increased markedly, and heat production is significant. This results in the devices running at temperatures in the range where grain-boundary diffusion occurs in certain of the component materials. A particular manifestation of grain-boundary diffusion in these devices is the phenomenon of electromigration, in which atoms drift in response to the electron "wind" in a conductor line. This drift is particularly problematical in aluminum interconnections, where the atom flux is carried along the grain boundaries and results in significant stress buildup in the material. A variety of different types of failure ensue from this phenomenon.

In addition to facilitating the buildup of stress in materials, grain-boundary diffusion can be an important mechanism by which a material responds to stress, as we shall see in the next section.

2.4 Strength of Materials

At low temperatures and high loading, grain boundaries act as barriers to crystal deformation, by impeding dislocation motion. As grain size decreases, the yield strength of a crystalline solid increases according to the Hall-Petch relationship (Hall, 1951; Petch, 1953):

$$\sigma_y = \sigma_i + k_y\, d^{-0.5}, \tag{4}$$

where σ_i is the lattice friction stress, d is the mean grain diameter, and k_y is a constant. Under these conditions, then, grain boundaries are desirable features in a material.

Grain boundaries promote brittle fracture, however, by providing an easy path for the propagation of cracks. The excess energy of fracture that is required to form two new surfaces is reduced by the grain-boundary energy that is released in fracturing the specimen. Grain-boundary fracture is especially promoted by segregating elements such as phosphorus and sulfur in iron. In the case of boron in the ordered alloy Ni_3Al, however, the impurity seems to have the opposite effect, removing the intrinsic propensity of the material to grain-boundary brittleness (Aoki and Izumi, 1979).

At high temperatures and low stresses grain boundaries become undesirable with respect to material strength. First, they become a locus of deformation because they allow the viscous sliding of one grain over another. This is a significant mode of failure in hot wires and filaments, which can develop a "bamboo" structure in which grain boundaries extend across the whole width of the material. Sliding on these boundaries causes a reduction of load- and current-bearing area as illustrated in Fig. 3. This increases the stress at the boundary (and also increases the temperature if the material is heated by the Coulomb effect), which eventually causes the separa-

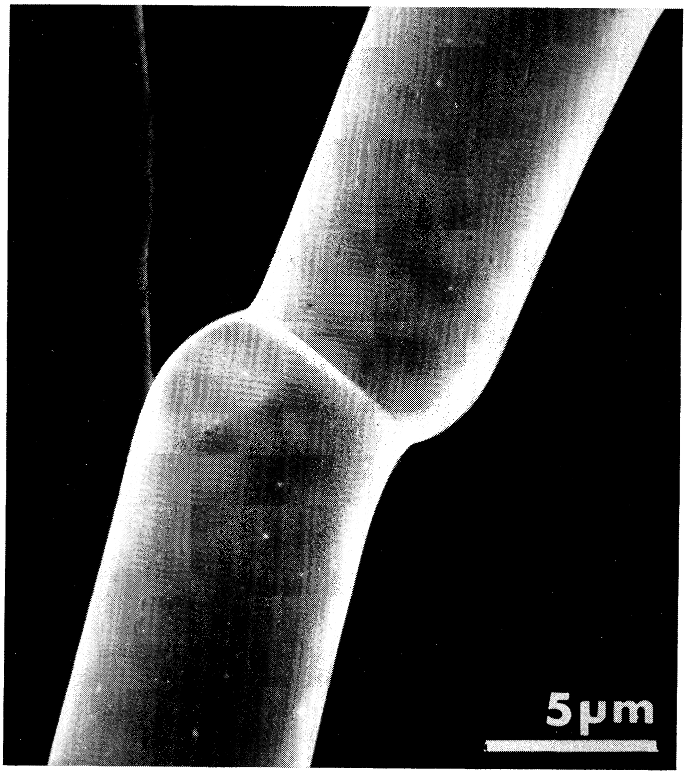

FIG. 3. Illustration of grain-boundary sliding in a thin gold wire. The grain boundaries in this "bamboo structured" specimen extended all the way across the wire and allowed for unconstrained sliding upon the application of stress at a temperature close to the melting point.

tion of the material. Filaments for incandescent lights are made from tungsten doped with impurities such as potassium that help to prevent grain growth and preserve a structure of long thin grains extending along the wire axis, which helps to prevent sliding of the type seen in Fig. 3. It is found that grain-boundary migration often accompanies sliding.

The second contribution of grain boundaries to creep deformation arises from their acting as sources and sinks for point defects, which makes possible diffusive deformation in which material is transferred from the sides of grains to their ends, making them longer and thinner in response to applied stress. When the mass transfer occurs through the grain interiors, Herring-Nabarro creep results at a strain rate that is proportional to the stress and d^{-2} (Herring, 1950; Nabarro, 1948). Coble creep is similar but occurs at lower temperatures by means of grain-boundary diffusion and has a strain rate proportional to d^{-3} (Coble, 1963). For these modes of deformation, then, large grain size is desirable for the resistance to deformation. This materials design paradigm is taken to the logical limit for gas-turbine blades as all grain boundaries are eliminated and the blades are cast as single crystals of high-performance superalloys.

There is a "threshold stress" of the order of 1 MPa below which no deformation occurs in diffusive creep. This is too small to be of any engineering significance in conventional materials. However, the threshold stress varies as $1/d$, and so there may be a regime of creep resistance at the extremely small grain sizes of nanocrystalline materials, provided that the problem of grain growth at high temperature can be solved.

Grain boundaries promote creep failure by acting as the sites of creep cavitation. At grain-boundary junctions, sliding allows one grain to act as a wedge forcing apart two others, giving rise to W-type, or wedge, cavities. At boundaries normal to an applied tensile stress, R-type or rounded cavities grow by accumulating vacancies emitted from the grain boundary. When creep cavities grow sufficiently large, they cause the failure of the material through the reduction of the cross-sectional area to the point where it can no longer bear the applied load.

Superplasticity is a phenomenon in which very large strains can be achieved at low strain rates, without fracture. This can be useful in forming complex shapes from sheet material, although the fabrication speed is typically slower than is usually associated with a manufacturing operation. Most applications of the phenomenon are found in the aerospace industry. Superplasticity requires fine grain sizes, and although the detailed mechanism by which it occurs is still not clear, it is certain that grain-boundary sliding plays a very important part.

2.5 Electronic Properties of Materials

Relatively little work has been done upon the effects of grain boundaries on electronic properties, especially in metals, where they add but a small contribution to the electrical resistivity. There is a greater wealth of information in the case of ceramic materials, some of which have rather more interesting electronic properties. For example, zinc oxide is used widely for its varistor properties, which make it very useful in the construction of electrical surge protection devices. Measurements made on a microscopic scale on this material reveal that the varistor effect only arises for current transport across grain boundaries, and within individual grains, there is no variation of resistance with current density. Thus the useful property of ZnO in this regard is effectively only a property of the grain boundaries themselves (Levinson, 1986).

In the high-temperature superconducting ceramics such as yttrium barium copper oxide, a major barrier to commercial device development is the small critical current density beyond which superconductivity breaks down. It has been found that the current-carrying capability of these materials is relatively good within single crystals but that the grain boundaries form weak links across which superconductivity is much less reliable. The weak-link behavior of the grain boundaries is strongly dependent upon the grain-boundary parameters, with a marked decrease in current-carrying ability being found as the misorientation angle increases (Dimos *et al.*, 1988).

3. SMALL-ANGLE GRAIN-BOUNDARY STRUCTURE

The structure of any grain boundary can be described formally in terms of arrays of dislocations lying in the boundary plane. This is perhaps most clearly seen in the cases of small-angle tilt and twist grain boundaries, which contain edge and screw dislocations, respectively, as shown in Fig. 2. For these simple cases, the dislocation spacing in the boundary plane is given by

$$S=\frac{b/2}{\sin(\theta/2)}, \tag{5}$$

where b is the Burgers-vector magnitude (i.e., the magnitude of the lattice displacement associated with each dislocation) and θ is the misorientation angle. This simplifies to $S\sim b/\theta$ in the small-angle approximation. For more general grain boundaries, it is necessary to adopt more complicated techniques that provide the overall dislocation content of the interface. Frank's formula is one such approach, which gives the total Burgers vector crossed by a "probe vector" lying in the grain-boundary plane (Frank, 1950):

$$\mathbf{B}=2\sin(\theta/2)(\mathbf{V}\times\mathbf{a}), \tag{6}$$

where $\mathbf{B}$ is the total Burgers vector crossed by the probe vector $\mathbf{V}$, and $\mathbf{a}$ is a unit vector parallel to the axis of rotation. Unfortunately, Frank's formula does not give unique results, because of the availability of different descriptions of the misorientation, which correspond to different choices of θ and $\mathbf{a}$. This ambiguity persists through all methods for determining the dislocation content of interfaces and gives rise to the concept of the "most physically meaningful" description of the grain boundary, which is the one that most closely describes the structure that nature selects. It is commonly held that the dislocation content that has the lowest elastic energy provides the most physically meaningful description of the grain boundary, and certainly where there is a "small-angle" description of a grain boundary, the disorientation is usually associated with the most meaningful description. In any case, it can be seen from Eq. (5) or (6) that as the misorientation angle increases, the dislocation content also increases or, equivalently, individual dislocation lines in the grain boundary get closer together. The conventional wisdom holds that for disorientations larger than about 15°, the dislocations that make up the grain boundary are so close together as to be physically indistinguishable from each other. It is this loss of individual dislocation character that marks the limit of small-angle grain-boundary behavior and the point beyond which dislocation descriptions of the grain-boundary structure are usually seen as having purely formal significance.

3.1 The O-Lattice

An important development in the geometrical theory of grain-boundary structure has been Bollmann's O-lattice approach (Bollman, 1970). The basis of the approach is the identification of locations of minimum strain between the two crystals, which are identified as those points of identical structure that occupy coincident positions. These "minimum strain sites" form a lattice known as the O-lattice, which extends throughout the bicrystal. The minimum strain sites are the O-lattice elements. The method is perhaps best developed rather abstractly by considering both crystals to extend to infinity, interpenetrating each other everywhere. This amounts to treating the crystals by considering their lattices only. The structures that are derived can be related to real grain boundaries by the formal artifice of selecting a grain-boundary plane and clothing the sites of lattice 1 with atoms or molecules on one side of the plane to create the crystal 1, and likewise clothing the sites of lattice 2 on the other side of the boundary plane.

With a lattice site chosen as an origin for the rotation that creates lattice 2 from lattice 1, the origin is by definition a point of good registry between the two crystals and thus a point of minimum strain. It is found that other "minimum strain points" exist at positions defined by the vectors $\mathbf{X}(\mathrm{O})$ that are solutions of the equation

$$(I-R^{-1})\mathbf{X}(\mathrm{O})=\mathbf{b}, \tag{7}$$

where $\mathbf{b}$ is a vector of crystal lattice 1. This equation also applies when the transformation between one crystal and the other is not a simple rotation, if R is replaced by a general transformation matrix. In the case of grain boundaries, where the transformation is most frequently a rotation, the O-elements take the form of infinite lines parallel to the rotation axis. Shear transformations give rise to O-lattices in the form of infinite planes parallel to the shear direction, and dilations and other general transformations (combinations of the three types listed here) create O-lattices made up of points.

The O-lattice elements, in fact, mark the possible origins for the transformation R, since they are positions at which the two crystal lattices are in perfect registry. It is this property of the O-elements that gave rise to the name of this formalism. The points of maximum disregistry between the two lattices are found on the planes that are perpendicular bisectors of the lines joining adjacent O-elements, as shown schematically in Fig. 4. Each O-element thus sits in a cell, surrounded by walls of maximum disregistry. In order to determine the structure of a grain boundary from the O-lattice, the boundary plane is drawn through the O-lattice construction, and lattice dislocations are placed at each intersection of the boundary plane with an O-cell wall, according to the "duality rules" given by Bollmann. This method for finding the dislocation structure of a grain boundary is essentially equivalent to Frank's formula, but instead of giving the average dislocation content, it allows us to determine individual dislocation positions at least on a purely geometrical basis.

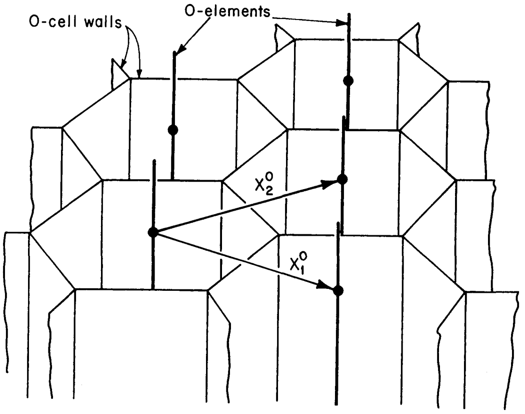

FIG. 4. Schematic illustration of the O-lattice construction. The O-elements are infinite lines parallel to the rotation axis for a grain boundary, and the O-cell walls situated between them are the positions of maximum disregistry between the two crystal lattices that make up the grain boundary. When a grain-boundary plane is chosen, the intersections of the boundary plane with the O-cell walls mark the positions of the dislocations in the boundary plane.

Energetic considerations can alter certain details of the predicted arrays, but cannot change the overall Burgers-vector content. This approach can also be extended to certain problems of large-angle grain-boundary structure, as we shall see below.

3.2 Grain-Boundary Energy

Once the dislocation structure of a grain boundary is known, the energy of the interface can be found, in principle, as the elastic strain energy of the array of dislocations that comprise the boundary. This was first done by Read and Shockley (1950), who obtained the result that the strain energy per unit area of interface, γ, for a symmetrical tilt boundary is

$$\gamma = A\theta(B - \ln\theta), \tag{8}$$

where A is a term proportional to the shear modulus and the Burgers vector, and B is a term that depends upon the exact geometry of the interface. This formula has been experimentally confirmed in numerous studies, and it applies well up to a misorientation of about 15°, especially in metal specimens. Beyond this angle, the formula predicts decreasing energy with increasing misorientation, but its basis is dubious because of the closeness of the lattice dislocations that make up the boundary. Grain boundaries in ceramic materials are amenable to this dislocation-based analysis, but only to the extent that they have crystalline grain boundaries. In many cases, particularly where the material is impure, the grain boundaries in ceramics are covered with a thin film of an amorphous (or glass-like) second phase, and the properties of the boundary then become simply the properties of this second phase.

An important finding of this theory is that the displacements due to the interfacial dislocations make up a strain-free rotation at distances from the boundary greater than about S, where S is the separation of the dislocations in the boundary plane. Within this distance from the boundary, there are strains and therefore stresses associated with the dislocations, but essentially all of energy of the interface is contained within a region with width of the order of S, surrounding the boundary.

4. LARGE-ANGLE GRAIN-BOUNDARY STRUCTURE

The structures of large-angle grain boundaries are more complicated than those of small-angle ones, and as yet there is no theory that allows for the analytical prediction of the structure of any randomly chosen boundary. However, observations of the properties of large-angle boundaries have yielded nearly constant results with respect to misorientation or boundary-plane inclination, except for a few "special" boundaries with particular misorientations. The structures of these special boundaries have been investigated with considerable vigor, both experimentally and theoretically.

4.1 Geometrical Characteristics of Special Large-Angle Grain Boundaries

The best known special interface is perhaps the twin boundary created by inversion of the stacking sequence of close-packed planes from ...*ABCABC*... to ...*CBACBA*... in face-centered-cubic metals. The boundary may also be described by the disorientation 60°;[111]. This boundary exhibits, *inter alia*, low interfacial energy, a high degree of sensitivity of interfacial energy to boundary orientation, low diffusivity, and limited effect upon various forms of deformation. Other special grain boundaries are found to be associated with other disorientations, such as 36.9°;[100] and 50.5°;[110] in cubic materials.

In each case of special boundary behavior, it is found that the misorientation of the boundary is at or close to a misorientation that creates a coincidence-site lattice, or CSL, which is a special form of coincidence between the two crystal lattices. In Sec. 3.1, we introduced the idea of lattice coincidence in a rather general form, in which a coincidence between any point of lattice 1 and a corresponding point of lattice 2 was considered as a point of minimum disregistry between the lattices. Such O-points can occur anywhere in the lattice, and they always form a lattice. The CSL, by contrast, considers only coincidences between lattice 1 and lattice 2 that occur at lattice points.

For simple crystal structures, the lattice points may also be taken to be atom sites. While the O-lattice exists at all misorientations and changes its unit-cell size, shape, and orientation continuously with crystal misorientation, coincidence-site lattices exist only at certain discrete misorientations. Obviously, where a CSL exists, at least some of the O-elements must also pass through coincident lattice sites. Figure 5 shows an example of a CSL in a simple-cubic material and a corresponding grain-boundary geometry (which we will shortly distinguish from grain-boundary structure) for a symmetrical tilt grain boundary. It can be seen that the grain-boundary geometry is periodic and highly ordered. The period of the grain-boundary geometry is related to the size of the CSL, and it is generally regarded as desirable to have an index of the size of the CSL unit cell. It is common to use the parameter Σ, which is defined as the ratio of the CSL unit-cell volume to the crystal unit-cell volume. For the example in Fig. 5, $\Sigma=5$, and greater values of Σ correspond to lower densities of coincident sites.

The misorientations giving rise to CSLs in cubic crystals are predicted by the generating function

$$\theta=2\arctan(pN^{1/2}/q), \tag{9}$$

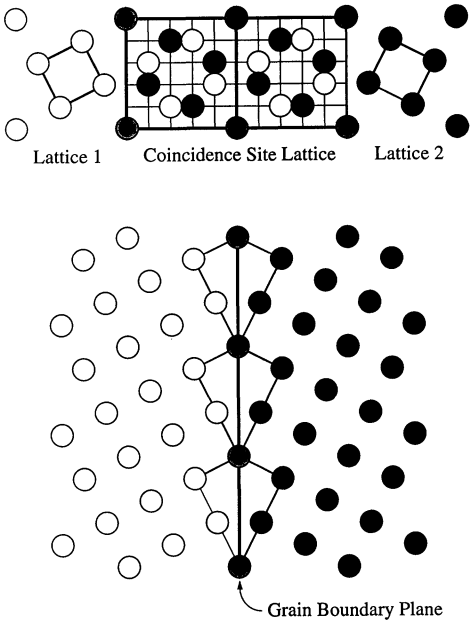

FIG. 5. A coincident-site lattice made up by the intersection of two crystal lattices, and a grain-boundary geometry based upon the CSL. Two unit cells of the CSL are shown, and the fine mesh within each CSL unit cell represents the DSC lattice.

where p and q are integers, and N is $h^2+k^2+l^2$ if $[hkl]$ is the rotation axis. The value of Σ is then given by

$$\Sigma=(p^2N+q^2)/2^n, \tag{10}$$

where n is chosen to make Σ an odd integer. A rotation matrix R that produces a CSL can be written in the form

$$R=(1/\Sigma)R', \tag{11}$$

where R' is a 3×3 matrix all of whose elements are integers.

Special grain-boundary properties are generally associated with small values of Σ (for example, the common twin in fcc crystals is a $\Sigma=3$ coincidence boundary, commonly denoted as $\Sigma3$) but not all small values of Σ necessarily predict special properties. Table 1 gives the disorientations leading to CSLs with Σ up to 25 for cubic crystals, and most of these correspond to large angles of misorientation between the two crystals. Even so, every boundary could, formally at least, be described in terms of lattice dislocations exactly as for small-angle grain boundaries. A special feature of boundaries that have CSL geometry is that the lattice dislocations they contain are always separated by integer numbers of exact climb distances (for the case of edge disloca-

Table 1. Disorientations corresponding to short-period CSLs in cubic crystals. Where more than a single, physically distinct rotation gives rise to the same value of Σ, the Σ value is conventionally postfixed with a lowercase letter, chosen in order of disorientation angle.

Σ	Axis	Angle (deg)
3	[111]	60.00
5	[100]	36.87
7	[111]	38.21
9	[110]	38.94
11	[110]	50.48
13a	[100]	22.62
13b	[111]	27.80
15	[210]	48.19
17a	[100]	28.07
17b	[221]	61.93
19a	[110]	26.53
19b	[111]	46.83
21a	[111]	21.79
21b	[211]	44.40
23	[311]	40.45
25a	[100]	16.25
25b	[331]	51.68

tions in tilt boundaries) or by integer multiples of the distance between Peierls valleys (for the case of screw dislocations in twist boundaries). Climb distances and the distances between Peierls valleys are the minimum separations between locally stable dislocation positions in the crystal lattice.

Alternatively stated,

$$S = MS^0, \tag{12}$$

where M is an integer and S^0 is the smallest distance between stable dislocation positions imposed by the crystal lattice. Thus for the case of coincidence boundaries, and only for that case, the lattice dislocations making up the boundary can simultaneously exist at uniform separation and at locally stable positions relative to the crystal lattice. For all other cases, one of those two conditions for a low-energy structure cannot be fulfilled: in small-angle grain boundaries, where the dislocations are relatively well separated, it appears that the perturbation of a uniform spacing, say S, into individual spacings of MS^0 and $(M \pm 1)S^0$ has relatively little effect upon the grain-boundary energy because the interactions between adjacent dislocations are relatively weak. For large-angle grain boundaries, though, the energy associated with such a perturbation can be large.

4.2 Structure of Coincidence-Related Boundaries

Although the periodic geometry of a grain boundary may be well described by a CSL, the actual structure of the boundary may not contain any atoms shared between the two crystals. This is because of possible local reconstruction of the boundary to minimize its energy. One form of reconstruction that is particularly important in the metals removes a certain amount of the free volume associated with a structure that contains atoms in coincident sites. Consider, for example, the geometry shown in Fig. 5. If we draw circles on each lattice site to indicate the ionic radii, we find gaps and overlaps in the structure as shown in Fig. 6(a). The overlaps in particular can be relieved if crystal 2 is translated with respect to crystal 1, as shown in Fig. 6(b), and computer simulation studies have consistently found that structures of lower symmetry but better space-filling properties, as in Fig. 6(b), have lower energy. This gives rise to the addition of three additional degrees of freedom for grain boundaries, corresponding to the magnitudes of the rigid translations of one crystal with respect to the other, in three orthogonal directions. These three degrees of freedom are often referred to as "microscopic" because they relate to translations that must be smaller than the unit cell of the crystal. A fourth microscopic degree of freedom exists when the crystal structure has more than one atom per lattice site in the unit cell: when this condition exists, as, for example, in ionic compounds, then there is also a nondegenerate freedom of choice of the site of the origin. This may equivalently be described as a degree of freedom related to the position of the grain boundary along its normal. A grain boundary may, then, have a total of nine degrees of freedom, five of which are considered macroscopic, and four microscopic.

The actual structure of the grain boundary will depend upon matters such as space filling in metals, bond-bending and dangling-bond avoidance in covalent solids, and Coulombic interactions in ionic materials. It is also found that energy differences between different stable structures can be small enough that the

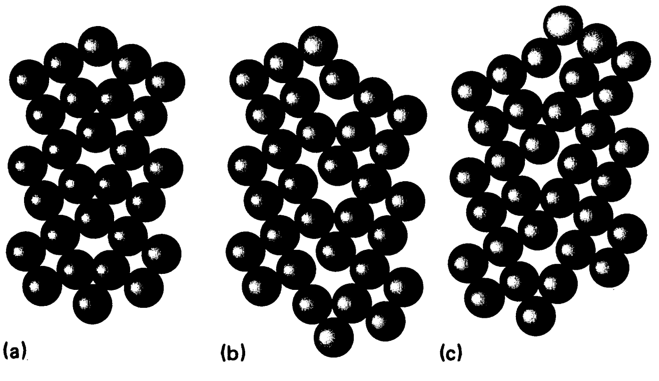

FIG. 6. Grain-boundary structures based upon the geometry illustrated in Fig. 5. (a) With the grains shown at the coincidence position (sharing atoms in the grain-boundary plane). (b) With the atomic overlaps in (a) removed by a rigid translation of the right-hand grain with respect to the left-hand grain. (c) Similar to (b) but with a different translation: this yields a structure that is symmetry related to the structure in (b) and can be transformed into (b) by a rotation of 180° about a vertical axis.

structures may coexist in a single grain boundary as adjacent patches of different structure. For certain cases, the structures may be related by symmetry and therefore embody identical energy: an example of two symmetry-related structures, based upon the Σ5 example, is given in Figs. 6(b) and 6(c).

4.3 Off-Coincidence Grain Boundaries

Although CSL-related grain boundaries occur at unique special orientations, it is found in practice that the special properties of these interfaces persist (although possibly diminished) over some range about the exact CSL-producing misorientation. For this reason, it is desirable to have a theory of the structure of "off-coincidence" or "near-coincidence" grain boundaries. As described above, CSL-related boundaries have lattice dislocations separated uniformly by distances MS^0. When the average dislocation spacing is slightly reduced or enlarged to change the angle of misorientation, a small number of spacings $(M-1)S^0$ or $(M+1)S^0$ are introduced into the lattice dislocation array, as shown in Fig. 7. These perturbations of the grain-boundary structure are defects of the defect array itself, and are sometimes referred to as "secondary defects." It can be shown that they correspond to additional dislocations superimposed upon the CSL-producing array (see, e.g., Balluffi, 1980). Their Burgers vectors can be defined, and their spacings (if uniform) are given by Eq. (4) with the simple substitution of $\Delta\theta$ for θ, where $\Delta\theta$ is the angle by which the misorientation deviates from the exact CSL misorientation.

Off-coincidence grain boundaries can therefore be described in very much the same way as small-angle grain boundaries, since the deviation from the exact coincidence misorientation is made up by superimposing a pseudo-small-angle grain boundary on the structure of the CSL interface. The energy of the grain boundary can be estimated simply by adding the energy of the superimposed small-angle interface to that of the coincidence boundary.

The O-lattice method can be applied to the so-called "secondary dislocation array" that accommodates the deviation from the ideal coincidence misorientation, by a simple modification of Eq. (7):

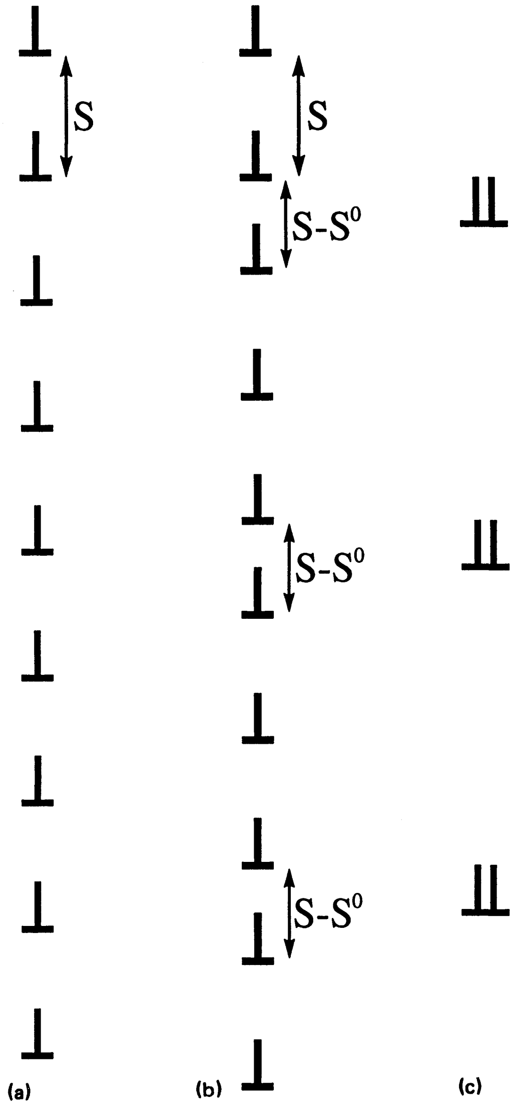

FIG. 7. (a) An equispaced array of primary edge dislocations making up a symmetrical tilt grain boundary. (b) The same array as in (a) but at a slightly greater angle of misorientation such that the dislocation array is forced to include two different spacings. (c) The array of secondary dislocations equivalent to the minority spacings in (b). The structure of the boundary can be considered as a superposition of (a) and (c), which is a particularly useful approach when (a) represents a coincidence boundary.

$$(I-R_cR^{-1})\mathbf{X}(O2)=\mathbf{d}. \qquad (13)$$

In this expression, R_c is the rotation that gives the CSL, $\mathbf{X}(O2)$ is a vector of the secondary O-lattice, and $\mathbf{d}$ is a Burgers vector appropriate to a secondary dislocation, as described below.

4.4 Topology of Secondary Dislocations

The dislocations that may exist in grain boundaries can be classified in a number of different ways. The most general of these is to consider which symmetry elements of the crystals are involved in the creation of the defect. In any crystal, the structure found on any particular plane is replicated at different locations on the same plane or parallel planes (via translational symmetry); or it may be replicated on the same plane but in a different orientation (via rotational or mirror symmetry); or it may be replicated on some nonparallel, equivalent plane (via rotational symmetry). In any case, it is possible to make a discontinuous displacement of one crystal that corresponds to one of these symmetry elements of the interfacial plane, and the line along which such a displacement is made is the core of an interfacial dislocation. The various possibilities of such defects according to Pond (1989) are illustrated schematically in Fig. 8. In each of these cases, the dislocations separate patches of identical interfacial structure and may thus be considered "perfect" dislocations within the interface. In most cases, the Burgers vectors of these dislocations are not translation vectors of the crystal lattice, and so they cannot exist in the grain interiors without the creation of high-energy stacking faults.

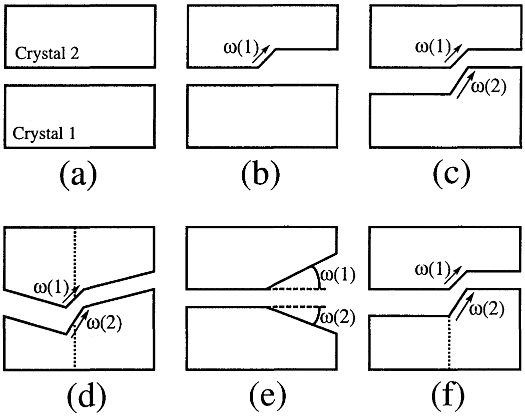

FIG. 8. Classification of line defects in interfaces according to the symmetry elements that they affect. In each case, ω is a symmetry operation in one of the crystals, and the crystals are shown as though separated along the interface. (a) The case of a perfect interface containing no defects; (b) a crystal lattice dislocation in grain 1; (c) a broken translation dislocation; (d) a "frustrated symmetry" dislocation; (e) an interfacial disclination; (f) an interfacial dislocation delineating the intersection of a domain boundary, or other stacking fault in crystal 2, with the interface [after Pond (1989)].

It is to be noted that in addition to the Burgers vector that characterizes the grain-boundary dislocations, these defects also embody a step in the interfacial plane in many cases. Motion of the step causes grain-boundary migration, and this is taken to explain the observation of grain-boundary migration in concert with sliding, when the sliding is mediated by the motion of grain-boundary dislocations. Any reaction of these dislocations must satisfy the usual node condition that $\Sigma \mathbf{b}_i = \mathbf{0}$, where $\mathbf{b}_i$ is the Burgers vector of the ith dislocation entering the node. There is, for interfacial dislocations, the additional requirement that $\Sigma h_i = 0$, where h_i is the height of the step in the boundary plane associated with the ith dislocation.

Perhaps the most widely studied class of secondary dislocations is the group defined by translational symmetries of the grain-boundary planes. This includes the so-called "displacement shift complete" (DSC) dislocations, whose Burgers vectors represent displacements of the crystals with respect to each other that shift the interfacial structure but leave it "complete." Various techniques have been developed to determine the Burgers vectors of these dislocations ("DSC vectors") and the step heights associated with them. An important result is that the DSC lattice is reciprocal to the CSL, for structures described by primitive unit cells. The structures of the DSC dislocations are specific to the CSL systems to which they relate, and their Burgers vectors are smaller than the Burgers vectors of crystal lattice dislocations. Since crystal lattice dislocations always correspond to translational symmetries of the crystal structure, they must also correspond to translational symmetries of the grain-boundary structure and may therefore be perfect dislocations within the grain boundary. In every case, the lattice Burgers vector is a sum of primitive DSC vectors so that lattice dislocations can dissociate upon entering grain-boundary planes, forming a number of DSC dislocations out of a single crystal lattice dislocation.

A special kind of defect can exist in a grain-boundary plane, and this is the "pure step," essentially a DSC dislocation with a null Burgers vector. The energies of interfacial dislocations, like ordinary crystal dislocations, have a component that is proportional to the square of the Burgers vector magnitude, but there is also a component that is proportional to the height of the step in the boundary plane. In certain cases the step energy can be of comparable magnitude to the strain energy, and so it can exert a controlling influence upon dislocation reactions (King and Smith, 1980).

4.5 Criteria of Specialness

As described above, the distinct properties of "special" grain boundaries extend over some range of misorientation close to certain exact coincidence values. Our understanding of this phenomenon is based upon the analogy with small-angle grain boundaries. Just as the dislocation description for a small-angle boundary loses utility when the constituent dislocations are too close together to be considered separately, the description of an "off-coincidence" large-angle boundary also breaks down, too. The point at which the secondary dislocations lose their individual physical identities is often considered to be the point at which the special properties of the boundary are lost. A variety of approximate formulas have been proposed to provide rules of thumb for the misorientation range over which special properties are retained.

The most commonly used of these criteria of specialness is the Brandon criterion (Brandon, 1966):

$$\Delta\theta_{\max}=15^\circ/\sqrt{\Sigma}. \tag{14}$$

This is the maximum rotational deviation from the coincidence misorientation for which a grain boundary is considered to be "coincidence related." It is based upon an approximate measure of dislocation spacing in the grain boundary and presumes that the secondary dislocation array comprises only primitive DSC dislocations. Like most rules of thumb, this provides a reasonable guideline but is by no means exception-free. The most significant problem with using the Brandon criterion, or any other developed from a consideration of dislocation spacing alone, is that all coincidence-related boundaries up to some arbitrarily chosen value of Σ are considered to embody special properties, and although a small value of Σ seems to be a prerequisite for a boundary to have special properties, not all low-Σ boundaries do in fact exhibit them. What is needed, but is still unavailable, is some rule of thumb to determine which interfaces are fundamentally "special."

4.6 Beyond CSL Geometry: Structural Unit Models

Computer simulation studies of grain-boundary structure (see Sec. 5) have shown that low-energy periodic boundaries are made up of recognizably shaped clusters containing between four and approximately ten atoms (Sutton and Vitek, 1983). The lower-energy boundaries tend to have structures that are characterized as being completely made up of a single type of cluster (say *P*) arranged periodically so as to occupy the whole interfacial area. The structure of this boundary might then be described as ...*PPPPP*..., and it would also be a coincidence grain boundary. At a different misorientation about the same rotation axis another low-energy boundary might be found with a structure made up entirely of a different structural unit, for example, ...*QQQQQ*... . This would also be a coincidence boundary. At misorientations between those of the *P* and *Q* boundaries, the structure is made up of simple "lever-rule" mixtures of the two types of structural unit. Close to the *P* misorientation, then, the structure would take the form ...*PPPPPQPPPPPQPPPPP*... . It is also found that the *Q* units in this interface correspond to the cores of DSC dislocations appropriate to the *P* CSL boundary, and thus the structural unit model and the CSL model for large-angle boundary structure are different descriptions of essentially the same thing.

Interestingly, at a misorientation halfway between the *P* and *Q* misorientations, the boundary structure is ...*PQPQPQP*..., which is also a coincidence boundary. It is found, however, that this boundary is one of those which does not embody special properties. Boundaries made up purely of a single structural unit are termed "favored" boundaries, and these have special properties as described above. The structures of other boundaries are essentially mixtures of the structures of favored boundaries. This provides some insight into the question of which values of Σ are significant, but still calls for computer simulation studies to determine which boundaries are likely to be special.

4.7 Noncubic Materials

It is not generally possible to form three-dimensional CSLs in noncubic crystal structures. It is required that there exist special relationships between the lengths of the unit-cell edges and the angles between them in order for CSL formation to be possible. In most cases, however, the crystal structure is quite close to such a special relationship, and then CSLs can be formed if there is a slight deformation or "constraint" of the crystal lattice. The boundary structure can then include an additional dislocation content that exists to accommodate the constraint of the crystal lattice. Constrained-CSL (CCSL) models have been successfully applied to explain observed grain-boundary structures in hexagonal, tetragonal, and orthorhombic crystals.

5. TECHNIQUES FOR THE STUDY OF GRAIN BOUNDARIES

A very wide range of techniques has been applied to the study of grain boundaries, and it would be impossible to list all of the types of experiments that have been performed. In many cases direct measurements of properties of interest are made either using bicrystals prepared with known boundary parameters or using polycrystalline specimens and accepting whatever boundaries are present in the material. Most studies of grain boundaries *per se* concentrate upon the structure or chemistry of the boundary in order to draw inferences about the ways in which structure and chemistry affect properties of interest. Three techniques stand out from the many that are applied.

5.1 Electron Microscopy

This accounts for the largest number of reported studies of grain boundaries and provides a wealth of different types of information. High-resolution transmission electron microscopy (TEM) can allow the determination of atomic-level structure of the grain boundary, provided that low-index crystallographic directions in each grain lie parallel to each other. Figure 1 is an example of an atomic-resolution transmission electron micrograph.

Conventional TEM, using diffraction imaging, can provide information about the primary and secondary dislocation structure of grain boundaries. When TEM is coupled with either an x-ray spectrometer or an electron energy-loss spectrometer, elemental analysis can be performed with spatial resolution depending upon the size of the electron probe—typically several nanometers. Analytical electron microscopy of this kind has provided much information about the segregation of solutes to grain boundaries.

5.2 Field-Ion Microscopy (FIM)

Although used much less than TEM, FIM provides atomic-level imaging of solid material without any constraints upon the boundary parameters and can be coupled with mass spectrometry to provide chemical analysis on an atom-by-atom basis. It has been applied to a number of grain-boundary segregation problems, but has provided structural information of lower content than TEM investigations. This is largely because it is not possible to create FIM specimens containing grain boundaries of known parameters with the ease that is available to the electron microscopist.

5.3 Computer Simulation

Modern computers have facilitated the construction of atomic models in which each atom is represented by a point and the interatomic forces are modeled with varying sophistication. The earliest models used pairwise central force approximations such as the Lennard-Jones potential, or more realistic pseudopotentials designed to model the properties of actual elements. More recently, certain many-body potentials have been developed that are still determined only by the atomic positions and apply very well to the metals. Potentials have been developed to include bond-bending effects applicable to covalently bonded solids such as silicon, and finally, complete quantum-mechanical solutions have also been used. In every case, the model consists of a finite block of atoms containing a section of grain boundary. Periodic boundary conditions are typically applied, which means that atoms close to one edge of the model are used, as though translated by the model dimension, to determine forces acting upon atoms close to the opposite edge of the model. The model thus becomes infinite and periodic. In principle, any grain boundary can be constructed as a computer model of this type, although available memory limits the size of the model and thus imposes a requirement that the boundary have a certain periodicity in its structure. Models can be "dynamic," allowing atomic vibration subject to a chosen total kinetic energy (equivalent to a chosen temperature), or they may be "static," effectively becoming models appropriate to absolute zero temperature. In either case, the models can be used to find lowest-energy structures and to measure internal stresses around the interface. In the case of dynamic models, the behavior of the boundary can also be investigated with respect to phenomena such as sliding, migration, and diffusion.

6. SUMMARY

Grain boundaries have been studied now for a century and their properties and behavior are quite familiar, although with as many as nine degrees of freedom, they are more different from each other than are snowflakes. Most research in the last two decades has centered on understanding grain-boundary structure and how it affects properties and behavior, but general rules are hard to find for such richly varied defects.

As individual grain boundaries become better understood, it becomes possible to consider how the network of boundaries that extends throughout a polycrystal controls the properties of the specimen. Important considerations then become the distribution of grain-boundary characters in the specimen and the behavior of "triple lines," where three grain boundaries meet. Much present research focuses on the possibility of controlling the distribution of grain-boundary parameters in polycrystalline specimens in order to optimize the properties of the material. This represents a significant refinement of traditional microstructure control, which typically concerns itself only with average grain size for single-phase materials. The payoff from such research is potentially very great, particularly with regard to electrical properties. For example, the critical current density in many ceramic superconductors can be radically increased by replacing "bad" grain boundaries with "good" ones. Techniques for controlling the grain-boundary character distribution include traditional thermal processing, which has been shown to have significant effects on the percentage of special boundaries. Other approaches include directional solidification from the melt and the sintering of prealigned powders into polycrystalline compacts. In the realm of thin-film technology, crystal orientation can often be controlled by suitable choice of substrate material, and where more than one crystal orientation is formed on a particular substrate, the result is a network of highly selected grain boundaries. As the ability to control the formation of various grain boundaries is perfected, it will be possible to build into materials the desirable properties of specific interfaces while avoiding the undesirable properties of others. This may result in the creation of usable electronic devices such as tunneling junctions out of individual grain boundaries, or may allow the creation of alloys with tailored grain boundaries that have enhanced resistance to embrittlement or other forms of metallurgical failure.

GLOSSARY

CCSL: Constrained coincidence-site lattice.

CSL: Coincidence-site lattice.

DIGM: Diffusion-induced grain-boundary migration.

Disorientation: The choice of misorientation (*q.v.*) embodying the smallest rotation angle and having a rotation axis described as [*uvw*], where $u \geqslant v \geqslant w \geqslant 0$.

DSC Dislocation: A dislocation whose Burgers vector is a vector of the DSC lattice.

DSC Lattice: "Displacement shift complete" lattice; lattice made up of vectors that can be Burgers vectors for perfect dislocations in grain-boundary planes.

Facet: Flat section of interface with well-defined extent and plane.

GBD: Grain-boundary dislocation.

General Boundary (Sometimes "Random Boundary"): Large-angle boundary embodying no special properties.

Large-Angle Boundary (Sometimes "High-Angle Boundary"): Grain boundary with a disorientation angle greater than approximately 15°.

Ledge: Linear discontinuity in a grain-boundary plane with either dislocation character, step character, or both.

Misorientation: Any description of the relative orientations of two crystals separated by a grain boundary.

O-Lattice: Lattice made up of possible origins for the mathematical transformation of one crystal into the other.

Secondary Dislocation: A DSC dislocation (*q.v.*) where the Burgers vector is not a lattice vector of either crystal.

Small-Angle Boundary (Sometimes "Low-Angle Boundary"): Grain boundary with a disorientation angle less than approximately 15°.

Special Boundary: A large-angle boundary with properties significantly different from general boundaries; usually corresponds to a CSL in cubic materials.

Triple Line: The intersection of three grain boundaries; topologically the only way a grain boundary can terminate other than at a free surface.

Twin Boundary: Any boundary with a 180° misorientation. Frequently applied to mean the $\Sigma3$ boundary, which is observed profusely in many cubic materials.

List of Works Cited

Aoki, K., Izumi, O (1979), *Nippon Kinzoku Gakkaishi* **43**, 1190.

Balluffi, R. W. (1980), in: R. W. Balluffi (Ed.), *Grain Boundary Structure and Kinetics*, Metals Park, OH: ASM.

Bollmann, W. (1970), *Crystal Defects and Crystalline Interfaces*, Berlin: Springer.

Brandon, D. G. (1966), *Acta Metall.* **14**, 1479–1484.

Coble, R. L. (1963), *J. Appl. Phys.* **34**, 1679–1682.

Dimos, D., Chaudhari, P. Mannhart, J., LeGoues, F. K. (1988), *Phys. Rev. Lett.* **61**, 219–222.

Ewing, J. A., Rosenhain, W. (1901), *Philos. Trans. R. Soc. London Ser. A* **195**, 279–301.

Frank, F. C. (1950), *Symposium on Plastic Deformation of Crystalline Solids*, Metals Park, OH: ASM, p. 150.

Hall, E. O. (1951), *Proc. Phys. Soc. London* **64**, 747–753.

Herbeuval, I., Biscondi, M., Goux, C. (1973) *Mem. Sci. Rev. Met.* **70**, 39–46.

Herring, C. (1950), *J. Appl. Phys.* **21**, 437–445.

Herrmann, G., Sautter, H. Baro, G., Gleiter, H. (1975), in: J. L. Walter, J. H. Westbrook, D. A. Woodford (Eds.), *Grain Boundaries in Engineering Materials*, Baton Rouge, LA: Claitor.

King, A. H. (1987) *Int. Mater. Revs.* **32**, 173–189.

King, A. H., Smith, D. A. (1980), *Acta Crystallogr.* A**36**, 335–343.

Levinson, L. M. (1986), in: C. S. Pande, D. A. Smith, A. H. King, J. L. Walter (Eds.), *Interface Migration and Control of Microstructure*, Metals Park, OH: ASM.

Nabarro, F. R. N. (1948), *Report of the Conference on Strength of Solids*, London: The Physical Society.

Petch, N. J. (1953), *J. Iron Steel Inst.* **174**, 25–28.

Pond, R. C. (1989), in: F. R. N. Nabarro (Ed.), *Dislocations in Solids*, Amsterdam: North-Holland.

Read, W. T., Shockley, W. (1950), *Phys. Rev.* **78**, 275–289.

Sutton, A. P., Vitek, V. (1983) *Philos. Trans. R. Soc. London Ser. A* **309**, 1–36.

Further Reading

Balluffi, R. W. (1980), *Grain Boundary Structure and Kinetics*, Metals Park, OH: ASM.

Bhadeshia, H. K. D. H. (1987), *Worked Examples in the Geometry of Crystals*, London: The Institute of Metals.

Bollman, W. (1970), *Crystal Defects and Crystalline Interfaces*, Berlin: Springer.

Chadwick, G. A., Smith, D. A. (1976), *Grain Boundary Structure and Properties*, New York: Academic.

Forwood, C. T., Clareborough, L. M. (1991), *Electron Microscopy of Interfaces in Metals and Alloys*, New York: Hilger.

Gleiter, H., Chalmers, B. (1972), *High Angle Grain Boundaries*, New York: Pergamon.

Kaur, I., Gust, W. (1989), *Fundamentals of Grain and Interphase Boundary Diffusion*, Stuttgart: Max-Planck-Institut für Metallforschung.

McLean, D. (1957), *Grain Boundaries in Metals*, Oxford: Oxford University Press.

Walter, J. L., King, A. H., Tangri, K. (1991), *Structure and Property Relationships for Interfaces*, Materials Park, OH: ASM.

GRAPHITE

M. S. DRESSELHAUS AND G. DRESSELHAUS, *Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*

Introduction 289
1. **Structure** 290
2. **Graphite-Related Materials** 291
2.1 Kish Graphite 291
2.2 Highly Oriented Pyrolytic Graphite 292
2.3 Graphitic Films 292
2.4 Carbon Fibers 292
2.5 Glassy Carbon 293
2.6 Graphite Intercalation Compounds 294
3. **Properties of Graphite** 294
3.1 Lattice Properties 294
3.2 Electronic and Transport Properties 296
3.3 Optical Properties 298
3.4 Thermal Properties 299
3.5 Mechanical Properties 299
Acknowledgments 300
Glossary 300
Works Cited 301
Further Reading 301

INTRODUCTION

Carbon materials are unique in many ways. One distinction is related to the various allotropic forms that carbon materials can assume. Under ambient conditions, the graphite phase with strong in-plane trigonal bonding is the stable phase, as indicated by the phase diagram of Fig. 1 (Bundy, 1985). Under the application of high pressure and high temperature, transformation to the tetrahedrally bonded cubic diamond structure takes place. Once the pressure is released, diamond remains essentially stable under ambient conditions although, in principle, it will very slowly transform to graphite. Figure 1 shows that for the direct transformations from graphite to diamond, high temperatures and pressures are required, but these high temperature and pressure requirements can be somewhat reduced by the use of catalytic particles, e.g., of iron. Another crystalline phase found at high temperatures and pressures is hexagonal diamond. The phase diagram further indicates that other carbon forms are obtained by exposure to shock waves and rapid solidification, though these carbon forms have not been studied in much detail. Although considerable attention has been given to study of the phase diagram of carbon, many unanswered questions remain concerning the high-temperature, high-pressure phases on the one hand, and the relation to the phase diagram of the recently discovered C_{60} and higher-mass fullerenes, on the other hand.

Graphite tends to be a material of extremes. The structure and properties of graphite are highly anisotropic, exhibiting, for example, strongly metallic behavior in the basal (*ab*) plane and poor electrical conductivity along the *c* axis (Kelly, 1981). In terms of mechanical properties, graphite is the stiffest material in nature and has the highest in-plane elastic modulus. Of all materials, graphite (in-plane) along with diamond exhibits the highest thermal conductivity (Kelly, 1981). Graphite is the highest–melting-point elemental material (Dresselhaus and Steinbeck, 1988).

Atomic carbon has an atomic number of 6 and is in the $1s^2 2s^2 2p^2$ electronic ground-state configuration, indicating that two of the six electrons are very tightly bound core electrons, and four are valence electrons. In the graphite structure, strong in-plane trigonal bonds are formed between a carbon atom and its three nearest neighbors from $2s$, $2p_x$, and $2p_y$ electronic orbitals and are called σ bonds; this bonding arrangement utilizes three of the four valence electrons and is denoted by sp^2. The remaining valence electron with a p_z orbital, called the π bond, provides only weak interplanar van der Waals bonding, but is responsible for the semimetallic electronic behavior in graphite. In contrast, the carbon

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

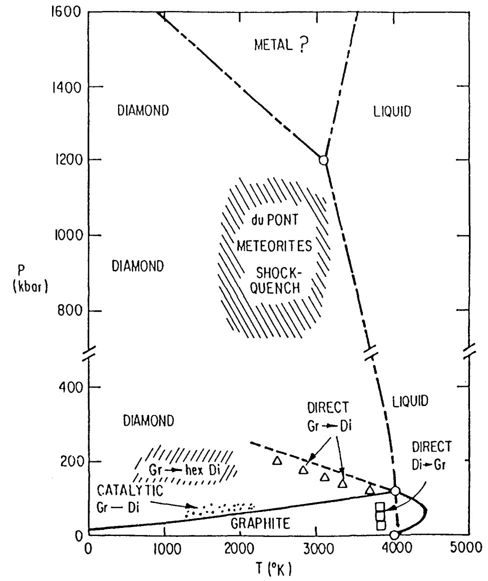

FIG. 1. The phase diagram of carbon suggested by Bundy. Under ambient conditions, graphite is the stable phase (Bundy, 1985).

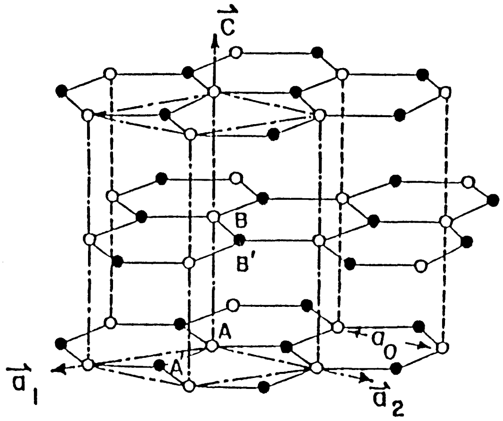

FIG. 2. The crystal structure of hexagonal single-crystal graphite. Planes of carbon hexagons are stacked in an ...*ABAB*... sequence (Wyckoff, 1964). The *A* and *B* carbon sites are denoted by open circles and the *A'* and *B'* sites by black circles, so that the *A* and *B* atoms are directly above one another, but the *A'* and *B'* atoms are not, as shown in the diagram. The in-plane lattice constant is denoted by a_0, and the vectors of the unit cell in the directions $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{c}$ are indicated (Wyckoff, 1964).

atoms in the diamond structure are tetrahedrally bonded to their four nearest neighbors using all four valence electrons in $2s$, $2p_x$, $2p_y$, and $2p_z$ orbitals in an electronic configuration denoted by sp^3. The difference in the structural arrangement of these allotropic forms gives rise to wide differences in their physical properties.

1. STRUCTURE

The ideal crystal structure of graphite (Fig. 2) consists of layers in which the carbon atoms are arranged in an open hexagonal network. Figure 2 shows two inequivalent layers with *A* and *A'* atoms on one of these layers and *B* and *B'* atoms on the other layer such that the *A* and *B* atoms are on top of one another, but the *A'* atom in one plane is over the unoccupied center of the adjacent layers, and similarly for the *B'* atoms (Wyckoff, 1964). This gives rise to an *ABAB* stacking arrangement of the layer planes as shown in Fig. 2, with an in-plane nearest-neighbor distance of 1.421 Å, an in-plane lattice constant of 2.462 Å, a *c*-axis lattice constant of 6.708 Å, and an interplanar distance of 3.3539 Å (Table 1). This structure is described by the D_{6h}^4 ($P6_3/mmc$) space group, which has four atoms per unit cell (see Fig. 2).

When graphite is deformed, a second layer stacking sequence, the *ABCABC* rhombohedral sequence, is introduced (Kelly, 1981). Graphite crystals extracted from mineral deposits or grown synthetically exhibit predominantly *ABAB* stacking shown in Fig. 2 with some *ABC* stacking faults. Such stacking faults tend to have little effect on the in-plane lattice constant, largely because the in-plane C—C bond is very strong and the nearest-neighbor C—C spacing is very small; stacking faults do, however, have a significant effect on the interplanar spacing because of the weak interplanar bonding. One consequence of this is that impurity species are unlikely to enter the in-plane lattice sites substitutionally, but rather will occupy some interstitial position between the layer planes.

Structural defects in the graphite lattice give rise to an increase in the interlayer distance, but essentially no change to the in-plane lattice structure. As the interlayer distance increases, the stacking order of the layers decreases until a value of about 3.44 Å is reached, at which distance the stacking of the individual carbon layers (called graphene layers) becomes uncorrelated; the resulting

Table 1. Properties of graphite.

Properties	Values[a]
Lattice structure	Hexagonal
Space group	$P6_3/mmc\text{–}D_{6h}^4$
Atom locations in unit cell	(000), $(00\frac{1}{2})$, $(\frac{1}{3}\frac{2}{3}0)$, $(\frac{1}{2}\frac{2}{3}\frac{1}{2})$
Lattice constant (RT) (Å)	2.462, 6.708
Atomic density (cm^{-3})	1.14×10^{23}
Specific gravity (g/cm^3)	2.26
Band overlap (eV)	0.04
Electron mobility (RT) [$cm^2/(V\ s)$]	20 000, 100
Hole mobility (RT) [$cm^2/(V\ s)$]	15 000, 90
Carrier density (cm^{-3})	6×10^{18}
Electrical resistivity (RT) (Ω cm)	4.0×10^{-5}, 0.12
Dielectric constant (RT, low frequency)	3.0, 5.0
Melting point (K)	4200
Thermal conductivity (RT) (W/cm K)	30, 0.06
Debye temperature (K)	2500, 950
Thermal expansion coefficient at RT (K^{-1})	-1×10^{-6}, $+29\times10^{6}$
Velocity of sound (cm/s)	$\sim2.63\times10^5$, $\sim1\times10^5$
Raman shift (cm^{-1})	42, 1582

[a] Anisotropic quantities are given first as the in-plane (*a*-axis) value, then the *c*-axis value, where the *a*-axis value refers to the value in the *ab* plane. Room-temperature values of the parameters are denoted by RT.

two-dimensional (2D) structure of uncorrelated graphene layers is called turbostratic graphite (Dresselhaus *et al.*, 1988). As is discussed below, the electronic structure of 2D turbostratic graphite is that of a zero-gap semiconductor, and is qualitatively different from that of ideal 3D graphite, a semimetal with a small band overlap of 0.04 eV between the lowest-energy occupied conduction-band states and the highest energy of the unoccupied valence-band states. From a structural standpoint, a graphite sample is characterized by the magnitudes of its interlayer separation and by the interlayer and intralayer crystallite lengths L_c and L_a, all of which can be obtained from the x-ray diffraction line positions and linewidths. The microstructure of graphite is typically determined by polarized-light microscopy, SEM micrographs, and high-resolution TEM lattice fringe patterns.

Graphite single crystals tend to be very small in size (on a millimeter scale) and very thin (on a 0.1-mm scale), and for this reason are not useful for many scientific studies and practical applications which require larger samples. For this reason a variety of graphite-related materials are used for both scientific studies and practical applications. A very brief review of these materials is given below.

2. GRAPHITE-RELATED MATERIALS

The comparison material in all graphite studies and applications is the natural single-crystal graphite flakes that are found in several locations around the world, especially in Madagascar, the former USSR, and the Ticonderoga area of New York State in the United States. The natural crystal flakes can sometimes be as large as several mm in the basal plane, and are typically much less than 0.1 mm in thickness. Natural graphite flakes usually contain several twinning planes, so that careful selection of the flakes is necessary for detailed structural studies. The natural flakes also contain impurities which are chemically removed in a purification step by boiling the flakes in concentrated HF and/or heating the flakes (to ~2000 °C) in flowing fluorine gas. Even after such chemical purification, metallic impurities (especially transition metals such as Fe) remain in the material at the ppm level (Spain, 1981).

2.1 Kish Graphite

A synthetic single-crystal graphite called kish graphite is commonly used in scientific investigations. Kish-graphite crystals form on the surface of high–carbon-content iron melts, and are harvested as crystals from such solutions (Austerman, 1968). The as-grown kish-graphite flakes are subsequently purified in much the same way as natural-graphite flakes (Spain, 1981). There are several reasons for using kish graphite in scientific studies. First, some kish-graphite flakes can be found with relatively large sizes in the basal plane, relative to natural-graphite flakes. Second, very small kish-graphite flakes ($\leqslant 1$ mm in size) without twin boundaries and with a relatively low density of defects can be found. Thus kish graphite is often used as a material of choice for sophisticated detailed structural studies. Third, the structural perfection of selected kish-graphite flakes can be very high relative to selected natural single-crystal graphite flakes.

2.2 Highly Oriented Pyrolytic Graphite

Many important experimental studies of graphite requiring larger size crystals have been carried out on highly oriented pyrolytic graphite (HOPG), which is prepared by the pyrolysis of hydrocarbons at temperatures above 2000 °C and is subsequently heat treated to higher temperatures (Übbelohde and Lewis, 1960; Moore, 1973; Moore, 1981). When stress annealed above 3300 °C, the HOPG exhibits electronic, transport, thermal, and mechanical properties close to those of single-crystal graphite. For the high-temperature, stress-annealed HOPG, the crystalline order extends to about 1 μm within the basal plane, and to about 0.1 μm along the *c* direction. HOPG has short-range in-plane order, with an in-plane crystallite size L_a up to ~ 1 μm depending on the heat-treatment temperature. HOPG, however, has no long-range in-plane *a*-axis alignment, and the *a* axes of adjacent crystallites are randomly arranged (Moore, 1973, 1981). In contrast, HOPG shows a high degree of *c*-axis alignment, with misalignment angles of a few tenths of a degree between adjacent crystallites.

The degree of in-plane structural order (usually expressed in terms of the magnitude of L_a) and *c*-axis alignment (expressed in terms of the magnitude of L_c and the half-width of the *c*-axis misalignment distribution) can be varied by control of the major processing parameters: heat-treatment temperature T_{HT}, residence time at T_{HT}, and applied stress during heat treatment. Turbostratic pyrolytic graphites are obtained for $T_{\mathrm{HT}} < 2300$ °C, and yet higher T_{HT} values are needed to establish 3D ordering. HOPG has been the material of choice for many of the fundamental studies of the structure and properties of graphite, as well as for industrial applications in x-ray monochromators and widespread use in neutron monochromators (Moore, 1973, 1981).

2.3 Graphitic Films

Recently, new precursor materials, such as polyimide (PI) and polyoxydiazole (POD) resins, have been used for the synthesis of graphite films. When heat treated to $T_{\mathrm{HT}} > 2800$ °C (Hishiyama *et al.*, 1991), these films show a high degree of 3D structural ordering. The structural perfection of these films, especially those based on the Kapton and Novax (polyimide) precursors, is rapidly improving and these films may soon become materials of choice for selected scientific studies or practical applications where high-quality thin films are needed. One potential application of these graphite films may be as a high–thermal-conductivity substrate which is electrically conducting, in contrast to diamond which has an equally high thermal conductivity but is electrically insulating.

2.4 Carbon Fibers

Carbon fibers represent another important class of graphite-related materials. Despite the many precursors that can be used to synthesize carbon fibers, each having different cross-sectional morphologies (Fig. 3), the preferred orientation of the fiber axes for all carbon fibers is close to an *a* axis in a graphene layer, thereby accounting for the high mechanical strength and stiffness of carbon fibers (Dresselhaus *et al.*, 1988). Referring to the various morphologies shown in Fig. 3 for carbon fibers, the as-prepared vapor-grown carbon fibers have an onion-skin morphology [Fig. 3(a)], which after heat treatment to about 3000 °C forms facets [Fig. 3(b)]; of all carbon fibers, these faceted fibers are closest to crystalline graphite in both crystal structure and physical properties. The commercially available mesophase pitch fibers with either the radial morphology [Fig. 3(c)] or the "Pan-Am" morphology [Fig. 3(d)] are exploited for their extremely high bulk modulus, while the com-

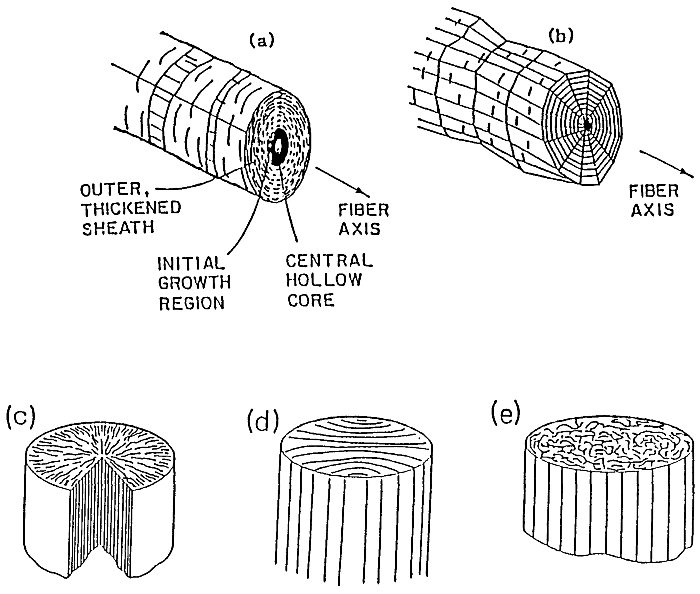

FIG. 3. Sketch illustrating the morphology of vapor-grown carbon fibers (VGCF): (a) as-deposited at 1100 °C, (b) after heat treatment to 3000 °C. The morphologies for commercial mesophase-pitch fibers are shown (c) for a "PAC-man" section with a radial arrangement of the straight graphene ribbons, (d) for a "Pan-Am" section showing a transverse alignment with nearly parallel graphene planes, and (e) for a PAN fiber, with a preferential circumferential arrangement of graphene ribbons in the sheath region and a random arrangement of ribbons in the core (Dresselhaus *et al.*, 1988).

mercial PAN (polyacrylonitrile) fibers [Fig. 3(e)] with circumferential texture are widely used for their high strength (Dresselhaus *et al.*, 1988). The high bulk modulus of carbon fibers is associated with the high degree of stiffness of graphite, and is an intrinsic property of crystalline graphite. The high tensile strength of carbon fibers is associated with the special defects in the microstructure which inhibit crack propagation. Typical fiber diameters are ~10 μm. The various morphologies shown in Fig. 3 exploit the highly anisotropic layered structure that is characteristic of graphite.

Scientific applications of carbon fibers for electromagnetic studies and electrical transport measurements typically exploit the large length-to-diameter ratio of carbon fibers. Practical applications of carbon fibers generally depend on their exceptional mechanical stiffness and strength. Large-scale applications of high-performance fibers are made in aircraft and spacecraft structures, and smaller-scale applications are made in sporting good equipment such as tennis racquets, racing bicycles, and pole vaulting poles. Many of the applications of carbon fibers exploit their high tensile strength, stiffness, and large strength-to-weight ratio. In many of these applications the fibers are assembled along with other materials to form composites. For high-temperature applications carbon fibers are often assembled with carbon-based host materials in so-called carbon-carbon composites. Lower-performance, lower-cost carbon fibers have been successfully used as reinforcement for concrete in building construction.

When activated at elevated temperatures (800–900 °C) in the presence of water vapor, the resulting activated carbon fiber can have enormous specific surface areas (SSA). Values of SSA up to 3000 m^2/g have been reported. Such fibers are used both for gas adsorption applications and for electrical applications as double-layer capacitors (Dresselhaus *et al.*, 1992).

2.5 Glassy Carbon

Glassy carbon (GC), another graphite-related carbon material, is manufactured as a commercial product by slow, controlled degradation of certain polymers at temperatures typically on the order of 900–1000 °C (Jenkins and Kawamura, 1976). This material is glass-like, granular, and moderately hard, can be easily polished, and is thermally conducting, impermeable, biocompatible, and stable at high temperatures (see CARBON MATERIALS; DIAMOND AND DIAMONDLIKE CARBON). The apparent density of GC is in the range 1.46–1.50 g/cm^3 irrespective of heat-treatment temperature, implying the existence of pores in the matrix. According to the Jenkins–Kawamura model, the microstructure of GC consists of a tangle of graphitelike ribbons or microfibrils, about 100 Å long and 30 Å in cross section, and resembles the polymer chain configuration from which the GC was derived. Because of the tangled ribbon microstructure, Jenkins and Kawamura have argued that glassy carbon does not fully crystallize into a three-dimensional graphite structure, even when heat treated above 3000 °C. For this reason glassy carbon has been used as a prototype hard (nongraphitizable) carbon for scientific investigations. X-ray diffraction studies of the radial distribution function show that the carbon atoms are well ordered in the honeycomb in-plane structure of the graphene layers, but that the 3D registry between the graphene layers is poor, so that the graphene ribbons form a turbostratic structure, typical of hard carbons (Oberlin, 1989). Recent structural studies have shown that the pores in glassy carbon form closed structures (Fig. 4) rather than an open interconnected network of ribbons (Yoshida *et al.*, 1991).

2.6 Graphite Intercalation Compounds

Graphite intercalation compounds (GICs) are formed by the insertion of layers of guest species between the layers of the graphite host material (Ubbelohde and Lewis, 1960; Dresselhaus and Dresselhaus, 1981; Zabel and Solin, 1990) (see INTERCALATION COMPOUNDS). The guest species between the graphite layers may be either atomic or molecular, and may act as either a donor or an acceptor of electrons to the surrounding graphite layers. Because of the weak interlayer forces associated with the sp^2 bonding in graphite, the intercalated species can easily be inserted between the graphite layers to form highly anisotropic layered graphite-based GICs. In diamond, on the other hand, the strong sp^3 bonding does not permit insertion of layers of guest species, and does not support intercalation. One practical application of graphite intercalation compounds is in the preparation of exfoliated graphite or Grafoil. This high–surface-area material is produced by heating selected acceptor GICs to introduce tiny gas bubbles between the layers, which, upon heating, expand rapidly, separating many of the graphite layers and greatly increasing the volume occupied by the material. Exfoliated graphite is used as an adsorption agent and as a packing material for valves.

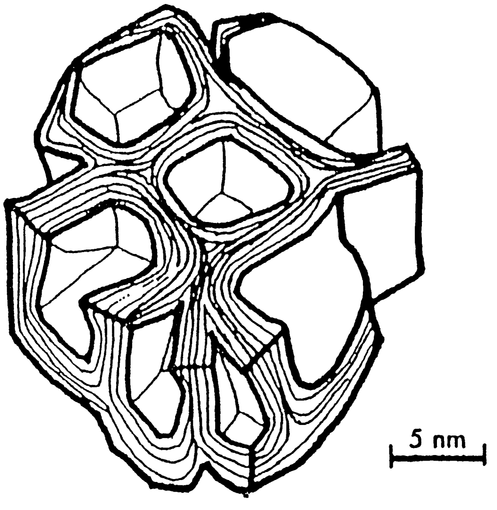

FIG. 4. Schematic diagram for the microstructure of the closed pore structure model for glassy carbon and other hard carbons (Yoshida *et al.*, 1991).

3. PROPERTIES OF GRAPHITE

The highly anisotropic structure of graphite gives rise to lattice, transport, optical, thermal, and mechanical properties that are likewise highly anisotropic. These are briefly reviewed below.

3.1 Lattice Properties

Because of the strong in-plane and weak interplanar bonding in graphite, the force constants controlling the atomic vibrations, the dispersion relations for the lattice vibrations, and the velocity of sound are all highly anisotropic (Table 1). Since there are four atoms/unit cell, there are $4 \times 3 = 12$ degrees of freedom giving rise to nine optical and three acoustic branches for the lattice vibrations (phonons). Of these, there are two infrared-active zone-center phonons with E_{1u} (1588 cm^{-1}) and A_{2u} (868 cm^{-1}) symmetries for the in-plane and out-of-plane vibrational modes, and two Raman-active in-plane vibrational modes denoted by E_{2g_1} (45 cm^{-1}) and E_{2g_2} (1582 cm^{-1}) (Dresselhaus and Dresselhaus, 1982). The atomic displacements associated with these normal-mode vibrations in graphite are shown in Fig. 5(a). Of particular importance to the characterization of graphite materials is the E_{2g_2} vibrational mode. In the case of 2D turbostratic graphite, the Raman-active vibrational mode at 1582 cm^{-1} persists.

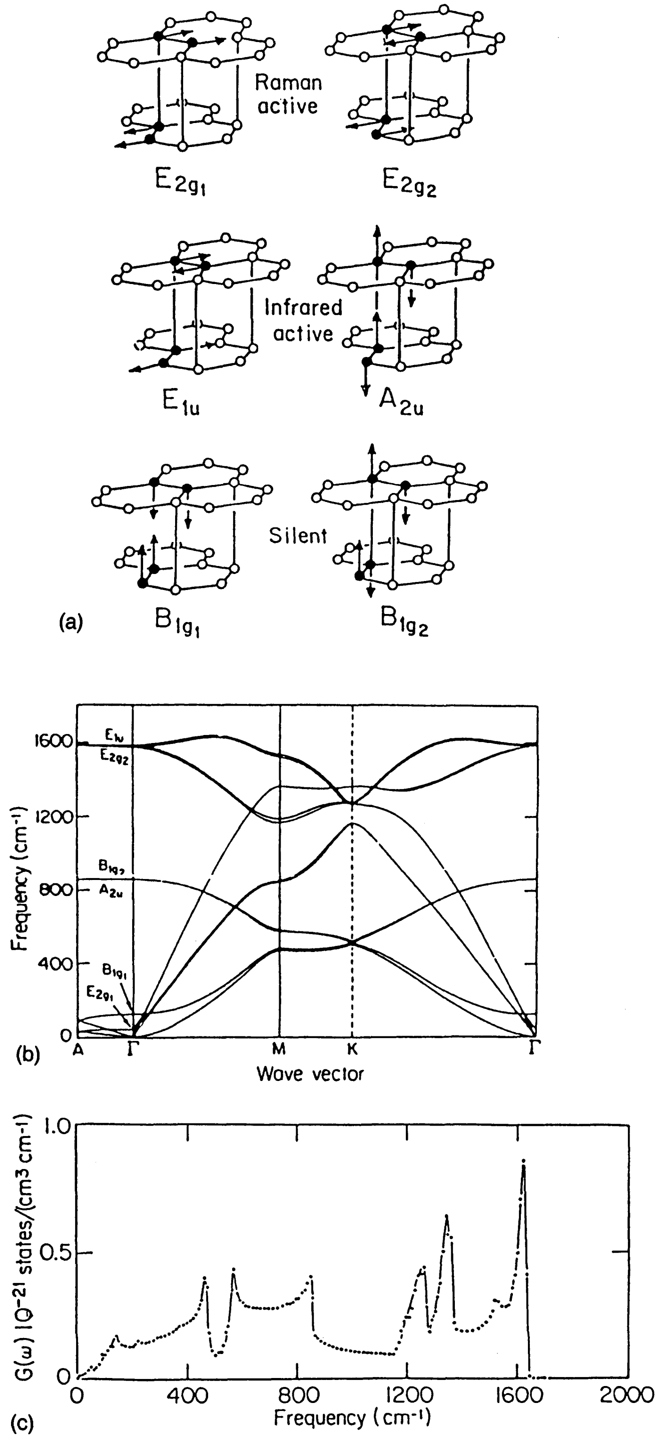

45p

FIG. 5. (a) Zone-center optical vibrational modes for carbon atoms in the graphite structure. For the in-plane vibrational modes (E_{1u}, E_{2g_1}, and E_{2g_2}), only one of the degenerate pairs of modes is shown. The *c*-axis modes (A_{2u}, B_{1g_1}, and B_{1g_2}) are nondegenerate. The zero-frequency acoustic modes (E_{1u} and A_{2u}) correspond to pure translations and are not shown. (b) Phonon dispersion curves for the lattice vibrations in graphite along certain high-symmetry axes in the hexagonal Brillouin zone. (c) Phonon density of states corresponding to the phonon dispersion curves for graphite. The phonon frequencies are expressed in cm^{-1} (Dresselhaus and Dresselhaus, 1982).

In addition, strong features in the Raman spectra are observed in disordered graphites at the maximum frequency in the graphite phonon dispersion relations at 1630 cm^{-1} [Fig. 5(b)] and the maximum in the phonon density of states near 1360 cm^{-1} [Fig. 5(c)] (Dresselhaus and Dresselhaus, 1982). Crystalline disorder breaks down the selection rules for Raman scattering, allowing contributions to the scattering intensity from non-zone-center phonons. Thus the integrated intensity of the disorder-induced lines at $\sim$1360 cm^{-1} and $\sim$1630 cm^{-1} relative to that for the Raman-allowed 1582-cm^{-1} line allows for characterization of disordered graphites in terms of their average in-plane crystalline size (L_a) (Knight and White, 1989). Information on the phonon dispersion relations $\omega(q)$ has been provided by inelastic neutron scattering measurements and a variety of other techniques, and this information has been used to obtain $\omega(q)$ for the 12 phonon branches and the phonon density of states shown in Fig. 5.

3.2 Electronic and Transport Properties

The electronic and transport properties of graphite have been extensively studied and reviewed (McClure, 1971; Spain, 1973; Spain, 1981). These properties are dominated by the strong in-plane sp^2 or σ bonding and by the large anisotropy of the graphite crystal structure. Corresponding to the strong sp^2 in-plane bonding of a single graphene sheet, three σ bonds are formed far below ($\sim$10 eV) the Fermi level E_F, and three antibonding σ bonds far above E_F. Thus the transport properties of 2D graphite are largely determined by the two π bands (one bonding and one antibonding) which lie close to E_F. By symmetry, the two π bands formed by the p_z orbitals are degenerate at the 2D hexagonal Brillouin zone corners, and the Fermi level E_F passes through this degenerate point in reciprocal space, giving rise to a zero-band-gap semiconductor. This zero-band-gap semiconductor model is appropriate for the description of turbostratic graphite, which has a well-developed crystalline structure within the layer planes but no correlation between the atomic positions of atoms on adjacent layer planes.

In contrast, perfectly ordered crystalline graphite has a weak interlayer interaction giving rise to semimetallic behavior with a band overlap between the conduction-band minimum energy and the valence-band maximum energy of about 40 meV. The band structure near the Fermi level is described by the Slonczewski–Weiss–McClure energy-band model for the four π bands in the 3D Brillouin zone in reciprocal space in terms of seven band parameters (McClure, 1971). These band parameters are identified with overlap integrals and transfer integrals arising from the tight-binding Fourier expansion of the electronic energy bands along k_z and a $\mathbf{k}\cdot\mathbf{p}$ treatment of the energy bands in the k_x, k_y directions away from the Brillouin-zone edges. Figure 6 shows the 3D hexagonal Brillouin zone and the location of the various high-symmetry points. The band-structure model near E_F focuses on the electronic dispersion relations in the vicinity of the *HKH* and *H′K′H′* edges of the Brillouin zone. In two-dimensional or turbostratic graphite the Brillouin zone of Fig. 6 collapses into a single hexagonal sheet and the *HKH* axis collapses into a single point at each corner of the hexagon where the π bands are degenerate as discussed above.

The Fermi surface for semimetallic 3D graphite consists of both holes and electrons, and these carriers are located in carrier pockets along the Brillouin zone edges as shown schematically in Fig. 6. Because the carrier pockets occupy only a small volume of the Brillouin zone, as indicated in the figure, graphite has a low concentration of holes (p) and electrons (n), with $n=p=3\times10^{18}$ cm^{-3} in the low-temperature limit of an ideal graphite crystal (McClure, 1971). The highly anisotropic Fermi surfaces (see Fig. 6) give rise to small effective masses for both electron and hole motion in the basal plane, but to very large effective masses for motion along the c axis. The high in-plane mobility of the elec-

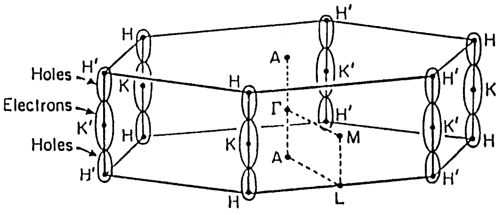

FIG. 6. Brillouin zone for three-dimensional graphite showing several high-symmetry points and a schematic version of the graphite electron and hole Fermi surfaces located along the *HKH* and *H′K′H′* axes of the Brillouin zone (Dresselhaus and Dresselhaus, 1981).

trons and holes in graphite allow accurate measurements of the Fermi surface, using the de Haas–van Alphen, Shubnikov–de Haas, and other magneto-oscillatory Fermi surface measurements (McClure, 1971).

The anisotropic Fermi surface is responsible for the high-mobility carriers [$\sim 1.3\times 10^4$ $cm^2/(V\,s)$] that dominate the in-plane transport properties and for the low-mobility carriers that dominate the electronic transport along the *c* axis. Anisotropy ratios of the in-plane to *c*-axis conductivity $\sigma_a/\sigma_c \sim 10^5$ are characteristic of low-temperature electronic transport in graphite (Spain, 1973). Since the Fermi level is only 24 meV above the *K*-point band extremum for electrons (see Fig. 6), there is significant thermal excitation of carriers at room temperature. Thus the temperature dependence of the in-plane electrical conductivity [$\sigma_a(T)$] for graphite is anomalous in comparison to that for conventional metallic conductors and is quite different from that along the *c* axis [$\sigma_c(T)$] (Spain, 1973). The temperature dependence of the electrical conductivity of single-crystal graphite in comparison with that for vapor-grown carbon fibers heat treated at various temperatures is shown in Fig. 7. At temperatures above $\sim$100 K lattice vibrations are the dominant scattering mechanism for electronic transport, while at low temperatures, scattering of electrons by crystalline defects is dominant. Thus the residual resistivity ratio (RRR), defined as the ratio between the room-temperature (T=300 K) and low-temperature (T =4.2 K) electrical resistivities ($\mathrm{RRR}=\rho_{300\,\mathrm{K}}/\rho_{4.2\,\mathrm{K}}$), is often used to characterize a conducting material for its crystalline defects. In Fig. 7 we see that single-crystal graphite can have residual resistivity ratios RRR of as

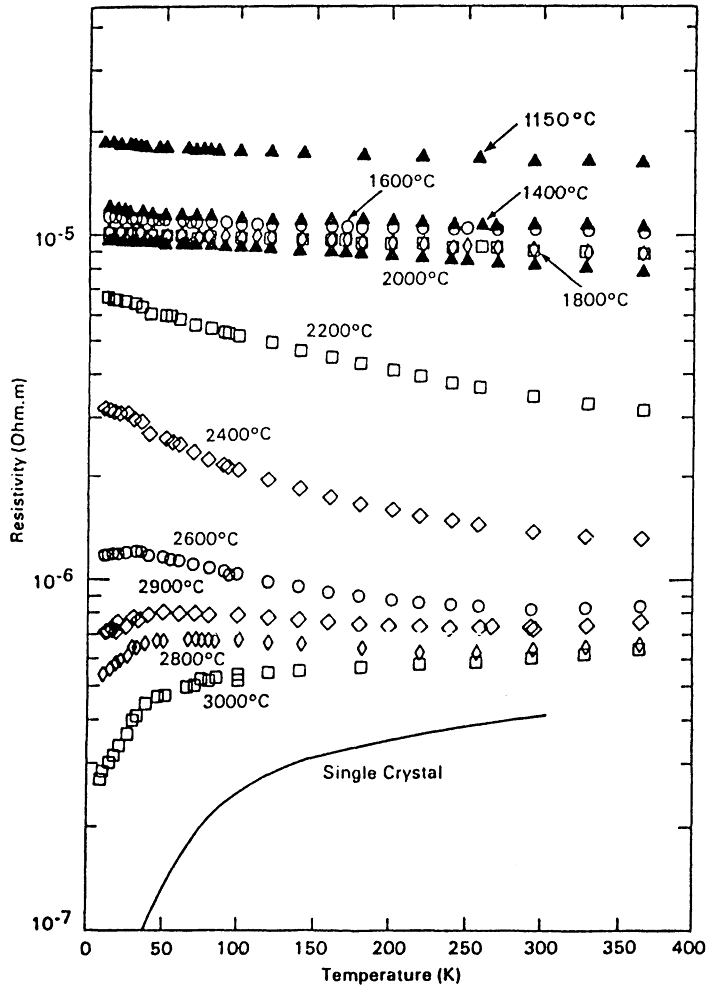

FIG. 7. The temperature dependence of the resistivity of single-crystal graphite and for benzene-derived vapor-grown graphite fibers heat treated to the various indicated temperatures (Dresselhaus *et al.*, 1988).

much as 10^2. Disordered carbons typically show RRR values approaching unity (see Fig. 7) and reduced anisotropy in $\sigma(T)$ as well. The special geometry of carbon fibers makes them attractive materials for transport studies. Vapor-grown carbon fibers when heat treated to high temperatures (e.g., 3000 °C) exhibit transport properties quite similar to graphite (see Fig. 7). However, in their as-prepared state ($T_{HT} \simeq 1150$ °C), their room-temperature resistivity is about 40 times higher and the resistivity is nearly temperature independent, since for these fibers the defect scattering of the electronic carriers is much larger than for crystalline graphite and dominates over the scattering by lattice vibrations over the entire temperature range of measurement in Fig. 7.

The introduction of crystalline defects tends to reduce the effective band overlap (thereby reducing the carrier density). Crystalline defects also reduce the carrier mobility, leading to a significant increase in resistivity ρ_a relative to its value ($\rho_a = 41\ \mu\Omega$ cm) in crystalline graphite (Spain, 1981). As the defect density increases, localized states are formed at the respective band edges for electrons and holes, and the conductivity drops in accordance with presently accepted models for weak localization (see ELECTRON STATES, LOCALIZED) and carrier-carrier interaction effects. Eventually, when a high degree of crystal disorder is reached, the Fermi level drops below the mobility edge (for holes), giving rise to a conduction process called variable-range hopping conduction across an effective mobility gap, typical of disordered semiconductors. Glassy carbon is an example of such a disordered material where the conductivity has dropped by a few orders of magnitude relative to that of crystalline graphite. A typical range of resistivities would be 4.5×10^{-3} to 4×10^{-4} Ω cm for GC heat treated to $T_{HT} \sim 2500$ °C. The temperature dependence of the conductivity of glassy carbon, which is due to carrier hopping, exhibits an $\exp[-(T_0/T)^{1/4}]$ law, characteristic of 3D variable-range hopping for $T_{HT} < 1000$ °C. As T_{HT} increases, $\sigma(T)$ increases in magnitude and becomes independent of temperature.

The magnetic susceptibility χ for crystalline 3D graphite has a large diamagnetic contribution arising from strong interband coupling terms (McClure, 1971), and this diamagnetic contribution can be used to provide an accurate determination of the nearest-neighbor π-orbital overlap integral γ_0. The introduction of disorder reduces the magnitude of the diamagnetic contribution, and in the 2D limit where adjacent graphene layers become decoupled (i.e., for turbostratic graphite), the diamagnetic contribution provides a measure of the interplanar coupling in graphite.

Electron-spin-resonance (ESR) measurements provide another common characterization tool for graphite and disordered graphite. For highly ordered 3D graphite, the ESR signal is typical of conduction-electron spin resonance in metallic systems with long electron mean free paths. The g factor in graphite is highly anisotropic at room temperature, where $g_\perp = 2.0026$ for $H \perp c$ axis, and $g_\parallel = 2.0495$ for $H \parallel c$ axis. The introduction of disorder reduces $g_\parallel$ sharply toward the free-electron value of 2.0023, thereby reducing the anisotropy of the g factor. For highly disordered graphite, where hopping conductivity predominates, the unpaired spins give rise to a Curie term in the low-temperature magnetic susceptibility, the ESR line becomes Lorentzian, and the g factor becomes isotropic and close to the free-electron value. Thus, ESR is also widely used for the characterization of disordered graphitic materials.

3.3 Optical Properties

Graphite is a relatively good electrical conductor; hence its optical properties are usually investigated by optical reflectivity rather than optical absorption measurements. Because of the low carrier density of graphite, the free carriers do not give rise to a clear plasma response in the reflectivity spectrum. Since the contributions to the dielectric function $\epsilon(\omega) = \epsilon_1(\omega) + i\epsilon_2(\omega)$ from the free carriers, interband transitions, and lattice vibrations all occur over the same frequency range, the frequency dependence for each of these three contributions to $\epsilon(\omega)$ must therefore be extracted from the measured reflectivity spectra through a line-shape analysis and a detailed curve-fitting procedure (Dresselhaus *et al.*, 1988).

Because of the large anisotropy of the electrical conductivity of graphite, the optical skin depth for the polarization **E** parallel to the c axis is an order of magnitude greater than for the polarization **E** perpendicular to the c axis. The resulting anisotropy of the free-carrier contribution to $\epsilon(\omega)$ together with major differences in the interband transitions for the two polarizations gives rise to highly anisotropic optical reflectivity spectra.

In a magnetic field, the optical spectra show well-defined structure associated with optical transitions between magnetic energy levels. These spectra have been extensively studied in crystalline graphite. These magnetoreflection spectra together with optical spectra in zero magnetic field and Fermi-surface measurements have all been used to determine the band parameters of the Slonczewski–Weiss–McClure model which describes the electronic structure of graphite within a few eV of the Fermi level (McClure, 1971).

3.4 Thermal Properties

The c-axis thermal conductivity is about 300 times smaller than in plane, giving rise to a material with extremely high anisotropy in the thermal conductivity; the highest thermal conductivities reported in the literature for graphite are given in Table 1. The thermal-conductivity and thermal-expansion coefficients of graphite are both highly anisotropic (Kelly, 1981; Dresselhaus *et al.*, 1988), as described below. These thermal properties of graphite (see Table 1) have been extensively studied and reviewed (Kelly, 1981). The in-plane thermal conductivity κ_a of graphite is the highest (together with that of diamond) of all materials in nature. Typically metallic materials exhibit contributions from both electrons and lattice vibrations to their thermal conductivity. Nevertheless, the thermal conductivity of graphite is dominated by lattice contributions because of the low carrier concentration of crystalline graphite. The very high in-plane thermal conductivity κ_a is reduced by defect scattering, and for this reason the best thermal conductors are highly crystalline. Likewise, the high degree of anisotropy in the thermal conductivity ($\kappa_a/\kappa_c \approx 300$) of graphite is reduced by the introduction of defects. Because of the quasi two-dimensional nature of graphite, the thermal conductivity roughly follows a temperature dependence of $T^{2.3}$ at low temperature, which is close to T^2, the temperature dependence expected for two-dimensional materials. Typically, materials in the solid state exhibit a T^3 temperature dependence of their thermal conductivity, associated with the temperature dependence of the excitation of lattice vibrations in 3D systems.

The thermal expansion coefficient α of graphite is also highly anisotropic, very large and positive along the c axis (α_c) and very small and negative in plane (α_a) at room temperature (Kelly, 1981; Dresselhaus *et al.*, 1988). As a function of temperature, α_c increases rapidly until about 400 K and much more slowly at higher temperatures. The in-plane behavior is, however, very different, with α_a first becoming more negative with increasing T until about 300 K where α_a is at its most negative value; as T is further increased, α_a becomes less negative, and finally becomes positive at about 800 K (Kelly, 1981). These unusual thermal expansion coefficients make it difficult to bond other materials to graphite over a broad temperature range.

The heat capacity of graphite is likewise dominated by the contribution from the lattice vibrations because of the low carrier density. Of particular interest is the temperature dependence of the heat capacity, which follows a T^2 dependence, indicative of the quasi-2D nature of graphite. The heat capacity C_p and the lattice thermal conductivity κ_l are related according to a kinetic model by $\kappa_l = \frac{1}{3}C_p v l$ where v is the velocity of sound and l is the mean free path for the lattice vibrations (Kelly, 1981).

The thermoelectric power for graphite is small and exhibits an unusual temperature dependence because of the importance of phonon-drag effects, which are dominant below ~100 K. These phonon-drag effects are associated with electronic motion resulting from electron carriers that are dragged along with their associated ions as the ions participate in lattice vibrations. The introduction of disorder suppresses phonon-drag effects and increases the magnitude of the thermoelectric power. For turbostratic graphite a linear temperature dependence of the thermopower can be observed (Tsuzuku and Sugihara, 1975).

3.5 Mechanical Properties

Because of its strong in-plane bonding, graphite is a very stiff material with an extremely high Young's modulus (~800 GPa)

Table 2. Elastic constants for single-crystal graphite.

Elastic moduli (GPa)		Elastic compliance (10^{-12} Pa^{-1})	
C_{11}	1060 ± 20	S_{11}	0.98 ± 0.03
C_{12}	180 ± 20	S_{12}	-0.16 ± 0.06
C_{13}	15 ± 5	S_{13}	-0.33 ± 0.08
C_{33}	36.5 ± 1	S_{33}	27.5 ± 1.0
C_{44}	4.5 ± 0.5	S_{44}	240 ± 30

and a very high in-plane tensile strength ($\sim$20 GPa) (see Table 1). These high values for Young's modulus and tensile strength were measured on carbon whiskers (Bacon, 1960) which are formed in a carbon arc discharge and consist of a highly ordered graphene sheet rolled up in the form of a scroll. This high tensile strength is exploited in most of the practical applications of carbon fibers. Because of the weak interplanar forces, graphite has a very low shear strength ($\sim 4.8\times10^5$ Pa) (Kelly, 1981); the compressional strength of graphite is also much lower than its tensile strength. Because of its hexagonal symmetry, graphite is described by stress-strain relations involving five independent elastic constants, as are listed in Table 2. In contrast, a cubic material is described by three independent elastic constants, while a completely isotropic material is described by two independent elastic constants, one relating to tension or compression, and the other relating to shear.

The mechanical properties of graphite are dominated by their high degree of anisotropy and reflect the high crystalline anisotropy. Since many of the graphite-based systems that are exploited for their mechanical properties are highly defective and are used in fibrous forms embedded in various matrix materials, the analysis of the mechanical properties of carbon fiber composites is complex.

ACKNOWLEDGMENTS

The authors wish to thank the National Science Foundation for Grant No. DMR-88-19896, and the Lawrence Livermore National Laboratory for a subcontract, for support for this work.

GLOSSARY

Activated Carbon Fibers: Carbon fibers with very large surface area per unit mass.

Basal Plane: The plane normal to the main symmetry axis, usually called the *c* axis.

Carbon Fiber: Fiber morphology showing graphitic planes preferentially aligned along the fiber axis.

Carbon Whisker: A scroll made with a single graphene plane.

Exfoliated Graphite: A graphitic material in which many layer planes are widely separated, giving rise to a very large–surface-area material.

Glassy Carbon: A highly disordered, weakly conducting form of carbon containing graphitic ribbons, having high porosity and good mechanical properties.

Graphene Layer: An isolated single layer of hexagonally arranged carbon atoms such as found in graphite.

Graphite Intercalation Compound (GIC): Compound formed by the insertion of a layer of guest species between the planes of graphite.

Graphitelike Ribbon: Ribbon formed of long thin strips of graphene layers.

High-Performance Fibers: Carbon fibers specially developed for their desirable mechanical properties.

Highly Oriented Pyrolytic Graphite (HOPG): A synthetic form of graphite with a high degree of *c*-axis alignment, but having in-plane crystallites ($\sim$1 μm in size) that are randomly oriented normal to the *c* axis.

Kish Graphite: A synthetic single-crystal graphite material prepared from precipitation in high-carbon steel.

Nongraphitizable: A form of disordered carbon which is a tangled ball of ribbons that will not graphitize even at high temperature and pressure.

Pyrolytic Graphite: Graphite formed by heating organic compounds.

Rhombohedral Graphite: Graphite with *ABC* layer stacking.

sp^2 Bonding: Three planar bonds derived from one *s* and two *p* atomic carbon orbitals making angles of 120° with each other.

sp^3 Bonding: Four bonds derived from one *s* and three *p* atomic carbon orbitals making tetrahedral angles.

Vapor-Grown Carbon Fibers: Carbon fibers synthesized from hydrocarbons in the gas phase, usually involving a catalyst.

List of Works Cited

Austerman, S. B. (1968), in: P. L. Walker, Jr. (Ed.), *Chemistry and Physics of Carbon,* Vol. 4, New York: Dekker, pp. 137–183.

Bacon, R. (1960), *J. Appl. Phys.* **31**, 283–290.

Bundy, F. P. (1985), in: S. Minomura (Ed.), *Solid State Physics Under Pressure: Recent Advance with Anvil Devices,* Dordrecht: Reidel, pp. 1–10.

Dresselhaus, M. S., Dresselhaus, G. (1981), *Adv. Phys.* **30**, 139–326.

Dresselhaus, M. S., Dresselhaus, G. (1982), in: M. Cardona, G. Güntherodt (Eds.), *Light Scattering in Solids III,* Topics in Applied Physics, Vol. 51, Berlin: Springer, pp. 3–57.

Dresselhaus, M. S., Dresselhaus, G., Sugihara, K., Spain, I. L., Goldberg, H. A. (1988), *Graphite Fibers and Filaments,* Springer Series in Materials Science, Vol. 5, Berlin: Springer.

Dresselhaus, M. S., Fung, A. W. P., Rao, A. M., di Vittorio, S. L., Kuriyama, K., Dresselhaus, G., Endo, M. (1992), *Carbon* **30**, 1065–1073.

Dresselhaus, M. S., Steinbeck, J. (1988), *Tanso* **132**, 44–56.

Hishiyama, Y., Kaburagi, Y., Inagaki, M. (1991), in: P. A. Thrower (Ed.), *Chemistry and Physics of Carbon,* Vol. 23, New York: Dekker, pp. 1–68.

Jenkins, G. M., Kawamura, K. (1976), in: *Polymeric Carbons—Carbon Fibre, Glass and Char,* Cambridge, England: Cambridge University Press.

Kelly, B. T. (1981), *Physics of Graphite,* London: Applied Science.

Knight, D. S., White, W. B. (1989), *J. Mater. Res.* **4**, 385–393.

McClure, J. W. (1971), in: D. L. Carter, R. T. Bate (Eds.), *Proceedings of the International Conference on Semimetals and Narrow Gap Semiconductors,* New York: Pergamon, pp. 127–137.

Moore, A. W. (1973), in: P. L. Walker, Jr., P. A. Thrower, (Eds.), *Chemistry and Physics of Carbon,* Vol. 11, New York: Dekker, pp. 69–187.

Moore, A. W. (1981), in: P. L. Walker, Jr., P. A. Thrower, (Eds.), *Chemistry and Physics of Carbon,* Vol. 17, New York: Dekker, pp. 233–286.

Oberlin, A. (1989), in: P. A. Thrower (Ed.), *Chemistry and Physics of Carbon,* Vol. 22, New York: Dekker, pp. 1–143.

Spain, I. L. (1973), in: P. L. Walker, Jr., P. A. Thrower (Eds.), *Chemistry and Physics of Carbon,* Vol. 8, New York: Dekker, pp. 1–150.

Spain, I. L. (1981), in: P. L. Walker, Jr., P. A. Thrower (Eds.), *Chemistry and Physics of Carbon,* Vol. 16, New York: Dekker, pp. 119–304.

Tsuzuku, T., Sugihara, K. (1975), in: P. L. Walker, Jr., P. A. Thrower (Eds.), *Chemistry and Physics of Carbon,* Vol. 12, New York: Dekker, pp. 109–170.

Ubbelohde, A. R., Lewis, F. A. (1960), *Graphite and its Crystal Compounds,* Oxford: Clarendon.

Wyckoff, R. W. G. (1964), in: *Crystal Structures,* Vol. 1, New York: Interscience.

Yoshida, A., Kaburagi, Y., Hishiyama, Y. (1991), *Carbon* **29**, 1107–1111.

Zabel, H., Solin, S. A. (Eds.) (1990), *Graphite Intercalation Compounds I and II: Structure and Dynamics,* Springer Series in Materials Science, Vols. 14 and 18, Berlin: Springer.

Further Reading

Dresselhaus, M. S., Dresselhaus, G. (1981), *Adv. Phys.* **30**, 139–326.

Dresselhaus, M. S., Dresselhaus, G., Sugihara, K., Spain, I. L., Goldberg, H. A. (1988), *Graphite Fibers and Filaments,* Springer Series in Materials Science, Vol. 5, Berlin: Springer.

Kelly, B. T. (1981), *Physics of Graphite,* London: Applied Science.

Moore, A. W. (1981), in: P. L. Walker, Jr., P. A. Thrower (Eds.), *Chemistry and Physics of Carbon,* Vol. 17, New York: Dekker, pp. 233–286.

Reynolds, W. N. (1968), *Physical Properties of Graphite,* New York: Elsevier.

Ruland, W. (1968), in: P. L. Walker, Jr. (Ed.), *Chemistry and Physics of Carbon,* Vol. 4, New York: Dekker, pp. 1–84.

Spain I. L. (1981), in: P. L. Walker, Jr., P. A. Thrower (Eds.), *Chemistry and Physics of Carbon,* Vol. 16, New York: Dekker, pp. 119–304.

Ubbelohde, A. R., Lewis, F. A. (1960), *Graphite and its Crystal Compounds,* Oxford: Clarendon.

GRAVITATION AND GENERAL RELATIVITY

BERNARD F. SCHUTZ, *Department of Physics and Astronomy, University of Wales College of Cardiff, Cardiff, Wales, United Kingdom*
CLIFFORD M. WILL, *Department of Physics, Washington University, St. Louis, Missouri, U.S.A.*

Introduction 304
1. **General Relativity and Relativistic Gravity** 305
1.1 Gravitation Theory: An Overview 305
1.1.1 Relativistic Gravity in Physics 305
1.1.1.1 Roots in Newtonian Gravity 305
1.1.1.2 The Impact of Special Relativity: Relativistic Gravity Is Required 306
1.1.1.3 The Impact of Quantum Mechanics: Quantum Gravity Is Required 306
1.1.2 Where Relativistic Gravity Is Important 307
1.1.2.1 Compact Objects of High Density 307
1.1.2.2 Cosmology: Relativistic Gravity at Low Densities 307
1.1.2.3 Small Relativistic Effects: Gravitational Waves and Equations of Motion 307
1.2 Fundamental Ideas and New Concepts of General Relativity 308
1.2.1 The Incorporation of Newtonian Gravity 308
1.2.1.1 The Equivalence Principle: How Gravity Affects Matter 308
1.2.1.2 Newton's Field Equation: How Matter Creates Gravity 308
1.2.1.3 Galileo's Principle of Relativity 308
1.2.2 Basic Ingredients of General Relativity 308
1.2.2.1 Special Relativity 308
1.2.2.2 The Einstein Equivalence Principle 310
1.2.2.3 Tidal Forces Are the Real Gravity 310
1.2.2.4 Gravitational Redshift 310
1.2.2.5 Gravitational Deflection of Light 311
1.2.3 General Relativity: Gravity as Geometry 312
1.2.3.1 Why Gravity Is Geometrical 312
1.2.3.2 The Metric of Gravity 312
1.2.3.3 Local Flatness of Spacetime and the Local Freely Falling Frames 313
1.2.3.4 Curvature and Tidal Forces 313
1.2.4 Sources of Gravity: How Matter Creates the Geometry 314
1.2.4.1 The Stress–Energy Tensor 314
1.2.4.2 Einstein's Field Equations 314
1.2.5 Other Theories of Gravity 315
1.2.6 A Cosmological Term in Einstein's Equations 315
2. **Some Consequences of Einstein's Field Equations** 316
2.1 Momentum and Stress Also Make Gravity 316
2.1.1 Gravitomagnetism 316
2.1.2 Gravitational Collapse 317
2.2 Black-Hole Theory 317
2.2.1 Black Holes in the Eighteenth Century 317
2.2.2 Black Holes in General Relativity 317
2.2.3 Singularities Inside the Hole 318
2.2.4 Black Holes Have No Hair 318
2.2.5 Black-Hole Thermodynamics 318
2.2.6 Wormholes 319
2.3 Gravitational Waves 319
2.3.1 The Necessity of Gravitational Waves 319
2.3.2 The Interaction of Gravitational Waves with Matter 319
2.3.3 Wave Emission: The Quadrupole Formula 321
2.4 Applications of General Relativity 322
2.4.1 Relativistic Stars (Pulsars) and Gravitational Collapse 322
2.4.1.1 Stellar Evolution 322
2.4.1.2 White Dwarfs 322

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

2.4.1.3 Neutron Stars 322
2.4.1.4 Black Holes 323
2.4.1.5 Pulsars 323
2.4.2 Black Holes in X-Ray Binaries and in Galactic Centers 323
2.4.2.1 X-Ray Binaries 323
2.4.2.2 Galactic Black Holes 324
2.4.3 Gravitational Lensing 324
2.4.4 Solar System and Stellar Orbits 325
2.4.5 Cosmology, Inflation, and the Origins of the Universe 325
2.4.5.1 Cosmological Models 326
2.4.5.2 Hubble's Law: The Big Bang .. 326
2.4.5.3 Cosmic Microwave Background Radiation 326
2.4.5.4 The Age of the Universe 327
2.4.5.5 Mass and Missing Mass 327
2.4.5.6 Microwave Anisotropies 327
2.4.5.7 Inflation and the Homogeneity Problem 328
3. Tests of Gravitational Theories and Their Technological Demands 328
3.1 Tests of the Einstein Equivalence Principle 328
3.1.1 Tests of Special Relativity 328
3.1.1.1 Michelson–Morley Experiments 329
3.1.1.2 Hughes–Drever Experiment ... 329
3.1.2 The Eötvös Experiment, the Weak Equivalence Principle, and the Fifth Force 329
3.1.2.1 Eötvös and the Weak Equivalence Principle 329
3.1.2.2 Fifth-Force Experiments 329
3.1.3 Gravitational Redshift 330
3.2 Solar-System Tests of General Relativity 330
3.2.1 The Deflection and Retardation of Light 330
3.2.1.1 Light Deflection 330
3.2.1.2 Shapiro Time Delay 331
3.2.2 Mercury's Perihelion Advance 331
3.2.3 Test of the Strong Equivalence Principle 332
3.3 The Binary Pulsar: An Astronomical Relativity Laboratory 332
3.4 Future Work in Experimental Gravitation 333
3.4.1 Search for Gravitomagnetic Effects 333
3.4.2 Tests of the Einstein Equivalence Principle 334
3.4.3 Further Fifth-Force Searches .. 334
4. Gravitational-Wave Detection: A Technological Frontier 334
4.1 Likely Sources of Detectable Waves 334
4.1.1 Supernovae 334
4.1.2 Coalescing Binaries 335
4.1.3 Pulsars 336
4.1.4 Ordinary Binaries 336
4.1.5 Cosmological Background 336
4.1.6 Unexpected Sources 336
4.2 Detectors 336
4.2.1 Bar-Type Detectors 337
4.2.2 Laser-Interferometric Detectors 337
4.2.3 Space-Based Detectors 338
Glossary 338
Works Cited 338
Further Reading 340

INTRODUCTION

Albert Einstein's development of general relativity in 1915 was one of the landmarks of twentieth-century theoretical physics. Einstein (1879–1955) had few experimental facts to guide him in replacing Newtonian gravity with a theory compatible with his earlier theory of special relativity (see RELATIVITY, SPECIAL). Using arguments of consistency, simplicity, and aesthetics, Einstein arrived at a theory that has subsequently passed stringent experimental tests, and that has predicted and explained a wide range of astronomical phenomena that were completely unknown at the time Einstein did his work.

Einstein's methodology had a great influence on the style of later attempts to simplify and unify theories of high-energy physics. Nevertheless, it is only since about 1970 that developments in astronomy and in experimental physics have placed general relativity into the mainstream of physics, part of the expected competence of any theoretical physicist.

General relativity is now the accepted theory of gravity. Although compatible with special relativity, it is incompatible with quantum theory. Presumably it will eventually be replaced by a quantum theory of gravity, and for this reason there is considerable interest in alternative theories of gravity and in testing gravity to find small departures from the classical theory. So far, all experiments and observations have been consistent with general relativity, many at a high level of precision.

This article discusses relativistic gravitation from a nonmathematical point of view, highlighting the underlying connections of general relativity with the rest of physics (including Newton's theory of gravity), examining the foundations of the theory (equivalence principle, geometrical approach), and emphasizing the observable and experimental consequences of the theory. These include astronomical phenomena, such as black holes, gravitational lenses, and the Big Bang; high-precision tests of the equivalence principle, of solar-system effects, and of gravitational radiation; and present efforts to make direct observations of gravitational waves from supernova explosions, colliding black holes and neutron stars, and the Big Bang itself.

Most aspects of experimental gravity push modern technology to its limits, and they are therefore fruitful areas for the application and development of new technology. Examples include the development of low-temperature superconducting quantum-interference device (SQUID) technology for space-based tests of magnetogravity and for gravitational-wave detectors; the routine employment of corrections for the gravitational redshift in the U.S. Air Force's Global Positioning System (GPS) navigation system; the application of dilution refrigeration on a large scale to cool gravitational-wave bar detectors weighing several tons to temperatures approaching 10 mK; and the use of high-power cw lasers and ultralow-loss optical components in interferometric gravitational-wave detectors. Current research into squeezed light for telecommunications (see OPTICAL COMMUNICATIONS) was stimulated in large part by theoretical work on ways of making gravitational-wave detectors more sensitive.

The field of experimental relativity is an active one; this article presents a snapshot of current developments as of the time of its completion, early in 1993.

1. GENERAL RELATIVITY AND RELATIVISTIC GRAVITY

Einstein's theory of general relativity (GR) is the accepted theory of gravity in modern physics. In the Solar System and other regions of weak gravitational fields, it differs little from Newton's theory of gravity; but when gravity is strong enough to accelerate particles to near the speed of light, its predictions are very different. General relativity describes gravity in geometrical terms; the curved geometry of space and time is an arena in which all the rest of physics takes place. The theory has become an everyday tool for astrophysicists who model the most unusual phenomena in astronomy: black holes, gravitational waves, and the theory of cosmology are all features of the theory that find application in astronomy. For textbook introductions to general relativity, see Misner *et al.* (1973) or Schutz (1985).

1.1 Gravitation Theory: An Overview

1.1.1 Relativistic Gravity in Physics

1.1.1.1 Roots in Newtonian Gravity. Newton's laws of motion and his theory of gravity had a revolutionary effect on science. They explained the motions of the planets and taught scientists that the heavenly bodies were subject to the same laws as bodies on Earth. Newton showed that masses exert forces on each other that diminish with the square of the distance between them. The success of Newtonian gravity in the Solar System, as well as in other stellar systems, is described in the article ASTRONOMY.

In Newton's theory, the gravitational forces between bodies act instantaneously across any distance. Space is flat and time is universal: once synchronized, identical clocks keep time with one another regardless of their state of motion. It was not until the nineteenth century that physicists had any experimental evidence that challenged this idea about time (the Michelson–Morley experiment). The first evidence for the *curvature* of space came from the 1919 eclipse expedition, described in Sec. 3.2.1 below.

1.1.1.2 The Impact of Special Relativity: Relativistic Gravity is Required. Einstein's development of special relativity (SR) changed notions of space and time (see RELATIVITY, SPECIAL). Space, while still flat, merged with time to form *spacetime*. Clocks in relative motion would not remain synchronized, and moving lengths would contract. Most importantly, no influences could travel faster than light.

This limiting speed undermines Newtonian gravity: gravity cannot "act at a distance" without a delay. In the solar system, where light crosses an orbit in a small fraction of the orbital period, the delay is small, and Newton's theory works well. But the theory had to be replaced.

It took Einstein another decade to find a suitable replacement: *general relativity*. It incorporates SR as the limit of vanishing gravity. It approaches Newton's theory as a limit when gravity is the dominant force but is still weak enough that the resulting motions are much slower than light. But in other regimes GR makes startlingly new predictions. Black holes, gravitational waves, the Big Bang: Einstein had no inkling of these when he devised the theory, and it is remarkable that these unforeseen consequences of the theory have been verified by astronomical observations in recent years.

It is interesting that Einstein was "beaten" to the correct form of the equations of GR by the great mathematician David Hilbert, whose paper (Hilbert, 1915) was submitted five days before Einstein presented his theory to the Prussian Academy (Einstein, 1915). Einstein is universally credited with the theory because his physical insight into its foundations, such as the equivalence principle and the geometric form of the theory, guided Hilbert as much as they guided Einstein; and it was Einstein who worked out the first physical consequences of the equations.

Despite GR's success, it is not the only possible relativistic generalization of Newtonian gravity. In many ways it is the simplest, but many others exist. Most, like the Brans–Dicke–Jordan class of theories (Sec. 1.2.5), introduce extra fields that have further gravitational effects. Others depart from GR by violating the equivalence principle (Sec. 1.2.2.2). The only way to decide which theory is right in detail is to test them experimentally. Testing relativistic gravity is one of the areas in which modern technology has made possible great advances (Sec. 3).

1.1.1.3 The Impact of Quantum Mechanics: Quantum Gravity is Required. Although GR reconciled the contradiction between SR and Newtonian gravity, it soon fell afoul of a different inconsistency. The theory of quantum mechanics matured only two decades after the birth of GR, and now almost all of physics is described by quantum theories; GR is the only major exception.

Much current research attempts to find an appropriate quantum theory of gravity. So far that effort, while producing many interesting new ideas, has not reached its goal. Because GR is a highly nonlinear theory, conventional methods of quantizing theories have so far failed to produce a consistent quantum theory of gravity. The search is made more difficult because of the paucity of experimental guidance. If there are properties of the known universe that give us decisive clues to quantum gravity, we have not yet recognized them.

Dimensional analysis is interesting in this respect. If quantum gravitational effects dominate in some situation, then the laws must involve the three fundamental constants of nature, G (Newton's constant of gravitation), h (Planck's constant), and c (the speed of light). From these constants one can construct numbers with any of the fundamental dimensions, e.g.,

$$(hc^5/G)^{1/2}=4.9\times10^9\ \mathrm{J}=3.1\times10^{28}\ \mathrm{eV}\quad \text{(Planck energy)}; \tag{1}$$

$$(hG/c^3)^{1/2}=4.1\times10^{-35}\ \mathrm{s}\quad \text{(Planck time)}; \tag{2}$$

$$(hG/c^5)^{1/2}=1.4\times10^{-43}\ \mathrm{m}\quad \text{(Planck length)}. \tag{3}$$

These are so far from accessible experimental domains that it is not surprising that we have little experimental guidance for quantum gravity.

It may well be, however, that effects related to quantum gravity are important on other scales, particularly if gravity is part of a *grand unified theory* of physical interactions. (See UNIFIED FIELD THEORIES.) There may be small corrections to Einstein's equations that are appropriate in a semiclassical limit, and these might take the form of extra gravitational fields. For this reason, there has been renewed interest of late in alternative theories of gravity and in the experimental limits that can be placed on them.

The regime of quantum gravity may be far from the experimental domain, but the theoretical motivation for finding the right theory is very strong. Besides the need simply for consistency with other theories, there is also the striking fact that GR seems to predict its own failure: many solutions, including apparently all black-hole solutions, contain *singularities*, which are places where GR fails to describe the Universe in a consistent way. (See Sec. 2.2.3.) The Big Bang begins with such a singularity. There is a strong feeling that quantum gravity will tell us what the singularities really mean.

1.1.2 Where Relativistic Gravity Is Important. Relativistic gravity is required where the speeds produced by gravitational attractions approach the speed of light. A system with mass M and overall size R will produce velocities of order

$$v_{\text{self-gravity}}=(GM/R)^{1/2}\ . \tag{4}$$

For example, the escape velocity from a spherical body is $\sqrt{2}v_{\text{self-gravity}}$, and the velocity of a planet in a circular orbit of radius R about a star of mass M is exactly $v_{\text{self-gravity}}$. A test for the importance of relativistic effects is

$$\Phi:=GM/Rc^2\ll 1 \Rightarrow \text{gravity is nonrelativistic.} \tag{5}$$

At the Sun's surface, $\Phi=2\times10^{-6}$; everywhere else in the Solar System it is smaller. This accounts for the accuracy of Newtonian gravity. One needs relativistic gravity when Φ is not small.

1.1.2.1 Compact Objects of High Density. To make Φ larger one can increase M and/or decrease R. If a body of fixed mass shrinks in size, relativistic effects become important when the radius nears the *gravitational radius* R_g:

$$R_g:=2GM/c^2\ . \tag{6}$$

The factor of 2 in this is conventional.

The gravitational radius of the Sun is about 3 km. Remarkably, stars do collapse to nearly this size: perhaps 1% of all stars are *neutron stars* (Sec. 2.4.1), with radii about 10 km. If a star actually reaches the gravitational radius, then it collapses further to a *black hole* (Sec. 2.2).

The density ρ of a neutron star is enormous. If we take $M=1M_\odot$ (astronomer's notation for the mass of the Sun) and $R=10$ km, we have $\rho_{\text{NS}}=5\times10^{17}$ kg m^{-3}. This is comparable to the density of a typical atomic nucleus.

The critical density ρ_c at which the object reaches its gravitational radius depends only on M:

$$\rho_c:=3c^6/32\pi G^3M^2\ . \tag{7}$$

For the Sun it is about 2×10^{19} kg m^{-3}, but such enormous densities are not required for bodies of larger mass. A cluster of some 10^8 stars will collapse to a black hole if it reaches the density of water. This seems to have happened in the centers of galaxies (Sec. 2.4.2.2).

1.1.2.2 Cosmology: Relativistic Gravity at Low Densities. A second interesting way of reaching a situation where relativistic gravity is required is by increasing the size of a system of fixed density. As the mass increases, the critical density drops. Eventually the critical density approaches the system's fixed density and the body becomes relativistic. For most systems, this size would be unrealistically large. But there is one system that we can make as large as we like: the Universe itself.

The average mass density of the Universe (or at least of the piece we can see) is highly uncertain, but an observational lower bound is 10^{-28} kg m^{-3}. A sphere of radius 1.3×10^{27} m, or about 10^{11} light years, would enclose enough matter to give a relativistic gravitational field. Such distances are the regime of *cosmology*, the study of the Universe as a whole (Sec. 2.4.5).

Being able to describe cosmology consistently was in fact one of the great triumphs of GR: Newtonian theory is highly ambiguous when applied to infinite systems, whereas GR provides well-defined models.

1.1.2.3 Small Relativistic Effects: Gravitational Waves and Equations of Motion. Relativistic gravity is essential for the description of compact objects and of cosmology, but it can also be important for situations where its effects make small but measurable changes to Newtonian gravity. These are easiest to measure when the new relativistic effect is absent from Newtonian gravity, so that its signature is easy to see. Some effects we will look at below include the following:

1. Gravitational radiation (Sec. 2.3) is a direct consequence of the limitation on the speed of propagation of gravitational influences: oscillating motions of stars (say, in a binary system) produce an oscillating field that moves through space, a wave in the gravitational field. These can be directly detected essentially by looking for time-dependent gravitational fields. The weakness of gravity makes this one of the greatest technical challenges in physics today (Sec. 4). Even when direct detection of the waves is impossible, the back-reaction effects of their emission can sometimes be observed. An example of great importance is the binary pulsar PSR1913+16 (Sec. 3.3).
2. Gravity deflects light and produces observable time delays, effects that can be directly observed (Sec. 3.2.1). Light deflection also leads to gravitational lensing of astronomical images (Sec. 2.4.3).
3. Relativistic gravity produces small anomalies in the motions of planets and satellites of the Earth. We will discuss these in Secs. 3.2.2 and 3.4.1.

1.2 Fundamental Ideas and New Concepts of General Relativity

1.2.1 The Incorporation of Newtonian Gravity. Any theory of gravity must address two issues: how gravity affects matter, and how matter creates gravity. We review these in the context of Newtonian gravity, and then look at what new ideas Einstein brought to the subject.

1.2.1.1 The Equivalence Principle: How Gravity Affects Matter. Gravity affects everything: unlike any other force in nature, it cannot be turned off or screened out. Galileo showed that it affects all things in the same way: all bodies fall with the same acceleration in a gravitational field. For this reason, one cannot experimentally distinguish a uniform gravitational field from a uniform acceleration of the experimenter. This equivalence between gravity and acceleration is called the *weak equivalence principle* (WEP).

1.2.1.2 Newton's Field Equation: How Matter Creates Gravity. Newton's law for the gravitational force between two bodies of masses m_1 and m_2, separated by a distance r,

$$F_{\text{grav}} = Gm_1m_2/r^2 , \tag{8}$$

incorporated the equivalence principle: the acceleration of body 1, F_{grav}/m_1, is independent of m_1. Newton arrived at the $1/r^2$ law by studying the motion of the Moon, but the resulting law turned out to describe all the known properties of the solar system. One of its great triumphs was that it predicted that planetary orbits should be ellipses with the Sun at a focus.

Later development of Newton's theory found that the simple law given by Eq. (8) was in an inconvenient form to deal with complicated bodies. The modern form introduces the *gravitational potential* created by particle 1 as the following integral over its density ρ_1:

$$\phi_1(\mathbf{r}) = -G \int \frac{\rho_1(\mathbf{r}')d^3x'}{|\mathbf{r}-\mathbf{r}'|} . \tag{9}$$

The force on particle 2 due to particle 1 is

$$\mathbf{F}_{12} = -m_2\nabla\phi_1(\mathbf{r}) . \tag{10}$$

The field ϕ solves the *Poisson equation*:

$$\nabla^2\phi = 4\pi G\rho . \tag{11}$$

This very compact form for Newton's potential is usually known as *Newton's field equation*. It is the equation that must be generalized in any new theory of gravity.

1.2.1.3 Galileo's Principle of Relativity. Galileo's other seminal contribution to modern gravitational theory is what is now called the *principle of relativity*. A simple statement of the principle is that the (uniform) speed of an experimenter does not affect the outcome of an experiment. Galileo used the example of a ship moving on a calm sea: if you stand on the deck and drop a ball, it falls straight down relative to you, not along a vertical line as judged by an observer on the shore. For modern travelers, the experience of flying in an airplane at 900 km/h is more compelling: unless one hits air turbulence or looks out the window, it is impossible to tell how fast one is going.

1.2.2 Basic Ingredients of General Relativity

1.2.2.1 Special Relativity. Newtonian gravity already incorporated the Galilean principle of relativity, that the laws of physics were independent of the state of motion of the experimenter. In SR, this becomes formalized into the idea of an *observer,* who is basically an entire reference frame with an information-gathering mechanism to determine what happens in experiments. We usually restrict ourselves to so-called Lorentz observers, who move at a uniform speed relative to one another, and who normally use a rectangular coordinate system for locating things in space. Their time coordinate consists of a set of synchronized clocks.

What Einstein added to Galileo's principle in order to get SR was the central role of the speed of light c. In SR, the speed of light has the status of a physical law: it is invariant under a change of experimenter. If two different experimenters measure the speed of the *same* beam of light, they will *each* measure $c=2.998...\times 10^8\ \mathrm{m\ s^{-1}}$, regardless of their speed relative to each other.

Special relativity made many conceptual changes in physics, which in turn carry over to any relativistic theory of gravity. Space and time are no longer distinct, but form a four-dimensional *spacetime,* whose points are called *events.* The set of all events experienced by a given particle through all its history is called its *world line.* Different observers make different measurements of simultaneity, of distance (the *Lorentz contraction*), and of time (*time dilation*), but there is still an invariant measure of distance, called the *spacetime interval* ds^2:

$$ds^2=dx^2+dy^2+dz^2-c^2dt^2\ . \qquad (12)$$

(There are many conventions for this definition. We adopt the one most commonly used in GR.)

This is a generalization of the Pythagorean theorem of Euclidean geometry. Just as in Euclidean geometry, the physical length of the hypotenuse of the right triangle is given by the square root is the sum of the squares of the sides, so also in SR is the square root of the interval a physical observable. For two events separated by a positive (*spacelike*) interval, its square root is the length of a physical ruler that would stretch between the events in a frame in which the events are simultaneous. This is called the *proper distance* between the events. For two events separated by a negative (*timelike*) interval, we define the *proper time* $d\tau$ by $d\tau^2=-ds^2/c^2$. The proper time is the time elapsed on a clock that experiences both events. Events separated by a zero (*null* or *lightlike*) interval have zero proper time between them.

Because spacetime is four dimensional, vectors have four components and are called *four-vectors.* A simple four-vector is a tangent vector to a timelike world line. All such tangent vectors at a given event are parallel to one another. A useful one, conventionally called the *four-velocity* $\boldsymbol{U}$, is defined to have components

$$(U^t,U^x,U^y,U^z)=\left(\frac{dt}{d\tau},\frac{dx}{d\tau},\frac{dy}{d\tau},\frac{dz}{d\tau}\right)\ . \qquad (13)$$

(Here we have introduced some conventions: vector indices are written as superscripts, and the time dimension is normally taken first.) Because τ is an invariant, the components of $\boldsymbol{U}$ change in the same way as the coordinates do when the observer changes. This is part of the definition of a four-vector. Note that for small velocities, the interval $d\tau$ is nearly equal to the time interval dt, and the spatial components of $\boldsymbol{U}$ are nearly equal to the ordinary velocity. At the other extreme, the four-velocity of a null world line (e.g., that of a photon) is undefined, because $d\tau=0$ along it.

Although SR is normally described in terms of Lorentz observers with rectangular coordinates, one can allow other coordinate systems for both space and time. This complicates the mathematics, such as the form of Eq. (12), but is a price worth paying in particular circumstances, such as when dealing with an accelerated particle. What distinguishes SR from GR is, fundamentally, that it is *possible* to find a single Lorentz frame that covers all of spacetime. The notion that Eq. (12) describes the way rulers and clocks behave everywhere in spacetime is very restrictive; in particular, since the interval depends only on the coordinate differences between events and not on their absolute location (no explicit dependence on x,y,z,t), it follows that SR can only hold in a spacetime that is perfectly uniform, or *homogeneous*.

Special relativity also unifies energy and momentum. Just as time is dilated under a change of observer, so is energy increased. Just as there is no invariant separation between space and time, so is there none between energy and momentum. They are parts of a single vector in spacetime, the energy–momentum four-vector, which equals the rest mass of the particle times its four-velocity. Its time component is the particle's energy, and its spatial components are the momentum. For photons, the energy–momentum vector exists, even though the four-velocity does not. Since this vector is null, photons have energy equal to c times their momentum.

The introduction of the energy–momentum vector was a great conceptual advance for physics. In Newtonian mechanics, neither energy nor momentum had any observer independence: both changed when the observer's velocity changed. In SR, there is a single energy–momentum four-vector, unique to the particle. Its time component (the energy) and spatial components (momentum) do change, but they are subsidiary to the unique four-vector.

The identification of an invariant four-vector is a way of thinking introduced by SR that has great importance for GR. It is easy to understand what it means for a scalar quantity to be invariant, such as the interval in Eq. (12): its value is the same for all observers. A four-vector is likewise an invariant under spacetime coordinate transformations, just as the electric field vector **E**, for example, is under spatial rotations; its orientation with respect to other vectors is unchanged, even though its components change. (Many authors call such a thing a *geometrical object*.)

1.2.2.2 The Einstein Equivalence Principle. Special relativity requires a new formulation of the equivalence principle. In the Galilean principle, we traded off gravity against acceleration. But what does acceleration mean for light, whose speed cannot change? Fortunately, there is a new way to state the principle that is equally at home in Galilean or Einsteinian relativity.

The key is to focus on an experimenter who falls freely in a uniform gravitational field. This experimenter could, say, be in a spaceship that is coasting through this field. Any other particle that falls freely (has no nongravitational forces on it) will keep a constant velocity with respect to the experimenter. This can then also apply to light: if a photon passes by the falling experimenter, it will maintain a constant relative velocity, maintaining not only its speed c but also its direction.

More generally, any observer who works in a *freely falling reference frame* in a uniform gravitational field will not be able to determine that there is a gravitational field at all. The results of any experimenters are just as in SR. This is the *Einstein equivalence principle* (EEP).

The EEP implies that SR can cope with *uniform* gravitational fields: just use freely falling observers and the effects of the fields go away. More importantly, it also means that SR is incompatible with nonuniform gravity, which is what one has in every real situation. Near the Earth, for example, the local acceleration of gravity changes from place to place. Since SR requires a Lorentz frame to exist and be rigid everywhere, it is not possible to identify such a frame in the presence of real gravity. Thus, SR cannot incorporate nonuniform gravity.

However, even in a nonuniform field, a *local* freely falling observer sees no gravity locally. Weightless astronauts feel no gravity, even though gravity holds them in orbit around the Earth. The EEP implies that gravity can still be removed *locally*, even if not globally. It is a powerful principle. Three of its consequences are explored in the next three sections.

1.2.2.3 Tidal Forces Are the Real Gravity. The EEP implies that all the real, irremovable effects of gravity must therefore have to do with the differences between nearby freely falling observers. Any relativistic theory of gravity must give special importance to these differences between local gravitational accelerations.

These differences are called the *tidal forces*. They get their name from the tides raised on the Earth by the Moon and Sun. The Earth's oceans bulge out on both the near side and the far side with respect to the Moon. The near side of the Earth wants to fall towards the Moon slightly faster than the average, while the far side falls slightly slower. The size of the bulge is determined by the *difference* in the acceleration of gravity due to the Moon across the Earth. This gradient is called the tidal force of the Moon. If we could not see the

Moon, we could still infer its existence from observations of the tides: tidal forces are the observer-independent gravitational forces.

1.2.2.4 Gravitational Redshift. Photons are redshifted (lose energy) as they climb out of a gravitational field. A simple thought experiment shows how this follows from the EEP. We send a photon up from the ground to the top of a tower of height h, where it is detected by equipment at rest with respect to the Earth. During the time h/c it takes the photon to reach the top of the tower, a local freely falling frame has acquired a speed $v=gh/c$ downwards, where g is the acceleration of gravity on the surface of the Earth. In this frame, the EEP tells us that the photon's frequency does not change, as measured by this observer. Since the top of the tower is *rising* at speed gh/c relative to this observer, there is a Doppler redshift that produces a lengthening of the wavelength λ by a fractional amount

$$\frac{\delta\lambda}{\lambda}=\frac{v}{c}=\frac{gh}{c^2}. \tag{14}$$

The EEP implies that an experimenter who remains at rest with respect to the Earth measures this redshift of the photon as it climbs upwards in the gravitational field. Since the product gh is the change in the Newtonian gravitational potential between the emitter and the absorber, the redshift is just a function of the potential difference.

The gravitational redshift has another important interpretation. Since the frequency of the photon is measured essentially by clocks located at the emitting and absorbing locations, it is impossible for these clocks to remain synchronized. During a given number of oscillations of the photon's electromagnetic field, the clock on the ground ticked less time than the one higher up. The clock nearer the ground runs more slowly. *The gravitational redshift implies a gravitational time dilation.* This is now routinely measured by the GPS of satellites (Sec. 3.1.3).

1.2.2.5 Gravitational Deflection of Light. Consider light passing a star. Since light must travel on a straight line with respect to a local freely falling observer, and since these observers all fall towards the center of the star, the light must continually bend its direction of travel in order to go on a straight line with respect to each observer it happens to pass. We can estimate the size of the effect roughly by the following argument, which uses Newtonian concepts.

Let us consider just one freely falling observer who is at rest with respect to a star of mass M at the point where the light beam makes its closest approach to the star as it passes it by. Let this closest distance be R. The observer's acceleration towards the star is $g=GM/R^2$. Traveling at speed c, the photon will experience most of its deflection in a time of order R/c, the time it takes for the photon to move significantly further away from the star. During this time, the observer has acquired a speed $v=gR/c=GM/Rc$ perpendicular to the motion of the photon. By the EEP, the photon must also have acquired roughly this same speed transverse to its original direction. This means it changes its direction of travel relative to the star by an angle $\alpha=v/c=GM/Rc^2$ rad.

The total deflection should be double this, since the photon will experience the same deflection in to the point of nearest approach as going out. A more careful integration of this effect along the whole of the light path gives, coincidentally, exactly the same as our rough estimate, a total Newtonian deflection of $2GM/Rc^2$. This calculation was first performed by Cavendish in 1784 and independently by von Soldner in 1801, on the assumption that light was a corpuscle moving at speed c.

In relativistic gravity, however, there is an added effect that increases the deflection. This is, as we will see later, that space is curved near the star, and so the actual natural path of light as it passes the star, even if it were to go at an infinite speed, is deflected. The result in GR is that light is deflected by a net angle of

$$\alpha=4GM/Rc^2\ , \tag{15}$$

double the Newtonian result.

For a light beam passing the Sun, just grazing its surface, the effect is small: α is less than 1.75 seconds of arc. This is measurable, as described in Sec. 3.2.1.

More dramatically, if the gravitational field is that of a faint cluster of galaxies, and the source of the light is a more distant but brighter quasar, then light from the quasar might reach the Earth by two or more different routes, giving separate images of the source separated by angles that are easy to resolve.

This is called *gravitational lensing*. By studying the distribution and relative brightness of multiple images, and the shape of distorted arclike images of a distant quasar, produced by an intervening galaxy or galactic cluster, astronomers can infer the distribution of mass in the cluster, particularly that of dark matter, which produces gravity, but not light. Indeed, one astronomer has dubbed this method *gravitational tomography*. By looking for differences in the arrival time of temporal features in the quasar signal (such as flare-ups) in the different images, a result of the light having traveled different path lengths, it may be possible to determine the distance of the lensed quasars and thereby obtain cosmological information. This is described more fully in Sec. 2.4.3.

1.2.3 General Relativity: Gravity as Geometry

1.2.3.1 Why Gravity Is Geometrical. The EEP tells us that the effect of gravity on a body does not depend on its internal properties. In a given gravitational field, a freely falling body will follow a straight line with respect to local freely falling observers. This means that its trajectory depends only on its initial position and velocity. Any body with the same initial position and velocity will follow the identical trajectory.

Einstein reasoned that the trajectories were more a property of the gravitational field itself than of its interaction with specific bodies. Gravity is not a force like electromagnetism or the strong interactions. It is a property of space and time.

It is natural to try to relate gravity to curvature. If the trajectories of freely falling bodies—bodies on which no nongravitational forces are acting—are curved lines, then maybe there is a natural way to associate these lines with the locally straight lines, called *geodesics*, of a curved space. This is exactly what Einstein did, but he had to include the curvature of space *and* time together. Consider a curved surface, such as that of an apple. By following as straight a line as possible, one traces out a curve that meanders over the surface. This curve is determined by one's initial position and direction. It is not determined, however, by the speed with which one travels along the apple. Since gravitational trajectories depend not only upon direction but also upon speed, they cannot be modeled by the curvature of space alone.

In a curved four-dimensional spacetime, the natural trajectories (the geodesics) will, by analogy, be determined by the initial position and the initial direction of the tangent vector, but not by the length of the tangent. Now, as we saw in our discussion of SR in Sec. 1.2.2.1, the four-velocity is a tangent vector to a particle's world line. Its components do contain the information about the ordinary velocity of the particle. For example, the ratio of the x component to the time component is $(dx/d\tau)/(dt/d\tau)=dx/dt=v^x$. Since any other tangent is a multiple of $\boldsymbol{U}$, this ratio will be the same for any tangent: it depends only on the direction and not the length of the tangent four-vector. Therefore, the information needed to determine a geodesic in a curved spacetime is exactly that needed to determine a gravitational trajectory: the position and ordinary velocity of the particle.

In this way, the EEP leads to the representation of gravity by the curvature of spacetime.

1.2.3.2 The Metric of Gravity. Curved spaces usually have no *natural* coordinate system. Special coordinates, such as Lorentz observers, do not exist in nonuniform gravitational fields, as we have seen. So we must admit arbitrary spacetime coordinate systems $\{x^\alpha,\ \alpha=0,1,2,3\}$ and we do not necessarily attach any physical significance to particular coordinate values or intervals. (Another convention here: we use Greek letters for spacetime indices, and the index 0 stands for the coordinate t.) Physical significance attaches only to things that can be measured. Coordinates are usually matters only of definition, not of independent measurement.

The intrinsic geometry of a curved spacetime is completely defined if one gives the distances between points (events, in our case). One only needs local distances, which may be integrated up along curves in the spacetime to give global ones. The distance information is contained in the *metric tensor* $g_{\alpha\beta}$:

$$ds^2=\sum_{\alpha=0}^{3}\ \sum_{\beta=0}^{3}\ g_{\alpha\beta}dx^\alpha dx^\beta\ . \tag{16}$$

This equation gives a prescription for calculating the invariant length ds^2 between nearby events separated by coordinate intervals dx^α in terms of the given set of functions $g_{\alpha\beta}$. The interval is an *invariant*; this property determines how the components $g_{\alpha\beta}$ change when the coordinates change. The term "tensor" refers to the fact that the 4×4 matrix of components must transform this way. (See GEOMETRICAL METHODS for more information on tensors.)

The metric tensor of SR is conventionally called $\eta_{\alpha\beta}$. By comparing Eq. (12) with Eq. (16), we find $\eta_{00}=-c^2$, $\eta_{xx}=\eta_{yy}=\eta_{zz}=+1$, and all other components of $\eta_{\alpha\beta}$ vanish. The spacetime of SR is called *Minkowski spacetime*, and the metric tensor $\eta_{\alpha\beta}$ is called the Minkowski metric.

The metric contains observable information. For example, the proper time on a clock that stays at fixed spatial coordinates ($dx=dy=dz=0$ on its world line) is related to the coordinate time interval dt by $d\tau^2=-g_{00}dt^2/c^2$. Given two clocks at different locations $\mathscr{A}$ and $\mathscr{B}$, then the ratio $g_{00}(\mathscr{A})/g_{00}(\mathscr{B})$ determines the time dilation between them.

We saw in Newtonian gravity that the gravitational time dilation depends on the difference between the potentials at the two points (Sec. 1.2.2.4). Since Newtonian theory only applies when gravity is weak and the metric is close to the Minkowski form, we have that $g_{00}=-(1+\epsilon)c^2$, where ϵ is small. Then the ratio of the two values of g_{00} depends to first order on the *difference* between the values of ϵ. Therefore, there is a close connection between ϕ and ϵ. If one follows the algebra through, one finds that, for a given Newtonian-type field, the metric needs to have $g_{00}\approx-(1+2\phi/c^2)c^2$. This establishes the role of the metric tensor in relativistic gravity: it is the analog of the Newtonian *potential.*

The metric has some simple properties that one can infer from Eq. (16). First, since exchanging dx^α and dx^β in this equation can be accomplished by relabeling the indices, there cannot be any difference between $g_{\alpha\beta}$ and $g_{\beta\alpha}$: the 4×4 matrix of components $g_{\alpha\beta}$ is *symmetric*. This means there are 10 independent components. Second, since one has complete freedom to change the coordinates, there are essentially four degrees of arbitrariness among the 10 components: a given curved spacetime can be described by a variety of metric tensors, all related by coordinate transformations. There are essentially $10-4=6$ geometrically independent functions that one can choose in order to determine the spacetime.

1.2.3.3 Local Flatness of Spacetime and the Local Freely Falling Frames. The freely falling observers are a special class of coordinate systems that make spacetime look as much like SR as possible in the region of one event. This idea is very similar to the observation that any smooth surface is nearly flat if we look at only a small region. The *local flatness theorem* states that we can choose any point as the origin of a locally flat (or in spacetime, freely falling) coordinate system and adjust the coordinates in such a way that if we make a Taylor expansion of the metric components about this point, the first departure from flatness will be at *second order* in the coordinates. This means that the coordinates can be chosen to make the first derivatives of the metric vanish at any desired point. Since the acceleration of gravity also vanishes in a freely falling frame, the local flatness theorem suggests that *the acceleration of gravity resides in the first derivatives of the metric tensor components* at the point. This is consistent with the relation we saw earlier between the metric and the Newtonian potential.

1.2.3.4 Curvature and Tidal Forces. The *curvature* of the spacetime is a function of the metric tensor. The relation is not immediately straightforward, since even in the flat (Minkowski) spacetime of SR, one can choose a funny coordinate system and make the metric components look very complicated. The local flatness theorem tells us that the curvature is not defined by the first derivatives of the metric, since these can be made to vanish. Therefore, the curvature information must reside in the second derivatives of the metric components. Now, if the acceleration of gravity is given by the first derivatives, then the true gravitational forces—the tidal forces—are given by the gradients of the acceleration of gravity, which are the second derivatives of the metric. Therefore, we conclude that *the tidal forces are represented mathematically by the curvature of spacetime.*

The fundamental measure of curvature is the *Riemann curvature tensor,* denoted by $R^\alpha{}_{\mu\beta\nu}$. It is an object with four indices, essentially because it must be a function of the second derivatives of the metric, which have four indices: $\partial^2 g_{\alpha\beta}/\partial x^\mu \partial x^\nu$. There are a number of symmetries among the indices, so that, among its $4^4=256$ components, only 20 are algebraically independent. This is far fewer than the number of independent second de rivatives of the metric tensor components ($10^2=100$) because most of the second derivatives can be adjusted or even set equal to zero by appropriate coordinate choices. The 20 Riemann components represent the irreducible, coordinate-free geometrical information about the curvature. It is a theorem that the spacetime is flat (Minkowskian) if and only if the Riemann tensor vanishes everywhere.

One can derive from this tensor another called the *Ricci curvature tensor* $R^{\alpha\beta}$ by taking a trace. (It turns out that there is only one independent trace of the Riemann tensor, so that the Ricci tensor is essentially unique). The Ricci tensor is a symmetric tensor, with 10 independent components. It has two indices. One can likewise take its trace to get an object with no indices: this is called the *Ricci scalar* R.

A further two-index tensor is the so-called *Einstein tensor* $G^{\alpha\beta}$, which is formed by subtracting one-half of the product of the Ricci scalar and the metric tensor from the Ricci tensor. Its derivatives in a locally flat coordinate system satisfy a remarkable identity, called the *Bianchi identity*: for any metric, the Einstein tensor is divergence-free. (The divergence of a symmetric two-index tensor is defined as the trace of its first derivative, the trace being taken on the derivative index and one of the tensor indices.) This identity is central to Einstein's theory of gravity, as we will see in the next section. Notice also that there is one other divergence-free two-index tensor: the metric tensor itself, whose first derivatives all vanish in a locally flat coordinate system.

1.2.4 Sources of Gravity: How Matter Creates the Geometry

1.2.4.1 The Stress–Energy Tensor. In Newtonian gravity, the source of gravity is simple: the density of mass. It creates the field (the Newtonian potential ϕ) through Newton's field equation, Eq. (11). Because of the interconvertibility of mass and energy, we would expect the density of total mass–energy to be the source. This is, however, incompatible with SR for two reasons. First, in SR, mass is a form of energy, and energy is not a scalar: it is only one component of the energy–momentum four-vector. Second, in a relativistic theory, the notion of a density is not frame independent. The Lorentz contraction means that volumes and therefore densities depend on the observer.

These two transformations of mass–energy density (one acting on *mass–energy*, the other on *density*) mean that the mass–energy density is actually only one component of a two-index tensor. The other components of this tensor include the momentum density and the *stress tensor*. Just as SR links momentum and energy in a single four-vector, so too does it link stress and energy–momentum density into a single four-tensor with two indices. The unified tensor is called the *stress–energy tensor*. It is denoted by $T^{\alpha\beta}$. The mass density is measured by T^{tt}, the density of x-momentum by T^{tx}, and the stresses by the spatial components.

Energy and momentum are conserved in ordinary physics, and by the EEP this must be true in a local freely falling frame. For energy, for example, the conservation law says that the rate of change of energy in any volume equals the total net rate at which energy enters the volume across its sides. This implies that the (four-)divergence of the stress–energy tensor vanishes.

If mass density creates gravity in Newton's theory, then it is natural to expect the stress–energy tensor to be the source for a relativistic theory of gravity. A simple counting argument reinforces this idea. The stress–energy tensor can be shown to be symmetric on its two indices, so it has 10 independent components. This matches the 10 components of the metric tensor that describe the gravitational field: 10 unknowns $g_{\alpha\beta}$ determined by the 10 sources $T^{\alpha\beta}$.

However, this simple argument is too hasty: we saw above that there are effectively only six components of the metric that can be determined by the physics, since there are four degrees of coordinate freedom. Would such a theory be overdetermined? The answer is no, because there are as well four identities among the components of the stress–energy tensor: the laws of conservation of energy and momentum. So the counting is indeed encouraging: there are as many independent "sources" of gravity as there are "fields."

1.2.4.2 Einstein's Field Equations. To turn the counting argument into a full theory of gravity, Einstein had to find the appropriate differential equations that would allow the fields to be found in terms of the sources. He made what seemed to be the simplest choice.

We shall see below that there are more complicated alternatives.

One wants a generalization of Eq. (11), which has the form of a differential equation involving second derivatives of the potential set equal to the mass density as a source. The analog of the potential in a geometrical theory is, as we have seen, the metric tensor. So we need an equation involving second derivatives of the metric. The natural tensorial object involving second derivatives is the curvature tensor. If we take the hint of the preceding section and assume that the source will be the stress–energy tensor, then what we want is to equate the stress–energy tensor to a two-index curvature tensor that, like the stress–energy tensor, is divergenceless. As we pointed out earlier, there is such a tensor: the Einstein tensor. Einstein's choice was, then, to adopt the following form of the field equations:

$$\text{Einstein tensor} = \kappa \times \text{stress–energy tensor} ,$$

where κ is a constant of proportionality that is not fixed by the argument so far.

We fix κ by demanding that the solution of the field equations when the density is low and the velocities of the sources are small should have geodesics that are the paths of freely falling particles in the Newtonian gravitational field of the same sources. This gives $\kappa = 8\pi G/c^4$. In mathematical notation, the field equations of GR are

$$G^{\alpha\beta} = (8\pi G/c^4) T^{\alpha\beta} . \tag{17}$$

These equations are determined only by the correspondence with Newtonian gravity: there are no free parameters. Therefore, experimental constraints on GR can be very strict. One new consequence of Einstein's field equations after another has been calculated: the perihelion shift of Mercury's orbit (Sec. 2.4.4); the gravitational deflection of light (Sec. 2.4.3); the energy radiated in gravitational waves by a binary star system (Sec. 2.4.4); the existence of black holes (Sec. 2.4.2); the Big Bang and its consequences (Sec. 2.4.5). Each one could have proved the undoing of the theory. Yet in each case, there has been excellent quantitative agreement with the predictions of GR.

Because the Einstein equations are exceedingly nonlinear in the metric tensor, it has not been easy to find solutions of the field equations. Some exact solutions are known, which fortunately correspond to some of the most interesting simple cases: the field outside an isolated neutron star, the field of a black hole, and the metric of the homogeneous universe as a whole. In other cases, approximations can be made to good effect, such as when looking for small corrections to Newtonian gravity or for gravitational waves. And increasingly, as computers improve, relativists are attempting to solve more complicated systems numerically.

1.2.5 Other Theories of Gravity. General relativity is not the only geometrical theory of gravity possible. By the Einstein equivalence principle (EEP), the spacetime metric is the only field that directly affects geodesics or the behavior of stress–energy. Nevertheless, there could be other gravitational fields in the universe. Most alternative metric theories introduce such auxiliary fields; they could be scalar fields, vector fields, tensor fields, and so on. These fields may mediate the manner in which matter generates the spacetime metric, but they do not act back directly on the matter. Theories of gravity in which auxiliary fields act directly on matter are called *nonmetric* theories and typically violate the EEP. Experimental tests of that principle, described in Sec. 3.1, place tight constraints on such theories.

The prototypical example of an alternative metric theory of gravity is the Brans–Dicke theory, in which a scalar field exists in addition to the metric. This field plays the role of allowing the gravitational "constant" to vary in space and time. The differences in the predictions for observable effects between Brans–Dicke theory and GR depend on how strongly coupled the scalar field is to the metric—the weaker the coupling, the smaller the differences.

When one focuses on the weak-field, slow-motion regime that is appropriate to the Solar System (the so-called post-Newtonian limit because it includes the first corrections to Newtonian gravity) one finds that, in a broad class of metric theories, the metric looks the same, except for the numerical values of coefficients in front of various terms. A framework for studying metric theories in general has been developed, called the parametrized post-Newtonian (PPN) framework, in which parameters take the place of the numerical coefficients, parameters whose values depend on the theory under study. We shall encounter the use of some of these parameters in Sec. 3.2.1.

1.2.6 A Cosmological Term in Einstein's Equations. When we derived the form of Einstein's equations, Eq. (17) above, we reasoned that we needed a curvature tensor that had two indices and was divergenceless, and so we followed Einstein and used the Einstein tensor. But we remarked in Sec. 1.2.3.4 that there is another divergenceless tensor, the metric tensor itself. So it is possible to amend Einstein's equations to the following form:

$$\text{Einstein tensor} + \Lambda \times \text{metric tensor} = \kappa \times \text{stress–energy tensor}, \qquad (18)$$

which satisfies all the properties we needed for consistency and which will reduce to Newton's field equation in the appropriate limit provided the constant Λ is sufficiently small. Einstein introduced this modification into his theory when he found that his original equations did not admit static cosmologies, because the astronomical prejudices of his day held that the Universe was static and unchanging. Einstein found that he could choose a value of Λ that would indeed allow a static cosmological solution. He called Λ the *cosmological constant.*

There was no experimental evidence for Λ in Einstein's time, and when the expansion of the Universe was established a decade or so later, Einstein disavowed the cosmological term, calling it one of his biggest blunders. This is understandable: had he stuck to the original form of the equations, he would have predicted the expansion of the Universe and the associated Big Bang long before there was astronomical evidence for it.

Today there is a different view of Λ. If we carry the cosmological term over to the right-hand side of the equation, it then plays the role of an effective stress–energy tensor of the form

$$T^{\text{cosmol}}_{\alpha\beta} = -\Lambda g_{\alpha\beta}. \qquad (19)$$

Because $g_{tt} < 0$, this means that Λ is an effective energy density, constant everywhere. Provided Λ is much smaller than the mean energy density of the solar system, its effects will not have been noticeable. The values discussed today are much smaller. Interestingly, since the stresses are given by the spatial parts of Eq. (19), we find that the effective pressure is $-\Lambda$: if the energy density is positive, the pressure is negative.

Strange as this may look for ordinary matter, it seems almost inevitable in some schemes for quantizing gravity. When a classical field is quantized, there are renormalization terms that affect the residual energy density of the field (see Birrell and Davies, 1982). If one quantizes the vacuum gravitational field, the stress–energy tensor that survives renormalization will have to be proportional to the metric tensor, since there are no other tensors around, and no preferred frames. (This effect has an electromagnetic analog, called the *Casimir effect.*) It could happen, of course, that the cosmological term turns out to be zero; but there is no reason for this to be inevitable, and there is an expectation among many cosmologists today that the term may be comparable to the smoothed-out energy density of the present Universe.

2. SOME CONSEQUENCES OF EINSTEIN'S FIELD EQUATIONS

2.1 Momentum and Stress Also Make Gravity

The fact that all components of the stress–energy tensor contribute to gravity has important consequences.

2.1.1 Gravitomagnetism. In the Solar System, Einstein's equations will give the Newtonian potential in g_{tt}, using the mass density (tt component of the stress–energy tensor) as the source (Sec. 1.2.3.2). But the momentum density is also a source, so that *moving* bodies create new parts of the gravitational field. These couple to the velocities of particles, so they are called *gravitomagnetic effects.*

An important example is rotation. The angular momentum density of the Earth generates a metric term whose value is proportional to the angular momentum and which falls off as $1/r^2$, just as the electromagnetic potential of a rotating conductor does.

This term has a number of consequences. A free particle falling straight down will acquire a small motion in the same sense as the Earth's rotation. This is sometimes called the "dragging of inertial frames" or the *Lense–Thirring effect,* after its discovers. For the same reason, the orbital period of a satellite orbiting from west to east is slightly less than that of one going in the opposite direction. A spinning gyroscope in orbit will couple to this effect in much the same way that the spin of an electron in an atom couples to the magnetic moment of the nucleus, producing a precession of the spin.

These effects are weak compared to the dominant Newtonian orbital effects and hence hard to measure, for two reasons. First, the angular momentum density is small compared to the mass density, basically by a factor of v/c, which for the Earth's rotational motion is about 1.5×10^{-6}; the gravitomagnetic metric terms are therefore smaller than ϕ by this factor. But the effect is smaller still because, like magnetic effects, it only couples to the speed of the orbiting body, not simply to its mass. This introduces yet another factor of v/c. Such effects that are predicted by Einstein's equations but are of order $(v/c)^2$ smaller than the Newtonian effects are called *post-Newtonian* effects. Measuring them is one way of testing the validity of GR (Sec. 3.4.1).

2.1.2 Gravitational Collapse. In the same way, the stresses inside a body also make a contribution to gravity. These have an even smaller effect in nonrelativistic bodies like the Earth, being down by order $(v/c)^4$. But in a relativistic star, like the neutron stars we will describe in Sec. 2.4.1 below, they can be important. The dominant stress inside a star is the ordinary pressure of the fluid. In a neutron star, pressure and density are comparable, in the sense that $p/c^2 \sim \rho$. Such a large pressure has the effect of *increasing* the effective gravitational field inside the star. The result in GR is that there is actually a limit on how compact and relativistic a star can be, regardless of what it is made of: no star of mass M and radius R can have $GM/Rc^2 > 2.25$.

This leads to a very interesting conclusion. Suppose a star is right at the limiting size, and then it is perturbed, to make its radius marginally smaller. It cannot restore itself to equilibrium, and so it will collapse. Once it begins to collapse, it cannot stabilize itself at a smaller radius, for this would violate the inequality. This is called *gravitational collapse*: the unstoppable collapse of a star. The result of such collapse is a *black hole.*

2.2 Black-Hole Theory

2.2.1 Black Holes in the Eighteenth Century. Probably the most dramatic prediction of Einstein's equations is the existence of black holes. It may seem surprising, therefore, that essentially the same concept was discussed in the context of Newtonian gravity two centuries ago!

Black holes are regions of space in which gravity is so strong that light is trapped. The boundary of this region is the *horizon*: no light emitted from inside the horizon can reach the outside. The idea that light could be trapped by gravity occurred to eighteenth-century physicists as well. The first to suggest it was the amateur scientist John Michell (1724–1793), who in 1784 reasoned that, since light traveled at a finite speed, it would in principle be possible for an astronomical body to exist whose escape velocity was greater than the speed of light. Similar conclusions were drawn by Pierre Laplace in 1796.

Since the escape velocity is $(2GM/R)^{1/2}$, it follows that a body will trap light in Newtonian gravity if $2GM/Rc^2 > 1$. This is, remarkably, exactly the condition for the formation of a black hole in Einstein's theory as well. [Michell's suggestion was no fluke of speculation: it was Michell who suggested to his friend Henry Cavendish (1731–1810) that Cavendish should perform the famous experiment to measure the value of G that now bears his name.]

In the Newtonian picture, the trapping star is dark, because light cannot reach us from it, but its gravitational field is unchanged, and it can be discovered and identified by its gravitational effects. The same is true for the relativistic black hole.

2.2.2 Black Holes in General Relativity. The modifications needed to turn the Newtonian picture into a black hole in GR arise because of the special role played by the speed of light in relativity. In the Newtonian picture, light would still leave the surface of the trapping body, but it would slow down progressively and eventually turn around and fall back to the surface. Light emitted somewhat further out would escape, passing at some point the previous photon where it turns around. In relativity, all photons travel at the same speed. If a photon leaves the surface of the body, it cannot turn around later while other photons pass it, still moving outwards. Instead, there is a *horizon*: any photons outside the horizon can escape to infinity, while those inside are trapped.

Moreover, in the Newtonian picture, an interested astronaut could still in principle explore the dark star, getting away on a sufficiently powerful rocket. In relativity, nothing can go faster than light, so anyone or anything that crosses inside the horizon is trapped there forever.

There are no markers at the horizon, no signs saying "Point of no return". The horizon is defined as the boundary between what is trapped and what is not, but this boundary may not be known until the whole dynamics of the collapse that formed the black hole has occurred. The horizon is generally a place in the vacuum gravitational field, locally flat, in every way locally unremarkable.

2.2.3 Singularities Inside the Hole. We saw in Sec. 2.1.2 that a star with a radius smaller than 9/8 of the horizon size of a black hole of the same mass as the star cannot support itself against gravity. This means that the mass that has collapsed to form the black hole cannot find an equilibrium radius inside the hole: the material must continue to move inwards. The details of what happens to it are conjectural at present, but there is a set of *singularity theorems*, proved within GR by the mathematical physicists Roger Penrose and Stephen Hawking (see Hawking and Ellis, 1973), that show that matter obeying rather weak conditions (such as that its energy density should be positive) will always generate some sort of singularity inside the horizon. The nature of this singularity is, in general, poorly understood. Quantizing gravity (or even the matter fields) may change or eliminate the singularity, but it is hard to see how quantum effects could get outside the horizon and change the external aspect of the hole of macroscopic size.

It is also not clear whether gravitational collapse can lead to a singularity that is not inside the horizon. Such an event would prove much more damaging to notions of causality: singularities inside the horizon are confined, but those outside might make unpredictable effects on their surroundings. Such *naked singularities* are an active subject of research today. Penrose has formulated the *cosmic censorship hypothesis*, which asserts that naked singularities do not form. There is no proof of it at present.

2.2.4 Black Holes Have No Hair. Although the inside of a black hole may be messy, the external gravitational field is extraordinarily simple. A series of "no hair" theorems by Hawking, Brandon Carter, and others (see Wald, 1984) has shown that black holes, once formed, quickly radiate away everything possible and settle into a stationary state characterized entirely by values of the conventional conserved quantities of physics: total energy (the mass of the hole), total angular momentum, electric charge, and magnetic monopole moment (should monopoles exist). These properties of the hole are all measurable externally: the mass from orbital periods of satellites, the angular momentum from the Lense–Thirring effect (Sec. 2.1.1), and the charge from electrostatic attraction.

Higher multiple moments of the electromagnetic or gravitational field that might have characterized the collapsing star are carried away from the hole in a burst of radiation that accompanies its formation. The total number of baryons and leptons that formed the hole is not measurable from outside, because the nuclear forces have short range.

The metric of the spherical uncharged black hole was the first exact solution found for Einstein's equations, derived by Karl Schwarzschild (1873–1916) in 1916. It was not completely understood until the 1960s, after work by Martin Kruskal, Peter Szekeres, John Wheeler, and others. (See Misner *et al.*, 1973, for references.) The rotating hole is described by the Kerr metric. The term "black hole" was coined by Wheeler in the 1960s.

2.2.5 Black-Hole Thermodynamics. Forming a black hole entails a considerable loss of information: by the "no hair" theorems, most of what goes down the hole is lost forever. It is not surprising, then, that one can assign an *entropy* to the black hole. In fact, the whole of thermodynamics can be generalized to include black holes, and *black-hole thermodynamics* is one of the most beautiful parts of black hole theory (see Thorne *et al.*, 1986). The entropy of the hole is proportional to the surface area of its horizon. A theorem of Hawking establishes that this area cannot decrease with time (Hawking and Ellis, 1973).

There is a corresponding temperature, which is proportional to the surface gravity of the hole.

This temperature is the most remarkable black-hole property of all. In black-hole thermodynamics, it plays a formal role in determining the energy (mass) change associated with a change in the entropy (surface area) of a black hole: $dM=TdS$. But in ordinary thermodynamics, objects also exhibit their temperature by emitting blackbody radiation. Despite the fact that black holes trap all photons, they also give off a *thermal radiation* whose spectrum is exactly that of a blackbody of the thermodynamic temperature. This is called the *Hawking radiation* (Hawking, 1974).

Hawking's demonstration of this radiation relies on quantum mechanics, but it requires only the quantum field theory of the electromagnetic field near the hole, not any quantum theory of gravity. The Hawking temperature is inversely proportional to the mass of the hole, and for a hole that has the mass of the Sun, it is entirely negligible. This means that the radiation is likely to be exceedingly difficult to observe. Its existence is accepted, however, partly because well-established techniques of quantum field theory predict it, and partly because it allows black-hole thermodynamics to fit so beautifully into standard thermodynamics.

Black holes have other macroscopic properties as well (Thorne *et al.*, 1986). For example, the horizon has a finite electrical conductivity. This comes essentially from the "no hair" theorems as well: if a charge is lowered near the horizon, it will have to look from a distance as if it is smeared uniformly over the horizon; otherwise there would be an electric dipole moment. The horizon therefore effectively conducts charge along its surface.

2.2.6 Wormholes. The simplest black-hole solution, the one found by Schwarzschild, describes a stationary black hole, i.e., one that has existed for all time. It has an interesting topology: it actually involves *two* "exterior" spaces connected by a throat, or wormhole. Although the gravitational field outside the hole in both exterior regions is static, the throat is not. It opens up to a maximum size and then closes off in a finite proper time. In fact, it closes off so fast that a particle falling into the hole would not have time to get through it: the wormhole is not a channel for communication between the two exteriors. Moreover, a black hole formed by the collapse of a star has a simpler topology, with no wormhole, and only one exterior region.

Nevertheless, wormholes have intriguing properties. If physicists could manipulate quantum fields to create and maintain a region of negative energy (as happens in the Casimir effect referred to in Sec. 1.2.6), then they could in principle maintain a wormhole open long enough for a particle to travel through it. If the two ends of the wormhole are open in the same space, then a particle could circle through the throat many times. If, in addition, the ends of the wormhole were made to move relative to one another, then, as Thorne (1991) has shown, the time dilation effect of special relativity could be arranged to allow the particle to travel backwards in time.

The realization that GR allows such behavior, even under what present technology suggests are implausible conditions, has created considerable interest. If such *closed timelike curves* exist, they raise questions about causality and especially the consistency of initial-value formulations of the law of physics, and they have to be taken into account in formulations of quantum gravity. The study of these rather exotic conditions may in the future shed new light on fundamental physics.

2.3 Gravitational Waves

2.3.1 The Necessity of Gravitational Waves. Gravitational waves (GWs) are present in any relativistic theory of gravity. When fields are weak and motions slow, the relativistic equations must reduce to Newton's field equation, Eq. (11). This involves the differential operator ∇^2. If fields are weak but motions may be rapid, then one expects the theory to be relativistically invariant. The operator ∇^2 will be replaced by its relativistic generalization

$$-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}+\nabla^2 .$$

This is the wave operator, with the characteristic wave speed c. We should expect wave effects in relativistic gravity, propagating at the speed of light. Different theories of gravity will differ in the details of the waves, but GWs will be present in all of them.

In GR, the waves bear a striking mathematical resemblance to electromagnetic waves. They are transverse, have two independent polarizations, and fall off in amplitude as $1/r$. Electromagnetic waves do not carry monopole radiation: for slow-motion sources they are dominated by dipole emission. Gravitational waves in Einstein's theory do not carry either monopole or dipole radiation, but are dominated by quadrupole emission (Sec. 2.3.3).

2.3.2 The Interaction of Gravitational Waves with Matter. Waves can only be detected through their time-dependent tidal forces. If a wave passes a single particle, one can always choose comoving coordinates, so that the particle remains at a fixed coordinate position. By the EEP, the particle will not feel anything locally from the wave: bowls of soup will not spill, pocket calculators will work normally. The particle will only be able to detect the wave if it looks at other things sufficiently far away for the tidal effects of gravity to be noticeable.

The absence of single-particle effects of GWs led to much misunderstanding in the early development of GR. The intuition developed in electromagnetism, where single charged particles can detect electromagnetic waves, was not helpful. Many relativists, including at times Einstein himself, believed that GWs were a mathematical illusion. Work in the 1950s and 1960s by Hermann Bondi, Richard Isaacson, Penrose, Joseph Weber, and many others showed that to detect GWs one needs at least two particles, but that the waves are very real, transferring energy and angular momentum away from their sources to their detectors. [See Misner *et al.* (1973) or Schutz (1985) for more details.]

The simplest way to detect a GW is to monitor the proper distance between two nearby free particles, particles that have no external forces on them. The distance can be monitored, for example, by measuring the round-trip light-travel time between the particles. A GW can affect this distance. Gravitational waves are transverse, and so if the wave moves in the z direction, then only distances in the x-y plane will be affected.

The pattern of distance changes produced by a given wave is shown in Fig. 1. The dots represent a ring of free particles. The circles and ellipses are not physical connections: they only illustrate the pattern of the placement of the particles. The first ring is the undisturbed position: a circular ring of particles. The second diagram shows the effect of a GW on proper distances. It has lengthened the x axis of the circle and shortened the y axis by the same fraction, preserving the area of the ring. When the wave reaches its opposite phase (final diagram), the effects on the axes are reversed. At the bottom of the diagram are

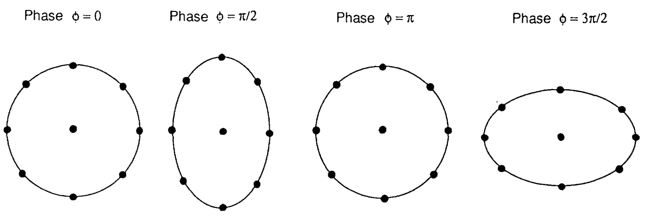
Polarization of a Gravitational Wave

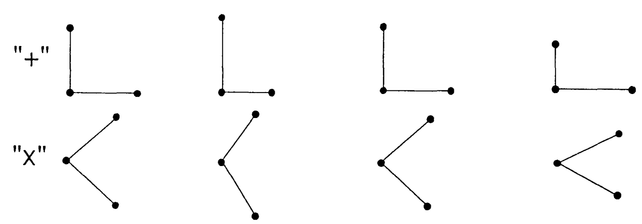
Response of a three-mass gravitational wave "detector" in two orientations

FIG. 1. Illustration of the action of a GW on a ring of free particles transverse to the direction of propagation.

shown the lines joining certain particles to the center of the ring. The top lines are for the particles at 12 o'clock and 3 o'clock (particles oriented in the "+" configuration relative to the ellipse's principal axes). Their lines remain perpendicular but change their relative length. The bottom lines are for the particles at 1:30 and 4:30 (particles oriented in the "×" orientation relative to the ellipse's principal axes). Their lines remain the same length but the angle between them changes.

Nothing is preferred about the x axis in this diagram. The entire ellipse can be rotated about the z axis to give possible patterns. A rotation of 90° effectively carries the pattern back into itself, which is characteristic of a tensor (spin-2) wave. The second independent polarization is obtained by a rotation of 45°.

Because tidal effects grow linearly with distance, *the shape of the ellipse in Fig. 1 is a property only of the wave, independent of the size of the ring.* The relative change in the axes is, therefore, a measure of the amplitude of the wave itself. It is conventionally called h, and can be defined from Fig. 1 as

$$h := 2(\delta l)_{\max}/l \,, \tag{20}$$

where l is the length of the semimajor axis of the ellipse. (The factor of 2 is conventional.) The relative distortions shown in the figure are, of course, exaggerated in order to make them easy to see. Modern GW detectors are being built to detect amplitudes as small as 10^{-22}. We shall consider the technical challenges of this in Sec. 4.2.

This tiny amplitude means that the effect of matter on the gravitational wave will also be of this order; the wave will lose or scatter only a fraction of order 10^{-22} of its energy as it passes through a detector. By extension, this means that the waves that arrive at our detectors have lost little of their original form: they carry information from their sources uncorrupted by scattering or absorption. In this they are a unique carrier of information in astronomy; even the neutrinos detected from the supernova event in 1987 (SN1987A) were from a thermal distribution that had scattered many times before leaving the collapsed core (Sec. 2.4.1.4). Traveling at the speed of light, gravitational waves follow geodesics through the Universe; they can be gravitationally lensed, but scattering and absorption are negligible.

2.3.3 Wave Emission: The Quadrupole Formula. Sources of gravitational waves respect the conservation laws for energy and momentum that are built into Einstein's equations. For weak GWs, these have a similar effect to that produced by the conservation of charge in electromagnetism, which ensures that there are no monopole electromagnetic waves: the lowest order of radiation is dipole. In GR, the law of conservation of energy similarly ensures there is no spherical gravitational radiation. Any oscillating spherical mass leaves its exterior gravitational field undisturbed in GR (just as in Newtonian gravity), and so it does not radiate GWs. Conservation of momentum has the same effect on dipole radiation: there is no dipole gravitational radiation.

From a fundamental point of view, charge conservation (and its consequence, the absence of monopole radiation) in electromagnetism follows from the gauge invariance of the theory. (See SYMMETRY AND CONSERVATION LAWS.) In GR, energy–momentum conservation (and the absence of monopole and dipole radiation) follow from the general coordinate invariance of the theory.

At quadrupole order, the formulas for the radiation emitted by a slow-motion system are remarkably similar to those for electric quadrupole radiation in electromagnetism. (For highly relativistic, fast-motion systems, the formulas below can only be used approximately.) We define the quadrupole moment tensor of the mass distribution:

$$I_{jk}(t) = \int \rho(t,\mathbf{r})\, x_j x_k\, d^3x \,, \tag{21}$$

where ρ is the mass density (rest mass dominates in a nonrelativistic system), and where we follow the usual convention of employing Latin indices for purely spatial values (1,2,3). This equation can only be used in a Cartesian spatial coordinate system. The trace-free or reduced quadrupole moment is derived from this by

$$\rlap{-}I_{jk} := I_{jk} - \tfrac{1}{3}\,\delta_{jk} I \,, \tag{22}$$

where I is the trace of I_{jk}.

The amplitude of the radiation at a distance r from the source is roughly

$$h \sim (4G/c^4)\,(\ddot{\rlap{-}I}/r) \,, \tag{23}$$

where $\ddot{I}$ stands for the second time derivative of a typical component of the quadrupole moment. If the motions that are responsible for the time derivatives are driven by the internal gravitational fields of the source, which is the case for all the realistic sources of GWs that we will consider later, then there is a convenient upper bound on this amplitude, given by

$$h \leqslant (GM/rc^4)\phi_{\text{internal}} , \tag{24}$$

where ϕ_{internal} is the typical size of the Newtonian gravitational potential within the source (Schutz, 1984).

Although the local conservation laws of energy and momentum eliminate the monopole and dipole radiation, it does not follow that GWs remove no energy or angular momentum from the source. Unlike electromagnetism, which is a linear theory, GR is nonlinear, and GWs can act as sources for gravity. So when they leave the system, there is a gradual decrease in its energy. This is compensated by a well-defined energy flux carried by the waves (Isaacson, 1968). Energy from this flux can be transferred to other systems, such as GW detectors. The GW luminosity (Landau and Lifshitz, 1962) is probably the most useful of the quadrupole formulas:

$$L_{\text{GW}} = \frac{G}{5c^5} \left\langle \sum_{j=1}^{3} \sum_{k=1}^{3} |\dddot{I}_{jk}|^2 \right\rangle , \tag{25}$$

where the angle brackets denote an average over one period of the motion of the source. Notice that this depends on the square of a third time derivative, just as the analogous formula for electric quadrupole radiation does. It is therefore very sensitive to the size of the velocities inside the source. As a source becomes more relativistic, the power radiated goes up very rapidly, as $(v/c)^6$.

2.4 Applications of General Relativity

For its first 50 years, GR was mainly the property of mathematicians and theoretical physicists who valued it for its intellectual challenge and its aesthetic beauty. It seemed to have little practical relevance to the rest of physics. But astounding discoveries in astronomy, beginning with the realization in 1963 that quasars have enormous luminosities, attracted physicists to the challenge of explaining the new phenomena, and they found that they needed to use GR as an everyday tool in making models of astronomical phenomena. In this section we describe some of the ways that GR is used. For more details of many of these phenomena, the reader is referred to the article on ASTROPHYSICS.

2.4.1 Relativistic Stars (Pulsars) and Gravitational Collapse

2.4.1.1 Stellar Evolution. An ordinary star nearing the end of its nuclear-burning lifetime undergoes many changes as it switches from its primary fuel (hydrogen) to short-lived secondary sources of energy (helium, then carbon, and so on). When it finally exhausts its nuclear fuel, it can no longer generate energy to replace that which it radiates from its surface, with the result that gas pressure can no longer hold the star up against gravity. The star then has only two options: it can rely on *degeneracy pressure* (the quantum-mechanical resistance of identical fermions to being squeezed into too small a volume), or it can collapse to a black hole. Which option it takes depends on its mass.

2.4.1.2 White Dwarfs. If the mass of the star when it reaches this point is less than about 1.4 times the mass of the Sun, then it can end its days quietly, supported by electron degeneracy pressure. It is called a *white dwarf*, and has a radius roughly that of the Earth, about 10^4 km.

However, Chandrasekhar (1939) showed that electron degeneracy cannot support more than about $1.4M_{\odot}$. The burnt-out core of a star that is larger than this will inevitably grow to exceed this size, and then it must gravitationally collapse. During the collapse, which lasts less than a tenth of a second, essentially all electrons and protons combine to form neutrons. This process is called *neutronization*. When the density is high enough for neutron degeneracy pressure to become effective, the collapse will halt, provided again that the star is not too massive. This *neutron star* has the density of a typical atomic nucleus, so that a $1M_{\odot}$ star has a radius of about 10 km.

2.4.1.3 Neutron Stars. We have met neutron stars (NSs) in Sec. 1.1.2.1. The upper mass limit of NSs is sensitive to poorly understood high-density nuclear physics, and to the details of the relativistic theory of gravity. It is probably between $2M_{\odot}$ and $3M_{\odot}$ in GR (Hartle, 1978).

The result of halting the collapse is a bounce; a shock travels outwards through the rest of the envelope of the star, blowing it away in an explosion called a *supernova*. Since the collapsing core originally had a rest mass of $1.4M_\odot$, one expects the supernova to leave behind a NS of about that rest mass or a little more. Because the gravitational redshift reduces all energies as measured from far away, and because the strong equivalence principle (Sec. 3.2.3) ensures that all energies contribute to gravity, it follows that the gravitational mass of the star as measured by orbits in its gravitational field will be some 10% or so less than its rest mass.

2.4.1.4 Black Holes. Collapse to a NS is likely to be complicated by a number of factors. One is rotation. If the collapsing core has as much angular momentum as the Sun (which is a slow rotator), centrifugal effects will halt the collapse before neutron-star densities. Smaller amounts of rotation may therefore have significant dynamical effects. Work is currently in progress to model rotating collapse on computers.

Another complication is neutrino physics. The collapse generates many neutrinos, first from the neutronization reaction $p+e\rightarrow n+\nu_e$, but more copiously from the thermal equilibrium reaction $\gamma+\gamma\leftrightarrow\nu+\bar{\nu}$ that develops as the collapse halts. These appear to exert a crucial pressure to power the shock outwards. The neutrinos detected from the nearby supernova event in 1987 (called SN1987A—see ASTROPHYSICS) came primarily from the thermal distribution.

Whatever the complications, it seems that on occasion the shock does not blow away the whole star, and enough further matter falls on the collapsed core to push it over the upper mass limit. It must then collapse further, and it forms a black hole (BH). In the next section we will see that there is good evidence for a number of $10M_\odot$ BHs in our Galaxy.

2.4.1.5 Pulsars. When a NS is formed, it is likely to be rotating. By what seems like a great stroke of luck, we are able to observe this rotation in hundreds of cases. During the collapse, not only is the rotation speed amplified, but so is the magnetic field of the star. It happens that the magnetic and rotation axes do not line up; in fact, they seem to prefer to be perpendicular to each other. The result is that rotation carries the magnetic poles around. Charged particles spiral around the magnetic field lines and crash into the star at its magnetic poles, in a colossal version of our terrestrial auroral displays. The result is a beam of radiation that emerges from the poles and is spun by the star's rotation in the same way a lighthouse beam turns. If we happen to sit in the direction of the beam, we see the radiation (radio, optical, and x-ray) turning on and off. This is a *pulsar*.

Pulsar periods must be longer than the orbital period at the surface of a NS, which is about 0.5–1 ms. In fact, the fastest known pulsar has a 1.6-ms period. However, all pulsars that we know are young (for example, those that can be identified with supernova explosions whose remnant clouds of gas are still visible) are rather slow rotators. For example, the pulsar in the Crab nebula spins "only" 30 times per second. We now believe that this is typical of new NSs, and that the really rapid rotators have been spun up by accreting gas from a companion star in a binary system (Sec. 2.4.2).

Pulsars in binaries are not uncommon, but they are hard to find because the orbital motion keeps changing their apparent pulse period. However, binary pulsars are prized discoveries, because by monitoring the Doppler shifts of the period one gets dynamical information about the system. In a few cases, we get enough information to determine the masses of the individual stars, as we describe in Sec. 3.3. Interestingly, all six neutron stars with well-determined masses are within 10% of $1.4M_\odot$.

2.4.2 Black Holes in X-Ray Binaries and in Galactic Centers

2.4.2.1 X-Ray Binaries. When gravity overwhelms neutron degeneracy, the star must collapse. Even though high-density nuclear physics is poorly understood, simple constraints (such as that the speed of sound in nuclear matter must be less than the speed of light) are enough to guarantee that nuclear physics cannot support more than about $3M_\odot$ at these densities (Hartle, 1978). This implies that a compact object with a significantly larger mass, such as $7M_\odot$ or $10M_\odot$, must be a BH. Since BHs do not emit observable radiation themselves, practically the only way we can identify holes of this size is by their effects in binary systems.

When NSs or BHs orbit ordinary stars, mass can sometimes flow from the companion onto the compact star. The angular momentum of the infalling gas forces it into a disk around the star, and friction in the disk leads to a more or less steady accretion of matter onto the compact object. As we noted above, NSs can be spun up by this process. Black holes also spin up, but this may not be observable. But the hot accretion disk is itself a source of observable radiation: x rays.

One of the big surprises of the early 1970s was the discovery, by x-ray satellite observatories, that our Galaxy has dozens of such binary x-ray sources. If the companion can be identified by optical observations and its mass estimated from its spectrum, then the mass of the compact object can be estimated. In this way, we now have three or four fairly secure identifications of BHs. The best identification, in a system called V404-Cyg, has a minimum mass of $8M_\odot$ (Casares *et al.*, 1992).

This certainly underestimates the number of BHs in the Galaxy. Most mass determinations are uncertain by 50% or so, and, given the uncertainty in the upper mass limit of NSs, there are a number of binaries where we are unable to be sure about the nature of the compact object. Therefore, while we are confident that about 10% of the stars in the Galaxy are white dwarfs and 1% are NSs, the fraction that are BHs is very uncertain. X-ray binary observations suggest a plausible lower limit of about one in 10^4.

2.4.2.2 Galactic Black Holes. Giant BHs seem to be relatively more common than this: it is possible that every normal galaxy has in its center a BH of mass between $10^4M_\odot$ and $10^{10}M_\odot$. These would probably have been formed at about the same time as the galaxy itself formed, either from a dense gas cloud in the center or from the collective collapse of a swarm of NSs and small BHs formed near the center. Recall that the density of a $10^8M_\odot$ cloud of gas when it forms a BH is only the density of water.

For general information about galaxies, the reader is referred to the article on GALAXIES AND COSMOLOGY. The evidence for BHs in the centers of galaxies is manifold, and in many ways better than that for individual x-ray binary BHs. The strongest evidence is velocity information. By measuring the Doppler shifts of spectral features in many wave bands, astronomers have accumulated evidence for rotation, collapse, and/or expansion with velocities of typically hundreds of km/s inside volumes no larger than a parsec (3×10^{16} m, which is a typical distance between individual stars in the neighborhood of the Sun). Using the order-of-magnitude relation $GM/R=v^2$ gives masses of the order we have quoted (Rees, 1990). There do not seem to be any mechanisms that would allow such concentrations of mass to remain stable for long without collapsing to BHs. The inference is that the collapse has already occurred in most cases, and we are seeing a small amount of gas swirling around in the resulting gravitational field.

Another line of evidence also points to BHs: the jets of relativistic particles that are expelled from the centers of galaxies. Some jets are narrow (opening angles of less than 1°) and run straight for millions of light years, which argues that their emission has been constant over millions of years. Jets are often found in opposing directions. The phenomenon is ubiquitous and at the same time highly variable. Some galaxies seem quite ordinary except for the immense radio emission that takes place at the ends of the jets, far outside the galaxies themselves. Other galaxies contain *quasars*, which are very small sources of immense luminosity, from which emerge jets (Blandford *et al.*, 1982). Quasars in particular seem to be associated with young galaxies: a far larger fraction of galaxies exhibited the quasar phenomenon when the Universe was $\frac{1}{3}$ of its present age than today.

These facts fit the BH model too, chiefly because only BHs seem to be able to form stable centers of activity, to supply sufficient energy, and (through their rotation) to provide a consistent direction for jets. Gas probably forms an accretion disk around a massive BH, just as it does in binary systems. Something then forms jets of relativistic particles and expels them perpendicular to the disk in both directions. That something probably involves twisted magnetic fields, but it may also involve the BHs directly. Since BHs have an electrical conductivity (Sec. 2.2.5) and will in general rotate, they can interact with magnetic fields to produce large voltage differences from pole to equator, which can accelerate and indeed even create pairs of charged particles. The energy would come from spinning down the rotating BH (Blandford and Znajek, 1977). Modeling the quasar and jet activity of galaxies is one of the most active areas of theoretical astrophysics today.

2.4.3 Gravitational Lensing.

In Sec. 1.2.2.5 we saw that gravitational fields bend light, and that this can lead to a form of lensing. The amount of deflection depends on the theory of gravity, but the qualitative features are the same in all theories. Lensing was predicted more than 50 years ago, but the phenomenon does not seem to have been taken very seriously by astronomers until the now-famous "double quasar" 0957+561 was discovered in 1979 (Walsh *et al.*, 1979). Since then, dozens of examples of gravitational lenses have been found.

Gravitational lensing is one of the most dramatic confirmations of relativistic gravity. Lensed images come in many forms. Some are simply multiple images of a single point-like source, each with a different magnification. (The number of such images must always be odd, although they may not all be bright enough to be observed.) Sometimes images are smeared out into arcs. Some images do not change position but are "microlensed" by the gravitational field of a single star of an intervening galaxy that happens to move across the image, magnifying it briefly.

Lensing takes place at all scales, and to some extent in every astronomical image, but to be observable it must stand out from the confusion of other properties of images. The ideal situation is a compact image that is lensed into two or more similar images separated by a few arc seconds, far enough apart to be distinguished but close enough to be noticed as unusual. Spectra provide the "fingerprints" that convince one that the images are of the same object and not neighboring distinct objects. The lensing mass is typically a cluster of galaxies somewhere between the source and us.

Unless the galaxies in the cluster can all be seen, it is hard to model the lens. However, modeling has potential rewards. The biggest is to predict the Shapiro time delay (the excess travel time, as described in Sec. 3.2.1.2) along each of the image paths. Since the only data we have about the distance to and thus the linear size of the lens and the source are the redshifts produced in their spectra by the cosmological expansion, we need the *Hubble constant* (Sec. 2.4.5) to predict the actual time delays. Conversely, the measurement of the time delays between various images of a variable source in a well-modeled lens can provide a determination of the Hubble constant, and from it an estimate of the age of the Universe. This is a very active area of research today.

2.4.4 Solar System and Stellar Orbits.

General relativity makes small corrections to the orbits of bodies in the solar system. The effect may be described by a small $1/r^3$ correction to the gravitational acceleration produced by the Sun. It is smaller than the Newtonian term by the ratio of the Sun's gravitational radius to r, which is of order 10^{-7} or less for the planets. Its effect is to make the planet's ellipsoidal orbits *precess*: the direction from the Sun to the point of a planet's closest approach to the Sun (its *perihelion*) rotates with time.

The precession of Mercury's perihelion provided an early confirmation of GR, as we describe in Sec. 3.2.2. But at 43 seconds of arc per century, it is hard to measure, since the gravitational fields of other planets produce similar but much larger effects on Mercury's orbit. In fact, the measurement of the relativistic precession of any other planet is hopeless, partly because of the dominance of planetary perturbations.

When stars orbit each other, there are similar effects. They are easiest to observe in binaries containing pulsars, where we have excellent dynamical information about the orbit. Indeed, the periastron shift (as it is called for binaries) is necessary if we are to determine the masses of the individual stars. Fortunately, the orbits of such stars bring them much closer to each other than Mercury gets to the Sun, so the effect is larger. For the pulsar binary system PSR1913+16 (Sec. 3.3), the rate of periastron advance is 4.2° per year, easily measurable.

Another dramatic effect of GR on the orbits of stars is caused by the emission of GWs by the orbital motion. The energy lost to waves brings the stars gradually closer together, and also tends to make orbits more circular. As the stars approach, their orbital period decreases, and the system actually speeds up. This effect has been observed in PSR1913+16. It is also thought to control the evolution of certain classes of binaries, where mass is transferred from one star to another when they get too close. The transfer pushes the stars apart, but they come closer together because of gravitational radiation reaction. The net effect is a steady transfer of mass from one star to another.

2.4.5 Cosmology, Inflation, and the Origins of the Universe. General relativity opened the door to the study of cosmology, the Universe in the large. In Newtonian gravity it is not even possible to formulate the gravitational force consistently for a Universe of infinite size, and if one takes the Universe to be finite but large, then the exact shape of its boundary affects the gravitational field everywhere in a measurable way. General relativity gives a consistent picture of the Universe. The reader should consult the article GALAXIES AND COSMOLOGY for a complete survey of cosmology; we concentrate here on the general relativistic aspects of the study.

2.4.5.1 Cosmological Models. Observations suggest that we can make a good first approximation to describing the Universe by assuming that it is homogeneous on the largest scales and isotropic around us. Since the Universe is also expanding, these assumptions imply the existence of a special choice of time everywhere: "now" has to be chosen so that the density, say, is the same everywhere at that time. This choice of time is unique: the Universe has a preferred reference frame. When cosmologists speak of the distance between galaxies, they are referring to proper distance at a particular time in this preferred frame. Similarly, the age of the Universe is the proper time since the Big Bang, as measured in this frame.

There are only three possible geometries for the spaces of constant time in a homogeneous and isotropic cosmology: flat, open, and closed. In the flat model, space at a constant time is Euclidean. In the closed model it is a three-sphere, which is the locus of all points equidistant from a fixed point in four Euclidean dimensions (equidistant as measured by the four-dimensional Euclidean distance). In the open model it is a three-hyperbola, which cannot be described as a subsurface of Euclidean space; rather, it has the same geometry as the hyperbola in Minkowski spacetime that consists of all points at a given timelike interval from the origin.

Without a cosmological constant, the field equations of GR imply that all models have a singularity at some time: they expand from a singular point of infinite density, contract to one, or both (the closed model). Since we observe the Universe to be expanding (see Sec. 2.5.4.2), we infer that there was a singular point a finite time in the past. This is the *Big Bang.*

If the initial expansion velocity was sufficiently great compared to the overall gravitational attraction, the expansion will continue forever. The critical measure here is the local density of mass–energy. It turns out that models that expand forever have hyperbolic space sections; those that recontract are three-spheres; and the marginal case is flat. Therefore, measuring the actual mass–energy density has implications for the large-scale structure of the Universe.

Although much is sometimes made of the philosophical implications of this, it must be emphasized that we can only observe a small section of our Universe, and any assertion about the large-scale topology of space depends on our assuming that the parts of the Universe that are too far away to observe are in fact identical to the section we can observe. This can, of course, never be proved.

2.4.5.2 Hubble's Law: The Big Bang. The expansion velocity of the Universe is given by the Hubble constant. Edwin Hubble was the first to show, by painstaking systematic observation of many galaxies, that all distant galaxies recede from us, and more distant galaxies recede faster (see GALAXIES AND COSMOLOGY). He expressed this as a linear relation between the distance d of a galaxy and the recession velocity v:

$$v = Hd\,. \tag{26}$$

We call the proportionality constant H the Hubble constant. This law is what one expects in a Universe that expands homogeneously: in any given interval of time, galaxies twice as far away should have moved by twice as much, leading to twice the recession velocity.

A frequently suggested analogy is an expanding balloon. If one draws dots on the balloon and measures their separations, then when the balloon is twice as large the points will have receded from one another in exactly this way. The balloon is also an excellent two-dimensional analog of the geometry of the three-spherical (closed) universe.

2.4.5.3 Cosmic Microwave Background Radiation. The Big Bang picture provides a simple model for our history. At a finite time in the past, the Universe was infinitely dense. It began expanding and maintained thermal equilibrium among all the exotic species of particles for a short time. Gradually, as the temperature dropped, different species froze out. When nuclei froze out, about 20% of the mass was in helium, the rest in hydrogen. Later, hydrogen atoms froze out and the expanding cloud of gas became neutral. The mean free path of photons increased dramatically, and most photons in existence at the time have not scattered since. This epoch of "decoupling" of photons and matter left behind a thermal distribution of photons that has remained thermal as the Universe has expanded. The temperature of this distribution is inversely proportional to the expansion. Today this temperature is about 2.7 K, and the resulting *microwave background radiation*, discovered by Penzias and Wilson (1965), is one of the strongest pieces of evidence favoring the Big Bang model.

2.4.5.4 The Age of the Universe. The measurement of the Hubble constant is fraught with difficulties. The velocity v is easy to obtain from spectra, but measuring the distance is hard. Astronomers look for "standard candles": objects of known intrinsic brightness, whose apparent brightness can be used to infer their distance. No really reliable standard candles exist, and even today astronomers differ by up to a factor of 2 in their determinations of H. In the astronomers' customary units, where distances between galaxies are measured in megaparsecs ($1\ \mathrm{Mpc}=3\times10^{22}\ \mathrm{m}=3\times10^{6}$ light-years) and velocity is measured in km/s, H is likely to lie in the following range:

$$45\ \mathrm{km\,s^{-1}\,Mpc^{-1}} < H < 100\ \mathrm{km\,s^{-1}\,Mpc^{-1}}\ .$$

It is usual to parametrize H by $h=H/(100\ \mathrm{km\,s^{-1}\,Mpc^{-1}})$.

If the expansion were at a constant rate, then the age of the Universe (time since the Big Bang) would be $1/H$. If $h=1$, then $1/H=10^{10}$ yr. Smaller values of H lead, of course, to older Universes. In fact, gravity has been slowing down the expansion rate, so $1/H$ overestimates the age of the Universe. Therefore one can also constrain H by measuring the age of, say, the Earth, or better, of an old (hopefully first-generation) star or cluster of stars. The Earth seems to be about 4.5×10^{9} years old, but some clusters may be older than 10^{10} years.

2.4.5.5 Mass and Missing Mass. Once we know H we know the so-called critical density, the density of mass–energy just needed to stop the expansion, to make the Universe flat. In terms of h, the critical density is

$$\rho_c=2\times10^{-26}h^2\ \mathrm{kg\,m^{-3}}\ .$$

Measuring the actual mass-energy density is even harder than measuring H; the uncertainty seems to be a factor of 10 at present. The problem is that the mass that is radiating light, mainly in the form of stars, accounts for less than 1% of the critical density, and there is much evidence for at least 10 times this amount of dark or hidden mass. The evidence is dynamical: the rotation rates of spiral galaxies and the internal motions of large clusters of galaxies are simply much too large for these objects to stay together if their masses were only what we could infer from the light they emit. This dark mass may consist of BHs or other low-brightness remnants of the early evolution of stars, or (and this is the more popular idea at present) it may consist of a cosmological distribution of elementary particles of an unknown kind (Knapp and Kormendy, 1986). Searches are now under way for such particles randomly passing through the laboratory. Their detection would have important implications for high-energy physics as well as for astronomy.

Indirect evidence of this dark matter may be sought in the dynamics of galaxy formation. It turns out that density perturbations in ordinary baryons do not grow rapidly in the early Universe, mainly because these particles are charged and couple too strongly to the equilibrium radiation field. Once the Universe has expanded and cooled off enough for hydrogen to become neutral, the perturbations may grow, but that does not leave enough time for galaxies as we observe them—well separated objects even at early times—to have evolved.

A pervasive background of neutral particles could provide seeds for galaxy formation that grew unhindered in the early Universe. If the particles had reasonably large rest mass, they would cool off rapidly and form strong condensations. Computer simulations performed within this *cold dark matter* hypothesis provide reasonable consistency with the present distribution of galaxies and their clustering. It must be stressed, however, that there are many other possibilities for providing seeds for galaxy formation. Cosmic strings, mentioned below, are also attractive. The real story of galaxy formation may involve several such factors.

2.4.5.6 Microwave Anisotropies. The microwave background ought, if the Big Bang is correct, to be a very good black-body distribution. Indeed, measurements by the recent COBE spacecraft show it to be (Smoot *et al.*, 1992). However, its temperature need not be the same in all directions. For one thing, the Earth may be moving with respect to the cosmological rest frame, the mean rest frame of the matter at decoupling. This would produce a dipole anisotropy in the temperature, and this has been measured. The inferred velocity of the Sun is 370 km s^{-1}, again from the COBE measurements.

In 1992 COBE also measured another effect on the microwave temperature: fluctuations that arise from the early density irregularities of the Universe. While not surprising, these provide striking "baby pictures" of the condensations that eventually evolved into large aggregates of clusters of galaxies.

2.4.5.7 Inflation and the Homogeneity Problem. The measurements performed by COBE highlight a conceptual problem for cosmology: the structures seen by it were on a scale so large that light could not have traveled across the structure in the time between the Big Bang and the time of decoupling, when the radiation was produced. Yet the structures seen by COBE were relatively consistent right around the sky. How could they be causally related? And if they were not causally related, why are they so similar?

The same problem exists for other measurements. We can see quasars in opposite directions on the sky that are today so far apart that they could not have communicated since the Big Bang, and yet the distributions of quasars and other galaxies near them look similar at both places. How indeed can the Universe be homogeneous over distances that could not communicate?

A simple answer is that it is all in the initial conditions of the Big Bang. But such an explanation is unsatisfying to many scientists, and an attractive alternative has been proposed: inflation. If, in the early universe, just after the Big Bang, the laws of high-energy physics at energy scales inaccessible to present-day experimentation were of a certain form, then the Universe could at some point have undergone a phase transition that pushed it into very rapid expansion, much more rapid than one would infer by working backwards from today's expansion. The expansion would in fact have been exponential, hence the term *inflation*. If it halted after many *e*-foldings, then parts of the Universe that now seem too far apart to have communicated would in fact have been much closer before inflation, and could have been causally connected. Inflation can solve, at least in this sense, the homogeneity problem (Guth and Steinhardt, 1984; Blau and Guth, 1987).

Inflation has side benefits, one of which is that it might produce exactly the spectrum of seed perturbations that COBE saw. Observational evidence is slim at present, but by the year 2000 one can expect much stronger constraints on inflationary models.

3. TESTS OF GRAVITATIONAL THEORIES AND THEIR TECHNOLOGICAL DEMANDS

For the first half-century of GR's existence, experimental tests were infrequent, largely because the smallness of the predicted effects made them extremely difficult to measure with any accuracy. The validity of the theory rested on two tests, the deflection of light, measured in 1919, and the resolution of an anomaly in Mercury's perihelion advance. But beginning in the 1960s, advances in technology made high-precision tests of the theory possible, and there followed a systematic effort to put the theory to the test (for a review of these tests see Will, 1993).

3.1 Tests of the Einstein Equivalence Principle

3.1.1 Tests of Special Relativity. The Einstein equivalence principle described in Sec. 1.2.2.2 demands that in any local, freely falling reference frame (in which gravity is absent locally), the nongravitational laws of physics (such as mechanics, electromagnetism, quantum mechanics) must be compatible with special relativity. As we have seen, a consequence of this principle is that gravity must be described by spacetime curvature.

Special relativity has become such a successful and integral part of such areas of modern physics as quantum field theory, nuclear physics, and particle physics that physicists often take its validity for granted. But in many of these subdisciplines of physics, the experiments are designed to test particular models of fields and interactions rather than the underlying special relativistic framework. The Michelson–Morley experiment and its modern-day descendants provide clean tests of SR in that they can constrain directly and quantitatively possible violations of SR (see RELATIVITY, SPECIAL).

3.1.1.1 Michelson–Morley Experiments. One way to understand the significance of these tests is to imagine an explicit violation of SR in electrodynamics, by permitting the limiting speed of material particles, c_m, to differ from the speed of electromagnetic waves, c_e. Since moving frames are no longer equivalent, this assumption establishes a preferred universal rest frame. From the field-theoretic point of view, such a violation is likely to be induced by some long-range field created by the matter in the universe; hence the preferred frame is likely to be that of the smoothed out distribution of matter, or equivalently of the cosmic microwave background. The resulting observable violations of SR in such a model depend on the fact that the Earth is moving through the universe at about 370 km/s, and are parametrized by $\delta=(c_m/c_e)^2-1$. By placing a limit on a difference in the speed of light in two perpendicular directions using an interferometer, the 1887 Michelson–Morley experiment set a limit $|\delta|<10^{-4}$. A 1979 laser-interferometric version of the experiment improved the limit to 10^{-9} (Brillet and Hall, 1979).

3.1.1.2 Hughes–Drever Experiment. But in 1960, a substantial improvement in the limit on δ resulted from experiments that tested the "isotropy of inertia", done independently by Hughes *et al.* (1960) and Drever (1961). If SR is violated in the way described above, then the energy levels of an atom or nucleus can depend both on the velocity of the nucleus through the preferred frame and on the orientation of the quantization axis relative to the direction of motion. Anomalous shifts of energy levels can then occur. The Hughes–Drever experiments used nuclear magnetic resonance techniques to look for such anomalous shifts as the Earth rotated in space. More recently, new techniques of atomic physics, such as laser-cooled ion and atom traps, have yielded improved limits on such variations with direction in the energy levels of beryllium ions, neon atoms, and mercury isotopes, ranging from 2×10^{-19} to 2×10^{-21} eV. These results give the bound $|\delta|<10^{-21}$, a truly high-precision confirmation of SR (Prestage *et al.*, 1985; Lamoreaux *et al.*, 1986; Chupp *et al.*, 1989).

3.1.2 The Eötvös Experiment, the Weak Equivalence Principle, and the Fifth Force

3.1.2.1 Eötvös and the Weak Equivalence Principle. Another experiment that helped lay the foundation for GR was the Eötvös experiment (1889,1908), which verified what we have called the weak equivalence principle (WEP), the equality of gravitational acceleration of objects of different composition. The precision achieved was a few parts in 10^9. Two new experiments, by Dicke at Princeton University in the early 1960s (Roll *et al.*, 1964), and by Braginski at Moscow State University in 1970 (Braginski and Panov, 1971), improved the accuracy by two orders of magnitude.

3.1.2.2 Fifth-Force Experiments. In 1986, there was renewed interest in the Eötvös experiment. As a result of a detailed reanalysis of Eötvös's original data, Fischbach *et al.* (1986) suggested the existence of a "fifth force" of nature, with a strength of about a percent of that of gravity, but with a range (as defined by the range λ of a Yukawa potential $e^{-r/\lambda}/r$ augmenting the usual Newtonian potential) of a few hundred meters. This proposal dovetailed with earlier hints of a deviation from the inverse-square law of Newtonian gravitation derived from measurements of the gravity profile down deep mines in Australia.

During the next four years, over a dozen new experiments looked for evidence of the fifth force by searching for composition-dependent differences in acceleration, with variants of the Eötvös experiment or with free-fall Galileo-type experiments. Many of the Eötvös-type torsion-balance experiments took advantage of new high-Q torsion suspension systems, advanced seismic isolation, and sophisticated arrangement of component masses to reduce gravity-gradient couplings.

One of the free-fall experiments used laser interferometry to measure the differential acceleration of two chambers, fitted with corner retroreflectors, containing different materials. Although two early experiments reported positive evidence, all others yielded null results. Over the range between 1 and 10^4 m, the null experiments produced upper limits on the strength of a postulated fifth force of between 10^{-3} and 10^{-6} the strength of gravity. Interpreted as tests of the WEP (corresponding to the limit of infinite-range forces), the results of a University of Washington experiment (dubbed Eöt-Wash) were comparable to that of Dicke (Adelberger *et al.*, 1991).

At the same time, tests of the inverse-square law of gravity were carried out by comparing variations in gravity measurements up tall towers or down mines or boreholes with gravity variations predicted using the inverse-square law together with Earth models and surface gravity data mathematically "continued" up the tower or down the hole. Despite early reports of anomalies, three independent tower measurements now show no evidence of a deviation (see, for example, Jekeli *et al.*, 1990). The consensus at present is that there is no credible experimental evidence for a fifth force of nature (for recent reviews see Adelberger *et al.*, 1991; Fischbach and Talmadge 1992; and Will 1990).

3.1.3 Gravitational Redshift. Although the gravitational redshift of light is a simple consequence of the EEP (Sec. 1.2.2.2), as Einstein found some eight years before he completed the full theory, it was not confirmed experimentally until the Pound–Rebka experiment of 1960, in which the frequency shifts of γ rays rising and falling in a tower were observed, making use of the Mössbauer effect to reduce recoil broadening of the emission and absorption γ-ray lines (Pound and Rebka, 1960).

Other tests included transporting atomic clocks on jet aircraft, and measuring the shift of solar spectral lines. The most accurate confirmation to date has been a 1976 rocket experiment, in which a hydrogen maser clock was launched on a Scout rocket to an altitude of 10 000 km, and its rate compared with an identical clock on the ground, resulting in a 0.02% test (Vessot *et al.*, 1980). Recently, a measurement of the shift of the rate of oscillator clocks on the Voyager spacecraft caused by Saturn's gravitational field yielded a 1% test (Krisher *et al.*, 1990).

The gravitational redshift now has practical consequences. In satellite-based navigation systems, such as the U.S. Air Force's Global Positioning System, the atomic clocks on the satellites tick faster, as a consequence of the gravitational frequency shift and the special relativistic time dilation, than do clocks on the ground by over 30 000 ns per day. Yet the navigational accuracy requirement of 10 m demands timekeeping accuracy to 30 ns at all times, and therefore general relativistic corrections *must* be taken into account in order for the system to function. Since GPS is now used extensively both for military and civilian navigation, this represents a new "applied" side of GR.

3.2 Solar-System Tests of General Relativity

3.2.1 The Deflection and Retardation of Light

3.2.1.1 Light Deflection. One of the first calculations that Einstein performed in November of 1915, when he had the final (vacuum) field equation of GR, was the deflection of light. Earlier, in 1911, he had determined the deflection in a preliminary theory based essentially purely on the EEP (Sec. 1.2.2.2) and got the answer, $2GM/Rc^2$, as in a Newtonian gravity theory in which light was treated as a corpuscle (see Sec. 1.2.2.5).

The result of Einstein's 1915 calculation was to double the prediction. For a light ray that grazes the Sun, for example, the deflection would be 1.75″ instead of 0.875″. The difference can be understood as follows: half the deflection indeed comes directly from the Einstein principle, or equivalently from a Newtonian ballistic calculation; the remaining part derives from the curvature of space near the Sun relative to space far away. The first contribution is the same in any theory of gravity that is a metric theory. The second, space-curvature contribution varies from one metric theory to another, and is conventionally represented by the PPN parameter γ, whose general-relativistic value is 1 (see Sec. 1.2.5). In this parametrized language, the deflection of a light ray by the Sun is given by

$$\Delta\theta=\tfrac{1}{2}(1+\gamma)1.75''/d\,, \tag{27}$$

where d is the distance of closest approach of the ray to the Sun, in units of a solar radius.

The measurement of this effect by British astronomers during a total solar eclipse in 1919 (Dyson *et al.*, 1920) catapulted Einstein and GR to worldwide fame. However, the accuracy was 20% at best. A few measurements during the next 45 years failed to yield substantial improvements. The development of radio interferometry during the 1960s coupled with the discovery of quasars led to dramatically better accuracy (see RADIO TELESCOPES). The technique involved monitoring the relative angle between a pair or group of quasars as they passed near the Sun as seen from Earth. During the passage, the light from a quasar whose image is closer to the Sun would be deflected more than that from a more distant one, leading to a displacement of one image relative to the other. Between 1969 and 1975 a dozen measurements of this sort were carried out systematically, culminating in confirmations of GR at about the 1.5% level (Fomalont and Sramek, 1976). After 1975, further direct measurements of the deflection of light to test relativity essentially ceased.

However, in the early 1980s systematic efforts were initiated by geophysicists to monitor the Earth's rotation state accurately by means of very-long-baseline interferometric (VBLI) measurements of the positions of radio galaxies and quasars. Using transcontinental and intercontinental baselines, and improved timing accuracy made possible by hydrogen maser and other atomic clocks, these measurements reached the several hundred *microarcsecond* level in accuracy, making it necessary to take the relativistic deflection of light into account over the entire celestial sphere, not just near the Sun. For a ray that approaches the Earth from a direction 90° away from the Sun, for example, the deflection is 4 milliarcseconds and is readily detectable. A by-product of this effort was a 0.1% confirmation of GR (Robertson *et al.*, 1991).

The European Space Agency (ESA) satellite HIPPARCOS, which is completing its mission as this is being written (1993), has similarly made precise determinations of the positions of some 10^5 stars, for which it has to take into account (and therefore measure) the light deflection over the whole celestial sphere. This will measure γ to similar accuracy.

3.2.1.2 *Shapiro Time Delay.* There is another important test of the propagation of light through curved spacetime, which was not around in 1915, yet which is closely related to the deflection of light. It was first predicted as a consequence of GR by Shapiro (1964), and is now commonly called the Shapiro time delay. It is an excess propagation delay of light passing through a region of curved space near a body compared to the analogous propagation time if the ray passes far from the body. A light ray that passes the Sun on a round-trip, say, from Earth to Mars at superior conjunction (when Mars is on the far side of the Sun) suffers a delay given by

$$\Delta t \approx \tfrac{1}{2}(1+\gamma)250(1-0.16\ \ln d)\ \mu\text{s}\ . \qquad (28)$$

The close relationship between this effect and the deflection of light is reflected in the factor $\frac{1}{2}(1+\gamma)$ and is to be expected, since any phenomenon that bends light (refraction, curved space) may be expected to alter its propagation time as well. Observations of the Shapiro time delay began in the mid-1960s with the use of radar echoes from Mercury and Venus. Later, use was made of interplanetary spacecraft equipped with radar transponders, such as Mariners 6, 7, and 9 and the Viking landers and orbiters. Data from Viking yielded a 0.1% test (Reasenberg *et al.*, 1979).

The result of these measurements on light is $\gamma=1.000\pm0.002$, in agreement with GR. This precise determination of the parameter γ is one of the crowning achievements of experimental gravitation.

3.2.2 Mercury's Perihelion Advance. The first effect that Einstein calculated in November 1915 using his new field equations was the advance of the perihelion of Mercury. The discrepancy between the observed advance and the amount that could be accounted for from the Newtonian gravitational perturbations of Mercury by the other planets was a problem that had bedeviled celestial mechanicians for the latter half of the nineteenth century. As mentioned in Sec. 2.4.4, GR predicted an amount that neatly accounted for the discrepancy. Einstein wrote later that he had palpitations of the heart upon finding this result. For another half a century, this stood as one of the triumphs of general relativity.

The predicted rate of advance of the perihelion of Mercury (excluding the part from planetary perturbations) can be written in the following form, in seconds of arc per century:

$$d\omega/dt = 42.98''\lambda_p, \tag{29}$$

$$\lambda_p := \tfrac{1}{3}(2+2\gamma-\beta) + 0.0003(J_2/10^{-7})\ . \tag{30}$$

The first term in the coefficient λ_p is the relativistic contribution to the advance, in a form that encompasses a wide class of alternative metric theories of gravity. The parameter γ is the same parameter that appeared in the deflection of light and the Shapiro time delay, while the parameter β is a rough measure of how "nonlinear" gravity is in a given theory. Both parameters are unity in GR. The second term comes from the Newtonian effect of a possible oblateness of the Sun, which will alter its external gravitational field from the pure inverse-square form of a spherical body. The oblateness is measured by the quantity $J_2 = (I_3 - I_1)/(M_\odot/R_\odot^2)$, where I_3 is the Sun's moment of inertia about its rotation axis and I_1 is the same about an equatorial axis; for a Sun that rotates uniformly with its observed surface angular velocity, so that the oblateness is caused by centrifugal flattening, J_2 is estimated to be of order 10^{-7}.

Now, the measured perihelion shift of Mercury is known very accurately because of the combination of two factors: improved radar ranging to Mercury since 1966, leading to a more accurate determination of its orbit, and improved data on the masses and orbits of the other planets from radar ranging and spacecraft encounters, leading to improved values for the planetary perturbations. After those perturbing effects have been accounted for, the excess shift is known to about 0.1%, with the result that $\lambda_p = 1.000 \pm 0.001$ (Shapiro, 1990).

If J_2 were indeed as small as 10^{-7} this would be in complete agreement with GR. However, in 1966, a value for J_2 of 2.5 ± 10^{-5} was inferred from visual solar-oblateness measurements made by Dicke and Goldenberg (1974), a result that, if confirmed, would have disagreed strongly with GR. Between 1966 and 1980 J_2 values ranging over two orders of magnitude were reported. Beginning around 1980, however, the observation and classification of modes of oscillation of the Sun ("helioseismology") made it possible to obtain information about its internal rotation rate, thereby constraining the possible centrifugal flattening; current results favor a value $J_2 \approx 1.7 \pm 10^{-7}$ (Brown *et al.*, 1989), making the perihelion shift of Mercury another success for general relativity.

3.2.3 Test of the Strong Equivalence Principle. Another important Solar-System experiment tests a generalization of the EEP, known as the strong equivalence principle (SEP). The strong equivalence principle states, for example, that *all* bodies should fall with the same acceleration in an external gravitational field, including bodies with significant internal gravitational binding energy, such as planets, stars, and so forth. In the WEP, one considers only laboratory-sized bodies, whose internal structures are dominated by nongravitational energies. Different theories of gravity can treat the effect of gravity on gravitational energy differently, and so could predict violations of the SEP by massive, self-gravitating bodies. General relativity is one of the few theories that actually obeys the SEP. The Brans–Dicke theory, for example, does not.

Since 1969, this principle has been tested using lunar laser ranging (LURE), in which laser pulses sent from Earth bounce off corner retroreflectors deposited on the Moon during U.S. and Soviet lunar landings. The goal is to look for the orbital effects of a possible difference in acceleration between the Earth and Moon toward the Sun (the *Nordtvedt effect*: Nordtvedt, 1968a,b). No orbital perturbation of this type has been found to date down to the 6-cm level, placing a limit of 7 parts in 10^{13} on a difference in acceleration between the two bodies (Dickey *et al.*, 1989; Shapiro, 1990; Müller *et al.*, 1991). The accuracy of lunar laser ranging is approaching the level of several millimeters, at which point the accuracy of this experiment as a test of the effect of gravity on gravitational energy (test of the SEP) will be limited by the accuracy of laboratory tests of the weak principle, because the composition of the Earth (iron rich) and Moon (iron poor) differ.

3.3 The Binary Pulsar: An Astronomical Relativity Laboratory

Until 1974, the solar system provided the principal testing ground for GR, because it is a "clean" system (few uncertain or messy physical processes to complicate the gravitational effects) and it is accessible to high-precision tools. However, the discovery of the binary pulsar PSR1913+16 in 1974 by astronomers at the Arecibo Radio Telescope in Puerto Rico (Hulse and Taylor, 1975) showed that certain kinds of distant astronomical systems may also provide precision laboratories for testing GR. The system consists of a 59-ms-period pulsar in an 8-h orbit with a companion that has not been seen directly, but that is generally believed to be another NS. The unexpected stability of the pulsar "clock" and the cleanliness of the orbit allowed radio astronomers to determine the orbital and other parameters of the system to extraordinary accuracy, by analyzing the variations in pulse arrival times caused by the orbital displacements (for the most recent data, see Taylor *et al.*, 1992).

Furthermore, the system is highly relativistic ($v_{\text{orbit}}/c \approx 10^{-3}$). Observation of the relativistic periastron advance (4.226 628 ±0.000 018 deg per year) and of the effects on pulse arrival times of the gravitational redshift caused by the companion's gravitational field and of the special relativistic time dilation caused by the pulsar's orbital motion (0.15% accuracy) have been used, assuming that GR is correct, to constrain the nature of the system. In GR, these two effects depend in a known way on measured orbital parameters and on the unknown masses m_p and m_c of the pulsar and companion (assuming that the companion is sufficiently compact that tidal and rotational distortion effects can be ignored), and consequently the two masses may be calculated with these two pieces of data, with the result $m_p = (1.4411 \pm 0.0007) M_\odot$ and $m_c = (1.3873 \pm 0.0007) M_\odot$. These are the most accurately known masses of any astronomical bodies outside the Solar System. It is interesting how GR plays a crucial role in this high-precision determination of astrophysical parameters.

This binary system should radiate GWs, which, unfortunately, are of too low a frequency to be detected directly by a ground-based GW detector. On the other hand, the loss of energy in GWs should cause the orbiting stars to spiral together, and the orbital period to *decrease*. The "quadrupole formula" (Sec. 2.3.3) determines the rate of loss of energy and the consequent orbital damping rate. It was first detected successfully in 1979, with about 10% precision. Using the measured orbital elements and the two masses, one can predict the rate of decrease of the period to be $dP/dt_{\text{predicted}} = -(2.402\,43 \pm 0.000\,05) \times 10^{-12}$ s/s. Because of the extraordinary stability of the pulsar, and with the added ability to transfer stable atomic time from the world's system of atomic clocks to the Arecibo Radio Telescope via the Global Positioning System, the observations of this period decrease have now reached 0.5% in accuracy, giving $dP/dt_{\text{observed}} = -(2.408 \pm 0.015) \times 10^{-12}$ (Taylor *et al.*, 1992). This agrees completely with the prediction and is very strong indirect evidence for the correctness of GR's predictions about GWs.

3.4 Future Work in Experimental Gravitation

3.4.1 Search for Gravitomagnetic Effects. According to GR, moving or rotating matter should produce a contribution to the gravitational field that is the analog of the magnetic field of a moving charge or a magnetic dipole (Sec. 2.1.1). Although gravitomagnetism plays a role in a variety of measured relativistic effects, it has not been seen to date, isolated from other post-Newtonian effects.

The Relativity Gyroscope Experiment at Stanford University (also known by the NASA terminology Gravity Probe B, or GP-B) is in the advanced stage of developing a space mission to detect this phenomenon directly (Everitt, 1988). A set of four superconducting-niobium-coated, spherical quartz gyroscopes will be flown in a low polar Earth orbit, and the precession of the gyroscopes relative to the distant stars will be measured. The predicted effect of gravitomagnetism is about 42 milliarcseconds per year, and the accuracy goal of the experiment is about 0.5 milliarcseconds per year.

To achieve this accuracy, which corresponds to a precession rate of 10^{-16} rad/s, numerous technical challenges have had to be met, including fabricating gyroscopes that are homogeneous and spherical to better than a part per million; developing and testing a "London moment" readout system that exploits the magnetic dipole moment developed by a spinning superconductor and uses SQUIDs to read out the varying currents in superconducting loops surrounding the gyroscope; and developing a magnetic shield of novel design to reduce the ambient magnetic field of the Earth below 10^{-7} G. Recently, a full-size flight prototype of the instrument package was tested as an integrated unit. Current plans call for a test of the final flight hardware on the Space Shuttle followed by a shuttle-launched science experiment.

Another proposal to look for an effect of gravitomagnetism is to measure the relative precession of the line of nodes of a pair of laser-ranged geodynamics satellites (LAGEOS), with supplementary inclination angles; the inclinations must be supplementary in order to cancel the dominant nodal precession caused by the Earth's Newtonian gravitational multipole moments (Ciufolini, 1989). A third proposal envisages orbiting an array of three mutually orthogonal, superconducting gravity gradiometers around the Earth to measure directly the contribution of the gravitomagnetic field to the tidal gravitational force (see, for example, Mashhoon *et al.*, 1989).

3.4.2 Tests of the Einstein Equivalence Principle. The concept of an Eötvös experiment in space has been developed as a possible joint NASA-ESA mission, with the potential to test the WEP to 10^{-17}, a millionfold improvement over current ground-based results (Worden, 1988). Such an experiment could also look for additional long-range, fifth-force-type interactions, with ranges in excess of about 40 km, and could lead to improvements in the value of the Newtonian constant G. It would also map the multipole moments of the Earth's Newtonian field to high accuracy.

The accuracy of measurements of gravitational redshift could be improved to the 10^{-9} level, and higher-order effects could be seen for the first time by placing a hydrogen maser clock on board Solar Probe, a proposed spacecraft that would travel to within four solar radii of the Sun (Vessot, 1989).

3.4.3 Further Fifth-Force Searches. Because they are relatively inexpensive and because they have the potential to constrain certain classes of particle-physics models, laboratory fifth-force experiments are likely to continue at some level for many years.

4. GRAVITATIONAL-WAVE DETECTION: A TECHNOLOGICAL FRONTIER

4.1 Likely Sources of Detectable Waves

In Sec. 2.3 we learned about GWs. Here we discuss their detection. Although detectors have been under development since Weber (1960) built the first one at the University of Maryland in the early 1960s, it is only recently that technology has permitted the design of a detector that meets theoretician's predictions about the likely strengths of expected waves. It is not unreasonable to expect that, by the year 2000, the first direct detection of a GW will have occurred.

Simple detection is not the main goal of the present detector development. GWs carry information about their sources that is obtainable in no other way. In order to extract maximum information from the waves, one needs to be able to infer their amplitude and direction. Since GW detectors are not directional, this can only be done with a worldwide network of detectors, which infer directions from the relative times of arrival of waves at different locations. Three detectors is the minimum for extraction of full information from the waves.

The first aspect of designing a detector is to estimate what amplitudes and frequencies one might expect from astronomical sources. These set targets for the experimental development.

We can dispose of one source right away: laboratory generators. It is not possible to build a laboratory generator of detectable GWs. One can make an apparatus that disturbs a detector by generating a time-depending gravitational field; this has been done in several laboratories (for example, Astone *et al.*, 1991a). But it is always the near-zone Newtonian field, and the disturbance would be the same in Newtonian gravity. To detect waves one must be at least one wavelength λ away from the source, and at this distance all reasonable sources are too weak.

We look therefore to relativistic sources in astronomy. All of these are powered by their internal gravitational forces, and so the key equation for estimating the strength of their emissions is Eq. (24), which sets a relatively simple upper bound on the amplitude we can expect. Here is a brief review of the principal candidates and the astronomical information we might expect to get from them. It is by no means an exhaustive list, but it represents the most conservative predictions that have been made about possible sources.

4.1.1 Supernovae. Supernovae are rare events, occurring once in perhaps 50 years in any galaxy. We would like to be able to detect them, therefore, in a volume of space containing perhaps 2000 galaxies, so that we have a reasonable chance of seeing one. This means that realistic detectors must reach as far as the Virgo cluster, a cluster of galaxies about 15 Mpc away, containing more than 1000 galaxies.

Supernovae will give off GWs if the collapse event is very nonspherical. The mass involved will be about a solar mass, and the size of the emission region is about 10 km. These give, from Eq. (24), an upper limit of a few times 10^{-21} for waves from the Virgo cluster. Given that this is an upper bound, the usual target is 10^{-21}. A detector with that sensitivity has some chance of seeing an occasional supernova explosion, unless they are all very symmetrical.

The expected frequency can be inferred from the collapse time scale. The whole collapse takes about 10 ms, but not much radiation comes off during most of this time. The bounce time scale is about 1 ms, and subsequent oscillations may also have this time scale. The relevant frequency is the inverse, about 1 kHz. One expects a broadband burst of this central frequency from a supernova explosion.

If such bursts are seen, they may allow us to identify the object (NS or BH) formed by the collapse; they may provide crucial information about high-density nuclear physics; they may allow rapid notification of other astronomers that a supernova has occurred at a particular position; and they will enlighten our understanding of the late stages of stellar evolution.

Bursts of radiation will also be emitted if galactic-size BHs are formed in a single event or if compact stars fall into such BHs. The frequency of the emitted radiation is inversely proportional to the mass of the hole, and for holes in the centers of galaxies this can be in the millihertz region or lower. Such events are not observable from the ground since ground vibrations and the near-zone Newtonian gravitational disturbances produced by atmospheric mass motions in this frequency range are too strong to screen out. But studies have shown (Bender *et al.*, 1989) that space-based detectors could be very sensitive at these frequencies. A suitable detector could see a giant BH formation event anywhere in the Universe; it could verify or rule out that mode of formation of galactic BHs.

4.1.2 Coalescing Binaries. The remarkable binary pulsar described in Sec. 3.3 will be even more remarkable in about 10^8 years, when the orbit has shrunk to the point that the two NSs in the system are orbiting 10 or more times per second. The system will be a strong source of GWs at relatively high frequency during the remaining few seconds before the stars coalesce. The frequency of the waves is twice the orbital frequency.

Such events are, of course, rare, occurring perhaps once every 10^5 to 10^7 years in any galaxy (Narayan *et al.*, 1991; Phinney, 1991; Tutukov and Yungelson, 1993). We therefore would need to be able to detect them in a volume of space up to 200 000 times larger than that for supernovae, or as much as 60 times further away. Fortunately, this is made possible by the regularity and predictability of the signal from such a system.

In contrast to supernovae, which are likely to be messy, binaries of compact objects emit a steady, almost monochromatic "chirp" signal, whose frequency increases with time in a predictable way. If one has N cycles of such a signal in one's data stream, then one can use pattern-matching techniques (such as matched filtering) to find it at an amplitude that is smaller by a factor of $\sqrt{N}$ than one could find if one had only a single cycle. Supernovae are basically single-cycle signals, so we can see binaries 60 times further away if we can detect 3600 cycles. The binary wave train will be followed by a much less predictable burst associated with the coalescence of the two stars.

The detector has to have a broad bandwidth at a few tens or hundreds of hertz to permit such filtering and following of the signal. In fact, the broadband interferometers described below are less noisy at the lower frequencies of binary signals, but on the other hand the intrinsic signal from a binary is a bit weaker than the strongest possible supernova (because ϕ_{internal} is smaller when the stars are still several radii apart), so that the calculation must be done carefully. The result is that detectors will have about 40 times the range for binaries as they have for moderate supernova explosions. Detectors that can just barely see supernovae in Virgo (such as the first-stage interferometers described below) are unlikely to see coalescing binaries. But the second stage of development, with 10 times better amplitude sensitivity, ought to see hundreds per year.

The binary signal contains much information, including the individual masses of the component stars. By observing them, we will directly identify NSs and BHs and get much better statistics on their mass distributions. Since they occur at great distances, these events sample the cosmological mass distribution on very large scales. The signal contains enough information to allow the absolute distance to the binary to be estimated: they are true standard candles. They therefore contain cosmological information, including the possibility of measuring the Hubble constant (Schutz, 1986). The actual coalescence of the two objects (NSs and/or BHs) may be associated with the mysterious γ-ray bursts seen by satellite detectors (Meegan *et al.*, 1992). If so, they will be easier to model and understand. Finally, the detection of the radiation from the merger of two BHs will provide a strong test of GR itself; modeling this event on computers is an area of current research.

4.1.3 Pulsars. A NS that is axially symmetric will not emit GWs when it rotates, but we know that pulsars are not symmetrical; they have off-axis magnetic fields. If they have other asymmetries, perhaps mass deformations that help to pin the magnetic poles in one place, then they may give off detectable radiation. There is an upper limit on the strength of this radiation for any pulsar whose spin-down rate has been measured: GWs probably do not carry away more than the rotational energy loss of the pulsar.

The signal will be steady, so one's ability to find it increases with the square root of time, for exactly the same reason as for coalescing binaries. Again, a broadband detector is desirable, so that signals of any frequency can be detected. Given a few months' observation, second-stage interferometers are expected to be able to beat the upper limits on several pulsars by factors of 10 or more.

It is possible to conduct searches for unknown pulsars through their GW emission. Here, as for known pulsars, one must remove the Doppler effects caused by the Earth's motion before the signal becomes periodic. This correction depends on the position of the pulsar. To perform a sensitive all-sky search is a demanding computational task, and the sensitivity will be limited by the capacity of available computers (Schutz, 1991).

4.1.4 Ordinary Binaries. Binary star systems emit radiation at twice the frequency of their orbit. For all except the coalescing binaries considered before, this is a very low frequency, not observable from the ground (Sec. 4.1.1). But space-based detectors would be able to reach as low as 10^{-4} Hz and could detect steady radiation from many binaries. One of the most prominent would be PSR1913 +16. Other compact binaries might be discovered this way.

Binaries containing white dwarfs are so numerous that at the lower frequencies they provide a chaotic background of waves that might mask other sources, at least if only one detector is used in space.

4.1.5 Cosmological Background. Just as there is a cosmological background of electromagnetic radiation, one expects a background of GWs. The thermal background may be too weak to see, but many other potential sources have been discussed. Most rely on aspects of high-energy physics that give rise to inflation or to topological anomalies.

Theories that produce cosmic strings have been studied extensively. Strings are massive linear regions of trapped field, which can seed galaxy formation and can emit GWs. If strings do provide the seeds for galaxy formation, then they ought to produce a background of GWs that is detectable by interferometers (Allen and Shellard, 1992).

Searches for such a background rely on cross-correlation of the output of two detectors, since in a single detector the background appears simply as another source of noise. If the detectors are close enough, the background will produce a correlated response in both that can stand out against other local sources of noise.

4.1.6 Unexpected Sources. One of the most exciting prospects is that there might be significant radiation from unpredicted sources. Whenever new windows on astronomy have been opened, such as radio or x-ray astronomy, completely unexpected objects have been found. This will undoubtedly be true for GWs as well, but of course one does not know what sensitivity will be required to reveal them. By going 10 times deeper in amplitude (100 times in energy) than the strongest predictions for bursts, second-stage interferometers must stand a good chance of making such discoveries.

4.2 Detectors

The reason for the current interest in GWs is technological: it now seems possible to build a detector that meets theoretical predictions about sources. But detectors of various types have been under development since 1960, and six or seven detectors around the world are able to make observations today. In this section we review the techniques, with an emphasis on the areas of technology that are relevant.

4.2.1 Bar-Type Detectors. Weber designed and built the first GW detector in the 1960s. It was based on a cylinder of aluminum, suspended and isolated from ground vibrations (Weber, 1967). When a GW hits the bar broadside, the tidal forces represented in Fig. 1 stretch it along its axis. By monitoring the excitation of the fundamental longitudinal mode of vibration of the bar, one can look for GWs.

Weber chose a bar that had a resonant frequency in the kilohertz region to look for supernova events. He showed that, even for a burst event, it is desirable for the vibrational mode to have a large Q; such a mode is weakly coupled to the thermal bath of the other modes, making it easier to recognize the rapid change in its excitation caused by a GW.

The dominant background is thermal vibrations. Subsequent generations of bars have been cooled to 4.2 K to reduce this background. A further generation of bars now under construction (1993) may reach 10 mK (for example, see Astone *et al.*, 1991b). Weighing several tons, such bars will be the coldest large objects the Universe has ever seen! Where Weber's original bar had a sensitivity limit of $h \sim 10^{-16}$ for a broadband burst around 1 kHz, cryogenic bars today reach about 10^{-18}, and the new bars may go to 10^{-20}.

For bars to go below this to the interesting level of 10^{-21} means defeating the *quantum limit*. This refers to the fact that, in classical terms, the excitation energy deposited in a bar by a GW of that amplitude will be smaller than the energy of one phonon of longitudinal excitation. Although one might think this would be an absolute limit, GW theorists showed that one can use the uncertainty principle to "squeeze" the uncertainty in a particular observable (not the phonon number) and deduce from that the amplitude of the incident GW (Caves *et al.*, 1980).

Incidentally, when the same theory was applied to the laser interferometric detectors, the notion of "squeezed light" was developed (Caves, 1981). Although it had been discussed before, GW detectors stimulated the current strong interest in the quantum optics and communications community in this technique for reducing photon noise.

Despite more than a decade of research into squeezing for bars, no practical scheme has emerged, and one cannot be hopeful that bars will break the 10^{-21} barrier. For the next five years, while interferometers are being constructed, the ultracryogenic bars will have the best chance of detecting GWs. If a supernova occurs in our Galaxy or in a nearby one, they may well succeed.

4.2.2 Laser-Interferometric Detectors. An interferometer is designed to measure the relative difference in two optical paths. If one places an interferometer in the center of the ring in Fig. 1, with the ends of two perpendicular arms on the ring, then the subsequent motion of the ends relative to the center can be detected by interferometry.

The relative motions are small, however, and so the technical challenge is large. For the end mirrors to respond as free masses, they must be hung from supports and isolated from ground vibration. Because the tidal gravitational forces scale with distance, it is desirable to make the arms as long as possible. Present prototypes (see the various articles in Blair, 1991) are in the 10–40-m range, but detectors now under construction (see below) will be as long as 4 km. It is this ability to take advantage of the tidal scaling that gives interferometers the edge over bars.

Even over 4 km, a disturbance of 10^{-21} translates into a mirror motion of 4×10^{-16} cm, less than 0.01 fm. To sense average displacements of the surface of a macroscopic object to this accuracy with optical or infrared photons whose wavelengths are 12 orders of magnitude larger requires many photons. In turn, this requires excellent mirrors and high-power continuous-wave lasers. The physical principles of such detectors have been reviewed by Giazotto (1989).

The detectors will be built in stages. The first stage will aim at a sensitivity to broadband bursts of 10^{-21}. The second stage will aim for 10^{-22}. Current designs anticipate mirrors with losses well below 10^{-4} per reflection, and neodymium-doped yttrium aluminum garnet (YAG) lasers as sources for the second stage. (See LASERS, SOLID STATE.) For the second stage, light will have to be conserved, and designers like R. W. P. Drever and the late B. J. Meers have devised clever methods of optimizing the use of light. It is now clear that, although the detectors are intrinsically broadband, they can also be tuned to narrow frequency ranges if desired.

The light must travel along the arms in a good vacuum, better than 10^{-8} Torr for the second stage. This is not hard in principle, but the volume to be evacuated is large: tubes 4 km in length and 1.5 m in diameter. The cost of this vacuum system is the dominant cost of the detectors.

Data will flow from these detectors at an enormous rate. Recent short observing runs using prototypes at Glasgow University and the Max Planck Institute for Quantum Optics in Garching, Germany, produced data rates approaching 1 Gbyte/h. The storage and analysis of data from a network of interferometers are formidable problems.

There are proposals for a number of interferometric detectors around the world. A project called LIGO to build two in the United States (Abramovici *et al.*, 1992) is funded, and should begin construction in 1993. If present (1993) schedules hold, LIGO could begin observing by 1998. A French–Italian detector called VIRGO, to be built near Pisa, Italy, is likely to get final approval in 1993 (Bradaschia *et al.*, 1990). There are further proposals for a British–German detector near Hannover, Germany, and an Australian detector near Perth that are awaiting funding. Japan is planning an intermediate-sized detector (300 m). Because detectors have broad quadrupolar antenna patterns, an accurate direction to a detected source can only be obtained by triangulation, using the time delays among various detectors. For this reason, a minimum of three detectors worldwide is required for the extraction of full information from detected signals.

4.2.3 Space-Based Detectors. As we noted in Sec. 4.1.1, the interesting frequency range around 1 mHz is only accessible to space-based detectors. Space-based searches for GWs have already been made using transponding data from interplanetary space probes. A passing GW would affect the time delays of round-trip signaling to the probes; the signature of this effect is unique, and sensitive searches can be made at very low frequencies. No positive detections have been reported (e.g., Armstrong *et al.*, 1987). As the principal limit on sensitivity is propagation disturbances of the radio communication signals, sensitivity can be increased by using multiple-frequency communication, by going to higher frequencies, and by using multiple spacecraft. A three-spacecraft experiment was performed in 1993 using NASA's Mars Observer and Galileo probes and the European Space Administration's Ulysses solar observatory. When the data are analyzed, they may improve previous limits by a factor of 10 or more.

Proposals have been made for purpose-built interferometers in space. Laser interferometers using the Earth–Moon or even the Sun–Earth Lagrangian points to stabilize their orbits look very promising.

GLOSSARY

Cosmology: The study of the large-scale properties of the observable universe as a whole.

Geodesics: Locally straight lines of a curved space or spacetime. They are also paths of extremal proper distance.

Homogeneous Cosmology: A cosmology in which there is a choice of time (a reference frame) such that all its properties are the same everywhere at a constant time.

Isotropic Cosmology: A cosmology that is homogeneous and in which, in addition, there is no preferred direction in space, such as a systematic velocity of the matter.

Reference Frame: A coordinate system for space and time that includes facilities for measuring and recording the exact location and time of any event. In special relativity, a *Lorentz* reference frame has rigid spacings between spatial coordinate locations, and its clocks are synchronized everywhere. In gen eral relativity, a *local* reference frame makes measurements only in a restricted region.

List of Works Cited

Abramovici, A., Althouse, W. E., Drever, R. W. P., Gürsel, Y., Kawamura, S., Raab, F. J., Shoemaker, D., Sievers, L., Spero, R. E., Thorne, K. S., Vogt, R. E., Weiss, R., Whitcomb, S. E., Zucker, M. E. (1992), *Science* **256**, 325–333.

Adelberger, E. G., Heckel, B. R., Stubbs, C. W., Rogers, W. F. (1991), *Annu. Rev. Nucl. Particle Sci.* **41**, 269–320.

Allen, B., Shellard, E. P. S. (1992), *Phys. Rev. D* **45**, 1898–1912.

Armstrong, J. W., Estabrook, F. B., Wahlquist, H. D. (1987), *Astrophys. J.* **318**, 536–541.

Astone, P., Bassan, M., Bates, S., Bizzarri, R., Bonifazi, F., Cardarelli, R., Cavallari, G., Coccia, E., Degasperis, A., Depedis, D., Frasca, S., Majorana, E., Merucci, L., Modena, I., Muratori, G., Pallottino, G. V., Patrignani, C., Pizzella, G., Price, M., Rapagnani, P., Ricci, F., Visco, M. (1991a), *Z. Phys.* C. **50**, 21–29.

Astone, P., Bassan, M., Bonifazi, F., Bronzini, F., Castellano, M. G., Coccia, E., Cosmelli, C., Fafone, V., Frasca, S., Majorana, E., Modena, I., Pallottino, G. V., Pizzella, G., Rapagnani, P., Ricci, F., Visco, M. (1991b), *Europhys. Lett.* **16**, 231–235.

Bender, P., Faller, J. E., Hall, J. L., Hils, D., Stebbins, R. T., Vincent, M. A. (1989), in: R. W. Hellings (Ed.), *Relativistic Gravitational Experiments in Space*, NASA Conference Publication 3046, Washington, D.C.: NASA, pp. 80–88.

Birrell, N. D., Davies, P. C. W. (1982), *Quantum Fields in Curved Space*, New York: Cambridge University Press.

Blair, D. G. (Ed.) (1991), *The Detection of Gravitational Waves*, Cambridge, U.K.: Cambridge University Press.

Blandford, R. D., Znajek, R. L. (1977), *Mon. Not. R. Astron. Soc.* **179**, 433–456.

Blandford, R. D., Begelman, M. C., Rees, M. J. (1982), *Sci. Am.*, May, 84–94.

Blau, S. K., Guth, A. H. (1987), in: S. W. Hawking, W. Israel (Eds.), *300 Years of Gravitation*, New York: Cambridge University Press, pp. 524–603.

Bradaschia, C., DelFabbro, R., DiVirgilio, A., Giazotto, A., Kautzky, H., Montelatici, V., Passuello, D., Brillet, A., Cregut, O., Hello, P., Man, C. N., Manh, P. T., Marraud, A., Shoemaker, D., Vinet, J. Y., Barone, F., Difiore, L., Milano, L., Russo, G., Aguirregabiria, J. M., Bel, H., Duruisseau, J. P., Ledenmat, G., Tourrenc, P., Capozzi, M., Longo, M., Lops, M., Pinto, I., Rotoli, G., Damour, T., Bonazzola, S., Marck, J. A., Gourghoulon, Y., Holloway, L. E., Fuligni, F., Iafolla, V., Natale, G. (1990), *Nucl. Instrum. Methods Phys. A* **289**, 518–525.

Braginski, V. B., Panov, V. I. (1971), *Zh. Eksp. Teor. Fiz.* **61**, 873–879 [*Sov. Phys. JETP* **34**, 463–466 (1972)].

Brillet, A., Hall, J. L. (1979), *Phys. Rev. Lett.* **42**, 549–552.

Brown, T. M., Christensen-Dalsgaard, J., Dziembowski, W. A., Goode, P., Gough, D. O., Morrow, C. A. (1989), *Astrophys. J.*, **343**, 526–546.

Casares, J., Charles, P. A., Naylor, T. (1992), *Nature* **355**, 614–617.

Caves, C. M. (1981), *Phys. Rev. D* **23**, 1693–1708.

Caves, C. M., Thorne, K. S., Drever, R. W. P., Sandberg, V. D., Zimmerman, M. (1980), *Rev. Mod. Phys.* **52**, 341–392.

Chandrasekhar, S. (1939), *An Introduction to the Study of Stellar Structure*, Chicago: University of Chicago Press. Reprint (1958), New York: Dover.

Chupp, T. E., Hoare, R. J., Loveman, R. A., Oteiza, E. R., Richardson, J. M., Wagshul, M. E., Thompson, A. K. (1989), *Phys. Rev. Lett.* **63**, 1541–1545.

Ciufolini, I. (1989), *Int. J. Mod. Phys. A* **4**, 3083–3145.

Dicke, R. H., Goldenberg, H. M. (1974), *Astrophys. J. Suppl.* **27**, 131–182.

Dickey, J. O., Newhall, X. X., Williams, J. G. (1989), *Adv. Space Res.* **9**, 75–78.

Drever, R. W. P. (1961), *Philos Mag.* **6**, 683–687.

Dyson, F. W., Eddington, A. S., Davidson, C. (1920), *Philos. Trans. R. Soc. London* **220A**, 291–333.

Einstein, A. (1915), *Preuss. Akad. Wiss. Berlin, Sitzber.* 844–847.

Everitt, C. W. F. (1988), in: J. D. Fairbank, B. S. Deaver, C. W. F. Everitt, P. F. Michelson (Eds.), *Near Zero: New Frontiers of Physics*, New York: Freeman, pp. 587–639.

Fischbach, E., Talmadge, C. (1992), *Nature* **356**, 207–215.

Fischbach, E., Sudarsky, D., Szafer, A., Talmadge, C., Aronson, S. H. (1986), *Phys. Rev. Lett.* **56**, 3–6.

Fomalont, E. B., Sramek, R. A. (1976), *Phys. Rev. Lett.* **36**, 1475–1478.

Giazotto, A. (1989). Phys. Rep. **182**, 365–425.

Guth, A. H., Steinhardt, P. J. (1984), *Sci. Am.*, May, 90–102.

Hartle, J. B. (1978), *Phys. Rep.* **46**, 201.

Hawking, S. W. (1974), *Nature* **248**, 30–31.

Hawking, S. W., Ellis, G. F. R. (1973), *The Large-Scale Structure of Space-Time*, New York: Cambridge University Press.

Hilbert, D. (1915), *Goett. Nachr.* 395–407.

Hughes, V. W., Robinson, H. G., Beltrán-Lopez, V. (1960), *Phys. Rev. Lett.* **4**, 342–344.

Hulse, R. A., Taylor, J. H. (1975), *Astrophys. J. Lett.* **195**, L51–L53.

Isaacson, R. (1968), *Phys. Rev.* **166**, 1272–1280.

Jekeli, C., Eckhardt, D. H., Romaides, A. J. (1990), *Phys. Rev. Lett.* **64**, 1204–1206.

Knapp, J., Kormendy, J. (1986), *Dark Matter in the Universe: I. A. U. Symposium Proceedings No. 117*, Dordrecht: Reidel.

Krisher, T. P., Anderson, J. D., Campbell, J. K. (1990), *Phys. Rev. Lett.* **64**, 1322–1325.

Lamoreaux, S. K., Jacobs, J. P., Heckel, B. R., Raab, F. J., Fortsen, E. N. (1986), *Phys. Rev. Lett.* **57**, 3125–3128.

Landau, L. D., Lifshitz, E. M. (1962), *The Classical Theory of Fields*, Oxford: Pergamon.

Mashhoon, B., Paik, H. J., Will, C. M. (1989), *Phys. Rev. D* **39**, 2825–2838.

Meegan, C. A., Fishman, G. J., Wilson, R. B., Paciesas, W. S. (1992), *Nature* **355**, 143–145.

Misner, C. W., Thorne, K. S., Wheeler, J. A. (1973), *Gravitation*, San Francisco: Freeman.

Müller, J., Schneider, M., Soffel, M., Ruder, H. (1991), *Astrophys. J. Lett.* **382**, L101–L103.

Narayan, R., Piran, T., Shemi, A. (1991), *Astrophys. J. Lett.* **379**, L17–L20.

Nordtvedt, K., Jr. (1968a), *Phys. Rev.* **169**, 1014–1016.

Nordtvedt, K., Jr. (1968b), *Phys. Rev.* **169**, 1017–1025.

Penzias, A. A., Wilson, R. W. (1965), *Astrophys. J.* **142**, 419–421.

Phinney, E. S. (1991), *Astrophys. J. Lett.* **380**, L17–L19.

Pound, R. V., Rebka, G. A., Jr. (1960), *Phys. Rev. Lett.* **4**, 337.

Prestage, J. D., Bollinger, J. J., Itano, W. M., Wineland, D. J. (1985), *Phys. Rev. Lett.* **54**, 2387–2390.

Reasenberg, R. D., Shapiro, I. I., MacNeil, P. E., Goldstein, R. B., Breidenthal, J. C., Brenkle, J. P., Cain, D. L., Kaufman, T. M., Komarek, T. A., Zygielbaum, A. I. (1979), *Astrophys. J. Lett.* **234**, L219–L221.

Rees, M. J. (1990), *Sci. Am.*, **26** (5), 56–66.

Robertson, D. S., Carter, W. E., Dillinger, W. H. (1991), *Nature* **349**, 768–770.

Roll, P. G., Krotkov, R. V., Dicke, R. H. (1964), *Ann. Phys. (NY)* **26**, 442–517.

Schutz, B. F. (1984), *Am. J. Phys.* **52**, 412–419.

Schutz, B. F. (1985), *A First Course in General Relativity*, New York: Cambridge University Press.

Schutz, B. F. (1986), *Nature* **323**, 310–311.

Schutz, B. F. (1991), in: D. G. Blair (Ed.), *The Detection of Gravitational Waves*, Cambridge: Cambridge University Press, pp. 406–452.

Shapiro, I. I. (1964), *Phys. Rev. Lett.* **13**, 789–791.

Shapiro, I. I. (1990), in: N. Ashby, D. F. Bartlett, W. Wyss (Ed.), *General Relativity and Gravitation*, Cambridge, U.K.: Cambridge University Press, pp. 313–330.

Smoot, G. F., Bennett, C. L., Kogut, A., Wright, E. L., Aymon, J., Boggess, N. W., Cheng, E. S., Deamici, G., Gulkis, S., Hauser, M. G., Hinshaw, G., Jackson, P. D., Janssen, M., Kaita, E., Kelsall, T., Keegstra, P., Lineweaver, C., Loewenstein, K., Lubin, P., Mather, J., Meyer, S. S., Moseley, S. H., Murdock, T., Rokke, L., Silverberg, R. F., Tenorio, L., Weiss, R., Wilkinson, D. T. (1992), *Astrophys. J Lett.* **396**, L1–L5.

Taylor, J. H., Wolszczan, A., Damour, T., Weisberg, J. M. (1992), *Nature* **355**, 132–136.

Thorne, K. S. (1991), *Ann. N. Y. Acad. Sci.* **631**, 182–193.

Thorne, K. S., Price, R. H., Macdonald, D. A. (1986), *Black Holes: the Membrane Paradigm*, New Haven: Yale University Press.

Tutukov, A. V., Yungelson, L. R. (1993), *Mon. Not. R. Astron. Soc.* **260**, 675–678.

Vessot, R. F. C. (1989), *Adv. Space. Res.* **9**, 21–28.

Vessot, R. F. C., Levine, M. W., Mattison, E. M., Blomberg, E. L., Hoffman, T. E., Nystrom, J. W., Farrel, B. F., Decher, R., Eby, P. B., Baugher, C. R., Watts, J. W., Teuber, D. L., Wills, F. O. (1980), *Phys. Rev. Lett.* **45**, 2081–2084.

Wald, R. M. (1984), *General Relativity*, Chicago: Univ. of Chicago Press.

Walsh, D., Carswell, R. E., Weymann, R. J. (1979), *Nature* **279**, 381.

Weber, J. (1960), *Phys. Rev.* **117**, 306–313.

Weber, J. (1967), *Phys. Rev. Lett.* **18**, 498–501.

Will, C. M. (1990), *Sky Telescope* **80**, 472–479.

Will, C. M. (1993), *Theory and Experiment in Gravitational Physics*, 2nd ed., Cambridge, U.K.: Cambridge University Press.

Worden, P. W. (1988) in: J. D. Fairbank, B. S. Deaver, C. W. F. Everitt, P. F. Michelson (Eds.), *Near Zero: New Frontiers of Physics*, New York: Freeman, pp. 766–783.

Further Reading

Blair, D. G. (Ed.) (1991), *The Detection of Gravitational Waves*, Cambridge, U.K.: Cambridge Univ. Press.

Davies, P. C. W. (1980), *The Search for Gravity Waves*, Cambridge, U.K.: Cambridge Univ. Press.

Hawking, S. W., Israel, W. (Eds.) (1987), *300 Years of Gravitation*, Cambridge, England: Cambridge Univ. Press.

Hoffmann, B. (1983), *Relativity and its Roots*, New York: Freeman.

Misner, C. W., Thorne, K. S., Wheeler, J. A. (1973), *Gravitation*, San Francisco: Freeman.

Pais, A. (1982), *Subtle is the Lord: the Science and Life of Albert Einstein*, Oxford: Oxford University Press.

Price, R. H. (1982), *Am. J. Phys.* **50**, 300–329.

Schutz, B. F. (1980), *Geometrical Methods of Mathematical Physics*, Cambridge, U.K.: Cambridge Univ. Press.

Schutz, B. F. (1985), *A First Course in General Relativity*, Cambridge, U.K.: Cambridge University Press.

Shapiro, S. L., Teukolsky, S. A. (1983), *Black Holes, White Dwarfs, and Neutron Stars*, New York: Wiley.

Taylor, E. F., Wheeler, J. A. (1992), *Spacetime Physics*, 2nd ed., New York: Freeman.

Wheeler, J. A. (1990), *A Journey into Gravity and Spacetime*, New York: Freeman.

Will, C. M. (1986), *Was Einstein Right?*, New York: Basic.

ERRATUM

GRAVITATION AND GENERAL RELATIVITY (Vol. 7, pp. 303–340). BERNARD F. SCHUTZ and CLIFFORD M. WILL.

Page 306. Eq. (2) and (3) should read

$$(hG/c^3)^{1/2} = 4.1 \times 10^{-35}\ \mathrm{m} \quad \text{(Planck length);} \tag{2}$$

$$(hG/c^5)^{1/2} = 1.4 \times 10^{-43}\ \mathrm{s} \quad \text{(Planck time);} \tag{3}$$

GREEN'S FUNCTIONS

KAZUO OHTAKA, *Laboratory of Applied Physics, Faculty of Engineering, Chiba University, Chiba-shi, Japan*

Introduction 342

1. **History of Green's Functions** 343

2. **Construction of Green's Functions** 345

2.1 One-Dimensional Equation of Sturm-Liouville Type with Dirichlet-Type Boundary Conditions 345

2.2 Retarded, Advanced, and Causal Green's Functions of the Helmholtz Equation 345

2.3 Green's Functions Obtained by Fourier Transform 346

2.3.1 Heat Equation 347

2.3.2 Time-Dependent Schrödinger Equation 347

2.3.3 Klein-Gordon Equation 347

2.4 Green's Functions Matching Homogeneous Boundary Conditions at the Boundary of a Finite Region

2.5 Spectral Representation of Green's Functions 348

3. **Green's Functions Used in Solving Initial- and Boundary-Value Problems** 349

3.1 Dirichlet and Neumann Problems of Poisson's Equation 349

3.2 Initial- and Boundary-Value Problem for the Heat Equation 349

4. **Boundary-Element Method** 350

4.1 Practical Boundary-Value Problems 350

4.2 Poisson's Equation as Treated by the Boundary-Element Method 350

4.3 Applications of the Boundary-Element Method 351

4.3.1 Fluid Mechanics 351

4.3.2 Sound and Electromagnetic Waves 351

4.3.3 Elasticity 351

5. **Green's Functions Having a Direct Relevance to Physical Reality** 352

5.1 Wave Front of Radiation Emitted from a Point Source and Huygens's Principle 352

5.2 Retarded Green's Function of Schrödinger's Equation 353

5.3 Dislocations 353

5.4 Magnetic Field around a Vortex Line in a Type-II Superconductor 353

6. **Perturbational Treatment to Obtain Green's Functions** 354

6.1 Slater-Koster Model 354

6.2 Scattering Cross Section of a Plane Wave from a Scatterer with Spherical Symmetry 355

6.3 Band Structure of Electrons and Photons and Diffraction Experiments 355

7. **Green's Functions in Many-Body Theories** 356

7.1 Single- and Two-Particle Green's Functions 357

7.2 Wick's Theorem and Feynman Diagrams 358

7.3 Temperature Green's Functions 359

7.4 Linear Response Theory of Kubo and Temperature Green's Functions 359

7.5 Green's Functions in Quantum Field Theories 360

Glossary 361

Works Cited 362

Further Reading 363

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

In mathematics the term Green's function is usually given to a solution of an initial- or boundary-value problem of a differential equation with a δ-function inhomogeneous term. Let us be more specific. Consider an ordinary or partial differential equation

$$L_x G(x,x') = \delta(x-x'), \tag{1}$$

with L_x a linear differential operator with respect to the variable x. Here, x may stand for either the position $\mathbf{r}$, the time t, or the pair $(\mathbf{r},t)$. Then the solution $G(x,x')$ is called the Green's function if it satisfies a given homogeneous boundary condition—a condition relating the value of G to its derivative G_x on the boundary of the domain, a simple example of which is $G(x,x')=0$ or $G_x(x,x')=0$. In physics and applied physics, however, the term Green's function is often used without explicitly referring to the boundary condition. For example, a fundamental solution in mathematics is often called simply a Green's function. Thus it is desired to seek an alternative and more generalized definition that enables us to deal with a wider class of functions encountered in various areas of physical science. This can be achieved by formulating the concept of Green's functions in terms of response theory: The Green's function is then a response function that connects the output signal $O(x)$ of the system with the input signal $I(x)$ in the form of a linear integral transform:

$$O(x) = \int G(x,x')I(x')dx', \tag{2}$$

the integral range over x' depending upon the problem under consideration. In the present article the term Green's function is employed in this generalized sense.

When the linear response of a system is described by a linear operator L_x, which may be differential, integral, integro-differential, or of any other kind, the two signals $O(x)$ and $I(x)$ are related through

$$L_x O(x) = I(x). \tag{3}$$

Comparing this equation with Eq. (2), we see that the Green's function is formally defined by L_x^{-1}. When, in particular, L_x is a differential operator in Eq. (3) and a homogeneous boundary condition, $u(x)=0$, for example, is imposed on the boundary Γ, the definition used in mathematics is recovered. This is because the superposition over x' in Eq. (2) solves the problem when G satisfies Eq. (1) with the boundary condition $G(x,x')=0$ for x on Γ. There are, however, many cases where it is difficult to specify the operator L_x for describing the response. The relationship between the responses of a black box to a δ-function type and distributed input signals is shown in Fig. 1.

Although the principle of superposition and hence the validity of the form given by Eq. (2) hold only when the solution satisfies a homogenous boundary condition, Green's functions are also central when one tries to construct a solution of a boundary-value problem with an inhomogeneous boundary condition—for example, a solution having a prescribed nonzero value on the boundary. This is one reason why Green's functions are so widely used.

The quantities called "resolvent," "resolvent kernel," "signal function," "point response function," or "transfer function," encountered in various fields of mathematics, physics, applied physics, and engineering, are nothing but the Green's functions in the generalized definition. We note that in Eq. (1) the Green's function $G(x,x')$ describes the response to a "point" input source and in Eq. (2) it "transfers" the input signal into the output response of the system in question. When one recalls that many problems in physics and applied physics ultimately reduce to finding the output $O(x)$ for a given input $I(x)$, one can

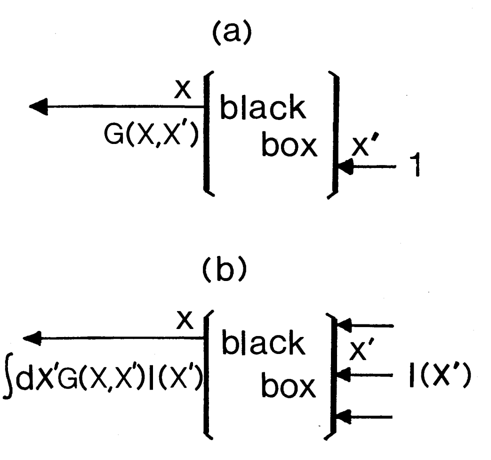

FIG. 1. Response at x of a black box to an input signal: (a) an input signal localized at x' with unit strength, (b) an input signal distributed with magnitude $I(x')$.

understand why Green's functions are very popular today in many fields—hydrodynamics, electrodynamics, acoustics, elasticity, quantum mechanics, solid-state physics, elementary-particle physics, and so on. To imagine how widely they are used, it is enough to remember the diverse names given to them, listed above. Their usefulness is still growing progressively today, as various numerical techniques continue to develop for calculations involving Green's functions.

The subjects of the present article are the definitions, significances, constructions, utilizations, and usefulness of Green's functions. We try to make the description as illustrative as possible. The Green's functions we deal with in this article range from those treated in mathematical textbooks to the ones used in many fields of pure and applied physics.

Although, unless stated otherwise, the concrete forms of the Green's functions will be given for the case of three-dimensional space, the reader should keep in mind that they depend intrinsically on the dimensionality of the space considered. The reader can consult the monographs quoted in the text for the case of other dimensionalities. Also, it should be noted that in defining a Green's function in this article the coefficient of a δ function of the point input signal is not always unity, reflecting some arbitrariness existing in the literature.

The present article is organized as follows. We sketch in Sec. 1 a brief history of Green's functions including their modern development. In Sec. 2 some typical methods of constructing Green's functions are explained for several differential equations. The usefulness of Green's functions in initial- and boundary-value problems is demonstrated in Sec. 3. The boundary-element method, devised to handle boundary-value problems for a nontrivial geometry, is explained in Sec. 4, together with the presentation of some of its applications. Up to this point, the description is given for the role of Green's functions as a convenient tool for solving a mathematical or physical problem. In Sec. 5, a number of Green's functions are given which have a direct relevance with a physical reality. The treatments as combined with perturbation method are described in Sec. 6. The Green's functions popular in many-body problems are described in Sec. 7, where, among other things, we review their extremely important application in linear response theory of condensed-matter physics. A brief sketch of their use in quantum field theory is also given.

1. HISTORY OF GREEN'S FUNCTIONS

In the history of Green's functions, it will be appropriate to go back to 1828, when N. Green put forth Green's formula (Kellog, 1939)

$$\int_{\Omega}[u(\mathbf{r})\Delta v(\mathbf{r})-v(\mathbf{r})\Delta u(\mathbf{r})]d^3\mathbf{r} = \int_{\Gamma}\left(u(\mathbf{r})\frac{\partial}{\partial n}v(\mathbf{r})-v(\mathbf{r})\frac{\partial}{\partial n}u(\mathbf{r})\right)ds. \tag{4}$$

It converts the volume integral within a region Ω of the left-hand side into the surface integral on its boundary Γ, with Δ the Laplacian and $(\partial/\partial n)v(\mathbf{r})=\hat{\mathbf{n}}\cdot\nabla v(\mathbf{r})$, $\hat{\mathbf{n}}$ being the outward normal with unit length to the boundary Γ. This formula holds for arbitrary $u(\mathbf{r})$ and $v(\mathbf{r})$. When, in particular, $u(\mathbf{r})$ is a harmonic function, $\Delta u(\mathbf{r})=0$, and $v(\mathbf{r})$ is the Green's function of the Laplace equation

$$\Delta v(\mathbf{r})=-\delta(\mathbf{r}-\mathbf{r}') \tag{5}$$

or

$$v(\mathbf{r})=1/(4\pi|\mathbf{r}-\mathbf{r}'|), \tag{6}$$

Green's formula yields for $\mathbf{r}$ within Ω

$$u(\mathbf{r})=\frac{1}{4\pi}\int_{\Gamma}\left(\frac{1}{|\mathbf{r}-\mathbf{r}'|}\frac{\partial}{\partial n'}u(\mathbf{r}') - u(\mathbf{r}')\frac{\partial}{\partial n'}\frac{1}{|\mathbf{r}-\mathbf{r}'|}\right)ds'. \tag{7}$$

It shows that, as in Cauchy's formula for regular functions of complex argument $z=x+iy$, we can express a harmonic function inside a region as an integral over its boundary: we may evaluate the value of $u(\mathbf{r})$ inside Ω only if we know values of both $u(\mathbf{r})$ and $\partial u(\mathbf{r})/\partial n$ on Γ. If we let the argument $\mathbf{r}$ approach the boundary Γ, Eq. (7) becomes a Fredholm integral equation, which allows us to express $u(\mathbf{r})$ in terms of $\partial u(\mathbf{r})/\partial n$. Likewise, differentiating Eq. (7) over $\mathbf{r}$ and letting $\mathbf{r}$ tend to Γ lead to the equation for $\partial u(\mathbf{r})/\partial n$. Thus $u(\mathbf{r})$ and $\partial u(\mathbf{r})/\partial n$ cannot both be specified freely on Γ. From this fact stem the

Dirichlet and Neumann problems of the Laplace equation, i.e., the problems of finding a harmonic function $u(\mathbf{r})$ that has a prescribed boundary value $u(\mathbf{r})$ and outward derivative $\partial u(\mathbf{r})/\partial n$, respectively, on the boundary Γ.

In the latter half of the 19th century, Green's functions played a fundamental role in the discussion of the existence and uniqueness of the solution of internal or external boundary-value problems (Courant and Hilbert, 1937). In addition to their significance in these basic problems, they were also the key quantities in the practical side of constructing the solutions of boundary-value problems, as analyzed fully by Lyapunov for the potential problem (Smirnov, 1965).

Towards the end of the 19th century and at the beginning of the 20th century, Green's functions were used in examining the completeness property (closure property) of the set of eigenfunctions of self-adjoint operators and in proving the expansion theorem for an arbitrary function in terms of the complete set obtained from, say, a Sturm-Liouville operator (Courant and Hilbert, 1937). In mathematics, these concepts opened the way to functional analysis, which has since refined and generalized greatly the theory of partial differential equations (Yosida, 1965). In the fields of pure and applied physics, not only were they adopted in solving various practical problems, but they were also used in clarifying many fundamental concepts underlying quantum mechanics, which was founded in 1925 and has been developing ever since. Indeed, one of the easiest ways to recognize a marked peculiarity of quantum mechanics contrasting with classical mechanics is to change the action integral for a classical motion into the form involving the Green's function of the Schrödinger equation (Sakurai, 1985). Also, in many quantum-mechanical applications, the Green's functions enable us to take into account a perturbation series to an infinite order to give a deep insight not attainable by a finite-order treatment.

As elementary-particle physics and solid-state physics began to develop rapidly after World War II, the extended applications of Green's functions were actively pursued. One example is seen in many-particle physics in which the Green's functions are defined in terms of field operators and used in conjunction with the graphical representation of many-body processes (see, e.g., Feynman, 1972). In these graphs, which have came to be called Feynman diagrams, each line standing for a Green's function describes temporal and spatial evolution of an elementary particle or excitation. The crossing or branching of the lines represents the interaction among particles, implying that the Green's function carries all the information on the "personal history" of an electron, proton, photon, etc. Because of this characteristic, these Green's functions are more often called "propagators." Their contributions in the development of quantum electrodynamics (Bogoliubov and Shirkov, 1959) and solid-state physics (Abrikosov *et al.*, 1963) have been quite remarkable. Despite the apparent difference in definition, Green's functions defined in terms of Feynman's path integral in quantum field theory constitute the second example belonging to this category (Feynman and Hibbs, 1965; Itzykson and Zuber, 1980). Since the functional integral seems to be the most powerful tool to date to quantize nonlinear Lagrangians, the Green's functions will continue to be a useful tool in the future development of this field. As a last example, we refer the reader to linear response theory applied widely in condensed-matter physics. In this example. too, Green's functions have been very useful in that the theory of Kubo is most powerful when it is applied in conjunction with the temperature Green's functions introduced by Matsubara (Kubo *et al.*, 1991).

Parallel to such generalizations in pure physics, the Green's functions of traditional usage have been refined in various ways, yielding many important concepts. Especially, many practical problems related to the Laplace, heat, wave, or Schrödinger equation, previously left untouched simply because the boundaries of the domains in question were too complex for analytical treatment, have come to be solved with the help of numerical techniques. The boundary-element method, one of the methods designed for just such problems, takes full advantage of the Green's-function approach (Brebbia, 1978; Brebbia and Walker, 1980). Green's functions are now so widely used everywhere that familiarity with them is becoming more and more important.

2. CONSTRUCTION OF GREEN'S FUNCTIONS

A number of typical methods of constructing Green's functions are illustrated.

2.1 One-Dimensional Equation of Sturm-Liouville Type with Dirichlet-Type Boundary Conditions

The Green's function for the Sturm-Liouville operator satisfies $[p(x)>0]$

$$L[G(x,x')]\equiv\frac{d}{dx}\left(p(x)\frac{d}{dx}G(x,x')\right)-q(x)G(x,x')=-\delta(x-x'),\tag{8}$$

where the one-dimensional region $0<x<1$ is assumed. Suppose that a Dirichlet-type boundary condition is imposed on G:

$$G(x,x')=0 \quad \text{at } x=0 \text{ and } 1.\tag{9}$$

The solution of this problem is constructed as follows:

$$G(x,x')=\begin{cases}\gamma u_<(x)u_>(x'), & 0\leqslant x\leqslant x'\leqslant 1,\\ \gamma u_>(x)u_<(x'), & 0\leqslant x'\leqslant x\leqslant 1.\end{cases}\tag{10}$$

Here $u_<(x)$ is a solution of $L[u_<(x)]=0$ with the boundary value $u_<(0)=0$ at the left boundary, while $u_>(x)$ satisfies $L[u_>(x)]=0$ with $u_>(1)=0$ at the right boundary. The constant γ in Eq. (10), independent of x and x', is given by

$$\gamma=\frac{1/p(x')}{u'_<(x')u_>(x')-u'_>(x')u_<(x')},\tag{11}$$

with $u'_<(x')=(d/dx')u_<(x')$. It is determined such that

$$p(x')[G'(x'^+,x')-G'(x'^-,x')]=-1,\tag{12}$$

with $G'(x'^\pm,x')=(d/dx)G(x,x')|_{x=x'^\pm}$, which is the condition obtained by integrating both sides of Eq. (8) in the infinitesimal interval $x'^-<x<x'^+$, where $x'^\pm=x'\pm\epsilon$ $(\epsilon\to 0+)$. When the left-hand solution $u_<(x)$ happens to satisfy simultaneously the condition $u_<(1)=0$, i.e., when $u_<(x)$ happens to be the true eigenfunction of the operator L with zero eigenvalue, the constant γ diverges, meaning that the system resonates with the point external force expressed by the δ function in Eq. (8). Still, in this case, one can redefine a generalized Green's function so that Eq. (2) remains valid in taking into account the inhomogeneous term (Smirnov, 1965). In the case of a Neumann- or mixed-type homogeneous boundary condition in place of Eq. (9), Eq. (10) still provides us with the Green's function if $u_<(x)$ and $u_>(x)$ therein satisfy the given boundary conditions. For various types of Sturm-Liouville operators and their generalizations, Green's functions are tabulated in many books. See, e.g., Butkovskiy (1982).

2.2 Retarded, Advanced, and Causal Green's Functions of the Helmholtz Equation

The Green's function of the Helmholtz equation is defined by

$$(\Delta+\kappa^2)G(\mathbf{r},\mathbf{r}';\kappa)=-\delta(\mathbf{r}-\mathbf{r}').\tag{13}$$

By Fourier transform we find

$$G(\mathbf{r},\mathbf{r}';\kappa)=\int\frac{d^3k}{(2\pi)^3}\frac{\exp[i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')]}{k^2-\kappa^2}.\tag{14}$$

As such the Green's function is undetermined because the integral is divergent. If we add a small imaginary part $i\epsilon$ to κ $(\epsilon\to 0+)$, one may construct the following three types of Green's functions:

1. By putting κ equal to $\kappa+i\epsilon$, one obtains

   $$G_R(\mathbf{r},\mathbf{r}';\kappa)=\exp(i\kappa|\mathbf{r}-\mathbf{r}'|)/4\pi|\mathbf{r}-\mathbf{r}'|,\tag{15}$$

   which is called the retarded Green's function. It is regular in the upper half of the complex κ plane.
2. By putting κ equal to $\kappa-i\epsilon$, one obtains a complex conjugate of the retarded function,

   $$G_A(\mathbf{r},\mathbf{r}';\kappa)=\exp(-i\kappa|\mathbf{r}-\mathbf{r}'|)/4\pi|\mathbf{r}-\mathbf{r}'|.\tag{16}$$

   This Green's function is called the advanced Green's function and is regular in the lower half of the complex κ plane.
3. By putting κ equal to $\kappa+i\epsilon\,\mathrm{sgn}\,\kappa$, $\mathrm{sgn}\,\kappa$ being $\kappa/|\kappa|$, (i.e., κ^2 to $\kappa^2+i\epsilon$), one obtains the causal Green's function

   $$G_C(\mathbf{r},\mathbf{r}';\kappa)=G_R(\mathbf{r},\mathbf{r}';\kappa)\theta(\kappa)+G_A(\mathbf{r},\mathbf{r}';\kappa)\times\theta(-\kappa),\tag{17}$$

the Heaviside step function $\theta(\kappa)$ being defined by

$$\theta(\kappa)=\begin{cases}1, & \kappa>0,\\ 0, & \kappa<0.\end{cases} \tag{18}$$

We see that G_R (G_A) is obtained by analytically continuing G_C of the range $\kappa>0$ ($\kappa<0$) to the upper (lower) half of the complex κ plane. Except in many-body theories, G_C is seldom used [see Eq. (88)]. The names "retarded" and "advanced" come from the time dependence of the Green's functions of the (time-dependent) wave equation

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2}-\Delta\right)G(\mathbf{r},t;\mathbf{r}',t')=\delta(\mathbf{r}-\mathbf{r}')\delta(t-t'), \tag{19}$$

c being a positive constant. Upon converting from t to the Fourier space ω, we obtain the Helmholtz equation given by Eq. (13) with $\kappa^2=\omega^2/c^2$. The inverse Fourier transform back to the variables t and t' of Eq. (14) [or Eq. (15) or (16)] shows that $G_R(\mathbf{r},t;\mathbf{r}',t')$ has nonzero values only in the case $t>t'$, while G_A is finite in the opposite case, $t<t'$. Namely, when G_R or G_A is substituted for G in the input-output relation (2), it turns out that

$$O(\mathbf{r},t)=\int_{-\infty}^{t} dt' \times\int d^3\mathbf{r}' G_R(\mathbf{r},t;\mathbf{r}',t')I(\mathbf{r}',t'),$$

$$O(\mathbf{r},t)=\int_{t}^{\infty} dt' \times\int d^3r' G_A(\mathbf{r},t;\mathbf{r}',t')I(\mathbf{r}',t'). \tag{20}$$

That is, the retarded Green's function duly satisfies the causality condition in the sense that a perturbation due to $I(\mathbf{r}',t')$ precedes its consequence observed at time t. The advanced Green's function describes the time-reversed process of the physical one. These features are also obvious in the time-independent version $G(\mathbf{r},\mathbf{r}';\kappa)$, which describes the scattering of, e.g., a monochromatic sound wave with a fixed frequency ω ($\kappa=\omega/c$, c being the velocity of sound). Here G_R given by Eq. (15) leads properly to the outgoing scattered wave, while G_A gives rise to an incoming scattered wave (Sommerfeld, 1949). To summarize, the three Green's functions of wave equation are defined by

$$\begin{aligned}
G_R(\mathbf{r},t;\mathbf{r}',t')&=\lim_{\epsilon\to 0}\int \frac{d^3k\,d\omega}{(2\pi)^4}\,\frac{\exp[i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')-i\omega(t-t')]}{k^2-(\omega+i\epsilon)^2/c^2},\\
G_A(\mathbf{r},t;\mathbf{r}',t')&=\lim_{\epsilon\to 0}\int \frac{d^3k\,d\omega}{(2\pi)^4}\,\frac{\exp[i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')-i\omega(t-t')]}{k^2-(\omega-i\epsilon)^2/c^2},\\
G_C(\mathbf{r},t;\mathbf{r}',t')&=\lim_{\epsilon\to 0}\int \frac{d^3k\,d\omega}{(2\pi)^4}\,\frac{\exp[i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')-i\omega(t-t')]}{k^2-\omega^2/c^2-i\epsilon}.
\end{aligned} \tag{21}$$

After evaluating the integrals, we find

$$\begin{aligned}
G_R(\mathbf{r},t;\mathbf{r}',t')&=(c/2\pi)\theta(t-t')\\
&\quad\times\delta(c^2(t-t')^2-|\mathbf{r}-\mathbf{r}'|^2),\\
G_A(\mathbf{r},t;\mathbf{r}',t')&=(c/2\pi)\theta(t'-t)\\
&\quad\times\delta(c^2(t-t')^2-|\mathbf{r}-\mathbf{r}'|^2),\\
G_C(\mathbf{r},t;\mathbf{r}',t')&=[ic/(2\pi)^2]/[|\mathbf{r}-\mathbf{r}|^2\\
&\quad-c^2(t-t')^2+i\epsilon].
\end{aligned} \tag{22}$$

The first of the three leads to Huygens's principle (See Sec. 5.1).

2.3 Green's Functions Obtained by Fourier Transform

As shown in Sec. 2.2, the Fourier transform is a convenient way to obtain Green's functions. It is powerful only for obtaining the Green's function for an infinite domain, however, i.e., a fundamental solution. Nevertheless, it should be noted that such a Green's function enables us to derive a Green's function subject to a homogeneous boundary condition on the boundary of a finite domain (see Sec. 2.4). Note also that the infinite-domain Green's functions are used very often to solve the problems for a finite domain [the boundary-element method is one of the examples (see

Secs. 4.2 and 4.3)]. With this remark in mind, we will in this section give some Green's functions obtained by Fourier transform.

2.3.1 Heat Equation. The Green's function of the heat equation is defined by

$$\left(\frac{\partial}{\partial t}-\sigma^2\Delta\right) G(\mathbf{r},t;\mathbf{r}',t')=\delta(\mathbf{r}-\mathbf{r}')\delta(t-t'). \tag{23}$$

In Fourier space, it holds that

$$\begin{aligned} G(\mathbf{r},t;\mathbf{r}',t') &= \int \frac{d^3k\, d\omega}{(2\pi)^4} \frac{\exp[i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')-i\omega(t-t')]}{\sigma^2k^2-i\omega} \\ &= \begin{cases} [4\pi\sigma^2(t-t')]^{-3/2}\exp\{-|\mathbf{r}-\mathbf{r}'|^2/[4\sigma^2(t-t')]\}, & t>t', \\ 0, \quad t<t', \end{cases} \end{aligned} \tag{24}$$

the second relation being obtained by calculating the residue in the ω integral. We should note that the integral over ω is well defined in contrast to the wave equation treated in Sec. 2.2. The finiteness of the Green's function only in the case $t>t'$ is in accord with the law of increase of entropy or the second law of thermodynamics.

2.3.2 Time-Dependent Schrödinger Equation. The Green's function for a free particle with mass m obeying Schrödinger's equation is defined by

$$\left(i\hbar\frac{\partial}{\partial t}+\frac{\hbar^2}{2m}\Delta\right) G(\mathbf{r},t;\mathbf{r}',t')=i\hbar\delta(\mathbf{r}-\mathbf{r}')\delta(t-t'). \tag{25}$$

Fourier transform then yields

$$\begin{aligned} G_R(\mathbf{r},t;\mathbf{r}',t') &= i\int \frac{d^3k\, d\omega}{(2\pi)^4} \frac{\exp[i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')-i\omega(t-t')]}{\omega-\hbar k^2/2m+i\epsilon} \\ &= \begin{cases} [m/2\pi i\hbar(t-t')]^{3/2}\exp[im|\mathbf{r}-\mathbf{r}'|^2/2\hbar(t-t')], & t>t', \\ 0, \quad t<t'. \end{cases} \end{aligned} \tag{26}$$

In the ω integral, we have replaced ω by $\omega+i\epsilon$ to obtain the retarded Green's function. The advanced Green's function, finite for $t<t'$, is obtained by putting ω equal to $\omega-i\epsilon$. Thus the Schrödinger equation allows a time-reversed solution, like what we have seen for the wave equation [Eq. (19)].

2.3.3 Klein-Gordon Equation. The Green's functions of the Klein-Gordon equation are defined by

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2}-\Delta+\mu^2\right) G(\mathbf{r},t;\mathbf{r}',t')=\delta(\mathbf{r}-\mathbf{r}')\delta(t-t'), \tag{27}$$

c and μ being two positive constants. Replacing μ^2 by $-\mu^2$ defines the Green's function of the telegraphic equation. By Fourier transform we find

$$G(\mathbf{r},t;\mathbf{r}',t')=\int \frac{d^3k\, d\omega}{(2\pi)^4} \frac{\exp[i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')-i\omega(t-t')]}{k^2+\mu^2-(\omega/c)^2}. \tag{28}$$

Since the integral is not well defined, we can construct three Green's functions, G_R, G_A, and G_C, by replacing ω as in Eq. (21). They are obtained, respectively, as (Bogoliubov and Shirkov, 1959)

$$\begin{aligned} G_R(x) &= -\theta(t-t')\Delta(x;\mu^2), \\ G_A(x) &= \theta(t'-t)\Delta(x;\mu^2), \\ G_C(x) &= \bar{\Delta}(x;\mu^2) + (i/2)\Delta^{(1)}(x;\mu^2). \end{aligned} \tag{29}$$

Here $G(\mathbf{r},t;\mathbf{r}',t')$ is expressed simply as $G(x)$, x standing for $(c(t-t'),\mathbf{r}-\mathbf{r}')$, and $\theta(t-t')$ is the step function defined by Eq. (18). The other quantities are defined by

$$\begin{aligned} \Delta(x;\mu^2) &= (-c/2\pi)\operatorname{sgn}(t-t')[\delta(x^2) \\ &\quad -\theta(x^2)(\mu^2/2)J_1(\mu x)/(\mu x)], \\ \bar{\Delta}(x;\mu^2) &= -\tfrac{1}{2}\operatorname{sgn}(t-t')\Delta(x;\mu^2), \\ \Delta^{(1)}(x;\mu^2) &= (c/4\pi|\mathbf{r}-\mathbf{r}'|)(\partial/\partial r) \\ &\quad \times[\theta(x^2)N_0(\mu x) - i\theta(-x^2) \\ &\quad \times H_0(i\mu(-x^2)^{1/2})], \end{aligned} \tag{30}$$

where $\operatorname{sgn}(t)=t/|t|$, $x^2=c^2(t-t')^2-|\mathbf{r}-\mathbf{r}'|^2$, $x=(x^2)^{1/2}$, J_1 is the first-order Bessel function, and N_0 and H_0 are the Neumann and first-kind Hankel functions of the zeroth order, respectively (Abramowitz and Stegun, 1965). In quantum field theory, the Green's functions in Fourier space [$(\mathbf{k},\omega)$ representation] are more often used than the $(\mathbf{r},t)$ representation given above. The Green's functions treated in Sec. 2.2 are reproduced by taking the limit $\mu \to 0$ in the above.

2.4 Green's Functions Matching Homogeneous Boundary Conditions at the Boundary of a Finite Region

As an illustration, let us consider the Green's function of the Laplace equation for a region Ω:

$$\Delta G(\mathbf{r},\mathbf{r}') = -\delta(\mathbf{r}-\mathbf{r}'), \tag{31}$$

with a homogeneous Dirichlet condition imposed on the boundary Γ:

$$G(\mathbf{r},\mathbf{r}')|_{\mathbf{r}\text{ on }\Gamma}=0. \tag{32}$$

Were it not for the restriction (32), the Green's function would be nothing more than the fundamental solution of the Laplace equation, the Coulomb potential given by Eq. (6). It satisfies the boundary condition $G(\mathbf{r},\mathbf{r}') \to 0$, as $r \to \infty$. To match the boundary condition (32), the Green's function must have the form

$$G(\mathbf{r},\mathbf{r}') = g(\mathbf{r},\mathbf{r}') + 1/4\pi|\mathbf{r}-\mathbf{r}'|, \tag{33}$$

where $g(\mathbf{r},\mathbf{r}')$ is the solution of the homogeneous equation

$$\Delta g(\mathbf{r},\mathbf{r}') = 0 \tag{34}$$

subject to the boundary condition

$$g(\mathbf{r},\mathbf{r}')|_{\mathbf{r}\text{ on }\Gamma} = -1/(4\pi|\mathbf{r}-\mathbf{r}'|)|_{\mathbf{r}\text{ on }\Gamma}. \tag{35}$$

The second term of Eq. (33) takes account of the δ function of the Poisson equation, while the first term $g(\mathbf{r},\mathbf{r}')$ incorporates the boundary condition. The problem of finding the Green's function thus reduces to an orthodox Dirichlet problem of Eqs. (34) and (35) of finding a harmonic function $g(\mathbf{r},\mathbf{r}')$ satisfying the inhomogeneous boundary condition. Although the existence and uniqueness of the solution is well established, a concrete expression for $g(\mathbf{r},\mathbf{r}')$ is hard to obtain analytically, unless the boundary Γ has a good symmetry (Kellog, 1939; Smirnov, 1965). Nevertheless, this method of obtaining the Green's functions satisfying the boundary condition has a wide applicability in many differential equations, and is not restricted to the Laplace equation treated here.

2.5 Spectral Representation of Green's Functions

As an example of Green's functions of three-dimensional self-adjoint operators, we treat here that of the Schrödinger equation defined by

$$[E-H(\mathbf{r})]G(\mathbf{r},\mathbf{r}';E) = i\hbar\delta(\mathbf{r}-\mathbf{r}'). \tag{36}$$

For an electron in a hydrogen atom, for example, the Hamiltonian H is given by

$$H(\mathbf{r}) = -\frac{\hbar^2}{2m}\Delta - \frac{e^2}{4\pi\epsilon_0 r}, \tag{37}$$

the proton being taken as the origin of coordinates. From Eq. (36) one obtains

$$G(\mathbf{r},\mathbf{r}';E) = \sum_n i\hbar\psi_n(\mathbf{r})\psi_n(\mathbf{r}')^*/(E-E_n), \tag{38}$$

with the eigenfunction ψ_n satisfying

$$H(\mathbf{r})\psi_n(\mathbf{r}) = E_n\psi_n(\mathbf{r}). \tag{39}$$

Equation (38) can easily be verified by applying $E-H(\mathbf{r})$ to both sides and using the completeness of $\{\psi_n\}$ for the self-adjoint operator $H(\mathbf{r})$. The set of states n includes not only the states with discrete energy eigenvalues but also the states within continuous spectra, if there are any, as in the case of a hydrogen atom. To describe a physical process occurring in a hydrogen atom, we must use the retarded version of the Green's function obtained by changing E to $E+i\epsilon$ in Eq. (38), in accordance with the remark made in 2.2.

Even in a finite-domain problem subject to a homogeneous restriction imposed on the boundary, the expression (38) remains valid, with the understanding that the ψ_n's are now the solutions for the eigenvalue problem with that boundary condition.

3. GREEN'S FUNCTIONS USED IN SOLVING INITIAL- AND BOUNDARY-VALUE PROBLEMS

In solving an initial- or boundary-value problem, the Green's function is useful in taking account of not only an inhomogeneous term but also an inhomogeneous initial or boundary condition. Some examples are given in Kellog (1939) and Morse and Feshbach (1953).

3.1 Dirichlet and Neumann Problems of Poisson's Equation

The internal Dirichlet problem of the Poisson equation is defined by

$$\Delta u(\mathbf{r})=-f(\mathbf{r}), \tag{40}$$

with the inhomogeneous boundary condition imposed on Γ,

$$u(\mathbf{r})|_\Gamma=g(\mathbf{r}). \tag{41}$$

The solution $u(\mathbf{r})$ of this problem may be written down if we know the Green's function $G_1(\mathbf{r},\mathbf{r}')$ of the Laplace equation satisfying the boundary condition $G_1(\mathbf{r},\mathbf{r}')|_\Gamma=0$ [i.e., the Green's function given by Eq. (33)]. It reads

$$u(\mathbf{r})=\int_\Omega G_1(\mathbf{r},\mathbf{r}')f(\mathbf{r}')d^3r' - \int_\Gamma \left(\frac{\partial}{\partial n'} G_1(\mathbf{r},\mathbf{r}')\right)g(\mathbf{r}')ds', \tag{42}$$

where n defines the outward normal to Γ as in Eq. (4). The right-hand side is being written solely in terms of the given boundary value $g(\mathbf{r})$. The reader can easily convince himself that this formula is correct by noting that the first term satisfies the Poisson equation with the boundary value $u(\mathbf{r})|_\Gamma=0$, while the second is the solution of the Laplace equation with $u(\mathbf{r})|_\Gamma=g(\mathbf{r})$, as can be checked by the use of the Green's formula (4) with $u=u(\mathbf{r})$ and $v=G_1(\mathbf{r},\mathbf{r}')$.

Let $G_2(\mathbf{r},\mathbf{r}')$ be the Green's function satisfying the homogeneous Neumann-type condition, $(\partial/\partial n)G_2(\mathbf{r},\mathbf{r}')=0$ for $\mathbf{r}$ on Γ. If we employ G_2 in place of G_1 and replace $-\partial G_1/\partial n'$ by G_2, the expression given by Eq. (42) gives the solution of the Poisson equation subject to the inhomogeneous Neumann-type boundary condition, $(\partial/\partial n)u(\mathbf{r})|_\Gamma=g(\mathbf{r})$, in place of Eq. (41). The external problems are treated analogously.

3.2 Initial- and Boundary-Value Problem for the Heat Equation

The Green's formula (4) is generalized to an arbitrary second-order differential operator L—to that of the heat equation, $L=\partial/\partial t-\sigma^2\Delta$, for example. By using it, we can express the solution for, say, the following problem of the heat equation for $\mathbf{r}\in\Omega$ and $t>0$:

$$\left(\frac{\partial}{\partial t}-\sigma^2\Delta\right) u(\mathbf{r},t)=f(\mathbf{r},t), \tag{43}$$

with the initial temperature distribution specified by

$$u(\mathbf{r},0)=g(\mathbf{r}) \tag{44}$$

and the boundary condition of Dirichlet type given by

$$u(\mathbf{r},t)|_\Gamma=h(\mathbf{r},t). \tag{45}$$

Suppose we already happen to know the Green's function $H_1(\mathbf{r},t;\mathbf{r}',t')$ for the operator M $(=-\partial/\partial t-\sigma^2\Delta)$, the adjoint of the operator L:

$$\left(-\frac{\partial}{\partial t}-\sigma^2\Delta\right)H_1(\mathbf{r},t;\mathbf{r}',t')=\delta(\mathbf{r}-\mathbf{r}')\delta(t-t'), \tag{46}$$

which satisfies the homogeneous boundary condition

$$H_1(\mathbf{r},t;\mathbf{r}',t')|_{\mathbf{r}\text{ on }\Gamma}=0. \tag{47}$$

Then it is shown that the solution of the problem (43)–(45) is given by

$$u(\mathbf{r},t)=\int_0^t dt'\int_\Omega d^3r' H_1(\mathbf{r}',t';\mathbf{r},t)f(\mathbf{r}',t')$$
$$+\int_\Omega d^3r' H_1(\mathbf{r}',0;\mathbf{r},t)g(\mathbf{r}')$$
$$-\sigma^2\int_0^t dt'\int_\Gamma ds'\frac{\partial H_1(\mathbf{r}',t',\mathbf{r},t)}{\partial n'}$$
$$\times h(\mathbf{r}',t'). \qquad (48)$$

If we employ, in place of H_1, another Green's function H_2 satisfying the boundary condition $[\partial H_2(\mathbf{r},\mathbf{r}')/\partial n]|_\Gamma=0$ instead of Eq. (47), we obtain the solution of Eqs. (43) and (44) with, in place of Eq. (45), the Neumann-type inhomogeneous boundary condition, $(\partial/\partial n)\times u(\mathbf{r})|_\Gamma=h(\mathbf{r},t)$. We should note that, as in Sec. 3.1, $-\partial H_1/\partial n'$ in Eq. (48) must be replaced by H_2.

The present examples given for operators that are not self-adjoint will suffice to illustrate the practical value of Green's functions in a rather wide class of boundary-value problems. An important point is that the Green's function used in the input-output relation (2) is not defined by a δ-function inhomogeneous term for the operator L but by the one for its adjoint operator M. Also, note that the reciprocity relation of Green's functions is in general established between the Green's functions G and H for the operator L and its adjoint M. Namely, for the Green's function G and H for the operators L and M, respectively, it holds that

$$G_i(\mathbf{r},t;\mathbf{r}',t')=H_i(\mathbf{r}',t';\mathbf{r},t), \qquad (49)$$

$i=1$ and 2 corresponding to the Dirichlet and Neumann boundary conditions, respectively. If the operator L is self-adjoint, with the relation $L=M$, the Green's function H is automatically identical to G, leading to the well-known reciprocity relation:

$$G_i(\mathbf{r},t;\mathbf{r}',t')=G_i(\mathbf{r}',t';\mathbf{r},t). \qquad (50)$$

The operator L for the Laplace, wave, Klein-Gordon, or Schrödinger equation is self-adjoint but that for the heat equation is not (Courant and Hilbert, 1937).

4. BOUNDARY-ELEMENT METHOD

4.1 Practical Boundary-Value Problems

In actual situations, we often encounter a complex boundary Γ. If we insist on applying the formulas given in Sec. 3, we will be forced to solve additional boundary-value problems in order to find Green's functions G_1, H_1, etc., as the example in Sec. 2.4 shows. Therefore these formulas are not very helpful in such problems, and more direct methods, taking full advantage of numerical techniques, are more often employed in practice. The difference method and finite-element method are two such popular examples. The boundary-element method, developed and applied widely in recent years, also belongs to this class. In contrast to the former two, which have nothing to do with Green's functions, this method is related deeply to Green's formula and hence Green's functions. Conceptually, it is a revival of the old method of expressing the solution of the Laplace equation in the form of the potential caused by a monopole or dipole layer on the boundary Γ, the unknown density of which is determined by solving the Fredholm integral equation (Courant and Hilbert, 1937). The Green's functions involved are the fundamental solution (6) (in the case of the Laplace equation), instead of the complicated Green's functions G_1, etc. Let us briefly see the characteristic points of the boundary-element method (BEM) through the following examples (Brebbia and Walker, 1980).

4.2 Poisson's Equation as Treated by the Boundary-Element Method

We return to the Poisson equation treated in Sec. 3.1. If the fundamental solution of the Laplace equation [Eq. (6)] is substituted for $v(\mathbf{r})$ in Green's formula (4), we can express the solution for the Poisson equation by using the boundary values of both $u(\mathbf{r})$ and $\partial u(\mathbf{r})/\partial n$

on Γ. The result is the extension of Eq. (7) to the Poisson equation. For $\mathbf{r}\in\Omega$, we find

$$u(\mathbf{r}) = \int_\Omega G(\mathbf{r},\mathbf{r}')f(\mathbf{r}')d^3\mathbf{r}' + \int_\Gamma \left(G(\mathbf{r},\mathbf{r}')\frac{\partial u(\mathbf{r}')}{\partial n'} - \frac{\partial G(\mathbf{r},\mathbf{r}')}{\partial n'}u(\mathbf{r}') \right) ds' \qquad (51)$$

with $G(\mathbf{r},\mathbf{r}') = 1/4\pi|\mathbf{r}-\mathbf{r}'|$. Since we know the value of $u(\mathbf{r})$ on Γ through the Dirichlet condition (41), $u(\mathbf{r})=g(\mathbf{r})$, this formula provides us with the solution of the original problem, if we somehow find the value of $\partial u(\mathbf{r})/\partial n$ on Γ on the right-hand side. The procedure characterizing the BEM is that the unknown quantity $\partial u/\partial n$ is determined from Eq. (51) by letting **r** tend to a point on Γ and setting $u(\mathbf{r})=g(\mathbf{r})$. The result is a Fredholm integral equation of the first kind for the unknown function $\partial u(\mathbf{r})/\partial n$ on Γ:

$$\tfrac{1}{2}g(\mathbf{r}) - \int_\Omega G(\mathbf{r},\mathbf{r}')f(\mathbf{r}')d^3\mathbf{r}' + \int_\Gamma \frac{\partial G(\mathbf{r},\mathbf{r}')}{\partial n'}g(\mathbf{r}')ds' = \int_\Gamma G(\mathbf{r},\mathbf{r}')\frac{\partial u(\mathbf{r}')}{\partial n'}ds'. \qquad (52)$$

Here the factor $\frac{1}{2}$ takes account of the discontinuous nature of $\partial G/\partial n$ in letting **r** tend to the boundary. What Eq. (52) shows is that we need to make a numerical calculation of the unknown quantity $[\partial u(\mathbf{r})/\partial n]|_\Gamma$. This can be carried out readily by discretizing the integral on Γ using the well-established algorithm of the finite-element method. Following similar steps, the internal Neumann problem and the external problems are eventually reduced to an integral equation on the boundary Γ as in this example.

4.3 Applications of the Boundary-Element Method

4.3.1 Fluid Mechanics. In fluid mechanics, this method has been known as the surface-singularity method. For an incompressible and irrotational fluid, it is well known that the velocity potential satisfies the Laplace equation. Hence Eq. (52), with the inhomogeneous term $f(\mathbf{r})$ dropped, is the key equation in analyzing various boundary problems for perfect fluids. In practical problems, such as the analyses for the air flow around an aircraft or a space shuttle flying with relatively low velocity, a distribution of vortices must often be taken into account. In such cases the final integral equation like Eq. (52) needs to be modified, but the BEM is still quite powerful. See for details the report by Morino *et al.* (1975).

4.3.2 Sound and Electromagnetic Waves. If the retarded Green's function G_R for the Helmholtz equation [Eq. (15)] is used in place of *G*, Eq. (52) turns out to be the key equation for the wave equation [here $f(\mathbf{r})$ therein is an inhomogeneous term of the wave equation]. Then if we let $u(\mathbf{r})$ stand for the velocity potential associated with a sound wave, the boundary values for $u(\mathbf{r})$ and $\partial u(\mathbf{r})/\partial n$ will be related, respectively, to the pressure and the velocity on Γ. For a region bounded by a rigid wall, it holds that $[\partial u(\mathbf{r})/\partial n]|_\Gamma=0$. When $f(\mathbf{r})=0$, Eq. (52) becomes a homogeneous integral equation for $g(\mathbf{r})$, yielding the eigenfrequencies for the sound modes established in that region, which can be, in an actual problem, an auditorium or a complicated resonator such as that of a violin. By converting the integral equation to the linear coupled equations, we can find the eigenvalues ω [involved in the Green's function through $\kappa=\omega/c$ in Eq. (15)].

An external problem may be formulated similarly to deal with sound propagation from a source with a complicated shape. Needless to say, the BEM for the wave equation is not limited to acoustics.

4.3.3 Elasticity. The final example we give on the application of the BEM is the problem of determining the strain tensor of an elastic body caused by a body force $\mathbf{f}(\mathbf{r})$ and a surface force $\mathbf{p}(\mathbf{r})$, both applied externally. Since the basic equation of elasticity is rather complicated, an analytical treatment is possible only for the exceptional case of very simple $\mathbf{f}(\mathbf{r})$ and $\mathbf{p}(\mathbf{r})$, applied to an elastic body whose shape, too, is very simple. In the BEM, these restrictions may be largely relaxed. First we need the Green's function for an infinite elastic body. For an isotropic and homogeneous system the tensor of the Green's functions satisfies (Landau and Lifshitz, 1986)

$$\Delta G_{ij}(\mathbf{r},\mathbf{r}')+\frac{1}{1-2\sigma}\sum_k\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_k}G_{kj}(\mathbf{r},\mathbf{r}')$$
$$=-\frac{2(1+\sigma)}{E}\delta_{ij}\delta(\mathbf{r}-\mathbf{r}') \qquad (53)$$

with Young's modulus E, Poisson's ratio σ, and Kronecker's delta δ_{ij}. Here G_{ij} is the ith component of the deformation field at $\mathbf{r}$ induced by the j-directed point force at $\mathbf{r}'$ with unit strength. The solution to Eq. (53) is known as Kelvin's solution, which reads

$$G_{ij}(\mathbf{r},\mathbf{r}')=\frac{1+\sigma}{8\pi E(1-\sigma)}[(3-4\sigma)\delta_{ij}+n_in_j]$$
$$\times\frac{1}{|\mathbf{r}-\mathbf{r}'|}, \qquad (54)$$

where $\mathbf{n}=(\mathbf{r}-\mathbf{r}')/|\mathbf{r}-\mathbf{r}'|$. In terms of G_{ij} one may obtain the basic integral equation on the external surface Γ, involving the body force $\mathbf{f}(\mathbf{r})$, an analog of Eq. (52) for the strain tensor $u_{ij}(\mathbf{r})$ (Brebbia and Walker, 1980). The boundary value for $\Sigma_j\partial u_{ij}/\partial x_j$ on Γ may be related to the given surface force $\mathbf{p}(\mathbf{r})$.

5. GREEN'S FUNCTIONS HAVING A DIRECT RELEVANCE TO PHYSICAL REALITY

The Green's functions treated in Secs. 3 and 4 were used mainly as a tool for solving a partial differential equation. The reader will recall that by definition they describe an output signal in response to a point input signal. This suggests that they are also usually related to a physical reality. That this is indeed so will be seen through the examples presented below.

5.1 Wave Front of Radiation Emitted from a Point Source and Huygens's Principle

The retarded Green's function given by Eq. (22) for the wave equation shows where the wave front of the radiation is found at time t, when it is emitted at a former time t' from the point source located at $\mathbf{r}'$. If we consider conversely a point of observation fixed at $\mathbf{r}$ and point sources distributed around that point, we can regard the radiation observed there at time t as a composition of the propagating wavelets that leave at $t=0$ the various point sources, whose distance from point $\mathbf{r}$ is ct (Fig. 2). In fact, this situation is well expressed by the solution of the following initial-value problem for the wave equation in three-dimensional space;

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2}-\Delta\right)u(\mathbf{r},t)=h(\mathbf{r},t) \qquad (55)$$

with the initial conditions

$$u(\mathbf{r},0)=f(\mathbf{r}),\quad u_t(\mathbf{r},0)=g(\mathbf{r}). \qquad (56)$$

The solution is given by Kirchhoff's formula (Baker and Copson, 1950; Morse and Feshbach, 1953)

$$u(\mathbf{r},t)=\int_{|\mathbf{r}-\mathbf{r}'|<ct}d^3r'\,\frac{h(\mathbf{r}',t-|\mathbf{r}-\mathbf{r}'|/c)}{4\pi|\mathbf{r}-\mathbf{r}'|}$$
$$+\frac{\partial}{\partial t}[tf(\mathbf{r})^{ct}]+tg(\mathbf{r})^{ct}. \qquad (57)$$

Here the first term takes into account the inhomogeneous term of Eq. (55) in the form of a retarded potential. The quantities f^{ct} and g^{ct} are the averages of their values on the surface of the sphere centered at $\mathbf{r}$ with radius ct: for example,

$$f(\mathbf{r})^{ct}=\int d\Omega\, f(\mathbf{r}+ct\hat{\mathbf{r}}_0)/4\pi, \qquad (58)$$

the integral being over the solid angle in the direction $\hat{\mathbf{r}}_0$ from the point $\mathbf{r}$. Equation (57) demonstrates in mathematical form the well-

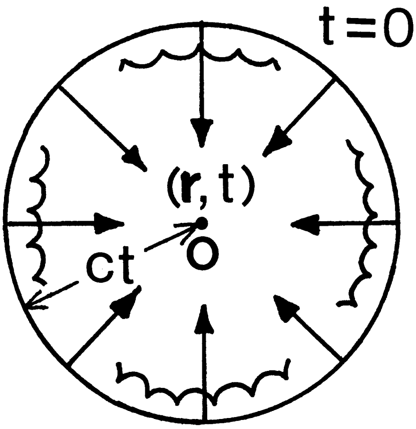

FIG. 2. Huygens's principle for wave propagation. The wave observed at the observation point $O(\mathbf{r},t)$ is the sum of the wavelets leaving, at $t=0$, the sources distributed on the sphere with radius ct, c being the velocity of the wave.

known Huygens's principle. Although the detailed derivation is omitted, the important role of the Green's function in this principle will be understood if the constraint imposed by Green's functions in Eq. (22) is compared with those involved in the integrals in Eqs. (57) and (58). The name "propagators" given very often to the Green's functions of the wave equation or its generalizations stems from the concepts involved in Huygens's principle.

5.2 Retarded Green's Function of Schrödinger's Equation

The inverse Fourier transform of the retarded version of the Green's function given by Eq. (38) is

$$G_R(\mathbf{r},t;\mathbf{r}',t')=\begin{cases}\Sigma_n\psi_n(\mathbf{r})\psi_n^*(\mathbf{r}')\exp[-(i/\hbar)E_n(t-t')], & t>t',\\ 0, & t<t'.\end{cases} \tag{59}$$

This quantity provides the probability amplitude for a particle observed at $(\mathbf{r}',t')$ to be found at $(\mathbf{r},t)$. To see this, we note that Eq. (59) is the solution of the Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{r},t)=H(\mathbf{r})\Psi(\mathbf{r},t), \tag{60}$$

subject to the condition that at time t'

$$\Psi(\mathbf{r},t')=\delta(\mathbf{r}-\mathbf{r}'). \tag{61}$$

This interpretation applies to an arbitrary Hamiltonian and is not restricted to that of an electron in a hydrogen atom, which we had in mind in Sec. 2.5. The above argument, when taken together with the brief explanation to be given in Sec. 7.5 on the path-integral approach to quantum mechanics, shows that the Green's function of the Schrödinger equation bridges the gap between quantum mechanics and classical dynamics (Sakurai, 1985). Also, in the perturbational treatment, we can follow an actual process of electron scattering in $(\mathbf{r},t)$ space, by making use of the Green's function (see Secs. 6.2 and 6.3).

5.3 Dislocations

A dislocation is a line of singularities in the deformation field of an elastic body. Equation (53) thus describes a dislocation if the strength and distribution of the δ functions of the right-hand side are specified appropriately. For example, around a straight edge dislocation on the z axis with a y-directed Burgers vector $\mathbf{b}$ ($|\mathbf{b}|=b$), the elastic deformation $\mathbf{u}(\boldsymbol{\rho})$ as a function of the two-dimensional coordinate $\boldsymbol{\rho}=(x,y)$ is given by (Landau and Lifshitz, 1986)

$$\mathbf{u}(\boldsymbol{\rho})=\mathbf{u}_0(\boldsymbol{\rho})+\mathbf{w}(\boldsymbol{\rho}). \tag{62}$$

Here the vector $\mathbf{u}_0(\boldsymbol{\rho})=(u_{0x}(\boldsymbol{\rho}),u_{0y}(\boldsymbol{\rho}),u_{0z}(\boldsymbol{\rho}))$ takes account of the multivaluedness of the deformation field and is defined by $u_{0z}(\boldsymbol{\rho})=0$ and

$$u_{0x}(\boldsymbol{\rho})+iu_{0y}(\boldsymbol{\rho})=i(b/2\pi)\ln(x-iy). \tag{63}$$

It is essentially a fundamental solution of the two-dimensional Laplace equation. On the other hand, the vector $\mathbf{w}(\boldsymbol{\rho})$ in Eq. (62) is obtained as

$$w_i(\boldsymbol{\rho})=\frac{Eb}{(1+\sigma)}\int_{-\infty}^{\infty}G_{iy}(\mathbf{r},\mathbf{r}')dz', \tag{64}$$

with $\mathbf{r}=(\boldsymbol{\rho},0)$, $\mathbf{r}'=(0,0,z')$, and G_{ij}, Kelvin's solution, given by Eq. (54). The integral over z' takes account of the line of δ-function singularities on the z axis. From Eqs. (63) and (64), one can determine the strain and stress fields around a straight edge dislocation.

5.4 Magnetic Field around a Vortex Line in a Type-II Superconductor

It is well known, as the Meissner effect, that a magnetic field applied externally cannot penetrate into a superconductor. In a type-II superconductor, however, superconducting and nonsuperconducting regions coexist in a phase called the mixed phase, which is realized above a certain threshold strength for the magnetic field: As the field strength increases to pass the threshold, the magnetic flux begins to penetrate the sample in the form of a quantized flux line, called a vortex. Well below the upper critical field, above which the superconducting phase can no longer exist, the density of such vortices is so low that one may treat each of them as independent. It is shown that around an isolated rectilinear vortex, the microscopic magnetic field configuration is determined by (in cgs units following the convection of this field)

$$(\Delta-\lambda^{-2})\mathbf{h}(\mathbf{r}) = -\phi_0\lambda^{-2}\hat{\mathbf{z}}\delta(\boldsymbol{\rho}), \qquad (65)$$

$\phi_0=hc/2e$ being the flux quantum and λ the penetration depth (Fetter and Hohenberg, 1969). Here $\hat{\mathbf{z}}$ is the unit vector in the z direction and the vortex line is assumed to be situated on the z axis $\boldsymbol{\rho}=0$, $\boldsymbol{\rho}=(x,y)$ being the two-dimensional position vector. The field $\mathbf{h}(\mathbf{r})$ is thus expressed by the Green's function of the two-dimensional Helmholtz equation [note, however, that κ^2 in Eq. (13) is replaced by $-\lambda^{-2}$]:

$$\mathbf{h}(\mathbf{r}) = (\phi_0\lambda^{-2}/2\pi)\hat{\mathbf{z}}K_0(\rho/\lambda), \qquad (66)$$

with K_0 the modified Hankel function of order zero. If the right-hand side of Eq. (65) is replaced by a two-dimensional periodic distribution of δ functions, one may determine the magnetic field set up in a vortex lattice and calculate the gain in the free-energy density. The discussion of the equilibrium lattice structure that minimizes the free-energy density provides us with a basis for the more complete analysis due to Abrikosov *et al.*, based on the Ginzburg-Landau equations. See for details Fetter and Hohenberg (1969).

6. PERTURBATIONAL TREATMENT TO OBTAIN GREEN'S FUNCTIONS

When an operator can be divided into two parts and the Green's function for one of the two is obtained straightforwardly, one may treat the remaining part perturbatively to obtain the full Green's function. For illustration, let us consider the stationary solution of the Schrödinger equation for the Hamiltonian of the following form:

$$H=H_0+V. \qquad (67)$$

It is very convenient to define the Green's function using the operator $(E-H)^{-1}$, which is sometimes called the Greenian, and to examine it as a function of the energy variable E. One may show that the quantity $\langle\mathbf{r}|(E-H)^{-1}|\mathbf{r}'\rangle$, the $(\mathbf{r},\mathbf{r}')$ matrix element of the Greenian, behaves as a Green's function, and one may reasonably denote it simply as $G(\mathbf{r},\mathbf{r}';E)$ (see, e.g., Schiff, 1968). [The present definition leads to the Green's function $(i\hbar)^{-1}$ times that defined in Eq. (38)]. Rewriting the operator $(E-H)^{-1}$ as a power series with respect to the perturbation V, we obtain the following integral equation for G:

$$G(\mathbf{r},\mathbf{r}';E)=G_0(\mathbf{r},\mathbf{r}';E)+\int d^3r_1 G_0(\mathbf{r},\mathbf{r}_1;E) \times V(\mathbf{r}_1)G(\mathbf{r}_1,\mathbf{r}';E), \qquad (68)$$

$G_0(\mathbf{r},\mathbf{r}';E)$ being the unperturbed Green's function $\langle\mathbf{r}|(E-H_0)^{-1}|\mathbf{r}'\rangle$. The expression is given for a local operator V, i.e., $\langle\mathbf{r}|V|\mathbf{r}'\rangle=V(\mathbf{r})\delta(\mathbf{r}-\mathbf{r}')$. Iterating the right-hand side, we find

$$G=G_0+G_0VG_0+G_0VG_0VG_0+\cdots=G_0+G_0TG_0. \qquad (69)$$

Here the simplified notation of G_0VG_0, etc., should be interpreted as in the second term of Eq. (68) by supplementing the arguments and integrals. The series in the first equality, the Born series, defines the operator T, which is usually called the t matrix by identifying its $(\mathbf{r},\mathbf{r}')$ matrix element $T(\mathbf{r},\mathbf{r}';E)$ with the operator itself. It is an effective scattering potential, taking into account the multiple scattering effect to all orders of V through the relation

$$\begin{aligned} T&=V+VG_0V+VG_0VG_0V+\cdots \\ &=V+VGV. \end{aligned} \qquad (70)$$

Note that the Green's function in the second equality is G, not G_0, which already takes full account of the effect of V.

6.1 Slater-Koster Model

As shown by Eqs. (38) and (39), the eigenvalues for a Hamiltonian operator H are given by the poles of the Green's function $\langle\mathbf{r}|(E-H)^{-1}|\mathbf{r}'\rangle$. The Slater-Koster model is a typical model in which the series (69) is exactly calculable. Usually, it is defined by the model Hamiltonian which has a δ-function-type (contact-type) perturbation, $V(\mathbf{r})=v_0\delta(\mathbf{r})$, v_0 being a constant for the potential strength. In this case, Eq. (69) leads to

$$G(\mathbf{r},\mathbf{r}';E)=G_0(\mathbf{r},\mathbf{r}';E)+G_0(\mathbf{r},0;E)\{v_0/[1-v_0G_0(0,0;E)]\}G_0(0,\mathbf{r}';E). \qquad (71)$$

Thus, the zeros of the denominator of the second term as a function of E provide the eigenvalues for the full Hamiltonian:

$$1-v_0G_0(0,0;E)=0. \tag{72}$$

Considering that G_0 has poles at the unperturbed energy eigenvalues [the eigenvalues of the Hamiltonian H_0 in Eq. (67)], we see that this equation has a form quite adequate to see how these eigenvalues are modified by the perturbation. Using Eq. (72), one can discuss, for example, how a plane-wave state $\mathbf{k}$ is affected by the potential $V(\mathbf{r})$ and whether or not a bound state appears for an attractive potential $V(\mathbf{r})$ for a given unperturbed energy spectrum. In condensed-matter physics these problems are very frequently encountered in dealing with electrons, phonons, magnons, etc. For a detailed analysis, see, e.g., the book by Economou (1990).

The treatment for a photon along this line is just the perturbational treatment of the Maxwell equations. For example, the reflection and transmission of light for a rough solid surface was analyzed successfully by this approach (see, e.g., Maradudin *et al.*, 1990).

6.2 Scattering Cross Section of a Plane Wave from a Scatterer with Spherical Symmetry

In the scattering of an incident wave ψ_0 of, say, an electron with energy E by a perturbing potential $V(\mathbf{r})$, the total wave function for the Schrödinger equation $H\psi=E\psi$ with H given by Eq. (67) may be expressed as (Schiff, 1968)

$$\begin{aligned}\psi(\mathbf{r})&=\psi_0(\mathbf{r})+\int d^3r'G_0(\mathbf{r},\mathbf{r}';E)\,V(\mathbf{r}')\psi(\mathbf{r}')\\&=\psi_0(\mathbf{r})+\iint d^3r_1d^3r_2G_0(\mathbf{r},\mathbf{r}_1;E)\\&\qquad\times T(\mathbf{r}_1,\mathbf{r}_2;E)\psi_0(\mathbf{r}_2).\end{aligned} \tag{73}$$

The second equality is obtained by iterating the first equality and using the definition (70) for the t matrix T. For a spherically symmetric potential $V(\mathbf{r})=V(r)$, we may go further by resolving $\psi(\mathbf{r})$ and $\psi_0(\mathbf{r})$ into partial waves according to the magnitudes of angular momentum $\hbar l$. The partial-wave expansion of the incident plane wave with wave vector $\mathbf{k}_i$ is

$$\psi_0(\mathbf{r})=e^{i\mathbf{k}_i\cdot\mathbf{r}}=\sum_{l=0}^{\infty}(2l+1)i^lj_l(k_ir)P_l(\cos\theta), \tag{74}$$

where $j_l(k_ir)$ is the spherical Bessel function of order l, the scattering angle θ is the angle between $\mathbf{k}_i$ and $\mathbf{r}$, and $P_l(\cos\theta)$ is the Legendre function of order l. Asymptotically, the solution ψ in Eq. (73) behaves as

$$\psi(\mathbf{r})=\psi_0(\mathbf{r})+f(\theta)e^{ik_sr}/r, \tag{75}$$

with k_s the wave number of the scattered spherical wave. Comparing Eqs. (73) and (75) and using the form given by Eq. (15), we find that the amplitude of the outgoing spherical wave is given by

$$f(\theta)=-(m/2\pi\hbar^2)T(\mathbf{k}_s,\mathbf{k}_i;E), \tag{76}$$

where

$$\begin{aligned}T(\mathbf{k}_s,\mathbf{k}_i;E)=\iint& e^{-i\mathbf{k}_s\cdot\mathbf{r}}T(\mathbf{r},\mathbf{r}';E)\\&\times e^{i\mathbf{k}_i\cdot\mathbf{r}'}d^3r\,d^3r',\end{aligned} \tag{77}$$

with $\mathbf{k}_s$ the wave vector of the scattered wave, directed from the origin to the observation point $\mathbf{r}$ ($k_s=k_i$ by energy conservation). The t matrix $T(\mathbf{k}_s,\mathbf{k}_i;E)$ for the two wave vectors, both related to the incident energy E through the relation $k_i=k_s=(2mE/\hbar^2)^{1/2}$, is called the t matrix on the energy shell and is known to be expressed in terms of the phase shift $\delta_l(E)$ of the partial wave l. In this way, $f(\theta)$ is finally expressed as

$$f(\theta)=\frac{1}{2ik_i}\sum_{l=0}^{\infty}(2l+1)(e^{2i\delta_l}-1)P_l(\cos\theta). \tag{78}$$

The differential and total cross sections are then obtained from $|f(\theta)|^2$.

To carry out the partial-wave expansion for the Maxwell equations, we need the tensor of the Green's functions (Mahan, 1990) and the vector spherical wave for each l (Stratton, 1941). With the difference that we now require two kinds of phase shifts for each l because of the character of light as a transverse vector field, the Mie and Rayleigh scatterings of light (Born and Wolf, 1975) may be treated compactly as in the electron case.

6.3 Band Structure of Electrons and Photons and Diffraction Experiments

The band structure of electron states in solids may be viewed as arising from the electron scattering by atoms arrayed periodically in a lattice. With the difference that the perturbation V is now due to a periodic array of scatterers, Eqs. (67)–(70) still hold here without modification. Since the poles of the full Green's function are identical to those of the t matrix [Eq. (69)], the calculation of the t matrix for arrayed scatterers is equivalent to the band-structure calculation. If we denote the multiple-scattering effect of the kth scatterer by the t matrix t_k, the scattering from the array as a whole is described by the following t matrix:

$$T = \sum_{k} t_k + \sum_{\substack{k,k' \\ (k \neq k')}} t_k G_0 t_{k'} + \cdots , \qquad (79)$$

where G_0 is the Green's function for free motion. The first term describes the scattering from a single site. The second exhibits the process where an electron, once scattered by the kth scatterer, propagates to another site k' and undergoes a multiple scattering there. The constraint $k \neq k'$ eliminates double counting, because the term $t_k G_0 t_k$ to be removed is already involved in the single t matrix t_k [see the series expansion of the first equality of Eq. (70)]. The higher-order terms in Eq. (79) incorporate processes involving three or more scatterers. These scattering processes are exhibited in Fig. 3. If the potentials are spherical and nonoverlapping with each other, as is usually the case to a good approximation, only the on-the-energy-shell values are involved in each t_k (see Sec. 6.2). Since G_0 sandwiched between t_k and $t_{k'}$ in Eq. (79) describes the free propagation of an electron from one atom k to another k', it depends solely upon the structure of the lattice. Therefore Eq. (79) shows that the band structure of electron energies in solids is determined quite generally by two quantities: the phase shifts of the atoms and the structure factor of the periodic array they form. This is indeed explicit in the Korringa-Kohn-Rostoker (KKR) eigenvalue equation for the band structure of electrons. See the monograph by Lukes (1969) for the derivation of the KKR equation based on Eq. (79).

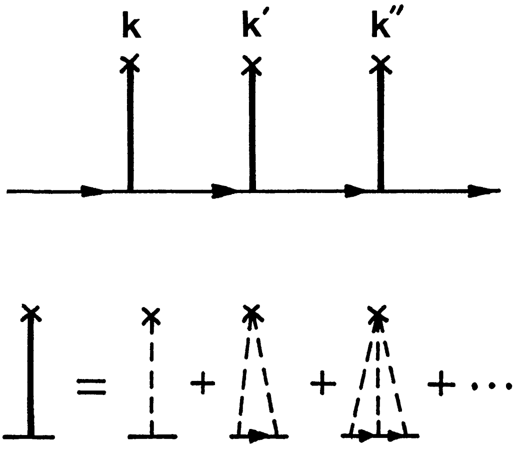

FIG. 3. Scattering of an electron in a solid. The vertical solid line at site k is the t matrix t_k used in Eq. (79) and the arrowed horizontal lines show G_0. The t matrix t_k is defined by the lower graph, which shows the series given by Eq. (70) with the dotted lines for the atomic potential V.

When we apply Eq. (73) to a periodic array of scatterers, the solution $\psi(\mathbf{r})$ describes how the atomic array responds to an incident electron. When, in particular, an incident electron from the outside has an energy that is not very different from the positions of the poles of the t matrix, the amplitude of the scattered wave reflects, through $T(\mathbf{r}_1,\mathbf{r}_2;E)$ of Eq. (73), the detailed band structure as a function of incident energy E and wave vector. This is essentially the origin of the fine structure observed in the usual low-energy-electron diffraction (Tong, 1975).

Similar fine structure is expected to arise in the diffraction of visible light, i.e., in the light scattered from a periodic array of scatterers with periodicity in the visible range. X-ray diffraction is simply its high-energy limit. For details of the diffraction of visible light reflecting the photon band structure, see Ohtaka (1980) and Inoue and Ohtaka (1982) and the analysis by Yablonovitch and Gmitter (1989).

7. GREEN'S FUNCTIONS IN MANY-BODY THEORIES

The developments of condensed-matter physics and elementary-particle physics owe considerably to the use of Green's functions in the treatment of many-body interactions. How they are defined and why they have been so widely used will be shown through several examples. For more details, see the books by, e.g., Abrikosov *et al.* (1963), Fetter and Walecka (1971), Economou (1990), and Mahan (1990). For illustration we have in mind a model Hamiltonian for interacting spinless particles, either bosons or fermions:

$$K=\int d^3r\,\Psi^\dagger(\mathbf{r})[(-\hbar^2/2m)\Delta-\mu]\Psi(\mathbf{r})$$
$$+\frac{1}{2}\int\int d^3r\,d^3r'\Psi^\dagger(\mathbf{r})\Psi^\dagger(\mathbf{r}')$$
$$\times v(|\mathbf{r}-\mathbf{r}'|)\Psi(\mathbf{r}')\Psi(\mathbf{r})$$
$$=K_0+K' \tag{80}$$

expressed in the second quantized form. Here $\Psi^\dagger(\mathbf{r})$ and $\Psi(\mathbf{r})$ are the field operators for creating and annihilating, respectively, a particle at position $\mathbf{r}$. The term proportional to the chemical potential μ in the first term on the right-hand side shows that we are considering the grand canonical ensemble. The two-body correlation described by the operator K' may be treated using Green's functions.

7.1 Single- and Two-Particle Green's Functions

The single-particle Green's function is defined in many-body theories by

$$G(x,x')=-\frac{i\langle\Phi_0|T(\Psi(x)\Psi^\dagger(x'))|\Phi_0\rangle}{\langle\Phi_0|\Phi_0\rangle}. \tag{81}$$

Here x and x' denote $(\mathbf{r},t)$ and $(\mathbf{r}',t')$, respectively, Φ_0 is the exact ground state for the Hamiltonian given by Eq. (80), and the time-dependent operators $\Psi(x)$ etc. are the operators in the Heisenberg picture defined by

$$\Psi(x)=e^{iKt/\hbar}\Psi(\mathbf{r})e^{-iKt/\hbar}. \tag{82}$$

The symbol T in Eq. (81) is the time-ordering operator that is defined by the rule

$$T(\Psi(x)\Psi^\dagger(x'))=\begin{cases}\Psi(x)\Psi^\dagger(x'), & t>t',\\ \pm\Psi^\dagger(x')\Psi(x), & t<t',\end{cases} \tag{83}$$

the upper (lower) sign referring to bosons (fermions). Aside from the presence of the time-ordering operator T, the Green's function defined by Eq. (81) can be given the physical interpretation of a probability amplitude, analogous to that given by Eq. (59). Indeed, when the particle interaction vanishes, i.e., $K=K_0$ in Eq. (80), G tends to the unperturbed Green's function

$$G_0(x,x')=-i\langle 0|T(\Psi(x)\Psi^\dagger(x'))|0\rangle, \tag{84}$$

$|0\rangle$ being the normalized ground state for K_0. It satisfies

$$\left(i\hbar\frac{\partial}{\partial t}+\frac{\hbar^2}{2m}\Delta+\mu\right)G_0(x,x')=\hbar\delta(x-x'). \tag{85}$$

The inhomogeneous term comes from the presence of the operator T, the origin of the step function via Eq.(83), and the commutation (anticommutation) relation of the field operators for bosons (fermions). Through Eq. (85), we see obviously that iG_0 is the many-body analog of the single-particle Green's function treated in Sec. 5.2.

The first reason why the single-particle Green's function is so useful is that many important observables such as the ground-state energy, number or current density, etc. of the many-body system, are expressed in terms of the Green's function (81). The second reason is that the effect of K' may be handled by applying Wick's theorem and interpreted visually with the help of Feynman diagrams (see Sec. 7.2). Third, part of the interaction process is taken into account to all orders with respect to K'.

For example, the modification of a single-particle energy due to the mutual interaction K' is taken into account by Dyson's equation

$$G(\mathbf{k},\omega)=G_0(\mathbf{k},\omega)+G_0(\mathbf{k},\omega)\Sigma(\mathbf{k},\omega)G(\mathbf{k},\omega) \tag{86}$$

in the Fourier space $(\mathbf{k},\omega)$. From this, we find

$$G(\mathbf{k},\omega)=[\omega-(\hbar k^2/2m)+(\mu/\hbar)-\Sigma(\mathbf{k},\omega)]^{-1}, \tag{87}$$

where we have used

$$G_0(k,\omega)=[\omega-(\hbar k^2/2m)+(\mu/\hbar)+i\epsilon\,\mathrm{sgn}(\omega)]^{-1} \tag{88}$$

given by Eq. (85). Note that G_0 has a causal form (see Sec. 2.2). The complex quantity $\Sigma(\mathbf{k},\omega)$ is called the self-energy part. The form (87) provides an exact expression for G, if the self-energy part is exactly given. In this sense, Dyson's equation (86) is one of the key equations in treating many-body interactions. The imaginary part of $\Sigma(\mathbf{k},\omega)$ determines the lifetime of a plane-wave state $\mathbf{k}$, caused by the many-body interaction. Although it is generally a hopeless task to attempt to make an exact calculation of $\Sigma(\mathbf{k},\omega)$ with all possible many-body processes included, an important subset may usually be taken into account. For a system of electrons interacting with Coulomb repulsion (the electron-gas model), for example, we may now say that some of the physical quantities have so far been obtained almost exactly.

The two-particle Green's function is defined by

$$G(x_1,x_2;x_1',x_2') = (-i)^2\langle\Phi_0|T(\Psi(x_1)\Psi(x_2) \times\Psi^\dagger(x_2')\Psi^\dagger(x_1'))|\Phi_0\rangle/\langle\Phi_0|\Phi_0\rangle. \quad (89)$$

It is so called because it deals with the two particles created at x_1' and x_2'. Since a correlation function between two physical quantities is usually expressed by the two-particle Green's function, the latter is an important quantity, yielding transport coefficients, conductivity, and dielectric constant, for example (see Sec. 7.4). As mentioned before, Wick's theorem is a key theorem for calculating the Green's function. The names "random-phase approximation," "ladder approximation," etc., are each assigned to a special choice of infinite series of Feynman diagrams considered in two-particle Green's functions. We give in Fig. 4 two typical examples, which exhibit how the two particles interact with each other, sometimes involving other particles.

7.2 Wick's Theorem and Feynman Diagrams

The Green's functions (81) and (89) are defined in terms of the operators in the Heisenberg picture for the full Hamiltonian K. In calculating them perturbationally, using the relationships between Φ_0 and $|0\rangle$ and between the operators in the Heisenberg and interaction representations, we encounter the following types of quantities:

$$\langle 0|T(\Psi_0(x_1)\Psi_0(x_2)\Psi_0(x_3)\cdots\Psi_0^\dagger(x_{2n-1}) \times\Psi_0^\dagger(x_{2n}))|0\rangle, \quad (90)$$

where $\Psi_0(x)$ etc., are the operators in the interaction picture defined by

$$\Psi_0(x) = e^{iK_0t/\hbar}\Psi(\mathbf{r})e^{-iK_0t/\hbar}. \quad (91)$$

Quantities such as (90) are the expectation values in the state $|0\rangle$ of the time-ordered product of an equal number of operators $\Psi_0(x)$ and $\Psi_0^\dagger(x)$. Wick's theorem handles just such quantities and guarantees that the quantity given above may be resolved into a sum of products of n unperturbed Green's functions G_0's, formed by pairing a Ψ_0 with a $\Psi_0^\dagger$ in all possible ways. For a rigorous proof, see, e.g., Fetter and Walecka (1971). In this way the perturbation series for a Green's function, either a single-particle one or a two- or many-particle one, is expressed as a sum of products of a number of G_0's, and each product is given a graphical representation called Feynman diagram (Feynman graph), where a line represents an unperturbed propagator G_0. Some examples were already given for the two-particle Green's functions in Fig. 4. For K given by Eq. (80), several second- and higher-order corrections for the single-particle Green's function are shown in Fig. 5.

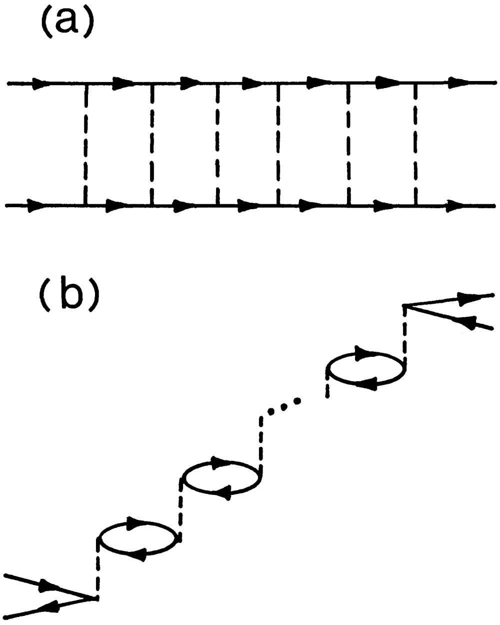

FIG. 4. Two-particle Green's function taking into account the particle interaction: (a) ladder diagram, (b) random-phase approximation. The dotted lines show the particle interaction, v of Eq. (80), and the arrowed solid lines show the unperturbed Green's function G_0, defined by Eq. (84). When v is an attractive short-range interaction for an narrow energy range, the ladder diagram for a singlet pair of electrons leads to superconductivity. When v is a repulsive Coulomb potential, the random-phase approximation leads, in the low-density limit, to the exact ground-state energy of interacting electrons.

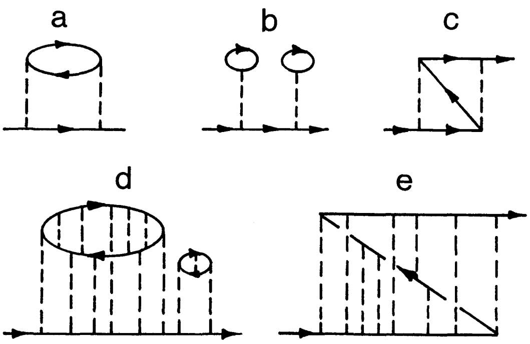

FIG. 5. Examples of Feynman diagrams. The graphs *a* through *c* show examples of second-order corrections to the single-particle Green's functions for the Hamiltonian given by Eq. (80). The vertical dotted lines show the particle interaction *v*. The graphs *d* and *e* are two higher-order corrections.

7.3 Temperature Green's Functions

The temperature Green's functions are essentially the Green's functions treated in Sec. 7.1 with the time argument t and t' replaced by the imaginary time $-i\tau$ and $-i\tau'$, respectively, and with the expectation value $\langle\Phi_0|\cdots|\Phi_0\rangle$ for the ground state replaced by the thermal average $\langle\cdots\rangle$. The temperature Green's function also goes under the name of Matsubara, who first introduced them (Kubo *et al.*, 1991). They are convenient tools for calculating various physical quantities in thermal equilibrium—for instance, the single-particle energy and thermodynamic potential. Wick's theorem holds here again. Because of the periodic properties with respect to the argument $\tau-\tau'$, the Fourier components of the Green's function $G(\tau-\tau')$ are defined at the discrete frequencies $\omega_n=(2n+1)\pi k_BT/\hbar$ for fermions and $\omega_n=2n\pi k_BT/\hbar$ for bosons, n being an integer, T the temperature, and k_B the Boltzmann constant. For those frequencies, the Dyson equation (86) holds and is solved algebraically. The single-particle energy is obtained by analytically continuing $G(\omega_n)$ from the discrete points $i\omega_n$ on the imaginary axis to the complex ω plane. For the BCS Hamiltonian of superconductivity, for example, all the predictions of the BCS theory (Bardeen *et al.*, 1957) have been reproduced straightforwardly in this way and the Ginzburg-Landau equations, originally introduced semiphenomenologically, have been given an unambiguous microscopic foundation. The developments of the theory and experiment of superconductivity owed much to the use of temperature Green's functions (see, e.g., Abrikosov *et al.*, 1963; Parks, 1969; Mahan, 1990). The advantage of the two-particle temperature Green's functions can be best appreciated in connection with the linear response theory of Kubo treated below.

7.4 Linear Response Theory of Kubo and Temperature Green's Functions

How the system responds to a small external signal is summarized by the Kubo formula of linear response theory. Its essence lies in that the response function is given by a spatial and temporal correlation function of fluctuations, calculated in thermal equilibrium without the external perturbation. For example, when a weak electric field, whose v component is $E_v(\mathbf{r}',t')$, of frequency ω is an input signal $I(\mathbf{r}',t')$ in the input-output relation given by Eq. (2), an Ohmic electric current at the position $(\mathbf{r},t)$ is the output signal $O(\mathbf{r},t)$. The response function, the Green's function $G(x,x')$ in Eq. (2), is in this case the ac conductivity tensor $\sigma_{\mu v}(x,x')$. Here the fluctuation in the Kubo formula is that of the current density. Letting $j_\alpha(x)$ be the operator for the current density in the α direction in the Heisenberg picture, the fluctuation is defined by $j_\alpha(x)-\langle j_\alpha(x)\rangle$ [$=j_\alpha(x)$, because $\langle j_\alpha(x)\rangle=0$ in equilibrium]. The response function is then expressed by

$$\sigma_{\mu v}(x,x')=i(ne^2/m\omega)\delta_{\mu v}\delta(\mathbf{r}-\mathbf{r}')+(-1/\hbar\omega)\times\langle[j_\mu(x),j_v(x')]\rangle\theta(t-t'), \qquad (92)$$

with n the electron density, m the mass, and the Heaviside step function in the second term guaranteeing causality (see Sec. 2.2). The square bracket in the second term is the commutator and $\langle\ \rangle$ denotes the thermal average. Because of the presence of the step function, quantities like that in the second term of Eq. (92) are called retarded Green's functions in linear response theory, in analogy with their counterpart for the wave equation in Sec. 2.2.

The way of calculating the Fourier transform of the correlation function involved in the second term of Eq. (92) is summarized by the theorem, sometimes called the Abrikosov-Gor'kov-Dzyaloshinski-Fradkin theorem, which relates a retarded Green's function to the corresponding temperature Green's function (Abrikosov *et al.*, 1963). According to this theorem, the first step is the calculation of the two-particle temperature Green's function

$$\tilde{G}_2(\mathbf{r},\tau;\mathbf{r}',\tau')=\langle T_\tau(j_\mu(\mathbf{r},\tau)j_\nu(\mathbf{r}',\tau'))\rangle \tag{93}$$

for the Fourier component $\tilde{G}_2(\mathbf{k},\omega_n)$, and the second is the analytic continuation of the result to the real frequency ω, carried out simply by changing $i\omega_n$ to $\omega+i\epsilon$. The first step is carried out with the help of Wick's theorem, and the second is a procedure analogous to that used in obtaining a retarded Green's function in Sec. 2.2 (Abrikosov *et al.*, 1963).

Since this method of calculation using the temperature Green's functions was introduced, the usefulness of the Kubo formula has increased remarkably. This was indeed one of the key steps in the development of condensed-matter physics.

7.5 Green's Functions in Quantum Field Theories

One typical way of quantizing a classical Lagrangian is based on the functional integral. One of the merits of this method is that we may treat the field theory and statistical physics on the same footing (Amit, 1984).

The Green's functions and the related quantities developed in this field are outlined here by taking an interacting scalar Bose field $\phi(x)$ in four-dimensional space as an example (Ramond, 1981). The Lagrangian density $L(\phi(x))$ for the model called the ϕ^4 model is defined [in natural units, where $c=\hbar=1$ and the dimension of the field $\phi(x)$ is $(\text{length})^{-1}$] by

$$L(\phi)=L_0(\phi)+L_{\text{int}}(\phi), \tag{94}$$

with

$$\begin{aligned}L_0(\phi)&=\phi_t(x)^2-\phi_x(x)^2-\mu^2\phi(x)^2,\\ L_{\text{int}}(\phi)&=(\lambda/4!)\phi^4(x),\end{aligned} \tag{95}$$

ϕ_t and ϕ_x being $\partial\phi/\partial t$ and $\partial\phi/\partial x$, respectively. The functional $Z(J)$ for generating the Green's functions is defined by

$$Z(J)=\mathcal{N}^{-1}\int \mathscr{D}\phi \exp\left(i\int [L(\phi)+J\phi]d^4x\right). \tag{96}$$

Here $J(x)$ is a fictitious external field linearly coupled to $\phi(x)$, the factor $\mathcal{N}^{-1}$ normalizes $Z(J)$ such that $Z(0)=1$, and the measure $\mathscr{D}\phi$ of the functional integral is defined by

$$\mathscr{D}\phi=\lim_{N\to\infty}\prod_{n=1}^{N} d\phi(x_n). \tag{97}$$

Here it is understood that the integral over x in Eq. (96) is treated as a sum over the integrand at N discrete points $\{x_n\}$ in four-dimensional space, and the limit for N is taken after the integrals over the N independent variables $\{\phi(x_n)\}$ have been carried out. Equation (96) is a generalization of the quantization scheme for the classical one-dimensional motion of a particle carried out through the quantity

$$F(t_2,t_1)=\int \mathscr{D}q \exp\left(\frac{i}{\hbar}\int_{t_1}^{t_2} L(q(t),q'(t))dt\right), \tag{98}$$

where the classical Lagrangian L is defined in terms of $q(t)$ and $q'(t)$, the particle position and velocity at time t, respectively. With the restriction $q(t_1)=q_1$ and $q(t_2)=q_2$ imposed on the path, it may be shown that, if the measure $\mathscr{D}q$ is defined appropriately (Sakurai, 1985), Eq. (98) is identical to the quantum-mechanical probability amplitude of a particle, just described by the Green's function treated in Secs. 2.3.2 and 5.2. Thus the quantization through the functional integral is well established.

From Eq. (96), the Green's function is obtained as follows:

$$\begin{aligned}G(x_1,x_2)&=\left.\frac{\delta^2 Z(J)}{\delta J(x_1)\delta J(x_2)}\right|_{J=0}\\ &=-\mathcal{N}^{-1}\int \mathscr{D}\phi\, T\left[\phi(x_1)\phi(x_2)\times\exp\left(i\int L(x)d^4x\right)\right],\end{aligned} \tag{99}$$

where the functional derivative $\delta Z(J)/\delta J(x_1)$ is defined by the ratio of the induced change of $Z(J)$ to the infinitesimal variation $\delta J(x_1)$ at x_1 of the external field. Many-point correlation functions, $G_2(x_1,x_2,x_3,x_4)$, etc., are defined similarly. The unperturbed Green's function is then calculated from $Z_0(J)$, the generating function for $L=L_0$. It is shown that

$$Z_0(J)=\exp\left(\frac{i}{2}\int\int J(x)g(x-y)\times J(y)d^4x\,d^4y\right), \qquad (100)$$

with $g(x-y)$ the Green's function for the Klein-Gordon equation, G_R, G_A, or G_C of Eq. (29), according to the way of avoiding the poles in the ω integral (see Secs. 2.2 and 2.3.3). From Eq. (99), the unperturbed Green's function G_0 is then obtained as

$$G_0(x_1,x_2)=ig(x_1-x_2). \qquad (101)$$

In order to take into account L_{int}, the following relation is useful:

$$Z(J)=\mathcal{N}^{-1}\exp\left[i\int d^4x\,L_{\text{int}}\left(-\frac{i\delta}{\delta J(x)}\right)\right]\times Z_0(J), \qquad (102)$$

$Z_0(J)$ being given by Eq. (100). With Eqs. (100) and (102), we can calculate Green's functions to any desired order. For example, to obtain the perturbation expansion of G with respect to the parameter λ involved in L_{int}, we have only to expand the exponential of Eq. (102) in terms of L_{int}.

It is remarkable that Wick's theorem, Dyson's equation, Feynman diagrams, etc., all apply here without any modification. The difficulty of ultraviolet divergence is thus handled and then a number of physical quantities are defined by a procedure that extracts the finite part of every term in the perturbation series. This is the regularization and renormalization program.

To connect field theory with statistical physics, we need only to change the time t to the imaginary time $-i\tau$ (called in this field the Wick rotation), just as in introducing the temperature Green's functions in Sec. 7.3. The point is that the generating function $Z(J)$ converts itself essentially to a partition function. In our example this will be seen by noting that the Lagrangian is then transformed to minus the Hamiltonian. Taking advantage of this remarkable connection, and combining it with renormalization-group procedures, the critical properties of a statistical system near a second-order phase transition can be discussed using field-theoretical methods. For example, the critical index of the specific heat of the Ising model can be successfully obtained in this way. See, for more detail, e.g., Amit (1984) and Itzykson and Drouffe (1989).

GLOSSARY

Adjoint Operator: For a second-order differential operator L with respect to x, y, and z, we may transform the integral in the region Ω,

$$\int_\Omega v(\mathbf{r})L[u(\mathbf{r})]d^3r,$$

by integrating by parts until all the derivatives on $u(\mathbf{r})$ in the volume integral are transferred to $v(\mathbf{r})$. Finally the volume integral reduces to the form

$$\int_\Omega u(\mathbf{r})M[v(\mathbf{r})]d^3r,$$

with a new differential operator M, called the adjoint operator of L. Thus we obtain the following relation, the generalized Green's formula:

$$\int_\Omega (vL[u]-uM[v])d^3r = \int_\Gamma (Px_n+Qy_n+Rz_n)ds.$$

The right-hand side shows the surface integral that remains in the integration by parts, where x_n, y_n, and z_n are the directional cosines of the outward normal $\hat{\mathbf{n}}$ to the boundary Γ, and P, Q, and R are functions of u, v, and their derivatives, determined dependent upon the form of L (see, e.g., Courant and Hilbert, 1937).

Dirichlet Problem: A boundary-value problem of a differential equation that seeks a solution with a specified boundary value. These types of boundary conditions are called the Dirichlet condition.

Fredholm Integral Equation: For an unknown function $u(\mathbf{r})$ in the domain Ω, the following integral equation is called the Fredholm integral equation of the second kind:

$$u(\mathbf{r})-\int_\Omega K(\mathbf{r},\mathbf{r}')u(\mathbf{r}')d^3r'=f(\mathbf{r}),$$

$f(\mathbf{r})$ being a given function and K the integral kernel. The Fredholm integral equation of the first kind is the one with the first term $u(\mathbf{r})$ missing on the left-hand side.

Generalized Green's Formula: For the Laplacian operator the Green's formula is given by Eq. (4). It is generalized to an arbitrary second-order differential operator L and its adjoint M in the form shown in the definition of the adjoint operator.

Homogeneous Boundary Condition: The boundary condition such as $u(\mathbf{r})|_\Gamma=0$ or $(\partial/\partial n)u(\mathbf{r})|_\Gamma=0$ on the boundary Γ of the domain under consideration is a homogeneous boundary condition. In general, homogeneous boundary conditions consist of relations between the value of $u(\mathbf{r})|_\Gamma$ and its derivative $(\partial/\partial n)u(\mathbf{r})|_\Gamma$. If $u(\mathbf{r})$ satisfies the homogeneous boundary condition, so does $cu(\mathbf{r})$, c being an arbitrary constant. Boundary conditions for which this does not hold are called inhomogeneous boundary conditions. Examples are $u(\mathbf{r})|_\Gamma=f(\mathbf{r})$ or $(\partial/\partial n)u(\mathbf{r})|_\Gamma=g(\mathbf{r})$, with given nonzero $f(\mathbf{r})$ or $g(\mathbf{r})$ defined on the boundary.

Internal or External Boundary-Value Problem: When the solution of a boundary-value problem is sought inside the boundary on which a boundary condition is imposed, it is called the internal boundary-value problem. The problem for the outside region is the external problem.

Neumann Problem: The boundary condition specifying $\partial u/\partial n$, the derivative in the direction of the outward normal to the boundary, is called the Neumann condition. A boundary-value problem with a Neumann condition is a Neumann problem.

Vector Spherical Waves: Let Φ be a scalar function satisfying the scalar Helmholtz equation

$$(\Delta+k^2)\Phi=0.$$

From Φ one may construct three vector fields, $\mathbf{L}=\mathrm{grad}\Phi$, $\mathbf{M}=\mathrm{rot}(\mathbf{a}\Phi)$, and $\mathbf{N}=(1/k)\mathrm{rot}\mathbf{M}$, with $\mathbf{a}$ any constant vector. They all satisfy the vector Helmholtz equation. The two vectors $\mathbf{M}$ and $\mathbf{N}$ are solenoidal, while $\mathbf{L}$ is irrotational. For $\Phi=C_l(kr)Y_{lm}(\hat{\mathbf{n}})$, C_l being the lth cylindrical function and $Y_{lm}(\hat{\mathbf{n}})$ the spherical harmonic, the vectors $\mathbf{L}$, $\mathbf{M}$, and $\mathbf{N}$ are called the lth vector spherical waves.

Wick's Theorem: The theorem transforming the time-ordered product $T(\Psi(x)\Psi(x')\cdots)$ of any number of field operators in the interaction picture into a sum of products of simpler quantities (see, e.g., Fetter and Walecka, 1971). The expectation value of a product of field operators in the noninteracting state may be calculated by use of this theorem.

List of Works Cited

Abramowitz, M., Stegun, I. A. (1965), *Handbook of Mathematical Functions*, New York: Dover.

Abrikosov, A. A., Gor'kov, L. P., Dzyaloshinski, I. E. (1963), *Methods of Quantum Field Theory in Statistical Physics*, Englewood Cliffs, NJ: Prentice-Hall.

Amit, D. J. (1984), *Field Theory, The Renormalization Group, and Critical Phenomena*, 2nd ed., Singapore: World Scientific.

Baker, B. B., Copson, E. T. (1950), *The Mathematical Theory of Huygens' Principle*, Oxford: Clarendon.

Bardeen, J., Cooper, L. N., Schrieffer, J. R. (1957), *Phys. Rev.* **108**, 1175–1204.

Bogoliubov, N. N., Shirkov, D. V. (1959), *Introduction to The Theory of Quantized Fields*, New York: Interscience.

Born, M., Wolf, E. (1975), *Principles of Optics*, 5th ed., New York: Pergamon.

Brebbia, C. A. (1978), *The Boundary Element Method for Engineers*, New York: Pentech.

Brebbia, C. A., Walker, S. (1980), *Boundary Element Techniques in Engineering*, New York: Newns-Butterworths.

Butkovskiy, A. G. (1982), *Green's Functions and Transfer Functions*, New York: Ellis Horwood.

Courant, R., Hilbert, D. (1937), *Methoden der Mathematischen Physik*, Berlin: Springer [reprint: (1989), New York: Interscience], Vols. 1 and 2.

Economou, E. N. (1990), *Green's Functions in Quantum Physics*, 2nd ed., Berlin: Springer.

Fetter, A. L., Hohenberg, P. C. (1969), in: R. D. Parks (Ed.), *Superconductivity*, New York: Dekker, Vol. 2.

Fetter, A. L., Walecka, J. D. (1971), *Quantum Theory of Many-Particle Systems*, New York: McGraw-Hill.

Feynman, R. P. (1972), *Statistical Mechanics*, New York: Benjamin.

Feynman, R. P., Hibbs, A. R. (1965), *Quantum Mechanics and Path Integrals*, New York: McGraw-Hill.

Inoue, M., Ohtaka, K. (1982), *Phys. Rev. B* **26**, 3487–3509.

Itzykson, C., Drouffe, J.-M. (1989), *Statistical Field Theory*, New York: Cambridge University Press, Vols. 1 and 2.

Itzykson, C., Zuber, J.-B. (1980), *Quantum Field Theory*, New York: McGraw-Hill.

Kellog, O. D. (1939), *Foundation of Potential*

Theory, Berlin: Springer [reprint: (1953), New York: Dover].

Kubo, R., Toda, M., Hashitsume, N. (1991), *Statistical Physics II*, 2nd ed., Berlin: Springer.

Landau, L. D., Lifshitz, E. M. (1986), *Theory of Elasticity*, 3rd ed., Oxford: Oxford University Press.

Lukes, T. (1969), in: P. T. Landsberg (Ed.), *Solid State Theory*, New York: Wiley-Interscience.

Mahan, G. D. (1990), *Many-Particle Physics*, 2nd ed., New York: Plenum.

Maradudin, A. A., Michel, T., McGurn, A. R., Méndez, E. R. (1990), *Ann. Phys. (N.Y.)* **203**, 255–307.

Morino, L., Chen, L. T., Suciu, E. O. (1975), *AIAA J.* **13**, 368–374.

Morse, P. M., Feshbach, H. (1953), *Methods of Theoretical Physics*, New York: McGraw-Hill.

Ohtaka, K. (1980), *J. Phys. C* **13**, 667–680.

Parks, R. D. (1969), *Superconductivity*, New York: Dekker, Vols. 1 and 2.

Ramond, P. (1981), *Field Theory*, Reading, MA: Benjamin.

Sakurai, J. J. (1985), *Modern Quantum Mechanics*, Reading, MA: Addison-Wesley.

Schiff, L. I. (1968), *Quantum Mechanics*, 3rd ed., New York: McGraw-Hill.

Smirnov, V. I. (1965), *A Course of Higher Mathematics*, Reading, MA: Addison-Wesley, Vol. IV.

Sommerfeld, A. (1949), *Partial Differential Equations in Physics*, New York: Academic.

Stratton, J. A. (1941), *Electromagnetic Theory*, New York: McGraw–Hill.

Tong, S. Y. (1975), *Prog. Surf. Sci.* **7**, 1–48.

Yablonovitch, E., Gmitter, T. J. (1989), *Phys. Rev. Lett.* **63**, 1950–1953.

Yosida, K. (1965), *Functional Analysis*, Berlin: Springer.

Further Reading

Bateman, H. (1932), *Partial Differential Equations of Mathematical Physics*, New York: Cambridge. An overview of Green's functions concentrating on the mathematical aspects.

Doniach, S., Sondheimer, E. H. (1974), *Green's Functions for Solid State Physicists*, Reading, MA: Addison-Wesley. A readable textbook for the Green's functions used in solid-state physics.

Hedin, L., Lundqvist, S. (1969), in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, New York: Academic, Vol. 23. A monograph for detailed use of Green's functions in many-electron systems.

Itzykson, C., Zuber, J.-B. (1980), *Quantum Field Theory*, New York: McGraw-Hill. A detailed textbook for quantum field theory.

Jackson, J. D. (1967), *Classical Electrodynamics*, New York: Wiley. Some examples of Green's functions as applied to electrodynamics are found.

Jaswon, M. A., Symm, G. T. (1977), *Integral Equation Methods in Potential and Electrostatics*, New York: Academic. A textbook for the foundation of the boundary-value problem.

Morse, P. M., Feshbach, H. (1953), *Methods of Theoretical Physics*, New York: McGraw-Hill. A detailed textbook for Green's functions used in many fields of theoretical physics.

Newton, R. G. (1966), *Scattering Theory of Waves and Particles*, New York: McGraw-Hill. A detailed overview of scattering theory.

GROUP THEORY

M. HAMERMESH, *School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota, U.S.A.*

	Introduction	365
1.	**Elementary Definitions**	366
1.1	Transformation Groups	366
1.1.1	Subgroups	367
1.1.2	Cosets	367
1.1.3	Conjugate Classes	367
1.1.4	Invariant Subgroups	368
1.1.5	Isomorphic Groups	369
1.1.6	Homomorphic Groups	369
1.1.7	Factor Groups	369
1.1.8	Direct Product	369
1.1.9	Finite Groups	369
1.1.10	Infinite Discrete Groups	370
1.2	Continuous Groups	370
1.3	Lie Groups	371
2.	**Linear Representations of Lie Groups**	373
2.1	Group Theory and Quantum Mechanics	373
2.2	Construction of Representations	374
2.2.1	Equivalent Representations	375
2.2.2	Addition of Representations. Reducibility and Irreducibility	375
2.2.3	Invariance of Functions and Operators	376
2.3	General Theorems	377
2.4	Kronecker Product of Representations	377
2.5	Analysis of Representations	377
3.	**Applications**	379
3.1	Atomic Structure	379
3.2	Nuclear Structure	381
3.3	Solid State and Crystal-Field Theory	381
4.	**Summary**	382
	Works Cited	382
	Further Reading	382

INTRODUCTION

Group theory is a mathematical technique for dealing with problems of symmetry. Such problems appear repeatedly in all branches of physics. The use of formal group theory is a very recent development, but notions of symmetry had been used extensively 1000 years ago. The design of ornaments with symmetries, the observation of periodic patterns, and the regular appearance of the Sun and other astronomical objects showed that symmetry was a useful concept. The first modern uses of symmetry were in crystallography. The first clear statement of the importance of symmetry was made by Pierre Curie around 1870. Since then group theory has become the principal tool for dealing with difficult problems in solid-state theory, relativity theory, atomic and nuclear spectroscopy, and the theory of elementary particles. In these problems we assert (or assume) that the laws describing the interactions of particles have some symmetry. In the simpler cases, such as the Coulomb field, the symmetry is easy to see. In nuclear physics the charge symmetry between neutrons and protons required a careful and bold extrapolation of experimental results. In elementary-particle physics we have very little clear understanding of the forces that determine the structure of the fundamental particles and their relations with one another. This has required us to assume some symmetry of the interactions, even though we know almost nothing about the details of the laws governing them. Once a symmetry is assumed we can then make predictions about the families of particles that are related to one another.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. ELEMENTARY DEFINITIONS

1.1 Transformation Groups

We first give the definition of a transformation group, because these are the groups of direct importance for physics. For example, if we are considering the motion of a particle in a central field we realize that a rotation about any axis through the center will take a given orbit into another orbit with the same energy. So we want to study rotations around the center. Two successive rotations around the center again give a rotation to an orbit with the same energy. A reverse rotation would bring the orbit back to its original position. We see that the set of rotations of the three-dimensional space form a set of transformations that is closed under composition and contains all inverses. Now we give the rigorous definition. We have a *linear vector space* on which a set of transformations act. The transformations form a group if

1. The "product" of any two elements a and b of the set is also a transformation in the set: $ab=c$. By product we mean the transformation c that results from the sequence of transformations b followed by a.
2. The product is associative, i.e., $((ab)c)=(a(bc))$. The product of a whole sequence of transformations gives the same final result if we split the sequence into any clumps that preserve the order.
3. The set of transformations includes the identity transformation e, that leaves all the coordinates unchanged.
4. If the set includes the transformation a, it must also contain the inverse transformation b, such that $ab=ba=e$. We usually write the inverse element of a as a^{-1} so that $aa^{-1}=a^{-1}a=e$.

In the example given above we note that in general, if we reverse the order of rotations a and b around different axes we get different results. In general the multiplication is not commutative, i.e., $ab\neq ba$. If the product of any two elements of the set is commutative ($ab=ba$ for all a,b in the set), we say that the group is *Abelian* (the group is a commutative group). In general the transformations of a group will not all commute with one another. Then we say that the group is non-Abelian. If we take the product of an element a of the group with itself, i.e., we form aa, we write the product as a^2, the "square" of a. Similarly we write successive products of a with itself as a^3, a^4, etc.—powers of a. If we perform repeated transformations using the inverse a^{-1} of a, we get the negative powers a^{-2}, a^{-3}, etc. The total number of distinct elements in the group is called the *order g* of the group. If the order of the group is finite we say that the group is a *finite group.* An infinite discrete group is one in which the distinct elements can be labeled by the integers. Often the elements of the group can only be labeled by one or more continuous parameters. Then we say that the group is a *continuous group.* If we talk about a group of objects with some product that satisfies our definitions we have an *abstract group.* We give some examples of groups.

The numbers 0,1,2,3 form a group if the product is addition modulo 4: $2+1=1+2=3$, $2+3=3+2\cong 5-4=1$, $2+2\cong 4-4=0$, etc. The inverse of 2 is 2, the inverse of 1 is 3. This is the group Z_4, an Abelian group of order 4. Similarly the numbers $0,1,\ldots,n-1$ give the Abelian group Z_n of order n. If we take all the integers $\ldots,-2,-1,0,1,2,\ldots$ we get the infinite discrete Abelian group Z, in which the identity element is 0 and the inverse of s is $-s$.

Cyclic group. This is a group that consists of positive powers of some element a, for which there is a finite positive n such that $a^n=e$. Thus a^{n-1} is the inverse of a. So the group consists of $a,a^2,a^3,\ldots,a^n=e$. One example of a cyclic group is the set of rotations about some axis through angles $\theta=2\pi m/n$ where n is a fixed integer and $m=1,2,\ldots,n$. This is the cyclic crystal group C_n. Here the product of rotations with $m=r$ and $m=s$ is the rotation through the angle $(2\pi/n)(r+s)$. This group C_n has the same abstract structure as Z_n. Clearly cyclic groups are Abelian.

An example of an infinite Abelian group is the group of translations along the x axis given by $x'=x+n\delta$, where δ is fixed and n is an integer, positive or negative or zero (which is the identity translation). This group has the same abstract structure as the group Z.

Permutation groups. Another important class of groups comprises the permutation groups on n letters. In these groups each letter is replaced by some letter, while no two are replaced by the same letter. If each letter is replaced by itself, we get the identity permutation e. We first illustrate for the case of $n=6$. A permutation of the numbers 1 to 6 replaces each number by one of the six numbers, and no two numbers are replaced by the same number. A simple notation for a permutation is

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ i_1 & i_2 & i_3 & i_4 & i_5 & i_6 \end{pmatrix},$$

where each number in the upper line is replaced by the number appearing below it. For example, the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 2 & 4 & 5 & 3 & 1 \end{pmatrix}$$

replaces 1 by 6 ($1\rightarrow6$), $2\rightarrow2$, $3\rightarrow4$, $4\rightarrow5$, $5\rightarrow3$, $6\rightarrow1$. Note that $2\rightarrow2$, so 2 is unchanged by this permutation. There are $6!=720$ different permutations. (If all the symbols are left unchanged we get the identity permutation with $1\rightarrow1$, $2\rightarrow2$, etc.). This same permutation can be written in *cyclic notation:* we start with any symbol, say 1. We record 1, and note that $1\rightarrow6$, so we write $1\rightarrow6$. $6\rightarrow1$, and we have a cycle (closed loop), so we put 1 and 6 in parentheses: (1 6). The number 2 is unchanged so we write (2). Next we start with 3 and find $3\rightarrow4\rightarrow5\rightarrow3$, giving the cycle (3 4 5) [which is the same as (4 5 3) or (5 3 4)]. Thus the original permutation is recorded as (16) (2) (3 4 5). The symbols inside a parenthesis are distinct. This particular permutation has the *cycle structure* [3,2,1]: it contains a cycle of three symbols, a cycle of two symbols, and a cycle of one symbol—a 3-cycle, a 2-cycle, and a 1-cycle. Often the permutation is written omitting its 1-cycles, so this would be (16) (345). Other examples are (134652) which consists of one 6-cycle, or (142)(365), with two 3-cycles (notation $[3^2]$), or (15)(26)(34) which contains three 2-cycles (notation $[2^3]$), or (135), where the three 1-cycles (unchanged symbols) are omitted, so its cycle structure is $[3\ 1^3]$. To find the inverse of a permutation expressed in cycle notation, we read the numbers from right to left, so the inverse of (142)(365) is (241)(563). The group that contains all the permutations on six symbols is called the symmetric group on six symbols S_6. We can form permutations on n letters and form the symmetric group S_n whose order is $n!$ This group S_n is very important in spectroscopy and especially when we consider identical (indistinguishable) particles, where the physical situation is unchanged when we make any permutation of the identical particles.

1.1.1 Subgroups. Suppose that we have a subset H of elements in the group G ($H\subset G$). If the elements of H form a group under the same product law as in G, we say that H is a *subgroup* of G. For example, the group S_6 contains all the permutations on five letters (e.g., we drop all permutations containing the number 6, or those that omit any single one of the six letters). Thus S_6 contains six subgroups that have the same structure as S_5. S_6 also contains many differently appearing subgroups that have the same structure as S_4. (We omit permutations that contain some two of the numbers 1–6). Such subgroups are isomorphic to one another. Clearly the group G is a subgroup of G. Also the group element e, the identity of G, is necessarily a subgroup. These two subgroups of G are said to be improper, while all other subgroups in G are called proper subgroups of G.

Suppose that H is a subgroup in G. If H does not exhaust G, take any element a not contained in H and form the set aH, which contains all elements of the form ah where h runs through H. The set aH contains no elements of H, since, if $ah=h'$ where h' is in H, applying h^{-1} on the right we would get $a=h'h^{-1}$, which says that a is a product of elements in H and is therefore in H. If the elements of G are not all included in H and aH, we choose one of the residual elements k of G and form kH. We continue this process until we have obtained all the elements of G. So G consists of the subsets of elements $H, aH, kH,\ldots,sH$, which have no elements in common and each contain m elements of G. We see then that the number m of elements in H (the order of H) must divide the order of G. This is called *Lagrange's theorem.* Note that the subgroups S_5 in S_6 have order $5!=120$, which divides the order $6!=720$ of S_6.

1.1.2 Cosets. The individual pieces in this decomposition of G are called the left cosets of H in G. We write this symbolically as $G=H+a_2H+a_3H+\cdots+a_rH$. If instead we formed the cosets by multiplying H from the right, giving disjoint sets Ha, we would get the right coset decomposition of G with respect to the subgroup H.

1.1.3 Conjugate Classes. If a and b are elements of the group G, the elements $bab^{-1}=c$, where b is any element of G, are said to be *conjugate* to a in G. We denote this by writing

$c \sim a$. If we let b run through all the elements of G, we get the set of *transforms* of a, the *conjugate class* of a in G. The determination of transforms gives a decomposition of G into conjugate classes that have no elements in common, i.e., the conjugation operation is an equivalence relation:

1. $a \sim a$.
2. If $a \sim b$ and $b \sim c$, then $a \sim c$.
3. If $a \sim b$, then $b \sim a$.

We prove 2. $a \sim b$ means that $a = kbk^{-1}$, where k is in G. Similarly $b \sim c$ means that $b = k'ck'^{-1}$, where k' is in G. So $a = kbk^{-1} = k(k'ck'^{-1})k^{-1} = (kk')c(k'^{-1}k^{-1}) = (kk') \times c(kk')^{-1}$, so $a \sim c$.

For geometrical transformations the conjugate has a simple interpretation. Suppose that a is a rotation through angle θ about a direction along the unit vector $\mathbf{n}$: $a = (\mathbf{n}, \theta)$. Let b be a rotation that turns the unit vector $\mathbf{n}$ into the direction of $\mathbf{n}'$. Then, since b turns $\mathbf{n}$ into $\mathbf{n}'$, in the conjugate bab^{-1}, b^{-1} first turns $\mathbf{n}'$ into $\mathbf{n}$, then a rotates about $\mathbf{n}$ through θ, then b brings $\mathbf{n}$ back to $\mathbf{n}'$, so the net result is a rotation through the angle θ about the unit vector $\mathbf{n}'$. Thus the conjugate of a rotation is always a rotation through the same angle θ about some other axis. Similarly, if a were a translation through a vector $\mathbf{s}$, the conjugate bab^{-1} would be a translation through the vector $\mathbf{s}'$, into which b takes $\mathbf{s}$.

For the permutation groups, the conjugate is also obtained simply. Suppose we want to find the conjugate c of the permutation a that results from taking $c = bab^{-1}$, where $a = (135)(24)$, $b = (325)$. Then $b^{-1} = (523)$, and we get $c = (325)(135)(24)(523) = (54)(123)$. We note that the conjugate of a has the same cycle structure as a; we get it by letting the permutation b act on the cycle symbol for a: apply (325) to a; it replaces 3 by 2, 2 by 5, and 5 by 3, so a is changed to $(123)(54)$, where 4 and 1 are untouched, since b does not contain 1 or 4. Another example would be the transform of $a = (142)(365)$ by $b = (16)(345)(2)$. Applying the permutation b to the cycle symbol for a replaces $1 \rightarrow 6 \rightarrow 1$, $3 \rightarrow 4 \rightarrow 5 \rightarrow 3$, giving $bab^{-1} = (652)(413)$.

We emphasize that the transforming element b must be in the group G. If the group is Abelian, then all its elements commute, and the conjugate class of an element a is just a itself. If an element a commutes with all the elements of the group G, its class consists of a alone. This is always the case for the identity element e for any group—the identity is a class containing the single element e. Suppose we consider the permutation group S_3 on three letters 1, 2, 3. There are six elements in this group: $e, (12), (23), (13); (123), (132)$. Only permutations with the same structure can be in the same conjugate class, so we have three classes: e; $(12), (23), (13)$; $(123), (132)$; with one, three, and two members, and cycle structures $[1^3]$, $[2\ 1]$, and $[3]$, respectively.

1.1.4 Invariant Subgroups. If H is a subgroup of G, we consider the set aHa^{-1}, containing all elements aha^{-1}, where h runs through the members of H. The elements of this set are all distinct, and the product of any two, $ah_1a^{-1}ah_2a^{-1} = ah_1h_2a^{-1}$, is also in aHa^{-1}, since h_1h_2 is in H. Clearly the set aHa^{-1} is a subgroup that looks like H in form. It is *a conjugate subgroup of H*. If all the conjugate subgroups of H in G are the same as H, we say that H is an invariant subgroup in G. But if $aHa^{-1} = H$ for all elements a, then multiplying on the right by a, we get $aH = Ha$, i.e., the sets of right and left cosets of H in G are identical. We also note that the invariant subgroup contains complete classes. Thus, in the example of the permutations on three letters, the group S_3 contains the proper subgroup e, $(123), (132)$. Since it consists of complete classes, it is an invariant subgroup. For the permutation group on four letters S_4, the group contains $4! = 24$ elements. The different classes are e; type (12), type (123), type (1234), and type $(12)(34)$, with numbers of elements 1,6,8,6,3, respectively. The classes correspond to the possible cycle structures given by the partitions of 4 into sums of positive integers. Thus e has four 1-cycles, partition $[1^4]$; (12) has one 2-cycle and two 1-cycles (partition $[21^2]$), (123) has one 3-cycle and one 1-cycle, (partition $[31]$), (1234) has one 4-cycle (partition $[4]$), and $(12)(34)$ has two 2-cycles (partition $[2^2]$).

As another example, we consider the group of all rotations and translations in a space with dimension m. We denote the elements by $(R \mid a)$, where this means that we first perform the rotation R and then the translation a. The translations a form a subgroup (Abelian), and we showed earlier that the transform of any a by a rotation R is again a translation. Thus the subgroup of the translations is invariant in the whole group.

1.1.5 Isomorphic Groups. Two groups G and G' are said to be *isomorphic* if their elements can be put into a one-to-one correspondence so that if $a \leftrightarrow a'$ and $b \leftrightarrow b'$ then $ab \leftrightarrow a'b'$. For example, consider the group consisting of the identity e and the inversion of coordinates i. Here $i^2=e$, so the group contains two elements. Now look at the group consisting of the numbers 1 and -1 with the product being ordinary multiplication, so that $(-1)(-1)=1$. We see that these groups can be put into one-to-one correspondence $e \leftrightarrow 1$, $i \leftrightarrow -1$, with $ii=e$ corresponding to $(-1)(-1)=1$. There can be many groups with the same abstract structure. They are isomorphic to one another.

1.1.6 Homomorphic Groups. Instead of a one-to-one correspondence, we may have two groups with a correspondence of products, but where several elements of the group G correspond to a single element of G'. We say that there is a *homomorphism* of the group G onto G'. For example, in the cyclic group of order 4, generated by an element a with $a^4=e$, the elements a^2 and $a^4=e$ are a subgroup H in G. If we take the odd powers a and a^3 and map them onto the number -1, and map the subgroup H onto 1, we have a two-to-one correspondence between the cyclic group and the two-element group.

1.1.7 Factor Groups. If the subgroup H is an invariant subgroup in G, we can write G as a "sum" of cosets of H in G: $G = H + a_2H + a_3H + \cdots + a_rH$, where $a_iH=Ha_i$. But then the product of cosets $(a_iH)(a_jH) = a_i(Ha_j)H = a_i(a_jH)H = (a_ia_j)H$, so the cosets themselves form a group, for which the identity is the coset H itself. This new group is called the factor group F of G by the (invariant) subgroup H. There is a homomorphism of the group G onto the subgroup H where all the elements in a coset are mapped on the same element of the factor group $F=G/H$.

1.1.8 Direct Product. If we have two groups, G and H, and the elements of one commute with all the elements of the other, we can form a new group, their direct product, by taking pairs (g,h) of elements g from G and h from H. The product of two elements (g_1,h_1) and (g_2,h_2) in this new group is the ordered pair (g_1g_2,h_1h_2). This group is called the direct product $G \otimes H$ of the two groups G and H. If we have a number of groups $G,H,K,\ldots$, and the elements of any one commute with all the elements of the others, we can form the direct product $G \otimes H \otimes K \otimes \ldots$, whose elements are obtained by taking one element from each group: $(g,h,k,\ldots)$. Clearly all the elements $(g,e_i,e_j,\ldots)$, where the $e_i,e_j,\ldots$ are the identity elements for the other factors, form a group isomorphic to G. This group can be identified with G and is an invariant subgroup of the direct product. The direct product is thus a product of the groups, $G,H,K\ldots$, and each of them is an invariant subgroup of the direct product.

For the crystal group that we described earlier, with elements $(R|a)$, the translations are an invariant subgroup, since RaR^{-1} is again a translation, while the subgroup of rotations is not invariant. Instead the rotations act on the translations and transform the translation vectors. Such a group is called the semidirect product of the groups R and A, where R contains all the pure rotations and A all the pure translations, and is written as $A \ltimes R$, where the invariant subgroup is the first symbol.

1.1.9 Finite Groups. For a finite group one can describe the group structure by recording the Cayley table for the group. We write the elements of the group as a matrix with each column headed by an element of the group, and similarly, each row. We then get a square matrix of products by placing the product g_ig_j at the i,j position.

The group formed by taking all products of r and s, where $r^3=s^2=(rs)^2=e$, is a group with six elements. Its Cayley table is shown in Table 1. Note that each row contains all six elements of the group, but they are rearranged. In other words, applying an element of the group from the left permutes the elements. We thus have a group of permutations that is isomorphic to the group in the table.

Table 1. A Cayley Table.

e	r	r^2	s	sr	sr^2
r	r^2	e	sr^2	s	sr
r^2	e	r	sr	sr^2	s
s	sr	sr^2	e	r	r^2
sr	sr^2	s	r^2	e	r
sr^2	s	sr	r	r^2	e

This is a group of permutations of six elements, and is a subgroup of the symmetric group S_6. In the same way, every finite group of order n is isomorphic to a subgroup of the symmetry group S_n. This result is called *Cayley's theorem.* An important collection of finite groups consists of the so-called point groups: rotations through submultiples of 2π that keep the origin fixed and reflections in planes passing through the origin (or inversion in the origin). These finite groups describe the point symmetry of crystals.

1.1.10 Infinite Discrete Groups. In three dimensions the translations through a displacement $\mathbf{r}=n_1\mathbf{a}_1+n_2\mathbf{a}_2+n_3\mathbf{a}_3$, where the n_i are integers and the vectors $\mathbf{a}_i$ are fixed, form an infinite discrete group. This is the translation group of a crystal lattice. Similar groups can be constructed for any space dimension. We can also have crystal space groups, that contain elements $(R|a)$ that preserve the crystal lattice. If the translations appearing in $(R|a)$ are always lattice vectors the crystal group is said to be *symmorphic.* There are some crystals in which there are additional symmetries which involve elements $(R|a)$ where the translation is some rational fraction of a lattice vector along the direction of the rotation axis, giving a screw motion. Or we may have glide motions where we translate parallel to some crystal plane through a rational fraction of a lattice vector and then reflect in the crystal plane. Crystals that have such symmetry elements are said to be *nonsymmorphic.* [For further information about space groups see Burns 1977, Chap. 11.] More detailed treatments of Sec. 1.1 can be found in Burns (1977), Chaps. 1, 2, and 13; Elliott and Dauber, (1979), Chaps. 1 and 2; and Hamermesh (1989), Chaps. 1 and 2.

1.2 Continuous Groups

When we consider continuous groups we combine the algebraic concept of "group" with the topological concept of "nearness." The law of combination of elements a, b of G now requires the product ab to depend continuously on its factors, and the inverse a^{-1} must depend continuously on a. We shall not discuss general topological groups but will assume that there is a metric (a measure of distance between group elements) on the group. So the group itself is a metric space on which the product and inverse are defined and continuous. We can look upon the group as a space in which the points are the elements of the group. We shall deal only with groups of transformations, so the group elements form a space whose points are the transformations of the group. Multiplying the element b on the left by some transformation a of the group moves the point b in the group space to the point ab. Thus the group elements can be regarded as operators that act on the group space itself and rearrange the points. The changed points also fill the space since any element of the group space c can be reached from b by applying the element cb^{-1} to it. Thus the group space looks the same everywhere. It is a homogeneous space, and we can obtain all our information by, e.g., working in the neighborhood of the identity e.

A group is said to be a mixed continuous group if the elements depend on discrete labels as well as continuous ones. For example, the translation group on the line G is the set of transformations $x'=x+a$, $(-\infty\leqslant a\leqslant\infty)$. If we adjoin the single transformation $x'=-x$ we also get the transformations $x'=-x+a$. The transformations $x'=x+a$ are in one piece and can be reached continuously from the identity. Thus this piece forms a group. The second piece, containing the transformations $x'=-x+a$, cannot be reached continuously from the identity, and is not a group. This piece is the coset of the first piece, obtained by multiplying elements of the first piece by the transformation $x'=-x$.

A *connected group* is a continuous group in which we can join any two points of the group by a continuous arc in the group space. Our last example showed a group consisting of two pieces (components). The piece that contains the identity is called the component of the identity. If a group is connected it consists of a single component. One basic theorem is the following: In a connected group G, every neighborhood of the identity generates the group. By taking products of elements near the identity we keep expanding the set of products and fill the whole connected piece containing the identity. For example, in the group of translations $x'=x+a$, if we take some neighborhood of the identity $a=0$, say $-\epsilon\leqslant a\leqslant\epsilon$, and keep applying these translations, we get the whole line.

1.3 Lie Groups

A Lie group is a group in which the neighborhood of the identity looks like (is homeomorphic to) a piece of an r-dimensional Euclidean space (an r-parameter Lie group). In other words, in the neighborhood of the identity the group elements can be parametrized using r parameters (coordinates on the group manifold). Thus each element can be labeled by r parameters $a_1,a_2,...,a_r$. Since we want continuity, this parametrization may not be possible over the whole group space, but we can use several maps that overlap to cover the whole space without singularities. (For example, to cover a sphere with continuous maps requires more than one map.) An element of the group is $R(a)=R(a_1,...,a_r)$ where the parameters $a_1,...,a_r$ are essential, i.e., all r parameters are needed to describe the space. The product of two elements with parameters a and b, respectively, $R(a)R(b)$, is the element $R(c)$ with parameters $c_1,...,c_r$ that are functions of $a_1,...,a_r$ and $b_1,...,b_2$, i.e.,

$$c_i=\varphi_i(a_1,...,a_r;b_1,...,b_r), \text{ or, symbolically,}$$

$$c=\varphi(a;b). \tag{1}$$

The simplest assumption is that the functions φ are analytic functions of the a's and b's. This was the original requirement for a Lie group. But this requirement is too strong. In fact, every parametric group is a Lie group. The group structure combined with continuity implies the analyticity of φ (Hilbert's "Fifth Problem").

For Lie groups with a finite number of parameters, the neighborhood of any point is bounded and contains all its limit points (the group is *locally compact*). If the parameter space is unbounded this may not be true as we move out to large distances in the parameter space. For most of the groups used in physics the group space is compact, but we shall see some cases of noncompact groups.

Consider transformations in n-space:

$$x_i'=f_i(x_1,...,x_n), \quad i=1,...,n, \tag{2}$$

or, symbolically, $x'=f(x)$. Suppose that we have a set of f's labeled by r parameters:

$$x_i'=f_i(x_1,...,x_n;a_1,...,a_r), \tag{3}$$

forming a Lie group. The transformation a is $x\to x'=f(x;a)$, and b is $x'\to x''=f(x';b)$. If we first apply a and then b, we get

$$\begin{aligned} x''&=f(x';b)=f(f(x;a);b)\\ &=f(x;\varphi(a,b))=f(x;c), \end{aligned} \tag{4}$$

where $c=\varphi(a,b)$ and φ is an analytic function of the parameters a and b. We give examples of Lie groups that comprise most of the groups of interest to physics.

1. The dilation group. In one dimension this is the group of transformations $x'=ax$, $a\neq 0$. The identity has $a=1$. All distances on the line are multiplied by the number a. The transformation with $a=-1$ reflects the line about the origin. The parameter a for the inverse is $\bar{a}=1/a$, and $c=\varphi(a,b)=ab$. The group space is the real line with the origin cut out, $\mathbb{R}\setminus\{0\}$. It consists of the two separate pieces $x\gtrless 0$. The group space for the dilations in two dimensions, $x'=ax$, $y'=ay$, is the real plane with the origin removed (the punctured plane $\mathbb{R}^2\setminus\{0\}$). In this case the group space remains a single connected piece. It is convenient to assign the parameter 0 to the identity. We write the group as $x'=(1+\alpha)x$, with $\alpha\neq -1$.
   The group $x'=ax, y'=a^2y$, with $a\neq 0$, has only one parameter, but now the transformations are in a space of dimension 2. We note that the number of parameters is not determined by the dimension of the space on which the transformations act.
2. GL(n), the general linear group in n dimensions, is the set of nonsingular linear transformations in n-dimensional space. For real entries we write GL(n,R), for complex entries GL(n,C): $x'=Ax$, or $x_i'=a_{ij}x_j$ ($i=1,...,n$), where we sum over the repeated index, and det $A\neq 0$ (the nonsingular transformations). The number of real parameters is n^2 for GL(n,R), and $2n^2$ for GL(n,C). $\bar{A}=A^{-1}$, $\varphi(A,B)=BA$. The elements a_{ij} vary over an infinite range ($-\infty\leqslant a_{ij}\leqslant\infty$), so the group is not compact. The elements of GL(n,C) are the subset of the $n\times n$ complex matrices with determinant $\neq 0$.
3. SL(n) is the special linear group (unimodular group) with det$=1$. SL(n,C) and SL(n,R) are subgroups of GL(n,C). Thus SL(2) is the collection of 2×2 matrices
   $$\begin{pmatrix} a & b\\ c & d\end{pmatrix} \text{ with } ad-bc=1.$$
4. $O(n)$ is the orthogonal group, a subgroup of GL(n), where we impose the requirement that the transformations leave scalar

products of any two vectors x,y unchanged; i.e., $(Ox,Oy)=(x,y)$, where the scalar product is

$$(x,y)=\sum_i x_iy_i.$$

Then $(x,y)=(Ox,Oy)=(x,\tilde{O}Oy)$, where $\tilde{O}$ is the transpose of the matrix O, so that

$$\tilde{O}O=1,\text{the unit matrix, or}\tilde{O}_{ij}O_{jk}=\delta_{ik}$$
$$=O_{ji}O_{jk}. \tag{5}$$

Taking determinants, $(\det O)^2=1$, or $\det O=\pm 1$. The column vectors of the orthogonal matrix are orthogonal unit vectors. Similarly $O\tilde{O}=1$, so the rows are orthogonal unit vectors. For $O(n,R)$, there are n conditions from the norm and $n(n-1)/2$ from the orthogonality, so the number of parameters is $r=n(n-1)/2$. The subgroup with positive determinant $+1$ is called $O^+(n,R)$ or $\mathrm{SO}(n,R)$, the special orthogonal group (or proper orthogonal group).

5. $U(n)$ is the unitary group that acts in a complex unitary space, where we have an inner product $(x,y)=\Sigma_i x_i^* y_i$, and the asterisk denotes the complex conjugate, so that $(x,y)=(Ux,Uy)=(x,U^\dagger Uy)$, where $U^\dagger=\tilde{U}^*$ is the adjoint of U, so $U^\dagger U=1=UU^\dagger$, $\Sigma_i U_{ij}^* U_{ik}=\delta_{jk}$, and $|\det U|^2=1$, so $\det U=\exp(i\epsilon)$. Note that for $j=k$ there are n conditions $\Sigma_i|U_{ij}|^2=1$, while for $j\neq k$ there are $2n(n-1)/2$ orthogonality conditions (real and imaginary parts both equal to zero). Thus the number of parameters is $r=n^2$. The sum of the absolute squares of all the matrix entries is n, so the group space is bounded and $U(n)$ is a compact group.
6. $SU(n)$ is the unitary unimodular group (special unitary group), a subgroup of $U(n)$ having $\det U=1$, so $r=n^2-1$.
7. The *Euclidean group $E(n)$* in real space, $x'=Ax+a$, where A is an orthogonal matrix and $\mathbf{a}$ is a translation vector, preserves the Euclidean distance, so $\tilde{A}A=1$, and the number of parameters is $r=n(n+1)/2$. For $n=3$, $r=6$. This group $E(3)$ can be regarded as a group of block matrices (A,a), where A is a 3×3 orthogonal matrix, O is the (1×3) null matrix,

$$\begin{pmatrix} & a_1 \\ A & a_2 \\ & a_3 \\ O & 1 \end{pmatrix},$$

and (A,a) acts on the column vectors

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}.$$

This group describes the rigid-body displacements in 3-space.

In obtaining $O(n,R)$ we required the invariance of the scalar product (x,y). If we had required the invariance of any positive definite symmetric bilinear form $g_{ij}x_iy_j=(x,gy)$, with $g=\tilde{g}$, we could bring it back to the unit form by a change of basis. Thus the group $O(n,R)$ is obtained for all of these. If we extend this argument to include all nonsingular symmetric bilinear forms, a change of basis will bring them (Sylvester's law of inertia) to diagonal form with p 1's and q (-1)'s, where $p+q=n$. We get the pseudo-orthogonal groups $O(p,q)$, defined over the real numbers. If we define a matrix

$$S_{pq}=\begin{pmatrix} \mathbf{1}_p & 0 \\ 0 & -\mathbf{1}_q \end{pmatrix},$$

where the $\mathbf{1}_p$ and $\mathbf{1}_q$ are the unit matrices for dimension p and q, these groups are defined by the requirement that $(x,sy)=(Ox,sOy)=(x,\tilde{O}sOy)$ for arbitrary x and y, so that $\tilde{O}sO=s$. Again we have $(\det O)^2=1$, so by choosing those transformations with $\det=1$, we can define the special groups $\mathrm{SO}(n-q,q)$. An important example of this kind of group is the *Lorentz group*, where we require the invariance of $x^2+y^2+z^2-\tau^2$ (where $\tau=ct$) under the group.

What about skew-symmetric bilinear forms? Given a nondegenerate skew-symmetric bilinear form, with

$$\{x,y\}=g_{ij}x_iy_j \quad\text{and}\quad g_{ij}=-g_{ji}, \tag{6}$$

$g=-\tilde{g}$, $\det g=(-1)^m\det g$, so the form will be nondegenerate only for even dimension m. If we want the transformations to preserve $\{x,y\}$, we must have $\tilde{O}gO=g$, so $(\det O)^2=1$. By a change of basis we can bring g to the canonical form

$$g=\begin{pmatrix} 0 & \mathbf{1}_n \\ -\mathbf{1}_n & 0 \end{pmatrix},$$

where O is the $n\times n$ null matrix and $\mathbf{1}_n$ is the $n\times n$ unit matrix. These groups are the symplectic groups Sp(2*n*). For further details and examples, see Hamermesh (1989), p. 283ff.

2. LINEAR REPRESENTATIONS OF LIE GROUPS

In Sec. 1 we gave an introduction to group theory and a description of some of the important transformation groups that have application to physics. Now we want to show how groups can be used in the solution of physical problems.

2.1 Group Theory and Quantum Mechanics

In quantum mechanics the states of a system of n particles are described by wave functions (state vectors) Ψ that are functions of the time t, the coordinate vectors $\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_n$ of the n-particle system, and any other internal coordinates of the particles. The changes of the system with time (the equations of motion) are governed by the Hamiltonian operator $H=T+V$, where T is the kinetic energy and V is the potential energy of the system, which includes the interaction of the particles with one another and with any external fields. V is a function of the coordinates, while T involves both the coordinates and derivatives with respect to the coordinates. The time behavior of the system will be given by the Schrödinger equation:

$$H\Psi=-\frac{\hbar}{i}\frac{\delta\Psi}{\partial t}, \tag{7}$$

where $\hbar=(1/2\pi)\times$(the Planck constant h). For most problems we look for the stationary states, i.e., states whose time dependence is given by an exponential factor $\exp(-iEt/\hbar)$:

$$\Psi(r_1,\ldots,r_n;t)=\psi(r_1,\ldots,r_n)\exp(-iEt/\hbar). \tag{8a}$$

E is the energy of this stationary state (energy eigenstate) and $H\psi=E\psi$. A particular problem will have many different solutions ψ_i with corresponding energies E_i. The set of all the eigenvalues E_i gives the spectrum of energy levels of the system. If there are several independent state vectors for a given energy level, we say that the state is degenerate. The theory is linear: we can superpose any states ψ_i to get a new state $\psi=\alpha\psi_r+\beta\psi_s$, where α and β are arbitrary complex numbers. If we assume that the state at time $t=0$ is given by

$$\psi=\sum_{i=1}^{m} c_i\psi_i,$$

then at time t, it will have developed into the function

$$\psi(t)=\sum_{i=1}^{m} c_i\psi_i(r_1,\ldots,r_n)\exp(-iE_it/\hbar). \tag{8b}$$

The quantities that are measured for a state ψ (the observables, such as position of the nth particle, the dipole moment of the system, etc.) are operators that act on the state vector, and in general change it to a new state. If the operator S leaves the state ψ unchanged except for a complex factor, so that $S\psi=s\psi$, we say that the state is an eigenstate of the operator with eigenvalue s. If we measure S on a state $\varphi=\Sigma_i b_i\varphi_i$, where the φ_i are normalized eigenstates of the operator S, the result will be one of the s_i, with a probability $|b_i|^2$, and the average value over many measurements (the expectation value of S in this state) will be

$$\langle S\rangle=|(\varphi,S\varphi)|^2=\sum_i |b_i|^2 s_i. \tag{9}$$

In addition to observables we can introduce other operators that act on the states of the system. For example, if the coordinates are changed by some transformation, there will be a corresponding change in the state vector $\psi(r)$. The new state vector ψ' will be obtained by acting on the state vector ψ with an operator related to the coordinate transformation. Some coordinate transformations may leave the physical situation unchanged. For example, for a particle in a central field, a rotation R about any axis through the center has no apparent effect on the physics, since the field depends only on the radius vector to the particle. The operator O_R corresponding to the rotation R acts on the state vector ψ to give a new state $\psi'=O_R\psi$ that looks the same in the rotated coordinates $r'=Rr$ as did the original ψ in terms of the coordinates r:

$$O_R\psi(r')=\psi(r),$$

$$O_R\psi(Rr)=\psi(r),$$

or

$$O_R\psi(r)=\psi(R^{-1}r). \tag{10a}$$

If the state ψ is an eigenstate of the Hamiltonian H with eigenvalue E, $H\psi=E\psi$, $O_R\psi$ will also be an eigenstate of H with the same energy E, i.e., $HO_R\psi=EO_R\psi$, so the states ψ and $O_R\psi$ will be degenerate (or differ by a phase). Applying the operator O_R^{-1} on the left, we get

$$O_R^{-1}HO_R\psi=E\psi=H\psi. \tag{10b}$$

Since this equation holds for any eigenstate, it holds for all states ψ, so the operators on both sides coincide:

$$O_R^{-1}HO_R=H; \quad HO_R=O_RH. \tag{10c}$$

Thus the operator O_R commutes with the Hamiltonian H. It is a symmetry operator for this Hamiltonian. If we choose some other rotation S we get a corresponding symmetry operator O_S that acts on the state vectors. The product of the operators O_SO_R is the operator O_{SR} that corresponds to the result of successive transformations by R and then by S. In general, if we have a collection of symmetry operators their products and inverses will also be symmetry operators. The identity operator is a symmetry operator and the operators satisfy the associative law. Thus the operators obtained by using Eq. (10a) for all the symmetry transformations will form a group of operators that is isomorphic to the group of symmetry transformations. The group of operators acting on some space of functions provides us with a representation of the symmetry group.

2.2 Construction of Representations

As a simple example of the construction of representations, we again consider the problem of a particle in a central field. The group of symmetry operators O_R will take an eigenfunction ψ into an eigenfunction $O_R\psi$ with the same energy E. If we apply all the rotation operators to the same ψ we will not always get a linearly independent $O_R\psi$ as the result. For example, if our $\psi(x,y,z)$ depends only on the radius r, $O_R\psi$ will leave ψ unchanged, since this state vector is spherically symmetric. If $\psi(x,y,z)=x$, the rotation operators will give only x,y,z or linear combinations of them. So this will be a representation in a three-dimensional carrier space, and we say that this is a three-dimensional representation of the rotation group. Calling these basis functions $\psi_i^{(1)}(i=1,2,3)$, $O_R\psi_i^{(1)}$ will be a linear combination of $\psi_i^{(1)}$:

$$O_R\psi_i^{(1)}=\sum_{j=1}^{3} D_{ji}^{(1)}(R)\psi_j^{(1)}, \quad (i=1,2,3), \tag{11}$$

where the superscript (1) labels the representation $D^{(1)}$, and $D_{ij}^{(1)}(R)$ is the matrix representative of the rotation in this representation. If we perform a second rotation S, we get

$$\begin{aligned} O_s(O_r\psi_i^{(1)}) &= \sum_j D_{ji}^{(1)}(R)O_s\psi_j^{(1)} \\ &= \sum_{k,j} D_{ji}^{(1)}(R)D_{kj}^{(1)}(R)\psi_k^{(1)} \\ &= \sum_k \left[\sum_j D_{kj}^{(1)}(R)D_{ji}^{(1)}(R)\right]\psi_k^{(1)} \\ &= \sum_k D_{ki}^{(1)}(SR)\psi_k^{(1)}. \end{aligned} \tag{12}$$

The matrices $D^{(1)}$ thus give a three-dimensional representation of the rotation group. If we take any linear subspace of this three-dimensional space, applying the rotation operators will automatically generate the rest of the three-dimensional space. There is no proper subspace that is invariant, i.e., is reproduced by the rotation operators. We say that the representation $D^{(1)}$ is an *irreducible* three-dimensional representation of the rotation group.

If we start with quadratic expressions in x,y,z: xy, yz, zx, x^2, y^2, z^2, we will obtain a six-dimensional carrier space on which the rotation operators will act, giving us a six-dimensional representation D. But this representation is *reducible:* we can find a proper subspace which is invariant. It is the one-dimensional subspace of the multiples of the function $\psi_{11}^{(0)}=x^2+y^2+z^2$, which is unchanged by rotations (orthogonal transformations). The remaining five-dimensional space with basis xy, yz, zx, x^2-z^2,y^2-z^2, contains no invariant subspace, so the matrices obtained by applying the rotation operators O_R to this subspace will give an irreducible five-dimensional representation of the rotation group. Thus by changing the basis we can decompose the representation into two independent irreducible representations.

In general, for any transformation group G with elements R, starting with some functions ψ, we can apply the corresponding operators O_R to the ψ's, find an independent basis set, and obtain an equation just like Eq. (11):

$$O_R\psi_i^{(\mu)}=\sum_{j=1}^{d_\mu} D_{ji}^{(\mu)}(R)\psi_j^{(\mu)},\quad (i,j=1,\ldots,d_\mu), \tag{13}$$

where μ labels the representation $D^{(\mu)}$, with dimension d_μ.

We mention again the two important steps:

1. We start with some functions and find the space of functions generated by the operator group.
2. We make changes of basis and look for invariant subspaces. If there are no invariant subspaces, the representation is irreducible. The basis functions must be degenerate with one another. If there are invariant subspaces, we check these separately. We finally end up with a collection of irreducible representations. The basis functions of each of these form a necessarily degenerate set.

In describing this process, it is convenient in general to use a "coordinate space" of dimension $d_\mu=m$, and call the basis set $y_1,\ldots,y_m$. The transformations R of the group G are represented by operators O_R that act on the coordinates $(y_1,\ldots,y_m)$ or by their matrix representatives $D^{(\mu)}(R)$.

An important method for constructing representations for transformation groups is to realize that the group supplies us with a representation of itself: each element R of the group is its own representative $D(R)$. If n is the dimension of the coordinate space on which the group acts, the representatives will be $n\times n$ matrices $D^{(1)}(R)$. This is the vector representation of the group G. Any quantity that transforms according to this representation is called a vector relative to the group G. The product of representatives of R and S is $D^{(1)}(RS)=D^{(1)}(R)D^{(1)}(S)$. We also sometimes refer to the basis of this representation as a tensor of rank 1 (relative to G). We can define tensors of higher rank relative to G as follows: If x and y are vectors we can form the n^2 products x_iy_j. Under the transformations of G these will transform so that $x_i'y_j'=D_{ik}^{(1)}(R)D_{jl}^{(1)}(R)x_ky_l$, where we always use the convention of summing over any repeated index. (Note that the same transformation R is applied to each factor.) Any set of quantities T_{ij} that transform like this product is called a second-rank tensor: $T_{ij}'=D_{ik}^{(1)}(R)D_{jl}^{(1)}(R)T_{kl}$. Similarly we can construct tensors of any rank:

$$T'_{i_1,\ldots,i_n}=D^{(1)}_{i_1j_1}(R)D^{(1)}_{i_2j_2}(R)\ldots D^{(1)}_{i_mj_m}(R)T_{j_1,\ldots,j_m}. \tag{14}$$

These representations will be reducible, but can be reduced by the methods described in Hamermesh (1989), Chap. 10.

2.2.1 Equivalent Representations. If we have a representation $D(R)$ of the group G and make a transformation of basis in the space of the representation, using some matrix P, we can form matrices $PD(R)P^{-1}=D'(R)$, and it is easy to verify that these also form a representation of G with the same dimension as D. We also see that $D'(E)=1$, and $D'(R)D'(S)=D'(RS)$, since $PD(R)P^{-1}\times PD(S)P^{-1}=PD(R)D(S)P^{-1}=PD(RS)P^{-1}$. We say that the two representations D and D' are equivalent. They are isomorphic and represent only a change of basis. Equivalent representations have the same set of eigenvalues. The diagonal sum (trace) is the same for $D(R)$ and $D'(R)$. We call this quantity which is the same for any two equivalent representations the character $\chi(R)$ of R in the representation D: $\mathrm{Tr}\,D'=D'_{ii}=D_{ii}=\mathrm{Tr}\,D$. If two elements R and S of the group G are in the same conjugate class, so that there is an element P of the group such that $S=PRP^{-1}$, then $D(S)=D(P)D(R)D(P^{-1})$ and, taking the trace, we get $\chi(S)=\chi(R)$. Thus all elements in the same class in G have the same character.

2.2.2 Addition of Representations. Reducibility and Irreducibility. Suppose that we have two representations $D^{(1)}$ and $D^{(2)}$ of the same group G, where $D^{(1)}$ has dimension n_1 and acts on coordinates x_i, $i=1,\ldots,n_1$, and $D^{(2)}$ has dimension n_2 and acts on some other space with coordinates x_i, $i=n_1+1,\ldots,n_1+n_2$. Thus the matrices $D^{(1)}(R)$ are $n_1\times n_1$ matrices and the $D^{(2)}(R)$ are $n_2\times n_2$. For each R in G, we construct a new representation D that has dimension n_1+n_2,

$$D(R)=\begin{pmatrix} D^{(1)}(R) & 0 \\ 0 & D^{(2)}(R)\end{pmatrix}$$

and acts on the column vectors

$$\begin{bmatrix} x_1 \\ \vdots \\ x_{n_1} \\ x_{n_1+1} \\ \vdots \\ x_{n_1+n_2} \end{bmatrix},$$

and $D(R)=D^{(1)}(R)\dot{+}D^{(2)}(R)$. The two representations act separately in the two subspaces. If we make a change of basis in the (n_1+n_2)-dimensional space so that the x's mix, the equivalent representation will no longer have this block form, and it will not be evident that the representation is actually a "sum" of two independent representations. Now we proceed in the reverse directions. Given a representation of the group in some space of dimension N, we ask whether it is possible to find a simultaneous decomposition of the matrices for all R into invariant subspaces of dimensions n_1, n_2 such that $n_1+n_2=N$, so that $D(R)$ acting on vectors in subspace 1 gives vectors in that subspace, and similarly for subspace 2. If this is true for all R in G we say that the representation D of dimension N is reducible and has been decomposed into the sum of $D^{(1)}$ and $D^{(2)}$. Next we try to find decompositions of the subspaces 1 and 2 into invariant subspaces. This process must end in a finite number of steps, and we finally have a decomposition of the original representation space into a sum of subspaces, and $D=D^{(1)}\dot{+}D^{(2)}\dot{+}\cdots\dot{+}D^{(r)}$, where none of the representations $D^{(i)}$ has an invariant subspace. Then we say that the original representation $D(R)$ has been *fully reduced.* If the carrier space of D contains no proper invariant subspace, we say that the representation is irreducible. Note that the character of the representation $D(R)$ is the sum of the characters of its irreducible component representations: $\chi(R)=\Sigma_i\chi^{(i)}(R)$. The same irreducible representation can occur several times in the decomposition, so that, in general, $D=\dot{\Sigma}_i a_i D^{(i)}$, and $\chi=\Sigma_i a_i\chi^{(i)}$, where a_i are positive integers.

2.2.3 Invariance of Functions and Operators. We found earlier that $O_R\psi(x)=\psi(R^{-1}x)$, so $O_R\psi$ is not the same as ψ. If $O_R\psi$ is the same as ψ, so that $\psi(Rx)=\psi(x)$, we see that the function ψ is invariant under the transformations. To test for invariance we replace the arguments x of any function $\psi(x)$ by Rx and see whether we get the same expression. If an operator T acts on a function ψ we get a function $\varphi=T\psi$, and applying the operator O_R to $T\psi$ we get

$$O_RT\psi(x)=\varphi(R^{-1}x)=T(R^{-1}x)\psi(R^{-1}x), \quad (15)$$

$$\begin{aligned} O_RT(x)O_R^{-1}O_R\psi(x)&=T(R^{-1}x)O_R\psi(x)\\ &=T'(x)O_R\psi(x), \end{aligned} \quad (16)$$

where $T'(x)=T(R^{-1}x)$. In general the operators T and T' are different. If they are the same, i.e., if

$$O_RT(x)O_R^{-1}=T(x), \quad (17)$$

the operator T is invariant under the transformation. In other words the operator T commutes with O_R. For example, the function x^2+y^2 is invariant under rotations around the z axis, and also under inversion. The operator $\partial^2/\partial x^2+\partial^2/\partial y^2$ is also invariant under these transformations. A function describing the state of a system of identical particles is invariant under any permutation of the particles.

Now we can see the connection of group theory with physics. The Hamiltonian H of a physical problem may be invariant under some operators T. The collection of these operators and all possible products form a group G—the symmetry group of the Hamiltonian H. The basis functions for an irreducible representation of this symmetry group must be transformed into vectors in this same space by the operators $D^{(\mu)}(R)$. This m-dimensional space provides us with an m-fold degeneracy of the particular energy level. For a reducible representation of the symmetry group G more states seem to be transformed among themselves, but we can reduce the representation by finding the irreducible components. The basis functions for an irreducible representation *must* transform into one another under the operations of the symmetry group. It may happen that a problem gives degeneracies that are greater than expected from the assumed symmetry group. We must then search for some symmetry operation beyond those assumed. Often this occurrence is labeled as "accidental" degeneracy. Note that the basis function $\psi_i^{(\mu)}$ for the irreducible representation $D^{(\mu)}$ is said to belong to the ith row of the representation.

2.3 General Theorems

We now list without proof the general theorems that enable us to reduce any representation of a group G into its irreducible constituents.

Schur Lemma 1. If D and D' are two irreducible representations of a group G, having different dimensions, and the matrix T satisfies $D(R)T=TD'(R)$ for all R in G, then the matrix T must be the null matrix $T=0$.

Schur Lemma 1a. If D and D' have the same dimensions and are irreducible representations of G, and if $D(R)T=TD'(R)$ for all R, then either D and D' are equivalent or $T=0$.

Schur Lemma 2. If the matrices $D(R)$ are an irreducible representation of G and if $TD(R)=D(R)T$ for all R, then T is a multiple of the unit matrix: $T=\text{const }1$. This lemma gives an easy test of irreducibility.

Next we present the orthogonality relations. The quantities $D_{ij}^{(\mu)}(R)$ for fixed μ, i, j form a vector (whose components are labeled by R) in a g-dimensional space (where g is the order of G). If $D^{(\mu)}$ and $D^{(\nu)}$ are two nonequivalent irreducible representations of G, then

$$\sum_R D_{il}^{(\mu)}(R)D_{mj}^{(\nu)}(R^{-1})=\frac{g}{n_\mu}\,\delta_{\mu\nu}\delta_{ij}\delta_{lm}, \tag{18}$$

where n_μ is the dimension of the representation $D^{(\mu)}$. If $D^{(\mu)}$ and $D^{(\nu)}$ are the same D, then

$$\sum_R D_{il}(R)D_{mj}(R^{-1})=\frac{g}{n}\,\delta_{ij}\delta_{lm}. \tag{19}$$

If the representations are unitary, we replace $D_{mj}(R^{-1})$ by $D_{jm}^{*}(R)$. Thus each irreducible representation $D^{(\mu)}$ gives us n_μ^2 vectors $D_{ij}^{(\mu)}(R)$ $(i, j=1,\dots,n_\mu)$ that are orthogonal to one another and to all such vectors formed from nonequivalent representations. Since the number of orthogonal vectors in a g-dimensional space cannot exceed g, we must have $\Sigma_\mu n_\mu^2\leqslant g$. We can obtain similar formulas for the characters $\chi^{(\mu)}$ by setting $i=l$ and $j=m$ and summing over i and j:

$$\sum_R \chi^{(\mu)}(R)\chi^{(\nu)}(R^{-1})=g\delta_{\mu\nu}\,, \tag{20}$$

or

$$\sum_R \chi^{(\mu)}(R)\chi^{(\nu)*}(R)=g\delta_{\mu\nu} \tag{21}$$

if the representation is unitary. We saw earlier that all the group elements in the same class have the same character. We label the classes K_1 to K_r, and denote by g_i the number in the class K_i. Then the last equation becomes

$$\sum_{i=1}^{r} \chi_i^{(\mu)}\chi_i^{(\nu)*}g_i=g\delta_{\mu\nu}, \tag{22}$$

where $\chi^{\mu}(R)=\chi_i^{(\mu)}$ for all elements in the class K_i.

2.4 Kronecker Product of Representations

If we have two irreducible representations $D^{(\mu)}$ and $D^{(\nu)}$ of the symmetry group, we can take products of their basis functions and get the Kronecker product representation $D^{(\mu\times\nu)}(R)$ with matrices

$$[D^{(\mu\times\nu)}(R)]_{ik,jl}=D_{ij}^{(\mu)}(R)D_{kl}^{(\nu)}(R). \tag{23}$$

The character of $D^{(\mu\times\nu)}$ can be found by setting $j=i$, $l=k$, and summing over i and k:

$$\chi^{(\mu\times\nu)}(R)=\chi^{(\mu)}(R)\chi^{(\nu)}(R). \tag{24}$$

All elements R in the same class K_i will have the same character $\chi_i^{(\mu\times\nu)}$. The scalar product of the basis functions is

$$(\psi_i^{(\mu)},\varphi_j^{(\nu)})=\int d\tau\, \psi_i^{(\mu)*}\varphi_j^{(\nu)}. \tag{25}$$

For unitary representations this expression is identical with $(D^{(\mu)}(R)\psi_i^{(\mu)},D^{(\nu)}(R)\varphi_j^{(\nu)})$ for any R. If we use Eq. (13) and the orthogonality relation of Eq. (18) we find

$$(\psi_i^{(\mu)},\varphi_j^{(\nu)})=\frac{1}{n_\mu}\sum_k (\psi_k^{(\mu)},\varphi_k^{(\nu)})\delta_{\mu\nu}\delta_{ij}. \tag{26}$$

The scalar product is zero for $i\neq j$, i.e., basis functions belonging to different rows are orthogonal. Setting $\mu=\nu$ and $i=j$, we find that $(\psi_i^{(\mu)},\varphi_i^{(\mu)})$ is independent of i. This means that the scalar product of two functions belonging to the same row of an irreducible representation is independent of the row. We shall see that this result is the basis of the use of perturbation theory and selection rules throughout the applications of group theory.

2.5 Analysis of Representations

If we know the characters of the irreducible representations of the group G, we can use the above theorems to find how a given representation decomposes into irreducible constituents. We found (see p. 376) the equation

$$D(R)=\sum_{\mu} a_\mu D^{(\mu)}(R). \tag{27}$$

Taking the trace for an element R in the class K_i, we get

$$\chi_i=\sum_{\mu} a_\mu \chi_i^{(\mu)}. \tag{28}$$

Next we multiply by $\chi_i^{(\nu)*}g_i$ and sum over i. Using the orthogonality relations found earlier, this gives

$$a_\nu=\frac{1}{g}\sum_{i} g_i \chi_i^{(\nu)*}\chi_i. \tag{29}$$

Thus the number of times a given irreducible representation is contained in D is given by this formula. In particular this shows that if two representations have the same set of characters, they are equivalent. Again, if we multiply Eq. (29) by g times its complex conjugate equation and sum over i, we find

$$\sum_{i} g_i|\chi_i|^2=g\sum_{\mu} a_\mu^2. \tag{30}$$

If the representation is irreducible, all the a_μ must be zero, except for one which is equal to 1. So if the representation is irreducible, its characters must satisfy the equation

$$\sum_{i} g_i|\chi_i|^2=g, \tag{31}$$

which gives a simple test for irreducibility. Finally one can show that the number of inequivalent irreducible representations is precisely equal to the number of conjugate classes in the group, and that

$$g=\sum_{\mu=1}^{r} n_\mu^2, \tag{32}$$

i.e., the sum of the squares of the dimensions of all the nonequivalent irreducible representations is equal to the order of the group.

We give some examples of finding characters for some groups.

1. Cyclic groups: $a,\ldots,a^n=e$. These groups are Abelian, so all irreducible representations have dimension 1. The matrices are 1×1 and the representative is just the character, a complex number. Since $a^n=e$, the nth power of the character $D(a)$ must equal 1, so $D(a)=\exp(2\pi im/n)$, $m=1,\ldots,n$, and $D(a^r)=\exp(2\pi imr/n)$.
2. General Abelian group. Again all irreducible representations are one-dimensional. We choose any element and take its powers. This gives a cyclic subgroup of G. We repeat this process with some other element. We see that the group is a direct product of cyclic subgroups. For example, if $G=C_2\otimes C_3\otimes C_5$, with $g=30$, we have generators a,b,c, with $a^2=b^3=c^5=e$, and the character of any element $a^m b^n c^p$ is $\exp[2\pi i(mr/2+ns/3+pt/5)]$.
3. Point groups. These are the groups of rotations about a fixed point (the origin) and reflections in planes through the origin. For crystals, only the rotations through multiples of $2\pi(\frac{1}{2};\frac{1}{3};\frac{1}{4};\frac{1}{6})$ are permitted, i.e., only cyclic subgroups C_2, C_3, C_4, C_6. As a typical example, we treat the octahedral (cubic) group O. This is the group of rotations about axes through the center of the cube that take the cube into itself. It consists of 24 rotations in 5 classes: 6 rotations C_4,C_4^3 around lines joining the midpoints of opposite faces; 3 rotations C_4^2 around these same axes; 6 rotations C_2 around lines joining the midpoints of opposite edges; 8 rotations C_3,C_3^2 around lines joining opposite vertices of the cube; and the identity E. There are five nonequivalent irreducible representations. Using Eq. (32) we find

$$24=\sum_{\mu=1}^{5} n_\mu^2$$

which has the unique solution $n_\mu=3,3,2,1,1$. The character table for this group is

	E	$C_3,C_3^2(8)$	$C_4^2(3)$	$C_2(6)$	$C_4,C_4^3(6)$
Γ_1	1	1	1	1	1
Γ_2	1	1	1	-1	-1
Γ_3	2	-1	2	0	0
Γ_4	3	0	-1	1	-1
Γ_5	3	0	-1	-1	1

where $\Gamma_1,\ldots,\Gamma_5$ label the irreducible representations. The column vectors are orthogonal:

$$1(1)+1(1)+2(-1)+3(0)+3(0)=0, \text{ etc.,}$$

and normalized to $g=24$:

$$1^2+1^2+2^2+3^2+3^2=24,$$
$$8[1^2+1^2+(-1)^2]=24, \text{ etc.}$$

The row vectors are also orthonormal when we include the factors g_i:

$$1(2)+8(1)(-1)+3(1)(2)+6(-1)(0)$$
$$+6(-1)(0)=0,$$
$$3^2+8(0)(0)+3(-1)^2+6(1)^2+6(-1)^2$$
$$=24, \text{ etc.}$$

4. The permutation groups S_n. These finite groups are important for dealing with identical particles, construction of irreducible tensors, and finding states in atomic or nuclear shell models. Earlier we described the conjugate classes in S_n. We found that there is a class corresponding to each partition of n. For example, for $n=3$, we have $r=3$ and $g=3!$, so $6=\Sigma_i n_i^2$, so $n_i=2,1,1$. There are two one-dimensional and one two-dimensional irreducible representations. The character table is

	E	$(12)_3$	$(123)_2$
Γ_1	1	1	1
Γ_2	1	-1	1
Γ_3	2	0	-1

For large n, the simple procedure becomes impractical. Instead we use Young diagrams. For each partition of $n=n_1+n_2+\cdots+n_r$, with $n_1 \geqslant n_2 \geqslant \ldots \geqslant n_r$, we draw a diagram with n_1 dots in the top row, n_2 in the second, etc. For example, for $n=5=3+1+1$ we get the Young diagram

• • •
•
•

Each such partition gives an irreducible representation of the group S_n. Next we enter the digits 1–5 in the boxes in all possible ways that keep the entries increasing to the right in rows and down in columns. These arrangements are the standard Young tableaux:

$$\begin{array}{cccccc} 123 & 124 & 125 & 134 & 135 & 145 \\ 4 & 3 & 3 & 2 & 2 & 2 \\ 5 & 5 & 4 & 5 & 4 & 3\,. \end{array}$$

There are six standard tableaux so the dimension of this irreducible representation is 6. The use of the symmetric group S_n and the construction of irreducible representations is discussed in Hamermesh (1989), Chaps. 7 and 10.

5. SO(3) is the group of pure rotations in three dimensions. All rotations through a given angle θ about any axis **n** are in the same class (see p. 368), so if we choose the z axis for **n** the rotation matrix is

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and the character for the vector representation is $\chi^{(1)}(\theta)=1+2\cos\theta=e^{i\theta}+e^0+e^{-i\theta}$. One can show that there is a single irreducible representation for each integer $l=0,1,\ldots$ with

$$\chi^{(l)}(\theta)=\sum_{m=-l}^{+l} e^{im\theta}, \tag{33}$$

and the dimension of the representation is $\chi^{(l)}(0)=2l+1$. There are also irreducible representations of SO(3) for $l=\frac{1}{2}, \frac{3}{2}$, etc. These are double-valued representations that can be derived by using the group SU(2), which is homomorphic to SO(3) with two elements corresponding to each element of SO(3). These irreducible representations give all information needed about states in a central field. For a detailed treatment of group representations, see Elliott and Dauber (1979), Chaps. 4 and 7; Hamermesh (1989), Chaps. 3, 4, and 5.

3. APPLICATIONS

3.1 Atomic Structure

The application of group theory to most physical problems requires the use of some model. It should be evident that we cannot consider the quantum problem exactly if we have more than two entities interacting with one another. We must find some method of successive approximations that enables us finally to get satisfactory agreement with experimental results. In atomic physics our procedure is first to consider the individual electrons to be moving in the field of the nucleus and some spherically symmetric averaged field of the other electrons in the atom. In this central field the Hamiltonian for the individual electron has the symmetry group SO(3), so the states of a single electron have quantum numbers l and $m=-l,\ldots,+l$. A single-particle level with quantum number l has degeneracy $2l+1$. There will be many states with a given lm, with energies $E=b/n^2$ approximately, where $n=1,2,\ldots$ giving us a triple of labels n, l, m, with $n\geqslant l+1$, and $l=0,1,\ldots$, where l is the orbital angular momentum. Often we use letters in place of l:

$l=$	0	1	2	3	4
state label	s	p	d	f	g.

For $l=0$, we have 1s, 2s, 3s,...; For $l=1$, we have 2p, 3p,...; For $l=2$, we have 3d, 4d,...,etc. The energy levels are given approximately by $E=b/n^2$, so the level ns is nondegenerate ($m=0$), the level np has $m=\pm 1,0$, etc. In addition we know that each electron carries an internal angular momentum (spin) with $s=\frac{1}{2}$. In light atoms the spin results in there being two states (with $m=\pm\frac{1}{2}$). This results in doubling the occupancy of each level. Thus we have a level sequence

Level	1s	2s	2p	3s	3p	3d	4s	4p	4d	4f
Degeneracy	2	2	6	2	6	10	2	6	10	14 ,

etc., where the atomic shells are labeled by n, and have a total number of electrons=2, 8, 18,... . The order of the levels goes with n, but some changes occur when the 3d and 4f levels become partly filled.

To study the spectrum of levels for an atom with atomic number Z, we fill the single-particle levels successively. For example, for $Z=16$, the only unfilled shell would be 3p, with two electrons. This state would be described by $(1s)^2(2s)^2(2p)^6(3s)^2(3p)^2$. The closed inner shells have spherical symmetry, so we consider only the two electrons in the 3p shell. The orbital states of the two electrons have $l_1=l_2=1$, while the spin states are $s_1=s_2=\frac{1}{2}$. For light atoms we use Russell-Saunders coupling, in which we first find the Kronecker product of all the space functions, and separately of all the spin functions, and then couple the two resultants. To keep the Coulomb force between the electrons unchanged requires that the same rotation R be applied to both electrons, so that we are taking the product representation $D^{(l_1)}(R)D^{(l_2)}(R)$. The basis functions are $\psi^{(1)}_{m_1}(1)\psi^{(1)}_{m_2}(2)$, giving nine coupled wave functions. This reducible representation can be decomposed into $L=2,1,0$. The two spins will couple to $S=1,0$. Finally we include the weaker coupling of orbit and spin to give the states of total angular momentum $J=L+S$, and find states $^3D,^3P,^3S$ and $^1D,^1P,^1S$. But we must also consider that the electrons are identical particles (fermions), and so the Pauli principle requires that the total wave function must be antisymmetric under the interchange of all coordinates of the two electrons. For this simple case of two electrons the nine functions $\psi^{(1)}_{m_1}(1)\psi^{(1)}_{m_2}(2)$ can be split into a symmetric second-rank tensor (with trace zero), an antisymmetric second-rank tensor, and a scalar. The orbital states with $L=2,0$ are symmetric while the state with $L=1$ is antisymmetric. The spin states with $S=1$ are symmetric (triplet states) while the $S=0$ singlet states are antisymmetric. Thus the Pauli principle allows only the states $^3P,^1D,^1S$. Note that this argument would not apply to the case of $(3p)(4p)$ because then the states of the two electrons are not the same, so the Pauli principle has no effect.

This simple procedure fails if there are more than two electrons in the same subshell. For example, for the case of $(3p)^3$ we would have 27 product functions $\psi^{(1)}_{m_1}(1)\times\psi^{(1)}_{m_2}(2)\psi^{(1)}_{m_3}(3)$ while the spin part would be the product of three spin-$\frac{1}{2}$ functions. The general procedure requires the use of irreducible tensors. If we have r electrons in a subshell with angular momentum l our spatial wave function for one electron is $\psi^{(l)}_m$, and for the r electrons the product is $\Psi=\psi^{(l)}_{m_1}(1)\times\ldots\psi^{(l)}_{m_r}(r)$ with $m_i=-l,\ldots,+l$. Thus Ψ is an rth-rank tensor in a space of dimension $2l+1$.

Since all these functions are degenerate, the transformations of the group SU$(2l+1)$ act on these ψ's. We then must classify them according to the irreducible representations of SU$(2l+1)$ with their particular symmetries. Similarly, for the spins we have tensors $\varphi^{(1/2)}_{\mu_1}(1),\dots,\varphi^{(1/2)}_{\mu_r}(r)$ of rank r in the space of dimension 2. Again we must find the irreducible pieces. Finally, to satisfy the Pauli principle we must assure that the total wave function is antisymmetric. For the detailed treatment, see Hamermesh (1989), Chap. 11, or Elliott and Dauber (1979), Chap. 8.

3.2 Nuclear Structure

In the case of atoms, we know that the interactions are electromagnetic, and we have just one type of identical particle, electrons. In the case of the nucleus we have two constituents, neutrons and protons, which have approximately equal masses and can transform into one another in β-ray processes. The Coulomb force between protons is small compared with the specific nuclear force, so we use a model in which we disregard the differences between n and p, and deal with a single type of identical particles, nucleons, in two charge states. This looks like the case of two possible spin states for electrons. Both n and p have an intrinsic angular momentum $s=\frac{1}{2}$, but now there is another intrinsic feature, the isospin $t=\frac{1}{2}$. The other important difference is that we have little basis for choice of a model. We start with no knowledge of the interaction and must try various shapes for its radial dependence and its dependence on spin and isospin.

The most successful model requires the use of j-j coupling. In such models we first couple (i.e., take the product of) the $\psi^{(l)}$ and $\psi^{(s)}$ for an individual nucleon to give a resultant $\psi^{(j)}$ and then couple the $\psi^{(j)}$s (i.e., take products of $\psi^{(j)}_m$ for the nucleons). Since the nuclear force is attractive, the state with lowest j for a given l,s will have the highest symmetry (since the particles can get closest to one another). The order of the levels will also depend on the choice of the radial dependence of the potential. One model is shown in the following diagram of single-particle levels and their degeneracies:

Nucleon states		Occupation	Total
$1g_{9/2}$		10	50
	$2p_{1/2}$	2	
$1f_{5/2}$		6	
	$2p_{3/2}$	4	
$1g_{7/2}$		8	28
$1d_{3/2}$		4	20
	$2s_{1/2}$	2	
$1d_{5/2}$		6	
$1p_{1/2}$		2	8
$1p_{3/2}$		4	
$1s_{1/2}$		2	2

The numbers in the right-hand column are the magic numbers corresponding to the closed shells.

As in the atomic case, we look only at the partially filled shells. If we consider a level $(j)^3$, the single-particle state has a wave function $\psi^{(j)}$ that is a vector in a space of dimension $2j+1$. Since the js are halves of odd integers, $2j+1$ is even. The wave function for $(j)^3$ is $\psi^{(j)}_{m_i}(1)\psi^{(j)}_{m_2}(2)\psi^{(j)}_{m_3}(3)$, a third-rank tensor in a space of dimension $2j+1$. Next we take the product of the three isospin functions which are vectors in a space of dimension 2. Just as in the atomic case, we must find the irreducible parts of each tensor. For the isospin tensor we get $T=\frac{3}{2},\frac{1}{2},\frac{1}{2}$. The completion of the problem would be the same as in the atomic case [see Hamermesh (1989), Chap. 11-9].

3.3 Solid State and Crystal-Field Theory

In discussing atomic structure we used a central-field model in which the Hamiltonian had the symmetry group SO(3). If we study the energy levels of electrons in a crystal, this is no longer the correct symmetry group. Now the electron moves in the periodic field of the crystal lattice. In the neighborhood of a lattice point the field is no longer spherically symmetric. It now has the symmetry of one of the crystal point groups. For example, the wave function of an electron in a cubic crystal will belong to an irreducible representation of the cubic point group O. Thus the possible degeneracy of a level will be 1, 2, or 3. (See the character table on p. 378.) If the crystal is distorted, the point symmetry group will be reduced from O to some lower symmetry and the levels may split. Or we may consider what happens to a level in a free atom when the atom is put at a crystal site. In the free atom the levels correspond to irreducible representations of the symmetry group SO(3). When the atom is placed in the crystal, we must use the crystal-field symmetry. The level with a given l may split into levels belonging to the symmetry group of the crystal ("crystal-field theory"). We illustrate this for the octahedral group. A level belonging to the irreducible representation $D^{(L)}$ of SO(3) has the character

$$\chi^{(L)}_{(\theta)} = \sum_{m=-L}^{+L} e^{im\theta}.$$

In the crystal this representation will be reducible. The cubic field has only rotations through angles $\theta=0$, $\pi/2$, π, and $2\pi/3$. In the crystal we need to record only the $\chi^{(L)}(\theta)$ for these values of θ. We repeat the character table of O from p. 376 and enter below the characters for $L=0$, 1, etc. Then we decompose using Eq. (29):

	E	$C_3,C_3^2(8)$	$C_4^2(3)$	$C_2(6)$	$C_4,C_4^3(6)$
Γ_1	1	1	1	1	1
Γ_2	1	1	1	-1	-1
Γ_3	2	-1	2	0	0
Γ_4	3	0	-1	1	-1
Γ_5	3	0	-1	-1	1
$L=0$	1	1	1	1	1
$L=1$	3	0	-1	-1	1
$L=2$	5	-1	1	1	-1
$L=3$	7	1	-1	-1	-1
$L=4$	9	0	1	1	1

For example, we see that $L=0$ gives Γ_1. $L=1$ is just Γ_5 so the vector does not split in a cubic field. For $L=2$ we find $\Gamma_3+\Gamma_4$, so the level splits into a doublet and a triplet. $L=3$ splits into $\Gamma_2+\Gamma_4+\Gamma_5$, while $L=4$ gives $\Gamma_1+\Gamma_3+\Gamma_4+\Gamma_5$. For details and applications to various point groups and the construction of wavefunctions, see Hamermesh (1989), Chap. 9; Elliott and Dauber (1979), Chap. 14; or Burns (1977), Chaps. 8, 11, and 12.

4. SUMMARY

Group theory has developed in the last 60 years to become an essential tool in many branches of physics and chemistry. After a mathematical introduction to the subject, we have given examples of its application to atomic and nuclear structure, and to crystal-field theory.

List of Works Cited

Burns, G. (1977), *Introduction to Group Theory with Applications,* New York: Academic.

Elliott, J. P., Dauber, P. G. (1979), *Symmetry in Physics,* Vols. 1 and 2, New York: Oxford.

Hamermesh, M. (1989), *Group Theory and Its Application to Physical Problems,* New York: Dover. [Original publication (1962), Reading, MA: Addison-Wesley.]

Further Reading

Burns, G. (1977), *Introduction to Group Theory with Applications,* New York: Academic.

Elliott, J. P., Dauber, P. G. (1979), *Symmetry in Physics,* Vols. 1 and 2, New York: Oxford.

Hamermesh, M. (1962), *Group Theory and Its Application to Physical Problems,* Reading, MA: Addison-Wesley; reprint (1989), New York: Dover.

Perkins, D. H. (1987), *Introduction to High Energy Physics,* Reading, MA: Addison-Wesley.

Tung, W.-K. (1985), *Group Theory in Physics,* Philadelphia: World Scientific.

HEAT ENGINES AND REFRIGERATORS

FRANCIS JOSEPH HALE, *North Carolina State University, Raleigh, North Carolina, U.S.A.*

Introduction 383
1. **Thermodynamic Considerations** 384
1.1 Thermodynamic Definitions 384
1.2 Thermodynamic Laws and Equations 385
1.3 Cycle Performance and Efficiency 386
1.4 The Carnot Cycle 387
2. **Vapor (Power) Cycles** 388
2.1 The Vapor Carnot Cycle 388
2.2 The Rankine Cycle 389
2.3 Reheat and Regeneration 390
2.4 Other Vapor Cycles 391
3. **Refrigeration (and Heat-Pump) Cycles** 391
3.1 The Vapor-Compression Refrigeration Cycle 391
3.2 The Absorption Refrigeration Cycle 393
3.3 Other Refrigeration Cycles 393
4. **Gas Cycles** 394
4.1 The Carnot Cycle 394
4.2 The Air-Standard Cycle 394
4.2.1 The Otto Cycle 395
4.2.2 The Diesel Cycle 395
4.2.3 The Brayton Cycle 397
4.3 The Gas Refrigeration Cycle 399
4.4 The Stirling and Ericsson Cycles 400
5. **Cogeneration and Environmental Considerations** 401
Glossary 402
Further Reading 403

INTRODUCTION

Thermodynamic devices that convert heat into work are called *heat engines*; these work producers require both a heat source and a heat sink and, to be useful, should operate in a cyclic manner. Examples are steam power plants, internal combustion engines, and gas turbines. On the other hand, thermodynamic devices that use work to transfer heat from a lower temperature to a higher temperature are also heat engines but are referred to as *refrigerators*. In addition to traditional kitchen refrigerators, cold rooms, and air conditioners, whose purpose is to transfer heat out of a space so as to lower its temperature, the term refrigerators also includes *heat pumps*, whose purpose is to transfer heat into a space so as to increase its temperature. The term heat engine is used at times to denote both work producers and refrigerators and at others to denote work producers only.

Each heat engine has its own descriptive cycle that has losses, and thus is irreversible, and that does not lend itself to simple modeling and analysis. Fortunately, however, it is possible to approximate an actual cycle by an ideal cycle, a cycle with reversible (lossless) processes that can be analyzed to yield insight into the performance and design of the actual cycle, particularly with respect to the influence of various parameters on the cyclic performance. Furthermore, the performance of the ideal cycle is a measure of the best performance that can be expected from the actual cycle and can therefore be used as a figure of merit when comparing different configurations of an actual cycle. This article is limited to a description of each major heat engine (and refrigerator) in terms of its ideal cycle and associated performance; several of the actual cycles and their associated hardware are discussed in separate articles. Since the reader is assumed to have some knowledge and understanding of engineering thermodynamics (see THERMODYNAMICS, EQUILIBRIUM), only a brief review of those elements deemed necessary to an understanding of the ideal cycles is presented in this article.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. THERMODYNAMIC CONSIDERATIONS

1.1 Thermodynamic Definitions

Engineering thermodynamics deals primarily with thermodynamic systems containing pure substances (and gaseous air) whose composition is the same whether a solid, liquid, or gas. A thermodynamic system is defined by boundaries that must be carefully drawn so as to identify the problem of interest; for example, the thermodynamic behavior of an automobile engine cylinder with the boundaries drawn just inside the cylinder walls differs from the behavior with the boundaries drawn so as to include the cylinder walls. Heat and work can cross the boundaries of a thermodynamic system; although input heat and output work are generally considered to be positive, other sign conventions may be encountered. If mass does not cross the boundaries (a *closed system*), any work, in or out, is accomplished by movement of the boundaries. If, on the other hand, mass does cross the boundaries (an *open system*), the boundaries remain stationary and any work is shaft or indicated work.

The condition of a system at any instant of time or location is called the *state* of the system and is established by a minimum number of independent properties. A change in one or more of these properties results in a change of state, and one or more changes of state constitute a *process*. When a system undergoes one or more changes of state and returns to its original state, the series of processes forms a *cycle*. A *reversible process* is one with no losses, and the processes of a reversible cycle are all individually reversible.

The three fundamental thermodynamic properties are pressure P, temperature T, and specific volume v (the inverse of the density). Other properties of interest are the specific internal energy u, the specific enthalpy h, and the specific entropy s. Enthalpy and entropy are both derived properties, enthalpy for convenience in analyzing open systems and entropy as a measure of the reversibility (or irreversibility) of a process. The differentials of the properties are exact and can be integrated without regard to the path traveled in going from one state to another. Heat and work, on the other hand, are not properties since their differentials are inexact and cannot be integrated without a knowledge of the path traveled. Heat and work cannot be stored, only transferred or transformed.

Pure substances can have different molecular structures, with clearly identifiable boundaries, which are known as *phases*. The three principal phases are solid, liquid, and vapor (at high temperatures a vapor is usually referred to as a superheated vapor or a gas), and there may be phases within a principal phase. The three principal phases of water, for example, are known as ice, water (the liquid phase), and steam (or vapor). When more than one phase is present, as with ice in water, the phases are separated from each other by boundaries. There is a single combination of temperature and pressure (the *triple point*) where all three phases can simultaneously exist in equilibrium.

In thermodynamic systems, the two phases of primary interest are the liquid and vapor phases and their combination, referred to as a (two-phase) mixture. Figure 1 is a typical (not to scale) T-v phase diagram that shows these phases and the lines of demarcation. The region labeled SMR is called the *saturated-mixture region* and is a two-phase combination of liquid and vapor; the boundaries of the SMR define what is known as the *vapor dome*. The *compressed-liquid region* (CLR) lies to the left of the SMR, and the line of demarcation is known as the *saturated-liquid line*, the combination of P-v-T at which a liquid begins to vaporize. The region to the right of the vapor dome labeled SHV is the *superheated-vapor region* (which blends into the gaseous region) and this line of demarcation is the *saturated-vapor line*. Within the SMR, the composition of the two-phase (saturated) mixture ranges from all liquid (the start of vaporization or boiling) to all vapor (no liquid remaining) and is described in terms of its *quality*, the percentage of vapor mass present.

The saturated-liquid and saturated-vapor lines come together at the *critical point* at the top of the vapor dome. At temperatures above the critical point, there is no distinction between the liquid and vapor phases; the substance becomes a *supercritical fluid*. Two isobars (lines of constant pressure) have been drawn to show the nature of the isobars in the various regions.

Tables of thermodynamic properties in these various regions are available for substances commonly used in thermodynamic

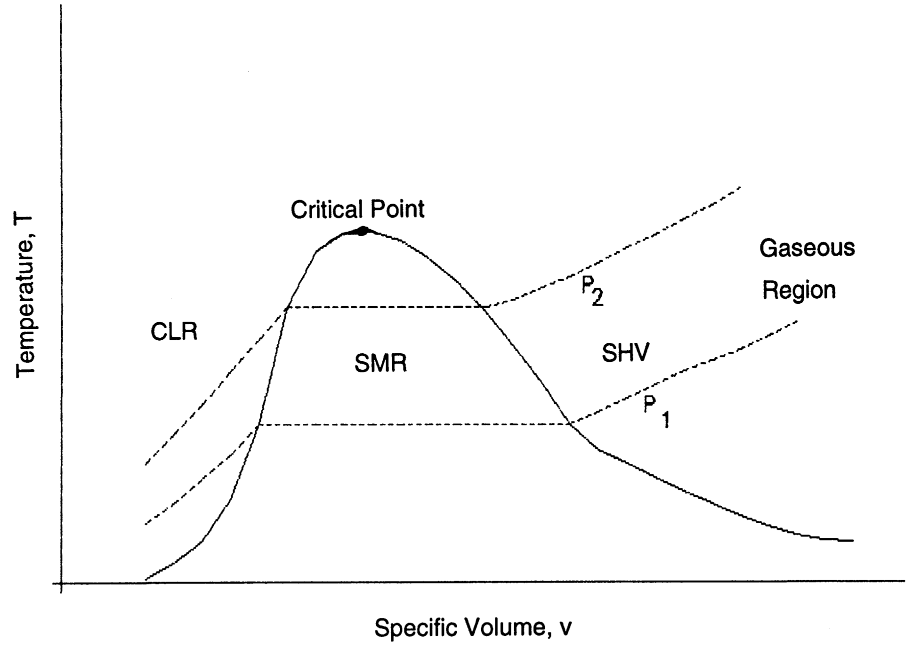

FIG. 1. A typical *T-v* diagram (not to scale) showing the regions of thermodynamic interest.

systems; for example, there are extensive tables for water (the steam tables) along with tables for such substances as ammonia, nitrogen, mercury, and the various Freon refrigerants. These tables have been constructed primarily from experimental data that have been correlated with and verified by thermodynamic relationships.

Tables are not needed for *ideal* or *perfect gases*, which are gases at high temperatures and low pressures whose intermolecular forces and molecular size can be neglected. Their properties can be found from relationships such as the *P-v-T* relationship,

$$Pv=RT, \tag{1}$$

where P is the pressure, v the specific volume, T the absolute (thermodynamic) temperature, and R the gas constant for the gas of interest. R can be found from tables or from the relationship that $R=R_u/M$, where R_u is the universal gas constant and M is the molecular mass of the gas.

Equation (1) is known as the *ideal-gas law* or the *equation of state for a perfect gas.* Although Eq. (1) may be used with acceptable error (along with other relationships) to determine the properties of many gases, including air, property tables should be used for the denser gases, such as water and refrigerant vapors in the vicinity of the SMR. The specific heats (C_p and C_v) are especially useful when dealing with changes of state of ideal gases as is the *ratio of specific heats k*, where $k=C_p/C_v$ and is always greater than unity.

1.2 Thermodynamic Laws and Equations

The continuity equation and the first law of thermodynamics are expressions of the conservation of mass and of energy, respectively. As for the mass balance of the continuity equation, the system mass is constant for a closed system; for an open system with steady-state flow the sum of the mass entering the system must be equal to the sum of the mass leaving the system.

As for the energy balance, the first law states that the net heat transferred across the boundaries is proportional to the net work done during a cycle (with SI units the constant of proportionality is unity). For a process, on the other hand, the net energy crossing the boundaries (the work and heat) will be equal to the change of the energy (internal, kinetic, and potential) of the system itself. For a closed system, the first law can be written on a per-unit-mass basis as

$$q=u_2-u_1+w=\Delta u-w, \tag{2}$$

where the values of q and w (the actual heat and work) are functions of how the change of state was accomplished (the path) whereas Δu, the change in internal energy, is a function only of the two states 1 and 2.

There is a large and important class of steady-state applications in which there is only one flow entering and leaving the system. They are known as *single-flow systems*, and the first law can be written as

$$q=(h_e-h_i)+(v_e^2/2-v_i^2/2)+w, \tag{3}$$

where h is the enthalpy per unit mass and the subscripts i and e denote the inlet and outlet states.

When the change in kinetic energy is negligible (as it is for well-designed systems, other than nozzles), Eq. (3) becomes

$$q=h_e-h_i+w=\Delta h+w. \tag{4}$$

Notice that Eq. (4) is similar in form to Eq. (2), the first law for a closed system, but with several differences. Whereas the closed system deals with the change in internal energy (Δu) over an interval of time, the open system deals with the change of enthalpy (Δh) between two locations, a spatial interval; in addition, the nature of the work involved differs, one being moving-boundary work and the other shaft work.

Although very useful, the first law has its shortcomings. For example, it does not indicate any limitations on the transformations between heat and work, such as the fact that all of the input heat cannot be transformed into work; the Kelvin-Planck statement affirms that some portion of the input heat must be transferred out of the system in order to complete a cycle (this "rejected heat" is commonly referred to as the *waste heat*). Neither does the first law show, in accordance with the Clausius statement, that work is required to transfer heat from a cooler body to a warmer body in order to complete a refrigeration cycle, nor does it indicate which processes can proceed in one direction only and may even be impossible.

These two statements, along with the inequality of Clausius, constitute what is known as *the second law of thermodynamics*. The second law leads to the derivation of a new property called *entropy*; the specific entropy can be used as an independent property to define a thermodynamic state. Since ds, the exact differential of the specific entropy, is defined as $\delta q/T$ for a reversible process, the amount of heat per unit mass that is transferred in a *reversible process* can be determined from the expression

$$q=\int_1^2 T\,ds, \tag{5}$$

where the limits of integration represent the start and end of the process and the relationship between T and s must be known.

Equation (5) shows that if a process is both reversible and adiabatic ($q=0$), the change in entropy will be zero; i.e., the process is *isentropic.* [If a reversible process is isothermal, then ${}_1q_2=T(s_2-s_1)$.]

Space does not permit a detailed discussion of the various expressions of the second law and their use, along with the entropy, to determine the nature (reversible, irreversible, or impossible)of cycles and processes, to determine a missing property, to serve as a baseline for evaluating process efficiencies, or to develop useful thermodynamic relationships.

1.3 Cycle Performance and Efficiency

The performance of a heat engine is described in terms of its *cycle (thermal) efficiency,* which is defined as the ratio of the net work output of the cycle to the input heat. Remembering that the first law for a cycle states that the net work is equal to the net heat transferred, the thermal efficiency of the cycle can be expressed on a per-unit-mass basis as

$$\eta_{\text{th}}=\frac{w_{\text{net}}}{q_{\text{in}}}=\frac{q_{\text{in}}-q_{\text{out}}}{q_{\text{in}}}=1-\frac{q_{\text{out}}}{q_{\text{in}}}, \tag{6}$$

where q_{out} includes the rejected (waste) heat (required by the second law) that keeps the efficiency less than 100%, as well as all other heat losses.

A refrigeration cycle is essentially a reversed heat-engine cycle whose objectives are to maximize the desired heat transfer and to minimize the required work input. Since the ratio of the transferred heat to the required work is greater than unity, the *coefficient of* *performance* (COP) is used to avoid efficiencies that would be greater than 100%.

With a refrigerator (or air conditioner) the desired heat transfer is that removed from the refrigerated space and is given the symbol q_L to denote transfer at the lower temperature; q_H denotes the necessary transfer of heat at the higher temperature. Consequently, the COP of a refrigerator is defined as

$$\mathrm{COP}=\frac{q_L}{w_{\mathrm{in}}}=\frac{q_L}{q_H-q_L}=\frac{1}{q_H/q_L-1}. \tag{7a}$$

With a heat pump, q_H, the heat transferred into the heated space, is the desired heat transfer, and the coefficient of performance $(\mathrm{COP})_{\mathrm{HP}}$ is given by

$$(\mathrm{COP})_{\mathrm{HP}}=\frac{q_H}{w_{\mathrm{in}}}=\frac{q_H}{q_H-q_L}=\frac{1}{1-q_L/q_H}. \tag{7b}$$

Note that there is a definite relationship between the two COP's, namely,

$$(\mathrm{COP})_{\mathrm{HP}}=1+\mathrm{COP}. \tag{7c}$$

This relationship should be considered when selecting a heat pump to be used as an air conditioner as well as a space heater.

1.4 The Carnot Cycle

The Carnot cycle serves as a baseline for evaluating cycle performance in that it has the highest efficiency (COP) of all heat engines (and refrigerators) operating between two temperatures and in that the Carnot efficiency is independent of the working substance. The Carnot cycle is distinguished by the fact that its processes are totally reversible in that, in addition to the absence of internal losses, all transfers of heat are isothermal and therefore externally reversible processes.

Figure 2 is the *T*-*s* diagram of a heat engine (work producer or refrigerator) operating in a Carnot cycle between two temperatures T_H and T_L and involving four totally reversible processes. This Carnot cycle is a work producer if traveled in a clockwise direction and a refrigerator if traveled in a counterclockwise (reversed) direction. For operation as a work producer, the first process, 1–2, is an isothermal transfer of heat q_H into the system from a high-temperature thermal reservoir; from Eq. (5) q_H is the area under the line connecting states 1 and 2. The second process, 2–3, is an isentropic ($ds=0$) expansion from T_H to T_L that produces work. The third pro-

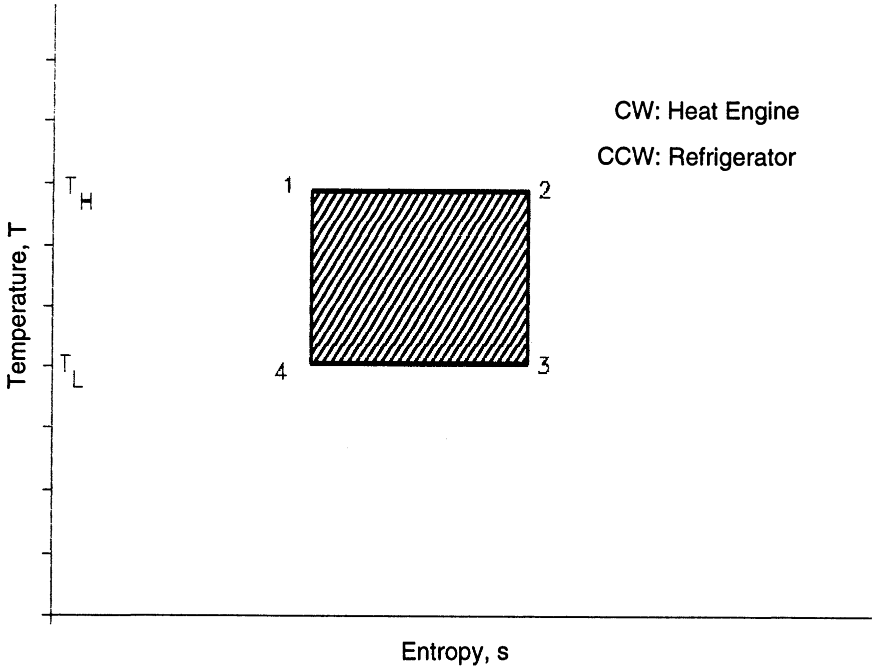

FIG. 2. The *T*-*s* diagram of a heat engine operating in a Carnot cycle.

cess, 3–4, is another isothermal transfer of heat, q_L, from the system to the low-temperature reservoir; q_L is the area under the line from 3 to 4. The fourth, and final, process is an isentropic compression, requiring a work input, that increases the temperature from T_L to T_H and completes the cycle. (The reversed cycle, 1–4–3–2–1, is that of a Carnot-cycle refrigerator.)

Since the thermal efficiency and COP expressions of Eqs. (6) and (7) are valid for all cycles, they are also valid for Carnot cycles but can be further defined in terms of the cycle temperatures, T_H and T_L. The relevant expressions are

$$\eta_{\text{th}}=\frac{w_{\text{net}}}{q_H}=\frac{T_H-T_L}{T_H}=1-\frac{T_L}{T_H}, \tag{8a}$$

$$\text{COP}=\frac{q_L}{w_{\text{in}}}=\frac{T_L}{T_H-T_L}=\frac{1}{T_H/T_L-1}, \tag{8b}$$

$$(\text{COP})_{\text{HP}}=\frac{q_H}{w_{\text{in}}}=\frac{T_H}{T_H-T_L}=\frac{1}{1-T_L/T_H}. \tag{8c}$$

Some words of caution are in order. These temperature relationships are valid for Carnot cycles only, and the net work is *not* the temperature difference nor is the heat transfer equal to the temperature; only the temperature *ratios* are valid, and the temperatures must be absolute (thermodynamic) temperatures.

The cross-hatched area in Fig. 2 represents the net work of the system: the output work for a heat engine and the required work for a refrigerator. The cyclic performance is improved by increasing the area for a heat engine and minimizing the area for a refrigerator when the heat transfer of interest is specified.

In the discussion of the cycles to follow, the appropriate Carnot cycle will serve as a starting point in developing an ideal cycle as well as providing a criterion for evaluating its performance. Since a real cycle is irreversible, its cycle efficiency or COP must be less than that of the Carnot cycle; if it is equal to or greater than the Carnot efficiency or COP, the cycle is not possible.

2. VAPOR (POWER) CYCLES

The isothermal heat transfers of the Carnot cycle can be approached by isobaric phase changes of a pure substance, by adding heat to a vaporizing liquid and removing heat from a condensing liquid while holding the pressure constant, i.e., by operating in the two-phase region (SMR). Such a vapor–liquid cycle is commonly referred to as a *vapor cycle* and is the basis of the Rankine cycle in its various forms. Rankine cycles can be used to generate power and, in a reversed mode, to provide refrigeration. Vapor cycles are characterized by a phase change of the working medium, by operation in the saturated-mixture (two-phase) region (the SMR of Fig. 1), and at times in the CLR.

Since the Rankine cycle has been used extensively in the past in steam power plants, this discussion of power-generation cycles will be limited to steam (see STEAM ENGINE). However, other fluids are under consideration along with other energy-conversion systems using geothermal energy, ocean thermal gradients, and solar energy.

2.1 The Vapor Carnot Cycle

The simplest steam plant has four major components; a steam generator (boiler), a turbine, a condenser, and a pump—two heat exchangers and two work devices. The vapor Carnot cycle represented by this steam plant is sketched in Fig. 3. The cycle remains within the SMR and starts with saturated (liquid) water entering the steam generator at state 1. The water is heated reversibly, isothermally, and isobarically until all of the water has turned to dry steam (vapor) at state 2, which is at the edge of the saturated-mixture region (SMR). This dry steam is then expanded isentropically in the turbine to state 3, producing work (and partially condensing) during the process. At state 3 this wet steam is further condensed reversibly, isothermally, and isobarically to state 4, where the still wet steam enters a pump that isentropically raises the pressure and temperature to state 1 and the cycle is complete.

Although this Carnot cycle has the highest possible cycle efficiency for the specified boiler and condenser pressures (and corresponding temperatures), it also has several serious practical drawbacks. The first drawback deals with

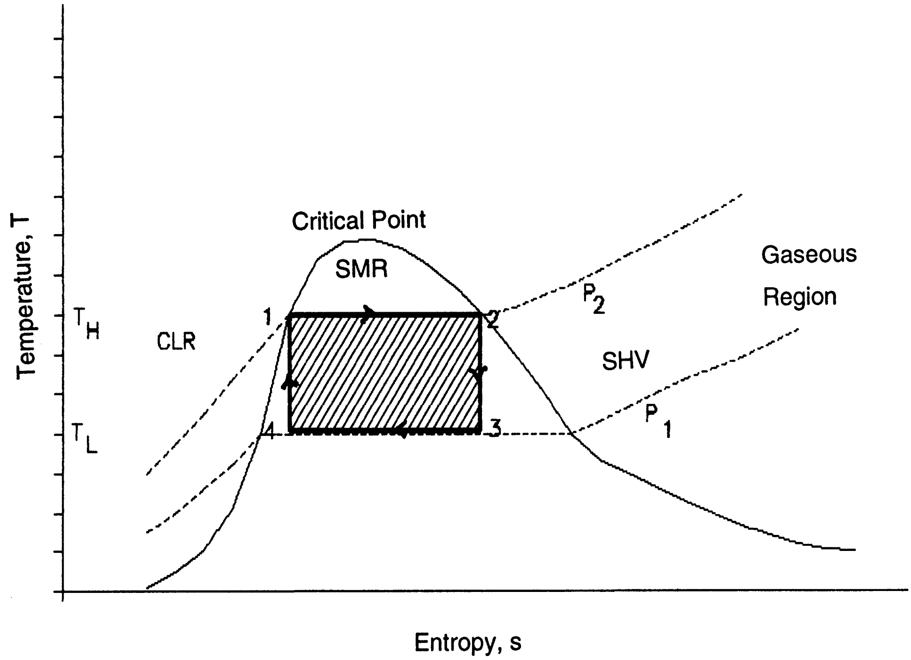

FIG. 3. The *T-s* diagram of a vapor (phase-change) Carnot cycle.

the turbine and pump, devices that work best with single-phase substances. In the turbine, for example, the drops of liquid water in the wet steam damage the blades and collide with the faster moving vapor particles for a loss in kinetic energy and a reduction of the turbine efficiency. In the pump, on the other hand, compressing a wet vapor to a liquid is not an efficient process, and the presence of the vapor leads to cavitation. Furthermore, in addition to the problem of stopping the condensation precisely at state 4, the net work per unit mass, as represented by the area 1–2–3–4–1, is small, requiring a large mass-flow rate and large power plant if the power output is to be of reasonable size.

2.2 The Rankine Cycle

The basic Rankine power cycle is the SMR Carnot cycle with modifications to correct its deficiencies. Referring to Fig. 4, where the Carnot-cycle states are indicated by primes, the first modification is to continue the condensation until all of the water vapor has been condensed and the water is a saturated liquid at state 4. The isentropic pump now compresses the liquid (with no vapor) to the boiler pressure at state 1. At this point heat transfer, which is isobaric and internally reversible but obviously not isothermal, commences. The isobaric addition of heat raises the temperature of the liquid water until the start of vaporization at state 1′, but heating does not stop at the saturated vapor line (state 2′); it continues isobarically into the superheated region to T_2, which normally is the highest allowable turbine inlet temperature. The additional heat, from 2′ to 2, is called ***superheat***, and its transfer is internally but not totally reversible since it is not isothermal; isothermal heat transfer in the superheated region (and in the CLR) requires a continuous change in the steam generator pressure, a difficult and impractical process. The superheated steam is then expanded isentropically in the turbine from state 2 to state 3, which is still in the SMR but is dryer than state 3′.

This additional heat, the superheat, and the higher average temperature accomplish three things: increase the net work, increase the thermal efficiency by increasing the average temperature at which heat is added, and reduce the amount of liquid water in the turbine and at state 3. Although the thermal efficiency with superheat can be greater than that of the

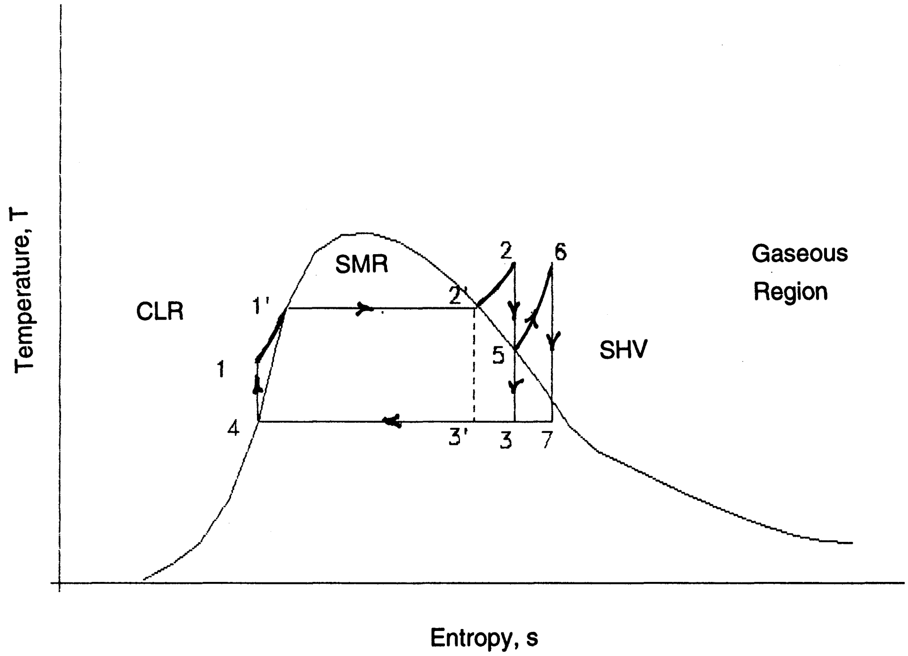

FIG. 4. The *T-s* diagram of an ideal Rankine power cycle.

Carnot cycle within the SMR, it is less than that of a Carnot cycle operating between T_2 and T_3, if such a cycle were possible.

The cycle efficiency with superheat can be evaluated from Eq. (6) with w_{net} and q_{in} determined from the appropriate use of Eq. (4). Neglecting the pump work ($h_4 \cong h_1$) and referring to Fig. 4 yield the expression

$$\eta_{\text{th}} = \frac{w_{\text{net}}}{q_{\text{in}}} \cong \frac{h_2 - h_3}{h_2 - h_1}. \tag{9}$$

Both the net work and thermal efficiency can be improved by increasing the temperature at which heat is added to the water and by decreasing the temperature at which the heat is removed in the condenser.

2.3 Reheat and Regeneration

The performance of the Rankine cycle can be further improved by adding what is known as *reheat* involving a two-stage turbine in which, as is shown in Fig. 4, the exhaust steam of the first (the high-pressure) stage is reheated isobarically but at a lower pressure from state 5 (usually still superheated) to state 6, the maximum turbine inlet temperature of state 2. This reheated steam is then expanded isentropically in the low-pressure stage to the condenser pressure at state 7, which usually is just within the SMR. This reheat cycle increases the net work per unit mass significantly, increases the thermal efficiency slightly, and reduces the moisture content of the turbine exhaust significantly. There may be more than one reheat cycle, particularly when the steam generator pressure is very high.

A second modification to the basic Rankine cycle is the addition of regeneration (not shown in Fig. 4) in which steam from the turbine is used to preheat the condensate before it reenters the steam generator. The simplest and commonly used device is the open *feedwater heater* in which steam extracted from the turbine is mixed directly with the condensate, whose pressure has been raised to that of the steam by an additional pump. The closed feedwater heater, on the other hand, resembles the more conventional heat exchanger in that the steam and condensate do not mix, eliminating the need for an additional condensate pump. A feedwater heater increases the thermal efficiency (reducing the fuel required for a given power output) and reduces the amount of coolant required for the condenser, both desirable results.

2.4 Other Vapor Cycles

Although reheat and regeneration have improved the thermal efficiency of steam power plants, the upper limit is restricted by the relatively low boiling temperature of water, even at high pressures, so that heat cannot be added at as high a temperature as desirable. In order to increase the temperature differential, the *binary vapor cycle* has been considered and even tested on a small scale using mercury in a Rankine cycle with a mercury-vapor generator, a mercury-vapor turbine, and a condenser (heat exchanger) that serves as the steam generator of a conventional steam cycle. The complications, costs, and complexities of using mercury (or other liquid metals such as potassium and rubidium) as well as the improvements in steam power plants have dampened interest in the binary vapor cycle.

There is, however, interest in using liquid metals in Rankine vapor cycles to produce auxiliary power in space. Since heat is rejected by radiation from the condenser, a high condenser temperature is needed in order to reduce radiator weight. This requirement in turn calls for supplying the vapor generator with higher-temperature heat from isotopes or solar panels.

At the other end of the power spectrum are Rankine vapor cycles in which the heat is supplied at relatively low temperatures from such "free" sources as geothermal energy, unfocused solar energy, ocean thermal energy, and waste heat. The resulting low thermal efficiencies require large mass-flow rates if the power output is to be significant. Since the specific volume of steam is large at the low pressures and temperatures of such systems, other fluids such as ammonia, isobutane, and possibly some of the Freons are more suitable than steam.

3. REFRIGERATION (AND HEAT-PUMP) CYCLES

Reversing the direction of the processes in a work-producing heat engine results in a refrigerator in which work is the input and a desired heat transfer is the output. Consider the two-phase Carnot cycle of Fig. 3, which when traveled in a clockwise direction (1–2–3–4–1) produces work with a heat input and a heat rejection. When traveled in a counterclockwise direction, however, this same cycle with a suitable working medium (the refrigerant) and a work input can produce two heat transfers. The cross-hatched area is the net work required whereas the area under the upper temperature line, 2–1, represents q_H, the heat transferred to the high-temperature reservoir, and q_L, the area under the lower-temperature line, 4–1, is the heat transferred from the low-temperature source. If the cooling of a space is the objective, q_L is to be maximized with respect to the net work input (the traditional refrigerator or air conditioner), whereas q_H is to be maximized if the heating of a space is the objective (the heat pump). The process from 3 to 2 is the isentropic compressor work, the work required to raise the pressure (and temperature) of the refrigerant from P_3 (T_L) to P_2 (T_H), and the process from 1 to 4 is the work produced by expanding the saturated liquid refrigerant through an appropriate device so as to reduce the pressure and temperature to that of the two-phase mixture at state 4.

This two-phase Carnot cycle provides the basis for the two major liquid-refrigerant refrigeration cycles, vapor-compression and absorption, as well as other lesser-known cycles such as the water-vapor, adsorption, and resorption refrigeration cycles. The vapor-compression cycle is favored for domestic and smaller-scale refrigeration and heating applications, whereas absorption refrigeration systems are better suited for larger, generally commercial, applications.

3.1 The Vapor-Compression Refrigeration Cycle

The vapor-compression refrigeration cycle, which is essentially a reversed Rankine power cycle without reheat or regeneration, is shown in Fig. 5. The two-phase refrigerant enters the evaporator at state 1, absorbs heat from the space to be cooled (q_L), and vaporizes until it reaches the saturated-vapor line (no liquid) at state 2 and enters a single-phase compressor. The vapor is compressed isentropically until it reaches the condenser pressure at state 3 and enters the condenser, where the heat of condensation (q_H) is transferred isobarically to the surroundings or to some other thermal

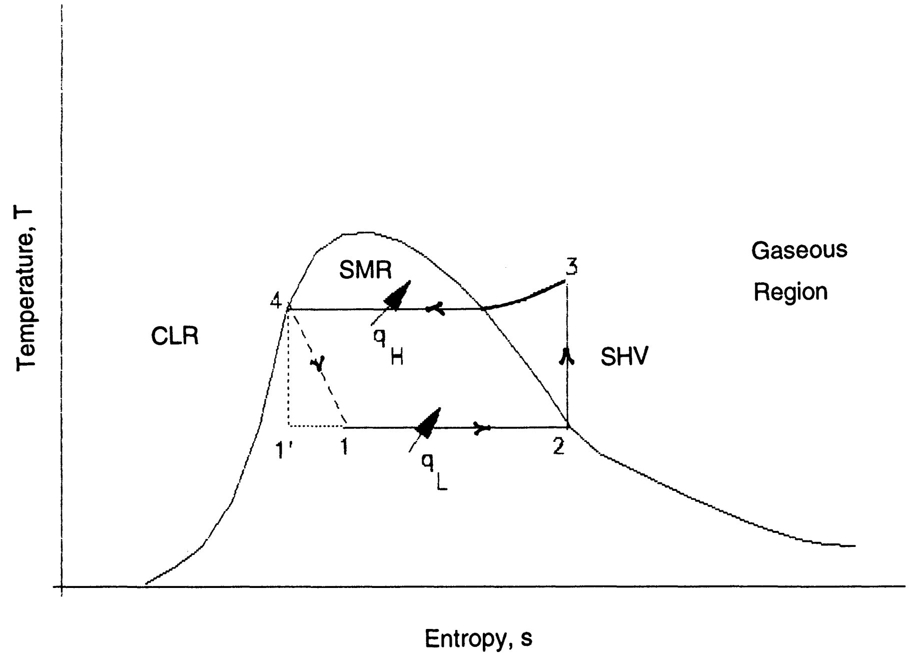

FIG. 5. The *T-s* diagram of an ideal vapor-compression refrigeration cycle.

reservoir. The liquid refrigerant of state 4 flows through an expansion valve (or capillary), changing phase and arriving back at state 1, the beginning of the cycle. The flow from 4 to 1 is a constant-enthalpy throttling process that is irreversible, as indicated by the dashed line. Although such a throttling device is lossy and has no work output, it has the advantages of simplicity and low cost when compared with an isentropic expander such as a two-phase turbine or a piston–cylinder combination whose work output would be small with respect to the required compressor work.

Referring to Fig. 5 and making use of Eqs. (7a) and (4) and the fact that $h_1 = h_4$, the COP for an ideal vapor-compression refrigerator can be written as

$$\mathrm{COP} = \frac{q_L}{w_{\mathrm{in}}} = \frac{h_1 - h_2}{h_2 - h_3} = \frac{h_4 - h_2}{h_2 - h_3}. \tag{10}$$

The cycles for space cooling and heating are identical; only the COP's and locations of the condenser and evaporator are different. For cooling (a refrigerator or air conditioner), the evaporator is located in the space to be cooled and the condenser usually discharges q_H into the atmosphere. This heat could be utilized, however, rather than be wasted—for example, to provide domestic hot water by using water rather than air in the condenser.

For heating (as a heat pump) the condenser is so located that q_H is transferred into the space to be heated and the evaporator obtains q_L from the surroundings, which might be ambient air, air from underground pipes, water from various sources, such as a well, a pond, or a swimming pool, or any other convenient thermal reservoir whose temperature is higher than that of the refrigerant, such as a solar collector or waste heat from another process.

Since heat pumps and air conditioners have the same components and operate in the same cycle, it is a common practice to use a heat pump as an air conditioner by adding a reversing valve to the cycle so that the condenser of the heat pump becomes the evaporator of the air conditioner and the evaporator of the heat pump becomes the condenser of the air conditioner. Sizing the heat-pump–air-conditioner combination to satisfy the cooling load sets the COP for the heating load and thus the temperature at which any supplementary heating will be required to maintain a desired space temperature.

The simple vapor-compression refrigerator cycle may be modified to improve its effectiveness and efficiency for large commercial applications. In *cascade refrigeration* systems two or more refrigeration systems are connected in series by a heat exchanger that contains the condenser of one system and the evaporator of the other; since there is no mixing in these cascaded systems, different refrigerants may be used in each stage. A second modification is the *multistage compression* system, which is in essence a cascaded system using a single refrigerant in which the heat exchanger connecting the stages is replaced by a mixing (flash) chamber in which part of the liquid refrigerant vaporizes.

With respect to refrigerants for vapor-compression cycles, any fluid that can be made to evaporate at pressures between 1 and 20 atm at reasonable temperatures is a possible candidate. However, there are other considerations that limit the choices, such as toxicity, corrosiveness, and environmental effects.

3.2 The Absorption Refrigeration Cycle

The *absorption refrigeration cycle* (not shown) is a variation of the vapor-compression refrigeration cycle in which the compressor is replaced by an absorption process in which a refrigerant vapor (such as ammonia) is absorbed in a liquid medium (such as water) and pumped as a liquid (sometimes referred to as a strong liquor) into a heat exchanger (called a generator), where heat (q_{gen}) is applied to vaporize most of the refrigerant along with some of the absorbent. This high-pressure vapor then passes through the condenser, the expansion valve, and the evaporator, as before, and returns to the absorber as a weak liquor, and the cycle is completed. Since the absorption process is exothermic, a coolant, such as water, is required to control the absorber temperature.

Since the specific volume of the liquid mixture to be pumped is a very small fraction of the specific volume of the refrigerant vapor, the pump work is also very small in comparison with that required for a vapor compressor, so small that it is often neglected in the cycle analysis. However, the amount of external heat required by the generator is quite large, and the COP is lower than that of the vapor-compression refrigerator. The COP can be improved somewhat, with added cost and complexity, by such techniques as adding a regenerator and an expansion valve to the absorption subsystem. In spite of a low COP, such absorption systems have certain operational advantages and can be very cost effective when a large source of inexpensive heat is available, such as waste heat from a power station or a process steam plant, geothermal energy, or solar energy.

Absorption refrigeration on a small scale is also attractive when electrical power is either limited or expensive as in isolated regions and campsites. If no electrical power is available, the pump can be replaced by the addition of hydrogen to maintain the total system pressure; this Electrolux absorption system, as it is known, has no moving parts or valves.

3.3 Other Refrigeration Cycles

There is a variation of the absorption refrigeration cycle known as the *resorption refrigeration cycle*, in which the condenser is replaced by a *resorber* where the vapor is not condensed to a pure liquid but is absorbed by a special weak solution while condensing. As the vapor is condensed and absorbed in the resorber, heat is transferred to and removed by the cooling water, and the solution flows through an expansion valve to the evaporator. As the refrigerant (ammonia) boils in the evaporator, the heats of vaporization and dissociation both contribute to the refrigerating effect. Greater cooling can be achieved than in the usual absorption system at the cost of an extra pump and some piping. Combination absorption and resorption cycles may be suitable for special applications.

Although chilled water can be provided from an evaporator heat exchanger, the *water-vapor refrigeration cycle* is used to provide large quantities of chilled water for cooling. The key element is an evaporator whose pressure has been lowered, by a vacuum pump, to the saturation pressure corresponding to the desired temperature of the chilled water. When the intake water is sprayed into the evaporator, the low pressure causes a portion of this water to flash into vapor, reducing the temperature of the resulting two-phase mixture to the saturation temperature for the evaporator pressure. The chilled water is easily drawn off from the bottom of the evaporator and pumped to its destination.

Removal and condensation of the vapor, however, are much more of a problem because of its large volume and high volume flow rate; the latter can easily be of the order of thousands of cubic meters per minute, and velocities can become supersonic if the cross-sectional area of the pipes is not large enough. Such large flow rates require large axial flow or centrifugal compressors (and large amounts of power) to move the vapor through the condenser. When steam at moderate pressures and low cost is available, a *steam jet compressor* may be used to provide a straightforward and low-cost alternative, simple to operate and maintain, to rotating compressors and blowers.

Even though the concept and system elements of an *adsorption refrigeration* system are similar to those of an absorption refrigeration system, adsorption refrigeration is not as well known and is still in the study and developmental phases and not in commercial use. Whereas in the absorption refrigeration cycle the refrigerant and absorbent form a liquid solution that can be pumped and from which the refrigerant can be vaporized, in the adsorption refrigeration cycle the refrigerant is adsorbed on the surface of a solid; otherwise the ideal cycles are identical.

4. GAS CYCLES

Although the vapor cycles involve phase changes and two-phase mixtures of the working fluid, the working fluid in a gas cycle remains in the gaseous phase throughout the cycle; there is no condensation or evaporation. Furthermore, many of the gas cycles are *open cycles* in that some or even all of the working fluid leaves the system and enters the surroundings during the cycle to be replaced by additional fluid entering the system at the same original state as the departing fluid. (The vapor cycles discussed in the previous sections were, with the possible exception of the water-vapor refrigerator, closed cycles.)

Although some of the vapor power cycles are reversed to operate as refrigerators (and heat pumps), the gas cycles are predominantly heat engines (work producers). An additional measure of performance for the piston–cylinder (reciprocating) engines is the *mean effective pressure* (mep), which is defined as the net cycle work divided by the volume change during one stroke; it can be an indicator of the size of an engine.

4.1 The Carnot Cycle

The *T-s* diagram for a gas-cycle engine operating in a Carnot cycle is shown in Fig. 6. Although this cycle has the highest thermal efficiency possible, it also has practical limitations that restrict its use. First of all, the isothermal heat transfers (from 1 to 2 and 3 to 4) are not isobaric so that some means of varying the pressure appropriately is required if the heat transfers are to be totally reversible; as mentioned earlier, maintaining an isothermal heat transfer without a phase change is not a simple task. Second, the net work per unit mass is small and the volume change is high. Consequently, the mep is quite low—so low that a Carnot engine might have trouble overcoming the starting friction.

Although the Carnot cycle is still a standard for performance evaluation, these two limitations are sufficient to preclude using the Carnot cycle as the basis for the ideal cycles to be used in the analyses of internal combustion engines, analyses that also assume air-standard cycles. External combustion devices with regenerators, however, such as those using the Stirling or Ericsson cycles, can approximate isothermal heat transfers and can have efficiencies approaching that of the Carnot cycle.

4.2 The Air-Standard Cycle

Many gas-cycle engines are classified as *internal combustion engines* (see AUTOMOTIVE PROPULSION SYSTEMS) because they obtain their thermal energy by compressing and then burning a mixture of fuel and air within the system boundaries. Although all of the combustion products comprise the working fluid, the predominant component is air and it is convenient to assume that the working fluid is dry air. Other convenient assumptions are that the air is an ideal gas with constant specific heats, that any combustion may be replaced by an equivalent heat transfer from an external thermal reservoir, and that any open cycles are thermodynamically closed by the transfer of heat to the surroundings. In addition, all processes are assumed to be

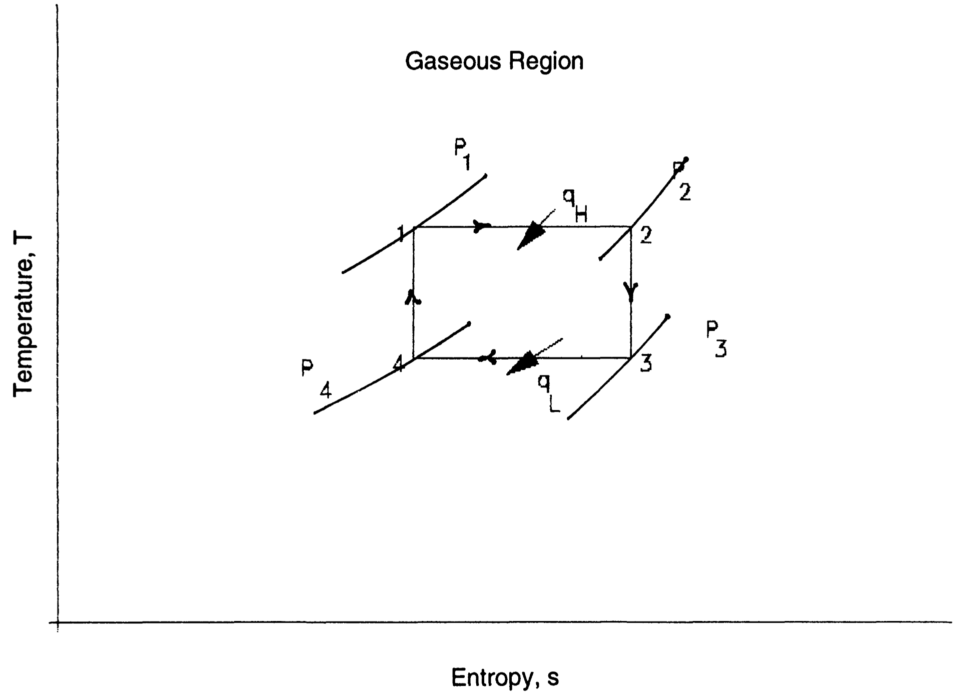

FIG. 6. The *T-s* diagram of a gas Carnot cycle (no phase change). The arcs labeled P_n are portions of isobars.

internally reversible. This collection of assumptions constitutes the *air-standard cycle,* which is incorporated in the analyses of the various ideal gas cycles that follow.

Since the specific heats are assumed to be constant, this set of assumptions may be thought of as the *cold-air-standard cycle* to distinguish it from the *hot-air-standard cycle* in which the variations in the specific heats that occur within an engine are taken into account.

4.2.1 The Otto Cycle. The air-standard *Otto cycle* is used to predict the performance of a *spark-ignition* (SI) gasoline piston–cylinder (reciprocating) engine. Referring to the *T-s* diagram of Fig. 7, the fuel–air mixture enters the cylinder at state 1 (with the piston at bottom dead center) and is compressed isentropically to state 2 (with the piston at top dead center) where ignition and combustion take place. With the assumption of insignificant piston movement during combustion, heat is added at constant volume from states 2 to 3 followed by an isentropic expansion to state 4, where the piston is again at bottom dead center. The process from states 4 to 1 is a constant-volume rejection of heat through the exhaust of the combustion products to the surroundings.

The piston–cylinder combination is a closed system, and the Otto-cycle efficiency can be expressed as

$$\eta_{\text{th}} = 1 - \frac{T_1}{T_2} = 1 - \frac{1}{r^{k-1}}, \qquad (11)$$

where k is the ratio of specific heats (1.4 for air) and r, the *compression ratio,* is the ratio of the cylinder volume before compression to the volume after compression (V_1/V_2); r is a volume ratio, *not* a pressure ratio.

Equation (11) shows that the efficiency of the ideal air-standard Otto cycle depends only on the compression ratio and the value of k and is independent of such factors as the heat added per cycle and the conditions at the start of compression. In addition to a reasonably high efficiency, this cycle has a high mep, i.e., a high net work output for a given displacement volume.

4.2.2 The Diesel Cycle. In *compression-ignition* (CI) engines, the intake air is so highly compressed that the temperature of the air is sufficiently high for self-ignition to occur

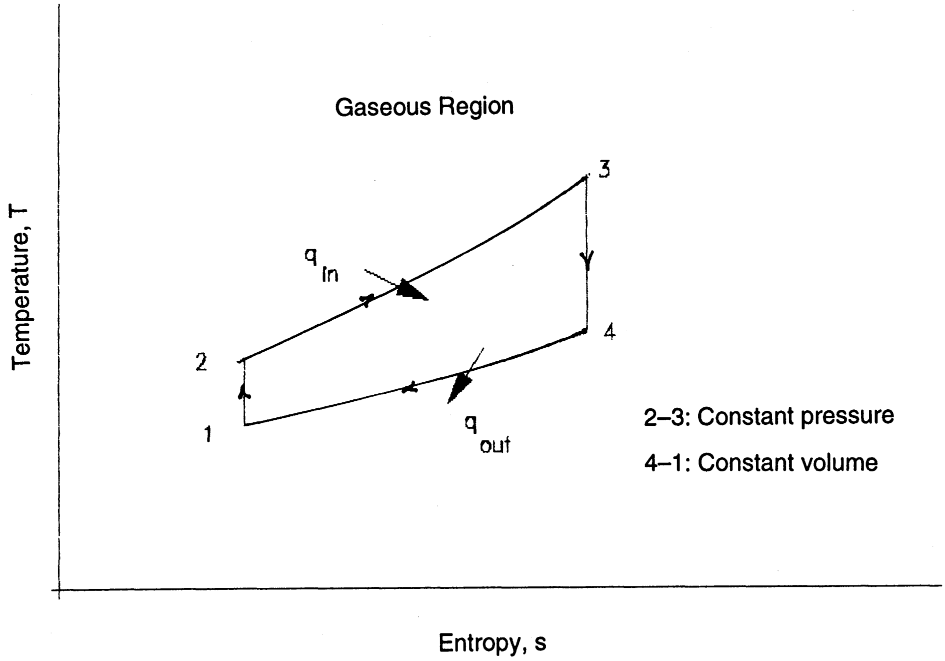

FIG. 7. The *T-s* diagram of an ideal Otto cycle.

when fuel is injected into the cylinder. The air-standard *Diesel cycle* is the model for low-speed (and usually large) compression-ignition engines. Referring to the *T-s* diagram of Fig. 8, the intake air is compressed isentropically from state 1 to 2, at which point fuel is introduced and self-ignites. Since the duration of combustion is longer than that for the Otto cycle, constant pressure is a more appropriate assumption; thus the heat is added isobarically in moving from state 2 to 3. Isentropic expansion from 3 to 4 (the power stroke) is followed by constant-volume rejection from state 4 back to 1, completing the cycle.

Notice the similarity to the Otto cycle; the only difference is the heat input process: constant volume in the Otto cycle and constant pressure in the Diesel cycle. The procedure for finding expressions for the thermal efficiency is similar to that used with the Otto cycle with the exception that q_{in} now involves moving-boundary work because of the volume change. At this point it is convenient to define two new parameters: the *expansion ratio*, $r_e = V_3/V_4$, and the *cutoff ratio*, $r_c = V_3/V_2$. Note that r_c, which is always greater than unity for the Diesel cycle, is also equal to the ratio of the compression ratio to the expansion ratio: $r_c = r/r_e$. (Since the expansion ratio of the Otto cycle is equal to the compression ratio, the cutoff ratio for the Otto cycle is unity.)

The Diesel-cycle efficiency expression can be written as

$$\eta_{\text{th}} = 1 - \frac{1}{r^{k-1}}\left[\frac{r_c^k - 1}{k(r_c - 1)}\right], \tag{12}$$

showing the efficiency to be a function of both the compression ratio and the cutoff ratio as well as of the ratio of specific heats. A comparison of Eq. (12) with Eq. (11) shows that only the bracketed term differentiates the efficiency of the Diesel cycle from that of the Otto cycle. With $r_c > 1$, this term is also greater than unity so that with identical compression ratios the thermal efficiency of the Otto cycle will always be greater than that of the Diesel cycle; increasing the cutoff ratio further decreases the Diesel efficiency. Note that when $r_c = 1$, the Otto-cycle value, the bracketed term becomes unity and the two efficiencies are identical.

As the rpm of a diesel engine is increased to increase the horsepower-to-weight ratio, the

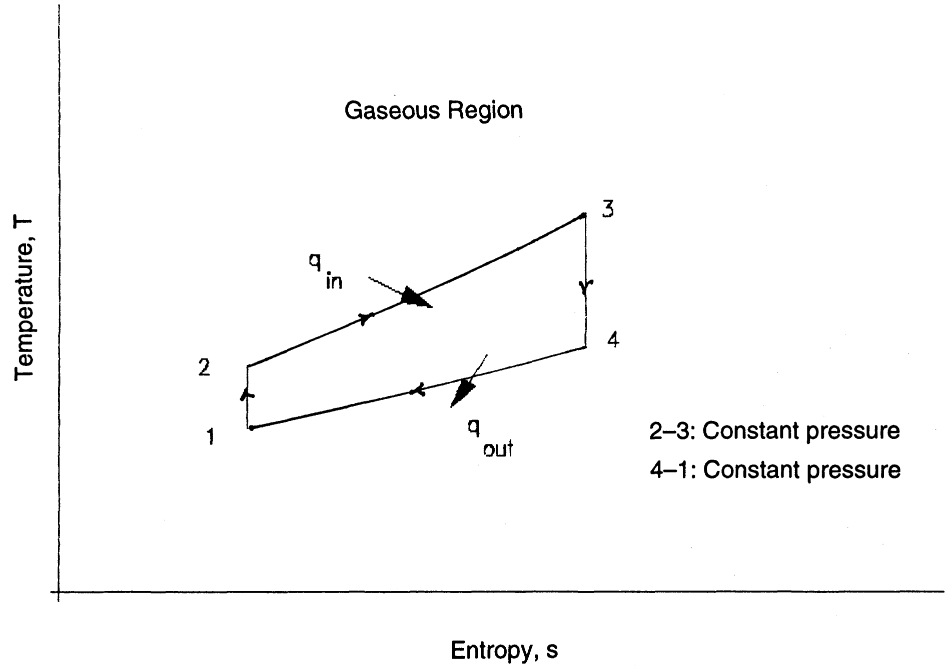

FIG. 8. The *T-s* diagram of an ideal Diesel cycle.

validity of the approximation of a constant-pressure heat input process becomes less valid. A more realistic representation is one in which heat is added in two steps, a constant-volume addition followed by a constant-pressure addition; the corresponding cycle is known as an *ideal dual cycle.* As the rpm increases, the constant-volume input increases and the constant-pressure input decreases so that the small high-speed compression-ignition engines used in automobiles today, although called diesel engines, approach the Otto cycle in their performance. This dual cycle is applicable to modern SI engines as well as CI engines, and the relative amounts of heat transferred during each process can be adjusted to approximate better the actual cycle. The ideal Otto and Diesel cycles are special cases of the dual cycle.

4.2.3 The Brayton Cycle. The air-standard *Brayton cycle,* also known as the *Joule cycle,* is the ideal cycle of the simple gas-turbine cycle; it can be an open cycle with internal combustion as the heat source and with heat rejection to the surroundings, or a closed cycle, utilizing an external heat source and transferring heat with heat exchangers. Either type of cycle can be represented by the *T-s* diagram of Fig. 9, where an isentropic compression from 1 to 2 is followed by an isobaric heat addition (in the combustion chamber for an open cycle) to state 3, which for best performance is the maximum allowable turbine inlet temperature. Process 3–4 is an isentropic expansion (through the turbine), and 4–1 is an isobaric heat rejection (to the surroundings for an open cycle).

Whereas the processes of the Otto and Diesel cycles represent closed systems, those of the Brayton cycle are open (single-flow) systems, and the cycle efficiency can be written as

$$\eta_{\text{th}} = 1 - \frac{T_1}{T_2} = 1 - \frac{1}{(P_2/P_1)^{(k-1)/k}}, \tag{13}$$

where P_2/P_1, the *pressure ratio* during compression (and expansion), is the parameter used to determine the efficiency of a Brayton-cycle engine; it should not be confused with the compression ratio r (V_1/V_2) used with the Otto and Diesel cycles. Note that the pressure ratio determines the cycle efficiency, not the maximum cycle temperature T_3.

Not only is the efficiency of a typical Brayton-cycle engine less than that of a comparable Otto-cycle engine, but the mep is also

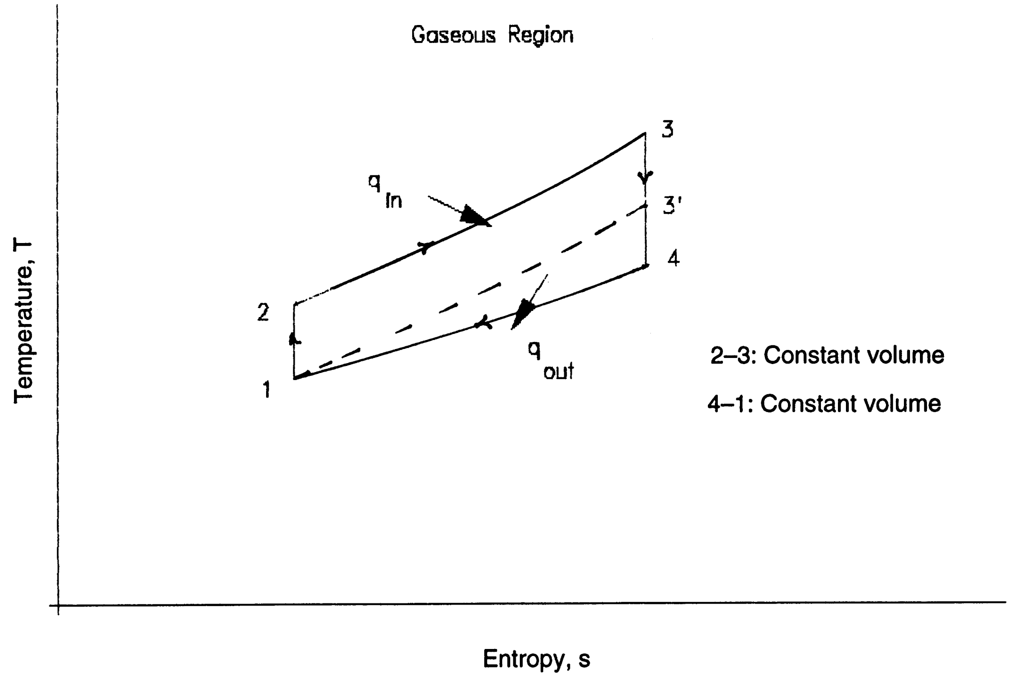

FIG. 9. The *T-s* diagram of an ideal Brayton cycle.

considerably lower, requiring the movement of large volumes of gases to obtain a useful amount of net work. The volume and weight of reciprocating compressors and expanders associated with a closed Brayton cycle are so large as to be impractical. Rotating machinery, on the other hand, can handle the large volumes of gases in a small volume of machinery so well that the open Brayton cycle has become the model for the various forms of gas-turbine engines, engines that have higher power-to-weight ratios than comparable SI and CI engines.

A large percentage of the turbine work output is required to drive the compressor (the *back work*). Whereas only 1% or 2% of the turbine work is required to drive the pump of a Rankine cycle, the compressor of a Brayton cycle can require from 40% to even 80% of the turbine output. The startling difference in the back-work ratios arises because the specific volume of the liquid being pumped in the Rankine cycle is very low compared with that of the gas being compressed in the gas turbine cycle.

Referring to Fig. 9, state 3′ represents the amount of the expansion in the turbine required to drive the compressor so that the area 1–2–3–3′–1 represents the compressor work and the remaining area, 1–3′–4–1, represents the net work, the available output work. Note that while an increase in T_3, the maximum turbine inlet temperature, does not improve the efficiency, it does increase the net work and reduce the back-work ratio, both desirable, since the divergence between the isobars increases with increasing temperature.

The energy remaining in the turbine at state 3′ is available to be used in various ways. It may be completely expanded in a nozzle to produce thrust (a *turbojet*), partially expanded in another turbine to drive a propeller (a *turboprop*) or a multibladed fan (a *turbofan*) with the residual energy transformed into jet thrust in a nozzle, or completely expanded in another turbine (a *turboshaft engine*) to drive a pump, an electrical generator, a helicopter rotor, a marine propeller, or a tank track.

In a closed Brayton cycle, heat is received and rejected via heat exchangers so that the working fluid is not limited to air (combustion products). One possible application of the closed cycle might be in a nuclear power plant using a gas with a higher C_p, such as helium.

As with the Rankine cycle, there are add-ons to improve the performance of the open gas-turbine cycle. A *regenerator* can be used when the temperature of the exhaust gases is higher than the temperature of the air leaving the compressor ($T_4 > T_2$). The exhaust gases enter a counterflow heat exchanger to preheat the air leaving the compressor before it enters the combustor. If the regenerator were ideal, the temperature entering the combustor would be T_4, and the ideal thermal efficiency with a regenerator is

$$\eta_{\mathrm{th}} = 1 - \frac{w_c}{w_t} = 1 - \left(\frac{T_1}{T_3}\right)\left(\frac{P_2}{P_1}\right)^{(k-1)/k} . \qquad (14)$$

Equation (14) shows that the thermal efficiency with regeneration depends not only upon the pressure ratio but also upon the ratio of the minimum to the maximum temperatures, and that the efficiency decreases with an increasing pressure ratio to the point where regeneration is no longer effective.

The efficiency of the Brayton cycle, either open or closed, can be further improved by the addition, in conjunction with a regenerator, of *multistage compression with intercooling between stages* and *multistage expansion with reheat between stages.* Although there is no theoretical limit to the number of stages, the economic and practical limit in view of the declining payoff and increasing complexity, cost, and losses is usually two or maybe three stages. It is interesting to note that with an unlimited number of stages, the thermal efficiency would approach that of the Ericsson cycle (discussed in Sec. 4.4) and therefore that of the Carnot cycle.

4.3 The Gas Refrigeration Cycle

The ideal *gas refrigeration cycle* is essentially a reversed Brayton cycle and may be referred to as *air-cycle refrigeration* or as the *air-standard refrigeration cycle* when air is used as the refrigerant. The gas cycle differs from the vapor-compression refrigeration cycles in that the condensation and evaporation of a liquid refrigerant are replaced by the sensible cooling and heating of a gas. In addition, the irreversible expansion valve is replaced by an isentropic expansion engine or turbine so as to achieve a greater drop in the gas temperature.

The air-standard refrigeration cycle, originally known as air-cycle refrigeration, was one of the earliest forms of cooling and was operated in either an open or closed (dense-air) cycle. Although air is a safe and inexpensive refrigerant, air-cycle refrigeration became obsolete for general use because of its low COP and high operating costs; it could not compete with vapor-compression refrigeration with its higher-performance refrigerants. It does, however, still have two specialized applications: to provide open-cycle cooling for pressurized aircraft and, in a closed cycle with regeneration, to liquefy air and other gases.

Referring to the ideal Brayton cycle of Fig. 9, in a closed cycle air enters the compressor at state 4 and is isentropically compressed to state 3. After being cooled isobarically from T_3 to T_2, the air expands in an isentropic process (doing work in the process) to the compressor inlet pressure at state 1. Heat is then transferred isobarically from the space to be cooled to raise the temperature of the air to T_4 at the compressor inlet. The amount of work done by the expander (the turbine) is significant and can be used, for example, to drive a circulating fan or to provide a portion of the compressor work if appropriate.

The COP of the ideal gas refrigeration cycle is related to the compressor pressure ratio by the expression

$$\mathrm{COP} = \frac{1}{(P_3/P_4)^{(k-1)/k} - 1} , \qquad (15)$$

where $P_3/P_4 = P_2/P_1$.

With regeneration, gas-cycle refrigeration can be used for the *liquefaction* of gases. With a simple gas cycle (without regeneration) the lowest possible turbine inlet temperature T_2 is the temperature of the cooling medium, the ambient temperature when ambient air is the coolant. With one or more regenerators [counterflow heat exchangers] between the compressor and the turbine, however, the inlet temperature can be lowered further so that the turbine exit temperature is below the critical temperature of the air (−140 °C) and the air enters the liquid phase. Since the critical temperatures of the major components of air (nitrogen and oxygen) are different (−147 and −119 °C, respectively), they can be removed separately from the liquid air by fractional distillation, as can the other components such as argon, helium, and neon.

The regenerative gas cycle can be applied to the liquefaction of other gases, such as helium, and to the solidification of carbon dioxide (dry ice). There are other techniques with varying configurations and complexities for the liquefaction of gases, such as the Linde, Claude, Collier, and Heylandt processes, and they all involve compression of a gas to a high temperature, successive cooling with regeneration, and then expansion through a turbine, an expansion valve, or a reciprocating expander.

4.4 The Stirling and Ericsson Cycles

The Stirling cycle (see STIRLING ENGINE) and the Ericsson cycle are ideal-gas cycles with external heat sources that make use of a regenerator to accomplish isothermal heat transfers into and out of the working fluid. Since the regenerator is an essential element of these closed cycles, they are often referred to as the regenerative cycles. The regenerator in these cycles is a thermal energy-storage device that has the capability of storing and releasing thermal energy on demand. It might, for example, be a porous plug with a high thermal mass through which the gas flows, or a counterflow heat exchanger. The porous plug is appropriate for intermittent devices, such as reciprocating expanders and compressors, whereas the heat exchanger is appropriate for continuous-flow devices, such as turbines and rotary compressors.

Consider Fig. 10, the *T*-*s* diagram of a Stirling cycle. Starting with the gas at T_H at state 1, heat from an external source is supplied during the process 1–2 to an isothermal expander (a turbine or a piston–cylinder combination); work must be done if the temperature is to remain constant. Then while the volume is kept constant, the hot gas is forced through the regenerator, transferring heat to the regenerator and cooling the gas to T_L at state 3. The cooled gas now enters an isothermal compressor, rejecting heat to an external sink to state 4, where $V_4 = V_1$. The gas is passed through the regenerator again in another constant-volume process (from 2 to 3), this time picking up the heat stored in the regenerator and raising the temperature back to T_H at state 1.

It can be shown that the heat transferred from the gas to the regenerator (the area

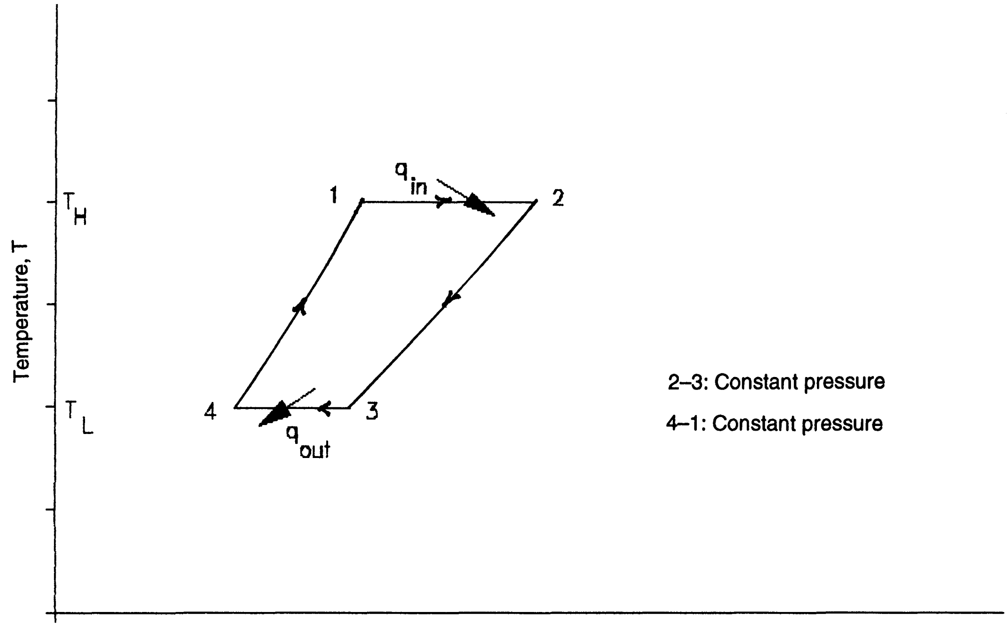

FIG. 10. The *T*-*s* diagram of an ideal Stirling cycle.

under the line from 2 to 3) is exactly equal to the heat transferred from the regenerator to the gas (the area under the line from 4 to 1). Therefore the transfer of heat from the external source to the gas takes place in the isothermal expander at T_H, and all of the heat rejected to the sink takes place in the isothermal compressor at T_L. Since these isothermal heat transfers are the conditions for a Carnot cycle, the efficiency of the ideal Stirling cycle will be that of a Carnot cycle operating between the same temperatures, namely

$$\eta_{\text{th}} = 1 - T_L/T_H. \tag{16}$$

The ideal Ericsson cycle is similar to the Stirling cycle, as is its *T-s* diagram, except that the heat exchanges with the regenerator take place at constant pressure rather than at constant volume. Since heat is transferred isothermally between the gas and the external heat source and the external sink, as in the Stirling cycle, the ideal thermal efficiencies are equal, namely, that of a Carnot cycle. These cycles provide an excellent example of the beneficial effects of regeneration upon thermal efficiency, as was seen with the Rankine cycle and with the Brayton cycle with multistage compression and expansion and intercooling, reheat, and regeneration.

The Stirling and Ericsson cycles share with the Carnot (and the Brayton) cycle the disadvantage of low mep's that lead to bulky reciprocating engines, which was the case with early *hot-air engines*. As with the Brayton cycle, the total work can be increased in spite of the low mep's by the use of rotating turbines and compressors.

The actual thermal efficiencies of these cycles are much less than the ideal values for several reasons. First of all, isothermal heat transfer is not practical as, without a temperature differential, it would require infinitely large surface areas or infinite transfer times. Furthermore, all four processes involve heat transfer through gas-surface films with large thermal resistances that lower the mean cycle temperatures and thus the efficiency. There are, as well, pressure drops in the components and the problems associated with achieving isothermal expansion and compression.

In spite of these drawbacks, there are advantages associated with external combustion and a closed cycle that have led to a renewed interest in Stirling and Ericsson cycles. With external combustion, any source of heat (combustion, solar, isotopes, nuclear, etc.) can be used. In the case of combustion, there is a greater choice of fuels, and the duration of the combustion process can be longer, leading to more complete combustion with greater energy output and less pollution. The closed cycle offers the opportunity to select a working fluid with the most suitable characteristics for the application of interest (in addition to air, gases such as hydrogen, nitrogen, and helium are suitable candidates) and does not require a reservoir of the working fluid to replace that exhausted in an open cycle.

These cycles can be reversed to be used as refrigerators; Stirling refrigerators, for example, are already being used in laboratories and in small-scale commercial applications to provide cooling, to liquefy air, and to recondense evaporated cryogens. Stirling and Ericsson engines and refrigerators can be combined to form heat pumps, such as a Stirling–Stirling heat pump. The *Vuilleumier cycle* is a related cycle that combines a Stirling engine and an Stirling refrigerator in a single device that cools and heats, and is characterized by three operating temperatures.

5. COGENERATION AND ENVIRONMENTAL CONSIDERATIONS

In a heat-engine cycle the major portion of the input heat is not transformed into work and is transferred out of the working fluid and out of the engine. If the residual energy in this rejected (waste) heat is not utilized in another heat engine or in some other manner but is merely rejected to the surroundings, it is truly wasted. Not only is this rejected energy wasted, but it also pollutes the environment, thermally if the cycle is closed and both thermally and with combustion products if the cycle is open and internal combustion provides the input heat. It should be noted that the external heat source of a closed cycle is also a source of environmental pollution, the type of pollution depending on the nature of the source, combustion or nuclear.

An ideal solution to the waste heat problem would be to extract all of the available energy from the input heat so that there would be no residual energy to affect the environment. Since this an impossibility under the second law, a realistic agenda calls for the maximum practicable reduction in the energy of the rejected heat and for the utilization of the lower-quality residual thermal energy in a meaningful and economic manner. If these two objectives can be met, then the rejected heat is no longer waste heat.

If the term "waste heat" is retained to identify the heat rejected from a cycle, waste heat can be categorized as *high-quality waste heat* when its temperature exceeds 93 °C (200 °F) and as *low-quality waste heat* when its temperature is less than 93 °C. If the temperature of the high-quality waste heat is sufficiently high, it can be used directly to provide all or a portion of the input heat for lower-temperature systems, to provide steam or power; if not, a heat pump can be used to raise the temperature to a useful level to generate steam or high-temperature liquids. The low-quality waste heat, particularly that represented by condenser cooling water, can be used for agricultural (greenhouses and soil warming), aquacultural (fish farming), and other similar purposes.

Before defining and discussing cogeneration, it should be noted that improving the combustion processes, both internal and external, so as to increase the energy extracted from the fuels not only improves the efficiency of the combustion process and the economics of the heat engine but also, by reducing the temperature of the combustion products, reduces the environmental pollution.

The major industrial needs are for heat (usually in the form of steam) for processes and heating, and for electrical and mechanical power. Consequently, *cogeneration* is generally defined as the use of a single integrated plant to generate both process steam and electrical power; such a cogeneration plant may also be described as a combined heat and power (CHP) plant. Actually, any multipurpose plant whose objectives are to increase the extraction of energy from the input heat and to reduce the amount of waste heat is a cogeneration plant. For example, a steam plant that provides steam for *district heating* (community space heating and domestic hot water), for an absorption refrigeration system, and to drive a turbine to generate electrical power to operate the plant is a cogeneration plant. Although the electrical power production in this particular plant, and in many like it, may be limited to that needed for plant operation, there is a growing trend for cogeneration plants to produce more electrical power than needed for plant operation and to use the excess or sell it to the public utilities when the utilities are willing to buy it.

Depending on how cogeneration is integrated into a system, particularly a CHP plant, the cogeneration technologies may be grouped as topping cycles and as bottoming cycles. In a *topping cycle*, the high-temperature input energy is first used for power generation and then the lower-temperature residual energy is used to produce process steam. In a *bottoming cycle*, the sequence is reversed; the high-temperature energy is used first to produce steam with the remaining energy used to produce power. Most currently available cogeneration technologies are not new and have been used in one form or another for many years.

There are many cogeneration configurations involving various combinations of steam and gas turbines and even large diesel engines with electrical generators and heat exchangers that maximize the extraction of energy and minimize the effect on the environment.

GLOSSARY

Adiabatic: Without heat transfer (thermally insulated).

Carnot Cycle: Ideal cycle with maximum efficiency or COP.

COP (Coefficient of Performance): Measure of the thermal performance of refrigerators and heat pumps.

Cycle: Series of processes with identical initial and end states.

Enthalpy: A derived property; sum of internal energy and the product of pressure and volume.

Entropy: A derived property related to the second law, and a measure of losses.

First Law: Thermodynamic statement of the principle of conservation of energy.

Heat Engine: Cyclic system with only heat and work crossing its boundaries. Often used to describe a system that receives and rejects heat and produces work.

Ideal Gas: A gas whose P-v-T relationship is determined by the ideal-gas equation of state, $Pv=RT$.

Isentropic: Adiabatic and reversible; taking place at constant entropy.

Isothermal: Taking place at constant temperature.

Process: Change in one or more properties.

Refrigerator: Heat engine that uses a work input to transfer heat from a lower to a higher temperature so as to cool or heat a space.

Reversible: Ideal with no losses.

Second Law: Complex set of relationships and equations that deal with the quality and direction of a process.

Further Reading

Baumeister, T., *et al.* (Eds.) (1978), *Marks' Standard Handbook for Mechanical Engineers,* 8th ed., New York: McGraw-Hill.

Doolittle, J. S., Hale, F. J. (1983), *Thermodynamics for Engineers,* New York: Wiley.

Granet, I. (1990), *Thermodynamics and Heat Power,* 4th ed., Englewood Cliffs, NJ: Prentice-Hall.

Haywood, R. W. (1991), *Analysis of Engineering Cycles,* 4th ed., Oxford: Pergamon.

Hatsopoulos, G. N., Keenan, J. H. (1981), *Principles of General Thermodynamics,* New York; Krieger.

Hu, S. D. (1985), *Cogeneration,* Reston, VA: Reston.

Jordan, R. C., Priester, G. B. (1956), *Refrigeration and Air Conditioning,* Englewood Cliffs, NJ: Prentice-Hall.

Reay, D. A., Macmichael, D. B. A. (1988), *Heat Pumps,* 2nd ed., Oxford: Pergamon.

Saito, T. (Ed.) (1990), *Heat Pumps, Solving Energy and Environmental Challenges,* Oxford: Pergamon.

Sengupta, S., Lee, S. S. (Eds.) (1983), *Waste Heat Utilization and Management,* Washington, DC: Hemisphere.

Van Wylen, G. J., Sonntag, R. E. (1985), *Fundamentals of Classical Thermodynamics,* 3rd ed., New York: Wiley.

Wurm, J., *et al.* (1991), *Stirling and Vuilleumier Heat Pumps,* New York: McGraw-Hill.

HEAT TRANSFER

Paul G. Klemens, *Department of Physics and Institute of Materials Science, University of Connecticut, Storrs, Connecticut, U.S.A.*

	Introduction	405
1.	**Thermal Conduction**	405
2.	**Thermal Conductivity**	406
2.1	Gases	407
2.2	Liquids	407
2.3	Solids	407
3.	**Radiation**	409
4.	**Convection**	410
4.1	Forced Convection	410
4.2	Free Convection	411
5.	**Interaction Between Transfer Modes**	412
6.	**Measurement of Thermal Conductivity**	412
7.	**Applications**	413
7.1	Enhanced Transfer	413
7.1.1	High-Conductivity Substrates	413
7.1.2	Hot Isostatic Pressing of Metal Powders	414
7.2	Reduced Transfer	414
7.2.1	Thermoelectric Energy Conversion	414
7.2.2	Thermal Insulations	415
	Works Cited	415
	Further Reading	416

INTRODUCTION

The transmission of heat between bodies at different temperatures, or between regions of the same medium at different temperatures, can take place without the movement of material, either by conduction through a body (solid or fluid) or by radiation through a transparent medium from one body to another. In addition, there can be heat transfer in fluids associated with the movement or circulation of matter. This process, convection, is often more effective than conduction in large bodies of fluid.

After discussing each mode of heat transfer individually, their combined effect will be discussed. Finally these principles will be illustrated by some applications.

1. THERMAL CONDUCTION

This is the transfer of heat within a continuous body, without overall or bulk motion of matter. It is through the motion of atoms or molecules that pass energy from regions of higher to lower temperature. According to the empirical Fourier law, which preceded a full understanding of the nature of heat, the local heat current density **J** (heat energy flowing per unit area per unit time) is proportional to the local temperature gradient grad$T$, T being the temperature, so that

$$\mathbf{J} = -\lambda \cdot \mathrm{grad} T. \qquad (1)$$

This relation defines the thermal conductivity. In the most general case λ is a symmetric tensor with six independent components. In crystals of cubic symmetry it has only three diagonal components, equal to each other, so that **J** and grad$T$ are parallel and λ appears isotropic, even though its tensor nature becomes apparent in derivatives such as the stress dependence of the thermal conductivity. Also λ is isotropic in materials without preferred direction (simple fluids, most amorphous solids, and solids composed of small crystallites arranged completely at random). In the case of isotropy, λ is replaced by a single parameter.

Measuring heat in units of energy and T in kelvins, the SI unit of thermal conductivity is $\mathrm{W\,m^{-1}\,K^{-1}}$. Older units are cgs ($\mathrm{erg\,cm^{-1}\,K^{-1}\,s^{-1}}$) and the convenient hybrid unit $\mathrm{W\,cm^{-1}\,K^{-1}}$, which frequently appears in the older literature and has not yet disappeared completely. Even older is the unit $\mathrm{cal\,cm^{-1}\,{}^{\circ}C^{-1}\,s^{-1}}$. Two practical British units are also used: $\mathrm{BTU\,ft^{-2}\,in.\,h^{-1}\,{}^{\circ}F^{-1}}$ and $\mathrm{BTU\,ft^{-1}\,h^{-1}\,{}^{\circ}F^{-1}}$. The reciprocal of the former defines

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

the R value of insulations. Table I lists some conversion factors.

An excess of heat flowing into a small region increases the local temperature, so that

$$\mathrm{div}\mathbf{J} = -C\rho\ \partial T/\partial t. \qquad (2)$$

Here $\partial T/\partial t$ is the rate of temperature increase at a fixed location, C is the specific heat, ρ is the density, and div**J** is the net rate of heat outflow per unit volume. For isotropic conductivity Eqs. (1) and (2) lead to

$$\mathrm{div}(\lambda\cdot\mathrm{grad}T) = C\rho\ \partial T/\partial t. \qquad (3)$$

Here λ may be a function of position (inhomogeneous medium). If λ is a constant, independent of position, and its T dependence can be neglected, Eq. (3) is a linear partial differential equation. The temperature T is a function of position and time. This equation can then be solved in terms of Fourier coefficients. Each Fourier component decays with a characteristic time that varies as the square of the wavelength, so that the spatial variation of the temperature field becomes progressively smoother with increasing time. Numerous special cases have been treated (Carslaw and Jaeger, 1959).

The conduction equation (3) is more complicated when λ is not constant. If it depends on T, the equation is no longer linear and numerical methods must be used to solve it. In the steady state when $\lambda(\mathbf{r})$ is a specified function of position, the problem is the same as for electrical conduction in inhomogeneous media, and various approximations have been proposed, as reviewed by Landauer (1978). For a medium that is inhomogeneous but globally uniform, the effective conductivity has upper and lower bounds given, respectively, by the volume average of $\lambda(\mathbf{r})$ and by the reciprocal of the volume average of the resistivity $1/\lambda(\mathbf{r})$.

The Fourier equation for thermal conduction is formally the same as that for diffusion. This is not surprising when one considers the microscopic picture of thermal conduction; heat transport can be considered as due to the random motion of carriers of thermal energy, with net transport in the direction toward the lower energy density due to such carriers.

It is thus useful in time-dependent problems to define thermal diffusivity by

$$\alpha = \lambda/C\rho. \qquad (4)$$

It has the same dimensions as ordinary diffusivity (length squared divided by time). It is an indicator of how long it takes for a temperature inhomogeneity to decay if that inhomogeneity extends over some given distance.

Table 1. Conversion of Units. Various units of heat and thermal conductivity converted to SI (MKS) units.

Units	SI
Heat	
1 cal	4.187 J
1 BTU	1055 J
Thermal Conductivity	
1 cal $\mathrm{s^{-1}\,cm^{-1}\,K^{-1}}$	418.7 $\mathrm{W\,m^{-1}\,K^{-1}}$
1 BTU $\mathrm{ft^{-2}}$ in. $\mathrm{h^{-1}\,{}^\circ F^{-1}}$	0.1442 $\mathrm{W\,m^{-1}\,K^{-1}}$
1 BTU $\mathrm{ft^{-1}\,h^{-1}\,{}^\circ F^{-1}}$	1.731 $\mathrm{W\,m^{-1}\,K^{-1}}$
1 W $\mathrm{cm^{-1}\,K^{-1}}$	100 $\mathrm{W\,m^{-1}\,K^{-1}}$
Thermal Resistance[a]	
1 R	0.176 $\mathrm{W^{-1}\,m^2\,K}$

[a] The R value of an insulation is the reciprocal of the conductance per square foot per °F, expressed in $\mathrm{BTU^{-1}\,ft^2\,h\,{}^\circ F}$.

2. THERMAL CONDUCTIVITY

On a microscopic or atomic basis, thermal conduction is due to the random diffusive motion of carriers of heat, moving with speed v and with a mean free path l. The net heat transport is proportional to the temperature gradient, and the thermal conductivity has the form

$$\lambda = \tfrac{1}{3}\,C'vl, \qquad (5)$$

where $C' = C\rho$ is the specific heat per unit volume of the material, ρ being the density. The factor $\frac{1}{3}$ accounts for the random direction of motion of the carriers. The concept of mean free path requires some elucidation. In a Boltzmann-equation approach, which examines the steady-state distribution of carriers in momentum space, one equates $l = v\tau$, where τ is the relaxation time that describes the approach of a perturbed distribution toward equilibrium as a result of the various random interactions between the carriers. When τ or l is a function of carrier energy, Eq. (5) is replaced by an appropriate integration over that energy. When there are several types of carriers, each type contributes additively to the conductivity. However, the mean free path of each carrier is possibly reduced by interaction with carriers of other types. For example, both electrons and phonons contribute a component each to the thermal conductivity of solids, but the electronic component is reduced by the interaction with phonons, while the phonon component is reduced by interaction with electrons, especially at low temperatures.

2.1 Gases

Here the atoms or molecules are themselves the carriers of heat. The heat energy is their kinetic energy and also the vibrational and rotational energy of molecules. The mean free path is comparable to the average distance traversed by a particle from an average initial position to the next collision, but not exactly so, because the relaxation time is really derived from the solution of an integral equation. If the gas consists of particles of only one kind, l is still a function of energy. However, the gas is subject to the conditions that pressure does not vary with local temperature and that there is no net flow of particles. This requires that τ is to a good approximation independent of energy, irrespective of the energy dependence of the collision cross section. Also C' in Eq. (5) is the specific heat per unit volume at constant pressure.

The average mean free path, when limited by interparticle collisions, is of the order of 0.1 μm at normal temperature and pressure. It varies inversely with pressure, but C' is proportional to pressure at fixed T, so that the thermal conductivity is practically independent of pressure. However, at sufficiently low pressures, particularly when the gas is confined in small spaces, the mean free path is also terminated by collisions with the walls and does not increase further with decreasing pressure. Then the thermal conductivity decreases with decreasing pressure (Knudsen effect). Also in that case the pressure in the presence of a temperature gradient is no longer constant, since $\tau = l(E)/v$ is no longer constant. In the extreme limit when l is limited only by boundaries and is independent of E, pressure varies as $T^{1/2}$ in a temperature gradient. This thermomolecular pressure can cause errors in gas thermometry or other pressure measurements when the gas space at one temperature is connected to the pressure sensor at another temperature by a narrow tube. A similar pressure variation would exist in a temperature gradient when the gas fills the interstices of a fine powder.

A more complex situation arises in gas mixtures. The mean free path of each species is limited by collisions with its own and with the other species. The total pressure stays constant in large spaces, but the partial pressure of any one component can vary with temperature, so that the composition varies with temperature. This thermal diffusion effect depends on the various collision cross sections and their energy dependence.

In molecular gases the internal degrees of freedom contribute to C', and for each such component there is, in principle, a different mean free path; collisions tend to bring all degrees of freedom into mutual equilibrium, but at different rates.

Thermal conductivity of gases is discussed, for example, by Hirschfelder *et al.* (1954) and more compactly by McLaughlin (1969).

2.2 Liquids

Liquids range in their properties from gas-like, but at high densities with strong interaction, to similar to solids, but with vanishing shear modulus over long times. The behavior of their thermal conductivity can correspondingly vary between these two extremes. Solids are discussed below. Hirschfelder *et al.* (1954) and McLaughlin (1969) also deal with liquids.

2.3 Solids

One can picture a solid as a uniform background that contains excitations; these behave like a gas of particles and carry thermal energy. The principal excitations are lattice vibrations, free electrons in electrically conducting solids, and electromagnetic radiation in transparent solids. The energy of electromagnetic waves consists of discrete quanta or photons, which can be regarded as particles, carrying energy and moving with the group velocity of the underlying wave. Similarly, the atomic vibrations can be represented by elastic or lattice waves; their quanta (phonons) carry energy with the group velocity of the waves. Each mobile excitation of type α makes a contribution C'_α to the specific heat per unit volume, and a corresponding contribution λ_α

to the thermal conductivity. If the excitations move with speed v_α and have a mean free path l_α, the thermal conductivity has the form

$$\lambda = \sum_\alpha \lambda_\alpha = \sum_\alpha \tfrac{1}{3} C'_\alpha v_\alpha l_\alpha. \tag{6}$$

The various mean free paths are limited by interaction processes. These include scattering of the mobile excitations or heat carriers by various defects of the crystal lattice and by boundaries, both external boundaries and internal or grain boundaries of polycrystalline aggregates. Phonons, and to a lesser extent electrons, do interact with each other. Also, when carriers of more than one type are present, they not only contribute additively to the conductivity, but they may interact with each other and limit each other's mean free path, as mentioned above.

In metals there are contributions from both electrons and phonons, but the electronic component frequently swamps the phonon or lattice component. The electronic component of the specific heat C_e is proportional to the absolute temperature T. In pure metals the electron mean free path is limited by interaction with lattice vibrations and varies as $1/T$; thus the thermal conductivity is almost independent of temperature from room temperature up. In the case of alloys with several-percent solute content, the mean free path is independent of T and $\lambda_e \propto T$, decreasing with increasing solute content. As a rough rule, one can regard the thermal resistivity $W = 1/\lambda_e$ as composed of an intrinsic component W_i independent of T, and a component W_0 that varies as $1/T$ and depends on solute content c, in continuous solid solutions varying as $c(1-c)$. Typically 1%–2% solutes make W_0 equal to W_i at room temperature. At low temperatures W_i decreases with decreasing temperature.

An equivalent and useful way of looking at the electronic thermal conductivity is to compare it to the electrical conductivity σ. Since the electrical conductivity is also proportional to the electron mean free path, and to the number of electrons per unit volume and to C_e, λ_e and σ are proportional to each other (Wiedemann-Franz law), and related by

$$\lambda_e = L_0 \sigma T, \tag{7}$$

where $L_0 = (\pi^2/3)(k/e)^2$ is the Lorenz ratio. Here k is the Boltzmann constant and e is the electronic charge. In mks units, $L_0 = 2.45 \times 10^{-8}\ \mathrm{V^2\,K^{-2}}$. At low temperatures, however, the mean free paths for thermal and electrical conduction are not equal and λ_e is often less than given by Eq. (7).

Equation (7) refers to the electronic component only; there is also a lattice component. If the electrical resistivity is high enough, λ_e is reduced and the lattice conductivity can become a significant fraction of the total. At room temperatures this may be the case when the electrical resistivity exceeds $10^{-3}\ \Omega\,\mathrm{m}$; below 90 K the resistivity needs only to exceed $10^{-4}\ \Omega\,\mathrm{m}$ for this to occur. The thermal conductivity of metals and alloys has been discussed by Klemens (1969) and by Klemens and Williams (1986).

In insulators and in most intrinsic semiconductors the electronic component is absent or negligible, and heat is carried by lattice waves or phonons. The phonon conductivity is

$$\lambda_p = \tfrac{1}{3} C'_p v_p l_p, \tag{8}$$

where C'_p is the vibrational specific heat per unit volume, v_p the group velocity of elastic waves, and l_p the phonon mean free path. When v_p and l_p depend significantly on frequency, this expression must be replaced by an integral over frequencies. The mean free path is limited by various processes:

1. the intrinsic interaction between the elastic waves due to the anharmonic nature of the interatomic forces;
2. the scattering of the elastic waves or phonons by various imperfections of the crystal lattice; and
3. the limitation due to the finite size of crystals, or of grains in the case of polycrystalline solids.

All these processes contribute additively to $1/l$, the scattering probability per unit path length.

Since these scattering probabilities depend on phonon frequency, with different processes having different frequency variations, one should add the contributions to $1/l_p$ for each frequency in turn, and consider the spectral distribution of C'_p to obtain the integral in the generalization of Eq. (8). While at ordinary and high temperatures the largest contribution to C'_p comes from the highest frequencies (with wavelength of about $\frac{3}{2}$ of the interatomic distance), the intrinsic mean free path increases with decreasing frequency, so that equal frequency intervals make equal contributions to the intrinsic phonon conductivity. Also $l_p \propto 1/T$, so that the intrinsic phonon conductivity λ_p varies as $1/T$. Point defects such as impurities or vacancies reduce mainly the high-frequency contribution; grain boundaries and extended defects, such as colloidal inclusions, reduce the low-frequency contribution.

At low temperatures the maximum contribution to C_p' shifts to lower frequency, and the temperature dependence of λ_p is also dependent on the frequency dependence of l_p. At very low temperatures the mean free path approaches a constant value, that of the smallest linear dimension of the specimen, when λ_p becomes proportional to C_p', and attains a T^3 dependence and a size dependence in many cases. This is the phonon-gas analog of the Knudsen effect in real gases.

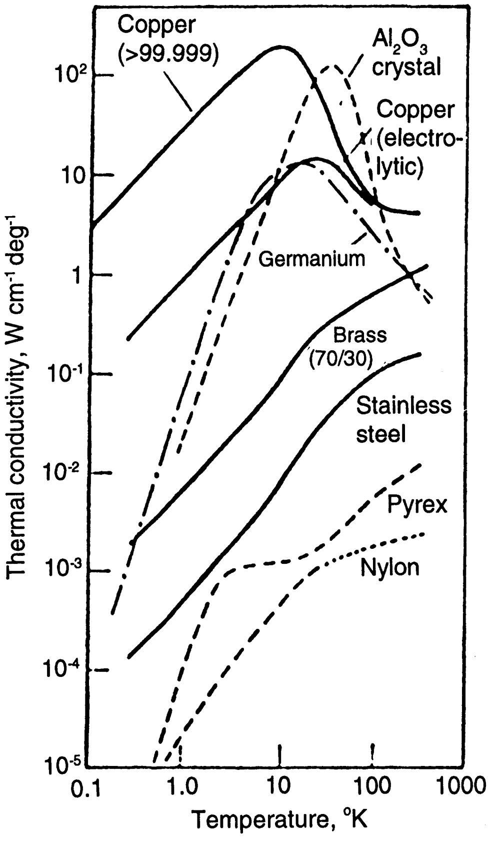

FIG. 1. Wide variability of the thermal conductivities of solids as functions of temperature for metals (full curves), semiconductors (dot-dashed curve), and insulators (dashed curves). From White (1969).

For a review of lattice thermal conductivity see Klemens (1969) and Berman (1976).

Figure 1 shows the thermal conductivities of several solids as functions of temperature, illustrating the variability of that property, which results from two conducting mechanisms, each subject to several different processes limiting the carrier mean free path.

3. RADIATION

Heat can be exchanged between solid bodies by electromagnetic radiation, even over great distances, if the intervening space is empty or filled with transparent material. The capacity of space to carry radiative energy increases rapidly with the temperature of the emitting body; also the energy peak of the emitted spectrum shifts to higher frequency as the temperature is increased. For room-temperature emission the peak lies in the far infrared, but moves to the visible at higher temperatures. The solar surface, at 6000 K, has its radiation maximum in the yellow, and this is also where the human eye has its peak sensitivity. Radiative heat transfer between bodies at different temperatures can be important even at ordinary temperatures; radiation from hot surfaces or to cold surfaces plays a part in human comfort. Radiation can also form a significant component of the thermal conductivity of condensed matter at high temperatures, provided the material has a sufficient spectral range of transparency.

The amount of radiant energy leaving unit surface area of a body varies with absolute temperature T and is given by

$$E = \sigma_{SB} \epsilon T^4, \qquad (9)$$

where $\sigma_{SB} = 5.670 \times 10^{-8}\ \mathrm{W\,m^{-2}\,K^{-4}}$ is the Stefan-Boltzmann radiation constant while the emissivity ϵ, a parameter generally less than 1, depends on the material, on surface roughness, and on the dominant wavelength of the radiation, hence on temperature. For a perfect blackbody, approximately realized by a small hole leading into a cavity, $\epsilon = 1$.

While radiation leaves every surface, other radiation reaches that surface from surrounding bodies, and a fraction ϵ is absorbed. While emissivity and absorption efficiency of a surface are the same at each wavelength, the average values depend on the spectral nature of the incoming radiation. The net energy transfer depends on the solid angle subtended by the other bodies, on the temperatures, and on the effective emissivity and absorptivity. For two bodies at temperature T_1 and T_2 the net energy transfer varies as $T_1^4 - \beta T_2^4$, where β depends on spectral differences and on geometry. When one body completely encloses another and their temperatures are equal, the net energy transfer must be zero.

In a partially transparent medium the photon mean free path, or the attenuation length of electromagnetic waves, is limited by absorption, by scattering from inhomogeneities, and by external boundaries. For moderate temperature differences one can define a radiative component of the thermal conductivity λ_r analogously to the phonon thermal conductivity, except that the specific heat per unit volume of space is smaller, and the mean free path is larger. As in the case of phonons, one must consider the frequency dependence of the mean free path, so that

$$\lambda_r = 4\sigma_{\mathrm{SB}} \frac{T^3}{26} \int n^2(f) l(f) x^4 e^x (e^x - 1)^{-2} dx. \tag{10}$$

Here $x = hf/kT$ is the reduced frequency and h is Planck's constant. Also $n(f)$ is the refractive index, generally a function of frequency f, and $l(f)$ is the mean free path of photons of frequency f. In general $1/l(f)$ is composed of contributions from absorption, from scattering, and from the limitation on the mean free path by external dimensions.

For a constant mean free path the peak in the radiative conductivity is at frequencies for which $x = 4$. In vacuum ($n = 1$) this corresponds to a wavelength of $2.898 \times 10^{-3}/T$ m, where T is in kelvins. In the presence of matter $n > 1$, and the wavelength is reduced by n. If $l(f)$ increases with decreasing frequency, the peak is shifted to lower frequencies. In extreme cases of strong and frequency-dependent scattering, λ_r can be size dependent with most of the heat transferred at those frequencies where $l(f)$ equals the smallest external dimension.

4. CONVECTION

Heat can be transported across a temperature gradient by fluid flow. In forced convection the flow arises from an external cause; in free convection the sources of the flow are density changes due to temperature differences in the presence of gravity. Convection transports heat within a body of fluid, and also transports heat from or to a solid body by conduction, through a boundary layer, into or out of the moving fluid.

When convection occurs, it is much more effective as a mode of heat transport over large distances than conduction. In conduction the heat current per unit area scales as $\Delta T/L$, where ΔT is the temperature difference over a distance L. In convection the heat current scales as $h\Delta T$, where the convection coefficient h is either independent of L or increases with L. It also depends on the flow rate and on local flow conditions at the interface. In free convection h varies roughly as L^3, but depends also on other factors.

An extreme example is heat transport in the mantle of the Earth, which is solid but flows with very high viscosity over geological time scales. Heat flow to the surface is primarily by free convection; conduction and radiation are unimportant and are dominant only in the thin crustal layer.

4.1 Forced Convection

The hydrodynamic equations that govern fluid flow are nonlinear, so that mathematical solutions can be obtained only in a few special cases, and the equations must be solved numerically. Many cases of engineering interest have also been investigated by scale models. For either case one can use dimensional analysis, and can scale solutions with similarities based on nondimensional parameters. The Reynolds number $Re = VL/\nu$ describes the transition between laminar and turbulent flow, being the ratio between inertial and viscous forces. Here V is the flow speed, L a linear dimension of the confining surface, and ν the kinetic viscosity (viscosity/density) that has the dimensions of diffusivity. A high value of Re results in turbulence, since laminar flow is stabilized by viscosity. Critical values of Re range from 10^3 to 10^4, depending on geometry and surface roughness.

Fluid flows parallel to a surface, with a boundary layer of laminar flow and regions of turbulence away from the surface. Within the boundary layer, heat is transferred by conduction. In the turbulent region a uniform temperature is reached by mixing, and the overall flow transports both turbulence and heat. Whether heat flow is limited by the boundary layer or by the rate of flow depends on the Prandtl number $Pr=\nu/\alpha$, where α is the thermal diffusivity. The Nusselt number is $Nu=hL/\lambda_f$, where h is the convection coefficient (heat flow per unit area and unit temperature difference). Now h/λ_f is the temperature gradient at the surface, where λ_f is the thermal conductivity of the fluid, since heat transfer in the boundary layer is by conduction; also $1/L$ is the average temperature gradient for unit temperature difference. Thus Nu is a dimensionless temperature gradient, relative to the average temperature gradient, and describes the relative importance of convection to conduction, or the inverse of the relative width of the thermal boundary layer to the laminar flow region. Finally, in the case of very poorly conducting solids, it is possible for heat flow to be limited by the thermal resistance within the solid. This is described by the Biot number $Bi=hL/\lambda_s$, where λ_s is the thermal conductivity of the solid.

All these dimensionless parameters define physically similar cases of heat transfer for the appropriate limiting mechanism.

4.2 Free Convection

Here the flow responsible for heat transfer is caused by differences in buoyancy due to thermal expansion. Since heat transfer varies as the product of flow velocity V and temperature difference ΔT, and V is caused by ΔT, the heat transfer by free convection would vary as ΔT^2, except for departures from that variation due to changes in the flow pattern with flow speed, e.g., from laminar to turbulent.

Free convection is the result of an instability of a horizontally layered structure of density variations. It thus occurs most easily when a heated surface or other heat source is near the bottom of the fluid.

For similarity considerations, the relevant dimensionless parameter is the Grashof number

$$Gr=g\rho_1(\rho_2-\rho_1)L^3/\nu^2$$
$$=g\rho_1^2\beta(T_1-T_2)/\nu^2, \qquad (11)$$

where g is the gravitational acceleration, β is the coefficient of volumetric thermal expansion, ρ is the density, and suffixes 1 and 2 refer to the hot and cold regions, respectively. The first version is preferable near the critical point, but the second one is the one frequently used. The Grashof number plays a role similar to the Reynolds number in forced convection.

The Rayleigh number $Ra=Gr\nu/\alpha$ describes whether free convection is more important than conduction; for $Ra<10^3$, free convection can be ignored. Apart from buoyancy, free convection depends sensitively on the linear dimension L. For $Ra>10^3$, an instability sets in and fluid motion consists of roll cells that, between parallel horizontal plates, are regularly spaced as shown in Fig. 2. This mode (Bénard-Rayleigh instability) persists for Ra values up to between 10^4 to 10^5, when the fluid motion becomes turbulent. Convection can be suppressed by dividing up the available space into smaller units, as is done in insulations.

Both forms of convection depend, of course, on the geometrical arrangement. For a fuller discussion see Burmeister (1983), or books on heat transfer. Free convection is also discussed by Ziebland (1969).

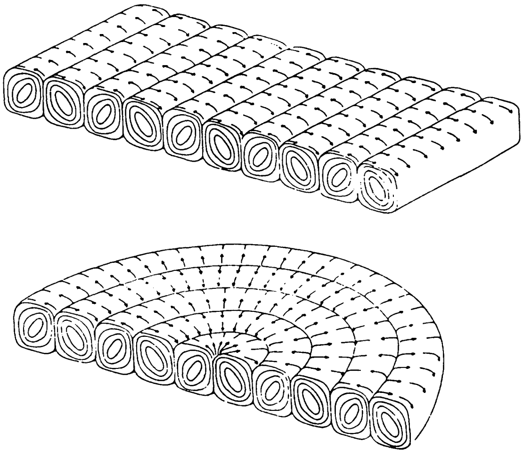

FIG. 2. Roll cells (Bénard-Rayleigh instabilities) in a fluid heated from below, for rectangular and circular symmetry. The fluid is confined at top and bottom by parallel plates. From Velarde and Normand (1980).

5. INTERACTION BETWEEN TRANSFER MODES

There are three basic mechanisms of heat transfer, conduction, radiation, and convection, and it is relatively rare that only one mode is important in any given situation. To a first approximation one treats the heat transfer modes as additive, though for large temperature differences this is unlikely to be true.

Thus, for small ΔT, radiative heat transfer is a linear function of ΔT, and it can be added to transfer by conduction. This holds particularly for radiative heat transport within a translucent solid, where radiation is important at higher temperatures and can be regarded as a component of the conductivity. However, when the attenuation length becomes comparable to the linear dimension and ΔT is no longer small compared to the absolute temperature T, conduction and radiation are no longer additive Also, in transient situations absorption and scattering of radiation play different roles.

Heat transport by phonons in dielectric crystals at very low temperatures, where at least part of the phonons can pass right through the specimen, shows similarity to radiation. The relative amounts of heat thus transported are in the ratio $T^3c^2v^{-2}$, where v and c are the phonon and light speeds, respectively. Phonons transport more heat. In the case of phonons it is also strictly speaking not possible to speak of a true conductivity (Klemens, 1969).

In fluids, free convection does not behave like conduction, since it depends strongly on the linear dimension of unbroken space. Thus the two modes are not additive in their action. In partially transparent fluids all three modes are present. The absorption of radiation in the fluid can cause local heating and enhance the onset of convection.

6. MEASUREMENT OF THERMAL CONDUCTIVITY

There are difficulties in the measurement of thermal conductivity that are often not fully appreciated and that have led in the past to substantial errors in reported values. In principle all that is needed is a measurement of heat current in a specimen of known geometry, together with a measurement of the temperature difference between two specified surfaces normal to the heat flow. Heat flow is known from the heat generated at one end by a heater, if one can be sure that all heat does indeed flow through the specimen. This, of course, is where errors can occur, because convection and radiation do cause some heat to bypass the specimen, at the same time also distorting the isothermal surfaces.

In the case of solid specimens, convection can be eliminated by placing the specimen in an evacuated container. When this is not possible, one should at least have the specimen mounted vertically with the heater on top. Evacuation is used at low temperatures, where one can use rod-shaped specimens. However, radiation can bypass the specimen. Depending on the conductivity, the rod diameter, and the emissivity, radiation becomes important above about 100 K. There is radiation from the sides of the specimen, and also from the heater and from the attached thermometers. Attempts to correct quantitatively for that radiative heat flow have usually not been successful. Packing the space between the specimen and the outer shield with evacuated powder or fiber insulations, using radiation baffles, and matching the temperature gradient of the specimen and that of an outer shield do mitigate the problem. However, as one goes to higher temperatures the geometry of the specimen must be changed, ideally platelike rather than rod-shaped. It is best to surround the plate with temperature-matched guard rings, so that edge effects may be minimized. This geometry is the most suitable for low-conductivity materials. However, it becomes difficult to measure the temperature gradient, since small errors in the effective position of the thermometers become important. Nevertheless this arrangement is currently the best way to measure the overall conductivity of insulations and other low-conductivity materials.

As one goes to high temperatures, radiation effects become increasingly troublesome, and it also becomes difficult to anchor thermometers to the specimen. Pyrometry has been used, sighting into small holes in the specimen. In recent years, however, transient techniques have been preferred. There are several versions. The most widely used one is the flash diffusivity method. A heat pulse is deposited on one side of a disk-shaped specimen, and the temperature rise on the back face is measured as a function of time, using a rapid-response thermometer or pyrometer. This yields the thermal diffusivity $\alpha=\lambda/\rho C$, where λ is the thermal conductivity, ρ the density, and C the specific heat. One still needs to know the specific-heat per unit mass and the density at the required temperature. High-temperature specific-heat measurements are far from trivial, but in general the specific heat and the thermal expansivity are less specimen dependent than the thermal conductivity, so that literature values can be used where available. The heat can be deposited by means of a laser, a flash lamp, or a pulsed current in a thin-film resistor. The temperature rise can be monitored by attached thermocouples or by infrared detectors. Nonlinearity in response can become important, and in some cases one must go beyond the one-dimensional conduction problem to consider spreading and edge effects. The technique is not sensitive to incomplete absorption or to heat losses at the front surface, or to small thermometer errors; the measurement of thermal diffusivity is mainly one of specimen thickness and response time. The fact that a small specimen can be used is often a great advantage. For transparent specimens one must place a thin absorbing film on the front face, and this can cause complications.

Transient methods can also be used to measure thermal conductivity and specific heat, but then one must pay careful attention to measure absolute temperature changes and the deposited heat.

Measurements of the thermal conductivity of fluids present all the problems just discussed, and in addition it is necessary to confine the fluid, both gases and liquids, in a container. Also, one must suppress free convection. A cylindrical geometry is often used, with a central wire that serves both as heater and as thermometer, and an outer cylindrical sheath. End effects are corrected for either by using two cells of different length, or by guard cylinders. The size must be chosen so as to make the Grashof number small enough. The hot-wire method can be adapted to conductivity measurements as function of pressure, and has also been used for powders. The main problem is accurate alignment. Experimental techniques to measure the thermal conductivity of fluids have been reviewed by Ziebland (1969).

The appreciation of the experimental difficulties in thermal conductivity measurements has led to a series of conferences on thermal conductivity, where these problems have been discussed in depth (Cremers and Fine, 1990; also earlier conferences in that series).

7. APPLICATIONS

Problems in heat transfer can be divided into two groups, either to obtain a maximum heat transfer between two bodies or regions, or to keep the heat transfer as low as possible. The applications given here divide into these two classes.

7.1 Enhanced Transfer

7.1.1 High-Conductivity Substrates. Although the amount of heat generated in electronic circuit elements is not large, the development of integrated circuits of high packing density does require high current densities and low temperature differences. Cooling by forced or free convection is not possible directly at these elements, but only at the far side of a solid substrate, often loosely called a heat sink. The development of electrically insulating solids of high thermal conductivity has thus become a priority. These substrates must have a high phonon conductivity, even in the form of thin plates or films.

High phonon conductivity at room temperature requires single crystals or large-grained polycrystalline materials. The basic material should be composed of light atoms, strongly bound, and should be free of lattice defects and impurities (to better than 3×10^{-3} per atom), and in the case of a polycrystalline aggregate the porosity or the amount of binder should be small. For a volume fraction c of pores or other low-conductivity inclusions the conductivity is reduced by a factor of $1-nc$, where n varies between $\frac{4}{3}$ and $\frac{5}{3}$. Candidate materials include boron nitride, beryllium oxide, and diamond. Oxides of the next row in the periodic table (MgO, Al_2O_3, SiO_2) have a lower thermal conductivity but are more readily available as single crystals. There is usually a 10%–15% difference between single crystals and polycrystalline aggregates of the same material (Touloukian *et al.*, 1970).

Natural diamond crystals have been considered as substrates for special applications, but only pure crystals have a high conductivity of 3000 W m^{-1} K^{-1} at room temperature. Their availability and cost would be a limitation. However, in recent years it has become possible to grow diamond films and free-standing thin plates by chemical vapor deposition. The key to the process is the removal of incipient graphite inclusions, using atomic hydrogen as an etchant. These films are composed of small grains, random or columnar, and also the lattice is not perfect but contains both point defects of atomic dimensions and large imperfections of 100-Å scale. All these depress the thermal conductivity substantially below 100 K, while at room temperature it has been possible to approach the conductivity of bulk diamond to within a factor of 2.

Diamond crystals can also be grown in pressure vessels using metals as pressure-transmitting media. These crystals are pure, but may contain traces of impurities. A surprising result was that the use of monisotopic carbon led to a significant increase in thermal conductivity. To what extent the room-temperature thermal conductivity is influenced by the natural 1% content of ^{13}C, and to what degree it is influenced by residual impurities, is still a matter of debate (Anthony *et al.*, 1990).

7.1.2 Hot Isostatic Pressing of Metal Powders. Many metal alloys are best prepared in the form of powders, as this allows rapid cooling of small droplets and retention of metastable phases. These powders are then compacted by pressing at temperatures high enough for plastic flow but low enough to retain the desired structures. Isostatic pressing allows control of the final shape. The powder is filled into a container that is evacuated and sealed, and hot gas is used to apply external pressure to the container and to the powder mass within it. One problem is to ensure that the powder has reached the desired temperature everywhere, particularly in the center of the container. Each particle is at one temperature, as a result of the high conductivity of the alloy, but the contact between particles has a small area, which varies as the hydrostatic pressure of the powder itself, and inversely as the yield stress. Gas conductivity is absent, and radiation transport is inhibited by the small photon mean free path in the interstices, so that the overall conductivity is very low. Preheating thus requires long times. One cannot directly measure but one can only estimate theoretically the time required, which can be of the order of a day for moderate sizes. Usually the last place to reach the pressing temperature is near the top of the powder mass, where contact areas are smallest, about as deep down as the horizontal radius of the form. This is indeed the region where pressing defects are most frequently found.

7.2 Reduced Transfer

In the following applications, one aims to reduce the heat transfer.

7.2.1 Thermoelectric Energy Conversion. A thermoelectric element can be used as a heat engine. The temperature difference causes a thermoelectric voltage that generates a current to deliver external work. Similarly an applied current can lift heat energy from the cold to the hot junction. The efficiency depends on the Carnot efficiency and a reducing factor that increases with a coefficient of performance $ZT = S^2\sigma T/\lambda$. Here S is the thermoelectric power, σ the electrical conductivity, T the absolute temperature, and λ the thermal conductivity. Until now it has been possible to obtain ZT values around 1. An increase in the coefficient of performance would widen the applicability of thermoelectric converters and refrigerators. Robustness and long life, so important in space applications, are achieved as a compromise at the expense of efficiency; here too an improvement in ZT is desirable. For a recent discussion see Allred *et al.* (1991).

Apart from external parasitic effects, the main limitation is the phonon component of the thermal conductivity, λ_p. In these conductors, which are between semiconductors and semimetals, $\lambda = \lambda_e + \lambda_p$, and the electronic component λ_e depends on both σ and S. If it were not for the phonon component, the quantity to be maximized, $S^2\sigma/\lambda_e$ would be the larger the lower the carrier concentration, since a small concentration of electrons or holes increases the magnitude of S. If $\lambda_p > \lambda_e$, one aims to increase $S^2\sigma$, which means increasing the carrier concentration. The actual optimum occurs for carrier concentrations such that λ_e is comparable to λ_p. The smaller one can make λ_p, the higher is the value of ZT for optimum carrier concentration. The aim is then to reduce λ_p without reducing σ for fixed S, i.e., without reducing the carrier mobility.

An obvious step is to make the semiconductor an alloy, replacing atoms by solutes from the same column of the periodic table, so that the mass and size difference scatters phonons, while electron scattering is minimized by the similarity in the ionic charge. However, this phonon scattering depends strongly on frequency, scattering high-frequency phonons but allowing low-frequency phonons to pass unimpeded. At those carrier concentrations that optimize performance, the phonons of lowest frequency are scattered by the carriers. This still leaves a frequency range or window in which phonons are only weakly scattered. This limits the attainable reduction in λ_p. Current research aims to close this window by introducing extended defects, which scatter phonons in that frequency range without degrading the mobility of the carriers.

7.2.2 Thermal Insulations. Efficient insulations consist of an array of solid particles of low conductivity or other elements, of small size, making contact with each other only over small areas, so that solid conduction takes a tortuous path and is reduced. At the same time the spaces between the particles are small, reducing convection. This still leaves gas conduction and radiation as important components.

The insulations can be powders, fibers, or foams. The last are cells of space separated by thin solid walls. The solid material should have a low bulk thermal conductivity.

The best insulations are those where gas has been removed from the interstices. These are the most expensive to produce, requiring prolonged pumping and vacuum-tight enclosures. The enclosures, in turn, require some form of mechanical support that act as conducting bridges; without spacers the enclosure walls could press onto the solid filler and increase the contact areas. In such evacuated insulations the solid filler still plays an important role, because the particles cut down the radiation. Unless the particles are very fine or the fibers extremely thin, the radiative resistivity $1/\lambda_r$ depends on the number of surfaces intersected by a random line traversing the insulation.

Evacuated panels are too expensive for many applications, and most common insulations, such as building insulations, contain gas at atmospheric pressure. It is then an advantage to use fine fibers, fine powders, or foams of fine structure, so that radiation is reduced by numerous interfaces, and the small interstices reduce gas conduction by limiting the molecular mean free path.

In the commonly used glass-fiber building insulations one accepts gas conduction, hoping to reduce the overall conductivity to approach that of air. The fibers serve to eliminate convection in the gas, and to reduce radiation by presenting interfaces and by some absorption. Since radiative resistance depends mainly on the total number of fiber surfaces in the total thickness of the insulation, a limited amount of settling can be tolerated.

Solid foams, which contain the blowing gas in the cells, also make good panel insulations, though they are fragile. It helps if the gas in the cells has a large molecular weight, though there may be environmental concerns. Unfortunately, air tends to leak or diffuse into the cells, and the blowing gas eventually leaks out. This increases conduction, so that foams deteriorate over a time that is too short for the long life required of building insulations. Encapsulation is a possible solution, but this limits the applicability of foams.

Aerogels are an interesting class of materials, developed in recent years. They consist of a skeleton of fine particles with interstices. The particles themselves contain substructures over a range of dimensions, so that the material has intrinsic interest as a fractal structure. The main interest lies in their low density and low thermal conductivity. Aerogels can be produced by supercritical evaporation of many liquid materials; commonly the residue is a form of amorphous silica. After evacuation, the material is lightweight and solid, has low thermal conductivity, and is translucent. For robust insulations some encapsulation or covering film are required. This is currently a field of intense research and development (Fricke, 1986, 1992).

List of Works Cited

Allred, D. D., Vining, C. B., Slack, G. A. (Eds.) (1991), *Modern Perspectives on Thermoelectric and Related Materials*, MRS Symposium Proceedings Vol. 234, Pittsburgh: Materials Research Society.

Anthony, T. R., *et al.* (1990), *Phys. Rev. B* **42**, 1104–1111.

Berman, R. (1976), *Thermal Conduction in Solids*, Oxford: Oxford University Press.

Burmeister, L. C. (1983), *Convective Heat Transfer*, New York: Wiley.

Carslaw, H. S., Jaeger, J. C. (1959), *Conduction of Heat in Solids*, 3rd ed., Oxford: Oxford University Press.

Cremers, C. J., Fine, H. A. (Eds.) (1990), *Thermal Conductivity 21*, New York: Plenum; also earlier conferences in that series.

Fricke, J. (1986), *Aerogels*, Berlin: Springer-Verlag.

Fricke, J. (Ed.) (1992), *J. Non-Cryst. Solids* **145**, 1–258.

Hirschfelder, J. O., Curtiss, C. F., Bird, R. B. (1954), *Molecular Theory of Gases and Liquids*, New York: Wiley.

Klemens, P. G. (1969), "Theory of Thermal Conductivity of Solids," in: R. P. Tye (Ed.), *Thermal Conductivity*, London: Academic, Vol. 1, pp. 1–68.

Klemens, P. G., Williams, R. K. (1986), *Int. Metals Rev.* **31**, 197–215.

Landauer, R. (1958), in: J. C. Garland, D. P. Tanner (Eds.), *Electrical Transport and Optical Properties of Inhomogeneous Media*, AIP Conf. Proc. No. 40, New York: American Institute of Physics, p. 2.

McLaughlin, E. (1969), "Theory of Thermal Conductivity of Fluids," in: R. P. Tye (Ed.), *Thermal Conductivity*, London: Academic, Vol. 2, pp. 1–64.

Touloukian, Y. S., Powell, R. W., Ho, C. Y., Klemens, P. G. (1970), *Thermophysical Properties of Matter*, Vol. 2: *Thermal Conductivity of Nonmetallic Solids*, New York: Plenum.

White, G. K. (1969), in: R. P. Tye (Ed.), *Thermal Conductivity*, London: Academic, Vol. 1, p. 70.

Velarde, M. G., Normand, C. (1980), *Sci. Am.* **243** (1), 92.

Ziebland, H. (1969), "Experimental Determination of the Thermal Conductivity of Fluids," in: R. P. Tye (Ed.), *Thermal Conductivity*, London: Academic, Vol. 2, pp. 65–148.

Further Reading

Berman, R. (1976), *Thermal Conductivity in Solids*, Oxford: Oxford University Press.

Burmeister, L. C. (1983), *Convective Heat Transfer*, New York: Wiley.

Incropera, F. P., DeWitt, D. P. (1990), *Introduction to Heat Transfer*, 2nd ed., New York: Wiley.

Kreith, F., Bohn, F. S. (1986), *Principles of Heat Transfer*, 4th ed., New York: Harper & Row.

Parrott, J. E., Stuckes, A. D. (1975), *Thermal Conductivity of Solids*, London: Pion.

HEAT-PULSE PROPAGATION IN SOLIDS

VENKATESH NARAYANAMURTI, *Department of Electrical and Computer Engineering, College of Engineering, University of California, Santa Barbara, Santa Barbara, California, U.S.A.*

Introduction and General Aspects 417
1. **Nature of Heat-Pulse Propagation** 420
1.1 Ballistic Phonons and Phonon Focusing 420
1.2 Second Sound 422
2. **Phonon Spectroscopy** 424
2.1 The Case of Donors and Acceptors in Semiconductors 424
2.2 Phonon Scattering by Dislocations 427
2.3 Phonon Scattering by Surfaces and Interfaces 428
2.4 Phonons in Superlattices—Mirrors, Filters, and Related Devices 429
3. **Optical Techniques** 430
3.1 Opto-Acoustic Spectroscopy 431
3.2 Picosecond Optical Techniques 431
4. **Conclusions** 431
Glossary 432
Works Cited 432
Further Reading 433

INTRODUCTION AND GENERAL ASPECTS

Acoustic waves or sound waves result from the vibrations of atoms or molecules and, at audible frequencies (lower than about 20 kHz), can be generated in many different ways. Human speech, for example, lies in this range. Sound waves at these frequencies can propagate reasonably well through the atmosphere, as evidenced by the fact that the waves can be heard from afar. But at higher frequencies (about 1 MHz), where sound waves are inaudible, the atmosphere is a poor medium for propagation. Ultrasonic waves are rapidly dissipated in the gases of the air and can be made to travel appreciable distances only in condensed media such as liquids and solids. Acoustic waves vibrating in condensed matter at ultrasonic frequencies are usually generated by radio-frequency excitation of piezoelectric materials such as quartz (see Fig. 1).

In the megahertz region ultrasonic waves have been widely used for the study of underwater objects, internal structural features of materials, and even organs inside the human body. The wavelength of megahertz ultrasonic waves in many solids is typically in the millimeter region, and thus such waves are sensitive to features that are macroscopic in size. Using microwave (Bömmel and Dransfeld, 1960) excitation of piezoelectric thin films or surfaces, one can extend the frequency to the gigahertz (10^9 Hz) regime, and the wavelength can be reduced to the submicrometer region. Since the early 1960s microwave ultrasonics has been widely used for the study of materials (Nowick and Berry, 1972). More recently an ultrasonic apparatus, the acoustic microscope, was developed (Quate, 1979) to generate "acoustic pictures" with a resolution comparable to that of visible light microscopy. Sound waves and light waves are reflected and absorbed differently by the same object and thus can often yield complementary microscopic information about the materials being studied.

Acoustic propagation (see Ilukor and Jacobsen, 1966) with conventional piezoelectric transducers becomes quite difficult at frequencies beyond about 10^{10} Hz. Sound waves with frequencies of hundreds of gigahertz and terahertz (10^{12} Hz) are well into the thermal vibrational spectrum of the solid. At these enormously high frequencies the discrete nature of the thermal vibrational spectrum becomes important, and the quantized lattice vibrations—also referred to as phonons—

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

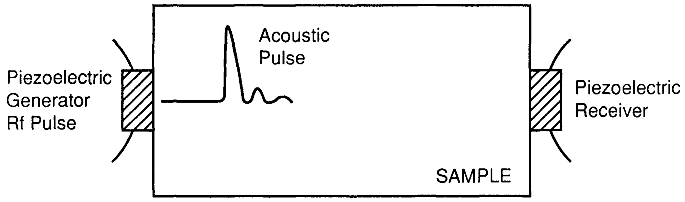

FIG. 1. Schematic of typical acoustic experiment using piezoelectric transducers.

have extremely short wavelengths, in the range 10–100 Å. Because of the short wavelength it becomes extremely difficult to preserve the phase coherence of a piezoelectrically generated wave as it traverses the boundary between the generator and the transmission medium under study. In addition, the propagation of such high-frequency phonons can be severely affected by the presence of defects and by phonon–phonon scattering.

An alternative to piezoelectric excitation of single-frequency lattice vibrational phonon modes would be to use a heat source to generate a broad spectrum of phonons locally on one end of the sample under study and "watch" their propagation with a temperature-sensitive detector at the other end of the sample [Fig. 2(a)]. The time constant and hence the speed with which this can be done is determined by the ratio of the heat capacity and the thermal conductivity of the thin-film transducers. For vacuum-deposited thin (~1000 Å) metal films this time constant is typically about 10^{-8} s, while the travel time across the sample is $\sim 10^{-6}$ s (depending on the phonon velocity and sample thickness). In the case of a normal-metal film such as gold or copper, the resistive heating results in a broad Planck frequency distribution [see Fig. 2(a)] with a peak in frequency that depends on the temperature T_h of the heater. For a value of $T_h \sim 10$ K the peak phonon frequency is ~0.6 THz. Such metal heaters are thus efficient sources of high-frequency phonons, but have a broad distribution similar to tungsten lamps in optics. The heat-generated phonons can be detected at the other end of the sample after a characteristic time of flight by means of a sensitive thin-film bolometer, that is, a device whose resistance is a strong function of temperature. Such heat-pulse experiments in solids were first reported by von Gutfeld (1968).

In order to get a more monochromatic distribution of high-frequency phonons, one has to make use of the detailed quantum-mechanical interaction between electrons and phonons in metals and semiconductors. In the case of superconductors, the existence of the energy gap, 2Δ, leads to the possibility that phonons can excite electrons across the gap, an analogy with the photoelectric effect. The energy gap itself arises because it takes an energy of 2Δ to bind two electrons of opposite momenta into what is known as a Cooper pair. The binding is provided through the mediation of the electron coupling by the strong electron–phonon interaction responsible for superconductivity. A superconducting device that can be used to generate high-frequency phonons is shown on the left-hand side of Fig. 2(b) and is known as a tunnel junction. It consists of a sandwich of extremely thin (~1000 Å) films of two superconductors with an insulator (~20 Å) in between. When a voltage V is applied across the sandwich, the insulator acts as a gate that transmits only electrons of energy $eV > 2\Delta$, where e is the electron charge. Because the electrons can couple to phonons by the electron–phonon interaction, the injected electrons transfer their energy to the lattice in the form of phonons with a spectral distribution that strongly deviates from a thermal Planck distribution. For voltages close to a value such that $eV = 2\Delta$, the recombination radiation (which involves the recombination of electrons of opposite momenta to form a Cooper pair) corresponds to phonons of frequency $\nu = 2\Delta/h$. For higher voltages, "relaxation" phonons of energy $eV - 2\Delta$ can also be emitted. This is because the electron loses its energy in two steps. It first reaches the gap edge by emitting $eV - 2\Delta$ phonons, and then recombines with an electron of opposite momentum to emit a 2Δ phonon. Thus the phonon spectrum depends in a characteristic way on the electron injection energy. The upper frequency limit of the relaxation phonon spectrum is determined by reabsorption processes in the superconducting film, and in favorable cases [see Eisenmenger (1976)] it can exceed the superconducting energy gap. The shape of

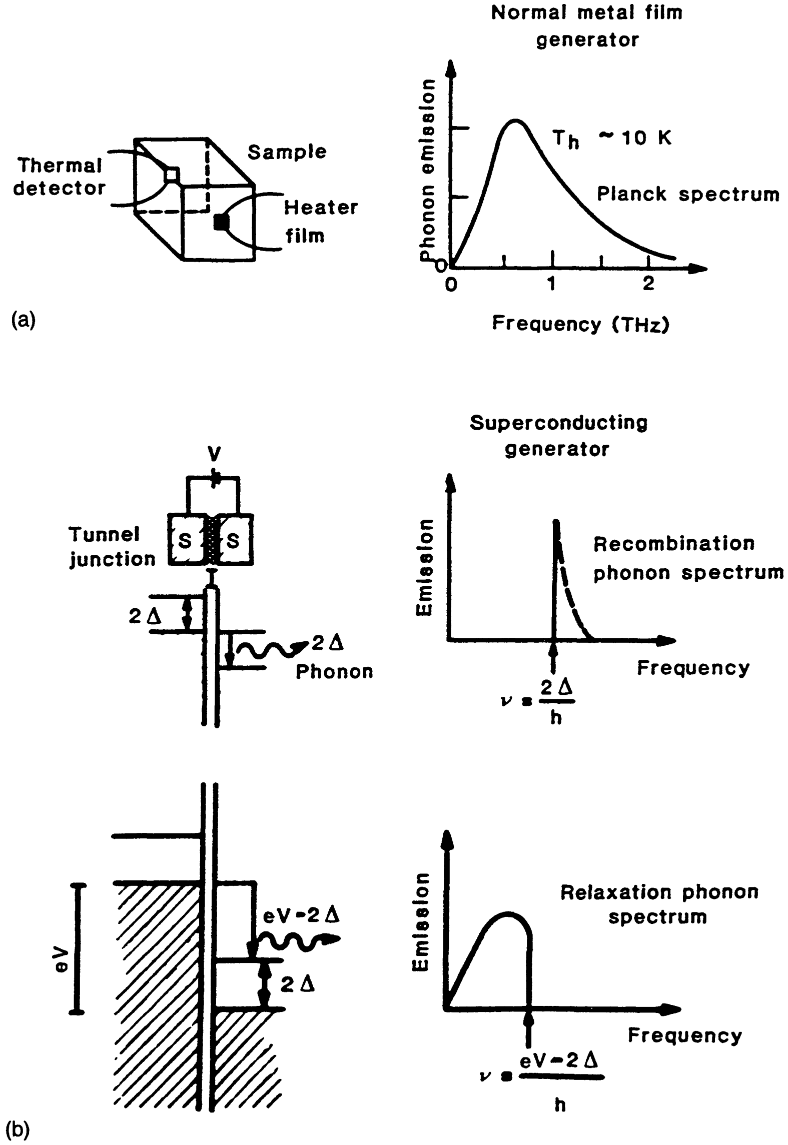
Thermal detector
Sample
Heater film
(a)
Normal metal film generator
Phonon emission
Th ~10 K
Planck spectrum
0
0
1
2
Frequency (THz)
V
Tunnel junction
S
S
2Δ
2Δ
Phonon
Superconducting generator
Emission
Recombination phonon spectrum
Frequency
ν = 2Δ/h
eV
eV-2Δ
2Δ
Emission
Relaxation phonon spectrum
Frequency
ν = eV-2Δ/h
(b)

FIG. 2. (a) Schematic of experimental arrangement with thin-metal-film transducers. Emission spectrum for a normal-metal film for a characteristic temperature of 10 K. (b) Spectrum of phonons generated by a superconducting tunnel-junction generator under different conditions of bias. For biases close to the gap edge, only recombination phonons are generated. For higher biases, the electron loses energy in two steps by emitting first a relaxation phonon and then a 2Δ phonon.

the generated recombination and relaxation phonon spectrum is determined in detail by the strength of the electron–lattice coupling and the density of states (i.e., the shape of lattice vibrational spectrum of the superconductor). Table 1 shows the transition temperature T_c and zero-temperature energy gap for a few common superconductors. It is clear that quantum processes in superconductors

Table 1. Transition temperature T_c, energy gap 2Δ, and gap frequency γ for several superconductors.

Superconductor	T_c (K)	2Δ (MeV)	γ (10^{11} Hz)
Al	1.18	0.36	0.9
Sn	3.8	1.2	2.9
Pb	7.2	2.7	6.7
$Pb_{0.85}Bi_{0.15}$	8.5	3.2	8.02

can be used for the generation and detection of phonons in the range 10^{11}–10^{12} Hz.

In addition to thin metal-film transducers, in the case of semiconductors and insulators one can excite the electronic energy states of donors, acceptors, and other atomic point defects, which, when preferentially excited by optical or electrical means, can lose their energy by phonon processes [see Narayanamurti (1981)]. Laser excitation in general can be a very useful method for studying phonons in solids. Laser beams can be focused to a small spot, scanned across the surface, tuned in frequency to excite specific levels, and pulsed to very short duration. The combination of laser excitation with phonon techniques can thus be very powerful. For example, one can directly excite impurity levels in crystals with far-infrared lasers and then watch the deexcitation through the emission of extremely high-frequency resonant terahertz phonons. In other cases, optical radiation can be absorbed in a thin film in an extremely short (picosecond) time scale, and thereby acoustic phenomena in an entirely new time domain can be observed.

In this article we will provide a survey of the nature of the propagation of both heat and high-frequency quasimonochromatic phonons in solids. We will show that the propagation in appropriate circumstances can be highly directional in nature, a property that can be useful for studying materials. Both the focusing of phonon energy and the use of high-frequency phonons for spectroscopic studies will be described. New applications analogous to photoacoustic spectroscopy will also be briefly presented. Thus the manipulation of phonons with novel transducer materials and the physical studies made possible may be best described as *phonon optics,* a term the author first used in 1979, which will form the major portion of this article.

1. NATURE OF HEAT-PULSE PROPAGATION

Propagation of heat in most materials is generally thought to be a slow diffusive process. This is true if the phonons that carry the heat suffer frequent scattering events. These scattering events are usually the result of the presence of defects and free carriers and of intrinsic phonon–phonon collisions. Phonon–phonon collisions can be minimized by cooling the sample to low temperatures where the population of thermally excited phonons is small. In "perfect" dielectric solids and semiconductors at low temperatures, heat propagation can thus be ballistic; that is, the phonons created by a pulse of heat at one point can travel rectilinearly for distances of the order of centimeters at the velocity of sound [see, for example, von Gutfeld (1968)].

The frequency distribution of the heat pulse, as described earlier, is characterized by a heater "temperature" T_h that can be substantially higher than the ambient temperature of the sample. The problem is like the radiation from the sun. On a clear day, we on Earth see a frequency distribution that is characterized by the temperature of the source (the sun in this case), even though the actual temperature on the Earth is much lower. In a moderate fog our eye detects a diffuse source of scattered photons, but may still observe some ballistic photons that sharply define a dim image of the sun. In a dense fog, only scattered photons can be seen.

As mentioned earlier, the phonons can also scatter along the way between the source and detector. Any perturbation that locally destroys the periodicity of the lattice will act as a scattering center. In an otherwise perfect crystal, even the random distribution of naturally occurring atomic isotopes can act as scattering centers. The isotope scattering is analogous to Rayleigh scattering of light by atoms, and has a strong frequency dependence (ν^4). High-frequency phonons scatter more readily than low-frequency ones. For similar reasons, the sky is blue.

1.1 Ballistic Phonons and Phonon Focusing

Ballistic phonons and their scattering have been observed in numerous solids over the last 20 years. In Fig. 3(a) we show typical heat pulses observed in the ⟨110⟩, ⟨111⟩, and ⟨100⟩ crystallographic directions in nominally pure GaAs for an almost identical propagation length of about 2.3 mm. The separate pulses correspond to ballistic traversal of longitudinal (L) and fast and slow transverse (FT and ST, respectively) phonons, whose velocity depends on the direction of phonon propagation. The longitudinal mode is a compression wave and has the highest velocity. The trans-

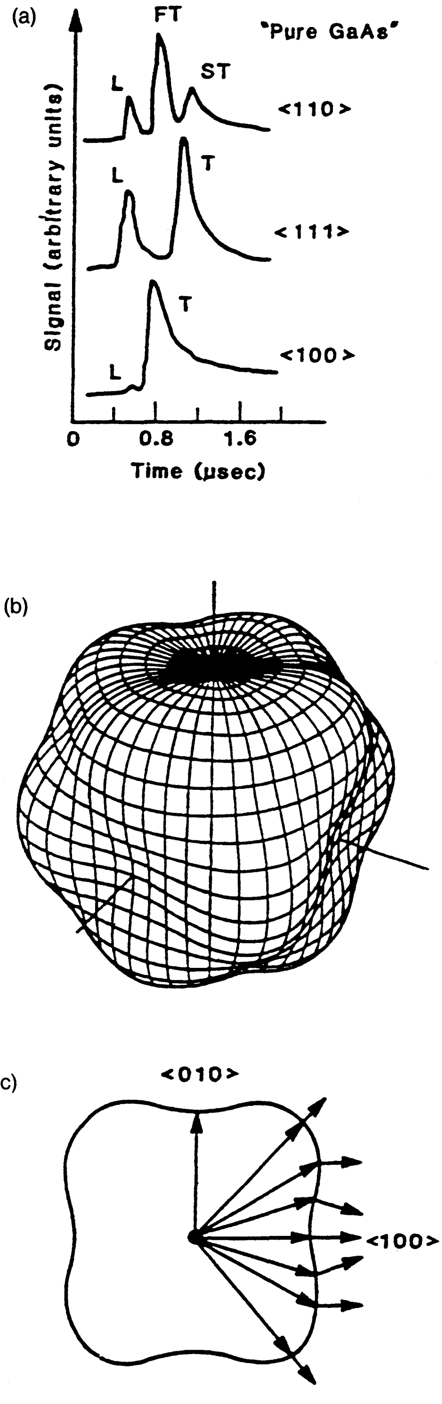

FIG. 3. (a) Ballistic phonons in "pure" GaAs for three different propagation directions; propagation length ~2.3 mm (after Narayanamurti, 1981). (b) Three-dimensional constant-frequency surface in wave-vector space for the slow transverse mode in GaAs. (c) Cross section of the same surface in the [100] plane and showing the focusing of the energy flux. The energy flux is determined by the group velocity $\partial\omega(k)/\partial k$.

verse modes are primarily shear waves and have a lower velocity. The two shear waves in the ⟨111⟩ and ⟨100⟩ directions have the same velocity. It is also clear from Fig. 3(a) that the relative intensity of the T and L pulses changes by more than an order of magnitude depending on the orientation. It is about 2 in the ⟨111⟩ direction and about 20 in the ⟨100⟩ direction. This focusing of the phonon intensity is due to the elastically anisotropic nature of the solid (see, for example, Maris, 1971). In anisotropic crystals the phase velocity $V=\omega(\mathbf{k})/|(\mathbf{k})|$ depends on the direction of the wave vector **k** and can be calculated from the constant-frequency or "slowness" surface $\omega(\mathbf{k})=\text{const}$. This surface is shown for the ST mode in GaAs in Fig. 3(b). The group velocity $V_g=\partial\omega(\mathbf{k})/\partial(\mathbf{k})$, which defines the direction of energy flow for the [100] plane, is shown in Fig. 3(c). In this case the energy flow is strongly enhanced or focused along the [100], [010], and [001] symmetry axes, which explains the high T-mode intensities relative to the L mode observed in the ⟨100⟩ direction in Fig. 3(a).

The focusing of phonon beams can be numerically calculated from a three-dimensional slowness surface such as that shown in Fig. 3(b). Maris (1971) showed that the energy flux enhancement A is determined by the ratio of the solid angles in **k** space and energy flux in ray (**r**) space: $A=d\Omega^k/d\Omega^r$. For a spherical surface $A=1$. For an anisotropic surface A can vary over a wide range, depending on the curvature of the surface.

Another major effect in determining focusing patterns is the effect of phonon dispersion. Figure 4 shows the frequency versus wave-vector relation for the TA mode in the [111] propagation direction of several III-V compound semiconductors. Unlike photons, the curvature of the frequency–wave-vector relation causes a reduction in the group velocity as a function of frequency. Furthermore, the focusing patterns now must depend markedly on frequency.

The effects of dispersion on phonon focusing in III-V semiconductors has been observed by Wolfe and co-workers (see Wolfe, 1989). An x-y laser scanning system is used to excite the heat pulse on one end of the sample under study. A bolometer or a superconducting tunnel junction is used as the detector. Since a tunnel junction has a threshold detection capability, the leading edge of the arriv-

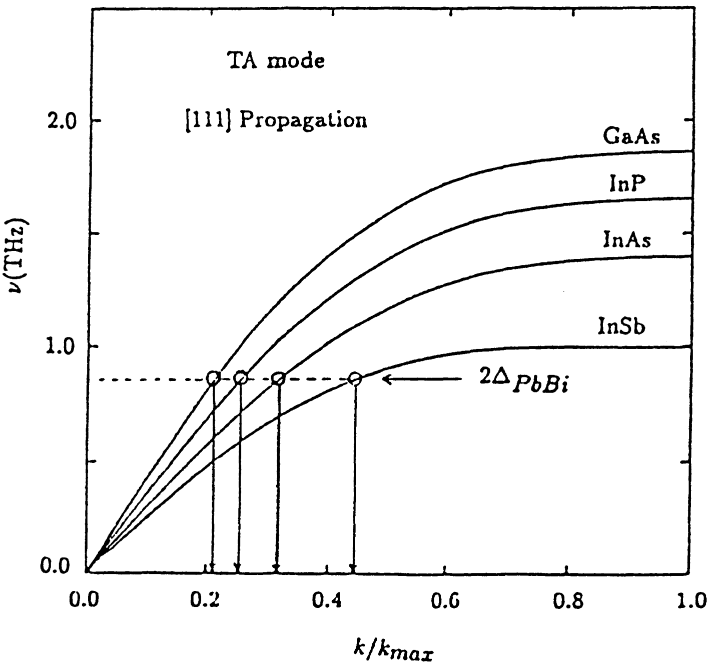

FIG. 4. Dispersion curves for several III-V semiconductors. The onset frequency of a Pb:Bi superconducting tunnel junction is also shown (after Wolfe, 1989).

ing pulse corresponds to the frequency of the energy gap of the superconductor. In the case of Pb:Bi alloys this corresponds to nearly 0.8 THz (see Table 1 and Fig. 4). Figure 5 shows the experimental phonon images for InSb (see Hebboul *et al.*, 1986) with a broadband Al bolometer [Fig. 5(a)] and with a Pb tunnel junction [Fig. 5(b)]. The two images are very different. The images for the tunnel junction require detailed knowledge of the dispersion curve for the solid. This in turn is a strong function of the interatomic forces that determine the lattice dynamics for the solid. By such focusing studies one can differentiate between different atomic force field models for compound semiconductors, as has been shown by Wolfe (1989).

1.2 Second Sound

We now turn to the situation when phonon–phonon scattering becomes important. This becomes true when the temperature is raised, i.e., when the phonon–phonon scattering time (τ) is less than the ballistic propagation time. The most likely phonon–phonon scattering process is a three-phonon process of the type

$$\omega_1+\omega_2\rightleftarrows\omega_3, \qquad (1)$$

$$\mathbf{k}_1+\mathbf{k}_2\rightleftarrows\mathbf{k}_3, \qquad (2)$$

where Eqs. (1) and (2) reflect conservation of energy and momentum. If the frequency of one of the phonons is large, then the final momentum can lie outside the Brillouin zone, and the process results in a directional change of the phonon (momentum nonconserving) with the extra momentum being taken up by the lattice. Such umklapp processes give rise to a decay in the heat flow and to a finite thermal conductivity, even in a perfect dielectric solid. However, if ω_1 and ω_2 are such that the final momentum lies within the Brillouin zone, then the process is referred to as a normal process (momentum conserving) and does not lead to diffusion of heat.

The propagation of a heat pulse (with a temperature T) down a bar of a dielectric solid (see, e.g., Ackerman and Guyer, 1968), can be described by two microscopic equations:

(a) energy conservation,

$$C_v\frac{\partial T}{\partial t}+\nabla\mathbf{Q}=0; \qquad (3)$$

(b) modified heat current equation,

$$\tau\frac{\partial Q}{\partial t}+\mathbf{Q}=-K\nabla T. \qquad (4)$$

Here T is the temperature, which is a function of the propagation direction and the time t, K

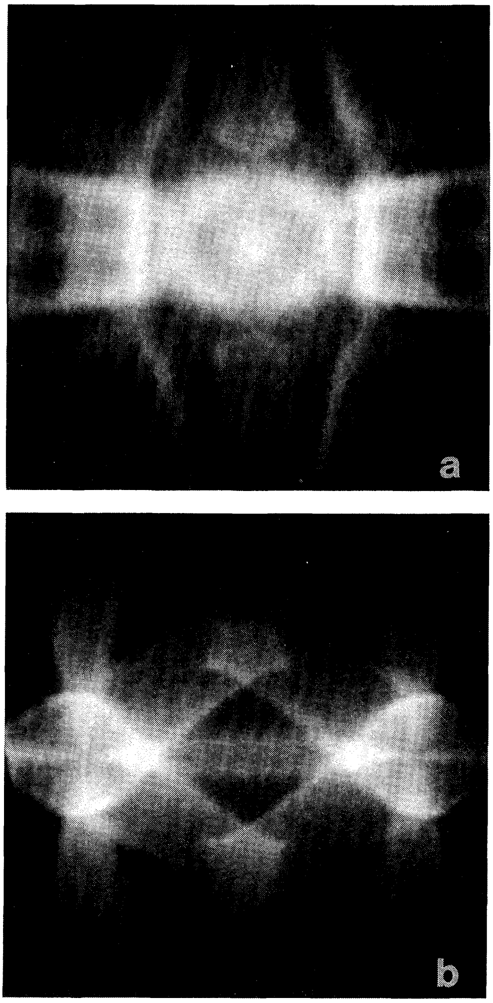

FIG. 5. Experimental phonon images for a 2-mm-thick InSb crystal: (a) for an Al bolometer detector: (b) for a Pb tunnel-junction detector (after Hebboul *et al.*, 1986).

is the thermal conductivity, τ is the relaxation time characterizing the response of the heat current, and C_v is the specific heat at constant volume. In the event $\tau \approx 0$, Eq. (4) describes the well-known Fourier heat law.

Equations (3) and (4) can be combined to lead to

$$\frac{\partial^2 T}{\partial t^2}+\frac{1}{\tau}\frac{\partial T}{\partial t}-\frac{K}{C_v\tau}\nabla^2 T=0. \tag{5}$$

The relaxation time τ consists of two parts: a part τ_N that refers to normal processes, and a term τ_u that refers to umklapp processes. The description of the heat flow when both τ_N and τ_u are finite and comparable to the propagation time is quite complex and is beyond the scope of this article. However, in the limit $\tau_u \to \infty$, Eq. (5) has the simple form

$$\frac{\partial^2 T}{\partial t^2}+\tfrac{1}{3}v_1^2\nabla^2 T=0 \tag{6}$$

for which we have plane-wave solutions for the temperature, $T=T_0+\delta T\exp[i(\mathbf{k}\cdot\mathbf{x}-\omega T)]$, with a dispersion relation $\omega^2=k^2v_1^2/3$; a true temperature wave propagates in the solid with a velocity $v_{\mathrm{II}}=v_1/\sqrt{3}$. Here, we have for simplicity assumed that the phonon velocity v_1 in the solid is isotropic and independent of polarization. Such temperature waves in solids were first predicted by Ward and Wilks in 1951 and experimentally observed in solid helium by Ackerman and Guyer in 1966 (see Ackerman and Guyer, 1968). The wavelike solutions predicted above get quickly damped in ordinary crystals where momentum-destroying defects, including naturally occurring mass defects, cause the heat current to decay with time. Since the velocity of propagation for temperature waves is considerably smaller (by a factor of $\sqrt{3}$) than that for ordinary (first) sound, the arrival time at a receiver is correspondingly delayed. This delay led to the coining of the name "second sound" for temperature waves. Even in pure solids there is only a narrow temperature window (where τ_N is short and τ_u is long) where second sound can be observed. To date second sound has been observed in solid ^{4}He, solid ^{3}He, NaF, and Bi (see, for example, Narayanamurti *et al.*, 1975).

The transition from ballistic phonon flow to second sound occurs in a temperature regime where $\omega\tau_N \ll 1$, i.e., when there are enough momentum-conserving phonon–phonon collisions to describe a true temperature in the gas of phonons. The dispersion relation for finite values of $\omega\tau_N$ can be approximately written as (see, e.g., Narayanamurti *et al.*, 1975):

$$k^2v_{\mathrm{II}}^2=\tfrac{1}{3}k^2v_{\mathrm{I}}^2=\omega^2+\frac{i\omega k^2\tau_N(v_{\mathrm{I}}^2-v_{\mathrm{II}}^2)}{1-i\omega\tau_N}\,. \tag{7}$$

Figure 6 shows the dispersion relation for the heat-pulse velocity from Eq. (7). By measuring the dispersion relation, one can get information on the temperature and/or frequency dependence of the normal-process

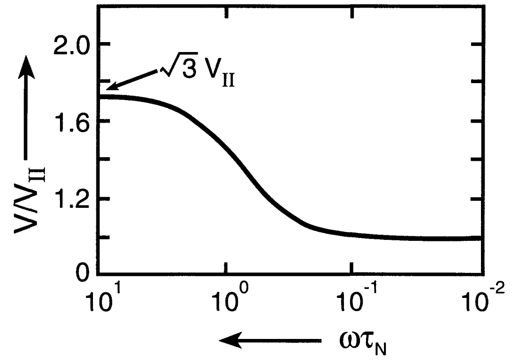

FIG. 6. Dispersion relation for heat-pulse velocity v as a function of $\omega\tau_N$ obtained using Eq. (7). v_{II} is the second-sound velocity and is reached only for very small values of $\omega\tau_N$.

relaxation time τ_N. The limiting second-sound velocity v_{II} is only reached for small values of $\omega\tau_N$ ($<10^{-1}$).

Figure 7 shows a series of traces as a function of temperature for heat-pulse propagation in solid ^{4}He (after Narayanamurti and Dynes, 1975). At very low temperatures one sees a strong transverse ballistic pulse that arrives at the ballistic velocity. As the temperature is raised from 0.07 to about 0.5 K, the pulse broadens and slows down and arrives at later times. By 0.5 K the normal-process collision time is short enough and the second-sound mode is fully developed and has many echoes [Fig. 7(f)]. Figure 8 shows the temperature dependence of the peak velocity. From such measurements detailed information on the temperature dependence of τ_N can be obtained.

2. PHONON SPECTROSCOPY

A great deal of work has been done in which ballistic heat-pulse and tunnel-junction techniques were used to study electronic states of donors, acceptors, and defects in semiconductors and insulators. The phonon coupling to such states depends markedly on the nature of the defect and on frequency, polarization, and propagation direction. Experiments that yield selective attenuation of groups of phonons can serve as a valuable tool for determining energy states and their structural symmetry. This field has been recently reviewed by Wybourne and Wigmore (1988). In the following we will consider a few illustrative examples.

2.1 The Case of Donors and Acceptors in Semiconductors

There are a large number of reported accounts of phonon studies applied to semiconductors. This is in large measure due to the technological importance of semiconductors and the important role defects play in the operation of electronic and optoelectronic devices.

Shallow donor and acceptor states are hydrogenlike and are constructed from the Bloch states of the conduction- or valence-band edge, respectively, multiplied by the approximate envelope function describing the location of the defect. Particularly interesting are donor states. Si and Ge, for example, are indirect

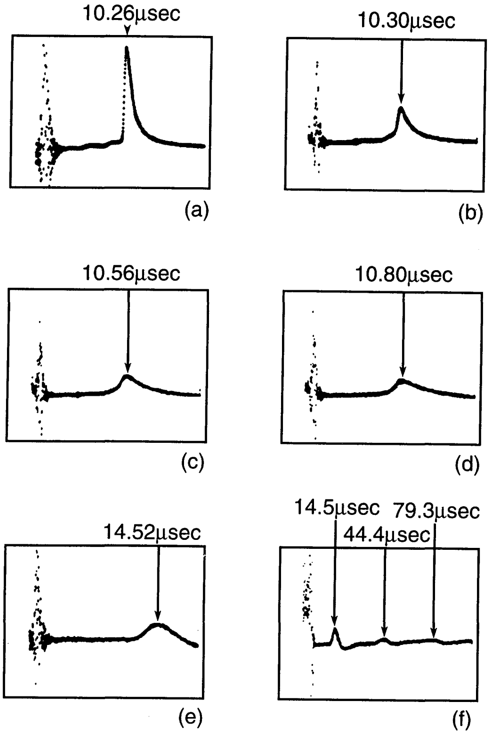

FIG. 7. Temperature dependence of heat-pulse propagation in solid ^{4}He. Molar volume ~21.0 cm^3/mol. (a) T=0.07 K; (b) T=0.18 K; (c) T=0.245 K; (d) T=0.27 K; (e) T=0.495 K; (f) T=0.495 K. The arrows indicate time of arrival of the peak. Time t=0 is determined from the initial electromagnetic ringing. In case (f) the time scale is such that two echoes of the second-sound pulse are readily seen. Propagation length was 2.33 mm (data from Narayanamurti and Dynes, 1975).

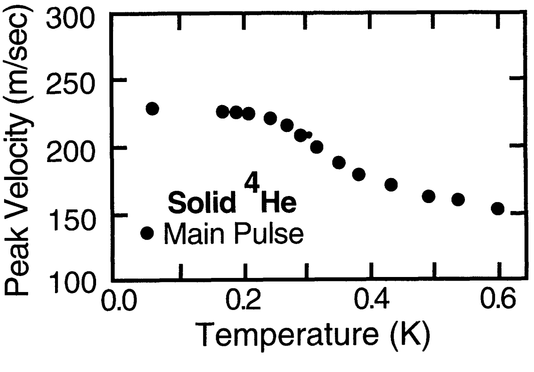

FIG. 8. Temperature dependence of the peak heat-pulse velocity for solid ^{4}He. Molar volume ~21.0 cm^3/mol. The measured second-sound velocity is about 157 m/s.

band-gap semiconductors whose conduction-band minima are sixfold and fourfold degenerate, respectively. The degeneracy of the ground state is lifted by interactions (the so-called valley–orbit interaction). The splitting of the ground and excited states is typically a few meV, which makes these systems ideal for acoustic-phonon resonance studies. Furthermore, the energy levels can be varied by applying pressure on the system.

The first direct phonon resonance absorption experiments in which superconducting tunnel junctions were used as sources and detectors of "quasimonochromatic" phonons were performed by Dynes, Narayanamurti, and Chin (see Narayanamurti, 1981). They used the pressure dependence of the energy levels of the shallow donor antimony in germanium to obtain a resonance absorption of 2Δ phonons generated by a superconducting tin tunnel junction. In Fig. 9 we show a schematic of the experimental arrangement as well as the theoretically predicted variation of the energy levels as a function of uniaxial pressure applied along the ⟨111⟩ direction. The generator was biased at a voltage $eV = 2\Delta = 1.2$ meV so that only recombination radiation was emitted by the junction. At a value of stress corresponding to an energy splitting of 1.2 meV a strong decrease in the phonon transmission was observed. Measurements as a function of phonon polarization and energy yielded the symmetry of the states and the strength of the defect phonon coupling. The observation of the resonance absorption of phonons in this experiment also served as the first direct experimental confirmation of the quantum nature (as distinct from heat) of the recombination process in superconducting junctions.

As a second example we consider the case of deep donor-related states in GaAlAs. In liquid-phase epitaxial growth of this material (see Lang, *et al.*, 1979), donors are believed to form a complex referred to as a *DX* center, where *D* denotes a donor and *X* the unknown second constituent. The *DX* center is believed to be strongly coupled to the lattice, so that when the center captures an electron it lowers its energy substantially. The electronic energy is shown as a function of the lattice configuration *Q*, on the right side of Fig. 10(a). The center can be emptied (ionized) by optical excitation at energy E_{opt}. The combined effect of the electron–lattice coupling and the defect potential produces a bound state when the defect is occupied.

Direct experimental confirmation of the strong defect–lattice coupling and the symmetry of the *DX* centers was observed for the first time by ballistic phonon attenuation tech-

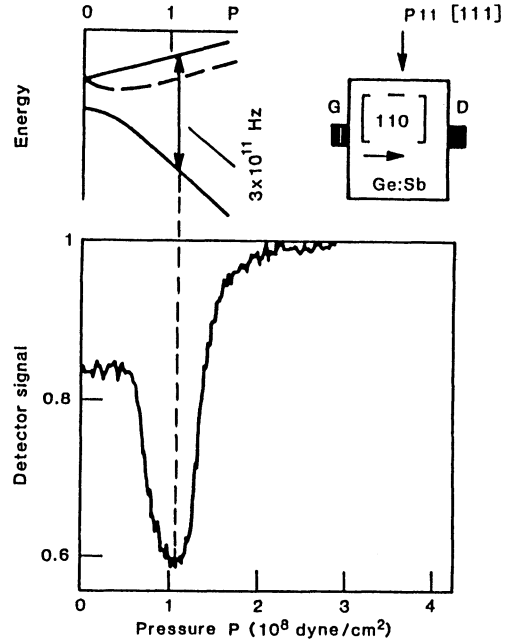

FIG. 9. Transmission of 1.2-meV phonons through a ~4-mm-thick sample of germanium doped with about 2×10^{15} antimony donors as a function of uniaxial pressure (after Narayanamurti, 1981). The top of the figure shows the calculated energy levels of Sb donors as a function of energy. When the energy-level splitting equals the 2Δ phonon energy, a resonance absorption is observed.

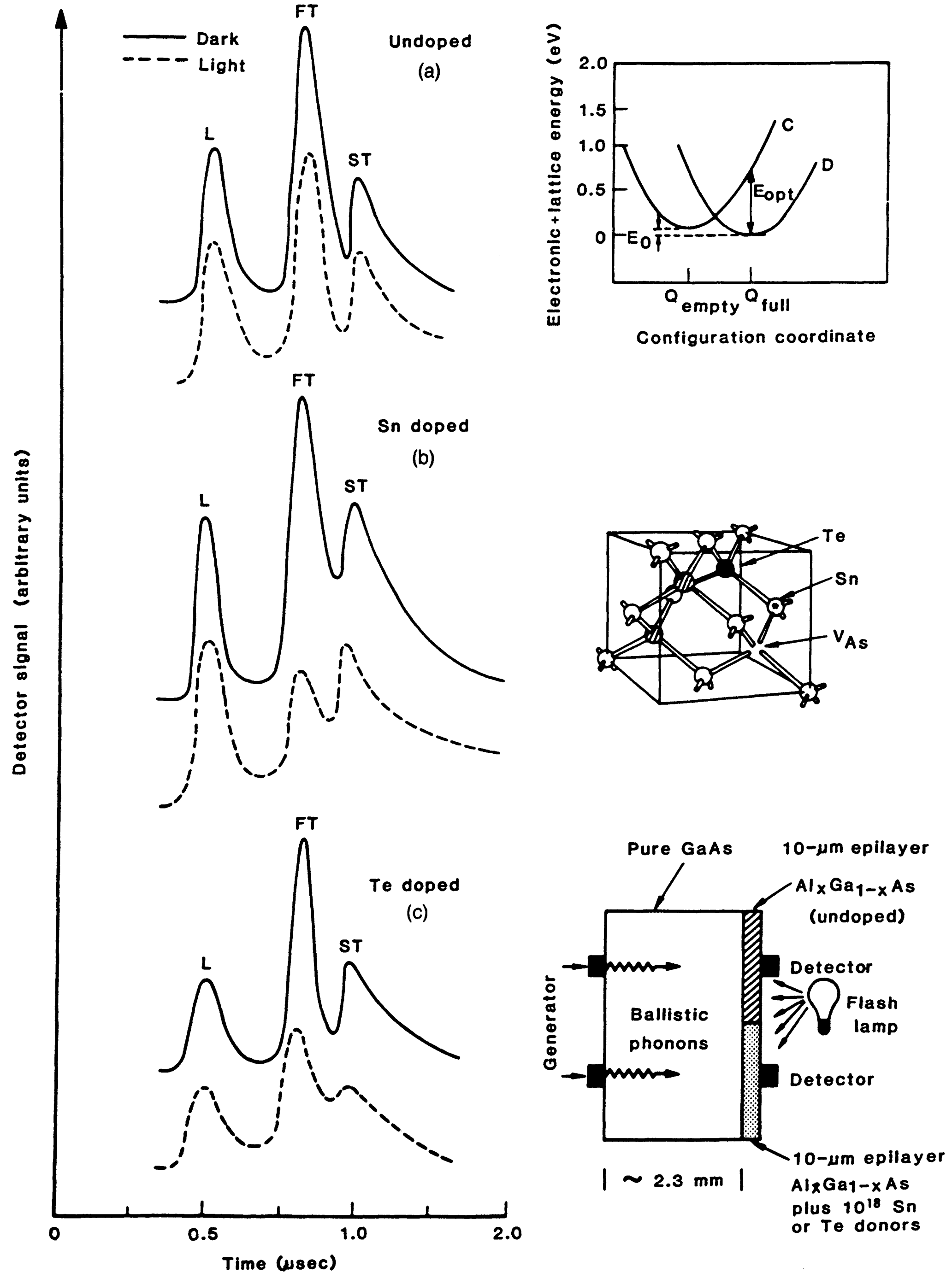

FIG. 10. Ballistic phonon intensities for three different samples of $Al_{0.5}Ga_{0.5}As$: (a) nominally undoped material; (b) with $\sim 10^{18}$ per cm^3 Sn donors; and (c) with $\sim 10^{18}$ per cm^3 Te donors. Solid lines are data taken in the dark; dashed lines are data taken after photoexcitation. The propagation direction is $\langle 110 \rangle$. The intensities of the ballistic phonons reveal selective attenuation of certain groups of transverse modes depending on the symmetry of the donor-related photoconductivity center. A configuration-coordinate model, a crystal model of GaAs, and a schematic of the experimental arrangement are shown on the right-hand sides of (a), (b), and (c). (After Narayanamurti, 1981.)

niques (see Narayanamurti *et al.*, 1979a). The donors used in the experiments were either tellurium (which takes the place of As in the lattice) or tin (which takes the place of gallium). A thick (2.5-mm) sample of GaAs was used as a transmitting medium, on one end of which was grown a $Ga_{0.5}Al_{0.5}As$ epitaxial layer of typical thickness 10 μm. Doped and undoped layers were grown side by side so that the effects of doping could be monitored. Photoexcitation by means of a flash lamp was used to change the occupation of the *DX* centers *in situ* in the epitaxial layers. The propagation direction was ⟨110⟩.

Figure 10 also shows typical ballistic pulses observed in the dark and after photoexcitation. In the case of tin donors strong attenuation of FT modes was observed while attenuation of the ST mode was negligible. In the case of tellurium donors both T modes suffered attenuation. From group-theoretical arguments the selective attenuation was interpreted as arising from trigonal centers in the case of tin doping and orthorhombic centers in the case of tellurium doping. This measurement is particularly useful, as other techniques such as electron spin resonance are difficult to use in this system.

The number of defects that have been studied include the acceptors In, B, Al, and Ga in Si; the donors P, Li, Sb, As, and P in Si; the acceptors Ga and In and Ge; and the different charge states of Cr impurities in GaAs. Detailed references can be found in the review by Wybourne and Wigmore (1988).

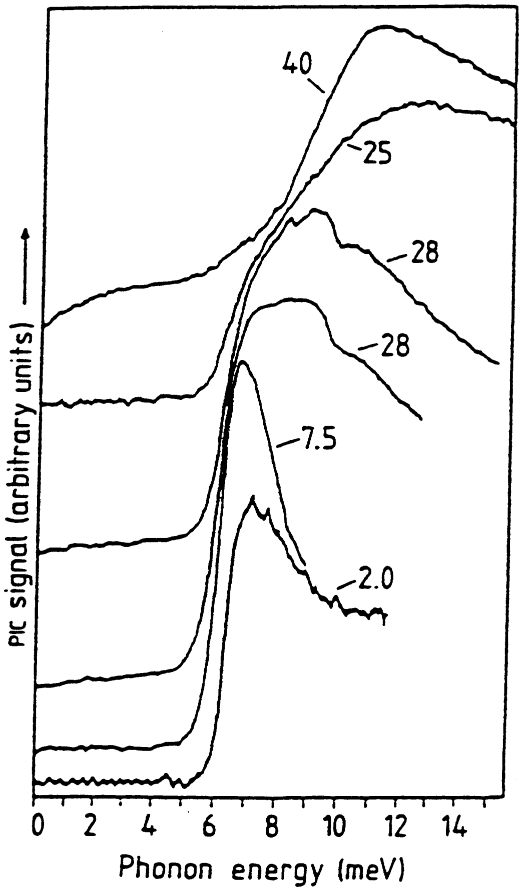

FIG. 11. Phonon-induced conductivity spectra from indium-doped silicon of various concentrations indicated in units of 10^{15} per cm^3. A threshold at the binding energy of 5.9 meV is clearly seen. (After Lassmann and Burger, 1986.)

A particularly interesting variant of the experiments described above involves measurements of the phonon-induced conductivity (PIC) to study the ionization energies of neutral donors and acceptors in Si that have a weakly bound carrier. This has analogy with the photoelectric effect for photons impinging on metals. The excitation of bound states by nonthermal phonons has been well developed as a spectroscopic tool by Lassmann and Burger (1986). The relaxation phonons from a superconducting Al tunnel junction are used to measure the PIC. A typical spectrum is shown in Fig. 11 for various In concentrations in silicon. The threshold in the spectrum at 5.9 MeV ($\sim$ 1.5 THz) corresponds to the binding energy of In in Si. The technique is very sensitive and can be used to detect impurity concentrations in the 10^{12} per cm^3 range and to perform accurate measurements of binding energies of donors and acceptors in Si and Ge.

2.2 Phonon Scattering by Dislocations

So far we have primarily considered scattering and spectroscopy by point impurities. In many crystals, dislocations, which are extended structural defects, are known to influence the thermal transport properties significantly. The static strain fields of the dislocations serve to scatter the phonons. In addition, incident phonons can cause dislocations to vibrate, thereby representing an exchange of energy. The inverse scattering time for slow transverse modes has been theoretically postulated to be

$$1/\tau_{st} \sim N_D \sigma^2 B^2 \gamma,$$

where N_D is the dislocation density per unit area, σ the Grüneisen constant, B the Burgers vector, and γ the phonon frequency (see Kneezel and Granato, 1982). Scattering of other modes is very weak. The preferential scattering of slow transverse phonons and the magnitude of the scattering has been studied in great detail by Northrop *et al.* (1983) for the case of LiF.

A particularly interesting case has been studied by the author. In the case of direct-band-gap semiconductors, such as GaAs, it is known that dislocations can reduce the luminescence efficiency of light-emitting diodes (LEDs) and lasers. This is because electron–hole recombination across the band gap with light emission is interrupted by nonradiative processes where the energy is released in the form of heat instead of light. This drop in efficiency was for many years presumed to be due to nonradiative processes occurring at defects or dislocations. Again phonon techniques provide a very direct way of monitoring such phenomena. Figure 12 shows the phonon emission from GaAs *p-n* junctions (see Narayanamurti *et al.*, 1978). The difference between the "control" and the dislocated junctions is due to enhancement of nonradiative recombination at dislocations. It is clear that the dislocated junction shows enhanced phonon output, particularly for the ST mode.

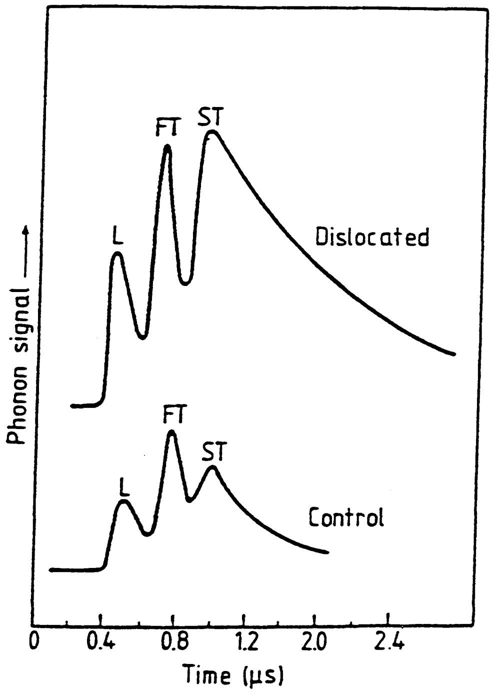

FIG. 12. Phonon emission from GaAs *p-n* junctions. The difference between the "control" and the dislocated junctions is due to enhancement of nonradiative recombination at dislocations. (After Narayanamurti, 1978.)

2.3 Phonon Scattering by Surfaces and Interfaces

The role of surfaces in determining the reflectivity of incident phonons has many analogies to photons. The quality of the surface, i.e., the degree of perfection and smoothness can be expected to have a profound effect. In the case of high-frequency phonons of typical wavelength of a few hundred angstroms, even ordinary "optically" smooth surfaces have been found to cause diffuse scattering and reflection. The damage just beneath the surface introduced during the preparation process has been shown to be very important. The effects of etching and annealing have been shown to be profound (for a review see Wigmore and Wybourne, 1988). For most ordinary surfaces the transmission of high-frequency phonons is found to be much greater than expected, while the reflectivity is much smaller. This anomaly, the so-called Kapitza anomaly, has been the subject of much study for many years.

Trumpp and co-workers (see Trumpp and Eisenmenger, 1977), for example, quantitatively showed that only 10% of the 2Δ phonons generated by a superconducting tin tunnel junction at 280 GHz could traverse the interface with a silicon crystal without degradation into heat. Reflection measurements show a serious degradation of the signal after one reflection. Thus, unlike conventional ultrasonics, it is virtually impossible to observe more than one echo in heat-pulse and/or tunnel-junction experiments.

In the next section we will, however, discuss the case of multiple interfaces that can be prepared to a high degree of perfection using modern crystal-growth techniques, and that obviate the problems mentioned above. This should allow the creation of a new generation of devices of potential use for characterization of materials in analogy with optics.

2.4 Phonons in Superlattices—Mirrors, Filters, and Related Devices

To control the reflectivity and/or transmissivity of photons, multilayer films of precise thickness and refractive index are used. For photons in the optical wavelength, $\lambda \sim 5000$ Å, the films constituting the multilayers are about $\lambda/2$ or 2500 Å thick (see, for example, Brillouin, 1946). In the case of phonons of wavelength 100 Å, the thicknesses are in the 50-Å range. Thus extremely high-quality films are required. Fortunately in the case of compound semiconductors, because of their technological importance, a great deal of effort has been spent in growing superlattices of a controlled period (see, for example, Panish, 1980). This has been made possible through the development of molecular beam epitaxy (MBE), which involves the controlled growth, layer by layer, of the constituent semiconductor material under ultrahigh-vacuum (10^{-10} Torr) conditions. Furthermore, the constituent materials are lattice matched (epitaxially grown) so that interfacial degradation is negligible. This is the case of GaAs/AlGaAs.

I now describe a novel thin-film spectrometer based on the selective transmission of high-frequency phonons through a superlattice of GaAs/$Al_{0.5}Ga_{0.5}As$. The filtering action arises from the selective reflection by the superlattice of phonons of wavelength λ_0 that fulfill the Bragg condition $\lambda_0 = 2d$, where d is the superlattice period (Brillouin, 1954). Quantitatively the magnitude of the selective filtering action depends on the ratios of the acoustic impedances ($Z = \rho v$, where ρ is the density and v the sound velocity) of the constituent materials of the superlattice. The acoustic impedance plays the same role for phonons as the refractive index plays in the case of photons. In view of the close analogy with optical dielectric mirrors, an appropriately designed thin-film superlattice may be called a dielectric phonon mirror. Reflectivity calculations, including the effects of small amounts of disorder, are described in the literature (see Narayanamurti *et al.*, 1979b).

The inset of Fig. 13 shows a schematic of the experimental arrangement used to test some of these ideas. A Sn tunnel junction is used as a source of phonons, which pass through a GaAs/$Al_{0.5}Ga_{0.5}As$ superlattice structure of period 122 Å before detection by an Al tunnel junction. The filter resonance thus corresponds to a wavelength of about 240 Å and a frequency of about 2.2×10^{11} Hz. Figure 13 shows the observed transmission (points) as a function of phonon frequency; also shown are the corresponding phonon wavelengths. The solid line corresponds to the theoretically calculated filter function for an acoustical impedance ratio $Z_1/Z_2 \approx 0.88$ and an overall linewidth of $\sim 0.2 \times 10^{11}$ Hz. Most of the linewidth arises from the width of the superconducting energy gap of the generator and detector used in these experiments. Experiments with superlattices with other periods show the position of the resonance changes as expected with the superlattice period.

The experiments described above involved propagation in the normal direction. For arbitrary propagation directions the Bragg condition becomes $n\lambda = 2d \cos\theta$, where θ is the angle between the phonon propagation direction and the normal direction. Figure 14 shows the differential phonon image for longitudinal phonons of frequency 850 GHz passing through an $In_{0.15}As_{0.85}As$/AlAs superlattice with an individual layer thickness of 20 Å consisting of 40 periods. The dark ring represents the selective reduction of 850-GHz phonons by the superlattice (Hurley *et al.*, 1987).

These results show that phonons of wavelength ~ 100 Å suffer relatively little interface degradation even though they had to pass through 50–100 layers of the multilayer structure. This suggests that interfaces produced by MBE are nearly ideal, and one can think of using superlattice phonon mirrors as phonon spectrometers for the manipulation of phonon diagnostic beams in semiconductors and for the further development of what one might call "phonon optics."

Clearly one can devise many interesting structures consistent with wave propagation in periodic structures. Much work has been done in this area over the last few years. Phonon transmission through quasiperiodic systems and through multisuperlattice structures consisting of two kinds of superlattices has been studied (Tamura, 1989). A particularly interesting example (Mizumo and Tamura, 1991) is the case of bulk material sandwiched between two identical periodic superlattices. This system is analogous to a double-barrier system for electrons where resonant tunneling can occur. Mizumo and Tamura have shown that sharp resonances that involve resonant transmission without

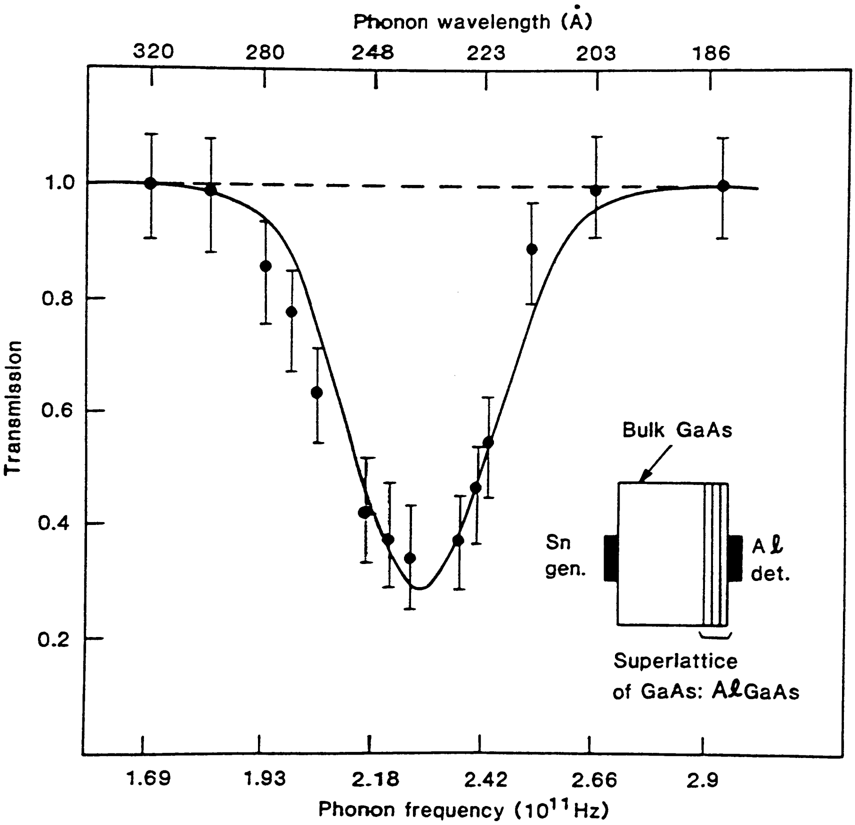

FIG. 13. Phonon transmission through a GaAs/$Al_{0.5}Ga_{0.5}As$ superlattice of period 122 Å and 50 double layers. Dots correspond to experimental points and the solid line corresponds to a theoretical fit for an acoustic impedance ratio $Z_1/Z_2 \sim 0.88$ and an overall linewidth of $\sim 0.24 \times 10^{11}$ Hz. (After Narayanamurti *et al.*, 1979a,b.)

attenuation can occur in the gap. Thus it appears quite plausible that phonon resonators can be designed, and that even a larger variety of phonon optics devices than those originally proposed by the author (see Chin *et al.*, 1982) is possible.

3. OPTICAL TECHNIQUES

We have so far primarily concerned ourselves with heat-pulse (occasionally with laser generation of heat) and quasimonochromatic superconducting tunnel-junction thin-film techniques. These techniques are analogous to ultrasonic techniques albeit in a much higher frequency domain. Optical techniques, on the other hand, generally involve excitation in the bulk of impurity levels (usually transition-metal ions in dielectric crystals), which lose energy after excitation through phonons. The energy level can usually be tuned by a magnetic field and therefore can serve as an *in situ*

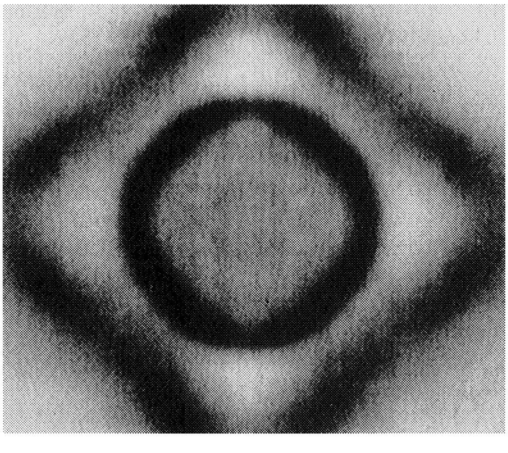

FIG. 14. Differential phonon image for longitudinal phonons of frequency 850 GHz passing through an $In_{0.15}Ga_{0.85}As$/AlAs superlattice. The dark ring represents the selective reduction of 850-GHz phonons by the superlattice. (After Hurley *et al.*, 1987.)

spectrometer. We do not discus these techniques further since they are beyond the scope of this article (see, e.g., the review by Renk, 1985). In the ensuing we will, however, briefly discuss applications primarily deriving from developments in the field of opto-acoustic spectroscopy and generation of phonons using picosecond optical techniques.

3.1 Opto-Acoustic Spectroscopy

We have already had occasion to mention briefly the studies of nonradiative (heat) generation measured by bolometric techniques. The field of opto-acoustic spectroscopy involves the measurement of optical absorption spectra through the detection of acoustical signals generated when the absorbed optical energy decays through nonradiative channels. The use of tunable lasers combined with sensitive acoustic techniques has led to the measurement of extremely weak absorption spectra in condensed matter (see, for example, Patel, 1981).

This technique has been further developed using superconducting tunnel junctions to study the optical absorption of adsorbed molecules, such as tetracene, on a silicon surface (see Dietsche, 1986). The extremely high sensitivity of tunnel-junction detectors has been used to study submonolayer coverages of adsorbed molecules.

3.2 Picosecond Optical Techniques

Conventional heat-pulse techniques are limited to samples that show macroscopic phonon mean free paths (at least fractions of a millimeter) because of the time domain in which such phonons are studied (10^{-7}–10^{-8} s). Such macroscopic mean free paths, as we have seen, are usually achievable only at low temperatures (below 10 K).

To make measurements at high frequencies and high temperatures one has to be prepared to study phonons with very short mean free paths (100–1000 Å). Thus if one wants to do phonon experiments in "real time," time resolution in the 10^{-11}–10^{-12} s range is required. This is not possible with conventional electronics. Maris and co-workers (see Maris, 1986) have used picosecond pulses to excite phonons in thin films at room temperature. Figure 15 shows a schematic of the experimental arrangement to generate and detect phonons in dielectric and semiconducting films. A picosecond optical pump pulse is used to generate a thermal stress (within an absorption distance ~300 Å), which in turn launches a coherent strain wave into the film. Since the optical constants of the material depend on strain, one can envision a change in the reflectivity of the film that depends on the strain. The amplitude of the strain pulse is measured through changes in the reflectivity as a function of time by means of a weak probe pulse also of duration 1 ps. This technique has been used to study phonon mean free paths in several amorphous materials (SiO_2, *a*-As_2Te_3, etc.).

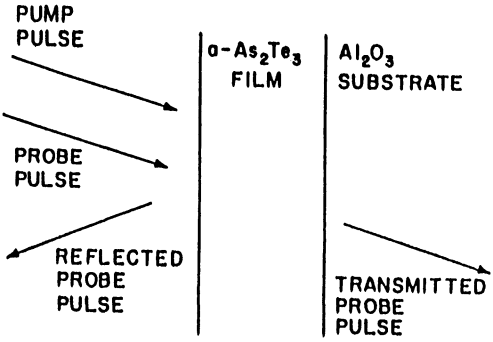

FIG. 15. Schematic diagram of the experiment to generate and detect phonons in the picosecond regime in a semiconducting film (after Maris, 1986).

4. CONCLUSIONS

In this review I have attempted to survey some of the applications of heat-pulse and quasimonochromatic phonon propagation in solids. Clearly the principal applications lie in the area of spectroscopy of both bulk material and of surfaces and interfaces.

The field has advanced sufficiently over the last 15 years that it now possesses many of the traditional aspects of optical spectroscopy experiments: monochromatic generation, rectilinear propagation, polarization dependence, manipulation and steering using superlattice filters, and focusing based on wave propagation in an anisotropic medium. Recent developments in the study of dispersive wave propagation and improved time resolution suggest that virtually the entire range of thermal phonon frequencies can now be studied.

Much work still remains to be done. Study of phonon interactions with lower-dimensional systems, superlattices, and one- and/or zero-dimensional systems clearly would be an exciting arena for study. Preparation of phonon optics devices based on novel double-barrier tunneling systems and their application to nanoscale physics appears feasible. Further developments and evolution in the field should lead to improved understanding of stress-sensitive properties on a truly atomic scale. When combined with molecular beam epitaxy and nanoscale fabrication, phonon microscopy in the 10–100-Å range appears to be a valuable tool for the study of materials.

GLOSSARY

Acceptor: Defect in a semiconductor that releases holes into the valence band.

Donor: Defect in a semiconductor that releases electrons into the conduction band.

Epitaxy: Crystal growth under lattice-matching conditions that permits crystalline growth without defects.

Molecular-Beam Epitaxy: Crystal-growth method that uses controlled deposition of materials in ultrahigh vacuum to grow materials layer by layer.

Phonon: Quantized lattice vibration.

Superconducting Energy Gap: Binding energy for pairing of electrons of opposite spins (Cooper pair), which leads to superconductivity.

Superconducting Tunnel Junction: Sandwich of two superconductors separated by an oxide barrier.

Superlattice: Periodic structure of multilayer films with an artificially introduced periodicity.

Transition Temperature T_c: Temperature below which material is superconducting.

List of Works Cited

Ackerman, C. C., Guyer, R. A. (1968), *Ann. Phys. (N.Y.)* **50**, 128–182.

Bömmel, H. E., Dransfeld, K. (1960), *Phys. Rev.* **117**, 1245–1252.

Brillouin, L. (1946), *Wave Propagation in Periodic Structure,* New York: McGraw-Hill.

Chin, M. A., Narayanamurti, V., Störmer, H. S. (1982), U.S. Patent No. 4349796.

Dietsche, W. (1986), in: A. C. Anderson, J. P. Wolfe (Eds), *Phonon Scattering in Condensed Matter,* Berlin: Springer, pp. 366–373.

Eisenmenger, W. (1976), in: W. P. Mason, R. N. Thurston (Eds.), *Physical Acoustics,* New York: Academic.

Hebboul, S. E., van Harlingen, D. J., Wolfe, J. P. (1986), in: A. C. Anderson, J. P. Wolfe (Eds.) *Phonon Scattering in Condensed Matter,* Berlin: Springer, pp. 309–311.

Hurley, D. C., Tamura, S., Wolfe, J. P., Morkoc, H. (1987), *Phys. Rev. Lett.* **58**, 2446–2449.

Ilukor, J., Jacobsen, E. H. (1966), *Science,* **153**, 1113–1114.

Kneezel, G. A., Granato, A. V. (1982), *Phys. Rev. B* **25**, 2851–2866.

Lang, D. V., Logan, R. A., Jaros, M. (1979), *Phys. Rev. B* **19**, 1015–1030.

Lassmann, K., Burger, W. (1986), in: A. C. Anderson, J. P. Wolfe (Eds.) *Phonon Scattering in Condensed Matter,* Berlin: Springer, pp. 116–122.

Maris, H. J. (1971), *J. Acoust. Soc. Am.* **50**, 812–818.

Maris, H. J. (1986), in: A. C. Anderson, J. P. Wolfe, (Eds.), *Phonon Scattering in Condensed Matter,* Berlin: Springer, pp. 374–380.

Mizumo, S., Tamura, S. (1991), *Phys. Rev. B.* **45**, 13 423–13 429.

Narayanamurti, V. (1981), in: K. E. Gray (Ed.), *Proceedings of the NATO Advanced Study Institute on Nonequilibrium Superconductivity Phonons and Kapitza Boundaries,* New York: Plenum.

Narayanamurti, V., Dynes, R. C. (1975), *Phys. Rev. B* **12**, 1731–1738.

Narayanamurti, V., Dynes, R. C., Andres, K. (1975), *Phys. Rev. B* **11**, 2500–2524.

Narayanamurti, V., Logan, R. A., Chin, M. A. (1978), *Phys. Rev. Lett.* **40**, 63–66.

Narayanamurti, V., Logan, R. A., Chin, M. A. (1979a), *Phys. Rev. Lett.* **43**, 1536–1539.

Narayanamurti, V., Störmer, H. L., Chin, M. A., Gossard, A. C., Weigmann, W. (1979b), *Phys. Rev. Lett.* **43**, 2012–2016.

Northrop, G. A., Cotts, E. J., Anderson, A. C., Wolfe, J. P. (1983), *Phys. Rev. B* **27**, 6395–6408.

Nowick, A. S., Berry, B. S. (1972), *Anelastic Relaxation in Crystalline Solids,* New York: Academic.

Panish, M. B. (1980), *Science* **208**, 916–922.

Patel, C. K. N. (1981), in: E. Lüscher, P. Korpiun, H. J. Coufal, R. Tilgner (Eds.), *Proceedings of the 1st International Conference on the Photoacoustic Effect,* Wiesbaden: Vieweg, pp. 21ff.

Quate, C. F. (1979), *Sci. Am.* **241**, 62–71.

Renk, K. F. (1985), in: W. E. Bron (Ed.), *Nonequilibrium Phonon Dynamics,* New York: Plenum, pp. 59ff.

Tamura, S. (1989), in: S. Hunklinger, W. Ludwig, G. Weiss (Eds.), *Phonons 89,* Singapore: World Scientific, Vol. 2, pp. 703–712.

Trumpp, H. J., Eisenmenger, W. (1977), *Z. Phys. B.* **28**, 159–171.

von Gutfeld, R. J. (1968), in: W. P. Mason (Ed.), *Physical Acoustics,* New York: Academic.

Wolfe, J. P. (1989), *Festkörperprobleme* **29**, 75–104.

Wybourne, N. M., Wigmore, J. (1988), *Rep. Prog. Phys.* **51**, 923–987.

Further Reading

Eisenmenger, W. (1976), in: W. P. Mason, R. N. Thurston (Eds.), *Physical Acoustics,* New York: Academic.

Narayanamurti, V. (1981), *Science* **213**, 717–723.

Pohl, R. O., Meissner, W. (Eds.) (1993), *Proceedings of the International Conference on Phonon Scattering in Condensed Matter VI,* Singapore: World Scientific.

Wybourne, N. M., Wigmore, J. (1988), *Rep. Prog. Phys.* **51**, 923–987.

Wolfe, J. P. (1989), *Festkörperprobleme* **29**, 75–104.

HEAVY-FERMION PHENOMENA

D. W. HESS, *Naval Research Laboratory, Washington, D.C., U.S.A.*
P. S. RISEBOROUGH, *Polytechnic University, Brooklyn, New York, U.S.A.*
J. L. SMITH, *Los Alamos National Laboratory, Los Alamos, New Mexico, U.S.A.*

Introduction 435
1. **Experiments** 437
1.1 Background 437
1.2 Normal-State Properties 438
1.2.1 Specific Heat 438
1.2.2 Magnetic Susceptibility 440
1.2.3 Thermal Expansion 441
1.2.4 de Haas–van Alphen Effect 441
1.2.5 Transport Properties 442
1.2.6 Ultrasonic Measurements 444
1.2.7 Nuclear Magnetic Resonance 444
1.2.8 Neutron Scattering 444
1.2.9 Photoemission 445
1.3 Properties of the Superconducting State 446
2. **Theory** 449
2.1 Phenomenology 449
2.1.1 Landau Fermi-Liquid Theory 449
2.1.2 Single-Impurity Model 451
2.1.3 Model Hamiltonians 453
2.2 Electronic Structure Calculations 456
2.3 Superconductivity 453
3. **Summary** 460
Acknowledgments 460
Glossary 460
Works Cited 461
Further Reading 462

INTRODUCTION

Solid-state compounds containing rare-earth or actinide elements can display a wide variety of unusual behavior, including magnetic order, mixed valence, and heavy-fermion phenomena. These various behaviors reflect the quantum state of the electrons derived from atomic *f* orbitals. The heavy-fermion systems have almost localized electronic states at the Fermi surface that are primarily *f* in character. The electrons in these states interact strongly and give rise to an unusual transformation evident at low temperature. Near room temperature (~300 K), the sublattice of rare-earth or actinide atoms, usually cerium or uranium, has properties resembling those of weakly interacting magnetic moments immersed in a sea of conduction electrons. The electronic transport properties are dominated by incoherent scattering of the conduction electrons by the local moments. As the temperature is lowered, local-moment behavior gives way to electronic properties that are consistent with those of a narrow band of conduction electrons. A crossover temperature, often referred to as the "coherence temperature" T^*, can be inferred from experiments. At sufficiently low temperature, the thermodynamic properties of a system of itinerant electrons may be described in terms of thermal effective masses for the electrons. The electronic effective masses in most metals do not differ much from the mass of a free electron. In contrast, the effective mass deduced from experiments on the heavy-fermion systems is up to 1000 times greater than that of a free electron. Hence these materials were christened with the somewhat whimsical name "heavy-fermion systems."

Heavy-fermion systems display a variety of ground states including antiferromagnetic, superconducting, paramagnetic (at least to the lowest temperatures at which experiments have been performed, ~20 mK), or, as recently observed, semiconducting. Table 1 lists some of the heavy-fermion systems with their ground states and ordering temperatures where appropriate. Note that superconductivity and antiferromagnetism often coexist, with the Néel temperature T_N typically being about an order of magnitude larger than the critical temperature T_c of the superconducting state.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Table 1. Some heavy-electron systems.[a]

Principal order	Compound	Ordering Temperature (K) T_c	T_N
Superconductors	$CeCu_2Si_2$	0.65	
	UBe_{13}	0.90	
	UPt_3	0.5	5.0
	URu_2Si_2	1.5	17.0
	UPd_2Al_3	2.0	14.0
Antiferromagnets	U_2Zn_{17}	—	9.7
	UCd_{11}	—	5.0
	UCu_5	—	15.2
	$CePt_2Sn_2$	—	0.88
No long-range ordering to 0.02 K	$CeAl_3$	—	—
	$CeCu_6$	—	—
	$YbAgCu_4$	—	—
	UAl_2	—	—
		Gap (K)	
Semiconductors	CeNiSn	6.0	
	$Ce_3Bi_4Pt_3$	35.0	

---

[a] Data from Fisk *et al.* (1988); $CePt_2Sn_2$, Beyermann *et al.* (1991); $YbAgCu_4$, Rossel *et al.* (1987); CeNiSn, Takabatake *et al.* (1990); $Ce_3Bi_4Pt_3$, Hundley *et al.* (1990); and UPd_2Al_3, Amato *et al.* (1992).

Heavy-fermion systems generally appear to be close to an instability, as suggested by the proclivity of many of these systems to display an ordered ground state on the addition of impurities, on the application of a magnetic field, and possibly on the application of pressure.

The low-temperature properties of some heavy-fermion materials display similarities to those of the archetypal Fermi liquid, liquid ^{3}He. In particular, the specific heat of liquid ^{3}He varies linearly with temperature, and the magnetic susceptibility is temperature independent. Both of these quantities are considerably larger than expected for a weakly interacting collection of ^{3}He atoms because of large many-body effects. The low-temperature properties of ^{3}He are well described by the phenomenological Landau Fermi-liquid theory. In the heavy-fermion systems, the specific heat and magnetic susceptibility are also greatly enhanced by many-body effects, and it is generally accepted that the low-temperature properties of these materials can also be described by Landau Fermi-liquid theory. It was hoped that analogies with liquid ^{3}He would help clarify the role of the many-body problem in understanding the physics of the "coherent" heavy-fermion state even though they suffer greatly from oversimplification. Liquid ^{3}He is after all isotropic and translationally invariant, and has only two spin states in contradistinction to the potentially manyfold-degenerate *f* electrons constrained to the symmetry of a crystalline lattice.

In seeking to describe not only the Fermi-liquid state, but also the process whereby a lattice of *f* moments immersed in a sea of conduction electrons becomes a Fermi liquid, microscopic approaches exploit the apparent similarity to the Kondo model. The Kondo model consists of a single impurity atom with a magnetic moment embedded in a metallic host. The heavy-fermion materials may be viewed as a lattice of *f* atoms, each with a magnetic moment, embedded in a metallic matrix; a compelling example is UBe_{13}. At high temperatures the lattice reflects the properties of independent magnetic impurities, and as the temperature is reduced one expects to observe the many-body effects characteristic of the single impurity model. At even lower temperatures, the interactions between impurities become important and the magnetic scattering from the lattice becomes coherent, leading to the formation of Bloch states with narrow bandwidths. A generalization of a system of dense, periodically arranged, single impurities, such as the Anderson lattice model described below, is believed to contain the essential physics of the heavy-fermion phenomenon.

Quite apart from the issue of the formation of the heavy-fermion state itself is the nature of the ordered states. In every case, the ground state listed in Table 1 is unusual as compared to the corresponding ground state encountered in most other materials. The magnetic state is characterized by anomalously small moments associated with the *f*-atom sites. The arrangement of these moments is simple in contrast with the spiral phases often observed in rare-earth magnets. The most studied ordered state is the unusual superconducting state, which is not described by the BCS theory for a conventional superconductor. The analogies between heavy-fermion systems and ^{3}He extend beyond the normal state. Below ~1 mK, the normal state of ^{3}He becomes unstable to the formation of a superfluid state in which the pairing of atoms gives rise to a condensate. A comparison of the properties of the heavy-fermion superconducting state to those of superfluid ^{3}He suggests the possibility of unconventional superconductivity in heavy-fermion systems. The conventional superconducting state is described by an order parameter that is a complex scalar function having the symmetry of the point group of the underlying crystalline lattice. In an unconventional superconductor, as in the unconventional superfluid ^{3}He, one or more of these symmetries may be broken (Rainer, 1988). The order parameter may have more degrees of freedom than its conventional counterpart, and different superconducting phases may be degenerate. The possible consequences of such an order parameter include the existence of domain walls separating different superconducting phases and exotic vortices with complicated core structures; neither of these occurs in a conventional superconductor. The superfluid state of ^{3}He is also unusual because spin fluctuations are believed to provide the attractive effective interaction that results in the pairing of ^{3}He atoms. In an ordinary superconductor where the electron-phonon interaction provides the effective attractive interaction between electrons, spin-fluctuation–mediated interactions are unfavorable to the pairing. The analogy with liquid ^{3}He further suggests that the attractive effective interaction between electrons in the heavy-fermion systems is not provided by the electron-phonon interaction as in ordinary superconductors, but rather is electronic in origin and mediated by spin fluctuations.

While much progress has been made over the past decade, a comprehensive understanding of the heavy-fermion phenomenon is still lacking. It should not be surprising, then, that some of the material presented below, particularly the simple interpretations of various experiments, appears in some cases to be contradictory; this is the nature of the field. Below, we attempt to provide a necessarily incomplete overview of the properties of these materials and of the physical picture that has evolved. One of the most exciting aspects of the heavy-fermion phenomenon is the possibility that they may be unconventional superconductors, which serves as the primary focus of our discussion of the superconducting state. Because of the nature of this article, it is not possible to discuss the many key investigations that have led to the current picture of this phenomenon. It is also not always possible to refer to the original scientific literature, and we direct the reader to the more extensive reviews listed in the Works Cited and under Further Reading.

1. EXPERIMENTS

1.1 Background

Heavy-fermion materials occur in a variety of crystal structures as shown by the examples in Table 2. Many of the heavy-fermion systems have crystal structures of uniaxial symmetry and, as expected, exhibit anisotropic properties. The degree of the anisotropy may be unusually large, as we will see below. A common feature of the crystal structures of heavy-fermion systems is that they all have a center of inversion symmetry, a fact that has important consequences for the classification of possible superconducting states. Ott and Fisk (1987) observe another common feature of the crystal structures of heavy-fermion systems: the nearest neighbors of an *f* atom do not include another *f* atom. The large separation between *f* atoms suggests that the transfer matrix elements directly connecting *f* orbitals on different sites are small.

In spite of the large separation between *f* atoms, these materials are not typical local-moment magnets. Specific-heat experiments played an important role in the discovery of the heavy-fermion phenomenon. While it was known in the 1970s that the properties of $CeAl_3$ below 0.1 K were consistent with those expected for a Fermi liquid characterized by quasiparticles with an anomalously large effective mass (Andres *et al.*, 1975), it was unclear whether this was the correct interpretation of the data. Steglich *et al.* (1979), showed that $CeCu_2Si_2$, a compound with low-temperature properties similar to those of $CeAl_3$, is a superconductor with a specific-heat jump at the transition temperature consistent with superconducting electrons with large effective masses. This was not conclusive at the time because $CeCu_2Si_2$ is difficult to prepare and may contain impurity phases. The large specific-heat jump could be the result of antiferromagnetic ordering (such as in $NpSn_3$), which has a temperature-dependent specific heat similar to that of a BCS superconductor. The "apparent" superconductivity could then be explained by the existence of some impurity phase. The discovery of superconductivity with a similarly large specific-heat jump in UBe_{13} by Ott *et al.* (1983) confirmed the interpretation based on large effective masses, because UBe_{13} lacks the materials problems associated with $CeCu_2Si_2$ but shows similar low-temperature behavior. Other heavy-fermion materials were soon discovered.

Table 2. Crystal structures of some heavy-fermion systems.

Compound	Symmetry	Full-symmetry group
$CeCu_2Si_2$	Tetragonal	$I4/mmm$
UBe_{13}	Cubic	$Fm\bar{3}c$
UPt_3	Hexagonal	$P6_3/mmc$
URu_2Si_2	Tetragonal	$I4/mmm$
UPd_2Al_3	Hexagonal	$P6/mmm$
U_2Zn_{17}	Rhombohedral	$P\bar{6}m2$
UCd_{11}	Cubic	$Pm\bar{3}m$
UCu_5	Cubic	$F\bar{4}3m$
$CeAl_3$	Hexagonal	$P6_3/mmc$
$CeCu_6$	Orthorhombic	$Pnma$
$YbAgCu_4$	Cubic	$F\bar{4}3m$
UAl_2	Cubic	$Fd3m$
CeNiSn	Tetragonal	$I4_1md$
$Ce_3Bi_4Pt_3$	Cubic	$I\bar{4}3d$
$CePt_2Sn_2$	Tetragonal	$P4/nmm$

1.2 Normal-State Properties

1.2.1 Specific Heat. The low-temperature specific heat of the metallic heavy-fermion systems differs dramatically from that of an ordinary metal. For temperatures much smaller than the Fermi energy and the Debye temperature, the specific heat of an ordinary metal consists of an electronic contribution linear in temperature and a contribution from phonons,

$$C(T)=\gamma T+\beta T^3. \qquad (1)$$

The specific heat of an ordinary metal may be understood by comparing γ with that expected for a gas of electrons neglecting interactions, γ_0. At low temperatures, only excitations in the vicinity of the Fermi surface are important. As a result of the interaction of an electron with the periodic potential of the lattice and with other electrons, these excitations are not those of single electrons but rather quasiparticles, which have an effective mass different from the mass of a free electron. Quasiparticle excitations are probed in a low-temperature specific-heat measurement, and the ratio γ/γ_0 provides a rough measure of the effect of interactions. In a metal such as sodium, the energy bands are broad and free-electron-like; the interaction of an electron with lattice excitations is relatively weak; the effective mass of the quasiparticle does not differ much from the free-electron mass; and γ/γ_0 is ~ 1.3. In contrast, enormous γ/γ_0 ratios up to 1000 are inferred for the metallic heavy-fermion systems.

The specific heat of copper is described well by Eq. (1) for temperatures below 10 K with $\gamma \sim 1$ mJ/mol K^2. In many of the heavy-fermion systems, notably UBe_{13}, $CeCu_2Si_2$, $CeAl_3$, and $CeCu_6$, $\gamma(T)\equiv C(T)/T$ shows a strong upturn with decreasing temperature below ~ 10 K, as shown in Fig. 1 for UBe_{13}. With further reduction in temperature $\gamma(T)$ may show a weak maximum, but eventually it tends to a value ~ 1 J/(mol *f*-atom K^2). Table 3 shows extrapolated values of $\gamma(0) = \lim_{T\to 0} C_p(T)/T$ for some heavy-fermion systems. The specific heat of UPt_3 and UAl_2 has been fitted by the expression

$$C(T)=\gamma T+\delta T^3 \ln T+\beta T^3 \qquad (2)$$

for temperatures up to 20 K. This form is suggestive of a strongly interacting Fermi liquid and will be discussed further below.

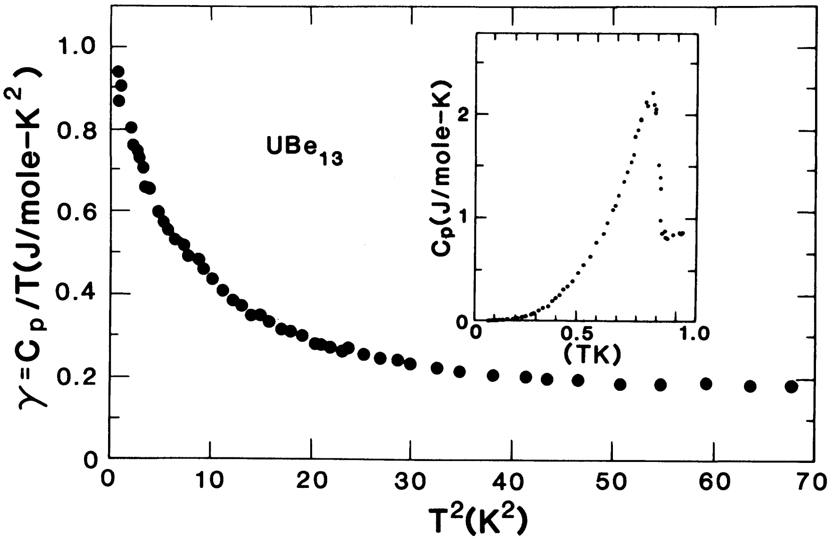

FIG. 1. The specific heat of UBe$_{13}$ divided by temperature, $\gamma(T)$ (Ott *et al.*, 1983), in the normal state. Note the upturn with decreasing temperature and the large extrapolated value of $\gamma(0)$ as compared with 0.07 mJ/mole-K^2 for copper. The inset shows the specific-heat jump at the superconducting transition, indicating that the enormous $\gamma(0)$ should be associated with itinerant electrons.

The antiferromagnets U_2Zn_{17} and UCd_{11} are unusual in that $\gamma(T)$ has already reached a saturation value by 10 K. As discussed in detail by Ott and Fisk (1987), UCu_5 is also unusual in that the characteristic heavy-fermion upturn in $\gamma(T)$ appears below the antiferromagnetic phase transition at 15 K. Thus although a gap associated with magnetic order appears over a major fraction of the Fermi surface, the mechanism that gives rise to the formation of the heavy-fermion state is still operative.

Specific-heat experiments at low temperature also show that the heavy-fermion state is sensitive to impurities and to the application of pressure. For example, substitution of thorium for 3.3% of the uranium atoms in UBe_{13} causes the upturn of $\gamma(T)$ to appear at

Table 3. Low-temperature properties of some heavy-electron systems. Properties for anisotropic systems are given along the principal directions in the form "*a/b/c*" when data on single crystals are available.[a]

Compound	$\gamma(0)$ (mJ/mol *f*-atom K^2)	$\chi(0)$ (10^{-3} emu cm^{-3})	ρ_0 ($\mu\Omega$ cm)	ρ_{ee} ($\mu\Omega$ cm K^{-2})
$CeCu_2Si_2$	1000	0.13	4.8	10.7
UBe_{13}	1100	0.18	18	—
UPt_3	450	0.19/0.19/0.10	0.5	0.5
URu_2Si_2	180	0.03/0.03/0.10	33	0.17/0.17/0.10
UPd_2Al_3	150	0.14	3.5	—
U_2Zn_{17}	500	0.19/0.19/0.24	—	—
UCd_{11}	840	0.24	—	—
UCu_5	>250	0.18	—	—
$CeAl_3$	1620	0.41	0.77	35
$CeCu_6$	1300	0.28	18.1/5.71/10.6	—
UAl_2	142	0.13	4.8	10.7
$CePt_2Sn_2$	10 000	0.30	57	—
$YbAgCu_4$	245	0.47	14	0.06

---

[a] Data from Fisk *et al.* (1988); $CePt_2Sn_2$, Beyermann *et al.* (1991); $YbAgCu_4$, Rossel *et al.* (1987); CeNiSn, Takabatake *et al.* (1990); $Ce_3Bi_4Pt_3$, Hundley *et al.* (1990): UPd_2Al_3, Amato *et al.* (1992).

lower temperatures and leads to a substantial increase in $\gamma(0)$. A similar replacement of uranium by lutetium causes the opposite effect, resulting in a ~25% decrease in $\gamma(0)$. In $CeAl_3$, UBe_{13}, $CeCu_6$, and $CeCu_2Si_2$ the application of 1 GPa pressure leads to a 30% to 50% decrease in $\gamma(0)$. Evidently some impurity substitutions lead to a contraction of the lattice and have essentially the same effect as the application of pressure.

1.2.2 Magnetic Susceptibility. An interesting crossover between local-moment and narrow-band regimes can be seen in the magnetic susceptibility as a function of temperature for the heavy-fermion systems as shown in Fig. 2 for UBe_{13} and $CeAl_3$. Near room temperature, the susceptibility $\chi(T)$ displays a Curie-Weiss temperature dependence,

$$\chi_{\mathrm{CW}}(T)=\frac{n}{3k_{\mathrm{B}}}\frac{\mu_{\mathrm{eff}}^2}{(T-\Theta_{\mathrm{CW}})}, \tag{3}$$

where μ_{eff} is the size of the local moment, Θ_{CW} is the Curie-Weiss temperature, n is the density of local moments, and k_{B} is Boltzmann's constant. The values of μ_{eff} and Θ_{CW} obtained from fitting $\chi(T)$ by Eq. (3) appear in Table 4 for some heavy-fermion systems. The effective moments deduced from the data are roughly the same as those expected from Hund's rules for free f atoms with integral f occupancy, which are also shown in Table 4 for comparison. The values of Θ_{CW} are always negative and lie between 8 K and room temperature. A detailed discussion of the sensitivity of these fits to the temperature range over which they are performed appears in Ott and Fisk (1987).

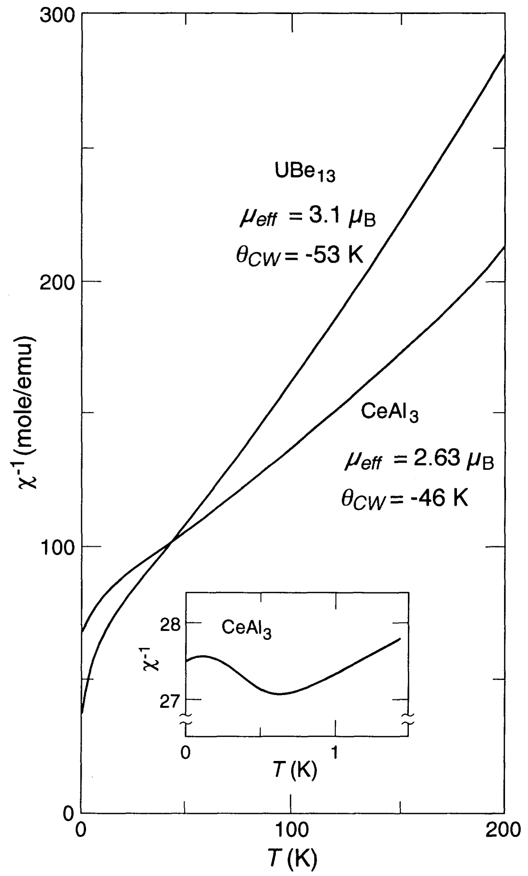

FIG. 2. The magnetic susceptibility $\chi(T)$ of UBe_{13} and $CeAl_3$ plotted as $\chi^{-1}(T)$. The effective moments and Curie-Weiss temperature from a fit by a Curie-Weiss form at high temperatures are also shown. The temperature dependence of $\chi(T)$ weakens with decreasing temperature and becomes Pauli type as shown for $CeAl_3$ in the inset. This is not apparent in UBe_{13}, as a transition to the superconducting state occurs before any behavior consistent with Landau Fermi-liquid theory is observed.

At low temperature, $\chi(T)$ is weakly temperature dependent and resembles a Pauli susceptibility. For a gas of noninteracting electrons, the Pauli susceptibility and linear specific-heat coefficient are related to the density of states at the Fermi energy $N(e_{\mathrm{F}})$ through

$$\chi_{\mathrm{P}}=\mu_{\mathrm{B}}^2N(e_{\mathrm{F}}), \tag{4}$$

$$\gamma=\frac{\pi^2}{3}k_{\mathrm{B}}^2N(e_{\mathrm{F}}), \tag{5}$$

and so a plot of γ vs χ_{P} for various free-electron-like systems lies along the line $\gamma=(\pi^2/3)(k_{\mathrm{B}}/\mu_{\mathrm{B}})^2\chi_{\mathrm{P}}$. The heavy-fermion systems, especially the superconductors, lie quite near this line (Fisk *et al.*, 1985), indicating that the low-temperature magnetic susceptibility and specific heat are enhanced by roughly the same amount. The crossover between these distinct temperature dependences occurs abruptly relative to the upturn in the specific heat, suggesting the existence of a characteristic energy scale.

The magnetic susceptibilities of many of the heavy-fermion compounds are anisotropic. A particularly striking example is that of URu_2Si_2, for which $\chi(T)$ for fields in the basal plane is nearly temperature independent up to room temperature whereas $\chi(T)$ for fields applied along the c axis displays the classic heavy-fermion behavior outlined above. In

Table 4. High-temperature properties of some heavy-electron systems. Properties for anisotropic systems are given along the principal directions in the form "*a/b/c*" when data on single crystals are available.[a]

Compound	$-\Theta_{CW}$ (K)	μ_{eff} (μ_B)	$\rho(T=300K)$ ($\mu\Omega$ cm)
$CeCu_2Si_2$	140	2.68	90
UBe_{13}	53	3.1	107
UPt_3	50/170	2.6/2.6	130
URu_2Si_2	—/—/65	—/—/3.51	324/324/169
UPd_2Al_3	47	3.2	140
U_2Zn_{17}	105	3.3	110
UCd_{11}	23	3.45	80
UCu_5	284	3.52	286
$CeAl_3$	46	2.63	65
$CeCu_6$	59/59/18	2.6/2.67/2.46	70/70/70
UAl_2	245	3.1	190
CeNiSn	—	—	206/187/125
$Ce_3Bi_4Pt_3$	125	~2.45	250
$CePt_2Sn_2$	25	2.59	~140
$YbAgCu_4$	31	4.28	~100

f configuration	Hund's-rule moment
f^1	2.54
f^2	3.58
f^3	3.62

[a] Data from Fisk *et al.* (1988); $CePt_2Sn_2$, Beyermann *et al.* (1991); $YbAgCu_4$, Rossel *et al.* (1987); CeNiSn, Takabatake *et al.* (1990); $Ce_3Bi_4Pt_3$, Hundley *et al.* (1990); UPd_2Al_3, Amato *et al.* (1992); UPt_3; de Visser (1986).

the limit of zero temperature, $\chi_c(0)/\chi_a(0) \sim 5$. A less dramatic but still large anisotropy exists for UPt_3. There the basal-plane susceptibility is nearly three times that along the *c* axis and shows a peak at 17 K.

1.2.3 Thermal Expansion. The thermal expansion coefficient of the heavy-fermion systems below room temperature is positive and strongly temperature dependent. A peak in the thermal expansion as a function of temperature is evident for some materials, while others show a thermal expansion that decreases monotonically with temperature. At low temperatures, the thermal expansion is anomalously large by two or more orders of magnitude.

The thermal expansion is often expressed as a Grüneisen parameter,

$$\Gamma=\frac{\alpha_v V_m}{\kappa C}, \qquad (6)$$

where $\alpha_v=(1/V)\partial V/\partial T$ is the coefficient of volume thermal expansion, κ is the (isothermal) compressibility, V_m is the molar volume, and C is the molar specific heat. In an ordinary metal, the low-temperature thermal expansion tracks the specific heat, and Γ is essentially temperature independent. The Grüneisen parameter for a heavy-fermion system increases strongly with decreasing temperature below 10 K. In analogy with the contribution to Γ from the lattice, which is often described by the volume derivative of the Debye temperature, the temperature dependence of Γ is interpreted as the result of a single electronic energy scale, T^*, and an electronic Grüneisen parameter is defined, $\Gamma_e = -\partial \ln T^*/\partial \ln V$. The elastic constants, magnetostriction, and thermal expansion are interrelated by this model as discussed in some detail by Fulde *et al.* (1988).

1.2.4 de Haas–van Alphen Effect. Detailed information on the geometry of the Fermi surface and the cyclotron effective masses can be obtained from the observation of oscillations in the magnetization as a function of the reciprocal magnetic induction. In ascertaining the shape of the Fermi surface, electronic band-structure calculations have traditionally been used as a guide, and for most ordinary metals, calculations and experiment are in good agreement. Generally speaking, the Fermi surface can have many sheets resulting from different bands crossing the Fermi energy. Consider a single sheet. In Fermi-liquid theories, the fundamental frequency of the oscillations gives the area of the extremal cross section of the Fermi surface perpendicular to the applied field. The temperature dependence of the amplitude is related to the cyclotron effective mass, which in turn is related to an average of the Fermi velocity over the perimeter of the extremal cross section. Although the cyclotron effective mass is not rigorously equal to the thermal effective mass, using the Fermi velocities deduced from a de Haas–van Alphen experiment to estimate the specific heat usually gives a result near the measured value.

To observe these oscillations, the mean free path l of a quasiparticle must be larger than a quasiparticle cyclotron radius,

$$l \gg v_F/\omega_H , \tag{7}$$

and the thermal smearing at the Fermi surface must be sufficiently small,

$$T \ll \omega_H . \tag{8}$$

Here v_F is the Fermi velocity, and $\omega_H = eH/m^*c$ is the effective cyclotron frequency of quasiparticles of charge e and effective mass m^* in a field H. For large effective masses and short mean free paths, these constraints suggest that the experiment should be performed at low temperature and high field. It is important to note that the amplitude of oscillation depends exponentially on the effective mass of the quasiparticles. Lighter quasiparticles produce a larger signal and are thus more easily observed.

Experiments have been performed on UPt_3, $CeRu_2Si_2$, and $CeCu_6$ for temperatures ~20–100 mK and fields up to ~14 T. In all cases unusually large effective masses have been observed. For UPt_3, the Fermi surface geometry is in good agreement with that obtained from electronic structure calculations (Taillefer *et al.*, 1987). The effective mass on any given sheet of the Fermi surface, however, is some 10 to 30 times larger than that obtained from band-structure calculations (Wang *et al.*, 1987) suggesting significant many-body effects. The effective masses range from 25 to 200 times the mass of a free electron. The ratio of the effective masses to those predicted by band theory is roughly the same as the ratio of the measured specific heat to that obtained using electronic structure calculations. UPt_3 shows only heavy masses at the Fermi surface, including the largest effective mass ever measured by the de Haas–van Alphen technique. In contrast, lighter electrons are evident in $CeCu_6$ and $CeRu_2Si_2$. Experiments on $CeCu_6$ (Chapman *et al.*, 1990) find effective masses in the range $(6\text{–}80)m_e$ and a Fermi-surface geometry that is neither similar to $LaCu_6$, which has no f electrons, nor to $PrCu_6$, which has f electrons localized in nearly atomic orbitals. Comparisons with specific-heat measurements suggest that unobserved heavier bands exist. Observed electron effective masses for $CeRu_2Si_2$ lie roughly in the range (1–20) m_e.

1.2.5 Transport Properties. At room temperature, the resistivity of a heavy-fermion compound is large, typically of order 100 μΩ/cm as shown in Table 4. In contrast to ordinary metals, where $\partial\rho/\partial T > 0$, $d\rho/dT$ can be of either sign. The qualitative behaviors observed for the resistivity as a function of

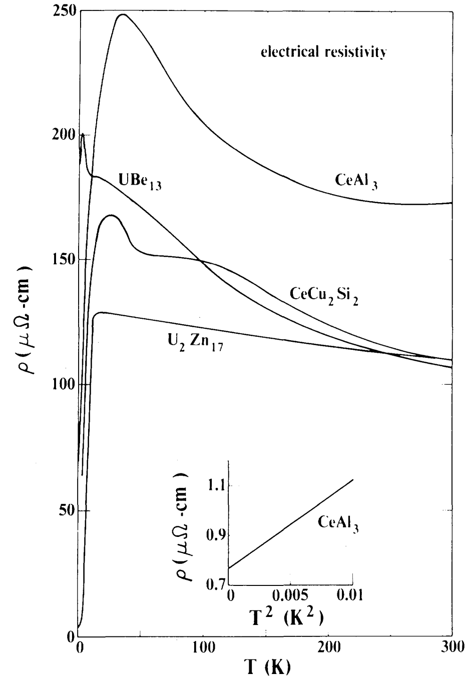

FIG. 3. The electrical resistivity as a function of temperature for the heavy-fermion systems U_2Zn_{17}, $CeAl_3$, $CeCu_2Si_2$, and UBe_{13}.

temperature in the heavy-fermion systems are summarized in Fig. 3. The large value of the resistivity and the often anomalous sign of its derivative are often attributed to the Kondo effect. Below room temperature, those heavy-fermion materials exhibiting the largest effective masses show an anomalous increase in the resistivity with decreasing temperature followed by a rapid decrease, which has been interpreted as the onset of "coherence" leading to the formation of the Fermi-liquid state. The temperature at which this occurs, the "coherence temperature," implies the existence of a characteristic energy scale, presumably the same energy scale inferred from the thermodynamic data. At low temperature, the resistivity of many heavy-fermion materials can be fitted by a form suggestive of strong quasiparticle-quasiparticle scattering,

$$\rho(T)=\rho_0+\rho_{ee}T^2, \qquad (9)$$

where ρ_{ee} is some six to seven orders of magnitude larger than that observed in ordinary metals. The low-temperature resistivity of $CeAl_3$ displays a clear T^2 contribution as shown in the inset to Fig. 3. The resistivities of UPt_3, UAl_2, and YbBiPt do not display a maximum below room temperature but are still described by the Fermi-liquid form of Eq. (9) below a material-dependent temperature less than 10 K.

The resistivity is sensitive to impurities, pressure, and magnetic fields at low temperature, as discussed in detail by Ott and Fisk (1987) and Stewart (1984).

The frequency-dependent resistivity of UPt_3, $CeAl_3$, and $CePd_3$ has been measured in the microwave, infrared, and optical regions of the spectrum. At low frequency, the conductivity can be fitted by a Drude form,

$$\sigma(\omega)=\sigma(0)/(1+\omega^2\tau^2), \qquad (10)$$

where $\sigma(0)$ is the dc conductivity, which is related to a relaxation time τ, a carrier density n, and an effective mass m^* by $\sigma(0)=ne^2\tau/m^*$. A narrow zero-frequency peak is observed at low temperatures; the relaxation time inferred from its width suggests that τ is enhanced compared to that for an ordinary metal by roughly the same factor as the effective mass that is obtained from thermodynamic properties. The narrow peak is not observed at room temperature, indicating that it is a signature of the heavy-fermion state. Complicated structure is observed at higher frequency, possibly due to interband transitions or plasmons.

In an ordinary metal, the Hall coefficient is weakly temperature dependent and provides information on the carrier charge and density. Experiments on the heavy-fermion systems find a strongly temperature-dependent Hall coefficient, as shown in Fig. 4. The Hall coefficient is large and positive at room temperature and increases with decreasing temperature. Usually the Hall coefficient displays a peak at nearly the same temperature as a similar feature in the resistivity. A peak is also observed in UPt_3 and UAl_2, which do not show an associated peak in the resistivity in the

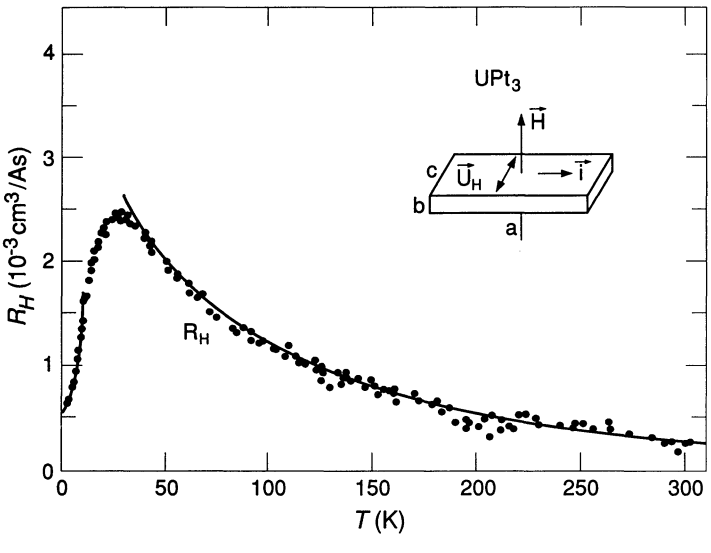

FIG. 4. The Hall coefficient as a function of temperature for UPt_3 measured in a 4-T field (from Schoenes and Franse, 1986). The solid line is a fit by the expression for incoherent magnetic scattering (see text).

corresponding temperature range. For metals containing small concentrations of impurity local moments, the Hall coefficient is often interpreted as the sum of an ordinary Hall coefficient, arising from the Lorentz force, and a generally large extraordinary term representing the incoherent scattering from local moments. At high temperature, this interpretation works well for the heavy-fermion systems (see solid line in Fig. 4) and is an aspect of the single-impurity phenomenology described below. In some heavy-fermion systems, the Hall coefficient changes sign at still lower temperatures. The rapid drop in R_H with temperature just below the peak has been interpreted as a signature of the onset of coherent scattering of the conduction electrons by the local *f*-atom moments.

1.2.6 Ultrasonic Measurements. The appearance of the heavy-electron state is evident in the temperature dependence of the elastic constants. Measurements on $CeAl_3$ (Lüthi, 1985), UPt_3, UBe_{13} (Yoshizawa *et al.*, 1985), and $CeCu_6$ (Weber, 1987) all show structure at low temperature and anomalies at high temperature, which contrast with the generally smooth and featureless elastic constants observed in an ordinary metal. Although some of these features signal structural phase transitions, as in the case of $CeCu_6$ at 220 K, others are interpreted as being due to magnetoelastic coupling to the crystal-field–split *J* multiplets (see Fulde *et al.* 1988). At low temperatures, where the heavy-electron state is developing, the temperature dependence of the elastic constants shows an anomalous change of slope for $CeCu_6$, UPt_3, and UBe_{13}.

The enormous density of states inferred from thermodynamic experiments suggests that the ultrasonic attenuation might be greatly enhanced. In fact, experiment finds that the magnitude of sound attenuation is of the order of that for an ordinary metal. In contrast to the comparatively smooth and featureless ultrasonic attenuation of an ordinary metal, an anomalous peak appears in UPt_3 at 12 K and in $CeCu_6$ at 2.5 K. In the latter case, the peak shifts to higher temperatures and broadens with increasing frequency. At still lower temperature, but above the superconducting transition, UPt_3 displays an attenuation that varies as T^2.

Specific-heat and resistivity data show that the coherent heavy-electron state is well developed at temperatures below ~ 10 K. In this regime, one expects that the phonon contribution to the ultrasonic attenuation should be negligible, and the energy of the sound wave should be dissipated primarily by the heavy-electronic liquid. As noted above, the ultrasonic attenuation is similar in magnitude to that measured for copper, which suggests that the electron-phonon coupling is reduced by an amount similar to the enhancement of the quasiparticle mass. In an ordinary metal, the sound speed v_s is much smaller than the Fermi velocity v_F, the electronic degrees of freedom are able to adjust rapidly to the lattice deformation presented by the sound wave, and the resulting attenuation process is isothermal. Lüthi (1985) pointed out that this is not the case in heavy-electron systems where $v_F \gtrsim v_s$, so that sound attenuation is adiabatic. As a consequence, the leading temperature dependence of the attenuation is due not only to the viscosity of the heavy-electron liquid, but also to the conduction of heat.

1.2.7 Nuclear Magnetic Resonance. At high temperature, the heavy-fermion systems show a nearly temperature-independent spin relaxation time T_1, consistent with the observed Curie-Weiss susceptibility. As the temperature is decreased, there is a regime below $\sim$ 10 K where T_1T is essentially constant (MacLaughlin, 1985), as expected for a Fermi liquid. As shown by Korringa, in a Fermi liquid the relaxation of polarization in the nuclear spin system is due to the coupling of the nuclear spin to the magnetic moment of the quasiparticles, so that $1/T_1T$ is proportional to the density of states at the Fermi surface (see Slichter, 1978). A comparison of T_1T for ^{13}Be in UBe_{13} and in $U_{0.967}Th_{0.033}Be_{13}$ just above T_c shows that $1/T_1T$ tracks the increase in γ observed in specific- heat experiments.

In URu_2Si_2, $T_1(T)$ decreases rapidly as the temperature is decreased through the antiferromagnetic transition at 17 K. It is interesting to note that a Korringa temperature dependence is evident above T_N and below $T \sim 10$ K.

1.2.8 Neutron Scattering. The results of inelastic neutron-scattering experiments on the magnetically ordered heavy-fermion systems U_2Zn_{17} and URu_2Si_2 and on the superconductor UPt_3 (see Aeppli *et al.*, 1988; Kjems and Broholm, 1988) show great similarities to those systems such as $CeCu_6$ and $CeRu_2Si_2$ (see Rossat-Mignod *et al.*, 1988) that remain paramagnetic down to the lowest temperatures. All of these systems have a response function with a large q-independent component that can be fitted by a relaxational form,

$$\chi(q,\omega)=\chi_0\Gamma/(\Gamma-i\omega). \tag{11}$$

The systems also have a smaller q-dependent contribution that peaks at a nonzero q, indicative of antiferromagnetic correlations. As most of the integrated intensity expected from sum rules is tied up in the above response, there is little intensity left to assign to the low-ω and -q peak associated with Pauli paramagnetism. This implies strong spin-orbit coupling between the quasiparticles of a Fermi liquid, and that most of the response is of local origin.

In paramagnetic heavy-fermion systems, the antiferromagnetic correlations only become apparent at low temperatures. The widths of the peaks in $S(q,\omega)$ at fixed ω indicate correlation lengths of the order of 3–10 Å. In the ordered systems, the single-site component of the response function remains comparable to those of the paramagnetic systems, and the phase transitions are considered to be the result of a slight increase in the effective exchange interaction at low temperatures. This interpretation is consistent with the observation that many of the paramagnetic systems can be driven into the ordered phase with the introduction of impurities. It is surprising that of the ordered magnetic systems only URu_2Si_2 shows the existence of sharp spin-wave-like excitations, albeit with a small dispersion. Perhaps the most surprising result of all is found in the elastic scattering peaks: in URu_2Si_2 and UPt_3 the strength of the antiferromagnetic Bragg peak is indicative of an ordered magnetic moment with very small magnitude, $\sim 0.03\mu_B$ and $\sim 0.02\mu_B$, respectively. It should also be pointed out that the Bragg peaks are broadened, indicating finite correlation ranges in the ordered state, probably due to defects.

1.2.9 Photoemission. The photoemission spectrum provides a measure of the occupied portion of the electronic density of states, if surface effects and the interaction between the observed electron and the system can both be neglected. Likewise the inverse photoemission spectrum is related to the unoccupied electronic density of states. Using the phenomenon of resonance, it is possible to extract the f component from the spectra. The resulting spectra show marked distinctions between the normal rare-earth and actinide systems. On one hand, the cerium systems show pronounced features that may be attributable to transitions between localized ionic configurations. The energy separation between these features is taken to be a measure of U_{ff}, the Coulomb interaction energy between the electrons in the $4f$ orbitals on the cerium ions. On the other hand, normal uranium-based systems, such as UIr_3 (Arko *et al.*, 1983), agree quite well with band theory, indicating much smaller values of U_{ff} for uranium. This satisfactory state of affairs does not extend to the heavy-fermion uranium systems. When the $5f$ spectra of UPt_3 and UAl_2 were compared to the density of states obtained from band-structure calculations in the local-density approximation (Allen *et al.*, 1985), it was found that the observed spectra were considerably broader. Similar discrepancies were found for UBe_{13}, U_2Zn_{17}, and URu_2Si_2 (Allen , 1988). It has been shown that these discrepancies in width exhibit a systematic variation (Arko *et al.*, 1988) that is most profound in the heavy-fermion systems and is attributable to the development of satellites in the occupied portion of the spectra. Similar analysis of inverse photoemission experiments on UBe_{13} shows some evidence for the existence of satellites in the unoccupied $5f$ density of states. Thus, in the heavy-fermion systems it appears as though the spectrum may have slight features corresponding to localized states superimposed on top of a $5f$ band that cuts across the Fermi level.

In contrast to the uranium-based heavy-fermion systems, there have been very few high-resolution studies of cerium heavy-fermion systems. The earlier studies seem to indicate that the features seen in the f spectrum are less sharp than those found in the cerium systems with more modestly enhanced electronic masses, perhaps being more akin to the uranium heavy-fermion systems. The interpretation of the photoemission spectra of cerium-based compounds is the subject of considerable controversy (Allen *et al.*, 1986; Liu *et al.*, 1982; Freeman *et al.*, 1987). The spectra show three major peaks: one above, one below, and one cutting across the Fermi level. One school of thought attributes the peaks above and below the Fermi level to transitions from an f^1 initial state to f^2 and f^0 final-state configurations, respectively. The peak near the Fermi level is associated with the Kondo effect whereby a spin polarization is induced in the conduction band at low temperatures, in order to compensate the magnetic moment of the f^1 configuration. The intensity in this peak is expected to scale with T_K, the characteristic energy scale of the Kondo model. A second school of thought attributes the existence of the two peaks at and below the Fermi level to two different mechanisms that are available to screen the charge deficiency associated with the f^0 final state. This school relates these screening channels to the systematic variation of the two peaks found in the photoemission of the praseodymium and neodymium rare-earth compounds, which have electronic mass enhancements of the order of those in ordinary metals.

1.3 Properties of the Superconducting State

Superconductivity in the heavy-fermion systems differs markedly from that in ordinary metals. Here, we briefly summarize some of the unusual properties of the heavy-fermion superconducting state, pointing out the differences between the heavy-fermion and conventional BCS superconducting states that lead to the conjecture that these are unconventional superconductors as defined in the Introduction.

At the transition temperature to the superconducting state T_c, the quasiparticles at the Fermi surface of a conventional superconductor are unstable to the formation of Cooper pairs. In the absence of a current, the Cooper pair consists of two electrons in states of momenta and spin $\mathbf{k},\sigma$ and $-\mathbf{k},-\sigma$ bound in a relative *s*-wave state by an attractive interaction mediated by phonons. Correlations between Cooper pairs generally lead to the formation of a gap in the quasiparticle dispersion relation,* which at sufficiently low temperature results in a thermally activated temperature dependence for specific heat, ultrasonic attenuation, nuclear spin relaxation, and thermal conductivity (see Schrieffer, 1964).

In contrast, the specific heat well below T_c for the heavy-fermion uranium compounds is described by

$$C(T)=\gamma_s T+\delta T^n \tag{12}$$

where γ_s is sample dependent and has often been attributed to conduction electrons that remain in the normal state, and n is 3 and 2 for UBe_{13} and UPt_3, respectively. Power-law temperature dependences have also been observed for ultrasonic attenuation, nuclear spin relaxation rates, thermal conductivity, and penetration depth. Sound attenuation in UPt_3 has been studied particularly thoroughly, and different exponents have been found, depending on the polarization and direction of propagation of the sound wave relative to the crystal axes. The exponents are also sample dependent. The power-law behavior observed in all of these properties strongly suggests that the gap function vanishes for some region on the Fermi surface and thus that, at the very least, the gap function depends strongly on the direction of **k**. The more exciting possibility is that the gap function has a lower symmetry than does the crystal lattice as a result of quasiparticles that pair in a nonzero angular momentum state, possibly in a triplet state as in liquid ^3He.

The many-body wave function of the superconducting state is a phase-coherent superposition of single-particle states, and interference effects can be observed in many experiments. The attenuation of sound waves and the relaxation rate of a nuclear spin show classic signatures of interference effects in a conventional BCS superconductor (Schrieffer, 1964). As the temperature is lowered through the superconducting transition, sound attenuation decreases abruptly and the relaxation rate of a nuclear spin displays a Hebel-Slichter peak as shown for ^{119}In in Fig. 5(a). In the heavy-fermion systems, the Hebel-Slichter peak is not observed, as shown for UBe_{13} in Fig. 5(b); the spin relaxation rate drops abruptly below T_c. The ultrasonic attenuation, on the other hand, does not plummet below T_c but shows a sharp peak for UBe_{13} and for the best samples of UPt_3. A broader peak is evident for URu_2Si_2.

We have already observed that the specific-heat discontinuity is a large fraction of the

---

*This does not occur in so-called gapless superconductors where the gap has been suppressed by impurities. The heavy-fermion superconductors are not intrinsically gapless.

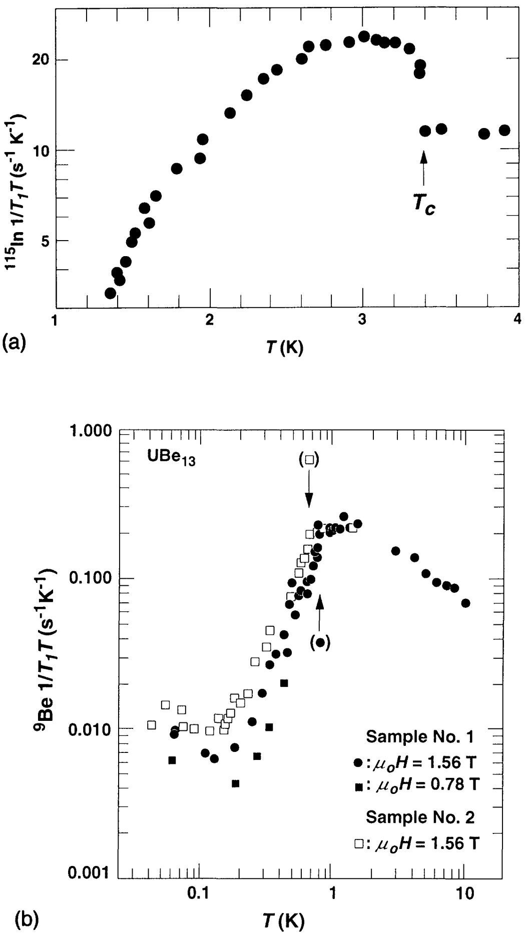

FIG. 5. Nuclear spin relaxation rate for (a) the conventional BCS superconductor ^{115}In and (b) the heavy-fermion system UBe$_{13}$. Note the absence of a Hebel-Slichter peak in UBe$_{13}$.

specific heat at T_c, consistent with the expectation for quasiparticles with a large effective mass condensing into the superconducting state. The normalized specific-heat discontinuities $\Delta C(T_c)/C_N(T_c)$ of several heavy-fermion superconductors are listed in Table 5. For most superconductors, $\Delta C(T_c)/C(T_c)$ lies within 20% of the weak-coupling BCS theory result of 1.43 (Parks, 1969). Larger values may occur as a result of strong coupling to lattice phonons. With the exception of UBe$_{13}$ and CeCu$_2$Si$_2$, the heavy-fermion superconductors show substantially smaller normalized specific-heat discontinuities. Moreover, the transition is often broad, as in UPt$_3$ and URu$_2$Si$_2$. We return to this point below.

The heavy-fermion superconducting state is very sensitive to impurities and defects. The transition temperature can be strongly depressed in UPt$_3$ by grinding the sample. The substitution of a small amount of thorium for uranium in UBe$_{13}$ leads to a rapid decrease of T_c. Further thorium substitution reveals the complicated phase diagram of U$_{1-x}$Th$_x$Be$_{13}$ in Fig. 6 that shows at least two and possibly more superconducting phases. For thorium concentrations lying roughly between 0.017 and 0.05, two phase transitions were observed in specific-heat experiments with discontinuities at the transitions of comparable size. The upper critical field of the higher-temperature phase behaves differently from that for pure UBe$_{13}$ when doped with small amounts of Gd impurities, suggesting that these are two different superconducting phases. Early sound attenuation experiments suggested that the low-temperature phase transition was to a magnetically ordered state. Subsequent neutron-scattering experiments failed to observe a magnetically ordered state. Muon-spin resonance experiments detect a small ordered moment $\leqslant 0.01\mu_B$. The small size of the moment suggests the possibility that the low-temperature transition may separate two unconventional superconducting phases. We discuss this possibility in Sec. 3 below.

The phase diagram of a conventional superconductor is determined by the ratio of the characteristic length scale over which the magnetic field varies, the penetration depth λ,

Table 5. Properties of some heavy-fermion superconductors.

Compound	T_c (K)	$\Delta C(T_c)/C(T_c)$	$-dH_{c2}/dT\|_{T=T_c}$ (T/K)	$H_{c2}(T=0)$ (T)
CeCu$_2$Si$_2$	~0.7	1.3	23	2.0
UBe$_{13}$	0.85	2.5	42	8
UPt$_3$	0.45	1.0	6.4/7.5	2.0/2.5
URu$_2$Si$_2$	1.1	0.82	4/14	8.0/2.0
UPd$_2$Al$_3$	2.0	1.2	4.3	~3.6

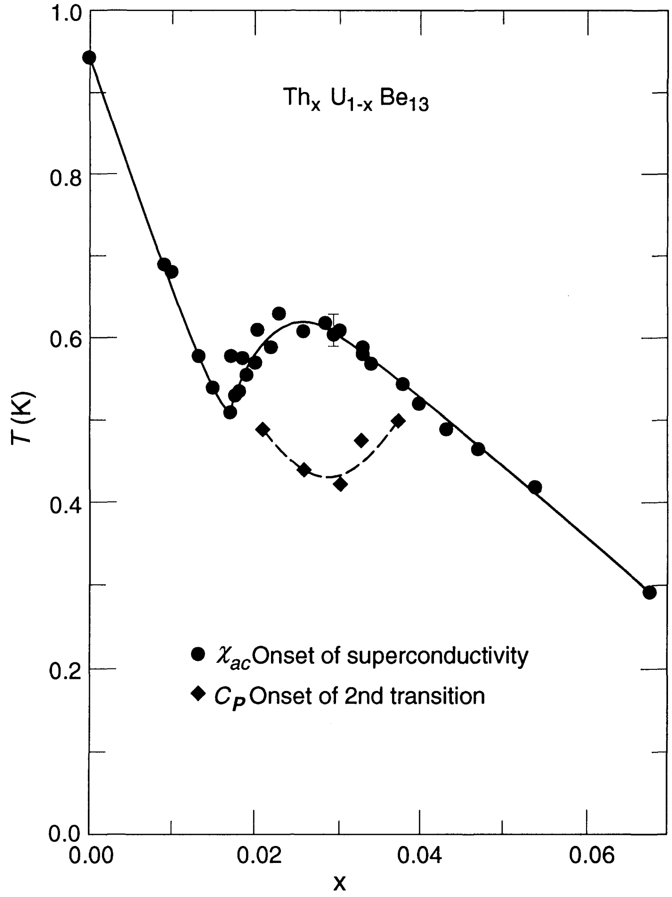

FIG. 6. The temperature–thorium concentration (x) phase diagram of $U_{1-x}Th_xBe_{13}$. Note the striking change in $T_c(x)$ (solid line) as x is increased through $x_c \simeq 0.018$ and the second phase transition at $T_{c2}(x)$ (dashed line).

to that over which the order parameter varies, the coherence length ξ. For a type-II superconductor, $\kappa = \lambda/\xi > \sqrt{2}$, and the H-T phase diagram shows two superconducting phases and one normal phase separated by the lower critical field $H_{c1}(T)$ and the upper critical field $H_{c2}(T)$. In the Meissner phase ($H < H_{c1}$), magnetic fields are excluded from the superconductor. As the field is increased above H_{c1}, magnetic flux penetrates the superconductor in a triangular lattice of flux tubes. Each flux tube is a vortex in the phase of the order parameter. The order parameter vanishes on the vortex axis and heals to its bulk equilibrium magnitude in a distance of the order of ξ. The magnetic field is a maximum at the center of the vortex and decays on the scale of λ. As the field is increased so too is the density of vortices until at H_{c2} the vortex cores begin to overlap and a transition occurs to the normal state.

The heavy-fermion superconductors are type-II superconductors with κ, as inferred from the slopes of H_{c2} and H_{c1} at T_c, as large as ~50 or more. These rank among the largest κ values found in conventional superconductors, for which the large value of κ occurs as a result of scattering from impurities. In contrast, the heavy-fermion substances are pure materials. The temperature dependence of $H_{c2}(T)$ for the heavy-fermion systems is generally unusual. As shown in Table 5, $dH_{c2}(T)/dT|_{T_c}$ for UBe_{13} is unusually large, possibly exceeded only by that for the high-T_c superconductors. In UBe_{13}, $H_{c2}(T)$ is nearly linear over a large range of temperature and shows a change in slope at about 0.5 K. In UPt_3 an anomalous crossover in the anisotropy of $H_{c2}(T)$ occurs with decreasing temperature. For temperatures near T_c, $H_{c2\parallel}/H_{c2\perp} \sim 1.3$ (where $\parallel$ and $\perp$ refer to the field direction with respect to the crystal c axis) and drops to

~0.8 as T approaches 0. The $T=0$ extrapolated value of H_{c2} exceeds simple estimates of the paramagnetic limit at which the spin-polarized normal state is favored over a conventional superconducting state. The upper and lower critical fields in the basal plane of UPt_3, and $H_{c1}(T)$ in $U_{1-x}Th_xBe_{13}$ at the low-temperature transition, show unusual sharp kinks.

The phase diagram of UPt_3 in the H-T plane shows five superconducting phases in contrast to the two observed in an ordinary superconductor. This is strong evidence, as we discuss below, that UPt_3 is an unconventional superconductor. The phase diagram as observed in the sound speed measurements of Adenwalla *et al.* (1990) is shown for field in the basal plane and along the c-axis in Fig. 7 (see also Bruls *et al.*, 1990). The figure shows three phases in the mixed state, but does not show the Meissner phases, which lie close to the temperature axis.

A connection between weak antiferromagnetic order and superconductivity is suggested in Table 1. For some of the heavy-fermion superconductors, a transition to an antiferromagnetic state with an extraordinarily small magnetic moment, of order $\sim 0.01\mu_B$, occurs at a Néel temperature $T_N \sim 10T_c$. A notable example is UPt_3, for which recent specific-heat (Fisher *et al.*, 1989) and sound-velocity experiments observe two superconducting transitions 60 mK apart, which may be a signature of the coupling of an unconventional superconducting-order parameter to a magnetic-order parameter (see Sec. 2.3 and Hayden *et al.*, 1992). It is interesting that prior experiments found only a single broad peak at the superconducting transition in UPt_3 as also observed for URu_2Si_2. This suggests that an aging effect may occur in UPt_3 and that antiferromagnetic order may appear with time (Smith, 1992). A possible mechanism for an aging effect is radiation damage due to uranium decay.

2. THEORY

2.1 Phenomenology

2.1.1 Landau Fermi-Liquid Theory. At low temperature, the linear dependence of the specific heat, the nearly temperature-independent susceptibility, and the large T^2 coefficient of the resistivity suggest that the heavy-fermion systems settle into a strongly renormalized Fermi-liquid state below a (material-dependent) characteristic temperature. Well

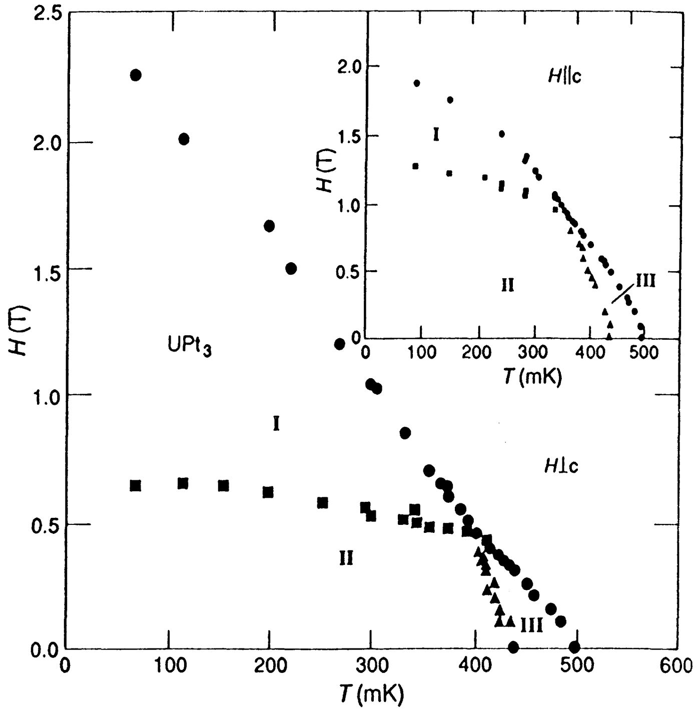

FIG. 7. The field-temperature phase diagram of the superconductor UPt_3 as deduced from sound-velocity measurements for fields in the basal plane (from Adenwalla *et al.*, 1990). The lower critical field separating the mixed phases from the Meissner phases is not shown but lies just above the temperature axis. In all there are five superconducting phases in the H-T plane. Inset: The corresponding phase diagram for fields along the c axis.

below an effective degeneracy temperature, the properties of diverse systems of strongly interacting fermions resemble those of a system of noninteracting fermions and are determined in large part by the low-energy excitations of the interacting system. The phenomenological Fermi-liquid theory of Landau may be used to describe such systems. The central concept of Landau Fermi-liquid theory is the quasiparticle, which is an elementary excitation of the interacting system that can be crudely viewed as a particle surrounded by a cloud of other particles. The lifetime of a quasiparticle near the Fermi energy [varying as $1/(e^2+\pi T^2)$, with e measured from the Fermi energy e_F] is sufficiently long to provide a useful description of the low-lying elementary excitations of the interacting system. The original system of strongly interacting "bare" particles is thought of as a system of interacting quasiparticles. The dynamical properties of these quasiparticles, such as the mass, are renormalized from those of the bare particles by the accompanying motions of the surrounding particle clouds. The Landau theory provides no prescription for calculating the properties of or interactions between quasiparticles; rather, these are to be determined from experiments or microscopic theories. The Landau theory provides relations between physical quantities in terms of the effects of residual screened interactions between the quasiparticles. A proper Landau theory for any metal should incorporate the symmetry of the underlying crystalline lattice as outlined by Fulde *et al.* (1988) for heavy-fermion systems. Such a description is complicated by the occurrence of a large number of parameters describing the quasiparticle interactions. So we describe an overly simplified theory that is similar in construction to the Fermi-liquid theory for liquid ^{3}He, for the purpose of introducing Fermi-liquid theory and also because the results of some calculations based on model Hamiltonians are expressed in terms of the Landau parameters of this model.

The free energy may be expanded about its zero-temperature limit in terms of the interactions between quasiparticles,

$$F(T)=F(0)+\sum_{\mathbf{p}} E^0_{\mathbf{p}}\delta n^\sigma_{\mathbf{p}}+\sum_{\mathbf{p},\mathbf{p}',\sigma,\sigma'} f^{\sigma,\sigma'}_{\mathbf{p},\mathbf{p}'}\delta n^\sigma_{\mathbf{p}}\delta n^{\sigma'}_{\mathbf{p}'} + \text{terms } O((\delta n)^3), \qquad (13)$$

where $\mathbf{p}$ and σ are momentum and spin variables and $\delta n^\sigma_{\mathbf{p}}$ is the deviation of the distribution function from its zero-temperature limit. We have retained only the interaction between two quasiparticles $f^{\sigma,\sigma'}_{\mathbf{p},\mathbf{p}'}$, which is dominant at sufficiently low temperature. Symmetries are used to simplify the form of f and to choose an appropriate set of basis functions; the coefficients appearing in the basis function expansion are the so-called Landau parameters. For example, rotational symmetry in spin space leads to the form $f^{\sigma,\sigma'}_{\mathbf{p},\mathbf{p}'} = \mathbf{1}f^s_{\mathbf{p},\mathbf{p}'} + \boldsymbol{\sigma}\cdot\boldsymbol{\sigma}' f^a_{\mathbf{p},\mathbf{p}'}$. For liquid ^{3}He, the appropriate basis functions are the usual spherical harmonics; $\mathbf{p}$ has the meaning of a quasiparticle momentum and σ the meaning of a quasiparticle spin. Various thermodynamic and transport properties of the system can be expressed in terms of the Landau parameters. In this case one finds for the compressibility

$$\chi_d=\frac{m^*/m}{1+F^s_0}\,\chi^0_d\,, \qquad (14)$$

and for the susceptibility

$$\chi_s=\frac{m^*/m}{1+F^a_0}\,\chi^0_s\,, \qquad (15)$$

where χ^0_d and χ^0_s are the corresponding susceptibilities in the noninteracting system, and the dimensionless Landau parameter $F^\lambda_l = N(e_F)f^\lambda_l$, where $N(e_F)$ is the density of states that appears in the linear coefficient of the specific heat,

$$\gamma=\frac{\pi^2}{3}N(e_F)k_B^2=\frac{m^*}{m}\gamma^0, \qquad (16)$$

m^* is the quasiparticle effective mass, and k_B is Boltzmann's constant.

Carneiro and Pethick (1975) showed that long-wavelength fluctuations of spin and density in a system of interacting fermions will generally contribute a term to the low-temperature specific heat that depends logarithmically on the temperature,

$$C(T)=\gamma T+\zeta T^3\ln T+\beta T^3+\ldots, \qquad (17)$$

where ζ may be expressed in terms of the Landau parameters.

The observation that UPt$_3$ has a specific heat of the form Eq. (17) led to the proposal that quasiparticle spin fluctuations play a similar role in UPt$_3$ as they do in ^{3}He (Stewart, 1984). The electron-phonon interaction can also contribute to the $T^3 \ln T$ term in the specific heat, but the sign of this contribution to ζ is opposite to that produced by the spin and density fluctuations and is also of opposite sign to that observed in UPt_3. A single-band spin-$\frac{1}{2}$ model was developed for UPt_3 and UAl_2, assuming an expression for ζ taken from the Landau theory of ^3He, estimating the effective radius of the Fermi surface, and neglecting scattering processes involving angular momentum channels of $l=2$ and higher (see Pethick and Pines, 1987, for a review). This allowed the analogy of UPt_3 and ^3He to be extended to the superconducting state. In ^3He the superfluidity is believed to occur due to an attractive interaction mediated by spin fluctuations. The analogy yields an attractive effective interaction between two electrons, which produces superconductivity below a critical temperature T_c. They calculated the pressure dependence of T_c and found it to be in agreement with experiment. The mechanism for the formation of Cooper pairs (see Sec. 3) in this model is of purely electronic origin, and the resulting superconducting state is unconventional. A further consequence of this model is that the effective moments of the quasiparticles are renormalized so that $\mu < \mu_B$.

2.1.2 Single-Impurity Model. For temperatures well above the "coherence temperature" T^*, the single-impurity Anderson model provides a useful phenomenological framework for the heavy-fermion systems. This model was originally introduced by Anderson to describe dilute magnetic impurity ions dissolved in a solid. However, the heavy-fermion systems have a relatively high concentration of magnetic (f) ions, and so the use of this single-ion phenomenology in the high-temperature regime implicitly assumes the decoupling of the f orbitals on different lattice sites.

The single-ion Anderson model describes a doubly degenerate localized level, e.g., an f level, that is bathed in a Fermi sea of conduction electrons. The Hamiltonian can be written as

$$H = H_f + H_d + H_{fd}\,, \tag{18}$$

where H_f describes the f-electron system, H_d describes the conduction-electron system, and H_{fd} describes the coupling between these systems. The Hamiltonian H_f is

$$H_f = \sum_{\alpha} E_f n_\alpha + \sum_{\alpha > \beta} U_{ff} n_\alpha n_\beta\,, \tag{19}$$

where n_α represents the number of electrons in the f orbital with spin index α. The first term, proportional to E_f, represents the binding energy of an electron of spin α in the f shell. The second term, proportional to U_{ff}, represents the Coulomb repulsion between pairs of electrons in the f shell. The Pauli exclusion principle makes this term inactive when $\alpha = \beta$. The conduction electrons described by H_d are noninteracting conduction electrons in band states:

$$H_d = \sum_{\mathbf{k}\alpha} e_{\mathbf{k}} n_{\mathbf{k}\alpha}\,, \tag{20}$$

where $n_{\mathbf{k}\alpha}$ is the number of electrons of spin α in the conduction-band state labeled by the Bloch wave vector $\mathbf{k}$ and $e_{\mathbf{k}}$ is the dispersion relation of the conduction-band states. The f- and d-electron systems are coupled by a spin-conserving hybridization interaction H_{fd}, which represents processes whereby individual electrons can tunnel out of the f level into the conduction-band states along with the reverse processes. This hybridization process has matrix elements between localized f states $|f\alpha\rangle$ and itinerant states $|d\mathbf{k},\alpha\rangle$ given by

$$V_{\mathbf{k}} = \langle f\alpha | H_{fd} | d\mathbf{k},\alpha \rangle. \tag{21}$$

Applying Fermi's "golden rule," one obtains an effective f-level width Δ, for the state in which the f electron is occupied by a single electron, given by

$$\Delta = \pi |V|^2 N(e_F), \tag{22}$$

where $N(e_F)$ is the conduction-band density of states, per spin, evaluated at the Fermi energy e_F.

One question of interest is whether the impurity ion can carry a magnetic moment in a metallic matrix. Anderson (1961), working within the Hartree-Fock approximation, showed that a magnetic moment might be formed for sufficiently large U_{ff}/Δ, whereas for smaller values of the ratio the impurity was nonmagnetic. These results can be understood by first considering two extreme limits that can be solved exactly: In the atomic limit, for which $\Delta = 0$, adding an electron of spin β to a localized level already occupied by an electron of spin α costs an energy of $E_f + U_{ff}$, whereas the energy for removing the electron of spin α is E_f. Thus, the one-electron density of states is split into two sharp levels; the β level occurs above the Fermi level, and the α sublevel occurs below the Fermi level. The other extreme limit is the noninteracting limit where $U_{ff} = 0$, in which case the f states admix with the conduction-band states through the hybridization term, giving rise to a resonant peak in the f density of states at E_f with width given by Δ. In this noninteracting case, the f states of spin α and spin β are occupied by equal numbers of electrons (possibly fractionally) and hence possess no net magnetic moment. For the same reason that the f level splits in the atomic limit, the Hartree-Fock solution also shows a splitting of the f density of states for large but finite values of the ratio U_{ff}/Δ. The two subbands are of different spin characters and are separated by an energy of the order of U_{ff}. This split density of states, shown in Fig. 8, can lead to the formation of a local f moment when occupied by electrons in accordance with the Pauli exclusion principle. For systems with a number of f electrons per f ion less than unity, $n_f < 1$, the Hartree-Fock treatment requires that $|E_f - e_F| < \Delta$, which places the lower subband close to the Fermi level and the upper subband above. For almost integer occupation of the f level the lower subband will lie at even lower energies and may have very little weight at the Fermi energy. This

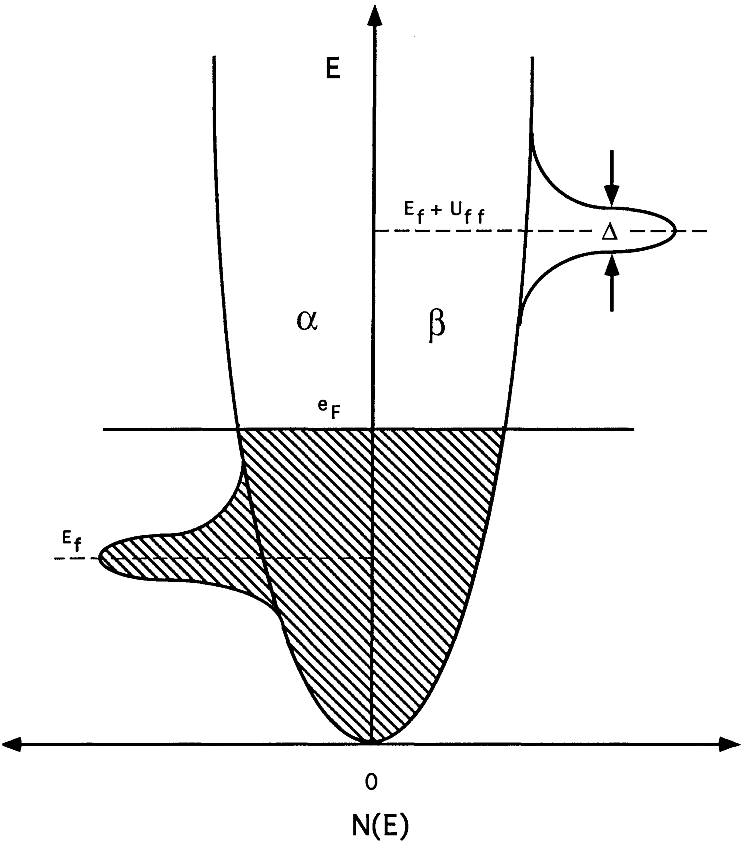

FIG. 8. The density of states from the Hartree-Fock solution to the single-impurity Anderson model showing that the effect of the intra-atomic Coulomb interaction is to split the original f level into two levels according to spin. One level occurs at the position of the original f level, the other at an energy U_{ff} higher and above the Fermi energy. Processes by which an electron can tunnel from an atomic f orbital to the conduction band and the reverse processes called hybridization are responsible for the finite level width $\sim\Delta$.

latter regime is defined as the Kondo regime, in which Hartree-Fock theory would predict well-developed local moments. Schrieffer and Wolff (1966) have shown that these local moments are coupled to the spin polarization of the conduction electrons by an antiferromagnetic exchange interaction of strength $J=V^2/E_f$. This coupling introduces quantum fluctuations in the directions of the local moments that are not included in the Hartree-Fock solution.

The most marked modifications to the Hartree-Fock results are required in the Kondo regime, as is suggested by renormalization-group scaling arguments (Haldane, 1978). What happens is that J produces a bound state that when occupied by an electron produces a compensating spin polarization in the conduction electrons. At low temperatures the resulting complex is a spin singlet, whereas at higher temperatures the bound state is ionized leading to the reappearance of the local moment. This low-energy behavior leads to the appearance of a narrow temperature-dependent resonance in the single-electron spectrum at the Fermi level known as the Kondo peak. Nozières (1974) has shown that the low-temperature properties can be described quite well by a Landau Fermi-liquid theory in which the impurity produces a phase shift for the conduction electrons that reaches $\delta_0(\mathbf{k}_\mathrm{F})=\pi/2$ at the Fermi energy. The picture of a low-energy, low-temperature Fermi liquid can also be obtained by a projection method for minimizing the Coulomb interaction energy, developed by Gutzwiller (1965), which basically renormalizes the hybridization width Δ by a factor of $1-n_f$ corresponding to the elimination of processes in which an electron may hop into an f shell that is already occupied. For cerium, with an f band less than half filled, the maximum intensity in the Kondo peak is above the Fermi energy. The resulting picture of the one-electron spectrum in the Kondo regime is that of a filled subband below the Fermi energy, separated by an energy U_{ff} from an unfilled upper subband, together with a temperature-dependent Kondo resonance located at the Fermi energy. The width of the Kondo resonance is related to the Kondo temperature T_K where $k_\mathrm{B}T_\mathrm{K}\sim\Delta(1-n_f)$ and its peak height is $\sim 1/\pi\Delta$.

Confidence in the above picture has been established by exact Bethe-*Ansatz* solutions (Schlottmann, 1989; Tsvelick and Wiegman, 1983) and by exact numerical renormalization-group studies (Krishnamurthy *et al.*, 1980). One of the most notable features of these solutions is that the thermodynamic properties only depend on a single energy scale given by the Kondo temperature.

For the heavy-electron systems, one imagines a lattice of such impurities (the cerium or uranium atoms). As the temperature is reduced it is possible that the compensating polarization clouds of each impurity lock together in a coherent heavy-fermion state. This possibility has motivated studies of the interactions between two Anderson impurities (Jones *et al.*, 1988). As a result of the occurrence of another energy scale, the exchange interaction J between the impurity moments, two possibilities arise. If $J\ll T_\mathrm{K}$, then the compensation of each local moment occurs independently, whereas if $J\gg T_\mathrm{K}$, the moments order before the screening clouds develop.

Nozières (1985) noted another complication of this picture. The formation of a compensating screening cloud only involves conduction electrons within T_K of the Fermi surface, but the total number of such electrons is much less than that required to screen every local moment in a concentrated compound.

Although much use can be made of the above phenomenologies, it remains unclear to what extent they provide a firm theoretical basis for understanding the properties of a lattice of f ions immersed in a Fermi sea of conduction electrons.

2.1.3 Model Hamiltonians. It is believed that the physics of the single Kondo impurity is related to the physics of the process by which the system of apparently weakly interacting local moments in a sea of conduction electrons is transformed to an itinerant band of heavy quasiparticles. The Anderson lattice model makes this notion more precise and quantitative, by generalizing the Hamiltonian of a single f impurity to a dense lattice of f sites.

The Anderson lattice Hamiltonian may be written as the sum of three terms as in Eq. (18), but with H_f representing a lattice of f ions:

$$H_f=E_f\sum_{i\alpha} n_{f\alpha i}+\sum_{i,\alpha<\beta} U_{ff}n_{f\alpha i}n_{f\beta i}\,, \qquad (23)$$

where i labels the f-lattice sites and α labels an appropriate basis set that reflects the full 14-fold spin- and orbital-degenerate f level. The hybridization terms are generalized by allowing the f level at each site to admix with the conduction-band states. The transformation from the Wannier representation of the f levels to a Bloch representation,

$$|f\mathbf{k},\alpha\rangle=\frac{1}{\sqrt{N_s}}\sum_i e^{i\mathbf{k}\cdot\mathbf{R}_i}|fi,\alpha\rangle, \tag{24}$$

together with conservation of crystal momentum, leads to the simple form for the hybridization matrix element

$$V_{\mathbf{k}}=\langle f\mathbf{k},\alpha|H_{fd}|d\mathbf{k},\alpha\rangle. \tag{25}$$

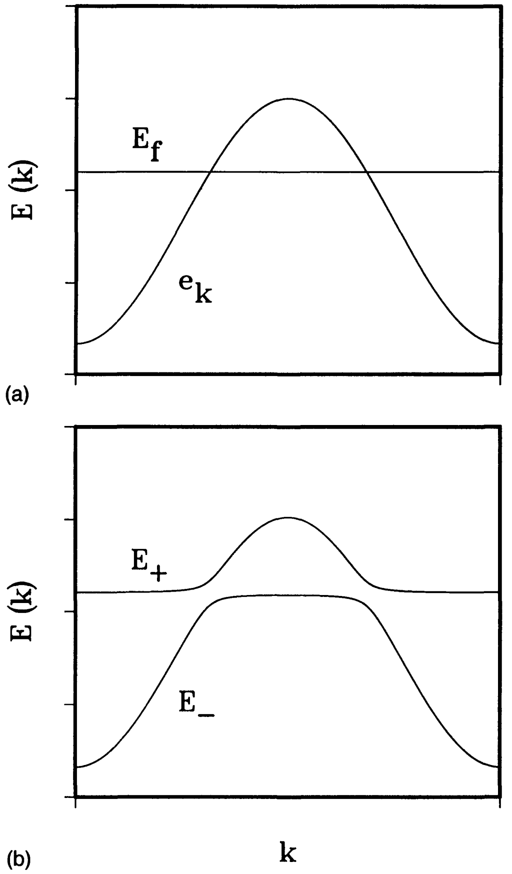

FIG. 9. Flat atomic f level, and a broad conduction band shown as a function of crystal momentum in (a), are admixed by the hybridization matrix element V, to produce the two spin-degenerate bands shown in (b).

In the absence of the intrasite Coulomb interaction, $U_{ff}=0$, the lattice Hamiltonian can be completely diagonalized. This yields two spin-degenerate bands with dispersion relations

$$E_{\pm}(\mathbf{k})=\tfrac{1}{2}\{E_f+e_{\mathbf{k}}\pm[(E_f-e_{\mathbf{k}})^2+|2V_{\mathbf{k}}|^2]^{1/2}\}, \tag{26}$$

which are sketched in Fig. 9. If the hybridization matrix element $V_{\mathbf{k}}$ is non-vanishing over the entire Brillouin zone, then the electronic-spectrum exhibits an indirect gap of magnitude V^2/W where W is the d-band width. The existence of a hybridization gap depends on the validity of many simplifying assumptions such as the neglect of direct f-f hopping, crystalline field effects, spin-orbit coupling, and so on. Nonetheless, the hybridization gap often plays an important role in many-body theories for the heavy-fermion semiconductors.

The introduction of the Coulomb interaction U_{ff} is expected to modify the above simple picture of the electronic states. Unfortunately, no exact solutions for the Anderson lattice model are known, leading researchers to rely on approximate methods.

At high temperature, the Anderson lattice Hamiltonian is expected to behave like an incoherent superposition of single Anderson impurities. As observed above, the properties of the heavy-electron systems at high temperature resemble those of a collection of single Anderson impurities in the Kondo regime. Thus one expects that in order for the Anderson lattice Hamiltonian to describe the heavy-fermion systems, the ratio U_{ff}/Δ should be large and the f level should lie well below the Fermi level. Moreover, the upper single-impurity band, E_f+U_{ff}, lies well above the Fermi level. In this situation, fluctuations of the f-site charge are strongly suppressed and fluctuations of the f-site moments are the important degrees of freedom.

Many different approximations have been developed for the Anderson lattice Hamiltonian, as reviewed by Lee *et al.* (1986) and Fulde *et al.* (1988). Of particular interest for the heavy-fermion systems are strong-coupling approximations that include direct decoupling of the hierarchy of equations of motion for the Green's functions, variational methods (Rice and Ueda, 1985), and expansions in powers of the spin and orbital degeneracy of the f levels, which are generalizations to the lattice of methods used for the single-impurity problem known as $1/N$ expansions. The decoupling schemes and variational methods, depending on initial choices of allowable physical states, are not systematic and have not been the methods of choice for most researchers. For technical reasons, much of the work on the Anderson lattice Hamiltonian (e.g., $1/N$ expansions) has been done for the limit $U_{ff} \rightarrow \infty$ in the hope of describing the essential features of strong correlations.

The expansions in powers of the spin-orbital degeneracy N are made with the assumption that N is a large number and that the expansion is rapidly convergent (Bickers 1987; Newns and Read, 1987). In the limit $N \rightarrow \infty$, a mean-field solution is expected to be exact. This mean-field solution produces a set of bands similar to those shown in Fig. 9(b). However, the position of the f level is renormalized upwards, and the magnitude of the hybridization is reduced in order to minimize the Coulomb interaction energy. Physically, the smaller the number of f electrons, the less the interaction energy. Furthermore, reducing the hybridization matrix element from its noninteracting value makes it less likely for a conduction electron to hop into a localized f orbital and interact with the electrons already present there, which also minimizes the effects of the Coulomb interaction. The above picture is that of renormalized bands, in which the width of the f states is narrowed by the Coulomb interactions. This picture is modified by including terms of higher order in $1/N$. In general, the above states are subject to further changes in their dispersion relations and also acquire a finite lifetime that is both temperature and energy dependent. The energy dependence of the lifetime can cause structure in the electronic density of states that is not directly related to the quasiparticle dispersion relations. For physical values of N, the contribution of the quasiparticle dispersion relations to the density of states is a small peak close to the Fermi energy, and the main structure occurs as a result of the energy dependence of the lifetime.

Many results obtained by the $1/N$ expansion technique can also be obtained by other methods. An example is a variational scheme based on a method of projecting out doubly occupied f states, thereby minimizing the energy of the Coulomb interaction. Rice and Ueda (1985) applied this method to the Anderson lattice model and produced an effective band model similar to that obtained by the $1/N$ expansion techniques. These authors argued that in the Kondo regime the physics should be similar to that of the Hubbard model near localization. The main results are discussed in the Landau Fermi-liquid language introduced above. They find that the divergence of the effective mass, as localization is approached, is balanced by a similar divergence in F_0^s, and thus the charge susceptibility remains unrenormalized. The parameter F_0^a that controls the tendency for magnetic ordering depends strongly on the total degeneracy of the f level. In the direct $1/N$ expansion techniques it has been found that the paramagnetic heavy-fermion state is only stable if the orbital degeneracy is large.

Although the strength of the Coulomb interaction is sufficiently large that series expansions in powers of U_{ff} are not expected to converge rapidly, studies along these lines have been pursued (Czycholl, 1986). The main motivation for applying perturbation theory to the Anderson lattice model is that for the single-impurity Anderson model with particle-hole symmetry ($2E_f + U_{ff} = 2e_{\mathrm{F}}$; see Fig. 8), perturbation expansions for certain quantities are rapidly convergent (Horvatić and Zlatić, 1985). Since the heavy-fermion state itself is apparently a Fermi liquid and the formal underpinnings of Fermi-liquid theory are described in terms of self-consistent perturbation theories that sum infinite classes of diagrams, it is further hoped that such an approach could at least describe the low-temperature limit. Numerical calculations are required to evaluate the expressions of self-consistent perturbation theories beyond Hartree-Fock theory, and it is possible to include the results of band-structure calculations without further approximation. The choice of classes of diagrams to be summed is to some extent arbitrary, although conservation laws provide guidance in selecting sets of terms. Apparently because of the difficulty of the numerical calculations (Schweitzer and Czycholl, 1990), fully self-consistent calculations have not been reported for the Anderson lattice model in three dimensions; the calculations that have been performed have been limited to second-order perturbation theory. The results show that the quasiparticle dispersion relation is strongly renormalized near the Fermi energy and that a temperature-dependent peak appears in the density of states at the Fermi level that bears a qualitative resemblance to the Kondo resonance in the single- impurity model. An unfortunate feature of these calculations is that they have been performed largely for models that include only one spatial dimension; such models may contain well-known pathologies related to low dimensionality.

Both the weak-coupling and strong-coupling schemes do yield results that have some qualitative similarities, namely a narrow temperature-dependent peak in the density of states close to the Fermi energy. However, no quantitative agreement has been reached between the strong- and weak-coupling schemes. This has led various groups to apply numerical techniques, which have given some insight into these problems. Conspicuous among these is the quantum Monte Carlo method, which has allowed the accuracy of some approximate solutions to be assessed. Unfortunately these calculations have been restricted to small systems, high temperatures, and near particle-hole symmetry. Thus, despite its deceptively simple appearance, the exact solution for the low-temperature, low-energy properties of the Anderson lattice Hamiltonian remains elusive.

2.2 Electronic Structure Calculations

The electronic structures of the heavy-fermion systems $CeAl_3$, $CeCu_2Si_2$, $CeCu_6$, UBe_{13}, UPt_3, U_2Zn_{17}, and URu_2Si_2 have all been calculated in the local-density approximation (Fulde *et al.*, 1988). This approach is based on a theorem due to Kohn and Sham (1965) that states that the energy of the fully interacting electronic system is a functional of the electronic density (see ELECTRON STATES IN SOLIDS). Electronic correlations are incorporated by the introduction of an exchange and correlation term in the energy functional, and the result of the calculations should be the exact real-space electronic charge density and the total energy. The exact form of the exchange and correlation functional is unknown, but in the local-density approximation (LDA) it is approximated by the form appropriate for a system with uniform electron density, but evaluated with the local value of the electronic density. As a by-product of the Kohn-Sham formalism a set of eigenfunctions and eigenvalues is generated; the spectrum of eigenvalues often exhibits a close resemblance to the electronic spectrum, as measured by, for example, angle-resolved photoemission. An advantage of these calculations is that crystal-field effects and spin-orbit interactions are automatically included. In the latter case, fully relativistic calculations are required. The results of these calculations share the common feature of a band with primarily f character straddling the Fermi energy. For cerium systems, the occupied $4f$ weight accounts for 1 to 1.3 f electrons, whereas for the uranium systems, the $5f$ count lies between 2 and 3 electrons per uranium atom. The total f bandwidths are of the order of 1 eV as the f character admixes to states in bands that are energetically removed from the Fermi energy. Although this is narrow when compared to the bandwidth of a typical conduction band, it is still too large to account for the order-meV energy scale deduced from the thermodynamic and transport properties of the heavy-fermion materials. Moreover, the calculated total density of states at the Fermi energy is still much smaller than that deduced from the linear temperature coefficient of the specific heat.

A tight-binding parametrization of LDA calculations is essential, if quantitative predictions incorporating many-body correlations are to be made; only a few such parametrizations have been performed. Although the entire parameter sets may vary from one fitting procedure to the next, the individual parameters do indicate the importance of hybridizing the f band with non-f bands.

2.3 Superconductivity

Perhaps the most exciting aspect of the heavy-fermion systems is that some of them might be unconventional superconductors. Various authors have proposed that an anisotropic but still conventional state occurs as a result of the strong coupling of f electrons to the lattice. This point of view is reviewed at length in Fulde *et al.* (1988).

The superconducting state is expected to be unusual at the very least. The standard theory of superconductivity relies in large part on the existence of small parameters. The difference in the thermodynamic potential between the superconducting state and normal state may be expressed as an asymptotic expansion in

T_c/T_F, where T_F is the Fermi energy. The leading term is of order $(T_c/T_F)^2$ and gives the weak-coupling BCS theory of superconductivity, which can allow both conventional and unconventional order parameters. Eliashberg theory, which treats strong electron-phonon coupling, is of order $(T_c/T_F)^2$ and also of leading order in the ratio of the electronic mass to the ionic mass. Strong-coupling corrections due to quasiparticle-quasiparticle scattering appear at the next order in (T_c/T_F). These corrections are known to be important in superfluid ^{3}He where $T_c/T_F \sim 10^{-2}$. Even though T_c is small in the heavy-fermion systems, the effective degeneracy temperature is also small, and $T_c/T_F \sim 10^{-2}$ or larger. Strong-coupling corrections due to quasiparticle-quasiparticle scattering may thus be important in the heavy-fermion superconductors; evidently they are unimportant in an ordinary superconductor where $T_c/T_F \sim 10^{-4}$ or smaller.

In a conventional BCS superconductor, the weak attractive interaction between electrons is mediated by phonons. The ratio T_D/T_F, where T_D is the Debye temperature, is small, allowing one to treat accurately even a strong electron-phonon interaction in second order. In the heavy-fermion systems, this ratio is evidently of order unity, and the validity of this theory is unclear.

The narrow bandwidth of the *f* electrons suggests that the repulsive interaction between quasiparticles due to the Coulomb interaction should be important in these materials. In liquid ^{3}He, strong short-ranged repulsive interactions between ^{3}He quasiparticles discourage Cooper pairing in a relative *s*-wave state, and the quasiparticles form a Cooper pair in a relative *p*-wave state. Analogies with liquid ^{3}He gave rise to the initial speculation that the quasiparticles forming the condensate wave function in the heavy-fermion superconductors are in a relative angular momentum state with $l > 0$. The strong intrasite repulsion between electrons on *f* sites also gives rise to spin fluctuations in the normal state. Analogy to liquid ^{3}He suggests that spin fluctuations, rather than phonons, may provide the attractive interaction between quasiparticles. The mechanism for pairing in the heavy-fermion systems and the microscopic description of the superconducting state are topics of considerable controversy and lie outside the scope of this work. We refer the reader to Lee *et al.* (1986) and Fulde *et al.* (1988) for reviews.

The determination of whether heavy-fermion superconductors are conventional or unconventional rests on the interpretation of experiments that probe the symmetry of the order parameter. The apparent absence of small parameters has prompted an approach that relies primarily on symmetry and on the most fundamental assumptions. Thus a considerable effort has been devoted to exploring the predictions of Ginzburg-Landau (GL) theories of unconventional superconductors. The connection between GL theory and microscopic theory has been studied carefully for superfluid ^{3}He, the archetypal unconventional "superconductor." Superfluid ^{3}He shows two distinct superfluid phases in the temperature-pressure phase diagram. At low pressure the *B* phase is stable, consistent with weak-coupling BCS theory. As a result of strong-coupling corrections, the *A* phase becomes stable at high pressure. The form of the Ginzburg-Landau free energy is dictated by symmetry; the value of the parameters that appear in the free-energy functional may be calculated from microscopic theory or deduced from experiments.

The normal-to-superconducting transformation can be viewed as a second-order phase transition in which quasiparticles form a condensate of Cooper pairs. A natural order parameter describing the superconducting state is the condensate wave function,

$$\langle \psi_\alpha(\mathbf{R}+\mathbf{r})\psi_\beta(\mathbf{R}-\mathbf{r}) \rangle, \tag{27}$$

where **R** and **r** are the center of mass and relative coordinates of the quasiparticles in the Cooper pair, and $\psi_\alpha(\mathbf{r})$ is a field operator that destroys a particle of spin α at **r**. It is a general property of a second-order phase transition that a symmetry of the normal state is broken in the ordered state. In the superconducting state, the gauge symmetry of the normal state is broken by the appearance of the condensate wave function. In an unconventional superconductor, not only is gauge symmetry broken, but a symmetry or symmetries of the crystalline lattice may also be broken because the condensate wave function has a lower symmetry than the crystal lattice. For example, because ^{3}He quasiparticles are paired in a spin-triplet orbital *p*-wave state, parity is broken, and ^{3}He is an unconventional superfluid.

In the heavy-fermion systems, strong spin-orbit coupling breaks full rotational symmetry in spin space, but the crystal structures have a center of inversion symmetry, so that the notions of "singlet" and "triplet" states should be replaced by "even-parity" and "odd-parity" states. The condensate wave function may still be written as above if α and β are interpreted as pseudospin labels.

The general theory of second-order phase transitions allows the possible unconventional superconducting states to be classified according to the symmetry of the condensate wave function, without making assumptions about the nature of the pairing interaction or the extent of strong-coupling corrections to the superconducting state. Here, we give a heuristic explanation of the symmetry classification of the superconducting order parameter, based on weak-coupling BCS theory. A gap matrix related to the condensate wave function is defined by the expression

$$\Delta_{\mathbf{k},\alpha\beta}=\sum_{\mathbf{k}'} V^{\alpha\beta;\gamma\delta}(\mathbf{k},\mathbf{k}')\langle\psi_\gamma(\mathbf{k})\psi_\delta(\mathbf{k}')\rangle, \qquad (28)$$

where $\langle\psi_\gamma(\mathbf{k})\psi_\delta(\mathbf{k}')\rangle$ is the Fourier transform of the condensate wave function and $V^{\alpha\beta;\gamma\delta}(\mathbf{k},\mathbf{k}')$ is the generally spin-dependent pairing interaction. The gap equation separates by symmetry representation, and there is a transition temperature T_c^i for the ith representation. Generally only one representation will occur at the physical T_c, although more than one may be present because of the occurrence of "accidental degeneracies." For order parameters of lower symmetry, the nonlinear gap equation can couple order parameters from different representations so that other representations may be induced below T_c. This has been discussed for ^{3}He by Leggett (1975) and reviewed in more detail along with the symmetry classification by Sigrist and Ueda (1991). A group-theoretical analysis for strong spin-orbit coupling is given by Volovik and Gor'kov (1985) and Blount (1985), and the weak spin-orbit coupling case is discussed by Ozaki *et al.* (1985). A GL theory of the superconducting state can be constructed once the symmetry of the order parameter is known.

As we have noted, at sufficiently low temperatures it is generally accepted that the heavy-fermion systems behave like Landau Fermi liquids. Many authors have used this as justification to proceed with a BCS description of the superconducting state (Sigrist and Ueda, 1991; Lee *et al.*, 1986). The connection between the condensate wave function and the quasiparticle dispersion relation is dependent on the theoretical description of the superconducting state. In BCS theory for a singlet superconductor, the quasiparticle energy is related to the gap function $\Delta(\mathbf{k})$ through

$$E_{\mathbf{k}}=\sqrt{\epsilon_{\mathbf{k}}^2+|\Delta(\mathbf{k})|^2}. \qquad (29)$$

The gap function of an unconventional superconductor may vanish for points or lines on the Fermi surface, which can be determined from the symmetry classification. Such an analysis shows (Blount, 1985) that for strong spin-orbit coupling, odd-parity states have no nodal lines. The thermodynamic and transport data described above suggest that the gap function of the heavy-fermion superconductors vanishes on points or lines, giving rise to power-law (rather than thermally activated) behavior at low temperatures. Detailed calculations of these properties for various unconventional superconducting states show a strong dependence of the "power law" on impurity scattering (see Sigrist and Ueda, 1991), making an unambiguous identification of the superconducting state impossible.

Other experiments have been proposed to probe the symmetry of the order parameter. A GL-theory analysis of upper critical fields by Gor'kov and by Burlachkov (see Gor'kov, 1987) showed that the reduced symmetry of an unconventional order parameter can lead to unusual upper-critical-field anisotropies. For example, the upper critical field in the basal plane of a tetragonal crystal might appear in contour like a four-leaf clover. Similar anisotropies would be apparent in H_{c1}.

The order parameter describing an unconventional superconductor often has more degrees of freedom than its conventional counterpart, and there often exist degenerate superconducting states. Some tests for unconventional superconductivity rest on the consequences of lifting a degeneracy between different superconducting states. Joynt and Rice (1985) showed that the condensation to an unconventional superconducting phase described by an order parameter with lower symmetry than the crystal lattice would be accompanied by a lattice distortion. This effect can be "inverted" by applying a uniaxial stress that induces a splitting of the superconducting transition. The transition to the unconventional superconducting state may break time-reversal symmetry, for example, resulting in a weak magnetic moment, which may explain the small magnetic moment observed below the lower-temperature transition in $U_{1-x}Th_xBe_{13}$. Tokuyasu *et al.* (1990) showed that broken time-reversal symmetry can also result in an unusual anisotropy of H_{c1}. Reversing the direction of the field applied along a high-symmetry direction in an unconventional superconductor with broken time-reversal symmetry creates a vortex state that is inequivalent to that for the original field direction, resulting in a different value for H_{c1}. For this effect to occur, it is crucial that one time-reversal symmetry-breaking state cannot be continuously deformed into the other as is the case in a crystal of uniaxial symmetry. Other tests are reviewed in Gor'kov (1987) and Sigrist and Ueda (1991).

As a superconductor is cooled through the phase transition, different states may nucleate in different areas of the sample and domain walls form. Generally, the presence of domains obscures the simple tests described above. For example, the anisotropy of the upper critical field can be obscured by randomly oriented superconducting domains.

Perhaps the most compelling evidence for unconventional superconductivity is the intriguing *H-T* phase diagram of UPt_3 (see Fig. 7) and the thorium-substitution phase diagram of $U_{1-x}Th_xBe_{13}$ (see Fig. 6). The explanation of these phase diagrams is controversial; however, we will discuss the most developed ideas in the area to illustrate some additional tests for unconventional superconductivity.

A promising model for UPt_3 assumes that the order parameter η has two complex components as allowed by the hexagonal symmetry of the crystal together with the assumption of strong spin-orbit coupling (Tokuyasu *et al.*, 1990). Numerical solutions of the Ginzburg-Landau equations (Tokuyasu *et al.*, 1990; Tokuyasu and Sauls, 1990) for vortices along the *c* axis in the time-reversal symmetry-breaking homogeneous phases $\eta^{\pm} \sim (1, \pm i)$ suggest an explanation for the line separating phases I and II in Fig. 7. For a range of parameters entering the GL free energy, Tokuyasu showed that the low-field solution of the GL equations is a triangular lattice of vortices each axially symmetric, threaded by two quanta of magnetic flux, and having a finite superfluid density in the core. At a higher field ($\sim 0.3 H_{c2}$), this lattice becomes unstable and abruptly transforms into a distorted lattice of single–flux-quantum vortices. A phase line like that shown separating phases II and III in Fig. 7 occurs as a result of the coupling of the order parameter to a symmetry-breaking field that lifts the twofold degeneracy of the pairing interaction (see for example, Hess *et al.*, 1989; Garg, 1992). In zero field, the phase transition is split into two: a transition from the normal state to a superconducting state described by a single complex component, e.g., $\eta \sim (1,0)$, and a lower-temperature transition from the superconducting phase to another superconducting phase that approaches η^+ or η^- with decreasing temperature. A natural candidate for the symmetry-breaking field is the weak antiferromagnetic order in the basal plane that has recently been correlated with the splitting of T_c (Hayden *et al.*, 1992). The additional phase lines in the mixed state may not be true phase lines but may rather mark regions of rapid crossover (Garg, 1992).

Attempts to understand the *x-T* phase diagram of $U_{1-x}Th_xBe_{13}$ assume that superconducting states belonging to two different representations are involved. A simple model with coupled *s*- and *d*-wave order parameters was considered by Kumar and Wölfle (1987). In their picture, pure UBe_{13} is a *d*-wave superconductor in the vicinity of the phase transition. As the temperature is lowered, *s*-wave character mixes in smoothly without introducing a second phase transition. Since a gap of lower symmetry is generally more sensitive to impurities, the effect of doping is to reduce the *d*-wave transition temperature dramatically so that above the critical concentration, the high-temperature transition is to an *s*-wave state. At lower temperatures a transition to a state described by a *d*-wave order parameter occurs. Sigrist and Rice (1989) took into account the cubic symmetry of the crystal and supposed that the pairing interaction leads to order parameters for two or more representations having nearly the same transition temperature. They explored the simple possibilities for various representations and concluded that the most promising phase diagram has an order parameter belonging to the Γ_1 (or Γ_5)

(one- or three-dimensional, respectively) representation for $x<x_c$, while for $x>x_c$ and $T>T_{c2}$ (see Fig. 6) the Γ_5 (or Γ_1) representation appears. For $T<T_{c2}$ the order parameter breaks time-reversal symmetry and belongs to a mixture of the two representations.

3. SUMMARY

The heavy-fermion materials are characterized by *f* electrons that exhibit localized behavior at high temperatures and itinerant behavior at low temperatures. The transformation from one regime to another suggests the existence of a characteristic energy scale. The entropy associated with the high-temperature magnetic moments is retained down to very low temperatures and appears in a specific heat that goes linearly with temperature. The slope of the specific heat at low temperature is anomalously large and suggests a narrow band of electrons characterized by an enormous effective mass. This state appears to be well described by Fermi-liquid theory, but it is unclear how such a state is formed from a lattice of weakly interacting magnetic moments. Taken as a dense lattice of local moments, the single-impurity Anderson model seems to be useful in discussing the high-temperature properties of these materials. Approximate methods applied to a model containing a dense lattice of Anderson impurities appear to yield the essential physics required to understand the heavy-fermion phenomena.

Many of these materials display unusual ordered ground states at low temperatures. Particularly interesting is the superconducting state that can apparently coexist with weak antiferromagnetic order. The unusual nature of the superconductivity suggests an order parameter of lower symmetry than that of the crystal lattice and a pairing interaction of electronic origin. The magnetic state is characterized by small magnetic moments associated with the *f*-atom sites and a simple magnetic structure. Although the magnitude of the magnetic moments is intimately connected with the removal of residual entropy, it is utterly unclear why so little entropy is associated with these transitions.

ACKNOWLEDGMENTS

We would like to thank J. D. Thompson for helpful conversations. We are also indebted to A. Y. Liu for useful suggestions and to J. W. Serene for a critical reading of the manuscript. Work at Los Alamos performed under the auspices of the U.S. Department of Energy.

GLOSSARY

BCS Theory of Superconductivity: Microscopic weak-coupling mean-field theory of superconductivity developed by Bardeen, Cooper, and Schrieffer in which electrons of opposite spin and momenta are paired by a weak attractive interaction.

Effective Mass: A parameter introduced as a renormalization of the electronic mass that takes into account the effect of the crystalline environment on the electronic energy spectrum. It includes the effect of interactions with other electrons and with the periodic potential of the atomic lattice.

Fermi Liquid: A macroscopic system of strongly interacting Fermi particles that can be described for energies and temperatures small compared to the degeneracy temperature as a gas of Fermi excitations (quasiparticles) with an effective mass different from that of a free constituent particle.

Hartree-Fock Approximation: An approximate treatment of the interactions between particles, obtained by replacing the effect of the other particles with a self-consistently chosen potential.

Hybridization: A quantum-mechanical admixture of atomic single-particle states, generally of differing symmetry.

Kondo Problem: The problem of describing a local magnetic impurity in a simple metallic environment.

Local-Density Approximation: An approximation made within the framework of density-functional theory for calculating electronic structure, in which the term representing the effects of electronic interactions is approximated by a quantity that depends only on the local electron density.

Mixed-Valence Materials: Materials containing elements of the lanthanide series, notably cerium, samarium, europium, thulium, and ytterbium, which show properties characteristic of a strong admixture of two or more *f* configurations in the ground state.

Unconventional Superconductivity: A superconducting state with an order parameter that breaks not only gauge symmetry (because the pair wave function is complex), but also some other symmetry of the crystal point group. In a conventional superconductor, only gauge symmetry is broken.

List of Works Cited

Adenwalla, S., Lin, S. W., Ran, Q. Z., Zhao, Z., Ketterson, J. B., Sauls, J. A., Taillefer, L., Hinks, D., Levy, M., Sarma, B. K. (1990), *Phys. Rev. Lett.* **65**, 2298–2301.

Aeppli, G., Bucher, E., Goldman, A. I., Shirane, G., Broholm, C., Kjems, J. K. (1988), *J. Magn. Magn. Mater.* **76&77**, 385–390.

Allen, J. W. (1988), *J. Magn. Magn. Mater.* **76&77**, 324–330.

Allen, J. W., Oh, S.-J., Cox, L. E., Ellis, W. P., Wire, M. S., Fisk, Z., Smith, J. L., Pate, B. B., Lindau, I., Arko, A. J. (1985), *Phys. Rev. Lett.* **54**, 2635–2638.

Allen, J. W., Oh, S.-J., Gunnarson, O., Schönhammer, K., Maple, M. B., Torikachvili, M. S., Lindau, I. (1986) *Adv. Phys.* **35**, 275–316.

Amato, A., Feyerherm, R., Gygax, F. N., Schenck, A., Weber, M., Caspary, R., Hellman, P., Schank, C., Geibel, C., Steiglich, F., MacLaughlin, D. E., Knetsch, E. A., Heffner, R. H. (1992), *Europhys. Lett.* **19**, 127–133.

Anderson, P. W. (1961) *Phys. Rev.* **124**, 41–53.

Andres, K., Graebner, J. E., Ott, H. R., (1975), *Phys. Rev. Lett.* **35**, 1779–1782.

Arko, A. J., Koelling, D. D., Reihl, B. (1983), *Phys. Rev. B* **27**, 3955–3961.

Arko, A. J., Koelling, D. D., Dunlap, B. D., Mitchell, A. W. (1988), *J. Appl. Phys.* **63**, 3680–3682.

Baym, G., Pethick, C. J. (1978), in: K. H. Bennemann, J. B. Ketterson (Eds.), *The Physics of Liquid and Solid Helium*, Vol. 2, New York: Wiley, p. 1.

Beyermann, W. P., Hundley, M. F., Canfield, P. C., Thompson, J. D., Fisk, Z., Smith, J. L., Selsane, M., Godart, C., Latroche, M. (1991), *Phys. Rev. Lett.* **66**, 3289–3692.

Bickers, N. E., (1987), *Rev. Mod. Phys.* **59**, 845–939.

Blount, E. I., (1985), *Phys. Rev. B* **32**, 2935–2944.

Brodale, G. E., Fisher, R. A., Phillips, N. E., Stewart, G. R., Giorgi, A. L. (1986), *Phys. Rev. Lett.* **57**, 234–237.

Bruls, G., Weber, D., Wolf, B., Thalmeier, P., Lüthi, B., de Visser, A., Menovsky, A. (1990), *Phys. Rev. Lett.* **65**, 2294–2297.

Carneiro, G. M., Pethick, C. J. (1975), *Phys. Rev. B* **11**, 1106–1124.

Chapman, S., Hunt, M., Meeson, P., Reinders, P. H. P., Springford, M., Norman, M. (1990), *J. Phys. Condens. Matter* **2**, 8123–8136.

Czycholl, G. (1986), *Phys. Rep.* **143**, 277–346.

Fisher, R. A., Kim, S., Woodfield, B. W., Phillips, N. E., Taillefer, L., Hasselbach, K., Floquet, J., Giorgi, A. L., Smith, J. L. (1989), *Phys. Rev. Lett.* **62**, 1411–1414.

Fisk, Z., Smith, J. L., Ott, H. R., Batlogg, B. (1985), *J. Mag. Magn. Mater.* **52**, 79–84.

Fisk, Z., Hess, D. W., Pethick, C. J., Pines, D., Smith, J. L., Thompson, J. D., Willis, J. O. (1988), *Science* **239**, 33–42.

Freeman, A. J., Min, B. I., Norman, M. R. (1987), in: K. A. Gschneidner, Jr., L. Eyring, S. Hufner (Eds.), *The Handbook on the Physics and Chemistry of the Rare-Earths*, Vol. 10, Amsterdam: North-Holland.

Fulde, P., Keller, J., Zwicknagl, G. (1988), in: H. Ehrenreich and D. Turnbull (Eds.), *Solid State Physics* Vol. 41, New York; Academic.

Garg, A. (1992), *Phys. Rev. Lett.* **69**, 676–679.

Geibel, C., Schank, C., Thies, S., Kitazawa, H., Bredl, C. D., Böhm, A., Rau, M., Grauel, A., Caspary, R., Helfrich, R., Ahlheim, U., Weber, G., Steglich, F. (1991) *Z. Phys. B.* **84**, 1-2.

Gor'kov, L. P. (1987), *Sov. Sci. Rev. A Phys.* **9**, 1-116.

Gutzwiller, M. C. (1965), *Phys. Rev.* **137**, A1726–A1735.

Haldane, F. D. M. (1978), *Phys. Rev. Lett.* **40**, 416–419.

Hayden, S. N., Taillefer, L., Vettier, C., Floquet, J. (1992) *Phys. Rev. B* **46**, 8675-8678.

Hess, D. W., Tokuyasu, T. A., Sauls, J. A. (1989), *J. Phys. Condens. Matter* **1**, 8135–8145.

Horvatić, B., Zlatić, V. (1985), *Solid State Commun.* **54**, 957–960.

Hundley, M. F., Canfield, P. C., Thompson, J. D., Fisk, Z. (1990), *Phys. Rev. B* **42**, 6842–6845.

Jones, B. A., Varma, C. M., Wilkins, J. W. (1988), *Phys. Rev. Lett.* **61**, 125–128.

Joynt, R., Rice, T. M. (1985), *Phys. Rev. B* **32**, 6074–6076.

Kjems, J. K., Broholm, C. (1988), *J. Magn. Magn. Mater.* **76&77**, 371–375.

Kohn, W., Sham, L. J. (1965), *Phys. Rev.* **140**, A1133–A1138.

Krishnamurthy, H. R., Wilkins, J. W., Wilson, K. G. (1980), *Phys. Rev. B* **21**, 1003–1044.

Kumar, P., Wölfle, P. (1987), *Phys. Rev. Lett.* **59**, 1954–1957.

Lee, P. A., Rice, T. M., Serene, J. W., Sham, L. J, Wilkins, J. W. (1986), *Comments Condens. Mater. Phys.* **12**, 99.

Leggett, A. J. (1975), *Rev. Mod. Phys.* 47, 331–414.

Liu, S. H., Ho, K.-M. (1982), *Phys. Rev. B* **26**, 7052–7055.

Lüthi, B. (1985), *J. Magn. Magn. Mater.* **52**, 70–78.

MacLaughlin, D. E. (1985), *J. Magn. Magn. Mater.* **47&48**, 121-126.

Newns, D. M., Read, N. (1987), *Adv. Phys.* **36**, 799–849.

Nozières, P. (1974), *J. Low Temp. Phys.* **17**, 31–42.

Nozières, P. (1985), *Ann. Phys. (Paris)* **10**, 19–35.

Ott, H. R., Fisk, Z. (1987) in: A. J. Freeman, G. H. Lander (Eds.), *Handbook on the Physics and Chemistry of the Actinides*, Vol. 5, Amsterdam: North-Holland.

Ott, H. R., Rudigier, H., Fisk, Z., Smith, J. L. (1983), *Phys. Rev. Lett.* **50**, 1595–1598.

Ozaki, M., Machida, K., Ohmi, T. (1985), *Prog. Theor. Phys.* **74**, 221–235.

Palstra, T. T. M., Menovsky, A. A., van den Berg, J., Dirkmaat, A. J., Kes, P. H., Nieuwenhuys, G. J., Mydosh, J. A. (1985) *Phys. Rev. Lett.* **55**, 2727-2730.

Parks, R. D. (1969), *Superconductivity*, 2 volumes, New York: Marcel Dekker.

Pethick, C. J., Pines, D. (1987) in: P. Siemens, R. Smith (Eds.), *Proceedings of the International Conference on Recent Progress in Many-Body Theoretical Physics*, Vol. 4, Berlin: Springer Verlag.

Rainer, D. (1988), *Phys. Scr.* **T23**, 106–112.

Rice, T. M., Ueda, K. (1985), *Phys. Rev. Lett.* **55**, 995–998.

Rossat-Mignod, J., Regnault, L. P., Jacoud, J. L., Vettier, C., Lejay, P., Floquet, J., Walker, E., Jaccard, D., Amato, A. (1988), *J. Magn. Magn. Mater.* **76&77**, 376–384.

Rossel, C., Yang, K. N., Maple, M. B., Fisk, Z., Zirngieble, E., Thompson, J. D. (1987), *Phys. Rev. B* **35**, 1914–1918.

Schlottmann, P. (1989), *Phys. Rep.* **181**, 1–119.

Schoenes, J., Franse, J. J. M. (1986), *Phys. Rev. B* **33**, 5138–5140.

Schrieffer, J. R. (1964), *Theory of Superconductivity*, Redwood City, CA: Addison-Wesley.

Schrieffer, J. R., Wolff, P. A. (1966), *Phys. Rev.* **149**, 491–492.

Schweitzer, H., Czycholl, B. (1990), *Solid State Commun.* **74**, 735–742.

Sigrist, M., Rice, T. M., (1989), *Phys. Rev. B* **39**, 2200–2216.

Sigrist, M., Ueda, K. (1991), *Rev. Mod. Phys.* **63**, 239–311.

Slichter, C. P. (1978), *Principles of Magnetic Resonance*, Berlin: Springer, p. 156.

Smith, J. L. (1992), *Philos Mag. B* **65**, 1367–1371.

Steglich, F., Aarts, J., Bredl, C. D., Lieke, W., Meschede, D., Franz, W., Schäfer, J. (1979), *Phys. Rev. Lett.* **43**, 1892–1896.

Stewart, G. R., (1984), *Rev. Mod. Phys.* **56**, 755–787.

Taillefer, L., Newbury, R., Lonzarich, G. G., Fisk, Z., Smith, J. L. (1987), *J. Magn. Magn. Mater.* **63&64**, 372–376.

Takabatake, T., Teshima, F., Fujii, H., Nishigori, S., Suzuki, T., Fujita, T., Yamaguchi, Y., Sakuri, J. (1990), *Phys. Rev. B* **41**, 9607–9610.

Tien, C., MacLaughlin, D. E., Lan, M. D., Clark, W. G., Fisk, Z., Smith, J. L., Ott, H. R. (1985), *Physica B 135*, 14–21.

Tokuyasu, T. A., Sauls, J. A. (1990), *Physica B* **165&166**, 347–348.

Tokuyasu, T. A., Hess, D. W., Sauls, J. A. (1990), *Phys. Rev. B* **41**, 8891–8903.

Tsvelick, A. M, Wiegman, P. B. (1983), *Adv. Phys.* **32**, 453–713.

de Visser, A. (1986), Ph.D. thesis, University of Amsterdam, unpublished.

Volovik, G. E., Gor'kov, L. P. (1985) *Zh. Eksp. Teor. Fiz.* **88**, 1412–1428 [*Sov. Phys. JETP* **61**, 843–854 (1985)].

Wang, C. S., Norman, M. R., Albers, R. C., Boring, A. M., Pickett, W. E., Krakauer, H., Christensen, N. E. (1987), *Phys. Rev. B* **35**, 7260–7263.

Weber, D., Yoshizawa, M. Kouroudis, I., Lüthi, B., Walker, E. (1987) *Europhys. Lett.* **3**, 827-831.

Yoshizawa, M., Lüthi, B., Goto, T., Suzuki, T., Renker, B., de Visser, A., Frings, P., Franse, J. J. M. (1985), *J. Magn. Magn. Mater.* **52**, 413-417.

Further Reading

Fisk, Z., Ott, H. R., Rice, T. M., Smith, J. L. (1986), *Nature* **320**, 124-129. A valuable experimental review with emphasis on the superconductors.

Fisk, Z., Hess, D. W., Pethick, C. J., Pines, D., Smith, J. L., Thompson, J. D., Willis, J. O. (1988), *Science* **239**, 33-42. A combined theory and experiment overview with some emphasis on the Fermi-liquid character of the low-temperature heavy-fermion state.

Fulde, P., Keller, J., Zwicknagel, G. (1988), in: H. Ehrenreich, D. Turnbull (Eds.), *Solid State Physics*, Vol. 41, New York: Academic. A summary of efforts to formulate a theory of heavy-fermion systems, more recent than the following.

Gor'kov, L. P. (1987), *Sov. Sci. Rev. A Phys.* **9**, 1-116. A review from a theoretical perspective emphasizing unconventional superconductivity.

Grewe, N., Steglich, F. (1990), in: K. A. Gschneidner, L. Eyring (Eds.), *Handbook on the Physics and Chemistry of the Rare Earths*, Vol. 13, New York: Elsevier. Another summary like that of Fulde *et al.*

Lee, P. A., Rice, T. M., Serene, J. W., Sham, L. J., Wilkins, J. W. (1986), *Comments Condens.-Matter Phys.* **12**, 99-154. A critical review of efforts to formulate a theory of heavy-fermion systems.

Ott, H. R., Fisk, Z. (1987), in: A. J. Freeman, G. H. Lander (Eds.), *Handbook on the Physics and Chemistry of the Actinides*, Vol. 5, Amsterdam:

North-Holland. This and the following provide valuable experimental overviews and insights.

Ott, H. R. (1987), in: D. F. Brewer (Ed.), *Progress in Low Temperature Physics XI,* Amsterdam: North-Holland, pp. 215-289.

Sarma, B. K., Levy, S., Adenwalla, S., Ketterson, B. (1992), in: M. Levy (Ed.), *Physical Acoustics,* Vol. 20, Boston: Academic. A review of heavy-fermion superconductors with particular emphasis on sound experiments.

Sigrist, M., Ueda, K. (1991), Rev. Mod. Phys. **63**, 239-311. A review at some length of unconventional superconductivity.

Stewart, G. R. (1984), Rev. Mod. Phys. **56**, 755-787. A review of experiments up to that time.

HETEROSTRUCTURES AND SUPERLATTICES, METALLIC

CHARLES M. FALCO, *Department of Physics, Optical Sciences Center, and Arizona Research Laboratories, University of Arizona, Tucson, Arizona, U.S.A.*

Introduction 465
1. **Preparation Techniques** 466
1.1 Introduction 466
1.2 Sputtering 467
1.3 Evaporation 467
1.4 Molecular-Beam Epitaxy 468
2. **Structural Perfection of Metallic Multilayers and Superlattices** 469
3. **Applications of Metallic Multilayers and Superlattices** 470
3.1 Introduction 470
3.2 Magnetism 470
3.3 Two-Dimensional Magnetic Behavior 471
3.4 Magneto-optics (Perpendicular Magnetization) 472
3.5 Superconductivity 472
3.6 X-Ray Optics 473
4. **Discussion** 474
5. **Future of Metallic Superlattices** 475
6. **Summary** 475
Glossary 475
Works Cited 476
Further Reading 476

INTRODUCTION

Many technical advances in vacuum technology and several vapor deposition techniques [see CVD (CHEMICAL VAPOR DEPOSITION)] over the past two decades have now made possible the sequential monolayer-by-monolayer deposition of more than one material (elements, alloys, or compounds). Exploitation of this capability for the synthesis of nearly flawless semiconductor superlattices has been highly successful, as discussed in the article HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR.

Early work on artifically layered metals was reported nearly sixty years ago by DuMond and Youtz (1935). Their goal was to mimic the periodic structure of atoms in natural crystals, but with the longer repeat distances necessary to diffract x rays of wavelength λ longer than ≈ 10 Å. This goal of making artificial crystals for soft–x-ray optics has motivated a series of workers up to the present. However, it was discovered very early that interdiffusion greatly degraded most of the multilayered structures produced. This led to a parallel series of studies, whereby many workers used these materials for the study of physics of diffusion.

During the decade starting in the late 1960s tremendous progress was made in the growth of high-quality semiconductor heterostructures, and so by the late 1970s interest in such compositionally modulated materials expanded to include metals. Advances in thin-film deposition technology over the past two decades also motivated renewed interest in attempting to prepare "artificial" multilayered metals that might exhibit different physical properties than found in materials that are naturally occurring. Actually, as in many fields of science, it is not possible to identify a single development that was the "start" of this field. As shown in Fig. 1, "recent" research on metallic multilayers and superlattices dates back to at least the early 1970s. It should be pointed out that the two curves in Fig. 1 included different types of data. For example, the upper curve includes all materials grown by molecular-beam epitaxy (MBE), not just heterostructures and superlattices. However, it is interesting to note how closely the fields have paralleled each other in growth. The exponential growth in research on MBE-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

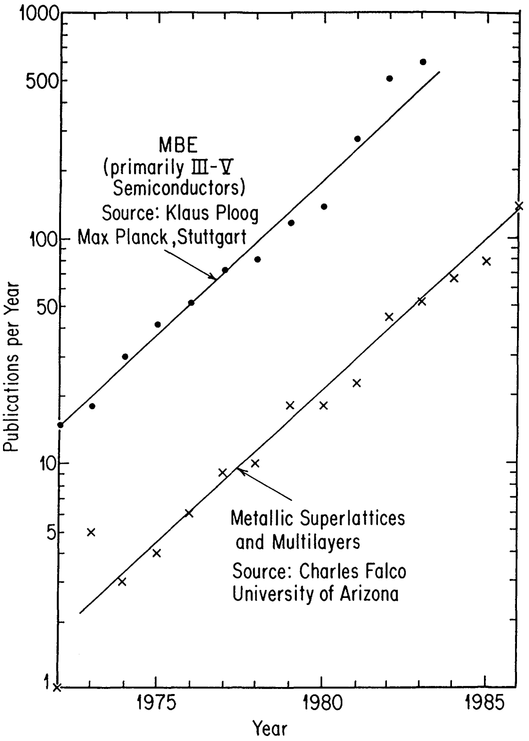

FIG. 1. Number of articles on metallic and semiconducting multilayers and superlattices vs. time.

grown semiconductors is in large part due to the practical devices arising from this research. Since various applications of metallic superlattices have started to appear over the past few years, this should stimulate additional research in this field, and should mean a strong future for both basic and applied research on metallic superlattices. A few examples of applications of metallic multilayers discussed in this article are optics for soft x rays, magneto-optic data recording, and magnetoresistive magnetic field sensors.

1. PREPARATION TECHNIQUES

1.1 Introduction

Metallic multilayers and superlattices were first made by evaporation, followed by use of sputtering techniques, and then molecular-beam epitaxy (*q.v.*). Since evaporation and MBE techniques share some features with each other, distinction between them in this article is made depending upon whether pressures are in the 10^{-10} Torr or below range, as well as whether *in situ* analysis techniques are available, such as reflected high- and low-energy electron diffraction (RHEED and LEED), Auger spectroscopy, etc. Simple evaporation is no longer used very much for preparing the types and quality of samples needed for research in the field of metallic superlattices, although such samples can exhibit very useful properties. With appropriate controls, sputtering remains a valuable deposition technique (Falco, 1984).

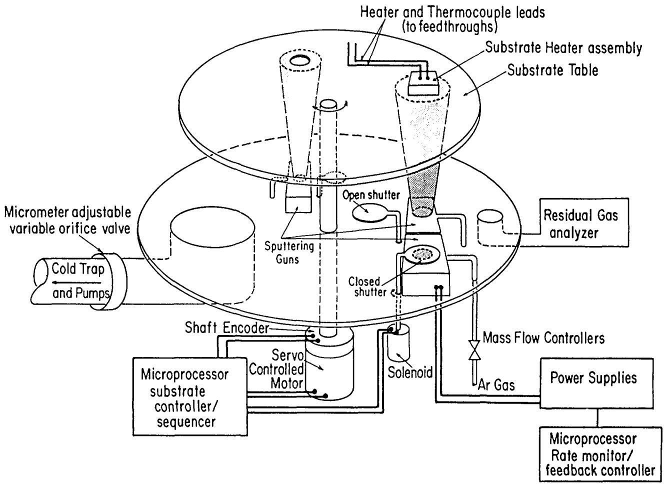
Heater and Thermocouple leads
(to feedthroughs)
Substrate Heater assembly
Substrate Table
Open shutter
Micrometer adjustable
variable orifice valve
Residual Gas
analyzer
Cold Trap
and Pumps
Sputtering
Guns
Closed
shutter
Shaft Encoder
Mass Flow Controllers
Servo
Controlled
Motor
Solenoid
Ar Gas
Power Supplies
Microprocessor
substrate
controller/
sequencer
Microprocessor
Rate monitor/
feedback controller

FIG. 2. A schematic diagram of a sputtering system used for the preparation of metallic superlattices.

1.2 Sputtering

A schematic diagram of a sputtering system used for the preparation of metallic superlattices is shown in Fig. 2. The base pressure of most such systems in current use by researchers in this field range from $<1\times10^{-7}$ to $<1\times10^{-9}$ Torr. Such deposition systems are capable of depositing many metals with purity equivalent to the starting target material. Argon is most commonly used as the sputtering gas, with pressures typically $\approx(1–5)\times10^{-3}$ Torr. Sputtering rates are only weakly pressure dependent in this pressure range, so that using gas flow controllers to hold pressure constant to $\pm0.5\%$ is adequate to reduce pressure-induced sputtering-rate variations to a negligible level.

One technique to produce multilayers is to hold a substrate stationary above two or more sputtering guns, and use solenoid-actuated shutters above each gun to expose the substrate alternately to the flux of particles from the guns. Another technique is to attach the substrate to a rotating platform, and use a motor to rotate the substrate alternately above each sputtering gun at controlled angular rotation rates. Variations in layer thickness due to rotation rate or to timing errors can easily be made negligible.

There are two sources of variation in superlattice layer thickness that occur for sputtered samples. The first is the variation during a deposition due to fluctuations or drift in the sputtering rate. Feedback control of the sputtering power supplies is capable of regulating the sputtering rates (typically 20–100 Å/s) to $\pm0.1\%$. This is the dominant source of layer-thickness error in sputtering these materials. The second error is caused by the spatial variation in the sputtering rates at the substrate because of sputtering targets that are not infinite in size. This latter problem also occurs for growth of multilayers in an MBE system. Because of the finite size of any sputtering target, absolute uniformity of layer thicknesses to arbitrary accuracy over the entire substrate is impossible to achieve unless a complex rotating holder is used for moving the substrates during deposition. In practice, variations in layer thickness of less than 1% over substrates several cm in size are easily obtainable.

A significant advantage of sputtering for the growth of metallic superlattices is that most sputtering systems are designed to allow rapid interchanging of targets. Because of this, high-quality sputtered superlattices of a wide variety of materials have been grown. A few of the metallic superlattices fabricated by sputtering and studied in detail are Nb/Cu, Nb/Al, Mo/Ta, Mo/Si, W/Si, Fe/Mo, Fe/W, Fe/Tb, Fe/Pd, Co/Cu, Co/Pd, and Co/Ti.

Although sputtering has many advantages, especially for refractory metals, there remain various limitations of the technique. For example, since *in situ* structural analysis techniques such as RHEED and LEED cannot be used during growth by sputtering, it is difficult to determine the optimum deposition conditions. Also, although the base pressure of a sputtering system can be very good (e.g., $<10^{-8}$ Torr), this is still much too high for studies of surface properties of freshly prepared films to take place before contamination occurs. Another problem is the high thermal conductivity of the Ar sputtering gas, which conducts a substantial amount of heat away from the substrate and therefore makes achieving substrate temperatures higher than a few hundred °C much more difficult than is possible in a high vacuum (where temperatures greater than 1000 °C are routinely obtainable). Such high substrate temperatures are expected to be required for many refractory materials, to give the material sufficient surface mobility to achieve epitaxy. Limitations such as these make it uncertain at present whether the sputtering process intrinsically is incapable of producing metallic superlattices with the highest purity, sharpest interfaces, largest crystalllites, etc.

1.3 Evaporation

Evaporation in vacuum systems with base pressures $>10^{-8}$ Torr using thermally heated ovens or electron-beam guns has been used successfully to prepare a number of metallic multilayers with interesting physical properties. Just as for sputtering, in the evaporation technique the materials are alternately deposited on the substrate by means of shutters, or by rotating the substrates successively over the evaporation sources. In principle, producing metallic multilayer samples with individual layer thicknesses down to the monolayer range is possible. Although base pressures $>10^{-8}$ Torr are used, the concentration of impurities incorporated in the metallic films due to the relatively high pressure, which would be unacceptable for a semiconductor heterostructure, still has allowed a number of interesting phenomena to be observed.

1.4 Molecular-Beam Epitaxy

Because of their high cost, fewer than 20 MBE machines are currently (1992) in use worldwide for the fabrication of metallic superlattices. However, just as for semiconductor heterostructures, MBE is a powerful growth technique, and should find increasing use as further applications of these materials are identified.

Most MBE machines currently in use were designed around the requirements for producing GaAs heterostructures. However, the requirements for a "metals MBE" machine are considerably different than for a GaAs machine. For example, many metals of interest have high melting points, so that reasonable vapor pressures cannot be achieved using standard thermal evaporation sources. Consequently, electron-beam evaporation guns must be incorporated in the growth chamber. Also, the substrate temperature necessary to achieve epitaxy of GaAs is relatively low, so that higher-temperature platforms for the epitaxy of refractory metals are needed. These and other requirements are very similar to those needed for the epitaxy of silicon. Consequently, MBE systems based on Si MBE machines are very appropriate for producing metallic superlattices. A schematic diagram of such a machine is shown in Fig. 3.

The MBE machine shown in Fig. 3 has a number of capabilities specifically for epitaxial growth of refractory materials. For example, the 3-in.-diam substrate is held by a manipulator inside the growth chamber, which operates at temperatures up to 1100 °C (with ±5 °C uniformity) while rotating at 120 rpm (resulting in less than ±0.5% spatial variation of thickness). The substrate may be electrically biased at up to 4 kV if desired.

Whereas vacuums in the 10^{-8}-Torr range are relatively easy to obtain, to reach the $<10^{-10}$ Torr range requires "baking" the entire MBE system at 200 °C for several days after each time it is opened to the room environment. This bakeout is necessary to drive off the water vapor that adsorbs on the vacuum chamber walls. Because of the necessity of this bakeout procedure, it is impractical to open the machine very often to change the materials being deposited. This is a disadvantage compared with sputtering. Consequently, MBE machines typically incorporate a number of deposition sources so that different materials may be grown without having to open the system to change targets. An MBE machine with eight separate deposition sources gives the capability to grow up to 28 different two-at-a-time combinations of materials without having to break vacuum to change sources. Such machines may be opened to the room environment only once or twice each year, to replenish or change materials and to conduct routine maintenance.

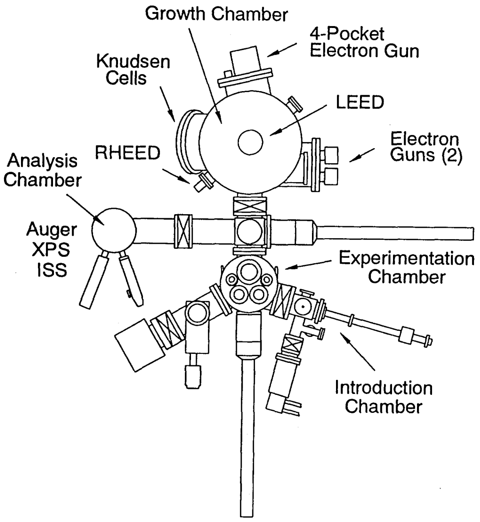

FIG. 3. Schematic diagram of an MBE system for producing metallic superlattices.

The ability to analyze the structural perfection and chemical composition of the multilayers as they are grown, or before they are removed from the ultrahigh-vacuum (UHV) environment, is essential. Because of the UHV environment of MBE machines, a number of *in situ* analytical instruments for studying the multilayers can be attached to the machine. Monitoring the physical structure normally is done during growth by RHEED, and quantitative determination of surface structure is obtained immediately after growth by LEED.

Immediately after growth, information on chemical composition can be obtained by analytical techniques such as ion-mill Auger-electron spectroscopy, x-ray photoelectron spectroscopy (XPS), and ion-scattering spectroscopy (ISS). Increasingly, UHV versions of scanning tunneling microscopes (STM) are being added to MBE machines to study the surface structure of the materials before and after deposition.

2. STRUCTURAL PERFECTION OF METALLIC MULTILAYERS AND SUPERLATTICES

As discussed earlier in this article, layered metallic structures with periodicities close to atomic dimensions have been produced from a variety of metals using sputtering, evaporation, and MBE techniques. Metallic multilayers can be produced with highly regular periodicity along the direction of the layering, but with varying degrees of perfection within the layers and at their interfaces. Multilayers and superlattices made with metallic components have degrees of structural perfection that can be roughly classified in three categories as described below, and as illustrated in Fig. 4. It is important to note that for many studies it is not necessary to have the highly ideal single-crystal epitaxy that can be achieved in certain semiconductor heterostructures. For this reason, preparation and study of metallic materials that have degrees of structural perfection falling into each class of this scheme is worthwhile.

A simple "multilayer" metallic film [Fig. 4(a)] is one that may have sharp interfaces between its two components and with very well controlled average spacing between layers. However, the structure within either or both layers may be polycrystalline or amorphous, rather than crystalline. Because of this, microscopic structural information is not transferred from a given layer to subsequent layers. Even though these "multilayers" have a relatively low degree of structural coherence, they exhibit interesting physical phenomena (e.g., superconductivity) and have properties that make them useful in applications such as x-ray optics and magneto-optic data storage. An example of such a multilayer is Mo/Si, which is of interest for x-ray optics in the wavelength range > 124 Å. Each Mo/Si interface in these multilayers reflects a small amount of incoming x rays. By proper choice and control of the multilayer period, these reflected x rays will coherently add to provide significant total reflectivity for this wavelength region (Haelbich *et al.*, 1979).

"Composition-modulated alloys" (CMAs) [Fig. 4(b)] have a single lattice for the entire sample, but one whose sites are occupied in a periodic manner by alternating elements. Typically, two metals that have the same crystal structure and nearly the same lattice constant are miscible in each other, and thus prone to interdiffusion when in contact. Thus, the boundaries between the two components in a CMA generally are diffuse and, depending upon how thin the layers are, the amplitude of composition modulation in the center of a layer can be in the range from 0 to 100%. Thus, even though a crystalline superlattice exists in these materials, the diffuse interfaces

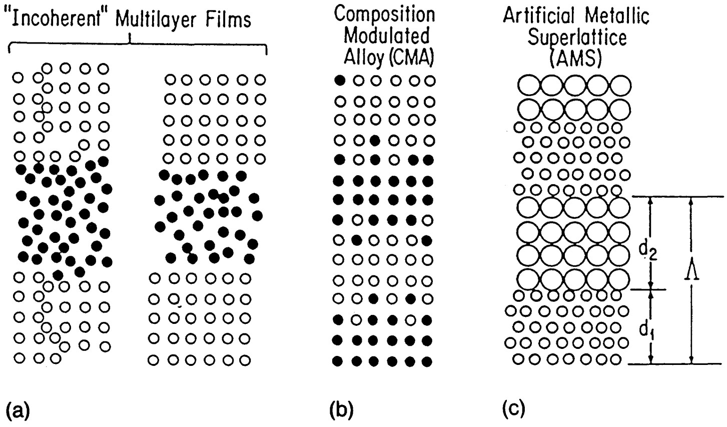
"Incoherent" Multilayer Films
Composition Modulated Alloy (CMA)
Artificial Metallic Superlattice (AMS)
d_2
Λ
d_1
(a)
(b)
(c)

FIG. 4. Diagram illustrating the degrees of structural coherence obtained in metallic multilayers and superlattices.

obstruct the possibility of observing many phenomena of interest. An example of a CMA is Cu/Ni, which exhibits interesting magnetic behavior.

Ideally, "metallic superlattices" [Fig. 4(c)] have atomically sharp boundaries between the two components as in the case of multilayered films. In addition, however, long-range structural coherence is maintained across many layers of the material, as in the case of a CMA. Thus, one has the roughly natural lattice spacing within the layer modulated by the imposed superlattice layer periodicity. Examples of superlattices include Mo/Ta (grown by sputtering) and Co/Pd (grown by MBE).

3. APPLICATIONS OF METALLIC MULTILAYERS AND SUPERLATTICES

3.1 Introduction

The growth of high-quality crystalline metallic superlattice samples is considerably more difficult than for semiconductors. This is largely due to the absence of the strong directional covalent bonding exhibited by semiconductors. However, in many ways metals are more forgiving of impurities and structural imperfections than are semiconductors. While defect- and impurity-free, highly perfect layering is required for semiconductors in order to exhibit phenomena of interest, metals can tolerate greater levels of structural and chemical imperfections and still have useful properties.

Although the nature of the metal atomic bond makes it more difficult to grow a "perfect" metal crystal than a perfect semiconductor crystal, fortunately it is usually less important. However, there are some physical properties that are sensitive to the fabrication of metallic superlattices. One example is superconductivity in superlattices containing Nb as one of the constituents. Oxygen reduces the superconducting transition temperature of niobium. Since Nb is a strong oxygen getter, ultrahigh vacuum or high deposition rates are required for the oxygen contamination to be negligible in the film. Another example where the physical properties can be affected by the fabrication conditions is in metal/nonmetal samples produced by sputtering, where Ar gas inclusion in the multilayers can be a problem. However, for the remainder of this article, we will assume the metallic multilayers have been fabricated in such a way that contamination or defects have negligible effect on the physical properties being studied.

3.2 Magnetism

An extremely interesting development in the field of magnetism in the past few years is the ability to study, and to some extent to control, the magnetic properties of surfaces and interfaces. Along with the experimental advances described earlier in techniques to prepare magnetic structures reliably by methods such as sputtering and MBE have been equally dramatic theoretical advances in the ability to calculate properties of realistic physical systems. Concurrently, remarkably sophisticated characterization techniques have been developed (e.g., spin-polarized scanning electron microscopy, intense synchrotron sources, pulsed sources for magnetic neutron scattering, scanning tunneling microscopy, etc.), which are now allowing sample characterization at a level not previously possible.

Single-crystal epitaxy of Fe on freshly cleaved NaCl substrates was reported over 50 years ago (Buck, 1936). As mentioned earlier, improvements in vacuum techniques continually broadened the range of materials being successfully studied until by the early 1980s developments in MBE deposition technology for semiconductor research began to be exploited for epitaxial growth of magnetic materials on a variety of substrates. Such work has lead to remarkable advances in this field over the past decade.

Techniques were developed to grow Fe epitaxially on (100)- and (110)-oriented GaAs substrates (Waldrop, 1979; Prinz and Krebs, 1980). More recently, epitaxial Fe films also have been grown on other semiconductor substrates, including ZnSe and lattice-matched $In_{0.2}Ga_{0.8}As$, using techniques based on commerical MBE equipment (Prinz *et al.*, 1986; Farrow *et al.*, 1988; Lottis *et al.*, 1988).

Although research on epitaxial magnetic films has been under way for over 50 years, particularly rapid experimental and theoretical advances have started to appear during the past decade. For example, MBE techniques have been successfully applied to produce magnetic metallic superlattices with high structural quality and very interesting physical properties. Single-crystal, epitaxial superlattices of Gd/Y (Kwo *et al.*, 1985) and Dy/Y (Salamon *et al.*, 1986) have been grown by MBE techniques, and their magnetic properties studied by neutron diffraction. Using these materials it was discovered that the coupling of successive magnetic layers could extend through 150 Å of the Y spacers. For example, epitaxial Dy/Y superlattices maintain the helical magnetic order of the Dy over several bilayers. The chirality of this helical order is conserved while the magnetic coherence length decreases with increasing Y thickness. For Gd/Y, either parallel or antiparallel alignment of the ferromagnetic Gd layers is found depending on the intervening Y-layer thickness. This oscillatory behavior is suggestive of the Ruderman–Kittel–Kasuya–Yoshida (RKKY) interaction. The RKKY model is a long-range indirect exchange interaction, mediated via the conduction electrons, that changes sign from ferromagnetic to antiferromagnetic with increasing radial distance between two interacting spins.

Another example of a discovery made with these materials is the "giant magnetoresistance" first observed in MBE-grown Fe/Cr multilayers (Baibich *et al.*, 1988), and now found in a number of other metallic multilayer systems (Parkin *et al.*, 1990; Mosca *et al.*, 1991). For many of these multilayers, the resistance can decrease by almost a factor of 2 in going from zero field to saturation. This effect has been attributed (Levy *et al.*, 1990) to spin-dependent transmission of the conduction electrons between the magnetic layers through the nonmagnetic layers. These superlattices order antiferromagnetically, so that alternate magnetic layers have opposite magnetization directions in zero field. The resistivity is at a maximum in this ordered state since the electrons interact with magnetic moments of both orientations and therefore have a higher probability of scattering. When a magnetic field of sufficient strength to align the magnetic moments is applied, one spin orientation is eliminated and the probability of scattering of the electrons is therefore reduced, concurrently reducing the resistivity. The large change in resistance observed with a small change in applied magnetic field makes these materials of great interest as possible magnetic field sensors and magnetic recording heads. This effect has attracted considerable attention, both because of the interesting physical properties, and also because of possible technological applications. Subsequent to the discovery of the giant magnetoresistance, it was found that in a wide variety of magnetic/nonmagnetic multilayer systems the exchange coupling between the magnetic layers oscillates between ferromagnetic and antiferromagnetic coupling (Parkin *et al.*, 1990, 1991). These oscillations depend strongly on the structure of the multilayers, and in most materials have an oscillation period of about 10 Å. That is, for each additional ≈10 Å thickness of the intervening nonmagnetic spacer layer, the coupling goes through one complete oscillation from ferromagnetic to antiferromagnetic. This is a fascinating discovery, and explanations of this phenomenon are leading to advances in the fundamental understanding of magnetic coupling (Wang *et al.*, 1990).

An important point to note is that to date there have been no electronic properties observed in metallic multilayers that are a consequence of collective superlattice behavior of the materials. This is to be contrasted to the case of semiconductor superlattices, where such collective effects have been observed for quite some time. As opposed to electronic effects, collective magnetic excitations due to the superlattice structure have been reported and understood theoretically. Thus, future work to study new combinations of metals using both MBE and sputtering may lead to observation of such electronic superlattice effects and additional collective magnetic effects.

3.3 Two-Dimensional Magnetic Behavior

The ability to deposit a perfect atomic monolayer of magnetic material has been a significant goal of surface magnetism research for some time, because of the possibility of testing theoretical predictions of intrinsic two-dimensional behavior in such an ideal film. Such predictions include the disappearance of long-range magnetic order above zero temperature, and substantially reduced values of the magnetization critical exponent as T_c is approached from below. In some ways, multilayers with very thin magnetic layers can be thought of as a collection of individual two-dimensional magnetic films. Thus, studying these predicted effects with multilayers rather than single films becomes possible. In addition, multilayers have several advantages over single, isolated films. The periodic multilayer structure lends itself to characterization by conventional x-ray diffraction techniques, whereas there is far too little material in a single monolayer to be studied with this technique. For the same reason, magnetic measurements on multilayers can be far easier than on single films.

An example of the usefulness of multilayers for the study of lower-dimensional magnetic behavior is the observation of finite-size effects in the spin-glass CuMn layered with Si (Kenning *et al.*, 1987). The Si layers in these multilayers decouple the thin CuMn layers both electrically and magnetically. In this way the spin-glass behavior of each CuMn layer can be amplified to within the sensitivity of conventional measurement techniques.

3.4 Magneto-optics (Perpendicular Magnetization)

The preferred axis of magnetization of a bulk magnetic material is determined by a competition between the anisotropies from the crystal structure, so called magnetocrystalline, and those from its shape, often referred to as magnetostatic demagnetization. The shape anisotropy in a thin film is very strong, and so it normally dominates and tends to constrain the magnetization to be in the plane of the film. However, a few ultrathin magnetic overlayers and superlattices with very thin magnetic layers were discovered that had their magnetization easy axis perpendicular to the plane of the film (Gradmann, 1974; Carcia *et al.*, 1985). Thus, other magnetic anisotropies are present in these materials that are sufficiently strong to overcome the shape and magnetocrystalline anisotropies. Studies of the thickness dependence of this effect have shown that it is the magnetic/nonmagnetic interface in these materials that gives rise to this anisotropy.

Although the precise mechanism is not yet well understood, calculations including the spin-orbit interaction of the magnetic anisotropy for the limiting case of unsupported ferromagnetic monolayers (Gay and Richter, 1986) and for certain overlayer-substrate combinations (Freeman and Wu, 1991) have been made. Magnetizations have been calculated that can be either perpendicular [e.g., Fe (001) on (001) Ag, Au, or Pd] or in-plane (free-standing Ni or Fe). However, it is not yet clear whether additional effects need to be included in the calculations of this effect.

Multilayers with perpendicular anisotropy are of considerable interest because of their potential technological usefulness as high-density data-storage media. It is possible to magnetize the film in domains of diameter $<0.5\ \mu m$ in order to store information, and then to read the stored data using the polar magneto-optical Kerr effect. This readout technique takes advantage of the fact that the plane of polarization of reflected polarized light is rotated by an amount approximately proportional to the perpendicular magnetization of the film. Using a focused laser beam to heat a small region of the film to above its Curie point while in the presence of a magnetic field, up or down magnetic domains can be written or erased. These domains can subsequently be read by a laser operated at a much lower power.

The advantages of magneto-optical data storage include a very high density, erasability, and no physical contact of the optical head with the medium. At present, the density appears to be limited by the spot size of the focused laser. However, fundamental limits on the size of the magnetic domains may exist due to their stability.

Co/Pt multilayers show considerable promise as future recording media due to their strong perpendicular anisotropy, resistance to oxidation and corrosion, and enhanced Kerr rotation at shorter wavelengths than those used in present media. The latter property is necessary for higher density storage, since the wavelength of the recording light ultimately limits the storage density by diffraction effects.

3.5 Superconductivity

Superconducting metallic multilayers are of interest for several reasons: as a route for synthesizing new superconducting materials; as a way to study certain superconducting properties in a controllable way; and for possible relevance for applications.

The characteristic length scale at which layering with a normal metal begins to modify significantly the physical properties of a superconductor is the temperature-dependent coherence length $\xi(T)$. In low-transition-temperature T_c superconducting materials, ≈ 50 Å $< \xi(0) < \approx 5000$ Å. Therefore, if superconductivity were the only physical phenomenon of interest in multilayer materials, layer perfection and interface sharpness on an atomic scale would not be required to observe interesting effects. For simplicity of discussion, we will implicitly assume in this section's discussion that the properties are independent of changes in the microscopic structure of the material, and result solely from the influence of spatial periodicity of the multilayer directly on the superconducting order parameter.

Layering can affect superconductivity in several ways. For this discussion, assume layering of a superconductor of thickness d_{super} with a nonsuperconductor of thickness d_{nonsuper}. For $\Lambda \gg \xi(0)$, where $\Lambda = d_{\text{super}} + d_{\text{nonsuper}}$, the individual layers retain their bulk properties. However, even in this regime of "thick" layers, spatially coherent planar defects are introduced by the layering, which can alter flux pinning by the superconductor. For $\Lambda \cong \xi(0)$ the superconducting order parameter becomes directly affected, and the superconductor will become anisotropic, resulting in the properties being altered significantly from those of the bulk. As a few examples, the critical magnetic fields (Banerjee *et al.*, 1983), critical currents (Raffy and Guyon, 1981), and even dimensionality (Ruggiero *et al.*, 1982) become dramatically different when $\Lambda \cong \xi(0)$. Through the proximity effect the microscopic properties of the superconducting layers also are affected (Banerjee *et al.*, 1982).

For multilayers with $\Lambda < \xi(0)$ superconductivity can be modified at the microscopic level. Control of the quality of the interfaces becomes crucial at this length scale. This is especially true if magnetism, with its much shorter length scale, is involved. Any size-effect modifications in the normal-state properties of the nonsuperconducting layers also may affect the microscopic properties of the superconductor. Examples of such effects are changes in the phonon structure due to interface phonons, density of states, cooperative electronic effects between layers, and epitaxial stress-induced changes. Epitaxial growth or stabilization of crystal phases with altered superconducting properties is also possible.

To date, effects of layering on the superconducting transition temperature, critical fields, and order parameter have received attention. These studies have shown that superconducting multilayer materials exhibit a variety of interesting, and possibly useful, effects.

Changes in the microscopic properties (e.g., Fermi-surface parameters or phonon spectrum) of a superconductor introduced by the multilayer structure also should affect its superconducting properties. Such changes could be caused by strains associated with epitaxial growth or through a high density of localized electron or phonon states at interfaces. However, because of the short Fermi wavelengths typical of metals, work to date has not resulted in materials of sufficient perfection for the additional periodicity to alter the electronic band structure enough to affect superconductivity.

One other possibility should be mentioned: that of stabilization of metastable metallurgical phases through the energy associated with the interfaces in these multilayers. This should lead to new superconducting properties in appropriate materials.

3.6 X-Ray Optics

In recent years there has been accelerating interest in optical elements that operate at soft-x-ray wavelengths ($\lambda_{\text{x ray}} = 10$–$500$ Å). In large part this interest has been driven by improvements in the technology for fabricating these elements, thus making possible several important applications, including soft-x-ray microscopy, astronomy, and microlithography. The development of x-ray microscopes that operate in the water window ($\lambda_{\text{x ray}} = 23$–$44$ Å) would be particularly significant since such microscopes might be able to image living cells in aqueous solutions. Another application, x-ray astronomy, already has provided information about high-energy processes in the universe. Finally, use of x-rays for microlithography should increase resolution over that possible with ultraviolet light, while having a higher throughput than with electron-beam writing techniques.

Conventional optical designs involving lenses and single-surface mirrors operating at near-normal incidence cease to function at soft–x-ray wavelengths. The reasons are twofold: the indexes of refraction of all materials approach unity, and absorption is high. These problems have resulted in solutions that fall into three categories of optics for soft x rays. Optical components are now available based on grazing-incidence reflection, Fresnel zone plates, and reflection from multilayer thin films. Each of these has certain applications for which it is particularly suited, although this article only will be concerned with the last.

There are several properties that make multilayer x-ray mirrors well suited for many applications. First, they are inherently wavelength selective. Also, they can be designed to function at a variety of angles including normal incidence. Finally, they can be made relatively large or small without serious limitations. Since the performance of multilayer mirrors is directly related to their structural perfection, understanding and controlling the structure and interface quality is central to improving their performance.

Just as for the other applications discussed in this article, a multilayer x-ray mirror is produced by alternately depositing layers of two materials. The thicknesses of the layers are designed so that the reflections from each interface add in phase to yield a high overall reflectivity. The best results are obtained by selecting two materials for which there is a large difference in the optical constants of the pair. The properties that make multilayer optics desirable for certain applications, high peak reflectivity and narrow wavelength pass band, result when large numbers of interfaces contribute to the overall reflectance of the mirror. Unfortunately, the nonzero absorption present in all materials at soft–x-ray wavelengths limits the depth of penetration of the x rays, and therefore limits the number of interfaces that contribute significantly to the overall reflectivity. Because of absorption, the reflectivity eventually saturates at some values less than 1, independent of how many total layers are added to the multilayer.

Because of structural imperfections and impurities, in real multilayers the peak reflectivity is always less than for an ideal structure. One type of imperfection is the variation in the layer thicknesses in the direction perpendicular to the layers due to variations in the deposition rates. Other imperfections include interfacial roughness, interdiffusion, and reaction between the materials that make up the multilayer. These imperfections become increasingly important as the layer thicknesses are decreased, because interfacial roughness or interdiffusion that blurs the interface over a thickness greater than $\approx 10\%$ of the wavelength of the light will significantly reduce the performance of the mirror. Since the average layer thickness in a multilayer mirror designed to reflect 40 Å light is no more than ten atomic spacings, it is critical to grow multilayer materials with well-controlled interfaces.

4. DISCUSSION

As illustrated by Fig. 1, there has been considerable work on metallic multilayers over the past several decades. In spite of this, the range of systems that can be successfully grown as metallic multilayers and superlattices is far from completely explored. Also, of those metal/metal systems that have been prepared to date, systematic investigation of the physical properties of many is lacking. Substrate temperature, deposition rate, base vacuum pressure, sputtering pressure, substrate material, etc., all can greatly influence the growth and quality of these materials. Consequently, given this wide range of parameter space, it is not surprising that only a few of the metal/metal systems have been thus far investigated with the required level of detail.

Various multilayers consisting of metal/amorphous semiconductors have been made successfully. Examples include Nb/Ge and W/Si. To date, layered structures of metals and crystalline semiconductors have not yet been produced, at least with the degree of purity required. However, such materials could be of technological importance. In this case, MBE machines specifically designed to incorporate electron-beam evaporation sources and high-temperature substrate platforms are likely to lead to important results.

Much of the work to date on metal/insulator multilayers has been motivated by the interest in efficient optical coatings for soft x rays. Evaporation and sputtering have been used to produce a number of these multilayer materials to date. As mentioned earlier, work has centered on alternating materials that have high electron density with those that have low electron density. Such combinations result in multilayers with the largest differences in optical constants, and hence the highest reflectivity at each interface. Amorphous C, B, and Si are commonly the low-electron-density materials of choice. Use of MBE techniques should result in crystalline Si layers with improved optical properties in certain regions of the soft-x-ray spectrum. An additional benefit of crystalline materials is that they should have the ability to withstand heat loads without recrystallization.

To date, no metallic multilayer or superlattice has been produced that has the same degree of perfection as their semiconductor counterparts. Factors thought to be important in determining structural quality of these materials are the binary phase diagram of the components being layered (including their mutual solubility and whether or not intermetallic compounds can form), interdiffusion rates, the existence of lattice matching and compatible crystalline symmetries, the availability of suitable substrates to promote epitaxial growth, ability to control substrate temperature over a wide range, and precise control over the deposition conditions. Understanding which factors allow the formation of high-quality metallic superlattices is an important research question at the present time.

5. FUTURE OF METALLIC SUPERLATTICES

From some of the applications discussed in this article, the future of metallic superlattices appears to be bright. Highly controllable layered structures in a wide variety of metallic systems have been achieved, and semiconductorlike crystalline properties have been approached in a few special cases. The precise structural nature of the interfaces (lattice matching, epitaxial strains, chemical composition, electronic structure, etc.) in metallic systems has not yet been entirely determined, and is the subject of active work in this field. The use of MBE growth techniques, which is becoming more frequent, with *in situ* determination of interfacial structure and physical properties, should lead to improvements in these materials.

Finally, it is important to emphasize that many of the interesting and important effects possible in metallic multilayers do not necessarily require the high degree of crystal perfection and chemical purity needed by their semiconductor counterparts.

6. SUMMARY

The study of the physical properties of metallic multilayers and superlattices has contributed to a number of significant experimental and theoretical developments over the past decade. Control of these structures at the atomic level recently has become possible, and now offers the opportunity to test a number of physical predictions using extremely well controlled materials.

GLOSSARY

Auger Electron Spectroscopy: Analysis of emitted "Auger" electrons at elemental-specific energies for chemical surface characterization.

Epitaxy: Continuation across the interface into deposited thin films of the underlying substrate crystalline order.

Low-Energy Electron Diffraction (LEED): Surface-sensitive structural probe using normal-incidence electrons with energies on the order of 100 eV.

Molecular-Beam Epitaxy (MBE): Ultrahigh-vacuum evaporation technique using single-crystal substrates and *in situ* surface structure characterization to obtain ordered crystalline thin-film growth.

Reflection High-Energy Electron Diffraction (RHEED): Surface-sensitive structural probe using grazing-incidence electrons with energies on the order of 10 keV.

Scanning Tunneling Microscope (STM): Microscopic surface topographic probe using electron tunneling current collected or emitted by a scanning tip.

Sputtering: Deposition technique that uses bombardment of a target material with an energetic gas plasma to create a vapor flux.

Superlattice: Multilayered structure in which crystalline coherence is maintained across several bilayers.

Ultrahigh Vacuum (UHV): Pressure regime $P \leqslant 10^{-9}$ Torr.

X-Ray Optics: The redirection and focussing of x rays through the interaction with reflecting and /or refracting media.

X-Ray Photoelectron Spectroscopy (XPS): Use of the photoelectric effect to determine the identities, concentrations, and bonding of atoms on a surface. This technique is sometimes called ESCA, for electron spectroscopy for chemical analysis.

List of Works Cited

Baibich, M. N., Broto, J. M., Fert, A., Nguyen Van Dau, F., Petroff, F., Eitene, P., Creuzet, G., Friederich, A. Chazelas, J. (1988), *Phys. Rev. Lett.* **61**, 2472–2475.

Banerjee, Indrajit, Yang, Q. S., Falco, Charles M., Schuller, Ivan K. (1982), *Solid State Commun.* **41**, 805–808.

Banerjee, I., Yang, Q. S., Falco, Charles M., Schuller, Ivan K. (1983), *Phys. Rev. B* **28**, 5037–5040.

Brück, L. (1936), *Ann. Phys.* **26**, 233–257.

Carcia, P. F., Suna, A., Onn, D. G., van Antwerp, R. (1985), *Superlatt. Microstruct.* **1**, 101–109.

DuMond, J. W. M., Youtz, J. P. (1935), *Phys. Rev.* **48**, 703 (1 page).

Falco, C. M., (1984) *J. Appl. Phys.* **56**, 1218–1219.

Farrow, R. F. C., Parkin, S. S. P., Speriosu, V. S., Wilts, C. H., Beyers, R. B., Pitner, P., Woodall, J. M., Wright, S. L., Kirchner, P. D., Pettin, G. D. (1988), in R. T. Tung, L. R. Dawson, R. L. Gunshor (Eds.), *Epitaxy of Semiconductor Layered Structures*, MRS Symposium Proceedings 102, Pittsburgh: Materials Research Society, p. 483.

Freeman, A. J., Wu, R. (1991), *J. Magn. Magn. Mater.* **100**, 497–514.

Gay, J. G., Richter, R. (1986), *Phys. Rev. Lett.* **56** 2728–2731.

Gradmann, U. (1974), *Appl. Phys.* **3**, 161–178.

Haelbich, R. P., Segmuller, A., Spiller, A. (1979), *Appl. Phys. Lett.* **34**, 184–186.

Kenning, G. G., Slaughter, J. M., Cowen, J. A. (1987), *Phys. Rev. Lett.* **59**, 2596–2599.

Kwo, J., Gyorgy, E. M., McWhan, D. B., Hong, M., DiSalvo, F. J., Vettier, C., Bower, J. E. (1985), *Phys. Rev. Lett.* **55**, 1402–1405.

Levy, P. M., Zhang, S., Fert, A. (1990), *Phys. Rev. Lett.* **65** 1643–1646.

Lottis, D. K., Florczak, J., Dahlberg, E. D., Batra, S., Wowchak, A. M., Cohen, P. I. (1988), *J. Appl. Phys.* **63**, 3662–3663.

Mosca, D. H., Petroff, F., Fert, A., Schroeder, P. A., Pratt, W. P., Jr., Laloee, R. (1991), *J. Magn. Magn. Mater.* **94**, L1–L5.

Parkin, S. S. P., More, N., Roche, K. P. (1990), *Phys. Rev. Lett.* **64**, 2304–2307.

Parkin, S. S. P., Bhadra, R., Roche, K. P. (1991), *Phys. Rev. Lett.* **66**, 2152–2155.

Prinz, G. A., Krebs, J. J. (1981), *Appl. Phys. Lett.* **39**, 397–399.

Prinz, G. A., Jonker, B. T., Krebs, J. J., Farrari, J. M., Kovanic, F. (1986), *Appl. Phys. Lett.* **48**, 1756–1758.

Raffy, H., Guyon, E. (1981), *Physica B* **108**, 947–948.

Ruggiero, S. T., Barbee, Jr., T. W., Beasley, M. R. (1982), *Phys. Rev. B* **26**, 4894–4908.

Salamon, M. B., Sinha, S., Rhyne, J. J., Cunningham, J. E., Erwin, R. W., Borchers, J., Flynn, C. P. (1986), *Phys. Rev. Lett.* **56**, 259–262.

Waldrop, J. R., Grant, R. W. (1979), *Appl. Phys. Lett.* **34**, 630–632.

Wang, Y., Levy, P. M., Fry, J. L. (1990) *Phys. Rev. Lett.* **65**, 2732–2735 .

Further Reading

Dhez, P., Weisbuch, C. (Eds.) (1987), *Physics, Fabrication, and Applications of Multilayered Structures*, NATO Advanced Study Institute Series B, Physics, No. 182, New York: Plenum.

Chang, L., Giessen, B. (Eds.) (1985), *Synthetic Modulated Structures*, New York: Academic.

Shinjo, T., Takada, T. (Eds.) (1987), *Metallic Superlattices: Artificially Structured Materials*, New York: Elsevier.

HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR

G. BASTARD, *Laboratoire de Physique de la Matière Condensée, Ecole Normale Supérieure, Paris, France*

	Introduction	477
1.	**Fabrication Techniques**	477
2.	**Bulk Band Structure**	478
3.	**Band Offset**	479
3.1	Type I and Type II Structures	480
3.2	Length Scales and Envelope Functions	481
4.	**Applications to Semiconductor Quantum Wells**	482
4.1	Size Quantization and Confinement Energy	482
4.2	Exciton Effects in Quantum-Well Absorption Spectra	485
4.3	Stark Effect in Quantum Wells	486
4.4	Modulation Doping of Semiconductor Heterostructures	487
5.	**Coupling Effect Between Quantum Wells**	488
5.1	Double Quantum Wells in an Electric Field	488
5.2	Semiconductor Superlattices in an Electric Field	489
6.	**Conclusion**	492
	Glossary	492
	Works Cited	493
	Further Reading	493

INTRODUCTION

In 1970, Esaki and Tsu predicted the emergence of a new field of semiconductor physics if the characteristic length scale of the semiconductor materials could be shrunk below the de Broglie wavelength of the charge carriers (Esaki and Tsu, 1970). Under such circumstances, the wavelike nature of the carriers would fully manifest itself and, for instance, a complete reorganization of the carriers' energy spectra due to the quantum size effects would show up. Even though quantum effects are already present in regular bulk semiconductors, most of the electronic phenomena are accurately described by semiclassical approaches. The implementation of a semiconductor physics based on quantum size effects poses a severe challenge to the fabrication techniques. One has to grow multilayers whose thicknesses can be controlled on the nanometer scale over macroscopic areas. The experimental challenges have been overcome, and one is now able to create artificial semiconductor materials or devices in which the characteristic length scale of the active regions is only a few nanometers. In this article, we briefly discuss some examples of the basic physical phenomena which are brought about by the quantum size effects.

1. FABRICATION TECHNIQUES

Crystalline semiconductors are the basic materials for modern microelectronics and optoelectronics. Conventional epitaxy techniques (e.g., liquid-phase epitaxy or the modified Bridgman method) are growth processes which occur near thermal equilibrium between the solid and liquid phases. They work extremely well for bulk single crystals (e.g., Si, GaAs). To be useful these single crystals have to be processed later on (oxidization, metallization, ...). This generally results in boundaries between the bulk of the active semiconductor component and its surroundings that are atomically ill defined, i.e., that occur on length scales considerably larger than the bulk lattice constant (0.566 nm in GaAs). In the case of heterostructures, i.e., the solids that result from the growth of two or

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

more semiconductors on top of each other, and that are needed, e.g., for field-effect transistors ($Si+SiO_2$) or double heterostructure lasers (e.g., $GaAs+Ga_{1-x}Al_xAs$), the interface boundaries may extend along the growth axis over several tens of nanometers and are not easily controlled. This may become troublesome when the dimensions of the components are shrunk (e.g., in order to increase their density).

In modern growth techniques such as molecular-beam epitaxy (*q.v.*) or metal-organic chemical vapor deposition [see CVD (CHEMICAL VAPOR DEPOSITION)] the growth rates are very slow (a few μm per hour), and the growth occurs far from thermal equilibrium. Schematically, in MBE the sample is grown atomic plane by atomic plane by sending fluxes of elemental elements (e.g., Ga or As) in a high-vacuum chamber onto a heated substrate ($T \sim 600\,°C$). Because of the slow growth rates the atomic species that have impinged on the surface have time to diffuse on this surface to find their equilibrium positions before the growth of another layer takes place. This layer-by-layer growth allows us (in ideal situations) to make atomically sharp interfaces. For instance, in a GaAs–AlAs multilayer grown along the [001] axis, the boundary between a GaAs and a AlAs layer is an As atomic plane. The nearest-neighbor planes are a Ga plane and an Al plane, respectively. Since, ideally, one can repeat the growth sequence at will, one can obtain with a nearly perfect accuracy any layer (or sequence of layers) of prescribed thicknesses. Thus, depending on the potential use, one can grow single *A*-*B* heterojunctions or double *BAB* heterojunctions, also termed quantum wells, or superlattices that consist of an infinite, periodic, sequence of *A* and *B* layers with thicknesses L_A and L_B, respectively. In the latter case, it is clear that one has superimposed an artificial periodicity d ($d=L_A+L_B$), which is controllable at will, from a few tenths to several tens of a nanometer, on the material hosts' periodicities. Notice that the layer thicknesses can be measured *in situ* (by the so-called reflection high-energy electron-diffraction oscillations technique). Further information on the growth and *in situ* characterization of semiconductor heterostructures can be found in the articles CVD (CHEMICAL VAPOR DEPOSITION) and MOLECULAR-BEAM EPITAXY.

From now on we shall assume that the growths are always ideal, i.e., that the interfaces are perfect. This is not always so, but in the case of lattice-matched host materials, like the pairs $GaAs$–$Ga_{1-x}Al_xAs$ or InP–$Ga_{0.47}In_{0.53}As$, this is a reasonable assumption. Note that the assumption of a perfect bidimensional growth (i.e., atomic plane after atomic plane) implies that in the layer plane the atomic arrangement and therefore the atomic potentials are identical to those found in the bulk, i.e., translational invariance is maintained in the layer plane.

2. BULK BAND STRUCTURE

The electronic structure of each host bulk material is characterized by a band structure, i.e., the set of allowed energy ranges where an electron can propagate in the crystal. These allowed energy bands are separated by forbidden gaps. This alternation of allowed and forbidden bands results from the translational invariance of a crystalline potential along the three basis vectors of the underlying Bravais lattice. The crystal periodicity gives rise to the Bloch theorem: the one-electron eigenstates are plane waves $\exp(i\mathbf{k}\cdot\mathbf{r})$ modulated by functions $u_{n\mathbf{k}}(\mathbf{r})$ that are periodic with the periodicity of the Bravais lattice. The corresponding eigenenergies $\epsilon_{n\mathbf{k}}$ are characterized by a discrete quantum number n, the band index, and a quasicontinuous real wave vector $\mathbf{k}$, which can be restricted to the first Brillouin zone of the reciprocal lattice. All these properties result from the translational invariance, and one should expect their modification when dealing with carrier propagation along the growth axis of the heterostructure where the translational invariance is broken. The new feature is the possible existence of allowed eigenstates in the forbidden gaps of one of the hosts: Having broken the periodicity of the crystalline potential there is no reason to discard the evanescent waves whose energies fall in the forbidden gaps. An evanescent wave is characterized by a wave vector $\mathbf{k}$ that is not purely real. In bulk materials, these evanescent waves have to be rejected as possible solutions (they are not normalizable in an infinite sample), but they have to be taken into account in heterostructures where the layers either have a finite thickness or are at best semi-infinite.

The band structure of many III–V materials in the vicinity of the fundamental band gap [i.e., the one separating the topmost occupied band (termed the valence band) and the lowest lying unoccupied band (termed the conduction band)] is relatively simple. The top of the valence band takes place at $k=0$ (Γ point) and is fourfold degenerate, while the bottom of the conduction band is either the zone center for heavy elements of the third column of the periodic chart (e.g., Ga) or along the [001] axis for the light elements (e.g., Al). Away from the zone center the valence band is split into a heavy-hole and a light-hole band. The Γ band gap shrinks with increasing heaviness of the cation (see, e.g., the sequence AlAs, GaAs, InAs where the Γ band gap decreases from 3.4 to 1.52 eV and then to 0.41 eV).

In a bulk semiconductor the carrier response to a perturbation $V_{\text{ext}}(\mathbf{r})$ that is slowly varying on the scale of the unit cell can be analyzed in a relatively simple way. For energy states that are close to an extremum $\mathbf{k}_0$ of the dispersion relations, $\epsilon_n(\mathbf{k})$ can be approximated by $\epsilon_n(\mathbf{k}_0)+\hbar^2/2\Sigma_{\alpha\beta}(k_\alpha-k_{0\alpha})(1/m_{\alpha\beta})(k_\beta-k_{0\beta})$ where $\alpha,\beta=x,y,z$. $m_{\alpha\beta}$ is the effective mass tensor, which is symmetrical: $m_{\alpha\beta}=m_{\beta\alpha}$. For the Γ conduction band $m_{\alpha\beta}=m_c\delta_{\alpha\beta}$. The carrier wave functions can be approximated by the product of their $\mathbf{k}_0$ Bloch functions modulated by a slowly varying envelope $F_n(\mathbf{r})$. The latter is the solution of a Schrödinger equation where the kinetic energy is $\frac{1}{2}\Sigma p_\alpha(1/m_{\alpha\beta})p_\beta$ instead of the usual $p^2/2m_0$ and where the potential energy consists only of the perturbing term. Thus, the relevant information on the carrier dynamics is embodied in the envelope function, and one may identify the motion of the electron in the perturbed crystal with that of a fictitious particle whose mass is not the bare (scalar) electron mass m_0 but the tensor $m_{\alpha\beta}$ and which moves under the effect of the perturbing potential. The crystalline potential disappears explicitly from the equations of motion, but its existence is manifested in the replacement of m_0 by $m_{\alpha\beta}$. For the filled valence band where $m_{\alpha\beta}<0$, the electronic motion is more easily analyzed if one defines as the fictitious particle the hole, which carries the positive effective mass $-m_{\alpha\beta}$ but a positive charge $+e$. Notice finally that an external disturbance that is attractive for electrons is repulsive for holes.

3. BAND OFFSET

One of the central questions of heterostructure electronic states is the magnitude and sign of the energy difference between the position of the tops of the valence bands of the various host materials. If the interfaces are sharp, we expect that plotting the position dependence of the valence-band edge ϵ_v versus the position z along the growth axis will result in a staircase variation: In the A layer, $\epsilon_v(z)$ coincides with $\epsilon_v^{(A)}$, the value in the bulk A layer, while in the B layer $\epsilon_v(z)$ coincides with $\epsilon_v^{(B)}$, the value in the bulk B layer. This is indeed what has been found for isovalent lattice-matched heterojunctions like GaAs/AlAs, but after considerable computations. The reason is that whatever the interface perfection there is no reason why the chemical bonds in the A and B materials should be exactly identical. This implies the formation of interface dipoles (e.g., at the Ga–As–Al interface the chemical bond Al–As is slightly more polar than the Ga–As one). These interface dipoles modify the electron energy levels and, in particular, shift the band edges. Until recently (van de Walle and Martin, 1987; Bylander and Kleinman, 1987), the unsolved question was to ascertain whether the dipole formation is a fairly localized perturbation around the interface or whether it extends over several lattice planes. For lattice-matched isovalent heterojunctions it appears that the assumption of piecewise constant band edges is nearly true, as shown in Fig. 1, where the valence charge distribution and electrostatic potential energy averaged over a unit cell are plotted versus the position along the [001] axis in a superlattice $(GaAs)_3(AlAs)_3$ (Baroni *et al.*, 1989). On the other hand, plots of the actual charge distributions (i.e., the nonaveraged ones) hardly give evidence of an interface.

Figure 1 helps to visualize a very important concept in heterostructure physics, which is that of wells and barriers. In the GaAs–AlAs superlattice of Fig. 1, the topmost valence electrons experience (in the GaAs layers) a higher cell-averaged electrostatic potential en-

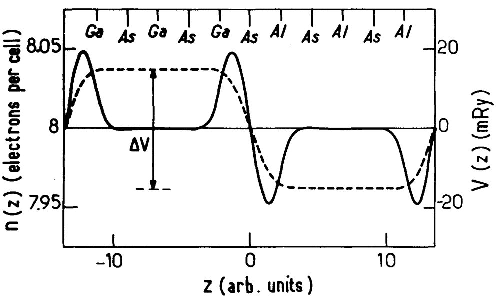

FIG. 1. Calculated cell-averaged charge distribution (solid line) and potential energy (dashed line) of dipolar origin along the [001] axis in a $(GaAs)_3$ $(AlAs)_3$ superlattice (adapted from Baroni *et al.*, 1989).

ergy $V(z)$ of dipolar origin than in the AlAs layers. Also $V(z)$ becomes constant a few tenths of a nanometer away from the interface. The energy difference ΔV of $V(z)$ on the two sides of the interface is not the only electrostatic contribution. One should take into account that GaAs and AlAs are different materials and add a term $\Delta E_v = E_v(\mathrm{GaAs}) - E_v(\mathrm{AlAs})$, which is the energy difference between the two bulk valence-band edges as measured from the (arbitrary) average of the electrostatic potential energy (see Fig. 2 for the general case of an A/B heterojunction). In the GaAs/AlAs pair, ΔV and ΔE_v have the same sign. The sum $\Delta E_v + \Delta V$ is the valence-band offset between GaAs and AlAs. In the GaAs/AlAs pair, it is such that a topmost valence electron with the lowest possible kinetic energy far from the GaAs–AlAs interface has a higher energy in the GaAs layer than in the AlAs one. Equivalently, the lower lying valence holes have a lower energy in the GaAs layer. For the holes, the GaAs layers act as potential wells and the AlAs layers act as potential barriers by analogy with the well known one-dimensional quantum mechanical problems. It is worth noticing that this analogy is facilitated because of our use of cell averages of the actual electrostatic potentials and of our neglect of interface details. We may therefore anticipate that this notion of wells and barriers will be fully operative for layers which are thick compared to the size of hosts' unit cells. In this case and for quantities that vary slowly on that scale, e.g., the cell-averaged carrier density, we shall be justified in neglecting the interface dipole effects. In the GaAs–AlAs case and more generally for lattice-matched isovalent multiple heterojunctions it appears that this is not a very stringent requirement. Indeed, models based on the assumption of flat band edges, i.e., piecewise constant functions $\epsilon_v(z)$ with abrupt changes at (infinitely thin) interfaces equal to $\Delta E_v + \Delta V$, work well, down to layer thicknesses of 1–1.5 nm.

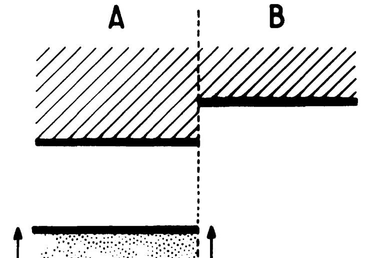

FIG. 2. Relationship between the valence-band offset (VBO), ΔV, and ΔE_v in a A/B heterojunction. In the specific case of the GaAs/AlAs pair, GaAs (AlAs) plays the part of the A (B) material (adapted from Baroni *et al.*, 1989).

3.1 Type I and Type II Structures

Once the valence-band offset is known, any other offset (lower conduction band, satellite conduction bands, and so on) is fixed since the various bulk band gaps are usually known very accurately. In particular, depending on the relative magnitudes of the valence-band offset and band-gap energy difference, one may classify the heterostructures into two main categories (see Fig. 3). In type I structures, e.g., $Ga_{0.47}In_{0.53}As/InP$, one kind of layer ($Ga_{0.47}In_{0.53}As$) confines both the electrons (in the lower conduction band) and the holes (in the upper valence band). In type II structures, e.g., $InP/Al_{0.48}In_{0.52}As$, the electrons and holes are spatially separated: The lowest-lying electron states are found in InP while the lowest-lying hole levels are in the $Al_{0.48}In_{0.52}As$ layers. The type I structures are

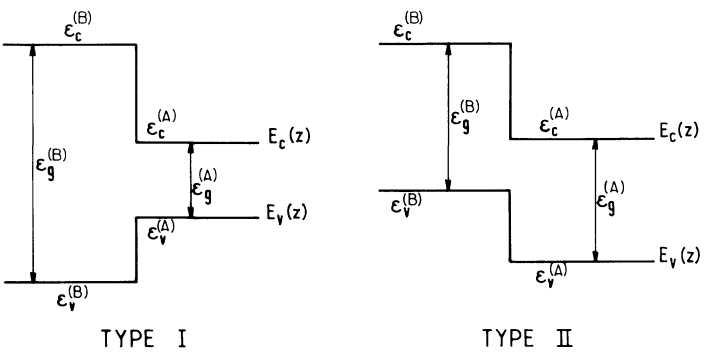

FIG. 3. Conduction- and valence-band edge profiles in type I and type II heterostructures.

by far the more numerous (and up to now the more useful). The pairs GaAs/$Ga_{1-x}Al_xAs$ ($x \leqslant 0.4$), $Ga_{0.47}In_{0.53}As$/InP, $Ga_{0.47}In_{0.52}As$/$Al_{0.48}In_{0.52}As$, and CdTe/$Cd_{1-x}Mn_xTe$ (at zero magnetic field) form type I structures. GaAs/AlAs, InP/$Al_{0.48}In_{0.52}As$, and InAs/GaSb belong to the type II family.

3.2 Length Scales and Envelope Functions

It appears from the previous discussions that different scales of lengths enter into heterostructure physics. The first relevant length is the natural host periodicity (a few tenths of a nanometer). The second is the range of interface dipole effects (also a few tenths of a nanometer in lattice-matched isovalent heterostructures). The third is the layer thicknesses. They can be varied from a few nanometers to hundreds of nanometers. Finally, because of the imperfections, or coupling to the phonons, that necessarily exist in heterostructures, the coherence lengths of the various quantum states come into play. We shall be mainly concerned with the heterostructures where the layer thicknesses are much larger than the natural host periodicity and the range of the interface dipole effects. This will enable us to assume infinitely sharp interfaces separating unperturbed layers (piecewise constant band edges). On the other hand, the layer thicknesses will be assumed smaller than the coherence length to make sure that if, for instance, we look upon the carrier motion in a quantum well, the quantum mechanical effects we expect from the interference of the carrier waves incident and reflected from the boundaries will be observable. Defects, if any, induce transitions between the eigenstates of perfect heterostructures. In particular, they lead to the existence of a finite electrical conductivity (Stern, 1993).

It happens that in many heterostructures the band offsets are a fraction of an eV, i.e., considerably smaller than the bulk conduction and valence bandwidths. This means that in the energy segment separating the same edge (i.e., with the same symmetry), both hosts' dispersion relations are relatively well described by a parabolic expansion. In addition the edge Bloch functions are known to be very similar among the III–V family. Then, for isovalent lattice-matched heterojunctions, it may be shown that the following simple "particle in a box" model works well (Altarelli, 1986). This model amounts to stating that in the heterostructure the complete wave function can be written as the product of the hosts' bulk Bloch function at the extremum by a slowly varying envelope function $F_l(\mathbf{r})$ solution of the effective Schrödinger equation:

$$\left\{-\frac{\hbar^2}{2}\sum_{\alpha\beta}\nabla_\alpha\frac{1}{m_{\alpha\beta}(z)}\nabla_\beta+\epsilon_l(z)+V_{\text{ext}}(\mathbf{r})\right\}F_l(\mathbf{r}) \\ =\epsilon F_l(\mathbf{r}). \qquad (1)$$

Equation (1) was first established by Ben Daniel and Duke (1966). The effective-mass tensor $m_{\alpha\beta}(z)$ is position dependent and actually is piecewise constant: $m_{\alpha\beta}(z)$ equals $m_{\alpha\beta}^{(A)}$ in the A layer and $m_{\alpha\beta}^{(B)}$ in the B layer. $\epsilon_l(z)$ is the position dependence of the energy extremum under consideration. It is also piecewise constant: $\epsilon_l(z)$ equals $\epsilon_l^{(A)}$ in the A layers and $\epsilon_l^{(B)}$ in the B layers. Finally, $V_{\text{ext}}(\mathbf{r})$ is any external perturbing potential, e.g., eFz if there is a constant electric field $\mathbf{F}$ applied along the growth axis of the heterostructure or $V_{\text{imp}}(\mathbf{r}-\mathbf{R}_i)$ if there is an impurity at $\mathbf{R}_i$.

4. APPLICATIONS TO SEMICONDUCTOR QUANTUM WELLS

4.1 Size Quantization and Confinement Energy

In quantum well structures, one realizes a basic quantum phenomenon: size quantization. The simpler quantum well, the rectangular one, corresponds to inserting a slab (thickness L) of an undoped well-acting material (A) between two semi-infinite undoped layers of barrier-acting materials (B). In Eq. (1), such a situation corresponds to $V_{\text{ext}}(\mathbf{r})=0$. Furthermore, let us for simplicity assume that the effective-mass tensor $m_{\alpha\beta}(z)$ is a scalar $m_c\delta_{\alpha\beta}$. The barrier height is the conduction band offset $\epsilon_c^{(B)}-\epsilon_c^{(A)}$.

We immediately notice that the in-plane variables (x,y) decouple from the z variable. Consequently, the envelope function $F(\mathbf{r})$ factorizes into a product of a z-dependent term $\chi(z)$ by an (x,y)-dependent one $\alpha(x,y)$. Since the potential energy is x,y independent, the in-plane motion is free and $\alpha(x,y)$ is the plane wave $S^{-1/2}\exp[i(k_x x+k_y y)]$, where S is the sample area and k_x,k_y the in-plane components of the carrier wave vector. A quasicontinuous parabolic dispersion $\hbar^2(k_x^2+k_y^2)/2m_c$ is associated with the in-plane free motion. This kinetic energy adds to each of the eigenvalues E_n (discrete or continuous) of the z-dependent problem. E_n is termed the edge of the nth two-dimensional sub-band describing the free in-plane motion, which occurs with increasing speed $\hbar k/m_c$ when $k[=(k_x^2+k_y^2)^{1/2}]$ increases.

Let us now discuss the solutions of the z-dependent problem. Analytical solutions exist. We refer the reader to Bastard (1988), where these solutions are given and discussed. For $\epsilon_c^{(A)}\leqslant\epsilon\leqslant\epsilon_c^{(B)}$ a classical particle would be bound in the A material and oscillate back and forth in the well at a constant velocity $[2(\epsilon-\epsilon_c^{(A)})/m_c]^{1/2}$. All energies would be allowed, i.e., one could find a possible classical motion for any ϵ. The quantum mechanical treatment brings two important new features. First, only discrete energies E_n are allowed (size quantization), and the lowest E_n, E_1, is always higher in energy than the bottom $\epsilon_c^{(A)}$ of the quantum well (confinement energy). Second, there is a finite penetration of the particle into the barrier (tunnel effect).

Trapping a quantum mechanical particle in a segment with length L means that the root mean square deviation of its position Δz ($\Delta z=[\langle z^2\rangle-\langle z\rangle^2]^{1/2}$) is of the order of L, where $\langle A\rangle$ denotes the quantum mechanical average of the operator A in a given eigenstate. For quantum particles, Δz and Δp_z cannot be simultaneously decreased to zero. As shown by Heisenberg the inequality $\Delta z\Delta p_z\geqslant\hbar/2$ holds. Knowing that $\Delta z\sim L$, we find $\Delta p_z\geqslant\hbar/2L$. Since in a bound state $\langle p_z\rangle$ vanishes we get $(1/2m_c)\langle p_z^2\rangle\geqslant\hbar^2/8m_cL^2$. Thus the carrier kinetic energy is never zero unless the quantum well is infinitely thick, which implies that the lowest eigenvalue E_1 is larger than $\epsilon_c^{(A)}$, the bottom of the well. This extra kinetic energy is often termed the confinement energy, for it arises from the squeezing of a quantum mechanical particle in a finite segment. We notice that it decreases (roughly) like L^{-2} with increasing L.

All the previous considerations apply as well to the valence band ($m_{\alpha\beta}=-m_v\delta_{\alpha\beta}$). In particular, the topmost valence eigenvalue is not $\epsilon_v^{(A)}$, the valence-band edge of the well-acting material, but is at lower energy ($-H_1+\epsilon_v^{(A)}\sim-\hbar^2/2m_vL^2$). Like the electrons, the holes display a confinement energy. Besides, the fourfold degeneracy of the bulk valence band edge is lifted. For a given L, the light holes have a larger confinement energy than the heavy holes. Therefore, the lowest-lying bound hole state is the heavy-hole state H_1 (in unstrained materials).

The confinement energy is very important with regard to the light-absorption (or -emission) edge of quantum wells versus the corresponding bulk ones. The absorption (emission) edge of a bulk semiconductor A is the lowest possible photon energy $\hbar\omega_g$ that can be absorbed (emitted) by the semiconductor. For a perfect material, at $T=0$ K, assuming that the band-to-band transition is optically allowed and neglecting the electron–hole interaction, one finds $\hbar\omega_g=\epsilon_c^{(A)}-\epsilon_v^{(A)}$. A quantum well of the same material clad between suitable barriers will display an absorption (emission) edge that is blue-shifted with respect to the bulk-A situation. This blue shift is equal to the sum of the confinement energies of the electron (E_1) and the hole (H_1). Therefore, it scales roughly like L^{-2} with the quantum well thickness. Thus it will be larger for the thinner well. In addition, by carefully selecting the thickness L, which is possible if the sample growth is well controlled, one may tailor the absorption or emission edges of the quantum well to a prescribed wavelength, a very desirable feature with regard to optoelectronics applications. We illustrate in Fig. 4 the effect of quantum well confinement energies in the case of $Ga_{0.47}In_{0.53}As$ where we show the L dependence of the shift E_1+H_1 of the absorption (emission) edge in the type I systems $Ga_{0.47}In_{0.53}As/InP$ and $Ga_{0.47}In_{0.53}As/Al_{0.48}In_{0.52}As$. In both cases $Ga_{0.47}In_{0.53}As$ is the well-acting material.

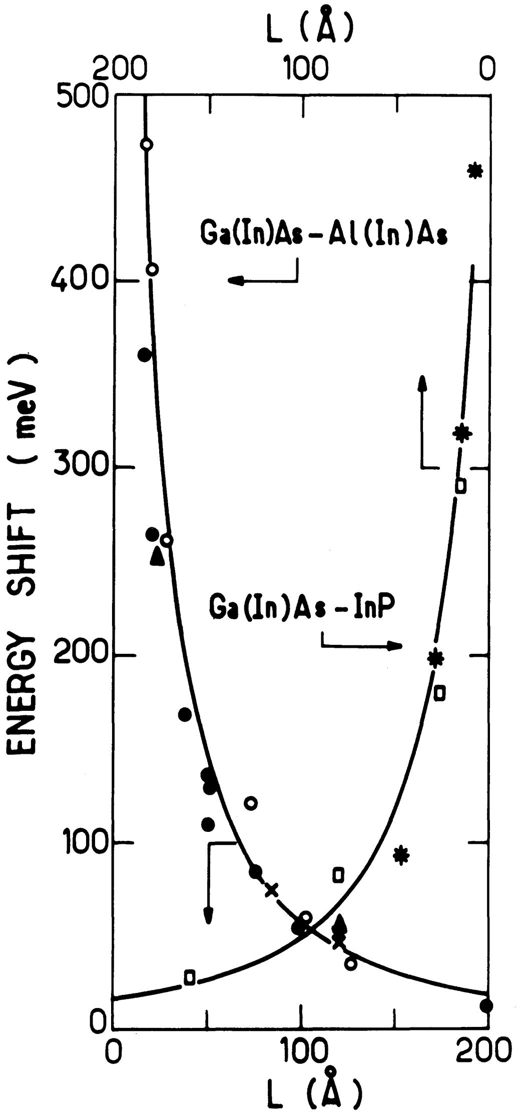

FIG. 4. The shifts of the absorption edge in $Ga_{0.47}In_{0.53}As/InP$ (right and upper scales) and $Ga_{0.47}In_{0.53}As/Al_{0.48}In_{0.52}As$ (left and lower scales) quantum wells, plotted versus the $Ga_{0.47}In_{0.53}As$ layer thickness. Solid lines: theory neglecting excitonic effects. Symbols: experimental data. From Bastard *et al.* (1991).

Why are the quantum mechanical allowed energies discrete for the bound states? This arises from the requirements of continuity that the solutions of a second-order differential equation must fulfill. In the effective Schrödinger equation for the z motion the potential-energy term and effective-mass tensor are piecewise constant functions. The integration of that equation around any point z leads to the continuity of $\chi(z)$ and $m_c^{-1}(z)\chi'(z)$. Notice that in the case of a constant $m_c(z)$, we recover the more familiar $\chi(z)$ and $\chi'(z)$ continuities. For a given energy ϵ, the solutions of Eq. (1) in each of the three layers can be written as a linear superposition of the free bulk solutions since in each layer both $m(z)$ and potential energy are constant. These solutions are evanescent waves in barriers, $\exp\{\pm z[2m_B/\hbar^2(\epsilon_c^{(B)}-\epsilon)]^{1/2}\}$, since $\epsilon \leqslant \epsilon_c^{(B)}$, and propagating waves in the well, $\exp\{\pm iz[2m_A/\hbar^2(\epsilon-\epsilon_c^{(A)})]^{1/2}\}$. We have therefore six unknown multiplicative coefficients. Two should be set equal to zero, for they multiply the exponentially growing solutions when $z\rightarrow\infty$ or $z\rightarrow-\infty$, which are not normalizable. We are therefore left with four multiplicative unknowns. Since four constraints have to be fulfilled (two per interface), we end up with a homogeneous system which can be satisfied only for discrete energies E_n.

It is easy to prove algebraically that a rectangular well admits at least one bound state and that in the large-L limit $E_n-\epsilon_c^{(A)}$ equals $\hbar^2n^2\pi^2/2m_cL^2$. The number of bound states that a rectangular well may accommodate is $1+\mathrm{Int}\{2m_cL^2(\epsilon_c^{(B)}-\epsilon_c^{(A)})/\hbar^2\pi^2\}^{1/2}$, where $\mathrm{Int}(x)$ denotes the integer part of x, i.e., this number is proportional to L in the large-L limit.

For eigenstates above the top of the well $(\epsilon>\epsilon_c^{(B)})$ the bulk solutions are propagating waves in the three layers. Thus no solution has to be discarded for normalizability reasons. Thus, we end up with four constraints versus six unknowns and each energy ϵ is allowed. The allowed spectrum is therefore continuous for energies above the top of the well. Each energy eigenvalue is twice degenerate corresponding to an electron either coming from $-\infty$, being partially reflected and transmitted

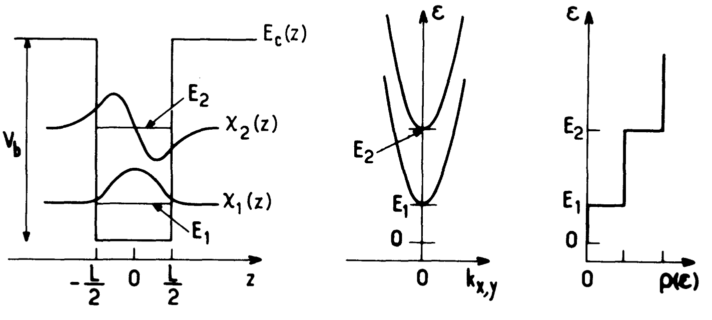

FIG. 5. A summary of quantum well electronic properties. Left panel: band-edge profile, and envelope functions $\chi_1(z)$, $\chi_2(z)$ of the first (E_1) and second (E_2) lowest lying bound states in the well. Middle panel: in-plane dispersion relations of the E_1 and E_2 sub-bands. Right panel: density of states. Adapted from Bastard (1988).

at the interfaces, and finally escaping to $+\infty$, or undergoing the reverse motion (Bohm, 1951).

The classical particle could never enter the barrier region if $\epsilon<\epsilon_c^{(B)}$, otherwise its kinetic energy would have to be negative, which is meaningless in classical mechanics. The quantum states do show a finite penetration of the carrier envelope functions in the barriers on account of the wavelike nature of the quantum particle. In this barrier, the waves are evanescent, i.e., they decay exponentially far away from the interfaces, but they are not zero. This is the well-known quantum mechanical tunnel effect. A summary of the quantum well properties is shown in Fig. 5.

An important quantity for evaluating a number of statistical properties is the density of states $\rho(\epsilon)$. By definition $\rho(\epsilon)d\epsilon$ is the number of allowed quantum states between ϵ and $\epsilon+d\epsilon$. One may easily work out that for a given sub-band n, $\rho_n(\epsilon)$ is a constant if $\epsilon>E_n$. Thus, for the sub-bands attached to the N bound states of the well one has

$$\rho(\epsilon)=\sum_{n=1}^{N}\rho_n(\epsilon)=\frac{m_cS}{\pi\hbar^2}\sum_{n}Y(\epsilon-E_n), \qquad (2)$$

where $Y(x)$ is the Heaviside (step) function and where a twofold spin degeneracy has been accounted for. The staircase aspect is noticeable and means that compared to a bulk situation where $\rho_{\text{bulk}}(\epsilon)\sim(\epsilon-\epsilon_c^{(A)})^{1/2}$ one expects sharper physical features when a sub-band becomes populated.

If we investigate the spatial localization of the ground bound state versus L, as, e.g., measured by the quantity Δz, we find that Δz first decreases with decreasing L (roughly like L) to reach a minimum value. Then, on further shrinking of the well, Δz increases again (see Fig. 6). The thickness L_{min} where Δz is minimum decreases with increasing barrier height, and for quantum wells clad between infinitely high barriers Δz is exactly proportional to L. The reason why Δz reaches a minimum versus L before increasing (to infinity) when L goes to zero is a consequence of the finite penetration in the barrier. We have seen that the confinement energy $E_1-\epsilon_c^{(A)}$ increases with decreasing L, roughly like L^{-2}. We expect on physical grounds that the ground state of a quantum well with vanishing thickness will coincide with the conduction-band edge of the barrier-acting material since in the

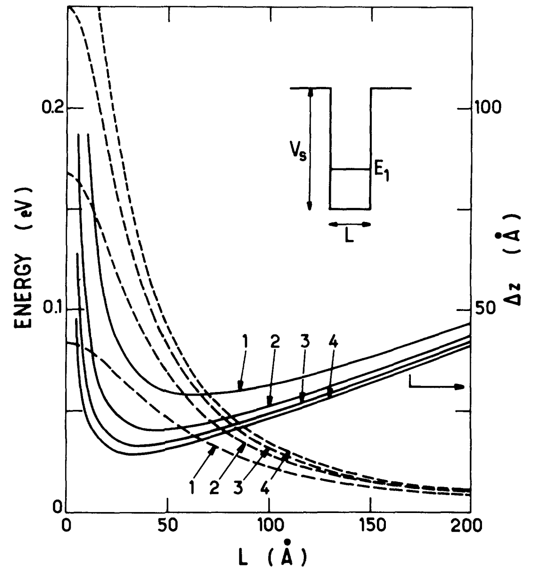

FIG. 6. Calculated confinement energy (left scale, dashed lines) of the E_1 state and root mean square deviation Δz (right scale, solid lines) of the electron position in the E_1 state versus the well thickness L. The curves labeled 1,2,3,4 respectively correspond to well depths V_b of 83.5, 167.1, 250.6, and 334.2 meV. The carrier effective mass is 0.07 m_0. From Bastard *et al.* (1991).

limit $L \to 0$ we deal with a bulk-like barrier. One can in fact show algebraically that $E_1(L)$ is a monotonically decreasing function of L from $\epsilon_c^{(B)}$ at $L=0$ to $\epsilon_c^{(A)}$ at infinite L. When L is large the barrier penetration is negligible and Δz, the uncertainty in the carrier position, is of the order of the quantum well thickness L (in fact $\Delta z \sim L$ at large L). In the narrow-well limit, instead, the spreading of the wave function takes place mostly in the barrier; with decreasing L [i.e., increasing $E_1(L)$] the evanescent wave vector in the barrier, which varies like $(\epsilon_c^{(B)} - E_1)^{1/2}$, decreases and the carrier wave function samples the barrier more and more, i.e., the particle becomes delocalized again.

4.2 Exciton Effects in Quantum Well Absorption Spectra

We have seen that the absorption or emission edge of a quantum well is blue shifted with respect to the corresponding bulk quantity. We neglected the electron–hole interaction in establishing this feature. Actually, the line shape of a quantum-well absorption edge does not display the predicted staircase behavior. Rather, it shows strong peaks at lower energies than the predicted plateaus, which persist even at room temperature (Fig. 7). Peaks are also seen in bulk materials at cryogenic temperatures. They arise from excitonic effects. The exciton is in fact a manifestation of many-body effects in solids (see COLLECTIVE PHENOMENA IN SOLIDS; EXCITONS). Let us for simplicity use the approximate picture of an exciton as the bound state formed between an electron and a hole. In a bulk semiconductor like GaAs the binding energy of such a pair is considerably smaller than one Rydberg (13.6 eV), which is the ionization energy of the ground (1*S*) state of atomic hydrogen in vacuum. This decrease comes from the lighter electron-hole reduced mass ($\sim 0.07 m_0$ in GaAs) compared with the bare-electron mass m_0 and from the weakening of the Coulombic electron–hole attraction which takes place not in the vacuum but in a dielectric medium of high permittivity ϵ_r (~ 12.4). The exciton binding energy is only ~ 5 meV in bulk GaAs, and the corresponding effective Bohr radius a^* is ~ 10 nm instead of 0.053 nm for hydrogen in vacuum. Thus the weakly bound excitons are easily dissociated, either by imperfections or random electric fields or by thermal excitation (kT is ~ 25 meV at room temperature).

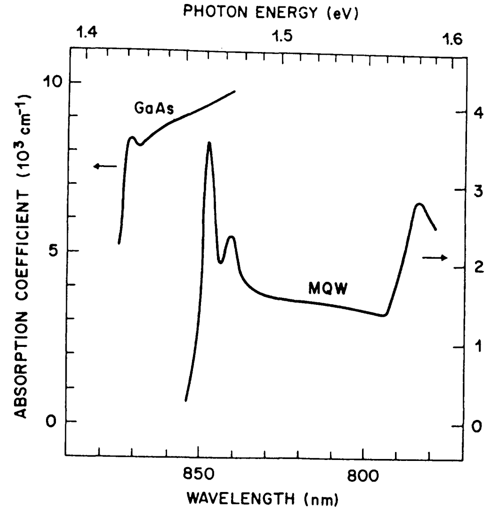

FIG. 7. Energy dependence of the room-temperature absorption coefficient in bulk GaAs and of a GaAs–Ga(Al)As multiple quantum well. From Chemla (1985).

Let us consider such an exciton in a quantum well. In the (ideal) purely bidimensional case, where the electron and hole would be constrained to move in the same plane, the exciton would be four times more bound than in three dimensions. The two-dimensional Bohr radius would be $\frac{1}{2}a^*$. This ideal situation is never achieved but we may anticipate that decreasing the well thickness will make the exciton quasibidimensional leading to an increase of its binding energy, if not by a factor of 4. The three-dimensional Coulombic potential $-e^2/4\pi\epsilon_0\epsilon_r r_{eh}$ mixes the free electron-hole pair eigenstates. The latter are labeled by a two-dimensional wave vector k of the reduced motion and by the sub-band indices n_e and n_h for the electron and hole, respectively (the in-plane wave vector K for the exciton center of mass is unaffected by the Coulombic term, which depends only on the relative electron-hole distance). We saw that by decreasing L we increase the sub-band separation. Thus by decreasing L the mixing between the ground sub-bands and the excited ones weakens, and if the sub-band spacing becomes larger than the bulk Rydberg this mixing becomes very weak. In this limit of strong quasibidimensionality the z motions of the electron and the hole become forced by the quantum-well potential, and what remains to be bound is the in-plane reduced motion of electron–hole pair.

This binding occurs not via the bulk Coulombic potential but by its average over the ground electron and hole probability densities along the growth axis. It is easy to show that this averaged Coulombic potential varies like ρ_{eh}^{-1} at large ρ_{eh}, the electron–hole distance in the layer plane. However, at small ρ_{eh} it varies more slowly than ρ_{eh}^{-1}, and more so when the sub-band separation decreases. Therefore, one does get an increase of the exciton binding energy when the quantum well thickness decreases, but it is less than that predicted by the purely bidimensional limit, as expected. Typical figures for the exciton binding energy and Bohr radius in GaAs quantum wells are 11 meV and 7 nm, respectively. This is apparently enough to produce very strong excitonic features in the absorption spectra of quantum wells, even at room temperature. Notice that the presence of strong absorption peaks with large oscillator strengths concentrated in a narrow energy range ($\leqslant$10 meV at room temperature) make the quantum well structures very attractive for optical nonlinear responses (Chemla, 1985).

4.3 Stark Effect in Quantum Wells

We saw that some tunability of the quantum-well band gap may be achieved through the well-thickness dependence of the electron and hole confinement energies. However, for a given quantum well, it is likely desirable (in view of optical applications) to control and tune its band gap by means of an external parameter. A convenient one is an electric field **F** is applied parallel to the growth axis. To produce such a field, either one evaporates a Schottky gate on the sample surface and biases it with respect to the doped substrate or one inserts (during the growth sequence) the quantum well in the intrinsic part of a *p-i-n* junction, which is reverse biased to achieved a large (several times 10^6 V/m) and homogeneous field.

The potential energy V_{ext} to be inserted in Eq. (1) is eFz_e $(-eFz_h)$, $e>0$, for an electron (a hole). Such potentials can be arbitrarily large and negative at large and negative z_e (positive z_h). This implies that the Schrödinger equation for the z motion admits no bound state. However, one feels that if the field is small enough, there should exist states that look more like quantum well bound states than do the other states of the continuum of allowed solutions. This feeling may be more quantitatively formulated if one uses the properties of the virtual (or metastable) bound states of the quantum well, i.e., states that, instead of being stationary, decay in time with a characteristic time constant τ_{ex}. The escape frequency τ_{ex}^{-1} is proportional to the transmission of the triangular barriers through which the electron or hole has to tunnel to escape to infinity (see Fig. 8). A long-lived virtual bound state is obtained if the barrier is very opaque. For the ground state E_1 of the electron in the tilted well (barrier height V_e) the criterion for low transparency is $eF \ll (V_e - E_1)^{3/2}(2m_c/\hbar^2)^{1/2}$. A similar criterion holds for holes with V_h, H_1, m_v replacing V_e, E_1, and m_c, respectively. Under usual circumstances

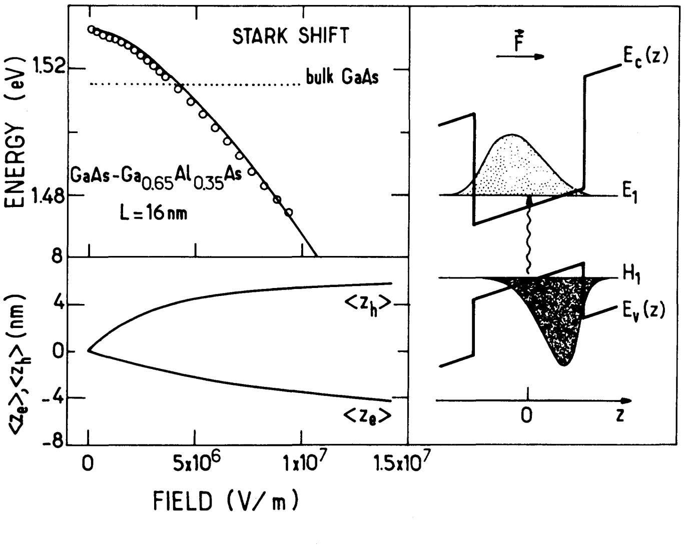

FIG. 8. Stark effect in quantum wells. Right panel: band-edge profile, and electron and hole envelope functions of the lowest lying quasistationary states (E_1 and H_1). Left panel, upper part: field dependence of the experimental (circles) and theoretical (solid line) lowest lying interband optical transition in a GaAs–Ga$_{0.65}$Al$_{0.35}$As quantum well (L $=16$ nm). Adapted from Vina *et al.* (1987). Left panel, lower part: calculated field dependence of the average positions of the electron ($\langle z_e \rangle$) and hole ($\langle z_h \rangle$) in the lowest lying electron (E_1) and hole (HH_1) state for the same well as in the upper panel.

($L\sim 15$ nm, $V_{e,h}\sim 0.2$ eV, $F\leqslant 6\times 10^6$ V/m) the criterion is very well fulfilled and the lifetime τ_{ex} is very long. For optical absorption experiments the finite lifetime effects, due to the field-induced ionization of the quantum well, are usually negligible.

The electric field not only induces the levels' metastability, it also shifts them, and this Stark effect is the useful feature for applications. At zero field the center of gravity of the electron and hole envelope probability densities is at the center of the quantum well by symmetry. If $F\neq 0$ the electron (hole) cloud will be shifted toward the left-hand side (right-hand side) interface of the well. In other words, the electric field induces for both particles an electric dipole, which, to the lowest order in F, is proportional to F. The induced dipole interacts with the field to give rise to energy shifts of either the electron or hole states that are quadratic in F (see Fig. 8). For the ground states, the induced dipoles are parallel to the field, and therefore, for both the electrons and holes, these quadratic Stark shifts are red shifts. In other words, the quantum-well band gap decreases with F. Using perturbation calculus, it is easy to show that the band-gap variation with F roughly scales like $F^2L^4(m_c+m_v)$ (Bastard *et al.*, 1983). This means that the quantum well thickness should not be too small, so as to allow sizable shifts. Representative figures are $L=16$ nm, $F=4\times 10^6$ V/m, leading to a Stark red shift of 16.2 meV in GaAs–$\text{Ga}_{0.7}\text{Al}_{0.3}\text{As}$.

The dipole formation has the disadvantage that it separates the electron and hole clouds, thereby weakening the exciton binding energy as well as the oscillator strength of optical transitions. Notice, however, that in the absence of quantum wells the electron and hole would be infinitely separated by the field. In quantum wells, instead, the exciton dissociation is inhibited by the presence of the quantum well barriers where the carriers accumulate in strong fields (but not so strong as to shorten significantly the carriers' lifetimes). Since L, roughly the maximum electron–hole separation, is comparable to the exciton Bohr radius, there remains enough of the exciton binding energy to ensure visible excitonic features in the room-temperature electroabsorption spectra of semiconductor quantum wells (Chemla, 1985). Since one can tune the excitonic absorption peak of a given quantum well with the field to make it match a prescribed wavelength, it has been possible to implement efficient electro-optical devices based on such a basic effect as the electric-field–induced Stark shifts of the quantum-well eigenstates (Miller *et al.*, 1985).

4.4 Modulation Doping of Semiconductor Heterostructures

When substitutional elements are introduced in a bulk semiconductor, the crystalline potential is no longer periodic. This gives rise most often to impurity states in the forbidden gaps of the pure material. For donor (acceptor) impurities that have one more (less) valence electron than the unit cell atom they replace, the impurity levels are energetically located below the conduction-band edge (above the valence band edge) of the pure material. In a simple hydrogenic model the binding energy ϵ_d of an electron bound to a donor is $13.6\text{ eV}\times[m_c/m_0\epsilon_r^2]$ in the $1S$ ground state. In GaAs, this corresponds to $\epsilon_d\sim 5$ meV for the bound electron ($\epsilon_a\sim 34$ meV for a bound hole). These impurities can be thermally ionized to supply the semiconductor conduction (valence) band with free electrons (holes), which ensure the electrical current flow. Since the impurities are located at random they scatter the free carriers (very effectively).

The characteristic feature of the shallow impurity states in bulk materials is that their binding energy or the scattering cross section of carriers by one such impurity does not depend on which unit cell hosts the impurity. This feature clearly arises from the translational invariance of the pure semiconductor. In heterostructures, this invariance remains as far as the in-plane location of the impurity is concerned. On the other hand, introducing impurities in the barrier-acting or in the well-acting material does lead to major differences both in the impurity binding energies and in the scattering cross section. Consider a single AB heterojunction and assume that only the barrier (B) is doped with donors. In the bulk B material such a donor would create shallow impurity levels below the conduction-band minimum. At $T=0$ K, the electron would be frozen on the lowest available state: the $1S$ bound state. There would be no electrical current because of a complete depopulation of the conduction band with free carriers. In the AB heterostructure, instead, the $1S$ barrier impurity state is no longer the ground electron state if the conduction-band offset exceeds the donor binding energy. There exists a whole electron continuum that corresponds to eigenfunctions mostly localized in the well-acting material. To ensure thermal equilibrium, the electrons have to be transferred to the well, leaving the ionized donors in the barrier. There is thus a spatial separation between the parent donors and the electrons. It gives rise to a band bending: The electrons are squeezed in a quasitriangular potential well formed on one side by the conduction-band offset and on the other side by the electrostatic potential energy (Ando, 1982). Bound states are formed in this well, and the electron motion becomes quasibidimensional. The scattering of charge carriers by the purposely set back impurities is much weakened compared to the bulk case (Dingle *et al.*, 1978). The low-temperature mobilities of such modulation-doped GaAs channels have reached unprecedented records: At liquid He temperature there is an improvement by more than two decades over the bulk GaAs figures. Notice, however, that at room temperature there is not much improvement obtained by using modulation-doped structures. This is because the dominant scattering mechanism at $T=300$ K is the scattering by the polar optical phonons in GaAs.

5. COUPLING EFFECTS BETWEEN QUANTUM WELLS

5.1 Double Quantum Wells in an Electric Field

Two identical rectangular quantum wells separated by a finite barrier are the simplest example of the quantum mechanical resonance effect. When isolated, each of the wells supports bound (at least one) and extended states. When brought into interaction—simply by the fact that the right-hand-side (rhs) well has a finite left-hand-side (lhs) barrier and vice versa, i.e., that the isolated well eigenstates are no longer eigenstates of the whole structure—the eigenstates delocalize over the two wells. That is to say, an electron in a stationary eigenstate of the double-well structure has an equal probability of being found in either well. If the two wells are different (e.g., because their thicknesses are not the same or because the bottoms of the two wells are not at the same energy) this perfect delocalization is broken. Some survives, however, if the difference remains small compared to the coupling.

To quantify these statements let us examine a simple one-level tight-binding analysis of the two wells. We mean that the two ground states of each well are well separated from their respective excited states, which allows us to approximate the ground state of the two-well system by a linear combination of the ground states $|\phi_l\rangle$, $|\phi_r\rangle$ of the lhs and rhs wells respectively:

$$|\psi\rangle=a_r|\phi_r\rangle+a_l\rangle|\phi_l\rangle, \tag{3}$$

$$(T+V_{l,r})|\phi_{l,r}\rangle=\epsilon_{l,r}|\phi_{l,r}\rangle. \tag{4}$$

The full Hamiltonian $T+V_r+V_l$ is left operating on $|\psi\rangle$. After projections on $\langle\phi_r|$ and $\langle\phi_l|$ one finds that Eq. (3) is an eigenstate of the double well with the eigenenergies

$$\epsilon_{\pm}=A\pm\sqrt{\delta^2+\lambda_{lr}\lambda_{rl}}\,, \tag{5}$$

where we have assumed $\langle\phi_l|\phi_r\rangle=\delta_{lr}$ for simplicity and

$$E_{l,r}=\epsilon_{l,r}+\langle\phi_{l,r}|V_{r,l}|\phi_{l,r}\rangle;$$

$$\lambda_{lr}=\langle\phi_l|V_l|\phi_r\rangle;$$

$$\lambda_{rl}=\langle\phi_r|V_r|\phi_l\rangle; \tag{6}$$

$$A=\tfrac{1}{2}(E_l+E_r);\quad \delta=\tfrac{1}{2}(E_r-E_l). \tag{7}$$

$E_{l,r}$ is the energy level of the isolated well corrected by a (downward) shift due to the presence of the other well. Also, A is the average of these two values and 2δ is their difference. Finally, λ_{lr} and λ_{rl} are transfer integrals (in the tight-binding terminology) that express the electron hoppings from one well to the other. Indeed, for identical wells ($\lambda_{lr}=\lambda_{rl}=\lambda$), if at $t=0$ the electron wave function is $|\phi_l\rangle$, the probability of finding it in the state $|\phi_r\rangle$ at time t is equal to $\sin^2(\lambda t/\hbar)$. Thus, $\pi\hbar/2|\lambda|$ is the jump time from one well to the other.

It is clear from Eq. (5) that the key dimensionless parameter of the problem is $(\lambda_{rl}\lambda_{lr})^{1/2}/\delta$. If this parameter is large the eigenstates are extended (over the double-well structure), while if it is small they are essentially localized in a given well. To use a chemistry analogy, the double-well bond is covalent or polar in these two limits.

Let us first discuss the case of symmetrical double wells where $\delta=0$, $\epsilon_l=\epsilon_r$, and $\lambda_{lr}=\lambda_{rl}=\lambda$. One finds immediately that $a_l^{\pm}=\pm a_r^{\pm}=2^{-1/2}$: The probability of finding the carrier in one well is evenly distributed between the two wells for both the ground (ϵ_-) and excited (ϵ_+) states. The twofold degeneracy of the infinitely separated wells is lifted by a quantity that is equal to twice the coupling term, i.e., the transfer integral. Like in the H_2^+ molecular ion, the ground state in the double-well problem corresponds to the symmetric combination of the isolated wells' eigenstates. This ensures the maximal gain of potential energy. The excited-state wave function instead has a node in the middle of the intermediate barrier.

It is remarkable that the electron delocalization persists even if the intermediate barrier is very thick, i.e., the coupling λ is very small (λ decreases exponentially with the barrier thickness). This shows that the coupling cannot be treated perturbatively in the $|\phi_l\rangle$, $|\phi_r\rangle$ basis. This is due to the energy degeneracy of the two interacting levels, a feature characteristic of resonances (cf. again the chemistry analogy with the π bonds in the benzenic ring).

In reality, it is seldom possible to achieve a strictly symmetrical double quantum well because of the imperfect control of layer thicknesses and barrier heights (many host materials are ternary random alloys). Thus one should investigate the effect of a nonzero δ on the carrier delocalization over the two wells. Besides, even if the wells are identical, it is always possible to create a controllable asymmetry by biasing the double well. This is actually realized by inserting the double well in the intrinsic part of a reverse-biased *p-i-n* junction. In this case 2δ is equal to $eF(L_w+L_b)$, where F is the magnitude of the constant electric field (**F** is parallel to the growth axis) and L_w (L_b) is the well (intermediate barrier) thickness.

One readily finds that an increase in δ leads to a convergence of ϵ_+, ϵ_- toward E_r and E_l respectively $[E_1\pm eF(L_w+L_b)/2]$ in the biased symmetrical double well. Concomitantly, the wave functions become more and more polarized on a given well: $a_l^+\to 0$, $a_r^+\to 1$ while $a_l^-\to 1$, $a_r^-\to 0$ in the large-δ limit. In the polar situation, the energy mismatch 2δ exceeds by far the coupling $(\lambda_{lr}\lambda_{rl})^{1/2}$ and the interwell coupling can be treated by perturbation calculus. This is because the unperturbed levels are nondegenerate ($\epsilon_r\neq\epsilon_l$) or off resonance.

We notice that an electric field induces a tunable transition from a resonant ($\delta=0$) to a nonresonant ($\delta\neq 0$) situation. Equivalently, one may state that the electric field transforms extended eigenstates ($\delta=0$) into localized ones ($\delta\neq 0$), a feature that is at first sight unusual and that leads to spectacular effects in superlattices (see below). Instead of misaligning degenerate levels one can start on the contrary with different ϵ_l and ϵ_r (i.e., an asymmetrical double well) and line them up with an appropriate bias. The field thereby induces a delocalization of the eigenstates over the two wells while, under flat-band conditions, they were heavily localized in one well or the other. The study of biased double wells has recently received considerable attention because they are the simplest fully quantum semiconductor structures where one can study the carriers (electron, hole) or exciton transfer from one well to the other (vertical transport).

5.2 Semiconductor Superlattices in an Electric Field

A superlattice is a periodic stacking of well-acting (thickness L_w) and barrier-acting (thickness L_b) identical layers (Esaki and Tsu, 1970). The band-edge profile in Eq. (1) becomes periodic with a tailorable periodicity d ($d=L_w+L_b$). This new periodicity implies the Bloch theorem for the envelope functions in Eq. (1):

$$F_l(x,y,z+nd)=e^{iqnd}F_l(x,y,z), \tag{8}$$

where q is the superlattice wave vector, which can be restricted to the first superlattice Brillouin zone $[-\pi/d,+\pi/d]$. It can be shown that the dispersion relations of the allowed energy bands in a one-dimensional superlattice are necessarily of the form

$$\cos qd=f(\epsilon), \tag{9}$$

where $f(\epsilon)$ is a function that depends on the band-edge profile inside a superlattice unit cell. For piecewise constant band edges (undoped layers, no external fields) one finds $f(\epsilon)$ explicitly:

$$f(\epsilon)=\cos k_w L_w \cos k_b L_b-\frac{1}{2}\left(\frac{k_w}{m_w}\frac{m_b}{k_b}+\frac{k_b}{m_b}\frac{m_w}{k_w}\right)\sin k_w L_w \sin k_b L_b\,, \qquad (10)$$

where m_w (m_b) is the electron effective mass in the well (barrier) and k_w, k_b are the corresponding carrier wave vectors at the energy ϵ.

In the absence of coupling between the wells ($L_b \to \infty$) each quantum well level E_1 is $(2N+1)$-fold degenerate where $(2N+1)d$ is the macroscopic superlattice thickness. This degeneracy is lifted at finite L_b (exactly like in the double-well situation), and one finds a miniband extending on each side of E_1 [the term miniband arises from the smallness of the superlattice bandwidths (~ 50 meV) and size of the Brillouin zone compared to those of bulk materials (a few eV)]. The one-level tight-binding analysis leads to the following dispersion relation and z-dependent envelope functions:

$$\epsilon(q)=E_1-2|\lambda|\cos qd, \qquad (11)$$

$$\chi_q(z)=(2N+1)^{-1/2}\sum_{-N}^{+N} e^{iqnd}\phi_1(z-nd), \qquad (12)$$

where E_1 is the isolated well eigenenergy corrected by the shift due to the other potential wells. $\phi_1(z-nd)$ is the associated envelope function for the well centered on the nth superlattice period and λ is the nearest-neighbor transfer integral. It is worth noticing that the lowest allowed eigenenergy ($q=0$) is red shifted by $2|\lambda|$ (i.e., half of the bandwidth) from the isolated well level E_1. This maximal gain of potential energy is realized with an envelope function $\chi_{q=0}$ that is nodeless, in close analogy with our findings in the double-well situation. The superlattice eigenstates are extended since the probability of finding the carrier in any of the $2N+1$ periods is $1/(2N+1)$, whatever q.

We restate that the delocalization occurs because we are considering the (resonant) coupling between exactly degenerate single-well levels. One may thus expect a partial or complete localization of the eigenstates if by any means the levels of consecutive wells become misaligned. This can occur in a number of ways. The most noticeable ones are the following:

1. A random fluctuation ΔE_1 of the isolated well levels that is too large in magnitude ($\Delta E_1 \geqslant 4|\lambda|$) leads to a complete localization of the carrier eigenstates (Anderson localization). Optical evidence of such a localization has been obtained by Chomette *et al.* (1986).
2. A constant energy shift eFd of one superlattice period with respect to the neighboring one, as achieved by biasing the superlattice, may lead if $eFd \geqslant 4|\lambda|$ to a complete localization. It is associated with a discrete (in fact evenly spaced) energy spectrum: the Wannier–Stark ladder.

These two situations are currently much investigated by means of electrical transport and optical experiments. The optical studies (absorption) are simpler since they probe directly the electronic density of states without any recourse to relaxation effects. The latter are of paramount importance in transport experiments. Let us restrict ourselves to a brief analysis of the consequence of the Wannier–Stark quantization in semiconductor superlattices for their optical properties.

In the presence of a longitudinal electric field and in the one-level tight-binding scheme outlined previously, one finds that the quasi-continuous band spectrum defined in Eq. (11) is destroyed and replaced by a set of evenly spaced levels:

$$\epsilon_\nu=E_1+\nu eFd, \quad \nu=0,\pm 1,\pm 2,\ldots, \qquad (13)$$

while in the envelope-function expansion of Eq. (12) the plane-wave term $(2N+1)^{-1/2} \times e^{iqnd}$ has to be changed into $J_{\nu-n}(-2|\lambda|/eFd)$, where $J_p(x)$ is the Bessel function of integer order p. These results were first established by Wannier in the late fifties (Wannier, 1959) and were much disputed on theoretical grounds. There were not clear cut experimental observations of these ladders in bulk solids essentially because $2|\lambda|/eFd \sim 10^2-10^3$ in bulk materials, which means that as far as localization effects are concerned, the bulk materials are always in the weak electric field regime. In the νth Wannier–Stark state the average position $\langle z\rangle$ of the carrier is equal to νd. The νth eigenstate extends over $\sim 2|\lambda|/eFd$ periods on each side of the nth period. Thus, with reasonable figures for bulk materials ($2|\lambda| \sim 1$ eV, $d \sim 0.5$ nm, $F \leqslant 10^7$ V/m) there is little hope to get a significant localization. In semiconductor superlattices instead

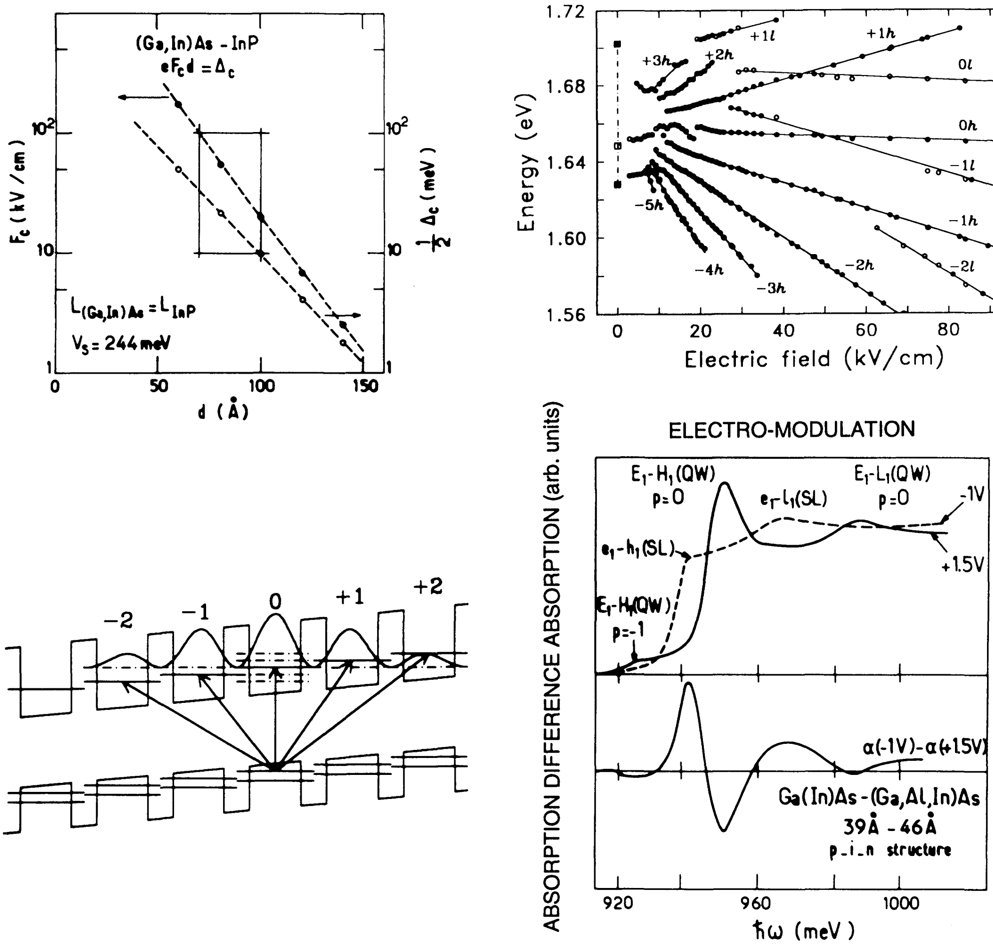

FIG. 9. Wannier–Stark quantization in superlattices. Right upper panel: fan chart of the observed interband optical transitions in a biased GaAs–$Ga_{0.65}Al_{0.35}As$ superlattice. $T=5$ K. The labels h and l correspond to heavy-hole–electron and light-hole–electron transitions, respectively (from Mendez and Agullo-Rueda, 1989). Right lower panel: absorption coefficient of a $Ga_{0.47}In_{0.35}As$–(Ga,Al,In)As superlattice for two different biases (adapted from Bleuse *et al.*, 1988b). The labels e_1h_1... identify superlattice transitions (zero electric field). The labels E_1-H_1, $p=0,-1$ correspond to vertical ($p=0$) and oblique ($p=-1$) transitions in (or between) nearly isolated quantum wells ($F\sim4\times10^6$ V/m). Left upper panel: half of the conduction sub-band width ($\frac{1}{2}\Delta_c$) and field strength (F_c) required to achieve a complete localization ($eF_cc=\delta_c$) plotted versus the superlattice period d in $Ga_{0.47}In_{0.53}As$/InP superlattices with equal layer thickness. Inside the rectangle a significant blue shift ($\frac{1}{2}\Delta_c>10$ meV) is achievable at a reasonable field ($F<10^7$ V/m). Left lower panel: band-edge profiles of a biased superlattice. A conduction envelope function centered at the 0th period is also shown. Several conduction Wannier–Stark states are displayed as well as vertical and oblique optical transitions from a Wannier–Stark valence state which is almost entirely localized in the 0th period ($\lambda_v \ll |\lambda_c|$).

$2|\lambda|\sim20$ meV, $d\sim5$ nm, and there is no difficulty to reach significant localization of the νth eigenstate around the νth period. A complete localization is achieved when the potential energy drop over a period eFd is equal to the zero-field bandwidth $4|\lambda|$.

At zero electric field, the oscillator strength of the optical transitions across the bandgap of the superlattice is nonzero only if the transition takes place vertically in the reciprocal space. This is due to the negligible wave vector carried by the photons. In addition the oscillator strength is proportional to the modulus squared of the overlap integral between the valence and conduction envelope functions. Let the absorbance $\alpha(\omega)$ be such that the transmitted (I_t) and incident light (I_0) intensities are linked together by the Lambert–Beer law: $I_t=I_0e^{-\alpha(\omega)}$, reflection losses being neglected. In the previous tight-binding scheme, one finds the frequency dependence of the superlattice absorbance $\alpha(\omega)$:

$$\alpha(\omega,F=0) = [(2N+1)/\pi]\alpha^{\mathrm{QW}}\arccos\{(-\hbar\omega + \epsilon_g^{\mathrm{QW}})/ 2(|\lambda_c|+\lambda_v)\}, \qquad (14)$$

where α^{QW} is the isolated well absorbance due to E_1-H_1 transitions ($\alpha^{\mathrm{QW}}\sim 0.3\%$ in usual III–V or II–VI materials). ϵ_g^{QW} is the band gap of the isolated quantum well (i.e., $\epsilon_g+E_1+H_1$ where ϵ_g is the band gap of the bulk well-acting material and E_1,H_1 the respective electron and hole confinement energies). $\alpha(\omega,F=0)$ merely reflects the joint density of states between the H_1 sub-band (width: $4\lambda_v$) and the E_1 sub-band (width: $4|\lambda_c|$). The absorption edge in the superlattice is $\epsilon_g^{\mathrm{QW}}-2(|\lambda_c|+\lambda_v)$, i.e., is red shifted by half the sum of the bandwidths with respect to the isolated quantum well edge. For $\hbar\omega>\epsilon_g^{\mathrm{QW}}+2(|\lambda_c+\lambda_v|)$ the absorbance is constant $[=(2N+1)\alpha^{\mathrm{QW}}]$. At nonzero F, one finds similarly

$$\alpha(\omega,F)=(2N+1)\alpha^{\mathrm{QW}} \times \sum_{\nu=-\infty}^{+\infty} J_\nu^2\left(\frac{2(|\lambda_c|+\lambda_v)}{eFd}\right) \times Y[\hbar\omega-\epsilon_g^{\mathrm{QW}}-\nu\, eFd]. \qquad (15)$$

From Eq. (15), it is clear that formally there is no more absorption edge in the biased superlattice. This is clearly due to the breakdown of the translational invariance of the superlattice potential imposed by the nonvanishing field. A staircase-shaped edge, linearly varying with F, is associated with any spatially oblique transitions ($\nu=\pm 1,\pm 2,\ldots$) where electron–hole pairs with a mean electron–hole separation *nd* are photogenerated (see Fig. 9). Since excitonic effects, not included in Eq. (15), sharpen these quasibidimensional edges, one expects to observe a series of evenly spaced peaks in electroabsorption experiments performed at fixed F. Conversely, oscillations periodic in F^{-1} should show up in experiments done at fixed ω and varying F. All these predictions were clearly confirmed in the last few years (Mendez *et al.*, 1988; Voisin *et al.*, 1988; Agullo-Rueda *et al.*, 1989).

A striking feature takes place in the strong-field regime where electron and hole envelope functions become electroshrunk in a single superlattice period. In that limit all the oblique transitions fade away [$J_\nu^2(x)\sim(x/2)^{2\nu}$ when $x\rightarrow 0$] to the benefit of the vertical transition [$\nu=0$ in Eq. (15)], and in effect the biased superlattice behaves like a collection of $2N+1$ isolated quantum wells. This effect is further strengthened by excitonic effects that are weak in superlattices at $F=0$ (three-dimensional materials) but are strong in isolated quantum wells (quasibidimensional materials). Once all the oblique transitions have collapsed, the absorption edge of the biased superlattice is effectively ϵ_g^{QW}, i.e., it is blue shifted with respect to the $F=0$ edge (Fig. 9). This blue shift can rightfully be termed a macroscopic quantum effect in that it arises from a purely quantum mechanical effect: the destruction by the electric field of the tunnel coupling between resonant (i.e., energy aligned) quantum levels. This spectacular effect has also been observed (Bleuse *et al.*, 1988b) and is currently used in the design of low-power electro-optical switches operating at room temperature (Bigan *et al.*, 1990).

6. CONCLUSION

This short article has, it is hoped, introduced the reader to a domain of semiconductor research, the heterostructures, that is explosively growing. The guiding line of the physics in these new materials is the preeminent part played by quantum mechanical effects (size quantization, tunnel effect, field-induced localization, etc.). It is expected that many more experimental breakthroughs and device implementations will result from semiconductor heterolayers.

GLOSSARY

Quantum Well: A double heterostructure that consists of a slab of a well-acting material inserted between two semi-infinite slabs of barrier-acting materials. Note that the notion of well and barrier is relative to a given band edge.

Resonance: A situation widely encountered in many branches of physics. It corresponds to the (fortuitous or induced) lining up of two characteristic frequencies of the system under consideration.

Stark Effect: The perturbation of electron states by an electric field.

Superlattice: An infinite, periodic sequence of slabs of two (or more) materials. Doping superlattices (i.e., alternate and periodic *n* and *p* doping of a given semiconductor) have also been studied.

Tunnel Effect: A quantum mechanical (wavelike) phenomenon that takes place when the carrier eigenenergy is locally smaller than the potential energy. The carrier wave function in these regions of space is exponentially varying with distance.

Tunnel Coupling (between Wells): The modification of the eigenenergy of a carrier moving in a multiple quantum-well structure over the one found in the isolated wells. It is due to the tunnel effect in the barriers separating the wells.

Valence-Band Offset: The energy difference between the topmost valence electron states that are localized on each side of (and far enough from) an interface between two semiconductors.

List of Works Cited

Agullo-Rueda, F., Mendez, E. E., Hong, J. M. (1989), *Phys. Rev. B* **40**, 1357–1360.

Altarelli, M. (1986), in: G. Allan, G. Bastard, N. Boccara, M. Lannoo, M. Voos (Eds.), *Heterojunctions and Semiconductor Superlattices*, Heidelberg: Springer-Verlag, pp. 12–37.

Ando, T. (1982), *J. Phys. Soc. Jpn.* **51**, 3893.

Baroni, S., Resta, R., Baldereschi, A., Peressi, M. (1989), in: G. Fasol, A. Fasolino, P. Lugli (Eds.), *Spectroscopy of Semiconductor Microstructures*, New York: Plenum, pp. 251–272.

Bastard G. (1988), *Wave Mechanics Applied to Semiconductor Heterostructures*, Les Ulis: Les Editions de Physique.

Bastard, G., Mendez, E. E., Chang, L. L., Esaki, L. (1983), *Phys. Rev. B* **28**, 3241–3245.

Bastard, G., Brum, J. A., Ferreira, R. (1991), in: H. Ehrenreich, D. Turnbull (Eds.), *Solid State Physics*, Vol. 44, New York: Academic, pp. 229–415.

Ben Daniel, D. J., Duke, C. B. (1966), *Phys. Rev.* **152**, 683–692.

Bigan, E., Allovon, M., Carre, M., Voisin, P. (1990), *Appl. Phys. Lett.* **57**, 327–329.

Bleuse, J., Bastard, G., Voisin, P. (1988a), *Phys. Rev. Lett.* **60**, 220–223.

Bleuse, J., Voisin, P., Allovon, M., Quillec, M. (1988b), *Appl. Phys. Lett.* **53**, 2632–2635.

Bohm, D. (1951), *Quantum Theory*, New York: Prentice-Hall.

Bylander, D. M., Kleinman L. (1987), *Phys. Rev. B* **36**, 3239–3236.

Chemla, D. S. (1985), in: M. H. Pilkuhn (Ed.), *High Excitation and Short Pulse Phenomena*, Amsterdam: North Holland, pp. 502–519.

Chomette, A., Deveaud, B., Regreny, A., Bastard, G. (1986), *Phys. Rev. Lett.* **57**, 1464–1467.

Dingle, R., Störmer, H. L., Gossard, A. C., Wiegmann, W. (1978), *Appl. Phys. Lett.* **33**, 665–667.

Esaki, L., Tsu, R. (1970), *IBM J. Res. Develop.* **14**, 61–68.

Mendez, E. E., Agullo-Rueda, F. (1989), *J. Lumin.* **44**, 233–231.

Mendez, E. E., Agullo-Rueda, F., Hong, J. M. (1988), *Phys. Rev. Lett.* **60**, 2426–2429.

Miller, D. A. B., Chemla, D. S., Damen, T. C., Wood, T. H., Burrus, C. A., Gossard, A. C., Wiegmann, W. (1985), *IEEE Trans. Quantum Electron.* **21**, 1462–1476.

van de Walle, C., Martin, R. M. (1987), *Phys. Rev. B* **35**, 8154–8165.

Vina, L., Mendez, E. E., Wang, W. I., Chang, L. L., Esaki, L. (1987), *J. Phys. C: Solid State Phys.* **20**, 2803–2815.

Voisin, P., Bleuse, J., Bouche, C., Gaillard, S., Allibert, C., Regreny, A. (1988), *Phys. Rev. Lett.* **61**, 1639–1642.

Wannier, G. H. (1959), *Elements of Solid State Theory*, Cambridge, U.K.: Cambridge Univ. Press.

Further Reading

Ando, T., Fowler, A. B., Stern, F. (1982), *Rev. Mod. Phys.* **54**, 437–672. A very complete survey of quasibidimensional systems, especially of Si inversion layers.

Bastard, G. (1988), *Wave Mechanics Applied to Semiconductor Heterostructures*, Les Ulis: Les Editions de Physique. An introductory textbook to the electronic properties of semiconductor quantum wells and superlattices.

Capasso, F., Margaritondo, G. (1987), *Heterojunction Band Discontinuities. Physics and Device Applications*, Amsterdam: North Holland. A comprehensive survey of the theory and measurements of band discontinuities.

Smith, D. L., Mailhiot, C. (1990), *Rev. Mod. Phys.* **62**, 173–234. A recent survey of the theory of superlattice structures, and in particular of strain effects.

Stern, F. (1993), *Electron States in Zero-, One- and Two-Dimensional Structures*, in: G. L. Trigg (Ed.), *Encyclopedia of Applied Physics*, Vol. 5, New York: VCH. An article more focused on transport properties of inversion layers, quantum wires, and quantum boxes.

HIGH-PRESSURE TECHNIQUES

WILLIAM B. DANIELS, *Department of Physics and Astronomy, University of Delaware, Newark, Delaware, U.S.A.*

	Introduction	495
1.	**Studies Using Static Techniques**	496
1.1	Principles	496
1.1.1	Strength of Pressure Chambers	497
1.2	Materials of Construction	497
1.3	Systems Using Pumped Fluids as the Pressure Medium	498
1.3.1	Seals	498
1.3.2	Tubing	499
1.3.3	Precompression	500
1.3.4	Intensifiers	500
1.3.5	Valves	500
1.3.6	Pressure Measurement	500
1.3.7	Dependent-Variable Techniques in Fluid Systems	501
1.3.7.1	Electrical Leads	501
1.3.7.2	Optical Windows	501
1.3.7.3	X-Ray Diffraction	501
1.3.7.4	Neutron Diffraction	501
1.3.7.5	Pumped Systems at Cryogenic Temperatures	502
1.4	Dry Systems—in Which Solids are Generally Used as Pressure-Transmitting Media	502
1.4.1	Piston-Cylinder Devices	502
1.4.2	Anvil Systems	503
1.4.2.1	Compressible-Gasket Anvils	503
1.4.2.2	Belt-Type Apparatus	503
1.4.2.3	Multiple-Anvil Designs	503
1.5	Diamond-Anvil Cells (DACs)	504
1.5.1	Pressure Measurement in Diamond-Anvil Cells	505
1.5.2	X-Ray Studies	506
1.5.3	Optical Studies	506
1.5.4	Electrical Transport	506
1.5.4.1	Miscellaneous Techniques—Magnetic, Mössbauer, Nuclear Magnetic Resonance	506
2.	**Dynamic Techniques—Shock Waves**	507
	Glossary	507
	Works Cited	508
	Further Reading	509

INTRODUCTION

This article presents techniques applicable to the study of a wide range of physical phenomena at pressures greater than 0.1 GPa (see Glossary for pressure unit conversions). It is characteristic of the field that many of the investigations involve small-scale studies in which details have been ingeniously worked out by individual investigators to make possible measurements over appropriate ranges not only of pressure, but of temperature, magnetic field, and other variables in a wide variety of materials. It is currently possible throughout the 1-GPa range to apply most of the techniques for investigation of condensed matter used at ambient pressure—for example, electrical resistivity, optical absorption and reflectivity, and Raman scattering. Techniques examined in this article shall not emphasize a lower limit to the term high pressure because some investigations at relatively low pressures require interesting methods and generate interesting results.

The boundaries keep shifting. Capability and sophistication are growing within all ranges, and the maximum attainable pressures have increased remarkably in recent years, as has the ability to carry out sophisticated experiments at those pressures. We shall describe some of the generally applicable high-pressure techniques, including pressure generation and measurement, and communication of the relevant dependent variables—e.g., electrical resistance—from samples within high-pressure cells.

High-pressure phenomena *per se* are not simply laboratory phenomena. Most of the condensed matter in the universe is at substantially elevated pressures. The observation

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

that the gravitationally generated pressures at the center of Earth and of Jupiter are about 364 GPa and 10^4 GPa, respectively, should be convincing evidence that the study of high-pressure phenomena is driven in part by the fields of geophysics and planetary physics.

There is a problem of determining what *is* "high" pressure in its interaction with disparate objects such as biological systems, atoms, and neutron stars. It is made easier by recognizing the identity of force per unit area and energy per unit volume (see Glossary). Thus it is possible to use dimensional analysis to make very crude, but sometimes illuminating, order-of-magnitude estimates of pressures required to produce various results by identifying characteristic energies and volume changes. For instance, pressures required to cause significant changes in chemical binding might be expected to be (chemical binding energies) ÷ (changes in volume in chemical reaction). These would be

$$\frac{\text{kilojoules/mole}}{\text{cm}^3\text{/mole}} \sim 1 \text{ GPa}. \qquad (1)$$

Or, the pressure at which pressure disruption of electronic "shell" structure of atoms into white dwarf stars would begin could be expected to be of the order of some characteristic electronic energy, e.g., 1 Ry, divided by the volume of an atom, e.g., $(4\pi/3)\times(\text{Bohr radius})^3$, i.e.,

$$P^* \simeq \frac{1 \text{ Ry}}{(4\pi/3)a_0^3} = 3511 \text{ GPa}. \qquad (2)$$

Going even further, the pressure to break down nuclei into neutron stars, i.e., ~(nuclear binding energies) ÷ (nuclear volumes), could be

$$\sim \frac{\text{millions of electron volts}}{(3 \text{ fermi})^3} \sim 10^{22} \text{ GPa}. \qquad (3)$$

These estimates are of course crude, but they do indicate the severe limitation of contemporary experimental maximum pressure techniques when examined on the scale of astrophysical phenomena, even though recent work has resulted in a reported static pressure of 416 GPa, greater than that at the center of the Earth (Ruoff *et al.*, 1990). Present (1992) upper limits of dynamic high-pressure experiments are even higher, at ~700 GPa for light gas guns (Holmes *et al.*, 1989) and pressures near 10 TPa using laser-driven shock waves (Cottet *et al.*, 1986). At the other end of the high-pressure scale, significant effects on living cells occur at pressures below 0.2 GPa (Zimmerman, 1970).

High-pressure techniques also form an important part of industrial technology. Very large-scale operations are applied to the manufacture of polyethylene, while one of the outstandingly successful examples of transformation of laboratory techniques to a major manufacturing activity is found in the synthesis of diamond. The former work involves large volumes and low pressures of a few tenths of a GPa, while the latter is carried out at about 5.5 GPa (simultaneously with temperatures near 1600 K). Polyethylene is one of the most important common plastics requiring production of an enormous yearly tonnage. See SOLIDS, HIGH-PRESSURE EFFECTS IN and DIAMOND AND DIAMONDLIKE CARBON for more detailed discussion of man-made diamonds.

Finally, the matter of safety of high-pressure systems, particularly large pressure vessels containing compressed gases, is an important and extensively studied subject (Dawson, 1977). Even small-scale laboratory apparatus using significant volumes of gas or liquids at high pressures must be treated with respect, using appropriate guards and shields, as well as common sense, to minimize chances of an injurious accident. Fortunately, small-scale "dry" systems such as the widely used diamond-anvil cells may be made safe by following relatively simple procedures, especially to protect eyes from chips that could spall from stressed brittle materials.

1. STUDIES USING STATIC TECHNIQUES

1.1 Principles

"Pressure" is given by normal force per unit area. The simplest means of applying pressure might be found in an apparatus in which a sample is squeezed between faces of two cylinders by a force F yielding $P=F/A$. If the length of the cylinders was about equal to the diameter, a definition of compressive strength σ_c of the cylinder material suggests that the maximum pressure on the sample would be about equal to σ_c. Although such a simple device could in fact be useful, pressure cells are usually more complex structures and can reach pressures higher than the simple compressive strength.

1.1.1 Strength of Pressure Chambers. In Fig. 1, we note that the splitting force generated by the internal pressure $F=Pdl$ in a simple cylinder must be successfully opposed by the tensile stresses in the wall of the cell. The wall stresses are not uniform: the tangential stress is largest near the interior wall and diminishes at larger radial distances. A cell made of a reasonably ductile material, as is usually the case, typically fails by a crack parallel to the axis that propagates outward from the inner cell wall. Quantitative treatments of the state of stress within simple pressure-vessel forms may be found in numerous places (Spain and Paauwe, 1977) but must always be used with caution as predictors of failure pressure. One must particularly note that the so-called thin-wall formula given by (average tensile stress) = (pressure) $\times d/(D-d)$ overestimates the bursting pressure of a thick-walled vessel if the average tensile stress is taken to be the tensile strength of the material of which the vessel is constructed. Not surprisingly, as interest has grown in reaching even higher pressures, stronger and stronger materials have been used in constructing high-pressure components, ultimately including diamond, which is the hardest material known at this time. Also, "geometrical" and statistical advantages have been applied. Of these, the so-called principle of massive support and Auerbach's law (Seal, 1984) are sufficiently evident in contemporary apparatus to demand description. The principle of massive support states that higher contact pressures may be obtained without failure when the contact area is supported by adjacent relatively unstressed regions, as in Fig. 2 showing a so-called Bridgman anvil. Auerbach's law will be described in Sec. 1.5 in relation to diamond-anvil cells. Calculating the stresses in a system such as a diamond-anvil cell is a complex problem, but a number of "finite-element" calculations have been carried out (Moss and Goettel, 1987) both to gain insight into maximizing performance of the cells and to improve confidence in values of indirectly measured pressures.

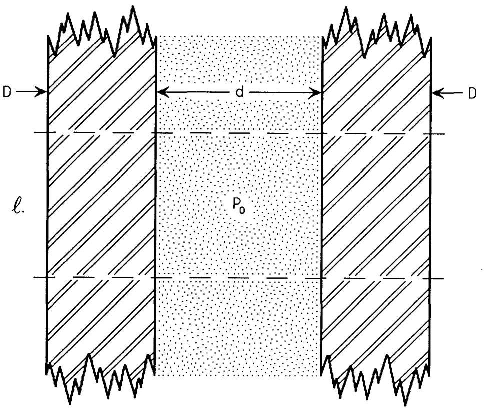

FIG. 1. Cross section of a long, hollow cylinder filled with fluid at a pressure P_0. d and D represent internal and external diameters, respectively. The splitting force P_0dl must be opposed by the tensile stresses in the cylinder walls.

1.2 Materials of Construction

Construction materials are quite varied, depending not only on the mechanical strength required but, in particular cases, on the need to resist corrosion, on performance at high temperature, on optical access, etc.

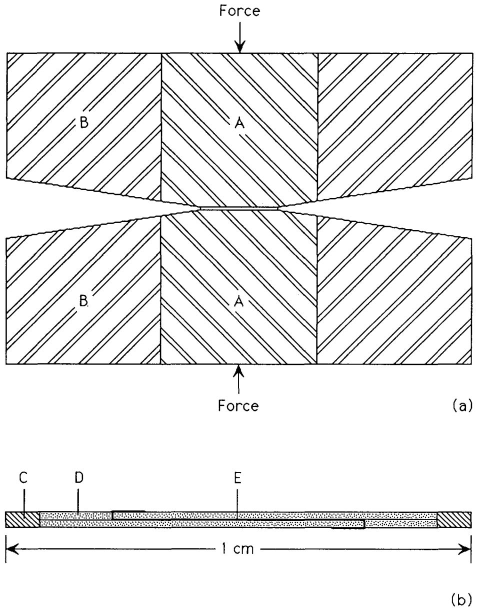

FIG. 2. (a) "Bridgman-anvil" apparatus: A, cemented tungsten carbide anvils; B, support rings of alloy steel. (b) Sample detail: C, compressible gasket; D, silver chloride or other "soft" material; E, electrical resistor wire sample. The diameter ϕ of the compressed region is typically about 1 cm.

Traditionally, for nonaggressive pressure media, alloy steels combining high strength and reasonable ductility such as SAE 4340 have been used. Alloys such as this requiring quenching from high temperatures show substantial deformation in the hardening process so that final machining often must be done after heat treatment. A more recently available class of alloys, the so-called nickel maraging steels, combine tensile strengths greater than 2 GPa with reasonable ductility, do not require quenching, and hold their dimensions well during the aging process by which they are hardened. Experience has shown, however, that the material frequently contains flaws, and pieces intended for critical construction should be sonically tested before a substantial investment is made in machining them. Various stainless steels are frequently used for components, most commonly type 316, which has good corrosion resistance properties. Pre-heat-treated type 416 stainless steel is a higher strength, less corrosion-resistant variant, which is useful as a general material for moderately stressed parts. It is sometimes forgotten that although the *strengths* of different steel alloys may be quite different, the elastic moduli that determine their *elastic* deformations under stress are practically the same, regardless of composition or condition of hardness. Thus it is often not necessary to use expensive superalloys for support purposes in which rigidity is required rather than extreme strength. Some pressure media affect pressure cells chemically or physically. Hydrogen in particular may severely attack steel alloys. Hydrogen embrittlement may be avoided by use of beryllium copper (BeCu) alloys or hydrogen-resistant steels such as A-286, which are, however, neither as strong nor as ductile as the maraging steels when both are heat treated to nearly maximum strength. Oxygen has been known to ignite a pressure vessel from within, with disastrous results. [It is often possible to use alloys resistant to attack such as BeCu as material for the core of a composite vessel, along with a material such as type 416 pre-heat-treated stainless steel as an external support—designed such that the latter is not exposed to the high-pressure fluid.] Various grades of cemented tungsten carbide are used where great compressive strength is needed but significant ductility is not required. These are commonly found in piston-cylinder and "anvil" applications, which will be discussed later. The ultimate high-strength material at present is diamond. This is extensively used in so-called diamond-anvil cells discussed in Sec. 1.5 of this article. Specialized, very simple cells are also being constructed using commercial sapphire spheres as anvils (Takano and Wakatsuki, 1991; Daniels *et al.*, 1991). Other materials used for particular purposes, such as electrically insulated lead-ins, optical windows, and neutron transmission will be discussed as they appear under those techniques.

1.3 Systems Using Pumped Fluids as the Pressure Medium

In spite of the relatively low pressure range, apparatus using pumped fluids demand the most complex equipment. Practically speaking, pumped-fluid apparatus are limited to about 3 GPa maximum pressures, but are more commonly used to only slightly above 1 GPa. The chief advantages of these systems are that the pressures can be truly hydrostatic (hence will not induce shear stresses even in soft single crystals) and accurately measured, and that they are very versatile. A wide variety of electrical, optical, magnetic, and other measurements may be made on compressed materials using these systems. Figure 3 illustrates the components of a typical system intended to compress a sample using helium as pressure medium to pressures near 1 GPa. It consists of a gas supply, a prepressurizing system, intensifiers, a pressure-measuring device, and a sample cell containing the material under study. The components are connected by high-pressure tubing. The sample cell is designed to accommodate appropriate electrical leads, optical windows, etc., relevant to the phenomena under study. The cell shown would be appropriate for Raman scattering. Many of the components, especially for the lower parts of the 1-GPa range, may be obtained commercially.

1.3.1 Seals. Prevention of leakage at pistons, closures, and connections is a fundamental problem in working with fluids at high pressures. The so-called Bridgman unsupported-area principle (Bridgman, 1952) is widely applied to provide leak-free closures. Figure 3(b) illustrates an application to a compression piston seal. The principle is based on the observation that fluid will not leak past

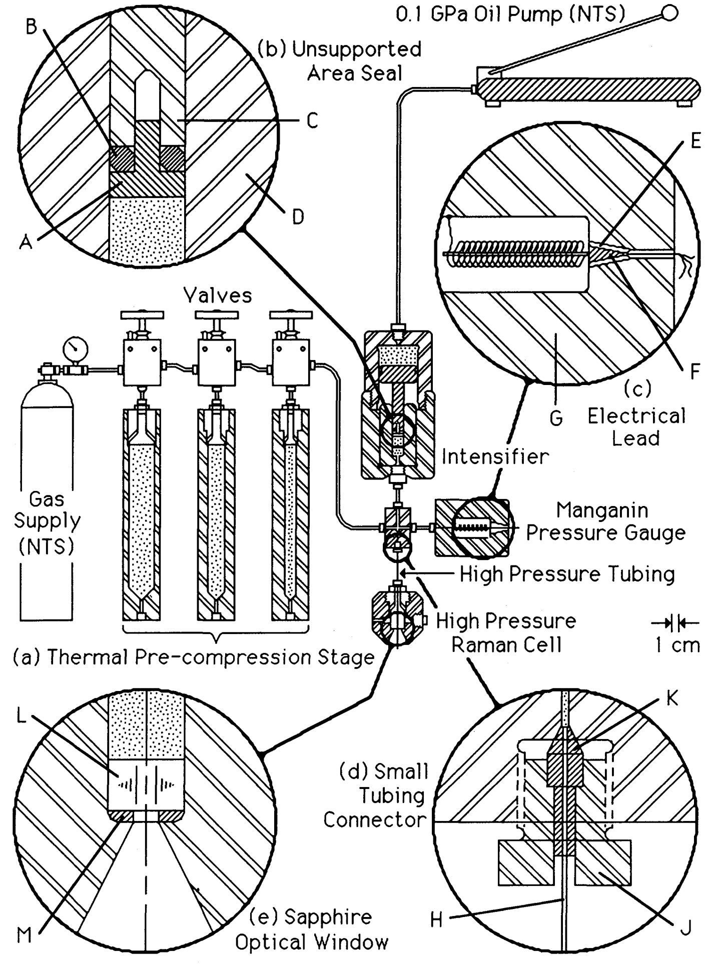

FIG. 3. System for generation of high pressure using helium as pressure fluid. The gas is supplied by a commercial high-pressure cylinder. The pressure cell shown is used for Raman scattering. In the details, *A* is the so-called mushroom of the unsupported area seal; *B* is the soft packing, guarded by antiextrusion rings; *C* is the high-pressure piston of the intensifier; *D* is the intensifier cylinder; *E* is the thin insulating core; *F* is the matching conical electrode; *G* is the pressure-gauge cell wall; *H* is the pressure tubing; *J* is the gland; *K* is the plug, with tubing hard-soldered into it; *L* is the single-crystal sapphire window, oriented with the crystallographic *c* axis along the cylinder axis; and *M* is the window seat. (NTS = not to scale.)

a soft solid packing in which the pressure is higher than the pressure in the fluid, together with a geometry for maintaining that condition. Ignoring frictional forces and nonhydrostatic stresses, the mushroom-shaped piece *A* is in equilibrium under the force applied by the pressure medium $F = P \times$ (area of the mushroom heads), and the force $F' = P'$ in the packing multiplied by the area of the annulus S' formed by the packing: thus $P'_{\text{packing}} = P_{\text{system}} \times S/S'$. Geometrically, $S' < S$ and so $P' > P$. That is, the packing pressure is automatically maintained higher than the fluid pressure and leakage will not occur. In practice, small scratches and irregularities can provide pathways for leakage, especially with light gases such as helium and hydrogen at cryogenic temperatures where sealing materials lose their plasticity. Careful construction and polished surfaces are beneficial. Packings are typically inert, soft, low-friction materials such as Teflon^{TM} or soft metals such as lead or indium (although the latter seems to embrittle some steels and beryllium-copper alloys). Antiextrusion rings, usually of phosphor bronze or brass, must be used to prevent the soft packing from extruding either out to the atmospheric pressure region or even into the pressure chamber (since $P' > P$!).

1.3.2 Tubing. Tubing suitable for high pressures is available from several suppliers. The reader is referred to commercial catalogs for details of its use, which typically involves threading a collar onto the tubing and forcing a conical end of the tubing into a mating cavity. Very small tubing [e.g., type 316 stainless steel tubing as small as 0.8 mm diam (0.32 in.) having a bore 0.15 mm (0.006 in.)] is convenient in coupling to sample cells, which are to be kept at very low temperatures, because of the mechanical flexibility and the small heat conduction by such tubes. Connections for this small tubing are shown in Fig. 3(d). The soldering operation must be done carefully to avoid overheating and annealing the tubing, which is usually more sensitive to temperature than the plug, which may be made of maraging steel. Maraging steel fittings, type 316 stainless steel tubing, and a silver solder have been used comfortably with inert pressure fluids (e.g., helium) to greater than 1 GPa. Beryllium copper fittings and tubing together with high-strength solder have been used satisfactorily, but never comfortably, with an aggressive material, hydrogen, to about 0.7 GPa.

1.3.3 Precompression. If the pressure fluid is a relatively incompressible material such as oil, precompression may take the form of a commercial oil pump that can fill the system with oil and drive the intensifier piston to an appropriate initial position. On the other hand, if the medium to be compressed is a gas, e.g., helium, it is often desirable to precompress it so that the desired high pressure may be reached in one stroke of the high-pressure intensifier. There exist 0.2-GPa diaphragm pumps for this purpose. An inexpensive alternative, satisfactory for light gases such as He or H_2, uses thermal compression. If one has convenient access to liquid nitrogen, the method may be a useful one. Commonly three cylinders, having decreasing volumes and increasing maximum pressure capability, are connected by tubing and valves as shown in Fig. 3(a). In operation, all three cells are initially immersed in liquid nitrogen and filled from a standard commercial helium or hydrogen high-pressure gas cylinder while connected to the system. After pressure and temperature have equilibrated, the valve between the supply tank and the first cylinder is closed, and that cylinder is warmed gently to room temperature or even higher, driving gas to subsequent cylinders and to the system. Then the valve between the first and second cylinders is closed and the second cylinder is warmed, etc. Three stages ($V_1=300$ cm^3, $V_2=100$ cm^3, $V_3=30$ cm^3) used with hydrogen generate about 0.2 to 0.3 GPa starting pressure in a system having a volume about 10 cm^3 with the intensifier initialized. Advantages of thermal compression are simplicity and absence of moving parts, which might introduce contaminants. A disadvantage is the relatively large volume of gas held at high pressure. *Safety shielding must be provided.*

1.3.4 Intensifiers. The intensifier is an elementary pressure multiplier having gas (system) pressure at one end and oil (pump) pressure at the other end. If the pressures and piston areas are P_1, A_1, and P_2, A_2 and the piston is in mechanical equilibrium, $P_1 A_1 = P_2 A_2$ ignoring friction. (Maximum oil pressure$\approx$0.1 GPa.) Ratios are commonly chosen so that a commercial oil pump can provide the low pressure needed. The high-pressure piston seal is usually a Bridgman seal, and the low-pressure side may use an O-ring or other packing together with an antiextrusion ring. Principally for safety, an external supporting sleeve of ductile material is usually pressed onto the high-pressure end of the intensifier. An appropriate valve arrangement may be used so the system can be multiply stroked to accommodate larger-volume sample cells.

1.3.5 Valves. Valves may be purchased or constructed in a variety of configurations. Commercial valves are available for pressures to about 1 GPa. Usually commercial valves require adaptation for use with very small-diameter pressure tubing.

1.3.6 Pressure Measurement. The fundamental measurement of pressure as a force per unit area is the concept embodied in the free piston gauge (Newhall *et al.*, 1963), sometimes also called a pressure balance or deadweight tester. The principle is simple—balancing of the force generated by pressure on a piston having a known area, against the gravitational force on weights placed on top of the piston—but the practice is complex. Sophisticated schemes to reduce friction by rotating the piston, application of controlled clearance seals, and calculations of corrections due to area changes and hydrodynamic stresses lie beyond the scope of this article (Newhall *et al.*, 1963). More practical secondary gauges include Bourdon-tube gauges and gauges based on pressure dependence of electrical resistance R of a coil of wire within a pressure cell. The Bourdon-tube gauges are available in various ranges, up to about 0.7 GPa with accuracy claimed to be as high as 0.1% of full scale. Electrical resistance gauges usually use Manganin wire, an alloy of copper, manganese, and nickel, because of its availability and its relatively small sensitivity to temperature change. Typical values of the pressure coefficient $d \ln R/dP$ are near 2.3 $\times 10^{-2}$ GPa^{-1}.

In the search for a calibration reference more widely accessible than the free-piston gauge, the melting pressure of pure mercury has been accurately determined as a function of temperature with free-piston gauges (Lloyd *et al.*, 1969). Using a reasonably high-quality resistance bridge and a temperature-controlled cell containing elemental mercury, together with a means for determining coexistence of the fluid and solid phases, any laboratory can calibrate its gauges to within better than 0.5% over a range to ~1 GPa. An alternative to the Manganin cell—useful for material such as hydrogen (which attacks Manganin)—is a small vessel with conventional strain gauges mounted on its *exterior* surface to sense expansion generated by the internal pressure.

1.3.7 Dependent-Variable Techniques in Fluid Systems. These techniques deal with communicating values of various physical quantities from within pressure cells. For example, electrical resistivity measurements require electrical leads into the system, optical absorption studies require transparent windows in appropriate regions of the spectrum, etc. The following are some of the most generally used techniques.

1.3.7.1 Electrical Leads. These are required for measurements involving electrical resistance or dielectric constants. The technique requires use of stressed electrically insulating material. A difficulty has been that insulating materials are usually not as strong as metals. The most widely used electrical lead is due to Bridgman (1952). It has the form of a 16° conical aperture open toward the high-pressure side of a closure, insulated by a thin (~0.2 mm) hollow cone of insulating material as shown in Fig. 3(c). Originally, mineral material such as pipestone or pyrophyllite were used as insulators; contemporarily a strong plastic such as nylon, Delrin^{TM}, or Lexan^{TM} is used. The unsupported-area principle is involved with this, i.e., the pressure in the insulator is higher than in the system, but attention to quality of the conical surfaces is required, especially when used with light gases as a pressure medium.

1.3.7.2 Optical Windows. These permit absorption and reflectivity measurements within their spectral transmission regions, and scattering experiments such as Raman and Brillouin scattering. Fortunately, single crystals of sapphire combine high mechanical strength and broad spectral transmission—in principle from about 6 μm in the infrared to about 150 nm in the ultraviolet, but impurities reduce the range significantly in typical commercially available crystals. Recently, however, nonlinear optics methods have been applied to high-pressure spectroscopy in order to extend the spectrum to excitations deeper in the ultraviolet (Lipp and Daniels, 1991). In these multiphoton methods, resonances at excited-state energies that are sums and differences of the probing photon energies become accessible. Thus, these processes can use photons that are transmitted by the windows to probe states at energies inaccessible to one-photon processes because of window absorption at those energies. The technique may be more important in anvil-type apparatus. The form of window most commonly used takes the form of a short cylinder, having the crystal *c* axis parallel to the cylinder axis, about 6 mm diam. by 4.8 mm thick covering an aperture about 3 mm diam. (or similar proportions) on the interior of a pressure cell or closure plug as shown in Fig. 3(e). Both the window and the seating surface must be very flat and highly polished. Sometimes an aluminum foil gasket is used between the window and its seat to improve sealing. The arrangement of the windows can be made to optimize the requirements of a particular technique. For example, for laser-excited Raman scattering, a window for collecting the weak scattered light may be at 90° to the exciting laser-beam window axis, and the laser windows may be much smaller in size than the collection window.

1.3.7.3 X-Ray Diffraction. This very useful technique is only rarely applied to pumped fluid systems—largely because of the large absorption of x rays by common high-strength materials and because diamond-anvil cells excel in this respect. On the other hand, a few investigations have been done below a few tenths of a GPa using very small beryllium cells on account of the low absorption of x rays by that material (So *et al.*, 1992).

1.3.7.4 Neutron Diffraction. In contrast with x rays, neutrons interact weakly with most materials; hence pressure cells relatively transparent to neutrons may be made. Unfortunately, as a corollary to this, samples must be relatively large to be studied successfully by neutron scattering, typically ~0.25 cm^3, requiring cells capable of accommodating them. Several designs have appeared to accomplish this, based on either aluminum alloys (Schmidt *et al.*, 1984) or high-strength alumina (Block *et al.*, 1976).

1.3.7.5 Pumped Systems at Cryogenic Temperatures. Use of low temperature is often a consideration in study of solid-state physics phenomena. A technique that has been used in a substantial number of combined high-pressure, low-temperature experiments requires solidifying helium around the sample under study, at high pressure, then cooling the combined sample+solid helium to the temperature of observation. After the helium freezes, such a process follows approximately an isochore (constant-volume trace) of the solid helium, yielding a final pressure dependent on the final temperature in an accurately predictable way (Schirber, 1970). In spite of the fact that the pressures are applied by *solid* helium, the stresses have been observed to be very hydrostatic and homogeneous.

1.4 Dry Systems—in Which Solids are Generally Used as Pressure-Transmitting Media

These fall into two categories, piston-cylinder devices and compressible-gasket systems. Both have the advantages of great simplicity and typically can be used at much higher pressures (12 GPa) than fluid systems. Readers are referred to a recent review covering these (Bundy, 1988). Some of the dry systems allow simultaneously high temperature together with high pressure. Consequently they often form the basis of geophysical investigations and are used in the commercially important process of diamond synthesis. The advantages are obtained at the expense of a system of stresses on the sample that is often nonhydrostatic and inhomogeneous as well as generating some difficulty in accurately defining and measuring the pressures applied. The article by Hall (1958) is still a useful reference for anyone contemplating work at high pressures and temperatures (although both the maximum pressures and temperatures suggested as possible are on an obsolete scale, values reported in early work being too high).

1.4.1 Piston-Cylinder Devices. As the name suggests, these consist of a pair of pistons loaded by an external press compressing a sample within a hollow cylinder filled with a solid or semisolid material. The latter may be the sample itself, as in the case of direct compression for equation-of-state measurements, or a material such as pyrophyllite, a hydrous aluminum silicate, or Teflon^{TM}, surrounding the sample under study. Pyrophyllite may easily be machined to accommodate specific samples, heaters, etc., and is a relatively good thermal insulator. The cylinders used for the highest pressure work may use cemented tungsten carbide cores, surrounded by one or more pressed-on steel support rings. The pressure limit of these devices, about 5 GPa, is usually set by compressive failure of the cemented tungsten carbide pistons. Pressure may be inferred from the force exerted on the pistons by the external press, calibrated in a independent experiment involving a known pressure dependence such as the bismuth I-II phase transitions at 2.55 GPa. The change of phase of the bismuth is signaled by large change in electrical resistance. The press load versus pressure on the sample in the piston-cylinder device usually has a large hysteresis upon cycling up and down in pressure. Solid lubricants such as molybdenum sulfide or powdered Teflon^{TM} are sometimes used to minimize the hysteresis.

Pyrophyllite is a good electrical and thermal insulator, making it possible to use internal electrical resistance heaters surrounding a sample region to generate temperatures ~1500 °C while keeping the stressed cylinder walls at an acceptably low temperature. Sample temperatures may be measured using thermocouples led into the solid matrix from the outside or by detecting melting points of materials in the cell. A modification of the piston-cylinder devices has been used extensively at low temperatures for superconductivity and magnetism studies. In it, a mechanical arrangement is made to clamp the piston in place after pressurizing the system at room temperature in a press, retaining the sample in its compressed state (Chester and Jones, 1953). The clamped piston-cylinder device is then removed from the press, placed in a cryostat, and cooled to the measurement temperature as a free-standing object.

1.4.2 Anvil Systems. Historically, in the hands of P. W. Bridgman, these generated a major step upward in the limits of attainable pressures. They introduce the principle that fractionally small surface areas on objects may be much more highly stressed than the usually defined compressive strength limit of a material. This is referred to as the principle of massive support (cf. Fig. 2). Further, since the stresses in the externally supported anvils are almost entirely compressive, it is possible to use in their construction hard but brittle materials such as cemented tungsten carbide, supported on their external cylindrical surfaces by pressed-on alloy steel rings (Hall, 1958).

1.4.2.1 Compressible-Gasket Anvils. In the most primitive form, the stressed areas may be small truncations on a pair of shallow conical "anvils" with the sample occupying a wafer-shaped space between the opposing faces. Mean pressure may be inferred approximately from the quotient of applied force and the area of the face. Addition of a compressible gasket of pyrophyllite as shown in Fig. 2 defines the sample better, and with use of a soft medium such as silver chloride or Teflon^{TM} surrounding the sample, a more homogeneous distribution of pressure is obtained, making it possible to measure semiquantitatively even structure-sensitive properties such as electrical resistance. Significant increases in pressure range and addition of optical access is provided by Drickamer cells (Drickamer and Balchan, 1962) in which the conical surfaces of the anvils are partially supported by extending and confining the gasket. If transparent material forms at least part of the gasket, these cells may be used for measurement of optical properties of materials to pressures greater than 10 GPa, and they form the basis of a large number of such studies.

1.4.2.2 Belt-Type Apparatus. Flat anvil devices may be heated externally as long as the temperatures remain below the softening points of the materials of construction. The very small dimensions of the compression space makes it difficult to use internal heating while keeping the anvil faces cool. A major improvement is provided by the so called "belt apparatus" (Bovenkirk *et al.*, 1959), which was used in the historic General Electric Company diamond synthesis. As shown in Fig. 4, the belt combines features of piston-cylinder and compressible-gasket cells, generating a larger and more isotropic sample space subjectable to both high temperature and high pressure while remaining somewhat thermally insulated from the stressed piston and pressure chamber walls.

1.4.2.3 Multiple-Anvil Designs. For reasons that form an interesting story in itself, Hall also reported the first of the multiple-anvil designs, the so-called tetrahedral anvil system (Hall, 1958). It consists of a set of four hydraulically driven tungsten carbide anvils oriented along normals to the faces of a regular tetrahedron. The anvils each have a triangular face and are shaped so as to leave an enclosure in the shape of a tetrahedron when they are advanced into mutual contact. In this case, the sample chamber is embedded in an oversized tetrahedron machined from pyrophyllite, and sample compression occurs by collapse of the tetrahedron. The structure is assembled so that the anvils uniformly contact corresponding faces of the sample block. Pressure is raised by advancing the rams, compressing the block. An internal electrical heating system may be used to generate high temperatures for geophysical studies. Experimentation requiring electrical leads is routine, and x-ray measurements have been

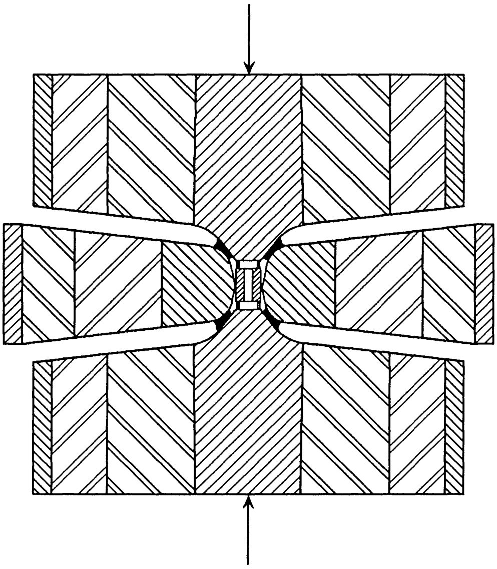

FIG. 4. Cross section of the General Electric "belt" apparatus. Note the tampered pistons and cylinders.

made by bringing the beams through the gasket material in the spaces between the anvils. More recently other multiple-anvil configurations have appeared, using six anvils in a cube configuration and, notably, an ingenious octahedral design involving eight relatively simple individual core anvils in the form of cubes with truncated corners enclosing a regular octahedron (Walker *et al.*, 1990). These devices provide electrical resistance heating of samples within the pyrophyllite octahedron. They are used extensively in geochemical studies.

1.5 Diamond-Anvil Cells (DACs)

Even though diamond-anvil cells are examples of the "anvil" format, a separate category is reserved for them because their application and refinement have been so extensive as to make them qualitatively different from other anvil systems (see Fig. 5). For example, they provide the highest static laboratory pressures. Commonly used techniques provide pressures in the 40-GPa range, while pressures above 400 GPa (Ruoff *et al.*, 1990) have been reported in some experiments. Laser heating has made possible studies to temperatures above several thousand kelvins for times less than a few milliseconds, together with pressures of several tens of GPa (Bassett and Weathers, 1987). In addition, a wide variety of dependent variables may be measured, not all, however, to the same ultrahigh pressures. There are several excellent and useful reviews available (Jamieson *et al.*, 1959; Jayaraman, 1983, 1986; Jephcoat *et al.*, 1987; Dunstan and Spain, 1989; Spain and Dunstan, 1989), and advances in technique appear very rapidly.

The fundamental reason that diamond is appropriate is that it has the highest compressive strength among known materials. The additional benefits of reasonable optical and

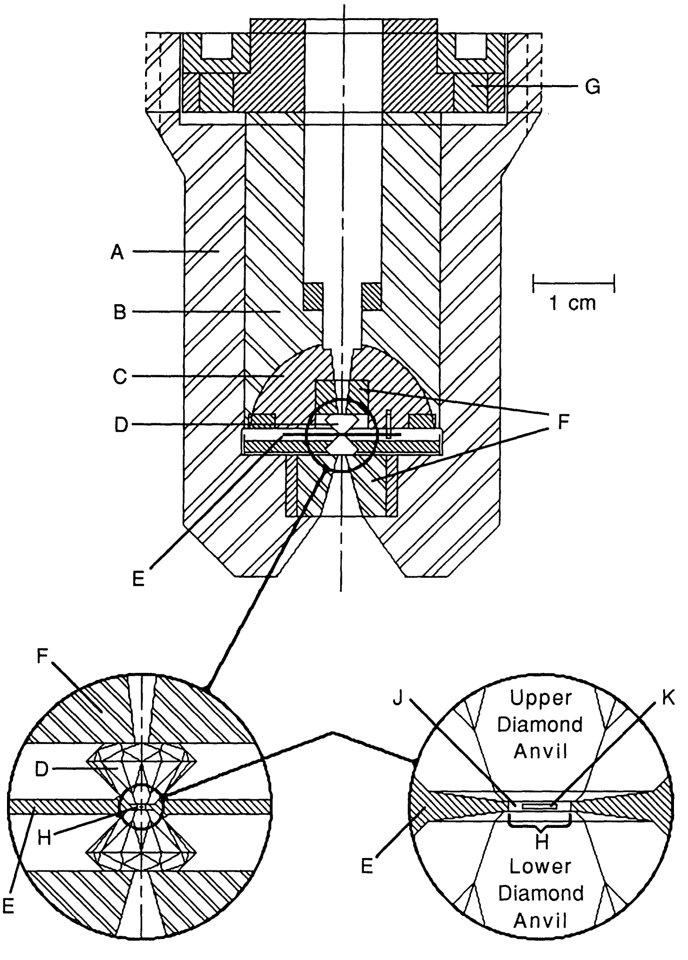

FIG. 5. Diamond-anvil cell, showing whole cell including screws used to generate the force between the diamonds, and successively enlarged details of the high-pressure regions. *A*, the cylinder, closely fitted to *B*, the sliding piston; *C*, the adjustment to make the anvil faces parallel; *D*, diamond anvil; *E*, gasket; *F*, tungsten carbide backing field; *G*, machine screws used to apply force to the diamonds; *H*, hole in gasket, to contain sample, pressure fluid, and ruby; *J*, ruby chip for pressure measurement; and *K*, sample. Most often the sample together with the ruby fills the entire gasket hole.

x-ray transparency over wide spectral ranges have made it a material of choice. Presence of a mature technology from the gemstone industry for shaping and polishing diamonds has also been a major positive factor. The advance to higher pressures has even seen selection of natural diamonds having appropriate impurities, e.g., nitrogen platelets, to provide higher strength. Design of details of the highly stressed tip of the anvil has advanced in a way reflecting collective experience. The ever-diminishing contact areas as attempts are made to reach even higher pressures reflect Auerbach's principle, which suggests that larger specific stresses can be applied without fracture to smaller objects than to larger ones constructed of the same material (Seal, 1984). Parallel development of lasers for optical studies and intense synchrotron sources for x-ray diffraction have helped mitigate the problems of diminishing sample volumes that have accompanied the advance to higher pressures. Referring to Fig. 4, the diamond-anvil cell consists of two diamonds, typically $\frac{1}{4}$ carat each, having the general form shown, mounted in a frame that permits alignment of the flat mutually contacting faces in both position and orientation. In addition, the frame provides a means of forcing those surfaces against one another. A substantial number of different configurations exist, but they share the essential features listed. Originally, samples to be studied were simply compressed between the small surfaces formed by grinding flats onto the culets (pointed ends) of two "brilliant cut" diamond gem stones mounted in opposition. Even this primitive process permitted x-ray-diffraction (Jamieson *et al.*, 1959) and infrared-absorption (Weir *et al.*, 1959) measurements to pressures of about 3 GPa, although pressures were neither very accurately determined nor very uniform. Subsequently, the "gasket" was developed at the National Bureau of Standards (now NIST) to contain the sample material, a pressure-transmitting medium (sometimes), and a tiny chip of ruby to serve as pressure sensor (Piermarini and Block, 1975). The procedure for preparing the system for experiment is as follows: first, position the diamonds and adjust their orientations to make the faces parallel; then, place a thin sheet of material (typically 300 μm thick) as the gasket between the faces and indent it with the diamonds. The indented gasket is removed and a small hole drilled in its center. The gasket is replaced on one of the diamonds; the sample, the pressure fluid, and a ruby chip are put in the hole; and the other diamond is brought into place. The pressure is raised by increasing the force pushing the diamonds together against the gasket-sample assembly. These developments and the addition of profiling (bevels) by the Carnegie group (Mao *et al.*, 1979) more or less completed the basic design shown in Fig. 5, which represents the crucial core of the DAC.

1.5.1 Pressure Measurement in Diamond-Anvil Cells. The earliest work depended on observation of phase transitions such as those in bismuth near 2.5 GPa and in various semiconductors. The invention of the ruby fluorescence method made the DAC a much more quantitative instrument (Barnett *et al.*, 1973). In this method, the pressure-dependent value of wavelength of sharp fluorescence lines in ruby (chromium-doped sapphire) crystal chips within the pressurized volume is measured. This fluorescence may be laser excited, and the wavelength measured using relatively straightforward spectroscopy to an accuracy of about 0.05 GPa at low pressure. The accepted method of calibration in the very high-pressure domain lies in connecting x-ray–determined equations of state of certain solids, for example gold and copper, with corresponding pressure-volume relations inferred from shock-wave studies—that is, setting $P_{\text{ruby}}(V) = P_{\text{shock}}(V)$, where the left-hand side of the equation involves static x-ray measurements of volume of a reference sample material together with the ruby fluorescence wavelength, while the right-hand side is a 300-K pressure-volume isotherm inferred from shock-wave compression experiments on the same material (Mao *et al.*, 1978). In a recent paper reporting pressures as high as 416 GPa, the x-ray measurements of volume of molybdenum and tungsten have *per se* been taken directly as the piezometric property to be calibrated by shock-wave equations of state (Ruoff *et al.*, 1990) at low pressures (<100 GPa). There is substantial agreement about the ruby scale, but its use in the domain above 200 GPa is the object of intense study and some controversy at this time (Ruoff *et al.*, 1991). Accurate measurement of pressure at high temperature remains a difficult problem, partly because of thermal effects on the ruby wavelength and increased spectral linewidth. Space limitations of this article force us to provide little more than a listing of the measurements that can be made in diamond-anvil cells and to give either reviews or recent articles as references.

1.5.2 X-Ray Studies. X-ray diffraction techniques, which may (in principle at least) uncover both crystal structures and equations of state, form a particularly important component of high-pressure research. Fortunately, diamond is a low–atomic-number material relatively transparent to x rays. For example, in pure diamond the $1/e$ intensity decay distance for molybdenum $K\alpha$ radiation is about 4 mm. Use of x rays as a probe in diamond pressure-cell studies dates from the earliest (Jamieson *et al.*, 1959) through some of the most recent work (Ruoff *et al.*, 1990) (which reports measurements to 416 GPa). Availability of synchrotron sources and energy-dispersive x-ray analysis have reduced data collection times and increased the general capability of the techniques by permitting very fine collimation ($\sim$4 μm) of the incoming beam (Ruoff *et al.*, 1990), and hence corresponding spatial resolution within the sample. Measurements reported in the highest pressure regimes have been made on relatively hard materials such as copper, gold, molybdenum, and tungsten. It is much more difficult to generate very high pressures in "soft" materials such as solid hydrogen and helium. For example, x-ray lattice parameter measurements in solid hydrogen (which is both "soft" and a weak x-ray scatterer) to 26.5 GPa represent a real tour de force (Hemley *et al.*, 1990).

1.5.3 Optical Studies. Typical natural diamonds used as anvils are transparent through most of the optical region of the spectrum. Specially selected diamonds at low pressures may be transparent from the near ultraviolet (5.5 eV) through 4 μm and from 5.5 to 10 μm in the infrared (Ferraro, 1984). However, most of the diamonds used for anvils are natural stones having a variety of impurities, nitrogen in particular, which dominate the optical properties. The optical absorption due to the impurities in fact forms the basis of classification of diamond types. The most common (Type I) diamonds only reach about 3.5 eV in the ultraviolet and show strong absorption in the 7–10-μm region. Optical transmission and reflectivity measurements as well as fluorescence and Raman and Brillouin scattering spectroscopies have been carried out. Raman spectroscopy, because it may be excited in the tiny samples by intense, fixed-frequency lasers, has been the most prominent of these techniques. Infrared spectroscopy suffers somewhat from the small sample geometry but, even so, has contributed significantly to high-pressure chemistry (Ferraro, 1984); and availability of synchrotron sources of infrared radiation promise to expand this area still further (Hanfland *et al.*, 1992). Reflectivity measurements in the visible and infrared have been extensively used as a diagnostic technique in recent studies of metallization of condensed permanent gases (Reichlin *et al.*, 1989). In addition to the various optical spectroscopies, microscopic imaging has been used to study growth and transformations of crystals at high pressures as well as direct estimation of volume changes under pressure. The combination of microscopy, video cameras, and recorders with diamond-anvil pressure cells has been especially useful in this regard.

1.5.4 Electrical Transport. Electrical conductivity measurements, historically an important probe in solids, have gained a new urgency in dealing with critical questions of the insulator-metal transition at high pressures. The status to about 1985 is reported in Jayaraman's review (Jayaraman, 1986). Most contemporary designs are based on use of MgO or alumina, Al_2O_3, either in sputtered form or as a powder in a matrix, as the stressed insulating material through which the conductors are led. In some investigations, microlithography has been used to form an array of electrical leads on the surface of the diamond itself (Grzybowski and Ruoff, 1984). Using the latter with Al_2O_3 in a sodium chloride matrix, conductivity measurements have been reported to 70 GPa in cryocrystals (van Straaten and Silvera, 1987). Major problems deal with open circuits in the stressed electrical wires and with the relatively low strength of insulating materials when compared with metals as gaskets.

1.5.4.1 Miscellaneous Techniques—Magnetic, Mössbauer, Nuclear Magnetic Resonance. The small sizes of samples in DACs make static magnetic measurements difficult, and in fact most of the magnetic susceptibility work has used larger-scale piston-cylinder pressure systems. Nonetheless, external coils have been used to measure static susceptibility using a SQUID (superconducting quantum interference device) detector to measure the small changes in magnetization (Webb *et al.*, 1976). Nuclear magnetic resonance (NMR) (Lee *et al.*, 1992) and electron paramagnetic resonance (EPR) (Bromberg and Chan, 1992) techniques have been used with DACs to pressures 5.2 and 10 GPa, respectively. ^{119}Sn Mössbauer measurements have been reported to about 29 GPa (Sterer *et al.*, 1990). These techniques have not yet seen as widespread application as x-ray and optical techniques, but do indicate the great versatility of the diamond-anvil cell.

2. DYNAMIC TECHNIQUES—SHOCK WAVES

Intense compression waves called shock waves may be generated in materials by explosives, by impact with high-velocity projectiles, or by absorption of intense light pulses (Holmes *et al.*, 1989; Cottet *et al.*, 1986). Shock pressures can be very high, exceeding 600 GPa. The high pressures in shocked material are accompanied by very high temperatures, ~2000–15 000 K. Shock waves are a transient phenomenon lasting typically $\sim 10^{-6}$ s. This requires data acquisition systems capable of measuring and recording physical phenomena on correspondingly short time scales. In spite of these difficulties, an impressive variety of measurements are made on shock-compressed material (see Solids, High-Pressure Effects in). Fundamental among these are measurements of pressure-volume-temperature (*PVT*) equations of state of materials. These data are based on the Rankine-Hugoniot equations that follow from conservation of momentum, mass, and energy in the response of the system to the impact. The equations take a simple form relating the thermodynamic variables pressure (P), volume (V), and energy (E) to the observable quantities shock propagation speed u_s and the particle speed (of the shocked material) u_p:

$$P_H - P_0 = P_0(u_s - u_0)(u_p - u_0),$$

$$V_H - V_0 = -(u_p - u_0)/(u_s - u_0),$$

$$E_H - E_0 = \tfrac{1}{2}(P_H + P_0)(V_0 - V_H). \qquad (4)$$

The subscripts 0 and H refer to initial and final conditions. The shock speed u_s is measured by timing the shock propagation over a known distance. The particle velocity is either inferred from collision properties of the impactor or obtained by observation of free-surface velocity as the shock wave emerges from the shocked material. Fast and accurate timing as well as well-defined planar shock fronts are required to produce accurate *PVT* data. Each shot, i.e., shock experiment, generates one point of the equation of state.

Other measurements that have been made under shock compression conditions include electrical resistance, laser x-ray diffraction, and spontaneous and coherent Raman scattering. Shock waves are also emerging as significant tools in materials synthesis and processing because the rapid quench rate can permit recovery of high-pressure phases such as diamond.

GLOSSARY

Cemented Tungsten Carbide: A composite material formed from tungsten carbide (WC) powder in a cobalt metal matrix.

Diamond: A tetrahedrally coordinated form of nominally pure carbon. Many of the characteristics of diamond crystals are dominated by impurities, e.g., nitrogen, which particularly affects its optical properties. The familiar "brilliant"-cut gem diamond has proved to be remarkably close to the ideal form for a diamond anvil if a small flat surface is cut at the pointed end of the stone (cf. Fig. 4).

Pressure Units: Pressure has units of force per unit area, or equivalently energy per unit volume. The SI unit of pressure is the pascal, Pa, defined to be one newton per square meter which is equivalent to one joule per cubic meter. The most commonly used high-pressure unit is the gigapascal (GPa), which is 10^9 Pa. The kilobar is 10^{-1} GPa. In English units, 1 GPa is about 9872 atm or 145 000 lb/in.^{2}.

Pyrophyllite: A naturally occurring hydrous aluminum silicate sometimes called Tennessee Grade A Lava.

Ruby: Chromium-doped sapphire crystal (see Sapphire below).

Sapphire (Corundum): A strong, hard (9 on Mohs scale) transparent mineral having the chemical formula Al_2O_3. In pure form it is transparent deeper into the ultraviolet region of the spectrum than diamond.

List of Works Cited

Barnett, J. D., Block, S., Piermarini, G. J. (1973), *Rev. Sci. Instrum.* **44**, 1–9.

Bassett, W. A., Weathers, M. A. (1987), in: M. H. Manghnani and Y. Syono (Eds.), *High Pressure Research in Mineral Physics*, Washington, DC: Terra Science.

Block, D., Paureau, J., Voiron, J., Parisot, G. (1976), *Rev. Sci. Instrum.* **47**, 296–298.

Bovenkirk, H. P., Bundy, F. P., Hall, H. T., Strong, H. M., Wentorf, R. H. (1959), *Nature* **184**, 1094–1098.

Bridgman, P. W. (1952), *The Physics of High Pressure*, London: Bell.

Bromberg, S. E., Chan, I. Y. (1992), *Rev. Sci. Instrum.* **63**, 3670–3673.

Bundy, F. P. (1988), *Phys. Rep.* **167**, 133–176.

Chester, P. F., Jones, G. O. (1953), *Philos. Mag.* **44**, 1281–1290.

Cottet, F., Hallouin, M., Romain, J. P., Fabbro, R., Faral, B., Pepin, H. (1986), in: Y. M. Gupta (Ed.), *Shock Waves in Condensed Matter*, New York: Plenum.

Daniels, W. B., Lipp, M., Strachan, D., Winters, D., Yu, Z.-L. (1991), in: A. K. Singh (Ed.), *Recent Trends in High Pressure Research*, New Delhi: Oxford & IBH, p. 809.

Dawson, V. D. C. (1977), in: I. L. Spain, J. Paauwe (Eds.), *High Pressure Technology*, New York: Dekker, Vol. 1, Chap. 3.

Drickamer, H. G., Balchan, A. S. (1962), in: R. H. Wentorf (Ed.), *Modern Very High Pressure Techniques*, Washington, DC: Butterworth, Chap. 2.

Dunstan, D. J., Spain, I. L. (1989), *J. Phys. E. Sci. Instrum.* **22**, 913–923.

Ferraro, J. R. (1984), *Vibrational Spectroscopy at High External Pressures*, New York: Academic.

Grzybowski, T. A., Ruoff, A. L. (1984), *Phys. Rev. Lett.* **53**, 489–492.

Hall, H. T. (1958), *Rev. Sci. Instrum.* **29**, 267–275.

Hanfland, M., Hemley, R. J., Mao, H.-K., Williams, G. P. (1992), *Phys. Rev. Lett.* **69**, 1129–1132.

Hemley, R. J., Mao, H.-K., Finger, L. W., Jephcoat, A. P., Hazen, R. M., Zha, C. S. (1990), *Phys. Rev. B* **42**, 6458–6470.

Holmes, N. C., Moriarty, J. A., Gathers, G. R., Nellis, W. J. (1989), *J. Appl. Phys.* **66**, 2962–2967.

Jamieson, J. C., Lawson, A. W., Nachtrieb, N. D. (1959), *Rev. Sci. Instrum.* **30**, 1016–1019.

Jayaraman, A. (1983), *Rev. Mod. Phys.* **55**, 65–108.

Jayaraman, A. (1986), *Rev. Sci. Instrum.* **57**, 1013–1031.

Jephcoat, A., Mao, H.-K., Bell, P. (1987), in: G. C. Ulmer, H. L. Barnes (Eds.), *Hydrothermal Experimental Techniques*, New York: Interscience.

Lee, S.-H., Conradi, M. S., Norberg, R. E. (1992), *Rev. Sci. Instrum.* **63**, 3674–3676.

Lipp, M. J., Daniels, W. B. (1991), *Phys. Rev. Lett.* **67**, 2810–2813.

Lloyd, E. C., Beckett, C. W., Boyd, F. R. (1969), *Science* **164**, 860–862.

Mao, H.-K., Bell, P. M., Shaner, J., Steinberg, D. (1978), *J. Appl. Phys.* **49**, 3276–3283.

Mao, H.-K., Bell, P. M., Dunn, K. J., Chrenko, R. M., DeVries, R. C. (1979), *Rev. Sci. Instrum.* **50**, 1002–1009.

Moss, W. C., Goettel, K. A. (1987), *J. Appl. Phys.* **61**, 4951–4954.

Newhall, D. H., Abbot, L. H., Dunn, R. A. (1963), in: A. A. Giardini and E. C. Lloyd (Eds.), *High Pressure Measurement*, Washington, DC: Butterworth, p. 339.

Piermarini, G. J., Block, S. (1975), *Rev. Sci. Instrum.* **46**, 973–979.

Reichlin, R., Brister, K. E., McMahan, A. K., Ross, M., Martin, S., Vohra, Y. H., Ruoff, A. L. (1989), *Phys. Rev. Lett.* **62**, 669–672.

Ruoff, A. L., Xia, H., Luo, H., Vohra, Y. K. (1990), *Rev. Sci. Instrum.* **61**, 3830–3833.

Ruoff, A. L., Luo, H., Vohra, Y. K. (1991), *J. Appl. Phys.* **69**, 6413–6416.

Schirber, J. E. (1970), *Cryogenics* **10**, 418–422.

Schmidt, J. W., Nielsen, M., Daniels, W. B. (1984), *Phys. Rev. B* **30**, 6308–6319.

Seal, M. (1984), *High Temp.-High Pressures* **16**, 573–579.

So, P. T. C., Gruner, S. M., Shyamuunder, E. (1992), *Rev. Sci. Instrum.* **63**, 1763–1770.

Spain, I. L., Dunstan, D. J. (1989), *J. Phys. E Sci. Instrum.* **22**, 923–933.

Sterer, E., Pasternak, M. P., Taylor, R. D. (1990), *Rev. Sci. Instrum.* **61**, 1117–1119.

Takano, K. J., Wakatsuki, M. (1991), *Rev. Sci. Instrum.* **62**, 1576–1580.

van Straaten, J., Silvera, I. F. (1987), *Rev. Sci. Instrum.* **58**, 994–996.

Walker, D., Carpenter, M. A., Hitch, C. M. (1990), *Am. Minerologist* **75**, 1020–1028.

Webb, A. W., Gubser, D. U., Towle, L. C. (1976), *Rev. Sci. Instrum.* **47**, 59–62.

Weir, C. E., Lippincott, E. R., Van Valpenburg, A., Bunting, E. N. (1959), *J. Res. Nat. Bur. Stand.* **63A**, 55–62.

Zimmerman, A. M. (Ed.) (1970), *High Pressure Effects on Cellular Processes*, New York: Academic.

Further Reading

Bundy, F. P. (1988), *Phys. Rep.* **167**, 133–176. Provides an especially good treatment of contemporary development of "belt" and other "dry" pressure systems.

Gupta, Y. M. (Ed.) (1986), *Shock Waves in Condensed Matter*, New York: Plenum. A conference proceedings covering all aspects of physics at extremely high temperatures and pressures studied using dynamic techniques.

Jannasch, H. W., Marquis, R. E., Zimmerman, A. M. (1987), *Current Perspectives in High Pressure Biology*, New York: Academic. A thorough treatment of this relatively unexploited field.

Jayaraman, A. J. (1984), *Sci. Am.* **250** (4), 42–50. A more informal but informative treatment of diamond anvil cells and their application.

Jayaraman, A. J. (1986), *Rev. Sci. Instrum.* **57**, 1013–1031. A particularly comprehensive overview of the applications of diamond and anvil cells.

Proceedings of the AIRAPT International Conference on High Pressure Science and Technology. These appear at about two-year intervals. The most recent is A. K. Singh (Ed.) (1991), *Recent Trends in High Pressure Research*, New Delhi: Oxford & IBH.

Ulmer, G. C., Barnes, H. L. (Eds.) (1987), *Hydrothermal Experimental Techniques*, New York: Wiley. Covers the interaction of high-pressure studies and geophysics.

HOLOGRAPHY, ACOUSTICAL

A. J. DEVANEY, *Northeastern University, Boston, Massachusetts, U.S.A.*

Introduction 511
1. **Fundamentals of Holography** 512
1.1 Acoustic Wave Equation 512
1.2 Forming the Hologram 513
1.3 Wave-Front Reconstruction 514
1.4 Hologram Classification 516
2. **Implementation of Acoustic Holography** 518
2.1 Scanned Acoustical Holography 518
2.2 Liquid-Surface Acoustical Holography 519
3. **Holographic Image Formation** 520
3.1 Analog Image Formation 520
3.2 Computer Generation of Holographic Images 521
3.3 Holographic Image Quality 522
3.4 Ideal Holographic Images 523
4. **Acoustic Diffraction Tomography** 523
4.1 Generalized Projection-Slice Theorem 525
5. **Multifrequency Holography** 526
5.1 Synthetic-Aperture Holography 527
Acknowledgments 528
Glossary 528
Works Cited 529
Further Reading 530

INTRODUCTION

The term "holography" refers to the procedure of reconstructing the amplitude and phase of a wave field from the recorded interference pattern between the wave and a known reference wave. Although the discipline is most commonly associated with optical waves and, in particular, with radiation from visible lasers, the wave field need not be an optical wave and can, for example, be an acoustic wave or even a quantum-mechanical (Schrödinger) wave function. The field of holography owes its origins to Gabor (1949, 1951), who called the holographic process "wave-front reconstruction," a term that is actually more descriptive than its popularized replacement term "holography." This latter term grew out of Gabor's usage of the term "hologram" for the recorded interference pattern, a term he chose to indicate that the interference pattern encodes the "whole message" contained in the field being recorded in the hologram.

Gabor's original application of holography was in electron microscopy of semitransparent objects. His goal was to perform electron microscopy via a two-step procedure, where in the first step a hologram of the electron wave field transmitted by the object was recorded on a photographic plate. This hologram consisted of the interference pattern between the electron wave incident on the object being imaged and the wave scattered by the object. In the second step the developed photographic plate (the hologram) was illuminated by a well-filtered and collimated thermal light source to generate an optical wave that could be shown to be a magnified and slightly distorted version of the original scattered electron wave. The overall procedure thus effected a "wave-front reconstruction," and the reconstructed wave front could be directly viewed to reveal a magnified virtual image of the original object. The major advantage of the imaging process over conventional electron microscopy was that it employed no lenses and, hence, was largely free of image defects introduced into electron microscopy by unavoidable aberrations of electron lenses. This advantage of holography over conventional imaging via lenses has led to the name "lensless photography" for the process of holographic image formation.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

After the introduction of the laser in the 1960s the theory and applications of optical holography developed rapidly. One of these applications was in acoustics (Hildebrand and Brenden, 1972; Mueller, 1971), where the idea was somewhat analogous to Gabor's original idea:

1. make a hologram of an acoustic wave and
2. reconstruct the hologram with an optical wave.

What might appear somewhat odd about this idea is that Gabor was driven to holography by the fact that fundamental physics prevents one from directly measuring the phase of an electron wave, so that holography is *essential.* This is not the case in acoustics where both the amplitude and phase of an acoustic wave can, in principle, be directly measured. However, it must be remembered that although the measurements are possible in principle, they were, at the time of this early work, difficult in practice, and this fact provided the impetus for the work. Even with the advent of accurate acoustic detectors and detector arrays this original form of acoustic holography remains valuable and, indeed, forms the basis for one form of acoustic microscopy employed today (Way and Kessler, 1990).

With the advent of modern digital computers and microprocessors the discipline of holography rapidly expanded well beyond its original boundaries of being a procedure for wave-front reconstruction. Indeed, in today's usage the term "holography" encompasses the entire process of converting measured field data to an "image" of the source of the field. In this evolution the discipline, especially in acoustics and electromagnetics, has also shifted in emphasis away from its original form as an essentially experimental method for reconstructing wave fronts to an essentially computational method for reconstructing the three-dimensional structure of sources and scatterers from field data. As such the discipline of acoustical holography is today more closely associated with coherent imaging, inverse scattering, and tomography than it is with conventional optical holography.

1. FUNDAMENTALS OF HOLOGRAPHY

1.1 Acoustic Wave Equation

The wave fields employed in acoustical holography are narrow-band pressure or displacement waves that can be well approximated by purely monochromatic waves of the general form

$$u(\mathbf{r},t)=\Re\{U(\mathbf{r})e^{-i\omega t}\}, \tag{1}$$

where $\Re\{\cdot\}$ denotes the real part of the quantity in brackets, and ω is the acoustic frequency, t is the time, and $\mathbf{r}$ denotes position. The space-dependent amplitude U is a complex-valued scalar wave field that satisfies the inhomogeneous Helmholtz equation

$$(\nabla^2+k^2)U(\mathbf{r})=\rho(\mathbf{r}), \tag{2}$$

where $k=2\pi/\lambda$ is the wavenumber, λ is the wavelength, and ρ represents the space-dependent amplitude of a monochromatic acoustic source $q(\mathbf{r},t)=\Re\{\rho(\mathbf{r})e^{-i\omega t}\}$. The Helmholtz equation (2) assumes that the acoustic wave propagates in a homogeneous fluid of constant density and velocity and is purely monochromatic with frequency $\omega=ck$, where c is the acoustic velocity of the medium.

In certain applications discussed later, broad-band (pulsed) acoustic wave fields are employed rather than narrow-band fields of the form (1). The following discussion applies only to "traditional" holograms constructed with nearly monochromatic wavefields of the form (1).

The source term ρ in Eq. (2) generally represents a secondary (induced) source generated by the interaction of an incident (insonifying) wave with an object placed in the background medium. There are basically two types of objects of interest: surface scatterers and penetrable scatterers. Surface scatterers interact with the incident wave only on their surface, while penetrable scatterers allow the incident wave to penetrate into the interior of the object. A perfectly reflecting acoustic object is an example of a surface scatterer, while an acoustic object having different acoustic parameters from those of the background medium is an example of a penetrable scatterer.

Surface scatterers are represented by induced sources of the general form

$$\rho(\mathbf{r})=\tau(\mathbf{r})\delta(\xi(\mathbf{r})), \tag{3}$$

where τ is a quantity that represents the scattering strength at the point $\mathbf{r}$ and that depends, in general, on both the object and the incident wave, $\delta(\cdot)$ is the one-dimensional Dirac delta function, and $\xi(\mathbf{r})=0$ defines the surface of the object. Penetrable objects are defined via a *scattering potential* $V(\mathbf{r})$ via the equation

$$\rho(\mathbf{r})=V(\mathbf{r})U(\mathbf{r}), \tag{4}$$

where U is the acoustic wave field within the interior of the object. The scattering potential is related to the acoustic parameters of the object through the equation

$$V(\mathbf{r})=k^2[1-n^2(\mathbf{r})], \tag{5}$$

where k is the wave number of the background and $n(\mathbf{r})$ is the (generally complex) *index of refraction* of the object. The index of refraction can be further expressed in terms of fundamental acoustic parameters such as the bulk modulus and density of the object and for nonabsorbing objects is purely real and equal to the ratio of the background velocity c to the phase velocity of the object.

The induced-source representation of a penetrable acoustic object given in Eqs. (4) and (5) tacitly assumes that the object's density is constant and equal to that of the fluid background in which it is embedded. For variable-density objects the relationship is somewhat more complicated and, in particular, depends on the gradient of the object's density distribution and the gradient of the acoustic field U within the volume of the object (Devaney, 1983). We will restrict our attention to the simpler case of constant-density objects throughout this article since the incorporation of variable density does not affect the holographic process in any significant way (Devaney, 1983).

An important class of penetrable scatterers in acoustic holographic applications is composed of *semitransparent* objects. Such objects are characterized by an index of refraction profile $n(\mathbf{r})$ that is pointwise close to that of the background so that $|V(\mathbf{r})| \ll k^2$. The wave field $U(\mathbf{r})$ within the interior of such objects can be approximated by the incident wave $U_{\text{in}}(\mathbf{r})$ with the result that the induced source simplifies to

$$\rho(\mathbf{r})\approx V(\mathbf{r})U_{\text{in}}(\mathbf{r}). \tag{6}$$

The importance of approximation (6), known as the *Born approximation* (Cowley, 1984), stems from the fact that the induced source depends on the incident wave in a known and simple way. This allows imaging of the scattering potential associated with such sources to be performed in a simple and elegant manner (see Sec. 4).

1.2 Forming the Hologram

In its traditional form holography is a two-step procedure where in the first step the intensity (amplitude squared) of a wave field is recorded over some surface and in the second step a wave-front reconstruction is performed from the intensity data. The wave field being recorded consists of the sum of a known "reference wave" U_r and an "object wave" U_o associated with an induced source. The recorded intensity distribution (hologram) includes an interference pattern between the reference and object waves in which is encoded the phase of the object wave, and the goal of the holographic reconstruction procedure is to regenerate this phase and hence a reconstruction of the object wave from the recorded hologram.

The object wave U_o is usually generated by the scattering of an incident wave from a surface or penetrable scatterer and can be represented in the form

$$U_o(r)= \int d^3r'\rho(\mathbf{r}')G(\mathbf{r}-\mathbf{r}'), \tag{7}$$

where ρ is an induced source of the general form given in Eq. (3), (4), or (6), and $G(\mathbf{R})$ is the outgoing-wave Green's function for the Helmholtz equation. The incident wave is employed as the reference wave in Gabor holograms while off-axis holograms (see below) employ a separate wave for this purpose. The formation of a Gabor hologram is illustrated in Fig. 1, where a penetrable acoustic object is immersed in a liquid bath and insonified by a diverging spherical wave generated by a point transducer. In this case the transducer represents a primary source that generates both the wave incident on the object and the reference wave for the hologram (the incident and reference waves are the same). The total field is the sum of the spherical reference wave and an object wave (scattered

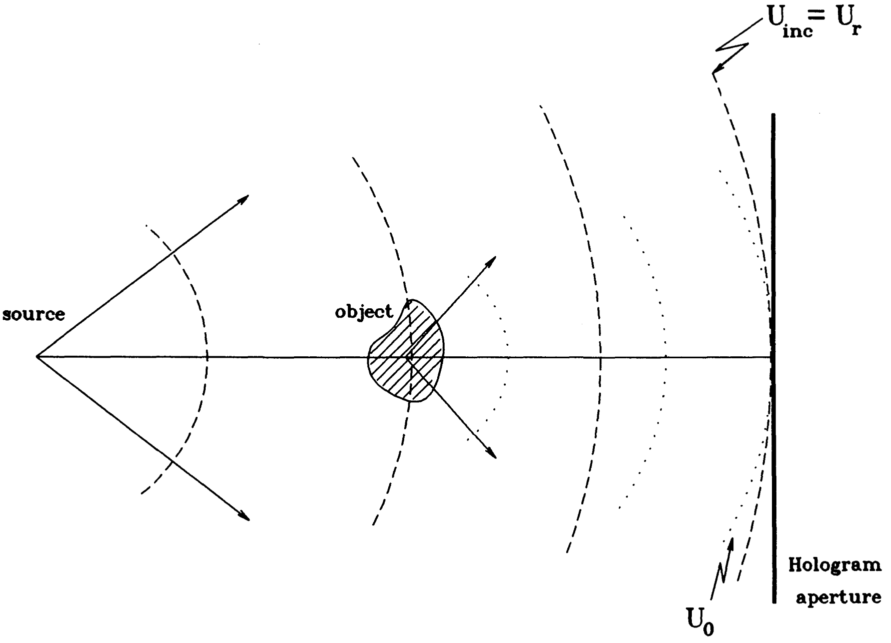

FIG. 1. Gabor hologram constructed using a spherical wave as both the wave incident on the object and the reference wave.

wave) generated by the scattering of the spherical wave from the object.

In its most basic form a hologram of the wave field $U = U_r + U_o$ is simply the intensity of this wave over some measurement surface and can be written in the form

$$\begin{aligned} I(\mathbf{r}_0) &= U^*(\mathbf{r}_0)U(\mathbf{r}_0) \\ &= I_r(\mathbf{r}_0) + I_o(\mathbf{r}_0) + U_r^*(\mathbf{r}_0)U_o(\mathbf{r}_0) \\ &\quad + U_r(\mathbf{r}_0)U_o^*(\mathbf{r}_0), \end{aligned} \tag{8}$$

where $\mathbf{r}_0$ denotes position over the measurement surface (e.g., hologram plane), the superscript asterisk * denotes the complex-conjugate operation, and I_r and I_o are the intensities of the reference and object waves, respectively. In writing Eq. (8) we have tacitly assumed that the wave field is perfectly spatially coherent over the hologram aperture. If this condition were violated, the interference terms as given by the last two terms in the equation would be modified and, in particular, would be replaced by the real part of the so-called *spatial coherence function* (Born and Wolf, 1980) between the reference and object wave fields. The narrow-band wave fields employed in acoustical holography are generated by sources having highly stabilized phases and are thus highly spatially and temporally coherent, so that Eq. (8) is a good approximation to the intensity distribution of an acoustical hologram.

The success of the holographic process depends on the presence and structure of the interference terms in the intensity distribution recorded in the hologram. In the best case the reference and object waves are perfectly coherent as assumed in Eq. (8), while in the worst case the two fields are perfectly incoherent resulting in no interference terms and hence no "hologram" and no possibility of wave-front reconstruction. Even if the field is perfectly coherent, the success of wave-front reconstruction is not guaranteed and depends on the relationship between the reference and object waves over the measurement surface.

1.3 Wave-Front Reconstruction

The second step in the holographic process is wave-front reconstruction. In Gabor's original work and in early acoustic and optical holography this was achieved in an analog fashion by physically illuminating the recorded hologram by an optical wave to generate the wave-front reconstruction. Consider,

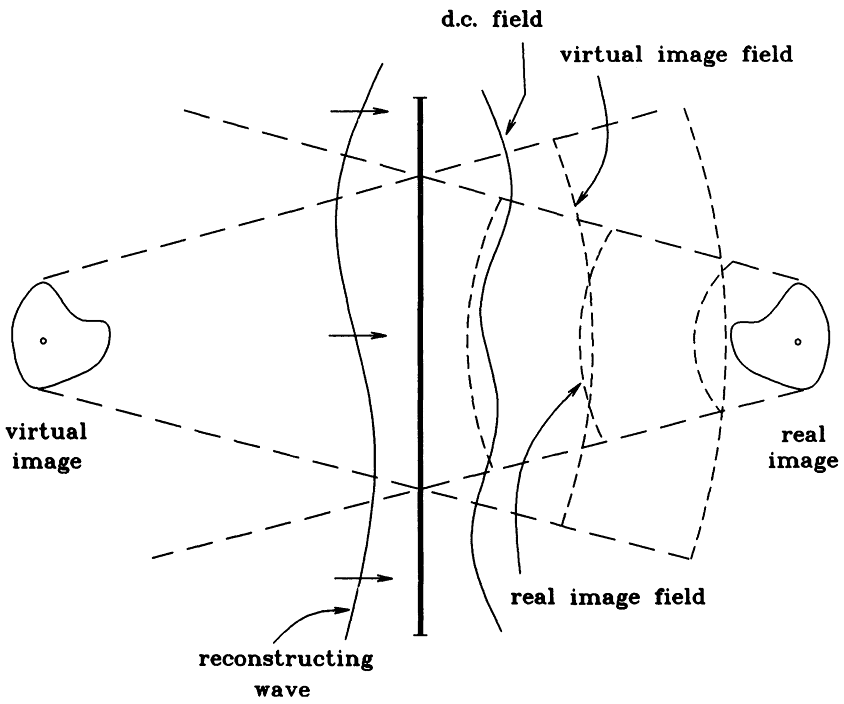

FIG. 2. Three components of the reconstructed wave field from a hologram.

for example, the situation illustrated in Fig. 2 where an optical wave is incident upon a planar transparency that has a transmittance function equal to the intensity distribution defined in Eq. (8). Within the scalar-wave Kirchoff approximation the wave field generated over the transparency plane by the interaction of the reconstructing (illuminating) wave with the hologram is equal to the product of the transmittance function and illuminating wave and is given by (Born and Wolf, 1980; Goodman, 1968)

$$\begin{aligned}\psi(\mathbf{r}_0) &= I(\mathbf{r}_0)U_r(\mathbf{r}_0)\\ &= [I_r(\mathbf{r}_0)+I_0(\mathbf{r}_0)]U_r(\mathbf{r}_0)+I_r(\mathbf{r}_0)U_o(\mathbf{r}_0)\\ &\quad + U_r^2(\mathbf{r}_0)U_o^*(\mathbf{r}_0),\end{aligned} \tag{9}$$

where we have selected the reconstructing wave to be the original reference wave and have, for the moment, neglected the difference between the acoustic and optical wavelengths.

The wave field ψ produced over the hologram surface is seen to consist of three terms:

1. the reconstructing wave spatially modulated by the sum of the intensities of the original object and reference waves,
2. the original object wave spatially modulated by the intensity of the reconstructing wave, and
3. the complex conjugate of the object wave spatially modulated by the square of the reconstructing wave.

Each of these components of the total wave ψ propagates away from the hologram plane in accordance with the wave equation and contributes to the total wave field observed, for example, by a viewer stationed to the right of the hologram. For virtually all cases of interest the intensities I_r and I_o are slowly varying functions of position over the hologram aperture while the phases of the reference and object waves can vary rapidly over this plane. It then follows that the character of the propagated components of the wave form are largely determined by the phases of these components on the hologram plane. The first term has a phase across the hologram equal to the phase of the reconstructing wave and thus would appear to a viewer stationed to the right of the hologram as a slightly distorted form of the reconstructing beam. The second term has a phase equal to the phase of the original object wave and thus would result in a wave that would appear to be very similar to the original object wave to a viewer. The third term is equal to the product of the square of the complex amplitude of the reference wave with the complex conjugate of the object wave. The complex-conjugate operation on the object wave has the effect of converting an outgoing (diverging) wave to an incoming (converging) wave. This third term then will result in a wave that focuses in the region to the right of the hologram. The three components of the reconstructed wave are illustrated in Fig. 2.

The three components of the reconstructed holographic field are termed the zero-order and first-order virtual and first-order real image fields, respectively. The success of the holographic process depends on the extent to which these components can be separated and, in particular, on the degree to which the first-order virtual image field can be isolated. Ideally, one would like the process to generate only the first-order virtual (or, in some applications, the first-order real) field. One limitation of Gabor's original (in-line) hologram is that these components achieve poor spatial separation so that a viewer looking into the reconstructed hologram would be distracted by the presence of the zero-order and first-order real image fields that would appear as background noise. The zero-order image field is sometimes referred to as the "dc" image field while the pair of first-order fields are sometimes termed the "twin" image fields.

The presence of the modulating factors associated with the reference and reconstructing wave fields in the reconstructed wave field over the hologram aperture introduce distortions into the first-order real and virtual image fields. These distortions manifest themselves as defects in the images generated by these fields and, in particular, cause the images to be stretched (or compressed) in directions orthogonal to the hologram plane (longitudinal magnification) and in directions lying in this plane (transverse magnification). The longitudinal and transverse magnification factors depend on the particular reference and reconstructing wave forms and on the ratio of the wavelengths of the two fields used in the holographic process. For the simplest case of plane reference and reconstructing waves, the transverse magnification is unity and the longitudinal magnification is constant and equal to $M_l = \pm\lambda_1/\lambda_2$, where λ_1 is the wavelength of the original wave field being recorded and λ_2 the wavelength of the reconstructing wave, and where the minus sign indicates that the image is formed in the space region occupied by the object (virtual image). The fact that the longitudinal magnification is negative for the real image means that this image is *pseudoscopic*, i.e., is mirror imaged relative to the hologram plane.

For the case of acoustic holography with optical reconstruction, the magnitude of the longitudinal magnification is on the order of 10^3 so that the real and virtual images experience considerable distortion in the reconstruction process. One way to minimize this distortion is to adjust the scale in the recorded hologram and thus artificially introduce a lateral magnification to balance the longitudinal magnification. Unfortunately, this process results in overall longitudinal and lateral magnifications having magnitudes on the order of 10^{-3} and hence in considerable demagnification and the consequent need for further optical imaging via a conventional imaging system (e.g., microscope). The use of direct optical reconstruction of acoustic holograms is thus of limited use in imaging applications.

1.4 Hologram Classification

Holograms are classified according to the manner in which they are recorded. There exist two major types of holograms termed, respectively, the Gabor or "in-line" hologram and the Leith-Upatnieks or "off-axis" hologram (Goodman, 1968; Leith and Upatnieks, 1962). Within each of these major categories the hologram is further classified to be a *Fresnel hologram* if it is recorded in the Fresnel region of the object or a *Fraunhofer hologram* if it is recorded in the Fraunhofer region or far field of the object. The Fraunhofer region of an object is roughly defined to be at distances $l \gg ka^2$, where $k=2\pi/\lambda$ is the wave number and a is the outer radius of the object (Goodman, 1968). Fraunhofer holograms are sometimes referred to as *Fourier-transform holograms* since the object wave in the Fraunhofer region (far field) is proportional to the spatial Fourier transform of this wave over a plane surface in the near field of the object (Goodman, 1968). Because of the relatively long wavelength of acoustic waves it is also possible to record acoustic holograms in the near field of an object (Williams and May-

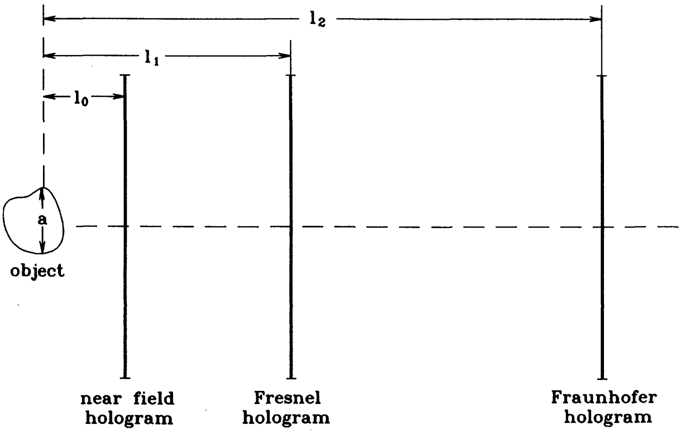

FIG. 3. A near-field hologram (left) is recorded at distances $l_0 \approx \lambda$ that are on the order of the acoustic wavelength, a Fresnel hologram (center) is recorded at intermediate distances $\lambda \ll l_1 \ll ka^2$, and a Fraunhofer hologram (right) is recorded in the far field $l_2 \gg ka^2$, where k is the wavenumber and a is the maximum transverse dimension of the object.

nard, 1980). Near-field holograms yield a reconstruction of the object field that includes evanescent plane-wave components of this field and can lead to "superresolution" in coherent imaging applications. We show in Fig. 3 geometries for the different types of holograms.

The Leith-Upatnieks off-axis hologram differs from the Gabor hologram in that the reference wave is specially selected to result in maximum separation between the three images generated in the reconstruction step of the holographic process. This is achieved by employing separate wave fields to illuminate the object and to act as the reference wave. Strictly speaking, the object wave in the Leith-Upatnieks hologram then consists of the sum of the wave illuminating the object plus the scattered wave. However, the object wave is usually generated in a backscattering configuration in off-axis holography so that the illuminating wave is effectively eliminated from the recorded hologram and the object wave then reduces to the scattered-field component defined in Eq. (7). The possibility of using a reference wave separate from the illuminating wave was not considered by Gabor because of the limited coherence of natural light sources and was realized only with the advent of the highly coherent optical laser. As mentioned earlier, narrow-band acoustic waves are by their very nature almost perfectly coherent so that generation of off-axis acoustic holograms poses no problem.

The freedom of selecting the reference wave in the off-axis hologram allows for the possibility of achieving good separation in the image fields generated in the reconstruction step of the holographic process. Consider, for example, an off-axis hologram constructed over the plane $z=0$ with a plane reference wave $\exp(ik\mathbf{s}_r \cdot \mathbf{r})$ propagating in the direction defined by the unit vector $\mathbf{s}_r$ and reconstructed with a plane wave $\exp(ikz)$ propagating normal to the surface of the hologram. The reconstructed wave field over the hologram surface is then given by

$$\psi(\mathbf{r}_0) = [I_r(\mathbf{r}_0) + I_o(\mathbf{r}_0)] + U_o(\mathbf{r}_0)e^{-ik\mathbf{s}_r \cdot \mathbf{r}_0} + U_o^*(\mathbf{r}_0)e^{ik\mathbf{s}_r \cdot \mathbf{r}_0}, \tag{10}$$

where $\mathbf{r}_0 = (x,y)$ denotes a point on the hologram surface (the plane $z=0$). The phases of the first-order real and virtual image fields are seen to include linear phase shifts of $k\mathbf{s}_r \cdot \mathbf{r}_0$ and $-k\mathbf{s}_r \cdot \mathbf{r}_0$, which cause these image fields to propagate as beams along the directions $\mathbf{s}_r$ and $-\mathbf{s}_r$, respectively, while the zero-order field propagates in the direction of the reconstructing wave. This then results in greater image separation than is possible with a Gabor hologram, where $\mathbf{s}_r = \mathbf{0}$ and the three image fields all propagate in more or less the same direction.

Good image-field separation will also occur if the off-axis hologram is reconstructed with either the original reference wave or the complex conjugate of this wave. In the first case a high-quality image of the virtual image field will be obtained while in the second case a high-quality image of the real image field will be obtained. Employing the reference wave or its complex conjugate as the reconstructing wave has certain advantages in optical holography due to the thickness of photographic emulsions (Goodman, 1968).

2. IMPLEMENTATION OF ACOUSTIC HOLOGRAPHY

2.1 Scanned Acoustical Holography

In this most basic form of acoustical holography the amplitude and phase of a spatially stationary acoustic wave field are recorded with a small (ideally a point) transducer that physically scans a measurement plane (Hildebrand and Brenden, 1972; Mueller, 1971). The recorded acoustic wave field is then converted into an intensity distribution on a photographic plate or film via a synchronously scanned light source as illustrated in Fig. 4. The reference wave in this type of hologram can be an actual acoustic primary wave that interferes with the acoustic object wave over the measurement surface (hologram surface) or can be electronically added at the output of the transducer prior to input into a square law detector (e.g., the photographic film). In either case an optical hologram is produced that has a transmittance function of the general form (8). The reconstruction is then performed using an optical reconstructing wave much in the same way that Gabor employed an optical wave to reconstruct holograms of electron wave fields.

An alternative form of scanned acoustical holography employs amplitude demodulation followed by low-pass filtering to convert the detected high-frequency, time-dependent, acoustic object wave to its spatial amplitude $U_o(\mathbf{r}_0)$ over the measurement plane. In this configuration no acoustic or electronic reference wave is used and the spatial amplitude of the object wave is directly stored on computer disk for later (digital) processing to generate a computer reconstruction of the acoustical wave field. In this form the use of the term "holography" might appear questionable since no reference wave is employed and the complex amplitude of the object wave is, in effect, directly measured. However, the process whereby the recorded object wave is "reconstructed" is mathematically the same as in holography with reference waves so that it is reasonable to consider this also to be a form of holography. An alternative and perhaps more accurate term for the process might be "synthetic aperture imaging" (cf. Sec. 5.1), although this latter term is usually reserved for reflection holography where the same transducer is employed both to transmit the illuminating wave and to detect the object (scattered) wave (Goodman, 1968).

A method closely related to scanned acoustical holography is fixed-transducer-array acoustical holography. In this method the hologram is not physically scanned with a single transducer but rather is recorded simultaneously with an array of transducer elements. The advantage of this method is speed of data acquisition, while its principal disadvantage is limited spatial resolution. In particular, the centers of the individual elements comprising the transducer array are physically separated by some minimum distance, and this distance then imposes a lower limit to the sample spacing employed in the recording of the hologram. Moreover, unless the separate array transducer elements are contiguous with each other, undersampling of the acoustic field being recorded can occur and result in aliasing being introduced in the reconstruction step.

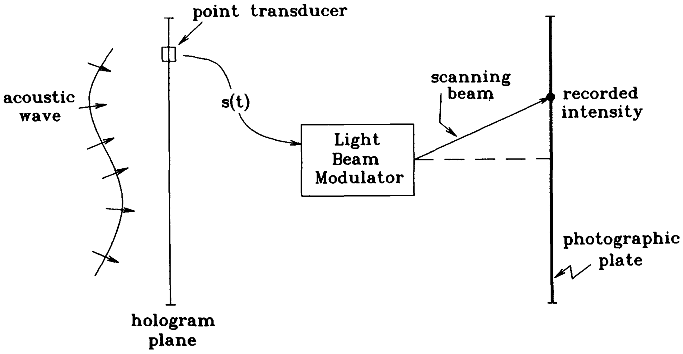

FIG. 4. Constructing an optical hologram using a light beam modulated by an acoustic signal. The resulting optical hologram is a scaled version of the acoustical hologram.

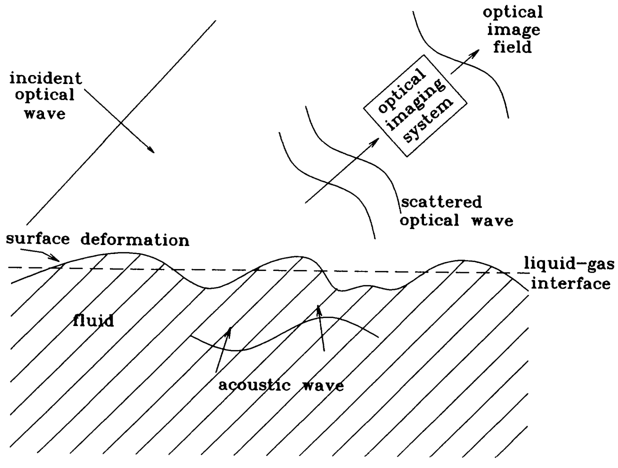
optical image field
incident optical wave
optical imaging system
scattered optical wave
surface deformation
liquid–gas interface
fluid
acoustic wave

FIG. 5. System for optical reconstruction of an acoustical hologram encoded in the surface deformation induced at a liquid–gas interface. The optical system generates an image field that is directly proportional to the acoustical intensity distribution over the interface.

2.2 Liquid-Surface Acoustical Holography

This form of acoustical holography makes use of the fact that an acoustic wave field propagating in a liquid bath induces a physical deformation at the liquid surface whose height is proportional to the intensity of the acoustic field. This deformation then acts as an optical phase grating to an incident optical wave and can thus be imaged using a specialized optical imaging system as illustrated in Fig. 5. The optical system converts the phase perturbation induced on the incident optical plane wave to an optical image field whose intensity distribution over an exit pupil is proportional to this phase. The optical system can be based on the central dark ground method of observation, the Schlieren method, or Zernike's phase contrast method (Born and Wolf, 1980). By selecting the acoustical wave to be the sum of a reference wave and object wave, the induced surface deformation is made proportional to the intensity pattern of an acoustical hologram taken over the liquid surface so that the optical field generated by the system has this same intensity profile and is thus an optical reconstruction of the (acoustic) hologram.

An interesting variation of liquid-surface acoustical holography is the scanning laser acoustic microscope (SLAM) (see MICROSCOPY, ACOUSTICAL). In its original form (Kessler and Yuhas, 1979) this microscope simply produced an optical image of the acoustic intensity distribution generated over a liquid–surface interface via a system such as that illustrated in Fig. 6. However, the device has recently been modified so as to generate holographic reconstructions by modulating the output of the photodiode with an electronic reference signal (Wey and Kessler, 1990). By

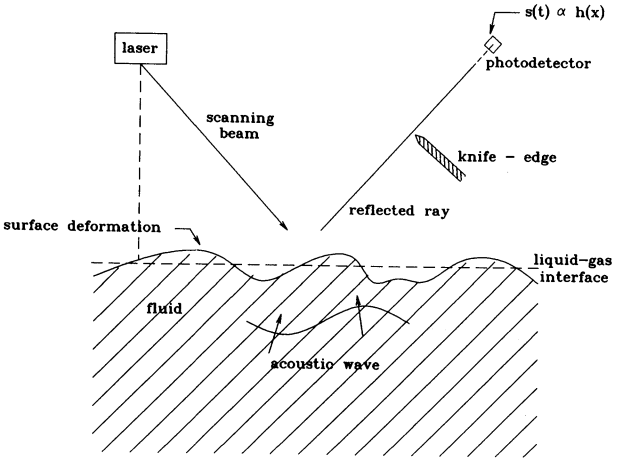
s(t) α h(x)
laser
photodetector
scanning beam
knife – edge
reflected ray
surface deformation
liquid–gas interface
fluid
acoustic wave

FIG. 6. Scanning laser acoustic microscope (SLAM). A scanning laser beam is modulated by the liquid surface perturbation and the scattered optical field is converted into an intensity proportional to the surface height by means of a knife edge.

means of this procedure it is possible to determine approximately the complex amplitude of the object wave over the liquid–surface interface; this is then used to generate an image of the object on a digital computer.

3. HOLOGRAPHIC IMAGE FORMATION

Images of the source ρ can be generated from holograms produced in any of the ways discussed above and of any of the types discussed earlier. Moreover, a high-quality virtual or real image of the source can be generated depending on whether the reconstructing wave is taken to be the original reference wave (virtual image) or the complex conjugate of this wave (real image). Of course in either case both real and virtual images as well as the "dc" image field will be present, but by appropriate selection of the reference and reconstructing waves it is possible to minimize the distortion in the desired (high-quality) image caused by the presence of these other image fields. In cases where the complex amplitude of the acoustic object wave is directly measured (e.g., by a scanned acoustical transducer or fixed transducer array) the unwanted image components will not be generated in the holographic reconstruction.

The acoustical image is formed by the conversion of the holographic boundary value field ψ as defined in Eq. (9) into an image field in the reconstruction step. The image field can be physically formed in an analog system or can be generated on a digital computer by appropriately processing the recorded hologram. In the case where the image is formed digitally there is of course more freedom in the image formation step than in the analog case where the image formation is determined entirely by the physics of wave propagation. However, even for the case of computer processing the image is largely determined by the physics of the problem and closely mimics what one would obtain via analog processing.

3.1 Analog Image Formation

To understand the image formation process in holography it is worth while to neglect, for the moment, the zero-order and first-order real image fields and to consider the spatial evolution of the object field U_o from its boundary value generated over the hologram surface by physical illumination of the hologram with the original reference wave. If we assume that the reference wave was a unit-amplitude plane wave, then this evolution is approximately described by the Rayleigh-Sommerfeld integral (Goodman, 1968; Stamnes, 1986)

$$U_o(\mathbf{r}) = \int_{\Sigma_0} d^2r_0\, \gamma(\mathbf{r}-\mathbf{r}_0) U_o(\mathbf{r}_0). \tag{11}$$

In this equation γ is a "propagator" that can be expressed in the form

$$\gamma(\mathbf{r}-\mathbf{r}_0) = \frac{\partial}{\partial n_0} G_d(\mathbf{r}-\mathbf{r}_0), \tag{12}$$

where G_d is the outgoing-wave Green's function for the Helmholtz equation that vanishes on the hologram surface and where $\partial/\partial n_0$ stands for the outward directed normal derivative to the hologram surface at the point $\mathbf{r}_0$. For a planar hologram this Green's function is well known to be twice the free-space Green's function G employed in Eq. (7) (Goodman, 1968; Stamnes, 1986). The integral is performed over the hologram surface, which we have denoted by Σ_0. We can express the virtual image field defined in Eq. (11) in terms of the source ρ by substituting for the boundary-value field $U_o(\mathbf{r}_0)$ from Eq. (7). We obtain

$$U_o(\mathbf{r}) = \int d^3r' h(\mathbf{r},\mathbf{r}') \rho(\mathbf{r}'), \tag{13}$$

where

$$h(\mathbf{r},\mathbf{r}') = \int_{\Sigma_0} d^2r_0\, \gamma(\mathbf{r}-\mathbf{r}_0) G(\mathbf{r}_0-\mathbf{r}'). \tag{14}$$

Equations (11) and (13) completely describe the formation of the virtual image in the reconstruction step. Equation (11) describes the evolution of the virtual image field from its boundary value over the hologram surface while Eq. (13) can be regarded to be an "imaging equation" that relates the virtual image field to the source ρ. With this interpretation the image kernel $h(\mathbf{r},\mathbf{r}')$ is thus identified as the *point-spread function* (PSF) of the imaging process. The virtual image will, of course, be concentrated in the region of space lying to the rear of the hologram and hence is not directly accessible. It can, however, be readily converted to a real image concentrated on the observer side of the hologram by a further image-forming step. Such a step is implemented, for example, by the human eye in viewing the reconstructed wave field generated in optical holographic reconstruction.

Equation (14) defining the PSF h is the well known Rayleigh-Sommerfeld integral representation (Born and Wolf, 1980; Goodman, 1968; Stamnes, 1986) of an outgoing spherical wave centered at the source point $\mathbf{r}'$ and diffracted by the hologram aperture. The PSF can thus be interpreted as being the (virtual) image of a point source located at $\mathbf{r}'$ and viewed through an aperture located over the hologram surface. The image formed by the source as given in Eq. (13) is then a coherent summation of such images weighted by the amplitude of the source $\rho(\mathbf{r}')$ at each source point. Because the PSF is not, in general, a function of the difference of the image and source vectors $\mathbf{r}$ and $\mathbf{r}'$, respectively, the image formation is *nonisoplanatic* (Born and Wolf, 1980; Goodman, 1968).

A real image is also formed in the holographic reconstruction step by the evolution of the first-order real image field. This evolution is described by Eq. (11), where, however, the boundary value of the first-order real image field is employed in place of U_o. When the reconstructing wave is taken to be the original reference wave, the image so formed will be corrupted by the presence of the square of the reference wave U_r^2 in the boundary value. For plane-wave illumination the phase of this quantity will cause the image location to shift laterally and will also introduce unwanted distortion in the image. To avoid this problem a reconstructing wave equal to the complex conjugate of the reference wave can be employed so that the boundary-value field ψ now includes a term equal to $I_r U_o^*$ that is not corrupted by the phase of the reference wave. For the usual case of plane reference waves this corresponds physically to illuminating the hologram from the opposite side from which it was recorded. However, as noted earlier, the image so generated will be pseudoscopic and will be an image of the complex conjugate of the source rather than of the source itself.

3.2 Computer Generation of Holographic Images

The image field $U_o(\mathbf{r})$ defined in Eq. (11) can be physically generated by actual illumination of the hologram by an acoustic (or optical) wave or can be mathematically computed by multiplying the recorded intensity pattern (hologram) $I(\mathbf{r}_0)$ by the reference wave and performing the integration (11) on a digital computer. Moreover, in the case of computer processing we are not limited to using the normal derivative of the Dirichlet Green's function G_d as the propagator in the image formation step, and we are free to choose this quantity to optimize image quality. We are thus led to a general image-formation procedure defined by an integral transform of the same general form as Eq. (11) but with an arbitrarily chosen propagator γ. The image formed in a computer reconstruction using a given propagator is related to the source by Eq. (13) with the PSF defined in Eq. (14).

The use of a computer to form the virtual-image field has certain advantages over analog image formation. For example, in cases where the complex amplitude of the object wave is directly recorded with either a scanned transducer or fixed transducer array, the virtual image field computed from Eq. (11) will not include the zero-order or first-order real image fields. In addition, it is possible to form a real image directly by *backpropagating* the virtual image field into the image space located *behind* the hologram plane. The backpropagation process can be thought of as an analytic continuation of the virtual image field from its value on the hologram surface *back* into the space region occupied by the source. This backward evolution can be achieved by the simple expedient of evaluating Eq. (11) directly for observation points $\mathbf{r}$ lying in object space (i.e., in the space lying to the rear of the hologram surface). However, it is important that the backpropagation be computationally stable, and so some care must be exercised in choice of the propagator γ. In particular, in the case where the Dirichlet Green's function is employed it is customary to exclude evanescent plane waves in the backpropagation computation. An alternative propagator that accomplishes the same purpose is given by the complex conjugate of the Dirichlet Green's function.

3.3 Holographic Image Quality

As discussed earlier, Eq. (14) defining the holographic PSF $h(\mathbf{r},\mathbf{r}')$ is the Rayleigh-Sommerfeld integral representation of an outgoing spherical wave diffracted by the hologram aperture (Goodman, 1968). This quantity can be approximated by the so-called Debye integral (Born and Wolf, 1980; Stamnes, 1986)

$$h(\mathbf{r},\mathbf{r}')\approx\frac{1}{4\pi}\int_{\delta\Omega} d\Omega\; e^{ik\mathbf{s}\cdot(\mathbf{r}-\mathbf{r}')}, \qquad (15)$$

where $\delta\Omega$ denotes the solid angle subtended by the hologram aperture when viewed from the point $\mathbf{r}'$ and the integration is over all unit vectors $\mathbf{s}$ within this solid angle. A further approximation is possible if the solid angle $\delta\Omega$ subtended by the hologram aperture is approximately constant for all source points $\mathbf{r}'$, as will occur, for example, if the hologram aperture is large compared to the size of the object. Within this approximation the PSF will only be a function of the difference $\mathbf{r}-\mathbf{r}'$ and the imaging will be *isoplanatic*. We will assume that this condition holds throughout the following discussion.

The quality of holographic image formation can be examined for various types of sources. Consider first a transparency $t(x,y)$ centered at $x=y=0$ on the plane $z=0$ and illuminated by a unit-amplitude plane wave e^{ikz} propagating along the positive z axis. In this case the induced source is given by Eq. (3) with the scattering strength $\tau=t(x,y)$ and $\xi(\mathbf{r})=z$. If the virtual image field generated by Eq. (11) is evaluated over the "image plane" $z=0$ we obtain an image $\hat{t}(x,y)$ of the transparency function given by

$$\hat{t}(x,y)=\int dx'dy'H(x-x',y-y')t(x',y'), \qquad (16)$$

where the image kernel H is equal to the holographic PSF h evaluated at $z=z'=0$.

The imaging equation (16) is most easily examined in the spatial frequency domain. In particular, taking the spatial Fourier transform of both sides of the equation we obtain

$$\tilde{\hat{t}}(K_x,K_y)=\tilde{H}(K_x,K_y)\tilde{t}(K_x,K_y), \qquad (17)$$

where the tilde denotes the Fourier transform. The spatial Fourier transform of the image kernel H is obtained from Eq. (15) upon setting $z=z'=0$ and converting from angular integration variables to the Cartesian spatial frequency variables K_x and K_y. If we assume that the holographic aperture is circular and centered at $x=y=0$ we obtain

$$\tilde{H}(K_x,K_y)=\begin{cases}\pi/k\gamma & \text{if } K_x^2+K_y^2\leqslant k^2\sin^2\theta,\\ 0 & \text{otherwise,}\end{cases}$$

where $\gamma=\sqrt{k^2-K_x^2-K_y^2}$ and where θ is the half-angular extent of the holographic aperture as viewed from the center of the transparency. The image (16) is seen to be a low-pass filtered version of the transparency function bandlimited to spatial frequencies lying within the circle $K_x^2+K_y^2\leqslant k^2\sin^2\theta$ and consequently has a spatial resolution of roughly $\lambda/(2\sin\theta)$.

A generalized version of the source considered above is generated by the scattering of an acoustic wave from a smooth object whose surface curvature is small on the scale of the acoustic wavelength. Such an object is a surface scatterer as defined in Sec. 1.2 and is characterized by an induced source of the general form given in Eq. (3). It is readily shown that the holographic image of this source is peaked at the surface of the object and yields a high-quality image of the scattering strength $\tau(\mathbf{r})$ as long as the acoustic wavelength λ is small compared with the scale over which τ and the surface profile vary.

Both of the sources considered above are two dimensional in that they are concentrated in a two-dimensional region of space (a plane in the first case and a general smooth surface in the second case). It is not surprising, then, that the images of these sources are of high quality since they are generated from the source's radiated fields recorded over the two-dimensional surface Σ_0. The recorded data (hologram) has two degrees of freedom, and hence can be expected to yield high-quality images of objects having this same number of degrees of freedom. The situation is considerably different for three-dimensional sources such as result from the scattering of an incident wave from a penetrable scatterer. Such sources have three degrees of freedom, and it is unreasonable to expect a holographic image of such a source to reveal a great deal of information concerning the source's three-dimensional spatial structure (Toraldo di Francia, 1969). We examine the quality of images of three-dimensional sources in the following section.

3.4 Ideal Holographic Images

In order to study the quality of holographic images of three-dimensional sources it is useful to consider an "ideal" holographic configuration (Porter and Schwab, 1971) that yields the best image theoretically possible. Such a configuration is obtained by letting the holographic surface Σ_0 completely surround the source ρ. In this case the PSF defined in Eq. (15) can be evaluated in closed form and we obtain

$$h(\mathbf{R})=\frac{\sin(kR)}{kR}=\operatorname{sinc}(kR), \tag{18}$$

where $R=|\mathbf{r}-\mathbf{r}_0|$. The PSF of the ideal holographic configuration is seen to be isoplanatic and spherically symmetric.

The lack of information content in the holographic image of a three-dimensional source is readily apparent from examining the three-dimensional spatial Fourier transform of the ideal image of such a source. The ideal image is given by Eq. (13) with the PSF given by Eq. (18),

$$\hat{\rho}(\mathbf{r})=\int d^3r' \operatorname{sinc}(k|\mathbf{r}-\mathbf{r}'|)\rho(\mathbf{r}'), \tag{19}$$

where we have denoted the ideal holographic image by $\hat{\rho}$. The three-dimensional spatial Fourier transform of the ideal image is found to be

$$\begin{aligned}\tilde{\hat{\rho}}(\mathbf{K})&=\int d^3r\,\hat{\rho}(\mathbf{r})e^{-i\mathbf{K}\cdot\mathbf{r}}\\&=(2\pi^2/k^2)\tilde{\rho}(\mathbf{K})\delta(K-k),\end{aligned} \tag{20}$$

where $\delta(\cdot)$ is the one-dimensional Dirac delta function.

We conclude from Eq. (20) that the ideal holographic image has a spatial Fourier transform that is concentrated on the surface of a sphere centered at the origin and having a radius equal to the wave number k. This image has only two degrees of freedom, corresponding to the two coordinates defining the surface of this sphere. Generation of high-quality images of such sources thus requires that new and independent data concerning the object being imaged be employed in the image formation process. We discuss two methods for generating such additional information in the following sections.

4. ACOUSTIC DIFFRACTION TOMOGRAPHY

We concluded in the preceding section that the virtual image formed of a three-dimensional source in acoustic holography contains limited structural information about the source. One method for constructing an image containing more information is to employ many holograms and to combine coherently the images generated from the separate holograms. However, for this procedure to succeed the separate holograms must contain independent or, at least, different information about the source. This then rules out the use of such a procedure with primary sources since an entire set of holograms of such a source contains no more information than that encoded in a single "ideal hologram" whose surface Σ_0 completely surrounds the source.

Although the use of multiple holograms yields no increase in image quality for primary sources, the procedure is applicable to induced sources that represent penetrable scattering objects and where the goal is to obtain a high-quality image of the scattering potential $V(\mathbf{r})$. For this class of objects it is possible to perform a suite of scattering experiments using different incident waves and to superimpose the holographic images generated in the set of experiments. Since the incident wave changes from experiment to experiment, each holographic image contains, in general, new information regarding the scattering potential, and an enhanced image of this quantity can thus be generated by properly combining the images produced by the separate holograms.

The procedure outlined above for generating high-quality images of the scattering potential of penetrable scatterers has become known by the name *diffraction tomography* (see ACOUSTICAL TOMOGRAPHY). This name is motivated by the fact that the data employed in generating the image arise from the diffraction (scattering) of incident waves by the object and that the coherent combination of the multiple holographic images is analogous to the combination of images used in computed tomography (CT) (Natterer, 1986) to generate a high-quality image of the interior of an object from multiple x-ray micrographs of the object. In diffraction tomography (DT) the holograms play the role of the micrographs and the holographic image formation process as encoded in Eq. (11) is the DT version of the image formation process in CT.

In order to understand the imaging procedure employed in DT, consider the geometry illustrated in Fig. 7 where a semitransparent acoustic object is insonified by a plane wave $\exp(ik\mathbf{s}_0\cdot\mathbf{r})$ and the scattered wave is measured over a holographic surface Σ_0 that is normal to the direction of propagation $\mathbf{s}_0$ of the insonifying plane wave and is large compared with the object's spatial extent. In most DT applications the wave scattered from the object will be highly concentrated within some small solid angle $\delta\Omega$ about the direction of propagation of the insonifying plane wave so that effectively an ideal hologram is formed as long as the recording surface Σ_0 is large enough to record the wave field within this solid angle. We will assume that this condition is satisfied in the following discussion.

The holographic image of the scattering potential is obtained by backpropagating the object wave (scattered wave) according to Eq. (11) and can be expressed, using Eq. (19), in the form

$$\hat{V}(\mathbf{r};\mathbf{s}_0) = \int d^3r'\, V(\mathbf{r}')e^{-ik\mathbf{s}_0\cdot(\mathbf{r}-\mathbf{r}')} \times \operatorname{sinc}(k|\mathbf{r}-\mathbf{r}'|), \tag{21}$$

where we have made use of Eq. (6) and where $\mathbf{s}_0$ has been included in the argument of the image to indicate its dependence on the propagation direction of the insonifying plane wave. If now multiple experiments employing insonifying plane waves having different directions of propagation are performed and the partial holographic images are coherently averaged over 4π steradians, we obtain the image

$$\hat{V}(\mathbf{r}) = \frac{1}{4\pi}\sum_{\mathbf{s}_0}\hat{V}(\mathbf{r};\mathbf{s}_0) = \int d^3r'\, H(\mathbf{r}-\mathbf{r}')\, V(\mathbf{r}'), \tag{22}$$

where

$$H(\mathbf{R}) = \operatorname{sinc}(k|\mathbf{R}|)\,\frac{1}{4\pi}\sum_{\mathbf{s}_0} e^{-ik\mathbf{s}_0\cdot\mathbf{R}}. \tag{23}$$

The quality of the image $\hat{V}(\mathbf{r})$ generated from the set of holograms increases monotonically with the number of experiments. In the ideal case where an infinite number of experiments are employed using insonifying unit vectors $\mathbf{s}_0$ completely covering the unit sphere, the summation in Eq. (23) becomes an integral over this sphere and we obtain

$$H(\mathbf{R}) = \operatorname{sinc}^2(k|\mathbf{R}|). \tag{24}$$

The image generated in this case is much superior to the image generated in a single or small number of experiments and is approximately equal to the image generated in a large number of experiments employing unit vectors equally spaced over the unit sphere. However, a further improvement is possible by post- (computer) processing so as to deconvolve the image kernel H from the image and thus arrive at a truly optimum image of the scattering potential. Alternatively, because of

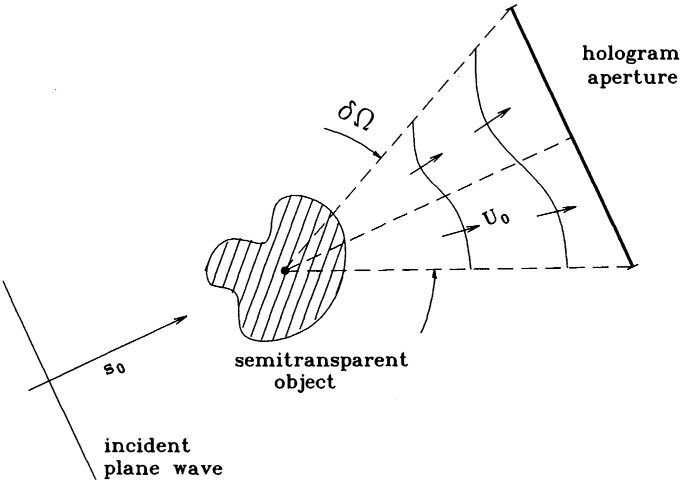

FIG. 7. Recording an effective ideal hologram in DT. In practice the scattered field will be highly concentrated within a solid angle $\delta\Omega$ centered on $\mathbf{s}_0$ so that the hologram surface Σ_0 need only extend over a limited region.

the linearity of the image formation process this postprocessing can be avoided entirely by *prespatial filtering* of the virtual image field on the measurement surface Σ_0 prior to backpropagation. When this is done the resulting image kernel is a three-dimensional δ function bandlimited to the interior of the so-called Ewald limiting sphere $|\mathbf{K}| \leqslant 2k$ familiar from x-ray crystallography (Cowley, 1984). In this case, then, an image of the scattering potential bandlimited to within this same sphere is obtained, corresponding roughly to a spatial resolution of a quarter-wavelength (Mueller *et al.* 1979; Wolf, 1969).

The image-formation process outlined above is seen to consist of three distinct steps:

1. spatial filtering of the holographic data (the scattered wave on the measurement surface Σ_0),
2. backpropagation of the filtered data according to Eq. (11), and
3. summing the partial images so obtained over the insonifying directions (view angles) $\mathbf{s}_0$.

This procedure has been called the filtered backpropagation algorithm (Devaney, 1983) and can be shown to be a generalization of the filtered *backprojection* algorithm (Natterer, 1986) of computed tomography (CT) to diffraction tomography. It is important to note that the spatial filters employed in the algorithm are independent of the object being imaged and only depend on the number and geometry of the holograms employed to record the scattered field data. The operation of the algorithm is illustrated in Fig. 8.

Diffraction tomography has been employed in a number of applications. Perhaps the first of these in acoustic tomography was studied by Greenleaf and co-workers (Greenleaf *et al.*, 1974) who used a simplified (nondiffracting) ray model similar to that employed in x-ray tomography. Later, Mueller and co-workers (Mueller *et al.*, 1979) replaced the ray model by the more accurate wave model that underlies diffraction tomography. Since this early work the theory and application of acoustic (ultrasound) tomography has progressed considerably (Greenleaf and Bahn, 1981), and at least one system is being readied for commercialization (Stamnes *et al.*, 1992). Acoustic diffraction tomography is also employed in geophysical tomographic applications (Devaney, 1984; Witten and Long, 1986) and also forms the basis for newer applications of the scanning laser acoustic microscope (Wey and Kessler, 1990).

4.1 Generalized Projection-Slice Theorem

Consider the partial image $\hat{V}(\mathbf{r};\mathbf{s}_0)$ defined in Eq. (21) of a semitransparent scattering potential formed from a hologram at some wave number k. A straightforward calculation shows that the three-dimensional spatial Fou-

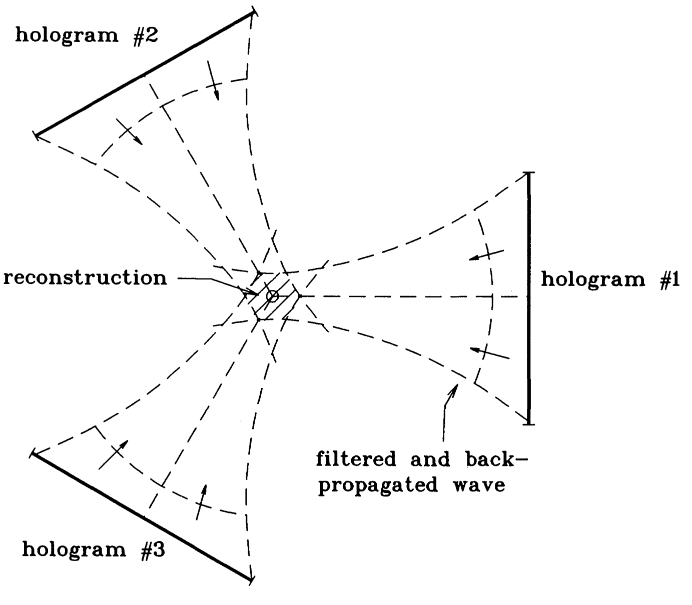

FIG. 8. Graphical illustration of the operation of the filtered backpropagation algorithm of diffraction tomography. The scattered wave recorded on each hologram is spatially filtered and backpropagated. The reconstruction is equal to the coherent sum of filtered and backpropagated images from the entire set of holograms.

rier transform of this image is confined to the surface of a sphere that is centered at the spatial frequency $\mathbf{K} = -k\mathbf{s}_0$ and has a radius equal to the wave number k. This sphere is the Ewald sphere well known from x-ray crystallography (Cowley, 1984), and the relationship between the spatial Fourier transform of the ideal holographic image and the scattering potential is a statement of the generalized projection-slice theorem of diffraction tomography (Mueller *et al.*, 1979).

In accordance with earlier discussion a single partial image $\hat{V}(\mathbf{r};\mathbf{s}_0)$ has only two degrees of freedom, corresponding to the two coordinates defining the surface of the Ewald sphere, and in order to obtain a high-quality image of the scattering potential it is necessary to "fill out" Fourier space by incorporation of additional data. In DT the filling-out process is achieved through the use of multiple Ewald spheres having different centers (different $\mathbf{s}_0$). In particular, the partial images generated at different propagation directions $\mathbf{s}_0$ of the incident plane wave define the Fourier structure of the object on different Ewald spheres, and by summing these images a reconstruction is generated that defines the structure over the union of Ewald spheres. In the ideal case where $\mathbf{s}_0$ completely covers the unit sphere, the object's structure is determined throughout the Ewald limiting sphere as illustrated in Fig. 9 and thus gives rise to a resolution of roughly a quarter-wavelength (Wolf, 1969). However, it should be pointed out that in practice only the forward-scattered field is recorded in DT (cf. Fig. 7), and this limits the ultimate resolution to be somewhat worse than this (Devaney, 1983; Mueller *et al.*, 1979).

5. MULTIFREQUENCY HOLOGRAPHY

The key to the success of diffraction tomography is the incorporation of data from multiple-scattering experiments into the image-formation process. In the case of DT each scattering experiment employs a narrow-band insonifying plane wave, and these "additional data" are generated by varying the direction of propagation of this plane wave. Another way to generate additional data is to employ a set of insonifying wave fields having different temporal frequencies ω or, equivalently, a single *broad-band* insonifying wave, in which case the "additional data" are generated by the various temporal frequencies ω within the band pass of the insonifying wavefield (Langenberg, 1987). However, for either of these latter procedures to be effective it is necessary that the object being imaged be nondispersive (i.e., frequency independent) or, at least, have a scattering potential that depends on temporal frequency in a known way.

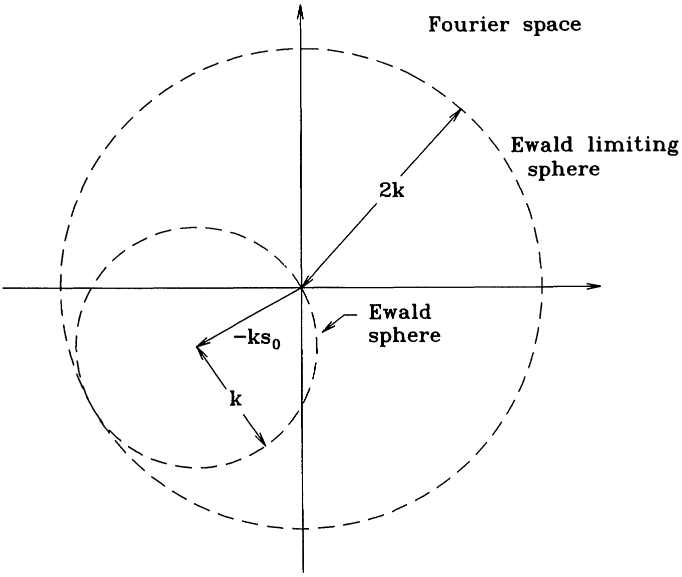

FIG. 9. The projection-slice theorem of DT. Each ideal hologram defines the spatial Fourier transform of the scattering potential *V*(**r**) over an Ewald sphere.

To examine the use of multiple frequencies in holographic image formation, consider the partial image $\hat{V}(\mathbf{r};\mathbf{s}_0)$ defined in Eq. (21) of a *nondispersive* semitransparent scattering potential formed from an ideal hologram at a single temporal frequency ω. As discussed above, the three-dimensional spatial Fourier transform of this image is confined to the surface of an Ewald sphere that is centered at the spatial frequency $\mathbf{K}=-k\mathbf{s}_0$ and has a radius equal to the wave number $k=\omega/c$, where c is the acoustic velocity in the background medium. A high-quality image of the object is obtained in DT by "filling out" Fourier space through the use of multiple Ewald spheres having fixed radii k but different centers (different $\mathbf{s}_0$). Alternatively, it is possible to employ multiple Ewald spheres whose centers are all aligned in the same direction $\mathbf{s}_0$ but which have different radii (different k). This then generates a Fourier coverage such as that illustrated in Fig. 10. A higher-quality reconstruction than obtainable from a single-frequency hologram is then possible by properly making use of these "additional data."

A multifrequency reconstruction can be generated in a manner completely parallel to the procedure employed in DT. In particular, the partial images $\hat{V}$ formed via Eq. (21) at each temporal frequency or wave number k can be coherently added to generate an image that is related to the scattering potential via Eq. (22) where the image kernel H is now given by

$$H(R)=\sum_k \operatorname{sinc}(k|\mathbf{R}|)e^{-ik\mathbf{s}_0\cdot\mathbf{R}}. \tag{25}$$

In this equation the summation is performed over the wave numbers employed in the set of experiments or in the passband of the acoustic insonifying plane wave pulse. As in the case of DT the multifrequency reconstruction can be further improved by deconvolving the image kernel (25) from the image $\hat{V}$. Alternatively, it is possible to prefilter the virtual image fields at each temporal frequency prior to the backpropagation step to obtain the same end result.

5.1 Synthetic-Aperture Holography

In synthetic-aperture holography a single point transducer located in the far field of the object is employed both to insonify the object and to record the scattered field as illustrated in Fig. 11. The measured object field is then the field *backscattered* by the object and, for semitransparent objects, is given by [cf. Eq. (7)]

$$\begin{aligned}U_o(\mathbf{r}_0)&=\int d^3r'\,V(\mathbf{r}')G(\mathbf{r}'-\mathbf{r}_0)G(\mathbf{r}_0-\mathbf{r}')\\&\sim\left[\left(\frac{1}{4\pi}\right)^2\int d^3r'\,V(\mathbf{r}')e^{-i2k\mathbf{s}_0\cdot\mathbf{r}'}\right]\\&\quad\times\frac{e^{i2kr_0}}{r_0^2},\end{aligned} \tag{26}$$

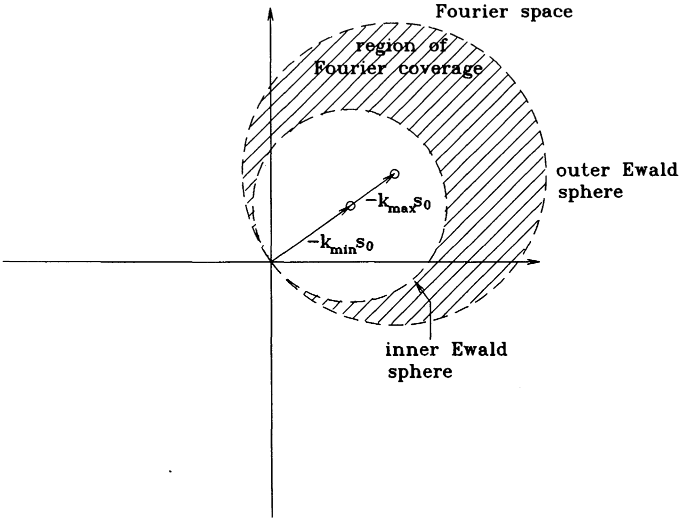

FIG. 10. Fourier coverage in multifrequency holography. The spatial Fourier transform of a nondispersive scattering potential is determined over the region bounded by an "outer" Ewald sphere of radius $k_{\max}$ and an "inner" Ewald sphere of radius $k_{\min}$, where $k_{\max}$ and $k_{\min}$ are the maximum and minimum wave numbers employed in the frequency range.

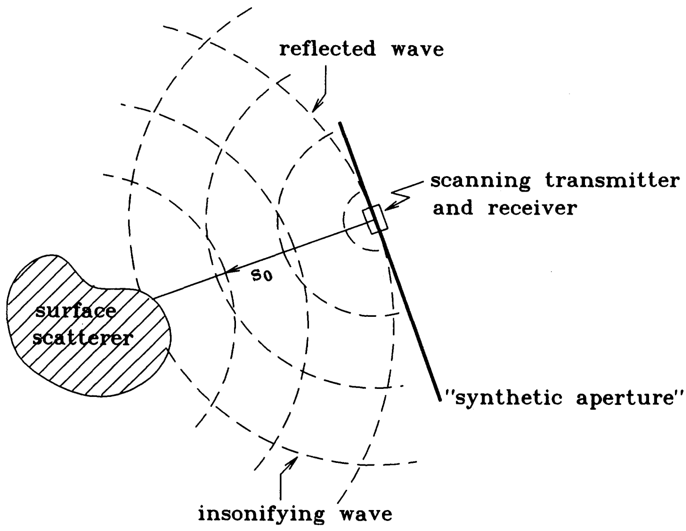

FIG. 11. Geometry for synthetic-aperture imaging.

where $\mathbf{r}_0 = r_0\mathbf{s}_0$ denotes the location of the transducer and where the far-field approximation has been employed. It is concluded from Eq. (26) that the measured object field determines the spatial Fourier transform of the scattering potential at the spatial frequency $\mathbf{K} = 2k\mathbf{s}_0$, which is a single point on an Ewald sphere having radius $2k$. Fourier space can then be filled out by using a broad-band transducer located at various points on some surface Σ_0. The result is similar to what would be obtained in multifrequency holography where the field is recorded over a holographic aperture lying on this surface. For this reason this imaging method is called synthetic-aperture imaging (Goodman, 1968).

ACKNOWLEDGEMENTS

The author wishes to thank Professor Jakob Stamnes and Dr. Alon Schatzberg for carefully reading the entire manuscript and making a number of suggestions that have improved the presentation of the material. He also wishes to thank Dr. Schatzberg for preparing the figures presented in the text.

GLOSSARY

Acoustics: The science of sound waves. Used in this article to denote applications where it is desirable to generate images of sources of sound waves usually at high (ultrasonic) frequencies.

Backpropagation: The mathematical process of computing the spatial distribution of a wave field within a source region from measured values of the wave field over an external aperture.

Born Approximation: A weak-scattering approximation used in scattering theory, valid when the scattering object has properties that closely match the background medium in which it is embedded.

Coherent Wave: Wave field whose amplitude and phase are nonrandom functions of time and position.

Computed Tomography (CT): The mathematical procedures used to reconstruct a multidimensional function from its set of projections. Used in x-ray tomography to compute a cross-sectional image of an object from multiple radiographs of the object.

Diffraction Tomography (DT): The mathematical procedures used to reconstruct a scattering potential from a set of scattering data.

Fraunhofer Hologram: A hologram recorded in the Fraunhofer region of the source. Same as Fourier-transform hologram.

Fresnel Hologram: A hologram recorded in the Fresnel region of the source.

Gabor Hologram: A hologram recorded using the illuminating wave as the reference wave. Sometimes called an "in-line" hologram.

Generalized Projection-Slice Theorem: The relationship between the spatial Fourier transforms of a source and the source's ideal holographic image.

Hologram: The recorded interference pattern between a wave generated from a source and a known reference wave.

Ideal Hologram: Hologram recorded over an aperture that completely surrounds the source.

Image: The spatial distribution of field amplitude generated by a lens or by a holographic reconstruction that resembles a source distribution. A *real* image is concentrated in space and is pointwise directly related to the source. A *virtual* image is not focused in space and can only be viewed after an additional imaging operation such as performed by a positive lens or on a computer via a backpropagation algorithm.

Near-Feld Hologram: A hologram recorded in the near field of a source. Such a hologram can include evanescent wave components of the source field and, hence, can lead to superresolution imaging.

Object Wave: The coherent wave generated by the source (object) that is to be imaged in holography.

Off-Axis Holography: A form of holography where the reference wave is different from the illuminating wave.

Reference Wave: The known wave (usually a plane or spherical wave) that is coherently added to the object wave to generate the interference pattern recorded in a hologram.

Scattering Potential: The spatial distribution of material properties determining the scattering of a wave from a penetrable object.

Synthetic-Aperture Holography: A specialized form of holography where a sequence of backscatter measurements is used to generate synthetically an image that would ordinarily be obtained from a hologram recorded over an extended aperture.

Lists of Works Cited

Born, M., Wolf, E. (1980), *Principles of Optics*, 6th ed., New York: Pergamon.

Cowley, J. M. (1984), *Diffraction Physics*, New York: North-Holland.

Devaney, A. J. (1983), "Inverse source and scattering problems in ultrasonics," *IEEE Trans. Sonics Ultrason.* **SU-30**, 355–364.

Devaney, A. J. (1984), "Geophysical diffraction tomography," *IEEE Trans. Geosci. Remote Sensing* **GE-22**, 3–13.

Gabor, D. (1949), "Microscopy by reconstructed wavefronts," *Proc. R. Soc. London Ser. A* **197**, 454–487.

Gabor, D. (1951), "Microscopy by reconstructed wavefronts: ii," *Proc. Phys. Soc. London Sec. B* **64**, 449–469.

Goodman, J. (1968), *Introduction to Fourier Optics*, New York: McGraw-Hill.

Greenleaf, J. F., Bahn, R. C. (1981), "Clinical imaging with transmissive ultrasonic computerized tomography," *IEEE Trans. Biomed. Eng.* **BME-28**, 177–185.

Greenleaf, J. F., Johnson, S. A., Lee, S. L., Herman, G. T., Wood, E. H. (1974), "Algebraic reconstruction of spatial distributions of acoustic absorption in tissues from their two-dimensional acoustic projections," in: P. S. Green, (Ed.), *Acoustic Holography*, New York: Plenum, pp. 591–603.

Hildebrand, B. P., Brenden, B. B. (1972), *An Introduction to Acoustical Holography*, London: Adam Hilger.

Kessler, L. W., Yuhas, D. E. (1979), "Acoustic microscopy—1979," *Proc. IEEE* **67**, 526–536.

Langenberg, K. J. (1987), "Applied inverse problems for acoustic, electromagnetic and elastic scattering," in: P. C. Sabatier (Ed.), *Basic Methods of Tomography and Inverse Problems*, London: Adam Hilger.

Leith, E. N., Upatnieks, J. (1962), "Reconstructed wavefronts and communication theory," *J. Opt. Soc. Am.* **52**, 1123–1130.

Mueller, R. K. (1971), "Acoustic holography," *Proc. IEEE* **59**, 1319–1335.

Mueller, R. K., Kaveh, M., Wade, G. (1979), "Reconstructive tomography and applications to ultrasonics," *Proc. IEEE* **67**, 567–587.

Natterer, F. (1986), *The Mathematics of computerized Tomography*, New York: Wiley.

Porter, R. P., Schwab, W. C. (1971), "Optimum imaging. Closed holograms and optical channel capacity," *J. Opt. Soc. Am.* **61**, 789–796.

Stamnes, J. J. (1986), *Waves in Focal Regions*, Bristol: Adam Hilger.

Stamnes, J. J., Gelius, L. J., Johnasen, I., Sponheim, N. (1992), "Diffraction tomography applications in seismics and medicine," in: M. Bertero, E. R. Pike (Eds.), *Inverse Problems in Scattering and Imaging*, London: Adam Hilger.

Toraldo di Francia, G. (1969), "Some recent progress in classical optics," *Riv. Nuovo Cimento I*, **1**, 460–484.

Wey, A. C., Kessler, L. W. (1990), "Holographic scanning laser acoustic microscope," in: Hua Lee, Glen Wade (Eds.), *Acoustical Imaging*, New York: Plenum, Vol. 18.

Williams, E. G., Maynard, J. D. (1980), "Holographic imaging without the wavelength resolution limit," *Phys. Rev. Lett.* **45**, 554–557.

Witten, A., Long, E. (1986), "Shallow applications of geophysical diffraction tomography," *IEEE Trans. Geosci. Remote Sensing* **GE-24**, 654–662.

Wolf, E. (1969), "Three-dimensional structure determination of semi-transparent objects from holographic data," *Opt. Commun.* **1**, 153–156.

Further Reading

General references in the area of acoustic holography can be found in the two series *Acoustic Holography* and *Acoustic Imaging*, both published by Plenum Press (see Greenleaf *et al.*, 1974, and Wey and Kessler, 1990, in List of Works Cited). Additional papers of interest follow.

Berger, M., Bruck, D., Fischer, M., Langenberg, K. J., Oberst, J., Schmitz, V. (1981), "Potential and limits of holographic reconstruction algorithms," *J. Nondestructive Eval.* **2**, 85–111.

Lohmann, A. W. (1978), "Three-dimensional properties of wave-fields," *Optik* **51**, 105–116.

Norton, S. J., Linzer, M. (1979), "Ultrasonic reflectivity imaging in three dimensions: reconstruction with spherical transducer arrays," *Ultrasonic Imaging* **1**, 210–231.

Porter, R. P. (1989), "Generalized holography with applications to inverse scattering and inverse source problems," in: E. Wolf (Ed.), *Progress in Optics*, Amsterdam: North-Holland.

Schulz, G. (1976), "Three-dimensional linear transfer by means of holographic depth selection," *Optic. Acta* **23**, 1029–1038.

Streibl, N. (1984), "Fundamental restrictions for 3-D light distributions," *Optik* **66**, 341–354.

Wolf, E., Shewell, J. R. (1970), "Diffraction theory of holography," *J. Math. Phys.* **11**, 2254–2267.

Zernike, F. (1948), "Diffraction and optical image formation," *Proc. Phys. Soc. London* **61**, 158–164.

HOLOGRAPHY, ELECTRON

Akira Tonomura, *Advanced Research Laboratory, Hitachi, Ltd., Hatoyama, Japan*

Introduction 531
1. **Brief History** 531
2. **Electron Waves** 532
3. **Interaction of Electron Waves with Objects** 533
4. **Electron Holography** 534
4.1 Types of Electron Holography 534
4.2 Electron Hologram Formation 535
4.3 Optical Image Reconstruction 536
4.4 Numerical Image Reconstruction 537
5. **Applications** 537
5.1 Interference Microscopy 537
5.1.1 Measurement of Object Thickness Distribution 537
5.1.2 Observation of Magnetic Fields 538
5.1.3 Observation of Electrostatic Fields 538
5.2 High-Resolution Microscopy 539
Glossary 540
Works Cited 540
Further Reading 540

INTRODUCTION

Holography is a two-step imaging method. In the first step, an interference pattern between an object wave and a known reference wave is recorded on film as a hologram. Then the image of the original object is reconstructed in three dimensions by illuminating the hologram film with a light beam equivalent to the reference wave. When a light beam is used in the hologram formation process, it is called optical holography, and when a sound wave is used it is called acoustical holography. Electron holography employs electrons rather than massless particles such as photons.

Optical images can be reconstructed from holograms formed with electron waves, even though electron waves seem to be quite different from light waves. For example, an electron wave is described by a quantum-mechanical wave function, while a light wave can be described by classical electromagnetic fields. Electron and light wavelengths differ by a factor of about 100,000. However, there is no essential difference between the two waves with regard to interference patterns. The holographic imaging capability of such different waves is based on the fact that only the most fundamental features of waves, i.e., interference and diffraction, are used for this imaging.

1. BRIEF HISTORY

The principle underlying holography was devised by Gabor (1949) for breaking through the resolution limit of electron microscopes, which were restricted by limits resulting from the unavoidable aberrations of electron lenses. Optical reconstruction was tried from optical, electron, and x-ray holograms. However, the images obtained in those days were rather poor mainly because of the lack of coherent beams and the disturbance from the conjugate images inevitably produced in holography.

In the early 1960s, holography greatly progressed in the field of optics with the advent of laser beams. Furthermore, clear images were obtained by the introduction of the off-axis method to remove the effect of the conjugate images (Leith and Upatnieks, 1962). Since

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

that time, applications of optical holography have been greatly expanded.

The situation is similar in electron holography. Although the feasibility of electron holography was demonstrated in 1968 (Tonomura *et al.*, 1968; Möllenstedt and Wahl, 1968), new application fields, beyond the reach of conventional electron microscopes, were not opened until 1979 when a coherent field-emission electron beam was developed (Tonomura *et al.*, 1979). This article explains only electron holography, but the principle behind it is the same as that behind holography using other massive particles.

2. ELECTRON WAVES

The wave nature of an electron is described by the Schrödinger equation

$$\{(\hbar\nabla/i+e\mathbf{A})^2/2m-eV\}\psi=E\psi, \qquad (1)$$

where m, $-e$, V, $\mathbf{A}$, E, $\hbar$, and ψ are the electron mass, the electric charge unit, the electric potential, the vector potential, the energy, Planck's constant divided by 2π, and the wave function. The wave function is completely different from any classical wave in that it does not correspond to any physical reality, and this wave function itself can never be detected in any conceivable experiment. One observable property is the probability $|\psi|^2$ of detecting single particles. The accumulation of detection records gives the probability $|\psi|^2$.

This process is demonstrated well in the electron double-slit-like experiment shown in Fig. 1. The wavelength λ of an electron wave in a vacuum ($V=\mathbf{A}=0$) is obtained from Eq. (1) as

$$\lambda=h/\sqrt{2mE}=h/p, \qquad (2)$$

where p is the electron momentum. Let us take the example of a 100-keV electron beam, which has a wavelength of 0.04 Å. An electron wave emitted from a source propagates towards the electron biprism (Möllenstedt and Düker, 1955), which consists of a central thin filament and two grounded electrodes. The application of a positive electric potential to the central filament makes the two electron waves passing the filament (one on each side) overlap to form an interference pattern in the lower plane. Continuous wave fronts are drawn just like a water ring on a pond surface, as in Fig. 1. However, the actual situation in this experiment is quite different from this image. Two electrons never exist in this apparatus at the same time; there is only a single electron if any. So why can we observe an interference pattern between two waves passing the two sides of the biprism filament when two electrons do not overlap? There is no choice but to think that a single electron is split into two (one on each side) of the biprism filament, which interfere with each other. We detect only a single electron and never one of the halves. This can be explained using wave functions: an electron wave function can pass through the two sides of the biprism filament and give the measurement probability of the interference pattern at the observation plane. Single electrons are detected according to this probability.

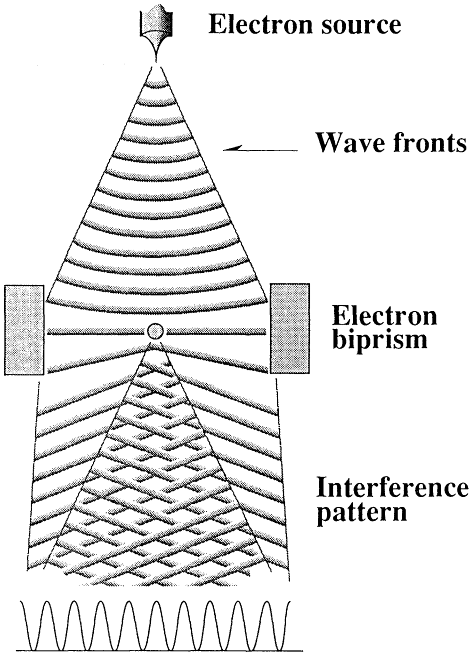

FIG. 1. Electron-wave interference. An electron wave emerges from a field-emission tip and propagates towards the electron biprism. Two waves, propagating on both sides of the biprism, overlap to form an interference pattern. Actually, however, coherent electrons are so sparse that only a single electron exists in the apparatus at a time.

This strange behavior of electron interference, recorded on videotape (Tonomura *et al.*, 1989), is shown in Fig. 2. Electrons are detected one by one as bright spots on the TV monitor. At first, their arrival seems quite at

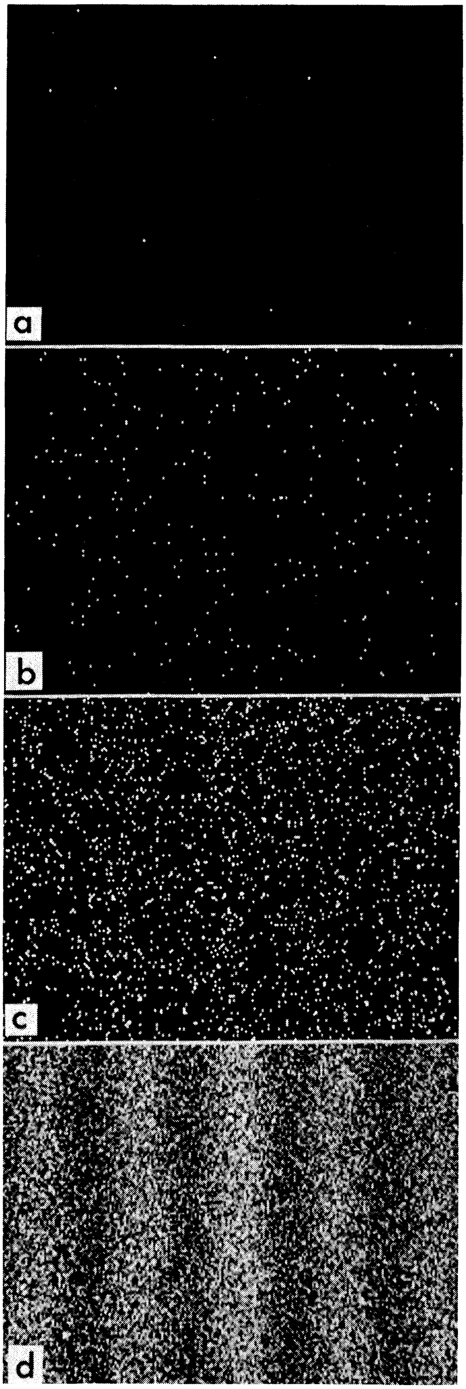

FIG. 2. Buildup of the electron interference pattern: (a) 8 electrons; (b) 270 electrons; (c) 2,000 electrons; and (d) 60,000 electrons.

random. Although the next electron is emitted long after an electron is detected, the accumulation of bright spots gradually forms the biprism interference pattern. Since the electron wavelength is extremely small, the fringe spacing is also narrow. For this reason, Feynman pointed out in his textbook that "this experiment has never been done in just this way, since the apparatus would have to be made on an impossibly small scale." However, such an experiment can now be performed using advanced technologies, such as two-dimensional single-electron counting systems combined with a field-emission electron microscope.

3. INTERACTION OF ELECTRON WAVES WITH OBJECTS

The interaction of an electron wave with an object is described by the Schrödinger equation (1). It can easily be seen from this equation that electrostatic potential V and vector potential $\mathbf{A}$ determine the behavior of the wave function. To take a specific example, the wave front of an electron plane wave incident on an object can be modulated through electromagnetic potentials as illustrated in Fig. 3. This modulation can be recorded in the interference pattern or the hologram as displacements of interference fringes if a reference plane wave is made to overlap this modulated wave. The modulation of the wave front, or the phase shift ϕ of an electron wave passing through weak electromagnetic potentials, is easily derived from Eq. (1) as

$$\phi=\frac{\sqrt{2me}}{\hbar}\int\left(\sqrt{V}-\sqrt{\frac{e}{2m}}\mathbf{t}\cdot\mathbf{A}\right)ds, \qquad (3)$$

where the integral is carried out along the electron trajectory, and $\mathbf{t}$ is the unit vector of the electron trajectory.

When an object is made of uniform material, the phase shift is caused by the change in the object's thickness. This is because when an electron passes through an object, it is accelerated by the inner potential V_0 and consequently its wavelength becomes a little shorter. The phase difference $\Delta\phi$ with and without the object is given by

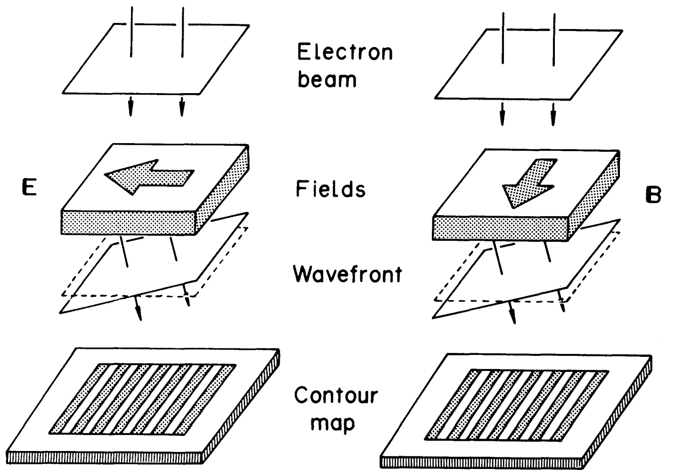

FIG. 3. Interaction of electron plane wave with electromagnetic fields: **E**, electric field; **B**, magnetic field. The transmitted wave front can be displayed as a contour map by overlapping a comparison beam (shown by dotted lines in the figure).

$$\Delta\phi=\frac{\sqrt{2\ \mathrm{MeV}}}{2\hbar}\int\frac{V_0}{V}\,ds=\frac{\sqrt{2\ \mathrm{MeV}}}{2\hbar}\frac{V_0}{V}\,d, \quad (4)$$

where d is the thickness of the object.

When an object is a magnetic field, the phase difference $\Delta\phi$ between two electron beams starting from one point and ending at another point is given by

$$\Delta\phi=-\frac{e}{\hbar}\int\mathbf{A}\cdot d\mathbf{s}=-\frac{e}{\hbar}\int\mathbf{B}\cdot d\mathbf{S}=-\frac{e}{\hbar}\,\Phi, \quad (5)$$

where the line integral is performed along the closed loop connecting the two electron trajectories and the surface integral in the second equation is performed over the surface enclosed by the loop. This equation indicates that the phase difference produced between the two beams is proportional to the enclosed magnetic flux even when the two beams never touch the magnetic field. This is called the Aharonov–Bohm effect (Aharonov and Bohm, 1959; Peshkin and Tonomura, 1989).

4. ELECTRON HOLOGRAPHY

4.1 Types of Electron Holography

Although the basic principle behind holography is very simple, many holography techniques have been developed, especially in the field of optics. Only a few of these techniques have been applied to electron holography.

Holography can be classified into two types according to how the reference wave is provided:

1. A background wave is used as the reference wave in in-line holography. Here, objects are restricted to a size small enough not to disturb the background wave. In addition, a hologram must be formed in a Fraunhofer diffraction pattern of the object in order to obtain a clear reconstructed image. In holography, an additional conjugate image is always produced since an interference pattern (hologram) can only record the absolute value of the phase, which reconstructs as twin images having amplitudes that are complex conjugates of each other. Therefore, the defocused pattern overlaps the reconstructed image, and in the early days, this was a persistent obstacle to holography. The effect of a conjugate image becomes a constant background near the reconstructed image in Fraunhofer in-line holography. Clear electron images were first obtained using this type of holography, since the coherence conditions of an electron beam are not stringent in this case. However, it is now used only for special purposes.
2. In off-axis holography, a known reference wave and an object wave are brought together at an angle. The twin images are reconstructed in different directions, which allows the twin image problem to be eliminated. There are no limitations on object size. Beam splitters are necessary to produce two coherent waves. Electron biprisms have usually been employed as beam splitters, but Bragg reflections at a single-crystalline thin film can also be used.

Holography can also be classified according to the plane of the object wave where holograms are formed, such as image-plane, Fresnel, Fraunhofer, and Fourier-transform holography. All of these types have already been demonstrated in electron holography.

If one classifies these techniques by the apparatus needed to form holograms, the standard one is the transmission electron microscope equipped with a coherent field-emission gun and an electron biprism. This is discussed in Sec. 4.2. Other types of apparatus include the scanning transmission electron microscope (STEM), the electron diffractometer (see Howie, 1990), and the projection electron microscope (Fink *et al.*, 1990).

4.2 Electron Hologram Formation

This section describes a typical experimental setup for forming an off-axis electron hologram using a transmission electron microscope (TEM).

Electrons are field-emitted from a sharp tungsten tip under an ambient pressure of 10^{-10} Torr (see Fig. 4). They are then accelerated through accelerating electrodes to, say, 100 kV, and illuminate an object with a collimation angle down to less than 1×10^{-7} rad, compared with 1×10^{-3} rad for conventional electron microscopes. A small illumination angle (2β) is crucial since this value determines the spatial coherence length, which is given by $\lambda/2\beta$. This has become feasible through the high brightness of a field-emission electron beam.

Hologram formation is schematically shown in Fig. 5. A collimated electron beam illuminates an object located in one half of the specimen plane, and the other half is for the reference beam. The image of the object is formed through the objective lens. The reference beam reaches the image plane without any disturbance from the object. These two beams are superimposed by the electron biprism situated below the objective lens to form an interference pattern.

The biprism is composed of a central thin filament and two grounded potential electrodes, one on each side. A positive potential up to 300 V is applied to the filament to make the object and reference beams deflect toward the center to overlap on the lower plane. The resultant interference pattern is

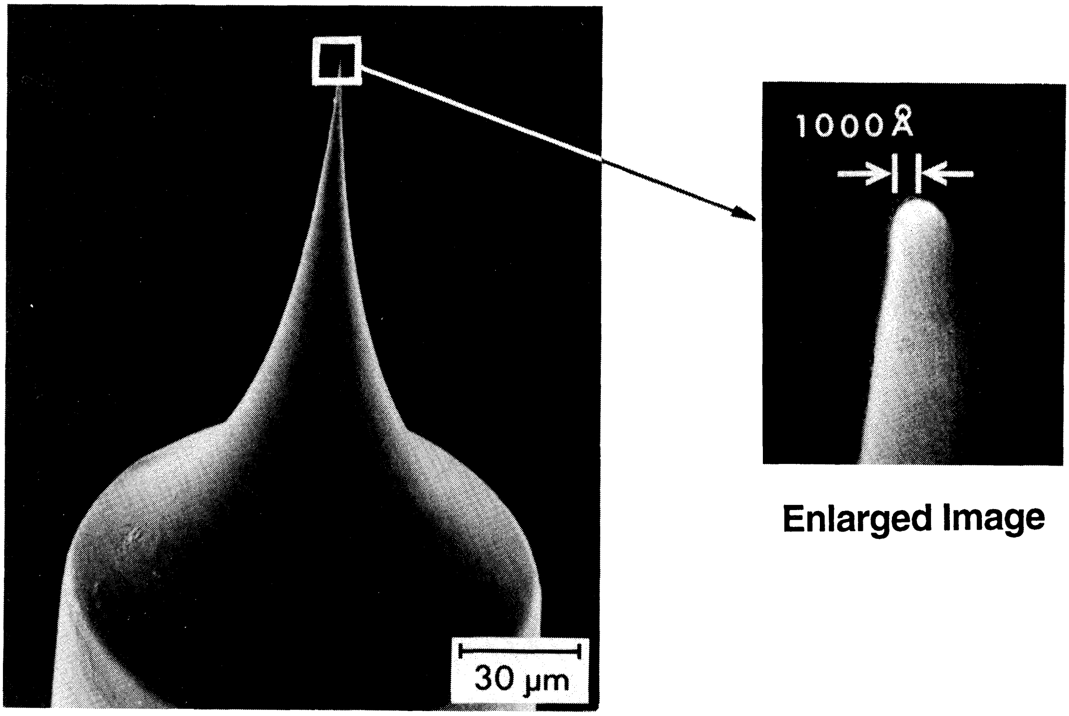

FIG. 4. Scanning electron micrograph of a field-emission tip.

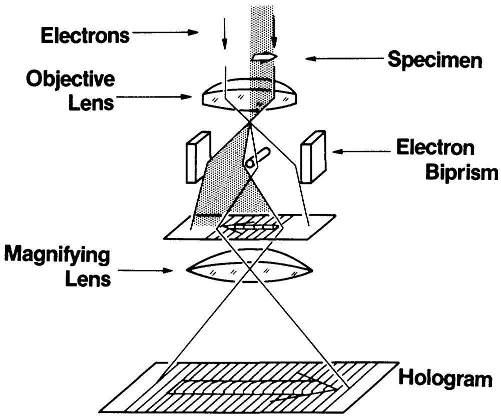

FIG. 5. Schematic diagram of electron hologram formation in an electron microscope.

enlarged through magnifying lenses and recorded as a hologram.

4.3 Optical Image Reconstruction

An optical image can be reconstructed in one of the two diffracted beams when a light beam corresponding to the reference electron beam illuminates a hologram. A conjugate image is also produced in the other diffracted beam. When the reconstructed image is photographed on film, only the intensity distribution is observed, which is equivalent to an electron micrograph.

The phase distribution is also reconstructed in an optical image, and consequently it can be displayed as an interference micrograph when an optical interferometer is combined in the optical reconstruction system as shown in Fig. 6. A light beam from a laser is split into two coherent beams by a Mach–Zehnder interferometer. The two beams, A and B, illuminate the electron hologram from slightly different directions. Each beam produces both a reconstructed image in its +1st-order diffracted beam and a conjugate image in its −1st-order diffracted beam. When only the transmitted beam (zeroth-order beam) of beam A and the first-order diffracted beam of beam B, or vice versa, are made to pass through the aperture to form an interference pattern between the reconstructed image and a comparison beam on the observation plane, an interference micrograph is obtained which displays the phase distribution (Fig. 6 does not show this case but the phase-amplification case described next).

In this interference micrograph, contour fringes appear at every phase change of 2π, and so the minimum detectable value is about $2\pi/4$. This precision is not always satisfactory for electrons, especially when we want to observe a specimen thickness variation on the order of atomic dimensions or a small magnetic flux. In such cases, however, a technique unique to holography can be employed to detect a smaller phase change. This technique uses the same optical system as in Fig. 6.

If the reconstructed image of beam A and the conjugate image of beam B, or vice versa, are overlapped by adjusting the interferometer, the resulting interference pattern produces doubly phase-amplified interference. This is because the phase difference between the twin images, which have phase distributions opposite in sign, is twice as large as that for a conventional interference micrograph. Higher amplification can be obtained when this procedure is repeated.

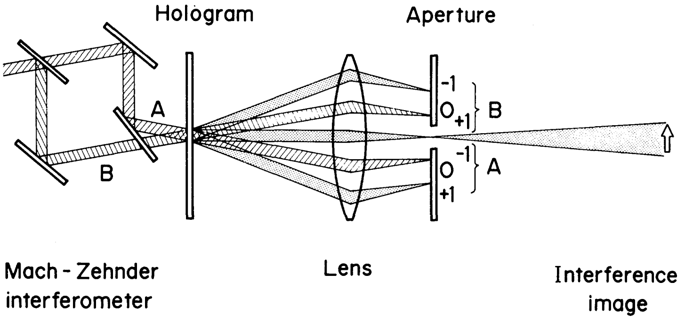

FIG. 6. Optical reconstruction system for interference microscopy.

4.4 Numerical Image Reconstruction

The image can be reconstructed by computation from holograms directly monitored with a TV camera. This method does not need wet processes such as film development. Consequently the time required for the reconstruction can be made much shorter, and even real-time observation becomes feasible when the computation time is short.

5. APPLICATIONS

This section introduces new application fields opened up by electron holography.

5.1 Interference Microscopy

An interference micrograph provides information which cannot be obtained by electron microscopy—phase information about the electron wave function as well as the intensity distribution. Furthermore, the precision in phase measurement can be increased from $2\pi/4$ up to $2\pi/100$ by using the phase-amplification technique. High-precision measurements and observation of unseen microscopic objects and fields have become feasible by holographic interference microscopy.

5.1.1 Measurement of Object Thickness Distribution. The phase shift of a uniform object is given by Eq. (4). For $V = 100$ kV and $V_0 = 10$–30 V, the object thickness for a phase shift of 2π is 200–700 Å. If the phase measurement precision is $2\pi/100$ as a result of phase amplification, the thickness precision is 2–7 Å.

An example for this application is shown in Fig. 7 (Tonomura *et al.*, 1985). The object is a molybdenite thin film. No contrast can be observed in a conventional electron micrograph. Even when observed as an interference micrograph (b), no fringe shift can be seen since the film surfaces are almost atomically flat. However, in the 24-times phase-amplified micrograph (c), fringe steps become recognizable along lines marked A, B, and C in the photography, which correspond to one, three, and five atomic steps on the surface. The thickness change of 6.2 Å due to the single atomic step A has become quantitatively observable, and it produces an electron phase shift of $2\pi/50$.

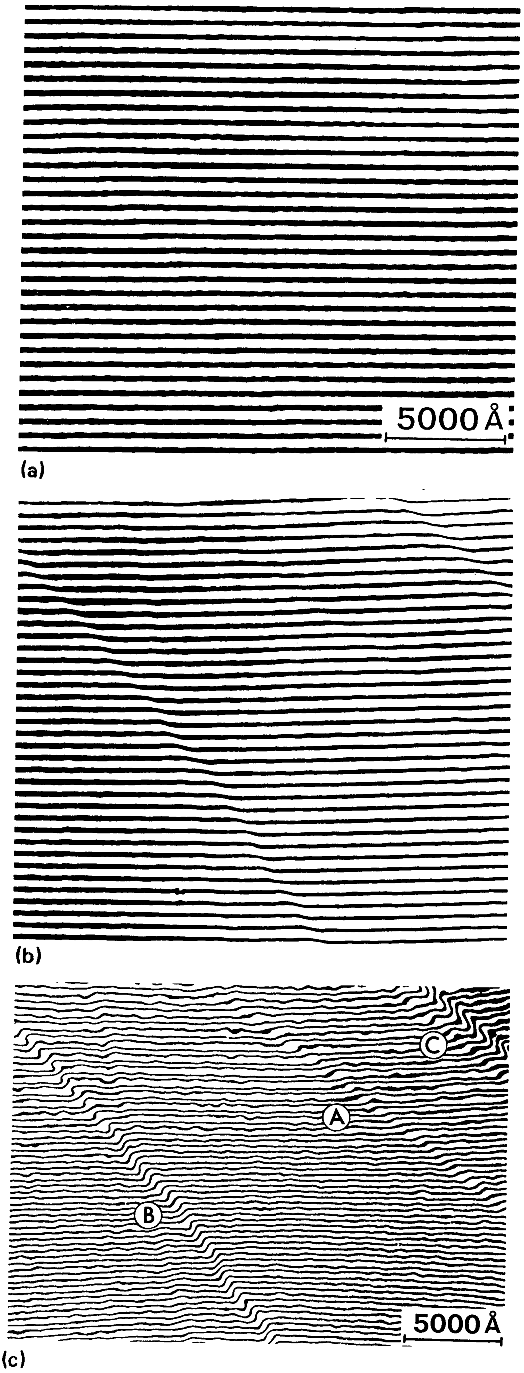

FIG. 7. Interference electron micrograph of molybdenite thin film: (a) conventional interference micrograph; (b) 4× phase-amplified interference micrograph; and (c) 24× phase-amplified interference micrograph. Fringe step A corresponds to a thickness change of monoatomic step height, 6.2 Å, and electron phase shift of $2\pi/50$.

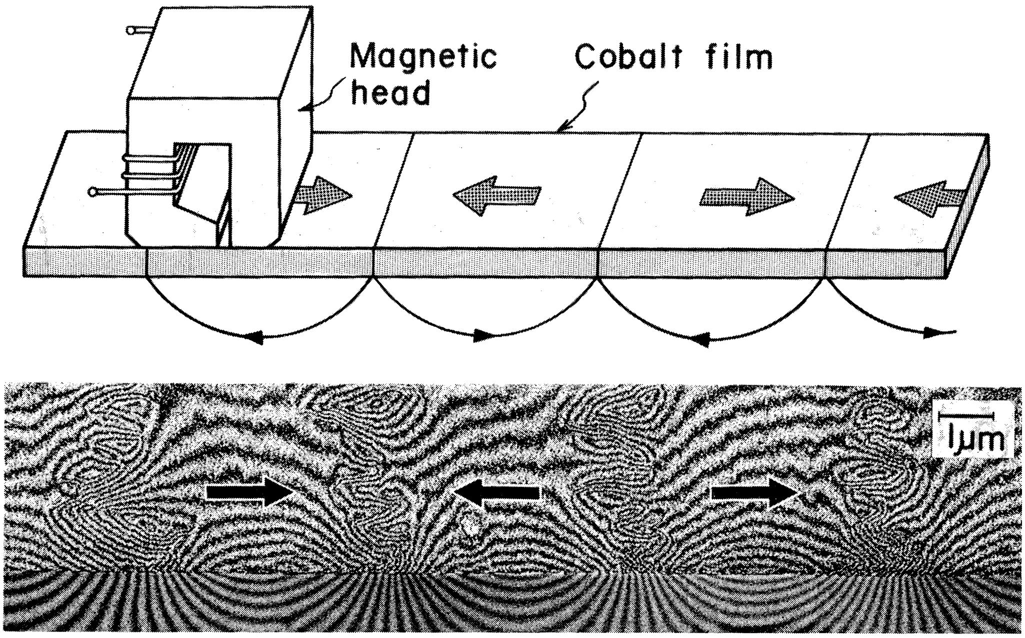

FIG. 8. Interference electron micrograph of recorded magnetic tape. A constant magnetic flux of h/e flows between two adjacent contour lines.

5.1.2 Observation of Magnetic Fields. An interference micrograph of a magnetic field can be interpreted in a straightforward way: contour fringes in an interference micrograph follow in-plane magnetic lines of force, and a constant magnetic flux of h/e ($=4\times10^{-15}$ Wb) flows between two adjacent contour fringes (Tonomura *et al.*, 1980).

An observation example of magnetic tape is shown in Fig. 8 (Osakabe *et al.*, 1983). A bit pattern was recorded on a cobalt film, and was observed as an interference micrograph. No contrast can be observed by electron microscopy, since this film is a pure phase object. However, in this interference micrograph, many fringes, i.e., magnetic lines of force, appear inside and outside the film. Two oppositely directed streams of magnetization collide head on, and produce vortices similar to those formed by streams of water. Such detailed observation is helpful for the development of high-density magnetic recording.

Another example is the observation of flux lines penetrating a superconducting thin film (Matsuda *et al.*, 1989). Figure 9 shows a doubly phase-amplified interference micrograph

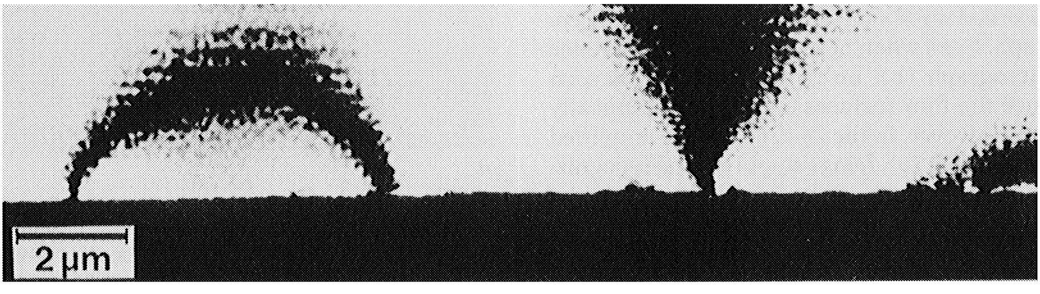

FIG. 9. Interference electron micrograph of single flux lines penetrating a superconducting Pb thin film 0.2 μm thick. One fringe corresponds to a flux of $h/2e$, i.e., a flux quantum, since the micrograph is phase-amplified by a factor of 2.

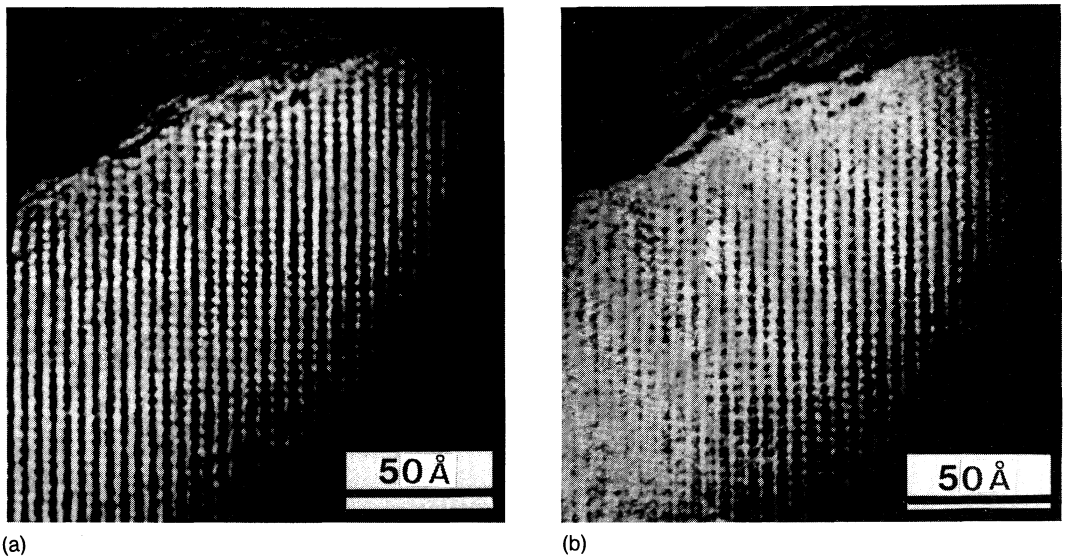

(a) (b)

FIG. 10. High-resolution image reconstruction: (a) reconstructed image of Nb_2O_5; (b) spherically corrected image (Lichte, 1986).

of Pb film 0.2 μm thick. When a magnetic field is applied perpendicularly to the film at 4.5 K, the magnetic field is excluded from the film due to the Meissner effect. A stronger field can penetrate it in the form of a thin filament, i.e., a flux line. An electron beam is incident perpendicular to the flux lines for observation. A single contour fringe in the picture corresponds to a flux line having a flux quantum of $h/2e$. A magnetic line of force is produced from a very small region (about 0.15 μm) of the film surface and fans out into the vacuum. In addition to this isolated flux line, an antiparallel pair of flux lines can also be observed on the left in the picture.

5.1.3 Observation of Electrostatic Fields. When the object is an electrostatic field, contour fringes in an interference micrograph indicate projected equipotential lines, as can be presumed from Eq. (4). This technique has been used to observe the microscopic distribution of electrostatic fields near *p-n* junctions for two values of applied reverse-bias voltages (Frabboni *et al.*, 1985).

5.2 High-Resolution Microscopy

The original objective of holography was to overcome image resolution limitations due to aberrations of electron lenses. Although electron microscope resolution has been greatly improved since the days of Gabor, the resolution is still limited by aberrations and not by wavelength. Aberrations such as spherical aberration and astigmatism in electron microscopes can be compensated for in the optical reconstruction stage.

Tonomura *et al.* (1979) attempted to compensate optically for a spherically aberrated image of a fine gold particle, and the lattice fringes of {111} planes (2.4 Å) and their half-spacing fringes (1.2 Å) were reconstructed inside the in-focus particle image. Lichte (1986) numerically reconstructed structure images of Nb_2O_5 crystals from electron holograms which had a 0.75-Å carrier fringe spacing, and obtained spherically corrected images for both phase and amplitude distributions (see Fig. 10). The carrier fringe spacing must be narrower than 1/3 of the required resolution. Spacings of 0.3 (Völkl and Lichte, 1990) and 0.09 Å (Kawasaki *et al.*, 1991) have been obtained for high-resolution purposes.

In spite of several trials, no images have yet been obtained that were unobtainable with conventional electron microscopes. We can expect that an ultrahigh resolution of better than 1 Å will be achieved with brighter electron beams.

GLOSSARY

Aharonov–Bohm Effect: Electron interference phenomenon where an electron wave passing through a field-free region receives an observable effect due to potentials.

Brightness: Beam current density per unit solid angle.

Carrier Fringe: Information about an object wave is recorded as the modulation of fine carrier fringes in off-axis holography.

Conjugate Image (Twin Image): An additional image inevitably reconstructed in the holography process. Its amplitude is the complex conjugate of the original amplitude.

Illumination Angle: The maximum variation of the propagation direction of a beam of particles.

Lattice Fringe: An electron microscope image of a crystalline thin film which is formed by the overlap of the transmitted electron beam and Bragg-reflected beams at some lattice planes. It provides an image of the lattice planes.

List of Works Cited

Aharonov, Y., Bohm, D. (1959), *Phys. Rev.* **115**, 485–491.

Feynman, R. P., Leighton, R. B., Sands, M. (1963), *The Feynman Lectures on Physics*, Reading, MA: Addison-Wesley.

Fink, H.-W., Stocker, W., Schmid, H. (1990), *Phys. Rev. Lett.* **65**, 1204–1206.

Frabboni, S., Matteucci, G., Pozzi, G., Vanzi, M. (1985), Phys. Rev. Lett. **55**, 2196–2199.

Gabor, D. (1949), *Proc. R. Soc. London Ser.* A **197**, 454–487.

Howie, A. (1990), *Nature* **345**, 386–387.

Kawasaki, T., Qing X. R., Matsuda, T., Bando, Y., Tonomura, A. (1991), *Jpn. J. Appl. Phys. Lett.* **30**, 1830–1832.

Leith, E. N., Upatnieks, J. (1962), *J. Opt. Soc. Am.* **52**, 1123–1832.

Lichte, H. (1986), *Ultramicroscopy* **20**, 293–304.

Matsuda, T., Hasegawa, S., Igarashi, M., Kobayashi, T., Naito, M., Kajiyama, H., Endo, J., Osakabe, N., Tonomura, A. (1989), *Phys. Rev. Lett.* **62**, 2519–2522.

Möllenstedt, G., Düker, H. (1955), *Naturwissenschaften* **42**, 41.

Möllenstadt, G., Wahl, H. (1968), *Naturwissenschaften* **55**, 340.

Osakabe, N., Yoshida, K., Horiuchi, Y., Matsuda, T., Tanabe, H., Okuwaki, T., Endo, J., Fujiwara, H., Tonomura, A. (1983), *Appl. Phys. Lett.* **42**, 746–748.

Peshkin, M., Tonomura, A. (1989), *The Aharonov–Bohm Effect*, Lecture Notes in Physics, Vol. 340, Berlin: Springer.

Tonomura, A., Fukuhara, A., Watanabe, H., Komoda, T. (1968), *Jpn. J. Appl. Phys.* **7**, 295.

Tonomura, A., Matsuda, T., Endo, J., Todokoro, H., Komoda, T. (1979), *J. Electron Microsc.* **28**, 1–11.

Tonomura, A., Matsuda, T., Endo, J., Arii, T., Mihama, K. (1980), *Phys. Rev. Lett.* **44**, 1430–1433.

Tonomura, A., Matsuda, T., Kawasaki, T., Endo, J., Osakabe, N. (1985), *Phys. Rev. Lett.* **54**, 60–62.

Tonomura, A., Endo, J., Matsuda, T., Kawasaki, T., Ezawa, H. (1989), *Am. J. Phys.* **57**, 117–120.

Völkl, E., Lichte, H. (1990), *Ultramicroscopy* **32**, 177–180.

Further Reading

Hanszen, K.-J. (1982), in: L. Marton (Ed.), *Advances in Electronics and Electron Physics*, Vol. 59, New York: Academic, pp. 1–77.

Lichte, H. (1982), in: Congress Organizing Committee (Eds.), *Proceedings of the 10th International Congress on Electron Microscopy, Hamburg*, Vol. 1, Frankfurt: Deutsche Gesellschaft für Elektronenmikroskopie, pp. 411–418.

Lichte, H. (1991), in: T. Mulvey, C. J. R. Sheppard (Eds.), *Advances in Optical and Electron Microscopy*, Vol. 12, London: Academic Press, pp. 25–91.

Missiroli, G. F., Pozzi, G., Valdré, U. (1981), *J. Phys. E* **14**, 649–671.

Tonomura, A. (1986), *Progress in Optics*, Vol. 23, North-Holland: Amsterdam, pp. 183–220.

Tonomura, A. (1987), *Rev. Mod. Phys.* **59**, 637–669.

Tonomura, A. (1989), *J. Electron Microsc. Suppl.* **38**, 43–50.

Tonomura, A. (1990), *Physics Today*, **22** (4), 22–29.

Zeitler, E. (1979), in: G. W. Bailey (Ed.), *Proceedings of the 37th Electron Microscopy Society of America Meeting*, Baton Rouge: Claitor, pp. 376–379.

Holography

Emmett N. Leith
Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan, USA
Phone: (734) 764 9545; Fax: (734) 647 2718; e-mail: leith@engin.umich.edu

Abstract

Holography, a method of photography that produces highly realistic images, fully three-dimensional, with the normal parallax relations of the actual object, is described. Basics of the theory are presented from both a mathematical and a physical viewpoint. Various areas of application are given, including three-dimensional displays, complex spatial filtering, hologram interferometry for nondestructive testing and metrology, product authentication and counterfeiting security, holographic optical elements, and imaging through inhomogeneities using holographic phase conjugation. The various forms of white light viewable holograms are described, including the Denisyuk or volume type, the Benton or rainbow type, and the integral type. The diffractive properties of holograms are analyzed, including Bragg diffraction, i.e., diffraction from a volume hologram. The construction of commercial display holograms is discussed. The present state of commercial and artistic holography is discussed, along with expected future developments.

Keywords

imaging; phase; interference; wavefront reconstruction; metrology; nondestructive testing; optical storage; displays; phase conjugation.

1 **Introduction** *774*
2 **Holographic Principles** *774*
3 **Mathematical Theory** *778*
4 **The Rise of Modern Holography** *780*

5 Some Early Applications *782*
5.1 Imaging Through Inhomogeneities Using Holographic Phase Conjugation *782*
5.2 Complex Spatial Filtering *783*
5.3 Hologram Interferometry *787*
5.4 Holographic Optical Elements *789*
6 Volume Holography *790*
7 Display Holography *793*
7.1 The Beginnings *793*
7.2 The Denisyuk Hologram *793*
7.3 The Benton Hologram *794*
7.4 The Multiplex and Related Holograms *795*
7.5 Construction of Commercial Display Holograms *796*
8 Holography Today *797*
Glossary *797*
References *799*
Further Reading *799*

1
Introduction

Holography is a process of recording a complete wave field, both amplitude and phase. It differs from conventional photography in which only the amplitude or equivalently, the irradiance of the field is recorded. By recording the entirety of the wave, some rather remarkable effects are produced. In particular, the images formed with holography are startlingly realistic, being fully three dimensional and containing all of the parallax relations that are encountered in everyday seeing but which are lost when an image is recorded photographically.

Holography was invented by the Hungarian – British scientist Dennis Gabor in 1947. His aim was to improve the resolution of the electron microscope, which at that time suffered from noncorrectable spherical aberration [1–3]. The idea was to photographically record the complete wave field of the electron waves, both phase and amplitude, and then make reconstructions using visible light, that is, to produce light waves that were replicas of the electron waves, complete with aberrations. The aberrations could then be corrected by appropriately designed lenses. The idea, while bold and imaginative, was unsuccessful. However, since then holography has burgeoned and has found many applications, all quite remote from Gabor's original intent.

2
Holographic Principles

Photographic film is inherently insensitive to phase, as are indeed all detectors of radiation. Yet the holographic process requires the recording of the entire wave, phase as well as amplitude. This feat is accomplished by converting phase variations into amplitude variations, which the film can then record. The process is described with reference to Figs. 1 through 3. Figure 1 shows a plane wave of light reflecting off an object (equivalently, for a transparent object, it could be transmitted through the object). The reflected light is a complicated wave pattern; the waves are of irregular shape, and the amplitude (brightness) varies from point to point. The phase is essentially the shape of the wave front, and the irregularities of the wave front of Fig. 1 show a phase that changes rapidly from point to point. To record the phase means essentially to record the shape of the wave front. But the argument is just as before: photographic film is insensitive to the shape of the wave front.

Figure 2 shows how the introduction of a background wave can result in recording of the total field. For simplicity, we have sketched a rather simple object wave front, where the shape varies slowly across the recording aperture. This wave front is propagating toward the recording medium. Point A of the wave front is ahead of point B, and arrives at the recording plane first. But, as we stress, the recording plate is insensitive to the time of arrival of a wave front, being sensitive only to the amplitude; hence the photographic record has no information about the shape of the wave front. However, the introduction of a second wave called the *reference wave*, here a plane wave, effectively converts phase into amplitude through the process of interference. The interference of two beams of light produces a fringe pattern – a sequence of light and dark areas, where the two beams interfere constructively and destructively. The fringe spacing depends on the angle between the two beams. Hence, near the top of Fig. 2, where the angle between the waves is large, the fringe pattern is fine. Further down, where the two beams are nearly parallel, the fringe pattern is coarse. Finally, near the bottom, the angle has increased again, and the fringe pattern is again fine. Hence, one can, by observing the fringe pattern, gain an idea of the shape of the wave front. In short,

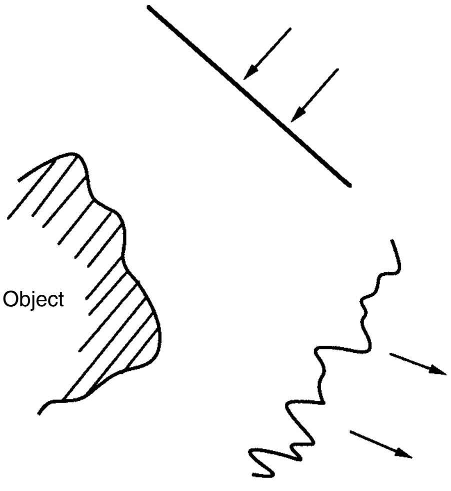

Fig. 1 Light reflecting off an object

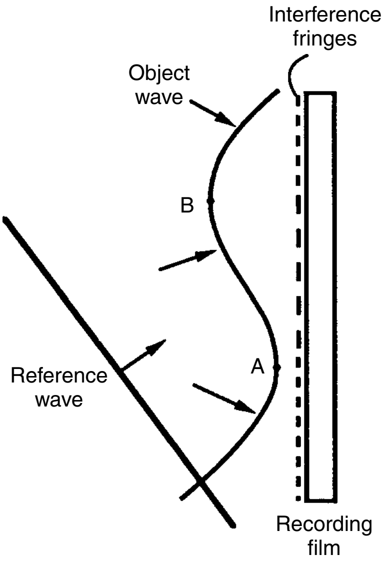

Fig. 2 Recording of a wavefront by interference with a reference wave

the effects of phase become visible. This phenomenon is, of course, the basis of all interferometry, including holography. The photographic film when developed becomes a hologram.

The hologram thus has, in an indirect way, made a record of the phase. The completion of the holographic process is the regeneration of the original wave front from the recorded pattern. We illuminate the hologram with a beam of light that, for simplicity, we take as a replica of the original reference wave. We think of the hologram now as a diffraction grating (Fig. 3), and consider that the angle of diffraction depends on the grating spacing – the closer the spacing, the greater the angle of diffraction. Then, at both the top and the bottom of the hologram, where the fringes are finest, the angle of diffraction is greatest. In the middle, where the fringes are coarser, the diffraction is at a smaller angle. Thus, the diffracted wave is a replica of the object wave that created the interference pattern when the hologram was recorded. The diffraction process is thus an inverse of the interference process, and the two processes in tandem lead back to the original wave front. Thus it is that phase can be recorded and wave fields reconstructed. This heuristic description of the holographic process basically says what mathematics says in a more quantitative and rigorous way.

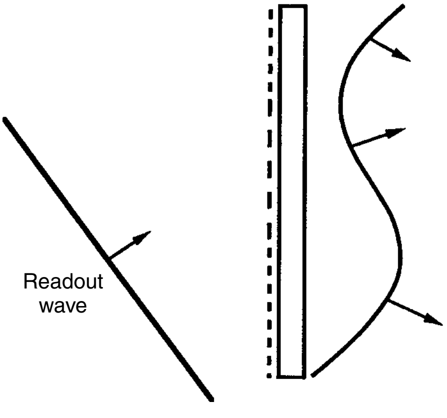

Fig. 3 Reconstruction of the wave by the process of diffraction. Only one diffracted order is shown

Consideration of the other diffracted order of the grating leads to a diffracted wave that is conjugate to the original wave; that is, in the conjugate wave, points A and B are reversed. On the original wave, as well as in the primary wave, which is the one we have described, point A is ahead of point B, that is, the phase at A leads that at point B. In the conjugate wave, this situation is reversed. Thus, the holographic process creates not just one image, but two, known as the *primary* and *conjugate images*. The conjugation results in the image forming downstream from the hologram, thus being a real image, as opposed to the primary, which is virtual. When the hologram is read out with a collimated beam of light, the two images lie in mirror-symmetric positions about the plane of the hologram, one image forming on the source side of the hologram and the other on the observer's side. However, this symmetry disappears when the illuminating beam is not collimated, and indeed, under illumination with highly divergent or convergent beams, it is possible for the images to be both real or both virtual. Hence, to call the two images the real and virtual images of the holographic process, as is sometimes done, can be misleading, and the terms primary and conjugate are generally preferred.

The existence of the extra image has been a problem in holography, since in the original Gabor process, where the reference beam was not introduced at an

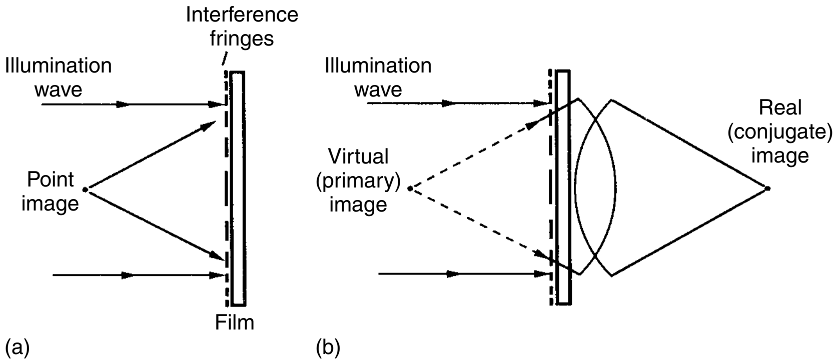

Fig. 4 Point image hologram made in the original Gabor configuration: (a) making the hologram; and (b) the reconstruction process, showing the emergence of two images

oblique angle, the images did not separate by propagation in different directions, but instead were superimposed, so that one image could be viewed only against the background of the other, giving a noisy reconstruction.

The origin of the twin image can be explained in various ways. A simple yet revealing way is described with respect to Fig. 4. Consider an object consisting of only a single point (a small opaque dot), and let this object be placed in a collimated beam of light. Most of the light passes undisturbed around the object, falls on the film, and serves as the reference beam. Light scattered by the object point also propagates to the recording film, where the interference between the two beams creates the interference pattern shown in Fig. 4(a). In the reconstruction process, with a collimated beam, we consider as before the hologram to be like a grating, one with a spacing (or spatial frequency) that changes with position. At each position, part of the incident light is diffracted in one direction and part in the opposite, at angles that depend on the spacing. The rays are seen to cross at two locations [Fig. 4(b)], one in front of the hologram and one behind. For the latter, the rays do not actually come together; rather, when projected back, they seem to emanate from a common point, thus producing a virtual image. Thus, the diffraction process has produced both real and virtual images, and the two images fall on top of each other so that one must always observe one image with the other giving an out-of-focus background, making the image noisy.

A somewhat philosophical way of explaining the origin of the twin image is to consider that a convergent wave of the same radius of curvature produces exactly the same interference pattern as the diverging wave when made to interfere with the planar background wave. Thus, the holographic process cannot distinguish between the two waves and cannot decide which wave produced the pattern. The reconstruction process resolves the quandary by reconstructing both.

Introducing a reference beam that passes around the object and impinges on the recording plate obliquely, as we have seen, sends the two reconstructed beams in different directions, thereby solving the problem.

3
Mathematical Theory

We offer a mathematical treatment that places the above viewpoint on a more rigorous basis (see also WAVE OPTICS).

The original theory, as developed by Gabor, assumed a transparent object s, which when illuminated with a coherent beam of light, that is, a monochromatic field from a point source, transmits some portion of the light without scatter and the remainder with scatter (Fig. 5). At some distance z_0 downstream from the object, let the field consist of a uniform component u_0, the background or reference wave, due to the unscattered portion of the wave, and u, the object beam produced by scatter from the object. The field has an irradiance $|u_0 + u|^2$. The field is recorded, typically by photographic film. If for simplicity we assume that the film has a transmittance proportional to the irradiance, thus ignoring the nonlinearity of the film and the contrast reversal that occurs upon recording, which makes the brighter parts of the field darkest on the record, we have the transmittance

$$\begin{aligned} t = I &= |u_0 + u|^2 \\ &= |u_0|^2 + |u|^2 + 2u_0|u|\cos\phi, \quad (1) \end{aligned}$$

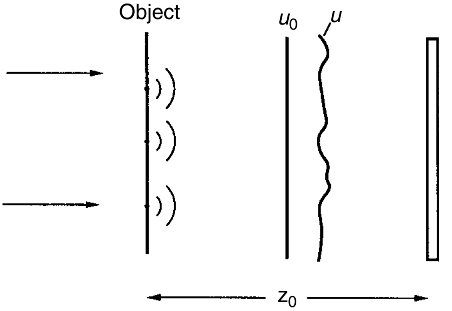

Fig. 5 Making a hologram in the Gabor configuration. The light transmitted by the object consists of a scattered wave u and an unscattered wave u_0

where we have dropped constants of proportionality and have assumed that u_0, the background wave, is uniform in amplitude and phase across the field, and that the varying part of the field, u, can be written in polar form as $u = |u|\exp(i\phi)$, where $|u|$ is the amplitude variation across the field and ϕ is the phase variation.

Upon illumination with a plane wave of monochromatic light of, for simplicity, unity amplitude, and at normal incidence, the field emanating from the hologram is

$$\begin{aligned} u_{\mathrm{EM}} &= |u_0|^2 + |u|^2 + u_0|u|\mathrm{e}^{i\phi} \\ &\quad + u_0|u|\mathrm{e}^{-i\phi}, \quad (2) \end{aligned}$$

where the final term of the film transmittance has been written in terms of the two complex components. The first term of Eq. (2) is just a uniform background. The second term is a noise term sometimes called the intermodulation product term, or sometimes the self-interference term. It is produced by interference among the various scattered waves that make up the object beam. If most of the light that passes through the object is unscattered, then u_0 will be much larger than u, and the quantity $|u|^2$ will be small and can be neglected. The term $u_0|u|\mathrm{e}^{i\phi} = u_0u$ is the replication of the original wave, and projects back to form a virtual image at the position occupied originally by the object. The final term is the complex conjugate of the original wave and can be written as u_0u^*. If, for example, the original object were a point scatterer, then u is a spherical wave centered on the object point. Similarly, u_0u^* is then a spherical wave with opposite curvature and is centered on a point lying downstream from the hologram. It thus forms a real image.

The conjugate image can be disturbing, as we have noted. Since the two image terms have the same energy, the reconstructed image will generally have a signal-to-noise ratio of unity, which is unacceptable for most applications.

The twin image of holography was a serious problem that resulted in considerable effort to remove it. Many of the early papers on holography suggested ways to eliminate the twin image. The problem is more than merely an imperfection; it is in fact rather fundamental. The complete wave front (ignoring polarization effects, which usually are of no concern) can be written as $u = |u|e^{i\phi}$, a product of an amplitude (or irradiance) part and a phase term. The ordinary photographic process records only the amplitude $|u|$, being totally insensitive to phase. The holographic process records both, but incompletely. That it is a rather fundamental defect can be appreciated by considering that, just as a complex quantity u can be written in the polar form $|u|e^{i\phi}$, so it can also be written as the summation of a real term and an imaginary term, $u = |u|\cos\phi + i|u|\sin\phi$. Inspection of Eq. (1) shows that the holographic process has recorded only the real part of the field. Thus, in a basic sense the process has failed to record any more of the total wave than does conventional photography; the mix is simply different. The holographic process has recorded part of the phase, at the expense of recording only part of the amplitude. This missing information manifests itself as the twin-image problem.

Gabor had a complete solution to the problem. He proposed to make two holograms, with one having the background shifted by 90°, which can readily be done by various phase-shifting means. The second hologram would have an information-bearing term of the form $u_0|u|\sin\phi$, which is just the missing information. The two holograms would be placed in separate branches of a two-beam interferometer and upon readout, would produce two reconstructed waves that are combined by the interferometer. If the interferometer has a 90° phase shift in one branch, then the reconstructed beams would combine in what is called *quadrature*, that is, with a 90° phase shift between them, giving a complete reconstructed image wave $u_0|u|\cos\phi + iu_0|u|\sin\phi$, which is the complete field.

This solution has practical difficulties. To make a pair of phase quadrature holograms and reconstruct them in quadrature in an interferometer would require a precision of phase control rather difficult to achieve in practice. In particular, the holograms would have to be recorded on optical flats with a uniformity much better than half a wavelength of the light over the entire hologram.

It would be preferable to have the entire field on a single hologram. The Leith–Upatnieks solution [4] arises from the recognition that the recording process has an aliasing, in that a spatial frequency component, $e^{i2\pi f_1 x}$, physically manifested as a plane wave propagating at an angle θ to the normal, and another spatial frequency component, $e^{i2\pi f_1 x}$, physically manifested as a plane wave propagating at an angle $-\theta$, produce exactly the same fringe pattern when interfering with the background wave. The solution is to modulate the entire wave field u onto a carrier. Each spatial frequency component of the object should be preserved, but should be shifted in spatial frequency, so that all spatial frequencies are of the same sign. Then after recording, the ambiguity would have been removed – every spatial frequency of the object wave will produce a different fringe pattern. In the reconstruction process, all of the positive spatial frequencies will belong to one image, and all negative

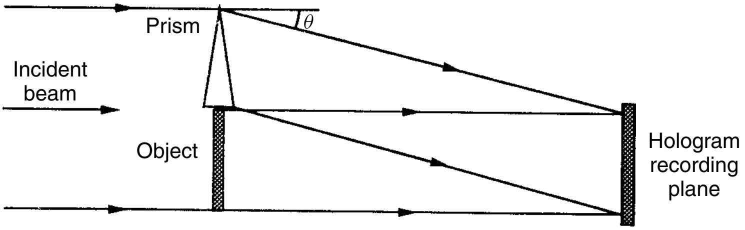

Fig. 6 Making of off-axis reference beam hologram

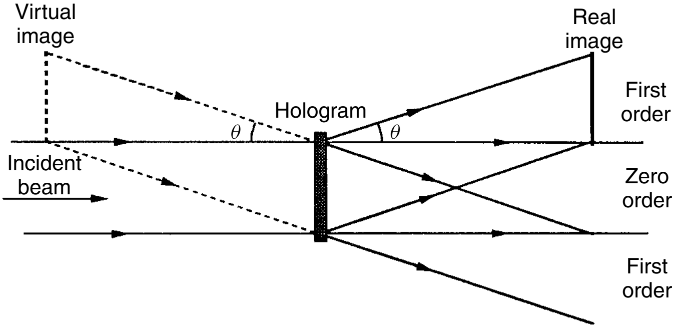

Fig. 7 Viewing of off-axis hologram. Low-quality images occur on hologram axis. High-quality images occur in the first-order diffracted waves

spatial frequencies to the other, and the images thereby become separable.

To accomplish this objective, the reference beam is introduced obliquely, as in Fig. 6. We let the reference beam be $a_0 \exp(2\pi i f_0 x)$, which describes a wave impinging obliquely on the hologram. Equation (1) then becomes

$$t = I = |u_0|^2 + |u|^2 + 2a_0|u|\cos(2\pi f_0 x - \phi). \quad (3)$$

The final term is on a spatial carrier f_0. The aliasing has been removed, in that the two previously noted spatial frequency components $\exp(2\pi i f_1 x)$ and $\exp(-2\pi i f_1 x)$ now produce distinguishable fringe patterns, one of spatial frequency $f_0 + f_1$ and the other of spatial frequency $f_0 - f_1$. In the reconstruction process, the two images, primary and conjugate, of Eq. (2) become $a_0|u|\exp[i(2\pi f_0 x - \phi)]$ and $a_0|u|\exp[-i(2\pi f_0 x + \phi)]$, which signify waves propagating in different directions. If we think of the hologram as being a diffraction grating, then the two terms become respectively the −1 and +1 orders of the grating, wherein the other extraneous term, $|u_0|^2 + |u|^2$, becomes the 0th order. Thus, as shown in Fig. 7, the beams representing the various terms become fully separated after some propagation distance. Thus, the complete wave field has been recorded and perfectly recovered.

4
The Rise of Modern Holography

The introduction of the off-axis or carrier-frequency technique in 1961 marked the beginning of the era of modern holography. In 1963, the technique was perfected and high-quality holographic imagery was reported, which created renewed interest in holography. Shortly thereafter [5], the formation of holograms of reflecting three-dimensional (3-D) objects, as opposed to the simple, two-dimensional transparencies that had been previously used in holography, was reported. In the reconstruction process, the observer looks into the hologram and sees an image that is a perfect likeness of the original object, fully three dimensional, with all of the normal parallax relations of real life. The observer, by moving his head, observes the same parallax relations that were observable in the original object scene. When looking from a near to a far point in the image, the observer must refocus his eyes. The observer can peer around objects to see what lies behind. The image, in short, is optically indistinguishable from the original object.

The hologram of the 3-D reflecting object, that is, the usual, real-world object, was made possible by the laser with its considerable coherence length compared with that of the mercury arc source that had been used in the prelaser era.

Coherence length is an important property of a light source used for holography and for interferometry in general (see INTERFEROMETRY). If a beam of light is split into two parts, which are then brought back together, the two beams have the possibility of producing interference effects. If the beams travel equal paths so that they are combined without a path difference being introduced, strong interference effects can be produced. If one beam travels a longer path than the other, the capability for interference diminishes. The amount of path difference that can be introduced without destroying the ability to interfere is called the *coherence length* and is determined by the spectral width of the source. A laser, being typically highly monochromatic, has a long coherence length, typically in the range of many centimeters. The mercury arc source, by comparison, has a coherence length of less than 1 mm.

Holograms of transparencies can easily be made with the mercury arc source, since the coherence length required for the basic holographic process is fairly small and can be handled very readily by the Hg arc source. However, the use of a reflecting, 3D object causes an increase of required coherence that can easily be several thousandfold. The problem is that in making a hologram of such an object, light reflected from all parts of the object must be coherent with the reference beam. If the object has a depth of 100 mm, which is fairly typical, then since the nearest point and the deepest points of the object, as well as all points in between, must interfere with the reference beam, the source must have a coherence length of the order of twice the object depth, or about 200 mm. This additional coherence length requirement, which has really nothing to do with the basic holographic process, is superimposed on the basic coherence requirement of the holographic process, which is typically about 0.1 mm. Thus, the transformation from 2-D objects to 3-D reflecting objects at once increases the needed coherence length to a level that can be met only with the laser.

Following the above events, interest in holography exploded, and holography became a research area in numerous laboratories throughout the world. Theory was developed, new ideas were conceived, and applications were sought.

In the meantime, Denisyuk in the Soviet Union described (in 1962) the volume hologram (see Sect. 7.2), in which the recording medium is sufficiently thick as to have a profound effect on the diffraction characteristics of the hologram. The Denisyuk, or volume hologram, is one of the main pillars of modern holography.

5
Some Early Applications

5.1
Imaging Through Inhomogeneities Using Holographic Phase Conjugation

The conjugate image of holography had traditionally been its greatest problem. Leith and Upatnieks [4] not only offered a practical way of avoiding the problem, but also proposed a use for the twin image: imagery through inhomogeneous media, such as highly aberrated lenses or scattering media, even as severe as frosted glass. Reduction to practice was reported in 1965 [6, 7].

Figure 8 shows a hologram-forming system in which a distorting medium has been placed between the object and the hologram. When the distorting medium is, as depicted, a diffuser such as ground glass, the object is not visible from the recording position. Similarly, from the hologram thus produced, the object cannot be imaged; the hologram reproduces only what can be seen from the recording plane. The hologram will, of course, form an image of the diffuser. Let the diffuser be described by a phase function $\exp[i\theta(x, y)]$, where θ is a noiselike function that describes the phase retardation introduced by the diffuser. The real image formed by the hologram is $\exp(-i\theta)$, since the real image is conjugate to the original object.

Placement of the original diffuser in the position where the holographic real image is formed, results in the superposition of the diffuser and its conjugate image, and the phase of one is cancelled by the complementary phase effects of the other or, mathematically, $\exp(-i\theta)\ \exp(i\theta) = 1$. The diffuser thus seems to disappear, and the object previously obscured by the diffuser is now revealed.

The process has an intriguing interpretation. As shown in Fig. 8, the readout beam retraces the reference beam path, but in reverse. The conjugate wave similarly retraces the original object beam path, but again in reverse. In the object beam, point

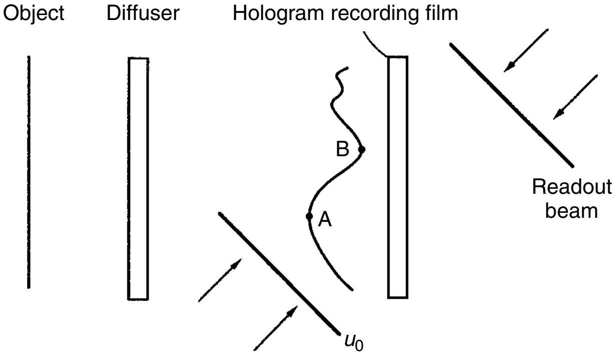

Fig. 8 Imaging through inhomogeneity, showing the hologram being made with a reference beam u_0 and subsequent readout with a beam retracing the original reference beam path

B led point A in phase. The reconstruction beam has exactly the same shape, but travels in the opposite direction. Therefore, point B now lags behind point A in phase; this is the meaning of conjugation. The wave travels over the original object beam path, and at each position along its path it assumes the same shape as the object beam did at that position, until finally it arrives at the object position, whereupon the wave forms an image. The process is somewhat analogous to time reversal. In a time-reversal situation – for example, a movie running backward – curious things happen. Scrambled eggs leap out of the pan, back into the shells, which then become whole again. Here, analogously, the wave front that had been mutilated by passage through the severe inhomogeneity is made whole again. The process has thus sometimes been called *wave-front healing.*

The process languished for many years; most of the envisioned applications involved the correction in real time of inhomogeneities that were constantly changing, such as atmospheric turbulence. Photographic film, however, needs processing and thus does not operate in real time. Phase conjugation eventually became an important area of research beginning in the 1970s, with the development of light-sensitive crystals and other devices that permitted the phase conjugation to be carried out on a much faster time scale than was possible with photographic film. Phase conjugation, today, is an important area of optics, and image formation through inhomogeneities is one of the major applications.

5.2
Complex Spatial Filtering

Holography has greatly enhanced spatial filtering technology by making practical the construction of spatial filters with arbitrary phase and amplitude control over the filter transfer function.

Spatial filtering is an optical analog of electrical filtering. An electrical signal can be decomposed into a collection of individual frequencies, and a simple filter selectively attenuates or rejects these frequencies; a very simple example is the frequency control on an audio system, which can, for example, be adjusted to emphasize the higher-frequency components of the signal. More sophisticated filters can be designed, which can not only operate on the amplitude of the frequency components but can also alter the phase relations among these components. Similarly, optical signals, such as object and image distributions, can be decomposed into spatial frequencies, and optical spatial filtering systems can be devised to alter the frequency components.

A basic filtering system is shown in Fig. 9. An object amplitude distribution s_0, typically a photographic transparency, is imaged to form the light distribution s_i. Under coherent illumination, there occurs at the common focal plane of the two lenses a light distribution that is Fourier transform–related to the object distribution. Put another way, the various spatial frequencies that constitute the

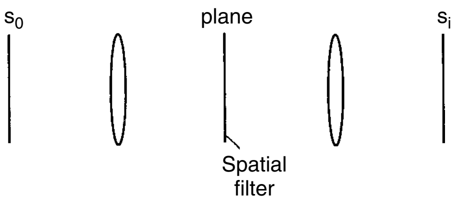

Fig. 9 A spatial filtering system. Object s_0 forms an image s_i, and the spatial frequency spectrum is modified by an attenuating and phase-shifting structure (a spatial filter) in the Fourier transform plane

object are displayed as a spectrum in that plane. For example, if the object is a simple sinusoid on a bias, $s_0 = \frac{1}{2} + \frac{1}{2}\cos 2\pi f_0 x$, the light distribution at the common focal plane consists of three points; the point on axis represents the bias term, and a pair of off-axis points represents the spatial frequency f_0. More precisely, if we write $\cos 2\pi f_0 x = \frac{1}{2}\exp(i2\pi f_0 x) + \frac{1}{2}\exp(-i2\pi f_0 x)$, that is, if we divide $\cos 2\pi f_0 x$ into its positive and negative complex frequency terms, the one off-axis point of light corresponds to the positive-frequency term and the other to the negative-frequency one.

A mask that alters the light distribution at this plane, either attenuating it or shifting its phase, is a spatial filter, and its behavior is analogous to that of an electrical filter. For example, a mask consisting of a small opening on axis is a low-pass filter, because it passes only the light representing low spatial frequencies. Similarly, a stop on axis would be a high-pass filter; such a filter in microscopy produces the well-known phenomenon of dark-field illumination. Other common spatial filters give rise to schlieren imagery and phase-contrast microscopy. In analogy with electrical systems, the optical spatial filtering system can be described in terms of a transfer function $H(f_x, f_y)$, which specifies how the system passes the various spatial frequency components (f_x, f_y).

The spatial filter mask can thus alter the object spatial frequency components by attenuating various spatial frequencies; this is accomplished by making the mask with an appropriate attenuation pattern. Similarly, the phase of the spectrum can be altered by making the mask with an arbitrary phase-delay pattern. Specifically, if the required transfer function is to be $H = |H|\exp(i\phi)$, with $|H|$ the amplitude part and ϕ the phase part, in principle, we need only to make a mask whose transmittance varies in accordance with $|H|$ and whose optical thickness varies in accordance with ϕ; the thicker the mask, the greater the phase delay.

The historical problem has been that, although amplitude control is easy, phase control is not. In principle, one could make an arbitrary phase characteristic $\exp(i\phi)$ by generating an arbitrary thickness function on some transparent material. Such precise thickness control is possible only in some simple cases, such as the Zernike phase-contrast filter.

A hologram, on the other hand, can easily produce exactly the needed transfer function, both in amplitude and in phase [8, 9]. The required transfer function $H = |H|\exp(i\phi)$ can be produced as a hologram having the transmittance $t = s_b + |H(f_x, f_y)|\cos[2\pi\alpha f_x + \phi(f_x, f_y)]$, where s_b is a bias term and where we use f_x and f_y as the variables, since the hologram is to be placed in a plane that displays the spatial frequency spectrum. The operation of the filter is described with reference to Fig. 10. The object s_0 is Fourier transformed, and the light emerging from the spatial filter mask H represents the product $S_0 t$, where S_0 is the Fourier transform of s_0. In expanded form, this product is written as

$$\begin{aligned} S_0 t = {} & S_0 s_b + \tfrac{1}{2} S_0 H \exp(2\pi i\alpha f_x) \\ & + \tfrac{1}{2} S_0 H^* \exp(-2\pi i\alpha f_x), \end{aligned} \qquad (4)$$

where the cosine term is written in exponential form.

The first term is of no interest, it simply represents a uniform filter that passes all spatial frequencies with uniform attenuation. The second term constitutes the object acted on by the complex filter $H = |H|\exp(i\phi)$ and multiplied by a factor $\exp(2\pi i\alpha f_x)$. Next, we see the usefulness

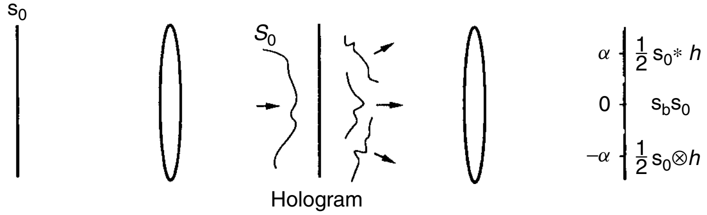

Fig. 10 Use of a hologram as a spatial filter

of this final factor. Similarly, the final term represents the object acted upon by a conjugate filter, $H^* = |H| \exp(-i\phi)$, and multiplied by a term $\exp(-2\pi i f \alpha_x)$. Thus, the holographic process has succeeded in producing an arbitrary transfer function H, as well as the conjugate transfer function H^*. We show that these three terms are physically separated in the image plane, just as the basic terms of the conventional hologram are separated. The second lens Fourier transforms these terms, yielding an image distribution

$$\begin{aligned} I_{\mathrm{im}} &= \mathrm{s_b s_0}|_{\text{centered on axis}} \\ &\quad + \tfrac{1}{2}\mathrm{s}_0^* h|_{\text{centered on }} x = a + \tfrac{1}{2}\mathrm{s_0} \\ &\quad \otimes h|_{\text{centered on }} x = -a, \qquad (5) \end{aligned}$$

where we have used some basic Fourier transform relations in obtaining this result. Here, * represents the well-known convolution, and ⊗ represents cross correlation. The latter is a mathematical method of comparing the similarity of two functions. Specifically, it is defined as

$$\begin{aligned} \mathrm{s_0} \otimes h = \int\!\!\int \mathrm{s_0}(x' + x, y' + y) \\ \times h(x', y')\, \mathrm{d}x'\, \mathrm{d}y'. \qquad (6) \end{aligned}$$

If $\mathrm{s_0}$ and h are very similar, then the integral will have a large value when $x = y = 0$, that is, when $\mathrm{s_0}$ and h are in alignment. We say then that $\mathrm{s_0}$ and h are highly correlated.

The term $\exp(2\pi i\alpha f_x)$ has resulted in a displacement α, a result following from a well-known theorem of Fourier transforms, namely, that linear phase shifts Fourier transform into displacements. From a physical viewpoint, the linear phase-shift terms represent wave-front tilts, and, consequently, the lens brings them to a focus at an off-axis location. If α is sufficiently large, the three terms will be completely separated, just as the various images are completely separated in conventional holography.

The matched filter is of special interest. A filter is said to be matched to a signal $\mathrm{s_0}$ if its transfer function is the complex conjugate of the object Fourier transform, or $H = S_0^*$. Matched filters are used for recognizing the presence of a particular signal that may be immersed in noise or other signals and therefore not readily seen. Pattern recognition is a common application of matched filtering.

To construct a matched filter for an object $\mathrm{s_0}$, a Fourier transform hologram is made of that object. The hologram is then used as a spatial filter, just as we have described above, with $H = S_0$. The output image is

$$\begin{aligned} i_{\mathrm{im}} \mathrm{s_0 s_b}|_{\text{centered on axis}} + \tfrac{1}{2}\mathrm{s_0} * \mathrm{s_0}|_{\text{centered on } \alpha} \\ + \tfrac{1}{2}\mathrm{s_0} \otimes \mathrm{s_0}|_{\text{centered on} -\alpha}. \qquad (7) \end{aligned}$$

The second term, generally uninteresting, is the self-convolution of the object.

The final term is the autocorrelation function of the object. It is typically a small, bright spot of light located at a position on the image corresponding to the position of the object.

Figure 11 shows experimental results using a matched filter of the letter *q*. In Fig. 11(a), we show a transparency containing an array of nine letters and numbers (alphanumerics), including *q*. The transparency is placed in the system shown in Fig. 10, with the aperture restricted so that only the letter *q* is illuminated. The Fourier transform of *q* [Fig. 11(b)] is formed at the frequency plane. A hologram of this distribution is formed by bringing in a reference beam. The hologram after development is placed back into its original position, and the array of nine letters is imaged through the hologram. Figure 11(c) shows an image of the correlation-forming first diffracted order. Each letter has been imaged as a light distribution that represents the cross correlation of that letter with the letter *q*. Since the letter *q* correlates strongly with itself, we see a very strong, concentrated spot of light at the position for *q*. The other letters are considerably different in structure from *q*, and the correlation is therefore much lower. A threshold detector, adjusted to respond only when the light level rises above some predetermined value, could then automatically recognize the presence of the letter *q*.

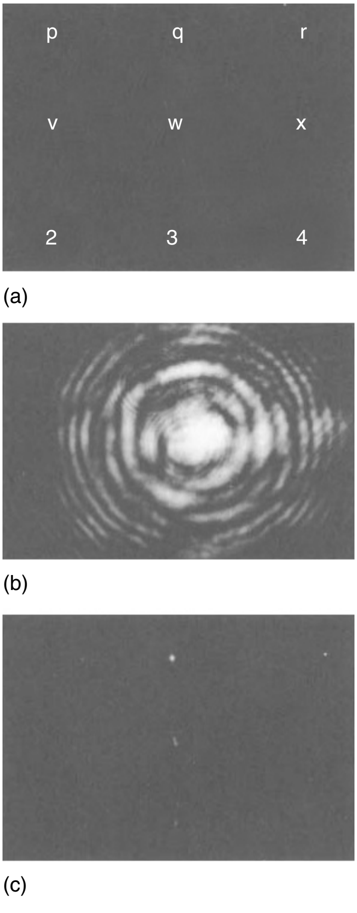

Fig. 11 Experimental results of matched filtering: (a) object, an array of numbers and letters; (b) Fourier transform of letter *q*; and (c) image formed by imaging object through the matched filter for the letter *q* and viewed in the correlation sideband of the filter

To give insight, a physical way of describing the operation of a matched filter is shown in Fig. 12. For simplicity, we show only the part of the filter mask. The spatial frequency spectrum S_0 of the object impinges on the filter mask as a wave front with some complex amplitude and phase distribution, the latter being depicted as wave-front irregularities. The wave passes through the spatial filter, where each part of the wave is phase adjusted so as to eliminate the irregularities. The wave emerges from the filter with uniform phase, essentially a plane wave; although, in general, not with uniform amplitude. This plane wave is then focused by the second lens into a small spot that is in fact the autocorrelation function we have previously described.

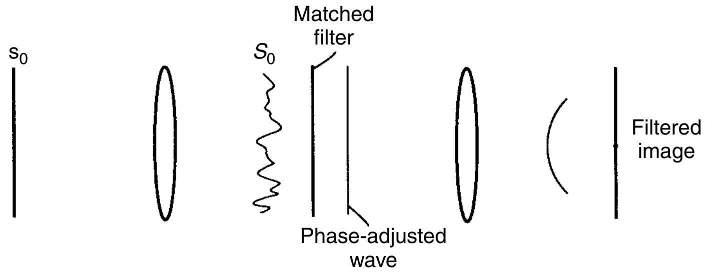

Fig. 12 Operation of a matched filter

The use of a hologram as a complex filter and, in particular, of a Fourier transform hologram of an object as a matched filter for that object was developed at the University of Michigan Willow Run Laboratories in the period 1961–1963 under military secrecy. The initial suggestion was given by C. J. Palermo and the initial reduction to practice was by Kozma and Kelly [9], who produced a matched filter for compression of a coded pulse of the sort used in radar systems. VanderLugt [8] extended the technique to pattern recognition and character reading, and in the process made some very sophisticated matched filters.

5.3
Hologram Interferometry

Hologram interferometry is another major application used primarily in scientific and engineering areas. There are three principal forms of hologram interferometry: time average, double exposure, and real time.

The earliest form, time average, was first reported by Powell and Stetson [10], who made holograms of vibrating objects. The reconstructed images were found to be overlain with light and dark bands. Such an occurrence is readily explained. It is well known that holography, being an interference process between two beams, requires stability in order that the fringes remain stationary. Thus, if the object moves during the exposure, the fringes will also move and will tend to be washed out, especially those fringes originating from portions of the object where the motion is greatest. Thus, motion of the object results in dark regions on the holographic image. A pessimistic assessment would conclude that the hologram is ruined. On the other hand, one might recognize that the banded pattern superimposed on the holographic image contains valuable information about how the object moved during the exposure.

In the case of the vibrating object, the pattern overlying the object depicts the vibration amplitude. In regions where the object is stationary (the nodes of the vibration pattern), the image is bright. Other object regions reconstruct less brightly. The strength of the reconstructed image is related to the vibration amplitude by a zero-order Bessel function, which is oscillatory in nature. Hence, the image is overlaid by a fringe pattern, the fringe peaks and valleys corresponding to contours of constant amplitude (Fig. 13). Thus, inspection of the image enables the vibrational amplitude of each point on the image to be accurately determined.

Double exposure is a second form of hologram interferometry. Typically, a

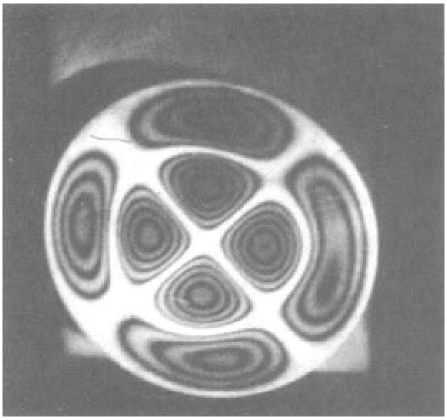

Fig. 13 Holographic image of a vibrating object, a flat plate held stationary at circumference

hologram is made of an object in an undistorted state, and then the object is distorted slightly, for example, by being stressed or heated. Then, a second exposure is made on the same photograph. On readout, two reconstruction fields are produced, one forming an image of the undistorted object, the other of the distorted object. The two wave fields interfere, forming an image with an overlay of fringes. In places where the object underwent no displacement, the two wave fields will be identical and will add in phase; the reconstruction will therefore be bright. At a position where the object shifted slightly, so that the two image waves are out of phase, destructive interference occurs and the image will be dark. Thus, the fringe pattern contains information about how the object distorted in response to the disturbance.

In the third basic form of hologram interferometry, a hologram is made of an object; then the hologram is placed back in its original position, while the object remains in its original position under continued illumination, and the reconstructed wave interferes with the object wave. The observer looking through the hologram thus sees a superposition of the object and the reconstructed image. Disturbing the object, by pressure, heating, and so on, produces a mismatch between the object wave and the reconstructed wave, causing a fringe pattern to be superimposed on the observed object–image combination. The result is akin to double-exposure hologram interferometry, except that the interference process occurs in real time, and the object can be distorted at will in many different ways. This process is appropriately called *real-time hologram interferometry*.

From the viewpoint of traditional interferometry, hologram interferometry is most remarkable. No longer is it necessary to bring two wave fields simultaneously into superposition in order to produce interference between them. The two waves can exit at different times or different places, and they can effectively be made to interfere simply by recording them holographically, thereby preserving both the amplitude and phase of each of the waves. In the context of conventional interferometry, such a process is absurd; interference could never be brought about by first recording one wave front and then at a later time and at another place, recording a second wave on the same film. Yet holography accomplishes exactly this feat.

Hologram interferometry, since its invention in 1965, has become a tool for science and engineering; its practical uses include the study of vibrations in automobile engines, the detection of defects such as cracks and debonds in airplane wings, the testing of tires, and the study of vibrational modes in musical instruments, such as violins.

Nowadays, hologram interferometry is generally carried out with a CCD (charge coupled device) camera as the recording medium, with computer reconstruction;

this method has significant advantages over photographic film. The recording can be done at lower light levels, and the hologram data can be available within a few seconds. The problems of wet film processing, including development, fixing, and drying are thus avoided. Also, with the computer, sophisticated image-processing procedures can be carried out. The CCD recording system is incorporated into an instrument consisting of a laser, beam splitter, and so on that are the usual parts of a hologram-forming system. The entire process can be carried out by persons unskilled in optics. Such systems are lightweight, highly portable, and relatively inexpensive. They have made hologram interferometer systems attractive for outside the laboratory, for example, in factory environments. The limitation of the CCD camera, which is of low resolution, is not a significant problem for hologram interferometry, since the large resolution requirements for holography are typically not present in hologram interferometry applications. The high-resolution requirements of holography arise from the recording of three-dimensional images with full parallax, and in hologram interferometry one is generally not required to form images that exhibit three dimensionality and parallax.

5.4
Holographic Optical Elements

Holograms have found a use very different from the production of fascinating three-dimensional imagery. They can be used as a special kind of optical element, analogous to lenses, but operating on the principle of diffraction instead of refraction.

A hologram of a point object made with a collimated reference beam can be illuminated with a collimated beam that duplicates the reference beam, whereupon a virtual point image is formed. Alternatively, the hologram can be illuminated with a conjugate reference beam that is collimated but counterpropagating, and the hologram will form a real point image. An alternative way of viewing the reconstruction process is to consider that the hologram is acting like a lens, bringing a collimated beam to a focus. Similarly, illuminating the hologram with a duplicate of the object beam reproduces the collimated reference beam. Now the hologram is a collimating lens. Finally, the hologram can be illuminated by light reflected from an extended object, and it will form an image of this object, just as does a lens. This phenomenon is not surprising if one considers that the hologram of a point object is just a Fresnel zone-plate lens, a commonplace device described in basic texts on physical optics.

As a lens, the point-object hologram is hardly competitive with a conventional lens since it has, because of its diffractive nature, severe chromatic aberrations, and therefore it is suitable only for use with monochromatic light. Besides, the zone-plate lens has all the basic aberrations of a simple refractive lens, such as spherical aberration, coma, and so on.

However, there are many specialty niches for the holographic optical elements. Diffractive elements disperse light in a way opposite to refractive lenses; that is, red is bent the most in a diffraction process, whereas the refraction process bends violet the most. Therefore, the hologram can sometimes be used to correct the dispersion of a glass lens, producing an achromatic combination [11]. In some situations, this method can be more effective than the usual method of using glasses of different dispersiveness for color correction. The holographic lens is especially effective in combination with refractive lenses to produce a compound Fourier transforming lens in which the dispersion requirements are quite different from those of image formation. Holograms can also be used as corrector plates to remove aberrations from a lens system [12]. Diffraction gratings are nowadays often made by recording on photoresists the interference of two collimated beams of light; these are usually called *holographic gratings*, although it is questionable whether the recorded patterns of interference of two plane waves should be called *holograms*. However, if one of the two beams contains some special irregularity that is to be incorporated into the grating, then hologram would be an appropriate term.

6
Volume Holography

The hologram has generally been treated as a two-dimensional structure with negligible thickness. Denisyuk [13] and van Heerden [14] departed from this view by describing holograms in which the thickness of the photographic emulsion, or other recording material, is sufficiently great to influence the diffraction properties of the hologram. When the recording material is sufficiently thick, the diffraction process is of a type known as *Bragg diffraction*.

The interference fringes formed when two light waves interfere are generally thought of as a two-dimensional pattern of light and dark lines, perhaps straight, perhaps curved. This is the way they are seen when viewed on a screen placed in the light path. However, it should be recognized that the interference typically occurs throughout the entire volume of space where the beams overlap, and the fringes are in fact three dimensional, being surfaces of brightness interspersed with surfaces of darkness. When the recording medium is thick, the recorded fringes become surfaces within the material. Figure 14 shows the two basic regimes of such recorded fringes. In the first, the two interfering beams impinge from the same side of the recording medium; in the second, from opposite sides. The fringes bisect the angle between the two beams and have a spacing that depends on the angle between the two beams. In the first regime, the fringes thus cut across the recording material and may have a spacing from very coarse to about λ, the wavelength of the illumination. In the second regime, the fringes are fine, on the order of $\lambda/2$ in spacing, and lie more nearly parallel to the emulsion surface, being exactly parallel to the surface when the beams are traveling exactly in opposite directions and normal to the surface. The criterion that distinguishes a thick recorded fringe pattern (i.e., a thick hologram) from a thin one is roughly that to be thick, the spacing between the fringes should be less than the thickness of the emulsion or other recording material. When two beams interfere with an angle of 5° between them and are recorded by a

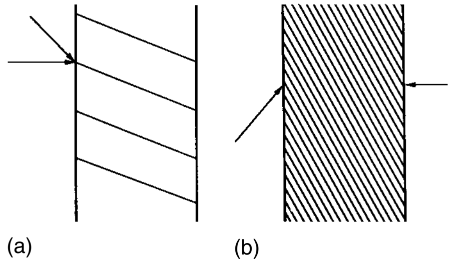

Fig. 14 Two regimes of volume-recorded fringes: (a) two beams impinging from same side of the recording medium; (b) two beams impinging from opposite sides of the recording medium

photographic emulsion of thickness 6 μm, a typical thickness, the fringe spacing is much greater than the thickness, and the law of diffraction for thin gratings holds. At an angle of 30°, the recorded fringes will usually be sufficiently fine to produce Bragg diffraction.

Bragg diffraction can be described in various ways. A simple approach is to consider it to evolve from conventional diffraction. Figure 15 shows diffraction from a thin grating. The incident beam strikes the grating at an angle θ_i to the normal and is diffracted at an angle θ_r from the normal. The diffraction angle can be calculated from the basic consideration that the first diffracted order occurs at an angle such that the optical path from a point on the impinging wave front to a point on the diffracted wave front is one full wavelength longer (or shorter) through one slit opening than through the adjacent one, from which we readily obtain $d_1 + d_2 = \lambda$, or $\sin\theta_\mathrm{i} + \sin\theta_\mathrm{r} = \lambda/d$, where d is the grating spacing. For the other diffracted order, the result is similar, except that one of the sine terms is negative. In either case, we require constructive interference between the light rays from the various openings. Examination of this diffraction equation shows that for a given angle of incidence, there will generally be a diffracted wave, whose angle changes with the incidence angle.

Now, suppose we let the emulsion become thick, and, for simplicity, let the fringes be normal to the emulsion surface (Fig. 16). The recorded fringes become surfaces. We consider each surface to be like a mirror that reflects some of the light incident on it, while the remainder is transmitted and undergoes partial reflection at the next surface. The diffraction process now takes on reflection properties. However, the same process that led to the previous diffraction equation still holds; the diffracted wave is still at the angle that leads to constructive interference from adjacent openings. Now, however, the openings have become surfaces that behave like mirrors, so that now the mirror law, that the angle of reflection equals the angle of incidence, must also apply. Thus we require also that $\theta_\mathrm{i} = \theta_\mathrm{r}$, causing the diffraction equation to become $2\sin\theta_\mathrm{i} = \lambda/d$, which describes Bragg diffraction. According to this equation, a beam of light falling on

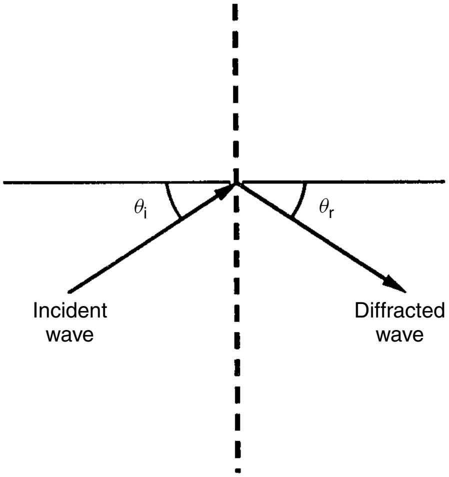

Fig. 15 Diffraction by a grating, arbitrary angle of incidence

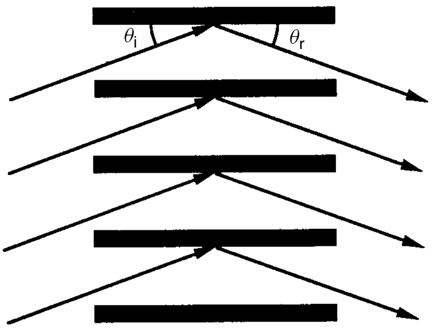

Fig. 16 Diffraction from a thick grating (Bragg diffraction)

the diffracting structure at an angle θ_i produces a diffracted wave if the equation is satisfied, but if the angle θ_i is changed, the diffracted wave does not change direction; it simply ceases to exist since the Bragg equation will not be satisfied. Thus, a Bragg-regime hologram must be illuminated at the proper angle in order to produce a diffracted wave. This effect is readily observed with holograms illuminated with a monochromatic source such as a laser beam. At the proper angle the image is bright, but rotation of the hologram by perhaps 5° or 10° will result in the disappearance of the image.

Alternatively, if the hologram is illuminated by a beam of white light, then in a given direction of viewing, only a narrow band of wavelengths will satisfy the Bragg angle close enough to give diffraction. Thus, the hologram can form an acceptably sharp image when viewed with white light; the hologram simply selects a narrow band of light for forming the image, while discarding the remainder.

In order that the band $\Delta\lambda$ of diffracted light be sufficiently narrow to form an acceptably sharp image, the hologram should be made with the object and reference beams coming from opposite sides of the film. Theory shows that the spectral width of the diffracted image depends on the number of fringe surfaces that contribute to a diffracted ray, in accordance with the equation $\Delta\lambda/\lambda = 1/N$, where N is the number of contributing surfaces and λ is the midband wavelength of the diffracted light (Fig. 17). This number is greatest when the reference and object beams are introduced from opposite sides since the fringes are finer, and the incident beam encounters more partially reflecting surfaces as it propagates through the emulsion.

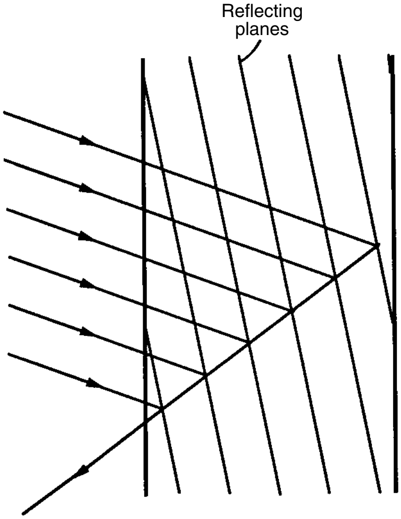

Fig. 17 Bragg diffraction from many surfaces. The diffracted ray is formed from many incident rays

Another characteristic of the hologram made with the beams coming from opposite sides of the plate is that the image is viewed from the same side of the hologram from which the illumination comes, as seen in Fig. 15. Hence, such holograms are sometimes described as *volume reflection holograms*. More often they are known as *Denisyuk holograms*. The reflection direction of the holograms can readily be seen by noting that the fringes are almost parallel to the emulsion; hence, considered as mirrors, they reflect the light back toward the source.

The most common way to make a Denisyuk hologram is shown in Fig. 18. A beam of laser light shines through a photographic plate to illuminate an object located on the other side. Light reflected from the object then is reflected back to the plate. Thus, the same beam of light exposes the plate twice, serving as the reference beam on the first pass

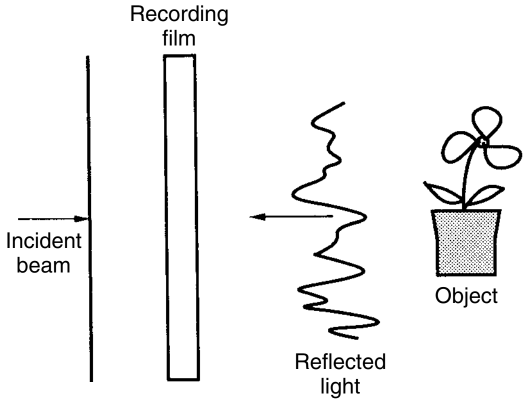

Fig. 18 Making of a Denisyuk hologram

and as the object beam on the second, a configuration that is both simple and effective.

7
Display Holography

7.1
The Beginnings

Display holography began in 1963, when holograms of 3D, diffusely reflecting objects were first made, as opposed to simple transparencies as objects. Such holograms were the first that could be viewed without the aid of eyepieces. Images such as these had never been produced previously, and people viewed them in astonishment.

However, it took many years for display holography to develop into a viable commercial activity. There were two principal problems. First, the early holograms had to be viewed with coherent sources, such as lasers or mercury arcs. Such sources were rather expensive. Second, the holograms themselves were rather expensive, since there were no ways to mass produce them. A hologram was made either as an original or by photographically contact printing from a master. In recent years, both limitations have been overcome, with the result that display holography, in many different forms, has prospered.

7.2
The Denisyuk Hologram

To be a successful display device, the hologram had to be white-light viewable. The first white-light viewable hologram, the volume reflection hologram of Denisyuk, had been described as early as 1962 [13], and high-quality Denisyuk holograms had been demonstrated by 1965. However, the early Denisyuk holograms lacked brightness, and the brightest ones were often afflicted with high levels of scattered light. The problem was one of chemistry; techniques of photographic development were required that would produce high diffraction efficiency with low noise scatter. These problems were gradually overcome over a period of many years, and today, bright, crisp, low-noise Denisyuk holograms are commonplace. A variety of recording media exist today for producing high-quality Denisyuk holograms. They can be made with very fine-grain emulsions and very special development techniques that have been perfected over many years. In addition, non–silver halide materials, such as dichromated gelatin and photopolymers, compete with silver halide and have captured a respectable percentage of the Denisyuk display hologram market. Denisyuk display holograms are generally the most expensive, since they are not easily mass produced. However, they certainly are the highest quality of display hologram. They seldom appear on covers of magazines, but rather are typically displayed in galleries and exhibitions.

The quality of the hologram is strongly dependent on the recording medium, especially for the Denisyuk holograms, since such holograms are the pinnacle of hologram perfection. The principal factors for judging hologram quality are: the brightness, or diffraction efficiency, the level of scattered light, or ambient background, and for color holograms, the degree of color saturation. Diffraction efficiency in holograms recorded on silver halide photographic plates is determined by the form of chemical processing used in the development and bleaching procedures. In general, as the diffraction efficiency is increased, the scattered light level increases, so the aim is to produce the highest diffraction efficiency in combination with the lowest light scatter. With color holography, the problem is especially critical since the blue light produces significantly higher scatter than the longer wavelengths. New photographic emulsions coming into use have silver grains smaller than the wavelength of light, and consequently provide extremely low levels of scatter, resulting in holograms of significantly higher quality that were available a decade ago. In such holograms, a portion of the image that should be black appears in these new holograms as being very black indeed.

7.3
The Benton Hologram

In 1969, a second kind of white-light viewable hologram, known as either the *rainbow* or *Benton hologram*, was invented [15]. The construction method has two steps (Fig. 19). First, a conventional laser transmission hologram is made. Next, a second hologram is made, using as the object the real image from the first hologram. In this process, a narrow horizontal slit is placed

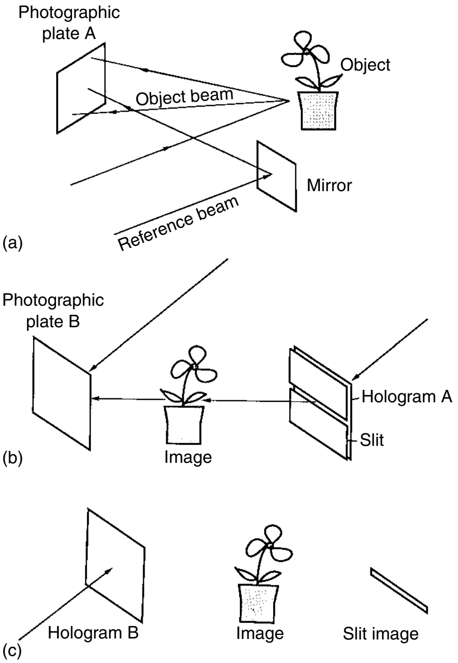

Fig. 19 Making and viewing of a Benton hologram. (a) First step, making a conventional hologram. (b) Second step, making a second hologram from real image of the first hologram. (c) Illumination of the second hologram (with monochromatic light). An image is formed both of the object and of the slit. Illumination with polychromatic light forms a spectrum at the slit image position. An observer places his eye somewhere in the spectrum; consequently, he sees an image in monochromatic light, and the image is therefore sharp

at the first hologram, thus eliminating all parallax in the vertical plane. The second hologram, under coherent illumination, forms an image of the original object, either virtual or real, as well as a real image of the slit. To see the image, the observer must place his eye at this slit image. If he moves his head vertically, the image disappears; hence there is no vertical parallax. The image, however, can be quite bright; all the light that ordinarily would be spread over a vertical distance of perhaps 1 m at the observer's position is now confined to the slit image, and the brightness increase can be several hundredfold.

There is, however, an even more important advantage to this method: white-light viewability. If one were to illuminate the hologram with light of two different wavelengths, the lateral dispersion would result in two vertically displaced slit images, one for each color. The observer would see the entire image in one color when he positioned his eyes at one slit image and in the other color when he positioned them at the other slit image. Extrapolating this process to white-light illumination, where the slit image would be smeared into a continuous spectrum, one finds that to view the hologram the observer may position his eyes at any location on the spectrum, whereupon he will see a sharp, clear image in the color corresponding to that part of the spectrum. As he moves his head vertically, the image changes color but remains sharp and clear. If his head is not in the image plane of the slit, he will still see the entire image, but the color will change from top to bottom. Appropriately, such holograms have been called *rainbow holograms*.

In exchange for giving up vertical parallax while retaining normal horizontal parallax, one has obtained holograms that can be viewed in white light without the thick emulsion requirement. The Benton hologram thus offers a favorable tradeoff, in which the comparatively unimportant feature of vertical parallax is exchanged for the important white-light viewing capability.

7.4

The Multiplex and Related Holograms

Another important development, the composite or multiplex hologram, overcomes some of the basic limitations of the more conventional holographic techniques. Here, a hologram is synthesized from a large number of ordinary photographs made in a conventional way. The pictures, made from different positions, constitute many views of the object, and they have in total all the essential information contained in a hologram. The composite method synthesizes the many pictures into a single hologram.

Composite holograms were first reported by various groups, including Pole [16], McCrickerd and George [17], and DeBettito [18]. In the next few years, many advances were made, culminating in a system perfected by Lloyd Cross and his colleagues. In the Cross method, the object is placed on a slowly rotating platform and a motion picture camera makes hundreds of pictures, just as in conventional cinematography. The pictures could then be made into a single hologram as follows. The first picture could be imaged onto a screen consisting of a piece of frosted glass. The light transmitted by the screen would be recorded as a hologram in the form of a long thin strip about 1 mm wide and perhaps 20 cm long, the purpose of the frosted glass being to direct light from all parts of the image to a recording slit that controls the width of this hologram. The actual system differs from this in that, instead of the diffuser, a large cylindrical lens at the image plane is used for directing the light to the slit; this method uses the light more efficiently and reduces laser speckle. After one frame is recorded as a hologram, the next frame is similarly projected, the recording film is moved one slit width, and a second hologram is made in an adjacent recording position, and so on.

The resulting hologram, the collection of all of the individual slit holograms, is then wrapped into an arc corresponding to the angle of rotation of the object and is illuminated as a conventional hologram. Each component hologram projects a virtual image without depth or parallax, just like the image from which the hologram was made. When an observer views the filmstrip, however, each eye looks at an image point through different holograms. The image formed on the retina of each eye is slightly different, as in a stereoscopic viewer, and consequently a three-dimensional image is perceived. As the observer moves his head laterally, different images are seen, thereby producing parallax. The Cross technique can be combined with the Benton technique, thus allowing white-light viewing. Essentially nothing is given up in so doing, since the composite holograms lack vertical parallax anyway.

In a variant of this method, the recording camera is translated laterally instead of the object rotating, resulting in what is called a *flat integral*, since the hologram is no longer curved into an arc for viewing. This flat integral can be used as a master for making a second hologram, where the image can be brought into the plane of the hologram, thus relaxing the viewing light source size requirements and giving a sharper image.

7.5
Construction of Commercial Display Holograms

Commercial holograms, whether for display on magazine covers, on credit cards, or in galleries, should be white light viewable. Various standard techniques have been developed for the production of commercial display holograms, all based on either the Denisyuk or Benton white-light viewing methods.

Typically, a master hologram is made as a conventional laser transmission hologram. The conjugate image may then be used as an object to produce a volume reflection hologram. The resulting hologram has an image that straddles the hologram, being partly in front of it and partly behind it, thus giving a sharper image than if the image were entirely on one side of the hologram. One of the limitations of the white-light-viewable hologram is that the image sharpness decreases with the distance of the image from the plane of the hologram.

Portrait holography can be done using this method, in which case the master hologram will have to be made with a short-pulsed laser so as to freeze object motions, since a living person cannot ever be stationary enough to be holographed, unless the exposure time is much less than 1 ms.

For inexpensive, mass-produced holograms, such as those found on magazine covers, and so on, the Benton technique is used, since Benton holograms do not require any volume effects for white-light readout and, consequently, can be readily produced by embossing methods, which are inherently surface phenomena.

To mass produce a Benton hologram, the original laser-produced master hologram is used to form a real image, which is then used as the object for making a second hologram of the Benton type. This hologram is recorded on photoresist, which when developed forms a surface relief pattern. This hologram is placed in a plating bath, where nickel is plated onto the hologram surface; when the photoresist is removed, a surface relief hologram made of nickel remains. This can be used in an embossing process, where the hologram is mounted in a press and used to stamp the surface relief pattern onto a sheet of heated plastic. Such holograms can be produced at a fast rate, thousands per hour, and the total production run can be in hundreds of thousands or even millions. These holograms can be bonded to paper by hot-stamping methods. This mass production method has made display holography a significant business.

8

Holography Today

Holography today is a relatively small but thriving and expanding business. Display holography is probably the leading commercial application area. Artists are employing holography as a new medium, and their works are displayed in museums and galleries. Increasingly, color holograms, as opposed to monochrome, are being seen.

The largest commercial activity is the mass production of Benton-type holograms in magazines, in advertising, for Christmas cards, for baseball cards, and so on, and as security devices on valuable objects such as credit cards and money.

Holographic optical elements are of ever-increasing importance in such applications as diffraction gratings and beam splitters. Head-up displays on fighter aircraft commonly use holograms. Holograms play a significant role in the development of all-optical or hybrid-optical electronic digital computers, an area of current intense research. Holograms have potential as memory elements in such computers, and as interconnections to connect various storage and processing elements of the computer. Other factors being equal, the more interconnections, the more powerful the computer, and holograms offer the possibility of an enormous number of interconnects in a very small volume.

Hologram interferometry continues as a solid area, with use in scientific and engineering applications. It shares its application areas with other new, laser-based types of interferometry, such as speckle interferometry and shearography.

Various other minor applications of holography exist that do not fall into these categories, but have the potential for becoming major applications. The outlook for further holographic application development is bright.

Glossary

Amplitude: For a light wave, the magnitude of the electric field, or equivalently, the square root of the irradiance (or intensity).

Bragg Diffraction: Diffraction from a volume, or thick, grating, characterized by a diffracted beam being produced only for certain incidence angles of the illuminating beam.

Coherence Length: The path difference allowable between two interfering beams before interference no longer occurs.

Coherent Light: Light that is monochromatic and derived from a point source, or equivalently, which can be focused to a point. Laser light is highly coherent; nonlaser light (e.g., Hg arc source light) can be made moderately coherent by passing it through a filter and a pinhole.

Composite Hologram: A hologram formed from a number of individual subholograms, each formed from a separate image.

Convolution: The convolution of two functions $g(x, y)$ and $h(x, y)$ is the mathematical operation

$$g(x, y) * h(x, y) = \iint_{-\infty}^{\infty} g(\alpha, \beta) \times h(x - \alpha, y - \beta)\, \mathrm{d}\alpha\, \mathrm{d}\beta$$

When an object is imaged through a spatial filter, the resulting image is the convolution of the object with the point spread function h of the filter, where the point spread function is the image produced by a point object.

Cross Correlation: Somewhat similar to a convolution, the cross correlation of two functions $g(x, y)$ and $h(x, y)$ is the mathematical operation.

$$\iint_{-\infty}^{\infty} g(\alpha, \beta) h^*(\alpha - x, \beta - y)\, \mathrm{d}\alpha\, \mathrm{d}\beta.$$

Diffraction: The tendency of light to spread out when passing through narrow apertures, or to bend when passing near an obstruction, such as a knife edge.

Fourier Transform: Mathematically, the Fourier transform of the spatial function $g(x, y)$ is defined as the integral

$$G(f_x, f_y) = \iint_{-\infty}^{\infty} g(x, y) \times \exp[-i2\pi(f_x x + f_y y)\, \mathrm{d}x\, \mathrm{d}y,$$

where f_x, f_y are new variables resulting from the transformation. More physically, a decomposition of the signal $g(x, y)$ into its spatial frequency components.

Frequency Plane: A plane in a coherently illuminated optical system where the light distribution is Fourier transform–related to the object light distribution.

Hologram Interferometry: A form of holography in which multiple images (two or more) of the same object are stored on the same recording medium, and in which, on readout, interference effects occur among the various recorded images.

Interference: The tendency of two or more light beams in superposition to combine in a reinforcing way at some positions to form a resultant intensity greater than the sum of the two intensities individually, and to produce cancellation effects at other positions.

Matched Filter: An important type of filter, used for detecting the presence of signals immersed in noise, clutter, or competing signals.

Multiplex Holography: A particular type of composite hologram, distinguished primarily by its cylindrical shape.

Phase: For a signal $a \cos(2\pi f_0 \tau + \phi)$, a is the amplitude, ϕ is the phase, and f_0 is the carrier frequency. The phase may be roughly described physically as the shape of the wavefront.

Phase Conjugation: The operation of changing a complex signal $u = a(x, y) \times \exp[i\phi(x, y)]$ into its complex conjugate $a(x, y)\exp[-i\phi(x, y)]$.

Spatial Filter: A mask placed in a plane in an optical imaging system where the light distribution is Fourier transform–related to the object distribution. The mask modifies the spatial frequency content of the image by attenuating or phase shifting the various spatial frequency components of the object.

Transfer Function: The transfer function of a system describes the way various frequency components of a signal pass through the system, that is, how much attenuation and phase shift the frequency components undergo. The system may be quite complicated, being formed from many elements, or it may be, in the simplest case, simply a filter.

References

[1] Gabor, D. (**1948**), *Nature* **161**, 777–778.
[2] Gabor, D. (**1949**), *Proc. R. Soc.* **A197**, 454–487.
[3] Gabor, D. (**1951**), *Proc. Phys. Soc.* **B64**, 449–469.
[4] Leith, E., Upatnieks, J. (**1962**), *J. Opt. Soc. Am.* **52**, 1123–1130.
[5] Leith, E., Upatnieks, J. (**1964**), *J. Opt. Soc. Am.* **54**, 1295–1301.
[6] Leith, E., Upatnieks, J. (**1965**), *SPIE J.* **4**, 3–6.
[7] Kogelnik, H. (**1965**), *Bell Syst. Tech. J.* **44**, 2451–1455.
[8] VanderLugt, A. (**1964**), *IEEE Trans. Inform. Theory* **IT-10**, 139–145.
[9] Kozma, A., Kelly, D. (**1964**), *J. Opt. Soc. Am.* **54**, 1395.
[10] Powell, R., Stetson, K. (**1965**), *J. Opt. Soc. Am.* **55**, 1593–1597.
[11] Morris, G. M., George, N. (**1980**), *Opt. Lett.* **5**, 446–448.
[12] Veldkamp, J., Swanson, G. J. (**1983**), *Proc. SPIE* **437**, 54–59.
[13] Denisyuk, Y. (**1962**), *Sov. Phys. Dokl.* **7**, 543–545.
[14] van Heerden, P. J. (**1963**), *Appl. Opt.* **2**, 387–392.
[15] Benton, S. (**1969**), *J. Opt. Soc. Am.* **59**, 1545.
[16] Pole, R. V. (**1972**), *Appl. Phys. Lett.* **10**, 20–22.
[17] McCrickerd, J. T., George, N. (**1968**), *Appl. Phys. Lett.* **2**, 10–12.
[18] De Bitteto, D. J. (**1969**), *Appl. Opt.* **8**, 1740–1741.

Further Reading

Abramson, N. (1981), *The Making and Evaluation of Holograms*. New York: Academic Press. Primarily deals with holographic interferometry, with emphasis on the author's own work. Full of very clever techniques.

Beiser, L. (1984), *Holographic Scanning*. New York: Wiley. Deals with a specialized application of holography.

Bjelkhagen, H. (1995), *Silver Halide Recording Materials for Holography and Its Processing*. Berlin: Springer.

Caulfield, H. J. (Ed.) (1979), *Handbook of Optical Holography*. New York: Academic Press. 32 well-known scientists have contributed to this rather thorough coverage of holography.

Collier, R. J., Burckhardt, C. B., Lin, L. H. (1971), *Optical Holography*. New York: Academic Press. A classic. Now somewhat dated, but still widely cited. For years the definitive text on holography.

Denisyuk, Yu. N. (1978), *Foundations of Holography*. St. Petersburg, Leningrad: State Optical, Order of Lenin, Order of October Revolution, S. I. Vavilov Institute. An enjoyable collection of lectures by the Soviet Union's greatest holographer.

Francon, M. (1974), *Holography*. New York: Academic Press (translated from French). A well-done basic book by a very famous French optical scientist.

Hariharan, P. (2002), *Basics of Holography*. Cambridge, UK: Cambridge University Press. A broad coverage of optical holography since its origin.

Iovine, J. (1997), *Holography for Photographers.* Boston: Focal Press.

Saxby, G. (1994), *Practical Holography*. New York: Prentice Hall.

Tonomura, A. (1999), *Electron Holography*. Berlin: Springer.

Vest, C. M. (1979), *Holographic Interferometry*. New York: Wiley. Thoroughly covers holographic interferometry.

Volkl, E. (1999), *Introduction to Electron Holography*. New York: Kluwer Academic/Plenum Publishers.

27

Biomedical Imaging Technologies[1]

William R. Hendee and Anthony B. Wolbarst

27.1 **Radiation as a Diagnostic Probe 897**
27.1.1 Creation of an Image 898
27.2 **X-ray Imaging 899**
27.2.1 Nature and Production of X-Rays 900
27.2.2 Interactions of X-rays with Matter 902
27.2.3 Creation of an X-Ray Image with a Screen and Film 905
27.2.4 Screen-Film Mammography 906
27.2.5 Analog Tomography 908
27.2.6 X-ray Fluoroscopy and Angiography 909
27.2.7 Image Quality 911
27.2.8 γ- and X-Ray Imaging: Radiation Dose and Health Effects 913
27.3 **The Joy of Digital 914**
27.3.1 PACS, RIS, HIS, DICOM, and so on 914
27.3.2 Digital Display 916
27.4 **Digital Planar X-Ray Imaging 917**
27.4.1 Computed Radiography (CR) 917
27.4.2 Digital Radiography (DR) 918
27.4.3 Digital Subtraction Radiography 919
27.4.4 Digital Tomosynthesis 919
27.5 **Computed Tomography 920**
27.5.1 Reconstruction Algorithms 925
27.5.2 Generations of Scanners 925
27.5.3 Helical and Multidetector-Row CT 927
27.5.4 Recent Developments in Computed Tomography 929
27.6 **Nuclear Medicine 929**
27.6.1 Radiopharmaceuticals and Technetium-99m in Particular 930
27.6.2 Scintillation Gamma Camera 931

---

[1] The authors are greatly appreciative that the editor of RADIOLOGY granted permission to draw extensively from their article "Evolving and Experimental Technologies in Medical Imaging". *Radiology* **238**, 16–39 (2006) in preparing this chapter.

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

27.6.3 Single Photon Emission Computed Tomography (SPECT) 932
27.6.4 Positron Emission Tomography (PET) 933
27.6.5 Recent Developments in Nuclear Medicine 934
27.7 Ultrasound Imaging 935
27.7.1 Phased-Array Transducers 938
27.7.2 Ultrasound Flow (Doppler) Imaging 939
27.7.3 Ultrasound Tissue Characterization 939
27.7.4 Recent Developments in Ultrasound 940
27.8 Magnetic Resonance Imaging 941
27.8.1 Nuclear Magnetic Resonance (NMR) 943
27.8.2 T1 (Longitudinal, Spin–Lattice) Proton Spin Relaxation 944
27.8.3 T2 (Transverse, Spin–Spin) Proton Spin Relaxation 945
27.8.4 Spin-Echo, Spin-Warp Voxel-Magnetization Localization 949
27.8.5 Flow Imaging: Magnetic Resonance Angiography (MRA) 951
27.8.6 Functional MRI (fMRI) 952
27.8.7 MR Spectroscopy (MRS) 953
27.8.8 Higher Magnetic Fields 954
27.8.9 Risks from MRI 955
27.8.10 Recent Developments in MRI 956
27.8.11 Electron Resonance Imaging 956
27.9 Laser Optical Imaging 957
27.10 Cellular and Molecular Imaging 960
27.11 Emission Imaging 963
27.11.1 Electrical Source Imaging: Electrocardiography and Electroencephalography 963
27.11.2 Tissue Electrical Impedance Tomography 963
27.11.3 Magnetic Source Imaging: Magnetoencephalography and Magnetocardiography 964
27.11.4 Infrared (IR) Thermography 965
27.11.5 Microwave Imaging 966
27.11.6 Acoustothermometry 966
27.11.7 Terahertz Imaging 966
27.12 Mini- and Nanotechnology 968
27.13 Roles of Computers in Medical Imaging 968
27.13.1 Medical Image Informatics 969
27.13.2 Image Interpretation by Means of Computer-Aided Detection and Diagnosis 969
27.13.3 Input of Computer Technologies from Other Fields 970
27.14 Radiation Treatment Planning 971
27.15 Conclusions 973
Glossary 973
References 975
Further Reading 982

27.1
Radiation as a Diagnostic Probe

Medicine is devoted to preventing disease or injury, detecting and diagnosing what is wrong when disease or injury may be present, and treating the patient to alleviate pain, restore function, and, when possible, cure the problem. Diagnosis is often a major challenge, because the patient's symptoms are frequently relatively unspecific and could be caused by a variety of conditions. Meeting this challenge usually requires conducting a variety of tests, including laboratory analyses and imaging procedures. These tests are nearly all products of technological advances in the practice of medicine that occurred in the twentieth and so far in the twenty-first century.

Before then, physicians were severely limited in the tools available to help them arrive at a correct diagnosis. Most of what they could learn came from a verbal description of symptoms by the patient, combined with the patient's history and of the physician's detection of abnormal odors, visual appearances, sounds, and feel in accessible organs. Indicators such as temperature, pulse and respiratory rates, and visual and muscular reflexes sometimes contributed to the process, as did rudimentary analysis of blood, urine, and so on. Exploratory surgery was a common but risky procedure that often contributed more to the problem than to the solution. On many occasions, the physician could provide little more than sympathetic understanding and emotional support.

Wilhelm Roentgen, a physicist at the University of Wurzburg in Germany, discovered X-rays in 1895, and thereby provided medicine with an astounding and totally unanticipated new life-saving tool for examining the inside of the body without having to cut into it (Figure 27.1) (Röntgen, 1895). And in the year following Roentgen's monumental finding, Henri Becquerel in Paris stumbled upon radioactivity and, in particular, the β-rays that it creates (Becquerel, 1896). While it was not apparent when they were discovered by Villard in 1900, γ-rays would also come to play an invaluable role in medicine.

Diagnostic X-rays are produced by an X-ray tube while γ-rays are emitted from radioactive materials, but they differ only in their origin: once in existence, they are identical. X- and γ-rays are a form of electromagnetic (EM) radiation similar to infrared (IR), visible, and ultraviolet light, except orders of magnitude higher in frequency and

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

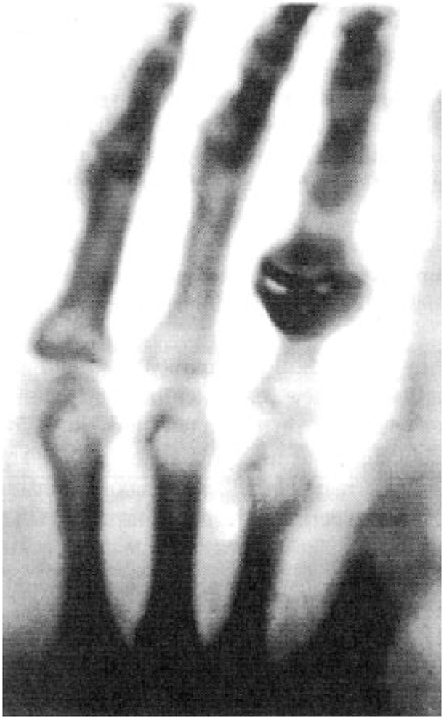

Fig. 27.1 A radiograph of the hand taken by Röntgen in December 1895. His wife may have been the subject. (From the Deutsches Röntgen Museum, Remscheid-Lennap, Germany.)

energy. Unlike lower-energy forms such as IR and visible light, γ- and X-rays have three critically important characteristics: they can penetrate substantial thicknesses of biological materials, with the degree of penetration depending on the nature of the tissue; they cause the emission of points of light (*scintillations*) when individual photons strike a fluorescent medium; and they ionize the tissues through which they pass. Their diagnostic usefulness depends crucially on the first two of these attributes, and the risk they pose arises because of the third.

27.1.1
Creation of an Image

Technology has long been propelling itself forward at an ever-faster pace, and medical imaging technology has been at the front edge of the changes, especially since the mid twentieth century. Today, methods to produce medical images involve the use of physical probes that are created, affected by the body, and detected in strikingly different ways. Some methods, such as X-ray imaging, computed tomography (CT), diaphanography, and tissue impedance tomography function by detecting how many of the probes pass through regions of the patient without interaction. Others, such as nuclear medicine (including single photon emission computed tomography (SPECT) and positron emission tomography (PET), thermography, electroencephalography (EEG), and magnetocardiography (MCG)), create images from extremely faint signals emerging from different regions of the body. Waves of ultrasound (US) energy are reflected from tissue interfaces and moving blood. Magnetic resonance imaging (MRI) involves the transmission of pulses of radio frequency (RF) EM radiation, their interaction with the protons of tissue water in a strong external magnetic field, and reception of RF signals the protons themselves subsequently produce. And an increasing number of optical imaging techniques employ visible or near-visible light to provide image information about patients.

Photons of all energies (X-ray, γ-ray, ultraviolet, optical, IR, microwave, and RF), weak electric and magnetic fields, US waves, and other probes differ substantially in their ability to penetrate the body, the types of background noise with which they must compete, and the ease with which they can be detected and localized. These factors, along with concerns related to radiation dose, acoustic power, intensity of static and dynamic magnetic fields, and so on, can influence the level of image contrast achievable among healthy and diseased tissues, the spatial and temporal resolution possible, the presence of noise or artifacts, and the overall clinical utility of an imaging modality. As a result, there are a number of widely accepted technologies to choose from in posing and addressing a particular clinical question; indeed, part of the task of the clinician is to be adept at selecting the imaging technology with the greatest likelihood of providing the correct diagnostic answer safely and at acceptable cost.

Biomedical research is becoming ever more interdisciplinary, as it depends increasingly on cross-fertilization between the physical sciences and engineering on the one hand, and biology and clinical medicine on the other. Examples of the fruits of such collaboration abound – such as recent developments in molecular imaging, biomedical informatics, nanobiotechnology, computer-aided detection (CAD) and computer-aided diagnosis (CADx), and the convergence of diagnosis and therapy in applications such as image-guided intervention. These issues comprise the subject of this chapter.

27.2
X-ray Imaging

There are two general categories of imaging with the high-energy photons produced by X-ray tubes. With either method, image information is obtained in the form of shadows cast in a uniform X-ray beam through differential attenuation by materials in the body that vary in thickness, density, or chemical composition (i.e. atomic number), leading to the creation of a two-dimensional X-ray shadowgram by tissues actually distributed in three dimensions. The first, in the form either of conventional *radiography* or of *fluoroscopy*, displays the resulting shadow information directly on a flat image receptor (such as a screen-film cassette, fluoroscopic image-intensifier tube plus electronic optical camera combination, or digital solid-state flat-panel imager) as a *planar* (flat), two-dimensional image. The second, X-ray transmission *CT*, captures the shadow with a bank of tiny radiation detectors, and then *reconstructs* the data to produce a true three-dimensional representation of the internal structures of the body – although clinical images are traditionally presented as transverse or other pancake-like tissue *slices*.

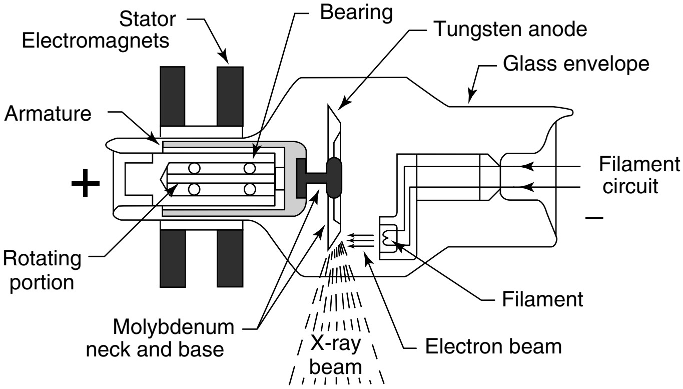

Fig. 27.2 Simplified X-ray tube with a rotating anode and a heated filament. (From Hendee, W.R. and Ritenour, E.R. (2002) *Medical Imaging Physics*, 4th edn, John Wiley & Sons, New York.)

27.2.1
Nature and Production of X-Rays

To generate clinically acceptable medical images, a source of X-rays has to provide a sufficient number of X-ray photons efficiently and in a reproducible fashion and, for radiography, in a short interval of time. The operator must be able to select the beam intensity, the exposure duration, and the maximum photon energy. And, of course, the device is legally bound to meet current standards of safety.

The standard source almost universally employed for X-ray imaging is the X-ray tube (Figure 27.2). As electrical current (on the order of several amperes) heats a tungsten filament in the cathode, electrons are "boiled off" (*thermionic emission*) and, when no high voltage is being applied, form an electron space-charge cloud near its surface; the emission causes the cathode to become slightly positive and that, along with the mutual repulsion among the electrons, prevents the cloud from spreading far. The situation changes radically when a potential of 20 000–150 000 V (20–150 kVp, where (*kVp*) denotes the *peak kilovoltage*) is applied directly between cathode and anode. Electrons are swept through the high vacuum and strike the thin tungsten target (high melting point) bonded to the surface of the massive copper anode (high thermal conductivity), all with a maximum kinetic energy (in *electron volts*[2], eV) numerically equal to the kVp setting on the tube's generator. A focusing cup around the filament helps direct the electrons onto the small *focal spot,* which traces out a narrow annulus on the rapidly rotating anode. Most (>99%) of the electron energy is expended through resistive heating of the target and anode; the massive anode disk rotates at several thousand revolutions per minute to spread the heat over a larger area of target metal, to keep its temperature below its melting point. The remaining 1% of the electron energy arriving at the focal spot is transformed into X-rays by two quite different mechanisms.

---

[2] The electron volt is a unit of energy equal to that gained by an electron accelerated through a potential difference of 1 V. keV stands for kilo-electron volt.

The first process, the creation of *bremsstrahlung* (from the German for "braking," as in hitting the brakes of a car), occurs through a mechanism similar to that which gives rise to radio waves. RF EM energy is radiated from a radio antenna when electrons are driven back and forth in it by the high alternating voltage produced by a radio transmitter. Likewise, the electrons striking the anode of an X-ray tube undergo Coulombic interactions with strongly charged nuclei of the target atoms, and are often deflected widely and also slowed down abruptly. Their rapid deceleration causes the emission of EM energy but, since the electron energies are so much higher, so also are the energies of the emitted X-ray photons. A bremsstrahlung spectrum is distinctive: although all electrons reach the anode with close to the maximum energy of kVp eV, only a tiny fraction of the resulting X-ray photons have the maximum energy, and the intensity of an "ideal" X-ray beam increases linearly with decreasing photon energy (Figure 27.3). Preferential absorption of low-energy photons by the anode itself, added *filtration*, and other materials *harden* the beam, causing the drop-off in beam intensity at low photon energies. Since the low-energy photons are not sufficiently penetrating to be of diagnostic value, but would deposit potentially harmful energy (dose) in a patient's tissues, their removal from the beam is a highly desirable effect.

The other mechanism for X-ray production involves the interaction of

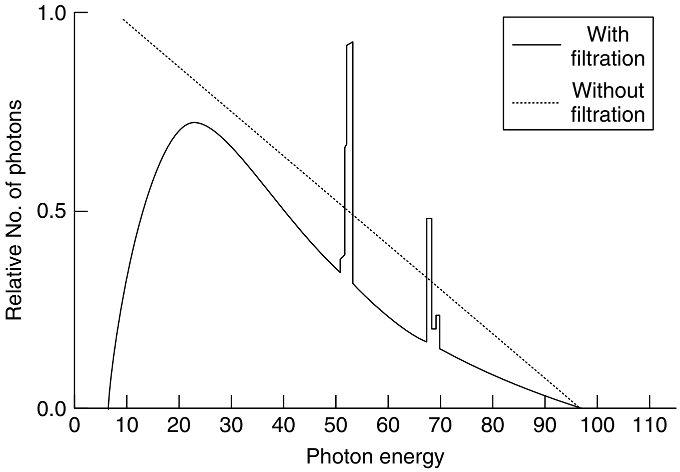

Fig. 27.3 Emission spectrum for a tungsten target X-ray tube operated at 95 kVp. Characteristic X-ray emission occurs for tungsten whenever the tube voltage exceeds 69 keV, the K shell binding energy for tungsten. The *dashed line* represents the theoretical bremsstrahlung emission from a tungsten target. The *solid line* represents the spectrum after self, inherent, and added filtration. The area under the spectrum represents the total number of X-rays. (From Hendee, W.R. and Ritenour, E.R. (2002) *Medical Imaging Physics*, 4th edn, John Wiley & Sons, New York.)

the incident, high-energy electrons with the orbital electrons, rather than the nuclei, of the target atoms. Target atoms are ionized with the ejection of inner-shell electrons, and X-ray photons are released as the affected atoms regain stability: outer electrons drop down into the vacancies, with a reordering of the electron shells, and emit photons in the process. These photons are termed *characteristic X-rays* because their energies are characteristic of the target metal material. Shown in Figure 27.3 are the 58- and 67-keV characteristic X-rays of tungsten.

For X-ray beams in nearly all medical applications, bremsstrahlung is a much more important contributor than characteristic X-rays. A notable exception is the primary role played by molybdenum characteristic X-rays in mammography (imaging of breast tissue), an application discussed later. Either way, X-rays are released in all directions as the electrons interact in the target. The X-rays are restricted by shielding and by collimators outside the X-ray tube to a beam to be directed at the anatomic region of interest in the patient. The tube is contained within an oil-filled shielded housing that provides electrical insulation and prevents X-rays from emerging in unwanted directions. Only about 1% of the original electron beam's energy is transformed into X-rays, and only a fraction of these X-rays escape through the tube window and housing to reach the patient.

During an X-ray exposure, the *technique factors* [tube voltage (kVp), tube current (mA), and exposure time (ms)] are controlled carefully to prevent overheating of the target and other components.

27.2.2
Interactions of X-rays with Matter

For a parallel, monoenergetic beam of X-rays, the intensity, I, of X-rays transmitted through a material of thickness z obeys de Beer's law:

$$I = I_0 \exp \int_0^{Z} \left[-\mu(z') dz' \right] \tag{1}$$

where the line integral is taken across all tissues along a "ray" of the X-ray beam, I_0 is the intensity with no material present, and $\mu(z)$ is the *linear X-ray attenuation coefficient* of the material, typically in units cm^{-1}. The linear attenuation coefficient depends on the effective energy of the beam, and on the density and effective atomic number of the material through which it passes, so it may vary along the path. If the material varies in composition laterally, as well, the intensity of the transmitted X-rays will vary across the face of the previously uniform X-ray beam, casting a shadow in it. This pattern of transmitted X-rays, which now bears information about the patient, can be captured on film, a fluorescent screen, or digital array detector, yielding a projection radiograph.

The linear attenuation coefficient is the sum of coefficients for several different types of X-ray interactions that may occur in the material. For the diagnostic energy range (20–150 keV), two of them are dominant: the photoelectric effect, described by the photoelectric linear attenuation coefficient, τ, and the Compton scatter effect with Compton linear attenuation coefficient, σ.

In a *photoelectric interaction*, the X-ray interacts almost always with a tightly bound electron in an inner shell of

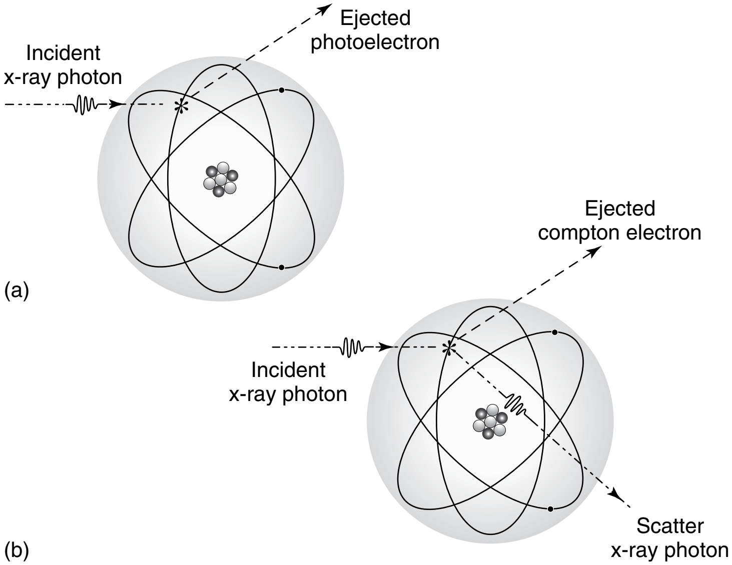

Fig. 27.4 (a) Photoelectric absorption of a photon with energy $h\upsilon$. The photon disappears and is replaced by an electron ejected from the atom with kinetic energy $E_k = h\nu - E_b$, where E_b is the binding energy of the electron. Characteristic radiation and Auger electrons are emitted as electrons cascade to replace the ejected photoelectron. (b) Compton scattering of an incident photon, with the photon scattered at an angle ø; the Compton electron is ejected at an angle θ with respect to the direction of the incident photon.

an atom of the tissue (Figure 27.4a). Photoelectric interactions are strongly influenced by the atomic number Z, with the probability of photoelectric interaction per unit mass of material varying with the third power of the effective atomic number of the material, as $\tau \sim Z^3$. In a collision, all the energy of the X-ray is transferred to the electron, which is ejected from the atom with a kinetic energy E_{KE} of

$$E_{\mathrm{KE}} = E_{\text{X-ray}} - E_b \tag{2a}$$

E_b is the binding energy of the ejected electron, and $E_{\text{X-ray}}$ is the X-ray energy, which can also be expressed as $h\nu$, where h is Planck's constant (6.62×10^{-34} J s) and ν the frequency of the wave-like X-ray. Since the X-ray is pure energy, it disappears during a photoelectric interaction and is essentially replaced by an ejected high-speed electron and an ionized atom. Photoelectric interactions occur predominantly with low-energy X-rays, and decrease in likelihood at higher energies ($\tau \sim (h\nu)^{-3}$). Overall, the probability of a photoelectric event occurring depends on density, atomic number, and photon energy as

$$\tau \sim \rho(Z/h\nu)^3 \tag{2b}$$

A *Compton interaction* results in the scattering of an X-ray in a new direction at reduced energy, rather than complete absorption of the X-ray (Figure 27.4b). Compton scattering produces scattered radiation that degrades the quality of

X-ray images. In this type of interaction, the X-ray photon collides with a loosely bound electron in an outer shell of an atom of the absorbing material. The electron is ejected with an energy equal to the difference in energy of the incoming and scattered X-rays, since the binding energy of the ejected electron is usually small enough to be neglected. Because the interaction occurs almost always with outer, loosely bound electrons, the Compton interaction is nearly independent of the atomic number Z of the absorbing material. The linear Compton attenuation coefficient is also nearly independent of photon energy in the diagnostic range, and becomes dominant at higher energies simply because of the rapid decline in photoelectron events.

The linear attenuation coefficients τ and σ are, in effect, probability rates for mutually exclusive events – that is, a photon that is removed from the beam by a photoelectric collision cannot also undergo a Compton scatter – so

$$\mu = \tau + \sigma \tag{3}$$

Depending on the kVp setting, an X-ray beam may yield a projected image that accentuates or suppresses effects due to differences in atomic number among different tissues, which can be of great advantage in imaging. Low-energy X-rays can be used to amplify small differences in soft tissue composition, as is standard practice in mammography. On the other hand, high-energy X-rays may be needed to suppress the consequences of relatively large differences in tissue composition, so that subtle features of soft tissue are not obscured by the shadows cast by strongly absorbing structures. In chest imaging, high-energy X-rays are used to penetrate the ribs and other strongly absorbing structures to reveal the intricacies of the underlying lung parenchyma (Figure 27.5).

For X-rays at any energy, those that are not transmitted through are absorbed or scattered in the patient. This results in energy deposition in the tissues, where the ionizing energy deposited per unit mass of absorbing material is

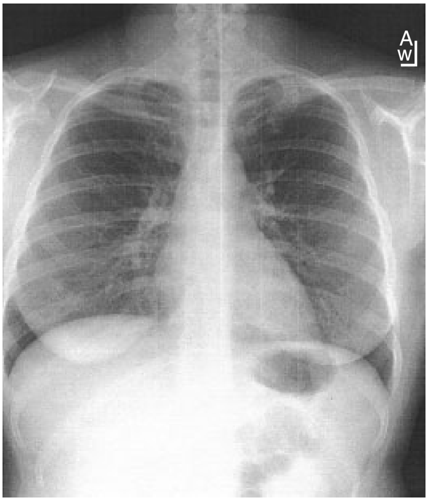

Fig. 27.5 A normal chest radiograph (courtesy of Lacey Washington, MD, Medical College of Wisconsin).

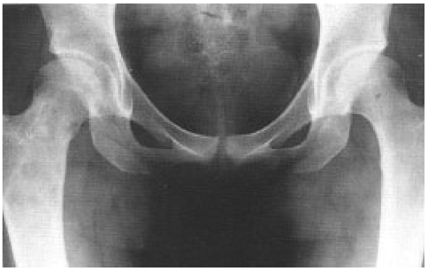

Fig. 27.6 Various appearances of fibrous dysplasia. Young adult with a sclerotic, osteolytic, and intracompartmental lesion confined to the femoral neck and proximal diaphysis. There is a fracture in the medial aspect of the femoral neck that accounts for symptoms. (From Bragg, D., Rubin, P., and Hricak, H. (eds) (2002) *Oncologic Imaging*, 2nd edn, W.B. Saunders Company, Philadelphia, with permission).

known as the *absorbed dose*. Because dose, like image quality, depends on beam energy, the choice of kVp requires the optimization of the balance between the two. This is but one of many trade-offs one confronts in medical imaging.

27.2.3
Creation of an X-Ray Image with a Screen and Film

When a diagnostic X-ray beam is directed into a region of the body, variations in thickness, density, and chemical makeup (effective atomic number) create shadow patterns in it, the relative shades of gray of which are influenced also by the effective energy of the beam. These spatial variations in intensity can be captured on a radiation-sensitive *image receptor* such as a light-tight *film cassette,* which sandwiches a sheet of X-ray film between a pair of flat *fluorescent screens,* currently almost always made of rare earth materials.3) When a screen is struck by an X-ray photon, it emits a point of light that exposes a microcluster of silver iodobromide crystals in the film's emulsion. The bits of silver halide that are exposed to sufficient light are later converted into flakes of pure silver upon chemical development, making the film darker. The resulting film is a negative, in a sense, since a part of the body where attenuation is small gives rise to a darker portion of film.

In the X-ray film of Figure 27.6, the bony structures are lighter because bone is a strong absorber of X-rays, and the portion of film behind it experiences less X radiation. Distinctions among the various soft tissues are much less obvious, however, because their attenuation coefficients are nearly the same. A major degrading influence on contrast is the incidence on the detector of Compton-scattered radiation in the patient, a form of random noise. Scatter can be reduced by placing a mechanical *grid* between the patient and the detector. A grid, consisting of thin metal foil sheets separated by very narrow gaps, allows unscattered X-rays to pass through to the detector, but obstructs scattered X-rays that have been deflected from their original paths through the patient.

The spatial resolution of a typical radiograph is exceptionally high, on the order of a tenth of a millimeter. This is so, in part, because the origin of the X-rays,

---

3) The sole purpose of the screens is to reduce patient exposure. This can be achieved very effectively, but unfortunately it also diminishes image sharpness somewhat. As in many situations in imaging, the operator is faced with a trade-off in selecting the screen, having to choose between higher sensitivity versus finer resolution.

the focal spot, is nearly a point source, on the order of a millimeter across, and the primary X-rays are projected from it in straight lines through the patient onto the image receptor. The blur that does arise comes mainly from the small but finite size of the focal spot (penumbra), motion of the patient or equipment, and the finite thickness of the fluorescent screens, within which fluorescent light can bounce about and, before exposing the film, diffuse away from the point where the X-ray struck.

27.2.4
Screen-Film Mammography

One of every eight women will develop breast cancer over the course of her lifetime, and some will die as a result. The most effective measures for early detection, and hence more likely cure, of this disease are frequent self-examinations and periodic X-ray screening of the breasts, which may detect a cancer before it has spread. X-ray imaging of the breasts is referred to as *mammography*.

A mammogram (an image of the breast) must be able to reveal very subtle irregularities in the composition of soft tissues, as well as the presence of tiny ($<100\,\mu m$) *calcifications*, specks of calcium hydroxyapatite that may (depending on their sizes and shapes) suggest the presence of malignant cells. That is, mammographic images must be able not only to reveal slight variations in contrast against a noisy background, but also to provide the exquisite detail needed to reveal the microcalcifications. Mammograms must depict the relatively thick tissues near the chest wall without loss of detail in the thinner regions near the edge of the breasts. And they must do so at low doses, so as not to induce new malignancies while searching for existing ones. These requirements place inordinate demands on breast imaging systems, and also necessitate stringent periodic evaluation of image quality and radiation dose by a physicist experienced in inspection and quality control specifically for mammography units.

A mammography device is like a radiography unit but with several significant differences. Subject contrast is enhanced in mammography by using low-energy X-rays that interact strongly with tissues by the photoelectric process. Low-energy X-rays yield the greatest differences in transmission among structures that are similar, although not identical in chemical composition. Mammography is performed most commonly with special, dedicated X-ray tubes with a molybdenum X-ray target. Operated at 20–35 kVp, these units produce predominantly K X-rays of 17.4 and 19.8 keV characteristic of molybdenum (Mo), superimposed on a bremsstrahlung X-ray spectrum. The X-ray beam is filtered with a thin sheet of molybdenum that is relatively transparent to the Mo characteristic X-rays but strongly absorbent of higher-energy photons because of the Mo K-absorption edge just above 20 keV. The image receptor consists of single-sided film within a single-screen cassette designed specifically for mammography. Screen-film mammography is exquisitely sensitive, but is being displaced by systems with digital detectors, which substantially improve the contrast sensitivity, an advantage that outweighs a slightly lower spatial resolution.

In earlier years, there was concern that screening large numbers of asymptomatic women for breast cancer could induce as many cancers as are cured

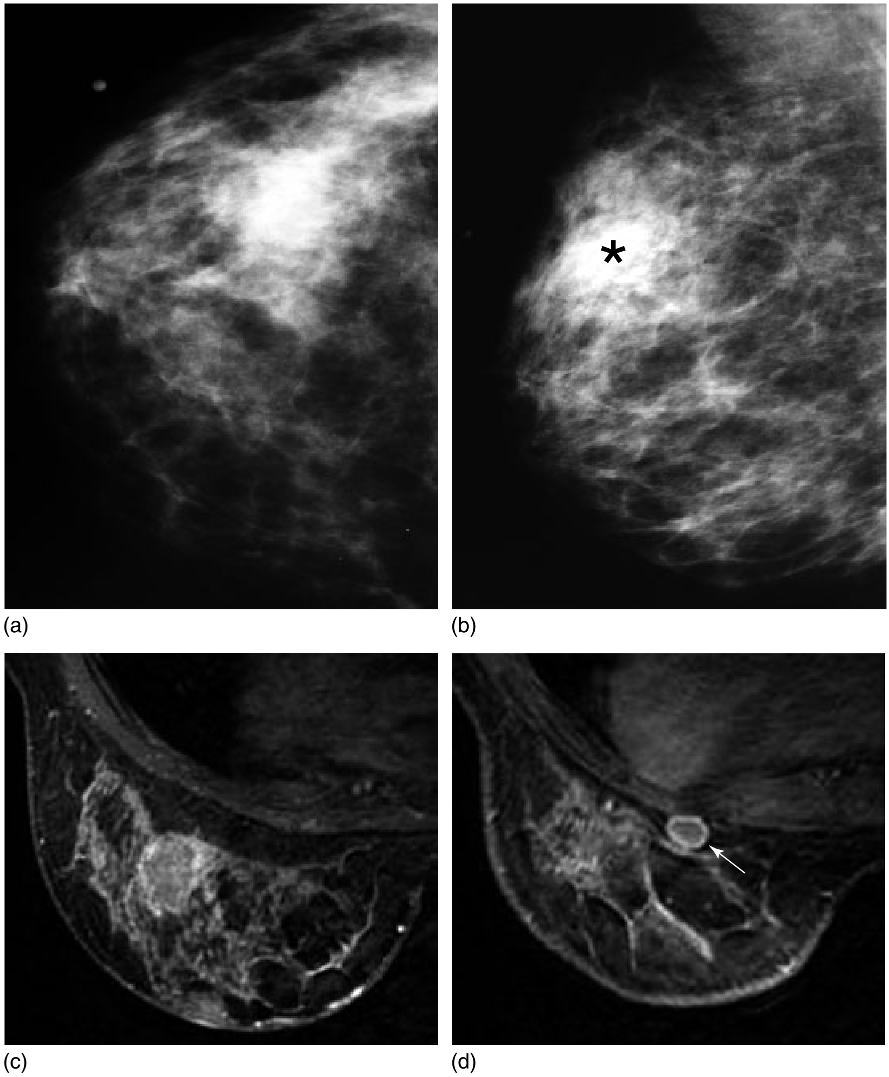

Fig. 27.7 Mammography and MR imaging of the breast. (a) Left craniocaudal and (b) left mediolateral oblique mammograms demonstrate an irregular high-density mass. Overall, breast tissue density is heterogeneous, which is consistent with the patient's age (35 years). Biopsy results show the mass to be grade III invasive ductal carcinoma. (c and d) Transverse three-dimensional fast spoiled gradient-echo MR imaging was performed to assess remaining breast tissue. (c) Enhancing mass corresponds to the known biopsy-proved breast cancer. (d) Incidental mass (arrow) was found posteriorly in the same breast. Biopsy was performed later with US guidance, and the mass found to be grade I invasive ductal carcinoma. This mass was not evident on the mammogram, either initially or in retrospect. (Image courtesy of Lonie Salkowski, MD, Medical College of Wisconsin, Milwaukee, WI.)

through early detection. This concern was primarily a result of the relatively high doses delivered to patients with earlier, much less sensitive methods of image production. With today's ultrasensitive screen-film combinations and digital image receptors, together with the use of low-energy, highly filtered X-ray beams for mammography, the induction of cancer is much less likely. The overwhelming balance of benefit over risk in screening programs for older women has largely dispelled concern over the possible induction of cancer in the screened women. Many voluntary and professional organizations now recommend annual screening with X-rays for women 50 years of age and older, and some recommend screening every 1–2 years for women between 40 and 50 years of age.

Nevertheless, efforts to reduce radiation dose further in mammography should be continued until each examination uses no more than the essential number of X-rays to reveal underlying pathology. Alternatively, there is evidence that (much more expensive) breast MRI (Liberman, 2004; Kriege *et al.*, 2004) is as effective in detecting malignancies as is dedicated mammography. It is especially promising for women who are at high risk for breast cancer or who have unusually dense breast tissue, which presents a severe visualization challenge to screen-film and digital mammography (Figure 27.7).

27.2.5
Analog Tomography

In a conventional radiograph, all structures within a three-dimensional volume of tissue are projected in focus onto the two-dimensional image receptor. As a result, an anatomic structure of interest may be obscured by the shadows of objects that lie above or below it. Removal of these shadows can greatly improve the visibility of the structure of interest. Two approaches, planar tomography and CT, are available to remove the shadows of overlying and underlying objects. Both are normally carried out digitally now, although some older planar tomography systems are still film based.

Planar tomographic images are produced by varying the geometry so as to keep structures in the selected image plane within the patient in focus while blurring out the images of everything above and below the image plane. The X-ray tube and image receptor are translated relative to one another during the exposure about a pivot (fulcrum) in the image plane (Figure 27.8). The image plane has a finite thickness (and is referred to more properly as an *image section*) that can be reduced by increasing the angle of motion. The location of the section within the patient may be moved up or down by positioning the patient higher or lower relative to the pivot. Tomosynthesis is a type of limited-angle (compared with CT) tomography that has relatively poor resolution in the depth direction. The level of detail in the imaging plane, however, can be superb.

Although the shadows of overlying and underlying structures are blurred, they are not removed completely from the image, and frequently appear as ghostlike artifacts. In addition, analog tomography has the standard limitation of film that it cannot distinguish small differences (1–2%) in the intensity of incident radiation, with implications for achievable contrast. Also, the use of wide-area X-ray beams in analog tomography creates problematic scatter

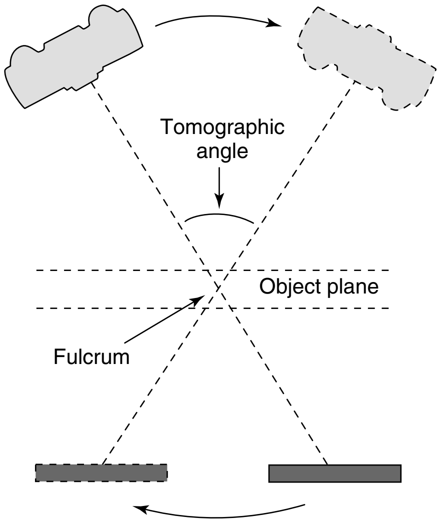

Fig. 27.8 Relationship of the fulcrum, object plane, and tomographic angle. (From Bushong, S.C. (2001) *Radiological Science for Technologists: Physics, Biology, and Protection*, 7th edn, Mosby, St. Louis, with permission.)

radiation. Planar tomography is still used for certain applications such as a complement to chest radiography, but for the most part is being displaced by tomosynthesis and the considerably more effective but expensive CT.

27.2.6
X-ray Fluoroscopy and Angiography

Analog *fluoroscopy*, unlike film, can capture dynamic X-ray images in studies where the organs and tissues are in motion, as with the heart and great vessels and the gastrointestinal tract.

A traditional fluoroscopic device is similar to a screen-film system except that the screen-film image receptor is replaced with an *image-intensifier* vacuum tube. The II converts the faint, life-sized X-ray pattern emerging from the patient into a bright and much smaller (2–4 cm in diameter) visible image, suitable for input into a TV or a charge-coupled device (CCD) electronic optical camera, which feeds signals into a computer. In systems marketed today, the II-tube plus optical camera combination is being replaced by a solid-state flat-panel imager.

The front end of the evacuated II tube is a fluorescent screen, usually CsI, coated with a thin veneer of photoemissive material that emits an electron when struck by a visible-light photon. Electrons liberated in this manner are accelerated toward the anode and compressed by focusing electrodes to strike the small output fluorescent screen at high velocity. The brightness ratio of the output screen compared with the input screen is known as the *brightness gain* of the intensifier, and is typically 3000–4000. It is a product of two factors: the minification gain accomplished by decreasing the size of the image and the electronic gain achieved by accelerating the electrons to high energy before the output image is produced.

From time to time during a fluoroscopic procedure, it may be necessary to record a few higher-quality static images.

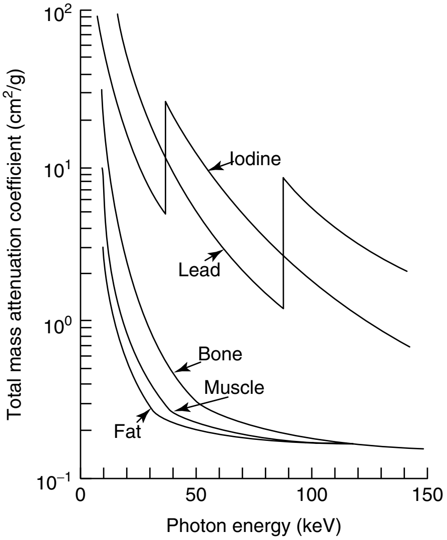

Fig. 27.9 Mass attenuation coefficients for fat, muscle, bone, iodine, and lead as a function of photon energy. (From Hendee, W.R. and Ritenour, E.R. (2002) *Medical Imaging Physics*, 4th edn, John Wiley & Sons, New York.)

Inserting a film cassette momentarily between the patient and the II provides a static *spot film*. Alternately, the light image from the output screen of the intensifier may be deflected briefly into a 100-mm *photospot camera*. Another option is to capture the image with a CCD or television camera and digitize the signal.

X-ray imaging is applied widely in real-time studies of the functioning of the heart and cardiovascular system. In studies of the circulatory system, for example, fluoroscopy can depict normal anatomy and identify abnormal conditions such as circulatory malformations, dysfunctional heart valves, coronary artery disease, atherosclerotic plaques, blood clots, circulation changes caused by injuries and diseases such as cancer, and a wide spectrum of other conditions. Because of the similarity of their X-ray attenuation characteristics, blood-containing vessels usually are not distinguishable from muscle and other soft tissues. This situation can be improved by filling the vessels with a high-atomic number contrast agent that has the same physical density and osmolarity as blood. Iodine-containing materials are currently the contrast agents of choice for cardiovascular studies, largely because iodine's 33-keV absorption edge is near the average photon energy for a typical diagnostic beam (Figure 27.9). X-ray studies of arteries and veins with contrast agents are referred to as *angiography*. Angiography confined to the study of arteries is called *arteriography*, while studies of the venous system are termed *venography*.

Atherosclerotic plaques can form, for example, in the coronary arteries, the vessels that supply blood to the heart muscle. This creates constrictions (narrowing) that impede and sometimes stop arterial blood flow. This condition, if not corrected, can cause angina and a heart attack. In *coronary angiography*, the contrast agent outlines the vessels and displays constrictions that may be present. If not too severe, the constrictions can be removed by mechanically opening the vessels, a procedure known as *angioplasty*. In more severe cases of coronary artery *stenosis*, it may be necessary to surgically install a vessel in parallel to the constricted artery, a procedure known as a *bypass graft*. Often a *stent* tube is placed inside the vessel after angioplasty to prevent it from constricting again. All these procedures are performed under continuous or pulsed fluoroscopic X-ray examination.

A momentous change in planar X-ray imaging is that a large and growing fraction of new radiographic and fluoroscopic equipment currently being installed in the United States contains image receptors that are fully digital.

27.2.7
Image Quality

The level of clinical information visible in a medical image results from a composite of four kinds of resolution: contrast, spatial, temporal, and statistical. *Contrast resolution* reflects differences in the tissue characteristics depicted in the image, and by the impact of extraneous information or structured or random noise in the receptor. *Spatial resolution* and sharpness are related to the geometry of the object being examined (e.g. sharp edges) and of the image-forming process (penumbra blurring from the finite size of a focal spot), and the inherent resolving capacity of the image receptor (blur from the diffusion of light in a fluorescent screen of finite thickness). *Temporal resolution* refers to the spatial blurring caused by the patient's voluntary or involuntary movement (cardiac motion, vascular pulsatile motion, breathing, and peristalsis) during data acquisition, or by shaking of the apparatus. *Statistical resolution* is related to the number of information carriers (X-ray or γ-ray quanta, US or MRI echoes, etc.) detected by the receptor to produce the image; random noise is inversely proportional to the square root of the number of information carriers used to form the image, and it is a common practice to improve the signal-to-noise ratio by acquiring data longer or averaging over several images, at higher dose, and so on.

In most cases, improving any one of these four resolution factors degrades one or more of the others. The necessary compromise among them represents a balance so that no single factor strongly dominates the degradation of detail in the image.

The challenge of medical imaging is to capture subtle differences in tissue composition with information carriers that probe the tissue, and to transfer the resulting patterns of contrast onto an image receptor. That is, the primary purpose of medical imaging is to transfer contrast *modulation* arising in the tissue (object) into visual modulation in the captured image. The degree to which this modulation is transferred faithfully by an imaging system, or by any independent component of it, is commonly described by a relatively simple measure of performance called the *modulation transfer function* (*MTF*). With most imaging systems, object contrast can be transferred to the image with high fidelity if the variations of physical properties within it are of relatively low frequency (i.e. the object is composed of relatively large structures that are well separated from adjacent structures). For smaller structures that are closer together, however, the modulation transfer may work less well, so that the MTF as a function of frequency, MTF(f), tends to fall off at higher frequency.

To measure the MTF of an X-ray tube focal spot, for example, one creates the image of a tiny hole (approximating a spatial delta function) in an otherwise radio-opaque thin sheet of metal. The image of the hole is scanned with a microdensitometer, and the resulting curve is described by a *point spread function* (*PSF*), the composition of which in frequency space can be determined through Fourier analysis. The result of this Fourier transform of the PSF is the MTF, expressed as a function of spatial frequency in cycles or line pairs per millimeter, just as temporal frequency may be reported in cycles per second or hertz.

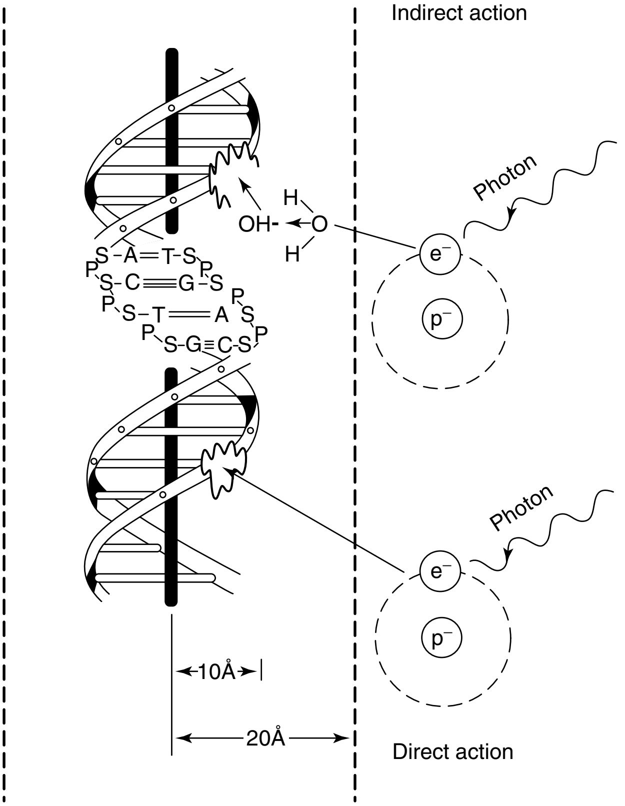

Fig. 27.10 Schematic of the direct and indirect actions of radiation on DNA. The letters S and P represent the sugar and phosphate backbone of the DNA molecule, and A, T, C, and G represent the four bases adenine, thymine, cytosine, and guanine. In direct action, an ejected electron produced by the absorption of radiation interacts directly with the DNA molecule, whereas in indirect action the ejected electron interacts with a water molecule, producing free radicals, which in turn damage the DNA molecule. (Redrawn from Hall E.J. (1988) *Radiobiology for the Radiologist*, 3rd edn, JB Lippincott, Philadelphia, p. 11, with permission.)

27.2.8

γ- and X-Ray Imaging: Radiation Dose and Health Effects

As a beam of X- or γ-rays traverses a body, it loses intensity as it transfers energy to the tissues it penetrates. This transfer of energy involves the ejection of photo- and Compton electrons from the atoms, resulting in a concentration of ions along the path of the beam. The ejection of an electron from an atom is termed an *ionizing event*, and X- and γ-rays and beams of high-energy charged particles are referred to as *ionizing radiation*.

It is the thousands of ionizations in tissue caused by the highly energetic photo- or Compton electrons that are of concern here. An ionizing event can injure or kill a cell if it occurs in an atom of a critical molecule of the cell, and in particular a DNA molecule. This effect, termed the *direct effect of radiation*, is responsible for some of the biological effects of radiation (Figure 27.10). More often, the ionizing event occurs in water, but this can still cause biological damage because the event may create *free radicals* that are highly reactive (because they contain unpaired electrons, as with the OH° or H° radicals) that can attack DNA and other important molecules *indirectly*. The indirect and direct mechanisms together are responsible for the harmful effects of ionizing radiation.

It is possible that sufficiently intense irradiation can kill so many cells and cell-precursors in an organ or tissue that it can no longer operate properly, resulting in a *deterministic health effect*. A deterministic effect typically occurs only above a tissue- and patient-specific threshold dose, but its severity increases with dose beyond that threshold. Examples are a burning of the skin (erythema), sloughing off of skin (moist desquamation), cataract generation, temporary or permanent sterilization or, in severe overexposure, *radiation sickness*, which can be fatal. Deterministic effects are caused only rarely in diagnostic imaging; a much greater worry are *stochastic effects*.

"Stochastic" is synonymous with random, or probabilistically determined, and here the word refers to the likelihood that someone exposed to ionizing radiation will experience a critical *DNA mutation* in a tissue cell (*not* a *killing* of the cell, as in deterministic effects) that leads at a later time to the induction of cancer[4] (carcinogenesis), the production of genetic consequences in offspring of exposed individuals (mutagenesis), and the development of abnormalities in persons exposed before birth (teratogenesis). With stochastic effects, it is not the severity of the disease that increases with dose, but rather the probability of its occurring at all. It is widely accepted, at least for radiation protection purposes, that the risk of stochastic injury increases nearly linearly with dose, and that there is no threshold below which effects cannot occur. The biological and epidemiological evidence for this *linear, no-threshold hypothesis* is an issue of considerable debate in the health physics profession.

---

[4] Cancer is a disease in which certain cells have been transformed so that they are dividing and multiplying at a much faster rate than normal. If these cells can be removed or destroyed without causing overly severe side-effects in the healthy tissues, the cancer patient often has a good chance of recovery. If the cells are confined to a specific region of the body and reasonably accessible, often they can be removed with the scalpel or destroyed with a beam of ionizing radiation. The latter technique is termed *radiation therapy*, and the physicians who utilize it are called *radiation oncologists*. If the cancer cells have spread (metastasized) from the initial (primary) site, the preferred treatment is usually to try to kill the cells with chemotherapy performed by a medical oncologist. A frequent approach to cancer treatment is to employ radiation to destroy the primary site of the cancer (sometimes after most of the cancer has been removed surgically), followed by chemotherapy in an effort to kill cancer cells that have spread to distant sites.

Radiogenic carcinogenesis is the health effect of greatest concern in imaging with γ- and X-rays, because at low doses its estimated rate of induction is much higher than for genetic effects or teratogenesis. It is the role of *medical and health physicists* to ensure that those who employ radiation do so in such a manner that virtually no deterministic health effects arise, and that the rate of stochastic health affects is kept below an "acceptable" level, where the meaning of "acceptable" is a social and regulatory call. The health physics literature is extensive, and several good texts provide an introduction to the field (Turner, 2007).

27.3
The Joy of Digital

First appearing in the clinic in the 1970s, computers have enabled novel computation-based imaging (CT, MRI, PET) – provided image processing and analysis tools ranging from windowing and noise reduction for all modalities to CAD or CADx for some – and allowed the instantaneous storage, retrieval, display, and transmission of images and other information with *picture archiving and communications systems* (*PACS*), electronic notepads, and so on. In so doing, they have radically changed the way that medicine is practiced. The deployment of digital imaging systems – high-bandwidth networks to transport images between acquisition, interpretation, consultation, and storage sites – and algorithms for computer recognition of image features for CADx are ushering in a whole new approach to medical imaging. The only certainty is that what we see today is just the beginning.

27.3.1
PACS, RIS, HIS, DICOM, and so on

Increasing numbers of radiology and other departments are installing PACS for both patient use and clinical research. Images from digital diagnostic devices are sent to the PACS as soon as they are acquired, where they can undergo image processing, computer-based analysis, distribution to other sites, and storage. Ideally, a PACS is integrated with the *radiology information system* (*RIS*) and the *hospital information system* (*HIS*) as part of an *integrated health care* (*IHC*) environment, so that the patient's electronic clinical record, and past and recent lab results, along with past imaging studies, can be pulled up immediately.

PACS got off to a shaky start years ago, largely because early systems were overhyped and user-unfriendly. With simpler user-PACS interfaces, better and more standardized PACS-machine connectivity such as that brought about by the *Digital Imaging and Communications in Medicine* (*DICOM*) protocol (DICOM, 2003), more effective training provided by vendors, better security, lower purchase and operational costs, and the availability of mini-PACS packages that can be upgraded as physicians choose to make fuller use of them, PACS are becoming indispensable for providing high-quality, cost-effective health care (Carrino, 2003; Huang, 2004; Smith, 2006).

In the common star network configuration, a PACS information management system receives images from any connected digital image-acquisition device, after which it can forward them and other information to any other sites (Figure 27.11). Some such sites consist of stand-alone imaging systems complete with their own image and display workstation, image archive, and image-processing computer. Others are "dumb" terminals that can receive transmitted images for viewing but lack the capacity for image modification. The latter are typically placed in clinical locations, such as critical-care units, emergency rooms, and outpatient clinics. The degree to which terminals for referring physicians can be less sophisticated (and less expensive) than those for radiologists is a highly debated issue.

Within a hospital or clinic, imaging data are usually transmitted on-line by copper or optical-fiber cable. For communication over greater distances, high transmission rates up to 2 megabits per second (Mbps) are achieved with fiber-optic cable where available, or with special "Digital Subscription Line (DSL)" and "T1" lines that can be leased from the telephone company. Also, microwave carriers can be used either directly for line-of-sight transmission or by satellite transmission for longer distances. The transmission of medical images over long distances is termed *teleradiology*. A teleradiology capability allows medical

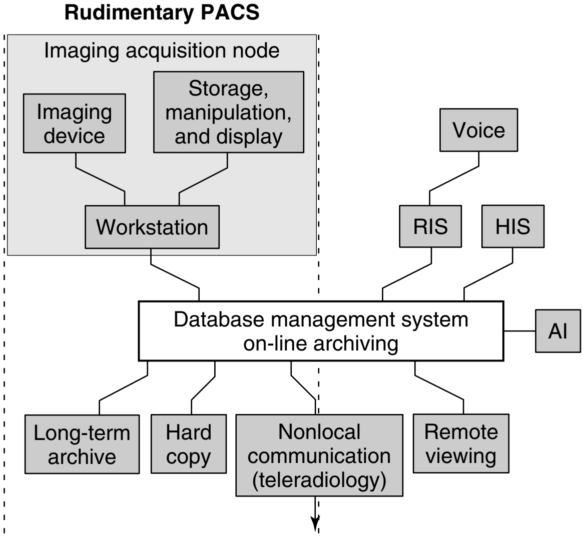
Rudimentary PACS
Imaging acquisition node
Imaging device
Storage, manipulation, and display
Workstation
Voice
RIS
HIS
Database management system on-line archiving
AI
Long-term archive
Hard copy
Nonlocal communication (teleradiology)
Remote viewing

Fig. 27.11 A picture archiving and communications system (PACS) in the "star" configuration. The imaging devices and workstations (each of which is under minicomputer control) and the other components all connect to a central computer. (From Wolbarst, A.B. (2000) *Physics of Radiology*, Medical Physics Publishing, Madison, WI.)

personnel to receive and respond to images on any compatible workstation, even at home; to communicate and share images effortlessly with other specialists; to obtain input from around the world on difficult cases; and to provide diagnostic expertise in remote areas. The Internet is already providing an invaluable conduit for telemedicine services, and further applications will evolve as the more powerful and versatile next-generation Internet and Internet2 become increasingly accessible.5) The medical value of next-generation Internet and Internet2 depends on there being in place agreements on a medical information (or more general) protocol to define interface uniformity, data formats, semantics, and other details.

Once in a PACS, an image may be routed to *image-processing* modules that can electronically enhance its contrast to accentuate variations in intensity among disjoint regions, smoothing out distracting random or systematic noise without excessive loss of real information, or by sharpening image resolution by emphasizing the edges of structures such as contrast-filled blood vessels. The image can then be sent to short- or long-term *storage*, and recalled almost instantly for display at any connected workstation. The image can be fetched, along with CT, nuclear medicine, or MRI studies of the patient, for comparison and correlation among the various studies. It can be examined by a *CAD* system to locate irregularities, and perhaps even a *CAD*x module that offers a statistical-based differential diagnosis. And that is just for starters!

27.3.2
Digital Display

The rapid evolution of digital technologies will continue to have a tremendous impact on imaging, and one ever-growing area of influence will undoubtedly be in systems for image display. For reasons of image quality, weight, size, required voltage, ruggedness, energy efficiency, and cost-effectiveness, flat-panel display solid-state technology (which greatly facilitated the development of the flat-panel imagers of DR) is rapidly displacing the traditional cathode-ray tube, not only in medical displays but also in computer monitors and high-definition television sets (Badano, 2004; Badano, Flynn and Kanicki, 2004; Samei, Badano, and Chakraborty, 2005). At present, liquid crystal panels, each with several million pixels, account for something like 99% of all flat-panel displays produced. However, other approaches, such as plasma, field emission, and electroluminescent panels, are also under development. Flexible, rewritable, plastic-based electronic "paper" is already available – it remains to be seen what applications arise for it in medical imaging.

True three-dimensional viewing, moreover, is now a reality. The reader may have seen stereoscopic drawings and movies years ago in which two slightly different perspectives of an object were presented in lines of blue and red; when one wore special glasses with one red and one blue lens, the spears and arrows appeared to come right out of

---

5) The next-generation Internet protocol (IPng), which is a new version of the current Internet protocol (IP), is designed to deal with the addressing and routing problems arising from the massive size and continuing expansion of the system. Internet2 is a not-for-profit consortium of universities and companies that are collaborating on these, and related, issues.

the screen! Now there are electro-optical glasses in which the lenses themselves can display slightly different electronic images for the two eyes – the images are perceived to be fully three dimensional and can be of value in virtual reality applications such as surgical planning or virtual endoscopy. Holographic displays, in which the viewer does not even need glasses, are in the early stages of development.

27.4
Digital Planar X-Ray Imaging

For many decades, the combination of intensifying screens and photographic film was the primary method for capturing and recording X-ray images. It is simple, portable, and inexpensive, and yields images with excellent anatomic detail. This approach, however, has disadvantages as well. It is limited to a narrow range of acceptable exposures, and it offers almost no flexibility for image processing. Since the same film both captures and displays the X-ray image, neither of the processes can be separately optimized. Film is bulky to store and easy to misplace, and it must be transported physically from one location to another. Replacing analog X-ray image production with digital eliminates nearly all these problems, and adds a number of clear benefits, as well.

Some clinics get a first taste of the capability of new imaging technologies by digitizing film radiographs with a scanning micro-optical densitometer. This device permits some compensation for regions of over- or underexposure, and also entry into a PACS, with all the advantages of digital technology. But to reap the full benefits of digital imaging, it is necessary to begin with a digital image receptor in the first place. While there are many variations on the theme, two general types of digital X-ray projection imaging dominate the scene at present: *computed radiography* (*CR*) and *digital radiography* (*DR*).

One great advantage of both the technologies (and of digital imaging in general) is the linearity of response of the radiation detectors over exposure ranges that are orders of magnitude broader than the latitude of any film, and the overall improvement in image quality that results from that property; there are no under- or overexposed regions of the image. Another is the separation of image acquisition from imaging processing (windowing, edge enhancement, noise reduction, etc.) and from display, so that these three steps can be independently optimized; with film, about the only processing available is the retake. And finally, there are the overall benefits of digital signal storage, communication, and analysis.

In planar X-ray imaging, new semiconductor-based digital image receptors are gaining rapid acceptance as viable replacements for screen-film combinations and for image intensifier plus television monitor or CCD optical camera (Goldman and Yester, 2004; Zhao, Andriole and Samei, 2006). Many imaging departments have gone filmless and fully digital, and are restructured around a central *PACS*. This trend is likely to continue wherever modern medical technologies are available (Thrall, 2005).

27.4.1
Computed Radiography (CR)

CR is the digital counterpart to the screen-film approach to X-ray imaging.

Appearing first in the 1970s, *photostimulable phosphor* (*PSP*) *imaging plates* (*IPs*), now composed typically of BaFBr and BaFI, are housed in an inert, light-tight cassette for exposure to X-rays. Areas exposed to higher levels of radiation will store excess charge in elevated energy states, from which they can later be dislodged with a suitable external stimulus.

Cassettes come in a range of sizes and shapes, and they are both easy to use and relatively cost effective. To convert from analog to CR, one need only exchange of a film cassette assembly with a very similar CR cassette, and the chemical film-processor for an electronic IP reader.

Inside the light-tight image processor, an exposed IP is scanned with a laser beam point of red light that, in turn, stimulates the IP to emit light of another color (blue). The emitted light is detected by a photomultiplier tube (PMT) and digitized with an 8- or 16-bit analog-to-digital converter. During scanning, the computer keeps track of the location and the light emission of each scan point, and transforms this information into a two-dimensional image. The digitized image is displayed on a video monitor and can be stored for later recall.

At present, the major limitation of the system is that, like a screen-film system, it can acquire only static images.

27.4.2
Digital Radiography (DR)

Active-matrix flat-panel imagers (*AMFPIs*) of DR and digital fluoroscopy (DF) overcome the static image limitation of CR. An AMFPI is an array of hundreds of thousands, or millions, of tiny, electrically independent semiconductor detectors that come in two general forms: they are either directly sensitive to high-energy photons (direct-detection array), or responsive to light from an overlying thin layer of fluorescent material (indirect-detection array) (Antonuk, 2004; Zhao and Zhao, 2005).

With *direct-detection arrays*, the image receptor consists of an array of several million voxel-sized photoconductive (e.g. amorphous selenium, cadmium zinc telluride, and lead iodide) tiles, each with a pair of electrodes on its face and a *thin-film transistor* (*TFT*) attached on the back side. When an incident X-ray photon undergoes a photoelectric interaction with the photoconductor layer of a volume element (voxel), there results a trail of ionizations through it. The high-voltage bias applied by the pair of electrodes sweeps the liberated electrons and holes rapidly through the photoconductor layer, and this process registers as a pulse of charge in the TFT. The set of TFTs is then read out through data lines threading throughout the detector. The data are organized by the computer and displayed as an image on the corresponding picture element (pixel) array of the display.

With *indirect conversion*, a scintillation layer such as CsI in the form of columnar (tall and thin) crystals converts the incident X-ray energy to flashes of light. An underlying matrix of amorphous-silicon photodetectors and their TFTs, one combination per voxel, converts the pattern of visible light to an electrical signal.

For most if not all X-ray applications, digital detectors yield images comparable in quality to those obtained with analog film-based techniques. Even for breast, where resolution is essential for the detection and assessment of microcalcifications, digital systems are becoming increasingly accepted (Pisano *et al.*, 2005; Fischmann *et al.*, 2005; Brateman and Karellas, 2006). A number of recent developments should significantly broaden the domain of DR. One of the more intriguing, ink-jet (rather than photographic) printing will permit fabrication of flat-panel imagers (and flat-panel displays!) on flexible plastic substrates of virtually any size and shape.

There are advocates of both CR and DR, but the two are complementary modalities and there is room for both (Monnin *et al.*, 2005). Both are likely to be around for awhile, providing quality images of teeth, chest, bone, gastrointestinal tract, extremities, and breast (Kapur *et al.*, 2004; Pisano and Yaffe, 2005). CR is considerably easier and less expensive to implement than DR, and its ruggedness makes it more suitable for mobile units. Only DR, however, can image medical events fluoroscopically in real time.

Two areas where DR has found considerable success are digital subtraction radiography and digital tomography.

27.4.3
Digital Subtraction Radiography

Computer subtraction of pairs of digital images taken at different times can often reveal changes, and even rates of change, that are not apparent in the separate images. In a simple example of *digital subtraction angiography* (*DSA*), an image is obtained midway during injection of an iodine-based contrast agent (Figure 27.12a). Three subsequent images (Figure 27.12b–d) are obtained at progressive times during contrast agent injection and subtracted from image (a) to yield three subtraction images that reveal only what has changed, namely the increased presence of contrast agent. Even fine blood vessels that are invisible in the prior images alone show up clearly in the *temporal subtraction* images.

Another technique, *energy subtraction*, involves the acquisition of images at two or more different kVps (Richard *et al.*, 2005).

In the resulting subtraction images, visual background noise, such as irrelevant bonelike or soft tissue structures that obscure what the clinician is looking for, is removed (Figure 27.13). A related application is dual-energy X-ray absorptiometry, which is the standard method for measurement of bone density, the major predictor of fracture from osteoporosis. Dual-energy X-ray absorptiometry is now being challenged, however, by US and other techniques (Khaw *et al.*, 2004; Wear *et al.*, 2005).

27.4.4
Digital Tomosynthesis

Digital tomosynthesis takes its analog predecessor into the modern era (Dobbins and Godfrey, 2003; Wu *et al.*, 2004). A series of DR images is acquired while the X-ray tube and beam are pivoted through a small angle. By largely blurring out the over- and underlying planes, high-quality images of each selected tomographic plane can be obtained with section thicknesses ranging from one millimeter to several centimeters (Figure 27.14). Digital tomosynthesis has been applied to chest, breast, angiography, orthopedic, dental, and other examinations. It is much less costly than CT and often produces equally useful information.

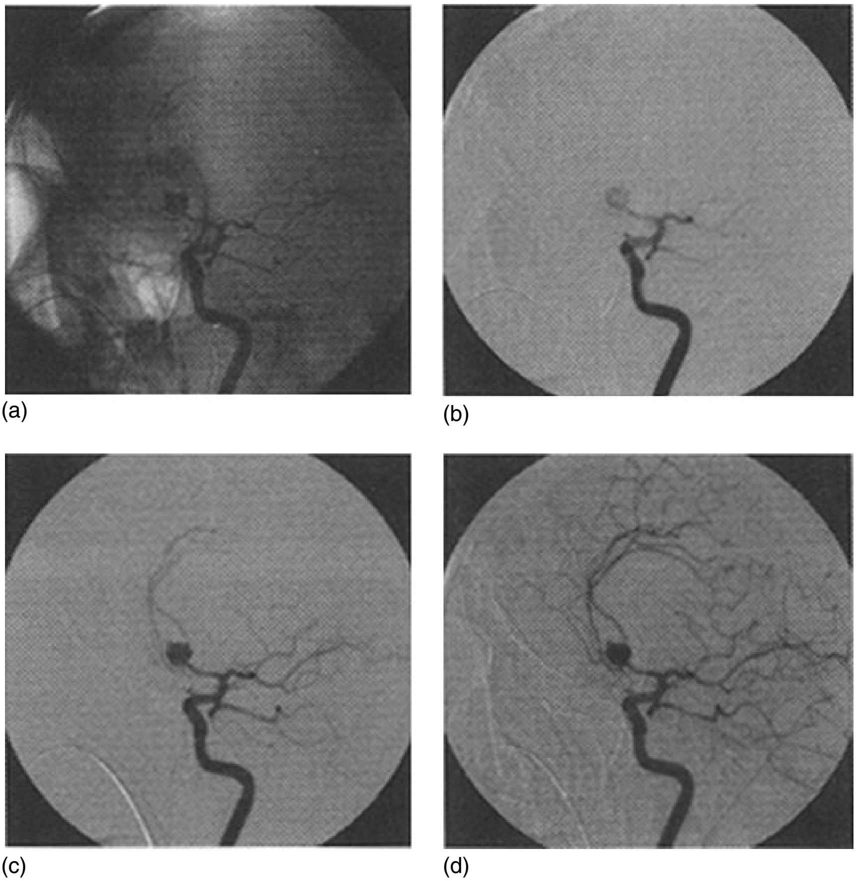

(a) (b) (c) (d)

Fig. 27.12 Digital subtraction angiography (DSA). Cerebral arteriogram (oblique transorbital view) shows an aneurysm of the anterior communicating artery (junction of the A_1 and A_2 segments of the anterior cerebral artery). (a) Unsubtracted original digital image obtained midway through the contrast material injection. (b–d) Subtracted DSA images obtained at three progressive time points during contrast material injections. (From Pooley, R.A., McKinney, J.M., and Miller, D.A. (2001) The AAPM/RSNA physics tutorial for residents: digital fluoroscopy. *Radiographics*, 21(2), 521–34, with permission.)

27.5
Computed Tomography

As with temporal or dual-energy digital subtraction techniques, the power of CT lies in its ability to remove irrelevant, unwanted, and interfering information.

A radiologist may be able to infer the three-dimensional shape of an object in a patient by viewing several radiographs taken from different angles. Likewise, CT involves acquisition of a large number of X-ray projection views. It then works backwards, in effect, to reconstruct the spatial distribution of materials – or, more precisely, the distribution of their X-ray photon attenuation characteristics – that gave rise

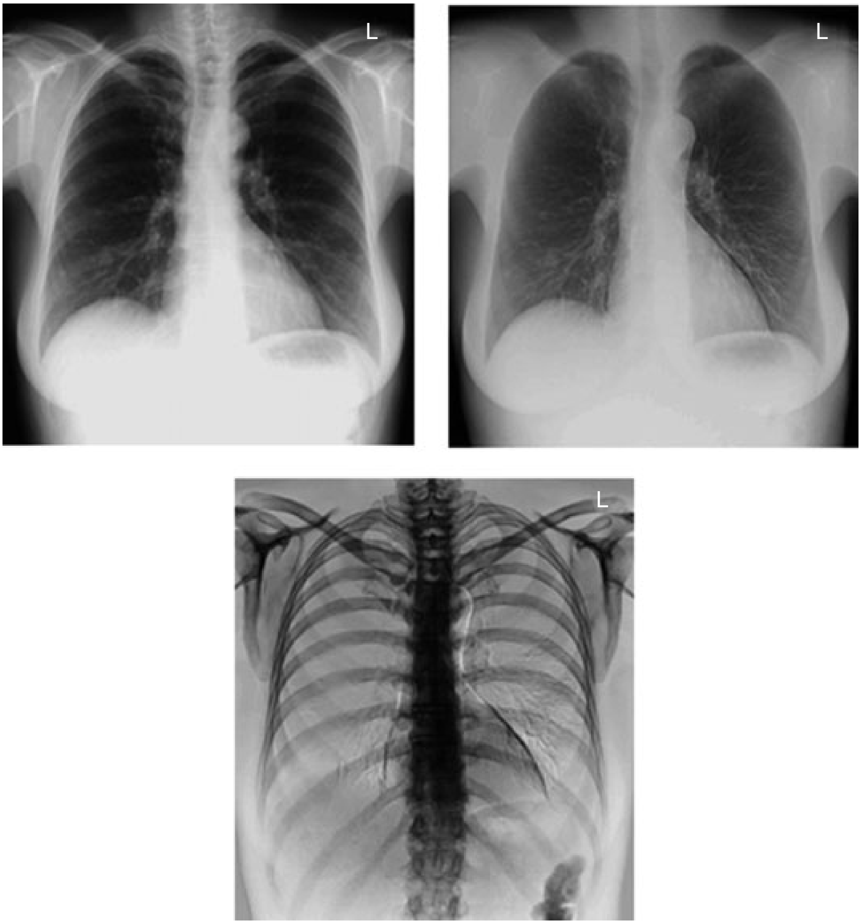

Fig. 27.13 X-ray images of the thorax in which bone (upper right) and soft tissue (lower center) have been subtracted.

to the projections (Fohr, Cody and McCollough, 2006).

In 1972 the EMI Group, a British company best known as home to the Beatles, announced the first commercially available X-ray transmission CT unit for studies of human anatomy. Seven years later, Hounsfield, the developer of this unit (Hounsfield, 1973), shared the Nobel Prize in Medicine with Cormack, a physicist (Cormack, 1963), who constructed a bench-scale prototype and did early work on CT image reconstruction from projections.

Hounsfield's device, oriented at some angle θ, scanned a narrow pencil beam of X-rays across the patient, thereby cutting out a thin slice of tissue, and moved a scintillation detector in synchrony with the X-ray beams on the opposite side of the patient (Figure 27.15). The angle was then incremented by a small amount,

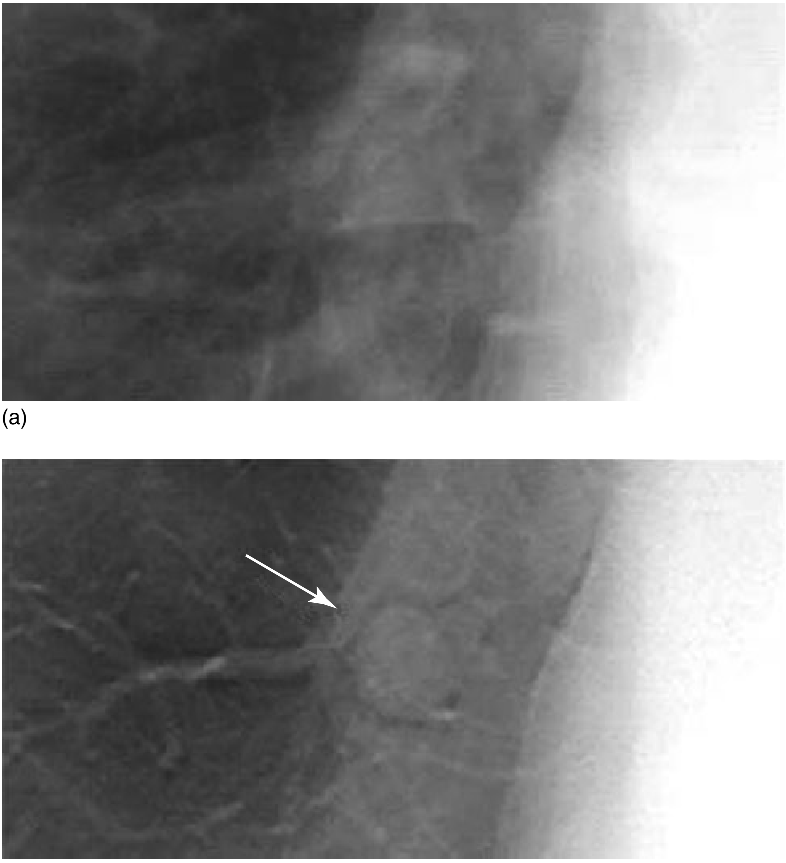

Fig. 27.14 Digital tomosynthesis of human lung. (a) Posteroanterior radiograph shows region of interest. (b) Digital tomosynthesis section demonstrates a 15-mm pulmonary nodule (arrow) that was not visualized in (a). Note also improved clarity of vascular detail in (b), which was reconstructed with a matrix inversion tomosynthesis technique that used 61 projection images acquired in 10 seconds over a total tube swing angle of 16°. Fifty-nine sections were generated with 3 mm spacing. Total subject entrance exposure was approximately the same as that for a screen-film lateral chest radiograph. (Image courtesy James T. Dobbins III, PhD, H. Page McAdams, MD, and Devon J. Godfrey, Ph.D., Duke University Medical Center.)

and the next scan begun. The intensity, $I_{\text{transmitted}}$, of X-rays that passed through the patient and were measured by the detector was, at any instant,

$$I_{\text{transmitted}}(\theta, q)/I_0 = \exp\left(-\sum \mu_i \Delta x_i\right) \quad (4)$$

where μ_i is the linear attenuation coefficient of the ith voxel lying on the beam path, q parameterizes the position of the beam as it carries out the scan, I_0 is the beam intensity in the absence of the patient, and $\sum \Delta x_i$ is the thickness of the patient for that particular pencil beam. With multiple

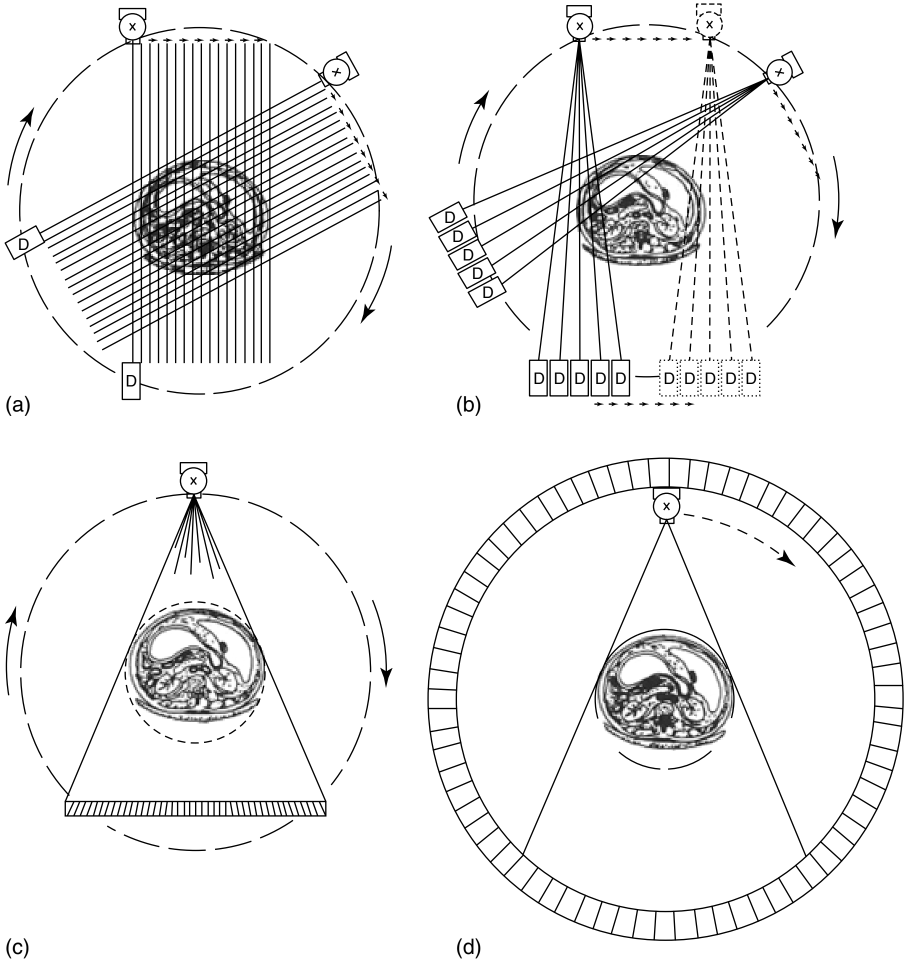

Fig. 27.15 Scan motions in computed tomography. (a) first-generation scanner using a pencil X-ray beam and a combination of translational and rotational motion; (b) second-generation scanner with a fan X-ray beam, multiple detectors, and a combination of translational and rotational motion; (c) third-generation scanner using a fan X-ray beam and smooth rotational motion of X-ray tube and detector array; (d) fourth-generation scanner with rotational motion of the X-ray tube within a stationary circular array of 600 or more detectors. (From Hendee, W.R. and Ritenour, E.R. (2002) *Medical Imaging Physics*, 4th edn, John Wiley & Sons, New York.)

transmission measurements obtained at different orientations, the data can be *reconstructed* to reveal all the individual values of μ_i. The reconstruction yields a two-dimensional map of linear attenuation coefficients throughout the slice under examination.

The X-ray tube and scintillation detector of the EMI *first-generation* CT unit made measurements of X-ray transmission every $\Delta q = 2.5$ mm over the 40 cm linear path to yield a total of 160 transmission data points. Then the angular positions of the X-ray tube and detector were advanced by $\Delta\theta = 1^\circ$, and another scan of 160 measurements was obtained. This process was continued over 180°, providing 28 800 transmission measurements. These data were then transmitted to a computer equipped with a mathematical package for reconstructing a two-dimensional matrix of attenuation coefficients across the scanned section on a 160 × 180 pixel matrix.

It is convenient to transform this matrix into an image based on *CT numbers*, defined as

$$\text{CT number}_i = 1000\ (\mu_i - \mu_w)/\mu_w \quad (5)$$

where μ_w is the linear attenuation coefficient of water for the X-ray energy, a process that almost entirely eliminates any photon-energy dependence of the image.[6] The CT number for water is zero *Hounsfield units* (*HUs*), and that for air nearly −1000. The CT number would be about +1000 for a structure such as compact bone or colon containing barium contrast medium, because the physical density would be almost twice that of water. The CT numbers can be represented as various shades of gray to provide a visual display of the distribution of tissues. As with any other digital modality, the image appearance can be optimized by *windowing*, or shifting the relationship between CT number and voxel brightness on the display (Figure 27.16). With dual- or multienergy techniques, CT can provide information on tissue parameters beyond simple pixel HUs.

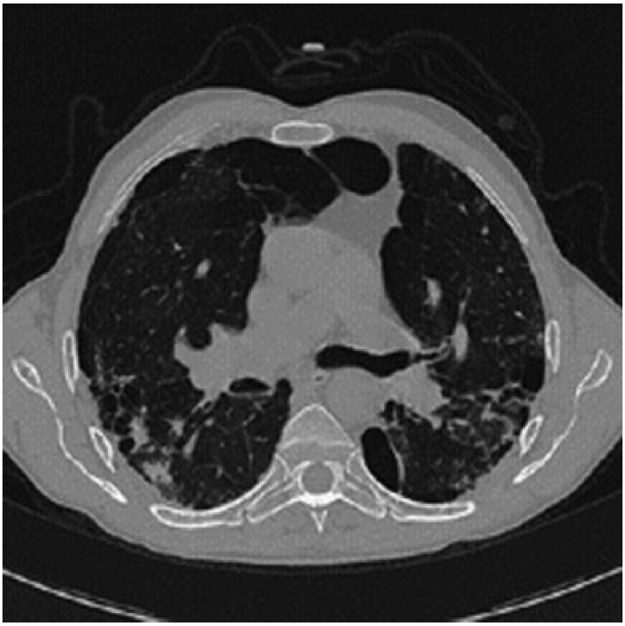

Fig. 27.16 Axial X-ray CT image of the thorax.

---

[6] The energy dependencies of the attenuation coefficients for soft tissue and water are nearly identical.

27.5.1
Reconstruction Algorithms

The general mathematical approach to CT image reconstruction was described in 1917 in a remarkable work by the Austrian mathematician Radon (Radon, 1917). The model was first used in applications of image reconstruction to radio astronomy, electron microscopy, and optics.

The process of data acquisition may be modeled by the operator equation

$$p = \hat{\mathbf{H}} f \tag{6a}$$

The function f refers to the spatial distribution of attenuation coefficient values throughout the object being examined, and $\boldsymbol{p}$ describes the set of X-ray intensity projection measurement values that it yields under the Radon operational process $\hat{\mathbf{H}}$.

For digital computations, the object space is discretized into a square (for convenience, here) matrix of $M \times M$ voxels. f may be viewed as an M^2-dimensional vector, each element of which corresponds to the attenuation coefficient value of the associated voxel. The angle θ can assume M values and, for each of them, the line parameterized by q is partitioned into M bins. There will be a total of M^2 bin values for the entire set of measurements, and it is these values that comprise the M^2-dimensional vector $\boldsymbol{p}$. An element, h_{ji}, of the *transfer* or *transition matrix*, $\hat{\mathbf{H}}$, indicates the contribution of the object's ith voxel, f_i, to the jth component of $\boldsymbol{p}$, p_j, for the corresponding values of θ and q.

The problem to be solved, the "inverse problem," is "given the set of measurements $\boldsymbol{p}$ and a model of the projection process $\hat{\mathbf{H}}$, estimate f." That is, find $\hat{\mathbf{H}}^{-1}$ such that, in effect,

$$f = \hat{\mathbf{H}}^{-1} \boldsymbol{p} \tag{6b}$$

Finding the solution exactly by way of the Radon transform is generally not feasible mathematically, and most approaches rely on various approximations. This clearly must be so when statistical noise is added to the process. Everything becomes even more complicated with complex geometries, such as 3D cone- (rather than a flat, 2D fan-) beam CT.

When the reconstruction problem is framed in this fashion, it is possible to rely on a rich body of work dedicated to solving linear systems of equations. Initially, reconstructed images were computed with *algebraic* techniques to solve the set of simultaneous linear equations of Eq. (1). As pixel matrices grew in size, however, the number of linear equations to be solved became unwieldy. Similarly, *iterative* mathematical models became too time consuming for the computers available at the time, even though they are now staging a comeback in SPECT and PET imaging. Faster algorithms designed for *filtered back-projection* (*convolution*) reconstruction have taken over the field (Kak and Slaney, 1988). These algorithms rely on a Fourier transformation of X-ray projection data into frequency space, and the use of ramp and cutoff-frequency filters to improve image quality (Barrett and Myers, 2004).

27.5.2
Generations of Scanners

The combination of translational and rotational motion employed in early CT units yielded satisfactory CT images, but required several minutes for data acquisition. This limited the method to studies of structures such as the head and extremities where voluntary and involuntary motion of the patient could be controlled. There was great interest in applying CT to thoracic and abdominal studies, of course, but the scan time had to be decreased from minutes to seconds for these areas to yield satisfactory images. Major efforts to reduce scan time led to the development of successive generations of CT scanners (Figure 27.15). *Second-generation* scanners, in which an array of detectors (rather than a single one) tracked the linear motion of the X-ray beam (Figure 27.15b) are now extinct, as are their Hounsfield-type predecessors.

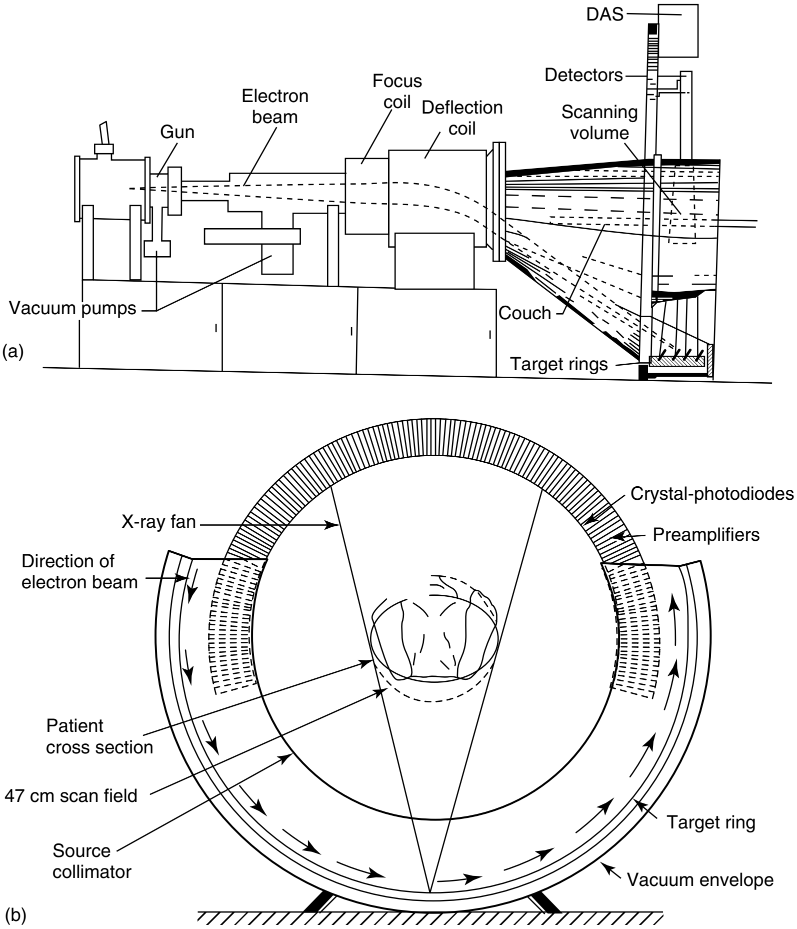
DAS
Detectors
Electron beam
Focus coil
Deflection coil
Gun
Scanning volume
Vacuum pumps
Couch
Target rings
(a)
Crystal-photodiodes
Preamplifiers
X-ray fan
Direction of electron beam
Patient cross section
47 cm scan field
Source collimator
Target ring
Vacuum envelope
(b)

Fig. 27.17 Longitudinal (a) and cross-sectional (b) view of the Imatron Ultra-Fast Computed Tomographic (UFCT) scanner. (From Gould R.G. Computed tomography overview and basics, in Vol 2, Proceedings of the American Association of Physicists in Medicine, *Specification, Acceptance Testing and Quality Control of Diagnostic X-ray Imaging Equipment*, July 1991, Santa Cruz, California, pp. 930–68, with permission.)

The purely rotational *third-generation* design (Figure 27.15c), and a *fourth-generation* scanner, in which an X-ray tube rotates but the ring of X-ray detectors remains stationary (Figure 27.15d), are still used in many clinics. With both types of rotational scanners, data for a single section can be acquired in just a few seconds, and there are true believers in each design.

Although newer CT scanning geometries shortened the time to acquire images, none permitted data acquisition in a time (<0.1 second) short enough to capture the heart and some other blood-perfused organs without significant degradation caused by motion. As the first significant attack on this problem, a *fifth-generation* CT machine produced X-ray projection data without mechanical motion of the scanner. The heart of the system is a very large, funnel-shaped X-ray tube in which the anode is a stationary, concave metal strip that partially circumscribes the patient (Figure 27.17). The electron beam is swung along the anode in under 50 ms, fast enough to freeze the heart in any part of its cycle. Also, its dose is low, about 1 mSv for angiography and even less for coronary calcium scoring. While revolutionary in its time, electron-beam CT images tend to be noisy, and the far more versatile multisection CT has better spatial resolution in all three directions (about 0.3 mm vs. 1.5 mm for electron-beam CT). Fifth-generation electrons-beam CT units are principally used today only for cardiac imaging, and even for those applications they are being displaced by multislice CT units.

27.5.3
Helical and Multidetector-Row CT

Two developments over the past few years have substantially increased the efficiency and speed of data acquisition in CT (Jiang, 2003).

In early *axial-mode* (nonhelical) CT systems, the X-ray tube is rotated once about an immobile table; the table is then advanced by a small amount in the *z*-axis direction, and the tube is rotated once in the opposite direction, to untangle the high-voltage cables. This procedure is repeated over and over, with each rotation producing data for a flat plane of tissue. *Helical* (*spiral*) *CT*, sometimes called *sixth generation*, allows the table to move continuously as the tube rotates in a constant direction

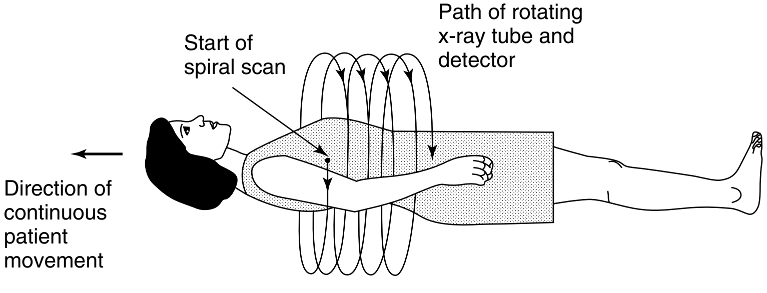
Path of rotating x-ray tube and detector
Start of spiral scan
Direction of continuous patient movement

Fig. 27.18 Spiral CT scanning (From Rowlands, J.A. (1999) X-ray imaging: radiography, fluoroscopy, computer tomography, in *Biomedical Uses of Radiation. Part A: Diagnostic Applications* (ed. W. Hendee), Wiley-VCH, Weinheim, Germany, with permission.)

around it. The technology is based on either third generation, with a rotating array of detectors, or fourth generation, in which a thousand or so very small detectors surround the patient and are fixed in space. *Slip ring* electrical contacts are needed to deliver power to the tube and, for third-generation machines, to retrieve signals from the detectors (Figure 27.18). Helical CT allows much faster scanning and provides images with higher resolution in the transverse plane.

The second important innovation is the ability to image multiple slices, rather than only one, for each rotation of the gantry (Cademartiri *et al.*, 2004; Saini, 2004). The single, narrow, transverse fan beam of X-rays used for axial-mode CT is broadened into a *cone beam*; each of the thousand or so detectors circling the patient is replaced with a row of 4–64 (128, 256, . . .) (or more in the future) separate detectors aligned parallel to the table axis – it is as if 4–64 thin belts, each with a row of a thousand small, closely spaced detectors, circumscribes the patient. This approach to efficiency in data acquisition makes it possible for a *multidetector computed tomography*[7] (*MDCT*) unit to acquire images of up to 64 sections of the body simultaneously at present with even more simultaneous slices predicted for the future.

MDCT (*seventh generation*) scanners use a classic third-generation scan geometry but with a cone- rather than a fan-shaped X-ray beam. For 16 or fewer sections, it is possible simply to employ well-established multiparallel-section approximate reconstruction algorithms. More sophisticated approximate and exact cone-beam algorithms have been created to yield better images, but they are computationally more complex and time consuming (Katsevich, 2004; Shechter *et al.*, 2004; Flohr *et al.*, 2005; Mori, Endo and Obata, 2005). These challenges are a target of considerable research at the present time.

CT images are displayed most commonly as cross-sectional representations of the body. By examining contiguous cross-sectional images in sequence, radiologists can intuitively integrate the imaging information into a three-dimensional model of anatomy. This integration can be accomplished also by reconfiguring the data set in three dimensions. By mathematically compiling data across planes that cut through this database in preferred directions, coronal and sagittal views, as well as those at oblique angles, can then be produced. Also, the three-dimensional database can be analyzed to yield three-dimensional images of high quality. Three-dimensional CT angiography, for example, is an increasingly important, and much less invasive and costly alternative to standard catheter-directed coronary arteriography for the demonstration of coronary artery disease, the detection and measurement of the volume of non-calcified plaques, and other applications. Similarly, virtual colonoscopy is coming to displace the real thing.

CT manufacturers continue to strive for greater scanning speed and more longitudinal anatomic coverage with each gantry rotation. Because centrifugal forces would become overwhelming, the speed of gantry rotation may not increase much beyond its present value of three cycles per second. This is good enough to examine the lung (Rietzel, Pan and Chen, 2005; Sonke *et al.*, 2005) or capture a heart at systole or diastole with little motion blurring. Shorter scan times might be achievable with multiple tubes. (Fast MRI can also support cardiac screening or image-guided intervention, and without ionizing radiation, but at significantly greater cost.) Also desirable in CT scanning is extended *z*-axis coverage that would allow functional imaging of the entire heart, lung, brain, and other structures with a single rotation. These innovations, and the development of better detector systems, will further improve resolution in all three directions so that CT will become a truly isotropic 3D imaging modality, similar to MRI. A potential difficulty with this is the increase in scatter radiation reaching the image receptors (Shikhaliev, Xu and Molloi, 2005; Siewerdsen *et al.*, 2004).

---

[7] Also known as a *multislice* and *multiring* scanner and by other names.

27.5.4
Recent Developments in Computed Tomography

Researchers have described devices in which the multiple rings of detectors have been replaced with a flat-panel imaging array (Siewerdsen *et al.*, 2004; Mahnken *et al.*, 2005). Several manufacturers have produced flat-panel prototypes, and quasi-CT C-arm devices are being marketed. However, there are daunting problems associated with the very high rates (gigabytes per second) of data transfer.

Meanwhile, a novel design of a standard X-ray tube has been introduced in which nearly everything *except* the anode, in effect, is rotated. This allows direct cooling of the anode, from its back side, that is five or so times faster than that possible with a standard rotating anode tube. This innovation permits either longer uninterrupted exposure durations or more intense (hence, briefer) pulses (Schardt *et al.*, 2004).

There is strong pressure to reverse the upward "radiation dose creep" for all X-ray technologies, and especially for CT of children and for mass screening applications. Fortunately, most operators of imaging devices are aware of the need to tailor technique factors to the dimensions of the patient, which can reduce the dose dramatically, especially to children (Kalra *et al.*, 2004; Vetter, 2005).

Finally, there is a downside to the overabundance of imaging riches: What does one do with all images it is now possible to obtain? For a single patient, it may be necessary to display hundreds, perhaps even thousands, of CT sections, and there may be a number of such patients each day. A radiologist clearly cannot examine all these images one by one, despite the possibility that a critical clinical sign may appear in any one of them – with serious medical and legal implications. Although some studies may require very high resolution, such as CT mammography or CT virtual colonoscopy (Imperiale, 2005), buying a multimillion dollar MDCT primarily to create smoother surface renderings of organs generally makes little sense. So it is essential to develop computer-based *bioinformatics* programs that can locate, and perhaps interpret, clinically essential visual items that are obscured within the reams of imaging data accumulating daily (Dodd *et al.*, 2004).

27.6
Nuclear Medicine

Many atomic nuclei are unstable because they contain combinations of neutrons and protons that cannot be held together by the combined action of nuclear and EM forces or because they possess more energy than needed to exist in their ground state. In the latter case, the nuclei can gain stability by releasing their excess energy, most commonly in the form of γ-rays. Because most γ-rays can penetrate tissue in a manner similar to X-rays, they can be detected outside the body after their emission from radioactive materials present inside the body. The procedure of administering a radioactive substance to a patient, and then using radiation detectors to measure or image the γ-rays emitted from the patient, is known as *nuclear medicine* (Cherry, Sorenson and Phelps, 2003; Chandra, 2004).

Nuclear medicine originated in the 1920s and 1930s during studies of the behavior of certain *radiopharmaceuticals* (chemical compounds containing radioactive atoms) in specific organs such as the thyroid and kidneys. A major impetus to nuclear medicine occurred after World War II (WW II) when nuclear reactors became available for the wide-scale production of *radionuclide*s (radioactive atoms) that were potentially useful in diagnostic medicine. Two inventions of the early 1960s greatly advanced nuclear medicine as a diagnostic specialty – the γ-ray scintillation imaging camera and the "molybdenum cow" for on-site generation of the particular radionuclide technetium-99m, which remains the standard workhorse of nuclear medicine. In the 1970s, the introduction of the computer permitted studies of the dynamics of physiologic processes and, in the 1980s, computers made possible the complex reconstruction computations underlying computed tomographic techniques. Today nuclear medicine, including its *SPECT* and *PET*, is one of the most sophisticated of the imaging technologies, and is essential to the detection and diagnosis of a variety of diseases and injuries.

Although a nuclear imaging study cannot compete with radiography, CT, or MRI in revealing precise anatomic detail, it may more than compensate by providing potentially invaluable information about the physiologic status of an organ or tissue.

27.6.1
Radiopharmaceuticals and Technetium-99m in Particular

A radiopharmaceutical consists of two parts: (i) an agent designed to concentrate, through chemical or other physiologic means, in a specific organ or compartment of the body and (ii) a radioisotope (most commonly by far, technetium-99m), which emits a medium-energy γ photon. The ideal radiopharmaceutical would satisfy a number of stringent requirements. First, it must be possible, usually with the aid of inexpensive commercial "kits," to combine the agent to the radionuclide quickly and easily. The radiopharmaceutical thus prepared must remain intact chemically after administration, concentrate rapidly in the organ or biochemical compartment of interest, and be eliminated rapidly from the organ after imaging. Preferably, the radioisotope (radionuclide) will emit only γ-rays (no α or β particles) and these should be energetic enough to be likely to escape the body, yet of low enough energy to interact with the detector material. And, its *half-life* (the time required for half of the atoms in a sample to undergo radioactive decay) should be long enough to allow performance of the necessary clinical chemistry and concentration in the biological compartment of interest, yet not so long that the patient continues to irradiate himself and others long after the study has been completed.

Several radiopharmaceuticals satisfy these criteria for the study of specific organs and tissues in the body (Owunwanue, Patel and Sadek, 1995; Saha, 2001). However, the nuclide employed most frequently is technetium-99m.[8] ^{99m}Tc emits 140-keV γ-rays, an almost ideal energy for escape from the body and for detection with a scintillation camera. There is little particulate radiation emitted along with the γ-rays. The half-life of technetium-99m is 6 hours, long enough to conveniently synthesize a variety of radiopharmaceuticals with desirable physiologic properties but with little hazard after a day or so.

Another major advantage of technetium is its production by decay of molybdenum-99, an easily produced radioactive nuclide with a half-life of 66 hours. The ^{99}Mo can be adsorbed to an ion-exchange column, and the ^{99m}Tc is removed daily in the clinic's laboratory by *eluting* (washing out) the column with saline solution – a routine commonly known as "milking the Mo cow." Most nuclear medicine departments receive one or more of these *technetium generators* (affectionately referred to as *technetium cows*) each week from a commercial supplier.

At least 85% of all nuclear medicine studies are performed with various technetium-based radiopharmaceuticals for studies of lung ventilation and perfusion, cardiac dynamics, liver integrity, kidney function, the presence of cancer metastases in bone and various organs, and a number of other conditions. Other radioactive nuclides that find specific applications include iodine-123 and iodine-131 for studies of the thyroid, thallium-201 for detection of impaired heart muscle, gallium-67 for identification of soft tissue tumors, and fluorine-18 for PET studies. The use of radiopharmaceuticals for medical diagnosis and evaluation is essential to the care of millions of patients each year, and the quality of their health care would be severely compromised if nuclear medicine procedures were not readily available.

27.6.2
Scintillation Gamma Camera

Almost all non-PET nuclear medicine images are obtained with the gamma camera. The action begins within the *scintillation detector*, a wide, thin single crystal typically of thallium-activated sodium iodine, NaI[Tl]. The crystal may be 0.6 cm (1/4 in.) to 1.0 cm (3/8 in.) thick, and 25 cm (10 in.) to 60 cm (24 in.) across. When struck by a γ-ray emerging from the patient's body, the scintillator emits a flash of light of intensity proportional to the photon energy. The backside of the crystal is observed by a honeycomb array of up to 91 *PMTs*. Each PMT views the flash from a different distance, and therefore sees a different intensity, and a comparison of the output voltages from all the tubes yields an estimate of where the flash occurred. This process also determines the total light emission for the flash, and hence whether it came from a primary γ-ray or a lower-energy Compton scatter photon from an interaction in the patient or elsewhere in the crystal. A pulse-height analyzer then rejects the smaller electronic signals produced by scattered γ-rays, thereby limiting the recorded interactions to only those produced by primary γ-rays.

---

[8] The nucleus left behind may still be unstable and therefore still radioactive.

Knowing where γ-ray interactions occur in the crystal must be correlated with the origin of the γ-rays in the patient. This correspondence is achieved by means of a metal *collimator* placed on the patient-side of the scintillation crystal. Most often employed is the parallel multihole collimator, and with such a device a flash detected in the crystal must have been produced by a γ-ray that originated directly below it and traveled straight along one of the collimator holes; γ-rays heading off in other directions are absorbed by the lead septa that separate the holes. In this fashion, the distribution of interactions across the crystal face reflects the distribution of radioactivity in the region of the patient being viewed by the camera. Other types of collimators that are useful in limited applications include the pinhole collimator (particularly helpful in thyroid imaging) and the converging collimator (used primarily for cardiac studies).

Planar nuclear images can reveal not only the static distribution of radioactivity in a particular organ or region but also the transient passage of radioactivity into and out of it. That is, data processing by computer permits quantitative analysis and display of the behavior of the radiopharmaceutical over time. One can partition the cardiac cycle into 25 sequential frames, for example, and, triggered by an electrocardiogram, accept counts in each for 40 ms – and repeat this enough times to acquire sufficient counts in each frame for a high-quality motion picture of the beating heart.

27.6.3
Single Photon Emission Computed Tomography (SPECT)

SPECT is the γ-ray counterpart to X-ray CT. In analogy to CT, images of the distribution of a γ-emitting radiopharmaceutical within the patient are obtained with several gamma camera heads that image for several minutes, rotate through a small angle about the patient (e.g. 6°), and then image again, and so on. A SPECT study is time consuming, but it produces three-dimensional images that can display physiology and anatomy with a high degree of sensitivity. Since the resolution is typically only 5–10 mm, SPECT (like PET) images are commonly overlaid on detailed anatomic maps of the patient obtained with CT.

Because the radiation source is inside the body, the task of image reconstruction is more difficult with SPECT than with X-ray CT. With transmission imaging, the position of the X-ray source and the intensity from it are known at all times during an imaging sequence, which is clearly not the case in nuclear medicine studies. In addition, CT directly measures the effect of main interest, photon attenuation; with SPECT, internal absorption and scatter of photons by the body are only a major complication to the primary task, which is the location of the origins of γ-rays. Prior to a SPECT examination, a CT-like transmission study is usually performed, using either an X-ray or a γ-ray beam, to acquire a relatively crude attenuation map to be used in correcting the SPECT data.

SPECT projection data are combined by computer to yield cross-sectional images of the distribution of radioactivity. Although CT normally employs filtered back-projection algorithms for reconstruction, SPECT fares better with statistical-based iterative approaches (Zimmerman and King, 2006).

SPECT has become nearly indispensable for the assessment of myocardial perfusion and other cardiac functions. Cross-sectional and 3D displays of thallium-201 and other radionuclides in the myocardium (or, more correctly, their *absence*) reveal damage to the heart muscle caused by diminished blood flow in the coronary arteries. More than half of SPECT studies are for coronary artery disease, another quarter are for bone, and the rest are for brain, prostate, thyroid, and other organs (Wernick and Aarsvold, 2004; Park, Moore and Kijewski, 2005). SPECT studies of the brain, for example, can shed light on conditions such as Alzheimer's dementia, epilepsy, and stroke.

27.6.4
Positron Emission Tomography (PET)

Radionuclei for which the proton-to-neutron number ratio is too high for nuclear stability may attempt to rectify the situation through the emission of a *positron*, the positively charged antiparticle of the electron. In positron decay, a proton in the nucleus transforms briefly into a neutron and a positron, in effect, and the positron escapes (along with an antineutrino). The emitted positron travels a few millimeters or less in tissue and collides with an atomic electron. The particles annihilate one another and create a pair of 511-keV *annihilation photons*[9] that fly off in almost exactly opposite directions.

In *PET*, a positron-emitting pharmaceutical is administered, and the pair of annihilation photons is detected with arrays of scintillation detectors surrounding, or on opposite sides of, the patient; in the latter case, the opposed array of detectors is slowly swung around by the gantry, as with SPECT. The arrays are operated in coincidence – that is, a positron-annihilation event is recorded only if two detectors on opposite sides of the patient are triggered virtually simultaneously. The origin of the pair of photons, hence the approximate site of positron emission, can be localized to the straight line between the two detectors. Coincidence events are recorded for all angles, allowing the creation of a 3D tomographic image of the radiopharmaceutical distribution. The operation of opposed detectors in coincidence improves both the spatial resolution and the signal-to-noise ratio relative to planar nuclear medicine and SPECT.

PET began with a prominent role in neurophysiological research. It made use of positron emitters such as oxygen-15 (^{15}O), nitrogen-13 (^{13}N), and carbon-11 (^{11}C), isotopes of elements that are in the small molecules of life. Most clinical and research work is now conducted with fluorine-18 (^{18}F), which has the benefit of a 1.8 hour half-life, long enough to transport it a good distance from the cyclotron in which it is produced.[10] Normally prepared in the form of the glucose analog fluorodeoxyglucose (^{18}FDG) and concentrating where glucose is being burned at an unusually high rate, ^{18}FDG is used widely

---

[9] The rest mass of either an electron or a positron, expressed in terms of energy, is 511 keV.

[10] The very short half lives of carbon-11, nitrogen-13, and oxygen-15 require a cyclotron on-site or nearby to produce the isotopes where they are to be deployed.

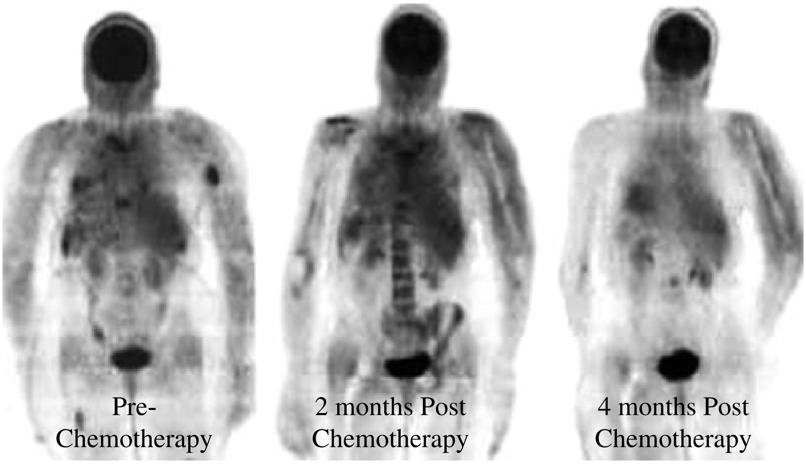

Fig. 27.19 Whole-body positron emission tomography (PET) FDG-fluorine 18 study of metastatic cancer treated by chemotherapy.

as a clinical tool for detection, localization, and staging of malignancies (Figure 27.19), and for monitoring of cancer patients after treatment (Wernick and Aarsvold, 2004; Valk *et al.*, 2003; Pevsner *et al.*, 2005; Nehmeh, Erdi and Pan, 2004).

Nearly every new PET machine now comes fused with its own dedicated CT, allowing the superpositioning of the relevant physiological information directly and precisely on a clear anatomic background. The number of clinical PET and fused PET/CT procedures in the United States has grown from about 250 000 in 2001 to more than a million in 2007. Ninety percent of the studies are performed in search of tumors; the remainder are split between cardiac and neurological studies. It is possible that, because of its better resolution, physiologically more interesting agents, and decreasing cost, PET will eventually displace SPECT and conventional nuclear medicine for many radiopharmaceutical studies.

27.6.5
Recent Developments in Nuclear Medicine

The fusion of SPECT and PET images with CT or MRI (Bacharach, 2004; Czernin, 2004) allows the projection of physiologic information on a detailed anatomic background so that tissue function can be correlated directly with tissue structure (Figure 27.19). With fusion, moreover, the CT attenuation data can be used to correct PET and SPECT images for attenuation of γ-rays within the body. In radiation treatment planning, it is proving useful to merge CT and magnetic resonance (MR) images.

Fused images can be acquired with a single, composite hybrid machine (e.g. CT/PET scanner) that acquires and integrates two separate imaging studies, one immediately after the other, with the patient in the same position. Alternatively, one can combine and register data from two independent devices with specialized software, although fusion is far more demanding in the body than in the head, and is subject to distortions due to changes in patient positioning. Common platforms for acquisition and display of fused image data from different imaging methods will undoubtedly continue to evolve as image-guided surgery and radiation therapy become increasingly desirable.

Meanwhile, advances in molecular biology are yielding receptor-specific radiopharmaceuticals and other pharmaceuticals that are opening up a new realm of opportunities for *in vivo* imaging at the molecular and cellular scale. Research in radiopharmacology and in the technologies for producing radionuclides is of growing importance, but funding of this research is becoming increasingly difficult to identify.

Production of radioactive pharmaceuticals, the creation of imaging devices and protocols, and the interpretation of nuclear images require the involvement of specialists in several disciplines, including radiopharmacology, physics, engineering, and medicine. Nuclear medicine is a true multidisciplinary specialty that represents the leading edge of high technology in medicine.

27.7
Ultrasound Imaging

Most *US* studies produce images from high-frequency sound echoes that are created at boundaries between tissues with different elastic properties or densities. With its ability to furnish images in everyone from fetuses to the elderly, and with no exposure to ionizing radiation, US is indeed a "womb-to-tomb" clinical imaging modality (Fowlkes and Wolbarst, 2006).

During WWII, echoes from high-frequency sound waves transmitted through water were used to detect and locate submarines. This technique, known as *Sound Navigation and Ranging* (*SONAR*), was developed from ideas suggested during World War I by the physicist Langevin. "Echo sounding" found additional applications during and after WWII, including the detection of flaws in welded joints and products such as steel castings. The first paper on the use of US to image human anatomy, specifically the female breast, appeared more than a half century ago (Wild and Reid, 1952). The motion- or M-mode technique for revealing the dynamics of moving structures in the heart was described 2 years later (Edler and Hertz, 1954). Since then, US has become, next to X-rays, the most widely used imaging modality. It provides clinically valuable diagnostic information, is highly adaptable, is less expensive than any other imaging technique, including projection radiography, and involves no use of ionizing radiation.

The frequencies used commonly for US imaging in medicine are between 2 and 15 MHz. This frequency range corresponds to wavelengths of a few tenths of a millimeter, which forms an upper limit to the spatial resolution achievable. Unfortunately, the absorption of US power increases rapidly with frequency, and its ability to penetrate tissue decreases, so high-resolution (short wavelength) studies are restricted to fairly superficial tissues. The 2–6 MHz region is used to examine deeper-seated organs in the abdomen and pelvis.

The velocity, v, of sound in a medium depends only on its physical density, ρ, and its adiabatic bulk elastic modulus, K,

$$v = (K/\rho)^{1/2} \tag{7a}$$

where K is, in effect, the converse of the tissue's bulk compressibility. The velocity of US in tissue, unlike its attenuation rate, depends only very weakly on its frequency, and is commonly taken to be 1540 m s^{-1} in soft tissue.

US images are produced by capturing echoes returning from reflecting surfaces in the body. Sound reflection occurs at any boundary between regions of different *acoustic impedance*, where the acoustic impedance, Z, of a tissue depends only on ρ and K:

$$Z = (\rho K)^{1/2} \tag{7b}$$

so the acoustic impedance is also seen as $Z = \rho v$.

The significance of acoustic impedance is that the *fractional reflection*, α_R, of US energy at a boundary separating two tissues with acoustic impedances Z_1 and Z_2 is

$$\alpha_R = [(Z_2 - Z_1)/(Z_2 + Z_1)]^2 \tag{7c}$$

The fractional transmission across the boundary is $(1 - \alpha_R)$. Z is much smaller for air than for tissue, for example, so US is reflected almost perfectly at the boundary of the chest wall and the lung. Consequently, little US energy is transmitted across the interface for imaging structures within the lung. Although US is not useful for examining normal lung, because of its irregular, fractal-like surface structure, it can image the heart since there is a path of solid tissue via the mediastinum. US imaging is used for imaging soft tissues in every organ system of the body other than the lung, adult brain, and bowel, where large impedance barriers impede its transmission. It is particularly helpful in identifying and characterizing cancer and cardiovascular disease.

To acquire a US image, a US transducer is placed on the skin of the patient, and air pockets are eliminated by means of a coupling gel between them. The active component in the transducer is a thin *crystal* or *piezoelectric* material that has unusual electromechanical properties: when a voltage is applied across it, it deforms and, conversely, when it is mechanically compressed or bent, a voltage appears across it. In the US system, an electronic impulse generator shocks the crystal several hundred times per second. Each time, the crystal contorts and oscillates at high frequency (MHz), thereby generating a short burst of US waves that propagate away from the crystal and into the tissue. Echoes from impedance discontinuities at tissue interfaces in the patient return to the crystal in the periods between shocks (the quiescent period). They apply pressure to the crystal, which responds by yielding a small electrical signal that is captured, amplified, processed, and displayed. An echo that returns at time t_{echo} was produced by a surface at depth

$$x_{\text{surface}} = 1/2 v t_{\text{echo}} \tag{8}$$

The earliest form of US, seldom used today, was *amplitude-mode* (*A-mode*), in which the US beam points in one, fixed direction through tissue (Figure 27.20). The display might involve sweeping a faint spot of light horizontally across a monitor in such a manner that its x position corresponds to depth within the body. For an echo created at depth x_{surface}, as determined by Eq. (9), the brightness of the spot will increase briefly with

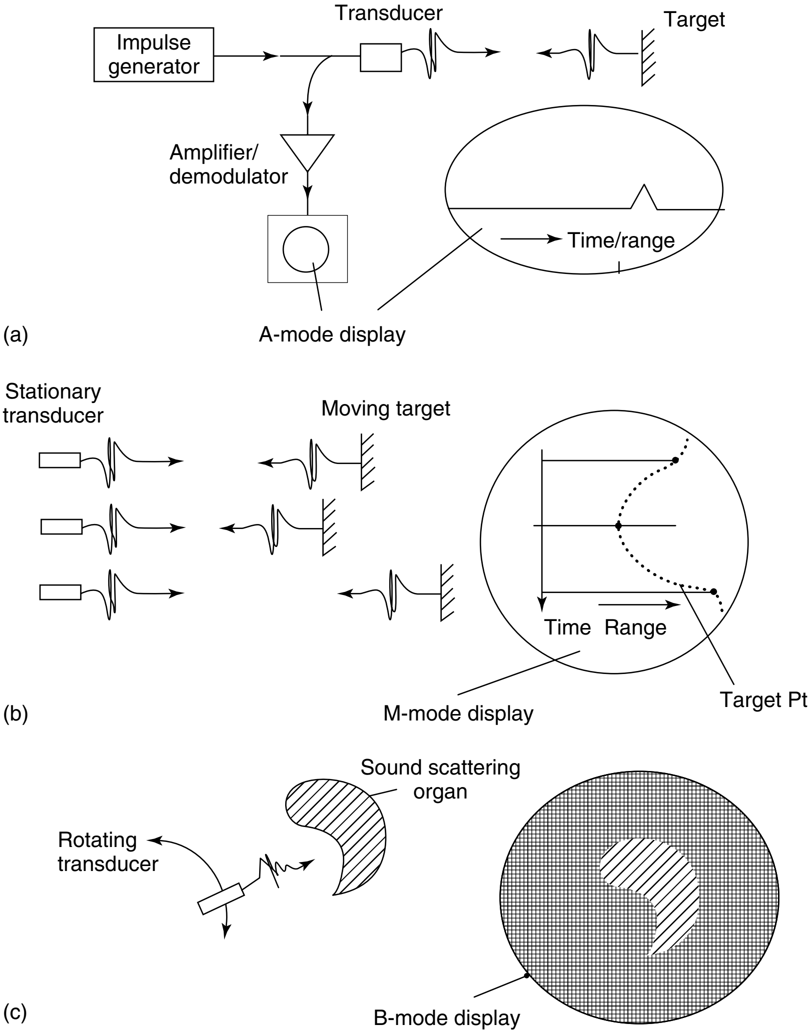
Impulse generator
Transducer
Target
Amplifier/ demodulator
Time/range
(a)
A-mode display
Stationary transducer
Moving target
Time
Range
Target Pt
(b)
M-mode display
Sound scattering organ
Rotating transducer
(c)
B-mode display

Fig. 27.20 Three basic ultrasonic imaging modes. (a) A-mode display, consisting of echoes from tissue boundaries encountered by a single ultrasound ray. (b) M-mode display, showing the motion, with time, of tissue boundaries along a single ultrasound ray. (c) B-mode display, showing an outline of tissue boundaries produced by scanning an ultrasound beam in a plane. (From Newhouse, V.L. (1993) Imaging techniques, biomedical, in *Encyclopedia of Applied Physics*, Vol. 7, VCH, New York, with permission.)

an intensity proportional to the echo signal voltage, that is, the Z mismatch at the reflecting surface. A *time gain compensation* (*TGC*) circuit will adjust the brightness to account for natural attenuation by the tissues.

The most common method of displaying US information is known as

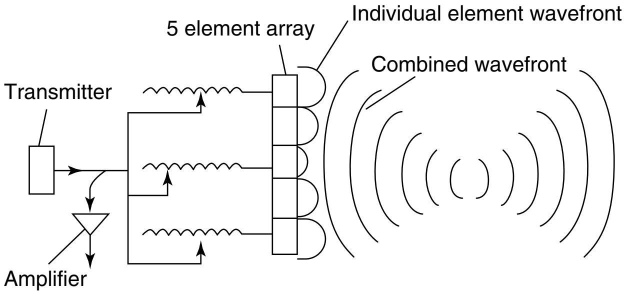

Fig. 27.21 Principle of phased-array operation, showing how a suitably phased five-element array launches a focused wave. (From Newhouse, V.L. (1993) Imaging techniques, biomedical, in *Encyclopedia of Applied Physics*, Vol. 7, VCH, New York, with permission.)

brightness mode (*B-mode*). The transducer transmits and receives US along a line through tissue, as in A-mode, and produces a line of varying brightness on the monitor depending on the strength of the returning echo. But now the US beam is swept back and forth through the tissue, perhaps 30 times per second, and the line on the display follows accordingly creating a B-mode image of the slice (Figure 27.20).

Echoes are produced not only at interfaces between organs in the body, but also from interference patterns resulting from countless reflections by microscopic structures within tissues. These echoes, often referred to as *speckle*, can be important for purposes of tissue identification and characterization.

27.7.1
Phased-Array Transducers

To acquire a B-mode image in real time, the US beam sweeps quickly through the tissue, and varies its depth of focus, as well. This is achieved with a phased array, which contains hundreds of tiny, separate, mechanically and electrically isolated piezoelectric elements (Figure 27.21) (Zagzebski, 1996). Each element is driven by an independent electronic delay circuit configured to vary the time of arrival of its own, unique voltage pulse.

The delay times depicted in Figure 27.21 cause the impulse from the transmitter to reach the top and bottom elements before it reaches those in the center of the array; US waves emitted from the top and bottom elements enter the tissue ahead of those from the center, and this timing pattern yields a curved ultrasonic wave front that converges to a focus at a particular depth in tissue. Changing the relative delays of the transducers alters the depth of focus of the US beam. A multielement array also allows us to steer the beam, as well. Increasing the delays of the top transducers relative to the bottom transducers, for example, deflects the US beam downwards. With the array transducer steering the beam back and forth while increasing the depth of focus more slowly, B-mode US can generate images that are of high resolution at all depths.

A phased array can function as a focusing mechanism not only while transmitting the US beam but also while receiving the echoes. The time it takes for the US energy created by any individual piezoelectric element in the array to reflect from a point on the tissue interface and then return to any other element depends on the relative positions of the two elements and the reflection point. The feature of *dynamic focus* corrects for this complication, which otherwise would cause significant image blur. Once a US burst has been transmitted out by the array, the delay settings can be continuously changed so that the *receive focus* is always optimized for echoes returning from specific depths. Images of Figure 27.22 were obtained with a variable depth-of-focus phased-array US transducer.

Figure 27.22 reveals an exceptionally high level of detail because the fetus is surrounded by uniform amniotic fluid, so there are no aberrations introduced by slight US velocity differences in inhomogeneous tissues in its immediate vicinity.

27.7.2
Ultrasound Flow (Doppler) Imaging

US can measure the speed of flow of blood in vessels and map flow patterns in the heart by means of the Doppler effect. The frequency of an echo reflected from a moving target is altered by the Doppler shift frequency, Δf_D, where

$$\Delta f_D / f_{\text{transmit}} = \pm 2\, v_{\text{blood}} / v \qquad (9)$$

where f_{transmit} is the frequency of the transmitted US, v_{blood} is the speed of the moving target (blood in a vessel), and v is the velocity of US in the tissue containing the blood vessel. By detecting the Doppler frequency shift, the velocity v_{blood} of the flowing blood can be determined, and by observing the polarity of the frequency shift, the direction of blood flow can be identified.

Through adoption of electronic processing techniques first used in weather radar, Doppler and B-mode have been combined to reveal the flow patterns of blood through a region (e.g. a heart chamber) in real time against a background of clearly delineated anatomy (Kasai *et al.*, 1985).

27.7.3
Ultrasound Tissue Characterization

Clinical US is used principally to obtain anatomical and blood-flow information, and it is discontinuous changes in elastic properties or the density at tissue interfaces that make this possible. The images generated, however, reveal little about tissue elasticity itself, which may nonetheless be of clinical interest. US can also discriminate between normal and diseased tissues through measurements of the frequency dependence of tissue attenuation and scattering (Greenleaf, 1986).

The hardness of tissue has been used for centuries as a method to detect abnormal conditions, usually through subjective measures such as finger pressure. A more objective approach is the use of a mechanical device to apply a known amplitude of vibration to the skin, and determination of the displacement of tissues at various depths by Doppler measurements. The resulting computations yield a measure of tissue rigidity (Krouskop, Dougherty and Vinson, 1987). More recently, *elastography*, or elastic imaging, allows direct measurement of tissue elasticity parameters.

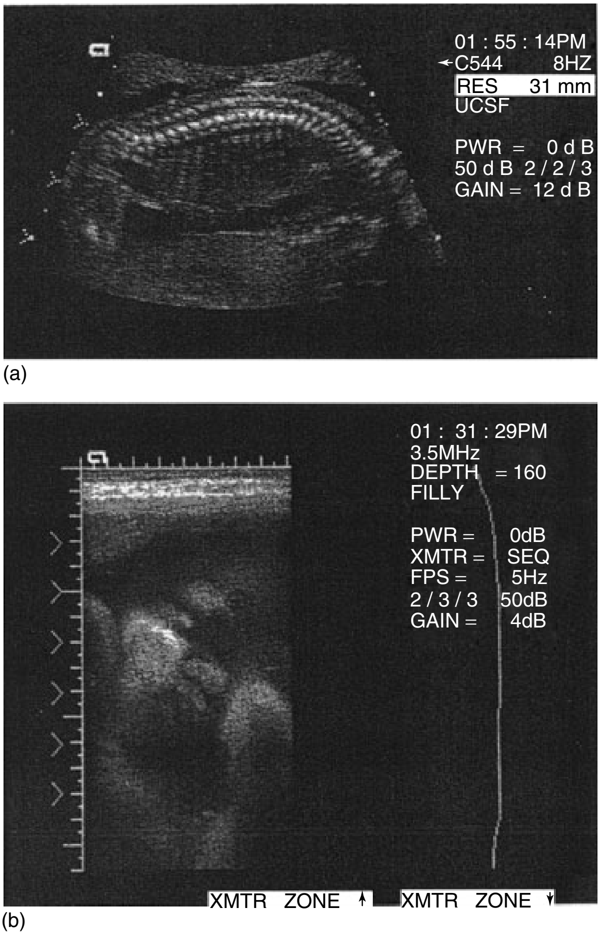
01 : 55 : 14PM
C544 8HZ
RES 31 mm
UCSF
PWR = 0 d B
50 d B 2 / 2 / 3
GAIN = 12 d B
(a)
01 : 31 : 29PM
3.5MHz
DEPTH = 160
FILLY
PWR = 0dB
XMTR = SEQ
FPS = 5Hz
2 / 3 / 3 50dB
GAIN = 4dB
XMTR ZONE
XMTR ZONE
(b)

Fig. 27.22 Ultrasound B-mode image of fetus produced with dynamic focusing. (a) View of 18-week-old fetal spine showing fine detail. (b) Soft tissue detail of a 37-week-old fetal face. (From Newhouse, V.L. (1993) Imaging techniques, biomedical, in *Encyclopedia of Applied Physics*, Vol. 7, VCH, New York, with permission.)

US images are compared before and after mechanical or acoustical application of static or dynamic pressure or shear, revealing the small amounts of compression or strain that are induced (Sandrin *et al.*, 2002). With the related technique of MR elastography, the tissues of interest are stressed with small (10–100 μm amplitude) vibrations, and the induced distortions are measured with a phase-contrast pulse sequence (Weaver *et al.*, 2001).

27.7.4
Recent Developments in Ultrasound

Over recent years, the evolution of diagnostic US has been quiet spectacular

(Harvey *et al.*, 2002; Wang *et al.*, 2003; Lee, 2003). There have been major strides in the design of miniaturized piezoelectric and capacitive transducers; contrast agents (which, at present, consist of fluid suspensions of erythrocyte-sized gas bubbles that may be coated with a tissue-targeting compound) (Hohmann *et al.*, 2003; Miller and Nanda, 2004); the exploitation of harmonic information, which is associated with higher frequency components created from nonlinear propagation of sound through tissues or with nonlinear oscillations of contrast agent gas bubbles (Desser and Jeffrey, 2001); the use of coded-excitation pulse trains and matched filters; image-processing software; techniques based on the absorption or speed of US radiation, rather than just on reflection; and small size and portability. In addition, two-dimensional arrays of piezoelectric elements make possible the creation of real-time images that appear truly three dimensional and that display high spatial, contrast, and temporal resolution over a large field of view (Smith *et al.*, 2004; Duric, Littrup and Babkin, 2005).

Clinical US systems normally operate in the 2–15 MHz range. Despite technical difficulties, devices are now being designed with much higher frequencies. The shorter wavelengths allow the imaging of correspondingly smaller objects. Micromachined US transducers consisting at present of tens of separate piezoelectric or capacitive elements a few micrometers across – fine enough to fit within a narrow-gauge catheter and then into a coronary artery – are being built by microelectromechanical semiconductor manufacturing technologies (Oralkan *et al.*, 2003). Vascular imaging is currently being performed in the 10–40 MHz range; at even higher frequencies (up to several 100 MHz), there are potential applications in ophthalmology, dermatology, and perhaps cellular-level imaging.

There are also early but promising results for true four-dimensional tomographic US imaging. US tomography may require hundreds of transducer elements covering a large-angle area of the body and perhaps firing and gathering echoes separately – although it is possible that a much simpler solution to four-dimensional image acquisition may be found. Transmission US tomography has been demonstrated in relatively homogeneous tissues such as the female breast, where (as in refraction seismography, used by geophysicists in the search for oil and gas) variations in the speed and/or absorption of sound are measured and processed to create images. Reflection US tomography, with its multiplicity of omni-directional echoes, continues to pose major technical problems that remain intractable at this time.

Finally, there is active research on biological effects (AIUM, 2003), especially as US applications move to much higher frequencies. This is ongoing not only to ensure patient safety in diagnosis but also to support novel applications in therapy (Curra and Crum, 2003).

27.8
Magnetic Resonance Imaging

X-ray and US imaging are termed *transmission* technologies because they involve the transmission of penetrating radiation into a region of the body from an external source. Nuclear medicine is an *emission* imaging technique because it utilizes radiation emerging from a biological compartment inside the body following administration of a radiopharmaceutical. *MRI* is a bit of both. It involves first the transmission of pulses of RF EM radiation into the body, which is situated in a strong external magnetic field, and then the detection of subsequent RF signals generated by magnetic nuclei within.

MRI not only reveals anatomic details of the various organs and tissues, as does CT, but it also provides information on their physiological status as does nuclear medicine.[11] But unlike CT and nuclear medicine, MRI does not involve X-ray or γ-ray energy, and it is frequently hailed for using no ionizing radiation. Instead, MRI employs magnetic fields and radio waves to probe the exquisitely choreographed motions of the magnetic nuclei of hydrogen atoms (almost exclusively) occurring naturally in the water (primarily) and lipid molecules within and around tissue cells.

Nuclei act rather like a spinning positively charged ball. As with any other moving charged body, they produce their own local magnetic field. An atomic nucleus therefore behaves in some ways like a tiny compass needle – not only producing its own very weak magnetic field but also aligning in an externally applied magnetic field.[12]

When a patient lies in the strong externally applied magnetic field of an MRI device, the hydrogen nuclei (which are single protons) of the water and lipid molecules in the tissues tend to align along it. Many protons settle into a lower-energy (ground) state, like a compass needle, with its spin axis and local field aligned along that of the external field. But some protons reside in a quasi-stable, higher-energy spin state, pointing in the opposite direction to the applied magnetic field.

Under conditions of thermal equilibrium, a slight excess of spin exists in the lower-energy spin state. For the magnetic field strength most commonly encountered in MRI, 1.5 T, this excess (the difference in numbers of nuclei pointing parallel and antiparallel to the external magnetic field) amounts to a few parts per million, as described by Boltzmann statistics. The spin excess in the jth tissue voxel gives rise to a relatively weak magnetic field, known as the voxel's *equilibrium net magnetization,* and is denoted by the vector $M_{j,z0}$:

$$M_{j,z0} = 10^{-6} N_j \mu \tag{10}$$

The subscript 0 indicates thermal equilibrium, and the z is a reminder that here, the net magnetization of the voxel points along the z axis defined by the external magnetic field. μ, the *nuclear moment* of the nucleus of interest, is a direct measure of the strength of the magnetic field it produces; there are N_j protons in the jth voxel; and there is an excess of about 10^{-6} nuclei (the exact value increases with the strength of the external magnetic field) in the lower-energy spin state.

When not in thermal equilibrium, which, during MRI, is almost always, the behavior of the voxel's net magnetization over time is written as $M_j(t)$. It is by monitoring the behavior of the $M_j(t)$ for

---

[11] Expanded treatments of MRI may be found in Bushong (1996); Bushberg *et al.* (2002); Hendee and Ritenour (2002); Wolbarst (2005); Weaver, Ballon and Wolbarst (2006).

[12] The only nuclei that do not act in this manner are the few that happen to contain an even number of protons (which pair up in opposite directions, so that their fields cancel) and an even number of neutrons. For better or worse, the common isotopes of carbon, oxygen, and calcium fall into this class; hence they contribute nothing to an MRI signal.

Tab. 27.1 Magnetic resonance properties of selected nuclei.

Nucleus	$\gamma/2\pi$ (MHz/T)	Natural abundance of isotope (%)	Mass in 1 g soft tissue (mg)	Relative sensitivity
^{1}H	42.58	99.98	–	100
^{13}C	10.71	1.11	–	1.59
^{19}F	40.05	100.00	–	83.30
^{23}Na	11.26	100.00	–	9.25
^{31}P	17.23	100.00	–	6.63
^{39}K	1.99	93.10	–	0.05

all the voxels in a section of tissue under various conditions that MRI images are created.

27.8.1
Nuclear Magnetic Resonance (NMR)

Nuclei can be intentionally flipped from their lower- into their higher-energy spin states through the process of *nuclear magnetic resonance* (*NMR*). This is achieved by irradiating the nuclei with a pulse of RF energy of the correct frequency. This *Larmor frequency* is easily deduced for the elementary, two-state quantum picture. The energy to flip a single nuclear spin depends on the strength of the external magnetic field, B_0, and on the "magnetness" of the particular isotope, as characterized by μ or by the closely related *gyromagnetic ratio*, γ:

$$E_{\text{flip}} = \gamma(h/2\pi)B_0 \tag{11a}$$

The presence of Planck's constant, h, suggests that the spin-orientation transition is a quantum mechanical event. Meanwhile, one of the fundamental pillars of quantum physics is the Einstein relation,

$$E_{\text{photon}} = hf_{\text{photon}} \tag{11b}$$

which states that the energy of a photon is proportional to its frequency. So for a photon to cause the spin transition of a nucleus in a magnetic field, it must be of the frequency[13]

$$f_{\text{Larmor}} = \gamma/2\pi\, B_0 \tag{11c}$$

known as the *Larmor frequency* (Table 27.1). For protons, $\gamma/2\pi = 42.58$. Therefore,

$$f_{\text{Larmor}}\ (\text{MHz}) = 42.58\ B_0 \quad (\text{T}) \tag{11d}$$

the central relationship of both NMR and MRI. In a 1.5-T field, water and tissue protons resonate at 63.87 MHz, and in a 3.0 T field, it is twice that frequency.

Quite remarkably, it is possible to arrive at exactly the same Larmor equation by a purely classical argument. Consider a top or gyroscope in outer space. Its rapid rotation and the Law of Conservation of Angular Momentum act to keep its axis pointing always in the same direction. But on the surface of the earth, the pivot point holding the top or gyroscope up and gravity acting at its center of mass together exert a twisting torque on it. Its equation of motion, a simple extension of Newton's law $F = ma$

---

[13] The Larmor equation is sometimes written in terms of angular frequency, in radians per seconds, as $\omega_{\text{Larmor}} = \gamma B_0$.

for a rotating object experiencing an applied torque, reveals that it will not simply topple over, but rather will *precess* around the direction of the applied field, with its axis sweeping around the surface of a cone. Similarly, a charged object, spinning with angular momentum and sitting in a magnetic field, is subject to the Lorentz force, and its resulting motion is a steady precession about the direction of the field, with the vector $M_j(t)$ circling it and constrained to the surface of a cone. And it will be precessing at the same Larmor frequency as was seen with the quantum picture!

An individual magnetization vector, $M_j(t)$, precesses in every one of the (typically) $256 \times 256 = 65$ k voxels in a thin slice.

By determining the magnitude of each vector, it should be possible, from the corresponding values of N_j of Eq. (11), to create a *proton density-weighted* (*PD-w*) MRI image that maps out the variation in numbers of protons in each voxel. The problem, however, is that the only thing we can measure in a pickup coil outside the body is the single RF signal produced by all the voxel magnetizations acting together. This problem, while perhaps seeming hopeless at first viewing, in fact has a clever and powerful solution that involves the application of various combinations of RF pulses and magnetic field gradients, as will be discussed below.

27.8.2
T1 (Longitudinal, Spin–Lattice) Proton Spin Relaxation

A PD-w image is conceptually straightforward, and of some clinical value, but of much greater medical importance are images that reflect spatial variations in the *spin relaxation times*, T1 and T2. Their meanings may not yet be apparent, but we can address T1 relaxation readily with our simple spin-up, spin-down picture. For simplicity, moreover, we shall focus hereafter on the protons of tissue water.

Imagine a tray covered with compasses with needles aligned along the Earth's magnetic field. Now, in your mind's eye, twist each needle through 180°, so that it points south, and then release them all at the same time. They will flip back over, oscillate a few times with diminishing swings, and eventually come to rest pointing north again. The average amount of time this settling down process takes for an ensemble of identical compasses is known as its *relaxation time*. It is determined both by the strength of the magnetic field interaction and by the nature of the frictional forces present, such as those occurring at the mechanical pivot points where the needles are supported.

A similar situation applies with protons. Suppose a voxel full of tissue protons is sitting between the poles of a turned-off magnet. They will be oriented randomly in all directions. Now quickly snap the field on. The protons will undergo a kind of spontaneous relaxation analogous to (but actually quite different from) that of the tray of compass needles, in which they flip into alignment along or against (because they obey quantum mechanics) the field. The time it takes them to go 63% of the way to their thermal equilibrium[14] condition is called the *longitudinal spin* (or *spin-lattice*) *relaxation time*, and designated T1:

---

[14] The time, dose, distance, energy, whatever, it takes for a system exhibiting (near-) exponential behavior to decay, be attenuated,

$$M_{j,z}(t) = M_{j,z0}(1 - \mathrm{e}^{-t/\mathrm{T1}}) \qquad (12)$$

T1 and a second related, but different, relaxation time, T2, are invaluable in creating clinical images, and it is primarily the spatial variations in the relaxation times T1 or T2 that MRI reveals.

The atomic-level mechanism that is responsible for T1-type relaxation also involves the Larmor equation, but the physical interpretation is different. For reasons explained by quantum mechanics, not only can Larmor-frequency RF photons kick protons in the ground spin state up into the excited state, with the absorption of $\gamma(h/2\pi)B_0$ of energy, but also *any* local magnetic field (not just that of a photon) that happens to be fluctuating at the Larmor frequency can "tickle" protons in the higher-energy state back down, where the proton loses its spin-orientation energy to the local environment. Magnetic "noise" occurs naturally in tissues if the water molecules are tumbling and interacting with one another and with their surroundings in such a manner that, at least intermittently, they are rotating at the Larmor frequency. In this condition, each proton experiences the magnetic field from its partner (on the same water molecule) as varying, again at least briefly, at the Larmor frequency, which can cause a downward spin transition. Unbound ("free") water, for example, tumbles much faster than the proton Larmor frequencies for the external fields employed in MRI, so T1 is long (4000 ms at 1.5 T), as is the case with other nearly clear fluids such as cerebrospinal fluid (T1 = 2.5 seconds). For water bound to intermediate-sized biologic molecules (which themselves may happen to be tumbling at the Larmor frequency), however, T1 can be shorter by an order of magnitude or more.

The proton relaxation times of the water in and between tissue cells are highly sensitive to the detailed nature of binding, and of friction-like mechanics and other interactions, with various biomolecules and membranes. The concentrations and biophysical characteristics of these biomolecules, in turn, depend on the type and physiologic status of the cells they inhabit. MRI is therefore able to distinguish not only among histologically different tissues, but even between healthy and diseased forms of the same tissue.

What a T1-weighted (T1-w) MR image actually reveals is variation in T1 throughout the tissue slice under examination. It can produce a degree of contrast among soft tissues that is often much better than that of CT, and it can reveal subtleties in the physiology of an organ that any form of X-ray imaging would completely miss. For some kinds of examinations, a T1-w may well be fully diagnostic – while for others, a T2-weighted image is even better.

27.8.3
T2 (Transverse, Spin–Spin) Proton Spin Relaxation

It is easiest to discuss T2 with the classical picture of NMR.

Under conditions of an unchanging external magnetic field, the protons in

---

be excited, relax, and so on, to $0.37 = e^{-1}$, or to $0.63 = 1 - e^{-1}$, of the original or final value is commonly taken as a characteristic "constant" parameter.

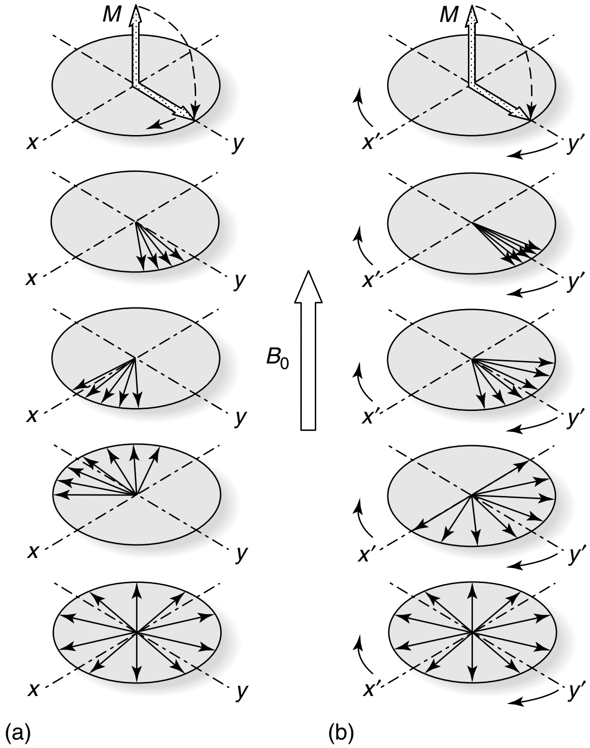

Fig. 27.23 Following a 90° pulse that drives the magnetization for a voxel, *M*, from the vertical down into the transverse plane, spin-spin interactions cause a dephasing of separate packets of protons within the voxel. This is seen (a) from the fixed frame of reference and (b) from a frame of reference rotating with *M*, at the Larmor frequency.

a voxel will be oriented in a direction close to that of the external field, yielding a magnetization $M_j(t)$ that precesses on the surface of an imaginary cone at the Larmor frequency (Figure 27.23a). If, in addition, an RF pulse with a magnetic field $B_1\cos(2\pi f_{\text{Larmor}}t)$ is applied at right angles to B_0, then the magnetization will *nutate* (rotate) away from its near-equilibrium alignment and spiral downward (Figure 27.23a) – as if the cone were slowly opening up (Figure 27.23b). This nutation, like precession, is readily explicable with a further extension of Newton's law of classical physics. For a Larmor-frequency γB_1 left on for time τ, the magnetization will spiral down through an angle of

$$\theta = \gamma B_1 \tau \qquad (13)$$

Of particular interest is the case in which $\gamma B_1 \times \tau$ yields $\theta = 90^\circ = \pi/2$, so that the magnetization of the *j*th ends up precessing in the transverse $(x - y)$ plane, and will be called $M_{j,x-y}(t)$.

With a receiver coil also located on the transverse plane, and ignoring the other voxels for the moment, $M_{j,x-y}(t)$ will induce a voltage in the coil every time the magnetization vector swings by. This process generates a Larmor-frequency signal whose voltage is proportional to the magnitude of $M_{j,x-y}(t)$. Precession of a magnetization in the transverse plane and its detection by a pickup coil, by the way, is the *only* way in which MRI data are ever obtained.

The signal produced following a 90° pulse decays exponentially with a time constant T2*:

$$M_{j,x-y}(t) = M_{j,x-y}(0)e^{-t/\mathrm{T2}^*} \qquad (14)$$

Such *free induction decay* (*FID*) corresponds to a *dephasing* of the spins over time caused primarily by two factors – random and nonrandom disturbances. The first includes the dephasing that accumulates because of slow, random magnetic interactions between a proton and its partner proton or some other spins, momentarily affecting its rate of precession in the $x - y$ plane. Another kind of random event occurs when a proton happens to lose its spin energy to its surroundings, and snaps back toward the z axis – the phenomenon of longitudinal relaxation.[15] Nonrandom, or quasi-constant disturbances arise from weak, nonchanging spatial inhomogeneities in the magnetic fields experienced even on the scale of a voxel; these are caused both by imperfections in the external magnetic field and by the presence of nearby, immobile biomolecules (the *susceptibility effect*) that have been made slightly magnetic by the external field. The random and nonrandom mechanisms act independently on the spins in any voxel, together giving rise to their dephasing over time, and causing the decline in $M_{j,x-y}(t)$ with time constant T2*.

Application of a single 90° FID pulse can provide a simple assessment of T2*, but what is usually desired are estimates of the biologically relevant *transverse* or *spin–spin relaxation time*, T2. There is nothing that can be done about the random dephasing (even if we wanted to do it, which we do not), but application of the *spin-echo* pulse sequence can actually *reverse* the dephasing effects due to the static, nonrandom magnetic field inhomogeneities. This leaves behind, in effect, only that part of the decline in $M_{j,x-y}(t)$ that is attributable to relaxation events arising from spin–spin interactions.

The decline in $M_{j,x-y}(t)$ may be understood more clearly by considering the *rates* of the several relaxation mechanisms, rather than their relaxation *times*. The overall decay of the transverse magnetization occurs at a rate proportional to 1/T2*. Contributing to it are the rates $1/\mathrm{T2}_{\mathrm{inhomo}}$ and 1/T2 corresponding, respectively, to speed-up and slow-down in precession caused by static field inhomogeneities, and to dephasing generated by random interactions among spins, as determined by the local biophysical environment:

---

[15] How can a proton precessing in the $x - y$ plane undergo a spin-up, spin-down transition? An apparent paradox arises here only because the quantum picture presented so far for a single proton is a gross over simplification for most purposes. A complete, rigorous quantum mechanical treatment reveals that the magnetization made up by the ensemble of all the spins in a voxel should be viewed as an *expectation value* that obeys the classical *Bloch equations.*

$$1/\mathrm{T2}^* = 1/\mathrm{T2} + 1/\mathrm{T2}_{\mathrm{inhomo}} \qquad (15)$$

By removing $1/\mathrm{T2}_{\mathrm{inhomo}}$ dephasing, we are left with $\mathrm{T2}^* = \mathrm{T2}$.

Now back to spin-echo, and let us ignore random, T2-type relaxation for the moment, concentrating only on the dephasing that comes from magnetic field inhomogeneities. Spin-echo begins with a 90° pulse that rotates a voxel's magnetization vector from its equilibrium configuration, $M_{j,z0}$, lying along the z axis, down into the $x - y$ plane, with virtually all the spins starting out precessing in phase (Figure 27.24a). Inhomogeneities in the external magnetic field cause some spins to precess faster than average and others slower, leading to their dephasing (Figure 27.24b). A 180° pulse, applied after the time 1/2TE[16] has passed, flips the spins through 180° within the $x - y$ plane (Figure 27.24c). This flip puts the spins that are precessing more slowly ahead of the faster ones, and vice versa (Fig. 27.24d). But because the field inhomogeneities are constant, each spin will continue to precess at the same rate

---

[16] The time TE is referred to as *the time to echo.*

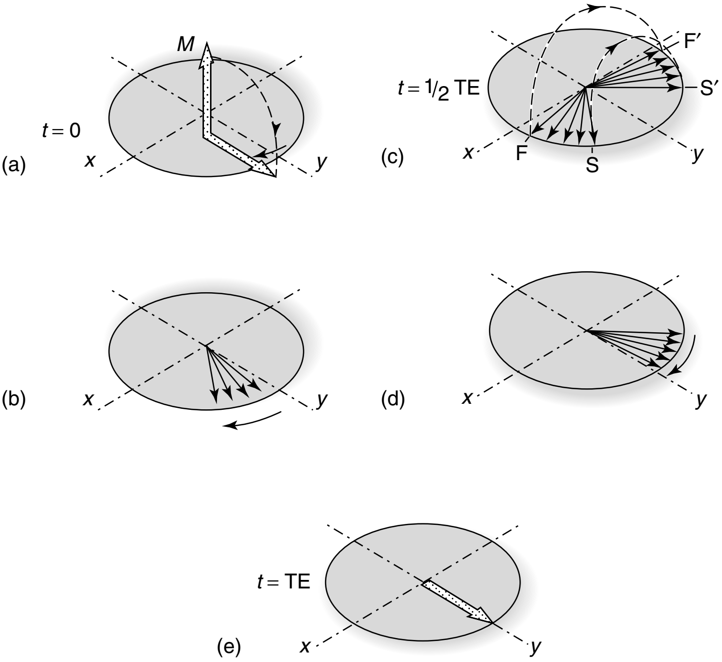

Fig. 27.24 In the spin-echo pulse sequence the 90° pulse (a) causes nutation of the bulk magnetization about the y axis. Dephasing then occurs (b) as slower components of the magnetization lag and faster components advance. A 180° rephasing pulse (c) causes nutation about the y axis and reverses the order of fast and slow components. The components then merge (d) to produce the maximum signal (e). Diagrams a–d are shown in the rotating frame of reference.

as before. So exactly at time TE, the faster spins will have caught up with the slower ones, and all will, very briefly, be back in phase again (Fig. 27.24e). As they precess at f_{Larmor} at time TE, their communal $M_{j,x-y}(\mathrm{TE})$ sweeps past the pickup coil and creates a brief but detectable ''echo'' voltage. This, in fact, is what makes signal acquisition possible – indeed, approaches other than spin-echo, such as *gradient-echo,* must find another way to create some sort of echo in order to record a signal.

In the spin-echo method, all the phase change arising during the first 1/2TE period caused by static field inhomogeneities is undone in the second 1/2TE and, in effect, these inhomogeneities are removed entirely from the picture. Now let us add T2 relaxation back in. Exactly at TE, the signal detected by the receiver coil will reflect the instantaneous value of $M_{j,x-y}(t)$ as determined by T2- (not T2*-) relaxation processes:

$$M_{j,x-y}(\mathrm{TE}) = M_{j,x-y}(0)e^{-\mathrm{TE}/\mathrm{T2}} \tag{16}$$

A sequence of measurements at TE intervals will reveal an exponential decline in $M_{j,x-y}(t)$, but one parameterized by T2, not T2*.

T2, like T1, depends on the biochemical microenvironments of resonant protons, and differences in these microenvironments among tissues strongly influence MRI image contrast. It is left as an exercise to ponder

$$\mathrm{T1} \geq \mathrm{T2} \geq \mathrm{T2}^* \tag{17a}$$

or in its perhaps more physically intuitive form

$$1/\mathrm{T1} \leq 1/\mathrm{T2} \leq 1/\mathrm{T2}^* \tag{17b}$$

and why paramagnetic *contrast agents* (most notable, gadolinium ions caged within the molecules of a chelating agent) can radically alter T1 and T2 (Allen and Meade, 2004).

27.8.4
Spin-Echo, Spin-Warp Voxel-Magnetization Localization

MRI involves manipulation of pulses of RF energy and magnetic field gradients to assess, in effect, T1 and T2 at each point in the part of the body being examined (Haake *et al.*, 1999; Liang and Lauterbur, 2000).

The signal induced in a pickup coil during resonance contains contributions from the magnetizations of all the voxels in the tissue subjected to the external magnetic field and the RF pulses. To produce an image from MR signals, the origin of each voxel-contribution must be localized to a tiny volume of tissue. With the standard *spin-echo spin-warp* approach, the voxel locations are determined by three separate processes that localize the signal origin along three orthogonal axes.

The first step is to select the transverse slice of tissue at some particular position along the z axis (the direction of the external magnetic field, B_0) for further investigation. This involves application of a *slice-select gradient* magnetic field, which points always along B_0, but whose strength varies linearly with z. This z gradient alters not only the applied magnetic field along the z axis but also the proton Larmor frequency. Thus, there is a direct and unique correspondence among z position, local field strength, and proton Larmor frequency for a slice. So with application of a narrowband 90° RF pulse, only those protons in the selected narrow (typically a few millimeter) transverse slice of tissues at the corresponding

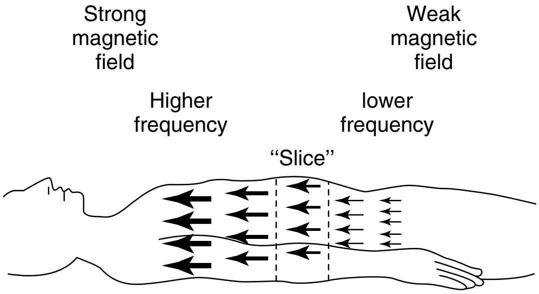

Fig. 27.25 A craniocaudal "gradient" or change in the magnetic field along the craniocaudal axis allows receiver coils to be "tuned in" to a slice.

z value will resonate and nutate down into the $x-y$ plane (Figure 27.25). Nuclei above and below the slice will be unaffected, and will not contribute to the MRI signal.

A second, *phase-encoding gradient* is now applied, also pointing along B_0 but with its field strength proportional to the y position. This gradient is applied for a brief interval before signal readout, and then removed. While it is on, the gradient alters the precessional frequencies of the nuclear spins in such a manner that those with higher y values precess faster than spins on the low field-strength end – and the amount of additional phase difference accumulated by a nuclear spin during the gradient-on interval depends on the position of the spin in the y direction. When the gradient is removed, these phase differences are locked in, and their determination during readout serves as a way to identify the location of voxels in the y direction. The amount of phase difference accumulated in a given voxel depends on the gradient amplitude, the y position, and the gradient on-time.

Locating the origin of an MRI signal component and its voxel along the x axis involves application of the *frequency-encoding* or *readout* gradient during signal readout, which spreads out the voxel-specific resonant frequencies according to their x positions. For an image matrix with 512 pixels along the patient's x axis, say, the x gradient must be strong enough for the RF receiver and frequency discriminator to be able to separate the acquired signal into components centered at 512 distinct frequencies.

A single reading, with one setting for the y gradient and one for the x gradient, does not provide anywhere near enough information to individually localize all of the voxels in the tissue slice selected along the z axis. So while the x-gradient amplitude is always set to the same value, the amplitude $G\gamma$ of the y gradient is systematically stepped upward in strength, with a readout (with the x gradient on) for each setting of the y-gradient amplitude. For a 512 × 512 matrix, for example, 512 readings are made, each with a different value for the y gradient or its on-time, and for every one of them, the signal is then separated into 512 frequency components.

Acquisition of phase- and frequency-encoded MRI signals over many cycles yields a complex set of data that is best analyzed by two-dimensional Fourier transform (2DFT) methods. This analysis constitutes the fourth and final step in the process of spatial localization of MRI signals.

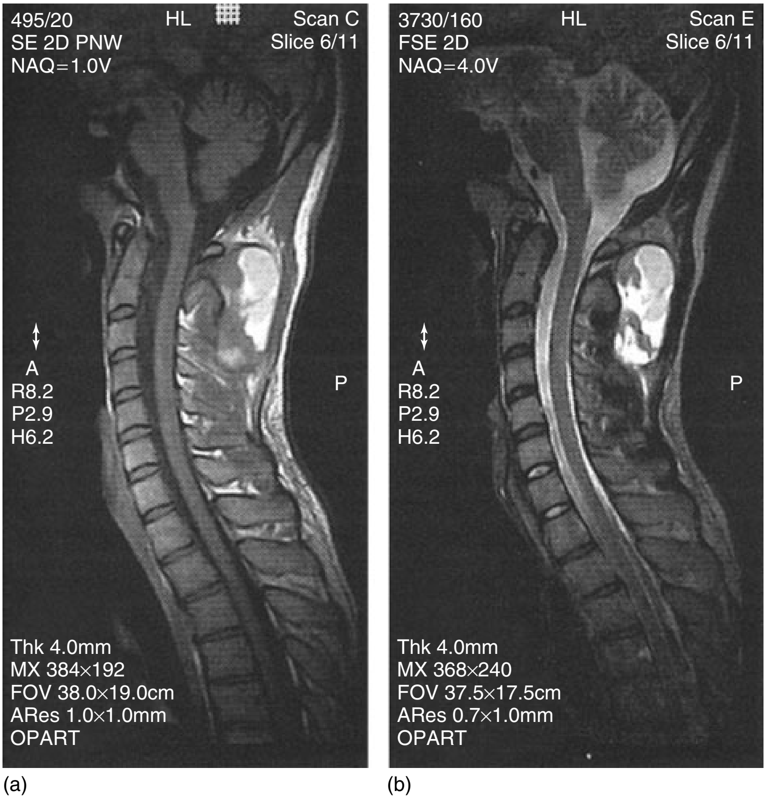
495/20
SE 2D PNW
NAQ=1.0V
HL
Scan C
Slice 6/11
A
R8.2
P2.9
H6.2
P
Thk 4.0mm
MX 384×192
FOV 38.0×19.0cm
ARes 1.0×1.0mm
OPART
(a)
3730/160
FSE 2D
NAQ=4.0V
HL
Scan E
Slice 6/11
A
R8.2
P2.9
H6.2
P
Thk 4.0mm
MX 368×240
FOV 37.5×17.5cm
ARes 0.7×1.0mm
OPART
(b)

Fig. 27.26 (a) Spin-echo T1-weighted sagittal cervical spine and (b) T2-weighted image fast spin-echo of the same patient. (From Woodward, P. (2001) MRI for Technologists, 2nd edn, McGraw-Hill, New York, with permission.)

The entire image creation process may be repeated several times, with the resulting images averaged to reduce image noise. Representative MRI images are shown in Figure 27.26.

27.8.5
Flow Imaging: Magnetic Resonance Angiography (MRA)

The acquisition of data for an MR image takes time, during which blood can move a considerable distance within a tissue slice. *Magnetic resonance angiography* (*MRA*), exploits this movement to produce maps of blood flow (Higgins *et al.*, 2005; Elkington, Gatehouse and Cannell, 2005; Slavin and Bluemke, 2005; Bulte, Kostura and Mackay, 2005).

Blood moving into a z-axis slice during signal acquisition will display a brightness different from that of stationary tissues present in the slice, primarily because the blood was outside the slice when the RF pulse was applied to excite the nuclear spins; blood moving within the slice will tend to "disappear" from the image, because the phase-encoding method relies on the tissues remaining stationary over multiple cycles of the signal acquisition process. Also, blood present in the tissue

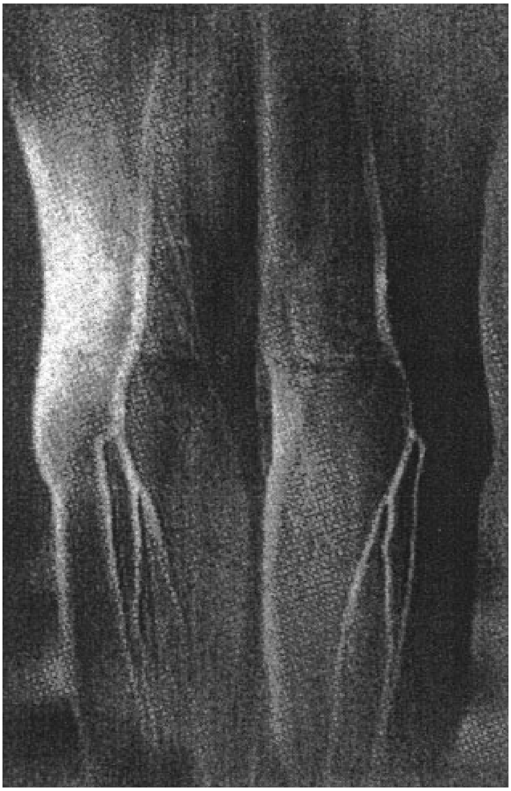

Fig. 27.27 Magnetic resonance image of blood vessels in the legs, obtained by subtraction of the systolic image from images during diastole. This is one of the earliest examples of magnetic resonance angiography. (From Newhouse, V.L. (1993) Imaging techniques, biomedical, in *Encyclopedia of Applied Physics*, Vol. 7, VCH, New York, with permission.)

slice during RF activation is "tagged," in the sense that the nuclear spins are oriented perpendicularly to the applied field; the path of this "tagged" blood can be followed in subsequent MR images.

Phase encoding can be applied in a manner that preserves the MR signal during diastole (when the heart is expanding and the blood is essentially stationary), and ablates the MR signal during systole (when the heart is contracting and the blood is moving rapidly). By subtracting the systolic image from the diastolic image, the moving blood in selected blood vessels can be revealed with excellent clarity (Figure 27.27).

At the other extreme from the rapid flow of blood through arteries and veins, *diffusion MRI*, which is becoming widely used in the assessment of acute stroke and malignancies, is able to reveal information (the diffusion tensor) about the actual directions of fluid (primarily water) diffusion in tissues (Holodny *et al.*, 2005; Yoshiura, Mihara and Tanaka, 2005).

27.8.6
Functional MRI (fMRI)

Functional magnetic resonance imaging (*fMRI*), and in particular *blood oxygen level-dependent* (*BOLD*) imaging, can determine where in the brain the instantaneous flow of oxygenated blood is unusually high and, by inference,

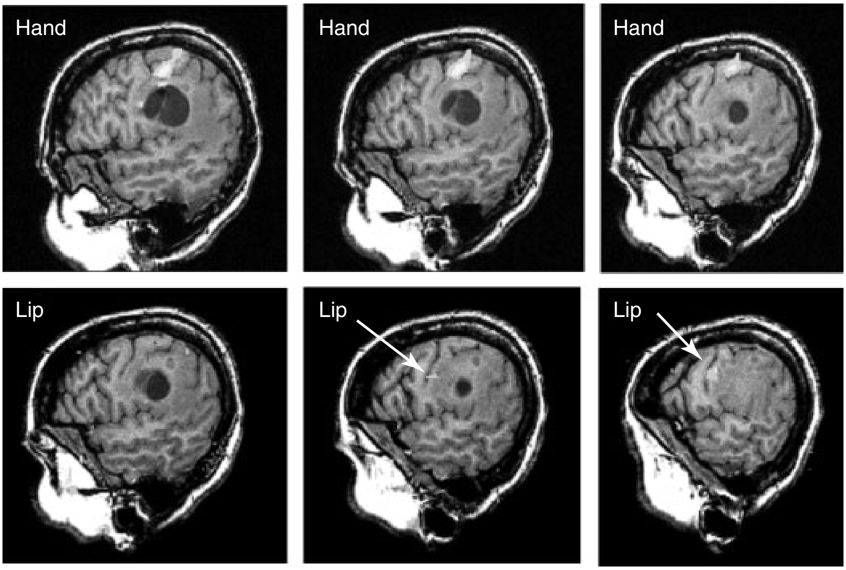

Fig. 27.28 44-year-old woman with a brain tumor. fMRI shows the location of the motor cortex that controls the hand and the lips on one side of the body. These cortical regions are important to avoid when resecting the brain tumor. (Obtained courtesy of John Ulmer, MD, Medical College of Wisconsin.)

where neurons are unusually busy (Figure 27.28).

The magnetic properties of deoxygenated blood (which is diamagnetic) differ from those of paramagnetic oxygenated blood, so in a region of tissue where deoxygenated blood is being replaced with oxygenated blood, the magnetic properties of the tissue are changing. In the brain, such a change may be triggered by increased blood flow to a region in which neuronal activity increases in response to an external stimulus (e.g. sound, light, and touch) or an internal stimulus (thinking, calculating, sending instructions to extremities, etc.). These changes can be detected and portrayed to yield a map of neuronal activity (Rosen *et al.*, 2000). The method provides information on neural activity that is complementary to contributions from PET and magnetic resonance spectroscopy (MRS) (Price *et al.*, 2002; Elas *et al.*, 2003).

As an example, in a study of potentially staggering implications, fMRI allowed researchers to determine which of several visual patterns that subjects were looking at – in effect, it allowed researchers to read the subjects' minds (Kamitani and Tong, 2005; Phan *et al.*, 2005; Faisan *et al.*, 2005).

27.8.7
MR Spectroscopy (MRS)

A proton on a molecule other than water resonates at a slightly different NMR frequency. This happens because the pattern of electron flow in its vicinity is not the same, and this leads to a slight difference in the local magnetic field that the proton experiences. This *chemical shift*,

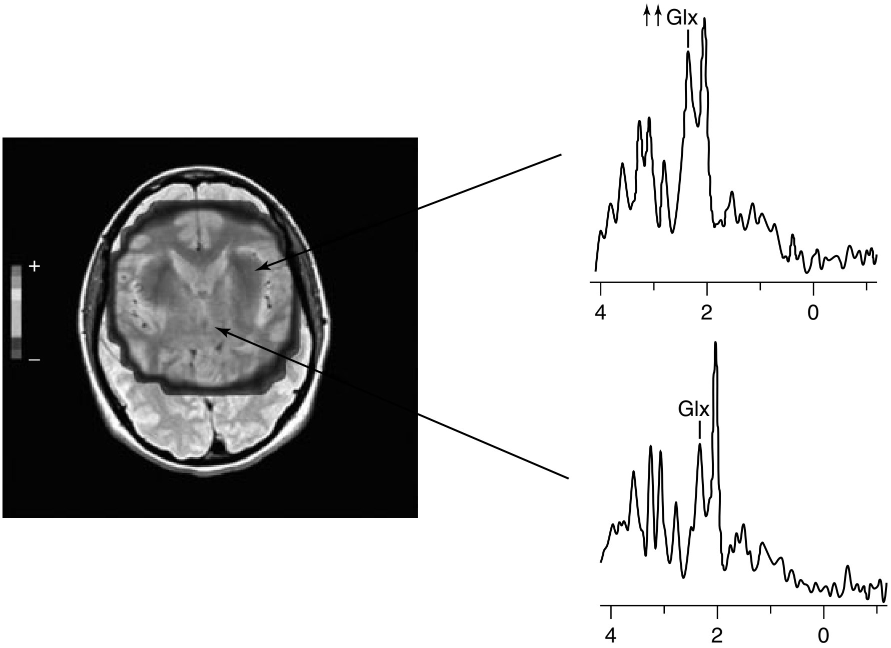

Fig. 27.29 Chemical shift/MR spectroscopic imaging (1000/18; bandwidth, 1000 Hz; 512 points acquired with two signals acquired). Transverse short echo-time chemical shift image (left) acquired at 0.5 T in a presymptomatic patient with Huntington disease shows strong elevation of glutamate (right: upper spectrum) in the head of the putamen. Unaffected thalamus (right: lower spectrum) is shown for comparison. (Image courtesy of Robert Prost, PhD, Medical College of Wisconsin, Milwaukee, WI.) (Please find a color version of this figure on the color plates.)

which is typically on the order of parts per million and increases with the strength of the external magnetic field, B_0, makes possible *MRS*. MRS can elucidate chemical processes occurring in small volumes of tissue, and is clinically proving effective in distinguishing among various forms of metastatic and benign tumors (Fayed and Modrego, 2005).

As with fusion PET/CT, MRS can be combined with high-resolution MRI to enhance sensitivity and specificity in the detection and identification of cancers and other abnormalities (Figure 27.29). And MRS of carbon-13, fluorine-19, and phosphorus-31 (rather than hydrogen-1) is also providing valuable insights into cellular metabolism and biochemistry (Stanley, 2002).

27.8.8
Higher Magnetic Fields

An area of considerable interest is the performance of MRI at higher magnetic fields ($\geq$3.0 T). As indicated by the Boltzmann factor, the magnetization in each voxel increases nearly linearly with the strength of the external magnetic field, B_0. As a result, the NMR signal strength and the signal-to-noise ratio also increase with magnetic field strength. The absolute amount of the chemical shift is also proportional to B_0, which can help to untangle messy spectra. The T1 relaxation times also may increase, and this may contribute to enhanced image contrast. The biophysical effects of operating at high magnetic fields are not well known yet, however. Also, stronger magnets are more costly.

27.8.9
Risks from MRI

MRI does not use ionizing radiation, and the absence of radiation dose to patient and staff is commonly touted as a significant plus. The effects on tissues of intense static magnetic fields, and of rapidly switched gradient fields and RF power, have been studied extensively. The Food and Drug Administration and various advisory bodies publish recommendations for limiting the principal magnetic field, the rate of switching gradient fields, RF power levels, and so on, which are thought to provide reasonable patient protection. None of these processes in and of themselves appear to pose serious risks to patients or staff under current clinical conditions when standard operational guidelines are followed.

Many new imagers use a field strength of 3 T, and a few even higher than that. The power of the pulsed RF fields is increasing, moreover, as are the strength and switching speeds of gradient fields. Studies of the long-term biologic effects of these changes continue, but meanwhile it clearly makes sense to exercise caution, especially in examining pregnant women, small infants, and patients on life support systems.

The most immediate and obvious source of danger is the principal magnetic field employed in MRI. The field is extremely strong, fills a large volume, and extends some distance beyond the bore. It can exert forces powerful enough to sling-shot a screwdriver into the magnet, or drag in an oxygen bottle or i.v. pole. Untrained personnel – housekeeping, maintenance, security, police, firemen, parents, non-MRI physicians, and other medical staff – may inadvertently bring clipboards, buckets, vacuum cleaners, scissors, car keys, or whatever into the vicinity of the magnet. If those objects have ferromagnetic parts, in an instant they can turn into lethal projectiles.

Patients or anyone else within whom there are aneurysm clips, shrapnel, or other metallic items should normally come nowhere near an MRI device until the objects are determined to be nonferromagnetic and safe. There is need for caution in examining patients with some types of surgical pins, metallic intrauterine devices, cochlear implants, and heart valves.

Another hazard is posed by the pulsed RF fields and rapidly switching gradients. Most importantly, they can disrupt the operation of cardiac pacemakers and other electronic devices.

Under some conditions, they can also induce currents that heat electrode wires or other metal objects on or within the patient, and this heating can lead to severe (albeit localized) burns. And for some pulse sequences and in tissues at certain positions within the bore, they can produce noticeable sensations even with no wires or other things present.

An MRI device can wipe out magnetic credit cards, watches, and certain medical electronic gear. It also is necessary to shield from the magnetic fields and RF noise produced by passing trucks and elevators and even flickering fluorescent lights. The RF problem can be solved through construction of a Faraday cage, a wire-mesh or solid metal panel enclosure around the patient (and RF coil) area. The MRI magnetic fields can be confined, and other fields excluded, with steel plates either in close proximity to the magnet or in the walls of the imaging room.

The objective is to prevent accidents related to magnetic or RF fields before they occur. Accomplishing this objective requires posting highly visible or audible warnings of the hazards; providing training for those who may be coming in and out of the facility or caring for patients; and, most importantly, establishing a system for scrutinizing all individuals seeking to gain entry to the device area, and enforcing appropriate restrictions on site access, even during an emergency.

27.8.10
Recent Developments in MRI

The ability of MR to provide flexible imaging parameters, and thereby images that demonstrate differences in a variety of tissue properties, surpasses that of any other imaging method. This capacity, together with the superb contrast differentiation among soft tissues, makes MRI a superb imaging modality with applications that are still evolving.

There is a wide range of variations and refinements in the general approach to signal generation, acquisition, and localization that permit MRI images to provide an ever-expanding range of information. Images can be presented for different slice thicknesses and at any desired orientations throughout the body. They can be designed to display information about hydrogen-atom (water or fat) concentrations, and about variations in T1 and T2 among tissues. A major objective of MR researchers and manufacturers has long been to reduce image-acquisition time without loss of image quality (Nitz, 2002; Jahng, Stables and Ebel, 2005). Obtaining signals from multiple RF coils simultaneously, a process known as *parallel imaging*, substantially reduces the number of phase-encoding pulses required and, therefore, the scan time (van den Brink *et al.*, 2003).

MRI is a vibrantly evolving technology, and its influence on clinical medicine and patient care continues to expand. It can be argued that the increasing flexibility, wider applications, declining costs, more subtly nuanced pulse sequences, and growing magnetic field strength (i.e. the increase from 1.5 to 3.0 T) of MRI, as well as the associated improvement in signal-to-noise ratio (Hu and Norris, 2004; Okada *et al.*, 2005), are placing the technology in a position to displace many of the applications of CT, PET, and conventional angiographic methods over the coming decade.

27.8.11
Electron Resonance Imaging

Electron-spin resonance (*ESR*) imaging, also known as *electron paramagnetic resonance* (*EPR*) imaging, is analogous to MRI, but involves stimulated and relaxation spin transitions of unpaired electrons rather than of unpaired protons (Elas *et al.*, 2003; Berliner, 2003). Since an electron is three orders of magnitude lighter than a proton, it has a correspondingly greater magnetic moment and higher Larmor frequency. At 1 T, an electron's MR occurs at 28 GHz, rather than at 42 MHz for a proton (1 GHz = 1000 MHz).

Microwave energy tends to be strongly absorbed in the process of excitation of rotational states in water – indeed, that is how microwave heating works – and at gigahertz frequencies, microwaves are absorbed in a millimeter or so of skin. At considerably lower resonance frequencies and field strengths (e.g. 0.01–0.04 T), however, the improved penetration permits imaging of free radicals *in vivo* in small animals. A variation on the theme, known as *proton–electron double resonance imaging* (*PEDRI*) involves the simultaneous performance of electron and nuclear MRI (Figure 27.30) (Liebgott *et al.*, 2003).

27.9
Laser Optical Imaging

The properties of visible and near-visible light transmitted through regions of the body have long intrigued scientists.

In the 1980s, interest surfaced in the use of transmitted light to image the female breast as a screening method for cancer. *Diaphanography* (*transillumination*) is the medical counterpart to shining a bright flashlight through the hand to display the bones. As with X-rays, beams of visible and near-IR light are differentially attenuated as they pass through different parts of a breast, and their shadows can be recorded as images on film or with an electronic camera. Transillumination of the breast can sometimes help distinguish benign from malignant masses – but this modality produces more false-positive and false-negative results than do conventional mammography and US, and it has not been accepted as a standard clinical tool.

The clinical utility of optical methods depends largely on the ability to extract image information about objects embedded in turbid media, so there have been serious efforts to reduce the level of scatter radiation reaching the image receptor. Improvements both in *time-resolved spectroscopy* techniques and in the theory of photon transport in tissues have contributed to advances in the field. One family of time-resolved spectroscopy methods, for example, employs gated detectors that are so fast that, following an extremely short laser pulse, only unscattered photons (which travel directly from source to detector and arrive first) are accepted. This approach can display tumors and other abnormalities by recording light transmission as a laser is stepped through a region of interest – a much more refined version of diaphanography.

Laser optical tomography or, for one particular application, tomographic laser mammography, yields cross-sectional images of tissue obtained by projecting laser beams inward from many directions. These images are useful in studies of blood perfusion, tissue oxygenation, and neovascularity in the brain, breast, and extremities.

In one interesting variation on this idea, *photographic acoustic (photoacoustic) tomography* exploits absorption differences among various tissues for laser light pulsed at megahertz frequencies. The pulses heat up and thermally expand small regions of tissue extremely rapidly, thereby producing ultrasonic waves that can be detected by sensitive piezoelectric receptors.

With a somewhat different approach, *confocal scanning laser tomography* can noninvasively acquire three-dimensional images of the posterior segment of the eye, creating a quantitative description of

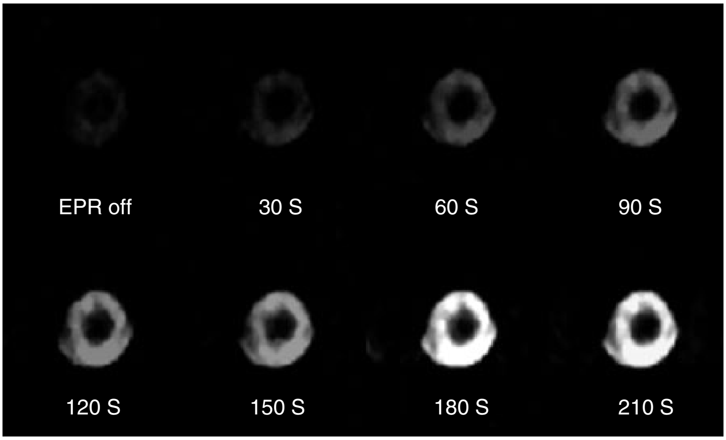

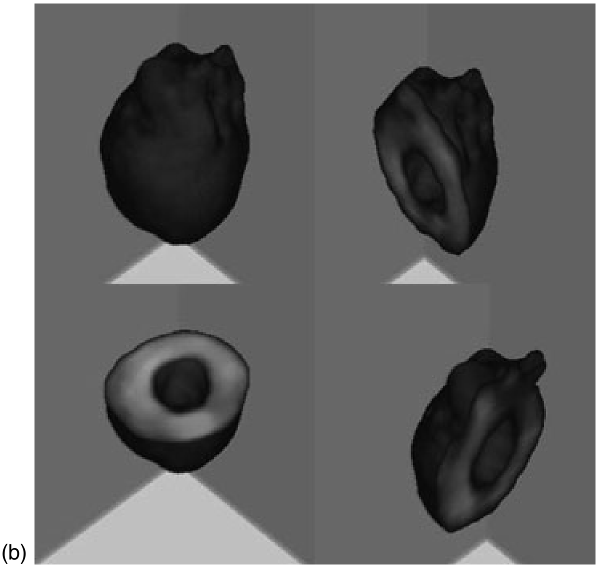

Fig. 27.30 Proton–electron double resonance imaging (PEDRI); also known as *Overhauser imaging*. (a) Time course of PEDRI study of myocardial uptake of free radical probe TEMPONE (4-oxo-2,2,6,6-tetramethylpiperidine-1-oxyl) by isolated perfused rat heart. TEMPONE as infused through a side arm proximal to the perfusion cannula at final concentration of about 3 mol l^{-1}. Two-dimensional PEDRI sections were then sequentially acquired every 30 seconds, with each scan taking 27 seconds. At low field strength of 0.02 T, the electron-spin resonance and nuclear MR frequencies are 567 MHz and 856 kHz, respectively, and an Overhauser enhancement of −13 was achieved. (b) Three-dimensional gradient-echo PEDRI images of isolated beating rat heart infused with 3 mol l^{-1} TEMPONE. Top left image shows complete three-dimensional surface-rendered image; the other images are cutaways to show internal structure. The image took 4 minutes 30 seconds to acquire at 0.02 T. (Images courtesy of Haihong Li and Jay Zweier, MD, Davis Heart and Lung Research Institute, Ohio State University, Columbus, Ohio.)

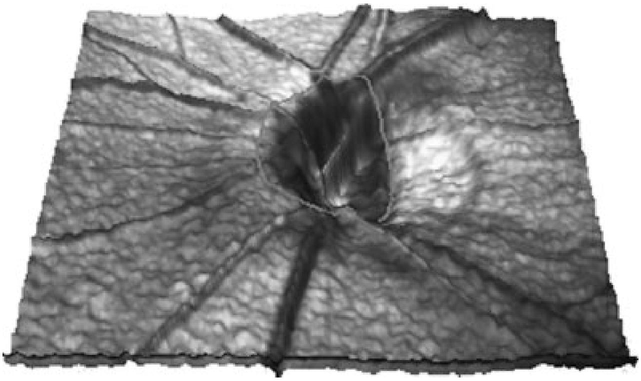

Fig. 27.31 Confocal scanning laser tomographic image of posterior segment of the eye shows topography of optic nerve head and of surrounding retinal surface in three dimensions. In less than 2 seconds, the 670-nm laser performed a sequence of 64 scans of the retina over a 15°–15° field, creating 64 386–386-pixel planar images out of light reflected from different depths. The shape of the edge of the indentation edge, emphasized by the drawn line, indicates nerve fiber layer defect at the rim of the optic nerve head. (Image courtesy of Heidelberg Engineering GmbH, Dossenheim, Germany.)

the optic nerve head and the surrounding retinal surface (Figure 27.31). A laser beam is focused to some depth within the eye and scans a two-dimensional plane. Only light from that focal plane is allowed to reach the detector. A sequence of such two-dimensional optical planar views is acquired for increasing depths of the focal plane, and the result can be displayed as a three-dimensional topographic image of the optic nerve head and peripapillary retinal nerve fiber layer. The approach is useful not only for detecting abnormalities directly but also for tracking subtle changes over time that would be difficult to detect clinically. A recent study suggests that, with expert interpretation, the power of the technique as an aid in the diagnosis and management of glaucoma is comparable to that of the accepted standard of stereofundus photography (Zangwill *et al.*, 2004).

Optical techniques such as optical coherence tomography and phase-resolved microscopy permit real-time micrometer-scale imaging and very sensitive display of cellular dynamics to depths of several millimeters in tissue. In *optical coherence tomography*, a thin beam of pulsed optical or IR laser light passes into an interferometer, one arm of which is directed at tissue through an optical fiber, which is, perhaps, part of an endoscope (Yu *et al.*, 2003). Light that scatters coherently in the top few millimeters of tissue contains information not only on the relative absorption of the radiation but also on changes in the light's phase. Scattered radiation returning from the tissue is caused to interfere with light from the reference arm, so that the phase shifts and degree of absorption can be measured. When such interference data are obtained for a 1024 × 1024-pixel matrix imposed on the tissue surface, the system can form 1-megapixel cross-sectional images in real time. Since the time of arrival of the scattered photons at the detector depends on the depth of the scatterer within the tissue, different levels in the tissue can be examined separately. Penetration is only several millimeters, but resolution can be 10 μm or better. Combination of the images with spectroscopic information can provide even more detail about the tissue for possible use as an optical biopsy.

Diffuse optical tomography provides measurements of hemodynamics and neural activation at depths of several centimeters in tissue. *Nonlinear microscopy*, which employs nonlinear optical methods such as mutiphoton molecular excitation, optical harmonic generation, and depletion of stimulated emission, can image subcellular morphology and trace molecular dynamics at subnanometer resolution at depths of up to a fraction of a millimeter in living tissue.

Tissue is nearly opaque throughout most of the IR, visible, and ultraviolet parts of the EM spectrum. One way to deal with this challenge is to use a very intense laser beam, as in diaphanography. Another is to take advantage of a narrow window of transparency in the near-IR region, with wavelengths in the 700–900 nm range. An important set of applications makes use of the fact that oxygenated and deoxygenated hemoglobin both absorb these wavelengths well but have spectra that differ enough to allow distinction between the two species. This distinction can be important in studies of blood flow and oxygen consumption. Each year, one million women undergo core-needle breast biopsy in the United States. While considerably easier on the patient than excisional biopsy, the core-needle technique suffers from a false-negative rate of up to 7%, despite needle guidance with X-ray fluoroscopy or US. But a more sophisticated probe currently under development may substantially reduce the false-negative rate (Breslin *et al.*, 2004; Lubawy and Ramanujam, 2004). After the probe has been guided to the tumor site within the breast, it transmits near-IR laser light into the breast tissue, and then senses the light coming back from it. The spectrum of the light returning from a tumor may differ from that from normal tissue because of irregularities in the degree of oxygenation, fluorescence characteristics, or other properties. In cases where no difference is detected, the tip of the probe can be repositioned to further search for disease. On finding tissue that yields a difference in returning light, the probe is used to acquire a core biopsy specimen. Unlike mammograms, the laser probe works well in dense tissue, such as in the breasts of young women, and may find applications in other organ systems as well.

27.10
Cellular and Molecular Imaging

Molecular imaging is the minimally invasive *in vivo* sensing, depiction, and characterization of spatially localized biologic processes at the cellular and molecular levels (Britz-Cunningham and Adelstein, 2003; Schultz *et al.*, 2004; Michalet, Bentolila and Weiss, 2006). Having evolved out of conventional nuclear medicine and still largely in a preclinical stage, molecular imaging now represents “the convergence of multiple image-capture techniques, basic cell/molecular biology, chemistry, medicine, pharmacology, medical physics, biomathematics, and bioinformatics into a new imaging paradigm'' (Massoud and Gambhir, 2003).

The development of molecular imaging is attributable, in part, to the synthesis of novel molecular agents that attach with high specificity to genes, proteins, and other biomolecular targets and that can then be detected with various imaging modalities (Schultz *et al.*, 2004). Among the agents are those that are activated only in response to particular changes in the local biochemical environment, such as specific gene expression or enzyme activity (Allen and Meade, 2004; Shah and Weissleder, 2005).

Molecular imaging is being driven also by the design of micro-CT, micro-MR, micro-PET, optical bioluminescence, fluorescence molecular tomography, and other methods for the study of small laboratory animals (Figure 27.32). In general, the corresponding clinical machines designed for 70-kg humans are scaled down, often with improved image quality and almost always at substantially reduced cost. The image receptor for micro-CT, for example, can be a commercially available ''off-the-shelf'' CCD (Wang, Zhao and Yu, 2005; Schelbert, Inubushi and Ross, 2003).

Molecular imaging technology – indeed, the whole field of biophotonics – is undergoing leaps in sophistication (Vo-Dinh, 2003). In addition to the new pharmaceuticals, imaging techniques, and methods for therapeutic monitoring, there are numerous recent reports on the monitoring or imaging of exogenous gene delivery and expression, protein–protein interactions, apoptosis as a predictor of therapeutic response, tumor growth, dopamine and serotonin neurotransmission, and much else. Also being studied are the dependence of absorption, scatter, and fluorescence of various wavelengths of light by tissues that may reflect the cellular integrity, matrix materials, and vasculature, all of which may depend on the patient's age, gender, state of health and other characteristics.

Fluorescent lifetime imaging microscopy, for example, exploits the dependence of the fluorescence decay time on tissue type to help identify abnormalities (Godavarty, Sevick-Muraca and Eppstein, 2005). Another approach makes use of *quantum dots* (*qdots*), which are fluorescent semiconductor nanocrystals (Gao *et al.*, 2004; Michalet *et al.*, 2005; Morgan, English and Chen, 2005). At nanometer sizes, a system begins to act somewhat like a large molecule or a set of interacting large molecules, with quantum mechanical properties quite different from those of a bulk piece of the same crystalline or amorphous material. Because of quantum confinement effects, the narrow, laser-stimulated emission spectra of quantum dots depend on their sizes, which can be precisely controlled during synthesis. When qdots emitting various colors are conjugated to antibodies, antibiotics, or other recognition molecules and are then injected intravenously, they can help detect and map out a number of specific cellular markers of diseased tissues simultaneously – a form of *in vivo* multiplexed optical biopsy. Also, MRI, PET, and CT contrast agents can be grafted onto their surface so that they, too, can provide molecular imaging diagnostics. It may even be possible to attach therapeutic agents to quantum dots, which could be activated once in proper position.

Considerable numbers and varieties of molecular imaging systems are being investigated and, as with any new

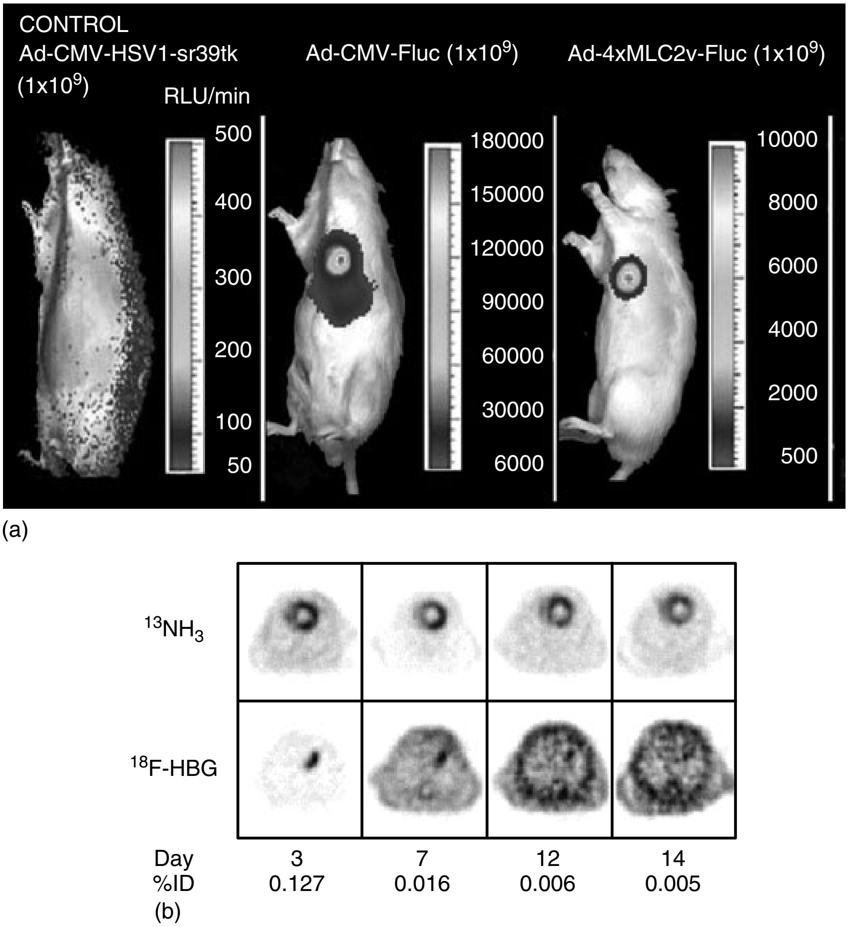
CONTROL
Ad-CMV-HSV1-sr39tk
(1x10^9)
RLU/min
500
400
300
200
100
50
Ad-CMV-Fluc (1x10^9)
180000
150000
120000
90000
60000
30000
6000
Ad-4xMLC2v-Fluc (1x10^9)
10000
8000
6000
4000
2000
500
(a)
$^{13}NH_3$
^{18}F-HBG
Day
3
7
12
14
%ID
0.127
0.016
0.006
0.005
(b)

Fig. 27.32 Biomolecular imaging. (a) Optical bioluminescence imaging of cardiac gene delivery shown on charge-coupled device optical images of three rats injected (directly into the lateral wall) with control adenovirus (*Ad*) (left), adenovirus carrying firefly luciferase (*Fluc*) driven by cytomegalovirus (*CMV*) promoter (middle), and adenovirus carrying firefly luciferase driven by myosin light chain (*MLC*) promoter (right). Middle: Image shows expression in the myocardium and liver due to leakage of adenovirus into the bloodstream from the injection site. Right: Image shows only cardiac expression of luciferase due to use of myosin light chain promoter, which is not markedly active in hepatocytes. Left: Control image shows background low activity. All images show light image of mouse onto which bioluminescence image is superimposed, in relative light units (*RLU*) per minute. (Image courtesy of Sanjiv Gambhir, MD, PhD, Stanford University, Palo Alto, CA.) (b) Serial transverse PET images of reporter gene expression in a rat as function of time in days after dose injection. Top row: Animal was injected with 13N-labeled ammonia. Bottom row: Injected agent is 18F-labeled *HBG* gene. %*ID* × percent of injected dose. (Image courtesy of Heinrich Schelbert, MD, PhD, David Geffen School of Medicine, University of California at Los Angeles.) (Please find a color version of this figure on the color plates.)

field, are undergoing rapid growth. Few are similar to one another in design. A plethora of recent publications describe these systems, including a recent issue of *Medical Physics* that contains abstracts from a major symposium on the subject (Medical Imaging Research Symposium, 2003). Topics include, in addition to the ones discussed above, diffuse optical tomography (Choe *et al.*, 2005); fluorescence resonance energy transfer, Fourier transform IR spectroscopy, autofluorescence imaging, hyperspectral imaging, video capillaroscopy, and others.

27.11
Emission Imaging

27.11.1
Electrical Source Imaging: Electrocardiography and Electroencephalography

The normal movement of ions within and between active nerve or muscle cells gives rise to weak, transient electric and magnetic fields that can extend to the surface of the body and beyond.

A general approach, *electrical source imaging* (*ESI*), has been devised to detect, and perhaps create images from, faint electrical fields with devices positioned external to the body. This method is termed *electrocardiography* (*ECG, EKG*) when applied to the heart and *EEG* for studies of the brain.

To map the surface potential on the scalp with reasonable resolution requires 20 or more electrodes, and 200 or more are needed for the torso. At present, sites of intense electrical activity in the heart and brain can be localized by image reconstruction techniques; the shapes of the voltage pulses as they change over time can also be clinically informative. Arrays of many electrodes spread over an area of the patient may eventually succeed in generating useful information in image form as well. One major problem is the rapid attenuation of electric signals in the body, so that those that do reach the electrodes are very weak and difficult to extract from background electrical noise.

ECG and EEG signals may be indicative of arrhythmias in the heart or epileptic foci in the brain. Attempts to identify more than a few sites of electrical activity are hampered by difficulties encountered with inverse computation of the origin from field measurements. Also, electrical fields are rapidly attenuated in tissue, so the fields reaching the surface are very weak.

27.11.2
Tissue Electrical Impedance Tomography

Although not exactly a form of ESI, the related discipline of tissue *electrical impedance tomography* (*EIT*) is receiving considerable research attention.

Electrons and ions in the fluids in and between cells can move about, to a greater or lesser extent, in response to an applied electric field. The standard measure of the resistance of tissue to this movement of electrons and ions is referred to as the tissue's *electrical impedance*. To assess the impedance (Z) of tissues in a region, numerous electrodes are attached to the skin, low voltages (V) are applied between pairs of electrodes, and the electrical currents (I) produced are assessed. Ohm's law ($Z = V/I$) then yields the impedance. After many such measurements have been obtained, EIT uses CT-like reconstruction calculations to generate three-dimensional maps of tissue impedance throughout the region (Poplack *et al.*, 2004; Grimnes and Martinsen, 2000; Ireland *et al.*, 2004). It is also possible to measure differences in the dielectric constant of the tissues, which is the reason why the method is referred to as impedance, rather than resistance, tomography. Spatial resolution is limited by the finite number of electrodes that can be fitted in place.

Two serious complications in impedance imaging are that only a relatively small number of electrodes can be fitted onto a given surface area, and that electric currents (unlike X-rays) do not follow straight lines. On the other hand, electric impedance tomography is safe, and can be used at the bedside, allows for long-term data acquisition, costs little, and images a new physiologic function that may turn out to be uniquely revealing for some disorders.

A*pplied potential tomography* (*APT*) is a closely related alternative for data acquisition. Voltages are applied at two points on the skin, and the resultant voltages are measured at other points on it. This process is repeated for the other pairs of electrodes. Another approach being explored is to place the region of interest in a strong magnetic field, after which passage of a high-frequency alternating current between electrodes generates a detectable US signal.

APT has been applied to the study of physiological processes such as gastric emptying, pulmonary ventilation, cardiac output, and blood flow in the thorax, head and extremities, and temperature changes during microwave hyperthermia treatments. Like MRI, tissue impedance imaging is sensitive to changes in the water content of tissues. Indeed, tissue impedance imaging is being developed by the military to monitor blood loss from wounded soldiers. Other applications include studies of gastric emptying, pulmonary ventilation, cardiac output, blood flow, and intracranial hemorrhaging in infants. Since malignant tumor tissue may be an order of magnitude or so more conductive than surrounding normal tissue, instruments have been built to facilitate the assessment of ambiguous breast masses.

27.11.3
Magnetic Source Imaging: Magnetoencephalography and Magnetocardiography

Like any other electrical currents, those within and between neurons and muscle cells generate weak magnetic fields. *Magnetoencephalography* (*MEG*) and *MCG* are similar to EEG and ECG except that they are used to detect and monitor the varying weak magnetic (rather than electric) fields produced in this fashion (Wheless *et al.*, 2004; Lesser, 2004; Barkley, 2004; Baumgartner, 2004). MEG measures the magnetic fields with detectors positioned around the head and back-calculates the locations and strengths of the neural currents that gave rise to the fields. These fields are about one billionth (10^{-9}) as strong as that of the Earth, so near-perfect magnetic shielding and highly sensitive detectors are required. Fortunately, the exquisitely sensitive *superconducting quantum interference device* (*SQUID*) fills the bill nicely: It is responsive enough to detect the movements of a refrigerator-door magnet a quarter mile away.

With millimeter and millisecond spatial and temporal resolutions, MEG demonstrates neural activity directly, as

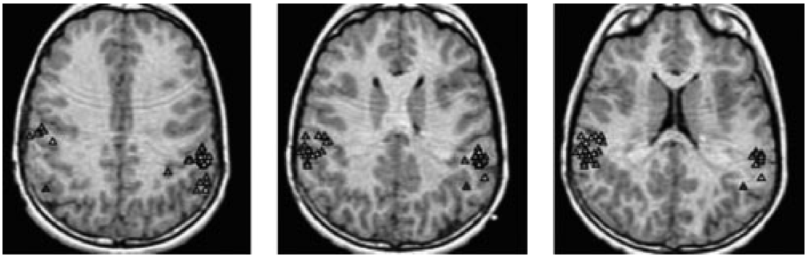

Fig. 27.33 Magnetoencephalographic spatial information superimposed on transverse T1-weighted MR sections (13/4.47). Magnetoencephalographic spike activity (*D*) in a 10-year-old patient with Landau–Kleffner syndrome. (Image courtesy of M. Funke, MD, PhD, Center for Advanced Medical Technologies, University of Utah, Salt Lake City, Utah.)

it takes place. fMRI and PET provide indirect assessments of brain function and they show only relatively slow variations, over seconds or even minutes, in metabolic activity. Unlike electrical fields, magnetic fields are transmitted through tissue with little attenuation or distortion, so magnetic source imaging (MSI) images tend to contain less noise or distortion. A major weakness of MSI, however, is that any magnetic field distribution on the surface of the body can be explained by an infinite variety of current distributions inside the body. This limitation is known as the *inverse computational problem*, and has a counterpart in ESI. Successful application of MSI (and ESI) usually requires prior knowledge of the object, such as the underlying anatomy, as obtained by MRI or X-ray CT.

MEG and MCG have already proven their clinical worth in noninvasive studies of well-localized sites of intense neuronal activity. Applications of MCG include identification of the origin of cardiac arrhythmias, and the guidance of a catheter directly to the site for treatment. MEG tracks the much weak magnetic field changes that accompany epileptic activity in the brain, migraine headache, and diabetic coma, and can detect functional areas of the brain that must be preserved during surgery (Figure 27.33). Like PET and fMRI, it can follow the brain's response to stimuli. It can be used to demonstrate subtle abnormalities in a region of the brain that appears to be structurally and physiologically normal when viewed with the aid of other modalities (Georgopoulos *et al.*, 2007).

On a different note: Under exploration as an alternative to electroshock therapy is the converse process of magnetic nerve stimulation, in which a pulse of electric current flow is generated in a well-localized region of the brain by rapidly changing the field from a powerful external magnet (Hallett and Chokoverty, 2005).

27.11.4
Infrared (IR) Thermography

The warmer something is, the more heat it emits, at a rate that is approximately proportional to the fourth power of the absolute temperature (i.e. approximately T^4). At room temperature, a body radiates EM energy that peaks in the IR region. Since IR radiation travels through soft tissues for distances only of about one wavelength (~0.01 mm), the emission rate depends on skin temperature. But skin temperature, and the local IR radiation intensity, may be influenced by disease conditions that lead to changes in superficial blood perfusion.

With one type of *radiative IR thermography* camera, a detector such as InSb (with short-wavelength IR sensitivity) or HgCdTe (long-wavelength IR) cooled to liquid-N_2 temperature scans a region of skin surface by means of a rotating prism and lens. It can operate at up to 25 frames per second, so that essentially real-time IR thermography is possible.

IR thermography has explored the angiogenesis and neovascularity that accompany breast disease, inflammatory conditions such as rheumatoid arthritis in joints, flow disturbances in blood vessels, injured muscles, diminished enervation of muscles, burns, and frostbite (Diamantopoulos, 2003).

27.11.5
Microwave Imaging

Room temperature objects also radiate energy in the microwave region, although at an intensity about 10 orders of magnitude less than that of IR radiation. Still, spontaneously emitted microwave radiation has imaging possibilities because it is more penetrating than IR, so that it can reach the skin surface from greater depths. The much longer wavelength of microwave, however, places a fundamental limitation on the spatial resolution achievable. For microwave radiation at 2 GHz, for example, the resolution and penetration depth in muscle are each about 1 cm, and the corresponding value for fat is closer to 10 cm (Newhouse, 1993). Also, because of its much lower intensity, microwave radiation data-acquisition times are considerably longer.

Microwave imaging is particularly sensitive to differences in absorption between fat and other soft tissues. The wavelength is too long to allow the degree of spatial resolution normally required for diagnostic imaging, but microwave devices may find other applications (Poplack *et al.*, 2004; Ciocan and Jiang, 2004). For example, promising preliminary clinical studies have been reported for a small device called a tissue resonance interferometer. This device generates low levels (less than that of a cell phone) of 400–1350 MHz radiation, and the returning signal may be altered by tissue irregularities, including tumors.

27.11.6
Acoustothermometry

Objects above absolute zero emit sound radiation that is described as *ultrasonic thermal noise*. The sound energy is emitted at very low intensity (e.g. 10^{-16} W cm^{-2} over a 100 kHz band at room temperature), and long data-acquisition times are required. An acoustothermometer that measures temperatures to an accuracy of 0.2 °C at depths of up to 10 cm in tissue with data-acquisition times of 30 seconds, for example, has been described (Godik and Gulyaev, 1991). Acoustothermometry has been proposed as a monitoring technique for hyperthermia treatment of tumors (Newhouse, 1993).

27.11.7
Terahertz Imaging

The terahertz portion of the EM spectrum lies between 300 and 100 mm in

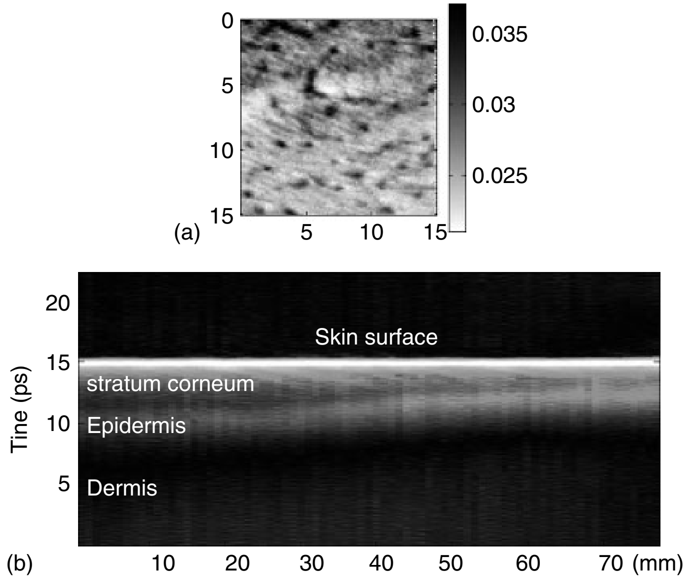

Fig. 27.34 Terahertz imaging. (a) *In vivo* terahertz image of volunteer's forearm acquired by using a handheld terahertz imaging system. The image shows a 15 × 15-mm region; a scar running from left to right can be seen in the top half of the image. The image was generated by plotting the electric field value reflected from beneath the skin surface. Dark circular regions are hair follicles of normal skin, which are not present at the scar. (b) Axial terahertz image (b scan) of edge of volunteer's hand. Gray scale indicates signal amplitude, which is plotted against optical delay (*y* axis) and position across the scanned area (*x* axis). Decrease in stratum corneum thickness across the *x* axis, from the palm-side (30 mm on *x* axis) to the backside (70 mm) of the hand, is evident. (Image courtesy of Vincent Wallace, PhD, TeraView Ltd, Cambridge, England.)

wavelength. The so-called terahertz rays, or T-rays, do not penetrate water or tissue well, so they would be of little use in examining deep-seated tissues. A substantial fraction of cancers lie in the epithelium, however, and while many of these are readily apparent to the trained eye, some that are small and flat can be overlooked or are simply not visible. Standard modalities are not adept at depicting or characterizing epithelial tumors – but *terahertz imaging,* because of the ability to recognize the spectral fingerprints of surface proteins that are markers for certain cancers, seems capable of detecting them at an early stage, when they can still be treated effectively (Figure 27.34). Terahertz rays are nonionizing and offer imaging resolution of less than a millimeter; the equipment is safe and portable. But there are major difficulties with terahertz sensing and imaging systems, including their inefficiency and high cost of production. However, recently created photonic band-gap materials may soon change that situation (Poplack *et al.*, 2004; Walker *et al.*, 2004; Wang and Mittleman, 2004).

27.12
Mini- and Nanotechnology

Small-scale technology offers exciting possibilities for medical imaging. There already exists a disposable video camera–transmitter–light source combination the size of a pill that sends color images to a receiver–recorder (strapped to the waist) as it wends its way through the bowel. Compared with endoscopy, limitations of the system are related to the inability to control the position of the camera and its direction of view and battery lifetime. These challenges, however, are surmountable.

At the smaller level, quantum dots and other nanoparticles that display specific physical or chemical characteristics, such as stimulated fluorescence or paramagnetism, or objects coated with nanoparticles, may be detected with appropriate imaging technologies (Wickline and Lanza, 2003). Ceramic, metallic, and organic nanotubes, nanowires, and other structures have been the subject of intense investigation, as have been (more recently) nanostructures such as minute sensors, motors, or other electromechanical devices (which might contain biologic components). They could enter a blood vessel and be directed to an area of interest, where they could acquire and transmit images; implement enzyme-, antibody-, or genome-based biologic tests; perform biopsies; build tissue-like materials and organs; and even carry out small-scale surgeries. While some might find this all to be a bit too Star Trek-ish, others would argue that it is already happening (Thrall, 2004; Silva, 2004). The National Cancer Institute, in particular, has recently announced a 5-year, $144 million program to develop nanotechnology in cancer detection and treatment.

27.13
Roles of Computers in Medical Imaging

Since it began playing a significant role in imaging four decades ago, the computer has made its presence felt primarily in image acquisition and processing, and through its tools to enhance contrast, sharpen edges, and reduce noise (Barrett and Myers, 2004; Bankman, 2000; Sonka and Fitzpatrick, 2000). But it appears that now and in the near future, there will be an increasing focus on faster, greater-capacity, and more reliable image communication, archiving, retrieval, and interpretation capabilities – that is, the development of ways to ‘‘tame the raging beast’’ of information overload in health care. (Berner, Detmer and Simborg, 2005; Blackmore, 2005; Morris *et al.*, 2005; Andriole, 2006). The PACS, as noted earlier, is a powerful starting point for helping all these to happen, especially as it becomes increasingly interconnected with the RIS, HIS, and the Web.

The advantages of placing all patient-related information onto Web-based interoperable data sets that link physicians, hospital medical records, and laboratories may seem self-evident. For most clinics and medical centers, however, this idea remains a distant dream. The health care industry invests only 2% of its gross revenues in information technology, compared with other information-intensive industries (e.g. finance, utilities, and communications), which typically spend about 10% (IT in the healthcare industry, 2005; Bates and Gawande, 2003). Furthermore, clinical information technology systems that exist often cannot readily connect with one another. The impact of miscommunication and misinformation on patient safety is a huge, closely related issue. The Institute of Medicine has estimated that 44 000–98 000 people die annually because of medical errors in the United States, making this the eighth leading cause of death, ahead of acquired immunodeficiency syndrome and breast cancer.

27.13.1
Medical Image Informatics

The amount of visual and other medical data being generated at the levels of molecules and cells, tissues and organs, and individuals and communities is populating huge databases at an ever-increasing and alarming rate. The relatively young discipline of biomedical informatics is beginning to provide tools to search for, identify, select, retrieve, integrate, analyze, model, display, store, communicate, and manipulate all these information (Berner, Detmer and Simborg, 2005; Blackmore 2005; Morris *et al.*, 2005; Andriole, 2006).

Biomedical informatics is perhaps best known for its successes in untangling the vast quantities of data coming from work on the genome. It may well come to play just as important a role in the clinic, such as in attempting to keep a cap on the impending CT slice-overload crisis. CT already produces hundreds of slices per patient, and that is bound to increase (256-slice machines are in the prototype stage), along with the swelling numbers of CT patients. Researchers in the field will continue to develop ways to integrate and scan effectively through the databases needed for clinical care, as they grow in size and complexity.

There are related applications, such as when there occurs a crisis on a spaceship halfway to Mars, or after detonation of a nuclear weapon in a city, where rapid access to the right essential medical information would prove to be invaluable. Here, too, effective medical informatics could make the difference between proper management of the situation and utter chaos.

27.13.2
Image Interpretation by Means of Computer-Aided Detection and Diagnosis

Interpretation of medical images is a complex process that is as much art and acquired intuition as science, and it requires years of training, much of which is provided in the "master and apprentice" style. Computers, on the other hand, excel at highly analytic jobs that involve logic and memory, such as playing chess or calculating just about anything; but even the most clever of them has great difficulty with even the most elementary tasks of everyday life that happen to require judgment or common sense, such as recognizing a face with a new beard or distinguishing a liver with a small lesion from one that is normal but just looks odd.

Still, there have been great strides in machine learning. Neural network algorithms and other species of computer program, for example, can improve their success rate at performing certain tasks by "learning" from past experiences; that is, they use feedback information from humans to build decision rules based on the successes and failures of past decisions.

Machine learning, pattern recognition, and statistical inference programs begin with the detection and characterization of a point, a straight line, or a circle on a blank background. Programs that search for irregularities in mammograms, chest CT sections, texture in US scans of the liver, and other images are greatly more complex and sophisticated, but the basic concepts are much the same. By comparing their analyses to databases compiled from a large number of normal and abnormal images, *CAD* systems can find and draw attention to questionable regions in an image. And a more highly evolved *CAD*x program can even provide a statistical-based tentative diagnosis to the interpreter for consideration.

Studies involving CAD have demonstrated a reduction of false negatives (missed identifications) and an increase in the accuracy of diagnosis of the typical radiologist (Doi *et al.*, 1993). Now that Medicare reimbursement is available, for example, CAD and CADx are being used routinely in many mammography clinics, with both digital and digitized images, as sharp-eyed and untiring second readers to pick out subtle cancers. Clinical successes here and in other applications are tantalizingly suggestive of even more wonderful things to come (Wagner *et al.*, 2002; Giger, Karssemeijer and Armato, 2001; Vyborny and Giger, 2002).

CADx is still a far step from the use of expert artificial-intelligence systems, however, to provide a primary diagnosis from general medical images without human intervention. Such systems would essentially have to mimic the thought processes of physicians interpreting images, including the incorporation of ancillary information such as the patient's history and data acquired by physical examination. These thought processes are largely intuitive, acquired through training and experience, and are almost impossible to articulate clearly. Indeed, complete replacement of the human interpreter of images with computer algorithms is a major challenge that, perhaps, we will be able to meet only partially in the foreseeable future. Fortunately, an expert system for image interpretation does not have to be perfect to be of value – but it must be able to function at least with the effectiveness of the average radiologist.

A voice record, like an electrocardiogram signal, may be viewed as a one-dimensional image. Likewise, verbal order-entry and real-time updating of patient charts by physicians with wireless tablet computers that incorporate dictation and transcription and error-prevention software is a kind of CAD. As many a secretary well knows, the ability of a computer to transcribe voice dictation of radiology reports automatically has undergone major advances. Radiologists in the United States dictate over 130 million reports annually, and computer-based technology reduces the costs associated with manual transcription considerably in a number of centers. Voice patterns are as distinctive as fingerprints, moreover, automatic dictation could enhance security, as well, and reduce unauthorized access of patient information.

It is not clear how quickly CAD and CADx will evolve, but a number of people involved in their development believe that the writing is on the wireless, flat-panel monitor on the wall.

27.13.3
Input of Computer Technologies from Other Fields

The National Aeronautics and Space Administration, the Defense Advanced

Research Projects Agency, the Central Intelligence Agency, the National Security Agency, and other government entities have developed highly sophisticated imaging technologies to assist astronomers in searching the heavens back to the beginning of time, to sharpen the capabilities of spy satellites, to select relevant messages out of billions of phone calls, to keep cruise missiles on track, to allow surveillance of individuals suspected of terrorism or criminal behavior, and to practice medicine in space (Fincke *et al.*, 2005). These agencies look for civilian applications of their nonclassified technologies, and they have already transferred numerous ideas and equipment designs to medicine.

Similarly, much of the technology that makes modern clinical imaging so vibrant and flexible was introduced first in the entertainment and communications industries for applications such as video games and special effects for movies and television. These technologies will evolve, and their spin-offs into medicine will continue to improve the quality and efficiency of patient care.

27.14
Radiation Treatment Planning

Cancer is characterized by an abnormally rapid proliferation of cells, initiated typically by the incorporation or activation of certain kinds of misinformation into some cells' DNA. Treatment normally involves some combination of surgery, chemotherapy, and radiation therapy, in which beams of high-energy photons or electrons from a *linear accelerator* (*linac*) are arranged with their cross-fire region centered on the lesion; in many circumstances, radiation treatments, usually delivered in separate daily sessions spread out over the course of weeks, are more effective at killing tumor cells than healthy tissues.

The successful treatment of a localized cancer with ionizing radiation requires that both the tumor and its microscopic extensions into normal tissue receive a lethal dose of ionizing radiation. At the same time, the dose to nearby normal tissues must be kept low enough to avoid serious complications. The art of the discipline is to find, for any patient, an optimal balance between these (and other) objectives (Figure 27.35). Suppose the probability $P_{\text{tumor}}(D)$ that the cancer will be controlled by killing tumor cells increases with D, the radiation dose. The probability $P_{\text{organ}}(D)$ that life-threatening or otherwise unacceptable radiation complications in a nearby critical organ can be avoided, on the other hand, is a decreasing function of D. The overall probability $P_{\text{patient}}(D)$ that the patient will survive both the disease

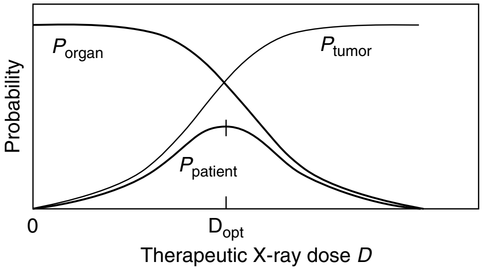

Fig. 27.35 Therapeutic outcome probabilities for cancer patients undergoing X-ray treatment, plotted schematically as functions of radiation dose D. The probability of success (P_{patient}) is the product of the probability that the tumor will be destroyed (P_{tumor}) and the probability (P_{organ}) that the treatment will not cause life-threatening damage to normal tissue.

and the treatment is (Wolbarst, Chin, and Stavrev, 2006)

$$P_{\text{patient}}(D) = P_{\text{tumor}}(D) \times P_{\text{organ}}(D) \tag{18}$$

If the tumor dose is too low, the cancer is likely to recur. At high doses, on the other hand, the therapy itself is likely to cause unacceptable complications, perhaps threatening the patient's life. Between these extremes the physician seeks to maximize the patient's survival of both the cancer and the treatment. If the radiation dose that optimizes the balance between curative treatment and complications is so high that the patient experiences extreme discomfort, or if the objective of the treatment is palliation rather than complete cure, the patient and physician may settle for a lower dose, even though the chance of surviving the cancer is reduced.

Radiation-dose optimization demands accurate delineation of the margins of the cancer and of radiosensitive organs. That, in turn, demands good medical images. Before CT and MRI, cancer localization and treatment planning depended principally upon projection radiography combined with clinical and surgical (if available) information and knowledge of the natural course of the disease. But it is frequently hard to see the cancer clearly with projection radiography, and to transcribe the essential information into a three-dimensional representation of the region of concern for designing a treatment plan, that is, for determining the orientation, relative weighting and shaping, and so on, of each of the radiation beams so as to give good coverage of the target while adequately sparing critical structures.

CT and MRI have largely overcome the problem of delineating anatomy, and PET/CT fusion machines are increasingly finding application in tumor localization; indeed, many treatment centers have a CT unit dedicated to treatment planning and monitoring. These modalities can provide a 3D picture of the patient's anatomy, which includes the site of the cancer and its immediate surroundings, the critical organs, and the body contour. The data, being in digital form, are entered immediately into the treatment-planning computer and, once an acceptable plan has been generated, it can be shipped semi-automatically to the linac controls.

Each treatment session must be carefully monitored to ensure proper setup and, in particular, correct positioning with respect to the treatment machine; when various forms of DR and flat-panel CT are used for this purpose, as is increasingly the case, the process is called *image-guided radiation therapy* (*IGRT*). Installing, managing, and monitoring are all the responsibilities of the medical physicist.

Most radiation treatments involve the use of multiple fixed radiation fields converging on the cancer from different directions. In this manner, the dose is concentrated in the cancer while much lower doses are administered to surrounding normal tissues. In some cases, a better dose distribution can be achieved by rotating the gantry of the treatment machine continually or in steps around the patient during treatment so that only the cancer is constantly in the path of the X-ray beam. One complication is that cancers are almost always asymmetrical. The size of the X-ray beam must therefore be expanded or contracted from one position of the gantry to the next, and its shape can be varied by means of *multileaf collimator*. This approach to external beam X-ray treatment is termed *conformal therapy*. The dose distribution can be improved further by simultaneously varying the dose rate. This *intensity-modulated radiation therapy* (*IMRT*) is becoming the norm at advanced centers for treatment of some forms of disease. It requires detailed three-dimensional knowledge of the anatomy of the irradiated tissue, together with exquisite computerized control of the linac.

27.15
Conclusions

Medical imaging is perhaps *the* prototypical example of "high-tech" medicine. Its successes illustrate the contributions of medical physicists and engineers working with physicians to improve the health and welfare of patients. But there is still much to be done, especially in refining the growing convergence of imaging with therapeutic techniques to cure disease and prevent the debilitating effects of injuries.

Physicists have played an instrumental role in the evolution of all branches of medical imaging. The American Association of Physicists in Medicine, a member society of the American Institute of Physics, includes over 6000 members. Graduate and postdoctoral training programs in medical physics increasingly attract bright young physicists, and among them are many of those who will bring about the advances yet to be made.

Glossary

Absorbed dose: the ionizing energy absorbed per unit mass of an irradiated material.

Active-matrix flat-panel imager (AMFPI): a solid-state semiconductor-based X-ray image receptor consisting of a matrix of hundreds of thousands or millions of separate X-ray detectors. Analogous to a CCD optical camera and used in digital radiography.

Artificial intelligence: the property of a machine capable of logical, systematic reasoning, by which it can learn functions normally associated with human intelligence.

Bremsstrahlung: the X-ray radiation (*Strahlung*) caused by the sudden "braking" (*Brems*) of electrons by a solid.

Acoustic impedance: a characteristic of a tissue, defined as $Z = (\rho/K)^{1/2}$, where ρ is its physical density and K its compressibility, a measure of its elasticity. At a boundary between two tissues, their Z values together determine the amount of reflectivity of ultrasound that will occur there.

Contrast agent: a substance placed in veins, arteries, the GI tract, and so on, that enhances the image contrast.

Characteristic X-rays: electromagnetic radiation emitted as a result of rearrangements of the electrons in the inner shells of atoms following a disruption; the spectrum consists of lines whose wavelengths depend only on the element concerned and the energy levels involved.

Charge-coupled device (CCD): a semiconductor-based electronic optical camera, comprising hundreds of thousands or millions of pixel-size photodetectors. The charges stored in the separate pixels are read out electronically row-by-row and immediately digitized.

Compton effect: scattering of a high-energy X-ray photon by an orbital electron of an atom, which is ionized in the process. A lower-energy Compton scatter photon is created, and the rest of the incident energy is transferred to the liberated electron as kinetic energy.

Computed radiography (CR): an X-ray imaging method in which a pattern of X-rays transmitted through the patient is recorded on a solid-state photostimulable phosphor (PSP) imaging plate (IP). The plate is read by a scanning laser beam that stimulates the emission of light of another color.

Computed tomography (CT): the process of producing high-contrast X-ray images of thin (usually transverse) slices of tissues within the body, with visual interference from over- and underlying tissues removed.

Digital radiography (DR): An X-ray imaging method in which a pattern of X-rays transmitted through the patient is recorded in digital form on an AMFPI, which is read in a manner similar to that of a CCD.

Doppler effect: the change in frequency of ultrasound waves reflected from a moving object.

Dynamic range: the amplitude range of signals that can be detected by an imaging system without distortion or disappearance into background noise.

Expert system: a computer system composed of algorithms that perform a specialized, usually difficult professional task at the level of (or sometimes beyond the level of) a human expert. For the basis for computer-aided detection (CAD) and computer-aided diagnosis (CADx), see *Artificial Intelligence*.

Fluoroscopy: use of an image intensifier or AMFPI to image receptor for obtaining dynamic, real-time X-ray images.

Gantry: an extended mechanical structure holding elements of an imaging system.

Image intensifier: an electronic vacuum-tube device that receives a life-sized, pale X-ray pattern from a patient and produces a bright, small-area optical image suitable for input into a TV, CCD, or other optical camera.

Intensifying screen: a luminescent panel within a screen-film cassette, in which incident X-rays cause the emission of points of light that, in turn, exposes a photographic film.

Linear attenuation coefficient: fractional rate of attenuation (by coherent scattering, the photoelectric effect, and Compton scattering events) of X-rays in an absorbing medium.

Mammography: X-ray imaging of the breast.

Modulation transfer function (MTF): a measure of the ability of an imaging system to register fine detail present in the object.

Neural network: an information-processing system that utilizes a large number of simple modules, to transmit, receive, and store digital information.

Photoelectric effect: the liberation of an inner-orbital atomic electron by a photon of electromagnetic radiation.

Pixel: the smallest area that can be displayed by a digital imaging system; a larger image is composed of many pixels.

Projection radiography: the technique of producing a common X-ray image. Transmits a uniform beam of X-rays through the body, thereby embedding an X-ray shadow in it that exposes a cassette and the photographic film within. The image results from variations in thickness, density, and chemical composition of the tissues, and depends on the effective energy of the X-ray beam.

Real-time display: a display that is able to follow changes in the object that occur at physiologic rates.

Spin–lattice relaxation: the process, parameterized by T1, by which nuclear spins (in MRI, primarily protons in tissue water molecules) that are disrupted return to their equilibrium condition lying along an external magnetic field.

Spin–spin relaxation: the process, parameterized by T2, by which nuclear spins precessing in a magnetic field exchange energy with other nuclear spins and thereby dephase.

Subtraction: temporal subtraction is a process in which the information content of two images, one obtained before and one after an intentional change (e.g. injection of contrast medium into a vein), are subtracted from one another; all that remains is the image of that which changed, with no distractions from irrelevant tissue background. Other forms of subtraction, such as with a rapid change of kVp, can serve other purposes.

Teleradiology: the transmission of digital medical images electronically over long distances.

References

Note to readers. The references in this paper comprise a mix of books, recent reviews in journals, and interesting, representative research articles. In-depth discussion of the standard and newer imaging technologies, and more extensive bibliographies, may be found in Hendee and Ritenour (2002); Wolbarst (2005); Wolbarst, Zamenhof and Hendee (2006). Nontechnical, introductory educational information on imaging for patients and the general public is presented in Wolbarst (1999) and at: www.nlm.nih.gov/medlineplus/diagnostic imaging.html.

Andriole, K.P. (2006) Overview of Medical Imaging Informatics, in *Advances in Medical Physics*, Medical Physic Publishing, Madison, WI, pp. 201–228.

AIUM (2003) *Bioeffects Committee: Evaluation of Research Reports: Ultrasound Bioeffects Literature Reviews (1992–2003)*, American Institute of Ultrasound in Medicine, Laurel, MD. www.aium.org.

Allen, M.J. and Meade, T.J. (2004) Magnetic resonance contrast agents for medical and molecular imaging. *Met. Ions Biol. Syst.*, **42**, 1–38.

Antonuk, L.E. (2004) a-Si:H TFT-based active matrix flat-panel imagers for medical X-ray applications, in *Materials and Processes: Amorphous Silicon Thin Film Transistors*, Vol. 1 (ed. K. Kuo), Kluwer Academic Publishers, Boston, MA, pp. 395–484.

Bacharach, S.L. (2004) The new generation positron emission tomography/computed tomography scanners: implications for cardiac imaging, in *Nuclear Cardiology: State of the Art and Future Directions*, 3rd edn (eds B. Zaret and G. Beller), Elsevier, New York, pp. 141–52.

Badano, A. (2004) AAPM/RSNA tutorial on equipment selection: PACS equipment overview: Display systems. *Radiographics*, **24**, 879–89.

Badano, A., Flynn, M.J., and Kanicki, J. (2004) *High-Fidelity Medical Imaging Displays*, SPIE Press, Bellingham, WA.

Bankman, I. (ed.) (2000). *Handbook of Medical Image Processing*, Academic Press, New York.

Barkley, G.L. (2004). Controversies in neurophysiology. MEG is superior to EEG in localization of interictal epileptiform activity: pro. *Clin. Neurophysiol.*, **115**, 1001–9.

Barrett, H.H. and Myers, K.J. (2004) *Foundations of Image Science*, Wiley-Interscience, Hoboken, NJ.

Bates, D.W. and Gawande, A.A. (2003) Improving safety with information technology. *N. Engl. J. Med.*, **348**, 2526–34.

Baumgartner, C. (2004) Controversies in clinical neurophysiology. MEG is superior to EEG in the localization of interictal epileptiform activity: con. *Clin. Neurophysiol.*, **115**, 1010–20.

Becquerel, H. (1896) Sur les radiation emises par phosphorescence. *Compt. Rend.*, **122**, 420.

Berliner, L.J. (ed.) (2003) *In Vivo EPR (ESR): Theory and Application*, Kluwer Academic/Plenum Publishers, New York.

Berner, E.S., Detmer, D.E., and Simborg, D. (2005) Will the wave finally break? A brief view of the adoption of electronic medical records in the United States. *J. Am. Med. Inform. Assoc.*, **12**, 3–7.

Berry, E., Arridge, S.R., Maidment, A.D.A., Claridge, E.I., Noble, (2003) Medical imaging research symposium: modalities using non-ionizing electromagnetic radiation. *Med. Phys.*, **30**, 1536–42.

Blackmore, C.C. (2005) Radiology and information technology: a health services research perspective. *Acad. Radiol.*, **12**, 397–98.

Brateman, L. and Karellas, A. (2006) Mammography and other breast imaging techniques, in *Advances in Medical Physics - 2006* (eds A.B. Wolbarst, R.G. Zamenhof, and W.R. Hendee), Medical Physics Publishing, Madison, pp. 25–57.

Breslin, T.M., Xu, F., Palmer, G.M., Zhu, C., Gilchrist, K.W., and Ramanujam N. (2004) Autofluorescence and diffuse reflectance properties of malignant and benign breast tissues. *Ann. Surg. Oncol.*, **11**, 65–70.

Britz-Cunningham, S.H. and Adelstein, S.J. (2003) Molecular targeting with radionuclides: state of the science. *J. Nucl. Med.*, **44**, 1945–61.

Bulte, J.W., Kostura, L., and Mackay, A. (2005) FerideX-labeled mesenchymal stem cells: cellular differentiation and MR assessment in a canine myocardial infarction model. *Acad. Radiol.*, **12**, 2–6.

Bushberg, J.T., Seibert, J.A., Leidholdt, E.M. and Boone, J.M. (2002) *The Essential Physics of Medical Imaging*, 2nd edn., Lippincott Williams & Wilkins, Philadelphia, PA.

Bushong, S.C. (1996) *Magnetic Resonance Imaging*, 2nd edn., Mosby, St. Louis.

Cademartiri, F., Lucchichenti, G., van Der Lugt, A., Pavone, P., Pattynama, P.M., de Feyter, P.J., and Krestin, G.P. (2004) Sixteenrow multislice computed tomography: basic concepts, protocols, and enhanced clinical applications. *Semin. Ultrasound CT MR*, **25**, 2–16.

Carrino, J.A. (2003) Digital imaging overview. *Semin. Roentgenol.*, **38**, 200–15.

Chandra, R. (2004) *Nuclear Medicine Physics*, 6th edn., Lippincott Williams & Wilkins, Philadelphia.

Cherry, S.R., Sorenson, J.A., and Phelps, M.E. (2003) *Physics in Nuclear Medicine*, 3rd edn., W.B. Saunders, Philadelphia.

Choe, R., Corlu, A., Lee, K., Durduran, T., Konecky, S.D., Grosicka-Koptyra, M., Arridge, S.R., Czerniecki, B.J., Fraker, D.L., DeMichele, A., Chance, B., Rosen, M.A., and Yodh, A.G. (2005) Diffuse optical tomography of breast cancer during neoadjuvant chemotherapy: a case study with comparison to MRI. *Med. Phys.*, **32**, 1128–39.

Ciocan, R. and Jiang, H. (2004) Model-based microwave image reconstruction: simulations and experiments. *Med. Phys.*, **31**, 3231–41.

Cormack, A. (1963) Representation of a function by its line integrals, with some radiological applications. *J. Appl. Phys.*, **34**, 2722–27.

Curra, F.P. and Crum, L.A. (2003) Therapeutic ultrasound: surgery and drug delivery. *Acoust. Sci. Technol.*, **24**, 343–48.

Czernin, J. (ed.) (2004) PET/CT imaging function and structure. *J. Nucl. Med.*, **45**(Suppl),, 1S–103S.

Desser, T.S. and Jeffrey, R.B. (2001) Tissue harmonic imaging techniques: physical principles and clinical applications. *Semin. Ultrasound CT MR*, **22**, 1–10.

Diamantopoulos, L. (2003) Arterial wall thermography. *J. Interv. Cardiol.*, **16**, 261–66.

Digital Imaging and Communications in Medicine (DICOM) (2003): DICOM PS 3 – 2003, National Electrical Manufacturers Association, NEMA, Rosslyn, VA. http://medical.nema.org/dicom/2003.html, http://medical.nema.org/dicom/2003.html.

Dobbins, J.T. III and Godfrey, D.J. (2003) Digital X-ray tomosynthesis: current state of the art and clinical potential. *Phys. Med. Biol.*, **48**(19), R65–R106.

Dodd, L.E., Wagner, R.F., Armato, S.G. III, Mc Nitt-Gray, M.F., Beiden, S., Chan, H.P., Gur, D., McLennan, G., Metz, C.E., Pet rick, N., Sahiner, B., Sayre, J., Lung Image Database Consortium Research Group (2004) Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography:

contemporary research topics relevant to the lung image database consortium (LIDC). *Acad. Radiol.*, **11**, 462–75.

Doi, K., Giger, M., Nishikawa, R., MacMahon, H., Schmidt, R. (1993) Artificial intelligence and neural networks in radiology: application to computer-aided diagnostic schemes, in *Digital Imaging* (eds W.R. Hendee and J. Trueblood), Medical Physics Publishing, Madison, WI, pp. 301–22.

Duric, N., Littrup, P., and Babkin, A. (2005) Development of ultrasound tomography for breast imaging: technical assessment. *Med. Phys.*, **32**, 1375–86.

Edler, I. and Hertz, C.H. (1954) The use of ultrasonic reflectoscope for the continuous recording of movements of heart walls. *Kungl. Fysiorgr. sällsk. I Lund förhandl.*, **24**, 5.

Elas, M., Williams, B.B., Parasca, A., Mailer, C., Pelizzari,C.A., Lewis, M.A., River, J.N., Karczmar, G.S., Barth, E.D., and Halpern, H.J. (2003) Quantitative tumor oxymetric images from 4D electron paramagnetic resonance imaging (EPRI): methodology and comparison with blood oxygen level-dependent (BOLD) MRI. *Magn. Reson. Med.*, **49**, 682–91.

Elkington, A.G., Gatehouse, P.D., and Cannell, T.M. (2005) Comparison of hybrid echoplanar imaging and FLASH myocardial perfusion cardiovascular MR imaging. *Radiology*, **235**, 237–43.

Faisan, S., Thoraval, L., Armspach, J.P., Foucher, J.R., Metz-Lutz, M.N., and Heitz, F. (2005) Hidden Markov event sequence models: toward unsupervised functional MRI brain mapping. *Acad. Radiol.*, **12**, 25–36.

Fayed, N. and Modrego, P.J. (2005) Comparative study of cerebral white matter in autism and attention-deficit/hyperactivity disorder by means of magnetic resonance spectroscopy. *Acad. Radiol.*, **12**, 566–69.

Fincke, E.M., Padalka, G., Lee, D., van Holsbeeck, M., Sargsyan, A.E., Hamilton, D.R., Martin, D., Melton, S.L., McFarlin, K., and Dulchavsky, S.A. (2005) Evaluation of shoulder integrity in space: first report of musculoskeletal US on the International Space Station. *Radiology*, **234**(2), 319–22.

Fischmann, A., Siegmann, K.C., Wersebe, A., Claussen, C.D., and Muller-Schimpfle (2005) Comparison of full-field digital mammography and film-screen mammography: image quality and lesion detection. *Br. J. Radiol.*, **78**, 312–15.

Flohr, T.G., Schaller, S., Stierstorfer, K., Bruder, H., Ohnesorge, B.M., and Schoepf, U.J. (2005) Multi-detector row CT systems and image-reconstruction techniques. *Radiology*, **235**, 756–73.

Fohr, T.G., Cody, D.D., and McCollough, C.H. (2006) Computed tomography, in *Advances in Medical Physics - 2006* (eds A.B. Wolbarst, R.G. Zamenhof, and W.R. Hendee), Medical Physics Publishing, Madison, pp. 59–100.

Fowlkes, J.B. and Wolbarst, A.B. (2006) Medical ultrasound imaging, in *Advances in Medical Physics - 2006* (eds A.B. Wolbarst, R.G. Zamenhof, and W.R. Hendee), Medical Physics Publishing, Madison, pp. 151–80.

Gao, X., Cui, T., Levenson, R.M., Chung, L.W., and Nie, S. (2004) In vivo cancer targeting and imaging with semiconductor quantum dots. *Nat. Biotechnol.*, **22**, 969–76.

Georgopoulos, A.P. *et al.* (2007) Synchronous neural interactions assessed by magnetoencephalography: a functional biomarker for brain disorders. *J. Neural Eng.*, **4**, 349–55.

Giger, M.L., Karssemeijer, N., and Armato, S.G. III (eds) (2001) Computer-aided diagnosis in medical imaging [editorial]. *IEEE Trans. Med. Imaging*, **20**, 11275–79.

Godavarty, A., Sevick-Muraca, E.M., and Eppstein, M.J. (2005) Three-dimensional fluorescence lifetime tomography. *Med. Phys.*, **32**, 992–1000.

Godik, E.E. and Gulyaev, Y.Y. (1991) Functional imaging of the human body. *IEEE Eng. Med. Biol. Mag.*, **10**, 21–29.

Goldman, L.W. and Yester, M.V. (eds) (2004). Specifications, performance evaluation, and quality assurance of radiographic and fluoroscopic systems in the digital era. *AAPM Medical Physics Monograph*, Vol. 30, Medical Physics Publishing, Madison, WI.

Greenleaf, J.F. (ed.) (1986) *Tissue Characterization with Ultrasound*, Vols I and II, CRC, Orlando, FL.

Grimnes, S. and Martinsen, Ø.G. (2000) *Bioimpedance and Bioelectricity Basics*, Academic Press, San Diego, CA.

Haake, E.M., Brown, R.W., Thompson, M.R., and Venkatesan, R. (1999) *Magnetic*

Resonance Imaging: Physical Principles and Sequence Design, John Wiley & Sons, Inc, New York.

Hallett, M. and Chokoverty, S. (eds) (2005) *Magnetic Stimulation in Clinical Neurophysiology*, 2nd edn., Elsevier Butterworth Heinemann Health, Philadelphia.

Harvey, C.J., Pilcher, J.M., Eckersley, R.J., Blomley, M.J., and Cosgrove, D.O. (2002) Advances in ultrasound. *Clin. Radiol.*, **57**, 157–77.

Hendee, W.R., Ibbott, G., Hendee, E. (2005) *Radiation Therapy Physics*, 3rd edn., John Wiley & Sons, New York.

Hendee, W.R. and Ritenour, E.R. (2002). *Medical Imaging Physics*, 4th edn., John Wiley & Sons, New York.

Higgins, D.M., Ridgway, J.P., Radjenovic, A., Sivananthan, U.M., and Smith, M.A. (2005) T1 measurement using a short acquisition period for quantitative cardiac applications. *Med. Phys.*, **32**, 1738–46.

Hohmann, J., Albrecht, T., Hoffmann, C.W., and Wolf, K.J. (2003) Ultrasonographic detection of focal liver lesions: increased sensitivity and specificity with microbubble contrast agents. *Eur. J. Radiol.*, **46**, 147–59.

Holodny, A.I., Gor, D.M., Watts, R., Gutin, P.H., and Ulu, A.M. (2005) Diffusion-tensor MR tractography of somatotopic organization of corticospinal tracts in the internal capsule: initial anatomic results in contradistinction to prior reports. *Radiology*, **234**, 649–53.

Hounsfield, G. (1973) Computerized transverse axial scanning (tomography). I. Description of system. *Br. J. Radiol.*, **46**, 1016–22.

Hu, X. and Norris, D.G. (2004). Advances in high-field magnetic resonance imaging. *Annu. Rev. Biomed. Eng.*, **6**, 157–84.

Huang, H.K. (2004) *PACS and Imaging Informatics: Basic Principles and Applications*, Wiley, Hoboken.

Imperiale, T.F. (2005) Can computed tomographic colonoscopy become a 'good' screening test? *Ann. Intern. Med.*, **142**, 669–70.

Ireland, R.H., Tozer, J.C., Barker, A.T., and Barber, D.C. (2004) Towards magnetic detection electrical impedance tomography: data acquisition and image reconstruction of current density in phantoms and in vivo. *Physiol. Meas.*, **25**, 775–96.

(2005) IT in the healthcare industry. *Am. Econ.*, Apr 30, 65–67.

Jahng, G.H., Stables, L., and Ebel, A. (2005). Sensitive and fast T1 mapping based on two inversion recovery images and a reference image. *Med. Phys.*, **32**, 1524–28.

Jiang, H. (2003) *Computed Tomography: Principle, Designs, Artifacts and Recent Advances*, International Society for Optical Engineering (SPIE) Press, Bellingham, WA.

Kak, A., Slaney, M. (1988) *Principles of Computerized Tomographic Imaging*, IEEE Press, New York.

Kalra, M.K., Maher, M.M., Toth, T.L., Hamberg, L.M., Blake, M.A., Shepard, J.A., and Saini, S. (2004) Strategies for CT radiation dose optimization. *Radiology*, **230**, 619–28.

Kamitani, Y. and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.*, **8**, 679–85.

Kapur, A., Carson, P.L., Eberhard, J., Goodsitt, M.M., Thomenius, K., Lokhand walla, M., Buckley, D., Roubidoux, M.A., Helvie, M.A., Booi, R.C., LeCarpentier, G.L., Erkamp, R.O., Chan, H.P., Fowlkes, J.B., Thomas, J.A., and Landberg, C.E. (2004) Combination of digital mammography with semiautomated 3D breast ultrasound. *Technol. Cancer Res. Treat.*, **3**, 325–34.

Kasai, C., Namekawa, K., Koyano, A., and Omoto, R. (1985) Real-time two dimensional blood flow imaging using an autocorrelation technique. *IEEE Trans. Sonics Ultrason.*, **SU-32**, 458–64.

Katsevich, A. (2004) An improved exact filtered backprojection algorithm for spiral computed tomography. *Adv. Appl. Math.*, **32**, 681–97.

Khaw, K.-T., Reeve, J., Luben, R., Bingham, S., Welch, A., Wareham, N., Oakes, S., and Day, N. (2004) Prediction of total and hip fracture risk in men and women by quantitative ultrasound of the calcaneus: EPIC-Norfolk prospective population study. *Lancet*, **363**, 197–202.

Kriege, M., Brekelmans, C.T., Boetes, C., Besnard, P.E., Zonderland, H.M., Obdeijn, I.M., Manoliu, R.A., Kok, T., Peterse, H., Tilanus-Linthorst, M.M., Muller, S.H., Meijer, S., Oosterwijk, J.C., Beex, L.V., Tollenaar, R.A., de Koning, H.J., Rutgers, E.J., Klijn, J.G., Magnetic Resonance Imaging Screening Study Group

(2004) Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition. *N. Engl. J. Med.*, **351**, 427–37.

Krouskop, T.A., Dougherty, D.R., and Vinson (Levinson), F.S. (1987) A pulsed doppler ultrasound system for making non-invasive measurement of the mechanical properties of soft tissue. *J. Rehabil. Res. Dev.*, **24**, 1–8.

Lee, W. (2003) 3D fetal ultrasonography. *Clin. Obstet. Gynecol.*, **46**, 850–67.

Lesser, R.P. (2004) MEG: good enough. *Clin. Neurophysiol.*, **115**, 995–97.

Liang, Z.P. and Lauterbur, P.C. (2000) *Principles of Magnetic Resonance Imaging: A Signal Processing Perspective*, IEEE Press, New York.

Liberman, L. (2004) Editorial: breast cancer screening with MRI – What are the data for patients at high risk? *N. Engl. J. Med.*, **351**, 497–500.

Liebgott, T., Li, H., Deng, Y., and Zweier, J.L. (2003) Proton electron double resonance imaging (PEDRI) of the isolated beating rat heart. *Magn. Reson. Med.*, **50**, 391–99.

Lubawy, C. and Ramanujam, N. (2004) Endoscopically compatible near-infrared photon migration probe. *Opt. Lett.*, **29**, 2022–24.

Mahnken, A.H., Seyfarth, T., Flohr, T. *et al.* (2005) *Flat-panel Detector CT for Assessment of Coronary Artery Stents: Comparison with 16-Slice Spiral CT*. Presented at the Annual Meeting of the European Congress of Radiology, March 4–8, 2005, Vienna, Austria.

Massoud, T.F. and Gambhir, S.S. (2003) Molecular imaging in living subjects: seeing fundamental biological processes in a new light. *Genes Dev.*, **17**, 545–80.

Michalet, X., Bentolila, L.A., and Weiss, S. (2006) Molecular imaging, in *Advances in Medical Physics - 2006* (eds A.B. Wolbarst, R.G. Zamenhof, and W.R. Hendee), Medical Physics Publishing, Madison, pp. 181–200.

Michalet, X., Pinaud, F.F., Bentolila, L.A., Tsay, J.M., Doose, S., Li, J.J., Sundaresan, G., Wu, A.M., Gambhir, S.S., and Weiss, S. (2005) Quantum dots for live cells and in vivo imaging, diagnostics and beyond. *Science*, **307**, 538–44.

Miller, A.P. and Nanda, C. (2004) Contrast echocardiography: new agents. *Ultrasound Med. Biol.*, **30**, 425–34.

Monnin, P., Gutierrez, D., Bulling, S., Lepori, D., Valley, J.-F., and Verdun, F.R. (2005) Performance comparison of an active matrix flat panel imager, computed radiography system, and a screen-film system at four standard radiation qualities. *Med. Phys.*, **32**, 343–50.

Morgan, N.Y., English, S., and Chen, W. (2005) Real time in vivo non-invasive optical imaging using near-infrared fluorescent quantum dots. *Acad. Radiol.*, **12**, 313–23.

Mori, S., Endo, M., and Obata, T. (2005) Clinical potentials of the prototype 256-detector row CT-scanner. *Acad. Radiol.*, **12**, 148–54.

Morris, R.W., Bean, C.A., Farber, G.K., Gallahan, D., Jakobsson, E., Liu, Y., Lyster, P.M., Peng, G.C., Roberts, F.S., Twery, M., Whitmarsh, J., and Skinner, K. (2005) Digital biology: an emerging and promising discipline. *Trends Biotechnol.*, **23**, 113–17.

Nehmeh, S.A., Erdi, Y.E., and Pan, T. (2004) Four-dimensional (4D) PET/CT imaging of the thorax. *Med. Phys.*, **31**, 3179–86.

Newhouse, V.L. (1993) Imaging techniques, biomedical, in *Encyclopedia of Applied Physics*, Vol. 7, VCH, New York.

Nitz, W.R. (2002) Fast and ultrafast non-echo-planar MR imaging techniques. *Eur. Radiol.*, **12**, 2866–82.

Okada, T., Yamada, H., Ito, H., Yonekura, Y., and Sadato, N. (2005) Magnetic field strength increase yields significantly greater contrast-to-noise ratio increase: measured using BOLD contrast in the primary visual area. *Acad. Radiol.*, **12**, 142–47.

Oralkan, O., Ergun, A.S., Cheng, C.H., Johnson, J.A., Karaman, M., Lee, T.H., and Khuri-Yakub, B.T. (2003) Volumetric ultrasound imaging using 2-D CMUT arrays. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, **50**, 1581–94.

Owunwanue, A., Patel, M., and Sadek, S. (1995) *The Handbook of Radiopharmaceuticals*, Chapman & Hall, London.

Park, M.A., Moore, S.C., and Kijewski, M.F. (2005) Brain SPECT with short focal-length cone-beam collimation. *Med. Phys.*, **32**, 2236–44.

Pevsner, A., Nehmeh, S.A., Humm, J.L., Mageras, G.S., and Erdi, Y.E. (2005) Effect of motion on tracer activity determination in CT attenuation corrected PET images: a lung phantom study. *Med. Phys.*, **32**, 2358–62.

Phan, K.L., Magalhaes, A., Ziemlewicz, T.J., Fitzgerald, D.A., Green, C., and Smith, W. (2005) Neural correlates of telling lies: a functional magnetic resonance imaging study at 4 Tesla. *Acad. Radiol.*, **12**, 164–72.

Pisano, E.D., Gatsonis, C., Hendrick, E., Yaffe, M., Baum, J.K., Acharyya, S., Conant,E.F., Fajardo,L.L., Bassett, L., D'Orsi, C., Jong, R., Rebner, M., Digital Mammographic Imaging Screening Trial (DMIST) Investigators Group (2005) Diagnostic performance of digital versus film mammography for breast-cancer screening. *N. Engl. J. Med.*, **353**(17), 1773–83.

Pisano, E.D. and Yaffe, M.J. (2005) Digital mammography. *Radiology*, **234**, 353–62.

Poplack, S.P., Paulsen, K.D., Hartov, A., Meaney, P.M., Pogue, B.W., Tosteson, T.D., Grove, M.R., Soho, S.K., and Wells, W.A. (2004) Electromagnetic breast imaging: average tissue property values in women with negative clinical findings. *Radiology*, **231**, 571–80.

Price, R.R., Allison, J., Massoth, R.J., Clarke, G.D., and Drost, D.J. (2002) Practical aspects of functional MRI (NMP Task Group #8). *Med. Phys.*, **29**, 1892–912.

Radon, J. (1917) Über die Bestimmung von Funktionen durch ihre Integralwerte laengs gewisser Mannigfaltigkeiten (on the determination of functions from their integrals along certain manifolds). *Ber. saechs. Akad. Wiss. Leipzig, Math.-Phys. Kl.*, **69**, 262–77.

Richard, S., Siewerdsen, J.H., Jaffray, D.A., Moseley, D.J., and Bakhtiar, B. (2005) Generalized DQE analysis of radiographic and dual-energy imaging using flat-panel detectors. *Med. Phys.*, **32**, 1397–413.

Rietzel, E., Pan, T., and Chen, G.T.Y. (2005) Four-dimensional computed tomography: image formation and clinical protocol. *Med. Phys.*, **32**, 874–89.

Röntgen, W. (1895) Über eine neue Art von Strahlen (vorläufige Mitteilung). *Sitzungs ber. physik.-med. Ges. Würzburg*, **9**, 132.

Rosen, H.J., Ojemann, J.G., Ollinger, J.M., and Petersen, S.E. (2000) Comparison of brain activation during word retrieval done silently and aloud using fMRI. *Brain Cogn.*, **42**(2), 201–17.

Saha, G.B. (2001) *Physics and Radiobiology of Nuclear Medicine*, 2nd edn., Springer-Verlag, New York.

Saini, S. (2004) Advances in multidetector CT. *J. Comput. Assist. Tomogr.*, **28**(Suppl 1), S1.

Samei, E., Badano, A., and Chakraborty, D. (2005) Assessment of display performance for medical imaging systems: executive summary of AAPM TG18 report. *Med. Phys.*, **32**, 1205–25.

Sandrin, L., Tanter, M., Catheline, S., and Fink, M. (2002) Shear modulus imaging with 2-D transient elastography. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, **49**, 426–35.

Schardt, P., Deuringer, J., Freudenberger, J., Hell, E., Knüpfer, W., Mattern, D., and Schild, M. (2004) New X-ray tube performance in computed tomography by introducing the rotating envelope tube technology. *Med. Phys.*, **31**, 2699–706.

Schelbert, H.R., Inubushi, M., and Ross, R.S. (2003) PET imaging in small animals. *J. Nucl. Cardiol.*, **10**, 513–20.

Schultz, J., Mrksich, M., Bhatia, S. *et al.* (2004) WTEC Panel Report on International Research and Development in Biosensing. http://www.wtec.org/biosensing/biosen sing-final.pdf (August 2004).

Shah, K. and Weissleder, R. (2005) Molecular optical imaging: applications leading to the development of present day therapeutics. *NeuroRx*, **2**, 215–25.

Shechter, G., Köhler, T., Altman, A., and Proksa, R. (2004) The frequency split method for helical cone-beam reconstruction. *Med. Phys.*, **31**, 2230–36.

Shikhaliev, P.M., Xu, T., and Molloi, S. (2005) Photon counting computed tomography: concept and initial results. *Med. Phys.*, **32**, 427–36.

Siewerdsen, J.H., Moseley, D.J., Bakhtiar, B., Richard, S., and Jaffray, D.A. (2004) The influence of antiscatter grids on soft-tissue detectability in cone-beam computed tomography with flat-panel detectors. *Med. Phys.*, **31**, 3506–20.

Silva, G.A. (2004) Introduction to nanotechnology and its applications to medicine. *Surg. Neurol.*, **61**, 216–20. See also: http://nano.cancer.gov.

Slavin, G.S. and Bluemke, D.A. (2005) Spatial and temporal resolution in cardiovascular

MR imaging: review and recommendations. *Radiology*, **234**, 330–38.

Smith, E.M. (ed.) (2006) PACS2006 : 6th Annual Conference: Digital Healthcare Information & Management Systems (DHIMS). http://www.urmc.rochester.edu/pacs2006.

Smith, S.W., Chu, K., Idriss, S., Ivancevich, N.M., Light, E.D., and Wolf, P.D. (2004) Feasibility study: real-time 3-D ultrasound imaging of the brain. *Ultrasound Med. Biol.*, **30**, 1365–71.

Sonka, M. and Fitzpatrick, J.M. (eds) (2000). *Handbook of Medical Imaging Volume 2. Medical Image Processing and Analysis*, SPIE Press, Bellingham, WA.

Sonke, J.J., Zijp, L., Remeijer, P., and van Herk, M. (2005) Respiratory cor related cone beam CT. *Med. Phys.*, **32**, 1176–86.

Stanley, J.A. (2002) In vivo magnetic resonance spectroscopy and its application to neuropsychiatric disorders. *Can. J. Psychiatry*, **47**, 315–26.

Thrall, J.H. (2004) Nanotechnology and medicine. *Radiology*, **230**, 315–18.

Thrall, J.H. (2005) Reinventing radiology in the digital age. Part I. The all-digital department. *Radiology*, **236**(2), 382–85.

Turner, J.E. (2007) Atoms, Radiation, and Radiation Protection. Wiley-Interscience, Somerset, NJ.

Valk, P.E., Bailey, D.L., Townsend, D.W., and Maisey, M.N. (2003) *Positron Emission Tomography: Basic Science and Clinical Practice*, Springer-Verlag, London.

van den Brink, J.S., Watanabe, Y., Kuhl, C.K., Chung, T., Muthupillai, R., Van Cauteren, M., Yamada, K., Dymarkowski, S., Bogaert, J., Maki, J.H., Matos, C., Casselman, J.W., and Hoogeveen, R.M. (2003) Implications of SENSE MR in routine clinical practice. *Eur. J. Radiol.*, **46**, 3–27.

Vetter, R.J. (2005) Medical health physics: a review. *Health Phys.*, **88**, 653–64.

Vo-Dinh, T. (ed.) (2003) *Biomedical Photonics Handbook: Monograph PM125*. SPIE Press, Bellingham, WA.

Vyborny, C.J. and Giger, M.L. (eds) (2002) Computer-aided detection and diagnosis. *Semin. Breast Dis.*, **5**, 180–258.

Wagner, R.F., Beiden, S.V., Campbell, G., Metz, C.E., and Sacks, W.M. (2002) Assessment of medical imaging and computerassist systems: lessons from recent experience. *Acad. Radiol.*, **9**, 1264–77.

Walker, G.C., Berry, E., Smye, S.W., Zinov'ev, N.N., Fitzgerald, A.J., Miles, R.E., Chamberlain, M., and Smith, M.A. (2004) Modelling the propagation of terahertz radiation through a tissue simulating phantom. *Phys. Med. Biol.*, **49**, 1853–64.

Wang, K. and Mittleman, D.M. (2004) Metal wires for terahertz wave guiding. *Nature*, **432**, 376–79.

Wang, X.F., Deng, Y.B., Nanda, N.C., Deng, J., Miller, A.P., and Xie, M.X. (2003) Live three-dimensional echocardiography: imaging principles and clinical application. *Echocardiography*, **20**, 593–604.

Wang, G., Zhao, S., and Yu, H. (2005) Design, analysis and simulation for development of the first clinical micro-CT scanner. *Acad. Radiol.*, **12**, 511–25.

Wear, K.A., Laib, A., Stuber, A.P., and Reynolds, J.C. (2005) Comparison of measurements of phase velocity in human calcaneus to Biot theory. *J. Acoust. Soc. Am.*, **117**, 3319–24.

Weaver, J.B., Ballon, D.J., and Wolbarst, A.B. (2006) Magnetic resonance imaging, in *Advances in Medical Physics - 2006*, (eds A.B. Wolbarst, R.G. Zamenhof, and W.R. Hendee), Medical Physics Publishing, Madison, pp. 123–49.

Weaver, J.B., Van Houten, E.E., Miga, M.I., Kennedy, F.E., and Paulsen, K.D. (2001) Magnetic resonance elastography using 3D gradient echo measurements of steady-state motion. *Med. Phys.*, **28**, 1620–28.

Wernick, M. and Aarsvold, J. (2004) *Emission Tomography: The Fundamentals of PET and SPECT*, Academic Press, Boston.

Wheless, J.W., Castillo, E., Maggio, V., Kim, H.L., Breier, J.I., Simos, P.G., and Papanicolaou, A.C. (2004) Magnetoencephalography (MEG) and magnetic source imaging (MSI). *Neurologist*, **10**, 138–53.

Wickline, S.A. and Lanza, G.M. (2003) Nanotechnology for molecular imaging and targeted therapy. *Circulation*, **107**, 1092–95.

Wild, J.J., Reid, J.M. (1952) The application of echo ranging techniques to the determination of structure of biological tissues. *Science*, **15**, 226–30.

Wolbarst, A.B. (1999) *Looking Within: How X-ray, CT, MRI, Ultrasound, and Other*

Medical Images Are Created, and How They Help Physicians Save Lives, University of California Press, Berkeley, CA.

Wolbarst, A.B. (2005) *Physics of Radiology*, 2nd edn., Medical Physics Publishing, Madison, WI.

Wolbarst, A.B., Chin, L., Stavrev, P. (2006) Optimization of radiotherapy treatment planning, in *Wiley Encyclopedia of Medical Devices and Instrumentation*, 2nd edn (J. Webster), John Wiley & Sons.

Wu, T., Moore, R.H., Rafferty, E.A., and Kopans, D.B. (2004) A comparison of reconstruction algorithms for breast tomosynthesis. *Med. Phys.*, **31**, 2636–47.

Yoshiura, T., Mihara, F., and Tanaka, A. (2005) Age-related structural changes in the young adult brain shown by magnetic resonance diffusion tensor imaging. *Acad. Radiol.*, **12**, 268–75.

Yu, P., Mustata, M., Turek, J.J., French, P.M.W., Mellocj, M.R., and Nolte, D.D. (2003) Holographic optical coherence imaging of tumor spheroids. *Appl. Phys. Lett.*, **83**, 575–77.

Zagzebski, J.A. (1996) *Essentials of Ultrasound Physics*, Mosby-Year Book, St. Louis.

Zangwill, L.M., Weinreb, R.N., Berry, C.C., Smith, A.R., Dirkes, K.A., Liebmann, J.M., Brandt, J.D., Trick, G., Cioffi, G.A., Coleman, A.L., Piltz-Seymour, J.R., Gordon, M.O., Kass, M.A., OHTS CSLO Ancillary Study Group (2004) The confocal scanning laser ophthalmoscopy ancillary study to the ocular hypertension treatment study: Study design and baseline factors. *Am. J. Ophthalmol.*, **137**, 219–27. See also: http://www.heidelbergengineering.com.

Zhao, W., Andriole, K.P., and Samei, E. (2006) Digital radiography and fluoroscopy, *Advances in Medical Physics - 2006* (eds A.B. Wolbarst, R.G. Zamenhof, and W.R. Hendee), Medical Physics Publishing, Madison, pp. 1–23.

Zhao, B. and Zhao, W. (2005) Temporal performance of amorphous selenium mammography detectors. *Med. Phys.*, **32**, 128–36.

Zimmerman, R.E. and King, M. (2006) Nuclear medicine, in *Advances in Medical Physics - 2006* (eds A.B. Wolbarst, R.G. Zam enhof, and W.R. Hendee), Medical Physics Publishing, Madison, pp. 101–21.

Further Reading

Andriols, K.P. (2006) Overview of Medical Imaging Informatics, in *Advances in Medical Physics - 2006* (eds A.B. Wolbarst, R.G. Zamenhof, and W.R. Hendee), Medical Physics Publishing, Madison, pp. 201–7.

Beiden, S.V., Wagner, R.F., and Campbell, G. (2001) Multiplereader studies, digital mammography, computer-aided diagnosis, and the holy grail of imaging physics: II, in *Medical Imaging 2001: Physics of Medical Imaging* (eds. L.E. Antonuk and M.J. Yaffe) SPIE, Bellingham, WA. *Proc SPIE* 4320, 619–27.

Beiden, S.V., Wagner, R.F., Doi, K., Nishikawa, R.M., Freedman, M., Lo, S.C., and Xu, X.W. (2002) Independent versus sequential reading in ROC studies of computer-assist modalities: analysis of components of variance. *Acad. Radiol.*, **9**, 1036–43.

Cacak, R.K., Winans, D.E., Edrich, J., and Hendee, W.R. (1981) Millimeter wavelength thermographic scanner. *Med. Phys.*, **8**(4), 462–65.

Chakraborty, D.P. and Berbaum, K.S. (2004) Observer studies involving detection and localization: modeling, analysis, and validation. *Med. Phys.*, **31**, 2313–30.

Chen, Z., Hou, T., Chen, Z.-W., Hwang, D.W., and Hwang, L.-P. (2004) Selective intermolecular zero-quantum coherence in high-resolution NMR under inhomogeneous fields. *Chem. Phys. Lett.*, **386**, 200–05.

Chotas, M.S., Dobbins, J.T., Ravin, C.E. (1999) Principles of digital radiography with large-area, electronically readable detectors: a review of the basics. *Radiology*, **210**, 595–99.

Cohen, A., Etienne-Cummings, R., Horiuchi, T. *et al.* (2003) Report on the 2003 Workshop on Neuromorphic Engineering, June 29–July 19, Telluride, CO. http://www.ini.unizh.ch/telluride/previous/report03.pdf.

Convery, D.J., Rosenbloom, M.E. (1995) Treatment delivery accuracy in intensity-modulated conformal radiotherapy. *Phys. Med. Biol.*, **40**, 979–99.

Doi, K., MacMahon, H., Giger, M.L., and Hoffmann K.R (eds) (1999) Computer-aided Diagnosis in Medical Imaging. *Proceedings of the First International Workshop on*

Computer-Aided Diagnosis, Elsevier Science, Amsterdam.

Donnelly, E.F., Price, R.R., and Pickens, D.R. (2005) Experimental validation of the Wigner distributions theory of phase-contrast imaging. *Med. Phys.*, **32**, 928–31.

Fahrig, R., Wen, Z., and Ganguly, A. (2005) Performance of a static-anode/flat-panel X-ray fluoroscopy system in a diagnostic strength magnetic field: a truly hybrid X-ray/MR imaging system. *Med. Phys.*, **32**, 1775–84.

Fishman, E., Magid, D., Ney, D., Macrae, K., Kuhlman, J., Robertson, D. (1989) Three-dimensional imaging: advanced medical applications, in *Innovations in Diagnostic Radiology* (J. Anderson), Springer-Verlag, New York, pp. 47–68.

Galindo, A. and Martin-Delgado, M.A. (2002) Information and computation: classical and quantum aspects. *Rev. Mod. Phys.*, **74**, 347–423.

Hendee, W. (ed.) (1999) *Biomedical Uses of Radiation*, Vol. A, VCH Publishers (subsidiary of John Wiley & Sons, Inc), New York.

Hendee, W. and Wells, P. (1997) *The Perception of Visual Information*, 2nd edn., Springer-Verlag, New York.

Hornak, J.P. (ed.) (2002) *Encyclopedia of Imaging Science and Technology*, Vols 1 and 2, John Wiley & Sons, New York.

Kimme-Smith, C. (1992) New and future developments in screen-film mammography equipment and techniques. *Radiol. Clin. North Am.*, **30**, 55–66.

Liu, B., Metz, C.E., and Jiang, Y. (2004) An ROC comparison of four methods of combining information from multiple images of the same patient. *Med. Phys.*, **31**, 2552–63.

Mackie, T.R., Holmes, T., Swerdloff, S., Reckwerdt, P., Deasy, J.O., Yang, J., Paliwal, B., Kinsella, T. (1993) Tomotherapy: a new concept for the delivery of dynamic conformal radiotherapy. *Med. Phys.*, **20**, 1709–19.

Montaldo, G., Aubry, J.-F., Tanter, M., and Fink, M. (2005) Spatiotemporal coding in complex media for optimum beamforming: the iterative time-reversal approach. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, **52**, 220–30.

Muramatsu, C., Li, Q., Suzuki, K., Schmidt, R.A., Shiraishi, J., Newstead, G.M., and Doi, K. (2005) Investigation of psychophysical measure for evaluation of similar images for mammographic masses: preliminary results. *Med. Phys.*, **32**, 2295–2304.

National Research Council Institute of Medicine (1996) *Mathematics and Physics of Emerging Biomedical Imaging*, National Academy Press, Washington DC.

Partain, C.L., Chan, H.-P., Gelovani, J.G., Giger, M.L., Izatt, J.A., Jolesz, F.A., Kandarpa, K., Li, K.C., McNitt-Gray, M., Napel, S., Summers, R.M., and Gazelle, G.S. (2005) Biomedical imaging research opportunities workshop II: Report and recommendations. *Radiology*, **236**, 389–403.

Rao, P., Santosh K., and Gregg, E. (1980) Computed tomography with microwaves. *Radiology*, **135**, 769–70.

Roco, M.C. and Bainbridge, W.S. (eds) (2003) *Converging Technologies for Improving Human Performance: Nanotechnology, Biotechnology, Information Technology, and Cognitive Science*, Kluwer Academic Publishers, Boston.

Su, X. and Smith, L.M. (2004) Demonstration of a universal surface DNA computer. *Nucleic Acids Res.*, **32**, 3115–23.

Waldrop, M.M. (1992) *Complexity*, Simon & Schuster, Inc, New York.

Wehrli, F.W., Leonard, M.B., Saha, P.K., and Gromberg, B.R. (2004) Quantitative high-resolution magnetic resonance imaging reveals structural implications of renal osteodystrophy on trabecular and cortical bone. *J. Magn. Reson. Imaging*, **20**, 83–89.

Westbrook, C. and Kaut, C. (1998) *MRI in Practice*, 2nd edn., Blackwell Science, Oxford, England.

Westphal, V. and Hell, S.W. (2005) Nanoscale resolution in the focal plane of an optical microscope. *Phys. Rev. Lett.*, **94**(14), article 143903. http://prl.aps.org/.

Wolbarst, A.B. and Hendee, W.R. (2006) Evolving and experimental technologies in medical imaging. *Radiology*, **238**, 16–39.

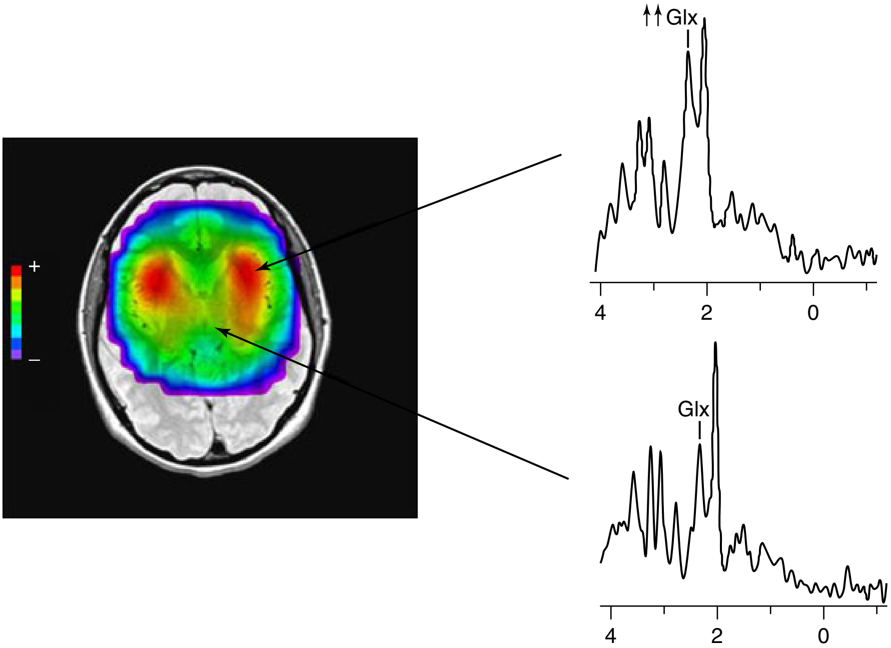

Fig. 27.29 Chemical shift/MR spectroscopic imaging (1000/18; bandwidth, 1000 Hz; 512 points acquired with two signals acquired). Transverse short echo-time chemical shift image (left) acquired at 0.5 T in a presymptomatic patient with Huntington disease shows strong elevation of glutamate (right: upper spectrum) in the head of the putamen. Unaffected thalamus (right: lower spectrum) is shown for comparison. (Image courtesy of Robert Prost, PhD, Medical College of Wisconsin, Milwaukee, WI.) (This figure also appears on page 954.)

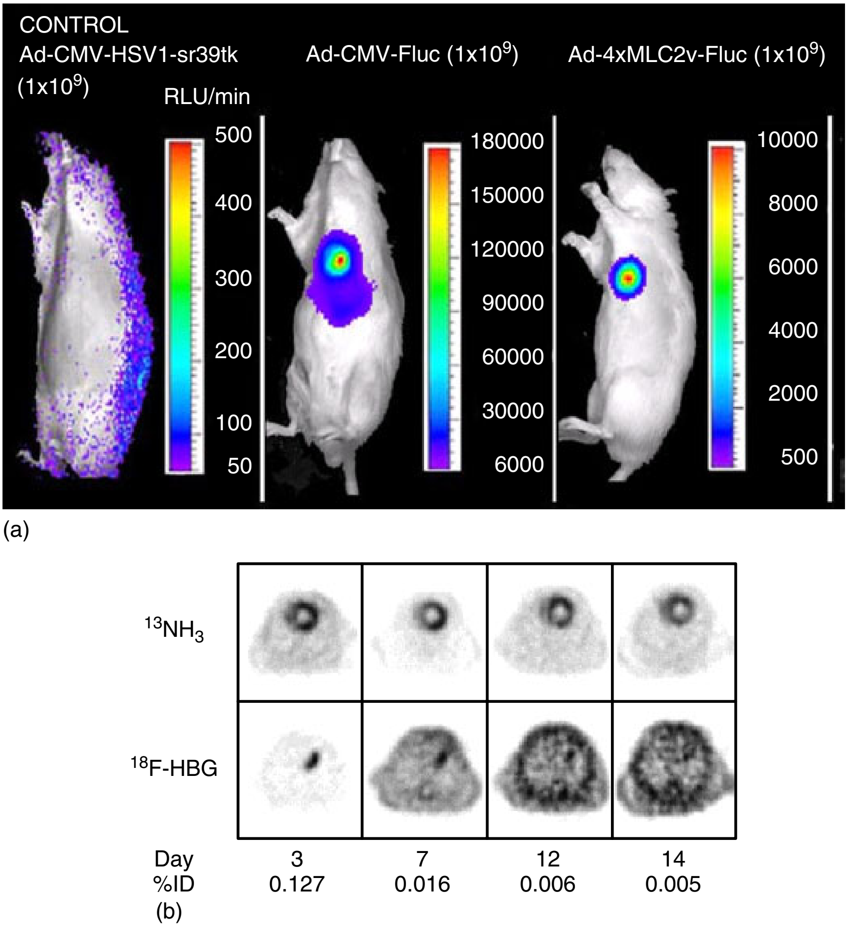
CONTROL
Ad-CMV-HSV1-sr39tk
(1x10^9)
RLU/min
500
400
300
200
100
50
Ad-CMV-Fluc (1x10^9)
180000
150000
120000
90000
60000
30000
6000
Ad-4xMLC2v-Fluc (1x10^9)
10000
8000
6000
4000
2000
500
(a)
13NH3
18F-HBG
Day
3
7
12
14
%ID
0.127
0.016
0.006
0.005
(b)

Fig. 27.32 Biomolecular imaging. (a) Optical bioluminescence imaging of cardiac gene delivery shown on charge-coupled device optical images of three rats injected (directly into the lateral wall) with control adenovirus (*Ad*) (left), adenovirus carrying firefly luciferase (*Fluc*) driven by cytomegalovirus (*CMV*) promoter (middle), and adenovirus carrying firefly luciferase driven by myosin light chain (*MLC*) promoter (right). Middle: Image shows expression in the myocardium and liver due to leakage of adenovirus into the bloodstream from the injection site. Right: Image shows only cardiac expression of luciferase due to use of myosin light chain promoter, which is not markedly active in hepatocytes. Left: Control image shows background low activity. All images show light image of mouse onto which bioluminescence image is superimposed, in relative light units (*RLU*) per minute. (Image courtesy of Sanjiv Gambhir, MD, PhD, Stanford University, Palo Alto, CA.) (b) Serial transverse PET images of reporter gene expression in a rat as function of time in days after dose injection. Top row: Animal was injected with 13N-labeled ammonia. Bottom row: Injected agent is 18F-labeled *HBG* gene. %*ID* × percent of injected dose. (Image courtesy of Heinrich Schelbert, MD, PhD, David Geffen School of Medicine, University of California at Los Angeles.) (This figure also appears on page 962.)

INDUCTORS

DENNIS PAVLIK, *Science and Technology Center, Westinghouse Electric Corporation, Pittsburgh, Pennsylvania, U.S.A.*

Introduction 1
1. **General Principles and Historical Perspective** 2
1.1 The Concept of Inductance 2
1.2 Analytical Methods 3
1.3 Types of Inductors 4
2. **Principles of Operation** 6
2.1 Fixed Inductors 6
2.1.1 Resonant Circuits 6
2.1.2 Chokes 6
2.1.3 Energy Storage 6
2.2 Adjustable Inductors 7
2.3 Variable Inductors 7
2.3.1 Moving Core 7
2.3.2 Saturable Reactors 8
2.3.3 Magnetic Switches 8
3. **Engineering Aspects** 10
3.1 Electromagnetics 11
3.2 Dielectrics 11
3.3 Thermodynamics 11
3.4 Mechanics 11
4. **Inductor Applications** 11
4.1 Power Reactors 12
4.2 Energy-Transfer Inductors 12
4.3 Resonant Charging for Particle Accelerators 13
4.4 Superconducting Magnetic Energy Storage 14
4.5 Magnetic Switches 15
4.6 Electronic Circuit Applications 16
4.6.1 Resonant Tuning 16
4.6.2 Electronic Filters 16
4.6.3 Current-Source Reactors 17
Glossary 18
Works Cited 18
Further Reading 19

INTRODUCTION

When a current flows through an electrical circuit, it creates a magnetic field. Inductance is the quantity that relates the energy stored in a magnetic field to the current flowing through that circuit. It defines the relationship between the voltage and current needed to add or extract energy from the magnetic field. *Inductor* is the generic name for a class of devices used for adding inductance to an electrical circuit. It covers a wide range of devices, applications, and configurations. Inductors range in size from small fractional-turn helixes having masses of several grams to devices hundreds of cubic meters in volume having masses in excess of 50 000 kg. Inductors are used as components of tuned circuits, electronic filters, magnetic energy storage devices, and measurement apparatus.

Because inductors usually comprise a number of turns or coils of conductor on a form, they are sometimes referred to as inductance coils, electric coils, or simply coils. There are other names used to describe inductors for specific applications. These include choke, choke coil, energy-storage choke, reactor, inductometer, induction coil, magnetic switch, pulsactor, and saturable reactor. A number of these will be described in detail to illustrate the distinction among different types and classes of inductors.

Inductors are classified according to their geometric configuration and whether or not a magnetic material is used to enhance the inductive effect. Inductors without magnetic materials are called air-core inductors. Those that use magnetic materials to enhance the inductive effect are referred to as iron-core inductors.

Inductors are typed according to whether or not they provide a fixed, adjustable, or variable inductance. The different classes and types are represented schematically by four different symbols as shown in Fig. 1. Combinations of these basic symbols are used to represent more complex inductor arrangements.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

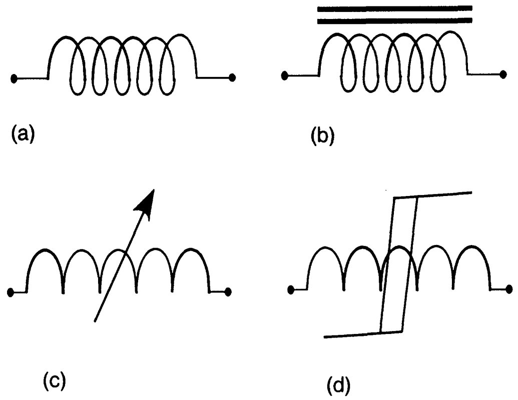

FIG. 1. Schematic symbols used to identify the different types of inductors: (a) air-core fixed inductor, (b) iron-core fixed inductor, (c) variable inductor, (d) saturating or nonlinear inductor.

The first section of this article explains the concept of inductance and the general principles and techniques used in its calculation. A historical perspective is taken to illustrate the development of different concepts and calculation procedures.

The class of air-core inductors is described and some formulas presented for the calculation of inductance. This is expanded to the more general class of linear inductors that can have iron cores but operate in the linear region. The operation of nonlinear and variable inductors is also presented.

The section on engineering aspects discusses how the electromagnetic system requirements are used in conjunction with the thermal and mechanical limitations of the available construction materials to synthesize a specific configuration. Examples of specific applications are presented in the last two sections.

1. GENERAL PRINCIPLES AND HISTORICAL PERSPECTIVE

1.1 The Concept of Inductance

In 1831 Faraday first showed that an electric current is induced in a closed circuit when the magnetic field in its vicinity is altered. This change in field could be caused by the movement of a permanent magnet in relation to the circuit or by a changing current in a neighboring electrical circuit. Lenz showed that the induced electromotive force (emf) or any current it generated in a closed circuit always acted to prevent the magnetic field linking it from changing. The amount of flux linking an electrical circuit is called the flux linkage λ. These two observations form Faraday's law:

$$\text{emf}=e=-\frac{d\lambda}{dt}. \tag{1}$$

Faraday (1839) reported that when a current flowing in a circuit was interrupted it produced an emf greater than that of the voltage source used to initiate the current. He showed that the inductive effect of a circuit on itself was part of the same phenomenon as that of a current affecting neighboring conductors. The inductive effect of a circuit on itself is called self-inductance. The inductive interaction between two or more distinct electric circuits is defined as the mutual inductance between the circuits (see TRANSFORMERS).

If the change in flux linkages is not due to a physical motion of conductors and the materials are linear, the induced emf can be related to changes in current by

$$\text{emf}=-L\frac{dI}{dt}, \tag{2a}$$

where L is the self-inductance of the circuit. For harmonic fields, Eq. (2a) becomes

$$V=j\omega LI=j(2\pi fL)I, \tag{2b}$$

where V is the magnitude of the counter emf, j is $\sqrt{(-1)}$ representing the 90° phase shift between voltage and current in phasor notation, ω is the angular frequency of the sinusoid equal to $2\pi f$, and I is the phasor current. The quantity $2\pi fL$ is known as the reactance of the inductive circuit element. The unit of inductance is a weber per ampere designated as the henry, after Joseph Henry, an American scientist and inventor. Assuming a constant inductance, Eqs. (1) and (2a) can be combined to give an equation for L in terms of λ and I:

$$L=\frac{d\lambda}{dI}, \tag{3a}$$

which, subject to linearity, is also

$$L=\lambda/I. \tag{3b}$$

The concept of flux linkages is often confusing and difficult to apply. For this reason, Maxwell (1891) preferred to explore inductance in terms of energy and geometrical functions giving an alternate definition of inductance:

$$L = 2W/I^2, \tag{4}$$

where W is the energy stored in the magnetic field as a result of the current I in the circuit. The stored magnetic energy can be related to the magnetic vector potential A. This is extremely useful in field calculations using numerical methods, such as the finite-element technique, as it allows for the direct calculation of inductance in very complicated arrangements and multicoil systems (see MAGNETOSTATICS).

Equations (2)–(4) give the same values of inductance and are useful when the surrounding media and materials are linear and there are no hysteresis, residual fluxes, or eddy-current effects. However, when these conditions are not met, using these equations interchangeably can result in serious errors. Rader and Litscher (1944) extensively treat the subject of inductance in the presence of nonlinear materials and eddy currents, giving examples showing when each formulation is appropriate.

The distinction between the formulas is illustrated in Fig. 2. Consider a nonlinear inductor operating at a current whose magnitude I is given by the length *oa* and total flux linkage is equal to length ***ab***. The linear or flux-linkage inductance L_b from Eq. (3b) is *ab/oa*. However, from Eq. (4) the inductance L_c is equal to twice the total stored energy (shaded area *obc*) divided by the square of the current. At operating point *b*, the incremental inductance L_e given by Eq. (3a) is the slope of the line *o′e′*, which is just *ae/oa*. In nonlinear circuits $L_b > L_c > L_e$, and which one is appropriate to use depends on the circuit application.

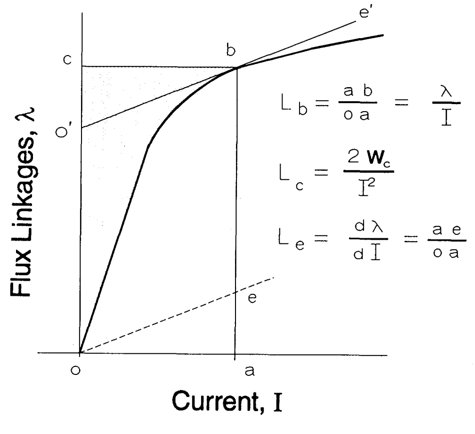

FIG. 2. Nonlinear magnetization curve illustrating how "inductance" differs depending on the selected definition and operating point.

In nonlinear devices such as saturable reactors and magnetic amplifiers the incremental inductance L_e is more appropriate. Magnetic switches are highly nonlinear, and their performance characteristics must be derived from a direct analysis of the *B-H* characteristics of the device (see related articles on MAGNETS and ENERGY CONVERSION).

1.2 Analytical Methods

The development of the concept of inductance was based on experimental observations. The defining equations are from this observational reference frame in terms of energy or voltage and current. Fundamentally, however, inductance is a measure of how an arrangement of conductors stores and exchanges electrical and magnetic energy. Therefore, it is a geometric quantity dependent on the physical arrangement of the currents and the permeability (inductive capacity) of the surrounding space.

The difficulty with magnetic materials arises because their inductive behavior is both nonlinear and hysteretic. When the surrounding space has materials that are conducting, currents are induced in them through mutual induction, and the observed inductance at the terminals is the resultant of the combined effects, making the calculation of inductance difficult.

Highly accurate formulas for inductance are available for cases where no magnetic materials, hysteresis, or eddy-current effects are present. Maxwell and his contemporaries developed exact solutions for the simple geometries comprising sheets or filaments of current. In all but the simplest cases, these formulas contained complicated inverse hyperbolic, trigonometric, and elliptic integrals. Analysis of these functions was difficult given the available tools. The approach taken was to represent the more difficult functions by rapidly converging series expansions that were then used to derive solutions for more complex geometries. Correction factors were developed to account for the fact that real inductors were made of finite-dimensioned wire with separation between turns. These complexities can best be illustrated by example.

Consider a circular loop of mean radius a, using a conductor with circular cross section of radius r. The inductance in henries with all dimensions in meters is given by

$$L=0.4\pi a[(1+r^2/8a^2)\ln(8a/r)(r^2/24a^2)$$
$$-1.75]\times 10^{-6}. \qquad (5)$$

Lorenz developed a closed formula for the inductance of a finite-length right circular cylindrical solenoid, with a current-sheet winding (infinitesimal radial thickness), in terms of elliptic integrals, given by

$$L=(3.2\pi N^2a^3/3b^2)$$
$$\times\{(2k^2-1)E/k^3$$
$$+(1-k^2)F/k^3-1\}\times 10^{-6}, \qquad (6)$$

where a is the radius of the solenoid, b is its axial length, and F and E are complete elliptic integrals of the first and second kind, respectively, and modulus $k=[4a^2/(4a^2+b^2)]^{1/2}$.

A highly accurate formula for a solenoid with a finite radial build (thickness of winding including insulation) c, length b, and mean radius a, developed by Weinstein, is given by

$$L=0.4\pi aN^2(p+s)\times 10^{-6} \qquad (7)$$

where $x=b/c$ and

$$p=\ln(8a/c)+\tfrac{1}{12}-\pi x/3+\ln(1+x^2)$$
$$\times[1/(12x^2)-\tfrac{1}{2}]+(x^2/12)\ln(1+1/x^2)$$
$$+\tfrac{2}{3}(x-1/x)\tan^{-1}(x) \qquad (8)$$

$$s=(c^2/96a^2)\{[\ln(8a/c)-0.5\ln(1+x^2)]$$
$$\times(1+3x^2)+3.45x^2+221/60-1.6\pi x^3$$
$$+3.2x^3\tan^{-1}(x)-\ln(1+x^2)/10x^2$$
$$+(x^4/2)\ln(1+1/x^2)\}. \qquad (9)$$

Equation (7) loses accuracy as the solenoid approaches a current-sheet type winding ($x>25$) or as c approaches a. Although very accurate for squarish type sections ($x<10$) and $a>c$, the terms in Eqs. (8) and (9) consist of large positive and negative numbers that nearly cancel. These equations illustrate the two major difficulties in using series expansions:

1. The equations are generally only valid over a particular range; and
2. the individual terms in the series nearly cancel, requiring a high degree of accuracy in the separate terms for even moderate accuracy in the final answer.

Rosa and Grover (1911), working for the U. S. National Bureau of Standards, made what is perhaps the most comprehensive investigation of the various formulas available at the time. Later, Grover (1946) refined this work and produced a book of working formulas and tables that can be used to calculate the inductances for the important geometric arrangements. Grover's formulas employing only one or two terms give accuracies to greater than one part in 1000 over extremely wide ranges of geometric variation. Corrections for winding space and insulation are readily incorporated into the formulations for all of the important geometries.

From a geometric perspective, inductance is just N^2 times the permeance $\mathscr{P}$ of the magnetic path. If the magnetic path can be clearly defined, it is possible to derive reasonable approximations for inductance. In general, defining the magnetic path is difficult, and the magnetic-circuit approach offers no calculational advantage. Iron-core inductors are an exception, since one of the purposes of using a magnetic material is to provide a clearly defined magnetic path. Therefore, the magnetic-circuit approach provides a reasonably simple and accurate approximation for the inductance in many iron-core inductor problems.

Today numerical methods and high-speed computers make it possible to calculate inductances accurately for complex conductor arrangements including the effects of eddy currents, nonlinear magnetic materials, and hysteresis.

1.3 Types of Inductors

The three main parts of an inductor are the winding, the winding form, and the core. The winding usually comprises a number of turns of conductor wound about some axis of symmetry. The winding form is the structure used to support the winding. The region enclosed

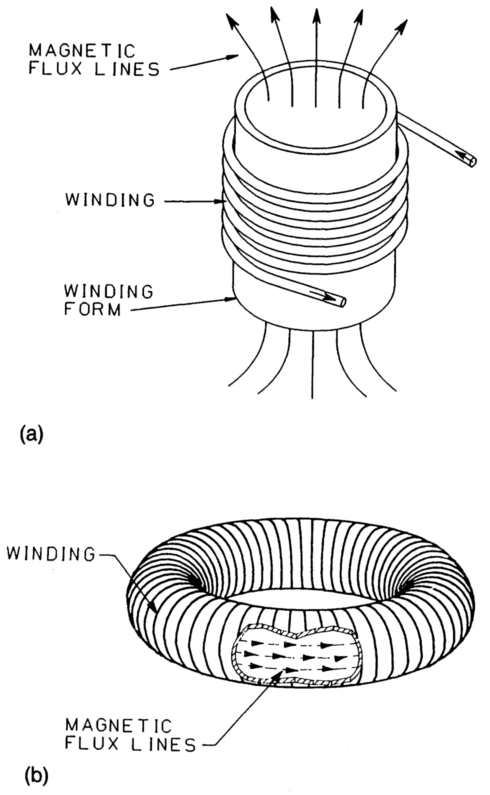

FIG. 3. The two most common types of air-core inductor configuration: (a) solenoid and (b) toroid.

by the winding in which the magnetic field is formed is called the core. Inductors with nonmagnetic core materials (wood, ceramic, plastic, air, oil, etc.) are called air-core inductors. If a magnetic material is used in the core to augment the inductive effect, the inductor is referred to as an iron-core inductor. The most common materials used in iron-core inductors are alloys of silicon steel. However, many different magnetic materials are used in the construction of inductors, especially at high frequencies, such as ferrites, powdered iron, cobalt-iron and nickel-iron alloys, and amorphous metals.

The two most common geometric forms of air-core inductors are solenoidal and toroidal. An inductor is solenoidal if the direction of the magnetic field is primarily straight within the body of the inductor and flux closure is external to the winding. Figure 3(a) shows a short solenoidal inductor of circular form. The winding form for a solenoid is generated by taking a closed planar figure (circle, rectangle, polygon) and extruding it along a straight line normal to its plane. If the length of the solenoid is short relative to its diameter, it is referred to as a pancake coil.

A toroidal inductor is one in which the direction of the magnetic field is predominantly a closed (generally circular) path so that flux closure is within the body of the winding. A toroidal winding form is generated by taking a closed planar surface and rotating it about a straight line in the same plane but external to the closed surface, as illustrated in Fig. 3(b). Because the winding closes on itself, toroidal inductors have a relatively small external magnetic field. In general, toroids have continuous smooth windings that approximate a current sheet over the surface of the winding form. It is not always possible or desirable to have a continuous winding over the surface of the form. An alternative to the smooth toroidal construction is to use discrete short solenoidal sections arranged in a generally circular pattern as illustrated by Fig. 4. This type of arrangement is known as a caged toroid and is used primarily in large energy-storage inductors and confining magnets used for fusion experiments.

In toroidal windings there are two components of electromagnetic forces acting on the windings. The first acts on the local cross section (minor radius) to expand the coil, as with any loop of wire. The second force acts to

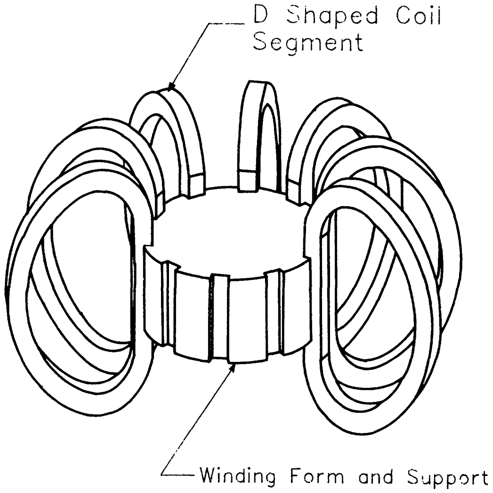

FIG. 4. Cage-type toroidal inductor with a 10-segment winding. The "D"-shaped coils provide a constant-tension winding that improves the inductor's ability to withstand electromagnetic forces.

compress the entire toroid inwardly toward the major axis. File *et al.* (1971) reported that a constant mechanical tension in the winding could be maintained by using a D-shaped cross section for the winding form. Shafranov (1973) later reported that this was also the optimum cross section for maximizing the inductance per given length of conductor for a toroidal arrangement. The D-shaped cross section is the shape taken by a sausage or water balloon resting on a flat surface. The internal expansive forces are like the local Lorentz forces in the coil, and the gravitational action flattening the bottom is analogous to the Lorentz forces acting to compress the toroid about the major axis.

The windings on iron-core inductors are generally either solenoidal or toroidal in form although alternative concepts are sometimes used. In addition to enhancing the inductive effect, the magnetic core also serves to contain the magnetic field within a controlled volume, limiting the "leakage flux." These inductors are generally named according to the shape of the magnetic core rather than winding type. A partial list of the more popular core configurations includes C-core, E-core, picture-frame, gapped-core, two-legged, three-legged, and racetrack.

Electrically conducting cores are sometimes used in variable-inductor high-frequency applications. Their operation is based upon a transformer action established between the main winding and eddy currents in the core, which act like a short-circuited secondary winding. As the core is inserted further into the winding, the magnetic coupling between winding and core increases and the inductance decreases, reaching a minimum when the core is fully inserted.

2. PRINCIPLES OF OPERATION

2.1 Fixed Inductors

These inductors provide a constant inductance over some range. There are many different configurations, but the essential feature is that the winding, core, and any air gaps remain fixed in position with respect to one another. Fixed inductors can be of the iron-core type, provided that their operating range is within the linear range of the core material. Many of these inductors use ferrites as the core material because of the high frequency (kHz to MHz) of operation.

2.1.1 Resonant Circuits. These are an important part of many electrical circuits and are the most common application for inductors. In these circuits, inductors and capacitors are selected to operate at a resonant frequency f_r given by

$$f_r = 1/2\pi\sqrt{(LC)}. \tag{10}$$

2.1.2 Chokes. These are inductors designed to isolate low-frequency from high-frequency signals in mixed-frequency electrical circuits. They are designed to "choke off" the flow of ac from one portion of a circuit and allow dc or relatively low-frequency ac power to flow to some other portion of the circuit. The inductance needed to achieve a desired level of isolation is derived from Faraday's law as

$$L = V/2\pi f_c I \tag{11}$$

where V is the voltage applied to the inductor and I is the maximum current that is to be allowed through the choke at the "cutoff" frequency f_c. Choke inductors come in a variety of geometric arrangements, but they are almost always of the iron-core type to maximize the inductance for minimum material and volume.

2.1.3 Energy-Storage Inductors. These inductors are designed to store energy in their magnetic fields (see STORAGE OF ENERGY). They are generally associated with power supplies of significant rating (several kVA to 1000 MVA) and are sometimes called inductive reactors or simply reactors. In the lower end of the energy ratings these devices are generally simple air-core solenoids. If stray fields are a problem, then an iron shield may be placed around the solenoid, or a toroidal arrangement may be selected having either a uniformly distributed or a cage-type winding construction (see Fig. 4).

In power systems these reactors are used

1. to supply energy for line charging in long transmission lines;
2. to limit fault currents through either series or neutral grounding;
3. to stabilize nonlinear loads in furnaces, welding, or other applications where arcs and nonlinear loads are present; and

4. to balance the current distribution between different parallel circuit branches.

They are also used in resonant power supplies for particle accelerator magnets (see Sec. 4.3), electromagnetic launchers, and some pulse-compression systems (see Accelerators, Linear; Cyclotrons; Synchrotrons).

These inductors are generally air-cored because of the high energy density achievable in nonmagnetic materials, equal to $B^2/2\mu$. Magnetic materials are often used for a portion of the magnetic flux path to concentrate the energy storage into small volumes of nonmagnetic material at a high flux density and to control the leakage flux. These reactors generally have solenoidal windings surrounded by an iron frame and are referred to as "solenoids in a box," or "picture frame" inductors as shown in Fig. 5(a). This type of inductor can still be classified as an air-core inductor as the volume surrounded by the winding is nonmagnetic. The air-core versus iron-core classification often becomes clouded as is the case with the inductor shown in Fig. 5(b). In this case, the volume surrounded by the winding as well as the outer legs consist of alternating magnetic and nonmagnetic regions. This arrangement is used to provide a linear inductance with high energy-storage capability and a well-defined magnetic path. The inductor in Fig. 5(c) uses two windings wound about the legs of a magnetic core and is clearly an iron-core-type inductor.

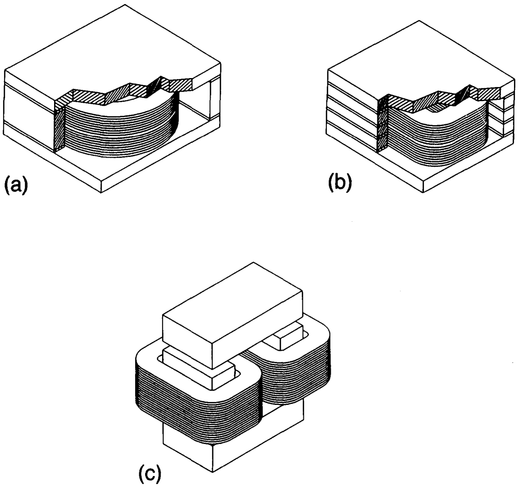

FIG. 5. Different construction methods used for large air-core and iron-core inductors. (a) Picture-frame construction with one large air gap in center of coil. (b) Picture-frame with continuous central leg and air gaps in the external legs. (c) Double-winding distributed air-gap construction.

The inductance required for a particular level of energy storage is found from Eq. (4). If the energy is to be stored in a region of constant permeability μ at a uniform flux density B, then the volume $\mathscr{V}$ needed is given by

$$\mathscr{V}=2\mu W/B^2. \qquad (12)$$

Equations (4) and (12) are used in combination with the characteristics of the power supply to define the optimum physical configuration for a particular application.

2.2 Adjustable Inductors

Adjustable inductors provide a variation in inductance through physical changes in the configuration of the winding arrangement. These inductors have either taps or sliding contacts that are used to alter the number of turns forming the inductor. They may also consist of several fixed inductors that are switched into various series or parallel combinations to provide a variation in inductance. In these inductors the variation in inductance is in discrete steps rather than a continuous variation.

2.3 Variable Inductors

Variable inductors provide an inductance that varies in a continuous manner over some range. This variation is achieved by

1. moving either all or part of the magnetic core with respect to the winding;
2. altering the permeability of the core material; or
3. moving one or more sections of the winding with respect to one another, thereby altering the mutual inductance between these sections and changing the inductance.

2.3.1 Moving Core. In this type of variable inductor a core is moved into or out of the center of a solenoidal coil. These devices are used in electronics applications where fine tuning of a resonant circuit is important. Typically the core is a low-loss ferrite material of cylindrical shape that moves on a threaded shaft. As the magnetic core is inserted further into the winding, the inductance increases over some range. In conducting-core inductors the inductance of the circuit decreases as the core is inserted further into the winding because of the eddy current generated in the conducting core material.

Moving-core inductors were formerly used to control the power flow to certain loads such as theater lights, where they operated in much the same manner as a rheostat without the attendant power loss. Solid-state devices have replaced variable inductors in most power delivery applications, but they are still widely used in communications circuits, using both magnetic- and conducting-core materials.

2.3.2 Saturable Reactors. These are devices that achieve a variable inductance by shifting the point on the magnetic saturation curve at which the inductor operates. This is accomplished by having an iron-core inductor with a control and load winding. The load winding is always divided into two separate windings as illustrated in Fig. 6. With this arrangement there is no ac voltage induced in the control winding; yet the control winding, shown on the central leg of the reactor, still functions to saturate the core. This can also be accomplished by using two separate magnetic cores, but the principle is the same.

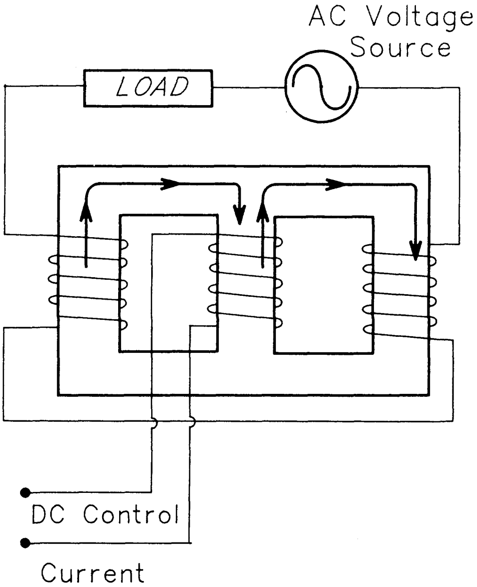

FIG. 6. Saturable reactor illustrating load and control winding arrangements. Note that the core and load windings are arranged so as to keep ac voltages from being induced in the dc control winding.

The inductance of the load winding is the incremental inductance determined by its operating point on the saturation curve (refer to Fig. 2). Shifting the operating point of the circuit is accomplished by varying a dc current through the control winding to cause either more or less saturation in the core. Saturable reactors were first introduced in the early 1900s, and their development proceeded along with that of vacuum tubes. Although saturable reactors tended to be heavy and bulky, they had many advantages over vacuum tubes. They are very simple, rugged, highly reliable devices capable of controlling large amounts of power. Saturable reactors can be designed to have essentially the same operating characteristics as silicon-controlled rectifiers (SCRs) and Triacs.

2.3.3 Magnetic Switches. These switches are saturable inductors capable of withstanding large voltages for short periods of time before saturating. Upon saturation the inductance of these devices decreases by two or three orders of magnitude. Magnetic switches are similar in operation to saturable reactors in that both depend on a large change in inductance caused by the saturation of a magnetic core material. Magnetic switches differ from saturable reactors in that they have a far greater swing in inductance, and it is the load circuit that drives the core to saturation, not a separate control winding. Schematically these devices are represented by the nonlinear inductor element shown in Fig. 1(d).

Melville (1951) first described the use of saturating reactors in pulse generators for radar modulators, nuclear particle acceleration, impulse testing in electrical apparatus, and auxiliary circuits in other power modulator applications. He coined the term "pulsactor" for these new devices that have since become known as magnetic switches. The principal advantages of these devices over electronic, spark-gap, or ignitron switching are that they have an indefinitely long and reliable life and can achieve high repetition rates, passive operation, and high efficiency. Their main disadvantage is that they are relatively heavy and difficult to operate in parallel because of triggering and timing problems.

Magnetic switches fell into relative obscurity until the 1980s, when the need developed for high-voltage (megavolt), high-power ($>10^{10}$ W), high-repetition-rate devices. Birx (1980) revisited the magnetic switch from the standpoint of these different application requirements taking advantage of new geometric configurations and better materials (amorphous metals; see METALLIC GLASSES).

A basic single-module magnetic switch circuit along with a magnetization curve and the voltage at several locations as functions of time are shown in Fig. 7. Operation of the switch is as follows: At time $t=0$ switch S_1 is closed and current begins to flow. Initially, the core is unmagnetized or is magnetized in the negative sense with respect to the applied voltage, and the inductance of the magnetic switch L_P is very large. As current begins to flow the flux in the magnetic core increases to the saturation flux density (B_s), and the impedance of the saturable reactor decreases rapidly. The inductance during the pulse is determined by the differential inductance as described in Sec. 1.1 and illustrated in Fig. 2. The initial inductance is the unsaturated value determined by the linear slope of the magnetization curve. The switch is driven so far into saturation that its saturated inductance can be calculated as if no iron core were present. The saturated inductance is typically 2000–3000 times less than the unsaturated inductance.

The time it takes for the switch to change from its "open" state to "closed" is given by

$$\int_0^{\tau_s} V_L(t)\,dt=NA\Delta B=NA[B_S-(-B_R)], \tag{13}$$

where V_L is the voltage across the magnetic switch, N is the number of turns on the core, A is the cross-sectional area of the core, ΔB is the change in flux density, and τ_s is the time to saturation. If the core is unpolarized, the flux swing will be from 0 to B_S. In general, however, magnetic switches are polarized so that the flux swing is from $-B_R$ to $+B_S$.

By using multiple stages in conjunction with capacitor storage, it is possible to get much higher changes in inductance as well as significant pulse compression compared to a single stage.

The Melville line is the classic example of a pulse-forming network using multiple magnetic switch modules, as shown in Fig. 8. Capacitor C_1 in Fig. 8 is charged through inductor L_0 at a rate roughly proportional to $\pi\sqrt{L_0C_1/2}$ for an ideal source. The charging continues until L_{P1} saturates. At this time C_1 will begin to discharge through L_{P1} and charge C_2. C_2 will charge much faster than C_1 because the saturated inductance of L_{P1} is designed to be much less than L_0. C_2 will then discharge into the load R once L_{P2} saturates. The saturated inductance of L_{P2} must be much less than that of L_{P1}. Figures 8(b) and 8(c) illustrate the voltage and current profiles at various points in the circuit. The voltage pulse is compressed in time but remains relatively constant in magnitude. The current pulse, on the other hand, is compressed in time but its amplitude is continually increasing from stage to stage. In an ideal system the energy transferred down the successive stages will remain constant with the magnitude of the current pulse increasing in direct proportion to the time compression it experiences. The energy

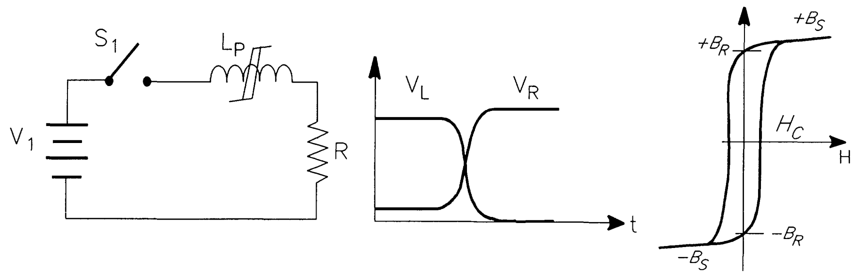

(a) (b) (c)

FIG. 7. Single-stage magnetic flux-compression circuit. (a) Schematic of the magnetic switch circuit. (b) Voltage wave forms across the magnetic switch and load as functions of time. (c) Magnetization curve of the switch core illustrating saturation and hysteresis.

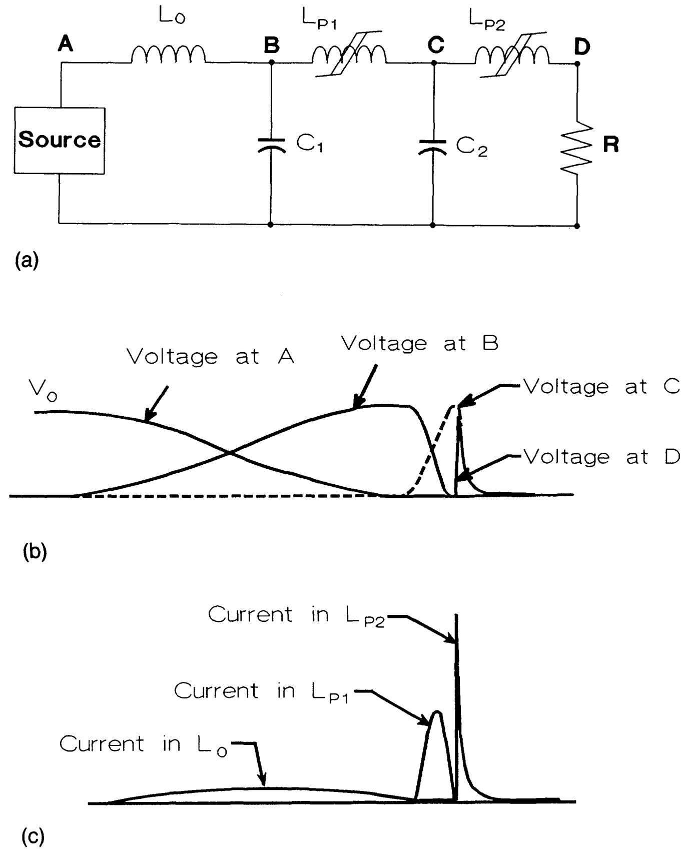

FIG. 8. Classic Melville line two-stage magnetic pulse compression circuit. (a) Schematic of compression line. (b) Voltage compression at selected points as a function of time. (c) Current through selected components illustrating compression of pulse width and increasing pulse magnitude.

lost per stage is typically less than 1% of the input energy.

The circuit described has only two magnetic switch/capacitor modules. In practice this arrangement can be extended to k elements where each C_n and L_{Pn} form a switching module. The gain g of an individual module is defined as the ratio of the capacitor's charge time to its discharge time. This is also equal to the square root of the ratio of the saturated inductance of module $n-1$ to that of module n. The overall gain G is equal to the product of the gains of each individual module.

For a given energy to be transferred, the minimum core volume is given by

$$\text{Minimum core volume} = (\pi/2)^2 \mu_{\text{sat}} W (g/\Delta B)^2 (1/sf). \qquad (14)$$

It depends on the gain, flux swing ΔB, energy W to be transferred, saturation permeability μ_{sat}, and the ratio of magnetic material to total core volume, called the core-space factor, sf.

3. ENGINEERING ASPECTS

The first task in any design problem is to establish the performance requirements and constraints for the particular application. For inductors, these requirements include the inductance, operating voltage, frequency, duty cycle (continuous, intermittent, pulsed), allowable energy loss, and tolerances on each parameter. The constraints define the environment in which the inductor will operate. These include maximum operating temperature, ambient temperature, available and acceptable sources of cooling, limitations on stray magnetic fields, size, weight, and the thermophysical properties of the construction materials. Economic factors such as the relative cost of materials and labor as well as the operating costs act to constrain the design further.

The objective is to produce a design that utilizes the minimum amount of resources while meeting the performance specifications within the constraints. This requires input from a broad spectrum of engineering disciplines including electromagnetics, dielectrics, thermodynamics, and mechanics. The specific functional requirements determine the relative importance of the different disciplines, but each must be addressed to some degree. Making tradeoffs between the operating requirements and environmental constraints constitutes the bulk of the design effort as described in the following sections.

3.1 Electromagnetics

The electromagnetic analysis defines the inductor configuration best suited to the specific set of specifications. Tradeoffs of different configurations establish the best configuration for an application. The electromagnetic analysis determines the main field profile, leakage field, losses, and forces within the device and its surroundings. The analysis may be relatively simple or may involve sophisticated three-dimensional nonlinear magnetic-field analysis techniques.

3.2 Dielectrics

The insulation system is a critical part of the overall design. The spacings between conductors and structural members, winding form selection, and insulating materials are important design parameters. The requirements of the insulation system also influence the inductance and electromagnetic performance, in particular the leakage fields. The insulation designer must ensure that the insulating materials are compatible with the environment and perform over the life of the inductor. Transient performance must be carefully evaluated as inductors, by their very nature, can generate voltages far in excess of the normal operating levels when subjected to transients.

3.3 Thermodynamics

Losses of one type or another are generated in every inductor and must be managed. The thermodynamic analysis ensures that temperatures will not exceed the allowable operating values for the materials in the inductor or surrounding components. All aspects of cooling, loss management, and the thermodynamic properties of the various components must be addressed. The general progression of cooling schemes in order of increasing cost and effectiveness is from free convection to forced air to forced liquid and then to cryogenic methods.

The components in electronic circuits are generally thought of as having low loss. However, the loss density in these small devices is often quite high. For instance, some inductors in a household television set reach temperatures in excess of 100 °C. The coatings on the wire and the form itself must be designed to withstand these temperatures.

Thermal management plays a key role in the design of cryogenic inductors. In cryogenically cooled designs the thermal analysis is primarily directed toward heat leak from the outside into the cryostat. Internal losses are primarily due to movements of the conductors or structural components. Although these losses are small, even the slightest amount of heat can cause significant increases in temperature that can cause the inductor to become unstable, especially if it is superconducting.

3.4 Mechanics

Currents flowing through conductors generate forces and stresses that must be analyzed and supported. This requires an accurate knowledge of the magnetic-field profile and current distribution. In ac devices the forces are oscillatory and can cause unacceptable noise and/or mechanical fatigue after some period. Inductors have been designed that carry currents of several million amperes. Although transient in nature, the forces generated are huge, requiring massive structural supports. The mechanical analysis and design must ensure that the inductor will be self-supporting and structurally sound.

4. INDUCTOR APPLICATIONS

The diversity of names by which inductors are known only suggests the wide range of size, shape, and function to which these devices are applied. Some of the more novel

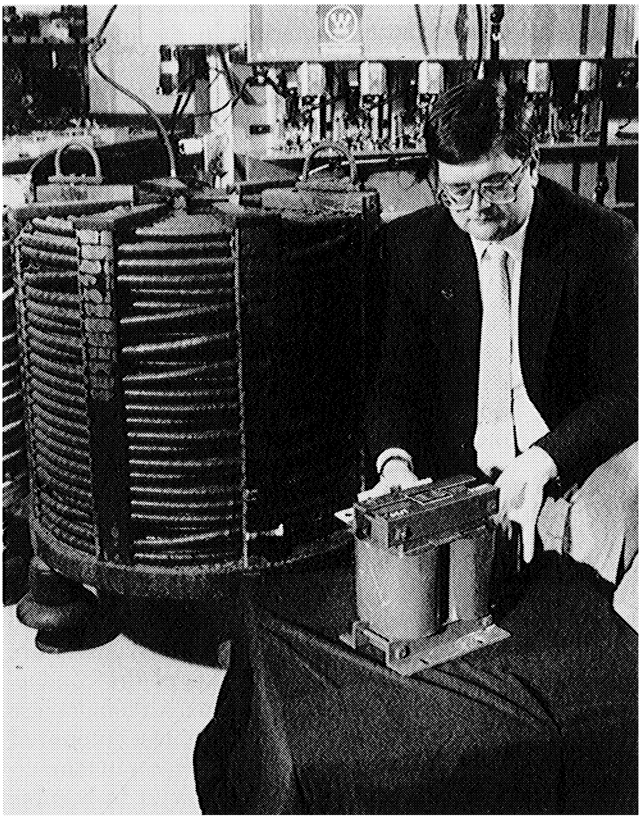

FIG. 9. An air-core power reactor used to provide a constant current for 100 μs during a switching operation is in the background, and an iron-core choke inductor used to limit ac current flow through a circuit is in the foreground.

aspects of inductor technology are presented to illustrate this diversity further.

4.1 Power Reactors

The loosely wrapped solenoidal air-core inductors shown in the background of Fig. 9 are typical of energy-storage reactors. These inductors have 94 turns in two parallel windings in a solenoidal arrangement 0.66 m long with a 0.7874-m outside diameter and 0.330-m inside diameter. If the current were uniformly distributed over the winding area, the inductance based on Eq. (7) would be 7.5 mH. However, the actual inductance, because of the looseness of the winding, is only 1.88 mH. This large variation between a theoretical uniformly distributed current assumption and reality illustrates the need for care in inductance calculations and why the correction factors for winding spacing must be considered in inductor design. These particular inductors are being used to provide a relatively flat current for a short period of time for a solid-state switching application. As the switch is opened, magnetic energy from the series-connected inductor is returned to the circuit. For time periods on the order of the switching time, 100 μs, the current remains essentially constant.

The relationship between the physical size of an inductor and its inductance can be misleading. For instance, the small iron-core inductor in the foreground of Fig. 9 is a 30-mH inductive choke, whose inductance is 15 times greater than that of the larger air-core reactors shown in the same figure. The volume of this inductor is 0.028 m^3 as compared to the 1.28 m^3 of the air-core reactors. This inductor is designed to limit the flow of ac current in a circuit.

4.2 Energy-Transfer Inductors

Inductors function as intermediate energy-storage and low-voltage to high-voltage load-matching devices in many pulsed-power applications. One such application is the hypervelocity electromagnetic macroparticle accelerator (or electromagnetic launcher). In

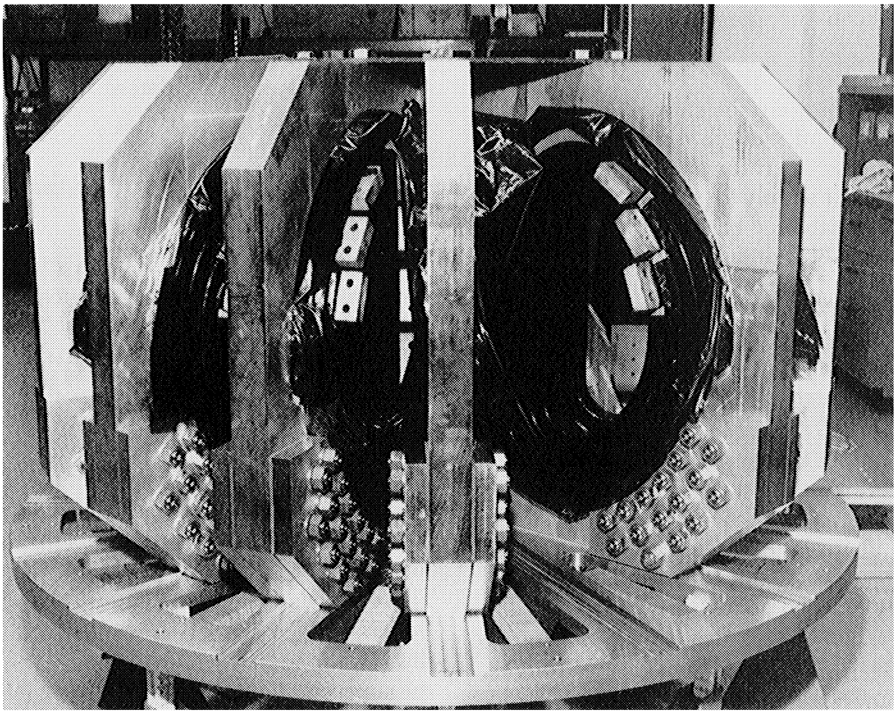

FIG. 10. Liquid-nitrogen–cooled, eight-segment caged-toroid energy-storage inductor used in a Westinghouse electromagnetic launcher power supply. The inductor will store 10 MJ of energy at 2 MA. Note the bracing structure needed to withstand the huge electromagnetically induced forces during operation.

this application, a low-voltage power supply is used to energize an inductor with high current over some relatively long period. At the appropriate time in the charging cycle, a commutating switch in series with the inductor and in parallel with the launcher is opened forcing current to flow into the high-voltage electromagnetic launcher. From Eq. (1) a sufficient voltage will develop across the inductor to keep the current flowing at its initial value. As the energy is extracted from the inductor the current will decay.

An inductor built for a 10-MJ particle-launcher system is shown in Fig. 10. It is a liquid-nitrogen–cooled toroidal cage-type inductor capable of storing 10 MJ of energy at 2 MA. Each of the eight coil segments measures 1.5 m high by 0.5 m across and is wound in the shape of a D to provide a constant-tension winding to maximize its energy-storage density. The Lorentz ($\mathbf{J}\times\mathbf{B}$) forces during operation are extremely high, which is the reason for the massive coil-support structure. The eight-section cage toroidal construction is a sufficiently close approximation to a pure toroid that the stray magnetic field is negligible.

4.3 Resonant Charging for Particle Accelerators

Synchrotron-type particle accelerators use powerful uniform-field electromagnets spaced periodically along the beam path to guide the particles along their equilibrium orbit. The field strength of these bending magnets must increase as the particles gain energy, and hence momentum, if they are to stay within their orbit. The field magnitude and rate of variation must be precisely controlled and balanced among all of the magnets for proper operation of the synchrotron.

White *et al.* (1956) introduced a resonant charging scheme that

1. reduced the cost and complexity of the magnet power supplies,
2. minimized the reactive power require-

ments from the electrical grid and,
3. could keep the magnets in synchronization.

The basic White circuit shown in Fig. 11 is designed with an inductor and capacitor resonant at the accelerator operating frequency and a dc power supply to pass the bias current. Energy is transferred back and forth between the magnet and inductor in precise synchronization with the particle acceleration. An ac source feeds power to the system via a secondary winding on the inductor to make up for losses. These resonant cells are connected in series to provide the synchronization among magnets. It is desirable to have all of the magnets connected in series; however, some accelerators require such a large number of magnets that this is not feasible.

The inductor is designed to store the energy of the magnets during the nonacceleration period and return it to the magnets as the particle accelerates. It also is used to provide makeup energy due to system losses through a secondary winding. Sasaki *et al.* (1989) designed a dual resonant power supply for the National Laboratory for High Energy Physics (KEK) in Japan that utilized a water-cooled air-core inductor comprising two pancake solenoidal coils enclosed in a picture-frame iron core as shown in Fig. 12. The secondary winding is directly interleaved with each primary turn to provide high magnetic coupling in this air-core inductor. Electrical losses are minimized by using a stranded cable transposed about a hollow conductor through which the cooling water circulates.

4.4 Superconducting Magnetic Energy Storage

Storage of large quantities of electrical power in a superconducting inductor for later recovery has been proposed by Boom and Peterson (1972) of the University of Wisconsin. In the superconducting magnetic energy storage (SMES) concept, electrical power from the ac grid is converted to dc, which is then used to energize a superconducting inductor. Since the inductor is superconducting, there are no losses and the current will circulate until extracted and converted back to ac power. There are significant refrigeration cost penalties associated with keeping the inductor at the low temperatures necessary to maintain the superconducting state. Studies have shown that efficiencies as great as 90% can be achieved by this method of energy storage as compared to 67% typical of hydroelectric pumped energy-storage systems. The greater efficiency is attributed to the ability of the inductor to store the electrical energy directly without intermediate energy conversion steps, as is the case with the pumped-storage concept.

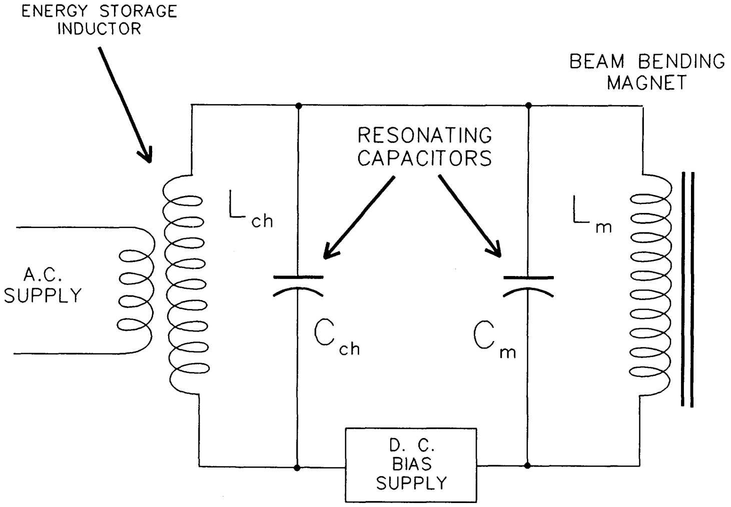

15p

FIG. 11. Circuit schematic of a single-module White circuit used for resonant charging of an accelerator magnet. Typically, these modules are connected in series to ensure that the same current flows in each magnet. Energy is shifted between magnet and inductor in precise synchronization with the beam acceleration profile.

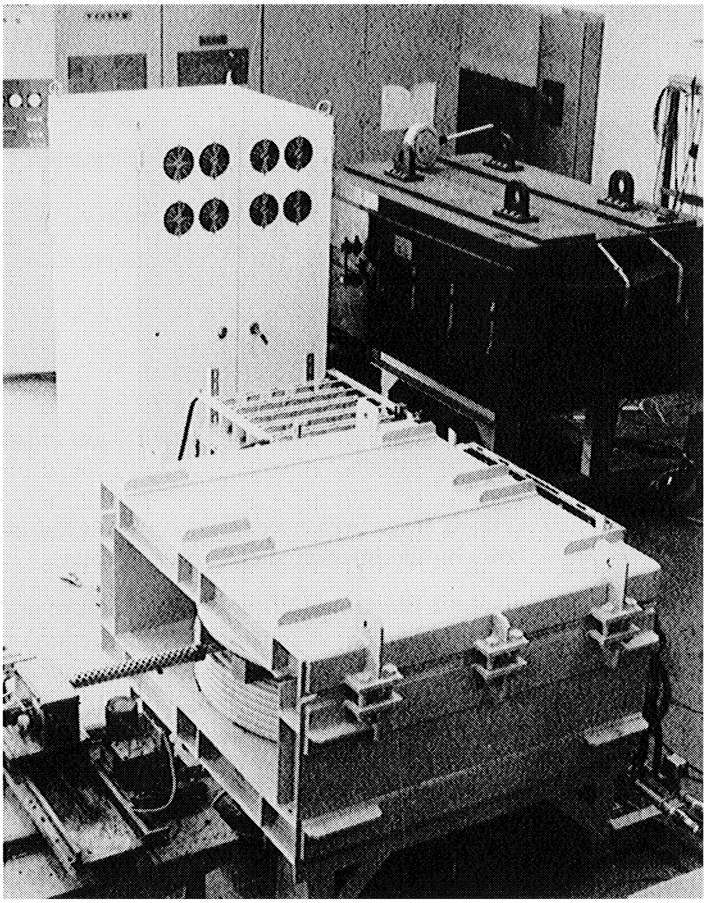

FIG. 12. Energy-storage inductor used in the Japanese National Laboratory for High Energy Physics (KEK) rapid-cycling proton synchrotron. The inductor is of picture-frame construction with two water-cooled pancake coils enclosed by an iron core. It is 11.8 mH with a peak current of 2150 A.

A 19.8-TJ (5500 MW h) full-scale SMES is currently being considered. An artist's conception of the concept being proposed by a team headed by Ebasco Services, Chicago Bridge & Iron, Westinghouse Electric, and the University of Wisconsin is shown in Fig. 13. This solenoidal inductor is 872 m (2864 ft) in diameter with a winding height of over 27 m (89 ft). The entire structure is enclosed in a cryostat filled with liquid helium designed to keep the heat leak to a minimum and to contain the tremendous tensile stresses on the winding. The entire structure is buried in a trench to provide additional support and stability.

4.5 Magnetic Switches

Magnetic switches have a wide range of applications in all types of pulse-forming networks including radar modulators, pulse compression lines for free-electron lasers, particle-beam accelerators, and solid-state switching circuits. The best performance is achieved in high–repetition-rate, high-voltage, continuous-duty applications. They have been used extensively in neutral-particle beam experiments such as PBFA and Comet.

With a core mass in excess of 2000 kg, the magnetic switch employed in the Comet experiment at the Sandia National Laboratories is perhaps the largest magnetic switch built to date. It is a single-shot device that has been tested to 2.7 MV, and 270 kJ per pulse at a peak current of 900 kA. The switch was a single-turn coaxial arrangement which formed an integral part of the transmission line. Figure 14 shows the switch in various stages of construction.

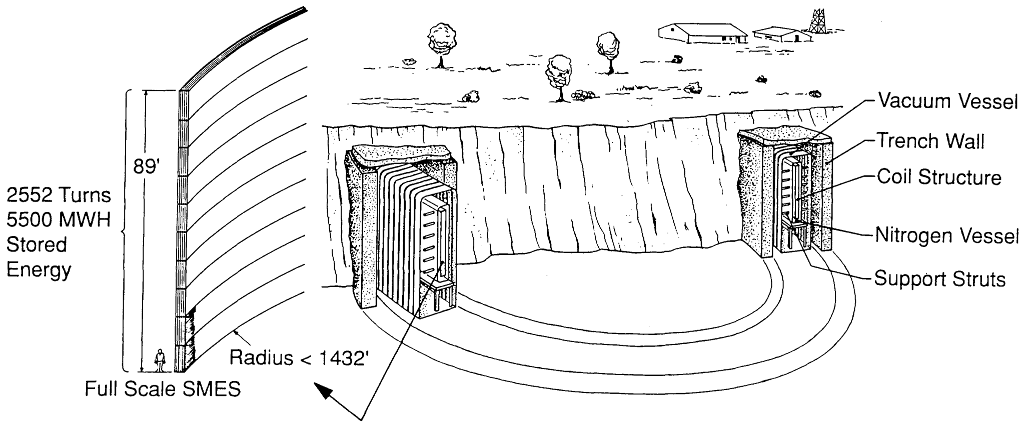

FIG. 13. Artist's conception of a full-scale superconducting magnetic energy-storage coil. The coil is over 870 m in diameter and stores 19.8×10^{12} J for later recovery by the power system.

4.6 Electronic Circuit Applications

Inductors in one form or another can be found in virtually every electronic circuit. Their uses in electronic circuits can be divided into three distinct functional categories: resonant tuning, filters, and current-source reactors.

4.6.1 Resonant Tuning. This type of tuning makes use of inductor-capacitor (*LC*) combinations to form a network selected to operate at a particular frequency as given by Eq. (10). These resonant networks provide regenerative feedback to an amplifier having nonlinear power characteristics to form a harmonic oscillator. Oscillator circuits are used in the generation, modulation, and demodulation of electromagnetic signals generally found in communications systems. Tuning to different frequencies is achieved by varying either the capacitance or inductance of the resonant combination. Variable inductors are used in circuits that may be susceptible to moisture or other contaminants because they are easier to seal against environmental factors than are variable capacitors.

In many applications, *LC* resonant elements have been replaced by circuits using piezoelectric crystals as the resonant elements because they are orders of magnitude more stable than inductor-capacitor combinations. Some integrated circuits have inductive elements built into them by the formation of a spiral conducting path. This arrangement, however, is generally avoided because of the relatively large area needed to construct inductors on the integrated circuit's substrate.

4.6.2 Electronic Filters. These make use of the variation in reactance as a function of frequency exhibited by inductors and capacitors to control which harmonics are either passed through or eliminated from a particular harmonic signal (see ELECTRONIC CIRCUITS). They are classified as either high-pass, low-pass, band-pass, or band-elimination depending on their function in the circuit. High-pass filters allow high-frequency signals through the system for further processing while rejecting frequencies below a particular value. The low-pass filter serves the opposite function. Band-pass filters select a band of frequencies for further processing, while band-elimination filters are used to reject a range of frequencies from the signal stream.

For inductors, the opposition to current flow, or inductive reactance, varies in direct proportion to the product of the inductance and frequency of the signal, as illustrated by Eq. (2b). In capacitors, the capacitive reactance varies inversely to the product of the capacitance and frequency. An inductor connected electrically in series with a load resistance represents a first-order low-pass filter because it presents a low series impedance to the lower frequencies of an applied harmonic signal. A second-order low-pass filter could use an inductor connected in series with the parallel combination of a capacitor and load. If the capacitor is placed in series with the parallel combination of an inductor and the

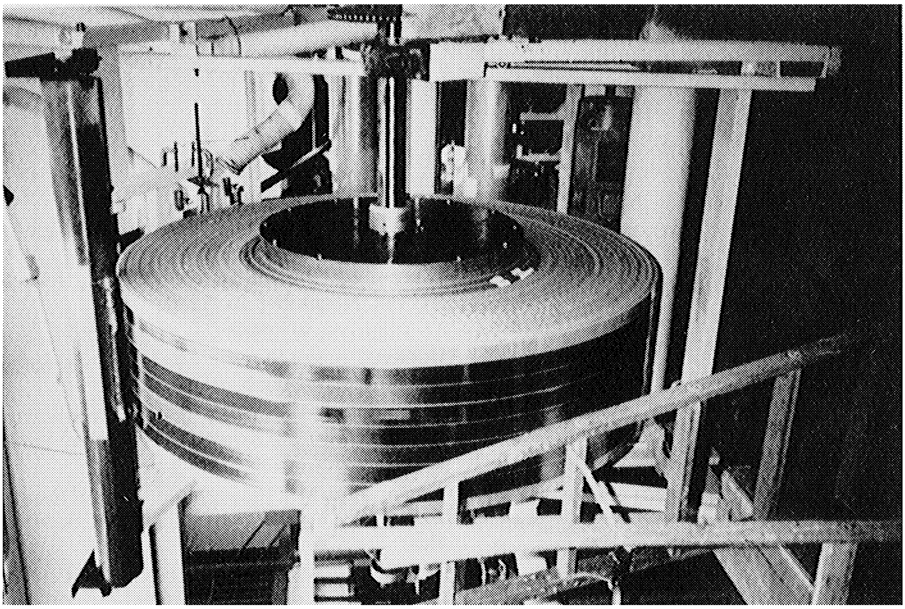

(a)

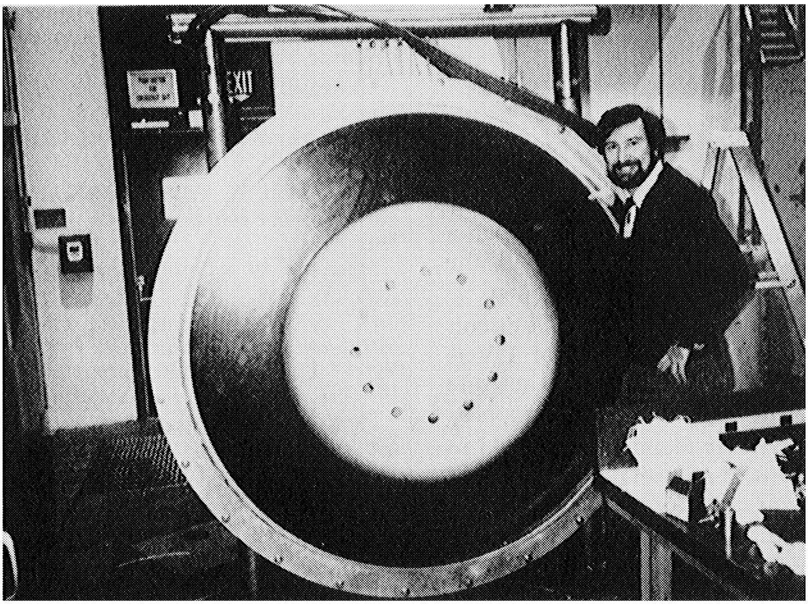

(b)

FIG. 14. Coaxial magnetic switch used in the Comet Project at Sandia National Laboratories at various stages of construction. (a) Magnetic core being wound from alternate layers of insulation and amorphous metal strip. (b) Finished magnetic switch ready for installation in the pulse compression line.

load, then a high-pass filter is obtained. Bandpass filters place an inductor and capacitor in series with the load, while band-elimination filters use a parallel combination of a capacitor and inductor in series with the load.

4.6.3 Current-Source Reactors. These reactors are used to control the magnitudes and rates of rise of current in electrical and electronic circuits. They are particularly useful in solid-state switching circuits used for power applications. Once triggered, solid-state devices can turn on in a relatively short space of time, generally 50 μs or less. If the current increases too rapidly, hot spots can be created that will thermally degrade the device. An inductor of appropriate size placed in series with the solid-state device will control the rate of rise of current through it to acceptable levels. Nonlinear inductors are often used for this application so that the series inductance decreases by several orders of magnitude after the solid-state device has been turned on. Many applications require that several solid-state devices operate in parallel. No matter how good the control system, not all devices will turn on at the same time. If this happens, all of the current will be flowing through only a few of the devices, resulting in an overload and premature failure of those devices. Eventually, the entire parallel combination will fail. By placing inductors in series with each device, the distribution of current can be controlled to ensure that the devices share the current load on an even basis.

GLOSSARY

Flux Linkage: The flux passing through a simple one-turn loop of wire. In a coil, the summation of the flux through each turn.

Henry: The unit of inductance. A circuit has 1 H of inductance if an emf of 1 V is induced when the current is changing at a rate of 1 A/s. Equivalent to a weber per ampere.

Hysteresis: The tendency for the field to persist in a magnetic material after the exciting field has been modified.

Permeability: A measure of the inductive capacity of a material given in units of henries per meter.

Phase Shift: The difference in phase angle between two sinusoidal waveforms.

Phasor notation: The technique of representing variables such as voltage, current, impedance, and resistance as complex numbers in the analysis of harmonic problems, having a magnitude and phase angle with respect to some reference.

Reactance: A complex constant of proportionality relating voltage to current, similar to resistance in Ohm's law. It is a function of frequency as well as inductance and capacitance.

List of Works Cited

Birx, D. L., *et al.*, (1980), "Basic Principles Governing the Design of Magnetic Switches," Lawrence Livermore Laboratory, UCID-18831, Berkeley, California, November 18, 1980.

Boom, R. W., Peterson, H. A. (1972), "Superconductive Energy Storage For Power Systems," *IEEE Trans. Magn.* **Mag-8**, 701–703.

Carter, G. W. (1967), *The Electromagnetic Field in Its Engineering Aspects,* 2d ed., New York: American Elsevier, Chaps. 8 and 9.

Faraday, M. (1839), *Experimental Researches in Electricity,* Vol. 1, Ser. 1, London: Bernard Quaritch.

File, J., Mills, R. G., Sheffield, G. V. (1971) "Large Superconducting Magnet Designs for Fusion Reactors," *IEEE Trans. Nucl. Sci.* **NS-18**, 277–282.

Grover, F. W. (1946), *Inductance Calculations, Working Formulas and Tables* (special printing by the Instrument Society of America with permission of Dover, New York, in 1982).

Maxwell, J. C. (1891), *A Treatise on Electricity and Magnetism,* 2nd ed., Vol. 2, Articles 693–706, London: Clarendon; reprint (1954), New York: Dover.

McLyman, W. T. (1978), *Transformer and Inductor Design Handbook,* New York: Marcel Dekker.

Melville, W. S. (1951), "The Use of Saturable Reactors as Discharge Devices for Pulse Generators," in: G. E. Williams (Ed.), *Proceedings of the Institution of Electrical Engineers (IEE)* (London), Vol. 98, Pt. 3 (Radio and Communications), No. 53, 185–207.

Rader, L. T., Litscher, E. C. (1944), "Some Aspects of Inductance When Iron Is Present," *Trans. Am. Inst. Elec. Eng. (AIEE)* **63**, 133–139.

Rosa, E. B., Grover, F. W. (1911), "Formulas and Tables for the Calculation of Mutual and Self-Inductance," U. S. Bureau of Standards Scientific Paper No. 169, Washington, DC: U.S. GPO, pp. 116–142.

Sasaki, H., Adachi, T., Someya, H., Sakai, I. (1989), "A DC Biased Rapid- cycling Magnet System Operating in a Dual Frequency Mode," in: A. Otter, J. Strathdee (Eds.), *Proceedings of the KAON PDS Magnet Design Workshop,* Vancouver, Univ. of British Columbia, B.C., pp. 33–40.

Shafranov, V. D. (1973), "Optimum Shape of a Toroidal Solenoid," *Sov. Phys. Tech. Phys.* **17**, 1433–1437.

Smith, S. (1985), *Magnetic Components Design and Applications,* New York: Van Nostrand Reinhold.

White, M. G., Shoemaker, F. C., O'Neill, G. K. (1956), "The 3 BEV High Intensity Proton Synchrotron," in: E. Regenstreif (Ed.), *Symposium du CERN sur les Accélérateurs de haute énergie et la Physique des Mesons, Genève, 11-23 Juin 1956*, Geneva, Switzerland: CERN, Vol. 1, pp. 525–529.

Further Reading

Information on design details and new applications and designs for inductors can generally be found in conference proceedings for applications to specific technology areas. The IEEE Pulsed Power Conference, the IEEE Power Modulator Symposium, the Symposium on Engineering Problems in Fusion Research, the *IEEE Transactions on Magnetics*, and the *IEEE Transactions on Nuclear Science* are good sources of related material for large inductor designs. The *Proceedings of the Institution of Electrical Engineers* (IEE) is also a good source of information. For accelerator applications there are numerous conferences sponsored by host laboratories throughout the world that can be good sources of information. For communications applications both the IEEE and IEE have societies that publish articles that may be of interest.

Carter (1967) provides a good physical insight into the electromagnetic field and treats the concept of inductance especially well. For anyone calculating inductances of standard geometries the Grover (1946) book is essential. Both Smith (1985) and McLyman (1978) do a thorough job of inductor design details for one wishing to build devices.

INFORMATION SCIENCE

CHARLES T. MEADOW, *Faculty of Library & Information Science, University of Toronto, Toronto, Ontario, Canada*

Introduction 21
1. **General Aspects** 22
1.1 Nature of Information 22
1.2 Related Concepts 23
1.3 Representing Information 23
1.4 Information as a Commodity 24
2. **Databases and Knowledge Bases** 25
2.1 General Aspects 25
2.2 Record Structure and Content 25
2.3 Representation of Attribute Values 25
2.4 Storage Structures and Indexes 26
2.5 Database Maintenance and Quality Control 26
2.6 Knowledge Bases 27
3. **Information Systems** 28
3.1 General Aspects 28
3.2 Major Components 28
3.2.1 Sources and Acquisition Subsystems 28
3.2.2 Processing Subsystems 28
3.2.3 Communication and Dissemination Subsystems 28
3.3 Control of Information Flow 29
3.3.1 Control of Input 29
3.3.2 Selection of Output 29
3.3.3 Delivery of Output to Users 30
3.4 Examples of Information Systems 30
3.4.1 Database Search 30
3.4.2 A Typical Accounting System 30
3.4.3 The American Institute of Physics—An Organization as an Information System 30
4. **Information Retrieval** 31
4.1 General Aspects and Models 31
4.2 Information Needs of Users 31
4.3 Query Languages 32
4.4 Intermediation—Assisting the User 32
4.5 Automatic Processing 33
4.6 Measurement and Evaluation of Retrieval Systems 33
5. **Informetrics** 34
5.1 General Aspects 34
5.2 Language Use 34
5.3 Publication Measures 34
5.4 Usage Measures 35
6. **Information Networks** 35
7. **The Users of Information Systems** 36
8. **Social and Commercial Aspects** 37
8.1 Use of Information Systems 37
8.2 Growth of the Information Economy 38
9. **Societies and Journals in Information Science** 38
Glossary 38
Works Cited 39
Further Reading 40

INTRODUCTION

Information science is a relatively new discipline, so new that the name is not widely known nor is there broad agreement as to what it means. Generally, it has grown out of studies of the design, use, and measurement of collections of information and of information systems, and, although rarely part of any

made that the systems studied are primarily computer systems. Computer science tends to concentrate on computer internals: design of hardware and algorithms, engineering of computer software, and the performance of the hardware or software components. Information science, on the other hand, tends to deal with systems as a whole, the interrelation of the systems with their users (indeed, even to

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

effect of information systems on their environment, and the extent to which they accomplish their objectives.

As a formal discipline, one branch of information science grew out of library science when, in the early 1960s, library science began to turn its attention increasingly to new methods of organizing and retrieving information, at least partly by the application of new technology to these problems. Other aspects are found in the disciplines of computer science, linguistics, communication (the social science), and telecommunications engineering. In Europe, the expression *information science* is rare; *informatics* is more likely to be used. Today, there are several respected scholarly societies and journals in the field, listed in Sec. 9.

After World War II there was a large amount of scientific documentation that had been created on the Allied side and more captured from other countries. In the 1950s the new "cold war" began, which had an accelerating effect on the growth of science and technology. Government agencies that supported scientific work believed that a vast amount of information was not being used because it was not accessible, and they invested heavily in developing systems to handle this information. *Handle* meant to organize it—classify and index it—and to develop means to search, retrieve, and disseminate it.

Technologically, information science evolved along with several streams of developments in other disciplines. These were

1. computers, primarily the evolution of time-shared large computers or, latterly, networked small ones;
2. telecommunications, especially packet-switched data communications networks;
3. large databases that became necessary for many disciplines, for example, in recording information about publications in science, and that demanded the use of computer–computer communications for remote searching;
4. new concepts in the organization and representation of information for more effective retrieval;
5. algorithms for interpreting and searching natural language and for measuring the relationships among words or texts; and
6. a movement toward more direct use of computers by end users, requiring new appreciation of user behavior.

This article is organized as follows: Section 1 discusses the nature of information as a mathematical concept and as a commodity, and the representation of information and coping with its ambiguities. Section 2 discusses databases, the logical configurations in which information is stored, and their structure; maintaining quality of the content of a database; and knowledge bases that record information about how to process or interpret data. Section 3 contains an overview of modern information systems; the components and flow of information; and some examples. Section 4 presents an overview of information retrieval, which today means mostly retrieval of text or coded representations of text. Section 5 discusses informetrics, or the measurement of information phenomena. Section 6 discusses information networks, which have made access to information systems not only physically easier but economically attractive. Section 7 considers users of information systems, and characteristics of importance and how they are studied. Section 8 explores the impact of information systems on society as a whole and some of the social issues of current concern. Section 9 discusses societies and journals in information science.

1. GENERAL ASPECTS

1.1 Nature of Information

Information is ubiquitous, so much so that it is difficult to define it in a way that satisfies both scientific rigor and popular conceptions. In the context of transmitting messages composed of strings of characters, Shannon (Shannon and Weaver, 1959) defined the information content of a communication system as

$$H = -\sum p_i \log_2 p_i , \qquad (1)$$

where p_i is the probability of symbol i being selected for transmission, and summation is over the number of possible symbols. The units of H are *bits*. Shannon was a communications theorist, concerned with such problems as computation of channel capacity. He emphasized that his work was not intended to apply to semantic aspects of information or to the meaning of messages transmitted. He also intended that H measure the amount of information in a *system* that generates symbols with the given probabilities, *not* the information content of a *single message*. The probabilities refer to the choice of a letter or a word, not of a message. The frequency distributions of letters in English are fairly uniform over a variety of writings, as are the frequencies of the most commonly used words. But no one knows or can predict the probability of a given message if the source is free to send any combination of English words.

The popular meanings of information (there are many) center around the related concepts of novelty, dispelling of uncertainty, and the probability of occurrence of a message. A message contains information if its contents were not already known to the recipient (novelty). A message contains information if it clarifies an issue (dispels uncertainty). An unexpected message contains more information than an expected one. That a heavy favorite wins a race is not much information, but that a long shot won *is* very informative. It is, of course, impossible to know the *a priori* probability of occurrence of a news item, which is one reason why Shannon warned against computing H for semantic information.

Yet another set of definitions shows the influence of cybernetics. Here, information is a message that changes the state of a system receiving it. Or, information is what enables decisions to be made, i.e., a choice among alternatives. Making a choice changes the state of a system; hence these definitions are compatible with each other. For more information on the formal information theory as applied to coding and transmission of information, see TELECOMMUNICATIONS.

1.2 Related Concepts

Terms closely related to *information* include *data, knowledge, meaning, intelligence,* and *wisdom.* As may be expected, there are no universally accepted definitions for these terms. *Data,* in common usage, tends to mean a set of symbols whose meanings or values are not necessarily known or accepted, but whose structure is known. Hence, it is possible to read data or operate upon it with a program, but without some form of interpretation or evaluation we cannot gauge its significance. The structural aspect is important. If we have simply a set of numbers with no known internal structure, it hardly even qualifies as data. What is data to one person may be highly significant information to another. The subjective nature of the distinction between data and information is becoming widely accepted.

Information is data that has been understood and evaluated, or is meaningful. Thinking, interpreting, or processing converts data into information. *Knowledge,* in general usage, is a higher form of information. The term tends to be applied to bodies of information rather than individual messages. When we use this word, we often mean to imply information possessed or accepted by an authority or an authoritative community. The sentence "We have knowledge of the laws of bankruptcy" is, in common usage, a far stronger statement than "We have information about bankruptcy law." Information acquires the status of knowledge by virtue of general acceptance or acceptance by respected authority, such as by a dictionary or encyclopedia or the *Handbook of Physics and Chemistry.* In Sec. 2.6 we discuss a particular form of knowledge called a *knowledge base.*

Intelligence has a number of meanings, one of which is approximately what we have called *information. Wisdom* is a term not routinely used in scientific work but often used colloquially. It tends to have much the same connotation as knowledge, but now knowledge held or promulgated by a person or community of special distinction.

1.3 Representing Information

An aspect common to all the various definitions is that information is *not* the ink and paper or magnetic disks with which it is represented. It is not something tangible. Information is a consequence of the receipt of a message or sequence of messages that may be recorded in any medium, using any symbol set mutually understood by sender and receiver. The symbolic representation of information consumes space, matter, and time to transmit, but the information itself does not. However, the material and time requirements for storage and transmission of symbols are of critical importance in information-system design.

In scientific work, information is often represented by numbers or codes made up of letters and numbers. There is little of the concept of syntax associated with such representations, but there is some. Given a vector or array, we know that each element is normally the same kind of entity (temperatures, weights, distances, etc., all represented as integers, decimal numbers, or floating-point numbers). Computer programs can easily be written to "understand" such data structures. However, if information is represented in natural language, the syntax is highly complex and often ambiguous. There is no complete definition of the syntax of a natural language nor of the composition of a picture. It is relatively easy to search a numeric array for occurrence of a stored value or a function of a stored value, such as the point where a second difference changes sign. However, it is a far more uncertain procedure to search a body of natural language text to determine whether the concept of the effect of global warming is discussed therein, or to search a picture to see if it portrays a cow. Clearly, there are many ways to describe global warming or to express concern about it, and certainly many ways to picture a cow. The ambiguity or uncertainty that accompanies this kind of searching in collections of information, called *databases*, makes up a large part of the domain of interest of information science. As Sir Walter Scott (1828?) put it, "The end of uncertainty is the death of interest." For example, does "She read her paper on time" refer to the subject matter of the paper or "her" adherence to a schedule? Only context can determine the answer. The sentence "I really liked that" can mean what the words seem literally to say or the exact opposite, depending on tone of voice, if spoken, or the general sense of a written text—is sarcasm used elsewhere? How do we program a computer to recognize sarcasm?

To find information in a large collection of text requires that users understand the vagaries of the text and seek alternative ways of expressing concepts. They must anticipate how authors have represented concepts, not just how they, themselves, would represent them. This, in turn, may require training of users, classification or indexing of the text, and computer programs that can not only search for what the user has requested but help the user decide what to request. See Sec. 7 for a summary of human considerations in the design and use of databases.

Because it is often too expensive or even impossible to store a complete text, graphic, or audio recording, we must often make do with classification codes or other descriptors to convey information about content. In bibliographic work selection of such terms is called cataloguing and indexing. It is crucially important if the original cannot be stored and searched. Today, we are less dependent on indexing for text, still very much so for graphics and sound images.

1.4 Information as a Commodity

Information has been bought and sold, surely for as long as *information* has existed, probably beginning with tips on good hunting and fishing locations and practices. Economists and information scientists have long been fascinated with the fact that, since it is not really the information, but its representation, that is sold, the information remains with the seller after the sale, unlike any other commodity. Hence, scarcity resulting from depletion of stock is not a basic element in determining price. The kinds of information packages that bring the highest prices seem to be those for which interest or demand is high, but access is limited. It is often to the benefit of the purchaser to maintain the access limitation. Thus market research reports, stock market newsletters, and spy reports cost much. Purchasers do not necessarily want their competitors to have equal access. Textbooks and newspapers cost relatively little, partially because a purchaser normally does not object to others owning a copy of the work.

A trial was held in Ontario (Potter, 1988) in which an agent of a labor union was charged with conspiracy to steal a copy of the name and address list of employees of a hotel the union wanted to organize. The judge ruled that no theft could have occurred because the hotel would still have had the information for its own use. What the hotel was potentially deprived of was not information but *exclusivity of access* to that information. Essentially the same logic applies to cases such as the infamous Pentagon Papers episode (Sheehan, 1971), in which classified U.S. government documents were copied. The government was not deprived of the information but of exclusive control of it.

In the context of the cost of access to databases, when such services first became available to the public, in 1972, the basis for charging was almost always *connect time*, the amount of time the user's computer or communications terminal was connected to the database service's computer. Computers had limited numbers of communication ports, and when one of these was in use, it was denied to others. Then, gradually, charges were placed on records or portions of records retrieved, which of course means *copied*. The trend is to charge for a combination of work done and information transferred.

2. DATABASES AND KNOWLEDGE BASES

2.1 General Aspects

Databases are collections of data, with some commonality of content, capable of being searched or processed, and usually but not necessarily containing large amounts of data. A knowledge base is similar, but the implication is that it is used to guide some process, such as searching or interpreting a database. A database may be the set of news articles appearing in the *Washington Post* over a span of 10 years. A knowledge base may be the set of information and rules used to edit a news article from the point of view of English grammatical usage as interpreted by that paper's editors. The difference in concept may not be apparent to a casual observer.

2.2 Record Structure and Content

A *record*, as the term is used here, is a collection of data elements that concern a common *entity*. An entity is the subject of an element of information, such as a laboratory animal or a bank depositor. Information about an entity is represented as a set of *values* of *attributes* or *variables*. Examples of attributes are the *name* of a person, *time* of an event, and *amount* of a bank deposit.

Attribute values are represented as numbers, words, or arbitrary strings of any characters. Attributes also have names, and they have relationships to the entity described. Sometimes, the name adequately describes the relationship. If the entity is a person and the attribute is *date of birth*, that is reasonably self-explanatory. If, however, an attribute is called *pay*, it is not clear whether this is an hourly rate of pay or the amount of a forthcoming check. Therefore, the design of a record should include the list of attributes represented, the manner of representation (integer, character string, etc.) and the relationship of attribute to entity. The record itself then consists of a set of values of the attributes. The name of the attribute, its relationship, and nature of representation could be a part of the record, but they normally are not.

All attribute values for an entity can be stored together or stored separately and brought together when and as needed. In manual systems, those based on paper or other non-computer-readable records, all information related to a given entity is normally stored together, on a single sheet of paper or in a single file folder. In early data-processing systems records took the form of punched cards or were sequences of bits recorded contiguously on a magnetic tape. Today, the information about an entity may be physically distributed over several files on different computers, possibly located in different cities.

One subclass of records is *surrogate records*. A common example is the library catalog card. The card is not itself the book it describes, and it contains virtually none of the content of the book. However, it is descriptive of the book and it represents it in user searches. Search decisions are made using the surrogate, not necessarily the actual document. Gradually, surrogate records, once of vital importance, are becoming less so because it is becoming feasible in many cases to store the entire entity in the record, obviating the need for a surrogate.

2.3 Representation of Attribute Values

We have defined three levels of information representation in a database: the entities about which there is information, the attributes or aspects of an entity about which there will be information, and the representation of the attribute values. Database designers must decide what attributes to represent—whether, for example, a medical bibliographic database should contain only records about research and medical practice or also tutorials and articles for the layperson. Researchers will consider the latter to be clutter, and laypersons will not understand the former. Given a decision about entities, it must be determined what information should be recorded about them. Bibliographic records are fairly standard but typically do *not* contain information about the type of article (research, review, opinion, tutorial, *etc.*), nor do they contain citations used in the article. Finally, the representation of attributes can vary considerably. For example, the subject of an article can be described by a hierarchical classification code, subject headings, or a set of key words extracted from the text. Which is best, which best fits the needs and knowledge of users, must be determined by the designer.

2.4 Storage Structures and Indexes

Users often have no idea how records or individual attributes are stored in computer memory. This condition is called *transparency*: the users see through the physical organization to a *logical organization* or a way of envisioning data that may actually differ from the physical organization but be easier to work with. As an example, text created by a word processor is typically stored in segments of about one line in length, but the segments are not necessarily physically stored in the order in which they appear in the text. Each is likely to be stored in the next space available at the time of storage, wherever on a disk that happens to be. But the users are not required to be concerned about this aspect of storage. They visualize and work with the text as a series of lines in the order of their appearance on the printed page.

Physical organization of data has important effects on the speed, hence cost, of any operation using a database. To explain briefly, there must be some order to records or sets of attributes when stored in memory. Records are put in order on the basis of the value of an attribute, called a *sort key*. If records are so ordered, then it is possible to determine the approximate location of a record, given its key, and thereby to limit search time. Or, if records are stored wherever space is available, then each record will have a *pointer* to the location in memory of the next record in logical (sort key) order. The alternative to ordering is to require a search of every record in a database in order to find just one attribute value.

Suppose a file of employee information is ordered on *name* or *employee number*, but a search is to be done for a physicist who speaks French. Records in this file would be randomly ordered with respect to *job title* and *language spoken*. To avoid an exhaustive search through all employee records, indexes can be created for any attributes likely to be used for searching. An index is created for one attribute. Each index entry consists of the value of the attribute and a pointer to the location of a record containing that value. Then, to find a physicist, the value PHYSICIST is looked up in the job-title index and the set of pointers to records containing that value is retrieved. A similar search for the value FRENCH in an index for languages is then done. Then, the two sets of pointers are combined, retaining only those from both lists, the Boolean algebra AND operation. The user is unaware of the pointer set retrieval and processing—it is transparent. The user is told how many records are represented in the final set and may retrieve any of the indicated records from the employee database.

There are other organizational methods such as the increasingly popular *relational model*. This permits users to use a database as if it contained whatever subset of attributes and in whatever order is convenient for the user at the time of use. The user need not be concerned with whether indexes exist or in what actual order records are stored in the original database. A database management program creates or appears to create the desired structure as needed. This method vastly simplifies computer programming for users of the database, or the composition of search requests, but does add to the operational cost. The economics of modern computer systems tend to encourage this approach, which saves human costs, which continue to rise, at the expense of hardware costs, which continue to decline in terms of unit cost of memory or computational performance. For more information on data structures see COMPUTER DATABASES.

2.5 Database Maintenance and Quality Control

In any information-producing endeavor, aspects such as timeliness, accuracy, reliability and completeness of the corpus of information, or database, are important. It is rare that a database can simply be created and then remain in a constant state. Far more likely, new information will continue to be presented for inclusion, information in the database will change or must be deleted, and errors in existing records will be detected. This means that the managers or operators of any information system must create policies on how to select new information for inclusion, when to delete outdated records, and when and on what basis to make changes. It should be readily appreciated that if a spurious signal causes an error in computing the track of an aircraft in an air-traffic-control system, it must be quickly corrected. By contrast, an error in a bibliographic record would probably not be corrected until a subsequent update cycle, which might occur anywhere from a month to half a year after discovery of the error. This can put users at a disadvantage but the cost of faster change may be prohibitive.

Thus there are some very practical questions: How much error can a database or its users tolerate? What is it worth to users to have a database maintained free of error or, since that is often impossible, to keep them below some threshold? How can error rate or content be measured and monitored in any but the most simple situations? Information science does not have theoretical answers for such questions.

Another aspect of database maintenance is the logic, time, and cost of making changes. The logic is bound up with the physical storage structure (Korth and Silbershatz, 1986). The key elements to consider are two:

1. There is usually a trade-off between optimizing a file structure for speed of search and retrieval versus speed of making changes. A database search service will normally lean toward optimizing search time because that is the dominant operation and the major source of revenue.
2. There is a difference between item-by-item database changes and complete replacement of one version with a new database. The latter is called *reloading* and is costly but sometimes necessary. If the amount of change is great enough, reloading becomes necessary. Either item-by-item change or reloading require the database or a portion of it to be kept out of use for a time, which may vary from fractions of a second to hours.

2.6 Knowledge Bases

In the new field of *artificial intelligence,* or AI, (see ARTIFICIAL INTELLIGENCE), one branch is concerned with *expert systems.* Variously defined, expert systems tend to be information systems that perform tasks traditionally considered performable only by a human intelligence. Of course, there is no strict definition of such tasks and the question of how much human intelligence was required depends on the sophistication of both the performer of the task and the observer. Nonetheless, such systems exist and their number and capabilities increase steadily.

An expert system has to have knowledge in its field, or *domain,* of application approximately equal to that of a human expert. The range of expertise is quite limited. While expert systems have been built or postulated to play chess, diagnose a medical complaint, or search a database and select the best five records meeting the user's needs, no single present-day expert system could conceivably do all these tasks. Indeed, even within the domain of information retrieval, the expert systems tend to be restricted to a single subject area.

The information an expert system uses to perform its task is, in a structural sense, just another database. The difference is that the program using it must be able to interpret the information well enough to base its action thereon. In this case, the database is called a *knowledge base,* which generally consists of a set of data, or facts, and a set of rules (Davis and Lenat, 1982). The facts are used in interpreting the rules. A fact states, for example, that a pawn may move only one or two squares forward and two squares on its first move only, excluding capture situations. A rule tells which piece to move at a given stage in a game and where to move it. A discipline called *knowledge engineering* is developing that is devoted to interrogating human experts to extract from them the information needed for a knowledge base, and then putting this into a meaningful and useful form (Hart, 1992).

The field of expert systems is still quite new and its future uncertain. Some programs have been brilliantly successful—those that can play chess, for example. But there are computer scientists and others who feel there are limits to what can be done, partially on the basis of the fact that while a computer program can interpret information in a knowledge base, it cannot have had the full range of human experience that is often required for difficult decision-making tasks (Weizenbaum, 1976).

3. INFORMATION SYSTEMS

3.1 General Aspects

In a sense, almost anything we might call a *system* can be called an *information system*. A system of equations stores information about the variables represented. A lever might be considered a system that senses the force applied to it and the position of application and transmits information about a resultant force to be applied to a mass. But the more general meaning of *information system* is a set of components capable of acquiring, processing, storing, and transmitting information. The sources from which information is acquired or the destinations to which it is transmitted may be other information systems, people, or even animals. The components usually are assembled for a defined purpose and are interconnected according to a plan. An information system also includes the data or information stored within it. Take two identical computers, loaded with identical programs, but with different sets of data, and we might have quite different information systems.

In the sections to follow, we will consider the major components of typical information systems, how information flows among them, and how that flow is controlled. First, there is a question of the boundaries of a system. For example, are sources part of the system? Are destinations? Are operators (those humans who control the system)? If all these questions are answered in the negative, then the system is a collection of mechanical parts and of information. We often find that, for purposes of evaluation, this definition is too restrictive. To assess an air-traffic-control system as if its human controllers were not an integral part would yield only a moot evaluation. Unfortunately, published descriptions and evaluations of systems are not always clear on specifying what is and is not included.

3.2 Major Components

3.2.1 Sources and Acquisition Subsystems. *Sources* provide information and may control a system's content, reliability, timeliness, etc. The procedure or basis for selecting sources and for deciding what information to ask of or accept from them may also be a part of the system. For a library, sources are publishers, wholesalers, and used-book dealers. The selection decisions are not trivial, because they involve considerations of user needs, discounts offered, reliability in delivery, and the library's desire to spread its custom to assure good service in case of future need. For a newspaper, direct sources are its own reporters or press associations, but *their* sources are varied and cannot be completely controlled by a central office. Questions of reliability arise frequently and cannot be fully answered in most cases. A seismographic information system uses instruments that are part of the system, and hence whose reliability can be periodically verified. It receives information from other systems whose reliability may or may not be open to question.

3.2.2 Processing Subsystems. *Processing* can mean almost any form of transformation of data: conversion of codes, formatting for display, performing mathematical or statistical calculations, or simply changing the order of records (sorting). Information science is typically more concerned with the selection than the processing of data but there is no sharply defined boundary between these functions. Almost any information system performs some form of both.

3.2.3 Communication and Dissemination Subsystems. Modern computer-based information systems typically consist of one or more central computers, or servers, connected with remote users' personal computers. The nature of the connection can be a relatively permanent connection, as in *local-area networks* (LANs), or dial-up connections typical of *wide-area networks* (see NETWORKS, COMPUTER; TELECOMMUNICATIONS). A voice-telephone connection is an example of a dial-up connection—the link to the destination is established on demand and is maintained as long, but only as long, as the user wants it to be.

In information retrieval, *dissemination* has a special meaning not directly related to telecommunication. It is the process by which an information system decides to whom to send records and then performs the physical distribution. This has been an important part of science information systems. It requires the user, or potential recipient, to provide a *profile* or statement of the kind of information that the user would like to receive. Then new records are compared with this statement and those that match are marked for distribution to the person who submitted the profile. Many scientific organizations and individual scientists maintain profiles with the large scientific bibliographic database distribution services. This process is sometimes called *selective dissemination of information* or SDI (Meadow, 1992, pp. 145–146). It is analogous to the older practice of preparing a routing sheet and distributing a copy of a newly arrived journal within an office or laboratory to those who requested it.

3.3 Control of Information Flow

The essence of information science lies in the study of how to control, or what governs the control of, the flow of information from its origin to its final destination. While we can use the phrase *final destination,* it is not final in the sense of reaching a terminal point and going no further. Information generation and use tend to form a giant feedback loop. The recipient of information may be a decision maker who, or which, will then use this information for decision-making purposes over a long period of time. Thus the incoming information serves as a control signal governing outgoing information, possibly in a continuous feedback cycle. Or, information published in a scholarly journal, then retrieved through a retrieval service, may flow to an author and be used in that author's *new* paper, which enters the published literature and is, itself, available to be retrieved in an indefinite cycle.

3.3.1 Control of Input. Information or data can be said to be created when an event occurs and some form of record of the event is made in symbolic form. An event will concern one or more entities. The "event" need not be instantaneous; it could well be a condition of long duration. Colloquially, we say we record the *event,* but we do not and cannot actually do this. We record *information about the event,* that is, we transform a set of perceptions about the event into a set of symbols, and these symbols must be perceived by someone or something. The classical Berkeley question about whether a tree falling unobserved makes a noise may be restated to ask whether a set of symbols constitutes information if it is never perceived. By most modern technical definitions, the answer is *no,* unperceived symbols are not information. Hence the record of the event depends on the perception as well as on the event itself.

Examples of events are a political campaign reported by a journalist into a newspaper or television news system, and the automatic recording of temperatures in a laboratory by a programmed device. In order for either of these reports to be sent, there must previously have been

1. *a policy decision on what kind of events to report*—which, of all possible ones, are wanted in the database;
2. *a design for the record*—the determination of what attributes of an entity are to be recorded and in what manner the values are to be represented;
3. *a selection of events to report*—the determination of what actual events or entities to include. Given the policy, the reality may be different. A newspaper does *not* record all the news; a laboratory sampling procedure may miss a key event.

3.3.2 Selection of Output. Information is selected from a store, not only in information retrieval, but in management information systems and numeric data systems as well. When selection is to be done, there must be some form of language for expressing selection criteria. Most frequently, the manner of selecting is to define a subset of the database in terms of the values of attributes, of the form ATTRIBUTE **RELATION** VALUE, where ATTRIBUTE is the name of an attribute, and **RELATION** is an equivalence relation such as =, >, or CONTAINS (the character string BCD *contains* the string BC). Then Boolean algebraic expressions can be formed from these basic set-defining expressions, producing a final statement such as AUTHOR CONTAINS 'ITH' AND DATE > 1991. Almost any high-order computer-programming language can interpret such expressions (except for CONTAINS, which is relatively rare except in database systems).

3.3.3 Delivery of Output to Users. Computer software must, in any case, interpret such statements as directives for how to examine a record and to determine whether or not it matches selection criteria. The system must then present its information as electronic displays, printed output, or electronic messages to other computers or devices.

3.4 Examples of Information Systems

Here are brief descriptions of three kinds of information systems: a remote-access database search system, a typical accounting system, and a professional organization acting as an information service.

3.4.1 Database Search. Dialog Information Services, Inc., is an example of a service that acquires databases under license from other organizations, called *database producers*, who are analogous to publishers of books or journals. The producers acquire the raw information, put it into usable and standard form, possibly publish it in paper form, and then also distribute it to organizations such as Dialog that will redistribute it electronically to the ultimate user. Dialog provides the computer system, telecommunications for user access, computer software to maintain and search the databases, documentation, training, and user assistance.

The service is highly diversified, offering over 400 databases whose subject matter ranges from *Peterson's College Database*, a directory of educational institutions, to *Standard & Poor's Corporate Descriptions*, providing information about officers, subsidiaries, income, bond ratings, and the like on publicly held corporations, to *Searchable Physics Information Notices (SPIN)*, a bibliographic database describing publications in physics. Users pay for the service based on a combination of time connected to the central computer, the number and form of records retrieved, and in some cases the number of search terms used.

3.4.2 A Typical Accounting System. Every company and institution and almost every family has some kind of accounting system, however primitive, whose basic purpose is to inform management of the financial state of the organization. Clearly, a laboratory with many projects, supported by different sources under different terms, needs a much different system than does a family deriving almost all its income from a single salary and having no unusual expenditures. But the purposes are similar: to account for all monies handled, for tax and other official purposes, and to enable management to make financial plans and adjust them to changing conditions.

An accounting system normally requires the creation of a record for every transaction involving an expense or a commitment, or the receipt of funds. Contrary to the database search system model, all the data are generated internally to the using organization, not acquired *in toto* from an outside source. Controlling input, both assuring that every item is reported and that input values are correct, is a critical responsibility of the system or its managers.

Most accounting systems generate a series of standard reports: a listing of all income and expense items (so that they can be checked for accuracy and so that managers can see where the money is spent), a comparison of each subdivision's expenses compared with spending plans (budget), and a tally of the status of the bank account. But there is also a need for *ad hoc* reports, enabling any manager to request data suited to a particular problem, such as all expenses connected with a particular project, independent of organization, or a time series of expenses in some category to see what changes are taking place.

3.4.3 The American Institute of Physics—An Organization as an Information System. Many organizations, such as professional societies, travel agencies, and stock brokerages, are today recognized primarily as information agencies. They not only operate information systems, they *are* information systems. The American Institute of Physics (AIP) is a federation of professional societies concerned with physics. It is a primary publisher of physics literature, a secondary publisher (indexes and abstracts), a database distributor, and the provider of an online, interactive information service. It is an organization that observes the needs of its users for information; acquires information; selects, processes, and edits it; and distributes it to its users.

As a primary publisher AIP publishes scholarly journals of research, monographs, reference books, conference proceedings, and a newsletter reporting on Washington news of interest to physicists. As a secondary publisher it publishes *General Physics Advance Abstracts* and *Current Physics Index*. As a database producer AIP publishes *Searchable Physics Information Notices* (SPIN), a database derived from *Current Physics Index* that contains bibliographic information and abstracts of articles and other published items. SPIN is then distributed to several database distributors, including Dialog. As a provider of an electronic information service, AIP operates the Physics Information Network (PINET) whose services include electronic mail, society notices of various types, and access to bibliographic databases, including the abstracts of articles in this Encyclopedia. See PHYSICS LITERATURE AND PUBLICATIONS for more detail on publications.

4. INFORMATION RETRIEVAL

4.1 General Aspects and Models

A large part of information science is concerned with *information retrieval*. The purpose of an information storage and retrieval system, or information retrieval system (IRS), is to organize and store information in such a way that it can be retrieved and to perform the retrieval. Retrieval implies selecting a subset of records from the database, according to a specification provided by a user or inquirer. An IRS is an information system that provides

1. a physical storage facility for data or information;
2. an organizational scheme for the information, manifesting both the internal composition of records and the ordering of records and indexes to them; and
3. computer software that interprets queries presented in one or more of a variety of languages, natural or artifical, and carries out the search for records conforming to specifications provided in a formal statement of information need.

There may also be

4. an intermediary function whose purpose is to convert a user's informal statement of information need into the formal version that can be processed by the IRS. The intermediary may be a person trained for this type of task or a computer program or a combination of the two.

There may be

5. assistance in analyzing and evaluating retrieved information,

and there is almost always

6. a subsystem that compiles indexes to records and that gives some indication of content or location.

4.2 Information Needs of Users

Questions to an IRS originate with a user's awareness of a lack of information. Awareness of a lack does not mean that the need is clearly defined. Belkin (1980) called this an *anomalous state of knowledge*, or ASK. A user needing information would approach an information service, perhaps a library or a computer-based IRS, where, in one form or another, it is necessary to explain the *information need*. We distinguish between the ASK and the statement of information need in that the latter is a translation or mapping of the former. This mapping may not be accurate, for a variety of reasons to be discussed in Sec. 4.4.

The user's information need must then be expressed as a *query* or question phrased in the usually formal language of the IRS (see Sec. 4.3). Barring error in the use of language or in the spelling of terms, users usually manage to retrieve *some* records. Then it becomes necessary to evaluate these, or a sample of them, decide whether they are relevant or useful, change the query if necessary, and iterate. The evaluation of record content by the user is presumably based on that person's understanding of the extent to which the record satisfies the ASK. If the needs statement from which the query is derived does not represent the ASK well, then it is unlikely the records will be of great value. Misunderstandings occur easily if the user cannot or does not express his or her actual need. Then there is dissatisfaction with the outcome of the retrieval operation, which could only work from the *expressed* need, not the unexpressed need.

4.3 Query Languages

A query language is the set of symbols and rules for their use by which a user conveys to an information retrieval system what information is wanted and other specifications such as format and order of presentation. It is a form of special-purpose computer-programming language. (See COMPUTER PROGRAMMING LANGUAGES). Most query languages are artificial, using a limited vocabulary and syntax. Many users and system designers feel that, at least for inexperienced users, natural language would be preferable, but this would require that the computer be able to translate natural-language expressions into terms it could use to direct a search. This is being done to a limited extent. Systems that employ this approach do not require full natural-language understanding; they often treat the query as a set of individual words, without consideration of syntax (Salton and McGill, 1983; Croft and Thompson, 1987).

Most programming and query languages are *procedural.* They specify a procedure to be carried out by the issuance of commands, which consist of imperative verbs, objects of verbs, and sometimes prepositional phrases and adverbs. For example, SELECT TITLE FOR AUTHOR=BOHR? might specify that the user wants to see only titles of works represented in a database that were written by an author whose last name (in last-name-first style) begins with the letters B-O-H-R. The "?" asks for retrieval of any author name with these first four letters, followed by any number of others, whether a continuation of the last name or initials. The alternative to procedural language is nonprocedural language, in which the user specifies the characteristics of output, rather than how to produce it. A specification of the format of output would show what information is to appear on a page, and where placed on the page, rather than providing the computer program that formats the display.

The conventional method of entering commands into the computer is to type them on the keyboard. *Menus* and sets of *icons* have become popular recently as a fast means of entering a command. Their use is not an alternative to commands but is a means of entering them. The language in use and the manner in which the computer system interacts with the user in presenting options, receiving commands, and responding to them are major parts of a collection of input–output programs, languages, and display designs called the *user interface,* which is increasingly being recognized as a vital factor in the success of computer usage.

4.4 Intermediation—Assisting the User

The reasons why information retrieval is iterative are that

1. natural-language text is inherently ambiguous and
2. users are uncertain of what they really want.

One reason for the latter uncertainty is that the user has an incomplete mental image of the problem at hand. A second reason is uncertainty about how best to explain or represent the information need. A third is that, to some extent, what information is seen as needed depends upon what is available, and hence a complete and satisfactory query cannot be stated until the database has been explored relative to this user's need.

Given this triple uncertainty, we generally look to a combination of help from an *intermediary* and *feedback* from the system to achieve a desirable outcome. The intermediary converses with the user, the person having the information need, with the objective of learning the user's needs, priorities, evaluation criteria, and even budget so that a query can be formulated. The intermediary then helps the user interpret output and, if necessary, revises the query. The discourse between the two is not unlike that between physician and patient or lawyer and client—the layperson's expression must be translated into the technical terms of a profession. In recent years there have been attempts, with varying degrees of success, to have some aspects of the intermediary role performed by computer software (Marcus, 1991; Marchionini, 1992; Meadow, 1992, pp. 253–269). While progress has been made, it cannot be said that the human role has been replaced.

Feedback is typically provided in several ways. One form is for the system to tell the user what the results of a search were—for example, how many records were retrieved or that a proffered input has a value out of acceptable range. A second is provided by reading the set of records or data retrieved. This, in its raw state, requires that the user who requested the information examine the output, decide whether or not it meets the objective, and decide how to revise the request or commands to the system to improve the outcome. A modern variation of this is to have the system analyze the retrieved data and either help the user to decide what changes may be needed, or directly initiate the next iteration itself. This is an example of the application of artificial intelligence to information retrieval.

4.5 Automatic Processing

One of the areas of greatest accomplishment in information retrieval research has been in the assistance the system can provide to the user—for example, in helping select the search terms that will lead to satisfactory results. Users cannot generally anticipate all the ways authors have chosen to describe a subject. In chemistry, for example, there can be dozens of different ways of representing a compound: as a chemical name, a molecular formula, a ring structure, a popular name, or a trade name, with some of these having many variations. Thesauri can be manually compiled, and word association matrices can be automatically compiled. These show the frequency with which certain words co-occur with others, and this can be done at the level of a sentence, a paragraph, or the entire document. Thus, if a user is interested in the term DISK STORAGE, a word co-occurrence table might suggest other terms that frequently co-occur with either word or both, resulting in an expanded query that might include MAGNETIC DISK, CD-ROM, HARD DISK, FLOPPY DISK, DISK DRIVE, COMPUTER MEMORY etc. Often, even experts in a field will not take the trouble to anticipate all these synonyms, near synonyms, or synonyms in limited contexts.

A second class of assistance techniques is concerned with relationships between records or texts. If a user, by whatever means, has been able to find a few good records, then a useful next step is to ask the system to find any other records that are closely related to these. "Closely related" is taken to mean related in terms of the vocabulary used. A typical function for computing a relationship is the cosine coefficient (Salton and McGill, 1983, p. 121), defined as

$$\cos(R_j,R_k)=\frac{\Sigma\, t_{ij}t_{ik}}{\sqrt{\Sigma\, t_{ij}^2\, \Sigma\, t_{ik}^2}}, \tag{2}$$

where R_j and R_k represent records, whether part of the database or a query, and t_{ij} is the frequency with which term i occurs in record j. The summations range from $i=1$ to n, the number of different terms in the combined documents. Using this measure a user can find a set of records, then ask for the subset most closely related to any given one.

In *relevance feedback* (Salton and McGill, 1983, pp. 140–145), a user retrieves some records and evaluates them. Then, these ratings are used as weights in performing a computation similar to $\cos(R_j,R_k)$, based not only on word co-occurrence but on the importance the particular user attached to a subset of the retrieved records. Other methods for computing relationships or making use of user judgements can be found in Salton and McGill (1983), pp. 199–256.

4.6 Measurement and Evaluation of Retrieval Systems

There is no measure of retrieval system performance that is at once well defined, used in research, and, perhaps the ultimate for a standard, accepted and used in trade. Evaluation is usually based on *relevance*, which can be a measure of *relatedness* between a query and a database record, or of *value* to the requestor. Unfortunately, users often do not have a clear objective in mind when using a retrieval system, leaving us with no clear measure of the degree of accomplishment. The measures most frequently used are *recall* and *precision*. Recall is the fraction of records of a database that are relevant to a query and that are retrieved by it. The difficulty of arriving at a value for recall is that, for a database of thousands to millions of records, it is virtually impossible to know which ones may be relevant but not retrieved. Precision measures the fraction of retrieved records that are relevant. This is a more practical measure, since a user or intermediary can be asked to make the necessary judgments. The difficulty is that a high value of precision is not necessarily the user's goal, although it is often assumed to be.

In practical use, for example, among commercial firms offering retrieval services, user surveys asking overall reaction to the service are more likely to be used and acted upon than measures of outcome or performance.

5. INFORMETRICS

5.1 General Aspects

Informetrics is a branch of information science that is "the study of the quantitative aspects of information in any form." (Tague-Sutcliffe, 1992). It includes patterns of language use, publication, and use of publications. It is related to, but broader than, *bibliometrics,* concerned with the measurement of various "aspects of the production, dissemination and use of recorded information" (Tague-Sutcliffe, 1992) and *scientometrics,* "the study of the quantitative aspects of science as a discipline or economic activity" (Tague-Sutcliffe, 1992). While these phenomena may seem, on the surface, to be almost random, they tend in fact to obey quite regular laws. This is a broad subject. We shall present only a few examples of the types of measures and their applications.

5.2 Language Use

The relationship of the frequency of occurrence of words in a text to the rank, by frequency, tends to follow a very smooth and mathematically simple curve. The relationship was first formalized by Condon (1928) but is commonly attributed to Zipf (1949). It states that $f=kr^{-1}$, where r is the rank, in descending order of word frequency f, and k is a constant, dependent on the sample size. This hyperbolic function is represented by a straight line, with slope -1, on a log-log scale. Most early studies of the phenomenon used "standard" English or French, but Whitehorn and Zipf (1943) also published some samples showing regular deviation from this geometric form, based upon Germanic language texts or English texts written by schizophrenic patients or by children. Mandelbrot (1953) modified it to

$$f=k(r+m)^B, \tag{3}$$

which fits most distributions better. The parameters k, m, and B are attributes of the language sampled or the sample selected. So far, there is little practical application of this relationship in information science, but it may be possible to use the Mandelbrot parameters to characterize languages, including artificial ones, and their users.

Some more traditional measures are the *type-token ratio* and Yule's *characteristic* (Yule 1944). A word *type* is a generic word appearing in a text. A *token* is one specific instance of its appearance. THE is a type, usually the most frequent word in an English text. Any one occurrence of it is a token. The ratio of types to tokens is broadly characteristic of the language or its use by an author. It is the inverse of the mean word frequency. Yule's characteristic is defined as

$$K=(\sigma^2-M)/M^2\,, \tag{4}$$

where σ^2 is the variance and M the mean of the word frequency distribution. Both these measures have been used in authorship studies to help determine the author of an anonymous text or to contest or verify that one person wrote several works attributed to a single person. Other authorship studies have used *discriminant analysis,* which involves finding relatively common words used with different frequencies by different authors (Mosteller and Wallace, 1963).

5.3 Publication Measures

A pressing problem for almost any library is to determine what scientific journals it should have in its collection to meet the demand of its users for maximum coverage of a discipline, while limiting its spending to meet budget limitations. Bradford's law (Kraft and Boyce, 1991) states that if journals in a discipline are ranked according to the number of articles on the subject in question, then organized into zones such that the journals in each zone contribute the same number of papers, then

$$N_k=N_1b^{k-1}, \tag{5}$$

where N_k is the number of journals in the kth zone, N_1 the number in the first zone, and b a *scattering factor* descriptive of and specific to the subject field of the journals. Once some basic investigation has been done in a field, then Bradford's Law can predict the number of journals required for complete coverage or, given an actual list of journals on hand, can predict the cost of an additional increment of coverage or the saving by giving up an increment.

5.4 Usage Measures

We shall illustrate one form of usage measure, *citation* of one publication or work by another, indicating that one author has made use of another's work. A citation is an assertion by an author that the *citing* and *cited* works are related. If two works are cited in the same work, then the two are *co-cited.* If the one work is cited by two others, then we have an instance of ***bibliographic coupling.*** If the two were often co-cited, then the relationship is strong. A person doing a search of bibliographic files and recovering one title might want also to recover all works that are strongly coupled to the first. This information is not contained in conventional bibliographic files or library catalogs. There are *citation indexes* published in science, social science, arts and humanities, and law that provide such information. More information can be found in Kessler (1963), Small (1973), Garfield (1979), and White and McCain (1989).

6. INFORMATION NETWORKS

A network, in the most general sense, is a set of nodes connected by links. The nodes are entities capable of sending or receiving a message, or both. The links are communication channels. The nodes may be computers or people and the links may be telephone lines or a postal delivery system.

From the point of view of information science, information networks can be classified by content and by structure. In terms of content, one major type of network is primarily devoted to sending mail or messages or even complete databases. Generally, these are characterized by *not* having the data permanently stored in the system. Messages are created and sent as needed. They tend to be specifically addressed to recipients, although group addresses are allowed, and they are deleted from the system when delivery is complete. *Electronic-mail* systems are the prime example of these. These "e-mail" systems are the modern version of a telegraph service. Individual users enter their messages and address them. A network relays a message from one node to another until the destination is reached. There are many different methods of controlling the routing. A message is ultimately delivered to a local computer to which the recipient has access and is held there until read. Then the user can delete or retain the message, reply to it, or forward to another node. Bulletin boards and electronic conferencing systems are exceptions to the storage procedure. In these applications, messages may be entered into the system and then stored for viewing by anyone authorized to have access, but reading is generally done after the fashion of reading mail, i.e., by browsing to see what is available and asking for what seems interesting. There is not necessarily a separate copy provided to each reader. Deletion is done by a person responsible for maintaining the files. An extension of this concept allows a user at one node to operate a computer program at another node, obviating the need for transmitting the program or its input data from one machine to another.

The other major content type contains typically large databases, stored relatively permanently in the system, and not for specific users, but for anyone authorized to use the system (See Sec. 3.4.1). Because there are so many records, there must be some kind of content-based search or retrieval process. Permanence of storage is a relative term; records need not stay forever, but until superseded or deleted by an editor. They are not deleted solely because they are read. Again, they have no specific addressee.

Classification by structure or control is typified by two subtypes: *local-area networks* (LAN) and *wide-area networks* (WAN). A LAN consists of a set of computers, printers, and similar equipment, linked together by cable and usually existing within a relatively short distance of each other, limited to a few hundred meters. Wide-area networks may have the components cabled together but typically make extensive use of dial-up telephone links. For more information on networks see COMPUTER NETWORKS and TELECOMMUNICATIONS.

7. THE USERS OF INFORMATION SYSTEMS

Information scientists have come to realize that to study an information system that involves humans as users or operators requires that these human "components" be included in the study. Indeed, it might be said that the domain of information science is those systems that involve human behavior as well as mechanical operation.

Some of the earliest research involving the activity of retrieving information was concerned with how people search in a library (Taylor, 1968) and how cataloguers or indexers perceive the materials they work with, in particular, how consistent they are in their decisions (Lancaster, 1991). Clearly, in any searching situation principally dependent upon surrogate records, the consistency, even more than accuracy, of indexing is critical. How to achieve consistency among indexers or catalogers remains a problem in any discipline.

Searchers using an IRS may be broadly classified into three groups:

1. *professional searchers*, whose function is performing searches on behalf of others, an activity that requires skill in eliciting the information needs of others;
2. *professional users of information*, who are members of a profession or specialized discipline and need information for their professional work; they are also able to judge the value of retrieved information; and
3. *others*, which includes students typically below the graduate level who need information but generally lack the professional's skill at evaluating it.

In general, the second and third groups do not have skill or experience in the use of computer-based information systems, but that statement is gradually becoming less true as home and routine office use of computers spreads. In characterizing users, these three types are not enough. Users vary on knowledge of the mechanical aspects of the system they are using, the language or set of commands available, the content and structure of the database, and the subject matter of the field encompassing the database. Their purpose in searching also affects performance: is it on behalf of another person (implying that the searcher may not know the subject as well as, or the exact information goal of, the end user); is it to find a few good items, or a comprehensive set?

Research has found that these three groups perform differently when doing searches (Marchionini *et al.*, 1991; Cherry and Clinton, 1991), as is to be expected. But, since most information systems do not have significantly different modes of searching to offer users of different levels, the fit between user skill and system design is important. Systems do have differences in user language or methods of providing specifications to the computer. Research suggests that performance differences are at the level of understanding what needs to be done rather than the mechanics of how.

There is not an established set of measures of the performance of information systems or of user performance. This lack hampers research because it is difficult to compare one user group with another. We recognize that there are *performance* measures and *outcome* measures. The former include such variables as time required to perform a task, number of errors (invalid commands issued, for example), or number of commands required to perform a task as meaningful measures.

Studies have investigated relative speeds of work, errors made, commands used, and sequences of commands used. As one would expect, experienced users work faster and make fewer errors. They tend to use more of the commands available to them, analogous to an experienced writer using a larger vocabulary than a novice. Usage studies have tried to identify those commands or output displays that seem to be ambiguous or give trouble to users, and to see how different classes of users respond to an error or ambiguity. They try to determine which types of users prefer to issue commands to systems, and which may prefer to select functions from a menu.

Outcome measures are concerned with results obtained, not with the manner of obtaining them. Outcome measures usually are based on relevance (see Sec. 4.6), essentially measuring the extent to which a system retrieved information useful to or subject-related to that requested. Here is where user skill in evaluating information is involved. Again, as a broad generality, the professional intermediary is not able to judge the value of a record to a user and will instead concentrate on relatedness. The professional user is more likely to judge value, and users in the third category may simply take anything offerred as acceptable.

Some research has focused on *mental models*, essentially the mental image a user has of the system being used. If two users have different perceptions of how a system operates, they will use it differently. Research is becoming more intensive on how to record and compare mental models (Gentner and Stevens, 1983). Continuing research is needed to refine the measures and the laboratory procedures for taking them and to identify the characteristics of users that are of importance. Other aspects of the role of users in information systems have been discussed in Secs. 1.3, 2.2, 3.3, 4.2, and 4.4.

8. SOCIAL AND COMMERICAL ASPECTS

In science, education, and law, and to a lesser extent in business, we have reached the point where users of information have the expectation that all published information can be accessed by anyone who can use the appropriate information networks. If information is not published, which implies lack of an intent to make it public, then it is normally protected by special access restrictions, requiring not only a password, but also membership in some select group, such as those holding a military security clearance. We have not yet reached the long-predicted point where every family and office has a computer tied into the various information networks, but we move closer every year. Here are some current social issues.

8.1 Use of Information Systems

If all published information in a field is available, then teachers, supervisors, and clients expect their students, employees, and consultants to find and use it. Not knowing how is no excuse. This, in turn, creates a need for education and training in how to find and evaluate information, which is not yet covered in the educational programs of most fields. Law is a rare discipline in which knowledge of how to search the literature of the field is a part of professional education. Conceivably, we shall see the day when failure to find information becomes grounds for a charge of professional malpractice. One subfield of information science, overlapping with sociology, studies the patterns of adoption and use of new technology by individuals and groups: who takes the lead and why, who lags and why (Rogers, 1971).

The ease of transmission and copying of files creates concern over copyright protection, not only of the electronic files but also of their printed counterparts. Will readers buy a printed journal or a computer program if a copy can be made electronically at a fraction of the retail cost? The store of information is not depleted by a sale. Should this affect the price? How can the value of information be assessed? For a general review of information economics see Hall (1981).

Privacy and freedom of access to information and media are concepts that existed before electronic media, but have become of more intense interest because of the speed of computers and large amounts of information potentially available. The prospect of virtually all information about each of us being available to anyone—faceless agents of government, insurance or credit companies, prospective employers, or the like—can be daunting. It raises the spectre of "Big Brother," popularized by Orwell (1949). That these fears have not been realized may be partially due to the vigor with which privacy legislation was supported to prevent it (Rosenberg, 1986, Williams, 1988).

Access to media has three aspects:

1. the right to see information about one's self,
2. the right to see information in the hands of government if there is no current reason for maintaining secrecy, and
3. the right to get information *into* a medium.

Rights of access to stored information are now secured by law in many countries. Less well known is the issue of access to media, typified by the right of a political party to free television air time for its candidate in an election. Broadcast frequencies are universally allotted under government control and many governments have rules requiring equal treatment of political viewpoints. The issue is highly sensitive. A similar issue arises in science publishing, when proponents of certain unpopular views claim to be barred from presenting them in major professional publications.

8.2 Growth of the Information Economy

Bell (1976) has pointed out that employment in the primary sectors of the United States' economy, agriculture and extraction, has decreased sharply since about 1900, and "industry" decreased slightly, as a percent of U.S. gross national product. But services and information-related industries have filled the gap. This has led to the terms "post-industrial society" and "the information society." In about 1970 there was virtually no industry devoted to selling databases in electronic or magnetic form. Today, this is a thriving industry. Revenues have grown from about $300 million in 1977 to over $10 billion in 1990 (Information Hotline, 1992). These figures should be treated as approximate, as the definitions of the products vary considerably. Clearly, though, information as a commodity has established itself in world markets. Science information, directly exchanged among scientists, and also science information used in business decision making play a major role in this.

9. SOCIETIES AND JOURNALS IN INFORMATION SCIENCE

The principal professional societies in information science are the American Society for Information Science, the Canadian Association for Information Science, and the Institute of Information Scientists (U.K.). Closely related are the Special Interest Group on Information Retrieval of the Association for Computing Machinery and the Division of Chemical Information of the American Chemical Society. Each holds at least an annual conference.

The major journals are *Journal of the American Society for Information Science* (U.S.), *Information Processing and Management* (U.S.), *Canadian Journal of Information Science, Journal of Information Science* (U.K.), and *Acta Informatica* and *Informatik: Theorie und Praxis der Wissenschaftlich-Technischen Information* (Germany). Some journals closely akin include *Forum of the Association for Computing Machinery Special Interest Group on Research in Information Retrieval* and *Journal of Chemical Information and Computer Sciences* (U.S.), and *Journal of Documentation* and *International Journal of Man-Machine Communication* (U.K.).

GLOSSARY

Anomalous State of Knowledge (ASK): An anomaly in one's state of knowledge, or lack of knowledge, with respect to a problem faced. The start of the information retrieval process.

Attribute: An element of information about or a variable descriptive of an entity.

Data: A *datum* is a value of an attribute. A set of data implies not only a collection of such values but that there is a known organiztion to elements of the set.

Database: A collection of sets of data about entities; or a collection of records, where each is concerned with an entity; or a set of related files, all concerned with a common set of entities.

Entity: The individual about which information is or may be gathered, such as a person, a sensor, or a philosophical concept.

Expert System: An information system capable of, or intended to perform, tasks otherwise considered to be demanding of human intelligence for performance.

Information: (1) Data that is understood or deemed meaningful or valuable by a recipient. (2) $H=-\Sigma p_i \log_2 = p_i$, the sum of the product of the probability of a symbol's occurrence times the logarithm to the base 2 of the probability, over the set of possible symbols.

Information Retrieval: The process of organizing and storing information such that it can be retrieved, and performing the retrieval by selecting a subset of records from the database, according to a specification provided by a user or inquirer.

Informetrics: The study of the quantitative aspects of information in any form; including patterns of language use, the production, dissemination and use of recorded information, and the quantitative aspects of science as a discipline or economic activity.

Knowledge: In general usage, similar in meaning to *information*, but tending to be applied to bodies of information rather than individual messages. Often meant to imply information possessed or accepted by an authority or an authoritative community.

Query: A statement of a user's information need, expressed in the usually formal language of an information retrieval system.

Query Language: The set of symbols and rules for their use by which a user conveys to an information retrieval system what information is wanted, and other specifications such as format and order of presentation. It may be a special-purpose computer programming language.

Relevance: A measure used in the evaluation of information retrieval systems, which can be of the subject *relatedness* between a query and a database record, or of *value* of a retrieved record to the requestor.

List of Works Cited

Belkin, N. J. (1980), *Can. J. Info. Sci.* **5** 133–143.

Bell, D. (1976), *The Coming of Post-Industrial Society*, New York: Basic, pp. 134–137.

Cherry, J. M., Clinton, M. (1991), *Can. J. Info. Sci.* **16**, 2–22.

Condon, E. U. (1928), *Science* **68**, 300.

Croft, W. B., Thompson, R. H. (1987), *J. Am. Soc. Info. Sci.* **38**, 389–404.

Davis, R., Lenat, D. B. (1982), *Knowledge-Based Systems in Artifical Intelligence*, New York: McGraw-Hill.

Garfield, E. (1979), *Citation Indexing—Its Theory and Application in Science, Technology and Humanities*, New York: Wiley.

Gentner, D., Stevens, A. L. (Eds.) (1983), *Mental Models*, Hilldale, NJ: Lawrence Earlbaum.

Hall, K. (1981), *Info. Soc.* **1**, 143–166.

Hart, A. (1992), *Knowledge Acquisition for Expert Systems*, 2nd ed. New York: McGraw-Hill.

Information Hotline (1992) **24**, 1.

Kessler, M. M. (1963), *Am. Documentation* **14**, 10–25.

Korth, H. F., Silbershatz, A. (1986), *Database Systems Concepts*, New York: McGraw-Hill.

Kraft, D. H., Boyce, B. R. (1991), *Operations Research for Libraries and Information Agencies*, San Diego: Academic, p. 115.

Lancaster, F. W. (1991), *Indexing and Abstracting in Theory and Practice*, Champaign: University of Illinois Graduate School of Library and Information Science.

Mandelbrot, B. (1953), in: W. Jackson (Ed.), *Communication Theory*, London: Butterworths, pp. 486–502.

Marchionini, G. (1992), *J. Am. Soc. Info. Sci.* **43**, 156–163.

Marchionini, G., Meadow, C. T., Dwiggins, S., Lin, X., Wang, J., Yuan, W. (1991), *Can. J. Info. Sci.* **16**, 23–59.

Marcus, R. (1991), J.-M. Griffiths (Ed.) in: *Proceedings of the 54th ASIS Annual Meeting* Medford, NJ: Learned Information Inc., (Ed.) **28**, 49–59.

Meadow, C. T. (1992), *Text Information Retrieval Systems*, San Diego: Academic.

Mosteller, F., Wallace, D. L., (1963), *J. Am. Stat. Assoc.* **53**, 275–309.

Orwell, George (1949), *1984: Nineteen Eighty-Four, a Novel*, New York: Harcourt Brace.

Potter, R. B. (1988), *The Globe and Mail* (Toronto), 22 Sept., p. A7.

Rogers, E. M. (1971), *Diffusion of Innovations*, 2nd ed., New York: Free Press.

Rosenberg, R. S. (1986), *Computers and the Information Society*, New York: Wiley.

Salton, G., McGill, M. J. (1983), *Introduction to Information Retrieval*, New York: McGraw-Hill.

Scott, Sir W., *The Heart of Mid-Lothian*, place and publisher unknown. Republished (1965), New York: Holt, Rinehart and Winston, p. 11.

Shannon, C. E., Weaver, W. (1959), *The Mathematical Theory of Communication*, Urbana: University of Illinois Press, p. 19.

Sheehan, N. (1971), *The Pentagon Papers as Published by the New York Times*, New York: Quadrangle.

Small, H. (1973), "Cocitation in the scientific literature: a new measure of the relationship between two documents," *J. Am. Soc. Info. Sci.*, **24**, 265–269.

Tague-Sutcliffe, J. (1992), *Info. Processing Management* **28**, 1–3

Taylor, R. S. (1968), *College Res. Lib.* **29**, 178–194.

Weizenbaum, J. (1976), *Computer Power and Human Reason*, San Francisco: Freeman, pp. 202–227.

White, H. D., McCain, K. W. (1989), "Bibliometrics," in: M. Williams (Ed.), *Annual Review of Information Science and Technology*, Medford, NJ: Learned Information, Inc., Vol. 24, pp. 119–186.

Whitehorn, J. C., Zipf, G. K. (1943), *Arch. Neurol. Psych.* **49**, 831–851.

Williams, S. (1988), in: H. Dyer, G. Tseng (Eds.), *New Horizons for the Information Profession*, Los Angeles: Taylor Graham, pp. 212–225.

Yule, G. U. (1944), *The Statistical Study of Literary Vocabulary*, Cambridge: Cambridge University Press, p. 53.

Zipf, G. K. (1949), *Human Behavior and the Principle of Least Effort*, Reading: Addison-Wesley.

Further Reading

Bell, D. (1976), *The Coming of Post-Industrial Society*, New York: Basic, pp. 134–137.

Garfield, E. (1979), *Citation Indexing—Its Theory and Application in Science, Technology and Humanities*, New York: Wiley.

Hall, K. (1981), "The economic nature of information," *Info. Soc.* **1**, 143–166.

Losee, R. M., Jr. (1990), *The Science of Information, Measurements and Applications*, San Diego: Academic.

Meadow, C. T. (1992), *Text Information Retrieval Systems*, San Diego: Academic.

Mosco, V., Wasko, J. (Eds.) (1988), *The Political Economics of Information*, Madison, WI: University of Wisconsin Press.

Parsaye, K., Chignell, M., Khashafian, S., Wong, H. (1989), *Intelligent Databases, Object Oriented, Deductive Hypermedia Technologies*, New York: Wiley.

Porat, M. U. (1977), *Information Economy: Definition and Measurement*, OT Special Publication 77-12(11), Washington: U.S. Department of Commerce, Office of Telecommunications.

Rosenberg, R. S. (1986), *Computers and the Information Society*, New York: Wiley.

Salton, G. (1989), *Automatic Text Processing*, Reading, MA: Addison-Wesley.

Salton, G., McGill, M. J. (1983), *Introduction to Information Retrieval*, New York: McGraw-Hill.

Schiller, D. (1981), *Telematics and Government*, Norwood, NJ: Ablex.

Shatt, S., Fox, S. (1990), *Voice Data Telecommunications for Business*, Englewood Cliffs, NJ: Prentice-Hall.

Sherman, K. (1990), *Data Communication: A User's Guide*, 3rd ed., Englewood Cliffs, NJ: Prentice-Hall.

Stonebraker, M. (1988), *Readings in Database Systems*, San Mateo, CA: Morgan Kaufmann.

INFRARED SENSORS

See SENSORS, INFRARED

INFRARED SPECTROMETERS

See SPECTROMETERS, INFRARED

INSULATION MATERIALS

See THERMAL AND ELECTRICAL INSULATION MATERIALS

INSULATORS AND SEMICONDUCTORS, CONDUCTIVITY IN

E. M. Conwell and M. A. Abkowitz, *Xerox Webster Research Center, Webster, New York, U.S.A.*

Introduction 42

1. Measurements 44
1.1 Equilibrium Conductivity 44
1.2 Phenomenological Classification of Contact Behavior 44
1.2.1 Preparation of Contacts 45
1.3 Two- and Four-Probe Conductivity Measurements 45
1.4 Mobility Measurements 46
1.5 Measurement of Carrier Concentration 46

2. Band Structure 47
2.1 What Determines Whether a Material is an Insulator, Conductor, or Semiconductor? 47
2.2 Hole Conduction 47
2.3 Energy, Momentum, and the Brillouin Zone 48
2.4 Band Structure of Column-IV and III-V Semiconductors 48

3. States Due to Defects in Crystalline Semiconductors 51
3.1 The Hydrogen Model 51
3.2 Effects at High Impurity Concentration 52

4. Density of Current Carriers 53
4.1 Intrinsic Carrier Concentration 53
4.2 Carrier Concentration in Doped Samples 54
4.2.1 Moderately Doped Samples 54
4.2.2 Heavily Doped Samples 56
4.3 Nonequilibrium Carrier Concentration: Photoconductivity 56
4.3.1 Intrinsic Photoconductivity 56
4.3.2 Extrinsic Photoconductivity 56

5. Formal Theory of Conductivity and the Hall Effect 56
5.1 Effects of Electric and Magnetic Fields on the Carriers 56
5.2 The Relaxation Time 57
5.3 Calculation of dc Conductivity 57
5.4 ac Conductivity 58
5.5 The Hall Effect 58
5.5.1 Low Magnetic Fields 58
5.5.2 The Hall Effect for Two Types of Carriers 58
5.5.3 High Magnetic Fields 59
5.6 Magnetoresistance 59
5.7 Thermoelectric Effect 59
5.8 Galvanomagnetic and Thermomagnetic Effects 60

6. Relaxation Times and Mobilities 60
6.1 Scattering Due to Lattice Vibrations 61
6.1.1 Acoustic and Optical Modes 61
6.1.2 Mobility Due to Acoustic-Mode Scattering 62
6.1.3 Mobility Due to Nonpolar Optical-Mode Scattering 62
6.1.4 Mobility Due to Intervalley Scattering 63
6.1.5 Mobility Due to Polar Optical-Mode Scattering 63
6.2 Impurity Scattering 64
6.3 Neutral-Impurity Scattering 64
6.4 Combination of Mobilities 65
6.5 Comparison with Experimental Data 65

7. High-Field Transport 65
7.1 Hot Carriers 65
7.2 Mobility Changes in High Fields 66
7.3 Impact Ionization 67
7.4 The Gunn Effect 68
7.5 Negative Differential Conductivity 68
7.6 Acoustoelectric Effect 69

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

8. **Impurity-Band Conduction** ... 70
8.1 Semiconductor Conductivity at Low Temperatures 70
8.2 Conduction Mechanisms at Very Low Temperatures 71
8.2.1 Variable-Range Hopping...... 72
8.3 Criterion for Localization..... 72
9. **Amorphous Semiconductors**.. 73
9.1 Atomic Structure 73
9.2 Electronic States 73
9.3 Native Defects................ 74
9.4 Conductivity Mechanisms..... 75
9.5 Doping 76
10. **Two-Dimensional and One-Dimensional Behavior** ... 76
10.1 Quantizing Magnetic Fields ... 77
10.1.1 Landau Levels 77
10.1.2 The Quantum Hall Effect 77
10.2 Transport in Inversion Layers....................... 77
10.2.1 Effect of Electric Field on Semiconductor Surface 77
10.2.2 Quantization in the Inversion Layer........................ 78
10.2.3 Transport in the Inversion Layer........................ 78
10.2.4 Hot Electrons................ 80
10.2.5 Localization.................. 80
10.2.6 Conductance Fluctuations..... 80
Glossary..................... 81
Works Cited................. 82
Further Reading............. 83

INTRODUCTION

An ideal insulator would be one that does not conduct electricity, i.e., carries no current when a voltage is applied across it. In practice, all materials provide a measurable amount of electrical conduction. The effect of a voltage is to set up an electric field **E** that exerts a force on the electrons and ions within the material. In this article we will consider only electronic conduction. Ionic conduction is taken up in the article IONIC CONDUCTION AND DIFFUSION IN IONIC SOLIDS. Electronic conduction in a semiconductor or insulator differs from that in the usual metal in various ways, one being that current may be carried simultaneously by two types of carrier. One of these has the charge $-e$ (1.6×10^{-19} C) expected for an electron, while the other, called the hole, behaves as though it had a charge $+e$. The origin of the holes will be discussed in Sec. 2 when we take up band theory. If the carriers are free to move, the force due to the field results in a net motion, specifically a drift velocity $\mathbf{v}_d$, in the direction of the force. We define the resulting current density **j** as the amount of charge crossing a unit area (perpendicular to the flow) in unit time. Thus if only electrons were conducting we would have

$$\mathbf{j}=ne\mathbf{v}_d, \tag{1}$$

where n is the number of free electrons per unit volume. If there were also hole conduction, that would add a similar term to Eq. (1) for the hole contribution. The drift velocity is a function of the size of the electric field. At low fields $\mathbf{v}_d$ increases linearly with **E** and we may write

$$\mathbf{v}_d=\mu\mathbf{E}, \tag{2}$$

where μ, the mobility, defined by Eq. (2), is a measure of how well the carriers in a particular material respond to the electric field. The mobility becomes a function of **E** at high fields, as will be discussed in later sections. Defining the conductivity κ and resistivity ρ of a material by $\kappa=j/E=1/\rho$, we deduce from Eqs. (1) and (2) that

$$\kappa=ne\mu=1/\rho\ . \tag{3}$$

If there were also hole conduction, κ would have an additional term. The statement that κ and ρ are constants, independent of E, which is generally true at low enough fields, is Ohm's law. What determines the carrier concentration and μ, and how they vary with temperature, electric field intensity, etc., are the major subjects of this article.

Rather arbitrarily, we term a material an *insulator* if its conductivity is of the order of or less than 10^{-8} S m^{-1} ($10^{-10}\,\Omega^{-1}$ cm^{-1}). As shown in Table 1, with this definition the conductivity of insulators covers a range of some 8 orders of magnitude. Good metals, such as silver and copper, have conductivities 16 or more orders of magnitude higher. It is seen from Table 1 that the class of insulators includes a wide variety of materials, mostly organic. Semiconductors also have a wide range of conductivities; these lie between the extremes of metals and insulators just cited. It

Table 1. Representative conductivities.

Material	Conductivity (S/m)
Silver[a]	6.8×10^{7}
Copper[a]	5.8×10^{7}
Germanium (intrinsic)[b]	2.1
Silicon (intrinsic)[b]	3.2×10^{-4}
Nylon (polyamide)	10^{-5}–10^{-6}
Methyl methacrylates (acrylic plastic)	10^{-6}–10^{-7}
Gallium arsenide (intrinsic)[b]	2.4×10^{-7}
Aluminum nitride (undoped)[b]	10^{-9}–10^{-11}
Ceramics[c]	10^{-12}–10^{-14}
Amorphous selenium	10^{-14}
Corning 7911 Glass (96% silica)	$< 10^{-15}$
Paraffin[c]	3×10^{-17}

[a] Gray (1957).
[b] Madelung (1991).
[c] *Handbook of Chemistry and Physics* (1975).

is characteristic of semiconductors and insulators that the conductivity is quite dependent on impurities and defects, and varies strongly with temperature. The values given in Table 1 are those for nominally pure specimens at room temperature. As shown in Table 1, the class of semiconductors also includes a wide variety of materials. Important elemental semiconductors are silicon (see SILICON, CRYSTALLINE; SILICON, AMORPHOUS; SILICON, POLYCRYSTALLINE), extensively used for semiconductor electronic devices (see MICROELECTRONICS), and germanium (*q.v.*). Also important for applications are the compound semiconductors made up by combining an element from column III of the periodic chart [particularly aluminum (Al), gallium (Ga), or indium (In)] with one from column V [particularly phosphorus (P), arsenic (As), or antimony (Sb)]. Of these the most important is gallium arsenide (GaAs), used for electronic devices, light-emitting diodes (*q.v.*), and semiconductor lasers (see LASERS, SEMICONDUCTOR). Compound semiconductors made up by combining an element from column II of the periodic chart [particularly zinc (Zn), cadmium (Cd), or mercury (Hg)] with one from column VI [particularly sulfur (S), selenium (Se), or tellurium (Te)] are also of importance commercially for electroluminescence (*q.v.*), photodetectors (see PHOTODETECTOR DEVICES), and photoconductors. There are numerous other semiconductors—elemental, binary and ternary compounds, etc.

The sensitivity of properties of semiconducting materials to impurities has been taken advantage of by deliberate addition of chosen impurities, the process called doping, to modify conductivity values or to achieve spatial variation of properties. The effects of doping will be discussed extensively in what follows. Another property that significantly affects κ is the structure of the semiconductor sample. The sample may be crystalline, i.e., having a regular structure in which a pattern involving a small number of atoms is repeated identically throughout (see CRYSTALLINE STATE). More typically, the regular structure may occur over large regions of the sample but with the orientation, i.e., the axis, of the structure varying in a random way from one such region to another; such a sample is called polycrystalline. Alternatively, the sample may be amorphous; in the neighborhood of any particular atom the structure may be fairly close to what it is in the crystalline case (short-range order) but this pattern is not repeated, i.e., long-range order does not exist (see Sec. 9 and AMORPHOUS MATERIALS, STRUCTURE OF).

In what follows we first describe how conductivity is measured, emphasizing the role of contacts. What makes a material an insulator or semiconductor rather than a metal was understood early in the century; the explanation is taken up in the section on band structure. That section includes also descriptions of the conduction and valence bands of typical semiconductors. The structure of these bands underlies many features of the conduction process. It is followed by discussions of what determines carrier concentration and mobility in low and high electric fields in typical semiconductors. A major part of the article deals with conduction in column-IV and III-V semiconductors, on which a great deal of work has been done. This is followed by a section on amorphous semiconductors and one on conduction in cases where the carrier motion is restricted to a thin layer (two dimensional) or to a thin strip within the layer (one dimensional). Conduction through layers of different materials, or heterostructures, is taken up in HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR, and in the various articles on semiconductor devices.

1. MEASUREMENTS

1.1 Equilibrium Conductivity

Measurements of electrical conductivity are usually carried out by attaching contacts to the sample to connect it to a dc voltage source, as indicated in Fig. 1(a). To measure the conductivity, one first measures the conductance G, defined as the ratio of the current I through the sample to the voltage drop V across it. Conductance is the reciprocal of the resistance R. For a sample of simple geometry, as indicated in Fig. 1, from the definition of j we find $I=jA$, where A is the cross-sectional area. Also, $V=EL$, where L is the length of the sample. Thus the measured conductance yields the value of κ through

$$G=\frac{I}{V}=\frac{jA}{EL}=\kappa\frac{A}{L}=\frac{1}{\rho}\frac{A}{L}, \tag{4}$$

from the definitions of κ and ρ given earlier.

For the measurements just described to yield a meaningful value of κ or ρ it is apparent that

1. the carrier concentration that exists in the sample in the absence of contacts, and of an applied field, must be retained under the conditions of the measurement, and
2. a uniform electric field must be sustained throughout the sample (with the possible exception of a very thin screening region at the contact–sample interface).

These conditions place severe requirements on the ability of the contacts to supply, or inject, charge, as well as on the sample. They are nevertheless satisfied with relative ease for metallic and high-conductivity semiconductor samples. For low-conductivity semiconductor samples and insulators they are much less likely to be satisfied. In those cases the current that flows in response to a voltage is mainly determined by the nature of the contacts.

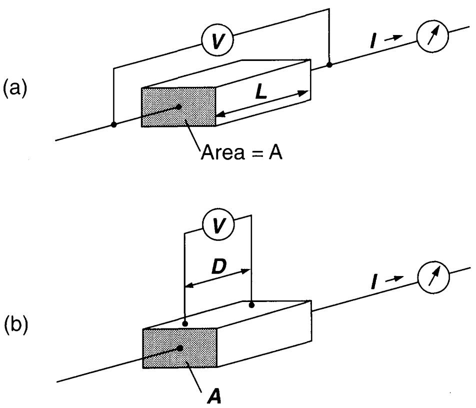

FIG. 1. Arrangements for measuring resistance or conductance. (a) Two-terminal method, (b) four-terminal method.

1.2 Phenomenological Classification of Contact Behavior

Contacts on an insulator or semiconductor may be operationally classified as Ohmic, finite injecting, or blocking. To understand these classifications it is necessary to review the elements of injection theory (see, for example, Lampert and Mark, 1970). We simplify the discussion by assuming that only one type of carrier is present. Consider the fate of a single excess carrier injected by a contact into a solid under the condition that the contact can act as an infinite reservoir of charge. The bulk dielectric relaxation time $\tau_r=\rho\epsilon$, where ϵ is the permittivity, may be defined as the time it takes for charges in the solid to rearrange themselves so as to neutralize any excess injected carrier. In a metal the relaxation is effectively instantaneous, while in a nearly perfect insulator τ_r is essentially infinite. Two regimes need to be considered, depending on the relative values of the injected carrier transit time (time to traverse the region between the contacts under the action of a field) and τ_r:

1. When the transit time is greater than the dielectric relaxation time, the carrier is neutralized before completing a transit, and the current at a given field is controlled in the usual way by the product of the resident thermal-equilibrium carrier population n and the mobility μ. In other words, Ohm's law is valid.
2. When the transit time is less than the dielectric relaxation time, the magnitude of the injected charge is limited only by the repulsion of the space charge already in the sample, including carriers that have been immobilized in traps. Such traps are quite common in high-resistivity materials.

In the simplest case of a trap-free specimen film configured as a plane parallel capacitor with capacitance C, under bias V the amount of charge injected per transit time to sustain this space-charge–limited current is about *CV* (Lampert and Mark, 1970). This injection rate can far exceed the rate required merely to sustain the equilibrium current. Thus, with increasing field, the onset of space-charge–limited injection occurs at the threshold field that just makes the transit time less than the dielectric relaxation time. The onset is marked by *I* increasing more rapidly than linearly with *V*, thus by Ohm's law no longer being valid. It is clear that the onset initiates a process that significantly increases the demand of the bulk on the interfacial source of injected charge. An Ohmic contact is generally defined as one fully capable of sustaining the demands of the bulk up to and including space-charge–limited conditions. A contact that is finite injecting imposes a supply limitation. The contact interface supply limitation can itself depend on field, temperature, and even the mobility of the solid. Thus there are circumstances in which it can be difficult experimentally to distinguish finite injecting from Ohmic contacts. A blocking contact is one that completely impedes injection. Usually a contact is blocking for only one sign of a carrier. Finally it is not uncommon for a given contact to evolve from Ohmic to finite injecting and then blocking with increasing field and for this entire process to be temperature dependent. For further discussion of the behavior of insulators see DIELECTRIC PROPERTIES OF INSULATORS.

1.2.1 Preparation of Contacts. Connection of specimens to measuring circuitry is made by contacts that are most commonly of the following types:

1. diffused;
2. painted (typically with metallic emulsions);
3. soldered, brazed, or spot welded;
4. alloyed;
5. evaporated or sputtered (commonly used on amorphous semiconductors, molecular crystals, and polymer films);
6. plated; or
7. pressure probe.

Diffused contacts are commonly used for semiconductors. A large dopant concentration is introduced by diffusion to bring their surfaces to a nearly metallic state. Metallization of the surface then produces a perfectly injecting electrical interface. It should be noted that contacts that work well on metals are relatively unpredictable on semiconductor surfaces because the interface typically undergoes uncontrolled compositional and thermal modification during the forming process. It is desirable to have current paths parallel to the surface of the specimen, which requires interelectrode spacing large compared to the specimen thickness. For further details on contacts and resistivity measurements, see Dunlap (1959).

1.3 Two- and Four-Probe Conductivity Measurements

The simplest arrangement for the measurement of dc conductivity is the terminal mode shown in Fig. 1(a). Here the voltage drop across the specimen and the current in series with the specimen are simultaneously recorded. It is necessary that the potential drops at the contacts to the specimen, i.e., the current contacts, be small compared to the drop across the specimen. In a typical configuration, the voltage source is an electronically regulated supply, sometimes buffered by a large storage capacitor. Voltage can be measured with an electrometer or electrostatic voltmeter, and current with an electrometer or picoammeter. When isolation or noise considerations are paramount, a shielded battery box still makes the best dc supply.

For high-resistivity materials, or cases where the contact potential drops are relatively large, measurements should be carried out in a four-terminal configuration, illustrated in Fig. 1(b). In this configuration the voltage drops across the contacts are unimportant because they are in series with much higher resistances in the voltage-measuring circuit. One such mode of measurement using a linear array of pin probes (commercially available) is commonly deployed for rapid routine evaluation of Si and Ge semiconductor specimens, or other materials with volume resistivities in the range 10^{-5} to about 10^{2} Ω m. Voltage measurements are taken across the innermost probes with both polarities of a constant-current source and averaged to eliminate extraneous thermal emf's or rectifying contact effects. If the probe spacing *D* is small compared to the smallest sample dimension and the contact area is negligible, the conductivity is given by $I/2\pi VD$. For high-resistivity materials, up to about 10^{9} Ω m, a particular four-probe configuration due to van der Pauw, which is commercially available, may be used (Dunlap, 1959). For $\rho > 10^9$ Ω m the two conditions specified in Sec. 1.1 are not likely to be satisfied. Values of $\rho > 10^9$ Ω m are probably qualitative at best.

1.4 Mobility Measurements

In addition to measuring conductivity or resistivity, it is possible in many materials to measure directly the mobility of the current carriers. The mobility may be obtained by measuring the transit time t_t over a known distance d of injected carriers in a known electric field. The drift velocity is $v_d = d/t_t$. Once v_d is measured, the mobility may be obtained from Eq. (2). There are basically two methods for measuring the mobility, one applicable to semiconductors, the other to insulators.

In the case of semiconductors the technique is applicable to a specimen in which one sign of carrier, either electrons or holes, is dominant under thermal-equilibrium conditions. This carrier is called the majority carrier. The technique relies on the fact that a suitable point contact applied to a semiconductor sample can inject or emit minority carriers into the sample for one sign of voltage relative to the sample, while it can collect, and thus detect, them for the reverse sign of voltage. The sample is fabricated as a uniform cylindrical rod and fitted with blocking contacts at each end. A dc voltage is applied between these contacts. Emitter and collector point contacts are located along the rod spaced a distance d apart. A pulse of minority carriers, typically of microsecond duration, is emitted into the sample at some time t_1 and drifts along the rod under the applied field, reaching the collector point contact at time t_2. The concentration of injected carriers is kept much less than the thermal-equilibrium carrier concentration to ensure that the transit occurs in a uniform field. The carrier mobility is then given by $\mu = d/E(t_2 - t_1)$, where t_1 and t_2 are determined by monitoring the voltage pulses induced across a sensing resistor in series with the specimen.

In the case of insulators, carriers are injected by means of light (see Sec. 9). The sample is prepared as a film of thickness d with semitransparent blocking contacts at the front and back through which the field $E = V/d$ is applied. Exposing the samples, through a contact, to a pulse of strongly absorbed light creates holes and electrons just below the contact. A sheet of either electrons or holes will be swept into the bulk, depending on the direction of the electric field, and through the sample to the other contact. To ensure that the transit occurs in a uniform field the amount of photoinjected charge must be kept small, below $0.1CV$ where C is the sample capacitance. The transiting carrier sheet induces a voltage across a series resistor that drops when the sheet exits through the other contact. The time between the creation of the sheet and its exit is the transit time t_t provided it is long compared to the duration of the light pulse, a condition that can usually be satisfied. Again, the mobility is given by $\mu = d/Et_t$.

If there are traps in the sample, μ measured as described above, called the drift mobility, will not agree with the conductivity mobility defined in Eq. (2). The conductivity mobility is a microscopic mobility, the mobility of an untrapped particle. For the case of shallow traps, where a carrier may be trapped and released from a trap many times during a transit, the drift mobility is the microscopic mobility times the fraction of time that the carrier is free. When the traps are not shallow, and there is a range of trap depths and thus a range of trapping times, the carriers straggle to the contact and it becomes impossible to define a transit time. Such dispersive transport is frequently seen for amorphous semiconductors and disordered materials.

1.5 Measurement of Carrier Concentration

Another important measurement for semiconductors is the Hall effect, from which the carrier concentration of samples with one type of carrier predominant may be determined. For good accuracy the sample should be long compared to the transverse dimensions. An electric field E_x applied along the long dimension, to be taken as x, sets up a current j_x. When a magnetic field is applied in a transverse direction, to be taken as z, it deflects the carriers in the y direction, whether toward $+y$ or $-y$ depending on their sign. During the initial transient phase a current flows in the y direction, causing charge to accumulate on the two faces perpendicular to y. This process continues until the field E_y due to this charge just balances the deflecting effect of the magnetic field. In the steady state the current again flows in the x direction. It is found that E_y is proportional to j_x and H_z, with the constant of proportionality being R_{H}, the Hall coefficient. The importance of the Hall effect lies in the fact that

$$R_{\mathrm{H}}=r/ne \quad \text{or} \quad R_{\mathrm{H}}=r/pe\,, \tag{5}$$

the proportionality constant r being of the order of unity, as will be shown in Sec. 5. Note that the sign of R_{H} is given by the sign of e. It is seen from Eqs. (3) and (5) that the product $R_{\mathrm{H}}\kappa$ equals $r\mu$. It is usual to call the product $R_{\mathrm{H}}\kappa$ the Hall mobility, μ_{H}.

2. BAND STRUCTURE

2.1 What Determines Whether a Material Is an Insulator, Conductor, or Semiconductor?

It is well known that the electrons in an atom or molecule occupy a set of discrete, well-separated energy levels (see ELECTRONIC STRUCTURE OF ATOMS AND MOLECULES). When atoms or molecules are brought close together to form a solid, electrons on different atoms or molecules interact and the energy levels spread into bands (see ELECTRONIC STRUCTURE OF SOLIDS). The deeper-level electrons, most tightly bound to the nuclei, interact least, forming low-lying narrow bands, well separated from each other. The electrons in the highest level (or levels) go into wide bands. These bands of allowed levels may be separated from each other, may overlap each other, or may overlap bands coming from levels ordinarily not filled in the atom.

It might be thought that, with the outermost electrons close to each other, they could move freely from one atom to another. All solids should then be conductors. The band picture given above can explain why this is not so. An electric field exerts a force on an electron, tending to increase its energy. This can only occur if there are unfilled levels nearby for the electron to move into. Thus if the uppermost band occupied by electrons were incompletely filled, or if it overlapped unfilled bands, the material would be a good conductor. This situation describes metals. It follows from these arguments that if the uppermost occupied band were completely filled, and did not overlap any unfilled bands, the material would not conduct. How then does one explain semiconductors? The band picture offers two possibilities. If the energy gap between the uppermost filled band, called the valence band, and the next higher unfilled band, called the conduction band, is small enough, electrons may be thermally excited into the unfilled band. This process will be discussed further in Sec. 4. This type of semiconductor is called an intrinsic semiconductor. The second possibility is that impurities or defects in the material give rise to energy levels in the energy gap between valence and conduction bands. Ordinarily electrons in these levels are bound or localized and cannot give rise to current. However, they can, at finite temperatures, produce unfilled bands in either of two ways. In one way an electron in a level close to the conduction band is excited into the conduction band by thermal energy. In the other way an electron from the valence band is thermally excited into an empty level close to the valence band. Both of these situations will be discussed further in Sec. 4. As a by-product of this discussion it is clear why some materials are insulators. An insulator is a material with a large gap between valence band and conduction band, typically greater than several electron volts. In materials with such large gaps it is not usual to find impurity levels close to the bands.

2.2 Hole Conduction

It has been found, e.g., by measurements of the Hall effect, that some metals and semiconductors act as though positive charges, rather than electrons, were the current carriers. Explanation of this apparent anomaly was another success of the band model. From the discussion in the last section it follows that semiconductors are characterized by a small fraction of filled states in the conduction band and/or a small fraction of unfilled states in the valence band. It is the latter states, called holes, that act like positive charges. This can be understood as follows. A filled band carries no current because every electron motion in a particular direction is canceled by an exactly opposite motion. When one electron is removed from the band the cancellation is no longer exact. In the presence of a field the net motion of all the remaining electrons is opposite to the field direction because of their negative charge. This results in the hole having a net motion in the field direction, like a positive charge. In fact, it can be shown that the behavior of all the remaining electrons in the band is equivalent to the behavior of a singly positively charged particle. The properties of this particle, the hole, can be deduced from the properties of the missing electron to which it corresponds. It is clear from its definition that the minimum energy for a hole occurs when it is at the top of the valence band.

2.3 Energy, Momentum, and the Brillouin Zone

The energy E of a free electron is given by $p^2/2m_0$, where p is its momentum and m_0 its mass. The electron may also be characterized by a wave vector k, which equals $p/\hbar$, $\hbar$ being Planck's constant, 6.63×10^{-34} J s, divided by 2π. An electron inside a crystal, which has a periodic arrangement of atoms, may still be characterized by a p value, or a k value. (Note, however, that it is no longer true that momentum is a constant as it is for a free electron.) However, unlike the free-electron case, for the electron in the crystal there is only a finite, although very closely spaced, set of allowed values of these quantities. That set is equal in number to the number of atoms in the crystal and is confined to a unit cell in k space. This unit cell is called the Brillouin zone (see ELECTRON STRUCTURE OF SOLIDS). The detailed boundaries, and thus the shape, of the Brillouin zone depend on the detailed geometry of the crystal lattice. For the diamond structure, possessed by Si and Ge, and the zinc-blende structure, which is the structure of GaAs, InSb, most of the other III-V compounds, and some of the II-VI compounds, the Brillouin zone is a truncated octahedron, shown in Fig. 2.

2.4 Band Structure of Column-IV and III-V Semiconductors

Although many features of the band structure may be obtained from experiments, complete determination of the band structure for any particular material requires solution of the wave equation for that material (see ELECTRON STRUCTURE OF SOLIDS). The result of the band-structure calculation is a detailed relation between electron energy E and k. Because we have confined the possible **k** values to the Brillouin zone, there is for each **k** a set of E values, each corresponding to a different band. The significant energy bands for conduction are the valence and conduction bands, and in fact only the upper and lower edges of these bands, respectively, for low electric fields.

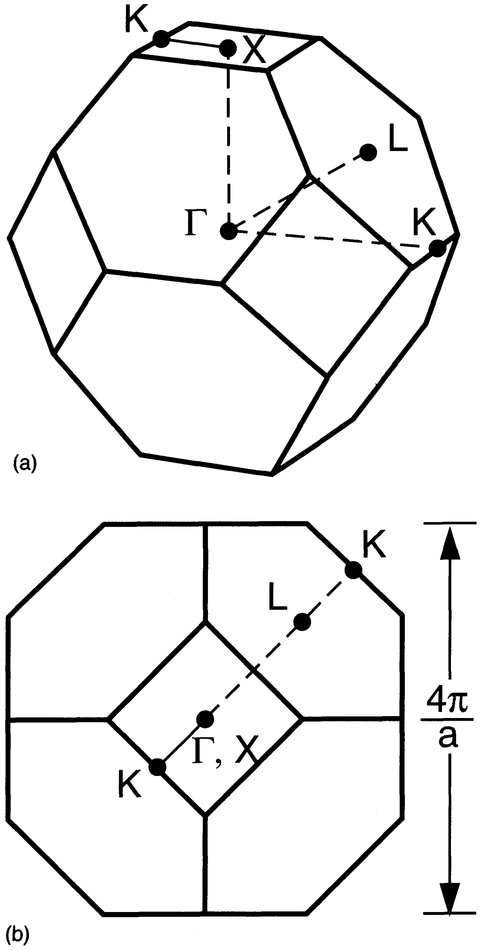

FIG. 2. The Brillouin zone and several symmetry points for the diamond and zinc-blende lattices. The view at the bottom is along a [110] (face diagonal) axis; that at the top has been tilted. The locations of the points shown are $\Gamma = (0,0,0)$; $X = (2\pi/a)(1,0,0)$; $L = (2\pi/a)(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$; $K = (2\pi/a)(\frac{3}{4}, \frac{3}{4}, 0)$, where a is the side of a cubic unit cell and the numbers in parentheses are the components of **k** along the cubic x, y, and z axes. (After Harrison, 1980).

It is instructive to consider what some possible results of band-structure calculations could be for the column-IV and III-V zinc-blende semiconductors before going on to the actual results for these materials. In Fig. 3 we show some possible results for the band edges.

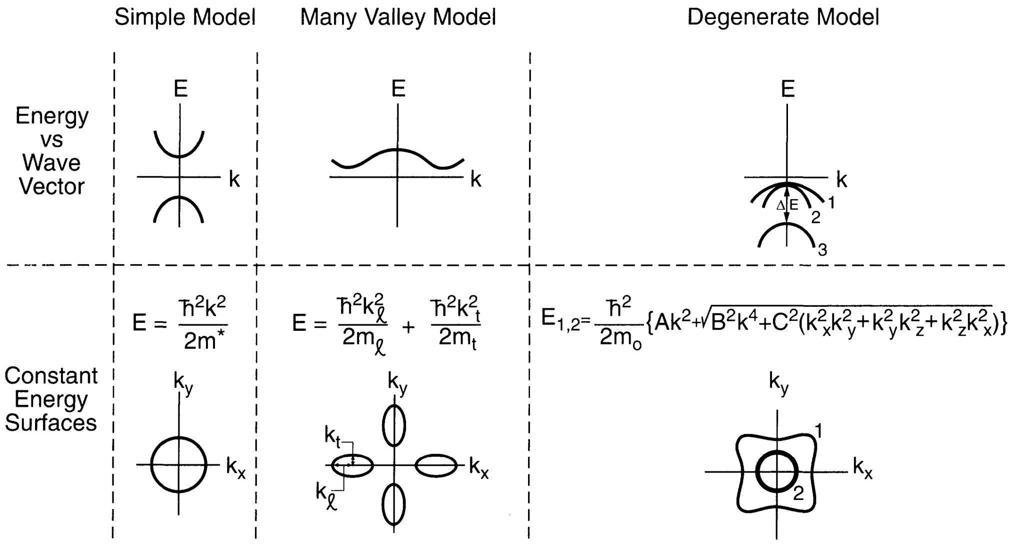
Simple Model
Many Valley Model
Degenerate Model
Energy vs Wave Vector
E
k
ΔE
1
2
3
Constant Energy Surfaces
$E = \frac{\hbar^2 k^2}{2m^*}$
$E = \frac{\hbar^2 k_\ell^2}{2m_\ell} + \frac{\hbar^2 k_t^2}{2m_t}$
$E_{1,2} = \frac{\hbar^2}{2m_0}\{Ak^2 \pm \sqrt{B^2k^4 + C^2(k_x^2k_y^2 + k_y^2k_z^2 + k_z^2k_x^2)}\}$
k_y
k_x
k_t
k_ℓ

FIG. 3. *E* vs **k** and constant-energy surfaces for some possible semiconductor band structures. (Adapted from Conwell, 1958).

In the first column is the so-called simple model, where the energy measured from the edge of the band is a constant times k^2. *E* vs *k* for electrons is then a parabola as in the free-electron case and an inverted parabola for holes. To maintain the analogy between this model and the free-electron case, the proportionality constant between *E* and k^2 is taken as $\hbar^2/2m^*$, the m^* thus defined being called the effective mass of the carrier in the crystal. In the simple model the same parabola for *E* vs *k*, thus the same m^*, is obtained for all directions of **k**. This type of band structure could occur if the band edge were at the center of the Brillouin zone, the Γ point. Another way of representing this result is shown at the bottom of the first column. If all **k**'s belonging to the same *E* are connected, a single sphere is obtained, just like that for a free electron. As a result of this simple relation between *E* and **k**, for this model of the band structure it is possible for most purposes to treat the electron inside the crystal as a free electron, using, however, a different number for the mass.

When the minimum energy is not at the center of the zone there is a further complication. Symmetry requires that all equivalent **k**'s i.e., all **k**'s reachable from each other by crystal symmetry operations (rotations, reflections, or inversions that transform the crystal into itself), have the same energy. Thus the band edge or position of minimum energy must be multiple, and the electron (or hole) population must be divided among the various minima, or valleys. This band structure is therefore called the many-valley model. Within each valley, not far from the minimum energy, *E* can be represented as a quadratic function of the components of **k** measured from the minimum energy point. Thus the points in **k** space belonging to a particular energy are on a set of ellipsoids, one for each minimum. Actually, these are ellipsoids of revolution here, and so the energy can be written as a constant times k_l^2 plus another constant times k_t^2, where k_l and k_t are longitudinal and transverse components. In Fig. 3 the lower half of the second column illustrates the case for the minimum energy located at a point on a cube axis, in which case there are six equivalent valleys. This is the case for the conduction band of silicon. Because $m_l \neq m_t$, we must deal with two effective masses in this case. In later sections we will discuss how this type of band structure complicates transport.

A third possible type of band structure for the semiconductors we consider is the degenerate band structure shown in the third column of Fig. 3 for the case of the valence band. This is a set of three bands, one split off by an energy ΔE from the other two. This is the structure for the top of the valence band in Si, Ge, and the III-V compounds with zinc-blende structure. It is the result of the wave function at the top of the valence band having within each cell the character of an atomic p state, which is triply degenerate. The interaction of spin and orbital motion of the electron partially breaks up the degeneracy of the p states, leaving two states stuck together at $k=0$ and the third split off by a small amount ΔE (see ELECTRON STRUCTURE OF SOLIDS).

As is the case for the many-valley model, E vs $\mathbf{k}$ for the degenerate model is different in different directions. This results in the constant-energy surfaces being warped rather than spherical. The surfaces from the two upper bands belonging to a given energy are shown, again rather schematically, in the lower part of the third column of Fig. 3; their equation is given also. The values of A, B, and C are different for different materials. Because of the warping of the surfaces, the effective masses are functions of direction. The anisotropy is not very great, however, and for most purposes the holes in bands 1 and 2 may be described by constant masses. The band with the smaller curvature has the larger mass. Holes in this band are called heavy holes, while those in the other band are called light holes. At temperatures low enough that the thermal energy $kT \ll \Delta E$, only these two bands will be populated.

The calculated energy-band structure for GaAs is shown in Fig. 4. Note that many features of the band structure have been verified by other calculations and by experiments. The band structures of the other column-IV and the other III-V zinc-blende semiconductors are quite similar to that in Fig. 4. (For a tabulation of band structures and other data for all the column-IV and III-V compounds see Madelung, 1991). Plotted in this figure is E vs $\mathbf{k}$ of a number of bands for some high-symmetry directions in the Brillouin zone. Thus, referring to Fig. 2, we see that the curves from Γ to X represent the cube axis direction and those from Γ to L the cube diagonal direction, etc. Looking at the borders of the crosshatched region, which represent the lowest conduction band and the highest valence band, we see that the only case where the simple model comes close to being realized is the conduction band of GaAs, centered at Γ_6. The mass for that band is quite small, $\cong 0.06m_0$, where m_0 is the free-electron mass.

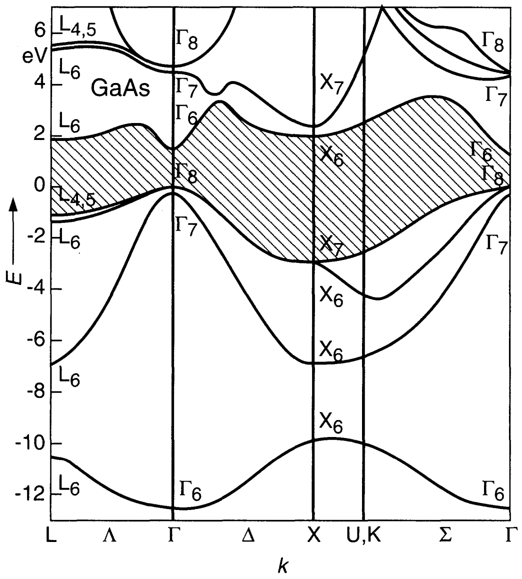

FIG. 4. Calculated band structure of GaAs. The crosshatched region represents the gap between the highest valence band and the lowest conduction band. (Adapted from Madelung, 1991.)

In the case of Si the conduction-band minimum is along the cube axis direction, actually close to the X point. Thus, as noted earlier, the conduction band of Si has six equivalent valleys, the situation depicted in Fig. 3. The ellipsoids are quite anisotropic, with $m_l = 0.92m_0$, $m_t = 0.19m_0$. The valence band of GaAs clearly fits the degenerate model of Fig. 3 with a spin-orbit splitting $\Delta E = 0.34$ eV. Light- and heavy-hole masses are well approximated by $0.08m_0$ and $0.05m_0$ (Madelung, 1991). The valence band of silicon also has the degenerate band structure, but the spin-orbit splitting is only 0.04 eV, so small that all three bands are populated at room temperature. Light- and heavy-hole masses are $0.15m_0$ and $0.54m_0$.

The band structures of semiconductors other than those in group IV and III-V compounds have not been studied as extensively. Regardless of the detailed lattice structure, band edges should still be described by the simple, many-valley, and degenerate models. Of course differences must be expected due to different symmetry when the lattice structure is not cubic.

3. STATES DUE TO DEFECTS IN CRYSTALLINE SEMICONDUCTORS

As indicated in the last section, impurities or defects give rise to energy levels in the forbidden gap that localize electrons. There is a great deal of information about such levels in silicon (see SEMICONDUCTORS, ELEMENTAL—ELECTRONIC PROPERTIES) and GaAs, for example (see SEMICONDUCTORS, COMPOUND—ELECTRONIC PROPERTIES). In this article we limit discussion to those impurities that are important for conduction. Specifically, we will discuss those that

1. provide energy levels close to the conduction- or valence-band edges, and
2. can be introduced into the semiconductor in relatively large concentrations.

The impurities that best satisfy these conditions are those that substitute for host atoms in the lattice and differ from the host in having one more or less valence electron. For Ge and Si these are the atoms of columns III and V of the periodic chart. For the III-V compounds such impurities include atoms from columns II, IV, and VI. As will be seen, the last electron in an impurity with one more electron than the atom it substitutes for is weakly bound to the impurity, i.e., occupies a level only a little below the conduction band. This last electron can easily be detached by thermal energy, going into the conduction band. Such an impurity is called a donor. By similar reasoning, an impurity with one less electron than the atom it substitutes for provides a level just above the valence band into which a valence electron can be excited by means of thermal energy, leaving a hole in the valence band. Such an impurity is called an acceptor. We show now that both donors and acceptors can be described by a simple intuitive picture that predicts reasonably well the energy levels they provide.

3.1 The Hydrogen Model

Consider as an example the case of a column-V donor atom, say P, substituting for a Si atom. P has five valence electrons. Four of these go into covalent bonds with the four nearest neighbors just as the electrons on a Si atom would. This results in a P^+ ion that attracts the fifth electron. This fifth electron uses wave functions from the conduction band. If the conduction-band structure were that of the simple model, the electron would move around the P^+ ion essentially as an electron in free space moves around the proton in a hydrogen atom, or more generally around a nucleus of charge Ze. The theory of such motion is well understood (see ELECTRONIC STRUCTURE OF ATOMS AND MOLECULES) It is found that the most probable distance a_H of the electron from the nucleus is given by $a_H = \hbar^2/Zm_0e^2$. The binding energy E_b of the electron to such a nucleus is $E_b = m_0e^4Z^2/2\hbar^2$. To utilize these results for the electron bound to a substitutional P^+ ion in Si it is necessary to take into account two important differences between the situations. Because the P^+ ion is embedded in a medium of permittivity ϵ, the attractive potential of the ion for the electron when they are separated by a distance r is reduced from e^2/r to $e^2/\epsilon r$. as though the nuclear charge $Z = 1/\epsilon$. The other difference arises from the band structure. In the case of the simple model (see Fig. 3) the free-electron mass m_0 may be replaced by m^*. For the simple model of the band structure, a_H then becomes

$$a_H^* = \varepsilon\hbar^2/m^*e^2, \tag{6}$$

and the binding energy becomes

$$E_b^* = m^*e^4/2\varepsilon^2\hbar^2. \tag{7}$$

E_b^* in this case represents the distance of the energy level below the conduction-band edge.

For typical semiconductors ε is large, ~12 to 16. This results in a_H^* being large and E_b^* quite small. For Si, for example, with $\varepsilon = 12$, by Eq. (6) the binding energy of the fifth electron to a P^+ ion would be $0.1m^*/m_0$ eV if the conduction-band structure were the simple model of Fig. 3. The measured value of this binding energy is 0.045 eV (Madelung, 1991), close to the value that would be obtained from Eq. (6) by inserting for m^* the geometric mean mass $(m_t^2 m_l)^{1/3}$ for a conduction electron in Si. Thus the simple H model we have used gives a reasonably good value for the binding energy. A detailed calculation for the many-valley model is a great deal more complicated [see Altarelli and Bassani (1982)]. Equations (6) and (7) should be valid for the appropriate impurities in GaAs, where the simple model of the band structure does apply reasonably well for the conduction band. Some impurities for which the theory should be valid are column-IV impurities—C, Si, and Ge—substituting for a Ga atom, and column-VI impurities—S and Se—substituting for an As atom. With $\varepsilon=12.4$ for GaAs, and $m^*=0.06m_0$, Eq. (7) leads to $E_b^*=5.3$ meV. This is in good agreement with the observed value of 6 meV for all the impurities listed but S, for which the observed E_b^* is closer to 7 meV. Refinements in the theory bring the theoretical values close to the experimental values (see Altarelli and Bassani, 1982).

In the case of a substitutional impurity with one less electron than the atom it replaces, when the four covalent bonds with nearest neighbors are filled, it results in a net charge of $-e$ in the region. This negative charge attracts a hole, which can be treated just as the bound electron was in the previous case. The most probable distance of the hole from the impurity is given by Eq. (6), and the binding energy of the hole by Eq. (7), with m^* the hole effective mass. Note that in this case E_b^* equals the distance above the valence band of the electronic energy level provided by the acceptor. Again the actual calculation is more complicated because of the complications of the band structure, in this case the degenerate model for the semiconductors of most interest. The measured binding energies for holes arising from the substitution for Si atoms of column-III acceptors—B, Al, Ga, and In—range from 40 to 150 meV. For holes arising from column-II elements—Be, Mg, Zn, and Cd—substituting for Ga in GaAs, the measured binding energies range from 28 to 35 meV. The tendency of the binding energies to increase with increasing atomic number is due to the fact, not acknowledged in the hydrogen model, that the bound electron or hole does actually spend some time close to the impurity, where ε has less than its full value.

The energy required to release a bound electron or hole from a donor or acceptor, respectively, is also called the activation energy because it is the energy required to give rise to conductivity. The activation energies cited for hydrogen-model donors and acceptors are typical for such donors and acceptors in the other column-IV and III-V semiconductors (see Madelung, 1991). They are also the smallest activation energies; other impurities may also supply conduction-band electrons or valence-band holes, but greater energy is required to achieve this. The activation energies for the hydrogen-model impurities cited above are generally <25 meV, the thermal energy at room temperature, with the consequence that the bound electrons and holes will be mostly or entirely free at room temperature and even below. The detailed variation of carrier concentration with temperature to which they give rise will be taken up in the next section.

3.2 Effects at High Impurity Concentration

In this section so far we have considered the activation for a single isolated impurity. When there is a high concentration of impurities, so that the wave functions of bound electrons or holes begin to overlap, the activation energy decreases. In a process similar to that described earlier for band formation, as the impurities come closer the energy levels spread into bands. Band formation occurs first for excited electronic levels, which we have not considered so far in this section, because their wave functions extend further. When the hydrogen model is strictly applicable the excited levels are analogous to those of the H atom (see ELECTRONIC STRUCTURE OF ATOMS AND MOLECULES). Otherwise they are different in detail. Overlap of the higher excited levels to form bands, which then overlap (and mix with) the conduction band, clearly decreases the activation energy, one of a number of effects that do so. For impurity concentration large enough that the ground states of adjacent impurities overlap, an impurity band is formed. We can estimate the impurity concentration N_I at which this occurs by equating a_{H}^* of Eq. (6) to $N_I^{-1/3}$, a rough measure of the average distance between impurities. For the conduction band of GaAs this yields $N_I \simeq 10^{21}\ \mathrm{m}^{-3}$ for formation of an impurity band. By this criterion a concentration several times as large would be required for impurity-band formation in Si or Ge. Such concentrations are not far below the maximum that can be incorporated into these semiconductor lattices for any impurity. The impurity band that results would be expected to have somewhat different properties from the bands we have been discussing because, apart from its being quite narrow at concentrations of 10^{21} or $10^{22}\ \mathrm{m}^{-3}$, the impurities are distributed randomly rather than on a regular lattice. Impurity-band conduction will be taken up in Sec. 8.

4. DENSITY OF CURRENT CARRIERS

A pure, perfect semiconductor crystal at the absolute zero of temperature would have no electrons in the conduction band and no holes in the valence band. The current carriers present at higher temperatures are excited both from the valence band and from donor and acceptor impurities by thermal energy. In what follows we will consider first the thermal-equilibrium carrier concentration in pure specimens, the so-called intrinsic carrier concentration, and then in doped or extrinsic specimens. The section will end with an introduction to nonequilibrium carrier concentrations. These lead to enhanced conductivity and other phenomena of great technological importance in semiconductors and insulators. The nonequilibrium carriers may be excited by light, in which case the enhanced conductivity is called photoconductivity. They may also be created by injection at contacts, as discussed in Sec. 1, and by ionizing radiation.

4.1 Intrinsic Carrier Concentration

The probability $f(E)$ of an electron occupying a state of energy E is given by the Fermi–Dirac distribution function (see STATISTICAL MECHANICS, QUANTUM)

$$f(E)=\{1+\exp[(E-E_F)/kT]\}^{-1}, \tag{8}$$

where E_F is the electrochemical potential, called the Fermi level in discussions of solids. The number of electrons per unit volume in the conduction band of a semiconductor is given by

$$n=\int f(E)N(E)dE, \tag{9}$$

where $N(E)dE$ is the number of available states per unit volume of the sample in the energy range dE at E, measured from the band edge, and the integration is taken over the conduction band. For the simple model of the band structure the states in the conduction band up to energy E occupy a sphere in k space of radius $k=[2m^*(E-E_c)/\hbar^2]^{1/2}$, where E_c is the energy of the conduction-band edge. From this, plus the fact (due to the uncertainty principle) that each state must occupy a volume of k space equal to $(2\pi)^3$/(volume of the sample), we deduce that

$$N(E)dE=(1/2\pi^2)(2m^*/\hbar^2)^{3/2} \times(E-E_c)^{1/2}dE. \tag{10}$$

This result may be formally generalized to other models by replacing m^* by $m^{(N)}$, called the density-of-states effective mass. For the many-valley model

$$m^{(N)}=g^{2/3}(m_t^2m_l)^{1/3}, \tag{11}$$

where g is the number of equivalent valleys.

We may now obtain n as a function of temperature by inserting Eqs. (8) and (10) in Eq. (9). Except for very heavily doped semiconductors, $E-E_F \gg kT$, and Eq. (9) may be approximated by the Maxwell–Boltzmann distribution

$$f(E)=\exp[(E_F-E)/kT]. \tag{12}$$

This allows the integration to be simplified by taking the limits from 0 to ∞. The result is

$$n=N_c\exp[-(E_c-E_F)/kT] \equiv 2(2\pi m_n^{(N)}kT/h^2)^{3/2}\exp[-(E_c-E_F)/kT], \tag{13}$$

which serves to define N_c, the effective density of states in the conduction band. Similar calculations may be made for holes in the valence band, starting from the probability $1-f(E)$ of finding a hole in the level at energy E. The result is that the hole concentration p varies with temperature according to

$$p=N_v\exp[-(E_F-E_v)/kT] \equiv 2(2\pi m_p^{(N)}kT/h^2)^{3/2}\exp[-(E_F-E_v)/kT], \tag{14}$$

which defines N_v, the effective density of states in the valence band. Because of the approximations made in the derivation, the expressions for n and p are accurate only when E_F is at least several kT from the respective band edge. As will be seen, this condition is always met for intrinsic samples. Multiplying together the expressions (13) and (14) we obtain the equilibrium value for the product np. For an intrinsic sample, in which the number of carriers is denoted n_i, $np=n_i^2$. We can use Eqs. (13) and (14) to obtain

$$n_i=p_i=(N_cN_v)^{1/2}e^{-E_g/2kT}. \tag{15}$$

Thus the intrinsic carrier concentration depends exponentially on the energy gap $E_g/$

$2kT$. Note that E_g is also somewhat temperature dependent, decreasing ~5% from 0 to 300 K for Si, for example. Using the fact that E_g for Si is 1.12 eV at 300 K (Madelung, 1991) and the effective masses given previously, we find that $n_i = 1\times10^{16}\ \mathrm{m}^{-3}$ ($1\times10^{10}\ \mathrm{cm}^{-3}$) at 300 K for Si. For GaAs, with a gap of 1.4 eV at 300 K, n_i is only $2\times10^{12}\ \mathrm{m}^{-3}$ at that temperature. Particularly in the case of GaAs the intrinsic number of carriers is so small that carriers coming from impurities, which cannot be reduced to such a low level, must always be dominant.

We can use the equations derived in this section to determine the location of E_F in intrinsic material, E_F^i. To do this it is convenient to use as our reference point, i.e., the zero of energy, the top of the valence band. Equating n_i obtained from Eq. (13) to p_i obtained from Eq. (14) we find

$$E_\mathrm{F}^i = (E_g/2) + \tfrac{3}{4}kT\ln(m_p^{(N)}/m_n^{(N)}). \qquad (16)$$

If $m_p^{(N)} = m_n^{(N)}$ the Fermi level for intrinsic material is at midgap, independent of temperature. Although this is generally not the case, E_F^i is usually not far from midgap because $m_p^{(N)}$ is usually relatively close in value to $m_n^{(N)}$.

4.2 Carrier Concentration in Doped Samples

4.2.1 Moderately Doped Samples. To illustrate the way in which carrier concentration varies with temperature in a doped sample, we will discuss the case of a donor-doped, or n-type, sample for specificity. The results apply to acceptor-doped samples with the obvious modifications. In the present section we discuss samples for which Maxwell–Boltzmann statistics, Eq. (12), are valid, typically samples with $N_d < 10^{22}\ \mathrm{m}^{-3}$. Actually, because it is not possible to control impurity content totally, the typical n-type sample with N_d donors/m^3 will also have some acceptors, present in concentration N_a/m^3. We will assume, as shown in Fig. 5, that the N_d donors all provide the same energy level E_d, while the N_a acceptors all provide the energy level E_a. In crystalline semiconductors that are not as well under control as column-IV and III-V semiconductors there may be different sets of donor and acceptor levels, or even a continuous range of levels across parts of the energy gap; the results we derive in this section will not apply in detail to such materials. As seen in Fig. 5, the energy required to raise an electron from a donor level to the conduction band, i.e., the electron binding energy, is $E_c - E_d$, while that required to raise it from the valence-band edge to the acceptor level, the hole binding energy, is $E_a - E_v$. Given N_d, N_a, N_c, N_v, and the various energy levels, it is in principle possible to determine the variation of n with T from Eq. (13), provided E_F is known as a function of T. A procedure for determining $E_\mathrm{F}(T)$ will be given later. It is useful to precede this with a qualitative view of the situation.

Let n_d represent the number of electrons in donor levels. In the limit of low temperatures ($kT \ll E_c - E_d$), n is negligible and $n_d = N_d - N_a$, i.e., there are N_a empty donors because electrons on the donors can lower their energy by going into the empty acceptor levels. For a moderately doped sample with $N_a < N_d/2$, a typical situation, the majority of the donor levels are filled. Consequently the probability of an electron having energy E_d, $f(E_d)$, must be not much less than unity and, by Eq. (8), E_F must be within a few kT of E_d. With E_F close to E_d, the number of electrons in the conduction band, according to Eq. (13), is given by

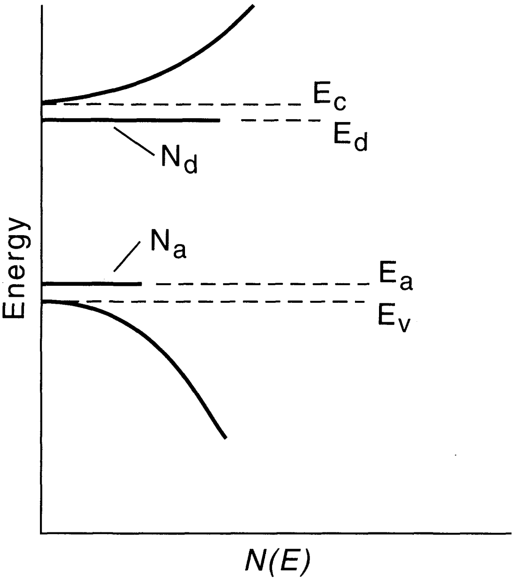

FIG. 5. Distribution of states in a lightly doped n-type crystalline semiconductor. The variation of $N(E)$ with E is given by Eq. (10).

$$n \simeq N_c \exp[-(E_c - E_d)/kT] \quad \text{for } kT \ll E_c - E_d. \qquad (17)$$

Note that relation (17) is frequently used for the experimental determination of $E_c - E_d$, or the activation energy. It is valid more generally than for the model of Fig. 5, requiring only that other donor and acceptor levels be more than several kT below E_d. What is required for its use is the observation of a range of temperatures in which ln n increases linearly with $1/T$.

When $E_c - E_d$ is small, as is true for many of the semiconductors of interest, the exponential increase in n with T predicted by Eq. (17) occurs over only a small temperature range. As the temperature is increased further, more electrons are excited to the conduction band. The conduction band is preferred over the donor levels in the competition for electrons because of its greater density of states. With more electrons in the conduction band E_F must move down; the total number of disposable electrons, $N_d - N_a$, is kept constant by reducing the number in the donor levels. There follows a temperature range in which, as E_F is moving down, there is substantially complete ionization of the donor levels, and $n \simeq N_d - N_a$. This temperature range starts as soon as E_F has fallen several kT below E_d, and is called the saturation range. With further increase in temperature E_F continues to drop until it reaches a position a few kT above midgap. When this occurs, holes begin to be produced, indicating the approach of intrinsic conductivity. Further increase in temperature brings E_F close to E_F of Eq. (16), thus essentially to the intrinsic condition.

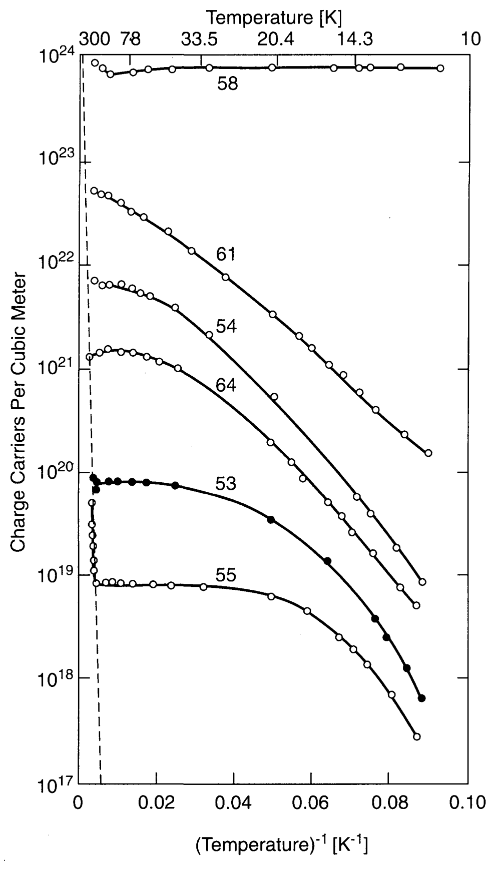

FIG. 6. Density of charge carriers vs the reciprocal of temperature for a set of As-doped Ge samples identified by numbers 55,53,.... The dashed line represents the density of intrinsic carriers. (Adapted from Debye and Conwell, 1954.)

Experimental data (obtained from Hall measurements) for n vs $1/T$ of a set of Ge samples doped with different amounts of As are shown in Fig. 6. In accordance with the above results it is seen that for all but the most heavily doped sample n is small and increases exponentially with $1/T$ at low temperatures. The increase becomes less rapid as T grows, and finally n reaches a saturation value, $N_d - N_a$, when essentially all the donors are ionized. With further increase in T the sample goes into the intrinsic range, n and p both increasing rapidly. For more heavily doped (or small-gap semiconductor) samples the intrinsic range may set in before saturation is achieved. The case of the most heavily doped sample, number 58, will be taken up in the next section.

A systematic procedure for finding E_F as a function of T is based on the requirement for electrical neutrality: the sum of the positive charges in the sample must equal the sum of the negative charges. The statement of this requirement for the situation of Fig. 5 is

$$p + p_a + N_d = n + n_d + N_a, \qquad (18)$$

where p_a is the number of holes in acceptor levels. From the definition of $f(E)$, $p_a = N_a[1 - f(E_a)]$ and $n_d = N_d f(E_d)$. With these expressions for p_a and N_a, and Eqs. (13) and (14), all quantities in Eq. (18) that depend on E_F are expressed in terms of E_F. Equation (18) may then be solved graphically for E_F for any set of N_d and N_a at any desired temperature (Shockley, 1950).

4.2.2 Heavily Doped Samples. For $N_d > 10^{24}$ m^{-3} it is found, as seen in the most heavily doped sample of Fig. 6, that $n \simeq N_d$ and essentially constant up to room temperature. (N_a should be negligible for such high doping.) This is the behavior of a metal. Material with such high doping is called degenerate because the carriers obey degenerate or Fermi–Dirac statistics, Eq. (8) rather than Eq. (12). The constancy of n with T means E_F must be constant with T in this case. This requires that E_F be at least several kT above the conduction-band edge. An exact description of the situation is not easily made because at such high concentration impurity levels overlap and mix with the conduction band. For further discussion see Sec. 8.

4.3 Nonequilibrium Carrier Concentration: Photoconductivity

4.3.1 Intrinsic Photoconductivity. In a semiconductor a photon of energy greater than the energy gap may raise an electron from the valence band to the conduction band, creating an electron and a hole (see Ryvkin, 1964). In some materials, particularly at low temperatures, the attraction of the electron and hole causes them to be bound to each other, the bound entity being called an exciton (see EXCITONS). In many materials, however, the electron and hole are free and may add to the conductivity. The number of electrons or holes generated per unit time per unit volume is given by the product of the number of photons (of energy greater than E_g) absorbed per unit time per unit volume and the probability β that an absorbed photon will create an electron–hole pair. β is called the quantum yield and in many cases has the value unity. For a thin sample and a wavelength of light not strongly absorbed, the number of photons absorbed per unit time per unit volume is the product of the number I incident on unit area of the surface per unit time and α, the fraction absorbed per unit distance the light travels in the sample. When the fraction of photons absorbed is not small it is necessary to allow for the decrease in the number as the light travels.

When a sample is illuminated with a light source of constant I it is found that the number of electrons and holes grows with time, at first rapidly and then more slowly, reaching a steady value or saturation. The rate of growth decreases because excess carriers tend to recombine. Saturation occurs when the recombination rate equals the generation rate. We can characterize the recombination rate by introducing average lifetimes τ_n and τ_p for electrons and holes, respectively. Using the generation rate $\beta\alpha I$ obtained above, we may write for the steady-state densities of electrons and holes $\beta\alpha I\tau_n$ and $\beta\alpha I\tau_p$, respectively. The lifetimes vary widely, typically ranging from 10^{-2} s to fractions of a microsecond.

Electrons and holes may recombine directly, i.e., an electron falls into a hole in the valence band, in which case $\tau_n = \tau_p$. The energy given up in this process, essentially E_g, may be in the form of heat (specifically phonons; see Sec. 6.1) or radiation. Radiative recombination is the process that results in light-emitting diodes (*q.v.*) or semiconductor lasers (see LASERS, SEMICONDUCTOR). The alternative to direct recombination is sequential capture of an electron and a hole, or vice versa, by a defect. Such defects are called recombination centers. When these centers mediate the recombination process, τ_n may be quite different from τ_p.

4.3.2 Extrinsic Photoconductivity. In a semiconductor with electrons in a donor level $E_c - E_d$ below the conduction band, a photon of energy $E_c - E_d$ or greater can raise electrons to the conduction band, resulting in a conductivity increase. Similarly, photons could excite electrons from the valence band into empty acceptor levels in the gap, resulting in an increase in hole concentration. The resulting increase in κ can be used for photodetection. (see PHOTODETECTOR DEVICES). The small amount of energy required to free an electron or a hole from a hydrogenic center in Ge or Si, for example, makes these and other semiconductors useful for infrared detectors.

5. FORMAL THEORY OF CONDUCTIVITY AND THE HALL EFFECT

5.1 Effects of Electric and Magnetic Fields on the Carriers

The force on an electron or hole due to an electric field **E** and a magnetic field **H** is given by

$$\mathbf{F}=e\mathbf{E}+e\mathbf{v}\times\mathbf{H}\,, \tag{19}$$

where e is negative for electrons, positive for holes, and **v** is the velocity of the electron or hole. The velocity of an electron or hole in a lattice is shown by band theory (see ELECTRONIC TRANSPORT IN SOLIDS) to be

$$\mathbf{v}=\nabla_p E=(1/\hbar)\nabla_k E\,. \tag{20}$$

For the simple model of band structure, with E given in Fig. 3, Eq. (20) leads to $\mathbf{v}=\mathbf{P}/m^*=\hbar\mathbf{k}/m^*$, the free-electron result with m_0 replaced by m^*. Thus **v** is parallel to **k**. For the many-valley model that need not be the case. By definition of the ∇ operator, **v** must be perpendicular to the constant-energy surfaces. Only for **k** along one of the ellipsoid axes will **v** be parallel to **k**.

As is expected by analogy with Newton's second law, the effect of a force on an electron or hole in a lattice is to change its momentum, i.e.,

$$\mathbf{F}=\frac{d\mathbf{P}}{dt}=\hbar\frac{d\mathbf{k}}{dt}\,. \tag{21}$$

5.2 The Relaxation Time

In the absence of an electric field the carriers are moving with random thermal velocities, but of course the net velocity vanishes when the velocities of all the carriers are summed up. In the presence of the field each carrier experiences the force $\mathbf{E}e$, which accelerates it and increases its velocity in the direction of the force. If the lattice were perfect the increase would continue indefinitely and the conductivity would be infinite. Of course the lattice is not perfect and the acceleration is interrupted on the average after a time τ when the carrier is scattered by, or collides with, a defect. We will assume, as turns out to be the case for most types of defect, that after each collision, i.e., after the time τ, the carrier has forgotten the acceleration it gained from the field and the acceleration process starts anew. We may therefore take the rate of loss of momentum due to collisions to be

$$\frac{d\mathbf{P}}{dt}=-\frac{\mathbf{P}}{\tau}\,. \tag{22}$$

Integration of this equation shows that excess momentum decays exponentially as $\exp(-t/\tau)$. τ is therefore called the momentum relaxation time. In the remainder of this section we will derive expressions for the conductivity, Hall effect, and magnetoresistance in terms of τ. Section 6 will be devoted to calculating τ for the most important scatterers: lattice atoms out of place as a result of their thermal vibrations, and foreign ions or atoms.

5.3 Calculation of dc Conductivity

Combining Eqs. (19) and (22) we obtain for the rate of momentum change of a single carrier

$$\frac{d\mathbf{P}}{dt}=e\mathbf{E}+e\mathbf{v}\times\mathbf{H}-\frac{\mathbf{P}}{\tau}\,. \tag{23}$$

Let us assume for the moment that all carriers have the same τ, i.e., that τ is not a function of the energy of the carrier. We may then sum Eq. (23) over all the carriers and obtain, for the simple model of the band structure,

$$\frac{d(m^*\mathbf{v}_d)}{dt}=e\mathbf{E}+e\mathbf{v}_d\times\mathbf{H}-\frac{m^*\mathbf{v}_d}{\tau}\,, \tag{24}$$

$\mathbf{v}_d$ being the drift velocity. For the steady state in a constant dc field $d(m^*\mathbf{v}_d)/dt=\mathbf{0}$. In the absence of a magnetic field Eq. (24) then gives

$$\mathbf{v}_d=(e/m^*)\tau\mathbf{E}\,. \tag{25}$$

With this equation and the definition, Eq. (2), of the mobility μ, we find for the simple model

$$\mu=(e/m^*)\tau\,, \tag{26}$$

and the dc conductivity is given by

$$\kappa_0=(ne^2/m^*)\tau\,. \tag{27}$$

Allowing for the variation of τ with energy E, we find with a more accurate theory (see, for example, Herring, 1955, or Conwell, 1982) that τ in Eqs. (25)–(27) is replaced by the weighted average $\langle E\tau\rangle/\langle E\rangle$.

For the many-valley model μ is a tensor. But the semiconductor crystals of greatest interest are cubic, and for low electric fields the mobility must be the same in all directions. To obtain the correct μ we must therefore average over the different valleys. The result is to replace $1/m^*$ in Eqs. (25)–(27) by an average inertial mass

$$1/m^{(I)}=\tfrac{1}{3}\,(2/m_t+1/m_l) \tag{28}$$

when the constant-energy surfaces are ellipsoids of revolution.

5.4 ac Conductivity

Consider now the case where the electric field applied to the sample is oscillatory, i.e., $\mathbf{E}=\mathbf{E}_0\, e^{i\omega t}$, where ω is the angular frequency. Such a field sets up an oscillatory current with the same frequency and thus a drift velocity $\mathbf{v}_d=\mathbf{v}_d^0\, e^{i\omega t}$. When these expressions for $\mathbf{E}$ and $\mathbf{v}_d$ are inserted into Eq. (24) with $\mathbf{H}=\mathbf{0}$, and the solution rationalized, we find the ac conductivity

$$\kappa(\omega)=\frac{ne^2\tau/m^*}{1+\omega^2\tau^2}-i\omega\,\frac{ne^2\tau^2/m^*}{1+\omega^2\tau^2}\ . \tag{29}$$

The conductivity being complex reflects the fact that, at all frequencies for which $\omega\tau$ is not negligible compared to unity, the carriers can no longer follow the field. The current then consists of a part in phase with the applied field, the first term, and a part 90° out of phase, the second term. Comparing with Eq. (27) we see that the in-phase conductivity is κ_0 reduced by the factor $(1+\omega^2\tau^2)^{-1}$. By inserting the current $\kappa(\omega)\mathbf{E}_0e^{i\omega t}$ into Maxwell's equations it is found that $1/\omega$ times the imaginary part of $\kappa(\omega)$ represents a contribution to the permittivity ε. This free-carrier contribution to ε is negative because it is 180° out of phase with the displacement current. For n and τ large and $\omega\tau$ not too large, this contribution may be large enough to make ε negative. Again, when τ is allowed to be a function of E the quantities involving τ must be replaced by averages. For the many-valley model $1/m^*$ is replaced by $1/m^{(I)}$ defined in Eq. (28).

5.5 The Hall Effect

5.5.1 Low Magnetic Fields. From Eq. (24), using Eqs. (1) and (27), we find for the steady state of electrons with the simple model of the band structure in a magnetic field

$$\mathbf{E}=(\mathbf{j}/\kappa_0)-(1/ne)\mathbf{j}\times\mathbf{H}\ . \tag{30}$$

If we assume the conditions of a Hall-effect experiment—$\mathbf{H}=H_z$, applied field $\mathbf{E}_x$, and current flowing only in the x direction—we find from Eq. (30) to terms linear in H that $E_x=j_x/\kappa_0$ and

$$E_y=(1/ne)j_xH_z,\quad \tau\neq\tau(E)\ . \tag{31}$$

According to Eq. (31) the Hall constant R_H, defined in Sec. 1.5 as E_y/j_xH_z, is

$$R_\mathrm{H}=-1/n|e|\ \text{for electrons,}\quad \tau\neq\tau(E)\ ; \tag{32}$$

$$R_\mathrm{H}=1/p|e|\ \text{for holes,}\quad \tau\neq\tau(E)\ . \tag{33}$$

If we allow τ to be a function of energy, the quantity $R_\mathrm{H}ne$ or $R_\mathrm{H}pe$ is no longer unity but is given by μ_H/μ, according to Sec. 1.5. Calculations (for example, Herring, 1955, or Conwell, 1982) lead to

$$\frac{\mu_\mathrm{H}}{\mu}=\frac{\langle E\tau^2\rangle\langle E\rangle}{\langle E\tau\rangle^2}\,B\,, \tag{34}$$

where B depends on the shape of the constant-energy surfaces. Such a dependence is to be expected because a magnetic field, which cannot change the energy of a carrier, causes the carriers to rotate around the constant-energy surfaces. $B=1$ for spherical constant-energy surfaces and is generally of the order of unity. The other factor in μ_H/μ will be seen to lie between 1 and 2 for the important scattering mechanisms.

5.5.2 The Hall Effect for Two Types of Carriers. There are two situations in which we must deal with more than one type of carrier. In intrinsic or nearly intrinsic samples both electrons and holes are present. In a semiconductor with degenerate bands we must deal with sets of electrons or holes with different effective masses or, more generally, different effective-mass tensors. It is usually a good approximation to treat the different types as independent. When an electric field is applied without a magnetic field, j and κ are given by the sum of the contributions of each kind of carrier regardless of sign. When a magnetic field is applied in addition to the electric field, this is no longer the case. Holes drifting in an electric field are pushed by a magnetic field in the same direction as electrons because both v_d and e are opposite in sign to those of electrons. The resulting Hall field is then the difference of the two fields.

Quite generally, for two types of carriers, denoted by subscripts 1 and 2, it is found that

$$R_\mathrm{H}=\frac{n_1e_1\mu_1\mu_{\mathrm{H}1}+n_2e_2\mu_2\mu_{\mathrm{H}2}}{(n_1|e_1|\mu_1+n_2|e_2|\mu_2)^2}\ . \tag{35}$$

If the two carriers are opposite in sign, the Hall voltage will go through zero when

$$n_1\mu_1\mu_{H1}=n_2\mu_2\mu_{H2}\ . \tag{36}$$

For the semiconductors we are discussing this will happen when p is somewhat greater than n because μ_n is generally larger than μ_p. If there are two carriers present with the same sign, as is generally the case for p-type column IV and III-V semiconductors, the ratio of R_H to the total carrier concentration may be increased by as much as a factor of 2 because μ of the light holes is much greater than μ of the heavy holes.

5.5.3 High Magnetic Fields. In the limit of high magnetic fields, although not so large that quantum effects occur, it is found theoretically that R_H goes to the value $1/ne$ or $1/pe$. Thus μ_H/μ is unity at high fields independent of the details of band structure and energy dependence of τ. For a sample with both electrons and holes, R_H in the high-field limit goes to $1/(n-p)e$.

As will be seen in Sec. 5.6, a high magnetic field is specified by the condition $\mu H\gg 1$. Thus, for example, the high-field condition is satisfied at lower values of **H** for light holes than for heavy holes. This can explain the fact that R_H for p-Ge, measured at a typical field (3000 Oe), was observed to decrease by a factor of 1.6 as T was decreased from 300 to 78 K. Clearly this decrease could not represent an increase in p with decreasing T. Nor, as will be seen, could it be accounted for by a change in $\tau(E)$. The explanation is that, with decreasing T, μ increases for both light and heavy holes. The light holes, with their higher μ, go into the high-**H** regime at a higher temperature than the heavy holes, decreasing their value of μ_H. As can be seen from Eq. (35) this decreases R_H considerably.

It can be concluded from this section that the carrier concentration deduced from measurements of the Hall effect is uncertain within about a factor of 2 unless one knows a great deal about the carrier mobilities and the detailed scattering mechanisms.

5.6 Magnetoresistance

It is readily seen from Eq. (30) that for the simple model with $\tau\neq\tau(E)$ there is no change in resistance in a magnetic field, i.e., no magnetoresistance. The resistivity may be calculated from the component of **E** along **j** divided by **j**, i.e., $\rho=\mathbf{j}\cdot\mathbf{E}/|j|^2$. According to Eq. (30) $\rho=1/\kappa_0$, independent of **H**. Magnetoresistance is a result of $\tau=\tau(E)$ or of nonspherical constant-energy surfaces.

To understand the situation for $\tau=\tau(E)$ we solve Eq. (30) for **j**, with the result

$$\mathbf{j}=\frac{\kappa_0\mathbf{E}}{1+\mu^2H^2}+\frac{\kappa_0\mu(\mathbf{E}\times\mathbf{H})}{1+\mu^2H^2}\ . \tag{37}$$

We have simplified the situation by assuming that $\mathbf{E}\perp\mathbf{H}$. For $\mathbf{E}=E_x$, $\mathbf{H}=H_z$, the first term on the right-hand side of Eq. (37) represents j_x, the second j_y. It is seen that

$$j_y=\mu Hj_x=(He/m^*)\tau j_x. \tag{38}$$

This equation expresses the fact that j_y is produced by rotation of j_x in the magnetic field. The quantity He/m^*, called the cyclotron frequency, is the rate at which carriers rotate around the constant-energy surfaces in a magnetic field, while τ is the mean time to undergo a collision, which redirects the carrier and effectively interrupts the rotation. Because carriers with different τ have different j_y, it is not possible to find an electric field that will keep currents due to carriers with different τ's in the same direction. The net current is therefore reduced, making the apparent resistance of the medium larger.

Because the magnetic field dependence of **j** occurs in the form μH, a small magnetic field is defined by $\mu H\ll 1$. From Eq. (37) it is seen that for small μH, where $(1+\mu^2H^2)^{-1}$ may be expanded as a power series in μ^2H^2, $\Delta\rho$ increases as H^2. It is also seen that, in the limit $\mu H\gg 1$, ρ does not keep on increasing, but saturates.

For ellipsoidal constant-energy surfaces the magnetoresistance is a function of the orientation and shape of the constant-energy ellipsoids. As a result, measurements of magnetoresistance may make it possible to determine the location and parameters of the ellipsoids in a many-valley structure. For detailed calculations of magnetoresistance for the many-valley model see Herring (1955).

5.7 Thermoelectric Effect

A temperature gradient across a sample causes the Fermi energy to vary with position. This can give rise to a current even in the absence of an applied field. In an isolated sample the flow of charge results in an electric field that prevents further charge from flowing. Thermoelectric power, or thermopower for short, is defined as the ratio of the field developed to the temperature gradient. The thermopower Q is related to the Peltier coefficient Π, defined as the energy flux, measured relative to E_F, per unit current. The relation is $Q = \Pi / T$. It follows from this relation that, when one type of carrier is present,

$$Q = \mp \frac{k}{|e|} \left(\frac{|E_F - E_b|}{kT} + \frac{|\Delta E_T|}{kT} \right), \tag{39}$$

where the upper sign is for electrons, the lower for holes, $|E_F - E_b|$ is the distance of E_F from the conduction- or valence-band edge, and $|\Delta E_T|$ is a weighted average of the energy transported per carrier (across unit area in unit time) relative to the band edge. Specifically $\Delta E_T = \langle v^2 (E - E_b) \tau \rangle / \langle v^2 \tau \rangle$. The dependence of Q on the sign of the carriers means that, like the Hall effect, it can be used to determine whether the dominant carriers are electrons or holes. For a moderately doped semiconductor sample, Q varies a great deal with T because, as we have seen, E_F does. Q can be expressed in terms of the carrier concentration by using relation (13) or (14) to express E_F in terms of n or p, respectively. Other things being equal, Q is smaller for more heavily doped samples. In samples where electrons and holes are present in comparable numbers we have

$$Q = (Q_n \kappa_n + Q_p \kappa_p) / (\kappa_n + \kappa_p) . \tag{40}$$

5.8 Galvanomagnetic and Thermomagnetic Effects

These effects occur when a conductor carrying an electric current or thermal current, which may arise as a result of a thermal gradient, is placed in a magnetic field. In the isothermal situation discussed earlier in this section, introduction of a transverse magnetic field was found to give rise to the Hall effect and magnetoresistance. When thermal gradients are permitted, many additional galvanomagnetic effects occur. One of these is the Ettingshausen effect, the appearance in a transverse magnetic field of a temperature gradient perpendicular to both **H** and the current direction. Another is the Nernst effect in which, due to the transverse magnetic field, a temperature gradient appears parallel to the current direction. This by no means exhausts the possible galvanomagnetic effects. Among the thermomagnetic effects are magnetic field effects on thermal conductivity (analogous to magnetoresistance) and thermopower. The Righi–Leduc effect, a thermal analog of the Hall effect, is the appearance in a transverse magnetic field of a thermal gradient perpendicular to both **H** and the thermal current direction. In the Ettingshausen–Nernst effect the combination of thermal current and transverse magnetic field produces an electric field perpendicular to both. Many additional thermomagnetic effects could be specified. For a systematic listing of these effects see, for example, Conwell (1982).

6. RELAXATION TIMES AND MOBILITIES

Mobility in semiconductors is found to vary greatly with temperature and doping. In Fig. 7 are shown Hall mobility data versus temperature for the set of Ge samples whose carrier concentration variation with T was shown in Fig. 6. The least doped, thus purest, sample has the highest μ_H. Comparing it with the dashed line indicating a $T^{-3/2}$ dependence, we see that above ~100 K, μ_H for this

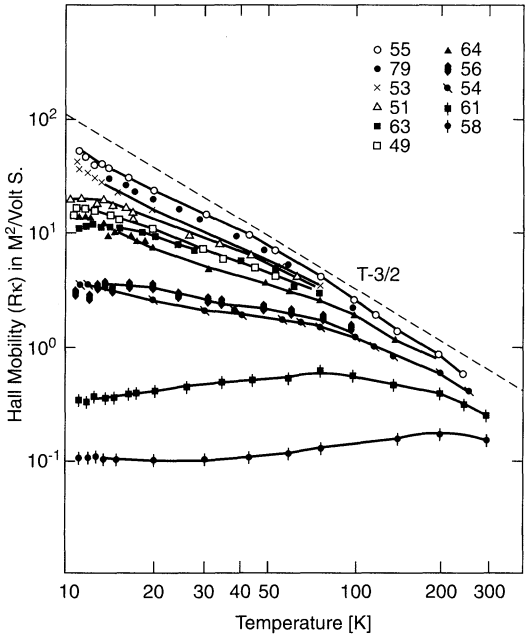

FIG. 7. Hall mobility vs temperature for the set of As-doped Ge samples for which n vs T is shown in Fig. 6. (After Debye and Conwell, 1954.)

sample decreases more rapidly with increasing T, while below ~100 K, the decrease is less rapid than $T^{-3/2}$. It will be shown shortly that the $T^{-3/2}$ dependence arises from scattering due to one kind of thermal vibration, the so-called acoustic modes. The steeper slope above 100 K indicates that another type of scattering, which can be attributed to another type of thermal vibration, is also operative. The progressively smaller mobility, and the less steep decrease with increasing temperature, seen with increasing doping will be shown to result from scattering by the doping ions.

In this section we discuss the momentum relaxation time τ defined in Eq. (22) for the most important scattering mechanisms. Also mobilities are given for these scattering mechanisms. At the end of the section we will present a more detailed comparison with experimental data.

6.1 Scattering Due to Lattice Vibrations

6.1.1 Acoustic and Optical Modes. The random thermal motion of the lattice atoms can be described as the superposition of a set of normal modes, each mode describing the displacement of every atom in the sample in terms of a propagating wave of a single frequency (see PHONONS IN CRYSTAL LATTICES). Because of the periodicity of the lattice, there is associated with each wave, as with each electron wave function, a wave vector in the direction of propagation. This wave vector will be denoted by **q** to distinguish it from the electron wave vector **k**. By the same reasoning as employed earlier for the **k**'s, the **q**'s have a closely spaced set of discrete values and occupy the first Brillouin zone. Frequency ν of the normal modes in Si versus wave vector is shown in Fig. 8. It is seen that in most directions there are four branches. These are labeled L for longitudinal and T for transverse, indicating that the vibrations are parallel to **q** and perpendicular to **q**, respectively. Actually the vibrations are not purely longitudinal or transverse in an actual crystal, but the difference is usually neglected. It is also usual, at least in conductivity theory, to neglect the differences between ν vs q for a given branch in different directions.

The branches that start at $q=0$ (the Γ point) are called acoustic branches. The low-frequency modes of the LA branch are familiar as sound waves. The other branch is called transverse acoustic (TA). In the low-frequency acoustic vibrations the two atoms in the unit cell move almost together, i.e., with only a small phase difference between them. Another type of long-wavelength vibration is possible: the two atoms in the unit cell move opposite to each other, with the same motion repeated in the next cell slightly earlier or later depending on the propagation direction of the wave, i.e., in a slightly different phase. These vibrations can also be both longitudinal and transverse. They are called optical modes (TO or LO) because, in a lattice with ions of opposite sign, they produce a polarization that can interact with light waves.

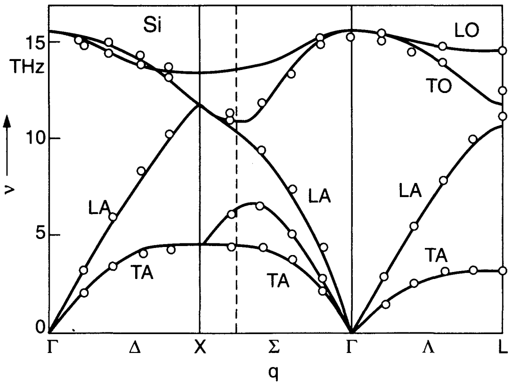

FIG. 8. Frequency ν vs wave vector **q** of normal modes of lattice vibration in Si for various directions in the Brillouin zone. (Adapted from Madelung, 1991.) See Fig. 2 for identification of the directions.

For the long-wavelength acoustic waves ν and the wavelength λ are related according to $\nu\lambda=\text{const}$. This constant is identified as the phase velocity, which for the LA branch is the sound velocity. In contrast to the acoustic branches, for the two optical branches ν is quite large for long wavelengths, i.e., close to $q=0$. The reason is that the opposing motion of the two atoms in the unit cell gives rise to a large restoring force and therefore a high frequency of vibration even for very large wavelengths. This has the further consequence that ν is not strongly dependent on λ, i.e., the optical branches are relatively flat, because λ only determines the relative phase of adjacent unit cells.

The vibrational energies of each normal mode can only have certain discrete values differing by amounts $h\nu$. The energy steps or quanta are called phonons, as suggested by the analogy with light quanta or photons. For a crystal in thermal equilibrium the average number of phonons n_q in a normal mode having frequency ν_q is given by Planck's law:

$$n_q = 1/(e^{h\nu_q/kT} - 1) \ . \qquad (40)$$

Thus around room temperature and below, the average number of optical phonons will be small because $h\nu$ for these phonons is larger than kT at room temperature.

6.1.2 Mobility Due to Acoustic-Mode Scattering. In the process of being scattered by the thermal vibrations, a carrier must emit or absorb a phonon. If the carrier wave vector changes from **k** to **k**′ in the process, while the phonon wave vector is **q**, conservation of momentum requires $\mathbf{k}' = \mathbf{k} \pm \mathbf{q}$. Energy must also be conserved, requiring $E_{\mathbf{k}'} = E_{\mathbf{k}} \pm h\nu_{\mathbf{q}}$, $h\nu_{\mathbf{q}}$ being the energy of the phonon with wave vector **q**. The upper sign is required for absorption and the lower for emission. The probability of absorption or emission of a phonon, thus of scattering, is proportional to the number of phonons, given by Eq. (40). Let us consider first the case of long-λ, or low-q, acoustic phonons. For these phonons, at all but extremely low temperatures, $h\nu_q$ is small compared to kT; Eq. (40) then gives $n_q \propto kT/h\nu_{\mathbf{q}}$. The small energy of the phonons, coupled with the conservation of momentum and energy conditions, results in these acoustic phonons causing scattering only within a valley.

The probability of scattering by any given vibrational mode also depends on how strongly the carrier is coupled to that mode, which in turn depends on the strain caused by that mode. For acoustic waves the coupling is described by a deformation potential E_1, which is the shift in the band-edge energy per unit strain (for a cubic crystal the strain is the fractional change in volume of a small region of the crystal) due to the wave. Combining the various factors described above, we find for the relaxation time τ_{ac} due to long-λ acoustic-mode scattering of a carrier with the simple model of band structure

$$\tau_{\text{ac}} = \frac{\pi\hbar^4 \rho u_l^2}{2^{1/2} E_1^2 (m^*)^{3/2} kT} E^{-1/2} \ , \qquad (41)$$

where ρ is the density of the material and u_l the sound velocity. (Note that TA waves do not scatter in this case.) It is seen that τ_{ac} depends inversely on $E^{1/2}$, or the velocity of the carrier, resulting in slow carriers having longer relaxation times or being scattered less.

Inserting Eq. (41) in Eq. (26), allowing for the averaging over carrier energy, thus over the Maxwell–Boltzmann distribution in this case, we obtain the mobility for long-λ acoustic-mode scattering:

$$\mu_{\text{ac}} = \frac{2(2\pi)^{1/2} e\hbar^4 \rho u_l^2}{3E_1^2 (m^*)^{5/2} (kT)^{3/2}} \ . \qquad (42)$$

Thus μ_{ac} decreases with increasing T as $T^{-3/2}$, $T^{-1/2}$ coming from the fact that $\langle E^{-1/2} \rangle \propto (kT)^{-1/2}$ for the Maxwell–Boltzmann distribution. The calculation of τ_{ac} and μ_{ac} for the many-valley model is more complicated because τ may not be constant over a constant-energy surface, and shear strains may also be important (see Herring and Vogt, 1956). However, the dependence on carrier energy and T remain the same.

In III-V and II-VI semiconductors, because they are somewhat polar and lack inversion symmetry, the strain of an acoustic wave produces also an electric polarization (see DIELECTRIC PROPERTIES OF INSULATORS). This causes scattering in addition to the scattering studied above, which may be called deformation-potential scattering. The additional scattering is called piezoelectric or piezoacoustic scattering. In contrast to τ_{ac} of Eq. (41), the relaxation time τ_{pe} for piezoelectric scattering is directly proportional to $E^{1/2}$. As a result, except for strongly piezoelectric material such as CdS or ZnO, it is expected to be important only for low-energy carriers, and thus for low temperatures only. It will be seen that, except for very pure samples, impurity scattering is the dominant scattering mechanism at low temperatures. Thus, in materials that are not strongly piezoelectric, piezoelectric scattering need only be considered at low temperatures in pure samples. For further discussion of piezoelectric scattering see Zawadzki (1982) and references therein.

6.1.3 Mobility Due to Nonpolar Optical-Mode Scattering. Optical-mode scattering in nonpolar crystals may be treated quite similarly to acoustic-mode scattering. For the coupling to the optical modes it is necessary to introduce a different deformation potential because the strain, in contrast to that due to an acoustic phonon, is not simply a change in linear dimensions or volume. In the case of a zinc-blende or diamond structure the strain is a displacement of the two sublattices, based on the two atoms in a unit cell, with respect to each other. The appropriate deformation potential is D, the shift of the band edge per unit relative displacement of the two sublattices. Another difference between optical- and acoustic-mode scattering is that, because of the larger size of the optical phonons, the exponential in Eq. (40) may not be expanded; n_q must be taken from Eq. (40). The relaxation time for optical-phonon scattering (all phonons assumed to have the same frequency ν_0) is given by

$$\tau_{\mathrm{op}}=\frac{2^{3/2}\pi^2\hbar^3\rho\nu_0}{D^2(m^*)^{3/2}}\,[n_q(E+h\nu_0)^{1/2} + (n_q+1)(E-h\nu_0)^{1/2}]^{-1}\,. \tag{43}$$

The first term in the square brackets represents relaxation due to phonon absorption and the second relaxation due to phonon emission. The emission probability is proportional to n_q+1, n_q representing stimulated emission and 1 spontaneous emission. Emission is only possible, however, if $E>h\nu_0$. If $h\nu_0$ were negligible compared to E, as is the case for long-λ acoustic phonons, n_q and n_q+1 could be replaced by $kT/h\nu_q$ and τ_{op} would have the same dependence on E and T as τ_{ac}. Generalizing τ_{op} to the many-valley model is simple; it requires only replacing m^* by the geometric mean mass $(m_l m_t^2)^{1/3}$.

The mobility μ_{op} obtained after averaging τ_{op} over energy is usually larger than μ_{ac} because of the relatively small number of optical phonons. This same feature, however, results in μ_{op} decreasing more rapidly with increasing T as n_q increases, according to Eq. (40).

6.1.4 Mobility Due to Intervalley Scattering. Because of conservation of momentum, the **q** of a phonon that causes an intervalley transition will be approximately equal to the difference between the **k** vectors of the valley minima. This difference is usually of the order of the radius of the Brillouin zone, which means that the energy of the phonon required is comparable to the optical-phonon energy. The intervalley scattering may then be treated much like the optical-phonon scattering, leading to an expression for the intervalley scattering relaxation time τ_{iv} of the same form as τ_{op}. Of course, the deformation potential is different for this case, and high-energy acoustic phonons as well as optical phonons may be involved in the transition. Note that the energy in Eq. (41) or (43) must be measured from the bottom of the valley into which the carrier is scattered, which need not be the same energy as the bottom of the valley in which it originates. Also m^* of Eq. (42) must be replaced by the geometric mean mass.

6.1.5 Mobility Due to Polar Optical-Mode Scattering. Nonpolar optical scattering, discussed in Sec. 6.1.3, may exist in polar as well as nonpolar crystals. In polar crystals, however, the polarization of the crystal arising from the vibrations of the LO branch results in an additional interaction with the carriers, or scattering, the polar optical scattering. In the III-V semiconductors of most interest, GaAs and InSb, the polar optical scattering dominates over nonpolar scattering for electrons in the conduction band because the latter vanishes for spherically symmetric wave functions, which is more or less the situation for the conduction-band minimum at the Γ point. Polar optical-mode scattering is much more difficult to treat theoretically than nonpolar. A relaxation time does not exist unless $E\gg h\nu_0$, which requires temperatures well above room temperature. It is possible to get a qualitative idea of what is important by considering the mean free time between collisions, τ_{po}. For the simple model

$$\tau_{po}=(2m^*E)^{1/2}/2eE_0\{n_q\sinh^{-1}(E/h\nu_0)^{1/2} + (n_q+1)\sinh^{-1}[(E-h\nu_0)/h\nu_0]^{1/2}\}, \tag{44}$$

where $E_0=m^*e^2h\nu_0(1/\epsilon_\infty-1/\epsilon_0)/\hbar^2$, ϵ_∞ and ϵ_0 being the high-frequency and static permittivities, respectively. Note that Eq. (44) applies only for low carrier concentration because it neglects the screening of the electric fields due to the free carriers, which decreases the scattering effect.

Equation (44) may be used to predict the temperature dependence for polar optical scattering. For $\langle E\rangle\ll h\nu_0$ the energy dependence cancels out and the temperature dependence of μ_{po} is that due to $1/n_q$. In the other limit, $\langle E\rangle\gg h\nu_0$, the E dependence arising from the

$\sinh^{-1}$ terms becomes logarithmic and may be neglected. The combination of $E^{1/2}$ and n_q^{-1} dependence in this limit yields $\mu_{\mathrm{po}} \propto T^{-1/2}$. For a detailed treatment of polar optical scattering see Zawadzki (1982) and references therein.

6.2 Impurity Scattering

The donors and acceptors that have given up an electron to the conduction band, or a hole to the valence band, are left charged. The charge results in an attraction or repulsion of the carriers, depending on the sign of the charge. In either case the carriers are deflected, thus scattered. There have been many calculations of the relaxation time for scattering by impurity ions, τ_I, varying in degree of sophistication and in which refinement of the process they concentrated on. A treatment that gives reasonably good agreement with experiment for not too heavily doped (nondegenerate) samples leads to

$$\tau_I = \frac{2^{1/2}}{\pi} \frac{(m^*)^{1/2}\epsilon^2}{e^4 N_I} \frac{E^{3/2}}{\ln(1+\xi) - \xi/(1+\xi)}, \tag{45}$$

where, for a sample with carrier concentration n, $\xi = (8\pi\epsilon m^* kT/ne^2h^2)E$, N_I being the number of ionized impurities per unit volume. The quantity ξ arises from the screening of the ions by the free carriers. For degenerate samples ξ is independent of T, although it still depends on E. Unless the sample has almost as many acceptors as donors (the reverse for a p-type sample), n may be taken equal to N_I. For moderately doped samples around room temperature $\xi \gg 1$, and Eq. (45) predicts that τ_I decreases as N_I increases, and that slow carriers are deflected or scattered more than fast carriers, both of these effects being expected. As the temperature decreases the terms involving ξ have more effect and τ_I increases less rapidly with E. The expression given above for ξ no longer represents the screening properly when many of the carriers have dropped out of the conduction band (or valence band), and n must be replaced by a quantity involving both the donor and acceptor concentrations. See, for example, Chattopadhyay and Queisser (1981), where also many of the improvements on the theory presented above are discussed.

Carrying out the averaging over carrier energy, we obtain from Eq. (45) the impurity mobility

$$\mu_I = \frac{2^{7/2}}{\pi^{3/2}} \frac{\epsilon^2 (kT)^{3/2}}{(m^*)^{1/2} e^3 N_I} \frac{1}{\ln(1+\bar{\xi}) - \bar{\xi}/(1+\bar{\xi})}, \tag{46}$$

where $\bar{\xi} = 24\pi\epsilon m^* (kT)^2/ne^2h^2$. In this form μ_I is valid for nondegenerate samples at temperatures where most of the carriers are still free. It increases with T less rapidly than $T^{3/2}$ because of the terms in $\bar{\xi}$. For degenerate samples $E \simeq E_{\mathrm{F}}$, independent of temperature, and with ξ independent of T, as pointed out earlier, μ_I is independent of T. This explains the fact that the Hall mobility of sample 58 in Fig. 8, which is both degenerate and so heavily doped that impurity scattering dominates, is essentially independent of T.

6.3 Neutral-Impurity Scattering

As the temperature falls carriers return to the donors or acceptors and the number of neutral impurities increases. When neutral the impurities still scatter carriers. Because of the analogy, discussed in Sec. 3, between a neutral impurity and a hydrogen atom, the scattering has been treated similarly to scattering of a free electron by hydrogen (see ELECTRON SCATTERING BY ATOMS AND MOLECULES). The result is that at low temperatures ($T < 20$ K for Si and < 4 K for Ge) the relaxation time τ_N for scattering by a concentration N_N of neutral impurities is given by

$$\tau_N = (m^*/2)^{1/2} \pi (a_{\mathrm{H}}^*)^2 N_N \times [(E/E_b^* + 0.0275)^{-1} + 10]^{-1} E^{-1/2}, \tag{47}$$

where a_{H}^* and E_b^* are defined in Eqs. (6) and (7), respectively (Zawadzki, 1982). Modifications for the many-valley model do not change τ_N greatly. Experimental verification of Eq. (47) is difficult because there are few free electrons at such low temperatures, and the effects of other scattering mechanisms must be subtracted. Nevertheless there are low-temperature data for Ge that indicate that μ_N

varies as $T^{-1/2}$, as predicted by Eq. (47). At higher temperatures μ_N is predicted to go to a constant value (Zawadzki, 1982).

6.4 Combination of Mobilities

In general, at any given temperature, more than one scattering mechanism is operative. To find the mobility one must first calculate an overall relaxation time τ by combining the τ's for the individual scattering mechanisms (which are essentially probabilities of scattering per unit time) according to $1/\tau = \Sigma_i(1/\tau_i)$. The resulting τ must then be averaged over energy to obtain μ. It is not unusual for this procedure to be simplified by calculating μ's for each scattering process and combining the reciprocals of these μ's to obtain the reciprocal of the overall μ. For the combination of acoustic-mode and ionized-impurity scattering the error due to this simplified procedure may be as much as 30%.

6.5 Comparison with Experimental Data

For well-characterized semiconductors, such as Si, Ge, and GaAs, the foregoing theory, with some refinements, agrees well with experiment for moderately doped and intrinsic samples. For *n*-Si around room temperature in very pure samples $\mu \propto T^{-2.42}$, resulting from an admixture of μ_{ac} ($\propto T^{-1.5}$) and μ_{iv} due to scattering between the [100] valleys. The latter has a much steeper T dependence arising from the large energy of the phonons required. Below 50 K the energy required for the intervalley process is too high to allow this process to occur, and in pure samples $\mu = \mu_{\text{ac}}$. For pure samples of *p*-Si and *p*-Ge around room temperature $\mu \propto T^{-2.33}$ from a mixture of μ_{ac} and nonpolar optical scattering. Plots of experimental and theoretical μ vs T for the cases just cited and for other intrinsic semiconductors may be found in Madelung (1991). Comparison of theory with experiment for a moderately doped *n*-type GaAs sample is shown in Fig. 9. It is seen that polar optical scattering dominates at high temperatures while impurity scattering dominates at low temperatures. In purer samples piezoelectric scattering may dominate at low temperatures.

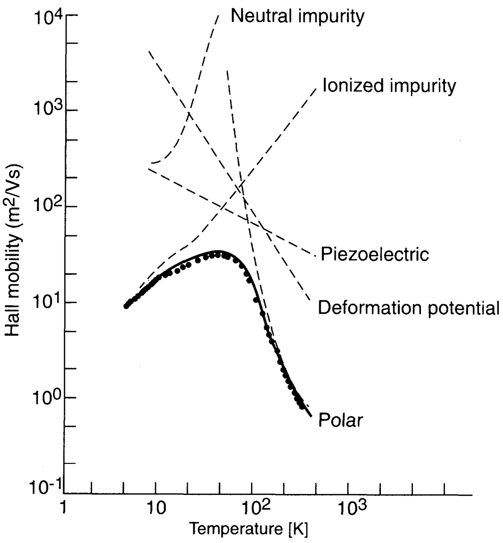

FIG. 9. Dots: The measured mobility of electrons in a sample of *n*-GaAs vs temperature. Dashed lines: Theoretical mobilities calculated for the various scattering process indicated. Solid line: The combined mobility calculated from all the scattering processes. (After Fletcher and Butcher, 1972.)

7. HIGH-FIELD TRANSPORT

7.1 Hot Carriers

In this section we consider only nondegenerate samples, i.e., samples in which the carrier energies in the absence of an electric field are given by the Maxwell–Boltzmann distribution, Eq. (12). The effects to be described cannot be achieved with heavily doped samples. In the absence of an electric field the carriers interact with phonons, emitting and absorbing them, while the Maxwell–Boltzmann distribution of carrier energies and the Planck distribution of phonon numbers, Eq. (40), are maintained. Application of an electric field, which accelerates the carriers, increases their energy. This results in an increase in their rate of phonon emission, giving rise to the well-known Joule heating that accompanies a current. Achievement of a steady state, the situation for dc current, requires that the rate at which the electric field puts energy into the carriers equals the rate at which they lose energy through phonon emission. The rate of energy gain is $\mathbf{F}\cdot\mathbf{v}_d$, which, given Eqs. (19) and (2), equals $e\mu|\mathbf{E}|^2$. With $\langle dE/dt\rangle$ representing the average rate of energy loss, the steady-state condition may be written

$$e\mu|\mathbf{E}|^2=\left\langle\frac{dE}{dt}\right\rangle . \qquad (48)$$

In small electric fields, satisfying the condition (48) requires little change in the distribution, the essential change being the addition of a small drift term to the Maxwell–Boltzmann distribution. A high electric field may be defined as one for which this is no longer true. In semiconductors a steady state is always possible, but in high fields the drift term may not be small and the part of the distribution that is independent of the field direction no longer corresponds to an average energy of $\frac{3}{2}kT$, where T is the lattice temperature. It may still be the case that the latter part of the high-field distribution is reasonably described by a Maxwell–Boltzmann distribution, but the carrier temperature involved, denoted by T_e, is greater than the lattice temperature. For this reason the carriers are called "hot." Even when the high-field distribution is not Maxwellian it is usual to speak of a carrier temperature, which may be defined by equating the average carrier energy to $\frac{3}{2}kT_e$. It should be noted that to perform clear-cut experiments to measure hot-carrier effects it is necessary to apply the high electric field in pulses of low repetition rate to avoid raising the lattice temperature. When that is not possible it is necessary to consider the effects of a nonequilibrium phonon distribution along with the hot-electron effects.

From the preceding discussion it is clear that the higher the mobility of a material, the lower will be the magnitude of the electric field required to create hot carriers. High mobility means that the field can give the carriers energy at a high rate. It also suggests that the coupling to phonons is not strong, with the consequence that the rate of energy loss to phonons is not large.

Hot electrons may also be created by light. When a photon is absorbed by an electron, the electron energy is increased by the amount of the photon energy. Another method of creating hot electrons is through a potential step. A junction of two semiconductors with different band gaps (heterojunction) constitutes such a step. When an electric field is applied in the appropriate direction, electrons moving with thermal energy in the wide-band-gap region cross into the narrow-band-gap region. There they find themselves well above the conduction-band edge. In other words, the energy difference between the conduction-band edges in the two regions is converted into kinetic energy. In such a situation, before the carriers have undergone the collisions that set up a steady state their velocities exceed the steady-state drift velocity. This phenomenon is called velocity overshoot. It occurs also when, because of a sudden change in electric field intensity, the distribution of carriers has not yet reached its steady state. Velocity overshoot has been detected by electrical methods (Hess, 1990) and optical methods (Auston, 1990). It is of practical interest because it could be used to make faster devices. A potential step such as the heterojunction described above can also be used as a barrier, making possible the measurement of electron energies. This procedure could be called hot-electron spectroscopy (Hess, 1990).

In the remainder of this section we discuss some of the more striking hot-electron effects. Computer simulations of some of these effects are displayed in Hess (1990).

7.2 Mobility Changes in High Fields

Because the relaxation times for most scattering processes depend on carrier energy, as shown in Sec. 6, it is expected that the mobility will change in high electric fields. At fields where the carrier "temperature" is not much increased over the lattice temperature, called the warm-electron range, it is expected that μ in a field $\mathbf{E}$ depends on μ_0, the low-field (Ohmic) mobility, according to

$$\mu=\mu_0(1+\beta|\mathbf{E}|^2) , \qquad (49)$$

where β is independent of field. The lowest-order deviation of μ from μ_0 must be quadratic in field because μ should not change on reversal of the field duration. The behavior of μ predicted by relation (49) has been verified by many experiments, in which lattice temperature, impurity content, and crystal orientation have been varied (for data and references for the material in this section see Conwell, 1967).

We can predict how μ changes from the energy dependence of τ for the scattering mechanisms that operate. Thus, if acoustic or nonpolar optical or intervalley scattering is dominant, according to Eqs. (41) or (43) μ must decrease on electron heating. This has been verified for Ge and Si, with reasonably good agreement with theory. For *n*-Ge at 300 K the decrease in μ starts at $\sim 10^5$ V/m, while at 77 K, where the low-field mobility in highly purified samples is higher by an order of magnitude, it starts at $\sim 10^4$ V/m. For *n*-Si, with lower mobility, higher fields are required to reach the non-Ohmic region.

A by-product of the mobility changes in the many-valley model is that current is no longer parallel to **E**, as it must be at low fields because we are dealing with cubic crystals. The nonparallelism occurs for **E** in directions that are not axes of rotational symmetry. Although the current direction for a given valley is at an angle to the valley axes because of the mass variation in the valley, in low fields the components perpendicular to **E** cancel out. In high fields, valleys with **k** vectors of small-mass (m_t) electrons more closely aligned with **E** get hotter, and their contribution decreases relative to the other valleys. This has the result that the perpendicular contributions to **j** no longer cancel out. The effect is magnified by the fact that the hotter valleys lose carriers by intervalley scattering to the cooler valleys.

As the field increases in the non-Ohmic region, v_d continues to increase, but more and more gradually. For Ge, Si, and many other semiconductors at 300 K, v_d tends to saturate at high fields at a value of $\sim 10^5$ m/s. This value is important for devices, which are frequently operated at high fields, because it determines the transit time through the device. Interestingly, this drift velocity is high enough so that, in the very small devices that are now being investigated, a sizable fraction of the carriers can make the transit through the device without a collision. Such collisionless transport is called ballistic transport. It is of particular interest because it provides a way of making faster devices.

In many of the group-IV and III-V semiconductors, when the field is raised beyond the range where v_d tends to saturate, to close to 10^7 V/m for *n*-Ge or 1.2×10^4 V/m for InSb to cite two examples, the current begins to rise very steeply. It has been shown that this rise represents an increase in carrier concentration due to impact ionization of the lattice atoms by the hot carriers. This process will be discussed in Sec. 7.3. A different phenomenon, called the Gunn effect, occurs at high fields in GaAs, GaP, InAs, and some other materials. This will be taken up in Sec. 7.4.

The discussion so far in this section has focused on phenomena from room temperature down to ~78 K in relatively pure samples. We consider now the situation at low temperatures, typically 20 K and below, where impurity scattering is important and a sizable fraction of the carriers have been frozen out of the conduction and valence bands. According to Eq. (45) τ_I and therefore μ_I should increase when the carriers are heated. Thus in this range β of Eq. (49) should be positive. This was in fact found for Ge samples at 20 K. However, there is another reason for conductivity to increase with field in this temperature range: hot carriers can cause impact ionization of neutral donors or acceptors. Hall measurements made as a function of field were able to show that both effects occur at this temperature. The initial non-Ohmic behavior is due to increase in mobility, but once the carriers are hot, impact ionization dominates, with the rise in current with field becoming so steep at temperatures where most of the carriers are frozen out as to be called a breakdown. The field required for this type of breakdown is less than 10^3 V/m for *n*-Ge.

7.3 Impact Ionization

In the temperature range where the donors or acceptors are fully ionized or almost so, an electron or hole that has gained sufficient energy from the field may give up its energy to a valence electron, raising the latter into the conduction band and leaving a hole in the valence band. The threshold energy required for an electron to create a hole–electron pair in this way is readily found for a semiconductor with the simple model of the band structure by applying the conditions of conservation of energy and momentum. For electron and hole masses m_n^* and $m_p^*=\gamma m_n^*$, respectively, the threshold energy for pair creation by an electron is $E_g(2+\gamma)/(1+\gamma)$. Thus for equal electron and hole masses an electron must have a minimum energy $3E_g/2$ to create an electron–hole pair. An initiating hole would also require only $3E_g/2$ for pair creation if $m_p^*=m_n^*$. With the gap in silicon, for example, 1.1 eV, at 300 K the field must raise the energy of an average electron, 0.025 eV, by a factor close to 60 to achieve pair creation. The field required is over 10^7 V/m. For InSb, with a much smaller gap and higher mobility carriers, only 1.2×10^4 V/m is required.

At fields just above threshold, ionization is due to a relatively few "lucky" carriers that have, fortuitously, been scattered less than others. This leads to an ionization rate changing roughly with field as $\exp(-A/|\mathbf{E}|)$, where A is a constant. At still higher fields, where most of the carriers are raised to a higher temperature T_e, the ionization rate is expected to depend on field as $\exp(-B|\mathbf{E}|^2)$.

7.4 The Gunn Effect

In n-type GaAs and a number of other materials, instead of the near saturation at high fields with subsequent impact ionization of the lattice found in Si, for example, a different and unique phenomenon called the Gunn effect is observed. Application of a field results in little departure from Ohm's law until after attainment of a threshold field in the range $(2\text{–}4)\times10^5$ V/m at 300 K in GaAs. Beyond this field the current is found to vary strongly with time, its average value dropping below the threshold value. Many different types of time variation have been seen in different samples, the simplest being an oscillation between the value at threshold and a value as low as half the threshold. The period of the oscillations was found to be about equal to the transit time of an electron between the electrodes as calculated from the threshold current, the number of electrons, and the length of the sample. The oscillations are typically of microwave frequency and the Gunn effect is used for microwave generation.

The current oscillations arise from the mobility change that occurs on heating of the carriers. As seen in Fig. 9, polar optical scattering is dominant in GaAs at high temperatures. According to Sec. 5, this means that electrons that acquire $E>h\nu_0$, ~400 K in GaAs, are scattered progressively less as their energy increases. In a strongly polar insulator with no other important scattering mechanisms, this would lead to the electron energy growing without limit as the field increases, causing impact ionization, severe lattice heating, changes in chemical and mechanical properties, and, most important, breakdown of the insulating properties. In fact, this mechanism has been suggested as the reason for breakdown of polar insulators. In GaAs, however, when an electron reaches 0.3 eV it can be scattered, by a phonon, from the Γ minimum to one of the L valleys in the conduction band (see Fig. 4). At 0.46 eV it can also be scattered into one of the X valleys in the conduction band. The probability of these processes is quite high because the densities of states in the X and L valleys are much higher than that in the Γ valley. This intervalley scattering has been detected by picosecond optical measurements (Auston, 1990). The mobility of electrons in the L and X valleys is also considerably smaller than that of electrons in Γ. Similar mobility changes due to intervalley electron transfers of hot electrons occur in GaAs-GaP alloys, InP, CdTe, ZnSe, InSb, and GaSb-InSb alloys, which also show the Gunn effect. In InSb the effect can be observed only for a nanosecond after the beginning of current flow because, the energy gap being smaller than the energy separation between valleys, impact ionization dominates after a short time. Interestingly, the Gunn effect is found under uniaxial stress in p-Ge. Although p-Ge has the degenerate band structure, under appropriate stress it goes over into one with two sets of valleys appropriately separated.

The result of the intervalley transfers is that beyond the threshold field further increase in $\mathbf{E}$ causes a decrease in current. This is an example of negative differential conductivity. How this phenomenon leads to a current that oscillates in time is discussed in the next section.

7.5 Negative Differential Conductivity

Negative differential conductivity (NDC) is characterized by regions of negative slope in plots of current density versus field. There are two types of NDC, shown schematically in Fig. 10, distinguished by their behavior in circuits.

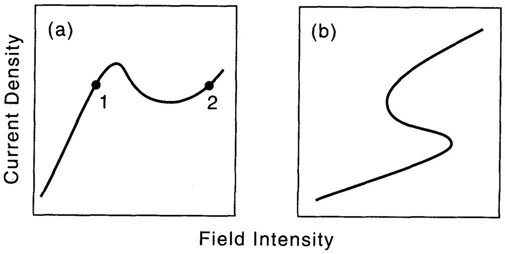

FIG. 10. The two types of negative differential conductivity. (a) The N-shaped curve representing voltage-controlled NDC, and (b) the S-shaped curve representing current-controlled NDC. (After Conwell, 1970.)

Unique electrical behavior results when a system operates in a region of NDC. In particular, a departure from equilibrium such as a fluctuation of field or of current will tend to grow rather than dissipate. In a sample with the voltage-controlled type of NDC, consider a small nonuniformity of the field at some point such that the field upstream (as regards current flow) is higher than the field downstream. This nonuniformity may be caused by a momentary deficiency of positive charge or accumulation of electrons, or by a permanent nonuniformity in impurity concentration or conductivity. If the differential conductivity of the sample is positive, a higher current will flow into the region than flows out, and any accumulated negative charge will be dissipated. If the differential conductivity is negative, however, a smaller current will flow into the region than out; more negative charge will accumulate and the original field discontinuity will be enhanced.

In a rectangular sample, as is ordinarily used, the equipotentials must be planes perpendicular to the sides of the sample, and the tendency will be to build up a plane of negative charge, an electron accumulation layer, at the discontinuity. If the initial nonuniformity or fluctuation had been of the opposite sign, there would be a tendency for a layer of positive charge, that is, an electron depletion layer, to grow. This is the type of instability that leads to the Gunn effect.

Imagine now a fluctuation in a sample with current-controlled NDC such that at some instant two adjacent lateral regions of filaments are carrying different currents. If the differential conductivity were positive, the filament with the higher current would have the higher field. A transverse current would then flow in such a direction as to equalize the currents. In a sample with NDC, however, the filament with the higher current would have the lower field, and transverse current would flow in the direction that increases the difference between the currents. Stability will be reached when the currents in the different filaments all correspond to the same value of electric-field intensity. Current-controlled NDC has been observed in Ge at low temperatures in situations where the carriers are hot enough to cause impact ionization of impurities, and in Ge and InSb at room temperature where the carriers are hot enough to ionize the lattice. In both these situations, once breakdown is achieved, the current can be carried at a lower voltage than was required for breakdown.

As has been seen, in the context of the Gunn effect, voltage-controlled NDC occurs in many materials because of mobility changes caused by the electric field. It also occurs in some semiconductors, notably CdS and high-resistivity GaAs, because the number of conduction electrons n decreases with increasing field. In CdS it is found that, beyond a threshold field $\sim 2 \times 10^4$ V/m, $n \propto E^{-m}$, where $m > 1$. This behavior has been attributed to "field quenching," in which the field serves to free holes from hole traps so that conduction electrons can combine with them. NDC in CdS is convenient for study because in the high-field regions E_g is decreased, making these regions opaque to light for which the low-field regions are transparent. It is found that with a blocking contact, which makes n at the cathode smaller than the value in the bulk, a high-voltage domain originates at the cathode at ~ 500 V and becomes larger with increasing voltage.

In GaAs exhibiting the Gunn effect a high-voltage region is also likely to form in the neighborhood of the cathode because of a small high-resistivity region there. The high-voltage region results from a dipole, with an accumulation of electrons in one region, a depletion in another. If the product of n and the sample length is large enough, the dipole will detach itself from the cathode and continue to grow until the upstream and downstream fields have a set of values such as 1 and 2 in Fig 10(a). The domain moves down the sample with the drift velocity because it is made of moving electrons. When it reaches the anode it exits from the sample, the current rises, and the process of building up another domain at the cathode begins. For further discussion and references see Conwell (1967, 1970).

7.6 Acoustoelectric Effect

In CdS, ZnO, and a number of other II-VI and III-V compounds for certain crystallographic directions, Ohm's law is obeyed only up to the field for which the drift velocity $v_d \simeq v_s$, the sound velocity. Note that this is not a high enough field for carrier heating. Further increase in **E** results either in saturation, i.e., a time-independent current that increases little with further increase in field, or in time-dependent behavior in which the current oscillates between the Ohmic value appropriate to the particular field and the lower, saturation value (Einspruch, 1965). In the latter case, probing the sample electrically led to the discovery that, similar to the Gunn effect, a high-resistance domain is generated at the cathode and moves down the sample. In this case, it moves with v_s. After it leaves at the anode a new domain is generated, etc. In the time-independent case, as in the Gunn effect, a high-resistance region is generated but the conditions for it to move are not satisfied.

A clue to the origin of these effects is that the directions for which they are observed are those for which there is piezoelectric coupling between an acoustic wave traveling in that direction and an electric field along the direction. An acoustic wave introduced from outside, through a transducer, for example, in such a direction produces an alternating electric field that bunches the conduction electrons or holes in its potential-energy minima. In the absence of an applied dc field the acoustic wave must drag the electron bunches along with it and experiences attenuation. If, however, a dc field is applied in this direction of sufficient magnitude to make the electrons drift with $v_d > v_s$, the electrons drag the wave, i.e., the wave experiences gain. Of course, there are acoustic waves, the phonons, present in these materials even when acoustic energy is not put in from the outside. Application of a dc field high enough to make $v_d > v_s$ in one of the piezoelectrically active directions results in amplifying acoustic waves that happen to be traveling in the field direction. The traveling domains have been found to be regions of high acoustic flux. The high resistance associated with these domains, or with stationary domains, is due to the large rate of momentum loss of the electrons in acoustic-wave generation. The acoustic-wave amplification we have been discussing may be usefully thought of as stimulated emission of phonons.

8. IMPURITY-BAND CONDUCTION

8.1 Semiconductor Conductivity at Low Temperatures

As discussed in Sec. 4, with decreasing temperature conductivity decreases rapidly in all but the most heavily doped (degenerate) semiconductor samples because electrons and holes fall back into the impurity levels. The decrease in n with T was shown in Fig. 6 for a set of n-Ge samples down to ~10 K. According to Eq. (17), the slope represents the activation energy, 0.01 eV for this case. A plot of $\ln\kappa$ vs $1/T$ for these samples would show the same decrease, with essentially the same activation energy, because the mobility has much weaker T dependence than n in this range. A set of p-Ge samples would give very similar data, with the same activation energy. As T is lowered further, carriers must continue to drop out of the conduction or valence band at the same rate, but the conductivity does not continue the rapid rate of decrease. This is shown in Fig. 11 where the logarithm of resistivity ρ is plotted versus $1/T$ for a set of p-Ge samples with acceptor concentrations ranging from 2×10^{22} to 7.3×10^{22} m^{-3}. From the minimum in ρ around 78 K, corresponding to all holes being in the valence band, ρ

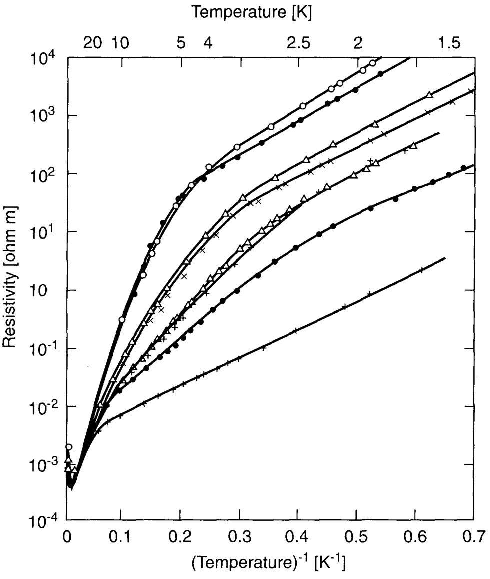

FIG. 11. Resistivity vs temperature for a set of p-Ge samples with acceptor concentrations ranging from 7.3×10^{22} m^{-3} for the lowest curve to 2×10^{22} m^{-3} for the uppermost curve. (After Fritzsche, 1955.)

rises exponentially for all samples, with a slope corresponding to the activation energy, 0.01 eV; this is a repetition of the data in Fig. 6 with a more compressed scale. However, beyond some temperature, ranging from 20 K for the lowest ρ (most doped) to 5 K for the highest ρ (least doped) sample, the rate of increase of resistivity with decreasing T is markedly smaller. Processes with lower activation energy are now dominating the conductivity.

The experimental curves of Fig. 11, from ~78 K on down, can be fitted by a sum of terms of the form

$$\rho=\rho_1 e^{\mathscr{E}_1/kT}+\rho_2 e^{\mathscr{E}_2/kT}+\rho_3 e^{\mathscr{E}_3/kT}, \tag{50}$$

where $\rho_1 \gg \rho_2 \gg \rho_3$ and $\mathscr{E}_1 > \mathscr{E}_2 > \mathscr{E}_3$. From the discussion above $\mathscr{E}_1=0.01$ eV. The second term is not seen for the lowest- and highest-resistivity samples but is seen in between. The third contribution, $\rho_3 \exp(\mathscr{E}_3/kT)$, is seen over the lowest temperature range for all samples.

8.2 Conduction Mechanisms at Very Low Temperatures

The $\mathscr{E}_2$ and $\mathscr{E}_3$ contributions can be understood in terms of the expected density of states. We will describe the situation in terms of electrons, but the description applies to holes with the obvious modifications. Figure 5 gives the density of states for a lightly doped semiconductor, Ge with $\sim 10^{20}$ donors/m^3, for example. At $\sim 10^{22}$ m^{-3}, as discussed in Sec. 3, the hydrogenic orbits of electrons bound on adjacent donors begin to overlap and the discrete level at E_d begins to spread into a band. The random arrangement of the impurities results in the overlap, and therefore the bandwidth, being somewhat greater than it would be for a regular arrangement. Nevertheless, the band is quite narrow at this concentration and the repulsion U_r for a second electron on a donor site is larger than the bandwidth B. In that situation the impurity band splits into two subbands, the upper band separated from the lower by an energy $\sim U_r - B$. In Fig. 12 the lower band is identified as the (a) band, the upper the (b) band. To explain the exponential temperature dependence of ρ in the (a) band, and other measured properties, it is necessary to assume that the electrons in this band are localized, one to a donor, for $N_d < 7\times10^{22}$ m^{-3}, the highest concentration in Fig. 11.

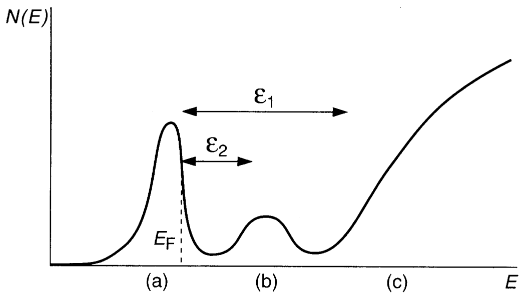

FIG. 12. Distribution in energy of the states in an *n*-type semiconductor showing the splitting of the impurity band into two bands (a) and (b). (c) is the conduction band. The position of the Fermi level is shown for the case $N_a < N_d/2$. (After Mott and Davis, 1971.)

In the limit of very low temperatures the (a) band would be completely filled with one electron per site if there were no acceptors, while the higher bands would be empty. Actually there are always some acceptors present in an *n*-type sample, with the result that in the low-T limit $E_{\rm F}$ lies in the (a) band as shown in Fig. 12, and the band has a number of vacancies equal to N_a. It is this fact that makes conduction possible within the (a) band. Conduction takes place by an electron hopping from a neutral donor, i.e., a filled donor site, to a positively charged donor, or empty donor site. It is clear that the amount of conductivity due to this mechanism depends on N_a/N_d. For $N_a \ll N_d$, κ will increase as N_a increases. However, if N_a is close to N_d, κ would decrease with further increase in N_a because there would be fewer electrons to hop.

The randomness of donor and acceptor distributions results in electrons on different donors having somewhat different energies. To enable an electron on site i to hop to a donor on site j, where the energy level is higher by W_{ij}, the energy W_{ij} must be provided by a phonon. The hopping probability therefore depends on the abundance of phonons of energy W_{ij}, which at low temperatures ($kT \ll W_{ij}$) is given by $\exp(-W_{ij}/kT)$. It also depends strongly on the overlap of the electron wave functions in the initial and final states. This dependence, for two donors separated by a distance R_{ij}, is given by $\exp(-2R_{ij}/a_{\rm H})$, where $a_{\rm H}$ is the radius of the ground-state (hydrogenic) orbit. Thus the probability ν_{ij} of a transition per unit time between sites i and j may be written

$$\nu_{ij}=\nu_0 \exp(-2R_{ij}/a_{\mathrm{H}}-W_{ij}/kT), \qquad (51)$$

where ν_0 depends on the coupling to the phonons, given by the deformation potential (see Sec. 6.1) in this case, R_{ij}, and other factors. In calculating κ or ρ from Eq. (51) it is useful to distinguish between two ranges:

1. the "high-temperature" range where kT is large enough so that an electron can usually find a phonon of sufficient energy to jump to a nearest neighbor;
2. the "low-temperature" range where kT is small enough so that the electron will preferentially jump to a further site for which W_{ij} is smaller despite the decrease in $\exp(-2R_{ij}/a_{\mathrm{H}})$.

The latter hopping process is called variable-range hopping (VRH) because the average hopping distance increases with decreasing T.

8.2.1 Variable-Range Hopping. It is readily shown that the optimum range $(R_{ij})_m$ is $[9a_{\mathrm{H}}/8\pi N(E_{\mathrm{F}})kT]^{1/4}$, where $N(E_{\mathrm{F}})$ is the density of states at $E=E_{\mathrm{F}}$. When $(R_{ij})_m$ and the W_{ij} that corresponds are inserted in Eq. (51) we find the hopping rate

$$\nu_{ij}=\nu_0 \exp[(-T_0/T)^{1/4}], \qquad (52)$$

where $T_0=16/kN(E_{\mathrm{F}})a_{\mathrm{H}}^3$. The diffusion coefficient D that results from the hopping is given by $\langle \nu_{ij} R_{ij}^2 \rangle$. From this the mobility may be obtained with the use of the Einstein relation $\mu=De/kT$. To calculate κ or ρ, n is taken as the number of electrons per unit volume within kT of E_{F} because essentially only those will be able to hop to an empty site by finding a phonon to assist them, or by emitting a phonon. This number is $2N(E_{\mathrm{F}})kT$, the factor 2 resulting from the fact that electrons of both spins may occupy each state.

In carrying out the averaging over R_{ij} and W_{ij} required to obtain D it is necessary to take into account not only the variations of these quantities but the fact that carriers may take different paths through the sample, avoiding those with particularly unfavorable R_{ij} or W_{ij}. For the high-temperature case, or nearest-neighbor hopping, the T dependence of κ will be dominated by the exponential in Eq. (51) with $\mathscr{E}_3$ representing a suitably averaged W_{ij}. For the "low-temperature" or VRH case the T dependence of κ will be essentially that shown in Eq. (52). The experimental T dependence seen in Fig. 11 in the lowest temperature range can clearly be well fitted by $\exp(\mathscr{E}_3/kT)$, consistent with nearest-neighbor hopping. Perhaps VRH can be seen at lower T's. Variable-range hopping will be found to fit the observations for amorphous semiconductors quite well.

8.3 Criterion for Localization

Conduction in the (b) band requires that an electron in (a) absorb a phonon of sufficient energy to make the transition to (b). Once the electron is in the (b) band, the conduction mechanism depends on whether the states are localized or extended. A criterion for determining this has been set up by Anderson (1958). He finds that if the actual energy spread ΔE of the states in a band is greater than ~5 times the bandwidth B that there would be due to overlap if the (isolated) states all had the same energy, all the states in the band are localized. If $\Delta E/B<5$ the states in the outer edges of the band may still be localized. The condition $\Delta E>5B$ apparently applies to the (a) band. It need not apply to the (b) band. It is not unlikely that the extent of the higher-energy state is larger, which would make B larger and the ratio $\Delta E/B$ smaller. The available data suggest that the (b) states are extended in the doping range of Fig. 11. In that case (see Fig. 11) $\mathscr{E}_2=E_i-E_{\mathrm{F}}$, the energy required to get the electron into the level E_i in the (b) band.

At concentrations higher than those of the Fig. 11 samples, probably $>10^{24}\ \mathrm{m}^{-3}$, the (a) and (b) bands have spread out and strongly overlap each other and the conduction band. The states at E_{F} are now extended, and the sample is degenerate, as can be seen from the experimental data for sample 58 in Figs. 6 and 9. The states in the low-energy tail of the impurity band are still expected to be localized, however; these correspond to electrons that, because of the randomness of the impurity distribution, are in particularly deep potential wells.

9. AMORPHOUS SEMICONDUCTORS

9.1 Atomic Structure

Amorphous materials lack the lattice periodicity that characterizes the crystalline solid state. The departure of the atomic positions in an amorphous solid from a periodic array is often represented by a radial distribution function (RDF). The RDF displays a number of peaks at r values giving the average distance between nearest neighbors, next-nearest neighbors, and so on. The radial fluctuation in these interatomic distances is manifested in the peak width. The lower part of Fig. 13 compares the RDFs for an assembly of identical atoms arranged in (a) crystalline form, (b) a glass, and (c) the gas phase (instantaneous picture). The RDF for the first few atomic positions in the amorphous solid is similar to that of the crystal because short-range order in the solid state is controlled by the dictates of chemical bonding. Beyond the first few neighbor distances in the amorphous solid, however, the RDF becomes progressively featureless, thus more gaslike than crystallike. Experimentally the RDF is determined by a complex procedure (Zallen, 1983) from the real-space Fourier transform of the x-ray scattering interference function.

One important result of the application of RDFs to the analysis of covalent amorphous semiconductors is the demonstration of the general applicability of the continuous random network (CRN) model first introduced by Zachariasen (1932) as an idealized description of covalent glasses such as silica. In the CRN model it is assumed that each constituent atom (including any impurity atom) is able to optimize its own bonding coordination needs. The optimum coordination number Z of any atom in a CRN is given by the Mott $8-N$ rule, where N is the number of valence electrons. According to the $8-N$ rule, which is based on elementary chemical reasoning, when $N \geqslant 4$ the coordination is $Z=8-N$, while for $N<4$ the optimum coordination is N. Figure 13 (upper part) illustrates the two-dimensional CRN of a three-coordinated (i.e., each atom has three bonds) glass. For a given composition all configurations of the glass that allow the constituent atoms to satisfy their local bonding requirements are equally probable. The effect of the long-range disorder on the electronic structure is expressed in terms of a randomly varying disorder potential (Mott and Davis, 1979).

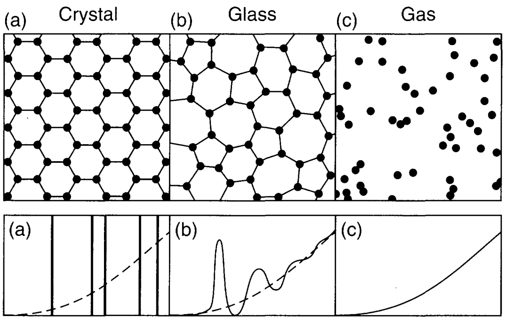

FIG. 13. Upper: Schematic representation of the atoms in (a) a covalent crystal, (b) a covalent glass network, and (c) a gas. Lower: Schematic represention of the corresponding RDFs (see text). (After Zallen, 1983.)

What, then, is the meaning of a structural defect in the CRN framework? In a crystal any atom not in a proper position in the lattice is a defect. In amorphous semiconductors, where CRN represents the ideal, the elementary defect is a miscoordinated atom, that is to say an atom with too few or too many bonds. Bonding miscoordination defects typically give rise to electronic states deep in the gap. These defects are now understood to play a major role in many of the key photoelectronic properties exploited in device applications (Street, 1991).

9.2 Electronic States

In the absence of lattice periodicity the vector **k** defined in Sec. 2.3 no longer has a meaning. For covalent-network amorphous solids whose crystalline counterparts have wide bands of order 5 eV or more, energy bands are still retained in the presence of disorder, but these must now contain localized as well as delocalized electronic states. Such disorder-modified bands can no longer be characterized by E-**k** dispersion relations such as that in Fig. 4, but must instead be described in terms of a density of states function $N(E)$. Important physical parameters like the carrier effective masses, defined in terms of the dispersion relations in the crystalline solid state, must be reformulated for amorphous semiconductors. The general problem of disorder-induced localization and its influence on the electronic band structure of amorphous covalent networks is very complex. It was elucidated by Anderson (1958), as discussed in Sec. 8. In hydrogenated amorphous silicon, *a*-Si:H, and in chalcogenide glasses, bandwidths are of order 5–10 eV, and the disorder estimated to be present, about 100–500 meV, is clearly too small to satisfy the Anderson criterion for complete localization. When the disorder potential is insufficient to induce complete localization of all states in the band, it can still act, as discussed earlier, to broaden the band edges into tails of localized states that extend into the gap. There is a significant difference in the mobility of carriers depending on whether they occupy extended or band-tail states. As a result, a demarcation energy separating extended states in the band from localized band-tail states was proposed (Mott, 1967). Because it was originally thought that the mobility should undergo a discontinuous change associated with the transition from band to hopping transport, this demarcation energy was said to define a mobility edge. It was suggested that for an amorphous semiconductor the appropriate analogy to the band gap of an intrinsic semiconductor crystal is the pseudogap bounded by the conduction- and valence-band mobility edges, denoted by E_C and E_V, respectively. The gap in an amorphous semiconductor is referred to as a pseudogap because it can be populated by a more or less continuous distribution of disorder-induced localized states (see Fig. 14). In addition to the band tails there are electronic states typically located deep in the gap, which arise from characteristic native defects associated with short-range disorder. Deep defects that can act as traps and recombination centers significantly influence the transport and photoelectronic properties of amorphous semiconductors and are believed to be involved in the mechanism of doping.

9.3 Native Defects

Every energy level of a defect can contain up to two electrons. For simplicity consider a defect that introduces a single level into the gap. That defect can exist in any of three charge states, corresponding to empty, singly, or doubly occupied, and its occupancy determines the position of the level it contributes to the gap relative to a mobility edge. The energies of the doubly occupied states differ from those of the singly occupied states because

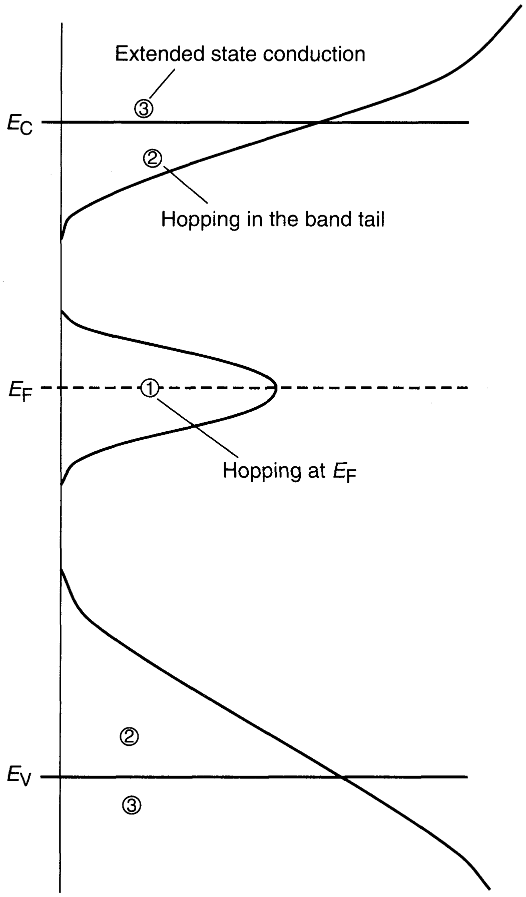

FIG. 14. Distribution in energy of the states, i.e., $N(E)$ vs E, in a prototypical undoped amorphous semiconductor.

1. the two electrons repel each other (Coulomb energy) and
2. addition of an electron can induce a bonding change or relaxation, which may either raise or lower the energy.

The algebraic sum of the two energies is called the correlation energy U. Both negative and positive correlation-energy defects exist and they have significantly different electronic properties. It can, for instance, be shown (Street, 1991) that for positive-U defects

1. the singly occupied, paramagnetic state has the minimum energy;
2. the Fermi level position is sensitive to dopants and there is minimal lattice relaxation.

On the other hand, for a negative-U defect

1. a large lattice relaxation is required for its formation;
2. the lowest energy state is diamagnetic; and
3. the Fermi level is constrained to lie between the defect energy levels.

The canonical example of a negative-U defect system is the valence alternation pair (VAP) that exists in chalcogenide glasses (Kastner *et al.*, 1976). *a*-Se, an elemental chalcogenide glass, provides the simplest example. According to the $8-N$ rule every Se atom should be bonded to two others, forming a one-dimensional network glass. A chain-terminating Se atom might be expected to have one unpaired or unbonded electron, called a dangling bond. Estimates based on viscosity data indicate that the number of such dangling bonds should exceed 10 ppm, which would be easily detectable by electron spin resonance (ESR). ESR is, however, not observed (Abkowitz, 1967), in *a*-Se or in any of the more complicated chalcogenide glasses. The neutral dangling-bond defect must be unstable against conversion to a diamagnetic species during glassy solidification. It has been proposed that this conversion process occurs by valence alternation and involves bond switching and electron transfer between chains (see AMORPHOUS MATERIALS, STRUCTURE OF).

9.4 Conductivity Mechanisms

It is clear from the preceding discussion that the analysis of conductivity in amorphous solids is greatly complicated by the broad energetic distribution and varied character of the states that can contribute channels for carrier motion. Instead of defining a single mobility for all states in the conduction band, or valence band, as done in Eqs. (2) and (26), it is useful to introduce a mobility $\mu(E)$ characteristic of the group of states in the range dE at E. The appropriate generalization of Eqs. (4) and (9) is then

$$\kappa=\int N(E)e\mu(E)f(E)dE\,. \tag{53}$$

Figure 14 is a schematic illustration of $N(E)$ vs E.

In Fig. 14 the Fermi level is shown contained in a submanifold of localized states near midgap. If the density of such states near the Fermi level is sufficiently high then, because their occupancy is heavily favored, these states will completely dominate the transport at low temperatures. For evaporated unhydrogenated amorphous Si, where the density of paramagnetic dangling-bond levels at the Fermi level is very large, it is found experimentally that $\ln\kappa\propto T^{-1/4}$ over a wide range of temperatures starting from the lowest (Street, 1991). As discussed in Sec. 8, this is the T dependence expected for VRH. In hydrogenated amorphous Si the dangling-bond defects from which the near-midgap states arise are in large measure eliminated by hydrogen complexation (i.e, passivation) and VRH is not observed.

At higher temperatures, nearest-neighbor hopping of carriers through band-tail states, regions 2 in Fig. 14, will contribute to the conductivity. When carriers of a single sign dominate, $1/\kappa$ vs T is activated, like the terms of Eq. (50). At still higher temperatures carrier transport is dominated by carriers that have been excited to states above the mobility edge E_C. κ may then be written

$$\kappa=\kappa_{E_C}\exp[-(E_C-E_F)/kT]\,, \tag{54}$$

where κ_{E_C} is the conductivity at the mobility edge and presumably also just above it. κ_{E_C} is called the minimum metallic conductivity. Mott and Davis (1979) have argued that for a three-dimensional case $\kappa_{E_C}=2.2e^2/Z^2\hbar a_E$, where Z is the coordination number and a_E the distance between localized states. It is difficult to verify this experimentally.

A reliable way of inferring the existence of localized states near band edges is via their influence on time-of-flight (TOF) drift mobility of injected carriers (Marshall, 1989). As discussed in Sec. 1.4, when the effective or drift mobility μ_d is limited by trapping it is reduced from the conductivity (free carrier) mobility μ_0 by the fraction of time that a carrier on average spends in traps, in this case the localized states adjacent to the mobility edge. When the injected carriers establish quasiequilibrium with trap and extended states in a small fraction of a transit time, the time of flight and the drift mobility are experimentally well defined. This condition is not satisfied, however, when there are deep traps with a range of depths, resulting in the carriers straggling to the contact. This behavior, called dispersive transport, is clearly exhibited by photoinjected holes in *a*-As_2Se_3. A successful statistical treatment of this type of transport has been carried out by Scher and Montroll (1975). On the basis of the Scher–Montroll treatment applied to time-of-flight and transient (bulk) photoconductivity data, *a*-As_2Se_3 is now generally believed to be characterized by a nearly exponential band tail, as suggested in Fig. 14.

9.5 Doping

It was thought earlier that the $8-N$ rule and the CRN model discussed in Sec. 9.1 would rule out the possibility of modifying the dc conductivity of an amorphous semiconductor by substitutional doping. In fact, *a*-Si:H exhibits doping phenomena (Spear and LeComber, 1975) with elements in groups V and III of the periodic table that are comparable to those in crystalline Si, though the underlying mechanism is believed to be quite different. Less widely appreciated and more subtle in its manifestations is a systematic doping effect in certain Se-rich chalcogenide glasses that is induced by both strong electropositive elements like the alkalis and electronegative substituents like the halogens. In both *a*-Si and *a*-Se (and related glasses) the underlying mechanism of doping is surprisingly similar and involves the interaction of dopant atoms with native defects.

The modification of conductivity by the doping of *a*-Si:H with P and B is dramatically illustrated in the data of Spear and LeComber (1975) shown in Fig. 15. Consider the case of P in *a*-Si:H. According to the $8-N$ rule P should be threefold coordinated (i.e., make three bonds) in its lowest-energy configuration. In fact the P doping efficiency in *a*-Si:H is very low, and it is generally believed that this is so because most of the substitutionally incorporated P is in threefold coordination, precisely as suggested by bonding and topology arguments. The smaller proportion of electronically active P dopant atoms that act as donors must be incorporated in fourfold coordination. The excess charge must somehow be compensated, and a large body of experimental evidence points to the dangling-bond defect D^- as the compensation center, with the Fermi level situated at an energy between P^+ and D^- levels. A complementary argument applies to the formation of ionized acceptors by B doping.

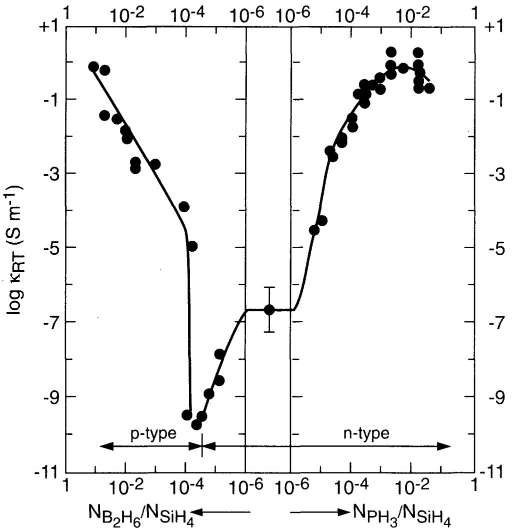

FIG. 15. Conductivity of *a*-Si:H at room temperature as a function of the dopant gases phosphine (PH_3) and borane (B_2H_6) introduced into silane (SiH_4), the deposition gas. (After Spear and LeComber, 1975.)

It should be noted that Mott (1976) suggested an analogous mechanism for the doping of *a*-Se by alkalis and halogens involving the native, negative-*U*, VAPs discussed earlier. In this case the VAPs, which normally exist in undoped material only as autocompensated defect pairs, would associate with some fraction of electropositive (e.g., Na) or electronegative (e.g., Cl) dopant atoms, thereby altering their relative populations while satisfying the requirement of overall charge neutrality. These predictions have been experimentally verified (Abkowitz, 1987). The doping effect of halogens on Se-rich chalcogenide alloys (e.g., Se:Te:As ternaries) is in fact systematically utilized in the design of the large-area photoconductors used in xerographic copiers and printers.

10. TWO-DIMENSIONAL AND ONE-DIMENSIONAL BEHAVIOR

The materials we have been discussing are all three dimensional (3D), in fact isotropic in most of their low-field properties. They can be made highly anisotropic, essentially 2D, by three different methods:

1. application of a high magnetic field;

2. confining the carriers in a thin layer at the surface; or
3. confining them in a thin layer of semiconductor between layers of different semiconductors, which may also be thin.

In the latter two cases further structuring the material to confine the carriers in one of the two dimensions can lead to quasi-one-dimensional behavior. This section is devoted to discussing the first two methods listed above; the third is the subject of HETEROSTRUCTURES, SEMICONDUCTOR.

10.1 Quantizing Magnetic Fields

10.1.1 Landau Levels. Solution of the wave equation has been carried out for a free electron in magnetic field. The same solution may be used for a carrier in the simple model of the band structure, with the replacement of m_0 by m^*. It is found that for a field **H** in the z direction the possible energy levels of the carrier are

$$E=(\hbar^2/2m^*)k_z^2+(\nu+\tfrac{1}{2})\hbar\omega_c, \tag{55}$$

where ν takes on integer values 0, 1, 2, ..., and ω_c is the cyclotron frequency He/m^*, the rate at which carriers rotate around the constant-energy surfaces in a magnetic field (see Sec. 5.6). It is seen that the energy of motion in the z direction is unaffected by the field, as expected because, as is seen in Eq. (19), there is no component of force parallel to **H**. The energy of motion perpendicular to the field, however, is quantized in steps of $\hbar\omega_c$. The set of all levels with a given ν and arbitrary k_z is referred to as the νth Landau level. The effect of the field has been to cause all the states with different k_x and k_y to "condense" on the nearest Landau level.

If an electromagnetic field of frequency ω_c is applied together with the magnetic field, a carrier can absorb a photon $\hbar\omega_c$, making a transition to the next Landau level. The transition corresponds to the carrier going from the orbit in **k** space characterized by ν to the $\nu+1$ orbit. This phenomenon is called cyclotron resonance. Its realization requires well-defined Landau levels ($kT\ll\hbar\omega_c$) and well-defined orbits (scattering time $\tau>2\pi/\omega$, the time of one rotation). Cyclotron resonance experiments, with **H** applied in different directions, have been very important for determining the band structure of semiconductors. In the electromagnetic field with frequency ω_c transitions are also possible from the level ν to $\nu-1$. The interlevel transitions are seen as magnetic-field–produced structure in optical transitions between bands, called magneto-optical absorption.

Quantization of the energy levels, Eq. (55), is accompanied by quantization of the areas of the orbits in **k** space corresponding to the allowed levels (Ziman, 1972). The quantization conditions result in a sharp oscillatory structure in the electronic density of states $N(E)$. $N(E)$ will have a sharp peak whenever E is equal to the energy corresponding to an orbit with maximum or minimum area, called an extremal orbit. This is an important effect for metals and for degenerate semiconductors, where many properties depend on $N(E_F)$, the density of states at the Fermi energy. At a value of **H** for which E of an extremal orbit coincides with E_F, $N(E_F)$ will be large. Thus it is expected that the values of many properties of metals and semiconductors, among them magnetoresistance, will have an oscillatory behavior as **H** varies. Specifically, the properties depending on $N(E_F)$ are periodic in $1/H$ (Ziman, 1972). The resulting oscillations in magnetoresistance, called the Shubnikov–de Haas effect, have been seen at low temperatures in HgSe and a number of other degenerate semiconductors. Oscillations with a period of $1/H$ have also been seen in interband optical absorption and magnetic susceptibility (de Haas–van Alphen effect).

10.1.2 The Quantum Hall Effect. At low temperatures and high magnetic fields, in a situation where a high density of electrons is confined in two dimensions (see Sec. 10.2), unusual behavior of the Hall effect has been found. The ratio of j_x to E_y, which by Eq. (31) should be ne/H_z, is found to change discontinuously with H_z, showing plateaus. It has been found that the plateaus correspond to the filling of successive Landau levels as the field changes. This effect, with accompanying phenomena, is known as the quantum Hall effect (*q.v.*).

10.2 Transport in Inversion Layers

10.2.1 Effect of Electric Field on Semiconductor Surface. Early in the modern history of semiconductors (1948) it was discovered that application of an electric field perpendicular to the surface of a semiconductor could increase or decrease carrier concentration at the surface, depending on the polarity of the field. Because it is required that, in thermal equilibrium, the Fermi level must be the same throughout a sample, the effect of the field is to bend the bands at the surface. In Fig. 16 we show some of the possible situations that can occur for a *p*-type semiconductor. The same situations occur, of course, for an *n*-type semiconductor. As can be seen from Eq. (14), when E_F is closer to the valence-band edge at the surface than in the interior, the hole density is greater at the surface, making it an accumulation layer. Where E_F is further from E_V in the surface than in the bulk, the surface is depleted. When the bands are bent further still, so that at the surface E_F is closer to E_C than E_V, the surface becomes *n* type, constituting an inversion layer. We should note, however, that the band bending may not be entirely due to the field; the bands are usually bent at the surface in the absence of the field because the surface has different defects and different defect levels from the bulk, and in fact may have a much larger number of defect levels.

10.2.2 Quantization in the Inversion Layer.

The case of the inversion layer is of particular interest. If the direction perpendicular to the surface is taken as *z*, electrons whose kinetic energy in the *z* direction lies below the bulk E_c are confined in a potential well. As is well known, if the well is deep enough, such electrons have quantized energy levels, the lowest level corresponding to a wave function with no nodes within the well, the next to a wave function with one node within the well, etc. (see QUANTUM MECHANICS). Some of these energy levels are indicated schematically in Fig. 16(c). The existence of these quantized levels was demonstrated by magnetoconductance measurements in which the Fermi level was swept through the Landau levels by varying the carrier concentration in the inversion layer as discussed in the last section. The characteristics of the Landau levels were those predicted for carriers confined to distances from the interface as small as 1–10 nm (Fowler, 1982). For such narrow wells the separation between levels is much larger than kT at 4 K and even somewhat higher temperatures, and the carriers occupy only the lowest quantum level at these temperatures. Of course, for each level there is a band of energies in the *x* and *y* directions, corresponding to all allowed wave vectors k_x and k_y, because motion parallel to the surface is unconfined. Thus 2D subbands are formed with electron energies, for the simple model of the band structure, given by the sum of E_n^0, the *n*th quantized energy level, and the kinetic energy $\hbar^2(k_x^2+k_y^2)/2m^*$.

10.2.3 Transport in the Inversion Layer.

Transport in inversion layers can be conveniently studied in a MOSFET, a metal-oxide-semiconductor field-effect transistor (see TRANSISTORS, FIELD EFFECT). The MOSFET structure is shown in Fig. 17. The silicon dioxide (SiO_2) thickness at the gate varies from ~10 nm to 1 μm. When $V_g=0$, electrons in the source cannot get to the drain, even though there is an attractive field due to the gate voltage V_D, because the *p*-Si region separating source and drain is a barrier. Application of a positive gate voltage creates an *n*-type channel, i.e., an inversion layer, at the *p*-Si surface with the SiO_2, permitting the electrons to go through. The voltage V_{sub} also affects the field at the SiO_2 interface and thus the number of carriers in the inversion layer. Thus varying V_{sub} and V_g appropriately allows a variation of electron concentration in the inversion layer at constant surface field or alternatively a variation of surface field at constant electron concentration.

Most of the transport measurements made on bulk samples may be made on the surface. Mobility may be determined from the measured conductivity of the surface layer in the usual way, using Eq. (3), but it is difficult to determine accurately the carrier concentration in the inversion layer. Hall measurements may be made and Hall mobility obtained from them, but, as discussed in Sec. 5.5, the ratio μ_H/μ is uncertain within about a factor of 2.

Conductivity and Hall-mobility measurements have been made for different orientations of the surface. Among the simplest to interpret is the [100] orientation. For that orientation in Si four of the valleys have the light mass in the *z* direction, while the other two have the heavy mass in that direction. The field in the potential well results in the latter two valleys having lower energies than the other four.

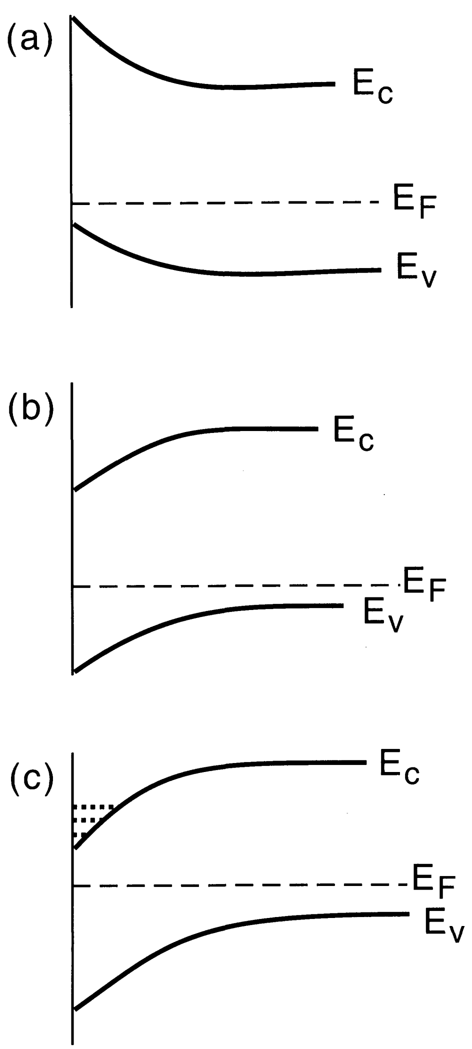

FIG. 16. Possible surface conditions for a *p*-type semiconductor: (a) accumulation layer; (b) depletion layer; (c) inversion layer. The dotted lines in (c) indicate, schematically, quantized energy levels.

The processes that scatter the carriers in the inversion layer are those discussed in Sec. 6, plus surface scattering. At low temperature, e.g., 4 K, the problem is relatively simple in that only the lowest quantized states are occupied and only impurity and surface scattering are significant. Impurity scattering can be due to either ions in the bulk or charges in the insulating SiO_2 layer. In practice, except for very heavy doping ($>10^{22}$ m^{-3}) in the bulk and for very clean oxides, the oxide ions are dominant. The samples in which impurity scattering is dominant at 4 K are likely to be degenerate, i.e., obey Fermi–Dirac statistics. For impurity scattering in a three-dimensional degenerate case we may use Eq. (45) for τ_I, replacing E by E_F. As the density of electrons in the inversion layer increases, E_F increases. Thus this equation predicts that at low temperatures impurity scattering gets less effective as the density of carriers in the inversion layer increases, which is in agreement with experiment. For detailed comparison with experiment, however, the Sec. 6 calculation of μ_I must be modified in a number of ways to describe better the inversion-layer situation (see Ando *et al.*, 1982).

As the carrier density increases and impurity scattering becomes less effective, surface roughness scattering comes to dominate. The surface roughness scattering is assumed to be due to random deviations from planarity of the Si–SiO_2 interface. It is characterized by two parameters, the rms deviation from surface planarity and the correlation length of the deviations. Information on these parameters may be obtained by transmission electron microscopy. The theory for this scattering process appears to be in reasonable agreement with experiment (Fowler, 1982).

At temperatures higher than 4 K phonon scattering is also important, specifically acoustic-phonon, nonpolar optical-phonon, and intervalley scattering. Extensive calculations are needed to take into account all the different phonons, and scattering within and between the different quantized bands and within and

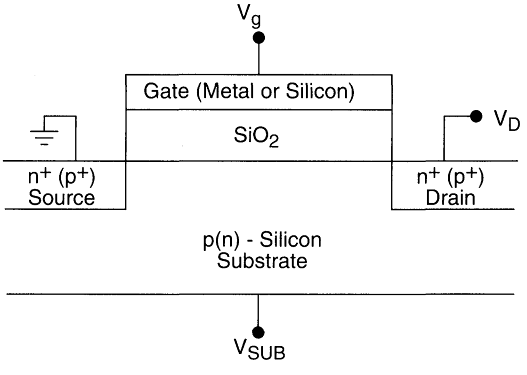

FIG. 17. Schematic MOSFET structure in cross section. The device may have a highly doped *n*-type (n^+) source and drain and a *p*-type substrate or a p^+ source and drain and an *n*-type substrate. (After Fowler, 1982.)

between valleys. For a review of the measurements and theoretical work see Ando *et al.* (1982).

10.2.4 Hot Electrons. Particularly at low temperatures, where the electron mobilities are high, the conductance is found to depend on the source–drain electric field. At 4 K mobility changes are observed at fields as low as a fraction of a V/cm. The mobility is found to rise initially when the electrons are heated by the field, as expected when the dominant scattering mechanism is ionized impurity scattering (see Sec. 7). The energy-loss mechanism that allows achievement of a steady state is acoustic-phonon scattering. For higher fields at 4 K impurity scattering becomes ineffective and optical-phonon scattering comes to dominate the scattering. In this field range mobility decreases with increasing field, as was described in Sec. 7 for hot carriers in the bulk. In the high-field limit drift velocity tends to saturate, as was also found for bulk *n*-Si. At 300 K, where impurity scattering is negligible, v_d vs $\mathbf{E}$ for the inversion layer is quite similar to that obtained for bulk *n*-Si, with no change up to $\sim 10^6$ V/m, beyond which v_d decreases with further increase in $\mathbf{E}$ and tends to saturate.

10.2.5 Localization. It is observed from ~2 to 50 K for samples with electron concentration in the inversion layer, n_{inv}, less than $10^{16}\ \mathrm{m}^{-2}$ that the conductivity is activated, i.e., proportional to $\exp(-W/kT)$ with W a constant. W decreases monotonically with increasing n_{inv} until it vanishes (Fowler, 1982). It was early suggested that these phenomena indicate localization, attributed to potential fluctuations due to charges known to be in the oxide layer. When the fluctuations are of short range they produce a band tail similar to the impurity band (see Sec. 8), as is found also in amorphous semiconductors. Indeed, the data for samples showing this type of behavior agree well with the prediction [Eq. (54)] for transport in a band with a localized tail. k_{E_C} is found to be approximately equal to the value $0.1e^2/\hbar$ predicted for electrons propagating above the mobility edge in a two-dimensional case. The activation energy decreases as n_{inv} increases, raising E_{F}, and vanishes when E_{F} goes through the mobility edge. Further evidence that this model is correct is provided by the fact that as the temperature is lowered κ is no longer activated, but rather $\ln\kappa \propto T^{-1/3}$. This is the temperature dependence predicted for variable-range hopping (see Sec. 8) in two dimensions.

10.2.6 Conductance Fluctuations. More recently work in this area has focused on transport properties of localized electrons at temperatures below 4 K in very small Si metal-oxide field-effect transistors, patterned to have widths as small as 50 nm. Conduction processes expected to be important for strongly localized electrons when $T < 1$ K are VRH and tunneling. Devices appropriate for studies of hopping conductivity are long ones (because tunneling decreases exponentially with distance), with narrow Si accumulation layers. In a system with dimensionality d when VRH dominates $\ln\kappa \propto -(T/T_{0d})^{1/d+1}$, where T_{0d} is a characteristic temperature that depends on the density of states $N(E_{\mathrm{F}})$ and d. It is found experimentally for these devices that, at temperatures $0.2\ \mathrm{K} < T < 3\ \mathrm{K}$, VRH may be either 1D or 2D depending on the value of V_g, which affects $N(E_{\mathrm{F}})$. (See ELECTRON STATES IN ZERO-, ONE-, AND TWO-DIMENSIONAL STRUCTURES for further discussion.)

A more striking effect in these samples is that at millikelvin temperatures conductance versus V_g has, superimposed on a general upward trend, a spiky behavior. The spikiness increases as T is lowered, the peaks being orders of magnitude above the background with widths at the top that are fractions of a millivolt. The spiky behavior results from the small size of the samples. As discussed in Sec. 8, an electron can hop only to states within a few kT of E_{F}. At 50 mK, for example, $2kT \simeq 10^{-5}$ eV. In a small device the number of states lying within this range is very small. Changing the gate voltage alters the energy of the electronic states in the channel relative to E_{F}. A small change in V_g may result in a large change in conductance G, for example, if it results in a particularly long hop being replaced by a short one. Further, each sample will have its own characteristic dependence of G on V_g because the locations of the levels, being determined by fluctuations in the oxide charge, are different in every sample.

Similar considerations apply to devices in which tunneling is dominant. In a short enough device an electron could tunnel directly from source to drain. Such short devices are difficult to make. In devices where the distance between source and drain is $\geqslant 500$ nm, resonant tunneling has been found of importance. In this process the electron tunnels from the source contact to a localized state in the channel and subsequently tunnels out of that state to the drain contact. For this process to occur the energy of the resonant localized state must equal, or be very close to, E_F in the source and drain contacts. Thus the situation is quite similar to that for VRH, and a similar spiky behavior is found in plots of G vs V_g (Fowler *et al.*, 1988).

GLOSSARY

Acceptor: Impurity that can accept an electron.

Accumulation Layer: Surface region in which the carrier density is greater than that in the bulk of the sample.

Activation Energy: When conductivity varies with temperature T as $\exp(-C/kT)$, C is called the activation energy.

Band Structure: The set of allowed energy E values of an electron in a crystal as a function of wave vector $\mathbf{k}$. The kinds of band structure referred to are (1) simple model—E vs $\mathbf{k}$, a parabola; (2) many valley—E vs $\mathbf{k}$ has a set of minima; (3) degenerate—E vs $\mathbf{k}$ consists of two bands, approximately parabolic, stuck together at $k=0$, and a third separated by a small energy interval.

Blocking Contact: A contact that completely cuts off the supply of charge to a sample.

Brillouin Zone: Unit cell in wave-vector or momentum space containing all the possible (discrete) wave-vector or momentum values.

Conduction Band: The lowest unfilled, or mostly unfilled, band of allowed energies.

Constant-Energy Surface: The surface or surfaces connecting points having the same energy in the Brillouin zone.

Dangling Bond: An unbonded electron. In amorphous silicon, for example, the paramagnetic defect state associated with a threefold-coordinated silicon atom.

Density of States: Number of states available for an electron per unit volume of a sample per unit energy range.

Density-of-States Effective Mass: The geometric mean effective mass for a given valley multiplied by the number of equivalent valleys to the 2/3 power.

Depletion Layer: Surface region in which the carrier density is smaller than that in the bulk of the sample.

Donor: Impurity that can give up an electron.

Doping: Addition of impurities that can increase conductivity of a semiconductor sample by adding electrons to the conduction band or holes to the valence band.

Effective Mass: A mass, which may be anisotropic, assigned to an electron or hole in a solid according to the energy–wave-vector relationship of the state it occupies.

Energy Gap E_g: The gap in energy between valence and conduction bands, empty of states in a sample without defects.

Fermi Energy E_F: The energy at which the probability of a state being occupied by an electron, under thermal equilibrium, is $\frac{1}{2}$.

Hall Effect: Application of a magnetic field perpendicular to the direction of current flow is found to give rise to an electric field perpendicular to both the magnetic field and the current. The Hall coefficient R_H, defined as the electric field developed per unit magnetic field per unit current density, is approximately equal to the reciprocal of the electron (hole) concentration times the charge on the electron.

Hall Mobility μ_H: The product of R_H (see **Hall Effect**) and conductivity, thus approximately equal to the mobility.

Hole: An unfilled state in a band more than half filled with electrons. The unfilled state acts as though it were filled with a charge equal in magnitude but opposite in sign to that of the electron.

Impurity Band: The group of states pulled out of the conduction (valence) band by the addition of donor (acceptor) impurities.

Inertial Mass: The reciprocal of the sum of reciprocals of the effective masses in the three principal-axis directions of the constant-energy surfaces for a given valley.

Injecting Contact: A contact able to supply charge to a semiconductor under the action of an electric field. An injecting contact that can be regarded as an infinite (i.e., nondepletable) charge reservoir, one able to fully satisfy the demands of the semiconductor even under space-charge–limited conditions, is called Ohmic.

Intrinsic Semiconductor: One in which carriers arise primarily from electrons that have crossed the gap into the conduction band, leaving behind holes in the valence band.

Inversion Layer: Surface region in which the current carriers have opposite sign to that of the carriers in the bulk.

Magnetoresistance: The change in resistance of a sample on application of a magnetic field.

Mobility μ: The net velocity in the electric field direction developed by an electron or hole due to acceleration by unit electric field.

Mobility Edge: The energy that separates localized from extended states near the band edge of an amorphous semiconductor.

Momentum Relaxation Time τ: The time for excess momentum, such as that created by an electric field, to decay to $1/e$ of its original value when the source of the excess is removed. τ is approximately equal to the mean free time of an electron or hole between scattering processes.

Ohm's Law: The statement that current increases linearly with applied voltage or electric field. This ceases to be true at high enough fields.

Phonon: A quantum of vibrational energy of one of the normal modes into which the random thermal motion of the lattice atoms may be decomposed.

Photoconductivity: An increase of electrical conductivity induced by the absorption of light.

Space-Charge–Limited Current (SCLC): The current demanded by a solid at fields that make the transit time of an excess injected carrier passing through the bulk to a collecting contact less than the time required to induce its local neutralization. Under these conditions the current is limited by the mutual repulsion of the transiting carriers.

Valence Band: The uppermost filled or nearly filled band of allowed energies.

Variable-Range Hopping: The transfer of an electron between localized sites at temperatures for which the thermal energy (required to make up the energy difference between sites) is so small that progressively larger hops are required with decreasing temperature for sites close enough in energy to the original to be found.

List of Works Cited

Abkowitz, M. A. (1967), *J. Chem. Phys.* **46**, 4537–4538.

Abkowitz, M. A. (1987), in: M. A. Kastner, G. A. Thomas, S. R. Ovshinsky (Eds.), *Disordered Semiconductors*, New York: Plenum, pp. 205–217.

Altarelli, M., Bassani, F. (1982), in: W. Paul (Ed.), *Handbook on Semiconductors*, Vol. 1, Amsterdam: North-Holland, Chap. 5.

Anderson, P. W. (1958), *Phys. Rev.* **109**, 1492–1505.

Ando, T., Fowler, A. B., Stern, F. (1982), *Rev. Mod. Phys.* **54**, 437–672.

Auston, D. H. (1990), *Phys. Today* **43** (2), 46–54.

Chattopadhyay, D., Queisser, H. J. (1981), *Rev. Mod. Phys.* **53**, 745–768.

Conwell, E. M. (1958), *Proc. IRE* **46**, 1281–1300.

Conwell, E. M. (1967), *High Field Transport in Semiconductors*, New York: Academic.

Conwell, E. M. (1970), *Phys. Today* **23** (6), 35–41.

Conwell, E. M. (1982), in: W. Paul (Ed.), *Handbook on Semiconductors*, Vol. 1, Amsterdam: North-Holland, Chap. 10.

Debye, P. P., Conwell, E. M. (1954), *Phys. Rev.* **93**, 693–706.

Dunlap, W. C. (1959), in: K. Lark-Horovitz, V. Johnson (Eds.), *Methods of Experimental Physics*, Vol. 6B, New York: Academic, pp. 32–70.

Einspruch, N. G. (1965), in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, Vol. 17, New York: Academic, pp. 217–268.

Fletcher, K., Butcher, P. N. (1972), *J. Phys. C* **5**, 212–224.

Fowler, A. B. (1982), in: W. Paul (Ed.), *Handbook on Semiconductors*, Vol. 1, Amsterdam: North-Holland, Chap. 11B.

Fowler, A. B., Wainer, J. J., Webb, R. A. (1988), *IBM J. Res. Develop.* **32**, 372–383.

Fritzsche, H. (1955), *Phys. Rev.* **99**, 406–419.

Gray, Dwight E. (Ed.) (1957), *American Institute of Physics Handbook*, New York: McGraw-Hill.

Handbook of Chemistry and Physics (1975), Cleveland: CRC.

Harrison, W. A. (1980), *Electronic Structure and the Properties of Solids*, San Francisco: Freeman.

Herring, C. (1955), *Bell Syst. Tech. J.* **34**, 237–290.

Herring, C., Vogt, E. (1956), *Phys. Rev.* **101**, 944–961.

Hess, K. (1990), *Phys. Today* **43** (2), 34–42.

Kastner, M., Adler, D., Fritzsche, H. (1976), *Phys. Rev. Lett.* **37**, 1504–1507.

Lampert, M. A., Mark, P. (1970), *Current Injection in Solids*, New York: Academic.

Madelung, O. (1991), "Semiconductors," in: R. Poerschke (Ed.), *Data in Science and Technology*, Berlin: Springer-Verlag.

Marshall, J. M. (1989), in: H. Fritzsche (Ed.), *Amorphous Silicon and Related Materials*, Vol. 1B, Singapore: World Scientific, pp. 799–824.

Mott, N. F. (1976), *Philos. Mag.* **34**, 1101–1108.

Mott, N. F., Davis, E. A. (1979), *Electronic Processes in Non-Crystalline Materials*, Oxford: Clarendon.

Ryvkin, S. M. (1964), *Photoelectric Effects in Semiconductors*, New York: Consultants Bureau.

Scher, H., Montroll, E. W. (1975), *Phys. Rev. B* **12**, 2455–2497.

Shockley, W. (1950), *Electrons and Holes in Semiconductors*, New York: Van Nostrand.

Spear, N. E., LeComber, P. G. (1975), *Solid State Commun.* **17**, 1193–1195.

Street, R. A. (1991), *Hydrogenated Amorphous Silicon*, Cambridge, U.K.: Cambridge University Press.

Zachariasen, W. H. (1932), *J. Am. Chem. Soc.* **54**, 3841–3851.

Zallen, R. (1983), *The Physics of Amorphous Semiconductors*, New York: Wiley.

Zawadzki, W. (1982), in: W. Paul (Ed.), *Handbook on Semiconductors*, Vol. 1, Amsterdam: North-Holland, Chap. 12.

Ziman, J. M. (1972), *Principles of the Theory of Solids*, Cambridge, U.K.: Cambridge University Press.

Further Reading

Beer, A. C. (1963), *Galvanomagnetic Effects in Semiconductors*, New York: Academic. Detailed treatment of theory and experiment for crystalline semiconductors.

Conwell, E. M. (1967), *High Field Transport in Semiconductors*, New York: Academic. Qualitative and quantitative discussion of hot electrons in crystalline semiconductors.

Lampert, M. A., Mark, P. (1970), *Current Injection in Solids*, New York: Academic. Standard text for physics of charge transfer across interfaces.

Mott, N. F., Davis, E. A. (1979), *Electronic Processes in Non-Crystalline Materials*, Oxford: Clarendon. Comprehensive description of electronic and optical properties of amorphous semiconductors.

Paul, W. (Ed.) (1982), *Handbook on Semiconductors*, Vol. 1, Amsterdam: North-Holland. General introduction to band structure and transport aimed at graduate students and young researchers in the field.

INSULATORS, STRUCTURE OF

CRONIN B. VINING, *Jet Propulsion Laboratory/California Institute of Technology, Pasadena, California, U.S.A.*

	Introduction	85
1.	**Properties of Insulators**	86
1.1	Dielectric Properties	86
1.1.1	The Complex Permittivity	86
1.1.2	Polarization Mechanisms	87
1.1.3	Typical Permittivity Behavior	88
1.1.4	Piezoelectricity	90
1.1.5	Pyroelectricity	91
1.1.6	Ferroelectricity	91
1.1.7	Lattice Dynamics and Soft Modes	92
1.1.8	Antiferroelectric and More Complex Order	92
1.2	Magnetism	93
1.2.1	Diamagnetism	93
1.2.2	Paramagnetism	93
1.2.3	Ferromagnetism	93
1.2.4	Antiferromagnetism and More Complex Order	94
1.3	Thermal Properties	94
2.	**Insulating Materials**	95
2.1	Elemental Insulators	95
2.2	Binary Compound Insulators	96
2.2.1	Covalent Insulators	96
2.2.2	Ionic Insulators	96
2.2.3	Binary Oxides and Other Binary Insulators	96
2.3	Multinary Compound Insulators	97
2.3.1	Rochelle Salts and Other Order/Disorder Systems	97
2.3.2	Perovskites and Displacive Transitions	98
2.3.3	Bronzes, Pyrochlores, and Nonmagnetic Oxides	99
2.3.4	Ferrites: Spinels and Garnets	100
2.3.5	Organics	100
3.	**Summary**	101
	Glossary	101
	Works Cited	102
	Further Reading	102

INTRODUCTION

Most of the substances around us, whether natural or synthetic, are electrical insulators. The distinction between insulators and metals has been recognized for centuries, and even children can tell the difference: metals are typically soft and malleable and can be polished to a high luster, while common insulators are hard, brittle, and often rather dull. Most electrical insulators are also thermal insulators. Hot beverages are served in electrical insulators, not metals, to retain the heat better and to prevent injury to the hand while drinking. These are among the more common properties, but insulators exhibit a broad range of characteristics that are perhaps less obvious, but just as important.

Insulators are needed whenever electrical current flow must be controlled, whether as a nerve impulse in a living organism, a bit of information in a digital computer, or the industrial transmission of electrical power. Certain insulators greatly enhance the performance of capacitors by increasing the amount of charge that can be maintained using a particular applied voltage. Other insulators respond to mechanical, thermal, magnetic, electrical, or optical stress in ways that make possible a wide variety of applications such as transducers, filters, motors, detectors, resonators, oscillators, imagers, memory devices, video displays, and many more. The tremendous variation of the properties of insulators plays a major role in the modern technological world, and it is important, therefore, to have some understanding of the relationship between their structure and their properties.

An insulator may be defined as any material that does not allow net movement of electric charge from one region of the material to any other region. The distinction between insulator and conductor, however, is relative and concerns both the conductivity of

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

the surrounding medium and the duration of the observation. Every material conducts electricity to some extent, so that given enough time even the best electrical insulators will allow some charge motion. Among the best electrical insulators are diamonds, with electrical conductivities of 10^{-13} S m^{-1} or less. But the best metals, like silver, have conductivities of around 10^{8} S m^{-1}, more than 20 orders of magnitude larger. As a rough guideline, anything with a conductivity less than 10^{-3} S m^{-1} may be considered an insulator.

All materials are composed of atomic nuclei and enough electrons to maintain charge neutrality. An electron in such a material may be tightly bound, always remaining near a particular atomic nucleus (or group of nuclei), or it may be free to migrate from one region in the material to another. A material with a large number of free electrons is either a semiconductor or a metal. In an insulator, there is a negligible number of free carriers and a substantial activation energy for the creation of free carriers from the bound electrons.

Electrical insulation can result from a number of distinct mechanisms. In gases, the thermal activation energy is a result of the high ionization energy of isolated atoms. In liquids, glasses, and amorphous solids the thermal activation energy is a result of the atomic disorder that makes migration of carriers a relatively tortuous process. Crystalline solids sometimes have exactly enough electrons to fill entire bands of electronic states completely, leaving the next state available for free electrons at a significantly higher energy, called the band-gap energy. In plastic, polymeric, and liquid-crystal materials various combinations of ionization energies, disorder, and band filling may contribute to a substantial activation energy.

This binding of electrons in an insulator implies that the electrical conductivity, which is proportional to the number of free carriers, will be thermally activated:

$$\sigma=\sigma_o e^{-E_g/2kT}, \tag{1}$$

where E_g is the band-gap energy (or possibly some other, related activation energy). The division is somewhat arbitrary, but typically the term "electrical insulator" is reserved for a material with a band gap of at least 3 eV. The following sections are most appropriate for the properties and structure of homogeneous, crystalline insulators, but some other insulators will be briefly mentioned, and the concepts involved are of general utility. For more detailed information on amorphous materials, electronic structure of solids, polymers, liquid crystals, ceramics, and composite materials see the related articles in this encyclopedia. An overview of the important properties of insulators follows, including a brief discussion of the "soft mode" concept. The section on materials discusses the occurrence of insulating solids among the elements, binary compounds, and certain classes of more complex materials.

1. PROPERTIES OF INSULATORS

1.1 Dielectric Properties

The term "dielectric" emphasizes the response of an insulator to an applied field. The principal properties of interest in this regard are the permittivity (ϵ) and the loss tangent ($\tan\delta$), which characterize both the degree of polarization in an applied field and the losses that result from application of an alternating frequency field. A brief outline of dielectric properties is provided below, but for a more complete discussion see DIELECTRIC PROPERTIES OF INSULATORS.

1.1.1 The Complex Permittivity. When an electric field E is applied to a dielectric material, the charges within the material experience forces that tend to redistribute the charges. This redistribution is characterized by the polarization P, which represents the electric dipole moment induced in the material by the applied electric field. It is also convenient to define a third quantity, called the displacement field D, which represents the effects of the actual charges in the material and is given by

$$D=\epsilon_{\mathrm{v}}E+P, \tag{2}$$

where ϵ_{v} is the permittivity of the vacuum.

In an ordinary dielectric, the polarization, the electric field, and the displacement field will all be proportional to each other. We define

$$D=\epsilon E=\epsilon_{\mathrm{v}}\epsilon_r E, \tag{3}$$

where ϵ_r is the relative permittivity, a unitless quantity sometimes called the dielectric constant, although this term is more properly reserved for situations in which ϵ_r is actually independent of frequency. The polarization is now given by

$$P=\epsilon_{\text{v}}(\epsilon_r-1)E. \tag{4}$$

The vacuum clearly has no polarization and $\epsilon_r^{\text{vacuum}}=1$.

As an example, consider a capacitor in a simple charging circuit. The effect of inserting a dielectric material between the plates of the capacitor is to increase the capacitance (the charge the capacitor can hold) by the factor ϵ_r. If the applied voltage is held constant, eventually only a small leakage current will flow because of parasitic conduction through the dielectric material. If the applied voltage is sinusoidal in time, there will be a leakage current in phase with the applied voltage as well as a charging current that precedes the applied voltage by a 90° phase angle.

Thus, the total current through the capacitor (Fig. 1) is given by

$$i=j\omega\epsilon' E+\omega\epsilon'' E=\epsilon\frac{dE}{dt}, \tag{5}$$

where the first term represents the charging current and the second term represents the loss current. The permittivity is now considered to be a complex quantity with both a real and an imaginary component:

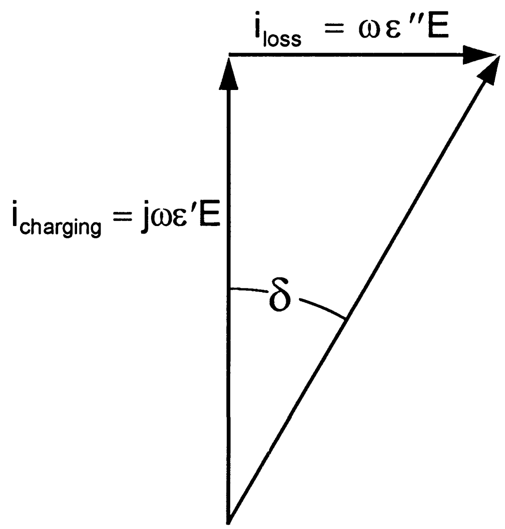
$i_{\text{loss}} = \omega\varepsilon'' E$
$i_{\text{charging}} = j\omega\varepsilon' E$
δ

FIG. 1. Charging and loss currents, where tanδ is the loss tangent.

$$\epsilon=\epsilon'-j\epsilon'', \tag{6}$$

where the imaginary component is connected to the frequency-dependent electrical conductivity as

$$\sigma=\omega\epsilon''. \tag{7}$$

A useful quantity for characterizing an insulator is the so-called loss tangent (also called the dissipation factor), defined as the ratio of the loss current to the charging current:

$$\tan\delta=\frac{i_{\text{loss}}}{i_{\text{charging}}}=\frac{\epsilon''}{\epsilon'}. \tag{8}$$

The loss tangent is a material property related to the complex permittivity, which characterizes the electrical losses in an insulator.

1.1.2 Polarization Mechanisms. To understand the origin of polarization effects in materials it is best to begin with the smallest structural units involved, either the individual atoms or small groups of atoms. The electric dipole moment of a single atom (or small group) is taken to be proportional to the local electric field as

$$p_i=\alpha_i E_{\text{local}}, \tag{9}$$

where α_i is the polarization factor. The polarization factor is a microscopic quantity and is convenient because it can be calculated from theory. However, the α_i's are connected to the local electric field, not to macroscopic quantities that may be measured in the laboratory: E, P, and D. In the simplest approximation, the local field is given by

$$E_{\text{local}}=E+P/3\epsilon_{\text{v}}. \tag{10}$$

Combining Eqs. (4), (9), and (10) leads to the Clausius-Mossotti equation, completing the required relationship between the microscopic polarization factors and the macroscopic relative permittivity:

$$\frac{\epsilon_r-1}{\epsilon_r+2}=\frac{1}{3\epsilon_{\text{v}}}\sum_i N_i\alpha_i, \tag{11}$$

where N_i is the number density of dipoles with polarization factor α_i.

The polarization has its origins in the structure of the insulator and may have a number of contributing factors. Figure 2 summarizes the most important mechanisms for polarization in a solid. All materials are subject to *electronic polarization* (also called *atomic* or *optical polarization*) because in an applied electric field the atomic nucleus is displaced with respect to the surrounding electrons, inducing a dipole moment. In the absence of an applied field, the charges within an atom will tend to restore the symmetry of the atom and remove the dipole moment. An atom, therefore, may be thought of as a simple harmonic oscillator with a rather high natural frequency, typically in the ultraviolet range due to the small mass of the electrons. Using this description, and including a damping term to represent losses in the oscillator, the electronic polarization can be shown to have the form given in Fig. 3.

Ionic polarization results from the displacement of cations (positive ions) in the direction of the applied field and anions (negative ions) in the opposite direction, as indicated schematically in Fig. 2. These processes also may be thought of as resulting from simple harmonic oscillators, in this case involving the relative motion of cations and anions in the crystal. Ionic polarization, then, has the same functional form as electronic polarization, but since the motion here involves entire atoms, the natural frequencies of the oscillators are much lower and correspond to lattice vibration frequencies, typically in the infrared or visible region.

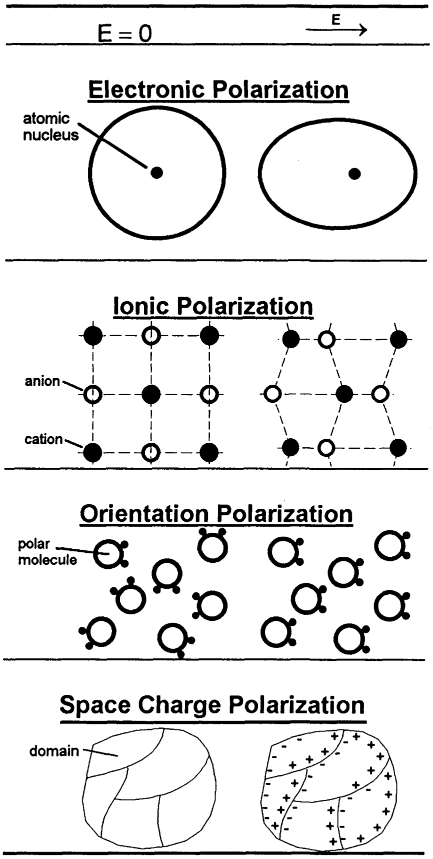

FIG. 2. Major polarization mechanisms. The displacement of the charges has been strongly exaggerated.

Orientation (or *dipole*) *polarization* occurs when a structural repeat unit or a polar domain with a permanent dipole moment (p_m) experiences a torque, tending to orient the dipole in the direction of the applied field. Thermal motion tends to randomize the orientation of dipoles. In the absence of an applied field a system of oriented dipoles will "relax" toward a random distribution of orientations with a characteristic time τ, called the relaxation time. Using this relaxation model, the orientation polarization can be shown to have the form shown in Fig. 3, with characteristic frequencies (τ^{-1}) ranging into the microwave region.

Space charge (or interfacial) polarization represents the contribution from charge carriers that normally are free to move through the solid, but become impeded by grain boundaries, electrodes, and the like as suggested in Fig. 2. In this case, a space charge eventually builds up, generating a field to counter the applied field until current ceases to flow.

1.1.3 Typical Permittivity Behavior. For the electronic, ionic, and orientational mechanisms three distinct frequency regimes may be identified: first, a low-frequency regime, where the applied field varies slowly enough for the polarization to track the applied field closely and the polarization factor is essentially a real constant; second, a high-frequency regime where the applied fields vary so rapidly that the polarization can not track the applied field at all and the polarization factor tends to zero; and finally, an intermediate regime near the characteristic frequency of the oscillators or relaxation mechanism, where both the real and imaginary components of the polarization factor vary relatively rapidly with frequency. It is only in this intermediate-frequency regime that α'' becomes of significant magnitude.

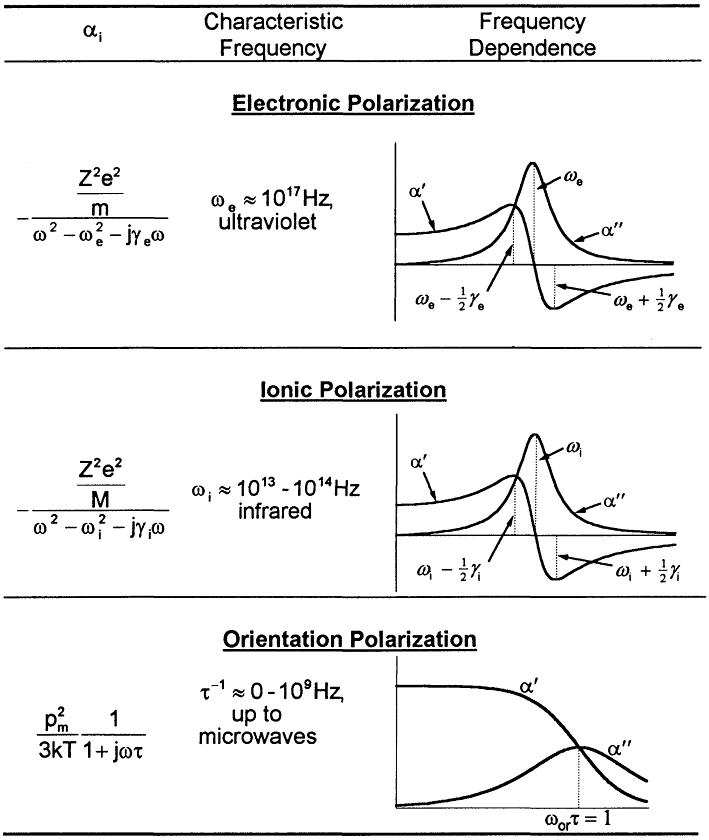
α_i
Characteristic Frequency
Frequency Dependence
Electronic Polarization
$-\frac{\frac{Z^2e^2}{m}}{\omega^2-\omega_e^2-j\gamma_e\omega}$
$\omega_e \approx 10^{17}$Hz, ultraviolet
α'
ω_e
α''
$\omega_e-\frac{1}{2}\gamma_e$
$\omega_e+\frac{1}{2}\gamma_e$
Ionic Polarization
$-\frac{\frac{Z^2e^2}{M}}{\omega^2-\omega_i^2-j\gamma_i\omega}$
$\omega_i \approx 10^{13}-10^{14}$Hz infrared
α'
ω_i
α''
$\omega_i-\frac{1}{2}\gamma_i$
$\omega_i+\frac{1}{2}\gamma_i$
Orientation Polarization
$\frac{p_m^2}{3kT}\frac{1}{1+j\omega\tau}$
$\tau^{-1} \approx 0-10^9$Hz, up to microwaves
α'
α''
$\omega_{or}\tau = 1$

FIG. 3. Frequency dependence of the major polarization mechanisms.

Ordinarily the four polarization mechanisms may be considered as acting independently, in which case the total polarization factor is given by a sum of the four contributions:

$$\alpha_i=\alpha_{i,e}+\alpha_{i,i}+\alpha_{i,o}+\alpha_{i,s}\ , \qquad (12)$$

where the second subscript indicates the contributions from the different mechanisms. Often, since the characteristic frequencies are so different for the different mechanisms, only one of the contributions will have a strong frequency dependence in any particular range of frequencies and the other mechanisms can be considered to provide a frequency-independent background.

The relative permittivity of a material can now be calculated using the sum of the individual contributions to the polarization factor and the Clausius-Mossotti relation. The relative permittivity typically has a frequency dependence similar to the frequency dependence of the contributing polarization factors, represented in Fig. 3. The real component of the relative permittivity is larger at low frequencies (usually called the static dielectric constant, ϵ_0) than at high frequencies (usually called the optical dielectric constant, ϵ_∞) and will vary rapidly only over a relatively narrow range around the characteristic frequency. At frequencies well beyond the ultraviolet, the relative permittivity would drop even further as even electronic polarization processes cannot track the applied field at these frequencies.

At frequencies below the ultraviolet, many materials exhibit a frequency-independent contribution to the polarization from electronic polarization processes and a frequency-dependent contribution to the polarization from ionic polarization processes. The relative permittivity in this case is given by

$$\epsilon_r(\omega)=\epsilon_\infty+\frac{\epsilon_\infty-\epsilon_0}{\omega^2/\omega_T^2-1}\ , \qquad (13)$$

where

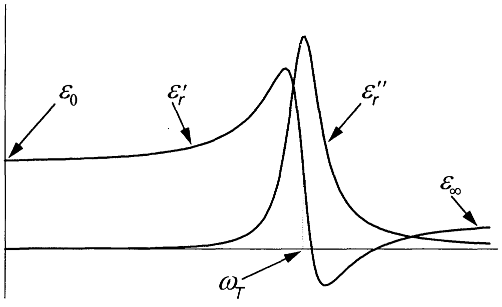

FIG. 4. Real and imaginary components of the relative permittivity typical of an ionic contribution to the polarization.

$$\omega_T^2=(\omega_i^2+j\gamma_i\omega)\left(\frac{\epsilon_\infty+2}{\epsilon_0+2}\right). \tag{14}$$

The real and imaginary components are shown in Fig. 4, where the change in sign of the real component just above ω_T is typical of an ionic contribution to the polarization.

Qualitatively different behavior characterizes the dipole orientation polarization mechanism, where the relative permittivity is given by

$$\epsilon_r(\omega)=\epsilon_\infty+(\epsilon_0-\epsilon_\infty)/(1+j\beta) \tag{15}$$

with

$$\beta=[(\epsilon_0+2)/(\epsilon_\infty+2)]\omega\tau\,. \tag{16}$$

Here $\epsilon'_r(\omega)$ decreases monotonically from ϵ_0 at low frequencies to ϵ_∞ at high frequencies, mimicking the frequency dependence of the orientation polarization factor shown in Fig. 3. In this case, a plot of $\epsilon''_r(\omega)$ as a function of $\epsilon'_r(\omega)$ has the particularly simple form of a circle, as indicated in Fig. 5. This type of graphical representation, called a Cole-Cole diagram, can be used to determine the relaxation time, τ. The procedure involves connecting a point $(\epsilon'_r(\omega),\ \epsilon''_r(\omega))$ at any particular frequency with either the high-frequency limit $(\epsilon_\infty,0)$ or the low-frequency limit $(\epsilon_0,0)$, as indicated by the straight dotted lines in Fig. 5. β may be estimated from the slope of either line and the relaxation time is then given by Eq. (16). Cole-Cole diagrams are useful in identifying relaxation-type polarization behavior, such as is typical of dipole orientation polarization, since other mechanisms appear more complex and qualitatively distinct when represented in this manner, as suggested in Fig. 6.

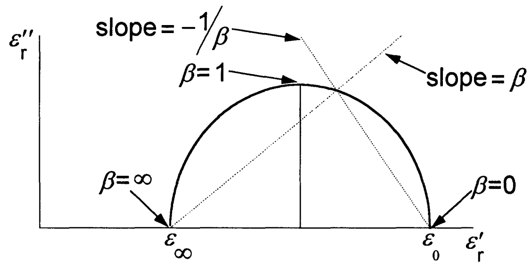

FIG. 5. Cole-Cole diagram for an ideal relaxation-time mechanism.

1.1.4 Piezoelectricity. Piezoelectric materials exhibit two characteristic effects: polarization in response to applied mechanical stress, and mechanical deformation in response to an applied electric field. Both effects are linear and change sign when the applied stress (mechanical or electrical) changes sign. A crystal with a center of symmetry cannot be piezoelectric, since mechanical stress alone will not disturb the balance among the dipole moments within a unit cell. Crystals with a center of symmetry will mechanically deform in an applied electric field, but the deformation will be proportional to the square of the applied electric field. This effect, also called electrostriction, is independent of the sign of the applied field and is usually quite small in magnitude.

Piezoelectricity is possible in any noncentrosymmetric crystal (which includes all the pyroelectric materials discussed below), as illustrated in Fig. 7. The Curie brothers (1880) discovered piezoelectricity in quartz, tourmaline $\{Na(Mg,Fe,Mn,Li,Al)_3Al_6[Si_6O_{18}](BO_3)_3(OH,F)_4\}$, and Rochelle salt. More recently (see PIEZOELECTRIC DEVICES) there has been great interest in organic materials such as polyvinylidene fluoride (PVDF).

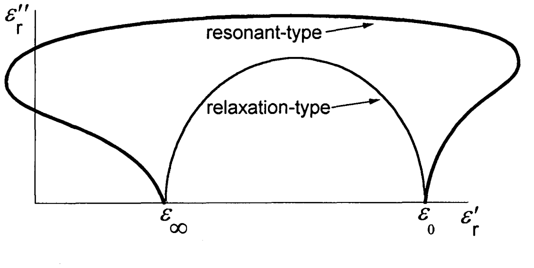

FIG. 6. Cole-Cole diagram illustrating qualitatively distinct behavior for different polarization mechanisms.

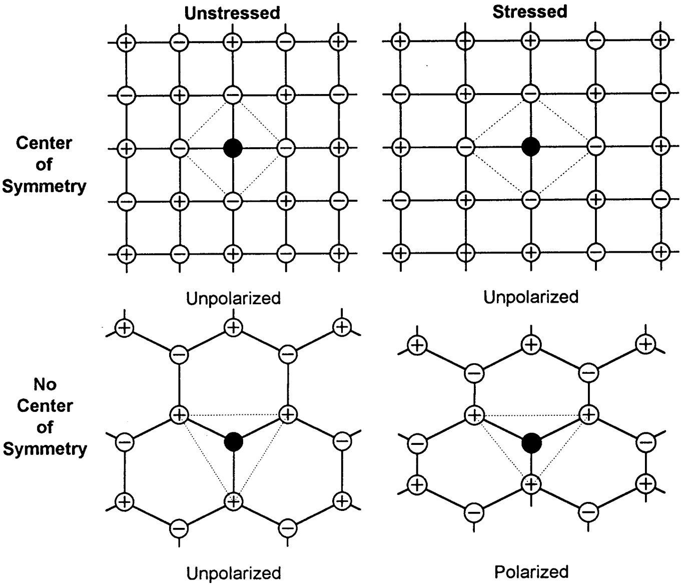

FIG. 7. Crystals with a center of symmetry are not polarized, even under stress. Piezoelectrics, without a center of symmetry, polarize under stress.

1.1.5 Pyroelectricity. Certain materials are called pyroelectric because they appear to generate a spontaneous polarization when heated (or cooled). In fact, polarization is always present internally in these materials, but is ordinarily masked by neutralizing charges collecting on the exposed surfaces. A net polarization is temporarily observed with temperature changes until the surface charges can reequilibrate with the temperature-dependent internal polarization. The observation of an internal polarization implies severe constraints on the symmetry of the crystal: there can be no mirror planes perpendicular to the axis of polarization and no rotational symmetry except along an axis parallel to the polarization axis.

1.1.6 Ferroelectricity. A material that is nonpyroelectric above a certain temperature T_C (the Curie temperature) and pyroelectric below T_C is called a ferroelectric. Just below T_C the spontaneous polarization is usually small and may be readily reversed in direction by an applied field opposing the original polarization, as shown by the characteristic hysteresis loop in Fig. 8. This hysteretic behavior in the polarization and the ability to reverse the polarization by an applied field is sometimes taken as the defining property of a ferroelectric. The term ferroelectric is in analogy to similar behavior observed in the magnetization of ferromagnetic materials.

The origin of ferroelectricity lies in the interaction between the local polarization and the macroscopic electric field. Solving Eqs. (9) and (10) for the total polarization yields

$$P = \frac{\Sigma_i N_i \alpha_i}{1-(1/3\epsilon_0)\Sigma_i N_i \alpha_i} E. \tag{17}$$

The orientational contribution to the polarization, $\alpha_{i,o}$, typically increases as $1/T$ as the temperature decreases, leading eventually to a zero in the denominator of Eq. (17) at some temperature, implying spontaneous polarization. Here, the local field produced by a dipole has a sufficiently strong effect on neighboring dipoles that all of the dipoles prefer to orient in the same direction, even in the absence of

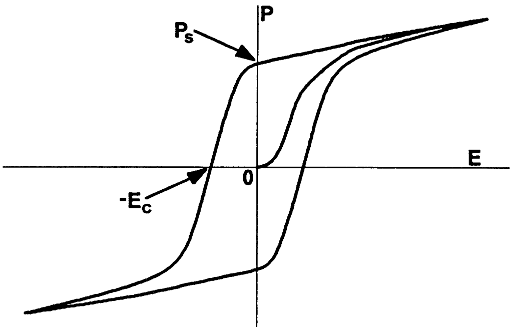

FIG. 8. The ferroelectric hysteresis loop indicating the spontaneous polarization (P_s) and coercive force ($-E_c$) required to reduce P_s to zero.

an applied field. This critical temperature is the Curie temperature T_C, and the dielectric susceptibility takes the form

$$\chi \equiv \epsilon_r - 1 = C/(T - T_C), \qquad (18)$$

where C is a constant, again in close analogy to the magnetic susceptibility of a ferromagnet.

The dipole moments that can be induced by applied fields or stress above T_C become permanent below T_C. The mechanisms responsible for ferroelectricity vary from material to material and will be discussed in the materials section below. Generally, however, ferroelectricity results from a lowering of the symmetry of the crystal lattice, similar to that suggested in Fig. 7 except that an applied stress is no longer required to maintain the polarized state. Blinc and Žekš (1974) list over 300 known inorganic ferroelectric compounds (not counting solid solutions), and Fousek (1991) estimates that over 50 families of liquid-crystal ferroelectrics are now known. A few representative materials exhibiting a variety of properties are listed in Table 1.

1.1.7 Lattice Dynamics and Soft Modes. The permanent polarization in the low-temperature form of a ferroelectric represents a modification of the lattice compared to the more symmetric, high-temperature form of the material. The lattice modification can be examined with experiments such as x-ray or neutron diffraction that are sensitive to the atomic positions. Cochran (1959) and Anderson (1959) introduced the concept that ferroelectricity results from an instability of one of the normal lattice vibrational modes. The characteristic frequency of the so-called soft mode, referring to the unstable lattice vibration, decreases on approach to the Curie temperature, T_C, and the restoring force tending to return the displaced atoms to their normal positions approaches zero. The soft mode then freezes in and the atoms become permanently displaced, resulting in the ferroelectric state.

While a detailed discussion of soft modes is well beyond the scope of this article, it is important to emphasize the close connection between the stability of the lattice with respect to distortion and the occurrence of ferroelectricity. Much of the study of ferroelectrics has been directed toward a detailed understanding of the occurrence, symmetry characteristics, and energetics of soft modes. Blinc and Zĕkš (1974) provide a review of theoretical and experimental techniques in the study of soft modes. Articles in this encyclopedia on phonons and lattice vibrations may also be of interest in this regard.

1.1.8 Antiferroelectric and More Complex Order. While individual dipoles orient in the same direction in a ferroelectric (at least within a single domain), this is by no means the only possible arrangement of dipoles. The first soft-phonon mode to freeze into a static displacement may have a periodicity identical to the original lattice unit cell or a simple multiple of the unit cell, or may even be incommensurate with the original lattice. In an antiferroelectric material, for example, there are two sets of dipole moments, each in a regular arrangement and alternating in direction. In this case there is no net polarization as one set of dipoles precisely cancels the other. $PbZrO_3$, the first known antiferroelec-

Table 1. Typical ferroelectric (F) and antiferroelectric (A) materials.

Family name	Formula		T_C (K)	P (C/m^2)	at T (K)
TGS (triglycine sulfate)	$(C_2H_5NO_2)_3 \cdot H_2SO_4$	F	322	0.03	293
KDP (potassium dihydrogen phosphate)	KH_2PO_4	F	123	0.05	96
ADP (ammonium dihydrogen phosphate)	$NH_4 \cdot H_2PO_4$	A	148	—	
Perovskite	$NaNbO_3$	A	627	—	
Perovskite	$PbTiO_3$	F	763	>0.50	300
Rochelle salt	$NaKC_4H_4O_6 \cdot 4D_2O$	F	297[a] 255[b]	0.003	278
PVDF (polyvinylidene fluoride)	$(CH_2CF_2)_n$	F	≈370	0.1	300

---

[a] Upper T_C.
[b] Lower T_C.

tric, was identified by Kittel (1951), who also coined the term.

$Gd_2(MoO_4)_3$ has been found to be a ferroelectric, but the transition to the ferroelectric state is not associated with any observable dielectric anomaly and the substance is called an "improper" ferroelectric since the spontaneous polarization does not alter the symmetry of the crystal. Other and increasingly subtle types of dipole ordering are now recognized, in both homogeneous and inhomogeneous materials, and have been reviewed by Fousek (1991).

1.2 Magnetism

Magnetite (Fe_3O_4) may have been used as a compass by the Chinese as early as 2500 B.C., and magnetic materials continue to play an important role today as inductor cores, magnetic recording media, and permanent magnets for use in motors, generators, loudspeakers, and much more. Insulating magnetic materials are preferred to electrically conducting magnetic materials for high-frequency applications because the absence of induced eddy currents, which reduces power losses, makes them especially important for the manufacture of inductors.

The magnetic behavior of materials is no less varied than the electrical behavior described in the preceding sections. A more complete discussion is provided in the articles on magnetism and magnetic devices, but in broad terms the magnetic field H, the magnetic induction B, magnetic permeability μ, and magnetic susceptibility χ_m parallel the roles of the electric field E, the displacement field D, the permittivity ϵ, and the electric susceptibility χ already described. Caution must be exercised in pressing the analogies too far, however, because the physical origin of local magnetic-dipole moments is quite distinct from the origin of local electric-dipole moments, as will be described in the following sections.

1.2.1 Diamagnetism. Insulators composed of ions with all electronic shells filled are typically diamagnetic, and the total diamagnetic susceptibility may be reliably estimated from the contributions of the individual ions. The effect of an applied magnetic field, in this case, has been described as inducing a microscopic current in the orbital electrons about each atom. According to Lenz's law, the induced currents produce a weak magnetic field tending to cancel the applied magnetic field. A complete description of this behavior requires a quantum-mechanical treatment, but the effect is called diamagnetism. Under certain conditions superconducting materials exhibit "perfect diamagnetism," where the induced currents completely cancel the applied field and prevent the penetration of magnetic field to the interior of the superconductor. Diamagnetism arising from the orbital motion of electrons within an atom is about a million times weaker than perfect diamagnetism and is temperature independent. Although typical of most insulators, this weak diamagnetism is of little technological interest.

1.2.2 Paramagnetism. Paramagnetism refers to an induced magnetic field in the same direction as the applied field, which is to say, a positive magnetic susceptibility. Paramagnetism occurs in insulators composed of ions with partially filled $3d$ (Ni, Fe, Mn, etc.) or $4f$ (Sm, Gd, Dy, etc.) shells, representing first-row transition-metal ions and the lanthanides, respectively. The electrons occupying an incompletely filled shell experience electron-electron Coulomb interactions as well as an interaction between the electron spin and the orbital motion. These two interactions tend to keep the electron spins aligned, with the result that the ion has a permanent magnetic-dipole moment. The magnetic response of a solid composed of such ions will depend on temperature following Curie's law, analogous to Eq. (18), where the decrease in the susceptibility at high temperatures reflects the inability of an applied field to overcome the increased thermal agitation of the local moments.

1.2.3 Ferromagnetism. When in close proximity, the magnetic field produced by one localized magnetic moment may directly influence the orientation of the magnetic moment of a neighboring ion. In this case, the localized moments no longer behave entirely independently, and large numbers of moments may spontaneously orient in the same direction. This behavior is called ferromagnetism [Fig. 9(a)], and the temperature at which ferromagnetic order sets in will depend on the strength of the interaction between

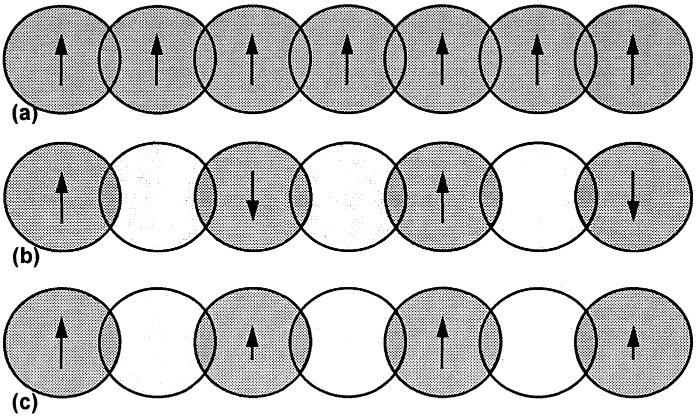

FIG. 9. (a) Direct exchange, local moments directly overlap leading to ferromagnetic order; (b) superexchange, local moments interact via an intermediate nonmagnetic ion leading to antiferromagnetic order; and (c) superexchange leading to ferrimagnetic order.

local moments. Above the ordering temperature, such materials exhibit paramagnetic behavior.

1.2.4 Antiferromagnetism and More Complex Order. In many crystal structures, local magnetic moments are separated by nonmagnetic ions, as illustrated in Fig. 9(b). Since the direct interaction between local moments decreases rapidly with increasing distance, the nonmagnetic ion can mediate a magnetic interaction that is stronger than the direct interaction. This type of interaction is called superexchange and is typically negative in sign, leading to the alternating spin-up, spin-down arrangement known as antiferromagnetism, as suggested in Fig. 9(b). An antiferromagnet exhibits no net magnetization, and the magnetic susceptibility, which peaks in a cusp at the antiferromagnetic ordering temperature, need not reach extremely large values. Complex crystal structures with more than one type of magnetic ion may exhibit a wide variety of magnetic ordering behavior with the magnetic moment varying in magnitude, as suggested by Fig. 9(c), or direction from one crystallographic site to another.

1.3 Thermal Properties

The heat capacity of insulators is remarkably well behaved, and even relatively simple models based on the statistical mechanics of phonons provides a quite accurate description. At high temperatures, each possible vibrational mode of an insulator is equally excited and the heat capacity approaches the temperature-independent Dulong-Petit value of $3k$ for each atom in the crystal. At low temperatures, the lattice vibrations decrease rapidly and the heat capacity varies as T^3 with relatively minor and well-understood deviations, except for materials with special considerations such as phase transitions.

Metals are typically good thermal conductors because of the ability of free electrons to conduct heat as well as electricity. The thermal conductivity of insulators is somewhat more complex, as discussed in more detail in Thermal Conduction in Solids, and can be either large or small depending on the material. Essentially, the heat carried by phonons in an insulator can be estimated from a knowledge of how fast and how far a phonon moves between scattering events. The thermal conductivity may be expressed as

$$\lambda=\tfrac{1}{3}Cv_sl, \qquad (19)$$

where C is the heat capacity per unit volume, v_s is the speed of sound, and l is the phonon mean free path. The speed of sound typically reflects the bond strength, and high–melting-point materials tend to have a high sound velocity and, therefore, high thermal conductivity. If there is only one atom per unit cell, all the phonon modes are acoustic modes with relatively high propagation velocities. As the crystal structure becomes more complex, the unit cell grows, and more phonon modes become optical in character with very low propagation velocity. Thus, the more complex the crystal structure, the lower the thermal conductivity is expected to be. Amorphous materials, such as fused silica, may be considered the limiting case in regard to structural complexity, and indeed these materials have among the lowest thermal conductivity values known.

In materials relatively free of defects, phonons may propagate a long distance before scattering, and in this case the long mean free path leads to a high thermal conductivity. Classic examples are diamond and aluminum nitride, which have thermal conductivity values greater than metals. Even if no other defects are present, the phonons themselves may be thought of as deviations from the ideal crystal structure and phonon-phonon scattering is often the limiting factor. At high temperatures, phonon-phonon scattering leads to a $1/T$ temperature dependence. At low temperatures, as phonon-phonon scattering decreases, the phonon mean free path becomes larger until the walls of the specimen prevent further propagation. Eventually, the thermal conductivity decreases as T^3 with decreasing temperature, reflecting the temperature dependence of the heat capacity discussed above.

2. INSULATING MATERIALS

Electrically insulating materials are far too common to enumerate in any complete way, but some appreciation of the trends and patterns may be achieved by a brief survey. Some of the more important materials will be discussed, as well as the connection between the structure of the materials and their properties.

2.1 Elemental Insulators

The rare gases, hydrogen, nitrogen, oxygen, and fluorine and chlorine are insulating gases under standard conditions and remain insulators even when they condense to liquids and finally solids at sufficiently low temperatures. Bromine is an insulating liquid and iodine and astatine are insulating solids under standard conditions. The common, stable form of selenium has infinite helical chains and is, like tellurium, semiconducting, but Se_8 polymorphs are red insulators similar to solid S_8, which is a very poor electrical conductor. Arsenic is a semimetal, and the electrical conductivity of phosphorus varies greatly among the four known allotropes, but the volatility and reactivity of these elements severely limit their use as solid insulators. Boron, the only refractory nonmetallic element other than carbon, occurs in two distinct crystal structures (the α and β forms), both of which are semiconductors with band gaps around 1.5 to 2.0 eV.

Carbon, in the diamond crystal structure shown in Fig. 10, is expected to become increasingly important since the development of diamondlike thin films. The diamond crystal structure may be described as two identical, interpenetrating face-centered-cubic (fcc) lattices. The second fcc lattice is displaced from the first by a translation along a diagonal of the cube by a distance of $\frac{1}{4}$ of the length of the diagonal. With its tightly bound electrons and strong covalent bonds, diamond is uniquely suited for many applications. The electrical properties of diamond are dominated by the very large band gap of 5.4 eV, which leads to electrical conductivity values as low as 10^{-13} S m^{-1}, although doping can increase this value to 10^{0} S m^{-1} or more. The strong bonds and small mass of carbon lead to sound velocity and thermal conductivity values among the highest known. For more information the reader is referred to DIAMOND AND DIAMONDLIKE CARBON.

Other forms of carbon are not electrical insulators. Graphite, for example, consists of two-dimensional sheets of hexagonally bonded carbon with only weak bonding between the sheets. Graphite is typically a fairly good electrical conductor, with conductivity values on the order of 10^5 S m^{-1}. Thus, by changing the crystal structure from the diamond arrangement to the graphite arrangement the

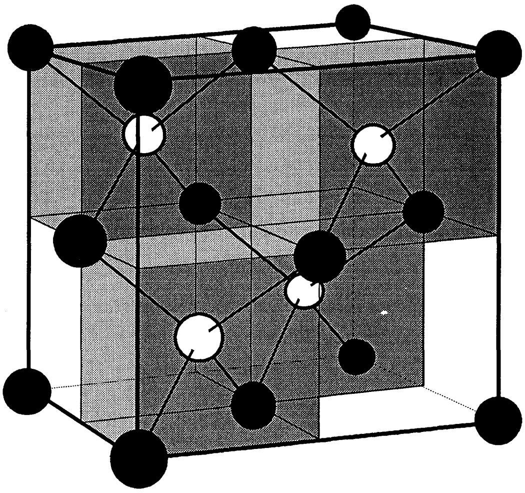

FIG. 10. The sphalerite (or zinc-blende) crystal structure. Also, the diamond crystal structure when the solid and open circles represent the same atom.

Table 2. Selected covalent insulators.

Structure type	Compound
Diamond	C
Sphalerite	BAs, BP, BeTe, ZnSe, BeSe, ZnS, CuBr, BN, BeS, ZnO, MgO
Wurtzite	BP, AgI, SiC, ZnS, BN, GaN, AlN, ZnO, BeO
Hexagonal	α-Si_3N_4, β-Si_3N_4, SiO_2

conductivity can increase by 18 orders of magnitude, which emphasizes the role of structure in determining the electrical properties of solids.

All of the remaining elements are semiconductors, semimetals, or true metals. Thus, among the elements, solid electrical insulators are relatively rare.

2.2 Binary Compound Insulators

2.2.1 Covalent Insulators. Many covalently bonded compounds are semiconductors, but some important insulators found in this group are listed in Table 2. The first group includes those compounds that share the same basic structure as diamond, except that the two interpenetrating sublattices shown in Fig. 10 are occupied by different atoms. This is the sphalerite (or zinc-blende) structure type, and Table 2 lists a number of common insulators with this structure. The sphalerite structure is symmetric about each atom in the unit cell and therefore is neither pyroelectric nor piezoelectric.

Other covalent insulators form in the wurtzite structure, which can be considered as two interpenetrating close-packed hexagonal lattices. As in the sphalerite structures, each atom is tetrahedrally coordinated with four atoms of the opposite type. BN can be found in the zinc-blende or wurtzite forms, as well as in a graphitelike hexagonal structure, all of which are good insulators. SiC forms in a large number of closely related structures composed of sphalerite and wurtzite layers, and although it has interesting semiconducting properties, all SiC polytypes may be considered insulators if sufficiently pure. The ease with which the covalent bond is polarized generally increases with atomic number, which can lead to somewhat larger static dielectric constants than expected for a nonpolarizable bond, but qualitatively the permittivity behaves very much as expected on the basis of the ionic polarization mechanism described in Figs. 2, 3, and 4.

2.2.2 Ionic Insulators. The ionic crystal structures for sodium chloride (NaCl) and cesium chloride (CsCl) are shown in Fig. 11. Under ordinary conditions, all of the alkali halides form in the NaCl-type structure, except CsCl, CsBr, and CsI, which form in the CsCl-type structure. Both are high-symmetry structures, exhibit no piezoelectric or pyroelectric effects, and have dielectric behavior in good accord with the ionic polarization mechanism described above. In Table 3 note the distinctly greater polarizability of the CsCl-type compounds compared to the trends among the NaCl-type as indicated by the relatively larger static dielectric constants ϵ_0 of the former compounds. This indicates that even closely related crystal structures with similar bonding can exhibit distinct properties.

2.2.3 Binary Oxides and Other Binary Insulators. Although both insulators and conductors can be found in almost any family of materials, many binary chalcogenides (especially oxides and sulfides) and pnictides

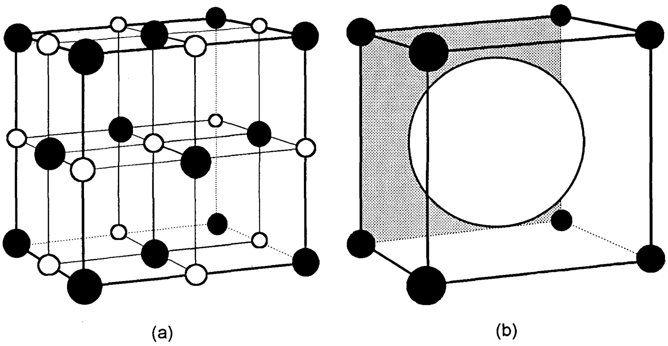

FIG. 11. (a) Sodium chloride and (b) CsCl crystal structure types.

(including nitrides and phosphides) and practically all halides are typically good electrical insulators. These compounds are neither purely covalent nor purely ionic, and while many are insulators, they tend to have relatively high symmetry crystal structures and fairly ordinary dielectric properties.

Most of the binary oxides are insulators, and materials such as alumina (Al_2O_3), magnesia (MgO), and beryllia (BeO) see widespread use for their dielectric properties. $3d$ transition-metal oxides such as Fe_3O_4 and CrO_2 are typically magnetic and exhibit the characteristic phenomena of saturation and hysteresis so important for magnetic recording applications. Many other $3d$ oxides such as MnO, FeO, and CrO are antiferromagnets, and in most cases the superexchange mechanism is responsible for the magnetic order.

Practically all the binary carbides and most of the borides not listed above, however, have at most relatively small band-gap energies and are semiconductors, and many, in fact, are metallic compounds. Only a handful of binary compounds (such as FeS, Bi_2S_3, WO_3 and possibly a few others) are ferroelectrics. Indeed, the scarcity of ferroelectric behavior among binary compounds left the impression for many years that ferroelectricity was rather rare, an impression later dispelled by studies of more complex materials. Finally, SiO_2 in its various forms deserves special attention and is discussed in dedicated articles on GLASSES and QUARTZ.

Table 3. Alkali halide ionic insulators in the NaCl structure type, except where indicated (after Knox and Teegarden, 1968).

Compound	ϵ_0	ϵ_∞	$\hbar\omega_T/k$ (K)
LiF	9.01	1.96	442
NaF	5.05	1.74	354
KF	5.46	1.85	274
RbF	6.48	1.96	224
CsF	—	2.16	125
LiCl	11.95	2.78	276
NaCl	5.90	2.34	245
KCl	4.84	2.19	215
RbCl	4.92	2.19	183
CsCl[a]	7.2	2.42	151
LiBr	13.25	3.17	229
NaBr	6.28	2.59	195
KBr	4.90	2.34	166
RbBr	4.86	2.34	139
CsBr[a]	6.67	2.42	114
LiI	16.85	3.80	—
NaI	7.28	2.93	167
KI	5.10	2.62	156
RbI	4.91	2.59	118
CsI[a]	6.59	2.62	95

[a] CsCl structure type.

2.3 Multinary Compound Insulators

As increasingly complex materials have come under investigation, new types of behavior have been observed and characterized. Often, these new behaviors have no precedent among the binary compounds or among the elements. As far as ordinary dielectric properties are concerned, polystyrene, polyethylene, Mylar, Teflon, and similar materials are not qualitatively superior to or even especially distinct from the classic binary compounds discussed above. Still, modern plastics and polymers are often vastly preferred in dielectric applications for cost, packaging, or other practical considerations.

Many of the more complex materials exhibit effects completely unobserved in the binary compounds. Table 4, for example, provides an overview of the ferroelectric materials now known to exist. Typically, as each class of materials has been developed, new theoretical concepts have been required as well.

2.3.1 Rochelle Salts and Other Order/Disorder Systems. The earliest known ferroelectrics were the Rochelle salts (also called Seignette salts), such as $NaKC_4H_4O_6 \cdot 4H_2O$ (Valasek, 1920, 1921). The orientation polarization mechanism is vividly demonstrated in these materials in that the hydrogen bonding is sufficiently weak to allow a type of hindered rotation of the water molecules. The orientation of one water molecule influences the preferred orientation of others nearby, and as the temperature is lowered through the Curie temperature, the water molecules tend to order with respect to each other, establishing a net polarization.

An interesting feature of the Rochelle salts is the presence of two critical temperatures, as indicated in Table 1. Above the upper transition temperature and below the lower temperature, two different sublattices of water mol-

Table 4. Overview of ferroelectric materials (modified from Fousek, 1991).

I. Inorganic compounds

	Nonhydrated	
Perovskites, $BaTiO_3$	K_2SO_4 type, K_2SeO_4	Barium fluoride type, $BaMnF_4$
$LiNbO_3$ type, $LiTaO_3$	Langbeinites, $(NH_4)_2Cd_2(SO_4)_3$	Halides, HCl
Manganites, $YMnO_3$	Sulfates, $(NH_4)_2SO_4$	Antimony sulfide iodine, SbSBr
Tungsten bronze type, $PbNb_2O_6$	Selenites, $LiH_3(SeO_3)_2$	Oxyfluorides, $Rb_3MoO_3F_3$
Pyrochlore type, Cd_2Nb	Nitrites, $NaNO_2$	Boracites, $Mg_3B_7O_{13}Cl$
Layer structure oxides, $Bi_4Ti_3O_{12}$	Nitrates, KNO_3	Iodates, NH_4IO_3
Molybdates, $Gd_2(MoO_4)_{32}O_7$	Thiophosphates, $Sn_2P_2S_6$	Oxyiodates, Sb_5O_7I
Vanadates, $Ca_3(VO_4)_2$	Potassium dihydrogen phosphate (KDP) type, KH_2PO_4	
Other double oxides, $SbNbO_4$		
	Hydrated	
Colemanite, $Ca_2B_6O_{11}\cdot 5H_2O$	Potassium cyanides, $K_4Mn(CN)_6\cdot 3H_2O$	Alums, $NH_4Fe(SO_4)_2\cdot 12H_2O$
Sulfates, $NaNH_4SO_4\cdot 2H_2O$	Uranospinite, $H_2(VO_2)_2(AsO_4)_2\cdot 8H_2O$	

II. Organic-inorganic compounds

Nonhydrated	
Triglycine sulfate (TGS), $(CH_2NH_2COOH)_3\cdot H_2SO_4$	Sarcosine salt, $(CH_3NHCH_2COOH)_3CaCl_2$
Betaines, $(CH_3)_3NCH_2COO\cdot H_3AsO_4$	Tetramethyl ammonium salt, $N(CH_3)_4\cdot HgCl_3$
Chloracetates, $(CH_2ClCOO)_2HNH_4$	Propionate, $(CH_3CH_2COO)_6Ca_2Sr$
Alkyl ammonium halogenantimonates, $(CH_3NH_3)_3Sb_2Br_9$	
Hydrated	
Rochelle salts (tartrates), $NaKC_4H_4O_64H_2O$	Formates, $Cu(HCOO)_2\cdot 4H_2O$
Triglycine sulfate (TGS), $(CH_2NH_2COOH)_2\cdot MnCl_2\cdot 2H_2O$	Alums, $CH_3NH_3Al(SO_4)_2\cdot 12H_2O$
Betaines, $(CH_3)_3NCH_2COO\cdot CaCl_2\cdot 2H_2O$	Guanidinium salts, $C(NH_2)_3Al(SO_4)_2\cdot 6H_2O$

III. Organics

Thioruea, $SC(NH_2)_2$	Tanane, $C_9H_{18}NO$	Benzil, $C_{14}H_{10}O_2$	Phenanthrene, $C_{14}H_{10}$
PVDF type, $(CH_2CF_2)_n$ and copolymers with $CHF{=}CF_2$, $CF_2{=}CF_2$			
Liquid crystals, ≈ 50 distinct homologues known			
Langmuir-Blodgett films: ω-tricosenoic acid $(CH_2{=}CHC_{21}H_{42}COOH)$ + docosylamine $(C_{22}H_{45}NH_2)$			

ecules are equally polarized, but in opposite directions resulting in no net polarization. Between the two transition temperatures the two sublattices are still polarized in opposite directions, but the polarizations do not exactly cancel. Fairly detailed models, which may be described as a competition between two distinct relaxation times rather than the single relaxation time assumed in Fig. 3, have been developed to describe the Rochelle salts and their heavy-water analogs (Blinc and Žekš, 1974).

Materials systems with related order-disorder transitions generally involve hydrogen bonding. Examples include the heavily studied potassium dihydrogen phosphate (KH_2PO_4 or KDP) type system, sulfates, selenites, nitrites, triglycine sulfate (TGS), formates, colemanite, alums, and more. These systems have been rather extensively studied and are reasonably well described by relaxation-time, order/disorder models.

2.3.2 Perovskites and Displacive Transitions. Figure 12 shows the typical perovskite crystal structure in the high-temperature, cubic phase. The oxygen atoms in the centers of the cube faces have two titanium nearest neighbors (atoms) along an axis perpendicular to the cube face, but there are no titanium atoms in the cube faces at all. Although the overall lattice is cubic, the oxygen atoms do not experience full cubic symmetry

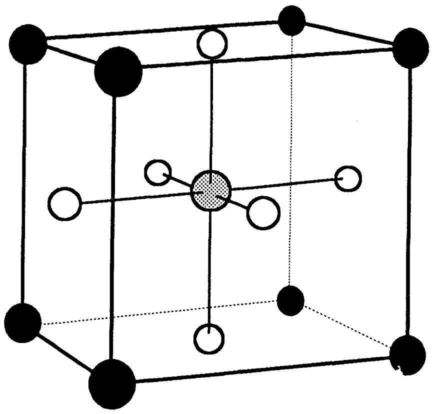

FIG. 12. The perovskite crystal structure for $BaTiO_3$. Ba atoms are at the cube corners, a Ti atom is at the cube center, and O atoms are at the centers of the cube faces. Note that each oxygen atom has only two titanium nearest neighbors, which is the origin of the ferroelectric behavior instability.

and can displace within the plane much more easily than in a direction perpendicular to the plane.

It is this asymmetric movement of the oxygen atoms that leads to an instability in the lattice, a rather clear example of a soft-mode instability. And indeed, upon cooling the unit cell distorts to a tetragonal structure and a net polarization develops along the direction of a cube edge, as shown in Fig. 13. On further cooling, the unit cell distorts again, this time along a diagonal in the face of the cube, and finally at still lower temperatures a rhombohedral distortion sets in, with a net polarization along a full cube diagonal.

The perovskites are called displacive ferroelectrics, since the mechanism involves a permanent displacement of atoms. Above the Curie temperature the resonant frequency (from Fig. 3) goes to zero as

$$\omega_i^2 \propto T - T_C , \tag{20}$$

a typical soft-mode signature.

2.3.3 Bronzes, Pyrochlores, and Nonmagnetic Oxides.

The metal oxides (niobates, manganites, tungsten bronzes, pyrochlores, molybdates, vanadates, and some others) listed in Table 4 undergo displacive transitions upon cooling and exhibit resonant behavior in the dielectric properties characteristic of an ionic soft mode, as with the perovskites. The theory of these ferroelectric materials can be said to be reasonably mature. Important for some time as commercial ceramics (see Table 5), many of these oxides continue to be the subject of research in the form of thin films, composites, and layered structures (see MATERIAL PROPERTIES OF CERAMICS).

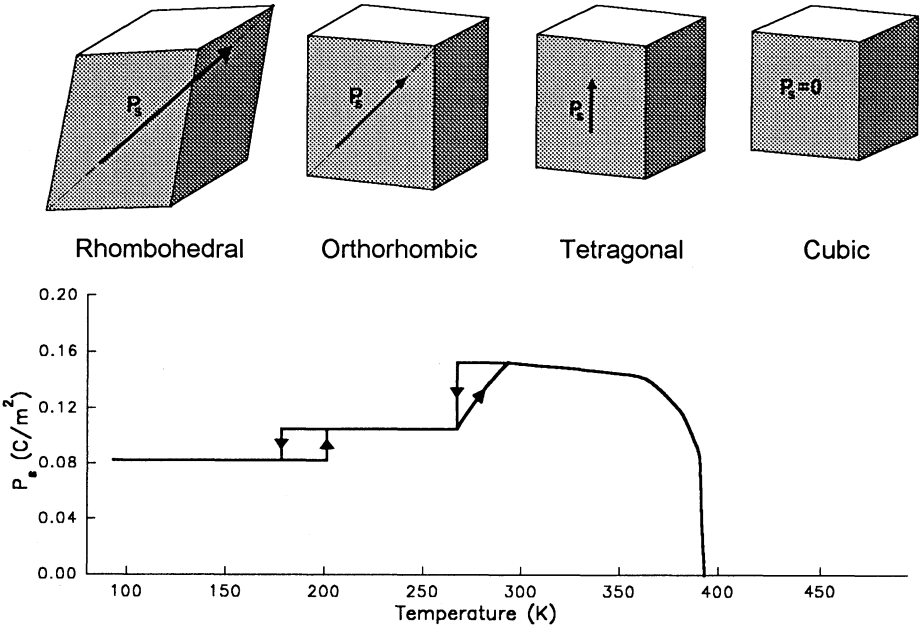

FIG. 13. Projection of the spontaneous polarization of $BaTiO_3$ onto one edge of the unit cell (modified from Merz, 1949).

2.3.4 Ferrites: Spinels and Garnets. The term ferrite is typically used to refer to ferrimagnetic oxides. These materials are hard, brittle, and poor electrical conductors widely used in inductors, particularly at high frequencies. An important family of ferrites are the spinels with the chemical formula MFe_2O_4, where the iron is trivalent and M denotes one of the divalent ions Mg^{2+}, Mn^{2+}, Fe^{2+}, Co^{2+}, Ni^{2+}, Cu^{2+}, Zn^{2+}, or Cd^{2+}. Spinels have a cubic crystal structure containing 8 chemical formula units with 32 oxygen sites per unit cell. The metal ions occupy one of two types of sites within the unit cell, designated A and B, and are coordinated by either four or six nearest-neighbor oxygen ions, respectively. Thus, there are two types of magnetic moments, corresponding to the Fe^{3+} ions or the divalent metal ions, and ferrimagnetic behavior due to the superexchange mechanism dominates the magnetic properties. There are two structural variants, called normal and inverse spinels, depending on whether the divalent ions occupy A sites or B sites, respectively, and real materials are typically mixtures of both structural variants.

The garnets, with chemical formula $3M_2O_3 \cdot 2Fe_2O_3 \cdot 3Fe_2O_3$, have 160 atoms per unit cell and three distinct types of crystallographic sites occupied by magnetic ions. In these materials M represents one of the trivalent rare-earth elements; perhaps the best known example is yttrium iron garnet, abbreviated YIG, which is notable for particularly low magnetic losses at high frequencies.

Table 5. Oxides often used in insulating ceramics (modified from von Hippel, 1954).

Kaolinite	$Al_2(OH)_4Si_2O_5$
Quartz	SiO_2
Feldspar	$K_xNa_{1-x}[Al \cdot Si_3]O_8$
Mullite	$Al_6Si_2O_{13}$
Cordierite	$Mg_2Al_2Si_5O_{15}$
Alumina (corundum)	Al_2O_3
Steatite (talc)	$Mg_3 \cdot (OH)_2 \cdot Si_4O_{10}$
Clinoenstatite	$MgSiO_3$
Forsterite	Mg_2SiO_4
Zircon	$ZrSiO_4$
Spodumene	$Li_2Al_2Si_4O_{12}$
Wollastonite	$CaSiO_3$
Rutile	TiO_2
Barium titanate	$BaTiO_3$
Sodium tantalate	$NaTaO_3$
Sodium niobate	$NaNbO_3$
Cadmium niobate	$Cd_2Nb_2O_7$
Magnetite	Fe_3O_4
Spinel	$MgAl_2O_4$
Ferrite spinel	$Ni_xZn_{1-x}Al_2O_4$

2.3.5 Organics. Plastics, polymers, and organic materials of many types have become increasingly important for a variety of applications, from packaging to textiles (see Polymers, Material Properties of). Organic chemistry techniques applied to ferroelectric materials have led to the development of novel structures with unique properties. A few examples follow.

Polyvinylidene fluoride (PVDF) is an organic polymer that can be prepared from solvents in the form of thin films. As prepared, PVDF crystallizes in a nonpolar α phase, which upon stretching transforms to a ferroelectric β phase. Mixtures of PVDF and trifluoroethylene (and other copolymers) will crystallize directly in the ferroelectric phase without further mechanical treatment, and the Curie temperature can be adjusted in the range of 330 K to 400 K. PVDF and related materials may be inexpensively prepared in thin layers ($< 10\ \mu m$) and have therefore received considerable attention for various piezoelectric and pyroelectric applications, particularly those requiring rapid response times (Gerhard-Multhaupt *et al.*, 1987).

Liquid crystals represent a second family of organic insulators with particularly interesting ferroelectric properties. Liquid-crystal ferroelectrics typically consist of relatively long, rodlike molecules which "crystallize" in regular layers, as suggested in Fig. 14. These layered phases, called smectic phases, come in a variety of types enumerated alphabetically, as in SmA for the smectic A-type arrangement and so on. The arrangement depicted in Fig. 14 is the SmC arrangement, indicating that the long axes of the molecules are not quite in the direction normal to the planes, but are tilted somewhat. Within the layers, however, the molecules have no long-range order and behave essentially as a liquid, but the molecules remain essentially parallel and from plane to plane behave much like a crystal.

If the molecules in a smectic-C type phase are also chiral, meaning that the molecule has distinct right-handed and left-handed versions, then the liquid crystal will exhibit ferroelectric properties. Liquid-crystal ferroelectrics have become increasingly important for their potential in low-cost flat-panel displays,

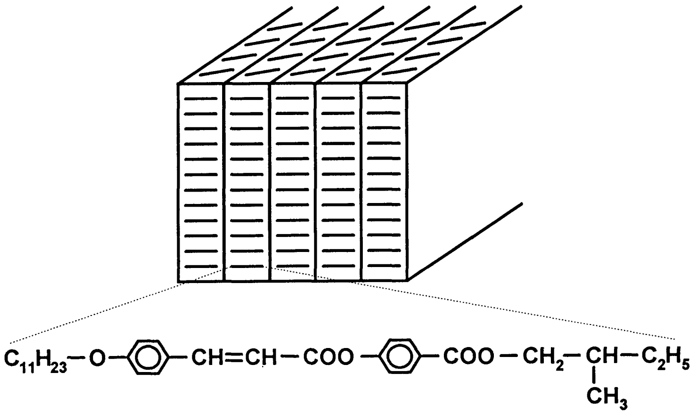

FIG. 14. Typical ferroelectric liquid crystal molecule, shown in the smectic-C arrangement.

and most of the new ferroelectric materials identified recently are liquid crystals (Fousek, 1991). For a more detailed discussion of liquid crystals, see LIQUID CRYSTALS, STRUCTURE OF.

Finally, the Langmuir-Blodgett technique is another powerful method for preparing organic thin-film structures with well-defined symmetry (Roberts, 1985). Typically, molecules with a hydrophilic head and a hydrophobic tail are dissolved in an organic solvent and spread on a water surface, and the solvent is evaporated. Monomolecular layers of the film-forming material can then be collected on a solid substrate by repeated passes through the water surface. A pyroelectric material can be produced by depositing alternating layers of fatty acids and amines, as shown in Fig. 15.

3. SUMMARY

The list of insulating materials has not been exhausted here. Rubbers, elastomers, plastics, polyvinyl chloride (PVC), gases, liquids, most minerals, glasses, and more have been nearly excluded. But the basic polarization mechanisms described (electronic, ionic, orientational, and space charges) provide a sound foundation for understanding the major insulating properties of a wide variety of materials. The principles of crystal structure and symmetry in allowing or forbidding various kinds of electric dipole ordering will also play a key role, regardless of the material or application in question.

The theoretical understanding of the classical insulating materials is quite mature and provides a firm foundation for further developments. The advent of liquid-crystal ferroelectrics and the ability to design and manufacture new structures nearly at will is sure to have an impact on future sensor, imaging, memory, and video technologies.

GLOSSARY

Critical Temperature: Typically a cusp or other sharp feature in the temperature dependence of an observable property, representing an abrupt change in the order among physical objects such as ions, electric dipoles, or magnetic dipoles.

Dielectric: A material that internally redistributes charges in response to an applied electric field.

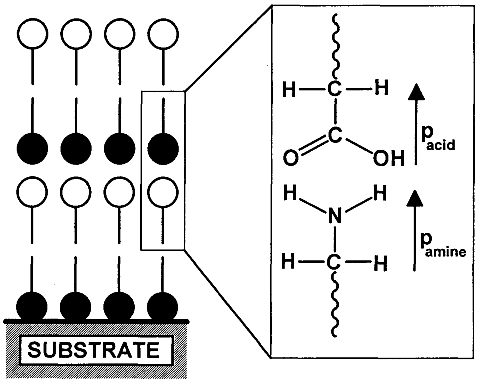

FIG. 15. Langmuir-Blodgett film of alternate acid/amine layers to produce a pyroelectric (redrawn from Roberts, 1985).

Phonon: A single lattice vibration mode representing physical displacement of atoms or ions about their equilibrium positions.

Polarization: The electric-dipole moment induced in a dielectric by an applied electric field.

Soft Mode: An unstable lattice vibration requiring little or no energy to displace atoms or ions from their equilibrium positions.

List of Works Cited

Anderson, P. W. (1959), in: G. I. Skanavi (Ed.) *Fizika Dielektrikov,* Moscow: Akad. Nauk SSSR, p. 290.

Blinc, R., Žekš, B. (1974), *Soft Modes in Ferroelectrics and Antiferroelectrics,* New York: North-Holland.

Curie, J., Curie, P. (1880), *C. R. Acad. Sci.* **91**, 294.

Cochran, W. (1959), *Phys. Rev. Lett.* **3**, 412–414.

Fousek, J. (1991), *Ferroelectrics* **113**, 3–20.

Gerhard-Multhaupt, R., Gross, B., Sessler, G. M. (1987), in: G. M. Sessler (Ed.), *Electrets,* Berlin: Springer, pp. 383–420.

Kittel, C., (1951), *Phys. Rev.* **82**, 729–732.

Knox, R. S., Teegarden, K. J. (1968), in: W. B. Fowler (Ed.), *Physics of Color Centers,* New York: Academic, p. 625–626.

Merz, W. J. (1949), *Phys. Rev.* **76**, 1221–1225.

Roberts, G. G. (1985), *Adv. Phys.* **34**, 475–512.

von Hippel, A. (1954), *Dielectric Materials and Applications,* New York: Wiley.

Valasek, V. (1920), *Phys. Rev.* **15**, 537–538.

Valasek, V. (1921), *Phys. Rev.* **17**, 475–481.

Further Reading

Blinc, R., Žekš, B. (1974), *Soft Modes in Ferroelectrics and Antiferroelectrics,* New York: North-Holland. A thorough discussion of lattice dynamics and soft modes in insulators.

Flügge, S. (1956), *Encyclopedia of Physics,* Vol. 16, *Dielectrics,* Berlin: Springer-Verlag. Classic presentation including detailed case studies of some of the most important ferroelectric systems.

Robert, P. (1988), *Electrical and Magnetic Properties of Materials,* Norwood, MA: Artech House. An introduction to the properties of materials with an emphasis on engineering applications.

von Hippel, A. (1954), *Dielectric Materials and Applications,* New York: Wiley. Somewhat dated, but an outstanding discussion of the basic concepts important for dielectric applications.

Wells, A. F. (1987), *Structural Inorganic Chemistry,* Oxford: Clarendon Press. A comprehensive discussion of the crystallographic aspects of inorganic materials, most of which are electrical insulators.

INTEGRATED CIRCUITS

See VLSI—INTEGRATED CIRCUITS

INTEGRATED FERROELECTRICS

See THIN-FILM INTEGRATED FERROELECTRICS

INTENSE PARTICLE BEAMS

JAMES BENFORD, *Physics International Company, San Leandro, California, U.S.A.*

Introduction 103
1. **Methods of Generation** 104
1.1 Acceleration Techniques 104
1.2 Electron Emitters 106
1.3 Accelerating Diodes 107
1.4 Ion-Beam Diodes 108
2. **Propagating Intense Beams** 109
2.1 Beams in Gas and Plasma 109
2.2 Beams in Drift Tubes 110
3. **Applications of Intense Particle Beams** 110
3.1 Inertial Fusion 111
3.2 High-Power Microwaves 112
3.3 X-Ray Production 115
3.4 Industrial Applications 116
Glossary 116
Works Cited 116
Further Reading 116

INTRODUCTION

Particle beams are characterized as intense if their self-fields are sufficiently strong to dominate their behavior. Such beams can be controlled with guiding magnetic fields, but they can also propagate freely in a variety of ways. They are characterized by the strength parameter ν/γ. The fundamental scale for intense beams is the Alfvén current

$$I_A = \frac{4\pi\epsilon_0 m_0 c^2}{e}\beta\gamma, \qquad (1)$$

where m_0 is the electron mass and the relativistic factor $\gamma^2 = (1-\beta^2)^{-1}$; $\beta = v/c$, and v is the particle velocity. In practical units

$$I_A = 17\beta\gamma \text{ kA}. \qquad (2)$$

The current of a beam can be written in terms of the Alfvén current as

$$I/I_A = \nu/\gamma, \qquad (3)$$

where ν is the Budker parameter, defined as the product of the number N of beam particles per unit length and the classical electron radius:

$$\nu = Ne^2/4\pi\epsilon_0 m_0 c^2. \qquad (4)$$

For $\nu/\gamma \ll 1$, electron trajectories are roughly sinusoidal and propagation is possible. Such a propagating beam has a characteristic "betatron wavelength" that is the characteristic oscillation length of the electron about the axis as it propagates. The betatron wavelength for a beam of radius a is

$$\lambda_\beta = a\gamma/\nu. \qquad (5)$$

As beam intensity is increased, a point is reached where the Larmor radius of beam electrons in the magnetic field of the current is about equal to the beam radius. Beam electrons turn completely around and flow stops. This limit occurs at $I \sim I_A$ for a uniform radial current distribution. So when $\nu/\gamma \approx 1$, self-fields strongly influence beam motion.

Intense particle beams were first developed in the 1960s when pulsed electrical systems producing tens of kiloamperes of electron current at megavolt energies succeeded in propagating beams at about the Alfvén current. Since that time, vigorous efforts have produced beams with megampere currents at several megavolts with typical pulse duration of 100 ns. Electrical beam impedance is low, $< 100\ \Omega$, and can be $< 1\ \Omega$. The large energy per pulse, as high as a megajoule, gives this technology many applications from nuclear effects simulation to inertial-confinement fusion, high-power microwave generation, and x-ray production. In this article we first discuss methods of intense-beam acceleration, then electron- and ion-beam emission, followed by beam propagation and, finally, the applications of intense beams.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

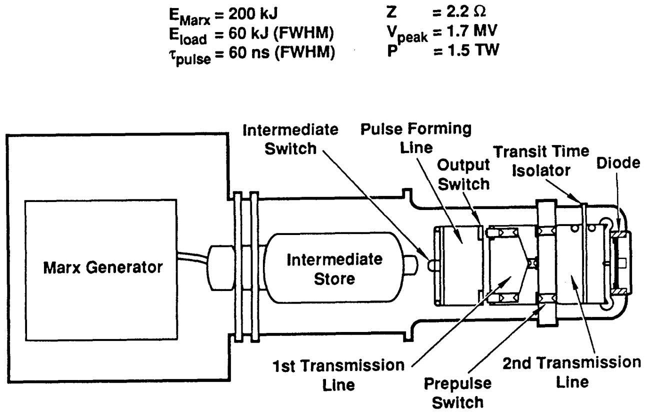

FIG. 1. Schematic of a typical pulsed-power machine used for generation of intense particle beams. The machine is the KALIF generator in Karlsruhe, Germany.

1. METHODS OF GENERATION

1.1 Acceleration Techniques

Intense beams can be produced by a variety of means, usually involving a voltage-multiplying Marx generator that pulse-charges a liquid dielectric intermediate line in a time scale of about a microsecond (Fig. 1). The Marx generator works by charging capacitors in parallel to a high voltage (~100 kV), switching them into a series arrangement, and discharging them into a load (see Miller, 1982). Firing a gas- or liquid-insulated intermediate switch then allows the voltage pulse from the intermediate line to be applied to the pulse-forming line. This line determines the output pulse duration: when the output switch fires, a wave propagates along the pulse-forming line, reflects at the end, and returns. Therefore the output pulse is twice the transit time of the line, typically 100 ns. There is a prepulse switch in some machines that fires early in the pulse, preventing any early voltage transients from reaching the output. Finally, the pulse passes through an insulator and enters a vacuum diode consisting of electrodes from which particles, usually electrons or protons, are accelerated across the potential difference and extracted as an intense beam. Intense electron beams were first generated in the 1960s, and intense ion beams appeared in the 1970s. Figure 2 shows the historical development of intense-beam generators. Machines approaching 100 TW have been built, consisting of tens of modules, each of a few terawatts each.

An important variant of the Marx generator–pulse-forming line–diode approach to beam generation is the linear induction accelerator (LIA). This technique is essentially a transformer distributed in vacuum and allows substantial downsizing of the pulsed power system. The LIA is in fact a series of one-to-one pulse transformers, each threaded by an electron beam or a solid cathode. In the latter mode it is called a voltage adder (Fig. 3). Each of the *N* transformer sections generates an

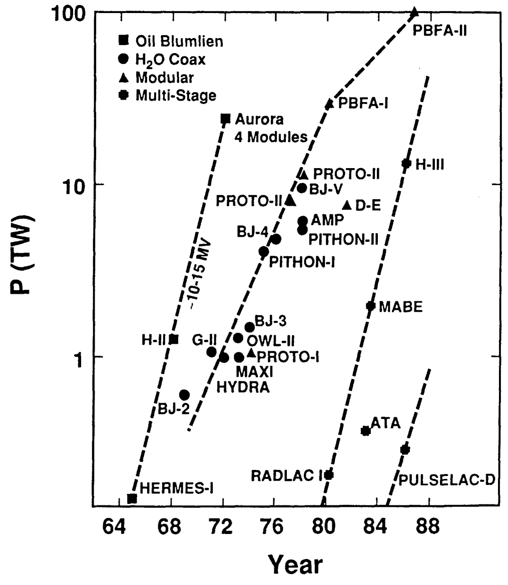

FIG. 2. Evolution of pulsed-power drivers of intense particle beams in the United States. The Blumlein is a type of pulse-forming line. Multistage devices accelerate beams incrementally; the LIA is of this type.

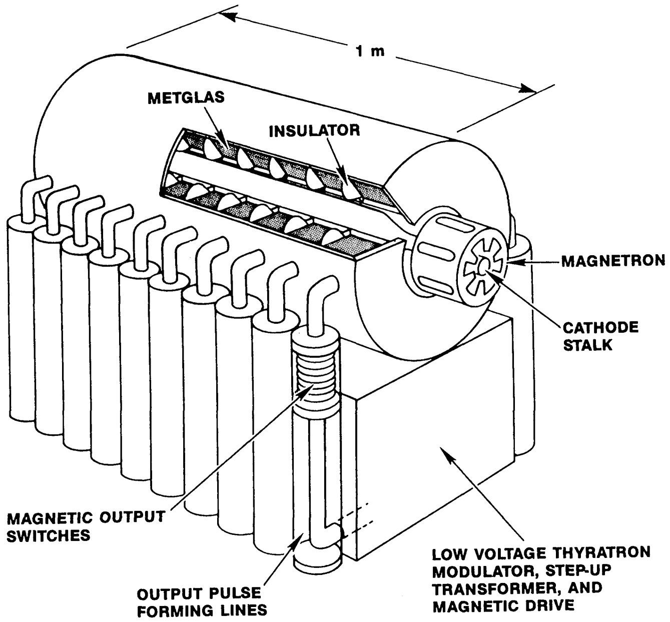
1 m
METGLAS
INSULATOR
MAGNETRON
CATHODE STALK
MAGNETIC OUTPUT SWITCHES
OUTPUT PULSE FORMING LINES
LOW VOLTAGE THYRATRON MODULATOR, STEP-UP TRANSFORMER, AND MAGNETIC DRIVE

FIG. 3. Compact version of voltage adder. Metglas is a ferrite material produced by Allied Corp. Figure courtesy Physics International Company.

increment of voltage, *V*, on the cathode or beam so that the peak voltage *NV* occurs at the tip of the cathode or on the exiting electron beam. Its principal virtue is that the required accelerating voltage is distributed among a number of pulse-forming subsystems so that the full voltage appears only on the beam or at a beam-emitting diode. LIAs are useful for short pulses, < 100 ns. Greater output voltage can be obtained by adding more cavities since the system is fundamentally modular. While diode machines are limited to voltages of a few MeV, LIAs have reached 50 MeV. For example, Hermes III is a very large voltage adder built from modules (Fig. 4). This is in contrast to the diode machines, with which much higher currents (megamperes) can be generated. The disadvantages of the LIA are

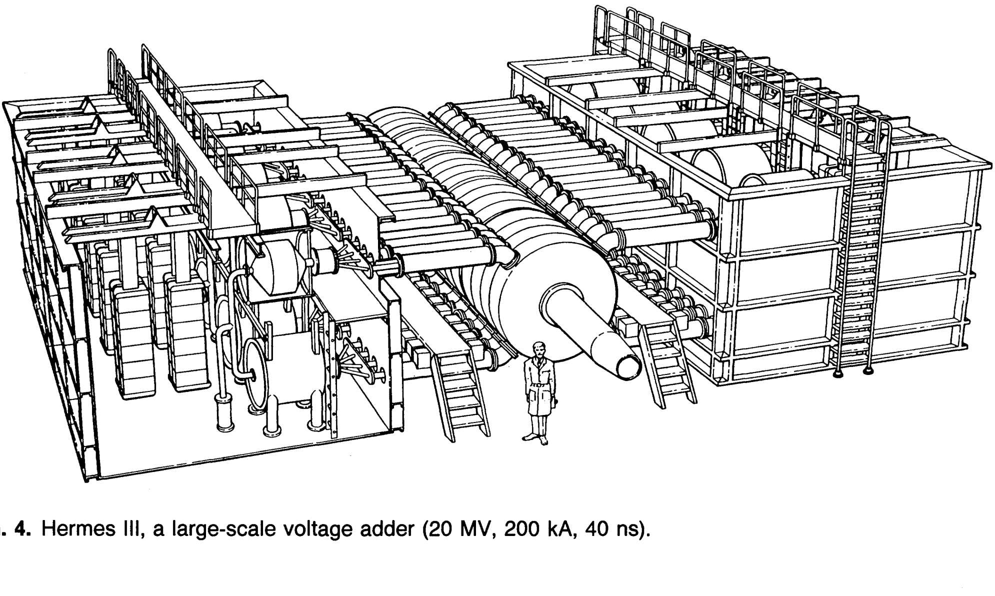

FIG. 4. Hermes III, a large-scale voltage adder (20 MV, 200 kA, 40 ns).

that it is expensive and complex compared to the diode machines.

Table 1 shows representative high-power beam-generation machines as of 1992. Most of these devices can generate intense particle beams for a wide variety of applications. For a discussion of these machines and those in Fig. 2, see Smith (1991).

1.2 Electron Emitters

The high currents needed make the electron source a vital factor in intense-particle-beam generators. We cover beam emission here; extraction and propagation are treated in Sec. 2. The most common electron sources are based on the following emitting phenomena, in order of decreasing emission current density: explosive emission, nonexplosive field emission, thermionic emission, and photoemission. In addition to the capacity for emitting at high current density, the attractiveness of each option depends on the requirements for "quality" of emission, as measured by spreads in the electron velocities and energies, ruggedness, and lifetime.

The most commonly used electron sources in present pulsed devices are based on the process of explosive emission. Applied average electric fields in excess of roughly 100 kV/cm at the emitting surface are further enhanced by the presence of naturally occurring micropoints. The enhanced fields lead to the flow of large field-emission currents through these micropoints, which rapidly heat and explode to form plasma flares within a few nanoseconds. Individual flares then expand and merge in 5–20 ns to form a uniform electron emitter capable of providing current densities of 10^7–10^8 A/m^2. The microscopic points are regenerated from one shot to the next, so that these sources can be reused. Explosive emission sources can be made from a wide variety of materials, including aluminum, graphite, stainless steel, and cloth emitters such as velvet. They produce the highest current densities of all options, are quite rugged, can tolerate poor vacuum conditions (the operationally limiting factor is arcing at $\sim 10^{-2}$ Pa), and have reasonable lifetimes. Undesirable features of explosive emission sources include the spotty character of the emission process at early times before the flares merge and the expansion of the dense plasma from which the electrons are drawn. This expansion causes the emitting surface to move with time, a phenomenon known as gap closure, which over the course of a long pulse will short out the accelerating voltage and terminate electron emission.

Field-emission sources use macroscopic points or grooves on a surface to get enhanced electric fields and produce nonexplosive field emission. Arrays of tungsten needles, grooved graphite, and polymer-metallic matrices are examples of emitters of this type; common velvet and cloth also make excellent field emitters. Current densities ranging from 10^6 to 10^7 A/cm^2 are achievable. Their advantages include operating at field levels below those required for explosive emission and avoidance of expanding plasma at the emitting surface. A disadvantage of these sources relative to explosive emission sources is the lower peak current density achievable; source lifetime is low.

Thermionic emission is based on heating an emitting surface to allow electrons to overcome the work function and escape the surface. No expanding surface plasma is required, and long-pulse to continuous electron beams can be produced. The most common thermionic emitter types are pure materials such as tungsten and LaB$_6$; oxide cathodes

Table 1. Representative intense beam generators, from Smith (1991).

	Pulse Power	Pulse energy	Duration (ns)	Voltage	Current
		Pulse Line Machines			
PBFA-II	70 TW	2.8 MJ	40	12 MV	5.8 MA
Double Eagle	7 TW	0.5 MJ	75	2 MV	3.5 MA
KALIF	1.5 TW	60 kJ	60	1.7 MV	1.3 MA
		Linear Induction Accelerators			
ATA	0.25 TW	10 kJ	40	50 MV	5 kA
Hermes III	16 TW	0.64 MJ	40	20 MV	0.8 MA

such as BaO; dispenser cathodes (tungsten impregnated with a mix of oxides such as BaO:CaO:Al_2O_3); cermet cathodes; and thorium-based cathodes. Thermionic emitters have the advantages of producing current densities up to 10^6 A/m^2 for a long or continuous pulses (lower than that for explosive and field emitters), and have good uniformity, long lifetime, and high beam quality. Disadvantages are the need for a cathode heater and susceptibility to poisoning at pressures above about 10^{-5} Pa.

In photoemission sources, the work-function energy for electron emission from materials such as GaAs and Cs_3Sb is supplied by laser illumination. These sources have the useful feature that electron emission can be turned on and off coincident with the laser pulse. Modulating the laser has the effect of producing a beam that is effectively bunched. Furthermore, because the electron emission is "cold," with very little excess energy over that required for emission, random thermal motions of the electrons imparted in thermionic sources can be eliminated. Current densities as high as 2×10^6 A/m^2 have been produced. The disadvantages of photoemitters are the possibility of plasma formation at the surface if the laser intensity is sufficiently high, the complication of adding a laser to the system, and the requirement for vacuums as low as 10^{-8} Pa.

Ion beams are produced by extraction of ions from cathode plasma; see Sec. 1.4.

1.3 Accelerating Diodes

The design of accelerating structures to produce electron or ion beams of high quality usually involves computer simulation in order to shape properly the electrodes and magnetic fields (if a field is used). See Septier (1983). Nevertheless, the basic features of diodes are illustrated by two idealized models. First is the Child-Langmuir diode in parallel planar geometry, which applies when the self-magnetic-field of the beam is small. The second is the pinched-flow model, parapotential flow, valid when the self-magnetic-field of the beam strongly influences the particle trajectories in the accelerating region. At nonrelativistic voltages below 500 kV, provided we can ignore the self-magnetic-field of the beam, the beam current density J depends on the diode voltage V according to the Child-Langmuir law. For an anode–cathode gap d, the law takes the form (in commonly used units)

$$J=(2.33\ \mathrm{kA\ MV^{-3/2}})\,V^{3/2}/d^2. \tag{6}$$

This expression is also based on the assumption that the cathode can supply enough current to be space-charge limited, in which case the electric field at the cathode is zero and the current is a maximum. The relativistically correct version of the Child-Langmuir law is complex, but for 500 kV and above the following approximate expression is quite accurate:

$$J=\frac{(2.71\ \mathrm{kA})}{d^2}\left[\left(1+\frac{V}{0.511\ \mathrm{MV}}\right)^{1/2}-0.847\right]^2. \tag{7}$$

Note the identical scaling with the anode–cathode gap spacing in both cases. The voltage scaling is quite different, however, with the familiar $V^{3/2}$ scaling at low voltages, while at high voltages $J\sim V$.

In real diodes of the Child-Langmuir type, several phenomena left out of this analysis can be important. First, the azimuthal self-magnetic-field of the beam causes electron beams to pinch at high currents due to the force $F_r=v_zB_0$. The critical beam current above which beam pinching occurs is estimated by equating the Larmor radius for an electron at the edge of the beam with the anode–cathode gap:

$$I_c=8.5(r_c/d)\gamma\ \mathrm{kA}, \tag{8}$$

where r_c is the cathode radius. Pinching leads to a very different flow pattern, the parapotential flow (Fig. 5). For $I>I_c$ electron flow is derived from a model in which electrons move along conical equipotentials, for which the parapotential current is,

$$I_P=I_c\ln[\gamma+(\gamma^2-1)^{1/2}]. \tag{9}$$

In order to extract the electron beam, the anode must be a screen or foil thin enough (i.e., much less than one electron range) that the electrons can pass through without undue angular scattering. The foil becomes a stress point in the design, since it will have a limited lifetime. Also, beam heating of the anode can create an anode plasma, which contributes to gap closure. Finally, the conducting anode

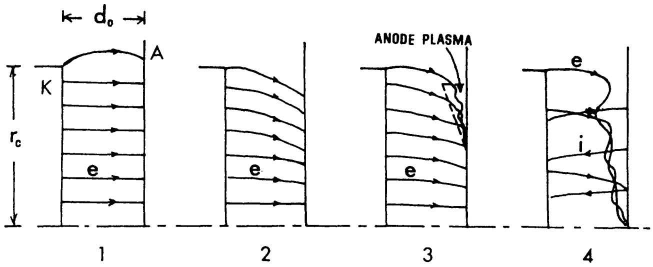

FIG. 5. Evolution of charge flow in an intense pinched electron-beam diode. (1) Child-Langmuir electron flow at low voltage; (2) electron flow at high voltage; (3) anode-plasma–induced collapsing electron ring; (4) electron and ion trajectories in pinched diode (parapotential flow).

will locally short-circuit the radial electric field of the beam, which disturbs the beam electron trajectories and introduces further spread in the transverse velocities of the electrons. Gap closure has the effect of making the diode gap a time-dependent quantity: $d=d_0 - v_p t$, where closure velocities are typically greater than 100 m/μs.

A means of avoiding the problems associated with foil or mesh anodes is to use a foil-less diode.* An example of such a diode is shown in Fig. 6. A strong magnetic field guides the electrons, preventing them from striking the wall so that they can be extracted down a drift tube. Foil-less diodes produce hollow electron beams, since emission at the outer edge of the cathode screens the smaller-radius portion of the cathode from the anode potential. The exact relation between current and voltage in these diodes depends on the details of the diode configuration. See Humphries (1990).

Just as with Child-Langmuir diodes, plasma expansion from the cathode is a problem with foil-less diodes. Plasma expansion both along and across the magnetic field can occur. Expansion along the field is faster, by roughly an order of magnitude, than cross-field expansion.

1.4 Ion-Beam Diodes

Intense ion beams at powers greater than a terawatt can be generated by megavolt-level voltages applied between two electrodes in vacuum if the electron flow in the diode is suppressed by using magnetic insulation. A magnetic field is supplied perpendicular to the electron path and to the electrodes so that it prevents electrons from crossing the gap. An anode plasma provides the ions that, being less affected by the magnetic field because of their greater mass, are accelerated by the electric field toward the cathode. The magnetic field in an ion diode serves two functions: it insulates the diode so that ions can be extracted, and it also traps electrons. The accumulated space charge of the electrons causes enhancement of the ion flux beyond the Child-Langmuir limit.

There are several ion diode types, depending upon how the field is generated and what geometry is chosen. One of the simplest types is the pinch reflex diode, which uses no static magnetic field. The flow develops as in Fig. 5. The electrons at first flow straight across the gap, then begin to pinch. They heat the anode, which begins to emit ions. Their space charge cancels that of the electrons, altering their flow. The result is the complex trajectories shown in the fourth part of Fig. 5. Since electrons move radially inward, the scale length for them is r_c. The ions cross directly, so their scale is d. This has the net result that the ion current is enhanced by roughly d/r_c

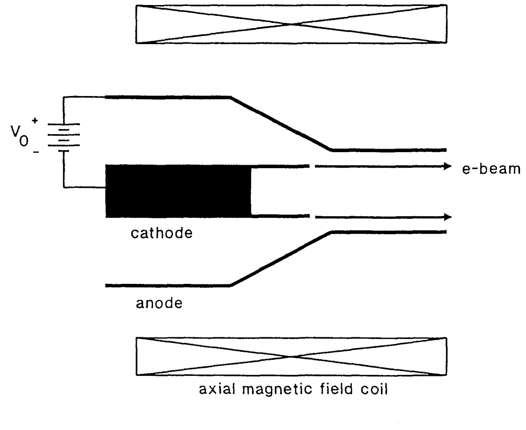

FIG. 6. Schematic of a foil-less diode.

---

*Diodes of this type are also commonly called *magnetically insulated diodes* in the Soviet literature.

relative to the Child-Langmuir diode. Therefore for large-aspect-ratio diodes ($r_c/d > 10$), large ion currents (~ 100 kA) can occur for megavolt applied voltages.

The pinch reflex diode has been largely superseded by applied-B-field diodes in which an external coil provides the insulating field, i.e., the magnetic field that must be applied to keep electrons drawn into the gap by the voltage from striking the anode. The two most common geometries are shown in Fig. 7. In Fig. 7(a), both anode and cathode are "barrel shaped," i.e., a structure symmetric about the center line. The ions cross the cathode and are focused in the center of the diode. The extractor diode, Fig. 7(b), is used in order to separate the target from the diode.

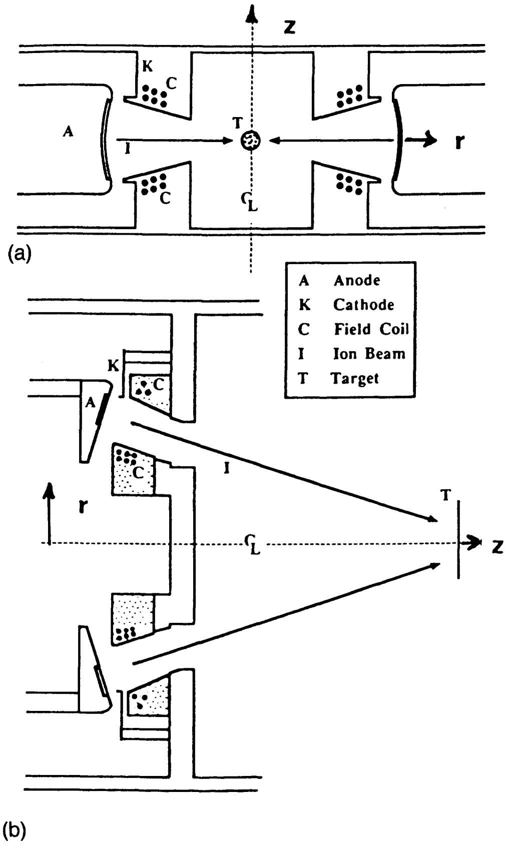

FIG. 7. Two forms of ion diodes used in inertial-confinement fusion studies. Above: barrel diode; below: extractor diode.

2. PROPAGATING INTENSE BEAMS

2.1 Beams in Gas and Plasma

Intense particle beams have self-fields sufficiently strong to dominate their behavior. Once a beam leaves the accelerator, its ability to travel a distance depends on canceling in some way the repulsive electrostatic force due to the charge of the beam electrons and the attractive magnetic force that arises from the beam current. In vacuum, the electrostatic force dominates, and the beam rapidly expands radially. Electron repulsion can be countered by supplying neutralizing ions from preexisting plasmas, by direct ionization of gas by beam electrons or by laser beam, or by applying a longitudinal magnetic guide field. In principle, fractional space-charge neutralization sufficient to provide force balance is possible, but in practice propagating beams very quickly become space-charge neutralized. When electrostatic fields have been eliminated by one or more of the above means, the magnetic force determines particle orbits.

There are three ways of exceeding the Alfvén limiting current. The naturally occurring method is "current neutralization," which arises from the tendency of plasma or any conductor to resist sudden changes in magnetic field. A fast-rising beam current induces an opposing axial electric field that drives a partially canceling return current in the plasma. The resulting net current (beam minus plasma current) determines electron orbits, so that v/γ is replaced by $(v/\gamma)_{\text{net}}$. For highly conducting plasma, the net current can be reduced to such a low value that beam electrons are not contained and expand to the boundaries of the chamber.

The second method of propagation is to employ a radially nonuniform current distribution. For example, a beam with most of the current flowing in a shell near the edge will propagate with $v/\gamma > 1$ by turning beam electrons only in the outer shell. Alternatively, in the Bennett distribution, the current density is peaked on the axis. It can be shown that v/γ is proportional to the ratio of the magnetic field energy within the beam to its kinetic energy. Nonuniform beam distributions reduce the magnetic energy, increasing the allowable current that can be propagated.

The third method, application of a longitudinal magnetic guide field, will prevent electrons from turning in their self-field, allowing propagation. The energy density of the field must be at least as great as that of the beam.

In recent years, experiments have demonstrated an electrostatic propagation aid. A laser ionizes a channel through a background gas. When the head of an electron beam passes, the electrons are expelled, leaving a background of ions. The remainder of the beam has its charge neutralized by the ions, so the beam propagates along the channel.

For further discussion of the above phenomena, see Humphries (1990).

2.2 Beams in Drift Tubes

A case of interest in many accelerators and beam-transport systems is the propagation of beams through metal pipes called drift tubes, usually in vacuum. Primary issues are space-charge depression of the energy of the beam electrons, limiting currents, beam equilibria, and beam quality.

Space-charge depression is a reduction of kinetic energy of the beam electrons by the self-electric-field arising from the beam space charge. Compare this to the ideal case of a low-current beam in which space-charge effects are negligible. In this case, the potential within the drift tube will be the same as the anode potential of the walls, and the kinetic energy of a beam electron will be given by the full diode voltage: $E_K = (\gamma - 1)mc^2 = eV_0$.

For an intense beam with significant space charge, the potential within the beam will be reduced from the anode potential by an amount determined by the beam space-charge density. Far from the diode region, the potential in the drift tube will be solely a function of radius, so that space-charge depression introduces shear (or radial variation) into the axial drift velocity of the electrons. As the beam current increases, the electrons are increasingly slowed down, until at some point a maximum current, known as the space-charge-limiting current, is achieved. One commonly occurring case is a thin annular beam drifting along a strong guiding magnetic field, for which the space-charge-limiting (SCL) current is

$$I_{\mathrm{SCL}} = (8.5\ \mathrm{kA})\,\frac{(\gamma^{2/3} - 1)^{3/2}}{1 + r_0/a}, \tag{10}$$

where a is the radius of the beam and r_0 is the radius of the drift tube wall. Note that I_{SCL} scales as $V_0^{3/2}$ at low voltages, and as V_0 at large voltages, similar to the Child-Langmuir diode.

For example, if a beam of 5 kA is injected at 425 keV with $a = 1.9$ cm and $r_0 = 2.4$ cm, then $I_{\mathrm{SCL}} = 12.6$ kA. The kinetic energy at injection is 340 keV, and the electrostatic potential energy is 85 keV, which is 20% of the total injected energy.

Electric and magnetic forces between the pipe wall and the beam can provide equilibrium in some cases. There is an induced image charge on the conducting wall. If the length of the beam pulse is shorter than the time it takes for magnetic field to penetrate the metal, there can be no magnetic beam-generated field outside the pipe, and therefore a return current exists on the wall equal and opposite to that of the beam. The image charges on the wall attract the beam and the return current repels it. If the beam is in the center of a cylindrical pipe, the image charges and currents are symmetric and no net force occurs on the beam. If the beam is off-center in the pipe, the beam shifts toward the outer wall and the repulsive force increases. An equilibrium ensues in which the force from the image charge balances the centripetal force. However, the beam can be destabilized if it encounters too abrupt a bend.

We have only touched on the large subject of beam instabilities. Many of the applications use instabilities to extract energy from beams in useful forms. For further discussion of beam instabilities and the above phenomena, see Humphries (1990) and Lawson (1988).

3. APPLICATIONS OF INTENSE PARTICLE BEAMS

The principal applications of intense particle beams are discussed here. The most studied applications are the topics described below. There is much interest today in industrial processes that could involve intense beams, but none of these has received sufficient attention in the research community to date.

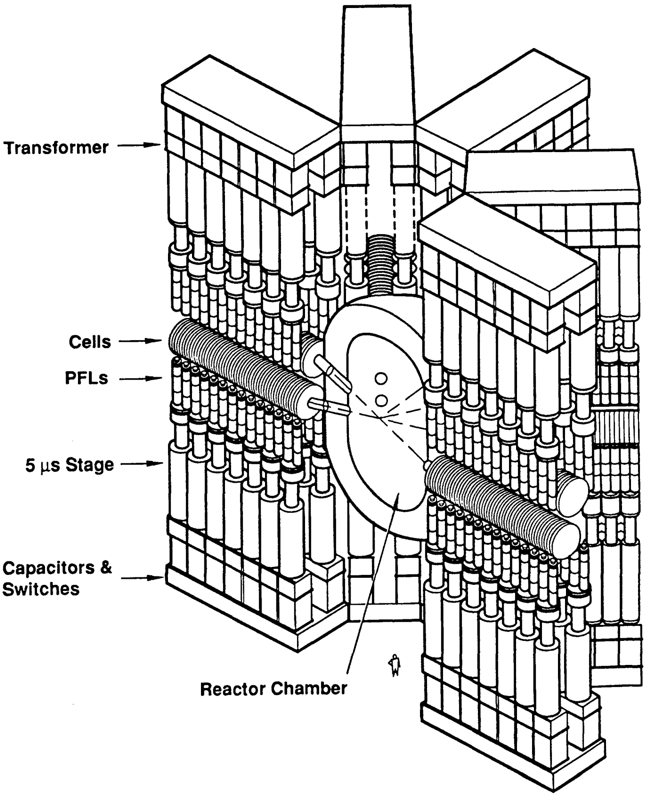

FIG. 8. LIBRA concept for a light-ion fusion reactor (30-MV, 4-MJ, 9-ns driver; 16 modules, 28 cells/module). A voltage adder injects beams into a reactor chamber. Note scale of human figure. PFL is pulse-forming line.

3.1 Inertial Fusion

Inertial fusion, as distinct from magnetically confined fusion techniques, centers around the rapid heating and compression of a small deuterium–tritium target at solid densities. The technique is called inertial-confinement fusion (ICF) because inertia of the target prevents its disassembly before fusion can occur (see FUSION, INERTIAL CONFINEMENT). Lasers and particle beams have been suggested for the driver. There are two approaches involving intense particle beams: one uses light ions and the other heavy ions.

The challenge of inertial fusion with particle beams is great. The particle-beam driver must deliver about 4 MJ of energy to the target, cost less than $100 per joule delivered, operate at an efficiency of at least 10%, avoid preheat of the fuel, and focus the beam to a 1-cm-diam target at about 100 TW/cm^2 with good uniformity. Any nonexpendable portion of the reactor must be at least 3–5 m from the pellet and must be built to survive the explosion. It would operate repetitively, at a few shots per second, in an environment with nuclear products.

The deuterium–tritium fuel must be compressed to approximately a thousand times its liquid density, and a small region in the center must reach a temperature of 4 keV. Ablatively driven spherical implosions compress and heat the fuel, driving a high-gain target. The key element is the ion-beam-producing diode, which is the juncture between the electrical energy produced by the pulsed power system and the transport system that carries the beam through the reactor cavity to the target. Heavy-ion fusion and laser fusion both have relative isolation because the transport methods to the target are ballistic and optical focusing, respectively, which place far fewer constraints on the total system than does light-ion focusing. The greatest challenge of light-ion fusion is achieving 100-TW/cm^2 power densities. An aid to focusing is that light ions at the approximately 30 MeV required are nonrelativistic, so that the voltage can be ramped upward to bunch the ion beam as it transits to the pellet. This technique also is used in heavy-ion fusion.

Many electrical technologies will be used in developing light-ion fusion, ranging from present-day, single-module, TW proton accelerators to repetitive, multimodule, multibeam, few-hundred-TW, multimegajoule lithium-ion accelerators, which will be needed to develop the final reactor chamber. The approach now being considered for the final reactor is the voltage-adder technique extended to much higher voltages, while keeping the current high enough for magnetic insulation. A conceptual view of such an accelerator is shown in Fig. 8.

Current concepts for an ion-beam diode are challenging. For example, the laboratory microfusion facility shown in Fig. 9 is conceived as 12 Hermes-III scale modules, each producing a 40-ns ion beam in which energy ramps from 18 to 20 MeV and current ramps from 460 to 720 kA. In addition, there are 24 accelerators delivering an ion pulse that ranges from 26 to 32 MeV and current that ranges from 800 to 1200 kA. Each beam bunches during its 4-m drift to the target, after which the pulse duration is 15 ns. The 36 lithium beams deposit 20 MJ onto a 2-cm-diam target. Extraction ion diodes are used with ballistic transport followed by focusing using magnetic lenses (Fig. 10). The challenges of efficient, low-cost, repetitive operation are immense.

3.2 High-Power Microwaves

High-power microwaves (HPM) is one of the most vigorous areas for applications of intense particle beams. Applications for high-power microwaves vary from particle accelerators to modification of the atmosphere, fusion heating, laser pumping, and electronic warfare. They generally require peak powers varying from 100 MW to 10 GW and average powers from 1 kW to as much as 1 MW radiated power. The more intense powers are required for large-scale applications such as modification of the atmosphere. For example, microwaves at a power level of 1 GW can be beamed into the lower ionosphere around 70 km to enhance local ionization. The moving microwave beam quickly paints an ionizing region, and a low-frequency (< 100-MHz) signal is bounced over the horizon, allowing long-range radar detection. Another suggestion is that microwaves can break up atmospheric contaminants such as chlorofluorocarbon (CFC), which is causing destruction of the ozone. Of course, the power required to clean the atmosphere is enormous. For a survey of HPM applications, see Benford and Swegle (1992).

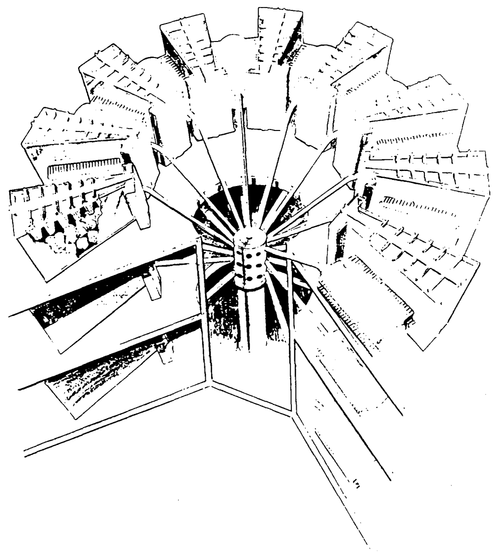

FIG. 9. Concept for a laboratory microfusion facility employing Hermes III–type technology.

Particle acceleration using rf fields has long been an application for conventional microwave devices. To achieve gradients of 100 MV/m requires raising the power or frequency. Typical values are > 100-MW sources operating above 10 GHz. Hundreds of such sources are necessary in order to drive next-generation accelerators.

The use of high voltages to produce relativistic electron beams has had several profound consequences for microwave generation. Very high power levels have been achieved. Figure 11 shows the peak power generated by a representative sample of high-power sources

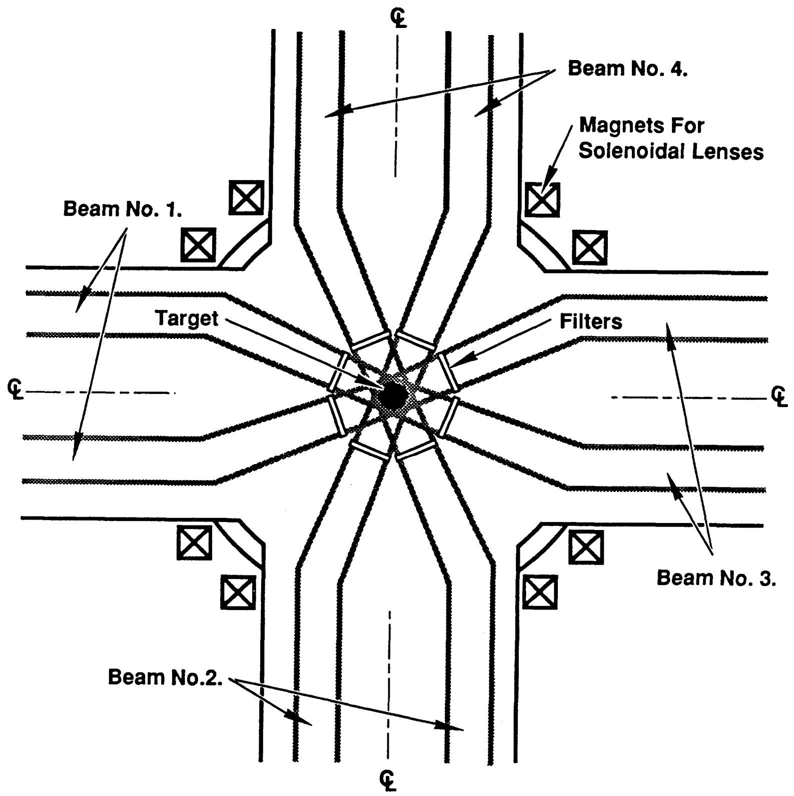

FIG. 10. Transport and focusing of multiple beams with ion lenses, filters, and overlap of beams on target.

as a function of the frequency. The factor P/λ^2 is a measure of the power density that can be produced. It appears that source power falls off roughly as f^{-2} at high frequencies; however, no clear trend emerges for the lower frequencies, below 10 GHz.

The limit on HPM source power is not known. Electrical pulses with power up to about 10 TW are available from a single pulse generator, and one can buy 1-TW generators commercially for a few million dollars. At a moderate extraction efficiency of 10%, one could therefore expect a peak power of 100 GW. Such a power will probably be achieved within the next decade. This will nevertheless require substantial improvements in the understanding of the limitations of specific devices and the overcoming of breakdown limits

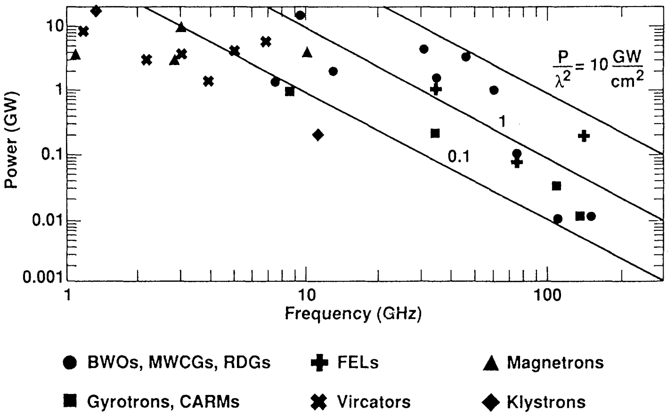

FIG. 11. Peak powers of pulsed high-power microwave sources.

Table 2. Classification of HPM sources.

	Slow wave	Fast wave
O type	Backward-wave oscillator (BWO) Traveling-wave tube (TWT) Surface-wave oscillator Relativistic diffraction generator Orotron Flimatron Multiwave Cherenkov generator Dielectric Cherenkov maser Plasma Cherenkov maser Relativistic klystron	Free-electron laser (ubitron) Optical klystron Gyrotron Gyro-BWO Gyro-TWT Cyclotron autoresonant maser Gyroklystron
M type	Relativistic magnetron Crossed-field amplifier Magnetically insulated line oscillator	Rippled-field magnetron
Space charge	Vircator Reflex triode	

and source-specific problems such as spurious mode generation.

Microwave sources can be classified into three basic types, as shown in Table 2:

1. O type: Devices in which the electrons drift parallel to an externally imposed magnetic field. This field helps to guide the intense beams used in HPM sources and, in some instances, plays an essential role in the microwave-generation process.
2. M type: Devices in which the electrons drift perpendicular to crossed electric and magnetic fields.
3. Space charge: Devices, possibly involving an applied magnetic field, in which the interaction producing microwaves is intrinsically traceable to intense space-charge interaction.

A consequence of the use of intense beams for HPM is the introduction of devices based explicitly on relativistic effects, such as the gyrotron. A second is stronger energy-, as opposed to velocity-, dependent tuning of the output frequencies in devices such as the free-electron laser (FEL) and cyclotron autoresonant maser (CARM). A third is the development of devices such as the virtual-cathode oscillator and the relativistic klystron that depend fundamentally on the very high beam currents that accompany high-voltage operation. Finally, a number of HPM sources are descendants of conventional microwave sources such as the magnetron, the backward-wave oscillator, and the traveling-wave tube. Nevertheless, in these devices, relativistic operation has brought increased power through the use of higher operating currents and stronger beam–field couplings within the interaction region.

An example of a space-charge device is the virtual-cathode oscillator, or vircator. An electron beam is generated in a diode with a thin foil anode that allows the beam to pass through to the drift space beyond. The maximum beam current that will pass through the drift region is the space-charge-limiting current I_{SCL}. One can ensure that the beam current exceeds this value by injecting the beam through a transmitting anode into a drift space of sufficiently large radius. In this event, a strong potential barrier will form at the head of the beam, from which a significant number of the beam electrons will be reflected back toward the anode. The reflection point for the electrons is known as the virtual cathode. Essentially, the excess space charge in the beam is rejected from the drift space. This situation is intrinsically different from other microwave-generation mechanisms. Instead of starting with a well defined initial beam state in which perturbations grow from small amplitudes at early times, in virtual-cathode formation there is a very fundamental breakdown of the beam state with the result that no beam equilibrium is possible. The position of the virtual cathode oscillates axially, shedding charge and behaving like a

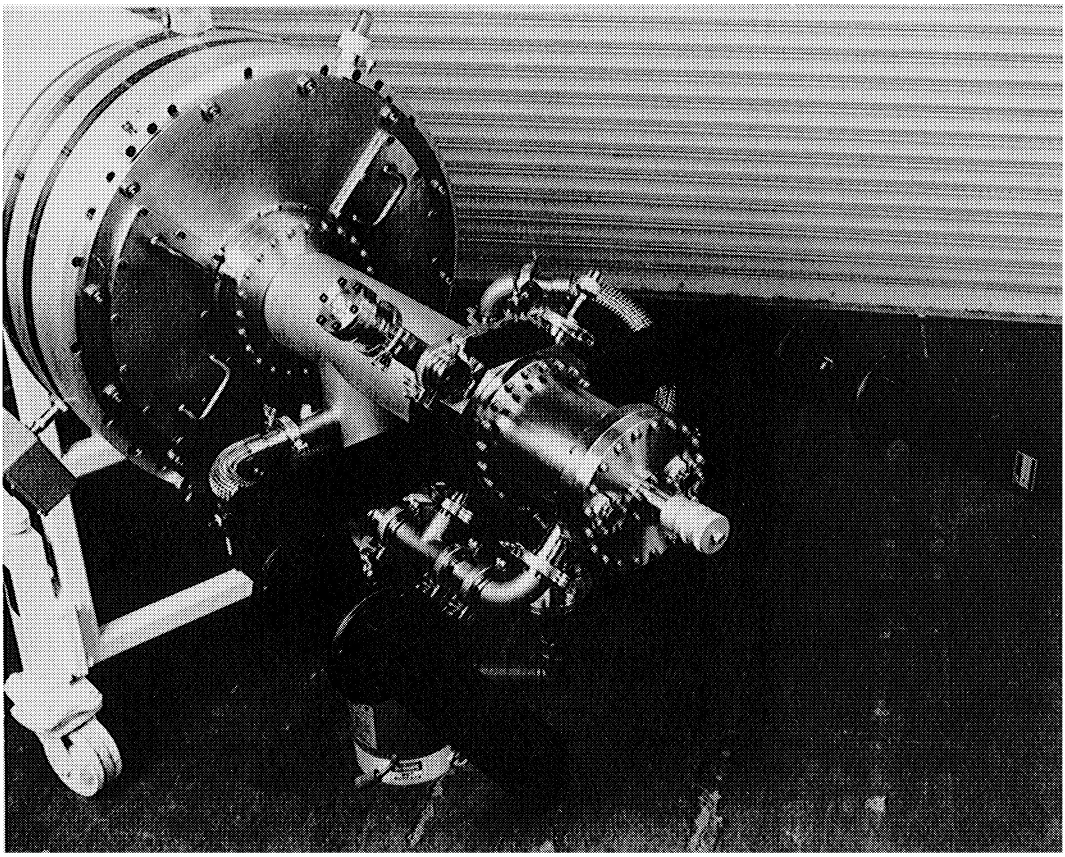

FIG. 12. Commercial vircator with tunable end wall and two extraction waveguides. Figure courtesy of Physics International Company.

relaxation oscillator. The motion of the space charge and potential well of the virtual cathode produces radiation at a frequency that is proportional to the square root of the beam current (of the order of the beam plasma frequency). The bandwidth of the vircator is quite large, though, and if the virtual cathode is formed in a resonant cavity, it will interact with the cavity modes to oscillate at a frequency tuned to a cavity mode. Figure 12 shows an advanced version of the vircator that is commercially available.

3.3 X-Ray Production

Intense beams allow production of very large fluences of x rays. At high voltages the efficiency of x-ray production can be large; for example, at 10 MeV x rays can be produced at about 10% efficiency in an optimized converter. Such converters contain a high–atomic-number material with a thickness of about a quarter of the electron range. Because the beam is intense, additional low–atomic-number material may be necessary to contain the impulse generated by the deposition of the beam. Efficiency is roughly proportional to voltage in this range, and therefore power will be proportional to the product of current and voltage squared. However, relativistic narrowing of the radiation cone enhances the radiated power in the forward direction so that power on target is more sensitive to voltage. At energies above 5 MeV, x-radiation power scales roughly as $V^{2.8}$. For example, at 10 MeV and 300 kA, x radiation can be generated at 0.3 TW and produce radiated energies of order 10 kJ. From such converters the mean x-ray energy is about a third of the energy of the accelerating potential of the electrons.

In order to produce much lower-energy x rays, pulsed-power devices are used to drive linear z pinches that, when their collapse stagnates on axis, generate copious amounts ($\sim$10 kJ) of low-temperature x rays with energy of order of 1 keV.

3.4 Industrial Applications

Intense particle beams have been developed for a variety of applications, usually involving defense, such as nuclear x-ray simulation and directed-energy weapons. Nonmilitary applications of particle beams are at an early stage of development. Examples are radiation treatment of food and waste, x-ray lithography, and powering of industrial lasers. These applications have a common thread of requiring high average power as well as high peak power. This is because they are performing high-volume processing. Continuous operation at high repetition rates is required. The ratio of peak to average power, usually referred to as the duty factor, must be large in order to make a system economically viable. Intense particle beams have not been developed with average power above about 100 kW, which is a threshold value for an industrial process. Therefore successful industrial applications of intense particle beams will have two basic requirements. The first is that the process must itself require high peak power in order for intense beams to be applicable. The second, the high average power of any industrial process, will stress the technology of intense particle beams that have been developed for single-shot work or low average powers. The future utility of intense particle beams in the emerging area of industrial processing lies in the ability to develop high-average-power, high-peak-power devices.

GLOSSARY

Bennett Beam Equilibrium: Model in which particle density is peaked on the axis.

Image Charge: Fictional charge used to describe electric fields in a region by replacement of the actual problem with charges external to the region of interest.

Larmor Radius: Radius of the circular orbit of an electron in a magnetic field.

Linear Induction Accelerator: Means of generating particle beams using vacuum pulse transformers based on ferrite magnetic cores.

Marx Generator: Means of producing high voltages by charging a group of capacitors in parallel and then firing switches to connect them; they then discharge in series.

Return Current: Conduction current on conducting surfaces near the beam that allows the current to return to its point of origin.

Space-Charge Fields: Electric fields due to distribution of electrons.

Voltage Adder: A linear induction accelerator with a metal cathode threading it, allowing peak voltage to appear only at its tip.

List of Works Cited

Benford, J., Swegle, J. (1992), *High Power Microwaves*, Norwood, MA: Artech House.

Humphries, S. (1990), *Charged Particle Beams*, New York: Wiley.

Lawson, J. D. (1988), *The Physics of Charged Particle Beams*, Oxford: Clarendon.

Miller, R. B. (1982), *An Introduction to the Physics of Intense Charged Particle Beams*, New York: Plenum.

Septier, A., (Ed.) (1983), *Applied Charged Particle Optics, Part C: Very-High-Density Beams*, New York: Academic.

Smith, I. (1991), "Pulsed Power in the United States," in: *Proceedings of the 8th IEEE International Pulsed Power Conference*, IEEE 91-CH52-8, p. 15.

Further Reading

Granatstein, V. L., Alexeff, I. (Eds.) (1987), *High Power Microwave Sources*, Norwood, MA: Artech House.

Humphries, S. (1986), *Physics of Charged Particle Acceleration*, New York: Wiley.

VanDevender, J. P. (Ed.) (1992), Special Section on Pulsed Power Technology, *Proc. IEEE* **80**, 931–1018.

Velurde, G., Martinez, J. (Eds.) (1991), *Nuclear Fusion by Inertial Confinement*, New York: CRC.

Yonas, G., Toepfer, A. (1978), "Relativistic Electron Beam Produced Plasmas," in: M. N. Hirsch and H. J. Oakam (Eds.), *Gaseous Electronics: Electrical Discharge*, New York: Academic, Vol. 1, Chap. 6, p. 399.

INTERACTION OF SOLIDS WITH PARTICLES AND RADIATION

S. E. SCHNATTERLY *Physics Department, University of Virginia, Charlottesville, Virginia, U.S.A.*

	Introduction	117
1.	**Particles**	117
1.1	Electrons	117
1.1.1	Low Energies	117
1.1.2	Intermediate Energies	118
1.1.2.1	Examples of Spectra for Various Solids	119
1.1.3	Radiation Damage	122
1.2	Ions	124
1.3	Neutrons	126
2.	**Electromagnetic Radiation**	126
2.1	Soft X Rays	126
2.2	Hard X Rays	128
2.3	γ Rays	128
	Glossary	130
	Works Cited	130
	Further Reading	131

INTRODUCTION

The interactions between solids and both energetic particles and ionizing radiation produce a wide range of phenomena and lead to numerous applications. These interactions form the basis of most detectors of particles and radiation. They also allow spectroscopies of the electronic structure of solids to be developed. In addition, various properties of the solid target can be altered, either by radiation damage or by the implantation of incident particles. In this article we describe and illustrate a number of these phenomena and applications. In Sec. 1 particles are discussed, including electrons, ions, and neutrons. In Sec. 2 electromagnetic radiation is discussed, including soft x rays, hard x rays, and γ rays.

1. PARTICLES

Several elementary particles and many ionized atoms can be formed into beams and directed onto a solid target. At high energies the interesting aspects of the collisions that occur have to do with the properties of elementary particles and the nuclei in the solid. Here we deal with lower energies where characteristics of the solid or the modification of its properties are the primary focus.

1.1 Electrons

Electrons are bound to the atoms out of which every solid is made. The outermost, or valence, electrons create the cohesive forces that bind the solid together (see COHESIVE ENERGY). These electrons are held within the solid by a binding energy of typically several electron volts (eV). An electron volt is the energy an electron gains or loses in passing through a potential difference of one volt. Valence electrons do not, however, all have the same total energy. They exist in distinct states having a range of energies. This range is referred to as the valence-band width, and varies from less than 1 eV to several eV in different solids. When an additional electron is injected into a solid the extent of its interaction with valence and core (very tightly bound) electrons depend on its own kinetic energy.

1.1.1 Low Energies. An electron in a solid with energy in the range of a few to several thousand eV interacts strongly with the valence electrons. Its electric field causes valence electrons to be excited to higher energy states. The energy gained by the valence electrons is lost by the moving electron, causing it to slow down rapidly. In a metal this rapid slowing down continues all the way to the Fermi energy, which is the lowest energy the extra electron can have. In an insulator or semiconductor the strong coupling to the va-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

lence electrons continues until the kinetic energy of the electron is equal to the band gap measured from the bottom of the conduction band. Below this point it does not have enough energy to excite an electron across the band gap and can therefore no longer lose energy by this process.

Transitions like these in which one electron loses energy while another gains energy are called Auger excitations. In this energy range Auger collisions occur so frequently that they can be thought of as causing a viscous friction force. The average distance an electron moves between Auger collisions is called the mean free path and depends strongly on the kinetic energy of the electron. The minimum value of the mean free path is typically 3–4 Å and occurs for a kinetic energy of about 30 eV. The mean free path is also sometimes referred to as the escape depth, since in photoemission experiments the important information comes from electrons excited by ultraviolet photons that emerge from the solid without losing energy.

1.1.2 Intermediate Energies. As the kinetic energy of the electron rises above 30 eV, the mean free path increases, and continues to do so to higher energies. The strength of the coupling between the fast electron and those in the solid decreases, and inelastic collisions become less frequent. It then becomes possible to produce sample materials in the form of thin films such that only one collision occurs as the fast electron passes through the sample. This means that the energy and momentum lost by the fast electron are transferred entirely to a single excitation in the solid, making it possible to demonstrate the existence of and study the dynamics of electronic excitations in solids by this means.

As the incident electron moves rapidly through the solid, the electrons in the material feel a pulse of electric field as it moves past. When an electric field is applied to a solid, its response can be characterized by a dielectric function $\epsilon(q,\omega)$, Here q is the wave vector of the applied electric field, equal to $2\pi/\lambda$, where λ is its wavelength and ω is its frequency. When the fast electron scatters in the solid, creating such an electric field, in order to conserve energy and momentum it must give up energy $\hbar q$, where $\hbar$ is Planck's constant divided by 2π. The relation between the scattering cross section and the dielectric function is (Schnatterly, 1977)

$$\frac{d^2\sigma}{d\Omega\, d\omega}=\frac{\hbar\gamma^2}{(\pi e a)^2 n q^2}\,\mathrm{Im}\left(\frac{-1}{\epsilon(q,\omega)}\right), \tag{1}$$

where n is the number density of electrons in the sample, a is the Bohr radius, and γ is the relativistic mass enhancement factor. Since the dielectric function has roughly the same magnitude for different momentum transfers q, this shows that the cross section falls off approximately as $1/q^2$.

The above equation for the scattering cross section is of great importance for our understanding of the dielectric properties of solids. It says that we can determine $\epsilon(q,\omega)$ by measuring the scattering of fast electrons through a thin film of a given material. Actually, what is directly measured is not the dielectric function itself, but the imaginary part of its inverse. The real part of the inverse dielectric function can be obtained from the imaginary part using a Kramers–Kronig integral relation. This requires knowing the imaginary part over a wide energy range. In the limit that q becomes small, the dielectric function obtained in this way becomes equal to that measured optically. To evaluate the total cross section for scattering through all angles and losing all possible energies, the above equation must be integrated with respect to Ω and ω. Using a simple statistical model of atoms to do so results in (Schnatterly, 1977)

$$\sigma=8\pi\gamma^2 Z^{1/2}\left[k^2\ln\left(\frac{Z^{3/2}}{(\Delta E/2E)^2(ka)^2}\right)\right]^{-1}. \tag{2}$$

This total inelastic cross section allows mean free paths to be estimated as a function of atomic number Z and electron beam energy E with forward momentum $\hbar k$.

In addition to the inelastic scattering processes described above, elastic scattering of the fast electron can occur from the atoms in the solid, with no loss of energy and no excitation of electrons in the material. The elastic cross section remains nearly constant with q out to a few Å^{-1}, while the inelastic cross section starts at a larger value and decreases rapidly with increasing q.

An important property of the dielectric function is that it obeys an important integral relation known as the oscillator strength sum rule (see Wooten, 1972):

$$\int \epsilon_2(q,\omega)\omega\, d\omega=\frac{\pi}{2}\omega_p^2. \quad (3)$$

Here $\omega_p^2=4\pi ne^2/m$ is called the plasma frequency of the electrons with number density n. Since the dielectric function is defined in terms of the polarizability per unit volume of the solid (see **Glossary**), the oscillator strength is proportional to the number of electrons per unit volume. This sum rule is a great help in making sure that measured cross sections or dielectric functions have the correct overall magnitude.

1.1.2.1 Examples of Spectra for Various Solids. Figure 1 shows inelastic electron scattering data for LiF, an ionic insulator with a very large band gap. LiF is sometimes regarded as the simplest of the alkali halides, since Li and F have fewer electrons than the other alkali-metal or halogen atoms. Figure 1 shows scattering power plotted against excitation energy in eV for four values of q in both the (100) and (110) directions. The lowest energy peak, at about 14 eV, is due to an exciton, which is an electron in the conduction band bound to a hole, or missing electron, in the valence band. An exciton is similar to a hydrogen atom with the hole playing the role of the proton. The excitation spectrum of an exciton is similar to that of a hydrogen atom but with a different energy scale for each solid. The dominant feature in the spectra shown in Fig. 1 is a strong peak near 25 eV, which is due to a plasma excitation. This is a collective motion of all the valence electrons in LiF. As q is increased the plasma peak loses strength relative to the other features in the spectrum and splits into two partially resolved peaks. This splitting is due to the interaction with the periodic field of the ions of the crystal lattice. The other features in the spectra are due to interband transitions that also change shape and position with q.

The change in position and shape of these features with q provides dynamical information about the electronic properties of LiF. For example, note that the exciton energy increases as q is increased. Figure 2 shows a graph of this exciton energy plotted against q for both the (100) and (110) directions. The curvature of this graph for small q gives an effective mass for the exciton. In other words, if you could reach into the solid and push on the exciton, how much inertia would it have? Since mass is a measure of the inertia of an object, the greater the mass (i.e., the less the curvature shown in Fig. 2), the greater the inertia of the exciton.

This question has a more fascinating answer than one might at first expect. In classical physics, when two particles are bound together, their total inertia is unaffected. This is a fundamental conclusion coming directly from Newton's laws of motion. Here we have an electron and a hole bound together, but it turns out that their inertia is affected by the binding. Figure 3 shows a measurement of E_p, the exciton peak energy, versus the square

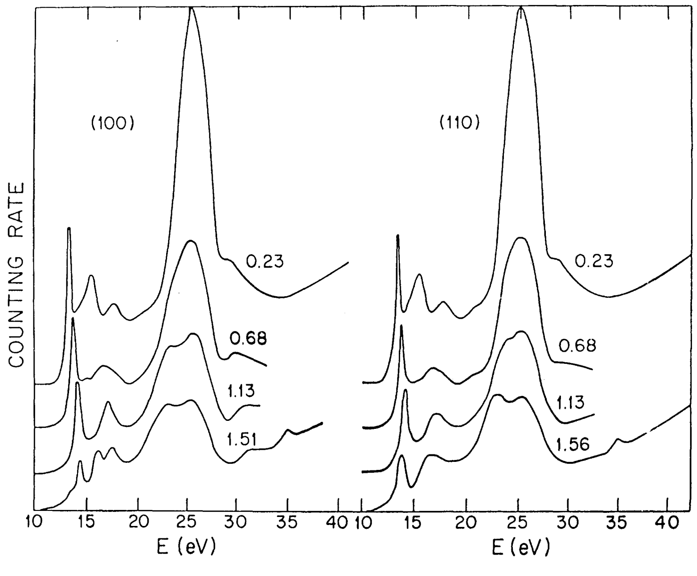

FIG. 1. Inelastic electron-scattering spectra for LiF. Momentum transfers in the (100) and (110) directions are given in Å^{-1} (Schnatterly, 1977).

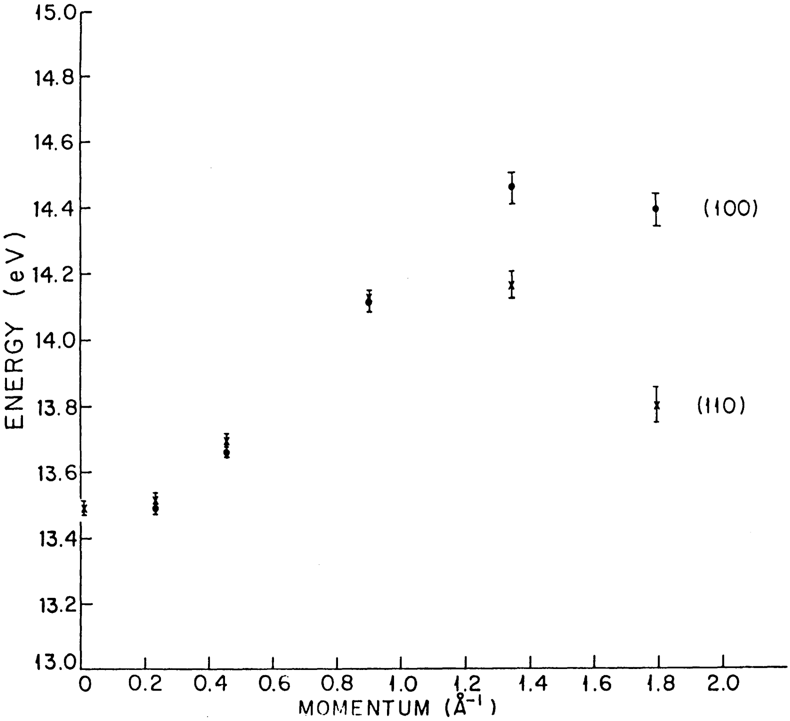

FIG. 2. Energy of the LiF valence exciton vs q for the (100) and (110) directions (Schnatterly, 1977).

of q for NaCl. The slope of this line gives a value for the mass of the exciton in this material. Plotted on the same graph is the energy to create an unbound electron and hole, E_g. The slope of this line is greater, indicating that the unbound electron–hole pair in NaCl is lighter than the bound pair. This is a quantum-mechanical effect due both to the wave nature of matter on the atomic scale, and to the fact that these electrons and holes are moving in a lattice with which they interact rather than in free space. This example stands in contrast to the other way that the mass of a system depends on its binding, namely special relativity. In that case the total mass of the composite system is decreased; here it is increased, and the effect is due to quantum mechanics, not relativity.

Figure 4 shows the real (dashed line) and imaginary (solid line) parts of the dielectric function for LiF plotted versus energy. These results were obtained from electron-scattering measurements, as described above, using a small momentum transfer q, and so they should agree with the results of optical experiments. In the 10–20-eV range are the exciton and valence excitations discussed above. At

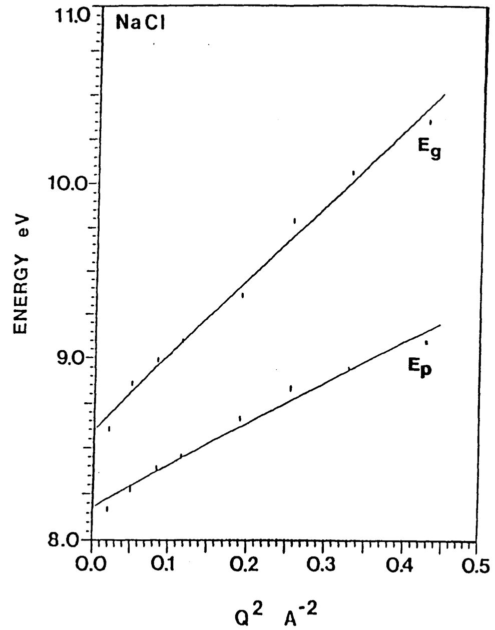

FIG. 3. E_g, the energy of an unbound electron–hole pair at threshold, and E_p, the energy of a bound electron–hole pair for NaCl vs q^2 (Schnatterly *et al.* (1988).

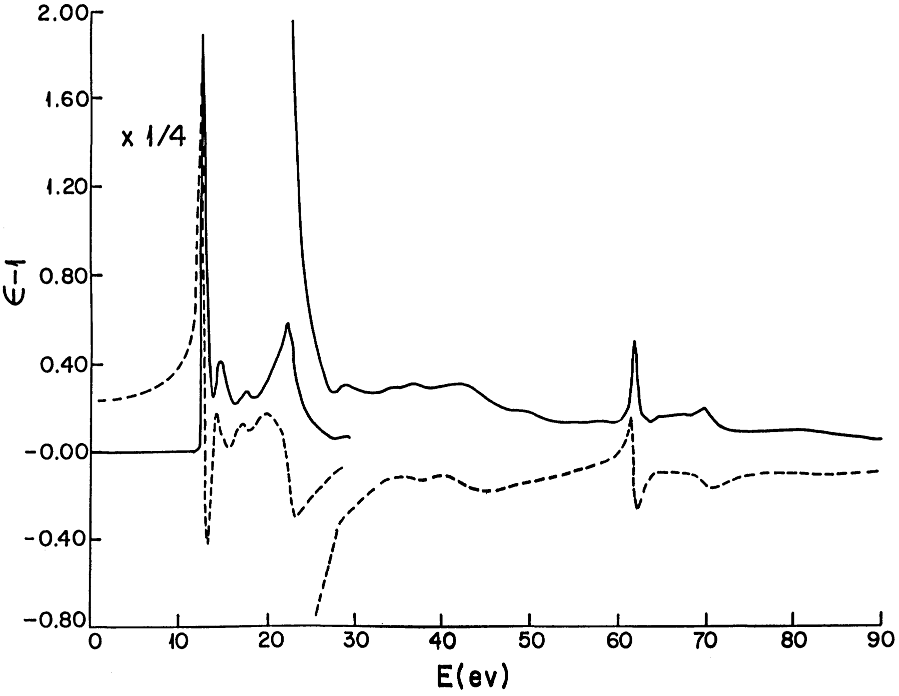

FIG. 4. $\epsilon - 1$ for small q vs E for LiF. The dashed line shows the real part and the solid line the imaginary part of the dielectric function (Schnatterly, 1977).

much higher energies, in the 60–70-eV range, is another set of features. These are due to the excitation of an electron out of the 1*s* core level of Li. The 1*s* level is the most tightly bound state on any atom. These electrons have a small-radius state and do not interact significantly with the other electrons in LiF. Core levels such as this play an important role in other spectroscopies of solids such as XPS (x-ray photoemission spectroscopy) and Auger experiments (see AUGER SPECTROSCOPY).

Simple metals such as Na and Al were the first kinds of materials studied by inelastic electron scattering. This is partly because some metals are easy to prepare in thin-film form by vacuum evaporation, and partly because the measured spectra were dominated by the plasma excitation, and so appeared simple to interpret. In simple metals the plasma peak stands out even more strongly than that for LiF above. Figure 5 shows an example for Al. The shoulder at 1.5 eV is an interband transition; at 7 eV is a surface plasma excitation; at 15 eV, and stronger by a factor of more than 100, is the bulk plasma excitation; at 30, 45, and 60 eV are multiples of this dominant feature in which electrons create two, three, and four plasma excitations. Beginning at 72 eV are features due to the 2*p* core of Al. All these excitations were measured with q small compared with the Fermi wave vector of Al.

The plasma excitation is a direct consequence of the strong tendency of electrons in solids to screen out applied electric fields. For metals this screening is complete for macroscopic fields at zero frequency. One of the important boundary conditions used in electrostatics is that there can be no electric field inside a metal. The highest frequency that the electrons in a metal can follow is the plasma frequency. Above that frequency metals become much more transparent to light; the reflectivity and absorption coefficient both drop sharply.

As the wavelength of an applied field decreases (the wave vector increases), the screening decreases even below the plasma frequency. It is in this regime that we need to understand the screening effect in order to develop quantitative models of the interactions between electrons as they move about inside a solid. The most direct way to study this experimentally is to measure the q dependence of the plasma excitation in simple met-

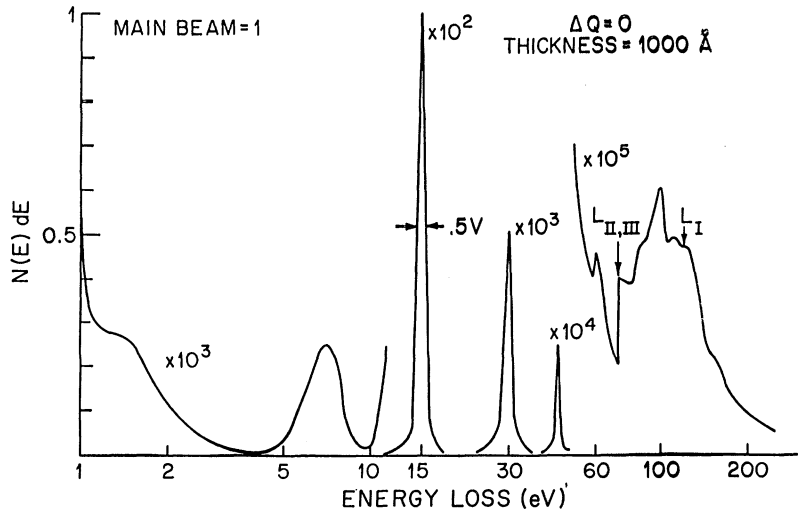

FIG. 5. Excitations in Al measured by inelastic electron scattering (Schnatterly, 1977).

als. As q is increased the plasma energy and the width of the peak both increase. Figure 6 shows the plasma energy of Al as a function of q. Figure 7 shows the width of the plasma peak in Al versus q. The peak width increases very rapidly for q greater than a cutoff wave vector q_c. In the simplest model, below this value the width is expected to be zero. Understanding the shape of this excitation in simple metals over a wide range of q values is a fundamental problem in condensed-matter physics. Much effort has been expended in developing models for this purpose (see Mahan, 1990).

Intermediate in terms of electronic properties between simple metals such as Al and good insulators like the alkali halides are the semiconductors. Excitons and interband transitions are much weaker in typical semiconductors than for good insulators. Plasma excitations in semiconductors often have widths for small q that are intermediate between those for metals and insulators. All atoms have core levels, and so semiconductors exhibit core spectra that often reveal information about the chemical state of the atom involved. Figure 8 shows the $2p$ core threshold for both Si and SiO_2. The threshold energy is shifted up in the oxide compared with the semiconductor and the shape is markedly different.

1.1.3 Radiation Damage. Electrons with energies below roughly 100 keV damage some materials rapidly and others not at all. In this energy range damage is caused by creation of electron–hole pairs in the solid. The number of electron–hole pairs created is roughly equal to the ionizing energy deposited divided by $2.8E_g$, where E_g is the band gap of the material. An electron excited out of the valence band of an insulator is a broken chemical bond. Whether the bond is reformed before a nearby atom moves from its normal lattice site is a question with no simple answer. Generally speaking, metals and covalent semiconductors do not damage rapidly, if at all, by

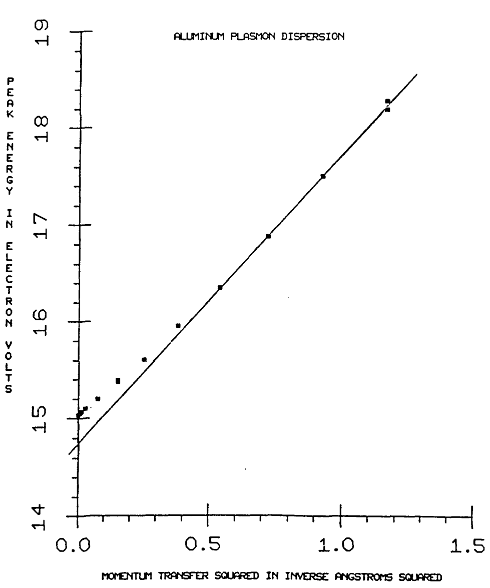

FIG. 6. Energy of the Al plasmon vs q^2.

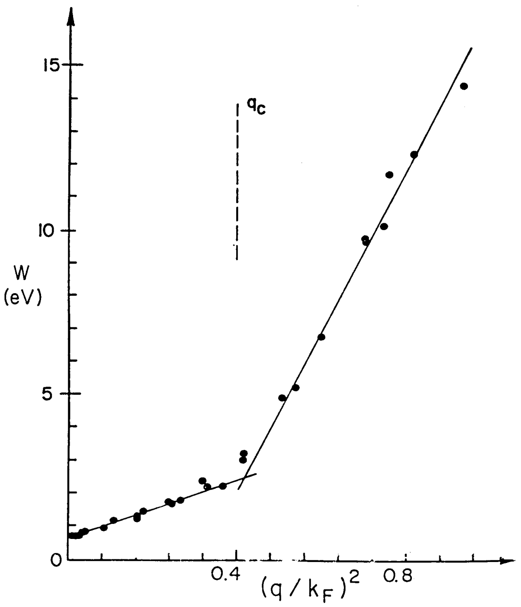

FIG. 7. Width of the Al plasmon vs q^2. q_c is the value of q below which the width should be zero according to the random-phase approximation (Schnatterly, 1977).

this mechanism. Ionic and organic materials are much more sensitive to damage by electron–hole pair excitation. In some materials two electron–hole pairs must be excited very near each other before damage occurs.

As the energy of an electron beam is increased, eventually the electrons have enough momentum to dislodge an atom from its lattice site by direct collision. Since the nucleus must be hit head-on, the cross section for this kind of damage is small. When an electron with energy E strikes an initially stationary atom of mass M, the maximum kinetic energy that can be transferred to the atom T_{max} is

$$T_{\text{max}}=2(E+2mc^2)E/Mc^2\ . \tag{4}$$

For each atom in a solid there is a threshold value of energy transferred to the atom, T_d, that will result in its being displaced from its normal lattice position. Values of T_d vary from 4 to 25 eV, with the value for a particular atom being typically about twice the bond energy for that atom. For example, for Si, T_d= 13 eV, which corresponds to 145 keV for the incident electron-beam energy.

Radiation damage effects in alkali halides are representative of those in insulators generally and have been extensively studied. The most readily produced atomic defect caused by electron–hole pair excitation in alkali halides is a negative-ion vacancy, which then traps an electron to produce a color center called an *F* center. The *F* center is an example on the atomic sale of a particle in a box. It is a direct consequence of the wave nature of matter that, as the box is made smaller, the energy of the electron in it increases. This also means that the energy difference between the ground and first excited states of the electron increases as the size of the box is reduced. The color associated with color centers comes from light being absorbed as the electron makes a transition between these two states. Changing the lattice constant of the crystal then changes the wavelength of light that is absorbed. This variation is described by the Mollwo–Ivey law: $\lambda_p = 703d^{1.84}$ where λ_p is the wavelength of the peak of the absorption band in Ångström units, and d is the lattice constant of the alkali halide in ångströms.

As the temperature is lowered, the *F*-center absorption peak shifts to shorter wavelengths and the width of the band decreases. The width is due to the interaction of the electron with the vibrations of the surrounding atoms, and the shift is due primarily to thermal contraction as the lattice constant becomes smaller at lower temperatures. The magnetic moment of the trapped electron allows *F* centers to be studied by electron paramagnetic resonance. *F* centers are a model system in solid-state physics, which have stimulated

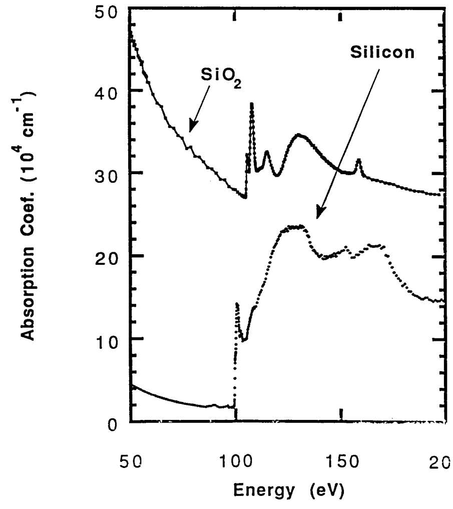

FIG. 8. Si 2*p* core threshold region for Si and SiO$_2$ (Husk *et al.*, 1991).

the development of new experimental techniques and theoretical concepts (see Fowler, 1968).

F centers interact with each other, and a kind of *F*-center chemistry occurs. Two *F* centers bound together form a new entity originally called the *M* center, which is the analog of the hydrogen molecule. Three and four *F* centers in nearest-neighbor positions also form stable aggregates. Higher concentrations of *F* centers are required to form these aggregate centers, which of course requires a greater dose of ionizing radiation. Eventually enough halogen atoms leave the crystal that alkali-metal particles are formed within the material. These particles also give rise to interesting optical effects since they scatter incident light strongly and in a manner that depends on wavelength.

Perhaps the most common material in the Earth's crust is SiO_2. It is also important in the semiconductor industry, since it is used to form insulating layers in Si integrated circuits. Exposure to ionizing radiation causes the formation of oxygen vacancies, creating the E_1' center, which absorbs light in the ultraviolet region of the spectrum. Defects such as this trap charge and cause inhomogeneities that interfere with the operation of high-density integrated circuits. (see Verweij and Wolters, 1983.)

Radiation damage produces defects in a number of insulators that can trap charge carriers and hold them for very long times, even millions of years. As such a material is heated, the trap lifetimes become rapidly shorter as a result of the Boltzmann factor, which describes the probability of thermally ionizing the trap. By observing the light emitted as the crystal is heated, the total number of defects caused by radiation can be determined. This technique is called thermoluminescence, and can be used to determine how long ago certain minerals were at elevated temperatures. The rate at which radiation produces traps can be measured for a given material. Then if the rate of cosmic-ray incidence or natural radioactivity is known, the material's thermoluminescent output indicates how long it has been subjected to the radiation at ambient temperatures. This is useful in geology (age since mineral formation) and anthropology (e.g., age of pottery).

Electrons with sufficiently high energies lose energy in solids primarily through bremsstrahlung. This is a scattering process in which the electron comes very close to the nucleus of an atom and is strongly accelerated. According to electromagnetic theory any accelerated charge must radiate. The radiation in this case is in the form of x rays. For example, in Pb, above about 10 MeV, the electron inelastic-scattering cross section is dominated by bremsstrahlung and is proportional to the energy to well above 1 GeV. This means that the energy lost by an energetic electron in a certain path length is proportional to the electron's energy, thus providing a method of measuring that energy.

As energetic electrons pass through Pb, or any other material, they generate x rays that propagate out in all directions from the path of the electron beam. This is called an electromagnetic shower. The x rays are scattered and eventually absorbed, creating electron–hole pairs. The number of pairs is proportional to the energy deposited in the material. Semiconductors allow the number of pairs to be measured by the current they generate in an electric field; phosphors (insulators doped with optically active charge traps) allow the number to be measured by the amount of light generated. In either case these interactions allow the energy of the electron to be determined through a variety of different experimental arrangements.

1.2 Ions

Ions incident on a solid surface cause many different physical and chemical processes to occur that are of both scientific and technological interest (see ION-BEAM MODIFICATION OF MATERIALS). Light ions of keV or more energy undergo a series of energy-loss collisions with both atoms and electrons, eventually stopping hundreds to thousands of atomic layers from the surface. Even for a monoenergetic beam of incident ions, there will be a distribution of penetration depths, approximately Gaussian in shape, due to the statistical nature of the collisions. This provides a means of doping the target material near the surface. Ion implantation, in contrast to chemical diffusion, allows a greater range of composition and concentration of dopant atoms because of the nonequilibrium nature of the process. The *p-n* junctions in many integrated circuits are formed using ion implantation.

Heavy ions with keV or more energy can create a collision cascade in which many hundreds of atoms in the target are displaced from their regular lattice sites. For example, 10-keV Bi ions incident on Si create amorphous tubes in which thousands of Si atoms along the flight path of a Bi ion have completely lost their local crystalline arrangement. These tube diameters are approximately 50 Å, and when the incident dose is sufficient for them to overlap, the entire surface region is amorphous.

Another process, important for surface studies, is sputtering (*q.v.*). Incident ions, often 3-keV Ar ions, deliver most of their energy very near the surface. Surface atoms, through multiple collisions with other target atoms, can be ejected into the vacuum. This provides a means of cleaning oxides and other contaminants from a surface. By using a mass spectrometer to identify the ejected ions, sputtering can be used to determine target composition as a function of depth. For some materials electron–hole pair excitation alone can cause sputtering without the need for direct momentum transfer from an incident ion.

In all of the above processes radiation damage plays a role. Often this is an undesirable aspect of the process, and it is desirable to anneal the damage away. Some fascinating things have been learned in studying these phenomena in semiconductors. If a heat pulse lasts less than about 1 ns and it contains sufficient energy, the result is to convert crystalline material to an amorphous structure. The heat pulse could come from a passing heavy ion, in which case it might last 10^{-11} s, or from a very short laser pulse. The physical origin of the lattice energy does not seem to matter. Longer heat pulses, usually produced by lasers, in the ns to μs regime, can produce local melting, followed by rapid crystal growth. Annealing can also be accomplished by the more conventional method of heating the entire target to an appropriate temperature.

In metals phenomena similar to those in semiconductors occur. Bulk radiation damage can be monitored by measuring the electrical resistivity of the metal. A metal with no structural disorder will have zero resistivity at a temperature of 0 K. A metal irradiated at low temperature and then warmed up while its resistivity is monitored reveals the temperatures at which various kinds of radiation damage anneal. Surprisingly, even for "hard" metals (e.g., Pt) annealing begins at very low temperatures, typically 10 K. X-ray diffraction and EXAFS (extended x-ray absorption fine structure) can also be used to monitor radiation damage.

For gas ions (e.g., H, He, Ar) incident on metals in the keV range an important phenomenon occurs at high dose levels. After the incident ions have been neutralized in the metal and have come to rest, they can usually diffuse around at room temperature. If the concentrations are high enough, they will aggregate to form bubbles. This often occurs for incident doses in the $(10^{16}\text{–}10^{17})/\text{cm}^2$ range. Bubbles cause blistering, which is an important source of degradation of thermonuclear fusion reactors. The densities of gas atoms or molecules in a bubble are comparable to those of the liquid, and yet the molecules are at room temperature, and so the pressures are very high. The bubbles near the surface eventually burst, causing a pockmarked appearance under a microscope and degraded vacuum and strength characteristics.

For incident ions in the MeV range deeper penetrations occur. For crystalline targets the depth of penetration is highly dependent on the target orientation relative to the incident beam direction. This is due to the occurrence in most crystal lattices of open regions within each unit cell, which line up with each other in particular directions in the lattice. If an incident ion beam is collimated and aimed along one of these directions, much of the beam penetrates considerably further than if the beam direction is changed by only 1° or 2°. This effect, referred to as channeling, can be observed by monitoring the fraction of backscattered ions as a function of target orientation. Channeling disappears completely for amorphous materials, and so it provides a means of monitoring the crystallinity of the surface of a target during experimentation or processing.

Energetic ions incident on thin targets will penetrate through and emerge from the other side, having lost some energy. The emerging ions often have lost a number of electrons and are in a highly excited state. Many excited-state lifetimes can be measured by monitoring the intensity of emitted light as a function of distance from the thin target. Experiments based on this technique are referred to as beam-foil spectroscopy. Another fascinating aspect of this kind of experiment is that electrons often emerge from the target having the same velocity as the ions, but not bound to them. These are referred to as convoy electrons. Convoy electrons are largely an atomic phenomenon, since their velocities are much greater than the Fermi velocity and they also occur in gases.

1.3 Neutrons

Neutrons interact with atoms in solids in a completely different manner than ions or electrons, since they have no charge. The cross sections for neutrons to scatter from atoms are about eight orders of magnitude smaller than those for keV ions or electrons, and so the mean free paths are that much longer. Thermal neutrons are available from research nuclear reactors in collimated beams that can be directed onto a sample. They have wavelengths of a few ångströms, comparable with the lattice constants of typical crystals. This means that thermal neutrons can be used to study crystal structures by diffraction, just as can x rays of a similar wavelength. A major difference is that the thermal-neutron scattering cross section varies irregularly from nucleus to nucleus in the periodic table, whereas the x-ray cross section varies as Z^2, where Z is the atomic number. This means that light elements are difficult to observe with x rays, while they may even dominate when observed with neutrons. Thermal neutrons also cause no radiation damage since they carry so little energy and momentum. Thermal neutrons are also used to study the vibrational energies of atoms in solids. For this they are uniquely suited, having both the correct momentum and energy for the task.

In addition to scattering directly from nuclei, neutrons possess a magnetic moment, which allows them to scatter from the magnetic moments of electrons in magnetic solids. In particular, antiferromagnetic materials, which have no net magnetization, could only be understood when thermal-neutron sources became available to study them after World War II.

Nuclear reactors also produce neutrons in the MeV energy range, which cause radiation damage by direct displacement of atoms from their regular lattice sites. Both point defects and collision cascades, in which a large number of atoms in a region are displaced, can occur. Samples can be irradiated by lowering them into the core region of a reactor and leaving them for a specific period of time. An interesting variation on this process is transmutation doping, in which the neutrons transmute one chemical element into another. For example, ^{30}Si, which has a natural abundance of about 3%, can be converted to P by absorbing a thermal neutron. This turns the Si into an *n*-type conductor with nearly perfect random placement of the P atoms and no introduced disorder. High-quality power diodes are doped in this way (see POWER ELECTRONICS).

2. ELECTROMAGNETIC RADIATION

The electromagnetic spectrum covers a wide range of wavelengths from radio waves to γ rays. The longer-wavelength regions are discussed in other articles in this Encyclopedia. Here we understand radiation to mean ionizing radiation, and begin with the soft–x-ray region.

2.1 Soft X Rays

X rays used for diffraction studies of solids have energies of 10 keV or more. Soft x rays have energies in the range of 20–30 eV to about 1000 eV. In this region of the spectrum man-made diffraction gratings can be used to resolve different wavelengths, but they must be used at grazing incidence to achieve reasonable reflectivities. Soft x rays are absorbed strongly by all solids, penetrating only 0.1–1 μm depending on wavelength.

Soft–x-ray absorption spectra exhibit sharp and detailed structure above the threshold energy for exciting an electron from a core level into empty states in the conduction band of the material being studied. Core spectra have sharp features in this energy range because these core levels have long lifetimes. At higher energies core states have shorter lifetimes, and in agreement with the energy uncertainty principle their spectra exhibit more broadening.

Core absorption spectra such as those shown in Fig. 8 reveal information about states available to electrons in the immediate vicinity of the atom being excited, in this case Si. As the figure shows, these states change significantly when the chemical state of the atom is altered. Spectra like those shown in Fig. 8 can be measured using soft x rays produced by a synchrotron, as well as by inelastic electron scattering, as was the case for the results shown in the figure.

Soft-x-ray emission spectra consist of soft x rays emitted by a sample after a core state has been ionized. An empty core state can be filled in two primary ways: A valence electron makes a transition into the core state while another valence electron is excited to higher energies. This is called an Auger transition, and typically dominates the process in this energy range. The second possibility is that a valence electron makes a transition into the core state and a soft x ray is emitted. The energy distribution of soft x rays then reveals the energy distribution of valence electrons since the core-state energy is fixed.

Figure 9 shows an example of soft-x-ray emission spectra for crystalline and amorphous Si. This is the inverse transition to that shown in Fig. 8. The energy position of the emission band is determined by the core-state binding energy, and thus will be different for each kind of atom. The shape of the spectrum is determined by the electronic structure of the valence band of the material. It is clear from Fig. 9 that crystalline Si has more structure than amorphous Si. The loss of long-range order in amorphous materials blurs the valence states of electrons, smearing out the structure in the density of states.

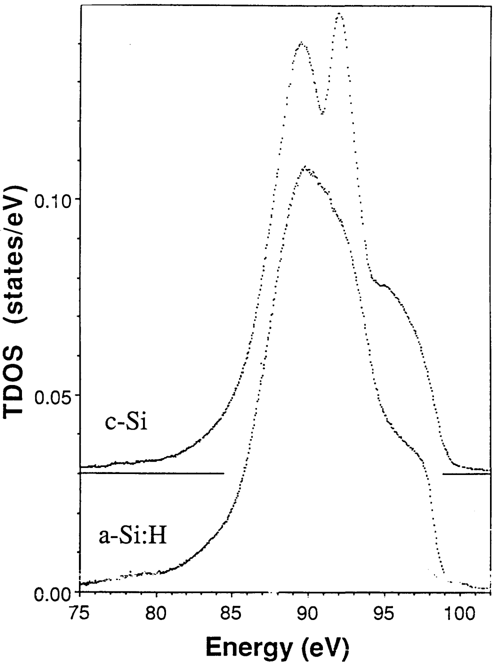

FIG. 9. Soft-x-ray emission of crystalline and amorphous Si due to transitions from the valence band to the $2p$ core level. The crystalline Si spectrum has been displaced upward by 0.03 units (Bruhwiler and Schnatterly, 1989).

Measurements such as these can be used as a check on our understanding of the valence electronic states in a material. This emission band in Si is due to transitions from valence states to the $2p$ core level. The dipole selection rule, which governs transitions such as these, requires that only s-like and d-like states in the valence band can contribute to the spectrum. The Si valence band is made up of $3s$ and $3p$ states, with a small contribution from $3d$ states. Since the $3p$ states are suppressed in this spectrum, it allows us to observe the smaller contribution of the $3d$ states, which is not possible in other spectroscopies such as photoemission, where the $3p$ states dominate.

Theoretical calculations of the shape of the Si emission spectrum of crystalline Si agree quite well with the result shown in Fig. 9. Calculations for amorphous Si have not yet been carried out.

Figure 9 also shows that some emission intensity remains at energies below the emission band. This is caused by a radiative–Auger transition in which both processes occur: A soft x ray is emitted and a valence electron is excited to higher energies, reducing the energy available to the soft x ray.

Figure 10 shows the emission spectrum of BN. This material has a layered structure, and its valence electrons have states either in the planes of the layers (σ states) or perpendicular to those planes (π states). This causes the emission spectrum to be anisotropic, allowing it to be separated into σ and π components as shown. The dashed line shows the radiative–Auger contribution for the π states. Also shown is a peak at about 192 eV in the π spectrum that is not part of the valence band. This is a core exciton, which for some materials contributes to the emission. Figure 9 shows a weak feature near 100 eV in the case of amorphous Si that is also due to a core exciton.

2.2 Hard X Rays

Higher-energy x rays are produced by both bremsstrahlung and transitions to core states. They are absorbed by matter predominantly by electron–hole pair creation. Structural studies in condensed-matter physics are greatly aided by x-ray diffraction, in which x rays scatter elastically from atoms, revealing their relative positions within a material.

Inelastic scattering of x rays in matter also occurs. In this case the incident x ray gives some of its energy and momentum to a single electron, emerging with a lower energy and longer wavelength. This process is known as Compton scattering and played an important role earlier in this century in verifying the reality of photons as the quanta of electromagnetic radiation. Compton showed that if λ_1 is the wavelength of the incident x ray, and λ_2 the wavelength of the scattered x ray,

$$\lambda_2 - \lambda_1 = (h/mc)(1 - \cos\theta) , \qquad (5)$$

where h is Planck's constant, m the mass of the electron, and c the speed of light. The combination h/mc is called the Compton wavelength of the electron and has the value 0.00243 nm.

Compton's formula above applies to electrons that have zero energy, either kinetic or potential, before interacting with the x ray. Electrons in solids have both. The interaction with the nucleus and the other atoms creates a binding energy, which can be very large for core states and is different for different atoms. Valence electrons have a range of kinetic energies as described above. This means that Compton's formula must be modified to include the effects of the initial energy and momentum of the electron. Measurements of inelastic x-ray scattering from solids then provide a means of probing these energies and momenta.

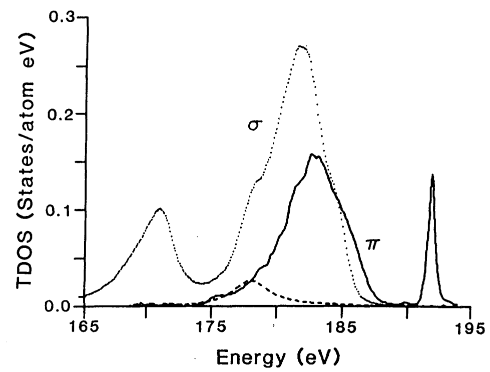

FIG. 10. Soft-x-ray emission spectrum of BN due to transitions from the valence band to the *B* 1s level (Mansour and Schnatterly, 1987).

The cross section for inelastic x-ray scattering from a solid has the same energy and momentum dependence as that for inelastic electron scattering, that is, it is proportional to the imaginary part of the inverse dielectric function. The coefficients multiplying the inverse dielectric function are completely different from those in Eq. (1) since the interaction of x rays with matter is different from that of charged particles. The size of the cross section is smaller than that for electrons by about 10 orders of magnitude. This means that macroscopic samples can be used as targets rather than the thin films used in electron scattering.

In principle, all the kinds of information described above obtainable about solids with electron scattering can also be found using x-ray scattering. In the past the counting rates in the measurements have been much lower, and the energy resolution obtainable has been worse, than that for electron scattering. As synchrotrons become more and more intense x-ray sources, this comparison will improve.

One interesting application of inelastic x-ray scattering derives from the fact that different atoms can be distinguished from one another because of their distinct electron energy levels. This makes it possible to scan an object, e.g., a package, to determine whether a given substance is present inside. With very sensitive x-ray detectors and a raster scanning source, the total radiation exposure is very low, and so this procedure can even be applied to human beings.

2.3 γ Rays

Higher-energy electromagnetic radiation is produced by nuclear transitions and collisions between elementary particles. As the energy increases, absorption of radiation by electron–hole pair creation decreases rapidly. Compton scattering becomes dominant typically above 0.1–1 MeV. At still higher energies, pair creation dominates the interaction of radiation with matter. At these energies, a photon near the nucleus of an atom can create an electron–positron pair. This requires a minimum of about 1 MeV, and the probability increases rapidly above that energy. For these and higher energies electromagnetic showers are produced by incident γ rays similar to those produced by energetic electrons.

Electron–hole pair creation, Compton scattering, and electron–positron pair creation are the three mechanisms by which γ rays interact with matter. These mechanisms underlie the production of detectors of γ rays, which can not only monitor their presence but also record their energies. In these detectors the following things must occur: first, the conversion of the γ-ray energy to kinetic energy of electrons by one of the above three mechanisms; second, the production of electron–hole pairs or excited molecular states by these electrons; and third, the collection and measurement of the charge carriers, or the light produced in the detector material.

Each γ ray then produces a pulse of either charge or light that can be measured. If the entire energy of the γ ray is deposited in the detector, then the size of the pulse is a measure of its energy.

Two such detectors will be described here briefly. NaI(Tl) is a scintillator material that has been used for years to detect x rays and γ rays. This consists of a crystal of NaI in which a small fraction of the Na atoms have been replaced with Tl atoms. The Tl impurities are optically active and can capture electrons and holes producing light. The cross section for stopping γ rays in a solid increases with atomic number, which is the reason the alkali iodide is used. These detectors are quite useful up to energies of several MeV. An important characteristic of these detectors is how much of the γ ray's energy is needed to produce each detected photon. For NaI(Tl) this is about 100 eV.

Ge is a semiconductor that is also widely used to detect γ rays. In pure Ge at low temperature no free carriers of charge exist within the material. Each of the four outer electrons of the Ge atom are contained in a valence band that cannot carry current. Above the top of the valence band is a gap region with no allowed electron states, followed by a conduction band. Only by exciting electrons into the conduction band can charge be carried through the Ge under the influence of an electric field. The gap region in Ge is about 0.7 eV wide.

A very pure crystal of Ge can be used as a γ-ray detector in the following way. An electric field is placed across the crystal, and the resulting current measured with an external circuit. Typically this is done at a temperature of 77 K, the boiling point of liquid nitrogen, so as to reduce the number of electrons thermally excited across the gap. When a γ-ray deposits energy in the Ge, many electron–hole pairs are created causing a pulse of current, which is measured. The average energy needed per created electron–hole pair is only about 3 eV.

The creation and detection of electron–hole pairs are statistical processes. This means that a beam of monoenergetic γ rays will produce pulses with a range of sizes. The energy resolution of the detector is determined by the width of this distribution of pulse sizes. This width increases approximately as the square root of the number of pairs detected. The γ-ray energy increases linearly with the number of pairs. This means that the relative resolution (energy divided by the uncertainty in energy) improves as the number of detected pairs increases. In this regard, Ge is a better detector than NaI(Tl), since more detectable events are created in Ge by the same energy γ ray than in NaI(Tl). In a typical detector, for example, a 1-MeV γ ray can be detected with a pulse width of about 1.6 keV.

Energy-sensitive detectors such as these are used in numerous applications involving radioactivity. For example, a company producing a radioactive source of ^{201}Tl, which is used in medical imaging, must specify not only how much activity there is from this isotope, but also how much is present due to impurities. Using a Ge detector, it is possible to detect impurity radionuclides down to the 0.001% activity level.

An important method used to identify the various chemical elements present in a sample of material is neutron-activation analysis. The sample is placed in a nuclear reactor, where it is subjected to a thermal neutron flux of up to 10^{14} cm^{-2} s^{-1} for perhaps a few days. The neutrons react with the nuclei present in the sample, changing them to new nuclides, many of which will be unstable and will emit radiation. γ-ray emitters are particularly valuable in this application because the cross sections for their production vary by large factors between radionuclides, making it easier to distinguish them. An important prerequisite for this to work well is either a low neutron cross section for the matrix material and a high cross section for the trace element of interest, or a large difference in half-lives of the produced radionuclides. Some elements, in particular the light ones, cannot be detected by neutron-activation analysis, since their nuclei do not become active under neutron exposure. This is an advantage in looking for trace elements in organic and geological materials, since the activation of the matrix (i.e., H, C, N, O, Si, P, S, and Ca) is negligible. For example, rare-earth elements can be detected in rock samples at the level of nanograms per gram. A Ge detector is a good choice for this application.

GLOSSARY

Auger Transition: A relaxation process in an atom, molecule, or solid in which one electron moves to an empty state with lower energy and another takes the energy gained in being excited to a higher energy state.

Band Gap: An energy region for which no electron states exist within a solid. The band gap most widely referred to is that just above the valence band.

Bremsstrahlung: The production of electromagnetic radiation, especially x rays, by the acceleration of an incident electron by the electric field of an atomic nucleus.

Channeling: Deep penetration of a beam of incident ions into a crystal when the direction of the beam is aligned along certain crystallographic axes.

Collision Cascade: The displacement of hundreds of atoms in a crystal by a single incident heavy ion with sufficient energy.

Compton Scattering: The scattering of an x ray by an electron with negligible initial energy.

Core Electron: An electron tightly bound to an atom in a small orbit so that in a molecule or solid it does not interact strongly with electrons on other atoms.

Dielectric Function: A coefficient describing the electric polarization of a solid, usually defined by $\epsilon E = E + 4\pi P$, where ϵ is the dielectric function, E the electric field, and P the induced dipole moment per unit volume. ϵ varies with q and ω where q is the wave vector and ω the frequency of the applied field.

Electron Volt: The energy an electron gains or loses upon passing through a potential difference of 1 V.

Exciton: An electron and a hole bound to each other in a semiconductor or insulator.

F Center: A negative-ion vacancy in an ionic solid that has trapped an electron.

Hole: A missing electron in a distinct state in an atom or solid. The hole behaves like a positively charged particle.

Ion Implantation: The process of adding atoms to a solid by exposing it to a beam of energetic ions.

Kramers–Kronig Relation: An integral equation relating the real and imaginary parts of the dielectric function. This allows the complex dielectric function to be evaluated if some measured property (reflectivity, inelastic electron scattering cross section) is known over a wide energy range.

Mean Free Path: Average distance between collisions for a particle moving in a medium (e.g., an electron in a solid).

Neutron-Activation Analysis: A test procedure for identifying chemical elements in a sample by exposing it to thermal neutrons in a reactor. This converts many elements to a radioactive form that can then be identified by their emissions.

Plasma Excitation: An excitation in solid in which a large number of electrons, e.g., all the valence electrons, move coherently together at a characteristic frequency (the plasma frequency) in response to an applied field.

Sputtering: The removal of surface atoms from a solid by bombarding with energetic ions, often Ar.

Thermoluminescence: Emission of light by a material as it is being gradually heated. The increasing temperature causes electrons and holes to be released from traps; as they recombine light is emitted.

Transmutation Doping: A method of doping Si in which one of the Si isotopes is transmuted to P by exposure to thermal neutrons in a reactor.

Valence Electrons: The outermost electrons on an atom, which form the chemical bonds that hold molecules and solids together.

List of Works Cited

Bruhwiler, P. A., Schnatterly, S. E. (1989), *Phys. Rev. B* **39**, 12 649–12 654.

Fowler, W. B. (1968), *Physics of Color Centers*, New York: Academic, Eq. (2-2).

Husk, D. E., Tarrio, C., Benitez, E., Schnatterly, S. E. (1991), *J. Appl. Phys.* **70**, 3338–3344.

Mahan, G. D. (1990), *Many Particle Physics*, 2nd ed., New York: Plenum, Chap. 5.

Mansour, A., Schnatterly, S. E. (1987), *Phys. Rev. B* **36**, 9234–9240.

Schnatterly, S. E. (1977) in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, New York: Academic, Vol. 34.

Schnatterly, S. E., Tarrio, C., Cafolla, A. A. (1988), *Excitons in Confined Systems*, Springer Proceedings in Physics Vol. 25, p. 68.

Verweij, J. F., Wolters, D. R. (1983), *Insulating Films on Semiconductors*, Amsterdam: North-Holland.

Wooten, F. (1972), *Optical Properties of Solids*, New York, Academic, Chap. 6.

Further Reading

Debertin, K., Helmer, R. G. (1988), *Gamma and X-ray Spectrometry with Semiconductor Detectors*, Amsterdam: North-Holland.

Johnson, R. A., Orlov, R. N. (1986), *Physics of Radiation Effects in Crystals*, Amsterdam: North-Holland.

Schnatterly, S. E. (1977), "Inelastic Electron Scattering in Solids," in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.) *Solid State Physics*, Vol. 34, New York: Academic.

Williams, J. S., Poate, J. M. (Eds.) (1984), *Ion Implantation and Beam Processing*, New York: Academic.

INTERCALATION COMPOUNDS

M. S. DRESSELHAUS, *Department of Electrical Engineering and Computer Science and Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*

G. DRESSELHAUS, *Francis Bitter National Magnet Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*

Introduction 133
1. **Structural Organization** 133
1.1 Layer Stacking Arrangements . 134
1.2 Requirements for Intercalation 136
1.3 Classification According to Rigidity 138
1.4 Classification of the Intercalate Layers 139
2. **Structural Characterization** .. 140
2.1 *c*-Axis Ordering 140
2.2 In-Plane Ordering 143
2.3 Interlayer Correlations 145
2.4 Intercalate Phase Transitions .. 145
3. **Electronic Properties** 146
3.1 Charge-Transfer Phenomena .. 146
3.2 Transport Properties 148
3.3 Electronic Structure 148
4. **Magnetic Intercalation Compounds** 148
5. **Superconducting Intercalation Compounds** 151
Acknowledgment 154
Glossary 154
Works Cited 154
Further Reading 155

INTRODUCTION

Intercalation compounds in general are formed by the insertion of atomic or molecular layers of a guest chemical species, called the intercalate, between layers of a host material. Intercalation compounds have received a great deal of attention from the physics, chemistry, and materials science communities because the intercalation process leads to the synthesis of new classes of materials, modifying the properties of both the host material and the guest species. In this article, the structure and properties of intercalation compounds are reviewed. These materials are of special interest because of their potential both for practical applications and for the study of novel physical phenomena. Shown in Fig. 1 are several schematic examples of layered intercalation compounds, with the basic structural unit of the host material ranging from a thin monolayer to a thick multilayer unit. In each case, the addition of the intercalant has little effect on the basic structural unit of the host material.

1. STRUCTURAL ORGANIZATION

To allow insertion of entire layers of guest species, the host material is usually a layered material, with very strong in-plane bonding and very weak interplanar bonding. There are many examples of host materials that support intercalation (see Fig. 1), including graphite (Dresselhaus, 1984; Dresselhaus and Dresselhaus, 1981; Solin, 1982, 1984; Kamimura, 1987), which has been studied most extensively; transition-metal dichalcogenides (Gamble and Geballe, 1976; Liang, 1987); some silicates and metal chlorides (Gamble and Geballe, 1974); some clays (Wada *et al.*, 1987; Newman, 1987); and some polymers and gels (Aldebert *et al.*, 1982). Intercalation modifies the properties of the host materials, giving rise to materials that are more conducting, more anisotropic, magnetic, or capable of storing large concentrations of hydrogen, to give some examples. Correspondingly, intercalation results in the modification of the intercalate, often changing the chemical reactivity drastically because of the bonding between the intercalate and the host material.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

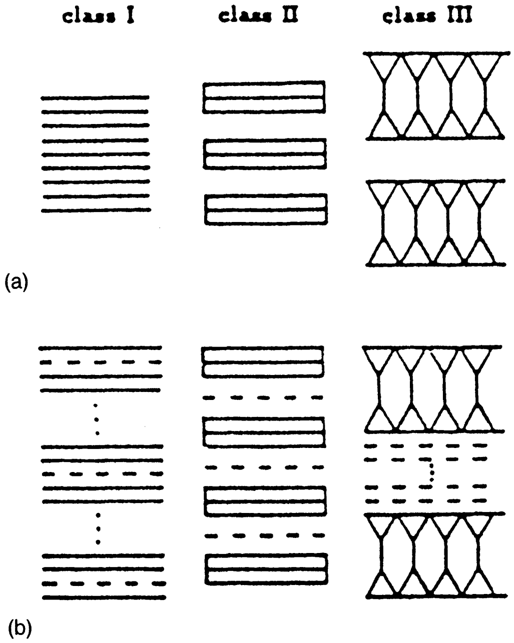

FIG. 1. Schematic classification of (a) pristine layered host materials and (b) the intercalated forms of (a). In (b) the dashed (solid) lines represent guest (host) layers. Three classes of intercalation compounds are shown, class I having the lowest layer rigidity and class III having the greatest rigidity for the host material (Solin, 1984).

1.1 Layer Stacking Arrangements

In discussing the structural organization of intercalation compounds it is necessary to comment on the structure of the host material and of the intercalate prior to intercalation, the effect of intercalation on these structures, and the structure of the entire intercalation compound. In this subsection, the layer stacking arrangements are discussed. First, host materials have their own specific structures prior to intercalation. In general, the basic structural unit of the host material is unaltered in the intercalation process. For example, the graphite host material is a layered compound with *ABAB* stacking and four atoms per unit cell (see GRAPHITE). After an intercalate layer has been inserted, the same intralayer atomic arrangements persist, and the same interlayer correlations between adjacent host layers are maintained within the basic structural unit of the intercalation compound, but the host-layer stacking sequence across basic structural units will in general be different if an intercalate layer is introduced between two layers of the host material.

Likewise the transition-metal dichalcogenide host material shown in Fig. 2 consists of a trilayer containing two chalcogenide monolayers *X* between which a transition-metal monolayer is sandwiched. This strongly bound trilayer unit maintains its structure as an intercalate layer is introduced in the van der Waals gap between the trilayer sandwiches, as indicated in Fig. 2. Similarly, the large, strongly bonded silicate basic structural unit also remains unaltered as the intercalate is inserted between silicate basic structural units as shown in Fig. 3. The intercalate may enter every available van der Waals gap (or intercalate gallery) between basic structural units of the host material in the case of a saturated intercalation compound with the maximum intercalate concentration. For more dilute intercalation compounds, only a fraction of the available van der Waals gaps are occupied by intercalate guest species. For those basic structural units of the host material connected by unoccupied van der Waals gaps, the stacking sequence remains un-

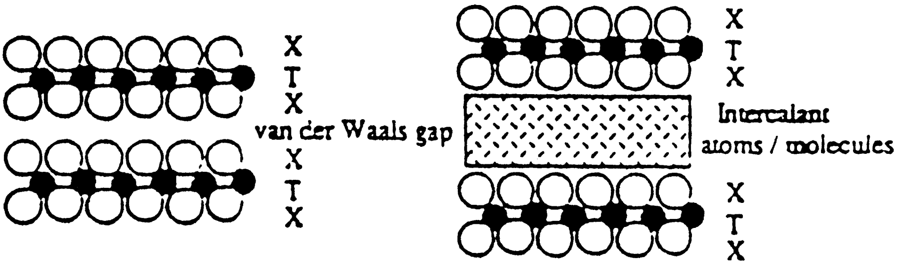

FIG. 2. Schematic representation of the trilayer transition-metal dichalcogenide TX_2 sandwiches and the interlayer van der Waals gap before and after addition of an intercalation complex (Liang, 1987) between layers of TX_2, where *T* denotes the transition-metal layer (usually from groups IV, V, and VI in the periodic table) and *X* is the chalcogenide layer (i.e., S, Se, or Te).

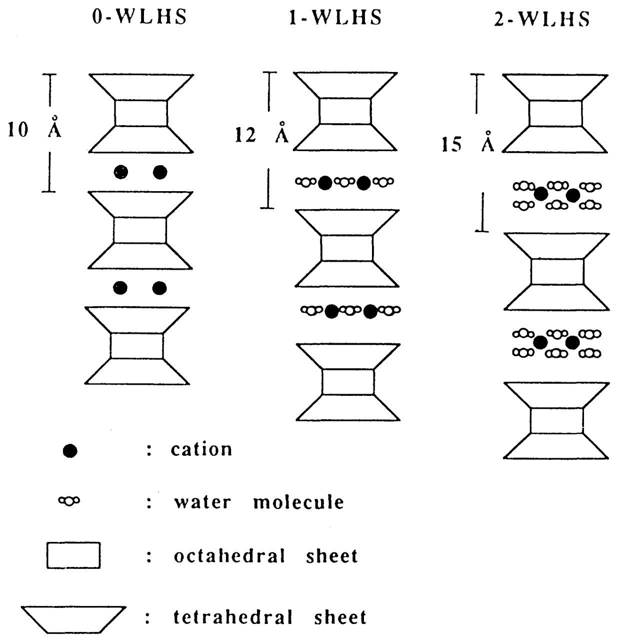

FIG. 3. Schematic representation of the vermiculite clay structure for zero-, one-, and two-water-layer hydration states (WLHS). Typical values for the *c*-axis repeat distance are indicated.

changed relative to the bulk host material prior to intercalation. The structure of the intercalate layer and of the resulting intercalation compound is more complex and is discussed in Sec. 2.2.

A special feature of intercalation compounds is the staging phenomenon, by which a constant number of layers of a host material are sandwiched between sequential intercalate layers, as shown in Fig. 4. The staging phenomenon thus gives rise to a periodic array of intercalate sheets [a one-dimensional (1D) superlattice normal to the layer planes] that can independently undergo 2D order–disorder and structural phase transitions within the layer planes (see Sec. 2.4). Staging allows control of the separation between sequential intercalate layers, thereby providing a means for control of the magnitude of the intercalate interplanar interactions. This control is exploited in studies of a variety of 2D phenomena including quasi-2D magnetism. The two layers of the host material between which the intercalated layer is sandwiched are called the *bounding* layers, while the layers of the host material with host layers on either side are called *interior* layers. Of all the host materials, graphite forms intercalation compounds (GICs) exhibiting the highest degree of staging fidelity.

Whereas intercalation normally refers to the insertion of complete atomic layers of guest species into a layered host material, the term intercalation is also used to denote the insertion of lines of atoms into host materials. For example, the high-T_c superconductor $YBa_2Cu_3O_7$ has chains of oxygen and copper (Cu2) atoms where the stoichiometry is CuO, and planes of the Cu and O species (Cu1) where the stoichiometry is CuO_2. The insertion of oxygen atoms into the chain planes, which contain available sites for O atoms, has also been called intercalation. It should be noted, however, that in the highly anisotropic high-T_c superconductors of the bismuth-based oxide series Bi-Sr-Ca-Cu-O [$Bi_2Sr_2CuO_x$ (2:2:0:1), $Bi_2Sr_2CaCu_2O_x$ (2:2:1:2), and $Bi_2Sr_2Ca_2Cu_3O_x$ (2:2:2:3)], layered intercalation compounds have been synthesized with the indicated basic structural units for the host material, using planes of iodine for the intercalate (Xiang *et al.*, 1991).

In some cases, such as fullerenes, the molecules of the host material are bound to each other by weak van der Waals forces, allowing the guest species to enter interstitial tetrahedral or octahedral sites. In this case the host material has been ordered crystals of C_{60} or C_{70} molecules, and the intercalate species

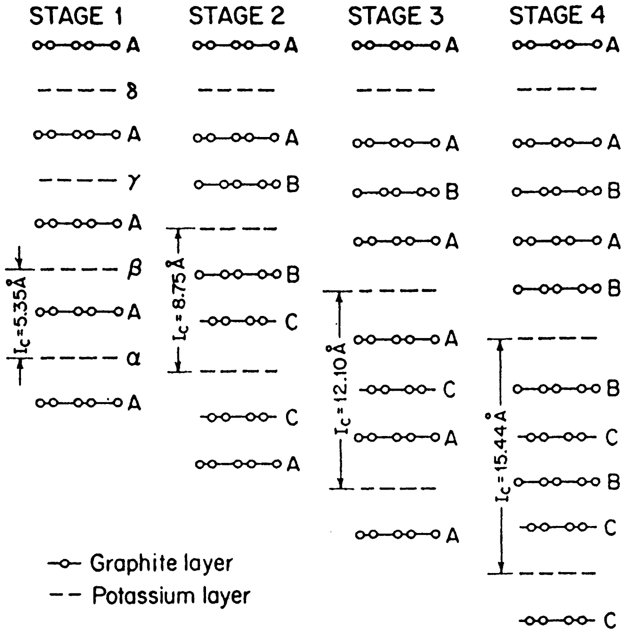

FIG. 4. Schematic diagram illustrating the staging phenomenon in certain layered intercalation compounds for stages $1 \geqslant n \geqslant 4$. In this figure the potassium intercalate layers are indicated by dashed lines and the host graphite layers by solid lines connecting open circles, and indicating schematically a projection of the carbon atom positions. For each stage, the distance I_c between adjacent intercalate layers is indicated.

have been various alkali metals (Na, K, Rb, Cs) and alkaline-earth metals (Ca, Ba) (see FULLERENES).

1.2 Requirements for Intercalation

The occurrence of the intercalation process in layered materials generally requires sufficient chemical activity to

1. transfer charge from the intercalate to the adjacent host layers in order to provide the electrostatic attraction between the intercalate and the host layers that is necessary for their mutual bonding, and
2. supply the elastic energy necessary to separate adjacent layers of the host material to accommodate the intercalate.

The insertion of isolated intercalate molecules into a layered host material introduces large local strains. These strains are greatly reduced by the insertion of a second similar intercalate molecule nearby. Thus both elastic forces and normal chemical bonding forces in solids give rise to an attractive in-plane intercalate-intercalate interaction, so that once nucleation of an intercalate layer is initiated, the subsequent growth of the intercalate layer is rapid, exploiting the enhanced in-plane diffusion coefficients of layered materials. In the case of intercalation compounds where the intercalate seeks specific sites (such as intercalated fullerenes) or specific atomic rows (such as certain cuprate compounds), the intercalation-induced change in lattice constant of the host material is quite small. In contrast, the repeat distance of a layered intercalation compound can be much larger than the basic structural unit of the host material, as discussed in Sec. 2.3.

A variety of chemical and structural considerations determine the facility with which a particular intercalate enters a given host material (Hérold, 1979). For example, graphite is a host material that is easily intercalated, forming intercalation compounds with several hundred chemical species. Although nature has synthesized intercalation compounds in the form of clays and minerals over geological time periods, and scientists have synthesized intercalation compounds for over 150 years (Schafhäutl, 1840), it is only recently that methods have been perfected to the point that one can prepare materials with specific structures and properties (Dresselhaus and Dresselhaus, 1981; Solin, 1982; Zabel and Solin, 1990). In the case of clay intercalation compounds, which are also layered materials, intercalation results from ion exchange between a guest ion and an ion in the host material. During this ion exchange process no charge transfer occurs.

A host material with very strong intraplanar binding and very weak interplanar binding provides a necessary condition for intercalation in layered compounds but not a sufficient condition. For example, BN has essentially the same crystal structure and lattice constants as graphite (see Table 1). However, guest species enter into graphite easily but do not enter BN easily. The major difference between the graphite and BN host materials is the semimetallic electronic structure of graphite with a 40-meV band overlap, while BN is an insulator with a band gap of 4.5 eV. Therefore graphite has a relatively high density of states near the Fermi level E_F to facilitate the charge transfer that provides the electrostatic attraction between the intercalate layer and the two adjacent graphite bounding layers. On the other hand, the wide-gap insulator BN does not have available states near E_F to facilitate charge transfer, so that the weak electrostatic interaction may account for the poor intercalation properties of BN host materials.

For a given host material, some species can easily be intercalated and other species cannot. With the exception of the clays where intercalation occurs by ion exchange, a strong electrostatic interaction and a large amount of charge transfer favors intercalation, while the separation of two sequential layers of the host materials to accept the intercalate layer requires the expenditure of elastic energy. The competition between the electrostatic process and the elastic process determines which species will intercalate easily and for which stages the intercalation process is favorable. Thus for the graphite intercalation compounds (GICs) with molecular bromine (denoted by Br_2-GIC), the stage-1 compound is energetically unfavorable while the stability of higher-stage compounds is favorable. As another example, alkali-metal–GICs can usually be prepared for many stages, though the formation of Na-GICs is energetically unfavorable except for high-stage ($n \geqslant 6$) compounds.

Table 1. A comparison between the crystalline and electronic properties of graphite and boron nitride.

Property	Graphite	BN
In-plane lattice constant	2.46 Å	2.49 Å
c-axis lattice constant	3.35 Å	3.33 Å
c-axis layer stacking	*ABAB*	*ABAB*
Electronic structure	Semimetallic	Insulating
Band gap (overlap)	−0.040 eV	4.5 eV

It is experimentally found that transition-metal dichalcogenide and fullerene host materials can be intercalated relatively easily with donor species, but not by acceptor intercalates. The intercalates that can be readily inserted into transition-metal dichalcogenides include metal atoms (such as alkali metals Li, Na, K, Rb, noble metals such as Ag, and polyvalent metals such as Ca, In, Sn, Pb), organic molecules (such as amines, pyridine, ammonia and hydrazine), and 3*d* transition metals (such as V, Cr, Fe, Co, Ni). Clays also permit a large number of species to be inserted into their van der Waals gap.

When an intercalation compound is energetically stable, it can often be prepared by a variety of techniques, such as vapor transport, liquid-phase growth, growth from solution, or electrochemical techniques (Hérold, 1979). Of these various possibilities, electrochemical techniques offer perhaps the greatest number of possibilities for preparing exotic intercalation compounds (Brec, 1987; Besenhard *et al.*, 1983), while vapor transport often results in materials with the greatest staging fidelity (Hérold, 1979).

While energetics determine whether or not an intercalation compound is formed, kinetic considerations are important for determining the time necessary to form a given compound. One of the most intriguing of the kinetic considerations concerns the transition from one stage to another as the intercalation process proceeds. The model due to Daumas and Hérold, which minimizes mass transport (see Fig. 5), provides a simple explanation for a staging transition from stage n to stage $n-1$ (Hérold, 1979), such as is commonly observed for graphite intercalation compounds. *In situ* x-ray (00l) diffraction or *in situ* Raman-scattering studies of the staging transition provide detailed information on the staging kinetics (Suematsu *et al.*, 1988). In such an *in situ* study, the intercalate species in the gas phase is brought into contact with the host material, which is held at constant temperature, and x-ray spectra are taken every few minutes to monitor the sample stage. The results presented in Fig. 6 for the K-GIC system show a stepwise dependence of the stage index on vapor pressure, with little evidence for multiple-stage coexistence.

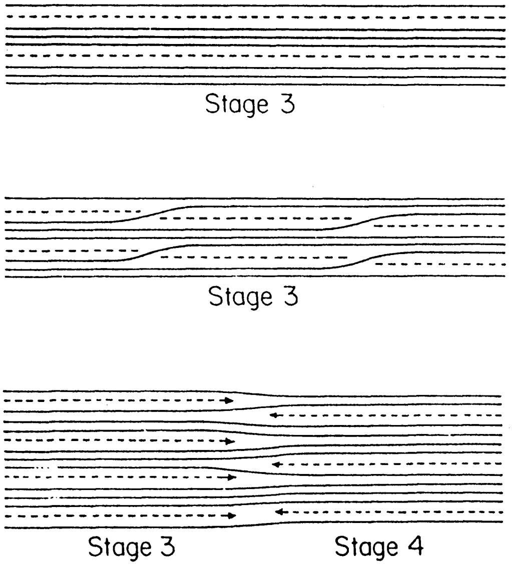

FIG. 5. (a) Single-domain model for a stage-3 compound with every third layer occupied by the intercalate species. (b) Daumas–Hérold domain model for a stage-3 compound with intercalate contained between all graphite planes. (c) Interchange of fourth- and third-stage regions as might occur during a stage transformation.

1.3 Classification According to Rigidity

The layered host materials that can be intercalated are conveniently classified according to their rigidity (Solin, 1984). For the class-I host materials, such as graphite and BN (see Fig. 1), the host material has a basic structural unit that is one monolayer thick (i.e., 3.35 Å for graphite), thereby giving rise to a relatively thin basic structural unit and to relatively flexible materials. A single isolated layer of the graphite structure is called a graphene layer. The basic structural unit of the host material for the transition-metal dichalcogenide intercalation compounds, which are examples of class-II compounds, consists of a trilayer (*XMX*), where *M* denotes the transition-metal layer flanked on either side by a chalcogen layer (see Fig. 2). This trilayer structural unit is intermediate between class I and class III with regard to layer rigidity. Other examples of host materials in the class-II category are the group III–VI (III: Ga, In; VI: S, Se) layered compound semiconductors, where the metal (*M*) and chalcogenide (*X*) layers are arranged in the sequence *XMMX* to form a four-layer slab with strong covalent bonding between each of the four layers and weak van der Waals bonding between these slabs (Julien *et al.*, 1986). Examples of the rigid class-III materials are the silicate clays, where the basic structural unit is typically ~10 Å thick, or thicker when the waters of hydration are included, as shown in Fig. 3 (Wada *et al.*, 1987).

In general, the less rigid the host material, the more likely that it can support the staging phenomenon, which depends also on the competition between the electrostatic and elastic interactions. Of the various host materials, graphite exhibits the highest degree of staging periodicity. Though some transition-metal dichalcogenide intercalation compounds exhibit evidence for staging, this occurs only for low-stage (i.e., $n=1,2$) compounds, and these compounds characteristically have low staging fidelity. Clay intercalation compounds (CICs), which show the largest layer rigidity, do not generally show any evidence for staging, except for stage-1 compounds. On the other hand, clay host materials permit the intercalation of multiple sequential intercalate layers (see Fig. 1), in contrast to the graphite and transition-metal dichalcogenide host materials, which only permit intercalation of a single intercalate unit (either a monolayer or a trilayer unit, as discussed in Sec. 1.4).

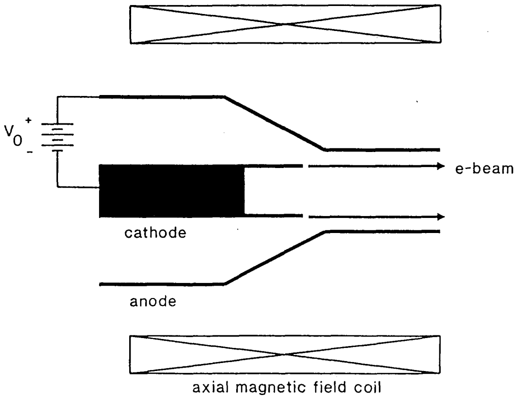

FIG. 6. *In situ* x-ray study of staging transitions (where T_G, the graphite temperature, is 400.1 °C). The kinetics of the staging process in K-GICs is obtained from the time dependence of (00*l*) reflections for transitions from a stage-7 compound to a stage-3 compound (Suematsu *et al.*, 1988).

1.4 Classification of the Intercalate Layers

There are several possible structural arrangements for the intercalate layers. The most common are binary compounds for which the intercalate is a single chemical species, such as an alkali metal, a halogen, or a metal chloride. In some cases, such as the alkali metals or halogens, the intercalate layer consists of a single atomic layer, sandwiched between the basic structural units of the host material. In other cases, such as transition-metal chlorides, the intercalate basic structural unit is a trilayer sandwich. For example, $FeCl_3$ forms a trilayer sandwich of Cl-Fe-Cl constituent layers. Not all compound intercalates form multilayer sandwiches; for example, $AuCl_3$ forms a monolayer intercalate.

A great deal of attention has also been given to the study of ternary intercalation compounds, which are formed by the insertion of a second intercalate species into a binary intercalation compound. Some common examples include the insertion of a second intercalate species into an alkali-metal graphite intercalation compound. Examples of intercalate species that intercalate into alkali-metal GICs include metal intercalates such as Ag and Tl, inorganics such as hydrogen or NH_3, and organic constituents such as benzene (C_6H_6) or tetrahydrofuran (THF) (Zabel and Solin, 1990). In general, the structural arrangement for these ternary intercalation compounds is that of the basic structural unit(s) of the host material, followed by an intercalate trilayer of *MNM* in which *M* refers to an alkali-metal layer and *N* refers to the new ternary constituent, subsequently added to the original binary intercalation compound. The classification of the ternary compounds is complicated by three issues. First, some of the ternary compounds, such as potassium hydride graphite intercalation compounds (KH-GICs), can also be prepared by direct intercalation; independent of whether KH-GICs are prepared sequentially or directly, they are classified as ternary compounds. (Here the intercalate contains two species, potassium and hydrogen.) Second, two-component metal alloy intercalates are said to form ternary intercalation compounds, even though the intercalate is inserted into the host material directly, and the intercalate occupies a single atomic layer as a random-alloy monolayer. The third complication in the ternary classification arises from the possibility of preparing bi-intercalation compounds where the two intercalate constituents are located in different van der Waals gaps or intercalate galleries. Further examples of the various types of binary and ternary systems are given later in this section.

For either binary or ternary intercalation compounds, the separation of two host layers d_s between which the intercalate is sandwiched can vary over wide ranges. In general d_s depends on the size of the guest species. The wide variation of d_s is most clearly seen in the GICs. One example of a GIC where d_s deviates only slightly from that of the unintercalated host material is the case of the Li donor intercalate in graphite where $d_s = 3.70$ Å and $\tilde{c}_0 = 3.35$ Å is the thickness of the basic structural unit of a single graphene sheet. In the opposite limit, d_s can be greater than $3\tilde{c}_0$ for trilayered intercalates into graphite such as transition-metal chlorides (see Table 2) (Dresselhaus and Dresselhaus, 1981; Solin, 1982). Ternary compounds such as KHg-GICs tend to have large d_s values as indicated in Table 2.

The separation between specific intercalate layers can be further increased by the synthe-

Table 2. Intercalate sandwich thickness d_s for several graphite intercalation compounds (GICs) showing that the repeat distance I_c for a low-stage GIC can be varied over a wide range by proper choice of intercalate.

Intercalate	d_s (Å)
Li	3.706
Eu	4.87
K	5.35
Rb	5.65
Cs	5.94
Br_2	7.04
HNO_3	7.84
AsF_5	8.15
SbF_5	8.46
$MnCl_2$	9.30
$NiCl_2$	9.30
$MoCl_5$	9.31
$FeCl_3$	9.37
$SbCl_5$	9.42
$CrCl_3$	9.45
$MnCl_2$	9.48
$CoCl_2$	9.50
$FeCl_2$	9.51
$AlCl_3$	9.54
KHg	10.22

sis of bi-intercalation compounds (Zabel and Solin, 1990). Bi-intercalation compounds contain two different intercalants that assume special ordered arrangements. Layered bi-intercalation compounds with intercalants M_1 and M_2 can be prepared, forming the layered sequence $M_2CM_1CM_2$ where C denotes carbon (a single graphene sheet). One synthesis route for such bi-intercalation compounds can be the intercalation of species M_2 into a stage-2 M_1-GIC. For example, in a bi-intercalation metal chloride compound, the separation between sequential M_1 layers is close to 20 Å. Such compounds are of particular interest for the synthesis of quasi-2D magnetic systems, and in this connection the compound $CoCl_2$-$GaCl_3$-GIC has been synthesized and studied magnetically (see Sec. 4).

More generally, ternary intercalation compounds provide a mechanism whereby more than one intercalate species can be introduced into a host material (Solin and Zabel, 1988). As mentioned above, aside from the bi-intercalation compounds, there are basically two different categories of ternary compounds: intercalates introduced as alloys, and intercalates introduced sequentially. An example of an intercalate alloy system is K_xRb_{1-x}, which intercalates into graphite as a substitutional random alloy on each intercalate monolayer (Solin and Zabel, 1988). Such substitutional alloy systems offer very interesting possibilities for studying the physics of quasi two-dimensional random alloys. Intercalate alloys can also be introduced into nonlayered host materials, such as fullerenes. For the case where the two intercalate species have nearly the same size, the two intercalate species enter the fullerene lattice randomly in either octahedral or tetrahedral sites, while for cases like Na_xCs_{1-x}, where the two species have very different radii, the Na ions tend to favor the smaller tetrahedral sites and the larger Cs ions strongly favor the octahedral sites, so that an ordered compound of stoichiometry Na_2CsC_{60} can be prepared.

The more widely studied classes of ternary compounds comprise the sequentially intercalated compounds (Solin and Zabel, 1988), which allow the intercalation of species M_2 that could not itself be intercalated into the host material. In such cases, the prior intercalation of species M_1, often an alkali metal, permits subsequent intercalation of species M_2. Examples of such ternary compounds formed with M_1 as an alkali-metal species include as the M_2 species tetrahydrofuran (THF), an organic species; NH_3, an inorganic species; and the metallic species Hg and Tl, which form superconducting ternary GICs (see Sec. 5). In the case of the hydrogen ternary compounds, both chemisorbed and physisorbed compounds are possible, where the structure of the chemisorbed compound is similar to that for the ionic KHg-GICs (see Fig. 7), while the physisorbed compound can be prepared from stage-2 K-GICs and is important for high-density hydrogen storage applications. Because of the large differences in the amount of physisorption of the various isotopic hydrogen species (tritium $> D_2 > H_2$), the physisorbed hydrogenated K-GICs offer important possibilities for isotope separation applications (Terai and Takahashi, 1981).

In the above discussion of the intercalate layer, it was assumed that the guest species in the intercalate layer is the same as the guest species prior to intercalation. This is not always the case. For compounds such as the $SbCl_5$-GICs, the intercalation process results in a chemical reaction called disproportionation, giving rise to different species in the intercalate layer, as, for example, the species $SbCl_3$, $SbCl_6^-$, $SbCl_5$, and $SbCl_4^-$ relative to the parent intercalate species $SbCl_5$ prior to intercalation (Zabel and Solin, 1990).

2. STRUCTURAL CHARACTERIZATION

In Sec. 1, the general structural organization and classification of intercalation compounds was summarized. In this section, the structural ordering along the *c* axis normal to the layer planes and in the *a-b* layer planes is reviewed. Interlayer correlations and phase transitions in intercalation compounds are also discussed.

2.1 *c*-Axis Ordering

The dominant ordering along the *c* axis is the staging periodicity, discussed in Sec. 1.2, and characterized by the stage index *n* and the staging fidelity, denoted by the average length of the stage-*n* material (both in-plane and along the *c* axis), and by the amount of each minority stage *n′*, expressed as a percentage of the sample volume. The simplest method

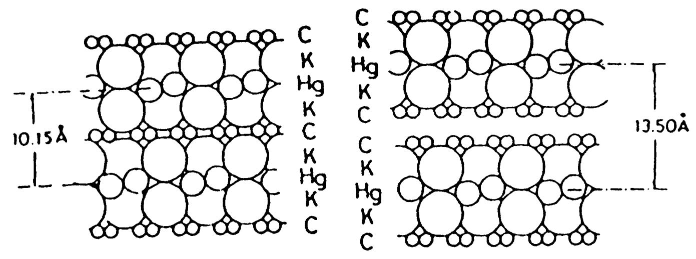

(a) stage 1 C_4KHg

(b) stage 2 C_8KHg

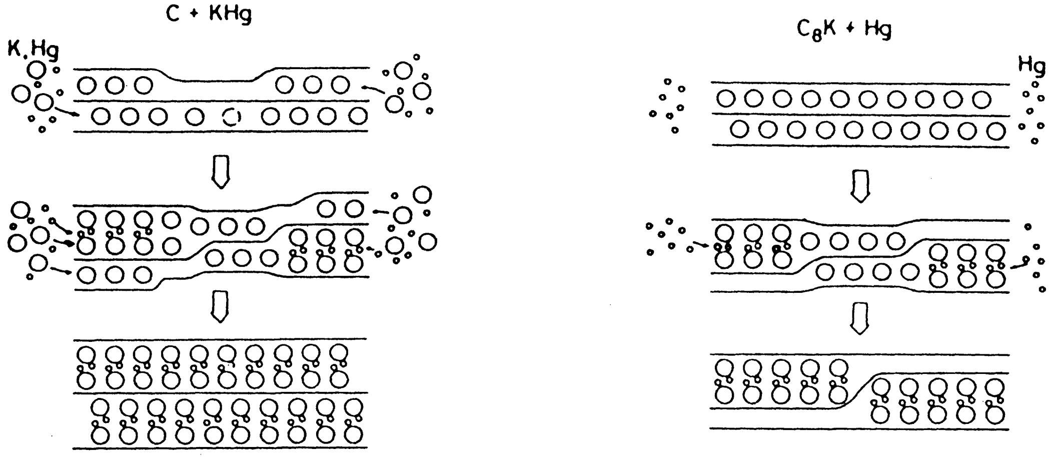

(c) stage 1 C_4KHg

(d) stage 2 C_8KHg

FIG. 7. (a),(b) The layer stacking order of the ternary KHg-GICs for stages 1 (C_4KHg) and 2 (C_8KHg), respectively, where the ideal chemical formulas are also given. (c),(d) A schematic diagram for the reaction process for stage-1 and stage-2 compounds, respectively.

for determining the stage index and information on the staging fidelity of an intercalation compound is the measurement of ($00l$) x-ray diffraction patterns. High-resolution transmission electron microscopy (TEM) studies of lattice fringe images provide a powerful technique for showing detailed aspects of the ordered lattice planes. Since the intercalate often has elements of higher atomic weight than the host material (e.g., carbon), the intercalate layer scatters the TEM electrons more effectively, thereby giving clear contrast between the intercalate and the light host materials. Such measurements show that many host materials exhibit some short-range staging correlations. For a few hosts, especially those with low rigidity (see Sec. 1.4), long-range staging is observed. For the case of graphite intercalation compounds, high-resolution TEM lattice fringe images show that a high degree of staging perfection (over a scale of ~100 Å) can be achieved under favorable circumstances (Zabel and Solin, 1990). For GICs, the c-axis repeat distance I_c between sequential intercalate layers follows the simple relation $I_c = d_s + (n-1)\tilde{c}_0$, where $\tilde{c}_0$ is the interlayer separation in pristine graphite, or the thickness of a single graphene sheet. The validity of this simple relation indicates that, to first order, GICs can be considered as a linear superposition of graphite and intercalate layers. Similar results are obtained in some cases for other host materials. Graphite intercalation compounds are, however, of particular interest with regard to staging because they can be synthesized with stages up to $n \sim 10$ and because of their relatively high degree of staging fidelity (Hérold, 1979; Zabel and Solin, 1990).

TEM studies also provide information about the coupling between the host material and intercalate at their interface. Some compounds such as $CuCl_2$-GICs show sharp interfaces with weak coupling between the intercalate and the graphite host, while K-GICs and F-GICs show very wavy fringes, indicating a strong interlayer coupling. In the case of F-GICs the coupling is so strong that the intercalate layers become puckered as covalent bonding is initiated between the fluorine intercalate and the graphite host material.

The basic structural units of the transition-metal dichalcogenide crystals TX_2 (see Fig. 2) also show *c*-axis ordering both before and after intercalation. The arrangements of atoms within a sheet and the stacking of the atomic TX_2 sheets are now described. Each sheet consists of atoms in a hexagonally close-packed network as shown in Fig. 8(a), which also identifies the three inequivalent close-packed sites *A*, *B*, and *C*. Two types of coordination structure are possible, and either one or both can form the basic structural units of the TX_2 crystal. In the *trigonal prismatic coordination*, the sandwich of *X-T-X* follows the ...*AbA*... sequence, as in Fig. 8(b), and in the octahedral coordination, the ...*AbC*... sequence, as in Fig. 8(c). Here capital letters are used to denote the chalcogen positions and lower-case letters for the metal atoms. In an ideal structure, the distance between chalcogen atoms across the sandwich is the same as that in the plane, which is also equal to the *a* and *b* lattice constants as well as the distance between metal atoms (in plane).

In addition to the two *intrasandwich* octahedral and trigonal prismatic coordination arrangements, there are many variations for *intersandwich* packings, and a large number of polytypes are known. The six most common polytypes for transition-metal chalcogenides are shown in Fig. 9. The nomenclature that identifies these polytypes specifies the number of trilayer sandwiches contained in the unit cell perpendicular to the plane of the layers, and the overall symmetry (trigonal, hexagonal, or rhombohedral) of the structure. Lower-case subscripts are sometimes further required to distinguish polytypes otherwise similarly labeled (e.g., $2H_a$ and $2H_b$). Thus the simplest polytype with octahedral coordination, labeled $1T$, has trigonal symmetry and a repeat distance, perpendicular to the layers, of one trilayer sandwich. The simplest trigonal prismatic coordination polytypes contain two trilayer sandwiches, and are designated $2H_a$ and $2H_b$ (see Fig. 9). The $2H_a$ structure is adopted by metallic group-V materials such as NbS_2 and stacks the metal atoms directly above each other along the *c* axis, while in the $2H_b$ structure, adopted by semiconducting group-VI materials such as MoS_2, the metal atoms are staggered. Octahedral interstitial sites between sandwiches are also shown in Fig. 9 in parentheses, and are the vacant sites into which intercalate atoms or molecules may enter. The crystallographic axes for the *T* and *H* structures are labeled by

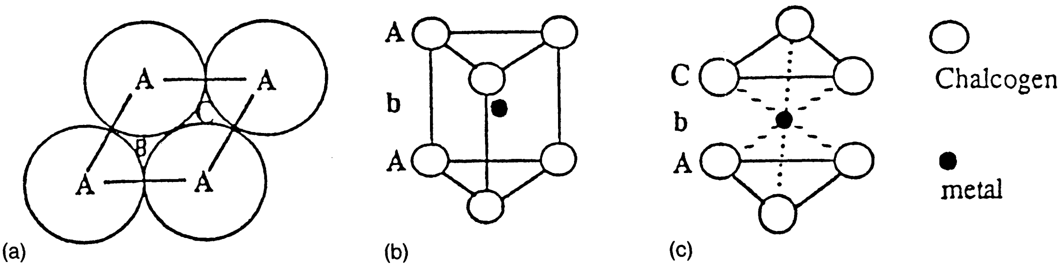

FIG. 8. The octahedral and trigonal prismatic structures of transition-metal dichalcogenides, TX_2. (a) *A*, *B*, and *C* represent the three nonequivalent positions for close-packed stacking; (b) the trigonal prismatic structure; and (c) the octahedral stacking arrangement. The capital letters designate chalcogen positions and the lower-case letters the metal positions.

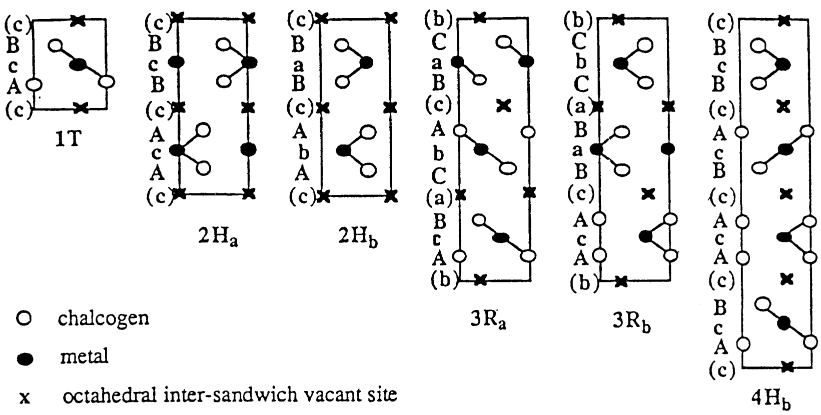
(c) B c A (c)
1T
(c) B c B (c) A c A (c)
2Ha
(c) B a B (c) A b A (c)
2Hb
(b) C a B (c) A b C (a) B c A (b)
3Ra
(b) C b C (a) B a B (c) A c A (b)
3Rb
(c) B c B (c) A c B (c) A c A (c) B c A (c)
4Hb
O chalcogen
● metal
x octahedral inter-sandwich vacant site

FIG. 9. Unit cells of simple polytypes shown with respect to their *c*-axis arrangement. The capital letters designate chalcogenide positions and the lower-case letters give the metal positions. Vacant octahedral sites for intercalation are indicated.

a(*x*,*y*) parallel and *c*(*z*) perpendicular to the layers. Typical values are 3–4 Å for the *a* lattice parameter, and 6 Å for the sandwich repeat distance along the *c* axis.

In addition to the octahedral interstitial sites shown in Fig. 9, there are also tetrahedrally coordinated intersandwich sites that lie immediately below and above the chalcogen atoms and are displaced from the symmetric plane between the trilayer sandwiches. These are shown as the "a" and "b" sites in Fig. 10. Octahedral interstitial sites are usually favored in the intercalation of alkali metals, Li, Na, etc., and the 3*d* transition metals, and the maximum intercalate uptake *M* gives a limiting stoichiometry of MTX_2. Tetrahedral site occupation is found in intercalation complexes formed with silver and copper, and with hydrazine; in this case the limiting stoichiometry is M_2TX_2. Li ions in VSe_2 can occupy both octahedral sites, forming $LiVSe_2$, and tetrahedral sites, as in Li_2VSe_2. Transition-metal dichalcogenide hosts are especially interesting for the possibility of introducing very large spacers between the TX_2 trilayers, as, for example, the insertion of *n*-octadecylamine into $TaSe_2$ leading to a separation of nearly 60 Å between sequential Ta layers in the layered structure of the intercalation compound.

2.2 In-Plane Ordering

Not only can intercalation compounds show long-range staging fidelity with regard to their *c*-axis ordering (1D superlattice), but intercalation compounds also show various in-plane orderings, which are of interest in their own right and bear importantly on the observed properties of this class of materials. The two major classes of in-plane structures are the commensurate and the incommensurate arrangements of the intercalate layers with respect to the adjacent host layers. In the commensurate structures (see Fig. 11), the intercalate atoms are in registry with the adjacent host, forming intercalate unit cells with areas that can be multiples of the areas of

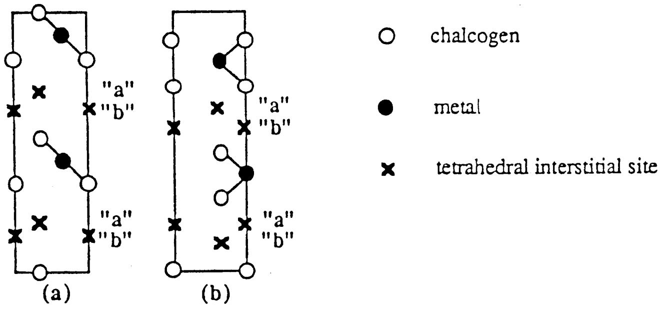

FIG. 10. Tetrahedral intersandwich sites *x* in the (a) 1*T* and (b) $2H_b$ TX_2 polytype structures.

the 2D unit cell of the host material. The following are common examples of commensurate structures that occur in GICs: $(\sqrt{3}\times\sqrt{3})R30°$, $(2\times 2)R0°$, $(\sqrt{7}\times\sqrt{7})R19.1°$, and $(\sqrt{3}\times 2)R(30°,0°)$, where the notation lists the lengths and angles of the intercalate basis vectors relative to those for the graphite lattice, and angular redundancies are not repeated. Commensurate structures are found in a wide variety of model GIC systems (see Table 3), including stage-1 binary donor compounds, such as C_6Li, C_6Eu, C_8K; coplanar molecular acceptor compounds, such as $C_{7n}Br_2$ for stages $n \geqslant 2$; triple-layer sandwich acceptors, such as the $SbCl_5$-GICs; and triple-layer sandwich donor ternary systems, such as KHg-GICs and KH-GICs (Zabel and Solin, 1990).

The formation of commensurate in-plane structures has a close analogy to the growth of semiconductor strained-layer superlattices by the molecular beam epitaxy (*q.v.*) (MBE) method or by metal-organochemical vapor deposition [see CVD (CHEMICAL VAPOR DEPOSITION)] (Dresselhaus, 1987). Because the intercalate layer in an intercalation compound is so thin (a single monolayer or a single trilayer sandwich), commensurate in-plane structures can be formed without the need for misfit dislocations to relieve local strains.

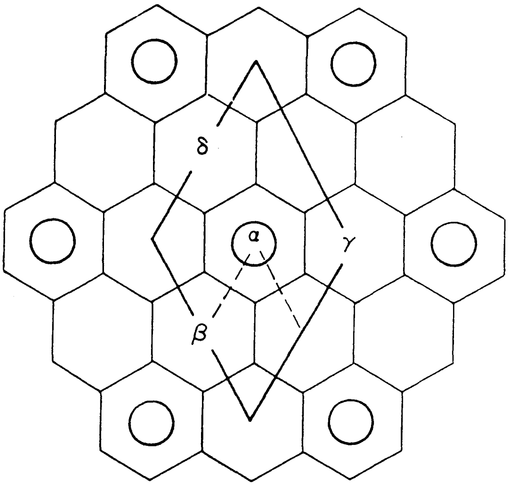

FIG. 11. Schematic diagram of an in-plane unit cell for the commensurate $(2\times 2)R0°$ structure. Carbon atoms from the host material are located at the corners of the hexagons on one layer and the intercalate species are shown as the large open circles, which are located on the adjacent layer. In this diagram the intercalate and graphite layers are projected onto a single plane. For this structure there are four equivalent intercalate sites, labeled α, β, γ, δ. On a given intercalate layer only one of these sites is occupied. The dashed lines represent the in-plane unit cell for pristine graphite. For the first-stage C_8K and C_8Rb, the intercalate on consecutive layers is arranged on sequential α, β, γ, δ sites.

Table 3. Structure of common commensurate GICs.

Intercalate	Stage	Formula[a]	Structure[b]
Li	1	C_6Li	$(\sqrt{3}\times\sqrt{3})R30°$
Eu	1	C_6Eu	$(\sqrt{3}\times\sqrt{3})R30°$
K	1	C_8K	$(2\times 2)R0°$
Br_2	$\geqslant 2$	$C_{7n}Br_2$	$(7\times\sqrt{3})R(0°,30°)$
$SbCl_5$	$\geqslant 1$	$C_{14n}SbCl_5$	$(\sqrt{7}\times\sqrt{7})R19.1°$
KHg	1	C_4KHg	$(2\times 2)R0°$
KH	1	C_4KH	$(2\times 2)\mathrm{R}0°$

[a] The nominal chemical formula for a stage-n compound.

[b] The notation for the 2D superlattice refers to the length scaling and the angle of rotation of the unit vectors for the intercalated superlattice relative to graphite.

The lowest energy state for an intercalate atom is obtained by placing it over (or below) the center of a graphite hexagon, thereby achieving the greatest packing density (Dresselhaus and Dresselhaus, 1981; Hérold, 1979). A close-packed in-plane configuration for the intercalate maximizes its chemical bonding and minimizes the strain energy introduced into the adjacent graphite layers from insertion of the intercalate. When both registry with the graphite hexagons and close packing of the intercalate layers can be simultaneously satisfied on a long-range basis, commensurate intercalate layers are favored. Otherwise, the close-packed configuration of the intercalate layer implied by chemical binding considerations prevails and an incommensurate solid forms. In many cases, the in-plane intercalate lattice constant for incommensurate interlayer stacking is close to that of the parent material. Examples of intercalates forming incommensurate GICs are transition-metal chlorides and bromides (Vogel and Hérold, 1977; Hérold *et al.*, 1983), some of which have especially interesting magnetic properties, as discussed in Sec. 4.

Intercalate registration (the tendency for an intercalate atom to be located in a position of minimum energy with respect to the adjacent graphite layers) affects the structural ordering of GICs in many ways besides the formation of commensurate GICs, as discussed above. For the case of incommensurate GICs, local registration of the intercalate with respect to the graphite layers leads to *orientational* alignment of the intercalate, as evidenced by the spot patterns observed by (*hk*0) diffraction spectra for many incommensurate GICs. In addition, many incommensurate intercalates (e.g., transition-metal chlorides) tend to form intercalate islands (~100 Å in diameter) to minimize the number of intercalate atoms not in registry with the graphite host material (Dresselhaus, 1987).

2.3 Interlayer Correlations

Not only is there intercalate ordering within the intercalate layer, but there is also 3D stacking order of intercalate layers with respect to one another. A clear example of this phenomenon is observed in stage-1 alkali-metal–GICs (Dresselhaus and Dresselhaus, 1981). Because the intercalate unit cells can be much larger than that for the host, there will be several equivalent sites for the intercalate atoms in a commensurate GIC. For example, for a $(2\times2)R0^\circ$ in-plane lattice (see Fig. 11), there will be four equivalent intercalate sites, labeled α, β, γ, δ. But once one of these sites is occupied by an intercalate atom (e.g., an α site), then the intercalate atoms on the same layer (within a single domain) will also tend to occupy the same type of site (an α site), thereby giving the maximum packing density for the intercalate (Dresselhaus and Dresselhaus, 1981). Because of the local strain introduced by placing an intercalate atom on site α, the intercalate atom on the adjacent intercalate layers will be less likely to be located on an α site than on a non-α site. This interlayer exclusion argument implies that when there is sufficient intercalate-intercalate interlayer correlation mediated by the intervening graphite layers, interlayer stacking order will result. A number of low-stage binary GICs and ternary GICs have been reported to exhibit interlayer stacking correlation effects (Solin and Zabel, 1988). Even when there is no long-range interplanar stacking order, there is often short-range interplanar correlation based on the interplanar intercalate exclusion argument given above. Because the interplanar intercalate-intercalate interaction energy is small, thermal excitation at room temperature is often sufficient to destroy interplanar intercalate correlation, thereby producing quasi-2D intercalate layers, uncorrelated in the *c* direction.

2.4 Intercalate Phase Transitions

In general, the host material and intercalate materials have different melting points. Thus unusual phase transitions can be observed, as one of the constituents melts while the other remains solid. Such behavior is, for example, observed in GICs, since graphite does not undergo structural phase transitions up to very high temperatures (~4000°C). Examples of such phase transitions are order–disorder transitions in the intercalate layer, as occur in the HNO_3-GIC system (Dresselhaus and Dresselhaus, 1981). The disordered intercalate phase often exhibits the characteristics of a lattice gas, whereby each intercalate atom is fluidlike in the intercalate layer, but has a high probability of being in registry with the adjacent graphite layers. Such effects are observed in high-resolution x-ray diffraction measurements (Zabel and Solin, 1990) of high-stage ($n>2$) alkali-metal GICs. Intercalate melting is a closely related phenomenon and has been observed for a variety of intercalates (Dresselhaus and Dresselhaus, 1981).

For some GICs, two or more intercalate orderings have approximately the same free energy, leading to multiphase coexistence, with domain walls that are either more dense (heavy walls) or less dense (light walls) than the ordered domains. An ordered array of domain walls may occur, giving rise to modulated structures, as occur in many high-stage ($n>2$) alkali-metal GICs (Zabel and Solin, 1990). Heavy and light walls also arise at the boundary between regions of different site occupation in a commensurate in-plane superlattice [e.g., the boundary between an α domain and a β domain in the same intercalate layer of a $(\sqrt{3}\times\sqrt{3})R30^\circ$ superlattice]. These phenomena have been widely studied for adsorbed rare-gas phases on graphite (Zabel and Solin, 1990).

A particularly interesting example of a modulated in-plane structure in GICs is the stripe domain phase (Zabel and Solin, 1990), which has been observed in a stage-4 Br_2-GIC with nominal stoichiometry $C_{28}Br_2$. This intercalation compound has a commensurate $(7\times\sqrt{3})R(0^\circ,30^\circ)$ in-plane structure with

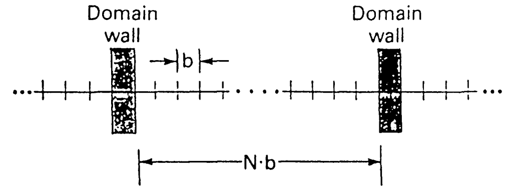

FIG. 12. Schematic diagram of a stripe domain phase. The intercalate lattice remains commensurate except in the domain wall where a phase slip occurs every *N* unit cells.

a rectangular unit cell containing four Br_2 molecules. The driving mechanism for this phase transition is the differential thermal expansion between the graphite basal plane, which has essentially zero thermal expansion, and that for a bromine layer, which has a normal thermal expansion coefficient. For the Br_2-GIC system, the coupling between the graphite and intercalate layers is sufficiently large so that the structure remains commensurate over a large temperature range. At a temperature $T_c=69$ °C, a phase transition occurs whereby the graphite and bromine layers retain long-range registry along the $\sqrt{3}$ direction, but form a stripe domain phase along the seven-fold direction. As the temperature increases above T_c, the average bromine lattice constant along the seven-fold direction increases, though the graphite and bromine remain locally commensurate over a domain containing N unit cells (see Fig. 12). These domains are arranged in a periodic array with lattice constant $Nb+\tau$, where τ is the thickness of the domain wall, and $\tau=2b/7$ where $b=7a_0$ and a_0 is the graphite in-plane lattice constant. The domain wall is also called a misfit dislocation in the terminology of the strained-layer superlattice. This stripe domain phase is one example of a more general discommensuration phenomenon, which features local commensurate behavior but long-range incommensurate behavior. Discommensuration effects are also observed in high-stage alkali-metal–GICs (Zabel and Solin, 1990).

3. ELECTRONIC PROPERTIES

In accordance with the structural ordering discussed above is the periodicity in the charge distribution perpendicular to the layer planes, which arises from the charge transfer associated with the intercalation process. In general, charge transfer is the driving mechanism for intercalation and is responsible for the observed transport and electronic properties of intercalation compounds. In this section the concept of charge transfer is discussed and the transport and electronic properties that result from charge transfer are also reviewed.

3.1 Charge-Transfer Phenomena

To illustrate the charge-transfer concept in a general intercalation compound, let us consider the charge transfer from an alkali metal to a host material such as graphite in a graphite intercalation compound (see Fig. 13). Let us suppose that each alkali-metal atom contains one nearly free electron. If this electron is donated to the graphite layers, the alkali-metal layers become positively charged, thereby attracting the negative charges in the graphite layers. Thus the majority of the electrons transferred to the graphite will be located in the graphite bounding layers adjacent to the intercalate layers, as shown in Fig. 13. Lower concentrations of transferred charge will be available for the graphite interior layers that are surrounded only by graphite layers (Dresselhaus, 1984). Calculations of the charge distribution in the graphite layers

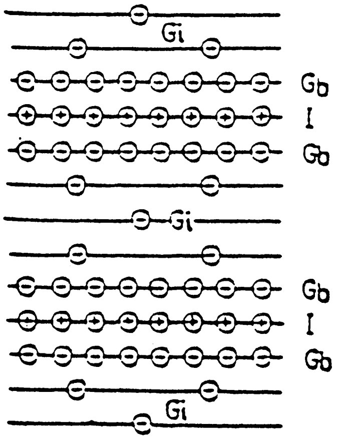

FIG. 13. Schematic representation of charge transfer in graphite intercalation compounds, showing the intercalate layer (*I*) for a donor compound transferring electrons to the graphite bounding (G_b) layer, with a small amount of charge also transferred to the graphite interior (G_i) layers.

show that the graphite bounding layers effectively screen the graphite interior layers from the intercalate layers with a power-law dependence of z^{-4} relative to the position of the intercalate layer. The charge-transfer concept for alkali-metal intercalation into other host materials is similar.

Since the carrier concentration in graphite is very low ($\sim 2\times 10^{-4}$ carrier/C atom at room temperature and $\sim 6\times 10^{-5}$ carrier/C atom at low temperature), the charge transferred to the graphite layers through intercalation causes a major increase in the mobile carrier concentration in the graphite layers. An estimate of three orders of magnitude increase in carrier concentration follows from assuming a charge transfer of one carrier per intercalate unit to the graphite layers and assuming an intercalate concentration one order of magnitude smaller than that for graphite. Since the mobility of pristine graphite is relatively high (13 000 cm^2/V s at 300 K), intercalation compounds with very high in-plane electrical conductivity can be synthesized. On the other hand, the conductivity in the intercalate layers is expected to be low because of the low mobility of ions generally, and the low density and low mobility of electrons (or holes) in the intercalate layers. Thus electrical transport in the basal planes of GICs is mainly by graphitic π-band electrons and holes. Intercalation with donors raises the Fermi level so that conduction is predominantly by electrons, while intercalation with acceptors lowers the Fermi level, giving rise to conduction by holes (Dresselhaus and Dresselhaus, 1981).

Because of the high electron affinity of the transition-metal dichalcogenides and fullerenes, these materials lead only to donor intercalation compounds. Charge transfer in the donor transition-metal dichalcogenide compounds is especially interesting, since TX_2 compounds containing a group-IV transition metal, such as ZrS_2, are semiconducting, so that, for example, intercalation with an alkali metal can partially fill the d_{z^2} level, making the material conducting. On the other hand, if the transition metal T is a group-V element such as Nb, the TX_2 compound is metallic; however, if sufficient charge is transferred, a filling of the d_{z^2} band may occur, resulting in a semiconductor. These various possibilities are considered in Fig. 14.

The current that flows during the electrochemical intercalation process provides one direct measure of the charge transfer in intercalation compounds. In principle, magneto-oscillatory Fermi-surface experiments should also yield accurate values of the charge transfer. This has been demonstrated by very careful magneto-oscillatory measurements made on several well-characterized alkali-metal–GICs and a very few acceptor GIC materials. Because of the difficulty of these experiments, it is not likely with present techniques that Fermi-surface studies will become a widespread characterization tool for charge transfer in GICs.

Charge transfer in ternary GICs is of particular interest. For example, the addition of Hg or H to the first-stage compound C_8K reduces the electron concentration in the graphite layers by a back-donation process, while the Hg or H layers become negatively charged. This can be more easily understood by referring to Fig. 7 where the trilayer structure of the intercalate sandwich for a ternary compound (such as KHg-GIC) is shown. Here the Hg or H layers behave somewhat like an acceptor intercalate, though residing in an otherwise donor compound. It is believed that the Hg conduction band gains electrons but remains partially occupied in order to support

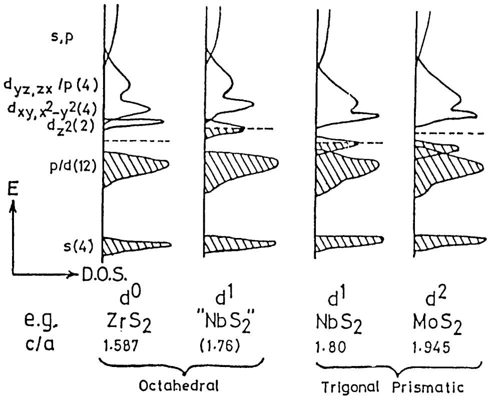

FIG. 14. Schematic density-of-states diagrams for groups IV, V, and VI transition-metal dichalcogenides. The numbers in parentheses following the atomic orbital notation indicate the number of states per formula unit that can be accommodated in each band. The compounds used as examples are taken from metals of the same row of the periodic table. NbS_2 does not actually crystallize in the octahedral form, and the second diagram represents an extrapolated hypothetical band diagram (Liang, 1987).

superconductivity in the KHg ternary compounds as discussed in Sec. 5.

3.2 Transport Properties

Because of the large increase in carrier concentration that often accompanies intercalation, large increases in the in-plane conductivity σ_a is often observed in intercalation compounds. For semiconducting group-IV or group-VI transition-metal dichalcogenides, the change in conductivity can be many orders of magnitude, since the material prior to intercalation has such a low electrical conductivity. For the GIC host material, the room-temperature in-plane conductivity can with suitable intercalates (e.g., AsF_5) become comparable to that of the best room-temperature conductors. Because of this potential for practical applications, most transport studies on intercalation compounds have been done on GICs. For most GICs, in-plane conductivity increases approaching an order of magnitude are frequently observed. Some exceptions to this general behavior are found in high-concentration F-GICs, where the intercalation process itself leads to disorder, the formation of covalent bonding, and a reduction in σ_a.

Not only is the electrical conductivity of layered intercalation compounds very high in the basal planes, but it is also highly anisotropic. The most anisotropic intercalation compounds are the GICs, a result in part of the anisotropy of the host material. The conductivity of pristine graphite in the *c* direction is between 3 and 4 orders of magnitude lower than that in the basal plane, depending on the temperature (Dresselhaus, 1987). Since intercalation with acceptors tends to introduce intercalate layers containing molecular species with very low electrical conductivity, the electrical anisotropy of graphite is increased by intercalation with acceptors, reaching values of σ_a/σ_c in the range 10^6–10^7 (McRae and Marêché, 1988). On the other hand, intercalation with alkali-metal donors often hybridizes the graphitic π electrons with intercalate electrons containing *s* and *p* character, thereby reducing the electrical anisotropy relative to pristine graphite, with anisotropy ratios as small as 10 for the case of stage-1 C_6Li (Dresselhaus and Dresselhaus, 1981). However, the ternary donor compounds (e.g., KHg-GICs and CsBi-GICs), show anisotropy ratios σ_a/σ_c comparable to the magnitudes previously reported for typical acceptor GICs (McRae and Marêché, 1988).

3.3 Electronic Structure

An extensive amount of work has been done in calculating the electronic structure of the various host materials supporting intercalation, such as graphite and the transition-metal dichalcogenides, and in some cases the electronic structure of the intercalation compounds has also been calculated (Zabel and Solin, 1992). From an experimental standpoint, photoemission has been extensively used for determinations of the electronic structure.

Although surface science techniques such as photoemission can, in principle, yield a great deal of valuable information about intercalation compounds, the near-surface sensitivity of these techniques causes them to emphasize the surface electronic structure rather than that typical of the bulk. This is a significant problem for intercalation compounds because of the large sizes of their *c*-axis unit cells. In addition, surface-science studies have thus far shown that the first few atomic layers of donor GICs have an excess intercalate concentration, while in acceptor GICs, the first few atomic layers exhibit a very low intercalate concentration (Dresselhaus, 1987). The cleavage of GICs at low temperature (e.g., ~77 K) has demonstrated the possibility of preparing surfaces retaining intercalate concentrations typical of the bulk for long enough times (minutes) to permit properties measurements to be carried out. Optical measurements, which probe many unit cells along the *c* axis, have also been widely applied to study interband transitions and phonon modes in various intercalation compounds. Because of the high degree of anisotropy in the structural properties, the electronic and optical properties are also highly anisotropic.

4. MAGNETIC INTERCALATION COMPOUNDS

A variety of magnetic materials has been prepared by the intercalation of magnetic species into various nonmagnetic host materials. In general, the intercalation of magnetic species into a host material results in a magnetically ordered intercalation compound below a magnetic ordering temperature T_c. By varying the separation between sequential magnetic layers through the staging mechanism, the interplanar magnetic coupling can be reduced in a controlled manner to very small values, thereby providing a convenient system for studying quasi-2D magnetism. Graphite intercalation compounds are of particular interest in the case of magnetic intercalation compounds because they provide novel systems for the study of quasi-2D magnetic behavior. The transition-metal dichalcogenides and clays also allow for the intercalation of layers of magnetic species in a nonmagnetic host material. The transition-metal dichalcogenides offer further possibilities insofar as by proper selection of the transition metal, the host material can itself be magnetically ordered; in this case, the intercalate can be either nonmagnetic or a different kind of magnetic layer. Some of the most interesting magnetic studies of 2D systems have so far been done on the graphite intercalation compounds because of their good staging characteristics.

Both donor (Eu) and acceptor ($FeCl_3$, $FeCl_2$, $CoCl_2$, $NiCl_2$, $MnCl_2$, $MoCl_5$) magnetic GICs have been synthesized and studied for their magnetic properties (Zabel and Solin, 1992). The magnetic ordering temperatures of the magnetic intercalation compounds are generally lower than those for their parent magnetic materials because of the smaller number of near neighbors imposed by the layer arrangement of the intercalation compounds. For the rare-earth donor GICs, the intercalate in-plane structure is commensurate with that of the adjacent graphite layers, forming a $(\sqrt{3}\times\sqrt{3})R30°$ superlattice structure, and thereby simplifying the treatment of the magnetic problem to some degree. Since the in-plane structures of incommensurate metal chloride acceptor GICs tend to be almost identical to those in the corresponding parent materials, the intraplanar magnetic exchange constants remain essentially unchanged upon intercalation, thereby providing a major simplification in the treatment of the magnetic problem in metal chloride acceptor GICs.

Of the various types of magnetic acceptor GICs, the quasi-2D XY systems (for which the spins are constrained to be in the layer planes) have been most widely studied, because of the special role these systems play in theoretical studies of critical phenomena. The absence of long-range order in 2D systems results in short-range correlations in the form of vortices. Close approximations to such magnetic phases can be observed in magnetic GICs over a limited range of temperatures and magnetic fields (Zabel and Solin, 1992).

Of the various magnetic acceptor GICs, the $CoCl_2$-GICs exhibit the largest *c*-axis magnetic anisotropy ($J_A/J=0.56$), so that these compounds are expected to provide the best approximation to a 2D XY model for GIC systems. Furthermore, it is possible to synthesize $CoCl_2$-GICs for a variety of stages, and thus to investigate the stage dependence of the observed phenomena. Moreover, the structural coherence lengths in the intercalate islands are large enough to contain a statistically meaningful number of spins ($>10^3$ spins/island), and yet small enough to permit a meaningful comparison with theoretical calculations for specific spin arrangements.

For all the acceptor 2D XY GIC systems that have been studied, two critical temperatures are identified (T_{cl} and T_{cu}), thereby delineating three temperature regimes, each exhibiting distinct magnetic behaviors: $T<T_{cl}$, $T_{cl}<T<T_{cu}$, and $T>T_{cu}$. At high temperature ($T\gg T_{cu}$), Curie–Weiss paramagnetic behavior is found, while at low temperature magnetically correlated states are observed (Zabel and Solin, 1992).

In the temperature range $T_{cl}<T<T_{cu}$ for the high-stage compounds, the spins are approximately confined to the basal plane and are arranged in a bound vortex configuration with strong in-plane spin-spin correlation but negligible interplanar correlation. Inelastic neutron-scattering studies provide strong evidence for the bound vortex state in the stage-2 $CoCl_2$-GIC. The absence of a remanent magnetization and the observation of a linear dependence of the magnetization on magnetic field for $T>T_{cu}$ identifies the phase above T_{cu} with paramagnetism.

Weak perturbations, such as the six-fold in-plane symmetry-breaking crystal field H_6 and the interplanar magnetic interaction J', become important at low temperatures and give rise to a 2D–3D crossover transition at T_{cl}. The spontaneous magnetization due to the coupling of the spins to H_6 gives rise to ferromagnetically ordered in-plane domains below a temperature of T_{cl}. The antiferromagnetic interplanar exchange interaction couples these ferromagnetic planar domains and results in an antiferromagnetic stacking arrangement of these domains, as implied by neutron scattering experiments (Zabel and Solin, 1992).

Magnetic phase transitions can be identified through the different field and temperature dependences that are found for the various properties such as the magnetic susceptibility, magnetization, and the magnetic contribution to the heat capacity. The most thoroughly studied magnetic donor GIC at the present time is the stage-1 europium compound, C_6Eu, for which the crystal structure is commensurate with the adjacent graphite layers. Each Eu atom has six nearest-neighbor Eu atoms at a distance of 4.31 Å within the layer plane, and six next-nearest-neighbor Eu atoms lie at a distance of 5.47 Å in *adjacent planes.* The configuration of the Eu^{2+} ion is $^8S_{7/2}$, which arises from a $(4f)^7$ electronic configuration, giving a total spin of 7/2 with no orbital angular momentum. Because the nearest-neighbor exchange interaction is antiferromagnetic, the spins form an antiferromagnetic frustrated triangular spin lattice. This spin system can easily be modified by introducing weak interactions, so that a variety of magnetic phases can be realized by small variations of the temperature and magnetic field. Field-dependent resistivity measurements (see Fig. 15) can be used to identify magnetic phase transitions, and the nature of

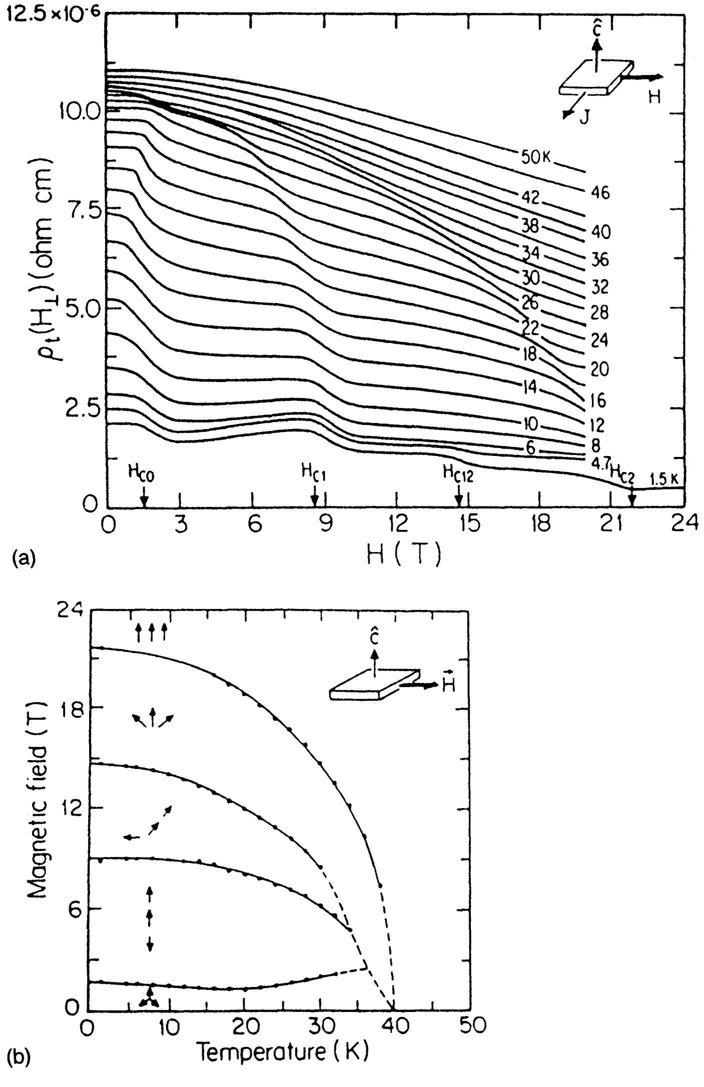

FIG. 15. (a) Anomalies in the field-dependent resistivity can be used to map out the magnetic phase diagram of a magnetic intercalation compound shown in (b). Specific results are shown here for the stage-1 donor compound C_6Eu. The orientations of three spins within the magnetic unit cell identify each of the magnetic phases. Ferromagnetic spin alignment along the magnetic field direction is achieved in the high-magnetic-field phase.

these magnetic phases is determined by magnetic measurements. On the basis of field-dependent resistivity measurements, a magnetic phase diagram has been developed for the low-temperature magnetic phases of C_6Eu (see Fig. 15).

The magnetic properties of donor GICs are distinctly different from those of the acceptor compounds because of differences relating to the basic magnetic coupling mechanism that occurs both in the pristine magnetic donor intercalates (which are metals) and in the intercalation compound (where the metallic behavior is associated predominantly with the graphite π bands). In the magnetic donor compounds, the mechanism for the magnetic exchange is through the conduction-band electrons, whereas the behavior of the magnetic acceptor GICs is dominated by the coupling between magnetic dipoles through induced dipoles on the nonmagnetic layers (i.e., by the superexchange mechanism).

5. SUPERCONDUCTING INTERCALATION COMPOUNDS

Superconductivity in intercalation compounds was first observed by Hannay *et al.* (1965), who identified superconducting transitions in the stage-1 heavy alkali-metal compounds with the intercalates K, Rb, and Cs in the graphite host material. The discovery of superconductivity in these materials attracted a great deal of interest insofar as neither the intercalate (an alkali metal) nor the host material (graphite) is by itself superconducting. Superconductivity in C_8K is also interesting with regard to its large anisotropy (the upper critical field H_{c2} is an order of magnitude larger for $H \perp c$ axis than for $H \parallel c$ axis), and for the unusual effect that for $H \parallel c$ axis, the intercalation compound exhibits type-I superconducting behavior (with no vortex formation), while for $H \perp c$ axis, the same sample can exhibit type-II behavior and vortex formation. Because of the low value for the superconducting transition temperature for C_8K (T_c=0.14 K), only a limited study of this system has been carried out (see Table 4). Other GICs with intercalates that are themselves superconducting (such as the stage-1 C_4KHg or the stage-2 C_8KHg) have T_c values in the range 1–2 K. Under pressure the alkali-metal intercalants enter the intercalate galleries at higher concentrations and give rise to T_c values up to 5 K (Zabel and Solin, 1992). The intercalation of fullerenes with alkali metals also leads to superconductivity (see FULLERENES), but with much higher transition temperatures (up to 33 K), because of the very much higher density of states near the Fermi level that can be achieved in this highly molecular solid.

Superconductivity is also observed in transition-metal dichalcogenides, but in this case the studies in the intercalation compounds have been focused in a different direction, since selected transition-metal dichalcogenides (such as TaS_2) are already superconducting prior to intercalation. In this case intercalation has been used to enhance the anisotropy of the superconducting phenomena.

Anisotropic superconductivity has been observed in a variety of layered materials (Dresselhaus and Chaiken, 1987), including deliberately structured materials, transition-metal dichalcogenides intercalated with large organic molecules, and graphite intercalation compounds. Because of the long range of the superconducting coherence distance ξ, the observation of 2D superconductivity focuses on samples with superlattice repeat distances I_c large compared with the superconducting coherence distance (in the *c*-axis direction) to

Table 4. Fundamental parameters[a] characterizing the superconductivity of C_8K (Zabel and Solin, 1992).

Parameter	Experimental value
T_c (mK)	128–198
Θ_D (K)	234.8
$N(0)V$	0.14
H_c (Oe)	10–13
$\kappa_\perp$ (T_c)	0.78
$\kappa_\parallel$ (T_c)	0.22
$\kappa_\perp / \kappa_\parallel$	3.55
$\lambda_\perp$ (0) (Å)	2300
$\lambda_\parallel$ (0) (Å)	8100
$\lambda_\perp / \lambda_\parallel$	0.28
$\xi_\perp$ (0) (Å)	10000
$\xi_\parallel$ (0) (Å)	2900
$\xi_\perp / \xi_\parallel$	3.45

---

[a] Θ_D is the Debye temperature, κ is the Landau–Ginzburg parameter, λ is the London penetration depth, ξ is the Ginzburg–Landau coherence length, and the subscripts $\parallel$ and $\perp$ refer to the directions parallel and perpendicular to the *c* axis, respectively.

achieve the condition of 2D superconductivity, $I_c > \xi_c$, where ξ_c denotes the coherence distance in the c direction for an anisotropic superconductor. Molecular-beam epitaxy and magnetron sputtering offer great promise for quantitative studies of 2D superconductivity and the 2D–3D crossover because of the flexibility of these synthesis techniques for the generation and control of large superlattice repeat distances in deliberately structured superconductors (Dresselhaus, 1987).

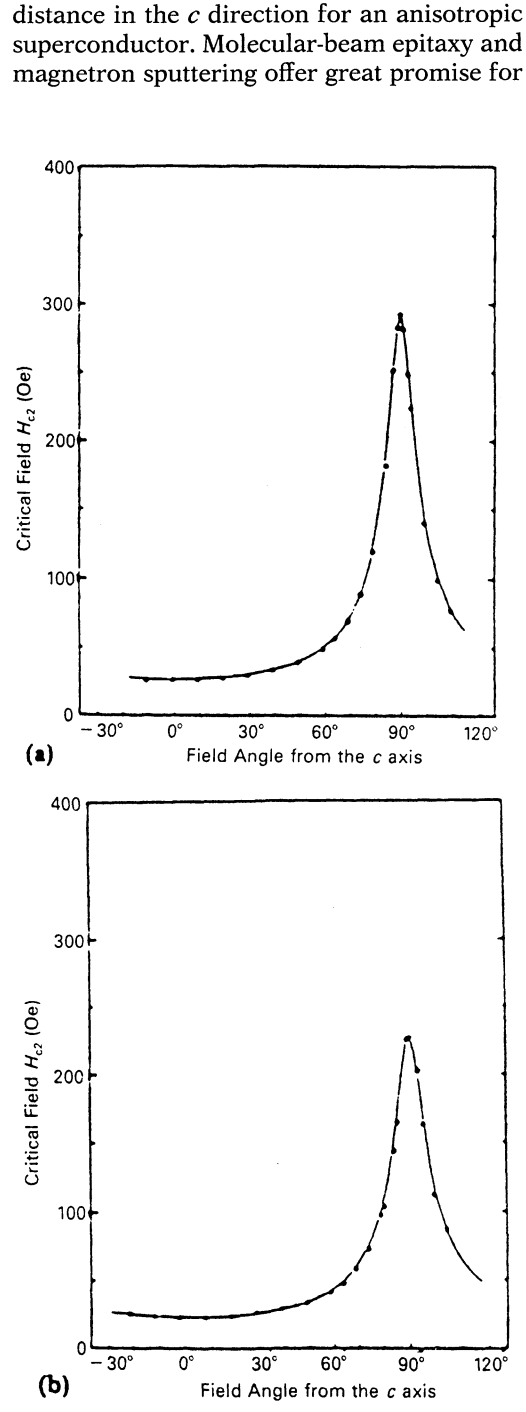

FIG. 16. Angular dependence of the upper critical field H_{c2} of the stage-1 ternary compounds (a) C_4KHg and (b) C_4RbHg. The curve in each case represents the best fit of the effective-mass model to the experimental points (Iye and Tanuma, 1984).

The earliest reports of superconductivity in layered transition-metal dichalcogenides intercalation compounds focused on the increase in T_c that occurred through intercalation with large organic molecules (Gamble and Geballe, 1976). The effect of lower dimensionality was observed in this work, especially in systems such as TaS_2 intercalated with *n*-octadecylamine, where a separation of nearly 60 Å was achieved between sequential Ta layers. This work was historically important in stimulating theoretical calculations of 2D superconductivity.

Most useful for studies of the 3D–2D superconducting transition have been the synthetic superlattice structures prepared by molecular beam epitaxy, because of the dimensional control that can be exercised in preparing the multilayers. Because of the abrupt interfaces and the possibility of preparing ordered sequences of superconducting monolayers, GICs provide a complementary tool for studying highly anisotropic superconducting materials. There is no known acceptor GIC that has shown evidence for superconductivity.

Critical-field studies for anisotropic superconductors have attracted considerable interest because of their similarity to the high-T_c superconducting cuprates, which are also layered compounds with anisotropic superconducting properties. For the intercalation compounds the angular dependence of the critical field is well approximated by the anisotropic Ginzburg–Landau formula (Zabel and Solin, 1992),

$$H_{c2}(\theta) = (\Phi_0/2\pi\xi_\perp^2) \times (\cos^2\theta + \epsilon^2 \sin^2\theta)^{-1/2}, \qquad (1)$$

in which $\Phi_0 = hc/2e = 2.07 \times 10^{-7}$ G cm^2 is the elementary flux quantum and θ is the angle between the magnetic field and the c axis. Here the parameter ϵ characterizes the anisotropic behavior and is given by

$$\epsilon^2 = m_\perp / m_\parallel = (H_{c2\perp} / H_{c2\parallel})^2 = (\xi_\parallel / \xi_\perp)^2, \qquad (2)$$

where ξ denotes the coherence length and m denotes the effective mass in the $\parallel$ and $\perp$

directions. We denote by $H_{c2\perp}$ the upper critical field for $\theta=0°$ whereby the diamagnetic screening currents lie in the layer planes. Similarly we denote by $H_{c2\parallel}$ the upper critical field for $\theta=90°$ whereby the diamagnetic screening currents must traverse different layer planes. These relations have been used to obtain qualitative fits to the experimental $H_{c2}(\theta)$ curves shown in Fig. 16. From these fits, values for the parameters $H_{c2\parallel}$, $H_{c2\perp}$, $\xi_{\parallel}$, and $\xi_{\perp}$ have been deduced. Specific values of these parameters are given in Table 5 for stage-1 KHg-GICs in comparison with values for pristine intercalate KHg, which is also superconducting, but shows no anisotropy. It is also of interest to note that intercalation increases the T_c value of a superconducting intercalate species (Zabel and Solin, 1992). For comparison a number of fundamental parameters characterizing the superconducting compound C_8K are listed in Table 4.

Also of interest is the temperature dependence of the critical fields $H_{c2}(T)$. Measurements of $H_{c2}(T)$ on GICs have been interpreted in terms of the anisotropic Ginzburg–Landau theory, which provides a qualitative fit to the data near T_c. Assuming an elliptical fluxoid to describe the anisotropy in the coherence length, the anisotropic Ginzburg–Landau theory yields the following expressions for the temperature dependence for the critical fields parallel and perpendicular to the c axis:

$$H_{c2\parallel}(T)=\frac{\Phi_0}{2\pi\xi_{\parallel}\,\xi_{\perp}}\left(1-\frac{T}{T_c}\right),$$
$$H_{c2\perp}(T)=\frac{\Phi_0}{2\pi\xi_{\perp}^2}\left(1-\frac{T}{T_c}\right). \qquad (3)$$

Characteristic of the Ginzburg–Landau theory is a linear T dependence of H_{c2} near T_c for both $H_{c2\perp}$ and $H_{c2\parallel}$. The anisotropy in H_{c2} comes in through the prefactor coefficients in Eq. (3). Critical-field measurements on stage-2 KHg-GICs for both $H\parallel$ c axis and $H\perp$ c axis are well approximated by a linear T dependence near T_c in accordance with Eq. (3). Fits to these data yield values for the parameters $\xi_{\parallel}$ and $\xi_{\perp}$, consistent with values obtained from fits to the angular dependence of $H_{c2}(\theta)$ discussed above.

The standard BCS model for superconductivity with only occupied graphite π bands cannot explain the observed transition temperatures because the π-band density of states is too small. On the other hand, a model with only heavy-mass intercalate-related bands cannot explain the large observed anisotropy in the superconducting properties of GICs. Therefore the model for superconductivity in GICs must include consideration of both the graphite π-band electrons and the occupied heavy-mass intercalate bands, showing that it is necessary to have at least two types of occupied bands or strong hybridization between them to produce superconductivity in GICs.

Table 5. Comparison of the superconducting parameters of stage-1 KHg-GICs and pristine KHg (Dresselhaus and Chaiken, 1987).

Parameter	Stage-1 KHg-GIC	Pristine KHg
I_c (Å)	10.22	—
T_c (K)	1.53	0.98
$H_{c2\perp}$ (Oe)	83	244
$H_{c2\parallel}$ (Oe)	850	244
$\xi_{\perp}$ (0) (Å)	1960	1135
$\xi_{\parallel}(0)$ (Å)	192	1135
$\xi_{\perp}/\xi_{\parallel}$	10.2	1

The theoretical models predict both the temperature dependence and anisotropy of H_{c2} for GICs. The models are generally consistent with the observed data. The models explain many of the detailed properties of superconducting GICs, including the absence of superconductivity in acceptors that lack occupied heavy-mass intercalate levels. The models show that the value of the superconducting transition temperature is more sensitive to changes in the π-band electronic density of states than to the corresponding changes in the heavy-mass band.

One key to gaining a detailed understanding of superconductivity has been through in-depth investigations of the tunneling characteristics in the normal and superconducting states, which in turn yield direct information on the density of states. The layered nature of many intercalation compounds makes them amenable for tunneling measurements. Such experiments have also been performed on intercalated fullerenes, thereby yielding a direct measurement of the temperature dependence of the superconducting band gap in these relatively high-T_c superconductors (see FULLERENES).

Since the critical fields for superconducting GICs are easily achieved in the laboratory, GICs provide model systems of interest for the study of anisotropic superconductors. Such studies have been of interest with regard to high-T_c superconductors for which the magnetic fields required to study the corresponding effects are difficult to achieve experimentally.

Acknowledgment

The preparation of this manuscript was supported by the National Science Foundation Grant No. 92-01878DMR.

GLOSSARY

Basal Plane: The plane normal to the main symmetry axis, usually called the *c* axis.

BCS Theory: Microscopic theory of superconductivity based on the pairing of electrons that are coupled by the electron–phonon interaction.

Bi-intercalation: The intercalation of a layered binary intercalation compound with a second intercalate species to form an intercalation compound that is ordered with respect to each of the two intercalate species.

Binary Compound: An intercalation compound for which the intercalate is a single atomic or molecular species.

Chalcogen: Element from column VI of the periodic table: O, S, Se, Te.

Clays: Compounds in a class of layered aluminosilicates such as vermiculite, which intercalate water between layers and take up other metal ions by ion exchange interactions.

Commensurate: Intercalation compound in which atoms in the intercalate layer and in the host material are in registry.

Covalent Bonding: A localized bond between two atoms to which the valence electrons from each atom contribute.

Discommensuration: A modulation in the structure of two materials with different lattice constants at an interface to achieve regions where the two structures are locally in registry, but separated by boundaries called misfit dislocations where the material with the smaller lattice constant has an extra row of atoms.

Disproportionation: Formation of chemical species in the intercalate galleries with proportions of atomic constituents different from that prior to intercalation.

Electron Affinity: Energy required to take an electron from the top of the valence band (in a semiconductor) to the lowest ionization state.

Ginzburg–Landau theory: Macroscopic phenomenological theory of superconductivity.

Graphene layer: An isolated single layer of hexagonally arranged carbon atoms in a honeycomb structure such as found in graphite.

Intercalate Gallery: van der Waals gap between basic structural units of a layered host material in an intercalation compound.

Ionic Bonding: The electrostatic bonding between two atoms caused by the transfer of charge from the donor atom to the acceptor atom.

Octahedral Sites: Sites with six nearest neighbors in the cubically symmetric locations $(\pm a,0,0)$, $(0,\pm a,0)$, $(0,0,\pm a)$.

Polytypes: The ensemble of ordered arrangements possible for a given basic structural unit.

Stage: The ordered periodic superlattice arrangement of the intercalate in an intercalation compound, whereby the intercalate layers are separated by a fixed number of host layers.

Stage Index: Integer specifying number of host layers between consecutive intercalate layers.

Ternary Compound: An intercalation compound into which two distinct intercalates are introduced.

Tetrahedral Sites: Sites in the lattice with nearest neighbors at the corners of a regular tetrahedron.

Transition-Metal Ion: Metal ions containing unfilled *d* shells (where *d* electrons denote electrons with angular momentum of $2\hbar$).

van der Waals Bonding: Weak bonding between two atoms, such as might arise from π bonding between electrons, falling off rapidly with distance.

Vortex: A flow of charge or an arrangement of spins with circulation about a singular point or line.

List of Works Cited

Aldebert, P., Baffier, N., Legendre, J. J., Livage, J. (1982), *Rev. Chim. Minèrale* **19**, 485–495.

Besenhard, J. O., Wudy, E., Moewald, H., Nickl, J. J., Biberacher, W., Foag, W. (1983), *Synth. Met.* **7**, 185–192.

Brec, R. (1987), in: M. S. Dresselhaus (Ed.), *Intercalation in Layered Materials*, NATO Advanced Study Institute, Series B: Physics, Vol. 148, New York: Plenum, pp. 93–134.

Dresselhaus, M. S. (1984), *Phys. Today* **37** (3), 60–68.

Dresselhaus, M. S. (Ed.) (1987), *Intercalation in Layered Materials*, NATO Advanced Study Institute, Series B: Physics, Vol. 148, New York: Plenum.

Dresselhaus, G., Chaiken, A. (1987), in: M. S. Dresselhaus (Ed.), *Intercalation in Layered Materials*, NATO Advanced Study Institute, Series B: Physics, Vol. 148, New York: Plenum, pp. 387–406.

Dresselhaus, M. S., Dresselhaus, G. (1981), *Adv. Phys.* **30**, 139–326.

Gamble, F. R., Geballe, T. H. (1976), in: N. B. Hannay (Ed.), *Treatise on Solid State Chemistry*, Vol. 3, New York: Plenum, pp. 89–166.

Hannay, N. B., Geballe, T. H., Matthias, B. T., Andres, K., Schmidt, P., MacNair, D. (1965), *Phys. Rev. Lett.* **14**, 225–226.

Hérold, A. (1979), in: F. Lévy (Ed.), *Physics and Chemistry of Materials with Layered Structures*, New York: Reidel, Vol. 6, pp. 323–421.

Hérold, A., Guérard, D., Lagrange, P. (1983), *Synth. Met.* **7 & 8**, Proceedings of the Third International Conference on Intercalation Compounds of Graphite.

Iye, Y., Tanuma, S. (1984), in: S. Tanuma, H. Kamimura (Eds.), *Graphite Intercalation Compounds*, Summary report of the Special Distinguished Research Project supported by the Ministry of Education, Japan, pp. 253–267. See also *Phys. Rev. B* **25**, 4583–4592 (1982).

Julien, C., Hatzikraniotis, E., Paraskevopoulos, K., Chevy, A., Balkanski, M. (1986), *Solid State Ionics* **18–19**, 859–864.

Kamimura, H. (1987), *Phys. Today* **40** (12), 64–71.

Liang, W. Y. (1987), in: M. S Dresselhaus (Ed.), *Intercalation in Layered Materials*, NATO Advanced Study Institute, Series B: Physics, New York: Plenum, Vol. 148, pp. 31–74.

McRae, E., Marêché, J. F. (1988), *J. Mater. Res.* **3**, 75–86.

Newman, A. C. D. (Ed.) (1987), *Chemistry of Clays and Clay Minerals*, New York: Wiley.

Schafhäutl, C. (1840), *J. Prakt. Chem.* **21**, 129–157.

Solin, S. A. (1982), *Adv. Chem. Phys.* **49**, 455–532.

Solin, S. A. (1984), *J. Molecular Catal.* **27**, 293–303.

Solin, S. A., Zabel, H. (1988), *Adv. Phys.* **37**, 87–254.

Suematsu, H., Suda, K., Metoki, N. (1988), *Synth. Met.* **23**, 7–12.

Terai, T., Takahashi, Y. (1981), *J. Nucl. Sci. Tech.* **18**, 643–648.

Vogel, F. L., Hérold, A. (Eds.) (1977), *Proceedings of the Conference on Intercalation Compounds of Graphite*, Lausanne: Elsevier Sequoia, Vol. 31.

Wada, N., Suzuki, M., Hines, D. R., Koga, K., Nishihara, H. (1987), *J. Mater. Res.* **2**, 864–870.

Xiang, X. D., Zettl, A., Vareka, W. A., Corkill, J. L., Barbee, T. W., III, Cohen, M. L. (1991), *Phys. Rev. B* **43**, 11 496–11 499.

Zabel, H., Solin, S. A. (Eds.) (1990), *Graphite Intercalation Compounds I: Stucture and Dynamics*, Springer Series in Materials Science, Vol. 14, Berlin: Springer.

Zabel, H., Solin, S. A. (Eds.) (1992), *Graphite Intercalation Compounds II: Transport and Electronic Properties*, Springer Series in Materials Science, Vol. 18, Berlin: Springer.

Further Reading

Dresselhaus, M. S. (1984), *Phys. Today* **37** (3), 60–68.

Dresselhaus, M. S. (Ed.) (1987), *Intercalation in Layered Materials*, NATO Advanced Studies Institute, Series B: Physics, Vol. 148, New York: Plenum.

Dresselhaus, M. S., Dresselhaus, G. (1981), "Intercalation compounds of graphite," *Adv. Phys.* **30**, 139–326.

Gamble, F. R., Geballe, T. H. (1976), in: N. B. Hannay (Ed.), *Treatise on Solid State Chemistry*, New York: Plenum, Vol. 3, pp. 89–166.

Hérold, A. (1979), in: F. Lévy (Ed.), *Physics and Chemistry of Materials with Layered Structures*, New York: Reidel, Vol. 6, pp. 323–421.

Kamimura, H. (1987), *Phys. Today* **40** (12), 64–71.

Solin, S. A. (1982), *Adv. Chem. Phys.* **49**, 455–532.

Solin, S. A., Zabel, H. (1988), *Adv. Phys.* **37**, 87–254.

Zabel, H., Solin, S. A. (Eds.) (1990), *Graphite Intercalation Compounds I: Structure and Dynamics*, Springer Series in Materials Science, Vol. 14, Berlin: Springer.

Zabel, H., Solin, S. A. (Eds.) (1992), *Graphite Intercalation Compounds II: Transport and Electronic Properties*, Springer Series in Materials Science, Vol. 18, Berlin: Springer.

ERRATUM

INTERCALATION COMPOUNDS (Vol. 8, pp. 133–155). M. S. Dresselhaus and G. Dresselhaus.

The figure shown as Fig. 6 is incorrect. The correct figure is shown here.

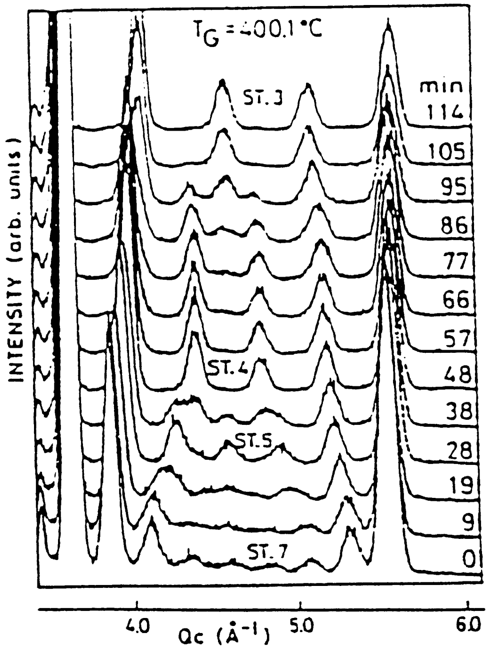

FIG. 6. *In situ* x-ray study of staging transitions (where T_G, the graphite temperature, is 400.1 °C). The kinetics of the staging process in K-GICs is obtained from the time dependence of (00*l*) reflections for transitions from a stage-7 compound to a stage-3 compound (Suematsu *et al.*, 1988).

ION-BEAM MODIFICATION OF MATERIALS

ROBERT S. AVERBACK, *Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, U.S.A.*

Introduction 173
1. **Depth Distributions of Implanted Ions**
1.1 Electronic Stopping Power 174
1.2 Nuclear Stopping Power 175
1.3 Implantation Range Distribution 176
1.3.1 Transport Theory 176
1.3.2 Computer Simulation 177
1.4 Channeling 178
2. **Implantation Damage** 179
2.1 Displacement Processes 179
2.2 Cascade Damage 179
2.3 Thermal Spikes 179
2.4 Applications of Ion Implantation 180
2.4.1 Doping of Semiconductor Devices 180
2.4.2 Buried Layers 181
2.4.3 Tribological Applications 182
3. **Ion-Beam Mixing** 183
3.1 Mechanisms of Ion-Beam Mixing 183
3.1.1 Recoil Implantation 183
3.1.2 Cascade Mixing 183
3.1.3 Thermal-Spike Mixing 184
3.1.4 Phase Formation During Ion-Beam Mixing 184
3.2 Ion-Beam–Assisted Deposition. 184
3.2.1 Dopant Incorporation 186
3.2.2 Improved Adhesion 186
4. **Implantation at Elevated Temperature** 186
4.1 Theory of Radiation-Enhanced Diffusion 186
4.2 Examples of Radiation-Enhanced Diffusion 187
5. **Summary** 189
Glossary 189
Works Cited 190
Further Reading 191

INTRODUCTION

The possibility of using ion beams for materials modification was first explored nearly 50 years ago. However, not until the early 1970s, when the electronics industry began their frenetic drive toward minaturization of integrated circuits, did ion implantation become a wide-scale commercial process. It was also about 50 years ago when studies of radiation damage in metals and ceramics were initiated. These early investigations were part of the research and development program on nuclear weapons, but they were soon expanded to include materials problems in fission and fusion reactors. These two fields of ion implantation and radiation damage have now become integrated in the far more comprehensive field of ion-beam modifications of materials. No longer is ion-beam processing of materials restricted to the implantation of electrically active impurities into Si for semiconductor applications, but rather ion beams are being employed as a basic processing step in the fabrication of many advanced materials. For example, ion implantation is being used to provide corrosion and oxidation resistance in metals; to create nonequilibrium surface phases with special properties; to improve the quality, epitaxy, and adhesion of vapor-deposited thin films; and to clean or pattern surfaces. One of the most celebrated recent applications of ion-beam modifications of materials has derived from the field of tribology, where it was shown that the implantation of nitrogen into titanium hip joint replacements can extend the wear resistance of these prostheses beyond their required service lifetime. Indeed, the field of ion–solid interactions has moved from the workbench of nuclear physicists to the design tables of materials scientists.

Several features of ion-beam processing underlie its rapid growth in recent years:

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. The concentration depth profile of an implanted species can be tailored to the materials requirements, whereas those obtained by the more conventional means of in-diffusing surface films are dictated by the diffusion equation.
2. Any ion species can be implanted into the solid. Because ion implantation is a high-energy process, thermodynamic and kinetic constraints are minimized. Consequently, nonequilibrium phases can be produced.
3. Ion implantation is intrinsically a low-temperature processing step, and so high-temperature anneals, which can cause interdiffusion in complex devices or phase transformations of metastable systems, can be avoided.
4. Ion irradiation can be employed to create structural defects in the surface region. Although defects are generally undesirable for implantation of dopants in semiconductors, they often have beneficial effects for other applications. Defects serve as nucleation sites for epitaxial growth, they modify the index of refraction for waveguide applications, and in some metals they lead to amorphization and concomitant improvement in tribology properties.
5. Ion implantation is conducive to lateral patterning using masks or focused ion beams. With its natural control of ion depth penetration, three-dimensional patterning becomes possible with submicron resolution.

Although ion implantation is often associated with the simple addition of foreign atoms into a substrate, doping is only one of many changes that occur in the material. As ions slow down in a target, they initiate cascades of energetic collisions between lattice atoms, as illustrated in Fig. 1. These collisions lead to the ejection of atoms from the surface (sputtering), the creation of Frenkel pairs (vacancies and interstitial atoms) in the interior of the crystal, and the relocation of many target atoms from one lattice site to another. These various processes, and their relative frequencies of occurrence, are indicated in Fig. 1. In this article, an elementary discussion of the physical principles underlying ion-beam modifications of materials is presented, and several examples are cited to illustrate the versatility of this new technology.

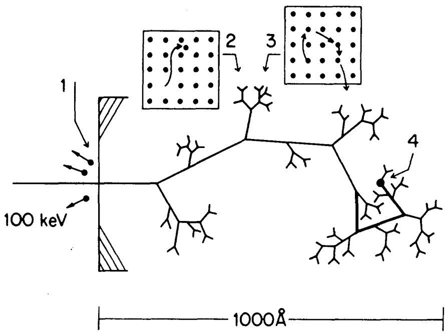

Summary of Events

Process	Relative Occurrence
1. Sputtered Atoms	~3
2. Frenkel Pairs	~500
3. Relocated Atoms	~5000
4. Implanted Atoms	1

FIG. 1. Schematic view of the different processes that occur and their frequency as an ion penetrates a solid and comes to rest. The values are typical for an ion implantation energy of approximately 100 keV.

1. DEPTH DISTRIBUTIONS OF IMPLANTED IONS

One of the primary advantages of ion-beam processing of materials is that the depth distribution of the implanted ions can be controlled with high precision, and with the use of a mask or a focused ion beam, lateral patterning of the substrate is also possible. It is now feasible to predict these implantation distributions accurately, greatly facilitating the design of new structures. In this section, the theory of the stopping of ions in materials and the methods for calculating implantation profiles are presented. In the next section, defect production during ion implantation is discussed, and several illustrative examples of the use of ion implantation in materials are provided.

1.1 Electronic Stopping Power

As energetic ions penetrate a solid, they lose energy by colliding with atomic nuclei and electrons. Theories of electronic stopping in the energy regime most relevant to ion-beam modifications of materials were developed by Firsov (1959) and Lindhard *et al.* (1963). The Firsov model, which is the easier to visualize, assumes that the ion velocity is slow relative to the electron orbital velocities, so that electrons behave "adiabatically," i.e., they are able fully to readjust their electronic states relative to the projectile position. Consequently, electrons are free to exchange between the projectile and target atom during collisions. Since the ion must perform work in accelerating target electrons up to its own velocity during an electron exchange, but without recapturing that energy when electrons leave, the ion loses energy. The energy loss is thus determined by calculating the electron currents flowing between the moving ion and the substrate atoms. Within the Thomas–Fermi model of an atom, the stopping power is given by the expression

$$\begin{aligned}\frac{dE}{dx}&=-NS_e(E)\\&=-N\{(0.336\ \mathrm{eV}^{1/2}\ \text{Å}^2)[Z_1+Z_2]\\&\quad\times[E/A_1]^{1/2}\},\end{aligned}\tag{1}$$

where $S_e(E)$ is the so-called electronic stopping cross section and has units of energy-area, Z_1 and Z_2 are the atomic numbers of the implanted ion and target atoms, respectively, A_1 is the atomic mass of the ion, N is the atomic density, E is the projectile energy, and x is the distance along the ion trajectory. Equation (1) shows that the electronic stopping power increases with increasing ion velocity and atomic number; the theory of Lindhard and Scharff yields similar results. It is interesting to note that because the stopping power depends linearly on the velocity of the ion, enormous electronic stopping powers can be obtained with heavy, GeV ions—energy losses exceeding 5 keV/Å. This incredibly high power density has recently been exploited for materials modification applications. In high-temperature superconductors, for example, it was shown that heavy GeV particles produce long, highly damaged tracks, over 20 μm in length and approximately 40 Å in diameter, and that these tracks are very effective in pinning magnetic flux lines. Thus, GeV particle irradiation can be employed to enhance the critical current densities of high-temperature superconductors (Roas *et al.* 1990). GeV irradiations are also being used to produce high-precision filters. For this application, long, porous tubes of uniform diameter are created by selective etching of the damage tracks (Spohr, 1990).

1.2 Nuclear Stopping Power

The loss of projectile energy due to the nuclear collisions is described by the expression

$$\frac{dE}{dx}=-NS_n(E)=-N\int_0^{T_{\max}}T\frac{d\sigma(E,T)}{dT}\,dT.\tag{2}$$

The nuclear stopping cross section $S_n(E)$ is defined by the right side of Eq. (2) (it also has units of energy-area), and $d\sigma(E,T)$ is the differential cross section for an ion of energy E to transfer an energy between T and $T+dT$ to a target atom. The cross section is governed by the appropriate interatomic potential, which for ion implantation is a screened Coulomb potential. Historically, Thomas–Fermi potentials have been employed for this purpose, although recently the more accurate Universal potential (Ziegler *et al.* 1985) has been preferred. The maximum energy transfer in Eq. (2), $T_{\max}$, occurs in a head-on collision; it is given by

$$T_{\max}=4\frac{A_1A_2}{(A_1+A_2)^2}E.\tag{3}$$

Equation (3) shows that the maximum energy transfer is largest when the ion and substrate atoms have similar masses. For this reason, energetic electrons, even those with energies in excess of 1 MeV, cause little structural damage in materials. Lindhard and co-workers (1963) showed that by employing reduced units, the nuclear stopping of all ion–solid combinations could be described by a single universal stopping function,

$$s_n(\epsilon)=-d\epsilon/d\rho,\tag{4}$$

where the reduced energy ϵ is

$$\epsilon=\left\{\frac{A_2E}{A_1+A_2}\right\}\left\{\frac{Z_1Z_2e^2}{a}\right\}^{-1}\tag{4a}$$

and the reduced length ρ is

$$\rho=N4\pi a^2A_1A_2x/(A_1+A_2)^2\ .\tag{4b}$$

The screening length a is

$$a=0.8853a_0/(Z_1^{2/3}+Z_2^{2/3})^{1/3}\ .\tag{4c}$$

where a_0 is the Bohr radius. The stopping function is shown in Fig. 2. Stopping powers for all ion–solid combinations can be determined from this one curve using the expression

$$\frac{dE}{dx} = -NS_n(E) = -N\left\{\frac{\pi a^2 T_m}{\epsilon}\right\} s_n(\epsilon). \tag{5}$$

Also shown in reduced units in Fig. 2 is the electronic stopping power. [$s_e(\epsilon)$ is not strictly a universal function, but it does not vary strongly with atomic number.] It is noteworthy that the total stopping power $s_{\text{tot}}(\epsilon) = s_n(\epsilon) + s_e(\epsilon)$ is rather constant over the range $0.2 < \epsilon < 2$, which is the range of energies of most implantations. The maximum nuclear stopping occurs at $\epsilon = 0.44$, and the electronic stopping becomes greater than the nuclear stopping at $\epsilon = 4$. This scaling with reduced energy units thus provides a natural way to compare implantations for all ion–target combinations. Table 1 lists the implantation energy at which $\epsilon = 1$ for various ion–solid combinations.

1.3 Implantation Range Distribution

1.3.1 Transport Theory. The distance that an ion travels before coming to rest within a solid, i.e., the path length, is given by the expression

$$L(E) = \int_E^0 \frac{dE}{dE/dx} = \frac{1}{N}\int_E^0 \frac{dE}{S_n(E) + S_e(E)}. \tag{6}$$

The path length, however, is usually of less interest for ion implantation than is the ion range, i.e., the distance between an ion's point of entry at the surface and its final location in the solid. Since the ranges of ions vary because of their different points of entry in the solid, lattice vibrations, and the discrete nature of collisions, a distribution of ranges is generated by an implantation, even for a single ion type, energy, and angle of incidence. The range distribution is comprised of longitudinal (normal to the surface) and transverse (parallel to the surface) components. Although the longitudinal component, or projected range R_P, and its variance are generally of primary concern for ion implantation, the transverse straggling ($\langle R_T^2 \rangle / \langle R_P^2 \rangle$) is of interest when implantations are performed through a mask, since it determines the minimum line width in circuit patterns (see, e.g., Tamura, 1991).

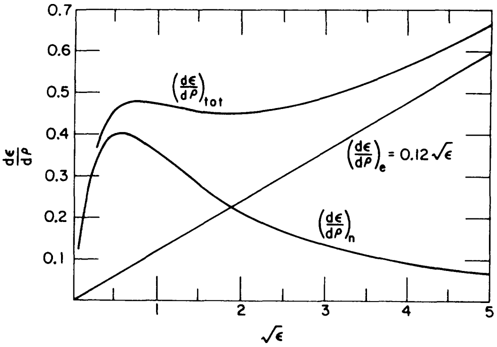

FIG. 2. Energy dependence of the nuclear and electronic stopping powers in reduced units. The nuclear curve is a universal function, but the electronic stopping depends somewhat on the ion–target combination. At low energies, most energy is dissipated through collisions with substrate atoms, but at high energies, most is lost by electronic excitation. (After Schiøtt, 1970.)

The range distributions of particles can be determined by solving the appropriate Boltzmann transport equation. Although the formalism is straightforward, solving the integro-differential equations is unwieldy. A more tractable procedure has been to calculate the moments of the distribution and then to construct the range distribution from an expansion of the various moments. The Edgeworth expansion has been found useful for this purpose (see Sigmund, 1969; Winterbon, 1975). The first two moments of the Edgeworth expansion, the mean range and its variance, determine a Gaussian distribution, and they often provide sufficient accuracy. The third moment becomes important for low-energy implantations since these distribu-

Table 1. Projectile energies (in keV) for which $\epsilon = 1$.

Projectile	Target				
	Be	Si	Ge	Ag	Au
H	0.256	1.16	3.32	5.47	10.8
C	2.44	11.2	25.1	39.0	72.1
Al	18.1	37.1	69.2	101.3	176
Ni	89.3	147	219	297	185
Au	1017	1347	1535	1810	2331

tions are skewed toward the surface. A convenient expression for the mean range is (Schiøtt, 1970)

$$R_P=C_1(\mu)A_2\,\frac{(Z_1^{2/3}+Z_2^{2/3})^{1/2}}{Z_1Z_2}\,E\mu\text{g/cm}^2\ \text{keV}, \tag{7}$$

where $\mu=A_2/A_1$ and $C_1(\mu)\approx 0.62+0.171\log\mu$. The rule of thumb (Davies, 1992)

$$\langle R_P^2\rangle^{1/2}/\langle R_P\rangle=0.35[2(A_1A_2)^{1/2}/(A_1+A_2)] \tag{8}$$

can be used for the variance. Figure 3 compares the range distributions generated from a Gaussian distribution, using Eqs. (7) and (8), with those determined more accurately by computer simulation, a method now described.

1.3.2 Computer Simulation. With the vast improvements in personal computers in recent years and the development of efficient codes, simulation has become the primary means of calculating range distributions. These codes also yield the depth distributions of defects and electronic excitations that are produced by the ions. The procedure employed by simulation programs is to follow the path of an implanted ion by calculating the scattering integrals and electronic losses of the ion as it collides with target atoms, one by one. Distributions are generated by compiling the histories of many trajectories, with each having a different point of entry in the target. The trajectories of the recoil atoms can also be obtained in a similar manner, although at much greater computer overhead due to their greater number. Only marginal improvement in the accuracy of the implantation profile is obtained by this procedure, but for calculations of sputtering, defect production, and ion-beam mixing (see below), the inclusion of recoil atoms is necessary. The most widely used simulation code, TRIM, was developed by Biersack and Haggmark (1980). It is distinguished by an ingenious scheme to reduce the computational effort in calculating the scattering integral. Additional computer economy is achieved in TRIM by the assumption that the target material is amorphous. This reduces the lengthy process of locating the ion's next scattering partner in a lattice to a simple application of a random number generator. Because of this inherent randomness, TRIM is a Monte Carlo simulation and is most useful for providing statistical averages. Accurate concentration distributions can presently be obtained running TRIM on a personal computer in under 1 h (see Zeigler *et al.*, 1985, for details). For some applications, the crystalline structure of the target has a large influence on the ion trajectories. In these cases, a lattice must be included in the simulation model, as in the code MARLOWE (Robinson and Torrens, 1974). Since these codes must search the lattice for the successive collision partners of the ion and recoil atoms, they are accordingly slower than TRIM.

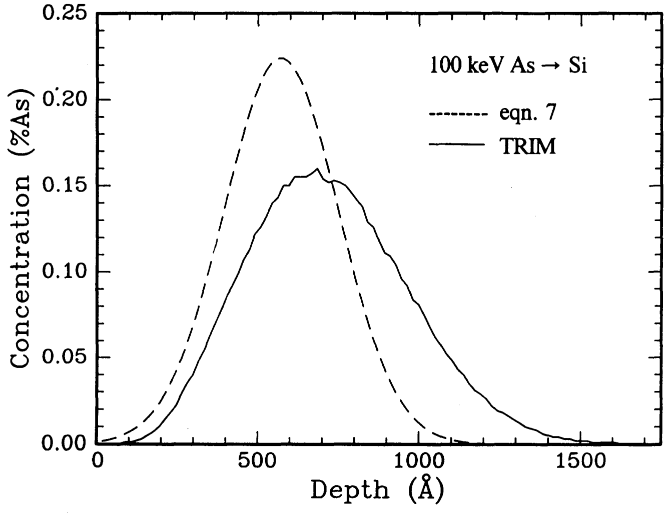

FIG. 3. Range distribution for the implantation of Si with 100-keV As calculated by TRIM and by Eqs. (7) and (8).

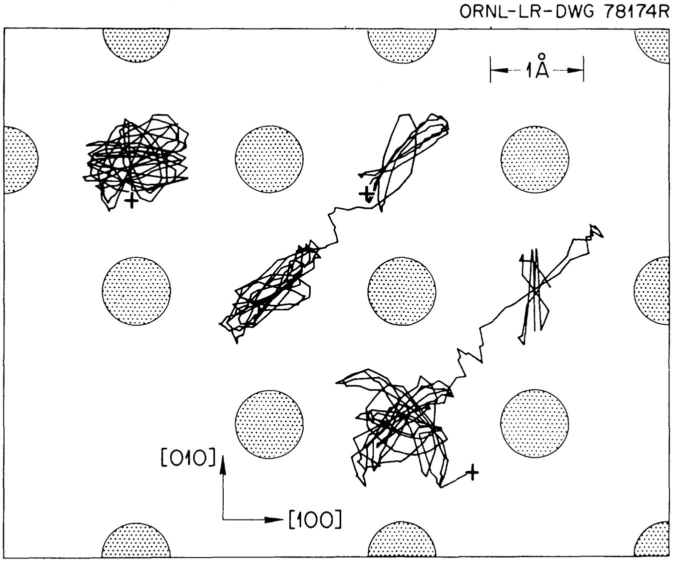

FIG. 4. Computer simulations showing the oscillatory motion of channeled ions in the (100) plane perpendicular to the ⟨001⟩ channeling direction. Radius of shaded circles is the distance of closest approach. (After Robinson and Oen, 1963.)

1.4 Channeling

Years ago, Stark (1912) predicted that the crystal lattice should influence the trajectory of ions in a crystalline solid, and by the early 1960s several experimenters began noticing long tails in their ion range distributions. But it was not until 1963 that the phenomenon of "channeling" was clearly elucidated by Robinson and Oen (1963), using computer simulation. In their simulations, which employed a lattice model, Robinson and Oen observed that some Cu projectiles had extremely long trajectories in a Cu lattice, particularly when the initial ion velocity was aligned with a major crystal axis. By following these special trajectories, they revealed that ions could travel through the relatively open spaces of a lattice as a result of a steering effect of the atomic rows. Figure 4 shows the transverse motion of a 1-keV Cu ion as it channels in a ⟨100⟩ direction of Cu. As the ion nears an atomic row, the repulsive interatomic potential nudges the ion back toward the center of the channel. The ion thus oscillates in two dimensions as it penetrates into the solid. Experimental evidence for channeling is shown in Fig. 5 (Piercy *et al.*, 1963), where the ranges of 40-keV Kr^+ in single-crystalline Al are shown for different incident ion directions. (Actual range distributions are the slopes of the curves in Fig. 5 as a function of depth.) The range in amorphous Al_2O_3 is shown for comparison. In the absence of channeling, the ranges of 40-keV Kr in Al and Al_2O_3 should be nearly the same. In many cases, the beam and substrate need only be aligned to within 3°–10° for channeling to occur, so that some fraction of the implanted ions can scatter into a channeling direction, independent of the initial orientation. This propensity for channeling poses serious problems for ion implantation applications where long tails in the range distribution are undesirable, such as shallow junction devices. Accordingly, much effort has been expended in finding crystallographic directions in Si, and other semiconductors, where channeling effects are minimized.

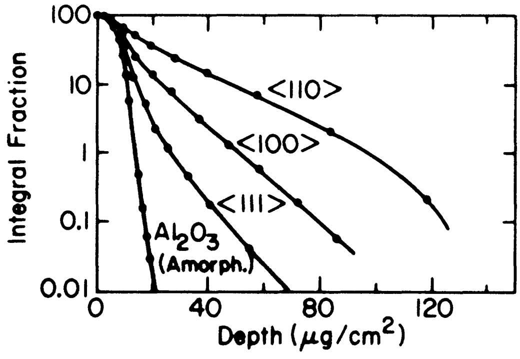

FIG. 5. Measurements of the range of 40-keV Kr in Al when the implantation beam is aligned with different crystallographic directions. Here the number of ions implanted deeper than a given depth is shown. For comparison, the range in amorphous Al_2O_3 is shown (after Piercy *et al.*, 1963).

2. IMPLANTATION DAMAGE

2.1 Displacement Processes

If a target atom recoils with an energy greater than the displacement energy E_d, it leaves its lattice site and creates a vacancy. The recoil atom may then fall into an interstitial site, or, if its energy is sufficiently high, it may displace additional atoms before coming to rest. Displacement energies in metals and semiconductors range from 5 to 40 eV (see Table 2). These values of E_d are surprisingly low in metals and derive from an efficient means of transporting energy and mass in a crystalline lattice by linear replacement collision sequences. Vineyard and co-workers (Gibson *et al.*, 1960) revealed, using molecular-dynamics computer simulations, that defects are created by series of collisions whereby one atom collides with its neighbor and replaces it in a linear sequence, as shown in Fig. 6.

2.2 Cascade Damage

High-energy particles can produce recoils with energies that are much greater than the minimum energy required for displacements. In this case, both the primary and the recoil atom leave the lattice site and produce additional displacements at other sites. If the energies of the secondary recoil atoms are sufficiently high, they too continue the multiplication process of displacing atoms, thus creating a displacement cascade. Since the cross section for atomic collisions increases rapidly with decreasing energy, because of the screened Coulomb interaction between atoms, a projectile will produce recoils closer and closer together as it slows down. The energies of the recoil atoms are usually low; therefore they too do not travel far before undergoing additional collisions in the localized vicinity of their creation (see Fig. 1). Consequently, defects produced in a displacement cascade are spatially correlated and often form defect clusters. This has been demonstrated experimentally using field ion microscopy, as shown in Fig. 7. In some semiconductor systems and intermetallic compounds, the high concentration of defects can lead to small amorphous zones.

Table 2. Threshold displacement energies for several elements or compounds.

Material	Energy (eV)
Al	16
Si	13
Cu	21
Ge	14
GaAs	10
Ni	23
Mo	33
Au	35

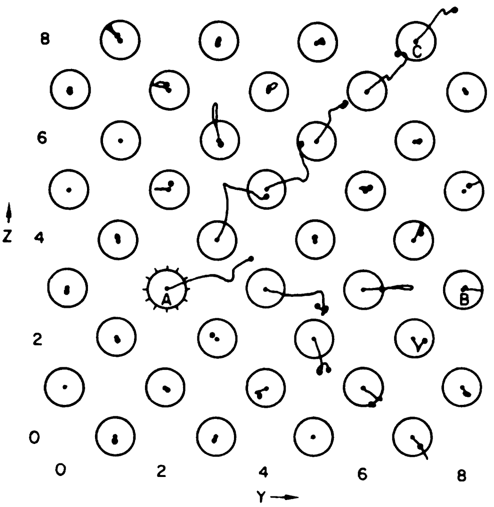

FIG. 6. Computer simulations illustrating replacement collision sequences in Cu (after Gibson *et al.*, 1960).

The number of Frenkel pairs that are produced in a cascade, $\nu_{FP}(T)$, can be estimated from the "modified Kinchin-Pease" expression (1955),

$$\nu_{FP}(T) = 0.8E_D(T)/2E_d, \tag{9}$$

where $E_D(T)$, the damage energy, is the total cascade energy T less the energy lost to electronic excitations. [The original Kinchin-Pease expression was $\nu_{FP}(T) = T/2E_d$.] Computer simulations, such as TRIM, are convenient for obtaining both the depth distribution and total number of defects produced by ion implantation.

2.3 Thermal Spikes

Although some of the kinetic energy of the implanted ion is lost to electronic excitations and some is stored as potential energy in the Frenkel pairs, most of the ion energy is dissipated in the lattice as heat. For 100-keV As

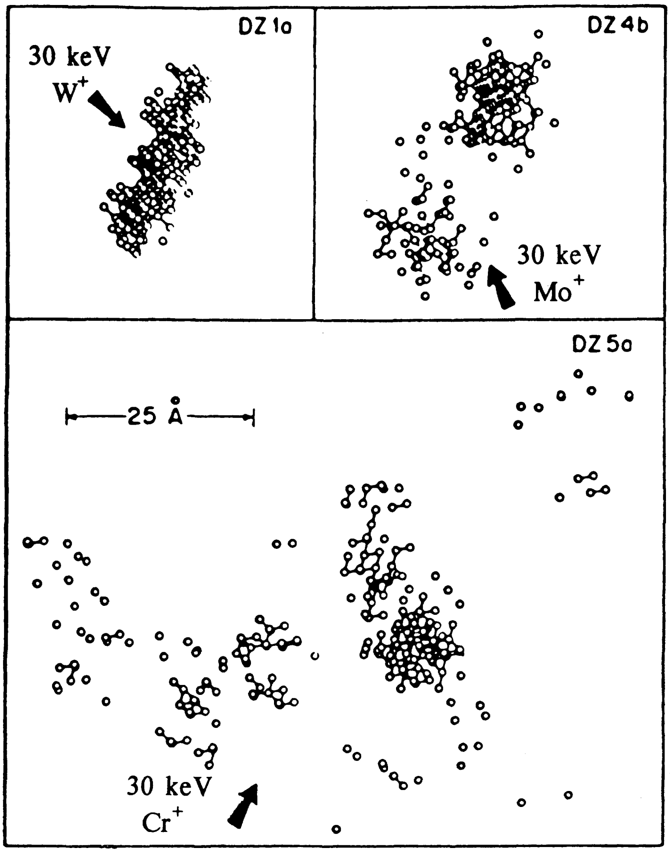

FIG. 7. Distribution of vacancies produced in displacement cascades of W. As the projectile mass increases, the density of defects also increases because of the increasing value of $S_n(E)$. Arrows indicate the initial directions of the incoming ions. (After Wei *et al.*, 1981.)

implantation of Ge, for example, 40 keV of energy is lost to electronic excitation, ≈2 keV is stored in defect energy, and the remainder, 58 keV, is converted to heat. The localization of such a large amount of energy in the small volume of a cascade can result in energy densities of several eV/atom. A long-standing question in the field of atomic collisions in solids has been whether this large localization of energy persists sufficiently long for thermal processes to evolve. The most vivid evidence for thermal spikes has derived from molecular-dynamics computer simulations. Figure 8, for example, shows the collision dynamics of a 10-keV Au ion impacting a Au surface (Ghaly and Averback, 1993). Clearly illustrated in Fig. 8(a), which shows a snapshot of the atoms in the vicinity of a cascade early in the event, are a shock wave emanating from the cascade center, replacement collision sequences, and a highly disordered cascade core. Shown in Fig. 8(b), which is a snapshot of the atoms some time later, is a well-defined local melt zone in the near surface; the melt is found to persist for several picoseconds. In materials with higher melting temperatures and smaller atomic numbers, the effects of thermal spikes are smaller.

2.4 Applications of Ion Implantation

The number of applications of ion implantation has grown enormously in the past decade, and they cannot be categorized here. A few examples, however, serve to illustrate the diversity of materials problems for which ion implantation has been successful.

2.4.1 Doping of Semiconductor Devices. The use of implantation for the doping of semiconductors has now been an industrial process for over 20 years, and a number of different devices include implantation as a processing step (see e.g., Ryssel and Ruge, 1986). As outlined above, implantation of semiconductors causes damage as well as increasing the dopant concentration, and damage generally degrades the electrical proper-

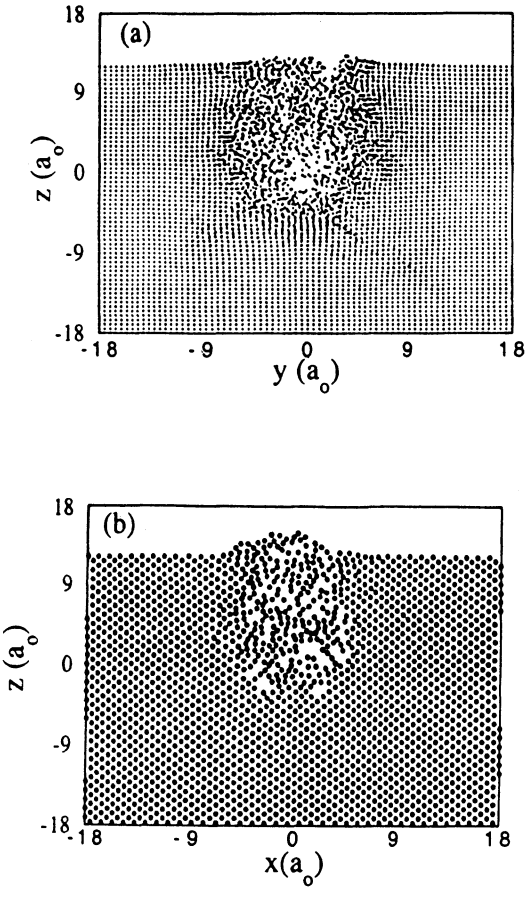

FIG. 8. Snapshots of atomic positions at two instants of time, 1.0 and 9.8 ps after a 10-keV Au atom impinges on a Au substrate. For clarity, only atoms in a cross-sectional slab containing four (100) lattice planes are shown. (a) A shock wave propagating from the cascade core, trails of replacement sequences, and atomic disorder are observed early in the event, while (b) local melting is observed sometime later. (After Ghaly and Averback, 1993.)

ties of the device. Moreover, the dopants do not always lie on substitutional sites in the lattice when initially implanted, and therefore they may not be electrically active. Consequently, annealing treatments are necessary to remove the damage and to activate the impurities. In Si, the damage can consist of defect clusters and dense dislocation networks. This type of damage is very resistant to annealing and typically requires annealing temperatures of 950 °C. Such high-temperature anneals are unpractical for most devices because of problems of interdiffusion. In many cases, however, the irradiation renders the implanted region amorphous. This can be advantageous because crystallization occurs at a relatively low annealing temperature, less than 650 °C in Si. The crystallization begins at the amorphous–crystal interface located at the end of the implanted ion range, and it advances toward the surface by a process called solid-phase epitaxy (SPE). SPE leaves very few defects in the reformed crystal, while dopants find lattice sites and become activated. Rates of regrowth on various cystallographic planes in Si are illustrated in Fig. 9.

2.4.2 Buried Layers. One of the advantages of ion implantation over diffusion doping is that the peak impurity concentration can be located beneath the surface. An interesting application of this capability is the production of planar waveguides in $KNbO_3$ and $LiNbO_3$ using He implantation (Townsend, 1987). Because He ions are very light, their stopping in a material occurs primarily via electronic excitation until they reach the end of their range, where the nuclear stopping becomes larger (see Fig. 2). Consequently, the damage produced by He implantation is mostly

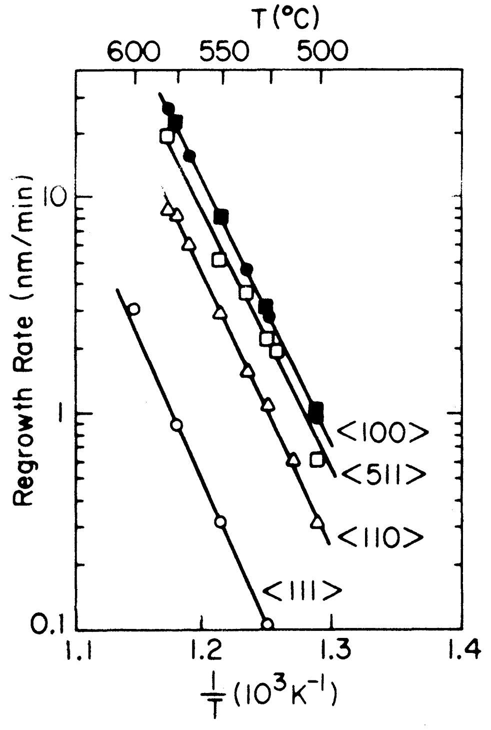

FIG. 9. The SPE regrowth rate at various amorphous-crystalline Si interfaces as a function of temperature (after Csepregi *et al.*, 1978).

confined to a narrow layer at a depth about equal to the projected range. This subsurface damage layer in $KNbO_3$ and $LiNbO_3$, in fact, becomes amorphous after sufficiently prolonged irradiation. Since the densities of these amorphous glasses are less than those of their corresponding crystalline phases, their indices of refraction are smaller. The buried amorphous layers thus create optical barriers or waveguides, with the thickness of the guiding region dictated by the energy of the He ion.

Buried layers are of interest for a variety of other applications as well. Co, for example, is implanted into Si to form a buried silicide layer, $CoSi_2$, a process called mesoepitaxy. These buried conductive layers serve as permeable base regions in transistor devices and interconnects for high-speed integrated circuits (Mantl, 1992). Oxygen is also used for implantation in Si to form buried insulator layers of SiO_2. These layers isolate the device from the silicon substrate. Consequently, they are useful for radiation hardening since the insulating layer prevents electron-hole pairs that are created by ionizing radiation anywhere in the thick Si substrate from entering the thin, active, device region at the substrate surface.

2.4.3 Tribological Applications. Materials selected for particular applications on account of such properties as strength, density, etc., often do not possess similarly attractive surface properties, such as oxidation resistance, wear resistance, or toughness. Ion implantation can be used to modify these surface properties without affecting the bulk material. Moreover, implantation has the advantages over surface plating and in-diffusion in that any material can be treated, any ion can be implanted, and discrete interfaces between the surface alloy and bulk material, which are susceptible to spalling, are avoided. One success of this method is illustrated in Fig. 10; it shows the improvement in sliding friction and wear resistance in 304 stainless steel resulting from implantation with Ti ions. The improved properties were attributed to the formation of a thin amorphous layer at the surface of the steel. Nitrogen implantation has been found effective for these purposes in a number of other metals, and because of its lower cost to implant, it is more commonly used than Ti (Rehn *et al.*, 1987).

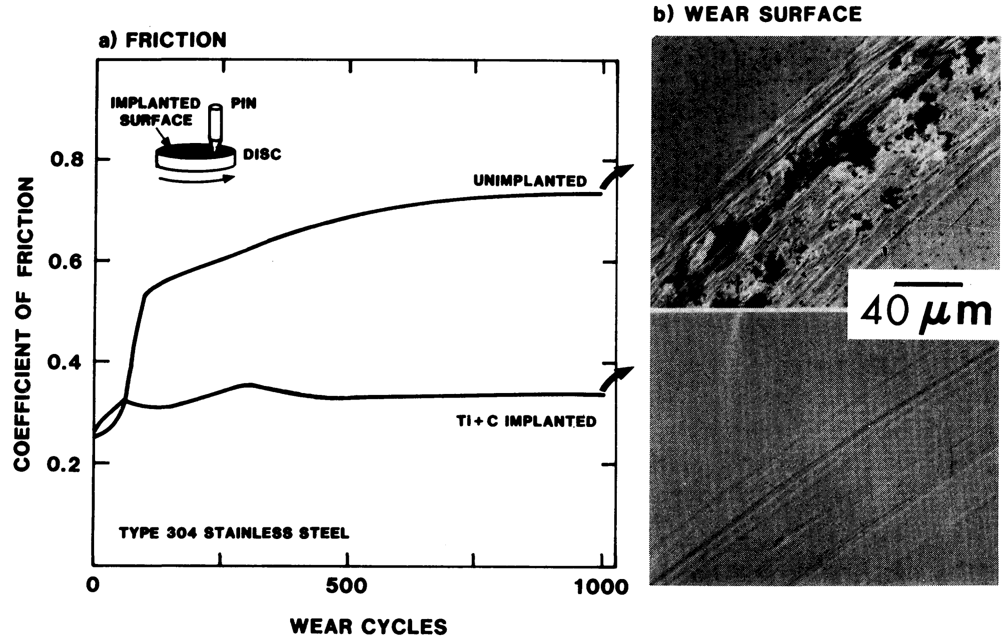

FIG. 10. Improvement of wear resistance in 304 stainless steel due to implantation with Ti and C. Doses of 2×10^{17} cm^{-2} 90–180-keV Ti and 2×10^{17} cm^{-2} 50-keV C were employed. (a) Coefficient of friction versus time (wear cycles); (b) scanning electron micrograph of the wear tracks. (After Follstaedt *et al.*, 1989.)

3. ION-BEAM MIXING

The large number of ions set into motion in a displacement cascade gives rise to a "mixing" of atoms in the solid. Three distinct mechanisms of mixing can be distinguished: recoil implantation, cascade mixing, and thermal spike mixing; these are schematically illustrated in Fig. 11.

3.1 Mechanisms of Ion-Beam Mixing

3.1.1 Recoil Implantation. The first target atom impacted by an implanted ion necessarily recoils in the forward direction, deeper into the solid. This process of recoil implantation can be important during ion implantation. For example, oxygen can be driven from a thin surface layer deeper into a target and degrade electrical properties of thin devices. Or, the stoichiometry of compound targets can be altered by preferential recoil implantation of one of the components. This is most significant in targets like InP or $LiNbO_3$ where the atomic numbers (and collision cross sections) of the components are substantially different. For a thin surface layer, the distribution of recoil implantation atoms can be approximated by the expression

$$S(x)=N^{2/3}\int dT\,\frac{d\sigma(E,T)}{dT}\,F(x-x',\theta), \qquad (10)$$

where $F(x-x',\theta)$, the range distribution of a recoil atom, can be obtained from Eq. (7) with a correction for scattering angle θ relative to the surface normal.

3.1.2 Cascade Mixing. After several collisions, the momenta of the implanted ion and recoil atoms become randomized so that the relocation of most displaced atoms is a stochastic process. Calculations of the mixing due to the random relocation of atoms is far more difficult than those for recoil mixing as they require solving the Boltzmann transport equation (Sigmund and Gras-Marti, 1981). Accurate results, however, can be obtained more simply using computer simulations based on TRIM. The parameter Q is often employed for quantifying ion-beam mixing; it is defined as

$$Q=\langle\Delta x^2\rangle/2\phi F_D, \qquad (11)$$

where $\langle\Delta x^2\rangle$ is the variance of the mixing profile, ϕ is the ion fluence, and F_D is the deposited damage energy per unit length normal to the surface. The parameter Q is a measure of mixing per unit damage energy per unit volume and thus is normalized for all ion–solid combinations. Calculated values for the mixing parameter are in the range $Q=1$–7 Å^5/eV. Table 3 shows measured values of Q. It is observed that the mixing parameter is much larger than the values calculated by cascade

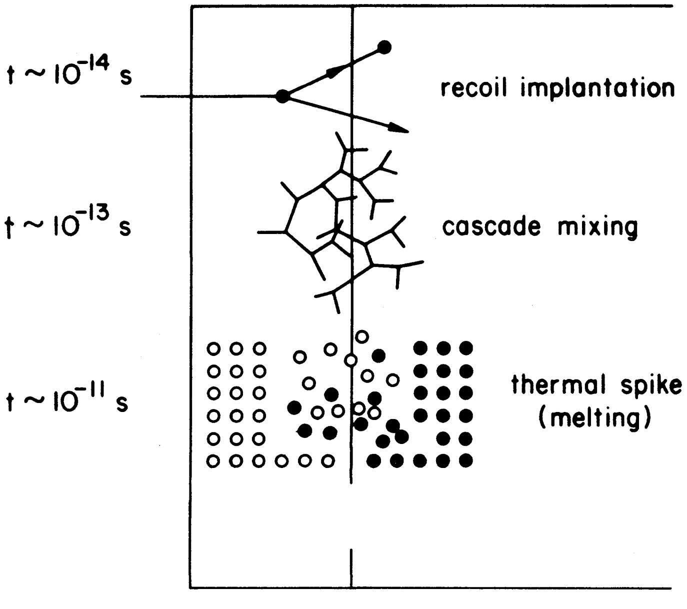

FIG. 11. Schematic illustration of the three contributions to ion-beam mixing: recoil implantation, cascade mixing, and thermal spike mixing.

Table 3. Mixing parameter Q in several metals.

Metal	Q (Å^5/eV)
Al	20
Ti	4
Ni	7
Cu	25
Ag	75
Pd	12
Au	120
Pt	19

mixing or recoil implantation for many metals and semiconductors.

3.1.3 Thermal-Spike Mixing. The possibility of local melting in a cascade provides another contribution to atomic mixing. Calculations of thermal-spike mixing have been performed using the equation

$$\langle \Delta x^2 \rangle = 2 \int \frac{d\sigma(E,T)}{dT}\, dT \int NV(T,t)D(T,t)\, dt, \tag{12}$$

where $V(T,t)$ and $D(T,t)$ are the instantaneous volume and diffusion coefficient of the hot cascade. (The outer integral sums over all the possible recoils that an ion makes with target atoms.) Since the diffusion coefficient of liquids is much larger than that of solids, most of the thermal-spike mixing takes place in the melt. Uncertainties in the cooling rates and appropriate diffusion coefficients make Eq. (12) difficult to evaluate accurately, but it nevertheless shows that mixing in the thermal spike is large in metals that have low melting temperatures and high energy densities (high atomic number). Values of the mixing parameter listed in Table 3 are consistent with these predictions. Johnson *et al.* (1985) have shown that because the characteristic energy of thermal-spike mixing is some tenths of electron volts, in contrast to the much higher energies of the implanted ion and recoil and cascade mixing, thermodynamics can sometimes play a role. This is a fundamentally new concept in the theory of ion–solid interactions, and it has helped significantly in understanding many details of ion-beam modifications of materials.

Calculations of the contributions to ion-beam mixing from thermal spikes cannot be obtained from simulation programs like TRIM, since these codes assume independent, two-body collisions and low densities of moving atoms. For thermal phenomena, however, the collective motion of atoms is vital. Recent studies of ion-beam mixing using molecular-dynamics computer simulations are beginning to clarify this process in metals (see Averback *et al.*, 1991), but the role of thermal spikes in ion-beam mixing in semiconductors and insulators is still poorly understood.

3.1.4 Phase Formation During Ion-Beam Mixing. The large number of atoms relocated during ion irradiation makes it possible to fabricate surface alloys and compounds far more efficiently than by direct ion implantation. The production of a 100-Å contact layer of Ni_2Si by direct implantation of Ni into Si requires an ion dose of over 2×10^{16} cm^{-2}. In comparison, a dose of only $\approx 1\times10^{15}$ cm^{-2} Xe ions is needed to mix a Ni film with a Si layer to produce a film of similar thickness. This technique has been applied to produce Ohmic contacts on semiconductor devices as well as surface alloys on metals for tribology applications.

An example of the use of ion-beam mixing to create metastable phases with specialized properties is illustrated by the mixing of a Cu/Bi diffusion couple. Cu and Bi are virtually immiscible in the solid state, so that CuBi alloys cannot be produced by conventional methods. Irradiation at low temperature, however, causes their intermixing and the formation of an amorphous phase as shown by the diffraction patterns in Fig. 12. The amorphous Cu-Bi phase has the interesting property of being superconducting at low temperatures, even though neither Cu nor Bi is itself a superconductor. On warming, however, the constituent metals precipitate and the superconductivity is lost.

3.2 Ion-Beam–Assisted Deposition

A variation of the ion-beam mixing process of growing technological significance is ion-beam–assisted deposition (IBAD), for which a low-energy ion beam is focused onto the specimen surface during the growth of a thin film. IBAD is employed for two distinct purposes:

1. creating a mixed interfacial layer between the substrate and the thin film to improve film adhesion;
2. improving the quality, stoichiometry, or

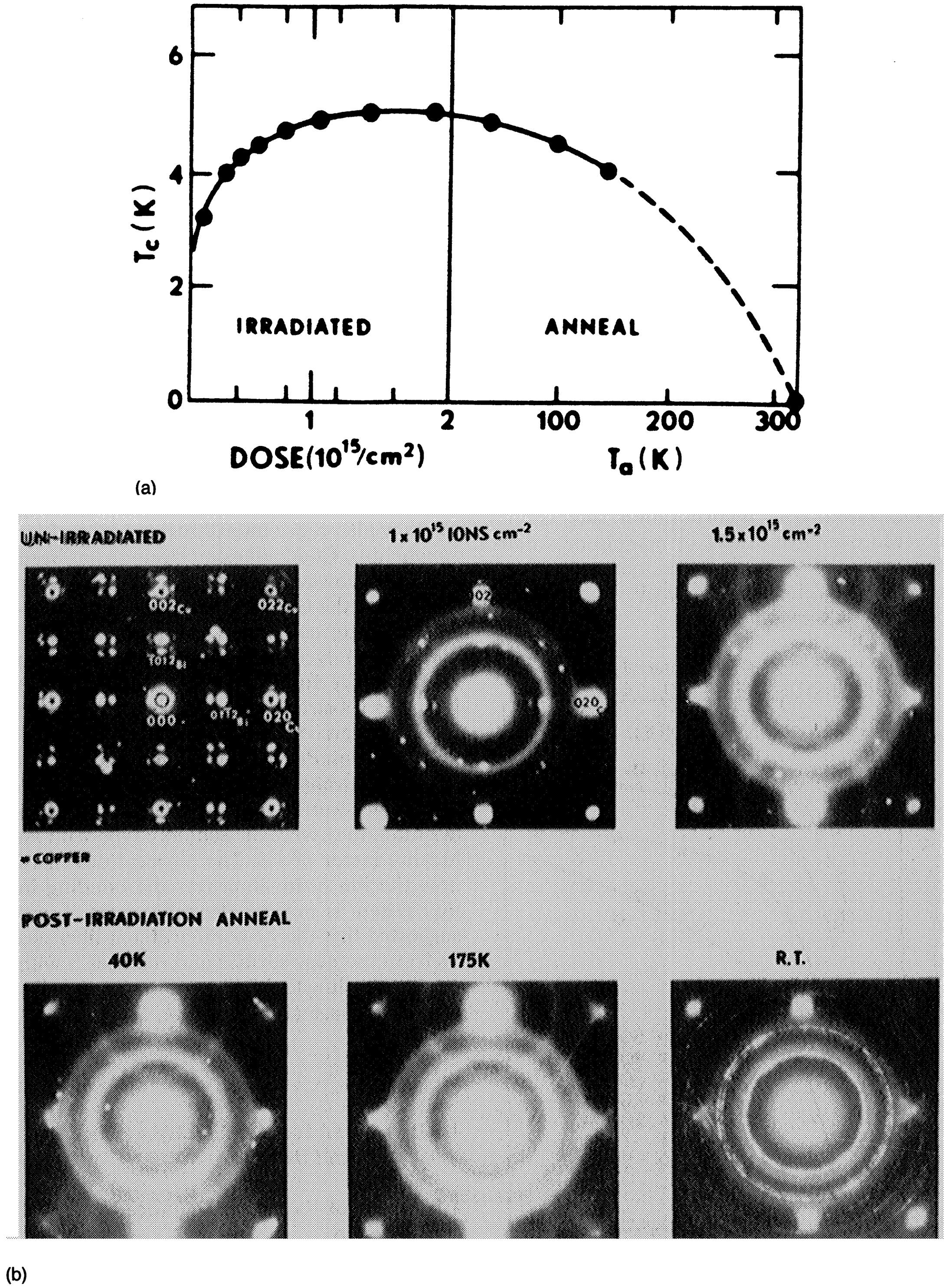

FIG. 12. Ion-beam mixing of Cu/Bi produces an amorphous alloy that is superconducting. (a) The variation of the critical temperature with 380-keV Kr fluence and annealing temperature; (b) electron diffraction patterns illustrating the amorphization during irradiation and subsequent crystallization during annealing. (After Averback *et al.*, 1985.)

epitaxy of the growing thin film.

A variation of the second process is direct ion-beam deposition (IBD), for which a high-intensity, low-energy ion-beam source is employed but without simultaneous vapor deposition. Two examples illustrate these purposes.

3.2.1 Dopant Incorporation. The synthesis of devices by thin-film deposition, e.g., molecular-beam epitaxy (MBE), is often limited by the difficulty of incorporating electrically active impurities in the growing film. As, Sb, In, and Ga are difficult to incorporate into Si, while Zn, Cd, Mn, S, and Te are difficult to incorporate into GaAs. The use of accelerated dopant beams during growth is proving to be a practical solution to this problem (Greene *et al.*, 1989). Shown in Fig. 13 is the incorporation probability of In in a growing Si film, plotted as a function of inverse temperature, for various energies of the In beam (Hasan *et al.*, 1989). The incorporation rate is increased over five orders of magnitude when the energy of the In beam is increased from thermal energies to over 50 eV. In addition, the incorporation was near unity for beam energies over 200 eV for growth temperatures up to 850 °C. These investigations also showed that the In atoms were located on substitutional sites and that they were electrically active for concentrations as high as 2×10^{18} cm^{-3}, which is well above the solubility limit. No evidence of radiation damage was observed in these experiments, and carrier (hole) mobilities were higher than those obtained using implantation and high-temperature annealing (Hasan *et al.*, 1989).

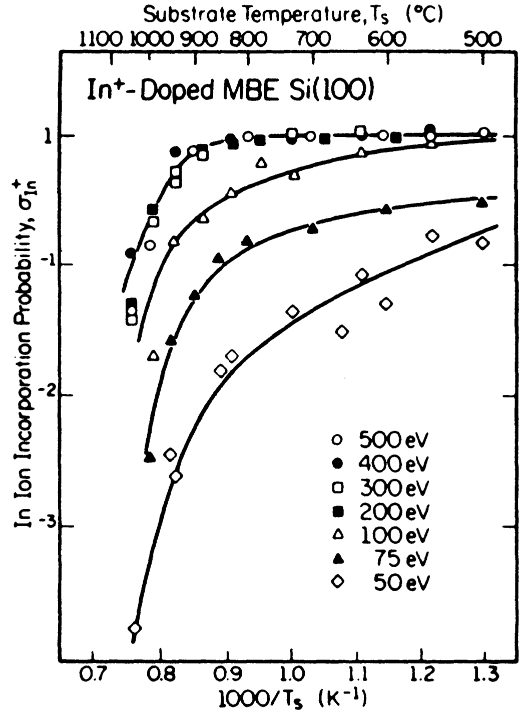

FIG. 13. The incorporation probability of In ions in MBE-grown Si as a function of the In energy and substrate temperature (after Hasan *et al.*, 1989).

3.3.2 Improved Adhesion. The fabrication of Schottky contacts that have good thermal stability, interface integrity, and adhesive strength on semiconductor devices poses several problems. In principle, Au should make an excellent Schottky barrier on GaAs; however, conventional deposition of Au on GaAs does not provide adequate bonding. High-temperature anneals improve the bonding, but they also cause interdiffusion and surface roughening. Good adhesion strength has been obtained, however, by directing a 300-eV Ar beam onto the GaAs surface during the deposition of Au, as shown in Fig. 14 (Kellock *et al.*, 1991). Here, the peel strength of the Au film, i.e., the force needed to peel the film from the substrate, is shown as a function of the Ar:Au arrival rate for several substrate temperatures. Far better adhesion is observed in the film prepared by beam-assisted growth. These films were stable at least up to 250 °C; in addition, they were all Schottky type with an ideality factor of 1.3. The physical details of how the ion beam improves the bonding in this system is not yet clear; however, it was suggested that the Ar bombardment liberates As from the near surface and replaces it with Au, thus leading to the formation of a complex Au-Ga-As phase (Kellock *et al.*, 1990).

4. IMPLANTATION AT ELEVATED TEMPERATURE

4.1 Theory of Radiation-Enhanced Diffusion

It is well established that diffusion in solids involves the motion of point defects—vacancies and interstitials. This is expressed through the equation

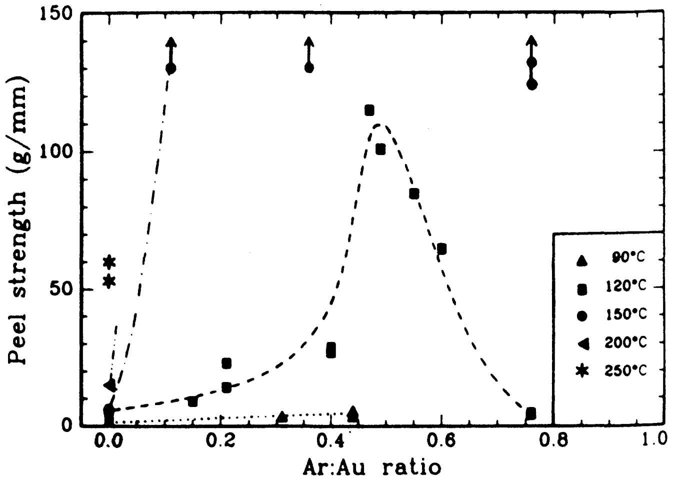

FIG. 14. The peel strength of a Au film vapor deposited on GaAs plotted as functions of the ratio of arrival of 300-eV Ar ions and Au atoms and temperature (after Kellock *et al.*, 1991).

$$D = D_i C_i + D_v C_v \tag{13}$$

where D_j and C_j are diffusivities and concentrations of vacancies and interstitials. The diffusivities have the form $D_j = D_{0j} \exp\{\Delta H_j / kT\}$, and ΔH_j is the activation enthalpy for defect migration; D_{0j} is a preexponential, or entropic, factor. In absence of irradiation, the concentrations of defects refer to their values in thermal equilibrium. However, during implantation, the displacement processes described above introduce large supersaturations of point defects, both interstitials and vacancies, and according to Eq. (13), they enhance diffusion. The instantaneous concentration of defects can be calculated using chemical rate theory (see, e.g., Sizmann, 1978). The diffusion coefficient that derives from this theory is plotted in Fig. 15 as a function of inverse temperature. At low temperatures, the defects are nearly immobile so that the radiation-enhanced diffusion (RED) becomes smaller than ion-beam mixing. At very high temperatures, normal thermal diffusion prevails. At intermediate temperatures, however, diffusion is controlled by RED. The temperature dependence of RED varies depending on a number of factors: temperature, ion flux, and the perfection of the crystal. The important point, however, is that RED occurs at substantially lower temperatures than normal thermal diffusion. Moreover, it is distinguished from ion-beam mixing in that

1. it can be orders of magnitude larger than ion beam mixing;
2. it is a "near" equilibrium process. This means that defects tend to anneal out during the irradiation and that equilibrium, rather than metastable, phases are usually produced.

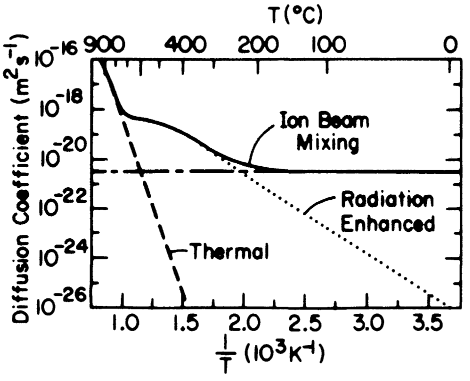

FIG. 15. Radiation-enhanced diffusion coefficient as a function of inverse temperature.

4.2 Examples of Radiation-Enhanced Diffusion

It may be expected from the current discussion that implantation at elevated temperatures may result in defect-free material and activated impurities in semiconductor devices. In some cases, doping of semiconductors is performed at elevated temperatures, but because of the enhanced diffusion of the impurity, precipitation, and the buildup of complex damage structures, this method is not widely used. Rather, rapid thermal annealing is employed to remove defects in implanted semiconductors. A more promising application of high-temperature implantation is the synthesis of buried $CoSi_2$ in Si discussed in Sec. 2.4.2. To produce good mesoepitaxial layers, it is imperative that the Si substrate remain crystalline during the Co implantation, otherwise the coherence between the silicide and Si substrate would be lost. Consequently, the Co is implanted at a temperature above which amorphous Si crystallizes, $\approx$300 °C. In addition, RED provides sufficient atomic mobility for the implanted Co to precipitate as an epitaxial layer of $CoSi_2$. The quality of these buried silicide layers is illustrated in Figs. 16 and 17. In Fig. 16(a), the cross-sectional transmission electron microscope image shows the flatness of the buried layer, while Fig. 16(b) reveals the atomic perfection at the interfaces. Figure 17 shows the effect of substrate temperature on crystal perfection. Here channeling of He ions is employed as a measure of the crystal quality (low yields implying little damage). For implantation below $\approx$300 °C the surface region becomes polycrystalline, even after high-temperature annealing. At too-high implantation temperatures, the crystal quality also diminishes since the $CoSi_2$ precipitates grow too rapidly and become three dimensional, rather than two-dimensional, flat films. Thus, careful control of RED is critical to this emerging technology.

A second example of RED is the formation of $NbSi_2$ using Ar irradiation at high temperatures. In Fig. 18 it is shown that $NbSi_2$ can be efficiently formed by ion-beam mixing at temperatures less than $\approx$300 °C. The formation of

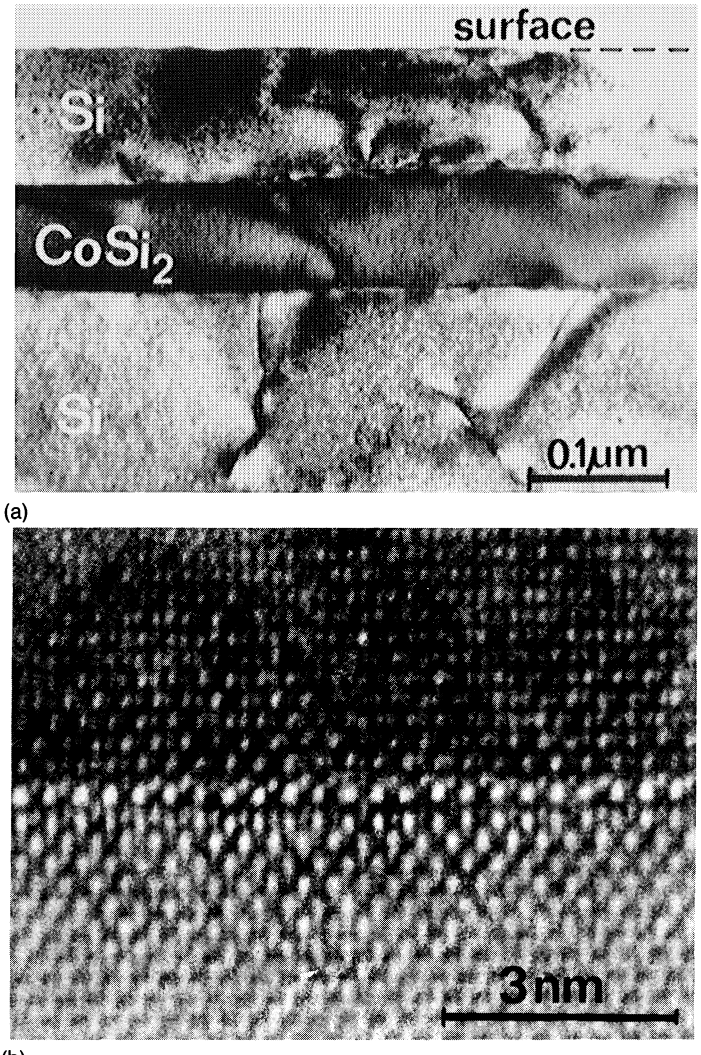

FIG. 16. Perfection of mesoepitaxial layer of $CoSi_2$ in Si. (a) Cross-sectional TEM image showing the flatness of the buried layer; (b) high-resolution TEM image showing the atomic perfection of the interfaces between the $CoSi_2$ and Si interfaces. (After Werner and Jaeger, 1993.)

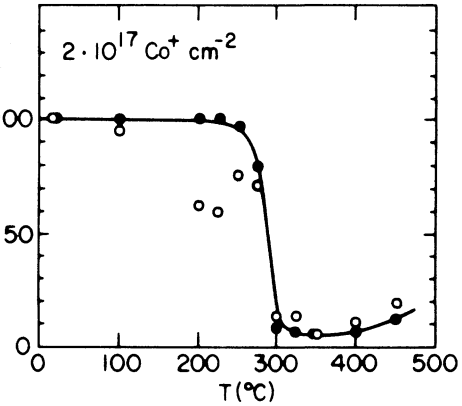

FIG. 17. The effect of implantation temperature on mesoepitaxial perfection. The quality of the film is characterized here by ion channeling measurements. A large yield of backscattered ions means that there is a large number of crystalline imperfections blocking the channels. (After Mantl, 1992.)

$NbSi_2$ by thermal annealing, however, does not take place until ≈450 °C. Ion-beam mixing of Nb and Si at low temperatures requires much higher fluences, and it does not yield the $NbSi_2$ phase.

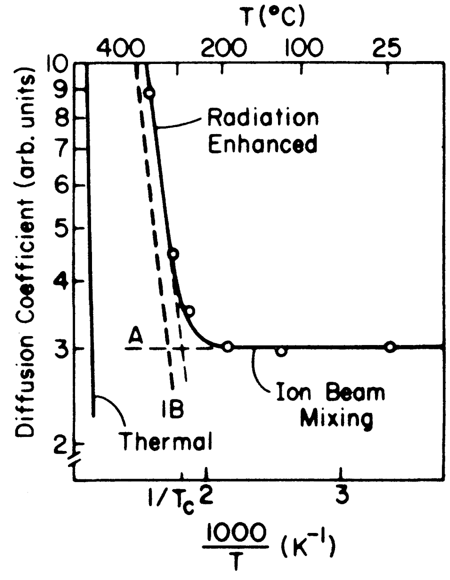

FIG. 18. Logarithmic plot of the intermixing of Nb and Si films to form $NbSi_2$ during Si irradiation as a function of inverse temperature. At low temperatures the intermixing is due to ion-beam mixing and is independent of temperature. At higher temperatures, the mixing is due to radiation-enhanced diffusion. Thermal diffusion occurs at much higher temperatures. (After Matteson *et al.*, 1979.)

5. SUMMARY

In recent years, ion-beam modification of materials has become an important processing step for a number of advanced materials. This development has become possible after many years of fundamental research on the interaction of energetic particles with solids and defects in crystals. Computer simulation programs are now readily available for accurately calculating the depth distributions of implanted ions, even in complex substrates. These programs can also reliably calculate sputtering yields and defect production. The damage created by the implantation is also becoming better understood, and along with this better understanding, methods to eliminate it, or to exploit it, are being developed. In this article, the fundamental processes involved in ion-beam modification were presented. Several applications were described that illustrate how various aspects of implantation can be utilized. These ranged from the doping of semiconductors, to ion-beam–assisted growth of thin films, to the formation of metastable phases at metal and ceramic surfaces for wear and oxidation resistance. In this age of advanced materials, where processing must operate on the submicron level, ion-beam modification of materials is certain to play an ever increasing role.

GLOSSARY

Channeling: The guidance of implanted atoms by lattice atoms to travel in the open spaces of axial or planar channels in a crystalline target.

Compositional Disordering: The intermixing of compositionally modulated structures.

Damage Energy: The fraction of an implanted ion's energy that is utilized in displacing atoms from their lattice sites. It is approximately the energy of the implantation particle less that dissipated in electronic excitations.

Displacement Energy: The energy required to displace an atom from its lattice site and create a Frenkel pair.

Frenkel Pair: An interstitial atom and its associated vacancy.

Ion-Beam–Assisted Deposition: The deposition of a thin film onto a surface while simultaneously bombarding the surface with ions.

Ion-Beam Mixing: The athermal rearrangement of atoms during ion irradiation. It is often associated with the intermixing of contiguous thin-film layers by ion irradiation.

Mesoepitaxy: The formation of a subsurface, coherent layer by ion implantation.

Radiation-Enhanced Diffusion: The enhancement of diffusion in a material at elevated temperatures by using irradiation to produce point defects.

Range Distribution: The concentration profile of an implanted species as a function of distance beneath the surface.

Recoil Implantation: The forward recoil of substrate atoms deeper into a substrate during ion implantation.

Replacement Collision Sequence: A linear chain of collisions along the close-packed directions of a lattice whereby one atom collides with and replaces its neighbor.

Sputtering: The erosion of a surface during ion irradiation. The sputtering yield is the number of target atoms ejected from the surface per incident ion.

Stopping Power: The loss of energy of an implanted ion per unit distance along its trajectory in a material.

Straggling: A measure of the variance of a range distribution. The straggling normal to the surface is the ratio of the square root of the variance to the mean range. The transverse straggling is the ratio of the variances in directions parallel and perpendicular to the surface.

Thermal Spike: Local, transient heating of a material around the trajectory of an implanted ion.

TRIM (Transport of Ions in Matter): A popular computer code for generating range damage and ionization profiles for ion implantation.

Lists of Work Cited

Averback, R. S., Okamoto, P. R., Baily, A. C., Strizker, B. (1985), *Nucl. Instrum. Methods. B* **7/8**, 556–560.

Averback, R. S., Diaz de la Rubia, T., Hsieh, H., Benedek, R. (1991), *Nucl. Instrum. Methods B* **59/60**, 709–717.

Biersack, J. P., Haggmark, L. G. (1990), *Nucl. Instrum. Methods* **174**, 257–269.

Csepregi, L., Kennedy, E. F., Mayer, J. W., Sigmon, T. W. (1978), *J. Appl. Phys.* **49**, 3906–3911.

Davies, J. A. (1992), *MRS Bull.* **17**, 26–29.

Firsov, O. B. (1959), *Zh. Eksp. Teor. Fiz.* **36**, 1517–1523 [*Sov. Phys. JETP* **36**, 1076–1080].

Follstaedt, D. M., Knapp, J. A., Pope, L. E. (1989), in; L. E. Rehn, J. Greene, F.A. Smidt (Eds.), *Processing and Characterization of Materials Using Ion Beams*, MRS Symposium Proceedings Vol. 128, Pittsburgh: Materials Research Society.

Ghaly, M., Averback, R. S. (1993) (unpublished).

Gibson, J. B., Goland, A. N. Milgram, M., Vineyard, G. H. (1960), *Phys. Rev.* **120**, 1229–1253.

Greene, J. E., Barnett, S. A., Sundgren, J. E., Rockett, A. (1989), in: T. Itoh (Ed.), *Ion Assisted Film Growth*, Amsterdam, Elsevier, p. 101.

Hasan, M. A., Knall, J., Barnett, S. A., Sundgren, J. E., Markert, L. C., Rockett, A., Greene, J. E. (1989), *J. Appl. Phys.* **65**, 172–179.

Johnson, W. L., Chang, Y. T., Van Rossum, M., Nicolet, M.-A. (1985), *Nucl. Instrum. Methods B* **7/8**, 657–665.

Kellock, A. J., Baglin, J. E. E., Bardin, T. T., Pronko, J. G. (1991), *Nucl. Instrum. Methods B* **59/60**, 249–253.

Kinchin G. H., Pease, R. S. (1955), *Rep. Prog. Phys.* **18**, 1–51.

Lindhard, J. L., Scharff, M., Schiøtt, H. E. (1963), *K. Dan. Videnskab. Selskab., Mat. Fys. Medd.* **33**, No. 10.

Mantl, S. (1992), *Mat. Sci. Rep.* **8**, 1–95.

Matteson, S., Roth, J., Nicolet, M.-A. (1979), *Radiat. Eff.* **42**, 217–226.

Piercy, G. R., Brown, F., Davies, J. A., McCargo, M. (1963), *Phys. Rev. Lett.* **10**, 399–400.

Rehn, L. E., Picraux, S. T., Wiedersich, H. (1987), *Surface Alloying by Ion, Electron, and Laser Beams*, Metals Park, OH: American Society for Metals.

Roas, B., Hensel, B., Henke, S., Klaumunzer, S., Kabius, B., Watanabe, W., Saemann-Ischenko, G., Schultz, L., Urban, K. (1990), *Europhys. Lett.* **11**, 669–674.

Robinson, M. T., Oen, O. S. (1963), *Phys. Rev.* **132**, 2385–2398.

Robinson, M. T., Torrens, I. M. (1974), *Phys. Rev. B* **9**, 5008–5024.

Ryssel, H., Ruge, I. (1986), *Ion Implantation*, Chichester: Wiley, p. 316.

Schiøtt, H. E. (1970), *Radiat. Eff.* **6**, 107–113.

Sigmund, P. (1969), *Phys. Rev.* **184**, 383–416.

Sigmund, P., Gras Marti, A. (1981), *Nucl. Instrum. Methods* **182/183**, 25–41.

Sizmann, R. (1978), *J. Nucl. Mater.* **69/70**, 386–412.

Spohr, R. (1990), *Ion Tracks and Microtechnology: Principles and Applications*, Braunschweig: Vieweg.

Stark, J. (1912), *Phys. Z.* **13**, 973–977.

Tamura, M. (1991), *Mat. Sci. Rep.* **6**, 1–214.

Townsend, P. D. (1987), *Rep. Prog. Phys.* **50**, 501–558.

Wei, C. Y., Current, M. I., Seidmann, D. N. (1981), *Philos. Mag.* **A44**, 459–491.

Werner, P., Jaeger, W. J. (1993), *J. Appl. Phys.* (to be published).

Winterbon, K. B. (1975), *Ion Implantation Range and Energy Deposition Distributions*, Vol. 2, New York: IFI/Plenum

Ziegler, J. F., Biersack, J. P., Littmark, U. (1985), *The Stopping and Range of Ions in Solids*, Vol. 1, New York: Pergamon.

Further Reading

The following books provide both a background to the theory of ion–solid interactions and many detailed examples:

Auciello, O., Kelly, R. (Eds.) (1984), *Ion Beam Modifications of Materials, Ion Beam Modifications of Surfaces*, Amsterdam: Elsevier. Sputtering applications.

Itoh, T. (Ed.) (1984), *Ion Beam Assisted Deposition*, Amsterdam: Elsevier. Beam-assisted deposition.

Ryssel, H., Ruge, I. (1986), *Ion Implantation*, Chichester: Wiley. Doping of semiconductors.

Townsend, P. D., Kelly, J. C., Hartley, N. E. W. (1976), *Ion Implantation, Sputtering and their Applications*, London: Academic. Sputtering applications.

Several international conference series are dedicated to ion-beam modifications of materials: International Conference on Ion-Beam Modification of Materials, Ion-Beam Analysis, Atomic Collisions in Solids, Ion-Implantation Technology. All meet biannually and their proceedings are published in *Nuclear Instruments and Methods, Part B.* In addition, several symposia of the Materials Research Society focus on ion-beam modification of materials, and their proceedings are published in the MRS Symposium Proceedings.

ION AND ELECTRON SOURCES

See SOURCES, ION AND ELECTRON

ION SPECTROMETERS

See SPECTROMETERS, ELECTRON AND ION

IONIC CONDUCTION AND DIFFUSION IN SOLIDS

A. V. CHADWICK, *Chemical Laboratory, University of Kent, Canterbury, Kent, United Kingdom*

Introduction: General Background to Diffusion in Solids.
1. **Theory of Solid-State Diffusion** .... ?
1.1 Phenomenological Theory of Diffusion..................... 194
1.1.1 Classical Approach 194
1.1.2 Irreversible Thermodynamics . 195
1.2 Atomistic Theory of Solid-State Diffusion.......... 196
1.2.1 Introduction to Point Defects.. 196
1.2.2 Thermodynamics of Point Defects in a Monatomic Solid . 197
1.2.3 Point-Defect Thermodynamics in an Ionic Solid 199
1.2.4 Point-Defect Concentrations in Solids........................ 202
1.2.5 Mechanisms of Diffusion...... 202
1.2.6 Random-Walk Theory 203
1.2.7 Relations between Macroscopic and Microscopic Parameters................... 204
1.2.8 Correlation Effects in Diffusion..................... 204
2. **Experimental Methods of Studying Diffusion**........... 205
2.1 Methods of Measuring Diffusion Coefficients 205
2.1.1 Study of the Diffusion Penetration Profile 205
2.1.2 Measurement of the Atomic Jump Frequency 206
2.2 Ionic Conductivity Measurements................ 207
2.3 Methods of Identifying Diffusion Mechanisms 211
2.4 Computational Techniques.... 212
3. **Diffusion in Different Solid Types** 213
3.1 Metals and Alloys 213
3.2 Molecular Solids 214
3.2.1 Rare-Gas Crystals............. 214
3.2.2 Plastic Crystals............... 215
3.2.3 Nonplastic Crystals........... 215
3.3 Covalent Solids............... 215
3.4 Ionic Solids.................. 215
3.4.1 Normal Ionic Solids.......... 216
3.4.2 Fast-Ion Conductors 217
4. **Diffusion in an Applied Field**. 218
5. **Summary**.................... 219
Glossary.................... 219
Works Cited 220
Further Reading 221

INTRODUCTION: GENERAL BACKGROUND TO DIFFUSION IN SOLIDS

The classical "textbook" picture of a solid is that the atoms in a solid stay located on their lattice sites, and the only motions that are present are the lattice vibrations. In fact, atoms in a real solid are continually changing their positions, and at temperatures close to the melting point an atom will change site about a million times per second. The study of these translational atomic movements, *diffusion*, in solids is a mature field of chemical physics with a long and rich history. In his pioneering studies of the electrical behavior of materials in the 1830s, Michael Faraday measured the conductivity of a number of ionic solids. It was almost a century ago that Sir William Roberts-Austen (a student of Thomas Graham, who formulated the laws of diffusion in gases) made the first diffusion measurement on a metallic system, gold in lead. Since those early days there has been a steady growth of interest in all aspects of diffusion in solids. From the viewpoint of basic science, the ability to model diffusion is a severe test of the theoretical concepts of

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

physics, particularly the understanding of the interatomic potentials, and there will be a continual striving to predict accurately the diffusion parameters of solids. A vast number of technological processes involve diffusion phenomena, including the manufacture of semiconductor devices by doping with impurities, sintering, plastic deformation, and corrosion. It is therefore important that physical scientists understand the basic phenomena and theories of solid-state diffusion and ionic conductivity. The objective of this article is to present an overview of the current state of knowledge that emphasizes the principal features. Readers wishing more detail are referred to the extensive bibliography at the end of the chapter, which will guide them to several authoritative texts and articles on specific topics. The focus is on diffusion in crystalline solids where the basic phenomena and ideas are now well established. Although there have been extensive studies of amorphous materials and glasses, the theory of diffusion in these systems is still in a state of flux (Greaves, 1990).

The first section treats the basic theories of diffusion in solids, both the macroscopic, phenomenological aspects and the atomistic mechanisms. The next section covers the variety of experimental techniques. A special emphasis is given to the measurement of conductivity of ionic solids, as this has proved particularly fruitful in the development of many basic concepts of solid-state diffusion. Also included in this section is a brief treatment of the methods of computer simulation, a technique that is proving invaluable in understanding the details of the diffusion processes. The third section reviews the current knowledge of the diffusion processes in the different solid types. A final section briefly introduces the difficult, but important, topic of diffusion in an applied field, i.e., a gradient of temperature (*thermomigration*) or of electric field (*electromigration*).

1. THEORY OF SOLID-STATE DIFFUSION

1.1 Phenomenological Theory of Diffusion

1.1.1 Classical Approach. Translational diffusion in classical physics is governed by *Fick's laws*. In a one-dimensional system the flux of a given type of particle, i.e., the number passing through unit cross-sectional area in unit time, can be expressed as

$$J=-D\left(\frac{\partial c}{\partial x}\right), \tag{1}$$

where c is the number concentration of particles, $\partial c/\partial x$ is the gradient of the concentration of particles along the x direction, and D is the diffusion coefficient. A dimensional analysis shows that D has units of L^2T^{-1} and is usually reported as $\mathrm{cm^2\,s^{-1}}$ or $\mathrm{m^2\,s^{-1}}$. Equation (1) is *Fick's first law*. In three dimensions this can be written as

$$\mathbf{J}=-\mathbf{D}\cdot\nabla c, \tag{2}$$

where D is now a second-rank tensor.

Fick's first law is analogous to Fourier's law of heat flow and Ohm's law and is a specific statement of the more general law of physics that a flow is directly proportional to a driving force, the driving force being a gradient of "potential." In the form of Eqs. (1) and (2), Fick's first law is not particularly useful to the experimentalist, as the evaluation of D would involve the measurement of a flux and a concentration gradient, both of which could be time dependent. If the particles are conserved, then the equation of material balance will apply, and therefore a further condition must be obeyed, i.e.,

$$\frac{\partial J}{\partial x}=-\frac{\partial c}{\partial t}. \tag{3}$$

This continuity equation is simply derived by considering the net flow into and out of a small volume along the diffusion axis. Combining Eqs. (1) and (3) yields

$$\frac{\partial c}{\partial t}=\frac{\partial}{\partial x}\left(D\frac{\partial c}{\partial x}\right). \tag{4}$$

If D is independent of concentration, then Eq. (4) reduces to

$$\frac{\partial c}{\partial t}=D\frac{\partial^2 c}{\partial x^2}, \tag{5}$$

which is often referred to as *Fick's second law*.

Equation (5) is a second-order differential equation to which a variety of solutions can be derived to correspond with the different experimental situations, i.e., different boundary conditions. These can be found in the standard texts (e.g., Crank, 1975), and the only solution quoted here is that for the initial condition of an *infinitely* thin source on the surface of an *infinite* sample. This geometry is commonly used in tracer diffusion experiments, and the appropriate solution to Eq. (5) is

$$c(x,t)=\frac{A}{\sqrt{\pi Dt}}\exp\left(-\frac{x^2}{4Dt}\right). \tag{6}$$

Here $c(x,t)$ is the concentration of particles at a penetration depth x into the surface after time t, and A is the initial concentration of particles on unit area of the surface. Thus the penetration profile, the plot of $c(x,t)$ vs x, will have a Gaussian shape, and the diffusion coefficient can be determined from the plot of $\ln c(x,t)$ vs x^2; i.e., the slope is $-1/4Dt$.

External forces acting on the system will perturb the diffusion, and intuitively Eq. (1) may be modified by including an additional flux. This term is referred to as a *drift* and can be simply equated to the concentration of particles multiplied by an average drift velocity, $\bar{v}$, imparted by the force. Thus the resulting total flux can be expressed as

$$J=-D\left(\frac{\partial c}{\partial x}\right)+\bar{v}c. \tag{7}$$

Examples of external driving forces are gradients of electrical potential, temperature, and stress. For particles of charge q in a gradient of electric potential E $(=-d\Phi/dx)$, the drift velocity is given by

$$\bar{v}=\mu E, \tag{8}$$

where μ is the mobility of the charged particle, i.e., the drift velocity in unit potential gradient. If the system is in a steady state with no net flux, then from Eq. (7)

$$D\left(\frac{\partial c}{\partial x}\right)=\bar{v}c=\mu Ec. \tag{9}$$

Assuming thermodynamic equilibrium in the system, the concentration of particles will obey Boltzmann statistics, and hence

$$c(x)=a\exp\left(\frac{-q\Phi(x)}{kT}\right), \tag{10}$$

where a is a constant, k is the Boltzmann constant, and T is the temperature. Differentiating Eq. (10) with respect to x yields

$$\frac{\delta c}{\delta x}=\frac{-cq}{kT}\frac{d\Phi}{dx}=\frac{cqE}{kT}. \tag{11}$$

Combining Eqs. (9) and (11) gives the result

$$\mu/D=q/kT; \tag{12}$$

and since the conductivity σ is defined by

$$\sigma=cq\mu, \tag{13}$$

the final result relating D and σ is

$$\sigma=cq^2D/kT. \tag{14}$$

Equation (14) is the *Nernst–Einstein equation* and is important in the study of atomic transport in ionic solids, as it provides a quantitative link between diffusion and conductivity experiments. This is discussed later in greater detail.

The treatment of diffusion in a temperature gradient, where the driving force is dT/dx, is not so straightforward. In the format of Eq. (7) the flux is given in this case by

$$J=-D\left(\frac{\partial c}{\partial x}\right)-\frac{cQ^{*\prime}}{kT^2}\frac{dT}{dx}. \tag{15}$$

Here $Q^{*\prime}$ is the reduced heat of transport and has the units of energy. By convention $Q^{*\prime}$ is positive if the particle moves down the temperature gradient. The meaning of $Q^{*\prime}$ is further explained in the treatment given below.

1.1.2 Irreversible Thermodynamics. The treatment of diffusion theory presented above is based on empirical and intuitive approaches to the phenomenon. A rigorous approach to the problem is possible using *irreversible* (or *nonequilibrium*) *thermodynamics*. Only the basic ideas can be treated here, and for more detail the reader is referred to standard texts (de Groot and Mazur, 1962; Prigogine, 1967), while the special applications to solid-state diffusion are treated by Howard and Lidiard (1964).

In a system in which diffusion is occurring there will be local fluxes of atoms or species of type i, $\mathbf{J}_i$, and there will be a heat flux $\mathbf{J}_q$. The thermodynamics of irreversible processes relates these fluxes linearly to certain conjugate thermodynamic forces, $\mathbf{X}_i$ and $\mathbf{X}_q$, to give the relations

$$\mathbf{J}_i = \sum_{j=1}^{n} L_{ij}\mathbf{X}_j + L_{iq}\mathbf{X}_q \tag{16}$$

and

$$\mathbf{J}_q = \sum_{j=1}^{n} \mathbf{L}_{qj}\mathbf{X}_j + L_{qq}\mathbf{X}_q. \tag{17}$$

The L_{ij}'s are called the *phenomenological coefficients* and are independent of the forces (but will depend on concentration and temperature).

The conjugate forces are given by

$$\mathbf{X}_i = -T\nabla(\mu_i/T) \tag{18}$$

and

$$\mathbf{X}_q = T\nabla(1/T), \tag{19}$$

where μ_i is the chemical potential of species i.

The diagonal coefficients (L_{ii}) are simply related to atomic mobilities and thermal conductivities. The off-diagonal or cross coefficients express the coupling between the various fluxes and forces. For example, a heat flux will give rise to a diffusional flux and vice versa. An important point in the general theory is the *Onsager reciprocal relations* between the coefficients, i.e.,

$$L_{ij} = L_{ji} \tag{20}$$

and

$$L_{iq} = L_{qi}. \tag{21}$$

It is more convenient to use the *reduced heat flux* $\mathbf{J}'_q$ in which the flux of enthalpy due to the diffusion of species is subtracted from $\mathbf{J}_q$, i.e.,

$$\mathbf{J}'_q = \mathbf{J}_q - \sum_{i=1}^{n} \mathbf{J}_i h_i. \tag{22}$$

Here h_i is the partial specific enthalpy of species i. In choosing the reduced heat flux the conjugate thermodynamic force needs to be redefined as $\mathbf{X}'_i$, given by

$$\mathbf{X}'_j = -(\nabla\mu_i)_T. \tag{23}$$

Thus $\mathbf{X}'_i$ is the gradient of chemical potential produced by the variation of composition in the absence of a temperature gradient.

It is useful to define the *heat of transport* Q^*_i by the relation

$$L_{iq} = \sum_{j=1}^{n} Q^*_j L_{ij}. \tag{24}$$

Thus Eqs. (16) and (17) can be reexpressed as

$$\mathbf{J}_i = \sum_{j=1}^{n} L_{ij}(\mathbf{X}_j + Q^*_j\mathbf{X}_q) \tag{25}$$

and

$$\mathbf{J}_q = \sum_{i=1}^{n} Q^*_i\mathbf{J}_i + \left(L_{qq} - \sum_{j=1}^{n} L_{jq}Q^*_j\right)\mathbf{X}_q, \tag{26}$$

which in the absence of a temperature gradient reduces to

$$\mathbf{J}_q = \sum_{j=1}^{n} Q^*_j\mathbf{J}_j. \tag{27}$$

In the framework of reduced heat flux a *reduced heat of transport* $Q^{*\prime}_j$ is defined by

$$Q^{*\prime}_j = Q^*_j - h_i, \tag{28}$$

and then Eqs. (25) and (27) are

$$\mathbf{J}_i = \sum_{j=1}^{n} L_{ij}(\mathbf{X}'_j + Q^{*\prime}_j\mathbf{X}'_q) \tag{29}$$

and

$$\mathbf{J}'_q = \sum_{j=1}^{n} Q^{*\prime}_j\mathbf{J}_j. \tag{30}$$

Equation (30) provides a physical meaning of the reduced heat of transport as the flux of heat in an isothermal system due to the flux of matter.

It is important to note that the fluxes of all species in a system cannot be independent. Thus measuring the fluxes relative to the lattice frame at rest yields

$$\sum_{i=1}^{n} \mathbf{J}_i = 0, \tag{31}$$

so that if there are n species in a system there are only $n-1$ *independent* heats of transport. This is important when considering diffusion due to point defects.

1.2 Atomistic Theory of Solid-State Diffusion

1.2.1 Introduction to Point Defects. All real crystals are imperfect, and they contain a variety of types of defects. These include *bulk defects,* such as grain boundaries and dislocations, which extend over many lattice spacings, *surfaces*, and *point defects*, the latter being at the atomic level. The mechanisms of diffusion in crystalline solids that move an atom from one lattice site to another, *lattice* or *bulk diffusion*, involve the point defects. Diffusion will occur along the other types of defect—in fact it should occur more readily because of the open nature of their structures; however, it will not be covered here. It is more difficult to treat dislocation and surface diffusion in a detailed analysis, as the atomic pathway is less well defined, although there has been some progress in recent years (Kaur and Gust, 1988; King, 1987). In addition, these defects are not thermodynamically inherent in crystals, and their concentrations, and hence the contribution to the total diffusion, will depend on the history of the sample. At high temperatures, lattice diffusion will usually dominate.

In a monatomic solid, there are four basic types of point defect, as illustrated in Fig. 1. In an absolutely pure crystal there may be *vacancies* [Fig. 1(a)], unoccupied lattice sites, and *interstitials* [Fig. 1(b)], atoms on interstitial positions in the lattice. These are sometimes referred to as *intrinsic* or *thermally activated* point defects as they will be present in any crystal above 0 K, arising from the fact that the energy of some atoms in the Boltzmann distribution will be sufficient to cause them to migrate from their normal lattice sites. A real crystal will also contain impurity atoms dissolved in the lattice. There are two types of these *extrinsic* or *impurity point defects: substitutional impurities* [Fig. 1(c)], where the impurity occupies a site normally filled by the host atom, and *interstitial impurities* [Fig. 1(d)], where the impurities occupy interstitial sites in the lattice.

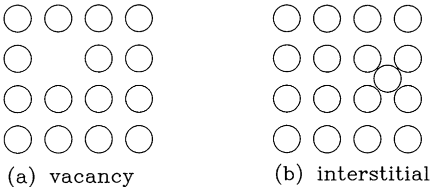

(a) vacancy

(b) interstitial atom

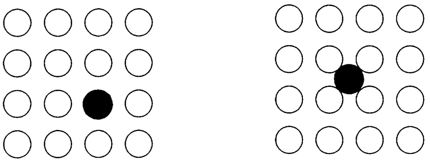

(c) substitutional impurity

(d) interstitial impurity

FIG. 1. Schematic representation of the basic point defects in monatomic solids.

In a real crystal, there will be interactions between the point defects, and this can lead to association of the simple defects to form more complex *aggregates* or *clusters*. Obvious examples are the *divacancy* (or vacancy pair), with two vacancies in adjacent lattice sites, and the *impurity–vacancy pair*, a substitutional impurity and vacancy on adjacent lattice sites. The concentration of a particular type of aggregate will depend on temperature, the concentration of the basic component defects, and their interaction energy. The last term will depend on the nature of the intermolecular forces, and the importance of aggregation, as will be seen later, depends very much on the type of solid under consideration.

1.2.2 Thermodynamics of Point Defects in a Monatomic Solid. The simplest case to deal with is a crystal containing vacancies in a single electronic state, and the treatment given below follows that developed in other texts (Howard and Lidiard, 1964; Chadwick and Glyde, 1977). The results for other types of defect will be described as appropriate.

The Gibbs free energy of a perfect crystal containing N atoms at a temperature T and uniform pressure P can be defined as $G_0(T,P)$. If n_v noninteracting, mobile vacancies are introduced into the crystal, the Gibbs free energy of this real crystal, $G(T,P)$, is given by

$$G(T,P)=G_0(T,P)+n_v g_v-kT\ \ln\Omega. \qquad (32)$$

Here g_v ($\equiv h_v-Ts_v$) is the Gibbs free energy of formation of a single vacancy, k is the Boltzmann constant, and Ω is the number of states of the system, i.e., the number of ways of arranging N atoms and n_v vacancies on $N+n_v$ lattice sites. Thus the final term in Eq. (32) is a *configurational entropy term*, given by

$$\ln\Omega=\ln\left(\frac{(N+n_v)!}{N!n_v!}\right). \qquad (33)$$

At thermal equilibrium $G(T,P)$ will be a minimum with respect to n_v, i.e.,

$$\left(\frac{\partial G}{\partial n_v}\right)_{T,P,N}=\mu_v=0, \qquad (34)$$

where μ_v is the *chemical potential of the vacancies*. The contributions to $G(T,P)$ from the vacancy free energy and the configurational entropy term are shown schematically in Fig. 2. Substituting from Eq. (33) into (32) and performing the differentiation (which requires the use of Stirling's theorem, i.e., $\ln N! = N \ln N - N$ for large values of N), yields

$$\mu_v = g_v + kT \ln\left(\frac{n_v}{N+n_v}\right) = 0 \tag{35}$$

or the site fraction of vacancies* c_v as

$$c_v = \frac{n_v}{N+n_v} = \exp\left(\frac{-g_v}{kT}\right). \tag{36}$$

Equation (36) shows that the vacancy concentration increases exponentially with temperature, reaching a maximum at the melting point. The physical process of vacancy formation may involve an atom *in* the surface layer jumping from its site to a new site *on* the surface and the subsequent migration of the vacancy so created into the bulk of the crystal. The exact reversal of the process leads to vacancy annihilation. In a crystal there is a dynamic equilibrium, the rates of vacancy formation and annihilation being equal. These rates are governed by the distances between the surfaces (*sources* and *sinks* for vacancies) and the diffusion coefficient of the vacancies. Fortunately for the experimentalist, internal surfaces, such as grain boundaries and dislocations, act as sources and sinks so that vacancy concentrations are usually in thermal equilibrium. In fact, special conditions have to be employed to perturb the equilibrium; e.g., rapid cooling (~ 1000 K s^{-1}) from near the melting point to low temperatures (where diffusion is very slow) will quench in high vacancy concentrations.

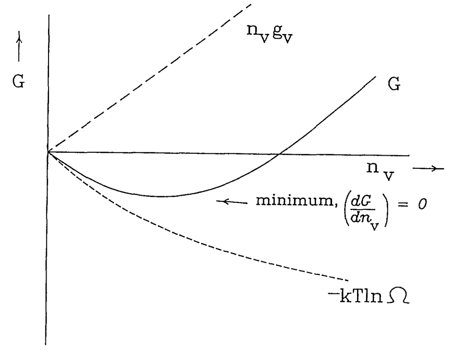

FIG. 2. Schematic representation of the vacancy contribution to the Gibbs free energy of a solid.

The derivation leading to Eq. (36) can easily be extended to a crystal containing n_i interstitial atoms to yield a site fraction (the fraction of interstitial sites in the crystal that are occupied by atoms) as

$$c_i = \frac{n_i}{\beta(N-n_i)} = \exp\left(\frac{-g_i}{kT}\right). \tag{37}$$

Here g_i ($\equiv h_i - Ts_i$) is the Gibbs free energy to form an interstitial atom, and β is the number of interstitial sites associated with a normal lattice site, i.e., a purely geometric factor depending on the crystal structure of the material. It is clear from Eqs. (36) and (37) that both vacancies and interstitial atoms will be present in any real crystal above absolute zero, although their relative concentrations will depend on g_v and g_i. Simple arguments based on atom sizes and compressibility lead to the prediction that in close-packed structures $g_i > g_v$ and the interstitial concentration is negligible with respect to vacancy concentrations.

It is worth considering the extension of the derivation of Eq. (36) to the case of vacancy pairs. This yields an expression for the *site fraction of pairs* c_p as

$$c_p = \frac{z}{2} \exp\left(\frac{-(2g_v + \Delta g_{v2})}{kT}\right), \tag{38}$$

where z is the coordination number of the lattice and Δg_{v2} is the binding energy of the pair. Written in this form Δg_{v2} is the free-energy change on bringing two vacancies from infinity to form a pair; i.e., if Δg_{v2} is negative, then divacancy formation is favorable. The term $z/2$ arises as there are $z/2$ ways of orienting a pair in the lattice. Combining Eqs. (36) and (38) yields

$$\frac{c_p}{c_v^2} = \frac{z}{2} \exp\left(\frac{-\Delta g_{v2}}{kT}\right), \tag{39}$$

---

*In the treatment of point defects it is more convenient to use *site fractions*, as defined in Eq. (36), as the expression of concentration rather than the more conventional number concentration. Since $N \gg n_v$ the site fraction is *virtually* equal to the mole fraction of vacancies.

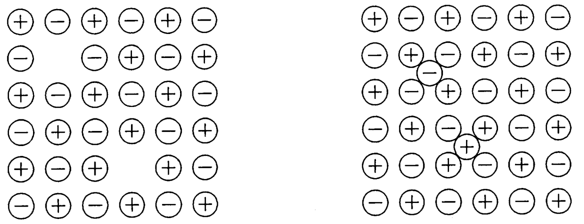

(a) Schottky pair

(b) anti Schottky pair

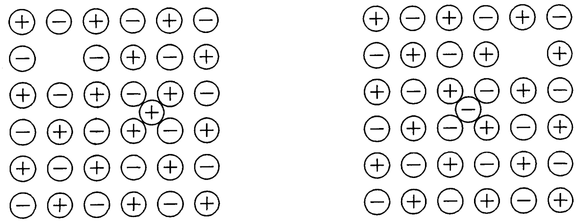

(c) cation Frenkel pair

(d) anion Frenkel pair

FIG. 3. Schematic representation of the point defects in a *MX* ionic crystal.

which can also be derived by considering the vacancy association reaction (vacancy + vacancy ⇆ divacancy) in terms of the mass action law. It should be noted that Eq. (38) will apply even when the vacancies are not in thermal equilibrium. Under normal circumstances, c_p will be expected to depend exponentially on temperature and to be significantly smaller than c_v.

Finally the perturbation on the system by the addition of impurities needs to be discussed. Assuming that the impurity concentration is low and that the impurity dissolves substitutionally, then the formation of impurity–vacancy pairs, i.e., an impurity and a vacancy on adjacent lattice sites, needs to be included in the treatment of the point-defect thermodynamics. In an analogous manner to the treatment of vacancy pairs, it can be shown that the site fraction of impurity–vacancy pairs, c_{Iv}, is given by

$$c_{Iv}=zc_I \exp\left(\frac{-(g_v+\Delta g_{Iv})}{kT}\right) \qquad (40)$$

or

$$\frac{c_{Iv}}{c_I c_v}=z \exp\left(\frac{-\Delta g_{Iv}}{kT}\right), \qquad (41)$$

where c_I is the site fraction of isolated impurity atoms and Δg_{Iv} is the binding energy of the pair. The total vacancy concentration $\overset{T}{c}_v$ $(=c_v+c_{Iv})$ is given by

$$\overset{T}{c}_v=\exp\left(\frac{g_v}{kT}\right)\left[1+zc_I \exp\left(\frac{-\Delta g_{Iv}}{kT}\right)\right]. \qquad (42)$$

Thus if Δg_{Iv} is negative and there is a tendency to form these pairs, the site fraction of vacancies is increased in comparison to the pure metal.

1.2.3 Point-Defect Thermodynamics in an Ionic Solid. In an ionic solid the point-defect picture is more complicated than that in a monatomic solid. This arises from the fact that an isolated point defect, such as a vacancy or an interstitial ion, will have an *effective* charge with respect to the lattice, and any perturbation must maintain electrical neutrality and preserve the lattice structure. In a 1:1 crystal *MX* that is strongly ionic the point defects in a pure crystal will occur in pairs.* The common types of intrinsic defects are shown in Fig. 3. The *Schottky pair* [Fig. 3(a)] consists of a cation vacancy and an anion vacancy, the *cation Frenkel pair* [Fig. 3(c)] consists of a cation vacancy and a cation interstitial, and the *anion Frenkel pair* [Fig. 3(d)] consists of an anion vacancy and an anion interstitial. The *anti-Schottky pair* [Fig. 3(b)], consisting of a cation and an anion interstitial, is a theoretically possible type of defect pair although it is not found in real systems.

---

*Only the case of a strongly ionic crystal, such as an alkali halide, is considered here. In semi-ionic solids, such as a transition-metal oxide, charge compensation on formation of a lattice defect may occur by creation of an electron or hole.

The concentration of point defects in an ionic crystal can be treated in a similar manner to that outlined in Sec. 1.2.2 for vacancies in a monatomic solid. Treatments for ionic crystals can be found in several standard texts (see, for example, Lidiard, 1957; Corish and Jacobs, 1973; Jacobs, 1983). The example taken here will be that of an MX crystal consisting of N_M cations and N_X anions and containing Schottky defects, n_{cv} cation vacancies, and n_{av} anion vacancies. The Gibbs free energy of the crystal can be expressed as

$$G(T,P)=G_0(T,P)n_{cv}g_{cv}+n_{av}g_{av}-kT\ \ln\Omega, \tag{43}$$

where g_{cv} ($\equiv h_{cv}-Ts_{cv}$) and g_{av} ($\equiv h_{av}-Ts_{av}$) are the Gibbs free energies of formation of cation vacancies and anion vacancies, respectively. In this case, with two sublattices, the configurational entropy is given by

$$kT\ \ln\Omega=kT\ \ln\left[\left(\frac{(N_M+n_{cv})!}{N_M!n_{cv}!}\right)\times\left(\frac{(N_X+n_{av})!}{N_X!n_{av}!}\right)\right]. \tag{44}$$

The minimization of $G(TP)$ with respect to the defect concentrations to find the equilibrium condition, i.e.,

$$\delta G=\sum_j\left(\frac{\partial G}{\partial n_j}\right)_{T,P,N}\delta n_j=0, \tag{45}$$

where n_j is the number of defects of type j, must be performed with the constraints that electrical neutrality is maintained and the number of sites on the two sublattices remains equal. If virtual chemical potentials of the defects are defined by

$$\mu_{cv}=g_{cv}+kT\ \ln\left(\frac{n_{cv}}{N_M+n_{cv}}\right) \tag{46}$$

and

$$\mu_{av}=g_{av}+kT\ \ln\left(\frac{n_{av}}{N_X+n_{av}}\right), \tag{46}$$

then the equilibrium condition can be rewritten as

$$\mu_{cv}+\mu_{av}=0, \tag{47}$$

which leads to

$$\frac{n_{cv}}{N_M+n_{cv}}\frac{n_{av}}{N_X+n_{av}}=c_{cv}c_{av}=\exp\left(\frac{-g_s}{kT}\right). \tag{48}$$

Here $g_s=g_{cv}+g_{av}$, the Gibbs free energy of formation of a Schottky pair, and c_{cv} and c_{av} are the site fractions of cation and anion vacancies, respectively.

Equation (48) has the form of a solubility product, and it is useful in ionic solids to consider defect equilibria in terms of quasi-chemical equilibria controlled by mass-action equations. Thus Schottky defect formation can be expressed as

$$\text{perfect lattice}\leftrightharpoons\text{cation vacancy}+\text{anion vacancy} \tag{49}$$

controlled by the mass-action equation

$$c_{cv}c_{av}=K_S=\exp(-g_S/kT), \tag{50}$$

where K_S is the equilibrium constant for the Schottky-pair formation equilibrium. Clearly it follows that in a pure crystal N_M will be equal to N_X and c_{cv} will be equal to c_{av} ($=K_S^{1/2}$).

Expressions similar to Eq. (50) can be obtained for other types of defect; for example, for cation Frenkel pairs the expression is

$$c_{cv}c_{ci}=\exp(-g_{cF}/kT), \tag{51}$$

where c_{ci} is the concentration of cation interstitials and g_{cF} ($\equiv h_{cF}-s_{cF}T$) is the Gibbs free energy of formation of cation Frenkel pairs. Clearly the type of point defect dominant in a particular crystal will depend on the relative magnitudes of their free energies of formation, which in turn depend on the ionic sizes and crystal structures. Among the simple materials the alkali halides contain Schottky pairs, the alkaline earth fluorides contain anion Frenkel pairs, and silver chloride and bromide contain cation Frenkel pairs.

A great deal of information on the point defects in an ionic crystal can be obtained by adding (or "doping") the crystal with aliovalent impurity. The additional charge introduced into the lattice has to be compensated, and this is usually achieved by perturbing the defect concentrations. Taking again the example of an MX crystal containing Schottky defects, the effect of adding substitutionally a small amount of a divalent cation, IX_2, is to introduce cation vacancies. The conditions of electrical neutrality, i.e.,

$$N_M+2n_I=N_X, \tag{52}$$

and site conservation, i.e.,

$$N_M+n_I+n_{cv}=N_X+n_{av}, \tag{53}$$

give the result

$$n_{cv}=n_I+n_{av} \tag{54}$$

and hence

$$c_{cv}=c_I+c_{av}. \tag{55}$$

Here n_I is the number of I^{2+} ions, and c_I is the corresponding site fraction. Combining Eq. (55) with (50) yields the result

$$c_{cv}=\tfrac{1}{2}[c_I+(c_I^2+4K_s)^{1/2}]. \tag{56}$$

This expression has two limits:

1. at low temperatures,

   $$c_I \gg 4K_S;\ \text{then}\ c_{cv}=c_I \tag{57}$$

   and c_{cv} is independent of temperature—the *extrinsic region*—and
2. at high temperatures,

   $$c_I \ll 4K_S;\ \text{then}\ c_{cv}=c_{av}=K_S^{1/2} = \exp(-g_S/2kT) \tag{58}$$

   and defects are thermally created—the *intrinsic region.*

Since the cation vacancy and the I^{2+} ion have opposite effective charges, they will tend to associate as a neutral pair on adjacent lattice sites (termed *impurity-vacancy pairs* or *I-V pairs*). The reaction can be written as

$$\text{cation vacancy}+I^{2+} \leftrightarrows I\text{-}V\ \text{pair} \tag{59}$$

and the mass-action relation as

$$\frac{c_{IV}}{c_I c_{cv}}=K_{IV}=\exp\left(\frac{-g_{IV}}{kT}\right). \tag{60}$$

Here c_{IV} is the site fraction of pairs, K_{IV} the equilibrium constant for association, and g_{IV} $(\equiv h_{IV}-Ts_{IV})$ the binding energy of the *I-V* pair.

The concentration of cation vacancies, c_{cv}, as a function of temperature in a *MX* crystal in which Schottky disorder dominates is shown schematically in Fig. 4. Three regions are shown: the intrinsic high-temperature region where the defects are thermally controlled, the extrinsic region at moderate temperatures where c_{cv} is equal to the divalent impurity concentration, and the association region where the cation vacancies form *I-V* pairs. As the temperature decreases a fourth region superimposes on the association region, the *precipitation region,* where the *I-V* pairs aggregate to large clusters and eventually form a separate phase, i.e., crystallites of IX_2. The plot is representative of the behavior of alkali halides, as they usually contain traces of divalent cations (e.g., calcium, magnesium, etc.) as natural impurities.

It is relatively straightforward to extend the above treatment of other types of defect, and examples can be found in several texts (Lidiard, 1957; Corish and Jacobs, 1973; Jacobs, 1983). The treatment can also be made more exact by replacing concentrations with

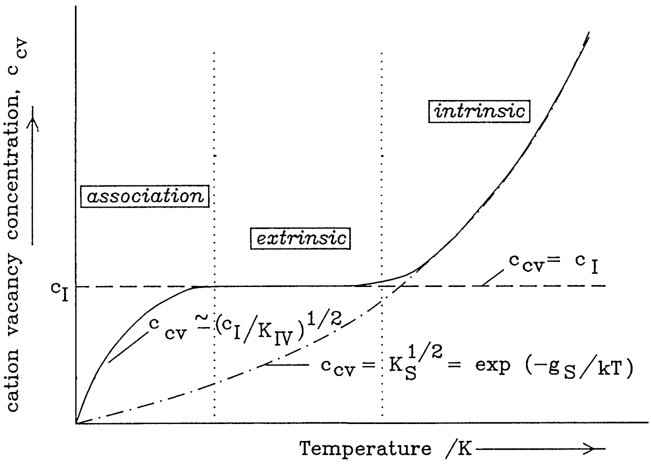

FIG. 4. Cation vacancy concentration in a *MX* crystal with Schottky disorder and doped with a divalent cation as a function of temperature.

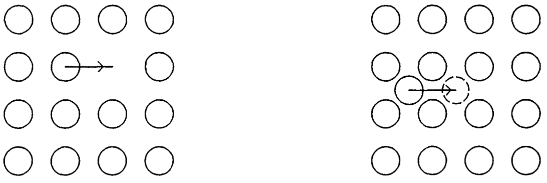

(a) vacancy (b) direct interstitial

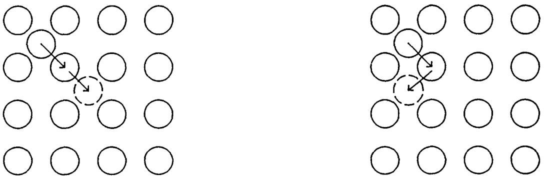

(c) collinear interstitialcy (d) non−collinear interstitialcy

FIG. 5. Schematic representation of the diffusion mechanisms in a solid.

thermodynamic activities. The nonideal behavior arises from the long-range Coulombic interactions between the defects, which are usually treated in a similar manner to electrolyte solutions by the Debye–Hückel–Lidiard theory.

1.2.4 Point-Defect Concentrations in Solids. The process of taking an atom from the lattice to form a vacancy is similar to that of sublimation. If there is no relaxation of the lattice around the vacancy then it is possible to a first approximation (Jost, 1960) to equate the enthalpy of vacancy formation, h_v, with the latent heat of sublimation, L_s, i.e.,

$$h_v \simeq L_s. \tag{61}$$

Estimation of the entropy of vacancy formation is more difficult, as it arises from the change in vibrational frequency spectrum of the crystal on forming the vacancy. Calculations for solid argon yield the result

$$s_v \simeq 2k. \tag{62}$$

Combining these values into g_v and substituting into Eq. (36) gives an estimate of c_v as about 1×10^{-4} for solid argon at the melting point T_m. This is quite a good estimate, and for many types of solid $c_v(T_m)$ lies in the range 10^{-4} to 10^{-3}. In close-packed solids c_i will be orders of magnitude smaller than c_v since $g_i > g_v$.

A variety of techniques can be used to determine the point-defect concentrations in solids, and the reader is referred to specific texts (see, for example, Flynn, 1972; Crawford and Slifkin, 1972). In general, this can be experimentally difficult because of the low magnitude of the concentrations and is usually only possible at temperatures close to T_m. An exception is the case of ionic solids where the measurement of ionic conductivity, which will be described in Sec. 2.2, can provide accurate information in a basically simple experiment.

1.2.5 Mechanisms of Diffusion. In nearly all crystalline solids where the mechanism of diffusion has been firmly established the mechanism involves point defects. The earliest theories of solid-state diffusion proposed that at the atomic level two atoms on neighboring sites simply exchanged positions. It was soon realized that this mechanism, known as the *direct exchange mechanism*, would be energetically unfavorable in close-packed solids, as it would require considerable dilation of the lattice to allow the atoms to move past each other. An extension of this mechanism is the *ring mechanism* where a group of atoms move around in a circular fashion, each atom in the group taking the place of its neighbor in the direction of motion. Although this reduces the energy of the process, the ring mechanism does not operate in most systems.

The diffusion mechanisms involving point defects are shown schematically in Fig. 5. The simplest of these is the *vacancy mechanism*, shown in Fig. 5(a), in which an atom "jumps" into a vacancy on a neighboring site. This process is energetically favorable and is the dominant diffusion process in most close-packed crystal structures. For example, this is the mechanism that operates for self-diffusion in fcc metals. In systems where the dominant defects are interstitial atoms, the diffusion can occur via these defects by a number of possible processes. In the *direct interstitial mechanism*, shown in Fig. 5(b), the interstitial atom jumps directly from interstitial site to interstitial site. This is the mechanism by which small impurity atoms that occupy the interstitial sites in metals diffuse—for example, carbon in iron. In pure solids interstitials can diffuse by the *interstitialcy mechanism*, shown in Figs. 5(c) and 5(d), where an interstitial moves onto a neighboring normal lattice site with the displacement of the original occupant atom onto a new interstitial site. There are two variants of this mechanism, the *collinear interstitialcy mechanism* [Fig. 5(c)], where the motion of the two atoms is along the same direction, and the *noncollinear interstitialcy mechanism*, where the two atoms move along different directions. Proven examples of the interstitialcy mechanism can be found in the motion of silver ions in silver halides and fluoride ions in the alkaline earth fluorides.

Variations on the basic point-defect mechanisms described above are also possible, such as the diffusion of atoms via divacancies and the *dumbbell interstitial* (two atoms located around the same lattice site).

It is important to note the use of the term *jump* to describe the process of the atom's motion. This implies that the movement is rapid, occurring over the time scale of the period of a lattice vibration, and that the time the atom is in "flight" between its initial and final sites is very short compared to its average "residence" time on a lattice site.

1.2.6 Random-Walk Theory. The diffusion of atoms in a solid with no applied driving force is analogous to Brownian motion, as each atom can be pictured as making jumps from lattice site to lattice site, the direction of each jump being in a random direction relative to preceding jumps. Thus the process can be treated by the random-walk theory developed originally by Einstein and Smoluchowski.

If an atom makes n consecutive jumps, each described by the translational vector $\mathbf{r}_i$, then the total displacement vector between the initial and final sites of the atom is given by

$$\mathbf{R}_n=\mathbf{r}_1+\mathbf{r}_2+\mathbf{r}_3\cdots=\sum_{i=1}^{n}\mathbf{r}_i. \tag{63}$$

It follows that the displacement squared is simply

$$\mathbf{R}_n^2=\sum_{i=1}^{n}\mathbf{r}_i^2+2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbf{r}_i\cdot\mathbf{r}_j. \tag{64}$$

The average displacement squared of many atoms after n jumps is given by

$$\langle\mathbf{R}_n^2\rangle=\sum_{i=1}^{n}\langle\mathbf{r}_i^2\rangle+2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\langle\mathbf{r}_i\cdot\mathbf{r}_j\rangle. \tag{65}$$

If there is only one mechanism of diffusion operative then in a cubic crystalline solid the length of each jump step will be the same, r. In addition, if the jumps are *truly random* then the last term in Eq. (65) will be averaged out to zero. For a cubic solid $\langle\mathbf{r}_i\cdot\mathbf{r}_j\rangle$ will be replaced by $\langle r^2\cos\theta_{ij}\rangle$, where θ_{ij} is the angle between the ith and jth jumps. Since all angles between jumps i and j will be found, the equation can be written simply as

$$\langle\mathbf{R}_n^2\rangle=nr^2. \tag{66}$$

Thus the average displacement, the root mean square displacement, is given by

$$\sqrt{\langle\mathbf{R}_n^2\rangle}=\sqrt{n}r. \tag{67}$$

If the n jumps take place in a time t, then the jump frequency Γ is given by

$$\Gamma=n/t \tag{68}$$

and

$$\langle\mathbf{R}_n^2\rangle=t\Gamma r^2. \tag{69}$$

The Einstein equation relates the diffusion equation to the mean square displacement, and for a cubic solid this is

$$D=\langle\mathbf{R}_n^2\rangle/6t. \tag{70}$$

Thus combining this equation with Eq. (67) yields

$$D=\Gamma r^2/6, \tag{71}$$

which is the *Einstein–Smoluchowski equation*.

1.2.7 Relations between Macroscopic and Microscopic Parameters. It is the Einstein–Smoluchowski equation that allows the connection to be made between the measured diffusion coefficient in an experiment and the parameters involved in the atomistic mechanism of diffusion. Here the case of self-diffusion of atoms in a monatomic solid by a vacancy mechanism is taken as the example.

The jump frequency of an atom, Γ, will depend on the probability that a vacant site is adjacent to the atom (the product of the site fraction of vacancies, c_v, and the coordination number z) and the jump frequency w_0 of the vacancy:

$$\Gamma = zc_v w_0. \tag{72}$$

Absolute reaction-rate theory, developed for chemical kinetics (Glasstone *et al.*, 1941), can be used to obtain an expression for w_0 in the form

$$w_0 = \nu \exp(-\Delta g_v^m/kT). \tag{73}$$

Here ν is the vibrational frequency of the atom in the direction of the vacancy (usually taken as the Debye frequency) and Δg_v^m ($\equiv \Delta h_v^m - T\Delta s_v^m$) is the Gibbs free energy of vacancy migration, i.e., the height of the potential-energy barrier that the atom has to overcome in passing from its initial site into its final site. There has been debate about the validity of this approach, as it implies that the jump process would have to occur slowly enough for thermodynamic equilibrium to be maintained. Thus "dynamic theories" have been proposed; however, it has been shown (Glyde, 1967) that these lead to essentially the same form of result.

Combining Eqs. (72) and (73) and substituting into Eq. (71) with the expansion of c_v yield the final expression

$$D = \frac{\nu z r^2}{6} \exp\left(\frac{s_v + \Delta s_v^m}{k}\right) \exp\left(\frac{-(h_v + \Delta h_v^m)}{kT}\right). \tag{74}$$

In experiments the temperature dependence of the diffusion coefficient in a crystalline solid usually obeys the *Arrhenius equation*:

$$D = D_0 \exp(-Q/kT), \tag{75}$$

where D_0 is called the preexponential factor and Q is the activation energy. Hence by comparing Eqs. (74) and (75) the following identities between the experimental and microscopic parameters can be made:

$$D_0 = \frac{\nu z r^2}{6} \exp\left(\frac{s_v + \Delta s_v^m}{k}\right) \tag{76}$$

and

$$Q = h_v + \Delta h_v^m. \tag{77}$$

It is possible to make a rough estimate of Δh_v^m at the same level of approximation used to estimate h_v, i.e., Eq. (61). As an atom moves from its site the number of bonds it needs to "break" with its neighbors is similar to those broken in vacancy formation. Thus Δh_v^m will also be approximately equal to L_s, i.e.,

$$\Delta h_v^m = L_s \tag{78}$$

and therefore

$$Q \simeq 2L_s. \tag{79}$$

This prediction should be reasonable where there is no relaxation, either electronic or lattice—for example, in molecular solids. If it is assumed that $\Delta s_v^m \simeq s_v \simeq 2k$, then it is possible to estimate the value of D at T_m as 10^{-8} $\mathrm{cm}^2\,\mathrm{s}^{-1}$, which is the value typically found experimentally for most solids.

1.2.8 Correlation Effects in Diffusion. In order to measure the diffusion coefficient it is necessary to "tag" atoms so that their motions can be followed; for example, they may be isotopically labeled as in a tracer experiment. If the motion of an individual atom that is tagged is followed, then it will be found that there can be correlations between successive diffusive jumps, depending on the mechanism. The simplest example that demonstrates correlation is that of diffusion in a pure solid via a vacancy mechanism. If the tagged atom makes a jump into a vacancy, then its *most probable next jump* is back to its original site, as there is a vacancy there ready to accept it. Thus the tracer jumps will be correlated although the jumps of the vacancy are random. For the tracer it is necessary to modify the diffusion coefficient to

$$D_t = fD_{\text{random}}. \tag{80}$$

Here D_t is the measured diffusion coefficient of the tracer, D_{random} is the diffusion coefficient given by random-walk expressions, and f is the correlation factor. Sometimes f is referred to as the *Bardeen–Herring correlation factor*, after the first workers (Bardeen and Herring, 1952) to note this correlation in the motion of tagged atoms. Since accurate values of f can be theoretically calculated for all the possible diffusion mechanisms in the different lattice structures, the determination of f provides an experimental route to the identification of diffusion mechanisms (LeClaire, 1970). Values of f are given in Table 1.

2. EXPERIMENTAL METHODS OF STUDYING DIFFUSION

A wide variety of methods can be employed in the investigation of diffusion. The focus here is on those that can provide direct, reliable information on the diffusion parameters and insights into the diffusion mechanisms. Of particular interest are clearly those methods that can yield accurate values of the diffusion coefficient. Thus techniques that provide indirect or limited diffusion information are not treated here. For example, the rate of plastic deformation (*creep*) under certain conditions is governed by atomic diffusion, and a study of the temperature dependence will yield good values of the activation energy for diffusion, Q, of the atoms. However, this technique, although useful, is limited in the detail it can provide and is not included in the discussion.

2.1 Methods of Measuring Diffusion Coefficients

2.1.1 Study of the Diffusion Penetration Profile. The most reliable method of measuring the diffusion coefficient is to determine the profile of labeled atoms that have been diffused from the surface of a sample in a known time at a constant temperature. In order to ensure that the true bulk diffusion coefficient is measured, the sample should be a single crystal with the minimum number of dislocations.

Table 1. Values of the correlation factor f for self-diffusion in simple lattices.

Mechanism of diffusion	Lattice	f
Vacancy	fcc	0.7815
	bcc	0.727
	sc	0.653
	Honeycomb	0.3333
Divacancy	fcc	0.458
Direct interstitial	Independent of structure	1
Collinear interstitialcy	fcc	0.9643
	sc	0.8000

The majority of the accurate diffusion coefficients that are available have been measured by the *radiotracer serial sectioning technique*, or its variants, and it is applicable to a wide range of solid types. The full details can be found in a number of texts (see, for example, Bénière, 1983; Mundy and Rothman, 1983). The basic procedures of this technique are shown in Fig. 6. First, a very thin layer of radiotracer is applied to a flat surface of the sample by vacuum sublimation or by deposition from a solution. The tracer may be an isotope of the host crystal (e.g., ^{22}Na to measure self-diffusion in Na, ^{36}Cl to measure chloride ion self-diffusion in NaCl, ^{14}C to measure carbon diffusion in steel, etc.). The sample is then annealed at constant temperature T for a known time t. Following the anneal the sample is sectioned parallel to the face of the tracer deposit and the sections subjected to a radioactive assay. This allows the determination of the penetration profile and the evaluation of the tracer diffusion coefficient from Eq. (6), i.e., using a plot of ln(tracer concentration) versus (penetration)^2. Two advantages of this technique are its simplicity and the relative cheapness of the equipment. In addition, the determination of D_t *directly* from the penetration profile provides a good test that bulk diffusion is being observed; the diffusion plot should be linear, and D_t should be independent of annealing time. An obvious disadvantage is the destructive nature of the sectioning, while the technique is limited to systems where convenient radiotracers, in terms of half-life and radiological safety, are available.

The controlling parameters in this experiment are basically the anneal time and the thickness of the sections. Reasonable times for the anneal lie in the range 10^3 s (any shorter and heating–cooling corrections become large) to 10^6 s (longer times would generally be inconvenient). A variety of techniques can be used to section the sample. For bulk samples, of the order of 10 mm dimensions, the sections are generally taken off with a microtome (for reasonably hard materials,

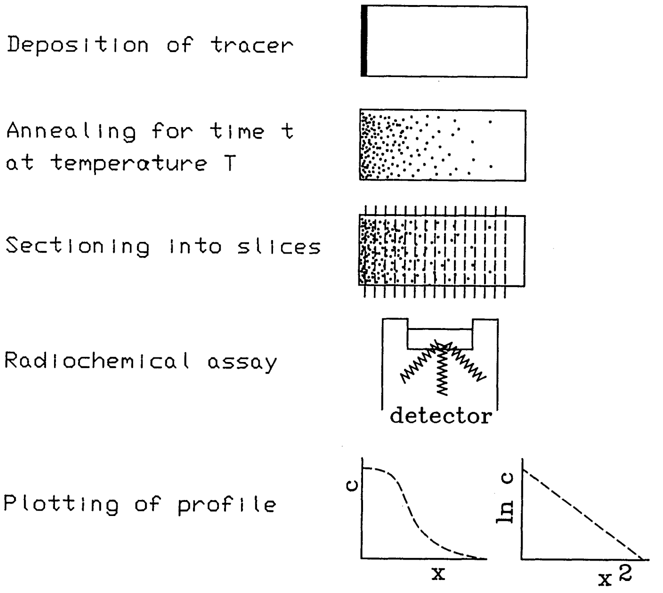

FIG. 6. Procedures of the radiotracer serial-sectioning diffusion experiment.

e.g., alkali halides), with a lathe (for metals), or by grinding (very hard materials), and the individual sections are usually in the 1- to 100-μm range. Using the criterion (Bénière, 1983) that the diffusion zone $[\sim 4(Dt)^{1/2}]$ should be divided into at least 13 sections over which the activity falls by a factor of 100 gives the useful range of D_t that can be studied with these types of sectioning as roughly 10^{-14} to 10^{-6} cm^2 s^{-1}. For smaller values of D_t, thinner sections need to be taken, and this is possible by sputtering (sections ~5 to 100 nm) and electrochemical etching (sections ~50 nm), putting the lower limit of D_t at around 10^{-17} cm^2 s^{-1}.

A very important variation of the technique is to use a stable isotope and secondary-ion mass spectroscopy (SIMS) to study the penetration profile (Macht and Naundorf, 1982; Lodding, 1987). This has proved particularly useful in the study of oxygen diffusion, as there is no radioactive isotope of this element and the diffusion is usually very slow. The sections taken in SIMS are in the range 1–100 nm allowing values of D_t as low as 10^{-19} cm^2 s^{-1} to be studied. In addition, very small samples, around 1 mm dimensions, can be used. However, the equipment required for a SIMS experiment is extremely expensive and requires a highly trained operator. Other means of determining the penetration profile of a tracer that are particularly useful in studying slow diffusion are the use of Rutherford backscattering and nuclear reaction analysis.

2.1.2 Measurement of the Atomic Jump Frequency. A wide variety of relaxation techniques can be used to deduce the jump frequency of atoms and hence allow the evaluation of D via the Einstein–Smoluchowski equation, i.e., Eq. (71). Generally these techniques rely on a theoretical model to relate the observed parameter to the microscopic atomic motion. This involves making some assumption about the nature of the *correlation function* (the function that expresses the probability that a particle originally at a position $\mathbf{r}=0$ at time $t=0$ is found at a position $\mathbf{r}=\mathbf{r}$ after a time t) and the *pair correlation function* (the probability that if an atom is at $\mathbf{r}=\mathbf{0}$ another atom will be found at $\mathbf{r}=\mathbf{r}$) (van Hove, 1954). Techniques that come into this category are dielectric relaxation spectroscopy, neutron scattering (Springer and Richter, 1987; Lechner, 1983), Mössbauer spectroscopy, and nuclear magnetic resonance (NMR) spectroscopy (Wolf, 1979; Strange, 1986; Kanert, 1982). The general advantages of these techniques are that they are nondestructive and rapid (the atoms only need to make a few jumps to give the effect and thus the observation time is short), and they are usually less demanding on sample requirements when compared to tracer methods (powders can be used in some cases). However, they often have limited applications to certain elements (e.g., Mössbauer spectroscopy can only be used for ^{57}Fe, ^{119}Sn, and a few other elements) and systems where diffusion is relatively fast, i.e., $D > 10^{-9}$ cm^2 s^{-1}. In addition, it is usually difficult to use these methods to give accurate values of D, because of deficiencies in the theoretical treatment of the correlation functions for an ensemble of atoms, although they will yield good values of Q. An exception to some of these current limitations is the NMR technique.

NMR diffusion experiments (see MAGNETIC RESONANCE, NUCLEAR) are best suited to systems where the NMR-active isotope is in high abundance and has nuclear spin s equal to $\frac{1}{2}$ (e.g., ^1H, ^{19}F) or a small quadrupolar moment (e.g., ^7Li, ^{23}Na). A reasonably large number of interesting systems fall into this category, particularly organic solids. There are a variety of NMR experimental procedures, but those in current common usage involve the measurement of a spin relaxation time in a pulsed spectrometer (Farrar and Becker, 1971; Fukushima and Roeder, 1981). The basis of the experiments involves disturbing the equilibrium spin population between a certain set of energy levels in the system by the application of a short burst (pulse) of radio-frequency power and then measuring the characteristic time required for the spins to return to equilibrium by monitoring the magnetization of the sample. For example, the spin–spin relaxation time T_1 is the time for the spins to equilibrate between the Zeeman levels whose separation is determined by the magnitude of the applied external magnetic field, $\mathbf{B}_0$. A spin in an excited state can only relax down to the lower level by losing a quantum of energy equal to the separation between the states, $\Delta E = \hbar\omega$, where ω is the Larmor precession frequency. This involves the spin interacting with oscillating magnetic fields in the system that fluctuate at a frequency equal to ω, i.e., it is a stimulated emission. In the system a spin will experience a *local field* due to dipole–dipole interactions with the neighboring spins. One source of oscillating magnetic field is the modulation of the local field caused by relative motion of the spins. Diffusive motion provides a fluctuating magnetic field whose frequency components are similar to those of a "noise spectrum." Analysis of this spectrum into its Fourier components yields the spectral density function, which depends on the atomic (spin) jump frequency Γ. The precise relationship between the measured relaxation time and Γ involves an assumption about the correlation function, and various models have been developed. However, it is clear that the relaxation will be most efficient (the relaxation time will be a minimum) when Γ is equal to ω. By measuring different relaxation times, i.e., probing different sets of energy levels for the spins, a wide range of Γ, and hence D, can be determined. Typically T_1 will monitor D in the range 10^{-11}–10^{-6} cm^2 s^{-1}; the spin–spin relaxation time T_2, the range 10^{-10}–10^{-6} cm^2 s^{-1}; the spin–lattice relaxation time in the rotating frame, $T_{1\rho}$, the range 10^{-14}–10^{-8} cm^2 s^{-1}; and the spin–lattice relaxation time in the local dipolar field, T_{1D}, the range 10^{-20}–10^{-12} cm^2 s^{-1}. Thus it is possible with one technique, using one spectrometer, to probe a very wide range of D.

Two further points concerning NMR methods are worthy of note. First, for cubic systems a precise theory relating relaxation times to Γ has been developed, tested, and shown to be reliable, and can be used to obtain accurate values of D. This is the *encounter model* as exploited by Wolf (1979) (see also Strange, 1986). Second, there is a further NMR technique, *pulsed-field gradient NMR*, that does not involve measurement of the relaxation time. In this experiment the spin is used to "label" atoms, like a tracer, and the time scale of the experiment is such that the atoms move macroscopic distances. Thus it is not subject to the criticisms outlined above and is particularly useful in the study of diffusion in liquids or solids where diffusion is rapid, i.e., $D > 10^{-9}$ cm^2 s^{-1} (Strange, 1986).

2.2 Ionic Conductivity Measurements

For materials that are ionic the measurement of the ionic conductivity provides a very powerful method of studying diffusion. As will be seen below, these experiments can yield not only information on the migration of the ions but also the concentration of point defects. This has been one of the major factors that has led to the current deep level of understanding of point defects and transport in ionic solids—a better understanding than in any other type of solid. Thus the measurement and interpretation of ionic conductivity has been described in a number of texts

(Lidiard, 1957; MacDonald, 1987; Chadwick, 1991; Corish and Jacobs, 1973; Jacobs, 1983) and is described here in some detail.

Samples for conductivity measurements are typically disks, about 100 mm^2 in cross-sectional area and 5 mm thick. Ideally the sample is a single crystal that has a known orientation, but this is not always available and it is common to use a compacted pellet. It is usually necessary to anneal the specimen close to the melting point prior to measurement to release any strains from the sample and to ensure a stable geometry. In the two-probe experiment most commonly in use, the large faces of the disk are placed in contact with the measuring electrodes in a cell with a controlled atmosphere and temperature. Many designs of conductivity cell have been published and the essential features of common systems are shown in Fig. 7. The key features of this design are that

1. the cell atmosphere can be controlled using either a purified inert gas stream or a dynamic vacuum (a very important feature when studying reactive materials like fluorides),
2. inert electrodes (usually graphite) are coated on the large faces of the sample to ensure good contact with the measuring electrodes and a true evaluation of σ from the measured conductance,
3. temperature is accurately measured by means of two thermocouples on each side of the sample, and
4. construction materials (high-purity silica, alumina, graphite, and platinum) that are involatile and unreactive to most materials are used in the hot zone.

This type of system can be used to make very precise conductivity measurements, and, provided the sample is thoroughly thermally equilibrated, the absolute accuracy is governed mainly by the accuracy of the thermocouples. The upper temperature limit of this cell, around 1000 °C, can be extended by replacing the silica components with alumina.

The conductivity is best measured by an ac technique to avoid polarization effects at the electrodes. Highly accurate ac conductivity–capacitance bridges, with errors around 1 part in 10^4, have been commercially available for many years. The earlier models were tedious to use as they were manually balanced and operated at a fixed, basic frequency, although separate signal generators and detectors could be attached to make frequency-dependence studies. Later generations of bridges were autobalancing, but still operated at fixed frequencies. The current generation of commercial systems are autobalancing, can rapidly scan a wide frequency range, and are readily interfaced to microcomputers to allow completely automated data collection. The evaluation from the raw data of the true "bulk" conductivity due to the motion of ions within the matrix of the lattice depends on the nature of the sample. The approach now in general use is that of complex-plane impedance analysis (Bauerle, 1969; Archer and Armstrong, 1981; MacDonald, 1987).

The application of a small, sinusoidally oscillating potential $\Delta E \sin \omega t$ with angular frequency ω ($=2\pi f$) across the electrodes will give rise to a current $\Delta i \sin(\omega t+\theta)$ that is

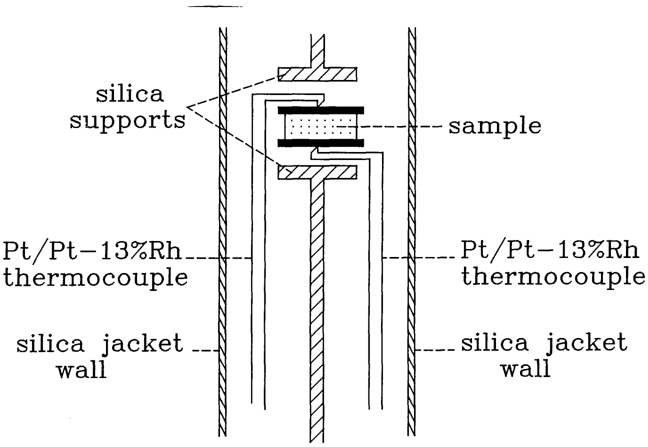

FIG. 7. Essential features of a cell for measuring the ionic conductivity of solids.

phase shifted by an angle θ. The impedance Z can be characterized by a combination of the ratio $\Delta E/\Delta i$ and the phase angle θ. It is convenient to represent Z at a given ω as a complex function and hence by a point on the complex impedance plane, where the y axis is the imaginary component, Z'', and the x axis is the real component, Z'. Thus $|Z|$ is the length of the line from the origin to the point and θ is the angle between this line and the x axis. The locus of Z as ω is varied will depend on the components in the circuit between the electrodes. The advantage of the complex-impedance representation is that a series combination of elements can be described by the vectorial addition of individual components. The situation is more complicated for a parallel combination of elements. For example, a simple combination of resistance R and capacitance C leads to a semicircle in the complex-impedance plot. For a parallel combination of elements it is often more useful to plot the admittance Y $(=1/Z)$ in the complex plane, i.e., the imaginary part B (the susceptance) versus the real part G (the conductance). All the parameters are interrelated, and it is possible to present the basic measurement of $|Z|$, ω, and θ in a variety of formats; the interconversion is a built-in feature of some of the commercial impedance analyzers. The problems in many studies are first to find the equivalent circuit that corresponds to the measurements and then to identify the circuit elements with physical features or chemical processes in the electrode-sample arrangement.

The simplest situation is that for a single crystal of a normal ionic conductor such as an alkali halide. In this case the circuit can be represented as a resistance and capacitance in parallel. Provided the frequency is high enough to avoid electrode polarization, usually $\geqslant 1$ kHz, then the measured conductance G will be the true bulk conductance of the sample, and extremely accurate measurements can be taken (the individual data points can be accurate to $\pm 0.1\%$). For materials with high ionic conductivities and for polycrystalline pellets the bulk conductance has to be extracted from the complex impedance or admittance plot. The ideal complex impedance spectrum for a ceramic pellet would consist of three semicircles, as shown in Fig. 8. Each of the contributions to the electrical behavior of the bulk material, i.e., the interior of the grains, the grain boundaries, and the sample–electrode interface, is represented by an RC component, and these are connected in series. Thus, in principle, it is possible to extract the true bulk conductance and to analyze it in terms of defect structures and transport mechanisms. However, the accuracy is necessarily lower than that achievable for alkali halides, and the number of data points that can be feasibly collected over a given temperature range is less. The complex-impedance spectra for many real pellets are more complicated than Fig. 8, and a number of strategies have to be adopted to identify the bulk resistance (e.g., varying the sample geometry, electrode preparation, impurity concentration, etc.). Good examples can be found in the original paper by Bauerle (1969).

Typical conductivity data for an ionic crystal, in this case sodium chloride, are shown in Fig. 9. It should be noted that this is very

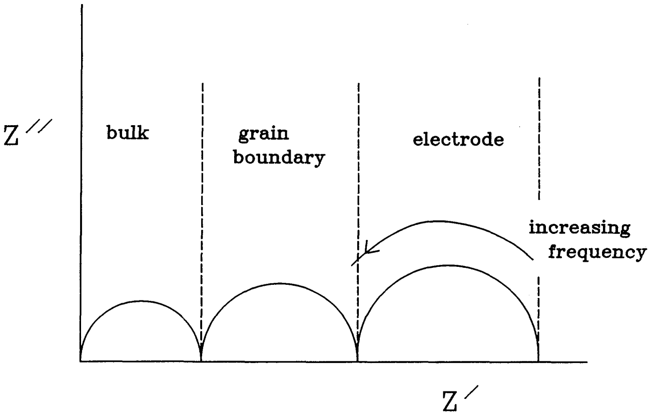

FIG. 8. Idealized complex-impedance spectrum for a ceramic ionic conductor.

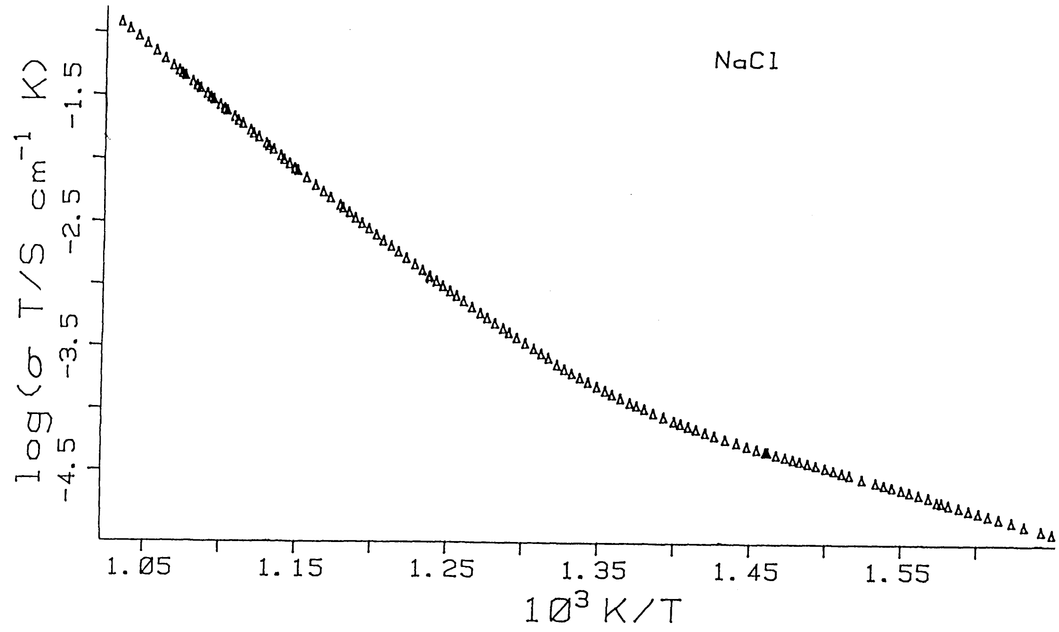

FIG. 9. Ionic conductivity data for a sodium chloride single crystal (after Hooton, 1988).

high-quality data with a high density of accurate points. The simplest first step in the analysis of data of this type is to determine the Arrhenius energies from the plot* of $\log(\sigma T)$ vs $1/T$. The next stage is to attempt to relate these to the microscopic conduction processes. The specific conductance of a substance is given by the relationship

$$\sigma = \sum_{\text{all } r} n_r |q_r| \mu_r, \tag{81}$$

where n_r is the number of conducting species of type r per unit volume, q_r and μ_r are its charge and mobility, respectively, and the sum is taken over all the types of mobile species. For a simple ionic crystal the species can be regarded either as the ions themselves or as the point defects by which they move. Thus values of n_r can be written down explicitly in terms of point-defect equilibria. Alkali halides will be considered as a simple example, and here the predominant point defects are Schottky pairs (a cation vacancy *cv* and an anion vacancy *av*) and the defect equilibrium is described by the equations in Sec. 1.2.2. It has already been mentioned that long-range defect interactions can be included by including activity coefficients γ_r, which can be calculated from the Debye–Hückel–Lidiard theory (Lidiard, 1957). This gives

$$\gamma_r = \exp\left(\frac{-q_r^2 \kappa}{2\epsilon kT(1+\kappa R)}\right), \tag{82}$$

where κ^{-1} is the Debye length, given by

$$\kappa^2 = 4\pi e^2 \sum_r \frac{c_r}{2a^3 \epsilon kT}. \tag{83}$$

Here, e is the protonic charge, ϵ is the permittivity, $2a^3$ is the volume per molecule of the salt, and R is the distance of closest approach between defects before the formation of a defect pair or complex. Thus it is possible to solve the mass-action equations using the constraints of charge and site balance and to write down an explicit expression for the concentration of any defect in terms of the thermodynamic parameters governing its formation and association with other defects and the corresponding activity coefficients.

The defect mobility μ_r is related to the defect jump frequency w_r by the expression

$$\mu_r = S_r |q_r| a_r^2 w_r / 6kT, \tag{84}$$

---

*In the older literature (and in some current papers) data are plotted as $\log\sigma$ vs $1/T$. However, this is not correct, as the mobility μ in Eq. (81) depends on $1/T$, and activation energies deduced from these plots will usually be slightly in error.

where S_r is the number of equivalent sites into which the defect can make a jump of distance a_r. The jump frequency w_r is given by Eq. (73). The mobilities can to be corrected for the drag experienced from the surrounding defect cloud by using the *Onsager–Pitts correction* (Lidiard, 1957).

Thus, given a model of the defect structure and transport mechanisms, it is possible to rewrite Eq. (81) with explicit expressions for the defect concentrations and mobilities. For an alkali halide crystal containing M^{2+} impurity this will contain nine parameters: the Schottky formation enthalpy and entropy, h_s and s_s; the migration enthalpies and entropies of the cation vacancy, $\Delta\mathscr{H}^{m}_{cv}$ and $\Delta\mathscr{S}^{m}_{cv}$, and anion vacancy, $\Delta\mathscr{H}^{m}_{av}$ and $\Delta\mathscr{S}^{m}_{av}$; the association enthalpy and entropy of the cation vacancy-M^{2+} complex, h_{IV} and s_{IV}; and the total M^{2+} concentration c_T. The modern procedure is to make a computer fit to the experimental data with these parameters as the variables in a nonlinear least-squares-fitting program. These programs will minimize the sum of the residual deviations squared, ϕ, i.e.,

$$\phi = \sum [\log(\sigma T)_{\text{calc}} - \log(\sigma T)_{\text{exptl}})]^2. \qquad (85)$$

The Debye–Hückel–Lidiard activity coefficients and the Onsager–Pitts correction are incorporated in the expressions, as are the temperature dependences of ϵ and a, if they are available. The advantages of the least-squares computer fitting approach are that it makes the best use of accurate data and that the *whole* conductivity plot is analyzed to yield a full set of defect parameters, for both defect formation and migration. Clearly, precautions in the fitting have to be taken (Jacobs, 1983; Chadwick, 1991); and to get really reliable defect parameters, crystals with a range of dopant concentrations have to be studied, and often other diffusion data are used as guidelines.

However, it is not always possible to obtain high-quality data over a wide temperature regime. In these situations a less detailed approach is required. It is usually reasonable to assume that only one type of the defect pair is present, and simple theoretical models will normally be able to identify which of the defects is the more mobile. It is then possible to obtain estimates of the enthalpy of defect migration and formation. For example, the more mobile defect in NaCl is the cation vacancy, and the dominant impurities in these crystals are divalent cations. Thus in the low-temperature region of Fig. 9 the cation vacancy concentration is determined by the impurities and, if *I-V* pairs are neglected, c_{cv} is given simply by Eq. (57). Substituting this into Eq. (81) shows that the slope of the *extrinsic conductivity region* is therefore equal to $-\Delta\mathscr{H}^{m}_{cv}/k$. Similarly, the value of c_{cv} in the high-temperature *intrinsic conductivity region* is given by Eq. (58), and the slope here is equal to $-(h_s/2+\Delta\mathscr{H}^{m}_{cv})/k$. Thus it is still possible to extract information on the defect-formation and -migration processes.

2.3 Methods of Identifying Diffusion Mechanisms

The ultimate aim of diffusion studies is to identify the nature of the atomic diffusion mechanism that is operative, i.e., vacancy, divacancy, interstitialcy, etc. Clearly, information on the nature of the dominant point defects in the solid will provide a starting point. For example, in close-packed metals the dominant defects are known to be vacancies, so that mechanisms based on interstitials can be ruled out. In less obvious situations one approach is to compare the experimentally determined diffusion parameters (predominantly activation energies but also activation volumes) with those obtained from theoretical calculations for the different mechanisms. In most cases this has been quite a sound approach, and with the advent of the modern computer-simulation procedures this is becoming even more reliable. For ionic solids the predictions from these calculations have been quite enlightening on some complex systems.

Direct experimental identification of the diffusion mechanism has usually relied on the determination of the correlation factor f. In principle, this can be determined from the study of the *isotope effect for diffusion* (LeClaire, 1970; Peterson, 1975), which involves the measurement of the differences in diffusion coefficient of two radiotracers, A and B, with different masses, e.g., ^{22}Na and ^{24}Na. However, there are only a few elements with convenient isotopes to use this approach, and the relations between f and D_A and D_B are not as simple as expected from the early work on this topic. The experiment will measure the product $f\Delta K$, where ΔK is the fraction of the kinetic energy of the jump that is carried by the tracer atom. In a jump many atoms around the tracer and defect will be involved; therefore, ΔK is not unity and is typically 0.8–0.9. In some well-understood systems where the diffusion mechanism and hence f are known, it is possible to work backward to determine ΔK; however, in an unknown system it can be hard to separate f and ΔK, and identification of the mechanism is not simple.

In ionic crystals the ionic conductivity arises from the movement of the effectively charged defects in the electric field. Since the diffusion of the defects is random, the diffusion coefficient calculated from the ionic conductivity D_σ via the Nernst–Einstein equation [Eq. (14)] will be the diffusion coefficient of the ions if they were moving randomly. Thus using Eq. (80) yields

$$f = D_t/D_\sigma, \tag{86}$$

and in principle the mechanism can be identified. However, the situation is often more complex. The ions may diffuse by more than one mechanism; e.g., the Ag^+ ions can diffuse by both vacancy- and interstitial-based mechanisms, and the ratio D_t/D_σ is not simply f. This ratio is still useful and is now referred to as the *Haven ratio.*

2.4 Computational Techniques

Computer simulation of defects and transport in solids is now well established, particularly for ionic materials, and the methodologies are described in a number of recent reviews (Catlow and Mackrodt, 1983; Catlow, 1987; Harding, 1990; Kehr and Binder, 1987). The simulation methods fall into three general types: *static lattice simulations, molecular dynamics* (MD) *methods,* and *Monte Carlo* (MC) *methods.*

Briefly, in the static lattice methods the atoms, with some appropriate mathematical representation of the interatomic potential, are placed on the sites of a representative portion of the lattice, and the energy of the lattice is calculated by pairwise summation of all the intermolecular forces. The calculation is then repeated with a defect in the center of the lattice to yield the energy of defect creation. The activation energy for diffusion can be calculated by placing the defect at the saddle point of the jump path—e.g., for the vacancy mechanism with the diffusing atom midway between the initial and final lattice sites. The success of the methods depends on the reliability of the interatomic potential and the ability to treat "edge" effects and relaxation in the system on defect creation. In ionic and semi-ionic solids the codes HADES and CASCADE have been thoroughly tested and shown to be remarkably successful for calculating defect formation and migration energies (Catlow and Mackrodt, 1983; Harding, 1990). These codes use the Mott–Littleton (1938) two-region procedure, where an inner region of ions (~ 100 ions) is treated explicitly and a large outer region is treated as an elastic and dielectric continuum. Fast minimization techniques are used to determine the minimum-energy configuration of the system. Quite sophisticated potentials, either empirical or *ab initio,* can be employed, and the effects of polarization of the ions have been incorporated using shell-model potentials. In essence the static lattice methods are calculations at 0 K and yield the internal energies of defect formation or migration. However, using the code SHEOL (Harding, 1990) the calculations for ionic solids can be extended to yield defect-free energies, enthalpies, and entropies. It should be noted that there are other methods of computing the defect entropies.

MD calculations (see MOLECULAR DYNAMICS) are often regarded as "experiments" (Bennett, 1984; Ciccotti and Hoover, 1986). An ensemble of atoms is arranged in the simulation box and the system allowed to evolve. Typically the box will contain 100 atoms with periodic boundary conditions to generate an infinite crystal. The atoms interact with a given potential, and they are started from coordinates close to lattice positions with trajectories and velocities corresponding to some given temperature. The system is allowed to "run," which means solving the classical equations of motion in a succession of small time steps (typically 10^{-15} s, which is less than the characteristic time of any motional process for the atoms, e.g., a lattice vibration) and continually updating the positions and velocities. Following an initial equilibration period the simulation should model all the properties, structural and dynamic, of the system. This method is useful for systems where there is rapid motion and a significant number of diffusive events can be observed in the time scale of the simulation. A limitation has been the demands on computing power due to the large memory requirement and therefore the necessity to use rather simplified potentials. With the developments in computers this limitation is not expected to be a major problem in the future.

MC methods (*q.v.*) (Kehr and Binder, 1987) are useful where high levels of disorder are present and a variety of types of jump of an atom may be possible in a system. The technique involves generating an ensemble of configurations by random moves and using an "acceptance criterion" to decide which of the configurations should be included in the ensemble. This leads to a picture of the most likely jump processes in the system.

3. DIFFUSION IN DIFFERENT SOLID TYPES

3.1 Metals and Alloys

Because of their technological importance there have been extensive studies of point defects and diffusion in metals (Brebec, 1983; Philibert, 1986). The most recent comprehensive and up-to-date collection of data can be found in the survey edited by Mehrer (1990). As expected for close-packed structures, the dominant disorder is single vacancies. Inevitably there will be vacancy aggregates (divacancies, trivacancies, etc.) present, but the concentrations are significantly smaller, and it is difficult to obtain reliable thermodynamic parameters of these defects. A variety of different techniques have been used to determine values of h_v in pure metals, and in the well-studied systems the values obtained from different experiments are consistent [see, for example, the comparison between the values from the positron-annihilation technique and quenching experiments (Eldrup, 1986)]. The values are considerably smaller than the simple prediction given in Sec. 1.2.3 that h_v would be $\sim L_s$; typically $h_v \sim (0.3–0.4)L_s$. The origin of this discrepancy is the atomic relaxation and electron charge redistribution that occurs on vacancy formation in a metal. Explicit calculations of defect parameters in metals are complicated by the complex electronic structure and the difficulty in describing the interatomic potential.

In the thoroughly studied fcc metals, like Al, Ag, Au, Cu, and Pt, the self-diffusion occurs predominantly via a vacancy mechanism. In cases where h_v and $\Delta \mathcal{H}_v^m$ can be determined separately, the sum of the two parameters agrees well with experimental values of Q, as illustrated in Table 2. The table also demonstrates the empirical observation that $\Delta \mathcal{H}_v^m \simeq 0.8 h_v$ for fcc metals. The Arrhenius plots of self-diffusion measurements for some of these metals exhibit slight curvature, the apparent activation energy increasing as T_m is approached. This can be seen in the data for Ag shown in Fig. 10. The most common interpretation of this effect is that there is a contribution to the diffusion from the migration via divacancies (Brebec, 1983; Philibert, 1986). Although the divacancy concentration is generally much lower than c_v, it is argued that they are more mobile than the single vacancies and hence have a significant effect on D as the temperature approaches T_m. It should be noted that an alternative interpretation is that single vacancies alone are responsible for self-diffusion but the defect enthalpies and entropies are slightly temperature dependent because of anharmonicity of the lattice vibrations (Mundy, 1992).

The self-diffusion in a number of bcc metals, notably β-Ti, β-Zr, β-Hf, and the rare-earth elements, reveals a strong curvature of the Arrhenius plots, and these have long been regarded as "anomalous" systems. A number of explanations have been proposed to explain the effect [see, for example, the recent reviews (Brebec, 1983; Philibert, 1986; Herzig and Kohler, 1987; Mundy, 1992)], including the presence of oxygen impurity, the presence of bulk defects, a contribution from divacancy diffusion, and temperature-dependent single-vacancy parameters. The contribution from divacancies has been a favored explanation, but recent work (Herzig and Kohler, 1987) related the self-diffusion to soft-phonon modes and provides a uniform explanation of the data in terms of single vacancies.

Table 2. Experimental values for the enthalpy of vacancy formation, h_v, enthalpy of vacancy migration, $\Delta \mathcal{H}_v^m$, and self-diffusion in fcc metals.

	Experimental enthalpies eV		
Metal	h_v	$\Delta \mathcal{H}_v^m$	Q
Cu	1.30	0.80	2.05
Ag	1.10	0.83	1.91
Au	0.98	0.82	1.81
Al	0.77	0.65	1.46

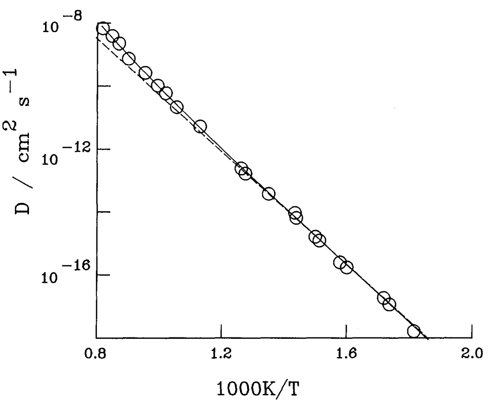

FIG. 10. Compiled self-diffusion coefficient of silver as a function of temperature (after Brebec, 1983).

The literature on metallic alloys is also very extensive, and a wide range of materials have been investigated (Brebec, 1983; Philibert, 1986). It is appropriate here to consider only the dilute alloys. The sites occupied in the host lattice depend mainly on the size of the impurity; atoms of a similar size will dissolve substitutionally, and small atoms can occupy interstitial sites (e.g., C, N, and O in bcc metals like α-Fe). Interstitial impurity diffusion in dilute alloys is consistent with the impurity migrating via the direct interstitial mechanism. In substitutional alloys the diffusion data are usually interpreted in terms of the *five-frequency model* (Brebec, 1983; Philibert, 1986). In this model the diffusion coefficients of the host impurity atoms are derived as functions of the defect concentrations and the five distinct types of jump of the vacancy:

1. vacancy adjacent to impurity exchanging with the impurity,
2. vacancy adjacent to impurity exchanging with host atom and remaining adjacent to the impurity,
3. vacancy adjacent to impurity exchanging with host atom and moving away from the impurity,
4. the reverse of **3**, and
5. vacancy exchanging with a host atom on sites not adjacent to the impurity.

There are some well-documented systems where the impurity exhibits anomalously fast diffusion in the alloy—for example, Cu, Pd, Au, Ni, and Ag in Pb, and Au in the alkali metals. In these cases the vacancy mechanism seems inappropriate and various forms of interstitial process have been suggested to explain the anomalies.

3.2 Molecular Solids

Molecular solids represent the largest group of materials, as they include all the organic and biological materials, which are virtually infinite in number given the facility of carbon to bond covalently with itself and a range of other elements (e.g., O, N, H, S, etc.). Information on the point defects in these materials is sparse, a result of both the experimental difficulties in studying these crystals (i.e., high vapor pressures, low purity) and the lack of technological interest in the systems as solids. The interest has been primarily of an academic nature, particularly in the case of the rare-earth gas solids where relative simplicity of the interatomic forces makes them ideal model systems.

It is convenient to divide molecular solids into three types:

1. *rare-gas crystals,*
2. *plastic crystals*—systems where the molecules are roughly spherical, undergo a rapid endospherical reorientation, and adopt highly symmetric crystal structures, usually fcc, bcc, and hcp (Sherwood, 1979), and
3. *nonplastic crystals*—the majority of organic crystals, where the molecular symmetry is low, the molecules are not reorienting, and the crystal structure is determined by packing factors.

In all three types the crystal structures are close packed and the expected dominant point defects are vacancies.

3.2.1 Rare-Gas Crystals. There have been a large number of theoretical estimates of the vacancy formation and migration parameters in rare-gas crystals (Chadwick and Glyde, 1977; Chadwick, 1983). The interatomic potentials employed range from Lennard–Jones potentials to sophisticated empirical potentials, and in most cases it is concluded that $h_v \simeq 1L_S$ and $\Delta \mathscr{H}_v^m \simeq 1L_S$, consistent with the simple predictions in Secs. 1.2.3 and 1.2.7. There have been several experimental studies of vacancy concentrations in rare-gas solids. The experiments are extremely difficult, and there is still some confusion over the experimental vacancy formation parameters. For example, the earlier experiments for solid krypton yielded $h_v \sim 0.7L_s$, whereas the value for neon was $1L_s$. The experiments for krypton were over a limited temperature range and the data analysis has been questioned, and there is a need for new studies on this system.

The diffusion data for rare-gas crystals from a number of sources, including radiotracers and NMR, suggest that these systems are well behaved (Chadwick and Glyde, 1977; Chadwick, 1983). The parameters are consistent with the simple theoretical predictions, i.e., Q is close to $2L_s$.

3.2.2 Plastic Crystals. In many respects the simpler plastic crystals can be considered as analogous to rare-gas crystals; the rapid molecular reorientation averages out the interactions between atoms on separate molecules, and these can be replaced by a single intermolecular potential acting between molecular centers of gravity. The simplest predictions (Chadwick, 1983) are therefore that $h_v \simeq 1L_s$ and $\Delta h_v^m \simeq 1L_s$. The same results have been obtained in the case of adamantane, where calculations have used atom–atom potentials (Meyer *et al.*, 1986). Reliable information on vacancy-formation parameters in plastic crystals is limited to a positron-annihilation study of two materials, adamantane and bicyclooctane, which yielded $h_v = 1L_s$ in both cases (Lightbody *et al.*, 1980).

There have been a large number of self-diffusion studies in plastic solids (Sherwood, 1979), as the motion is relatively rapid at moderate temperatures (D at T_m is typically 10^{-9}–10^{-8} cm^2 s^{-1} and T_m is less than 300 °C) and a wide number of diffusion techniques can be employed. NMR measurements for these materials are particularly appropriate, as D is in the required range and the organic plastic crystals usually contain a suitable nucleus (e.g., ^{1}H). In many cases the simple prediction of $Q \simeq 2L_s$ seems to be observed, but there are a number of systems where NMR methods yield lower values of Q (e.g., $Q \simeq 1L_s$ in cyclohexane and white phosphorus). This apparent anomaly is still unresolved.

3.2.3 Nonplastic Crystals. The diffusion in nonplastic solids is around 4 to 5 orders of magnitude *slower* than that in plastic crystals. The diffusion experiments are therefore difficult, and only a few systems, mainly simple aromatic ring systems like anthracene and naphthalene, have been studied. The general results are that $Q \simeq 2L_s$ and that the dominant point defects are single vacancies (Chadwick, 1983).

3.3 Covalent Solids

There has been considerable debate over the nature of the point defects in both elemental (e.g., C, Si, Ge) and compound covalent crystals. In essence the problems arise in assigning the electronic structure around the defect, e.g., what happens to the "dangling bonds" when an atom is removed to form a vacancy. Each broken bond contains one electron, which may occupy lower energy levels by forming new bonds across the vacancy with the dangling bonds of neighboring atoms. Self-diffusion and impurity diffusion have been very actively studied in these materials, particularly for systems where there are semiconductor applications, as the information is useful in device fabrication. In the elemental crystals the self-diffusion is extremely slow (D at T_m is typically 10^{-13}–10^{-12} cm^2 s^{-1}), and the mechanisms that are operative have not been fully resolved (Gosele and Tan, 1992; Leroy, 1990). It is now believed that both self-interstitials and vacancies are involved in the self-diffusion and most impurity diffusion processes in Si and GaAs, a suggestion made over 20 years ago (Seeger and Chik, 1968) that is now being substantiated by experiments.

3.4 Ionic Solids

Of all the solid types, the point defects and diffusion are best understood in ionic crystals. This is due to a combination of factors, some of which are evident from the foregoing sections. From the experimental viewpoint there are a number of good model systems, like the alkali and silver halides, that are readily available in the form of large single crystals. In addition, it is possible to control the defect concentrations precisely by means of doping with aliovalent impurities. However, the greatest advantage over other materials in defect and diffusion studies is that the technique of ionic conductivity provides a relatively simple and accurate means of determining defect-formation and -migration parameters. The combination of ionic conductivity and diffusion measurements also provides a means of identifying mechanisms via the Haven-ratio method. From the theoretical viewpoint the interionic potential can be well represented by fairly simple mathematical functions, and therefore computer simulation methods have proved highly successful.

In the last two decades it has been convenient to consider ionic solids as being of two types, *normal ionic conductors* and *fast-ion conductors*, and these will be discussed separately.

3.4.1 Normal Ionic Solids. Normal ionic conductors are materials like the alkali halides where the conductivity is well described by jump mechanisms of simple point defects and the ionic conductivity is around 10^{-3} S cm^{-1} at the melting point. Most of the detailed conductivity and diffusion work has focused on the simple systems like the metal halides, particularly those with the rock-salt and fluorite structures, because of the ease of experiment, although very similar behavior is expected for the higher-melting-point oxides. For a large number of systems there is now a wealth of experimental information, including the nature of the dominant point defects, the thermodynamic parameters for point-defect formation and migration, and the nature of the self-diffusion mechanisms (Lidiard, 1957; Corish and Jacobs, 1973; Chadwick, 1991; Jacobs, 1983). In many of these systems static lattice computer simulations have been extremely successful in calculating the defect parameters (Catlow, 1987).

The systems for which there is the best experimental information are the alkali halides, the silver halides, and the alkaline-earth fluorides. Some of these data are shown as an example in Table 3. Only the principal point-defect enthalpies are shown; however, in many cases the entropies and the parameters for defect association are available. The dominant defects in the alkali halides are Schottky pairs, and both ions diffuse by a vacancy mechanism. The diffusion coefficients of both ions are similar. For systems like NaCl and KCl there is a wealth of data, and the level of interpretation is now extremely sophisticated (see, for example, Hooton and Jacobs, 1988), including contributions to the ion transport from vacancy pairs and Frenkel defects (which are present in very low concentrations but are mobile enough to make a significant effect on the transport at high temperatures). In silver chloride and bromide the dominant point defects are cation Frenkel pairs, and not surprisingly the cation diffusion coefficient is about 1000 times higher than that for the anion. In the intrinsic region the silver ions move predominantly by interstitialcy mechanisms, both collinear and noncollinear, while in the extrinsic region of crystals doped with divalent cations (i.e., doped with excess cation

Table 3. Point-defect enthalpies (in eV) for simple ionic crystals.

Material	Point-defect enthalpy	Experimental value determined from ionic conductivity	Calculated value from computer simulation
NaCl	h_S	2.41	2.32
	Δh^m_{cv}	0.67	0.65
	Δh^m_{av}	0.77	0.71
KCl	h_S	2.50	2.50
	Δh^m_{cv}	0.67	0.71
	h^m_{av}	0.85	0.69
AgCl	h_{cF}	1.49	1.4
	Δh^m_{ci} (noncollinear)	0.10–0.12	
	Δh^m_{ci} (collinear)	0.002–0.004	
	Δh^m_{cv}	0.27–0.28	
CaF_2	h_{aF}	1.85	1.98
	Δh^m_{av}	0.52	0.40
	Δh^m_{ai} (noncollinear)	0.79	0.72
BaF_2	h_{aF}	1.91	1.98
	Δh^m_{av}	0.57	0.46
	Δh^m_{ai} (noncollinear)	0.76	0.72
MgO	h_S	5–7	7.5
	Δh^m_{cv}	2.2–2.6	2.18

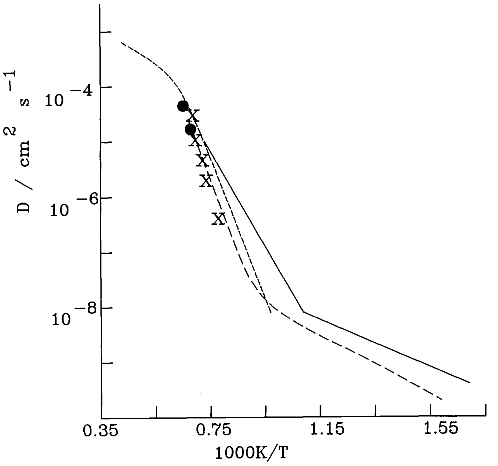

FIG. 11. Ionic transport data for Li_2O. A plot of the Li^+ diffusion coefficient from ionic conductivity (long-dashed curve), molecular-dynamics simulations (short-dashed curve), static lattice simulations of HADES and SHEOL (solid line), tracer (crosses), and neutron data (filled circles) (after Flack, 1990).

vacancies) the silver ions diffuse by a vacancy mechanism. The alkaline-earth fluorides at moderate temperature, up to about $0.8T_m$, behave as normal ionic crystals with anion Frenkel pairs as the dominant defects, and the fluoride ion diffuses predominantly by the vacancy mechanism, although there is a significant contribution from the noncollinear interstitialcy process (Jacobs, 1983).

The success of the computer modeling when applied to the point-defect parameters of ionic crystals is also seen in Table 3. In a recent study (Chadwick *et al.*, 1988; Flack, 1990) of lithium oxide, Li_2O (a material with the antifluorite structure), it was possible to compare experimental conductivity and diffusion data with a range of computer-simulation methods. The results are shown in Fig. 11 and the agreement is excellent. The successful testing of the computer calculations in simple systems means that they can now be used with confidence in predicting the behavior of more complex ionic materials.

3.4.2 Fast-Ion Conductors. There are a number of ionic solids with unusually high ionic conductivities, the magnitude of the conductivity approaching that found in molten salts and aqueous electrolyte solutions, and these are now variously referred to as *fast-ion conductors* or *superionic conductors* or *solid electrolytes*. Tremendous interest has been shown in these materials for technological applications where a purely ionic conductor is required, e.g., in solid-state batteries and fuel cells. There is an extensive literature on these materials, including books (Chandra, 1981; Laskar and Chandra, 1990; Stoneham, 1989), reviews (Schulz, 1982; Catlow, 1990; Chadwick, 1992), and a specialized journal (*Solid State Ionics*). Conductivity plots for a few fast-ion conductors are shown in Fig. 12. It can be seen that a wide range of solid types exhibit the phenomenon, i.e., fluorite-structured solids at high temperature (e.g., PbF_2), silver iodide and related compounds (e.g., $RbAg_4I_5$), and layered structures (e.g., β-aluminas). Clearly in terms of Eq. (81), in order for there to be a high ionic conductivity the point defect in the solid must have exceptionally high concentrations and/or exceptionally high mobilities.

The types of ion that show fast-ion conduction are usually small with low formal charges, e.g., Li^+, Na^+, F^-, H^+, and O^{2-}. It has now been realized that there is no unique origin to fast-ion conduction in the different systems, and therefore a variety of theoretical approaches have to be adopted. There are several ways of classifying fast-ion conductors, depending on their structure (open, tunnel, layered, glasses, polymers, etc.) or their applications. A very useful approach developed by Catlow (1990) and explored by others has been to use a hybrid classification based on structure and mechanism. Thus the types of class include solids that are fast-ion conductors as a result of a phase transition (like AgI), massively disordered and heavily doped materials (such as ZrO_2 stabilized in the cubic

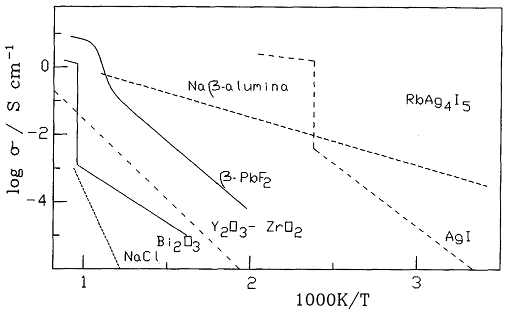

FIG. 12. Conductivity plots for several fast-ion conductors.

phase by impurity), and layered structures (like the β-aluminas).

There are still many problems to be resolved in fast-ion conductors, particularly the precise nature of the diffusion mechanisms. In many cases the diffusion is so rapid that many of the conventional experimental probes are inadequate. An additional problem is the lack of availability of good single crystals, and a measurement of the ionic conductivity is virtually the only transport technique that is applicable. Thus there is often a lack of precise information on which to base deductions concerning mechanisms. In much of the early work it was believed that the diffusion of the ions was liquidlike, with ions in continuous motion, and that defect-jump models would be inappropriate. However, a picture is now emerging, particularly from molecular-dynamics computer simulations (Gillan, 1989), that in most fast-ion conductors the mobile ion spends most of the time on a lattice site and a relatively short time "in transit" between sites. Molecular-dynamics simulations are particularly applicable to fast-ion conductors, as the diffusion is sufficiently rapid that a significant number of diffusive jumps can be observed in a computational time that is not prohibitive.

4. DIFFUSION IN AN APPLIED FIELD

The investigation of diffusion in an applied field, either an electric field (*electromigration*) or a temperature gradient (*thermomigration*), is difficult from both experimental and theoretical viewpoints (Allnatt and Chadwick, 1967; Wever, 1973; Hummel and Huntington, 1977; Gillan, 1983). However, the phenomena are extremely important technologically as they can lead to "unimixing" of a system. When there are large electric fields and current densities, as in a thin-film metal connection on a semiconductor device, electromigration can lead to a net movement of the metal in one direction and an eventual break in the connection. Similarly, a large temperature gradient, as is found in the walls of a furnace or in the fuel element of a nuclear reactor, can lead to segregation of one of the components of a material and ultimately mechanical or chemical breakdown. The academic interest in these phenomena lies in the fact that they provide a severe test of the atomistic models of diffusion in solids.

There have been few studies of electromigration. One method of study of impurity electromigration is to place a tracer between two slices of material and measure the drift of the tracer after the application of a large field across (and large current through) the sandwich. Self-electromigration is studied by placing markers along a specimen and noting their relative movement in the applied field. In metals the movement of the atoms can be considered as due to the *effective charge* of the atom interacting with the flow of electrons, termed the *electron wind effect*.

Thermomigration in pure solids can also be studied in marker-movement experiments, and these have been used to study metals. However, the experiments are extremely difficult to perform, as distortion of the samples complicates the interpretation. For impurity thermomigration it is possible to use the Soret effect, the steady concentration gradient of impurity developed in the temperature gradient. From Eq. (15) at steady state the flux will be zero and hence $Q^{*\prime}$ can be deduced from the concentration gradient, i.e.,

$$\frac{1}{c}\left(\frac{\delta c}{\delta x}\right)=\frac{Q^{*\prime}}{kT^2}\left(\frac{dT}{dx}\right) . \tag{87}$$

This experimental approach has been used to study thermomigration of impurities in a variety of types of solids. In ionic crystals the effect of thermomigration is to tend to separate the charged species. The incipient separation is opposed by an electric field, the *thermoelectric power*, and this provides a very convenient means of studying thermomigration (Allnatt and Chadwick, 1967; Gillan, 1983) and determining the heats of transport.

From the earlier description of the heat of transport, the heat carried from the initial site to the final site in an isothermal jump of an atom, it is clearly a complex parameter to evaluate, since at the atomic level the energy of the jump is spread over many atoms and $Q^{*\prime}$ depends on how this energy is redistributed after the jump. An early, simple theory, due to Wirtz (1943), predicted that $Q^{*\prime}$ would be equal to or less than the activation enthalpy of migration of the atom, Δh^m. Later this was simplified to $Q^{*\prime}=\Delta h^m$ in the LeClaire–Brinkman theory (LeClaire, 1954; Brinkman, 1954), i.e., the atom carries all the migration enthalpy from initial site to final site. In systems where there is simple vacancy diffusion, e.g., in alkali halide crystals, this seems to be reasonable as experiments show $Q^{*\prime} \sim \Delta h^m$. However, in cases where interstitial migration is known then this theory fails and does not even predict the correct sign for $Q^{*\prime}$; in these systems experimental values of $Q^{*\prime}$ are usually negative (Allnatt and Chadwick, 1967; Gillan, 1983). This probably reflects the highly cooperative nature of interstitial diffusion processes.

5. SUMMARY

The objective of this article has been to present the basics of the diffusion and ionic conductivity in solids to form a foundation on which to build by further reading. The nature of the different types of point-defect diffusion has been presented along with an introduction to defect thermodynamics. The situation in the different types of solid can be summarized as follows:

Metals: In pure metals the dominant defects are single vacancies. The defect physics of fcc metals is well understood as a result of reliable measurements of defect concentrations and diffusion. In bcc metals the dominant defects are also single vacancies, although there has been some debate concerning the role of divacancies in diffusion. The lattice positions occupied by impurities depends mainly on their size, small impurities (with respect to the host atoms) occupying interstitial sites and larger impurities dissolving substitutionally.

Covalent solids: The defects are less well understood in these systems because of the role of "dangling bonds" and the very low diffusion coefficients. The diffusion is complicated and appears to involve both vacancies and interstitial defects in some form.

Molecular solids: These have not been thoroughly studied, mainly because of the lack of technological applications. As the structures are close packed it is generally assumed that the dominant defects are single vacancies and diffusion is via a vacancy mechanism; these assumptions are upheld by the very limited amount of experimental data.

Ionic solids: These are extremely well understood in terms of their defect physics as the result of the ability to vary the defect concentrations deliberately, experimentally, by doping with aliovalent ions. In addition, ionic conductivity measurements provide an excellent probe of both the defect concentrations and mobilities.

The references given in the list of Further Reading will provide plenty of material for the interested reader to extend his or her knowledge and to consider the various processes where the defects and diffusion play a vital role, for example, in corrosion, catalysis, the photographic process, materials design, etc.

GLOSSARY

Correlation Factor: The degree of correlation between successive diffusive jumps of a labeled atom in a solid (more correctly termed the *Bardeen–Herring correlation factor*).

Divacancy: Two vacancies on neighboring lattice sites.

Doping: The deliberate addition of impurities to a solid.

Electromigration: Diffusion in an applied electric field.

Frenkel Defect: Defect in an ionic crystal consisting of a vacancy and an interstitial of one of the ions (hence two types are possible: *cation Frenkel defects* and *anion Frenkel defects*).

Fast-Ion Conductors: Solid with an unusually high ionic conductivity (sometimes referred to as *superionic conductors*).

Haven Ratio: For an ionic solid, the ratio of the tracer diffusion coefficient to the diffusion coefficient calculated from the ionic conductivity via the Nernst–Einstein equation.

Interstitial Atom: An atom on an interstitial lattice site in a crystal.

Interstitial Diffusion: The diffusion process in which an interstitial atom jumps from interstitial site to interstitial site.

Interstitialcy Diffusion: The diffusion process in which an interstitial atom jumps onto a normal lattice site as the atom on that site moves onto an interstitial site.

I-V Pairs: An impurity atom and a vacancy on neighboring lattice sites.

NMR: Nuclear magnetic resonance.

Plastic Crystal: A molecular solid with globular molecules that are undergoing rapid endospherical reorientation. The crystals have high symmetry (cubic or hexagonal lattices) and tend to be easily plastically deformed.

Point Defect: A defect in a crystal at the atomic level (see **Vacancy** and **Interstitial Atom**).

Schottky Defects: Point defect in an ionic solid consisting of cation and anion vacancies.

SIMS: Secondary-ion mass spectrometry.

Site Fraction: The concentration of a species in a crystal measured as the ratio of the number of molecules of the species to the total number of sites in the crystal the species could occupy.

Thermomigration: Diffusion in an applied temperature gradient.

Tracer Diffusion: The measurement of diffusion coefficients by monitoring the penetration of isotopic tracer atoms into a solid.

Vacancy: An unoccupied lattice site in a crystal.

Vacancy Diffusion: The diffusion process in which an atom moves by jumping into a neighboring vacancy.

List of Works Cited

Allnatt, A. R., Chadwick, A. V. (1967), *Chem. Rev.* **67**, 681–705.

Archer, W. I., Armstrong, R. D. (1981), in: H. R. Thirsk (Ed.), *Electrochemistry*, London: The Chemical Society, p. 157.

Bardeen, J., Herring, C. (1952), in: W. Shockley (Ed.), *Imperfections in Nearly Perfect Crystals*, New York: Wiley, p. 261.

Bauerle, J. E. (1969), *J. Phys. Chem. Solids* **30**, 2657–2670.

Bénière, F. (1983), in: F. Bénière, C. R. A. Catlow (Eds.), *Matter Transport in Solids*, NATO Advanced Study Institute, Series B, Physics, Vol. 97, New York: Plenum, p. 21.

Bennett, C. H. (1984), in: G. E. Murch, A. S. Nowick (Eds.), *Diffusion in Crystalline Solids*, New York: Academic, p. 73.

Brebec, G. (1983), in: F. Bénière, C. R. A. Catlow (Eds.), *Matter Transport in Solids*, NATO Advanced Study Institute, Series B, Physics, Vol. 97,New York: Plenum, p. 251.

Brinkman, J. A. (1954), *Phys. Rev.* **93**, 345.

Catlow, C. R. A. (1987), in: A. K. Cheetham, P. Day (Eds.), *Solid State Chemistry: Techniques*, Oxford: Clarendon, p. 231.

Catlow, C. R. A. (1990), *J. Chem. Soc. Faraday Trans.* **86**, 1167–1176.

Catlow, C. R. A., Mackrodt, W. C. (Eds.) (1983), *Computer Simulation of Solids*, Berlin: Springer.

Chadwick, A. V. (1983), in: F. Bénière, C. R. A. Catlow (Eds.), *Matter Transport in Solids*, NATO Advanced Study Institute, Series B, Physics, Vol. 97, New York: Plenum, p. 285.

Chadwick, A. V. (1991), *Philos. Mag. A* **64**, 983–998.

Chadwick, A. V. (1992), in: G. E. Murch (Ed.), *Unsolved Problems in Diffusion*, Zurich: Trans. Tech., p. 489.

Chadwick, A. V., Glyde, H. R. (1977), in: M. L. Klein, J. A. Venables (Eds.), *Rare Gas Solids*, New York: Academic, Vol. 2, Chap. 19.

Chadwick, A. V., Flack, K. W., Strange, J. H., Harding, J. H. (1988), *Solid State Ionics*, **28-30**, 185–188.

Chandra, S. (1981), *Superionic Solids*, Amsterdam: North-Holland.

Ciccotti, G., Hoover, W. G. (Eds.) (1986), *Molecular Dynamics Simulation of Statistical Mechanical Systems*, Amsterdam: North-Holland.

Corish, J., Jacobs, P. W. M. (1973), in: M. W. Roberts, J. M. Thomas (Eds.), *Surface and Defect Properties of Solids*, London: The Chemical Society, Vol. 2, p. 184.

Crank, J. (1975), *The Mathematics of Diffusion*, Oxford: Clarendon.

Crawford, J. H., Slifkin, L. M. (Eds.) (1972), *Point Defects in Solids*, New York: Plenum, Vols. 1–3.

de Groot, S. R., Mazur, P. (1962), *Non-Equilibrium Thermodynamics*, Amsterdam: North-Holland.

Eldrup, M. (1986), in: A. V. Chadwick, M. Terenzi (Eds.), *Defects in Solids—Modern Techniques*, NATO Advanced Study Institute, Series B, Physics, Vol. 147, New York: Plenum, p. 145.

Farrar, T. C., Becker, E. D. (1971), *Pulse and Fourier Transform NMR*, New York: Academic.

Flack, K. W. (1990), Ph.D. thesis, University of Kent (unpublished).

Flynn, C. P. (1972), *Point Defects and Diffusion*, Oxford: Clarendon.

Fukushima, E., Roeder, S. B. W. (1981), *Experimental Pulse NMR: A Nuts and Bolts Approach*, Reading, MA: Addison-Wesley.

Gillan, M. J. (1983), in: F. Bénière, C. R. A. Catlow (Eds.), *Matter Transport in Solids*, NATO Advanced Study Institute, Series B, Physics, Vol. 97, New York: Plenum, p. 227.

Gillan, M. J. (1989), in: A. M. Stoneham (Ed.), *Ionic Solids at High Temperatures*, Singapore: World Scientific.

Glasstone, S., Eyring, H., Laidler, K. J. (1941), *The Theory of Rate Processes*, New York: McGraw-Hill.

Glyde, H. R. (1967), *Rev. Mod. Phys.* **39**, 373–382.

Gosele, U., Tan, T. Y. (1992), in: G. E. Murch

(Ed.), *Unsolved Problems in Diffusion*, Zurich: Trans. Tech., p. 189.

Greaves, G. N. (1990), in: D. R. Uhlmann, N. Kreidl (Eds.), *Glass Science and Technology*, London: Academic, Vol. 4B, p. 1.

Harding, J. H. (1990), *Rep. Prog. Phys.* **53**, 1403–1466.

Herzig, C., Kohler, U. (1987), *Mater. Sci. Forum* **1–3**, 301–322.

Hooton, I. E. (1988), Ph.D. thesis, University of Western Ontario (unpublished).

Hooton, I. E., Jacobs, P. W. M. (1988), *Can. J. Chem.* **66**, 830–835.

van Hove, L. (1954), *Phys. Rev.* **95**, 249–262.

Howard, R. E., Lidiard, A. B. (1964), *Rep. Prog. Phys.* **27**, 161–240.

Hummel, R. E., Huntington, H. B. (Eds.) (1977), *Electro- and Thermotransport in Alloys*, New York: AIME.

Jacobs, P. W. M. (1983), in: F. Bénière, C. R. A. Catlow (Eds.), *Matter Transport in Solids*, NATO Advanced Study Institute, Series B, Physics, Vol. 97, New York: Plenum, p. 81.

Jost, W. (1960), *Diffusion in Solid, Liquids and Gases*, New York: Academic.

Kanert, O. (1982), *Phys. Rep.* **91**, 183–232.

Kaur, I., Gust, W. (1988), *Fundamentals of Boundary and Interface Boundary Diffusion*, Stuttgart: Zeigler.

Kehr, R., Binder, K. (1987), in: K. Binder (Ed.), *Applications of Monte Carlo Simulation and Statistical Physics*, Topics in Current Physics Vol. 36, Berlin: Springer, p. 181.

King, A. H. (1987), *Int. Mater. Rev.* **32**, 173–189.

Laskar, A. L., Chandra, S. (Eds.) (1990), *Supersonic Solids and Solid Electrolytes*, New York: Academic.

Lechner, R. (1983), in: F. Bénière, C. R. A. Catlow (Eds.), *Matter Transport in Solids*, NATO Advanced Study Institute, Series B, Physics, Vol. 97, New York: Plenum, p. 169.

LeClaire, A. D. (1954), *Phys. Rev.* **93**, 344.

LeClaire, A. D. (1970), in: H. Eyring, D. Henderson, W. Jost (Eds.), *Physical Chemistry—An Advanced Treatise*, New York: Academic, Vol. 10, p. 261.

Leroy, B. (1990), in: A. L. Laskar, G. Brebec, J. L. Boucquet, C. Monty (Eds.), *Diffusion in Materials*, NATO Advanced Study Institute, Series E, Applied Sciences, Vol. 179, Dordrecht: Kluwer, p. 525.

Lidiard, A. B. (1957), in S. Flügge (Ed.), *Handbuch der Physik*, Berlin: Springer, Vol. 20, p. 246.

Lightbody, D., Eldrup, M., Sherwood, J. N. (1980), *Chem. Phys. Lett.* **70**, 487–491.

Lodding, A. (1987), in: F. Adams, R. Gijbels, R. van Grieken (Eds.), *Inorganic Mass Spectroscopy*, New York: Wiley, p. 125.

MacDonald, J. R. (1987), *Impedance Spectroscopy*, New York: Wiley.

Macht, M. P., Naundorf, V. (1982), *J. Appl. Phys.* **53**, 7551–7557.

Meyer, M., Mahric, C., Ciccotti, G. (1986), *Mol. Phys.* **58**, 723–733.

Mehrer, H. (Ed.) (1990), *Diffusion in Solid Metals and Alloys*, Landolt-Bőrnstein, New Series Vol. III/26, Berlin: Springer.

Mott, N. F, Littleton, M. J. (1938), *Trans. Faraday Soc.* **34**, 485–499.

Mundy, J. N. (1992), in: G. E. Murch (Ed.), *Unsolved Problems in Diffusion*, Zurich: Trans. Tech., p. 1.

Mundy, J. N., Rothman, S. J. (1983), in: J. N. Mundy, S. J. Rothman, M. J. Fluss, L. C. Smedskjaer (Eds.), *Methods in Experimental Physics*, New York: Academic, Vol. 21, p. 1.

Peterson, N. L. (1975), in: A. S. Nowick, J. J. Burton (Eds.), *Diffusion in Solids, Recent Developments*, New York: Academic, p. 115.

Philibert, J. (1986), in: A. V. Chadwick, M. Terenzi (Eds.), *Defects in Solids—Modern Techniques*, NATO Advanced Study Institute, Series B, Physics, Vol. 147, New York: Plenum, p. 349.

Prigogine, I. (1967), *Introduction to the Thermodynamics of Irreversible Processes*, New York: Wiley-Interscience.

Schulz, H. (1982), *Annu. Rev. Mater. Sci.* **12**, 351–376.

Seeger, A., Chik, C. P. (1968), *Phys. Status Solidi* **29**, 455–542.

Sherwood, J. N. (Ed.) (1979), *The Plastically Crystalline State*, London: Wiley.

Springer, T., Richter, D. (1987), in: D. L. Price, K. Sköld (Eds.), *Methods of Experimental Physics*, Orlando, FL: Academic, Vol. 23B, p. 131.

Stoneham, A. M. (Ed.) (1989), *Ionic Solids at High Temperatures*, Singapore: World Scientific.

Strange, J. H. (1986), in: A. V. Chadwick, M. Terenzi (Eds.), *Defects in Solids—Modern Techniques*, NATO Advanced Study Institute, Series B, Physics, Vol. 147, New York: Plenum, p. 243.

Wever, H. (1973), *Elektro- und Thermotransport in Metallen*, Leipzig: Barth.

Wirtz, K. (1943), *Physik. Z.* **44**, 221–231 and 369–392.

Wolf, D. (1979), *Spin Temperature and Nuclear Spin Relaxation in Matter*, Oxford: Clarendon.

Further Reading

Textbooks Devoted to Diffusion in Solids

Adda, Y., Philibert, J. (1966), *La Diffusion dans les Solides*, Paris: Presses Universitaire de France, Vols. 1 and 2.

Flynn, C. P. (1972), *Point Defects and Diffusion*, Oxford: Clarendon Press.

Kofstad, P. (1972), *Non-Stoichiometry, Diffusion and Electrical Conduction in Binary Metal Oxides*, New York: Wiley.

Manning, J. R. (1968), *Diffusion Kinetics for Atoms in Crystals*, Princeton, NJ: Van Nostrand.

Philibert, J. (1985), *Atom Movements*, Paris: Les Editions de Physique: translation into English by S. J. Rothman (1991) with additional material.

Shewmon, P. G. (1963), *Diffusion in Solids*, New York: McGraw-Hill. 2nd ed. (1989), Warrendale, PA: The Metallurgical Society of AIME.

Edited Texts on Diffusion in Solids

Abromeit, C., Wollenberger, H. (Eds.) (1988), *Interstitials and Vacancies in Metals and Alloys*, Aedermansdorf, Switzerland: Trans. Tech.

Crawford, J. H., Slifkin, L. M. (Eds.) (1972), *Point Defects in Solids*, New York: Plenum, Vols. 1–3.

Dayananda, M., Murch, G. E. (Eds.) (1985), *Diffusion in Solids, Recent Developments*, Warrendale, PA: The Metallurgical Society of AIME.

Lodding, A., Lagerwall, T. A. (Eds.) (1970), *Atomic Transport in Solids and Liquids*, Tübingen, Germany: Zeitschrift für Naturforschung.

Murch, G. E. (Ed.) (1992), *Unsolved Problems in Diffusion*, Zurich: Trans. Tech.

Murch, G. E., Nowick, A. S. (Eds.) (1984), *Diffusion in Crystalline Solids*, New York: Academic.

Murch, G. E., Birnbaum, H. K., Cost, J. (Eds.) (1984), *Non-traditional Methods in Diffusion*, Warrendale, PA: The Metallurgical Society of AIME.

Nowick, A. S., Burton, J. J. (Eds.) (1975), *Diffusion in Solids, Recent Developments*, New York: Academic.

Shaw, D. (Ed.) (1973), *Atomic Diffusion in Semiconductors*, New York: Plenum.

Sherwood, J. N., Chadwick, A. V., Muir, W. M., Swinton, F. S. (Eds.) (1971), *Diffusion Processes*, London: Gordon and Breach, Vols. 1 and 2.

Proceedings of Conferences Devoted to Diffusion in Solids

Bénière, F., Catlow, C. R. A. (Eds.) (1983), *Matter Transport in Solids*, NATO Advanced Study Institute, Series B, Physics, Vol. 97, New York: Plenum.

Chadwick, A. V., Terenzi, M. (Eds.) (1986), *Defects in Solids—Modern Techniques*, NATO Advanced Study Institute, Series B, Physics, Vol. 147, New York: Plenum.

Kedves, F. J., Beke, D. L. (Eds.) (1983), *DIMETA-82, Diffusion in Metals and Alloys*, Aedermansdorf, Switzerland: Trans. Tech.

Kedves, F. J., Beke, D. L. (Eds.) (1990), *DIMETA-88, Diffusion in Metals and Alloys*, Aedermansdorf, Switzerland: Trans. Tech.

Laskar, A. L., Brebec, G., Bocquet, J. L., Monty, C. (Eds.) (1990), *Diffusion in Materials*, NATO Advanced Study Institute, Series E, Applied Sciences, Vol. 179, Dordrecht: Kluwer.

IONOSPHERE

Alv Egeland, *University of Oslo, Oslo, Norway*
Nelson C. Maynard, *Phillips Laboratory, Hanscom Air Force Base, Massachusetts, U.S.A.*

Introduction 223
1. **Definition and Historical Preamble** 223
2. **The Earth's Upper Atmosphere and Magnetic Field** 224
3. **Ionization Sources** 228
4. **The Ionospheric Layers** 229
4.1 The Topside Ionosphere (>400 km) 230
4.2 The F Region (>140 km) 230
4.3 The E Layer (90–140 km) 231
4.4 The D Region (50–90 km) 231
5. **Global Distribution of Ionization** 232
6. **Electric Fields, Conductivity, and Currents** 233
7. **Aurora—The Polar Light** 236
8. **Dynamics of the Ionosphere** 239
9. **Radio Propagation Effects** 240
10. **Planetary Ionospheres** 240
11. **Questions for the Future** 241
Glossary 242
Works Cited 242
Further Reading 243

INTRODUCTION

The formation, structure, and dynamics of the ionosphere—a layer of partially ionized gas extending from about 50 to beyond 1000 km—are the central topics of this article. Up to the start of the space age—around the 1960s—our main information about the ionosphere was obtained by the use of radio waves reflected from the different ionospheric layers. During the last three decades instrumented rockets and satellites have been widely used in the studies of the ionosphere, particularly the ionosphere poleward of approximately 60° latitude.

The polar ionosphere is much more complicated than the ionosphere at lower latitudes. While solar radiation with wavelengths below ~250 nm is the principal ionization source of the low-latitude ionosphere, energetic electrons and ions from the magnetosphere and solar wind significantly contribute to the production of the polar ionosphere. These charged particles are also the main source of the aurora, contributing to intense electric fields and currents within the ionosphere. The electric and magnetic fields control the motion of electrons and ions within the ionosphere and thereby influence the global distribution of ionization.

After a short historical preamble and a brief review of the Earth's neutral atmosphere and its magnetic field, we will concentrate on the production and the characteristics of the ionosphere. Physical and chemical processes as well as interactions between the different ionospheric regions and the magnetosphere will be discussed. A section on ionospheres of other planets together with a short summary of items that should have high priority during the coming decade conclude this article.

1. DEFINITION AND HISTORICAL PREAMBLE

The Earth's ionosphere extends from ~50 to over 1000 km and is a partially ionized gas that envelops the Earth, interfacing the upper atmosphere with the magnetosphere. Electrons in concentrations from 10^8 to greater than 10^{12} m^{-3} roam free from their parent ions and move with their thermal speed under the influence of the ambient electric (**E**) and magnetic (**B**) fields in an anisotropic manner relative to the geomagnetic field. As the atmo-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

spheric density decreases with increasing altitude, the motion of charged particles changes from collision-dominated to collisionless motion following **E**×**B** drift paths. The electrons carry currents and influence radio-wave propagation, an effect that historically provided the first clues of the ionosphere's existence.

When Appleton and co-workers in 1925 had demonstrated the existence of the Earth's ionosphere by the use of radio waves and obtained reliable measurements of its height, studies were begun of both the nature of the upper atmosphere and the ionospheric formation (cf. Ratcliffe, 1972). Naturally, one first studied the simpler ionospheric aspects and only later concentrated on the more complicated polar ionosphere. Actually, between 1925 and 1955 ground-based radio measurements led to a fairly complete understanding of the average characteristics of the ionosphere and the atmosphere between the heights of about 90 and 300 km. The name ionosphere for this region was coined by R. A. Watson-Watt in 1926.

Quite a bit was known about the "ionosphere" even before it was discovered by the radio technique. The First International Polar Year (1882–83) could be regarded as marking the beginning of modern ionospheric and auroral research. Studies of the Earth's magnetic field by Stewart (1882) suggested that the upper atmosphere acted as the armature of an "atmospheric dynamo." Thus, the current hypothesized by Stewart explains the regular daily variations (S_q; S is short for Sun and q for quiet) of the geomagnetic field observed on the ground at mid and low latitudes. Two important suggestions regarding the ionosphere were made around the turn of our century, namely:

1. A. E. Kennelly and O. Heaviside suggested the existence of a layer of ionized gas in the upper atmosphere to explain G. Marconi's transmission of radio signals from England to Newfoundland.
2. On the basis of data from a network of geomagnetic observatories in the auroral regions, Birkeland (1908) introduced and estimated the heights of "ionospheric currents" 25 years before the basic character of the ionosphere was actually established (cf. Ratcliffe, 1972; Chapman and Bartels, 1940).

Probably the most spectacular of nature's phenomena, the aurora (also called the northern light and/or the polar light) has stirred man's imagination, curiosity, and fear since the dawn of man. The aurora is the visible means by which we can observe the coupling of energy from the Sun via the magnetosphere to the ionosphere and atmosphere. These processes involve field-aligned currents (now referred to as Birkeland currents) and horizontal currents (the auroral electrojet) in the polar ionosphere. It was from studies of the aurora around the turn of the century that sunspots and charged particles were first tied to ionospheric dynamics by Birkeland (1908).

Man's interest in the ionosphere mainly developed because of the need for long-distance radio communications and out of general curiosity. A fairly detailed ionospheric knowledge would be desirable in improving and extending communication quality. Today, modern, *in situ* ionospheric research is mainly conducted through the use of sophisticated instruments on board rockets and satellites, closely coordinated with data from advanced ground-based equipment. Even active controlled experiments—e.g., using radio transmitters to heat the ionosphere from the ground and/or chemical and charged-particle injections from rockets and satellites—are today regularly performed. The use of instrumented rockets and satellites, beginning about 1955, at first mainly confirmed previous findings and ideas about the ionosphere. However, after about 1960 new information about the topside ionosphere (i.e., above about 400 km) as well as the lower region (i.e., below 90 km) was obtained. In addition, new ground-based techniques, such as partial-reflection, cross-modulation, coherent-scatter, and incoherent-scatter radars, have yielded a lot of new information about the ionosphere during the last 30 years.

2. THE EARTH'S UPPER ATMOSPHERE AND MAGNETIC FIELD

The Earth's upper atmosphere and its magnetic field strongly influence the structure and dynamics of the ionosphere. The atmospheric composition affects the density of charged particles while the geomagnetic field influences charged-particle transport. Only those aspects important to the ionosphere will be discussed here.

We define the upper atmosphere as the atmosphere above the stratopause, i.e., above ~50 km. This is also where ionization levels become significant. The atmosphere has no sharp upper limit, but here we will concentrate on the atmosphere below ~1000 km.

Density, composition, and temperature are the atmospheric properties of primary importance to the ionosphere because of their effects on collision frequency and recombination processes, respectively. The average density profiles for the major atmospheric constituents of an International Standard Atmosphere are sketched in Fig. 1 (see also ATMOSPHERIC PHYSICS; ATMOSPHERIC STRUCTURE). Typical temperature curves are shown in Fig. 2. The density curves represent idealized, year-round, mid-latitude conditions under mean solar activity. Large temporal and spatial variations occur, especially above 200 km. The primary gases up to ~120 km are molecular oxygen (≈20%) and nitrogen (78%), while atomic oxygen is very important above 120 km. Minor constituents such as water vapor (H_2O), nitric oxide (NO), hydroxyl (OH), and sodium (Na) are all important for photochemical processes in the upper atmosphere up to 100 km. Notice that density and pressure decrease rapidly (approximately exponentially) with height up to approximately 100 km. Above this height the decrease is significantly slower (see Fig. 1). The atmosphere is considered to rotate with the Earth (cf. *U.S. Standard Atmosphere*, 1976).

The energy budget for the upper atmosphere, whose main energy source is ultraviolet radiation from the Sun, is sketched in Fig. 3. Electromagnetic radiation from the Sun (in units of $\mathrm{W\,m^{-2}\,nm^{-1}}$) outside the atmosphere and at the Earth's surface is shown in Fig. 4. The albedo is approximately 0.3, which means that ~30% of the radiation is reflected back into space. Roughly 45% of the energy is found within the visible spectrum, normally defined as from 400 to 700 nm, while approximately 10% of the energy is within the ultraviolet (uv) and x-ray bands. The total energy—integrated over all wavelengths—is called the solar constant because it is practically independent of solar activity; however, the uv segment of the spectrum varies significantly with solar activity. The solar constant outside the atmosphere is $\approx 1.370\ \mathrm{kW\,m^{-2}}$. If averaged over the whole sphere, a realistic value at the Earth's surface of $250\ \mathrm{W\,m^{-2}}$ is obtained. The upper atmosphere acts as a filter because, as it becomes ionized, it absorbs practically all x rays and uv radiation from the Sun (i.e.,

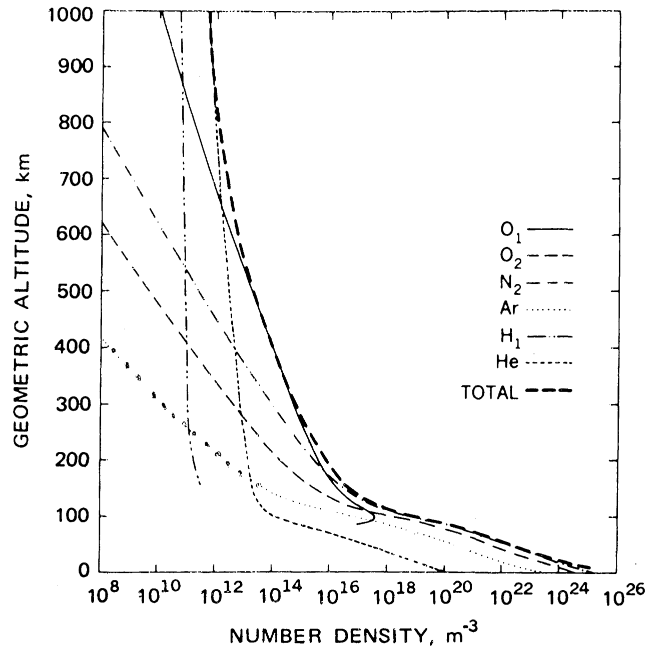

FIG. 1. Number densities of individual species and total number density as functions of altitude for the International Standard Atmosphere (from *U.S. Standard Atmosphere*, 1976).

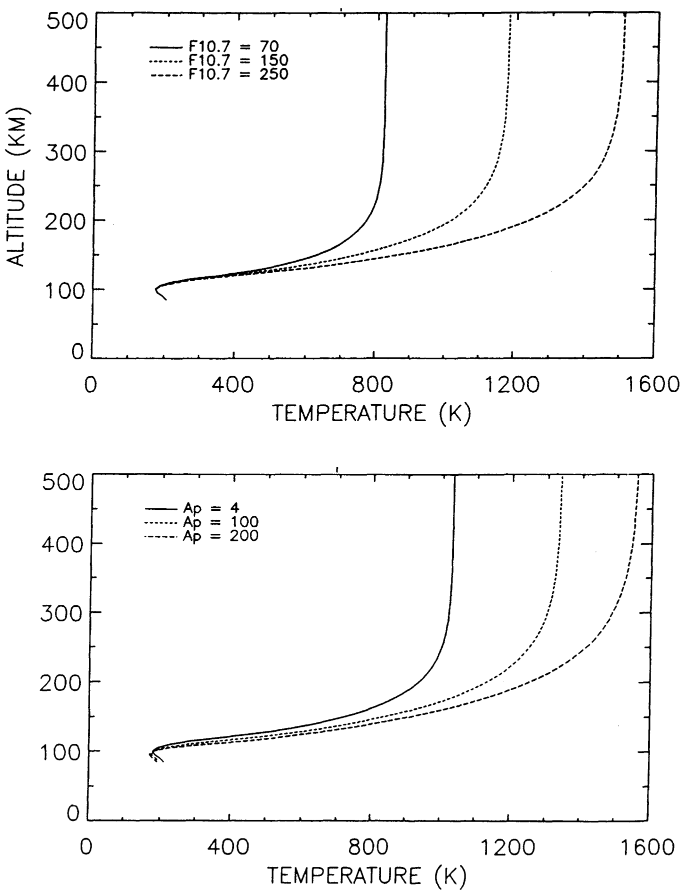

FIG. 2. Neutral temperature as a function of altitude. Top: The variation at 14 h LT and latitude 30° with solar activity, represented by the magnitude of the F10.7 flux (F10.7 is the electromagnetic radiation from the Sun at the wavelength 10.7 cm). Bottom: The variation at 2 h LT and latitude 65° with changes in magnetic storm activity represented by the global magnetic activity index A_p. The solar flux was held constant at the middle level of the top panel. The profiles are from the MSIS-90 model of Hedin (1991) [courtesy of F. Marcos, Geophysics Directorate, Philips Laboratory (USAF)].

solar radiation below ~250 nm, which we in the following will call the xuv band).

Notice that atmospheric waves generated close to the Earth contribute to the energy budget around 100 km. Also, as will be discussed later, energy from the magnetosphere, in the form of charged-particle precipitation and Joule heating, heats the upper atmosphere in the high-latitude regions. Neutral winds are also important for the energy balance. Winds and waves, as well as turbulence, affect the transport of energy to different regions of the ionosphere (see Fig. 3).

The temperature is variable and decreases in the mesosphere, reaching a minimum near 90 km. Above this height the temperature gradient is always positive (Fig. 2). A large fraction of the absorbed ultraviolet radiation from the Sun goes into heating of the upper atmosphere. Infrared radiation to outer space is a major energy loss (Fig. 3). The vertical temperature distribution, which results from the delicate balance of the various sources of energy and energy transport, varies considerably as it increases between 100 and 200 km. The temperature continues to increase up to approximately 500 km, where it approaches an asymptotic value (above about 500 km temperature no longer has a clear physical meaning). The top panel of Fig. 2 shows the variation of the dayside temperature with increasing solar radiation at 1400 LT (local time) and latitude 30°. The bottom panel shows the high-latitude, nightside temperature variability with magnetic activity at the middle level of solar flux, but at 0200 LT and latitude 65° in the auroral zone. Energy is deposited at high latitudes from the magnetosphere by energetic particle precipitation and field-aligned currents. Local heating affects both neutral temperature and composition, which in turn affect the ionization level (discussed in Secs. 4 and 5).

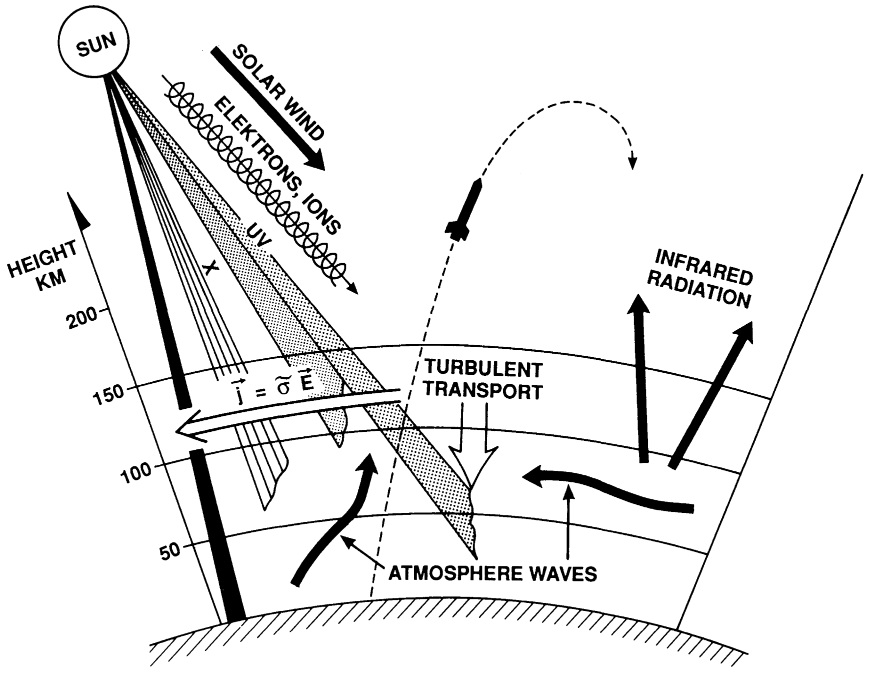
SUN
SOLAR WIND
ELEKTRONS, IONS
UV
X
HEIGHT KM
200
150
100
50
$\vec{j} = \tilde{\sigma} \vec{E}$
TURBULENT TRANSPORT
INFRARED RADIATION
ATMOSPHERE WAVES

FIG. 3. A schematic representation of the energy budget of the upper atmosphere. The main energy source is electromagnetic and particle radiation from the Sun, but even atmospheric waves contribute to the energy budget around 100 km. The main energy sink is infrared radiation.

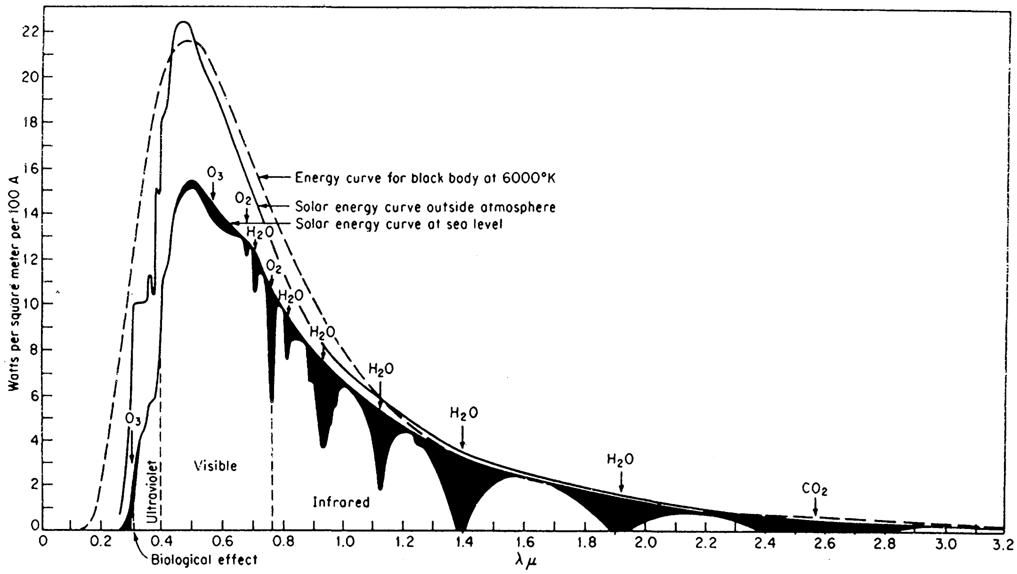

FIG. 4. Spectral energy curves of the solar energy outside the atmosphere and at sea level. Atmospheric absorption is indicated. For comparison the energy curve for a black body at 6000 K is also drawn. (From Egeland *et al.*, 1973.)

The solar xuv band significantly influences the upper atmosphere and changes its composition dramatically, particularly through photodissociation of atmospheric molecules (X_2) according to the formula

$$X_2 + h\nu \rightarrow X + X, \qquad (1)$$

where $h\nu$ is a photon—within the xuv band—with energy sufficient to dissociate atmospheric molecules. Equation (1) explains why molecules dominate the region below ~150 km, whereas atoms (nitrogen and oxygen) are the primary particles above this altitude (Fig. 1).

Diffusive separation becomes increasingly important with increasing altitudes, with the heavier constituents experiencing more gravitational attraction and the lighter ones less. As a result, the light gases dominate the upper atmosphere. Thus two elements that are rare near the Earth [i.e., helium (He) and hydrogen (H)] become relatively more and more abundant at higher altitudes. Above 600 km, layers of He and H dominate (cf. Fig. 1) because of their low masses. More important for the physics of the upper ionosphere is the existence of a several-hundred-kilometer-thick layer of atomic oxygen (O) above about 250 km. Even between 120 and 250 km, O is a very important constituent.

Because the Earth's magnetic field controls charged-particle motion in the ionosphere, a few of its main characteristics will be given here. The basic shape of this field is that of a dipole—similar to that of a bar magnet. However, significant differences exist that cause the magnetic field to deviate from the dipole shape and require a 15th-order spherical harmonic expansion to describe it accurately in the ionosphere. The intensity of the geomagnetic field gradually increases from the magnetic equator to the pole, where its strength of roughly 60 000 nT is nearly twice that at the equator. The Earth's magnetic axis is offset from its geographic axis in the northern hemisphere by ~11.5° and in the southern hemisphere by nearly 15°. Thus the dipole axis is offset from the center of the Earth. Magnetic ore deposits cause small spatial perturbations. Strong, variable electric currents in the ionosphere and magnetosphere further perturb the dipole shape. Because these time-varying currents (especially those associated with the aurora; cf. Sec. 8) cause large temporal perturbations in the Earth's magnetic field, a magnetic compass is not always an accurate aid in the auroral zones.

The dipole field decreases with distance from the Earth's surface as R^{-3}. At a distance of more than ~4 Earth radii the magnetic field is significantly compressed on the dayside of the Earth, while it is stretched out into a tail on the nightside by the continuous interaction of the magnetic field with the solar wind (cf. Fig. 4 of GEOELECTROMAGNETISM; also MAGNETOSPHERE). The resulting paraboloid-shaped surface with its nose at local noon encloses the region called the Earth's magnetosphere; i.e., the region where charged particles have greater energies than thermal particles and where their motion is controlled mainly by the Earth's magnetic field. The position of the nose depends on the balance of the solar-wind pressure against the magnetic pressure of the Earth's magnetic field, which dominates over the plasma pressure in the outer magnetosphere. The standoff distance is typically at 10 Earth radii. The magnetic field lines that connect the magnetosphere to the ionosphere guide the motion of charged particles and thus provide the paths for the interchange of currents and energy. At polar latitudes, where the magnetic field lines are nearly vertical, ions and electrons can escape from the ionosphere in a process termed the polar wind.

3. IONIZATION SOURCES

The atmosphere above 50 km and extending out to the order of 1000 km is called the ionosphere, because in this region charged particles—free electrons and ions—exist in large enough numbers to influence the propagation of radio waves. Ions are created by attachment of free electrons, by collisions with energetic particles, or by photoionization. The latter two processes create free electrons. The density of charged particles in any particular region depends on the balance between production and loss processes. All of these processes are dependent on the upper atmosphere composition and density.

The solar xuv band is particularly important for producing the ionosphere—via the process called photoionization, which can be written symbolically

$$X + h\nu \rightarrow X^{+} + e_n, \qquad (2)$$

where X is an atmospheric gas, atom, or molecule, $h\nu$ is solar radiation within the xuv band, while $X^+ + e_n$ is an ion–electron pair. Only photons within the xuv band have sufficient energy to ionize the atmospheric gases. Thus the xuv radiation changes the upper atmosphere from electrically neutral to electrically conducting. X^+ is a positive ion while e_n is a free, low-energy electron. The closer the solar radiation reaches the Earth, the denser the atmosphere. The production of ion pairs increases until a large fraction of the radiation is absorbed. It is therefore a combination of atmospheric density and composition affecting solar xuv radiation absorption that determines the height of the ionosphere.

The density of electrons and ions is determined by the balance between production and loss. Electrons and ions are lost by recombination processes

$$X^+ + e_n \rightarrow X, \tag{3a}$$

$$YX^+ + e_n \rightarrow X + Y, \tag{3b}$$

in which X is an atom and XY a molecule. Electrons and ions may also be transported away and thus lost to a particular area. Because solar xuv radiations vary with time, season, and location, large temporal variations are found in the concentrations of ion pairs in the ionosphere. Other variations result from different recombination rates for different atmospheric constituents and from variations in the magnetospheric particle input. For instance, NO^+ recombines with electrons much more quickly than O_2^+, and so spatial changes in the relative abundance of these constituents create local variations in the ion or electron density.

At low and medium latitudes photoionization [Eq. (2)] is mainly responsible for the creation of charged-particle distributions. At higher latitudes precipitating energetic electrons and ions from the magnetosphere and from the solar wind (a stream of charged particles with an average speed of 400 km s^{-1} and a density of $\sim 10^7\ m^{-3}$ constantly blowing towards the Earth's magnetosphere) significantly contribute to the production of ionospheric electrons via the following processes:

$$X + e \rightarrow X^+ + e_n + e', \tag{4a}$$

$$X + i \rightarrow X^+ + e_n + i' \tag{4b}$$

$$(e' = e - 36\ \text{eV},\quad i' = i - 36\ \text{eV}),$$

where e and i are energetic ($\geqslant$ 100 eV) electrons and ions, respectively. [One electron volt (1 eV) is the energy an electron gets when passing through a voltage drop of 1 V.] Each collision process of the type shown in Eq. (4) decreases the energy of an incoming particle by an average 36 eV. A 10-keV particle bombarding the atmosphere will on the average participate in 300 collisions and thus produce 300 ion pairs before it comes to a complete stop at or above ~90-km altitude.

Notice that in the processes listed in Eqs. (2) and (4) the molecules and/or atoms will have lost an electron and thereby become positively charged. A gas that contains free electrons and positive ions is called a plasma. Because an ionospheric plasma contains roughly the same number density of electrons and ions, it is for all practical purposes electrically neutral.

In the polar regions (i.e., poleward of ~60° magnetic latitude), the energy source from the solar wind and its interaction with the magnetosphere is of particular importance. The solar wind is indirectly the main generator of auroras, electric fields, and currents in the upper atmosphere (i.e., ⩾100 km) (cf. Secs. 6 and 7).

4. THE IONOSPHERIC LAYERS

On the basis of the results of ionospheric studies by radio technique, it has been agreed to divide the ionosphere into different regions as illustrated in Table 1. The first to be identified was a layer of ionization near 100 km, the E layer, so named because of the electric field (i.e., the E field) of the reflected radio waves with which it was measured. Above the E region lies the F region, while the region underlying was named D.

A typical ionospheric electron density profile for mid-latitude, daytime conditions is shown in Fig. 5. The different regions reflect radio waves with different wavelengths. While short waves (i.e., ≃100-m wavelengths equivalent to a frequency ⩾3 MHz) are reflected from the F region, the longer radio waves (i.e., ~1000 m) require less density for reflection and are reflected from the E region. These reflection differences allow the ionosondes to measure the ionization profile with height by

Table 1. Characteristics of the ionospheric regions.

		Electron density		
Region	Height (km)	Day (m^{-3})	Night (m^{-3})	Production sources
D	50–90	10^9	10^8	xuv+cosmic rays
E	90–140	10^{11}	$<10^{10}$	xuv+auroral particles
F_1	140–200	3×10^{11}	10^{10}	uv+auroral particles
F_2	200–400	10^{12}	10^{11}	uv+transport from below

measuring the variation of travel-time delay of the reflected radio signal with frequency.

At higher frequencies or shorter wavelengths signals pass through and are not reflected. Thus the variable ionosphere has a pronounced influence on long-distance radio communication and navigation. By comparing the electron density given in Table 1 for the D and E regions with Fig. 1, it is clearly seen that only a tiny fraction of the atmospheric gases are ionized. The majority of the molecules remain neutral in the lower ionosphere.

4.1 The Topside Ionosphere (> 400 km)

The region above the peak density of the F layer is called the topside ionosphere. The density decreases slowly with increasing altitude, with the rate of decrease being a function of temperature. This region was called the topside because its density profile could not be determined by radio-wave sounding techniques from the ground. Its characteristics were first determined by *in situ* sounding from above with transmitters on satellites and by sounding rocket measurements. In the lower part O^+ and O are the major ion and neutral species, respectively. Above approximately 600 km the lighter atomic ions He^+ and H^+ can dominate. In the upper part of the topside ionosphere, the plasma is effectively fully ionized.

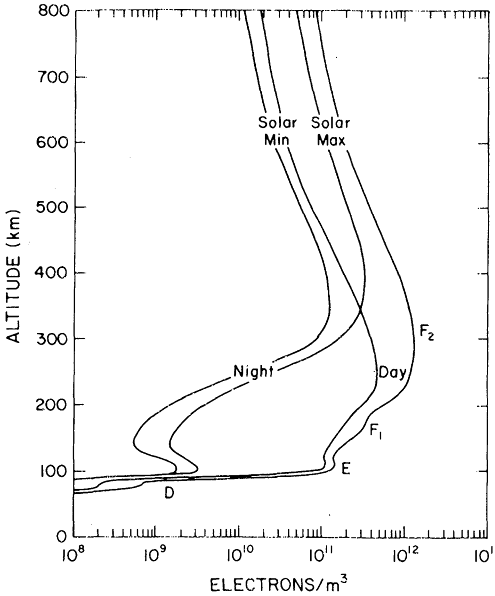

FIG. 5. Typical mid-latitude electron density profiles for solar maximum and minimum for both daytime and nighttime conditions. The different ionospheric layers are marked off in this figure (from Rees, 1989).

4.2 The F Region (> 140 km)

Above the E layer (i.e., > 140 km up to a height arguable in the scientific community) lies the F region. In the following we will define the upper height as that where the light ions (helium and hydrogen) begin to dominate over the heavier ions, in particular oxygen (between 500 and 1000 km depending on conditions). This height will vary with temperature. Note that the F region overlaps the lower part of the topside ionosphere. Even if production of ionization peaks below 200 km, the maximum electron density is found at considerably greater (up to 350 km) heights. In 1933 it was found that the F region during daytime often has two maxima, thus splitting into F_1 and F_2 layers. Typical peak electron densities are 3×10^{11} and 10^{12} m^{-3} in the F_1 and F_2 layers, respectively, while the neutral density is approximately 10^{15} m^{-3}. The most important ions in the F region are O^+, NO^+, O_2^+, and N_2^+. The important F-layer ionization source is found in the xuv band between 10 and 90 nm. It should be pointed out that N_2^+ reacts quickly with O via the process

$$N_2^+ + O \rightarrow NO^+ + N. \quad (5)$$

In the F_1 region (normally only a ledge in the lower F region with peak density near 180 km) the following reaction is also important:

$$O^+ + O_2 \rightarrow O_2^+ + O. \tag{6}$$

The conclusion is that NO^+ and O_2^+ together with O^+ are the most numerous ions below 200 km, while O^+ and the light ions dominate in the upper part of the F region (i.e., F_2 layer). Thus most of the ions in the F_2 layer are atomic.

The continuity equation for the F_2 layer can be written

$$\frac{dn_e}{dt} = q - \beta n_e + \mathbf{v} \cdot \nabla n_e, \tag{7}$$

where q is the production rate of electrons, n_e is the electron density, and β is the recombination (attachment) coefficient. Note that the loss of electrons depends linearly on electron density. In addition, the last term in Eq. (7) shows that there are pronounced drift effects. Electrons and ions are also able to escape from the ionosphere along open magnetic fields at polar latitudes. This loss of plasma—called the polar wind—can have a significant effect in the polar ionosphere. Plasma flow from one hemisphere to the magnetic conjugate ionosphere is important at midlatitudes. The inner magnetosphere or plasmasphere is populated by outflow of plasma along magnetic field lines from the ionosphere at mid latitudes.

4.3 The E Layer (90–140 km)

The E layer was the first ionospheric layer to be recognized. Since the typical daytime electron density is 10^{11} m^{-3}, and the neutral density at 120 km is 10^{17} m^{-3}, the E region is weakly ionized, and collisions between charged particles and neutral species are important in determining the conductivity and ion motion. The chief ions are O_2^+ and NO^+. The dominant neutral species in this region are N_2 and O_2. In the upper part of the layer O^+ occurs in significant quantities. N_2^+ is also produced but rapidly lost through charge exchange via the reaction

$$N_2^+ + O_2 \rightarrow N_2 + O_2^+. \tag{8}$$

In the lower part of the layer, NO^+ dominates.

Because NO has such a low ionization potential, Lyman-α radiation ($\lambda = 121.6$ nm) is the main ionization source. The predominant loss term is dissociative recombination, which takes the form [cf. also Eq. 3(b)]

$$O_2^+ + e_n \rightarrow O_2 \rightarrow O + O. \tag{9}$$

[In process (9) both momentum and energy are conserved, which makes this process very fast.] The continuity equation for this layer can be written

$$\frac{dn_e}{dt} = q - \alpha\, n_e^2, \tag{10}$$

where α is the recombination coefficient for all reactions of the type described by Eq. (9). Because of the high neutral density, drift effects of type shown by the gradient term in Eq. (7) can be neglected. More recent results based upon equilibrium between production and loss (i.e., no time variations) have given α values in good agreement with laboratory measurements. Also, rocket measurements have been very useful in testing Eq. (10).

Near 100 km a thin layer of enhanced ionization, known as sporadic E, sometimes exists. This layer can contain enough electrons to reflect high-frequency (HF) radio signals that would normally propagate through to the F region.

4.4 The D Region (50–90 km)

Because of the relatively high neutral density in this region, the low ion densities, and the dominance of collisions, the classical radio techniques used to study the E and F regions of the ionosphere are not applicable. Rocket techniques provide the only *in situ* measurements since this region is not accessable to satellites and balloons. Electron density does not show a marked peak, but rather a gradual increase with height (cf. Fig. 5).

The important ionization sources are listed in Table 1. X rays between 0.1- and 1-nm wavelengths are a principal source below 90 km. Solar protons from flares on the Sun also create D-region ionization. The bottom of the D region, where the electron density is very low (i.e., 10^8 to 10^9 m^{-3}), is produced by galactic cosmic rays. Thus there are hardly any diurnal variations in the lower D region during quiet conditions.

The primary ions are NO^+ and O_2^+. In addition, negative molecular ions and hydrated ions (water-vapor complexes) have been observed. One reaction producing negative ions may be

$$2O_2+e_n \rightarrow O_2^- + O_2. \qquad (11)$$

Production (of particularly complicated positive and negative ions), recombination, and other loss processes in the D region are complicated and not fully understood.

Because the D region is important for the energy balance as well as a principal region for absorption of radio waves propagating over long distances, further studies by extending modern incoherent-scatter radar and lidar techniques to this region have just begun.

5. GLOBAL DISTRIBUTION OF IONIZATION

Detailed studies of the ionosphere by simultaneous, coordinated satellite, rocket, and ground observations have clearly demonstrated large temporal and spatial ionospheric variations. Before the space age, studies were limited mainly to the regular ionospheric variations; i.e., seasonal, diurnal, and solar-cycle responses. These variations are at least partially due to the rotation of the Earth with its geographic axis tilted relative to the Sun and the geomagnetic axis tilted relative to the geographic axis, and long-term activity variations on the Sun. Both the photoionization, the main source for maintaining the medium- and low-latitude ionosphere, and magnetospheric and solar-wind ionization contributions will also vary considerably and thereby add to the complexity of the global distribution.

As seen from Eq. (10), the electron density in the E region during stationary conditions is

$$n_e=(q/\alpha)^{1/2}=c_1(\cos\chi)^{1/2}, \qquad (12)$$

because the ion production rate q is proportional to the cosine of the zenith angle (χ) of the Sun. Because of the regular diurnal variation alone, the electron density will thus vary as $\cos\chi$. During winter in the northern hemisphere, when the polar region is completely in darkness for a few months, photoionization is practically absent, and precipitating particles become the principal ionization source for the E region.

Because the lifetime of ions is much longer in the F regions than below, they can be transported considerable distances from their origin. The polar F-region densities vary greatly depending on the proximity to auroral particle and dayside uv sources. Just equatorward of the nocturnal auroral oval a minimum in electron density, called the electron trough, is very pronounced during the winter months. This trough has a limited latitudinal extent and it is very deep from 16, to 0600 MLT (magnetic local time). During this time, the F-region density can be as low as 10^9 m^{-3} (cf. Table 1).

To illustrate the global distribution of ionization and its seasonal variation, Plate 1 presents the local time-averaged global distribution of plasma density versus month of the year. These measurements are from the Defense Meteorological Satellite Program (DMSP) satellite orbiting at 830 km and are averaged over a 2-y period. The distribution at the topside ionosphere is representative of other F-region altitudes, although variation with altitude will occur. All p.m. local times are averaged together. The dayside low-latitude maximum, which stretches to higher latitudes in the summer hemisphere, is clearly evident, as is the depletion in the winter polar cap. The trough at subauroral latitudes does not appear because this latitude variation is lost by averaging. A polar view of a typical ionization distribution in the southern high-latitude regions at the altitude of F-region maximum is shown in Plate 2(a) along with a similar plot in Plate 2(b) for 12 h later. This is a simulation using the Utah State University ionospheric model (Schunk, 1988) that has been run over a 24-h period with constant inputs. The polar hole, ionization from auroral oval precipitation, and the mid-latitude trough are clearly seen.

An additional diurnal variation occurs at high latitudes because of the offset of the geomagnetic pole from the geographic pole. Electrodynamically driven effects are controlled by the geomagnetic field, while neutral atmospheric and solar ionization effects are dominated by the geographic axis. As a result the ionosphere is stirred daily with the rotation of the geomagnetic axis about the geographic axis in universal time (UT). Comparing Plate 2(a) with 2(b) illustrates the maximum differences from this effect as these times were chosen when the magnetic pole was toward and away from the Sun relative to the geographic pole. These variations from stirring are added to those from solar and magnetic activity to produce very dynamic behavior of the ionosphere in the auroral and polar cap regions.

6. ELECTRIC FIELDS, CONDUCTIVITY, AND CURRENTS

There are three principal sources of electric fields in the global electrical circuit: the thunderstorm-generated atmospheric electric field, the tidal-wind–generated S_q ionospheric electric field, and the magnetospheric convection electric field driven by the interaction of the solar wind with the magnetosphere. The last two are important to the upper ionosphere, while all three interact at the lower boundary. The coupling of these generators is dependent on the ionospheric conductivity.

The motion of charged particles in the ionosphere depends on electric and magnetic fields, on the ionization density, and on collisions. Thus conductivity is directionally dependent. It may be divided into three terms: σ_0 is the specific conductivity along the magnetic field **B**, σ_P is the Pedersen conductivity in the direction perpendicular to the magnetic field and along the perpendicular component of the electric field **E**, and σ_H is the Hall conductivity in the direction perpendicular to both the magnetic and electric fields. The profiles of the conductivities with altitude are shown in Fig. 6. Note that the specific conductivity scale is six orders of magnitude greater than the other two. For this reason the electric field parallel to the magnetic field is, to first approximation, zero, and the perpendicular electric field maps along magnetic field lines. The Hall conductivity typically peaks near 110 km while the Pedersen conductivity peaks near 130 km. Above 200 km the collision frequencies are so small that the charged-particle motion is controlled strictly by the electric and magnetic fields. The ion velocity at these altitudes can be expressed by

$$\mathbf{v}_i = (\mathbf{E} \times \mathbf{B})/B^2. \qquad (13)$$

At high latitudes the magnetic field lines project out into the magnetosphere. The magnetospheric electric fields project along the magnetic field into the ionosphere and impose a potential pattern on the high-latitude regions. Electric fields at high latitudes are typically 20 to 60 mV m^{-1} and can exceed 200 mV m^{-1} during disturbed conditions. The ion flow is along the equipotentials, and this flow is also the principal driving force for neutral winds at F-region altitudes. When the interplanetary magnetic field has a southward component, the flow patterns form two convection cells with a negative potential minimum near 16 h and a positive potential maximum near 04 h. The shape of the patterns is dependent on the dawn–dusk component of the interplanetary magnetic field. Representative patterns are shown in Fig. 7. For positive (negative) B_y the BC pattern is applicable to the northern (southern) hemisphere while the DE pattern is applicable to the southern (northern) hemisphere. Typical values of the difference in potential between the maximum and minimum range from 30 to 150 kV depending on magnetic activity. In extreme cases

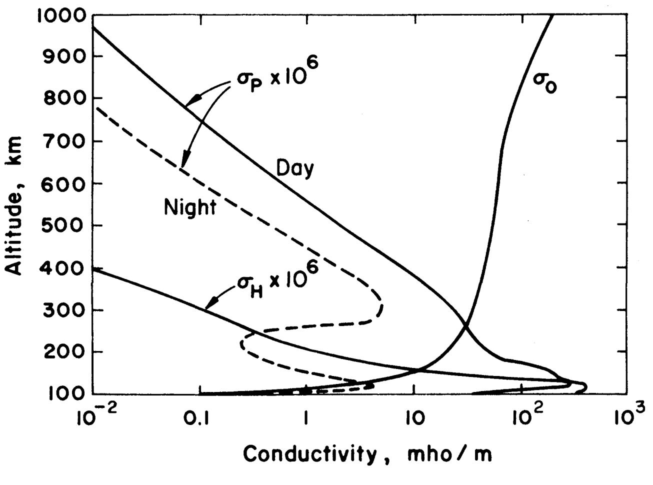

FIG. 6. Typical conductivity values for the mid-latitude daytime ionosphere. Notice the change of scale for the Hall and Pedersen conductivities. The dashed curve is a typical nighttime profile of the Pedersen conductivity also with the scale change (from Kelley, 1989; adapted from Johnson, 1961).

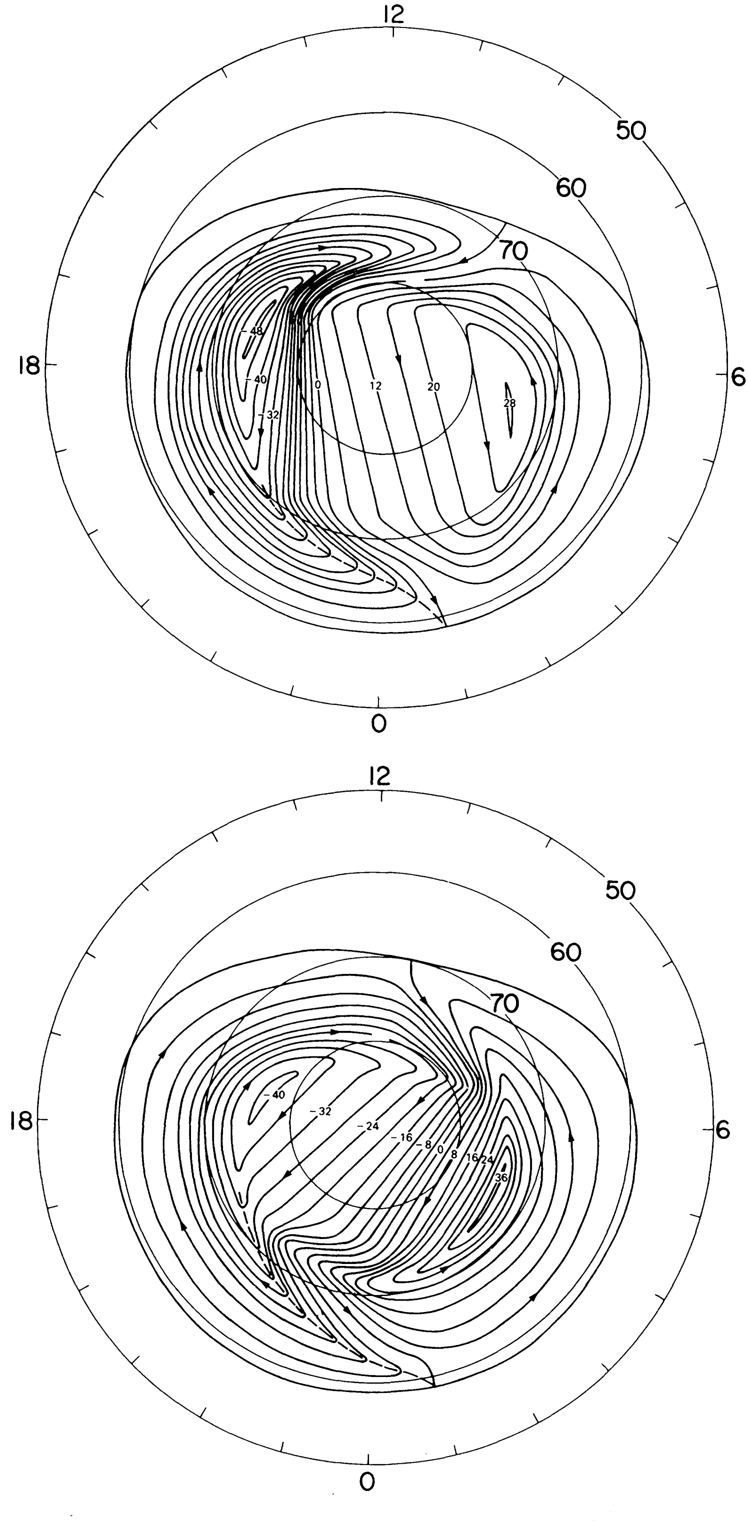

FIG. 7. Two convection patterns of Heppner and Maynard (1987) for a southward interplanetary magnetic field to illustrate the variation in the pattern with the east–west component (B_y) of the interplanetary magnetic field (IMF). The BC pattern (bottom) is for the northern hemisphere for positive B_y. The DE pattern (top) is for the northern hemisphere during negative B_y. For the southern hemisphere the same patterns apply with the opposite sign of the IMF B_y component. The contours represent equipotentials spaced at 4 kV and are representative of average activity conditions. The magnitude of the potential difference between the potential minimum on the evening side and the potential maximum on the morning side (76 kV in the above patterns) varies from ~20 to over 200 kV with varying magnetic activity.

it may exceed 200 kV. When the interplanetary magnetic field component B_z is northward, the two-cell patterns become distorted and can evolve into more than two cells. The potential differences are generally very small and the electric fields are extremely variable, in some cases making it difficult to recognize a pattern (Heppner and Maynard, 1987).

Under steady-state conditions the electric fields from the magnetospheric generator are poleward of about 60° magnetic latitude. A shielding layer in the inner magnetosphere prevents penetration of the magnetospheric electric fields inside of about 4 Earth radii, which projects along the Earth's magnetic field to 60° latitude. As magnetic activity changes, the shielding must be reestablished. Magnetospheric electric fields temporarily penetrate closer to the Earth and project along the Earth's nearly dipolar magnetic field lines, which are equipotentials, to lower latitudes in the ionosphere. At middle latitudes the ionospheric electric field is a combination of these penetration electric fields and S_q dynamo electric fields generated by neutral wind and tidal motions of the neutral atmosphere, which through collisions drive ions across magnetic field lines. This ionospheric E-region tidal generator is driven by a complex combination of solar-heating–induced neutral winds and gravitational effects from the Sun and Moon. Electric fields are small, of the order of a few mV m^{-1}, with a total potential drop of about 10 kV.

At the lower edge of the D region the gas is only weakly ionized and the conductivity becomes isotropic. The atmospheric electric field is vertical and couples to **E** fields from the other two sources. The atmospheric vertical electric field in the lower D region is small since 90% of the 300-kV potential drop between the Earth and the ionosphere exists below 10 km and the electric field is inversely proportional to the conductivity. Under conditions where all electrons are tied to heavy negative ions, gravitationally driven and/or differential neutral-wind–driven ion motion can generate anomalously large (V m^{-1}) local electric fields.

The ionospheric electric fields drive horizontal currents in the ionosphere, which are observed in the ground magnetometer recordings. These are principally Hall currents in a layer around 110 km (Fig. 6). At midlatitudes these are strongest during the day when the conductivity peaks. At the magnetic equator the Hall current would be principally downward and is impeded from flowing because of the decreasing conductivity. A polarization electric field results, causing an apparent enhancement of conductivity (Cowling conductivity) and an enhanced east–west current called the equatorial electrojet. The magnitude of the electrojet current peaks with the conductivity near 1100 LT (see Chapman and Bartels, 1940).

In high-latitude regions ionization from solar xuv radiation is supplemented by ionization from the energetic particles producing the aurora (see Sec. 8). Within the auroral oval the conductivity due to auroral particle precipitation dominates, and the conductivity varies with magnetic activity. The Hall currents driven by the magnetospheric convection electric fields are large. They are referred to as auroral electrojets because of the primary currents being confined in altitude to the peak of the Hall conductivity profile. On the evening side the currents are eastward, while after midnight they are westward. The two electrojets overlap in the midnight region at the Harang discontinuity. The conductivity, electric fields, and currents are extremely variable. Current continuity under these spatially and temporally variable conditions requires field-aligned currents, or Birkeland currents, to flow.

Birkeland currents couple the ionosphere with the magnetosphere. The upward currents carried by the precipitating auroral electrons are balanced by downward currents, primarily carried by upward-flowing, low-energy electrons. The large-scale pattern of field-aligned currents (shown in Figure 8) can be divided at most local times into two principal regions. Region-2 currents couple to the inner magnetosphere, and currents flow into the ionosphere on the evening side and out of the ionosphere on the morning side. Region-1 currents couple to the magnetospheric boundary layers and flow out of the ionosphere on the evening side and into the ionosphere on the morning side. Near noon and midnight these patterns overlap, and three current layers are common (Iijima and Potemra, 1978; Potemra, 1990). Additional weaker current systems are observed poleward of the oval especially during periods when the interplanetary magnetic field has a northward component. Field-aligned currents are closed in the

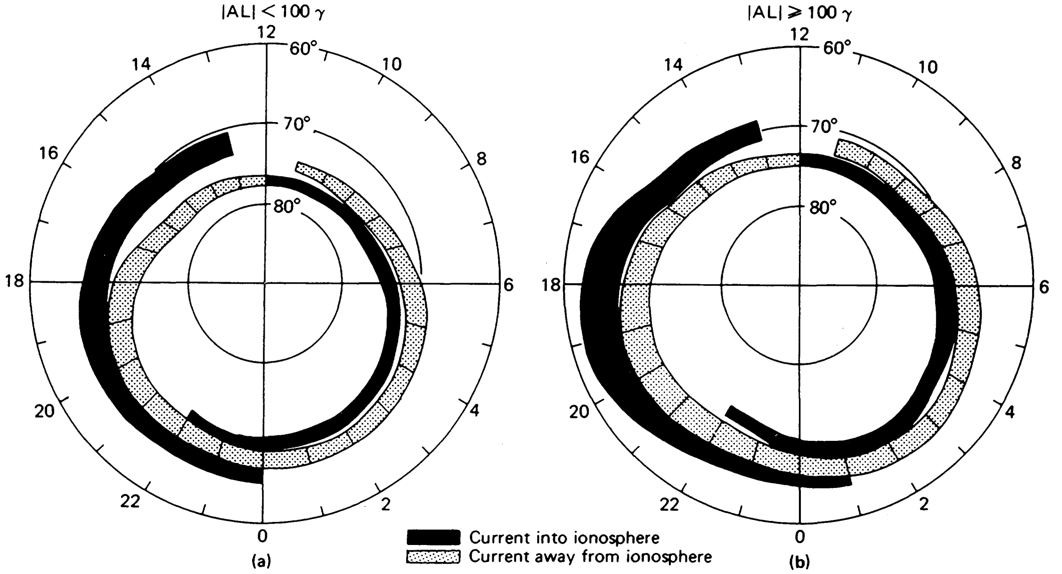

FIG. 8. A summary of the distributions and flow directions of large-scale field-aligned currents based on Triad data for weakly disturbed and active conditions (from Iijima and Potemra, 1978). AL is related to the maximum current intensity in the westward auroral electrojet and is the larger of the two components of the auroral activity index AE. The units are in nT (1 $\gamma = 1$ nT). Thus the left plot is for magnetically quiet conditions and the right plot is for active conditions.

ionosphere principally by Pedersen currents. In all these current regions the direction of the current is the average direction. Structured field-aligned currents in both directions exist side by side in all regions, a necessary consequence of the variability of the conductivity and the electric fields.

7. AURORA—THE POLAR LIGHT

The aurora, as illustrated in Plate 3, is certainly one of the most spectacular of nature's phenomena. In earlier times, inhabitants of northern Canada, Greenland, and the Nordic countries saw the northern lights as omens from the gods portending disaster, signs from deceased relatives, a battle of the gods, or a weather sign. For those interested in the history of the aurora, monographs by, e.g., Brekke and Egeland (1983) and Eather (1980) are available.

Those who admire the beauty of wonders in the dark sky may find nothing comparable to a night with magnificent auroral displays. Unfortunately, this impressive and imagination-inspiring display can only be seen regularly by people living near the polar region. At these latitudes, Aurora Borealis (i.e., in the north) and Aurora Australis (in the southern hemisphere) routinely appear in the so-called auroral zones. Consequently, auroras are far from most population centers. Even today, the area of the southern auroral zone (roughly around Antarctica) is inhabited only intermittently. In the northern auroral zone, which crosses Alaska, central Canada, northern Scandinavia, and northern Siberia, these majestic lights are a routine part of the daily life of these northern people.

The instantaneous distribution of aurora versus time and latitude is given by an oval-shaped belt of maximum activity called the auroral oval (cf. Fig. 9). These zones represent the average statistical occurrence of aurora and are approximately centered about the geomagnetic poles. Notice that both the location and width vary significantly from midnight to midday. They also vary with solar and magnetic disturbances as shown in Fig. 9. The oval expands equatorward during disturbed periods and contracts during quiet periods. Auroral emissions called polar cap arcs also appear poleward of the oval, but with markedly less frequency and intensity. The frequency of auroral occurrence, as well

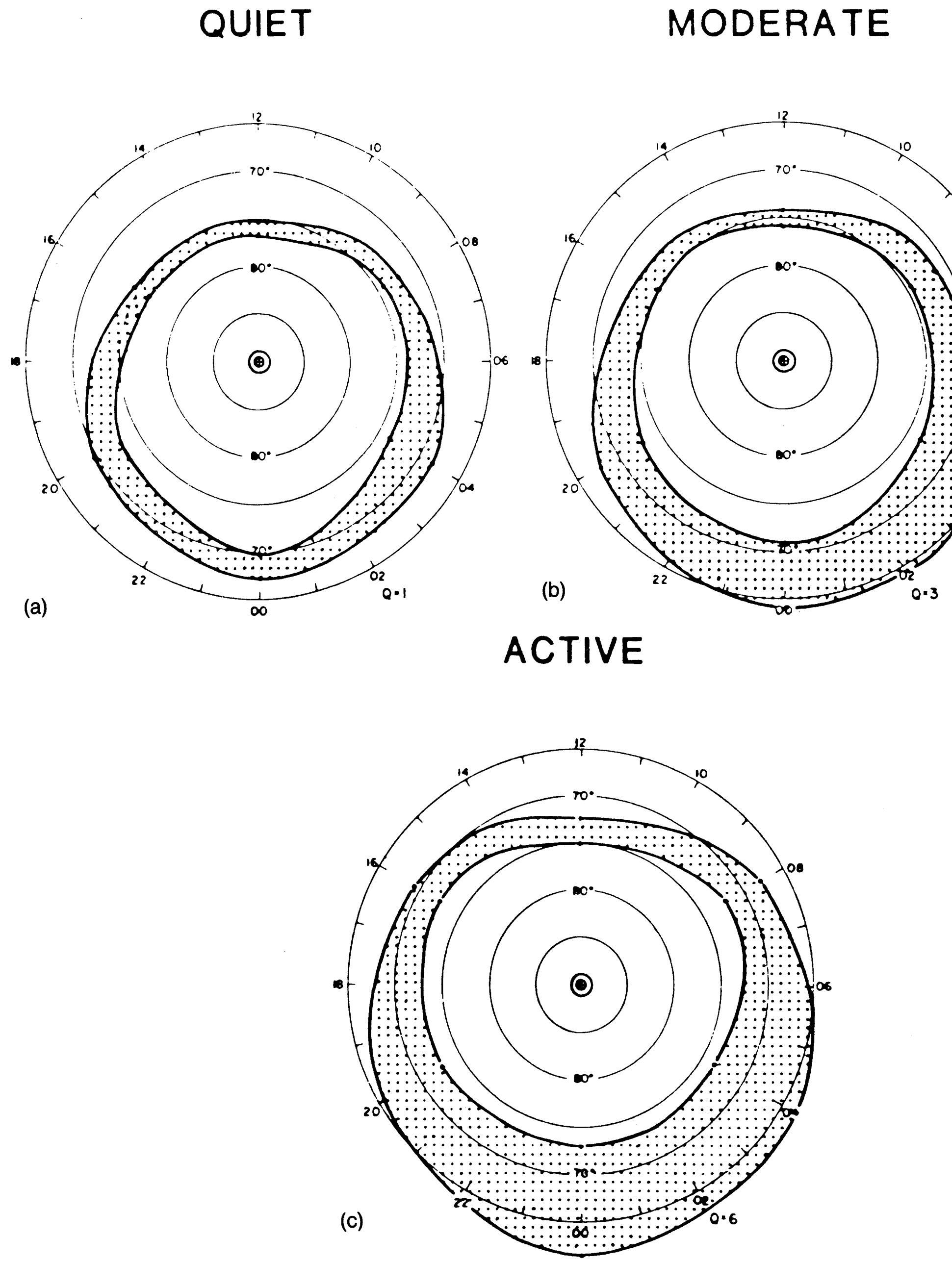

FIG. 9. Variation in the size of the auroral oval for different solar or magnetic activity; quiet, moderate, and disturbed conditions.

as its intensity and location, is clearly correlated with the degree of activity on the Sun. This is also borne out by the 27-d recurrence tendency (i.e., the rotation period of the Sun) as well as the 11-y sunspot cycle. However, auroral activity peaks one to two years after sunspot maximum. The oval width and location for quiet, moderate, and active conditions are illustrated in Fig. 9.

Auroral intensity can vary from being barely detectable among the stars to a luminance bright enough to veil the stars completely.

Thus the intensity of this polar light can vary over more than four orders of magnitude. At maximum intensity the illumination on the ground is comparable to moonlight, and ~100 times stronger than starlight. With a modern 35-mm camera, an f/1.4 lens, and a fast color film (e.g., 400 ASA), an exposure time of ~1 s is needed to photograph a bright aurora.

By watching the aurora, one may get the impression that a confusingly large number of different auroral forms and structures occur. However, by categorizing them into four primary forms—i.e., arcs and bands, auroral rays, spiral structures, and diffuse surfaces—the customary occurrence forms can be explained.

Of all characteristic properties of the northern light, none have been discussed and studied more than its height. On the basis of about 20 000 parallactic measurements between 1910 and 1940 (Störmer, 1955), it was proved that auroral arcs and bands, auroral glow, and surfaces lie predominantly within the height interval 95–150 km, whereas isolated rays or bundles of rays may have an altitude considerably higher (i.e., >200 km). The average height of dayside aurora is typically 150 km higher than nighttime aurora. A few auroral forms—called sunlit aurora—are found up to 1000 km. Today, the luminosity height profile may be measured *in situ* by instrumented rockets. A few tens of such flights have been carried out, showing that the height profile is directly related to the energy of the precipitating particles from the magnetosphere. The lower border near 100 km is excited by electrons of 5 to 15 keV. The peak emissions on the dayside, which typically are near 300 km, are from particles precipitating with energies of ten to a few hundred eV.

Although a weak aurora appears to the eye as white, increased intensity reveals the predominant green color together with red and blue. As illustrated in Plate 4, the optical spectrum of an aurora consists of a great number of spectral lines and bands—from ultraviolet to infrared wavelengths. These emissions are primarily due to a two-step process; i.e., precipitating energetic particles collide with the Earth's upper atmosphere, converting their kinetic energy into chemically excited states of atoms and molecules that in turn generate photons of wavelengths determined by the specific gas (i.e., oxygen, nitrogen, helium, etc.) and the relaxation processes. The auroral emissions of atomic lines and molecular bands can therefore be considered as the "fingerprints" of atmospheric constituents.

Symbolically, this can be written

$$N_2+(e \text{ and/or } i)\rightarrow(N_2^+)^*+(e' \text{ or } i')+e_n, \tag{14}$$

where e' $(i')=e$ $(i)-36$ eV and the asterisk denotes an excited state. Equation (14) is followed by radiation of the first negative bands:

$$(N_2^+)^*\rightarrow N_2^+ + h\nu \quad \text{(aurora at 391.4 and 427.8 nm).} \tag{15}$$

Because the probability of process (14) followed by (15) (spontaneous emission as excited species relax to a lower energy level) is nearly independent of energy of the primary particles, these blue auroral bands at 391.4 and/or 427.8 nm can be used to determine the net downward particle energy. Quantitative considerations show that one photon is emitted at $\lambda=391.4$ nm per 25 ion pairs produced.

The brightest auroral emission is the "green line" at 557.7 nm, which is due to atomic oxygen. But, as these emission photons are not emitted spontaneously as in process (15), it is a forbidden transition. Another commonly observed line—particularly in the polar cusp—is the red line at 630 nm, which is also emitted from excited oxygen. This is also a forbidden transition. Allowed or permitted transitions occur very rapidly (i.e., 10^{-7} s). Forbidden transitions, which violate the quantum-mechanical selection rules, still do occur, but only after a much longer time, i.e., ≈ 0.7 and 110 s for the green and the red lines, respectively. For the nearly 2-min lifetime of the red-line transition, collision deactivation (without light emission) is very likely for altitudes below about 200 km. Thus, the red 630-nm [OI] emission will always peak at altitudes above 200 km.

The intensity, height, and spectral content of the aurora are projections of magnetospheric processes onto the high-latitude ionosphere, much like the cathode-ray tube in a television set. The magnetic field lines to which the particles are tied trace out to the plasma sheet in the nightside and magnetospheric boundary layers and the cusp on the dayside. The aurora provides a continuous monitor of these activities. The advantage of using the aurora in the ionosphere as a monitor of these near-Earth processes over any conceivable system of *in situ* measurements is that the size over which the phenomena are occurring in the ionosphere differs in scale from that in the magnetosphere by a factor of perhaps as much as 10^6. Continuous high spatial and temporal resolution observations provide another advantage of studying ground measurements of aurora.

8. DYNAMICS OF THE IONOSPHERE

During recent years ionospheric variations, both small-scale irregularities and large-scale variations connected with specific solar events, such as solar flares—and more recently coronal mass ejections—have received special attention (cf. SOLAR RADIATION). Flares are the most dramatic of the solar events. Here we will briefly list in chronological order the most dramatic ionospheric effects of a large solar event.

Approximately 8 min after a flare occurs, the first effect in the Earth's ionosphere—the so-called sudden ionospheric disturbance (SID)—is observed on the sunlit portion of the Earth. The existing source is enhanced x rays from the Sun with a factor of 100 and even more increase in solar radiation between 0.1 and 1 nm. As a result, the electron density in the D region is significantly enhanced. Long-distance radio-wave propagation is seriously disturbed, while the propagation conditions for frequencies with longer wavelengths may be improved. The SID events normally last less than 1 h.

Major flares are normally connected with large fluxes of relativistic protons (typically 100 MeV, i.e., comparable to the energies of cosmic rays). Within one to a few hours these energetic solar particles reach the Earth's ionosphere poleward of typically 50° magnetic latitude. The result is a complete "blackout" of all long-distance radio communications at high latitudes, which often lasts for many days. These are called polar cap absorption (PCA) events, named for their effect upon radio-wave transmissions. The physical explanation is that these high energy particles cause a new ionospheric layer in or below the regular D region. Even with an electron density of only 10^8 to 10^9 m^{-3}, the absorption effects are so large because of the high collision frequency at these low altitudes. Again, the enhanced D region will improve long-wavelength transmissions that are reflected from the D layer.

The third class of ionospheric disturbance is called auroral absorption (AA) in radio terminology, or, more generally, magnetic substorm events. A major solar event will produce both a magnetic storm and substorm events (while a magnetic storm is a global phenomenon lasting days, substorms are episodic releases of magnetospheric energy into the ionosphere primarily to the auroral oval, lasting on the order of one hour). AA events occur not only from flares, but can also depend on other solar and solar-wind irregularities (not always well defined). As both x rays and the solar xuv radiation vary in phase with the solar activity, marked changes in electron densities and irregularities are also caused by these radiations. About 30 to 60 h after this solar activity, AA effects are observed particularly in the oval within ±5–6 h around magnetic midnight. The ionospheric disturbances from magnetic storms and substorms are primarily caused by auroral electrons; i.e., electrons between 0.1 and 100 keV. The occurrence of individual substorm events is normally irregular, and there may be several events in one disturbed night. Particle bombardment introduces local changes in E- and F-region ionization. The electric fields are generally enhanced and variable, and activity expands equatorward. Since auroral particles are the principal source of nighttime E- region ionization, changes in this region are large and rapid during the substorm. Transport increases in importance for determining F-region densities. As the substorm activity passes its peak, large electric fields, called subauroral ion drift (SAID) events, develop on the equatorward edge of the auroral region in the trough. In this region, also, the subauroral red arc or SAR arc is sometimes seen as a result of heating of the upper F region by energy coupled along magnetic field lines from the inner edge of the plasma sheet in the inner magnetosphere.

Solar abnormalities may exist for several solar rotation periods, and ionospheric disturbances may peak at 27-d intervals, i.e., when the same place on the Sun faces the Earth again. Not all magnetic storms are caused by solar flares. Recent research points to coronal mass ejections, which may come from flares or may be from other active regions on the Sun, as the carriers of the enhanced plasma in the solar wind that causes magnetic storms. The magnetic storm typically lasts a few days with enhanced substorm activity during the period.

9. RADIO PROPAGATION EFFECTS

For the same reasons that radio waves were important in the discovery of the ionosphere, the ionosphere greatly influences radiowave propagation. For the longer wavelengths that cannot penetrate the ionosphere, the reflection of these waves to a first approximation obeys Snell's law. Signals can propagate within the Earth–ionosphere cavity to great distances. Amateur radio operators quickly learn that different frequency bands propagate better depending on daytime or nighttime conditions. The E layer is normally greatly reduced at night because of higher neutral densities, shortening the recombination time, and signals that would be reflected by the E layer in the daytime pass through to the F region at night. This allows communication to greater distances while bypassing nearer regions, a process often referred to as "skip."

Reflection from a smooth layer is well behaved, while reflection from an irregular region can greatly distort the signal. During magnetic storms the high-latitude regions are disturbed, which in turn disrupts or greatly changes communication characteristics. Skip becomes irregular. In the case of a PCA event (cf. Sec. 8), the energetic solar protons create extra ionization in the D region. This results in communication blackouts for high frequencies.

The ionosphere also influences the propagation of VHF and UHF radio waves to and from satellites at frequencies that normally are not reflected. Irregularities in the ionization density create both phase and amplitude distortions in radio waves as they propagate through the ionosphere. Increasing the bit error rate can make communications marginal. These distortions are referred to as scintillations. Scintillations vary widely in space and time; however, two regions of high scintillation activity deserve specific mention. At low latitudes scintillations are especially prevalent in the evening hours. After sunset the E region disappears because of recombination, leaving a sharp gradient in the plasma density. This results in a more dense fluid over a less dense fluid, which is subject to a Rayleigh-Taylor instability. Bubbles of low-density plasma form and rise through the F region creating irregularities referred to as equatorial spread F. This causes major communications problems because of the resulting scintillations. Auroral precipitation, especially in the auroral oval, is also a major source of irregularities and therefore scintillations. In addition, the polar cap can also have significant amounts of irregularities. Patches of ionization are created in the auroral-zone F region near noon and are convected across the polar cap to the nightside by the electric fields. These patches are scources of scintillations.

10. PLANETARY IONOSPHERES

Since the flyby of the planet Venus in the 1960s by the Mariner probes, a wealth of information on planetary ionospheres has been gathered from other spacecraft by *in situ* and remote sounding techniques. The outer planets have been explored by Voyagers 1 and 2. Pioneer Venus orbited Venus for nearly a decade, providing extensive information about ionospheric structure and dynamics. During the coming decade planetary ionospheres will receive more and more attention. In this section we will only list a few important ionospheric facts about the planets.

Even though several of the basic physical and chemical processes (i.e., production, loss, and transport) that control planetary ionospheres may be similar to those on Earth, major differences occur in the composition and height distribution of the neutral atmospheres as compared to the Earth. One other important factor is the difference in the magnetic fields of the planets. The magnetic moments on the other inner planets, particularly Venus, are low, while the outer planets have large magnetic fields when compared to the Earth.

The ionospheres of the inner planets have several similarities to the Earth's E layer. The most abundant ion is O_2^+ even if there is practically no neutral O_2 in their upper atmospheres. This is due to photodissociation of CO_2, which is a major gas in the planets' atmospheres. Even the maximum plasma densities are found between 100 and 200 km. In a similar way to the terrestrial ionosphere, transport processes are important for determining electron and ion densities at greater altitudes. However, because their magnetic fields are so weak, magnetic conjugate effects are generally nonexistent. The solar wind interacts directly with the ionosphere on Venus, with the plasmapause and bow shock standoff distance being determined by the balance of the plasma pressure with the solar-wind pressure.

Both the densities and, especially, the vertical height distributions of the ionospheres are significantly different at the major outer planets. Taking Jupiter as a typical example, the major ion is H^+ with its peak density above 500 km. Also H_3^+ (H_2^+ reacts with H_2 to form H_3^+) is of special importance in Jupiter's ionosphere. The upper atmosphere of Jupiter consists mainly of molecular hydrogen, while H, He, and CH_4 are also important constituents. Again, both photoionization and solar-wind–ionospheric interactions are principal ionization sources. Because of Jupiter's large magnetic field, a significant ionospheric difference is expected from mid- and low-latitude to polar regions where energetic particles also become important. Particularly for the outer planets, information about background atmospheres is insufficient, and their chemical reactions are poorly understood. Furthermore, there is no direct information on electron and ion temperatures in the ionosphere.

11. QUESTIONS FOR THE FUTURE

Although the fundamental processes related to the large-scale formation and nature of the ionosphere are fairly well understood, many problems concerning structure and dynamics, energy balance, physical and chemical couplings, etc., still remain. Here we will only briefly mention some items that will have a high priority during the coming decade.

Coupling and transport of ionization between the auroral zones and regions further poleward and equatorward are not well mapped and explained. Both horizontal and particularly vertical transport (the so-called polar wind) from high latitudes need further experimental and theoretical attention. New incoherent-scatter radars for the central polar regions, where even the electron density profiles are poorly known, closely coordinated with simultaneous rocket and satellite measurements, should have priority.

It may surprise readers that the ionospheric layer closest to the Earth—i.e., the D region between 50 and 90 km—is least known and understood. Here the atmospheric pressure is large and collisions totally dominate so that the usual ground techniques for studying the ionosphere are not applicable. Because collision frequencies exceed the critical frequencies for the region, absorption dominates.

Progress in the next decade will be toward exploration of the lower ionosphere and understanding the complexities of this intricate coupling between regions. Global scale models like the Utah State University ionospheric model (Schunk, 1988; Schunk and Szuszczewicz, 1988) along with *in situ* satellite measurements and ground-based radar sounding of the ionosphere should greatly increase our understanding of processes in the upper ionosphere. However, data at lower altitudes are limited in scope and coverage by the inaccessibility of the D and E regions to satellite measurements. Sounding rockets have provided limited coverage, but cannot provide a large-scale picture of the dynamics of these regions necessary to address the complicated feedback mechanisms associated with geomagnetic storms and substorms. Models can simulate most instantaneous conditions, but have too many unconstrained variables to specify the time dependence accurately. NASA is planning a mission in the latter part of the decade called TIMED (thermosphere–ionosphere–mesosphere energetics and dynamics) to address some of these problems. TIMED will perform an exploratory study of the physical and chemical processes acting within and upon the coupled mesospheric, thermospheric, and ionospheric system between 50 and 180 km. The energetics of these regions, the global circulation and dynamics, the chemistry and composition, and the ionospheric structure and electrodynamics will be primary objectives of this mission. It will also begin to address possible anthropogenic changes in this region, questions raised because of global change that appears to be taking place in the atmosphere below.

GLOSSARY

Aurora: Optical emissions in the upper atmosphere (> 90 km) due to excitation of molecules and atoms by precipitating energetic particles from the magnetosphere and solar wind.

Auroral Oval: An oval-shaped belt around the geomagnetic pole in each hemisphere between about 60° and 80° geomagnetic latitude that is the location of the majority of the auroral emissions. The oval expands and contracts with changes in magnetic activity.

Conjugate Auroras/Disturbances: Simultaneous phenomena in both hemispheres at locations connected by the same geomagnetic field lines.

Convection: The global scale motion of plasma in the ionosphere driven by electric fields. The motion is generally sunward in the auroral zone and antisunward over the polar cap following lines of equipotentials in the electric-field pattern.

Ionosonde: A transmit and receive system for determining the structure of the ionosphere by measuring the reflection height of radio waves at various frequencies.

Ionosphere: A region of partially ionized gas between 50 and 1000 km where free electrons and ions coexist in an electrically neutral plasma. Subdivided into **D region:** ionospheric region between 50 and 90 km above the Earth; **E region:** ionospheric region between 90 and 140 km; **F region:** ionospheric region above 140 km; **topside ionosphere:** ionospheric region above the peak of the F region.

Magnetic Local Time: Local time defined in a magnetic latitude and longitude system centered on the Earth's geomagnetic axis.

Magnetic Storm: Large-scale disturbance of the magnetosphere and ionosphere caused by the arrival of an enhanced region of solar-wind density and velocity caused by a solar disturbance. The magnetic storm typically lasts 1–3 d from onset to recovery.

Magnetic Substorm: An episodic release of energy into the auroral region from the magnetosphere causing enhanced auroral activity and magnetic disturbances in the polar ionosphere lasting approximately 1 h.

Magnetosphere: The region between the ionosphere and the magnetopause, a boundary determined by the balance of the plasma pressure of the solar wind and the magnetic pressure of the Earth's magnetic field. The Earth's dipole magnetic field is pushed in on the dayside and stretched out into a tail on the nightside. The typical standoff distance of the magnetopause on the dayside is 10 Earth radii.

Plasma: An electrically neutral, partially or fully ionized gas containing free electrons and ions.

Polar Cap: The region poleward of the auroral oval.

Polar Ionosphere: The ionosphere poleward of about 60° magnetic latitude (encompassing the auroral oval and polar cap).

Polar Wind: The outflow of ionized particles from the polar ionosphere.

Solar Wind: Low-density supersonic plasma flowing out from the Sun which interacts with the Earth's magnetosphere.

Trough: A region of low electron density equatorward of the nighttime auroral oval.

Universal Time (UT): Time referenced to 0° longitude or the Greenwich meridian.

List of Works Cited

Birkeland, Kr. (1908), *Norwegian Aurora Polaris Expedition, 1902–3,* Christiana, Norway: Aschehoug, Vol. 1, pp. 1–315.

Brekke, A., Egeland, A. (1983), *The Northern Light,* Heidelberg: Springer.

Chapman, S., Bartels, J. (1940), *Geomagnetism,* Oxford: Clarendon.

Eather, R. H. (1980), *Majestic Lights,* Washington, DC: American Geophysical Union.

Egeland, A., Holter, Ø., Omholt, A. (1973), *Cosmical Geophysics,* Oslo, Norway: Universitetsforlaget, p. 30.

Hedin, A. E. (1991), "Extension of the MSIS Thermospheric Model into the Middle and Lower Atmosphere," *J. Geophys. Res.* **96**, 1159–1172.

Heppner, J. P., Maynard, N. C. (1987), "Empirical High Latitude Electric Field Models," *J. Geophys. Res.* **92**, 4467–4489.

Iijima, T., Potemra, T. A. (1978), "Large Scale Characteristics of Field Aligned Currents Associated with Substorms," *J. Geophys. Res.* **83**, 599–615.

Johnson, F. S. (Ed.) (1961), *Satellite Environment Handbook,* Stanford, CA: Stanford University Press.

Kelley, M. C. (1989), *The Earth's Ionosphere,* San Diego: Academic, p. 40.

Potemra, T. (1990), *Satellite Measurements of Birkeland Currents,* Birkeland Lecture No. 3, Oslo: The Norwegian Academy of Science and Letters.

Ratcliffe, J. A. (1972), "The Formation of the Ionosphere," *Geophys. Norveg.,* **29**, 13–27.

Rees, M. H. (1989), *Physics and Chemistry of the Upper Atmosphere,* Cambridge, U.K.: Cambridge University Press.

Schunk, R. W. (1988), "A Mathematical Model of the Middle and High Latitude Ionosphere," *Pure Appl. Geophys.* **127**, 255–303.

Schunk, R. W., Szuszczewicz, E. P. (1988), "First Principle and Empirical Modelling of the Global Scale Ionosphere," *Ann. Geophysicae* **6**, 19–30.

Stewart, B. (1882), "Terrestrial Magnetism," *Encyclopedia Brittanica,* 9th ed., Edinburgh: Adam Black.

Störmer, C. (1955), *The Polar Aurora,* Oxford: Clarendon.

U.S. Standard Atmosphere (1976), National Oceanographic and Atmospheric Administration, National Aeronautics and Space Administration, and U.S. Air Force, Washington, DC.

Further Reading

Akasofu, S.-I., Chapman, S. (1972), *Solar-Terrestrial Physics,* Oxford: Clarendon.

Banks, P. M., Kokarts, G. (1973), *Aeronomy,* New York: Academic.

Brekke, A., Egeland, A. (1983), *The Northern Light,* Heidelberg: Springer.

Chapman, S., Bartels, J. (1940), *Geomagnetism,* Oxford: Clarendon.

Eather, R. H. (1980), *Majestic Lights,* Washington, DC: American Geophysical Union.

Jursa, A. S. (Ed.) (1985), *Handbook of Geophysics and Space Environment,* Air Force Geophysics Laboratory, Air Force Systems Command, Springfield, VA: National Technical Information Service.

Heppner, J. P., Maynard, N. C. (1987), "Empirical High Latitude Electric Field Models," *J. Geophys. Res.* **92**, 4467–4489.

Kelley, M. C. (1989), *The Earth's Ionosphere,* San Diego: Academic.

Ratcliffe, J. A. (1972), *An Introduction to the Ionosphere and Magnetosphere,* Cambridge, U.K.: Cambridge University Press.

Rees, M. H. (1989), *Physics and Chemistry of the Upper Atmosphere,* Cambridge, U.K.: Cambridge University Press.

Rishbeth, H., Garriott, D. K. (1969), *Introduction to Ionospheric Physics,* Amsterdam: Academic Publishers Assn.

Vallance Jones, A. (1974), *Aurora,* Dordrecht, Holland: Reidel.

IRON AND STEEL

See STEEL

LABORATORY INSTRUMENTATION, CALIBRATION AND MAINTENANCE

See MAINTENANCE AND CALIBRATION OF LABORATORY INSTRUMENTATION

LANs

See NETWORKS, COMPUTER

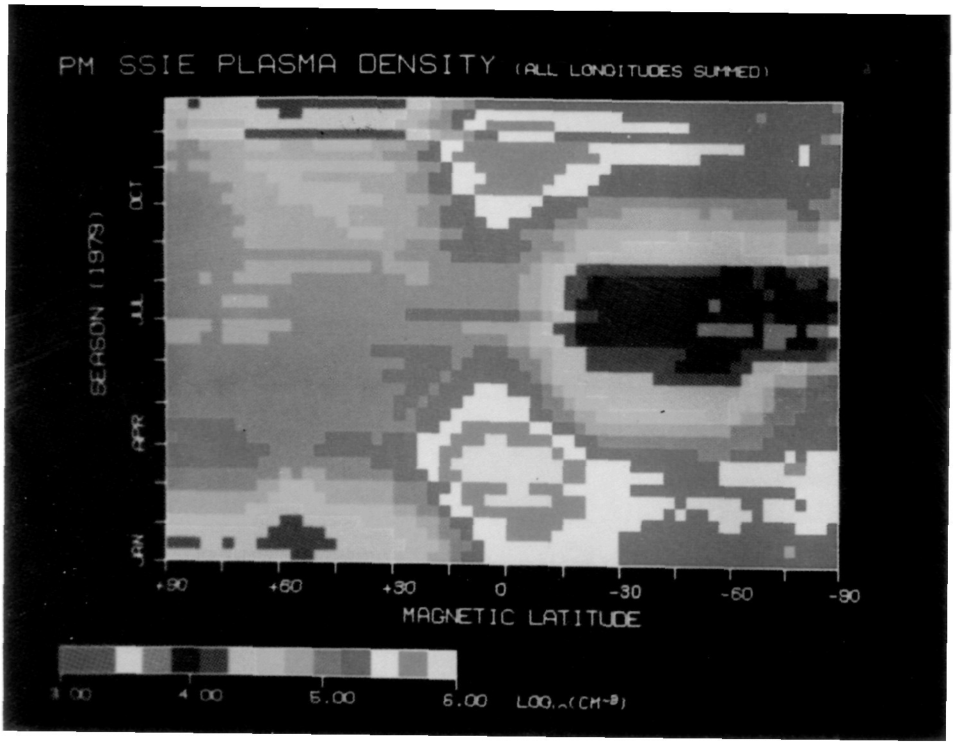

Vol. 8 – PLATE 1. An illustration of the global distribution of ionization and its seasonal variation at 830 km. This is an averaged distribution over all p.m. local times from the DMSP F2 and F4 satellites over a 2-y period (courtesy of F. J. Rich, Geographics Directorate, Phillips Laboratory, USAF).

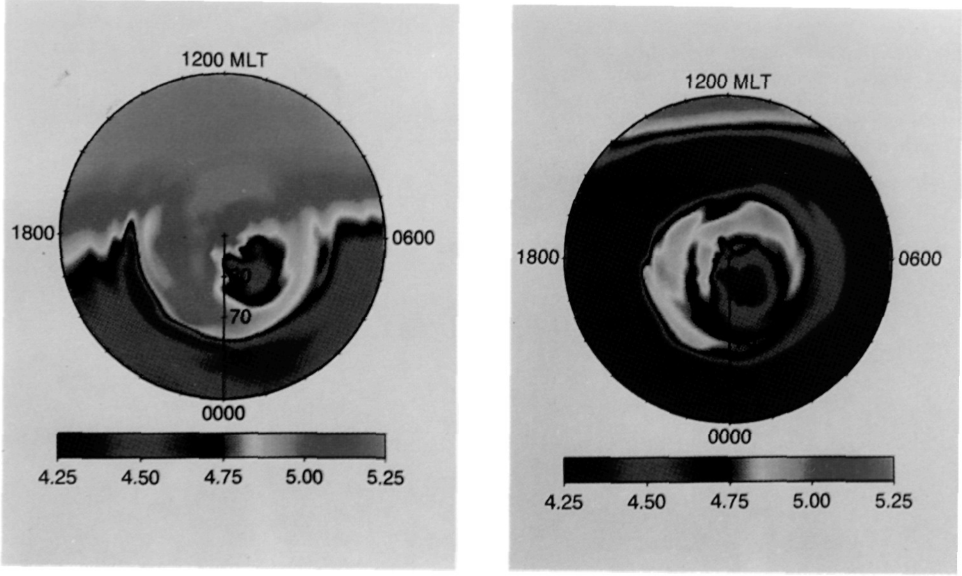

Vol. 8 – PLATE 2. Polar views of a typical ionization distribution for southern-hemisphere winter for low solar activity and quiet geomagnetic conditions. (a) For 0500 UT when the geomagnetic pole is tipped toward the Sun; (b) for 12 h later when the geomagnetic pole is tipped away from the Sun. These distributions are calculated for the altitude of the F-region peak using the Utah State University ionospheric model (courtesy of R. W. Schunk and J. J. Sojka, Utah State University).

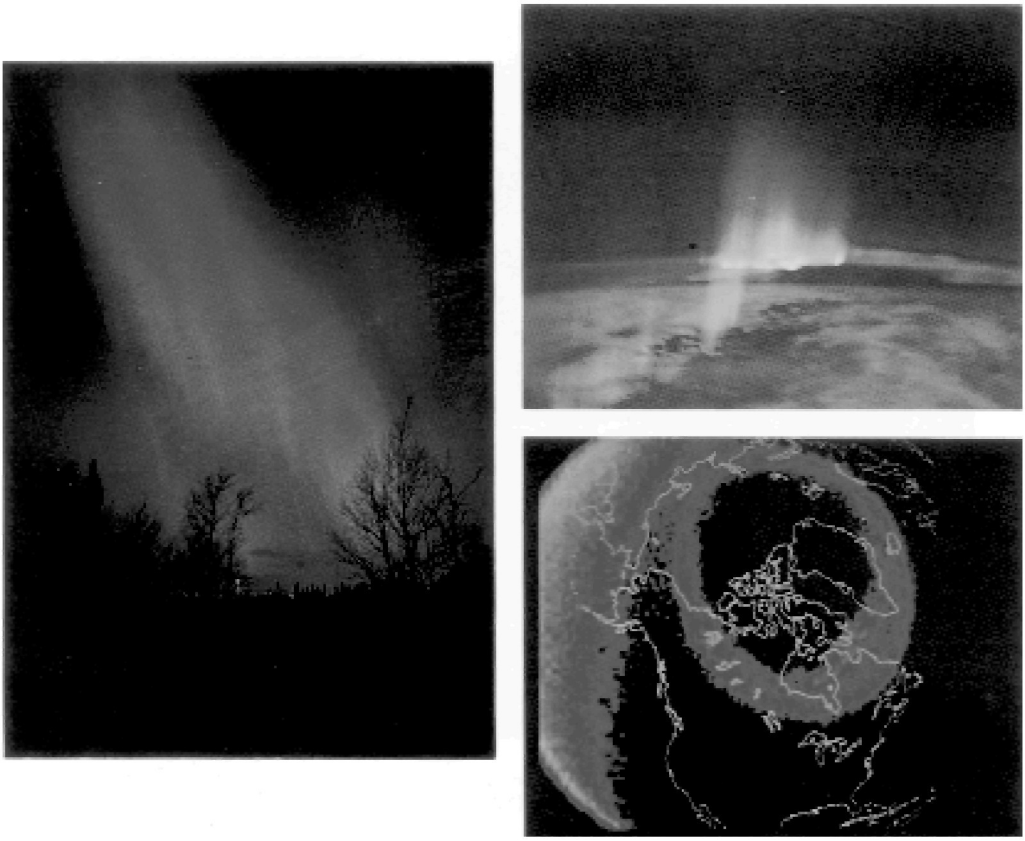

Vol. 8 – PLATE 3. Auroral pictures–(a) Rayed aurora band photographed in Northern Scandinavia. (b) A view of the aurora australis photographed from the flight deck of NASA Space Shuttle *Discovery* by one of its seven crew members on Mission STS-39. Note the vertical extent of the rays and the red color of the higher altitude portions. (Photo courtesy of the National Aeronautics and Space Administration). (c) The aurora as seen from deep space by the DE-1 spacecraft (from nearly 4 Earth radii) showing the sunlit dayside of the Earth and the auroal oval over the United States and Scandinavia (courtesy of L. A. Frank University of Iowa).

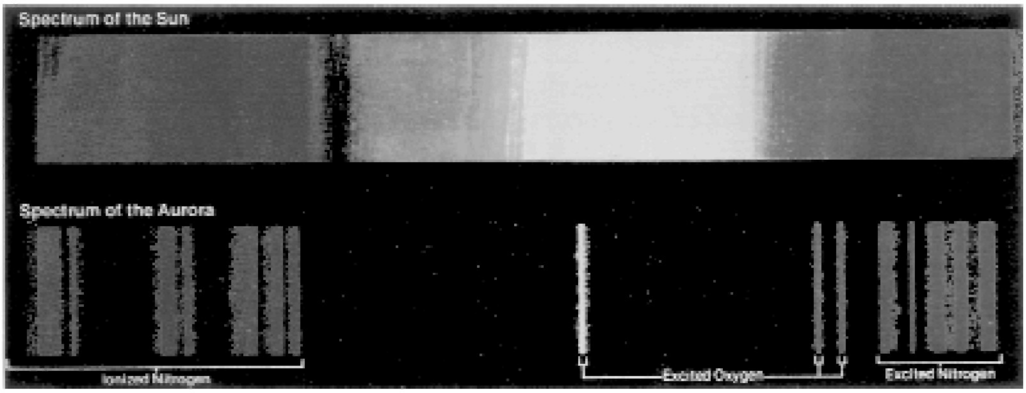

Vol. 8 – PLATE 4. The auroral spectrum, which consists of distinct lines and bands. For comparison the spectrum of the Sun shown.

LABORATORY INSTRUMENTATION

AYMAN A. SHOURBAJI, *Instrumentation and Controls Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A.*

	Introduction	247
1.	**Test and Measurement Instrumentation**	248
1.1	Digital Voltmeters	248
1.2	Oscilloscopes	250
1.2.1	Analog Oscilloscopes	252
1.2.2	Analog Storage Oscilloscopes	252
1.2.3	Digital Oscilloscopes	253
1.3	Function Generators	253
1.4	Recorders	256
1.4.1	Analog Recorders	256
1.4.2	Digital Recorders	258
1.4.2.1	Digital Wave Recorders	258
1.4.2.2	Programmable Recorders	258
1.4.3	*X-Y* Recorders	258
1.5	Spectrum Analysis	259
1.5.1	Fourier Analyzers	260
1.5.2	Spectrum Analyzers	261
1.6	Counters	262
1.7	Boxcar Integrators	265
1.8	Lock-In Amplifiers	266
1.9	Time-Domain Reflectometry	267
2.	**Data-Acquisition Instrumentation**	268
2.1	Data Loggers	268
2.2	PC-Based Data-Acquisition Systems	270
2.2.1	Software	271
2.3	Instrument Interfaces	272
3.	**Measurement Interface**	275
3.1	Digital-to-Analog Converters	275
3.2	Analog-to-Digital Converters	276
3.3	Signal Transmission	278
4.	**Conclusion**	279
	Glossary	279
	Works Cited	279
	Further Reading	279

INTRODUCTION

The main mission of every laboratory is to conduct scientific investigations to develop an aspect of a new technology or product or to improve an existing technology or product. The success of the experiments, which are an important part of the scientific investigative efforts, depends to a large extent on the availability of the appropriate test and measurement tools to quantify physical phenomena in complex measurements. Consequently, instrumentation users are faced with the ultimate responsibility of selecting the appropriate instrument for the scientific task in view of the diversified and rapidly changing technologies associated with the instrumentation field. The instrumentation industry, particularly, has undergone major changes that are directly related to the technlogical advances in electronics that evolved over the past five decades.

One major change was marked by the invention of the transistor in the 1940s. That was the beginning of semiconductor electronics. By replacing the vacuum tube, the transistor (solid state device) sent the instrument designer back to the drawing board, so to speak, and a new generation of instruments emerged. The newly developed instruments afforded better accuracy and improved stability, were more efficient and more compact, and offered faster response than their predecessors that were based on vacuum-tube technology. In the 1960s and the 1970s, electronics research developed solid-state devices capable of performing complex functions that included many discrete components such as chips, transistors, resistors, diodes, and capacitors integrated in a single silicon substrate. Today, these devices are known as integrated circuits (ICs). Further research and development in the IC field led to the microprocessor, which can be considered as a second turning point—if we consider the transistor as being the first—in the electronics and the instrumentation industries. Most modern test and measuring devices use one to three microproces-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

sors to add "intelligence" (i.e., capabilities that allow these devices to perform complex functions that previously required several instruments and off-line calculations). In addition, with microprocessors, test and measurement instruments combined the best of two worlds—the digital and the analog.

This article, in its scope and content, is intended for readers in a wide range of technical disciplines. The topics covered and their treatment, therefore, were chosen with that in mind. The article introduces a class of instrumentation that is most commonly used in the laboratory for conducting a variety of tests and performing a wide range of measurements as well as measurement systems for large-scale data monitoring and processing. The discussion includes principles of operation, and the underlying technologies associated with the latest development and functionality of these instruments. Also discussed are issues that confront instrument users when interfacing analog and digital devices for measurement applications necessitated by modern technologies.

It should be noted that all standards, communication protocols, and conventions cited in this article are based on those adopted in the United States—which may or may not be the same as those adopted by other countries.

1. TEST AND MEASUREMENT INSTRUMENTATION

1.1 Digital Voltmeters

The digital voltmeter (DVM), also known as the digital multimeter (DMM), is one of the most versatile instruments used in the laboratory. It is a multifunction instrument capable of measuring ac and dc electrical quantities such as volt, current, and resistance. The analog signal to be measured is introduced to the instrument, and it produces nearly instantaneously a direct visual digital display of the measured quantity. This type of instrument is available in analog form, but DVMs are now more commonly used than the analog counterpart. The DVM displays the readings on a liquid crystal display (LCD) or on a light-emitting diode (LED) device, which is an integral part of the meter. (See DISPLAY DEVICES for more details on digital display.)

In dc measurement, the dc quantity is an average value taken over some time interval *T* to smooth any fluctuations. The average value is computed by a simple averaging circuit in one of two ways. The first approach is to apply the signal to a true integrating circuit (true averaging) by using an operational amplifier with a capacitor in its feedback path (Fig. 1). In the second approach, the average value is obtained by using a low-pass filter circuit,

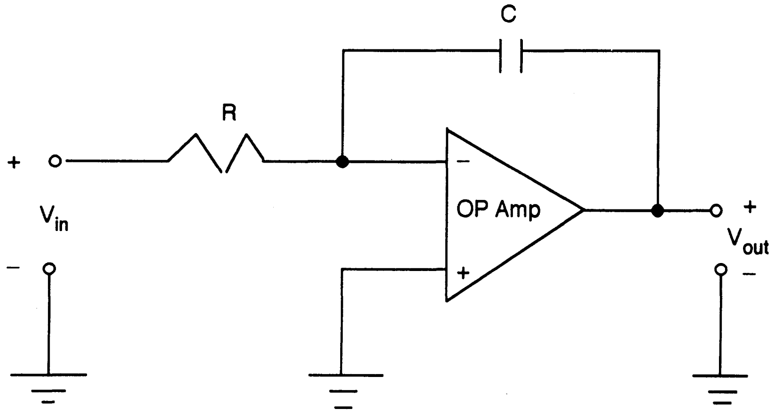

$$V_{out} = -\frac{1}{RC}\int V_{in}dt = -\frac{t}{RC}V_{in}$$

time constant (τ) in seconds = RC ; R in ohms, C in farads

FIG. 1. Active integrator using operational amplifier (op-amp). It is used in some voltmeters as true averaging circuits in root-mean-square measurements.

which is simply a resistance–capacitor (*RC*) network (Fig. 2). The *RC* network continuously smooths the signal fluctuations to produce, after approximately three to five time constants, an average value of the signal (*RC* averaging). Some other digital voltmeters consist simply of an analog-to-digital converter (ADC) (discussed in Sec. 3.1) and an integrated display device such as a LED or LCD. Some DVM designs employ other techniques to compute the average value for a dc signal, but they will not be discussed here; the reader can consult ELECTRICAL INSTRUMENTATION or Helfrick and Cooper (1990).

Ordinarily, ac measurement is given in terms of a root mean square (rms) value, also known as the effective value. Mathematically, the rms value of a voltage signal $v(t)$ is given by

$$V_{\mathrm{rms}}=\left(\frac{1}{T}\int_0^T v^2(t)dt\right)^{1/2}. \qquad (1)$$

If $v(t)$ in Eq. (1) is a sine wave, that is, given by

$$v(t)=V_m\sin\omega t, \quad \omega=2\pi f, \qquad (2)$$

where V_m is the sine-wave amplitude, ω is the radian frequency in radians per second, and f is the frequency in cycles per second (cps) or hertz (Hz), then the rms value (V_{rms}) can be computed from Eq. (1) to be $V_m/\sqrt{2}$.

DVMs, which provide rms measurements based on the concept expressed by Eq. (1), are known as ***true mean square*** (TMS) instruments. The underlying TMS concept can be simply outlined by

1. squaring the instantaneous value of the signal,
2. averaging the squared instantaneous value over the sampling time, and
3. taking the square root of the measured mean square value.

One type of TMS DVM squares the voltage of the applied signal by using a detector circuit with a square-law characteristic. The squared voltage signal is time averaged by using either a true averaging circuit or an *RC* averaging circuit. The square-root function can be accomplished by a proper readout scale calibration. Other types of TMS-type DVMs use detectors that compute rms value directly. When highly accurate rms measurements of nonsinusoidal signals are required (e.g., random signals), methods based on the heating power of the wave forms are used since heating power is directly proportional to the mean squared voltage.

Some voltmeters are calibrated to read the rms value of a sinusoidal signal, but they do not actually measure the rms value of the applied signal. These instruments usually consist of a simple rectifier followed by an equivalent low-pass smoothing filter such that a quantity proportional to either the mean absolute value (average rectified value) or peak value of the applied signal is measured. The mean absolute value for a sinusoidal voltage signal $v(t)$ is defined as

$$\overline{|v(t)|}=\frac{1}{T}\int_0^T |v(t)|dt, \qquad (3)$$

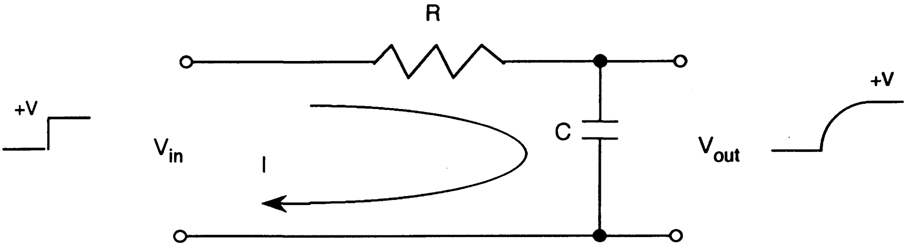

$$V_{out}(t) = V_{in}\,(1 - e^{-t/RC})$$

time constant (τ) in seconds = RC; R in ohms, C in farads

FIG. 2. Low-pass resistance–capacitor (RC) filter used in some voltmeters as RC averaging circuits in root-mean-square measurements.

where $v(t)$ is the absolute value and $\overline{|v(t)|}$ is the mean of the absolute value. For either a peak or mean absolute-value measurement, a readout in terms of the rms value of a sine wave can be obtained by a proper readout scale calibration. This calibration is easy to accomplish since the peak, mean absolute, and rms values of a sine wave have fixed relationships. For the sine wave in Eq. (2), the peak value is V_m, the mean absolute value is $\overline{|v(t)|}=(2/\pi)V_m$ and the rms value is $V_m/\sqrt{2}$. Some instrument designs take advantage of these relationships because peak- or mean absolute-value detection circuits are rather cheaper to build than true rms detection circuits. These simple instruments will produce inaccurate readings if they are used to measure complex periodic signals that are not sinusoidal.

DVMs are available in a wide range of capabilities to measure dc and ac voltages from <1 mV to >1000 V, currents from less than a milliampere to several amperes, and dc resistances from less than an ohm to up to several gigohms. Inaccuracy of the meters is generally determined by the degree of stability of the reference voltage used in the circuit and the inherent error of ±1 unit in the least significant digit (LSD). Better accuracy can be realized with an increased number of display digits because of the higher resolution that can be attained. Digital voltmeters are available with up to $8\frac{1}{2}$ digits for bench-type versions and up to $4\frac{1}{2}$ digits for the portable-type (hand-held) versions. A meter with $4\frac{1}{2}$ digits can display four full digits ranging from 0 to 9, and one digit (leftmost) that displays only a 1 or 0. Therefore, this meter can display up to 19 999 counts of resolution, whereas a meter with $5\frac{1}{2}$ digits can display 199 999 counts of resolution, as shown in Table 1. Most recently, frequency measurement has been added to the DVM's measurement capability. Hand-held DVMs with built-in frequency counters that can measure from 1 Hz to >20 kHz are now commercially available.

1.2 Oscilloscopes

The oscilloscope, also commonly known as a scope, has been and is still considered to be one of the most versatile general-purpose electronic measuring instruments available for industrial applications as well as scientific experimentation. The scope is capable of capturing, displaying, and analyzing analog signals of various wave forms. The applied signal is displayed on a phosphorescent screen as an illuminated two-dimensional x-y trace. The signal magnitude is always displayed on the vertical y axis (vertical deflection), while the x axis (horizontal deflection) is used either as the time base or to display a second input signal to produce an X-Y plot of the two signals.

The most common use of an oscilloscope is to produce a plot of the input signal against time. The input signal is sensed and applied to the scope through a probe designed for that purpose. Most laboratory oscilloscopes have a 12.7-cm-diam (5-in.-diam) screen, but larger screens up to 53.3 cm (21 in.) are also available. Large-screen oscilloscopes usually cannot attain the high-frequency capability of the 12.7-cm (5-in.) types.

The modern oscilloscope, regardless of its sophistication, consists basically of the same building blocks, or components used in earlier models, as illustrated in Fig. 3. The cathode-ray tube (CRT) is the main output device that displays a trace of the signal wave form. The signal trace is produced by a focused narrow beam of electrons projected from an electron gun through a set of vertical and horizontal deflection plates. Application of voltages to these plates creates an electric field that deflects the electron beam, causing vertical and horizontal displacement of its focused beam

Table 1. Relationship between the number of display digits and the resolution.

Number of display digits	Number of counts of full scale	Resolution (%)
$3\frac{1}{2}$	1999	1 part in 1999=0.05
$4\frac{1}{2}$	19 999	1 part in 19 999=0.005
$5\frac{1}{2}$	199 999	1 part in 199 999=0.0005
$6\frac{1}{2}$	1 999 999	1 part in 1 999 999=0.000 05

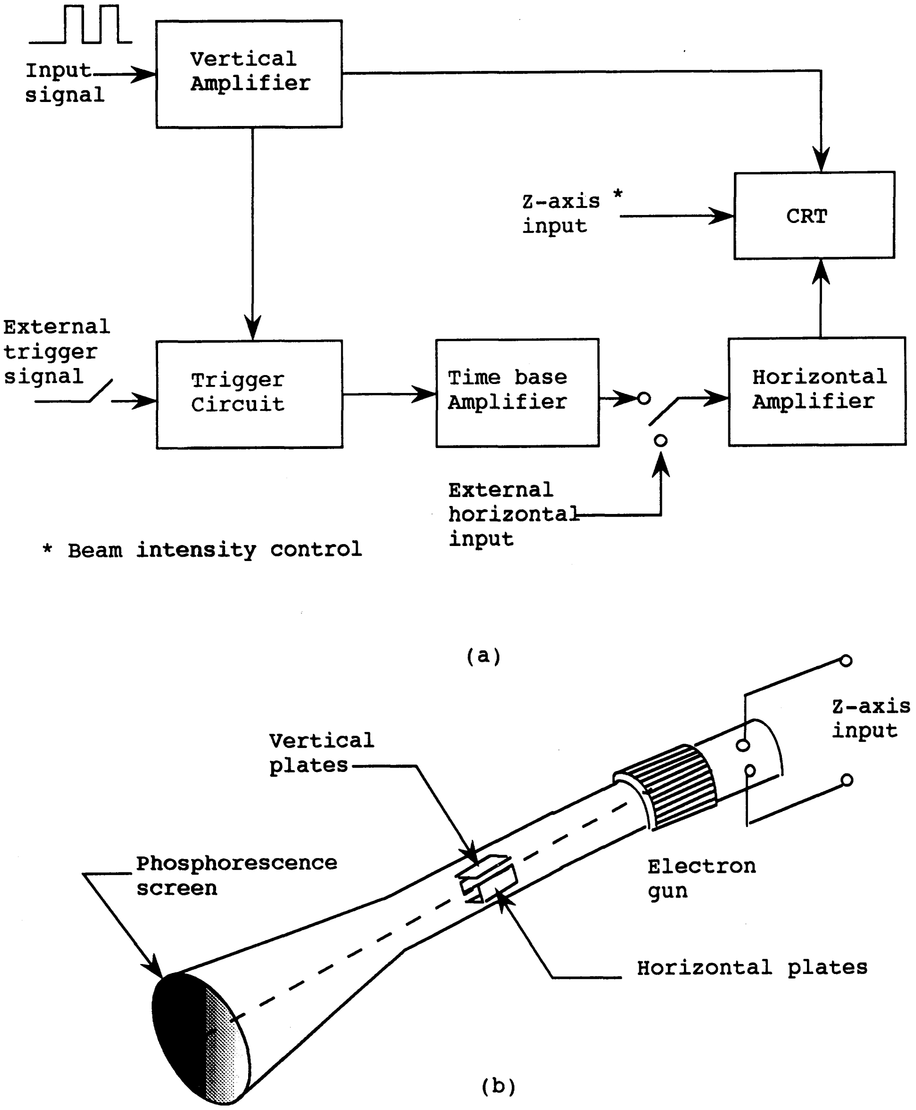
Input signal
Vertical Amplifier
Z-axis input *
CRT
External trigger signal
Trigger Circuit
Time base Amplifier
Horizontal Amplifier
External horizontal input
* Beam intensity control
(a)
Vertical plates
Phosphorescence screen
Z-axis input
Electron gun
Horizontal plates
(b)

FIG. 3. Basic components of oscilloscopes: (a) block diagram; (b) cathode-ray tube basic structure.

point on the phosphorescent screen. With the proper design, the displacement can be made closely linear with the applied voltages. In standard CRTs, the phosphor is electrically refreshed before the phosphorescence decays to avoid blinking or flickering of the displayed trace. Some CRTs, storage CRTs, do not require continual refreshing to maintain the displayed trace. This type of CRT is discussed in Sec. 1.2.2.

The horizontal amplifier drives the horizontal plates with a ramp voltage, causing the focused spot to sweep from left to right at a constant speed. The trigger or synchronizing circuit provides the mechanism that defines the starting point of every successive sweep of the displayed trace. For proper synchronization of the sweep and the input signal applied to the vertical deflection plates, triggering of the sweep is initiated by energizing the trigger circuit from the leading edge of the input signal itself. External sources also can be used for sweep triggering.

Laboratory oscilloscopes may be classified in many ways. Usually, the distinctions are based either on frequency-response capability or on CRT characteristics.

1.2.1 Analog Oscilloscopes. An analog oscilloscope usually deploys an electrostatically deflected CRT with a standard refresh phosphor and real-time display circuitry. With this circuitry, the CRT beam deflection traces exactly the signal across the screen as the event occurs. For repetitive signals, the beam traces the signal pattern over and over, resulting in a bright wave form displayed on the screen.

Analog oscilloscopes can be classified into low- and high-frequency categories. The low-frequency oscilloscope typically has video-frequency capability ranges from dc to about 10 MHz (10-MHz bandwidth), and the high-frequency oscilloscope is capable of displaying single-shot signals with rise times as fast as 0.7 ns ($\sim$500 MHz).

Two of the most important considerations when evaluating a scope are the bandwidth and the rise time. The bandwidth is defined as the frequency range between dc and the maximum frequency at which the scope could display a signal with an attenuation of no more than 3 dB, and it determines the signal capturing performance that could be expected from the scope. The rise time is defined as the time it takes the signal to change from 10% to 90% of its final value, as shown in Fig. 4. The rise time determines the ability of the scope to deal with short events such as fast transients with short durations. The bandwidth and the rise time t_r are related by the approximation

$$\text{bandwidth} = 0.35/t_r. \tag{4}$$

It should also be realized that the rise time of a signal as observed on the screen is not just the rise time of the signal but also includes the rise time contributed by the scope and according to the approximate relationship

$$t_r(\text{observed}) = [t_r^2(\text{signal}) + t_r^2(\text{oscilloscope})]^{1/2}. \tag{5}$$

1.2.2 Analog Storage Oscilloscopes. Analog storage oscilloscopes can display signals of very low frequencies or low repetition rates, or even single fast events. They can temporarily store (for several minutes) wave forms in their display CRTs for extended visual analysis. The storage CRT, unlike the standard type, does not require continual refreshing to maintain a displayed trace. The storage capabilities of this type of CRT are attributed to the type of phosphor used with the screen. Phosphor types with different characteristics are classified according to their storage capabilities as short, medium, or long persistence, as indicated by Table 2.

Storage oscilloscopes are useful in applications where display measurements are difficult to observe on a standard oscilloscope CRT. Single-shot events, such as the wave

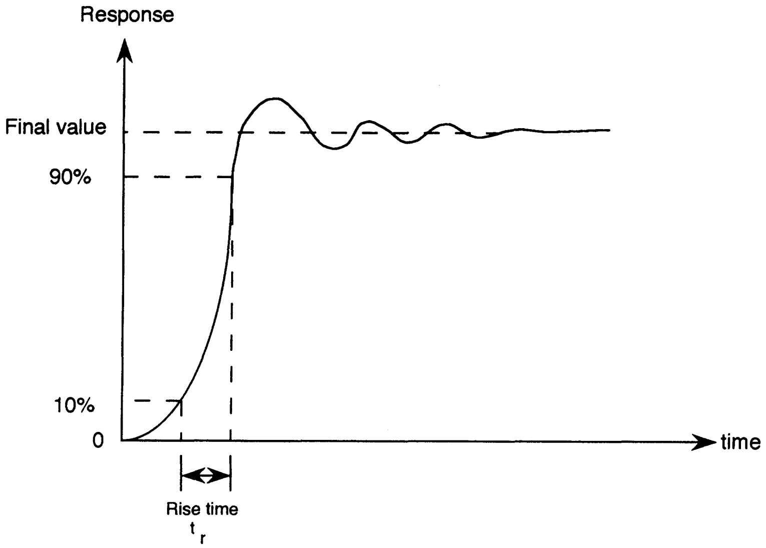

FIG. 4. Example of a typical oscillatory response of a transducer showing the rise time.

Table 2. Some of the commonly used phosphors.

Phosphor type	Color		Persistence	Display application
	Fluorescence[a]	Phosphorescence[a]		
P1	Yellow-green	Yellow-green	Medium (<2 s)	Scopes, radar
P2	Blue-green	Green	Medium	Scopes
P7	Blue-white	Yellow-green	Variable[b]	Radar, medical
P11	Blue-violet	Blue	Medium to short (<1 s)	Photographic recording
P19	Orange	Orange	Long (few min)	Radar

[a] Fluorescence is the light emitted as a beam strikes the phosphor, whereas phosphorescence is the light after an excitation beam is removed.

[b] Persistence can vary from short to long depending on filter color.

form resulting from some destructive testing or fast-occurring phenomena, are transient in nature and are quickly lost unless the wave form can be either photographed or stored on the CRT itself. Signals with very slow repetition rates are also difficult to view on standard oscilloscopes because of the phosphorescence decay. Comparison of two wave forms is often difficult when a standard oscilloscope is used since CRT refresh characteristics do require simultaneous testing of the two circuits. In storage oscilloscopes, one signal trace is stored for comparison with a second signal obtained later. Analog storage oscilloscopes have their own limitations because of the high cost of storage CRTs and the limited storage time depending on the type of phosphor used.

1.2.3 Digital Oscilloscopes. The digital oscilloscope is capable of measuring and displaying a repetitive wave form by sampling the input signal with an ADC at regular intervals and internally storing the values in solid-state memory. Nonvolatile memory can provide data storage without degradation or loss of data for up to five years with the power removed from the instrument. The stored data can be retrieved later to reconstruct the signal by a digital-to-analog converter (DAC) (discussed in Sec. 3.2) long after the signal is removed from the input. The reconstructed signal can be displayed on the screen as a series of dots, representing signal values, located closely enough to appear as a continuous trace. The number of these values on the screen determines the vertical resolution as dictated by the input ADC. A 10-bit ADC provides a full scale of 1024 discrete values that, when used to represent a single-shot signal, results in a resolution comparable to that of an analog storage oscilloscope.

The sampling rate for sampling a real-time signal needs to be carefully considered to reconstruct the signal accurately. As shown in Fig. 5, sampling at a rate less than the signal's own frequency gives false indication of a much slower signal. This phenomenon is known in sampling theory as *aliasing*. Theoretically, the sampling frequency f_s, also known as the Nyquist frequency, should be at least twice that of the input signal's highest-frequency component to ensure accurate results.

The number of ADCs used in the sampling process is an important consideration. When one ADC is shared by two or more input channels and one channel is in use, it can sample the input signal at full speed. When two channels are in use, the effective sample rate for each channel is reduced by 50%. The result is that a single-shot frequency performance can be reduced in half. Some manufacturers recently introduced a digital storage oscilloscope and DVM integrated into one compact hand-held meter. The scope is dual trace with 50-MHz bandwidth and sampling rate of 25 megasamples per second. The DVM can measure the true rms of ac or ac-plus-dc signals with 8-bit resolution.

1.3 Function Generators

The function generator instrument is commonly used in testing, calibration, and signal simulation applications. The generator can deliver a variety of wave forms whose frequencies are adjustable over a wide range with high accuracy. The only input to the instrument is the ac power required for the electronic circuitry. The output can be selected from a set of commonly used standard wave-

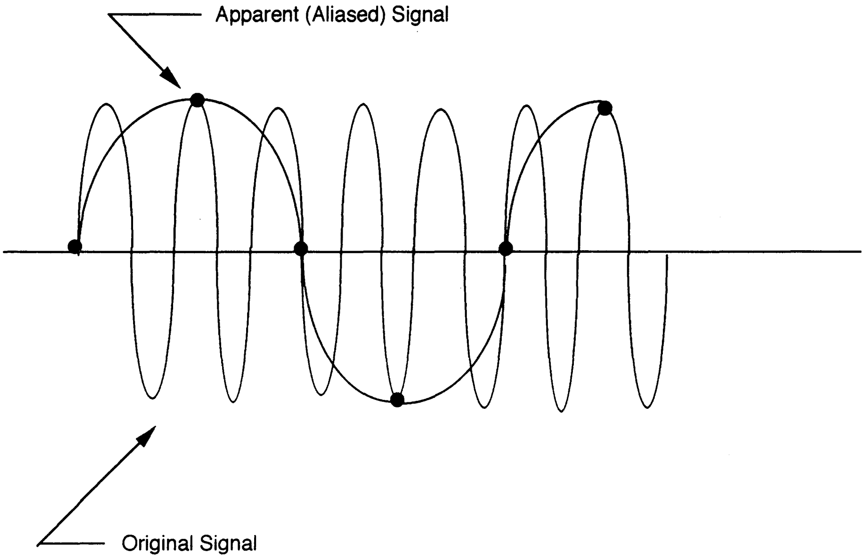

FIG. 5. Aliasing. A 16-kHz sine wave sampled at 20 kHz appears to be a 4-kHz sine wave.

forms such as the sine wave, haversine, triangular, square, and sawtooth (Fig. 6). Modern function generators can offer waveforms with frequencies adjustable from 1 μHz to 80 MHz. In addition to standard wave forms, arbitrary or user-defined waveforms (for special applications) can be created on a personal computer (PC) by using a specific software developed for this purpose and then downloaded to the generator by a standard instrument interface (IEEE-488/GPIB interface discussed in Sec. 2.3). Arbitrary signals such as video test signals and mechanical vibration can be used in simulating responses of some physical systems. When equipped with the IEEE-488 interface, the function generator can also provide remote control of various functions and allows the generator to be integrated with automatic testing systems.

Some function generators have a very useful feature for phase locking the generator with external signal sources. With phase lock, one function generator can be phase locked to a harmonic of the sine wave of a second function generator. By adjusting the phase and amplitude of the harmonics, many other wave forms can be generated as the result of summing the fundamental frequency from one generator with the harmonics of the other. The function generator can also be phase locked to a frequency standard. With such arrangements, the generated outputs have the standard source frequency accuracy and stability.

It is very difficult to give a unified description of the internal structure of all available generators since the structure varies from one

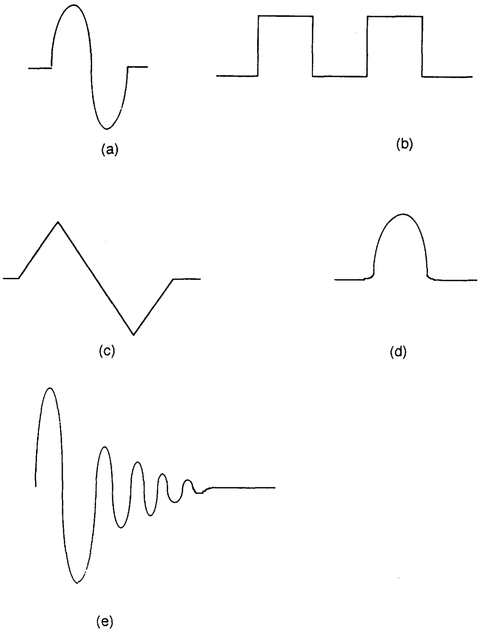

FIG. 6. Typical output wave forms produced by function generators: (a) sine wave; (b) square wave; (c) triangular wave; (d) haversine; (e) example of an arbitrary (user-defined) wave form.

instrument designer to another. However, the underlying concept can be illustrated by the structure of the function generator in Fig. 7 (Helfrick and Cooper, 1990).

This structure represents a function generator with three outputs (sine, triangular, and square waves) with a frequency range of 0.01 Hz to 100 kHz. The frequency can be adjusted either manually from the instrument front panel or from an external source. The frequency control circuit has the function of regulating two current sources.

The upper current source supplies a constant current i to the integrator A whose output voltage V_{out} increases linearly with time according to the relationship

$$V_{\text{out}} = -\frac{1}{C}\int i\,dt, \qquad (6)$$

where C is the capacitance of the condenser in the feedback path of the operational amplifier K. The slope of V_{out} varies proportionally with the upper current. The integrator output is connected to a comparator whose state changes at a predetermined level on the positive slope of the integrator's output. This change of state results in cutting off the upper current supply to the integrator, and the lower current source is switched in.

The lower current source supplies a reverse current to the integrator, resulting in a linear decrease of its output with time. At a predetermined level on the negative slope of the output, the comparator switches to cut off the lower current source while simultaneously switching in the upper source once again. The switching action between the two current sources results in a triangular wave form at the output of the integrator. The switching action of the comparator causes its output to toggle between two fixed voltage levels, resulting in a square wave. The sine wave can be generated from the triangular signal by a diode-resistance network. Such a network can be designed to alter the slope of the triangular wave in a piecewise linear approximation fashion as its amplitude changes, with < 1% distortion. Figure 8 illustrates the basic idea of piecewise-linear approximation of nonlinear functions by using the diode-resistance circuit. The accuracy of such an approximation is determined by the number of line segments used. The approximate curve is obtained by the summation of the individual line segments, whose "breakpoint" voltages and slopes are determined separately for each segment. The operational amplifier function is to sum up the segments by summing up the currents from the individual diode-resistance circuits at the summing junction. The feedback resistor R_f provides the scaling function.

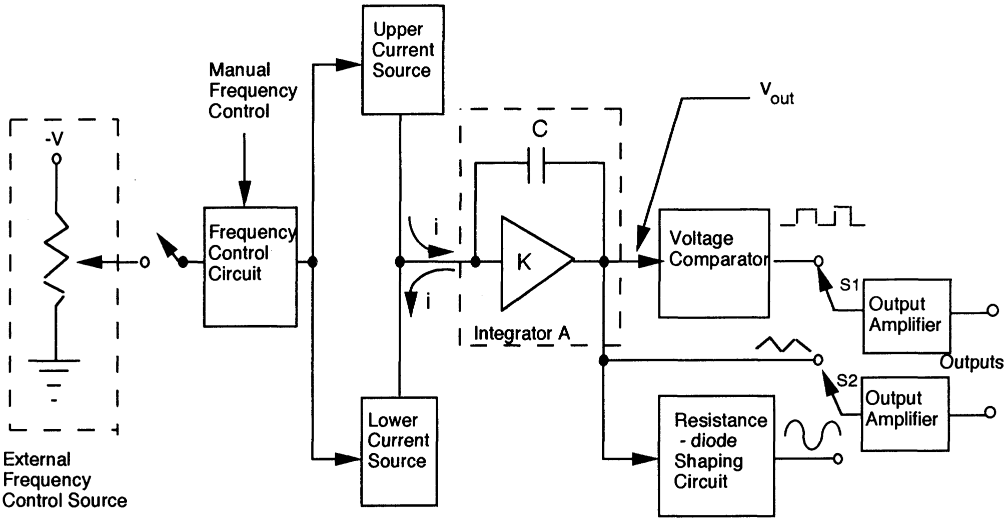

FIG. 7. Basic elements of a function generator providing three standard wave forms through two separate outputs. Some generators have only one output.

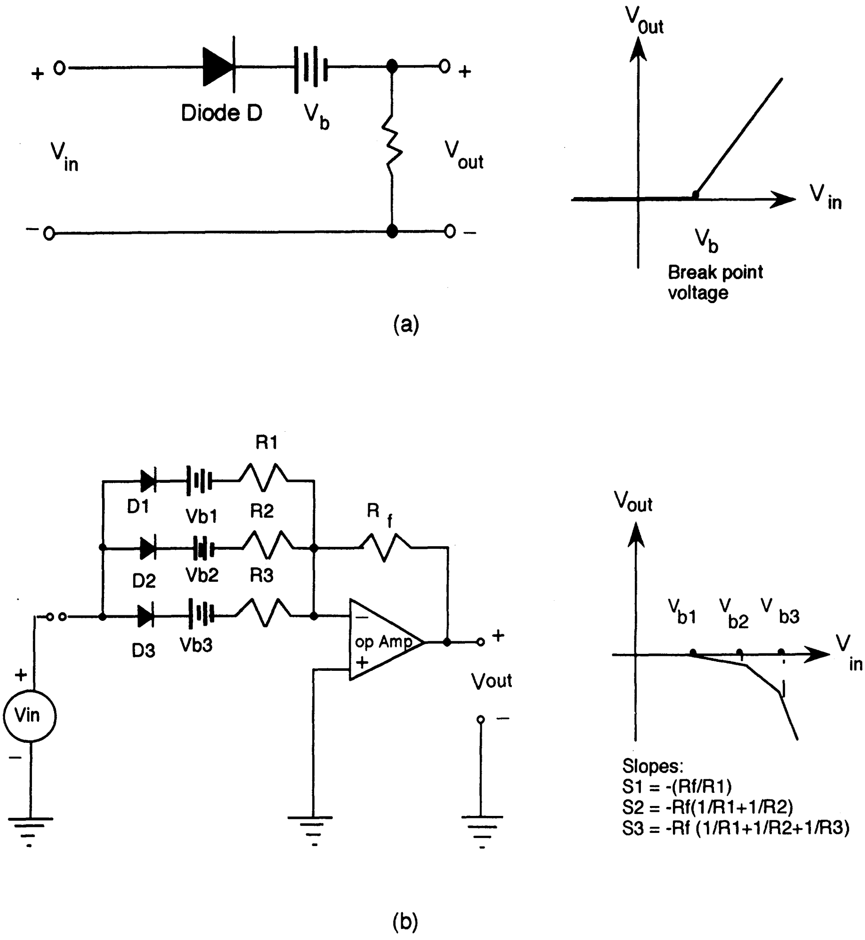

FIG. 8. Piecewise approximation: (a) simple line segment; (b) multiline segment approximation with summing amplifier.

1.4 Recorders

Recorders, in general, represent a fairly inexpensive class of recording devices primarily used for capturing and preserving data long after the signal has ceased. These recorders can be classified on the basis of their principle of operation and function. In this section, three conventional types of recorders that can be described as stand-alone instruments are discussed. The other forms of recorders require interfacing with some peripheral devices to provide the user with the level of man–machine interaction necessary for recalling and displaying the data. These types of recorders are discussed in the ELECTRICAL INSTRUMENTATION and RECORDING DEVICES and, therefore, will not be covered here.

1.4.1 Analog Recorders. The analog recorder receives analog (real-time) signals and produces a continuous time record or trace on a strip of calibrated paper (chart) traveling at a constant velocity under a motor-driven writing pen (Fig. 9). The widely known strip-chart recorder, commonly used in laboratory applications, has chart motion perpendicular to the pen deflection. The basic recorder has a single writing pen dedicated for a single input. The multipen version, usually up to six pens, produces traces of several signals simultaneously with different colors. Also available are multi-input recorders, each accommodating as many as 30 input signals and using a dot printing head instead of the writing pen. Circular-chart recorders are also available and are used primarily in industrial applications.

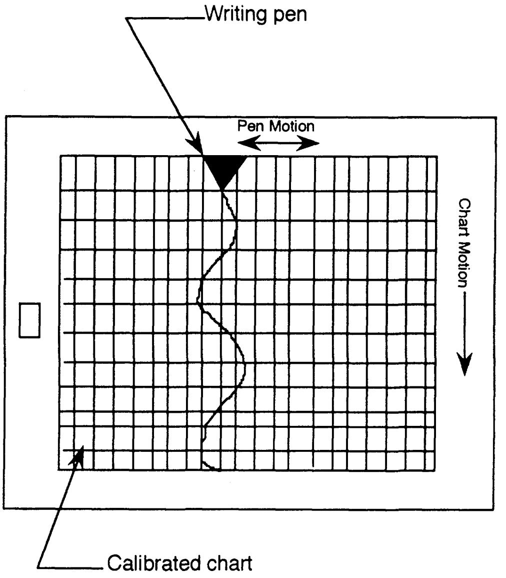

FIG. 9. Strip-chart recorder.

Some modern recorders use a pen-deflection mechanism based on improved versions of the moving-coil galvanometer type widely recognized as the D'Arsonval movement. More modern designs employ a digital servomechanism to drive the writing pen. Although servo systems vary in their design and hardware implementation, their principle of operation is very much the same. Figure 10 is a block diagram illustrating the principle of a closed-loop null-balance servo system used for driving the writing pen in recorders. The modern servomechanism consists basically of a dc motor and associated components (gears or linkages) to drive the pen and an encoder (position transducer) to detect and feed back the pen position. As a null-balance positioning system, the balance position is always being sought by the servo system as the input signal changes. The error produced is proportional to the signal change, causing the servo to drive the pen in the appropriate direction to offset the error. Scaling circuits are used to scale the input and the position signals appropriately before they are summed to produce the positional error.

With the advances in microprocessor technology, analog recorders are provided with advanced features unsurpassed by the earlier generation. Such features provide the analog recorder with improved flexibility useful for laboratory experimentation. One of the new powerful features is the programmability of input types, full-scale ranges, and chart speed without the need for operator programming skills. With this feature, the recorder can be programmed to record temperature by using a thermocouple or resistance-thermometer detector (RTD) in one application and can be easily reprogrammed to use the same recorder hardware with different input signals such as voltage or current in another application. Some computational capabilities are also now possible, such as computing the mean, minimum, and maximum values as well as having an integral digital readout and printout of the programmed setup. Although modern analog recorders are reliable, economical, and accurate (better than ±1%), their response is limited by the inherently slow pen mechanism. Typical frequency response is on the order of a few hundred hertz. Analog

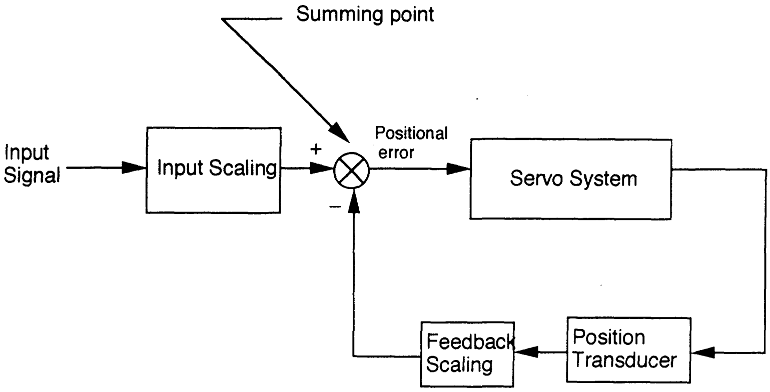

FIG. 10. Functional block diagram of a null-balance servo.

oscilloscopes (discussed in Sec. 1.2.1) also fall into the class of analog recording instruments.

1.4.2 Digital Recorders

1.4.2.1 Digital Wave Recorders. The development of high-speed ADCs (discussed in Sec. 3.1) together with the ready availability, small size, and reasonable cost of digital memories and associated digital-processing hardware has resulted in the development of the digital wave recorder.

The incoming analog voltage signal is digitized by an ADC with a resolution dependent on the digital output capability of the converter. An 8-bit ADC has a resolution of 1 part in 256 (0.4% accuracy), and a 10-bit ADC has a resolution of 1 part in 1024 (0.1% accuracy). The analog input can be sampled at adjustable rates (100 000 samples/s or more is possible). The sampled values are stored in a nonvolatile solid state memory for later processing of the data. The sampling rate and memory size must be selected to suit the duration and wave form of the physical event being recorded.

To view the wave form, stored data are read from memory through a DAC (discussed in Sec. 3.2), and the resultant signal is displayed on an integral CRT screen (discussed in Sec. 1.2). The data can be read from memory at a slow rate to allow the original event (which may have been rapidly changing, such as a one-shot transient) to be reproduced at a rate that is easy to view or even to record on one of the slower recording devices (such as a strip-chart or *X-Y* recorder). Data can also be read out from memory at a faster rate repetitively where, in such a case, a one-shot transient can appear as a continuous wave form that can be easily observed and analyzed. Digital memory can also be accessed directly (by bypassing the DAC) by a PC-based data acquisition system (DAS) (discussed in Sec. 2.3) for manipulation and processing of the recorded data.

Other developments in digital recording devices include visual trend display, where several channels of information are continuously and simultaneously updated, stored, and displayed on the CRT as bar charts or in pen-recorder-type traces.

1.4.2.2 Programmable Recorders. Combining digital technology, microcomputing, and analog recording capabilities has allowed a programmable class of recorders to become available. The result is a digital potentiometric recorder that is superior to the analog recorder. The programmability of the device provides the kind of flexibility and diversity required by the experimental nature of laboratory investigations. The sampled and stored data can be plotted against time or one channel against another with 0.1% full-scale accuracy. Additional capabilities attributed to the intelligence provided by on-board microprocessors are computing functions such as addition, subtraction, multiplication, division, averaging any two curves point by point, differentiation, integration, and normalization. In addition to these functions, the data for plotting can be offset, scaled, or expanded. Another form of digital recording device is the digital oscilloscope discussed in Sec. 1.2.2.

1.4.3 *X-Y* Recorders. The *X-Y* recorder, also known as the *X-Y* plotter, is another type of analog recorder, in which one variable can be plotted against a second variable in two-dimensional Cartesian coordinates (Fig. 11). The function plot is produced by simultaneously introducing two signals to the instrument, where each signal serves as an input to independent pen-driving mechanisms. The writing pen is driven in a two-dimensional plane by the two systems (*X* axis and *Y* axis) to produce the plot on chart paper. The most common chart paper sizes for this type of recorder are the standard 21.6×27.9 cm ($8\frac{1}{2}\times11$ in.) or 27.9×43.2 cm (11×17 in.). The chart paper in this case is held in place on a flat surface by an electrostatic field. Other instruments are available with roll chart paper with a variable speed selectable over a wide range to a maximum of 80 cm/s (31.5 in./s) (for example), with accuracy > 1% of full scale.

Two common mechanisms are available for driving the writing pen: (1) the servo system, described earlier, mostly employed in modern *X-Y* recorder designs, and (2) the optical galvanometer. The optical galvanometer offers significant improvements by eliminating the friction and inertia problems associated with the conventional galvanometer. As illustrated by Fig. 12, a small fixed mirror fastened to a moving coil is positioned to reflect a beam of light (from a light source) on light-sensitive, moving chart paper. When the coil turns, the light beam deflects over the paper, causing a trace on the chart. The re-

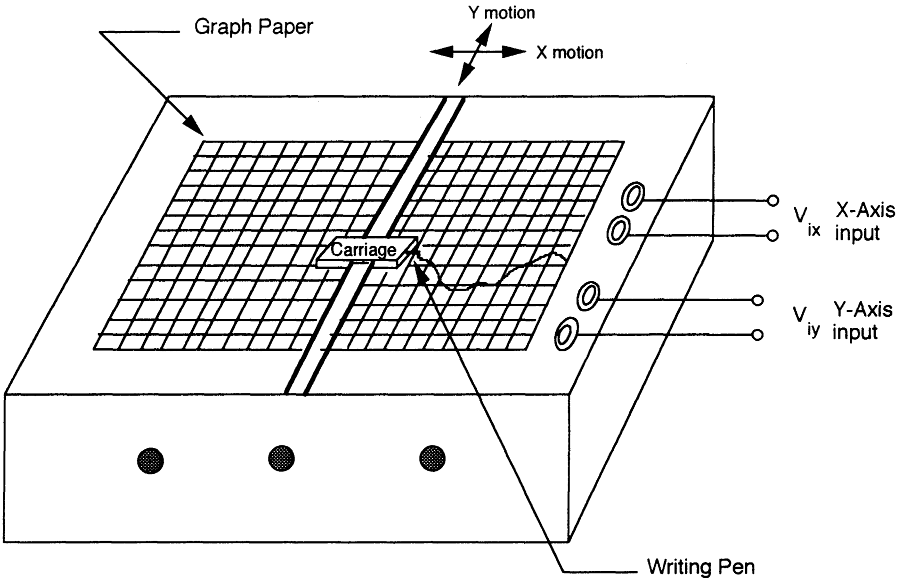

FIG. 11. *X-Y* recorder (plotter).

cording speed varies somewhat between self-balancing potentiometer types (earliest form of recording mechanism) and oscilloscopes (analog form). The frequency response of the structure shown is flat from dc to 100 Hz at full-scale deflections and nonlinearity of 1% of full scale (Doebelin, 1983).

Many *X-Y* recorders also can perform as strip-chart recorders by allowing for one of the axes to serve as a time base if so desired. The strip-chart recording function can be accomplished by introducing a voltage ramp to one of the axes and the signal to the other. The ramp can be generated by charging a capacitor with a constant current, thus causing the voltage to change linearly with time. The corresponding servo system will then cause its axis to translate at a constant speed. Most modern *X-Y* recorders offer the choice of a wide range of dc input voltages for each axis. As many as 19 different ranges and a sensitivity from as low as 0.1 mV/cm to 50 V/cm is not uncommon.

1.5 Spectrum Analysis

In many measurement applications, the analysis of signals in the frequency domain in terms of amplitude versus frequency is quite important and desirable; vibration analysis is a good example. The distribution of signal amplitude against frequency is known as the spectrum. Spectrum analysis and computation depend on signal behavior whether periodic or nonperiodic. Almost all physical measurements that we deal with can be classified as *periodic* or *nonperiodic.*

Fourier (the mathematician) showed that periodic signals (e.g., sine, square, triangular wave) expressed in the continuous-time domain can be represented by a sum of components at harmonic frequencies. If $x(t)$ is a periodic function [i.e., $x(t+T)=x(t)$, where T is the time period], then it can be approximated by Fourier transform by the summation

$$x(t)=\sum_{n=0}^{\infty} a_n \cos\omega_n t+b_n \sin\omega_n t, \tag{7}$$

where

$$a_n=\frac{1}{T}\int_0^T x(t)\cos\omega_n t\, dt,$$

$$b_n=\frac{1}{T}\int_0^T x(t)\sin\omega_n t\, dt,$$

$$\omega_n=n\omega_0,$$

ω_0=fundamental radian frequency,

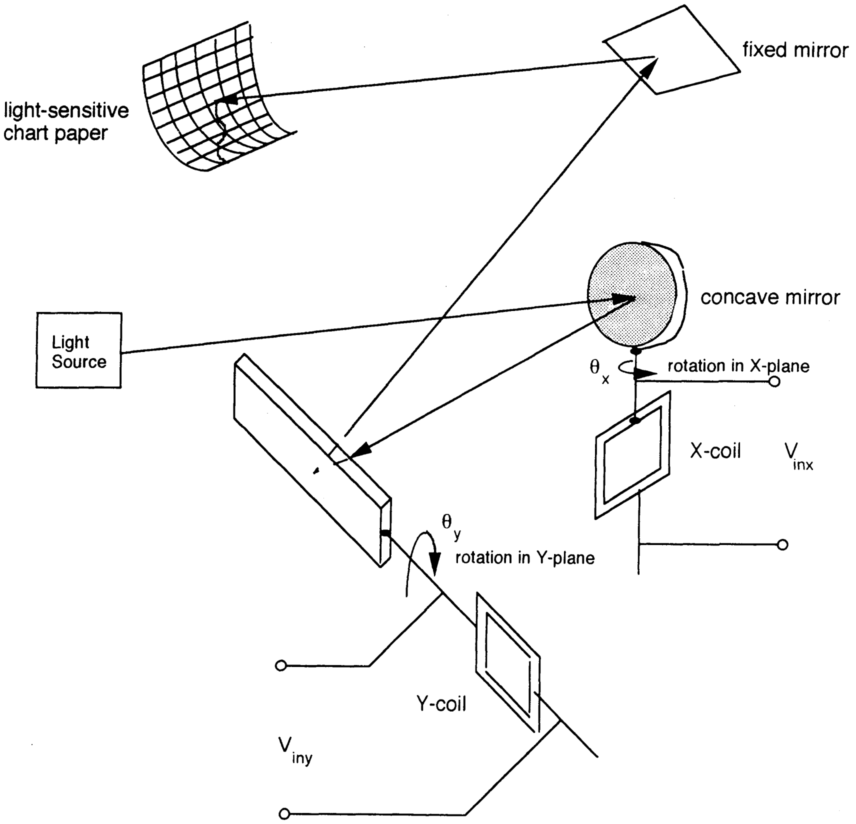

FIG. 12. A simplified structure of an optical galvanometric *X-Y* recorder.

which results in a discrete amplitude–frequency spectrum where frequency contents exist at discrete values only, as shown in Fig. 13(a).

With the advent of computer technology and its growth in the 1960s, a fast-Fourier-transform (FFT) algorithm [based on the discrete Fourier form (DFT)] was developed by Tuckey and Colley to compute the FFT by using digital computers [see FOURIER AND OTHER MATHEMATICAL TRANSFORMS or one of the textbooks on the subject (e.g., Brigham, 1974)]. The FFT algorithm is simply an efficient means of calculating the DFT $X(n)$ of a discrete signal $x(kT)$ (signal expressed in discrete-time domain) by

$$X(nf_s/N) = \sum_{k=0}^{N-1} x(kT_s)\,W^{nk}, \tag{8}$$

where $W = e^{-j(2\pi/N)}$ is a complex exponential weighting term, and the sampling frequency $f_s = 1/T_s$, which for sampling interval $T_s = 1$ can be written in the simpler form

$$X(n/N) = \sum_{k=0}^{N-1} x(k)\,W^{nk}. \tag{9}$$

For nonperiodic signals, typical of most physical measurements, the amplitude–frequency spectrum is continuous where frequency contents are distributed over some range depending on the signal, as shown in Fig. 13(b). To calculate the frequency spectrum of nonperiodic signals, another technique based on the concept of power spectrum density function is used, as will be discussed in Sec. 1.5.2.

1.5.1 Fourier Analyzers. The development of the computationally efficient FFT technique led to the development of the Fourier analyzer. FFT-based analyzers provide fast, high-resolution spectrum and network analysis with the information presented on a CRT. Unlike the spectrum analyzer discussed below (swept tune), the Fourier analyzer can measure dynamic signals (in real time) because they measure all frequencies simulta-

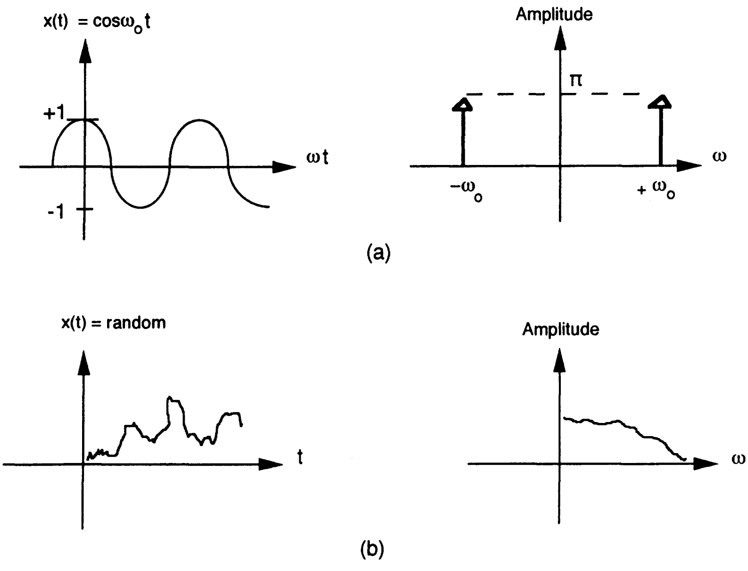

FIG. 13. Examples illustrating the frequency spectrum of periodic and nonperiodic functions: (a) periodic function; (b) nonperiodic function.

neously (not one at a time). Also, this analyzer provides both the magnitude and the phase information of each frequency component. The analyzer is useful for measuring signals from a few microhertz to about 100 kHz. In transient–wave-form analysis, signals can be sampled, digitized, and stored in internal memory or sent by one of several interfaces [e.g., general-purpose interface bus (GPIB), discussed in Sec. 2.3] to an external storage device. The stored information can be recalled and analyzed in time, frequency, and amplitude domains.

Some modern designs incorporate measurement of the statistical properties of signals as well as the joint properties of two or more signals. These measurements are made possible because of the availability of phase information afforded by the analyzer.

1.5.2 Spectrum Analyzers. For nonperiodic functions (signals), the spectrum is calculated on the basis of the power spectral density concept. For a complex, nonperiodic function $x(t)$, its power spectral density $S_{xx}(\omega)$ can be expressed as (Sydenham, 1982)

$$S_{xx}(\omega)=\frac{1}{2\pi}\lim_{T\to\infty}\frac{|X(j\omega)|^2}{2T}, \qquad (10)$$

where $X\ (j\omega)$ is the Fourier transform of $x(t)$ given by

$$X(j\omega)=\int_{-\infty}^{+\infty} x(t)e^{-j\omega t}dt. \qquad (11)$$

It is clear from Eq. (10) that the power spectral density is a real-valued function of ω and even, so that $S_{xx}(\omega)=S_{xx}(-\omega)$. The power P of the entire signal in the frequency domain is related to the power spectral density $S_{xx}(\omega)$ by Parseval's equation (Sydenham, 1982) as

$$P=\overline{x^2(t)}=\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{+T} x^2(t)dt$$
$$=\int_{-\infty}^{+\infty} S_{xx}(\omega)d\omega. \qquad (12)$$

Equation (12) shows that the spectrum of a signal at various frequencies is directly related to the signal power at these frequencies. Such a relationship provides the basic concept in spectrum analyzer design.

Design implementation of spectrum analyzers can be generally summarized in the following steps:

1. frequency filtering of the signal by a narrow-band–pass filter,
2. squaring the instantaneous value of the filtered signal,
3. averaging the squared instantaneous value over the sampling time, and
4. computing the rms value of the result.

As shown by Fig. 14, most modern spectrum analyzers perform the filtering function by a technique known as heterodyning, which al-

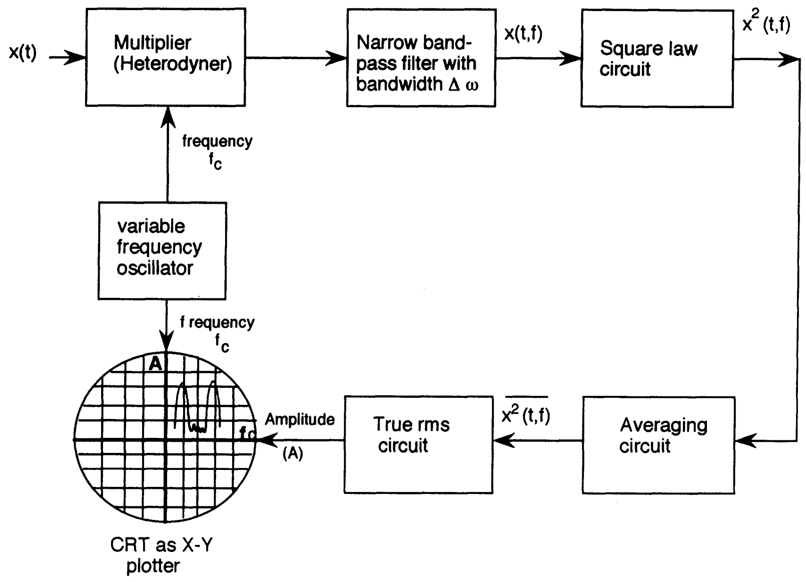

FIG. 14. Functional block diagram of a spectrum analyzer.

lows the use of a single fixed-band–pass filter throughout the entire range of frequencies. Without heterodyning, several filters of fixed bandwidth staggered in frequency would have to be used to cover the desired frequency range. The heterodyning operation consists of simply multiplying the signal with a sine wave whose frequency f_c can be accurately and quickly changed through a range of values. This multiplication process is equivalent to amplitude modulation that shifts any frequency content of the signal at, say, frequency f_s to f_c+f_s.

The filtered signal $x(t,f)$ is that part of $x(t)$ at frequency f. After squaring and averaging $x(t,f)$, the spectrum can be computed by using a calibrated TMS voltmeter circuit, as described in Sec. 1.1. The variable frequency from the oscillator provides the sweep signal for the synchronization with the x-axis sweep of the CRT to generate the amplitude spectrum on an oscilloscopelike display.

Spectrum analyzers can provide a broad frequency range from a few hertz to >300 GHz with a wide amplitude range from less than −100 to about +30 dBm. With the proper instrument interface protocol, the analyzers can be directly connected to plotters or printers to provide hard copies of the CRT displays.

1.6 Counters

A most convenient and widely used instrument for accurate measurement of frequency and time intervals is the electronic counter. Counters can perform a wide range of functions including high-accuracy frequency calibration of radio-frequency (rf) signals, frequency measurements on noisy signals, accurate propagation delay measurements, and phase shift in filter signals.

Modern counters employ the state of the art in digital technology and ICs by providing high-precision measurement and powerful analyzing capabilities—all in one compact, cost-effective instrument. With on-board microprocessors, counters can even provide some computational capabilities with built-in mathematical and statistical functions such as inversion, normalization, scaling, mean value, standard deviation, and minimum and maximum values. In addition to the integrated digital display of the measurement, the counter can have other outputs such as binary-coded decimal (BCD) (discussed below) and analog output for strip-chart recording. The versatility of the counter can also be extended by adding the IEEE-488 standard instrument interface bus for communication with other instruments and remote programmability of the counter.

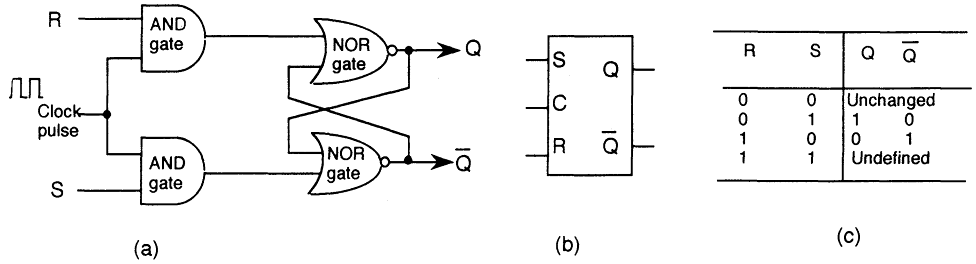

(a) (b) (c)

R	S	Q $\bar{Q}$
0	0	Unchanged
0	1	1 0
1	0	0 1
1	1	Undefined

FIG. 15. Clocked SR flip-flop using NOR gates: (a) circuit diagram; (b) block diagram; (c) state table.

The most basic element of the counter is a digital device called the *flip-flop* (FF). The flip-flop, also called a bistable multivibrator, is an extremely versatile logic device that is used extensively in most digital circuitry including digital computers. The basic FF, known as the RS FF, can be formed by cross-connecting two NOR gates or two NAND gates (Fig. 15 shows the NOR configuration to illustrate the basic operation of the FF). The RS FF device has three inputs: S (SET), R (RESET), and CP (CLOCK PULSE). The S and R inputs are normally at logic 0 (also referred to in the digital field as logic LOW or 0 state). If the input applied to S is changed from logic 0 to logic 1 (referred to in the digital field as logic HIGH or 1 state), the Q output goes to logic 1 and the complemented (inverted) output $\bar{Q}$ goes to logic 0. Even if the S input returns to the 0 state, Q remains at the 1 state because of the internal feedback between the two NOR gates. This operation is called *setting the FF.* Similarly, when the input applied to R changes to the 1 state, Q goes to the 0 state (and $\bar{Q}=1$ state) and remains in that state. This operation is called *resetting the FF,* as indicated by the table of Fig. 15(c). Important to note is that the S and R inputs should not go to logic 1 simultaneously, because the output state in this case is not defined. In practice, the FF would be one component in a digital circuit performing a specific function. Therefore, the operation of the FF must be synchronized with the rest of the circuitry. Synchronization is accomplished by a CLOCK PULSE at the CP input. This pulse is usually a square pulse with

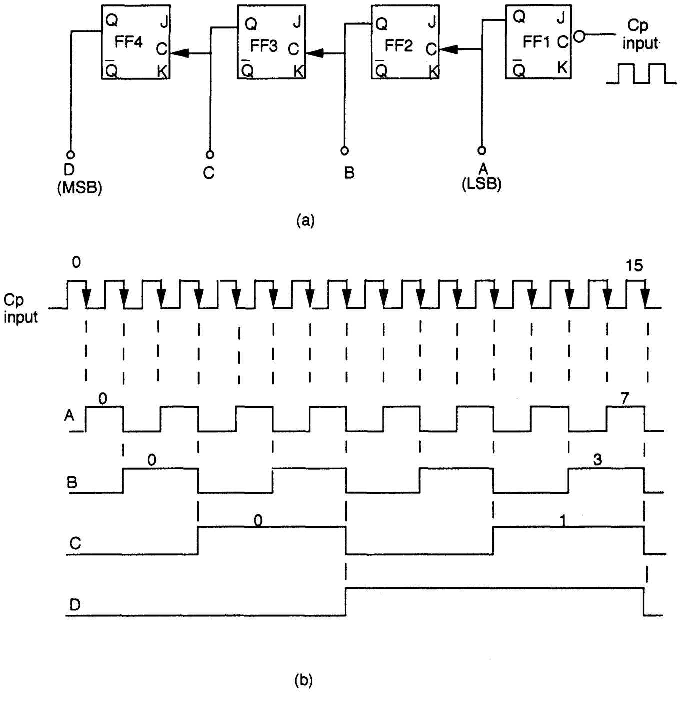

FIG. 16. Four flip-flops forming a 4-bit (MOD-16) binary counter: (a) interconnection diagram; (b) wave forms showing frequency division.

specific frequency. It allows the FFs to change states, on the basis of the present states of inputs R and S, only once for each clock pulse. Other FF forms exist, but their discussion is beyond the scope of this article. For further information on digital logic devices, refer to a textbook on the subject (e.g., Kohavi, 1978). A wide variety of counters (in their basic forms) are available as standard IC packages, and it is seldom necessary to construct a counter from individual FFs. Therefore, coverage of internal operation of counters will be limited to discussing external operating characteristics of currently available counters.

Flip-flops can be connected in various arrangements to function as binary counters that count input clock pulses or events. A 4-bit counter, for example, contains four FFs, one per bit, with the number of outputs equal to the number of bits (i.e., four). As shown by Fig. 16, the counter basically has one input, the clock pulse CP. It is assumed in this counter structure that the FFs are of a form known as JK, where J and K are the FF inputs, and that both J and K are kept at logic 1. In this configuration, the outputs of the JK FF change states on the high-to-low transition of the clock pulse, as indicated by the small circle at the input CP. Some counters trigger on low-to-high transition, in which case the input CP would not have the small circle. The states of the four FFs represent a binary number equal to the number of pulses that have been applied to the CP input. In the diagram, the A output represents the least significant bit (LSB) and D is the most significant bit (MSB) of the binary count. For example, after the fifth input pulse, the output DCBA=0101, which is the binary equivalent of five.

The 4-bit counter can count from 0000_2 (zero) to 1111_2 (fifteen), for a total of 16 distinct states. After the count reaches the maximum, the next input pulse causes the FF outputs to recycle to the 0000 count, and the counting sequence is then repeated with subsequent pulses. Generally, a counter with N FFs can count from 0 up to $2^N - 1$, for a total of 2^N distinct states.

The total number of states is known as the MOD number. The counting sequence can be cut short at any time by applying logic HIGH to a second input (RESET) that can be incorporated in the design. The decade counter is such an example, which counts from 0000 to 1001 (0 to 9) and then recycles to 0000. It has 10 distinct states and thus is referred to as a MOD-10 counter. It is also called a BCD counter because the FF outputs count through 10 in the common digital code known as the BCD code. The BCD is a weighted code where four bits are used to represent each decimal digit. The bits carry weights 8, 4, 2, 1 in the order from the MSB to the LSB, as shown in Table 3.

The counting capability of FFs can be demonstrated by examining the wave form in Fig. 16(a). It is evident that the frequency of the square wave (at CP) is divided by 2 successively by each FF. Therefore, cascading more FF units will produce division by 4, 8, 16,..., etc. In general, the wave form out of the MSB FF of a counter will divide the input frequency by the MOD number. For example, the MSB FF of a 4-bit counter has a frequency that is 1/16 of the clock input, as is evident by comparing waveform D to the clock pulse CP.

Counters can count up from zero to some maximum count and then reset to zero, and up/down counters can count in either direction. Some counters have a single CP input that is used for both count-up and count-down operations, whereas others have two separate CP inputs—one for counting up and another for counting down. For the latter type, only one input can be pulsed at one time to avoid

Table 3. The 8-4-2-1 binary-coded decimal (BCD) code representation of decimal numbers.

Decimal number	8-4-2-1 BCD code
0	0 0 0 0
1	0 0 0 1
2	0 0 1 0
3	0 0 1 1
4	0 1 0 0
5	0 1 0 1
6	0 1 1 0
7	0 1 1 1
8	1 0 0 0
9	1 0 0 1
10	0 0 0 1 0 0 0 0
11	0 0 0 1 0 0 0 1
12	0 0 0 1 0 0 1 0
⋮	⋮
19	0 0 0 1 1 0 0 1
20	0 0 1 0 0 0 0 0
⋮	⋮
99	1 0 0 1 1 0 0 1
100	0 0 0 1 0 0 0 0 0 0 0 0
⋮	⋮

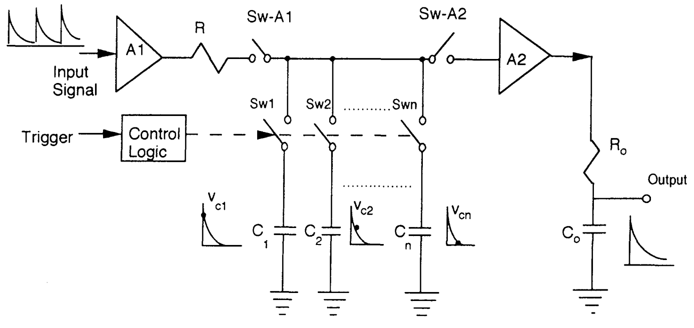

FIG. 17. Descriptive schematic diagram of a signal averager. Upon receiving a trigger, the control logic closes the switches one at a time in sequential order.

erratic operations. Some IC counters are designed to give the user the capability of varying the counting sequence through appropriate external connections. These types of counters are usually referred to as variable-MOD counters.

Commercially, counters are classified into frequency counters and general-purpose timer–counters. The frequency counter has a single input channel and sometimes an additional prescaled input for measuring high frequencies. In addition to frequency measurements, most frequency counters can also perform period measurements. The timer–counter version can perform time-interval measurements between two events in addition to frequency and period functions. For a given counter, a frequency can be used as a guide in deciding the suitability of frequency versus period measurement. For a 1-s sampling period, this frequency f_0 is given by (Doebelin, 1983)

$$f_0=(f_c)^{1/2}, \tag{13}$$

where f_c is the frequency of the crystal oscillator ("clock" frequency).

For example, a 10-MHz clock-frequency counter has $f_0=3160$ Hz, above which frequency measurements should be used and below which period measurement should be used.

A more advanced version can provide pulse-width measurements; totalizing capabilities (pulse counting); for the timer/counter type, rise/fall time computation; and phase delay between two events.

1.7 Boxcar Integrators

In some measuring applications, it is desirable to trigger a process repeatedly, such that on repetition, the resulting signal will be identical each time and will appear to be superimposed upon itself. To enhance the signal-to-noise ratio (SNR) of the repetitive signal, an averaging process is performed to smooth out the signal. Two known analog techniques used for this purpose are the signal averager and the boxcar integrator.

The signal averager, as shown in Fig. 17, consists of a bank of parallel capacitors connected to the output amplifier A2 through switch SW-A2. The capacitors are charged by closing in a specific sequence a set of electronic switches known as field-effect transistors (FETs), which are solid-state devices with very high input impedance. The sequence is determined by the timing control circuit controlling the FETs in response to an external trigger signal. The sequence always begins at a fixed time following application of the trigger signal. Consequently, upon repetition, each capacitor circuit receives part of the applied signal through amplifier A1 with the same phase each time by charging the capacitors through the resistance *R* by a small amount each time. The result after a large number of repetitions is that each capacitor contains an averaged sum corresponding to a point on the signal trace. The information stored on the capacitors can be retrieved by allowing switches SW1–SW*n* to be closed one at a time while SW-A1 is opened and SW-A2 is closed. The output resistance–capacitance combination R_0C_0 is used to smooth the final response, which otherwise would look like changing steps because of the switching action from one channel to the next in succession.

The boxcar integrator, also known as the boxcar averager or detector, performs the same function as that of the signal averager, except with a rather simpler circuit. Instead of using a bank of many capacitors, the boxcar employs a single capacitor circuit, as shown in Fig. 18, to take only one sample during each

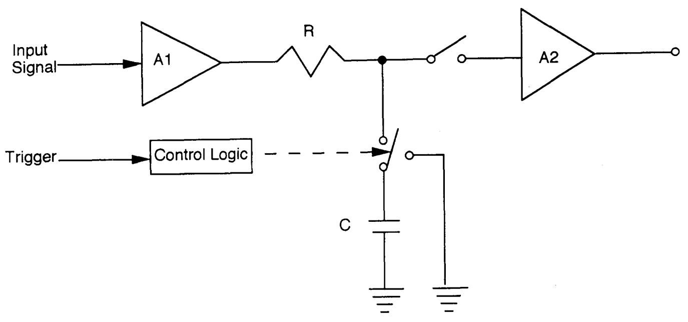
Input Signal
A1
R
A2
Trigger
Control Logic
C

FIG. 18. A simplified schematic of a boxcar integrator.

signal occurrence and requires a trigger signal at a fixed time interval.

The trigger pulse is used to delay by a fixed time interval the closing of the switch. The capacitor in this case is clearly taking an average of a set of readings for one single point. The entire process is then repeated with different delay-time increments until all points are averaged. [See Sydenham (1982) for more details.]

1.8 Lock-In Amplifiers

The lock-in amplifier is used to recover weak ac signals corrupted with noise. One common application of lock-in amplifiers is spectroscopy, where the generated signal is typically weak. By amplitude-modulating signals at some reference or carrier frequency, the noise often can be reduced and dc drift problems avoided. The main principle of signal recovery is based on the synchronous detector (also known as the phase-sensitive detector or PSD) part of the circuit, which in essence functions as a demodulator. As shown by Fig. 19, the PSD circuit consists primarily of a mixer (multiplier) followed by a low-pass filter. The output of the mixer is the product of two signals x and y. The phase difference between the two inputs can be adjusted precisely by the phase shifter. Mixer circuits used in lock-in amplifiers must be capable of withstanding large amounts of noise without overloading. The square wave used by the mixer is supplied by a sine-to-square-wave converter (this could be simply a comparator—sensing zero crossing). If the square-wave frequency is precisely the same as that of the input signal and if it coincides in phase with its zero crossing, the result would be a type of full-wave rectification. The SNR of the noisy amplitude-modulated signal can be improved by reducing the bandwidth with a band-pass filter tuned to the carrier frequency. The signal can finally be recovered by a low-pass filter, which removes all components of the mixer output that have frequencies beyond the filter cutoff frequency. Some lock-in amplifiers are commercially available with built-in oscillators that can be used to supply excita-

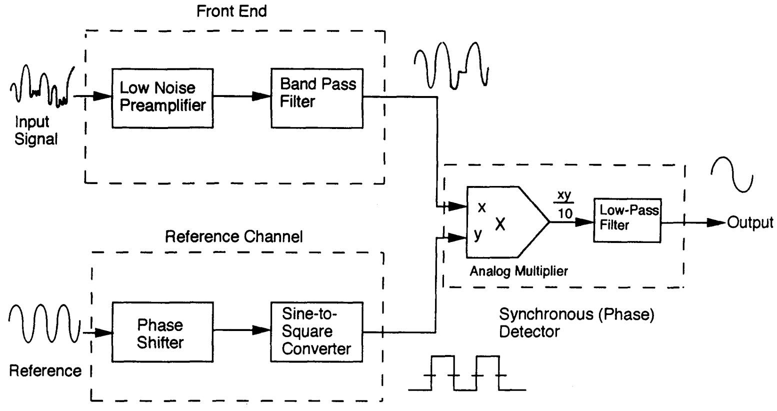
Front End
Input Signal
Low Noise Preamplifier
Band Pass Filter
Reference Channel
Reference
Phase Shifter
Sine-to-Square Converter
x
y
X
xy/10
Low-Pass Filter
Output
Analog Multiplier
Synchronous (Phase) Detector

FIG. 19. Functional block diagram of a lock-in amplifier.

tion to an external system as well as the needed reference signal such as in strain gauge measurement (Vassos and Ewing, 1985; Sydenham, 1982).

1.9 Time-Domain Reflectometry

Time-domain reflectometry (TDR) is an echo technique used in transmission-line analysis to characterize and detect discontinuities in the line. This method also can identify the location and nature of the mismatch resulting from the discontinuity. This identification can be accomplished by applying a pulse of energy and monitoring the reflection of that energy by an oscilloscope. In the laboratory, the TDR is used primarily to test electrical cable (e.g., coaxial cable) and to check thermocouple characteristics in a thermopile.

The TDR system consists primarily of a step generator and an associated oscilloscope designed for this type of application, as shown by Fig. 20. The step generator produces a positive-going wave (incident wave) that is applied to the transmission system under test. The wave propagates through the transmission line at the velocity of propagation of the line. If the load impedance Z_L is equal to the characteristic impedance Z_0 of the line, then no wave reflection occurs, and the incident step voltage is displayed on the oscilloscope without change in its shape. If a mismatch exists at the load, part of the incident wave is reflected. The reflected wave is displayed on the oscilloscope's CRT as algebraically added to the incident wave, as depicted in Figs. 21 and 22 (Hewlett-Packard, 1988).

Analyzing the time delay, magnitude, and shape of the reflected wave form permits determination of location and nature of the mismatch whether the mismatch is at the load or somewhere along the transmission line.

Time-domain reflectometry is much more useful and advantageous than the classical standing-wave-ratio technique (frequency-domain measurement), which requires measurement at many frequencies to identify the mismatch—a rather tedious and time-consuming procedure. Also, with the technological advances in the optical field and the increased applications of fiber optics both in the laboratory and in industry, the optical TDR presents a powerful tool for testing the fiber. This device can analyze the reflected light energy to determine the existence and location of breaks in the fiber and losses at splices and connectors as well as the total loss of the system (Helfrick and Cooper, 1990).

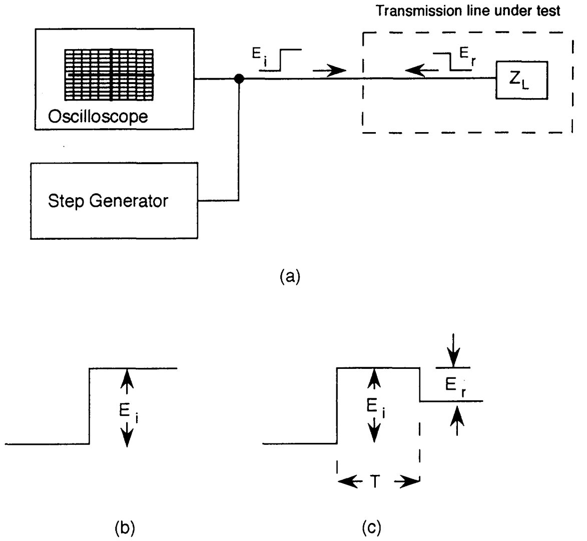

E_i : incident voltage

E_r: reflected voltage

Z_o: characteristic impedance

Z_L: load impedance

$$\frac{E_r}{E_i} = \frac{Z_L + Z_o}{Z_L + Z_o}$$

FIG. 20. Time-domain reflectometer system: (a) typical reflectometer measurement set-up; (b) oscilloscope display when $E_r = 0$; (c) oscilloscope display when $E_r \neq 0$.

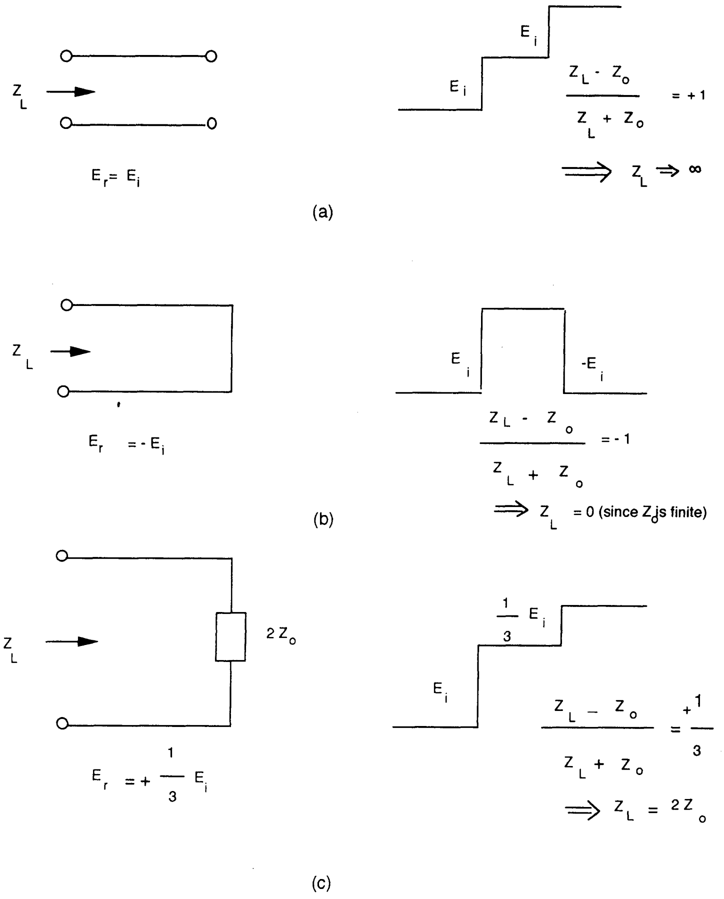
Z_L
E_r = E_i
E_i
E_i
(Z_L - Z_o)/(Z_L + Z_o) = +1
⇒ Z_L ⇒ ∞
(a)
Z_L
E_r = -E_i
E_i
-E_i
(Z_L - Z_o)/(Z_L + Z_o) = -1
⇒ Z_L = 0 (since Z_o is finite)
(b)
Z_L
2Z_o
E_r = +1/3 E_i
1/3 E_i
E_i
(Z_L - Z_o)/(Z_L + Z_o) = +1/3
⇒ Z_L = 2Z_o
(c)

FIG. 21. Time-domain reflectometer displays for typical loads: (a) open-circuit termination; (b) short-circuit termination; (c) line terminated with $Z_L = 2Z$.

2. DATA ACQUISITION INSTRUMENTATION

In many measurement situations, one is faced with vast quantities of information—mostly in analog form—generated by transducers. Gathering and processing the measured data with the speed and accuracy demanded by the challenging and complex scientific development tasks necessitates a more complex, more sophisticated type of instrumentation. Data acquisition systems (*q.v.*) (DASs) are used for that purpose. A DAS may be as simple as a DVM or it may be complex enough to contain a large-scale computer as part of its hardware. Ordinarily, DASs are digital systems since, with today's technology, complex circuits are low cost, accurate, and relatively simple to implement and easy to interface with human operators. As depicted by Fig. 23, the DAS generally accepts a wide mix of analog signals representing physical parameters; performs signal conditioning functions (filtering, linearization, scaling, and square root); converts the signals into digital form; performs various computations; analyzes data; and interfaces with many external peripheral devices (e.g., CRTs, digital recorders, printers, and mass storage devices). The analog multiplexer is a high-speed scanning device allowing inputs to be switched sequentially one at a time for processing and converted to digital form by using the same ADC.

2.1 Data Loggers

The data logger is a data acquisition system whose architecture resembles very closely that shown in Fig. 23. It provides an automated method of performing measurements and recording data as well as performing some computations with high speeds and accuracy. These functions can be performed at precise,

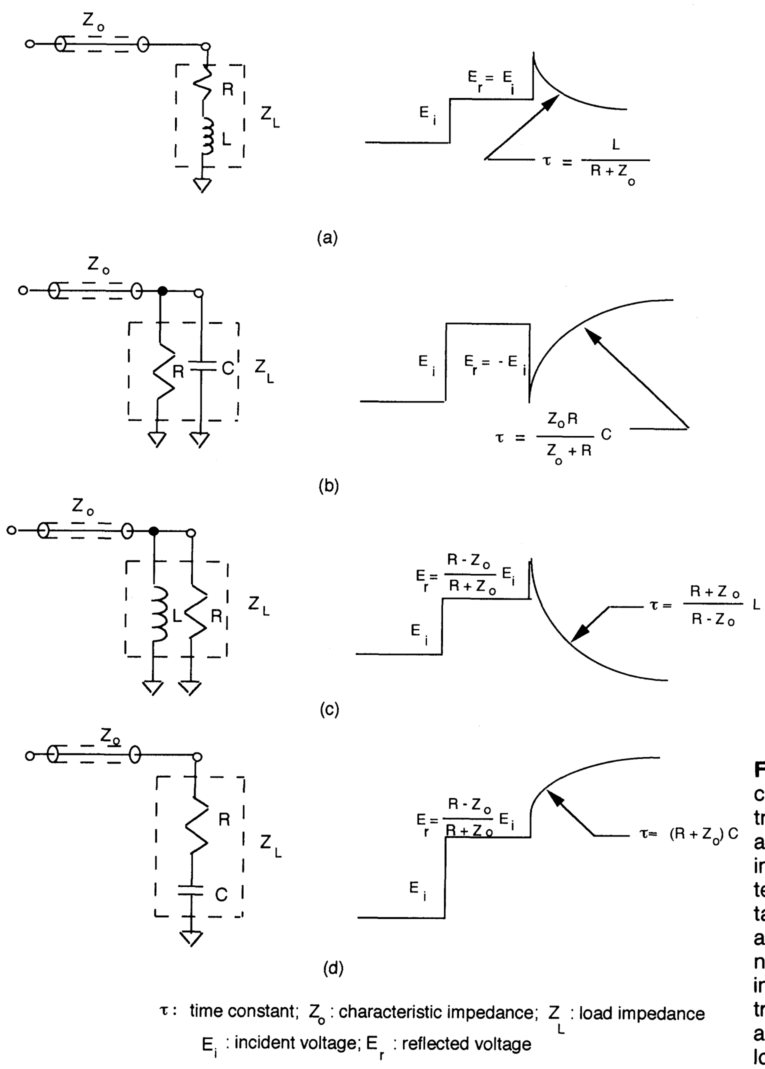
Z₀
R
L
Z_L
E_r = E_i
E_i
τ = L / (R + Z₀)
(a)
Z₀
R
C
Z_L
E_i
E_r = - E_i
τ = Z₀R / (Z₀ + R) C
(b)
Z₀
L
R
Z_L
E_r = (R - Z₀)/(R + Z₀) E_i
E_i
τ = (R + Z₀)/(R - Z₀) L
(c)
Z₀
R
C
Z_L
E_r = (R - Z₀)/(R + Z₀) E_i
E_i
τ = (R + Z₀) C
(d)
τ : time constant; Z₀ : characteristic impedance; Z_L : load impedance
E_i : incident voltage; E_r : reflected voltage

FIG. 22. Oscilloscope displays for complex load impedance Z_L: (a) transmission line terminated with a series resistance-inductance load impedance; (b) transmission line terminated with a parallel resistance-capacitance load impedance; (c) transmission line terminated with parallel resistance-inductance load impedance; (d) transmission line terminated with a series resistance-capacitance load impedance.

programmable intervals; once the logger is configured for the application, it will carry out the measurement precisely, unattended.

Data loggers offer good flexibility such as mixing signals representing various physical parameters and display each signal with the appropriate engineering units (pascals, liters per second, degrees Celsius or Fahrenheit). In addition, the same logger can be easily configured for different applications. Data loggers, however, have limitations compared with other DAS forms. Loggers are limited in their ability to provide histograms, trends, time function plots, and curve fits of data required for analyzing data. Although data loggers can interface with many external devices (output devices), they are still limited to some extent in their ability to interface directly with mass storage devices such as hard-disk drives. Some modern data-logger designs include microprocessors and two separate internal memories. One memory (RAM) for the system and user program can run without interference from a host, while the other memory can provide > 100 kbytes of data storage. In addition, an RS-232 or IEEE-488 interface can be included for communication with a host computer for data transfer and remote control of the logger. With a 12-bit ADC, the logger can provide up to 62 500 readings/s; while with a 16-bit ADC, it can provide up to 50 000 readings/s.

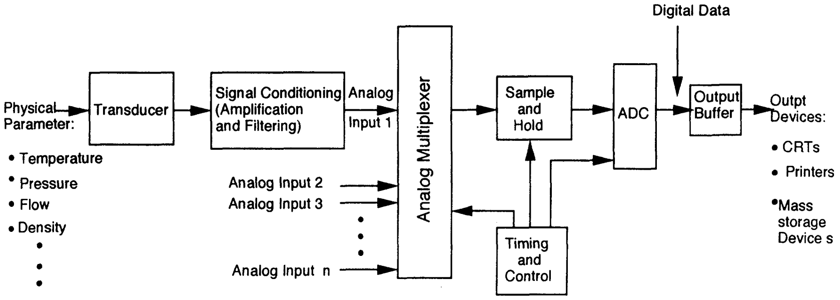
Digital Data
Physical Parameter:
Transducer
Signal Conditioning (Amplification and Filtering)
Analog Input 1
Analog Input 2
Analog Input 3
Analog Input n
• Temperature
• Pressure
• Flow
• Density
Analog Multiplexer
Sample and Hold
Timing and Control
ADC
Output Buffer
Outpt Devices:
• CRTs
• Printers
•Mass storage Device s

FIG. 23. Typical architecture of data acquisition system.

2.2 PC-Based Data Acquisition Systems

The high speed and low cost of personal computers (PCs) along with extensive hardware and software support made PCs an ideal choice for automated data acquisition systems in the laboratory and even in industrial applications. Most of today's PC-based DASs are based on either the IBM PC or the Apple Macintosh. As illustrated by Fig. 24, the basic elements of a PC-based DAS are listed below.

1. A PC with an operating system. The common operating system for the IBM PC is the PC-Disk Operating System (PC-DOS) and the Microsoft Disk Operating System (MS-DOS) for IBM-compatible PCs (commercially known as *clones*).
2. DAS boards for analog measurements and control. Analog input boards include an ADC, which is necessary to convert analog signals to digital form for further processing by the PC. Analog output boards also include a DAC, which is necessary to interface the DAS with the analog world. A host of other input/output (I/O) boards are available for practically any application. DAS boards are designed to be inserted directly into available expansion slots inside the PC chassis. In this case, the boards connect directly (through a standard connector, which is part of the board) into the PC bus. Some manufacturers offer stand-alone systems in which DAS boards can be inserted into a separate chassis, which in turn connects to the PC through an RS-232 or an IEEE-488 interface.
3. An instrument interface IEEE-488. This interface provides complete automated measurement and control capabilities of IEEE-type instruments such as DVMs.
4. Output devices. The PC can offer a wide choice of output devices from many differ-

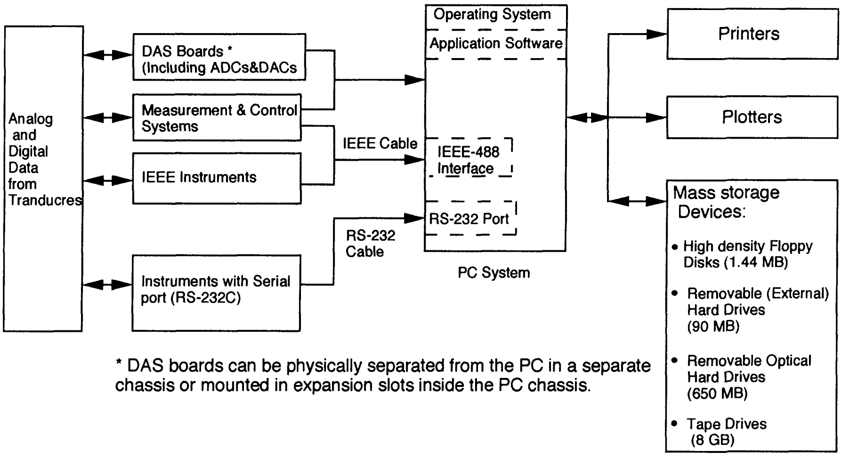
Analog and Digital Data from Tranducres
DAS Boards * (Including ADCs&DACs
Measurement & Control Systems
IEEE Instruments
Instruments with Serial port (RS-232C)
IEEE Cable
RS-232 Cable
Operating System
Application Software
IEEE-488 Interface
RS-232 Port
PC System
Printers
Plotters
Mass storage Devices:
• High density Floppy Disks (1.44 MB)
• Removable (External) Hard Drives (90 MB)
• Removable Optical Hard Drives (650 MB)
• Tape Drives (8 GB)

* DAS boards can be physically separated from the PC in a separate chassis or mounted in expansion slots inside the PC chassis.

FIG. 24. Personal-computer-based data acquisition system structure.

ent manufacturers. In addition to the PC's own CRT, external devices such as printers, plotters, and many mass storage devices with various capacities are available from numerous manufacturing sources.

5. Application software. A wide range of software has been developed specifically for PC-based DASs for both the IBM PC and the Apple Macintosh at reasonable cost. Plug-in boards also are available that include the necessary hardware and resident software that can turn the PC into an oscilloscope, a DVM, a signal generator, a logic analyzer, etc. These simulated instruments, known as virtual instruments (VIs), can be controlled from the PC keyboard, which emulates the control knobs and buttons used with the actual instruments by means of software modules known as instrument drivers. These instrument drivers are essential building blocks in the construction of VIs. The drivers handle the details of controlling and communicating with specific instruments. With the instrument drivers, the user can create VIs by combining several instruments with data analysis, data presentation, and user interface code. A multitasking software development system for DOS called LabWindows is becoming the industry standard. LabWindows primarily serves three purposes. It provides an integrated development environment, a high-level interface to the instrumentation hardware, and the user interface to the virtual instruments. LabWindows also has extensive libraries of functions for data acquisition, data analysis, and data presentation along with tools for editing and debugging user programs written in BASIC or C (discussed in the following). PC-based DASs offer a high sampling rate [100 000 samples/s (i.e., 10-μs sample period)].

2.2.1 Software. Numerous programming languages have already been developed and exist in a variety of dialects (see COMPUTER PROGRAMMING LANGUAGES). Some of these languages were developed specifically for scientific applications such as those performed in the laboratory from customized PC-based DASs to data analysis routines.

BASIC (Beginner's All-purpose Symbolic Instruction Code) is an interpreting, high-level, unstructured, interactive, easy to learn and use language. It became popular among programmers because of its English-like commands. The difficulty of the original version of BASIC is the global internal structure, which does not lend itself to reusing or sharing parts of the program. Many PCs include improved versions of BASIC (e.g., Microsoft BASIC) by incorporating the BASIC interpreter in ROM, which is also available on disk. Another version of BASIC known as *QuickBASIC* was developed and became very popular for developing specialized DASs. Although QuickBASIC is an interpreting language, it is still considered fast enough in comparison with a compiled language such as FORTRAN.

Recently, a programming language called Visual Basic, for Windows, was introduced, primarily to improve users' interface development of virtual instrumentation as well as of general-purpose languages.

C is a block-structured, general-purpose programming language developed by Bell Laboratory. It has been closely associated with the UNIX system because it was developed on that system; also, UNIX and its software are written in C. However, the language is not tied to any one operating system or machine. Because of its portability, C can be used to compose programs that can run on a host of microcomputers with minimal alteration. C is quickly becoming the language of choice among professional application and system developers for the PC, other microbased systems, minicomputers, and mainframe computers. C can be used with LabWindows (for DOS) to develop customized PC-based ADSs. Recently, a programming language called C++ was introduced to improve users' ability to interface with virtual instrumentation as well as handling interfacing requirements of general-purpose languages.

Assembly generates the tightest code of any language and is considered to be the fastest to execute except for machine language. It is used to design programs such as the well-known spreadsheet software, Lotus 1-2-3. Assembly programs are generated from individual instructions represented by three- or four-letter mnemonics like MOV, JMP, and PUSH. When assembled, a program translates directly into machine code that the processor can read from memory and execute. Although the language is somewhat difficult to learn and use, once the programming is complete the results are hard to match.

2.3 Instrument Interfaces

Transmitting and capturing data among instruments in a measurement system requires some physical links or interface to integrate instruments from different manufacturers in a system. Also, the increased application of computers in test, measurement, and control in the instrumentation field requires some form of communication between instruments and computers. Several interface standards adopted by industry are discussed below (see also BUS SYSTEMS AND COMPUTER INTERFACES).

1. IEEE-488/GPIB interface: a standard communication bus adopted by the Institute of Electrical and Electronics Engineers (IEEE) in 1978. This interface was initially developed in 1972 by Hewlett-Packard (HP) as a computer–to–instrument-interface bus (HP-IB) to standardize interface configuration and operating protocol for all HP equipment supplied with digital control interfaces. The HP-IB was adopted by IEEE in 1975 and became IEEE Standard 488-1975 (Digital Interface for Programmable Instrumentation) and was later reissued as IEEE 488-1978 and 488-1987. The standard is also known as the general-purpose interface bus (GPIB). It defines the physical and data-link structure of a local area network for digital communications between a main controller (e.g., computer) and various pieces of instrumentation and laboratory equipment. Instruments (e.g., scopes, analyzers) equipped with this standard can be controlled remotely as part of automated test and measurement systems. The interface bus is available as a module (printed circuit board) containing all the necessary hardware and firmware to implement the specified communication protocol. The IEEE board includes a standard connector for interfacing with the IEEE instruments through a standard cable. The basic bus structure as it interfaces with the instruments is shown in Fig. 25. The diagram shows a typical computer-controlled measurement system in which the various instruments are connected by the IEEE-488 bus. The diagram clearly shows that all the instruments are connected in parallel to the 16 different bus lines. The various instruments can perform one or more of the following functions: (a)Talker, which can transfer data to other instruments by the data bus; (b) Listener, which can receive data from other instruments by the data bus; (c) Controller, which manages the operation of the bus system, primarily by sending addresses and commands over the data bus and sending control signals over the control bus. Only one Talker is allowed to put data on the data bus at a given time, but more than one Listener can receive data simultaneously. The Controller sends a Talk address out over the data bus while activating the appropriate control line. The instrument with that particular address is the only Talker that is activated, and all other Talkers are disabled. A similar procedure selects the one or more Listeners that can receive data from the active Talker.
2. VXIbus interface. The VXIbus interface is a modular instrument standard developed in 1987 to define a bus structure architecture that allows a wide range of instruments, interface cards, or computers from different manufacturers to coexist as modules in the same enclosure (cardcage). The VXIbus specification describes in detail the technical requirements of VXIbus-compatible components such as mainframes, backplanes, power supplies, and functional plug-in instrument and communication interface modules. The VXIbus standard is used in custom building powerful microcomputer-based DASs with distributed intelligence and extremely high data-transfer rates. The VXIbus instrument standard is an extension to the VMEbus that is open to all manufacturers and is compatible with present industry standards. The VMEbus architecture was developed primarily with focus on computer systems and with limited instrumentation capabilities. A new open architecture instrumentation bus, based on VMEbus, the Eurostandard (printed circuit board standard IEC 297-3), and other instrumentation standards such as IEEE-488.2, was created and became known as the VXIbus. A typical VXIbus (abbreviation for VMEbus eXtension for Instrumentation) system consists of a single frame [48.3-cm (19-in.) rack] with up to 13 modules plugged into numbered slots. A common configuration places into slot 0 a system resource module such as the VXIbus-

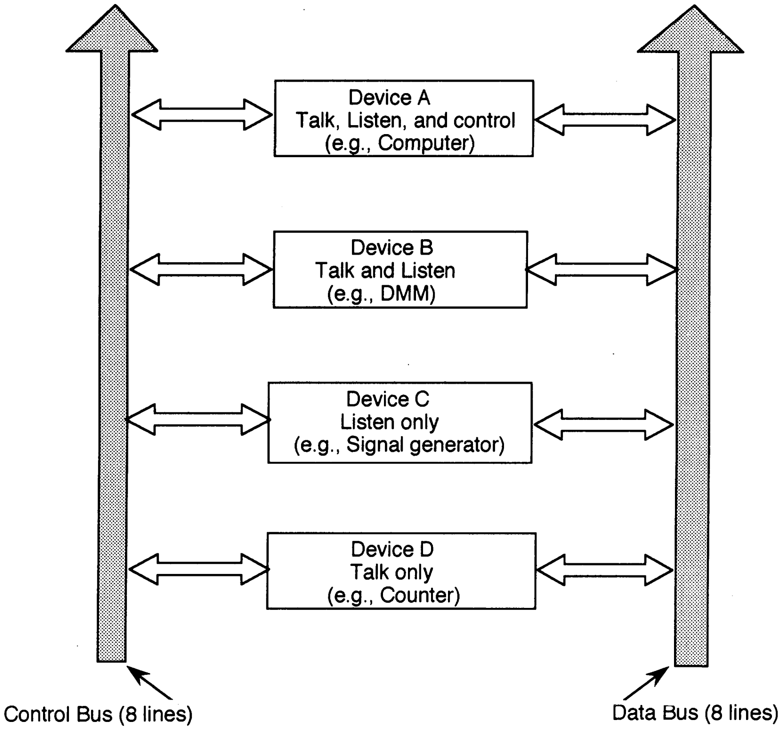
Device A
Talk, Listen, and control
(e.g., Computer)
Device B
Talk and Listen
(e.g., DMM)
Device C
Listen only
(e.g., Signal generator)
Device D
Talk only
(e.g., Counter)
Control Bus (8 lines)
Data Bus (8 lines)

FIG. 25. Typical IEEE-488 bus system.

mandated timing generation or communication port such as IEEE-488 or RS-232. The other positions are general-purpose slots for the user to mix and match. Figure 26 shows some of the possible system configurations and illustrates the flexibility that is afforded by the VXIbus standard. [For more details, see *The System Engineer's Handbook,* a guide to building VMEbus and VXIbus systems (Black, 1992).]

3. RS-232C interface. The RS-232 serial (asynchronous) communication standard was developed initially by the Electronic Industries Association (EIA) to standardize the interface between modems and data terminal equipment (e.g., computers and termi-

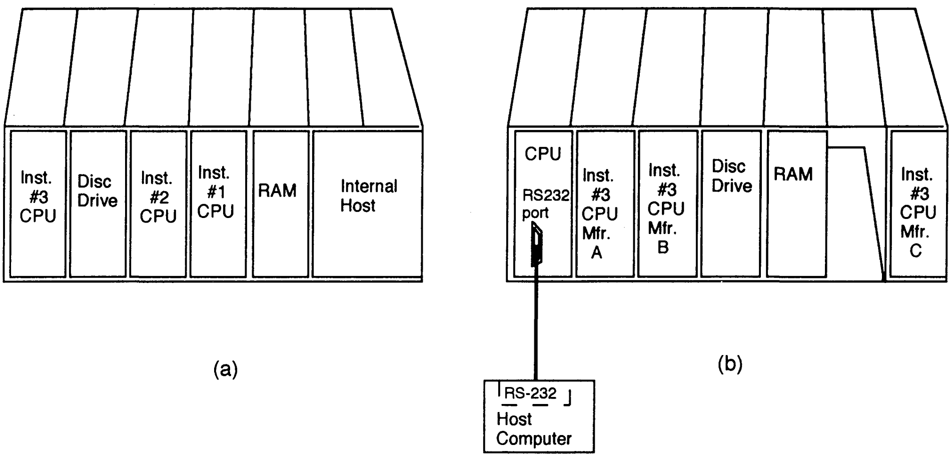
Inst. #3 CPU
Disc Drive
Inst. #2 CPU
Inst. #1 CPU
RAM
Internal Host
(a)
CPU
RS232 port
Inst. #3 CPU Mfr. A
Inst. #3 CPU Mfr. B
Disc Drive
RAM
Inst. #3 CPU Mfr. C
(b)
RS-232
Host Computer

FIG. 26. Typical VXIbus system configurations: (a) stand-alone system; (b) multiple-CPM system.

nals) connected to them. The latest version, RS-232C, was adopted in 1969 and has become the serial interface standard when one microprocessor system communicates with another microprocessor system over long distances or with peripheral devices such as printers or CRT terminals. Serial transmission requires only two wires to carry all the necessary data, address, and control information. The information is transmitted one bit at a time, making serial transmission slower than parallel transmission. The RS-232C standard is typically used for baud rates up to 20 kb/s using signal lines up to 15 m (50 ft). For higher transmission rates over longer distances, EIA has adopted other standards such as the RS-422, RS-423, and RS-485, listed in Table 4 along with their electrical characteristics. Almost all modern instrumentation manufacturers offer serial interface as an option for their products to communicate with computers and other devices. An interface can be purchased as a card (printed circuit board), commonly known as a serial port, with all the necessary hardware and software (serial interface protocol). It is usually inserted into an open slot in the instrument or in the computer. The connection among devices equipped with RS-232C ports is through standard 25-wire cable that connects directly into a 25-pin connector mounted on the card.

4. CAMAC interface. The CAMAC is a modular instrumentation and digital interface system developed to address instrument incompatibility that results from employing instrumentation from various manufacturers in a given measurement system both in the laboratory and in industrial fields. An initialism for Computer-Automated Measurement And Control, the CAMAC is an ANSI/IEEE standard (Standard 583-1982) that is concerned with mechanical, electrical, and signal interfaces. As shown in Fig. 27, measurement devices are in modular units and are housed in a standard card enclosure (cage or crate) providing a common dc power required by these devices to function. Data input and output from the modules are handled by a specified bus called a data highway (Dataway) at the rear of the cage. The system initially has a 1-μs Dataway cycle time and a high address capability. The Dataway is a parallel bus with 24 read lines and 24 write lines so that words of 24 bits or less can be handled in a single data transfer. Additional levels of compatibility are achieved through the use of standardized parallel and serial highways that have been developed for use with the basic CAMAC system. The serial highway can be used to interconnect up to 62

Table 4. Communication Electronic Industries Association standards comparison.

	Interface			
Mode of operation	RS-232C Single-ended	RS-422 Single-ended	RS-423A Differential	RS-485 Differential
Number of drivers and receivers allowed on-line	1 driver, 1 receiver	1 driver, 10 receivers	1 driver, 10 receivers	32 drivers, 32 receivers
Maximum cable length [m (ft)]	15 (50)	1200 (4000)	1200 (4000)	1200 (4000)
Maximum common-mode voltage	±25 V	±6 V −0.25 V	±6 V	±12 V −7 V
Driver output	±5 V min ±15 V max	±2 V min	±3.6 V min ±6.0 V max	±1.5 V min
Driver load	3 to 7 KΩ	100 Ω min	450 Ω min	60 Ω min
Driver slew rate	30 V/μs max	NA[a]	Externally controlled	NA[a]
Driver output short-circuit current limit	500 mA to V_{cc} or ground	150 mA to ground	150 mA to ground	150 mA to ground
Receiver input resistance	3 or 7 KΩ	4 KΩ	4 KΩ	12 KΩ
Receiver sensitivity	±3 V	±200 mV	±200 mV	±200 mV

---

[a] NA=Not applicable.

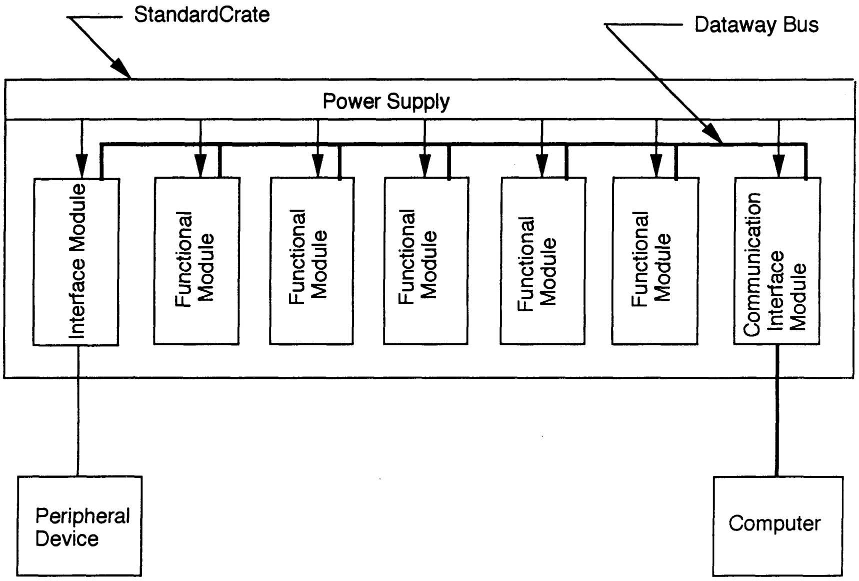

FIG. 27. Basic computer automated measurement and control (CAMAC) standard crate shown with a computer as the control device. A self-contained CAMAC crate system can be used with a CAMAC-type crate controller in place of the communication module.

crates, providing a widely distributed system, and the parallel highway can handle up to 7 crates. (See CAMAC Instrumentation and Interface Standards, IEEE 1982, or latest version for more details.)

3. MEASUREMENT INTERFACE

Every measurement application consists of a number of components that together form a system used to perform a measurement and processes the result. The application could be data storage, recording, display, controlling functions, or analysis. The components within the measurement system are most likely to be a mix of analog and digital devices. Such a mix in interdomain devices gives rise to two important factors that must be considered and well understood for the measurement system to perform its intended function in the most satisfactory manner. These factors are data conversion from analog to digital and vice versa and signal transmission within the system.

The coverage in this section of interdomain conversion and signal transmission is intended to be brief and to serve as an introduction of this subject to the reader. A complete and more comprehensive coverage is provided in ELECTRICAL INSTRUMENTATION.

3.1 Digital-to-Analog Converters

The digital-to-analog converter (DAC), also known as the D/A converter, is an electronic device that accepts digital information and transforms it into an analog voltage. The digital information is in the form of a binary number with some fixed number of digits. In the digital field, the binary number is called a binary word, and the digits are called bits. For example, the 8-bit word 10101011_2 would be a binary number having eight digits consisting of 0s and 1s.

The DAC converts the digital word into an analog voltage by scaling the analog output to be zero when all bits are 0 and some maximum value when all bits are 1, as illustrated by the simplified functional diagram of Fig. 28. This function can be mathematically represented by expressing the binary number as a fractional number according to the relationship

$$V_x \simeq V_r(a_1 2^{-1} + a_2 2^{-2} + \ldots + a_n 2^{-n}), \quad (14)$$

where V_x is the analog voltage output, V_r is some voltage reference in the DAC internal

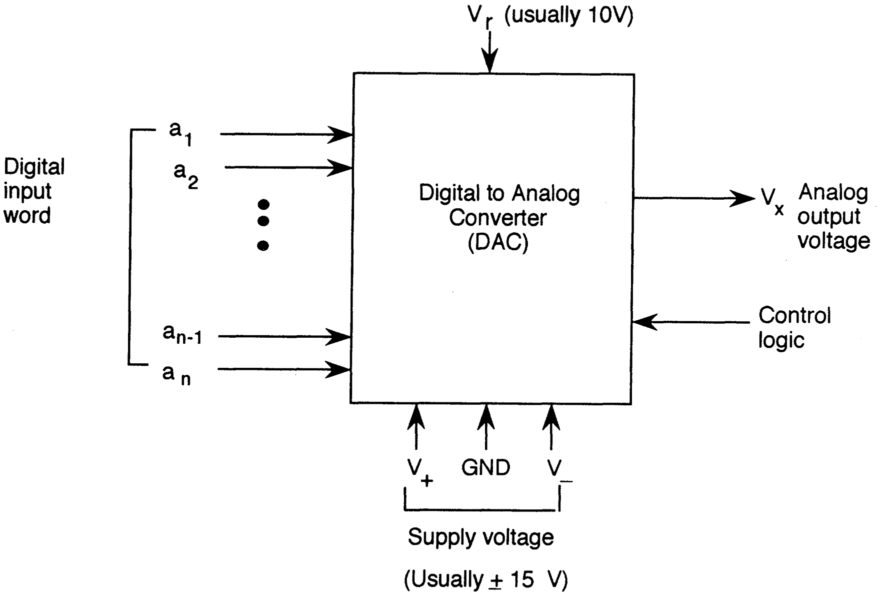

FIG. 28. Input/output diagram of an *n*-bit digital-to-analog converter.

circuit, and the coefficients $a_1,...,a_n$ represent the n-bit binary word. For a 10-bit DAC ($n=10$) with a 10-V reference, the output can have any of 1024 evenly spaced voltages, including 0 and 10.0 V, thus offering a resolution (1/1024) of slightly better than 0.1%.

DACs can be used in various applications other than for data conversion from digital to analog form. A DAC, for example, when driven by a binary counter, can produce a *staircase.* This staircase is useful in a variety of experiments. One form of DAC (based on a $R/2R$ resistor ladder network structure), when controlled with a BCD-coded input, can also be used as a digitally controlled voltage divider with high resolution. One such device is designed to provide an output between 0 and 1.999 times the input voltage, with a resolution of 0.001.

3.2 Analog-to-Digital Converters

The analog-to-digital converter (ADC), also known as the A/D converter, is an electronic circuit that accepts an analog voltage signal and transforms it into an encoded digital word, as illustrated by the simplified function diagram of Fig. 29. The relationship between the analog input voltage and the digital output sequence of the ADC is the same as that expressed by Eq. (14), where V_x is the analog input voltage, V_r is a reference voltage, and the coefficients $a_1,...,a_n$ represent the n-bit digital output (digital word). It is clear from Eq. (14) that the voltage, as expressed by the right-hand side, can change only by a finite step size given by

$$\Delta V = V_r 2^{-n}. \qquad (15)$$

The reference voltage V_r for the ADC also represents the full-scale value of the input voltage that can be applied to the converter. It is for that reason that Eq. (15) often appears in publications as

$$\Delta V = V_{\mathrm{fs}} 2^{-n}. \qquad (16)$$

The step size ΔV thus constitutes the smallest change in the analog input voltage that the ADC can resolve. Because of this resolution, the approximate equality in Eq. (14) is used rather than the equality sign. It is also clear that the resolution increases with an increase in the number of bits of the digital output. Historically, the first generation of ADCs was developed with 4-bit digital output. With the increased demand for higher performance, ADCs with 8, 10, 12, 14, and 16 bits offering superior resolution capabilities were developed.

The ADC sampling of the analog signal for conversion gives rise to two important considerations that could adversely affect its perfor-

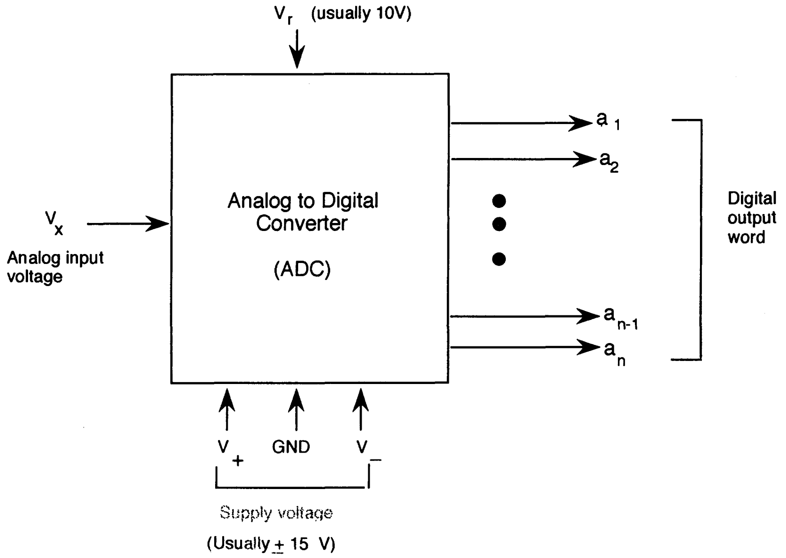

FIG. 29. Input/output diagram of an *n*-bit analog-to-digital converter.

mance if not properly addressed. The first is the sampling frequency, which must be at least twice the greatest frequency component of the signal to avoid the aliasing problem discussed in Sec. 1.2.3. Sampling the analog signal at a frequency less than that would lead to erroneous results when the sampled values are reconstructed. The second consideration concerns the finite time required for the ADC to complete its conversion. This time is known as aperture time, which is defined as the actual time the ADC requires to obtain an accurate reading of the analog signal. The aperture time must not be greater than the period of the greatest frequency component of the signal multiplied by the resolution. If the ADC is not fast enough to make a conversion within the aperture time for the given signal, then a circuit known as a sample-and-hold should be used. The sample-and-hold circuit, shown in its simplest form by Fig. 30, samples the signal by closing switch S1 long enough to charge capacitor C to the present signal value and hold that value for the time required by the ADC to make its conversion.

Many design methods exist for performing the ADC function. These methods vary in their techniques for realizing improvement in performance associated mainly with conversion speed and noise rejection. These designs resulted in the development of the following ADCs, which are listed for reference only because they are discussed in more details in ELECTRICAL INSTRUMENTATION:

1. successive approximation,
2. single-slope integrating,
3. dual-slope integrating,
4. servo,
5. ramp,

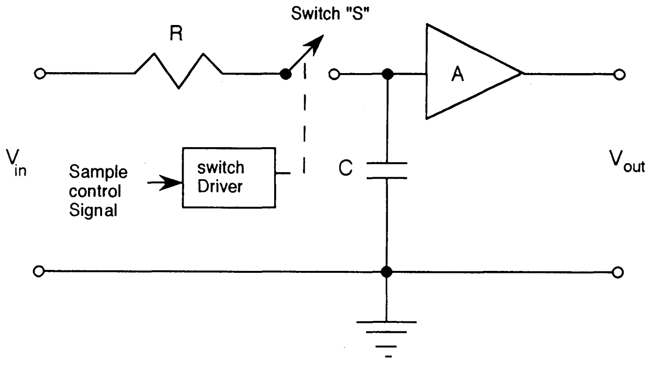

FIG. 30. Basic sample-and-hold circuit. The input is sampled periodically by closing switch S until capacitance C is charged to the new input value. The switch is then opened to hold the value to be delivered to the analog-to-digital converter through high-input-impedance amplifier A.

6. voltage-to-frequency,
7. parallel or flash,
8. cyclic, and
9. sigma-delta.

Until recently, the successive approximation and the dual-slope integrating-type ADCs have dominated because of their compromise in performance versus cost benefits. The dual-slope integrating-type ADC is not as fast as some of the others, but its inherent excellent noise immunity as well as its cost effectiveness due to design simplicity makes it ideal for many applications.

The most recent development in data-conversion technology is the sigma-delta ADC. It is considered state of the art and uses the latest development in digital technology to offer simplicity combined with powerful performance and resolution of up to 20 bits and high SNR.

3.3 Signal Transmission

Measurement starts with a transducer to quantify the physical measurement in terms of an electrical signal. Most transducers have analog outputs that need to be transmitted to other instrumentation for display, recording, analysis, or other purposes. It is crucial that the transmitting medium preserves signal integrity between the transmitting and the receiving instruments. Noise corruption from surrounding sources [e.g., electromagmetic interference (EMI)] and transmitting medium characteristics are two limitations that must be carefully considered.

The transmitting media can be classified in three categories: analog, digital, and optical. In industrial applications, the instruments within the same loop are separated by large distances, and the application environment is more noisy than the laboratory environment. Digital transmission methods [e.g., voltage-to-frequency (V/F), RS-232C, RS-422, RS-423, RS-425] offer a practical solution in industrial applications because of the inherent noise immunity and protection against signal degradation due to voltage drops along the transmitting media (wires). Analog transmission using a 4- to 20-mA current signal is also used for limited distances and when practical.

In laboratory applications, analog transmission of signals is the most practical and thus most commonly used method. Among the various signals used in analog data transmission (10–50 mA, 0–20 mA, 4–20 mA, 0–5 V, 0–10 V), the 4- to 20-mA current has been widely accepted as the industry standard signal because of the advantages that it has to offer. The 4- to 20-mA current is usually scaled to represent the signal range where the 4-mA current corresponds to the lowest range value, and the 20-mA current corresponds to the maximum range value. The 4-mA current, compared with the 0-mA current, value is recognized as the "live zero" because of the distinction it provides between loss of signal (electrical zero) due to physical interruption in the transmission medium and signal zero (representing the minimum range value). In addition to this advantage, current signals in general are less susceptible to noise pickup than voltage signals are, and degradation in signal due to voltage drops does not apply. The transmitting medium is typically a twisted pair of wires to reduce noise corruption further. Figure 31 illustrates a typical transmission configuration known as two-wire transmission. In this configuration, the 4- to 20-mA current is transmitted around a loop, allowing

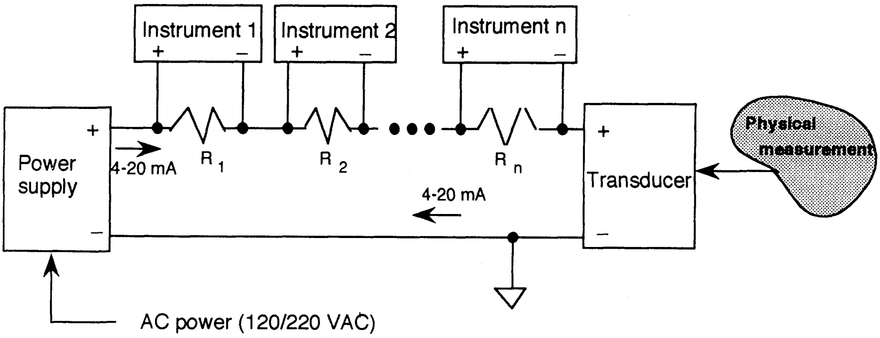

FIG. 31. Two-wire signal transmission configuration.

several instruments to share the same signal simultaneously. The signal is introduced to each instrument through an individual resistance R. The total number of instruments that can share the signal within the loop is limited by the sum of resistances around the loop. The maximum value of the total resistance in ohms is calculated by

$$R(\max)=R_1+R_2+\ldots+R_n$$
$$=(V_{\text{supply}}-12\text{ V})/(20\text{ mA}). \quad (17)$$

The transducer in the two-wire configuration is self-powered since its connection with the power supply serves to provide the dc power to and carry the signal from the transducer at the same time.

4. CONCLUSION

As was stated in the Introduction, this article is intended to acquaint readers within the scientific community with the kind of instruments and related technologies that are available to carry out scientific endeavors. In view of the rapidly changing technologies associated with the instrumentation industry, it is difficult to predict what will happen over the next few years. However, in observing the trend in the research and development in the electronics industry including the digital field and the their explosive markets, we can see that the basic ideas and theories will probably remain the same while details on device or instrument implementation will change. For example, in the area of instrument interface, our discussion of the VXIbus noted that the idea was not totally new, but it was an improvement over the VMEbus developed earlier. Another example is the future development in laboratory automation. PCs (such as the i386- and i486-based IBM-compatible PCs and the recently announced Pentium and the Macintosh) have invaded the workplace. Future development in this area is toward improving the processing power, networking [such as Ethernet and local area network (LAN)] capabilities, and data storage capacity. Workstations with 10^9 floating-point operations per second (Gflops) have evolved as a result of reduced instruction-set computer (RISC) processors.

GLOSSARY

ADC (Analog-to-Digital Converter): A device that changes analog voltage or current signal into a digitized number that can be digitally displayed or further analyzed by digital instruments.

Algorithm: A step-by-step procedure for solving a problem or performing a computation.

Analog device: A device that produces or processes analog signals.

Analog signal: A signal that varies continuously and can manage an infinity of values in any given range.

Aperture time: The time window, or time uncertainty, in making a measurement. For an analog-to-digital converter, it is the conversion time; for a sample-and-hold circuit, it is the signal averaging time during the sample-to-hold transition.

BCD (Binary Coded Decimal): A binary code used to represent decimal numbers in which each digit from 0 to 9 is represented by four bits weighted 8-4-2-1. Only 10 of the 16 possible states that are represented by the four bits are used.

Binary: A number representation system with a radix of 2.

Binary digit: In binary notation, either of the digits 0 or 1.

Bit: An abbreviation of "binary digit."

Bus: A group of wires used as a common path connecting various components within a microprocessor system (or computer) for transferring data by using control signals–for example, for interfacing the central processing unit of a digital computer to memory and inputs/outputs.

Byte: A group of adjacent binary digits operated on by the computer as a unit. The most common byte size is eight binary digits.

CRT (Cathode-Ray Tube): An electronic vacuum tube with a screen for graphical or alphanumeric visual display of information. Display is produced by means of proportionally deflected electron beams.

DAC (Digital-to-Analog Converter): A device that converts a digital signal, often from a computer, into a proportional analog signal (voltage or current).

Data Conversion: The process of changing the form of data representation. For example, converting analog signal to digital form and vice versa.

dB (Decibel): A logarithmic unit that expresses the magnitude of change in the level of a signal. The gain or attenuation of a device is commonly expressed in decibels defined in terms of output/input ratio as voltage (V): $dB_V = 20 \log V_{out}/V_{in}$; current ($I$): $dB_I = 20 \log I_{out}/I_{in}$; power ($P$): $dB_P = 10 \log P_{out}/P_{in}$. The decibel scale has the advantage of using small numbers to represent large ratios. Consecutive stages of amplification or attenuation lead to a sum (rather than a product) of individual gains.

dBm: A logarithmic measure in decibels of absolute signal level above or below some reference. For power measurement, m signifies a reference power of 1 mW. For example, 20 dBm is 100 mW. For other measurements (e.g., volts), the reference could be 1 V (dBV) or 1 mV (dBmV).

Digital Device: A device that produces or processes digital signals.

Digital Signal: A signal that varies in discrete steps and can manage only a finite number of different values in any given range.

EPROM (Erasable Programmable Read-Only Memory): A solid-state memory device that can be erased with ultraviolet light and reprogrammed by using equipment specialized for this purpose.

GPIB (General-Purpose Interface Bus): The IEEE-488 bus (HP-IB) designed by Hewlett-Packard in 1965 and adopted by the Institute of Electrical and Electronic Engineers in 1974.

LCD: Liquid crystal display.

LED: Light-emitting diode.

LSB (Least-Significant Bit): The rightmost binary digit in a digital word.

MSB (Most-Significant Bit): The leftmost binary digit in a digital word.

Peripherals: Auxiliary equipment that can be attached to and controlled by a central processing unit or a computer (e.g., CRTs, printers, magnetic tape transports).

Port: An electronic device that allows a computer to communicate with other external devices having communication capabilities.

Protocol: A formalized set of conventions including wiring arrangements necessary to establish and maintain communication between two devices.

RAM (Random Access Memory): This term, through common usage, refers to high-speed, volatile, solid-state (semiconductor) memory used to store and recall information. This type of memory can be read from or written to.

Refresh: Update information on the cathode-ray tube to maintain data on the screen.

Resolution: The magnitude of step changes distinguished by an instrument or device as the measured signal is continuously varied over the instrument range.

ROM (Read-Only Memory): A memory used for permanently stored computer programs and data.

Sampling Rate: The number of analog-to-digital conversions per second that are actually performed.

Sampling Theorem: A theorem by Nyquist that states that if a continuous bandwidth-limited signal contains no frequency components higher than clock frequency (f_c), then the original can be recovered without distortion if it is sampled at a rate of at least $2f_c$ samples per second.

Serial Communication: A form of communication that allows only one bit at a time to be sent or received between two communicating devices.

Serial Port: An electronic device that sends and receives data one bit at a time (also known as asynchronous communication port).

Spectrum: The spectrum of an analog signal is the amplitude versus frequency for all components within the signal.

Transducer: A primary device that converts physical measurements into low-level electrical signals.

Transmitter: A primary device that translates physical measurements into electrical signals, usually 4 to 20 mA.

Transmitter (Two-Wire): A device that is used to transmit data by a two-wire current loop. The loop usually has an external direct current power supply, and the transmitter acts as a variable resistor with respect to its input signal.

Volatility: An inability of memory to retain stored data during and after a power failure. An example of volatile memory is random access memory.

List of Works Cited

Black, J. (1992), *System Engineer's Handbook,* New York: Academic.

Brigham, O. E. (1974), *The Fast Fourier Transform,* Englewood Cliffs, NJ: Prentice-Hall.

Doebelin, E. O. (1983), *Measurement Systems: Application and Design,* New York: McGraw-Hill.

Helfrick, A. D., Cooper, W. D. (1990), *Modern Electronic Instrumentation and Measurement Techniques,* Englewood Cliffs, NJ: Prentice-Hall.

Hewlett-Packard (1988), *Fundamentals of Time-Domain Reflectometry,* Application Note No. 62.

Institute of Electrical and Electronics Engineers (1982), *CAMAC Instrumentation and Interface Standards.*

Kohavi, Z. (1978), *Switching and Finite Automata Theory,* 2nd ed., New York: McGraw-Hill.

Sydenham, P. H. (1982), *Handbook of Measurement Science,* New York: Wiley, Vol. 1.

Vassos, B. H., Ewing, G. W. (1985), *Analog and Digital Electronics for Scientists,* 3rd ed., NewYork: Wiley.

Further Reading

Bendat, J. S., Piersol, A. G. (1971), *Random Data: Analysis and Measurement Procedures,* New-York: Wiley.

Black, J. (1992), *System Engineer's Handbook,* New York: Academic.

Brigham, O. E. (1974), *The Fast Fourier Transform,* Englewood Cliffs, NJ: Prentice-Hall.

Kohavi, Z. (1978), *Switching and Finite Automata Theory,* 2nd ed., New York: McGraw-Hill.

Ouchi, G. I. (1987), *Personal Computers for Scientists,* Washington, DC: American Chemical Society.

Sydenham, P. H. (1982), *Handbook of Measurement Science,* New York: Wiley, Vol. 1.

VanDoren, A. H. (1982), *Data Acquisition Systems,* Reston, VA: Prentice-Hall.

LASERS, MEDICAL USE OF

See MEDICAL USE OF LASERS

LCDs

See DISPLAY TECHNOLOGY

LEDs

See LIGHT-EMITTING DIODES

ERRATUM

LABORATORY INSTRUMENTATION (Vol. 8, pp. 247–281). AYMAN A. SHOURBAJI.

Page 279. Column 1, line 20 from the bottom, the word "the" should be deleted.

LASER PHOTOCHEMISTRY

Paul Engelking, *Department of Chemistry, University of Oregon, Eugene, Oregon, U.S.A.*

Introduction 283
1. **Photophysics** 284
1.1 Laser Light 284
1.2 Interaction with Matter 285
1.2.1 Linear Absorption 285
1.2.2 Multiphoton Absorption 285
1.2.3 Level Shifts 286
1.2.4 Hole Burning 286
2. **Multiphoton Processes** 286
2.1 Electronic Multiphoton Absorption 286
2.1.1 Multiphoton Selection Rules 286
2.1.2 Multiphoton Ionization (MPI) 287
2.1.3 Multiphoton Dissociation (MPD) 288
2.2 Vibrational Multiphoton Absorption 289
2.2.1 Vibrational-State Ladders 289
2.2.2 Isolated Molecules 289
2.2.3 Effects of Collisions 290
3. **Energy Partitioning** 290
3.1 Randomization of Energy in Molecules 290
3.2 Energy Localization 291
4. **Chemical Changes** 292
4.1 Ionization and Detachment 292
4.2 Dissociation 292
4.3 Isomerization 294
4.4 Condensation and Addition 294
5. **Applications** 294
5.1 Holography 294
5.2 Isotope Selection 294
5.3 Surface Modification 295
5.3.1 Etching 295
5.3.2 Deposition 295
5.4 Information Storage 295
5.5 Chemical Analysis 295
5.6 Biochemistry 295
Glossary 296
Works Cited 297
Further Reading 298

INTRODUCTION

Photochemistry transforms matter with energy in the form of light; laser photochemistry transforms matter with light from lasers. While most chemistry emphasizes the production of bulk materials, much photochemistry aims at the production of specialty materials, the manufacture of structures, or the processing of materials to obtain, transfer, or store information. Laser photochemistry retains these further emphases.

Manmade photons are expensive chemical reagents; economics restricts photochemistry to high-value–added processes, or to where only a few photons selectively initiate a process. Photons are very effective at causing chemical changes, but current photon sources are inefficient. While a mole of visible photons contains energy valued at only a cent at today's electric rates, inefficiencies of conventional light sources can quickly push the cost of photons to over a dollar per mole. These economics generally make photochemistry unattractive for the production of bulk chemicals. Photochemistry is more useful for high-value products such as drugs (currently vitamin D is produced photochemically), insecticides, perfumes, and polymer resins. Established photochemical technologies are found in photography, coating (especially of thermoplastics, such as used in magnetic recording or adhesive tapes), photocuring of optical and dental cements, drug phototherapy, uv disinfection, sterilization, and surface cleaning, printing, lithography, and photoresist masking in semiconductor and printed-circuit manufacturing.

The capital costs associated with most laser sources—particularly those operating in the photochemical uv—can bring the cost of photons up to $1000 per mole. Production of exceptionally valuable materials, such as isotopes or isotopically labeled molecules, can

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

benefit from laser photochemistry. More commonly, laser photochemical processes transfer or store information within materials. Lasers find applications in photographic rendering of computer-drawn images, laser marking, and photoresist masking. In stereolithography, laser photochemistry is used to realize complex three-dimensional plastic prototypes or mold models directly from computer-aided designs. The latter complements automated machining: solid material adds to an object held on a mandrel wherever a laser beam initiates curing of a polymer resin.

Lasers have found extensive application in the investigation of chemical processes themselves, where they have given us invaluable information about the details of chemical reactions. Photoinitiation is much more selective than thermal initiation of reactions and provides greater control in the investigation of chemical kinetics (now often called "chemical dynamics" to emphasize the higher precision afforded to prepare and investigate reactions of single quantum states). Lasers have moved into nearly every chemical kineticist's laboratory—if not to initiate reactions, then to analyze products. A whole field of "half reactions" has developed, in which a reaction intermediate, normally prepared by a precursor reaction or collision process, is instead prepared directly by a photoprocess. This allows the experimenter to prepare an intermediate of known chemical conformation, having a given energy, at a specified time.

The availability of short-duration (picosecond and femtosecond) pulses has shortened the time scale directly accessible for kinetic experiments in the time domain (Mukamel, 1990; Khundkar and Zewail, 1990). For example, nearly a dozen transient molecular states and conformers (different, relatively stable arrangements of atoms) have been identified by pump-probe laser experiments subsequent to the absorption of light by rhodopsin, the optical pigment in the retina (Birge, 1990).

Fundamental reaction dynamics also can be investigated in the frequency domain. For example, photoemissions from the dissociating states of the molecules methyl iodide and ozone have been observed. The structure of each intermediate, changing as it dissociates, is impressed upon the spectra as a series of resonances, while the dynamics and time scales of the reaction appear in the pattern of intensities and spectral widths of these features (Imre *et al.*, 1984). A laser provides a high flux of photons of a single frequency, making such investigation of weak emissions possible.

Information from the analysis of materials is becoming more in demand, and laser techniques have become more popular there. Laser spectroscopy is useful, but often lasers are used to prepare the sample by a mixture of physical and photochemical effects. Solid material may be introduced into the gas phase by laser ablation, for example. Photoionization of a sample may prepare ions for identification by mass spectrometry, or photodissociation may prepare atoms for atomic spectroscopy.

In much of photochemistry brought about by lasers, the fundamental uniqueness of laser radiation has not yet been used; lasers often represent specialized "light bulbs" that, in principle, could be replaced by thermal light sources. As such, much of this laser photochemistry is not qualitatively different from photochemistry conducted with weaker light sources, the subject of the article on PHOTOCHEMISTRY. The present article concentrates on the unique properties of laser radiation and focuses on the qualitative differences in the photophysics, energy partitioning, and chemistry that can appear when laser radiation is used.

1. PHOTOPHYSICS

1.1 Laser Light

The radiation field within a laser must stimulate emission in competition with spontaneous emission. This establishes a characteristic brightness of laser radiation.

Spontaneous emission can be viewed as emission stimulated by the random, zero-point fluctuations of the radiation field. The spectral density of these spontaneous fluctuations approximates that of the density of states per wavelength interval, $4\lambda^{-4}$. In terms of photon flux, or brightness, this is

$$\frac{d(\text{photons})}{dA\, d\Omega\, d\lambda\, dt} = \frac{c}{3\pi\lambda^4}, \tag{1}$$

or, in terms of flux power density M, solid angle Ω, and wavelength spread $\Delta\lambda$ of a laser beam,

$$M/\Omega\Delta\lambda = hc^2/3\pi\lambda^4. \qquad (2)$$

Laser light differs from light from lower-brightness, thermal sources in one or more of these quantities:

- high power flux density,
- low angular spread,
- narrow frequency distribution.

These give laser light several advantages for photochemistry.

1.2 Interaction with Matter

1.2.1 Linear Absorption. Until intense laser sources of radiation became available, the fundamental photochemical reaction step, absorption, depended linearly upon intensity. At low power density, light interacts with matter by the absorption of one quantum of radiation at a time.

Even without entailing direct multiphoton processes, lasers may still significantly enhance rates of photoabsorption. At low light levels, using thermal light sources, the absorption of light is usually the rate-limiting step in a sequence of steps that compose the mechanism of a photochemical process. At high light levels, resonant absorption may become the fastest step in a photochemical mechanism.

While the line shapes of individual transitions and the spectral distribution of transition bands depend upon numerous details of the molecule and its interaction with the radiation field, the overall line strengths, or band intensities, can be related to simple perturbation expressions. The absorption cross section σ for transition to the final state $|f\rangle$, when integrated over frequency ν, is given theoretically by the leading, first-order perturbation of the initial state $|i\rangle$ in the electric dipole approximation,

$$\int \sigma_{fi}(\nu)d\nu = 4\pi^2\alpha\bar{\nu}\left|\left\langle f\middle|\sum_e \mathbf{r}_e\cdot\mathbf{n}\middle|i\right\rangle\right|^2. \qquad (3)$$

The sum is taken over all charges e at distances $\mathbf{r}_e$; α is the fine-structure constant, and $\mathbf{n}$ is a unit polarization vector. The bar over the frequency indicates averaging.

Each electron contributes to the total integrated absorption cross section, averaged over orientations and summed over possible final states, one "electron oscillator," $\pi r_0 c \approx 0.03$ cm^2 s^{-1}; r_0 is the classical electron radius. Individual electronic absorption bands typically contain the transition strength corresponding to 0.01–0.5 of an electron oscillator, while weaker vibrational transitions contain perhaps 10^{-6}–10^{-4} of an electron oscillator in each band.

1.2.2 Multiphoton Absorption. At the high power densities achievable with lasers (easily 10^{18} W/m^2 in focused pulses), light interacts with matter through the additional mechanism of direct multiphoton absorption.

The second-order perturbation correction gives the two-photon contribution to the absorption. An absorption cross section $\sigma^{(2)}N_1$ for a photon of frequency ν_2 is induced by an off-resonance monochromatic field of frequency ν_1 and photon flux N_1. When integrated over the frequency of the second photon, the second-order cross section can be related to the dipole matrix elements:

$$\int \sigma_{fi}^{(2)}(\nu_1,\nu_2)N_1(\nu_1)d\nu_2 = \frac{4\pi^2\alpha^2}{\nu_1\bar{\nu}_2}\left|\sum_m \frac{\nu_{fm}\nu_{mi}\langle r_{fm}\rangle\langle r_{mi}\rangle}{\nu_{mi}-\nu_1}\right|^2, \qquad (4)$$

where

$$\langle r_{fm}\rangle = \left\langle f\middle|\sum_e \mathbf{r}_e\cdot\mathbf{n}_2\middle|m\right\rangle,$$
$$\langle r_{mi}\rangle = \left\langle m\middle|\sum_e \mathbf{r}_e\cdot\mathbf{n}_1\middle|i\right\rangle, \qquad (5)$$

and the sum is taken over intermediate states $|m\rangle$ having frequencies ν_{mi} and ν_{fm} for transitions to the initial and final states. The energy for the overall transition comes from two photons; hence the two-photon resonance condition $\nu_{fi} = \nu_1 + \nu_2$. A special case often occurs when only one radiation frequency is used: then $\nu_1 = \nu_2$ and two-photon resonance is achieved when the energy of the transition corresponds to twice the frequency of the radiation field.

This integrated induced cross section is roughly of the order $\alpha(\pi c r_0)^2 N_1/(\bar{\nu}_1\nu_2)$. Making the two-photon absorption comparable to the one-photon absorption requires off-resonance fields on the order of

$$N_1 \approx \nu_1\nu_2/\alpha\pi c r_0 \qquad (6)$$

or about 10^{21} photons s^{-1} cm^{-2} (10^{17} W m^{-2} s^{-1} for typical green light).

Expressions for higher-order, multiphoton cross sections are readily derived. Typical dipole-allowed molecular multiphoton absorption cross sections have been found on the orders of 10^{-58} m^4 s for two photons and 10^{-94} m^6 s^2 for three photons.

1.2.3 Level Shifts. High radiation powers can cause atomic and molecular resonant frequencies to vary, responding to the ac Stark effect (Carmichael and Walls, 1976; Li *et al.*, 1985).

In atoms and diatomic molecules, effects of high power can be rather striking, even at small laser powers. In sparse spectra, associated with atoms and small molecules, each individual resonance has a relatively large oscillator strength. Even moderate fields, tuned near resonance, interact strongly with an atom, which undergoes rapid excitation to, and stimulated emission from, excited states at a characteristic, power-dependent rate, the Rabi frequency. The fluorescence from such an interacting system has a characteristic "head and shoulders" appearance, best understood as radiation at the fundamental frequency undergoing amplitude modulation at the Rabi frequency (Mallow, 1969). Such a spectrum may be "smeared" to higher or lower wavelength by an ac Stark effect (Li *et al.*, 1985).

Large molecules, in contrast to small molecules, spread the oscillator strength of an electronic or vibrational transition over a large number of molecular "rovibronic" states. Especially when a number of initial states are populated at finite temperatures, transitions are heterogeneously broadened. Each individual state-to-state transition may have only a small fraction of the original oscillator strength; at any given frequency, many resonances with widely varying transition strengths are likely to overlap. Thus, power broadening and ac Stark shifts require much higher powers (and are less observable) in the multiphoton photochemistry of large polyatomic molecules. However, power-driven state "pulling," field-induced state mixing, and spontaneous and stimulated Raman effects often explain the excitation to higher states that would otherwise appear to be inaccessible.

1.2.4 Hole Burning. At high power densities, a general excitation of rotational and vibrational levels accompanies excitation and deexcitation of states by the radiation field. Absorbing states that interact resonantly with radiation become depleted as molecules enter and remain in "dark" states having weaker interactions with the radiation field. These optical-pumping effects can make resonance features in an absorption spectrum disappear at high power levels (Rothberg *et al.*, 1981).

Radiation at a particular frequency generally moves population out of states that interact strongly with that radiation. This depletion may appear as a "hole" in the absorption spectrum. "Hole burning" is the basis for Doppler-free Lamb-dip spectroscopy. Atoms or molecules having little or no velocity component along the axis of two counterpropagating beams are temporarily depleted from the population distribution of velocities. At exactly the frequency center of an absorption line, the absorption cross section acquires a power-dependent decrease or "dip." This technique is commonly used for laser frequency stabilization, such as the iodine-locked He-Ne laser. Recently, interest has shifted to permanent hole burning as a method of information storage in materials.

2. MULTIPHOTON PROCESSES

The most striking examples of new photophysics and photochemistry that occur when lasers are used in place of conventional light sources involve multiphoton absorption.

2.1 Electronic Multiphoton Absorption

Electronic multiphoton absorption, excitation to electronic states energetically inaccessable by one-photon absorption, was one of the first effects explored with lasers. Two-photon absorption was first reported for inorganic crystals containing europium ions (Kaiser and Garrett, 1961). Blue and ultraviolet fluorescence appeared in the interaction of red, ruby-laser light with organic compounds (Peticolas *et al.*, 1963; Peticolas and Reickhoff, 1963; Singh and Stoicheff, 1963; Weisz *et al.*, 1964). Others observed two-photon absorption directly (Hopfield *et al.*, 1963; Staginnus *et al.*, 1968).

2.1.1 Multiphoton Selection Rules. Using the symmetry of the absorbing molecule, and the electronic-state symmetry assignments, McClain (1974) derived explicit selection rules for two-photon electric-dipole interactions. Most notable in two-photon excitations, even-parity transitions appear for centrosymmetric molecules, normally forbidden in single-photon electric-dipole transitions. Light polarization characteristics may also affect absorption, even for collections of randomly oriented molecules, especially those possessing high molecular symmetry.

2.1.2 Multiphoton Ionization (MPI). One striking multiphoton effect, runaway electrical breakdown of air, often occurs at the focus of high-power, pulsed laser light. Initiated by direct multiphoton absorption, the resulting molecular excited states undergo additional single-photon absorption directly into the ionization continuum. In collisions with molecular and ionic excited states, slow electrons efficiently extract energy: slow electrons "superelastically" scatter to become fast electrons; cross sections for this process diverge for low electron-impact velocities. Thus initiated, high-pressure ionization rapidly transforms the initially transparent gas into an optically thick plasma, in which most of the absorbing power is derived from electronic transitions originating on excited states of molecules and ions, and terminating on states consisting of additional free electrons and ions. Taken to its ultimate, a totally ionized plasma produced by laser radiation is now the object of laser fusion research.

Some researchers began to use ionization as a tool to investigate weak multiphoton absorption. Two groups almost simultaneously, Johnson *et al.* (1975) and Petty *et al.* (1975), devised a simple technique. A tunable laser was focused into a proportional-counter cell containing an inert gas and a trace amount of the molecule being studied (Fig. 1). For two-photon spectroscopy, when the energy of two photons was resonant with an allowed, two-photon transition, the energy of a third photon would be sufficient to ionize the resulting excited molecules, as represented in Fig. 2(a). Since ionization usually has a relatively unstructured absorption spectrum, and the power levels employed to produce two-photon absorption saturate the one-photon ionization step, the overall rate of ionization depends predominantly upon the rate of the initial two-photon step. Monitoring the amount of ionization produced in the proportional-counter cell as a function of wavelength generates a "two-photon" ionization spectrum. One of the first molecules investigated this

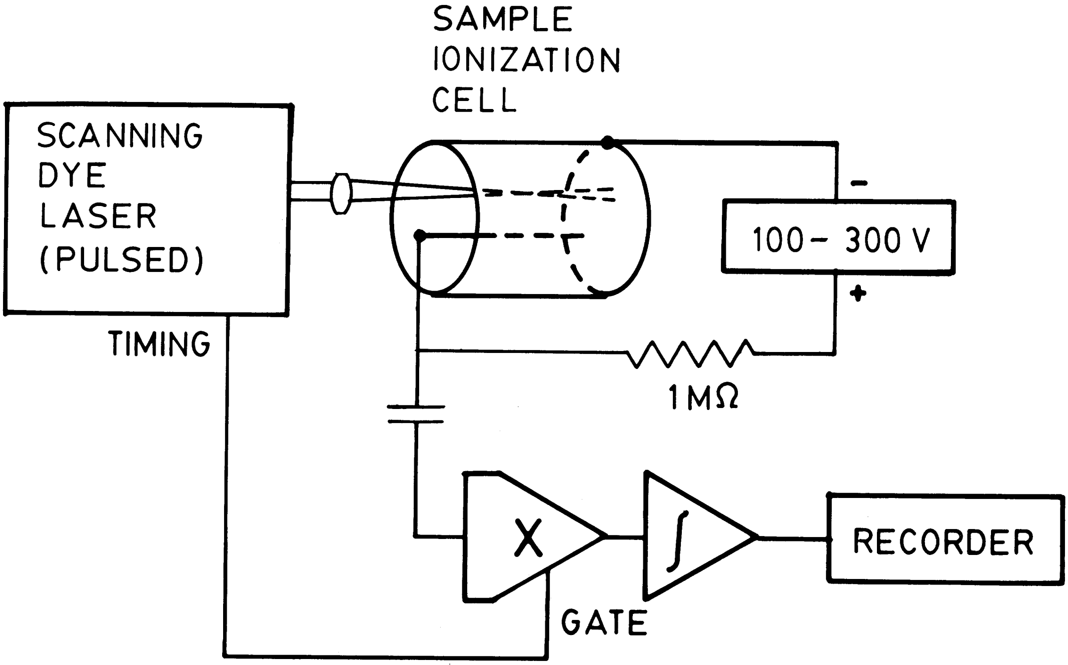

FIG. 1. Multiphoton ionization apparatus (schematic).

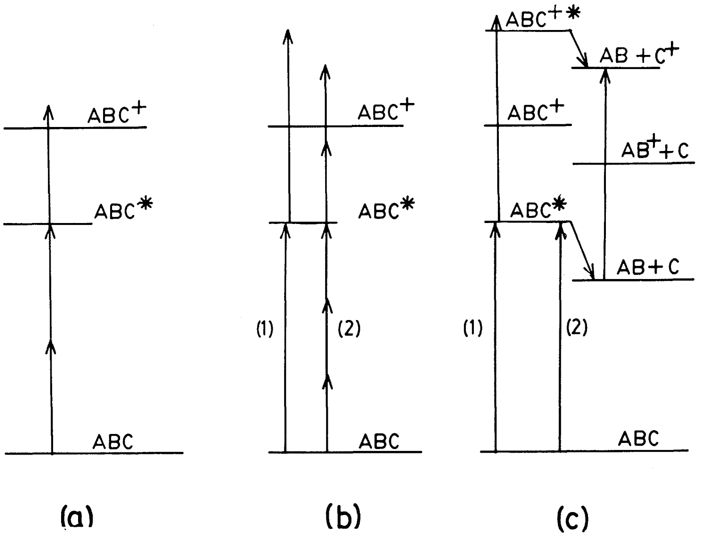

FIG. 2. Multiphoton ionization processes. (a) 2+1 multiphoton ionization occuring by two-photon absorption to a molecular excited state and subsequent one-photon ionization. (b) 1+1 and 3+2 multiphoton ionization processes. (c) Multiphoton ionization either (1) preceding dissociation of the ion, or (2) subsequent to dissociation of the neutral.

way was nitric oxide, NO; one of the first polyatomic molecules investigated was benzene, C_6H_6 (Johnson, 1975, 1976). Molecular states near the ionization limit (principally molecular Rydberg states) appear, states that are usually invisible because of the selection rules of one-photon spectroscopy.

In addition to this simple "2+1" process (two-photon absorption followed by one-photon ionization), numerous other *n*-photon combinations leading to absorption and ionization could, and did, occur, as exemplified in Fig. 2(b). Any of these processes are subsumed under the name "multiphoton ionization." Processes that depended upon resonant one-, two-, or many-photon intermediate processes were generally indicated by the additional term "resonant" or "selective."

2.1.3 Multiphoton Dissociation (MPD). With the recognition that intense visible or UV radiation rapidly excites molecules to high electronic states, culminating in ionization, a seductive argument can be made that saturated electronic absorption at high power densities would completely dominate "slower" processes involving nuclear displacements, such as molecular rearrangement and dissociation. Until the early 1980s, multiphoton dissociation was generally thought to compete ineffectively with multiphoton ionization; molecular fragmentation was thought to occur only after ionization of the parent molecule.

Studies of ions produced by multiphoton ionization of organic molecules suggested that rearrangement and dissociation could occur in the neutral molecule, before ionization, but poor understanding of the absorption spectra of the intermediate, neutral fragments did not allow definitive proof. Clear evidence that dissociation may precede one or more slow steps of absorption, and that it certainly precedes ionization, came from organometallic photochemistry. Multiphoton ionization of transition-metal compounds occurs by multiple molecular absorption events and rapid dissociation of the neutral molecule, followed by multiphoton ionization of naked, neutral transition-metal atoms (Gerrity *et al.*, 1980; Engelking, 1980; Leutwyler *et al.*, 1980).

In such cases, the MPI spectrum of a molecule actually expresses the atomic MPI spectrum of one or more of the molecule's constituent atoms, which have been ejected from the molecule by MPD. Subsequent work has shown the prevalence of MPD in the photochemistry of polyatomic organic molecules, such as benzene, where laser-induced fluorescence and resonance Raman spectroscopy show the presence of C_2, CH, and carbon atoms at high power levels.

Now we understand that the situation in a photochemical mechanism involving laser radiation is complicated by numerous competing pathways, and that photophysical and chemical steps can be intertwined, as represented in Fig. 2(c). In retrospect, the reason for this is clear: the electric fields within a molecule exceed those imposed by radiation, and any time an electron is displaced from the bonding regions between nuclei for any reason, nuclei are exposed to strong Coulomb repulsion. Two reviews in the same journal published within months of each other (Bernstein, 1982; Gedanken *et al.*, 1982) show the rapidity of this change of perspective.

2.2 Vibrational Multiphoton Absorption

2.2.1 Vibrational-State Ladders. Molecules couple to infrared (IR) radiation principally through the overall vibrational motion of molecular charge. Typically involving only a fraction of an electronic charge moving through only a fraction of a picometer, the resulting strength of a vibrational transition is commonly 10^{-4} of an electronic transition.

Low-energy vibrations of molecules usually take place in nearly harmonic potentials, containing "ladders" of nearly equally spaced vibrational energy levels. An IR laser frequency can almost coincide with a sequence of vibrational transitions corresponding to the intervals between levels of a particular normal mode, as shown in Fig. 3. At high laser powers, these nearly resonant intermediate states enhance multiphoton processes. Since vibrational spacing generally decreases with increasing mode energy, laser radiation that is resonant with molecular transitions at the bottom of a vibrational ladder becomes less resonant and couples less effectively with molecules already excited to higher energies, as shown in Fig. 3.

2.2.2 Isolated Molecules. The effectiveness of multiple resonances in assisting IR multiphoton transitions in isolated molecules depends to a large degree upon molecular size. For small molecules, such as diatomic molecules, with only one vibrational mode, anharmonicity quickly moves subsequent absorptions out of resonance. Except at extreme power densities, IR excitation of small molecules saturates lower transitions, but cannot give access to higher levels unless measures are taken to account for anharmonicity. In medium-sized molecules, it is easy to excite molecules rapidly to high vibrational levels, but difficult to drive the molecules to dissociation or rearrangement. In a process akin to hole burning, the energy ensemble of vibrating molecules acquires a bimodal, non-Boltzmann distribution: some low-energy molecules occupy vibrational and rotational states not effectively coupled to the radiation, while a high-energy group of molecules in a broader distribution of states interacts with the field. Bimodal population distributions are most often produced when molecular modes of vibration that occur at isolated frequencies are excited by IR laser radiation.

Large molecules, having a great number of vibrational modes, present many nearly coincident vibrations and overtones, even at moderate levels of excitation. Mixing of the numerous states reduces the interaction of any one state with the radiation field, giving rise to a dense cluster of states around each resonance. This is especially true at high energies. States that fill the gaps in a vibrational-state ladder in large molecules compensate for the anharmonicity that interrupts resonant multiphoton excitation to higher levels in small

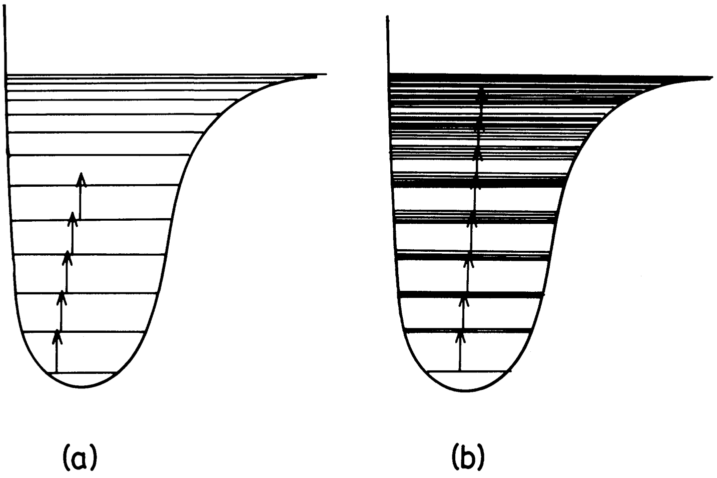

FIG. 3. IR multiphoton events with single-frequency radiation. (a) Resonant excitation drives small molecules up a ladder of nearly evenly spaced energy levels until anharmonicity moves states from further resonance. (b) Large molecules have high densities of states, providing near resonances all the way to dissociation.

molecules. With the increasing likelihood that each transition will be assisted by a resonance, excitation to dissociation becomes easier yet. Furthermore, with the dilution of resonances and the complication of the absorption spectrum, it is less likely that a narrow internal energy distribution will result, and the distribution of excitation in energy levels becomes more statistical, resembling a thermal Boltzmann distribution.

Vibrational transitions are weaker and more numerous than electronic transitions, making sequential resonant, or nearly resonant, transitions more important in the infrared and direct nonresonant multiphoton excitation less important in the infrared. Typically, one expects to need 10^{14} W m^{-2} or more to dissociate a diatomic molecule by near-resonant multiple transitions, and 10^{8}–10^{12} W m^{-2} for dissociation of a large molecule.

2.2.3 Effects of Collisions. If molecules are allowed to collide, the rates for IR ladder climbing increase. Colliding pairs of vibrationally excited molecules, which are no longer resonant with the radiation field, may "pool" energy into higher vibrational levels of one collision partner, leaving the other collision partner in lower vibrational levels. This brings the second molecule again in resonance with the exciting IR.

Collisional redistribution of vibrational energy is a statistical process: one expects to obtain a Boltzmann energy distribution. The persistence of a fraction of molecules in states that have transitions resonant with the radiation allows efficient heating. Power levels of 10^{4}–10^{7} W m^{-2} may lead to dissociation, rearrangement, or other thermally initiated chemistry.

IR heating of gases has proven to be effective in thermally initiating reactions for kinetic studies. This may be done either by directly coupling IR to the reactive molecules, or by indirectly heating the reactive molecules seeded in a bath of strongly absorbing molecules (Golden *et al.*, 1981).

3. ENERGY PARTITIONING

Chemists have, since the 1930s, known that certain bonds between atoms gave rise to characteristic frequencies in the IR, corresponding to bond stretches and bends. Thermal sources of radiation proved ineffective in breaking specific bonds, since energy could not be supplied fast enough to stretch a bond to breaking, in competition with intramolecular and intermolecular energy relaxation.

In recent years, the topic of mode-specific photochemistry has revived in the context of a better understanding of the energy redistribution that may occur inside molecules.

3.1 Randomization of Energy in Molecules

Energy flow in molecules falls between two extremes. In the "small-molecule" limit, radiation excites isolated states; the interaction with the radiation field is more important than the interaction between internal molecular states. In the "large-molecule" limit, radiation excites a dense band of internal states strongly mixed with the absorbing state; the interaction between numerous molecular states exceeds that with the radiation field. The difference is quite striking: small-molecule states typically fluoresce; large-molecule states typically retain the energy within the molecule.

The coupling between states within molecules is mediated by numerous mechanisms: vibrational mixing, Coriolis interactions, velocity and acceleration coupling of electronic and vibrational motion, and spin-orbit coupling.

The density of internal states is as important as the strength of interaction between them in determining the large-molecule limit. A single radiating molecular state interacting strongly with just a few molecular states produces simply a few, widely separated states. The large-molecule limit must be defined a bit more carefully by the two conditions

$$\Gamma \ll \gamma \tag{7}$$

and

$$\Delta E \ll \Gamma, \tag{8}$$

where Γ is the natural width of the radiative state, γ is the average strength of the interaction of the radiative state with other perturbing molecular states, and ΔE is the average spacing between perturbing states (the reciprocal of the density of states). The effect of various combinations of these parameters is outlined in Fig. 4.

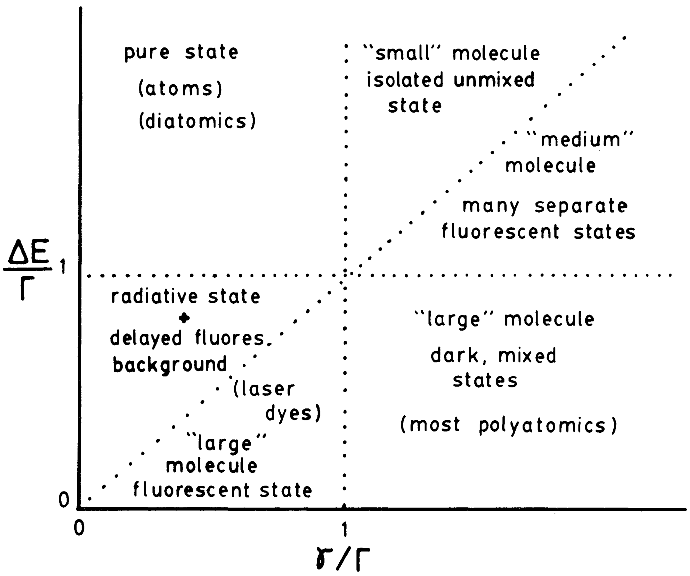

FIG. 4. Regimes of "large" and "small" molecule behavior. Here γ is the strength with which a radiative state interacts with other molecular nonradiative states; Γ is the strength of the radiative interaction; ΔE is the average spacing between mixing states.

Under the large-molecule conditions, where both relationships hold, the high density of perturbing states provides the molecule with its own thermodynamic bath. If a single radiative state couples with a large number of nonradiative bath states, the mixture of perturbed states will rapidly dephase into a mixture of incoherently radiating states, each with a small fraction of the original transition strength. It appears that the energy absorbed by the initial state has dissipated into the bath states.

The phenomenon of a high density of internal molecular vibrations acting as an intramolecular thermodynamic bath has been carefully investigated in the case of alkyl benzenes in which the length *n* of the "floppy" carbon chain attached to the rigid benzene ring is systematically increased:

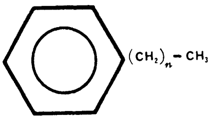

$(CH_2)_n-CH_3$

The electronic transition occurs primarily on the benzene ring, but couples with the molecular vibrations. As might be expected, the fluorescence quantum yields and the radiative lifetimes of molecules decrease as the chain length, and hence the density of vibrational states, increases (Smalley, 1982).

This ability of large molecules to transfer energy between electronic, vibrational, and spin-rotational manifolds of states has a profound effect upon the chemistry of molecules. In laser photochemistry, the rate at which this happens often competes with the rate at which subsequent photons are absorbed. In effect, the rate of intersystem crossing, internal conversion, and vibrational mixing defines the rate by which energy deposited in a given state, or mode, naturally "randomizes" within molecules, and it is against this rate that photoexcitation must compete.

In the condensed phase, vibrational relaxation typically occurs on the order of tens of femtoseconds to hundreds of picoseconds, as shown in laser pump-probe experiments (Elsaesser and Kaiser, 1991). One should be aware, however, that in simple liquids, such as liquid nitrogen or inert gases, vibrational relaxation times on the order of minutes have been found (Brueck and Osgood, 1976)!

3.2 Energy Localization

Recent work on molecules has shown the conditions for maintaining local concentrations of energy in spite of relatively high densities of vibrational states. In some molecules, at high levels of vibrational excitation the energy coupling between modes becomes weak. Under these conditions, it is possible to have vibrational energy localized primarily in one or more modes for many vibrational periods.

Examples of mode-specific chemistry include the work of Sinha *et al.* (1990) in which singly deuterated water was alternatively excited along the O—H bond or along the O—D bond. Upon reaction with hydrogen atoms,

$$H+HOD(4\nu_{OH})\rightarrow\begin{cases}OD+H_2, & (9)\\ OH+HD, & (10)\end{cases}$$

reaction (9) is enhanced, compared to (10), when the O—H bond is excited to the fourth overtone of the OH stretch. Bar *et al.* (1991) found that photodissociation of the same molecule at 193 nm depended upon the mode in which one quantum of vibration was placed. In this case, HOD excited to $\nu_3=1$ produced a different mix of photodissociation products from that normally produced by dissociation of HOD ground state, while there was no significant change in the product distribution when HOD was excited in the vibrational mode $\nu_1=1$.

Not all molecules are suitable for mode-specific photochemistry, as pointed out theoretically in the work of Kellman *et al.* (1991) on the HOD system. For example, Marquardt and Quack (1991), working on CHX_3 theoretical models, point out that the energy initially localized in a wave packet on the CH stretch mixes with CH bends in 80 fs, and is randomly distributed in hundreds of femtoseconds.

4. CHEMICAL CHANGES

4.1 Ionization and Detachment

One of the more striking effects of laser photochemistry is the frequent appearance of ions, even when working with photon energies less than the molecular ionization potential. This ionization usually results from the processes, shown in Fig. 2, of direct multiphoton absorption, resonant multiphoton ionization involving several intermediate states, or ionization of chemical products resulting from rearrangements or elimination of groups of atoms.

An analogous process for negative ions is the production of neutral species by multiphoton photodetachment. In fact, one of the first structured multiphoton spectra observed was that of photodetachment of C_2^- (Lineberger and Patterson, 1972). Detachment is reviewed (along with ion dissociation) by Miller (1981).

At high pressures, collisional mechanisms allow molecules to pool energy, and "chemi-ionization" may result. This is thought to be one possible mechanism to account for multiphoton ionization by IR absorption, similar to the production of ionization in flames (Avouris *et al.*, 1980).

4.2 Dissociation

The major photochemical pathway is that of dissociation. Lasers have been instrumental in exploring photodissociation by their ability to initiate and probe the sequence of events as molecules come apart. For example, it has been possible to excite, and observe, the emission by molecules while they are dissociating. Imre *et al.* (1984) reviewed the seminal experiments, along with the accompanying advances in theory that allow these experiments to be understood.

Lasers have also provided insight into the distribution of products, in terms of energy release, angular distributions of products, and the internal rotation and vibrational states of the fragments (Leone, 1985).

One study looked at the details of predissociation in formaldehyde by carefully comparing the rotational state distribution expected from an RRKM (Rice-Ramsperger-Kassels-Marcus) statistical theory with the state populations measured for CO (Bamford *et al.*, 1985). The RRKM model assumes an equipartitioning of energy between all molecular modes. Although there are only four atoms present, the excited electronic state efficiently couples to the vibrational manifold associated with the electronic ground state, which then decomposes by elimination of hydrogen:

$$H_2CO+h\nu\rightarrow H_2CO(S_1)\rightarrow H_2CO(S_0,\text{vib.})$$
$$\rightarrow H_2+CO(\text{rot.}).\quad(11)$$

In liquid solutions, photodissociation may be followed immediately by recombination of species trapped by a solvent cage. The effect of this cage has been the object of much study since it was first observed, and the prototypical system here is the photodissociation of I_2.

This system is still not fully understood, but the use of picosecond and faster lasers has allowed us a better understanding than in the past. A recent review is provided by Harris *et al.* (1988). As currently understood, dissociation happens on the subpicosecond time scale, followed by recombination on the order of several picoseconds. Vibrational relaxation occurs within several hundred picoseconds in polyatomic solvents. The most difficult step, which is not entirely understood, is the I-atom escape from the solvent cage, which can take place in anywhere from tens of picoseconds to tens of nanoseconds, depending upon the solvent in a way that is not yet understood.

One area of investigation that would be nearly impossible without lasers is photodissociation of isolated molecular ions in molecular beams or ion traps (Moseley, 1985). Examples of these studies include work on aromatic molecular ions of benzene (Freiser and Beauchamp, 1975),

$$C_6H_6^+ + h\nu\ (2.5\ \text{eV}) \rightarrow C_6H_6^{+*},$$

$$C_6H_6^{+*} + h\nu\ (1.5\ \text{eV}) \rightarrow C_6H_5^+ + H, \qquad (12)$$

and toluene (Eyler, 1976),

$$C_7H_8^+ + h\nu \rightarrow C_7H_7^+ + H. \qquad (13)$$

In the first case, the benzene cation

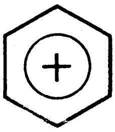

requires sequential electronic absorption of two photons in order to surmount the barrier to dissociation. This is one example of large-molecule photoabsorption, as discussed above, followed by dissociation. Interestingly, the intermediate formed by the absorption of one photon is very long-lived (seconds) and appears to be a vibrationally hot ion, but in its electronic ground state. Relaxation of the initially produced, electronically excited state into a mixture of vibrational levels of the ground electronic state prevents the excited molecular ion from radiating, and prepares the ion for absorption of an additional quantum of radiation. The combined energy of two photons now exceeds the threshold for dissociation of about 4 eV.

The toluene ion,

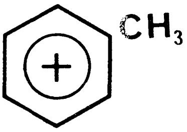

illuminated at almost the same energy, dissociates in a one-photon process. Direct elimination of a hydrogen from the methyl group to produce the benzyl cation

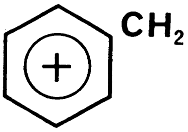

is energetically allowed when 2.3 eV is supplied by radiation. However, isomerization of the carbon atom backbone of the molecule, enlarging the ring to seven members,

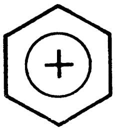

in concert with the elimination of a hydrogen, would be energetically possible with the addition of only 1.7 eV of energy from radiation. The near coincidence of the photothreshold (2.3 eV) with the predicted thermodynamic threshold for the production of benzyl appears to indicate that direct dissociation occurs without (at least initially) isomerization of the molecular ion fragment.

Dissociation is also the usual step initiating radical chain reactions, such as polymerization. Commercially important are reactions involving azides,

$$RN_3 + h\nu \rightarrow RN{:} + N_2, \qquad (14)$$

triazines, and electron-transfer initiators such as diaryliodonium and triarylsulfonium salts.

4.3 Isomerization

Isomerizations, the rearrangement of atoms within single molecules, are some of the cleanest photochemical transformations. Classical photochemistry has focussed intently on *cis-* to *trans-* isomerizations, and these are the object of laser photochemistry, too.

IR photochemistry of the isomers of 1,2-dichloroethene at 981 cm^{-1} results in a photostationary mixture in which the higher-energy *cis-* conformation dominates, in contrast to the result produced by simple heating (Ambartzumian *et al.*, 1976).

Photophysical and photochemical processes accompanying isomerizations of the visual pigment rhodopsin have been studied on time scales ranging from femtoseconds to seconds with lasers. The key transformation occurring there is the isomerization of the conjugated chain of retinal between conformations (a) and (b):

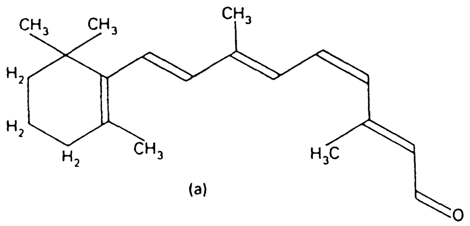

(a)

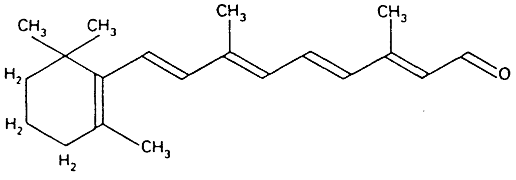

(b)

4.4 Condensation and Addition

A common photochemical process is the initiation of polymerization, the condensation of many monomer molecules into larger polymer molecules. While the usual mechanism is radical-initiated or radical-ion–initiated chain reaction, occasionally a direct addition of an excited reactant with another is causative.

One of the more striking effects observed upon passing laser light through an absorbing gas sample is that of photoinduced condensation, producing brilliant opalescence along the laser beam. Examples of this include alkali-metal vapors through which light from a dye laser is directed. In this configuration, a significant fraction of alkali-metal-atom excited states are produced, which then may act as nucleation centers for the formation of diatomic molecules and clusters. Several mechanisms appear to be operating; the most striking occurs when the radiation density is sufficient to cause ionization. Ion-atom potentials are always attractive, and they operate at long range, and this electrostatic force, in contrast to covalent bonding forces, does not saturate, allowing the growth of large clusters of molecules that otherwise might not even react.

5. APPLICATIONS

Their cost per mole makes laser photons an expensive reagent, economical in only high-value products. Uses that can be foreseen for laser photons include the production of drugs, the isolation of isotopes, processing of materials on or near surfaces, and the storage and retrieval of information.

5.1 Holography

Laser photochemistry proved important early in the field of holography (see HOLOGRAPHY, OPTICAL). Lasers provided the coherent light, and the interference patterns were recorded by the photochemistry of either silver halide photographic emulsions or organic photoresists. Holographic technology now provides nearly ghost-free commercial diffraction gratings (see MONOCHROMATORS).

5.2 Isotope Selection

High wavelength specificity of lasers, along with sufficient power to cause chemistry on the mole scale, led to interest in using lasers for isotope production. Some possible methods are reviewed by Letokhov (1983) and some of the concerns are raised by Krauss (1977). Numerous processes have become available for laser isotope separation (Cantrell *et al.*, 1979; Hartford, 1979; Jensen *et al.*, 1980; Letokhov, 1979; Letokhov and Moore, 1976, 1977; Zare, 1977). The application to the separation of uranium isotopes has been the focus of studies on both atomic U and its compounds (Dietz *et al.*, 1982).

5.3 Surface Modification

The most common industrial use of lasers in materials processing is for cutting or machining, applying the high energy densities of laser beams to burn or ablate materials. Secondary to this use, lasers have found their way into hardening of materials, thermal treatments, and exposure of photoresists.

5.3.1 Etching. One possibility for increasing the effectiveness of laser ablation techniques is in modification of a nonabsorptive material to make it absorptive. High–band-gap materials will develop color centers upon electron irradiation. At low power levels, lasers can be used to photobleach these color centers; at high power levels, these same color centers can be used to provide the initial absorption necessary to initiate ablation.

Lasers have been found to enhance greatly the rate at which a chemical process removes material from surfaces. Plasma etching and electrolytic etching are examples where lasers enhance the rate of reaction. One effect of lasers that is different from most other techniques is the ability to etch long channels in materials along the axis of light propagation with diameters narrower than the beam profile (Podlesnik and Gilgen, 1989).

5.3.2 Deposition. Laser decomposition of organometallic compounds over surfaces generates metal atoms and deposits metal layers at the adjacent surface (Osgood, 1983; Ehrlich and Tsao, 1983; Jones *et al.*, 1985). Direct dissociation of gaseous organometallic compounds occurs promptly, and the metal ions often possess significant translational energy, as observed in the Doppler profiles of optical transitions of these ions (Marinero and Jones, 1985). This would seriously complicate the ability to deposit patterns at surfaces, since metal atoms generated in the gas phase would tend to spread over the surface, away from the region of photodissociation. Dissociation has been found, however, to be enhanced at surfaces, which permits the writing of structures directly on the surface with dimensions on the order of 10^{-7} m (Wilson and Houle, 1985; Houle *et al.*, 1986).

The field of laser deposition writing is summarized in the book edited by Ibbs and Osgood (1989).

5.4 Information Storage

Many lasers in the information industry provide controlled light for exposure of optical recording media. Most photochemical processes are based upon photographic or photoresist technology, although currently information storage is dominated by photophysical techniques such as xerography (*q.v.*) or thermal melting, such as in the now familiar commercial CD and "laser" disks. Currently, the relative lack of cw uv lasers operating at wavelengths below 300 nm, compared to the ready availability of IR and visible lasers, limits the types of photochemical recording methods available to information technology.

5.5 Chemical Analysis

Lasers have been used to ablate surfaces, thereby generating gaseous ions, atoms, and molecular fragments from samples of surfaces and molecules at surfaces. One use to which multiphoton ionization has been applied is in the detection of molecules eluting from a gas chromatograph (Klimcak and Wessel, 1980) (see CHROMATOGRAPHY).

5.6 Biochemistry

Since a number of biological processes are photochemical, it has been natural to use light in the study of these. The photochemistry of rhodopsins, the pigments of the eye, has been extensively investigated with ever-shortening laser pulses. For a review, see Birge (1990). The major chemical mechanism of vision is isomerization of the long, conjugated-chain molecule (see Sec. 4.3), but numerous conformational changes occur in this molecule before and after this isomerization occurs.

One area of concern to molecular biologists is the rate at which proteins, especially enzymes, change configurations. Optical techniques have been used to study these changes. Miller (1991) reviews how transient optical gratings have provided information about the transformations occurring in hemoglobin with the addition and release of oxygen or CO, and the rate at which various conformations of the heme group relax.

GLOSSARY

Ablation: Removal of material from a surface caused by the impact of light, atoms, molecules, or ions. Generally, a thermal process involving local evaporation.

Bond-Specific Photochemistry: Photochemistry having products determined by a specific bond or combination of bonds set into vibrational motion.

Chemi-ionization: Ionization caused by a chemical reaction.

Conformation: The (possibly unstable) arrangement of atoms inside a molecule.

Conformers: Different, relatively stable, arrangements of atoms, each consisting of identical numbers and types of atoms. A dissociating or rearranging molecule may pass through different conformers. Conformers need not be capable of being physically or chemically isolated.

Hole Burning: Depletion of part of an inhomogeneous distribution of molecules, caused by the interaction of light with those molecules; selective photobleaching.

Internal Conversion: Mixing of the vibrational levels of an electronic state with the vibrational levels of another electronic state.

Intersystem Crossing: Mixing between states of different spin.

Isomers: Different molecular conformations, consisting of identical atoms, that are stable, or metastable long enough to be physically or chemically separated.

Lamb Dip: The reduction in absorption (hole burning) observed when two counterpropagating laser beams interact with the same atoms. Usually observed at the center of a Doppler-broadened transition.

"Large" Molecule: A molecule having a high density of states, forming numerous overlapping resonances that interact simultaneously with radiation of a single frequency.

Local Mode: Classically, relative motion of one atom, or a group of atoms, in a molecule. Quantum mechanically, an eigenfunction or superposition of degenerate, or nearly degenerate, eigenfunctions of the vibrational Hamiltonian with considerable amplitude on one or a few atoms.

Mechanism: Sequence of physical and chemical steps leading to a chemical change. Sometimes used to refer to the most unique or characteristic step of an overall mechanism.

Mode-Specific Photochemistry: Photochemistry having products determined by the vibrational mode set into motion.

Multiphoton Absorption: Direct multiphoton absorption occurs when two or more photons are absorbed at once, mediated by higher-order perturbation terms of the molecular Hamiltonian interaction with the radiation field. Sequential one-photon events are sometimes called by this name, but more properly may be called "multiple photon absorption."

Multiphoton Ionization (MPI): Ionization caused by multiphoton absorption, either directly into the ionization continuum, or by absorption to an intermediate resonant state, followed by photoionization. The latter is termed "resonant" or "selective."

Normal Mode: Classically, molecular vibration in which (nearly) all atoms move concertedly and periodically with the same frequency, but possibly not with the same phase. Quantum mechanically, eigenfunction of the vibrational Hamiltonian.

Organometallic: A compound containing organic carbon and metal atoms.

Photodetachment: Ejection of an electron from a negative ion by the effect of light.

Photodissociation: Fragmentation of a molecule into two or more pieces by the effect of light.

Photoresist: An organic resin that polymerizes under the influence of light.

Rabi Frequency: The rate at which a stimulated electronic transition would occur in response to a radiation field. A simple two-level system will alternately absorb and emit, toggling between excited and ground states at this frequency.

Rovibronic State: An individual state of an isolated molecule, including contributions from rotational, vibrational, and electronic motion.

RRKM: Rice-Ramsperger-Kassel-Marcus theory of unimolecular reactions, based upon the statistical flow of vibrational energy within a molecule. The rate at which fluctuations localize energy in critical, reactive bonds determines the rate of reaction.

"Small" Molecule: A molecule having a sparse density of states in which isolated resonances interact individually with a radiation field of a single frequency.

Spontaneous Raman Scattering: Inelastic scattering by matter of photons into a random final state.

Stimulated Raman Scattering: Inelastic scattering of photons stimulated by the radiation field interacting with matter.

Vibrational Relaxation: Loss of internal vibrational energy into other modes of motion.

List of Works Cited

Ambartzumian, R. V., Chekalin, N. V., Doljikov, V. S., Letokohov, V. S., Lokhman, V. N. (1976), *Opt. Commun.* **18**, 400–402.

Avouris, P., Chan, I. Y., Loy, M. T. (1980), *J. Chem. Phys.* **72**, 3522–3527.

Bamford, D. J., Filseth, S. V., Foltz, M. F., Hepburn, J. W., Moore, C. B. (1985), *J. Chem. Phys.* **82**, 3032–3041.

Bar, I., Cohen, Y., David, D., Arusi-Parpar, T., Rosenwaks, S., Valentini, J. J. (1991), *J. Chem. Phys.* **95**, 3341–3347.

Bernstein, R. B. (1982), *J. Phys. Chem.* **86**, 1178–1184.

Birge, R. R. (1990), *Annu. Rev. Phys. Chem.* **41**, 683–733.

Brueck, S. R. L., Osgood, R. M., Jr. (1976), *Chem. Phys. Lett.* **39**, 568–572.

Cantrell, C. D., Freund, S. M., Lyman, J. L. (1979), in: M. L. Stitch (Ed.), *Laser Handbook*, Vol. 3, Boca Raton, FL: CRC Press, pp. 485–576.

Carmichael, H. J., Walls, D. F. (1976), *J. Phys. B: Atom. Mol. Phys.* **9**, 1199-1219.

Dietz, T. G., Duncan, M. A., Smalley, R. E., Cox, D. M., Horsley, J. A., Kaldor, A. (1982), *J. Chem. Phys.* **77**, 4417–4426.

Ehrlich, D. J., Tsao, J. Y. (1983), *J. Vac. Sci. Technol. B* **1**, 969–984.

Elsaesser, T., Kaiser, W. (1991), *Annu. Rev. Phys. Chem.* **42**, 83–107.

Engelking, P. C. (1980), *Chem. Phys. Lett.* **74**, 207–210.

Eyler, J. R. (1976), *J. Am. Chem. Soc.* **98**, 6831–6834.

Freiser, B. S., Beauchamp, J. L. (1975), *Chem. Phys. Lett.* **35**, 35–40.

Gedanken, A., Robin, M. B., Kuebler, N. A. (1982), *J. Phys. Chem.* **86**, 4096-4107.

Gerrity, D. P., Rothberg, L. J., Vaida, V. (1980), *Chem. Phys. Lett.* **74**, 1–5.

Golden, D. M., Rossi, M. J., Baldwin, A. C., Barker, J. R. (1981), *Acc. Chem. Res.* **14**, 56–62.

Harris, A. L., Brown, J. K., Harris, C. B. (1988), *Annu. Rev. Phys. Chem.* **39**, 341-366.

Hartford, A., Jr. (1979), *AIChE Symp. Ser.* **75**, 44–49.

Hopfield, J. J., Warlock, J. M., Park, K. (1963), *Phys. Rev. Lett.* **11**, 414–417.

Houle, F. A., Wilson, R. J., Baum, T. H. (1986), *J. Vac. Sci. Technol. A* **4**, 2452–2458.

Ibbs, K. G., Osgood, R. M. (Eds.) (1989), *Laser Chemical Processing for Microelectronics*, Cambridge, U.K.: Cambridge University Press.

Imre, D., Kinsey, J. L., Sinha, A., Krenos, J. (1984), *J. Phys. Chem.* **88**, 3956–3964.

Jensen, R. J., Sullivan, J. A., Finch, F. T. (1980), *Separation Sci. Technol.* **15**, 509–532.

Johnson, P. M. (1975), *J. Chem. Phys.* **62**, 4562–4563

Johnson, P. M. (1976), *J. Chem. Phys.* **64**, 4143–4148.

Johnson, P. M., Berman, M. R., Zakheim, D. (1975), *J. Chem. Phys.* **62**, 2500–2502.

Jones, C. R., Houle, F. A., Kovac, C. A., Baum, T. H. (1985), *Appl. Phys. Lett.* **46**, 97–99.

Kaiser, W., Garrett, C. G. B. (1961), *Phys. Rev. Lett.* **7**, 229–231.

Kellman, M. E., Li, Z., Xiao, L. (1991), *Chem. Phys. Lett.* **181**, 141–150.

Khundkar, L. R., Zewail, A. H. (1990), *Annu. Rev. Phys. Chem.* **41**, 15–60.

Klimcak, C. M., Wessel, J. E. (1980), *Anal. Chem.* **52**, 1233–1239.

Kraus, A. S. (1977), *Science* **196**, 721–731.

Leone, S. (1985), *Science* **227**, 889–895.

Letokhov, V. S. (1979), *Nature* **277**, 605–610.

Letokhov, V. S. (1983), *Nonlinear Laser Chemistry*, Berlin: Springer-Verlag.

Letokhov, V. S., Moore, C. B. (1977), in: C. B. Moore, (Ed.), *Chemical and Biochemical Applications of Lasers*, Vol. 3, New York: Academic, pp. 1–165.

Leutwyler, S., Even, U., Jortner, J. (1980), *Chem. Phys. Lett.* **74**, 11–14.

Li, Leping, Yang, Bing-Xin, Johnson, P. M. (1985), *J. Opt. Soc. Am. B* **2**, 748-752.

Lineberger, W. C., Paterson, T. A. (1972), *Chem. Phys. Lett.* **13**, 40–44.

Mallow, B. R. (1969), *Phys. Rev.* **188**, 1969–1975.

Marinero, E. E., Jones, C. R. (1985), *J. Chem. Phys.* **82**, 1608–1609.

Marquardt, R., Quack, M. (1991), *J. Chem. Phys.* **95**, 4854–4876.

McClain, W. M. (1974), *Acc. Chem. Res.* **7**, 129–135.

Miller, R. J. D. (1991), *Annu. Rev. Phys. Chem.* **42**, 581–614.

Miller, T. M. (1981), *Adv. Elec. Electron. Phys.* **55**, 119–188.

Moseley, J. T. (1985), in: K. P. Lawley (Ed.), *Photodissociation and Photoionization*, New York: Wiley, pp. 245–298.

Mukamel, S. (1990), *Annu. Rev. Phys. Chem.* **41**, 647–681.

Osgood, R. M., Jr. (1983), *Annu. Rev. Phys. Chem.* **34**, 77–101.

Peticolas, W. L., Rieckhoff, K.E. (1963), *J. Chem. Phys.* **39**, 1347–1348.

Peticolas, W. L., Goldsborough, J. P., Rieckhoff, K. E. (1963), *Phys. Rev. Lett.* **10**, 43–45.

Petty, G., Tai, C., Dalby, F. W. (1975), *Phys. Rev. Lett.* **34**, 1207–1209.

Podlesnik, D. V., Gilgen H. H. (1989), in: K. G. Ibbs, R. M. Osgood (Eds.), *Laser Chemical Processing for Microelectronics*, Cambridge, U.K.: Cambridge University Press.

Rothberg, L. J., Gerrity, D. P., Vaida, V. (1981), *J. Chem. Phys.* **75**, 4403–4412.

Singh, S., Stoicheff, B. P. (1963), *J. Chem. Phys.* **38**, 2032–2033.

Sinha, A., Hsiao, M. C., Crim, F. F. (1990), *J. Chem. Phys.* **92**, 6333–6335.

Smalley, R. E. (1982), *J. Phys. Chem.* **86**, 3504–3512.

Staginnus, B., Frölich, D., Caps, T. (1968), *Rev. Sci. Instrum.* **39**, 1129–1130.

Weisz, S. Z., Zahlen, A. B., Gilreath, J., Jarnagin, R. C., Silver, M. (1964), *J. Chem. Phys.* **41**, 3491–3495.

Wilson, R. J., Houle, F. A. (1985), *Phys. Rev. Lett.* **55**, 2184–2187.

Zare, R. N. (1977), *Sci. Am.* **236** (2), 86–98.

Further Reading

Albertson, B. (1969), *Photochemical Processes*, Park Ridge, NJ: Noyes Data Corp.

Ambartzumian, R.V., Letokhov, V. S. (1977), *Acc. Chem. Res.* **10**, 61–67.

Andrews, D. L. (1986), *Lasers in Chemistry*, Berlin: Springer-Verlag.

Bauerle, D. (1986), *Chemical Processing with Lasers*, Berlin: Springer-Verlag.

Bloembergen, N., Yablonovich, E. (1978), *Phys. Today* **31** (5), 23–30.

Braun, A. M. (1991), *Photochemical Technology*, New York: Wiley.

Ibbs, K. G., Osgood, R. M. (Eds.) (1989), *Laser Chemical Processing for Microelectronics*, Cambridge, U.K.: Cambridge Univ. Press.

Letokhov, V. S. (1983), *Nonlinear Laser Chemistry*, Berlin: Springer-Verlag.

Luxon, J. T., Parker, D. E. (1987), *Lasers in Manufacturing*, Kempston, Bedford, U.K.: IFS.

Phillips, R. (1983), *Sources and Applications of Ultraviolet Radiation*, New York: Academic.

Ranney, M.W. (1975), *Irradiation in Chemical Processes: Recent Developments*, Park Ridge, NJ: Noyes Data.

Journals specializing in laser photochemistry include *Applied Physics B: Photophysics and Laser Chemistry* and *Laser Chemistry*.

18

Laser Sources

David L. Andrews and Robert H. Lipson

18.1 Introduction 657
18.2 Optically Pumped Solid-State Lasers 657
18.2.1 Ruby Laser 659
18.2.2 Neodymium Lasers 660
18.2.3 Widely Tunable Lasers 662
18.2.4 Color-Center Lasers 662
18.2.5 Fiber Lasers 662
18.3 Diode Lasers 664
18.3.1 Vertical Cavity Surface Emitter Lasers 666
18.3.2 Visible Laser Diodes 666
18.3.3 Quantum Dot Lasers 667
18.3.4 Quantum Cascade Lasers 667
18.3.5 Quantum Microcavity Lasers 668
18.4 Atomic and Ionic Gas Lasers 669
18.4.1 Helium–Neon Laser 669
18.4.2 Argon and Krypton Lasers 671
18.5 Molecular Gas Lasers 672
18.5.1 Carbon Dioxide Laser 672
18.5.2 Nitrogen Laser 675
18.5.3 Iodine Laser 675
18.5.4 Excimer Lasers 676
18.5.5 Dye Lasers 677
18.6 Pulsing Techniques 680
18.6.1 Cavity Dumping 680
18.6.2 Q-Switching 681
18.6.3 Mode-Locking 681
18.6.4 Ultrafast Pulse Compression 683
18.6.5 Continuum Generation 684
18.7 Short Wavelength Coherent Radiation 684
18.7.1 Frequency Conversion 684
18.7.2 Frequency Doubling 685

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

18.7.3 Optical Parametric Oscillators 687
18.7.4 Frequency Tripling and Four-Wave Mixing 687
18.7.5 Higher Order Harmonic Generation 689
18.7.6 Free-Electron Laser 691
18.7.7 Tabletop X-ray Lasers 692
References 692

18.1
Introduction

The uses of lasers in spectroscopy are extensive. The key characteristics of laser light – particularly the unparalleled combination of an essentially monochromatic output at high intensity levels – are extremely well suited to a wide variety of spectroscopic techniques. Often, high laser powers are exploited in the determination of weak or transient spectral features, or to produce optically nonlinear processes; equally, the narrow bandwidth of many laser sources is utilized for high-resolution studies. A wide variety of optical materials is employed in the lasers themselves, and the detailed nature of these materials determines the principal output characteristics such as wavelength and achievable power. Since the construction of the first laser based on ruby, widely ranging materials have been adopted as laser media. The range is still being extended to provide output at new wavelengths and, as can be seen from Table 18.1, the output available from commercial lasers now covers most of the electromagnetic spectrum.

By a variety of means, laser light can also be tailored to specific applications by frequency-conversion and pulse-generation techniques; the facility to modify these characteristics is again important for many spectroscopic investigations. This chapter reviews the basic mechanisms and principles of operation of the most important laser sources and beam management techniques, focusing on systems that are especially well suited to spectroscopic applications. Space does not permit a thorough description of the entire range of laser systems now available; the more limited scope of the following account is a discussion of representative laser instrumentation.

18.2
Optically Pumped Solid-State Lasers

The active medium in most solid-state lasers, capable of delivering 1 W or more power in the UV to near-IR range, is generally a transparent crystal or glass doped with a small amount of transition metal; transitions in the metal ions are responsible for the laser action. The two most common dopants are chromium (in the ruby and alexandrite lasers) and neodymium (Nd : YAG and Nd : glass lasers), with

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

Tab. 18.1 Wavelengths available from commercial lasers.[a]

λ/nm	Laser	λ/nm	Laser	λ/nm	Laser
157	Fluorine	428	Nitrogen	568	Krypton
174	Ruby × 4	437	Argon	578	Copper
193	Argon fluoride	442	Helium–cadmium	596	Xenon
213	Nd:YAG × 5	455	Argon	628	Gold
222	Krypton chloride	458	Krypton	633	Helium–neon
231	Ruby × 3	458	Argon	647	Krypton
244	Argon × 2	462	Krypton	657	Krypton
248	Krypton fluoride	463	Krypton	677	Krypton
257	Argon × 2	466	Argon	680	GaAlInphosphide
263	Nd:YLF × 4	468	Krypton	687	Krypton
266	Nd:YAG × 4	473	Argon	694	Ruby
275	Argon	476	Krypton	723	Lead
305	Argon	477	Argon	753	Krypton
308	Xenon chloride	477	Krypton	799	Krypton
312	Gold	483	Krypton	852	Calcium
325	Helium–cadmium	485	Krypton	866	Calcium
334	Argon	488	Argon	904	Gallium arsenide
337	Nitrogen	496	Xenon	1053	Nd:YLF
347	Ruby × 2	497	Argon	1060	Nd:glass
351	Krypton	502	Argon	1064	Nd:YAG
351	Nd:YLF × 3	511	Copper	1092	Argon
351	Xenon fluoride	515	Argon	1130	Barium
351	Argon	521	Krypton	1152	Helium–neon
353	Xenon fluoride	527	Nd:YLF × 2	1290	Manganese
355	Nd:YAG × 3	529	Argon	1300	In Ga As phosphide
356	Krypton	531	Krypton	1313	Nd:YLF
364	Argon	532	Nd:YAG × 2	1315	Iodine
407	Krypton	534	Manganese	1319	Nd:YAG
413	Krypton	540	Xenon	1500	In Ga As phosphide
415	Krypton	544	Helium–neon	1523	Helium–neon

[a] Emission lines of commonly available discrete-wavelength lasers over the range 100–1600 nm, harmonics indicated by × 2, × 3 etc. The box indicates wavelengths encompassed by the tuning range of a titanium:sapphire source.

the dopant concentration typically 1% or less. Such lasers are optically pumped by a broadband flash-lamp source, and can be pulsed to produce the high intensities.

18.2.1
Ruby Laser

The ruby laser has an important place in history as it was the first type ever constructed (in 1960). In construction, the modern ruby laser comprises a rod of commercial ruby (0.05% Cr_2O_3 in an Al_2O_3 lattice) of between 3 and 25 mm diameter and up to 20 cm length. The chromium ions are excited by the emission from a flash-lamp coiled around it, or placed alongside it within an elliptical reflector as shown in Figure 18.1.

From the energy level scheme shown in Figure 18.2, it is apparent that the ruby laser may be regarded as a *pseudo three-level system*, because in the course of excitation, decay, and emission, only three Cr^{3+} levels are involved; $^4A_2$, $^4T_1$ or $^4T_2$, and ^2E. These energy levels are quite different from those of a free chromium atom, due to crystal field splitting. The initial flash-lamp excitation takes the Cr^{3+} ions up from the ground state E_l ($^4A_2$) to one of the two E_3 (^4F) levels. Both of these levels have subnanosecond lifetimes, and rapidly decay into the E_2 level (^2E). The decay takes place by nonradiative processes that channel energy into lattice vibrations, heating the crystal. Because the E_2 level has a much longer decay time of around 4 ms, a population inversion is created between the E_2 and E_l states, leading to laser emission at a wavelength of 694.3 nm as the majority of Cr^{3+} ions cascade down to the ground state. The lasing process thus generates a single, intense pulse of light, typically between 0.3 and 3 ms duration; the delay between successive pulses usually lasts several seconds, and can be as long as a full minute. A very different kind of pulsing can be created by the technique known as *Q-switching*, discussed in Section 18.6.2

A problem with the ruby laser, common to all lasers of this type, is the damage caused by the repeated cycle of heating and cooling associated with the generation of each pulse, which

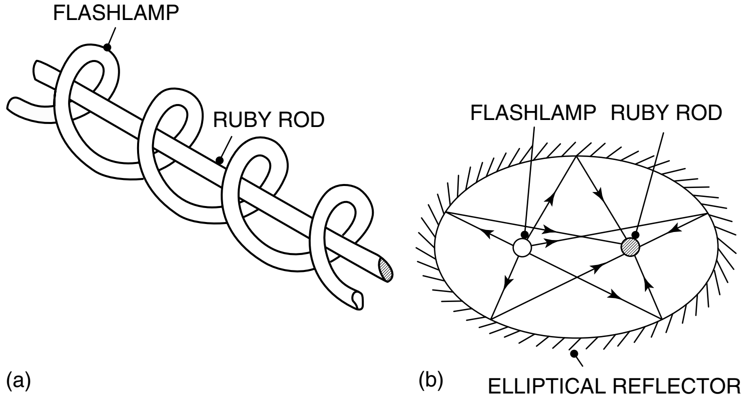

Fig. 18.1 Arrangement of flash-lamp tube and ruby rod in a ruby laser. (a) Helical flash-lamp option; (b) confocal flash-lamp and laser rod within an elliptical reflector.

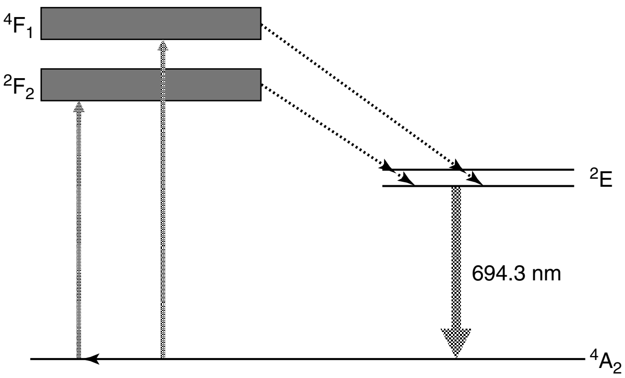

Fig. 18.2 Energetics of the three-level ruby laser. The vertical arrows upward on the left hand side refer to pumping transitions. The dashed lines refer to nonradiative relaxation from the pump bands into the upper lasing level. The broad vertical arrow downward corresponds to the laser line at 649.3 nm.

ultimately necessitates the replacement of the ruby rod. To improve performance, the rod usually needs to be cooled by circulation of water in a jacket around it. Despite its drawbacks, with pulse energies as high as 200 J, the ruby laser represents a powerful source of monochromatic light in the optical region. The emission bandwidth is typically about 0.5 nm, but this can be reduced by a factor of up to 10^4 by suitable optics. The beam diameter of a low-power ruby laser can be as little as 1 mm, with 0.25 mrad divergence; the most powerful lasers may have beams up to 25 mm in diameter and a divergence of several mrad.

18.2.2
Neodymium Lasers

Neodymium lasers are of two main types. In one, the host lattice for the neodymium ions is a crystal – usually yttrium aluminum garnet crystal ($Y_3Al_5O_{12}$) – and in the other, the host is an amorphous glass. These are referred to as *Nd : YAG* and *Nd : glass*, respectively. Although transitions in the neodymium ions are responsible for laser action in both cases, the emission characteristics differ because of the influence of the host lattice on the neodymium energy levels. Also, glass does not have the excellent thermal conductivity properties of YAG crystal, so that it is more suitable for pulsed than for continuous-wave (CW) operation. Like the ruby laser, both types of neodymium laser are normally pumped by a flash-lamp arranged confocally alongside a rod of the laser material in an elliptical reflector. The energy levels of neodymium ions involved in laser action are split by interaction with the crystal field, as illustrated in Figure 18.3. As a result, transitions between components of the $^4F_{3,2}$ and $^4I_{11/2}$ states, forbidden in the free state, are permitted and can give rise to laser emission. The $^4F_{3,2}$ levels are initially populated following nonradiative decay from higher energy levels excited by the flash-lamp, and because the terminal $^4I_{11/2}$ laser level lies above

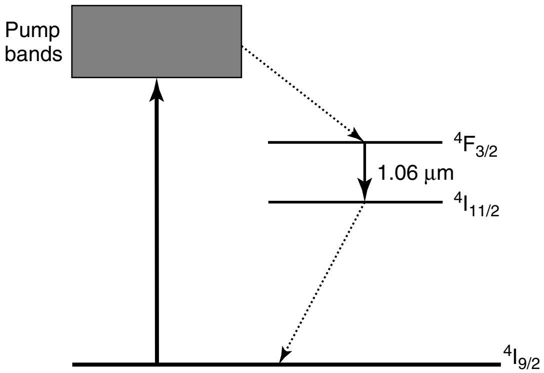

Fig. 18.3 Simplified energy levels of a four-level neodymium laser lasing at 1.06 μm.

the $^4I_{9/2}$ ground state, we thus have a pseudo- four-level system.

The principal emission wavelength for both types of neodymium laser is around 1.064 μm, in the near-IR (or 1.053 μm in the less common Nd : YLF lasers, where lithium replaces aluminum in the host lattice). Some commercial lasers can also operate on a different transition, producing 1.319 μm output. The YAG and glass host materials impose very different characteristics on the emission, however. Quite apart from the differences in thermal conductivity, which determine the question of continuous or pulsed operation, one of the main differences appears in the line width. Because glass has an amorphous structure, the crystal field splitting also varies from site to site. Hence, the line width is much broader than in the Nd : YAG laser, where the lattice is much more regular and the field splitting is more uniform. However, the concentration of neodymium dopant in glass may be as large as 6%, compared with 1.5% in a YAG host, so that a much higher energy output can be obtained. For both these reasons, the Nd : glass laser is ideally suited for the production of extremely high intensity ultrashort pulses by the technique of mode-locking (see Section 18.6.3). In fact, amongst all commercially available lasers, it is the neodymium lasers that deliver the highest available laser beam intensities.

The output power of a typical CW Nd : YAG laser is several watts, and can exceed 200 W. When operating in a pulsed mode, the energy per pulse varies according to the method of pulsing and the pulse repetition rate, but can be anywhere between a small fraction of a joule and 100 J for a single pulse. The majority of Nd : YAG research applications in spectroscopy make use not of the 1.064 μm radiation as such, but rather the high-intensity visible light that can be produced by frequency-conversion methods. Particularly important in this respect are the wavelengths of 532, 355, and 266 nm obtained by harmonic generation (see Section 18.7.1).

Diode-pumped solid-state lasers represent a very different laser technology. These miniature-scale devices, which can often fit comfortably into the hand, incorporate a semiconductor diode laser (Section 18.3) to directly pump a small Nd : YAG crystal. Cavity modifications enable line-widths of less than 5 kHz (2×10^{-7} cm^{-1}) to be attained in certain models. While CW output powers are generally low (milliwatt scale), the

1-W target is exceeded in some commercially available systems. Also important is the common inclusion of frequency-conversion crystals in these devices, so providing milliwatt emission in the visible region, especially a 532-nm line in the green. Such integrated solid-state lasers have the advantages of smaller size, higher efficiency, better stability, and lower noise levels than most gas discharge lasers.

18.2.3
Widely Tunable Lasers

Recently, other types of solid-state laser have been developed which, although again based on optical transitions in transition metal ions embedded in a host ionic crystal, produce *tunable* radiation. For a number of spectroscopic applications, such lasers represent an attractive alternative to tunable lasers based on organic dyes. Their chief advantages are an appreciably more rugged and compact construction and operation without use of toxic chemicals.

The recently developed titanium : sapphire laser, commonly pumped by a primary laser beam such as the all-lines output of an ion gas laser, is the best known example in this category, and is today one of the most widely used commercially available solid-state lasers. Crucially, the coupling of electronic levels with lattice vibrations here leads to a broad, continuous band of *vibronic* energy levels. On the basis of the Al_2O_3 crystal doped with approximately 0.1% Ti_2O_3, lasing occurs on vibronic $^2E \rightarrow {}^2T_2$ transitions in the Ti^{3+} ions, offering high-power tunable CW or pulsed emission over the exceptionally large wavelength range 650–1100 nm in the near-infrared, peaking near 800 nm. At that wavelength, over 5 W CW outputs can be achieved. The titanium : sapphire laser is finding an increasingly wide range of applications, particularly in spectroscopy. Other commercially available vibronic lasers offering pulsed output in the infrared are Co^{2+}-doped MgF_2 (the first tunable solid-state laser) tunable over the range 1.75–2.50 μm, and chromium-doped Mg_2SiO_4 (forsterite), in which the laser-active species is Cr^{4+} and the corresponding emission range 1.15–1.35 μm.

18.2.4
Color-Center Lasers

Another type of solid-state laser in which the active medium is a regular ionic crystal does not involve transition metal ions. This is the so-called *F-center* (or *color center*) laser that operates on optical transitions at defect sites in alkali halide crystals, as for example in KCl doped with thallium; the color centers are typically excited by a pump Nd : YAG or argon/krypton ion laser. Such lasers produce radiation that is tunable over a small range of wavelengths in the overall region 0.8–3.4 μm; different crystals are required for operation over different parts of this range. One disadvantage of this type of laser is that the crystals need to be held at cryogenic temperatures. Color center lasers have found limited applications, although they have proven use in the high-resolution spectroscopy of small molecules.

18.2.5
Fiber Lasers

As the name suggests, the gain medium of these devices are silica or fluoride glass fibers doped with active rare-earth ions

(King, 2004). Some of dopants used include Nd, Tm, or Yb. The fiber, which has an external cladding with an index of refraction which is lower than that of the core, confines the pump and laser light to the core. Mirrors that are attached at the ends of the fiber or Fresnel reflections at each end complete the cavity. The devices are usually end-pumped through one of the mirrors (longitudinal pumping), although side pumping is possible. The long interaction lengths and wave guiding nature of the cladded fibers translates into devices with high gain, low thresholds, high efficiencies, and compactness. Unlike other solid-state lasers, the beam quality in these devises is determined here mainly by the refractive index profile of the cladding materials. Gain broadening in the fiber allows for limited tunability. Coherent emission can be obtained between ≈0.3 and 4.0 μm using different ions and different fibers.

A particularly important dopant is erbium, Er^{3+}, which under laser-diode pumping at 0.81, 0.98, or 1.48 μm can produce coherent radiation between 1.52 and 1.62 μm. This output range overlaps $\lambda = 1.55$ μm, which corresponds to the wavelength of lowest transmission losses in fiber-optic systems. This makes Er^{3+}-doped fiber amplifiers a key component in optical communications systems. A schematic energy level diagram of Er^{3+} is shown in Figure 18.4. Overall, the behavior of the $^4I_{15/2}$ lower laser state is similar to that of a dye laser molecule (Section 18.5.5). The higher energy M_J components are sparsely populated at typical operating temperatures, which results in a population inversion with respect to the $^4I_{13/2}$ upper lasing state. The inversion is maintained by rapid lower-level nonradiative relaxation.

One parasitic effect that can reduce the infrared gain in a fiber laser is the process of upconversion where light is emitted at higher photon energies than that of the pump photon generating the excitation. However, at the same time, this has led to another class of solid-state fiber lasers, called *upconversion lasers*, that operate at different wavelengths in the visible, depending on the rare-earth ion dopant(s) involved (Auzel, 2004).

Several mechanisms can lead to upconversion. The first is sequential absorption of pump laser photons through metastable (relatively long lived) ion excited states. In this case, low dopant concentrations but high pump powers

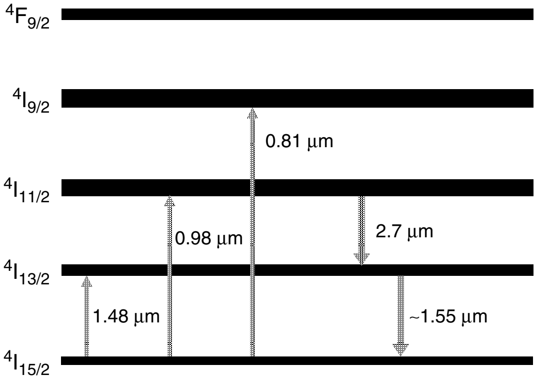

Fig. 18.4 Energy level diagram of Er^{3+} showing pumping transitions as vertical upward arrows and lasing transitions as vertical downward arrows. The thickness of the lines indicates unresolved fine structure.

are required. A second mechanism involves sequential energy transfers between different rare-earth ions. This could involve a simple radiative transfer where the emission of a donor ion is absorbed by an acceptor ion. Such an interaction has an R^{-2} interatomic distance dependence. Alternatively, the process could involve a nonradiative mechanism such as a Forster resonant energy transfer (FRET) or a Dexter transfer. Again, the emission of the donor ion must overlap the absorption profile of the acceptor species. However, the Forster mechanism is mediated by a dipole–dipole interaction between the donor and acceptor, and has a steep R^{-6} distance dependence. Dexter energy transfers are controlled by even higher order multipole interactions or very short-range exchange forces, and therefore are expected to more important in systems with forbidden transitions. In both cases, the donor and acceptor ion concentrations need to be relatively high and the energy transfer can be resonant or phonon assisted. A third mechanism involving cooperative upconversion is possible when the donor and acceptor ions are the same, and the dopant concentration has led to clustering. Here, a resonant dipole-dipole interaction between the two excited ions having an R^{-3} distance dependence cause one ion to be promoted to an even higher excited state, whereas the other is de-excited.

The choice of fiber is important for upconversion lasers. First, they must be transparent at the wavelengths used to pump the ions. Perhaps more importantly, they should have relatively large phonon energies to minimize phonon-assisted nonradiative relaxation of the metastable ions states involved in the upconversion process. Some of the more successful devices have been constructed using fibers made from the five component fluoride glass ZBLAN: (53 mol % ZrF_4; 10 mol % BaF_2; 4.5 mol % LaF_3, 3.5 mol % AlF_3, and 20 mol % NaF) (Jordan, Jha and Ryan, 1993). On the other hand, Si fibers can be problematic because they tend to absorb at wavelengths longer than around 2 μm.

18.3
Diode Lasers

In the solid-state lasers considered above, the energy levels are associated with low concentrations of dopant. Because, under these conditions, the dopant atoms are essentially isolated from each other, their energy levels remain discrete and the emission exhibits the same kind of line spectrum that we associate with isolated atoms or molecules. In the case of semiconductor diode lasers, however, it is energy bands, rather than discrete energy levels that are involved in the laser action.

Solid-state electronic devices generally make use of junctions between *p*-type and *n*-type semiconductors. The most familiar semiconductor materials are Class IV elements such as silicon and germanium; however, binary compounds between Class III and Class V elements, such as gallium arsenide (GaAs), exhibit similar behavior. In the latter cases, the lattice consists of two interpenetrating face-centered cubic lattices, the *p*- and *n*-type crystal reflecting a stoichiometry that departs from precisely 1 : 1. In *p*-type material, some arsenic atoms are replaced by gallium; in *n*-type crystal, the converse applies. On the basis of a junction between these materials, diode lasers operate in a broadly similar way to light-emitting diodes. When an electrical potential is applied across the very thin diode junction formed between *p*- and *n*-type semiconductor layers, the electrons crossing the semiconductor boundary recombine with holes and drop down from the conduction band to the valence band. In doing so, they emit radiation in the process as illustrated in Figure 18.5. The emission is most often in the infrared, and the optical properties of the crystal at such wavelengths make it possible for the crystal end faces to form the confines of the resonant cavity. One advantage of this kind of edge emitter is its extremely small size, usually about half a millimeter. However, the geometry of the gain region does result in very poor beam quality and collimation; divergences of 10° are by no means unusual. However, this does not preclude edge emitters from being used to pump fiber or other solid-state lasers to obtain better beam outputs, albeit, at the expense of overall efficiency.

There are two main types of semiconductor lasers: those operating at fixed wavelengths and others that are tunable. The three most common fixed-wavelength types are gallium arsenide, gallium aluminum arsenide, and indium gallium arsenide phosphide. Gallium arsenide lasers emit at a wavelength of around 0.904 μm; the wide range of flexibility in the stoichiometry of the other types makes it possible to produce a range of lasers operating at various fixed wavelengths in the region 0.8–1.3 μm. “Lead salt” diode lasers, which are derived from nonstoichiometric binary compounds of lead, cadmium, and tin with tellurium, selenium, and sulfur, emit in the range 2.8–30 μm (3500–330 cm^{-1}) depending on the exact composition. Although these lasers require a very low operating temperature, typically in the range 15–90 K, the operating wavelength is very temperature-dependent, and the wavelength can be tuned by varying the temperature. The tuning range for a lead

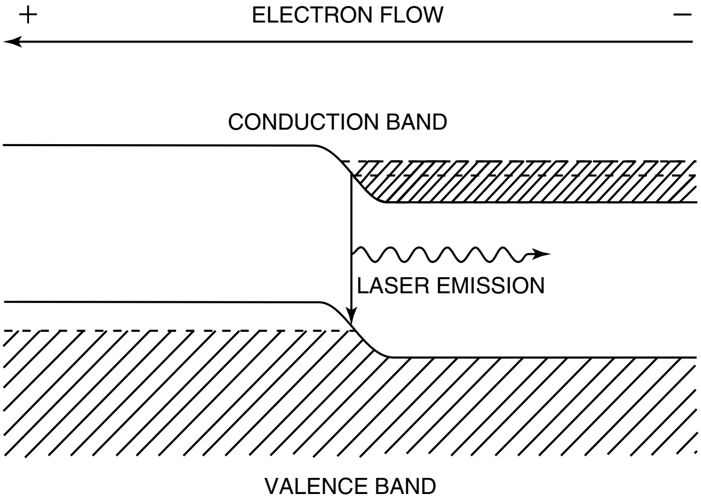

Fig. 18.5 Energetics of a diode laser; energy levels of *p*-type material on the left and *n*-type on the right.

salt diode laser of a particular composition is typically about 100 cm^{-1}.

Modes in a diode laser are typically separated by 1–2 cm^{-1}, and each individual mode generally has a very narrow line width, $10^{-3}\,\mathrm{cm}^{-1}$ or less. The output power of CW diode lasers is generally on the milliwatt scale, though some array devices are capable of emitting as much as 10 W. Diode lasers are very well suited to high-resolution IR spectroscopy, because the line width is sufficiently small to enable the rotational structure of vibrational transitions to be resolved for many small molecules. This method has proved particularly valuable in characterizing short-lived intermediates in chemical reactions. More diverse applications can be expected as the range of diode laser emission wavelengths encroaches further into the visible region – devices emitting at the blue end of the spectrum are now available, in some cases through frequency-doubling methods (Section 7.1).

18.3.1
Vertical Cavity Surface Emitter Lasers

Vertical cavity surface emitter lasers (VCSEL) are diode lasers where the axis of coherent beam propagation is perpendicular to the gain region at the p–n junction. In this geometry, the width of the active area exceeds the height, so that beam divergence due to diffraction is much reduced (Hecht, 2005). However, because the active volume is small, high-reflectivity mirrors must be attached to the ends of the cavity as part of the fabrication process. Alternatively, distributed-Bragg-reflection gratings can be used to deflect light oscillating in the horizontal plane of an edge emitter out through the surface. Grating-coupled surface emitters have also been designed where a second-order grating in the active layer diffracts the light through the surface. The coherent emission from this configuration is bright with good spatial quality.

18.3.2
Visible Laser Diodes

Doped III–V nitride-based semiconductors have a wide direct band gap for blue emission (Leinen *et al.*, 2000). For example, the band gap of AlGaInN can be varied between 2 and 6.2 eV at room temperature depending on its composition. Commercial edge and surface emitting lasers based on InGaN emit light by the recombination of injected electrons and holes in an active layer (Nakamura, 1998). The In allows a strong conduction band to valence band emission at room temperature which otherwise would be difficult to obtain using GaN alone. These devices are expected to have a large impact on the development of full color laser displays, which require at least three primary colors: red, green, and blue to produce any other visible wavelength. Blue-emitting diode lasers also provide an approximately three time enhancement in the capacity of digital versatile disks (DVDs) over those read using red diode lasers. Presently, ''Blu-ray'' discs and ''HD–DVDs'' are competing formats that both capitalize on the use of blue-emitting diode lasers for reading and writing, and for providing digital movies with a higher resolution than conventional DVDs.

18.3.3
Quantum Dot Lasers

Quantum dots are nanometer-sized semiconducting assemblies whose energy levels are fully discrete and atomic-like (Asada, Miyamoto and Suematsu, 1986). The energy level spacings depend on the radius of the dot. Whereas bulk semiconductors are three dimensional, quantum films are two dimensional, and quantum wires are one dimensional, quantum dots are often classified as zero dimensional. Considerable effort has been expended on fabricating arrays of uncapped (no organic surface terminators) InAs quantum dot lasers grown on GaAs that can lase at 1.3 μm (Mowbray, 2005; Fiore *et al.*, 2006). InAs self assembles on a GaAs substrate into a quantum dot array upon epitaxial growth because the two semiconductors have significantly different lattice constants. Although the individual dots have narrow spectral lines, size dispersion in an assembly of dots results in an inhomogeneous broadening of several tens of nanometers. The resultant quantum dot arrays have low gain for stimulated emission due to their lower density of states and large spectral widths. However, this problem can be overcome by stacking several layers of quantum dots within one assembly. On a positive note, these devices have minimal sensitivity to operating temperatures between 20 and 85 °C because their energy level spacings are larger than the Boltzmann energy, kT. They require ultra low threshold current densities for lasing, and the inhomogeneous broadening provides some degree of tunability.

18.3.4
Quantum Cascade Lasers

Quantum cascade laser emission is a result of electronic transitions between the discrete energy levels of a quantum well within the conduction band of a semiconductor. A quantum well is an ultrathin (typically nanometer scale) layer of one semiconductor sandwiched between two layers of a

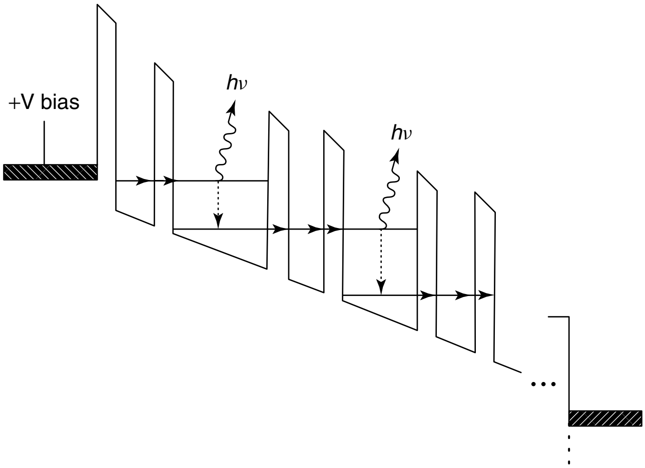

Fig. 18.6 Well structure of a quantum cascade laser.

second semiconductor with a wider band gap. This two-dimensional arrangement effectively confines injected electrons within the well. Because the discrete energy level spacing with the quantum well can be varied by changing its thickness, lasers based on this geometry are tuned between 4 and 24 μm in the mid- to far-infrared.

Although these devices require sophisticated molecular beam epitaxy to be fabricated, they are unique in that unlike other semiconductor devices the laser wavelength of the continuous output is not dependent on the chemical properties or the composition of the medium (Faist *et al.*, 1994). As shown in Figure 18.6, a typical device consists of a series of quantum well. Under an applied voltage, an electron injected into the first quantum well will emit a photon when it drops from higher to lower energy. Because the well is within the conduction band, the electron can tunnel into an identical adjacent well where it produces a second photon. This process is repeated for each well. The effect of multiple quantum wells is that one electron can produce many photons as it cascades down an energy "staircase" created by the applied voltage. This avalanche effect results in output lasers powers which are ~1000 X that of a diode laser.

18.3.5
Quantum Microcavity Lasers

Microcavities are resonators that are characterized by high quality factors, Q, and small volumes, V. Like all resonators, microcavities support discrete optical mode frequencies for photons propagating along its axis. The lowest frequency mode supported by the resonator is called the *fundamental resonator mode.* Because the resonator mode frequency spectrum is size-dependent, microcavities, by virtue of their small dimensions, have mode frequency distributions that are considerably sparser than those for larger cavities.

There are three common geometries associated with microcavities (Vahala, 2003). The first is a Fabry–Perot resonator with two coplanar mirrors separated by one of more multiples of $\lambda/2$ where λ is the emission wavelength. Here, the Q-factor, (which is related to the photon lifetime in the cavity) is determined primarily by the reflectivity of the mirrors and is typically of the order of 2×10^3. The typical volume of a Fabry–Perot microcavity of refractive index n, operating at wavelength λ is $V = 5(\lambda/n)^3$.

The second configuration consists of spherical or a disk–shaped dielectric structures that confine optical surface waves by continuous total internal reflection. In general, the resultant atomic-like optical modes can be classified by three indices: n, ℓ, and m (Righini *et al.*, 2005). The radial mode numbers, n, corresponds to the number of maxima in the radial component, while the angular momentum index, ℓ, is related to the number of wavelengths contained in the equatorial length. The combination $\ell - |m| + 1$ gives the number of maxima in the polar component. The smallest volume modes, called *whispering gallery modes,* are low order radial modes with $\ell \approx |m|$ that orbit in the equatorial ring. Typical Q-factors and volumes for these so-called whispering gallery resonators are 12×10^3 and $6(\lambda/n)^3$, respectively. The Q-factor is limited by losses due to surface roughness.

The third type of microcavity is based on photonic crystals. Photonic crystals are materials with an index of refraction periodicity on the order of λ. The photonic band gap which opens if the index contrast of the periodicity is sufficiently large corresponds to a frequency range where light with wavelengths ∼2λ will not propagate (Yablonovitch, 1987). This is the optical analog of the electronic band gap in semiconductors. However, microcavities can be formed by introducing a defect into the periodic structure (for example by drilling an additional hole or omitting a hole). These cavities can be very small ($V \sim 1.2(\lambda/n)^3$), and their Q-factors can very large ($\sim 13 \times 10^3$) provided there is an adequate spatial extent of the index periodicity. However, photonic crystals can be challenging to fabricate for operation in the visible region of the spectrum.

When an optically active medium is placed inside a microcavity, its spontaneous emission can be dramatically changed because the spontaneous emission rate depends on the optical mode density at the emission wavelength. For example, media in cavities having a fundamental mode frequency that is resonant with its emission frequency will exhibit an enhanced emission rate. This is the so-called Purcell effect. Conversely, a cavity can be designed to suppress spontaneous emission. Similarly, media can experience an enhanced or reduced absorption depending on whether or not the medium's absorption frequency is resonant with a mode of the cavity.

Microcavity physics has led to the development of several laser sources. Lasing has been observed in dye-filled water droplets, dye-doped silica and polystyrene spheres, semiconductor microdisks, vertical cavity micropillars containing a quantum dot, and doped photonic crystals. One unique property of lasers based on microcavities is their low thresholds relative to conventional devices. Indeed, it has been argued that an ideal microcavity laser, where spontaneous emission has been suppressed, is actually thresholdless (Rice and Carmichael, 1994).

18.4
Atomic and Ionic Gas Lasers

The class of lasers in which the active medium is a gas covers a wide variety of devices. Generally, the gas is monatomic, or else it is composed of very simple molecules. In both cases, because laser emission results from optical transitions in free atoms or molecules, usually at low pressures, the emission line width is small. The gas is often contained in a sealed tube, with the initial excitation provided by an electrical discharge, so that in many cases, the innermost part of the laser bears a superficial resemblance to a conventional fluorescent light. The tube can be constructed from various materials, and need not necessarily be transparent; silica is commonly employed – and also beryllium oxide, which has some advantages in its high thermal conductivity. It is quite a common feature to have the laser tube contain a mixture of two gases, one of which is involved in the pump step and the other in the laser emission. Such gas lasers are extremely reliable, and they are the most widely used type for routine purposes.

18.4.1
Helium–Neon Laser

The helium–neon laser was the first CW laser ever constructed, and it was also the first laser to be made available to a commercial market, in 1962. The active medium is a mixture of the two gases contained in a glass tube at low pressure; the partial pressure of helium is approximately 1 mbar, and that of neon 0.1 mbar. The initial excitation, provided by an electrical discharge, primarily serves to excite helium atoms by electron impact. The excited helium atoms subsequently undergo a process of collisional energy transfer to neon atoms; certain levels of helium and neon are very close in energy, so that this transfer takes place with a high degree of efficiency. Because the levels of neon that are so populated lie above the lowest excited states, a population inversion is created relative to these levels, enabling laser emission to occur as shown in Figure 18.7.

Three distinct wavelengths can be produced in the laser emission stage; one visible wavelength, typically of milliwatt power, appears in the red at 632.8 nm, and there are two IR wavelengths of somewhat lower power at 1.152 and 3.391 μm. Following emission, the lasing cycle is completed as the neon undergoes a two-step radiationless decay back down to its ground state. This involves transition to a metastable $2p^5 3s^1$ level, followed by collisional deactivation at the inner surface of the tube. The last step has to be rapid for the laser to operate efficiently; for this reason, the surface/volume ratio of the laser tube has to be kept as large as possible, which generally means keeping the tube diameter small. In practice, tubes are commonly only a few millimeters in diameter. Other very weak transitions are utilized in a 1-mW helium–neon laser emitting at various wavelengths including 543.5 nm in the green; its principal virtue is that it is substantially cheaper than any other green laser.

Helium–neon lasers operate continuously, and despite their low output power, they have the twin virtues of being

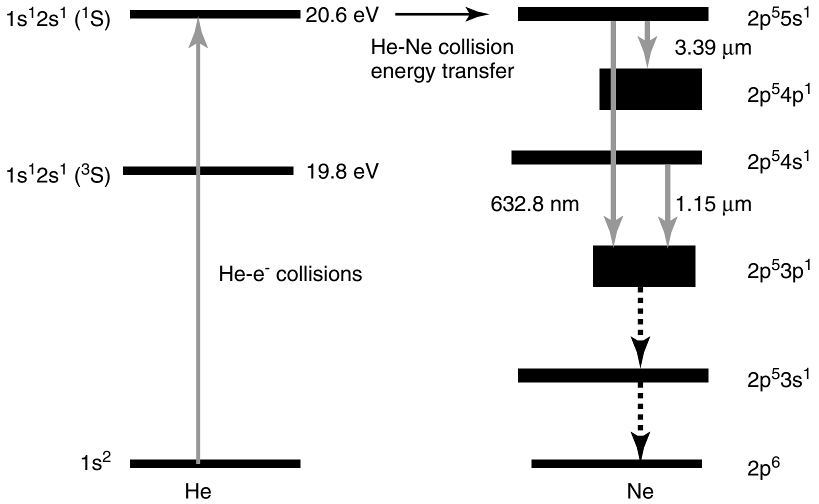

Fig. 18.7 Simplified energy levels involved in He–Ne lasing. The broadness of the levels is to indicate unresolved fine structure.

both small and relatively inexpensive. Consequently, they have more widely ranging applications than any other kind of laser. Their principal applications are those that hinge on the typically narrow laser beam width, where power is not very important. Another laser of similar type is the helium–cadmium laser, in which transitions of cadmium atoms result in milliwatt emission at 442 nm in the blue and 325 nm in the ultraviolet.

18.4.2
Argon and Krypton Lasers

The argon laser is the most common example of a family of ion lasers in which the active medium is a single-component inert gas. The gas, at a pressure of approximately 0.5 mbar, is contained in a plasma tube of 2–3 mm bore, and is excited by a continuous electrical discharge. Argon atoms are ionized and further excited by electron impact; the pumping process populates several ionic states. The population inversion between these states and others of lower energy results in emission at a series of discrete wavelengths over the range 350–530 nm (see Table 18.1), the two strongest lines appearing at 488.0 and 514.5 nm. These two wavelengths are emitted as the result of transitions from the singly ionized states with electron configuration $3s^2 3p^4 4p^1$ down to the $3s^2 3p^4 4s^1$ state. Further radiative decay to the multiplet associated with the ionic ground configuration $3s^2 3p^5$ then occurs, and the cycle is completed either by electron capture or by further impact excitation. Doubly ionized Ar^{2+} ions contribute to the near-ultraviolet laser emission.

Because several wavelengths are produced by this laser, dispersive optics are generally placed between the end mirrors to select one particular wavelength for amplification; the output can thus be varied. By selecting a single longitudinal mode, an output line width as small as 0.0001 cm^{-1} is obtainable. The pumping of ionic levels required for laser action requires a large and continuous input of energy, and the relatively low efficiency of the device means that a large amount of thermal energy has to be dissipated. Cooling is commonly achieved by the circulation of water in a jacket around the tube. The output power of a CW argon laser usually ranges from milliwatt emission up to about 25 W. Such lasers have found widespread research applications in spectroscopy, where they are often employed to pump dye lasers (see Section 18.5.5).

The other common member of the family of ion lasers is the krypton laser. In most respects, it is very similar to the argon laser, emitting wavelengths over the range 350–800 nm, although because of its lower efficiency the output takes place at somewhat lower power levels, up to around 5 W. The strongest emission is at a wavelength of 647.1 nm. In fact, the strong similarity in physical requirements and performance between the argon and krypton lasers makes it possible to construct a laser containing a mixture of the two gases, which provides a very good range of wavelengths across the whole of the visible spectrum – see Table 18.1. Such lasers emit different wavelengths useful for biomedical applications, the blue-green argon lines being especially useful in many biological applications.

Laser operation does not by itself automatically produce polarized light, although the output is commonly at least partially polarized as a result of

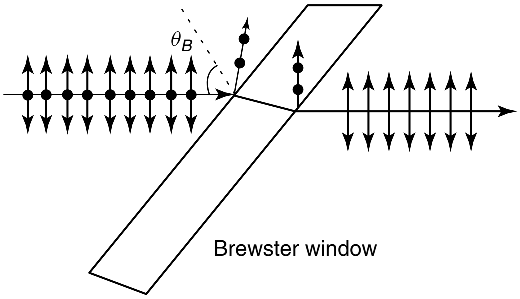

Fig. 18.8 Brewster angle laser window as used in a gas laser. Light polarized in the plane of the figure (double headed arrows) has a low reflection loss at the Brewster angle (θ_B) while light polarized perpendicular to the plane of the figure (indicated by dots) is highly reflected off both surfaces of the window. Only the polarization in the plane of the figure will be strongly amplified in the laser medium.

interactions with the numerous optical components. In these lasers (and others), it is desirable to obtain highly polarized emission, and it is usual to incorporate optics into the laser cavity to accomplish this. The standard method is to use a Brewster angle window. This is a parallel-sided piece of glass set at the Brewster angle, defined as $\theta_B = \tan^{-1} n$, where n is the refractive index of the glass for the appropriate wavelength. The Brewster angle is the angle at which reflected light is completely plane polarized. When such elements are used for the end windows of a gas laser tube, for example, one selected polarization experiences gain and is amplified within the cavity. Polarized emission thus ensues, as shown in Figure 18.8.

An alternative is to make use of an optically anisotropic or *birefringent* crystal such as calcite, in which orthogonal polarizations propagate in slightly different directions and can thus be separated. Calcite is a good example of such a material. Crystals of this type have a number of other important uses in controlling laser polarization; mostly, these stem from the fact that in such media, different polarizations propagate at slightly different velocities. By passing *polarized* light in a suitable direction through such a crystal, the birefringence effects a change in the polarization state, the extent of which depends on the distance the beam travels. With a slice of crystal of the correct thickness, plane polarized light thus emerges circularly polarized, or *vice versa,* and with a crystal of twice the thickness, plane polarized light emerges with a rotated plane of polarization, or circularly polarized light reverses its handedness. The two optical elements based on these principles are known as a *quarter-wave plate* and a *half-wave plate,* respectively. These optical elements can be used intracavity or external to the resonator.

18.5
Molecular Gas Lasers

18.5.1
Carbon Dioxide Laser

In the CO_2 laser, the energy levels involved in laser action are rotation–vibration levels, and emission therefore occurs at much longer wavelengths, well into the IR. The lasing medium consists of a mixture of CO_2, N_2, and He gas approximately in the ratio of partial pressures 1 : 4 : 5; the helium being added to improve the lasing efficiency. The sequence of excitation is illustrated in Figure 18.9. The first step is the excitation, by electron impact, of

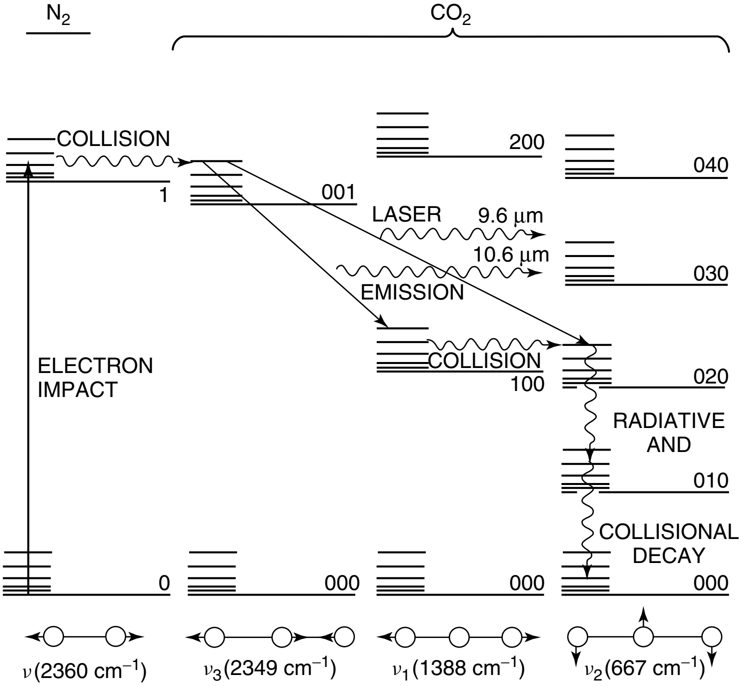
N$_2$
CO$_2$
COLLISION
1
001
LASER 9.6 μm
10.6 μm
EMISSION
200
040
030
COLLISION
100
020
RADIATIVE AND
010
COLLISIONAL DECAY
ELECTRON IMPACT
0
000
000
000
ν(2360 cm^{-1})
ν$_3$(2349 cm^{-1})
ν$_1$(1388 cm^{-1})
ν$_2$(667 cm^{-1})

Fig. 18.9 Energetics of the carbon dioxide laser, rotational structure of the vibrational levels shown schematically. (Adapted from Andrews (1997).)

various metastable rotational sublevels in the first vibrationally excited level of nitrogen. The rotational sublevels of one vibrationally excited state (001) of carbon dioxide (having one quantum of energy in the ν_3 vibration, the antisymmetric stretch) has almost the same energy as the vibrationally excited nitrogen. Consequently, collision between the two molecules results in very efficient transfer of energy to the CO$_2$. Laser emission then occurs by two routes, involving radiative decay to sublevels of the (100) and (020) states. The two laser transitions result in emission in bands centered on wavelengths at around 10.6 and 9.6 μm, respectively. Because various rotational sublevels can be involved in the emissive transitions, use of suitable optical elements enables the laser to be operated at various discrete frequencies within the 10.6 and 9.6 μm bands as listed in Table 18.2. Because the CO$_2$ energy level positions vary for different isotopic species, other wavelengths may be obtained by using isotopically substituted gas.

A small carbon dioxide laser, with a discharge tube about half a meter in length, may have an efficiency rating as high as 30%, and produce a continuous output of 20 W; even a battery-powered hand-held model can produce 10 W CW output. Much higher powers are available from longer tubes; although efficiency drops, outputs in the kilowatt range are obtainable from the largest devices. Apart from extending the cavity

Tab. 18.2 Emission of the carbon dioxide laser. The lines (wave number units $\tilde{\nu}$) result from rotation–vibration transitions, the rotational quantum number J changing by one unit. Transitions $J \rightarrow J - 1$ are denoted as P(J); transitions $J \rightarrow J + 1$ are denoted as R(J).

Line	$\tilde{\nu}$ (cm^{-1})	Line	$\tilde{\nu}$ (cm^{-1})	Line	$\tilde{\nu}$ (cm^{-1})
10.6 μm *band* (001 → 100 *transitions*)					
P(56)	907.78	P(22)	942.38	R(18)	974.62
P(54)	910.02	P(20)	944.19	R(20)	975.93
P(52)	912.23	P(18)	945.98	R(22)	977.21
P(50)	914.42	P(16)	947.74	R(24)	978.47
P(48)	916.58	P(14)	949.48	R(26)	979.71
P(46)	918.72	P(12)	951.19	R(28)	980.91
P(44)	920.83	P(10)	952.88	R(30)	982.10
P(42)	922.92	P(8)	954.55	R(32)	983.25
P(40)	924.97	P(6)	956.19	R(34)	984.38
P(38)	927.01	P(4)	957.80	R(36)	985.49
P(36)	929.02	R(4)	964.77	R(38)	986.57
P(34)	931.00	R(6)	966.25	R(40)	987.62
P(32)	932.96	R(8)	967.71	R(42)	988.65
P(30)	934.90	R(10)	969.14	R(44)	989.65
P(28)	936.80	R(12)	970.55	R(46)	990.62
P(26)	938.69	R(14)	971.93	R(48)	991.57
P(24)	940.55	R(16)	973.29	R(50)	992.49
9.6 μm *band* (001 → 020 *transitions*)					
P(50)	1016.72	P(20)	1046.85	R(16)	1075.99
P(48)	1018.90	P(18)	1048.66	R(18)	1077.30
P(46)	1021.06	P(16)	1050.44	R(20)	1078.59
P(44)	1023.19	P(14)	1052.20	R(22)	1079.85
P(42)	1025.30	P(12)	1053.92	R(24)	1081.09
P(40)	1027.38	P(10)	1055.63	R(26)	1082.30
P(38)	1029.44	P(8)	1057.30	R(28)	1083.48
P(36)	1031.48	P(6)	1058.95	R(30)	1084.64
P(34)	1033.49	P(4)	1060.57	R(32)	1085.77
P(32)	1035.47	R(4)	1067.54	R(34)	1086.87
P(30)	1037.43	R(6)	1069.01	R(36)	1087.95
P(28)	1039.37	R(8)	1070.46	R(38)	1089.00
P(26)	1041.28	R(10)	1071.88	R(40)	1090.03
P(24)	1043.16	R(12)	1073.28	R(42)	1091.03
P(22)	1045.02	R(14)	1074.65	R(44)	1092.01

length, another means used to increase the output power is to increase the pressure of carbon dioxide. Carbon dioxide lasers can be made to operate at or above atmospheric pressure. To produce sufficiently strong fields without using dangerously high voltages, it is necessary to apply a potential across, rather than along the tube; such a laser is referred to as a *transverse excitation atmospheric* (*TEA*) *laser*. In devices where the gas pressure is raised to around 15 atmospheres, pressure broadening results in the formation of a quasi-continuum of emission frequencies, enabling the laser output to be continuously tuned over a range of approximately 910–1100 cm^{-1}.

18.5.2
Nitrogen Laser

The nitrogen laser operates on electronic, rather than vibrational transitions. A high-voltage electrical discharge populates the $C^3\Pi_u$ triplet electronic excited state, and the laser transition is to the lower energy metastable $B^3\Pi_g$ state. The upper laser level has a very short lifetime of only 40 nanoseconds, and consequently population inversion cannot be sustained; essentially all the excited nitrogen molecules undergo radiative decay together over a very short period of time, effectively emptying the cavity of all its energy. This *super-radiant* emission is sufficiently powerful that a highly intense pulse is produced without the need for repeatedly passing the light back and forth between end mirrors. The laser automatically operates in a pulsed mode, producing pulses of about 10 nanoseconds or less duration at a wavelength of 337.1 nm. Bandwidth is approximately 0.1 nm, and pulse repetition frequency is 1–200 Hz. Pulses tend to have a relatively unstable temporal profile as a result of the low residence time of photons in the laser cavity. Because the laser can produce peak intensities in the 10^{10} Wm^{-2} range, the nitrogen laser is one of the most powerful commercial sources of ultraviolet radiation, and it is frequently used in photochemical studies. It has also been a popular choice as a pump for dye lasers, although here it is now often displaced by the more powerful wavelengths from excimer lasers or the second or third harmonics of the Nd : YAG laser.

18.5.3
Iodine Laser

Another laser operating on broadly similar principles is the iodine laser, more correctly termed as the *atomic iodine photodissociation laser*. Unusually, polyatomic chemistry is involved in its operation, but the crucial laser transition takes place in free atomic iodine, so that it is arguable whether or not it is truly designated as a molecular laser. The laser is activated by the photolysis of an iodohydrocarbon or iodofluorocarbon gas such as 1-iodoheptafluoropropane, C_3F_7I, by ultraviolet light from a flashlamp. The gas is introduced into a silica laser tube at a pressure of 30–300 mbar; the ensuing mechanisms are illustrated in Figure 18.10. Laser emission involves transition between an excited metastable $^2P_{1/2}$ state and the ground $^2P_{3/2}$ state of atomic iodine; this produces narrow line width output at a wavelength of 1.315 μm (7605 cm^{-1}), consisting of closely spaced hyperfine components encompassing a range of less than 1 cm^{-1}. In the absence of any other pulsing mechanism, the laser typically produces pulses of microsecond duration, and pulse energies of several joules; however the output is often modified by Q-switching or mode-locking (see Section 8) to produce pulse trains of nanosecond or subnanosecond duration.

An important advantage of the iodine laser is the fact that the active medium is comparatively cheap and available in large quantities. Moreover, its emission wavelength lies in a region offering excellent atmospheric transmission. One variation is a device known as the *chemical oxygen iodine laser*, in which collisional energy transfer occurs from electronically excited ($^1\Delta$) oxygen

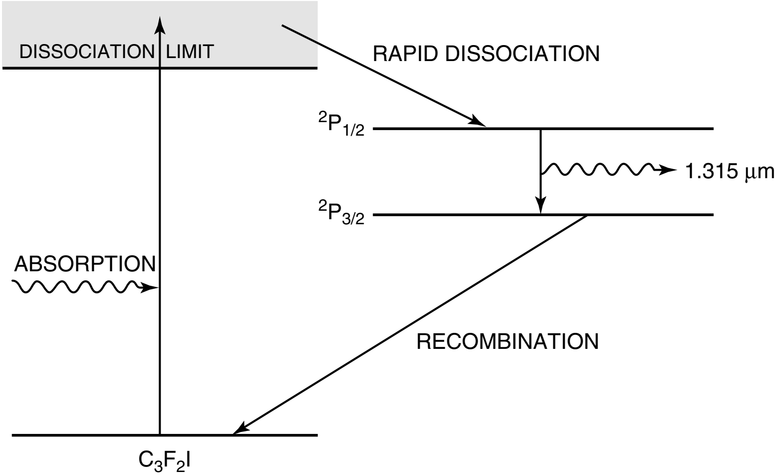

Fig. 18.10 Energetics of the atomic iodine photodissociation laser (not showing fine structure or side reactions).

molecules to atomic iodine, here formed by dissociation of molecular iodine, populates the $^2P_{1/2}$ state and so leads to the same 1.315 µm laser emission. By employing supersonically flowing gases, multikilowatt CW outputs can be achieved.

18.5.4
Excimer Lasers

In excimer lasers, the active medium is an *exciplex* or an excited complex. The crucial feature of an exciplex is that only when it is electronically excited, does it exist in a bound state with a well-defined potential energy minimum; the ground electronic state generally has no potential energy minimum, or else only a very shallow one. Most examples are found in the inert gas halides such as KrF. In principle, it is only homonuclear diatomics, such as Xe_2, that should be termed *excimers*, but in connection with lasers the designation is widely applied to heteronuclear systems. The exciplex is generally formed by chemical reaction between inert gas and halide ions produced by an electrical discharge. In the case of krypton fluoride, KrF, for example, the exciplex has a lifetime of about 2.5 nanoseconds. The decay transition can be stimulated to produce laser emission with a high efficiency, typically around 20%.

Excimer lasers are super-radiant, producing radiation with pulse durations of 10–20 nanoseconds and pulse repetition frequencies generally in the 1–500 Hz range. Pulse energies can be up to 1 J, with peak powers in the milliwatt region and average power between 20 and 100 W. The emission wavelengths of commercially available systems are: F_2 157 nm; ArF 193 nm; KrCl 222 nm; KrF 248 nm; XeCl 308 nm; XeF 351 and 353 nm. These short ultraviolet wavelengths lie in a region of the electromagnetic spectrum absorbed by a wide range of materials; photoabsorption often results in the rupture of chemical bonds and some degree of sample vaporization.

18.5.5
Dye Lasers

Dye lasers are essentially different from every other common category of laser. Major differences arise due to the unusual nature of the active medium, a solution of a chemically complex organic dye. A wide range of more than 200 dyes can be used for this purpose; the only general requirements are an absorption band in the visible and/or the ultraviolet, and a broad fluorescence spectrum. The kind of compounds that best satisfy these criteria consist of comparatively large polyatomic molecules with extensive electron delocalization. One of the most widely used systems is the dye known as *Rhodamine* 6G ($C_{28}H_{31}N_2O_3Cl$), whose broad absorption and fluorescence are shown in Figure 18.11.

Generally, the absorption of visible light first results in a transition from the ground singlet state S_0 to the energy continuum belonging to the first excited singlet state S_1. This is immediately followed by rapid radiationless decay to the lowest energy level within the S_1 continuum, as illustrated in Figure 18.12; in the case of Rhodamine 6G, this process is complete within 20 picoseconds of the initial excitation. Fluorescent emission then results in a downward transition to levels within the S_0 continuum, followed by further radiationless decay. It is the fluorescent emission process that has made the basis of laser action, provided a population inversion is set up between the upper and lower levels involved in the transition; this is essentially a four-level laser. Clearly, because the emitted photons have less energy, the fluorescence must always occur at longer wavelengths than the initial excitation.

One significant loss mechanism is quenching of the excited states through interaction with other molecules; triplet quenchers such as dimethylsulfoxide (DMSO) are often added to increase the output power by repopulating singlet

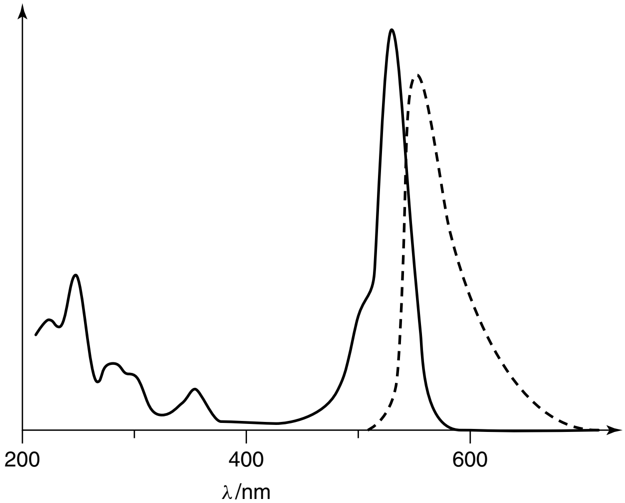

Fig. 18.11 Solution spectra of Rhodamine 6G in ethanol. The solid curve shows absorption and the dotted curve shows the fluorescence at longer wavelength.

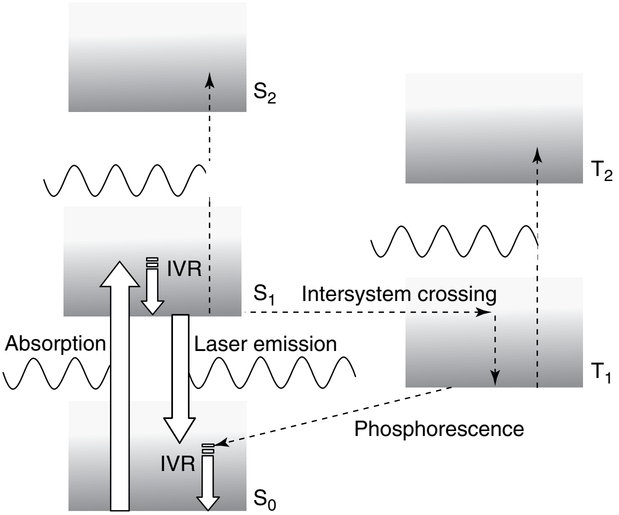

Fig. 18.12 Jablonski diagram for a laser dye. Solid vertical arrows show transitions involved in laser action; dotted arrows illustrate some competing transitions. IVR: intramolecular vibrational redistribution.

states. Another problem is the heat produced by the radiationless decay transitions, a necessary part of the laser cycle. This can rapidly degrade a dye, and therefore it is common practice to continuously circulate the dye solution, to facilitate cooling. A common setup is shown in Figure 18.13. The pump radiation from a flash-lamp, or primary laser source with emission in the visible or near-ultraviolet range, is focussed at a point traversed by a jet of dye solution, which typically has a concentration in the range 10^{-2}–10^{-4} mol 1^{-1}. The solvent is usually based on ethylene glycol, which provides the viscosity needed to maintain an optically flat jet stream.

Fluorescent emission from the dye jet is stimulated by the formation of an optical cavity, with two parallel end mirrors placed on either side of the jet. However, because the fluorescence occurs over a range of wavelengths, monochromatic laser emission can only be obtained by the use of additional dispersive optics within the cavity; rotation of this element can change the wavelength amplified within the dye laser, so that *tunable emission* is obtained. A dye laser based on a solution of Rhodamine 6G in methanol, for example, is continuously tunable over the range 570–660 nm. The full range of commercially available dye lasers provides complete coverage of the range 200 nm–1 μm: typical tuning curves for some of the more important dyes are shown in Figure 18.14.

The dye laser is often used as a means of providing continously tunable radiation Commonly, an inert gas or nitrogen laser is used as the pump, and the output is tunable over a range of wavelengths (determined by the choice of dye), all longer than the pump. The facility for tuning provides a very useful method of obtaining laser emission

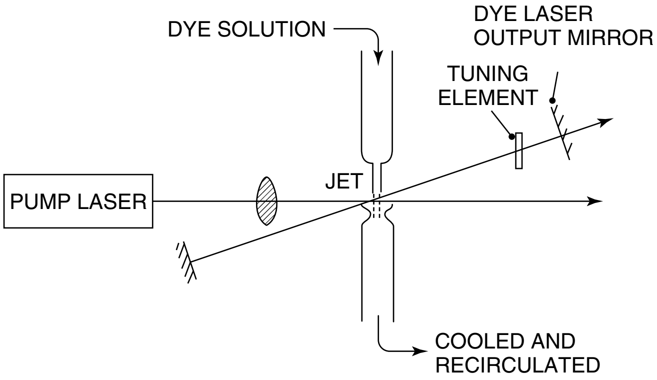

Fig. 18.13 Schematic diagram of a laser-pumped dye laser. (Adapted from Andrews (1997).)

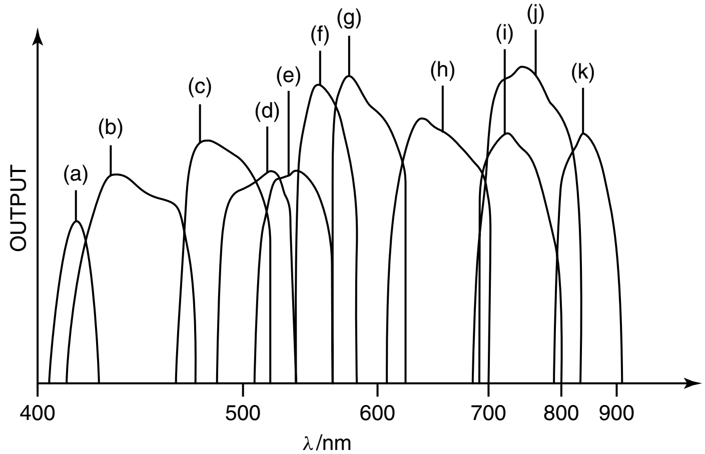

Fig. 18.14 Tuning curves for a dye laser pumped by lines from a krypton/argon laser. The dye and pump wavelength for each curve are given in Table 18.3. (Reproduced by kind permission of Coherent Radiation Ltd.)

at a chosen wavelength. Conversion efficiencies are generally somewhat low; 5–10% is common, although with certain dyes figures of 20% or more are possible. The highest powers are obtained by pumping with harmonics of a Nd laser. CW dye lasers produce emission with line width in the range 10–20 GHz (around 0.5 cm^{-1}), although with suitable optics this can be reduced to about 1 GHz. The combination of narrow line width, good frequency stability and tunability is particularly attractive for spectroscopic applications. One disadvantage of the dye laser is that it tends to have rather poorer amplitude stability than a gas laser, so that indirect spectroscopic methods such as fluorescence or the photoacoustic effect are often the most appropriate.

Tab. 18.3 Dyes and pump wavelengths for the tuning curves of Figure 18.14.

Curve	Dye	Pump laser; wavelength λ (nm)
(a)	Stilbene 1	Argon; 333.6, 351.1, 363.8
(b)	Stilbene 3	Argon; 333.6, 351.1, 363.8
(c)	Coumarin 102	Krypton; 406.7, 413.1, 415.4
(d)	Coumarin 30	Krypton; 406.7, 413.1, 415.4
(e)	Coumarin 6[a]	Argon; 488.0
(f)	Rhodamine 110	Argon; 514.5
(g)	Rhodamine 6G	Argon; 514.5
(h)	DCM[b]	Argon; 488.0
(i)	Pyridine 2	Argon; 488.0, 496.5, 501.7, 514.5
(j)	Rhodamine 700	Krypton; 647.1, 657.0, 676.5
(k)	Styryl 9M	Argon; 514.5

[a] With cyclo-octatetraene and 9-methylanthracene.

[b] 4-dicyanomethylene 2-methyl 6-(p-dimethylaminostyrol) 4H-pyran.

One important variation on the concept is the *ring dye laser*, in which the laser radiation travels around a cyclic route between a series of mirrors, rather than simply back and forth between two. When both clockwise and anticlockwise traveling waves are present, their frequencies are normally identical. However, any rotation of the laser itself results in a small difference between these two frequencies, and detection of this difference can be used as the basis for very accurate measurement of the rotation; this is the principle behind the ring laser gyroscope. Alternatively, an optical diode can be placed within the cavity to select one particular direction of propagation (clockwise or anticlockwise). In this case, the ring laser has an emission line width typically at least a factor of ten smaller than a conventional dye, but in state-of-the-art actively stabilized cases the line width may be as small as 10^{-15} cm^{-1}.

18.6
Pulsing Techniques

There are several reasons why it is common practice to make use of a pulsing device, either to convert a CW laser to pulsed output or to shorten the duration of the pulses emitted by a naturally pulsed laser. One objective is to obtain high peak intensities, to optimize multiphoton or other nonlinear optical processes. Another motive is to produce pulses of very short duration to make measurements of processes that occur on the same timescale. There are three widely used pulsing methods, reviewed below.

18.6.1
Cavity Dumping

Cavity dumping refers to a method of rapidly emptying a laser cavity of the energy stored within it. The simplest means of achieving this with a CW laser would be to have both of the end mirrors fully reflective, but with a third coupling mirror able to be switched into the beam to reflect light out of the cavity. This configuration would deliver a single pulse of light, terminating laser action until the coupling mirror was again switched out of the beam. In practice, an acousto-optic modulator is usually employed. This double-pass device is driven by a radio-frequency electric field, and generates an acoustic wave that produces off-axis diffraction of the laser beam. The cavity-dumped output of the laser in this case consists of an essentially sinusoidal temporal profile; the frequency of oscillation is twice the acoustic frequency of the cavity dumper, typically of the order of megahertz. This method of pulsing is often used in conjunction with mode locking (Section 18.6.3).

18.6.2
Q-Switching

The term *Q-switching* refers to a method that essentially involves first reducing, and then swiftly increasing the quality, or Q-factor of the laser, formally defined as $Q = 2\pi \times$ energy stored in cavity/energy loss per optical cycle (where the optical cycle length is the inverse of the resonant frequency). Q-switching thus represents the effect of suddenly reducing the rate of energy loss within the cavity. In practice, the pumping rate has to exceed the rate of spontaneous decay to build the required population inversion, and the time taken for Q-switching has to be sufficiently short to produce a single pulse. There are several commonly used methods. One of the most common is an electro-optical method based on a Pockels cell, which rapidly switches the cavity transmissivity of a selected polarization, effecting a shutter action when voltage is applied. A Pockels cell consists of a crystalline material such as potassium dihydrogen phosphate (KDP) that exhibits the Pockels effect, essentially a proportional change in refractive index on application of an electric field. A Pockels cell can be used to rotate the plane of polarization of light passing through it, or to reverse the handedness of circularly polarized light. However, its particular advantage is that, because the effect takes place only while the electric field is applied, modulation of the field at a suitable frequency results in the corresponding modulation of the beam polarization. The Q of the resonator can be *actively* switched from low to high when the Pockel cell is used in conjunction with a polarizer by applying a voltage to the Pockel cell. Pockel cells are also a particularly useful for increasing detection sensitivity in many techniques based on optical activity, such as polarimetry and circular differential Raman scattering.

Another common method for Q-switching involves the use of a *saturable absorber* dye such as cryptocyanine. With a cell containing a saturable absorber solution placed inside the laser cavity, the intensity of emission increases to a point where the dye is bleached only after a large population inversion has been achieved in the active medium; stimulated emission therefore results in the emission of a pulse of light. This last method of Q-switching is known as *passive*, because the switching time is not determined by external constraints. Q-switching generally produces pulses with a duration of the order of 10^{-9}–10^{-8} s.

18.6.3
Mode-Locking

The third, widely used method of producing laser pulses is known as *mode locking*. This produces pulses of much shorter duration than those produced by either cavity dumping or Q-switching, the output typically being measured on the picosecond or femtosecond scale and referred to as *ultrashort*. The technique has very important applications in chemistry in the study of ultrafast reactions, for example.

The mechanism of mode locking is as follows. Laser emission generally occurs over a small range of line frequencies within the emission line width of the active medium, separated $\Delta\nu = c/2L$, (the free spectral range, determined by the cavity length L), which is typically around 10^8 Hz. In the case of solid-state or dye lasers, where the fluorescence is broad, the number of longitudinal modes N sustained within the laser cavity may thus be as many as $10^3 - 10^4$ if frequency-selective elements are absent. Mode locking establishes a phase relationship that leads to constructive interference between all the modes at just one point, with destructive interference everywhere else. Consequently, a single pulse of travels back and forth between the end mirrors, giving a shot of output each time it encounters the output mirror. The pulse duration (the pulse-width at half power) is given by $\Delta t = 4\pi L/(2N+1)c$, and the interval between successive pulses is the cavity round-trip time. Hence, mode-locked pulses are typically separated by an interval of around 10 nanoseconds, and have a duration of $1 - 10$ picoseconds; the greater the number of modes, the shorter the pulse length, in accordance with Fourier principles. The temporal

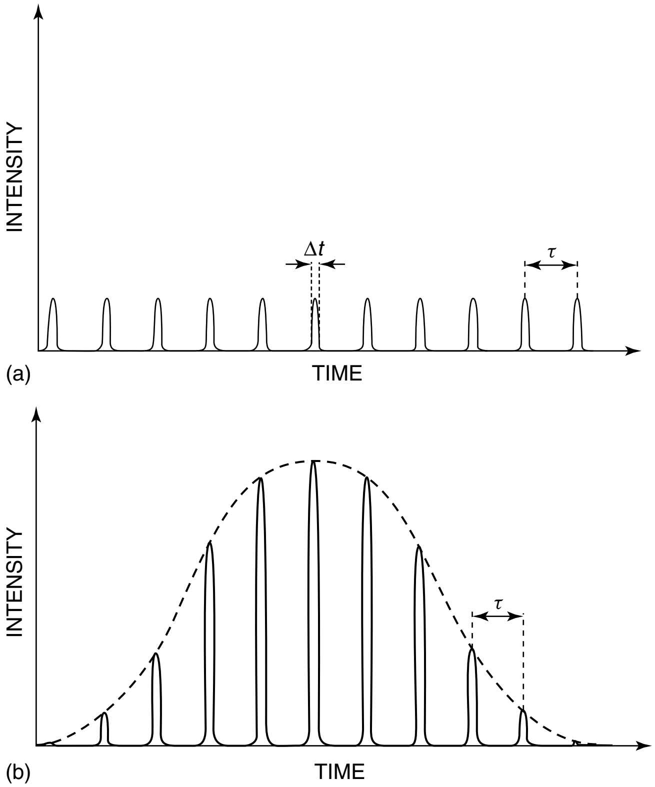

Fig. 18.15 Temporal profile of mode-locked emission from: (a) CW laser; (b) pulsed laser. In the second case, several pulses than illustrated would normally be found within the overall pulse envelope.

profile of the emission from mode-locked CW and pulsed lasers is as shown in Figure 18.15, and in each case consists of a pulse train. The instantaneous intensity of a mode-locked laser pulse can be readily computed on the assumption that the overall energy emitted within the duration of one laser round-trip is the same before and after mode-locking. Thus, with a pulse-length reduction from 10 nanoseconds to 1 picosecond, the pulse power can be increased by a factor of approximately 10^4, increasing the power of GW pulses up to the region of 10^{13} W.

As with Q-switching, there are several means of accomplishing mode-locking, both active and passive methods. The active methods generally involve modulating the gain of the laser cavity using an electro-optic or acousto-optic switch driven at the frequency $c/2L$. Passive mode-locking is generally accomplished with a saturable absorber cell – although, in the exceptional case of the titanium : sapphire laser, automatic mode-locking can be observed without the inclusion of any such element.

Another mode-locking technique principally employed with tunable dye lasers is *synchronous pumping*. Here, a mode-locked primary laser pumps a dye laser whose optical cavity length is adjusted to be equal to that of the pump, so that the round-trip time of the dye laser matches the interval between pump pulses. In this way, frequency-converted pulses propagating within the dye laser cavity arrive at the dye jet at the same rate as the pump pulses. The period over which amplification is effective in the dye laser is thus very short, and once again picosecond pulses emerge. In practice, the dye laser cavity length needs to be kept constant to within a few microns, so that special mounting and temperature control are necessary. Finally, there are various means by which pulse lengths can be reduced, as for example in *colliding-pulse* ring dye lasers.

18.6.4
Ultrafast Pulse Compression

Mode-locking commercially available lasers in conjunction with pulse compression techniques can now provide pulses of a few picoseconds duration in the infrared, and ultrashort pulses measured on the femtosecond (10^{-15}s) timescale in the UV/visible range. A common approach in generating ultrashort pulses is to first pass a longer pulse (on the order of 1–100 picoseconds) through an optical fiber to impose a chirp on the pulse carrier, and then compress the output using a dispersive delay line often made from two gratings (Figure 18.16).

A chirped pulse is one in which the wavelength of the carrier changes continuously throughout the pulse. A chirp

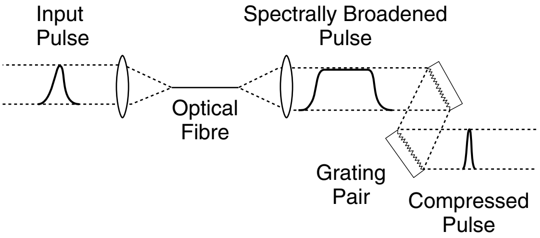

Fig. 18.16 Experimental arrangement used for pulse compression.

can occur in an optical fiber due to the nonlinear process of self-phase modulation, where the index of refraction of the fiber is instantaneously modified in the presence of an intense light pulse through the nonlinear optical Kerr effect (Monerie, 1989). This variation in the index of refraction produces a shift in the instantaneous phase of the pulse. The effect is that as the pulse passes through the fiber, its leading edge experiences a shift to lower frequencies, while its trailing edge experiences a shift to higher frequencies. In a normally dispersive medium, the front of the pulse will move faster than the back of the pulse, resulting in both a spectral and temporal broadening. The effect of the grating pair after the fiber is to produce a phase shift that is opposite to the chirp (Treacy, 1969). This compresses the output of the fiber. Geometrically, the phase shift is a result of time delays that arise because the optical path length due to Bragg diffraction off the gratings increases with increasing wavelength.

Early work using a fiber/grating pair combination yielded 30-femtosecond pulses at 619 nm, corresponding to 14 optical cycles (Shank *et al.*, 1982). More sophisticated arrangements involving prisms and diffraction gratings have led to pulses as short as 6 femtoseconds at 620 nm (Fork *et al.*, 1987). Recently, chirped mirrors have been developed, which have the same function as a grating pair (Steinmeyer, 2006). These are dielectric mirrors multilayered with low and high index coatings whose thicknesses increase monotonically from the surface to the substrate. The reflectivity of such a structure takes place at a depth within the mirror stack that depends on wavelength. This imposes a negative group delay on the pulse that compensates for any material dispersion in the laser system leading to ultrashort pulse generation.

18.6.5
Continuum Generation

It is a notable feature of ultrashort (picosecond) laser pulses that, on passage through certain media, a broad continuum of light is generated, typically covering a range of about 10 000 cm^{-1} of the laser frequency. The facility to produce such a broad continuum of light with such short duration has a number of distinctive uses, as for example in the flash photolysis study of ultrafast reactions. The light produced by this method, which results from a mechanism known as *self-phase modulation,* is referred to as an *ultrafast supercontinuum laser source* or as *picosecond continuum.* A typical spectral profile is shown in Figure 18.17. Such a source is an invaluable component of the instrumentation for ultrafast pump-probe spectroscopy.

18.7
Short Wavelength Coherent Radiation

18.7.1
Frequency Conversion

In tailoring the wavelength of laser output for different applications, a variety of methods can be utilized. One method involves the use of dye lasers, as discussed above. Another is Raman shifting, in which the stimulated Raman effect is employed for conversion to either shorter or longer wavelength. The latter process is

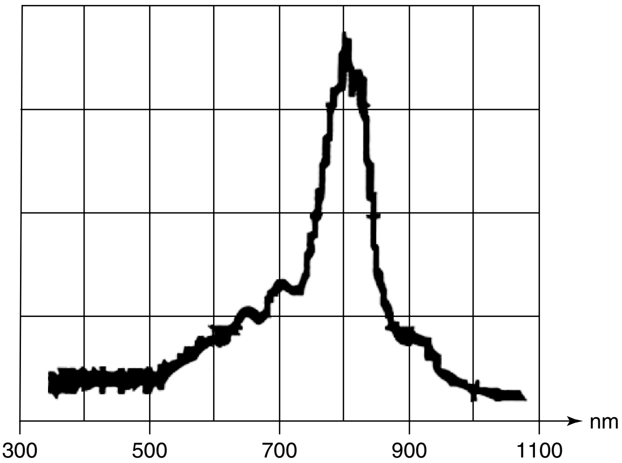

Fig. 18.17 Typical spectrum of the supercontinuum generated by a titanium : sapphire laser, drawn on a logarithmic vertical scale. (Redrawn from an original, courtesy of Coherent Laser Group.)

usually accomplished by the passage of the laser light through a suitable crystal or a stainless steel cell containing gas at a pressure of several atmospheres. Conversion efficiency for the principal Stokes shift to longer wavelength can be as high as 35%. The nature of the crystal or gas determines the frequency increment; $Ba(NO_3)_2$ gives a shift of 1047 cm^{-1}, while the most commonly used gases H_2, D_2, and CH_4 produce shifts of 4155, 2987, and 2917 cm^{-1} respectively. However, by far the most widely adopted method of frequency conversion involves frequency doubling.

18.7.2
Frequency Doubling

Frequency doubling is the best known example of a *nonlinear optical* process. The latter term refers to a wide range of frequency-conversion effects that are nonlinearly dependent on laser intensity, such that their conversion efficiencies are generally improved when laser power is increased. The energetics of frequency doubling (second harmonic generation) are illustrated in Figure 18.18.

Two photons of laser light with frequency ν are assumed to be absorbed by a substance in its ground state, and a single photon of frequency ν' is emitted through a return transition to the ground state. However, this process is parametric. The high intensity of electric field of the incident laser forces the electrons to oscillate in the ground state at the harmonic frequency. This induced nonlinear polarization is a source of *coherent* emission at the doubled frequency. As frequency doubling is a concerted process; there is no intermediate excited state with a measurable lifetime; hence, the energy-time uncertainty principle allows the process to take place regardless of whether there are energy levels of the substance at energies $h\nu$ or 2 $h\nu$ above the ground state. Indeed, it is usually better if there are not, because the presence of such levels can lead to competing absorption processes. Symmetry

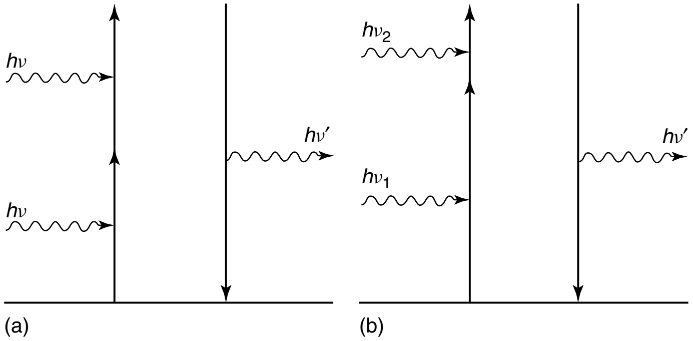

Fig. 18.18 Energetics of (a) frequency doubling, where the emitted photon has frequency $\nu' = 2\nu$, and (b) frequency mixing, where the emitted photon has frequency $\nu' = \nu_1 + \nu_2$.

considerations reveal that the frequency-doubling process can take place only in media that lack a center of symmetry – generally a noncentrosymmetric crystal is employed. Although liquids can very weakly generate second harmonics where their bulk symmetry is broken, a feature that can incidentally prove useful for the characterization of liquid surfaces, they are ineffective for efficient frequency conversion.

One singular advantage of harmonic generation is that it is a *coherent* process, so that with laser input the emission also has typically laserlike characteristics even though no population inversion was created in the material. To sustain coherence, however, it is necessary for the conversion process to conserve photon momentum. For harmonic emission in the forward direction, momentum is conserved only if the harmonic photons have wave vector $\mathbf{k}' = 2\mathbf{k}$, where $\mathbf{k}$ is the wave vector of the laser photons; this leads to a requirement for the refractive indices of the pump beam and the harmonic to be equal. Fulfillment of this wave-vector matching condition can only be accomplished in birefringent solids where refractive index is a function of the polarization and direction of propagation. Here, it is often possible to obtain index matching by careful orientation of the crystal. Refractive indices are often very heat sensitive, and fine control of the temperature is normally required; the crystal is accordingly mounted inside a temperature-controlled heating cell. Either the crystal orientation or its temperature can be varied for any particular incident wavelength, to achieve maximum efficiency. Conversion efficiencies may be as high as 20–30%, and under optimum conditions, it is still higher. Frequency doubling is particularly useful for generating powerful visible (532 nm) radiation using an infrared Nd laser pump. Two of the most widely used crystals are KDP (KH_2PO_4) and lithium niobate (Li_3NbO). The better choice for tunable UV generation is β-barium borate, with a transparency range of 190–3500 nm.

The generally high conversion efficiency of frequency doubling makes it possible to use a series of crystals to produce $4\times, 8\times, 16\times, \ldots$, the pump frequency, and so to obtain coherent radiation at very short wavelengths. Similarly, third harmonic generation (which can take place in either centrosymmetric or noncentrosymmetric media) can be used to generate 355 nm radiation from Nd : YAG sources. More commonly, however, suitable crystals can be used to generate odd harmonics (3×, 5×, 7×, etc.) by two-colour mixing (see below). Indeed 355 nm is usually generated in Nd: YAG systems by mixing 532 nm +1.06 micron light. For some spectroscopic applications, it is also significant, that, a doubling crystal coupled with a dye laser provides tunable emission over a range of wavelengths, below the operating wavelength of the pump.

18.7.3
Optical Parametric Oscillators

In frequency mixing, a variation of the frequency-doubling process permits a coupling between two beams of laser radiation with different frequencies to produce output at the sum frequency, as illustrated in Figure 18.18(b). Because wavelength is inversely proportional to frequency, the emitted wavelength is given by $1/\lambda' = 1/\lambda_1 + 1/\lambda_2$, hence shorter than either of the input wavelengths. A closely related application of nonlinear optics concerns the *optical parametric oscillator*, essentially a difference-frequency technique. Often the third harmonic at 355 nm from a Nd : YAG laser pumps the nonlinear crystal, placed along with suitable frequency-selective elements in a secondary cavity (as in a dye laser). The result is that two new frequencies (the signal and idler frequency) are formed whose sum equals that of the pump. Wide tunability and high conversion efficiencies can be obtained in the IR by this method because phase matching for a specific wavelength can be achieved by rotating the nonlinear crystal to specific angles with respect to the optic axis. The same principles can be applied in the visible and UV by the adoption of a different pump, such as one of the other harmonics of a Nd : YAG laser. For example, pumping beta-barium borate crystal, with fourth-harmonic 266 nm radiation, can produce tunable output covering the range 300 nm to 2.5 μm.

18.7.4
Frequency Tripling and Four-Wave Mixing

Higher order frequency mixing can be employed as means of generating shorter wavelength radiation. Two methods, which exploit third-order nonlinearities in centrosymmetric media (media where second order effects are absent), are nonresonant third harmonic generation, THG, and two-photon resonance enhanced four-wave mixing (Lipson, Shi and Lacey, 2002). The energy level schemes for these processes are shown in Figure 18.19. Radiation in the vacuum ultraviolet (VUV, $100\,\text{nm} \leq \lambda \leq 200\,\text{nm}$) or the extreme ultraviolet (XUV, $\lambda < 100\,\text{nm}$) can be obtained by using visible and UV fundamentals. Although technically the radiation produced by THG or four-wave mixing is not the result of a population inversion, the radiation has all the characteristics of a laser; namely, it is coherent, directional, monochromatic, and, even more significant for molecular spectroscopy, is widely tunable. There are no other "real" VUV or XUV laser sources that are competitive in this regard.

Nonresonant THG arises when intense lasers pulses at a frequency ν induce a nonlinear polarization in an atomic medium that acts as a source of radiation at 3ν. Solids and liquids

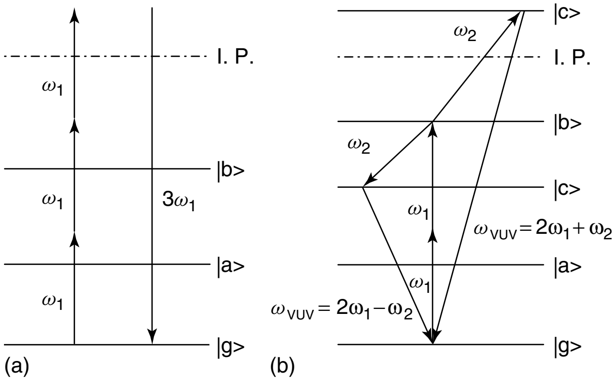

Fig. 18.19 Energetics of (a) frequency tripling, where the emitted photon has frequency $\omega' = 3\omega_1$, and (b) two-photon resonance-enhanced four-wave mixing, where the emitted photon has frequency $\omega' = 2\omega_1 \pm \omega_2$.

are usually never used as they become strongly absorbing below 200 nm. In almost all cases, pulsed lasers are used to generate the third harmonic radiation because its intensity is proportional to the cube of the input fundamental laser power, $P_\omega{}^3$. The medium must also be negatively dispersive; that is, the wave vector mismatch between the light at 3ν and the light at ν must be <0. This limits the tunability of the generated light to the red of allowed Rydberg ← ground state transitions in the medium. Still as shown in Figure 18.20, significant portions of the VUV and XUV can be covered using rare gases and Hg as nonlinear media. Typical conversion efficiencies range from 10^{-6} to 10^{-3}%, which while low, still translate into intense coherent VUV and XUV outputs given that commercially available dye lasers for example can readily provide fundamental beams with high peak powers (1–10 MW; ~10-nanosecond pulse durations). One particularly simple and popular source is VUV light at ~118 nm radiation generated by tripling the third harmonic of a Nd : YAG laser at 355 nm. Such a source can have wide applications in the soft-ionization of organic molecules, that is, ionization without fragmentation.

Two-photon resonantly enhanced four-wave mixing is a two-color third-order nonlinear process where one laser at frequency ν_1 is tuned to be in resonance with a two-photon allowed transition in a nonlinear medium. A second laser operating at ν_2 can be tuned to generate coherent radiation at the sum and difference frequencies $2\nu_1 \pm \nu_2$. The expression four-wave mixing indicates that three input waves generate a fourth output signal. The nonlinear output wavelengths lie in the VUV and XUV when using visible and UV fundamentals.

The coherent output radiation is widely tunable because there is no inherent restriction on the frequency ν_2. Phase matching for sum-frequency mixing requires the medium to be negatively dispersive. However, the medium can be either positively or negatively dispersive for four-wave difference mixing. Most atomic media that have been used for four-wave mixing

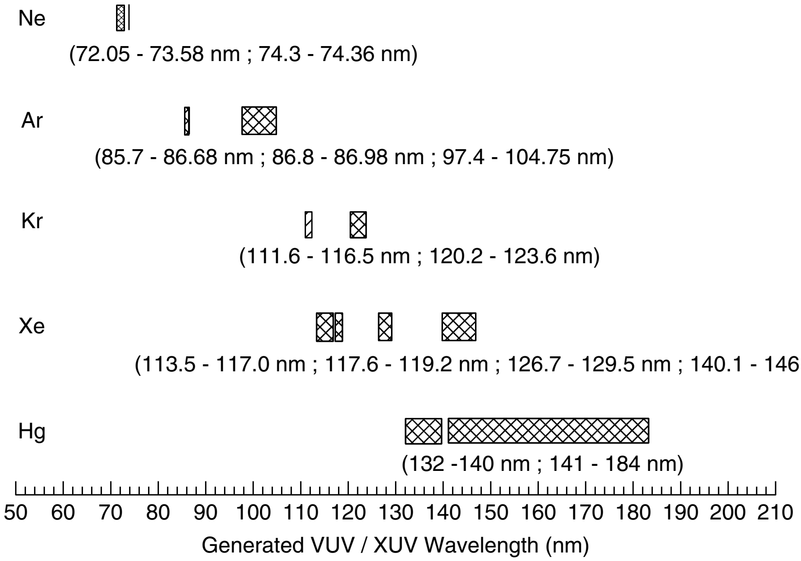

Fig. 18.20 VUV and XUV spectral tuning regions for nonresonant THG in Ne, Ar, Kr, Xe, and Hg.

are intrinsically negatively dispersive because the generated light lies higher in energy than the energy of the main resonance line of the medium, whereas the fundamental energies lie below.

Pulsed fundamental laser beams are almost always used since both sum- and difference-mixing scale as $P_1{}^2P_2$. The advantage of two-photon resonance enhanced four-wave mixing over nonresonant THG however, is that the third-order nonlinear susceptibility of the process can be enhanced many orders of magnitude by tuning one of the fundamental laser frequencies to be in resonance with a two-photon allowed transition in the nonlinear medium. In this way, intense VUV and XUV radiation can be generated with only moderate incident laser powers.

The tuning ranges available using Kr gas, Xe gas, Hg vapor and Mg vapor are shown in Figure 18.21. Metal vapors need to be generated at high temperatures, usually in heat pipes, while rare gas are easily handled in a gas cell. Rare gases have the additional advantage that they can also be expanded as supersonic expansions in a vacuum system. These gas jets are spatially localized and therefore can be supported in a windowless arrangement, provided the vacuum system has adequate differential gas pumping capacity. The elimination of windows is ideal for XUV generation as there are no broadly transmitting materials in these spectral regions.

18.7.5
Higher Order Harmonic Generation

The high intensities ($10^{13}-10^{16}$ W cm^{-2}) associated with femtosecond sources such as Ti : sapphire lasers can be used to extend the generation of short wavelength light into the soft X-ray region by

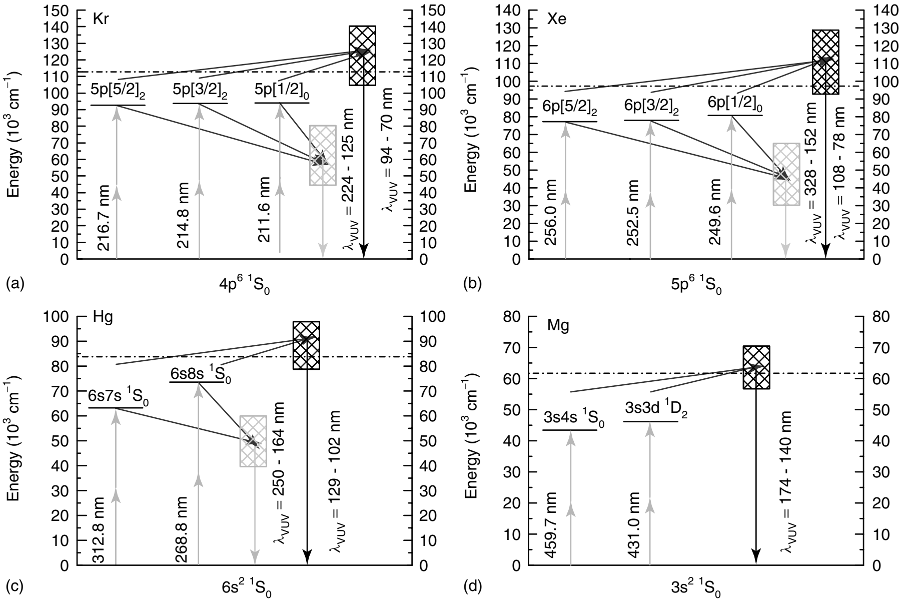
Kr
Energy (10^3 cm^{-1})
5p[5/2]$_2$
5p[3/2]$_2$
5p[1/2]$_0$
216.7 nm
214.8 nm
211.6 nm
λ_{VUV} = 224 - 125 nm
λ_{VUV} = 94 - 70 nm
(a)
4p^6 ^{1}S$_0$
Xe
6p[5/2]$_2$
6p[3/2]$_2$
6p[1/2]$_0$
256.0 nm
252.5 nm
249.6 nm
λ_{VUV} = 328 - 152 nm
λ_{VUV} = 108 - 78 nm
(b)
5p^6 ^{1}S$_0$
Hg
6s7s ^{1}S$_0$
6s8s ^{1}S$_0$
312.8 nm
268.8 nm
λ_{VUV} = 250 - 164 nm
λ_{VUV} = 129 - 102 nm
(c)
6s^2 ^{1}S$_0$
Mg
3s4s ^{1}S$_0$
3s3d ^{1}D$_2$
459.7 nm
431.0 nm
λ_{VUV} = 174 - 140 nm
(d)
3s^2 ^{1}S$_0$

Fig. 18.21 VUV and XUV spectral tuning regions for two-photon resonance-enhanced four-wave sum and difference mixing in (a) Kr, (b) Xe, (c) Hg, and (d) Mg.

high-order harmonic generation. When the intensity of the input laser pulses is sufficiently high, electrons from a gas-phase atomic medium are ionized and begin to move away from the resultant ions at relativistic velocities. However, because the distance they travel is small within one or more laser periods, they still respond to the electric field by undergoing ponderomotive free space oscillation. Those electrons that recombine with the ions within a fraction of one optical cycle give up their energy as high order multiples of the input laser frequency. However, a plateau exists on the frequency of harmonic light that can be generated, which is determined by the ionization potential of the atom being ionized and the maximum kinetic energy of the recombining electron (Corkum, 1993). Remarkably, the pulse duration of the light generated by this strong field interaction is on the attosecond (10^{-18} s) timescale that provides new opportunities to probe complex electron dynamics in molecules and materials (Kapteyn *et al.,* 2007). The output efficiency scales with the gas density, and phase matching is approximately independent of the order of the process. Wavelengths as short as 2.7 nm have been generated in He using 26-femtosecond 800-nm pulses from a Ti : sapphire laser (Chang *et al.*, 1997). Coherent harmonic generation up to order 221 and shorter continuum range of wavelengths up to order 297 were resolved. The shortest wavelength lies within the "water window" region of X-ray transmission, making these sources useful for the study of biological samples.

18.7.6
Free-Electron Laser

The free-electron laser has a highly accelerated electron beam as its active medium. One of the most common experimental arrangements, illustrated in Figure 18.22, involves passing a beam of very high energy (10^7–10^8 MeV) relativistic electrons from the accelerator of a synchrotron ring between the poles of a series of regularly spaced magnets of alternating polarity called an *undulator*. The electrons are repeatedly accelerated and decelerated in a direction perpendicular to their direction of travel, resulting in an oscillatory path, as shown. The effect of this process is to produce emission of coherent *bremsstrahlung* radiation along the axis of the laser at a wavelength that depends on the periodicity of the undulator and the energy of the electrons. The radiation is then trapped between parallel mirrors and stimulates further emission.

High-energy electrons travel at an appreciable fraction of the speed of light, and can be characterized by a parameter f, denoting the ratio of their relativistic total energy to their rest-mass energy. If the magnet separation is d, then the laser emission wavelength is given by the simple formula $\lambda = d/2f^2$. Because the energy of the electrons emerging from the accelerator can be continuously varied, the result is again a laser with tuning capability. There is, in principle, no limit to the range of tuning right across the infrared, visible, and ultraviolet regions of the electromagnetic spectrum. Moreover, this kind of laser has been shown to produce high average powers and to be capable of a reasonable efficiency. For example, a setup in the Naval Research Laboratory in the USA. can produce 75 MW pulses of 4 mm radiation with an efficiency of about 6%. The efficiency in the visible and ultraviolet is generally lower. The downside of these devices is that they are expensive and require access to a large synchrotron facility.

Plans have been made to develop free-electron lasers that can operate in the soft and hard X-ray regions. For example, a second generation free-electron laser has been proposed for the BESSY beamline in Germany where a seed laser pulse copropagates with the electron bunch through the undulator to allow other laser sources to be synchronized for pump-probe experiments (Follath, 2007). The seed laser pulses at 51–1.24 nm are generated by higher harmonic generation of a tunable laser operating between 230 and 460 nm. Nonlinear interactions generate higher harmonics of the seed laser, and therefore additional insertion

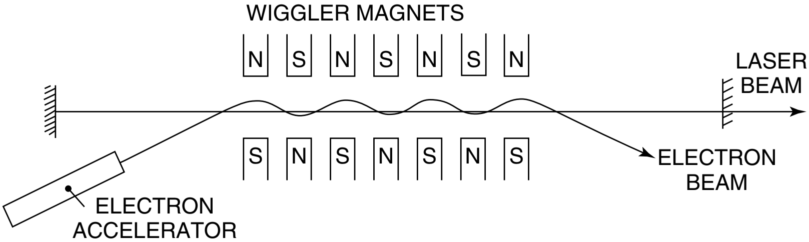

Fig. 18.22 Configuration of a free-electron laser.

devices are required to extract our one particular wavelength. Proof of principle of this approach has been realized by researchers at Brookhaven National Laboratory who demonstrated a high-gain harmonic-generation free-electron laser operating in the ultraviolet at 266 nm (Yu *et al.,* 2003).

18.7.7
Tabletop X-ray Lasers

Despite the obstacles inherent in generating a population inversion that can lead to gain in the X-ray region, there have been many notable successes (Rocca, 1999). While the field is too broad to cover in this article, it is worth noting that Rocca and coworkers have developed fixed-wavelength soft X-ray lasers that use fast pulsed discharge excitation of capillary channels to create plasmas that exhibit gain by collisional recombination. For example, this scheme was used to generate coherent radiation on the 46.9 nm 3p-3s ($J = 0-1$) line of Ne-like Ar (Rocca *et al.*, 1994). These lasers are attractive because of their small size, earning them the moniker "tabletop" devices.

References

Andrews, D.L. (1997) Lasers in Chemistry, 3rd edn. Springer-Verlag, Berlin.

Asasda, M., Miyamoto, Y., and Suematsu, Y. (1986) *IEEE J. Quantum Electron.*, **QE-22**, 1915–1921.

Auzel, F. (2004) *Chem. Rev.*, **104**, 139–173.

Chang, Z., Rundquist, A., Wang, H., Murmane, M.M., and Kapteyn, H.C. (1997) *Phys. Rev. Lett.*, **79**, 2967–2970.

Corkum, P.B. (1993) *Phys. Rev. Lett.*, **71**, 1994–1997.

Faist, J., Capasso, F., Sivco, D.L., Sirtori, C., Hutchinson, A.L., and Cho, A.Y. (1994) *Science*, **264**, 553–556.

Fiore, A., Markus, A., Rossetti, M., and Li, L.H. (2006) *Laser World Focus*, **42**, 124–127.

Follath, R. (2007) *Proc. SPIE*, **6586**, 658604.

Fork, R.L., Brito Cruz, C.H., Becker, P.C., and Shank, C.V. (1987) *Opt. Lett.*, **12**, 483–485.

Hecht, J. (2005) *Laser Focus World*, **41**, 143147.

Jordan, W.G., Jha, A., and Ryan, J. (1993) *J. Mat. Sci. Lett.*, **11**, 771–773.

Kapteyn, H., Cohen, O., Christov, I., and Murnane, M. (2007) *Science*, **317**, 775–778.

King, T.A. (2004) Fibre lasers, in *Comprehensive Series in Photochemistry and Photobiology, 4* (eds G. Palumbo and R. Pratesi), Royal Society of Chemistry, Cambridge.

Leinen, H., Gläβner, D., Metcalf, H., Wynands, R. Haubrich, D., and Meschede, D. (2000) *Appl. Phys. B*, **70**, 567–571.

Lipson, R.H., Shi, Y.J., and Lacey, D. (2002) in *An Introduction to Laser Spectroscopy*, 2nd edn (eds D.L. Andrews and A.A. Demidov), Chapter 9, Kluwer Academic Plenum Publishers, New York, pp. 257–309.

Monerie, M. (1989) *Phys. Scripta.*, **T29**, 218–222.

Mowbray, D. (2005) *Laser Focus World*, **41**, 157–159.

Nakamura, S. (1998) *Science*, **281**, 956–961.

Rice, P.R. and Carmichael, H.J. (1994) *Phys. Rev. A*, **50**, 4318–4329.

Righini, G.C., Brenci, M., Chiasera, A., Feron, P., Ferrari, M., Nunzi, J.-M., Conti, G., and Pelli, S. (2005) *Proc. SPIE*, **6029**, 603903.

Rocca, J.J. (1999) *Rev. Sci. Instrum.*, **70**, 3799–3827.

Rocca, J.J., Shlyaptsev, V., Tomasel, F.G., Cortázar, O.D., Hartshorn, D., and Chilla, J.L.A. (1994) *Phys. Rev. Lett.*, **73**, 2192–2195.

Shank, C.V., Fork, R.L., Yen, R., and Stolen, R.H. (1982) *Appl. Phys. Lett.*, **40**, 761–763.

Steinmeyer, G. (2006) *Appl. Opt.*, **45**, 1484–1490.

Treacy, E.B. (1969) *IEE J. Quantum Electron.*, **QE-5**, 454–458.

Vahala, K.J. (2003) *Nature*, **424**, 839–846.

Yablonovitch, E. (1987) *Phys. Rev. Lett.*, **58**, 2059–2062.

Yu, L.H., DiMauro, L., Doyuran, A., Graves, W.S., Johnson, E.D., Heese, R., Krinsky, S., Loos, H., Murphy, J.B., Rakowsky, G., Rose, J., Shaftan, T., Sheehy, B., Skaritka, J., Wang, X.J., and Wu, Z. (2003) *Phys. Rev. Lett.*, **91**, 074801–1–074801–4.

LASER TECHNOLOGY

Hiroshi Takuma, *The Institute for Laser Science, University of Electro-Communications, Tokyo, Japan*

Introduction 321

1. **Measurements of Basic Characteristics of Lasers** 321

1.1 Mode of Oscillation and Beam Divergence 321

1.2 Output Power 322

1.3 Temporal Brightness 322

1.4 Spectral Brightness 323

1.4.1 Grating Spectrometer and Fabry–Pérot Interferometer .. 323

1.4.2 Scanning Confocal Fabry–Pérot Optical Spectrum Analyzers 323

1.4.3 Homodyne Detector and Electronic Spectrum Analyzer. 323

1.4.4 Autocorrelation Measurements

2. **Wavelength Measurements** ... 323

2.1 Moving Michelson-Type Wavelength Meter 324

2.2 Direct Frequency Comparison with Microwave Frequency Standard 324

2.3 Fizeau Wavelength Meter 325

3. **Measurement of the Characteristic Parameters of Laser Media** 325

4. **Stabilization of the Laser Output** 327

4.1 Amplitude Stabilization 327

4.2 Frequency Stabilization 327

5. **Squeezing** 328

6. **Manipulation of Laser Beams** 328

6.1 Q Switching 328

6.2 Modulation and Switching of Laser Beams 329

6.3 Optical Harmonic Generation 329

Works Cited 330

Further Reading 330

INTRODUCTION

Laser technology deals with operation of lasers and the characterization and the use of their output beams. Although lasers are used in various kinds of systems that are important in various practical applications, only basic techniques are presented here. For more detailed descriptions on various kinds of lasers and application systems, the reader is advised to consult appropriate other articles of this Encyclopedia and the references listed at the end of the present article.

1. MEASUREMENT OF BASIC CHARACTERISTICS OF LASERS

1.1 Mode of Oscillation and Beam Divergence

Normally a laser is designed to operate in its fundamental TEM_{00} transverse mode so that an output beam having Gaussian intensity distribution is generated, although sometimes a higher-mode or mixed-mode output may be sought for some very specialized applications. In practice some amount of higher modes is mixed with the fundamental mode, and the beam pattern shows some deviation from the pure Gaussian distribution. Precise measurement of the beam pattern can be made by observing the intensity distribution of the cross section of the laser beam with a charge-coupled-device (CCD) camera and analyzing the two-dimensional pattern with a computer.

The Gaussian beam behaves just like a part of a spherical wave at a long distance from the beam waist, which is the point where the spot size of the beam has its minimum value. A beam waist corresponds to the focal point in geometrical optics and has a nonzero size because of diffraction of the light wave. The spot size of a Gaussian beam can be defined as

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

the radius of the contour of the spot where the intensity is $1/e^2$ of the peak value. When the beam is focused by a lens of focal length f, the minimum spot size, which shall be designated as d, is given by the relationship $d = f\theta$, where θ is the half-angle beam divergence in radians. For a Gaussian beam, θ is given by the equation

$$\theta = [(\lambda/\pi w_1)^2 + w_1^2/R_1^2]^{1/2}, \qquad (1)$$

where λ, w_1, and R_1 are the wavelength, the beam spot size, and the radius of curvature of the laser beam at the input face of the lens, respectively. In practice, it is convenient to make the second term negligibly small by making the input laser beam nearly parallel using a beam expander, which is an inverted Galilean telescope. In a practical laser beam the measured beam divergence is always larger than that given by (1) because of the presence of the higher modes.

In above discussion, we assumed axial symmetry of the laser beam. However, in some kinds of lasers, the active part of the laser medium strongly deviates from axial symmetry. The output beam of such lasers may not be axially symmetric. The typical example of such is the semiconductor diode laser: the active part of it is very thin in the direction perpendicular to the *pn* junction layer. The output of a diode laser reflects such a peculiar structure, showing an asymmetric diffraction pattern. The typical beam divergence of a diffraction-limited, single-mode diode laser is 10° in the direction parallel to the junction layer, and 30°–40° in the perpendicular direction. Because even such a large angular spread is due to diffraction, a single-mode diode laser output can be converted into a parallel axially symmetric beam by the use of properly designed optics.

1.2 Output Power

The absolute value of the average output power is normally obtained by a calorimeter. Let the laser beam or a well-defined part of it be absorbed by a body, the heat capacity of which is well known, for a known duration of time. The average power is determined by measuring the temperature rise of the body due to the absorbed laser energy. A more conventional method is to use a commercially available calibrated bolometer. In carrying out the measurement, care must be taken that the effective area of the detector is large enough to collect all the input beam, the absorption rate of the detector is well known, and the intensity of the laser-beam is not strong enough to burn the detector surface or to cause any nonlinear optical effect (which scatters part of the energy of the laser beam out of the detector as photons of a different wavelength).

In a pulsed laser, the output energy per pulse can also be measured by the same technique by knowing how many pulses are incident on the detector. The average output, which is directly measured by the bolometer, is also a very important quantity for many kinds of applications.

We assumed that only a coherent output beam is coming out of a laser in the above discussion. However, this is not true in most practical cases, and a typical laser output includes an incoherent part of amplified spontaneous emission in addition to the coherent output. In order to remove the incoherent part of the output beam, which is spread wide over the coherent beam, the use of a spatial filter is strongly recommended. This is a device with two focusing lenses aligned on the axis of the laser beam at a distance where the two lenses share a common focal point, at which a pinhole is placed so that only the focusable component of the laser beam is passed through.

It is important in designing a laser to eliminate the incoherent part of the output. In a high-gain laser such as rare-gas halide lasers it is especially important, and sometimes the incoherent portion exceeds the coherent portion in integrated energy. However, in lasers designed for some special applications in which uniform illumination of a surface is essentially important, the lasers are designed so as to make the output incoherent while still keeping the beam divergence small.

1.3 Temporal Brightness

The temporal brightness, which is defined as the output power per unit solid angle, is a very important physical quantity that specifies the output characteristics of lasers. The higher the brightness, the higher the intensity of the focused laser beam. When the output power of a laser is rapidly changing, simultaneous measurements of the instantaneous output power, which is expressed as $f(t)$, the integrated output power W, and the beam divergence θ for a given time interval t give the brightness of the beam $B(t)$ as follows:

$$B(t)=Cf(t)/\pi\theta^2, \tag{2}$$

where

$$C=W\Big/\int_0^t f(t)dt^{-1},$$

assuming that only the relative value of $f(t)$ is known. In the case where a time resolution of the order of 1 ns is sufficient, the instantaneous output power can be measured by detecting the relative value of the instantaneous intensity by a photodiode and displaying it on an oscilloscope. A high-speed streak camera provides a high time resolution up to the order of 1 ps. Those interested in subpicosecond resolution phenomena and their measurements should refer to the article ULTRAFAST PULSE PHENOMENA.

1.4 Spectral Brightness

Suppose that the output of a laser is stationary for the sake of simplicity. Although a laser beam is generally regarded as monochromatic light, the output energy of a laser is more or less spread in the frequency domain. By expressing the output power of a laser in a frequency domain $[\nu, \nu+d\nu]$ as

$$dI=I_0S(\nu)d\nu\,, \tag{3}$$

where I_0 is a constant, $S(\nu)$ is defined as the normalized spectroscopic brightness. The following techniques are the most standard ones to measure the spectroscopic brightness.

1.4.1 Grating Spectrometer and Fabry–Pérot Interferometer. For the measurement of a light source having rather broad spectral width and where a resolution of only 0.01 nm (12 GHz) or lower is needed, a conventional grating spectrometer is the most suitable measuring device. A Fabry–Pérot étalon (see INTERFEROMETERS AND INTERFEROMETRY) is appropriate to measure a spectral width of 30–1 GHz. It should be noted here that both of the above instruments are made for incoherent light sources, and one should remove the spatial coherence of the beam before using them. Because a rather long time interval is normally needed in measurements with the above two instruments, inserting a ground glass or a sheet of tissue paper in front of the monochromator slit or the Fabry–Pérot étalon is sufficient.

1.4.2 Scanning Confocal Fabry–Pérot Optical Spectrum Analyzers. Units having finesse of higher than 200 are commercially available, providing a versatile way to measure the power spectrum of coherent sources having spectral widths of a few megahertz to a few gigahertz.

1.4.3 Homodyne Detector and Electronic Spectrum Analyzer. When a resolution of better than a few MHz is needed, homodyne detection by the use of a photodiode as a mixer is convenient. The key to obtaining a good signal-to-noise ratio in mixing two coherent optical beams is to align the wave fronts of the two beams with high accuracy.

1.4.4 Autocorrelation Measurements. According to the Wiener–Khinchin theorem, the Fourier transform of the autocorrelation function of a signal gives its power spectrum, and the spectral brightness of a laser beam can be obtained by measuring its autocorrelation. That is, if we describe the normalized temporal evolution of a laser output by $f(t)$, the autocorrelation function, represented by $A(\tau)$, is defined by

$$A(\tau)=\int_{-\infty}^{+\infty} f(t+\tau)f^*(t)dt. \tag{4}$$

In principle, the spectral brightness is obtained by obtaining the experimental value of $A(\tau)$ by varying the delay time τ and then carrying out numerical Fourier transformation by a computer.

2. WAVELENGTH MEASUREMENTS

As the length standard is connected to the frequency standard by the speed of light, which is defined as $2.997\,924\,58\times10^8$ m/s, a laser wavelength measurement is equivalent to a laser frequency measurement. The wavelength of a laser relative to another is directly measured by comparing the interference fringes by the use of various types of interferometer arrangements.

2.1 Moving Michelson-Type Wavelength Meter

This most popular method essentially counts the number of fringes when one arm of a Michelson interferometer travels a known distance. The distance is normally measured by counting fringes made by another laser beam, the wavelength of which is established as a secondary standard (Hall and Lee, 1976).

A basic setup is schematically shown in Fig. 1. If the length of the fixed arm is l_1 and the movable one l_2, the difference in the optical path length for a round trip of the two arms is given by $2l_1-2l_2$, and interference of the two beams shows maxima when the following relationship holds:

$$2l_1-2l_2=n\lambda/2\ . \tag{5}$$

Therefore if two laser beams having respective wavelengths λ_1 and λ_2 are introduced to the interferometer at the same time separately and fringe counts of n_1 and n_2, respectively, are obtained for the same value of $2l_1-2l_2$, then we have

$$n_1\lambda_1/2=n_2\lambda_2/2\ . \tag{6}$$

If the wavelength of the one of the laser beams, for example, λ_1, is known, the other is given by

$$\lambda_2=(n_1/n_2)\lambda_1. \tag{7}$$

The technical difficulty in such a measurement is to move the end reflector M_2 smoothly without changing the direction of the reflected beam. The effect of an inevitable mechanical fluctuation of the moving reflector may be avoided by the use of a retroreflector such as a corner cube. Several wavelength standards are established on laser lines as listed in Table 1 by comparing their frequency with that of the cesium clock with a frequency chain method described below. The accuracy attained by this method ranges from a few parts in 10^7 to 10^8 for commercially available systems to a few parts in 10^{10} for the most elaborate laboratory systems.

2.2 Direct Frequency Comparison with Microwave Frequency Standard

This is not a conventional method to be applied in a routine measurement, but an elaborate way to establish secondary fre-

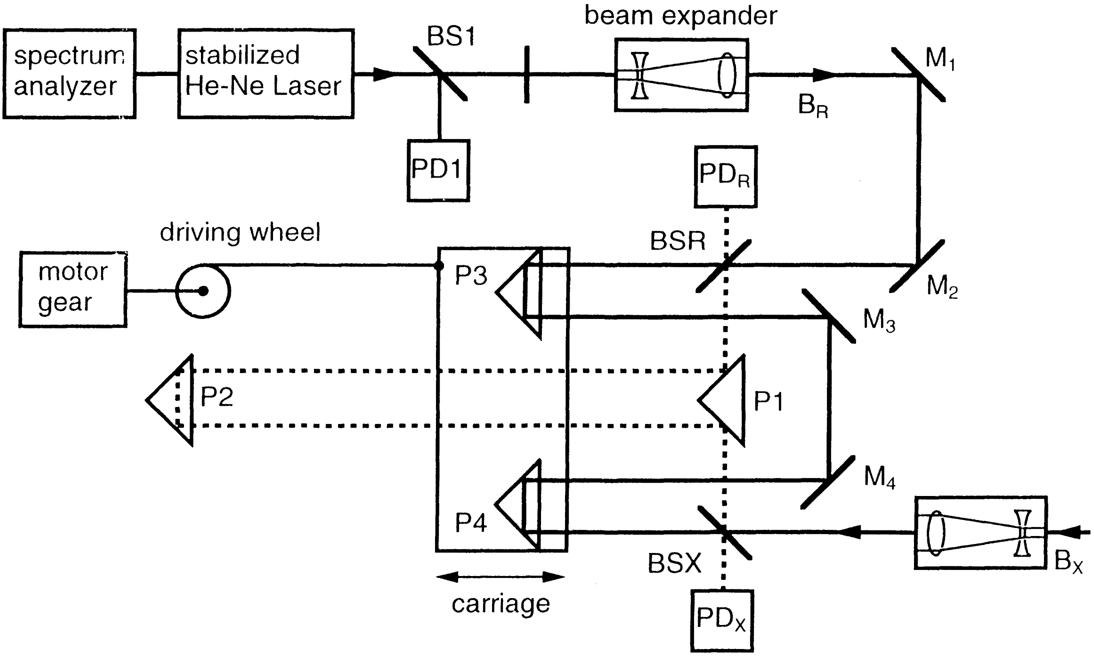

FIG. 1. A variation of the moving Michelson interferometer (Kowalski, *et al.*, 1977). The interference fringes of the reference and the sample beams are detected with PD_R and PD_X, respectively, as the carriage moves back and forth in the horizontal direction.

Table 1. Laser frequencies measured by direct comparison with calibrated microwave frequencies.

Atom or molecule	Frequency (THz)
HCN	0.890 760 550
H_2O	10.718 068 6
$^{12}C^{16}O_2$ *R*[20]	30.950 409
$^{12}C^{16}O_2$ *R*[22]	32.373 156
^{20}Ne (CH_4 *P*[7])	88.376 181 627[a]
^{20}Ne	196.780 372
^{20}Ne	260.103 264
$^{127}I_2$ [17–1 *P*(62)]	520.206 528[b]

[a] Frequency of the laser locked to the *P*[7] line of CH_4.

[b] Frequency of 17–1 *P*(62) transition of $^{127}I_2$ measured by the second harmonic of the 260-THz line of ^{20}Ne.

quency standards in optical region. Although the principle is to compare the higher harmonics of the microwave frequency standard with a laser frequency, direct comparison is impossible in the present state of art, and the frequency comparison chain must include several intermediate-frequency lasers ranging from far infrared to infrared. The frequencies of those lasers that have been determined by such a frequency comparison chain are tabulated in Table 1.

2.3 Fizeau Wavelength Meter

This is a method that can be used to determine the wavelength of a laser with pulsed output, in contrast to the two methods described above. As shown in Fig. 2, the principle is simply to compare the fringe separation observed by a Fizeau interferometer with that formed by a wavelength standard laser. The fringes can be counted by a photodiode array and the result is processed by a small computer. This is a versatile method to be used in the visible to near-infrared region, giving an accuracy of up to a few parts in 10^8.

3. MEASUREMENT OF THE CHARACTERISTIC PARAMETERS OF LASER MEDIA

The phenomenological characteristics of a laser medium are generally described by three parameters: small-signal gain, absorption coefficient, and saturation intensity, which will

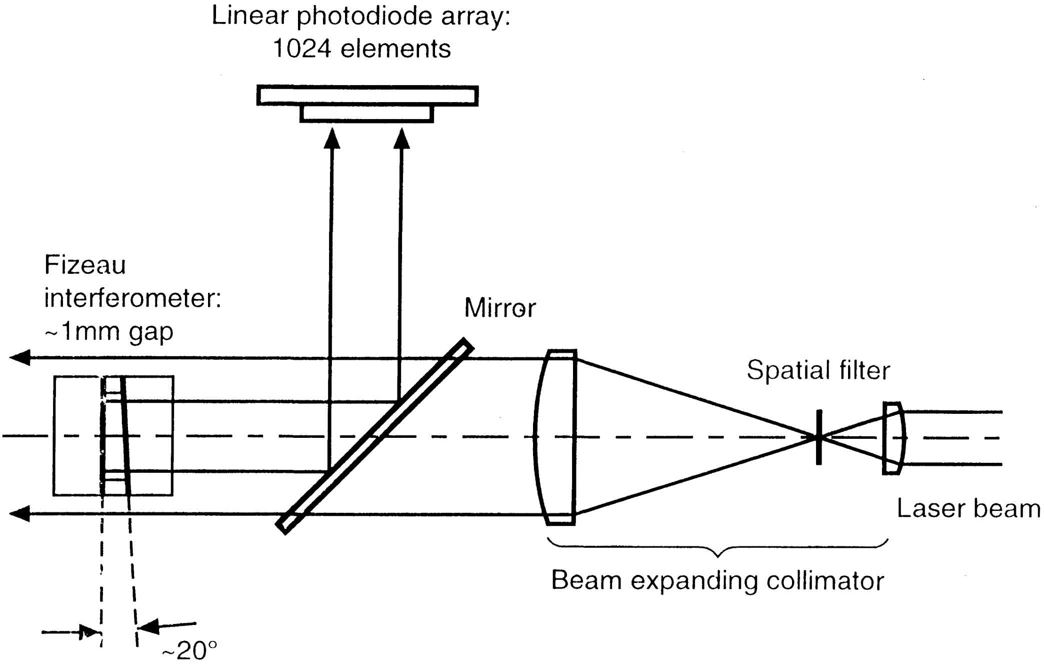

FIG. 2. Schematic diagram of the Fizeau-type wavelength meter (Snyder, *et al.*, 1977).

be designated by g_0, a, and I_S, respectively. The gain coefficient of the laser medium in operation with a laser radiation intensity of I is given by

$$g=\frac{g_0}{1+I/I_S} . \tag{8}$$

The effective gain coefficient, which is represented by g_e, is defined as

$$g_e=g-a . \tag{9}$$

Since the field intensity in the laser medium is dependent on the quality factor of the optical resonator, it is possible to determine the three constants by fitting the experimental values of the output obtained for various values of quality factor by replacing a set of reflectors that compose the resonator. However, the accuracy obtained by such measurements is rather poor, and they are not recommended if the following measurement is possible.

The most accurate determination of the above three amplifier constants can be made by realizing conditions under which the effective gain is expected to be constant in the medium. Such a condition is generally satisfied in the following three cases:

1. small-signal condition, that is, the input power is so small that saturation is negligible even after the laser beam has made a single excursion of the laser medium;
2. zero-gain condition, where the input power happens to make the saturated gain equal to the absorption coefficient, making the effective gain exactly zero; and
3. perfectly saturated gain condition, where the input gain is so large that the saturated gain is negligibly smaller than the absorption coefficient.

Figure 3 shows the output of an amplifier for a wide range of input intensity.

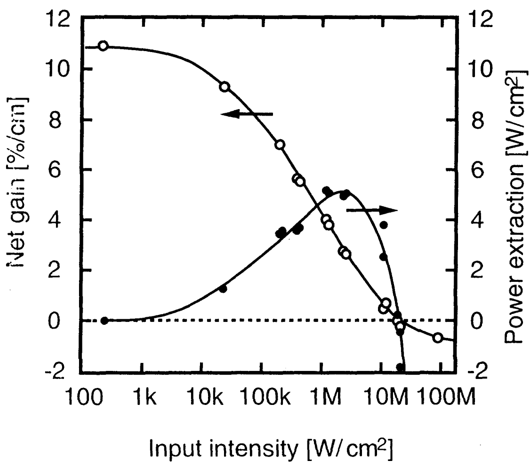

FIG. 3. An example of the output of a laser amplifier for a broad range of input intensity (Ueda and Takuma, 1987).

The small-signal condition may be checked easily by confirming the proportionality of the output and the input by making the input power level as low as possible, and g_0 is simply determined as the ratio of the output to the input power. The third condition is also defined accurately by confirming that the laser medium shows linear absorption by applying an extremely intense input power: the experimentally measured absorption coefficient under such a condition gives us the value of a with good accuracy. A good number of experimental runs with various input intensities may be needed, but in principle it is certainly possible to find the input power at which the saturated gain is exactly equal to the linear absorption and the zero effective gain condition is realized. Let us define such an intensity as I_n. Then we expect the following relationship to hold:

$$I_S=I_n/(g_0-a) , \tag{10}$$

which gives us the saturation intensity I_S from the values of g_0 and a obtained above. Thus the three key parameters can all be determined with a good accuracy.

Although the above method gives complete knowledge of the three characteristic parameters of an amplifier, it is not guaranteed that it can be applied to any laser medium. The measurement of gain under small-signal condition imposes the difficulty of obtaining a good signal-to-noise ratio in measuring the intensity of a laser beam at a very low power level. For some kinds of laser media, such a high intensity of input is needed to realize the third condition, and sometimes even the second condition, that the laser media are damaged by the laser beam, making accurate measurements impossible.

A typical example in which such measurements gave excellent results is the excimer lasers, especially the KrF laser, having high small-signal gain and rather low saturation intensity.

4. STABILIZATION OF THE LASER OUTPUT

Stabilization of the laser output is needed quite often when it is used in high-accuracy measurements. For some applications, stabilization of the output power of the laser source is needed. In many cases frequency stabilization of the laser output is needed.

4.1 Amplitude Stabilization

Amplitude stabilization can be accomplished by the use of a continuously variable attenuator and installation of a feedback loop to adjust the attenuation rate to make the output power of the attenuator constant. In practice, it can be accomplished easily when the laser beam is linearly polarized by a system as illustrated schematically in Fig. 4. The beam is passed through a Pockels cell that rotates the polarization plane of the beam in proportion to the increase of the output power. A polarizer placed behind the cell selects only the component having the same polarization as the original one. A small fraction of the output power is detected by a photodiode and the amplified detected signal is used to control the voltage applied to the Pockels cell.

The above method is the conventional way to obtain an amplitude-stabilized laser beam. However, one can expect to generate an amplitude-stabilized output directly from lasers or nonlinear devices by suppressing the amplitude noise. For example, a diode laser driven by a current-stabilized dc source shows amplitude squeezing at a pumping level much higher than the threshold value. This will be discussed in Sec. 5.3.

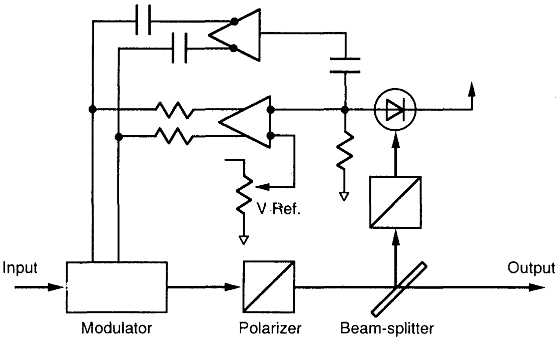

FIG. 4. Schematic diagram of an intensity stabilizer.

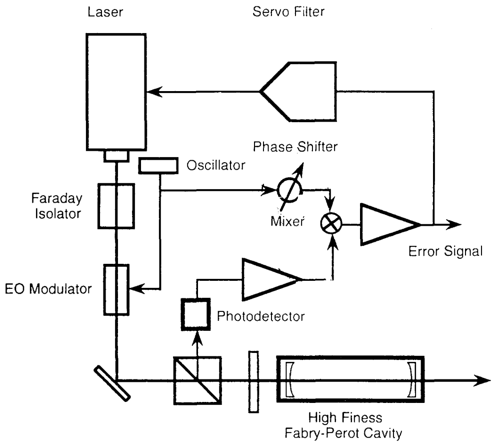

FIG. 5. Schematic diagram of a typical frequency-stabilized laser.

4.2 Frequency Stabilization

The general principle of laser frequency stabilization is to pass the laser beam through a frequency discriminator having a peak transmission or peak attenuation at a given frequency, and to control the laser frequency so that it always stays at the frequency defined by the frequency discriminator as schematically shown in Fig. 5. A Fabry–Pérot type optical resonator is widely used as the frequency discriminator, although it cannot be an absolute frequency standard unless it is calibrated with respect to some kind of absolute standard such as atomic or molecular spectral lines.

Recent progress in optical coating technology makes it possible to obtain low-loss, high-reflectivity reflectors, for example, 99.99% reflectivity and a few ppm loss, the use of which allows us to make a Fabry–Pérot resonator having a finesse of several hundred thousands by keeping a fairly high transmission rate for the beam of resonant frequency. The use of such an optical resonator gives us the possibility of controlling a laser frequency to be stabilized at a given frequency with subhertz accuracy. As a matter of fact, He-Ne lasers and laser-diode-pumped Nd:YAG lasers have been stabilized to a level of relative stability on the order of mHz, exceeding the theoretical limit of a free-running laser given by Schawlow and Townes. On the other hand, stabilization of absolute frequency involves several unknowns, such as the effect of heating of the mirror surface by the resonant radiation, the intensity of which is enhanced to an extremely high level in high-finesse resonator.

Stabilization of frequency-tunable lasers such as dye lasers, Ti sapphire lasers, or diode lasers poses difficult extra problems, such as keeping single-mode oscillation for a wide range of tuning. However, frequency stabilization up to the level of 1 MHz to 100 kHz can be realized rather easily even for those lasers.

5. SQUEEZING

According to the quantum mechanics of optical fields, uncertainties in the amplitude and phase are governed by the uncertainty relationships. That is, representing the root mean square deviations of the photon density and the phase as Δn and $\Delta\phi$, respectively, we have the relationship

$$\Delta n \, \Delta\phi \geqslant \tfrac{1}{2} \,. \tag{11}$$

This means that photon-density noise in the laser resonator can be suppressed by allowing the phase noise to increase. A normal laser oscillator generates coherent photons, which show equal uncertainty in the photon density and the phase. The process that generates photons having less uncertainty in the photon density is called *amplitude squeezing*, and that which generates less phase noise, *phase squeezing*.

Normal laser oscillation, which is built up from spontaneous emission, generates shot-noise–limited light. However, it is possible to generate squeezed-state photons by suppressing the fluctuations. The most commonly observed example is a semiconductor diode laser driven by a current-stabilized power source at a pump level much higher than its threshold. The general scheme to generate squeezed states is to use photon generation by nonlinear effects such as degenerate optical parametric oscillation, four-photon mixing, second-harmonic generation, or the stimulated Raman effect. An experimental setup for squeezed-state generation by a degenerate parametric oscillator is shown in Fig. 6. In this mechanism, the generated photons have a strong correlation with the pumping waves, and this is the key to generation of the squeezed state. Therefore quantum noise included in the pumping wave such as spontaneous emission tends to destroy the squeezed state. This makes experimental generation of the squeezed states difficult.

The common method to detect the squeezed state is a balanced homodyne detector as shown schematically in Fig. 7. The detected signal changes periodically with the phase of the local oscillator, and squeezing appears as the minimum detected noise being much less than the noise of the coherent photons.

The use of squeezed-state photons gives us the possibility of obtaining high sensitivity or accuracy in optical measurements. In practice, a squeezed state is easily destroyed by introducing loss or spontaneous emission, and so high precaution is needed. Even so, improvement in signal-to-noise ratio has been reported in a spectroscopic measurement.

6. MANIPULATION OF LASER BEAMS

6.1 *Q* Switching

In many laser applications, only a short pulse with high peak power is needed. In such cases, *Q* switching provides a convenient way of generating a pulse of a few to a few tens of nanosecond duration by extracting a large portion of the stored energy of the laser medium. A normal laser pumped by a pulsed source exhibits pulsed oscillation, generating pulses of the laser beam reflecting the temporal behavior of the pumping source. In such an oscillation, the population inversion is limited by the induced emission. However, if the laser oscillation is suppressed by inserting a loss in the laser cavity, an inverted population much higher than the threshold value can be built up. Then the cavity loss is suddenly

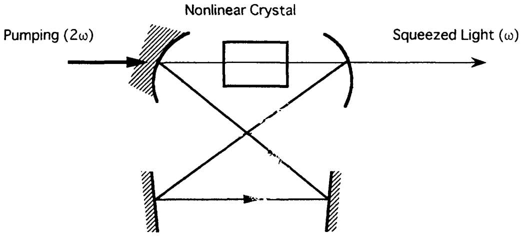

FIG. 6. Schematic diagram of a degenerate optical parametric oscillator that generates a squeezed output.

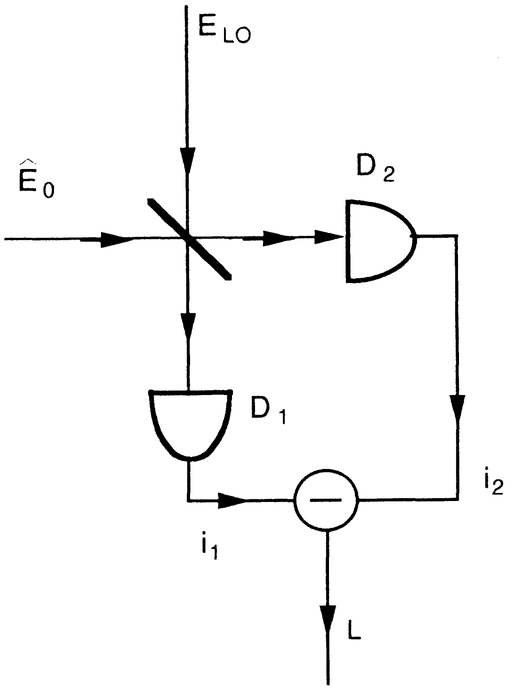

FIG. 7. Optical homodyne detector for detection of squeezed light. E_0 and E_{LO} stand for the sample beam and the reference beam, respectively. L stands for the local frequency output. D_1, D_2: detectors.

removed, and the energy stored in the medium as the inverted population can be extracted in a short time by building up the laser radiation. Such devices are often called *Q*-switched lasers and are widely used when a concentration of laser energy in a short time interval is needed.

As a *Q*-switching device, that is, a very fast shutter installed in a laser cavity, the Pockels cell is most frequently used in solid-state lasers. The laser is constructed in a way so that only linearly polarized light can be built up in the cavity by inserting a polarizer in the laser cavity, for example, and the Pockels cell rotates the direction of polarization 90° for the laser photon, making a round trip. Thus the oscillation of the laser is stopped until the population inversion reaches its maximum value by pulsed pumping and then the voltage applied on the Pockels cell is suddenly turned off. The laser radiation is built up rapidly by intense stimulated emission due to a large density of population inversion.

6.2 Modulation and Switching of Laser Beams

Many methods for laser applications require modulation or turning off and on the laser beams. There are two methods of frequency or phase modulation of laser beams. One is the acousto-optical modulator, which shows high modulation efficiency for frequencies lower than 100 MHz. However, only the electro-optical modulator is suitable for high-frequency modulation, although it is difficult to obtain high modulation efficiency.

An acousto-optical modulator consists of a transducer attached to a low-loss optical material having a high acousto-optical coefficient such as $PbMoO_4$, which is transparent to visible and near-infrared radiation. A supersonic wave is generated and propagates through the acousto-optical material in the direction perpendicular to the laser beam, and the latter is diffracted by a density grating made by the acoustic wave. The diffracted beam of the first order is frequency shifted by an amount equal to the frequency of the acoustic waves. The diffraction efficiency is approximately proportional to the power of the acoustic waves and can be higher than 50%.

For a frequency higher than 100 MHz, electro-optical modulation is the only method applicable at present. Electro-optical materials such as $LiNbO_3$ and $LiTaO_3$ have relatively high electro-optical coefficients and are most frequently used for such purpose. The electro-optical materials change their refractive indices when an electric field is applied. In most cases the electric field is applied in the direction perpendicular to the propagation direction of the laser beam and therefore the phase of the laser beam is shifted by an amount proportional to the applied electric field. The electro-optical modulator can be used as a fast intensity modulator and switching device having a high extinction rate. For more details, refer to OPTOELECTRONIC DEVICES and MODULATORS AND DEMODULATORS.

6.3 Optical Harmonic Generation

Second or higher-order harmonics of a laser beam can be generated in nonlinear optical materials. Second-harmonic generation is possible only in a medium that does not have point symmetry. Therefore piezoelectric crystals are good candidate nonlinear materials for second-harmonic generation. The intense electric field of the laser induces dipoles that oscillate at second or higher harmonics of the original frequency. The conversion efficiency is proportional to the nonlinear polarizability.

The second order nonlinear polarization, which we shall designate as $\mathbf{P}_2$, can be represented as follows when a plane wave of angular frequency ω is propagating in the z direction:

$$\mathbf{P}_2=\chi^{(2)}(\mathbf{E}_0e^{i\omega t-ikz})^2=\chi^{(2)}\mathbf{E}_0^2e^{2i\omega t-2ikz}, \quad (12)$$

where $\chi^{(2)}$ is the second-order nonlinear susceptibility of the medium. The above equation shows that the polarization wave of angular frequency 2ω propagates in the medium in the same direction as the fundamental laser beam with wave number $2k$. On the other hand, the wave number of the electromagnetic wave of angular frequency 2ω in the medium, which shall be represented by $k_{2\omega}$, is not equal to $2k$ because of dispersion of the medium. Thus special care must be taken in order to obtain phase matching between the polarization wave and the electromagnetic wave of the second-harmonic frequency.

The phase-matching condition can be established by the use of birefringence of anisotropic crystals. That is, in a negative uniaxial crystal, for example, one may be able to find an angle θ in refractive index space for which

$$n_e^{2\omega}(\theta)=n_o^{\omega}, \quad (13)$$

where the left-hand side represents the index of refraction of the extraordinary ray at angular frequency 2ω in the direction at an angle θ with the optical axis, and the right-hand side the index of refraction of the ordinary ray at ω. In a positive uniaxial crystal, the roles of extraordinary ray and ordinary ray should be exchanged.

The angle θ changes as a function of temperature of the crystal and the frequency of the laser beam, and it is sometimes possible to find a temperature at which the above condition is satisfied for $\theta=2\pi$. Such a case is sometimes called temperature phase matching, and is preferentially applied in the case where the phase matching is obtained at a temperature that the crystal can be easily kept at.

By choosing an appropriate nonlinear crystal, the efficiency of the second-harmonic generation can be made higher than 75% for a pulsed laser beam having high temporal spectroscopic brightness. For a continuously oscillating laser, high-efficiency generation of the second harmonic is rather difficult. The efficiency can be improved significantly in such a case by placing the nonlinear crystal in an optical cavity that is resonant at the laser frequency.

Generation of harmonics higher than the second can normally be done only with low efficiency. The efficiency can be improved greatly, but is still much lower than that of the second harmonic, by selecting a medium in which some resonance effect can be expected. In spite of the low efficiency that one expects in the high-order harmonic generation, it is useful for generating coherent radiation at extremely short wavelengths.

List of Works Cited

Hall, J. L., Lee, S. A. (1976), *Appl. Phys. Lett.* **29**, 367–340.

Jennings, D. A., Peterson, F. R., Evenson, K. M. (1979), in: H. Walther and K. W. Roth (Eds.), *Laser Spectroscopy IV*, Heidelberg: Springer, pp. 39–48.

Kowalsky, F. V., Demtröder, W., Schawlow, A. L., (1977), in: J. L. Hall and J. L. Carlsten (Eds.), *Laser Spectroscopy III*, Heidelberg: Springer, pp. 412–413.

Snyder, J. J. (1977), in: J. L. Hall and J. L. Carlsten (Eds.), *Laser Spectroscopy III*, Heidelberg: Springer, pp. 419–420.

Ueda, K., Takuma, H. (1987), in: C. Yamanaka (Ed.) *Short-Wavelength Lasers*, Heidelberg: Springer, pp. 178–185.

Further Reading

Arecchi, F. T., Schultz-Dubois, E. O. (Eds.) (1972), *Laser Handbook*, Amsterdam: North-Holland, Vols. 1 and 2.

Bass, M. Stitch, M. L. (Eds.) (1985). *Laser Handbook*, Amsterdam: North-Holland, Vols. 4 and 5.

Hecht, E. (1987), *Optics*, 2nd ed., Reading, MA: Addison-Wesley.

Mystre, P., Walls, D. F. (Eds.) (1991), *Nonclassical Effects in Quantum Optics: Key Papers in Physics*, New York: AIP.

Stitch, M. L. (Ed.) (1979), *Laser Handbook*, Amsterdam: North-Holland, Vol. 3.

Yariv, A. (1989), *Quantum Electronics*, 3rd ed., New York: Wiley.

LATTICE VIBRATIONS, STATISTICS OF

I. P. Ipatova, *A. F. Ioffe Physico-Technical Institute, Russian Academy of Sciences, St. Petersburg, Russia*
A. A. Maradudin, *Department of Physics, University of California, Irvine, California, U.S.A.*

Introduction 465
1. **Dynamics of a Crystal Lattice** 466
2. **Quantization of Crystal Vibrations** 470
3. **The Debye Specific Heat** 472
4. **Lattice Thermal Conductivity of Insulators** 476
Works Cited 484
Further Reading 484

INTRODUCTION

The atoms that comprise a crystal are constantly executing small vibrations about equilibrium positions that form a structure periodic in each of three mutually perpendicular directions. At nonzero temperatures these atomic vibrations are due to thermal fluctuations in the crystal, while at the absolute zero of temperature they are due to the zero-point motion of the atoms that is associated with the quantum-mechanical uncertainty principle.

The vibrations of the atoms in a crystal play a significant role in many physical properties of the crystal. In insulating crystals, the thermodynamic, optical, elastic, and transport properties are determined by these atomic vibrations. Even in metallic and semiconducting crystals, in which the conduction electrons contribute to each of these properties, the contribution from the atomic vibrations is a significant part of the total.

The discussion of the vibrational properties of a crystal is based on the assumption that the potential energy of the crystal is a function of the instantaneous positions of the atoms in it. Because the displacements of the atoms from their equilibrium positions, as they oscillate about them, are small compared with the distance between nearest-neighbor equilibrium positions, the potential energy can be expanded in powers of these small displacements. The linear terms in this expansion are absent, since the expansion is about an equilibrium configuration. If the expansion is terminated at the terms quadratic in the atomic displacements, we have what is called the *harmonic approximation* to the potential energy. The higher-order terms in the expansion of the potential energy in powers of the atomic displacements, of third, fourth, ... order, constitute the *anharmonic* terms in the potential energy. In many cases these anharmonic terms contribute quantitative and qualitative corrections to physical properties of a crystal that are already calculated with good accuracy in the harmonic approximation, such as the thermodynamic functions. These corrections are usually small enough that they can be obtained by perturbation theory. However, as in the case of thermal expansion, they can also give rise to effects that have no counterpart in the harmonic approximation. Nevertheless, the harmonic approximation is generally a very good first approximation to the calculation of a vibrational property of a crystal, and serves as the starting point for perturbative calculations of anharmonic effects. (The exceptions to these statements are quantum crystals, e.g., solid helium, which we will not consider here.) Consequently, in what follows most of our discussion will be based on this approximation, and the effects of the anharmonic terms in the potential energy of a crystal will be discussed only briefly.

Although in principle the vibrational contribution to a physical property of a crystal can be calculated by considering the motions of the individual atoms in it, it turns out to be simpler, and physically more transparent, to exploit the special nature of the harmonic approximation—viz., that the potential energy is a positive-definite quadratic form in the

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

atomic displacements—to reformulate the calculation in such a way that the given physical property is expressed as the sum of contributions from collective motions of *all* the atoms in a crystal. Each of these collective motions, called a *normal vibration mode* of the crystal, is a phased linear superposition of the vibrations of the individual atoms, each with the same frequency, and is independent of all the other normal modes of the crystal. The number of these normal modes is equal to the number of degrees of freedom in the crystal. The statistical properties of the normal modes are central to the way in which they contribute to a given physical property of the crystal. In this article we describe these statistical properties, and their consequences, in particular for the thermodynamic and transport properties of a crystal.

1. DYNAMICS OF A CRYSTAL LATTICE

Let us consider an infinite crystal that is periodic in three mutually perpendicular directions. The equilibrium position of the κth atom in the lth primitive unit cell is given by the vector

$$\mathbf{x}(l\kappa)=\mathbf{x}(l)+\mathbf{x}(\kappa). \tag{1}$$

In this equation the vector $\mathbf{x}(l)$ is

$$\mathbf{x}(l)=l_1\mathbf{a}_1+l_2\mathbf{a}_2+l_3\mathbf{a}_3, \tag{2}$$

where l_i $(i=1,2,3)$ is an integer, and $\mathbf{a}_i$ $(i=1,2,3)$ are the primitive translation vectors of the crystal. $\mathbf{x}(l)$ defines the equilibrium position of the lth unit cell with respect to the origin of coordinates, chosen to coincide with one of the lattice sites. The vector $\mathbf{x}(\kappa)$ $(\kappa=1,2,\ldots,r)$ gives the positions of the r atoms in the unit cell relative to the equilibrium position of the latter. The volume of a primitive unit cell is given by $v_0=\mathbf{a}_1\cdot(\mathbf{a}_2\times\mathbf{a}_3)$.

The thermal motion of the atoms of the crystal consists of small displacements of all the atoms from their equilibrium positions, and we denote the displacement of the atom $(l\kappa)$ by $\mathbf{u}(l\kappa)$. The momentum of this atom will be denoted by $\mathbf{p}(l\kappa)$.

The total kinetic energy of the crystal can be expressed in the form

$$T=\sum_{l\kappa\alpha}\frac{p_\alpha^2(l\kappa)}{2M_\kappa}, \tag{3}$$

where M_κ is the mass of the atom of type κ, and $\alpha=x,y,z$. The total potential energy Φ of the crystal is a function of the instantaneous positions of all the atoms. It can be expanded in a series of powers of the displacements $\mathbf{u}(l\kappa)$ of the atoms from their equilibrium positions,

$$\Phi=\Phi_0+\sum_{l\kappa\alpha}\Phi_\alpha(l\kappa)u_\alpha(l\kappa)+\frac{1}{2}\sum_{l\kappa\alpha}\sum_{l'\kappa'\beta}\Phi_{\alpha\beta}(l\kappa;l'\kappa')u_\alpha(l\kappa)u_\beta(l'\kappa')+\cdots \tag{4}$$

where

$$\Phi_\alpha(l\kappa)=\left(\frac{\partial\Phi}{\partial u_\alpha(l\kappa)}\right)_0,$$

$$\Phi_{\alpha\beta}(l\kappa;l'\kappa')=\left(\frac{\partial^2\Phi}{\partial u_\alpha(l\kappa)\partial u_\beta(l'\kappa')}\right)_0,\quad \text{etc.} \tag{5}$$

The approximation in which the expansion (4) contains only terms through second order in $\mathbf{u}(l\kappa)$ is called the *harmonic approximation.* The terms neglected, of higher order in $\mathbf{u}(l\kappa)$, are called *anharmonic.* They are neglected because of their smallness.

From the definition (5) for $\Phi_\alpha(l\kappa)$ we see that it has the physical meaning of the negative of the force acting in the direction α on the atom $(l\kappa)$ when all the atoms of the crystal are at their equilibrium positions. Since in equilibrium no net force can be exerted on any atom, we must have

$$\Phi_\alpha(l\kappa)\equiv 0. \tag{6}$$

The Hamiltonian of the crystal $H=T+\Phi$ in the harmonic approximation therefore has the form

$$H=\Phi_0+\sum_{l\kappa\alpha}\frac{p_\alpha^2(l\kappa)}{2M_\kappa}+\frac{1}{2}\sum_{l\kappa\alpha}\sum_{l'\kappa'\beta}\Phi_{\alpha\beta}(l\kappa;l'\kappa')u_\alpha(l\kappa)u_\beta(l'\kappa'). \tag{7}$$

The coefficients $\{\Phi_{\alpha\beta}(l\kappa;l'\kappa')\}$ are called *atomic force constants.*

From their definition, Eq. (5), the atomic force constants possess the symmetry property

$$\Phi_{\alpha\beta}(l\kappa;l'\kappa')=\Phi_{\beta\alpha}(l'\kappa';l\kappa). \tag{8}$$

In addition, they are subject to restrictions that are imposed by the invariance of the potential energy against rigid-body translations and rotations of the crystal, which are expressed by (Maradudin *et al.*, 1971, pp. 8–10)

$$\sum_{l'\kappa'} \Phi_{\alpha\beta}(l\kappa;l'\kappa')=0, \tag{9}$$

$$\sum_{l'\kappa'} \Phi_{\alpha\beta}(l\kappa;l'\kappa')x_{\gamma}(l'\kappa') = \sum_{l'\kappa'} \Phi_{\alpha\gamma}(l\kappa;l'\kappa')x_{\beta}(l'\kappa'), \tag{10}$$

respectively. Finally, the symmetry and structure of a crystal impose further restrictions on the atomic force constants. The most important of these for our purposes in this article is a consequence of the periodicity of the equilibrium crystal structure. The potential energy (4) cannot change under a displacement of this structure through any of the vectors $\mathbf{x}(l)$ given by Eq. (2). This has the consequence that the atomic force constants $\{\Phi_{\alpha\beta}(l\kappa;l'\kappa')\}$ depend on the cell indices l and l' only through their difference,

$$\Phi_{\alpha\beta}(l\kappa;l'\kappa')=\Phi_{\alpha\beta}(l-l',\kappa;0\kappa') \tag{11a}$$

$$=\Phi_{\alpha\beta}(0\kappa;l'-l,\kappa'). \tag{11b}$$

Additional restrictions on the atomic force constants due to the symmetry of a crystal are discussed on pp. 10–14 of Maradudin *et al.* (1971).

From Hamilton's equations of motion that follow from Eq. (7),

$$\dot{u}_{\alpha}(l\kappa)=\partial H/\partial p_{\alpha}(l\kappa) \tag{12a}$$

$$\dot{p}_{\alpha}(l\kappa)=-\partial H/\partial u_{\alpha}(l\kappa), \tag{12b}$$

we immediately obtain the classical equations of motion of the atoms of the crystal,

$$M_{\kappa}\ddot{u}_{\alpha}(l\kappa)=-\sum_{l'\kappa'\beta} \Phi_{\alpha\beta}(l\kappa;l'\kappa')u_{\beta}(l'\kappa). \tag{13}$$

The solution of the system of equations (13) can be simplified if we exploit the periodicity of the equilibrium crystal structure expressed through Eqs. (11). Since the infinite crystal is periodic, it is natural to seek a solution of the system (13) in the form of a traveling wave,

$$u_{\alpha}(l\kappa)=\frac{u_{\alpha}(\kappa)}{\sqrt{M_{\kappa}}}\exp[i\mathbf{k}\cdot\mathbf{x}(l)-i\omega t], \tag{14}$$

where $\mathbf{k}$ is its wave vector, ω is its frequency, and the amplitude $u_{\alpha}(\kappa)$ is independent of l. On substituting Eq. (14) into Eqs. (13) we obtain as the equation for the $\{u_{\alpha}(\kappa)\}$

$$\omega^2 u_{\alpha}(\kappa)=\sum_{\kappa'\beta} D_{\alpha\beta}(\kappa\kappa'|\mathbf{k})u_{\beta}(\kappa'), \tag{15}$$

where the matrix $D_{\alpha\beta}(\kappa\kappa'|\mathbf{k})$ does not depend on the index l, and is equal to

$$D_{\alpha\beta}(\kappa\kappa'|\mathbf{k})=\frac{1}{\sqrt{M_{\kappa}M_{\kappa'}}}\sum_{l'} \Phi_{\alpha\beta}(l-l',\kappa;0\kappa') \times\exp\{-i\mathbf{k}\cdot[\mathbf{x}(l)-\mathbf{x}(l')]\}. \tag{16}$$

In contrast with the system of equations (13), which is a system of an infinite number of equations in an infinite number of unknowns, the system (15) consists of $3r$ equations for each value of the wave vector $\mathbf{k}$. This result is due to the fact that the force constants $\{\Phi_{\alpha\beta}(l\kappa;l'\kappa')\}$ depend on l and l' only through their difference.

The solvability condition for the system of equations (15) is the vanishing of its determinant:

$$|D_{\alpha\beta}(\kappa\kappa'|\mathbf{k})-\omega^2\delta_{\alpha\beta}\delta_{\kappa\kappa'}|=0. \tag{17}$$

Equation (17) is an equation of degree $3r$ in ω^2. We will denote the $3r$ solutions for each value of $\mathbf{k}$ by $\omega_j^2(\mathbf{k})$, where $j=1,2,...,3r$. From its definition (16) the matrix $D_{\alpha\beta}(\kappa\kappa'|\mathbf{k})$ can be shown to be Hermitian. From this it follows that all of its eigenvalues $\omega_j^2(\mathbf{k})$ are real, while the frequencies $\omega_j(\mathbf{k})$ themselves can be either purely real or purely imaginary. The stability of the crystal translates into the requirement that all the frequencies $\omega_j(\mathbf{k})$ be real. The $3r$ functions $\omega_j^2(\mathbf{k})$ for each value of $\mathbf{k}$ can be regarded as the branches of a multivalued function $\omega^2(\mathbf{k})$. Equation (15) can therefore be rewritten in a form that makes the dependence of the frequency ω and of the corresponding eigenvector $\mathbf{u}(\kappa)$ on the wave vector $\mathbf{k}$ and branch index j explicit,

$$\omega_j^2(\mathbf{k})u_{\alpha}(\kappa|\mathbf{k}j)=\sum_{\kappa'\beta} D_{\alpha\beta}(\kappa\kappa'|\mathbf{k})u_{\beta}(\kappa'|\mathbf{k}j). \tag{18}$$

The relation

$$\omega=\omega_j(\mathbf{k}) \tag{19}$$

is called the *dispersion law* for the vibrations of the crystal. Of the $3r$ solutions of Eq. (18) for each value of $\mathbf{k}$, three go to zero as $\mathbf{k}$ goes to zero. This can be seen as follows. On setting $\mathbf{k}=0$ in Eqs. (16) and (18), and dividing the latter by $\sqrt{M_\kappa}$, we obtain

$$\omega_j^2(\mathbf{0})\,\frac{u_\alpha(\kappa|\mathbf{0}j)}{\sqrt{M_\kappa}}=\frac{1}{M_\kappa}\sum_{l'\kappa'\beta}\Phi_{\alpha\beta}(l\kappa;l'\kappa')\times\frac{u_\beta(\kappa'|\mathbf{0}j)}{\sqrt{M_{\kappa'}}}. \tag{20}$$

If for each of the three values β can assume, viz., x, y, and z, $u_\beta(\kappa'|\mathbf{0}j)/\sqrt{M_{\kappa'}}$ is independent of κ', then in view of Eq. (9) we find that the right-hand side of Eq. (20) vanishes, implying the vanishing of $\omega_j^2(\mathbf{0})$. Thus we have three solutions for $\omega_j^2(\mathbf{k})$, one for each of the coordinate axes, that vanish with vanishing $\mathbf{k}$. From the form of Eq. (14) at $\mathbf{k}=0$ we see that the condition that $u_\alpha(\kappa|\mathbf{0}j)/\sqrt{M_\kappa}$ be independent of κ means that all r atoms in a primitive unit cell move in parallel and with equal amplitudes, which corresponds to a rigid-body displacement of the entire crystal. The three branches of $\omega^2(\mathbf{k})$ that vanish as $\mathbf{k}\to\mathbf{0}$ are termed *acoustic modes,* because the same property is shared by the frequencies of a continuous elastic medium. The remaining $3r-3$ solutions, which do not go to zero as $\mathbf{k}$ goes to zero, are called the frequencies of *optical modes.* The latter name is due to the fact that an electromagnetic wave incident on the crystal can interact only with the $\mathbf{k}=0$ modes whose frequencies are nonzero, and are generally in the far-infrared region of the optical spectrum.

The frequencies $\{\omega_j(\mathbf{k})\}$ belonging to the $3r$ branches of the vibrational spectrum possess the important property that they are even functions of $\mathbf{k}$,

$$\omega_j(-\mathbf{k})=\omega_j(\mathbf{k}). \tag{21}$$

For we see from Eq. (16) that

$$D_{\alpha\beta}(\kappa\kappa'|-\mathbf{k})=D^*_{\alpha\beta}(\kappa\kappa'|\mathbf{k}). \tag{22}$$

Consequently, if we replace $\mathbf{k}$ by $-\mathbf{k}$ in Eq. (18), and take the complex conjugate of the resulting equation, we obtain

$$\omega_j^2(-\mathbf{k})u^*_\alpha(\kappa|-\mathbf{k}j)=\sum_{\kappa'\beta}D_{\alpha\beta}(\kappa\kappa'|\mathbf{k})\times u^*_\beta(\kappa'|-\mathbf{k}j), \tag{23}$$

where we have used the fact that $\omega_j^2(\mathbf{k})$ is real because $D_{\alpha\beta}(\kappa\kappa'|\mathbf{k})$ is a Hermitian matrix. Thus $\omega_j^2(-\mathbf{k})$ and $\omega_j^2(\mathbf{k})$ are eigenvalues of the same matrix. With no loss of generality we can equate $\omega_j^2(-\mathbf{k})$ and $\omega_j^2(\mathbf{k})$, even for values of $\mathbf{k}$ at which degenerate eigenvalues occur, which leads to Eq. (21).

To each of the $3r$ values of $\omega_j(\mathbf{k})$ corresponds a vector $u_\alpha(\kappa|\mathbf{k}j)$ whose components are the solutions of Eq. (18). The homogeneous equations (18) define $u_\alpha(\kappa|\mathbf{k}j)$ to within an arbitrary constant factor. This factor can be chosen in such a way that $u_\alpha(\kappa|\mathbf{k}j)$ satisfies the orthonormality and closure conditions

$$\sum_{\kappa,\alpha}u^*_\alpha(\kappa|\mathbf{k}j)u_\alpha(\kappa|\mathbf{k}j')=\delta_{jj'}, \tag{24a}$$

$$\sum_{j}u^*_\alpha(\kappa|\mathbf{k}j)u_\beta(\kappa'|\mathbf{k}j)=\delta_{\alpha\beta}\delta_{\kappa\kappa'}. \tag{24b}$$

We have not yet specified the values that the wave vector $\mathbf{k}$ introduced in Eq. (14) can assume. They are clearly determined by the boundary conditions imposed on the atomic displacements $\{\mathbf{u}(l\kappa)\}$. The most convenient boundary conditions, which preserve the simplicity of the theory developed so far, are obtained if we subdivide the infinite crystal into "macrocrystals," parallelepipeds whose edges are defined by the vectors $L\mathbf{a}_1, L\mathbf{a}_2, L\mathbf{a}_3$, where L is a large number, of the order of 10^8, and which thus contain $L^3=N$ primitive unit cells. Any one of these macrocrystals can be regarded as the physical crystal whose properties we are studying. We then postulate that the atomic displacements are periodic with the periodicity of the macrocrystals. These conditions on the displacements are called the *cyclic boundary conditions* (Born and Huang, 1954, pp. 45–46). An equivalent set of boundary conditions require that in a finite crystal the atoms at corresponding points on opposite faces of the crystal move in exactly the same way. These conditions are called *periodic boundary conditions,* and are due to Born and von Kármán (1912, 1913).

Applied to the displacements given by Eq. (14), the cyclic boundary conditions require that

$$\exp(i\mathbf{k}\cdot L\mathbf{a}_1)=\exp(i\mathbf{k}\cdot L\mathbf{a}_2)=\exp(i\mathbf{k}\cdot L\mathbf{a}_3)=1. \tag{25}$$

To satisfy these conditions we introduce the translation vectors of the lattice reciprocal to the direct lattice defined by the vectors $\{\mathbf{x}(l)\}$:

$$\mathbf{G}(h)=h_1\mathbf{b}_1+h_2\mathbf{b}_2+h_3\mathbf{b}_3, \tag{26}$$

where h_i $(i=1,2,3)$ is an integer, and the primitive translation vectors of this *reciprocal lattice* $\mathbf{b}_i$ $(i=1,2,3)$ are defined by the set of equations

$$\mathbf{a}_i\cdot\mathbf{b}_j=2\pi\delta_{ij}. \tag{27}$$

Then from the result which follows from Eqs. (2), (26), and (27) that

$$\mathbf{x}(l)\cdot\mathbf{G}(h)=2\pi(l_1h_1+l_2h_2+l_3h_3)=2\pi(\text{integer}), \tag{28}$$

we see that we can write the vector $\mathbf{k}$ that satisfies Eqs. (25) as

$$\mathbf{k}=\frac{1}{L}\mathbf{G}(h)=\frac{h_1}{L}\mathbf{b}_1+\frac{h_2}{L}\mathbf{b}_2+\frac{h_3}{L}\mathbf{b}_3. \tag{29}$$

The values of the integers $\{h_i\}$ are no longer unrestricted, however. From Eqs. (14) and (28) we see that adding to $\mathbf{k}$ any translation vector $\mathbf{G}(h)$ of the reciprocal lattice leaves the value of $\mathbf{u}(l\kappa)$ unchanged. This means that all the distinct solutions to Eq. (18) are obtained if we restrict $\mathbf{k}$ to lie in one unit cell of the reciprocal lattice:

$$\mathbf{k}=\frac{h_1}{L}\mathbf{b}_2+\frac{h_2}{L}\mathbf{b}_2+\frac{h_3}{L}\mathbf{b}_3,$$

$$h_1,h_2,h_3=0,1,2,\ldots,L-1. \tag{30}$$

Thus, there are $L^3=N$ values of $\mathbf{k}$ allowed by the cyclic boundary conditions. Since there are $3r$ solutions of Eq. (18) for each value of $\mathbf{k}$, we see that the macrocrystal has as many vibration frequencies $\{\omega_j(\mathbf{k})\}$ as it has degrees of freedom.

A related consequence of Eq. (28) is that the dynamical matrix $D_{\alpha\beta}(\kappa\kappa'|\mathbf{k})$ defined by Eq. (16) is unchanged if we add to the wave vector $\mathbf{k}$ any reciprocal-lattice vector $\mathbf{G}(h)$. It follows that the normal-mode frequency $\omega_j(\mathbf{k})$ possesses the same property: $\omega_j(\mathbf{k})=\omega_j(\mathbf{k}+\mathbf{G}(h))$. Thus, all the distinct normal-mode frequencies are obtained by restricting $\mathbf{k}$ to one unit cell of the reciprocal lattice.

In practice, it is more convenient to restrict the allowed values of the wave vector $\mathbf{k}$ to the *first Brillouin zone* of the lattice defined by the vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, rather than to a unit cell of the reciprocal lattice. The first Brillouin zone is obtained by drawing vectors from the origin of the reciprocal lattice to all the points of that lattice and constructing the planes that are the perpendicular bisectors of these vectors. The smallest volume containing the origin that is enclosed by these planes is the first Brillouin zone. It has the same volume as a primitive unit cell of the reciprocal lattice, and so contains the same number N of allowed values of $\mathbf{k}$ as the latter does, and it is invariant under the operations of the crystal class to which the crystal belongs. In what follows, therefore, sums over the wave vector $\mathbf{k}$ will be understood to be over those N values allowed by the cyclic boundary conditions that are contained in the first Brillouin zone.

In most calculations the precise values that $\mathbf{k}$ is allowed to have are not important. What is important is that they are distributed uniformly and densely throughout the Brillouin zone with a density equal to $V/(2\pi)^3$, where $V=Nv_0$ is the volume of the crystal. Thus, in sums over $\mathbf{k}$ we can treat $\mathbf{k}$ as a continuous variable and replace summation by integration according to

$$\sum_k f(\mathbf{k})=\frac{V}{(2\pi)^3}\int d^3k f(\mathbf{k}), \tag{31}$$

where the integration is over the volume of the first Brillouin zone.

Since the rather artificial cyclic boundary conditions on the atomic displacements have been adopted primarily for reasons of mathematical convenience, it has to be shown that the calculation of extensive vibrational properties of the crystal is insensitive to the imposition of these or any other boundary conditions. If this were not so, it would be virtually impossible to develop a theory of crystalline vibrations that possesses any degree of generality. Such a demonstration was given by Ledermann (1944), who showed that in the limit of large N the distribution of vibration frequencies of a crystal is affected negligibly by the assumption of any particular boundary conditions in the case that the range of the interatomic forces is finite. Subsequently, Hardy (1962) provided a justification for the use of the cyclic boundary conditions in calculations of extensive properties of ionic or polar crystals, in which the constituent atoms or ions interact through long-range Coulomb forces.

2. QUANTIZATION OF CRYSTAL VIBRATIONS

The vibrational Hamiltonian of a crystal,

$$H=\sum_{l\kappa\alpha}\frac{p_\alpha^2(l\kappa)}{2M_\kappa}+\tfrac{1}{2}\sum_{l\kappa\alpha}\sum_{l'\kappa'\beta}\Phi_{\alpha\beta}(l\kappa;l'\kappa')\times u_\alpha(l\kappa)u_\beta(l'\kappa'), \qquad (32)$$

is the sum of quadratic forms in the displacement components $\{u_\alpha(l\kappa)\}$ and in the conjugate momenta $\{p_\alpha(l\kappa)\}$. It can be diagonalized if we introduce the following expansions of $u_\alpha(l\kappa)$ and $p_\alpha(l\kappa)$ in traveling plane waves:

$$u_\alpha(l\kappa)=\frac{1}{\sqrt{NM_\kappa}}\sum_{\mathbf{k}j}u_\alpha(\kappa|\mathbf{k}j)\left\{\frac{1}{2}\,[q(\mathbf{k}j)+q(-\mathbf{k}j)]+\frac{1}{2}\left[\frac{i}{\omega_j(\mathbf{k})}\,\{\mathbf{p}(\mathbf{k}j)-p(-\mathbf{k}j)\}\right]\right\}e^{i\mathbf{k}\cdot\mathbf{x}(l)}, \qquad (33a)$$

$$p_\alpha(l\kappa)=\frac{1}{\sqrt{NM_\kappa}}\sum_{\mathbf{k}j}u_\alpha(\kappa|\mathbf{k}j)\left\{\frac{1}{2}\,[p(\mathbf{k}j)+p(-\mathbf{k}j)]+\frac{1}{2}\left[\frac{i}{\omega_j(\mathbf{k})}\,\{q(\mathbf{k}j)-q(-\mathbf{k}j)\}\right]\right\}e^{i\mathbf{k}\cdot\mathbf{x}(l)}. \qquad (33b)$$

In these expansions the quantities in square brackets are real displacements and momenta, corresponding to waves propagating with the wave vector **k**. In the new variables $p(\mathbf{k}j)$ and $q(\mathbf{k}j)$ the diagonalized Hamiltonian becomes the sum of the energies of independent harmonic oscillations,

$$H=\frac{1}{2}\sum_{\mathbf{k}j}[p^2(\mathbf{k}j)+\omega_j^2(\mathbf{k})q^2(\mathbf{k}j)]. \qquad (34)$$

In going from Eq. (32) to Eq. (34) we have used the result that

$$\sum_l e^{i\mathbf{k}\cdot\mathbf{x}(l)}=N\Delta(\mathbf{k}), \qquad (35)$$

where

$$\Delta(\mathbf{k})=\begin{cases}1 & \text{if } \mathbf{k}=\mathbf{G}(h),\\ 0 & \text{otherwise.}\end{cases} \qquad (36)$$

The vibrations described by the Hamiltonian (34) are called *normal modes,* and the corresponding coordinates $q(\mathbf{k}j)$ and momenta $p(\mathbf{k}j)$ are called *normal coordinates.* They are related by $p(\mathbf{k}j)=\dot{q}(\mathbf{k}j)$.

The Hamiltonian (34) is written in a form that permits a straightforward transition to a quantum-mechanical description of the vibrations of the crystal. For this we must regard $q(\mathbf{k}j)$ and $p(\mathbf{k}j)$ as quantum-mechanical operators obeying commutation relations that contain Planck's constant,

$$[q(\mathbf{k}j),p(\mathbf{k}j)]=i\hbar\Delta(\mathbf{k}-\mathbf{k}')\delta_{jj'}. \qquad (37)$$

It follows from Eq. (37) that

$$p(\mathbf{k}j)=-i\hbar\,\partial/\partial q(\mathbf{k}j). \qquad (38)$$

On substituting Eq. (38) into Eq. (34) we obtain the quantum-mechanical Hamiltonian operator of the crystalline vibrations,

$$\hat{H}=\frac{1}{2}\sum_{\mathbf{k}j}\left[-\hbar^2\frac{\partial^2}{\partial q^2(\mathbf{k}j)}+\omega_j^2(\mathbf{k})q^2(\mathbf{k}j)\right]. \qquad (39)$$

The corresponding Schrödinger equation has the form

$$\sum_{\mathbf{k}j}\left[-\frac{\hbar^2}{2}\frac{\partial^2}{\partial q^2(\mathbf{k}j)}+\tfrac{1}{2}\omega_j^2(\mathbf{k})q^2(\mathbf{k}j)\right]\Psi=E\Psi. \qquad (40)$$

Here E and Ψ are the energy and wave function of the vibrations, respectively. Since the Hamiltonian in Eq. (39) has the form of a sum of independent harmonic-oscillator Hamiltonians, one for each of the normal modes, the wave function Ψ can be represented as the product of the wave functions of the individual oscillators,

$$\Psi=\prod_{\mathbf{k}j}\psi_{n_j(\mathbf{k})}(q(\mathbf{k}j)), \qquad (41)$$

while the energy E is the sum of the energies of the individual oscillators,

$$E = \sum_{\mathbf{k}j} E_{n_j(\mathbf{k})}. \tag{42}$$

In Eqs. (41) and (42) $n_j(\mathbf{k})$ is a quantum number that labels the energy level of the oscillator ($\mathbf{k}j$). By substituting Eqs. (41) and (42) into Eq. (40) we find that the Schrödinger equation for an individual normal mode has the form

$$\left[-\frac{\hbar^2}{2} \frac{\partial^2}{\partial q^2(\mathbf{k}j)} + \frac{1}{2}\omega_j^2(\mathbf{k}) q^2(\mathbf{k}j) \right] \psi_{n_j(\mathbf{k})}(q(\mathbf{k}j)) = E_{n_j(\mathbf{k})} \psi_{n_j(\mathbf{k})}(q(\mathbf{k}j)). \tag{43}$$

The solution of this equation is well known from quantum mechanics (Schiff, 1949, p. 64):

$$E_{n_j(\mathbf{k})} = [n_j(\mathbf{k}) + \tfrac{1}{2}]\hbar\omega_j(\mathbf{k}) \tag{44}$$

and

$$\psi_n(q) = \left(\frac{\alpha}{\pi^{1/2} 2^n n!} \right)^{1/2} \exp(-\tfrac{1}{2}\alpha^2 q^2) H_n(\alpha q), \tag{45}$$

where $\alpha_j^2(\mathbf{k}) = \omega_j(k)/\hbar$, and $H_n(\alpha q)$ is the Hermite polynomial of nth degree, where $n = 0,1,2,\ldots$. The total vibrational energy of the crystal for the state characterized by the quantum numbers $\{n_j(\mathbf{k})\}$ is then given by

$$E_{\{n_j(\mathbf{k})\}} = \sum_{\mathbf{k}j} [n_j(\mathbf{k}) + \tfrac{1}{2}]\hbar\omega_j(\mathbf{k}). \tag{46}$$

The integer $n_j(\mathbf{k})$ is the number of quanta of vibrational energy $\hbar\omega_j(\mathbf{k})$ in the normal mode ($\mathbf{k}j$). This is often expressed by the statement that there are $n_j(\mathbf{k})$ *phonons* in the mode ($\mathbf{k}j$). The average number of phonons in this mode, $\bar{n}(\omega_j(\mathbf{k}))$, is temperature-dependent, and is simply obtained. Since the contribution to the total vibrational energy of the crystal is the sum of the energies of each of the vibrational modes, from the general principles of statistical physics we have that

$$\bar{n}(\omega_j(\mathbf{k})) = \frac{\sum_{n_j(\mathbf{k})=0}^{\infty} n_j(\mathbf{k}) e^{-[n_j(\mathbf{k}) + \frac{1}{2}]\hbar\omega_j(\mathbf{k})/T}}{\sum_{n_j(\mathbf{k})=0}^{\infty} e^{-[n_j(\mathbf{k}) + \frac{1}{2}]\hbar\omega_j(\mathbf{k})/T}}, \tag{47}$$

where T is the absolute temperature in energy units, i.e., Boltzmann's constant $k_B \equiv 1$. The sums are readily evaluated, and we obtain the result

$$\bar{n}(\omega_j(\mathbf{k})) = 1/(e^{\hbar\omega_j(\mathbf{k})/T} - 1). \tag{48}$$

Thus we find that the equilibrium distribution of phonons in the normal mode ($\mathbf{k}j$) is described by the Bose-Einstein distribution function, in which the chemical potential μ is set equal to zero,

$$\mu = 0. \tag{49}$$

The fact that the chemical potential of phonons is zero indicates that the creation and destruction of a phonon are accomplished without the expenditure of any energy.

The distribution function (48) is called the Planck distribution function. It has the following limiting forms:

$$\bar{n}(\omega_j(\mathbf{k})) = \begin{cases} \exp[-\hbar\omega_j(\mathbf{k})/T] \text{ as } T \to 0, & \text{(50a)} \\ T/\hbar\omega_j(\mathbf{k}) \text{ as } T \to \infty. & \text{(50b)} \end{cases}$$

Thus, in the limit of low temperatures, $T \ll \hbar\omega_j(\mathbf{k})$, very few phonons are present in the mode ($\mathbf{k}j$). However, in the high-temperature limit, $T \gg \hbar\omega_j(\mathbf{k})$, the mean number of phonons in this mode is proportional to the absolute temperature T.

The thermodynamic properties of the crystal can be completely described with the aid of the partition function, which in the present case has the form

$$Z = \sum_{\{n_j(\mathbf{k})\}} \exp(-E_{\{n_j(\mathbf{k})\}}/T). \tag{51}$$

If we substitute into this expression the sum (46), we obtain

$$Z = \prod_{\mathbf{k}j} \frac{\exp[-\hbar\omega_j(\mathbf{k})/T]}{1 - \exp[-\hbar\omega_j(\mathbf{k})/T]}. \tag{52}$$

The Helmholtz free energy of the crystalline vibrations can be obtained from the partition function

$$F = -T \ln Z = T \sum_{\mathbf{k}j} \ln\left[2 \sinh \frac{\hbar\omega_j(\mathbf{k})}{2T} \right], \tag{53}$$

as can the internal energy

$$E=F-T\left(\frac{\partial F}{\partial T}\right)_v$$

$$=\sum_{\mathbf{k}j}\left\{\frac{\hbar\omega_j(\mathbf{k})}{2}+\frac{\hbar\omega_j(\mathbf{k})}{\exp[\hbar\omega_j(\mathbf{k})/T]-1}\right\}, \quad (54)$$

the specific heat at constant volume

$$C_v=\left(\frac{\partial E}{\partial T}\right)_v=\sum_{\mathbf{k}j}\frac{[\hbar\omega_j(\mathbf{k})/2T]^2}{\sinh^2[\hbar\omega_j(\mathbf{k})/2T]}, \quad (55)$$

and the entropy

$$S=-\left(\frac{\partial F}{\partial T}\right)_v$$

$$=\sum_{\mathbf{k}j}\left\{\frac{\hbar\omega_j(\mathbf{k})}{2T}\coth\frac{\hbar\omega_j(\mathbf{k})}{2T}-\ln\left[2\sinh\frac{\hbar\omega_j(\mathbf{k})}{2T}\right]\right\}. \quad (56)$$

If the vibrational spectrum of the crystal $\omega_j(\mathbf{k})$ for $j=1,2,\ldots,3r$ is known, then Eqs. (53)–(56) in principle allow all four thermodynamic functions F, E, C_v, S to be calculated.

However, the number of atoms in a crystal is enormous, and the vibrational frequencies are very densely distributed in the spectrum. Therefore, it is more convenient in the calculation of these functions to use the frequency distribution function $g(\omega)$, which can be introduced in the following fashion. Let us consider an arbitrary additive function of the normal-mode frequencies,

$$S=\sum_{\mathbf{k}j}F(\omega_j(\mathbf{k})). \quad (57)$$

It can be transformed in the following way:

$$S=\int d\omega\sum_{\mathbf{k}j}\delta(\omega-\omega_j(\mathbf{k}))F(\omega)$$

$$=\int d\omega g(\omega)F(\omega), \quad (58)$$

where $\delta(\omega)$ is the Dirac delta function. The frequency distribution function $g(\omega)$ is therefore given by

$$g(\omega)=\sum_{\mathbf{k}j}\delta(\omega-\omega_j(\mathbf{k})). \quad (59)$$

The function $g(\omega)$ has the meaning of the density of vibrational states, where $g(\omega)d\omega$ is the number of normal modes with frequencies in the infinitesimally narrow interval between ω and $\omega+d\omega$. The frequency distribution function $g(\omega)$ therefore satisfies the normalization condition

$$\int_0^{\omega_{\max}} d\omega\, g(\omega)=3rN, \quad (60)$$

which indicates that the total number of normal modes of the crystal is $3rN$. In Eq. (60) $\omega_{\max}$ is the largest frequency of the vibrational spectrum.

The expressions for the free energy, the internal energy, the specific heat, and the entropy can also be expressed in terms of $g(\omega)$, and have the forms of averages over the frequency distribution function:

$$F=T\int_0^{\omega_{\max}} d\omega\, g(\omega)\ln\left[2\sinh\frac{\hbar\omega}{2T}\right], \quad (61)$$

$$E=\int_0^{\omega_{\max}} d\omega\, g(\omega)\hbar\omega\left[\frac{1}{2}+\frac{1}{\exp(\hbar\omega/T)-1}\right], \quad (62)$$

$$C_v=\int_0^{\omega_{\max}} d\omega\, g(\omega)\left(\frac{\hbar\omega}{2T}\right)^2\frac{1}{\sinh^2(\hbar\omega/2T)}, \quad (63)$$

$$S=\int_0^{\omega_{\max}} d\omega\, g(\omega)\left[\frac{\hbar\omega}{2T}\coth\frac{\hbar\omega}{2T}-\ln\left(2\sinh\frac{\hbar\omega}{2T}\right)\right]. \quad (64)$$

3. THE DEBYE SPECIFIC HEAT

According to classical theory the specific heat of a solid at constant volume is determined solely by the number of degrees of freedom of the system, with each of which is associated an energy T. The resulting specific heat $C_v=\partial E/\partial T=3rN$ has a constant value, independent of temperature. This constant value was observed experimentally at elevated temperatures, and the constancy of the specific heat at such temperatures became known as the law of Dulong and Petit. At lower temperatures, however, it turned out that the specific heat is strongly temperature dependent, specifically that it decreases with decreasing temperature.

In 1893, to give a theoretical basis for the law of Dulong and Petit, Richarz (1893, 1899) was led to characterize a crystal as an assembly of atoms each of which is executing a simple harmonic motion about its mean position in the crystal. Several years later Einstein (1907), in attempting to explain the optical properties of diamond, adopted the same model of a crystalline solid, and assigned the characteristic frequency ω_0 to the motion of each atom. In this way he was led to his well known theory of lattice specific heats. The frequency distribution function $g(\omega)$ in this case can be represented by a suitably normalized delta function,

$$g(\omega)=3rN\delta(\omega-\omega_0). \tag{65}$$

On substituting Eq. (65) into Eq. (63) we obtain an exponential dependence of the specific heat on temperature of the form

$$C_v\sim\exp(-\hbar\omega_0/T), \tag{66}$$

in the low-temperature limit. However, more detailed experimental investigations of the temperature dependence of the specific heat did not confirm this strong, exponential, temperature dependence.

This discrepancy prompted several investigations of the dynamical properties of crystals, which culminated with the publication, in 1912, of fundamental papers by Debye (1912) and by Born and von Kármán (1912, 1913), which laid the basis for the modern theory of lattice dynamics. The theories of Debye and of Born and von Kármán differed from that of Einstein in their recognition that the atoms of a crystal do not all vibrate with the same frequency, but that there is a distribution of frequencies available to the atomic motions. [Even before the work of Debye and of Born and von Kármán, Einstein (1911) had suggested that this is the case to explain the disagreement between the predictions of his theory and experiment.] Debye regarded a crystal as a continuous elastic medium, whose motion was described by the equations of elasticity theory in calculating the distribution of vibrational frequencies of a crystal. It is known that the theory of elasticity describes the long-wavelength, low-frequency vibration modes of a crystal correctly [see, e.g., Maradudin *et al.* (1971), pp. 21–31]. Debye made the additional, simplifying assumption that the elastic medium is elastically isotropic. In contrast, Born and von Kármán took explicit account of the discrete and periodic nature of a real crystal. It is essentially the latter theory that has been outlined in the preceding two sections of this article. Although the theory of Born and von Kármán was more comprehensive than that of Debye, the simplicity of the latter in comparison to the former had the consequence that Debye's theory eclipsed that of Born and von Kármán for more than 20 years, and even today finds widespread application in the study of the dynamical properties of solids. Therefore, we devote the remainder of this section to a discussion of Debye's approach.

If we denote by $\mathbf{u}(\mathbf{x},t)$ the displacement of an arbitrary point $\mathbf{x}$ of an isotropic elastic medium at time t, the classical equations of motion of such a medium have the form

$$\rho\,\frac{\partial^2\mathbf{u}}{\partial t^2}=\mu\nabla^2\mathbf{u}+(\lambda+\mu)\nabla(\nabla\cdot\mathbf{u}). \tag{67}$$

Here $\rho=v_0^{-1}\,\Sigma_\kappa\,M_\kappa$ is the mass density of the medium, where v_0 is the volume of a primitive unit cell. The quantities μ and λ are called the Lamé elastic coefficients. They describe the shear and hydrostatic deformations existing in a moving isotropic continuum. The right-hand side of Eq. (67) is the force acting on an infinitesimally small volume element of the elastic medium; the left-hand side is the inertial force.

A traveling elastic wave propagating in the medium can be described by a displacement vector of the form

$$\mathbf{u}(\mathbf{x};t)=\mathbf{u}_0\,\exp(i\mathbf{k}\cdot\mathbf{x}-i\omega t). \tag{68}$$

On substituting this expression into Eq. (67) we obtain

$$\omega^2\rho\mathbf{u}_0=\mu k^2\mathbf{u}_0+(\lambda+\mu)\mathbf{k}(\mathbf{k}\cdot\mathbf{u}_0). \tag{69}$$

If the amplitude of the wave is parallel to the wave vector $\mathbf{k}$, $\mathbf{u}_{0l}\|\ \mathbf{k}$, we speak of a longitudinal wave. In this case Eq. (69) takes the form

$$[\omega_l^2\rho-(\lambda+2\mu)k^2]\mathbf{u}_{0l}=0. \tag{70}$$

The frequency of a longitudinal wave is therefore equal to

$$\omega_l(\mathbf{k})=[(\lambda+2\mu)/\rho]^{1/2}k. \tag{71}$$

For two transverse waves the amplitude is perpendicular to the wave vector $\mathbf{k}$, $\mathbf{u}_{0t}\perp\ \mathbf{k}$. In this case the second term on the right-hand side of Eq. (69) vanishes. The frequency of the two degenerate transverse waves is therefore obtained from the equation

$$(\omega_t^2\rho-\mu k^2)\mathbf{u}_{0t}=0, \tag{72}$$

and equals

$$\omega_t(\mathbf{k})=(\mu/\rho)^{1/2}k. \tag{73}$$

The quantities

$$c_l=[(\lambda+2\mu)/\rho]^{1/2} \quad \text{and} \quad c_t=(\mu/\rho)^{1/2} \tag{74}$$

are the speeds of one longitudinal and two transverse elastic waves, respectively.

We can now obtain the frequency distribution function $g(\omega)$ for Debye's model of a crystal. The general definition of $g(\omega)$ is given in Eq. (59). In Debye's model there is one longitudinal and two transverse modes, so that the frequency distribution function assumes the form

$$g(\omega)=\sum_{\mathbf{k}}\delta(\omega-c_lk)+2\sum_{\mathbf{k}}\delta(\omega-c_tk). \tag{75}$$

By going over from summation over $\mathbf{k}$ to integration, with the use of Eq. (31), we obtain

$$g(\omega)=\frac{V}{(2\pi)^3}\int d^3k\,\delta(\omega-c_lk)+\frac{2V}{(2\pi)^3}\int d^3k\,\delta(\omega-c_tk). \tag{76}$$

On carrying out the integrations we find

$$g(\omega)=\frac{V}{(2\pi)^3}4\pi\left(\frac{1}{c_l^3}+\frac{2}{c_t^3}\right)\omega^2. \tag{77}$$

From this result we see that $g(\omega)$ increases without limit with increasing ω. However, a real crystal possesses only $3rN$ normal mode frequencies, among which, of course, is a maximum frequency. Therefore, to impart a greater physical sense to the Debye model it was necessary to impose a normalization condition on the frequency distribution function obtained through its use, viz.,

$$\int_0^{\omega_D}\frac{V}{(2\pi)^3}4\pi\left(\frac{1}{c_l^3}+\frac{2}{c_t^3}\right)\omega^2d\omega=3Nr, \tag{78}$$

where the maximum frequency of the Debye model, ω_D, is defined by this condition. On evaluating the integral on the left-hand side of this equation we find that

$$\omega_D=2\pi\left(\frac{9Nr}{4\pi V}\right)^{1/3}\left(\frac{1}{c_l^3}+\frac{2}{c_t^3}\right)^{-1/3}. \tag{79}$$

It is convenient to rewrite $g(\omega)$ given by Eq. (77) in terms of ω_D, when it becomes

$$g(\omega)=\begin{cases}3rN(3\omega^2/\omega_D^3) & 0\leqslant\omega\leqslant\omega_D\\ 0 & \omega>\omega_D.\end{cases} \tag{80}$$

It is useful to note that ω_D is an artificial limiting frequency that is introduced into the Debye model so that we can take into account correctly the total number of degrees of freedom in the system. It may not coincide with the true maximum frequency of the vibrational spectrum of the crystal.

The frequency distribution function (80) can be used to calculate the specific heat. The result obtained describes well the temperature dependence of the low-temperature specific heat, since only the low-frequency part of the frequency distribution function is significant for the determination of the latter. On substituting Eq. (80) into Eq. (63) we obtain for the specific heat

$$C_v(T)=3rN\left(\frac{T}{\Theta_D}\right)^3\int_0^{\Theta_D/T}dx\,\frac{x^4e^x}{(e^x-1)^2}, \tag{81}$$

where we have introduced the Debye characteristic temperature

$$\Theta_D=\hbar\omega_D. \tag{82}$$

In the limiting case of low temperatures, $T\ll\Theta_D$, the upper limit on the integral in Eq. (81) can be replaced by infinity. In this case the integral can be evaluated exactly, and we obtain the following dependence of the specific heat on temperature:

$$C_v(T)=3rN\frac{4\pi^4}{5}\left(\frac{T}{\Theta_D}\right)^3. \tag{83}$$

This T^3 dependence of the low-temperature specific heat was in agreement with experimental results.

From Eq. (81) we see that in the Debye theory the specific heat is characterized by the single, dimensionless parameter T/Θ_D. This means that the dependence of C_v on T given by Eq. (81) is universal. This means in turn that all experimental curves of the temperature dependence of the specific heat will coincide if the appropriate value of Θ_D is assigned to each solid.

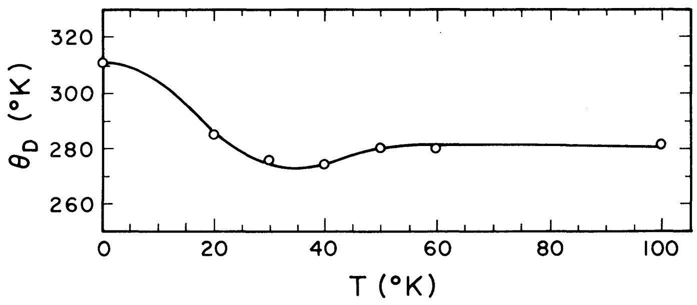

FIG. 1. The temperature dependence of the Debye characteristic temperature Θ_D for NaCl (Kellermann, 1941).

It turned out, however, that in order to obtain agreement between the Debye specific heat and an experimentally measured specific heat it was necessary to use different values of Θ_D at different temperatures. It was shown by Blackman (1934a, 1934b, 1936, 1937a, 1937b) that this temperature dependence of Θ_D can be obtained by equating the specific heat calculated numerically from the Born–von Kármán theory of lattice dynamics for a particular solid to the Debye expression for the specific heat (81). If the equation obtained in this way is regarded as an equation for Θ_D, one can establish the theoretical dependence of Θ_D on temperature. The comparison of this theoretical temperature dependence of Θ_D with that obtained from experimental data can serve as a sensitive criterion for the quality of the force-constant models used in the calculation of vibrational frequency distribution functions of crystals.

In Fig. 1 we show the temperature dependence of Θ_D for NaCl calculated by Kellermann (1941) by the method described above. From the figure we see that in the low-temperature limit Θ_D is a constant independent of temperature only at the very lowest temperatures, up to 5–7 K. Consequently, it is only at these low temperatures that the Debye T^3 law (83) is valid. The Debye temperature Θ_D is also constant over a broad temperature range at elevated temperatures. The reason for this will be discussed below.

If we compare the frequency distribution function used by Kellermann with the Debye frequency distribution function (Fig. 2), we see that they coincide only for frequencies up to the order of $\hbar\omega \sim 0.1\Theta_D$. At first glance it would seem, therefore, that the Debye expression for the specific heat, Eq. (81), should describe the temperature dependence of the specific heat in the temperature range $0 \leqslant T \leqslant 0.1\Theta_D$ ($T \lesssim 20$–50 K), in which only those normal modes of the crystal are excited that are well described by the Debye theory. As is seen from Fig. 1, the experimental data establish a different, smaller, upper limit for the applicability of the Debye theory ($T \lesssim 5$–7 K). This circumstance is related to the form of the integrand in the Debye expression (81), which is $x^4e^x/(e^x-1)^2$. For small values of x this function increases as x^2, while for large x it decreases as x^4e^{-x}. The maximum of this function occurs at $x \cong 3.922$. Consequently, the departure of the Debye frequency distribution function from the exact frequency distribution function will manifest itself in the temperature dependence of the specific heat at temperatures $T \gtrsim \frac{1}{4}(\hbar\omega/2)$. Thus, if the exact frequency distribution function coincides with the Debye frequency distribution function up to frequencies $\hbar\omega \lesssim 0.1\Theta_D$, then the Debye expression for the specific heat will describe the temperature dependence of the specific heat exactly only up to temperatures $T \lesssim 0.1\Theta_D/8 \cong 0.01\Theta_D$.

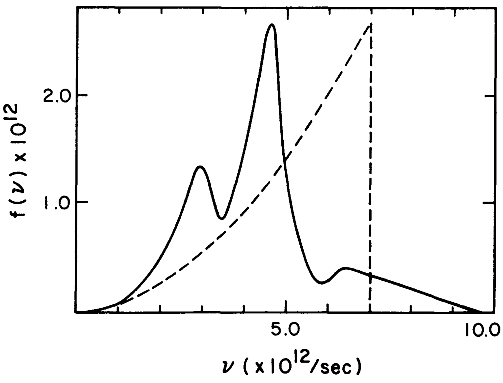

FIG. 2. The frequency distribution function $g(\omega)/3rN$ of NaCl (solid curve) compared with the corresponding Debye approximation to it (Born and Huang, 1954, p. 78).

In the high-temperature limit, $T > \hbar\omega_{\max}$, the specific heat of an arbitrary crystal, in the harmonic approximation, can be obtained by expanding the integrand in Eq. (63) in powers of T^{-1} and integrating the result term by term. The result can be expressed in the form

$$C_v = 3rN\left[1 - \sum_{n=1}^{\infty} (-1)^{n+1} \frac{2n-1}{(2n)!} \times B_{2n} u_{2n} \left(\frac{\hbar\omega_{\max}}{T}\right)^{2n}\right]. \tag{84}$$

In this expression, known as *Thirring's expansion* (Thirring, 1913, 1914a, 1914b), the B_{2n} are the Bernoulli numbers (Adams, 1957),

$$B_2 = \tfrac{1}{6},\quad B_4 = \tfrac{1}{30},\quad B_6 = \tfrac{1}{42},\quad B_8 = \tfrac{1}{30},$$
$$B_{10} = \tfrac{5}{66},\quad B_{12} = \tfrac{691}{2730},\quad B_{14} = \tfrac{7}{6},\ \ldots, \tag{85}$$

while u_{2n} is the dimensionless $2n$th moment of the frequency distribution function $g(\omega)$,

$$u_{2n} = \frac{1}{3rN\omega_{\max}^{2n}} \int_0^{\omega_{\max}} d\omega\, g(\omega)\omega^{2n} g(\omega). \tag{86}$$

The attractiveness of the expansion (84) is that the moments $\{u_{2n}\}$ can be calculated without an explicit knowledge of $g(\omega)$ from the result that

$$u_{2n} = \frac{1}{3rN\omega_{\omega_{\max}}^{2n}} \sum_{\mathbf{k}j} \omega_j^{2n}(\mathbf{k}) = \frac{1}{3rN\omega_{\omega_{\max}}^{2n}} \sum_{\mathbf{k}} \mathrm{Tr} D^n(\mathbf{k}), \tag{87}$$

where $D(\mathbf{k})$ is the dynamical matrix of the crystal, whose elements are given by Eq. (16). In the Debye approximation, it is found from Eqs. (80) and (86) that

$$u_{2n} = 3/(2n+3), \tag{88}$$

so that Eq. (84) in this approximation becomes

$$C_v = 3rN\left[1 - \sum_{n=1}^{\infty} (-1)^{n+1} \frac{3(2n-1)}{2n+3} \times \frac{B_{2n}}{(2n)!}\left(\frac{\Theta_D}{T}\right)^{2n}\right]. \tag{89}$$

If we equate Eqs. (84) and (89) and regard the result as an equation for Θ_D, we find that in the high-temperature limit the latter is given by

$$\Theta_D(T) = \Theta_\infty\left[1 - \frac{1}{80}\left(\frac{1}{3}\frac{u_4}{u_2} - \frac{25}{14}u_2\right)\left(\frac{\hbar\omega_{\max}}{T}\right)^2 + O\left(\frac{1}{T^4}\right)\right], \tag{90a}$$

where Θ_∞, the Debye characteristic frequency at infinite temperature, is

$$\Theta_\infty = \left(\tfrac{5}{4}u_2\right)^{1/2}\hbar\omega_{\max}. \tag{90b}$$

From the result given by Eq. (90a) it appears as if the near constancy of the Debye temperature of NaCl at high temperature seen in Fig. 1 is due to the smallness of the coefficient of $(\hbar\omega_{\max}/T)^2$ for this crystal.

The series (84) converges only for $T > \hbar\omega_{\max}/2\pi$. A rearrangement of this series, based on Euler's summation formula, that yields a series convergent for all T has been presented by Sack *et al.* (1961).

An expansion similar to Thirring's for the high-temperature specific heat was obtained for the entropy by Stern (1916). The internal energy at high temperatures also possesses an expansion of this type.

4. LATTICE THERMAL CONDUCTIVITY OF INSULATORS

Up to now in this article we have considered only the contribution of phonons to equilibrium vibrational properties of crystals. However, these vibrational excitations of a crystal play an essential role in a variety of nonequilibrium processes as well. Among these are the electrical and thermal conductivities of metals and semiconductors, in which the departures from the perfect periodicity of a crystal caused by the atomic displacements scatter the conduction electrons and give rise to nonzero electrical and thermal resistivities. In such materials the phonons themselves also contribute to the thermal conductivity, but their contribution usually is smaller than that due to the conduction electrons. The situation is quite different in the case of insulating crystals, which lack conduction electrons. The thermal conductivity of these materials is due solely to the phonons in them, and it is this case that we will use to illustrate the role of phonons in nonequilibrium processes.

In discussing the lattice thermal conductivity of an insulator it is convenient to think of the phonons in it as particles that can propagate through it and be scattered by collisions with each other (through the anharmonic terms in the potential energy of the crystal that we have neglected until now), with crystalline imperfections, and with the boundaries of the crystal. Since a phonon with wave vector **k** and branch index j is an excitation of the normal vibration mode ($\mathbf{k}j$) of the crystal that involves all the atoms in the crystal, it may seem at first glance that a particlelike description of it is inappropriate. It can be justified, however, by the use of a wave packet synthesized from phonons whose wave vectors are within a narrow range $\Delta\mathbf{k}$ about the wave vector **k**. This wave packet then has a width in space of the order of $\Delta x \cong 1/\Delta k$, and can be regarded as a particle (Peierls, 1929).

Because the chemical potential of phonons is zero, Eq. (49), in phonon-mediated thermal conductivity phonons are created at the hot end of the sample and are absorbed at the cold end. The thermal conductivity tensor $\kappa_{\alpha\beta}(T)$ relates the components of the heat current in the crystal, **Q**, to the components of the temperature gradient ∇T present in it:

$$Q_\alpha = -\sum_\beta \kappa_{\alpha\beta}\frac{\partial T}{\partial x_\beta}, \qquad (91)$$

where the minus sign reflects the fact that heat flows from the hotter to the colder side of the crystal. It is an interesting property of this lattice thermal conduction that the heat flux, or equivalently the flux of phonons, is not accompanied by mass transport, as is the case, for example, in electrical conduction.

The thermal conductivity tensor is a symmetric, second-rank tensor $\kappa_{\alpha\beta}=\kappa_{\beta\alpha}$.

The heat current carried by all the phonons can be written as

$$\mathbf{Q} = \frac{1}{V}\sum_{\mathbf{k}j} E_{n_j(\mathbf{k})}\mathbf{v}_j(\mathbf{k}), \qquad (92)$$

where $E_{n_j(\mathbf{k})}$ is the energy of the phonon ($\mathbf{k}j$) and is given by $E_{n_j(\mathbf{k})} = \hbar\omega_j(\mathbf{k})[n_j(\mathbf{k}) + \frac{1}{2}]$ [Eq. (44)], $\mathbf{v}(\mathbf{k}j) = \nabla_{\mathbf{k}}\omega_j(\mathbf{k})$ is the group velocity of this phonon, and V is the volume of the crystal. Because the frequency $\omega_j(\mathbf{k})$ is an even function of the wave vector **k**, the group velocity $\mathbf{v}(\mathbf{k}j)$ is an odd function of **k**, $\mathbf{v}(-\mathbf{k}j) = -\mathbf{v}(\mathbf{k}j)$. Consequently, the zero-point energy does not contribute to the heat current, which takes the form

$$\mathbf{Q} = \frac{1}{V}\sum_{\mathbf{k}j} n_j(\mathbf{k})\hbar\omega_j(\mathbf{k})\mathbf{v}_j(\mathbf{k}). \qquad (93)$$

When we consider the motion of phonons of wave vector **k** and branch index j through a crystal in the presence of a temperature gradient, their number changes. This is because (Callaway, 1991, p. 169)

1. the number of phonons depends on the temperature gradient, so that $n_j(\mathbf{k})$ becomes a function of position, $n_j(\mathbf{k}) \to n_j(\mathbf{k};\mathbf{x})$, and the phonons tend to drift down the gradient; and
2. phonons collide with each other, with imperfections, and with the boundaries of the crystal.

As a result of these effects $n_j(\mathbf{k};\mathbf{x})$ deviates from its equilibrium value in the absence of the temperature gradient, $\bar{n}(\omega_j(\mathbf{k}))$. In a stationary state the total change in the phonon occupation number due to the temperature gradient, $(\partial n_j(\mathbf{k};\mathbf{x})/\partial t)_T$, and due to collisions, $(\partial n_j(\mathbf{k};\mathbf{x})/\partial t)_c$, must vanish:

$$\left(\frac{\partial n_j(\mathbf{k};\mathbf{x})}{\partial t}\right)_T + \left(\frac{\partial n_j(\mathbf{k};\mathbf{x})}{\partial t}\right)_c = 0. \qquad (94)$$

This balance equation is called the *Boltzmann equation*. It expresses the fact that in a stationary state the number of phonons $n_j(\mathbf{k};\mathbf{x})d^3x$ contained in the volume elements d^3x at the point **x** in the crystal must be independent of the time.

The rate of change of $n_j(\mathbf{k};\mathbf{x})$ due to the temperature gradient can be obtained from the argument that as the phonons drift in the presence of the temperature gradient, the same number of phonons are in the volume element d^3x at **x** at the time t as were in the volume element d^3x at $\mathbf{x}-\mathbf{v}_j(\mathbf{k})t$ at the time $t-\Delta t$. The change in $n_j(\mathbf{k};\mathbf{x})$ in the time interval Δt is therefore equal to the difference of the number of phonons in these two volume elements:

$$\left(\frac{\partial n_j(\mathbf{k};\mathbf{x})}{\partial t}\right)_T \Delta t = n_j(\mathbf{k};\mathbf{x}-\mathbf{v}_j(\mathbf{k})\Delta t) - n_j(\mathbf{k};\mathbf{x}) \qquad (95)$$

in the limit as $\Delta t \to 0$. On expanding the right-hand side of this equation to first order in Δt we obtain

$$\left(\frac{\partial n_j(\mathbf{k};\mathbf{x})}{\partial t}\right)_T = -\mathbf{v}_j(\mathbf{k})\cdot\nabla n_j(\mathbf{k};\mathbf{x}). \tag{96}$$

However, $n_j(\mathbf{k};\mathbf{x})$ depends on $\mathbf{x}$ only through its dependence on the temperature, which is a function of $\mathbf{x}$, so that

$$\begin{aligned}\left(\frac{\partial n_j(\mathbf{k};\mathbf{x})}{\partial t}\right)_T &= -\mathbf{v}_j(\mathbf{k})\cdot\frac{\partial n_j(\mathbf{k};\mathbf{x})}{\partial T}\nabla T\\ &\cong -\frac{\partial \bar{n}(\omega_j(\mathbf{k}))}{\partial T}\,\mathbf{v}_j(\mathbf{k})\cdot\nabla T\\ &= -\bar{n}(\omega_j(\mathbf{k}))[\bar{n}(\omega_j(\mathbf{k}))+1]\\ &\quad\times\frac{\hbar\omega_j(\mathbf{k})}{T^2}\mathbf{v}_j(\mathbf{k})\cdot\nabla T.\end{aligned} \tag{97}$$

In obtaining $(\partial n_j(\mathbf{k};\mathbf{x})/\partial t)_c$, which is due to the scattering of phonons from crystalline imperfections and from the boundaries of the crystal, and also to collisions among the phonons themselves, we have to differentiate between processes in which the total wave vector of the participating phonons is conserved and processes in which it is not. The scatterings of a phonon by crystalline imperfections (point defects, dislocations, etc.) and by the boundaries of the crystal are examples of processes in which the wave vector of the phonon scattered is not conserved. The collisions of phonons with each other are more complicated in that processes in which the total wave vectors of the initial and final phonons are the same can occur, together with processes in which they differ. This can be seen by considering the lowest-order term in the expansion of the potential energy of the crystal in powers of the atomic displacements that was neglected in obtaining the harmonic approximation, Eq. (4), viz.,

$$\Phi_3=\frac{1}{6}\sum_{l\kappa\alpha}\sum_{l'\kappa'\beta}\sum_{l''\kappa''\gamma}\Phi_{\alpha\beta\gamma}(l\kappa;l'\kappa';l''\kappa'')u_\alpha(l\kappa)u_\beta(l'\kappa')u_\gamma(l''\kappa''), \tag{98}$$

where $\Phi_{\alpha\beta\gamma}(l\kappa;l'\kappa';l''\kappa'')$ is the third-order atomic force constant,

$$\Phi_{\alpha\beta\gamma}(l\kappa;l'\kappa';l''\kappa'')=\left(\frac{\partial^3\Phi}{\partial u_\alpha(l\kappa)\partial u_\beta(l'\kappa')\partial u_\gamma(l''\kappa'')}\right)_0. \tag{99}$$

If we expand $u_\alpha(l\kappa)$ according to Eq. (33a),

$$u_\alpha(l\kappa)=\frac{1}{\sqrt{NM_\kappa}}\sum_{\mathbf{k}j}u_\alpha(\kappa|\mathbf{k}j)Q(\mathbf{k}j)e^{i\mathbf{k}\cdot\mathbf{x}(l)}, \tag{100}$$

where, to simplify the notation, we have defined

$$Q(\mathbf{k}j)=\tfrac{1}{2}[q(\mathbf{k}j)+q(-\mathbf{k}j)]+\frac{1}{2}\frac{i}{\omega_j(\mathbf{k})}[p(\mathbf{k}j)-p(-\mathbf{k}j)], \tag{101}$$

Eq. (98) takes the form

$$\Phi_3=\sum_{\mathbf{k}j}\sum_{\mathbf{k}'j'}\sum_{\mathbf{k}''j''}V^{(3)}(\mathbf{k}j;\mathbf{k}'j';\mathbf{k}''j'')Q(\mathbf{k}j)Q(\mathbf{k}'j')Q(\mathbf{k}''j''), \tag{102}$$

with

$$\begin{aligned}V^{(3)}(\mathbf{k}j;\mathbf{k}'j';\mathbf{k}''j'')&=\frac{1}{6N^{3/2}}\sum_{l\kappa\alpha}\sum_{l'\kappa'\beta}\sum_{l''\kappa''\gamma}\Phi_{\alpha\beta\gamma}(l\kappa;l'\kappa';l''\kappa'')\\ &\quad\times\frac{u_\alpha(\kappa|\mathbf{k}j)}{\sqrt{M_\kappa}}\frac{u_\beta(\kappa'|\mathbf{k}'j')}{\sqrt{M_{\kappa'}}}\frac{u_\gamma(\kappa''|\mathbf{k}''j'')}{\sqrt{M_{\kappa''}}}e^{i[\mathbf{k}\cdot\mathbf{x}(l)+\mathbf{k}'\cdot\mathbf{x}(l')+\mathbf{k}''\cdot\mathbf{x}(l'')]}.\end{aligned} \tag{103}$$

The coefficient $V^{(3)}(\mathbf{k}j;\mathbf{k}'j'\mathbf{k}''j'')$ is important in phonon–phonon interactions involving three phonons because time-dependent perturbation theory tells us, for example, that the rate at which phonons $(\mathbf{k}'j')$ and $(\mathbf{k}''j'')$ collide to create a phonon $(\mathbf{k}j)$ is proportional to $|V^{(3)}(-\mathbf{k}j;\mathbf{k}'j';\mathbf{k}''j'')|^2$ (Peierls, 1929; Leibfried, 1955; Gurevich, 1986, pp. 54ff.). Now, the atomic force constant $\Phi_{\alpha\beta\gamma}(l\kappa;l'\kappa';l''\kappa'')$ appearing in the expression (103) for $V^{(3)}(\mathbf{k}j;\mathbf{k}'j';\mathbf{k}''j'')$ is invariant against the addition of the same arbitrary cell index m to each of the indices l,l',l'':

$$\Phi_{\alpha\beta\gamma}(l+m,\kappa;l'+m,\kappa';l''+m,\kappa'') \\ =\Phi_{\alpha\beta\gamma}(l\kappa;l'\kappa';l''\kappa''), \qquad (104)$$

because of the invariance of the crystal against the displacement through the translation vector $\mathbf{x}(m)$. As a consequence, $V^{(3)}(\mathbf{k}j;\mathbf{k}'j';\mathbf{k}''\mathrm{j}'')$ and $|V^{(3)}(\mathbf{k}j;\mathbf{k}'j';\mathbf{k}''\mathrm{j}'')|^2$ are proportional to $\Delta(\mathbf{k}+\mathbf{k}'+\mathbf{k}'')$, where $\Delta(\mathbf{k})$ has been defined in Eq. (36). Three-phonon scattering processes proportional to $|V^{(3)}(\mathbf{k}j;\mathbf{k}'j';\mathbf{k}''j'')|^2$ in which $\mathbf{k}+\mathbf{k}'+\mathbf{k}''=0$ have been termed *normal processes;* those in which $\mathbf{k}+\mathbf{k}'+\mathbf{k}''=\mathbf{G}(h)\neq 0$ have been termed *Umklapp (flipping-over) processes.* An Umklapp process is a scattering process in which momentum is transferred to the crystal as a whole.

The central role in lattice thermal conduction in insulators is played by the scattering processes in which the wave vector of the participating phonons is not conserved. This is particularly readily seen at very low temperatures when only very low-energy acoustic phonons are excited. Thus, in this case, if we rewrite Eq. (93) as

$$\mathbf{Q}=\frac{1}{2V}\sum_{\mathbf{k}}\sum_{j=1}^{3} n_j(\mathbf{k})\hbar\nabla_{\mathbf{k}}\omega_j^2(\mathbf{k}), \qquad (105)$$

and make the simplifying assumption that the frequencies of acoustic phonons are independent of their direction of propagation,

$$\omega_j(\mathbf{k})=c_jk, \quad j=1,2,3, \qquad (106)$$

where c_j is the speed of sound for the jth branch of the phonon spectrum, we obtain for the heat current

$$\mathbf{Q}=\frac{1}{V}\sum_{j=1}^{3} c_j^2 \sum_{\mathbf{k}} n_j(\mathbf{k})\hbar\mathbf{k}=\sum_{j=1}^{3} c_j^2\mathbf{P}_j, \qquad (107)$$

where

$$\mathbf{P}_j=\frac{1}{V}\sum_{\mathbf{k}} n_j(\mathbf{k})\hbar\mathbf{k} \qquad (108)$$

is the total wave vector of the phonons in the branch j. If we assume further that the speeds of sound are the same for all three acoustic branches, viz., that $c_j=c$ for $j=1,2,3$, then the expression for the heat current $\mathbf{Q}$ reduces to

$$\mathbf{Q}=c^2\mathbf{P}, \qquad (109)$$

where

$$\mathbf{P}=\sum_j \mathbf{P}_j \qquad (110)$$

is the sum of the wave vectors of all the phonons in the crystal.

If only wave-vector–conserving scattering processes existed, the total wave vector of all the phonons in the crystal, $\mathbf{P}$, could not change. Thus, in this case collisions could not lead from a state with nonzero $\mathbf{P}$ to a state with zero $\mathbf{P}$. In equilibrium we must have $\mathbf{P}=\mathbf{0}$. In a state with $\mathbf{P}\neq 0$ a heat current exists, according to Eq. (109). Therefore, in the case of wave-vector–conserving processes only, if a heat current were established it would continue to flow even in the absence of a thermal gradient, which implies zero thermal resistance or an infinite thermal conductivity. However, wave-vector–destroying processes change the total wave vector of the phonon system, and thus are essential for the existence of a finite thermal conductivity of a perfect, unbounded crystal.

Although this conclusion has been reached on the basis of several simplifying assumptions, it has a more general validity. Thus in obtaining the contribution to $(\partial n_j(\mathbf{k};\mathbf{x})/\partial t)_c$ due to wave-vector–destroying processes we exploit the fact that they cause any deviation of $n_j(\mathbf{k};\mathbf{x})$ from its equilibrium value $\bar{n}(\omega_j(\mathbf{k}))$ to disappear if the temperature gradient is turned off. A commonly made approximation that describes this is to assume that because of the scattering of phonons by processes that change the total wave vector of the crystal, $n_j(\mathbf{k};\mathbf{x})$ relaxes to its equilibrium value $\bar{n}(\omega_j(\mathbf{k}))$ according to an exponential law,

$$\left(\frac{\partial n_j(\mathbf{k};\mathbf{x})}{\partial t}\right)_U=-\frac{n_j(\mathbf{k};\mathbf{x})-\bar{n}(\omega_j(\mathbf{k}))}{\tau_j^{(u)}(\mathbf{k})}, \qquad (111)$$

where $\tau_j^{(u)}(\mathbf{k})$ is the relaxation time for the mode $(\mathbf{k}j)$. It includes the contributions from all scattering processes that do not conserve the phonon wave vectors, e.g., scattering from point defects, and from sample boundaries and Umklapp processes. The result represented by Eq. (111) is called the *relaxation-time approximation.* Although Eq. (111) describes in a qualitative way how the deviation $n_j(\mathbf{k};\mathbf{x}) - \bar{n}(\omega_j(\mathbf{k}))$ decays to zero as a result of wave-vector–destroying phonon-scattering processes, in general it cannot be justified from first principles. An exception to this statement, however, is provided by elastic scattering processes in which the number of phonons is conserved, such as the scattering of phonons by point defects and by the boundaries of the crystal. For such processes Eq. (111) is exact, and explicit expressions for the inverse relaxation time $\tau_j^{(u)}(\mathbf{k})$ can be obtained (Maradudin, 1966, pp. 335–338; Grimm *et al.*, 1970).

The normal processes, which conserve the total wave vector of the phonons in the crystal, cannot relax the phonon distribution $n_j(\mathbf{k};\mathbf{x})$ to the equilibrium distribution. Instead they relax it to a displaced equilibrium distribution $\bar{n}(\omega_j(\mathbf{k}) - \mathbf{k}\cdot\mathbf{V})$ (Klemens, 1955):

$$\left(\frac{\partial n_j(\mathbf{k};\mathbf{x})}{\partial t}\right)_N = -\frac{n_j(\mathbf{k};\mathbf{x}) - \bar{n}(\omega_j(\mathbf{k}) - \mathbf{k}\cdot\mathbf{V})}{\tau_j^{(n)}(\mathbf{k})}, \tag{112}$$

where

$$\bar{n}(\omega_j(\mathbf{k}) - \mathbf{k}\cdot\mathbf{V}) = \{\exp(\hbar[\omega_j(\mathbf{k}) - \mathbf{k}\cdot\mathbf{V}]/T) - 1\}^{-1}, \tag{113}$$

and $\tau_j^{(n)}(\mathbf{k})$ is the corresponding relaxation time. The vector $\mathbf{V}$ has the physical interpretation of the drift velocity of the phonon gas in the temperature gradient (Gurevich, 1986, p. 120), and can be regarded as small (we will see that it is proportional to ∇T). Thus we can write to first order in $\mathbf{V}$ that

$$\bar{n}(\omega_j(\mathbf{k}) - \mathbf{k}\cdot\mathbf{V}) = \bar{n}(\omega_j(\mathbf{k})) + \frac{\hbar}{T}(\mathbf{V}\cdot\mathbf{k}) \times \bar{n}(\omega_j(\mathbf{k}))[\bar{n}(\omega_j(\mathbf{k})) + 1]. \tag{114}$$

The total time rate of change of $n_j(\mathbf{k};\mathbf{x})$ due to scattering and collision processes is the sum of Eqs. (111) and (112) which, with the use of Eq. (114), is

$$\left(\frac{\partial n_j(\mathbf{k};\mathbf{x})}{\partial t}\right)_c = -\left(\frac{1}{\tau_j^{(n)}(\mathbf{k})} + \frac{1}{\tau_j^{(u)}(\mathbf{k})}\right)[n_j(\mathbf{k};\mathbf{x}) - \bar{n}(\omega_j(\mathbf{k}))] + \frac{(\mathbf{V}\cdot\mathbf{k})}{\tau_j^{(n)}(\mathbf{k})}\frac{\hbar}{T}\bar{n}(\omega_j(\mathbf{k}))[\bar{n}(\omega_j(\mathbf{k})) + 1]. \tag{115}$$

With this result and that given by Eq. (97), the Boltzmann equation (94) takes the form

$$-\bar{n}(\omega_j(\mathbf{k}))[\bar{n}(\omega_j(\mathbf{k})) + 1]\frac{\hbar\omega_j(\mathbf{k})}{T^2}\mathbf{v}_j(\mathbf{k})\cdot\nabla T - \left(\frac{1}{\tau_j^{(u)}(\mathbf{k})} + \frac{1}{\tau_j^{(n)}(\mathbf{k})}\right)[n_j(\mathbf{k};\mathbf{x}) - \bar{n}(\omega_j(\mathbf{k}))] + \frac{(\mathbf{V}\cdot\mathbf{k})}{\tau_j^{(n)}(\mathbf{k})}\frac{\hbar}{T}\bar{n}(\omega_j(\mathbf{k}))[\bar{n}(\omega_j(\mathbf{k})) + 1] = 0, \tag{116}$$

whose solution is

$$n_j(\mathbf{k};\mathbf{x}) = \bar{n}(\omega_j(\mathbf{k})) + \frac{\tau_j^{(n)}(\mathbf{k})\tau_j^{(u)}(\mathbf{k})}{\tau_j^{(n)}(\mathbf{k}) + \tau_j^{(u)}(\mathbf{k})}\bar{n}(\omega_j(\mathbf{k}))[\bar{n}(\omega_j(\mathbf{k})) + 1]\frac{\hbar}{T}\left\{\frac{\mathbf{V}\cdot\mathbf{k}}{\tau_j^{(n)}(\mathbf{k})} - \frac{\omega_j(\mathbf{k})}{T}\mathbf{v}_j(\mathbf{k})\cdot\nabla T\right\}. \tag{117}$$

The phonon drift velocity $\mathbf{V}$ is then obtained from the condition that the rate of change of the total phonon wave vector due to normal processes vanishes. This can be expressed by the statement

$$\sum_{\mathbf{k}j}\left(\frac{\partial n_j(\mathbf{k};\mathbf{x})}{\partial t}\right)_N \mathbf{k} = \sum_{\mathbf{k}j}\left(\frac{\bar{n}(\omega_j(\mathbf{k}) - \mathbf{k}\cdot\mathbf{V}) - n_j(\mathbf{k};\mathbf{x})}{\tau_j^{(n)}(\mathbf{k})}\right)\mathbf{k} = \mathbf{0}, \tag{118}$$

or, with the use of Eqs. (114) and (117),

$$\sum_{\mathbf{k}j} \frac{\bar{n}(\omega_j(\mathbf{k}))[\bar{n}(\omega_j(\mathbf{k}))+1]}{\tau_j^{(n)}(\mathbf{k})+\tau_j^{(u)}(\mathbf{k})} \frac{\hbar}{T}(\mathbf{V}\cdot\mathbf{k})\mathbf{k} = -\sum_{\mathbf{k}j} \frac{\tau_j^{(u)}(\mathbf{k})\bar{n}(\omega_j(\mathbf{k}))[\bar{n}(\omega_j(\mathbf{k}))+1]}{\tau_j^{(n)}(\mathbf{k})+\tau_j^{(u)}(\mathbf{k})} \frac{\hbar\omega_j(\mathbf{k})}{T^2}\mathbf{v}_j(\mathbf{k})\cdot\nabla T\mathbf{k}. \tag{119}$$

Thus, if we write the solution of this equation formally as

$$V_\alpha = -\frac{1}{T}\sum_\beta \beta_{\alpha\beta}\frac{\partial T}{\partial x_\beta}, \tag{120}$$

then from Eqs. (91), (93), and (117) we obtain for the thermal conductivity tensor

$$\kappa_{\alpha\beta} = \frac{1}{V}\sum_{\mathbf{k}j} \tau_j^{(c)}(\mathbf{k})C_j(\mathbf{k})v_{j\alpha}(\mathbf{k})\left[v_{j\beta}(\mathbf{k}) + \frac{1}{\omega_j(\mathbf{k})\tau_j^{(n)}(\mathbf{k})}\sum_\gamma \kappa_\gamma\beta_{\gamma\beta}\right], \tag{121}$$

where $C_j(\mathbf{k}) = [\hbar\omega_j(\mathbf{k})/T]^2\bar{n}(\omega_j(\mathbf{k}))[\bar{n}(\omega_j(\mathbf{k}))+1]$ is the contribution to the specific heat of the crystal from the mode ($\mathbf{k}j$), and $\tau_j^{(c)}(\mathbf{k})$ is the combined relaxation time defined by

$$1/\tau_j^{(c)}(\mathbf{k}) = [1/\tau_j^{(n)}(\mathbf{k})] + [1/\tau_j^{(u)}(\mathbf{k})]. \tag{122}$$

If it were not for the second term in brackets on the right-hand side of Eq. (121), this expression for $\kappa_{\alpha\beta}$ would have the form predicted by elementary kinetic theory (Kittel, 1976, pp. 143–144). The presence of the second term in brackets is a consequence of the fact that normal processes cannot relax a given phonon distribution $n_j(\mathbf{k};\mathbf{x})$ to the equilibrium distribution $\bar{n}(\omega_j(\mathbf{k}))$, but only to the displaced distribution $\bar{n}(\omega_j(\mathbf{k})-\mathbf{k}\cdot\mathbf{V})$.

It is clear that a central role in this approach to the thermal conductivity of an insulator is played by the relaxation times $\tau_j^{(n)}(\mathbf{k})$ and $\tau_j^{(u)}(\mathbf{k})$. It was shown by Herring (1954) that the relaxation time of a single low-energy mode of wave vector **k** due to three-phonon processes depends on **k** and the temperature T according to

$$1/\tau_j(\mathbf{k}) = k^a T^{5-a}, \tag{123}$$

where a is an exponent determined by crystal symmetry. The relaxation time for normal scattering processes $\tau_j^{(n)}(\mathbf{k})$ is usually written in the form (Callaway, 1959)

$$1/\tau_j^{(n)}(\mathbf{k}) = BT^3\omega_j^2(\mathbf{k}), \tag{124}$$

although an alternative form (Gurevich, 1986, p. 98),

$$1/\tau_j(\mathbf{k}) = B[T/v_j(\mathbf{k})]^5, \tag{125}$$

is also used in the limit of very low temperatures, $T \ll \Theta_{\mathrm{D}}$. Both of these forms are consistent with Herring's result, Eq. (123).

For the scattering of phonons by point defects the relaxation time $\tau_j^{(d)}(\mathbf{k})$ has the temperature-independent form (Pomeranchuk, 1942; Klemens, 1951, 1955, 1957)

$$1/\tau_j^{(d)}(\mathbf{k}) = A\omega_j^4(\mathbf{k}), \tag{126}$$

where the constant coefficient A is proportional to the concentration of point defects. Equation (126) is the Rayleigh law for the scattering of long-wavelength phonons by defects whose linear dimensions are smaller than the phonon wavelength. A derivation of Eq. (126) on the basis of defect lattice dynamics for the particular case of the scattering of phonons by isotopic impurities has been given, e.g., by Maradudin (1966, pp. 335–338).

For the scattering of phonons by the boundaries of the sample, the relaxation time $\tau_j^{(b)}(\mathbf{k})$ is frequency independent and temperature independent (Casimir, 1938),

$$1/\tau_j^{(b)}(\mathbf{k}) = v_j(\mathbf{k})/L, \tag{127}$$

where L is a length that is of the order of the linear dimensions of the sample. In a microscopic derivation of Eq. (127) by Grimm *et al.* (1970), $v_j(\mathbf{k})$ is replaced by the component of the group velocity $\mathbf{v}_j(\mathbf{k})$ normal to the boundary.

Finally, the relaxation time for the scattering of phonons by three-phonon Umklapp processes has the form

$$1/\tau_j^{(\mathrm{U})}(\mathbf{k}) = C\exp(-\alpha\Theta_{\mathrm{D}}/T) \tag{128}$$

at low temperatures (Gurevich, 1986, p. 123; Brüesch, 1987, p. 96), where α is a numerical factor of the order of, and smaller than, unity. The origin of this form can be understood by considering a scattering process in which two phonons with wave vectors $\mathbf{k}_1$ and $\mathbf{k}_2$ interact to create a phonon whose wave vector $\mathbf{k}_1+\mathbf{k}_2$ extends outside the first Brillouin zone. At the low temperatures we are considering, only acoustic phonons are excited, and for simplicity we can ignore the distinction between longitudinal and transverse phonons, and can suppress the branch indices of the interacting phonons. This process is illustrated for a two-dimensional square lattice of lattice constant a in Fig. 3. The addition to $\mathbf{k}_1+\mathbf{k}_2$ of a reciprocal-lattice vector $\mathbf{G}=(-2\pi/a,0,0)$, whose magnitude is equal to the length of an edge of the Brillouin zone, transforms this vector into a vector $\mathbf{k}_3$ inside the first Brillouin zone. Here k_{1x} and k_{2x} are both positive, but k_{3x} is negative. This is the origin of the name "Umklapp" for scattering processes of this type. For an Umklapp process to be possible, $|\mathbf{k}_1+\mathbf{k}_2|$ must exceed $\frac{1}{2}G_{\min}$, where $G_{\min}$ is the magnitude of the shortest reciprocal lattice vector, which in the present case is $2\pi/a$. This requires in turn that the magnitude of at least one of the two wave vectors $\mathbf{k}_1$ and $\mathbf{k}_2$ must exceed $\frac{1}{4}G_{\min}$. In the Debye approximation the Brillouin zone is replaced by a circle of radius k_{D} (in the two-dimensional case we are considering), whose area is equal to that of the Brillouin zone. Therefore k_{D} has a value $k_{\mathrm{D}}=2\sqrt{\pi}/a \gtrsim \frac{1}{2}G_{\min}$. From Fig. 3 we see that the frequency of the phonon created in this scattering process is therefore given by $\hbar\omega(\mathbf{k}_3) \lesssim \Theta_{\mathrm{D}}$, in the Debye approximation, where Θ_{D} is the Debye temperature, since $|\mathbf{k}_3| \lesssim k_{\mathrm{D}}$. The collision frequency of an Umklapp process is proportional to the product of the occupation numbers of the colliding phonons, $\bar{n}(\omega(\mathbf{k}_1))$ and $\bar{n}(\omega(\mathbf{k}_2))$, which, in the low-temperature limit, $T \ll \Theta_{\mathrm{D}}$, are given by Eq. (50):

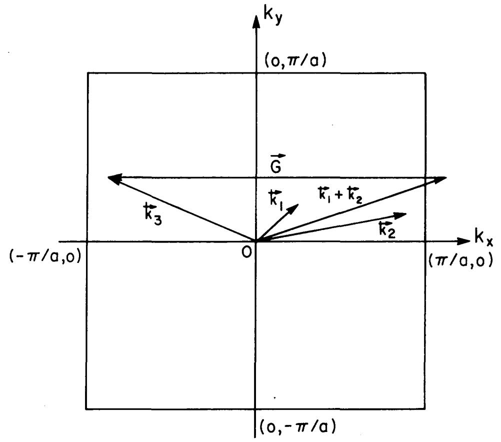

FIG. 3. The first Brillouin zone for a square lattice of lattice constant a, and the wave vectors of phonons participating in an Umklapp process.

$$\bar{n}(\omega(\mathbf{k}_1)) \cong \exp[-\hbar\omega(\mathbf{k}_1)/T],$$
$$\bar{n}(\omega(\mathbf{k}_2)) \cong \exp[-\hbar\omega(\mathbf{k}_2)/T], \qquad (129)$$

respectively. The product $\bar{n}(\omega(\mathbf{k}_1))\bar{n}(\omega(\mathbf{k}_2))$ is therefore given by

$$\bar{n}(\omega(\mathbf{k}_1))\bar{n}(\omega(\mathbf{k}_2)) \cong \exp\{-\hbar[\omega(\mathbf{k}_1)+\omega(\mathbf{k}_2)]/T\}. \qquad (130)$$

However, energy is conserved in the interaction, so that $\omega(\mathbf{k}_1)+\omega(\mathbf{k}_2)=\omega(\mathbf{k}_1+\mathbf{k}_2)=\omega(\mathbf{k}_1+\mathbf{k}_2+\mathbf{G})=\omega(\mathbf{k}_3)$, where we have used the result that the frequency $\omega(\mathbf{k})$ is a periodic function of $\mathbf{k}$ with period $\mathbf{G}$. It follows that the collision rate is proportional to

$$\bar{n}(\omega(\mathbf{k}_1))\bar{n}(\omega(\mathbf{k}_2)) \cong \exp[-\hbar\omega(\mathbf{k}_3)/T] = \exp(-\alpha\Theta_{\mathrm{D}}/T), \qquad (131)$$

where α is a number of the order of, and smaller than, unity. This result establishes Eq. (128).

In the high-temperature limit, $T \gg \Theta_{\mathrm{D}}$, an approximate microscopic calculation (Ludwig, 1964; Gurevich, 1986, p. 96) yields the result that the relaxation time for the scattering of phonons by three-phonon Umklapp processes has the form

$$1/\tau_j^{(\mathrm{U})}(\mathbf{k}) = D_j(\mathbf{k})T. \qquad (132)$$

This dependence on temperature can be understood (Kittel, 1976, p. 145) in terms of the number of phonons with which a given phonon can interact. At high temperatures, when all phonons are excited, the magnitudes of the relaxation times for normal and Umklapp processes become comparable, and we need not distinguish between these two types of scattering processes. In this limit the total number of excited phonons is proportional to the absolute temperature T [Eq. (50b)]. The collision frequency of a given phonon should be proportional to the number of phonons with which it can interact, and Eq. (132) follows from this argument.

The various scattering processes just described are assumed to be independent, so that the probabilities of their occurrence add, and consequently so do the inverse relaxation times. Thus we have that

$$\frac{1}{\tau_j^{(u)}(\mathbf{k})}=\frac{1}{\tau_j^{(d)}(\mathbf{k})}+\frac{1}{\tau_j^{(b)}(\mathbf{k})}+\frac{1}{\tau_j^{(\mathrm{U})}(\mathbf{k})}. \qquad (133)$$

The calculation of the thermal conductivity of an insulator by the procedure just described is usually carried out on the basis of the assumption of a Debye spectrum for the phonons with speeds of sound that are independent of the direction of propagation, and even that the speed of sound is the same for all three acoustic branches (Callaway, 1959). With such approximations it follows from Eq. (121) that at very low temperatures (usually below approximately 10 K), where the inverse lifetimes for the scattering of phonons by other phonons and by crystalline defects and impurities have essentially vanished, and the only scattering mechanism limiting the thermal conductivity is the scattering of phonons by the boundaries of the crystal, the thermal conductivity becomes proportional to the specific heat of the crystal, and thus increases with increasing temperature according to a T^3 law. For, in this case, if we denote by c the common speed of sound, from Eqs. (121), (127), and (133) we obtain

$$\begin{aligned}\kappa_{\alpha\beta}&=\frac{1}{V}\sum_{\mathbf{k}j}\tau_j^{(b)}(\mathbf{k})C_j(\mathbf{k})v_{j\alpha}(\mathbf{k})v_{j\beta}(\mathbf{k})\\ &=\delta_{\alpha\beta}\frac{cL}{3}\frac{1}{V}\sum_{\mathbf{k}j}C_j(\mathbf{k})\\ &=\delta_{\alpha\beta}\frac{cL}{3}\frac{C_v(T)}{V}. \end{aligned}\qquad (134)$$

At high temperatures, $T \gg \Theta_{\mathrm{D}}$, the dominant scattering processes leading to a finite lattice thermal conductivity are the three-phonon Umklapp scattering processes. In this limit, we obtain from Eqs. (121), (132), (133), and the fact that $C_j(\mathbf{k})\to 1$ as $T\to\infty$, the result that

$$\begin{aligned}\kappa_{\alpha\beta}&=\frac{1}{V}\sum_{\mathbf{k}j}\tau_j^{(U)}(\mathbf{k})C_j(\mathbf{k})v_{j\alpha}(\mathbf{k})v_{j\beta}(k)\\ &\xrightarrow[T\to\infty]{}\frac{1}{T}\frac{1}{V}\sum_{\mathbf{k}j}\frac{v_{j\alpha}(\mathbf{k})v_{j\beta}(\mathbf{k})}{D_j(\mathbf{k})}\\ &=\delta_{\alpha\beta}\frac{1}{3}\frac{c^2}{T}\frac{1}{V}\sum_{\mathbf{k}j}\frac{1}{D_j(\mathbf{k})}. \end{aligned}\qquad (135)$$

Thus, in this limit the thermal conductivity of all insulating crystals must decrease according to a T^{-1} law (Brüesch, 1987, p. 97; Callaway, 1991, p. 175). A maximum therefore appears at intermediate temperatures in the temperature dependence of the thermal conductivity. The thermal conductivity in the temperature range where this maximum occurs is determined primarily by the scattering of the phonons by crystalline defects and impurities. In quantitative calculations of the conductivity as a function of temperature, the coefficients in the expressions for the several relaxation times entering the theory are usually obtained by fitting the resulting theoretical curves to experimental data in the limiting cases of low and high temperatures. As an example of the kinds of results that can be obtained in this way, we present in Fig. 4 plots of the lattice thermal conductivity of almost isotopically pure ^{74}Ge and of normal Ge (Callaway, 1959), calculated with the inclusion of all the scattering processes described above. The thermal conductivity of the normal Ge is seen to be smaller than that of the isotopically pure Ge because of the additional thermal resistivity of the former sample due to the scattering from the isotopic impurities in it. The agreement between theory and experiment is quite good, suggesting that the essential physics has been incorporated into the theory. The discrepancies, particularly in the vicinity of the maximum, are presumably due to the use of a relaxation-time approximation

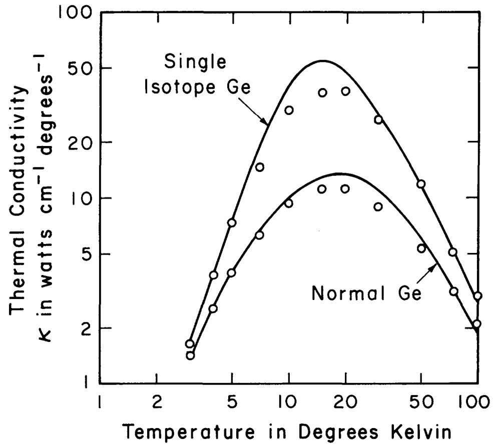

FIG. 4. The thermal conductivity of germanium. The solid curves are theoretical results calculated in the manner described in the text. The open circles are experimental data of Geballe and Hull (1958) (Callaway, 1959).

to describe three-phonon collision processes and to the neglect of the distinction between longitudinal and transverse phonons.

List of Works Cited

Adams, E. P. (1957), *Smithsonian Mathematical Formulae and Tables of Elliptic Functions,* Washington, DC: The Smithsonian Institution. Note that Adams denotes by B_n what we have called B_{2n}.

Blackman, M. (1934a), *Proc. R. Soc. London Ser. A* **148**, 365–383.

Blackman, M. (1934b), *Proc. R. Soc. London Ser. A* **148**, 384–406.

Blackman, M. (1936), *Philos. Trans. R. Soc. London Ser. A* **236**, 103–131.

Blackman, M. (1937a), *Proc. Cambridge Philos. Soc.* **33**, 94–103.

Blackman, M. (1937b), *Proc. R. Soc. London Ser. A* **159**, 416–431.

Born, M., Huang, K. (1954), *Dynamical Theory of Crystal Lattices,* Oxford: Clarendon.

Born, M., von Kármán, Th. (1912), *Phys. Z.* **13**, 297–309.

Born, M., von Kármán,Th. (1913), *Phys. Z.* **14**, 15–19.

Brüesch, P. (1987), *Phonons: Theory and Experiments III,* Berlin: Springer-Verlag.

Callaway, J. (1959), *Phys. Rev.* **113**, 1046–1051.

Callaway, J. (1991), *Quantum Theory of the Solid State,* 2nd ed., New York: Academic.

Casimir, H. B. G. (1938), *Physica* **5**, 495–500.

Debye, P. (1912), *Ann. Phys. (Leipzig)* **39**, 789–839.

Einstein, A. (1907), *Ann. Phys. (Leipzig)* **22**, 180–190.

Einstein, A. (1911), *Ann. Phys. (Leipzig)* **35**, 679–694.

Geballe, T. H., Hull, G. W. (1958), *Phys. Rev.* **110**, 773–775.

Grimm, A., Maradudin, A. A., Tong, S. Y. (1970), *J. Phys. (Paris), Colloque* **31**, C1-9–C1-19.

Gurevich, V. L. (1986), *Transport in Phonon Systems,* Amsterdam: North-Holland.

Hardy, J. R. (1962), *Philos Mag.* **7**, 315–336.

Herring, C. (1954), *Phys. Rev.* **95**, 954–965.

Kellermann, E. W. (1941), *Proc. R. Soc. London Ser. A* **178**, 17–24.

Kittel, C. (1976), *Introduction to Solid State Physics,* 5th ed., New York: Wiley.

Klemens, P. G. (1951), *Proc. R. Soc. London Ser. A* **208**, 108–133.

Klemens, P. G. (1955), *Proc. Phys. Soc. Ser. A* **68**, 1113–1128.

Klemens, P. G. (1957), *Proc. Phys. Soc. Ser. A* **70**, 833–838.

Ledermann, W. (1944), *Proc. R. Soc. London Ser. A* **182**, 362–377.

Leibfried, G. (1955), "Gittertheorie der mechanischen und thermischen Eigenschaften der Kristalle," in: S. Flügge (Ed.), *Encyclopedia of Physics,* Vol. VII, Part 1, Berlin: Springer, pp. 104–324.

Ludwig, W. (1964), "Some Aspects of Phonon-Phonon-Interaction," in: T. A. Bak (Ed.), *Phonons and Phonon Interactions,* New York: Benjamin, pp. 23–75.

Maradudin, A. A. (1966), "Theoretical and Experimental Aspects of the Effects of Point Defects and Disorder on the Vibrations of Crystals—1," in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics,* Vol. 18, New York: Academic, pp. 273–420.

Maradudin, A. A., Montroll, E. W., Weiss, G. H., Ipatova, I. P. (1971), *Theory of Lattice Dynamics in the Harmonic Approximation,* New York: Academic.

Peierls, R. (1929), *Ann. Phys. (Leipzig)* **3**, 1055–1101.

Pomeranchuk, I. Ya. (1942), *J. Phys. USSR* **6**, 237–250.

Richarz, F. (1893), *Ann. Phys. (Leipzig)* **48**, 708–716.

Richarz, F. (1899), *Ann. Phys. (Leipzig)* **67**, 702–706.

Sack, R. A., Maradudin, A. A., Weiss, G. H. (1961), *Phys. Rev.* **124**, 717–722.

Schiff, L. I. (1949), *Quantum Mechanics,* New York: McGraw-Hill.

Stern, O. (1916), *Ann. Phys. (Leipzig)* **51**, 237–260.

Thirring, H. (1913), *Phys. Z.* **14**, 867–873.

Thirring, H. (1914a), *Phys. Z.* **15**, 127–133.

Thirring, H. (1914b), *Phys. Z.* **15**, 180–185.

Further Reading

Beck, H. (1975), "Second Sound and Related Thermal Conduction Phenomena," in: G. K. Horton, A. A. Maradudin (Eds.), *Dynamical Properties of Solids,* Amsterdam: North Holland, pp. 205–284.

Bilz, H., Kress, W. (1979), *Phonon Dispersion Relations in Insulators,* Berlin: Springer-Verlag.

Böttger, H. (1983), *Principles of the Theory of Lattice Dynamics,* Berlin: Akademie-Verlag.

Hardy, J. R., Karo, A. M. (1979), *The Lattice Dynamics and Statics of Alkali-Halide Crystals,* New York: Plenum.

Reissland, J. A. (1973), *The Physics of Phonons,* New York: Wiley.

LIQUID CRYSTALS, STRUCTURE OF

J. D. LITSTER, *Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*
R. SHASHIDHAR, *Center for Biomolecular Science and Engineering, Naval Research Laboratory, Washington, DC, U.S.A.*

Introduction 515
1. **Nematic Liquid Crystals** 516
1.1 Optical Properties of Nematics 516
1.2 Viscoelastic Properties of Nematics 517
1.3 Device Applications of Nematics 518
1.3.1 Twisted Nematic Displays 518
2. **Cholesteric Liquid Crystals** 519
2.1 Optical Properties of Cholesterics 519
2.2 Response of Cholesterics to Applied Fields 519
2.3 Applications of Cholesterics 520
2.4 Cholesteric "Blue" Phases 520
3. **Smectic Liquid-Crystal Phases** 521
3.1 Smectic-*A* Phases 521
3.1.1 Analogy to Superconductivity 521
3.1.2 Polymorphic Smectic-*A* and Smectic-*C* Phases 522
3.2 Smectic-*C* Phases 522
3.2.1 The Absence of True Long-Range Order 522
3.3 Ferroelectric Smectic-*C* Phases 523
3.4 Bond-Orientationally Ordered Phases 524
3.5 Higher-Order Smectic Phases 525
3.6 Applications of Smectic Phases 525
3.6.1 Smectic-*A* Devices 525
3.6.2 Ferroelectric Smectic-*C* Applications 526
4. **Discotic Liquid Crystals** 526
5. **Polymer and Polymer-Dispersed Liquid Crystals** 527
5.1 Polymer Nematic and Smectic Liquid Crystals 527
5.1.1 Mesogenic Main-Chain and Side-Chain Polymers 527
5.1.2 Applications of Polymer Liquid Crystals 528
5.2 Polymer-Dispersed Liquid Crystals 528
6. **Lyotropic Phases** 528
6.1 Phases with Ordered Micelles 529
6.2 Lyotropic Lamellar Phases 529
6.3 Applications of Lyotropic Liquid Crystals 530
7. **Conclusions** 530
Acknowledgments 531
Glossary 530
Works Cited 531
Further Reading 532

INTRODUCTION

True liquid crystals are fluids, in the sense that they cannot support shear stress. They possess order intermediate between the high degree of order of a perfect crystal and the disorder of a common isotropic liquid. For this reason they are often called *mesophases*, and molecules that tend to form liquid crystals are said to be mesogenic. The first reported observation of the liquid crystalline state was by the Austrian botanist Friederich Reinitzer (1888). In the intervening century these materials have been investigated not only for their potential applications, but also for their intellectual interest as a window into the variety of phases that can exist in condensed matter. The latter was recognized by the award of the 1991 Nobel Prize in physics to Pierre-Gilles de Gennes of the Collège de France.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

In this article we shall describe the nature of the order found in the various known liquid crystal phases, their structure, and some of their optical and viscoelastic properties. As this is an encyclopedia of *applied* physics, we shall pay special attention to their applications in display devices or for other purposes.

The many liquid-crystal phases can be divided broadly into two classes. The first liquid crystals discovered, and the most studied, are those classified as *thermotropic*. These are commonly formed from a single organic compound, and the liquid-crystal phases are obtained as the temperature is changed. Many thermotropic phases are formed from "mixtures" of different compounds in order to obtain the liquid-crystal order over a desired temperature range, but these are really solutions in which the different components are distributed quite uniformly. Thermotropic liquid crystals can also be formed from polymers. These are usually classified into two types, depending on whether the polymer main-chain or attached side-chain units are mesogenic.

Most commonly, molecules exhibiting liquid-crystalline phases are rodlike, but disk-shaped molecules have also been found to have liquid-crystal phases (Chandrasekhar *et al.*, 1979). The important thing is that the molecules should have considerable shape anisotropy. For example, for mesomorphic rodlike molecules, the ratio of the length of the molecule to its breadth is typically about 5:1. Two typical thermotropic molecules are shown in Fig. 1. The thermotropic phases are broadly classified as nematic, cholesteric, and smectic; about a dozen different smectic phases have been identified.

Liquid crystals that involve molecules in solution, usually amphiphiles dissolved in water, are called *lyotropic*. As they are systems with at least two components, this provides an extra degree of freedom. The types of order that occur in lyotropic phases are analogous to those in thermotropic phases, but usually involve the ordering of aggregates of molecules. Materials that show lyotropic order may also, with somewhat different amounts of solvent or solute, show non-liquid-crystal phases, such as microemulsions, that are technologically useful.

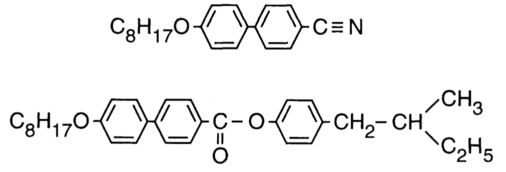

FIG. 1. The structure of two typical thermotropic liquid-crystal molecules.

1. NEMATIC LIQUID CRYSTALS

All liquid crystal phases have long-range orientational order of the molecules. The nematic phase, which is the simplest phase, has only this type of order to distinguish it from a disordered isotropic liquid. The name nematic comes from the Greek word for thread, and was given to this phase because defects in the order, somewhat analogous to the domain walls of ferromagnets, resemble threads when observed under a microscope. A schematic of the molecular order in a nematic phase is shown in Fig. 2. While the molecules are orientationally ordered, the nematic phase has continuous translational symmetry.

1.1 Optical Properties of Nematics

The optical polarizability of the nematic molecules is anisotropic, and so a phase in which molecules tend to orient parallel to one another will be birefringent. Most nematics are uniaxial; that is, there is a single optic axis that the molecules tend to align parallel to, and the material has rotational symmetry about that axis. The optic axis of the material is usually denoted by a unit vector field **n**, called the director, which is also the principal axis of the refractive-index tensor. (Recently, biaxial lyotropic nematics have been discovered; these would require two directors.) Choosing **n** to lie along the z direction, we may let θ be the angle between the symmetry axis of a molecule and the director as shown in Fig. 3. The nematic order may be described quantitatively by the order parameter

$$\mathscr{S}=\tfrac{1}{2}\langle 3\cos^2\theta-1\rangle, \qquad (1)$$

where the average is taken over a small, but macroscopic, volume. In Eq. (1), $\mathscr{S}$ has been defined to be 0 for randomly oriented molecules and 1 if all were perfectly aligned. It is also possible for $\mathscr{S}$ to be negative; the minimum value $\mathscr{S}$ can have is $-\frac{1}{2}$, which occurs when all molecules are randomly oriented in

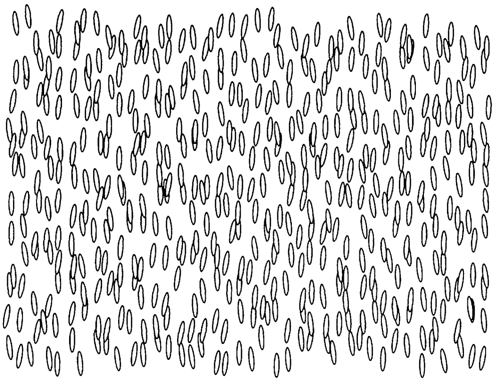

FIG. 2. Schematic of the local arrangements of molecules in a nematic liquid-crystal phase.

a plane perpendicular to the optic axis. The molecules also have an anisotropic diamagnetic susceptibility, largely due to the benzene rings most of them contain, and the diamagnetic susceptibility of the nematic will be given by

$$\chi_{\alpha\beta}=\tfrac{1}{3}\Delta\chi\mathcal{S}(3n_\alpha n_\beta-\delta_{\alpha\beta}), \tag{2}$$

where the subscripts α and β denote Cartesian components, and $\Delta\chi$ would be the anisotropy for a perfectly ordered nematic. A similar phenomenological relation can be defined to represent the optical birefringence by expressing the dielectric constant tensor as

$$\epsilon_{\alpha\beta}=\tfrac{1}{3}\Delta\epsilon\mathcal{S}(3n_\alpha n_\beta-\delta_{\alpha\beta}). \tag{3}$$

Equation (2) provides, in principle, a means to determine $\mathcal{S}$ from a macroscopic measurement if the fundamental diamagnetic properties of the molecule are known. The uncertainty of local field corrections makes it impossible to relate $\Delta\epsilon$ to molecular properties, and so Eq. (3) cannot be used in the same way. However, Eq. (3) does provide a quantitative description of the birefringence of nematic phases, which is large; a typical nematic has $\Delta\epsilon\simeq 1$. Because nematics are fluids, the birefringence can be modified by relatively modest applied forces, which makes them useful for displays.

1.2 Viscoelastic Properties of Nematics

The effect of electric fields, or other perturbations, on nematic liquid crystals is usually described in terms of distortions of the director field **n**. The lowest energy state of the nematic occurs when **n** is spatially uniform. Although the molecules themselves are not end-for-end symmetrical, all known nematic phases are unchanged if the direction of **n** is reversed. (If this were not so, they would be ferroelectric, and no ferroelectric nematics have been discovered.) A general distortion can be expressed in terms of three fundamental distortions commonly called splay, twist, and bend. Mathematically, these are represented by $(\nabla\cdot\mathbf{n})^2$, $(\mathbf{n}\cdot\nabla\times\mathbf{n})^2$, and $(\mathbf{n}\times\nabla\times\mathbf{n})^2$, respectively (de Gennes, 1974). The energy per unit volume of a distorted nematic in the presence of applied electric and magnetic fields is

$$\begin{aligned}\mathcal{F}_{el}=&\tfrac{1}{2}[K_1(\nabla\cdot\mathbf{n})^2+K_2(\mathbf{n}\cdot\nabla\times\mathbf{n})^2\\&+K_3(\mathbf{n}\times\nabla\times\mathbf{n})^2]+\tfrac{1}{2}\epsilon_a(\mathbf{n}\cdot\mathbf{E})^2\\&+\tfrac{1}{2}\chi_a(\mathbf{n}\cdot\mathbf{H})^2,\end{aligned} \tag{4}$$

where ϵ_a is the anisotropy of the dielectric constant at low frequencies, $\chi_a=\chi_\parallel-\chi_\perp=\Delta\chi\mathcal{S}$ from Eq. (2), and K_1, K_2, and K_3 are called the splay, twist, and bend elastic constants. Typical values for the elastic constants are $\sim 10^{-6}$ dyn, χ_a is about 10^{-7} cgs units for molecules with benzene rings, and ϵ_a is approximately 10 for a molecule with a cyano group, like the upper one in Fig. 1. Equation (4) can be used to calculate the size of applied

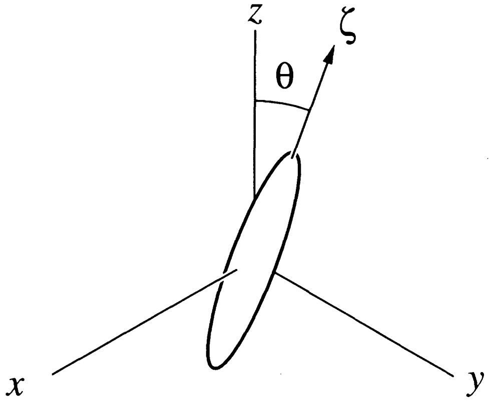

FIG. 3. The coordinates used to describe the orientation of nematic liquid-crystal molecules.

electric or magnetic fields required to cause changes in the orientation of the optic axis **n** (de Gennes, 1974). The sign of χ_a for most nematics is such that the director will orient parallel to an applied magnetic field. A molecule like the upper one in Fig. 1 has a large dipole moment due to the cyano group, and will readily orient parallel to an applied electric field. There are also nematic materials whose optic axis will orient perpendicular to an electric field, but they are little used in modern devices.

The viscosity of nematic liquid crystals is also a tensor, with five independent components (de Gennes, 1974). All of the components are roughly of equal size, ~10 cP, and they, in combination with the elastic constants, determine the frequencies of the director normal modes (Martin *et al.*, 1972) and the response time of devices using nematic liquid crystals. Thermal excitation of these normal modes gives rise to substantial scattering of light, so that the mean free path of a photon can be less than 1 mm. The attenuation of light by scattering in even perfectly aligned nematic materials can be an important consideration in the design of optical devices using these materials.

1.3 Device Applications of Nematics

Most liquid-crystal display (LCD) applications use the liquid crystal to alter the transmission of light through the device. A number of effects have been investigated for use in devices. One of these is the "guest–host" display, in which strongly absorbing dye molecules are dissolved in a nematic liquid crystal that orients them; by reorienting the director variable dichroism can be obtained. However, the dominant type of liquid-crystal display in use today is the kind known as "twisted nematic" (Schadt and Helfrich, 1971), or TN, illustrated in Fig. 4. These devices have become practical thanks to the pioneering work of Gray and co-workers (Gray, 1975) at the University of Hull in England in developing materials with the required properties.

1.3.1 Twisted Nematic Displays. The display cell consists of nematic material between two transparent plates separated by a distance *d*, which is typically about 10 μm. The inner surface of each plate is coated with a transparent conductor such as indium-tin oxide so that an electric field may be applied across the liquid crystal. The plate surfaces are also treated to orient the director of the material in contact with them so that the optic axis will follow a particular direction in the plane of the plate. The orientation directions of the two plates are generally set at a 90° angle, which causes the director of the material between the plates to follow smoothly a 90° twist. The cell is generally used to transmit light and placed between a pair of polarizers. Light incident upon the cell will have its plane of polarization rotated to follow the twist as it passes through the cell; it will pass through if the polarizers are crossed and be blocked if they are parallel. The materials used have $\epsilon_a > 0$, so that a voltage applied between the electrodes will try to align the molecules normal to the plates. Those in contact with the plates will retain their orientation, but at a rather sharply defined threshold most of the liquid crystal between the plates will switch orientation. When this happens, the cell will no longer rotate the polarization of light passing through it. By placing the cell between crossed or parallel polarizers, one may choose to have the electric field either block or permit the transmission of light. The chief advantage of this display is that the liquid crystals are excellent electrical insulators, so that the power consumption is very small. Another advantage is that the response is determined by the voltage between the electrodes, rather than the electric field, so that TN devices are less sensitive to minor variations in the electrode spacing than those using ferroelectric liquid crystals. Several considerations are important for practical TN devices. To avoid electrochemical polarization effects at the electrodes, low-

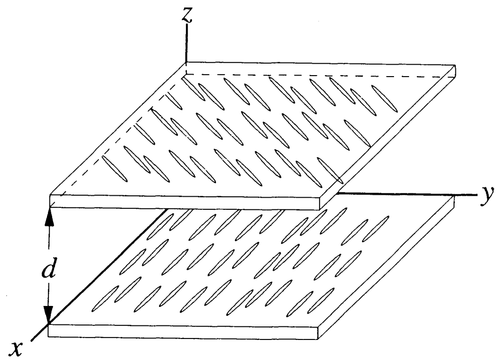

FIG. 4. Schematic of a twisted nematic display cell.

frequency ac fields must be used, rather than dc. The liquid crystal is often doped with small amounts of a chiral material to influence the choice of the twist direction. High information content displays for computer screens often employ a twist angle greater than 90°, frequently as high as 270°. Known as "supertwist," this improves the threshold response and makes possible greater levels of multiplexing and good contrast over a wider reviewing angle. The best displays now use thin-film transistors at each pixel to improve performance.

2. CHOLESTERIC LIQUID CRYSTALS

The name "cholesteric" comes from the historical fact that for a long time most of the known examples of this phase were esters of cholesterol. The molecules of cholesteric phases tend to order parallel to one another as the molecules of nematics do, but they are formed from chiral molecules. The chirality of the molecules causes them to pack in a locally helical structure; in the ordered liquid crystal phase there is a macroscopic precession of the director throughout the sample. Cholesterics have a helical axis to which the director is perpendicular and around which it precesses. The distance travelled along the helical axis for one full turn of the director is called the pitch. Both right- and left-hand cholesterics exist, and the length of the pitch is often comparable to the wavelength of visible light. A cholesteric phase formed from a racemic mixture (equal numbers of molecules with right- and left-handed chirality) will be nematic, and, thermodynamically, cholesteric liquid crystals are a class of nematic.

2.1 Optical Properties of Cholesterics

The magnitude of the pitch plays an important role in determining the optical, dielectric, and magnetic properties of cholesteric phases. Even when disordered, most cholesteric materials have a high value of optical activity; polarization rotations of 300° mm^{-1} are common. Cholesteric phases have a strong circular birefringence, and propagation of light through them (Chandrasekhar, 1977) can exhibit unusual effects. When the pitch is comparable to the wavelength of light, wavelength- and polarization-selective reflections occur, so that cholesteric materials often display bright colors when reflecting white light. The color is determined by the pitch, and is used in applications we shall now discuss.

2.2 Response of Cholesterics to Applied Fields

Because cholesterics are nematics with a macroscopic twist to the director, their response to applied fields can be discussed using a modified form of Eq. (4) in which the twist term is replaced by $K_2(\mathbf{n}\cdot\nabla\times\mathbf{n}+q_c)^2$, where the pitch length $2\pi/q_c$ is long compared to molecular dimensions. The esters of cholesterol have diamagnetic anisotropies about 100 times smaller than those of common nematic materials, and the effects of magnetic fields on them are correspondingly weaker. More recently developed materials in which the chirality comes from a branched chain, as seen in the second molecule in Fig. 1, have greater response to magnetic fields. For cholesterics, we should remember that the susceptibility for fields along the helical axis is $\chi_\perp$ or $\epsilon_\perp$ of Eqs. (2) and (3). For fields perpendicular to the pitch direction, the average susceptibilities $(\chi_\parallel+\chi_\perp)/2$ and $(\epsilon_\parallel+\epsilon_\perp)/2$ will be observed. So, in applying Eq. (4) to cholesterics, $\chi_a=(\chi_\parallel-\chi_\perp)/2$ and $\epsilon_a=(\epsilon_\parallel-\epsilon_\perp)/2$ should be used. If χ_a or ϵ_a is negative, the director will align perpendicular to the appropriate field and the pitch direction will align parallel to it. When χ_a or ϵ_a is positive, the system will align with the pitch direction normal to the field, and the effect of the field will be to unwind the helix. In fact, the pitch diverges logarithmically to ∞ at a critical field that is inversely proportional to the unperturbed pitch. At this point, the cholesteric has been converted to a nematic phase by the field.

Because the pitch modulates a number of properties of the cholesteric phase as one travels along the pitch direction, cholesterics can exhibit viscoelastic behavior analogous to smectic phases. These are discussed by de Gennes (1974).

2.3 Applications of Cholesterics

Most of the applications cholesterics have found to date stem from their selective reflection of colors of light. The color reflected depends upon the pitch, which can vary with temperature through a process analogous to thermal expansion. Cholesteric materials can therefore be used as thermometers with visible readout. They have found application in measuring skin temperatures and the detection of poor contacts and hot spots on printed circuit boards. When composite materials are subjected to a step in heat flux, the resulting surface temperature distribution can reveal information about internal flaws; it has been shown (Williams, 1983) that cholesteric liquid crystals can be used in this application to provide a nondestructive evaluation of flaws in fiberglass boat hulls. By synthesizing and mixing various materials, the temperature at which visible reflections occur can be varied from roughly 0 to 50 °C, and the temperature range over which the reflected color changes from red to blue can be adjusted from about 10 to 1 °C. The materials can also be obtained commercially encapsulated in micron-sized polymer capsules, which makes them easier to handle in many applications.

2.4 Cholesteric "Blue" Phases

An interesting puzzle that has been solved during the last decade is the nature of the "blue" phases exhibited by some cholesteric materials over a narrow temperature range just below their transition to the isotropic phase. The name derives from the bluish appearance they often have; they were first observed by Reinitzer (1888). Blue phases have proved interesting to theorists seeking to understand the significance and broader implications of the chiral term in the free energy, and they have provided a challenge to experimenters seeking to elucidate their structure (Collings, 1990).

By a combination of calculations and polarization-sensitive light scattering, it has been shown that the blue phases consist of an ordered lattice of defects in the cholesteric texture (the spatial orientation field of the director) that are able to accomodate a double twist. Three blue phases are known: blue phase I has body-centered symmetry, blue phase II is simple cubic, and the structure of blue phase III, sometimes called the "fog phase," is yet unknown. It turns out that the state in which the director throughout the sample precesses about a uniform pitch axis is not the lowest elastic free-energy state for these materials. The reason is that the cholesteric would have a lower free energy of interaction between the molecules if the director could twist in two directions at once. This is not possible topologically, but it can occur through the creation of singular line defects (regions where the order parameter goes to zero), which have, however, a high free-energy cost. The tendency to form a blue phase increases as the pitch becomes shorter. The blue phases exist only over narrow temperature ranges and have yet to find applications.

3. SMECTIC LIQUID-CRYSTAL PHASES

The smectic phases represent the next step towards crystalline order beyond the nematic phase. Each smectic phase is named by a letter; these were assigned roughly in the order in which they were discovered. At this writing, smectics *A* through *L* have been identified. In the past, the structure and molecular arrangements of many smectic phases were not known. Phases were identified by their appearance under a polarizing microscope, which gave some indication of the texture (the spatial orientation field of the director), and a new phase was identified as one which did not mix with other known smectic phases. The refusal to mix was a fairly reliable indication the phases had different symmetries. As more has been learned about the structure of smectics, it has become clear that about a third of them are not truly liquid-crystal phases, because they have three-dimensional long-range crystalline order. They are crystals with especially soft elastic constants. The smectic liquid crystal phases all have some translational order in addition to the molecular orientational long-range order of the nematic and cholesteric phases. Some also have a new kind of order, called bond-orientational order, which we shall now discuss.

3.1 Smectic-*A* Phases

The simplest smectic phase is called smectic-*A*. In this phase, the molecules tend to form "layers" that are perpendicular to the director, as illustrated in Fig. 5. The term "layer" is an overstatement of the ordering. In most smectic-*A* materials the layers are not well defined, and what one really has is more accurately described as a modulation of the density

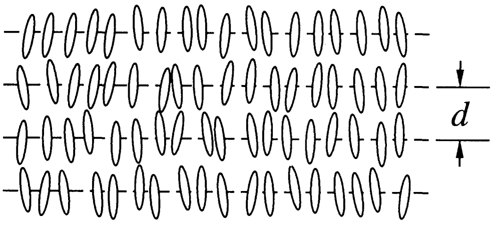

FIG. 5. Molecular arrangement in a smectic-*A* liquid crystal. The director is parallel to the density wave vector.

whose wave vector is along the director. In fact, the smectic-*A* order is commonly represented (de Gennes, 1972) by the order parameter ψ in the following expression for the density of the liquid:

$$\rho(\mathbf{r})=\rho_0[1+\mathrm{Re}\,\psi(\mathbf{r})e^{iq_0z}], \tag{5}$$

where $d = 2\pi/q_0$ is the wavelength of the density wave or the spacing of the smectic "layers." The order parameter ψ is a complex number with two degrees of freedom. The magnitude $|\psi|$ is the amplitude of the density modulation, and the phase represents a displacement of the "layers" in the wave-vector direction, which is usually taken to be z. Smectic-*A* order is really a one-dimensional density wave in a three-dimensional liquid, and can be said to be the simplest kind of freezing or melting to occur in nature. The smectic-*A* phase remains translationally invariant (liquid-like) in directions perpendicular to the density wave vector.

3.1.1 Analogy to Superconductivity.

de Gennes (1972), Kobayashi (1970), and McMillan (1971) all recognized that the two-component order parameter of the smectic-*A* phase was similar to the order parameter used in the Ginzburg-Landau phenomenological model (Ginzburg and Landau, 1950) of superconductivity. The analogy may be seen through the following argument. Suppose, while moving through the smectic-*A* material from point *A* to point *B*, we compute the line integral

$$\frac{1}{d}\int_A^B \mathbf{n}\cdot d\mathbf{r}. \tag{6}$$

This integral calculates the number of "layers" that have been crossed. If the smectic-*A* is free of defects, the result must be independent of the path followed, which means that the integral carried out around a closed path must yield 0. Applying Stokes's theorem, we find that $\nabla \times \mathbf{n} = 0$ in the smectic-*A* phase, and so, referring to Eq. (4), it is apparent that bend and twist director distortions are excluded from the smectic-*A* phase. This is analogous to the exclusion of magnetic fields from superconductors (the Meissner effect). In fact the phenomenological theories of the smectic-*A* to nematic phase transition and the superconducting to normal-state transition are mathematically similar. The liquid crystal director **n** is quite analogous to the magnetic vector potential **A** of the superconductor. As a smectic-*A* phase is approached in the nematic phase, the presence of smectic short-range order causes a divergence of the bend and twist elastic constants, which is the analog of fluctuation diamagnetism in superconductors. Recently a twisted grain-boundary phase has been observed (Goodby *et al.,* 1988) in smectic-*A* materials. This phase is analogous to a lattice of flux penetration vortices in superconductors (Toner, 1991).

3.1.2 Polymorphic Smectic-*A* and Smectic-*C* Phases.

The smectic-*A* phase discussed above has been assumed to have a layer spacing *d* that is approximately equal to the length *l* of a molecule. However, if the molecules have a strong longitudinal dipole moment, such as the top cyanobiphenyl one in Fig. 1, there will be near-neighbor antiparallel correlations. These correlations can lead to different polymorphic forms of the basic smectic-*A* symmetry, which can be characterized by the ratio d/l. These have largely been elucidated through the pioneering work of the Bordeaux group (Sigaud *et al,* 1979; Hardouin *et al.,* 1983). Four forms of smectic-*A* are now recognized: smectic-A_1 with $d \simeq l$, smectic-A_2 with $d \simeq 2l$, smectic-A_d with $l<d<2l$, and smectic-$\tilde{A}$ which has the same local symmetry as smectic-A_2 but with a transverse dipolar modulation of the structure. The existence of these polymorphic smectic-*A* phases has been explained by a phenomenological model (Prost and Barois, 1983) that has two coupled order parameters, the mass-density wave and a dipolar-density wave. This approach has predicted novel critical phenomena associated with the transitions between different polymorphic types. Many of these predictions have been verified experimentally (Garland, 1990). Analogous modulated variants of the smectic-*C* phase, introduced below, have also been found to exist (Hardouin *et al.*, 1983; Huang *et al.*, 1984; Safinya *et al.*, 1986a).

3.2 Smectic-*C* Phases

In smectic-*C* liquid crystals the molecules are disordered within the layers as in a smectic-*A* phase, but they are also tilted with respect to the layer normal (Fig. 6); in other words, the density wave vector and the director are not parallel. This means that smectic-*C* phases are optically biaxial; often this complication can be approximated by simply referring to separate components of a director in the plane of the smectic layers and normal to them. In a well-established smectic-*C* phase enough energy is required to change the tilt angle with respect to the layers (it is coupled to the layer thickness) that it fluctuates very little. The component of the smectic-*C* director in the plane of the layers is almost as free to fluctuate in orientation as the director of a nematic phase, and these fluctuations provide a time and space dependence to the refractive indices that can scatter light strongly.

In many cases, the smectic-*C* phase will form from a smectic-*A* by a continuous tilt of the molecules. This is a second-order phase transition that has the same symmetry properties as transition to a superfluid phase in helium. It should, therefore, exhibit critical behavior. Careful experiments have shown, however, that the growth of the tilt can be described by a mean-field model (Birgeneau *et al.*, 1983). The reason for this is that a large bare coherence length reduces the effect of the thermal fluctuations responsible for critical behavior to a region very close to the transition temperature (Ginzburg, 1960). In some materials, the smectic-*C* phase forms directly from the nematic without an intervening smectic-*A* phase. In these cases the tilt angle is usually large, about 45°, and only weakly temperature dependent. The smectic-*C* to nematic transition is first order because of the effect of thermal fluctuations.

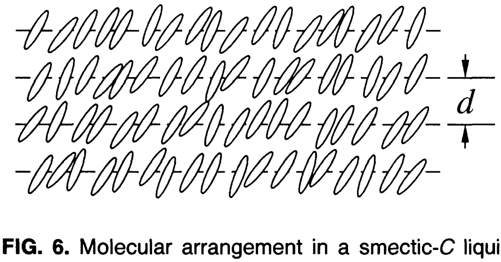

FIG. 6. Molecular arrangement in a smectic-*C* liquid crystal. The director is tilted with respect to the density-wave vector.

3.2.1 The Absence of True Long-Range Order. As the number of spatial dimensions is reduced, ordered phases lose their order. No long-range order is believed possible in one dimension, because the entropy of the disordered state is overwhelmingly higher. Many of the common three-dimensionally ordered phases cannot have long-range order in two dimensions at finite temperature. That is because thermally excited fluctuations will destroy the order. Solids, for example, cannot exist in two or fewer dimensions because the fluctuations of the atomic positions in the crystal lattice due to thermally excited lattice vibrations become infinite in two dimensions (Landau and Lifshitz, 1980). In the language of phase transitions, we say that the lower marginal dimensionality for the existence of solid crystals is two.

Smectic-*A* and smectic-*C* liquid crystals are examples of a one-dimensional density wave in a three-dimensional liquid. The marginal dimensionality for such a material is three (Caillé, 1972), so these smectic phases can, strictly speaking, only exist in more than three dimensions. Precisely at the marginal dimensionality, the effects of the fluctuations are subtle. Whether or not long-range order exists can be discussed quantitatively by defining a correlation function

$$\mathscr{G}(\mathbf{r}) = \langle e^{i\mathbf{G}\cdot[\mathbf{u}(0)-\mathbf{u}(\mathbf{r})]}\rangle, \qquad (7)$$

where $\mathbf{G}$ is a reciprocal-lattice vector of the crystal, and $\mathbf{u}(\mathbf{r})$ is the displacement of an atom at $\mathbf{r}$ from its equilibrium position. In the case of a smectic-*A* phase, $\mathbf{G}$ would be the smectic density wave vector q_0 and $\mathbf{u}$ the displacement of a smectic layer from its proper position. This function is useful because its Fourier transform gives the x-ray Bragg scattering from the crystal lattice or smectic density wave. If there is long-range order, $\mathscr{G}(\mathbf{r})$ will have a finite value as $r \to \infty$. If there is only short-range order, $\mathscr{G}(\mathbf{r})$ decays exponentially to zero as $e^{-\kappa r}$. At the marginal dimensionality, $\mathscr{G}(\mathbf{r})$ goes to zero at large distances, but does so algebraically as $r^{-\eta}$, where η is a small number ~ 0.1. When there is long-range order, the Bragg scattering will give a resolution-limited peak. Short-range order will give a broad peak whose width is proportional to κ. The algebraic decay at the marginal dimensionality gives a scattering peak that is a cusp and is very difficult to distinguish experimentally from the resolution-limited peak of true long-range order. The experimental trick is to use a high-resolution spectrometer with a sharply decaying resolution function (Als-Nielsen *et al.*, 1980) obtained from multiple Bragg reflections in perfect collimating crystals. Most of the properties of smectic-*A* and smectic-*C* liquid crystals are not significantly different than they would be if the order were truly long range. However, they have been of fundamental interest to demonstrate the existence of marginal dimensionality and to study in three dimensions behavior that occurs only in two dimensions in most other materials.

3.3 Ferroelectric Smectic-*C* Phases

Many liquid crystal molecules have dipole moments, and there appears to be no intrinsic reason why nematic phases could not be ferroelectric, but none has ever been observed. Meyer realized that it should be possible to obtain ferroelectricity in a smectic-*C* liquid crystal if the molecules were chiral (Meyer *et al.*, 1975). We can illustrate the idea by referring to Fig. 7. In the figure we represent the chirality by choosing molecules shaped like a hockey stick with a dot on the left side. Suppose the dot represents an electric dipole pointing from the side of the molecule. The second molecule in Fig. 1 is, in fact such a molecule. The top panel, Fig. 7(a), shows a smectic-*A* phase of these molecules. Two alternatives for a smectic-*C* phase are shown in the other two panels of the figure. Both have the end-for-end symmetry of liquid crystal phases, but the arrangement in Fig. 7(c) permits a much closer packing of the molecules and is the one that might be expected to occur once the molecules tilt. It also has the dots on the molecules visible and will have a net electric dipole moment in the plane of the smectic-*C* layers. This is the origin of the ferroelectricity that enables applied fields to control the molecular orientation and leads to the development of a device we shall discuss presently. Ferroelectric smectic-*C* phases are commonly referred to by the notation smectic-*C**. In a racemic mixture, half the molecules would have the dot on the left side, and half would have it on the right; hence there would be no net dipole moment.

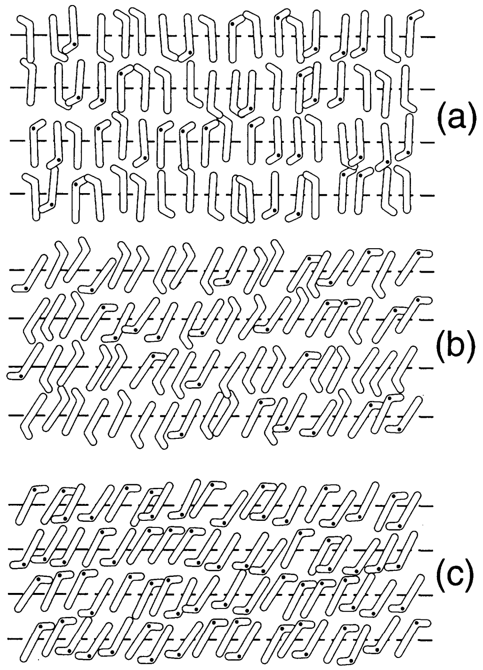

FIG. 7. Smectic phases of chiral molecules. (*a*) A smectic-*A* phase, (*b*) a tilted smectic-*C* phase with no ferroelectric order, and (*c*) a ferroelectric smectic-*C* phase.

3.4 Bond-Orientationally Ordered Phases

We mentioned earlier that the structure of many smectic liquid crystal phases was unknown until recently. The concept of bond-orientational order was the missing idea that was necessary to characterize the structure of the different smectic phases. The idea originated in theoretical investigations of melting in two dimensions. Liquids have both continuous rotational and continuous translational symmetries. In crystalline solids, both of these symmetries have been "broken" and are discrete. As we mentioned in Sec. 3.2.1 above, solids cannot have true long-range order in two dimensions. However, they can have a phase, different from a two-dimensional liquid, which has power-law decay of the correlation function given in Eq. (7). This is commonly called a Landau–Peierls state, in order to recognize earlier work of Peierls and Landau on the topic. In 1973, Kosterlitz and Thouless proposed a theory of melting in two dimensions, a dislocation-mediated transition between a two-dimensional Landau–Peierls state and a two-dimensional liquid. This was a model for breaking the continuous translational symmetry. Extending it to consider the rotational symmetry as well, Halperin, Nelson, and Young predicted the existence of an intermediate "hexatic" phase that had the discrete rotational symmetry of a crystal combined with the continuous translational symmetry of a liquid. The idea is illustrated by the sketch shown in Fig. 8. Even in a two-dimensional liquid, one might expect local arrangement of the molecules to be in hexagonal clusters whose size is a correlation length ξ, as shown in Fig. 8. There would, however, be no correlation between the positions or orientations of such clusters if they are separated by more than ξ in distance. If the material were to form a hexagonal crystal, such local clusters would be expected to have the same orientation of the imaginary lines joining near neighbors, and to be separated by an integral number of lattice vectors. In effect the translational and rotational corelation lengths have become infinite. In the hexatic phase, which has "bond orientational" order, there is no positional correlation between widely separated clusters, but lines joining near neighbors have the same orientation throughout the sample. Only the rotational correlation length is infinite. Thus the rotational symmetry is discrete although the continuous translational symmetry remains unbroken. It was recognized by Landau, and subsequently by Mermin, that it was theoretically possible for a phase with these properties to exist in nature. In effect, it would be a liquid in which one knew in which directions it was necessary to go to encounter a neighboring molecule, and these directions were the same throughout the liquid, but one did not know how far to go. Birgeneau and Litster (1978) proposed that extending these two-dimensional ideas to three dimensions might provide an explanation for some of the higher-ordered smectic phases, such as smectic-*B*, whose structure was unknown. They proposed that within the smectic layers there was bond-orientational order, but no translational order.

Although some smectic-*B* phases have three-dimensional long-range translational order (Pershan *et al.*, 1979), others have been found with only long-range bond-orientational order (Moncton and Pindak, 1979; Pindak *et al.*, 1981). These are usually called smectic-B_H to indicate the presence of hexatic order. Tilted analogs of the smectic-B_H are also known; these are the smectic-*F* and smectic-*I* phases. In the smectic-*I* the molecules tilt towards the position of their nearest neighbors, while in the case of smectic-*F* phases the tilt direction is between the nearest neighbors. There are no interlayer correlations in the smectic-B_H, but

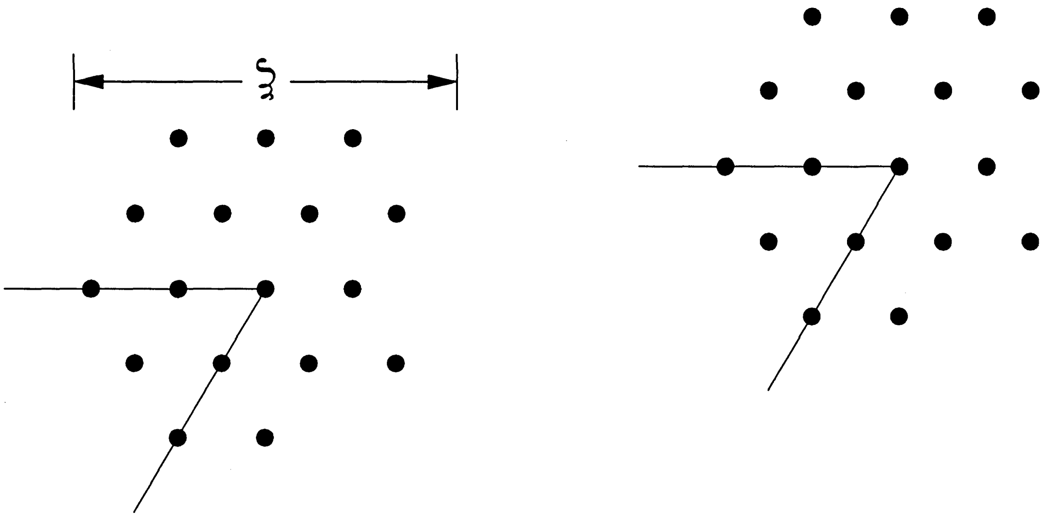

FIG. 8. Sketch to illustrate the concept of bond-orientational order. The two clusters have the same orientation but their positions are not correlated.

in the tilted phases the tilt applies an effective field to the bond-orientational order and causes interlayer correlations. In principle, the tilt will induce bond-orientational order in the smectic-*C* phase, changing it to an *F* or *I* phase. Strictly speaking, therefore, the smectic-*C* phase does not exist. This effect, so far, has been too small to detect and seems to have few practical consequences.

The smectic phases with bond-orientational order exist only over a narrow temperature range, and have not seen applications. They have, however, been studied extensively (Brock *et al.*, 1989) in order to elucidate the properties of a phase of matter whose existence would have greatly surprised most solid state physicists 15 years ago.

3.5 Higher-Order Smectic Phases

In addition to the fluid phases described so far, there are a number of smectic phases with true long-range translational order in three dimensions. These are the smectic-*B*, -*G*, -*J*, -*E*, and -*K* phases. There are differences in the molecular tilt, stacking of layers, and the extent of rotational disorder about the long axis of the molecules among these phases, but they are not true liquid crystals. They are crystalline solids, although in many cases the elastic constants can be as much as 10^6 weaker than the elastic constants of more common crystals.

3.6 Applications of Smectic Phases

3.6.1 Smectic-*A* Devices. Although the higher viscosity and greater force required to alter the texture of smectic-*A* phases renders them less suited for electro-optic displays, these are features that make them attractive for different types of memory devices (Coates, 1990). In one type of memory device the smectic-*A* is aligned "homeotropically" (molecules perpendicular to the substrate surface). This will provide a clear state for transmission of light between two glass substrates. A focused laser beam is used to heat a small area of liquid crystal into the isotropic state. When the laser is turned off, the liquid crystal cools rapidly to the smectic-*A* phase. The rapid cooling results in misalignment and the different texture scatters light, thus storing information that can be read out optically. The texture can be restored to homeotropic, and the stored information erased, either by the application of an electric field or by heating the entire liquid crystal to the isotropic state and cooling slowly. This makes possible a very high–information-content, although not particularly fast, read-and-write optical storage device.

The electroclinic effect in chiral smectic-*A* liquid crystals, such as in Fig. 7(a), can also provide device capability. In the electroclinic effect (Garoff and Meyer, 1977) an applied electric field can induce a tilt in the optic axis that is proportional to the applied voltage. This can, in principle, make possible features such as variable gray scale and analog phase modulation in a chiral smectic-*A* device.

3.6.2 Ferroelectric Smectic-*C* Applications. Ferroelectric smectic-*C** liquid crystals (FLCs) appear to have excellent potential for applications in fast electro-optic switching devices with memory. In such a device the basic configuration is the surface-stabilized ferroelectric liquid crystal (SSFLC) geometry in which the smectic-*C** material is contained between two transparent electrodes separated by about 2μm. The surfaces are treated to obtain a planar orientation of the molecules. Because they are chiral, smectic-*C** phases will have a precessing optical axis analogous to cholesteric phases. But, if the thickness of the SSFLC cell is less than the pitch of the optical axis helix, the interaction with the surfaces will unwind the helix. This produces a "bookshelf" geometry of Fig. 9 in which the smectic layers are perpendicular to the substrates and the molecular orientation is in the plane of the substrate making the tilt angle ω with respect to the layer normal. By reversing the polarity of the applied electric field, the molecules can be caused to flip to an orientation $-\omega$ with respect to the layer normal, hence rotating the optic axis by 2ω. The optimal value for ω is 22.5°, and switching times as fast as 0.1 μs have been demonstrated (Clark and Lagerwall, 1980). The switched state is also stable, and so the device provides storage, which greatly reduces the multiplexing problems of TN (twisted nematic) devices. The small electrode spacing and the fact that the response is determined by the electric field, not the applied voltage, has made these devices more difficult to manufacture. Efforts are also under way to synthesize ferroelectric liquid crystals with a significant dipole mo-

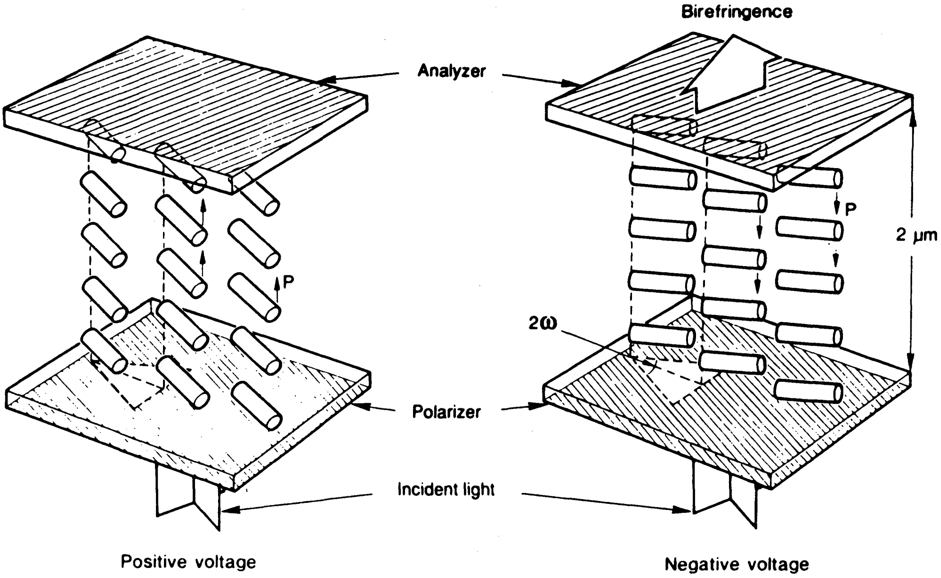

FIG. 9. The "bookshelf" geometry of a surface-stabilized ferroelectric liquid crystal display cell.

ment parallel to the long axis of the molecule; if the efforts are successful these materials could play an important role in FLC devices. Despite their fast switching times and the storage capability, SSFLC devices do not yet appear to have mounted a serious challenge to TN displays. However, this may change; at least one company (Canon Electric Company) has introduced ferroelectric liquid-crystal displays for its notebook computers.

4. DISCOTIC LIQUID CRYSTALS

Until recently it was believed that a prerequisite for the existence of a thermotropic liquid crystal phase was a rodlike shape of the molecules. However, liquid crystal phases have been demonstrated for disklike molecules (Chandrasekhar, 1979). Generally, the molecules have flat cores with long-chain substituents attached to the core. These molecules have been shown to exhibit two distinct types of mesophases: columnar and nematic. In the nematic phase, the disks are orientationally ordered without any positional order.

The columnar phases are formed by the molecules being stacked in columns, like poker chips, and then the columns are arranged in a hexagonal or rectangular fashion. A sketch of the local structure of a columnar discotic phase is shown in Fig. 10. Unlike the nematic phases of rodlike molecules, discotic nematics are usually optically negative ($\epsilon_{\parallel} < \epsilon_{\perp}$). Tilted columnar phases also exist. There is no translational order along the length of the columns. These phases are analogous to the smectic phases, except that they have a two-dimensional density wave in a three-dimensional liquid. This arrangement lacks the Landau-Peierls instability of the smectic-*A* phase and can have true long-range order.

5. POLYMER AND POLYMER-DISPERSED LIQUID CRYSTALS

5.1 Polymer Nematic and Smectic Liquid Crystals

5.1.1 Mesogenic Main-Chain and Side-Chain Polymers. Thermotropic liquid-crystal polymers evolved by researchers incorporating mesogenic monomers into polymer chains (Donald, 1992). They can be incorporated into the main polymer backbone chain, or by adding mesogenic side chains. Polymers with disklike mesogenic units are also known. The main-chain liquid crystal polymers tend to have very high melting temperatures and to be very slow in response to applied electric or

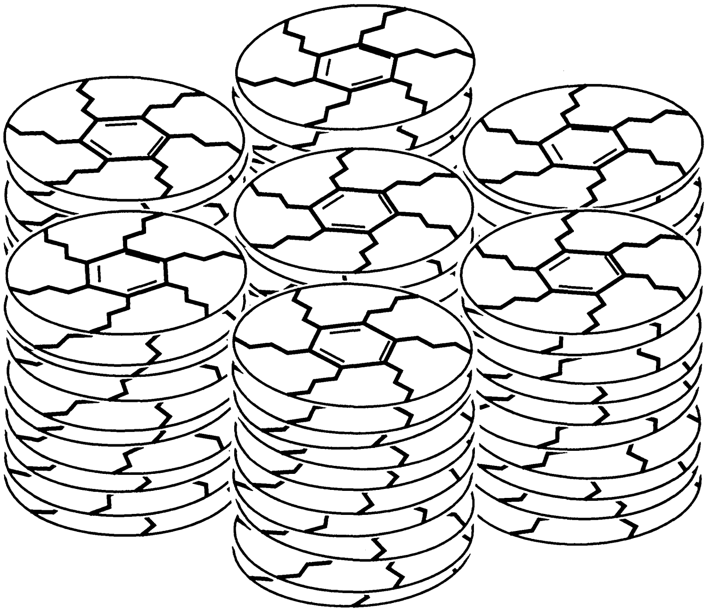

FIG. 10. Schematic of molecular arrangements in a columnar discotic liquid-crystal phase.

magnetic fields. However, one of the viscoelastic features of nematic liquid crystals is a very strong tendency to flow alignment. This can be exploited in polymer processing to produce fibers in which the polymers are oriented; such materials can have a much greater elastic modulus than those prepared conventionally from an isotropic melt. The best known example is Dupont's Kevlar®, which is spun from a nematic phase and has an elastic modulus comparable to that of steel.

The polymers with mesogenic side chains often offer response times to applied fields that are comparable to those of small-molecule liquid crystals. They can have not only nematic, but also cholesteric and smectic phases. Their viscosities and glass transition temperatures are much lower than main-chain polymers. They can be aligned easily by the same techniques (surface rubbing, skew evaporation, and surface texturing of the substrate) used for small-molecule liquid crystals. There is a potential for the polymer main chain to provide added stability to the liquid-crystal order, perhaps preserving some of the features in a solid phase, or to assist in obtaining better alignment of the liquid-crystal phase.

More recently, combined liquid crystalline polymers with mesogenic groups in both the main-chain and side groups have been synthesized. The degree of coupling of the main-chain and side-chain mesogenic groups depends on the molecular "spacer" used to attach the side groups to the main chain.

5.1.2 Applications of Polymer Liquid Crystals. Most of the work on polymeric liquid crystals has been motivated by the belief that these materials have many potential technological applications. The example of Kevlar® has inspired much of the effort. Mesogenic side-chain polymers offer potential for electro-optic displays, nonlinear optical applications, and pyroelectric applications. They may be attractive candidates for erasable holographic elements. This is an area of high current activity, and we shall have to await the results to see how valuable they turn out to be.

5.2 Polymer-Dispersed Liquid Crystals

These materials were the serendipitous result of efforts to improve the stability of liquid crystals aligned for NMR studies (Doane *et al.*, 1986). Liquid crystals can be mixed with the components of an epoxy resin; as the epoxy cures, a phase separation takes place in which much of the liquid crystal is expelled from the epoxy matrix. The result is approximately micron-sized droplets of small-molecule nematic liquid crystals dispersed in a polymer binder. Polymer-dispersed liquid crystals

(PDLCs) are now produced in several ways: solvent-induced precipitation, thermally induced precipitation, emulsion polymerization, or phase separation.

A material with a positive dielectric anisotropy can be aligned by a field applied when the PDLC is sandwiched between transparent electrodes. Normally the nematic droplets will scatter light strongly, and the material appears like a translucent shutter. If the ordinary refractive index of the liquid crystal is matched to that of the polymer matrix, the scattering will disappear when the field aligns the director. PDLCs offer the potential for high-transmission, large-area light shutters and displays with low power consumption. The absence of polarizers increases the transmission considerably, which makes them attractive for projection displays.

6. LYOTROPIC PHASES

The lyotropic liquid crystal phases contain amphiphilic or surfactant molecules. Because one end of the molecule prefers to be in contact with water and the other prefers oil, the molecules will go to the oil-water interface of a "solution" that contains both. If only oil or water is present along with the surfactant, various structures will form in order to segregate the oil-loving moiety, say, from water. The simplest type of aggregate formed by surfactant molecules in water is a micelle, shown in Fig. 11. The hydrophobic tails of the molecules can minimize contact with the water by segregating themselves in the center of the micelle. There is a price in reduced entropy as well as crowding of the hydrophobic tails to be be paid when this structure is formed. Usually double-tailed surfactants do not form micelles. In an oil solution, inverted micelles can form in which the hydrophilic moiety goes to the center of the micelle. When oil, water, and surfactants are mixed a wide variety of phases can form (Fontell, 1981). Only some of these phases have liquid crystalline order, but all are probably closely related and their existence and properties are determined by the same physical principles (Israelachvili, 1985). The surfactant phases are interesting because, as in biological systems, entropy plays a dominant role in their properties and behavior; some also provide interesting examples of two-dimensional physics. Lyotropic liquid-crystal phases are not yet as well studied or fundamentally understood as their thermotropic counterparts are.

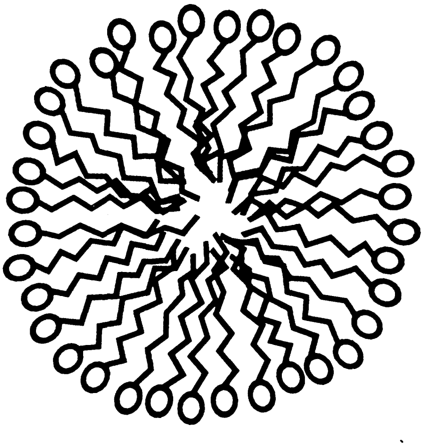

FIG. 11. Cross-sectional sketch of a micelle formed by amphiphilic molecules in water.

6.1 Phases with Ordered Micelles

Micelles are usually spherical, but under conditions that are not entirely understood, the micelle may not be spherical. Nonspherical micelles can order orientationally to form a nematic phase, and micellar nematic phases can be formed from either rodlike or disklike (Fujiwara and Reeves, 1980) micelles. As well as the usual hydrocarbon-tailed surfactants, micellar liquid crystals can be formed by surfactants with fluorocarbon tails (Tiddy, 1972; Boden *et al.*, 1979). The micellar nematic phases have viscoelastic properties (Haven *et al.*, 1981) and normal modes (Kumar *et al.*, 1983) that are closely similar to those of small-molecule thermotropic phases. In some materials the rodlike micelles can grow very long and order in a hexagonal array; these "middle soap" phases are probably the lyotropic analogs of the columnar discotic thermotropic phases. With chiral surfactants, lyotropic cholesteric phases are possible (Covello *et al.*, 1983).

6.2 Lyotropic Lamellar Phases

Many surfactants, especially biologically derived ones with double aliphatic tails, will form bilayers in aqueous solution. The phase formed depends sensitively on the spontaneous curvature of the bilayer. If it tends to be flat, then the system will form a lamellar structure as in Fig. 12; the resulting phase is closely similar to the thermotropic smectic-*A* phase; in the language of lyotropic liquid crystals (Tiddy, 1980), this phase is called L_α. If there is a tendency for the bilayer to curve, then interesting bicontinuous phases may form in which the bilayer forms a cubic lattice of intersecting curved surfaces (Gruner, 1989), sometimes referred to as the "plumber's nightmare."

The L_α phase has the same symmetry and normal-mode structure as the thermotropic smectic-*A* phase, and it also has the same lack of true long-range order in three dimensions (Safinya *et al.*, 1986b). The bilayers are more sharply defined than the smectic-*A* density wave, and the higher-order Bragg peaks are usually present in the x-ray scattering. With nonionic surfactants in water, such as the polyethylene glycol ethers (Strey *et al.*, 1990), or bilayers separated by oil (Appell *et al.*, 1989), it is possible to prepare highly swollen L_α phases in which bilayers a few nm thick are separated by many hundreds of nm. These systems are stabilized by repulsive forces between the bilayers, which can act over a very long range. In fact their properties are dominated by a previously unrecognized force predicted theoretically by Helfrich (1978). This force comes into play when any electrical forces between the bilayers are extremely weak and the bilayers are rather flexible. Its origin is entropic; the thermal fluctuations of the layers is sufficiently great that they repel through colliding with each other as they fluctuate. The highly swollen L_α phases are interesting systems in which to study the Helfrich force, as it is the dominant force stabilizing them. They are also interesting experimentally because Bragg scattering, which normally requires x rays, can be done with visible light. The energy resolution of 10^{15} available in light scattering enables a careful investigation of the dynamical behavior that one cannot obtain with x rays.

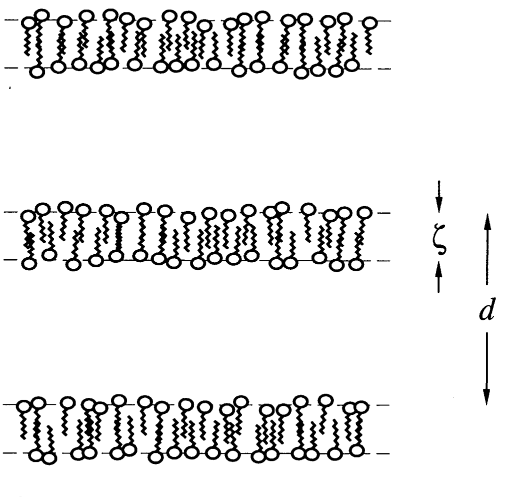

FIG. 12. A smectic-*A* lyotropic lamellar liquid-crystal phase. The surfactant bilayers of thickness ζ are separated by regions of water of thickness $d-\zeta$.

6.3 Applications of Lyotropic Liquid Crystals

Lyotropic liquid crystal phases usually occupy only a small part of the complicated phase diagrams that result when oil, water, salt, and one or more surfactants are combined. Surfactants are extraordinarily useful materials with many technological applications. These are usually connected with the ability of the surfactant molecules to reduce the interfacial energy (surface tension) between two liquids that would not ordinarily mix because of this high interfacial energy. A common example is the use of microemulsions in tertiary oil recovery. A microemulsion is a thermodynamically stable phase (Winsor, 1954; Kahlweit *et al.*, 1987) in which oil and water are separated by a random interface of surfactant. The surfactant can reduce the interfacial energy between oil and water, normally about 30 erg cm^{-2}, by 1000 times or more. This is important, for water is pumped into oil wells to force the oil out of small cracks in the rock, and reducing the oil–water surface tension greatly facilitates this process. The practice is not economical with current oil prices, but we shall undoubtedly see it at some time in the future.

The ordered liquid crystal lyotropic phases and their disordered relatives are interesting objects to study for those who seek a better understanding of classical statistical mechanics and random surfaces (Huse and Leibler, 1988), for they are systems in which thermal fluctuations and entropy play an important role. Understanding the ordered liquid-crystal phases can lead to better understanding of the forces and interactions that determine properties of other phases and materials containing biological surfactants such as phospholipids. For example, the viscoelastic properties of lyotropic liquid crystals can explain the shape of red blood cells (Deuling and Helfrich, 1976), and are believed to provide insight into biological membrane functions such as endocytosis, membrane fusion, and cell fission (Gruner, 1989) as well as the formation of vesicles and tubules (Lubensky and Prost, 1992). It also seems that studies of lyotropic liquid-crystal ordering should offer some insight into the self-assembly process (Deckman *et al.*, 1988) of Langmuir–Blodgett films.

7. CONCLUSIONS

In this article we have felt it was important to give an introduction to the structure of the various ordered phases, to indicate their role in the intellectual development of both physical chemistry and condensed matter physics, to point out special viscoelastic and optical properties of thermotropic liquid-crystal phases, and to comment upon the applications they have found. The reader should consult the literature references for more information. The journals of the Institute for Electrical and Electronic Engineers, the Society for Information Displays, and SPIE–The International Society for Optical Engineering (Jacobs, 1992) are especially good sources of information about display uses of thermotropic liquid crystals, which remains their major area of application.

ACKNOWLEDGMENTS

One of us (J.D.L.) would like to express thanks to the National Science Foundation for support from Grants No. DMR-9022933 and No. DMR-9014886 during the preparation of this article and for support of the research some of it is based on.

GLOSSARY

Amphiphile: A molecule consisting of two parts, one being soluble in water, the other in oil.

Birefringence: The condition of having two different refractive indices; the one observed depends on the polarization state of the light.

Bond-Orientational Order: A type of order in which the imaginary lines joining the centers of neighboring molecules are ordered, but the positions of the molecules are not. (See Sec. 3.4.)

Chirality: A chiral object is one which has either right- or left-handed symmetry, analogous to the difference between right- and left-handed screw threads.

Correlation Length: Sometimes called coherence length, the distance over which fluctuations in a quantity, usually an order parameter, are correlated. Commonly, the relative strength of correlations at two points separated by a distance r weakens as $e^{r/\xi}$, where ξ is the correlation length.

Director: A name given to a unit vector field that points along the local orientation of the optic axis of a liquid crystal. (See Sec. 1.1.)

Ferroelectric: A material that posesses a permanent electric polarization.

Mesogen: A term used to indicate a molecule that tends to form a liquid-crystal phase, which is sometimes called a mesophase. (Mesophase means "in-between phase.")

Micelle: Roughly sperical aggregate of amphiphilic molecules. (See Sec. 6.1.)

Optical Activity (Circular Birefringence): An optically active material is one whose refractive index differs for right and left circularly polarized light.

Order parameter: A quantity used to describe quantitatively the amount of order in a phase; see Sec. 1.1 for an example describing molecular orientational order in a nematic liquid crystal. It is defined to be zero in a disordered phase.

List of Works Cited

Als-Nielsen, J., Litster, J. D., Birgeneau, R. J., Kaplan, M., Safinya, C. R., Lindegaard-Andersen, A., Mathiesen, S. (1980), *Phys. Rev. B* **22**, 312–340.

Appell, J., Bassereau, P., Marignan, J., Porte, G. (1989), *Coll. Pol. Sci.* **267**, 600–647.

Birgeneau, R. J., Litster, J. D. (1978), *J. Phys. Lett.* **38**, 1399–1407.

Birgeneau, R. J., Garland, C. W., Kortan, A. R., Litster, J. D., Meichle, M., Ocko, B. M., Rosenblatt, C., Yu, L.-J. (1983), *Phys. Rev. A* **27**, 1251–1284.

Boden, N., Jackson, P. H., McMullen, K., Holmes, M. C. (1979), *Chem. Phys. Lett.* **65**, 476–511.

Brock, J. D., Birgeneau, R. J., Litster, J. D., Aharony, A. (1989), *Phys. Today,* **42** (7), 52–59.

Caillé, A. (1972), *C. R. Acad. Sci. Ser. B* **274**, 891–902.

Chandrasekhar, S. (1977), *Liquid Crystals*, Cambridge: Cambridge University Press.

Chandrasekhar, S., Sadashiva, B. K., Suresh, K. A., Madhusudana, N. V., Kumar, S., Shashidhar, R., Venkatesh, G. (1979), *J. Phys. (Paris) Colloq.* **40**, C3-120–3-124.

Clark, N. A., Lagerwall, S. T. (1980), *Appl. Phys. Lett.* **36**, 899–912.

Coates, D. (1990), in: B. Bahadur (Ed.), *Liquid Crystals: Applications and Uses*, Singapore: World Scientific, p. 275.

Collings, P. J. (1990), *Liquid Crystals: Nature's Delicate Phase of Matter*, Princeton: Princeton University Press, Chap. 10.

Covello, P. S., Forrest, B. J., Marcondes-Helene, M. E., Reeves, L. W., Vist, M. (1983), *J. Phys. Chem* **87**, 176–191.

Deckman, H. W., Dunsmuir, J. H., Garoff, S., McHenry, J. A., Peiffer, D. G. (1988), *J. Vac. Sci. Tech. B* **6**, 333.

Deuling, H. J., Helfrich, W. (1976), *Biophys. J.* **16**, 861–873.

Doane, J. W., Vaz, N. A., Wu, B.-G., Zumer, S. (1986), *Appl. Phys. Lett.* **48**, 269–274.

Donald, A. M. (1992), *Liquid Crystalline Polymers*, Cambridge, U.K.: Cambridge University Press.

Fontell, K. (1981), *Mol. Cryst. Liquid Cryst.* **63**, 59–72.

Fujiwara, F. Y. Reeves, L. W. (1980), *J. Phys. Chem.* **84**, 653–670.

Garland, C. W. (1990), in: J. C. Tolédano (Ed.), *Geometry Thermodynamics*, New York: Plenum.

Garoff, S., Meyer, R. B. (1977), *Phys. Rev. Lett.* **38**, 848–861.

de Gennes, P.-G. (1972), *Solid State Comm.* **10**, 753–767.

de Gennes, P.-G. (1974), *The Physics of Liquid Crystals*, London: Oxford University Press.

Ginzburg, V. L. Landau, L. D. (1950), *Sov. Phys. JETP* **20**, 1064.

Ginzburg, V. L. (1960), *Sov. Phys. Sol. State* **2**, 1824–1834.

Goodby, J., Waugh, M. A., Stein, S. M., Chin, E., Pindak, R., Patel, J. S. (1988), *Nature* **337**, 449–452.

Gray, G. W. (1975), *J. Phys. (Paris) Colloq.* **36**, C1-337–C1-347.

Gruner, S. M. (1989), *J. Phys. Chem.* **93**, 7562.

Hardouin, F., Levelut, A. M., Achard, M. F., Sigaud, G. (1983), *J. Chim. Phys.* **80**, 65–xx.

Haven, T., Armitage, D., Saupe, A. (1981), *J. Chem. Phys.* **75**, 352–xxx.

Helfrich, W. (1978), *Z. Naturforsch.* **33a**, 305–315.

Huang, C. C., Lien, S. C. (1981), *Phys. Rev. Lett.*

Huang, C. C. Viner, J. M. (1982), *Phys. Rev. A* **25**, 3385–3396.

Huang, C. C., Lien, S. C., Dumrongrattana, S., Chiang, L. Y. (1984), *Phys. Rev. A* **30**, 965.

Huse, D. A., Leibler, S. (1988), *J. Phys. (Paris)* **49**, 605–621.

Israelachvili, J. N. (1985), *Intermolecular Surface Forces*, New York: Academic.

Jacobs S. D. (Ed.) (1992), *Selected Papers on Liquid Crystals for Optics*, Bellingham: WA: SPIE Optical Engineering.

Kahlweit, M., Strey, R., Haase, D., Kuneida, H., Schmeling, T., Faulhaber, B., Borkovec, M., Eicke, H.-F., Busse, G., Eggers, F., Funck, Th., Richmann, H., Magid, L., Söderman, O., Stilbs, P., Winkler, L., Dittrich, A., Jahn, M. (1987), *J. Coll. Int. Sci.* **118**, 436–453.

Kobayashi, K. K. (1970), *J. Phys. Soc. Jpn.* **29**, 101–109.

Kumar, S., Yu, L.-J., Litster J. D. (1983), *Phys. Rev. Lett.* **50**, 1672–1677.

Landau, L. D. Lifshitz, E. M. (1980), *Statistical Physics*, 3rd ed., Pt. 1, New York: Pergamon, Chap. 13.

Lubensky,T. C., Prost, J. (1982), *J. Phys. II France* **2**, 371–387.

Martin, P. C., Parodi, O., Pershan, P. S. (1972), *Phys. Rev. A* **6**, 2401–2413.

McMillan, W. (1971), *Phys. Rev. A* **4**, 1238.

Meyer, R. B., Liebert, L., Strzelecki, L., Keller, P. (1975), *J. Phys. (Paris) Lett.* **36**, L-69–L-71.

Moncton D. E., Pindak, R. (1979), *Phys. Rev. Lett.* **43**, 701–724.

Pershan, P. S., Aeppli, G., Litster, J. D., Birgeneau, R. J. (1981), *Mol. Cryst. Liquid Cryst.* **67**, 205.

Pindak, R., Moncton, D. E., Davey, S. C., Goodby, J. W. (1981), *Phys. Rev. Lett.* **46**, 1135–1144.

Prost, J. Barois, P. (1983), *J. Chim. Phys.* **80**, 65–72.

Reinitzer, F. (1888), *Monatsh. Chem.* **9**, 421.

Safinya, C. R., Roux, D., Smith, G. S., Sinha, S. K., Dimon, P., Clark, N. A., Bellocq, A. M. (1986a), *Phys. Rev. Lett.* **57**, 2718.

Safinya, C. R., Varaday, W. A., Chiang, L. Y., Dimon, P. (1986b), *Phys. Rev. Lett.* **57**, 432.

Schadt M., Helfrich, W. (1971), *Appl. Phys. Lett.* **18**, 127–131.

Sigaud, G., Hardouin, F., Achard, M. F., Gasparoux, H. (1979), *J. Phys. (Paris) Colloq.* C3-356–C3-368.

Strey, R., Schomäcker, R., Roux, D., Nallet, F., Olsson, U. (1990), *J. Chem. Soc. Faraday Trans.* **86**, 2253–2861.

Tiddy, G. J. T. (1972), *J. Chem. Soc. Faraday Trans. I* **68**, 608–612.

Tiddy, G. J. T. (1980), *Phys. Rep.* **57**, 1–46.

Toner, J. (1991), *Phys. Rev. B*, 8289–8296.

Williams, J. H., Jr. (1983), *Mater. Eval.* **41**, 202–210.

Winsor, P. A. (1954), *Solvent Properties of Amphiphilic Compounds*, London: Butterworths.

Further Reading

Belyakov, V. A. (1992), *Diffraction Optics of Complex-Structured Media*, New York: Springer. Discussion of the optical properties of cholesteric liquid crystals, and their relation to other scattering processes in physics.

Brock, J. D., Birgeneau, R. J., Litster, J. D., Aharony, A. (1989), *Contemp. Phys.* **30**, 321. Bond-orientational order.

Flack, J. H., Crooker, P. P., Johnson, D. L., Long, S. (1984), in: A. C. Griffin, J. F. Johnson (Eds.), *Ordered Fluids and Liquid Crystals*, New York: Plenum. Cholesteric blue phases.

Kosterlitz, J. M., Thouless, D. J. (1973), *J. Phys. C* **6**, 1181–1203. Bond-orientational order.

Landau, L. D. (1937), *Phys. Z. Sowjetunion* **11**, 545–555; republished in D. ter Haar (Ed.) (1965), *The Collected Papers of L. D. Landau*, New York: Gordon and Breach. The Landau-Peierls state of matter, which is manifested in three dimensions in smectic-*A* liquid crystals.

Meiboom, S., Sethna, J. P., Anderson, P. W., Brinkman, W. F. (1981), *Phys. Rev. Lett.* **46**, 1216;

Mermin, N. D. (1968), *Phys. Rev.* **176**, 250–254. The Landau–Peierls state of matter.

Nelson, D. R., Halperin, B. I. (1979), *Phys. Rev. B* **19**, 2457. Bond-orientational order.

Onusseit, H. Stegemeyer, H. (1981), *Z. Naturforsch.* **36a**, 1083. Cholesteric blue phases.

Pershan, P. S. (1988), *Structure of Liquid Crystal Phases*, Teaneck, NJ: World Scientific. A good overview of the structure of the various liquid crystal phases, combined with a selection of articles reprinted from the original literature.

Peierls, R. E. (1934), *Helv. Phys. Acta Suppl.* **7**, 81–xx. The Landau–Peierls state of matter.

Strandburg, K. J. (Ed.) (1992), *Bond-Orientational Order in Condensed Matter Systems*, New York: Springer. Bond-orientational order.

Wright, D. C. Mermin, N. D. (1985), *Phys. Rev.* **A 31**, 3498–3500. Cholesteric blue phases.

Young, A. P. (1979), *Phys. Rev. B* **19**, 1855–1866. Bond-orientational order.

LIQUIDS, SIMPLE, STRUCTURE OF

N. E. CUSACK, *School of Physics, University of East Anglia, Norwich, England*

	Introduction	533
1.	**What is Meant by the Structure?**	534
2.	**Distribution Functions**	534
2.1	The Pair Distribution Function	534
2.2	Long- and Short-Range Order	535
2.3	More Distribution Functions	536
2.4	Pair Distributions in Binary Systems	536
3.	**Diffraction Studies of Pair Distribution Functions**	536
3.1	The Structure Factor	537
3.2	Structure Factors and the PDF	538
3.3	Partial Structure Factors in Binary Systems	538
3.4	Neutron Diffraction and Isotopic Substitution	539
4.	**Intermolecular Forces and Liquid-State Theory**	539
4.1	Liquid-State Theory	540
4.2	Computational Aids	541
5.	**Liquid Argon and Similar Systems**	542
6.	**Metals: Some Observations**	543
7.	**Chemical Short-Range Order**	545
7.1	Chemical Thermodynamic Models	545
7.2	Chemical SRO and Structure Factors	546
8.	**Structure and the Theory of Metals**	547
8.1	The General Concept of a Metal	548
8.2	Forces and Energy	548
8.3	Structure from $\phi_{\text{eff}}(r)$ in Metals	549
8.4	A Successful Example	550
9.	**Molecular Liquids**	551
9.1	Structure Factors of Molecular Liquids	552
9.2	A Representative Experiment	553
9.3	Some General Remarks about Structure in Molecular Liquids	554
10.	**Molten Salts**	555
10.1	Halides	555
10.2	Computer Simulation and Theory	556
11.	**Liquids and Glasses**	558
12.	**Conclusion**	558
	Glossary	558
	Works Cited	559
	Further Reading	560

INTRODUCTION

Although all liquids, and solid glasses too, are structurally disordered, or noncrystalline, there are many varieties of structure, some of which may be called *simple*. What differentiates one great class of liquids from another—say, liquid metals from molten salts—is the different nature of the intermolecular forces or, one could equally well have said, the different kinds of molecule. To see which liquids might reasonably be labeled as simple, we must therefore refer to the interparticle forces. We shall adopt four criteria of simplicity which, though not logically equivalent, are such that, if one holds, it will not be surprising if the others do too.

First, the molecules should be small. Second, they should interact by force fields which, if not spherically symmetrical, depart from this by such a small amount that it may be specified by at most a few nonspherical terms. Third, it should be true to a high degree of approximation that the internal structure of the molecule is an independent entity not affected by the position of neighbors. Fourth, the molecules should act by pairwise additive forces—a well-tried assumption in the physics

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

of simple liquids. Thus the ideal case might well be liquid argon composed of stable monatomic molecules with spherical force fields. Among the simple liquids are metals—although it was a considerable achievement of metal physics to show under what conditions the fourth criterion holds. Also simple are those molten salts in which the long-range interaction is Coulomb attraction or repulsion between ionized atoms as in NaCl. The second criterion allows us to admit small molecules like N_2 or CCl_4, but molecules with quite small numbers of atoms, such as glycerol ($CH_2OHCHOHCH_2OH$), can have shapes and dipole moments which call all four criteria into question. The remaining sections will therefore discuss liquid noble gases, a few molecular liquids, metals, and molten salts.

The criteria exclude several kinds of liquid. The structures of many nonmetallic glasses, both liquid and solid, are formed from ramifying networks of directed covalent bonds; even some elements, such as Te and Se, are excluded for this reason. Liquid crystals have rodlike or otherwise nonspherical molecules that cause structural order in one or two dimensions. Liquid polymers or solutions containing highly complex molecules such as proteins are clearly not simple. Additional complications that exclude liquids from present consideration are macroscopic quantum behavior, as in liquid He; the intermolecular bridges formed by hydrogen bonds, as in water; or large solute particles, as in colloidal solutions. Nevertheless, the diffraction techniques mentioned below for simple liquids remain important even for the complex cases we are excluding. The reader will find discussions of more complex liquids in the articles on LIQUID CRYSTALS, STRUCTURE OF; POLYMERS, STRUCTURE OF; and GLASSES.

1. WHAT IS MEANT BY THE STRUCTURE?

The structure of a pure perfect crystal is totally defined by its unit cell, and the elegant science of crystallography has solved the problem of determining unit cells. Liquids do not have unit cells, and to visualize the atomic arrangement as billiard balls jumbled in a bag is to make a good first guess. The numerous symmetries that control crystal structure mean that the environment of an atom in its unit cell is identical with that of every corresponding atom in every other unit cell. In a liquid no two atoms have the same environment or, if they happened by chance to do so, it would only be transiently, because the atoms are continually interdiffusing and changing neighbors. It follows that structure must be a statistical concept in liquids. For example, in a crystalline binary alloy of *A*-type and *B*-type atoms one can answer the question: Where exactly are the *A* atoms relative to the *B* atoms? In the liquid alloy one can answer a limited number of questions such as: What is the average number of *B* atoms that are nearest neighbors to an *A* atom? How far away are they on the average? In what follows we shall assume that our materials are homogeneous and isotropic. This not only reduces the amount of vector symbolism but is intrinsically reasonable for liquids in thermodynamic equilibrium. It might well be inadequate for strained glasses, which are not isotropic, or multiphase systems, which are not homogeneous.

2. DISTRIBUTION FUNCTIONS

2.1 The Pair Distribution Function

Let us define with some care one of the fundamental quantities for a pure liquid. Extensions of the formulas can then be made more summarily. Suppose N atoms are confined in a macroscopic volume V. Then $n_0 \equiv N/V$ is the average number density. The position of a spherical atom means the point occupied by its center. Imagine a spherical shell of radius r and thickness δr centered on any atomic position. The volume of the shell is $4\pi r^2 \delta r$, but the number of atomic centers in the shell is not necessarily $4\pi r^2 n_0 \delta r$. For example, if r is approximately the radius of an atom, the number of centers in the shell will be zero because atoms strongly resist overlap. On the other hand, if r is about the atomic diameter, the number could be up to 12 because twelve spheres can touch simultaneously a central one of equal size. In general, therefore, the number of centers in the shell is $4\pi r^2 n_0 g(r) \delta r$, where the factor $g(r)$ allows the actual number density in the shell, i.e., $n_0 g(r)$, to depart from n_0 for the geometrical packing reasons just referred to or, more generally, because of the forces of attraction and repulsion between atoms, which determine the number and distance of neighbors. $g(r)$ is to be an average. It can be looked upon as the time average of its value around any one atom, or the average at any one time of the values around all the atoms. In thermodynamic equilibrium these will be the same. $g(r)$ is called the *pair distribution function* or PDF. The quantity $4\pi r^2 g(r)$ is given the symbol $\rho(r)$ and the name *radial distribution function* or RDF: $n_0\rho(r)\delta r$ is the average number of atoms positioned in the shell of radius r.

Figure 1 shows the PDF of a simple metal. It is zero for $r \leqslant 2.4$ Å, which represents the distance of closest approach of two atoms. At larger distances the curve is *continuous*, which shows that two atoms may be at any distance apart; this is in marked contrast to crystal spacings, which are precise and discontinuous. Strictly, the crystal spacings are between abstract points about which the atoms vibrate; the liquid spacings are between the actual atomic positions. The high peak of $g(r)$ indicates the crowding of neighbors around the atom at the origin. Their number is given by the integral

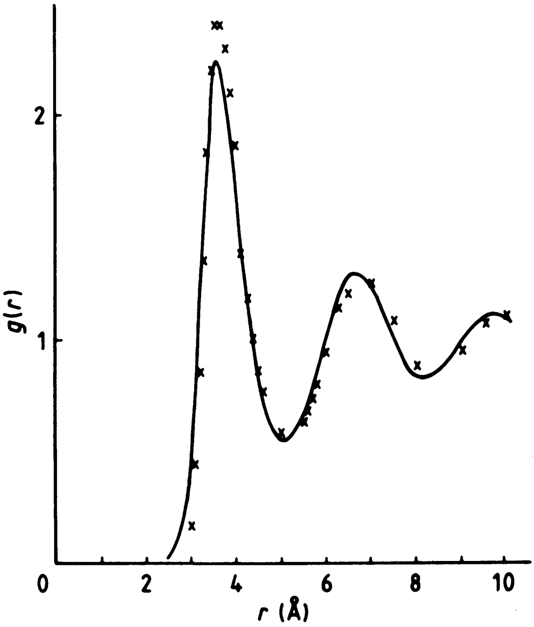

FIG. 1. The pair distribution function $g(r)$ of liquid Na. Full curve, molecular-dynamics calculation with a ϕ_{eff} (see Sec. 8.3). Crosses, experimental points from Waseda (1980).

$$n_1 = n_0 \int_{r_l}^{r_r} \rho(r)dr,$$

where r_l and r_r are the left- and right-hand boundaries of the peak. The values of both r_l and r_r are to some extent arbitrary, particularly if the peak has distortions or shoulders. In principle, n_1, often called the *coordination number*, is not a well-defined integer in a liquid; in a simple liquid it may be about 10. The radial distance to the first peak is the *nearest-neighbor distance*. Other peaks represent successive shells of neighbors, but not many can be seen. For $r > 10$ Å, the curve smooths out and $g(r) \rightarrow 1$.

2.2 Long- and Short-Range Order

By defining $g(r)$ we have taken one step towards saying what structure means in a liquid. It is helpful to think of $n_0 g(r) d^3r$ as the probability of finding an atom positioned at the infinitesimal volume d^3r at a distance r from another atom at the origin. Where $g(r) = 1$, this probability is just $n_0 d^3r$ and the atom may be anywhere with equal probability. Thus at large r randomness has replaced order and there is said to be no *long-range order*, (LRO). Crystals do have LRO, which in a large single crystal extends over macroscopic distances. The first high peak of $g(r)$ reveals a clear departure from randomness, which is called *short-range order* (SRO). Peaks further out indicate *intermediate-range order* (IRO), which is currently much discussed in more complex materials such as glasses (see AMORPHOUS MATERIALS, STRUCTURE OF).

If there is an atom at the origin at time $t = 0$, the probability of another in the volume element d^3r at a distance r is, as we have seen, $n_0 g(r) d^3r$. If we ask the probability of the second one at time $t > 0$, a whole new field is opened up because the motion of the atoms is involved, not just their positions. In thermodynamic equilibrium this is the natural thermal motion associated with the temperature, and, to a first approximation in a simple liquid, it is an irregular oscillation about a center whose position wanders diffusively. We shall see in Sec. 3 that the PDF is explored by elastic scattering. It is *inelastic scattering* that is required to examine the *time-dependent PDF*, which is the generalization of $g(r)$ taking thermal motion into account. The articles on LATTICE VIBRATIONS, STATISTICS OF; PHONONS IN CRYSTAL LATTICES; and VIBRATIONAL STATES IN DISORDERED SYSTEMS deal with this quantity.

2.3 More Distribution Functions

$g(r)$ refers only to the distance between two atoms; it offers no directional information. It is not difficult to conceive of more informative distribution functions. For instance, the probability of finding three atomic positions in a triangle in the volume elements d^3r_1, d^3r_2, and d^3r_3 is written $n_0^3 g^{(3)}(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)d^3r_1d^3r_2d^3r_3$, and $g^{(3)}$ is called the *triplet distribution*. $g^{(4)}$, $g^{(5)}$, etc., can be correspondingly defined. The more of these functions we know, the more completely defined is the structure. In fact, very little is known even about $g^{(3)}$. It is not possible to derive $g^{(3)}$ from $g(r)$; indeed, an infinite number of $g^{(3)}$'s are compatible with the same $g(r)$. Experiments yielding any information about $g^{(3)}$ are few and far between, but theory and computer simulations are used to study it. However, the vast majority of our knowledge of the structure of simple liquids and mixtures is confined to $g(r)$'s and to numbers such as n_1 that can be derived from them.

2.4 Pair Distributions in Binary Systems

Suppose we have a liquid made of atoms of two types, A and B. After the discussion of $g(r)$, the reader will probably grant that there will now be four pair distribution functions to be labeled $g_{AA}(r)$, $g_{BB}(r)$, $g_{AB}(r)$, and $g_{BA}(r)$. The first two measure the probability that two like atoms are at distance r; the last two do the same for unlike atoms. Since the number of A–B and B–A interparticle distances is the same, $g_{AB}=g_{BA}$, and only three of the four g's are independent. The g_{ij}'s are called *partial pair distribution functions*, where i, j can be A or B. From these, *partial RDFs* $\rho_{ij}(r)$ can be defined. To take an example: if there are N_B B-type atoms in the volume V, their number density is $n_B \equiv N_B/V$. The number of B atoms within a distance r_1 of an A atom is

$$n_B \int_0^{r_1} g_{AB}(r)4\pi r^2 dr = n_B \int_0^{r_1} \rho_{AB}(r)dr. \tag{1}$$

If this integral were taken over the first peak, we should have the *partial coordination number* n_{AB}, giving the number of B-type nearest neighbors to an A atom. There are four partial coordination numbers in a binary system. If g_{AB} has a large peak at small r, the $A \leftrightarrow B$ attraction favors SRO with B atoms near A atoms and vice versa. If g_{BB} has a large peak at small r, then B atoms are tending to cluster together.

In principle, triplet and higher distributions can be defined for ternary and more complex liquids. To know all these functions would be to know completely the structure of the system. Such knowledge is scarce enough even for the partial pair distributions of binary liquids, and nothing more complicated will be considered here. Therefore we now move on to consider how $g(r)$ and g_{ij} are measured.

3. DIFFRACTION STUDIES OF PAIR DISTRIBUTION FUNCTIONS

"Measured" is too strong a word; $g(r)$ is inferred by a long train of reasoning from diffraction experiments. It is true that clues to the existence of SRO may be picked up from, *inter alia*, densities, known chemical affinities, x-ray absorption, analogy with crystals of similar composition, nuclear magnetic resonance (see MAGNETIC RESONANCE AND QUADRUPOLE RESONANCE, NUCLEAR; NUCLEAR MAGNETIC RESONANCE, DETERMINATION OF STRUCTURES BY), or the dependence of electrical resistivity or volume on concentration in alloys. But most knowledge comes from measuring the scattered intensity as a function of angle when a collimated monochromatic beam is transmitted or reflected by the sample. If the beam is of x rays, the technique corresponds to x-ray crystallography (*q.v.*), though the lack of crystallinity implies the absence of Bragg peaks. To help avoid confusion we refer first to x-ray diffraction from pure monatomic liquids.

3.1 The Structure Factor

The fundamental diagram is Fig. 2. $\hbar\mathbf{k}_i$ and $\hbar\mathbf{k}_f$ are the momenta of incident and deflected photons scattered elastically by the atoms in the liquid. $\mathbf{k}$ is the wave vector of the radiation. The momentum change $\hbar\mathbf{Q}$ and its relation to the scattering angle θ are

$$\hbar\mathbf{Q}\equiv\hbar(\mathbf{k}_i-\mathbf{k}_f),\quad |\mathbf{Q}|=2k\sin(\theta/2). \tag{2}$$

If there were only one scattering atom, the amplitude and intensity at the distant detector would be proportional to $f(\theta)$ and $|f(\theta)|^2$, respectively, where $f(\theta)$ specifies the probability with which x rays are scattered by one atom in the θ direction and is known and tabulated for all atoms. The scattered waves from all the atoms arrive at the detector with different phases and interfere with one another more or less destructively depending on θ, resulting in an angular intensity variation, $I(\theta)$. The phase differences depend on the atom-to-detector path differences and therefore on the atomic positions, which can be denoted by vectors $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N$. Consequently the function $I(\theta)$ contains information about the positions. This fact underlies the possibility of inferring structural information from diffraction. The formulas are derived with care in numerous references [see, e.g., Waseda (1980); Cusack (1987); Guinier (1963); we shall refer to these again as "the cited texts"]. The fundamental result for the detected intensity is

$$I(Q)=\alpha N|f(\theta)|^2S(Q). \tag{3a}$$

N and f have already been defined and f is characteristic of the element in the sample. $S(Q)$, called the *structure factor*, depends only on the positions of the atoms according to the formula

$$NS(Q)\equiv\left|\sum_{i=1}^{N}\exp(i\mathbf{Q}\cdot\mathbf{R}_i)\right|^2. \tag{3b}$$

Equation (3b) results from summing the amplitudes of all the scattered waves and then squaring to obtain the intensity. It is usual to use Q instead of θ to express the angular dependence; and, in Eq. (3b), $S(Q)$ depends only on the magnitude of Q, not the direction, because of the isotropy of liquids. Note that no units are specified for the intensity here, and the constant of proportionality, α, will depend on how the intensity is measured, e.g., by photon counting rates, and on the sample and equipment geometry and numerous experimental conditions such as absorption of the beam by the sample and its container (essential for liquids).

A great deal of subsidiary experimentation, calibration, corrections, and experience go into establishing a reliable value for α. We refer the reader to the cited texts for these details and will assume that $S(Q)$ can eventually be obtained from laboratory measurements of $I(\theta)$. The observed $S(Q)$ for liquid sodium is shown in Fig. 3, and an overall accuracy of $\sim\pm2\%$ in $S(Q)$ results from good practice in this field. The problem is how to relate $S(Q)$ to the structure.

A crucial point is that the derivation of Eq. (3) assumes that scattering is improbable and that a photon scattered by one atom is not scattered by another. In other words, *multiple scattering* is absent. This is a reasonable assumption for x-rays; but multiple scattering always occurs to some extent and has to be ignored or subtracted in the data processing. This leaves only the distances between any two atoms to produce the path differences that determine $I(\theta)$. The latter therefore contains information only about these distances, i.e., about $g(r)$. Multiple scattering would involve $g^{(3)}$, $g^{(4)}$, etc., but as well as being a minor effect in these experiments, it is hard to relate in a usable way to the higher distribution functions. The problem therefore reduces to relating $S(Q)$ to $g(r)$.

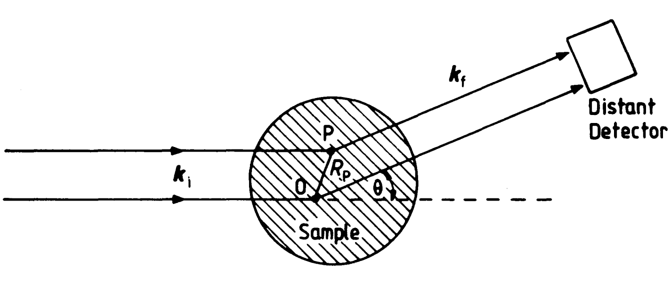

FIG. 2. Definitions of quantities in the coherent scattering calculation. O and P are two atomic positions illustrating the path difference.

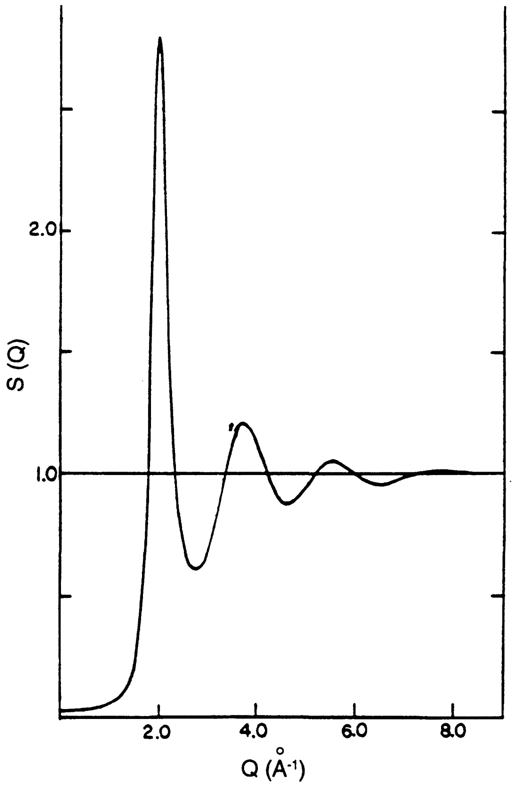

FIG. 3. $S(Q)$ for liquid Na. This curve and that of Fig. 1 mutually transform by Eq. (4).

3.2 Structure Factors and the PDF

$S(Q)$ and $g(r)$ are related by the following Fourier transform equations:

$$g(r)=1+\frac{1}{8\pi^3 n_0}\int_0^\infty [S(Q)-1] \times\frac{\sin Qr}{Qr}\,4\pi Q^2 dQ; \tag{4a}$$

$$S(Q)=1+n_0\int_0^\infty [g(r)-1]\frac{\sin Qr}{Qr}\,4\pi r^2 dr; \tag{4b}$$

$$S(Q)\to 1 \quad \text{as } Q\to\infty; \quad g(r)\to 1 \quad \text{as } r\to\infty. \tag{4c}$$

These are theoretical results demonstrated in the cited texts. Equation (4a) shows that $g(r)$ can be found by an integration over the observed $S(Q)$. Although such calculations are easily done nowadays, the most powerful computer cannot remedy a lack of experimental data, and it is clear that $S(Q)$ will not be measured as $Q\to 0$ (because observation is impeded by the strong unscattered beam at $\theta=0°$) nor as $Q\to\infty$ [because $I(\theta)$ is very small at large angles]. In many x-ray experiments $\sim 0.5 < Q\,(\text{Å}^{-1}) < 12$, but modern sources such as synchrotrons for x rays and spallation sources for neutrons extend Q to 40 Å^{-1} and more. Even so, some extrapolations or approximations in the integration, and the intrinsic errors in $S(Q)$, are unavoidable and lead to uncertainty in $g(r)$.

3.3 Partial Structure Factors in Binary Systems

A-type and B-type atoms will have different scattering amplitudes f_A and f_B. In the sum of scattered amplitudes at the detector, leading to Eq. (3), there will be terms in f_A with phases dependent on the positions of the A atoms, and corresponding terms in f_B. On squaring to get the intensity, terms in f_A^2, f_B^2, and $2f_Af_B$ will appear. As in Eq. (3), these terms will have multiplying factors depending on the A–A, B–B, and A–B distances, respectively. In other words, there will be three *partial structure factors*, S_{ij}. These are related to the $g_{ij}(r)$'s of Sec. 2.4 by relations like Eq. (4). For binary mixtures with concentrations $c_A=N_A/(N_A+N_B)$ and $c_B=N_B/(N_A+N_B)$, the detailed calculations leads to

$$I(Q)=\beta F(Q), \tag{5a}$$

$$\begin{aligned}F(Q)\equiv{}& f_A^2[c_Ac_B+c_A^2S_{AA}(Q)]\\ &+f_B^2[c_Ac_B+c_B^2S_{BB}(Q)]\\ &+2f_Af_B\,c_Ac_B[S_{AB}(Q)-1],\end{aligned} \tag{5b}$$

$$S_{ij}(Q)=1+n_0\int_0^\infty [g_{ij}(r)-1] \times\frac{\sin Qr}{Qr}\,4\pi r^2 dr, \tag{5c}$$

$$g_{ij}(r)=1+\frac{1}{8\pi^3 n_0}\int_0^\infty [S_{ij}(Q)-1] \times\frac{\sin Qr}{Qr}\,4\pi Q^2 dQ, \tag{5d}$$

$$S_{AB}=S_{BA}, \tag{5e}$$

where β is a calibration constant like α in Eq. (3a). This shows that, if experiments could provide the S_{ij} separately, Eq. (5d) would provide the partial pair distributions and coordination numbers.

3.4 Neutron Diffraction and Isotopic Substitution

The S_{ij} were measured separately for the first time in 1966 (Enderby *et al.*, 1966). Since the experiment relied on neutron diffraction we refer to that first.

Equations (3)–(5) apply to any kind of scattered waves if the quantities f, α, and β are reinterpreted and if the data processing is altered to accommodate different kinds of correction. Neutron beams of wavelength suited to structure studies are available at international facilities, and spectrometers are well developed. The x-ray quantities f and $|f|^2$ must be replaced by b and $|b|^2$ where b, the *scattering length*, expresses the probability of neutron scattering by the nucleus. At the wavelengths used, b is independent of θ and is usually well known. We note that $F(Q)$ contains one contribution, called *coherent scattering*, dependent on the structure factors and another, $c_A c_B (f_A - f_B)^2$, called *incoherent scattering*, dependent on the difference between the two f's. Since the system (neutron +nucleus) can exist in two spin states with two different b's, even a single isotope shows incoherent scattering. Moreover, a chemically pure element will usually contain several isotopes with a range of b values. In short, b_A and b_B in the formulas must be understood as averages weighted for both spin states and isotopic mix. Multiple scattering occurs with neutrons, and so does inelastic scattering, which is quite serious for low-Z samples; both must be subtracted from the data. In these and other ways, the data processing for neutrons and x rays differs, but both are aimed at correcting the measured $I(Q)$ to reveal its coherent elastic part, which gives $S(Q)$ and thence $g(r)$.

For binary liquids, separating the three unknown functions S_{ij} requires, for each concentration, three measurements of $F(Q)$, which are independent of each other on account of significantly different ratios of b_A to b_B or of f_A to f_B. Varying the concentration is not an option because this will in general change S_{ij}. Varying b is practicable nowadays because isotopic abundances can sometimes be changed at will in samples of constant chemical nature. Three mixtures $A_x B_{1-x}$, in which either or both A and B vary in isotopic composition, could then give three functions $F(Q)$ solvable as simultaneous equations for S_{AA}, S_{BB}, and S_{AB}. There are other possibilities. Three experiments using x-ray, neutron, and electron scattering might serve the purpose. It is also possible to use x rays alone by exploiting the fact that, for the anomalous scattering close to an x-ray absorption edge, f for a given Q depends quite strongly on wavelength. Thus measurement at three different wavelengths effectively varies the ratio of f_A to f_B.

In practice, none of these methods is without substantial difficulties, ranging from the high cost of isotopes to the enhancement of errors when three functions, each subject to experimental inaccuracy, are combined. The methods have been applied to glasses, aqueous solutions, molten salts, and alloys, but rather few partial structure factors for liquid systems have been thoroughly investigated.

The reader may wonder whether $S_x(Q)$ from x-ray diffraction should be the same as $S_n(Q)$ from neutrons. It is true that x rays are scattered by the electron density distribution and that in metals this is mainly concentrated in the ion cores. $f(Q)$ relates to spherical ions or atoms centered on the nucleus, and if this were the only scattering, the two $S(Q)$'s would be the same. But the free electrons characteristic of metals permeate the entire system including the interstices between the ions, and their density distribution will be distorted by all the electrostatic forces. Insofar as they cause scattering, they make $S_x(Q) \neq S_n(Q)$. The difference is just detectable in experiments designed to find it and can be safely neglected for the present purpose. (See Takeda *et al.*, 1990, for metals and Dore, 1990, for molecular liquids.)

4. INTERMOLECULAR FORCES AND LIQUID-STATE THEORY

Later sections will deal separately with noble-gas liquids, metals, molecular liquids, and molten salts, referring to their special characteristics. This section makes some points relevant to all of them.

The preceding sections explain what structure means in a liquid and outline the way $S(Q)$ and $g(r)$ can be found from diffraction experiments. The PDF and RDF are often referred to loosely as "the structure," but it was emphasized that they relate only to the average radial interparticle distances. Since $S(Q)$ and $g(r)$ are mutually transformable, they contain the same information. A rise in temperature increases motion and disorder, but the repulsive forces still define a nearest-neighbor distance. Heating broadens and lowers the first peak in $g(r)$ but alters its position very little. This shows that thermal expansion reorganizes the SRO and reduces n_1 without changing the nearest-neighbor distance—an important difference from the behavior of crystals.

The theory of liquids aims to derive the macroscopic properties, including the structure, from the intermolecular force. If we assume this to be pairwise and additive—not strictly true but a generally accepted approximation for simple liquids—the potential energy of two molecules i and j may be written

$$\phi_{ij}=\phi_{ij}(\mathbf{r}_{ij};\ \omega_i,\omega_j), \tag{6}$$

where $\mathbf{r}_{ij}$ is the separation of representative points or centers in the two molecules and ω symbolizes the three angles necessary to specify a molecular orientation. For linear molecules, ω reduces to two polar angles, and for spherical atoms or ions ω is irrelevant.

The dependence of ϕ on $\mathbf{r}_{ij}$ contains attractive and repulsive parts. We take for granted that there is a strong short-range repulsion, though this is in fact a subtle and profound property of matter (see, e.g., Lieb, 1976, who discusses why a stone, say, is so big). The repulsion is often approximated by an infinite force, which, without an attractive force, gives us the hard-sphere (HS) model. The structure of a dense random packing of hard spheres has been much investigated by computer simulation, and its $g(r)$ is well known. It is not dissimilar from Fig. 1 and represents approximately the structure of those simple liquids whose particles have undirected attractive forces that simply pack the molecules together as closely as the repulsive forces allow. When they freeze, such liquids generally adopt one of the closer-packed crystal structures such as face-centered cubic, which are a few percent denser than their melts.

A more realistic potential represents the attractive and repulsive parts by power laws, and another much used model, named after Lennard-Jones (LJ), has two parameters characteristic of the species, viz.,

$$\phi(r)=4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12}-\left(\frac{\sigma}{r}\right)^{6}\right]. \tag{7}$$

It is shown in Fig. 4. This serves well for noble gases. Many more complicated potentials have been devised for other liquids and will be exemplified below. From any chosen intermolecular potential there are two routes to the structure: statistical mechanical theory and computer simulation.

4.1 Liquid-State Theory

Statistical mechanical texts show that all thermodynamic properties, including the structure, may be obtained from the partition function, of which the classical form suits all but a few quantum liquids like He at low temperatures. For a liquid of N similar particles at temperature T, acting by a pairwise additive potential $\phi(r)$, some formulas which

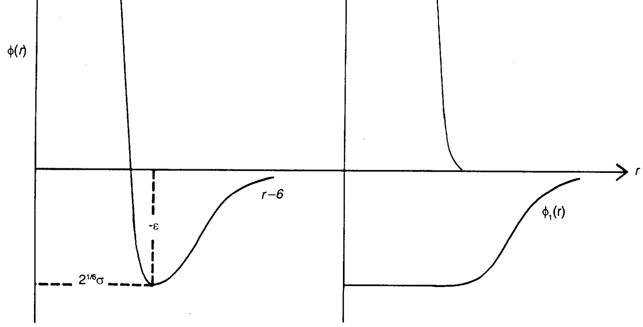

FIG. 4. The Lennard-Jones potential of Eq. (7) and its decomposition according to Eq. (11).

relate $g(r)$ to $\phi(r)$ are as follows. The total potential energy is

$$\Phi=\frac{1}{2}\sum_{i,j}\phi(r_{ij}). \tag{8}$$

Then

$$g(r_{12})=n_0^{-2}Q^{-1}N(N-1)\times\int\exp(-\Phi/k_BT)d^3r_3\cdots d^3r_N, \tag{9}$$

$$Q(V,T)=\int\exp(-\Phi/k_BT)d^3r^{(N)}. \tag{10}$$

Q is called the configuration integral. k_B is Boltzmann's constant and $\mathbf{r}^{(N)}$ stands for the set of N position vectors $\mathbf{r}_1,\ldots,\mathbf{r}_N$ of the particles.

Equation (9) can be generalized by including orientational coordinates and a variety of particle species; but, even without this, its solution is a central problem of liquid-state theory, leading to hierarchies of equations, density expansions, and variation and perturbation methods for which the reader will require one of the treatises on liquid-state theory such as Hansen and McDonald (1986). This will show that there is no inherent mystery in the subject but that the technical complexity is great. The various approximate solutions give very good agreement with observed $g(r)$'s when well chosen $\phi(r)$'s are used. Specimen cases will be outlined below.

4.2 Computational Aids

Liquid studies, like the rest of condensed-matter physics, have benefited from two decades of computer simulation. *Molecular dynamics* (MD) (*q.v.*) is much used and, reduced to its barest outline, it works as follows (see also MONTE CARLO METHODS).

A computer is given the following information: coordinates and velocities for N particles of specified nature in a volume V, often cubic; the interparticle potential, $\phi(r)$; and the boundary conditions—frequently periodic to ensure that N is representative of a much larger number and that if a trajectory takes a particle out at one face, an identical one enters at the opposite face to keep n_0 constant. With use of the classical equations of motion, the computer then

1. calculates the resultant interparticle force acting on each particle at time $t=0$;
2. finds all the new positions and velocities Δt seconds later;
3. repeats for successive Δt's.

When sufficient time has elapsed for the system to be in thermodynamic equilibrium, some property, say the total kinetic energy or $g(r)$, is computed after each of many successive steps and the average taken. Thus the average equilibrium property is found. N may be $\leqslant 10^3$, $\Delta t \sim 10^{-14}$ s, and there will be many hundreds of steps. The method is of great power and versatility and there are many versions and techniques.

A new and already valuable computational aid to the interpretation of scattering data works as follows. Suppose the N particles in the box with periodic boundary conditions are allocated, not $\phi(r)$'s, but neutron scattering lengths b. Starting with an arbitrary configuration, the computer calculates $g(r)$ and Fourier transforms it to $S(Q)$. The next task is to compute a suitable statistical measure, χ^2, of the difference between $S(Q)$(calc.) and $S(Q)$(expt.) in terms of the experimental error in $S(Q)$. One particle is now displaced randomly, subject to a minimum distance of approach, and χ^2 recalculated. The new configuration is accepted in place of the old with probability $p=1$, if $\chi^2(\text{new})<\chi^2(\text{old})$, or $p=\exp[\chi^2(\text{old})-\chi^2(\text{new})]<1$ if $\chi^2(\text{new})>\chi^2(\text{old})$. Repetition of this process makes χ^2 approach a minimum and the configuration for $\chi^2(\text{min})$ is saved. A number of saved configurations could be so generated and are deemed to be statistically independent. They are all structures that reproduce the observed $S(Q)$ within experimental error. The $g(r)$'s and, in the binary case, the $g_{ij}(r)$'s can be extracted and averaged and regarded as the experimental results. Other properties of the structure, even the triplet distribution, are also available for study, and no intermolecular potential has been involved in this. The method is called *reverse Monte Carlo* (RMC) and for its implementation, including a discussion of the definition of χ^2, see, e.g., McGreevy and Pusztai (1988), McGreevy and Pusztai (1990) for molten salts, Sec. 10.2, and Howe (1990) for molecular liquids, Sec. 9.3.

5. LIQUID ARGON AND SIMILAR SYSTEMS

Liquid argon has been much studied because its liquid range is not too difficult to attain experimentally and its interatomic potential is relatively easy to envisage. At large separations the attractive force originates in fluctuations in the electron clouds, which mutually induce electric multipoles in otherwise spherical charge distributions. The dipole-dipole interaction is the strongest of these and varies as r^{-6} as in Eq. (7). At close quarters the resistance of the electron clouds to overlap creates the strong repulsion approximated by the r^{-12} term in Eq. (7) but theoretically more like $\exp(-r/r_0)$, where r_0 is a parameter. The intuitive expectation therefore is that the attractive force, which is spherically symmetrical, simply draws in the neighbors and provides the cohesion while the SRO is the result of the dense random packing of the spheres delineated by the strong repulsions. Crystalline Ar has the densest possible packing, namely face-centered cubic, but on melting near the triple point the order decreases and the density falls by about 14%. This is approximately the same as the corresponding change for hard spheres. It is now widely accepted that the shapes of curves like Figs. 1 and 5 are well approximated by hard-sphere packings for both noble-gas liquids and simple liquid metals. Structure factors have been computed for LJ and HS fluids and are almost identical (see, e.g., Hansen and McDonald, 1986). These general remarks apply to dense liquids; near the liquid-vapor critical point, the relative effect of the attractive forces on the structure increases as the distinction between the liquid and gas phases vanishes.

A good example of the application of the methods of Sec. 3 to liquid Ar is the neutron-scattering experiment of Yarnell *et al.* (1973). Figure 5 shows $g(r)$ obtained experimentally and also points calculated by computer simulation. $\phi(r)$ of LJ form is quite capable of reproducing $g(r)$ very well, and the simulation confirms that the curve is largely determined by the repulsive force.

We must omit general discussion of how to calculate $g(r)$ from $\phi(r)$ since this is a major problem of liquid-state theory. One class of approaches to it uses perturbation theory, and

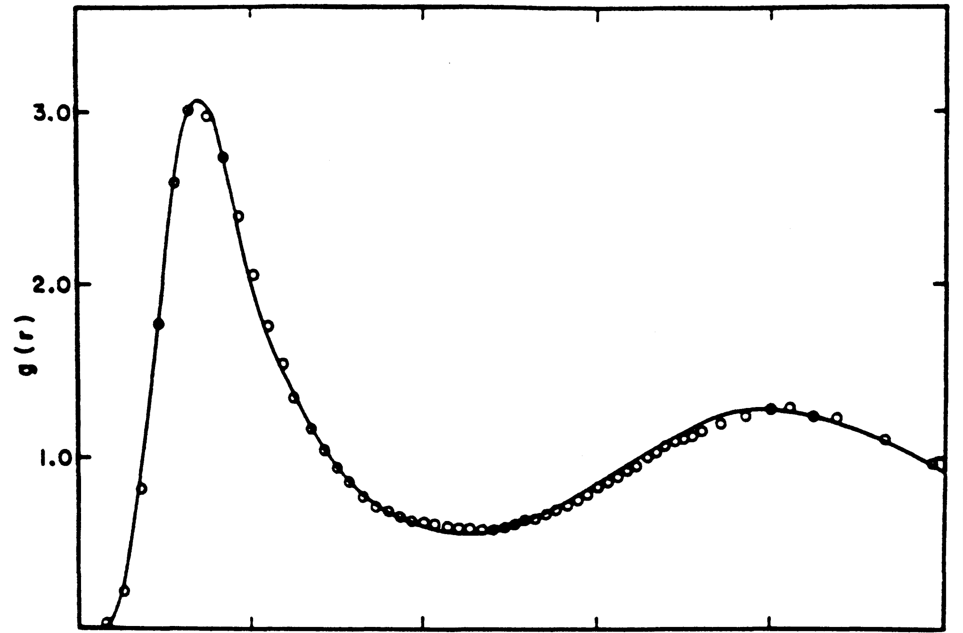

FIG. 5. $g(r)$ for liquid Ar at 85 K, $n_0 = 0.02125$ atom/Å^3. Full curve, experimental results; circles, computer simulation points from Eq. (7).

to give a flavor of this we refer briefly to the influential work of Weeks *et al.* (1971). The idea is to divide $\phi(r)$ into repulsive and attractive parts, ϕ_0 and ϕ_1, respectively, and to treat the hypothetical system represented by the subscript 0 as a reference system of well-known properties. This is then perturbed or corrected by the effect of ϕ_1 to describe the real system. The theory will be convenient and accurate if the perturbation corrections are small and preferably confined to one or two terms of a perturbation expansion. For the LJ potential, $\phi_{\min}=-\epsilon$ and occurs at $r=2^{1/6}\sigma$; so, for the present case (see Fig. 4),

$$\phi_0(r)=\begin{cases}\phi(r)+\epsilon, & r<2^{1/6}\sigma,\\ 0, & r\geqslant 2^{1/6}\sigma;\end{cases}$$
$$\phi_1(r)=\begin{cases}-\epsilon, & r<2^{1/6}\sigma,\\ \phi(r), & r\geqslant 2^{1/6}\sigma.\end{cases} \tag{11}$$

Now define $y_0(r)$ by

$$g_0(r)=y_0(r)\exp[-\phi_0(r)/k_BT].$$

Since $\phi_0=0$ outside the repulsion, $y_0(r)$ is the PDF of the reference system outside the repulsion. But in precisely this region it is plausible that the PDF resembles that of a hard-sphere system, so the next step is to replace y_0 by y_d, where d refers to a system of hard spheres with diameter d; thus

$$g_0(r)\simeq y_d(r)\exp[-\phi_0(r)/k_BT]. \tag{12}$$

Once d is chosen, $g_0(r)$ can be calculated from $\phi_0(r)$ and known properties of the HS system and can then be used as the first approximation to the theoretical $g(r)$—at least at those densities for which the structure reflects the geometry of sphere packing. In practice, Weeks *et al.* showed that, near the triple point, $g_0(r)$ represented excellently the exact result for the LJ potential given by computer simulation and therefore the experimental result too (e.g., Fig. 5). It remains, however, to see how to choose d. This is done by requiring the HS system to have the same equation of state as the reference system. Imposing this condition selects d because there is a connection between the equation of state and $g(r)$ given by statistical mechanics, viz.,

$$k_BT\left(\frac{\partial n_0}{\partial p}\right)_T=1+n_0\int[g(r)-1]d^3r, \tag{13}$$

where p is the pressure. The equation for d is therefore

$$\int[y_d(r)\exp(-\phi_0/k_BT)-1]d^3r$$
$$=\int[y_d(r)\exp(-\phi_d/k_BT)-1]d^3r$$

in which $\phi_d=\phi_{\mathrm{HS}}$ with diameter d.

It might be expected that $g_0(r)$ and the exact $g(r)$ agree less well when the density is lower and repulsive forces cease to dominate the structure. This is true, and corrections involving ϕ_1 become necessary. Furthermore, even at high density, thermodynamic quantities, such as the Helmholtz free energy F, require F_0 (known) to be supplemented by the first-order perturbation term $(Nn_0/2)\times\int\phi_1(r)g_0(r)d^3r$ before good results are obtained. But it is not possible to pursue thermodynamic properties here.

Similar but fewer structure studies exist for other liquified noble gases, and $\phi(r)$ functions carefully constructed from solid and gaseous state data are available for them. For example, $S(Q)$ for Kr over a range of densities was measured by Egelstaff *et al.* (1983) but does not quite agree with the computer simulation using the ϕ's just referred to. Presumably the pairwise interaction assumed for the liquid is in question here. Even if three- or many-body forces are significant, it may be possible to *adjust* the parameters in a two-body expression like Eq. (7) so that it seems sufficient by itself. But the two-body ϕ would then be to some extent fictitious, and so the interplay of the quantum theory of atomic interactions and observation of the liquid-state continues.

6. METALS: SOME OBSERVATIONS

Sections 6–8 will deal with the structures of pure metals and binary alloys. Nevertheless, such examples as there will be of chemical thermodynamics and liquid-state theory, as opposed to metal theory, are in principle applicable to nonmetals also and can be regarded as examples of general approaches.

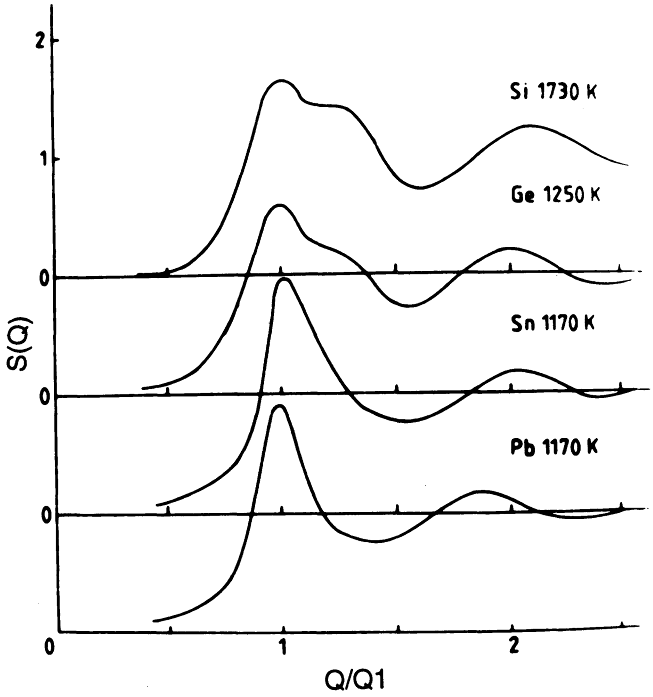

FIG. 6. Rather complicated $S(Q)$ shapes in Si and Ge; distorted first peak in Sn, tending to more close packed hard-sphere structure in Pb (see Sec. 8.4). From Hafner and Kahl (1984).

Figures 1 and 3 are typical of metals in group I of the periodic table for which $9.0 < n_1 < 10.5$. They are the simplest metals and, like noble-gas liquids, have PDFs quite close to that of a close-packed disordered system of hard spheres. The first peaks are symmetrical and the radial distances of the first and second peaks in $g(r)$ are in the ratio ~1.91. Be, Mg, Al, and Pb are like this also. In group-IIB metals, Zn, Cd, and Hg, the high-r and high-Q sides of the first peaks are spread out giving an asymmetrical peak though n_1 is still about 10. Although molten Si and Ge are metallic, not semiconducting like their crystals, n_1 is lower at ~6.6 and the ratio of the first two peak positions is higher at 2.0 to 2.3. These group-IV elements, and Ga and Sn also, have even more distorted peaks with shoulders or subsidiary maxima. Ge is discussed in detail below.

Thus there are noticeable structure trends through the periodic table and some examples are given in Fig. 6. It is plausible to see correspondence between these trends and those in crystalline elements, viz., close-packed metallic structure in groups I to III; distorted structures in group IIB and Ga, but more metallic close packing again in In and Tl; open covalent structure in group IV (Si and Ge) tending to metallic close packing at higher atomic number (Sn and Pb). Fundamental metal theory ought to account for this in both solids and liquids—see Sec. 8 below.

It is not easy to classify binary alloy structures. A random mixture of hard spheres of two different sizes would have the nearest-neighbor A–B distance between the A–A and B–B distances; the first peaks of the S_{ij} would behave similarly. The S_{AA} and S_{BB} would resemble those of pure A and B. The alloy Na-K in Fig. 7 is an example of this, and there are others, e.g., Al-Mg, Pb-Sn. There are cases, however, where the first peak of S_{AB}, but not S_{AA} or S_{BB}, has a well defined prepeak that has been interpreted as evidence of a larger size structural unit probably due to unlike-atom attraction. Figure 8 is an example with strong unlike-pair attractions.

Ordering (a preference for unlike neighbors) or *clustering* (preference for like neighbors), both departures from random mixing that should show up in S_{ij} and g_{ij}, must result from interparticle attractions and repulsions characteristic of the chemical nature of the components. The remarkable properties of some binary alloys at special compositions strongly reinforce this idea.

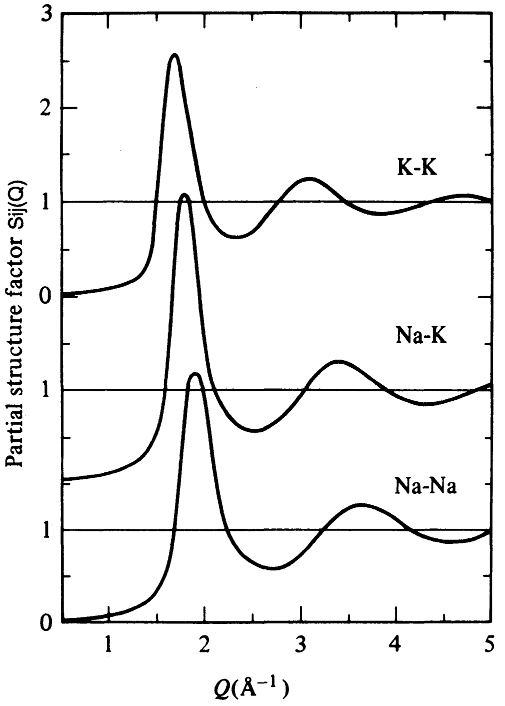

FIG. 7. Partial structure factors in Na-K; from Waseda (1980).

7. CHEMICAL SHORT-RANGE ORDER

There are surprises in the behavior of some liquid alloys that raise questions about the structure. For example, Ga and Hg form a two-phase liquid for some ranges of composition and temperature; liquid $Au_{0.5}Cs_{0.5}$ has very low electrical resistivity and behaves like a molten salt, which suggests that oppositely charged Au and Cs ions have each other as nearest neighbors (Kilic *et al.*, 1990). Anomalies such as very low electrical resistivity are often accompanied by extrema in thermodynamic properties, such as the entropy or volume of mixing, and in magnetic or thermoelectric behavior. Liquid systems well known for this include Li-M where M=Pb,Bi,Tl; Na-Sn; Cs-Pb; Mg-Bi; and Tl-Te. The significant point is that these systems have a special composition (Li_4Pb; Mg_3Bi_2; Tl_2Te, etc.), or possibly two, at which the anomalies appear most intense and that these compositions have stoichiometric proportions which suggest a chemical relation of the elements. There are usually crystalline compounds of the same composition. These liquids are often called *compound-forming alloys*. This leads to the hypothesis that SRO is caused by associations or complexes of chemically determined composition that are most in evidence when the overall concentration has the stoichiometric composition, e.g., Li_4Pb. This is called *chemical SRO*. The local associations are not conceived as stable molecules but as groupings of finite lifetime whose existence affects the $g_{ij}(r)$. This idea emphasizes the distinction between alloys of chemically attracting species and merely random mixtures.

7.1 Chemical Thermodynamic Models

On this basis many workers have tried to fit the observed thermodynamic properties by a chemical SRO model of the system. The fitted parameters enable the concentration of the hypothetical complexes to be calculated, and this knowledge allows discussion of other properties such as electrical conductivity. This is a very indirect and phenomenological way of investigating SRO, and there is considerable latitude in the fitting of numerous parameters. Nevertheless, the approach is an interesting halfway house towards a fundamental *ab initio* theory of structure, which presents profound difficulties (see below). Most workers are convinced that chemical SRO exists in the compound-forming alloys.

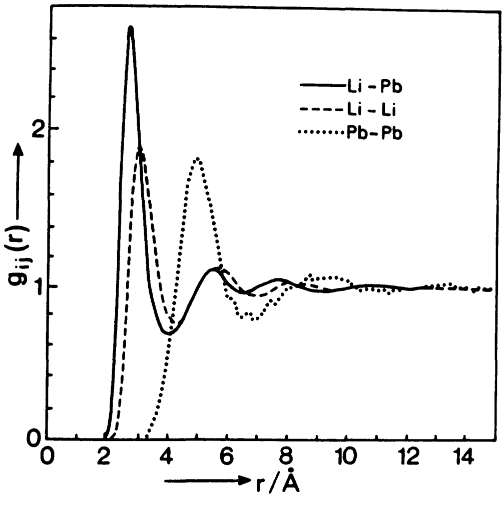

FIG. 8. Partial pair distribution functions in Li_4Pb derived by molecular dynamics. Note the preference for unlike near neighbors shown by the Li-Pb peak at 2.6 Å. From Jacucci *et al.* (1985).

Omitting detailed calculations, which can be found in Bhatia and Hargrove (1974) and Gray *et al.* (1980), the scheme is as follows. Elements A and B are attributed a tendency to form complexes $A_\mu B_\nu$, which have a Gibbs free energy of formation of $g(T)$ per mole at temperature T. The thermodynamic properties of mixing, such as the heat of mixing ΔH, can be derived from an expression for the total Gibbs free energy $G(p,T,c,N)$ of N moles of the binary mixture in volume V. This is

$$G=\sum_{i=1}^{3} n_i G_i^{(0)}+RT\sum_i n_i \ln \varphi_i + \tfrac{1}{2}\sum\sum_{i\neq j} n_i\varphi_j\chi_{ij}, \qquad (14a)$$

$$\sum_{i=1}^{3} n_i G_i^{(0)}=NcG_1^{(0)}+N(1-c)G_2^{(0)}-n_3 g, \qquad (14b)$$

where the symbols have the following meanings: i or $j=1, 2, 3$ denote respectively atoms A and B and the complex $A_\mu B_\nu$; χ_{ij} is an interaction energy between i and j; $\varphi_i \equiv n_i v_i/V$, where n_i is the number of moles and v_i the partial molar volume of species i; c is the fraction of A atoms in the original binary mixture; $G^{(0)}$ is the free energy per mole of a pure constituent. The condition for chemical equilibrium is

$$\left(\frac{\partial G}{\partial n_3}\right)_{T,p,N,c}=0. \qquad (15)$$

A plausible assumption based on chemical knowledge or observed anomalous composition is made for μ and ν; e.g., $\mu=4$, $\nu=1$ for Li_4Pb. Then $\Delta H(c)$, say, is computed from Eqs. (14) and (15) in terms of the parameters g and χ_{ij}, which are evaluated by fitting the observed $\Delta H(c)$. From this the three n_i can be calculated as functions of c. Having done this, Hoshino and Young (1980), for example, assumed that Li_4Pb conducted by virtue of n_1 free electrons from the n_1 Li atoms (valency 1) and $4n_2$ from the n_2 Pb atoms (valency 4) while the complexes contributed no free electrons. The thermodynamic model thus led to a curve of resistivity versus concentration which reproduced quite well the large peak observed at Li_4Pb.

7.2 Chemical SRO and Structure Factors

There is an interesting relation between the chemical SRO of compound-forming alloys and the $S_{ij}(Q)$'s of Sec. 3.3. Mathematically, a set of three new structure factors, formed by linear combinations of the $S_{ij}(Q)$'s, can be achieved in an infinite number of ways. A particular set turns out to have a useful physical interpretation and was first given a statistical mechanical basis by Bhatia and Thornton (1970). They are known as the *Bhatia-Thornton structure factors* and, omitting the argument Q, they are

$$S_{NN}=c_A^2 S_{AA}+c_B^2 S_{BB}+2c_A c_B S_{AB}, \qquad (16a)$$

$$S_{Nc}=c_A c_B[c_A(S_{AA}-S_{AB})-c_B(S_{BB}-S_{AB})], \qquad (16b)$$

$$S_{cc}=c_A c_B[1+c_A c_B(S_{AA}+S_{BB}-2S_{AB})]. \qquad (16c)$$

The subscripts have the following significance. If many small equal volumes of alloy were sampled for the number of atoms in them, and for their concentration, both properties would reveal *fluctuations*, which are a fundamental property of macroscopic systems. Let ΔN and Δc stand for these fluctuations, i.e., for the departures of N and c from their average values. S_{NN} is a measure of the fluctuations in number density. S_{cc} expresses the concentration fluctuations. S_{Nc} represents the correlation between the two. In the limit $Q\rightarrow 0$, usually called the *long-wavelength limit*, $S_{NN}(0)=\langle(\Delta N)^2\rangle/\langle N\rangle$, where N is the number of atoms in the small volume in question and angular brackets indicate a statistical or ensemble average. Thus $S_{NN}(0)$ gives the mean square number density fluctuation. Similarly, $S_{Nc}(0)=\langle\Delta N\Delta c\rangle$ and $S_{cc}(0)=\langle N\rangle\langle(\Delta c)^2\rangle$. The value of these results is enhanced by the fact that for $Q\rightarrow 0$ the Bhatia-Thornton structure factors have thermodynamic expressions. The most widely used of these is

$$S_{cc}(0)=\frac{RT}{(\partial^2 G/\partial c_A^2)_{T,p,N}}=\frac{k_B T(1-c_A)}{(\partial\mu_A/\partial c_A)_{T,p,N}}. \qquad (17)$$

In Eq. (17), k_B is Boltzmann's constant and μ_A is the thermodynamic quantity known as the activity of species A.

There are various thermodynamic and electrochemical ways of measuring $G(c,p,T)$ and $\mu(c,p,T)$. Equation (17) then gives $S_{cc}(0)$ as a function of c for one temperature at a time, usually for $p \sim 0$. The value of this in our present context is that the resulting curve can give strong indications about the SRO. For example, it can be shown that $S_{cc} = c_A c_B$ for a random mixture—often called an ideal solution; consequently, departures from this indicate some degree of ordering. In the extreme case $S_{cc}(0) = 0$, concentration fluctuations must have vanished—which is to say that a well-defined compound exists at the composition for which $S_{cc}(0) = 0$. At the other extreme, $S_{cc}(0) \rightarrow \infty$, the concentration fluctuations are as large as conceivable, i.e., the alloy separates into two phases. Figure 9 shows a hypothetical but not absurd case in which two tendencies to compound formation and one to phase separation are suggested. Measurements of $S_{cc}(0)$ have often stimulated the use of the chemical model of Eq. (14) by suggesting the values of μ and ν for a complex. Figure 10 shows a measured $S_{cc}(0)$ fitted by a model for a compound-forming alloy.

Since the Bhatia-Thornton structure factors are linear combinations of S_{ij} and therefore contain the same structural information, it must be possible to express the quantity $F(Q)$ of Eq. (5b) in terms of them instead of S_{ij}. Thus for neutron scattering we have

$$F(Q) = \langle\langle b \rangle\rangle^2 S_{NN} + (\Delta b)^2 S_{cc} + 2\Delta b \langle\langle b \rangle\rangle S_{Nc}, \tag{18}$$

where $\langle\langle b \rangle\rangle \equiv c_A b_A + c_B b_B$ and $\Delta b \equiv b_A - b_B$. Some elements have isotopes with both $b<0$ and $b>0$. Suppose the isotopes were mixed to give $\langle\langle b \rangle\rangle = 0$. Then one measurement of $F(Q)$ gives $S_{cc}(Q)$ directly since the other terms in Eq. (18) vanish. Likewise, an isotope mixture devised to have $\Delta b = 0$ would give $S_{NN}(Q)$ directly from $F(Q)$. Naturally the possibilities of such manipulations are limited but are being tried out, particularly with metallic glasses, and are potentially powerful ways of measuring the partial structure factors separately.

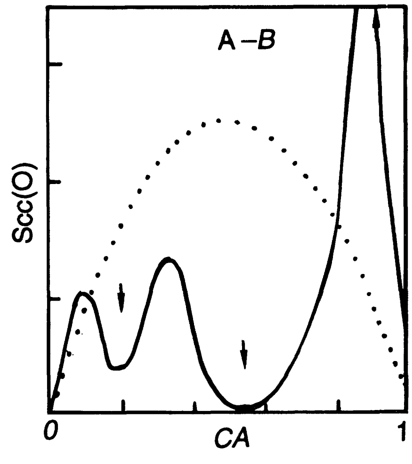

FIG. 9. A hypothetical $S_{cc}(0)$. The arrows indicate possible compositions for complex formation and the peak shows a tendency to phase separation. The dotted curve is for a random mixture or ideal solution for which $S_{cc}(0) = c_A c_B$.

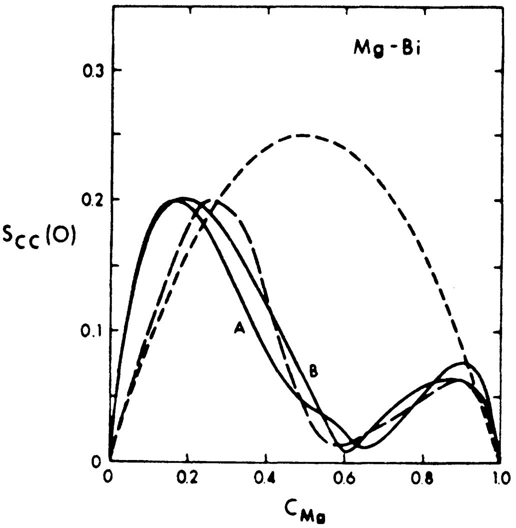

FIG. 10. A measured $S_{cc}(0)$ for Mg-Bi (long dashed curve), with $S_{cc}(0)$ ideal (short dashes) and two alternative model fittings *A* and *B* based on the ideas of Sec. 7.2. Note the tendency to compound formation at Mg_3Bi_2. From Bhatia and Hargrove (1974).

8. STRUCTURE AND THE THEORY OF METALS

It is possible in principle to take the properties of an atom and deduce theoretically the structure of the condensed element. Perhaps the least exacting such exercise is to infer the structure of simple ionic crystals like NaCl from the nature of the Na^+ and Cl^- ions. Liquid metals probably present the most demanding of such tasks for three reasons: the interatomic forces are very complicated in metals; the structures are very disordered in liquids; and these two problems interact. Nevertheless, since about 1970, ways of combining liquid-state theory with metal physics have been devised, and by 1990 there was a reasonable understanding of why metallic elements, crystalline and liquid, have the structures they do.

This required a major effort of theoretical physics and the following sections therefore have the limited objective of setting out the nature of the arguments and giving a few conclusions. The following ingredients are needed for the discussion:

1. the general concept of a metal;
2. the pseudopotentials, the interionic potential, and the energy;
3. perturbation and similar techniques in liquid-state theory;
4. molecular dynamics.

8.1 The General Concept of a Metal

A pure liquid metal is a gas of electrons permeating a fluid of positively charged ions. There are about 10^{28} electrons/m^3, one per ion in Na, two in Ca, and so on according to valency z. They are loosely called *free electrons* and move through the entire system. Free electrons are responsible for the characteristic properties such as conductivity and metallic reflectivity. In a binary alloy the electron density n_e is $n_A z_A + n_B z_B$, unless complicated by compound formation as in Sec. 7.

8.2 Forces and Energy

The ions contain electrons tightly bound to nuclei. The free electrons flow through the ions under the control of two major influences:

1. the electrostatic attraction of the positive ions, and
2. a complicated response to the other electrons bound in the ions.

A powerful development of the wave-mechanical theory of solids made it possible to treat influence **2** mathematically as if it were an electrostatic force like **1** except that it is a repulsion that partially cancels the attraction **1**. The net effect of the two is quite small and is called a *weak pseudopotential.*

An electron passing through an ion may therefore be thought of as deviated by the appropriate pseudopotential. However, the ion does not suddenly exert its influence as the electron crosses a well-defined ionic boundary. As the ions are positively charged and the electron gas is mobile, the ions exert long-range influence on the gas by attracting electrons to themselves. Meanwhile the electrons are also repelling each other and a complex balance of forces sets in whose ultimate effect is to envelop each ion in a free-electron cloud. This cloud, being negative, neutralizes the effect of the positive ion rather completely at large distances, say a few atomic diameters, and less completely at smaller distances. The cloud of electrons piled up near an ion is called the ion's *screening cloud* and the total effect on an approaching electron of the ion and its cloud is called the *screened pseudopotential.*

In summary we may say that, from the point of view of one electron passing through a liquid metal, it encounters the screened pseudopotential of the ions (the nearest ones being felt more strongly) and is scattered by them. The total effect of all the ions, acting though their weak screened pseudopotentials, on all the free electrons is to produce the very electron clouds that do the screening.

Any two positive ions repel one another electrostatically, but in a metal, they have another, indirect, mutual influence because they both exert forces on the same electron gas, e.g., on each other's screening cloud. The latter contribution must depend on the density of the system because this determines the free-electron density, n_e. This contrasts with intermolecular forces in nonmetals, which are independent of density to a good approximation. However, *for a fixed density,* it turns out that the total ion-ion interaction in a metal can be treated just like the forces in nonmetallic liquids. In particular, the ions have *pairwise interaction.* The exact form of the ion-ion potential function will vary from metal to metal and differ from those of nonmetals; it depends on the screened pseudopotential. In general, it will have, as in nonmetals, a short-range repulsive core and a long-range attractive part, which, however, unlike nonmetals, exhibits maxima and minima called *Friedel oscillations* (see Fig. 11). This effective pairwise potential, $\phi_{\text{eff}}(r,n_e)$, controls the structure for any given density. It can be inserted into liquid-state theory to calculate thermodynamic properties and structure.

The expression for the energy of a metal is complicated but its structure is simple, viz.,

$$E(\text{per ion}) = E_{i,\text{ke}} + E_{\text{vol}} + E_{\text{str}}. \qquad (19)$$

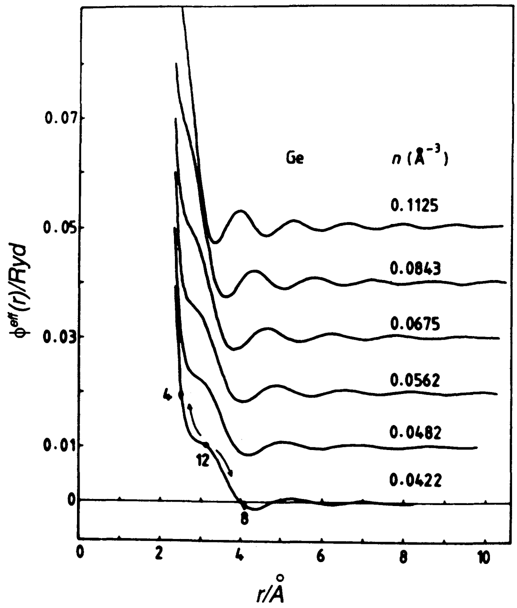

FIG. 11. ϕ_{eff} for Ge at various densities. From Arnold *et al.* (1989).

Here $E_{i,\mathrm{ke}}$ is just the kinetic energy of the ions and is $3k_BT/2$ because their motion, unlike that of the electrons, can be treated classically. E_{vol} is independent of structure and embodies the idea that a free-electron gas has a kinetic energy that rises if the volume falls. E_{str} is due to the mutual forces of the ions. Because ϕ_{eff} is pairwise, E depends on the interparticle distances and therefore contains $g(r)$ or $S(Q)$, but also involves n_e. So far these remarks apply equally to crystals and liquids.

8.3 Structure from $\phi_{\mathrm{eff}}(r)$ in Metals

The theory referred to in Sec. 4 assumes pairwise potentials and can therefore be taken over for liquid metals once pairwise potentials for these have been devised—provided, of course, due respect is paid to the volume-dependent aspects.

To exemplify the great progress of the 1970s (see, e.g., Ashcroft and Stroud, 1978), we first note that all the thermodynamic properties are derivable from the Helmholtz free energy, $F(\text{per atom})=E-TS$ where S is the entropy per atom. Now there is a remarkable theorem, true in classical and quantum statistical mechanics, which works as follows. Suppose the real system, in our case a liquid metal, has its free energy conceptually divided into that of a hypothetical reference system (F_0) plus the rest (F_1) so that $F=F_0+F_1$. Choose the reference system to consist of hard nonattracting spheres. Then the theorem says that $F\leqslant F_0+\langle H_1\rangle_0$, where H_1 is the difference in energy between the real liquid metal and the reference system. $\langle H_1\rangle_0$ means the average of H_1 over all configurations of the reference system. The choice of hard-sphere reference system has advantages for this exposition, viz.,

1. the potential is zero for $r>\sigma$, the hard-sphere diameter, so that H_1 means simply the terms $E_{\mathrm{str}}+E_{\mathrm{vol}}$ while $E_{i,\mathrm{ke}}$ just contributes to F_0;
2. all the properties of the hard-sphere (HS) reference system are well known and tabulated, including $S(Q)$, and depend only on σ and the density;
3. the hard-sphere $S(Q)$ is known to fit some liquid metals well, and so the reference system is structurally near to the real system.

Inserting the entropy term TS into the free energy we have

$$F\leqslant F_0+F_1$$
$$=(3k_{\mathrm{B}}T/2-TS_{\mathrm{HS}})+(E_{\mathrm{vol}}+E_{\mathrm{str}}-TS_{\mathrm{eg}}), \tag{20}$$

where eg stands for electron gas. In calculating E_{str}, the $S(Q)$ for the HS system must be used because the theorem requires F_1 to be averaged over the *reference* system. The final step is a computing exercise to find the value $\sigma=\sigma_{\min}$ that minimizes F_0+F_1, keeping the volume at the observed value for whatever temperature is assumed. The minimum F_0+F_1 should be a good approximation to F and can be used to derive all the thermodynamic properties such as S, specific heat, etc. Furthermore, the hard-sphere $S(Q)$ for $\sigma=\sigma_{\min}$ is the theoretical prediction for the metal. Of course, the possibility of doing this depends on there being expressions for E_{str} that depend on $S(Q)$—which is exactly what the pseudopotential theory of metals provides. The same arguments apply to binary alloys, though the three S_{ij} and the three $(\phi_{ij})_{\mathrm{eff}}$ cause technical complication.

This outlines the use of the Gibbs-Bogoliubov theorem with the hard-sphere reference system. Naturally, improvements proliferated. Better choices of reference system led to more general and reliable results at the expense of denser theory and heavier computing. But the essence remained of dividing ϕ_{eff} into an inner repulsive reference part (but not necessarily infinitely repulsive like hard spheres) and a long-range oscillatory part (like Fig. 11), together with a minimization or optimization procedure to choose the best parameters. For detailed examples see, e.g., Hafner (1977) and Hafner and Kahl (1984). The reader will have noticed that this variation method is similar in spirit to the perturbation approach of Sec. 5.

The concept of $\phi_{\text{eff}}(r)$ in metals, *for a given density*, makes molecular dynamics readily available for liquid metals. $g(r)$ and $S(Q)$ are therefore available in three ways: by diffraction experiments; by MD using $\phi_{\text{eff}}(r)$; and from liquid-state theory using $\phi_{\text{eff}}(r)$. An example comparing all three will be given below.

The problem is that $\phi_{\text{eff}}(r,n_e)$ depends on the free-electron states that are the solutions of Schrödinger's equation of motion. But these states depend on the structure, which in turn depends on ϕ_{eff}. In short, the structure and the electronic states ought to be solved in a self-consistent way. Procedures have been tried that avoid the excessive computation needed for a truly self-consistent calculation. For example, ϕ_{eff} is first obtained from a simple form of the pseudopotential theory that does not need details of the electron states. Two branches are now possible: use ϕ_{eff} with one of the forms of liquid-state theory to derive $g(r)$(theor.); or use ϕ_{eff} and MD to find $g(r)$(comp.). Either or both of these $g(r)$'s may now be compared with $g(r)$(expt.). Assuming agreement is encouraging, use the now known structure from MD to recalculate or improve knowledge of the electron states. This is a formidable program even for simple metals [see Hafner (1990)]. For transition metals (Fe, Ni, etc.) with partially occupied 3*d* shells of electrons, there is the additional rather intractable complication of incorporating the not wholly free *d* electrons into the expression for ϕ_{eff}. [See, e.g., Hausleiter and Hafner (1990) for an approach to this problem for the glassy transition metal alloy $Ni_{33}Y_{67}$.]

8.4 A Successful Example

Figure 11 shows $\phi_{\text{eff}}(r)$ for various densities of Ge calculated from the ideas of Sec. 8.2. The features are

1. a steep repulsive screened ion core—though not infinitely steep as in hard spheres;
2. an intermediate- and long-range oscillating part characteristic of metals and depending on the density of the electron gas;
3. an evolving relation of these as the changing density alters the position of the oscillations and spreads out the core.

If liquid Ge were controlled by the repulsive core acting alone, its structure would resemble a hard-sphere packing with $\sigma \sim 2.72$ Å$=r_1$ say, a second peak at $\sim 1.9r_1$, and $n_1 \sim 10$ to 11. But the density near the melting point is about 0.045 atom/Å^3, and Fig. 11 shows that ϕ_{eff} is distinctly structured. The observed peaks in $g(r)$ are at $1.6r_1$ and $2.25r_1$, which coincide with minima in the oscillatory part of ϕ_{eff}. Thus the atoms reduce overall energy by adjusting their positions in response to the superimposed demands of geometric sphere packing (with a periodicity fixed by σ) and an oscillatory tail (with wavelength $\lambda \propto n_e^{-1/3}$). The conflicting effects of the two different periodicities cause the peaks in $g(r)$ to peter out after ~7 Å. The atoms in the oscillatory troughs reduce what would have been the high peak of close-packed neighbors at $r_1 \sim 2.72$ Å and thus lower n_1 to the observed value of ~6.5. Using the ϕ_{eff} of Fig. 11, $g(r)$ and $S(Q)$ can be calculated by MD or liquid-state theory and compared with experimental results. Figures 12 and 13 show that

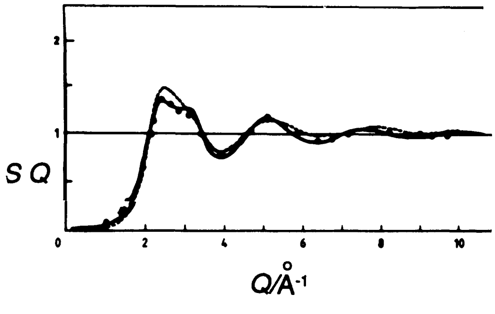

FIG. 12. $S(Q)$ for liquid Ge at ~1300 °C showing agreement between MD (full curve) and neutron diffraction experiments (dotted curve and circles). From Arnold *et al.* (1989).

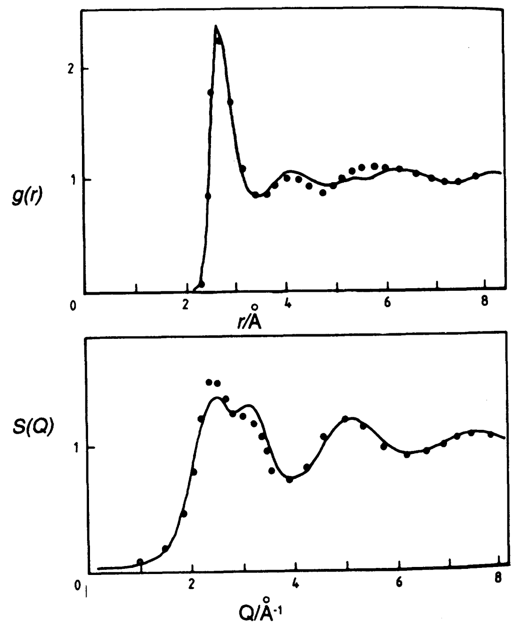

FIG. 13. *g*(*r*) and *S*(*Q*) in liquid Ge showing good but not perfect agreement between MD (full curve) and liquid-state theory (points). From Arnold *et al.* (1989).

these three approaches are beginning to agree well. [For the full theory, see Arnold *et al.* (1989).]

Figure 11 also shows that if the core radius and n_e are appropriate (e.g., the highest curve), the geometrical packing at $r \sim \sigma$ is encouraged rather than reduced by the position of the oscillatory minima: the two periodicities (σ and λ) are compatible, and structure resembling hard-sphere packing results. Calculations with suitable pseudopotentials and densities for the group I and IIA metals of the periodic table show that this is just what happens, leading to $g(r)$'s like Fig. 1. In this way detailed examination, element by element, can rationalize the trends through the periodic table referred to in Sec. 6 [see Hafner and Kahl, 1984). Both MD and liquid-state theory can be extended to alloys with the increasing complication due to three interionic potentials all dependent on the electron density, which is itself a function of composition. If chemical SRO in compound-forming alloys is due to ionic or covalent bonding forces becoming important at special compositions, then additional hypotheses for the forms of these forces are required. (For the use of exponentially screened ionic forces and application of MD to Li_4Pb, see, e.g., Jacucci *et al.*, 1985.)

9. MOLECULAR LIQUIDS

Even the simplest diatomic molecule, say N_2 or CO, involves Eq. (6) and raises the question whether the PDF is affected by the relative orientation of the molecules or whether, on the other hand, liquids of stable molecules approximate to hard-sphere packings to the extent that argon does.

Although homonuclear diatomic molecules and other special cases will not be dipoles, molecules in general will have electric dipole and quadrupole moments. The electric force fields will be nonspherical on this account and can be attractive or repulsive depending on orientation. The short-range repulsive force may also have an angular dependence. In special cases there can be a nonspherical magnetic force as in O_2. Much of the somewhat limited structural study of molecular

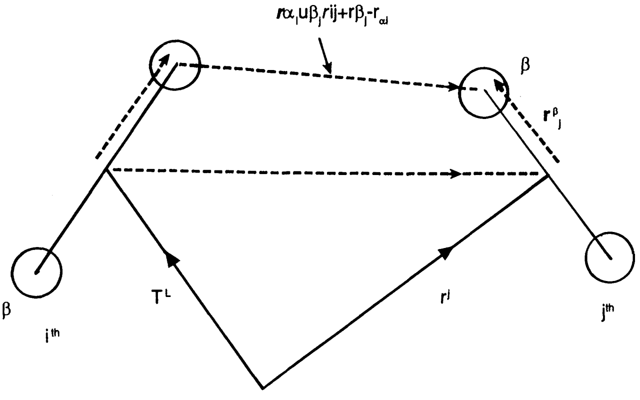

FIG. 14. The angles and distances between the *i*th and *j*th diatomic molecules composed of atoms of types α and β.

liquids has addressed the question of the orientation dependence of the pair distribution $g(\mathbf{r}_{ij}; \omega_i, \omega_j)$. This is a complicated quantity, and Fig. 14 shows the angles and distances relevant to neighboring diatomic molecules to which this account will largely be restricted. Even if the three $g_{ij}(r)$ for the two kinds of atoms in, say, CCl_4 were fully determined by the methods of Sec. 3.4, they would not contain enough information to give $g(\mathbf{r}_{ij}; \omega_i, \omega_j)$ uniquely, and a single diffraction experiment is even less informative.

The interaction potential can be modeled in various ways. For example, a simple and much used suggestion is

$$\phi_{ij}(\mathbf{r}_{ij}; \omega_i,\omega_j) = \sum_{\alpha,\beta} \phi_{\alpha\beta}(\mathbf{r}_{\alpha_i \beta_j}) + \text{electric multipole terms.} \tag{21}$$

Here $r_{\alpha_i \beta_j}$ is the intermolecular distance between atoms of types α and β, and $\phi_{\alpha\beta}$ could be given the LJ form of Eq. (7). In a diatomic molecule the summation would have four terms. For comparison we quote without detailed discussion a potential used by Rodger *et al.* (1988) for Cl_2, viz.,

$$\phi(r_{ij}) = K \exp\{-\alpha[r-\sigma_0-\sigma(\Omega)]\} + C[r+\xi_0-\sigma(\Omega)]^{-6}+E_s, \tag{22}$$

where E_s is a dipole plus quadrupole term on each atom chosen to give the correct overall quadrupole for the molecule; Ω is the relative orientation of the two molecules, and $\sigma(\Omega)$ is expressed in terms of several spherical harmonic functions and contains a number of parameters which, with the other parameters in the expression, were chosen to give agreement with known properties of chlorine such as the crystal structure. This model implies orientation dependence of both attractive and repulsive forces and shows that a simple molecule can demand a quite complex potential. It is, of course, intuitively reasonable that a diatomic molecule should be dumbbell shaped, and at close quarters some relative orientations should favor closer approaches than others.

9.1 Structure Factors of Molecular Liquids

As in Sec. 3.3, the structure factor of a system of two species of atom, say CO or $GeBr_4$, is found by summing the scattered amplitudes and squaring to get the intensity. For a molecular assembly it is helpful to separate those contributions of interfering wavelets originating *within* one molecule from those coming from *different* molecules. $S_M(Q)$, now given a subscript to indicate a molecular liquid, may then be written

$$S_M(Q) = f_1(Q) + D_M(Q), \tag{23a}$$

where

$$f_1(Q) = \left(\sum_\alpha b_\alpha\right)^{-2} \left\langle \left| \sum_\alpha b_\alpha \exp(i\mathbf{Q}\cdot\mathbf{r}_\alpha)\right|^2 \right\rangle \tag{23b}$$

and

$$D_M(Q) = N_M^{-1} \left(\sum_\alpha b_\alpha\right)^{-2} \left\langle \sum_{i\neq j} \exp(i\mathbf{Q}\cdot\mathbf{r}_{ij}) \times \sum_{\alpha_i,\beta_j} b_\alpha b_\beta \exp[i\mathbf{Q}\cdot(\mathbf{r}_{\beta_j}-\mathbf{r}_{\alpha_i})] \right\rangle. \tag{23c}$$

In these formulas N_M is the number of molecules and α or β can be either A or B for molecules of two different kinds of atom. $f_1(Q)$, called the molecular form factor, is often calculable from the known molecular structure or available from electron diffraction studies of the gas phase. If so, a measurement of $S_M(Q)$ yields $D_M(Q)$, which includes the effect, if any, of orientation. $f_1(Q)$ can be refined by an additional factor $\exp(-\gamma_{\alpha\beta}^2 Q^2/2)$, where $\gamma_{\alpha\beta}$ is the rms thermal-vibration amplitude of the bond length, but this is often put equal to zero, thus treating the molecule as rigid with respect to internal motion; by hypothesis rigidity under intermolecular forces is already assumed. $D_M(Q)$ takes on a special form if there is no correlation of orientation at any r_{ij}. In that case, r_{ij}, r_{α_i}, and r_{β_j} are all statistically independent, and from Eq. (23c) it follows that

$$D_M(Q) = f_{2u}(Q) N_M^{-1} \left\langle \sum_{i\neq j} \exp(i\mathbf{Q}\cdot\mathbf{r}_{ij}) \right\rangle, \tag{24a}$$

where

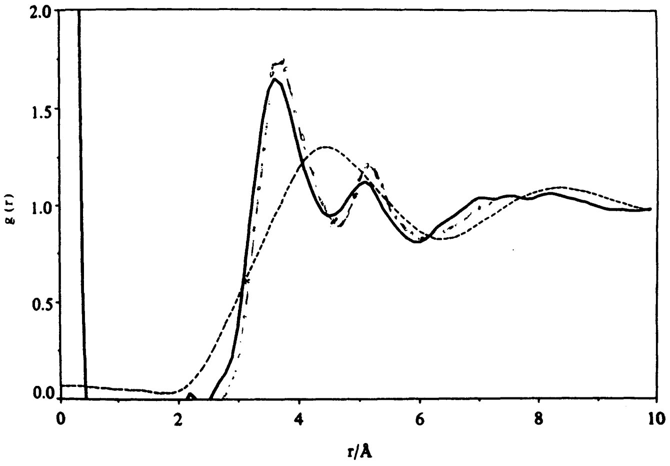

FIG. 15. Intermolecular $g(r)$ in liquid chlorine. Full line, from neutron diffraction; dashed line, calculated for an uncorrelated model; dot-dashed line, MD calculation (see text).

$$f_{2u}(Q) \equiv \left| \sum_{\alpha} b_{\alpha} \right|^{-2} \left| \sum_{\alpha} b_{\alpha} \langle \exp(i\mathbf{Q} \cdot \mathbf{r}_{\alpha}) \rangle \right|^{2} . \tag{24b}$$

Like $f_1(Q)$, $f_{2u}(Q)$ is calculable from the molecular structure: e.g., for a homonuclear diatomic molecule it is $\sin^2(Qd/2)/(Qd/2)^2$, where d is the bond length. Thus, measuring $D_M(Q)$ leads to the quantity

$$N_M^{-1} \left\langle \sum_{i \neq j} \exp(i\mathbf{Q} \cdot \mathbf{r}_{ij}) \right\rangle .$$

By direct analogy with Eq. (3b), this expression could be called $S_c(Q)-1$, where the subscript indicates a structure factor for molecular centers and 1 is subtracted because the $i=j$ term is omitted from the summation. For uncorrelated orientation, therefore,

$$D_M(Q) = f_{2u}(Q)\,[S_c(Q)-1]. \tag{25}$$

These equations have been written with neutron-scattering lengths. Similar formulas have been used for x rays, though it is a dubious approximation to treat the x-ray scattering factors for molecules as simple summations of the effect of the constituent atoms because the bonding will distort the electron distribution somewhat.

9.2 A Representative Experiment

Recent practice in this field is exemplified in a neutron experiment by Bellissent-Funel *et al.* (1990) on Cl_2. For values of $Q \leqslant 16$ Å^{-1}, at temperatures of 202 and 286 K, $S_M(Q)$ was measured and $D_M(Q)$ obtained by Eq. (23). By a Fourier transformation of $D_M(Q)$ analogous to Eq. (4a), a $g(r)$ was found, which, however, excludes the intramolecular Cl-to-Cl distance by virtue of the definition, Eq. (23c). Suppose that S_c is now calculated for a hypothetical set of molecular centers with no orientation correlation. It could then be inserted, together with f_{2u} for Cl_2, into Eq. (25) and an uncorrelated D_M calculated. The hypothetical $S_c(Q)$ was computed by MD for a monatomic liquid with an LJ potential [Eq. (7)] in thermodynamic states corresponding to those of the Cl_2 samples. The result shows that the observed $D_M(Q)$ differs markedly, having two peaks where the uncorrelated curve has one. Indeed the two peaks are thought to be a hallmark of correlated orientation in diatomic molecules. Figure 15 shows the corresponding $g(r)$'s and also one calculated by MD from the potential in Eq. (22). This potential, which builds orientation correlation into the interaction, gives a rather good fit. Furthermore, the ratio r_1/r_2 of the two peak positions in $g(r)$ is ~ 1.4, much lower than the ~ 1.95 characteristic of uncorrelated diatomic, and also monatomic, liquids. The peaks may therefore be interpreted as the effect of orientation modifying the first shell, not as two separate shells of neighbors. From these and other related arguments, it seems clear that orientational correlation exists in liquid Cl_2, but the conclusion remains rather qualitative. It is not obvious which correlations are preferred: e.g., do neighboring molecules tend to be parallel or antiparallel, or in perpendicular formation?

9.3 Some General Remarks about Structure in Molecular Liquids

It should be evident from the preceding paragraphs that even homonuclear diatomic molecules present substantial difficulties in the interpretation of data and the formulation of potentials. Hydrogen and deuterium halides evade our criteria of simplicity because hydrogen bonding plays a strong role in them; it cannot be said that their structure is well understood. Liquid N_2, CO, O_2, and the halogens have been studied by methods comparable with the Cl_2 experiment just referred to, and it appears that whereas definite orientation correlation is detectable in Cl_2, Br_2, and, especially, I_2, there is little or none in N_2, CO, and F_2. A detailed review is presented by Andreani *et al.* (1991). As indicated for Cl_2, MD simulations with comprehensive potentials are becoming available, and the understanding of molecular liquids is bound to grow as MD and experiments interact in future. In this connection an interesting development is the application of RMC (see Sec. 4.2) to halogen liquids by Howe (1990). From experimental data available up to about 1988 and RMC molecular distributions that can reproduce them, it appears that orientation correlation in the halogens persists little further than the first shell of neighbors.

Although this account has emphasized diatomic molecules for reasons of space and simplicity, the MD techniques are powerful enough to handle more complex cases, and diffraction experiments can likewise be involved with advantage, especially if isotopic substitution can be achieved. Thus there have been studies of linear triatomic molecules such as CS_2 (e.g., Tildesley and Madden, 1981), the tetrahedral molecule CCl_4 (e.g., McDonald *et al.*, 1982), and even mixtures of the two (Mittag *et al.*, 1989). In these cases the intermolecular potential had the form of Eq. (21) with LJ $\phi_{\alpha\beta}$'s and no multipole term. Thus between two CS_2 molecules there will be four S-S, four S-C, and one C-C interactions. It would be fair to say that in these cases the S_c's and $g(r)$'s from neutron and x-ray experiments, including some with isotopic substitution, are well reproduced by the computer simulations but that comparison is sometimes made difficult by discrepancies between one experiment and another and by inaccuracies in the computations. It is a measure of the difficulty of interpreting the data that firm statements about the orientation correlations cannot yet be confidently made.

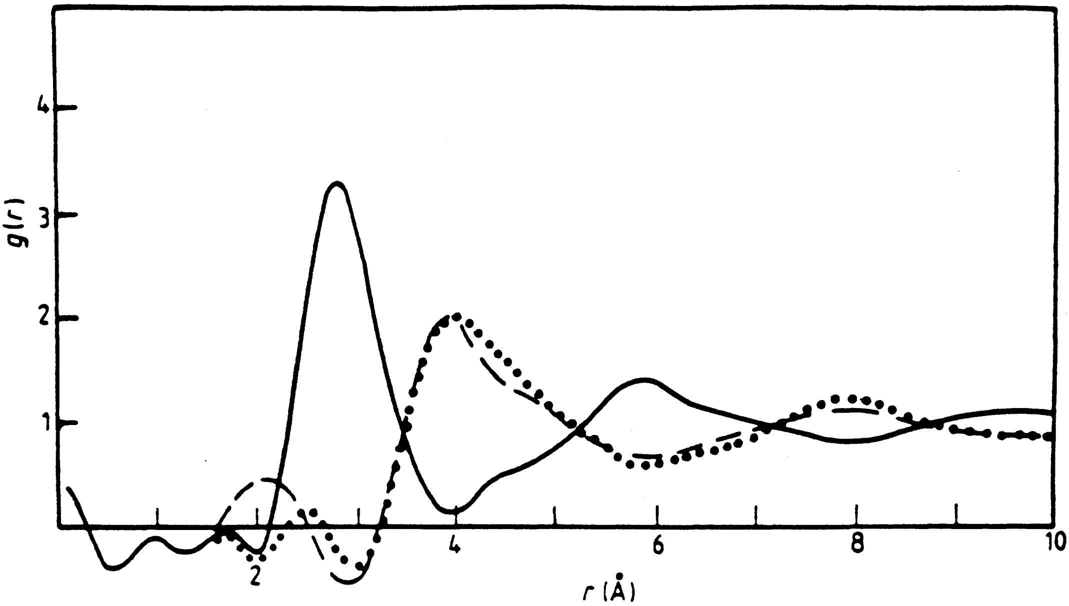

FIG. 16. Partial pair distributions in liquid NaCl; from Biggin and Enderby (1982). Full curve, g_{NaCl}; broken curve, g_{NaNa}; dotted curve, g_{ClCl}.

10. MOLTEN SALTS

The first crystalline structures determined by x-ray crystallography were those of alkali halides, and nowadays it seems natural that they should be interpenetrating lattices of positive and negative ions such that each ion has nearest neighbors of opposite sign. There seems no obvious reason why melting should alter this general picture since it does not radically change relevant properties, e.g., ionic conductivity persists. But it was not until 1975 that this was explored directly by the isotopic substitution method of Sec. 3.4 (Edwards *et al.*, 1975; Derrien and Dupuy, 1975). As in previous sections, the accumulated work in this field will be illustrated by representative examples.

10.1 Halides

Enderby and co-workers (see, e.g., Biggin and Enderby, 1982) changed the isotopic state of Cl in NaCl and obtained the three partial PDFs in Fig. 16. Using + and − to indicate cations (metal) and anions (Cl), we see that g_{++} and g_{--} are very similar with a major peak at ~4 Å; that the *nearest*-neighbor peak is in g_{+-} at 2.72 Å; and that for $r \geqslant 4$ Å, the phases of the curves conspire to cancel the charge almost completely. The similarity in g_{++} and g_{--} results from a corresponding similarity in S_{++} and S_{--}, and the first peaks in these structure factors coincide with a sharp dip in S_{+-} (see Fig. 17). There is, however, a small intrusion of like species into the first shell of unlike neighbors. The coordination number n_1, now more explicitly called n_{+-}, is 3.9. Thus the overall picture is indeed

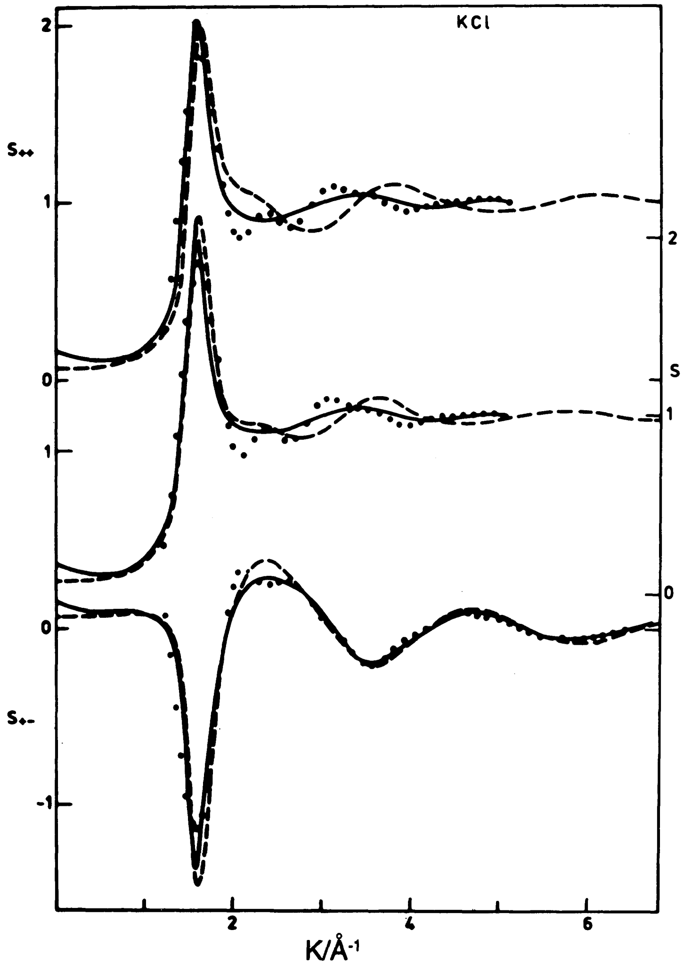

FIG. 17. $S_{ij}(Q)$ in liquid KCl. Full curve, HNC theory (see text); points, MD calculation; dashed curve, a less satisfactory theoretical result discussed by the authors. From Abernethy *et al.* (1981).

of ions surrounded by about four neighbors of opposite sign and with a next shell of mostly similar sign. However, this is only SRO, and the disordering effect of melting means that interparticle distances $\gtrsim 6$ Å are equally likely to be between like or unlike charges.

The clear interdependence of the g_{ij} in molten salts is in marked contrast to the behavior of these quantities in binary alloys. The explanation lies in the fact that the concentrations of positive and negative species in ionic melts are not independent because electrical neutrality must prevail. In binary alloys neutrality is preserved by the free electrons and the concentrations of the species can vary; there is then nothing to secure any unique relation between the peaks and troughs in g_{ij} or S_{ij}.

Group-II halides—$ZnCl_2$, $CaCl_2$, $SrCl_2$, and $BaCl_2$, etc.—are somewhat different (see, e.g., McGreevy and Mitchell, 1982). As the cation size increases from Zn to Ba there are trends of behavior believed to be associated with this size change. Let us assume a Cl^- ion at the origin; then, from Zn to Ba, while the first peak of g_{--} remains at about the same place (3.7 Å) and yields a roughly constant coordination number (3.7 to 3.8), it diminishes in height and spreads out. At the same time the penetration of Cl^- ions into the first shell of cations increases, as shown in Fig. 18. In $BaCl_2$, unlike NaCl, g_{++} and g_{--} are quite different. This is partly due to the different size of charge on Ba^{2+} and Cl^- but is also thought to be a consequence of the greater mobility of the anion. With the smallest cation, Zn^{2+}, there is a well-defined peak in g_{+-}, implying $\sim 4.3 \pm 0.3$ Cl^- ions around each Zn^{2+} ion, and the sharpness of this peak decreases progressively from Zn to Ba. These experiments yield no evidence of special Zn–Cl distances to be associated with complexes of substantial lifetime, such as Zn^+Cl or $ZnCl_3^-$ (compare Sec. 7).

10.2 Computer Simulation and Theory

The foregoing shows some of the inferences that can be drawn from neutron scattering with the isotopic method. Deriving them theoretically requires appropriate interionic potentials. As with the other types of liquid, there are model potentials of various degrees of complexity, but, for molten salts, they will be dominated by the Coulomb forces. These account for the bulk of the cohesive energy and cause the SRO of the charge distribution. A much used potential is

$$\phi_{ij}(r)=A\,\frac{z_iz_je^2}{r}+\beta\exp[-\alpha_{ij}(\sigma_{ij}-r)]-\frac{C_{ij}}{r^6}-\frac{D_{ij}}{r^8}, \qquad (26)$$

in which z_ie and z_je are the ionic charges giving electrostatic attraction or repulsion $\propto r^{-2}$; the second term is the strong repulsion

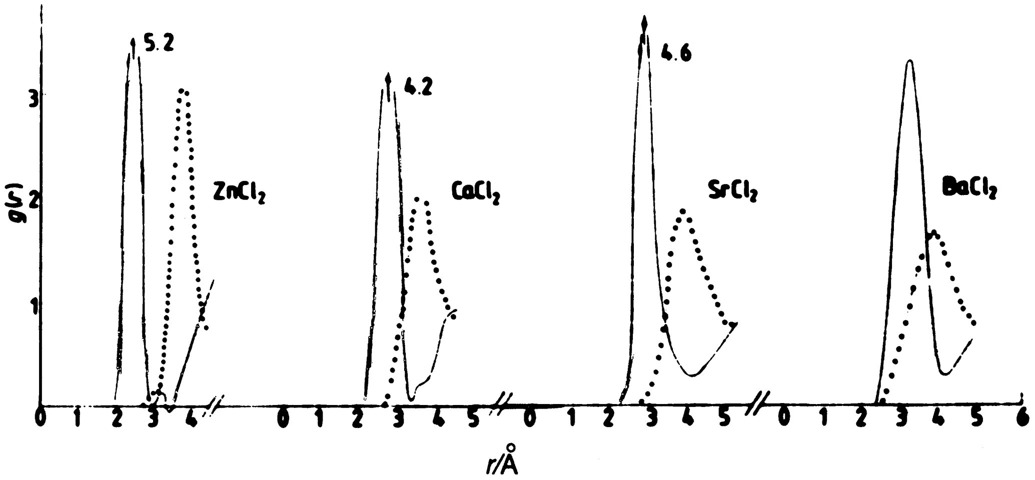

FIG. 18. Development of the penetration of like ions into the first shell of unlike neighbors as cation size increases, according to Biggin and Enderby (1981). The $SrCl_2$ curve is from molecular simulation, the others from neutron scattering. Full line, $g_{+-}(r)$; dotted line, $g_{--}(r)$.

due to overlap (an alternative to the r^{-12} term in the LJ model), and the remaining terms are the induced multipole terms (see Sec. 5) giving interactions of longer range than the repulsion but much shorter than the Coulomb field. A depends on the units, and the other symbols are parameters usually fixed by making ϕ compatible with the properties of the crystalline phases (see, e.g., Sangster and Dixon, 1976, for a detailed discussion of ϕ). With this potential, MD calculations of $g_{ij}(r)$ can be made, and the outcome is satisfactory for the alkali halides but less so for the group-II halides, suggesting that ϕ needs improvement.

Indications were given above of the perturbation method for $g(r)$ in liquid Ar and the Gibbs-Bogoliubov variation method in metals and alloys. For ionic liquids a third of the very numerous approximate methods in the theory of liquids will be invoked. The reader will appreciate that that justice can hardly be done to any of these methods short of a systematic exposition of liquid theory. What is needed is a solution of Eq. (9). Discussion is often facilitated by introducing two more correlation functions: $h(r)$, the *pair correlation function,* and $c(r)$, *the direct,* or *Ornstein-Zernike, correlation function.* These are defined for liquids by

$$h(r) \equiv g(r) - 1, \tag{27}$$

$$h(r_{12}) = c(r_{12}) + n_0 \int c(r_{13}) h(r_{23}) d^3 r_3. \tag{28}$$

$c(r_{12})$ has the effect of splitting the positional correlation $h(r_{12})$ of atoms 1 and 2 into a part depending directly on their positions and an indirect part depending through the integral on the correlated positions of all the other atoms. The range of $c(r)$ is much shorter than that of $h(r)$ or $g(r)$ and is comparable with that of $\phi(r)$. A considerable chain of reasoning, with approximations (see, e.g., Hansen and McDonald, 1986), then introduces another relation between $h(r)$, $c(r)$ and $\phi(r)$, viz.,

$$c(r) \simeq h(r) - \ln g(r) - \phi(r)/k_B T. \tag{29}$$

If this is inserted into Eq. (28), there results

$$\ln g(r_{12}) + \frac{\phi(r_{12})}{k_B T} = n_0 \int \left(h(r_{13}) - \ln g(r_{13}) - \frac{\phi(r_{13})}{k_B T} \right) h(r_{23}) d^3 r_3. \tag{30}$$

This is an integral equation soluble numerically for $g(r)$ once $\phi(r)$ is given. Equations of similar form hold for the $g_{ij}(r)$ in binary systems. This is known as the *hypernetted chain (HNC) approximation.* Unlike other approximate theories with which it is often compared in the literature, it works quite well for liquids with long-range Coulomb interactions. The $g_{ij}(r)$ and $S_{ij}(Q)$ for all the molten alkali halides have been computed by HNC and by MD, using the ϕ of Eq. (26), and the agreement, exemplified in Fig. 17, is good (Abernethy *et al.*, 1981).

With the help of the measured $S_{ij}(Q)$ and the RMC method (Sec. 4.2), it is possible to extend insight into molten salt structures because the RMC results are three-dimensional arrays consistent with the isotopic substitution experiments but not dependent on any assumed potential. For example, the first shell of neighbors can be analyzed by means of a quantity $C_{ij}(n)$, which is the fraction of ions of type i that have n ions of type j in their nearest-neighbor shell. $C_{ij}(n)$ can be extracted from the RMC structure. It turns out, for example, that ~ 0.4 of Na^+ ions have five Cl^- ions in the first shell in liquid NaCl; the corresponding numbers in CsCl are ~ 0.38 and ~ 6. Other Na^+ ions may have, though with lower probability, three, four, six, or seven Cl^- neighbors, giving an overall coordination number $n_{+-} = \Sigma n C_{+-}(n) = 4.85$. This emphasizes how disordered the shell is in the liquid compared with the well-defined one in the crystal and how limited the information is in a single coordination number. We also recall from Sec. 2.1 that particular n_{+-} values depend on the definition adopted by their authors. Another limitation partly removed by the RMC structures, as also by MD and MC, is the absence of angular information. If two lines are drawn from one Na^+ ion to two of its nearest-neighbor like ions, what is the distribution of the angle θ_{+++} between them? This, and the corresponding question for θ_{ijk}, where i, j, k can each be either + or −, can be answered at least approximately and begins to elucidate the triplet distribution. These angular distributions are in fact rather broad for θ_{ijk} in the alkali halides but show peaks at ~48° for θ_{+++} and θ_{---} compared with 45° in the crystal, and near 90° and 180° for θ_{-+-} and θ_{+-+}. (For a more detailed discussion of such inferences, see McGreevy and Pusztai, 1990.)

We refer finally to $ZnCl_2$ for which the isotopic method, with or without RMC, suggests strongly that each Zn^+ ion is surrounded by a tetrahedron of four Cl^- ions. Since four anions per cation are not available, there is some vertex sharing of the tetrahedra similar to, though less orderly than, that in the crystal from which the material melts. Unlike the alkali halides, $ZnCl_2$ is very viscous just above its freezing point; its ionic conductivity is low and it readily forms a glass. These facts begin to suggest that its liquid and glassy forms depart from purely ionic interparticle forces towards ramifying networks of directional bonds more like those in glassy materials such as $GeSe_2$ or even the most well-known glass SiO_2. If so, $ZnCl_2$ would exclude itself from "simple liquids" as defined by our original criteria, and it is not surprising that, at the time of writing, it is one of several materials attracting more detailed study (Salmon, 1991.)

11. LIQUIDS AND GLASSES

One of the numerous ways of making glasses, indeed the normal way for most glasses, is to quench a liquid so rapidly that it has no time to crystallize. The result is a structurally disordered solid. All the concepts of Sec. 1–3.4 apply as well to glasses as to liquids. Much of Secs. 4.2 and 7–10 do also; e.g., glasses may be simulated by MD or discussed with pseudopotentials. The basic difference between liquids and glasses is that only the former are in thermodynamic equilibrium and therefore have their structures fully determined by the prevailing temperature and pressure. Since glasses are not in thermal equilibrium, their structure is not stable and in general changes slowly with time. Thus it is not uniquely defined: it depends on the history of the sample and especially on its mode of preparation and heat treatment. While the $g_{ij}(r)$ of a glass may well bear a general resemblance to those of a liquid of the same composition, they are rarely identical, and the glass may show more detailed features characteristic of short- or intermediate-range order. (See GLASSES.)

12. CONCLUSION

The reader may feel that the structure of simple liquids is somewhat unspectacular—all PDFs are continuous curves showing a few rather featureless peaks which soon peter out at r greater than a few atomic diameters. This just observation results from the fact that liquids are, to a first approximation, irregular packings of spheres. Size differences, the Friedel oscillation in ϕ_{eff} in metals, chemical binding forces in alloys, the orientation of nonspherical molecules, or Coulomb forces between ions all impose on the general randomness various forms of SRO and IRO that give the PDFs such features as they have. Understanding these is part of the general challenge presented by condensed matter including crystals. Determining structure from basic theory has never been easy because the energy differences between various alternative structures are small, so that quite accurate theories are necessary to establish which is the preferred structure of lowest energy. The period from 1970 to 1990 saw sufficient development in the theory of liquids and in computer simulations to justify a claim that structures are beginning to be understood from first principles even in disordered liquids. The subject is not closed, however. Apart from the problems of embracing more complex cases such as larger molecules or transition metals, there remain both experimental and theoretical difficulties in the way of understanding how structures change with density on the way to the liquid-vapor critical point. The triplet distribution and the effect on particle interactions of many-body forces also remain as challenges.

GLOSSARY

Ensemble Average: Average of a property over a large number of replicas of a system called collectively an ensemble; a basic device in statistical mechanics (*q.v.*).

Fourier Transform (FT): $F(y)$ is the FT of the function $f(x)$ if

$$F(y)=\int_{-\infty}^{\infty} f(x)\exp(-2\pi ixy)dx.$$

f is then the inverse FT of F and

$$f(x)=\int_{-\infty}^{\infty} F(y)\exp(2\pi ixy)dy.$$

There are many equivalent and related forms.

Free Energy: A fundamental thermodynamic property of any macroscopic system from which other properties may be derived. Gibbs free energy $G=E-TS+pV$; Helmholtz free energy $F=E-TS$.

Heat of Mixing: The heat given out, or taken in, when a mixture is formed from constituents.

Pairwise Interaction: Force between two particles that is unaffected by the presence of other particles.

Partition Function: The quantity $\Sigma \exp(-E_n/k_BT)$, where E_n is the energy of the nth allowed quantum state of a macroscopic system and the sum is over all states. It is fundamental for the calculation of thermodynamic properties by statistical mechanics (*q.v.*).

Stoichiometric Proportion: The proportion of constituents involved in chemical combinations.

List of Works Cited

Abernethy, G. M. Dixon, M., Gillan, M. J. (1981), *Philos. Mag.* B **43**, 1113–1123.

Andreani, C., Dore, J. C., Ricci, F. P. (1991), *Rep. Prog. Phys.* **54**, 731–788.

Arnold, A., Mauser, N., Hafner, J. (1989), *J. Phys. Condens. Matter* **1**, 965–980.

Ashcroft, N. W., Stroud, D. (1978), in: F. Seitz, H. Ehrenreich, D. Turnbull (Eds.), *Solid State Physics*, Vol. 35, New York: Academic, pp. 2–81.

Bellissent-Funel, M.-C., Buontempo, U., Petrillo, C., Ricci, F. P. (1990), *Mol. Phys.* **71**, 253–268.

Bhatia, A. B., Hargrove, W. H. (1974), *Phys. Rev. B* **10**, 3186–3196.

Bhatia, A. B., Thornton, D. E. (1970), *Phys. Rev. B* **2**, 3004–3012.

Biggin, S., Enderby, J. E. (1981), *J. Phys. C* **14**, 3577–3583.

Biggin, S., Enderby, J. E. (1982), *J. Phys. C* **15**, L305–L309.

Cusack, N. E. (1987), *The Physics of Structurally Disordered Matter*, Bristol, England: Hilger.

Derrien, J. Y., Dupuy, J. (1975), *J. Phys. (Paris)* **36**, 191–198.

Dore, J. C. (1990), in: M. C. Fairbanks, A. N. North, R. J. Newport (Eds.), *Neutron and X-Ray Scattering: Complementary Techniques*, Institute of Physics Conference Proceedings Series No. 101, Bristol, U. K.: IOP.

Edwards, F. G., Enderby, J. E., Howe, R. A., Page, D. J. (1975), *J. Phys. C* **8**, 3483–3490.

Egelstaff, P. A., Gläser, W., Litchinsky, D., Schneider, E., Suck, J.-B. (1983), *Phys. Rev. A* **27**, 1106–1165.

Enderby, J. E., North, D. M., Egelstaff, P. A. (1966), *Philos. Mag.* **14**, 961–970.

Gray, P., Cusack, N. E., Tamaki, S., Tsuchiya, Y. (1980), *Phys. Chem. Liq.* **9**, 307–322.

Guinier, A. (1963), *X-Ray Diffraction*, San Francisco: Freeman.

Hafner, J. (1977), *Phys. Rev. A* **16**, 351–364.

Hafner, J. (1990), in: H. Endo (Ed.), *Liquid and Amorphous Metals VII*, Amsterdam: North-Holland, p. 18.

Hafner, J., Kahl, G. (1984), *J. Phys. F* **14**, 2259–2278.

Hansen, J.-P., McDonald, I. R. (1986), *Theory of Simple Liquids*, London: Academic.

Hausleiter, Ch., Hafner, J. (1990), *J. Phys. Condens. Matter* **2**, 6651–6657.

Hoshino, K., Young, W. H. (1980), *J. Phys. F* **10**, 1365–1374.

Howe, M. A. (1990), *Mol. Phys.* **69**, 161–174.

Jacucci, G., Ronchetti, M., Schirmacher, W. (1985), *J. Phys. (Paris), Colloque* **46**, C8-385–C8-389.

Kilic, H. H., Pietzko, S., Schmutzler, R. W. (1990), in: H. Endo (Ed.), *Liquid and Amorphous Metals VII*, Amsterdam: North-Holland, p. 521.

Lieb, E. H. (1976), *Rev. Mod. Phys.* **48**, 553–569.

McDonald, I. R., Bounds, D. G., Klein, M. L. (1982), *Mol. Phys.* **45**, 521–542.

McGreevy, R. L., Mitchell, E. W. J. (1982), *J. Phys. C* **15**, 5537–5550.

McGreevy, R. L., Pusztai, L. (1988), *Mol. Simul.* **1**, 359–367.

McGreevy, R. L., Pusztai, L. (1990), *Proc. R. Soc. London Ser. A* **430**, 241–261.

Mittag, U., Samios, J., Dorfmuller, T., Guenster, S., Zeidler, M. D., Chieux, P. (1989), *Mol. Phys.* **67**, 1141–1151.

Rodger, P. M., Stone A. J., Tildesley, D. J. (1988), *Mol. Phys.* **63**, 173–188.

Salmon, P. S. (1992), *Proc. R. Soc. London Ser. A* **437**, 591–606.

Sangster, M. J. L., Dixon, A. (1976), *Adv. Phys.* **25**, 247–342.

Takeda, S., Harada, S., Tamaki, S., Waseda, Y. (1990), *J. Non-Cryst. Solids* **117/118**, 80–83.

Tildesley, D. J., Madden, P. A. (1981), *Mol. Phys.* **42**, 1137–1156.

Waseda, Y. (1980), *The Structure of Non-Crystalline Materials,* New York: McGraw-Hill.

Weeks, J. D., Chandler, D., Andersen, H. C. (1971), *J. Chem. Phys.* **54**, 5237–5247.

Yarnell, J. L., Katz, M. J., Wentzel, R. G., Koenig, S. H. (1973), *Phys. Rev. A* **7**, 2130–2144.

Further Reading

Cusack, N. E. (1987), *The Physics of Structurally Disordered Matter,* Bristol, England: Hilger.

Faber, T. E. (1972), *Introduction to the Theory of Liquid Metals,* Cambridge, England: Cambridge University Press.

Hafner, J. (1987), *From Hamiltonians to Phase Diagrams,* Berlin: Springer.

Hansen, J.-P. McDonald, I. R. (1986), *Theory of Simple Liquids,* London: Academic.

March, N. H. (1990), *Liquid Metals,* Cambridge, England: Cambridge University Press.

McGreevy, R. L. (1987), in: F. Seitz, H. Ehrenreich, D. Turnbull (Eds.), *Solid State Physics,* Vol. 40, New York: Academic, pp. 247–325.

Waseda, Y. (1980), *The Structure of Non-Crystalline Materials,* New York: McGraw-Hill.

LOCALIZED ELECTRON STATES

See ELECTRON STATES, LOCALIZED

LOGIC METHODS

See QUANTUM LOGIC METHODS; SEMANTIC APPROACH AND OTHER LOGIC METHODS

LOW-DIMENSIONAL PHYSICS

See ELECTRON STATES IN ZERO-, One-, and TWO-DIMENSIONAL SYSTEMS

LPCVD

See CVD (CHEMICAL VAPOR DEPOSITION)

MAGNETIC CONFINEMENT FUSION

See FUSION, MAGNETIC CONFINEMENT

MACHINES, ELECTRICAL

F. J. FRIEDLAENDER, *School of Electrical Engineering, Purdue University, West Lafayette, Indiana, U.S.A.*
D. W. NOVOTNY, *Department of Electrical and Computer Engineering, University of Wisconsin, Madison, Wisconsin, U.S.A.*

	Introduction	561
1.	**Principles**	562
2.	**Design Considerations**	563
2.1	Thermal	563
2.2	Electric Breakdown	564
2.3	Mechanical Failure	564
3.	**Direct-Current Motors and Generators**	564
3.1	General Description	564
3.2	Classification of dc Motors	566
3.3	Characteristics of dc Motors	567
3.3.1	Shunt Motors	567
3.3.2	Series Motors	567
3.3.3	Starting of Motors	568
3.4	Direct-Current Generators	568
4.	**Alternating-Current Machines**	568
4.1	Rotating Field	568
4.2	Induction Motors	569
4.2.1	Induction-Machine Equivalent Circuit and Characteristics	569
4.2.2	Single-Phase ac motors	571
4.3	Hysteresis Motors	571
4.4	Universal Motors	571
4.5	Three-Phase Synchronous Machines	572
4.5.1	Reluctance Motors	574
4.5.2	Amortisseur or Damper Windings	574
4.5.3	Transients in Synchronous Machines	574
4.6	Linear Motors	574
5.	**Permanent-Magnet Excitation**	574
6.	**Electronic Control of Electrical Machines**	575
6.1	Electronic Power Converters for Electric Drives	575
6.1.1	Controlled Rectifiers	575
6.1.2	Choppers	576
6.1.3	Phase Control of ac Voltage	576
6.1.4	Inverters	576
6.2	dc Drives	577
6.2.1	Capability Curves	578
6.2.2	High–Dynamic-Performance dc Drives (Servo Drives)	579
6.2.3	Open-Loop dc Drives	579
6.3	Variable-Speed ac Drives	580
6.3.1	Synchronous-Motor Speed Control	580
6.3.2	Induction-Motor Speed Control	581
6.3.2.1	Voltage-Controlled Induction Motor Drives	581
6.3.2.2	Adjustable-Speed, Constant-Volts/Hertz Induction-Motor Drives	581
6.4	Adjustable-Torque, Vector-Controlled ac Drives	582
6.4.1	ac Current Regulators	582
6.4.2	Field Orientation	583
6.4.3	Brushless dc Machines	584
6.5	Other ac Drives	584
	Glossary	585
	Further Reading	585

INTRODUCTION

The class of devices to be considered here are used for energy conversion with electrical energy as one component. In motors, conversion is from electrical to mechanical energy. In generators and alternators, the energy flow is reversed. Usually, the mechanical energy is supplied by means of a rotating drive, and the majority of motors supply rotational motion, but there are an increasing number of applications in which electrical energy is used to

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

provide linear motion. In the following, motors utilizing dc, ac, and modulated power supplies are discussed, and dc generators and ac generators, usually called alternators, are considered. A final section treats the electronic control of electrical machines.

1. PRINCIPLES

All of the devices to be discussed employ a magnetic flux to produce a voltage, with the basic underlying principle being Faraday's law of magnetic induction. The calculation of the magnetic flux requires the application of Ampère's circuital law, and forces are calculated by using suitable formulations of the Lorentz force equations.

Computation of machine characteristics requires the calculation of fluxes in magnetic circuits containing ferromagnetic elements, often a difficult task complicated by the geometrical structures encountered. Hence, simplifying assumptions are usually made and magnetic circuit models employed, with circuit parameters obtained from tests on the machine. Also, computational methods are becoming increasingly available for magnetic flux computations; these are considered in the articles NUMERICAL METHODS and PHYSICS APPLICATIONS OF COMPUTERS—MODELING. Ampère's circuital law is the basic tool used for flux calculations:

$$\oint_{\mathscr{L}} \mathbf{H} \cdot d\mathbf{l} = \int_S \mathbf{J} \cdot d\mathbf{s}, \tag{1}$$

where **H** is the field intensity vector, integrated over a closed path $\mathscr{L}$, **dl** is a length increment tangential to the path, and **J** is the current density vector, which is integrated normal to the area S enclosed by the path $\mathscr{L}$ over which the field is integrated. When **J** consists only of currents in windings (wires), i.e., there are no substantial eddy currents in the iron, the above equation is often simplified to

$$\sum Hl = \sum Ni. \tag{2}$$

Here the individual terms Hl, known as magnetomotive forces (mmfs), are summed around a closed magnetic circuit, represented in a simplified manner by lumped magnetic circuit elements, and the current in each winding linking the magnetic circuit contributes a term Ni equal to the number of ampere turns carried by it. It should be noted that the above holds independent of the source of the current, external or induced by changing magnetic fluxes in the device.

In the magnetic circuits used in machines, ferromagnetic materials such as iron constitute the major portion of the magnetic flux path. These materials display a very nonlinear B vs H relation, where B is the flux density vector, or flux per unit area. B saturates at a flux density typical for each magnetic material (actually, not B but the magnetization M saturates—the reader is referred to the article MAGNETOSTATICS for a discussion of these concepts).

Faraday's law for lumped circuits simply states that the induced voltage V across an N-turn winding is given by

$$V = N\frac{d\phi}{dt}, \tag{3}$$

where $d\phi/dt$ is the time rate of change of the magnetic flux ϕ linking all N turns. For sinusoidal flux variation, $\phi = \phi_m \sin\omega t$ where ϕ_m is the amplitude of the total flux and $\omega = 2\pi f$ where f is the frequency of the flux variation. In the sinusoidal case, therefore,

$$V = \omega N \phi_m \cos\omega t. \tag{4}$$

The amplitude of the voltage $V_M = \omega N \phi_m$, and

$$V_M = \sqrt{2} V_{\text{rms}} = 2\pi f N \phi_m. \tag{5}$$

With $\phi_m = B_m A$, where B_m is the amplitude of the flux density, assumed to be uniform over the area A enclosed by the coil,

$$\begin{aligned} V_{\text{rms}} &= (2\pi/\sqrt{2}) f N \phi_m \\ &= 4.44 f N \phi_m \\ &= 4.44 f N B_m A. \end{aligned} \tag{6}$$

In case of relative motion between flux sources and windings in which voltages are induced, it can be shown that these induced voltages can be calculated in the same manner, even if sliding contacts are involved. The literature often uses the concept of motional emf (electromotive force), which leads to the simple formula

$$V = u\,B\,l \tag{7}$$

where V is the voltage across a straight conductor of length l moving at a velocity u perpendicular to a magnetic flux density B (i.e., u, B, and l are mutually perpendicular); in a more general case

$$V = \oint_{\mathscr{L}} \mathbf{u} \times \mathbf{B} \cdot d\mathbf{l}. \tag{8}$$

The calculation of voltages induced in case of "general motion" requires care, but its discussion is beyond the scope of this article. In most rotating or linear-motion machines, the simple arguments mentioned in the previous paragraph are quite adequate. Equation (4) is only meaningful for a closed circuit (only a few elements may contribute to the voltage V; which ones may depend on the choice of observer in this formulation).

Forces are calculated by using the Lorentz force equation, which takes the form

$$d\mathbf{f} = I\, d\mathbf{l} \times \mathbf{B}, \tag{9}$$

where $d\mathbf{f}$ is the force exerted on a conductor element $d\mathbf{l}$ carrying a current I, located in a field B—note vector relations. Analogously to the case of an induced voltage, the force on a straight conductor of length l carrying current I, located in a field B, is given by

$$f = I\, lB, \tag{10}$$

with f, l, and B mutually perpendicular.

2. DESIGN CONSIDERATIONS

Limiting design considerations are those leading to premature failure of the machine. There are essentially three failure mechanisms: thermal, mechanical, and electrical breakdown (there is no equivalent magnetic breakdown, except due to excessive forces, covered under mechanical breakdown).

It must be understood that the concern here is not only with immediate failure but also "decreased lifetime" of the device. Hence, the design decision involves economic factors as well as "years of expected use." Sound scientific judgment is not always available for the last criterion.

2.1 Thermal

In a stationary device, there are copper (I^2R) losses in the conductors, magnetic losses due to cyclic magnetization of the (iron) core, and stray losses due to induced currents or alternating magnetic fields in the adjacent structure. To these must be added frictional and other losses due to motion in rotating machines. The copper losses are obtained from an integration of $J^2\rho$ (J is current density, ρ is effective resistivity) over the entire copper volume. In this calculation, proper account has to be taken of the skin effect in the conductors. Dielectric losses are usually small at power frequencies and are neglected here.

Losses in the iron path of an electromagnetic device depend on the magnetic flux density and frequency. Just as the current is not uniform over the cross section of a conductor, the flux density is largest at the surface of the magnetic material, if it is conducting. Loss calculations are much more difficult in this case and are usually based on empirical knowledge (test data) of, for instance, the particular lamination thickness of the magnetic steel used in different portions of the magnetic circuit. The situation is further complicated by the fact that the losses are anisotropic in the laminations, i.e., they depend strongly on the direction of the magnetic flux density vector $\mathbf{B}$, as do the magnetic properties, in general, in the iron parts of an electrical machine. Iron losses are described in detail in the article MAGNETIC MATERIALS. The maximum value of B that is appropriate in a given portion of the magnetic circuit is not only dictated by thermal considerations but is limited by the saturation of the magnetic flux density—too high a value of B will lead to larger magnetizing currents, which increase the no-load loss and decrease the amount of current that can be safely carried for energy conversion purposes.

The total heat generated, at given values of magnetic loading (B) and electric loading (J), depends on the volume (L^3, where L is the "linear dimension") of iron and copper. Without artificial cooling, dissipation of heat is a function of surface area, which increases only as L^2. Hence, it is clear that for optimum use of a machine, additional means of cooling must be provided as the device becomes larger. In many electric machines, the surface area is increased by the provision of fins in larger units, and fan blades are usually mounted on the armature/rotor shaft to provide additional air flow for cooling larger units. More elaborate methods of cooling may also be used.

2.2 Electric Breakdown

In the design of electrical machines, the maximum voltage must not produce electric fields that cause dielectric breakdown (failure of the insulation), corona, or sparkover. Clearly, the actual voltages must be such as to provide a margin of safety. As in other matters, the effect of aging on insulation has to be taken into account to provide trouble-free operation over the estimated lifetime of the machine. Transients are another important consideration, as the applied voltage may not be distributed evenly over all the turns during rapid voltage changes.

2.3 Mechanical Failure

Mechanical failure results from the forces that act on current-carrying conductors in a magnetic field—the same forces that provide the driving force in electric motors. Forces also arise in motors and generators as a result of their motion, specifically centrifugal forces in rotating machines. Torques on drive shafts cause torsion forces that can lead to shaft breakage. Electromagnetic forces can be calculated from the Lorentz force equation [Eqs. (9) and (10)].

Generally speaking, the important parameters in the design of a machine are the maximum flux density, determined by material properties (saturation magnetization and iron losses), and the current density J, usually dictated by thermal considerations. In the operation of the machine, the total flux and total current rather than their densities are of concern. Hence, the amount of iron and copper (dictated, in part by their cross-sectional areas) and the size of a machine are all related to the magnetic and electric "loading" of the iron and copper, at a given voltage and current rating. In ac machines, with the voltage dependent on the frequency—Eqs. (4) to (6) clearly show how voltage relates to both frequency and magnetic flux—the size of a machine at a given rating varies inversely with frequency. One can also have different ratios of copper to iron volumes, a fact emphasized in transformer design. Hence, economic factors play an important part not only in the construction but also in the operation of electric machines.

This illustrates both the economy of scale and difficulties with scale changes in a general way. An important result is the increase in power handling capability at a given size of device with frequency.

3. DIRECT-CURRENT MOTORS AND GENERATORS

3.1 General Description

A dc machine usually consists of a fixed field structure with evenly spaced salient-pole pieces (in multiples of two) with current-carrying windings to provide a magnetic field. Additional, smaller pole pieces are placed in the space between the main poles in larger dc machines (interpoles—see later discussion). Permanent magnets instead of current-carrying coils are used in many modern electric machines, especially in the smaller sizes, to produce the desired magnetic flux. Laminated pole faces are often provided in large machines to reduce eddy currents produced by magnetic field fluctuation due to armature slotting (see below).

The rotating armature consists of a steel shaft with a steel hub on which sheet-steel disks with slots on the outer periphery are placed. There is a small air gap between the field structure and the laminated armature core surface (see Fig. 1).

Windings (insulated copper conductors) are placed in the armature slots, and current is carried from the conductors through a commutator—copper conductor segments insulated from each other—to external circuits by means of brushes arranged at spacings equal to the spacing of pole pieces on the field structure. Some examples of typical armature windings are given in Figs. 2 and 3; there are many other variations.

A conductor on the rotating armature experiences a force f since it carries a current I_A in the presence of the magnetic field of density B produced by the stationary field structure. The induced voltage e and the force per conductor f, moving at a velocity u and having a

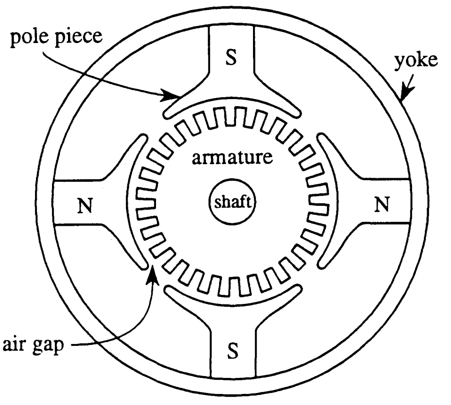

FIG. 1. dc motor. Commutator is mounted on shaft (not shown); see Fig. 2 for details of windings. Coils are placed on pole pieces to provide field excitation.

projected length l on the plane containing the **B** and **u** vectors, are given by [see Eqs. (7), (10)]

$$e=u\,B\,l \tag{11}$$

and

$$f=i\,l\,B. \tag{12}$$

More conveniently, the total generated voltage e_g for all the conductors in series between a pair of brushes is expressed by

$$e_g=K\phi\omega, \tag{13}$$

where ω is the radian velocity of the conductors (with $\omega R=u$, where R is the radius of the armature at the conductor position) and ϕ is the total air-gap flux per pole [$B\times$ (pole area)]; the coefficient K is determined by geometrical factors such as armature length, the number of turns, and their arrangement between brushes, etc., and is a constant for a given machine. $\omega/2\pi=n$ is the speed of rotation in revolutions per second, or rpm if multiplied by 60. e_g is also referred to as back or counter emf (in motors).

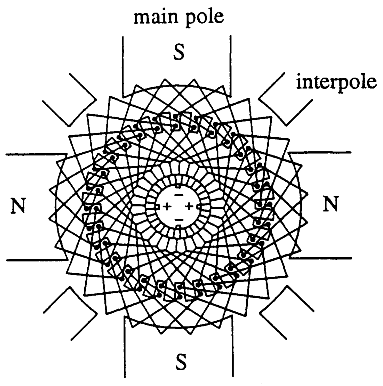

FIG. 2. dc motor with schematic of a lap armature winding shown.

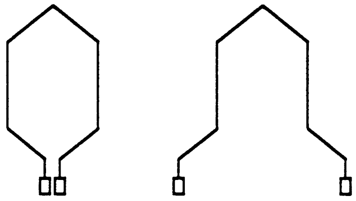
Coil for lap wound armature

Coil for wave wound armature

FIG. 3. Typical armature coil shapes.

The torque T_A produced by the interaction of the armature current with the field is similarly expressed as

$$T_A=K\phi I_a, \tag{14}$$

which gives the expected result

$$e_g I_A=T_A\omega \tag{15}$$

as required by energy conservation. T_A is the total torque, including friction and windage torques T_{loss} (shaft bearing friction and torques due to air motion produced by the rotating armature, often enhanced by fan blades to provide cooling). The net torque available at the shaft is $T_0=T_A-T_{\text{loss}}$.

Basic equivalent circuits for dc machines are given in Fig. 4. From the circuit equation and torque balance (in steady state)

$$V_0-R_A I_A=K\phi\omega, \tag{16}$$

$$T_0+T_{\text{loss}}=T_A=K\phi I_A, \tag{17}$$

speed–torque and voltage–current relations are readily obtained for a known value of ϕ. ϕ is determined by the field excitation. ϕ and, therefore, e_g depend on the field current I_f in the nonlinear fashion shown in Fig. 5. This nonlinearity is a consequence of the magnetization characteristics of iron. A machine is considered to be a motor if $T_0>0$ and a generator if $I_A<0$. In other words, energy flow from electrical energy to mechanical energy characterizes a motor, vice versa for a generator.

The magnetic flux ϕ due to the field excitation—different schemes are discussed

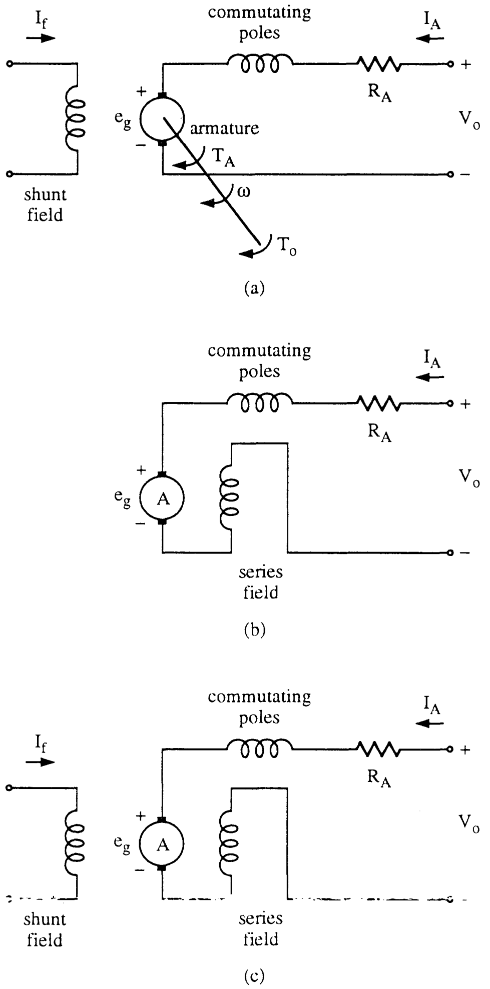

FIG. 4. Direct-current motor circuits: (a) dc shunt motor, (b) dc series motor, (c) dc compound motor.

below—can be readily obtained if the *B-H* characteristics and dimensions of each part of the magnetic path are known but are mostly determined by the air gap because of its higher reluctance (the effective permeability μ is much larger in the iron until it saturates). In practice, the experimental determination of e_g vs I_f is simple and convenient (and the results are readily adapted to all means of field excitation).

Currents flowing in the armature conductors produce an mmf that distorts the field produced by the poles. This effect is known as armature reaction. The armature current mmf essentially decreases the flux density in the trailing pole tip (in the case of a motor), and increases it in the leading pole tip. Because of saturation effects, the increase of flux density is smaller than the decrease, and the net effect of the armature reaction is to decrease the total available magnetic flux with increasing load.

Another factor that influences the operation and design of dc motors is commutation of currents as the commutator segments pass the brushes. Inductance effects cause sparking at the commutator brushes and, if not corrected, in all but the smallest motors or generators would lead to the destruction of the commutator. To counteract the effect, pole shoes (interpoles) are used in the space between the main poles (see Fig. 2), with windings connected in series with the armature in such a direction as to reduce the inductance of the armature coils under commutation to a small value. In very large machines, interpoles (also called commutating poles) do not suffice and a compensating winding, carrying armature current, is placed in slots on the main pole pieces, to oppose and hence reduce the net flux produced by the (commutating) armature conductors and avoid sparking at the commutator brushes.

3.2 Classification of dc Motors

Usually, dc motors are classified by the method used to obtain field excitation. Broadly speaking, one has to distinguish between *shunt excitation*, in which the field coils are supplied from constant or variable external dc voltage sources, and *series excitation*, in which the armature current flows through a set of field coils connected in series with the armature. Shunt coils are usually high-resistance, many-turn windings carrying a relatively small cur-

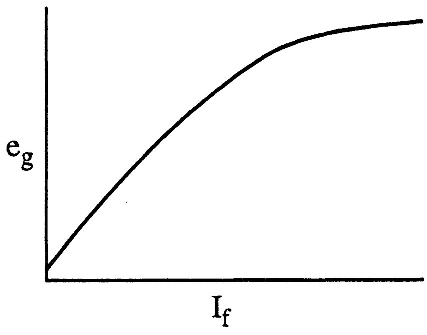

FIG. 5. Generated voltage e_g versus field current. Because of the remanent magnetic flux, e_g is not zero when I_f=0.

rent. On the other hand, series coils consist of few turns of wire, capable of carrying the total armature current at all loads.

A combination of shunt and series excitation is often used, with both a high-resistance shunt coil and a low-resistance series coil placed on each pole. Such a motor is called a *compound motor.* In a compound motor, the field excitation is given by the sum of the ampere turns of the shunt and series windings (or the difference in the case of a motor with negative compounding, which is rarely used). In Fig. 4 circuit diagrams are given for all the above.

dc motors may have fields provided by permanent magnets. Such motors behave like motors with shunt excitation, since a permanent magnet is, roughly speaking, equivalent to a winding carrying constant current.

3.3 Characteristics of dc Motors

The speed-torque characteristics of a dc motor depend on its mode of field excitation—shunt, series, or compound. Typical speed–torque curves are shown in Fig. 6.

3.3.1 Shunt Motors. As can be seen from Eq. (16),

$$\omega \approx (V_0/K\phi) \text{ if } R_A I_A \ll V_0,$$

as is usually the case. The effect of $R_A I_A$, at constant field excitation, is a small reduction in speed with increasing load. However, this effect is partly compensated by the decrease in flux ϕ due to armature reaction. A shunt motor provides, therefore, relatively constant speed; i.e., the slope of the speed–torque curve shown in Fig. 6 is very steep. Moreover, the motor speed level can be easily controlled by having the "shunt" field separately excited by means of a variable-voltage source (or a constant-voltage source with a variable series resistance or an electronic control; see Sec. 6). If the armature is also supplied from a variable-voltage source, then both V_0 and ϕ are independently controlled, resulting in a wide range of possible speeds. The easy speed control of dc shunt motors has been their traditional advantage over induction motors, offsetting their higher cost, though this advantage is now diminished by the advent of electronic control of induction motors.

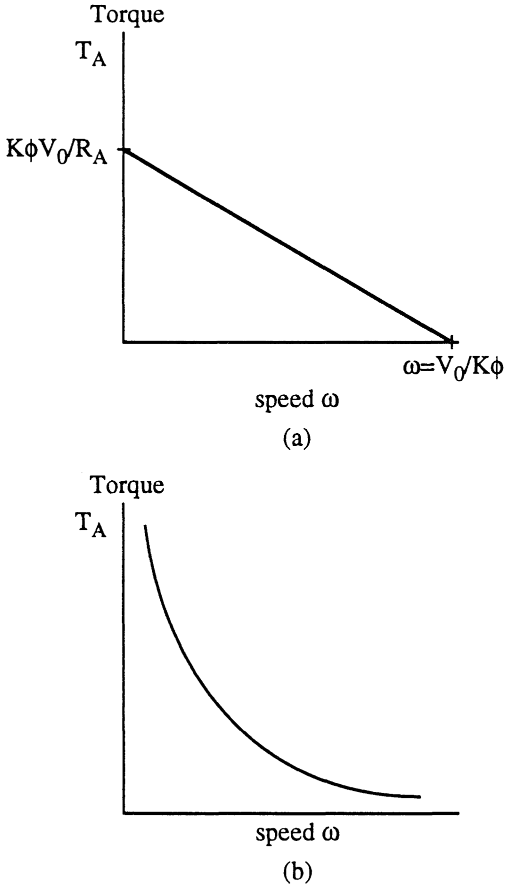

FIG. 6. Torque vs speed curves. (a) Shunt motor, (b) series motor.

3.3.2 Series Motors. When the field excitation is supplied by the armature current, ϕ becomes a function of I_A. If we neglect saturation effects, we have $\phi \propto I_A$, $T \approx K I_A^2$, and $\omega \approx (V_0/K I_A^2)$. Series motors exhibit very high torques at low speeds, at which armature current has to be limited to a safe value by means of external resistances (see Sec. 3.2.3). On the other hand, they have to be protected against excessive speeds at no load, since ϕ will drop almost to the remanent value, allowing excessive speeds. Series motors are used frequently in traction applications where the load is never removed and high starting torques are required.

The characteristics of a *compound motor* lie somewhere between those of shunt and series motors, depending on the amount of compounding (ampere turns of series winding relative to those on shunt winding, at a given load).

Operating points are readily found for any motor by plotting load-torque versus speed characteristics on the same graph as the motor torque-speed curves, and finding the intersection of the two curves, for each operating point (i.e., for each load condition and/or motor condition).

3.3.3 Starting of Motors. Transients in motors depend on mechanical and electrical time constants, with the mechanical time constants often much larger than the electrical time constants. For this reason, the starting currents of a dc motor have to be reduced by placing decreasing values of external resistance R_e in the armature circuit, as the speed ω gradually increases. I_A has to be limited to some safe value, a value that will not overheat the armature and series circuit during the starting period and will not allow excessive forces. R_e is usually decreased in steps so that I_A varies between an upper limit and the full-load value (or some similar value).

3.4 Direct-Current Generators

dc generators are essentially identical to dc motors, except that the power flow is reversed. Hence, the signs of T_A and I_A change (as well as the sign of T_0, of course). Then the applied torque T_0 must equal the sum of T_A and T_{loss}. Similarly, the output voltage V_0 is now given by the generated voltage $K\phi\omega$ minus the $R_A I_A$ voltage drop [see Eq. (16)].

Again, generators can be classified by their field connections. The separately excited generator has its field (with a variable series resistance for voltage adjustment) connected to a constant-voltage source. Hence the generated voltage $K\phi\omega$ is a constant for a given ω (actually, armature reaction will cause some decrease of ϕ at higher load currents). A self-excited machine has its shunt field connected across the armature terminals. The resulting operating point is readily determined from the magnetic characteristic e_g vs I_f of the machine, shown in Fig. 5, and its intersection with the field load line, i.e., Ohm's law applied to e_g and I_f: $e_g = R_f I_f$ where R_f is the total resistance of the shunt field circuit. The fact that there is usually some permanent magnetic flux plays an essential role in making a self-excited generator possible (as seen in Fig. 5, e_g is not equal to zero when I_f is zero). As the armature starts rotating the remanent flux will induce a voltage in the armature, causing some current to flow, allowing the flux, and hence, voltage to "bootstrap" up to the desired value if appropriate resistance values are chosen for the field circuit.

Compounding i.e., additional field windings carrying the armature current, can be added, producing a compound dc generator. The two windings usually are aiding—both produce magnetic fluxes in the same direction. Hence, a compound generator displays a flatter V_0 vs I_A characteristic, since the series winding increases the flux at larger armature current, thus compensating for the larger $R_A I_A$ voltage drop. Series generators are not often used, but their characteristics are readily derived from the principles given above. Typical curves for dc generator terminal voltage versus load current at a given speed are shown in Fig. 7 for various amounts of compounding.

4. ALTERNATING-CURRENT MACHINES

4.1 Rotating Field

A brief derivation of the rotating-field concept is given here, since it is the basis of most ac rotating machines. A three-phase arrangement will be considered, in which three separate sets of coils are supplied with three-phase power—three voltage sources with 120° time-phase difference between them.

We shall assume that each set of coils, arranged on the surface of a cylinder, pro-

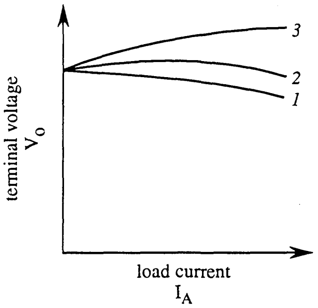

FIG. 7. Terminal voltage vs load (armature) current. (1) Shunt generator, (2)compound generator with flat compounding, (3) overcompounded generator

duces a sinusoidally distributed flux at the cylindrical surface, described by

$$B=B(t)\cos(p/2)\theta \tag{18}$$

with

$$B(t)=B_m\cos(\omega t), \tag{19}$$

where p is the number of poles, determined by the winding pitch (i.e., the coil span; i.e., 180° for $p=2$, 90° for $p=4$, etc.) and placement of windings and θ defines the angular position on the cylinder. Obviously p has to be an even number. One period of the pole structure (i.e., "north pole" to nearest "north pole") spans $2\pi/(p/2)$ radians. The quantity $p\theta/2=\theta_e$ is referred to as "electrical degrees." In a typical three-phase machine, where the supply voltages differ in phase by 120°, the three identical sets of windings are displaced by $\theta_e=120°$ from each other.

The resultant field B at an angle θ and time t is given by the addition of the three components:

$$\begin{aligned}B&=B_m\cos\omega t\ \cos(p\theta/2)+B_m\cos(\omega t+2\pi/3)\\&\quad\times\cos(p\theta/2+2\pi/3)\\&\quad+B_m\cos(\omega t-2\pi/3)\\&\quad\times\cos(p\theta/2-2\pi/3)\\&=(B_m/2[\cos(\omega t+p\theta/2)+\cos(\omega t-p\theta/2)\\&\quad+\cos(\omega t+p\theta/2+4\pi/3)\\&\quad+\cos(\omega t-p\theta/2)\\&\quad+\cos(\omega t+p\theta/2-4\pi/3)\\&\quad+\cos(\omega t-p\theta/2)]\\&=\tfrac{3}{2}B_m\cos(\omega t-p\theta/2).\end{aligned} \tag{20}$$

Thus, the resultant field is a single field of constant amplitude $\frac{3}{2}B_m$ rotating at a speed

$$\frac{d\theta}{dt}=\omega/(p/2)=2\pi n_0, \tag{21a}$$

where n_0 is called the synchronous speed (in revolutions per second). Expressed in rpm,

$$\frac{d\theta}{dt}=\frac{\omega}{p/2}\,\frac{60}{2\pi}=(120f/p)\ \text{rpm}. \tag{21b}$$

Reversal of the connection to one of the three phases will reverse the direction of rotation. Note that each phase—each set of coils—produces an alternating field that can be resolved into two half-amplitude oppositely rotating fields. In this arrangement, the fields rotating in one direction have a vector sum of zero; those rotating in the opposite direction are exactly in phase and hence add to give a rotating field of amplitude $3B_m/2$. n-phase systems for $n\geqslant 3$ (n an integer), with coils displaced by $\theta_e=2\pi/n$ radians and supplied from sources with phase differences of $2\pi/n$ radians between adjacent phases, behave similarly. $n=2$ is a single-phase system with polarity reversal.

4.2 Induction Motors

An induction motor usually has a three-phase winding arranged in slots in the laminated stator structure, producing a rotating field in the air gap, that is, the space between the stator lamination surface and that of the cylindrical rotor. There are slots symmetrically placed around the rotor armature, into which copper or aluminum bars are inserted. These bars are shorted by solid metal (copper or aluminum) rings at each end. The conductor structure, if visualized without the iron (it is often cast as a unit on the rotor and not removable), looks like a cage; hence the name "squirrel cage" induction motor is used to describe this type of motor

In an induction machine the rotating field generated by the stator windings induces voltages and, hence, currents in the usually short-circuited rotor conductors. The field caused by the rotor currents interacts with the stator field and causes the rotor to start rotating. A simple explanation: the rotor field "chases" the stator field but never catches up—if it did, there would be no induced currents in the motor and, hence, no interaction [but supplying a dc current to the rotor conductors would cause the motor to "lock" in with the rotating stator field and become a synchronous motor, rotating at $(120f/p)$rpm].

4.2.1 Induction-Machine Equivalent Circuit and Characteristics. The equivalent circuit of the "locked rotor" (rotor not turning) ac motor is essentially that of a transformer (with an air gap). As the rotor starts turning, the voltage induced in the rotor windings decreases, and goes to zero if a speed of

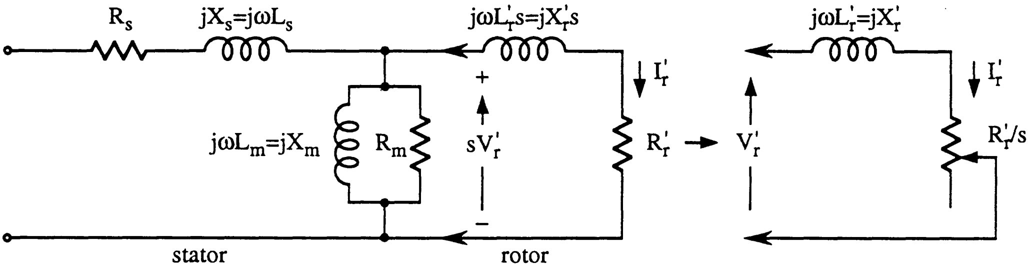

FIG. 8. Approximate equivalent circuit of induction motor showing evolution to modified rotor representation.

$2f/p=n_0$ rps is reached (i.e., the rotor moves at the speed of the rotating field). If we define slip as

$$s=(n_0-n)/n_0, \tag{22}$$

where n is the actual speed of rotation, the voltage induced in the rotor can be shown to be equal to sV_r, where V_r is the voltage induced in the "locked" rotor.

An equivalent circuit is shown in Fig. 8. It should be noted that the secondary (rotor) circuit has its values referred to the primary; hence the primed parameters (see TRANSFORMERS). As the speed of the rotor increases, and hence its speed relative to the rotating field decreases, the frequency of the voltage induced in the rotor decreases–it is always equal to $f(n_0-n)/n_0$, or sf, where f is the frequency of the stator ac supply.

It is convenient to "scale" rotor values by s, i.e.,

$$I'_r=sV_r/(R'_r+js\omega L'_r)=V'_r/(R'_r/s+j\omega L'_r) \tag{23}$$

without affecting the current (or the torques produced by these currents).

The evolution of the rotor circuit, referred to primary stator windings, is also shown in Fig. 8. Since $R'_r/s=R'_r+[(1-s)/s]R'_r$, Fig. 8 allows a simple interpretation of the power flow. I'_r and R'_r are the physical values of the parameters (actually R'_r may have to be adjusted for the change in the frequency of rotor currents because of skin-depth variations). The term $[(1-s)/s]R'_r$ is a fictitious resistance that models the power flow from the motor to the mechanical system via the rotating rotor shaft:

$$P=Tn_0(1-s)=I'^2_r[(1-s)/s]R'_r. \tag{24}$$

Since I'_r can be found from the equivalent circuit, torque versus speed curves for a fixed-voltage excitation of the stator are readily obtained from Eq. (24).

Typical speed–torque curves for an induction motor are shown in Fig. 9, for a series of values of rotor resistance R_r. Increasing the value of R_r shifts the point of maximum torque T_m to lower speeds, without affecting the value of maximum torque, as is readily derived from the above equations. Increasing the value of R_r consequently increases the starting torque. Squirrel-cage motors are often built to have a larger resistance upon starting than during normal operation. This is achieved by taking advantage of the fact that the depth of penetration of the currents varies with the speed: Cast aluminum or copper bars are placed in deep rotor slots, and a second winding with a higher resistance (usually cast aluminum) is placed at the top of the slots, near the air gap. As a result, the rotor resistance is higher near the surface where the currents concentrate at starting. At higher

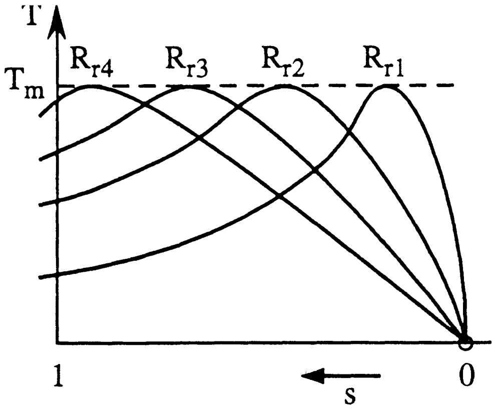

FIG. 9. Induction-motor torque-vs-slip curves for a number of values of rotor resistance R_r, with $R_{r_4}>R_{r_3}>R_{r_2}>R_{r_1}$. T_m is the maximum torque independent of value for R_r for constant voltage.

speeds, the frequency of the current decreases and, consequently, the current spreads deeper and a large portion of it is carried by the lower resistance copper bars.

As s reverses its sign (motor is driven above synchronous speed n_0) an induction motor becomes an induction alternator—it will convert mechanical energy to electrical energy. However, the stator has to be supplied by an ac source: this machine is not "self-excited" unless a bank of capacitors is connected to the terminals. Figure 10 shows torque and current versus s over a wide range of values of s defining different ranges of operation.

In some cases, instead of a squirrel-cage, separate three-phase windings are placed on the rotor and connected through slip rings to an outside circuit. This scheme allows a better control of the induction machine, but is becoming rare as more effective electronic means are being adopted to control the machine via the stator power supply.

4.2.2 Single-Phase ac Motors. The principle of operation of single-phase motors is readily explained by using an observation made in connection with the derivation of Eq. (20). Each phase produces an alternating field that can be resolved into two half-amplitude oppositely rotating fields. Superposing the torque speed curves of the two oppositely rotating fields, one finds that at the small value of slip at which induction motors operate, only the torque due to the appropriate component is significant. To get the motor started in the desired direction several schemes are used. The most common is the use of an auxiliary winding, displaced by 90° electrical from the main winding, and phase shifted by close to 90° by placing a large capacitor in series with it. A centrifugal switch may be used to disconnect this starting winding as the motor approaches the design speed. It should be noted that a single-phase motor is much noisier (pulsating rather than rotating field), larger, and less efficient than a three-phase motor of the same rating. Its main advantage is that it can be used where three-phase power is not available. A torque versus speed curve for a single-phase induction motor without an auxiliary starting winding is shown in Fig. 11.

In a *shaded-pole motor*, used for small single-phase motors with low starting torques, a short circuited copper strap is placed on a fraction of each pole (the "lagging" end of the pole). The induced currents provide a "lagging" field that, in turn, provides a small unidirectional torque, sufficient to start a fan motor, for instance.

4.3 Hysteresis Motors

Hysteresis motors (usually very small single-phase motors) derive their torque from the lagging flux produced by the motor magnetization, during subsynchronous operation. Special magnetic materials are used that are semihard (i.e., they are "permanent magnets" that can be cycled around their hysteresis loop by the stator currents). When the rotor reaches synchronous speed, the rotor flux no longer cycles and the permanent magnet rotor behaves as in a synchronous machine.

4.4 Universal Motors

Universal motors are series motors designed with relatively low field inductance, to allow them to function on either dc or ac supplies. Both stator and armature need to be laminated, and the design is usually more difficult since coils under commutation have voltages induced in them due to the ac fluxes (if used on an ac supply). Because of the higher ac impedance of a shunt field, as compared with its resistance, and because the high inductance of the shunt would lead to an undesirable phase difference between the stator and armature fields, universal motors are

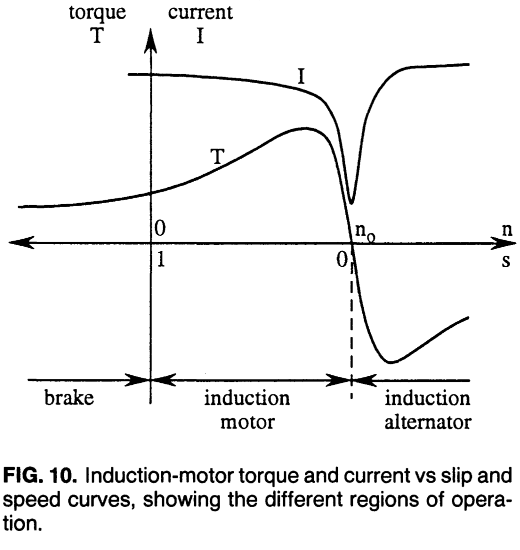

FIG. 10. Induction-motor torque and current vs slip and speed curves, showing the different regions of operation.

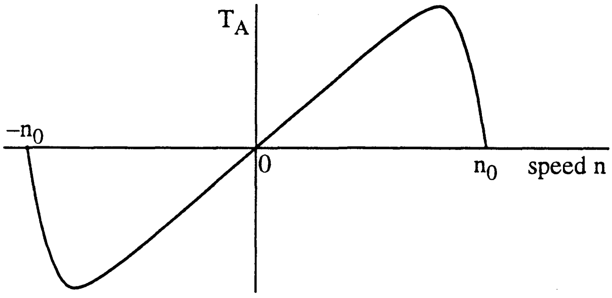

FIG. 11. Torque–speed curve for a single-phase induction motor without auxiliary starting winding.

not practical in a shunt-field configuration. Universal motors are commonly used in hand tools and similar applications.

4.5 Three-Phase Synchronous Machines

In synchronous machines, dc-excited poles rotate at a speed ω_s, where $\omega_s = 2\pi n_0/60$ rad/s, relative to stator windings similar to those described for an induction machine rotor. In a three-phase machine, there will be three windings displaced by 120° electrical from each other. The rotor may have a smooth cylindrical surface with slots to hold the dc windings (in large turbine alternators) or may consist of salient poles excited by windings, or the windings may be replaced altogether by permanent magnets. In this discussion we will explain the behavior of synchronous machines from a simple point of view without detailed considerations of harmonics—pure sine waves are only encountered in ideal machines—or of saturation effects. The reverse configuration (dc field on the stator and three-phase winding on the rotor) is possible, but rare, and can be described in a similar manner.

For symmetrical three-phase circuits, one needs to consider only one of the three phases—the other two will only differ by having their currents and voltages displaced by $+120°$ and $-120°$, respectively. For smooth-rotor (cylindrical rotor) synchronous machines, the reluctance of the magnetic paths will not depend on the position of the rotating member (if one neglects the effect of the slots, which will produce harmonics but will not directly affect the fundamental component in a major way).

An understanding of the principles of operation is most easily gained by considering the magnetic fluxes produced by the field current I_f and by the currents in the three-phase windings. Suppose the two fluxes are exactly in phase so that a rotor "south" pole faces a stator "north" pole. In this situation no torque can result. The two fluxes must be out of phase, and in fact the stator windings carrying the currents that produce the mmf will ideally be situated 90° (electrical) away from the center of the rotor pole, in a region of zero flux. For the rotor field to lead the stator field (the air-gap field), a torque, that is, power, needs to be applied to the rotor shaft; the synchronous machine, then, is an alternator. If, on the other hand, the rotor field lags behind the stator (air gap) field by some angle, then a torque is applied to the rotor by the field—it is "pulled along" by the air-gap field—and the machine becomes a synchronous motor. It can be shown that the torque T (or power flow $T\omega_s$, with ω_s a constant) is proportional to the product of the stator current I_s, the rotor field ϕ_f, and sin γ where γ is the electrical angle between the fields produced by the rotor and the stator current I_s, respectively. A very simple model results if one neglects stator resistance (reasonable for large machines), leakage reactance, and saturation so that one can assume that the total mmf produced relates linearly to the total voltage V_t at the terminals of the ideal machine. This voltage is the sum of E_f, the voltage generated by the rotor magnetic flux, plus the "armature reaction" voltage due to the stator current. $V_s = jX_sI_s$ where X_s is the magnetizing reactance of the stator and δ is the angle between E_f and V_t (between the direction of the rotor field and the resultant field, i.e., the rotating field in the air gap due to all mmfs. One obtains

$$(V_t - E_f)/jX_s = I_s. \qquad (25)$$

It follows from the above that the power P per phase of this ideal machine is

$$P = (V_t E_f \sin\delta)/X_s \tag{26}$$

with the sign of δ determining the direction of power flow. The same qualitative result was obtained from the physical reasoning above.

We can use this ideal model to understand the power flow, which has a maximum value of $V_t E_f/X_s$—if this is exceeded, the machine will fall out of synchronism (that is, the stator and rotor fields do not rotate at the same speed any more, and the average torque is zero). One can also study the use of the machine as a *synchronous condenser*, the operation when $P=0$ and hence $\delta=0$ (still neglecting losses) and V_t is fixed (supplied by a "bus," which is essentially a connection to a system that determines both V_t and the frequency f, and hence ω_s). For one value of field current I_f, $V_t=E_f$ and $I_s=0$. If I_f is increased, I_s has to increase in such a direction and phase that $\mathbf{E}_f+jX_s\mathbf{I}_s=\mathbf{V}_t$ for $E_f>V_t$ ($\delta=0$ since we are assuming a lossless, ideal machine). $\mathbf{I}_s$ has to lead $\mathbf{E}_f$ and $\mathbf{V}_t$ by 90°, i.e, $\mathbf{E}_f$ $E_f\angle 0°$ and $\mathbf{V}_t=V_t\angle 0°$ in phasor notation with $\mathbf{I}_s=I_s\angle 90°$. This machine behaves just like a capacitor, and hence this mode of operation is called that of a "synchronous condenser." A decrease in field current would cause a 90° lagging current to flow in the stator, a less desirable mode since a purely inductive load would lower the power factor of the system (assumed to be lagging), normally a poor option.

In a "real" machine, the terminal voltage V_t would be modified by additional voltage drops due to stator resistance and leakage reactance, conceptually similar to the transformer circuit model. Though complicating the circuit somewhat (see Fig. 12), the calculations needed to take account of the more realistic situation are straightforward.

Also, the nonlinear effects due to saturation in the iron have to be taken into account. The field excitation curve E_f vs I_f displays the effects of saturation—it is quite similar to that of a dc machine. Values of E_f have to be adjusted accordingly in regions of higher mmf and, hence, higher E_f values. With saturation, the value of X_s also decreases.

Experimental determination of cylindrical-rotor synchronous-machine parameters is similar to that of transformer parameters. E_f is determined by means of an open-circuit test, as a function of the field mmf. The short-circuit current is obtained, also as a function of the field mmf. The ratio of the open-circuit voltage divided by short-circuit current is the synchronous impedance at the value of the field mmf at which the ratio is taken. The synchronous impedance is the sum of the resistance plus the total reactance of the machine. The latter is also called synchronous reactance.

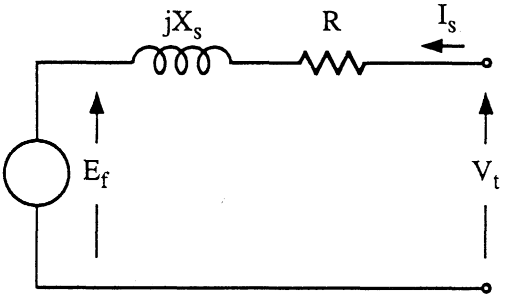

FIG. 12. Approximate equivalent circuit of a synchronous machine (per phase). X_s is the synchronous reactance and R is the total resistance of the stator circuit.

A salient-pole synchronous machine has separate pole pieces, with gaps between the poles. Hence, the reluctance of the magnetic path varies as the rotor rotates, with a complete period of the variation occurring as the rotor advances by 180° electrical. This situation is described most conveniently by defining two separate values of the stator reactance, the direct-axis synchronous reactance X_d for the position of the stator windings so that their mmf acts in the low-reluctance path, and a quadrature-axis synchronous reactance X_0, 90° electrical from the direct axis, when the reluctance is a maximum. The stator current is expressed in terms of its components, one in phase with the direct-axis position and the other (at 90° from the first) in phase with the quadrature-axis position of the rotor relative to the stator.

If the resistance and leakage reactance of the stator windings are neglected, one can write, approximately, for the salient-pole motor,

$$\begin{aligned} P = {} & (V_t E_f/X_d)\sin\delta + (V_t^2/2X_d X_q) \\ & \times (X_d - X_q)\sin 2\delta. \end{aligned} \tag{27}$$

As can be seen, the saliency adds a term called reluctance torque (or power when multiplied by ω_s) that only depends on the angle δ

between the center of the salient field structure and the peak of the air-gap flux variation of the fundamental term. Since the saliency introduces a 180° (electrical) periodicity, a "2δ" function appears.

4.5.1 Reluctance Motors. A *reluctance motor* is one that uses no dc field but relies entirely on the variation of reluctance of the structure to produce a torque. In principle, a *stepping motor* also is based on similar reluctance torques with the essential difference that, instead of three-phase excitation, two or three phase windings (arranged appropriately 90° or 120° electrical from each other) are excited successively by pulses, at a rate consistent with the speed at which the motor is to be "stepped" (and within a frequency range consistent with the capability of the system in terms of electrical and mechanical time constants).

4.5.2 Amortisseur or Damper Windings. Slots may be provided on the air-gap surface of the field structure (assumed to be the rotor here and in most of the previous discussion), and copper or aluminum bars placed in the slots and short-circuited at each end by means of a copper or aluminum ring, respectively, as in an induction (squirrel-cage) motor. The purpose is twofold. Currents are induced in these conductors when there is any "slip" between the rotating dc field structure and the synchronously rotating ac field, and will serve as a means of damping such motion. In a synchronous motor, this short-circuited winding will start the motor as an induction motor until it can be synchronized as the slip becomes small. Synchronization of synchronous machines is an important, practical topic beyond the scope of this article.

5.4.3 Transients in Synchronous Machines. These are a somewhat formidable subject. In the previous treatment, reactances have been discussed, but only for situations when all magnetic fluxes are stationary with respect to each other in a rotating frame of reference (expressed above by the term "synchronous speed"). In transients, this is no longer the case and additional "transient" and "subtransient" reactances have to be introduced, to account for the flux linkages between the dc field windings and the ac stator windings and damper windings, respectively. In salient-pole machines, direct- and quadrature-axis transients and subtransients have to be taken into account. The basic method used is based on the "conservation of flux linkages," in which one examines the rate of change of flux linkages in each circuit according to its time constant and uses the fact that flux linkages are conserved for times that are short compared to the time constant of the circuit considered.

4.6 Linear Motors

Linear motors are used to produce linear rather than rotational motion. Linear motors are built to operate as induction, synchronous, or reluctance motors. But instead of windings arranged in a cylindrical geometry, a planar geometry is used. The design may differ somewhat and larger air gaps are customary in traction devices. More sophisticated controls are needed. Only the windings in the region of the "moving part" of the motor (previously called rotor) are excited, and speed control, especially in synchronous motor design, is usually accomplished by means of variable-frequency excitation. One prospective application is the use of linear motors in high-speed transportation, with levitation integrated into the motor in some designs.

5. PERMANENT-MAGNET EXCITATION

Both dc machines and synchronous machines require a dc field structure. Conventionally, these fields were supplied by current-carrying coils placed on a pole structure (iron pole pieces) as explained above. However, with the advent of new permanent-magnet materials, and the availability of inexpensive ferrite magnets, dc fields can be supplied by substituting permanent magnets for the electromagnets in the field structure. In the automotive industry, ferrite permanent-magnet motors are commonly used in the dozens of small motors used in each automobile. Greater torque requirements are placed on some applications such as starting motors in a car or truck, so that neodymium-iron-boron permanent magnets, which provide much higher magnetic flux densities and have higher coercive forces, are used in these and other more demanding applications.

In modern practice, conventional dc motors are being replaced by so-called brushless dc motors (see Sec. 6.4.3). These are really synchronous motors with a permanent-magnet field structure. They are driven by variable ac polyphase sources, derived from solid-state supplies powered by a dc power source (or rectified ac sources). Total speed control over the entire range from starting to full speed is achieved by controlling the frequency.

6. ELECTRONIC CONTROL OF ELECTRICAL MACHINES

When supplied from conventional power sources, characterized by constant frequency and nearly constant voltage, electrical machines operate with their characteristic speed-torque curves. For all but series-connected dc machines (Sec. 4.2) and the ac equivalent, the universal motor, this results in very nearly constant-speed operation.

Adjustable-speed operation requires, as a minimum, the ability to control voltage. For efficient control of induction machines and for synchronous machines, coordinated voltage and frequency control is necessary. Because of the practical problems associated with control of frequency, traditional adjustable-speed drives utilized dc machines. However, advances in power electronics have made variable-frequency inverters quite feasible, and adjustable-speed ac drives are now readily available.

6.1 Electronic Power Converters for Electric Drives

The size and cost of all electronic power converters are strongly affected by both their continuous and short-time current capacities, which must therefore be carefully chosen and limited in the design process. An important characteristic of all electronic power converters is, therefore, very limited current capacity compared to conventional power sources. In all cases the controller must be designed to avoid the typical high inrush currents associated with normal "line starting" of motors. Current-limiting circuits and instantaneous electronic trip features are necessary to avoid damage to the electronic converter switching devices.

6.1.1 Controlled Rectifiers. Controlled rectifiers are used to provide variable dc voltage for motor control from a single-phase or three-phase ac supply. Thyristors (silicon-controlled rectifiers) are generally favored for higher power levels because of their high current and voltage ratings and low cost despite the attendant need for relatively complex timing and control circuits. Figure 13 shows a very good form of controlled bridge rectifier, called a dual bridge, capable of full four-quadrant operation. The circuit consists of two full bridges, one to handle forward (motor) current and one to handle reverse (regenerating) current. Avoiding having both bridges on simultaneously is imperative since this would result in a three-phase input fault. Reactors are often employed to limit fault current and allow recovery without shutdown. Because of control complexity and reliability considerations, the dual bridge is only used when necessitated by regeneration requirements. More typically only one bridge is used, and limited regeneration is allowed by employing a switched resistor on the dc bus. Other simplifications include semiconverters using three diodes in place of three of the thyristors, half-wave converters using only three thyristors, and single-phase versions of the various alternatives. There are also more sophisticated ac-to-dc converters (rectifiers) using gate turnoff devices such as transistors or GTOs (gate turnoff thyristors) to provide some control of the input ac wave form and power factor, but these are not currently widely used and will not be further described here.

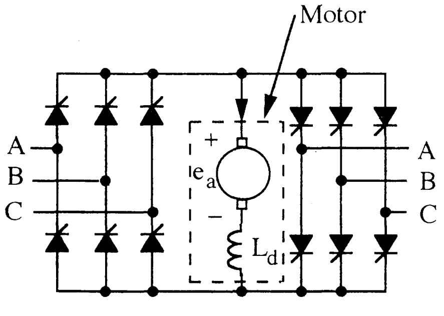

FIG. 13. Dual-bridge three-phase controlled rectifier.

When thyristors are used, the output voltage of a controlled rectifier is determined by the phase delay angle α of the thyristors. This angle is measured between the actual instant of initiation of conduction and the instant of conduction if the thyristor were replaced by a diode. Both positive and negative output voltage is available from a single bridge, and for the three-phase full bridge the voltage can be expressed as

$$V_{\mathrm{dc}}=(3\sqrt{6}/\pi)V_{\phi}\cos\alpha, \tag{28}$$

where V_{ϕ} is the rms phase voltage. There is also an equivalent resistance, related to the ac-side line inductance, which causes a reduction in output voltage dependent on the dc load current.

6.1.2 Choppers. Power converters used to convert a constant-amplitude dc source to variable amplitude are called choppers. The *H* bridge illustrated in Fig. 14 is a four-quadrant chopper. The switches are operated in pairs, 1 and 4 yielding a positive output, 2 and 3 a negative output, and 1 and 3 or 2 and 4 a zero output. The variable voltage output is obtained by pulse-width modulation (PWM) with the output wave form consisting of variable-width voltage pulses of the desired polarity separated by segments of zero voltage. There are simpler variants that do not offer four-quadrant operation. When combined with a simple diode rectifier, choppers offer an alternative to controlled rectifiers for ac to variable-voltage dc conversion.

6.1.3 Phase Control of ac Voltage. Variable ac voltage for motor control is accomplished by utilizing back-to-back connected thyristors or TRIACS, as is done in the dimmer control of lighting. Figure 15 illustrates the scheme and a typical wave form in a three-phase system. Voltage control is obtained by varying the phase delay angle much as in a phase-controlled rectifier.

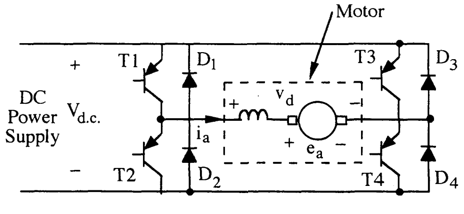

FIG. 14. Four-quadrant *H*-bridge chopper using transistors as the switching elements.

6.1.4 Inverters. Inverters are used to convert from dc to variable-frequency ac using either six-step operation (peak ac amplitude equal to the dc amplitude) or PWM operation (variable ac amplitude). At present, essentially all ac drives employ an inverter and hence require a dc supply. For ac drives operating from an ac power source, the resulting dc link system uses a simple rectifier (for PWM operation), or one of the various variable-voltage dc converters described previously, as the input converter. Figure 16 illustrates the two basic types of circuits in common use for three-phase systems.

The two circuits can be distinguished by the presence of the inverse connected diodes and the dc bus capacitor in the VSI and the dc link inductor in the CSI. Both switch blocks of the input dc to the output to create a three-phase output of rectangular wave form. The two systems differ in that the voltage source inverter (VSI) operates with three switches on at each instant and creates a direct instantaneous correspondence between dc voltage amplitude and the peak output ac voltage. In contrast, the current source inverter (CSI) only has two switches on at each instant and produces instantaneous correspondence between the dc current amplitude and the peak output ac current. The two systems are mathematical duals. Because of cost and certain technical limitations, the CSI is presently used only in very large sizes employing thyristors, and primarily with synchronous machines where leading power factor operation can be used to attain self-commutation. Some retrofit applications to large induction motors with added power factor correction capacitors also use thyristor-based CSIs.

The line-to-line output wave form of the VSI for six-step operation is a rectangular wave with two periods of zero voltage each of a duration of one-sixth of a cycle as shown in Fig. 17. The peak value is equal to the dc-side voltage resulting in a fundamental output voltage of

$$V_L=(\sqrt{6}/\pi)V_{\mathrm{dc}}, \tag{29}$$

where V_L is the fundamental rms line voltage. The line-to-neutral wave form for the six-step VSI has a peak amplitude of two-thirds of the

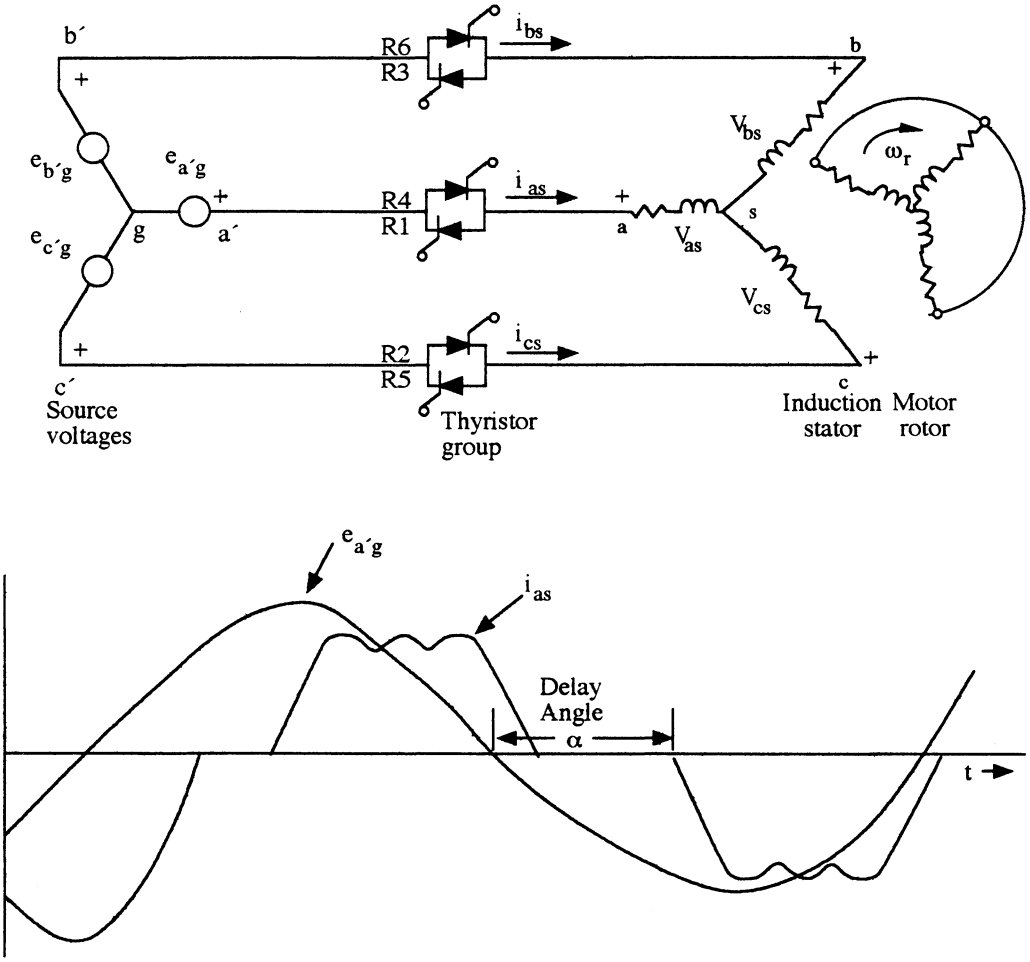
b′
+
$e_{b'g}$
$e_{a'g}$
+
a′
g
$e_{c'g}$
+
c′
Source voltages
R6
R3
R4
R1
R2
R5
i_{bs}
i_{as}
i_{cs}
Thyristor group
b
+
V_{bs}
+
a
V_{as}
s
V_{cs}
+
c
ω_r
Induction Motor
stator
rotor
$e_{a'g}$
i_{as}
Delay Angle
α
t

FIG. 15. Thyristor phase control of motor voltage.

dc voltage amplitude as shown in Fig. 17 and has a fundamental component of

$$V_\phi = (\sqrt{2}/\pi)V_{dc}, \tag{30}$$

where V_ϕ is the fundamental rms phase voltage. The wave forms for the CSI are also shown in Fig. 17, and the values of the fundamental components of the line and phase currents (delta connection) are given by Eqs. (29) and (30) with I replacing V. These rectangular waves have only odd harmonics without any triple multiples of the fundamentals, and the harmonic amplitudes are inversely related to harmonic order. Only the fundamental produces useful output torque; the harmonics create torque pulsations and extra heating of the motor. Control of the output frequency is obtained by varying the stepping rate of the switches. The required voltage control can be accomplished by controlling the dc-side voltage or, now more commonly, by PWM operation of the switches as in the dc chopper. Both types of inverters are fully bilateral, allowing power flow in either direction. The VSI requires reversal of the dc current and thus requires a dual converter at the input to the dc link. In contrast the CSI regenerates by reversing the dc voltage and only requires one input-controlled rectifier bridge.

6.2 dc Drives

Solving Eq. (16) for the speed ω yields

$$\omega = (V_0 - R_A I_A)/K\phi, \tag{31}$$

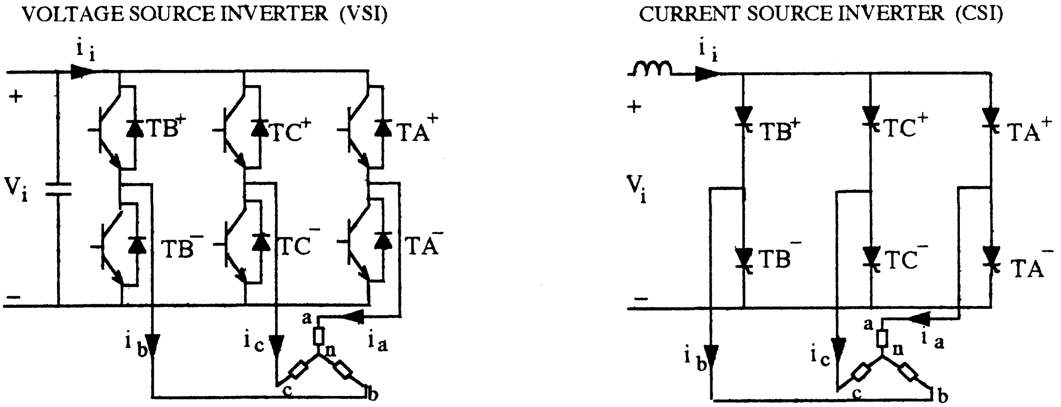

FIG. 16. Power-circuit configuration of VSI and CSI inverters.

which indicates that the dc-motor linear speed–torque characteristic can be controlled by

1. varying the armature voltage V_0, which controls the zero-speed intercept,
2. varying the armature resistance R_A, which controls the slope, or
3. varying the field flux ϕ, which controls both the slope and the zero-speed intercept.

Armature-resistance speed control is very inefficient and yields poor speed regulation and is therefore seldom used except in very small motors. Field flux control is the only way to attain operation above rated speed without exceeding rated voltage and is widely used in combination with other techniques to yield high-speed operation. It is also the only speed control method that does not require handling the full motor output power or current. It is, however, only useful at high speeds since it reduces the torque per ampere and thus the available output power unless the speed is high.

The simplest and most common electronic dc motor drive utilizes armature voltage control with a separate field supply (or permanent-magnet field). Before power electronic converters were available, the variable armature voltage was supplied by a dc generator with field control (Ward-Leonard system). Replacing the generator with an electronic converter provides much faster response and improved efficiency. However, the inherent regeneration capability of the dc generator requires a dual-bridge converter to provide equivalent performance. Also, the system isolation and input power-factor control provided by employing an ac synchronous motor to drive the dc generator are difficult to emulate in the electronic version of the system.

Power level, required control functions, and cost constraints are the primary factors that dictate selection of the type of converter used. Generally, high-power drives require thyristor-based converters and thus tend to be controlled-rectifier systems. Low-cost drives also are generally controlled-rectifier based using semiconverters or half-wave systems.

6.2.1 Capability Curves. The "capability curve" of a dc machine shows the available torque as a function of speed when the arma-

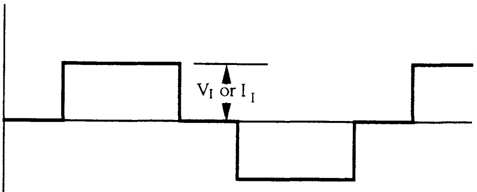

VSI Line Voltage or CSI Line Current

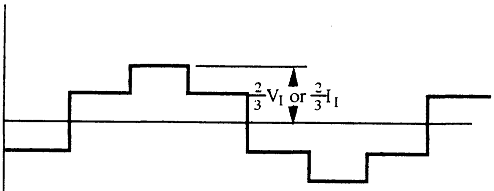

VSI Phase Voltage (Wye-Connection)
or
CSI Phase Current (Delta-Connection)

FIG. 17. Wave forms for six-step VSI and CSI inverters.

ture voltage V_0 and current I_A are limited to their rated values. As shown in Fig. 18, the total speed range is divided into two parts: the range from zero to the rated speed, where armature voltage control with full field current is employed, and the range from rated speed to the maximum allowable speed, where operation is with full-rated armature voltage and steadily reducing field current to keep the back emf e_g smaller than the terminal voltage V_0. The armature voltage control region below rated speed is called the "constant torque region" since the machine is *capable* of full-rated torque at any speed in this region. Above rated speed the torque *capability* is reduced since the field flux is smaller. This region is called the "constant (horse)power region" since with both the voltage and current held at their limit values, the power capability is constant. The term "field weakening region" is also used to describe this mode of operation at high speeds.

6.2.2 High-Dynamic-Performance dc Drives (Servo Drives). When fast response to speed or torque commands is required, a chopper-controlled system is employed except at very high power levels. This avoids the inherent dead time associated with controlled rectifiers. In high-power drives using controlled rectifiers this dead time can be reduced by employing multiple rectifier bridges phased to produce more pulses per input frequency cycle.

High-performance drives nearly always utilize an inner armature-current control loop to provide current regulation as shown in Fig. 19. With controlled current the dc motor becomes a controlled torque source since the torque and armature current are proportional when the field flux is constant. This makes the overall drive performance independent of the armature resistance, inductance, and back emf since these are all compensated by the action of the current loop.

There is an upper speed limit beyond which there is insufficient source voltage to overcome the back emf and maintain current control. Reducing the field flux (field weakening) reduces the back emf at any speed and thus extends the maximum available speed for the drive. Avoiding overcurrent in the converter is easily accomplished by simply limiting the command to the current regulator. This torque command signal is typically derived from the output of the outer loop regulator controlling speed, position, or other physical variables.

6.2.3 Open-Loop dc Drives. In applications where response time and/or precise control of speed are less important, the inner

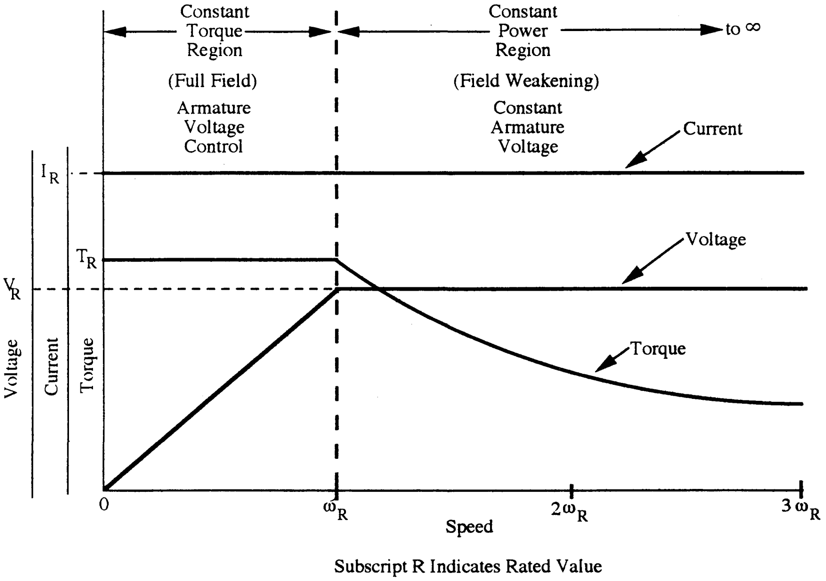

FIG. 18. Capability curves of a dc motor.

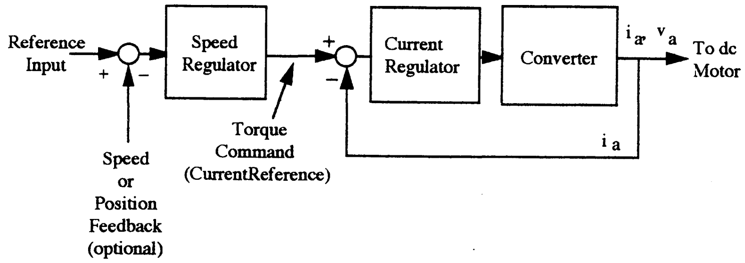

FIG. 19. High-Performance dc servo drive schematic block diagram.

current loop and sometimes even the outer speed feedback loop are eliminated. To avoid overcurrent resulting from the low dynamic impedance of the motor, an electronic current limit, as illustrated in Fig. 20, and, often, a slow ramping of the speed command that directly controls the converter output voltage are employed. The speed command can be the output of a regulator fed by speed error or simply an open-loop command. In this latter case, there is often an *IR*-drop compensator, which partially corrects for the speed droop caused by armature and converter resistance voltage drop. In most cases these simple, open-loop drives utilize some form of rectifier converter, often of the semiconverter of half-wave type.

6.3 Variable-Speed ac Drives

Unlike dc drives where there are only a few basic types, there are a large number of viable ac drive types. This variety stems from the many possible combinations of induction and synchronous machines, VSI and CSI inverters, and speed-control strategies. Only the most commonly used systems are described here.

6.3.1 Synchronous-Motor Speed Control. The only mechanism for speed control in synchronous machines is varying the frequency, which directly varies the speed of the rotating magnetic field [Eq. (21a)]. To avoid changing the flux density in the iron core, and particularly saturation, changes in frequency are typically accompanied by proportional changes in the applied voltage [Eq. (6)] leading to the term "constant-volts/hertz control" used to describe this mode of operation. When operated with constant volts/hertz and constant field current (and neglecting the armature resistance *R*; see Fig. 12), a synchronous machine develops the same torque at the same torque angle δ and the same armature current I_s at any frequency f. This very desirable result occurs because the voltage V_t, the induced voltage E_f, and the reactance drop I_sX_s in Fig. 12 are all linearly proportional to the frequency, and hence the phasor diagram relating the three quantities simply shrinks or grows as the frequency varies without altering the angle δ. Under these conditions the torque, given by the power divided by the speed, is [from Eq. (26)]

$$T=(V_tE_f/\omega X_s)\sin\delta \qquad (32)$$

and is constant. At low frequencies the armature resistance alters this simple relation and the torque is reduced unless the voltage V_t is increased above the constant-volts/hertz value (referred to as "voltage boost").

Above rated frequency the voltage is typically held constant at its rated value, and the flux is allowed to decrease [Eq. (6)]. The power capability can be maintained at the rated value for these higher frequencies if the

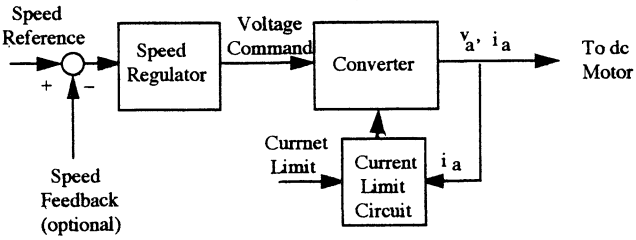

FIG. 20. Adjustable–speed dc drive schematic block diagram.

field current is appropriately reduced so as to maintain the power factor angle between V_t and I_s at a constant value. The synchronous machine thus has a "capability curve" identical to that of the dc machine illustrated in Fig. 18.

6.3.2 Induction-Motor Speed Control. Speed control in induction machines can be attained by varying the frequency and voltage as in the synchronous machine and, to a much more limited extent, by varying only the voltage.

6.3.2.1 Voltage-Controlled Induction-Motor Drives. Voltage variation alters the torque produced at any specific slip in proportion to the voltage squared and thus produces a family of torque–speed curves as illustrated in Fig. 21. Depending on the load torque–speed curve, the equilibrium point where motor torque and load torque are equal will vary with voltage, yielding speed control over a limited speed range. Since the synchronous speed is constant, speed control by voltage variation is really slip control, which results in high rotor I^2R loss as the speed is reduced.

Voltage control is therefore used only for speed control of small induction motors with fan loads in which the torque drops rapidly with speed (as speed squared). These drives are inherently inefficient since they require high-slip operation and are viable only because the rapid drop in torque makes the high percentage loss an acceptably small absolute number. Typically the voltage is controlled by varying the amount of phase delay in series connected TRIACS or back-to-back connected thyristors. This introduces additional losses as a result of the time harmonics in the voltage wave form. Phase control of voltage is also widely used in speed control of single-phase universal motors as in, for example, portable electric drills.

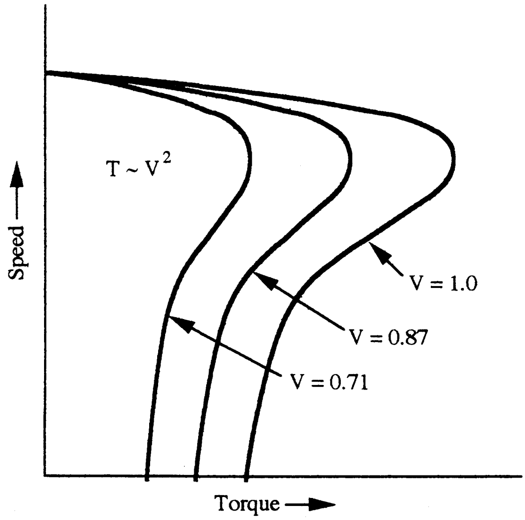

FIG. 21. Torque–speed curves of induction motor for voltage control.

Voltage control, typically via series-connected thyristors in which initiation of conduction is delayed by phase control of the gate voltage, is also used to limit starting current in induction and synchronous motors and for light-load loss reduction in motors subject to variable load cycles. This last application is limited to motors that operate at very light load for long periods, since the voltage controller creates its own loss at all loads and this must be overcome by the reduction of motor core losses at light loads.

6.3.2.2 Adjustable-Speed, Constant-Volts/Hertz Induction-Motor Drives. As in the case of synchronous-machine drives, the synchronous speed of an induction motor can be varied by varying the frequency. Maintaining constant motor flux density to avoid saturation is again accomplished by keeping the ratio of voltage and frequency (volts/hertz) constant. Under this condition the induction machine exhibits essentially the same speed–torque characteristic at all frequencies. This is readily shown by rewriting the resistance R_r'/s in Fig. 8 as $R_r'/s\omega$ times ω and observing that, except for R_s, for a fixed value of $s\omega$ (slip frequency) every element in the circuit is a linear function of ω. By making the voltage vary linearly with ω, the current and torque are independent of frequency for any fixed value of $s\omega$. At very low frequencies (below 5 Hz) the stator resistance becomes significant, and voltage boost (increased volts/hertz) is employed to maintain torque capability. The result is a family of torque–speed curves, as illustrated in Fig. 22, where the various curves are simply translated up or down as the frequency is varied. Operation above base frequency at constant output voltage (and increased values of $s\omega$) allows the flux to decrease and corresponds approximately to the constant-power, field-weakening region of dc drives as illustrated in Fig. 22. The most common system at low and moderate power levels employs a PWM VSI as the variable-

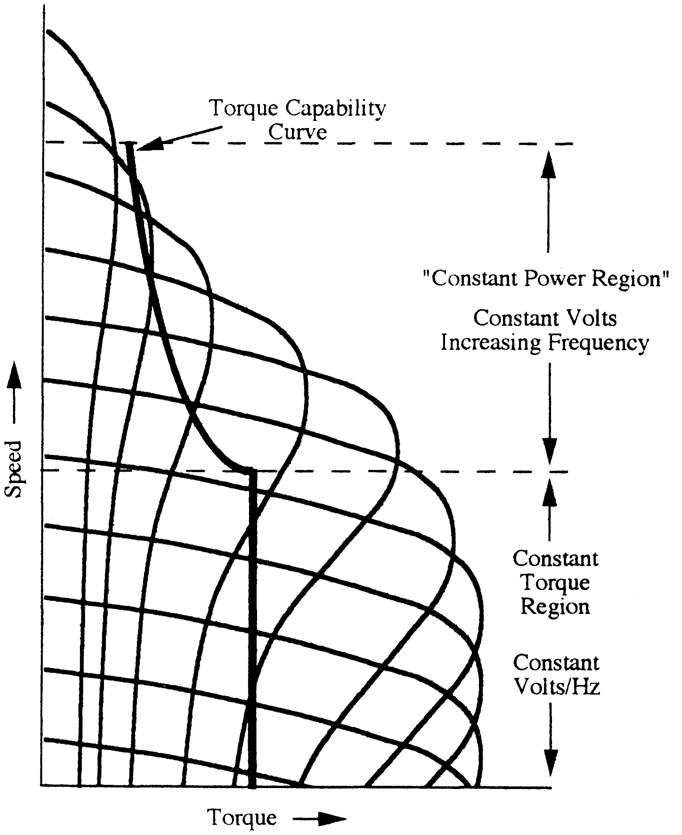

FIG. 22. Capability curves of induction motor for frequency control.

frequency supply. At higher power levels, six-step VSI or CSI inverters fed by controlled rectifiers are used. In the CSI case a voltage feedback loop is necessary to convert the CSI output to a controlled, constant-volts/hertz supply.

Motor starting is accomplished by initially applying a low frequency and slowly ramping up to the desired speed. This is necessary to keep the current within the converter rating. Unfortunately, the proper ramp rate and voltage boost depend on machine parameters and load characteristics, and effective converter utilization during starting is difficult to attain. A variety of control strategies utilizing internal converter variables such as output voltage or current, link dc current, or link power are used to assist under starting conditions or during load or speed changes. However, in the absence of feedback from the motor (speed or flux) these open-loop constant-volts/hertz drives have the fundamental problem of basically appearing as a voltage source to the machine and thus having to cope with the normal constant-voltage properties of the machine.

6.4 Adjustable-Torque, Vector-Controlled ac Drives

Torque control in ac motors is accomplished by emulating the regulated–armature-current dc motor drive (see Sec. 6.2.2). This requires controlling the spatial angle between field flux and armature mmf (leading to the name "field orientation"), which in turn requires control of both amplitude and phase of the ac excitation and leads to the name "vector control."

6.4.1 ac Current Regulators. Most torque controllers use an ac current regulator based on time-domain current feedback and a PWM inverter as shown in Fig. 23, although CSI inverters are also used at high power levels. A variety of schemes for generating the inverter voltage commands in the current regulator

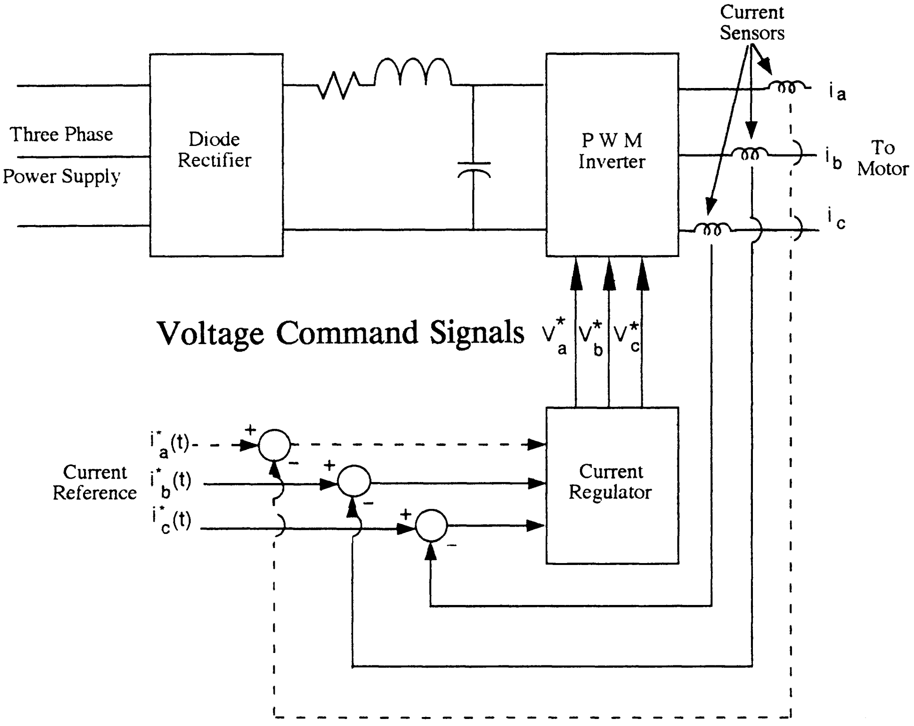

FIG. 23. Current-regulated PWM inverter for "vector control" of ac current.

block of Fig. 23 are employed, with simple on–off control with hysteresis and sine-triangle comparison (natural sampled PWM; see POWER ELECTRONICS) being the most common. The ac current-regulated inverter of Fig. 23 (or its CSI equivalent) allows positioning the machine armature (stator) mmf in space as well as eliminating the effects of armature (stator) resistance, inductance, and back emf as in the dc machine.

6.4.2 Field Orientation. The location of the field flux can be directly determined by measuring rotor position in a synchronous machine but must be estimated from indirect measurements in induction machines. The most common method in induction machines uses rotor position measurement and an estimate of slip frequency to estimate rotor flux position indirectly (indirect field orientation). As illustrated in Fig. 24, the slip estimation depends on the rotor time constant and hence is sensitive to operating conditions. The equation for the slip frequency used in the slip calculator block in Fig. 24 is

$$s\omega = \frac{R_r'}{L_r}\frac{I_T}{I_\phi} = \frac{1}{T_r}\frac{I_T}{I_\phi}, \tag{33}$$

where L_r is the total rotor self inductance, T_r the rotor (open circuit) time constant, I_T the torque command current, and I_ϕ the flux command current. Errors in the time constant cause steady-state flux and torque errors as well as noninstantaneous, usually oscillatory torque response. The sensitivity of the controller to errors in the time constant depends on the inductance-to-resistance ratio of the machine and is therefore more serious in large, high-efficiency machines. In small machines no adaptation to time-constant changes is usually necessary. For large machines a variety of adaptive controllers have been proposed, but the problem remains an area of active research.

For a synchronous machine the slip calculator in Fig. 24 is eliminated since rotor position and rotor flux position are identical (i.e., the slip frequency is zero). Direct field orientation in induction machines, in which the rotor flux position is estimated from flux

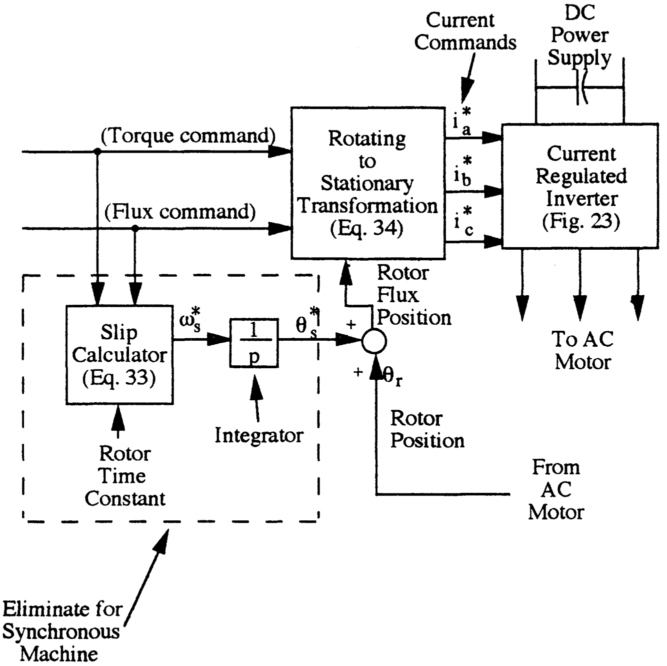
Current Commands
DC Power Supply
(Torque command)
(Flux command)
Rotating to Stationary Transformation (Eq. 34)
i_a^*
i_b^*
i_c^*
Current Regulated Inverter (Fig. 23)
Rotor Flux Position
To AC Motor
Slip Calculator (Eq. 33)
ω_s^*
$\frac{1}{p}$
θ_s^*
+
+
θ_r
Integrator
Rotor Time Constant
Rotor Position
From AC Motor
Eliminate for Synchronous Machine

FIG. 24. Torque-controlled ac drive schematic diagram.

estimates obtained from flux coils or from armature (stator) voltage and current, is also used, but is much less common.

Torque-controlled ac drives based on field orientation are fully equivalent to torque-controlled dc servo drives. They offer the advantages of higher torque/inertia ratios, higher operating speed, and lower maintenance since there is no mechanical commutator. These drives can also be built in larger sizes than for dc machines by employing CSI inverters and synchronous motors (drives of several thousand horsepower are not uncommon) and for higher speeds since the limitations of a mechanical commutator are eliminated.

6.4.3 Brushless dc Machines. There are a large number of variants of the torque-controlled synchronous machine described in Sec. 6.4.2 and illustrated in Fig. 24 that are called brushless dc motors. They all follow the basic principle of controlling the spatial angle between flux and mmf, and all utilize some form of rotor position measurement. Most use permanent magnets for field excitation and some form of simplified position sensing using Hall-effect sensors or position estimation based on back emf. In some cases the motor back emf is intentionally caused to be trapezoidal in wave form to allow effective use of limited position-sensing accuracy and to increase the torque per ampere. The most common limitation of many of these machines is high torque ripple.

6.5 Other ac Drives

There are a number of other, less widely applied, ac drives utilizing special types of converters or special motors. For example, very high-power, low-speed applications sometimes employ cycloconverters for direct conversion of line frequency (50 or 60 Hz) to the low frequency needed to drive the machine without an intermediate dc link. Such systems are limited to frequencies below about one-third of the input frequency and utilize a large number of power-switching devices (usually thyristors) with 36 or more devices being common. Because of the complexity of controlling the switching of such systems, they are typically only used at power levels where other types of converters are unavailable because of power-switching–device limitations (for example, for large, high-voltage motors).

Another, somewhat more common, example is the use of a wound-rotor induction motor with a frequency converter feeding the rotor circuits in high-power applications requiring only a small speed-control range (20% to 40% variation). Here the motivation is to be able to use a lower power rating in the converter since it only needs to be rated to handle the slip power (i.e., 20% of the motor rating if only a ±20% speed range about synchronous speed is required), though this advantage must be balanced against the more expensive machine construction.

GLOSSARY

ac: Alternating current—a power source that alternates sinusoidally with time.

Alternator: A device to change mechanical energy to ac electrical energy.

Armature: The member of a machine in which the energy conversion takes place.

Brushless dc Motor: Synchronous motor operating on a dc supply which is converted to ac of variable frequency and voltage by means of electronic controls.

Commutator: Copper bars insulated from each other, mounted on the armature shaft, and connected to armature windings to carry currents, through brushes riding on the smooth cylindrical outer surface of the commutator, to external circuits.

Compound Motor or Generator: A dc motor or generator that has both shunt and series field coils.

Converter: A generic term for a device that alters the nature of electrical energy (for example, changes the frequency).

Diode: A power-switching device which allows current to pass in only one direction.

dc: Direct current—a power source that has a constant polarity of voltage (or current).

Four-quadrant drive: A drive capable of operating the driven machine both as a motor and as a generator and in either direction.

Generator: A device to change mechanical to electrical energy.

Hall-Effect Sensor: A magnetic-field–sensitive semiconductor device used to sense rotor position in electric drives.

Induction Motor: An ac motor in which rotor currents are set up by magnetic induction (often in a squirrel-cage conductor arrangement).

Inverter: A device for converting dc to ac.

Linear Motor: An electric motor providing linear motion rather than rotation.

Motor: A device to change electrical energy to mechanical energy.

Rectifier: A device for converting ac to dc.

Rotor: The rotating member of an ac machine.

Semiconverter: A controlled rectifier in which half of the thyristors are replaced with diodes.

Series Motor or Generator: A dc motor or generator that has the field coils connected in series with the armature.

Shunt Motor or Generator: A dc motor or generator which has the field coils connected in parallel with the armature.

Synchronous Motor or Alternator: An ac motor or alternator that runs at a constant speed dictated by the frequency of the power supply and the number of pole pairs of the machine.

Thyristor: A power-switching device which can be turned on with a small control signal but must have its voltage reversed to be turned off.

TRIAC: A bidirectional device equivalent to two thyristors connected in inverse parallel.

Universal Motor: A motor that will operate on ac or dc.

Volts/Hertz Control: A control strategy that maintains approximately constant magnetic flux in ac machines.

Wound-Rotor Induction Motor: An induction motor in which the rotor circuits are windings as in the stator and are made accessible by slip rings.

Further Reading

Bose, B. K. (1986), *Power Electronics and AC Drives*, Englewood Cliffs: Prentice-Hall.

Fitzgerald, A. E., Kingsley, C., Umans, S. D. (1990), *Electric Machinery*, 5th ed., New York: McGraw-Hill.

Jordan, Howard E. (1983), *Energy Efficient Electric Motors and Their Application*, New York: Van Nostrand-Reinhold.

Krause, P. C. (1986), *Analysis of Electric Machinery*, New York: McGraw-Hill.

Miller, T. J. E. (1989), *Brushless Permanent-Magnet and Reluctance Motor Drives*, New York: Oxford University Press.

Mohan, Ned, Undeland, Tore M., Robbins, William P. (1989), *Power Electronics Converters, Applications, and Design*, New York: Wiley.

Murphy, J. M. D. and Turnbull, F. G. (1988), *Power Electronic Control of AC Motors*, New York: Pergamon.

Nasar, S. A. (1984), *Electric Machines and Transformers*, New York: Macmillan.

Nasar, S. A. (1987), *Handbook of Electric Machines*, New York: McGraw-Hill.

Say, M. G. (1983), *Alternating Current Machines*, 5th ed., London: Pitman.

Shepherd, W., Hulley, L. N. (1987), *Power Electronics and Motor Control*, Cambridge, U.K.: Cambridge Univ. Press.

Slemon, G. R. (1953), "Equivalent Circuits for Transformers and Machines, Including Non-Linear Effects," Proc. IEE **100**, Pt. 4 (5), 129–143.

Slemon, G. R. (1992), *Electric Machines and Drives*, Reading, MA: Addison-Wesley.

MAGNETIC FILMS

See PLATES AND FILMS, MAGNETIC

MAGNETIC DEVICES

JOSEPH HEREMANS, *General Motors Research Laboratories, Warren, Michigan, U.S.A.*

Introduction 587
1. **Magnetic Field Sensors** 588
1.1 Galvanomagnetic Effects in Semiconductors 588
1.2 Semiconductor Magnetoresistors 590
1.3 Hall Generators 591
1.4 Magnetodiodes 593
1.5 Magnetotransistors 593
1.6 Integrated Semiconductor Magnetic Field Sensors 595
1.7 Ferromagnetic Metal Film Magnetoresistors 595
1.8 Variable-Reluctance Sensors 597
1.9 Fluxgates 598
1.10 Other Magnetometers 599
1.11 Comparison of Magnetic Field Sensors 599
2. **Applications of Magnetic Field Sensors** 600
2.1 Magnetic Recording Heads 600
2.2 Position and Speed Sensors 601
2.3 Rotary Shaft Encoders 601
2.4 Brushless Motors 602
2.5 Shuntless dc Current Measurements 602
3. **High-Frequency Devices** 602
3.1 General Aspects 602
3.2 Microwave Devices 604
3.2.1 Filters 604
3.2.2 Isolators 605
3.2.3 Circulators 606
3.2.4 Strip-Line, MMIC, and Thin-Film Technologies 607
3.3 Other Magnetic Devices 607
3.3.1 Antenna, Inductor, and Transformer Cores 607
3.3.2 Magnetic Memories 608
4. **Conclusion** 609
Acknowledgments 609
Glossary 609
Works Cited 610
Further Reading 611

INTRODUCTION

Iron ferrite or magnetite was "lodestone" to the ancients. Its magnetic properties were known to the thirteenth-century Chinese, as was its property of pointing to the north (Bai Shouyi, 1982). The stone was ground and shaped into a needle, kept aloft inside a bowl of water by a floating wick, and used for guidance. The resulting compass (see Fig. 1) is the first example of a magnetic device. The torque exerted on the compass needle is the result of the action of the Earth's magnetic field on the ferrite. It exemplifies the two elements that define a "magnetic device": an element using the measurement of the intensity or direction of a magnetic field, or an element using a magnetic material.

This article is organized in the following way. The principles of operation of magnetic field-measuring devices, such as magnetoresistors, Hall generators, magnetodiodes, and magnetotransistors, are described, as well as some integrated circuits that use them. Also described are pickup coils, and fluxgates that are used in compasses. A special paragraph is included in which the performance of the different sensors is compared. A second section is devoted to applications of magnetometers. Magnetic field-measuring devices are used in compasses and attitude detectors, in magnetic recording heads (see also MAGNETIC STORAGE OF INFORMATION), in position and speed sensors, in rotary encoders, in contactless potentiometers, in dc current measurements, and in scientific instrumentation.

Microwave devices, such as filters, phase shifters, isolators, and circulators, use ferrimagnetic resonances in ferrites and garnets. Other ferromagnetic or ferrimagnetic materi-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

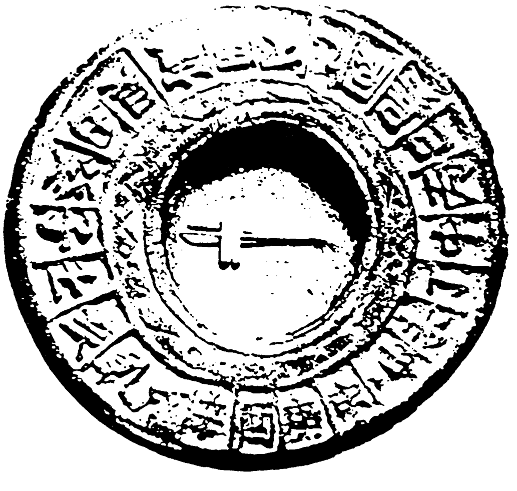

FIG. 1. A Song dynasty compass in the collection of the Museum of Chinese History, Beijing (from Bai Shouyi, 1982).

als are used in antenna cores, inductors, or even simple electromagnets. These devices are described.

Magnetic materials are also used in electrical machinery (see MACHINES, ELECTRICAL), transformers and inductors, thermometry (see MAGNETOTHERMAL DEVICES), and magnetic lenses (see MAGNETOELECTRICAL DEVICES). The reader is referred to the appropriate articles.

Recent developments in microelectronic technology have revolutionized the fabrication techniques of magnetic devices and broadened the range of useful applications. Even pickup coils can be made using thin-film technology. Magnetic field sensors with on-chip integrated amplifiers and signal conditioners are sensitive, reliable, and inexpensive, and have made magnetic position detection a practical reality. Microwave circuit elements, including gallium arsenide (GaAs) amplifiers, are integrated on semiconductor chips called monolithic microwave integrated circuits (MMICs). Soon magnetic circuit elements will be integrated (Prinz, 1990) directly in MMICs, thus lowering their cost and opening new fields of applications.

I have attempted to describe the physical principles of operation of the devices by using simplified schematic outlines. The reader can find more detailed descriptions and mathematical formulation in the references.

1. MAGNETIC FIELD SENSORS

1.1 Galvanomagnetic Effects in Semiconductors

The fundamental principles of galvanomagnetic effects are discussed in the articles ELECTRONIC TRANSPORT IN SOLIDS and TRANSPORT PHENOMENA. In this section I point out the difference between the intrinsic magnetoresistance that many electrically conductive materials exhibit and the geometrical magnetoresistance effects that are the basis of magnetoresistive sensing elements (Heremans, 1993).

Electrons or holes moving under an electric field **E** and a transverse magnetic field **H** are subject to the Lorentz force **F** proportional to the velocity **v** of the carrier and the magnetic induction **B**:

$$\mathbf{F}=\pm e(\mathbf{v}\times\mathbf{B}), \tag{1}$$

where $e=1.6\times10^{-19}$ Coulomb; the + sign is used if the carriers are holes, and the − sign if they are electrons. The Lorentz force is exerted at right angles to both **v** and **B**. As a consequence, the path of these charge carriers, or the current density vector **j**, is not collinear with the applied electric field. The angle θ between **j** and **E** is the Hall angle. Consider the case of a slab of semiconductor, of length l, width W, and thickness t along the x, y, and z directions, respectively, as shown in plan view in Fig. 2. An applied voltage V_{xx} forces a current I through the slab. A magnetic field **H** is applied normal to the plane of the slab, and the induction is **B**. The Lorentz force will distort the current lines close to the injecting contacts. Some distance along the x axis into the slab, there is an accumulation of charge carriers to one side of the slab. This results in the generation of a transverse voltage, the Hall voltage, V_{xy} in Fig. 2. In a slab of arbitrary length and width, the current lines follow the paths shown in Fig. 2. Conformal mapping techniques are used to calculate these lines (Lippmann and Kuhrt, 1958). If the slab is long enough, the transverse, or Hall, electric field balances the Lorentz force, and the current lines become parallel to the side of the sample along the x axis. In an infinitely long slab (not shown in Fig. 2) the electric fields associated with V_{xx} and V_{xy} are

$$E_x=V_{xx}/l, \tag{2}$$

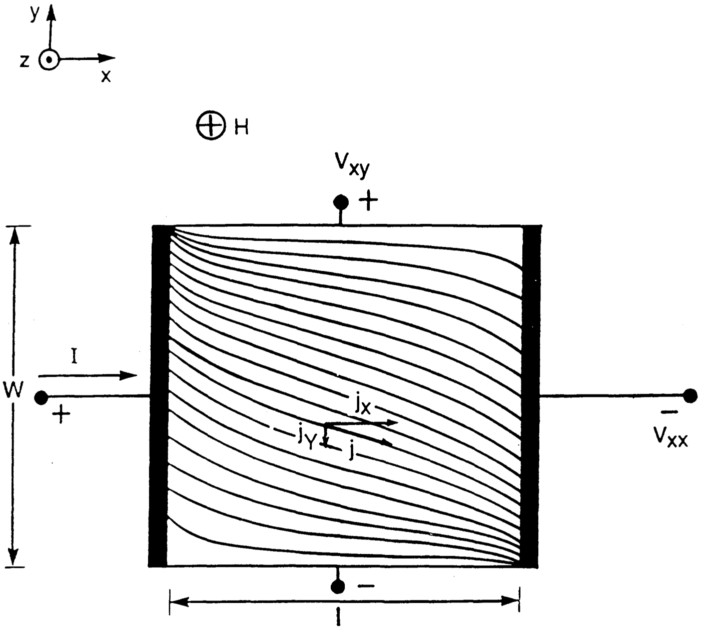

FIG. 2. Projection of a slab of semiconductor with current density (j) lines flowing between the two current electrodes (dark regions) in the presence of a magnetic field H pointing into the plane of the slab. The total current flowing is I, V_{xx} is the longitudinal voltage, and V_{xy} is the transverse, or Hall, voltage.

$$E_y = V_{xy}/W, \tag{3}$$

and the current density is mostly parallel to x:

$$j \simeq j_x \simeq I/Wt. \tag{4}$$

This geometry is commonly used to make resistivity and Hall effect measurements on materials, since these intrinsic material properties are then easily defined. The resistivity is

$$\rho_{xx} = E_x/j = (V_{xx}/I)(Wt/l). \tag{5}$$

The Hall resistivity is

$$\rho_{xy} = E_y/j = (V_{xy}/I)t \tag{6}$$

and the Hall coefficient is

$$R_{\mathrm{H}} = \rho_{xy}/B. \tag{7}$$

In a homogeneous, nonmagnetic material, with a single type of charge carrier, R_{H} is a constant related to the density n of these carriers by

$$R_{\mathrm{H}} = \pm 1/ne, \tag{8}$$

where the + sign is used if the carriers are holes, and the − sign if they are electrons. The resistivity ρ_{xx} is then given by

$$\rho_{xx} = 1/ne\mu, \tag{9}$$

where μ is the carrier mobility. The Hall angle is then θ, where

$$\tan\theta = \mu B. \tag{10}$$

If, furthermore, the carrier's Fermi surface is isotropic, the resistivity ρ_{xx} is independent of B. All the conditions needed for this are almost never met, and most materials have a dependence of ρ_{xx} on B that contains information about the physics of the electrical transport in that material. This is the *intrinsic magnetoresistance* and has a different origin than the geometrical magnetoresistance used in sensors.

The Hall angle can become large even in moderate magnetic fields if the mobility of a semiconductor is very large. In a rectangularly shaped slab of semiconductor, shown in Fig. 3, with $l \ll W$, there is not much length available for the current lines to become parallel to the x axis, and the current lines are similar to what they are near the contacts in Fig. 2. The current path in such a short, wide sample is considerably longer than in the same sample at zero magnetic field. This leads to a large increase in resistance $R = V_{xx}/I$. It is the *geometrical magnetoresistance* effect. The Hall voltage is not fully developed and does not counteract the Lorentz force effectively. The Corbino disk (Fig. 3) is another geometry in which the Hall voltage is shorted out and a large geometrical magnetoresistance is observed. Conformal mapping techniques can show that the Corbino disk is equivalent to an

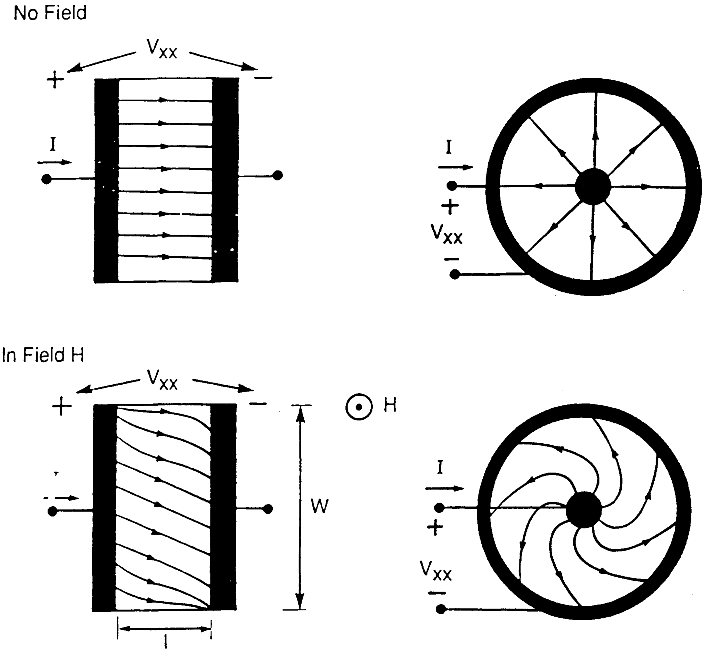

FIG. 3. Elements showing the geometrical magnetoresistance effect. A short, wide rectangle, at left, can be viewed as consisting only of the regions close to the contacts in Fig. 2. In absence of an applied magnetic field *H*, the current lines are normal to the contacts, but in presence of the field, they distort as shown, leading to an increase in current path length and resistance. In the Corbino disk at right, the current lines are radial in the absence of field, but lengthen as shown in the presence of field.

infinitely wide or short rectangular element, in which the Hall voltage is totally shorted out, and the geometrical magnetoresistance is maximal. The resistance of the device is then

$$R(B)=R(B{=}0)\,[1+(\rho_{xy}/\rho_{xx})^2]\times[\rho_{xx}(B)/\rho_{xx}(0)].$$

In practical devices, Corbino disks are not used, because they have a low resistance, which leads to a large power dissipation. Most devices are made from rectangular elements such as shown in Fig. 3 (left). The response of such elements is then given by

$$R(B)=R(B{=}0)\,[1+g(W/l)(\rho_{xy}/\rho_{xx})^2]\times[\rho_{xx}(B)/\rho_{xx}(0)] \tag{11}$$

or, under the conditions under which Eqs. (8) and (9) hold, and for small Hall angles $\theta<\frac{1}{8}\pi$, or $\rho_{xy}/\rho_{xx}<0.2$,

$$R(B)=R(B{=}0)\,[1+g(W/l)\mu^2B^2]\times[\rho_{xx}(B)/\rho_{xx}(0)]. \tag{12}$$

In these equations, the factor $\rho_{xx}(B)/\rho_{xx}(0)$ accounts for the intrinsic magnetoresistance, while the factor $g(W/l)$ accounts for the geometrical influence of the width-to-length ratio and is given in Fig. 4.

1.2 Semiconductor Magnetoresistors

Practical magnetoresistive devices are made out of high-mobility semiconductors, such as *n*-type InSb or InAs, because the device sensitivity is essentially a quadratic function of mobility [see Eq. (12)]. Unfortunately, these are also narrow-gap semiconductors, in which the carrier density is a strong function of temperature at room temperature. To compensate for this, manufacturers dope the semiconductors *n*-type. This technique has two disadvantages:

1. In most semiconductors, the mobility de-

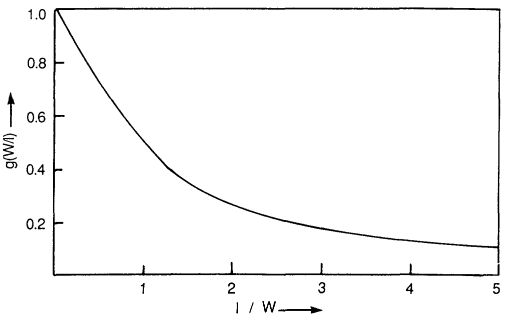

FIG. 4. The geometrical coefficient *g*(*W*/*l*) that relates the geometrical magnetoresistance to the length/width (*l*/*W*) ratio of a rectangular element (Lippmann and Kuhrt, 1958).

creases as the doping level is increased, because of scattering of electrons by ionized impurities, and

2. when both the electron density and the electron mobility are high, the electrical conductivity is also high.

In order to keep current consumption and heat dissipation low, the ideal resistance of a magnetoresistor is in the 500- to 10 000-Ω range. Given the high conductivity of the active semiconductor used, obtaining such high device resistances requires the use of very thin slabs or films of semiconductor. Recent breakthroughs in growing thin InSb and InAs films epitaxially on insulating III-V substrates, by molecular-beam epitaxy and metal-organic chemical vapor deposition (Partin and Heremans, 1992), have made thin-film magnetoresistors very practical. Devices are made by connecting multiple rectangular elements in series (Fig. 5) to increase the resistance. The magnetoresistor shown in Fig. 5 was described by Heremans *et al.* (1993); its resistance is shown in Fig. 6 as a function of applied magnetic field at different temperatures.

The recent use of two-dimensional electron gases (see HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR) in magnetoresistors has made it possible to achieve the best compromise between high electron density and high mobility, while maintaining a reasonably large resistance per device (Heremans *et al.*, 1990). Devices based on two-dimensional electron gases have so far been limited to InAs, and work is under way to extend this to InSb-based magnetoresistors.

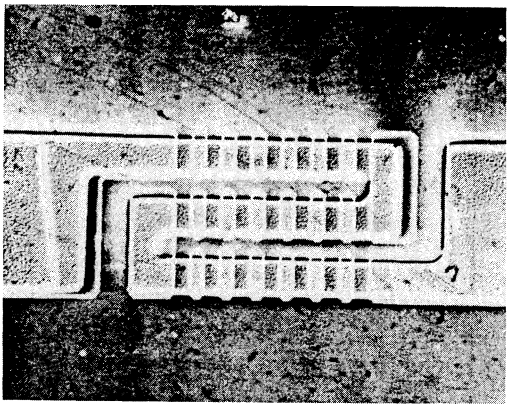

FIG. 5. A multielement magnetoresistor made from an InSb mesa etched from an epitaxial film of InSb on InP. Each rectangular element not covered by the contact metal behaves as one element described in Fig. 3.

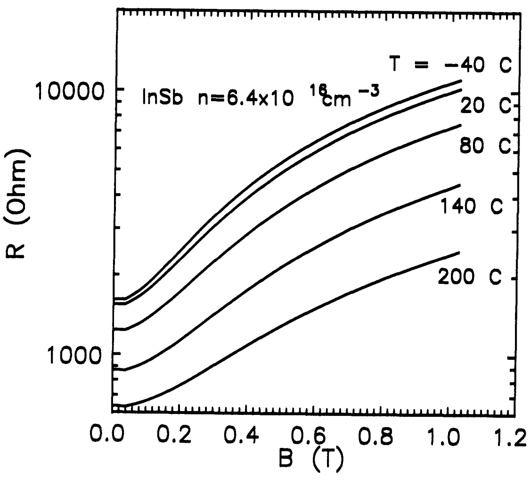

FIG. 6. The resistance of a magnetoresistor made from *n*-type InSb doped at 6.4×10^{16} cm^{-3}, as a function of applied magnetic field at the different temperatures indicated (Heremans *et al.*, 1993).

Siemens AG uses a different manufacturing process based on InSb boules that contain needlelike precipitates of metallic NiSb. The metallic needles serve the same purpose as the lateral contact stripes in the device shown in Fig. 5: they short out the Hall voltage, so that a large magnetoresistance appears. Devices are made from the boules by slicing wafers and lapping them to the desired thickness (Göpel *et al.*, 1989).

1.3 Hall Generators

The Hall voltage itself is commonly used to measure the magnetic field. Hall generators are often made in the geometry shown in Fig. 2. The Hall resistance for an infinitely long slab is defined by

$$R_{xy} = V_{xy}/I = \rho_{xy}/t. \tag{13}$$

The Hall resistance of a generator made from a material for which Eq. (7) holds is

$$R_{xy} = B/net. \tag{14}$$

A four-terminal device analysis (Heremans, 1993) of a Hall device (Fig. 2) shows that the ratio of the power output ($V_{xy}I_{\text{out}}$, where I_{out} is the current drawn from the Hall electrodes) to the power input ($V_{xx}I$) is almost proportional to the mobility of the semiconductor. Equation (14) shows that the output of a Hall generator is linear in the applied magnetic field, and that the most sensitive devices are made from thin, lightly doped slabs of material. Again, the best compromise between sensitivity, temperature stability, and moderate current consumption is obtained when Hall generators are made from thin doped slabs of high-mobility semiconductors. Two-dimensional electron gases are very well suited, but not widely used yet.

Silicon Hall generators (Popovic, 1991) can be made using an inversion layer in a metal-oxide-semiconductor (MOS) structure. Hall devices also can be made by ion-implanting of an *n*-type doped region of the desired geometry into *p*-type or semi-insulating GaAs, or *p*-type Si. Sometimes, *n*-type epitaxial Si, GaAs, or InAs grown on *p*-type or semi-insulating substrates is used. A mesa of the desired geometry is etched into it and equipped with current and voltage contacts. Thin lapped layers of InAs or InSb are sometimes used. Silicon is the preferred semiconductor for room-temperature applications that are very cost sensitive and do not require operation above 10 kHz, because amplifying and signal-processing electronics can be integrated on the same chip as the Hall sensor. *n*-type channels or implanted regions are preferred, since the electron mobility is larger than the hole mobility in Si. In applications in which a substantial current is to be drawn from the sensor's output, InAs- or even InSb-based sensors are used, because of their better power efficiency. InAs and InSb Hall sensors are also preferred in metrology applications or when high signal-to-noise ratios are desired. GaAs is used in some high-temperature applications. Figure 7 shows the output of an ion-implanted GaAs Hall generator (Siemens KSY-10). The ordinate is the Hall resistance, i.e., V_{xy}/I, the ratio of the output voltage to the input current. This device is quite stable with respect to temperature in the range from −40 to 160 °C. GaAs integrated Hall sensors for high-temperature applications are also being developed (Itakura *et al.*, 1989).

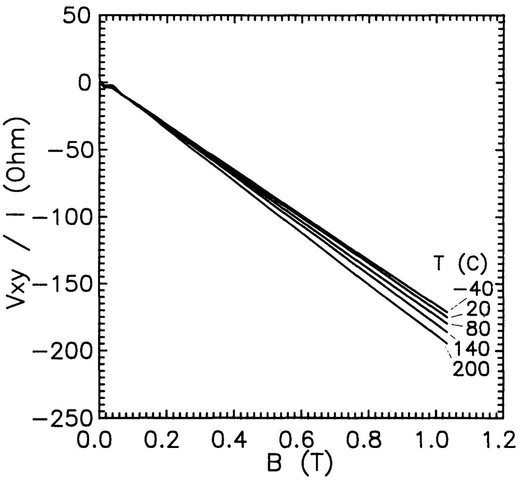

FIG. 7. Ratio of the Hall voltage to the input current, for a GaAs Hall sensor (Siemens KSY-10) made by ion implantation, shown as a function of magnetic field and temperature (after Heremans, 1993).

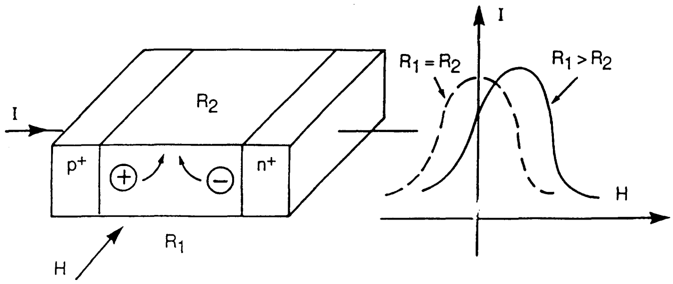

FIG. 8. A magnetodiode, made from a slab of lightly doped semiconductor sandwiched between heavily doped *n* and *p* regions. The surfaces of the lightly doped region have different recombination rates R_1 and R_2. The longitudinal voltage applied to the diode in forward biasing injects both electrons and holes in the lightly doped region. In presence of a magnetic field, the Lorentz force pushes both types of carriers in the same direction. The magnetic field *H* can push the carriers either toward the high-recombination-rate surface, in which case the current decreases, or toward the low-recombination-rate surface, and then the current is increased. The current versus field relation under a given bias is schematically given at right (after Cristoloveanu *et al.*, 1983).

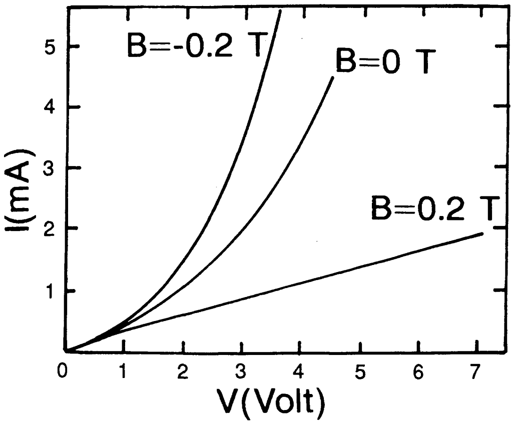

FIG. 9. Magnetic field dependence of the current-voltage characteristics of a Ge magnetodiode (after Yamada, 1968).

1.4 Magnetodiodes

Magnetodiodes (Cristoloveanu *et al.*, 1983) are *p-i-n* diodes, schematically represented in Fig. 8. They consist of an intrinsic region, contacted to a heavily *n*-type region on one side and a heavily *p*-type region on the other. The magnetodiode differs from a simple *p-i-n* diode because one of its lateral faces has been treated to form a surface at which electrons and holes have a high recombination rate R_1. This treatment can be mechanical abrasion or damaging by ion implantation. Some magnetodiodes exploit the high recombination rates at the silicon/sapphire interface. The other face is made as nearly perfect as possible, by etching for instance, to keep the recombination rate R_2 low. When the diode is forward biased, and a current I flows as shown in Fig. 8, electrons are injected from the n^+ electrode into the intrinsic region, and holes from the p^+ region. In the presence of a magnetic field H oriented as shown, the Lorentz force acting on both types of carriers pushes them in the same direction, because the Lorentz force is sensitive to the sign of the charge of the carrier; see Eq. (1). If the carriers are deflected by the magnetic field toward the face with a high recombination rate R_1, they recombine there, and the current flow is impeded. If they are deflected toward the other face, much less recombination takes place and more current flows. Current-voltage characteristics of a Ge-based *p-i-n* diode at room temperature at different field values are shown in Fig. 9. Sensitivities can be up to 1000 times higher than those of Hall generators made from similar materials. Discrete magnetodiodes have been made from Ge, Si, GaAs, silicon-on-sapphire, and CMOS technology (Baltes and Popovic, 1986).

1.5 Magnetotransistors

Many geometries have been used to exploit the influence of the Lorentz force on the flow of charge carriers in transistors. For a more complete review, the reader is referred to Baltes and Popovic (1986). Magnetotransistors can be made as FET transistors with two drains, or as bipolar transistors with two collectors. The applied magnetic field distorts the current lines and favors current flow into one of the two collectors or drains.

A dual-drain Si magnetic FET is shown in Fig. 10. A magnetic field normal to the chip plane causes a deflection of the current lines toward one of the two drains, as is shown in Fig. 11. Complementary magnetic FETs can be made using CMOS technology, and consist of a dual-drain *n*-channel device and a dual-drain *p*-channel device. The gate electrode voltage does not need to be varied during

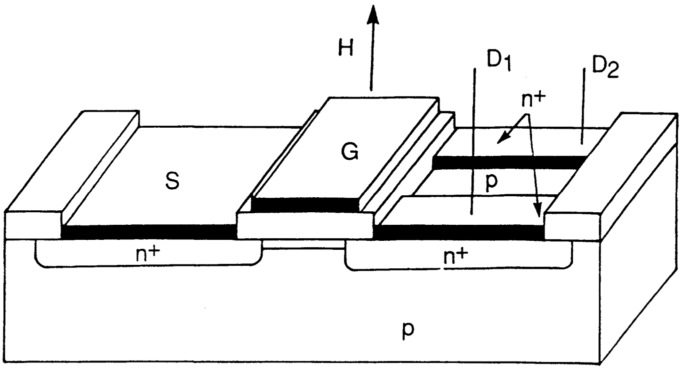

FIG. 10. Two-drain (D_1 and D_2) magnetic FET transistor. A magnetic field H exerts a Lorentz force that pushes the electrons from the source (S) preferentially toward one of the drains. The resulting current imbalance is a measure of the field H. The gate is labeled G (after Baltes and Popovic, 1986).

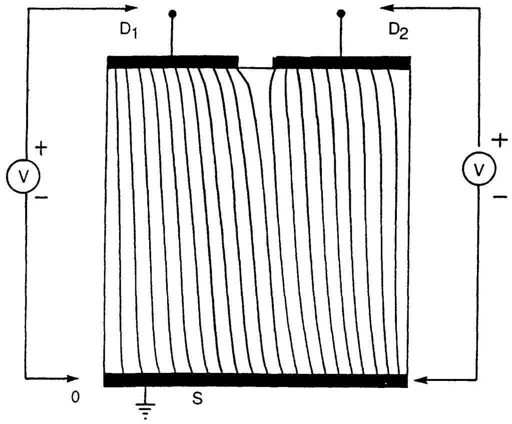

FIG. 11. Current density lines in a two-drain magnetotransistor as shown in Fig. 10 (after Baltes and Popovic, 1986).

operation; its only purpose is to create a channel of the desired surface density. In that sense, these devices are not strictly transistors. InAs tends to have a natural accumulation layer of electrons at its interface with air. Three-terminal magnetic field sensors with geometries similar to those in Fig. 10 can be etched as mesas into epitaxial layers of InAs grown on insulating GaAs, and no gate electrode is necessary (Heremans *et al.*, 1992). Similarly, GaAs/AlAs interfaces with active two-dimensional electron gases have been used by Nathan *et al.* (1990). In dual-drain magnetic FETs, the useful output voltage is measured as a voltage drop between load resistors connected to the drains; see Fig. 12.

Bipolar transistors can be designed to be sensitive to magnetic fields in different ways: some transistors have field-dependent current gains; others are designed with dual collectors. They can be made vertically or horizontally. The magnetic field can be designed to act in three different ways:

1. The Lorentz force can act on the minority carriers in the base region, or in the base-collector depletion layer, or in the low-doped collector region itself.
2. The Hall voltage can be designed to deflect the majority carriers in the base region.
3. The Lorentz force can act on both types of carriers, which then will adjust their concentration to maintain space-charge neutrality at each point.

In general, all three effects coexist to some degree. A *pnp* double-collector magnetotransistor with large sensitivity was described by Popovic and Baltes (1983). A schematic of the device is shown in Fig. 13. The output voltage is the differential voltage obtained between two load resistors connected to the two bases; see Fig. 12. The device in Fig. 13 has an emitter purposely designed with a low efficiency. The magnetic field operates on the majority-carrier current in the base (effect **2**), yielding a larger Hall field than if the Lorentz force were operating on the minority carriers alone. The base region is surrounded by a deep diffusion ring, which confines the base current to a thin slab perpendicular to the direction of the magnetic field.

Magnetotransistors can achieve sensitivities $(dI_C/dB)/I_C$ (I_C is the collector current in Fig. 12) as high as 4 T^{-1}. This seems superior to those of Si-based Hall plates, but such bipolar magnetotransistors require high base currents. When the sensitivities of existing magnetotransistors, bipolar or FET-like, are calculated including the sum of all currents, they are similar to those of Hall generators. Given the large variety of effects involved in

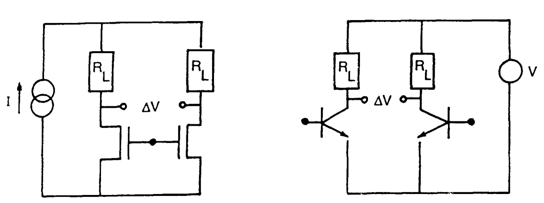

FIG. 12. The output of dual-drain FETs (left) and of bipolar magnetotransistors (right) is a current imbalance, which translates as a voltage ΔV across load resistors R_L in these diagrams.

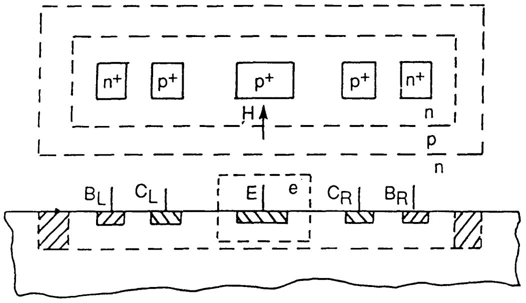

FIG. 13. Double-collector magnetotransistor. The magnetic field *H* modulates the ratio of the currents between the emitter (*E*) and each of the collectors (C_L at left and C_R at right). A diffusion ring (dashed line) confines the current. The base contacts are B_L and B_R (after Popovic and Baltes, 1983).

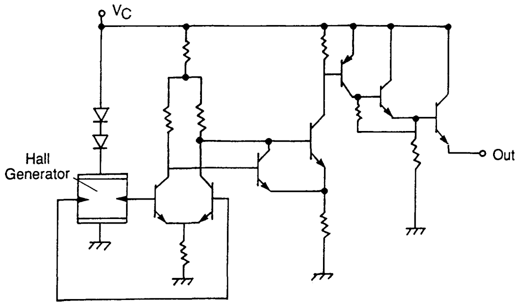

FIG. 14. Schematic diagram for an integrated silicon Hall element with amplifier (after Nakamura and Maenaka, 1990).

magnetodiode and magnetotransistor operation, these devices still have a large potential for optimization (Nathan *et al.*, 1989).

1.6 Integrated Semiconductor Magnetic Field Sensors

Silicon Hall generators are mostly used with on-chip amplification and signal-conditioning electronics. A simple circuit, used in solid-state keyboards (Nakamura and Maenaka, 1990), is shown in Fig. 14. Most integrated magnetic field sensors are rather simple combinations of the sensor with an amplifier and, when hysteresis is desired, with a latching circuit. The voltage output of a Si-based Hall generator is usually low, especially when small current consumption is also a design requirement. This implies that high-gain amplifiers must be used, with limited bandwidth. Hall generators also give a differential output, which often has to be converted to a single-ended output. For high-temperature applications, such as many automotive ones, GaAs sensors with integrated electronics are being developed (Itakura *et al.*, 1989).

InAs- or InSb-based magnetoresistors grown on an insulating GaAs substrate, in which the amplifiers and signal conditioners can be integrated, become very competitive with GaAs Hall generators, because they generate ten times higher output voltage in moderate magnetic fields (see Sec. 1.11, Comparison of Magnetic Field Sensors). Magnetoresistors give a single-ended output voltage with an offset bias.

Magnetotransistors can readily be integrated. Complementary pairs of magnetic FETs can be incorporated into a configuration reminiscent of a CMOS differential amplifier, shown in Fig. 15. Each of the usual two complementary MOS transistor pairs is replaced by the corresponding magnetic FET. Their drains are cross-coupled. When a magnetic field is applied perpendicularly to the plane of the devices, the current in one drain of one transistor increases, while the current in the drain of the other transistor connected to it decreases. The gain built into the circuit is such that small imbalances in the drain currents lead to large output-voltage variations. This circuit converts the differential output of the magnetic FET into a single-ended, amplified signal. The bipolar dual-collector magnetotransistor shown in Fig. 13 can be integrated using standard analog bipolar technology (Maenaka *et al.*, 1990).

1.7 Ferromagnetic Metal Film Magnetoresistors

The electrical resistance of certain ferromagnetic alloys is anisotropic in a magnetic field (see MAGNETIC ORDERING IN SOLIDS; MAGNETIC MATERIALS). The theoretical background

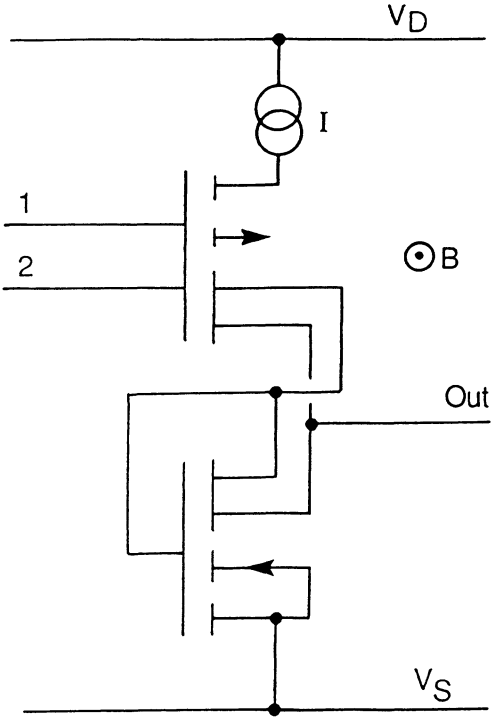

FIG. 15. Two complementary dual-drain magnetic-field effect transistors integrated with a conventional CMOS differential amplifier (after Baltes and Popovic, 1986).

for this effect is summarized by McGuire and Potter (1975). In a film of a ferromagnetic metal thin enough to be single-domain, the magnetic moment vector tends to lie in the plane because of the large demagnetization factor. The electrical resistivity measured when the current-density vector is parallel to the magnetic moment, $\rho_{\parallel}$, is different from the resistivity measured with the current perpendicular to the moment, $\rho_{\perp}$; see Fig. 16. This experimental fact has been explained by an anisotropy in the scattering of the electrons (Campbell *et al.*, 1970).

The anisotropy in the magnetoresistance of Permalloys ($Ni_{1-x}Fe_x$, $x \sim 0.2$) is used to make magnetoresistors (Dibbern, 1986). These devices are made from thin films because the electrical resistance of thin-film devices is higher than that of bulk samples, which keeps the current consumption low. A thin layer can also be made to behave like a single magnetic domain. Its magnetization with no external field applied has one single, well-defined direction in the plane, called the easy axis, also introduced during manufacture. A conceptual diagram of a device is shown in Fig. 17. There is an angle θ between the direction of current flow and the direction of the magnetization **M** along the easy axis. As an external magnetic field **H** is applied to the sample, the direction of the magnetization **M** is rotated away from the easy axis, and the angle θ changes. Since the resistivity along the magnetization direction is $\rho_{\parallel}$, different from that perpendicular to **M**, $\rho_{\perp}$, the total resistance of the film becomes a function $R(H)$ of the applied field H, via the dependence of $\theta(H)$ and $\rho_{\perp}$ on H:

$$R(H) = (l/Wt)[\rho_{\perp} + (\rho_{\parallel} - \rho_{\perp})\cos^2\theta]. \qquad (15)$$

Several techniques exist to linearize the output of these sensors; see Dibbern (1986). Ferromagnetic magnetoresistor sensors are best suited to the measurement of low fields, because they are not sensitive at fields above the saturation field of the layer.

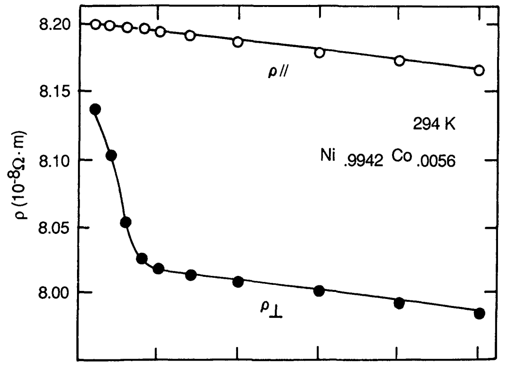

FIG. 16. Magnetoresistivity of Ni-Co at room temperature as a function of field, applied parallel ($\rho_{\parallel}$) or perpendicular ($\rho_{\perp}$) to the easy axis of magnetization (after McGuire and Potter, 1975).

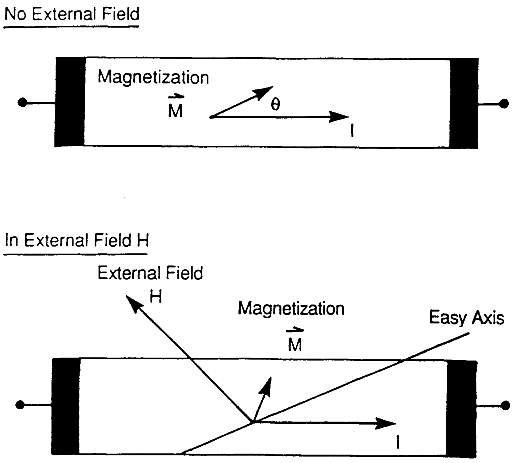

FIG. 17. Magnetoresistor made from a thin film of Permalloy. In the absence of field (top), the current makes an angle θ with respect to the magnetization **M**, which lies along the easy axis. An in-plane applied field **H** (bottom) rotates the magnetization **M**. The anisotropy in the magnetoresistivity shown in Fig. 16, and the fact that the angle between current *I* and moment **M** is influenced by the field, lead to a change in resistance of the strip (adapted from Lenz, 1990).

A recent development, which may force a reassessment of the previous statement, is the discovery of giant negative magnetoresistance in Fe/Cr superlattices (Baibich *et al.*, 1988) consisting of alternating layers of magnetic and nonmagnetic metals. The magnetic moments of adjacent magnetic layers can be spontaneously aligned either ferromagnetically or antiferromagnetically, depending on the thickness of the layers. When the magnetic layers are aligned antiferromagnetically, the resistance of the structure is greater than when they are aligned ferromagnetically, because of the interplay between the spin dependence of the scattering and the magnetic orientation in the structure. An external magnetic field can change the magnetic orientation of the layers with respect to one another, and thus change the resistance of the multilayers (Barthelemy and Fert, 1991). Relative changes in resistance of 50% at applied fields of 2 T have been reported, and are shown in Fig. 18. To date, no publications have appeared describing the use of this effect in sensors, but that is likely to change. The large saturation field H_S for the Fe/Cr system is a disadvantage for sensor applications, but other systems, such as Co/Cu (Mosca *et al.*, 1991), offer better sensitivity. Also recently discovered are giant magnetoresistive effects in metal multilayers that are not magnetically coupled (Dieny *et al.*, 1991), and these too may lead to future devices.

1.8 Variable-Reluctance Sensors

This type of sensor is based on the measurement of the current induced in a coiled conductor, which is proportional to the rate of change of the magnetic flux through the coil (Göpel *et al.*, 1989, Chap. 6). A core of high-permeability magnetic material can be inserted in the coil to gather the surrounding magnetic flux. An example of such a sensor, implemented with thin-film technologies and used in a magnetic recording head, is shown in Fig. 19. The simplest output of such a device is a voltage or a current appearing between the coil end contacts as the magnetic flux through the coil is varied with time. In the example of Fig. 19, a magnetic recording medium passes over the gap (see the next section), and closes the magnetic loop that is coupled to the search coil. The sensitivity of such sensors depends on the permeability of the core material, the area of the coil, the number of turns, and the rate of change of the magnetic flux through the coil. The core material's desired properties include a high permeability for optimal sensitivity, but also a large saturation field, so that large fields can be used in the gap. A small coercivity, low hysteresis losses, and a low magnetostriction are desired to minimize noise. The core material should also have a small but nonzero uniaxial anisotropy so that the domain structure is controllable, and a high electrical resistivity to minimize eddy-current losses. The most used materials are Permalloys (Ni-rich NiFe alloys), and newer materials are being developed such as FeAlSi, Fe/C or Ni/Fe superlattices, and other materials described by Jagielinski (1990). The frequency response of search-coil sensors is limited by the ratio of the coil's inductance to its resistance, though the capacitance between the windings also can be a limiting factor. Their useful range is typically from 1 Hz to 1 MHz. The low-frequency limit implies that variable-reluctance sensors cannot be used as steady-

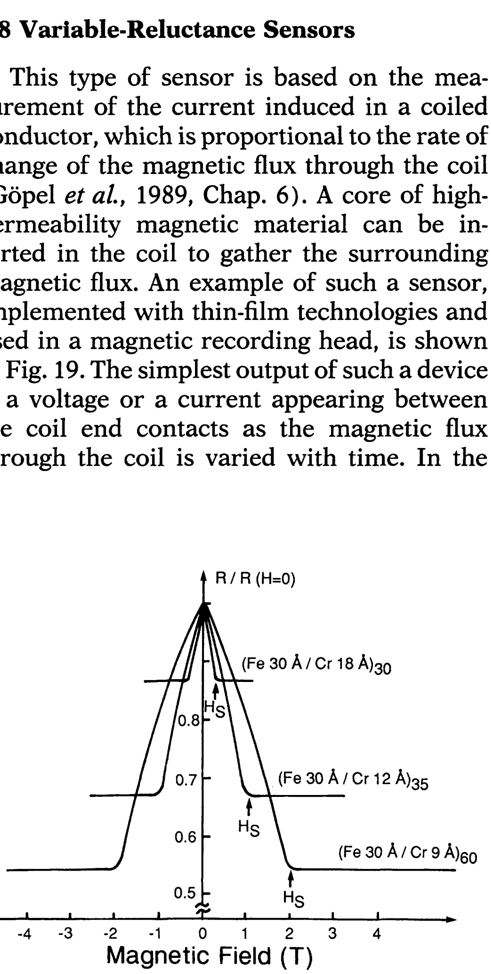

FIG. 18. Relative change in resistance of Fe/Cr multilayers as a function of magnetic field, at 4.2 K. The thicknesses of the individual layers (in Å) are given in the figure, along with the number of layers. The magnetoresistance saturates at a field H_S, but the effect persists to room temperature (after Baibich *et al.*, 1988).

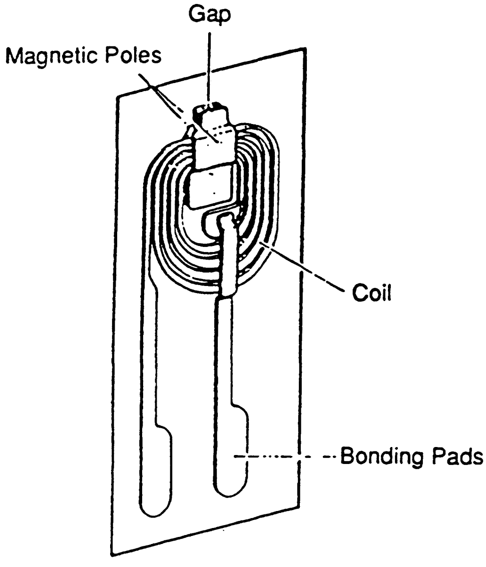

FIG. 19. Inductive head for magnetic recording/readout, made using thin-film technology (after Jagielinski, 1990).

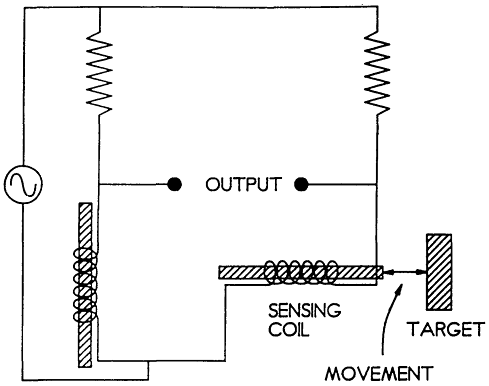

FIG. 20. Search-coil magnetometer in an inductive bridge. Movement of the ferromagnetic target changes the inductance of the coil at right, and unbalances the bridge.

state magnetometers, or in position- sensing applications. They can, however, detect field changes as small as 10^{-10} T.

Another possible way to obtain a readout from a search-coil magnetometer is to measure the coil's inductance, which will change as the magnetic circuit around the coil is changed, for instance by the proximity of a target made from ferromagnetic metal (see Fig. 20). There are two typical readout circuits used to do that. In the circuit shown in Fig. 20, the coil is one leg of a balanced inductive bridge, and the output voltage of the bridge measures the change of inductance of the coil. The coil also can be incorporated in an oscillating circuit, in such a way that a change in inductance results in a change in resonant frequency of the oscillator. These "active" search coils can be used in true position sensors.

1.9 Fluxgates

Fluxgates are often used in compasses. They exploit the magnetic hysteresis exhibited by ferromagnetic materials. They typically consist of cores of a high-permeability ferromagnet, such as Permalloy, each equipped with two coils (Geyger, 1954). There are many types of possible configurations, but only one will be described here (see Fig. 21), built around two rods of a ferromagnetic material. The long thin geometry of the rods is designed to concentrate the flux Φ from the external magnetic field **H**, which is the quantity to be measured. A sinusoidal ac current is applied to the primary windings W'_p and W''_p, which are connected in series but generate magnetic fluxes of opposing polarity in the rods. There are two secondary windings W'_s and W''_s, connected in series as shown. A third coil can eventually be added to supply a dc current I_{dc}. The primary coils are driven so that the core is saturated at each half-cycle of the current. Figure 22 shows very schematically how the voltages induced in W'_s and W''_s are not sinusoidal at frequency f, but clipped. With no external flux Φ or dc current I_{dc}, the voltages V'_s and V''_s induced in W'_s and W''_s cancel each other identically. When an external dc flux Φ or a dc current I_{dc} is applied, the voltages induced in the secondary coils become asymmetrical. A voltage consisting largely of even harmonics of f appears at the output of the two secondary coils. The voltage $V'_s + V''_s$ can be detected by an amplifier locked in at a frequency $2f$. The output can be used as a measurement of Φ, and is sensitive to the polarity of Φ. The dc coil can eventually be used to balance the external field, and the fluxgate is then used as a null detector. The

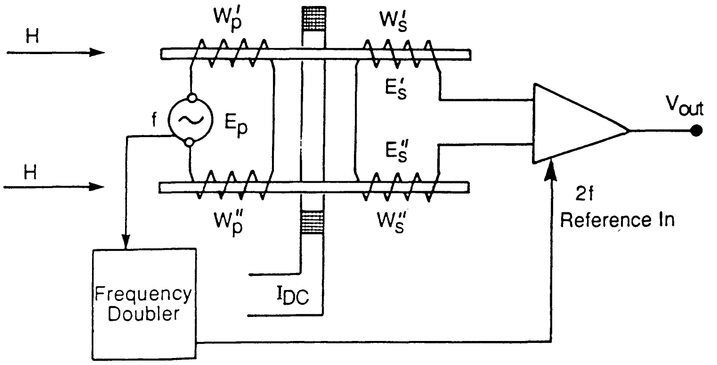

FIG. 21. A fluxgate magnetometer (after Geyger, 1954).

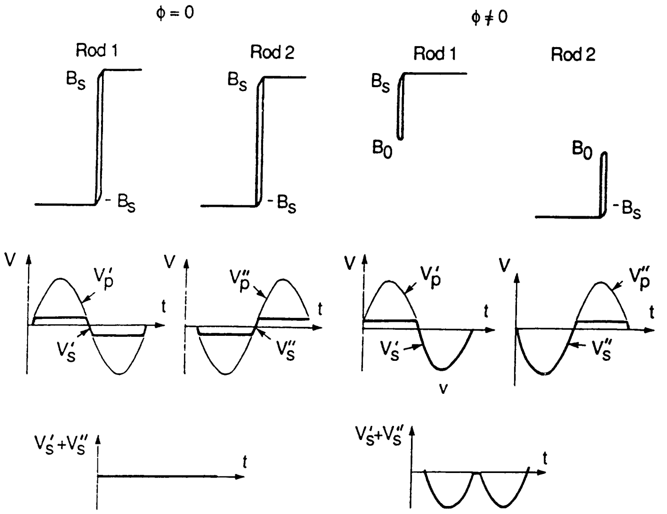

FIG. 22. Principle of operation of the fluxgate. The top figures show which part of the hysteresis curve of the core material is swept in each of the two rods at zero external flux Φ (left), and when flux Φ is present (right). The bottom shows schematically the time dependence of the voltages across the primary (V'_p and V''_p) and secondary (V'_s and V''_s) coils, again without (left) and with (right) flux Φ. The voltage at the input of the amplifier, $V'_s + V''_s$, is shown on the last line. In presence of Φ, $V'_s + V''_s$ has a large component at double the excitation frequency.

sensitivity of fluxgate magnetometers depends on the shape of the hysteresis curves, which should ideally be square. The use of a core material with low coercivity and saturation field values minimizes the power consumption of the device. Sensitivities range from 10^{-10} to 10^{-2} T, but the operation frequency is limited to about 10 kHz, because of the need to have an excitation source at a higher frequency. The power consumption of fluxgate magnetometers can be decreased by operating them on a minor hysteresis loop, because then the core does not have to be driven from saturation to saturation. Flux gates based on thin-film technology have been described by Seitz (1990).

Magnetic amplifiers are devices similar to fluxgates, in which a magnetically controlled saturation of the core of a winding is used to modulate the amount of power delivered from a source to a load. The reader is referred to descriptions by Geyger (1954) and Shea (1968).

1.10 Other Magnetometers

The devices described above are mostly small solid-state type devices. Many complex magnetometer systems exist. Scientific magnetometer systems used in the measurement of magnetic properties of materials have been reviewed by Foner (1981).

Superconductive quantum interference devices (SQUIDs) are the most sensitive instruments for measuring magnetic fields. SQUIDs are described in a separate article in this encyclopedia, and the reader is referred thence.

Optical magnetometers are based on the Zeeman effect in gases such as helium or cesium vapor; they are typically large devices with a high power consumption (Lenz, 1990). There are nuclear precession magnetometers, based on the precession of proton magnetic moments around an applied magnetic field. This type of sensor makes use of hydrocarbon fluids or ^{3}He, and the precession is sensed by a pickup coil. These magnetometers are potentially very sensitive (Lenz, 1990). Magnetooptical sensors are based on Faraday rotation of polarized light traveling through a magnetic material, for instance a terbium gallium garnet (Lenz, 1990). Unfortunately, the Faraday effect is rather small in moderate fields, except in ferromagnets. Dandridge *et al.* (1980) describe a magnetic field sensor based on a conventional single-mode optical fiber coated with a magnetostrictive layer. Under an applied field, the magnetically induced strain changes the optical path length, which can be monitored by interferometry.

1.11 Comparison of Magnetic Field Sensors

The magnetic field range to be detected determines the choice of sensor. The relative sensitivity of the sensor in its field range is the primary factor that must be considered, but other factors are the operating temperature range, the power consumption, the size, the linearity of the sensor, and whether or not sensitivity to the polarity of the field is needed.

Solid-state sensors, such as magnetoresistors, Hall generators, magnetodiodes, and magnetoresistors, have the great advantage of being integratable with amplification and signal-processing circuits, resulting in very cost-competitive systems. In Fig. 23, the sensitivity of a number of solid-state sensors is shown at 20 °C as a function of applied magnetic field. These sensors are a thin-film InSb-based magnetoresistor [n-doped at 6×10^{16} cm^{-3} (Partin *et al.*, 1992)], a thin-film Permalloy-based magnetoresistor (Nippondenso 076100-0202), an ion-implanted GaAs Hall generator (Siemens KSY-10), and a Si-based Hall generator with integrated on-chip amplification (Allegro 3501). In order to make a meaningful comparison, the "sensitivity" of each sensor is defined as the difference between the output voltage of the sensor at the field given in the abscissa and at zero field divided by the supply current (Heremans *et al.*, 1992). At the lowest fields, magnetoresistors based on Permalloy are most sensitive, but their sensitivity saturates at fields around 5×10^{-3} T. Hall generators give an almost linear response to field, while InSb-based magnetoresistors give a quadratic one. As a result, Hall generators are always more sensitive than magnetoresistors at low fields, and the opposite is true at high fields. The crossover field depends on the technology used: unamplified InSb-based magnetoresistors are more sensitive than the Si-based Hall generator shown with integrated amplifier above 0.03 T. The latter sensor has a cutoff frequency (3 dB) of 25 kHz, because of the large-gain amplifier used on the chip. Many other ways of conditioning the output signal can be integrated with the Si Hall generator, such as trigger circuits, latching circuits, or circuits that linearize the output and compensate for temperature over a limited range. Similar functions could be integrated on the GaAs substrates used in the InSb- or GaAs-based sensors. To summarize, Permalloy-based magnetoresistors are preferred for sensing magnetic fields below about 0.005 T, Si Hall sensors with integrated amplification for fields between 0.005 and 0.03 T, if the frequency cutoff is not a problem, and InSb-based magnetoresistors for fields above 0.03 T. These are only crude guidelines. When high-temperature (above 170 to 200 °C) operation is required, such as in automotive applications, GaAs Hall sensors can be considered along with heavily doped InSb or InAs magnetoresistors, but magnetoresistors are more sensitive.

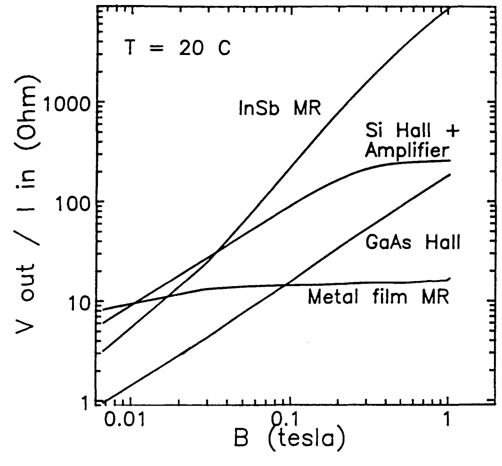

FIG. 23. Comparison of the sensitivities of different solid-state magnetometers at room temperature. The sensitivity is defined as the change in output voltage, divided by the input current, when the applied magnetic field is changed from zero to the value shown in abscissa. At very low fields, the magnetic metal film magnetoresistance (MR) is the most sensitive. It is instructive to note that when field changes above 0.03 T have to be detected, unamplified InSb-based magnetoresistors are more sensitive than Si Hall sensors with integrated amplifiers. The gain of such amplifiers is limited by the signal-to-noise ratio and the bandwidth (Heremans *et al.*, 1993).

2. APPLICATIONS OF MAGNETIC FIELD SENSORS

2.1 Magnetic Recording Heads

Magnetic recording is described elsewhere (see MAGNETIC STORAGE OF INFORMATION). In short, thin layers of magnetic recording media can be magnetized in the direction perpendicular to the film, or in the plane of the film ("longitudinal recording"). Recording heads can use search coils, magnetoresistors, or Hall generators. Figure 24 shows an example of a magnetic circuit that uses a Hall sensor to read codes magnetized into a magnetic data carrier (Siemens, 1989). Magnetoresistive heads operate on much the same principle, but use the more sensitive metal magnetoresistors (Sec. 1.7). A review of biasing methods is given by Markham and Jeffers

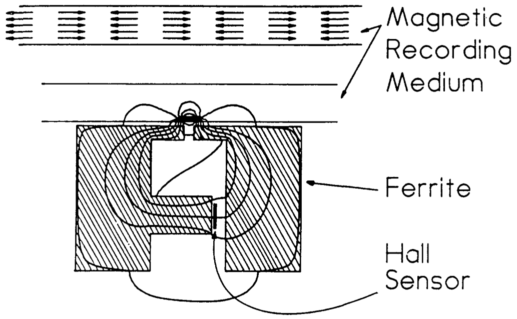

FIG. 24. Data can be encoded in a medium, the top foil, by magnetizing regions of the foil in the longitudinal direction (top). The reading head (bottom) consists of a closed ferrite flux path with two gaps: one for the magnetometer (a Hall sensor) and one close to the medium. The magnetized area (dark rectangle) in the medium generates the flux lines shown.

(1990). Inductive heads are also used, and Jagielinski (1990) shows an example of the parallel readout of ten recording tracks, using thin-film pickup devices such as shown in Fig. 19. Current technology trends are toward multitrack devices, response frequencies up to 100 MHz, and head-media spacings less than 50 nm. In longitudinal recording an increase in linear information density requires a corresponding decrease in gap width, and a higher magnetic field in the gap.

2.2 Position and Speed Sensors

With the widespread use of microprocessors in cars and airplanes, sensors that translate mechanical movement into an electrical signal are becoming very common. The simplest device that does this is the potentiometer; but the friction involved in conventional potentiometers makes them less reliable than frictionless methods. Optical techniques (a light-emitting diode and a photodetector, with or without fiber optics, for instance) are sensitive to contamination, and their applications are limited to protected environments. Magnetic position sensors are the technology of choice when reliability is a critical issue. They are based on the use of a permanent magnet whose position is changed with respect of that of a magnetic field sensor. Alternatively, the magnetic field sensor can be mounted on the magnet, and a ferromagnetic metal part moved with respect to that assembly. In either design, the mechanical movement is translated into a change in the flux through the magnetic field sensor. An example of the design of a rotary contactless potentiometer is shown in Fig. 25. Linear position sensors using a sliding magnet moving with respect to a magnetoresistor or a Hall sensor can be designed. Proximity sensors based on similar designs are also common, for instance to verify door and hatch closures (Lenz, 1990).

2.3 Rotary Shaft Encoders

Magnetoresistors or Hall sensors are widely used in rotary shaft encoders when the speed and the position of the shaft have to be detected. Pickup coils can be used when the information is not needed at zero rotation speed. A typical magnetic circuit used for sensing the position of gear teeth machined into an iron or steel wheel is shown in Fig. 26.

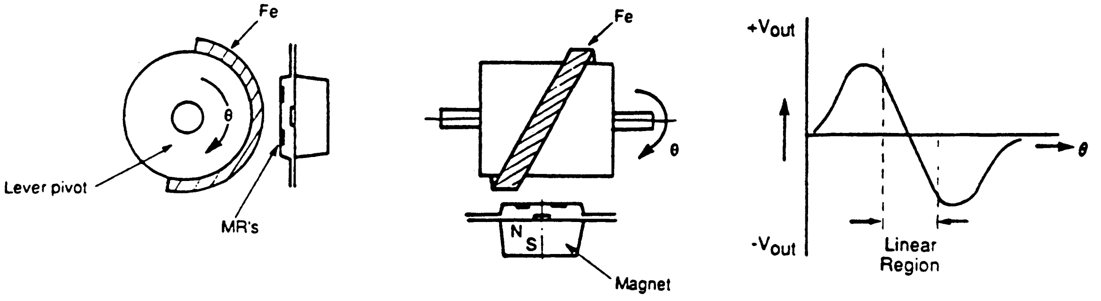

FIG. 25. A contactless potentiometer using an iron film on a mandrel, and two magnetoresistors (MRs) mounted on a biasing magnet. The angular position θ of the mandrel determines the amount of flux through each magnetoresistor. When the two magnetoresistors are connected in series, the voltage V_{out} of the midpoint varies with θ as shown. In the region in which V_{out} varies linearly with θ, the device can be used as a potentiometer (adapted from Siemens, 1989).

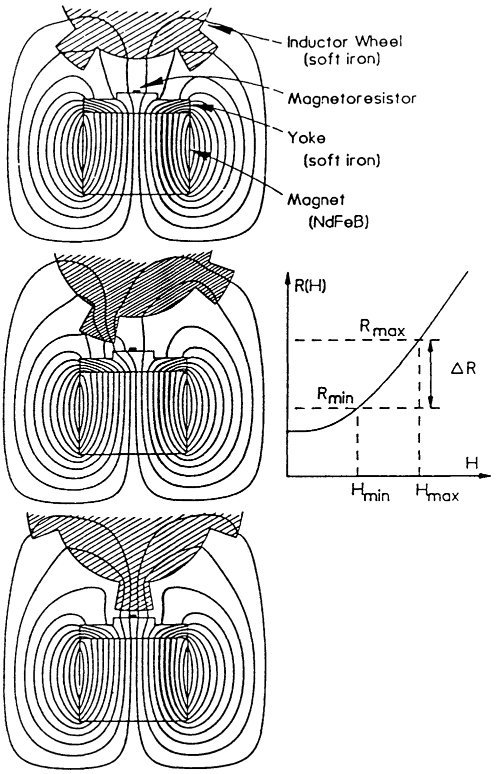

FIG. 26. Rotary shaft position sensor using a magnetoresistor mounted on an iron yoke on a bias magnet, and a cogged inductor wheel. The flux lines are shown at left, for different positions of the inductor wheel. The magnetic induction *H* at the sensor location changes with the inductor wheel angle, and so does the electrical resistance *R*(*H*) of the magnetoresistor.

The size and distance of the teeth can be used to encode absolute position sensing. This type of sensor is widely used to measure the position of the crankshaft and camshaft in a gasoline engine (see Fig. 27). Its information is fed into the engine control computer, which then determines which cylinder fires and when it does so. The ignition system shown in Fig. 27 now replaces the mechanical switches and "points" used in older car engines.

2.4 Brushless Motors

Figure 28 shows a schematic diagram of a brushless dc electric motor. The rotor is a permanent magnet (labels indicate the north and south poles). The position of the rotor is monitored by Hall sensors H_1 and H_2. These control the base currents into the bipolar transistors that control the current flow through the stator windings. Such motors have easy starting characteristics. The absence of brushes results in better reliability and the elimination of the electrical noise due to sparking. They find applications in video recorders, disk drives, turbopumps, and automotive motors. The control of the absolute rotor position can be done by a microprocessor chip, and complicated rotation patterns can be programmed in a flexible manner.

2.5 Shuntless dc Current Measurements

Shuntless dc ammeters can be made as follows: A primary coil is wound on a soft iron core that contains an air gap. The current to be measured, I_{in}, passes through this winding. The resulting flux generates a field that can be detected by the Hall sensor located in the air gap. The output of the Hall sensor is a measurement of the current I_{in}. For more precision and wider dynamic range, a secondary coil is wound around the core. A current can be made to flow through this secondary, so as to cancel the field in the gap. The Hall sensor is then used as a null detector, and the current through the secondary coil is a measure of the current through the primary.

3. HIGH-FREQUENCY DEVICES

3.1 General Aspects

The magnetic component of an electromagnetic wave couples readily to magnetic materials. Ferrites and powdered ferromagnetic materials are used as magnetic field concentrators in antenna cores, inductors and transformers. Devices operating in the microwave-frequency range also use the ferromagnetic (or ferrimagnetic) resonance. In ferromagnetic solids, the adjacent magnetic moments of the atoms interact very strongly and align parallel to each other, as long as the solid is below a critical temperature, the Curie temperature T_C. In ferrimagnetic materials, the spins are aligned antiparallel, as in antiferromagnets, but there are more spins pointing in one direction than in the other. The material thus has a net magnetic moment, but smaller than it would have if all spins were parallel. Their critical ordering temperature is called the Néel temperature, T_N. Unfortunately for device applications, solids that are strictly

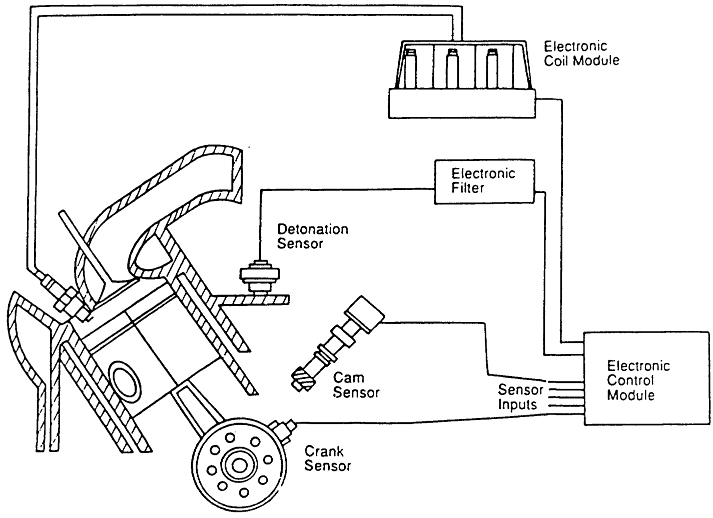

FIG. 27. Ignition system of a Buick engine. Magnetic position systems such as shown in Fig. 26 measure the position of crankshaft and camshaft, and transmit the information to a microprocessor (the electronic control module). The microprocessor, which also receives data on the air flow into the engine, the coolant temperature, the knock sensor, and the transmission, determines the spark timing and sequence. At the proper time, it switches low-voltage current to one of three coils. The secondary windings of each of these coils are connected to two spark plugs. When the current to a given coil is interrupted, both sparks fire, one in an empty cylinder, and one in a fully fueled cylinder. Such ignition systems avoid the mechanical commutation of high voltage.

ferromagnetic and with Curie temperatures above room temperature are metals. Eddy currents are generated in these electrically conducting materials at high frequencies. This leads to unacceptable energy dissipation. Magnetic materials with a high electrical resistivity, such as powdered metals, ferrites, and garnets, are preferred for high-frequency devices.

Powdered ferromagnetic materials, compacted with an insulating binder, can have reasonably low eddy-current losses. The reduction in losses in these materials usually comes at the expense of the permeability of the compact. Powdered cores are made using Permalloy, pure iron, or Sendust powder. Some properties of these powders are summarized in Table 1.

Ferrites (Smit and Wijn, 1959) are ferrimagnetic. The general chemical composition of soft ferrites is $(MO) \cdot (Fe_2O_3)$, where M is a divalent metal such as Fe, Ni, Co, Mn, Mg, Zn, Cd, or Cu. The simplest member of this class is thus iron ferrite, Fe_3O_4, or magnetite, mentioned in Fig. 1. Ferrites have high resistivities because the constituent atoms have no free 3s or 4s electrons available for conduction. Ferrites are selected on the basis of a number of properties, such as a high permeability at low field, or low losses at high frequency, or a square hysteresis loop. Impurities tend to decrease the permeability. The properties of some ferrites are given in Table 1.

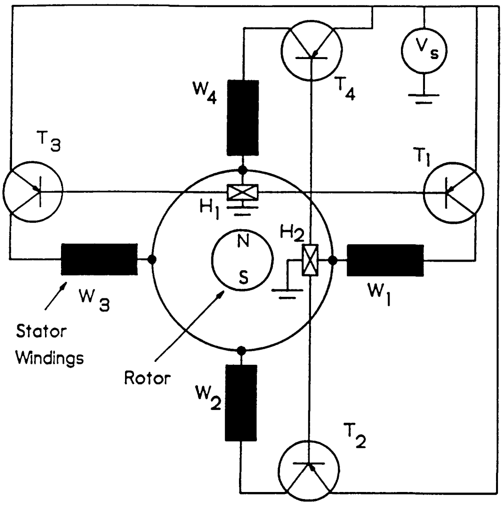

FIG. 28. Brushless electrical motor, with a permanent-magnet rotor and Hall sensors. The control currents are provided by the Hall sensors H_1 and H_2 to the transistors T which supply the windings W (adapted from Siemens, 1989).

Table 1. Magnetic properties of ferrites and pressed powders used in magnetic cores (Chikazumi, 1964). The balance for composition in the powders is Fe.

Material	Composition	Permeability, μ	Saturation magnetization (G)	Curie or Néel temp. (°C)
Pressed powders				T_C
Iron	Carbonyl Fe	20	1240	770
Mo-Permalloy	2 Mo, 81 Ni	125	560	480
Sendust	5 Al, 10 Si	80	360	500
Ferrites				T_N
Mn-Zn	50 Mn, 50 Zn	2000	200	110
Cu-Zn	40 Cu, 60 Zn	1100	—	90
Cu-Mn	40 Cu, 60 Mn	—	230	—
Ni-Zn	30 Ni, 70 Zn	80	320	130
Mg-Zn	50 Mg, 50 Zn	500	207	120
Mg-Mn	50 Mg, 50 Mn	1400	215	130

Other materials commonly used are rare-earth garnets, mostly yttrium iron garnet (YIG) or gallium-substituted YIG (GaYIG). Their general chemical formula is $(5Fe_2O_3) \cdot (3M_2O_3)$, where M represents Y or a rare-earth ion (Sm to Lu). Materials used in practice can deviate from these simple formulas. Both polycrystalline and single-crystal materials are used.

Microwave applications use the ferromagnetic resonance, whereby the spin system absorbs energy from the electromagnetic wave at a specific, resonant energy. Figure 29 shows schematically a sphere of YIG in a vertical dc field H_0. Ferromagnetic resonance can be viewed as the tendency for the spins in the garnet to precess about the vertical field H_0. For a YIG sphere, the resonance frequency of the precession is empirically given by $f_R = (28\ \mathrm{GHz/T})H_0$. The spin precession results in a rotating magnetic dipole **m** shown in Fig. 29(a). A circularly polarized microwave (rf) radiation with a magnetic vector $\mathbf{h}^+$ shown in Fig. 29(b) can excite that precession, if the radiation is at or near the frequency f_R. rf magnetic fields will build up about the sphere in the same way as the magnetic field builds up in the inductance of an *L-C* tuned circuit at resonance. Circularly polarized rf radiation with a magnetic component in the direction $\mathbf{h}^-$ shown in Fig. 29(c) does not couple to the sphere, even if the radiation is at the resonant frequency. The device is thus not "reciprocal," and will be selective with regard to the direction of propagation of a traveling rf wave. Linearly polarized fields may be constructed as the sum of two circularly polarized fields of equal strengths rotating in opposite directions. When the sphere is subjected to a linearly polarized field, it will respond to only one of the two circularly polarized components, and therefore emit only that circularly polarized field at resonance.

Some relevant properties of a few single-crystal materials used in practice are given in Table 2. The saturation magnetization limits the range of dc magnetic field available for tuning the resonance frequency. The linewidth of the resonance can be expressed as the range of dc magnetic fields ΔH over which the resonator responds (within 3 dB) at frequency f_R. This quantity depends on material parameters, and also on the shape and surface finish of the device. The resonant frequency is a function of the shape of the resonator (via the demagnetization factor; see MAGNETOSTATICS), the crystalline anisotropy (in single crystals), and interactions of the ferrite or garnet with currents in the metallic walls surrounding the device. Details are given by Soohoo (1985) and Matthaei *et al.* (1980).

3.2 Microwave Devices

3.2.1 Filters. Figure 30 shows a resonator filter built around a ferrite sphere. The incoming and outgoing waves are coupled through two loops, slightly elliptical in shape so that they can be nested perpendicularly to each other. The ferrimagnetic resonator is located in the center. The biasing dc magnetic field H_0 is parallel to the plane of both loops.

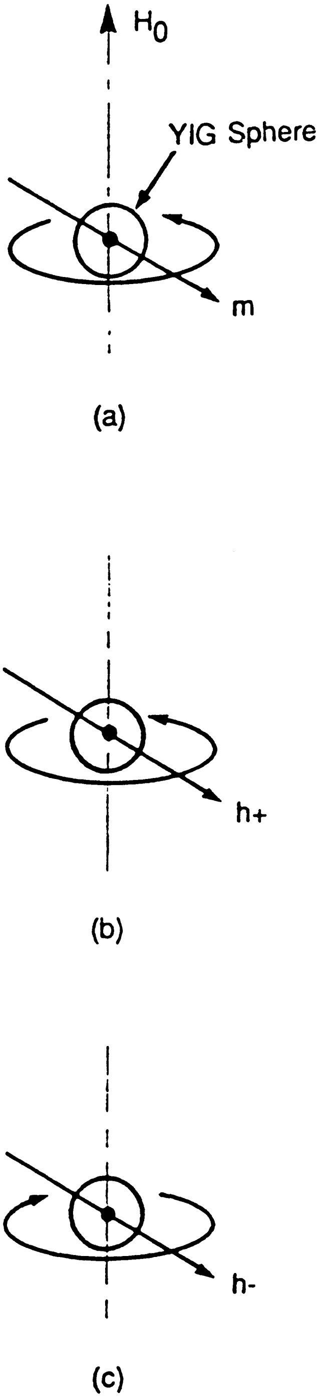

FIG. 29. (a) A YIG sphere with a circularly polarized ac moment **m**. (b) External ac fields at the resonant frequency of the sphere and circularly polarized magnetic field component $\mathbf{h}^+$ couple to the sphere. (c) If the polarization of the external field is in the wrong direction, $\mathbf{h}^-$, the wave does not couple to the YIG sphere. (After Matthaei *et al.*, 1980.)

When an rf magnetic field h_x at the resonant frequency f_R is generated by the x loop, the ferrite emits an rf moment m with components m_x and m_y. The m_y component couples to the y loop, and the wave propagates into the termination R_y. At frequencies other than f_R, no wave propagates. The frequency f_R can be tuned by changing the biasing field H_0. Such loop-coupled filters are easy to understand, but in practice there is considerable stray coupling between the loops, and the rf fields generated are not very uniform, which tends to excite higher-order modes. A much more practical, waveguide-based, device is shown in Fig. 31. Two ferrimagnetic resonating spheres are placed near the walls in the waveguide at locations such that the spheres see a strong rf magnetic field. The spheres are coupled to one another through a slot parallel to the waveguide axis; the geometry of the slot is designed to minimize coupling between the waveguides except when the spheres are excited into resonance. The spheres are placed in reduced-height waveguides to obtain adequate coupling between them and the waveguide. This also reduces the air gap for the dc biasing magnet. The steps in the waveguide height form a Tchebyscheff transformer, designed to provide impedance matching between the normal waveguide and the reduced-height portion (Matthaei *et al.*, 1980). This filter, tunable via variations of H_0, has the same attenuation and phase characteristics for both directions of transmission. Similar structures can be designed in strip-line technology. Filters can also be designed that induce a phase shift in the propagating wave, or that have different transmission coefficients for waves propagating in different directions (Matthaei *et al.*, 1980).

3.2.2 Isolators. Isolators are devices that allow wave propagation in one direction, but not in the opposite one. They are particularly useful to isolate a microwave source from the waves reflected from a load that perturbs the amplitude or frequency stability of the source. A waveguide-based ferrite isolator is shown in Fig. 32. In an empty rectangular waveguide, Fig. 32(a), the components of the rf magnetic field for the TE_{10} mode are h_x and h_z, shown in

Table 2. Properties of single-crystal ferrimagnetic materials used in microwave devices (Matthaei *et al.*, 1980).

Material	Néel temp., T_N (°C)	Saturation magnetization, $4\pi M_s$ (G)	Linewidth, ΔH (Oe)
Yttrium iron garnet (YIG)	292	1750	0.22 (@ 4 GHz)
Gallium-substituted YIGs	—	50–1750	—
(GaYIGs)	160	600	—
	206	950±50	0.7–2 (@ 4.4 GHz)
Lithium ferrite[a]	—	3550±40	3 (@ 5 GHz)
Planar ferrite, $Ba_{12}Zn_2Fe_{12}O_{22}$	—	2850	16 (8–12 GHz)

[a] Nielsen *et al.* (1962)

Fig. 32(b) as functions of x for $y=0$. Since h_x and h_z are in space and time quadrature, circular polarization is obtained at the locations where $|h_x| = |h_z|$, i.e., the points x_1 and $a-x_1$ in Fig. 32. The value of x_1 is calculated by Soohoo (1985). An isolator can be made by applying a ferrimagnetic layer as shown in Fig. 38(c) at those points, and biasing them with opposing dc magnetic fields H_0 as shown. The nonreciprocity of this structure arises from the fact that each ferrite layer couples to only one direction of polarization of the rf magnetic field, as shown in Fig. 29. Not all four strips of ferrite need to be present to make an isolator. As described, the device in Fig. 32 is an efficient isolator, giving a large attenuation for waves traveling backward (backward attenuation), and a low attenuation for forward propagating waves (low forward attenuation, or insertion loss) only at the resonant frequency. If the bias magnetic field H_0 is far away from resonance, one can obtain a phase shift at low attenuation. Such devices are "phase shifters." A phase shifter with reciprocal action can be made if the H_0 values applied to the strips at x_1 and $a-x_1$ are of the same polarity. If the bias magnets produce H_0 values of opposite polarity, as shown in Fig. 32, the phase shift is nonreciprocal.

Both the bandwidth and the reverse-to-forward absorption ratio of the isolator can be increased by adding a strip of dielectric in the waveguide, next to the ferrite strip and on the inside of it. This technique makes the fields inside the waveguide more circularly polarized throughout the cross section and over a broader wavelength range. It is described by Soohoo (1985).

3.2.3 Circulators. A circulator is a microwave-circuit component with three or more ports. In a circulator, microwave energy can propagate from one port to another only in

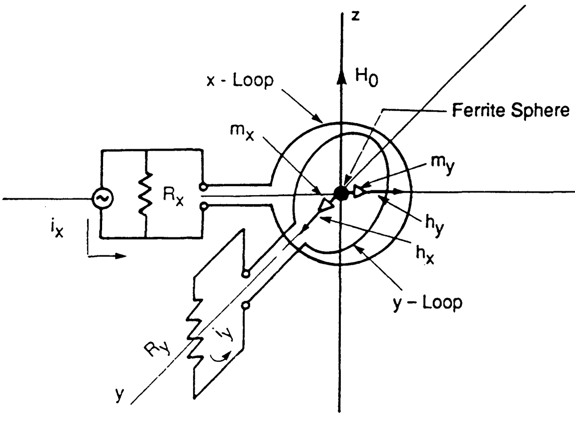

FIG. 30. A magnetically tunable filter using a single ferrite resonator in a dc biasing field H_0, and loops x and y for coupling the waves (after Matthaei *et al.*, 1980).

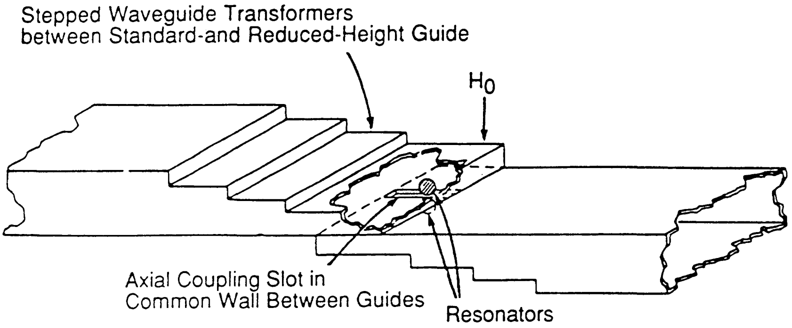
Stepped Waveguide Transformers
between Standard-and Reduced-Height Guide
H_0
Axial Coupling Slot in
Common Wall Between Guides
Resonators

FIG. 31. A magnetically tunable waveguide filter with two resonators and an elongated axial coupling slot (after Matthaei *et al.*,et al. 1980).

one direction. In the three-port circulator shown in Fig. 33, with the dc field H_0 applied in one direction, energy can go from port 1 to port 2, from 2 to 3, and from 3 to 1, but not from 1 to 3, 3 to 2, or 2 to 1. A pedagogical example of the use of these devices is an elementary radar: attach a source to port 1, an antenna to port 2, and a detector to port 3. The wave generated at 1 propagates out through port 2 and out of the antenna. The wave reflected back from the target is received by the antenna and propagates from port 2 to port 3, where it can be detected. A Doppler radar can be made by mixing a small amount of source radiation with the wave reflected from the target. The carrier-wave frequency is filtered out at port 3 and the detector is made sensitive only to the difference frequency. This frequency is then a measure of the relative velocity of the target with respect to the radar. The circulator itself can be built as shown in Fig. 33 (top). The ferrite is attached to the top or bottom plate of the cavity at the location where the wave is circularly polarized, as in the design of the isolator, Fig. 32. A dc biasing magnetic field H_0 is applied normal to the plane of Fig. 33. Depending on the polarity of H_0, circulation is made possible either in the clockwise direction, $1 \rightarrow 2 \rightarrow 3$, or in the counterclockwise direction. This circulator is strictly a resonant device again. Dielectrics can be used, as in the design of isolators, to broaden the device's bandwidth (Fig. 33, bottom). There are many other circulator designs possible, and the reader is referred to Matthaei *et al.* (1980), Nielsen *et al.* (1962), and the article on MICROWAVE CIRCUITS.

3.2.4 Strip-Line, MMIC, and Thin-Film Technologies. The microwave-circuit elements described above used waveguide technology, for pedagogical reasons. The component elements described above are copied in strip-line technology, using ferrite or YIG thin films, and even thin-film magnets (Stancil, 1989). Ferromagnetic metals can be considered in thin-film devices, because the eddy currents are much reduced when the rf magnetic field is applied in the plane. This allows for the design of novel devices, such as the strip-line notch filter shown in Fig. 34. This filter is based on the ferromagnetic resonance in iron. The iron film shown has a resonance at $f_R \simeq 10$ GHz without external biasing. In Fig. 34, the radiation is polarized with the ac electric and magnetic fields **e** and **h** as shown. It propagates from left to right. In the iron film, which has a magnetization vector **M** as shown, radiation at the resonant frequency f_R is absorbed by the iron. Radiation at other frequencies is transmitted through the strip with much less attenuation, so that the strip serves as a notch filter.

Modern thin-film techniques can combine device-grade semiconductors with ferromagnetic films. This opens fascinating new device opportunities in monolithic microwave integrated circuits (MMICs) (Prinz, 1990).

3.3 Other Magnetic Devices

3.3.1 Antenna, Inductor, and Transformer Cores. A winding distributed about the length of a rod of high-permeability material forms an antenna in broadcast radio receivers. The core concentrates the ac magnetic flux into the antenna. Similarly, inductors and

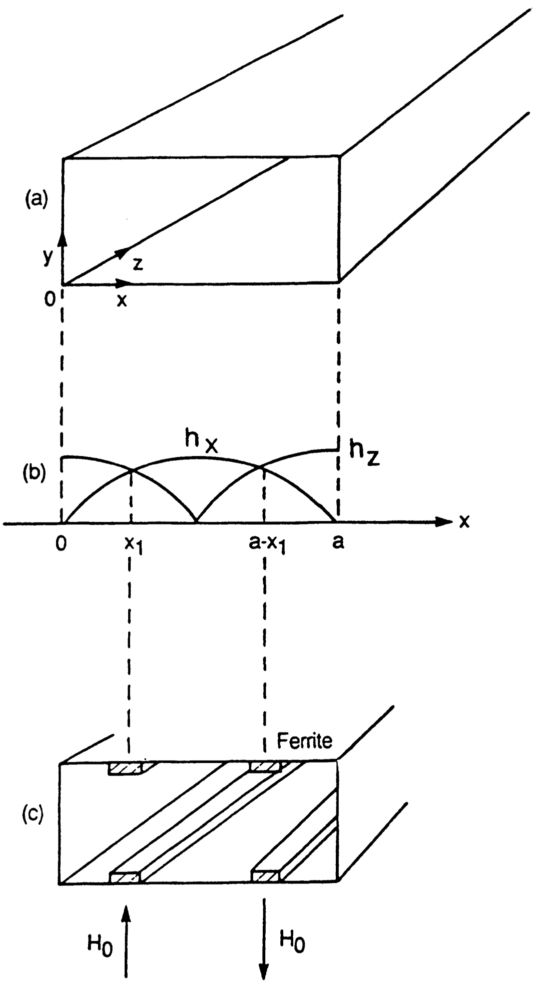

FIG. 32. A waveguide-based ferrite isolator. (a), (b) The magnetic field distribution of the TE$_{10}$ mode in a rectangular waveguide. (c) The ferrite strips are inserted at the location where the wave is circularly polarized, i.e., $|h_z| = h_x$. The dc biasing field is H_0. (After Soohoo, 1985.)

transformers are wound around magnetic material cores. Adjustable inductors are used in many circuits. The most common form of adjustment involves changing the position of the high-permeability core material with respect to the windings. Conventional variable capacitors are difficult to seal with respect to ambient humidity; this is much less a problem in tuned inductors. The windings can easily be sealed, and moisture has little effect on the behavior of the core. Designs of antennas, inductors, and transformers use a compromise between high permeability and low losses at the intended frequency of operation. Powdered magnets and ferrites are used for cores at the higher end of the frequency spectrum. Ferromagnetic metals, such as the Ni-Fe alloys, Mumetal, Permalloy, Perminvar, and thin metglasses can be used at lower frequencies. Tape-wound cores from these alloys are available. Fe-Co, Fe-Si, Fe-Al, and pure Fe are sometimes also used in electromagnet and relay cores, since these are nearly dc applications (Fink and McKenzie, 1975).

3.3.2 Magnetic Memories. Magnetic materials are used in memories, and this type of devices is the subject of a separate article (MEMORIES). The synergistic effects of new film-growth techniques of both semiconductors and magnetic materials, already mentioned in the section on microwave thin-film techniques, also influences this field. For in-

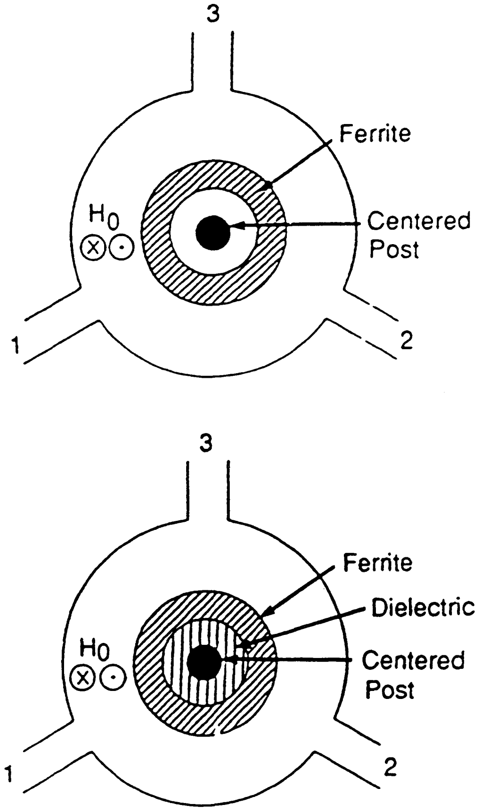

FIG. 33. (top) A waveguide circulator with three ports, 1, 2, and 3. Microwave energy can go from 1 to 2, 2 to 3, and 3 to 1, but not vice versa. The circulator is made by inserting a ferrite at the location where the wave is circularly polarized. (bottom) Dielectric loading of the circulator increases the bandwidth. (After Soohoo, 1985.)

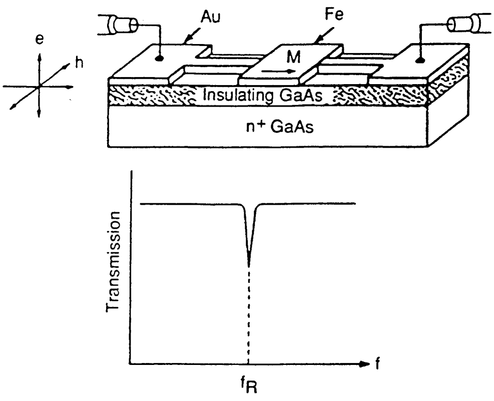

FIG. 34. Strip-line microwave notch filter, using an epitaxial iron film as resonator (after Prinz, 1990).

stance, an integrated magnetic memory element was proposed by Prinz (1990), and is shown in Fig. 35. It is the thin-film analog of the ferrite core. The element can be made from epitaxial magnetic metal deposited on a semiconductor substrate, such as Fe on Si or GaAs. In the example chosen, the iron film is patterned so that the magnetic moments align with the magnetically easy axes of Fe, here the ⟨100⟩ axes. The film can be magnetized with the moments aligned clockwise or counterclockwise, representing the 0 or 1 states of the bit. The readout of the magnetic state of the memory element can be integrated in the semiconductor, for instance as a Hall element. The magnetic loop is closed, and very little leakage flux and cross-talk are expected. The down-scaling of this memory element is only limited by semiconductor technology, because the magnetic properties of Fe are independent of size, down to maybe 200 Å.

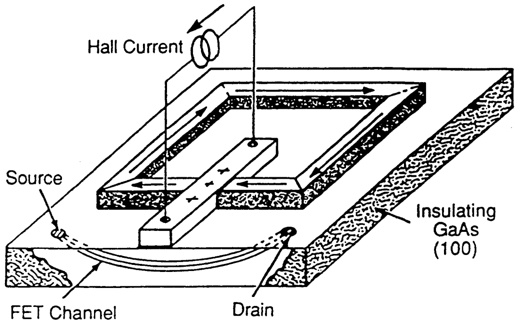

FIG. 35. Picture-frame memory element made from a thin epitaxial iron film on GaAs with a readout Hall sensor integrated in the substrate (after Prinz, 1990).

Another intergrated magnetic storage technology that uses magnetoresistors has been described by Daughton (1992), and a bipolar spin switch is described by Johnson (1993).

4. CONCLUSION

Magnetic effects and materials are widely used in electrical machines, but also in the conversion of mechanical movement into an electrical signal (position sensors) and in electronics. Magnetic device technology parallels semiconductor technology, and the synergism between the two will continue to lead to yet new devices. The ultimate synergism would be in the field of magnetic semiconductors, which were not discussed here. Only "dilute" magnetic semiconductors have been studied recently (see SEMICONDUCTORS, DILUTE MAGNETIC), and their applications are limited, because their magnetic properties are not very pronounced, especially at room temperature. Optical isolators, which perform in the visible light range the same function as their microwave equivalents described in Sec. 3.2.2, and are needed in high-data-rate optical communication systems, could be made from such materials (Furdyna *et al.*, 1991).

Acknowledgments

I want to acknowledge the help of Dr. J. Herbst and Dr. S. Foner in reading the manuscript, Dr. D. L. Partin for frequent discussions, and Mr. R. C. Dinsmore, from Dinsmore Instruments Company, for reviewing the section on fluxgates.

GLOSSARY

Circulator: Microwave device having three or more ports, labeled 1, 2, 3, etc. Microwave radiation can only propagate from port 1 to 2, 2 to 3, etc., in that order, and cannot propagate in the reverse order.

Corbino Disk: Circularly shaped slab of semiconductor, with two electrodes, one in the middle of the circle, the other on the periphery. The current lines in this element flow radially through the semiconductor. This geometry gives the largest geometrical magnetoresistance. (Can also be made from semimetals and metals.)

Fluxgate: Sensitive magnetometer using a core of magnetic material driven close to saturation, and often used in compasses.

Garnet: Ferrimagnetic mixed oxide of iron and rare-earth elements or yttrium, electrically insulating. Used as magnetic material for high-frequency applications, in which eddy-current losses have to be minimal.

Geometrical Magnetoresistance: Magnetoresistance induced in a semiconductor resistive element by choosing the geometry of the element in such a way that an applied magnetic field lengthens the current paths of the charge carriers.

Hall Generator: Device exploiting the Hall effect to generate a voltage that is used to measure the magnetic field.

Isolator (microwave or optical): Device that lets an electromagnetic wave propagate only in one direction. It is commonly placed between a source (microwave generator or laser) and its load to prevent energy reflected from the load from perturbing the source.

Magnetic Amplifier: Transformerlike device that modulates the amount of power that can be delivered from a source to a load by saturating the core of the transformer using a separate dc control current.

Magnetodiode: *p-i-n* diode in which an external applied magnetic field controls the amount of recombination in the intrinsic region.

Magnetoresistor: Resistor having a resistance that is a strong function of magnetic field, used in magnetometers. It exploits either the geometrical magnetoresistance in semiconductors, or the magnetoresistance anisotropy in magnetic metals.

Phase Shifter: Microwave device that shifts the phase of a wave propagating through it. Nonreciprocal phase shifters give different amounts of shift for waves propagating in opposing directions.

List of Works Cited

Bai Shouyi (1982), *An Outline History of China*, Beijing: Foreign Languages Press.

Baibich, M. N., Broto, J. M., Fert, A., Nguyen Van Dau, F., Petroff, F., Etienne, P., Creuzet, G., Friederich, A., Chazelas, A. (1988), *Phys. Rev. Lett.* **61**, 2472–2475.

Baltes, H. P., Popovic, R. S. (1986), *Proc. IEEE* **74**, 1107–1132.

Barthelemy, A., Fert, A. (1991), *Phys. Rev. B* **43**, 13124–13129.

Campbell, I. A., Fert, A., Jaoul, O. (1970), *J. Phys. C* **3** (Suppl. 1), S95–S101.

Chikazumi, S. (1964), *Physics of Magnetism*, New York: Wiley.

Cristoloveanu, S., Chovet, A., Malyutenko, V. K. (1983), *Sens. Actuators* **4**, 165–171.

Dandridge, A., Treten, A. B., Sipel, G. H., Jr., West, E. J., Giallorenzi, T. G. (1980), *Electron. Lett.* **16**, 408–409.

Daughton, J. M. (1992), *Thin Solid Films* **216**, 162–168.

Dibbern, U. (1986), *Sens. Actuators* **10**, 127–140.

Dieny, B., Speriosu, V. S., Parkin, S. S. P., Gurney, B. A., Wilhoit, D. R., Mauri, D. (1991), *Phys. Rev. B* **43**, 1297–1300.

Fink, D. G., McKenzie, A. A. (Eds.) (1975), *Electronic Engineers' Handbook*, New York: McGraw-Hill.

Foner, S. (1981), *IEEE Trans. Magn.* **MAG-17**, 3358–3363.

Furdyna, J., Aggarwal, R. L., Awshalom, D. D., Ehrenreich, H. H., Geschwind, S., Gunshor, R. L., Hass, K. C., Heremans, J. P., Peercy, P. S., Prinz, G. A., Ramdas, A. K., von Molnar, S., Wolff, P. A. (1991), "Diluted Magnetic Semiconductors," Panel Report, Washington, DC: National Academy Press.

Geyger, W. A. (1954), *Magnetic-Amplifier Circuits*, New York: McGraw-Hill.

Göpel, W., Hesse, J., Zemel, J. N. (Eds.) (1989), *Sensors*, Vol. 5, Weinheim, Germany: VCH.

Heremans, J. (1993), *J. Phys. D: Appl. Phys.* **26**, 1–20.

Heremans, J., Partin, D. L., Morelli, D. T., Fuller, B. K., Thrush, C. M. (1990), *Appl. Phys. Lett.* **57**, 291–293.

Heremans, J., Partin, D. L., Thrush, C. M., Green, L. (1993), *Semicond. Sci. Technol.* **8**, 424–430.

Itakura, K., Ueda, D., Hagio, M., Kazumura, M. (1989), *Electron. Lett.* **25**, 1493–1495.

Jagielinski, T. (1990), *MRS Bull.* **15** (3), 36–43.

Johnson, M. (1993), *Science* **260**, 320–323.

Lenz, J. E. (1990), *Proc. IEEE* **78**, 973–989.

Lippmann, H. J., Kuhrt, K. (1958), *Z. Naturforsch.* **13A**, 462–474.

Maenaka, K., Okada, H., Nakamura, T. (1990), *Sens. Actuators* **A21–A23**, 807–811.

Markham, D., Jeffers, F. (1990), *Proc. Am. Electrochem. Soc.* **90**, 185–204.

Matthaei, G., Young, L., Jones, E. M. T. (1980), *Microwave Filters, Impedance-Matching Networks, and Coupling Structures*, Dedham, MA: Artech House Books.

McGuire, T. R., Potter, R. I. (1975), *IEEE Trans. Magn.* **MAG-11**, 1018–1038.

Mosca, D. H., Petroff, F., Fert, A., Schroeder, P. A., Pratt, W. P., Jr., Laloee, R. (1991), *J. Magn. Magn. Mater.* **94**, L1–L5.

Nakamura, T., Maenaka, K. (1990), *Sens. Actuators* **A21–A23**, 762–769.

Nathan, A., Maenaka, K., Allegretto, W., Baltes, H. P., Nakamura, T. (1989), *IEEE Trans. Electron Devices* **36**, 108–117.

Nathan, A., Baltes, H. P., Castagnetti, R., Sugiyama, Y., Briglio, D. R. (1990), *Sens. Actuators* **A21–A23**, 776–779.

Nielsen, J. W., Lepore, D. A., Zneimer, J., Townsend, G. B. (1962), *J. Appl. Phys.* **33S**, 1379–1380.

Partin, D. L., Heremans, J. (1992), "Growth of Narow-Gap Semiconductors," in: S. Mahajan (Ed.), *Handbook on Semiconductors*, Vol. 3, 2nd ed., New York: Elsevier,

Partin, D. L., Heremans, J., Thrush, C. M., Green, L. (1992), in: *Technical Digest, IEEE Solid-State Sensor and Actuator Workshop, Hilton Head Island, SC, 22–25 June 1992*, New York: IEEE, pp. 35–38.

Popovic, R. S., Baltes, H. P. (1983), *Sensors and Actuators* **4**, 155–163.

Prinz, G. A. (1990), *Science* **250**, 1092–1097.

Seitz, T. (1990), *Sens. Actuators* **A22**, 799–802.

Shea, R. F. (1968), *Amplifier Handbook*, New York: McGraw-Hill, Chaps. 8–21.

Siemens AG (1989), *Magnetic Sensor Data Book*, München: Siemens AG.

Smit, J., Wijn, H. P. (1959), *Ferrites*, New York: Wiley.

Soohoo, R. F. (1985), *Microwave Magnetics*, New York: Harper & Row.

Stancil, D. D. (1989), *Microwave Opt. Tech. Lett.* **2**, 81–85.

Weiss, H. (1969), *Structure and Applications of Galvanomagnetic Devices*, New York: Pergamon.

Yamada, T. (1968), in: *Proceedings of the IX International Conference on the Physics of Semiconductors*, Moscow, St. Petersburg: Publishing House Nauka, pp. 672–675.

Further Reading

Baltes H. P., Popovic, R. S. (1986), *Proc. IEEE* **74**, 1107–1132.

Fink, D. G., McKenzie, A. A. (Eds.) (1975), *Electronic Engineers' Handbook*, New York: McGraw-Hill.

Furdyna, J., Aggarwal, R. L., Awshalom, D. D., Ehrenreich, H. H., Geschwind, S., Gunshor, R. L., Hass, K. C., Heremans, J., Peercy, P. S., Prinz, G. A., Ramdas, A. K., von Molnar, S., Wolff, P. A. (1991), "Diluted Magnetic Semiconductors," Panel Report, Washington, DC: National Academy Press.

Geyger, W. A. (1954), *Magnetic-Amplifier Circuits*, New York: McGraw-Hill.

Göpel, W., Hesse, J., Zemel, J. N. (Eds.) (1989), *Sensors*, Vol. 5, *Magnetic Sensors*, Weinheim, Germany: VCH.

Heremans, J. (1993), "Solid state magnetic field sensors and applications," *J. Phys. D: Appl. Phys.* **26**, 1–20.

Matthaei, G., Young, L., Jones, E. M. T. (1980), *Microwave Filters, Impedance-Matching Networks, and Coupling Structures*, Dedham, MA: Artech House Books.

Popovic, R. S. (1991), *Hall Effect Devices*, Bristol, U.K.: Hilger.

Prinz, G. A. (1990), *Science* **250**, 1092–1097.

Soohoo, R. F. (1985), *Microwave Magnetics*, New York: Harper & Row.

MAGNETIC INSTRUMENTATION

See Measurement of Magnetic Properties

MAGNETIC MATERIALS

C. D. GRAHAM, JR., *Department of Materials Science, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.*

Introduction ... 1
1. **Classes of Magnetic Materials** ... 1
1.1 Paramagnets and Diamagnets ... 1
1.2 Ferromagnets, Antiferromagnets, and Ferrimagnets ... 2
2. **Fundamental Magnetic Properties** ... 3
2.1 Saturation Magnetization ... 3
2.2 Magnetic Field ... 4
2.3 Magnetic Induction ... 4
2.4 Curie Temperature; Néel Temperature ... 4
2.5 Anisotropy ... 4
2.6 Magnetostriction ... 5
2.7 Magnetic Domains ... 6
3. **Technical Magnetic Properties** ... 7
3.1 Remanence ... 7
3.2 Coercive Field (Coercive Force) ... 7
3.3 Permeability ... 8
3.4 Hysteresis Loop and Power Loss ... 9
3.5 Energy Product ... 9
4. **Engineering Magnetic Materials** ... 9
4.1 Soft Magnetic Materials ... 9
4.1.1 Iron and Steel ... 10
4.1.2 Nickel-Iron Alloys (Permalloys) ... 10
4.1.3 Glassy Metals or Amorphous Alloys ... 10
4.1.4 Ferrites ... 11
4.2 Hard Magnetic Materials ... 11
4.2.1 Alnico ... 11
4.2.2 Hard Ferrites ... 11
4.2.3 Rare-Earth Compounds ... 11
4.2.4 Bonded Magnets ... 12
4.3 Information Storage Materials ... 12
4.3.1 Particulate Media ... 12
4.3.2 Continuous Media ... 12
4.3.3 Magneto-optical Materials ... 12
Glossary ... 13
Further Reading ... 13

INTRODUCTION

Magnetic materials are widely used in industrial societies: they are essential for the generation and transmission of electric power, for electric motors, for recording and playback of sound, pictures, and numerical data, and for a great many other purposes. This article reviews the basic phenomena of magnetism, including units and terminology, with emphasis on the quantities that are important for applications. No treatment of the basic physics of magnetism is attempted (see MAGNETOSTATICS; MAGNETS). The magnetic materials of greatest usefulness for magnetic purposes are described, and some representative properties are given.

1. CLASSES OF MAGNETIC MATERIALS

1.1 Paramagnets and Diamagnets

All materials may be classified according to their magnetic behavior as paramagnetic, diamagnetic, or ferromagnetic. Paramagnetic and diamagnetic materials are only weakly magnetic; they are not obviously attracted to or repelled by a permanent magnet, and sensitive instruments are required to detect their magnetization. Paramagnets show positive magnetization in a positive field; diamagnets show negative magnetization in a positive field; that is, they become (very slightly) magnetized in a direction op-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

posite to that of the applied field. The magnetic behavior of paramagnets and diamagnets provides information about the electronic structure and bonding of the constituent atoms, but these materials have few if any practical uses based on their magnetic properties except when strongly magnetic behavior must be avoided. *Superconductors*, however, are strongly diamagnetic.

1.2 Ferromagnets, Antiferromagnets, and Ferrimagnets

This article will concentrate on strongly magnetic materials, which show *ferromagnetic* behavior. Some of these materials belong to a subcategory known as *ferrimagnets*, as explained later. These strongly magnetic materials are used for a very wide range of applications—most notably the production and transmission of electric power in generators and transformers; the conversion of electric power to mechanical motion in motors, relays, and loudspeakers; and the recording of audio, video, and digital information on magnetic tapes and disks.

In visualizing and understanding the behavior of strongly magnetic materials, it is useful to regard each "magnetic" atom as carrying a fixed magnetic dipole moment. That is, each magnetic atom acts like a tiny permanent magnet, with a north and a south pole, or alternatively as a tiny loop carrying a constant electric current. These atomic moments are conventionally represented by arrows, as shown in Fig. 1.

In ferromagnetic materials (iron, Latin *ferrum*, is the prototype), the atomic moments are held in parallel alignment by a coupling energy known as the exchange energy. Figure 2 illustrates this arrangement. A positive value of the exchange energy corresponds to parallel alignment of the atomic moments. Only three elements are ferromagnetic at room temperature: iron, nickel, and cobalt. The rare-earth metal gadolinium becomes ferromagnetic just below room temperature, and several other rare-earth elements are ferromagnetic at lower temperatures (often with more complex spin structures). It is also possible to have a ferromagnetic arrangement of magnetic atoms in a compound where some of the atoms have no moment, as shown schematically in Fig. 3. The best-known example is chromium dioxide, CrO_2, which can be used as a video magnetic recording material.

The exchange energy can be negative, leading to antiparallel coupling between magnetic atoms. This occurs most commonly in oxides and other compounds, and is shown schematically in Fig. 4. These materials are called *antiferromagnets*; their magnetic be-

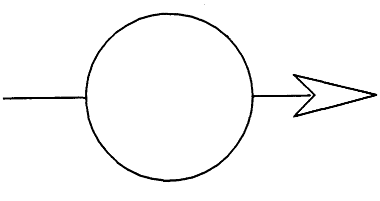

FIG. 1. Isolated atom carrying a magnetic dipole moment (schematic).

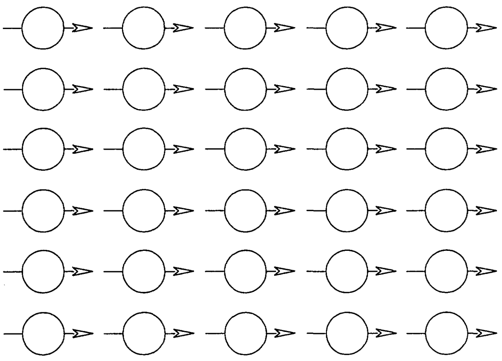

FIG. 2. Arrangement of moments in a ferromagnetic element (schematic).

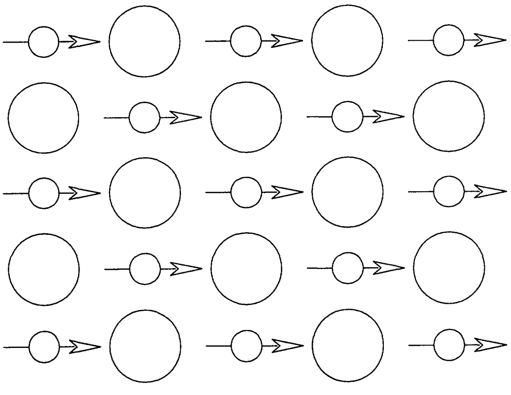

FIG. 3. Arrangement of moments in a ferromagnetic compound containing magnetic and non-magnetic atoms (schematic).

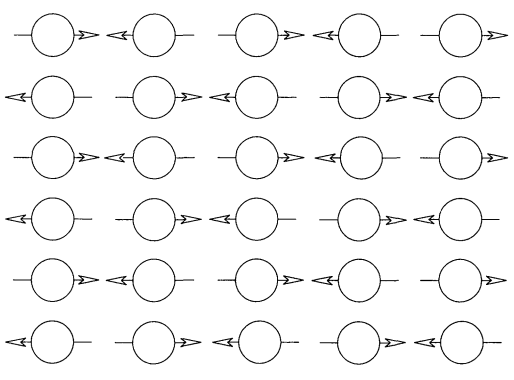

FIG. 4. Arrangement of moments in an antiferromagnetic compound (schematic).

havior is generally similar to that of paramagnets, although with a different temperature dependence (discussed in more detail later). Antiferromagnets are of little importance in applications, but a special case of antiferromagnetism, known as *ferrimagnetism,* has great practical utility. If two different magnetic atoms, each with a different value of the atomic moment, are arranged antiparallel, the net or resultant moment is not zero. This case is shown schematically in Fig. 5.

Practical ferrimagnets are oxides. The most important group of ferrimagnets are called ferrites; they are oxides having the spinel structure, with the general formula $M\mathrm{O} \cdot \mathrm{Fe_2O_3}$. The simplest example is $\mathrm{FeO} \cdot \mathrm{Fe_2O_3}$ or $\mathrm{Fe_3O_4}$ (magnetite). In this case the two different magnetic atoms are $\mathrm{Fe^{2+}}$ and $\mathrm{Fe^{3+}}$. A second class of useful ferrimagnets comprises the garnets, which have the general formula $\mathrm{RE_3Fe_5O_{12}}$, where RE represents a rare-earth metal, and which are used in microwave devices. Oxide ferrimagnets are electrical insulators, whereas most ferromagnets are metals, and hence are electrical conductors. For high-frequency applications, non-conductors have overwhelming advantages. For more detailed consideration of ferromagnets, antiferromagnets, and ferrimagnets, see the article MAGNETIC ORDERING IN SOLIDS.

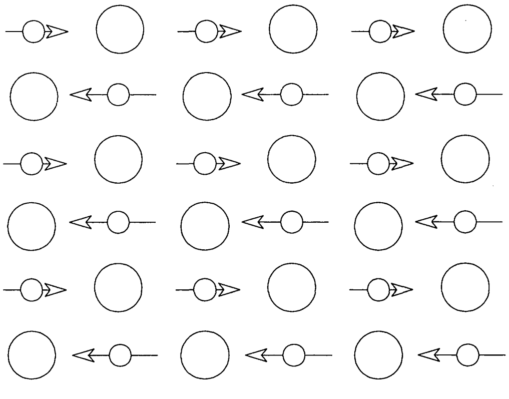

FIG. 5. Arrangement of moments in a ferrimagnetic compound; net magnetic moment is not zero (schematic).

2. FUNDAMENTAL MAGNETIC PROPERTIES

2.1 Saturation Magnetization

When a ferromagnet or ferrimagnet is subjected to a strong magnetic field, the net magnetization aligns parallel to the field and the material is said to be magnetically saturated. The macroscopic saturation magnetization is simply the vector sum of the atomic moments divided by the volume of the sample:

$$\mathbf{M}_s = (\Sigma \mathbf{m})V^{-1}. \tag{1}$$

Here $\mathbf{m}$ is an atomic moment, V is the sample volume, and $\mathbf{M}_s$ is the saturation magnetization. In SI units, energy moment is measured in $\mathrm{A \cdot m^2}$, reflecting the fact that a magnetic moment is equivalent to a current flowing in a circular loop. Magnetization is then moment/volume, $\mathrm{A \cdot m^2/m^3}$ or A/m. In cgs units, still widely used in applied magnetism, the magnetic moment is defined by the work done in rotating the moment in a fixed magnetic field. The electromagnetic unit of magnetic moment is then the erg/Oe (the oersted, abbreviated Oe, is the cgs unit of magnetic field strength), commonly called the emu (electromagnetic unit of moment). Magnetization in cgs units is therefore in emu/$\mathrm{cm^3}$, or (less commonly) erg/$\mathrm{Oe \cdot cm^3}$.

It is sometimes useful, especially when considering changes in magnetization with temperature, to eliminate thermal expansion effects by defining magnetization per unit mass, σ. The SI units are $\mathrm{A \cdot m^2/kg}$, and the cgs units are emu/g.

Physicists commonly express the moment of a single atom in a unit called the *Bohr* *magneton*, corresponding to the moment of a single unpaired electron spin. The Bohr magneton is 0.927×10^{-20} emu, or 0.927×10^{-23} $\mathrm{A \cdot m^2}$.

2.2 Magnetic Field

A magnetic field is produced by a current-carrying conductor. The field at the center of a long solenoid of length l, containing N turns of wire, and carrying a current i, is

$$H = Ni/l \quad \text{A/m [SI]} \quad \text{or}$$

$$H = (4\pi/10)Ni/l \quad \text{Oe [cgs]},$$

where SI length is in meters and cgs length is in centimeters, and current in both cases is in amperes.

2.3 Magnetic Induction

The quantity called magnetic induction or magnetic flux density (symbol B) is the vector sum of the magnetization and the field:

$$B = \mu_0(H + M) \text{ tesla [SI]},$$

$$B = H + 4\pi M \text{ gauss [cgs]}.$$

The quantity μ_0 is the magnetic constant or the permeability of free space. Its value is $4\pi \times 10^{-7}$ H/m or V s/A. The quantity B is necessary and useful primarily because changes in B generate a voltage in a surrounding wire, according to Faraday's law: $E = -k(dB/dt)$. This equation is the basis for the operation of generators and transformers.

2.4 Curie Temperature; Néel Temperature

As the temperature of a ferromagnet increases, the measured saturation magnetization decreases. The decrease is slow at first, but then increasingly rapid until at a temperature called the *Curie temperature* or *Curie point*, the magnetization disappears (Fig. 6). The thermal energy kT tends to randomize the direction of the atomic moments, while the exchange energy acts to keep the moments parallel. When the effect of thermal energy exceeds the exchange energy, the atomic moments are no longer aligned, and the material behaves as a paramagnet with a strongly temperature-dependent magnetization.

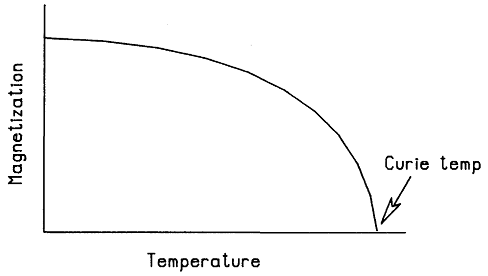

FIG. 6. Variation of magnetization with temperature in a ferromagnet, showing the Curie temperature.

The ideal curve of Fig. 6 is rather more complicated in practice; there is an increasingly strong field dependence of the magnetization at temperatures near the Curie point. This leads to experimental difficulty in measuring the Curie point with an accuracy better than a few K. Various methods have been proposed and used, but there is no generally accepted standard method. However, for practical purposes an uncertainty of a few K is seldom a major problem.

The disordering temperature for an antiferromagnetic is called the *Néel temperature*; experimentally it is evident as a maximum in the magnetic susceptibility $\chi = M/H$. In ferrimagnets, the two magnetic sublattices can each have a different temperature dependence of the magnetization, and so the curve of saturation magnetization versus temperature can take various forms, including the appearance of a compensation point (Fig. 7) where the magnetization goes to zero but then increases again before disappearing at the Curie point. The magnetization is very sensitive to both field and temperature near the compensation point, and this behavior can be used in some techniques for magnetic recording.

2.5 Anisotropy

The magnetic properties of a material may vary with the direction of measurement in the sample. This anisotropy may arise for several reasons. One source of anisotropy is magnetostatic or shape anisotropy: the magnetization prefers to lie parallel to a plane surface and parallel to the long axis of an elongated sample such as a wire or ribbon. This effect arises because surface magnetic charges or free magnetic poles exist wherever the mag-

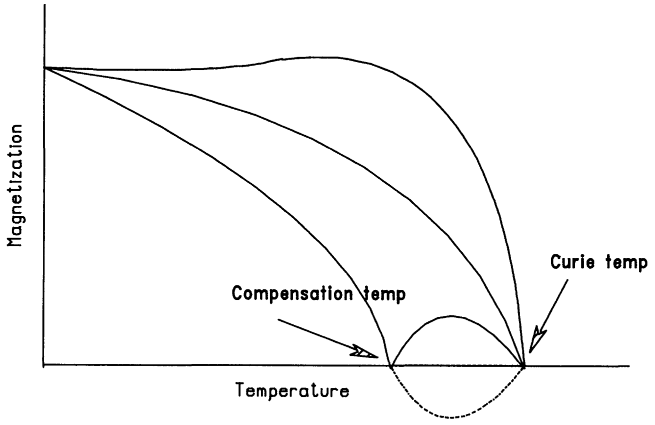

FIG. 7. Possible variations of net magnetization with temperature in ferrimagnets. Curves for three different materials, normalized at 0 K. The different behavior is due to different temperature variation of the moment of two oppositely-directed magnetic sublattices. The lowest curve shows a compensation point at which the net magnetization is zero; the thin line indicates that the net moment of the sample changes sign at the compensation point, although this will not be directly observed (thick line).

netization has a component normal to the sample surface. These free poles create magnetic fields in space, known sometimes as stray fields and sometimes as demagnetizing fields. The energy density of these fields is minimized when magnetization is parallel to the sample surface and when the free poles are as far apart as possible, which is why magnetization in the "long" direction is favored.

The demagnetizing field acting on a magnetized sample as a result of the free poles at its ends is given by

$$\mathbf{H}_d = N\mathbf{M}, \tag{2}$$

where N is a number called the demagnetizing factor. The value of N depends on the sample geometry. The demagnetizing factor has its maximum value (unity in SI; 4π in cgs) when the magnetization is normal to an infinite thin plate, and is zero in a rod or ribbon of infinite length. In real samples, the demagnetizing field is often comparable to or larger than the measuring field, and so must be corrected for in making measurements.

Most magnetic materials are crystalline, and in general the magnetization has a preferred crystallographic direction. This phenomenon is called the magnetocrystalline anisotropy, or just the crystal anisotropy. It can be measured by observing the torque exerted on a single crystal sample by a magnetic field, or from the difference in resonance fields in different crystallographic directions in a ferromagnetic resonance experiment (see NUCLEAR MAGNETIC RESONANCE, DETERMINATION OF STRUCTURE BY). If a polycrystalline material has its constituent crystals, or grains, oriented at random, there will be no macroscopic crystal anisotropy in the sample. Often, however, there is a preferred orientation, or texture, in the crystals, which is reflected in the macroscopic directional properties. A preferred orientation is beneficial in some cases, and may be deliberately induced.

The crystal anisotropy energy is expressed in a power series reflecting the symmetry of the crystal structure. In a uniaxial structure (hexagonal, tetragonal), the leading term has the form

$$E = K \sin^2\theta, \tag{3}$$

where θ is the angle between the magnetization and the uniaxial axis. In a cubic structure, the anisotropy is given by

$$E = K(\alpha_1^2\alpha_2^2 + \alpha_2^2\alpha_3^2 + \alpha_3^2\alpha_1^2), \tag{4}$$

where the α's are the direction cosines of the magnetization with respect to the three crystal axes. In both equations, E is the energy of the system as it depends on the direction of magnetization. The higher-order terms are usually not significant.

A third source of anisotropy arises from the interaction of mechanical stresses and the magnetostriction (see the next section).

2.6 Magnetostriction

When a magnetic material is cooled through its Curie temperature, it undergoes an elastic distortion (usually very slight) whose symmetry axis corresponds to the local direction of magnetization. If the direction of magnetization is changed, as by an external field, the distortion follows the magnetization, and the sample undergoes a macroscopic change in dimensions. This change in dimensions with the state and direction of magnetization is called *magnetostriction*. The measured strain (change in length per unit length, $\Delta l/l$), depends both on the direction of magnetization in the crystal and on the direction in which the strain is measured. In a cubic system, the simplest expression is

$$\begin{aligned}\Delta l/l &= \lambda_s\\ &= \tfrac{3}{2}\lambda_{100}(\alpha_1^2\beta_1^2 + \alpha_2^2\beta_2^2 + \alpha_3^2\beta_3^2 - \tfrac{1}{3})\\ &\quad + 3\lambda_{111}(\alpha_1\alpha_2\beta_1\beta_2 + \alpha_2\alpha_3\beta_2\beta_3\\ &\quad + \alpha_3\alpha_1\beta_3\beta_1),\end{aligned} \tag{5}$$

where the α's are the direction cosines of the magnetization and the β's are the direction cosines of the direction of the strain. The constants λ_{100} and λ_{111} are called the magnetostriction constants or magnetostriction coefficients. In the case of an isotropic material, such as a random polycrystal, a much simpler expression holds:

$$\Delta l/l = \tfrac{3}{2}\lambda_s(\cos^2\Theta - \tfrac{1}{3}), \tag{6}$$

where Θ is the angle between the magnetization and the strain axis.

These equations are based on an assumed demagnetized state in which the magnetization is completely random; the actual measured change in length depends strongly on the initial magnetic state of the sample, and may be much less than the value predicted by Eqs. (5) and (6).

In most magnetic materials the magnetostrictive strains are small, of the order of 10^{-5}, although there are materials where the strains can approach 10^{-2}. The dimensional changes can produce acoustical noise (generally unwanted) when magnetic materials are magnetized with ac fields; the magnetostriction also can make the properties of a material sensitive to applied stress. In particular, the interaction of an applied elastic stress with the magnetostriction creates another source of magnetic anisotropy, given (in the isotropic case) by

$$E = \tfrac{3}{2}\lambda_s\sigma\cos^2\Theta, \tag{7}$$

where σ is the applied stress and Θ is the angle between the magnetization and the stress axis.

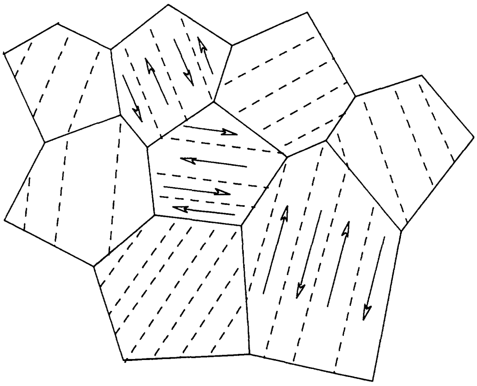

FIG. 8. Polycrystalline material containing multiple magnetic domains (schematic).

2.7 Magnetic Domains

According to Fig. 2, in a ferromagnetic material all the atomic moments are aligned parallel, so that one would expect the material to be magnetically saturated at all times. The demagnetized, or partially magnetized, state is achieved by the material being subdivided into small regions, called *domains*, each magnetized locally to saturation but adding vectorially to a lower value, or to zero. Figure 8 indicates schematically how domains may subdivide the crystals of a polycrystal; Fig. 9 shows how the motion of domain walls changes the magnetization.

A domain wall is a region in which the lo-

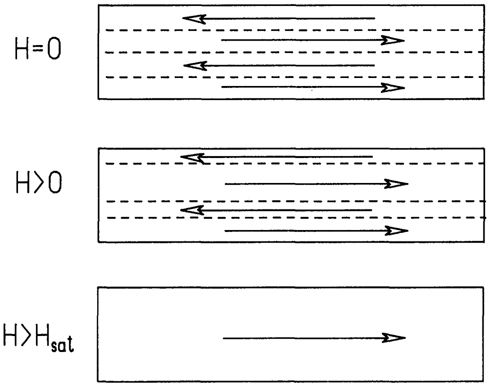

FIG. 9. Change in magnetization caused by domain wall motion (schematic).

cal magnetization rotates from one easy direction to another. In a cubic material with $K > 0$, the easy directions are either 90° or 180° apart, and the domain walls are classified as 90° or 180° walls. Figure 10 shows the structure of a 180° wall. In most ferromagnetic and ferrimagnetic materials, the domain-wall thickness is several hundred atomic spacings. This means that there is little interaction between a domain wall and a single impurity atom or a lattice vacancy or an isolated dislocation. There is, however, a strong interaction with a second-phase particle whose magnetization is different from that of the matrix, or with a region of local elastic strain.

The magnetization process at low fields consists primarily of the motion of domain walls. In general, the walls do not move smoothly and continuously, but in a series of local displacements called Barkhausen jumps, resulting from local pinning of the walls by magnetostatic and magnetoelastic forces. In an electrically conducting material, domain-wall motion is further limited by the fact that wall motion causes a local change in magnetic induction, and this change in induction generates local voltages (by Maxwell's equations) which in turn drive local *eddy currents*. These eddy currents dissipate energy by generating heat, and act to retard domain-wall motion. In electrically insulating materials, this effect is practically absent, although at higher frequencies there are energy losses associated with the rapid change in the direction of magnetization.

At high fields, as the magnetization approaches saturation, the magnetization process consists primarily of the slow rotation of the local magnetization away from the local easy direction toward the direction of the applied field. This is a basically reversible process, which causes very little energy loss.

3. TECHNICAL MAGNETIC PROPERTIES

3.1 Remanence

The curve of magnetization versus field of a ferromagnet or ferrimagnet is not only nonlinear (Fig. 11) but also irreversible (Fig. 12). When a magnetic material is saturated, and the field is then removed, some magnetization remains when the field is zero. The remaining magnetization is called the remanence M_r; it is often expressed as the remanence ratio M_r/M_s. In permanent magnets and in magnetic recording materials, high remanence is desirable; in materials for ac applications such as motors and transformers, low remanence is usually wanted.

3.2 Coercive Field (Coercive Force)

Since the magnetization is not zero when the field is zero (Fig. 12), it is necessary to apply a negative or reverse field in order to make the magnetization zero. The negative field required to reduce the magnetization to zero is called the coercive field or coercive force H_c (regarded as a positive number).

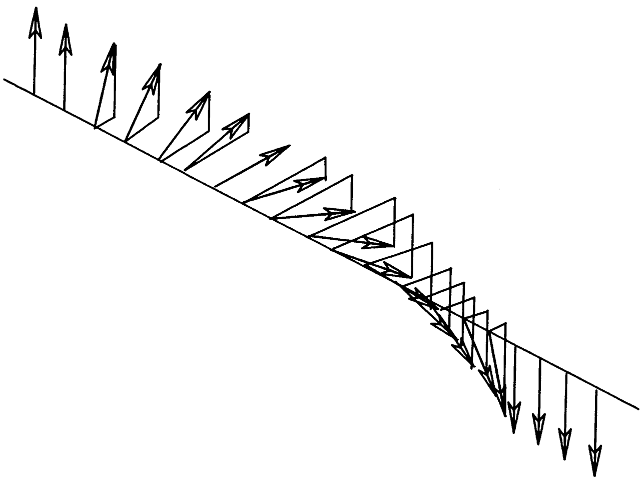

FIG. 10. Arrangement of atomic moments through a 180° domain wall.

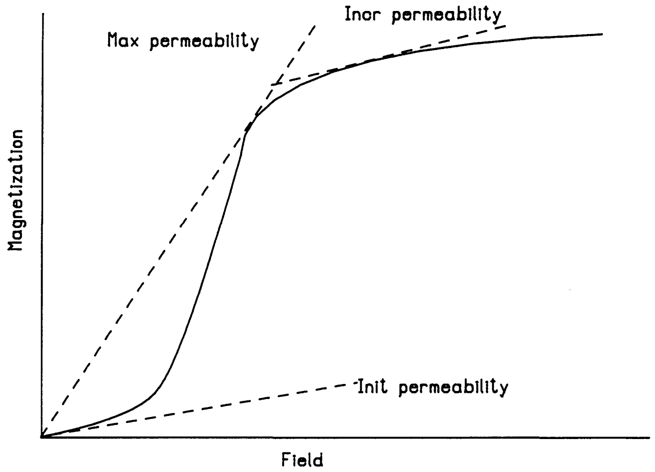

FIG. 11. Magnetization curve (B vs H) showing initial permeability, maximum permeability, and incremental permeability.

Strictly, this quantity is the intrinsic coercive field H_{ci} or H_c^i. The ordinary or normal coercive field H_c is the field required to reduce the magnetic induction $B = H + M$ to zero. (We ignore the complexities of magnetic units here.) Except in permanent-magnet materials, H is generally small compared to M, and the distinction between H_c and H_{ci} is unimportant. In permanent magnets, however, H can be comparable to or larger than M, and H_c and H_{ci} can be very different. H_{ci} is always larger than H_c.

3.3 Permeability

If a magnetic field is applied to a demagnetized sample (that is, a sample whose net magnetization is zero), the resulting curve of B vs H is known as the magnetization curve. There is a small field region over which the induction varies nearly linearly and reversibly with field. The slope of this region is known as the initial permeability, μ_0 (note that this is not the magnetic constant, for which the same symbol is used). With increasing field, the magnetization rises rapidly and irreversibly. The maximum slope of a line joining the origin and the magnetization curve is the maximum permeability $\mu_{\max}$, sometimes called simply the permeability. This is the quantity often used as a measure of magnetic quality for soft magnetic materials, although it is seldom directly useful in design. At any point on the magnetization curve, a small ac field may be applied, and the slope of the resulting minor loop is the

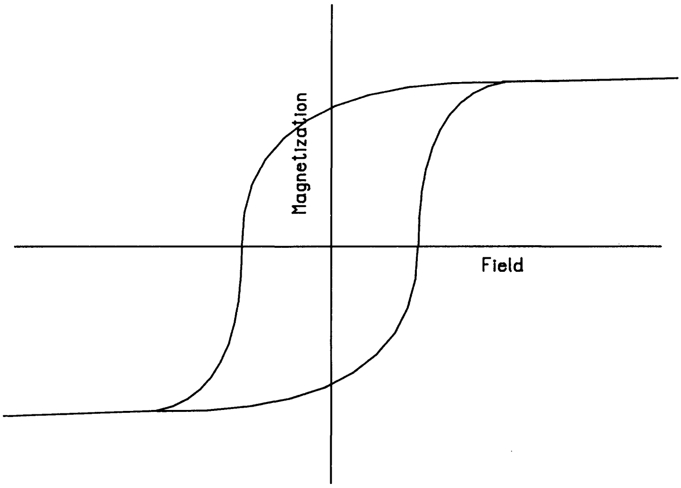

FIG. 12. Hysteresis loop of a ferromagnetic or ferrimagnetic material.

incremental or reversible permeability, μ_i. These quantities are shown in Fig. 11.

3.4 Hysteresis Loop and Power Loss

Work must be done in order to magnetize or demagnetize a magnetic material. The area enclosed by the hysteresis loop (Fig. 12) is proportional to the energy lost as heat when the material is magnetized to saturation alternately in the positive and negative directions. This energy loss is of primary importance in many ac applications of magnetic materials. The energy lost as heat reduces the efficiency of the device, and the heat may cause a local rise in temperature that is damaging.

The hysteresis loop is usually measured slowly, giving what is called a dc loop, but it may be measured at any desired frequency. The loop changes shape as the frequency is increased, notably by an increase in the coercive field and consequently of the enclosed area. This effect is especially marked in electrically conducting materials, because the energy lost to eddy currents increases as the frequency increases.

Most of the useful information about the magnetic properties of a ferromagnetic or ferrimagnetic material is contained in the hysteresis loop. The fact that the magnetization is both nonlinear and irreversible makes the behavior of magnetic materials difficult to handle mathematically, either analytically or by computer approximation.

3.5 Energy Product

A useful and widely used figure of merit for permanent-magnet materials is the maximum energy product $(BH)_{\max}$. It is defined as the area of the largest rectangle under the B vs H hysteresis loop in the second quadrant ($M > 0, H < 0$). The second quadrant is important because permanent magnets are always used to create a magnetic field in space, and therefore always operate in the presence of their own demagnetizing field. The maximum energy product defines the amount of magnetic material needed to create a given field in a given volume of space. Higher energy product means less material is required.

4. ENGINEERING MAGNETIC MATERIALS

4.1 Soft Magnetic Materials

This class of materials is used where the magnetization is required to change frequently in magnitude and/or direction; these uses are broadly considered ac applications. Examples are transformers, motors, generators, tape recorder heads, microwave equipment, and many other items. The name "soft" derives from the fact that the original material for this class of use was a mechanically soft, relatively pure iron, as opposed to the first permanent magnets, which were hardened steel. The terminology is no longer correct in its original sense, because there are mechanically hard materials that are magnetically soft, and also moderately soft and ductile materials that are magnetically hard.

A soft magnetic material has a low coercive field, and so it can be magnetized and demagnetized easily. This naturally leads to a hysteresis loop with small enclosed area, which means a small loss of energy in the form of heat during each magnetization cycle. The desired value of remanence may be large or small depending on the application. These requirements are met when the domain walls can move as freely as possible. This is accomplished by making the material of high purity and large crystal or grain size, and avoiding mechanical stresses or plastic strain. If the material has nonzero crystalline anisotropy, the use of a preferred anisotropy, or crystal texture, is desirable; in such a material, the individual crystals are not oriented at random, but the magnetic easy directions are preferentially aligned in the direction of the desired magnetization when the material is in use. This is not always easy or practical, but is done on a large scale in the production of high-quality transformer steel, consisting of Fe plus about 3 wt. % silicon.

It is also desirable to have the electrical resistivity as high as possible. This reduces the power losses due to circulating electric eddy currents. Metallic materials for ac use are always in the form of electrically insulated sheets or laminations, arranged with the sheets parallel to the direction of magnetization. Since the eddy currents flow perpendicular to the direction of magnetization, the layered structure interrupts the current path and reduces the eddy-current loss. Electrically insulating magnetic materials, known generically as ferrites, do not need to be laminated and are used as solid blocks.

Table 1 gives typical values for saturation induction (flux density) and coercive field for several common soft magnetic materials.

4.1.1 Iron and Steel The cheapest soft magnetic material is commercially pure iron, or very low-carbon steel. It is used when cost is a major consideration, or when efficiency is not important. The latter condition may apply, for example, in very small motors or in motors that operate only occasionally (like vacuum cleaners for home use). Higher-quality grades of magnetic steel contain added silicon, to increase the electrical resistivity; the upper limit of silicon content (about 3%) is set by the increasing brittleness of the steel as silicon is added. The best grades of magnetic steel are treated in a fairly elaborate series of rolling and annealing treatments to develop a preferred crystallographic orientation; this material is known as grain-oriented steel. The usual production processes for this material result in a very large grain or crystal size in the sheet; this is undesirable because it leads to a large magnetic domain size, and increases the alternating-current losses. This can be overcome by "laser scribing," which consists of rastering a fine laser beam over the sheet, creating a series of fine lines on the surface of each grain that act as nucleation sites for domains.

4.1.2 Nickel-Iron Alloys (Permalloys) Better soft magnetic properties are obtained in the nickel-iron alloys, known by their original trade name of Permalloy. There are two families of permalloys, one containing approximately equal amounts of iron and nickel, the other containing about 75%Fe-25%Ni. Both may contain small amounts of additional elements, to increase resistivity or control chemical ordering. The magnetic properties of these alloys depend quite sensitively on the details of their heat treatment, and they are mechanically soft and easily damaged by bending or other deformation.

Table 1. Soft magnetic materials (representative values).

Material	B_{sat}(T)	H_c (A/m)
Iron (commercial purity)	2.17	80
Steel (low carbon)	2.17	100
3% silicon steel (unoriented)	2.0	40
3% silicon steel (oriented)	2.0	8
Permalloy (79% Ni, 17% Fe, 4% Mo)	0.8	1
MnZn ferrite	0.4	5
NiZn ferrite (high electrical resistivity)	0.3	80
Amorphous $Fe_{81}B_{13.5}Si_{3.5}C_2$ (Allied Corp Metglas 2605SC)	1.6	3

4.1.3 Glassy Metals or Amorphous Alloys A relatively small number of metallic alloy compositions may be made in the form of noncrystalline, or amorphous, material. These are solids in which the atoms are not arranged in a regular crystal structure, but are densely packed in a random structure; these materials are also known as glassy metals. A number of these alloys containing large fractions of Fe or Co are ferromagnetic, and have useful magnetic properties. They have the general composition $(TM)_xM_{1-x}$, where TM represents one or more of the transition-metal elements, and M represents one or more of the metalloid or glass-forming elements, such as B, Si, C, or P. The value of x is in the range of 75 to 85 at. %. A widely used composition is $Fe_{78}Si_{10}B_{12}$, which combines relatively high magnetization with low losses, and is used in power transformers. The composition $(Co_{93}Fe_7)_{80}(SiB)_{20}$ has lower magnetization but near zero magnetostriction, and is superior for some applications.

These materials are made by rapid solidification: a pool of liquid alloy is cooled by a rapidly rotating cold metal wheel, which carries with it a thin layer of solid alloy. The resulting strip is about 25 μm thick and may be 10 cm or more in width. The electrical resistivity is many times higher than that of iron. Since these materials have no crystal structure, they have no crystal anisotropy, although they do develop a small but significant anisotropy when annealed in a magnetic field. The combination of high electrical resistivity and small thickness leads to very low power losses under ac excitation. They are, however, relatively costly and tend to be brittle after annealing; also the small thickness makes them difficult to use in some conventional manufacturing operations.

4.1.4 Ferrites The ferrimagnetic spinels known as ferrites have relatively low magnetization, but are electrical insulators. They are used exclusively for high-frequency applications where the power losses in metallic materials would be intolerable. These are ceramic materials, prepared by pressing powders into the desired shape and then sintering at high temperature to bond the particles together by solid-state atomic diffusion. By long experience, a relatively small number of compositions have emerged as standard, although variations in preparation conditions can influence the magnetic properties in useful ways.

4.2 Hard Magnetic Materials

Table 2 lists representative values of magnetic properties of some well-known hard magnetic materials. They are discussed in more detail in the following sections.

4.2.1 Alnico The first widely used permanent magnets (aside from naturally occurring magnetite lodestones) were hardened steels of various compositions. They were bulky and not very stable in their magnetic properties, and were quickly replaced when better materials became available.

The first of these better materials emerged in the 1930s from experiments on Fe-Ni-Al alloys intended as high-temperature structural materials. Eventually there resulted a family of materials known as the Alnicos, containing iron plus aluminum, nickel, and cobalt (AlNiCo), plus often other elements such as Ti or Zr. These are brittle multiphase alloys, made usually by casting but sometimes by powder compaction and sintering. They are much superior to steel in their permanent-magnet properties, but have now largely been replaced by newer materials except (because of their relatively high Curie temperatures) in applications where the magnetic properties must be maintained at temperatures above room temperature.

4.2.2 Hard Ferrites The most widely-used permanent-magnet materials are a family of ferrimagnetic oxides with hexagonal crystal structure and the general composition $MFe_{12}O_{19}$ (M = Ba or Sr), known generically as hard ferrites. They have high coercive fields, but relatively low magnetization, and they are low in cost. They are used in most loudspeakers, small motors, holding magnets, etc., unless the special properties of more expensive materials are required.

4.2.3 Rare-Earth Compounds The rare-earth elements became available in industrial quantities as a result of research carried out during World War II, and it was quickly established that many of these elements are strongly magnetic with large anisotropy. Their Curie temperatures are all below room temperature, and so they are not useful as magnetic materials in pure form. However, there are numerous binary and ternary intermetallic compounds of the rare earths with transition metals and other elements, some of which are ferromagnetic with Curie temperatures substantially above room temperature. Of these, a few have large positive crystal anisotropy, and can make excellent permanent magnets. There are three major families of rare-earth permanent magnets, based on the compounds $SmCo_5$, Sm_2Co_{17}, and $Nd_2Fe_{14}B$ (Sm is samarium and Nd is neodymium).

All three of these materials are magnetically superior to any of the previously available permanent magnets. Their only serious disadvantage, aside from cost, is their relatively low Curie temperature. This means that

Table 2. Hard magnetic materials (representative values).

Material	B_r (T)	H_c (kA/m)	$(BH)_{max}$ (kJ/m^3)
Cobalt steel	1.0	12.5	6
Alnico (isotropic)	0.6–0.8	60–100	40–60
Alnico (anisotropic)	1.1–1.4	60–120	50–70
Ba or Sr ferrite	0.43	180–240	35
$SmCo_5$	0.85	650	150
Sm_2Co_{17} [$Sm_2(Co_{.68}Cu_{.10}Fe_{.21}Zr_{.01})_{14.8}$]	1.1	525	250
$Nd_2Fe_{14}B$	1.25	900	300

their magnetic properties are fairly sensitive to temperature, degrading as the temperature rises. This problem is the most serious with $Nd_2Fe_{14}B$, which is otherwise the best and cheapest of the rare-earth magnets. The Curie temperature can be raised by replacing Fe with Co, but this unfortunately lowers the coercive field. NdFeB magnets are also susceptible to corrosion in humid atmospheres, but various protective schemes can be used to minimize this problem.

These materials are produced by pressing and sintering (heating below the melting point) small particles at high temperatures. The particles may be produced in various ways, but commonly are mechanically ground from bulk materials. The NdFeB magnets may also be made from flake material made by a rapid-quenching process (see Sec. 4.1.3).

4.2.4 Bonded Magnets All of the common permanent-magnet materials are mechanically hard and brittle. Simple shapes can be made by pressing and sintering (or by casting in the case of Alnico), but complex shapes can only be made by abrasive cutting and grinding. This is slow and expensive, and so in cases where the highest magnetic properties are not required, either ferrite or rare-earth permanent magnets may be ground into coarse powders and embedded in a matrix of a rigid or flexible polymer. These bonded magnets are easily molded by standard polymer equipment, and are furthermore not subject to breakage. Bonded magnets can also be made by embedding magnetic particles in a metal matrix.

4.3 Information Storage Materials

Conventional magnetic recording, as used in audio and video cassettes and in hard and floppy computer disks, employs a highly localized magnetic field produced by a *recording head* to magnetize a small region on a magnetically coated tape or disk. When the magnetized region is moved past a *playback head*, the local magnetization induces a voltage in a coil on the head, which is amplified to recreate the recorded signal. The same basic system is used for both analog and digital recording. The recording and playback heads (sometimes both functions are combined in a single head) must be magnetically soft materials with good wear resistance: ferrites, Permalloys, and amorphous alloys have been used. The recording material must have a low enough coercive field to be magnetized by the field from the recording head, and a high enough coercive field to retain its magnetization in the presence of negative fields (from the surroundings, or from neighboring magnetized regions). Thus properties intermediate between the usual soft magnetic and hard magnetic materials are required.

4.3.1 Particulate Media The commonest material for both audio and digital recording is an assembly of small particles of γ-Fe_2O_3, a metastable oxide prepared by a complex series of chemical steps. The particles are irregular elongated single crystals, about 100 nm long and 20 nm in diameter. They are embedded in a polymer matrix on a flexible or rigid substrate to make recording tape, floppy disks, or rigid (hard) disks. For increased coercivity, the particles may be given a surface coating containing Co ions.

Particles of other compositions are sometimes used, including chromium dioxide (CrO_2) and metallic Fe or FeCo. For high-density digital recording, particles of the hard magnetic oxide barium ferrite can be used. The recording characteristics of the various materials are different, so that they cannot be used interchangeably.

4.3.2 Continuous Media In order to obtain higher recording densities (more information in less space), continuous metal films, produced by vapor deposition in a vacuum chamber, may also be used for video and digital recording. Alloys of Co and Cr are usually employed for video tape, and Co-P or CoNi-P alloys for digital recording.

4.3.3 Magneto-optical Materials The demand for even higher storage densities has led to the development of optical and magneto-optical digital information storage. Pure optical storage (as in the audio or video compact disk) does not use magnetic materials and is not properly part of this article, but magneto-optical storage is a close relative. In this system, bits of digital information are written onto a magnetic film by pulses of laser light, which locally heat the storage material through its Curie temperature or compensation temperature, after which it cools in a controlled magnetic field.

Reading the stored information is done with a lower-power laser beam: the plane of polarization of the reflected laser light is rotated clockwise or counterclockwise depending in the direction of magnetization of the recorded bit. This optical phenomenon is known as the magnetic Kerr effect. This system has the advantage that only a tiny laser beam affects the magnetic recording material; there is no mechanical contact with a recording head and consequently no friction or wear. The recorded track is also much narrower than in conventional recording with a head, which permits much higher storage densities. The major problems with this system are concerns about the stability of the recording material after many cycles of recording, and the fact that erasing and recording are significantly slower than reading the stored information. A variety of materials have been suggested and tried for magneto-optical recording, and the best material is not yet established. Commercial systems using magneto-optical recording are just becoming available at the time of this writing (1992).

GLOSSARY

Antiferromagnet: Material in which the atomic moments are aligned antiparallel, so that no net magnetization results.

Anisotropy (Magnetic): Directional dependence of magnetic properties, due to sample shape, crystal structure, or stress.

Coercive Field (Coercive Force): Field necessary to reduce the magnetic induction (H_c) or magnetization (H_{ci}) to zero, after applying a saturating field.

Curie Temperature: Temperature at which the saturation magnetization drops to zero.

Diamagnet: Material that has no magnetization in the absence of a field, and develops a weak negative magnetization in an applied field.

Domain (Magnetic): Region in which all the atomic moments are aligned along a single axis.

Domain Wall: Boundary between two domains having different alignment directions.

Energy Product: Quality measure of a permanent magnet material, given by the area of the largest rectangle inscribed under the second quadrant of the hysteresis loop.

Ferrimagnet: Material in which the atomic moments are aligned antiparallel, but do not cancel because the number and/or magnitude of opposing moments is not equal; may show strongly magnetic behavior.

Ferromagnet: Material in which all atomic moments are or can be aligned parallel; may have magnetization in the absence of a field; shows strongly magnetic behavior.

Hysteresis Loop: Plot of magnetic induction or magnetization against field, as the field is cycled from positive to negative values; the area enclosed by the loop is proportional to the energy lost as heat.

Magnetostriction: Change in dimensions of a material when the direction or magnitude of its magnetization is changed.

Néel Temperature: Temperature at which an antiferromagnet loses its magnetic ordering.

Paramagnet: Material that has no magnetization in the absence of a field, and develops a weak positive magnetization in an applied field.

Permeability: Ratio of magnetic induction to field, under some specified condition; if not otherwise specified, usually the maximum value of permeability.

Remanence: Magnetization or induction remaining at zero applied field after applying a saturating field.

Saturation Magnetization: Limiting value of magnetization reached in a strong applied field; value depends on temperature.

Further Reading

Brief Introductory Works

Crangle, J. (1977), *The Magnetic Properties of Solids*, London: Edward Arnold.

Jacubovics, J. P. (1987), *Magnetism and Magnetic Materials*, London: The Institute of Metals.

General Texts

Chikazumi, S. (1964), *Physics of Magnetism*, New York: Wiley.

Cullity, B. D. (1972), *Introduction to Magnetic Materials*, Reading, MA: Addison-Wesley. Emphasizes materials and applications.

Jiles, David (1991), *Introduction to Magnetism and Magnetic Materials*, London: Chapman and Hall. The most recent general text.

Specialized Encyclopedia

Evetts, Jan (Ed.) (1992), *Concise Encyclopedia of Magnetic & Superconducting Materials*, Oxford: Pergamon. Excellent, up-to-date, general and specialized articles.

Multivolume Reference

Wohlfarth, E. P. (Ed.) (1980–1988), *Ferromagnetic Materials*, Amsterdam: North-Holland. Multivolume treatise, with advanced coverage of theory and experiment.

Coverage of Specific Types of Magnetic Materials

Chen, Chih-Wen (1977), *Magnetism and Metallurgy of Soft Magnetic Materials*, Amsterdam: North Holland, Reprint (1987), New York: Dover.

McCaig, Malcolm (1977), *Permanent Magnets in Theory and Practice*, London & Plymouth: Pentech.

Mee, C. Denis, Daniel, Eric D. (Eds.) (1987), *Magnetic Recording*, Vol. 1: *Technology*, New York: McGraw-Hill.

MAGNETIC ORDERING IN SOLIDS

B. R. Coles, *Physics Department, Imperial College, London, England*

Introduction: Ordered Magnetic Structures 15
1. **The Nature of the Ordering Moments** 17
2. **Interactions between Moments** 18
3. **Molecular Field Models, Magnetizations, and Susceptibilities above Ordering Temperatures** 20
4. **Magnetic Order with Itinerant Electrons; Band Ferromagnetism and Spin-Density Waves** 22
5. **Excitations in Magnetically Ordered Solids** 24
6. **Magnetic Phase Transitions and Critical Phenomena** 24
7. **Magnetic Order in Intermetallic Compounds** ... 25
8. **Magnetic Order and Superconductivity** 27
9. **Ferromagnetic Domains and Magnetization Curves** 27
Works Cited 30
Further Reading 30

INTRODUCTION: ORDERED MAGNETIC STRUCTURES

In no other branch of condensed-matter physics has there been such a proliferation of new ground states as has been observed in studies of magnetic order in recent decades. Until 1932, when Néel postulated the existence of antiferromagnets, ferromagnetism enjoyed its solitary glory, but when neutron scattering made possible the investigation of detailed magnetic moment structures and their excitations, the real proliferation began. In the pure rare-earth metals alone there are helices, fans, cones, and *c*-axis modulated structures; and more structures can be added to the list if one includes the alloys and intermetallic compounds of the rare-earth metal. In all of those the moments that order are associated with electrons in inner shells localized on the atoms, but the itinerant 3*d* electrons of the metals of the first transition group made their own novel contribution when Cr justified Overhauser's prediction that a band of such electrons could prove unstable against the formation of a spin-density wave, with a wave vector incommensurate with the lattice periodicity and dictated by dimensions of the Fermi surface.

The richness of magnetic ordering phenomena, especially when one considers also the nature of the low-lying excitations and the character of the transition from the high-temperature disordered state, makes the field a happy hunting ground for theoretical physicists eager to try out the latest weapons in their conceptual and mathematical armories. In this article, in an encyclopedia of applied physics, however, it seems more appropriate to present mainly the experimental situation and the general concepts that provide some order (in the simplest sense) to what might otherwise seem to be a confusing and arbitrary collection of phenomena.

The principal ground states with which we are concerned are the following:

Ferromagnetism—Spontaneous parallel alignment of atomic moments (normally constrained by magnetocrystalline anisotropy to lie in some specific crystallographic direction). Parameters of interest are the size of the atomic moments, the Curie temperature above which order is destroyed, and the orientation of the moments in the crystal. We deal in a later section with the domain theory required to explain the nature of the room-temperature "demagnetized state" of iron.

Antiferromagnetism—Ordering of magnetic moments on sublattices such that the net magnetization vanishes. The simplest such

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

structures are those where the moments are arranged on crystal lattices like the simple cube or body-centered cube such that all moments that are nearest neighbors of a given up moment have their moments pointing down. However, even within the family of antiferromagnets that can be analyzed in term of sublattices a huge variety of structures is possible, including alternating planes of up and down moments and even a tetrahedral subdivision of the face-centered cubic lattice such that the moments of all four atoms constituting the basis $(0,0,0;\ 0,\frac{1}{2},\frac{1}{2};\ \frac{1}{2},0,\frac{1}{2};\ \frac{1}{2},\frac{1}{2},0)$ point toward the center of the tetrahedron they define. It is clear that the magnetic symmetry will in general differ from the lattice symmetry; the monatomic bcc antiferromagnetic is clearly only simple cubic with a basis, since the atom at $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$ is no longer the same as that at the origin.

Ferrimagnetism—A structure that is basically antiferromagnetic but where two sublattices carry different numbers of atomic moments, or moments of different sizes, so that in the ordered state a net magnetization remains. The name derives from the family of ferrites $MO \cdot Fe_2O_3$, of which magnetite Fe_3O_4 is the most famous example.

Helimagnetism and related structures—The wide variety of magnetic structures found in the rare-earth metals and their alloys, which was referred to above. Some of the relevant structures are shown in Fig. 1, and full details are given in various chapters in Wohlfarth (1980). These structures arise because the atomic moments are provided by well-localized $4f$ electrons (see Sec. 2), which can only couple via the conduction electrons (see Sec. 5). Their wave-vector–dependent susceptibility, $\chi(q)$, in conjunction with the character of the $4f^x$ state imposed by the crystal fields, dictates the magnetic structure.

Spin-density waves—Just as ferromagnetism can be associated with either localized or itinerant (band) electrons, so can antiferromagnetism. The sublattice approach described above encourages a local-moment viewpoint, but the itinerant $3d$ electrons of chromium sustain spin-density waves such that at any point in the crystal there is a net spin, the difference between spin-up and spin-down local densities. In some dilute alloys of

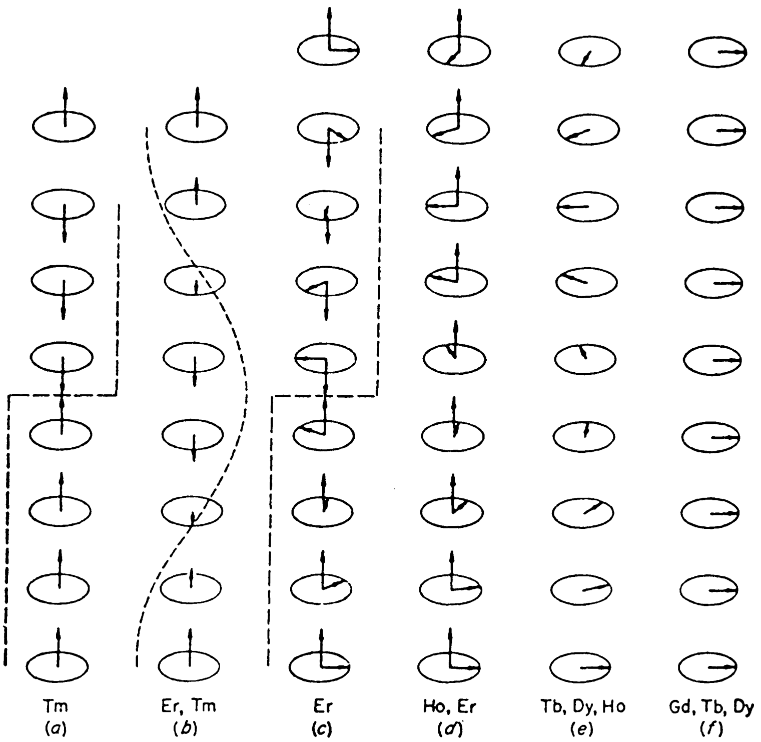

FIG. 1. Magnetic structures of the heavy rare-earth metals (after Koehler, 1972).

other metals in chromium the maximum in spin-up density is at the 0,0,0 position and the maximum in spin-down at $\frac{1}{2},\frac{1}{2},\frac{1}{2}$, and the structure is described as commensurate. Neutron diffraction then shows scattering intensity at the Bragg peaks such as (100), (111), etc., which would not be seen for body-centered cubic symmetry. In pure chromium, however, the wave vector of the magnetic modulation differs slightly from the lattice vector and satellite peaks appear in the neutron diffraction pattern; the structure is then described as incommensurate. (It can be recalled that in many of the helimagnetic rare-earths incommensurate structures are found, although the dominant ordered moments are localized.). Incommensurate structures have also been found in heavy-fermion intermetallic compounds (see HEAVY-FERMION PHENOMENA) with small ordered moments in narrow bands.

Spin glasses—Intensive studies have been made in recent years of the magnetic ground states associated with magnetic moments in amorphous materials or distributed at random over some points of the lattice of a nonmagnetic host. If, as they normally do, the interactions between the moments have both ferromagnetic and antiferromagnetic signs, the ground state, when the material is cooled in zero field below some temperature characteristic of the mean interaction strength, will have no net magnetization, but cannot show the sublattice structure of a conventional antiferromagnet. It is not clear whether this state, in which the moments are frozen, but frozen at random in one of a large number of possible configurations of similar free energy, should be regarded as "ordered"; and arguments continue about the nature of the transition from the paramagnetic state. Recent reviews are given by Binder and Young (1986) and Mydosh (1993). It is clear that if all interactions were ferromagnetic the randomness essential to spin-glass behavior would not prevent ferromagnetism; the other essential ingredient is the competition between ferro- and antiferromagnetic interactions.

1. THE NATURE OF THE ORDERING MOMENTS

In the above discussion of ordered magnetic structures the object becoming aligned was referred to as a magnetic moment, although it became a "spin" in the discussion of itinerant antiferromagnets. The use of "spin" and "moment" as synonyms is common in the magnetic literature, but is not of universal correctness. The origin lies in the evidence that the angular momentum that yields the magnetic moment of iron, like that yielding the moment of a free electron, is the spin angular momentum. More generally the magnetic moment μ of an atom is given by $g_J\sqrt{J(J+1)}\mu_B$, where μ_B is the Bohr magneton, the atomic unit of magnetic moment of size 9.27×10^{-24} J/T, J is the total angular momentum in units of $\hbar$, and g_J is the Landé spectroscopic g factor that takes into account the fact that J can have contributions from both the orbital angular momentum vector (**L**) and the spin angular momentum vector (**S**) and that the latter produces twice as much magnetic moment per unit of angular momentum as the former. Thus

$$g_J = \frac{3}{2} + \frac{S(S+1) - L(L+1)}{2J(J+1)}, \tag{1}$$

which is 1 for $S = 0$ and 2 for $L = 0$.

In materials where atomic or ionic magnetic moments do not interact, the moment given above is the quantity entering the expression for the magnetic susceptibility and is often called the effective moment, μ_{eff}, since the low-field susceptibility, which in the classical Curie-law expression was $N\mu^2/3k_BT$, becomes $N\mu_{\text{eff}}^2/3k_BT$ in a quantum-mechanical treatment. (N is simply the number of magnetic atoms or ions per unit volume.) This topic is discussed in detail in the article PARAMAGNETISM, and here we need only note that in most solid compounds of the $3d$ transition elements, strong crystal fields quench the orbital contribution to μ_{eff}, so that the moments responding to external fields or interacting with one another are "spin-only" moments; while in solids containing $4f$ elements crystal fields are in general smaller than spin-orbit interactions. One can then have situations where, at low temperatures, crystal field splittings, k_BT, and moment-moment interaction energies can be comparable, so that ordered moments have to be considered carefully. Compounds containing actinide ($5f$) elements that order magnetically are still more complicated, for while metals and alloys of the $3d$ elements are generally most appropriately treated with band models, and $4f$ elements with local-moment models, $5f$ electrons may display different characters, depending both on the element concerned and on those with which it is combined.

In transition-metal atoms the electrons in the partly filled shell are d electrons ($l = 2$) so that the effective moments are as shown in Fig. 2 for free atoms or ions, while for the rare-earth elements with $4f$ electrons ($l = 3$) they are as given in Fig. 3.

Even for noninteracting moments two complications to the above picture can arise. One arises from the effects of crystalline electric fields, and is so important in solids containing $3d$ atoms that the orbital moment is normally quenched and the effective moment is given by setting $J=S$ and $g_J = 2$, with the results shown in Fig. 2, since by Hund's rule S always takes its maximum possible value. For the rare-earth elements spin-orbit interactions are normally appreciably larger than either crystal field effects or k_BT, and μ_{eff} has to be calculated (for higher temperatures) by applying all Hund's rules to obtain the values of S, L, and J, but in a few cases the spin-orbit levels are separated by energies comparable with k_BT and J states other than the ground state play a role, which is most apparent when $J = 0$ for the ground state as in Eu^{3+}, since that would yield diamagnetism without the admixture in applied fields of contributions from higher states.

2. INTERACTIONS BETWEEN MOMENTS

The temperatures below which magnetic order is found vary enormously from very low values to temperatures above 10^3 K, but it is immediately clear that very rarely is the responsible interaction a simple magnetic dipole interaction, since the strength of such an interaction between moments at normal in-

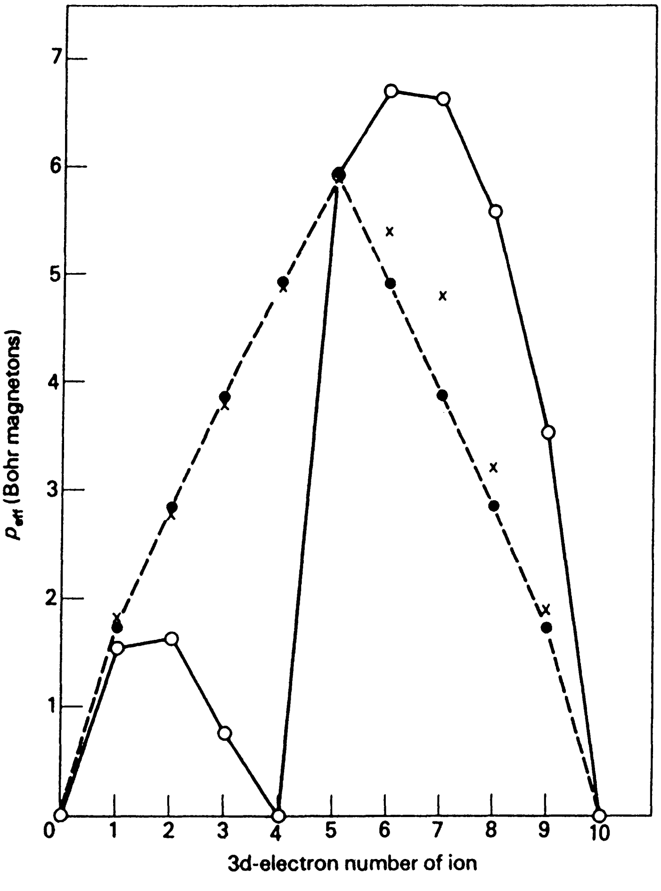

FIG. 2. The effective moments [expressed as $p_{\text{eff}}(= \mu_{\text{eff}}/\mu_B)$] for ions of the $3d$ series. ×, experimental values for salts. ○, theoretical values for full J; ●, theoretical spin only values.

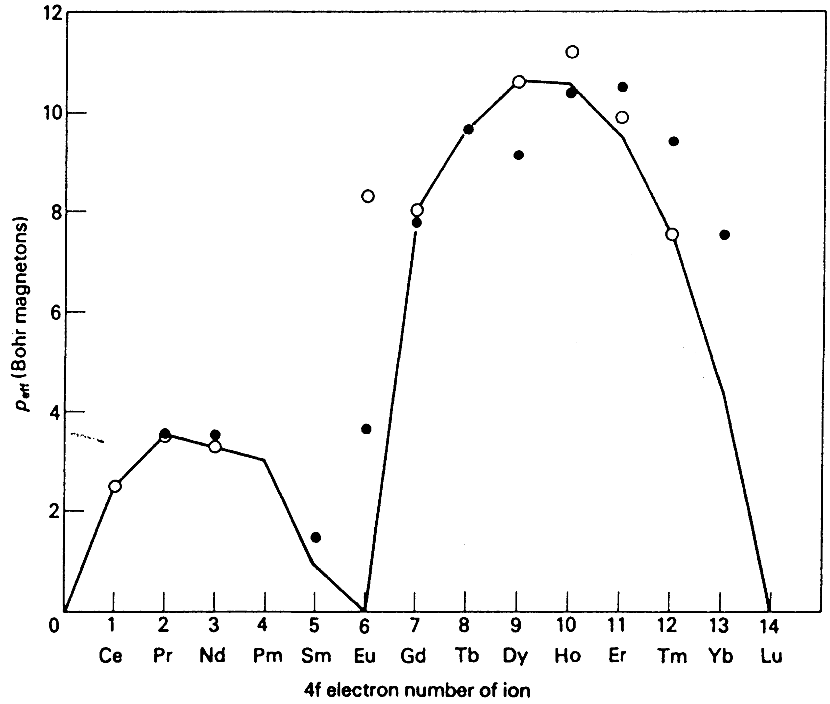

FIG. 3. The effective moments for ions of the 4*f* series (given as μ_{eff}/μ_B). ○, metals; ●, R_2O_3; —, theoretical full *J* value.

teratomic distances would yield ordering temperatures well below 1 K. The origin of interactions is instead the effect described as *exchange interaction*. This has its origin in Coulomb interactions between electrons, which in combination with the Pauli principle can favor either parallel or antiparallel moment alignment depending on the nature of the electron states. The situation is complicated by the fact that the concept of an exchange integral $\mathcal{J}$ emerged first in the Heitler-London treatment of the hydrogen molecule where $\mathcal{J}$ is negative, favoring antiparallel spins of the two electrons in the lowest energy (bonding) molecular orbital. In the carbon atom, however (with configuration $1s^22s^22p^2$), the Coulomb energy is reduced if the two *p* electrons occupy different atomic orbitals and Hund's rule makes the state of parallel spins lowest in energy. The same holds for two of the *p* electrons of the oxygen atom ($1s^22s^22p^4$).

The simplest object where this effect leads to alignment (positive $\mathcal{J}$) between spins on different atoms is the oxygen molecule O_2. Here bonding *molecular* orbitals are occupied by pairs of electrons with opposite spins, but two electrons remain to be placed in the two antibonding π orbitals and will reduce the Coulomb energy by occupying different orbitals with the parallel spins required by Hund's rule. (A complete treatment of correlation efforts justifies this empirical rule that the triplet state is lower in energy for two electrons in different orbitals.) The resultant magnetic moment makes oxygen gas paramagnetic. (In solid oxygen these moments take up an antiferromagnetic structure through a somewhat more complicated exchange interaction.) It may be noted that the traditional 3*d* ferromagnetic metals, Fe, Co, and Ni, have nearly full 3*d* bands where the highest-energy electrons occupy states of antibonding character. Unlike the O_2 situation, however, these electrons are now in energy bands and ferromagnetism occurs only if the exchange-energy gain is not offset by the need for the electrons yielding the net spin to occupy slightly higher energy states in the band.
]

We have seen how intra-atomic exchange gives a net moment to carbon and oxygen atoms, and the largest spin state is found for the Gd^{3+} ion ($4f^7$) where all seven spins are aligned (and $L = 0$). In gadolinium metal these electrons form part of the ion core (inside the $5s^25p^6$ full subshells) and cannot yield direct interatomic exchange interactions, so that the ferromagnetism of Gd has to be ascribed to *indirect exchange*. Here the polarization of the conduction electrons (in states derived from 6*s* and 5*d* atomic levels) by exchange interaction with a given $4f^7$ moment propagates through the lattice to couple neighboring moments. As with the well-known Friedel *charge*-screening clouds around impurities in metals, this *spin* polarization of the conduction electrons falls off with distance from the local moment and oscillates in sign, so that the effective interaction between two local moments in a bath of conduction electrons can be ferromagnetic or antiferromagnetic.

A final type of exchange of great importance in magnetic oxides is termed *superexchange*. In its simplest form this couples (antiparallel) the spins of two 3*d* ions (e.g., Fe^{2+}) through an intermediate O^{2-} ion. An up-spin 3*d* moment (unshielded by 4*s* electrons) produces an imbalance between the up-spin and down-spin charge clouds of the oxygen 2*p* electrons, the down spins being slightly pushed away toward the other 3*d* ion whose spin they align with their own.

Although it is a topic of rather specialized interest, mention might be made at this point of *nuclear magnetic ordering* where the moments interacting are those given by the spins of nuclei. These moments are smaller by factors of $\sim 2 \times 10^3$ than electronic moments, and no exchangelike mechanism is normally present, so that ordering occurs only at very low temperatures ($\sim 10^{-7}$ K). It has, however, been detected in a number of substances, including metallic copper, where the dipole interaction is assisted by indirect interaction through the conduction electrons. The latter is a Ruderman-Kittel interaction like that which couples the local 4*f* moments in Gd metal, but here the necessary coupling between the nuclear moment and the conduction electrons is the much weaker hyperfine interaction.

3. MOLECULAR FIELD MODELS, MAGNETIZATIONS, AND SUSCEPTIBILITIES ABOVE ORDERING TEMPERATURES

The original Weiss theory of ferromagnetism involved the phenomenological introduction of a molecular field $H_{\text{eff}} = N_W M$ where M is the magnetization and N_W is the molecular field coefficient. If the field acting on the moments is regarded as $\mathbf{H}_{\text{ext}} + \mathbf{H}_{\text{eff}}$ the susceptibility can be calculated by assuming that this is the field to be put into Curie's law $M/H = C/T$. (We have mentioned that C, which was $N\mu^2/3k_B$ in a classical treatment, becomes $N\mu_{\text{eff}}^2/3k_B$ in correct treatments of the low field susceptibility. For a further treatment see PARAMAGNETISM.)

Thus we have

$$M/(H + H_{\text{eff}}) = C/T, \tag{2}$$

so that χ ($= M/H$) is given by

$$\chi = C/(T - N_W C), \tag{3}$$

and the quantity $N_W C$ is normally denoted θ and is called the Curie-Weiss temperature. Thus the susceptibility in this simple approach is expected to diverge as T is lowered to θ.

To examine the behavior at lower temperatures one needs the more complete treatment of paramagnetism that yields the expression

$$M = N g_J J \mu_B B_J(x)$$

where $B_J(x)$ (the Brillouin function) is given by

$$B_J(x) = \frac{2J+1}{2J} \coth\left(\frac{2J+1}{2J} x\right) - \frac{1}{2J} \coth\left(\frac{x}{2J}\right) \tag{4}$$

and $x = g_J J \mu_B H / k_B T$.

From Fig. 4, where we have plotted the spontaneous magnetization M_{sp} (for $H_{\text{ext}} = 0$) versus x with $H = N_W M$, (i.e., $M_{\text{sp}} = x \cdot kT / g_J J \mu_B N_W$), it is clear that below a certain temperature, when the straight line giving M_{sp} falls below the initial slope of the Brillouin-function expression for M, a finite value of

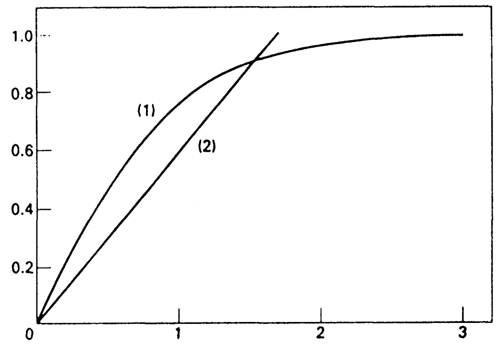

FIG. 4. Graphical solution for the spontaneous magnetization in the molecular-field model. The reduced magnetization M/M_{sat} is plotted as (curve 1) the Brillouin function of x ($= g_J J\mu_B H/kT$), (curve 2) H_{eff}/N_W, which for $H_{\text{ext}} = 0$ is linear in T and x.

M_{sp} is found for $H=0$. This spontaneous magnetization is plotted against T in Fig. 5, and it falls to zero at $T_c = N_W \mu_{\text{eff}}^2/3k_B = \theta$. This identity of T_c (the Curie temperature) and θ is a result of the mean-field treatment, and to explain the normally observed deviations of χ from $C/(T - T_c)$ just above T_c, account has to be taken of correlated fluctuations.

The approach to saturation at low temperatures ($M_s = Ng_J J\mu_B$)* is also not correctly given by the mean-field treatment for the same reason, and a spin-wave approach in that regime yields

$$M = M_0[1 - A(T/T_c)^{3/2}] \tag{5}$$

in good agreement with experiment.

As mentioned above, it was first pointed out by Néel that if a lattice of magnetic moments can be subdivided into two sublattices A and B with antiparallel interactions between them, a similar mean-field treatment, taking $H_{\text{eff A}} = H + N_W M_B$ and a corresponding expression for $H_{\text{eff B}}$, yields

$$TM_A = C'(H - N_W M_B) \quad \text{and}$$
$$TM_B = C'(H - N_W M_A), \tag{6}$$

giving

$$TM = 2C'(H - N_W M) \tag{7}$$

and

$$M(T/2C' + N_W) = H \tag{8}$$

---

[*] Note that this is not $N\mu_{\text{eff}}^2$, since the maximum value of J in the field direction is J, not $\sqrt{J(J+1)}$.

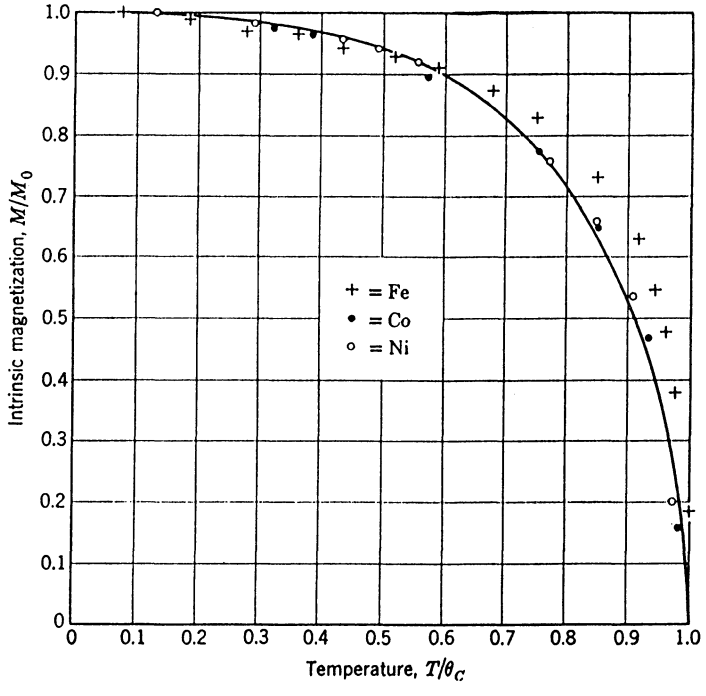

FIG. 5. The spontaneous (or intrinsic) magnetization given by molecular-field theory and the values for Fe, Co, and Ni.

so that, above the temperature of $C'N_W$, $\chi = 2C'/(T + \theta)$ where $\theta = N_W C'$.

Below a temperature T_N ($= C'N_W$) the sublattice magnetization of simple antiferromagnets is found by neutron scattering to follow the type of temperature dependence found for the spontaneous magnetization of simple ferromagnets. Since there is no net magnetization, the susceptibility is linear in field (for low fields), and its magnitude depends on the orientation of the field relative to the direction of the sublattice magnetization; $\chi_\parallel$ tends to zero as T tends to zero, and $\chi_\perp$ is only weakly T-dependent below T_N.

For both the ferromagnetic and antiferromagnetic cases it should be remembered that the linearity of $1/\chi$ with T will not be preserved if crystal-field-split levels have spacings comparable with T, and in some antiferromagnets with this character the ordered moments are given by the appropriate (to the symmetry) combination of M_J values that characterizes the ground state. When crystal fields produce a ground-state singlet i.e., a nonmagnetic atom) a complicated situation will exist with the possibility of magnetic order at low temperatures if the exchange interaction is larger than the excitation energy to higher crystal field levels.

4. MAGNETIC ORDER WITH ITINERANT ELECTRONS: BAND MAGNETISM AND SPIN DENSITY WAVES

Since it is known that the $3d$ electrons of Fe, Co, and Ni occupy bands in the metallic state, it is clear that the above local-moment treatment cannot be appropriate, although it should be remembered that not all the on-site correlations that provide the moments of ions of these elements in paramagnetic salts will be lost in the metallic state. Measurements of g factors show that in these metals the magnetic moments arise almost wholly from electron spin, and the saturation moments, expressed in Bohr magnetons, therefore give the number of unpaired electrons. The basic parameters are shown in the following table:

	Curie temperature (K)	Saturation moment (μ_B)
Fe	1044	2.22
Co	1398	1.72
Ni	627	0.61

In earlier work it was assumed that one-half (spin up) of the $3d$ band was fully occupied at spontaneous saturation at $T = 0$ and that the saturation moment gave the number of empty states in the spin-down band [Fig. 6(a)]. (It should be remembered that even this

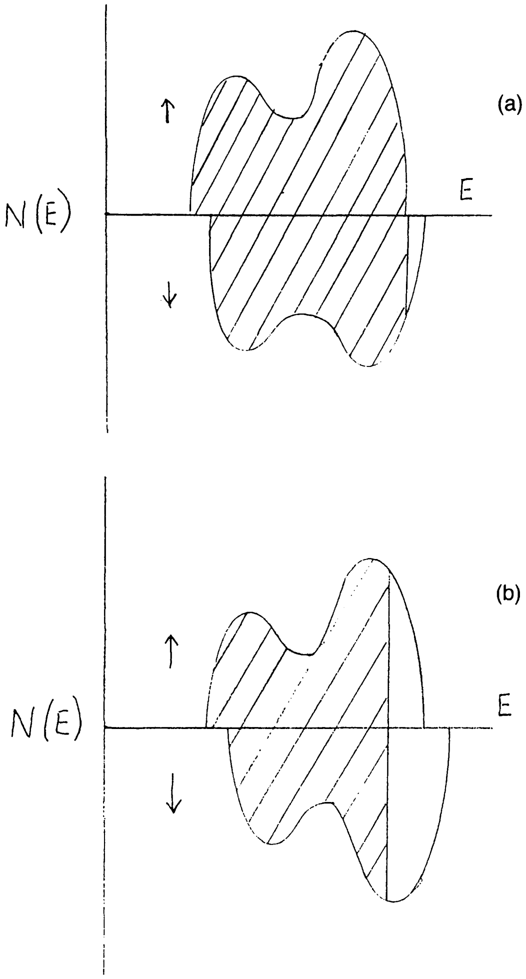

FIG. 6. Occupied and empty d-band states in ferromagnetic transition metals (schematic). (a) A "strong" (fully magnetized) metal, e.g., Ni; (b) a "weak" ferromagnetic, e.g., Fe.

need not be an integer since the 3*d* band overlaps a partly filled 4*s* band). It is now recognized that in Fe the Fermi surface lies in both spin bands, and the saturation moment is $(n_\uparrow - n_\downarrow)\mu_B$ where the *n*'s are the numbers of holes in the two half *d* bands [Fig. 6(b)]. For alloys between the 3*d* metals the maximum saturation moment is found at Fe 30% Co with $2.44\mu_B$. Figure 7 shows the so-called Slater-Pauling curve for binary alloys, and for the sequence Fe → Co → Ni → Cu this supports a collective *d*-band view of the electronic structures, which breaks down for such alloys as Co-Cr and Ni-Cr. It must also fail for Ni-Cu, since the termination of ferromagnetism at 60% Cu does not correspond to the filling of all 3*d* holes, and at higher Cu concentrations they are associated only with Ni atoms, but with $n_\uparrow = n_\downarrow$.

The theory of the $T = 0$ state of the itinerant-electron ferromagnets is now in a fairly satisfactory state (see Mattis, 1988), and the expression for the susceptibility,

$$\chi(0) = \chi_0(0)/[1 - I\chi_0(0)], \tag{9}$$

where $\chi_0(0)$ is the susceptibility at $q = 0$, which is a measure of the density of states that can be calculated from a good band-structure calculation, and I is the exchange energy, formalizes the Stoner criterion for ferromagnetism, $I\rho > 1$, where ρ is the electronic density of states. The situation at finite temperatures is less satisfactory, but the low-lying excitations are discussed in Sec. 5.

Itinerant-electron ferromagnetism is best considered through the wave-vector–dependent susceptibility $\chi(q)$. If the exchange interaction is such that as the temperature is decreased, the susceptibility enhanced by it diverges for the q value corresponding to twice the crystal lattice period (i.e., $q = \pi/a$ for a simple cube), the magnetic ground state will be antiferromagnetic. No pure element satisfies this condition, although a number of fairly dilute alloys of other transition metals in Cr do. Cr itself, as indicated in the Introduction, has an antiferromagnetic ground state ($T_N = 312$ K) characterized by a q value that couples two sheets of the Fermi surface but deviates slightly from the Brillouin-zone vector. (The stabilization energy arising from the introduction of a periodicity that matches the Fermi surface is very similar to that arising in a Peierls distortion of a lattice; with such a match, an energy gap appears that depresses the energy of the occupied states just below the Fermi level while raising that of empty states just above it.)

Even when the dominant contribution to the magnetic moments is made by the 4*f* electrons in the rare-earth metals, the incommensurate periodicities of the exotic structures (Fig. 1) are related to Fermi wave vectors associated with the conduction electrons that mediate the indirect exchange, as described in Sec. 2.

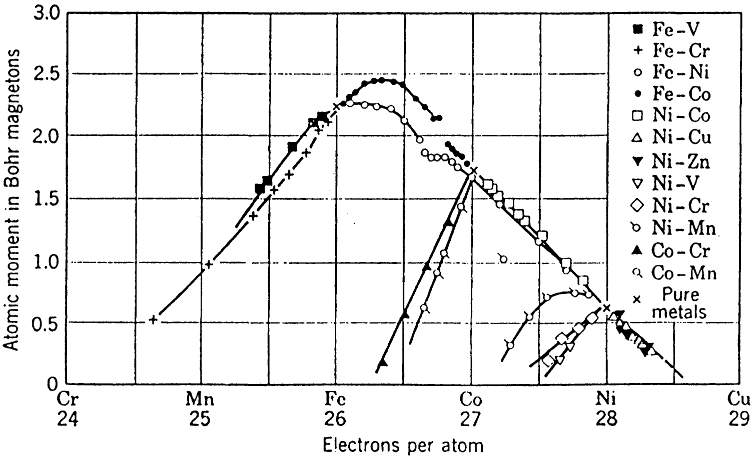

FIG. 7. Average atomic moments for iron-group metals and alloys as a function of electron concentration.

5. EXCITATIONS IN MAGNETICALLY ORDERED SOLIDS

In a simple molecular-field model of a ferromagnet, the natural excitation to consider is that given by turning over a single spin, requiring an energy $E \sim Jz$ where z is the number of nearest neighbors. As in the Einstein model of the specific heat of independently vibrating atoms on a lattice, this gives the (wrong) result of an exponential specific heat. In fact, the excitation is of the whole lattice, and although in the magnetic case this is quantized to give a magnetization less by μ than $N\mu$, the energy is reduced by the fact that the magnetization loss propagates through the lattice, as atomic vibrations do as phonons. The resultant excitation is thus a magnon or spin wave which for low energies can be shown to have an energy $\epsilon(q)$ equal to $\hbar\omega_q$ where $\omega_q = Dq^2$; D is called the spin-wave stiffness and, as with phonons, can be derived from inelastic neutron-scattering measurements. Some branches of ω versus q for ferromagnetic spin waves are shown in Fig. 8(a), and for antiferromagnetic spin waves in Fig. 8(b). It should be noted that in antiferromagnets there is an energy gap in the spin-wave spectrum, related to anisotropy, without which the initial region would be linear in q.

For itinerant-electron magnets, spin-wave-like excitations coexist with excitations from conduction-electron states below the Fermi level to those above, and these now include Stoner excitations, those from one spin half-band to the other. We then have the situations shown in Fig. 9 for the spin nonconserving excitations in (a) a "strong" itinerant-electron ferromagnet (one half-band full) and (b) a "weak" one like Fe where the Fermi level lies in both half-bands. It should be noted that in both cases spin waves above a certain energy will be strongly damped by dissociation into Stoner excitations. (At lower energies magnon-magnon interactions will give some broadening as do anharmonic effects in phonon spectra.)

6. MAGNETIC PHASE TRANSITIONS AND CRITICAL PHENOMENA

Magnetic ordering supplies some illuminating examples of phase transitions that differ from the first-order transitions with latent heats $\Delta L > 0$ such as those between liquid and solid or two crystallographically distinct phases. All such ($\Delta L = 0$) phase transitions can be characterized by an order parameter that goes smoothly to zero as the temperature is raised. The molecular-field model used for the magnetization gives M as a suitable order parameter and predicts a specific heat for a ferromagnet of the form shown in Fig. 10. Close to T_c this fails rather badly, as did the predicted behavior of $1/\chi$, and for the same reason. Correlated fluctuations are at the root of the problem, and a great deal of effort has been expended in attempts to incorporate these within theories of the magnetic and thermodynamic behavior at magnetic ordering temperatures. (For a thorough

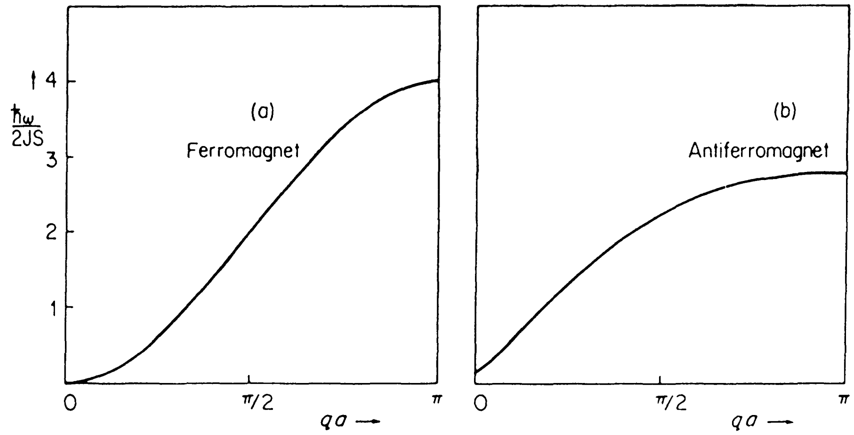

FIG. 8. Spin-wave dispersion curves for (a) a ferromagnetic, (b) an antiferromagnetic.

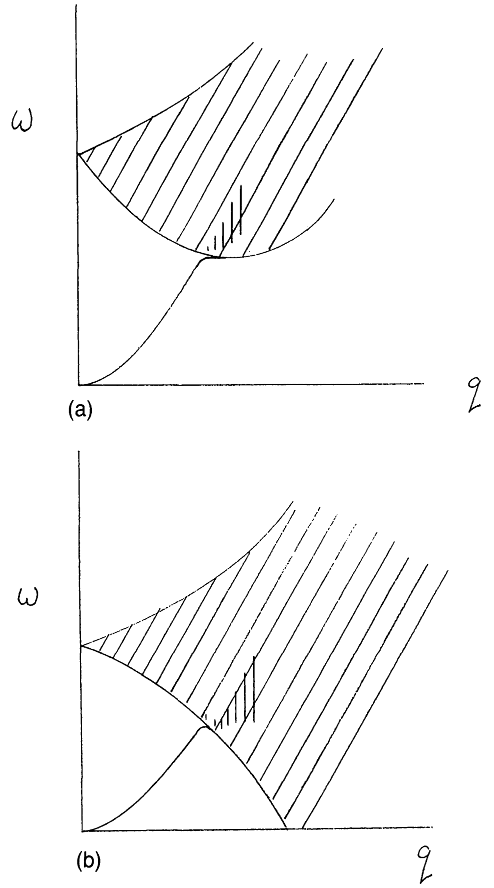

FIG. 9. Excitations in itinerant-electron ferromagnets showing spin-wave modes intersecting Stoner excitations (schematic); vertical hatching corresponds to strongly damped resonances. (a) A "strong" ferromagnet [see Fig. 6(a)]; (b) a "weak" ferromagnet [see Fig. 6(b)].

account see Stanley, 1971, and for an elementary treatment, Goodstein, 1985.)

In all such transitions the theoretical discussions lead to the examination of critical exponents, which express the dependence of some measured property on external variables such as the temperature. An appropriate variable is often $(T - T_c)/T_c = \epsilon$, and critical exponents relate such quantities as the susceptibility, the magnetization (below T_c), and the specific heat to ϵ by expressions such as $\chi \propto \epsilon^{-\gamma}$, where we have seen that $\gamma = 1$ for the simple molecular-field model.

A great deal of effort has been applied to the calculation of critical exponents for models that have specific dimensionalities and types of interaction, but only rarely can measured values of critical exponents provide insights into the fundamental nature of magnetically ordering materials. Comparisons of theory and experiment do exist, however, for a number of model systems (see de Jongh and Miedema, 1974).

7. MAGNETIC ORDER IN INTERMETALLIC COMPOUNDS

Many of the important developments in magnetism in recent years, both in the development of theoretical understanding and in that of new materials of importance in applied magnetism, have been concerned with intermetallic compounds. Such materials make it possible to study in a systematic way the effects of variations in crystal structure and electronic structure on the character and characteristic temperatures of magnetic order, and on the appropriate characteristics (such as remanence and energy product) of materials of potential value as novel permanent-magnet materials.

A particular structure that has been an especially rich source of interesting materials is the cubic Laves-phase structure $MgCu_2$, where the RAl$_2$ alloys in which R is a rare-earth metal have been intensively studied. (Much detailed information on these and other rare-earth compounds is assembled in a valuable review by Buschow, 1980.) In this, as in other series of compounds of the rare-earth metals, the two aspects that play major roles are the crystal-field splitting of the lowest-J multiplet of the R^{3+} ion and the indirect exchange interaction through the conduction electrons, the character of which varies only slightly across the series. A simple analysis of the data is prevented, however, by the fact that in a number of the compounds the crystal-field splittings are of the same order of magnitude as the strength of the indirect exchange.

The same crystal structure is found for a number of RT_2 compounds where T is a transition metal, and while considerations of the sort relevant in the RAl$_2$ compounds are appropriate when T is a 4d or 5d element, a whole new range of considerations enter for those where T is a 3d element. There are then R-R, T-T and R-T interactions to be considered, and many such compounds are ferromagnets with quite high Curie temperatures,

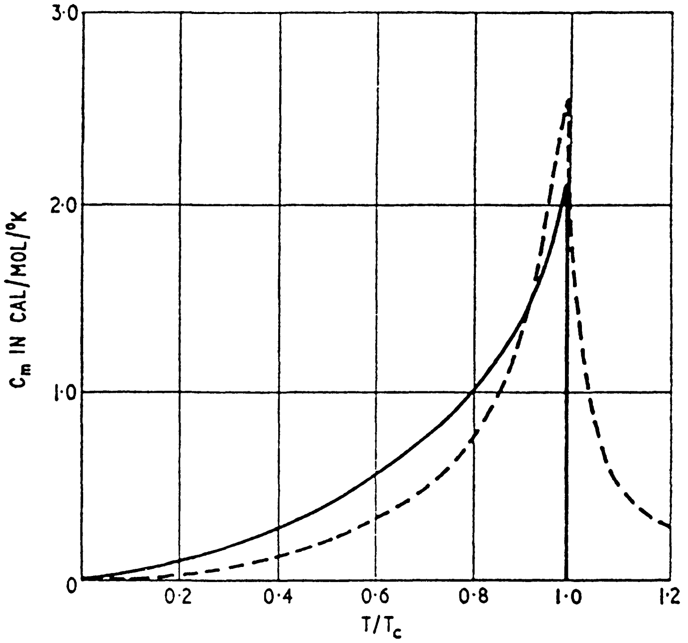

FIG. 10. Magnetic contributions to the specific heat in the molecular-field model (solid line) and in nickel (broken line).

although in the RNi_2 compounds the Ni atoms carry no moment in the presence of the strongly electropositive R atoms.

In the R-3d alloy systems a large number of other intermetallic compounds are also found, ranging from R_3T to R_2T_{17} and RT_{12}, and those with Co have been extensively studied because of the potential as permanent-magnet materials of such compounds as $SmCo_5$, which has a very large coercive force and energy product.

Another group of 4f-3d compounds that has attracted much recent attention because of their even larger coercive forces and energy products is that related to $Nd_2Fe_{14}B$. In some quasi-amorphous or rapidly quenched materials of a range of compositions, coercive fields of around 20 kOe and energy products $(BH)_{\max}$ of more than 10 MG Oe have been reported.

Because of the ability of the 4f configuration of Ce^{3+} to hybridize to a varying extent with conduction electrons, the intermetallic compounds of Ce show a variety of interesting behaviors ranging from superconducting ($CeRu_2$, $CeCo_2$) to antiferromagnetic ($CeAl_2$) and even semiconducting character (CeNiSn). These materials, along with related uranium compounds, have been extensively studied in recent years because of the intense theoretical interest aroused by their very large electronic specific heats which have given them the name "heavy fermions" (see HEAVY-FERMION PHENOMENA). A detailed discussion of these, even of those showing magnetic order, lies outside the scope of this article, and the reader is referred to such review articles as that of Ott (1987). It should be emphasized that such Kondo-effect anomalies arise from the antiparallel effective exchange between the f electrons and the conduction electrons produced by their hybridization, which acts to diminish the local magnetic moments, thereby acting against the onset of magnetic order. However, this moment compensation competes with an indirect interaction favoring magnetic order between f electrons on different sites *through* the conduction electrons, and a delicate balance sometimes exists, which can be pushed one way or the other by "chemical" pressure (i.e., alloying with atoms of different size) or by external pressure.

Alloys of the 3d transition metals with metals other than the rare earths are also a rich source of varying magnetic behavior, much of which was reviewed some years ago by Kouvel (1967). Mention may be made of Ni_3Fe and Ni_3Mn where the effects of the atomic order-disorder transition can be studied, of the equiatomic FeRh alloy, which is ferromagnetic below its Curie temperature of

675 K down to about 350 K where it undergoes a first-order transition to antiferromagnetism, and of the classic Heusler alloy Cu_2MnAl, which was the first example of a ferromagnetic intermetallic compound ($T_c \sim 700$ K, $\mu_{Mn} \sim 4\mu_B$) containing no ferromagnetic element.

8. MAGNETIC ORDER AND SUPERCONDUCTIVITY

Magnetically ordered and superconductive ground states of metals and alloys were long considered to be alternative and mutually incompatible conditions, not surprisingly since for most superconductors applied fields weaker than the molecular fields in ferromagnets will return the material to its normal state. In recent years, however, a number of materials have shown that some types of magnetic order can coexist with superconductivity, and in other materials small changes in composition can replace a magnetic ground state with superconductivity.

Prominent among the latter are the cuprate materials (e.g., La_2CuO_4 and $YBa_2Cu_3O_7$) where changes in the oxygen stoichiometry, in the former by Sr doping, can lead to high superconducting transition temperatures. It has also been known for some time that the temperature of the itinerant-electron antiferromagnet transition of Cr is initially raised to quite high values by alloying with Ru or Re but then collapses at ~20% to be replaced by superconductivity at fairly low temperatures.

Actual coexistence of magnetic order and superconductivity has also been found in two groups of rare-earth compounds, the Chevrel phases (RMo_6S_8 and RMo_6Se_8) and the rare-earth rhodium borides (RRh_4B_4). Some examples of both structures show the coexistence of long-range antiferromagnetic order of $4f$ rare-earth moments and superconductivity associated with the transition-metal clusters. The compound $ErRh_4B_4$ is of particular interest, in that slightly below the range of such coexistence the magnetic order becomes ferromagnetic and the superconductivity is destroyed.

A unique material, Y_9Co_7, shows the presence of bulk superconductivity ($T_c = 2.7$ K) after strong magnetic correlations of ferromagnetic character have shown up at higher temperatures (~5 K), making it the only material where such coexistence involves the magnetism of itinerant $3d$ electrons. It seems likely, however, that the remanence shown by some specimens at ~4.2 K is associated with imperfections in crystallographic order even in samples that are metallurgically single phase.

Recently the heavy-fermion group of Ce and U compounds has yielded examples where neutron scattering shows unambiguously that the small-moment antiferromagnetism setting in on cooling (at 17.5 K for URu_2Si_2) is still present below a much lower superconductive transition (0.9 K for URu_2Si_2). Here both magnetism and superconductivity seem to be associated with the same group of heavy (hybridized f-conduction band) electrons, and this encourages speculation that spin fluctuations rather than phonons may be providing the superconducting coupling mechanism. For a range of Th dopings in the superconducting $U_{1-x}Th_xBe_{13}$ alloys, specific-heat measurements show two transitions, the upper of which seems purely superconductive, but the lower one is found by muon relaxation studies to have some magnetic character.

9. FERROMAGNETIC DOMAINS AND MAGNETIZATION CURVES

In materials that clearly satisfy the conditions for ferromagnetic order, the measured magnetization of a particular sample cannot be explained in terms of the considerations of Sec. 3 and represented by the plot given in Fig. 5. The most striking manifestation of this failure is the possibility of having zero total magnetization associated with a sample of a substance which, by being an ordered ferromagnet well below its Curie temperature, would be expected to have a large spontaneous magnetization.

It has long been recognized that the minimization of the total energy associated with a piece of ferromagnetic material requires the reduction of the energy stored in the magnetic field around the sample, even at the cost of local increases in energy within the sample. This is achieved by the decomposition of the magnetization into domains (Fig. 11) within which the magnetization is given by curves like Fig. 5, but with orientations within

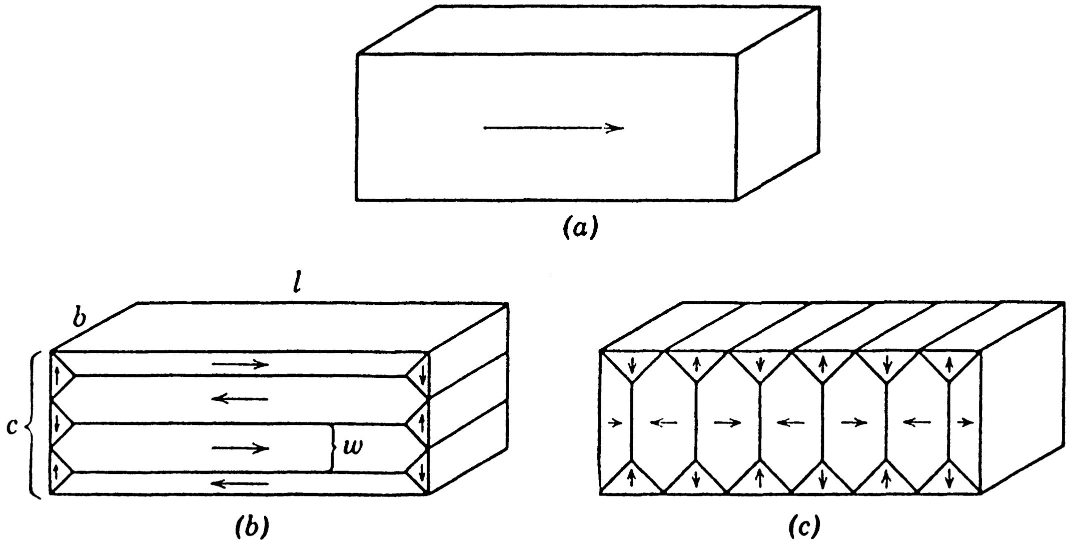

FIG. 11. Possible domain structures that lower the energy from that of the completely magnetized state (a). Configuration (c) has the lower energy.

the domains being such as to reduce the total magnetization to zero. It is clear that a positive contribution to the free energy will be associated with the domain boundaries, where there is a failure to minimize the energy by aligning neighboring atomic moments. This energy increase can be minimized by gradual rotations of atomic moments from one atomic plane to the next (Fig. 12), but the width of the transition region from spin up to spin down will clearly be given by a balance between the saving in exchange energy by broadening this region (the domain wall) and the increase in energy associated with the larger number of moments not lying in the direction favored by magnetocrystalline anisotropy.

The shape and size of the domains that minimize the total energy depend on the crystal structure and the configuration of the specimen, and these can be examined by a number of experimental techniques, such as the Bitter patterns produced by magnetic colloids spread on the surface of the sample or by magneto-optical methods.

The response of a ferromagnetic specimen to applied magnetic fields can be understood in terms of the response of the domain structure to such fields and the possible constraints on boundary movement imposed by imperfections, crystal grain boundaries, etc., within the specimen.

For a small applied field in some arbitrary direction, domains with magnetizations close to the field direction will grow at the expense of other domains (see Fig. 13). For small field increments the boundary movements required will be reversible, but at larger displacements the boundaries are likely to lock at pinning points when the field is reduced. At large fields, where the specimen is effectively one domain, with the magnetization aligned in an easy direction of magnetization, the final approach to saturation consists of the rotation of the magnetization into the direction of the applied field.

When the field is reduced the magnetization will remain finite when the applied field reaches zero at what is called the remanent magnetization, because of the pinning of domain boundaries. A finite field (the coercive force) will be required in the opposite direction to reduce the magnetization to zero. The applications of these principles to the development of materials for specific applications are discussed in MAGNETS AND MAGNETOSTATICS. Traditionally the two classes of magnetic materials of applied significance were "soft" materials, which acquire large magnetizations in small applied fields (i.e., have large permeabilities) and have applications in transformers, generators, and motors, and "hard" magnetic materials, which retain large remanent magnetizations and are suitable for

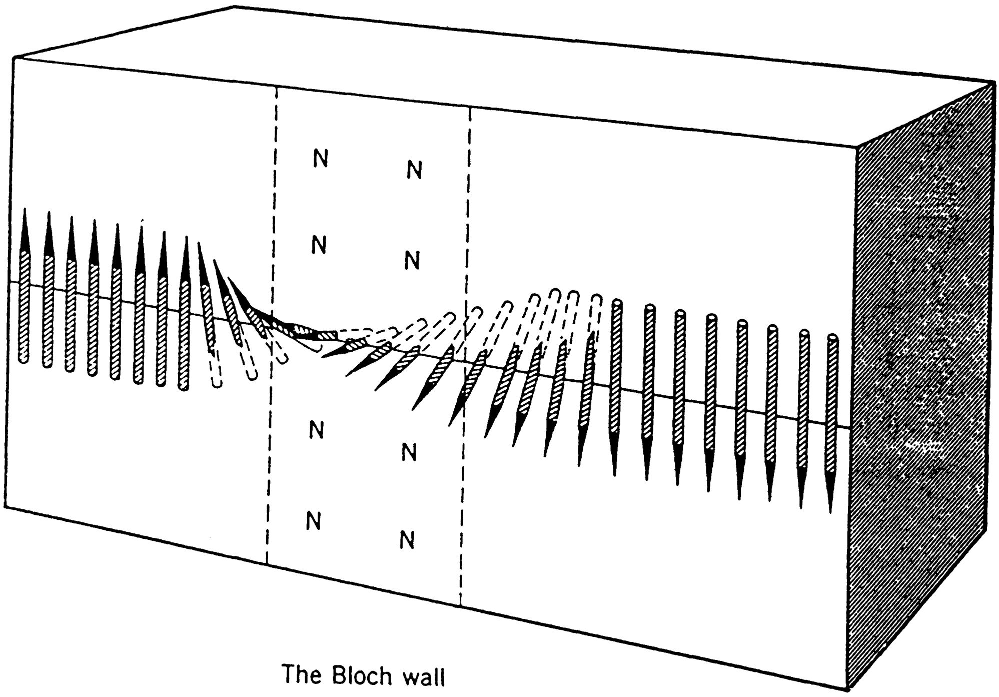
N
N
N
N
N
N
N
N
The Bloch wall

FIG. 12. A domain boundary or Bloch wall.

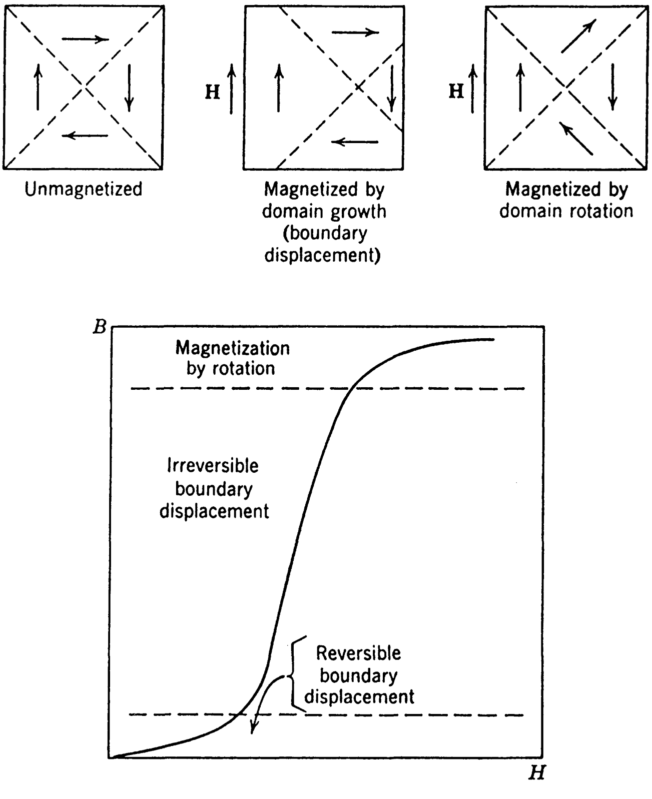
H
H
Unmagnetized
Magnetized by domain growth (boundary displacement)
Magnetized by domain rotation
B
Magnetization by rotation
Irreversible boundary displacement
Reversible boundary displacement
H

FIG. 13. Magnetization curve–domain response relationships (schematic).

permanent magnets. Although both types were originally based on metallic alloys, insulating materials such as ferrites and garnets have been extensively applied both for traditional uses and for more recent applications in magnetic recording media.

Works Cited

Binder, K., Young, A. P. (1986), *Rev. Mod. Phys.* **58**, 801.

Buschow, K. H. J. (1980), in: E. P. Wohlfarth (Ed.), *Handbook of Ferromagnetic Materials*, Vol. 1, Amsterdam: North-Holland, p. 297.

Elliott, R. J. (Ed.) (1972), *Magnetic Properties of Rare Earth Metals*, London: Plenum.

Goodstein, D. L. (1985), *States of Matter*, New York: Dover, Chap. 6.

de Jongh, L., Miedema, A. R. (1974), *Adv. Physics*, **23**, 1.

Kouvel, J. S. (1967), in: J. H. Westbrook (Ed.), *Intermetallic Compounds*, New York: Wiley, p. 529. (A new version is due to appear shortly.)

Mattis, D. C. (1988), *Theory of Magnetism*, Berlin: Springer.

Mydosh, J. A. (1993), *Spin Glasses*, London: Taylor and Francis.

Ott, H. R. (1987), in: D. F. Brewer (Ed.), *Progress in Low Temperature Physics*, Vol. 11, Amsterdam: North-Holland, p. 215.

Stanley, E. R. (1971), *Introduction to Phase Transitions and Critical Phenomena*, Oxford: Oxford University Press.

Wohlfarth, E. P. (Ed.) (1980), *Handbook of Ferromagnetic Materials*, Amsterdam: North-Holland.

Further Reading

Bozorth, R. L. (1978), *Ferromagnetism*, Princeton: Dover.

Evetts, J. (Ed.) (1992), *Magnetic and Superconducting Materials*, Oxford: Pergamon.

Martin, D. H. (1967), *Magnetism in Solids*, London: Iliffe.

Wohlfarth, E. P. (Ed.) (1980), *Handbook of Ferromagnetic Materials*, Vol. 1 and later volumes, Amsterdam: North-Holland.

MAGNETIC RESONANCE AND QUADRUPOLE RESONANCE, NUCLEAR

Jane F. Emerson and P. J. Bray, *Department of Physics, Brown University, Providence, Rhode Island, U.S.A.*

	Introduction	31
1.	**NMR in Solids**	32
1.1	Nuclear Interactions	32
1.1.1	Zeeman Interaction	32
1.1.2	Dipolar Interaction	32
1.1.3	Quadrupolar Interaction	32
1.1.4	Chemical- and Knight-Shift Interactions	33
1.1.4.1	Chemical Shifts	34
1.1.4.2	Knight Shift in Metals	34
1.2	Effects of the rf Field: Absorption of Energy and Relaxation	34
1.3	Detection Methods	35
1.3.1	Continuous-Wave NMR	35
1.3.2	Pulsed NMR	36
1.3.2.1	Determination of Relaxation Times	37
1.3.2.2	Magic-Angle Spinning (MAS)	39
1.4	Applications of NMR in the Study of Solids	39
1.4.1	NMR Studies of Nonmetallic Glasses and Polycrystalline Powders	39
1.4.2	NMR Investigations of Magnetic Materials	41
2.	**NQR in Solids**	41
2.1	Quadrupole Interaction in Zero Magnetic Field	43
2.2	Detection Methods	43
2.2.1	Continuous-Wave NQR	43
2.2.2	Pulsed NQR	44
2.3	Applications of NQR in the Study of Solids	44
	Glossary	44
	Works Cited	44
	Further Reading	45

INTRODUCTION

Nuclear magnetic resonance (NMR) and nuclear quadrupole resonance (NQR) are two valuable techniques for probing structure and dynamics in solids (see Nuclear Magnetic Resonance, Determination of Structures By). Resonance methods, in general, involve inducing transitions between neighboring energy levels available to a system and detecting a signal resulting from the transitions. In NMR and NQR, electromagnetic radiation of the appropriate frequency is used to induce transitions between nuclear energy levels, and the recording of the detected signal versus frequency results in the resonance spectrum.

In NMR, the energy levels available to the nucleus (and thus the shape of the resonance spectrum) depend on the various interactions of a nucleus in an external magnetic field. Because these interactions are dependent on the structural environment of the nucleus, the NMR line shape is a sensitive probe of the nuclear site in a solid. Dynamics (i.e., ionic motion, network vibration, molecular rotation, etc.) can be inferred from the time dependence of these interactions as determined by relaxation-time measurements.

In NQR, no external magnetic field (or a very weak field) is used, and the structural information is gained from the effect on the resonance line shape that results solely from the nuclear quadrupole interaction.

The first part of this article is a general discussion of NMR with emphasis on those aspects that are most important in the study of solids. The basic quantum-mechanical interactions of a nucleus in an external static magnetic field, including the interaction required to produce resonance, along with relaxation mechanisms are discussed in the first two sections; basic detection methods and some applications of NMR to the study of solids are discussed in the last two sections of this part.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

The second part of the article deals with the theory, detection methods, and applications of NQR to the study of solids.

1. NMR IN SOLIDS

1.1 Nuclear Interactions

1.1.1 Zeeman Interaction A nucleus with non-zero spin angular momentum, $\hbar\mathbf{I}$, is a magnetic dipole with magnetic moment, $\boldsymbol{\mu} = \gamma\hbar\mathbf{I}$. $\mathbf{I}$ is the nuclear spin vector, whose magnitude is $\sqrt{I(I+1)}$, where the scalar I is an integer or half-integer and is called the spin of the nucleus; γ is the gyromagnetic ratio of the nucleus, and $\hbar$ is Planck's constant h divided by 2π. When the nucleus is in an external static magnetic field of magnitude H, the otherwise degenerate energy levels of the nucleus are split into $2I + 1$ energy levels; the separation between adjacent levels is

$$\Delta E = \gamma\hbar H. \tag{1}$$

Figure 1(a) depicts the Zeeman energy levels for a nucleus of spin $I = \frac{3}{2}$, where m is the magnetic quantum number ($-I \leq m \leq +I$) and increases from $-I$ in integer steps.

In NMR, inducing transitions between energy levels is effected by introducing a time-dependent magnetic field of frequency ν_0, such that $h\nu_0 = \Delta E$. This gives the Bohr resonance condition

$$\nu_0 = \gamma H/2\pi, \tag{2}$$

where ν_0 is typically in the radio-frequency (rf) region. The resulting absorption spectrum is a single sharp peak at ν_0 as seen in Fig. 1(b).

1.1.2 Dipolar Interaction For a system of spins, there is an energy of interaction due to magnetic dipole coupling. Dipole-dipole interactions represent the effect of neighboring nuclear magnetic moments and their associated small magnetic fields on energy levels of the probe nucleus. The effect on the NMR spectrum depends on the magnitudes of the magnetic moments $\boldsymbol{\mu}_i$, their spatial separations $\mathbf{r}_i$, and the orientation of the $\mathbf{r}_i$ with respect to the direction of the external magnetic field, but not on the value of the magnetic field. Because of the random orientation of the dipoles, the effect on the absorption line shape is an overall broadening [Fig. 1(c)]. The dipolar contribution to the full width at half maximum, δ, can be calculated for a Gaussian line shape using Van Vleck's method of moments (Van Vleck, 1948):

$$\delta = 2.36\sqrt{M_2^I}, \tag{3}$$

where M_2^I is the "second moment" for a nuclear spin in the presence of like spins I and unlike spins S. Van Vleck showed that for polycrystalline and glassy samples the second moment can be written

$$M_2^I = \Delta_{II} + \Delta_{IS}, \tag{4}$$

where

$$\Delta_{II} = \tfrac{3}{5}\gamma^4\hbar^2 I(I+1)\sum_i \frac{1}{r_i^6},$$

$$\Delta_{IS} = \tfrac{4}{15}\gamma_I^2\gamma_S^2\hbar^2 S(S+1)\sum_i \frac{1}{r_i^6}. \tag{5}$$

Equations (4) and (5) can be used to determine interatomic separations in solids from the experimental line shape.

1.1.3 Quadrupolar Interaction Nuclei with spin $I = 1$ or greater possess an electric quadrupole moment, denoted eQ, which pro-

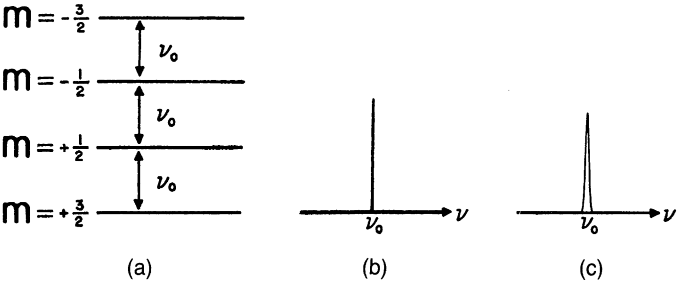

FIG. 1. (a) Energy levels for a spin-$\frac{3}{2}$ nucleus in the presence of a magnetic field. (b) The resultant NMR absorption spectrum is a single peak at the frequency ν_0. (c) Dipolar interactions broaden the resonance symmetrically so that it is not the δ-function response depicted in (b).

vides a measure of the departure from spherical symmetry of the nuclear charge distribution. The quadrupole moment interacts with any electric field gradient (EFG) present at the nuclear site. The energy levels of the nucleus depend on the orientation of the charge distribution with respect to the principal axes of the EFG. Since closed shells of electrons do not generate an EFG, and since the EFG falls off as $1/r^3$, the major contribution arises from bonding electrons. Hence, the quadrupolar interaction is a very sensitive probe of chemical bonding and local atomic arrangements. When the quadrupole interaction is treated as a first-order perturbation to the Zeeman energy levels, the transition frequencies become

$$\nu_{m\to m-1} = \nu_0 - \nu_Q \frac{m - 1/2}{2}(3\cos^2\theta - 1 - \eta \sin^2\theta \cos 2\phi), \qquad (6)$$

where

$$\nu_Q = 3Q_{cc}/2I(2I - 1), \qquad (7)$$

and Q_{cc} is the quadrupole coupling constant:

$$Q_{cc} = eQV_{zz}/h. \qquad (8)$$

The asymmetry parameter η is a measure of the departure of the EFG from cylindrical symmetry and is given by

$$\eta = (V_{xx} - V_{yy})/V_{zz} \qquad (9)$$

where V_{xx}, V_{yy}, and V_{zz} are the diagonal components of the EFG tensor expressed in its principal-axis system. With the labeling of the axes chosen such that $|V_{zz}| \geq |V_{yy}| \geq |V_{xx}|$, one finds that $0 \leq \eta \leq 1$. The variables θ and ϕ are the polar angles of the magnetic field as referred to the principal-axis system of the EFG tensor. Note that quadrupole effects on the NMR spectrum, to first order, are independent of the magnitude of the external magnetic field. In a glass or polycrystalline sample, all orientations of the principal-axis system will exist, and the NMR spectrum will be the envelope of the responses for all values of θ and ϕ. The resulting line shape is a powder pattern. Figure 2 is an example of a powder pattern due to a first-order quadrupole interaction for a spin-$\frac{3}{2}$ nucleus from Eq. (6). (Note that the $m = +\frac{1}{2} \leftrightarrow m = -\frac{1}{2}$ transition is not affected in first order by a quadrupolar interaction.) Locating the positions of the shoulders and divergences in an uncomplicated spectrum experimentally determines the values of Q_{cc} and η, the parameters conventionally measured in NMR spectroscopy.

When the strength of the quadrupole interaction is such that second-order effects must be included, the change in the NMR spectrum for a glass or polycrystalline powder is a broadening that varies inversely with the magnitude of the applied field. Most notably, the central transition ($m = \frac{1}{2} \leftrightarrow m = -\frac{1}{2}$) is broadened by an amount that depends on η and Q_{cc} (Fig. 3).

1.1.4 Chemical- and Knight-Shift Interactions

When a magnetic field is applied to a sample, there are induced changes in the electron orbital motion in atoms and, in the case of metals, in the conduction-electron distribution. These effects give rise to chem-

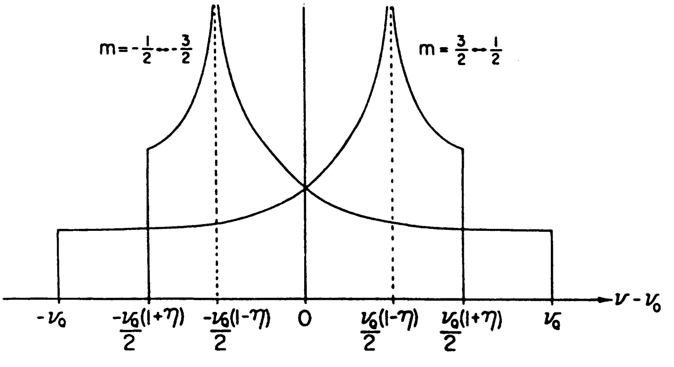
$m = \frac{1}{2} \leftrightarrow -\frac{1}{2}$
$m = -\frac{1}{2} \leftrightarrow -\frac{3}{2}$
$m = \frac{3}{2} \leftrightarrow \frac{1}{2}$
$-\nu_Q$
$-\frac{\nu_Q(1+\eta)}{2}$
$-\frac{\nu_Q(1-\eta)}{2}$
0
$\frac{\nu_Q(1-\eta)}{2}$
$\frac{\nu_Q(1+\eta)}{2}$
ν_Q
$\nu - \nu_0$

FIG. 2. First-order quadrupolar powder pattern for a spin-$\frac{3}{2}$ nucleus. Broadening due to dipolar or other interactions has not been included.

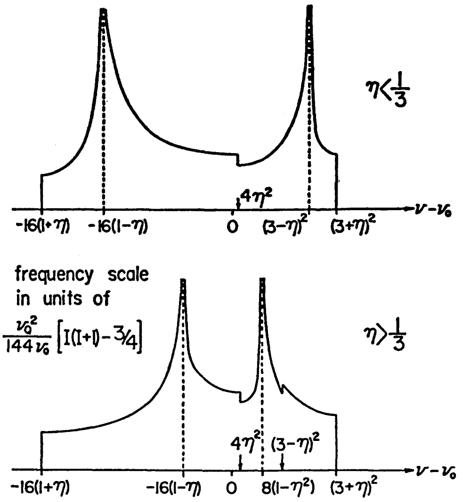

FIG. 3. Second-order quadrupolar powder patterns for the central transition ($m = \frac{1}{2} \leftrightarrow m = -\frac{1}{2}$). Here $\nu_Q = 3Q_{cc}2I(2I-1)$.

ical and Knight shifts in the resonance frequency.

1.1.4.1 Chemical Shifts. The motion of the atomic electrons induced by the applied magnetic field will cause a small shift in the magnetic field experienced by the nucleus. The resulting change in the resonant frequency is termed the chemical shift because it is sensitive to bonding electrons, and is measured in parts per million (ppm). The energy of interaction can be written as the coupling of the spin vector **I** with the magnetic field **H** via the chemical-shift tensor σ:

$$E_{CS} = \gamma\hbar\mathbf{I}\cdot\boldsymbol{\sigma}\cdot\mathbf{H}. \tag{10}$$

Equation (10) is a compact notation wherein the tensor $\boldsymbol{\sigma}$ contains the spatial information derived from the distribution of electronic current densities surrounding the probe nucleus. The components of the chemical-shift tensor, then, are a measure of the symmetry of the nuclear site. The observed spectral frequency is

$$\nu = \nu_0(1 - \sigma_{zz}), \tag{11}$$

where ν_0 is the resonant frequency of the bare nucleus and σ_{zz} is the component of the shift tensor along the direction of the external field. Rewriting this in terms of the diagonal components of the chemical-shift tensor in its principal-axis system, σ_1, σ_2, and σ_3, gives

$$\begin{aligned}\nu = \nu_0[1 &- \sigma_{\text{iso}} - \sigma_{\text{ax}}(3\cos^2\theta - 1)\\ &- \sigma_{\text{aniso}}(\sin^2\theta\cos 2\phi)],\end{aligned} \tag{12}$$

where

$$\begin{aligned}\sigma_{\text{iso}} &= \tfrac{1}{3}(\sigma_1 + \sigma_2 + \sigma_3),\\ \sigma_{\text{ax}} &= \tfrac{1}{6}(2\sigma_3 - \sigma_1 - \sigma_2),\\ \sigma_{\text{aniso}} &= \tfrac{1}{2}(\sigma_2 - \sigma_1).\end{aligned} \tag{13}$$

Because the frequency ν depends on θ and ϕ, the polar angles of the field **H** as referred to the principal axes of the chemical-shift tensor, the spectrum of a powdered crystalline sample or glassy sample will be the envelope of responses for all values θ and ϕ. Representative examples of the resulting powder pattern are shown in Fig. 4.

1.1.4.2 Knight Shift in Metals. The Knight shift (Townes *et al.* 1950; Knight, 1956) in the resonant frequency is also the result of the coupling of electrons with nuclear spins. However, in contrast to the relatively localized bonding electrons in a nonmetal giving rise to the chemical shift in NMR frequency, the conduction electrons in a metal interact collectively with the nuclei, superimposing an internal magnetic field at the nuclear site on the externally applied magnetic field. The magnitude of the Knight shift is typically an order of magnitude larger than chemical shifts, and rather than providing information with regard to the local structure at the nuclear site, it is sensitive to the electronic structure and properties of metals. The components of the Knift-shift tensor $\boldsymbol{\kappa}$ are related to a powder pattern as in the case for chemical shifts described above.

1.2 Effects of the rf Field: Absorption of Energy and Relaxation

In a solid, the collective nuclear spins in their respective sites can be visualized as a spin system coupled to a lattice. (Note that here, use of the term *lattice* is not restricted to the case of regularly arranged atoms as in a crystal, but it also used to denote the more disordered network of atoms in an amorphous solid.) The vector sum of the individual magnetic moments $\boldsymbol{\mu}$ is **M**, the magnetization vector of the spin system. At equilibrium in an external magnetic field, **M**

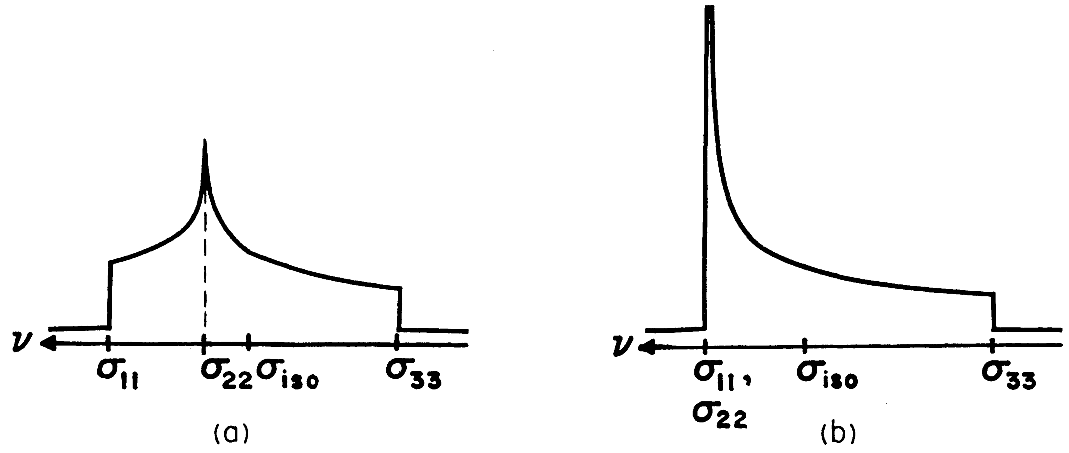

FIG. 4. Chemical-shift powder patterns for (a) a completely anisotropic chemical-shift tensor and (b) an axially symmetric tensor.

is collinear with the axis of the field. Occupation of the Zeeman energy levels (and thus the magnitude of **M**) is determined by the temperature T of the lattice and the strength of the applied field. The various states n with energies E_n will be occupied with the fractional probabilities p_n given by the Boltzmann factor

$$p_n = e^{-E_n/kT}/Z, \tag{14}$$

where Z is the partition function

$$Z = \sum_n e^{-E_n/kT}, \tag{15}$$

and k is Boltzmann's constant.

When rf power is used to excite the spin system, the energy is distributed among the spins via spin–spin interactions characterized by the time constant T_2, the spin–spin (or transverse) relaxation time.

For an external static magnetic field to change the Boltzmann distribution of states, conservation of energy requires that the spin system be coupled to a lattice that can make a transition to an energetically lower state. The mechanisms that couple the spins to the lattice are collectively characterized by T_1, the spin–lattice or longitudinal relaxation time. Thus T_1 determines the time for the magnetization vector to grow in an external field. When rf power is used to excite the spin system, the spin-lattice mechanisms dissipate the energy to the lattice, such that T_1 also characterizes the time for the magnetization vector to return to equilibrium. In solids, T_2 is typically orders of magnitude shorter than T_1; thus for a time on the order of T_2, the spins are in thermal equilibrium with each other and not the lattice. The Boltzmann distribution of states is governed by Eq. (14) provided the temperature T now refers to T_s, the spin temperature (Bloembergen, 1949). As the spin system gives up thermal energy to the lattice, T_s returns towards the lattice temperature T at a rate determined by T_1. For high rf levels and/or long T_1 relaxation times, the spins are unable to transfer energy with sufficient rapidity to the lattice, the occupancies of the energy levels equalize, and the system absorbs no more energy; the system is said to be saturated.

1.3 Detection Methods

The two basic techniques in NMR, continuous-wave (cw) and pulsed spectroscopy, involve exciting the nuclear magnetization with radio-frequency (rf) energy and detecting the subsequent response of the spin system. Each method of excitation has advantages and disadvantages depending on the width of the NMR line and the type of information sought.

1.3.1 Continuous-Wave NMR In cw NMR the nuclear magnetization is irradiated at a constant level, and either the frequency of irradiation or the magnetic field is swept through the region of the resonance condition given in Eq. (2). In the crossed-coil method the transmitting coil, with its axis perpendicular to the external magnetic field, feeds rf into the sample. Recall that a magnetic moment, and hence the magnetization vector for a collection of magnetic moments, will precess about an applied magnetic field with angular frequency $\omega = \gamma H$. When the applied rf is of the same angular velocity as the precessing spins, the angle of inclination is changed, and a small signal will be induced in a receiver coil that is placed such that its axis is perpendicular to both the transmitting coil and the external field. This signal is very small and is usually overwhelmed by random noise. Signal averaging and lock-in detection are the usual means of improving the signal-to-noise (S/N) ratio. One limitation of cw NMR is that the signal is proportional to the rf power used, but rf levels must be low enough to avoid saturating the spin system and thereby distorting the line shape or eliminating the signal. Doping the solid with paramagnetic impurities (which can absorb energy from the spin system through electron interactions) can effectively shorten T_1 and reduce saturation without significantly altering the structure of the sample or the absorption line shape. Signal averaging when used with lock-in detection further increases S/N; however, because of the modulation of the magnetic field there is an upper limit on how fast one can sweep through resonance without distortion. Nonadiabatic superfast passage (NASP) (Segal *et al.*, 1983) has been used to solve this problem. In this method neither lock-in detection nor modulation of the magnetic field is used, so that the external field can be linearly swept through the region of interest in a short time (usually less than 50 ms). The spectrometer output is then fed directly to a signal averager. Higher rf levels can be used than with lock-in detection because the measured line shape is not distorted if the resonance is slightly saturated. This, in combination with the rapid accumulation of signals, allows for better S/N ratios in short time. NASP is limited by the external field sweep range, which is on the order of a few hundred gauss.

Both of these cw techniques are useful for broad lines, which commonly occur in amorphous solids. NASP is preferable unless the line is broader than a few hundred gauss.

1.3.2 Pulsed NMR In pulsed NMR spectroscopy, in contrast to cw NMR in which the frequency or magnetic field is swept through the region of the resonance condition, the entire frequency range of interest is irradiated at once by a short intense rf pulse having a magnetic field magnitude of H_1. For a pulse length of t_p, with angular frequency $\omega = 2\pi\nu = \gamma H_1$ [see Eq. (2)], the sample magnetization will be rotated from its equilibrium position along the z axis by the angle

$$\theta_p = \gamma H_1 t_p, \tag{16}$$

as seen in Fig. 5, and will precess about the z axis with angular frequency $\omega = \gamma H$. The magnetization vector is thus stationary in a frame rotating about the z axis with that frequency. The components of **M** that lie in the xy plane will induce a signal called the free induction in a receiver coil. Immediately following the pulse, the individual spins are aligned, but spin–spin interactions and magnetic field inhomogeneities cause dephasing of the spins with a reduction in the magnitude of **M**. This, along with spin–lattice relaxation mechanisms that return the magnetization vector toward equilibrium, diminishes the free induction. The decaying signal is termed the free-induction decay (FID) and is the one that is accumulated and signal averaged. The Fourier transform (FT) of the real part of the FID gives the NMR absorption spectrum.

The exponential decay of the FID is characterized by a time constant T_2^*, or the apparent T_2. It is inversely related to the line-

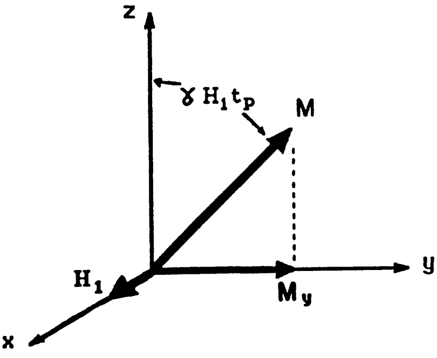

FIG. 5. A pulse of length t_p rotates the magnetization vector from its equilibrium position by an angle $\theta_p = \gamma H_1 t_p$.

width of the absorption spectrum as shown in Fig. 6. In general,

$$\frac{1}{T_2^*} \approx \frac{1}{2T_1} + \frac{1}{T_2} + \gamma\delta H_0. \tag{17}$$

In solids T_1 is usually much longer than T_2; hence, the signal decays before the magnetization has returned to its equilibrium position, and the T_2 term dominates the decay rate. The last term in Eq. (17) represents broadening due to magnetic field inhomogeneities of extent δH_0.

If the magnetization vector has not returned fully to its equilibrium position along the z axis, subsequent pulsing will result in saturation unless the repeated pulse is delayed by a time on the order of several times T_1. In the study of solids, if the structural information obtained from the line shape is the objective, rather than the dynamics inferred from relaxation-time measurements (see Sec. 1.3.2.1 below), the sample is doped with paramagnetic impurities to shorten T_1. Ensuring full relaxation between pulsing is important because nuclei in different structural sites often have different relaxation rates. Even partial saturation will distort the relative spectral intensities due to these sites.

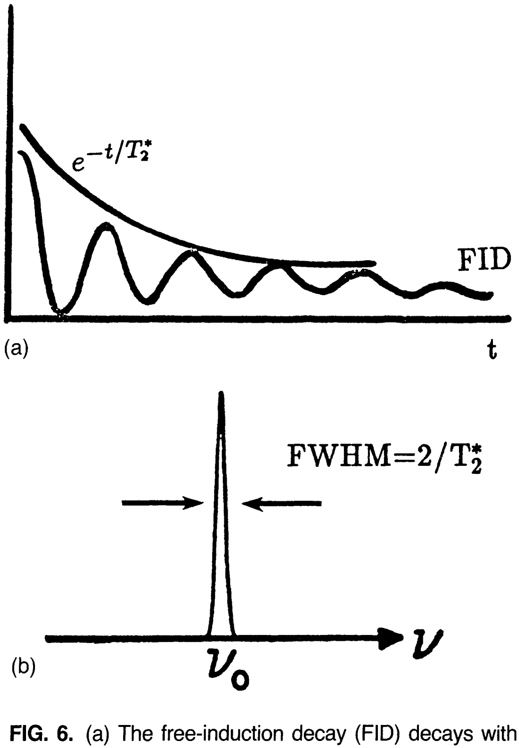

FIG. 6. (a) The free-induction decay (FID) decays with time constant T_2^*. (b) Fourier transform of the FID results in an absorption spectrum with a full width at half maximum (FWHM) inversely proportional to T_2^*. For Lorentzian broadening the FWHM = $2/T_2^*$; for Gaussian broadening the FWHM = $3.34/T_2^*$.

The frequency range of observation is inversely proportional to the pulse length t_p; shorter pulse lengths permit illumination of a broader frequency range. From Eq. (16), a short pulse requires higher rf power to obtain a given tip angle of the magnetization. Pulse lengths are typically on the order of microseconds.

To avoid the rf pulse obscuring part of the FID and to allow for electronic recovery time, data acquisition must be delayed beyond the end of the pulse. For broad lines, when the FID decays rapidly, significant information is lost. Spin-echo techniques can often recreate the early part of the FID, which is obscured by receiver dead-time. In this method (Fig. 7), a 90° pulse is applied at $t = t_0$. [A 90° pulse is defined by Eq. (16) as the pulse length such that $\theta = \pi/2$.] After a certain time interval τ, during which time the spins are dephasing and the FID is decaying, the spin system receives a 180° pulse. This rotates the individual spins 180° around the x axis. The inhomogeneous mechanisms that caused the dephasing of the spins in the xy plane now act to rephase the spins, and at a time 2τ later a spin echo is formed. The echo consists of two FIDs back to back, and the FT of the second half of the echo gives the absorption spectrum. There have been many pulse sequences designed to observe echos in solids, and descriptions can be found in the sources in the Further Reading list.

1.3.2.1 Determination of Relaxation Times. Measuring T_2 is the object of studies in which the aim is to characterize the time dependence of spin–spin interactions in the sample. T_2^* as measured by the FID in a one-pulse NMR experiment differs from the intrinsic spin–spin relaxation time T_2 as seen in Eq. (17). In solids, with $T_1 \gg T_2$, T_2 differs from T_2^* primarily because of inhomogeneities in the applied magnetic field. The Carr–Purcell spin-echo train (Carr and Purcell, 1954) and modifications based on it are designed to eliminate the effects of the magnetic inhomogeneity. In this method, a repeated sequence of 90°–180° pulses as described above results in a train of echos;

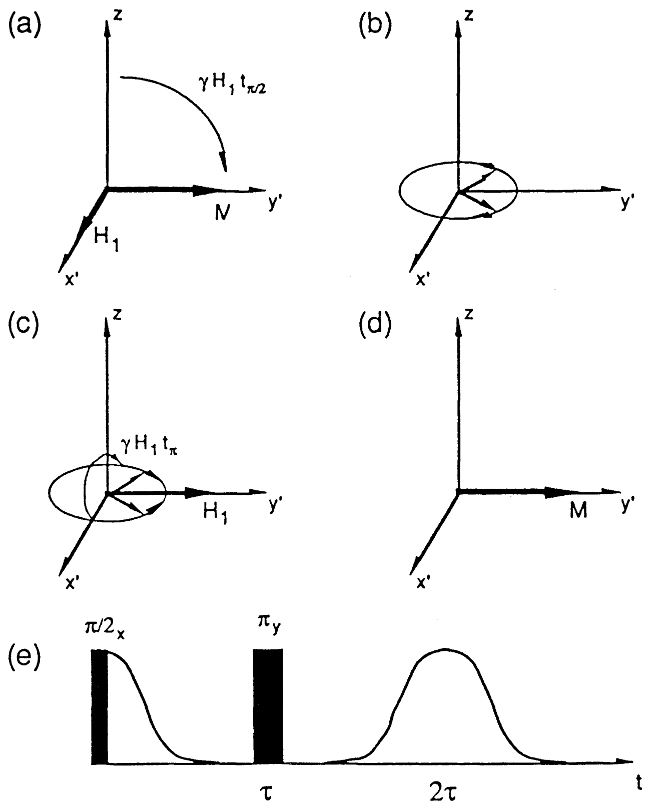

FIG. 7. Formation of the spin echo. (a) At $t = 0$ a 90° pulse is applied, tipping the magnetization vector into the $x'y'$ plane. (b) Because the rate of precession varies for the individual magnetic moments, the spins become dephased and the magnetization decays. (c) At time τ a 180° pulse is applied, thereby flipping all the spins 180° about the y' axis. (d) The individual magnetic moments then become rephased, restoring the magnetization. (e) The spin-echo pulse sequence and the detected signal. The axis along which the magnetic component of the pulse is directed is designated by subscripts; a $\pi/2_x$ pulse indicates that $\mathbf{H}_1$ is directed along the x' axis.

the 180° pulse corrects for dephasing of the spins caused by magnet inhomogeneities, so that the decaying amplitudes of the echo represent dephasing due to intrinsic spin–spin interactions. The decay is characterized by the time constant T_2.

Bloembergen *et al.* (1948) and Torrey (1953) used the temperature dependence of spin–lattice interaction times to determine activation energies and jump times for ionic motion. There are many methods for measuring T_1 values by pulsed NMR; one general category consists of an initial pulse to prepare the spin system followed by a second pulse to monitor the recovery of the spins as a function of waiting time. Two commonly used sequences are known as the saturation-recovery (SR) sequence and the inversion-recovery (IR) sequence. The SR sequences consists of a 90° pulse followed by a second 90° pulse after a time τ. When the interval τ between pulses is greater than T_1, the magnetization vector $\mathbf{M}$ has recovered towards equilibrium such that a second 90° pulse results in a stronger signal. In the IR sequence a 180° pulse is used to prepare (in this case invert) the spins, and a 90° pulse is used to measure signal strength as a function of τ. Another general method of measuring T_1 is to create a "steady state" condition in which the signal achieves a constant amplitude that depends on T_1. One example that is useful for solids is a pulse train of 90° pulses separated by a time τ (Fig. 8).

Slichter and Ailon (1964) and Look and Lowe (1966) showed that slow atomic motions can be studied by experiments analyzed in a reference frame rotating with the Larmor frequency. Measurements of the rotating-frame spin-lattice relations time $T_{1\rho}$, have the advantage that low-frequency motion can be studied while maintaining the higher S/N that goes with stronger externally applied magnetic fields. In these experiments, a 90° pulse is used to tip $\mathbf{M}$ into the xy plane; It is followed by a second rf pulse, now phase shifted such that its magnetic field $\mathbf{H}_1$ lies along $\mathbf{M}$ in the rotating frame. In this frame, the spins experience only the magnetic field of the rf pulse and the

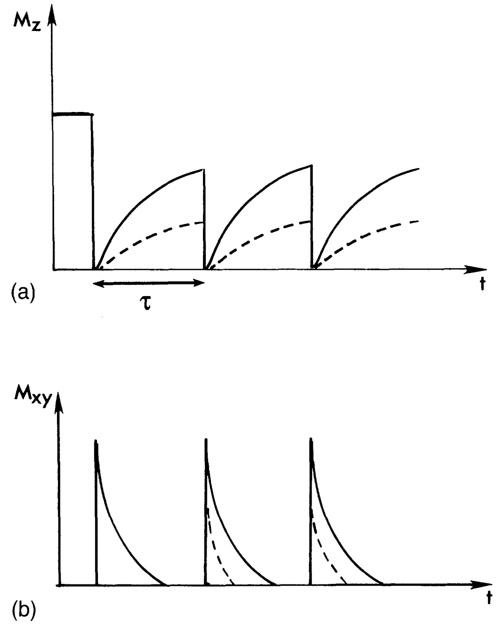

FIG. 8. (a) The z component of the magnetization vector for a sample subjected to 90° pulses separated by an interval τ that is on the order of T_1 (solid line) or shorter than T_1 (dashed line). (b) The xy component of **M** (and the strength of the signal) for the above two cases.

magnetization is said to be spin locked to $\mathbf{H}_1$. Now the spins relax to a smaller equilibrium value of **M** (corresponding to $\mathbf{H}_1$ rather than $\mathbf{H}_0$) with a time constant $T_{1\rho}$. The rf field is turned off after a spin-locking period, and the initial height of the signal as a function of the length of the spin-locking pulse is recorded. $T_{1\rho}$ measurements thus probe motions typified by frequencies on the order of $\nu_1 = \gamma H_1/2\pi$ (usually kHz) rather than $\nu_0 = \gamma H_0 2\pi$ (MHz).

1.3.2.2 Magic-Angle Spinning (MAS). Resolving signals from different nuclear sites requires that the broadening for each line be less than the frequency differences between them. Dipolar, chemical-shift, and first-order quadrupolar interactions that can contribute to the broadening contain terms with an angular dependence of $1 - 3\cos^2\theta$, where θ is the angle between the external field and the internuclear vector (for the dipolar interaction) or is one of the angles characterizing the orientation of the principal axes of the appropriate tensor (chemical shift or EFG) with respect to the field. For $\theta = 54.74°$ these terms are zero, thus narrowing the overall NMR response (Andrew, 1971). In MAS, the sample is placed in a rotor that spins the sample at this angle relative to the static field. Thus any internuclear vector will, on the average, be along the rotation axis and any terms with this angular dependence will average to zero (Fig. 9). To narrow the line effectively, the rate of rotation must exceed the linewidth; otherwise, the transverse magnetization decays faster than the time it takes to complete one revolution.

1.4 Applications of NMR in the Study of Solids

1.4.1 NMR Studies of Nonmetallic Glasses and Polycrystalline Powders Because glasses lack long-range order, only those techniques that probe the short-range atomic and molecular order are useful for determining their structure. NMR has been a valuable technique in the study of both glass structure and dynamics for over 30 years. In glasses (and in polycrystalline powders) the component structural units are randomly oriented with respect to the applied field so that the resonance condition is averaged over all possible orientations. The line shape is then a

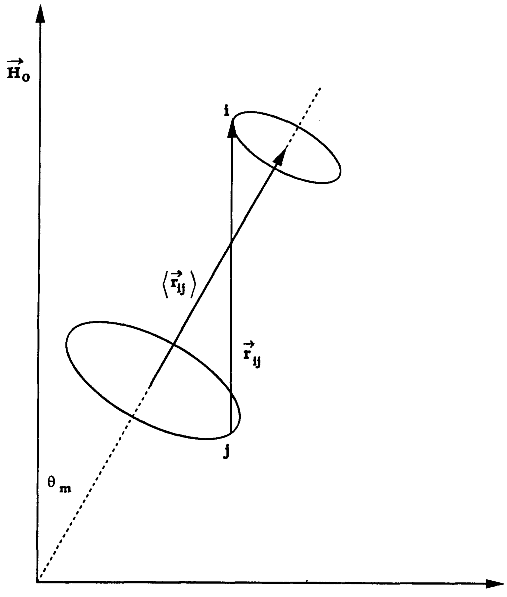

FIG. 9. Sample spinning about an axis at angle θ_m with respect to the external field brings the average positions of the nuclei i and j (and the average internuclear vector, $\langle \mathbf{r}_{ij} \rangle$) to lie along that axis.

composite of the powder patterns due to chemical-shift and quadrupolar interactions. In an uncomplicated spectrum, or one in which the structural sites are well resolved, the positions of the shoulders and divergences in the powder pattern determine chemical-shift tensor values for the different structural units present in the glass and/or the quadrupolar parameters that characterize the nuclear sites. In more complicated cases, structural information can be obtained by computer simulation of the line shape wherein the relative contributions to the spectral intensity of various sites along with their chemical shift or quadrupolar parameters are varied until a best fit to the observed line shape is obtained. Broadening mechanisms are taken into account by convoluting the line shape function with an isotropic broadening function. Most commonly the broadening function is taken to be a Gaussian, which should be exact for the case of dipolar broadening.

Many glass systems have been and are being studied by these methods. An example is the determination of structural sites in silicate glasses. The basic structural unit is a tetrahedron, a silicon atom in fourfold coordination with oxygen, and is denoted Q^n, where n indicates the number of oxygen atoms bridging to other tetrahedral units. The ^{29}Si nucleus is a useful probe for determining structure in these glasses because it is a spin-$\frac{1}{2}$ nucleus and is subject only to chemical-shift and dipolar interactions. Figure 10 schematically displays the various Q^n units along with the ^{29}Si resonance line shape for each. Structural units in multicomponent silicate glasses can often be determined by computer simulation of experimental spectra.

Borate glasses have been extensively studied by NMR. Addition of modifier oxide (e.g.,

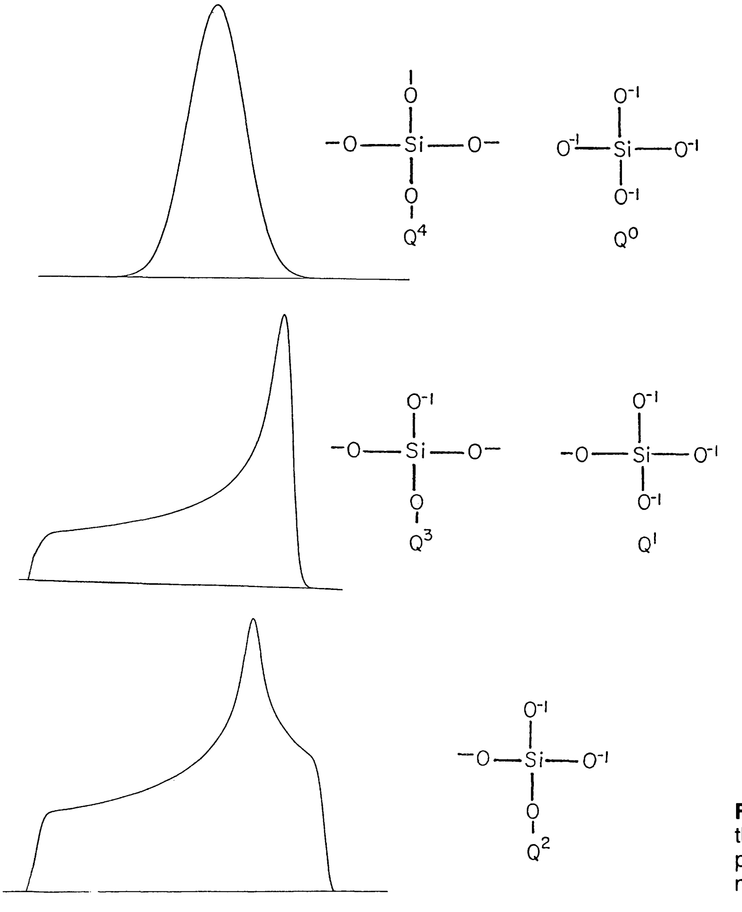

FIG. 10. NMR powder patterns for the Q^n structural units, including dipolar broadening. O^{-1} indicates a non-bridging oxygen (NBO).

alkali or alkaline-earth oxides) to the borate network results in the formation of negatively charged borate units consisting of borons in threefold or fourfold coordination with oxygen. The amounts and types of borate structural units (Fig. 11) vary with the modifier oxide content. Borate glasses are often characterized by the fraction of boron atoms in fourfold-coordination (N_4), the fraction in asymmetric units of threefold-coordination (N_{3A}), and the fraction in symmetric threefold-coordination (N_{3S}). ^{11}B is a spin-$\frac{3}{2}$ nucleus and thus is subject to quadrupolar interactions. The magnitude of the interaction for threefold-coordinated boron atoms is such that it can be treated with second-order perturbation theory at field strengths of 1 T or higher. Figure 12 shows the expected derivative spectrum (i.e., $dP/d\nu$ versus ν where P is the power absorption and ν is the frequency) from a sample containing boron atoms in distinct environments. By matching computer simulations to the observed response, one can extract N_4, N_{3S}, and N_{3A}, along with values of Q_{cc} and η for the latter two cases. The parameters Q_{cc} and η can be obtained for fourfold-coordinated boron atoms in some cases by working at low frequencies, for which second-order effects become apparent, or by observing the satellite transitions (those other than the central transition) arising from the BO_4 units.

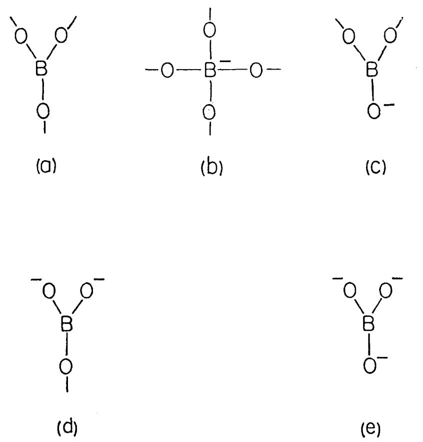

FIG. 11. Structural units found in modified borate glasses. (a) Trigonal $BO_{1.5}$; (b) tetrahedral BO_2^{-1}; (c) asymmetric trigonal BO_2^{-1} with one nonbridging oxygen (NBO); (d) asymmetric trigonal $BO_{2.5}^{-2}$ with two NBOs; (e) symmetric trigonal BO_3^{-3} with three NBOs.

1.4.2 NMR Investigations of Magnetic Materials NMR in magnetic materials was first observed over 30 years ago (Gossard and Portis, 1959) and since then has been used to further the understanding of the correlation between magnetic order and short-range atomic features. Magnetic order of any type in a solid superimposes magnetic fields at the nuclear site. Typically the magnetic hyperfine fields are stronger than the externally applied fields such that resonance for a nucleus with a magnetic moment can be detected even in the absence of the applied field. The shifts of NMR responses by hyperfine interactions are determined by structural symmetries, interatomic distances, type and number of nearest neighbors, magnetocrystalline anisotropy, etc.; thus the shifts provide structural information. Spin-echo techniques are commonly used because the spectra are broad and have very short T_2 relaxation times. Oscillations in the intrinsic magnetic fields of the samples are induced by the rf irradiation, thus increasing the power experienced by the nucleus. The enhancement factor decreases the requirement on the rf power levels compared to NMR of nonmagnetic materials.

2. NQR IN SOLIDS

As in the case of NMR, NQR studies of solids provide both structural and dynamic information. For example, crystallographically nonequivalent nuclear sites can be distinguished by differences in the values of the quadrupole parameters. These parameters are often obtained more readily and accurately using NQR spectroscopy rather than NMR spectroscopy. In pure NQR spectroscopy, the quadrupole interaction does not just give rise to a perturbation of the Zeeman levels; no magnetic field is employed, and so the energy levels are generated entirely by the quadrupole interaction. The NQR resonance is unaffected by chemical-shift anisotropies, and therefore the complicated line shapes found in some cases in NMR are avoided. Because, in NMR, all of the responses from nuclei in sites with different quadrupole interactions cluster around the Larmor frequency

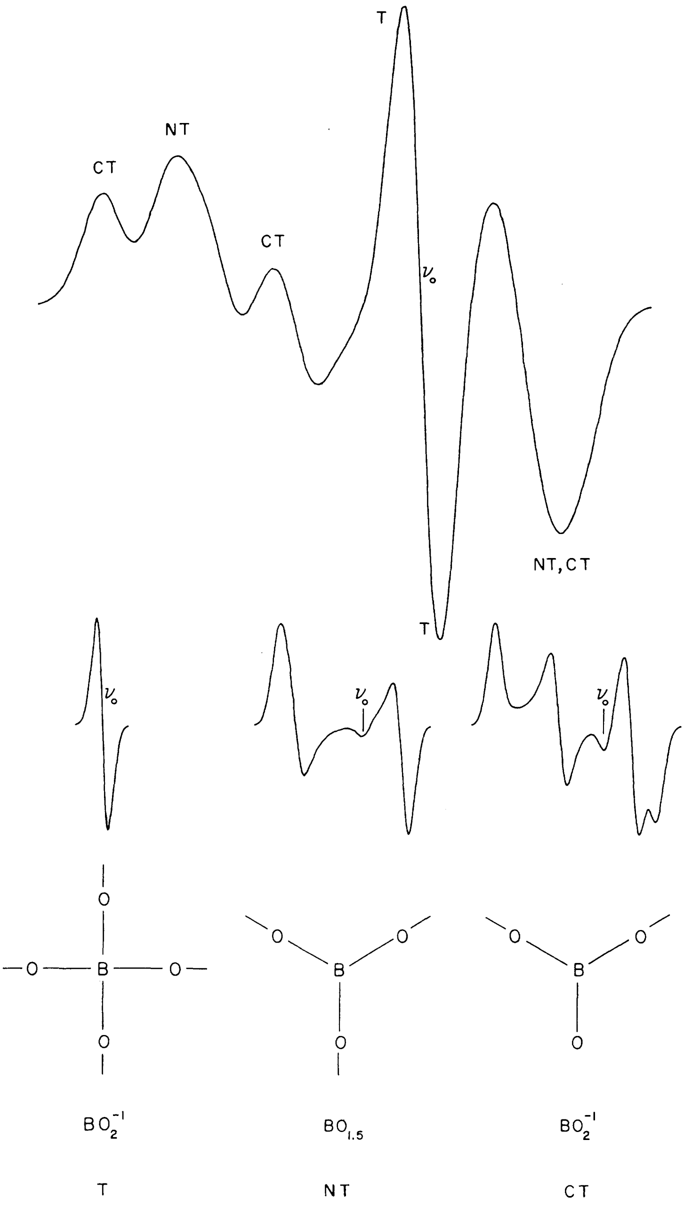

FIG. 12. Derivative spectrum of a dipolar broadened and second-order quadrupolar broadened ^{11}B NMR response. The spectrum is comprised of a weighted sum of spectra due to different sites. The features in the spectrum are labeled T (tetrahedral), NT (neutral trigonal), and CT (charged trigonal), corresponding to the boron units as shown below the spectrum.

ν_0 [see Eq. (2) and Fig. 12], some responses may be unresolved and undetected. This is generally not true in NQR; each site gives rise to one or more resonances at various frequencies determined by Q_{cc} and η. Additionally, for some nuclei such as ^{35}Cl, measurement of the quadrupole parameters by NMR may not be possible because the quadrupole interaction is too large and the resonance is broadened beyond detection.

In addition to measuring Q_{cc} and η, NQR is a valuable technique for obtaining T_1 and T_2. Phase transitions can be detected by their measurement since structural phases are characterized by different relaxation times.

The NQR responses for many nuclei in many solids occur at low frequencies (e.g., less than 5 MHz), a range where resonance methods are relatively insensitive. Since the first successful pure quadrupole resonance experiment was performed (Dehmelt and Krüger, 1950), there have been many variations on the two basic detection methods, cw and pulsed NQR, all aimed at providing a greater sensitivity. These will be summarized after a discussion of the basic theory of pure quadrupole resonance.

2.1 Quadrupole Interaction in Zero Magnetic Field

As noted above, the energy levels are generated entirely by the quadrupole interaction. The magnetic quantum number *m* is used to label the quadrupolar energy levels (although *m* is not strictly a good quantum number unless $\eta = 0$) and allowed transitions are given by $\Delta m = \pm 1$. For spin-1 nuclei the three energy levels are given by

$$E_{\pm} = (hQ_{cc}/4)(1 \pm \eta),$$
$$E_0 = -hQ_{cc}/2, \tag{18}$$

giving rise to three NQR frequencies:

$$\nu_{\pm} = (3Q_{cc}/4)(1 \pm \eta/3),$$
$$\nu_0 = (\eta/2)Q_{cc}. \tag{19}$$

Spin-$\frac{3}{2}$ nuclei have four (two doubly degenerate) available energy levels,

$$E_{\pm 3/2} = (hQ_{cc}/4)\sqrt{1 + \eta^2/3}$$
$$E_{\pm 1/2} = -(hQ_{cc}/4)\sqrt{1 + \eta^2/3}, \tag{20}$$

and only one NQR transition frequency:

$$\nu = (Q_{cc}/2)\sqrt{1 + \eta^2/3}. \tag{21}$$

The transition frequencies for nuclei with other spin values are more complicated and are not reproduced here. In general, for half-integer spins, there are $I + \frac{1}{2}$ energy levels, all doubly degenerate, and there are $I - \frac{1}{2}$ transition frequencies. For a spin-3 nucleus, there are 18 possible transitions. Tabulations of transition frequencies as functions of Q_{cc} and η can be found elsewhere (Butler and Brown, 1981; Creel, 1982).

2.2 Detection Methods

The detection methods that have been developed have various advantages and disadvantages depending on the particular nuclei, the material being studied, and the problem of interest. Historically, pulsed techniques have been widely employed for NQR studies, and are clearly preferable where relaxation times are desired. The quadrupole resonances for some nuclei (^{11}B and ^{27}Al), however, have not been successfully studied by pulsed NQR, and cw methods must be used.

2.2.1 Continuous-Wave NQR There are two categories of cw NQR spectrometers based on the type of generator-detector used. A super-regenerative oscillator was used in the first observation of a pure quadrupole resonance. An oscillator is an *LC* (tank) circuit in which the frequency of oscillation is given by $\omega = 1\ \sqrt{LC}$. A frequency range can be swept with a variable capacitor. A super-regenerative oscillator is one in which the rf amplitude is quenched at a low rate compared to the oscillator frequency by a superimposed wave form. After a quench, while the amplitude of oscillations is building, the oscillator is more sensitive to external signals as its characteristic frequency. In an NQR spectrometer, the super-regenerative oscillator is operated in the coherent mode. That is, the quenching action is incomplete so that the phase of the rf in the coil is maintained. The rf spectrum is then made up of the fundamental frequency plus a number of discrete components that occur at integral multiples of the quenching frequency, all of which can excite nuclear quadrupole transitions. This type of spectrometer is very sensitive but has several disadvantages. The rf output is high, easily saturating many systems. The observed line shape is not predictable, being generally a mixture of absorption and dispersion components, and is complicated by sidebands that result from rf components related to the quenching frequency; thus, the method is not applicable to the detection of closely spaced multiplets or the study of line shapes. To address these problems, a number of modifications to the super-regenerative scheme have been devised, the particular approach depending on the frequency range of interest.

Marginal oscillators were first developed by Pound and Knight (1950) and then by Wang (1955). A significantly improved version was developed by Robinson (1963, 1982), reportedly approaching the theoretical limit in signal-to-noise ratios. In the Robinson spectrometer, the sample is placed in the inductor coil of the *LC* oscillator, and the frequency is swept by tuning the capacitor. When the oscillating frequency is tuned to the NQR frequency, the coupling between the magnetic moments of the nuclei and the rf electromagnetic field induces transitions between the quadrupole energy levels, resulting in the absorption of rf energy by the nuclear spin system. Thus the rf voltage across the coil becomes a function of nuclear absorption. These signals are weak, and lock-in detection is used to improve the S/N. Recent improvements on the Robinson oscillator have resulted in sensitive NQR spectrometers, operable in low-frequency regions, that produce undistorted line shapes. An additional advantage of the marginal oscillator over the super-regenerative oscillator is the low rf output; systems characterized by long relaxation times are not as easily saturated using spectrometers.

2.2.2 Pulsed NQR Pulsed techniques similar to those used in NMR were first developed in 1954 (Bloom *et al.*, 1955) and offer many advantages over cw NQR. There is an overall increase in sensitivity, especially to broad lines, because spin-echo methods can be employed, and the problem of saturation is overcome. Additionally, the line shapes are undistorted and relaxation times can be accurately determined. The chief disadvantage is the unexplained failure of pulsed spectrometers to detect the resonances of certain nuclei such as ^{11}B and ^{27}Al.

2.3 Applications of NQR in the Study of Solids

Because the nuclear quadrupole parameters Q_{cc} and η characterize the nuclear site, experimentally determined NQR transition frequencies provide important information about the local atomic arrangements in a solid. When the probe nucleus is spin-$\frac{3}{2}$ and hence gives rise to only one transition frequency [Eq. (21)] for each site, independent measurement of both Q_{cc} and η is precluded. If however, the nucleus of interest has another isotope of spin 1 or greater than $\frac{3}{2}$, both parameters can be determined precisely. Because the EFG for both isotopes is identical, η has the same value in each case. The ratio of the values for Q_{cc} becomes simply the ratio of the electric quadrupole moments for each nucleus. Therefore, observing the transition frequencies for both isotopes uniquely evaluates Q_{cc} and η. ^{11}B (spin $\frac{3}{2}$) and ^{10}B (spin 3) cw NQR have been used in this manner to study the structure of boron-containing solids.

^{14}N, a spin-1 nucleus, has been used as a probe in pulsed NQR studies of structure and of phase transitions in a large number of nitrogen-containing compounds. The NQR transition frequencies given by Eq. (19) uniquely determine the quadrupole parameters for each crystallographically independent N site. Line shapes, relaxation times, and values of Q_{cc} and η are obtained at different temperatures, thus characterizing structural phases.

GLOSSARY

EFG: Electric field gradient.

FID: Free-induction decay. The net magnetization, precessing freely after an rf pulse, induces a decaying signal in a receiver coil.

Gyromagnetic Ratio (γ): Ratio of the magnetic moment of a particle to its spin moment.

MAS: Magic-angle spinning.

Powder Pattern: The envelope of NMR responses from an averaging of the resonance condition over all possible orientations.

Saturation: Condition in which the spin system is unable to absorb more energy because of insufficient rate of transfer of energy to surroundings. This occurs when the spin system has been irradiated with high rf levels and/or when longitudinal relaxation times are long.

Spin Echo: The reappearance of a NMR signal resulting from rephasing the components of magnetization in the *x,y* plane.

Works Cited

Andrew, E. R. (1971), *NMR Spectrosc.* **8**, 1–8.

Bloembergen, N. (1949), *Physica* **1**, 386–426.

Bloembergen, N., Purcell, E. M., Pound, R. V. (1948), *Phys. Rev.* **73**, 679–712.

Bloom, M., Hahn, E. L., Herzog, B. (1955), *Phys. Rev.* **97**, 1699–1707.

Butler, L. G., Brown, T. L. (1981), *J. Magn. Res.* **42**, 120–131.

Carr, H. Y., Purcell, E. M. (1954), *Phys. Rev.* **94**, 630–638.

Creel, R. B. (1982), *J. Magn. Res.* **50**, 81–85.

Dehmelt, H. G., Krüger, H. (1950), *Naturwissenschaften* **37**, 111–121.

Gossard, A. C., Portis, A. M. (1959), *Phys. Rev. Lett.* **3**, 164–166.

Knight, W. D. (1956), in: F. Seitz, D. Turnbull, (Eds.), *Solid State Physics*, New York: Academic, Vol. 2, pp. 93–136.

Look, D. C., Lowe, I. J. (1966), *J. Chem. Phys.* **44**, 2995–3000.

Pound, R. V., Knight, W. D. (1950), *Rev. Sci. Instrum.* **21**, 219–225.

Robinson, F. N. H. (1963), *Rev. Sci. Instrum.* **34**, 1260.

Robinson, F. N. H. (1982), *J. Phys. E* **15**, 814–823.

Segal, S. L., Creel, R. B., Torgensen, D. R. (1983), *J. Molec. Struc.* **111**, 79–84.

Slichter, C. P., Ailon, D. (1964), *Phys. Rev.* **135**, A1099–A1110.

Torrey, H. C. (1953), *Phys. Rev.* **92**, 962–969.

Townes, C. H., Herring, C., Knight, W. D. (1950), *Phys. Rev.* **77**, 852–853.

Van Vleck, J. H. (1948), *Phys. Rev.* **74**, 1168–1183.

Wang, T. C. (1955), *Phys. Rev.* **99**, 566–577.

Further Reading

Dean, C. (1954), *Phys. Rev.* **96**, 1053–1057. A discussion of super-regenerative circuits.

Goldman, M. (1970), *Spin Temperature and Nuclear Magnetic Resonance in Solids*, London: Oxford University Press.

Lucken, E. A. C. (1969), *Nuclear Quadrupole Coupling Constants*, New York: Academic.

Mehring, M. (1976), *High Resolution NMR Spectroscopy in Solids*, Berlin: Springer. Concise treatment of the principles.

Segal, S. L., Barnes, R. G. (1962), *Catalog of Nuclear Quadrupole Interactions and Resonant Frequencies in Solids*, US-AEC Report No. IS-520, Parts 1 and 2.

Slichter, C. P. (1963), *Principles of Magnetic Resonance*, New York: Harper and Row. The selected bibliography serves as a guide to the subject.

Winter, J. (1971), *Magnetic Resonance in Metals*, Oxford: Clarendon Press. Useful discussion of relaxation in metals.

MAGNETIC RESONANCE IMAGING

J. J. Attard and L. D. Hall, *Herchel Smith Laboratory for Medicinal Chemistry, The University of Cambridge, University Forvie Site, Cambridge, United Kingdom*

Introduction 47

1. **Historical Overview** 48

2. **General Principles** 48

2.1 Influence of Magnetic-Field Gradients 49

2.2 k-Space Representation of MRI and its Optical Analogy 50

2.3 Selective Excitation 51

2.4 Two-Dimensional Imaging and Reconstruction 52

2.4.1 Fourier Imaging 52

2.4.2 Projection Reconstruction 53

2.5 Three-Dimensional Imaging and Reconstruction 54

2.6 Image Contrast and Resolution 54

2.6.1 Tissue-Intrinsic Contrast Mechanisms 54

2.6.2 Tissue-Extrinsic Contrast Mechanisms 55

2.6.3 Resolution 56

3. **Basic Equipment** 57

3.1 Magnets 57

3.2 Gradient System 58

3.2.1 Magnetic Resonance Imaging Gradient Designs 58

3.2.2 Eddy Currents and Gradient Shielding 58

3.2.3 Gradient Drivers 60

3.3 Radio-Frequency Probes 60

3.4 Transmitter and Receiver 60

3.5 Computer Requirements 61

4. **Image Techniques and Applications** 62

4.1 The Spin Echo 62

4.2 Rapid Scan Techniques 63

4.3 *In Vivo* Spectroscopy 64

4.3.1 Localization Techniques for Spectroscopy 65

4.3.2 Spectroscopic Imaging 65

4.3.3 Proton Chemical-Shift Imaging 66

5. **Image Artifacts** 67

6. **Safety** 67

Glossary 68

Works Cited 69

Further Reading 69

INTRODUCTION

The application of nuclear magnetic resonance (NMR) imaging to medicine has had an enormous impact on diagnostic radiology. High–spatial-resolution, diagnostic quality, medical images can be routinely and noninvasively generated at any angle to, and projection of, the human body. Information can be obtained that depends on the complex interaction of many parameters, allowing, for example, tissue discrimination between normal and diseased states. The NMR signal not only provides morphological data, but also *in vivo* data concerning human biochemistry and pathophysiology based upon the principles and interactions of atomic nuclei, rather than the electrons, as is the case for conventional x-ray radiology.

NMR imaging has become a multibillion dollar industry, with the result that expensive, highly automated imaging systems are rapidly pervading the hospitals of the world. In the interests of public relations, the word "nuclear" was dropped from NMR, and the subject is now commonly known in the medical field by the acronym MRI, or magnetic resonance imaging.

Section 1 gives a brief historical overview of the development of MRI. The rest of the article is devoted to detailing the principles of MRI (Sec. 2) via its analogy to optics, and the equipment needed to perform such experiments (Sec. 3); the latter concentrates on apparatus relevant for imaging, and glosses over general aspects of the spectrometer. Section 4 gives a brief overview of some imaging applications, chosen to show the di-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

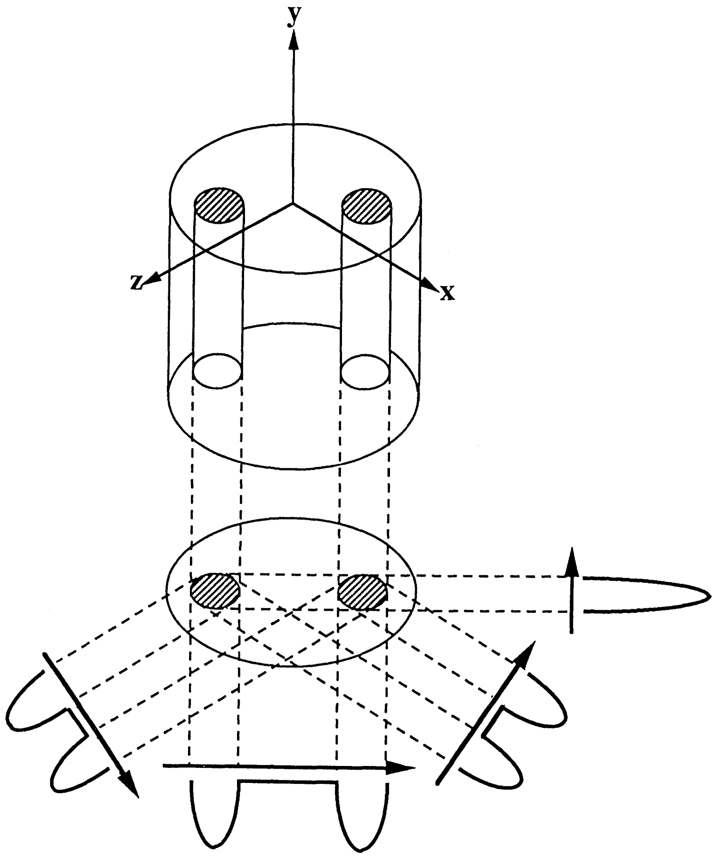

FIG. 1. Schematic of the projection-reconstruction method used by Lauterbur (1973) to create an MR image of two tubes of water. The arrows represent the direction of the magnetic-field gradients used to encode the NMR signal with spatial information, and the various shapes below them represent the one-dimensional projections of the tubes onto the gradients. Projecting the shapes back along the gradients (dotted lines) creates the two-dimensional MR image of the tubes.

component directed along the field, with the direction determined by the allowed quantization levels of the moment. In the presence of a strong magnetic field, a Boltzmann population difference is induced between these levels, with a very small excess of nuclei (about 1 in 10^6–10^9, depending on H_0) aligned parallel to the field. It is this slight excess that is exploited by the magnetic resonance practitioner, making magnetic resonance an intrinsically insensitive technique.

2.1 Influence of Magnetic-Field Gradients

The nuclear environment generates additional magnetic fields that co-add vectorially to $\mathbf{H}_0$ and so influence the precessional frequency of the nuclei, as dictated by Eq. (1). In conventional nuclear magnetic resonance, the spectrum of nuclear precession frequencies produced in this way gives information about the local environment of the spins, and so it is important to ensure that the static field is as homogeneous as possible across the sample. In contrast, imaging is concerned with the application of magnetic-field gradients that are tailored to vary linearly across the sample space, evoking a similar spatial dependence in the Larmor frequency of the spins. The linearly varying magnetic field is applied independently of the static field by means of specially shaped coils. As the magnitude of the field gradient is in practice much less than H_0, ω_0 is only affected by any components parallel to the static field. The local Larmor frequency is then defined as

$$\omega(\mathbf{r}) = \gamma H_0 + \gamma \mathbf{G} \cdot \mathbf{r}, \tag{2}$$

where $\mathbf{G}$ represents the gradient of the applied field component parallel to $\mathbf{H}_0$, and $\mathbf{r}$ is the nuclear spin position vector. In practice, image resolution is not at the nuclear level, and we are dealing with ensembles of spins within finite volume elements centered at positions $\mathbf{r}$.

2.2 k-Space Representation of MRI and its Optical Analogy

In 1973, Mansfield and Grannell applied the concept of a reciprocal-space vector **k** to demonstrate the link between MRI and conventional optics. The **k**-space analysis has since become a way of developing novel imaging schemes by allowing an easier visualization of the scanning modes. A reciprocal vector in **k** space is defined by

$$\mathbf{k} = (2\pi)^{-1}\,\gamma\mathbf{G}t, \tag{3}$$

where t is the duration of the gradient.

As can be seen, both **G** and t are involved in determining **k** space because, unlike x-ray tomography, MR data are acquired in the conjugate (time) domain. The time-domain signal $S(t)$ written in the **k**-space format is given by

$$S(\mathbf{k}) = \int_{-\infty}^{+\infty} d^3r\,\rho(\mathbf{r})\exp(i\gamma H_0 + i\mathbf{k}\cdot\mathbf{r}), \tag{4}$$

where $\rho(\mathbf{r})$ represents the nuclear spin density within the volume element at **r** and i is $\sqrt{-1}$.

Equation (4) is analogous to that of a plane wave scattered from some point P displaced by **r** from the origin O, shown diagrammatically in Fig. 2. The wavy line represents the total accumulated phase γH_0 of the "scattered" wave at the origin. The straight lines represent the additional phase accumulation arising from the "scattering" in the gradient at P. The theoretical wavelength of the fictional MRI wave can be written as $\lambda = 2\pi/k$, and just as in the diffraction limit in optics, it is λ that determines the spatial resolution obtainable in MRI, and not the wavelength of the radio-frequency (rf) radiation at the Larmor frequency. As a result, MRI is able to resolve spatial structures far smaller than would be possible from normal scattering of the radio-frequency electromagnetic wave.

To draw the link between MRI and scattering more clearly (Mansfield and Grannell, 1973), consider a discrete distribution of static spins forming an orthorhombic lattice with unit-cell dimensions a,b,c. The displacement vector $\mathbf{r}_{lmn}(q)$ from the origin to the qth spin site in the (l, m, n)th unit cell is given by

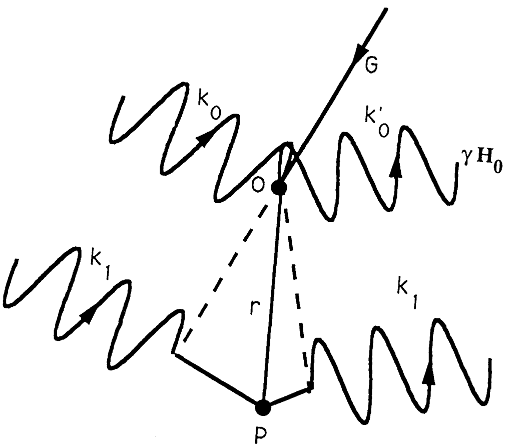

FIG. 2. Schematic of the scattering of the fictitious "NMR wave" from some point P displaced by **r** from the origin O, using an optical analogy to describe the MRI technique. The wavy line represents the total accumulated phase of the scattered wave at the origin ($k'_0\gamma H_0$), while the straight line represents the additional phase accumulation from scattering in the gradient G at P, K'_1.

$$\mathbf{r}_{lmn}(q) = (l + u_q)\mathbf{a} + (m + v_q)\mathbf{b} + (n + w_q)\mathbf{c}, \tag{5}$$

where $\mathbf{a} = a\mathbf{i}$ (**i** is the unit vector), etc. The u_q, etc., are fractions of the primitive-cell dimensions and l,m,n are integers. The NMR signal can now be rewritten as a discrete scattering equation:

$$S = \sum_{l,m,n} a_{lmn}\exp[i2\pi(le + mf + ng)] \times \sum_q f_q \exp[i2\pi(u_qe + v_qf + w_qg)], \tag{6}$$

where $e = ak_x/2\pi$, etc., and $a_{lmn} = 0, 1$. The dimensionless quantities e,f,g are the Miller indices at time t. The term involving the summation over q, which corresponds to signal contributions from within the unit cell, is equivalent to the geometrical structure factor S_k in electron or neutron scattering. The NMR scattering cross section $f_q = 0$ or 1, according to whether resonant spins are absent from, or present within, the cell. This gives plane-wave diffraction from a regular lattice. Hence, the NMR signal $S(\mathbf{k})$ gives rise to a diffraction pattern that, for regular structures, can be used to determine directly the lattice constants, just as in x-ray diffraction studies. More importantly, NMR diffraction measurements carry both phase and amplitude information, which means that the inverse Fourier transformation of the NMR diffraction pattern yields directly the spatial distribution of spins or (more exactly) ensembles of spins.

This conjugacy of signal and spin density is the fundamental relationship of MRI . The NMR signal $S(k_z)$ from a three-dimensional object encoded using a gradient along the z direction can be written in the **k** space formalism as

$$S(k_z) = \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \rho(x,y,z) dx\, dy \right) \times \exp(i2\pi k_z z) dz, \tag{7}$$

where the term in large parentheses is called the spectrum of the sample in the presence of the gradient. The spectrum, $P_{xy}(z)$, can be considered as one-dimensional (1D) projection of the three-dimensional (3D) nuclear spin density onto the direction of the field gradient, and for this reason is referred to as the *projection profile*. All nuclei in a plane perpendicular to the field gradient will experience the same field and contribute to the signal amplitude at the same frequency. The projection can then be written as the Fourier transformation of the signal $S(k_z)$:

$$P_{xy}(z) = \int_{-\infty}^{+\infty} S(k_z) \exp(-i2\pi k_z z) dk_z. \tag{8}$$

As can be inferred from Eq. (3), **k** space may be traversed by moving either in time or in gradient magnitude, and the direction of the traverse is always determined by the direction of the gradient G. The way in which the sampled points are distributed in **k** space is known as a raster; two commonly used rasters are those based on Cartesian and polar coordinates.

2.3 Selective Excitation

Efficient and precise excitation of a selected region of the sample is a vital component of most MRI techniques. This is because the imager or radiographer may only be concerned with a particular "dissected" region of the body and may wish to exclude possible conflicting information from other regions. The excitation principle is described below. The bandwidth of frequencies contained in an rf excitation pulse is inversely proportional to pulse duration, i.e., if a pulse has a duration of 1 ms, then only those spins with a resonant frequency within approximately a 1-kHz bandwidth of the Larmor frequency will be stimulated. In pulsed NMR the whole frequency spectrum is normally irradiated using broadband, or "hard," rf pulses, usually intense rectangular (in the time domain) pulses of short duration. The excitation profile of these pulses, a sinc function, is given by the Fourier transform of the pulse envelope; see Fig. 3(a). In MRI what is required is as close to an excitation of a rectangular band of frequencies as is possible. This means that tailored, narrow-band (or "soft") rf pulses have to be employed. Figures 3(b) and 3(c) show the excitation profiles of two widely used soft pulses, a Gaussian pulse and a double-lobed, truncated sinc pulse.

In order to excite a rectangular frequency profile within a required layer, the soft rf pulse has to be applied in the presence of a field gradient. Figure 3(d) shows a homogeneous spin distribution in the form of a cylinder, the axis of which lies along the z axis, with a linear static gradient G_z existing over the sample. Spins that lie in the plane z are at exact resonance and interact strongly with the rf pulse, while those spins on either side of this plane will be progressively less affected the further they lie away from it. In this way, a spatially defined layer of spins has been selected. As can be inferred from the diagram, adjusting the parameters $\Delta\omega$ and G_z changes the width of the selected slice—a thin slice can be obtained using either strong gradients or a narrow-band pulse. By choosing an appropriate direction for the slice gradient, the orientation of the plane can be chosen at will—this is an advantage of MRI over x-ray CT where slices are necessarily normal to the detector-plane axis (the terms *axial, coronal,* and *sagittal* are used to describe the slice orientation). The use of three slice-selective pulses, in quick succession and in conjunction with appropriately oriented gradients, will select a volume of nuclei within the patient.

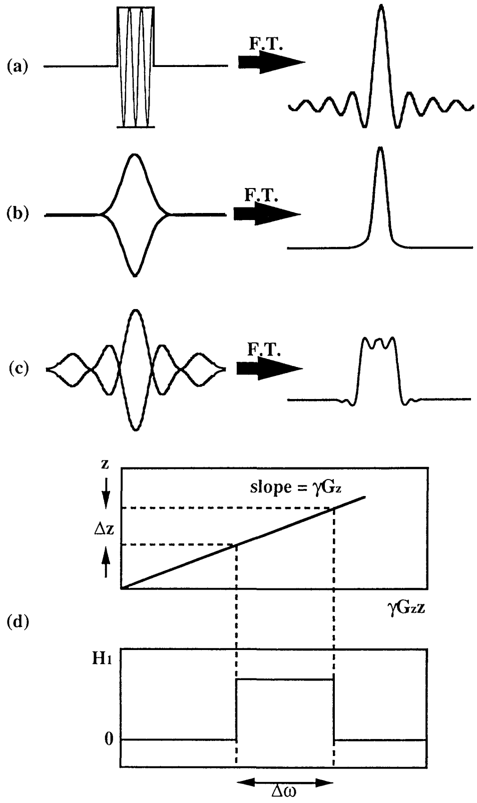

FIG. 3. Schematic of some NMR pulses. (a) A wide-band, "hard," top-hat pulse that excites a profile given by the Fourier transformation (FT) of its envelope, a sinc function. (b), (c) Some commonly used narrow-band, "soft," slice-selective NMR pulses, (b) a Gaussian pulse and (c) a double-lobed, truncated sinc pulse, and their Fourier transformations, a Gaussian and a top-hat excitation profile, respectively. (d) A schematic of the process of slice selection, where the combination of a pulse exciting a range of frequencies $\Delta\omega$ and a magnetic-field gradient of slope γG_z will select a slice, or plane, of thickness Δz in the sample.

2.4 Two-Dimensional Imaging and Reconstruction

If the equilibrium magnetization in a 3D sample is disturbed, for example, by selective excitation, then image reconstruction may be performed in the two dimensions of the slice plane. The resulting image is an average of contributions from across the layer, and the volume element (voxel) will have its depth dimension defined by the slice thickness.

2.4.1 Fourier Imaging Fourier imaging uses the Cartesian raster of Kumar *et al.* (1975), and encodes phase and amplitude information in the NMR signal by means of specifically defined evolution periods. One of the possible timing diagrams and illustration of the method can be seen in Fig. 4. Following slice selection, there is a period of *phase encoding,* so called because the (in this case) G_y gradient imparts a phase modulation to the signal, dependent on the position of the volume elements along the y axis. Phase encoding then corresponds to scanning **k** space along the k_y axis, and since $k_y = (2\pi)^{-1}\gamma G_y t_y$, there exist two extreme cases for accumulating the phase:

1. when G_y is constant and t_y is incremented (Kumar *et al.,* 1975), and
2. when t_y is fixed and G_y is incremented in equal steps (Edelstein *et al.,* 1980).

Following phase encoding, the NMR signal is further spatially encoded in the "read" period, t_x, a process known as *frequency encoding.* As the signal is sampled in the presence of G_x, signal points can be obtained along a single line in **k** space given by $k_x = (2\pi)^{-1}\gamma G_x t_x$.

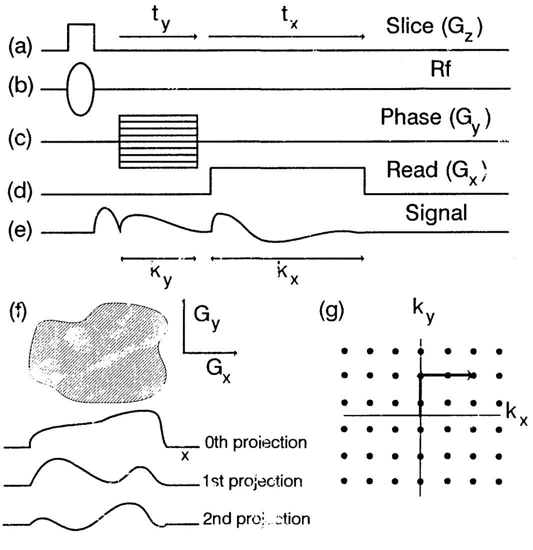

FIG. 4. Timing diagram of one of the possible means of Fourier imaging. After (a)–(b) slice selection, there follows a period of spatial encoding of the NMR signal, using (c) a "phase"-encoding gradient G_y on for a period t_y, and (d) a "frequency"-encoding gradient G_x on for a period t_x. (e) The change in amplitude and phase of the NMR signal due to spatial encoding by the gradients. (f) Projections taken along the x direction for different values of the G_y gradient. (g) An indication of how the **k**-space Cartesian raster is traversed using this technique.

In practice, **k** space is sampled along the x axis under a fixed read gradient, and the process repeated for successively increased values of k_y intercept so that the first quadrant of **k** space is sampled on a raster with the signal sampled along k_x being stored as a row in a matrix. Reversing the y gradient allows the fourth quadrant to be mapped, and since the image density

$$\rho(x,y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} S(k_x,k_y) \times \exp[-i2\pi(k_x x + k_y y)]dk_x\, dk_y \qquad (9)$$

is (ideally) a real function, the signal $S(k_x,k_y) = S(-k_x,-k_y) = S^*(k_x,k_y)$. A 1D Fourier transformation of the signals obtained in the read period from successive experiments yields a set of profiles partially encoded along k_x; see Fig. 4(f). The set is further transformed along k_y by a second 1D Fourier transform to yield the image of the object.

2.4.2 Projection Reconstruction If the NMR signal is acquired while the imaging gradient **G** is applied in some arbitrary direction, a series of signal points are obtained along a radial direction in **k** space. By varying the direction of the gradient, **k** space may be sampled on a polar raster (see Fig. 5 for a timing diagram and illustration of the method). In practice this is done by applying gradients in the x and y directions simultaneously and adjusting their relative values so that the variable polar angle and constant magnitude are, respectively,

$$\phi = \tan^{-1}G_y/G_x \quad \text{and} \quad |\mathbf{G}| = (G_x^2 + G_y^2)^{1/2}. \qquad (10)$$

Reconstruction can be performed in two ways. The first involves initial interpolation on a Cartesian raster, and subsequent Fourier transformation with respect to k_x and k_y. The second, projection reconstruction, is the same as that performed on x-ray CT data, where the interpolation is performed in real space. The image in polar form is given by

$$\rho(x,y) = \int_0^{\pi}\left(\int_{-\infty}^{+\infty} S(k,\phi) \times \exp(-i2\pi\mathbf{k}\cdot r)|k|dk\right)d\phi, \qquad (11)$$

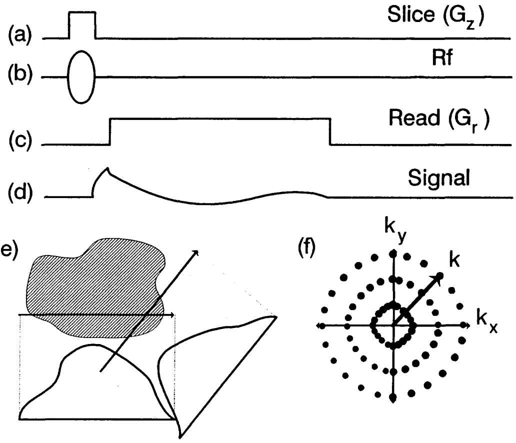

FIG. 5. Timing diagram for the projection-reconstruction technique. After (a)–(b) slice selection there follows (c) a period of spatial encoding of the NMR signal using the gradient G_r, which, in practice, is the vector addition of the gradients G_x and G_y applied simultaneously. (d) The change in amplitude and phase of the NMR signal due to spatial encoding by the gradient, (e) projections taken along different directions in space according to the G_r gradient. (f) An indication of how the **k**-space polar raster is traversed using this technique.

where the inner integral is the filtered profile. In practice, following slice selection, the gradient is swept in angular increments over two quadrants, and at each position the filtered profile is calculated. For each position (x,y) in the image matrix, the projection r along **G** is calculated, and the closest possible value to the filtered profile assigned. This is a process of back-projection and interpolation. The rule for conjugacy of the signal S (Sec. 2.4.1) holds.

In general, Fourier imaging is the preferred technique for two-dimensional MRI. The reasons for this are many and include the following:

1. Fourier imaging samples **k** space more effectively;
2. sampling **k** space in equal steps means that gradient calibration and matching of the driver amplifiers is very important in projection reconstruction (it can be difficult to achieve this in practice);
3. $\mathbf{H}_0$ inhomogeneities manifest themselves as curves in the projected lines making reconstruction difficult (in Fourier imaging they appear as geometric distortion).

2.5 Three-Dimensional Imaging and Reconstruction

The Fourier imaging technique can employ a second phase-encode gradient, with, in this case, magnitude G_z and duration t_z. One method in which it can be employed is to step it in conjunction with the stepped y gradient, i.e., for each step in the z gradient, G_y is allowed to step through its whole range of values, and in this case a full (k_x,k_y) plane is imaged per value of k_z. The image is then reconstructed by a 3D Fourier transformation.

The mathematics of 3D projection reconstruction is essentially the same as for two dimensions except that now a spherical polar, rather than cylindrical polar, coordinate system is used. Three-dimensional projection reconstruction was developed by Lauterbur and his colleagues (Lai and Lauterbur, 1980), and one way of performing this technique is to direct the field gradient $\mathbf{G}_r$ around some path so as to produce a series of 2D projections by reconstruction in the conventional manner. These can be combined using the same algorithm to yield a 3D image.

Both of the above methods employ hard pulses. A third method of generating a three-dimensional image involves sequential imaging of adjacent or interleaved slices. This multislice approach relies on selective excitation ideally disturbing the magnetization in the slice being imaged, so that nuclear spins in other slices remain in thermal equilibrium, available for immediate excitation following the first acquisition. The images can then be separately Fourier transformed and viewed as a series of slices; alternatively, the 2D images may be collated and then interpolated to generate a 3D image. The acquisition of multi(thick)slice 2D images is the preferred method of choice for most clinical imaging; however, it is less appealing (except in the total imaging time sense) as a means of generating 3D images since slice definition is rarely accurate (resulting in out-of-slice excitation and subsequent artifacts when the data are interpolated) and in order to generate slice thicknesses equivalent to 3D pixel resolution, very strong gradients in the slice direction have to be generated [with all the attendant problems (Sec. 3.2)].

2.6 Image Contrast and Resolution

A tissue's "brightness" is directly proportional to the MR signal strength arising from that tissue. In order to distinguish adjacent tissues in an MR image, each must possess different image brightness. This concept is expressed in the term *contrast*, and it is used by the clinician for, for example, differentiating between healthy and diseased tissue. Unlike other imaging modalities, there are a number of parameters that affect image contrast in MRI. These can be divided into two basic groups: parameters intrinsic to the tissue, and extrinsic, operator-dependent factors. In the first category, the principal factors include spin density and characteristic signal-decay time constants T_1 and T_2. The latter category includes the type of imaging sequence, the "repeat" time (T_R), the "echo" time (T_E), and the administration of contrast agents. Manipulation of the extrinsic parameters to maximize the effect of the intrinsic parameters is the great strength of magnetic resonance.

2.6.1 Tissue-Intrinsic Contrast Mechanisms Nuclear spin densities of many tissues in the body only differ by a few percent, and so they do not provide an efficient contrast parameter. More useful is MR relaxation. The nuclei in the body do not exist in isolation, but interact with the fluctuating magnetic fields of neighboring nuclei or the lattice environment. When nuclei in an excited state interact with the fluctuating fields generated by their local molecular environment, they "relax" back to their equilibrium state in a manner determined by the physical and chemical nature of their environments. This is the physical basis of MR, and the strength of these interactions is measured by the "relaxation"-time constants associated with them.

When a sample is placed in a magnet, the population difference is not induced instantaneously. The nuclei nonradiatively transfer some of their magnetic energy to their surroundings, and the population difference rises exponentially and asymptotically towards the Boltzmann equilibrium condition; see Fig. 6(a). The time constant is known as the spin-lattice relaxation time T_1, where the term *lattice* is a general term for the thermal motions of the molecules of which the nuclei are a part. Typical human tissue T_1 values for hydrogen

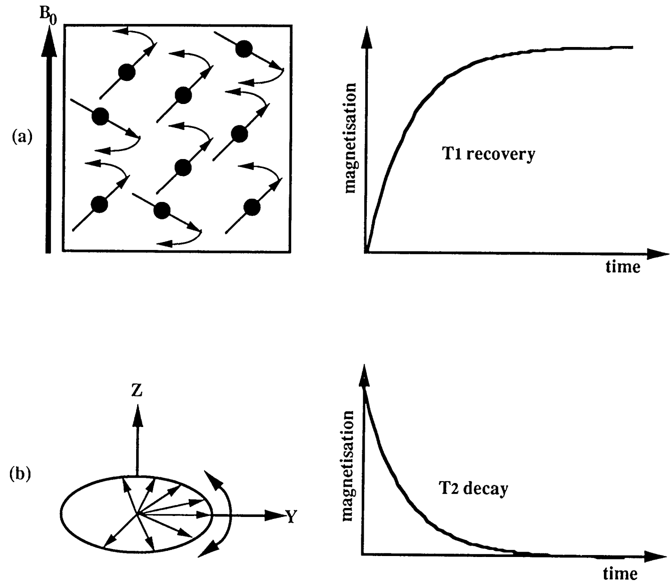

FIG. 6. Schematic of the two most common means of nuclear relaxation. These give rise to contrast changes across MR images. (a) The nuclear spins becoming aligned along the main magnetic field $\mathbf{H}_0$, traditionally taken as the *z* direction. This alignment is not instantaneous, but is exponential and asymptotic with a time constant T_1, the spin–lattice relaxation time. (b) How the nuclear spins, in the transverse *x*-*y* plane, lose phase coherence within a time T_2, the spin–spin relaxation time.

nuclei measured *in vivo* at 0.15 T (they are a function of H_0) are 250 ms for the liver, 600 ms for the spleen, 520 ms for brain grey matter, 380 ms for brain white matter, 700–1200 ms for a brain tumor, and 170 ms for fat.

By applying an rf pulse of the right frequency to the nuclear system, interactions can be induced between the nuclei perturbed to an excited state. At the end of the pulse, the nuclei start precessing in phase; however, they are influenced by the presence of small magnetic interactions that cause the nuclei gradually to lose phase coherence with a time constant known as the spin–spin relaxation time T_2; see Fig. 6(b). In contrast to T_1, T_2 interactions do not involve exchange of energy with the lattice, but between the spins themselves, via flip-flop type interactions. Typical T_2 values for human tissue measured *in vivo* at 0.15 T are 50 ms for the liver, 100 ms for the spleen, 95 ms for brain grey matter, 85 ms for brain white matter, 120–300 ms for a brain tumor, and 80 ms for fat. Note that the slightly higher hydrogen nucleus-density of tumors, and other pathology, tends to oppose contrast due to T_1 weighting and aids contrast due to T_2 weighting.

2.6.2 Tissue-Extrinsic Contrast Mechanisms In order to understand how to manipulate contrast, it is necessary to examine signal strength in a little more detail. Signal strength is proportional to the following:

1. nuclear spin density (ρ);
2. a T_1-weighting term, which is a function of T_1 and T_R;
3. a T_2-weighting term, which is a function of T_2 and T_E; and
4. the usual imaging gradient term.

These all act simultaneously, and the MR signal from within a voxel is more accurately given by

$$\text{Signal} \propto \rho[1 - a\exp(-T_R/T_1)] \times [\exp(-T_E/T_2)]. \tag{12}$$

where a is a pulse-sequence-dependent variable.

The repeat time T_R is concerned with the manipulation of T_1 contrast. After spin excitation using rf pulses, a period of recovery is needed for the spin system to return to equilibrium with its surroundings [$\approx(3\text{–}5)T_1$]. If the next rf pulse (or module of pulses) is applied too soon, the spin system is said to become *saturated* and the numbers of spins aligned parallel and antiparallel to the main field become equal, resulting in no population difference and hence no observable signal. T_1 weighting is increased as T_R is reduced, and in practice this means that certain tissues can have their signals nulled by repeating the experiments so fast that their magnetization has no time to recover and they are excluded from the final image. If T_R is too short there is not enough time for signal recovery in any tissue and, in general, T_R's of 300–800 ms are used.

If magnet inhomogeneities are a dominant source of spin dephasing, the MR signal is said to decay with a time constant T_2^* ($<T_2$). *Shimming* can help reduce the effect, or else a phenomenon known as the *spin echo* may be employed. Spin echoes result from a certain class of experiment that refocuses phase coherence at some time after the initial spin perturbation. One of the simplest experiments is known as the Carr–Purcell (Carr and Purcell, 1954) sequence [Fig. 7(a)]. The first "90°" pulse rotates the magnetization aligned along the z axis (normally taken as the direction of the main field) through 90° about an axis defined by the phase of the rf constituting the pulse, and down on to the x-y plane, where the "transverse" magnetization decays during the period τ because of loss of phase coherence. The second, "180°," pulse flips the relative orientations of the individual spin vectors through 180° [Fig. 7(b)], and the spins rephase during the second τ period, creating a spin echo. Additional rephasings are possible by applying a stream of 180° pulses with intervals 2τ. Note that only loss of phase coherence due to $\mathbf{H}_0$ inhomogeneities can be refocussed in this way. T_E is given by 2τ, and by varying this interval, differential T_2 weighting can be created across an image as a result of the different tissue T_2's. The best T_2 contrast is obtained when T_E is intermediate between the T_2 relaxation times of the tissues, since for very long T_E's, tissue contrast is worsened because all the signals become weak and their differences may be masked by background noise. T_E's vary between 20 and 150 ms in practice, and Fig. 8 shows a simplified representation of image contrast as a function of T_R and T_E.

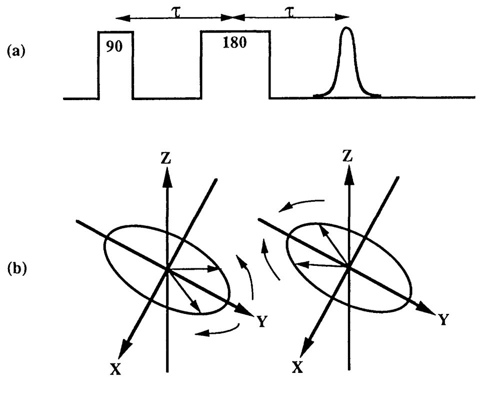

FIG. 7. Schematic of the Carr–Purcell (1954) spin-echo pulse sequence. The 90° rf pulse rotates the nuclei down from the z axis through 90° and onto the x-y plane, about an axis defined by the phase of the rf constituting the pulse. The nuclei are allowed to evolve for a period τ, before the 180° pulse flips the dephasing nuclei through 180°. The nuclei then rephase during the second period τ. Note that only dephasing caused by magnetic field inhomogeneities is rephased in this way.

A common method of enhancing tissue contrast is the use of contrast agents, analogous to the iodine-based ones used in x-ray CT, and most effort has gone into the application of gadolinium diethylene pentaacetic acid (Gd-DPTA), which is still the only one clinically approved. The effect of contrast agents is to reduce the relaxation-time coefficients (more T_1 than T_2), since they are strongly paramagnetic with one or more free electrons, which have a much larger magnetic moment than is associated with such nuclei as hydrogen. The metal atoms of contrast agents are toxic, and they are chelated with very high binding efficiency in order to make them acceptable for human consumption. This means that the number of attractive alternatives for *in vivo* contrast agents (ferromagnetic contrast agents are now being developed) is fairly limited.

2.6.3 Resolution In a conventional imaging system, the resolution is roughly equal to the wavelength of the irradiation used. If this were the case for MRI, then the resolution would be tens or even hundreds of meters. However, as mentioned in Sec. 2.2, MRI resolution is dictated by the strength of the gradient field, and can be written as

$$\Delta x = 2\pi \Delta \nu / \gamma G, \qquad (13)$$

where $\Delta \nu$ is the linewidth associated with the spin interactions taking place within the sample, and is related in some manner to the T_2 (or T_2^*) of the sample. For a typical T_2 of 100 ms and a typical gradient strength of 1 G/cm, the pixel resolution is of order 0.5 mm. In practice, on whole-body systems, in-slice MRI resolutions of 0.25 mm are readily achievable, though with slice thicknesses of 3 mm, a degree of volume averaging within the slice is unavoidable. Equation (13) would seem to indicate that there is no lower limit

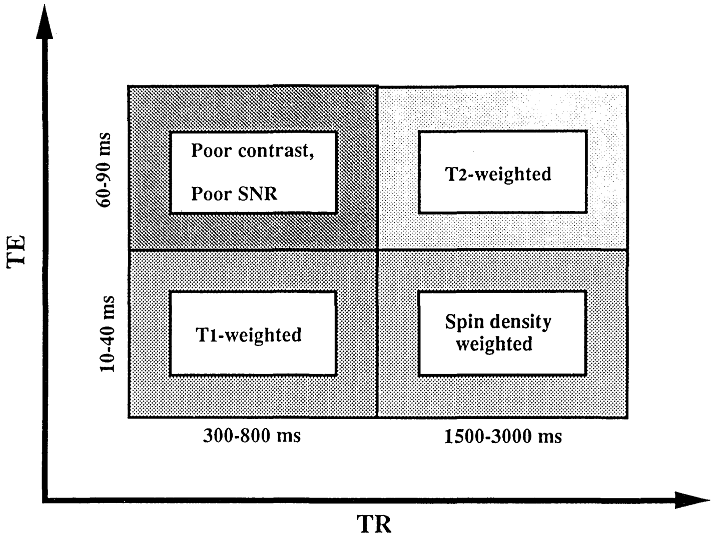

FIG. 8. Representation of how MR image contrast can be manipulated by varying the experimental parameters T_E, the echo time, and T_R, the repeat time. Lengthening T_E increases the T_2 weighting of the images, while shortening T_R increased the T_1 weighting of the images.

to the resolution that can be achieved, provided that enough gradient strength was available. This is, however, not the case. There is a fundamental limit to resolution based on the signal-to-noise ratio (SNR), and the relationship also includes the time taken to acquire the image. The relationship in full is

$$t \approx \left(\frac{S}{N}\right)^2 a^2 \left(\frac{T_1}{T_2}\right) \frac{2.8 \times 10^{-15}}{f^{7/2}} \left(\frac{1}{\Delta x}\right)^6, \qquad (14)$$

where t is the imaging time, a is the rf coil radius, S/N is the desired SNR, and f is the spectrometer frequency (the other terms have their usual meanings). This expression underlines the prime importance of efficient gathering of MR information if imaging times are to be kept as short as possible without sacrifice of SNR or anatomic resolution. The implied dependence of the imaging time on the sixth power of the resolution demonstrates that the sensitivity limit acts as a "brick wall" for the improvement in resolving power by signal averaging (there is a limit to how long a sick patient is able to remain in the magnet!). Even when examining a slice there is a fourth-power dependence, since $(\Delta x)^6$ becomes $(\Delta z)^2(\Delta x)^4$ for a Δz thick slice.

The problem can be alleviated somewhat in terms of the SNR by moving to higher field as $S/N \propto H_0^{3/2}$. However, this is, in turn, offset by the proportionality dependence of T_1 on the applied field, with the result that the recycle delay T_R also increases. Fast imaging techniques are able to produce images (relatively poor resolution) in a few tens or hundreds of milliseconds, but many of them suffer from intrinsically low SNR's. Whichever way the problem is regarded, there is almost always a trade-off involved. Even if clinical MRI resolution were to be brought down to MR microscopy level by some means, other restrictions would come in to play, such as diffusion, which act to decrease the SNR.

3. BASIC EQUIPMENT

In this section we will be concentrating on the apparatus relevant for imaging, and the reader is referred to the encyclopedia entry on MAGNETIC RESONANCE SPECTROMETERS for a more general discussion on NMR spectrometers.

3.1 Magnets

The magnet is the key element in an MRI system. Its role is to provide a high, uniform, and stable polarizing field in which to carry out the scan and a number of approaches to generating these fields have been developed. The advantages of permanent magnets lie in their cheaper cost and much lower stray field. The problems lie in their weight (approaching 100 tonnes, although 5–7-tonne versions are being devised) and poor thermal stability. For these reasons, permanent magnets for

MRI are still relatively uncommon, except in Japan where they have recently taken a significant share of the market. They are built of blocks of magnetic material, e.g., samarium cobalt compounds, ferrites, or more recently, iron–rare-earth–boron materials, and can generate fields of up to about 0.3 T.

By contrast, resistive magnets generate high fringing fields and internal fields up to 0.15 T. These can consume up to 60 kW of electrical power, and that power in the form of heat, has to be dissipated by cooling water. They are usually of the four/six-coil, complex Helmholtz type, consisting of two large central coils of typical size 1.5 m o.d. (outer diameter), and 1 m i.d. (inner diameter). The coils are mounted on a suspension system that allows independent movement of the four coils, thus allowing the magnet to be shimmed (active shims within the magnet may be used). They generate an acceptably homogeneous region over a sufficiently large volume, usually about 1 in 10^5 for 1-m i.d. central coils. The coils are constructed by winding thin bands of aluminum (insulated from one another), of the order of 2 mm thick and 15 cm wide. Aluminum is chosen as a compromise among weight, resistance, and cost, and most designs weigh between 1 and 10 tonnes.

Superconducting magnets represent the majority of magnets deployed in Europe and the United States. Ring magnets, generally made from twisted filaments of niobium-tin alloy in a copper matrix, are enclosed and surrounded by liquid helium to lower the temperature of the windings to 4.2 K. This is then surrounded by a further layer of cooling material, this time liquid nitrogen at 77.4 K, which is used to reduce the boiloff rate of the liquid helium. Superconducting magnets with field strengths of between 0.15 and 4 T have been used for MRI, with room-temperature bores of about 1 m diameter. The magnets are shimmed electrically using an array of shim coils within the magnet, and typical homogeneities are between 10 and 50 ppm (parts per million) over a 45-cm-diam region.

3.2 Gradient System

A gradient system consists essentially of three elements: a control system that defines the required gradient profiles and any compensation that may be required; power supplies to drive sufficient current to achieve the required gradient strength; and the gradients coils that generate the gradient fields.

3.2.1 MRI Gradient Designs There are many varieties of gradient winding designs, the choice being somewhat reduced by the necessity of allowing unrestricted patient access. The geometry of the gradient coils and the direction of the electric current determine the orientation of the gradient, and the design is to produce gradients that vary in a linear fashion, parallel to H_0, i.e., G_z varying as $\partial H_0/\partial z$, etc.

Among the more common designs are those inherited from NMR spectrometers where they were used for magnet shimming. The simplest is the Maxwell pair [Fig. 9(c)], a pair of circular coils, each with the same radius a, but with opposing currents generating a linearly varying field with zero field in the midplane. These are generally used for G_z. The conductor configurations for generating x and y gradients are more complex. The same design can be used for both, one being oriented at 90° to the other. A planar coil configuration was used in early MRI systems, where the linear gradient is generated by four approximately spaced straight wires, all with currents applied in the same direction. The return paths of the coils had to be sufficiently far away so as not to interfere with the main gradient field, and the dimensions of this design were often far too unwieldy for practical whole-body MRI. A third example, Fig. 9(b), is more commonly used in MRI and is known as the saddle configuration. Saddle geometry consists of two saddle coils with four current arcs, two on each end of a cylindrical shaped former, each with associated return paths. It is also possible to modify all three basic coils using a technique known as distributed winding, where the individual wires are distributed over the surface of a cylinder. In this way, field concentrations caused by bunching of the wires, and consequential lowering of the field energy for a given gradient strength, are avoided.

3.2.2 Eddy Currents and Gradient Shielding Two major problems are encountered when gradient fields are rapidly switched on and off. Apart from the large instantaneous voltages present across the coils when the currents through them are quickly altered, there are metal structures in close proximity to the gradient coils. As the gradient field

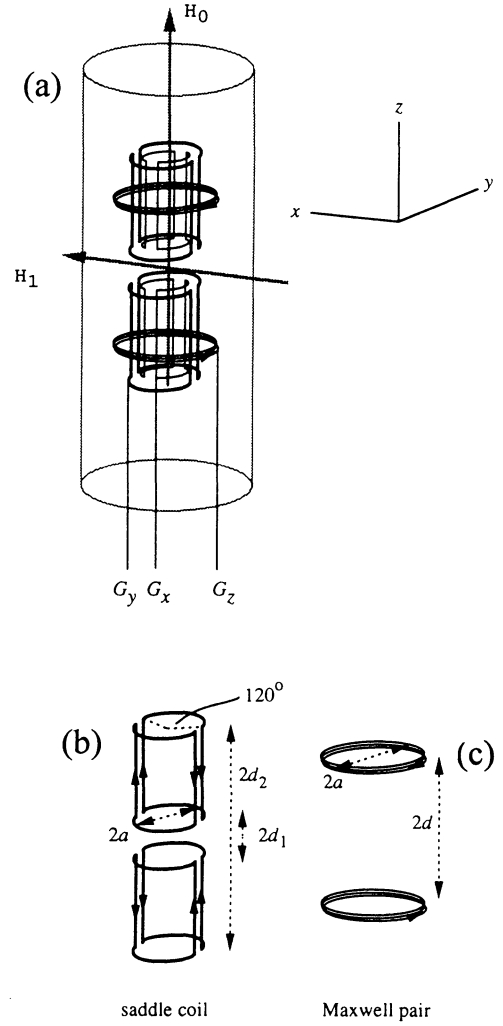

FIG. 9. Schematic of two commonly used magnetic-field gradient coil configurations. (a) The composite gradient set, which breaks down into (b) saddle coil or Golay geometry used for generating both *x* and *y* gradients depending on its orientations within the magnet bore, or the directions of the current flow; (c) Maxwell pair geometry used to generate the *z* gradient. (Reproduced with permission from Callaghan, 1991.)

changes, the changing flux lines passing through the metal surface induce *eddy currents,* which can persist for many milliseconds. Eddy currents cause additional unwanted magnetic fields that cause image artifacts and lengthen the switching time of the gradients. The problem is most severe for superconducting magnets, as a large continuous area of metal in the form of the thermal radiation shield is presented to the gradient fields.

If eddy currents merely changed the value of the gradient following the switch, it would be a simple matter to provide compensation with an appropriate time-dependent cancellation current in the gradient coil. However, many transient harmonics are produced, and each decays with its own decay constant. Several options are available for dealing with this problem. One is deliberately to overide the drive current at the leading and falling edges of the gradient pulse; or the coils themselves may be used to produce fields that compensate for unwanted induced currents. This preemphasis or deemphasis of the coil current relies on Lenz's-law requirement that the sign of the fields associated with eddy currents will be opposed to the change that produces them. The optimization of pre- and deemphasis is a complex process requiring adjustments of a multitude of time constants and amplitudes of exponential currents that are added to the desired wave form, and this compensation can never be perfect. Software correction of images may also be performed but requires exact and detailed knowledge of the eddy-current behavior, which is rarely available. An alternative technique is to employ *reduced-bore* gradients, where the gradients are wound on a former of reduced size as compared to the internal diameter of the magnet, designed to minimize the interaction with the metal surfaces, and hence the amplitude of the transients.

A more effective, and increasingly widely used, method is to ensure eddy-current suppression by using gradient sets whose fields are designed to be zero outside the gradient coil boundaries. *Active shielding* (Mansfield and Chapman, 1987, and references therein) can suppress external fields by up to two orders of magnitude, aided by rapid attenuation of the field with increasing distance from the coil exterior surface. The principle is to add a second layer of current density outside the primary-coil surface, which produces a magnetic field that precisely compensates the primary-coil field at all points in space outside the screen. This involves an entirely new design, as the cancellation currents also produce fields within the sample volume that may act at odds to the field gradients.

3.2.3 Gradient Drivers Whichever gradient coil system is adopted, the design of switched gradient systems is still a difficult problem, involving basic questions of magnet, coil, and power-supply design. The instantaneous power present when a gradient is switched may be many kilowatts, and the energy involved must be dissipated as heat, or stored. Energy-storage techniques involve holding the energy in either a capacitor or an inductor, and such power supplies need sophisticated engineering in order to stabilize the current in the coil once the energy has been transferred with the throw of an electronic switch. The power supplies used to drive each gradient coil can be operated in either constant-voltage or constant-current modes, where the output voltage or current, respectively, follows the drive voltage. Constant-voltage mode is easily implemented, but needs input voltage profiling to overcome coil back emf's. The constant-current mode setup senses the voltage across a resistance in series with the load, and presents this to one input of a difference amplifier that controls the output current. The other input voltage is provided by the voltage from the wave-form generator digital-to-analog converter, so that this voltage wave-form will be faithfully represented as an output current. The maximum current switching speed, di/dt, is limited partly by the power-supply voltage, which must equal $Ri + Ldi/dt$, where L and R are the load inductance and resistance, respectively. The other limitation rate is the bandwidth of the power supply, expressed in the time domain as the *slew rate*, or maximum rate of change of current. In MRI where the large gradient coil sets present a substantial inductive load to the current supplies, the gradient pulses must be turned on slowly using a ramp function. Gradient switching times of 1 ms seem to be standard.

3.3 Radio-Frequency Probes

The rf probe (or antenna) is an essential device, both for transmitting the rf power generated by the transmitter to the sample, and for detecting the nuclear response. Detection of the signal relies on the process of induction, where the precessing nuclei generate a fluctuating magnetic field that induces an emf in a loop of conductor through which the flux passes. The aim is to manufacture a receiving coil in which the maximum emf is generated, and at the same time, when used for transmission, the same coil should generate a homogeneous rf field, $\mathbf{H}_1$.

The most fundamental and important parameters of the rf coil are its quality (Q) factor and its filling factor. The closer the coil is coupled to the patient, the greater is its filling factor (the ratio of the volume occupied by the patient to the total coil volume) and hence the larger the detected emf. Against this must be set the needs of patient access and comfort, and a reasonable homogeneity of $\mathbf{H}_1$ over the region of interest. The Q factor is a measure of how sharply the coil is tuned to the NMR frequency and is the ratio of the reactance (ωL) to the resistance. The higher the Q, the lower the contribution of the coil to the thermal (Johnson) noise on the NMR signal. Unloaded Q factors of several hundred are normal in whole-body imagers. However, the presence of a patient within the coil (human tissue is an ionic conductor at megahertz frequencies) would decrease the Q factor, increasing the coil "resistance" and, hence, the noise. In addition, the rf may not completely penetrate the body, and the so-called *penetration depth* depends on tissue resistivity and the rf frequency. This is not a problem at field strengths of 1.5 T but becomes one at 4 T.

Different coil structures are required for different orientations of the main field. In order to excite the nuclear spins and for efficient flux cutting, the magnetic field lines of the rf coil must be perpendicular to the main field direction. Thus, with a permanent magnet, a solenoidal coil [Fig. 10(a)] gives the best SNR. For the vast majority of superconducting systems no such favorable orientation exists, and other designs are used. The conditions for maximum sensitivity and homogeneous $\mathbf{H}_1$ may be mutually exclusive, and may result in one probe being used for transmission (volume resonators such as slotted-tube, saddle [Fig. 10(b)], and bird cage [Fig. 10(c)] coils) and a second for reception (surface coils [Fig. 10(d)]).

3.4 Transmitter and Receiver

The rf transmitter consists essentially of a frequency source and amplifiers to boost the rf signals to the desired level. RF pulses are formed at low power by gating of a contin-

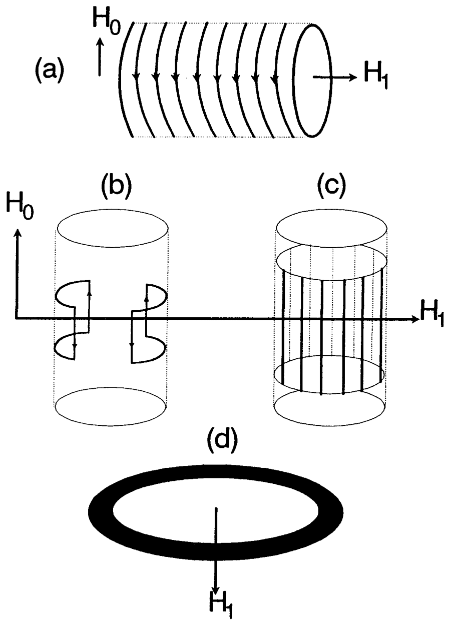

FIG. 10. Schematic of three commonly used radiofrequency coil configurations. (a) A solenoidal coil, which is oriented perpendicular to the main magnetic field, $\mathbf{H}_0$. (b), (c) The "saddle" and "bird-cage" coils, respectively, which are oriented parallel to the main magnetic field. (d) A fourth type of rf coil the surface coil, which, as its name suggests, is placed on the surface of the skin, close to the object which is to be imaged.

uous rf signal on and off with a switch, and are then fed through amplifiers to the sample via the probe. The switch usually has several stages, including sections acting on the amplifier, which then becomes *gated.* This is to ensure that rf does not leak through the switch to the sample between pulses, which would compete with the NMR processes or attenuate the signal by saturating it. Typical on–off ratios are 80–120 dB, equivalent to a voltage attenuation factor of 4 to 6 orders of magnitude.

The receiver consists of amplifier stages to make the MR signals large enough for detection, followed by the detector, which removes the carrier frequency so that only the signal envelope remains. The signal is then further amplified to a size appropriate for digitization. Because NMR is so insensitive, the receiver must amplify and detect the signal without introducing distortion and additional noise. Since the NMR signal in an imaging experiment is spread over a large range of frequencies by the action of the field gradients, a "window" has to be made by using an appropriate filter centered at the carrier frequency, so that only the signal and corresponding part of the noise spectrum will be let through. After amplification, the signal is demodulated into two components in phase quadrature. These are then converted into digital form by analog-to-digital converters (ADC's, normally one for each signal channel), suitable for processing by computer. The sampling rate (typically 10 kHz) is determined by the Nyquist theorem (Black, 1953), to a resolution of between 12 and 14 bits.

The transmitter and receiver are connected to each other, and the rf probe, via a matching network. Because NMR signals are in the microvolt range, whereas the transmitter generates hundreds of volts or more, the receiver has to be decoupled from the network during pulsing, otherwise the receiver electronics become paralyzed for a short time, the "dead time." This can be achieved (as has already been mentioned) by using separate probes for transmission and reception, alternatively using crossed diodes at the receiver input, or using a tuned limiter to the receiver.

3.5 Computer Requirements

The MRI experiment requires a succession of different operations for

1. data acquisition (such as pulse generation, gradient switching, analog-to-digital conversion, etc.), which must be performed according to a strict timing schedule, and
2. data analysis, display, and storage.

All of these require different levels of computer software and hardware, and, in the past, all this was performed by the system's minicomputer. However, it is increasingly the trend nowadays to divide the labor between a dedicated pulse controller for data acquisition and a workstation for data processing, data transfer taking place via some suitable network system. The pulse controller is described in MAGNETIC RESONANCE SPECTROMETERS, and we will confine this discussion to image transfer, display, and storage.

The ability to transfer data quickly and efficiently from the spectrometer is necessary, both for efficient use of the spectrometer and to take advantage of general-purpose computer facilities. The ideal transfer medium in terms of speed and functionality is a network system, and choices include Ethernet, token ring, and fiber-distributed data interface (FDDI) operating at speeds between 4 and 100 Mbps over a variety of cable types. The data then have to be stored, and the demands placed upon archival software and hardware are considerable (for example, image- and time-domain data for a single 256×256×128 3D data set occupies 24 Mwords, or, typically, 96 Mbytes of data). Few standards for archival software as yet exist. The choice for archival hardware is also complicated, consisting of removable magnetic disk, magnetic tape, write once, read many (WORM) optical disks, and rewritable magneto-optical disks. Processing can be divided into five main groups:

1. Reconstruction—creating the image from the time domain data. The major algorithms used are back-projection, and fast Fourier transformation (FFT) and maximum entropy (Press *et al.*, 1988).
2. Processing—including restoration and enhancement. Restoration is used to remove or correct image artifacts; see Sec. 5. Enhancement involves improving the qualitative appearance of the image in order to permit more efficient interpretation by, for example, contrast manipulation, edge enhancement, and voxel interpolation.
3. Analysis—converting image data into quantitative information.
4. Two-dimensional visualization—mapping the pixel's dynamic range (typically 16 bit) to a range of gray levels that can be perceived by an observer (typically 8 bit).
5. Three-dimensional visualization—as above, often with dedicated hardware to achieve interactive speeds. 3D data are often manipulated by interactive selection of a viewing slice extracted from the source data, interpolated, and either displayed separately or "texture mapped" (geometrically distorted) onto the side of a viewing cube or cuboid. The external surface may also be visualized using methods such as surface or volume rendering via ray casting. A "movie" of different viewing directions may be created and replayed to allow the user to construct a 3D impression of the data.

4. IMAGE TECHNIQUES AND APPLICATIONS

The main goal of the radiologist is to seek the best possible image quality in the time available. Ideally, this requires thin, contiguous, nonoverlapping slices with the greatest possible spatial resolution with T_1-weighted, proton-density-weighted, and T_2-weighted images for optimal tissue contrast. At the same time, motion-induced artifacts (see Sec. 5) should be minimized by reducing the imaging time, and patient throughput should be maximized. When specifying a particular MR sequence, the radiologist must decide which "secondary" parameters (contrast, spatial resolution, signal-to-noise ratio, acquisition time, etc.) take precedence for best solving a given clinical question. Determining these specifies the "primary" variables (sequence type, T_R, T_E, slice thickness, number of phase-encode steps, etc.) needed to optimize the secondary parameters. As always, parameter choice requires trade-offs, and consideration must always be taken of the cost involved in changing or selecting a particular parameter.

There is a huge number of possible MR sequences, with the most commonly used sequence on a day-to-day basis being the spin echo. More sophisticated sequences are required for rapid scanning and *in vivo* spectroscopy and spectroscopic imaging, and these are outlined below. Lack of space necessitates leaving out discussions of, for example, flow imaging, which, while not routinely employed, is the subject of much clinical interest. For an in-depth treatise of clinical MRI, the reader is referred to two excellent volumes edited by Stark and Bradley (1992).

4.1 The Spin Echo

The spin-echo phenomenon has already been discussed in Sec. 2.6.2, and reference should also be made to Fig. 8. For spin-echo imaging, tissue contrast is determined by varying T_R and T_E. T_R determines the amount of T_1 or spin-density weighting at short and long values of T_R, respectively. T_1-weighted images (short T_R, short T_E) (see Fig. 11) are best suited for demonstrating anatomic detail, and the shortest possible T_E (usually 10–25 ms) should be used to minimize T_2 effects. In the brain, for instance, spin-density–

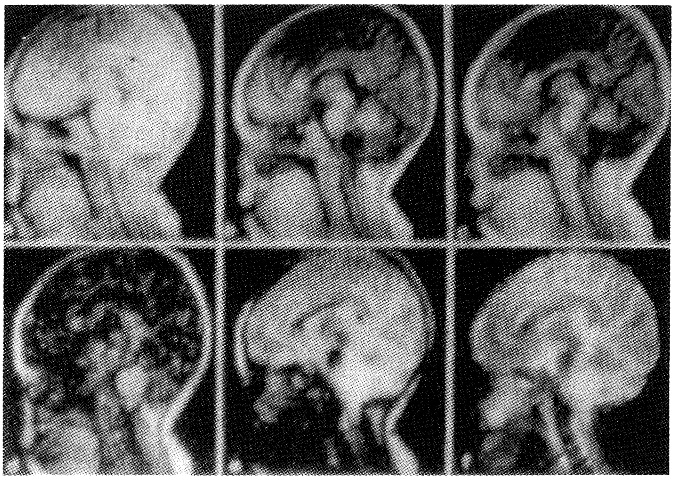

FIG. 11. Proton images of the head and thorax of a 13-year-old boy with a tumor in the medulla (marked by the arrow). (a) A proton density image; (b)–(e) the effect of different T_1 weighting on the image, 500, 300, 200, and 140 ms, respectively [the tumor is most clearly seen in (d)–(e)]. (f) A calculated T_1 image, generated by fitting the T_1-weighted data by a particular function. (Reproduced with permission from Smith, 1987.)

weighted images (long T_R, short T_E) are best suited both to demonstrate lesions and to differentiate between grey and white matter, while T_2-weighted images (long T_R, long T_E) are usually better for demonstrating pathology in the brain or cord. If measuring time were not an issue, the spin-echo Fourier imaging would be the method of choice for most clinical imaging studies. However, spin-echo sequences cannot be accelerated sufficiently, because reducing T_R leads to progressive spin saturation and low SNR, possible significant slice profile deformation ($T_R < T_1$), and considerable rf heating (see Sec. 6) due to fast pulsing of the 180° pulses.

4.2 Rapid Scan Techniques

The original rationale for rapid MRI was for shorter investigational times to improve throughput and patient comfort as well as shorter scan times to reduce motion-related artifacts. With the advent of rapid scanning, however, the possibilities arose for dynamic scanning, cine studies, high-resolution 3D imaging, and flow imaging (angiography). MRI sequences may be classified according to typical measuring times for the acquisition of one cross-sectional image, and are conventional (minutes), fast scan (seconds), and high speed (fractions of seconds).

The bulk of *fast-scan* MRI is based on imaging sequences that combine low flip-angle ($\theta \leq 20°$) rf excitation and gradient-recalled echo (GRE) acquisitions. The low-angle pulse means that only part of the magnetization ($\propto \sin \theta$) is brought down into the x-y plane, leaving an amount ($\propto \cos \theta$) left along the z axis for the next excitation. Consequently, there is no need to wait for full magnetization recovery before the next excitation, and high repetition rates are possible. GRE sequences rely on spin refocussing by a sign change in the frequency-encoding read gradient, eliminating the need for 180° pulses. T_2^*-weighted images generated in this way do not suffer a loss of spatial resolution in comparison to conventional spin-echo images nor suffer from an unacceptable trade-off in SNR. In addition, typical measurement times compare favorably with all time constants of physiologic motions in humans, with the exception of the heart (here data acquisition can be synchronized with the periodicity of the heart motion, a process known as *cardiac gating*). These techniques provide access to functional information on blood flow (angiography) and myocardial movements in cine-MRI movie loops with an apparent temporal resolution of a few milliseconds. An example of a fast-scan imaging sequence is fast low-angle shot (FLASH; Frahm *et al.*, 1986); see Fig. 12.

Scan times of *high-speed* MRI may be as fast as a few milliseconds. In general, such imaging times are achieved not only by technical means but also at the expense of image quality. Spatial resolution may be reduced from the more typical 256 × 256 matrix size to a 64 × 128 matrix size, while SNR penalties are compensated for by significant increases in slice thickness. SNR may be furversity of such applications to which magnetic resonance imaging can be applied and also to give an indication of the multitude of techniques that is available. The last two sections describe common image artifacts (Sec. 5) and the safety aspects of MRI (Sec. 6).

1. HISTORICAL OVERVIEW

The discovery of NMR was made in the United States in 1946, independently by the research groups of Felix Bloch at Stanford (Bloch, 1946) and Edward Purcell at Harvard Universities (Purcell *et al.*, 1946), who were jointly awarded the Nobel prize for physics in 1952 for their discovery. Since then, NMR has had a tremendous impact on chemistry, physics, and other disciplines, and has become an important method for the spectroscopic investigation of matter in many forms. The possibility of applying the newly discovered technique to the study of biological systems was realized early on by Bloch, who obtained a strong proton signal from his finger, but it was only in the early 1970s that the first accounts of the new technique of nuclear MRI were published.

There is still some contention as to who discovered the technique. Gabillard in the early 1950s showed that a magnetic-field gradient can give spatial information about the distribution of an object (Gabillard, 1952), the fundamental principle of nuclear MRI, and in 1959, a Czechoslovakian emigre, Kudravcev, devised a technique for displaying the projection of a quail egg with a surface coil. A patent application submitted by Damadian in 1972 (Damadian, 1974) was the first to show, in principle, the potential scope of MRI, but lacked any implementation. Another patent, submitted in Japan in 1973 by Abe *et al.* (1976), has an implementation similar to that of Kudravcev, and in it they outline many of the techniques used today in MRI.

It was Lauterbur's classic imaging paper of 1973 that established the subject (Lauterbur, 1973). The first practical realization of nuclear MRI using projection–reconstruction [similar to processes used in x-ray computer tomography (CT) scanning] gave an efficient method for generating a two-dimensional representation of proton density and NMR relaxation times; see Fig. 1. In the same year Mansfield and Grannell demonstrated the Fourier relationship between spin density and the NMR signal acquired in the presence of a field gradient (Mansfield and Grannell, 1973).

Industry had, up to that time, played a minimal role in advancing the technique, and it was left to university researchers to develop NMR imaging. This they rapidly proceeded to do, with the concept of slice selection using frequency-selective pulses in conjunction with field gradients coming from Nottingham University, U.K. (Garroway *et al.*, 1974) and the State University of New York, U.S. (Lauterbur *et al.*, 1975), a new efficient method of imaging using two-dimensional Fourier transformation at the Technische Hochschule in Zurich, Switzerland (Kumar *et al.*, 1975), spin-warp imaging by Edelstein *et al.* (1980) at Aberdeen University, U.K.; and fast imaging techniques, such as echo-planar imaging (Mansfield and Pyckett, 1978) at Nottingham University, U.K. Clinical trials of the various techniques began in 1980. More recently, industry (such as Brukermedizintechnik and Siemens of Germany, Philips of Holland, Varian Associates and General Electric of the United States, and Picker International and Oxford Instruments of the United Kingdom) and various universities have collaborated closely on the design and development of MRI hardware and techniques.

2. GENERAL PRINCIPLES

The nuclei of some isotopes contain odd numbers of protons and neutrons and as a result have quantum-mechanical "spin." The combination of positive nuclear charge and spin generates a magnetic dipole moment that interacts with magnetic fields. When a static magnetic field is applied to a nucleus, the nucleus processes about it with a frequency proportional to the magnitude of the applied field, a condition known as Larmor precession and expressed as

$$\omega_0 = \gamma H_0, \tag{1}$$

where ω_0 is the frequency of angular precession, γ is the magnetogyric ratio of the nucleus, and H_0 is the magnitude of the applied field. The precessing magnetic moment has a

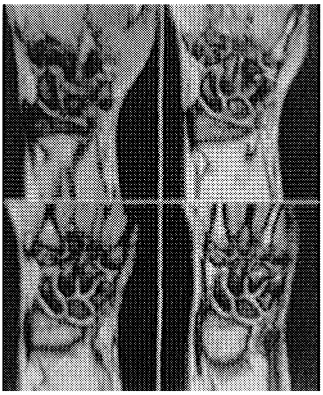

FIG. 12. High-resolution FLASH images of the wrist. These four images (512 × 512) taken at different levels have slice thicknesses of 2 mm, and were obtained with T_R = 150 ms and T_E = 11 ms. (Reproduced with permission from Stark and Bradley, 1992.)

ther compromised by *relaxation convolution*, where successive lines in **k** space may be differently relaxation weighted. T_2-weighted rapid acquisition with relaxation enhancement (RARE; Hennig *et al.*, 1986) and T_1-weighted stimulated echo acquisition mode (STEAM; Frahm *et al.*, 1985) rely on spin rephasing using rf pulses; RARE employs 180° pulses to create multiple spin echoes, with individual echoes phase encoded in a different way by reversing the action of phase encoding in each echo interval [the main application is selective imaging of the cerebrospinal fluid (i.e., myelography) and other body fluids (i.e., urography)]. High-speed STEAM uses two 90° pulses followed by a series of low flip-angle pulses, creating stimulated echoes arising from the action of the low-angle pulses on magnetization stored along the z axis by the action of the second rf pulse. Again, individual echoes are differently phase encoded, and it can be used for cardiac applications and diffusion imaging. T_2^*-weighted echo-planar imaging (EPI; Mansfield, 1977) and high-speed FLASH (saturation weighted) are based on the acquisition of gradient-recalled echoes; EPI employs single rf excitation followed by multiply refocused gradient echoes; high-speed FLASH uses multiple rf excitation and the acquisition of one gradient echo per repetition cycle. Basic EPI employs oscillating gradients along one **k**-space direction in the presence of a static gradient in an orthogonal direction, with the result that **k** space is traversed rapidly in a zigzag fashion. This high-speed imaging technique enables a single cardiac cycle to be imaged at video rates up to approximately 20 frames per second; see Fig. 13. Subsecond FLASH is used for abdominal imaging in which peristaltic motion is effectively frozen. Single cardiac cycle studies of the heart are also possible.

4.3 *In Vivo* Spectroscopy

Magnetic resonance spectroscopy (MRS) is the only noninvasive technique capable of yielding quantitative chemical information from within the body, and the development

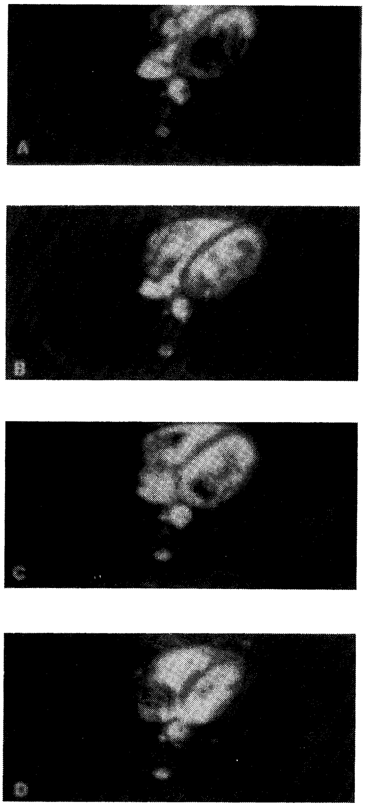

FIG. 13. Four cross-sectional images taken from a 32-frame movie sequence triggered from consecutive heartbeats to span all phases of the cardiac cycle uniformly using MBEST (a variant of EPI). Each image was taken in a single shot lasting 64 ms. The phases of the heart captured are (a) early diastole, (b) mid-diastole, (c) late diastole, and (d) systole. (Reproduced with permission from Ordidge *et al.*, 1989.)

of magnetic resonance spectroscopic imaging (MRSI) provides metabolic information within an imaging format. At a given, highly homogeneous, external field, every chemically distinct nucleus of a given species resonates at a slightly different frequency, resulting in an MR spectrum. This arises from the simultaneous interaction of a nucleus with an electron and of the electron with the static field, such that the electrons effectively shield the nucleus from the static field. Chemical shift allows the spectroscopist to detect a variety of individual protons on a given protein, different phosphorus signals from phosphates on adenosine triphosphate (ATP), or many different carbons of metabolic intermediates: see Fig. 14.

4.3.1 Localization Techniques for Spectroscopy The interpretation of MRS results is greatly facilitated if the volume of tissue giving rise to the signal is controlled so as to correspond to a limited region of interest. For instance, examination of liver is hampered if the spectra also contain signals from overlying muscle. Confining the signal to originate from a desired volume is known as localization, and this requires the use of magnetic-field gradients either in the H_0 field (as in imaging) or in the H_1 field. In the surface-coil, H_1 gradient method, a loop of wire placed close to the region of interest affords rough localization in that metabolites in tissues close to the coil are better detected than metabolites further away, with the degree of localization depending on the size of the coil. This can be improved with the use of *depth pulses*, pioneered by Bendall and Aue (1983), allowing detection of signals from a volume subjected to a particular value of H_1 field strength. Damadian *et al.* (1977) developed FONAR (a static H_0 field gradient method) in which a static magnetic field produces a high-field homogeneity over only a small volume. An analogous approach known as *topical magnetic resonance* (TMR) was developed by Oxford Research Systems (U.K.).

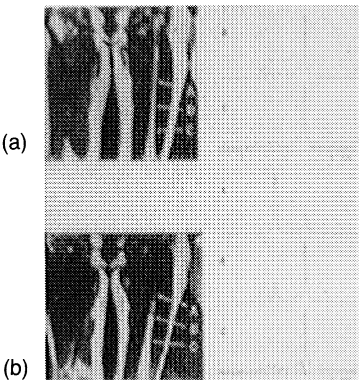

FIG. 14. Proton-stimulated echo images and corresponding spectra from 1-cm^3 voxels in the femoral marrow of (a) a healthy volunteer and (b) a patient with chronic neutropenia, showing differences in cellularity along the length of bone. (Reproduced with permission from Ballon *et al.*, 1991.)

Slice-selective H_0 gradient methods use frequency-selective rf pulses in conjunction with magnetic-field gradients. Depth-resolved surface spectroscopy (DRESS; Bottomley *et al.*, 1984) uses a frequency-selective 90° pulse, with the slice-selective gradient reversed after the pulse to refocus the selected nuclei. Image-selected *in vivo* spectroscopy (ISIS; Ordidge *et al.*, 1986) employs inversion frequency-selective pulses so that a series of acquisitions (adding or subtracting) are needed to achieve localization and remove signal from the surrounding regions. STEAM is the most popular approach to proton spectroscopy, using three 90° selective rf pulses, each applied in succession in the presence of an orthogonal gradient to generate a stimulated echo from the volume defined by the intersection of the three slices.

4.3.2 Spectroscopic Imaging In MRSI, spatial information is obtained by (1, 2, or 3D) phase encoding, and the data are processed to provide a visual image that reflects the distribution of a particular nucleus (such as phosphorus-31) or metabolic species (such as phosphocreatine). To preserve the chemical-shift information, the data are collected in the absence of a frequency-encoding gradient, in contrast to MRI.

Because the concentration of metabolites in tissue is typically more than three orders of magnitude less than that of the protons in the tissue water, high-resolution imaging of metabolite tissues with resolution comparable to MRI is not feasible at the current level of MRS technology. In addition, other nuclei of interest have smaller magnetogyric ratios and hence are less sensitive than protons, an exception being sodium-23, which exists in the body at relatively high concentrations so that low-resolution sodium images have been generated; see Fig. 15. Because of the difficulty, MRS voxel volumes are typically several milliliters or more (MRI may have voxel volumes of 5 μL or less), and the technique is characterized by highest spectral resolution but limited spatial resolution.

4.3.3 Proton Chemical-Shift Imaging Proton chemical-shift imaging (CSI) is primarily used for the separation of fat and water signals and has the high spatial resolution of MRI with good spectral resolution (only two spectral components need to be distinguished). The clinical potential and physical principles of these techniques were demonstrated early in the development of MRI, but sequences that could take advantage of the improvements in magnet, rf coils, and pulse technology have only recently been developed. The frequency (chemical) shift between the water and fat resonances can appear in an image as a spatial mismapping of the two components. If the MR system is tuned to the water frequency, fat will be misplaced in the image relative to its true spatial position in the body (and vice versa). The magnitude of spatial shift depends on H_0 and the strength of the magnetic-field gradients, the bandwidth of the slice-selective pulses, the data-sampling bandwidth, the voxel dimensions, and the chemical shift itself. These chemical-shift misregistration artifacts (CSMA's) are easily visualized in the frequency-encoding direction, but may also occur in the slice-select and phase-encoding directions.

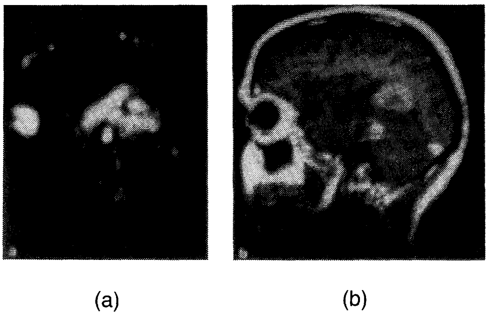
(a) (b)

FIG. 15. Comparison of sodium and proton images of a brain tumor. (a) Sodium image. Note the loss in resolution due to the low concentration of sodium ions and the smaller magnetogyric ratio. There is significant signal intensity in the region of the eye, due to the presence of sodium ions in the humors. (b) A T_1-weighted image enhanced with Gd-DTPA, clearly showing the tumor. (Reproduced with permission from Schuierer *et al.*, 1991.)

Discrimination can be obtained by exploiting the T_1 differences between water and fat by choosing a value of τ that would effectively null the fat signal, short T_1 inversion recovery (STIR; Bydder and Young, 1985), or by using paramagnetic contrast agents to suppress the fat signal, paramagnetic enhancement accentuation by chemical-shift imaging (PEACH, Simon and Szurnowski, 1989); see Fig. 16. Usually, true chemical-shift methods are employed; spectroscopy-based techniques have already been outlined in Secs. 4.3.2 and 4.3.3; phase-contrast techniques, such as the Dixon method (Dixon, 1984), combine "in-phase" (water-fat phase difference = 0°) and "out-of-phase" (water-fat phase difference = 180°) images to generate water-only or fat-only images; frequency-selective pulse techniques, such as CHESS (chemical-shift selective; Haase *et al.*, 1985), are based on narrow-bandwidth, frequency-selective pulses (in the absence of any gradient) to either excite or saturate a specific chemical

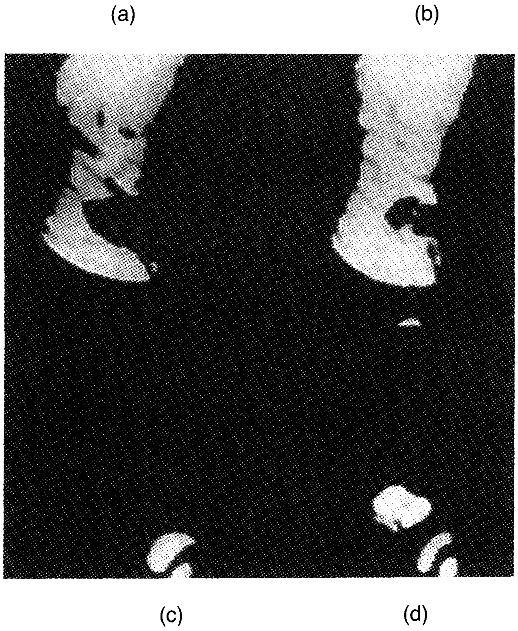

FIG. 16. Conventional spin-echo vs PEACH imaging of the breast. Enhancement is apparent in the breast mass on (b) the post Gd-DTPA image when compared to (a) the precontrast image. There is a much improved enhancing lesion–normal breast contrast using (d) the PEACH technique, particularly when compared to (c) the precontrast fat-suppressed image. (Reproduced with permission from Stark and Bradley, 1992.)

species (these pulses are frequently used as a means of suppressing the dominant water signal in most applications of MRS).

5. IMAGE ARTIFACTS

Artifacts are spurious features appearing on the image, which may arise from the MRI equipment, from some unwanted interaction between the patient and the imaging equipment, or from some external cause. What follows is necessarily a brief introduction, and the reader is recommended to read a comprehensive review by Henkelman and Bronskill (1987).

Anything that distorts or changes the main magnetic field will cause artifacts. Pieces of ferromagnetic material residing in, on, or close to the patient cause local field distortion that can obliterate the image or cause geometrical distortion along with altered brightness, depending on the distance of the object. Temporal variations of the magnetic field also cause artifacts, and random field variation over the time scale of the data collection for the image will give a smearing artifact in the phase-encode direction, the axis in the image defined by the phase-encode gradient.

Any physiological motion in the subject that changes in a regular cyclic manner, particularly if its frequency is harmonically related to the imaging sequence's basic repetition rate, can cause artifacts in the phase-encode direction. Heart motion and relaxed breathing are the main causes. Irregular motion such as fidgeting by the patient or peristalsis of the gut causes featureless smears.

Blood flow in major vessels and the heart can give rise to a number of different artifacts in normal imaging (as outlined in Sec. 4.3, the effects of flow can be put to good use in flow imaging). Constant and pulsatile flow produces changes in intensity due to fresh, unselected spins flowing into the slice between consecutive pulses, and artifacts similar to those due to cyclic motion.

In high-field imagers, the chemical shift between water and fat may cause the "water" and "fat" images to be slightly misplaced from each other in the frequency-encoding direction. At 1.5 T the two Larmor frequencies differ by 200 Hz, equivalent to several pixels of shift. Clinically, this can be confusing and open to misinterpretation. An example is the displacement of an image of the kidney relative to an image of the perinephric fat. Susceptibility effects due to gas pockets within the body can produce nearby field changes of the same order as chemical shift, with consequent distortion of the image.

Machine-generated or machine-specific artifacts cover a very large range. A common source is rf interference, which may arise from external sources or from components of the imaging spectrometer itself. These manifest themselves in the image as single bright dots, lines in the phase-encoding direction, or wide bands, depending on the stability of the source. Transient interference (for example, light switches or thermostats) often result in a characteristic "herringbone" pattern across the image.

Imbalance of the two channels of the quadrature detector (in phase or amplitude) will give rise to ghost-image aliases of the true image, mirrored about the center of the image frame. If the detected signal has been overamplified, the ADC may become overloaded, resulting in truncation of some of the data. This is reflected in nonzero or nonconstant background levels outside the image, edge ringing, and fringing in the outer regions.

6. SAFETY

Potential hazards associated with any medical technique should be considered for the patient, for equipment operators, and for the general public. For all these groups, risks should be kept to a minimum. MRI, by virtue of its nonionizing nature, is considered inherently safe when compared to, e.g., x-ray CT, but possible hazards may arise from several sources:

1. Chemicals—the only chemicals used in clinical studies are contrast agents (and at the moment only Gd-DPTA is in widespread use), and since they are intended for internal use they should be of low toxicity.
2. The biological situation—MRI, being a technique performed upon the intact body, offers no more biological hazard than is generally met in a hospital environment. Special precautions may need to be taken when high-risk patients are to be imaged,

e.g., patients with transmittable infections or with low immunity, such as premature babies or transplant patients undergoing immune suppression.

3. Electrical and electronic equipment—these should always be subject to regular safety checking according to local safety standards.
4. The inherent nature of MR equipment and procedure—this can be further subdivided into two sections:
   - **a.** Effects on nonbiological material—the presence of substantial magnetic fields means that any ferromagnetic object (e.g., steel screwdrivers, scissors, and scalpels) may become projectiles. The torque and attraction of magnetic fields will also affect metal inside the patient, and could have serious repercussions, e.g., twisting magnetically sensitive clips off blood vessels, resulting in hemorrhage. It is widely accepted that patients (or the general public) with cardiac pacemakers should not be exposed to (or allowed close to) MRI equipment.
   - **b.** Biomagnetic interactions—these involve the direct action of one or more of the three types of magnetic field upon the body. Various mechanisms have been put forward for possible effects from *static* field: changes in rate constants of metabolic reactions due to magnetically sensitive free-radical intermediates; alignment of body constituents containing anisotropic magnetic susceptibilities (e.g., DNA molecules, retinal rods, and entire sickled red blood cells); slowing down of ion-diffusion rates in the direction of the field; and induction of flow potentials in arteries. At the field strengths currently in use, the available literature suggests that these are not a source of hazard to patients or volunteers. However, guidelines dictate that the general public is not to be exposed to fields greater than 5 G. Slow, time-varying fields associated with gradient switching could produce eddy currents in conductive loops, possibly leading to "hot spots" where loop boundaries meet, or muscle twitching. These require very strong switching gradient fields (e.g., > 2000 T s^{-1}), not normally experienced in clinical imaging. A lower threshold effect is the generation of visual phosphenes (bright flashes in the eyes), which are not regarded as harmful. Finally, there are the effects of rf fields, whose main effect is generally considered to be local tissue heating. There are guidelines for exposure to rf fields in MRI [the Food and Drug Administration (FDA) in the United States and National Radiological Protective Board (NRPB) in the United Kingdom], based upon the assumption of heating. In both cases, a recommendation is made in terms of a maximum specific absorption rate (SAR) that yields a very small increase in body temperature, stated as about half the basic metabolic rate by the FDA, and as resulting in a body temperature rise of less than 1 °C by the NRPB.

GLOSSARY

Magnet Shimming: The way magnets are constructed and the presence of metallic surfaces close to the magnet mean that the main, static magnetic field is rarely exceptionally homogeneous. In order to compensate for these inhomogeneities, the main field is shimmed. This can be done in two ways: *Passive* shimming involves placing pieces of iron at strategic positions with respect to the main field, while *active* shimming uses magnetic fields generated from coils placed close to, or surrounding, the main magnet.

Magnetogyric Ratio: This is defined as the ratio of the magnetic moment of a nucleus to its angular momentum.

Nuclear Precession: When a nucleus is placed in a magnetic field, the magnetic moment experiences a torque that would tend to align it along the field. However, the angular momentum of the nucleus modifies the situation and causes the nucleus to behave like a gyroscope. In classical terms this means that the magnetic moment precesses about the magnetic field.

Radio-Frequency Pulses: The radio-frequency (rf) pulse is a burst of rf that is applied to the sample by way of an rf probe, and is usually made by gating a continuous source of rf on-and-off for a defined period. Quantum mechanically, the pulse is used to induce transitions between the various energy levels of the nuclear spin system, where the transitions are governed by the Bohr frequency condition. Classically, the action of the pulse manifests itself as a rotation of the nucleus through an angle θ about an axis defined by the phase of the rf. The angle $\theta = 2\pi\gamma H_1 t$, where t is the pulse duration, and the other terms are defined in the text.

Stray Field: This term refers to the magnetic field that is found outside the actual, physical magnet.

Works Cited

Abe, Z., Tanaka, K., Hotta, M., Imai, M. (1976), U.S. Patent No. 3932805, "Method of Obtaining Internal Information of Measuring a Target from the Outside by the Application of a Nuclear Magnetic Resonance Phenomenon."

Ballon, D., Jakubowski, A., Gabrilove, J., Graham, M. C., Zakowski, M., Sheridan, C., Koutcher, J. A. (1991), *Magn. Res. Med.* **19**, 85–95.

Bendall, M. R., Aue, W. P. (1983), *J. Magn. Res.* **54**, 149–152.

Black, H. S. (1953), *Modulation Theory*, Princeton, NJ: Van Nostrand.

Bloch, F. (1946), *Phys. Rev.* **70**, 460–474.

Bottomley, P. A., Foster, T. B., Darrow, R. D. (1984). *J. Magn. Res.* **59**, 338–342.

Bydder, G. M., Young, I. R. (1985), *J. Comput. Assist. Tomogr.* **9**, 659–675.

Callaghan, P. T. (1991), *Principles of Nuclear Magnetic Resonance Microscopy*, Oxford: Clarendon.

Carr, H. Y., Purcell, E. M. (1954), *Phys. Rev.* **94**, 630–638.

Damadian, R. (1974), U.S. Patent No. 3789832, "Apparatus and Method for Detecting Cancer."

Damadian, R., Goldsmith, M., Minkoff, L. (1977), *Physiol. Chem. Phys. Med. NMR* **9**, 97.

Dixon, W. T. (1984), *Radiology*, **153**, 189–194.

Edelstein, W. A., Hutchinson, J. M. S., Johnson, G., Redpath, T. (1980), *Phys. Med. Biol.* **25**, 751–756.

Frahm, J., Haase, A., Matthaei, D. (1986), *Magn. Res. Med.* **3**, 321–327.

Frahm, J., Hänicke, W., Merboldt, K.-D., Haase, A. (1985), *J. Magn. Res.* **64**, 81–93.

Gabillard, R. (1952), *Phys. Rev.* **85**, 694–695.

Garroway, A. N., Grannell, P. K. Mansfield, P. (1974), *J. Phys. C* **7**, L457–L462.

Haase, A., Frahm, J., Hänicke, W., Matthaei, D. (1985), *Phys. Med. Biol.* **30**, 341–344.

Henkelman, A. M., Bronskill, M. J. (1987), *Rev. Magn. Res. Med.* **2**, 5.

Hennig, J., Nauerth, A., Freidburg, H. (1986), *Magn. Res. Med.* **3**, 823–833.

Kumar, A., Welti, D., Ernst, R. R. (1975), *J. Magn. Res.* **18**, 69–83.

Lai, C. M., Lauterbur, P. C. (1980), *J. Phys. E* **13**, 747–750.

Lauterbur, P. C. (1973), *Nature* **242**, 190–191.

Lauterbur, P. C., Kramer, D. M., House, W. V., Chen, C.-N. (1975), *J. Am. Chem. Soc.* **97**, 6866–6888.

Mansfield, P. (1977), *J. Phys. C* **10**, L55–L58.

Mansfield, P., Chapman, B. (1987), *J. Magn. Res.* **72**, 211–223.

Mansfield, P., Grannell, P. K. (1973), *J. Phys. C* **6**, L422–L426.

Mansfield, P., Pykett, I. L. (1978), *J. Magn. Res.* **29**, 355–374.

Ordidge, R. J., Connelly, A., Lohman, J. A. B. (1986), *J. Magn. Res.* **66**, 283–294.

Ordidge, R. J., Howseman, A., Coxon, R., Turner, R., Chapman, B., Glover, P., Stehling, M., Mansfield, P. (1989). *Magn. Res. Med.* **10**, 227–240.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T. (1988), *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge: Cambridge University Press.

Purcell, E. M. Torrey, H. C., Pound, R. V. (1946), *Phys. Rev.* **69**, 37–38.

Schuierer, G., Ladebeck, R., Barfuss, H., Hentschel, D., Huk, W. J. (1991), *Magn. Res. Med.* **22**, 1–9.

Simon, J. H., Szumowski, J. (1989), *Radiology*, **171**, 539–543.

Smith, F. W. (1987), *Magn. Res. Med.* **5**, 118–128.

Stark, D. D., Bradley, W. G. (Eds.) (1992), *Magnetic Resonance Imaging*, St. Louis, MO: Mosby, Vols. 1 and 2.

Further Reading

Brey, W. S. (Ed.), *Journal of Magnetic Resonance*, Orlando, FL: Academic. Recent advances in NMR, published monthly.

Chen, C.-N., Hould, D. I. (1989), *Biomedical Magnetic Resonance Technology*, Bristol: Adam Hilger. Further details on NMR hardware.

Emsley, J. W., Feeney, J., Sutcliffe, L. H. (Eds.), *Progress in Nuclear Magnetic Resonance Spectroscopy*, Oxford: Pergamon. Review articles on various aspects of NMR, published bimonthly.

Fukushima, E., Roeder, S. B. W. (1981), *Experimental Pulse NMR: A Nuts and Bolts Approach*. Reading, MA: Addison-Wesley. A very readable and useful book on basic NMR and NMR hardware.

Stark, D. D., Bradley, W. G. (Eds.) (1992), *Magnetic Resonance Imaging*, St. Louis, MO: Mosby. Two volumes describing in detail the physics and clinical applications of MRI.

Traficante, D. D. (Sr. Ed.), *Concepts in Magnetic Resonance*, Kingston, RI: NMR Concepts. Detailed and informative explanations of NMR concepts, published quarterly.

Wehrli, F. W. (Ed.-in-chief), *Magnetic Resonance in Medicine*, Orlando, FL: Academic. Monthly periodical about NMR in medical research.

Wehrli, F. W., Shaw, D., Kneeland, J. B. (1988), *Biomedical Magnetic Resonance Imaging: Principles, Methodology and Applications*, New York: VCH.

27

Nuclear Magnetic Resonance Spectrometry

Vladimir I. Bakhmutov

27.1 Introduction 935
27.2 NMR: General Considerations 936
27.2.1 Nuclei in the External Magnetic Field 936
27.2.2 Detection of an NMR Signal – General Principles 938
27.2.3 NMR Equipment 939
27.2.3.1 Magnets 939
27.2.3.2 NMR Probes 940
27.3 NMR Spectral Parameters 940
27.3.1 Chemical Shift 940
27.3.1.1 Anisotropy of Chemical Shift 942
27.3.1.2 Quadrupolar Splitting 942
27.3.2 Influence of an Unpaired Electron on the Chemical Shift 942
27.3.3 Spin–Spin Coupling 943
27.3.4 Nuclear Relaxation 945
27.3.4.1 Relaxation Mechanisms in Liquids 945
27.3.4.2 Paramagnetic Relaxation 947
27.3.4.3 Nuclear Relaxation in Solids: Spin Diffusion 947
27.3.4.4 Relaxation Time Measurements 948
27.4 Spectroscopic NMR Experiments in Liquids 948
27.4.1 Routine One-Dimensional (1D) NMR spectra 948
27.4.2 Polarization Transfer in NMR (SPI, INEPT, and DEPT) 949
27.4.3 Two-Dimensional (2D) NMR Spectra 950
27.4.4 Chemical Exchanges: the NMR Time Scale 950
27.4.5 Translational Molecular Diffusion via NMR 951
27.4.6 Isomeric and Conformational Analysis in Solutions by NMR 953
27.5 NMR Spectroscopic Experiments in the Solid State 954
27.5.1 Wide-Line NMR 954
27.5.1.1 Dead Time in Wide-Line NMR 955
27.5.1.2 Relaxation Experiments in Solids 955
27.5.2 High-Resolution Solid-State NMR 956
27.5.2.1 High-Power Decoupling and Double Resonance 956

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

27.5.3 Cross Polarization Technique 957
27.5.4 Internuclear Distances from the Solid-State NMR Data 957
27.5.5 Zero-Field Solid-State NMR Experiments 958
27.6 Sensitivity Enhancement by Dynamic Nuclear Polarization (DNP) 959
27.7 Imaging NMR 960
References 961
Further Reading 962

27.1
Introduction

The NMR spectrometry is an extraordinarily powerful analytical method, which originally had applications in physics and later to chemistry, biochemistry, biology, pharmacy, geophysics, materials science, and archaeology. Nowadays, this method is widely used in veterinary science and medicine, particularly in clinical research for routine human brain NMR imaging (magnetic resonance imaging (MRI)).

As a method, the NMR spectroscopy is based on resonance phenomenon independently discovered by Bloch *et al.* (1946) and Purcell, Torrey, and Pound (1946), which involves placing nuclear spins in an external magnetic field and their excitation by a radio-frequency irradiation. The energy, thus absorbed by the nuclei, measured in frequency units, is characteristic of the nuclei and their environments and can be used for the structural analysis of the compounds investigated.

Initially, the NMR experiments were performed by physicists to experimentally verify the basic postulates of quantum mechanics that require detailed knowledge and accurate measurements of nuclear magnetic moments. However, the development of the relaxation theory (Bloembergen, Purcell, and Pound, 1948) and also the discovery of chemical shift (Proktor and Yu, 1950) and spin–spin coupling patterns of nuclei have shown the great potential of NMR to describe solids, liquids, and gaseous matter in terms of structural and dynamic properties. Then, fast successes in the radio-frequency technology resulted in the appearance of the first commercial NMR spectrometers (1953), which strongly impacted the research in chemistry and biochemistry. In turn, more complex compounds, such as biochemical molecular systems and aggregates, required more powerful NMR spectrometers, modifications in magnet technology, and developments of new spectroscopic approaches. Therefore, the modern commercial NMR spectrometers represent multiple-pulse devices, capable of experiments in solutions and the solid state, which can operate at very high magnetic fields up to 21.14 T and detect nuclei with a small natural abundance (rare nuclei). In addition, increasing the speed of computers, performing very fast Fourier transformations (FTs), has resulted in routine use of the two- or three-dimensional NMR experiments.

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

There are numerous techniques and approaches in NMR spectrometry. They are constantly modified and published in original articles, books, and the NMR Encyclopedia. Even a small portion of the NMR methods cannot be covered in this relatively short chapter, the aim of which is to present only the general principles of NMR experiments and their practical applications.

27.2
NMR: General Considerations

After publications of the pioneering works, describing NMR, its practical applications can be grouped under two categories: NMR spectrometry and NMR relaxation. In the first group, an NMR spectrum is studied, which is a frequency-dependent pattern, showing a number of nonequivalent nuclei and their environments. In contrast, the resulting data, collected by the relaxation measurements, are time-dependent and are rather related to dynamics in molecular systems. The first application is more popular among practical chemists or biochemists while the relaxation aspects attract researchers working in molecular physics, physics of solids, and biophysics. However, it should be emphasized that the frequency- and time-dependent data are the essence of the same phenomenon, and knowledge of nuclear relaxation is needed even for recording a simplest NMR spectrum because obtaining a spectrum itself can become difficult or impossible, when relaxation times are extremely long. It is probable that the first attempt to observe NMR for ^{1}H and ^{7}Li nuclei in 1936 was unsuccessful (Gorter *et al.*, 1942) because of long relaxation times. In contrast, short relaxation times broaden NMR lines and often mask their fine structures. Sometimes, the broadenings are very strong and lines become "invisible." This is obviously important even for users applying NMR only for an express structural analysis.

27.2.1
Nuclei in the External Magnetic Field

According to Abragam (1961), a nucleus with nonzero angular momentum, P, possesses a nuclear magnetic moment, $\mu = \gamma P$, capable of interaction with the external magnetic field, where γ is the nuclear magnetogyric ratio, which depends on the nature of nuclei (Table 27.1). The P and the μ are quantized. For example, the projections of the angular nuclear momentum, P_Z, on the OZ axis can be written as

$$P_Z = \hbar \times m_I \tag{27.1}$$

where h ($\hbar = h/2\pi$) is the Planck's constant ($h/2\pi = 1.05457266 \times 10^{-34}$ J s) and m_I is the magnetic quantum number. The latter depends on nuclear spin I and takes values from I to $-I$: I, $I-1$, $I-2$, …, $-I$. In turn, I is a multiple of ½, for example, at $I = ½$, the angular momentum and the magnetic moment are

$$P_Z = \pm\left(\frac{1}{2}\right)\hbar \text{ and } \mu_Z = \pm\left(\frac{1}{2}\right)\gamma\hbar \tag{27.2}$$

Then, the μ_Z (or P_Z) projections, represented as vectors, can be parallel or antiparallel relative to the above Z axis. If the external magnetic field B_0 is applied along this axis, the μ_Z (or P_Z) orientations become energetically nonequivalent. This nonequivalency leads to the, so-called, Zeeman's energy splitting ΔE

Tab. 27.1 NMR properties of some nuclei.

Nucleus	**Spin I**	**Natural abundance (%)**	**NMR frequency, ν_0, in megahertz at $B_0 = 2.3488$ T**	**γ (10^7 rad T^{-1} s^{-1})**	**Sensitivity relative to ^{1}H**
^{1}H	$^1/_2$	99.98	100	26.752	1.000
^{2}H	1	0.016	15.35	4.107	1.45×10^{-6}
^{11}B	$^3/_2$	80.42	32.08	8.584	0.133
^{13}C	$^1/_2$	1.108	25.14	6.728	1.76×10^{-4}
^{14}N	1	99.63	7.22	1.934	1.00×10^{-3}
^{19}F	$^1/_2$	100	94.08	25.18	0.834
^{31}P	$^1/_2$	100	40.48	10.84	0.065
^{17}O	$^5/_2$	0.037	13.557	−3.628	1.08×10^{-5}
^{93}Nb	$^9/_2$	100	24.549	6.567	0.487
^{117}Sn	$^1/_2$	7.61	35.63	−9.578	3.49×10^{-3}
^{119}Sn	$^1/_2$	8.58	37.29	−10.02	4.51×10^{-3}
^{199}Hg	$^1/_2$	16.84	17.91	4.815	9.82×10^{-4}
^{205}Tl	$^1/_2$	70.5	57.63	15.589	0.140

(Figure 27.1) proportional to the μ_Z magnitude and the strength of the applied magnetic field, B_0:

$$\Delta E = 2\,\mu_Z B_0 \tag{27.3}$$

Then, formulating ΔE, as an energy quantum, $\Delta E = h\nu$, leads to the resonance condition written as

$$h\nu_0 = 2\,\mu_Z B_0 = \gamma \hbar B_0$$

$$\nu_0 = \gamma B_0 \tag{27.4}$$

This condition shows that the nuclear spin, placed into the external magnetic field, B_0, undergoes a single-quantum transition (i.e., the m_I number changes from $-1/2$ to $+1/2$), when an irradiation frequency reaches ν_0 (the Larmor's frequency). The transition results in an energy absorption by lattice, which is registered as an NMR signal. Table 27.1 shows the ν_0 frequencies for different nuclei.

An ensemble of nuclei will populate the Zeeman's energy levels according to factor $\exp(-\Delta E/kT)$, where T is the temperature and k is the Boltzmann's constant (1.380658×10^{-23} J K^{-1}). Thus the low-energy levels are more populated than the high levels, which give raise to macroscopic magnetization, undergoing precession around the B_0 direction

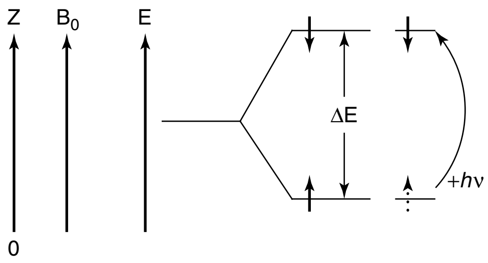

Fig. 27.1 The Zeeman energy levels of nuclear spins $I = {}^1/_2$ in the external magnetic field B_0.

at the ν_0 frequency, responsible for observations of NMR signals. It should be emphasized that the population difference, ΔN, compared with a total number of the nuclei, N, is less: a $\Delta N/N$ ratio is of 8×10^{-6} for ^{1}H nuclei in a field of 2.35 T at 300 K. This condition defines relatively low sensitivity of NMR experiments. As seen in Table 27.1, ^{1}H and ^{19}F nuclei are mostly preferable as target nuclei in this context.

Nuclei with spin > ½ behave similarly in the external magnetic filed. However, a number of the Zeeman's levels increases accordingly. (Such nuclei have nonspherical charge distributions responsible for appearance of nuclear quadrupole moments Q). For example, ^{2}H or ^{14}N nuclei ($I = 1$) give three energetically equidistant levels, corresponding to the μ_Z values of $+\gamma\hbar(1)$, 0 and $-\gamma\hbar(1)$. Because the double quantum transitions (the m_I value changes from +1 to −1) are not observed, the nuclei again produce a single line in an NMR spectrum.

27.2.2
Detection of an NMR Signal – General Principles

Technically, observation of an NMR signal can be realized at the ν_0 irradiation of a sample by the continuous wave (CW) method. The principle schemes of such CW spectrometers can be found in Harris (1986), Sanders and Hunter (1994), or Duel (2002). Here, a spinning sample is situated between the poles of a magnet, radio frequency is applied via an antenna coil, connected to a transmitter, and a receiver coil, surrounding the sample, registers the absorbed energy by electronic devices and a computer. The CW spectrometers imply a sweeping of the magnetic field or the radio frequency when the external magnetic field remains constant.

A simple NMR experiment, where the ν_0 irradiation is performed by a radio-frequency pulse (RFP), is shown in Figure 27.2. The experiment includes three time sections: (i) relaxation delay (RD), (ii) action of RFP, and (iii) recording the NMR data as free induction decays (FID)s, during acquisition time (AT). According to the scheme, the NMR data can be accumulated, when sections from (i) to (iii) are repeated. The resulting FIDs are time-dependent, $f(t)$, and should be converted into the NMR spectra, as the frequency-dependent data, $F(\nu)$, by FT:

$$F(\nu) = \int f(t)\ \exp(-i2\pi\nu t)\ dt$$

$$f(t) = \int F(\nu)\exp(+i2\pi\nu t)\ d\nu \quad (27.5)$$

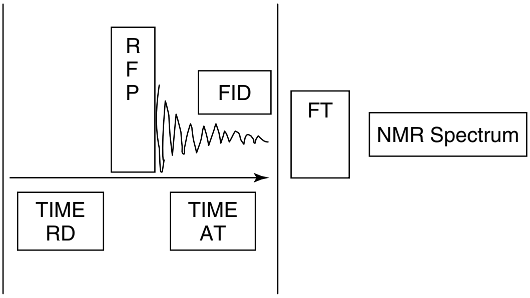

Fig. 27.2 Schematic presentation of the 1D NMR experiment where typical durations of radio-frequency pulse (RFP), relaxation delay (RD), and acquisition time (AT) are of 5–10 microseconds, 1–4 seconds, and 1–3 seconds, respectively.

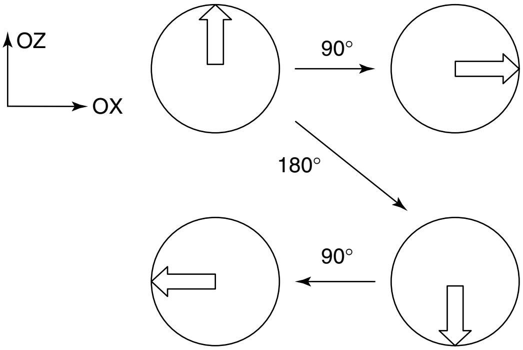

Fig. 27.3 The schematic presentation of the behavior of the macroscopic magnetization vector, **M**, in a rotating coordinate system after action of 90°- and 180°-radio-frequency pulses or two consequent 180°- and 90°-radio-frequency pulses.

Such an experiment is realized on an FT NMR spectrometer equipped with a computer capable of fast FT. It should be emphasized that durations and amplitudes of RFP should be sufficient to rotate the macroscopic magnetization, for example, by an angle α from the magnetic field B_0. According to Ernst *et al.* (1966), the pulse angle, α, is defined as $\alpha = \gamma B_1 t_P$, where B_1 is the power of the pulse and t_p is its duration. Then, the magnetization goes to thermal equilibrium. Figure 27.3 schematically shows the behavior of the macroscopic magnetization vector, **M**, in a rotating coordinate system due to actions of 90°- or 180°-RFPs. Here, the **M** orientation along the positive OX axis corresponds to the maximal intensity of a registered NMR signal. Thus, after a 180° pulse, the signal is not observed or its intensity becomes negative after a combination of 180 and 90° pulses.

27.2.3
NMR Equipment

As shown above, an NMR spectrometer includes the following basic units: a radio-frequency-reference generator, a transmitter, a receiver, a pulse generator, a computer, a magnet, and an NMR probe. All of the units remarkably affect the quality of the NMR spectra and the signal/noise ratio as a measure of sensitivity. However, for experimentalists, a magnet and an NMR probe are most important.

27.2.3.1 Magnets

A magnet is a source of strong, stable, and homogeneous magnetic field to obtain the high-resolution NMR spectra. When characteristics of transmitters, receivers, and other devices of NMR spectrometers do not change, then increasing the magnetic field, B_0, increases the energy, ΔE, population difference $\Delta N/N$ and hence the sensitivity of an NMR spectrometer. Iron magnets (permanent or electromagnets) are relatively unstable, sensitive to the temperature, and create moderate field strengths, corresponding to the working ^{1}H frequency of $\leq$100 MHz. In contrast, modern superconducting electromagnets show excellent stability, good homogeneity and their fields are capable of working at the proton frequency up to 900 MHz. For both types of magnets, stability of the magnetic field needed for the high-resolution NMR experiments can be additionally increased by field-locking. Because deuterium NMR signals coming from deuterated solvents can be well filtered from signals of ^{1}H or other nuclei, they are successfully applied for additional field stabilization.

27.2.3.2 NMR Probes

Generally, an NMR probe represents a single tank circuit needed for irradiation and detection of an NMR signal. According to Kisman and Armstrong (1974), the radio-frequency, coming from the transmitter, must be distinguished from the signal, coming from the probe. In other words, a weak signal has to be measured immediately after the action of a high-power pulse. Such a switching problem gives rise to the, so-called, deed time of the spectrometer. If the deed time is long and the FID is short, the NMR signal loses intensity. Technically, this problem is minimized in modern NMR spectrometers and their standard components reduce the deed times to few microseconds.

The design of NMR probes significantly depends on desired NMR experiments. For example, variable-temperature measurements (usually between −120 and +120 °C) require the corresponding heating/cooling equipments, which provide accurate temperature settings and the absence of significant temperature gradients. Double or triple resonance and gradient-field experiments are also available when NMR probes have a special design. Finally, an NMR probe should be equipped with tools for the mechanical rotation of samples. Low spinning rates (approximately 20 Hz) are generally used for liquids to compensate the nonhomogeneity of the external magnetic field. Very high spinning rates (tens of kilohertz) are needed to suppress strong dipole–dipole interactions in solids. In the latter case, the NMR rotors are oriented at the, so-called, magic angle setting relative to the external magnetic field (see below).

27.3
NMR Spectral Parameters

Nuclear spins in molecules are not isolated and this circumstance gives rise to the phenomena of chemical shift, spin–spin coupling, and relaxation, directly related to chemical structures. In turn, these spectral parameters lead to reliable structural determinations even in the complex cases providing, thus, wide applications of the NMR spectroscopy as an analytical method.

27.3.1
Chemical Shift

Electrons, circulating around nuclei, placed into the external magnetic field, B_0, create a new magnetic field, opposing the initial field. Then, a local magnetic field, B_{loc}, modifies the resonance frequency ν_0, which leads to shielding constant σ, $B_{\text{loc}} = (1 - \sigma)B_0$, or chemical shift $\delta = -\sigma$. Chemical shift δ (Figure 27.4) is measured in parts per million and referred to *internal (or external) standards:* $\delta = \{\nu$ (sample) − ν (reference)$\}$ $10^6/\nu$ (reference) (generally to TMS for ^{1}H and ^{13}C nuclei, H_3PO_4 for ^{31}P nuclei, CF_3COOH for ^{19}F nuclei, and so on).

According to Abragam, three main terms contribute to total constant σ: diamagnetic term σ^{local}(dia), paramagnetic term σ^{local}(para), and magnetic anisotropy term σ^*. The first contribution, caused by the circulation of

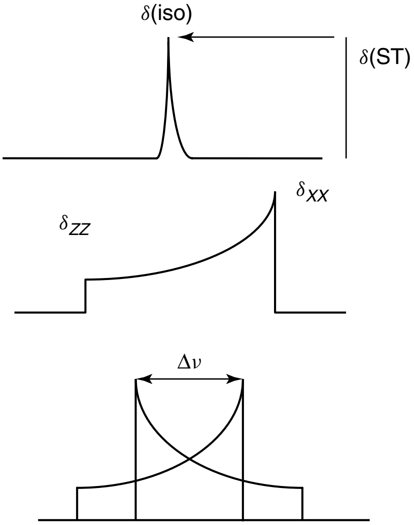

Fig. 27.4 The line shape of an NMR resonance typical of a liquid (a), a static solid with an axially symmetrical screening tensor, $\sigma_{ZZ} \neq \sigma_{XX} = \sigma_{YY}$, (b) and a powder ^{2}H sample with an axially symmetric electric field gradient at deuterium.

unperturbed spherical electrons (s) with mass m_e, is expressed in a simplest form by Lamb's equation:

$$\sigma^{\text{local}}(\text{dia}) = \left(\frac{\mu_0 e^2}{8\pi m_e}\right) \int r\rho_e \, dr \quad (27.6)$$

where μ_0 is the permeability constant, ρ_e is the charge density, and r is the electron-nucleus distance. The second term, formulated by McConnel *et al.* (1958), appears due to perturbed non-spherical electrons (p or d) and it can be described via the average energy of the electron excitation, ΔU, as

$$\sigma^{\text{local}}(\text{para}) = -\left(\frac{\mu_0 e^2 \hbar^2}{6\pi m^2 \Delta U}\right)\langle \text{r}^{-3}\rangle P_U \quad (27.7)$$

Here P_U is named as the p-electron "imbalance" expressed via charge densities and bond orders (Jameson and Gutowsky, 1964). The last (relatively small) term, σ^*, comes from electrons situated at neighboring groups. These groups (for example, bonds C=C or C≡C) create magnetically anisotropic fields and therefore the observed effects depend on orientations of observed nuclei. Combinations of these terms lead to large variations in chemical shifts δ, which depend on the nature of nuclei (Table 27.2). A correct quantum-chemical analysis of chemical shifts is possible, but the calculations are not simple. Therefore, in practice,

Tab. 27.2 Chemical shift ranges, δ (SCS), for different nuclei.

Nucleus	**SCS (parts per million)**[a]
^{1}H	−50 to +20
^{11}B	−130 to +95
^{13}C	−300 to +250
^{19}F	−250 to +400
^{31}P	−125 to +500
^{15}N	−500 to +850
^{195}Pt	−1370 to +12 000

[a] The positive values correspond to low-field displacements.

structural determinations or predictions of the δ values for known structures are based on numerous spectrum/structure relationships or increment equations, which can be found in Harris (1986) or Sanders and Hunter (1994). Finally, the so-called, isotopic chemical shifts (for example, for ^{1}H at ^{12}C and ^{1}H at ^{13}C) are relatively small and their role in the structural analysis is not significant.

27.3.1.1 Anisotropy of Chemical Shift

Screening constants σ are three dimensional and can be characterized by components σ_{XX}, σ_{YY}, and σ_{ZZ} that are also associated with chemical structures (Strub *et al.*, 1983). However, generally, this coordinate system does not correspond completely to a molecular coordinate system, though the Z coordinate is often situated along chemical bonds.

Generally, the σ components differ leading to chemical shift anisotropy, $\Delta\sigma$, defined as $\Delta\sigma = \{2\sigma_{ZZ} - (\sigma_{XX} + \sigma_{YY})\}/3$. The high-amplitude molecular motions in liquids produce the Lorenz-shaped resonances, where the σ_{XX}, σ_{YY}, and σ_{ZZ} components are not observed. Here, the resonances are characterized by isotropic chemical shifts $\delta(ISO)$ (Figure 27.4) or isotropic screening constants $\sigma(ISO)$ determined as $\sigma(ISO) = (1/3)(\sigma_{XX} + \sigma_{YY} + \sigma_{ZZ})$. More complex lines are observed in rigid solids and their shapes directly depend on the σ (or δ) components that are determined from static NMR patterns (Duel, 2002). Figure 27.4 shows a solid-state NMR signal, typical of an axial-symmetrical screening-tensor where $\sigma_{XX} = \sigma_{YY} \neq \sigma_{ZZ}$. A lower symmetry ($\sigma_{XX} \neq \sigma_{YY} \neq \sigma_{ZZ}$) obviously leads to more complicated signals. This is particularly important for heavy nuclei such as ^{19}F, ^{31}P or ^{117,119}Sn.

27.3.1.2 Quadrupolar Splitting

Spins of nuclei, having quadrupole moments Q (for example ^{2}H, ^{14}N, ^{81}Br, ^{127}I, and ^{35}Cl), interact not only with the external and local magnetic fields but also with electric field gradients (EFG) at these nuclei. The energy of interactions is expressed through nuclear quadrupole coupling constant (NQCC), $NQCC = e^2 q_{zz} Q/h$, where eq_{ZZ} is the principal component of the EFG tensor. In rigid solids, the quadrupolar interaction directly affects the shapes of NMR resonances. Figure 27.4 shows a typical ^{2}H NMR spectrum, recorded in a static powder sample, where the signal exhibits the, so-called, quadrupolar splitting, measured in frequency units (kilohertz) and marked as $\Delta\nu$. This shape is observed for systems with the axially symmetric EFG. In the absence of motions, the splitting provides determinations of the NQCC values via $\Delta\nu = 3/4(e^2 q_{ZZ} Q/h)$. In liquids, however, the quadrupolar interaction is not observed because of the fast molecular motions and NMR signals become again Lorenz-shaped.

27.3.2

Influence of an Unpaired Electron on the Chemical Shift

The very high magnetic moment of an unpaired electron affects an NMR frequency via effective magnetic field B_{EF}:

$$\begin{aligned} B_{\mathrm{EF}} &= 2\pi A \mu_B g B_0 S \frac{(S+1)}{3\gamma_N kT} \\ &\quad + \mu_0 {\mu_B}^2 g^2 B_0 S(S+1) \\ &\quad \times \frac{(3\ \cos^2\theta - 1)}{3kTr^3} + B_{\mathrm{BMS}} \quad (27.8) \end{aligned}$$

Here, according to Poople (1983), A is the isotropic electron/nucleus hyperfine coupling constant, μ_B is the Bohr magneton (9.2741×10^{-24} J T^{-1}), r is the electron/nucleus distance, θ is the angle between the electron-nucleus vector and the external magnetic field $B_{0,}$ and g is the electron g factor. The first term produces the, so-called, Fermi-contact (FC) chemical shifts (McConnel *et al.*, 1958). Owing to the direct delocalization of the unpaired electron density on the nuclei, the FC shifts in solids can reach 3000 ppm or higher as a function of the nature of the nuclei and paramagnetic centers. Generally, the chemical shift tensors are axially symmetric at large anisotropies covering ranges up to 3500–4000 ppm (Mali *et al.*, 2005). In a combination with short relaxation times, the nuclei affected by FC interaction are difficult to observe experimentally. In liquids, in spite of the fast molecular motions, the FC shifts can be also significant (up to 100 ppm), mainly due to the very large electron/nucleus coupling.

The second field, causing smaller shifts, corresponds to dipolar electron/nucleus interaction. The interaction strongly depends on distance r and includes an angle-dependent part, $(3\cos^2\theta - 1)$. In liquids, where molecular reorientations with respect to the external magnetic filed are fast, θ is defined as an angle formed by the r vector and an axis of the symmetry of the g factor (Harris, 1986). The corresponding chemical shifts are named as *pseudo-contact*, which are again associated with chemical structure. The last term, B_{BMS}, represents a demagnetization field appearing in a sample due to bulk magnetic susceptibility (BMS). This effect, generally insignificant in diamagnetic samples, plays a very important role in paramagnetic systems (Kubo, Spaniol, and Terao, 1998).

27.3.3
Spin–Spin Coupling

According to the theory, which can be found in Abragam (1961) or Harris (1986), the spin Hamiltonian, describing NMR in isotropic fluids in the limits of the quantum-mechanical formalism, takes the form:

$$h^{-1}\hat{H} = -\sum\nolimits_j (2\pi)^{-1}\gamma_j B_0(1-\sigma_j)\hat{I}_j^z + \sum\nolimits_{\mathrm{j<k}} J_{jk}\ \hat{I}_j\hat{I}_k \qquad (27.9)$$

The first part, modified by screening constant σ, corresponds to isolated nuclear spins in the field, B_0 (see Eq. (27.4)). The second term is spin–spin coupling and J_{jk} is the spin–spin coupling constant. Because the J_{jk} constants are relatively small (typically from 1–2 to 200–300 Hz), spin–spin coupling is considered as perturbation of the first term.

Effects of perturbation on a number of Zeeman's energy levels can be easily shown by attributing the symbols α and β to the parallel and antiparallel mutual orientations for nuclear spins of ½. Then, the spin states for a system, consisting of two nuclei, are $\beta\beta(m_I = -1)$, $\alpha\beta(m_I = 0)$, $\beta\alpha(m_I = 0)$, and $\alpha\alpha(m_I = 1)$. Thus, the first nuclear spin can "see" the second one in two states and its energy levels undergo an additional spitting by energy J measured in hertz. The J constant is a scalar magnitude, has a sign, and is independent of the external magnetic field B_0.

Two magnetically nonequivalent nuclear spins, A and X, form a simple spin system, AX, when the chemical shift difference $\Delta\delta^{AX}$ (measured in hertz) is significantly larger than the spin–spin coupling constant, $J(A-X)$: $\Delta\delta^{AX}(\text{hertz})/J^{AX}(\text{hertz}) > 10$. Generally, line multiplicity N of the A resonance is $N = 2nI + 1$, where n is the number of the X nuclei with spin I. Relative intensities within multiplets are distributed according to the so-called Pascal's triangle. For two protons ($I = ½$), the AX spin system shows two doublets at 1 : 1 intensity ratio for each doublet. In these terms, group $CH_2{-}CH_3$ should exhibit two ^1H resonances as a quartet and triplet, respectively. Decreasing the $\Delta\delta^{AX}(\text{hertz})/J^{AX}(\text{hertz})$ ratio transforms the AX spin system to a strongly coupled system, named as AB, to emphasize the small chemical shift difference. Similarly, simple spin systems AX_2 or AXX' are converted to strongly coupled systems ABX, AMX, AB_2 or ABC, which show complex NMR patterns. Owing to the presence of combinational transitions, their intensities and multiplicities depend on values $\Delta\delta$ and J. Such systems can be analyzed via simulations or fittings to the experimental data by convenient computer programs, using quantum-mechanical formalism. Because the J constant does not depend on B_0, the strongly coupled spin systems are simplified in higher external magnetic fields.

Spin–spin coupling is transmitted via electrons, surrounding the nuclei, and therefore the J constant is a function of chemical structures: electron distributions, atomic hybridizations, and interbond angles. Generally, the total J energy consists of three contributions coming from the spin-orbital (SO), spin-dipolar (SD), and FC interactions (Abragam, 1961) written as

$$H_{\text{SO}} = \left(\frac{\mu_0}{4\pi}\right) g_L \mu_B \gamma_N \hbar \sum 2\, L_e\, I_N\, r_{\text{eN}}^{-3} \tag{27.10}$$

$$H_{\text{SD}} = -\left(\frac{\mu_0}{4\pi}\right) g_S \mu_B \gamma_N \hbar \sum \{ r^{-3}{}_{\text{eN}} I_N S_e - 3\, r^{-5}{}_{\text{eN}} (I_N r_{\text{eN}})(S_e r_{\text{eN}}) \} \tag{27.11}$$

$$H_{\text{FC}} = \left(\frac{2}{3}\right) \mu_0 g_S \mu_B \gamma_N \hbar \sum \delta(r_{\text{eN}}) I_N S_e \tag{27.12}$$

where e and N correspond to electron and nucleus, respectively, L_e is the orbital angular moment operator, r_{eN} is the distance between electron and nucleus, and $\delta(r_{\text{eN}})$ is the Dirac delta function. The function takes a value of 1 at $r_{\text{eN}} = 0$ and 0 otherwise, requiring, thus, the presence of electron density directly on the nuclei. Calculations show that FC contribution usually dominates (Table 27.3) and hence J constants can be interpreted in terms of electron densities.

Spatial proximity of some nuclei, for example, ^{19}F, leads to the, so-called, through-space spin–spin coupling constants (sometimes reaching 100 Hz and more). Here, according to Contreras *et al.* (2000), again FC interaction dominates

Tab. 27.3 Contributions to total spin–spin coupling constants (in hertz).

Coupling constant	Molecule	FC	SO	SD
$^1J(C{-}C)$	C_2H_6	11.99	−0.55	0.3
	C_2H_2	69.43	4.13	2.84
$^1J(F{-}C)$	CH_3F	−126.30	−4.15	6.92
$^2J(F{-}F)$	CF_4	−1.09	6.20	16.73

but the coupling goes via nonbonding electron orbitals.

27.3.4
Nuclear Relaxation

The nuclear spins, excited by a radio-frequency irradiation, are capable of an energy exchange with their environments to reach the initial equilibrium state. There are two principally different relaxation mechanisms. The first mechanism, named as *spin-lattice relaxation,* represents the energetic exchange between nuclear spins and lattice, formulated as a continuum of nuclear and/or electron magnetic moments of any sort. The second mechanism, named as *spin–spin relaxation,* occurs when two (or more) spins undergo "flip–flop motions," which shorten, thus, the spin lifetimes while the total energy of the system does not change (for this reason, such flips are often named as the *energy-conserving spin transitions.* Finally, the relaxation of macroscopic magnetization is given by the Bloch's equations:

$$\frac{dM_Z}{dt} = \frac{-(M_Z - M_Z{}^0)}{T_1} \quad \text{and}$$

$$\frac{dM_{X,Y}}{dt} = \frac{-M_{X,Y}}{T_2} \qquad (27.13)$$

and formulated as recovery of the Z-(longitudinal) component with a time constant T_1 and the Y- and X-(transverse) components with a time constant T_2.

27.3.4.1 Relaxation Mechanisms in Liquids

Nuclear relaxation will be effective if magnetic fields in the spin environments fluctuate at frequencies close to the Larmor's resonance frequency ν_0 (or $\omega_0 = 2\pi\nu_0$). In liquids, such fluctuations are directly associated with thermal molecular motions: rotational reorientations, translational motions, or their combinations. Rapid intramolecular motions also contribute to nuclear relaxation. Molecular motions are characterized by correlation times τ and activation energies ΔE that can be determined from the relaxations data if the mechanism of the process is known.

According to the theory of Bloembergen, Purcell and Pound (1948), nuclei can relax via dipole–dipole, quadrupole, spin-rotation, scalar, and chemical shift anisotropy interactions. Generally, the relaxation rates are written as $1/T_{1,2} = CJ(\omega_0, \tau)$, where C is the coupling strength and $J(\omega_0, \tau)$ is the spectral density function. Relaxation due to dipolar interaction of two identical nuclei can be expressed as

$$\frac{1}{T^{\mathrm{DD}}{}_1} = \left(\frac{2}{5}\right)\gamma^4\hbar^2 r^{-6} I(I+1) \times \left\{\frac{\tau}{(1+\omega_I{}^2\tau^2)} + \frac{4\tau}{(1+4\omega_I{}^2\tau^2)}\right\} \qquad (27.14)$$

$$\frac{1}{T^{\mathrm{DD}}{}_2} = \left(\frac{1}{5}\right)\gamma^4\hbar^2 r^{-6} I(I+1) \times \left\{3\tau + \frac{5\tau}{(1+\omega_I{}^2\tau^2)} + \frac{2\tau}{(1+4\omega_I{}^2\tau^2)}\right\} \qquad (27.15)$$

where ω_I is the resonance frequency. This mechanism is particularly powerful in strong dipolar coupling, $\gamma^4\hbar^2 r^{-6} I(I+1)$, particularly at short internuclear distances r and at molecular motional frequencies close to ω_I. Similarly, relaxation of nucleus A via hetero-nuclear dipolar coupling by nucleus B is described by the following equations:

$$\frac{1}{T_1} = \left(\frac{4}{30}\right)\left(\frac{\mu_0}{4\pi}\right)^2 r(A-B)^{-6}{\gamma_A}^2{\gamma_B}^2 \times \hbar^2 I_B(I_B+1)\left\{\frac{3\tau}{(1+{\omega_A}^2\tau^2)} + \frac{6\tau}{(1+(\omega_A+\omega_B)^2\tau^2)} + \frac{\tau}{(1+(\omega_A-\omega_B)^2\tau^2)}\right\} \quad (27.16)$$

$$\frac{1}{T_2} = \left(\frac{4}{30}\right)\left(\frac{\mu_0}{4\pi}\right)^2 r(A-B)^{-6}{\gamma_A}^2{\gamma_B}^2 \times \hbar^2 I_B(I_B+1)\left\{4\tau + \frac{3\tau}{(1+{\omega_A}^2\tau^2)} + \frac{6\tau}{(1+{\omega_B}^2\tau^2)} + \frac{\tau}{(1+(\omega_A-\omega_B)^2\tau^2)} + \frac{6\tau}{(1+(\omega_A+\omega_B)^2\tau^2)}\right\} \quad (27.17)$$

Because the molecular motion correlation times τ are temperature-dependent, that is, $\tau = \tau_0 \exp(E_a/RT)$, the T_1 plots of Eqs. (27.14) and (27.16) in semilogarithmic coordinates are symmetrical, V-shaped and go through minima. The slops of the curves correspond to the E_a values and the τ-independent minima provide determinations of internuclear distances. Finally, if target nuclei interact with a number of neighboring nuclei, then the corresponding contributions should be summed up.

Nuclei with spins > ½, interacting with the EFG, eq_{ZZ}, relax by the quadrupolar mechanism:

$$\frac{1}{T_1}(Q) = \left(\frac{3}{50}\right)\pi^2(2I+3)(I^2(2I-1))^{-1} \times \left(\frac{e^2 q_{zz} Q}{h}\right)^2\left(1+\frac{\eta^2}{3}\right)\left(\frac{\tau}{(1+{\omega_Q}^2\tau^2)} + \frac{4\tau}{(1+4{\omega_Q}^2\tau^2)}\right) \quad (27.18)$$

where η is the asymmetry parameter (corresponding to 0 at an axial symmetry). The T_1(Q) times can change in large limits ($10^{-4} - 10^4$ ms), strongly depend on the nature of nuclei and particularly on the symmetry of their environments. For example, the ^{14}N T_1 time is as long as 10^4 ms in molecule $Me_4{}^{14}NBr$ due to the completely symmetrical environment of ^{14}N nuclei.

The chemical-shift-anisotropy mechanism (CSA)

$$\frac{1}{T_1}(\text{CSA}) = \left(\frac{1}{15}\right){\gamma_I}^2{B_0}^2(\Delta\sigma)^2 \times \left(\frac{2\tau}{(1+{\omega_I}^2\tau^2)}\right) \quad (27.19)$$

$$\frac{1}{T_2}(\text{CSA}) = \left(\frac{1}{90}\right){\gamma_I}^2{B_0}^2(\Delta\sigma)^2 \times \left(8\tau + \frac{6\tau}{(1+{\omega_I}^2\tau^2)}\right) \quad (27.20)$$

is effective for nuclei with large values $\Delta\sigma$ (^{13}C, ^{15}N, ^{19}F and ^{31}P, for example). As it is seen, at $\omega^2\tau^2 \ll 1$ the relaxation rates become field-dependent and proportional to ${B_0}^2$. This is an experimental test for domination of the CSA mechanism. According to Abragam (1961), the spin-rotation mechanism is remarkable only for small molecules undergoing fast rotation in nonviscous media. This mechanism is valuable when the T_1 times reduce at increasing the temperature.

Generally, different relaxation mechanisms can operate simultaneously. In such cases, contributions that are important for studying the molecular mobility or structural features should be accurately evaluated (Bakhmutov, 2005). It should be emphasized, however, that two “independent” mechanisms can interfere to give cross-relaxation contributions.

27.3.4.2 **Paramagnetic Relaxation**

Fluctuating magnetic fields created by the large magnetic moment of an unpaired electron, $\mu_S = -g_e\beta eS$, can strongly reduce time of nuclear relaxation via dipolar and/or contact interactions. The dipolar mechanism, formulated by Solomon (Solomon, 1955) as controlled by reorientations of electron-nucleus vectors r, is written as (for $S = ½$)

$$\frac{1}{T_1{}^{\mathrm{DD}}} = 0.1\gamma_I{}^2\gamma_S{}^2\hbar^2 r^{-6} \times \left\{ \frac{\tau}{(1+(\omega_I-\omega_S)^2\tau^2)} + \frac{\tau}{(1+\omega_I{}^2\tau^2)} + \frac{\tau}{(1+(\omega_I+\omega_S)^2\tau^2)} \right\} \quad (27.21)$$

$$\frac{1}{T_2{}^{\mathrm{DD}}} = \left(\frac{1}{20}\right)\gamma_I{}^2\gamma_S{}^2\hbar^2\, r^{-6} \times \left\{ 4\tau + \frac{\tau}{(1+(\omega_I-\omega_S)^2\tau^2)} + \frac{3\tau}{(1+\omega_I{}^2\tau^2)} + \frac{6\tau}{(1+\omega_S{}^2\tau^2)} + \frac{6\tau}{(1+(\omega_I+\omega_S)^2\tau^2)} \right\} \quad (27.22)$$

Here, τ is the molecular motion correlation time at rapid molecular motions or the electron relaxation time at slow reorientations. Because the nuclei, situated closer to paramagnetic centers, relax much faster (see the r^{-6} factor), these $T_{1,2}$ times can be rationalized in structural terms.

Bloembergen (1957) has shown that strong FC coupling can lead to this situation when electron-spin flips are directly accompanied by nuclear-spin flips. This mechanism, arising due to delocalization of the unpaired spin density in a nucleus, is given by the following equations:

$$\frac{1}{T_1{}^{\mathrm{CON}}} = \left(\frac{2}{3}\right) S(S+1)\left(\frac{A}{\hbar}\right)^2 \times \frac{(\tau_{E2})}{(1+\omega_I{}^2\tau_{E2}{}^2)} \quad (27.23)$$

$$\frac{1}{T_2{}^{\mathrm{CON}}} = \left(\frac{1}{3}\right) S(S+1)\left(\frac{A}{\hbar}\right)^2 \times \left(\tau_{E1} + \frac{(\tau_{E2})}{(1+\omega_I{}^2\tau_{E2}{}^2)}\right) \quad (27.24)$$

where τ_{E1} and τ_{E2} are the longitudinal and transverse electron-spin relaxation times, respectively. The FC mechanism dominates if under conditions $\omega_S\tau > 1 > \omega_I\tau$ the T_2 time is much shorter than T_1.

27.3.4.3 **Nuclear Relaxation in Solids: Spin Diffusion**

When solids exhibit some kind of molecular motions, then nuclei can relax by all of the above mechanisms. However, the relaxation process is usually nonexponential and described by a stretched exponential, $\exp(-(\tau/T_1)^{\beta})$. Tse and Hartmann (1968) have shown that β of 0.5–0.6 is typical of powder samples. In rigid solids with very restricted motions, nuclear relaxation should be extremely slow. However, even at very small concentrations of paramagnetic impurities, the relaxation rates can increase by the, so-called, spin diffusion, suggested by

Bloembergen (1949). Owing to energy-conserving transitions of spins (mutual spin flips), belonging to dipolar-coupled nuclei, a target spin is "transferred" to a paramagnetic center, where it relaxes effectively by electron-nuclear dipolar interaction. Then, relaxation times are controlled by spin-diffusion coefficients D:

$$\frac{1}{T_1{}^{\mathrm{SD}}} = \left(\frac{1}{3}\right) 8\,\pi NpC^{1/4}D^{3/4} \quad (27.25)$$

$$C = \left(\frac{2}{5}\right) \gamma_I{}^2\gamma_S{}^2\hbar^2\ S(S+1)$$

In turn, D is defined as $D = (\mathbf{M}^{1/2}/30)a^2$, where $C = (2/5)\gamma_I{}^2\gamma_S{}^2\hbar^2\ S(S+1)$, Np is the paramagnetic center density, $\mathbf{M}$ is the second moment of the dipolar internuclear interaction, and a is the internuclear distance. There are two distinguished cases: rapid spin diffusion and diffusion-limited relaxation. $\mathrm{T}_1{}^{\mathrm{SD}}$ relaxation via fast spin-diffusion is always exponential, whereas diffusion-limited relaxation is nonexponential and can be treated by a stretched exponential at $0.5 < \beta < 1$. Kessemeier and Norberg (1967) have shown that the spin-diffusion coefficient D decreases in spinning solids. Thus the $T_{1,2}{}^{\mathrm{SD}}$ times can depend on rates of mechanical rotation. Recently, Grutzer and Dybowski (Grutzer *et al.*, 2001) have found that heavy nuclei, such as ^{207}Pb, can relax via a spin-phonon Raman scattering mechanism.

27.3.4.4 Relaxation Time Measurements

Modern commercial NMR spectrometers provide simple and convenient procedures for measurements of relaxation times. Relaxation times T_1 can be measured by the inversion-recovery (180°-τ-90° *at the varied delay*s τ), saturation-recovery or progressive saturation experiments. To improve the quality of spin inversion, the 180° pulse can be replaced with a composite pulse cluster $90^\circ_\phi\ 240^\circ_\phi\ 90^\circ_\phi$ (see, for example, Harris, 1986). According to Sanders and Hunter (1994), the Hahn-echo (90°-τ-180°-τ-*FID*) or Carr–Purcell pulse sequences (90°x'-τ-180°y'-τ(*first echo*)-τ-180°y'-τ (*second echo*), *and so on*) can be applied to determine the spin–spin relaxation times T_2.

Spin-locking experiments in a rotating coordinate system create conditions providing nuclear magnetization to be aligned along the direction of the applied radio-frequency field. Because this field is smaller by several orders than the external magnetic field B_0, the magnetization decays with a specific time constant, $T_{1\rho}$, determined by significantly slower molecular motions. In solids, where molecular motions are strongly restricted, $T_{1\rho}$ values can differ strongly from T_1 and T_2 times. In contrast, the T_1, T_2, and $T_{1\rho}$ times are practically identical in solutions and nonviscous liquids, where molecular tumbling is fast.

27.4
Spectroscopic NMR Experiments in Liquids

27.4.1
Routine One-Dimensional (1D) NMR spectra

Generally, the routine 1D NMR spectra imply a direct excitation of nuclei with a single-pulse sequence (Figure 27.3). The best sensitivity in such experiments, often requiring a large number of accumulations at long delay times, can be reached when the acting angle of

RF pulses is optimized. In accordance with the Ernst rule (Ernst *et al.*, 1966), at the total RD of 1 s, the angle should be of 8, 25, 33, 68, 86, and 90° if the T_1 time is 100, 10, 4, 1, 0.4, and 0.1 s, respectively.

Rare nuclei, for example, ^{13} C, are often spin–spin coupled by other nuclei, having a high natural abundance, for example, ^{1} H. The ^{13}C lines splitting and their observations are complicated. The NMR spectra of such nuclei are simplified with a broad-band-decoupler technique, which irradiates ^{1}H nuclei during the (RD + AT) time and suppresses, thus, the spin–spin coupling. Besides ^{1}H-decoupling, the signal intensities of carbons attached to protons (for example, groups CH, CH_2 or CH_3) increase due to the nuclear overhauser effect (NOE) (Overhauser, 1953). This effect connects with a perturbation of an equilibrium magnetization of X nuclei at irradiation of A nuclei, when nuclei A and X are coupled by dipole–dipole interaction. Commonly, when signs of γ_A and γ_X are identical, the X intensity increases according to: $I^X\{A\}/I^X{}_0 = 1 + \gamma_A/2\ \gamma_X$. To avoid the NOE, the ^{1}H- decoupler is applied during the AT time only.

Owing to the numerous and well-established relationships "spectrum/ structure," even the routine 1D NMR spectra (multinuclear NMR particularly) are good tools for the determinations of chemical structures. The NOE measurements are also useful in this context. For example, Harris (1986) notes that in a system, consisting of three protons, H(A)···H(B)···H(C), coupled by dipolar interaction, irradiating the H nucleus results in the positive and negative NOE enhancement for H(B) and H(C) nuclei, respectively.

27.4.2
Polarization Transfer in NMR (SPI, INEPT, and DEPT)

When rare nuclei, for example,^{13}C, ^{15}N, ^{29}Si, and so on are coupled by high-γ–nuclei, for example, ^{1}H, their equilibrium population difference can be perturbed by a polarization transfer technique to increase the NMR sensitivity. The physical principles of this technique can be found in the book of Friebolin (1993). Generally, an 180° pulse (selective or nonselective), which inverts the level populations, plays a main role in such experiments, named as signal enhancement by polarization transfer (SPI), insensitive nuclei enhanced by polarization transfer (INEPT) and distortionless enhancement by polarization transfer (DEPT). Among them, the DEPT sequence, shown in Figure 27.5, is most important, particularly for ^{13}C nuclei. Here an evolution time, τ, is calculated as 1/ $2\,^1J$(C−H) and pulse $\theta = 135°$ leads to an ^{1}H-decoupled ^{13}C NMR spectrum with "negative" signals for CH_2 groups and "positive" signals for CH and CH_3 groups.

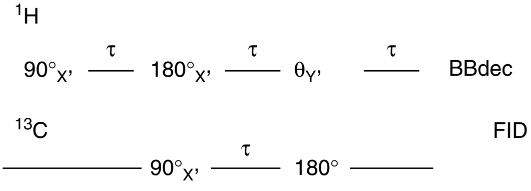

Fig. 27.5 Presentation of the DEPT pulse sequence.

27.4.3
Two-Dimensional (2D) NMR Spectra

The pulse sequences, applied for the 2D NMR experiments, are various and are dependent on target tasks. Many of them, for example, the 2D heteronuclear *J*-resolved ^{13}C NMR spectra, the 2D homonuclear *J*-resolved ^{1}H NMR spectra, the 2D heteronuclear correlated NMR spectra, the 2D relayed NMR spectra, the 2D exchange NMR spectra, and the 2D INADEQUATE NMR spectra can be found in the book of Friebolin (1993). The more complex pulse trains are applied for NMR studies of biomolecules in solutions or the solid state (Duel, 2002). General principles of 2D NMR can be demonstrated by one of the simplest ^{1}H NMR experiments, ^{1}H–^{1}H COSY (correlated spectroscopy), shown in Figure 27.6.

After the first 90° pulse, a nuclear system is developed by proton–proton spin–spin coupling during time t_1. Thus, the first set of time-dependent data, $f(t_2)$, can be recorded after the action of the second RFP during time t_2. The $f(t_1)$ sets are collected at various t_1 values. Then, a double FT, performed for domains t_1 and t_2, leads to two frequency-axes. The resulting spectrum, shown in Figure 27.7, is a square plot, where the diagonal corresponds to the 1D ^{1}H NMR spectrum and cross peaks appear due to scalar spin–spin coupling. Similarly, one can perform a 2D NOE experiment, NOESY, (Figure 27.7) which is a good tool to establish spatial proximity of nuclei or an exchange process between them.

27.4.4
Chemical Exchanges: the NMR Time Scale

Besides nuclear relaxation and spin–spin coupling, reversible exchange processes between distinguishable spin states can also affect line shapes in NMR. These effects are particularly strong when life times (τ) of these states are in a range from 10^{-4} to 10^{1} s, named as the NMR time scale.

Commonly, a chemical exchange between spins *A* and *X* perturbs shapes of their resonances, if the exchange frequency,ν_E, is comparable to chemical shift difference ($\nu_A - \nu_X$), expressed in Hertz. Then, the NMR spectra show a typical temperature evolution (Figure 27.8): the signals broaden and coalesce. A fast exchange, $\nu_E \gg (\nu_A - \nu_X)$, results in a single resonance. If spin states *A* and *X* are equally populated, this resonance appears at frequency $\nu = (\nu^A + \nu^X)/2$.

Generally, spin states can have different populations, they can belong to the same molecule (an intramolecular exchange) or different molecules (an intermolecular exchange) and finally,

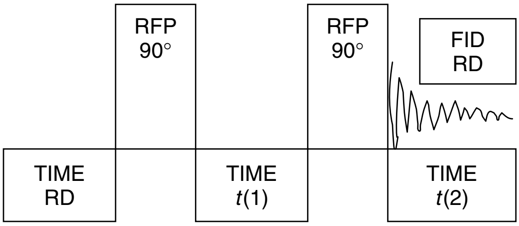

Fig. 27.6 The pulse sequence for a homonuclear proton–proton correlation (COSY) NMR experiment: the second pulse can be 90 or 45°.

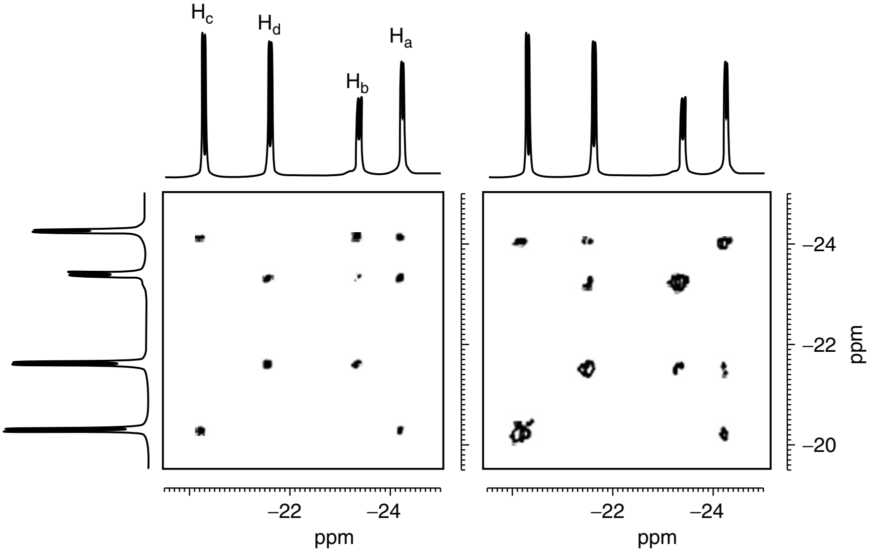

Fig. 27.7 A high-field region of the ^{1}H COSY (left) and NOESY (right) NMR spectrum of the complex $[Ir_2(\mu\text{-}H^a)(\mu\text{-}Pz)_2(H^b)(H^c)(H^d)(NCCH_3)(PPr^i{}_3)_2]$ in C_6D_6: pairs of nuclei H^a/H^c, H^a/H^b and H^b/H^d are spin coupled, while pairs H^a/H^c and H^b/H^d are spatially approximated.

multicenter exchanges are also possible. Because the NMR line shapes in liquids are well established, exchange rates are accurately determined by a full line-shape-computer analysis, which can be found in NMR books cited in this chapter.

At slow exchanges, $\nu_E < \nu_A - \nu_X$, the NMR signals are well separated and remain narrow. However, relaxation times of the exchanging spin states can be averaged partially. Then, the FID's collected by the inversion-recovery experiments will depend on life times of these spin states which, thus, can be determined. Slow exchanges in ^{1}H NMR spectra can be easily established by saturation transfer experiments: irradiation of X nuclei leads to the saturation of the A resonance. For nuclei other than protons, 2D NOESY NMR, known as *two-dimensional exchange spectroscopy* (*EXSY*), is more preferable. Some examples of applications of the above techniques can be found in the books of Friebolin (1993), Harris (1986), and Bakhmutov (2005). It should be emphasized that the effects of chemical exchanges are particularly strong when one of the exchanging states is paramagnetic. In such cases, quantitative characterizations are reached by relaxation time measurements.

27.4.5
Translational Molecular Diffusion via NMR

The influence of magnetic-field inhomogeneity on forming the spin echo has been noted by Hahn (1950). If a static field gradient G is actually present, then a coherent transverse magnetization, created by the first 90° pulse in the Hahn-echo pulse sequence ($90°\text{-}G(\tau)\text{-}180°\text{-}G(2\tau)\text{-}FID$), undergoes an evolution in

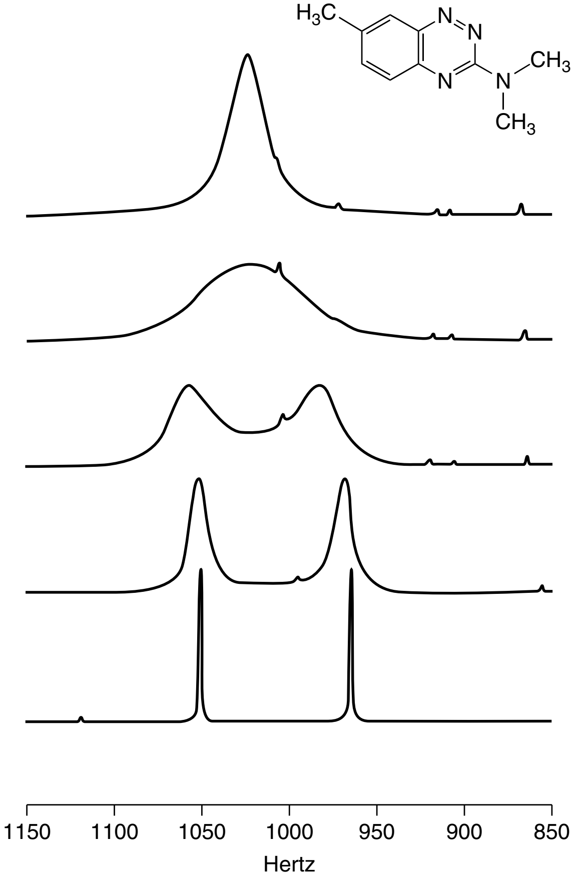

Fig. 27.8 Temperature evolution of the NCH_3 lines in the variable-temperature ^1H NMR spectra typical of a rotation around C—N bond.

this gradient. After a dephasing time, τ, the second 180° pulse inverts the sign of the net accumulated phase and the spins precess at the Larmor frequencies, reducing their accumulated phases. When the rephasing and dephasing rates are equal, then the magnetization reaches its coherent initial state after time 2τ, forming, thus, a spin echo. Then, the spin echo decays due to T_1/T_2 relaxation and additional factors are directly related to the molecular displacement on the time scale of the spin-echo experiments. At well-defined field gradients, the time correlation functions, describing the spin-echo decays, can be applied for studies of translational molecular diffusion.

Technically, these NMR experiments can be performed with the static-field gradients (SFG) or the pulsed-field gradients (PFG) suggested by McCall *et al.* (1963). As noted by Geil (1998), a combination of large static field gradients with short radio-frequency-pulse-spacings provides measurements of very small diffusion coefficients (10^{-15} m^2 s^{-1}) at a high spatial resolution (10 nm).

27.4.6
Isomeric and Conformational Analysis in Solutions by NMR

One of the key areas of NMR applications in solutions is the molecular conformational/configurational analysis. Because generally the analysis of cyclic molecules involves various NMR methods, such as the variable-temperature NMR spectra, the NOESY or other 2D NMR experiments, T_1 measurements, and determinations of the spin–spin coupling constants and their calculations by a convenient computer program, this aspect of the NMR spectrometry can also be a good illustration of its practical application. A methodology, typical of these studies, is demonstrated here by five-membered cycles 1 and 2, depicted in Scheme 27.1 (Guizado-Rodriguez *et al.*, 2001).

Initial structural formulations of these compounds can be easily carried out on the basis of the ^{1}H, ^{13}C, ^{11}B, ^{19}F, and 2D ^{1}H–^{13}C HETCOR NMR spectra. However, *a priori*, the cycle molecules can exist as trans-(1*A*, 2*A*) and cis-isomers (1*B*, 2*B*) and their cycles can take an envelope (*E*) or twist (*T*) conformations. In addition, usually five-membered cycles undergo fast pseudorotation with insignificant transformation energies and therefore the envelope and twist conformations can be equilibrated leading to observations of the average room-temperature (RT) NMR spectra. According to the RT multinuclear NMR spectra, recorded in C_6D_6, compound 1 gives two isomeric products, 1*A* and 1*B*, in a ratio of 9 : 1. Protons 2 and 2′ in the RT ^{1}H NMR spectrum of the major product are equivalent, whereas protons 5(4′) and 5′(4) are nonequivalent to show a complicated pattern in Figure 27.9. Simulation procedures for the AA′BB′ pattern show that the multiplet can be calculated when the geminal ^{1}H–^{1}H couplings are assumed to be negative and the *J*(5–4′) and *J*(5′–4) constants take positive but different values: 9.5 and 2.5 Hz, respectively. In terms of the modified Karplus equation (Haasnoot *et al.*, 1980), the corresponding dihedral angles should be strongly different. The NOESY ^{1}H NMR spectra of 1*A* in toluene-d_8 recorded at room-temperature and −90°C reveal cross peaks 2(2′)-CH_3, 5′(4)-CH_3 and 5–5′, showing spatial proximity of the above protons. Protons 2 and 2′ remain equivalent even at low temperatures. In addition, calculations of the 5(4′)–5′ (4) patterns, observed in toluene-d_8 at 25 and −90 °C, lead to

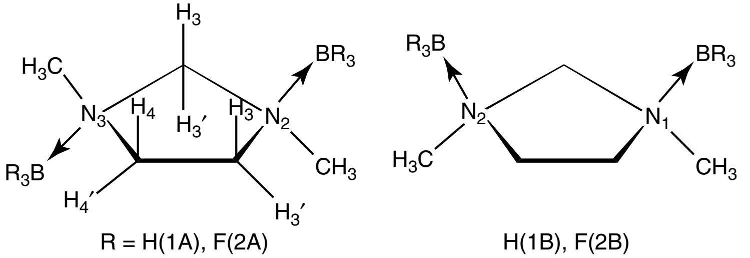

Scheme 27.1

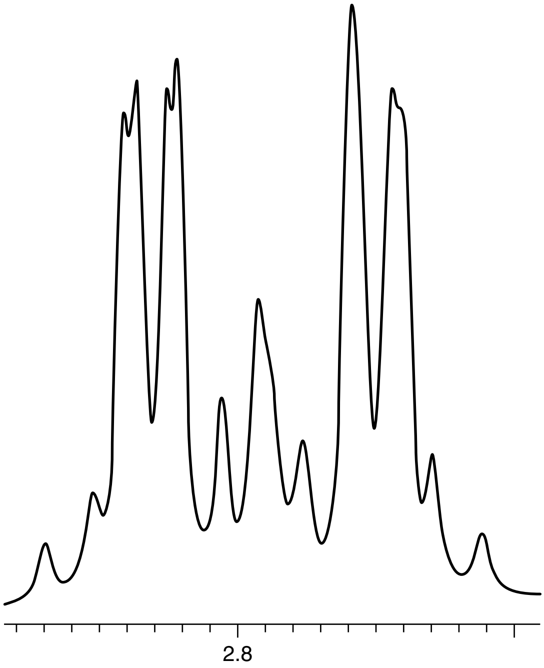

Fig. 27.9 Pattern AA′BB′ observed for protons 5(4′) and 5′(4) in the RT ^{1}H NMR spectrum of 1*A* in C_6D_6.

identical results: $^2J(5-5') = -10.5$ Hz, $^3J(5-4') = 9.5-10.0$ Hz, $^3J(5'-4) = 2.0-3.0$ Hz, $^3J(5-4) = 7.0$ Hz, $^3J(5'-4') = 7.0$ Hz, and $^2J(4-4') = -10.5$ Hz. Thus, only a trans BH_3/BH_3 arrangement is possible in 1*A* and its cycle takes twist conformation $T-1$ in Scheme 27.2, which is preferable even at room temperature. Compound 2 shows identical spatial properties. The same (or similar) methodology can be successfully used in the studies of other molecules.

27.5
NMR Spectroscopic Experiments in the Solid State

27.5.1
Wide-Line NMR

Resonance lines in NMR spectra of liquids or solutions are sharp due to the fast high-amplitude molecular motions. For example, line-width of a ^{1}H resonance in liquid water is as small as 0.1 Hz. In contrast, the resonance of ice is dramatically broadened up to $\sim10^5$ Hz. According to

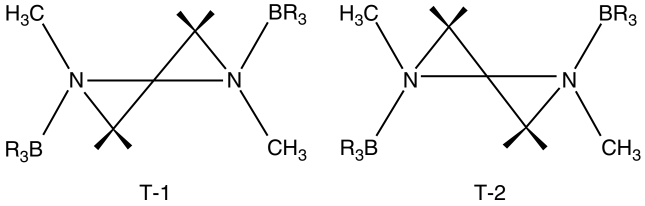

Scheme 27.2

Andrew and Szczesniak, 1995, a primary source of this broadening is magnetic dipolar proton–proton interactions, governed by a factor $(3\ \cos^2\theta - 1)/r^3$, where θ is the angle between the external magnetic field and internuclear vector r. It is obvious that the broadening factor is particularly large for powder samples. A single crystal sample can be oriented in the external magnetic field and its NMR spectrum is analyzed better. For example, a single crystal of $CaSO_4 \cdot 2\ H_2O$ with the relatively isolated protons shows the so-called Pake doublet providing determination of the proton–proton distance (Pake, 1948). Because dipolar interactions depend on the nature of nuclei, line widths of their resonances change drastically. For example, $^{29}Si–^{29}Si$ dipolar coupling produces line widths of 2–3 kHz only. Similarly, hetero-nuclear dipolar and CSA interactions also lead to wide lines in solid NMR. Finally, quadrupolar interactions exceed dipolar interactions by a factor e^2Q/μ^2, where eQ and μ are electric quadrupole moment and nuclear magnetic moment, respectively. All of these effects, particularly important for powder samples, give rise to the solid-state NMR spectra, where directional information "nucleus-nucleus" is lost and line widths mask the presence of chemically different nuclei.

Nevertheless, wide-line NMR can provide structural conclusions based on an approach suggested by Van Vleck (1948). In the limits of the approach, the dipolar second moment (or square width) is calculated in terms of the sum $\Sigma(3\ \cos^2\theta_{jk} - 1)^2/r_{jk}{}^6$ over all nuclei pairs in crystals or the sum $\boldsymbol{\Sigma} r_{jk}{}^6$ in polycrystalline samples. Because of the $r_{jk}{}^6$ factor, nearest dipole–dipole contacts dominate and thus structural information about molecules, groups, or ions can be obtained by comparing the experimental and calculated NMR spectra.

27.5.1.1 Dead Time in Wide-Line NMR

When a magnetization, created by RFP, decays rapidly with respect to the dead time of a spectrometer, FT converts the FID into an NMR spectrum where signal intensities are remarkably reduced and the base line shows strong distortions. This problem, particularly important for solid-state NMR, can be minimized by the applications of pulse sequences Hahn-echo ($90°_X$-τ-$180°_Y$-τ-FID) or solid echo ($90°_X$-τ-$90°_Y$-τ-FID), where τ is echo-delay time. According to Duel (2002), the FIDs can be collected immediately after 180° pulses. For spinning samples, the experiments require echo-delays synchronized with rotor periods.

In paramagnetic samples, resonance lines can be too broad and even unobservable in regular NMR spectra. In such cases, as noted by Mali *et al.* (2005), carrier frequencies in the Hahn-echo experiments can be varied and the resulting Hahn-echo NMR spectra can be summed up to show these wide lines.

27.5.1.2 Relaxation Experiments in Solids

Despite the relatively poor structural information, the wide-line NMR technique is successfully applied for the studies of molecular motions and phase transitions in solids by relaxation time measurements and/or a line-width analysis, briefly reviewed by Andrew and Szczesniak, 1995. The experiments can be performed in a very large temperature range, which is sufficient to characterize the molecular dynamics in materials of all kinds: amino acids, polypeptides, proteins, polymers, rubber, metabolites and steroids, and so on. In solid proteins, for example, it is possible to identify and characterize energetically a number of motions contributing to relaxation rates such as methyl-group reorientations, segmental motions, side-chain motions, and bound-water reorientations. It should be emphasized that solid-state NMR measurements, including T_1 and T_{1D} times and also the rotating-frame relaxation time, $T_{1\rho}$, is sensitive to motions covering the range of frequencies from 10^{-1} to 10^{11} Hz.

27.5.2
High-Resolution Solid-State NMR

At an angle θ of 54.4°, called as *magic angle* by Andrew, factor $(3\ \cos^2\theta - 1)/r^3$, characterizing dipole–dipole interaction, transforms to 0. It has been shown that the broadening effects, observed in static solids, can be reduced by their spinning when the given samples are situated at this angle relative to the external magnetic field. The narrowing effects will be more pronounced at spinning rates close or larger than line widths in static NMR. This principle is basic in the magic-angle-spinning (MAS) NMR spectroscopy leading to high resolution in the solid-state spectra. In practice, the collected spectra show isotropic resonances accompanied by spinning sidebands, which can be quite intense. However, positions of these bands can be changed by rotation at different rates and thus isotropic chemical shifts can be determined. In addition, there are special pulse sequences, which are capable of suppressing the sidebands. According to Duel (2002), one of such scheme is the TOSS pulse sequence consisting of four 180° pulses in a period prior to acquisition.

At intermediate spinning rates ν_R with respect to static line widths $\Delta\nu$ (for example, $\nu_R \sim \Delta\nu/4$) and domination of homonuclear dipolar coupling, the lines observed in MAS NMR spectra are still broadened due to the remaining time-dependent terms. In contrast, the MAS NMR spectra of molecular systems with dominant hetero-dipolar and CSA interactions show sharp lines even at intermediate spinning rates. According to Strub *et al.* (1983), distributions of sidebands in such cases can be used to determine magnetic screening tensors.

Nowadays, the modern commercial NMR spectrometers are equipped with MAS probes, where the spinning rates of 30–50 kHz are routinely achieved. Under these conditions, broadenings caused by the magnetic dipolar interactions, chemical shift anisotropy, Knight-shifts, pseudodipolar interactions, and quadrupolar interactions are canceled and the resulting spectra are determined by the isotropic chemical shifts and spin–spin coupling constants as in liquids and solutions.

27.5.2.1 High-Power Decoupling and Double Resonance

Typically, even in relatively simple molecules, NMR detections of rare nuclei, such as ^{13}C or ^{15}N, and so on, are seriously problematic because of the very strong dipolar interactions at short bond distances $^1H{-}^{13}C$ or $^1H{-}^{15}N$. Generally, these interactions are suppressed by a high-power decoupling technique producing ^1H irradiation when the ^{13}C or ^{15}N NMR FIDs are recorded. The decoupling can be realized by a continuous irradiation at a power of 100–1000 W or by special pulse sequences operating at proton frequency. Among them, sequences Two pulse phase modulation (TPPM), Waugh-Huber-Haeberlen (WAHUHA), and MREV-8 are most applicable (Duel, 2002).

Although, in rigid solids, the homonuclear dipolar interaction $^1H-^1H$ is particularly strong, lines in the ^1H MAS NMR spectra, recorded even at high spinning rates, are still broadened due to remaining dipolar interaction. However, such broadenings can be weakened or completely removed by the application of combined rotation and multiple pulse spectroscopy (CRAMPS), pioneered by Gerstein (1981). The CRAMPS combines sample spinning and multiple-pulse-decoupling sequences (for example, MREV-8).

27.5.3
Cross Polarization Technique

This technique helps strongly to improve a signal/noise ratio in the NMR spectra of ^{13}C, ^{15}N, and other nuclei with long relaxation times, low natural abundance, and low γ values. The task of a pulse sequence

$$^1\text{H } 90^\circ_X \text{ - contact pulse}_{-Y}\ ^1\text{H decoupling}$$

$$^{13}\text{C - contact pulse}_{-Y}\text{ -}^{13}\text{C FID} \qquad (27.26)$$

is a transfer magnetization from the ^1H spins with a high natural abundance to the ^{13}C (or ^{15}N) spins via the dipolar coupling between them. The amplitudes of two contact pulses, operating at proton and carbon frequencies, should be adjusted to achieve the Hartmann–Hahn conditions (Hartmann and Hahn, 1962):

$$\gamma_H B_1(^1\text{H}) = \gamma_C B_1(^{13}\text{C}) \qquad (27.27)$$

Under these conditions, the energy gaps between the ^1H and ^{13}C rotating-frame spin states are equal and the transition requiring energy on ^1H spin can be exactly compensated by the transition releasing energy on ^{13}C spin. This pulse sequence in combination with the MAS technique leads to rapid accumulations of the ^{13}C NMR spectra because the recycling delay is decided by the ^1H (and not ^{13}C) relaxation.

27.5.4
Internuclear Distances from the Solid-State NMR Data

When molecular systems are amorphous or insoluble in common solvents, solid state NMR is the most efficient tool in their structural studies. Traditional NMR applications, based on the solid-state 1D and 2D NMR spectra, can be found in Duel (2002). Technically, collections of the NMR spectra in the solid state and solutions remarkably differ but their interpretations are very similar. It should be noted that most chemists are satisfied with obtaining such data. The other way of structural characterizations by NMR techniques is the high-precision measurement of interatomic distances, illustrated in this section.

Theoretically, information about distances between nuclei can be evaluated from the magnitudes of long-range dipole–dipole interactions. The classical Pake's powder spectrum, (Pake, 1948), expected for two nuclei in a static sample, is shown in Figure 27.10. Here, the shape of the resonance is formed by the superposition of two subspectra (dashed lines) and the distance between the two singularities corresponds to the dipolar coupling constant ω_D. Recalculations of ω_D gives finally a distance between the nuclei. A weak point in this simple approach is obviously connected with a low NMR

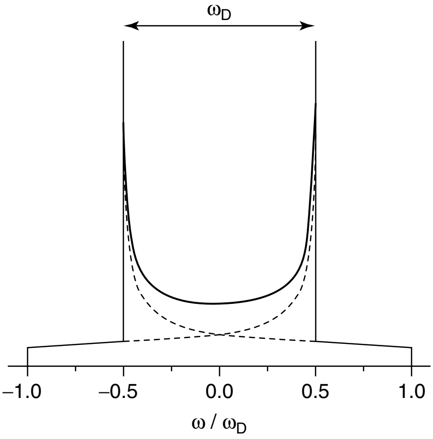

Fig. 27.10 The Pack dipolar powder spectrum. (Reproduced with permission from Lee, J-A and Khitrin, AK (2008) *Concepts Magn. Reson. Part A*, **32A**, 56.)

sensitivity particularly important for rare nuclei. Development of less direct NMR techniques, such as *SLF, PISEMA PITANSEMA, REDOR, TEDOR*, and so on, simplifies the determinations of dipolar couplings and hence internuclear distances. Some methods, their physical basis, and applications can be found, for example, in the book of Duel (2002). Lee and Khitrin (2008) have suggested a technique, specially developed for static samples, which is based on a two-dimensional single-echo pulse sequence, where adiabatic cross-polarization increases sensitivity. This technique, applied for isotopically labeled molecules, for example, glycine, provides the accuracy in measuring internuclear distances C—C and C—N, which is well compared with that in the X-ray and neutron-diffraction methods. For example, distance C_α—N is determined as 1.496 ± 0.002 by NMR and 1.476 ± 0.001 by ND.

27.5.5
Zero-Field Solid-State NMR Experiments

An initial nuclear polarization, needed for detecting NMR signals, is generally created by the strong external magnetic field. On the other hand, nuclei can be polarized by internal magnetic fields, which appear in samples due to their nature. For example, such a zero-field NMR spectrum has been observed by Reif and Purcell (1953) in solid hydrogen due to very strong proton–proton dipolar interactions. Similarly, even in the absence of an external magnetic field, the internal magnetic fields polarize the ^{57}Fe and ^{59}Co nuclei in ferromagnetic iron and cobalt, again resulting in the zero-field NMR spectra.

Other physical principles, considered by Andrew and Szczesniak (1995), are used in the zero-field NMR experiments directed to enhancements of resolution in the NMR spectra of polycrystalline samples. In such experiments, an initial nuclear polarization is created in a sample when it is placed into a strong external magnetic field. This polarized sample is transferred mechanically to an intermediate magnetic field, which, however, is still remarkably larger than local fields existing in the sample due to dipolar interactions. Then, this intermediate field is rapidly switched off and the nuclear spins undergo an evolution, during time τ_1, initiated obviously by the dipolar forces. After this time, the intermediate field is turned on and the sample is again placed into the high magnetic field, where a high-field NMR signal is detected as $I(\tau_1)$. FT of $I(\tau_1)$ gives, thus, a zero-Field NMR spectrum. In other words, this experiment leads to the high-magnetic-field NMR spectrum, recorded, at a high sensitivity, after an evolution of the nuclear spin system in the absence of the magnetic field. Because during the evolution period all directions are equivalent, numerous crystallites in polycrystalline samples will behave identically and the resulting NMR spectra will show sharp lines. Under these conditions even powder samples exhibit dipolar couplings with high resolution, usually reached only in field-oriented single crystals. Andrew and Szczesniak, 1995 note the obvious limitations: the zero-field technique can be successfully used when the spin-lattice relaxation time, T_1, in a low magnetic field is long enough to complete the experimental cycle. Similar zero-Field experiments can be performed for spinning samples. However, acting pulses and rotation rates should be synchronized in such cases.

27.6
Sensitivity Enhancement by Dynamic Nuclear Polarization (DNP)

Applications of the solution and solid-state NMR spectroscopy for many fields of chemistry, biology, and materials science often require a very high sensitivity in NMR spectra. When probing rare nuclei, for example ^{13}C, this is particularly critical when using the techniques considered in the previous sections is not enough. It should be recalled that the sensitivity depends on the natural initial nuclear polarization, which can be enhanced by the specially developed approaches based on optical pumping (Walker and Happer, 1997), parahydrogen-induced-polarization (Bargon and Natterer, 1977), and dynamic nuclear polarization (DNP) (Wind *et al.*, 1985). For example, Bifone *et al.* (1996) have found that optical pumping with laser light can enhance the nuclear polarization of gaseous xenon by 5 orders of magnitude. Among different techniques, the DNP seems to be most important and applicable to solids and solutions containing species with unpaired electrons (paramagnetic metal ions, organic groups, and so on). It is obvious that the systems investigated can be specially doped by such paramagnetic species.

Generally, an electron-nucleus system is described via the Hamiltonian:

$$H = -\omega_e S_Z - \omega_I I_Z + H_{\mathrm{ee}} + H_{\mathrm{en}} + H_{\mathrm{nn}} \qquad (27.28)$$

where the first two terms correspond to the Zeeman's interactions of electron and nuclear spins, respectively, H_{ee} and H_{nn} describe electron–electron and nucleus–nucleus coupling, and H_{en} describes the hyperfine electron-nucleus interaction. In the context of DNP generation, the latter term plays a main role.

Irradiating the samples at (or near) the electron Larmor frequency, ω_e, can lead to simultaneous electron-nucleus spin flips due to electron-nuclear coupling H_{en}. In turn, these flips enhance the absolute value of the nuclear polarization via three mechanisms: an Overhauser effect, a solid effect, and a thermal mixing effect, described in the review of Wind *et al.* (1985). The enhancement factors observed in the nuclear polarization depend on the number of unpaired electrons, the ESR line widths, the nuclear relaxation rates in the absence of electrons, and the amplitude of the microwave field, applied to the sample. Generally, in solids, the enhancement factors can reach 660 and 2600 for ^{1}H and ^{13}C nuclei, respectively.

Technically, the DNP experiments require a special design of an NMR spectrometer and particularly an NMR probe, which should be equipped with a coil that can be double-tuned to ^{13}C and ^{1}H frequencies, for example, and a horn antenna to transmit the microwaves. The simplest enhancement experiments with ^{13}C nuclei imply their direct polarization. A proton polarization via DNP and its transfer to the ^{13}C spins via ^{1}H$-$^{13}C cross-polarization is also possible.

Addenkjaer-Larsen *et al.* (2003) have applied the similar principle for the NMR spectra in liquids to increase the signal/noise ratio by a factor of $>10^3$. This method is based on a low-temperature high-field nuclear polarization performed by a DNP polarizer where a frozen sample is irradiated with microwaves. This instrument is equipped with a superconducting magnet, a liquid-helium cooled sample space, and a special injection tool inside the DNP magnet. This injection tool provides a rapid dissolving of the solid sample in a suitable solvent and thus the resulting solution contains molecules with hyperpolarized nuclear spins. Then the hyperpolarized solution is rapidly transferred into a high-resolution NMR spectrometer where an enhanced NMR signal is registered by a convenient procedure. This powerful method can be successfully applied for different DNP-enhanced NMR experiments *in vitro* and *in vivo*.

27.7
Imaging NMR

High resolution in the NMR spectra, very important for a detailed structural analysis, can be obtained when the strong external magnetic fields are maximally homogenous. Regularly, such fields are created by magnets with bores of a few centimeters, capable of keeping the resonance condition (Eq. (27.4)) in each point through the volume of a sample investigated. In large and heterogeneous samples, requiring obviously wider bore magnets, the resonance condition cannot be satisfied throughout the whole object. Nevertheless, in such cases, the magnetic field can be homogenized only in a small volume of the object and thus a high-resolved NMR signal can be still recorded. Then, this small volume can be "spatially shifted" to collect the next signal. Repeating these experiments finally leads to the distribution of detected nuclei through the volume of the object. This principle is basic in the imaging NMR spectroscopy. The first imaging NMR experiments have been performed more than 30 years ago by the research group of Lauterbur (1973) and resulted in a special design of magnets, NMR probes and also methods for computer treatments of the collected data to construct the image.

Among the various modifications, one of the central places belongs to the variable-field-gradient technique, which focuses the magnetic filed. This technique defines the volume selectivity, important in the practical NMR imaging, and hence the spatial resolution depending on object sizes. At present, the highest spatial resolution, which can be obtained in small objects, corresponds to an element volume of 10^{-5} mm^3.

Because the time needed for the collection of NMR data from such a small element volume significantly increases, the experiments should be highly sensitive. Therefore, ^{1}H NMR plays a main role in the imaging spectroscopy particularly when the objects of studies are living systems. Owing to the discovery that relaxation times ^{1}H T_1 and ^{1}H T_2 in water molecules depend on the characteristics of their bonding, the development of the imaging technique was strongly accelerated to result in magnetic resonance tomography. As it is emphasized by Minati and Weglarz (2007), Echo-imaging, based on the ^{1}H T_2 time measurements by spin-echo sequences, leads to quality of the T_2 images well comparable to that obtained by X-ray tomography.

It is known that at a given gradient strength, a wider NMR spectrum produces a poorer image resolution. Since, for example, ^{1}H NMR spectra in static solids are generally broader than in liquids by a factor of $10^5 - 10^6$, the solid-state imaging experiments require the much stronger gradients focusing the field (Andrew and Szczesniak, 1995). Therefore, solid-state NMR imaging includes the MAS technique and the multipulse coherent averaging technique or their combination reducing the dipolar, CSA, and quadrupolar interactions. However, again, spinning rates and the field-gradient pulses should be synchronized.

References

Abragam, A (1961) *Principles of Nuclear Magnetism*, Oxford University Press, Oxford.

Addenkjaer-Larsen, J.H, Fridlund, B, Gram, A, Hansson, G, Hansson, L, Lerche, M.H, Servin, R, Thanong, M, and Goldman, K (2003) *PNAS*, **100**, 10158–10163.

Andrew, E.R and Szczesniak, E (1995) *Progr. Nucl. Magn. Reson. Spectr.*, **28**, 11–36.

Bakhmutov, V.I (2005) *Practical NMR Relaxation for Chemists*, John Wiley & Sons, Ltd, Chichester.

Bargon, J and Natterer, J (1977) *Progr. Nucl. Magn. Reson. Spectr.*, **31**, 293–315.

Bifone, A, Song, Y.O, Seydoux, R, Taylor, R.E, Goodson, B.M, Pietrass, T, Budinger, T.F, and Navon, G, Pines, A(1996) *Proc. Natl. Acad. Sci. USA*, **93**, 12932–12936.

Bloch, F, Hansen and W.W, Packard, M (1946) *Phys. Rev.*, **70**, 474–485.

Bloembergen, N, Purcell, and E.M, Pound, R.V (1948) *Phys. Rev.*, **73**, 679–712.

Bloembergen, N (1957) *J. Chem. Phys.*, **27**, 572–573.

Bloembergen, N (1949) *Physica*, **15**, 386–426.

Contreras, R.H and Peralta, J.E (2000) *Progr. Nucl. Magn. Reson. Spec.*, **37**, 321–425.

Duel, M.J (2002) *Solid-State NMR Spectroscopy: Principles and Applications*, Blackwell Sciences.

Ernst, R.R and Anderson, W.A (1966) *Rev. Sci. Instr.*, **37**, 93–102.

Friebolin, H (1993) *Basic One- and Two-Dimensional NMR Spectroscopy*, Weinheim, VCH.

Geil, B (1998) *Concepts Magn. Reson.*, **10**, 299–321.

Gerstein, B.C (1981) *Philos. Trans. R. Soc. Lond. Ser. A*, **299**, 521–546.

Gorter, C.J and Broer, L.J.F (1942) *Physica*, **9**, 591–596.

Grutzer, J.B, Stewart, K.W, Wasilishen, R.E, Lurnsden, M.D, Dybowski, C, and Beckmann, P.A (2001) *J. Am. Chem. Soc.*, **123**, 7094–7100.

Guizado-Rodriguez, M, Ariza-Castolo, A, Merino, G, Vela, A, Noth, H, Bakhmutov, and V.I, Contreras, R (2001) *J. Am. Chem. Soc.*, **123**, 9144–9152.

Haasnoot, C.A.G, Leeau, F.A.M, Altona, C (1980) *Tetrahedron*, **36**, 2783–2785.

Hahn, E.L (1950) *Phys. Rev.*, **80**, 580–594.

Harris, R.K (1986) *Nuclear Magnetic Resonance Spectroscopy: A Physicochemical View*, The Bath Press, Avon.

Hartmann, S.R and Hahn E.L (1962) *Phys. Rev.*, **128**, 2042–2053.

Jameson, C.J and Gutowsky, H.S (1964) *J. Chem. Phys.*, **40**, 1714–1724.

Kessemeier, H and Norberg, R.E (1967) *Phys. Rev.*, **155**, 321–337.
Kisman, K.E and Armstrong, R.L (1974) *Rev. Sci. Instr.*, **45**, 1159–1163.
Kubo, A, Spaniol, T.P, and Terao, T (1998) *J. Magn. Reson.*, **133**, 330–340.
Kubo, R and Tomita, H (1954) *J. Phys. Soc. Jpn.*, **9**, 888–919.
Lauterbur, P.C (1973) *Nature*, **242**, 190–193.
Lee, J.-A and Khitrin, A.K (2008) *Concepts Magn. Reson. A*, **32A**, 56–67.
Mali, G, Ristic, A and Kaucic, V (2005) *J. Phys. Chem. B.*, **109**, 10711–10716.
McCall, D.W, Douglass, D.C, and Anderson, E.W (1963) *Ber. Bunsenes. Physik. Chem.*, **67**, 336–340.
McConnel, H.M and Chesnut, D.B (1958) *J. Chem. Phys.*, **28**, 107–117.
Minati, L and Weglarz, W.P (2007) *Concepts Magn. Reson. A*, **30A**, 278–307.
Overhauser, A.W (1953) *Phys. Rev.*, **92**, 411–415.
Pake, G.E (1948) *J. Chem. Phys.*, **16**, 327–336.
Poople, C.P (1983) *Electron Spin Resonance*, John Wiley & Sons, Ltd, New York.
Proktor, W.G and Yu, F.C (1950) *Phys. Rev.*, **77**, 717–717.
Purcell, E.M, Torrey, H.C, and Pound, R.V (1946) *Phys. Rev.*, **69**, 37–38.
Reif, F and Purcell, E.M (1953) *Phys. Rev.*, **91**, 631–641.
Sanders, J.KM and Hunter, B.K (1994) *Modern NMR Spectroscopy: A Guide for Chemists*,Oxford University Press, Oxford.
Solomon, I (1955) *Phys. Rev.*, **99**, 559–565.
Strub, H, Beeler, A.J, Grant, D.M, Michel, J, Cutts, P.W, and Zilm, K.W (1983) *J. Am. Chem. Soc.*, **105**, 3333–3334.
Tse, D and Hartmann S.R (1968) *Phys. Rev. Lett.*, **21**, 511–514.
Van Vleck, J.H (1948) *Phys. Rev.*, **74**, 1168–1183.
Walker, T.G and Happer, W (1997) *Rev. Mod. Phys.*, **69**, 629–642.
Wind, R.A, Duijvestun, M.J, Van Der Lugt, C, Manenschijn, A, and Vriend, J (1985) *Prog. NMR Spectrosc.*, **17**, 33–67.

Further Reading

In addition to the very limited number of references cited, advances in magnetic resonance spectroscopy can be found in Malcolm Levitt (2001) *Spin Dynamics: Basics of Nuclear Magnetic Resonance*, John Wiley & Sons, Ltd, Chichester and James Keeler (2005) *Understanding NMR Spectroscopy*, John Wiley & Sons, Ltd, Chichester and also within the periodicals: *Journal of Magnetic Resonance, Magnetic Resonance in Chem*istry, *Magnetic Resonance in Medicine, Progress in NMR spectroscopy, Concepts in Magnetic Resonance, Journal of Magnetic Resonance Imaging* and *the NMR Encyclopedia*.

MAGNETIC STORAGE OF INFORMATION

MARK H. KRYDER, *Data Storage Systems Center, Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A.*

Introduction 95
1. **Magnetic Recording** 95
1.1 Fundamentals of Magnetic Recording 96
1.2 The Magnetic Recording Process 96
1.3 The Magnetic-Recording Readback Process 98
1.4 Magnetic Recording Media 98
1.4.1 Particulate Recording Media 99
1.4.2 Thin-Film Recording Media 99
1.5 Magnetic-Recording Heads 100
1.5.1 Ferrite Heads 100
1.5.2 Metal-in-Gap Heads 101
1.5.3 Thin-Film Heads 101
1.5.4 Magnetoresistive Heads 102
2. **Magneto-optic Recording** 102
2.1 Fundamentals of Magneto-optic Recording 103
2.2 Magneto-optic Recording Media 104
2.3 Magneto-optical Recording Heads 105
2.4 Direct Overwrite 107
3. **Conclusions** 108
3.1 Acknowledgements 108
Glossary 108
Works Cited 109
Further Reading 110

INTRODUCTION

Storage has of late been the fastest growing segment of the computer hardware industry. The overwhelming majority of storage products today are based on magnetic recording technology. Floppy disk drives, rigid disk drives, and tape drives all enjoy multibillion dollar markets for computer data storage. Magnetic bubble memories, which attracted much attention in the 1970s, have made it to the marketplace, but only in niche markets where the customer is willing to pay a relatively high cost per bit for a nonvolatile solid-state memory with no moving parts. Magneto-optic recording, after many years of research and development, has also been introduced into the marketplace and is finding significant applications, although its market is still orders of magnitude smaller than that of magnetic recording products. In contrast to some early expectations, the market acceptance of magneto-optic recording is not based primarily on its high areal density, but rather on the fact it offers a removable high-capacity storage medium in a disk format.

Future markets for both magnetic and magneto-optic recording are expected to grow substantially as storage densities increase by orders of magnitude and data rates and access times also improve. In this article the basic principles, state of the art, and future trends for these two recording technologies are described.

1. MAGNETIC RECORDING

Magnetic recording today provides extremely high density and reliable storage of information at a cost that is approaching that of paper. With a storage density of about 20 Mb/cm^2 (1 Mb = 2^{20} bits $\cong 10^6$ bits), magnetic recording media today can store the equivalent of about 1000 pages of text on 1 cm^2. This is more than 50 000 times the storage density on the RAMAC, which was introduced in 1957 by IBM as the first disk drive for storage of digital information. The original 130-mm magnetic disk drive, introduced in 1980, stored just 5 MB [1 B (byte) = 8 bits]. Today, instead of storing megabytes, 130-mm drives store gigabytes, and drives as small as 45 mm store 80 MB.

This astounding rate of progress over a relatively long lifetime of magnetic record-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ing shows no sign of slowing. Fundamental limits to magnetic recording density are still several orders of magnitude away, and recent product announcements and laboratory demonstrations indicate that the industry is increasing the rate of progress rather than approaching practical limits. In the past few years, both IBM (Tsang *et al.*, 1990) and Hitachi (Futamoto *et al.*, 1991) have demonstrated the feasibility of storing information at densities beyond 100 Mb/cm^2 on a disk. Similar advances can also be expected in audio and video recording.

1.1 Fundamentals of Magnetic Recording

Although magnetic recording is practised in a wide variety of formats and serves a wide variety of applications, the fundamental principles by which it operates are similar in all cases. The fundamental magnetic recording configuration is illustrated in Fig. 1. The recording head consists of a toroidally shaped core of soft magnetic material with a few turns of conductor around it. The magnetic medium below the head could be either tape or disk, and the substrate could be either flexible (for tape and floppy disks) or rigid (for rigid disks). To record on the medium, current is applied to the coil around the core of the head, causing the high-permeability magnetic core to magnetize. Because of the gap in the recording head, magnetic flux emanates from the head and penetrates the medium. If the field produced by the head is sufficient to overcome the coercive force of the medium, the medium will be magnetized by the head field. Thus a representation of the current wave form applied to the head is stored in the magnetization pattern in the medium.

Readout of previously recorded information is typically accomplished by using the head to sense the magnetic stray fields produced by the recorded patterns in the medium. The recorded patterns in the medium cause magnetic stray fields to emanate from the medium and to flow through the core of the head. Thus, if the medium is moved with respect to the head, the flux passing through the coil around the head will change in a manner that is representative of the recorded magnetization pattern in the medium. By Faraday's law of induction, a voltage representative of the recorded information is thus induced in the coil.

1.2 The Magnetic Recording Process

During recording the head is used to produce large magnetic fields that magnetize the medium. It was shown by Karlqvist (1954) that, in the case where the track width and length of the poles along the gap are both large compared to the gap length, the fields produced by a recording head could be de-

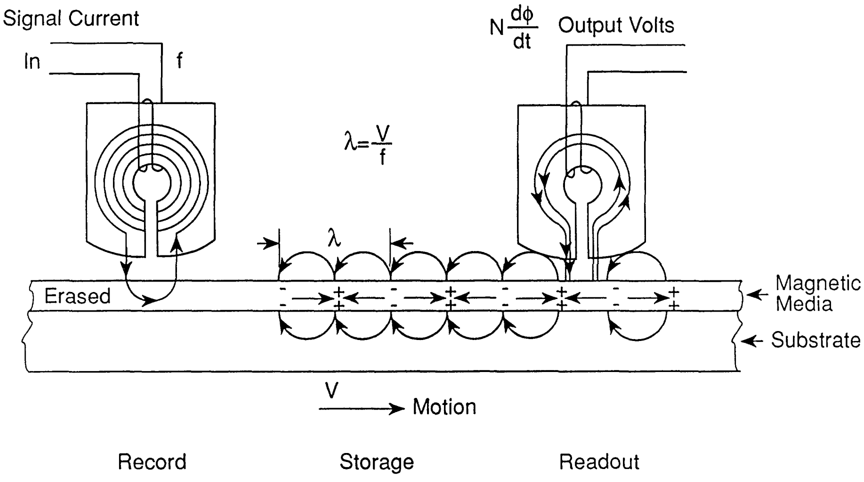

FIG. 1. A diagram illustrating the fundamental magnetic recording configuration.

scribed by

$$H_x = \frac{NI}{\pi g}\left[\tan^{-1}\left(\frac{x + g/2}{y}\right) - \tan^{-1}\left(\frac{x - g/2}{y}\right)\right], \tag{1a}$$

$$H_y = \frac{NI}{\pi g}\ln\left[\frac{(x + g/2)^2 + y^2}{(x - g/2)^2 + y^2}\right], \tag{1b}$$

where H_x and H_y are the longitudinal and perpendicular components of field, as indicated by the coordinates in Fig. 2, N is the the number of turns on the head, I is the current driving the head, and g is the gap width of the head. In this approximation, the contours of equal longitudinal field are described by arcs of circles passing through the gap corners as shown in Fig. 2.

In digital or saturation recording, the recording head is driven with sufficiently large currents that a portion of the recording medium is driven into saturation. However, because of the gradient in the head fields, other portions of the medium see fields less than those required for saturation. This is illustrated in Fig. 2 where the contours for three different longitudinal fields were drawn. In this figure H_{cr} is the remanence coercivity or the field required to produce zero remanent magnetization in the medium after it was saturated, and H_1 and H_2 are fields that would produce negative and positive remanent magnetization, respectively. Note that the head field gradient is the sharpest near the pole tips of the head. This means that smaller head-to-medium spacing and thinner media both lead to narrower transitions being recorded.

Modeling the recording process involves convolving the head field contours with the very nonlinear and hysteretic magnetic properties of the recording medium. A typical *M-H* hysteresis loop for longitudinal magnetic recording media is shown in Fig. 3. Whether the medium has positive or negative magnetization depends upon not only the magnetic field applied, but the past history of the magnetization. If the medium was previously saturated at $-M_s$, then when the magnetic field H is reduced to zero, the remanent magnetization will be $-M_r$; however, if it were previously saturated at $+M_s$, then the remanent magnetization would be $+M_r$. Similarly, if the medium were initially saturated to $-M_s$, then magnetized by a field $+H_1$, and finally allowed to go to a remanent state, the magnetization would go to value M_1. This hysteretic behavior is the basis for the use of the medium for long-term storage of information, but makes the recording process highly nonlinear.

An additional complicating factor in determining the actual recorded pattern is the demagnetizing field of the medium itself. As

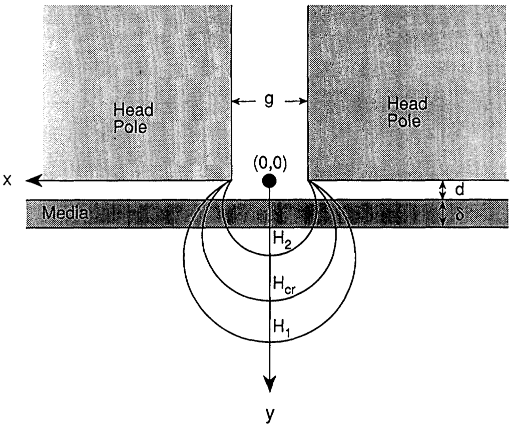

FIG. 2. A diagram illustrating the constant longitudinal field contours in the gap region of a recording head.

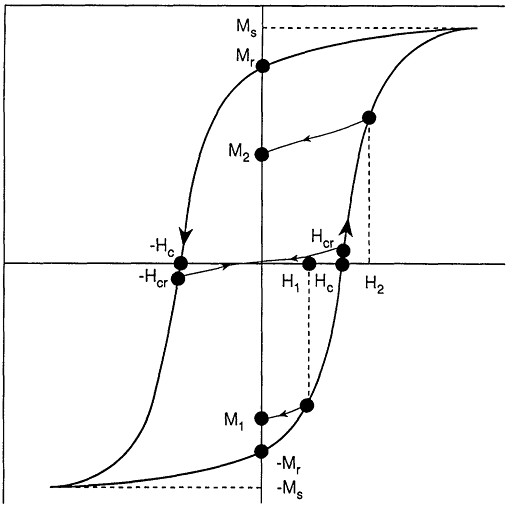

FIG. 3. A remanent *M-H* hysteresis loop for longitudinal recording media.

indicated by the + and − signs in Fig. 1, transitions in the recorded magnetization direction produce effective magnetostatic charge given by

$$\rho_M = -\mu_0 \nabla \cdot \mathbf{M}, \tag{2}$$

which in turn results in demagnetizing fields. The demagnetizing fields outside the medium are what is sensed by the head during readback, but demagnetizing fields also exist inside the medium and act to alter the total field seen by the medium during the record process from that of the head field alone.

Taking into account the head field gradients, the nonlinear *M-H* loop characteristics of the medium, and the demagnetizing fields, Williams and Comstock (1971) developed a model for the recording process. This model predicts the width of a recorded transition, in a material with a square hysteresis loop, to be

$$a = \sqrt{M_r \delta d / \pi H_c}, \tag{3}$$

where δ is the medium thickness, d is the head-to-medium spacing, M_r is the remanent magnetization of the medium, and H_c is the coercivity of the medium. That the transition widens with the product $M_r\delta$ is because the demagnetizing fields increase linearly with this quantity. Similarly, the transition narrows as H_c is increased, because with high coercivity, the medium can resist the transition broadening due to the demagnetizing fields. The increase in transition width with d is because a poorer head-field gradient is obtained with larger head-to-medium spacing.

The nonlinearities of the recording process can be largely removed by a technique referred to as ac bias recording. This is frequently used in analog recording in audio and video recorders. In this technique a high-frequency ac bias is added to the signal to be recorded. This ac bias signal is ramped from a value much larger than the coercivity of the medium to zero. This removes the hysteretic behavior of the medium and causes it to assume a magnetization state that represents the minimum energy state determined by the amount of field produced by the signal to be recorded.

1.3 The Magnetic-Recording Readback Process

As opposed to the recording process, the readback process can usually be modeled as a linear process. This is because the changes in magnetization that occur in either the head or the medium during readback are typically small.

The most common way to model the readback process is to use the principle of reciprocity, which states that the flux produced by the head through a cross section of an element of medium, normalized by the number of ampere turns of current driving the head, is equal to the flux produced in the head by the element of medium, normalized by the equivalent current required to produce the magnetization of that element. For a magnetic recording head, which produces the longitudinal field $H_x(x,y)$ when driven by NI ampere turns of current, this principle leads to the following expression for the voltage induced in the head by a recording medium with magnetization $M(x,y)$ and moving with velocity v relative to the head:

$$e = \frac{\mu_0 W v}{I} \int_d^{d+\delta} \int_{-\infty}^{\infty} \frac{\partial M_x(x-\bar{x},y)}{\partial \bar{x}} \times H_x(x,y)\,dx\,dy, \tag{4}$$

where W is the track width of the head.

This expression shows that the readback voltage induced in the recording head is linearly dependent upon the magnitude of the magnetization in the recording medium being sensed and the relative head-to-medium velocity. The linearity of the readback process ensures that analog recordings such as those recorded on audio or video tapes are faithfully reproduced.

1.4 Magnetic Recording Media

There are a wide variety of magnetic recording media available today. Different applications require different media; however, in many cases the same application will be able to utilize a variety of different competitive media.

Just a decade ago, essentially all recording media consisted of fine acicular magnetic particles embedded in a polymer and coated either onto flexible substrates such as Mylar for floppy disks and tapes or onto rigid Al-

alloy substrates for rigid disks. Today, although such particulate media are still widely used for tape and floppy disks, thin-film media have almost entirely taken over the rigid disk business and metal evaporated thin-film media have been introduced into the tape marketplace. Furthermore, many new particle types have been introduced.

1.4.1 Particulate Recording Media The most common particulate recording media today are γ-Fe_2O_3, Co surface-modified γ-Fe_2O_3 (γ-Fe_2O_3 that has a surface coating of Co), CrO_2, and metal particle media. All of these particles are acicular in shape with aspect ratios of the order of 5 or 10 to 1. The particles are sufficiently small that it is energetically most favorable for them to remain in a single-domain saturated state. Because of demagnetizing effects caused by the acicular shape, the magnetization prefers to align along the long axis of the particle.

As was noted in the discussion of Eq. (3), to achieve higher recording densities requires media with higher coercivity. The coercivity of a particle is determined by the field required to cause the magnetization to switch by 180°. If the magnetization remained in a single-domain state during the switching process, then the coercivity should be given by

$$H_c = (N_a - N_b)M_s, \tag{5}$$

where N_a and N_b are the demagnetizing factors in the directions transverse and parallel to the particle axis, respectively. In practice the coercivity is measured to be less than this. This has been explained as being due to the fact that the magnetization does not remain uniform during the switching process, but switches inhomogeneously (White, 1984). In addition to the effects that the shape anisotropy of the particles have on the coercivity, crystalline anisotropy can also be used to control coercivity.

The coercivity of a medium that is made from particles is determined by the distribution of coercivities of the particles, their orientation in the medium relative to the fields from the head, and their interactions among each other. The coercivities of a variety of particulate recording media are summarized in Table 1.

Although coercivity is indeed an important parameter for magnetic-recording media, it is by no means the only one. Particle size affects medium noise because, at any time, the head is sensing a fixed volume of the medium. Since the particles are quantized and there are statistical variations in their switching behavior, the power signal-to-noise ratio of the medium varies linearly with the number of particles contained in that volume. To reduce particulate-medium noise, it is therefore generally desirable to use small particles.

There is a limit, however, to how small particles may be made and still remain stable. When the thermal energy kT is comparable in magnitude to the energy required to switch a particle, $\mu_0 M H_c$, the particle becomes unstable and may switch as a result of thermal excitation. This phenomenon is known as superparamagnetism and can lead to decay of recorded magnetization patterns over time.

The remanent magnetization of the medium is important because it directly affects the signal level during readback, as shown by Eq. (4). The remanent magnetizations of several particulate media are listed in Table 1. Obtaining high remanent magnetization in a particulate medium requires the use of particles with high saturation magnetization and a high volume packing fraction of particles in the polymer binder. Obtaining a high volume packing fraction of particles in the binder, however, can lead to nonuniform distributions of particles and agglomerates of many particles, which switch together, also causing noise during readback.

Generally, then, to obtain a good high-density particulate recording medium it is desired to have adequate coercivity (to achieve the required recording density), to have small particles (for low noise) with a very narrow switching field distribution (to obtain a narrow transition), to have them oriented along the direction of recording (to obtain a large remanence), and to have them uniformly dispersed (to obtain low modulation noise) with high packing density (to obtain large signals).

1.4.2 Thin-Film Recording Media Thin-film recording media generally have excellent magnetic properties for high-density recording. Since they are nearly 100% dense (voids at the grain boundaries reduce the

Table 1. Magnetic material, saturation remanence $M_r(\infty)$, coercivity H_c, switching-field distribution Δh_r, and number of particles per unit volume, N, of various particulate magnetic-recording media (Köster and Arnoldussen, 1987).

Application	Material	$M_r(\infty)$ kA/m (emu/cm^3)	H_c, kA/m (4π Oe)	Δh_r	N ($10^3/\mu m^3$)
Reel-to-reel audio tape	γ-Fe_2O_3	100–120	23–28	0.30–0.35	0.3
Audio tape IEC I	γ-Fe_2O_3	120–140	27–32	0.25–0.35	0.6
Audio tape IEC II	CrO_2	120–140	38–42	0.25–0.35	1.4
	γ-Fe_2O_3+Co	120–140	45–52	0.25–0.35	0.6
Audio tape IEC IV	Fe	230–260	80–95	0.30–0.37	3
Professional video tape	γ-Fe_2O_3	75	24	0.4	0.1
	CrO_2	110	42	0.3	1.5
	γ-Fe_2O_3+Co	90	52	0.35	1
Home video tape	CrO_2	110	45–50	0.35	2
	γ-Fe_2O_3+Co	105	52–57	0.35	1
	Fe	220	110–120	0.38	4
Instrumentation tape	γ-Fe_2O_3	90	27	0.35	0.6
	γ-Fe_2O_3+Co	105	56	0.50	0.8
Computer tape	γ-Fe_2O_3	87	23	0.30	0.16
	CrO_2	120	40	0.29	1.4
Flexible disk	γ-Fe_2O_3	56	27	0.34	0.3
	γ-Fe_2O_3+Co	60	50	0.34	0.5
Computer disk	γ-Fe_2O_3	56	26–30	0.30	0.3
	γ-Fe_2O_3+Co	60	44–55	0.30	0.5

density somewhat), they can be made to have the highest possible magnetization. Because of their high magnetization they can be made extremely thin and still provide adequate signal during readback. This helps narrow the recorded transition since the head-field gradient is sharper for thinner media, as was discussed in reference to Fig. 2.

Thin-film media can also be made extremely smooth. To achieve the smallest possible head-to-medium spacing and therefore the sharpest head-field gradient and the least spacing loss, smooth media are required.

Thin-film media were first made by plating, but today sputtering is generally preferred. Sputtering enables one to incorporate more elements into the films and to achieve higher coercivity. The coercivity of sputtered thin-film media can be made very high. In volume production today are media with coercivities of 100 kA/m; however, media with coercivities up to 250 kA/m have been made and appear promising (Velu and Lambeth, 1992). Such high coercivities are adequate to achieve more than an order-of-magnitude higher recording density than today.

Numerical models indicate that noise in thin-film media increases when the grains in polycrystalline films are strongly coupled by exchange effects (Zhu and Bertram, 1988). Exchange-coupled films tend to exhibit zigzag transitions, such as the ones indicated in Fig. 4. With exchange coupling the zigzags are much larger and produce more jitter in the transition position relative to the location where the record current in the head goes through zero. Experimental studies have indicated that the introduction of nonmagnetic elements that segregate to the grain boundaries and careful control of the sputtering conditions to achieve a porous microstructure at the grain boundaries, to reduce the exchange coupling, reduce such transition jitter (Chen and Yamashita, 1988).

1.5 Magnetic Recording Heads

1.5.1 Ferrite Heads Early recording heads consisted of toroids of magnetically soft ferrites, such as NiZn-ferrite and MnZn-ferrite, with a few turns of wire around them. For high-density recording applications, however, ferrite can no longer be used, because the saturation magnetization of ferrite is limited to about 400 kA/m. Saturation of the pole tips of a ferrite head begins to occur when the deep gap field in the head approaches one-half the saturation magnetization of the ferrite. Since the fields seen by the medium are one-half to one-quarter of the deep gap field, media with coercivities above

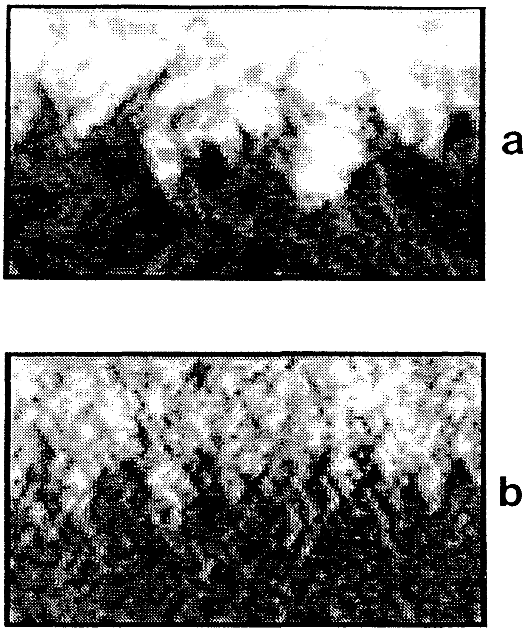

FIG. 4. Magnetization patterns of simulated recorded transitions (a) in a film having exchange coupling between grains, and (b) in a film having no exchange coupling between grains (Zhu, 1992).

about 80 kA/m cannot be reliably written with a ferrite head. High-density thin-film disk media, metal particle media, and metal evaporated media therefore cannot be written with a ferrite head.

1.5.2 Metal-in-Gap Heads Magnetically soft alloys of metals such as Permalloy (NiFe) and Sendust (FeAlSi) have saturation magnetizations of the order of 800 kA/m, about twice that of ferrites, but, because they are metallic, may suffer from eddy-current losses when operated at high frequencies. To overcome the limitations imposed by eddy currents, they are used in layers thinner than a skin depth at their operating frequency. To prevent saturation of the ferrite heads, the high-magnetization metals are applied to the pole faces of the ferrite, making a so-called metal-in-gap (MiG) recording head, as shown in Fig. 5. Since the corners of the pole faces are the first parts of a ferrite head to saturate, the high-magnetization metals enable these MiG heads to be operated to nearly twice the field to which a ferrite head can be operated. Because the layer of metal is thin, it can furthermore be less than a skin depth, and eddy-current losses do not limit performance at high frequencies.

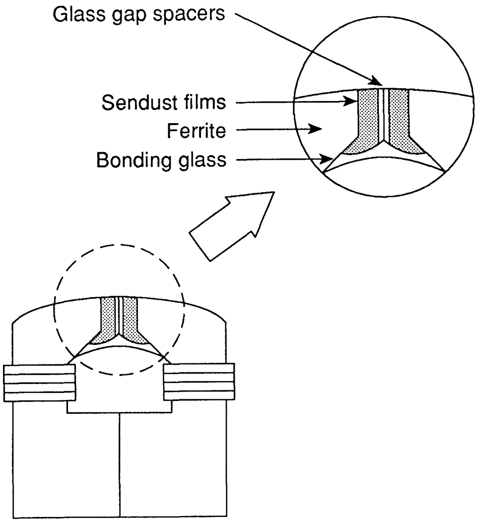

FIG. 5. A diagram of a metal-in-gap (MiG) recording head (Hoagland and Monson, 1991).

1.5.3 Thin-Film Heads Yet another solution to the saturation problem of ferrite heads is to use thin-film heads. Thin-film heads are made of Permalloy and are therefore metallic, but the films are made thinner than the skin depth, and, consequently, the heads operate well at high frequencies. A diagram of a thin-film head is shown in Fig. 6. It consists of a bottom yoke of Permalloy, some insulating layers, a spiral conductor, and a top of Permalloy that is joined to the bottom yoke at the back gap, but separated from it by a thin insulator at the recording gap. These thin-film heads are made using photolithography and microfabrication techniques similar to those used in the manufacture of semiconductor devices. The thin pole tips of these heads actually sharpen the headfield function and, consequently, the pulse shape produced by an isolated transition, although at the expense of some undershoot, as illustrated in Fig. 7. Because thin-film heads are made by photolithographic techniques, they can be made extremely small and with low inductance. This, too, helps extend the frequency of operation.

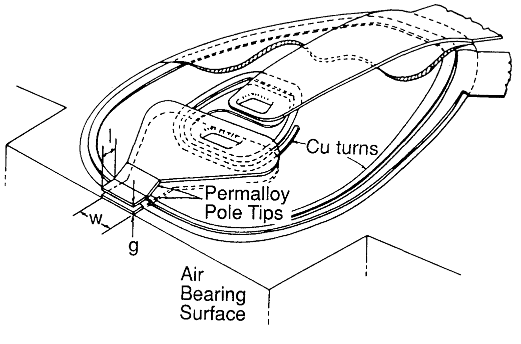

FIG. 6. A diagram of a thin-film head (White, 1984).

1.5.4 Magnetoresistive Heads A relatively new head that is now being used for readback of information in high-density recording is the magnetoresistive head. Magnetoresistive (MR) heads are based on the phenomenon of magnetoresistance in which the electrical resistance of a magnetic material is dependent upon the direction of magnetization in the material relative to the direction of current flow. An unshielded magnetoresistive head is depicted in Fig. 8. Current flows in one end of the head and out the other. The resistivity of Permalloy, from which the head is made, varies as

$$\rho = \rho_0 + \Delta\rho \cos^2\theta, \qquad (6)$$

where θ is the angle that the magnetization in the Permalloy makes relative to the direction of current flow, ρ_0 is the isotropic resistivity, and $\Delta\rho$ is the magnetoresistivity. When the recording medium with a changing magnetization pattern moves under the MR head, the stray fields from the medium cause a change in the direction of magnetization and, consequently, a change in resistance in the head. With a constant-current source driving the head, the head will therefore exhibit a change in voltage across its terminals.

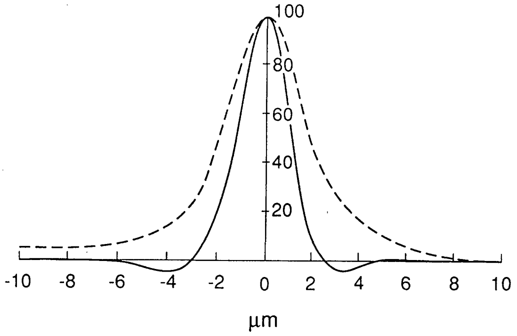

FIG. 7. Pulse shapes for (dashed curve) long- and (solid curve) short-pole heads normalized for equal amplitude (Valstyn and Shew, 1973).

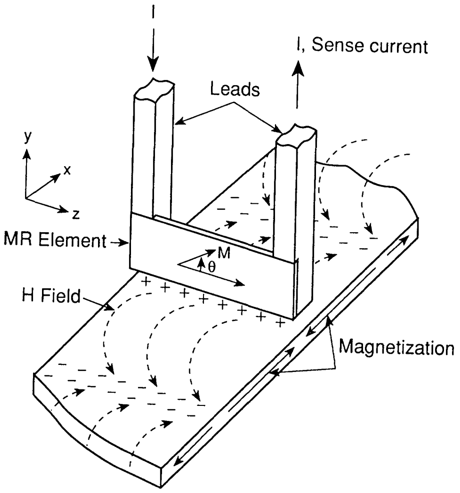

FIG. 8. An unshielded magnetoresistive element (Hoagland and Monson, 1991).

Magnetoresistive heads are typically more sensitive than inductive heads and therefore produce larger signal amplitudes during readback. The increased sensitivity and the fact that the read head is independent of the write head can be used to make a write/read head combination in which the write head writes a wider track than the read head senses. Thus adjacent track interference is reduced during the readback process.

Another advantage of the MR head is that it senses magnetic flux ϕ, not the time rate of change of flux $d\phi/dt$ as an inductive head does. Consequently, whereas the inductive head output voltage is dependent upon the head-to-medium velocity as was shown by Eq. (4), the output voltage of a MR head is independent of velocity.

2. MAGNETO-OPTIC RECORDING

The initial magneto-optic recording products introduced in late 1989 to the market place stored about 600 MB of information on a double-sided 130-mm removable disk. Typically these drives are too large to fit in the standard drive bays on a personal computer and therefore are stand-alone units. Data rates obtainable from these devices are about 1 MB/s, and access times range from about 50 to 100 ms. Both of these performance factors are slow compared to rigid magnetic disk drives, where data rates of 2–3 MB/s and access times of 10–15 ms are commonly obtained.

More recently a second generation of magneto-optic drives has been introduced. These drives use single-sided 90-mm removable disks to store 127 MB of data. Data rates and access times are still about 1 MB/s and 50–100 ms, respectively, but the drives have been made small enough to fit into a standard 90-mm drive bay on a personal computer.

Neither the 130-mm nor the 90-mm magneto-optic drive products enjoy a market that is significant compared to that of magnetic hard drives, but the markets for these products are growing rapidly. Sales were estimated at US$261 million in 1991 and are projected to grow at a 26% rate in the next six years (Abraham and Freeman, 1992).

In the future it can be expected that magneto-optic recording will advance significantly. Standards for the next generation of magneto-optic drives call for doubling the disk rotation speed from 1800 to 3600 rpm. This will double the data rate and shorten latency times, even without increasing the linear bit density. Density increases will be achieved through the use of higher numerical aperture objectives to produce a smaller spot size, run-length limited codes, and more advanced signal processing. Seek times will likely also be reduced, leading to access times in the 20- to 30-ms range. Direct overwrite, which will shorten the time to update a record, will also likely be introduced. For the more distant future, researchers are working on laser diode sources and media that will operate with blue (400-nm wavelength) rather than infrared (800-nm) light, thus permitting a quadrupling of areal density. Thus it appears that substantial advances in magneto-optic recording technology can be expected.

2.1 Fundamentals of Magneto-optic Recording

In magneto-optic recording technology, a focused laser beam is used to write and read reverse magnetic domains in a magnetic thin film having a strongly temperature-dependent coercivity and magnetization perpendicular to the film plane. The writing and reading processes are schematically illustrated in Figs. 9(a) and 9(b), respectively. To write a reverse domain into a magnetic film with all its magnetization pointing in the upward direction, a magnetic field **H** is applied antiparallel to the magnetization. This field is significantly less in magnitude than the room-temperature coercive force of the thin film, however, and so it alone has no effect on the magnetization. If the beam of a diode laser is then focused to a micrometer-sized spot and pulsed onto the film surface, the film temperature in the illuminated region will increase, typically by amounts of the order of 300 °C. Since the coercivity decreases as the Curie temperature (which is typically selected to be about 200 °C) is approached, the magnetization reverses within the heated area. When the thin film cools after the laser pulse,

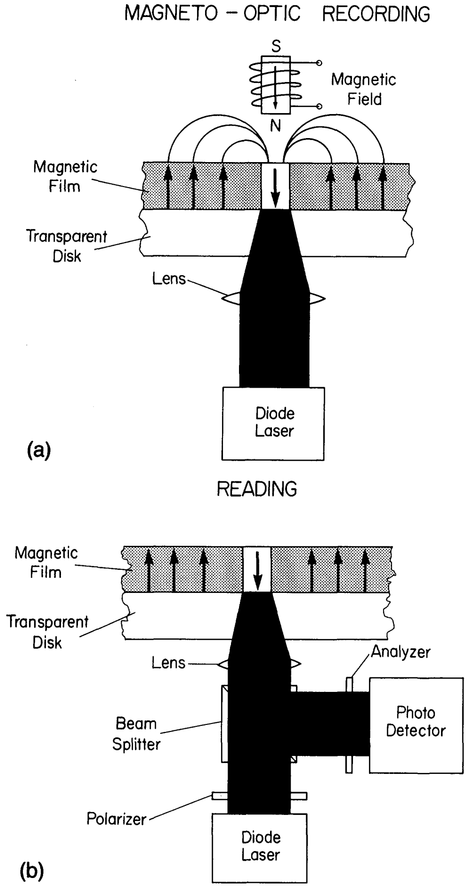

FIG. 9. (a) Magneto-optic recording process. (b) Magneto-optic reading process.

the reverse domain of magnetization is frozen in. Such a reverse domain at a particular location on a circumferential track on a disk could represent a 1 in a binary memory, while the absence of a reverse domain could represent a 0.

To read the stored information the polar Kerr magneto-optic effect is used, as illustrated in Fig. 9(b). A reduced light level from the diode laser is used so that the laser beam does not disturb the recorded information during readout. The diode laser light is plane polarized and focused onto the magneto-optic disk. Upon reflection from the magnetic layer, the plane of polarization of the light is rotated through a small angle θ_K by the polar Kerr magneto-optic effect. The direction of rotation depends upon whether the magnetization is directed upward or downward. The reflected light is analyzed with a second plane polarizer (the analyzer), and the changes in reflected polarization are converted into changes in light intensity that are then detected with the photodetector. Thus the recorded binary pattern of magnetic domains may be read out using the polar Kerr magneto-optic effect.

To overwrite information in the magneto-optical drives available today it is first necessary to erase the old information. This is done by reversing the magnetic field used to write the domains in Fig. 9(a) and turning on the focused laser beam continuously as the sector to be erased passes by. Then, on the second rotation of the disk, the field is again reversed so that it is once again in the direction shown in Fig. 9(a), and the diode laser is pulsed to write the new information. This double-pass overwrite procedure increases the time required to update a record and is a significant disadvantage for magneto-optic recording as compared to magnetic recording where the new data are written directly over the old. The reason this double-pass overwrite is necessary is that to overwrite random data directly, it would be necessary to modulate the magnetic field used to determine the direction of magnetization during the writing process at the full data rate. This requires a small recording head close to the magnetic thin film and eliminates a key advantage of magneto-optic recording—having a large head-to-medium spacing. Methods to achieve direct overwrite, which are being aggressively pursued in research laboratories throughout the world, will be discussed later.

2.2 Magneto-optic Recording Media

The manufactured magneto-optic media today consist of sputtered amorphous alloys of rare earths and transition metals. These materials exhibit a uniaxial anisotropy such that the magnetization prefers to lie perpendicular to the film plane. In these alloys the rare-earth and transition-metal subnetwork magnetizations are antiferromagnetically coupled. Thus rare-earth-transition-metal (RE-TM) films are ferrimagnetic. The temperature dependences of the rare-earth and transition-metal magnetizations are schematically illustrated in Fig. 10. At low temperatures the magnetization of the rare earths dominates and the net magnetization points in the same direction as the rare-earth magnetization. However, at high temperature the transition-metal magnetization dominates and the net magnetization points in the direction of the transition-metal magnetization. At an intermediate temperature, called the compensation temperature T_{comp}, the two subnetwork magnetizations are equal and opposite in direction, causing the net magnetization to go to zero. At sufficiently high temperatures, thermal energy over-

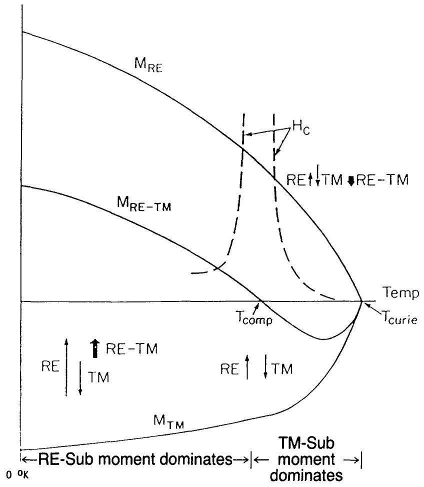

FIG. 10. The temperature dependences of the rare-earth magnetization (M_{RE}), transition-metal magnetization (M_{TM}), net magnetization ($M_{RE\text{-}TM}$), and coercivity (H_c), of an amorphous rare-earth-transition-metal alloy.

comes the exchange coupling and the ferrimagnetic behavior vanishes at the Curie temperature.

The compensation behavior of these materials gives rise to the strong temperature dependence of coercivity H_c that is required for magneto-optic recording. As the compensation temperature is approached, the net magnetization **M** tends toward zero. Since the energy coupled into the magnetization by an externally applied field **H** is $-\mu_0\mathbf{M}\cdot\mathbf{H}$, as the compensation temperature is approached, increasingly large amounts of field must be applied to overcome energy barriers to reversal of the magnetization such as anisotropy. Hence as the compensation temperature is approached the coercive force tends toward infinity. Conversely, as the Curie temperature is approached, thermal energy tends to help in destabilizing the magnetization and only a small field is required to cause switching. Hence the coercive force H_c tends toward infinity at the compensation temperature and toward zero at the Curie temperature, as suggested in Fig. 10.

The most widely used RE-TM materials are alloys of Gd, Tb, Fe, and Co. As explained in detail elsewhere (Kryder, 1989), these alloy systems allow one to adjust independently the compensation temperature, the Curie temperature, the magnetization-coercive force product MH_c, and the uniaxial anisotropy perpendicular to the film plane.

For future high-density recording it is anticipated that 400-nm-wavelength laser light sources will be used. By shortening the wavelength of the light by a factor of 2, both the linear bit density and the track density may be improved by a factor of 2, leading to an increase in areal density by a factor of 4. However, the Kerr magneto-optic rotation θ_K produced by RE-TM films typically decreases as the wavelength of the light is shortened. This, combined with the fact that the read laser intensity must be reduced when a smaller spot is read and that there are fewer photons and therefore higher shot noise when shorter-wavelength light is used, means that the signal-to-noise ratio degrades in RE-TM films when shorter-wavelength light is used.

To overcome the problems with RE-TM films when 400-nm light is used, other materials are being investigated. Garnets (Gomi and Abe, 1989; Shono *et al.*, 1989) and transition-metal multilayers such as Pt/Co and Pd/Co (Zeper *et al.*, 1989; Lin and Do, 1990) appear promising. Both provide enhanced magneto-optic effects at short wavelengths of light.

2.3 Magneto-optic Recording Heads

A magneto-optic recording head today contains a large number of discrete optical components. A schematic diagram of a typical head is shown in Fig. 11. A laser diode, typically with a wavelength of about 800 nm, is used as the light source. Since a laser diode emits an elliptically shaped beam, some means are typically used to circularize the beam after collimation by a collector lens. In Fig. 11 this is illustrated as being done by an anamorphic prism. After passing through the anamorphic prism and a beam splitter, the laser beam is focused by an objective lens through the transparent substrate of the magneto-optic disk onto the magnetic thin film, in which the information is stored in the form of reverse domains. After reflecting off the disk, the light returns to the beam splitter and is deflected toward a polarizing beam splitter. This polarizing beam splitter is set at 45° to the incident plane of polarization of the light from the laser. Thus, if there were no rotation of the plane of polarization of the light by the disk, 50% of the light would go to each quadrant photodiode assembly. On summing the signals to the four photodiodes from each quadrant photodiode and differentially detecting the sums from each quadrant photodiode assembly, the output would then be zero. With a magneto-optic signal, however, the signal at one quadrant photodiode would go up by a certain amount Δ, while the signal at the other goes down by the same amount; thus the differential signal would be 2Δ. This differential detection system also cancels out, to first order, all intensity variations that arise from the laser or changes in reflectivity of the disk. This common-mode noise rejection is very important to obtaining a good signal-to-noise ratio from the very small magneto-optic effect.

For focus control, the objective lens is mounted in a voice-coil assembly that moves it vertically relative to the disk (not shown in Fig. 11). The system is designed so that when the laser beam is properly focused on the disk, the laser beam also comes to a focus on the aperture plane of the cylindrical len-

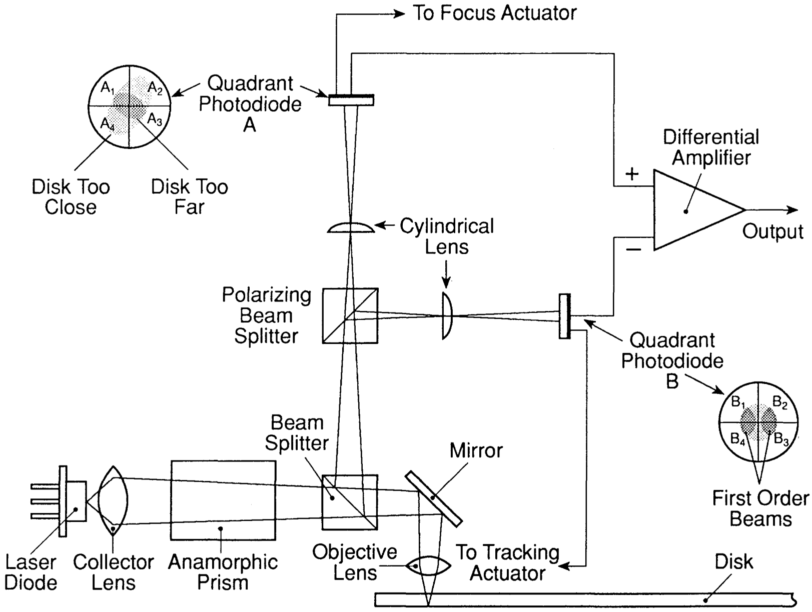
To Focus Actuator
Quadrant Photodiode A
A1
A2
A4
A3
Disk Too Close
Disk Too Far
Differential Amplifier
+
−
Output
Cylindrical Lens
Polarizing Beam Splitter
Cylindrical Lens
Quadrant Photodiode B
B1
B2
B4
B3
First Order Beams
Beam Splitter
Mirror
Laser Diode
Collector Lens
Anamorphic Prism
Objective Lens
To Tracking Actuator
Disk

FIG. 11. A magneto-optical recording head.

ses before the quadrant photodiode detectors. In that case the cylindrical lenses have no effect on the beam and the quadrant photodiode detectors are illuminated with a circular beam. The focus error signal is derived by summing the signals from photodiodes A_1 and A_3 and subtracting that sum from the sum of the signals from photodiodes A_2 and A_4. With a circular beam, as is the case when the disk is properly focused, $A_1 + A_3 = A_2 + A_4$ and there is no focus error signal. However, when the disk is too close to or too far from the objective lens, the light will no longer focus on the aperture of the cylindrical lens, and the quadrant photodiode detector will be illuminated with an elliptical beam with its major axis along $A_1 + A_3$ or $A_2 + A_4$, depending upon whether the objective is too close to or too far from the disk, as indicated in the diagram of quadrant photodiode A in Fig. 11. Thus the difference signal $(A_1 + A_3) - (A_2 + A_4)$ will provide a focus error signal, which can be used for feedback to the voice-coil assembly that moves the objective lens back into focus.

For track following the so-called push-pull tracking sensor is frequently used. With this sensor the zero-order and first-order diffracted beams from a grooved disk return through the objective lens and interfere with one another, causing illumination on the quadrant photodiodes as illustrated in the diagram of quadrant photodiode B in Fig. 11. When the beam is centered directly on a groove or midway between two grooves, the two first-order beams are identical, and they interfere equally with the zero-order beam; however, when the beam is off center, they interfere differently. By aligning the quadrant photodiode detector so that one first-order beam falls on the left half of the detector $(B_1 + B_4)$ and the other falls on the right half of the detector $(B_2 + B_3)$, one can then derive a tracking error signal from the difference $(B_1 + B_4) - (B_2 + B_3)$. This tracking error signal can then be used to control a voice-coil assembly to translate the objective in the radial direction on the disk, thus providing a servo system for track following.

The use of discrete optical components makes a magneto-optical recording head massive and bulky in comparison to a magnetic-recording head. This has two very significant drawbacks:

1. the large mass limits the acceleration that the head can undergo and thus slows the seek time,
2. the size of the head makes it impossible to place several disks closely together on

a spindle, as is commonly done in magnetic hard drives.

Thus the large size of magneto-optic heads are a major reason why magneto-optic drives have longer seek times and have lower capacity than magnetic hard drives.

Recently manufacturers have been introducing magneto-optic drives with split heads to reduce the seek time. In these heads the laser and detector optics are in a stationary location beside the disk, and a collimated beam is sent to and from a much smaller moving module containing a 90° mirror, focusing lens, and voice-coil fine servo. By minimizing the mass of the moving element in this way, seek times can be significantly reduced.

To reduce the size and mass of optical heads, a number of alternative technologies are under consideration. These include the use of optical fibers (Opsasnick *et al.*, 1991), thin-film holographic optical elements (Sincerbox, 1989; Ono *et al.*, 1989), nonmechanical scanners (Arimoto *et al.*, 1990), and integrated head structures (Ura *et al.*, 1986; Valetta, 1989; Ura *et al.*, 1988; Arimoto *et al.*, 1990).

2.4 Direct Overwrite

As noted above, present magneto-optic recording drives require two passes of the disk to overwrite old information. This places them at a significant disadvantage compared to rigid disk drives that can simply write the new record over the old. There are currently at least three approaches to solving this problem and providing direct overwrite in a magneto-optical drive. The most straightforward approach is to use a magnetic recording head flying close to the disk to provide the rapidly modulated magnetic field required to overwrite data at the full data rate (Nakao *et al.*, 1987).

This technique reportedly works well at the data rates of present drives. Reportedly, the jitter in the position of the domain walls is actually smaller with this technique than when the laser is pulsed. This has led to the suggestion that such field-modulation overwrite could utilize run-length limited codes that pack more than 1 bit per transition.

The disadvantages of this technique include the fact that only single-sided disks may be used, since the magnetic head must be close to the magnetic layer. In addition, the double-sided head is more complex and adds to the cost of the drive, and the inductance of the head may place limitations on the frequency at which this technique may be applied.

The first direct overwrite scheme to be proposed (Shieh and Kryder, 1986) uses a laser beam modulated between two different energy levels to record or erase domains selectively in a single magnetic layer. This technique utilizes specially designed materials in which a high-energy laser pulse writes a domain, whereas a low-energy pulse destabilizes the domain and causes it to collapse. Critical to the operation of this technique is a high domain-wall mobility and a significant difference in the thermal gradients achieved during writing and erasing.

This single-layer laser modulation technique is the most attractive to implement from a systems point of view, because it utilizes a single magnetic layer, not substantially more difficult to manufacture than present media, and because it also could use present magneto-optic heads (Kryder and Schultz, 1989). Double-sided media would be allowed, and data collected thus far indicate that the technique does not restrict the data rate. In fact, the technique presently works better at higher data rates than at low data rates.

The problem with this technique is that an adequate signal-to-noise ratio has not yet been achieved. The accepted standard calls for a 45-dB carrier-to-noise ratio (CNR) in a 30-kHz slot bandwidth. Recently this standard has been met, but only barely (Schultz *et al.*, 1993). Further improvements are expected.

The third overwrite technique also uses laser modulation, but it utilizes multilayer exchange-coupled media (Saito *et al.*, 1987; Kaneko *et al.*, 1989; Fukami *et al.*, 1990). With three- and four-layer media, CNR in excess of 50 dB has been achieved. Because of the multiple layers and the criticality of their parameters, this technique is expected to add significantly to the cost and to degrade from the yield of manufactured media. On the other hand, if adequate yield can be achieved, the technique does not require any other changes to the drive.

In summary, at least three direct overwrite schemes are being actively pursued. Each of them has its own advantages and shortcomings. It is too early to select a definite winner, but it appears likely that magneto-optic drives will offer direct overwrite in the near future.

3. CONCLUSIONS

Magnetic recording today finds a wide spectrum of applications and formats ranging from relatively low-density, low-cost floppy-disk drives and audio recorders to high-density video cassette recorders, digital audio tape recorders, computer tape drives, rigid disk drives, and instrumentation recorders. Magnetic recording media are used for storing audio, video, and computer data and a wide variety of data from instrumentation. They are used for both temporary and long-term archival storage. Formats range from tape, which has the highest volumetric packing density and lowest cost per bit stored, to rigid disks, which provide fast access to a large volume of data. The performance of this technology and the applications for it are expected to expand even more rapidly well into the next century.

Magnetic recording technology has increased storage density by more than a factor of 50 000 over the past 35 years since it was first used in a disk format for computer data storage; however, there is no sign that this rapid pace is slowing. Indeed, fundamental limits, set by superparamagnetism, are estimated to be yet several orders of magnitude from where we are today. Furthermore, recent product announcements and developments in research laboratories suggest that the rate of progress is likely to increase. Storage densities of over 100 $\mathrm{Mb/cm^2}$ are likely by the end of this decade, and densities of 1 $\mathrm{Gb/cm^2}$ appear likely in the early twenty-first century. Multilayer thin-film media, thin-film write heads utilizing high-magnetization thin-film multilayer materials, and magnetoresistive write heads are expected to be among the new technologies which make this occur.

Magneto-optic recording has only recently been introduced to the marketplace. Areal storage densities are high, but the volumetric storage densities are no higher than in rigid magnetic disk drives containing multiple disks. Data rates and access times are currently about a factor of 2 or 3 times slower than on rigid disk drives, and costs at the drive level are higher than for rigid disks. This leaves the removability of the magneto-optic medium as its main advantage. Presently the medium costs about \$0.50/MB. The technology, however, is young, and both cost reductions and improvements in storage density and performance are expected. Higher numerical aperture objectives, improved signal processing, improved track-following servos, and shorter-wavelength laser sources combined with improved media for short wavelength recording are expected to increase the storage density. Over an order of magnitude improvement is likely. Lighter and smaller heads will lead to improved seek times. Direct overwrite will reduce the time to update records, and both higher bit density and higher spindle rotation rates will lead to higher data rates. The higher spindle rotation speed will also shorten the latency time. A removable 50-mm magneto-optic disk that will store a few GB of data with data rates and access times only slightly poorer than those of rigid magnetic drives appears likely by the year 2000.

ACKNOWLEDGMENT

This material is based in part upon work supported by the National Science Foundation under Grant No. ECD-8907068. Portions of this article have been previously published in an article entitled "Magnetic Recording" in the *Electrical Engineering Handbook*, published by CRC Press, Boca Raton, FL, in 1993, and in *Annu. Rev. Mater. Sci.* (1993), **23**, 411–436.

GLOSSARY

Coercive Force or Coercivity: The magnetic field required to reduce the mean magnetization of a sample to zero after it was saturated.

Compensation Temperature: The temperature at which the subnetwork magnetizations of a ferrimagnetic material cancel each other, producing zero net magnetization.

Curie Temperature: The temperature at which thermal excitation becomes sufficiently large that a magnetic material has no spontaneous magnetization in zero applied field.

Demagnetizing Field: The magnetic field produced by divergences in the magnetization of a magnetic sample.

Direct Overwrite: The process of writing new information over old with no intervening erasure process.

Longitudinal Recording: Magnetic recording in which the magnetization lies in the plane of the medium and parallel to the track.

Magnetoresistance: The resistance change produced in a magnetic sample when its magnetization is changed.

Perpendicular Recording: Magnetic recording in which the magnetization lies perpendicular to the plane of the medium.

Remanent Coercivity: The magnetic field required to produce zero remanent magnetization in a material after the material was saturated.

Remanent Magnetization: The magnetic moment per unit value of a material in zero field.

Saturation Magnetization: The magnetic moment per unit volume of a material when the magnetization in the sample is aligned (saturated) by a large magnetic field.

Superparamagnetism: A form of magnetism in which the spins in small particles are exchange coupled, but may be collectively switched by thermal energy.

Uniaxial Anisotropy: An energetic preference for the magnetization to lie parallel (or at right angles to) a single axis in a magnetic material.

Works Cited

Abraham, Robert C., Freeman, Raymond C., Jr. (1992), *Freeman Reports: Optical Data Storage Outlook,* Santa Barbara, CA: Freeman Associates.

Arimoto, Akira, Muraoka, Kouji, Shimano, Takeshi, Senda, Kazuto, Ishikawa, Sachiko (1990), *Appl. Opt.* **29**, 247–265.

Chen, Tu, Yamashita, Tom (1988), *IEEE Trans. Magn.* **MAG-24**, 2700–2705.

Fukami, T., Nakaki, Y., Tokumagu, T., Tagachi, M., Tsutsumi, K. (1990), *J. Appl. Phys.* **67**, 4415–4416.

Futamoto, M., Kugiya, F., Suzuki, M., Takano, H., Matsuda, Y., Inaba, N., Miyamura, Y., Aragi, K., Nakao, T., Sawaguchi, H., Fukuoka, F., Munemoto, T., Takagaki, T. (1991), *IEEE Trans. Magn.* **MAG-27**, 5280–5285.

Gomi, M., Abe, M. (1989), *Proceedings of the Materials Research Society Symposium,* Pittsburgh: Materials Research Society, Vol. 150, pp. 121–130.

Hoagland, A. S., Monson, J. E. (1991), *Digital Magnetic Recording,* New York: Wiley.

Kaneko, A. S., Aratani, K., Mutoh, Y., Nakaoki, A., Watanabe, K., Makino, H. (1989), *Jpn. J. Appl. Phys.* **28**, 27–31.

Karlqvist, O. (1954), *Trans. R. Inst. Technol.,* No. 86; reprint (1985), in: R. M. White (Ed.), *Introduction to Magnetic Recording,* New York: IEEE.

Köster, E., Arnoldussen, T. C. (1987), "Recording Media," in: C. Denis Mee, E. D. Daniel (Eds.), *Magnetic Recording,* New York: McGraw-Hill.

Kryder, Mark H. (1989), in *Proceedings of the Materials Research Society Symposium,* Pittsburgh: Materials Research Society, Vol. 150, pp. 3–14.

Kryder, Mark H., Schultz, M. D. (1989), *Jpn. J. Appl. Phys. Suppl.* **28–3**, 3–9.

Lin, C. J., Do, H. V. (1990), *IEEE Trans. Magn.* **MAG-26**, 1700–1705.

Nakao, T., Ojima, M., Miyamura, Y., Okamine, S., Sukeda, H., Ohta, N., Takeuchi, Y. (1987), *Jpn. J. Appl. Phys. Suppl.* **26–4**, 149–154.

Ono, Yuzo, Kimura, Yasuo, Nishida, Nobuo (1989), in: Ivan N. Cindrich, Sing H. Lee (Eds.), *Holographic Optics: Optically and Computer Generated,* SPIE Proceedings Vol. 1052, Bellingham, WA: SPIE, p. 150.

Opsasnick, Michael N., Stancil, Daniel D., White, Sean T., Tsai, Ming-Horn (1991), in: J. J. Burke, N. Imemura, T. A. Schull (Eds.), *Optical Data Storage '91,* SPIE Proceedings Vol. 1499, Bellingham, WA: SPIE, p. 276.

Saito, J., Sato, M., Matsumoto, H., Akasaka, H. (1987), *Jpn. J. Appl. Phys. Suppl.* **26-4**, 155–159.

Schultz, Mark D., Kryder, Mark H., Sekiya, M., Chiba, K. (1993), "True Direct Overwrite in Single Layer Direct Overwrite Magneto-Optic Recording," in: Proceedings of the Intermag Conference, Stockholm, Sweden, *IEEE Trans. Magn.* (to be published).

Shieh, H.-P. D., Kryder, M. H. (1986), *Appl. Phys. Lett.* **49**, 473–474.

Shono, K., Kuroda, S., Kano, H., Koshino, N., Ogano, S. (1989), in: *Proceedings of the Materials Research Society Symposium,* Pittsburgh: Materials Research Society, Vol. 150, p. 131–135.

Sincerbox, Glenn T. (1989), in: G. Michael Morris (Ed.), *Holographic Optics II: Principles and Applications,* SPIE Proceedings Vol. 1136, Bellingham, WA: SPIE. p. 80.

Tsang, M.-M., Chen, T., Yogi, T., Ju, K. (1990), *IEEE Trans. Magn.* **MAG-26**, 1689–1693.

Ura, Shogo, Suhara, Toshiaki, Nishihara, Hiroshi, Koyama, Jiro (1986), *IEEE J. Lightwave Technol.* **LT-4**, 913–918.

Ura, Shogo, Sunagawa, Hiroshi, Suhara, Toshiaki, Nishihara, Hiroshi (1988), *IEEE J. Lightwave Technol.* **LT-6**, 1028–1033.

Valetta, S. (1989), "Reading Head in Integrated

Optics for Reading Information Recorded on a Magnetic Support," U.S. Patent No. 4 796 226.

Valstyn, E. P., Shew, L. F. (1973), *IEEE Trans. Magn.* **MAG-9**, 317–322.

Velu, E., Lambeth, D. (1992), *IEEE Trans. Magn.* **MAG-28**, 3249–3254.

White, R. M. (1984), *Introduction to Magnetic Recording*, New York: IEEE, p. 14.

Williams, M. L., Comstock, R. L. (1971), in: C. D. Graham, Jr., J. J. Rhyne (Eds.), *Magnetism and Magnetic Materials*, AIP Conference Proceedings Vol. 5, New York: American Institute of Physics, pp. 738–742.

Zeper, W. B., Greidanus, F. J. A. M., Carcia, P. F. (1989), *IEEE Trans. Magn.* **MAG-25**, 3764–3766.

Zhu, J.-G. (1992), "Micromagnetic Modeling of Thin Film Recording Media," in: T. C. Arnoldussen, L. L. Nunneley (Eds.), *Noise in Digital Recording*, Singapore: World Scientific.

Zhu, J.-G., Bertram, H. N. (1988), *IEEE Trans. Magn.* **MAG-24**, 2706–2708.

Further Reading

There are several books that provide additional information on magnetic and magneto-optic recording. They include the following:

Hoagland, Albert S., Monson, James E. (1991), *Digital Magnetic Recording*, New York: Wiley.

Marchant, Alan B. (1990), *Optical Recording*, Menlo Park, CA: Addison-Wesley.

Mee, C. Denis, Daniel, Eric D. (1987), *Magnetic Recording*, New York: McGraw-Hill.

White, R. M. (1984), *Introduction to Magnetic Recording*, New York: IEEE.

A number of articles on magneto-optic recording media are contained in the following:

Robinson, C. J., Suzuki, T., Falco, C. M. (Eds.) (1989), *Materials for Magneto-Optic Data Storage*, Pittsburgh: Materials Research Society.

MAGNETOHYDRODYNAMICS

PAUL H. ROBERTS, *Department of Mathematics, University of California at Los Angeles, Los Angeles, California, U.S.A.*

Introduction ... 111
1. **Electrodynamics** ... 113
1.1 Basic Equations ... 113
1.2 Stress and Energy ... 115
1.3 Induction Equation; Perfect Conductors ... 115
1.4 Continuity Properties ... 116
2. **Fluid Mechanics** ... 117
2.1 The Ideal Fluid ... 117
2.2 Continuity Conditions for an Ideal Fluid; Shocks ... 118
2.3 The Incompressible Viscous Fluid ... 119
3. **Steady Parallel Flows** ... 120
3.1 Hartmann Flow ... 121
3.2 Flow at Large Hartmann Number ... 124
3.3 The Hartmann Layer and Applications ... 125
4. **Steady Nonparallel Flows** ... 126
5. **High Frequencies; Levitation, Heating, and Stirring** ... 128
6. **Infinitely Conducting Fluids** ... 130
6.1 Alfvén's Theorem ... 130
6.2 Waves and Shocks ... 130
6.3 Creation of Magnetic Energy ... 133
6.4 Magnetic Helicity ... 134
6.5 The Lundquist Number ... 135
7. **Magnetohydrostatics** ... 135
7.1 Balancing Nonzero Pressure Gradients ... 135
7.2 Force-free Fields ... 137
7.2.1 First Variational Property ... 138
7.2.2 Second Variational Property ... 138
8. **Ideal Instabilities** ... 139
8.1 Approaches to Stability ... 139
8.2 The Energy Method ... 140
8.2.1 The Flowers–Ruderman Mechanism ... 140
8.2.2 The *Z* Pinch ... 140
8.2.3 Interchange Instability ... 142
8.2.4 Screw Pinch; Suydam's Criterion ... 142
9. **Resistive Effects** ... 143
9.1 Flux Concentration and Destruction ... 143
9.2 Flux Expulsion ... 144
9.3 Field-Line Reconnection ... 145
9.4 Resistive Instabilities; Tearing and Coalescence ... 147
10. **MHD in the Large** ... 148
10.1 Sunspots; Magnetoconvection; Formation of Flux Tubes ... 148
10.2 Current Sheets; Solar Flares ... 149
10.3 Fluid Dynamos ... 149
10.4 RMHD ... 150
Acknowledgments ... 152
Notation ... 152
Glossary ... 153
Works Cited ... 155
Further Reading ... 155

INTRODUCTION

Magnetohydrodynamics (MHD) is the study of the motion of electrically conducting fluid in the presence of a magnetic field. It is woven from two threads. The first is electrodynamic: the motion of the fluid creates, from the prevailing magnetic field $\mathbf{B}$, an electromotive force (emf) $\mathbf{V} \times \mathbf{B}$, where $\mathbf{V}$ is the fluid velocity. This emf sets up an electric field $\mathbf{E}$ and drives electric currents of density $\mathbf{J}$ which (by Ampère's law) modify the magnetic field. The second thread is fluid dynamical: the field and current create a Lorentz force $\mathbf{J} \times \mathbf{B}$, which modifies the fluid flow.

The two threads together constitute more than the mere sum of the two. New time scales and new phenomena, particularly the Alfvén wave, arise in MHD that are completely absent in electrodynamics and fluid dynamics

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

separately. MHD differs essentially from electrohydrodynamics (EHD) in the nature and magnitude of **E**. In EHD, large electric fields are externally applied; in MHD, the electric field is produced by the system itself from the emf created by the motion and is therefore of order $\mathcal{V}\mathcal{B}$, where $\mathcal{V}$ and $\mathcal{B}$ are typical magnitudes of **V** and **B**.

MHD phenomena arise in two main contexts: through *controlled*, man-made engineering applications and through the *uncontrollable* evolution of naturally occurring objects, such as planets, stars, and galaxies. Clearly, the latter should not occupy a large place in an encyclopedia devoted to applied physics, and only one section of this review (Sec. 10) is devoted to it. Controlled MHD may itself be divided into two main parts: the MHD of liquid metals and the MHD of ionized plasmas. Both these areas have important technological applications.

Liquid metals cool the cores of fast fission reactors; their use in the blankets of fusion reactors is also envisaged, both to remove heat and to breed tritium. The possibility of generating electricity in conventional, fossil-fuel power stations by direct MHD conversion (i.e., without the use of turbines) has also been extensively investigated. Metallurgical processing provides interesting applications of MHD, including the levitation, heating, and stirring of casts by high-frequency magnetic fields.

Fusion reactors confine the reacting plasmas by strong magnetic fields. Such "magnetic bottles" are prone to instabilities that shorten the confinement time of the plasmas so drastically that thermonuclear fusions in the reactor are too few for its commercial viability. The understanding and removal of such instabilities has for many years been a central target of MHD research. And today plans are afoot for the construction of large reactors in which the energy released by thermonuclear reactions exceeds that expended in their operation.

MHD is a comparatively young subject, a landmark paper by Alfvén in 1942 often being considered to mark its birth. Today it is such a vast and rich field that the choice of topics here has had to be constrained. The lack of a full discussion of MHD turbulence is particularly regretted. Sources that deal with this topic are cited at the end of the article. It was also impractical to reference even all the most significant papers, and the literature cited below consists principally of books and reviews.

The first two sections of this article summarize the underlying theory it uses; they may be of little interest to readers familiar with electromagnetism and fluid mechanics, who may wish to skip to Secs. 3–5 in which the MHD of liquid metals is studied. Experiments on MHD are notoriously difficult to perform, and the emphasis is therefore on theory and its interpretation. Readers anxious to know more about the current experimental situation might profit by studying a recent book by Moreau (1990). Theory relevant to magnetically confined plasmas is expounded in Secs. 6–9. Readers wishing to study this field in greater depth should benefit by reading a book by Freidberg (1987). Topics of astrophysical MHD are adumbrated in Sec. 10.

One aim of this article is to describe MHD phenomena in as simple terms as possible. For example, although an MHD theory has been developed that is relativistically covariant, we shall be concerned here with conducting fluids moving with velocities $\mathcal{V}$ much less than the speed (c_0) of light. Since it would be inconsistent to supplement nonrelativistic equations of fluid motion with relativistic equations of electromagnetism, we shall use only the reduced "pre-Maxwell equations," in which the displacement current is absent (see Sec. 1). This action requires little defense, since most (though not all) MHD research assumes that relativistic effects are negligible.

Somewhat more questionable was the decision to limit the discussion here to the MHD of dense conductors that are well described by Ohm's law and move according to the laws governing a Newtonian fluid. In defense of this simplification, one may argue that, even though to obtain a quantitatively good description it may be necessary, especially in the case of rarefied plasmas, to replace our dense fluid by a more realistic though a necessarily more complicated model, the dense fluid is sufficient to demonstrate many of the interesting phenomena of MHD in a qualitatively accurate way. Even the full theory of a Newtonian fluid is unnecessarily intricate for our purposes, and we therefore concentrate on two special cases that suffice for almost all our needs: the *incompressible* viscous fluid and the *inviscid* compressible fluid (see Sec. 2). Deviations from these models are described in qualitative terms only.

Table 1 lists some of the MHD parameters, together with their values in different MHD situations. A table of notation appears at the end of the article. SI units are used throughout the article, and (as above) typical sizes of physical variables are denoted by calligraphic letters; e.g., a length characteristic of a body is written as $\mathcal{L}$; a typical flow speed is, as above, $\mathcal{V}$. A familiarity with vector calculus is assumed, including use of the summation convention for repeated suffices.

1. ELECTRODYNAMICS

This section summarizes the electromagnetic ("em") basis of MHD in SI units. Readers for whom this summary is unnecessary may wish to refer to the table of notation at the end of this article and then skip forward to Sec. 2.

1.1 Basic Equations

MHD is principally concerned with fluids moving with speeds $\mathcal{V}$ small compared with c_0, the velocity of light in vacuo. Naturally, the basic dynamics is then Newtonian, and the basic electrodynamics is also nonrelativistic; i.e., it is pre-Maxwellian, so that

$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E}, \quad \mathbf{J} = \nabla \times \mathbf{H}, \tag{1, 2}$$

$$\nabla \cdot \mathbf{B} = 0, \quad \nabla \cdot \mathbf{D} = \vartheta. \tag{3, 4}$$

Equation (1) is Faraday's law of induction; Eq. (2) is Ampère's law; Eq. (3) expresses the absence of magnetic poles; and Eq. (4) shows how free charges create **D**. Equation (2) lacks the displacement current present in full Maxwell theory, which makes it relativistically covariant.

Equation (1) shows that, if Eq. (3) holds initially, it holds for all time. By Eq. (2),

$$\nabla \cdot \mathbf{J} = 0, \tag{5}$$

which shows that the term $\partial\vartheta/\partial t$ present in the full equation of charge conservation ($\partial\vartheta/\partial t + \nabla \cdot \mathbf{J} = 0$) is negligible in prerelativistic theory; its absence represents an $O(\mathcal{V}^2/c_0^2)$ error, and errors of the same order are consistently present throughout pre-Maxwell theory.

The pre-Maxwell equations are invariant under the Galilean transformation $\mathbf{x} \rightarrow \mathbf{x}' = \mathbf{x} - \mathbf{w}t$ to a reference frame moving with a constant relative speed **w**. To see this, recall that according to pre-Maxwell theory,

$$\mathbf{B}' = \mathbf{B}, \quad \mathbf{E}' = \mathbf{E} + \mathbf{w} \times \mathbf{B}, \tag{6}$$

$$\mathbf{H}' = \mathbf{H}, \quad \mathbf{D}' = \mathbf{D} + \mathbf{w} \times \mathbf{H}/c_0^2, \tag{7}$$

$$\mathbf{J}' = \mathbf{J}, \quad \vartheta' = \vartheta - \mathbf{w} \cdot \mathbf{J}/c_0^2. \tag{8}$$

Since $\partial/\partial t = \partial/\partial t' - \mathbf{w} \cdot \nabla'$, it follows that Eqs. (1)–(5) are also satisfied in the moving frame.

Conservation of charge is as fundamental as conservation of mass, and it may at first seem strange that, according to Eq. (8), ϑ' differs from ϑ unless **J** vanishes. This apparent paradox, which is also faced by the full Maxwell theory, may seem less mysterious when the operational problem of measuring charge is considered. Clearly, the measurement of ϑ by a probe inserted at **x** would be corrupted unless **J** were zero there and if, in an attempt to prevent current flowing continuously into the probe, a small domain, of volume v, containing **x** were isolated—for example, by surrounding the domain with a thin insulating sheath—a charge distribution would be set up on the sheath to reflect the incoming current and the charge contained by v would no longer be $v\vartheta$; i.e., ϑ would be corrupted. Perhaps the clearest demonstration that pre-Maxwell theory does not conflict with charge conservation is the evaluation in both reference frames of the net charge carried by an isolated body V surrounded by an insulator. It is easily shown [see Eq. (28)] that the difference between the surface charge densities Θ' and Θ on the boundary S of V as evaluated in the two reference frames exactly compensates for the difference (8) between ϑ' and ϑ so that the net charge carried by the body is the same in both frames even though ϑ and Θ differ in each.

The pre-Maxwell equations must be supplemented by constitutive relations, i.e., relations that specify the constitution of the conductor. We assume that the conductor is isotropic so that, if we transform to a reference frame moving with the velocity $\mathbf{w} = \mathbf{V}(\mathbf{x})$ of a particular point **x** in the conductor, the

Table 1. Typical values of parameters in MHD systems.[a]

System	Ionized H	Hg	Na	Nak	Earth's core	Solar convection zone	Solar corona	Magnetic stars	Galactic disk
Length scale, $\mathcal{L}$ (m)	0.1	0.1	0.1	10^{-1}	10^{6}	10^{8}	10^{8}	10^{9}	10^{20}
Velocity scale, $\mathcal{V}$ (m/s)	1	0.1	0.1	0.1	10^{-4}	10^{2}	10^{5}	10?	10^{-4}
Field magnitude, $\mathcal{B}$ (T)	0.1	1	1	1	10^{-2}	10^{-1}	10^{-4}	1	10^{-9}
Typical density, ρ (kg/m^3)	10^{-7}	10^{4}	10^{3}	9×10^{2}	10^{4}	10^{-2}	10^{-15}	10^{3}?	10^{-21}
Magnetic diffusivity, η (m^2/s)	15	1	8×10^{-2}	0.3	1	(10^{7})	1	1	(10^{24})
Kinematic viscosity, ν (m^2/s)	10	10^{-7}	8×10^{-7}	6×10^{-7}	10^{-2}?	(10^{7})	10^{16}	10^{-6}?	(10^{24})
Electromagnetic diffusion time, $\tau_\eta = \mathcal{L}^2/\eta$(s)	7×10^{-4}	10^{-2}	10^{-1}	3×10^{-2}	10^{12}	(10^{9})	10^{16}	10^{18}	(10^{16})
Magnetic response time, $\tau_m = \eta/\mathcal{V}_A^2$ (s)	10^{-10}	10^{-2}	9×10^{-5}	3×10^{-4}	10^{2}	(10)	10^{-12}	10^{-2}	(10^{16})
Alfvén velocity (m/s) $\mathcal{V}_A = \mathcal{B}/\sqrt{\mu_0\rho}$	3×10^{5}	8	30	30	0.1	10^{3}	10^{6}	10	10^{4}
Magnetic Prandtl number, $p_m = \nu/\eta$	0.7	10^{-7}	10^{-5}	2×10^{-6}	10^{-2}?	(1)	10^{16}	10^{-6}?	(1)
Hartmann number, $M = \mathcal{V}_A\mathcal{L}/(\nu\eta)^{1/2}$	2×10^{3}	3×10^{3}	10^{4}	7×10^{3}	10^{-6}?	(10^{4})	10^{6}	10^{13}	(1)
Magnetic Reynolds number, $R = \mathcal{V}\mathcal{L}/\eta$	7×10^{-3}	10^{-2}	1	3×10^{-2}	10^{2}?	(10^{3})	10^{13}	10^{10}	(1)
Lundquist number, $S = \mathcal{V}_A\mathcal{L}/\eta$	10^{3}	1	40	10	10^{5}	(10^{4})	10^{14}	10^{10}	(1)
Alfvén number, $A = \mathcal{V}/\mathcal{V}_A$	3×10^{-6}	10^{-2}	3×10^{-3}	3×10^{-3}	10^{-3}	0.1	0.1	1	1

---

[a] Astrophysical quantities are given representative values which in some cases (denoted by ?) are extremely uncertain. Quantities based on turbulent transport coefficients are in parentheses.

relations between fields at $\mathbf{x}$ are isotropic and are, in fact, the well-known relations of classical em theory:

$$\mathbf{H}' = \mathbf{B}'/\mu_0, \quad \mathbf{D}' = \epsilon\mathbf{E}', \quad \mathbf{J}' = \sigma\mathbf{E}'. \tag{9}$$

We assume in Eq. $(9)_1$ and throughout that the permeability of the fluid is that of free space; the greater generality of supposing otherwise carries too high a price in added complexity. Since $c_0^2 = (\mu_0\epsilon_0)^{-1}$, where ϵ_0 is the permittivity of free space, we find from Eqs. (6)–(9) that

$$\mathbf{H} = \mathbf{B}/\mu_0, \quad \mathbf{J} = \sigma(\mathbf{E} + \mathbf{V} \times \mathbf{B}), \tag{10, 11}$$
$$\mathbf{D} = \epsilon\mathbf{E} + (\epsilon - \epsilon_0)\mathbf{V} \times \mathbf{B}, \tag{12}$$

the first two of which are especially significant. Equation (11) expresses Ohm's law in a nonrelativistically moving conductor; the term $\mathbf{V} \times \mathbf{B}$ is known as the induced emf.

1.2 Stress and Energy

The nonrelativistic approximation has further significant consequences. The force per unit volume on the conductor due to the free charges it carries is $\vartheta\mathbf{E} = O(\epsilon\mathscr{E}^2/\mathscr{L}) = O(\epsilon\mathscr{V}^2\mathscr{B}^2/\mathscr{L})$ by Eqs. (4) and (12). Through current flow, it also experiences a *Lorentz force* $\mathbf{J} \times \mathbf{B} = O(\mathscr{B}^2/\mu_0\mathscr{L})$ by Eqs. (2) and (10). The ratio of the first force to the second is $O(\mathscr{V}^2/c_0^2)$ so that, with negligible error,* we have, for the total force per unit volume,

$$\mathbf{L} = \mathbf{J} \times \mathbf{B}. \tag{13}$$

This is often conveniently thought of in terms of Faraday's equivalent stresses:

$$L_i = \nabla_j m_{ij}, \tag{14}$$

where

$$m_{ij} = B_iB_j/\mu_0 - (B^2/2\mu_0)\delta_{ij} \tag{15}$$

are the components of the (magnetic) stress tensor. It follows that

$$\mathbf{L} = \mu_0^{-1}\mathbf{B} \cdot \nabla\mathbf{B} - \nabla(B^2/2\mu_0). \tag{16}$$

The term B_iB_j/μ_0 in Eq. (15) represents a tension in (magnetic) lines of force; the last term imparts an isotropic *magnetic pressure,* which is about 4 atm for a field of 1 T.

Because $\vartheta\mathbf{E}$ could be omitted from Eq. (13), the electric stress tensor has, to the same $O(\mathscr{V}^2/c_0^2)$ inaccuracy, no part to play in the total stress tensor. In a similar way and with the same $O(\mathscr{V}^2/c_0^2)$ error, the term $\vartheta\mathbf{V}$ representing convection of free charge could be omitted from Ohm's law (11) and the contribution $\frac{1}{2}\,\epsilon E^2$ made by the electric field to the total electromagnetic energy density E_M may be ignored. Thus,

$$E_M = B^2/2\mu_0 \tag{17}$$

is the magnetic field energy density. The Poynting flux $\mathbf{S} = \mathbf{E} \times \mathbf{H}$ of em energy is the same in both relativistic and nonrelativistic theories,

$$\mathbf{S} = \mathbf{E} \times \mathbf{B}/\mu_0, \tag{18}$$

and is $O(E_M\mathscr{V})$, corresponding to the advection of E_M by $\mathbf{V}$. The momentum of the electromagnetic field is negligible in pre-Maxwell theory.

The rate per unit volume at which Ohmic resistance heats the fluid is

$$\epsilon_\sigma = J^2/\sigma. \tag{19}$$

Of the three expressions for ϵ_σ used interchangeably in the familiar em theory of stationary conductors, namely, J^2/σ, σE^2, and $\mathbf{J} \cdot \mathbf{E}$, only J^2/σ is correct in MHD. Note, however, that, by Eqs. (11) and (13),

$$\mathbf{J} \cdot \mathbf{E} = \epsilon_\sigma + \mathbf{V} \cdot \mathbf{L} \tag{20}$$

gives the em-power loss per unit volume through both Ohmic dissipation and the rate of working of the Lorentz force. Although ϵ_σ is non-negative, $\mathbf{V} \cdot \mathbf{L}$ can have either sign and, if negative, represents the generation of em energy from kinetic energy.

1.3 Induction Equation; Perfect Conductors

Equation (4) serves only to determine the charge density in the conductor once $\mathbf{E}$ has been determined by solving the remaining MHD equations. It is of little interest, although it is worth noting that ϑ is nonzero in general. In the nonconducting exterior $\hat{V}$ of the conductor, $\hat{\vartheta}$ may be known (e.g., $\hat{\vartheta} = 0$ if $\hat{V}$ is a vacuum), and Eq. (4) is then needed to determine $\hat{\mathbf{E}}$; but $\mathbf{E}$ plays such a minor role in MHD compared with $\mathbf{B}$ that one usually speaks of $\mathbf{B}$ as "the field," the adjective "magnetic" being implied. The emphasis is always to determine $\mathbf{B}$ and $\mathbf{V}$, and, once these have been found, the solution is seldom completed by the evaluation of $\mathbf{E}$ and ϑ. For this reason, we shall consider ϑ no further and eliminate $\mathbf{E}$ in favor of $\mathbf{B}$ wherever possible.

---

*Strictly speaking, before claiming this, we should also require that the Debye length be small compared with all other lengths of interest.

For example, by Eqs. (2), (10), and (11),

$$\begin{aligned}\mathbf{E} &= -\mathbf{V} \times \mathbf{B} + \mathbf{J}/\sigma \\ &= -\mathbf{V} \times \mathbf{B} + \eta\nabla \times \mathbf{B},\end{aligned} \tag{21}$$

and, substituting this into Eq. (1), using Eq. (3) and the vector identity $\nabla \times \nabla \times \mathbf{B} = \nabla(\nabla \cdot \mathbf{B}) - \nabla^2\mathbf{B}$, we obtain the *induction equation*:

$$\frac{\partial \mathbf{B}}{\partial t} = \nabla \times (\mathbf{V} \times \mathbf{B}) + \eta\nabla^2\mathbf{B}, \tag{22}$$

where η is the *magnetic diffusivity*,

$$\eta = 1/\mu_0\sigma, \tag{23}$$

which is constant because σ is constant by assumption.

A useful idealization, which will arise several times in subsequent sections, is the perfectly conducting fluid ($\sigma = \infty$; i.e., $\eta = 0$). In this case, Eqs. (21) and (22) reduce to

$$\mathbf{E} = -\mathbf{V} \times \mathbf{B}, \quad \frac{\partial \mathbf{B}}{\partial t} = \nabla \times (\mathbf{V} \times \mathbf{B}). \tag{24, 25}$$

According to Eqs. (15), (17), (18), and (24), in a perfect conductor,

$$\begin{aligned}&\mathbf{S} = \mathbf{B} \times (\mathbf{V} \times \mathbf{B})/\mu_0 \quad \text{or} \\ &S_i = E_M V_i - m_{ij} V_j,\end{aligned} \tag{26}$$

which has a clear interpretation.

1.4 Continuity Properties

To make the solutions of the electrodynamic equations unique, one must apply continuity conditions at any discontinuity in electrodynamic properties. These conditions rest on the same fundamental physical laws as do the differential equations. The laws take the form of integral statements. Equation (3) is the consequence of statement that the net flux of $\mathbf{B}$ from any closed volume is zero, and, applying this to a small "pill-box" spanning the discontinuity surface S, one finds that $\mathbf{n} \cdot \mathbf{B}$ must be continuous across S, where $\mathbf{n}$ is the unit normal to S directed from side 1 into side 2. Similarly, Eq. (4) is the consequence of the statement that the flux of $\mathbf{D}$ from a closed volume is equal to the net charge that that volume contains, and condition (28) also follows. Equation (5) follows from the demand that the net current leaving a closed volume is zero, and this also leads to conditions $(27)_2$ and (31). Equation (2) results from the requirement that the integral of $\mathbf{H}$ around any closed curve is equal to the current flowing through it. Applying this to a small rectangular loop with its long sides on either side of S, one obtains condition (30). Similarly, Eqs. (1), $(27)_3$, and (32) follow from the demand that the rate of change of magnetic flux through a closed curve is equal to the integral of $-\mathbf{E}$ around its perimeter (taken in the sense of Stokes's theorem).

If the fluid is not a perfect conductor ($\sigma \neq \infty$), $\mathbf{B}$ is everywhere differentiable, but $\nabla\mathbf{B}$, in general, is discontinuous on any surface S where the conductivity changes abruptly. One has

$$\begin{aligned}&[\![\mathbf{B}]\!] = 0, \quad [\![\mathbf{n} \cdot \mathbf{J}]\!] = 0, \\ &[\![\mathbf{n} \times \mathbf{E}]\!] = 0, \quad \text{on S},\end{aligned} \tag{27}$$

where $[\![Q]\!] = Q_2 - Q_1$ denotes the jump in a field Q across S. Although it is of no greater interest than ϑ, we may note that a free surface charge of density Θ per unit area exists on S given by

$$\Theta = [\![\mathbf{n} \cdot \mathbf{D}]\!], \quad \text{on S}. \tag{28}$$

Conditions (27) are not independent. The second follows from the first, and the third shows that, if

$$[\![\mathbf{n} \cdot \mathbf{B}]\!] = 0, \quad \text{on S}, \tag{29}$$

is obeyed initially, it will be obeyed for all time.

In a perfect conductor, $\mathbf{n} \times \mathbf{B}$ may be discontinuous, and

$$[\![\mathbf{n} \times \mathbf{H}]\!] = \mathbf{I}, \quad \text{on S}, \tag{30}$$

where $\mathbf{I}$ is a *surface current*. This is an infinitely thin sheet carrying along it a finite net current (per unit length across **I**). It may be thought of as the limit, as its thickness δ tends to zero, of a thin *current layer* in which **J** is $O(\mathcal{J}/\delta)$, i.e., very large. Continuity of charge now requires that

$$[\![\mathbf{n}\cdot\mathbf{J}]\!] = -\nabla_2\cdot\mathbf{I}, \quad \text{on S}, \tag{31}$$

where $\nabla_2\cdot$ denotes the surface divergence. Equation $(27)_3$ is also altered. Since $\mathbf{n}\times\mathbf{E}'$ is continuous in a frame moving with S, we have

$$[\![\mathbf{n}\times(\mathbf{E}+\mathbf{U}\times\mathbf{B})]\!] = 0, \quad \text{on S}. \tag{32}$$

If S separates two perfect conductors of different physical properties, then

$$V_{1n} = V_{2n} = U_n, \quad \text{on S}, \tag{33}$$

where $V_{\alpha n} = \mathbf{n}\cdot\mathbf{V}_\alpha$ and $\mathbf{V}_\alpha$ is the velocity on side α of S; Eqs. (24) and (33) reduce condition (32) to

$$B_n[\![\mathbf{V}]\!] = 0, \quad \text{on S}. \tag{34}$$

Thus, **V** is continuous when $B_n \equiv \mathbf{n}\cdot\mathbf{B} \neq 0$ on S. If, however, S is a shock, in which a material exchanges one thermodynamic state for another as it passes through S, Eqs. (33) and (34) do not hold [see Eqs. (54) and (59)].

2. FLUID MECHANICS

This section summarizes the theory governing two models of fluid behavior: the incompressible Newtonian (or linear viscous) fluid and the Euler fluid, which is compressible but inviscid; these models are generalized to the MHD context merely by the addition of the em body force **L** or the equivalent em stresses m_{ij}. Readers for whom this summary is unnecessary may wish to refer to the table of notation at the end of this article and then skip forward to Sec. 3.

2.1 The Ideal Fluid

The Euler fluid is often used to model fluid behaviors when there is reason to believe that viscous forces and thermal diffusion are insignificant. Ignoring these processes and also ignoring the possibility that the fluid is being internally heated, the equations expressing conservation of mass, momentum, and energy may be written as

$$\frac{D\rho}{Dt} = -\rho\nabla\cdot\mathbf{V}, \tag{35}$$

$$\rho\frac{D\mathbf{V}}{Dt} = -\nabla p + \mathbf{J}\times\mathbf{B} + \rho\mathbf{g}, \tag{36}$$

$$\frac{DS}{Dt} = 0, \tag{37}$$

where D/Dt stands for the motional derivative, i.e., the rate of change following the motion of the fluid:

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{V}\cdot\nabla; \tag{38}$$

g is an externally applied body force per unit mass, such as gravity. An alternative and useful way of writing Eq. (36) makes use of the representation (16) of $\mathbf{L} = \mathbf{J}\times\mathbf{B}$:

$$\rho\frac{D\mathbf{V}}{Dt} = -\nabla P + \frac{\mathbf{B}\cdot\nabla\mathbf{B}}{\mu_0} + \rho\mathbf{g}, \tag{39}$$

where

$$P = p + B^2/2\mu_0 \tag{40}$$

is the *total pressure*, i.e., the sum of thermodynamic and magnetic pressures.

Equation (37) excludes entropy production, such as might have arisen from dissolved heat sources (such as radioactivity) if present and would certainly arise from the diffusion of heat and momentum had we not already excluded those processes. When the Euler fluid is generalized to become the *ideal fluid* of MHD, the entropy source ϵ_σ/T, where T is the (absolute) temperature, must also be excluded from Eq. (37), and this requires that the Ohmic heating ϵ_σ, and therefore η, must be negligible; i.e., the fluid is a perfect electrical conductor. Thus, the ideal fluid is defined by*

$$\nu = \kappa = 0 \quad \text{and} \quad \sigma = \infty, \tag{41}$$

---

*Strictly, we should also add to Eqs. (41) the demand that the bulk viscosity is zero. The effects of bulk viscosity are excluded from this article.

where ν is the kinematic viscosity and κ is the thermal diffusivity.

The concept of an ideal fluid is not entirely straightforward. It is always advisable to consider Eqs. (41) as a limit: $\nu \to 0$ and $\eta \to 0$, with their ratio, the *magnetic Prandtl number,*

$$p_m = \nu/\eta = \mu_0 \sigma \nu, \tag{42}$$

held fixed. We shall recognize in Sec. 3.3 that, even though the ideal equations do not involve ν and η, their solutions may be required to depend on p_m.

The adiabatic condition (37) ensures that the entropy of each fluid element remains at its initial value, which, however, may differ from one fluid element to the next. Two alternative expressions of Eq. (37) follow from the thermodynamic relations

$$\begin{aligned} dp &= a^2 d\rho + (\gamma T/K_p C_v) dS, \\ dI &= (p/\rho^2) d\rho + T dS, \end{aligned} \tag{43}$$

where I is the internal energy per unit mass, a is the (adiabatic) speed of sound, γ is the coefficient of volume expansion, K_p is the isothermal compressibility, and C_v is the specific heat at constant volume. By Eqs. (35), (37), and (43), we have

$$\frac{Dp}{Dt} = -a^2 \rho \nabla \cdot \mathbf{V}, \tag{44}$$

$$\frac{DI}{Dt} = -\frac{p}{\rho} \nabla \cdot \mathbf{V}. \tag{45}$$

Equations (35), (36), (44), and (25), together with Eq. (2) and appropriate boundary conditions (see Sec. 2.2), suffice to determine the evolution of ρ, $\mathbf{V}$, p, and $\mathbf{B}$.

The situation is yet simpler when the system is *isentropic*, i.e., when S is the same for all $\mathbf{x}$. Then the equation of state $p = p(\rho, S)$ reduces to $p = p(\rho)$, i.e., the fluid is *barotropic*, and Eq. (36) becomes

$$\frac{D\mathbf{V}}{Dt} = -\nabla \varpi + \frac{\mathbf{J} \times \mathbf{B}}{\rho} + \mathbf{g}, \tag{46}$$

where ϖ is the *reduced pressure*:

$$\varpi(\rho) = \int_0^{\rho} \rho^{-1} dp(\rho). \tag{47}$$

Since $dH = TdS + \rho^{-1} dp$, where $H = I + p/\rho$ is the enthalpy per unit mass, we see that ϖ essentially coincides with H in an isentropic system.

The conservation laws (35)–(37) have useful alternative expressions:

$$\begin{aligned} &\frac{\partial \rho}{\partial t} = -\nabla \cdot \mathbf{M}, \quad \frac{\partial M_i}{\partial t} = -\nabla_j \Pi_{ij}, \\ &\frac{\partial E}{\partial t} = -\nabla \cdot \mathbf{Q}, \end{aligned} \tag{48}$$

where $\mathbf{M} = \rho \mathbf{V}$ is the momentum density, E is the total energy density per unit volume, Π_{ij} is the total stress tensor, and $\mathbf{Q}$ is the total energy flux. With the help of the alternative form (45) of Eq. (37), we find that

$$E = \rho(I + \tfrac{1}{2} V^2) + B^2/2\mu_0, \tag{49}$$

$$\Pi_{ij} = \rho V_i V_j + P \delta_{ij} - B_i B_j/\mu_0, \tag{50}$$

$$\mathbf{Q} = \rho(H + \tfrac{1}{2} V^2)\, \mathbf{V} + \mathbf{E} \times \mathbf{B}/\mu_0. \tag{51}$$

In deriving Eqs. (48), we have assumed that $\mathbf{g} = 0$. When $\mathbf{g} \neq 0$, not only must $\rho \mathbf{g}$ be restored to the right-hand side of Eq. $(48)_2$ but an extra term $\rho \mathbf{V} \cdot \mathbf{g}$ must be added to the right-hand side of Eq. $(48)_3$ representing the rate of working of the external forces. If, however, $\mathbf{g} = -\nabla \varphi$ is a *time-independent* conservative force, Eq. $(48)_3$ holds once again, but with E and $\mathbf{Q}$ redefined as

$$E = \rho(I + \tfrac{1}{2} V^2 + \varphi) + B^2/2\mu_0, \tag{52}$$

$$\mathbf{Q} = \rho(H + \tfrac{1}{2} V^2 + \varphi)\mathbf{V} + \mathbf{E} \times \mathbf{B}/\mu_0. \tag{53}$$

In Eq. (52), we see the sum of internal, kinetic, and em energies and the potential energy (φ) of the applied forces; in Eq. (53), we see enthalpy, kinetic energy, and external potential energy advected with the fluid motion, while em energy is transmitted by the Poynting flux. Provided no energy is being injected from outside a system (e.g., from external current-carrying coils), its total energy (the integral of E) is conserved. This forms the basis of an energy principle described in Sec. 8.

2.2 Continuity Conditions for an Ideal Fluid; Shocks

The assumptions (41) defining the ideal model permit discontinuities (including shocks) to arise at a stationary or moving surface S in the fluid. For convenience, let us transform to a reference frame in which $U_n = 0$. The integral forms of Eqs. (48) tell us that M_n, Π_{in}, and Q_n are continuous across S. The first of these gives

$$\rho_1 V_{1n} = \rho_2 V_{2n} = j, \quad \text{say}. \tag{54}$$

The second, when separated into components parallel and perpendicular to **n**, gives, by Eq. (50),

$$[\![p]\!] + j^2[\![\rho^{-1}]\!] + [\![B_H^2/2\mu_0]\!] = 0, \tag{55}$$

$$j[\![\mathbf{V}_H]\!] = B_n[\![\mathbf{B}_H/\mu_0]\!], \tag{56}$$

where $\mathbf{V}_H = \mathbf{V} - V_n\mathbf{n}$ and $\mathbf{B}_H = \mathbf{B} - B_n\mathbf{n}$. The third gives, by Eq. (51),

$$j\left[\!\!\left[H + \frac{j^2}{2\rho^2} + \tfrac{1}{2}V_H^2 + \frac{B_H^2}{\mu_0\rho}\right]\!\!\right] = B_n\left[\!\!\left[\frac{\mathbf{V}_H\cdot\mathbf{B}_H}{\mu_0}\right]\!\!\right]. \tag{57}$$

In writing Eqs. (55)–(57), we have made use of two results from Sec. 1.4: first, by Eq. (29),

$$B_{1n} = B_{2n} = B_n, \quad \text{say}; \tag{58}$$

second, by Eq. (24), continuity of $\mathbf{n} \times \mathbf{E}$ requires that

$$B_n[\![\mathbf{V}_H]\!] = j[\![\mathbf{B}_H/\rho]\!]. \tag{59}$$

When $j \neq 0$ and $B_n \neq 0$, the discontinuity is an MHD shock and Eqs. (54)–(59) are the MHD generalizations of the *Rankine–Hugoniot conditions* of classical fluid mechanics. And, as in that subject, they must be supplemented by the demand that fluid passing through the shock cannot suffer a diminution in entropy; *i.e.*,

$$j[\![S]\!] \geq 0. \tag{60}$$

Thus, the basis of Eq. (37), conservation of entropy, generally fails as fluid passes through S. This is called *shock heating*.

There are three principal types of discontinuity in which shock heating does not occur, either because $[\![S]\!] = 0$ or because $j = 0$. The former possibility arises in the so-called *rotational discontinuity* in which $[\![S]\!] = [\![\rho]\!] = 0$ so that $[\![H]\!] = [\![p]\!] = 0$. By Eq. (54), $[\![V_n]\!] = 0$ and Eqs. (56) and (59) show that

$$j = \pm\left(\frac{\rho}{\mu_0}\right)^{1/2} B_n, \qquad [\![\mathbf{V}_H]\!] = \pm\frac{1}{\sqrt{\mu_0\rho}}[\![\mathbf{B}_H]\!]. \tag{61}$$

We shall recognize in Sec. 6.2 that the rotational discontinuity is merely a finite-amplitude Alfvén wave with a discontinuous wave form.

When $j = 0$ but $B_n \neq 0$, we see from Eqs. (56)–(59) that

$$[\![p]\!] = 0, \quad [\![\mathbf{V}]\!] = 0, \quad [\![\mathbf{B}]\!] = 0, \tag{62}$$

though the values of ρ, S, T, ... may jump across such a *contact discontinuity*. Because $[\![S]\!] \neq 0$, in general, the contact discontinuity is one example of an entropy wave. The other is the *tangential discontinuity* for which $j = 0$ and $B_n = 0$. Then the values of $\mathbf{V}_H$, $\mathbf{B}_H$, ρ, S, T, ... may jump in crossing S, and we require only that

$$[\![P]\!] = 0. \tag{63}$$

We return to the topic of shock waves in Sec. 6.2.

Contact and tangential discontinuities also arise when S is a surface separating materials of different physical properties, such as two immiscible fluids of different composition. Conditions (62) or (63) apply depending on whether $B_n \neq 0$ or $B_n = 0$; interfacial tension may also have to be included. If one side of S is a rigid solid, conditions on the stresses are meaningless: a rigid solid can balance fluid-created stresses on its boundary by accelerating as a solid body and by *infinitesimal* internal deformations. Conditions $(62)_1$ and (63) then cease to apply, although the integral of the fluid stresses involved may have to be included when determining how the rigid body moves. Condition $(62)_1$ or (63) applies when one side of S is a vacuum, p then being zero in the vacuum.

2.3 The Incompressible Viscous Fluid

Using the imprecise terminology common in fluid mechanics, we say that a fluid is *incompressible* when its density is unaffected both by changing pressure and by changing temperature. Since then $D\rho/Dt = 0$, the expression (35) of mass conservation becomes

$$\nabla \cdot \mathbf{V} = 0. \tag{64}$$

Equation (64) is equivalent to assuming that the sound speed a is infinite; shocks, therefore, do not exist in an incompressible fluid, although if $\nu = 0$, discontinuities in $\mathbf{V}_H$ are allowed; if $\kappa = 0$, discontinuities in T are allowed; and if $\eta = 0$, discontinuities in $\mathbf{B}_H$ are allowed. The assumption $D\rho/Dt = 0$ only ensures that the density of each fluid element remains the same and not that the density is uniform; however, we will suppose that ρ is the same everywhere.

Approximation (64) simplifies the mathematical expressions for the viscous forces and the viscous dissipation of heat, ϵ_ν. Assuming that $\nu(\geq 0)$ is constant, conservation of momentum (36) may be written as

$$\frac{D\mathbf{V}}{Dt} = -\nabla\varpi + \frac{\mathbf{J} \times \mathbf{B}}{\rho} + \nu\nabla^2\mathbf{V} + \mathbf{g}, \tag{65}$$

where now $\varpi = p/\rho$. The rate at which kinetic energy is lost in heating the fluid through friction is

$$\epsilon_\nu = \tfrac{1}{2}\rho\nu(\nabla_i V_j + \nabla_j V_i)(\nabla_i V_j + \nabla_j V_i) \tag{66}$$

and is clearly non-negative.

Energy conservation plays a subservient role in the theory of incompressible fluids. The continuity, momentum, and induction equations can be solved for $\mathbf{V}$, $\mathbf{B}$, and p; and the energy equation may subsequently be solved if it is of interest to do so, e.g., if the temperature of the fluid resulting from viscous, Ohmic, and external heating is required. For this reason, it is unnecessary to set down the energy equation here.

If $\nu \neq 0$, $\mathbf{V}$ is everywhere continuous and differentiable, but $\nabla\mathbf{V}$ will, in general, jump in value in crossing a surface S where the material properties change abruptly. The *no-slip condition* $(62)_2$ applies both when S is the boundary of a solid and when it is an interface between two immiscible viscous fluids. In the latter case, viscous stresses must be included in condition (50) and so

$$[\![P\delta_{ij} - B_iB_j/\mu_0 - \rho\nu(\nabla_i V_j + \nabla_j V_i)]\!]n_j = 0. \tag{67}$$

If $\eta \neq 0$, so that $[\![\mathbf{B}]\!] = 0$, condition (67) simplifies to

$$[\![p\delta_{ij} - \rho\nu(\nabla_i V_j + \nabla_j V_i)]\!]n_j = 0. \tag{68}$$

Again, corrections for interfacial tension may have to be included in conditions (67) and (68).

3. STEADY PARALLEL FLOWS

It is natural to ask how the Lorentz force affects well-known and well-understood situations encountered in classical fluid mechanics. (When we use the word "classical," we imply that magnetic fields are absent so that electrical conduction is irrelevant.) The simplest examples provide the most revealing demonstrations of how drastically MHD flows can differ from classical flows. This section will exhibit this through the study of simple, steady unidirectional flows of the form

$$\mathbf{V} = V(y, z)\mathbf{e}_x. \tag{69}$$

Moreover, laboratory investigations on the flow of liquid metals, particularly mercury, have provided support for the theoretical findings. Theory and experiment have significant technological applications to power generation, plasma propulsion, and the flow of liquid metals in fast reactors and in the blankets surrounding thermonuclear reactors.

One of the simplest cases is one-dimensional (1D) *plane Poiseuille flow*, for which

$$V(y) = (k/2\rho\nu)(a^2 - y^2). \tag{70}$$

This classical solution of the Navier–Stokes equations depends on y alone and results from the application of the pressure gradient $-k\mathbf{e}_x$ to an incompressible fluid contained between fixed parallel walls, $y = \pm a$. The profile of the velocity (70) is said to be "parabolic" for obvious reasons; see the uppermost curve of Fig. 1. The mean velocity,

$$\bar{V} = \frac{1}{2a}\int_{-a}^{a} V(y)dy, \tag{71}$$

is proportional to k: it is $\bar{V}_0 = ka^2/3\rho\nu$.

The MHD flow corresponding to solution (70) arises when the fluid is an electrical conductor and a uniform magnetic field, $\mathbf{B}_0 =$

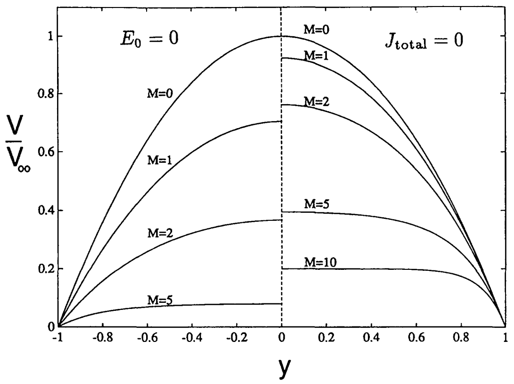

FIG. 1. Velocity profiles for Poiseulle flow ($M = 0$) and for Hartmann flow ($M \neq 0$). Velocity profiles $V(y)$ are shown for various Hartmann numbers M in two cases: $E_0 = 0$ (perfectly conducting walls) on left, and $J_{\text{total}} = 0$ (insulating walls) on right. Here, $V_{00} = ka^2/2\rho\nu$ is the velocity at $y = 0$ for $M = 0$. Note the reduction and flattening of the profiles as M, a measure of the applied field, increases.

$B_0\mathbf{e}_y$, is externally applied perpendicularly to the layer. In studying the resulting *Hartmann flow,* we shall suppose, for simplicity, that the "top and bottom walls" $y = \pm a$ are insulators.

Another very simple 1D classical solution of the Navier–Stokes equation is *plane Couette flow,* in which $k = 0$ but in which the fluid attaches to the walls $y = \pm a$, which move with velocity $\pm U\mathbf{e}_x$, so creating the velocity $\mathbf{V} = (Uy/a)\mathbf{e}_x$ between them. The corresponding MHD flow will not be studied here (e.g., see Sec. II.3 of Moreau, 1990).

3.1 Hartmann Flow

One of the most profound differences between classical and MHD flows arises from the long-range nature of em forces. For example, plane Poiseuille flow provides a good approximation to the flow far from the "side walls" $z = \pm c$ of a duct having a rectangular cross section ($-a \leq y \leq a$; $-c \leq z \leq c$) of large aspect ratio c/a (see Fig. 2). In contrast, the effect of the side walls on Hartmann flow pervades the entire duct. To see this, note that the flow (69) produces an emf, $\mathscr{E} = \mathbf{V} \times \mathbf{B} = VB_0\mathbf{e}_z$, which will, in general, create charges on the side walls and an electrostatic field $\mathbf{E}$. In the case of the 1D flow (70), $\mathbf{E}$ is parallel or antiparallel to $\mathscr{E}$, so that $\mathbf{E} = E(y)\mathbf{e}_z$. Since conditions are steady, Eq. (1) becomes $\nabla \times \mathbf{E} = 0$ so that $E(y)$ is a constant ($-E_0$, say). The value of E_0 is decided by whether the side walls are insulating or conducting and (if the latter) what electrical circuitry connects them to one another. Thus, V is *not* uniquely determined by B_0 and k (where k is now the gradient of *total* pressure in the x direction); it depends through E_0 also on the side walls. At one extreme, we shall focus on the case in which the side walls are perfect conductors that are connected together by a wire of zero resistance so that $E_0 = 0$; at the other extreme, we shall consider insulating side walls so that the net current

$$J_{\text{total}} = \int_{-a}^{a} J_z dy \tag{72}$$

carried across the duct is zero. In this case, $E_0 = \bar{V}B_0$ and a measurement of the potential difference $2cE_0$ between the side walls directly meters $\bar{V}$ noninvasively, making it a potentially interesting way of, for example, measuring blood flow in the body (see Shercliff, 1962). Perhaps the first flow meter to make use of induced currents was that of

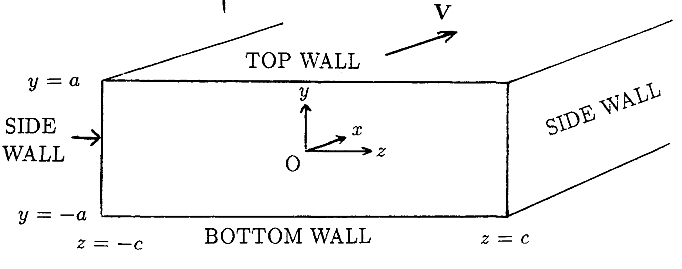

FIG. 2. The rectangular duct. When the aspect ratio c/a of the duct is large, plane Poiseuille flow ($\mathbf{B} = 0$) is a good approximation for $\mathbf{V}$ sufficiently far from the side walls. In the case of Hartmann flow (the MHD generalization of Poiseuille flow in which $\mathbf{B}_0 \neq 0$ is parallel to the side walls), the solution throughout the duct depends on the electrical properties of the side walls.

Faraday, who in 1843 measured flow in the Thames from the potential difference of its banks, an experiment which, like that of Ritchie mentioned below, demonstrated one-half of the MHD relationship and which is sometimes described as the first MHD experiment.

The general form of Hartmann flow is easily derived from the theory of Secs. 1 and 2 and will not be given here (e.g., see Sec. II.2 of Moreau, 1990). The solution depends crucially on the *Hartmann number*

$$M = (\sigma/\rho\nu)^{1/2} B_0 a, \qquad (73)$$

which measures the relative sizes of the Lorentz force $\mathbf{J} \times \mathbf{B} \approx \sigma(\mathbf{V} \times \mathbf{B}) \times \mathbf{B} \approx \sigma\mathcal{B}^2\mathcal{V}$ and the viscous force $\rho\nu\nabla^2\mathbf{V} \approx \rho\nu\mathcal{V}/a^2$; the ratio of these terms is M^2, a parameter also called the *Chandrasekhar number*. The dependence of the velocity profile on the magnetic field is illustrated in Fig. 1. When M is small, the profile differs little from profile (70), but, as M increases, the flow is increasingly suppressed, the effect being more pronounced at the $E_0 = 0$ extreme (shown on the left of Fig. 1) than at the $J_{\text{total}} = 0$ extreme (on the right). This is also illustrated in Fig. 3, where $\bar{V}$ is shown in each extreme case; here, $\bar{V}_0$ is the mean velocity for zero M (see above). Another striking feature of Fig. 1, which will be explicated in Sec. 3.2, is the way that the velocity profile flattens as M increases.

In the general case, the Poynting energy flux (per unit xz area) out of the duct through its top and bottom surfaces is

$$S = (E_0/\mu_0)[b(-a) - b(a)] = E_0 J_{\text{total}}. \qquad (74)$$

If $S < 0$, the duct converts externally provided em energy into kinetic energy; i.e., it is a pump. The applied pressure gradient, $-k\mathbf{e}_x$, can then be dispensed with, and flow is down or up the duct depending on whether E_0/B_0 is positive or negative. One example of an electromagnetic pump is sketched in Fig. 4: the annular linear induction pump. The cross section shown should be imagined rotated about the axis of symmetry. Only 12 of perhaps 30 windings are depicted; these carry ac currents that produce fields that close in the central laminated core and are phased to provide a magnetic field pattern that travels from left to right in the figure (The field directions shown apply, of course, only at one instant of time.) The induced currents flow in bands, as indicated, and the associated Lorentz force drives the liquid metal in the stainless steel pipe in the direction shown. For a recent discussion of the technology of electromagnetic pumping, see Baker and Tessier (1987). Ritchie performed an experiment in 1843 which may have been the first in which a fluid (water) was pumped electromagnetically, so demonstrating the other half of the MHD relationship.

If $S > 0$, the duct acts as a generator of electricity, i.e., a dynamo. This fact lies behind several interesting proposals for MHD generators, including a coal-fired open-cycle liquid-metal MHD energy conversion system, designed as a retrofit to a pre-existing conventional power plant. A liquid metal is mixed with the combustion gas and expanded across a transverse magnetic field. Electrical power is drawn off in the direction perpendicular to both **V** and **B**. After leaving this MHD generator (one of perhaps eight similar generators), the liquid metal is separated and purified before being returned to the mixer; the gas is directed to a conventional steam plant. The eight generators are connected in series to produce high-voltage dc, later converted to ac. Liquid copper is a good choice for the working metal, for not only is it compatible with the combustion gases but it also carries away the particulates in those gases, such as sulfur and nitrogen oxides, and the slag/ash. These can be easily and inexpensively removed so that no stack cleanup or slag/ash

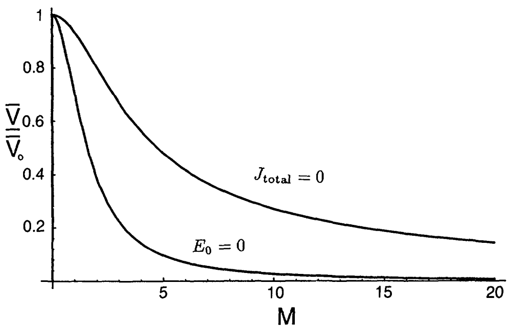

FIG. 3. Variation of mean velocity with increasing field strength in Hartmann flow. The variation of the mean flow down the duct $\bar{V}$ with Hartmann number M is shown in the two cases $J_{\text{total}} = 0$ (insulating walls) and $E_0 = 0$ (perfectly conducting walls). Here, $\bar{V}_0 = ka^2/3\rho\nu$ is the mean velocity for $M = 0$.

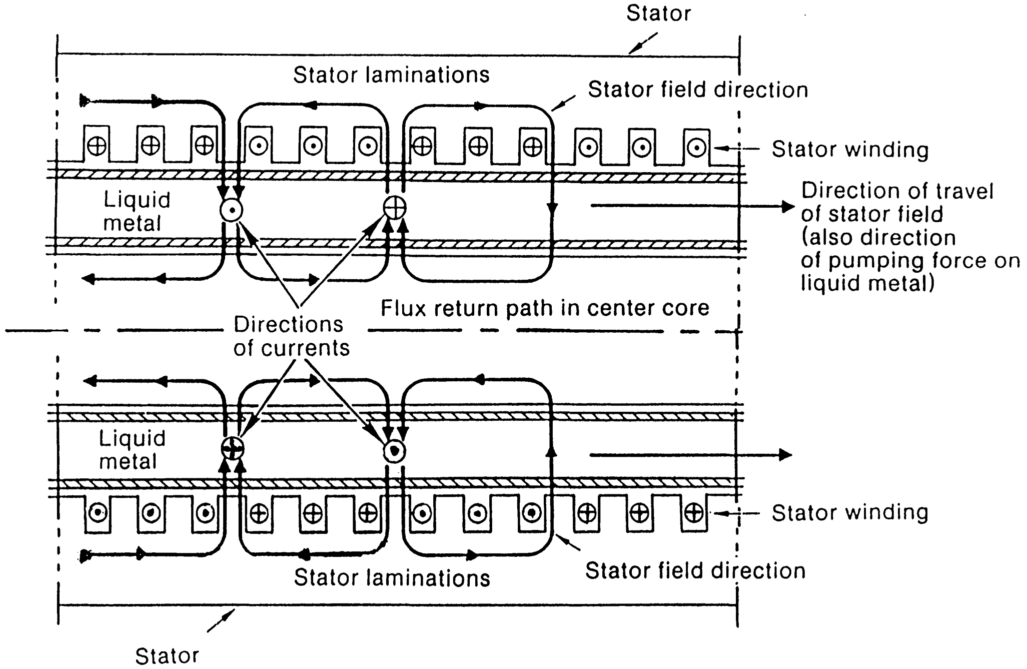
Stator
Stator laminations
Stator field direction
Stator winding
Liquid metal
Direction of travel of stator field (also direction of pumping force on liquid metal)
Flux return path in center core
Directions of currents
Liquid metal
Stator winding
Stator laminations
Stator field direction
Stator

FIG. 4. Schematic of an annular linear induction pump. The cross section sketched should be imagined rotated about the axis of symmetry and extended further to the right and left to include perhaps 30 stator windings. The Lorentz force on the induced currents shown drives liquid metal from left to right.

removal from the boiler is required; clearly the system is environmentally clean. See Pierson *et al.* (1983). In another type of MHD generator, the air is seeded before being mixed with coal and burnt so that the gas leaving the combustor is a plasma, i.e., has a high electrical conductivity. It is expanded across a magnetic field in an MHD generator in much the same way as described above [see Jackson (1983), where earlier references will be found]. Systems that make use of solar power have also been proposed.

It is clear from these examples that the MHD concepts so clearly exemplified by Hartmann flow have significant technological applications. But the added complexities of those applications require the theory to be developed further, as we shall do in the remainder of this section.

We should note first that to confine attention to the two special cases of insulating and perfectly conducting walls would be too restrictive for many applications, but to introduce finitely conducting walls would add complications connected with their geometry and spatial extent. A commonly adopted simplification supposes that the walls are very highly conducting but extremely thin, thinner in fact than every length scale of interest (even δ_H, see Sec. 3.2). Their integrated conductivity defines a conductance $\bar{S}$ that is supposed finite; $\bar{C} = \bar{S}/\sigma a$ is called the *conductance ratio*. The currents flowing in the layer act as a current sheet of strength $\mathbf{I}_{\text{total}} = \bar{S}\mathbf{E}_H$ by Ohm's law. Also, Eq. (11) gives $\mathbf{J}_H = \sigma\mathbf{E}_H$, because $[\![\mathbf{E}_H]\!] = \mathbf{V} = 0$ on the wall. (As in Sec. 2.2, the suffix H refers to components tangential to the boundary.) It follows that $\mathbf{I}_{\text{total}}/\bar{S} = \mathbf{J}_H/\sigma$. Moreover, condition (30) gives $\mu_0\mathbf{I}_{\text{total}} = \mathbf{n} \times \mathbf{b}$, where $\mathbf{n}$ points into the duct. Here, $\mathbf{b} \equiv \mathbf{B} - \mathbf{B}_0$ is evaluated on the fluid edge of the thin conducting layer; on the other edge, it is zero. We, therefore, have

$$\mathbf{n} \times \mathbf{b} = a\bar{C}\mu_0\mathbf{J}_H, \tag{75}$$

which, in the case of parallel flow, gives $b = a\bar{C}\mathbf{n}\cdot\nabla b$, so that $b(\pm a) = \mp\bar{S}\mu_0E_0$. It follows that

$$M(k - \bar{C}\sigma B_0E_0) = (k + \sigma B_0E_0)\tanh M,$$

which determines E_0.

3.2 Flow at Large Hartmann Number

Not only are large Hartmann numbers attainable in the laboratory, they are also relevant to many technological applications. The corresponding theory is also particularly amenable to solution, by the application of asymptotic methods (see, for example, Kulikovsky, 1973). Moreover, a knowledge of flows for large M can be helpful in qualitatively interpreting situations in which $M = O(1)$. A large-M solution is pictured as a main stream and boundary layers that match it to the walls, in much the same way as is routinely done at large Reynolds numbers,

$$Re = \mathcal{V}\mathcal{L}/\nu, \qquad (76)$$

in classical fluid mechanics. The nature of the main stream (or "core flow" as it is usually called in the MHD context) and of the boundary layers for large M is, however, very different from those arising classically for $Re \gg 1$.

Hartmann flow provides an ideally simple test bed for MHD asymptotics and for explaining the flattening of the velocity profile observed in Fig. 1. For uniform $\mathbf{B}_0$ (and for $\mathbf{g} = 0$), the steady forms of Eqs. (22) and (39) are to leading order

$$\mathbf{B}_0 \cdot \nabla\mathbf{V} = 0, \quad 0 = -\nabla p + \mathbf{J} \times \mathbf{B}, \qquad (77,78)$$

the latter of which gives

$$\mathbf{B}_0 \cdot \nabla\mathbf{J} = 0. \qquad (79)$$

Equations (77) and (79) are the basis of a *two-dimensional theorem*: If $M \gg 1$ and $\mathbf{B}_0$ is uniform, the fluid velocity and electric current density are independent of the coordinate parallel to $\mathbf{B}_0$. This theorem is analogous to the classical *Proudman–Taylor theorem* of rotating fluids, which states that the slow, steady motion of inviscid fluid in a rotating fluid is two-dimensional (2D) with respect to the direction of the angular velocity $\mathbf{\Omega}$ (see Sec. 10.4 and Greenspan, 1968).

Equation (78) is the equation of magnetostatic balance. It shows that a tiny $\mathbf{J}$ arises (tiny because $\mathbf{B}_0$ is large) for the sole purpose of creating the Lorentz force necessary to offset the applied pressure gradient. The large $\mathbf{B}_0$ can generate from a very small $\mathbf{V}$ the potential differences required to produce the necessary $\mathbf{J}$. In application to Hartmann flow, Eq. (78) gives at once $J_z^{\text{core}} = k/B_0$, from which it follows by Ampère's law (2) that $B_y^{\text{core}} = -k\mu_0 y/B_0$. Consistent with Eq. (77), Ohm's law (11) then gives a y-independent V, namely, $V^{\text{core}} = (k + \sigma B_0 E_0)/\sigma B_0^2$, so accounting for the flattening of the profiles seen in Fig. 1. Clearly, if $E_0 \neq 0$, then $V^{\text{core}} = E_0/B_0$ for large M, while if $E_0 = 0$, $V^{\text{core}} = k/\sigma B_0^2$.

It is clear that V^{core} does not satisfy the no-slip conditions on $y = \pm a$. This is rectified by thin *Hartmann layers*, across which viscous stresses diminish V monotonically from V^{core} to zero. The thickness δ_{H} of these layers may be estimated by equating the viscous force $\rho\nu\nabla^2\mathbf{V} = O(\rho\nu\mathcal{V}/\delta_{\text{H}}^2)$ to $\mathbf{J} \times \mathbf{B} = O(\mathcal{J}\mathcal{B}) = O(\sigma\mathcal{V}\mathcal{B}^2)$. In this way, we obtain

$$\delta_{\text{H}} = a/M = \sqrt{\rho\nu/\sigma B_0^2}. \qquad (80)$$

The same dynamical balance can be used to relate the discontinuities $\mathbf{B}_2 - \mathbf{B}_1$ and $\mathbf{V}_2 - \mathbf{V}_1$ in field and velocity from one side of the boundary layer to the other. Let us suppose that $\mathbf{B}_2 - \mathbf{B}_1$ is in the z direction, $\mathbf{B}_0$ being in the y direction as before. According to Eq. (30), the net current carried by the Hartmann layer is then $(B_{2z} - B_{1z})\mathbf{e}_x/\mu_0$, which creates from $\mathbf{B}_0$ a Lorentz force $B_0(B_{2z} - B_{1z})\mathbf{e}_z/\mu_0$ per unit area, that acts to accelerate the entire layer in the z direction. In a steady state, this must be balanced by the difference in viscous stress at the edges of the layer, which is of order $\rho\nu(\mathbf{V}_2 - \mathbf{V}_1)/\delta_{\text{H}}$. This must therefore be in the z direction, showing that $V_{2z} - V_{1z} = O(\delta_{\text{H}} B_0/\mu_0\rho\nu)(B_{2z} - B_{1z}) = (\text{sgn}B_0)O((\eta/\mu_0\rho\nu)^{1/2})(B_{2z} - B_{1z})$. This argument is supported by precise analysis (Stewartson, 1960):

$$\frac{\mathbf{B}^{\text{core}} - \mathbf{B}^{\text{wall}}}{\sqrt{\mu_0\rho}} = p_m^{1/2}(\text{sgn}B_n)(\mathbf{V}^{\text{core}} - \mathbf{V}^{\text{wall}}), \qquad (81)$$

where $p_m = \nu/\eta$ is the magnetic Prandtl number. In condition (81), we have specialized to the duct by identifying sides 1 and 2 with the wall and core, respectively. Also, we have generalized to ducts of arbitrary cross section by replacing B_0 by B_n, the component of $\mathbf{B}$ normal to the wall and directed into the duct; each side of the n component of condition (81) vanishes separately.

If we integrate J_z^{core} across the layer, we apparently obtain a net z-current flow of $2ak/$

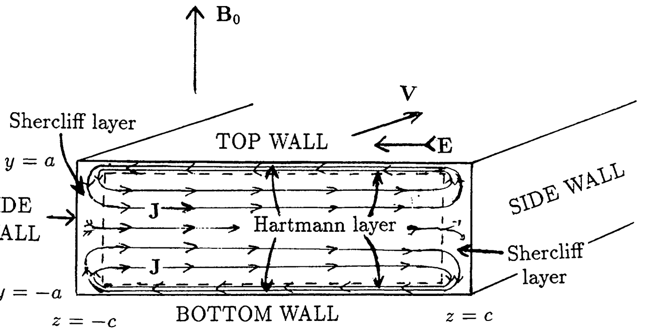

FIG. 5. An insulating rectangular duct when M is large. The Hartmann layers are shown on the top and bottom walls and the Shercliff layers on the side walls. The current generated in the core of the duct is reflected from the $z = c$ boundary and, driven by the electric field $\mathbf{E}$ generated by the free charges at the sides, returns equally along the Hartmann layers and is reflected at the $z = -c$ boundary to return to the core.

B_0, irrespective of the conductivity of the side walls, even though, in the case of insulating side walls, we know that $J_{\text{total}} = 0$. The current $2ak/B_0$ must therefore be totally reflected at the side walls and return in the Hartmann layers (see Fig. 5). Since V is much less than V^{core} in these layers, the current within them is driven primarily by $\mathbf{E} = -E_0\mathbf{e}_z$, which is indeed in the correct (negative z) direction and is of the right magnitude if $-J_z\delta_H \approx \sigma E_0 \delta_H \approx ak/B_0$. It follows that E_0 is independent of B_0 and M; in fact, $E_0 \approx ka/\sqrt{\rho\nu\sigma}$. Thus, V and $\bar{V}$ are proportional to M^{-1} for $M \gg 1$. In contrast, they are clearly proportional to M^{-2} in the case ($E_0 = 0$) of perfectly conducting side walls when (see above) $V^{\text{core}} = k/\sigma B_0^2$. This accounts for the fact that the profiles on the left side of Fig. 1 are depressed, compared with the curves for corresponding M on the right. It also explains the much smaller values of $\bar{V}$ shown in Fig. 3. In the case of the insulating duct, the Lorentz force on the return current is, in apparent contradiction to Lenz's law, in the *positive x* direction. When the current returns elsewhere, it exerts no Lorentz force on the fluid and the flow speed for $E_0 = 0$ is then $O(M^{-1})$ times smaller.

3.3 The Hartmann Layer and Applications

Hartmann layers arise when $\mathcal{B}$ is large in far more general situations than those considered in Sec. 3.2. And, even when $\mathcal{B}$ is not large, insight is gained from solutions for $M \gg 1$. A Hartmann layer forms on a wall wherever $|B_n|$ is large compared with $\sqrt{\rho\nu/\sigma\mathcal{L}^2}$. Its thickness is given by Eq. (80), with B_n replacing B_0, i.e., $\delta_H = \sqrt{\rho\nu/\sigma B_n^2}$. It is, therefore, a function of position on the walls and tends to infinity as the curves (or surfaces) on which $B_n = 0$ are approached, such as the points T_1 and T_2 shown in Fig. 6. The infinity is, however, illusory, since the Hartmann layers must themselves match to different boundary-layer structures that exist at the curves (or surfaces) where $B_n = 0$. These structures are passive and not of Hartmann layer type. They are hard to analyze, and fortunately they usually make only a small correction to the solution obtained by ignoring them.

Hartmann layers exist when all geometrical length scales $\mathcal{L}$ of interest are large compared with typical values of δ_H. They also arise in time-dependent situations and, provided that external conditions change little during the *magnetic response time* $\tau_m = \rho/\sigma\mathcal{B}^2$ and during the time scale $\nu\mu_0\rho/\mathcal{B}^2$, they can be treated as steady.

As an illustration of the usefulness of condition (81), we determine to leading order the flow in a duct with an insulating wall of arbitrary cross section but with just two points,

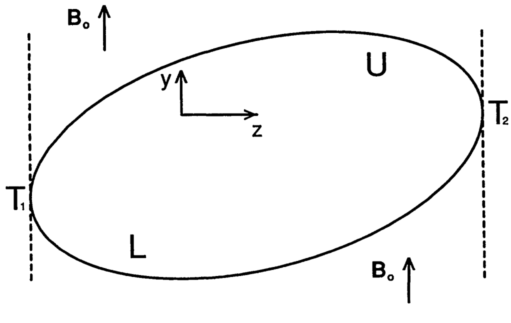

FIG. 6. Flow in a duct of arbitrary cross section. The points T_1 and T_2 where $\mathbf{B}$ is tangent to the cross section divide its perimeter into two segments: the upper part T_1UT_2, where condition (81) applies with $\text{sgn}B_n < 0$, and the lower part T_1LT_2, where $\text{sgn}B_n > 0$ in condition (81).

T_1 and T_2, at which $z = z_1$ and $z = z_2$ and $B_n = 0$ (see Fig. 6). Let the upper boundary, T_1UT_2, have the equation $y = y_1(z)$, the lower boundary, T_1LT_2, being $y = y_2(z)$. To leading order, the core is in magnetostatic balance (78) and $J_z^{\text{core}} = k/B_0$, as before. Again, $\mathbf{E}$ is predominantly created by em induction: $\mathbf{E} \approx -B_0V\mathbf{e}_z$. Since $\nabla \times \mathbf{E} = 0$, we have

$$\frac{\partial b^{\text{core}}}{\partial y} = -k\mu_0/B_0, \quad \frac{\partial V^{\text{core}}}{\partial y} = 0, \tag{82}$$

so that

$$b^{\text{core}} = -(k\mu_0/B_0)y + f(z), \quad V^{\text{core}} = g(z), \tag{83}$$

where $f(z)$ and $g(z)$ are functions that remain arbitrary until condition (81) is applied at $y = y_1(z)$ and $y = y_2(z)$. This means that the role of the Hartmann layers is decisive; they *control* the core flow. In fact, condition (81), which applies at every point on the cross section (except T_1 and T_2), requires that

$$\begin{aligned} f(z) &= \frac{k\mu_0}{2B_0}[y_1(z) + y_2(z)], \\ g(z) &= \frac{k}{2B_0\sqrt{\rho\sigma\nu}}[y_1(z) - y_2(z)]. \end{aligned} \tag{84}$$

If we set $y_1 = -y_2 = a$, we recover our earlier results for Hartmann flow at large M. If we take $y_1 = -y_2 = \sqrt{a^2 - z^2}$, we obtain the core flow for a circular duct of radius a: to leading order,

$$\begin{aligned} b^{\text{core}} &= -\frac{k\mu_0}{B_0}y, \\ V^{\text{core}} &= \frac{k}{B_0\sqrt{\rho\sigma\nu}}\sqrt{a^2 - z^2}, \end{aligned} \tag{85}$$

where the origin is on the axis of the duct. Equation $(85)_2$ provides another example of the 2D theorem (77).

If $\mathbf{B}_0$ is parallel to one of the sides of a duct of rectangular cross section, the singularities T_1 and T_2 of Fig. 6 extend along the entire side walls and boundary layers are set up along their entire length. The thickness of these *Shercliff layers* (also called "side-wall layers") is $\delta_S = M^{-1/2}a$, and within them the core flow is brought to rest in obedience to the no-slip conditions on $z = \pm c$. The corresponding flux deficit depends on the electrical properties of the walls and is not necessarily positive. Indeed, in one striking case, the side walls may make a greater contribution to $\bar{V}$ than does the entire core. This is because the current $J^{\text{core}} = k/B_0$ flowing across the duct must, when the side wall is an insulator, be deflected within the Shercliff layer toward the top and bottom walls (see Fig. 5). Being then substantially parallel to $\mathbf{B}_0$, the Lorentz force it creates is greatly reduced and is insufficient to balance the applied pressure gradient, which must therefore be offset by the viscous force. This can only happen if the velocity shear within the Shercliff layer is large. This results in a jet of fluid along the side wall of speed $O(M^{-1})$ which makes a positive $O(M^{-3/2})$ contribution to $\bar{V}$. When the top and bottom walls of the duct are perfect conductors [so that the core flow is $O(M^{-2})$], the boundary layer "deficit" is $O(aM^{1/2}/c)$ times larger than the contribution made by the core.

Side-wall layers are essentially passive shear layers that smooth out discontinuities that are permitted by the 2D theorem, i.e., those in which $\mathbf{V}(x, z)$ and $\mathbf{J}(x, z)$ are different on either side of a surface $x = x(z)$, where Oy is parallel to $\mathbf{B}_0$. They arise in more general contexts than Hartmann flow (see Sec. 4). Their theory was initiated by Braginsky (1960) and developed by Waechter (1979). The velocity profile at a shear layer usually possesses an inflection point and may therefore be subject to inertial instabilities of the type described in, for instance, Ch. 4 of Drazin and Reid (1981) (see Reed and Picologlou, 1989).

Rectangular ducts have been studied both theoretically and experimentally for a variety of different electrical wall conditions. For a more detailed account, see Ch. IV of Moreau (1990). Flow parallel to a cylinder in the presence of a transverse magnetic field is also discussed in Sec. V.V of that book.

4. STEADY NONPARALLEL FLOWS

In the situations examined in Sec. 3, the induced field $\mathbf{b}$ is everywhere parallel to the flow $\mathbf{V}$, so that it does not contribute to the emf $\mathbf{V} \times \mathbf{B}$, which is therefore $\mathbf{V} \times \mathbf{B}_0$. In the situations now described, $\mathbf{V} \times \mathbf{b} \neq 0$, so that a further field $\mathbf{b}_1$ arises which, in turn, creates another field $\mathbf{b}_2$ and so on. The ratio of successive terms is $O(|\mathbf{b}|/B_0) = O(\mu_0\sigma\mathscr{V}\mathscr{L})$, so that the sum converges rapidly if $R \ll 1$, where

$$R = \mathcal{V}\mathcal{L}/\eta \tag{86}$$

is the *magnetic Reynolds number*. In this case, all fields except $\mathbf{B}_0$ and $\mathbf{b}_1$ may usually be safely ignored.

In the situations examined in Sec. 3, the inertial term in the equation of motion $\rho\mathbf{V} \cdot \nabla\mathbf{V}$ is identically zero. In what follows, this is not the case, and the inertial force influences the flow. The ratio of this to the viscous force is of order $Re = \mathcal{V}\mathcal{L}/\nu$, the usual *(hydrodynamic) Reynolds number*. If $Re \ll 1$, the inertial term is negligible in comparison with the viscous force, and if $N \gg 1$, where $N = M^2/Re = \sigma\mathcal{B}^2\mathcal{L}/\rho\mathcal{V}$ is the *interaction parameter*, the inertial term may be discarded in comparison with the Lorentz force. In short, if $Re \ll \max(1, M^2)$, inertial forces may be ignored. The situation is then much as in Sec. 3, and, in particular, if $M \gg 1$, viscous forces may be neglected except in Hartmann or shear layers. In what follows, we shall assume that $N \gg 1$, $M \gg 1$, and $R \ll 1$, since such situations are often the simplest and most illuminating. It is worth noting, however, that, even then, inertial forces may sometimes be significant. When duct flow changes direction abruptly at an elbow joint, N is, in effect, reduced near it and inertial forces usually cannot be ignored. Also, as mentioned in Sec. 3.3, shear layers are prone to inertial instability.

As a first example (Alpher *et al.*, 1960), consider steady flow under the pressure gradient $-k\mathbf{e}_x$ between horizontally infinite top and bottom boundaries that are insulating, except for a highly conducting disk that is inset into one of them. Far upstream and far downstream of the disk, $\mathbf{E} = -E_0\mathbf{e}_z$ and $\mathbf{J} = J\mathbf{e}_z$, as in Sec. 3; but $\mathbf{E}$ and $\mathbf{J}$ are distorted near the disk, which is an equipotential. The return current takes the path of least resistance and returns (where possible) through the disk rather than within the Hartmann layers. Where this happens, the flow is reduced by the $O(M^{-1})$ factor of Sec. 3. The plug flow, therefore, passes around the cylinder whose surface touches the perimeter of the disk and whose axis is parallel to $\mathbf{B}_0$, almost as though that cylinder were solid rather than fluid (Fig. 7). This is another striking example of the two-dimensionality imposed by a large $\mathbf{B}_0$. The surface of the cylinder is not infinitely sharp but is a free shear layer of the type described in Sec. 3.3. Its thickness would be $O(\delta_S)$ ($\propto M^{-1/2}$) were it not for instabilities that create turbulence that thickens the layer (see Sec. IV.V.4 of Moreau, 1990).

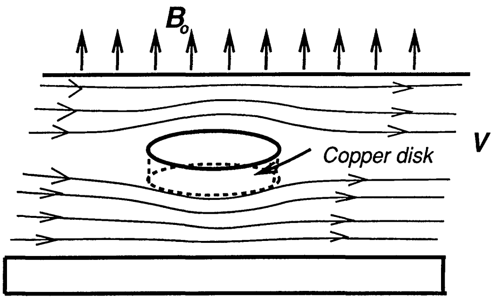

FIG. 7. Flow down an insulating rectangular duct in which the bottom wall contains a copper disk. When the applied field is strong, the flow is almost stagnant within a cylinder with axis parallel to $\mathbf{B}_0$ and passing through the perimeter of the disk, and the flow elsewhere passes around that cylinder, almost as though it were a solid obstacle.

As a second and similar example, consider flow past an insulating three-dimensional (3D) obstacle placed in an otherwise unidirectional flow between parallel insulating planes in the presence of a strong ($M \gg 1$) uniform applied field. Again, the flow is almost 2D and flows around the cylinder formed by the tangents to the obstacle that are parallel to the field, that cylinder being itself surrounded by a free shear layer that is singular on the curves of tangency. An expression for the drag on the obstacle can be derived theoretically and tested experimentally (see Sec. 5.5 of Moreau, 1990).

Numerous cases of practical importance involve ducts of varying cross section lying in spatially varying magnetic fields; e.g., $\mathbf{B}_0 = \mathbf{B}_0(x)$. Then $\mathbf{V}$, $\mathbf{b}$, and $\mathbf{J}$ also vary with x but in a way that can often be simply understood when $M \gg 1$ by applying a simple physical principle (see Sec. 6.3): Suppose that the changes in $\mathbf{B}_0$ and cross section are "abrupt" or, more precisely, that they occur over an x length short compared with $\mathcal{L}M^{1/2}$, where $\mathcal{L}$ is a cross-section dimension. Then, in the vicinity of the abrupt changes, streamlines and current lines follow *characteristic surfaces*. These are defined as surfaces composed of lines of force of $\mathbf{B}_0$ and are such that

$$\int \frac{dl}{B_0} = \text{constant}, \qquad (87)$$

where the integration is taken along the segment of the line of force lying within the duct. Each value of the constant in Eq. (87) defines a different characteristic surface.

As an application, suppose that the field applied across the circular duct considered in Sec. 3.3 depends on x. Since $\mathbf{B}_0$ is a potential field, we should not simply assume that $\mathbf{B}_0 = B_0(x)\mathbf{e}_y$; we should include a y-dependent x component in $\mathbf{B}_0$. So, for definiteness, suppose that the change in B_0 occurs in an interval $-a < x < a$ in which B_{0x} is nonzero and that $\mathbf{B}_0 = B_0(x)\mathbf{e}_y$ in $x > a$ is only half what it is in $x < -a$. The characteristic surfaces are planes of constant z in $|x| > a$. According to Eq. (87), the plane at $z = z'$ (say) for $x < -a$ becomes the plane $z = \frac{1}{2}\sqrt{3a^2 + z'^2}$ in $x > a$. Thus, the streamlines in the segment $-a < x < a$ move outward toward either T_1 (for $z' < 0$) or T_2 (for $z' > 0$); the situation in the xz plane is sketched in Fig. 8. Since the width of the characteristic surface is reduced by two, the flow speed must increase correspondingly, and the velocity profile becomes "M-shaped." The central streamline (shown dashed) splits into two streamlines, and according to Eq. (87), the fluid is stagnant between them, $\mathbf{V}$ being (weakly) singular as they are approached from the outside. In reality, since Eq. (87) is not exact, the flow is nowhere infinite and there are slowly moving trapped eddies between the streamlines. Of course, the profile must approach the solution (85), with the appropriately reduced B_0, as $x \to +\infty$; and, in fact, for a distance of order $aM^{1/2}$ both ahead of and behind the interval $-a < x < a$, there are regions of adjustment. For more details, see Ch. IV of Moreau (1990).

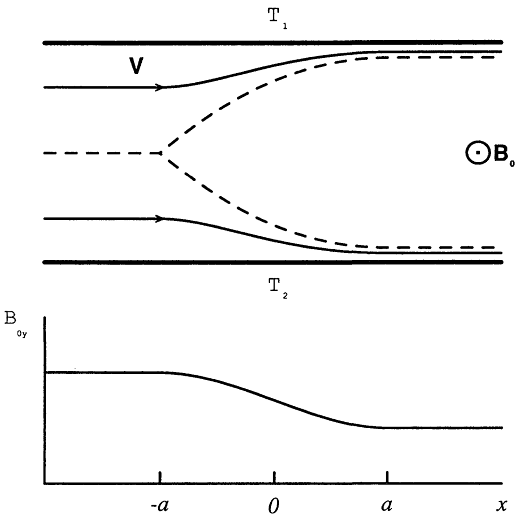

FIG. 8. Flow down a circular pipe when the applied field is halved. When the applied field is strong, the flow approximately follows the characteristic surfaces (87), and the streamlines therefore crowd toward the edges of the duct, leaving a stagnant area in the middle, so producing an M-shaped velocity profile.

5. HIGH FREQUENCIES; LEVITATION, HEATING, AND STIRRING

Suppose that current-carrying coils create in free space a nonuniform ac field, $\mathbf{B}_0 = \Re[\hat{\mathbf{B}}_0 \exp(i\omega t)]$, so that outside the coils, $\hat{\mathbf{B}}_0$ is a potential field ($\nabla \cdot \hat{\mathbf{B}}_0 = 0, \nabla \times \hat{\mathbf{B}}_0 = 0$). (In what follows, $\Re$, which stands for the real part of the expression it precedes, will be omitted but implied.) Now suppose that a solid block of material, V, is introduced. The field $\mathbf{B}_0$ will diffuse into V, and if $\omega = 0$, it will do so completely, the final field being everywhere the same as the initial field. If $\omega \neq 0$, the penetration of $\mathbf{B}_0$ will be less complete, and if $R_\omega \gg 1$, where

$$R_\omega = \omega \mathscr{L}^2/\eta \qquad (88)$$

is a new magnetic Reynolds number (sometimes called the "magnetic shielding parameter" to distinguish it from R), the field will penetrate the conductor only to a small *skin depth*,

$$\delta_\eta = (2\eta/\omega)^{1/2}. \qquad (89)$$

It is the strong induced current flowing in this skin that creates an induced field $\mathbf{b} = \tilde{\mathbf{b}} \times \exp(i\omega t)$ equal and opposite to $\hat{\mathbf{B}}_0 \exp(i\omega t)$ in V, i.e., which *expels* the applied field from V, apart from the thin skin on its surface S (see also Sec. 9.2). The component B_n of the total field $\mathbf{B} = \mathbf{B}_0 + \mathbf{b}$ normal to S is zero to a first approximation, and the tangential component $\mathbf{B}_H = \hat{\mathbf{B}}_H \exp(i\omega t)$ differs from the tangential component of $\mathbf{B}_0$. The skin currents are of order $\mathbf{J} = \mathbf{B}_H/\mu_0\delta_\eta$ and produce a field (**b**) that is comparable with $\mathbf{B}_0$ on S but that decreases with distance from V. They dissipate energy at an average rate $|\overline{J^2}|/\sigma \approx \sigma^{-1}(|\hat{\mathbf{B}}_H|/2\mu_0\delta_\eta)^2 \exp(-2z/\delta_\eta)$, and, integrated through the depth of the skin, they heat S at the rate $|\hat{\mathbf{B}}_H|^2/2\mu_0^2\sigma\delta_\eta$ per unit area.

On the outer (vacuum) side of the skin layer, the magnetic pressure is $B_H^2/2\mu_0$, and on the inner, field-free side, there is no magnetic stress. The net force per unit area on the skin, and hence on V, is therefore $B_H^2/2\mu_0$, and this has a nonzero mean: $|\hat{\mathbf{B}}_H|^2/4\mu_0$. When placed in a suitably chosen $\mathbf{B}_0$, the force can *levitate* the solid. This forms the basic idea of an almost perfect furnace, in which the metal is both levitated and heated by the field and is kept free of contamination from walls (see Fig. 9). (The composition of the atmosphere surrounding the furnace is carefully controlled.)

When the metal has been heated past its melting point, it might flow out of the levitating field and a challenging problem arises in *magnetohydrostatics*, namely, how to design the external coils and so "shape" $\hat{\mathbf{B}}_0$ that it can levitate a *fluid* conductor. This is a nontrivial free-boundary problem, i.e., to find S such that the following conditions are satisfied:

1. $\hat{\mathbf{B}}$ is everywhere tangential to S;
2. the interior V of S contains the correct volume of (incompressible) conductor;
3. the stresses across each element of S balance; i.e.,

$$\frac{|\hat{\mathbf{B}}_H|^2}{4\mu_0} + p_{\mathrm{atm}} = p_0 - \rho g z - \tilde{\gamma}\left(\frac{1}{R_1} + \frac{1}{R_2}\right). \tag{90}$$

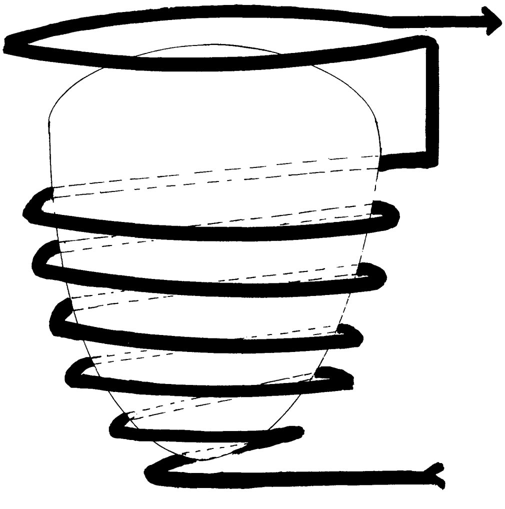

FIG. 9. Schematic of a levitated furnace. The metal is melted and supported by ac currents flowing in the lower coil and is stabilized by the current loop at the top.

Here, p_{atm} is the (constant) atmospheric pressure; p_0 is constant, $p_0 - \rho g z$ being the hydrostatic pressure in V; g is the acceleration due to gravity; z is height; $\tilde{\gamma}$ is the coefficient of surface tension; and R_1 and R_2 are the principal radii of curvature of S. Surface tension is required to prevent the metal from flowing out of the bottom of the furnace at which, since it lies on the axis of symmetry, $\mathbf{B} = 0$.

There is an equivalent, and very useful, variational way of finding S, namely, that of minimizing $W_g + W_\gamma - W_m$, where

$$\begin{aligned} W_g &= \int_V \rho g z\, dV, \quad W_\gamma = \tilde{\gamma} \oint_H dS, \\ W_m &= \int_{\hat{V}} \frac{\overline{\mathbf{B}^2}}{2\mu_0}\, dV, \end{aligned} \tag{91}$$

with the proviso that the currents in the external coils are fixed so that $\hat{\mathbf{B}}_0$ is constant. This proviso accounts for the at first surprising minus sign in the quantity to be minimized; the sources maintaining $\mathbf{B}_0$ supply energy $2W_m$, on average. Clearly, Eqs. (91) define the total gravitational, surface, and magnetic energies.

Even after S has been found, one important question should be addressed: Is the configuration stable? A similar question is even more significant in our later discussion of magnetostatic equilibria (Sec. 8).

There are two reasons why the furnace is not strictly in magnetostatic balance. First, in addition to the mean magnetic pressure that levitates V, there are fluctuating stresses that create oscillations in V of frequency 2ω. Because of the inertia of the fluid, the amplitude of these oscillations is usually small and can be safely ignored. Second, $B_n = 0$ on S only to leading order; in the next approximation, $B_n = O(\delta_\eta B_H/\mathcal{L})$. There is, then, in addition to the magnetic pressure, a smaller tangential magnetic stress of order $B_n B_H/\mu_0 \approx \delta_\eta B_H^2/\mu_0\mathcal{L}$. This stress has a nonzero average that causes slow circulations which stir the interior of the furnace (see Sec. VI.V of Moreau, 1990).

Suppose that the external coils are placed symmetrically about Oz; let the currents they carry have frequency ω but be phase shifted in such a way that a *rotating* axisymmetric $\mathbf{B}_0$ is created: i.e., $\mathbf{B}_0 = \hat{\mathbf{B}}_0(s, z) \exp[i(m\phi - \omega t)]$, where (s, ϕ, z) are cylindrical coordinates and $m > 0$ is an integer. When the body V is axisymmetric about Oz and $\delta_\eta \ll \mathcal{L}$, an electromagnetic skin again forms on S that expels $\mathbf{B}_0$ and heats V. As before, a magnetic pressure $B_H^2/2\mu_0$ acts on S, which has a nonzero average and can levitate V. In addition, there is a tangential stress that exerts a torque about Oz of order $m\delta_\eta B_H^2/2\mu_0$ per unit area of S; the rotating field imparts a rotation to V in the same sense—an example of Lenz's law. For further discussion, including reports on experiments and analyses of cases in which R_ω is not large, see Sec. VI.VI of Moreau (1990).

6. INFINITELY CONDUCTING FLUIDS

6.1 Alfvén's Theorem

At the start of Sec. 5, we noted that, if a conducting *solid* V were placed in a steady field $\hat{\mathbf{B}}_0$, that field would diffuse into V and ultimately take up its initial configuration everywhere. Suppose that the currents in the coils generating the field *subsequently* start to oscillate with frequency ω, where $R_\omega \gg 1$. The steady $\hat{\mathbf{B}}_0$ is then replaced outside V by $\mathbf{B}_0 = \hat{\mathbf{B}}_0 \exp(i\omega t)$. As in Sec. 5, an electromagnetic skin forms on S that shields the interior of V from the externally changing $\mathbf{B}_0$; i.e., the field $\hat{\mathbf{B}}_0$ present in V before the oscillation commenced is unchanged beneath the skin on S. This may also be seen from the induction equation (22) for $\mathbf{V} = 0$:

$$\frac{\partial \mathbf{B}}{\partial t} = \eta \nabla^2 \mathbf{B} = \frac{\omega \delta_\eta^2}{2} \nabla^2 \mathbf{B}. \tag{92}$$

Beneath the skin, $\nabla = O(\mathcal{L}^{-1})$. Since $\delta_\eta \ll \mathcal{L}$, Eq. (92) reduces in leading order to $\partial \mathbf{B}/\partial t \approx 0$, just as though one had formally set $\sigma = \infty$ in Eq. (22), i.e., just as though the solid were a perfect conductor.

There is another way of interpreting this result: the time scale over which the flux threading a conductor can change significantly is $\tau_\eta = \mu_0 \sigma \mathcal{L}^2$, where $\mathcal{L}$ is the length scale over which the field varies. For example (see Sec. 10.3), the field in a stationary conductor that has no sources to maintain it will disappear in a time of order τ_η. In the present case, the period of the external sources is $\tau_P = 2\pi/\omega$, and since $R_\omega \equiv 2\pi\tau_\eta/\tau_P \gg 1$, the flux of field through V has no time to change before those sources reverse.

The omission of $\nabla^2 \mathbf{B}$ from Eq. (22) for regions far from boundary layers (and other singular layers; see Sec. 9) applies more generally to a fluid, provided that $R_\omega \gg 1$ or $R \gg 1$. Equation (22) then reduces to Eq. (25); i.e.,

$$\frac{\partial \mathbf{B}}{\partial t} = \nabla \times (\mathbf{V} \times \mathbf{B}), \tag{93}$$

which is analogous to the equation of classical fluid mechanics that governs the vorticity $\boldsymbol{\omega}$ $(= \nabla \times \mathbf{V})$ in an inviscid, incompressible (or barotropic) fluid:

$$\frac{\partial \boldsymbol{\omega}}{\partial t} = \nabla \times (\mathbf{V} \times \boldsymbol{\omega}). \tag{94}$$

From Eq. (94), the *Kelvin–Helmholtz theorem* follows: vortex lines are material curves; i.e., vortex lines are "frozen" into the fluid as it moves. Similarly, *Alfvén's theorem* follows from Eq. (92): *Magnetic lines of force are frozen into an infinitely conducting fluid as it moves.* This property allows a completely new type of wave motion, one that belongs to MHD but not to classical fluid mechanics or to em theory: the *Alfvén wave,* sometimes also called the *hydromagnetic wave.*

6.2 Waves and Shocks

We recall that the Maxwell stresses associated with a field $\mathbf{B}$ consist of an isotropic (magnetic) pressure $B^2/2\mu_0$, which is for the moment irrelevant, together with a tension B^2/μ_0 along field lines. A flux tube, composed of a bundle of field lines of cross-sectional area $\mathcal{A}$, is therefore in a tension of $\mathcal{A}B^2/\mu_0$ and carries frozen within it a mass per unit length of $\mathcal{A}\rho$. Thus, like an elastic string, it can transmit a transverse wave in either direction along the tube with speed $V_{\mathrm{A}} = \sqrt{(\mathcal{A}B^2/\mu_0)/(\mathcal{A}\rho)}$, i.e., with the Alfvén velocity

$$\mathbf{V}_{\mathrm{A}} = \mathbf{B}/\sqrt{\mu_0 \rho}. \tag{95}$$

These Alfvén waves are transverse to a uniform **B** and are nondispersive. They define a new time scale, $\tau_A = \mathcal{L}/V_A$, the time taken by an Alfvén wave to cross a system of dimensions $\mathcal{L}$.

Returning now to the conductor V containing the trapped field $\hat{\mathbf{B}}_0$ and lying in the oscillatory field $\mathbf{B}_0 = \hat{\mathbf{B}}_0 \exp(i\omega t)$, where $R_\omega \gg 1$, we now suppose that V is an initially stationary *fluid* and we recognize a remarkable fact: the boundary layer on S (which is actually now a Hartmann layer) is no longer able to shield the interior of V from $\mathbf{B}_0$. News of the external variations is radiated throughout V by Alfvén waves. This radiation consists of vibrations of the field lines of the trapped $\hat{\mathbf{B}}_0$; by Alfvén's theorem, the flux threading the conductor beneath the Hartmann layer cannot change.

In a moving medium, the frequency of Alfvén waves is Doppler shifted in the obvious way. For example, the velocities of waves that ride on an *aligned flow*, i.e., one in which **V** is parallel to **B**, are $\mathbf{V} \pm \mathbf{V}_A$. If the *Alfvén number* (also called the *magnetic Mach number*), defined by

$$A = \mathcal{V}/\mathcal{V}_A, \tag{96}$$

is less than 1, one wave moves upstream and the other downstream, but if $A > 1$, both Alfvén waves travel downstream. If a body is placed in an aligned flow, it will create an upstream wake if $A < 1$ (see Sec. V.III of Moreau, 1990).

When $\mathbf{B}_0 = B_0\mathbf{e}_z$ is uniform,

$$\begin{aligned} \mathbf{V} &= \mathbf{V}(x, y, z \mp V_A t), \\ \mathbf{b} \equiv \mathbf{B} - \mathbf{B}_0 &= \mp \frac{1}{\sqrt{\mu_0 \rho}} \mathbf{V}(x, y, z \mp V_A t) \end{aligned} \tag{97}$$

satisfy the MHD equations exactly for **V** of arbitrary amplitude, provided $\nabla \cdot \mathbf{V} = 0$; their passage leaves the magnetic pressure unchanged. They correspond to waves traveling parallel (upper sign) or antiparallel (lower sign) to $\mathbf{B}_0$. Since the fluid is dissipationless, it may be assumed that **V** is discontinuous; if we then transform to a frame moving with the wave carrying the discontinuity and compare Eq. (61) with Eqs. (95) and (97), we see that the rotational discontinuity of Sec. 2.2 is merely a finite-amplitude Alfvén wave having a discontinuous profile. A linear combination of waves (97) does not satisfy the MHD equations unless the waves are of infinitesimal amplitude.

When the fluid is compressible, two further types of plane waves arise: *fast waves* and *slow waves*. Assuming that **B** and ρ are uniform and that the waves are infinitesimal, they have phase velocities

$$\begin{aligned} c_{f,s} &= \tfrac{1}{2}\sqrt{a^2 + V_A^2 + 2aV_A|\cos\theta|} \\ &\quad \pm \tfrac{1}{2}\sqrt{a^2 + V_A^2 - 2aV_A|\cos\theta|}, \end{aligned} \tag{98}$$

where a is the speed of sound, θ is the angle between **B** and the wave vector **k**, and the upper (lower) sign refers to the fast (slow) wave. Clearly, in a compressible medium, there are in addition to τ_A two further significant time scales: $\tau_f = \mathcal{L}/c_f$ and $\tau_s = \mathcal{L}/c_s$.

The fast and slow waves are dispersive. Their polarizations (defined by the direction of **V**) lie in the plane of **B** and **k**; the direction perpendicular to that plane defines the direction of polarization of the pure Alfvén wave discussed earlier, which exists unchanged in a compressible medium (since it does not affect ρ) and which is often called an *intermediate wave*; its phase speed is, as before, $V_A \times |\cos\theta| = c_f c_s/a$. This is also approximately the phase speed of the slow wave when $a \gg V_A$, and in this case, the polarizations of slow and intermediate waves are almost perpendicular to one another and to **B**. They are then essentially the Alfvén waves of the incompressible fluid, while the fast wave becomes an acoustic wave.

MHD shocks occur in many contexts. Perhaps the best known example is the bow shock, created where the solar wind encounters the magnetosphere and resembling the shock that occurs in Earth's atmosphere ahead of a supersonically moving blunt object. Figure 10 is a conceptual (overidealized) sketch of Earth's vicinity, the Dungey (1961) picture of the flow of the solar wind around the Earth in the interesting case in which the interplanetary magnetic field (IMF), generated in the Sun and carried to Earth by the solar wind, is in a direction opposite to the magnetic field of Earth. (It would be more realistic, but less clear for our purposes, to show Earth's dipole and the IMF tilted with respect to the ecliptic and to include components of **B** out of the plane shown.) The bow shock is a fast MHD shock of the type considered below. The magnetopause is the boundary of the mag-

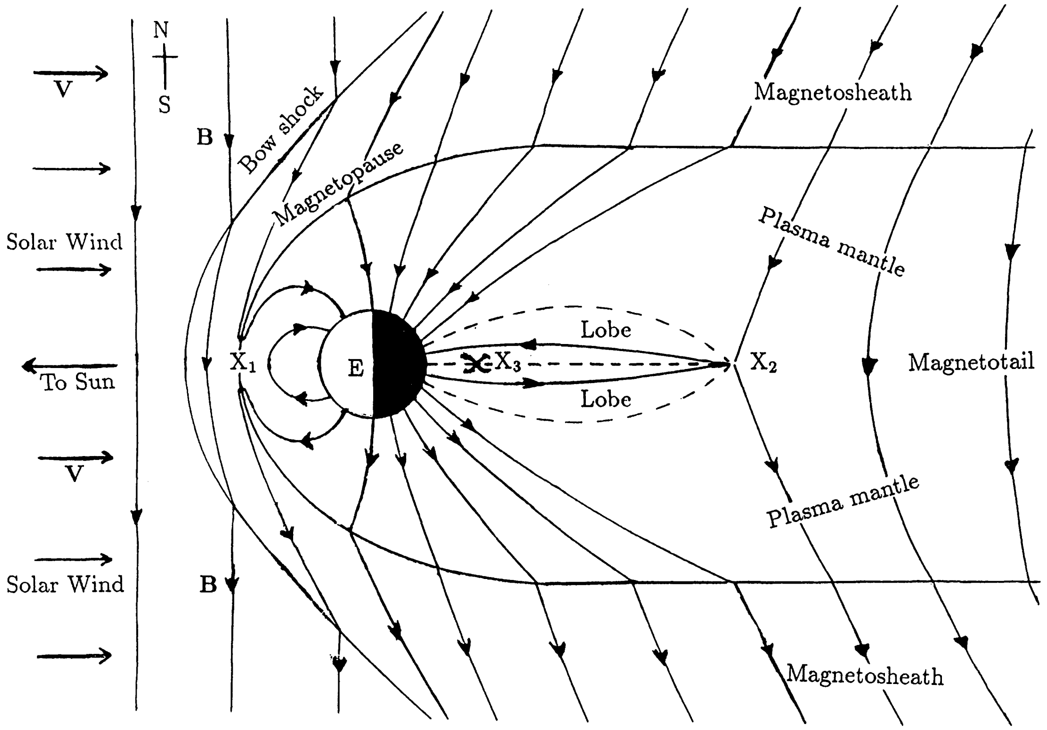

FIG. 10. Schematic of the Dungey model for the magnetic field in the vicinity of Earth (not to scale). The case in which the interplanetary magnetic field is antiparallel to the geomagnetic field is sketched. The bow shock is shown. Field line reconnection occurs at X_1 and X_2; further reconnection may take place on the neutral line between Earth (E) and X_2, as indicated by X_3 (see Sec. 9.4).

netosphere. There is little exchange of plasma across the magnetopause, and in a first approximation, one can think of the solar wind as flowing past a solid body whose boundary is the magnetopause. The shape of that body is, however, constantly altering as the MHD state of the solar wind changes. It appears that the magnetopause can be thought of as a contact discontinuity (Sec. 2.2) or, when material crosses from one side to the other, as an intermediate shock (see below). For future reference, we may observe three further features of Fig. 10. First, reconnection of field lines takes place at X_1, the head of the bow shock. Second, field line reconnection also occurs in the magnetotail, where the IMF detaches from Earth's field at a second neutral point X_2. Reconnection at these X-type neutral points is discussed in Sec. 9.3, where it is noted that slow shocks of the type described below are involved in the process of reconnection. Third, along the neutral line EX_2 of Fig. 10, there is a strong "northward" shear in the magnetic field, making it a favorable environment for tearing instabilities that create further points of reconnection, such as the one indicated by X_3 in Fig. 10 (see Sec. 9.4).

When treated as discontinuities, MHD shocks must satisfy the relations derived in Sec. 2.2. It follows from conditions (54), (56), and (59) that

$$\left(\frac{B_n^2}{\mu_0 j} - V_{n1}\right)\mathbf{B}_{H1} = \left(\frac{B_n^2}{\mu_0 j} - V_{n2}\right)\mathbf{B}_{H2}, \qquad (99)$$

so that **B** on one side of the shock lies in the plane defined by **B** and **n** on the other; conditions (56) and (59) show that **V** lies in the same plane Π on each side of the shock. In this coplanarity, MHD shocks differ fundamentally from the rotational discontinuity described in Sec. 2.2.

In categorizing MHD shocks, it has proven helpful to divide possible fluid states into four classes, differing by the normal flow speed in a reference frame in which the shock is at rest:

State 1. $V_n > c_f$;
State 2. $V_A < V_n < c_f$;
State 3. $c_s < V_n < V_A$; and
State 4. $V_n < c_s$.

Condition (60) allows six possibilities for the incoming → outgoing transitions: 1 → 2; 1 → 3; 1 → 4; 2 → 3; 2 → 4; and 3 → 4, all of which are compressive, i.e., the density of the gas is increased in passing through the shock.

The easiest transitions to understand are the first and the last, the *fast shock* and the *slow shock*, respectively; the remainder are called *intermediate shocks*. It follows from Eq. (99) that $\mathbf{B}_{H1}$ and $\mathbf{B}_{H2}$ are parallel for a fast shock and a slow shock and that they are antiparallel for intermediate shocks. Further, the strength of the transverse field is increased by a fast shock but decreased by a slow shock. Two marginal states can be approached:

1. $V_n = V_{A1}$, i.e., $1 \rightarrow (2 = 3)$, defines the "switch-on shock," so named because, by Eq. (99), $\mathbf{B}_{H1} = 0$, and
2. $V_n = V_{A2}$, i.e., $(2 = 3) \rightarrow 4$, defines the "switch-off shock," so named because $\mathbf{B}_{H2} = 0$.

See also Landau *et al.* (1984).

Situations in which the front separates states of oppositely directed $\mathbf{B}_H$ are particularly significant for reconnection processes and are, therefore, of special theoretical interest. Akhiezer *et al.* (1959), Germain (1960), and Kantrowitz and Petschek (1966) argued, from three different points of view, that intermediate shocks are unphysical and cannot arise. Consider, for example, the point of view of Germain (1960). He generalized to MHD the very successful theory of weak, non-MHD shocks [see, for example, Sec. 89 of Landau and Lifshitz (1987)]. According to this "classical" theory, it is always possible to connect two neighboring states on the same Hugoniot [i.e., related by $I_1 - I_2 + \frac{1}{2}(p_1 + p_2)(\rho_1^{-1} - \rho_2^{-1}) = 0$ and, therefore, satisfying the Rankine–Hugoniot conditions] by a thick shock front governed by the dissipative continuum equations. Although there is still no satisfactory theory of a strong shock front, it is an article of faith that solutions of the Boltzmann equation exist that similarly connect two distant states on the same Hugoniot by a shock front having a thickness of the order of a mean free path. In this way, the idea arises naturally that the discontinuity theory of shocks always gives results that can in principle arise in nature.

In the MHD case, a solution of the dissipative MHD equations can always be found to connect a state 1 to a neighboring state 2 or to connect a state 3 to a neighboring state 4. Thus, the discontinuity theory is on firm ground for weak fast and weak slow shocks. Germain (1960) proved, however, that, for *some* specifications of the transport parameters (shear and bulk viscosities and electrical and thermal conductivities), no solution of the dissipative MHD equations exists to bring about an intermediate transition (such as 1 → 3). In *these* circumstances, the discontinuity theory fails, and this led (erroneously) to the conclusion that intermediate shocks are unphysical in *all* circumstances. Rotational discontinuities were invoked to connect states where $\mathbf{B}_H$ switches direction at the front.

Recent research has led to an amusing reversal of ideas. Computer solutions for weak fronts have shown that, in situations such as those described above in which Germain (1960) had shown that no intermediate shock is possible, two or more shocks evolve, one intermediate and the other(s) fast or slow. These together effect the required intermediate transition (such as 1 → 3). It was also shown that rotational discontinuities cannot be involved in any way. Paradoxically, however, $\mathbf{B}_H$ may have a nonzero component perpendicular to Π that is confined to the interior of an intermediate shock front. The topic is one of some complexity (see Wu, 1992).

6.3 Creation of Magnetic Energy

Alfvén's theorem provides an almost indispensable aid to MHD thought: the evolution of $\mathbf{B}$ is more readily pictured in concrete terms as the motion of field lines frozen into the fluid than in abstract mathematical terms as a solution of Eq. (25).

As an example, consider two ways in which energy is converted from kinetic to magnetic forms. Imagine a tube of magnetic flux of radius a_0 and length L_0 containing a uniform field B_0 and surrounded by force-free fluid. Suppose that an inward radial motion toward the axis of the tube compresses the tube symmetrically until its radius is reduced to $a < a_0$ and the (uniform) field it contains becomes B. Since the strength κ of the tube (defined as the flux of magnetic field through any cross section) remains the same by Alfvén's theorem, $\pi a^2 B = \pi a_0^2 B_0 = \kappa$. The trapped magnetic energy per unit length of tube is increased, however, by a factor of $(a_0/a)^2$, from $\pi a_0^2 B_0^2/2\mu_0 = \kappa^2/2\pi\mu_0 a_0^2$ to $\pi a^2 B^2/2\mu_0 = \kappa^2/2\pi\mu_0 a^2$. This increase represents the work done by the motion compressing the tube against the magnetic pressure in the tube (which increases during the compression). During the compression, B/ρ does not change. (This is a consequence of Cauchy's integral in MHD.)

Magnetic energy can be increased by stretching the flux tube to a length of $L > L_0$, say. This mechanism also works in an incompressible fluid for which volume is preserved and for which the stretching requires that the cross section is reduced by a factor of L_0/L; it follows that B is enhanced by a factor of L/L_0 and the magnetic energy of the tube is increased by the factor $(L/L_0)^2$, without altering the pressure within the tube. The increase in magnetic energy is now supplied by the work done by the motion against the tension in the tube (which increases during the stretching).

The proportionality of B and L during the stretching of field lines lies at the heart of the principle invoked in Sec. 4 to explain how duct flow changes at abrupt variations in cross section or applied field. The strength of electric current in a duct at large M is dynamically determined and may be small compared with both the induced emf and **E**. Then, despite the smallness of R, the last term of Eq. (21) is negligible, Alfvén's theorem follows, and the characteristic surfaces defined by Eq. (87) become significant.

6.4 Magnetic Helicity

The topology of field lines cannot change when $\sigma = \infty$. This has attracted much theoretical interest [see, for example, Moffatt *et al.* (1992)]. One of the best known topological connections arises from the constancy of the *magnetic helicity*. This quantity, sometimes also called the "Woltjer invariant" after its discoverer, is defined by

$$H_M = \oint \mathbf{A} \cdot \mathbf{B} dV \qquad (100)$$

taken over the interior of any closed *magnetic surface*, which is a surface S such that

$$\mathbf{n} \cdot \mathbf{B} = 0, \quad \text{on S}. \qquad (101)$$

Here, **A** is a vector potential which by Eq. (25) obeys

$$\frac{\partial \mathbf{A}}{\partial t} = \mathbf{V} \times (\nabla \times \mathbf{A}) + \nabla \Phi \qquad (102)$$

for some Φ. The integral (100) is independent of the gauge chosen for the vector potential **A** of **B**, i.e., of the arbitrary gradient that can be added to **A** without disturbing the relationship $\mathbf{B} = \nabla \times \mathbf{A}$. From Eqs. (25), (101), and (102), it follows that $dH/dt = 0$; i.e., H is constant in a perfect fluid.

The topological connection can be clarified with the help of the Biot–Savart representation

$$\mathbf{A}(\mathbf{x}) = \frac{1}{4\pi} \int_{\text{all space}} \frac{\mathbf{B}(\mathbf{x}') \times (\mathbf{x} - \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|^3} dV. \qquad (103)$$

Suppose that **B** vanishes everywhere except in the two flux tubes, of strength κ_1 and κ_2 and filling volumes V_1 and V_2, depicted in Fig. 11. Then,

$$H = H_1 + H_2 + H_{12} + H_{21}, \qquad (104)$$

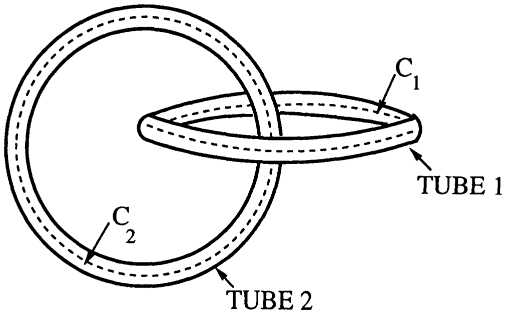

FIG. 11. Linked flux tubes with nonzero mutual helicity. In the limit in which the tubes have infinitesimal cross sections, the mutual helicity is the product of the strengths of the tubes and the winding number (here, ±1) which is Gauss' measure of topological linkage of two curves.

where

$$H_l = \frac{1}{4\pi}\int_{V_l}\int_{V_l} \frac{(\mathbf{x}-\mathbf{x}')\cdot[\mathbf{B}(\mathbf{x})\times\mathbf{B}(\mathbf{x}')]}{|\mathbf{x}-\mathbf{x}'|^3}\,dV dV' \tag{105}$$

is the self-helicity of flux tube l, where $l = 1$ or 2 and

$$H_{12} = H_{21} = \frac{1}{4\pi}\int_{V_1}\int_{V_2} \frac{(\mathbf{x}_1-\mathbf{x}_2)\cdot[\mathbf{B}(\mathbf{x}_1)\times\mathbf{B}(\mathbf{x}_2)]}{|\mathbf{x}_1-\mathbf{x}_2|^3}\,dV_1 dV_2 \tag{106}$$

is their mutual helicity.

Now suppose the flux tubes to be of infinitesimal cross section, surrounding the closed field lines C_1 and C_2. Then, Eq. (106) reduces to

$$H_{12} = H_{22} = \kappa_1\kappa_2 N_{12}, \tag{107}$$

where

$$N_{12} = \frac{1}{4\pi}\oint_{C_1}\oint_{C_2} \frac{(\mathbf{x}_1-\mathbf{x}_2)\cdot(d\mathbf{x}_1\times d\mathbf{x}_2)}{|\mathbf{x}_1-\mathbf{x}_2|^3}\,dC_1 dC_2 \tag{108}$$

is the "winding number," first isolated by Gauss, who showed that $|N_{12}|$ is an integer equal to the number of times C_1 threads C_2 before closing. If the field lines are disconnected, $|N_{12}| = 0$; in the topology of Fig. 11, $|N_{12}| = 1$.

Ohmic dissipation causes magnetic helicity to dissipate at a rate proportional to $\mathbf{B}\cdot\mathbf{J}/\sigma$ per unit volume, which may be compared with the rate of dissipation of magnetic energy $\epsilon_\eta = J^2/\sigma$. In a turbulent conductor, in which $\mathcal{J} = O(\mathcal{B}/\mu_0 l)$, where the correlation length l of the turbulence is small compared with $\mathcal{L}$, the scale of the system, the rate of dissipation of H_M is much smaller than that of E_M (see Sec. 9.3).

6.5 The Lundquist Number

Despite the usefulness of Alfvén's theorem in visualizing MHD processes, it is wise to bear constantly in mind that, in reality, σ is not infinite, that field lines diffuse relative to the fluid at a rate proportional to η, and that the frozen-field picture may be seriously misleading if applied over time scales comparable with τ_η, or perhaps even shorter (see Sec. 9.4).

Alfvén waves are damped by Ohmic resistance and by viscosity. In a frame moving with the wave, they decay on a time scale of $\tau_D = \mathcal{L}^2/(\nu + \eta)$, where $\mathcal{L}$ is their wavelength. Often, $p_m \ll 1$ and we may therefore replace τ_D by τ_η. The waves then travel a distance of $\mathcal{V}_A\tau_\eta$ through a conductor V before dissipating. An important parameter is clearly the *Lundquist number*:

$$S = \mathcal{V}_A\tau_\eta/\mathcal{L} = \tau_\eta/\tau_A = p_m^{1/2}M = \mathcal{V}_A\mathcal{L}/\eta, \tag{109}$$

where $\mathcal{L}$ is the scale of V. Thus, S resembles the magnetic Reynolds number (86), but with $\mathcal{V}$ replaced by $\mathcal{V}_A$. If $S \gg 1$, Alfvén waves cross V in an *Alfvén time*, $\tau_A = \mathcal{L}/\mathcal{V}_A$, short compared with τ_η and are, therefore, scarcely attenuated; if $S \ll 1$, they are damped out after traveling only a small fraction of $\mathcal{L}$. Although large M can be attained in liquid-metal facilities, it is more difficult to make S large since $p_m \ll 1$ for most available fluids. Large S can be attained in plasma experiments. Usually $S \gg 1$ in geophysical and astrophysical systems because of their enormous scale $\mathcal{L}$ (see Sec. 10).

7. MAGNETOHYDROSTATICS

7.1 Balancing Nonzero Pressure Gradients

A conducting fluid is in magneto(hydro) static equilibrium when Eq. (78) holds. The fluid is unaccelerated and either is stationary or, as in Sec. 3.2, moves so slowly that inertial forces are negligible, the dynamical balance being between the pressure gradient, the Lorentz force, and externally applied forces such as gravity, if any. In the simplest case of no external forces,

$$\mathbf{J}\times\mathbf{B} = \nabla p. \tag{110}$$

If such an equilibrium is disturbed, it will respond on the dynamic time scales of the three types of waves introduced in Sec. 6.2: $\tau_A = \mathcal{L}/V_A$, $\tau_f = \mathcal{L}/c_f$, and $\tau_s = \mathcal{L}/c_s$. The result may, or may not, be an instability that takes the system far from its initial configuration during these time scales (see Sec. 8). Equation (110) expresses a dynamical balance over the three time scales and in particular over τ_A. Clearly, such a balance has little meaning if **B** has evolved during that time through Ohmic diffusion. It is, therefore, implicitly assumed in the study of magnetostatics that $\tau_A \ll \tau_\eta$, i.e., that $S \gg 1$. Here and in Sec. 8, the fluid is assumed to be perfectly conducting: $S = \infty$.

The left-hand side of Eq. (110) is at most $O(\mathcal{B}^2/\mu_0\mathcal{L})$, and the right-hand side is $O(p/\mathcal{L})$, where p is a typical pressure *difference* within the system. It follows that

$$\beta = \mu_0 p/\mathcal{B}^2 \tag{111}$$

is $O(1)$ at most, but it may be zero, as when **B** and **J** are parallel; **B** is then a *force-free field* and

$$\mathbf{J} \times \mathbf{B} = 0. \tag{112}$$

In fusion-reaction technology, where the aim is to promote as many reactions as possible during a "confinement time" in which Eq. (110) approximately holds, it is desirable to design a magnetic bottle that has as large a β as possible, and β is in fact one of the *figures of merit* by which a design is judged. Unfortunately, the larger the β, the more prone the device is to instability. A compromise is necessary, and in practical devices, β generally lies between 0.01 and 0.1. For other figures of merit, see Freidberg (1987).

According to Eq. (110), the lines of force and current lines lie on isobaric magnetic surfaces, which are necessarily multiply connected and which, in 2D plasma configurations, may resemble the axisymmetric set of nested surfaces sketched in Fig. 12. The surfaces surround a single degenerate toroid (the dotted circle shown). The magnetic field has *poloidal* and *toroidal* components, and their sum produces a field that spirals around each isobaric surface without, in general, closing; it is "ergodic." The current lines similarly fill each magnetic surface.

Insight into these *toroidal equilibria*, especially those of large aspect ratio, R/a (see Fig. 12), can be gained from studies of "1D equilibria" such as the *cylindrical equilibria* ($R \rightarrow \infty$) of which the *screw pinch* depicted in Fig. 13(a) is the general case. The word "pinch" refers to the compression of the fluid (or plasma) by the magnetic stresses. The fluid (or plasma) pressure p increases toward the axis $s = 0$, both because of the increase in magnetic pressure $B^2/2\mu_0$ away from the axis and because the poloidal field lines, which girdle the cylinder like hoops around a barrel, exert a *hoop stress* B_ϕ^2/μ_0.

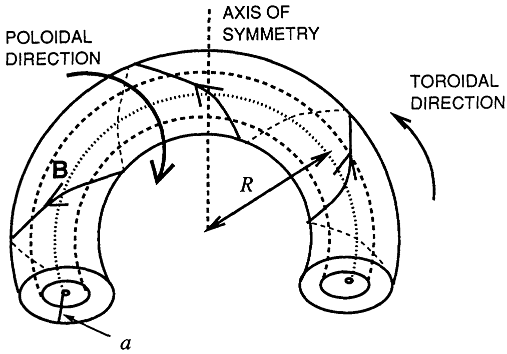

FIG. 12. A toroidal magnetostatic equilibrium. The magnetic surfaces are a set of nested toroids, on which the pressure is constant and about which the lines of force and current lines wind ergodically. Two surfaces are shown (full lines and dashed lines) as well as the degenerate surface (the dotted curve) forming the center of the nest.

Two special cases of the screw pinch have received particular attention in the field of controlled thermonuclear research: the θ

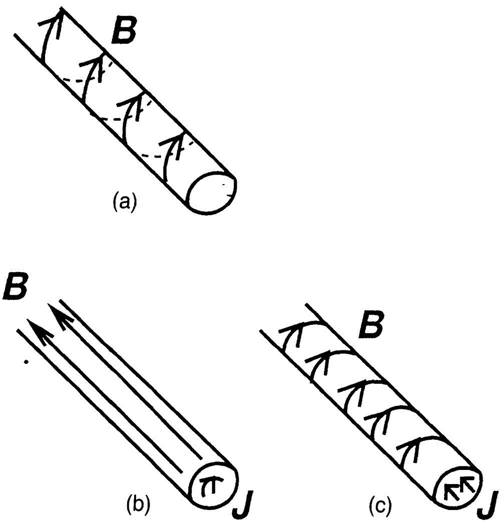

FIG. 13. The general screw pinch and two limiting cases. (a) The general cylindrical pinch, which has nonzero toroidal and poloidal fields. (b) The θ pinch, which has no poloidal field or toroidal current. (c) The Z pinch, which has no toroidal field or poloidal current.

pinch, in which B_ϕ and J_z are zero [Fig. 13(b)], and the Z pinch, in which $B_z = J_\phi = 0$ [Fig. 13(c)]. The latter is the 1D analogue of Ohmically heated Tokomaks, such as ITER (pronounced ē'ta, as in Greek), proposed as the way to controlled thermonuclear fusion in the future. The design envisages a blanket of liquid metal around the reaction chamber, both to remove heat and breed tritium [see Fig. 14 and, for example, Malang *et al.* (1991)]. Such a blanket operates on the same physical principles as those described in Secs. 3 and 4, and, in this sense, the technology of ITER spans MHD more broadly than does the MHD usually encountered in plasma physics.

Also of interest are 3D configurations such as the stellarator, which has no symmetry about an axis and which will not be described here (see Freidberg, 1987).

7.2 Force-Free Fields

In a low-density fluid (such as the solar magnetosphere; see Sec. 10), the pressure gradient can be neglected in comparison with the Lorentz force, and the internal energy density is negligible compared with the magnetic energy density. To a good first approximation, $\beta = 0$. Equation (112) then holds so

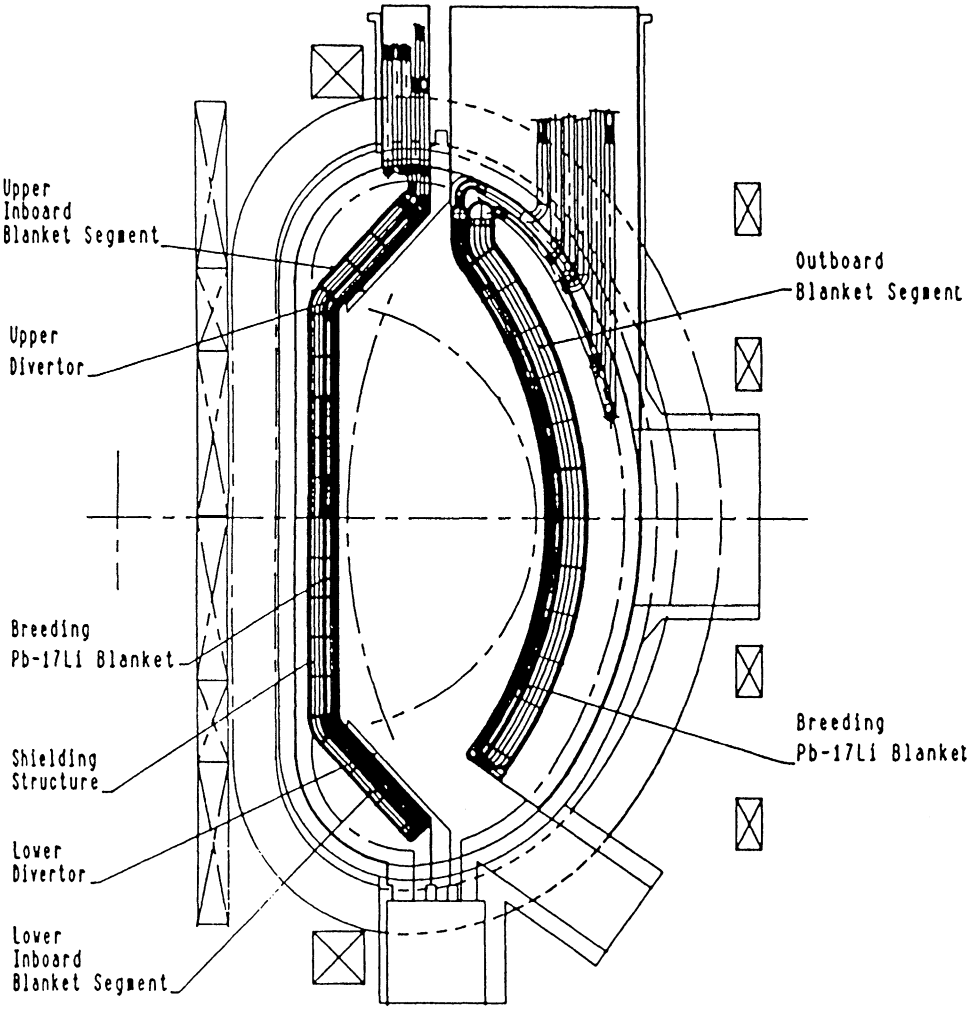
Upper Inboard Blanket Segment
Upper Divertor
Breeding Pb-17Li Blanket
Shielding Structure
Lower Divertor
Lower Inboard Blanket Segment
Outboard Blanket Segment
Breeding Pb-17Li Blanket

FIG. 14. Schematic of ITER (the International Thermonuclear Experimental Reactor). The sketch shows a cross section of the tokomak including the oval reaction chamber and the arrangement of segments of the water-cooled lead-lithium blanket that breeds tritium. (After Malang *et al.*, 1991.)

that, for some field α,

$$\nabla \times \mathbf{B} = \alpha \mathbf{B}. \tag{113}$$

Operating on this with $\nabla\cdot$ using Eq. (3), we see that

$$\mathbf{B} \cdot \nabla \alpha = 0; \tag{114}$$

i.e., α is constant on each line of force. The lines of force and current lines lie on magnetic surfaces of constant α on which they are usually ergodic; i.e., $\alpha = \alpha(v)$, where v is some label for a magnetic surface, e.g., the volume v it contains.

If α is constant in a volume, then the lines of force and current lines will in general be chaotic in that volume, i.e., they will usually be space-filling rather than surface-filling. Such fields, for which

$$\nabla \times \mathbf{B} = \alpha \mathbf{B}, \quad \alpha = \text{constant}, \tag{115}$$

are known as *Beltrami fields*.

A simple Cartesian example illustrating both magnetostatic and force-free (Beltrami) fields was given by Fadeev *et al.* (1965):

$$\mathbf{B} = \frac{B_0}{\cosh kz + \epsilon \cos kx} [\sinh kz, \alpha_0 \sqrt{1 - \epsilon^2}, \epsilon \sin kx], \tag{116}$$

where B_0, k, α_0, and $\epsilon (<1)$ are constants. This is force-free if $\alpha_0 = \pm 1$ and is hydrostatic otherwise, with

$$p = p_\infty + \frac{B_0^2 (1 - \alpha_0^2)(1 - \epsilon^2)}{2\mu_0 (\cosh kz + \epsilon \cos kx)^2},$$

where p_∞ is the pressure at $z = \pm\infty$. Field lines projected onto the xz plane are sketched in Fig. 15. The $\epsilon = 0$ field, $B_0 = B_0 (e_x \tanh kz + \alpha_0 e_y \operatorname{sech} kz)$, is prone to instability through a resistive instability, known as the "tearing mode" (see Sec. 9.4). This instability reconnects the field lines of $\mathbf{B}_0$, and the final result might be a steady state that could be modeled by Eq. (116) with $\epsilon \neq 0$. It has been shown, however, that state (116) is itself prone to a coalescence instability (see Sec. 9.4).

Force-free fields have useful variational properties. We mention two, to which we return in Sec. 9.

7.2.1 First Variational Property Let V be a simply connected volume bounded by a surface S over which B_n is specified. By a well-known theorem of potential theory, the field for which the magnetic energy contained in V is a minimum is harmonic: $\nabla \times \mathbf{B} = 0$ (and as always $\nabla \cdot \mathbf{B} = 0$). Now suppose that, through every point inside V, a line of force passes that meets S in two "footpoints." Suppose not only that B_n is specified over S but also that the correspondence between every such pair of footpoints is fixed; i.e., we envisage that each line of force crossing V can move, but only between fixed end points on S. It can then be shown that the field of minimum magnetic energy is force-free. Thus, if a perfectly conducting fluid is started from an arbitrary initial state and evolves consistently with the constraints, and if the fluid has a dissipation mechanism available such as viscosity, then it will equilibrate to a force-free configuration satisfying Eqs. (113) and (114) for some $\alpha(v)$. Clearly, the constrained minimum energy of the force-free field cannot be less than the unconstrained minimum of the harmonic field. The difference between the two is significant (see Sec. 10.2).

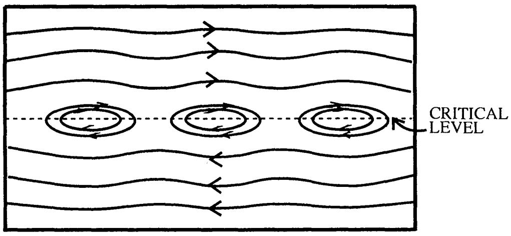

FIG. 15. An example illustrating both magnetostatic and force-free equilibria. This example (Fadeev *et al.*, 1965) has both closed and open field lines. The former, called cat's eyes or magnetic islands, may have been produced by the tearing instability of the sheared field (open field lines) as described in Sec. 9.4.

7.2.2 Second Variational Property Suppose V is not simply connected and that its boundary S is a magnetic surface. For example, the magnetic field configuration might initially resemble that shown in Fig. 12, with the outermost toroid a perfectly conducting wall. The initial configuration is not force-free and will evolve in time, but each magnetic surface possesses a magnetic helicity $H_M(v)$, which, for a perfectly conducting fluid, is conserved during the evolution (see Sec. 6.4). The variational property is this: the state of minimum (magnetic) energy that preserves $H_M(v)$ for each surface is force-free. Thus, if the fluid has a dissipation mechanism available such as viscosity, it will equilibrate to a force-free configuration satisfying Eqs. (113) and (114) for some $\alpha(v)$. But this is not the minimum energy state that it can reach if the constraints are partially lifted. The difference between the two is significant (see Sec. 9.3).

8. IDEAL INSTABILITIES

8.1 Approaches to Stability

It is of vital importance in fusion reactor technology to design stable field configurations that will confine a plasma for a sufficiently long time for a significant number of thermonuclear reactions to take place. A configuration may be unstable through MHD processes or through microscale processes beyond the scope of this review. The most obvious MHD processes are ideal instabilities that can destroy the equilibria discussed in Sec. 7 in a time of order τ_A, which is too short for them to be interesting as potential reactors. Less obviously, a configuration may be threatened by resistive instabilities; i.e., it may seem to be stable when diffusive processes are ignored but is in fact unstable for $\eta \neq 0$. Resistive modes will be discussed in Sec. 9.4. Here, we suppose that

$$\eta = \nu = \kappa = 0. \tag{117}$$

The first step is to study the linear stability of the configuration. It is supposed that the equilibrium state, in which ρ, p, $\mathbf{B}$, ... are denoted by ρ_0, p_0, $\mathbf{B}_0$, ..., is perturbed. As a result, a new state is created in which

$$\rho(\mathrm{x}, t) = \rho_0(\mathrm{x}) + \delta\rho(\mathrm{x}, t), \ldots \tag{118}$$

and which is governed by the ideal MHD equations. This system is *linearized* by supposing that $\mathbf{V}$, $\delta\rho$, δp, $\delta\mathbf{B}$, ... are so small that their squares and products are negligible. The resulting equations can be reduced to a single equation governing the vector displacement $\boldsymbol{\xi}(\mathbf{x}, t)$ of the fluid element originally situated at $\mathbf{x}$:

$$\rho_0 \frac{\partial^2 \boldsymbol{\xi}}{\partial t^2} = \mathbf{F}(\boldsymbol{\xi}), \tag{119}$$

where $\mathbf{F}$ is linear in ξ and its gradients and depends on the equilibrium fields, but not explicitly on t. We shall confine attention to two cases in which the operator $\mathbf{F}$ is self-adjoint:

$$\int_v \boldsymbol{\xi}_1 \cdot \mathbf{F}(\boldsymbol{\xi}_2) dV = \int_v \boldsymbol{\xi}_2 \cdot \mathbf{F}(\boldsymbol{\xi}_1) dV. \tag{120}$$

This happens in a number of cases of interest, including the following:

1. when the fluid conductor V is confined by infinitely conducting boundaries S; and
2. when S is a free surface separating V from a nonconducting exterior $\hat{\mathrm{V}}$ free of sources of energy such as current-carrying coils.

An investigation of linear stability consists of solving the initial-value problem, i.e., of solving Eq. (119) and appropriate boundary conditions starting from arbitrary initial $\boldsymbol{\xi}$ and $\partial\boldsymbol{\xi}/\partial t$. If the solution remains bounded for all time, the configuration is linearly stable; if any solution grows without bound, it is unstable. (Of course, in reality, no solution can grow without bound; it can grow only until its amplitude violates the assumed smallness of ξ, and nonlinear interactions become significant in limiting its growth.)

The initial-value problem is rarely easy to solve. An attractive, but seldom useful, alternative is to seek normal-mode solutions in which $\boldsymbol{\xi} \propto \exp(i\omega t)$ and for which Eq. (119) becomes

$$-\omega^2 \rho_0 \boldsymbol{\xi} = \mathbf{F}(\boldsymbol{\xi}), \tag{121}$$

which, together with the boundary conditions, poses an eigenvalue problem for ω^2. When Eq. (120) holds, ω^2 is real. If the spectrum of ω^2 of the normal modes is discrete and complete, which is rarely the case, the arbitrary initial state can be expanded as an infinite sum over the eigenfunctions. If one or more of these ω^2 is negative, the system is unstable; if all are positive, it is linearly stable.

Equation (118) and the relevant boundary conditions may possess only a finite number of discrete eigenvalues or perhaps none at all; the remainder belong to a continuous spectrum. For example, consider the result of perturbing an infinite ocean of perfectly conducting fluid $z < 0$ lying under gravity ($\mathbf{g} =$

$-g\mathbf{e}_z$) and permeated by a uniform vertical field $\mathbf{B}_0 = B_0\mathbf{e}_z$. It is obvious on physical grounds that the system is stable; surface waves are damped as Alfvén waves radiate their energy to $z = -\infty$, and this may be confirmed by solving the (non-self-adjoint) initial-value problem. There are, however, no discrete normal modes whatsoever.

Although the unstable modes are usually found in a finite discrete spectrum, there is no general theorem that makes this invariably true; the unstable modes may in principle lie in the continuous spectrum. To solve for the continuous spectrum adds complications; its existence also makes the already difficult task of solving the initial-value problem even harder (see Lifschitz, 1989). For this reason, energy methods are often employed for their relative simplicity.

8.2 The Energy Method

In the configurations **1** and **2** described above, the total energy is conserved; i.e., $K + W$ is constant, where K is the kinetic energy and W is the potential energy of the system (see Sec. 2.1). The situation then resembles that of a ball rolling on a surface. A stable configuration is shown in Fig. 16(a), one that is stable for small displacements but not large is sketched in Fig. 16(b), and an unstable equilibrium is shown in Fig. 16(c). The possibilities shown in Figs. 16(a) and (b) cannot be distinguished in a linear stability analysis, in which K and δW are quadratically small:

$$K = \frac{1}{2}\int_V \rho_0 \left(\frac{\partial \boldsymbol{\xi}}{\partial t}\right)^2 dV,$$

$$\delta W = -\frac{1}{2}\int_V \boldsymbol{\xi}\cdot\mathbf{F}(\boldsymbol{\xi})dV, \qquad (122)$$

where the integrations are over the unperturbed equilibrium configuration, $\delta W = W - W_0$ and W_0 is the equilibrium potential energy. If *any* displacement $\boldsymbol{\xi}$ can be found that satisfies the boundary conditions and for which $\delta W < 0$, that potential energy is transformed into an increase in kinetic energy; i.e., the system is unstable. If $\delta W > 0$ for *all* $\boldsymbol{\xi}$ satisfying the boundary conditions, the system is (linearly) stable.

For displacements that deform S, the evaluation of δW from Eq. (122) is hampered by the need to choose only those $\boldsymbol{\xi}$ satisfying boundary conditions on S, such as Eq. (29) and either Eq. (62) or (63). In a celebrated reformulation of Eq. $(122)_2$, Bernstein *et al.* (1958) showed how to rewrite δW in such a way that only Eq. (29) had to be enforced; we may call such $\boldsymbol{\xi}$ "eligible." For further details, see Freidberg (1987).

In the simplest cases (e.g., Secs. 8.2.1–8.2.3), it is comparatively easy to guess an eligible $\boldsymbol{\xi}$ for which $\delta W < 0$ and so prove instability. In other cases (e.g., Sec. 8.2.4), it is necessary to appeal to the calculus of variations to find the minimum of δW and to deduce stability or instability from its sign.

8.2.1 The Flowers–Ruderman Mechanism

This mechanism was advanced in an attempt to explain certain perplexing data from pulsars (Flowers and Ruderman, 1977). The idea is illustrated in Fig. 17. Two identical bar magnets are side by side and free to rotate about a common pivot through their centers; initially their N poles are adjacent, as are their S poles [Fig. 17(a)]. This configuration is unstable. One magnet swings relative to the other to bring the N pole of one next to the S pole of the other [Fig. 17(b)].

Now suppose that, instead of bar magnets, we have two hemispheres of a magnetic "star" V, containing an axisymmetric field which in the exterior, $\hat{\mathrm{V}}$, of V is a potential field of dipole type. If one hemisphere rotates relative to the other through 180° about an axis through the equator, the dipole moment is reduced and a quadrupole moment is increased. The magnetic energy $\hat{W}_M$ of $\hat{\mathrm{V}}$ is less for the quadrupole than for the dipole, and the magnetic energy W_M of V is unaltered. Indeed, further motions on S may, by rearranging the areas where magnetic flux enters and leaves V, reduce the energy of the configuration even more by the destruction of the lower harmonics of $\mathbf{B}$ and the creation of higher harmonics. Such motions may, however, increase the distortion of $\mathbf{B}$ within V and so offset the reduction in $\hat{W}_M$ by an even greater increase of W_M. Moreover, if V is rotating, such motions may be inhibited by Coriolis forces. Nevertheless, the general point is made: $\hat{W}_M$ can be reduced by suitable surface motions, and this may suffice to make the configuration unstable.

8.2.2 The *Z* Pinch The *Z* pinch shown in Fig. 13(c) is vulnerable to an axisymmetric

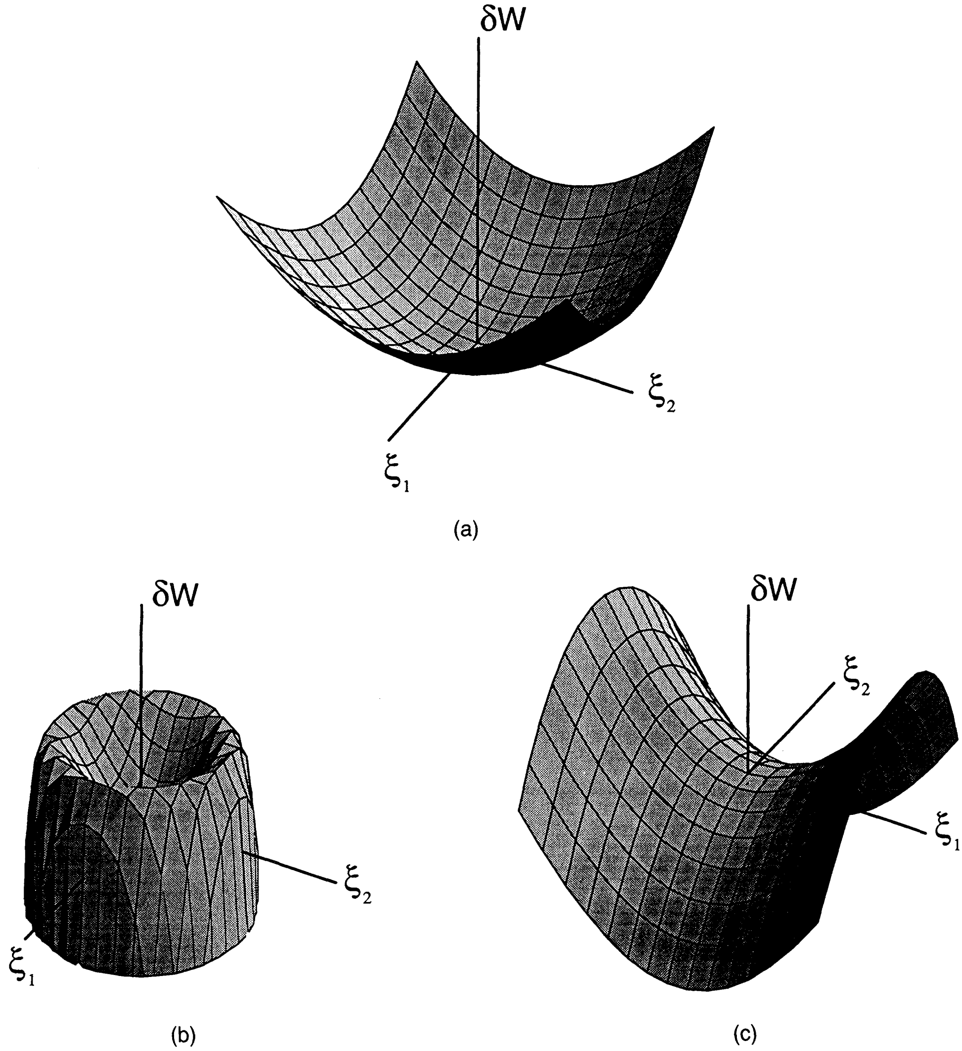

FIG. 16. Schematic explanation of the energy principle. These depict an analogy with the motion of a particle over a smooth surface under gravity: (a) a stable state in which $\delta W > 0$ for all displacements **gj**; (b) a state that is stable for all sufficiently small $\boldsymbol{\xi}$ but is unstable for larger $\boldsymbol{\xi}$; (c) the saddle, an unstable case in which δW is positive for some $\boldsymbol{\xi}$ but negative for others.

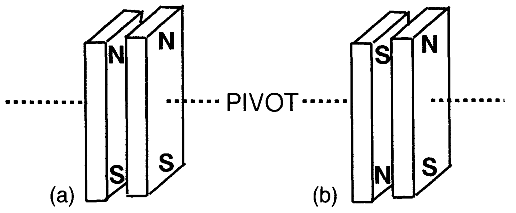

FIG. 17. (a),(b) The Flowers-Ruderman instability of two bar magnets. The configuration on the left is unstable to a displacement that carries it into the configuration on the right.

perturbation $\boldsymbol{\xi}$ that deforms its surface, $s = a$, into something vaguely resembling a string of sausages (see Fig. 18). Such a mode is sometimes called a "sausage" or "necking" instability.

We consider the simplest case ($\mathbf{B}_0 \equiv 0$) of a field-free plasma V, carrying a uniform surface current $\mathbf{I} = I\mathbf{e}_z$ that has been induced by the sudden creation of an external field $\hat{\mathbf{B}} = aI\mathbf{e}_\phi/\mu_0 s$ in the vacuum $\hat{\mathrm{V}}$ surrounding V. The net surface current along the pinch is

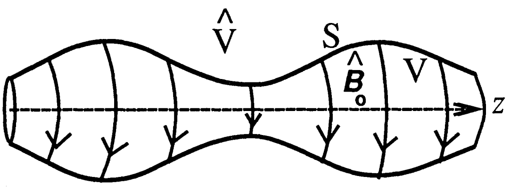

FIG. 18. Necking instability of the Z pinch.

not altered by the axisymmetric $\boldsymbol{\xi}$, so that $\hat{\mathbf{B}}$ is unchanged. The areas of maximum (minimum) surface displacement from Oz, therefore, are areas of reduced (enhanced) field and magnetic pressure. By choosing ξ_z so that $\nabla \cdot \boldsymbol{\xi} = 0$, we ensure that ρ and p are also unchanged in V, and so the unbalanced forces on the areas of maximum and minimum displacement act to increase the displacement further.

Necking instability can be inhibited by setting up a field B_{0z} in V prior to inducing the surface current on S. The compression of this field, by the $\boldsymbol{\xi}$ considered above, opposes the formation of necks. Thus, the general screw pinch shown in Fig. 13(a) is less prone to this particular kind of instability. It may, however, be prone to a *kink instability*. This is the $m = 1$ case of the general displacement $\boldsymbol{\xi} \propto \exp[i(m\phi + kz)]$, where m is a positive or negative integer. (Necking instabilities are axisymmetric: $m = 0$.) In the kink (or "Shafranov") instability, the central axis of the cylinder is displaced from Oz and winds about it as a helix. The instability resembles the one that creates loops on the cord of a telephone handset when it is twisted.

8.2.3 Interchange Instability We consider the field-free plasma more generally, supposing it is confined by a field $\hat{\mathbf{B}}_0$ in the surrounding vacuum $\hat{V}$ (see Fig. 19). A *local stability criterion* can be developed which decides whether the configuration is unstable or (linearly) stable at a particular point P on S. More precisely, displacements are envisaged that vanish everywhere except in a small neighborhood of P, which is then shrunk to zero. If $\delta W > 0$ for all such $\boldsymbol{\xi}$, the configuration is locally stable at P; if $\delta W < 0$ for any $\boldsymbol{\xi}$, it is locally unstable. Local instability is enough to establish overall instability, but local stability establishes very little; it is always possible that a global $\boldsymbol{\xi}$ might make $\delta W < 0$ even when the configuration is everywhere locally stable.

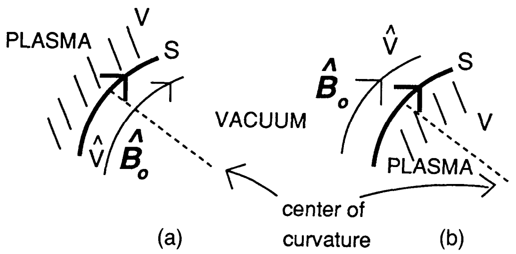

FIG. 19. Instability of a current-free plasma contained in a magnetic bottle. (a) Points on the plasma surface where $\hat{\mathbf{B}}_0$ is concave to the surface of the plasma are locally stable, but (b) points where it is convex are unstable.

It may be shown that a configuration in which $\hat{\mathbf{B}}_0$ is concave to the plasma, i.e., in which the center of curvature of the field line through P lies in the vacuum [Fig. 19(a)], is locally stable; if $\hat{\mathbf{B}}_0$ is convex to the plasma [Fig. 19(b)], it is unstable. The deformation $\boldsymbol{\xi}$ that establishes local instability is said to be of *interchange* type. The situation is sketched in Fig. 20. The displacement "corrugates" S near P along the direction of $\hat{\mathbf{B}}_0$ and so tends to raise and depress the field lines while bending them only slightly. Bending entails a positive contribution to δW that is stabilizing.

8.2.4 Screw Pinch; Suydam's Criterion So far we have considered instabilities that require S to be deformed. We now consider one that is local within V and that, therefore, exists even when S is undisplaced. It is again of interchange type, and it threatens the sta-

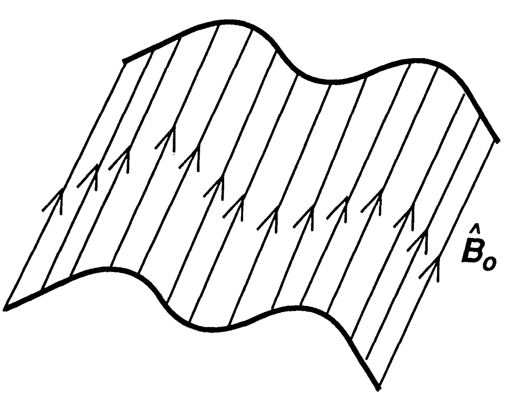

FIG. 20. Interchange instability. The displacement is dangerous for the stability of a system, because it moves the field lines with little bending and therefore wastes little in increasing the magnetic energy.

bility of the general screw pinch shown in Fig. 13(a). The putative instability involves a singular surface, S_0, within the plasma. The word "singular" is especially significant when the stability of the configuration is studied by the normal-mode method: S_0 corresponds to a singularity of the differential equation governing ξ_s (see Sec. 9.4). In the energy method, the instability is studied by the calculus of variations, and S_0 corresponds to a singularity of the Euler–Lagrange equation governing the minimizing ξ_s.

The displacement envisaged has the form $\boldsymbol{\xi} \propto \exp[i(m\phi + kz)]$, and the singular surface S_0 is defined by $F = 0$, where

$$F = kB_{0z} + mB_{0\phi}/s; \qquad (123)$$

$F = 0$ means that the pitch, $s\zeta = B_{0\phi}/B_{0z}$, of the helix defined by $\mathbf{B}_0$ on S_0 coincides with the pitch, $-ks/m$, of the helix on which the phase of displacement is constant, i.e, the perturbation moves the field lines on S_0 without bending them. The perturbation also bends them very little in a singular layer surrounding S_0 (see Sec. 9). The larger the shear, $d\zeta/ds$, the thinner the singular layer and the more stable the equilibrium. A variational argument reveals that the level s is unstable to the modes $\boldsymbol{\xi} \propto \exp[i(m\phi + kz)]$ for which $F = 0$ if *Suydam's criterion* is satisfied:

$$-\frac{dp_0}{ds} > \frac{s}{8\mu_0}\left[\frac{B_{0z}}{\zeta}\frac{d\zeta}{ds}\right]^2. \qquad (124)$$

9. RESISTIVE EFFECTS

Magnetic flux is "frozen" to a perfect conductor according to Alfvén's theorem, but finite resistivity ($\eta \neq 0$) allows flux to drift relative to the moving conductor at a speed of order $\mathcal{V}_d = \eta/\mathcal{L}$. The topology of field lines cannot be altered in a perfect conductor, but a finite resistivity allows the field lines to be severed and reconnected in new ways that change the field-line topology. Not unexpectedly, both these phenomena are very significant over the τ_η time scale. What is more surprising is that they can also be significant over periods of much shorter duration.

It may be recalled here that τ_η was defined to be $\mathcal{L}^2/\mathcal{V}$, where $\mathcal{L}$ was, somewhat vaguely, said to be "a length characteristic of the field." The central idea was that $\mathcal{B}/\mathcal{L}$ should be of the order of the field gradient $\nabla\mathbf{B}$ so that $1/\tau_\eta$ is a good estimate of the diffusion term $\eta\nabla^2\mathbf{B}$ in the induction equation (22). Similarly, the magnetic Reynolds number, $R = \mathcal{V}/\mathcal{V}_d$, involves $\mathcal{V}$, "a velocity characteristic of the flow," the central idea being that it should measure the relative sizes of the induction and diffusion terms in Eq. (22).

Portmanteau descriptions, such as τ_η and R, cannot reliably describe a system in detail, simply because no single $\mathcal{L}$ can represent the spectrum of length scales it possesses and no single $\mathcal{V}$ can reflect the range of $|\mathbf{V}|$ present. For example, rapid fluid motions can generate high field gradients by sweeping together fluid carrying similarly directed flux, so creating flux tubes or flux sheets (see Secs. 9.1 and 9.2); similarly, by bringing differently directed fields together, they can create regions where flux is destroyed or reconnected (see Secs. 9.1 and 9.3). The field gradients may be large in some parts of the system but comparatively small in others, and em diffusion will proceed rapidly in the former but comparatively slowly in the latter. In regions where $\mathcal{V}$ is large, the field will be swept along with the flow, much as envisaged by Alfvén's theorem, but where $\mathcal{V}$ is small, the diffusion of field will be more significant than its advection. Qualitatively, we may regard R, τ_η, and $\mathcal{V}_d$ as functions of position $\mathbf{x}$. Depending on their values, there may be considerable diffusion and rearrangement of magnetic flux in some parts of the system, but virtually none elsewhere. And the influence of the diffusively dominated regions may be far reaching.

9.1 Flux Concentration and Destruction

We consider here two simple but illuminating examples in which the steady flow

$$\mathbf{V} = \zeta(-x\mathbf{e}_x + y\mathbf{e}_y), \qquad \zeta = \text{constant} \qquad (125)$$

maintains a steady unidirectional magnetic field $\mathbf{B} = B(x)\mathbf{e}_y$. Equation (21) gives

$$\eta\frac{dB}{dx} + \zeta xB = E_0, \qquad (126)$$

where $E_0\mathbf{e}_z$ is a (constant) electric field. The steady equation of motion (39) is also obeyed (for $\mathbf{g} = 0$), with

$$p = p_0 - \tfrac{1}{2}\rho V^2 - \frac{B^2}{2\mu_0}, \quad p_0 = \text{constant}. \tag{127}$$

In the first example, we suppose as in Sec. 3 that the implied distant walls, $z = \pm c$, are perfectly conducting and shortcircuited; i.e., $E_0 = 0$. Then Eq. (126) shows that

$$B = B_0 \exp(-x^2/l^2), \quad B_0 = \text{constant}, \tag{128}$$

where $l = \sqrt{2\eta/\zeta}$. This solution describes a *flux sheet* of thickness l created by inflow of fluid toward $x = 0$, the inward advection of flux being equal but opposite to the outward diffusion of flux relative to the conductor. As in Sec. 6.3, magnetic energy is created from kinetic energy by the compression of the flux sheet, this energy being dissipated by Ohmic resistance. In a similar way, the motion $\mathbf{V} = \zeta(-s\mathbf{e}_s + 2z\mathbf{e}_z)$ can steadily maintain the *flux rope*, $\mathbf{B} = B_0\mathbf{e}_z \exp(-s^2/l^2)$.

In the second example, we take $E_0 = \eta B_0/l$. Equation (126) gives

$$B = B_0 e^{-x^2/l^2} \int_0^{x/l} e^{\xi^2}\, d\xi. \tag{129}$$

Clearly, $B(0) = 0$, and the field is oppositely directed on each side of $x = 0$. It is advected toward $x = 0$, where the oppositely directed fields cancel each other out mostly in a layer of thickness l surrounding $x = 0$. Again, the Poynting flux is everywhere directed toward $x = 0$. Again, the overall Ohmic dissipation is made good by the rate of working of the fluid against the Lorentz forces, but now this is not true in detail: the Lorentz forces are driving for $x < 0.925l$. A similar phenomenon will be met in Sec. 9.3.

9.2 Flux Expulsion

Suppose that space is filled with a solid conductor that contains a spherical inclusion, of radius $\mathcal{L}$, in perfect electrical contact with its surroundings; initially, a uniform field $\mathbf{B}_0 = B_0\mathbf{e}_y$ is present. At $t = 0$, the sphere starts to rotate with angular velocity $\boldsymbol{\omega} = \omega\mathbf{e}_z$, where $R\omega \gg 1$. An observer sitting on the surface of the sphere would see the applied field as oscillating with frequency ω, and (as in Sec. 5) that field would eventually only penetrate into the conductor to a depth $O(\delta_\eta)$. The field initially present within the sphere has been *expelled*.

The rotating conductor need not be solid; it could equally be a circulation in a fluid, such as a convection cell. If the overturning time of the cell is short compared with τ_η, any flux that threaded the cell initially, before overturning commenced, would be swept out of the cell, to its sides, corners (if any), and (in the case of hexagonal cells) center. There are many examples of such *flux expulsion* in MHD; one 2D case is shown in Fig. 21. This solution of the induction equation (22) illustrates what happens in a convection roll of square cross section (side d) to a uniform applied field when a circulation is "switched on" at $t = 0$; the value of the magnetic Reynolds number, $R = Ud/\eta$, based on the vertical velocity U at the center of the cell, is 1000. It may be seen that, during the evolution of $\mathbf{B}$ shown, the field lines sever and reconnect (see Sec. 9.3). By time $t = 4d^2/\eta$, the flux originally threading the cell has been concentrated by the flow into sheets confined to boundary layers on the surfaces $x = \pm\frac{1}{2}d$. It has been totally expelled from the interior of the cell. It is by such processes of flux expulsion that a magnetic field becomes *intermittent* when R is large. Such intermittency is observed in the turbulent solar wind.

Fluid motions can also expel field from a fluid into an insulator. Imagine, for example, an ocean $y < 0$ lying beneath a solid insulator $y > 0$ and suppose that a field $\mathbf{B}_0 = -\zeta y\mathbf{e}_x$ is initially present in the fluid (ζ = constant), with $\mathbf{B} = 0$ in $y > 0$. Now, let a steady upwelling of fluid start that presses lines of force near $x = 0$ against the interface $y = 0$; see Fig. 22(a), where streamlines of this $\mathbf{V}$ and the initial field $\mathbf{B}_0$ are indicated. If the fluid is perfectly conducting, the field is unable to detach itself from the fluid and pass into the insulator. If $\eta \neq 0$, however, the field near $y = 0$ will diffuse into the insulator, as sketched in Fig. 22(b). It has been proposed that such a mechanism carries magnetic flux generated in Earth's core into Earth's mantle. In the case where $y > 0$ is free space, there is a second possibility: even if $\eta = 0$, the inertia of the upwelling fluid will carry it, and the trapped flux it contains, above $y = 0$. Gravity now drains the raised field lines of fluid, which returns to the ocean, leaving the flux tube evacuated. In this way, field is carried into

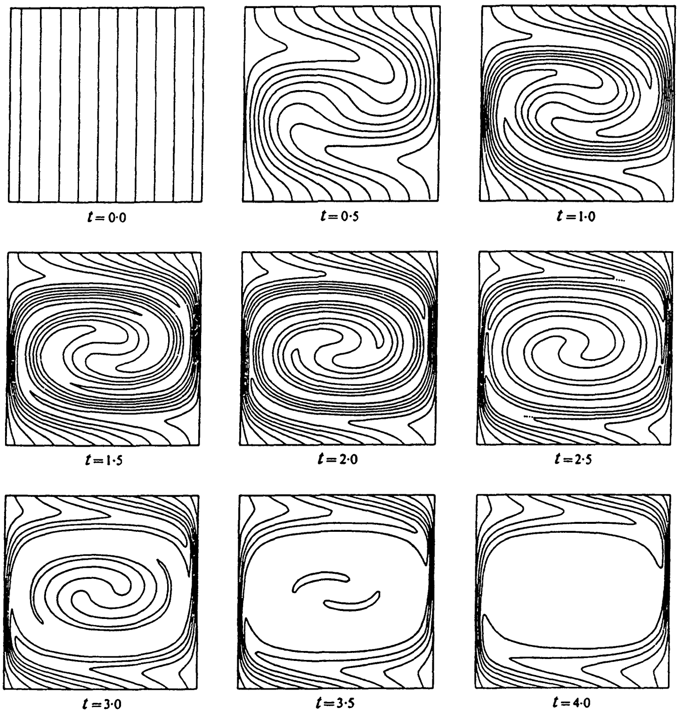

FIG. 21. Flux expulsion from a convective roll. The flow $\mathbf{V} = (\partial\psi/\partial y, -\partial\psi/\partial x, 0)$ is assumed as a model of flow in a convection roll parallel to $\mathbf{e}_z$, where Ox and Oz are horizontal and Oy is vertically upward. The cell is defined by $-\frac{1}{2}d < x < \frac{1}{2}d$, $-\frac{1}{2}d < y < \frac{1}{2}d$, and $-\infty < z < \infty$, and the externally applied field $B_0\mathbf{e}_y$ is vertical. The flow, $\psi = -(U/\pi d)(d^2 - 4y^2)^4 \cos(\pi x/d)$, is switched on at $t = 0$. The subsequent evolution of $\mathbf{B}$ is shown at equally spaced intervals of time, made dimensionless by division by d^2/η. The magnetic Reynolds number $R = Ud/\eta$ is 1000. (Figures courtesy of N. O. Weiss.)

the vacuum where no field previously existed. A similar process is thought to accelerate significantly the rise (due to magnetic buoyancy; see Sec. 10.1) of magnetic flux in the solar convection zone and to explain how it bursts out of the solar surface into its atmosphere.

9.3 Field-Line Reconnection

One of the most celebrated examples of field-line reconnection and its consequences is the so-called "stretch-twist-fold" mechanism sketched in Fig. 23. Three steps (a)–(c) are performed so rapidly that diffusion has

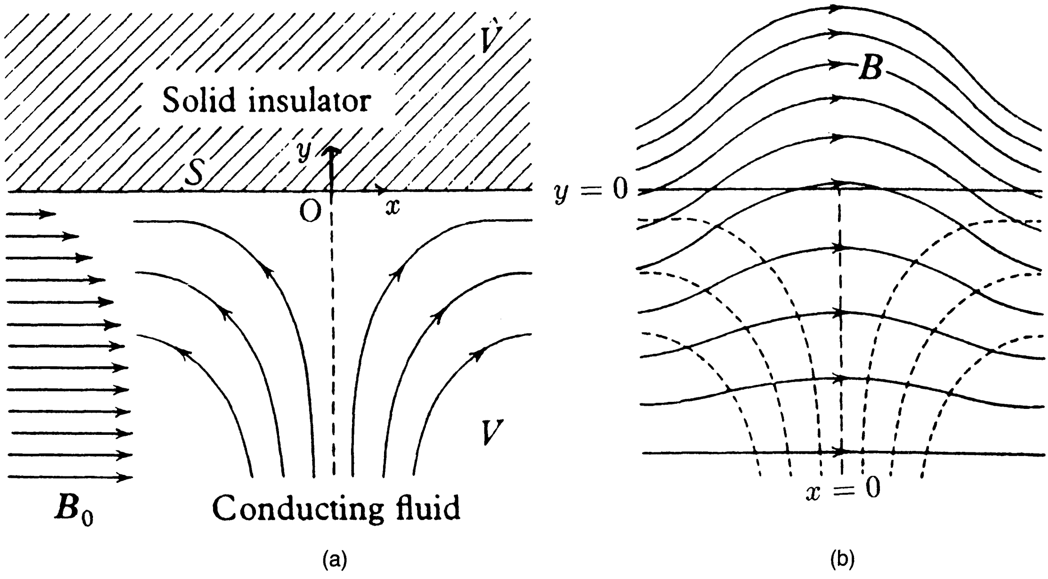

FIG. 22. Flux expulsion from a fluid into an insulator. (a) The initial field and flow; and (b) the final steady field.

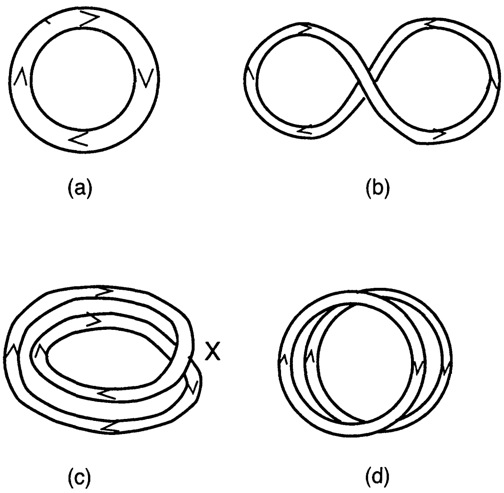

FIG. 23. The stretch-twist-fold process. The process described in the text creates, from one flux tube, two flux tubes, each of twice the energy. The loop (a) is stretched and twisted to form (b). The figure eight (b) is then folded over to produce (c). The field gradients are then high near X, and rapid field diffusion severs and reconnects the field lines near X to produce (d), i.e., two loops like (a).

no time to operate and Alfvén's theorem applies. The flux tube shown in (a) is stretched to double its length and four times its magnetic energy (see Sec. 6.3); one-half of it is twisted through 180°, as illustrated in (b). Then, like closing an open book, one-half is folded over onto the other, as shown in (c). The two halves of the loop are still connected at X on panel (c), where the field gradients are large and diffusion operates rapidly. It is plausible that, through diffusion, the two halves of the loop will separate from one another at X, so creating two flux tubes (d), each of twice the energy of the starting loop (a) (see also Sec. 10.3).

Processes of reconnection are significant in plasma physics and astrophysics (see Secs. 6.2 and 10.2). They are often completed during a time short compared with τ_η, as defined by the dimension $\mathscr{L}$ of the large-scale field structures. Indeed, it may take a time not much longer than $\tau_A = \mathscr{L}/V_A$, independent of η. This is because reconnection can be confined to regions small compared with $\mathscr{L}$, as the 2D Petschek mechanism sketched in Fig. 24 illustrates. This is a model of field-line reconnection in an incompressible fluid in the neighborhood of an X-type neutral point (or more precisely a line, Oz, of X-type neutral points). Flow toward the origin O from the right and left is so rapid that diffusive effects are at first negligible, and the oppositely directed flux is swept with the fluid toward O where it creates such large field gradients that diffusion can no longer be ignored. The field lines are reconnected in a region symbolized by the dashed rectangle. In their new configuration (above and below O in Fig. 24), they are highly bent and, in straightening, they catapult the fluid out in the general directions of $y = \pm\infty$ and at the same time generate a suction that assists new flux to be drawn in from $x = \pm\infty$.

The rapidity of reconnection in Petschek's mechanism arises from the Lorentz forces, which (as for the second model of Sec. 9.1) drive the fluid motions but which (unlike the second model of Sec. 9.1) cannot be balanced by a pressure gradient. They accelerate the outgoing fluid to Alfvénic speeds, i.e., $A \sim 1$, while the ingoing fluid is almost in magnetostatic balance. The inflow and outflow regions are separated by *Alfvén lines*, OA$_i$ (i = 1–4), defined as curves through O to which

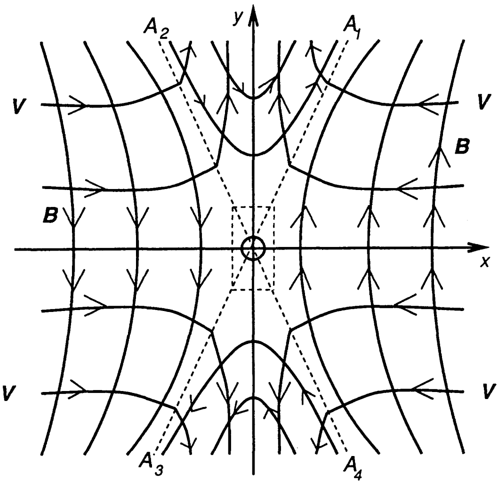

FIG. 24. The Petschek mechanism of field line reconnection at an X-type neutral point. The 2D flow is from the right and left toward the reconnection region (indicated by the dashed rectangle). The reconnection results in Lorentz forces that propel the fluid outward, toward the top and bottom of the figure. The tangential velocity is discontinuous along the Alfvén lines, OA_i (i = 1–4).

$\mathbf{V} - \mathbf{V}_A$ are tangential. The field direction changes rapidly across these slow "shocks." The current densities are large within them, and Joule losses are therefore significant (see Sec. 10.2).

The Petschek mechanism has been very thoroughly analyzed, not only for incompressible but also for compressible fluids, in which the discontinuities arise at slow shock fronts. See Soward (1982), where earlier references may be found. Slow shocks have been detected in the vicinity of the neutral point X_2 in the magnetotail of Earth (Fig. 10), suggesting that the Petschek mechanism is operating there.

The phenomenon of field-line reconnection has a bearing on the variational properties of force-free fields. It was remarked in Sec. 7.2.1 that, when the correspondence between the points of entrance and exit of field lines crossing a volume V is specified, along with B_n everywhere on S (the boundary of V), the state of minimum energy is force-free but that minimum is greater than the unconstrained minimum possible when B_n alone is given. The addition of resistivity opens up the possibility that, by reconnecting field lines, the system can release magnetic energy and evolve toward the lower energy state of the potential field (see Sec. 10.2).

In Sec. 7.2.2, a field confined within a volume V bounded by a perfectly conducting surface S was considered, such as that shown in Fig. 12. It was remarked in Sec. 6.4 that the rate of dissipation of magnetic helicity is much smaller in a turbulent conductor than the rate of dissipation of magnetic energy. It is reasonable to suppose that the *net* helicity, H_{total}, of the entire volume V is approximately conserved during the time that the system takes to reach its minimum-energy state. The field having the least magnetic energy subject to specified H_{total} is a Beltrami field [i.e., a force-free field of constant α; see Eq. (115)], and its energy is smaller than that of the more constrained minimum, discussed in Sec. 7.2.2, in which $H(v)$ is preserved on every magnetic surface in V; the addition of Ohmic dissipation permits field-line reconnection, and the system can release magnetic energy and evolve toward the Beltrami state. This possibility, often called "Taylor relaxation," is significant in thermonuclear research for the formation and stability of the so-called "reversed-field pinch"; see Taylor (1986) and Sec. 9.5.7 of Freidberg (1987). For an astrophysical application, see Aly (1992).

9.4 Resistive Instabilities; Tearing and Coalescence

The sheet pinch provides one of the simplest examples of a resistive instability, the *tearing mode*. The fluid is incompressible and inviscid and is confined between two perfectly conducting planes $z = 0$ and $\mathcal{L}$. Magnetostatic equilibrium prevails in a field of

$$\mathbf{B}_0 = B_{0x}(z)\mathbf{e}_x + B_{0y}(z)\mathbf{e}_y, \tag{130}$$

which alters its direction as z changes; see, for example, the field (116) when $\epsilon = 0$. As in Sec. 7, it is supposed that the Lundquist number S is large.

The field (130) is ideally stable, and its instability relies on field-line reconnection. The growth rate λ of the instability must therefore tend to zero as $\eta \rightarrow 0$; i.e., it must be much less than $O(\tau_A^{-1})$, which is the growth rate of the dynamical instabilities discussed in Sec. 8. But the equilibrium (130) will itself evolve through Ohmic diffusion during a time of order τ_η. It is found (see below) that $\lambda^{-1} \ll \tau_\eta$, and it is therefore permissible to treat $\mathbf{B}_0$ as constant in time. In short: because the instability develops over a length scale δ small compared with $\mathcal{L}$, λ is intermediate between τ_A^{-1} and τ_η^{-1}.

The linear stability of the state (130) is determined by studying nontrivial perturbations proportional to $\exp(ik_xx + ik_yy + \lambda t)$; it is found that the relevant λ are real. The key to the instability is a diffusive layer (of thickness δ, say), which surrounds a *critical level* and in which lines of force are detached from the fluid and reconnected. A critical level arises at any z where $\mathbf{B}_0$ is perpendicular to the wave vector $\mathbf{k} = k_x\mathbf{e}_x + k_y\mathbf{e}_y$. It marks a singularity of the ideal stability equations (121) for $\omega = 0$. The reconnection creates a string of *magnetic islands* (or cat's eyes), as illustrated in Fig. 15, where the lines of force of the components of $\mathbf{B}$ lying in the plane defined by Oz and $\mathbf{k}$ are shown. By contracting, the closed field loops can release magnetic energy to fuel the instability.

The growth rate of the instability is partly decided by the rate at which field lines can detach and reconnect themselves from fluid in the critical layer. This depends on the field gradient, δ^{-1}, in the layer, rather than on $\mathcal{L}^{-1}$. Thus, $\lambda \approx \mathcal{L}/\mathcal{V}_d$, where $\mathcal{V}_d \approx \eta/\delta$. At the critical level, the perturbation lifts lines of force bodily, i.e., it is of interchange type. The emf that balances magnetic diffusion in the layer is therefore generated by the small component of field, of order $B_0\delta/\mathcal{L}$, perpendicular to **k** in the critical layer. This emf is of order $\mathcal{V}(B_0\delta/\mathcal{L})$ rather than $\mathcal{V}B_0$. Since Ohmic diffusion destroys flux at the rate η/δ^2, the velocity perturbation must be of order $\eta\mathcal{L}^2\mathcal{B}/B_0\delta^3$, where $\mathcal{B}$ is the field perturbation. Similarly, the Lorentz force is of order $\mathcal{B}(B_0\delta/\mathcal{L})/\mu_0\mathcal{L}$ rather than $\mathcal{B}B_0/\mu_0\mathcal{L}$. Setting this against the inertial force $\lambda\rho\mathcal{V} \approx \eta\rho\mathcal{V}/\mathcal{L}\delta \approx \eta^2\rho\mathcal{L}/B_0\delta^4$, we obtain

$$\delta \approx (\tau_A/\tau_\eta)^{2/5}\,\mathcal{L} = S^{-2/5}\,\mathcal{L}, \tag{131}$$

$$\lambda \approx \mathcal{L}/\tau_\eta\delta = S^{2/5}\tau_\eta^{-1} = S^{-3/5}\tau_A^{-1}. \tag{132}$$

The effects of nonlinearities are considered in the review of White (1983). It might be expected that the linear instability of state (130) results in a new magnetostatic state of finite amplitude that depends on x and y. For example, it might seem plausible that, after an infinitesimal perturbation of the state (116) with $\epsilon = 0$ grew to finite amplitude, it would equilibrate to a configuration resembling state (116), but of finite ϵ. This does not necessarily seem to be the case. A study of state (116) has shown that it too is unstable to a *coalescence instability*, in which cat's eyes merge together to form one larger eye. Despite the implied severing and coalescence of field lines, the time scale of this instability appears scarcely greater than τ_A. It has been suggested that this instability operates in the magnetosphere and explains the presence of giant "plasmoids" that have been observed in the magnetotail, structures that consist of loops of closed field lines and hot plasma [see Fig. 10 and Richard *et al.* (1989)]. In a steady state, the rate at which the lines of force of the interplanetary magnetic field detach at X_1 must equal the rate at which they reconnect in the magnetotail. The speed of the reconnection is limited by the efficiency of the Petschek mechanism, and, if a surge of additional flux in the IMF is brought to X_1 and detached, it will, while waiting to be reconnected at X_2, cause a large increase in the S → N field gradient near the central neutral sheet joining E to X_2, so promoting instabilities of tearing type and further neutral points, such as the one indicated by X_3 in Fig. 10. These both enhance the rate at which the IMF reconnects and explain the observed plasmoids.

Other types of resistive instabilities can also arise in the sheet pinch, notably *rippling modes*, which depend on spatial variations in σ; *g* modes, which depend on the existence of a gravitational field in the $-z$ direction and a density gradient in the $+z$ direction but which ignore the diffusion of density differences; and *convective modes*, which restore density diffusion to the *g* modes, with dramatically different results. An important class of resistive instabilities arises from the curvature of the field lines of $\mathbf{B}_0$ in the cylindrical pinch, i.e., instabilities that occur even when Suydam's criterion would indicate stability. Detailed discussions of some of the associated linear stability problems may be found in Lifschitz (1989).

10. MHD IN THE LARGE

The physics of the Earth, planets, Sun, stars, and galaxies provide a rich fabric in which MHD phenomena are deeply woven. While it would be inappropriate in an encyclopedia devoted to applied physics to enter into these topics in depth, a brief discussion of a few of them supplies not only illustrations of the ideas developed in preceding sections of this article but also deeper insights into MHD processes. For further reading about the topics adumbrated below, see, for example, Parker (1979) and Priest (1982).

10.1 Sunspots; Magnetoconvection; Formation of Flux Tubes

Sunspots are the signatures of a large zonal field in the Sun. It is thought that flux tubes are generated deep within the Sun and that they rise toward the surface; a segment breaks through the surface approximately along a line of latitude and gives rise to two dark areas where it leaves and re-enters the Sun. How are the flux tubes created? Why do they rise? Why are sunspots dark?

Flux tubes rise through *magnetic buoyancy*. Radiative transfer of heat is highly efficient in the Sun except near its surface; i.e., the thermal diffusivity κ is large and the temperature within a flux tube is that of its surroundings. Likewise, mechanical equilibrium ensures that the *total* pressures $P = p + B^2/2\mu_0$ inside and outside the tube are almost the same. But the field within the tube is larger than the field outside, so that the magnetic pressure is greater and the kinetic pressure is less. Since the temperature is the same, the density is less, by the gas law. The tube, therefore, is buoyant. It rises and breaks through the solar surface in the way indicated at the end of Sec. 9.2.

Near the solar surface, radiative transfer is ineffective and heat is transported by thermal convection. Magnetic fields tend to suppress convection, and less heat per unit area is carried out of the Sun through a sunspot than elsewhere on the solar surface. The sunspot therefore appears dark.

A simple example illustrates the inhibiting effect of a magnetic field on convection: the Rayleigh–Bénard problem. This employs a theoretical idealization, the Boussinesq fluid, that avoids inessential complications while preserving the essence of the buoyancy mechanism. The density of this fluid is independent of the pressure but decreases slightly with increasing temperature T, i.e., $d\rho/dT = -\gamma\rho$, where γ, the coefficient of volume expansion, is assumed to be positive and constant. An unbounded horizontal layer, $0 < z < \mathscr{L}$, is filled with Boussinesq fluid, and the temperature of its lower wall ($z = 0$) exceeds that of its upper wall ($z = \mathscr{L}$) by ΔT. A linear stability analysis shows that, if ΔT is increased quasistatically from zero, convection first occurs, as a horizontally tessellated pattern of infinitely weak convection cells, when the *Rayleigh number*,

$$Ra = g\gamma\Delta T\mathscr{L}^3/\nu\kappa, \tag{133}$$

attains a critical value, which is $Ra_c \approx 1707.76$ if $z = 0$ and $z = \mathscr{L}$ are no-slip walls.

If the fluid is electrically conducting and a strong vertical magnetic field, $\mathbf{B}_0 = B_0\mathbf{e}_z$, is present, that field will try to enforce the two-dimensionality of Sec. 3.2. But motions $\mathbf{V} = \mathbf{V}(x, y)$ independent of z can only be horizontal; they cannot carry heat across the layer. A convective circulation must necessarily depend on z. To achieve that three-dimensionality, a larger temperature gradient is needed to defeat the putative two-dimensionality. Thus, $Ra_c(M)$ increases with M and, when $p_m < 1$, it may be shown that $Ra_c \sim \pi^2M^2$ for $M \gg 1$. Evidently, the magnetic field delays the onset of convection.

This model of magnetoconvection may also be used to demonstrate that the shortcomings of linear stability theory are no less severe in MHD than in other sciences. It was already noted in Sec. 8.1 that a state judged stable on linear theory might nevertheless be unstable when perturbed with finite amplitude. When Ra sufficiently exceeds $Ra_c(M)$, the convective circulations are so rapid that the magnetic flux passing through a cell is expelled to the sides, corners, and center of the cell (see Sec. 9.2). Once removed from the interior of a cell, the magnetic field is no longer present to inhibit convection, which can therefore continue even if Ra is reduced far below $Ra_c(M)$; i.e., the layer is *subcritically unstable.*

This gives a second, dynamical reason why intermittency is prevalent in turbulent fluids when R is large: the "preferred" mode of magnetoconvection is one in which the available magnetic flux is concentrated rather than being smoothly spread out.

10.2 Current Sheets; Solar Flares

The atmosphere of the Sun is light and highly conducting. We may expect, therefore, that its magnetic field will be approximately force-free. Matters raised in Secs. 7.2.1 and 9.2 assume a new significance. We are dealing, however, with a domain that is essentially infinite, and a line of force leaving the Sun need not return, being blown away, in effect to infinity, by the solar wind (see Fig. 10). Where a correspondence between footpoints on the solar surface exists, the field is continually modified as the footpoints are moved over the surface of the Sun by convective motion. This motion enhances the energy of the magnetic field in the atmosphere, but that field also continually evolves toward lower energy states by the processes described in Sec. 9.3, dissipating its energy to heat as it does so through the formation of current sheets and electrical discharges, known as solar flares. These processes are even more active in so-called "flare stars."

10.3 Fluid Dynamos

In Sec. 10.1, we asked how the magnetic field associated with sunspots is generated, and we may wonder similarly what the source of the main magnetic fields of the Sun is, to what some stars and planets owe their magnetism, and where the magnetic fields of galaxies originate. This is the provenance of dynamo theory, one of the largest and most rapidly growing branches of MHD, but one that is beyond the scope of this review, although an experiment is currently being constructed.

The geomagnetic field emanates from Earth's core but, if that core were a solid, the field would disappear in a time of order $\tau_\eta \sim$ 14 000 years. Motions in the core must act as a self-excited dynamo. Self-excited dynamos exist; conceptually, the simplest is the homopolar dynamo sketched in Fig. 25. The emf created by a sufficiently rapid rotation of the disk in the magnetic field created by the current in the wire loop generates this current; when the seed field initially necessary to start the regeneration is removed, the induced and inducing field are the same. The energy dissipated Ohmically in the circuitry (and in any external load, indicated by the light bulb in Fig. 25) is made good by the rate of working of the power source maintaining the motion of the disk against the opposing Lorentz force.

If a homogeneous mass of fluid V is to act as a dynamo, the magnetic Reynolds number of the motions within it must be sufficiently large. Since the ratio of the induced field to the inducing field is $O(R)$, it is necessary (but not sufficient) that $R \geq O(1)$. This condition is very hard to satisfy in laboratory conditions but may be attained in the flow of liquid sodium in the cores of fast reactors where $R \approx 50$ (see, for example, Bevir, 1973). Fortunately for the safety of the reactors, no dynamos have yet been excited in their operation, probably because the flow within them has been insufficiently helical. The point here is the following.

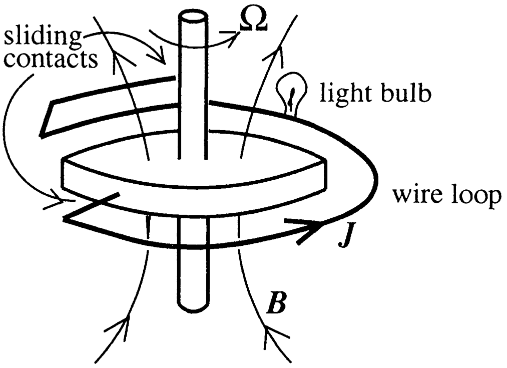

FIG. 25. The homopolar dynamo. The dynamo is self-excited. A magnetic field is created by the current flowing in the almost complete wire loop and that current is created by em induction through the rotation of the disk in the magnetic field. The device does not violate energy conservation: the current flowing between the axis and the perimeter of disk, together with the field produced by the wire loop, produces a Lorentz force that opposes the motion of the disk, and a source of energy is needed to maintain the rotation of the disk against that force.

The homopolar dynamo is man-made and highly asymmetric in structure, but the homogeneous masses of conducting fluid encountered in nature are generally highly symmetric (e.g., spherical). For these to regenerate field, the motions within them must, by *their* asymmetry, make good the lack of asymmetry of the container. Helical motions possess an asymmetry conducive to field generation. These are defined as motions of nonzero *kinetic helicity* H_K, where

$$H_K = \mathbf{V}\cdot\nabla\times\mathbf{V} = \mathbf{V}\cdot\boldsymbol{\omega}. \tag{134}$$

Beltrami motions [in which $\boldsymbol{\omega}$ and $\mathbf{V}$ are parallel; see Eq. (115)] provide some of the simplest models of working fluid dynamos; e.g., the ABC flow (Arnold–Beltrami–Childress),

$$\mathbf{V} = (C\sin z + B\cos y, A\sin x + C\cos z, B\sin y + A\cos x), \tag{135}$$

is the basis of an efficient dynamo. Other dynamos make use of the stretch-twist-fold process sketched in Fig. 23 and the kink instabilities discussed in Sec. 8.2.2. For a recent review of dynamo theory, see Roberts and Soward (1992).

10.4 RMHD

RMHD stands for "Rotating MHD", i.e., the MHD of rotating fluid conductors. Most large naturally occurring bodies are rapidly rotating, in a sense to be made more precise below, and rotating MHD therefore finds significant applications in astrophysics and planetary physics. The importance of Coriolis forces on the dynamics of Earth's fluid core explains why the magnetic compass needle points approximately North–South over

Earth's surface. Their action within the Sun explains why sunspots appear in pairs approximately along one line of latitude (Sec. 10.1). Compared with the classical MHD of nonrotating fluids, RMHD exhibits strange and sometimes almost paradoxical behaviors, even when the working fluid is of constant density and incompressible, as we shall suppose here. For example, consider again the Rayleigh–Bénard convection problem of Sec. 10.1, but now suppose that the configuration rotates rapidly about the vertical. One might have expected a priori that, since rotation is also associated with two-dimensionality (the Proudman–Taylor theorem stated in Sec. 3.2), Ra_c would be increased even beyond the $O(M^2)$ value imposed by the MHD 2D theorem. This is not the case. The magnetic field "breaks the constraint" of the Proudman–Taylor theorem (or alternatively we could say that rotation breaks the constraint of MHD two-dimensionality), and Ra_c is smaller than it would be if either rotation or magnetic field were absent.

Rotating systems are usually most simply described in a frame of reference rotating with a constant angular velocity $\mathbf{\Omega}$ characteristic of the conductor as a whole; $\mathbf{V}$ and $\partial/\partial t$ are now measured in the rotating frame but $\mathbf{B}$ and D/Dt are the same as in the inertial frame; the induction equation, therefore, is unaltered, but the equation of motion (65) becomes

$$\begin{aligned}\partial\mathbf{V}/\partial t + \mathbf{V}\cdot\nabla\mathbf{V} + 2\mathbf{\Omega}\times\mathbf{V}\\ = -\nabla\mathbf{\Pi} + \mathbf{J}\times\mathbf{B}/\rho + \nu\nabla^2\mathbf{V} + \mathbf{g}.\end{aligned} \qquad (136)$$

Here, $2\rho\mathbf{\Omega}\times\mathbf{V}$ is the so-called "Coriolis force"; the centrifugal force, $-\rho\mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{r})$, can be expressed as $\nabla[\frac{1}{2}\rho(\mathbf{\Omega}\times\mathbf{r})^2]$ and included in the gradient of the pressure $\rho\Pi$.

The theory of rotating but nonconducting fluids is a well-developed and important branch of fluid mechanics, with applications to atmospheric and ocean physics (see, for example, Greenspan, 1968). Two nondimensional measures of Ω are particularly significant, the *Ekman number Ek* and the *Rossby number Ro,* that measure, respectively, the viscous and inertial forces against the Coriolis force:

$$Ek = \nu/\Omega\mathcal{L}^2, \qquad Ro = \mathcal{V}/2\Omega\mathcal{L}. \qquad (137)$$

When $Ek \ll 1$ and $Ro \ll 1$, steady flow is almost two-dimensional with respect to the direction of $\mathbf{\Omega}$. This is the Proudman–Taylor theorem referred to in Sec. 3.2. When a magnetic field is present, two other dimensionless groupings arise, the *magnetic coupling parameter C* and the *Elsasser number Λ*:

$$C = V_A^2/2\Omega\mathcal{L}\mathcal{U}, \quad \Lambda = \sigma\mathcal{B}^2/2\Omega\rho = RC. \qquad (138)$$

The former is the ratio of Lorentz force to Coriolis force, with $\mathbf{J}$ estimated as $\mathcal{J} = \mathcal{B}/\mu_0\mathcal{L}$; the latter appeals to Ohm's law and uses instead $\mathcal{J} = \sigma\mathcal{V}\mathcal{B}$.

Rotation mixes the two transverse polarizations of the Alfvén wave, and two different types of wave motion replace the Alfvén wave of Sec. 6. In the simplest case in which the wave vector $\mathbf{k}$ is parallel to both $\mathbf{\Omega}$ and $\mathbf{B}_0$, the frequencies of these waves are $\pm\omega_{F,S}$ where

$$\omega_{F,S} = \sqrt{\Omega^2 + V_A^2k^2} \pm \Omega. \qquad (139)$$

The subscripts F and S refer particularly to the often relevant, "highly rotating" case in which $\Omega \gg V_Ak$ and for which $\omega_F \approx 2\Omega \gg \omega_S \approx V_A^2k^2/2\Omega$. The fast wave is then essentially an inertial oscillation, of the type much studied in the classical theory of rotating fluids, that arises from the vibration of vortex lines that (as seen in the inertial frame) fill the system (see, for example, Greenspan, 1968). Greater interest centers on the slow waves, which have a phase speed of

$$c_S \approx V_A^2/2\Omega\mathcal{L}, \qquad (140)$$

where $\mathcal{L} = 1/k$. This defines a new dynamic time scale $\tau_S = \mathcal{L}/c_S$, and many mechanisms that work on the Alfvén time scale τ_A in a nonrotating fluid (see Sec. 8) operate in a highly rotating fluid on the time scale τ_S instead. (The resistive instabilities studied in Sec. 9.4 are similarly retarded.) The Elsasser number, $\Lambda = \tau_\eta/\tau_S$, is the analogue for the rotating system of the Lundquist number, $S = \tau_\eta/\tau_A$.

RMHD has a particular significance in the studies of Sec. 10.3. Rotation imparts kinetic helicity to a rotating fluid, which is beneficial for dynamo action. Consider, for example, a sinking plume of hot plasma in a rotating star. This plume will be compressed as it moves into higher pressure environments, and, in attempting to conserve its angular momentum, it will start to spin about the vertical, i.e., it will acquire vertical vorticity and therefore nonzero kinetic helicity; the same sense of helicity would be associated with an ascending, expanding plume. By Alfvén's theorem, the helicity of either plume would tend to twist the zonal field of the star into meridian planes, and many such plumes would create a general stellar meridional field, which (Alfvén's theorem again) would be stretched out into the zonal direction by differential rotation in the interior of the star, so recreating the zonal field. If the magnetic Reynolds number is sufficiently large, the dynamo cycle is complete, and the star maintains a permanent magnetic field.[*]

ACKNOWLEDGMENTS

This review has benefited from the advice of Stanislav Braginsky, Andrew Soward, Mark Tillack, Ray Walker, Cheng Chin Wu, and (for the Encyclopedia) George Trigg. I am also grateful to four postgraduate students, Weijia Kuang, Sikun Lan, Neil Morley, and Martin St. Pierre, for acting as guinea pigs and to Maureen Roberts for preparing PostScript[†] files for most of the figures.

NOTATION

a is the speed of sound (m s^{-1})

$\mathbf{A}$ is vector potential for magnetic field (T m)

$\mathbf{B}$ is magnetic field (T)

$\hat{\mathbf{B}}$ is magnetic field in $\hat{\mathrm{V}}$

$\mathcal{B}$ is a characteristic magnetic field strength

c_0 is the speed of light in vacuo (m s^{-1})

C_v and C_p are the specific heats at constant volume and pressure (J kg^{-1} K^{-1})

$\mathbf{D}$ is electric displacement (C m^{-2})

$\mathbf{e}_c$ is a unit vector in the direction of c increasing

E is the total energy density per unit volume (J m^{-3})

E_M is the (electro)magnetic energy density per unit volume (J m^{-3})

$\mathbf{E}$ is electric field (V m^{-1})

$\mathbf{g}$ is the externally applied body force per unit mass, such as gravity (m s^{-2})

$\mathbf{H}$ is magnetizing force (A m^{-1})

H is the enthalpy per unit mass (J kg^{-1})

I is the internal per unit mass (J kg^{-1})

$\mathbf{I}$ is surface current (A m^{-1})

$\mathbf{J}$ is electric current density (A m^{-2})

$\mathbf{k}$ is wave vector; k is wave number (m^{-1})

K is the kinetic energy of a system (J)

K_p is the isothermal compressibility (Pa^{-1})

$\mathbf{L} = \mathbf{J} \times \mathbf{B}$ is the force per unit volume exerted by the (electro)magnetic field on a conductor (N m^{-3})

$\mathcal{L}$ is a characteristic length (m)

m_{ij} is the (electro)magnetic stress tensor (N m^{-2})

p is kinetic pressure (Pa)

$P = p + B^2/2\mu_0$ is the total pressure (Pa)

$p_m = \nu/\eta$ is the magnetic Prandtl number

S is specific entropy (J kg^{-1} K^{-1})

$\mathbf{S} = \mathbf{E} \times \mathbf{B}/\mu_0$ is the em-energy flux (W m^{-2})

t is time (s)

T is temperature (K)

$V = 1/\rho$ is specific volume (m^3 kg^{-1})

V is the volume occupied by conducting fluid

$\hat{\mathrm{V}}$ is the exterior of V, usually assumed insulating

$\mathbf{V}$ is fluid velocity (m s^{-1})

$\mathcal{V}$ is a characteristic flow speed (m s^{-1})

$\mathcal{V}_A = \mathcal{B}/\sqrt{\mu_0\rho}$ is a characteristic Alfvén speed (m s^{-1})

W is the total energy of a system (J), i.e., the integral of E

$\mathbf{x}$ is position (m)

β is the ratio of kinetic pressure to magnetic pressure, i.e., $\mu_0 p/\mathcal{B}^2$

γ is the coefficient of volume expansion (K^{-1})

$\tilde{\gamma}$ is the coefficient of surface tension (kg s^{-2})

δ is boundary layer thickness (m)

ϵ is the permittivity of a material (F m^{-1})

ϵ_ν is the rate of viscous heating per unit volume (W m^{-3})

$\epsilon_\sigma = J^2/\sigma$ is the rate of Ohmic dissipation per unit volume (W m^{-3})

$\eta = 1/\mu_0\sigma$ is magnetic diffusivity (m^2 s^{-1})

Θ is surface charge density (C m^{-2})

ϑ is volume charge density (C m^{-3})

κ is thermal diffusivity (m^2 s^{-1})

λ is growth rate (s^{-1})

μ is the permeability of a material (H m^{-1}); taken to be μ_0 everywhere in the text

---

[*] Recent research indicates that this is not a viable explanation of the Sun's magnetic field e.g., see Krause *et al.* (1993).

[†] PostScript is a trademark of Adobe Systems Inc.

$\mu_0 = 4\pi 10^{-7}$ H m^{-1} is the permeability of free space

ν is kinematic (shear) viscosity (m^2 s^{-1})

$\boldsymbol{\xi}$ is fluid displacement (m)

ϖ is the reduced pressure (m^2 s^{-2}); for a uniform density fluid, this is p/ρ; for a barotropic fluid, for which $p = p(\rho)$, it is $\varpi = \int \rho^{-1} dp$

ρ is (mass) density (kg m^{-3})

σ is electrical conductivity (S m^{-1})

τ is a characteristic time scale (s)

Φ is electrostatic potential (V)

φ is the potential energy per unit mass (m^2 s^{-2}), when $\mathbf{g} = -\nabla\varphi$ is conservative

$\boldsymbol{\omega} = \nabla \times \mathbf{V}$ is vorticity (s^{-1})

Ω is angular velocity (s^{-1})

GLOSSARY

Alfvén Number: See **Mach numbers.**

Alfvén's Theorem: The magnetic field lines in a fluid that is a perfect electrical conductor ($\sigma = \infty$; $\eta = 0$) are carried with it in its motion, as though they were frozen to it.

Alfvén Velocity/Speed: The velocity, $\mathbf{V}_A = \pm\mathbf{B}/\sqrt{\mu_0\rho}$, with which an Alfvén wave moves relative to a fluid conductor; the wave speed, therefore, is $V_A = |\mathbf{V}_A| = B/\sqrt{\mu_0\rho}$.

Alfvén Wave: A wave moving with the Alfvén velocity, relative to a conducting fluid. It travels along, and is transverse to, the magnetic lines of force.

Beltrami Field: A vector field that is a constant multiple of its curl. It is an important special case of the **Force-Free Field.**

Chandrasekhar Number: The square of the **Hartmann number.**

Couette Flow: Flow that is set up in viscous fluid contained between two coaxial, differential rotating cylinders and is everywhere zonal with respect to the common axis of the cylinders; see also **Plane Couette Flow.**

Cylindrical Pinch: An infinitely long circular cylinder of plasma through which passes an axisymmetric magnetic field having a zero component perpendicular to the axis of the cylinder. Also called "the general screw pinch," special cases of which are the **Theta Pinch** and the **Z Pinch.**

Electromagnetic Diffusion Time: The characteristic time, $\tau_\eta = \mathcal{L}^2/\eta$, in which a magnetic field of scale $\mathcal{L}$ in a stationary conductor of magnetic diffusivity η disappears through Ohmic losses.

Ekman Number: A dimensional number, $Ek = \nu/\Omega\mathcal{L}^2$, measuring the relative importance of viscous and Coriolis forces in a rotating fluid.

Elsasser Number: A dimensional number, $\Omega = \sigma\mathcal{B}^2/2\Omega\rho$, measuring the relative importance of Lorentz and Coriolis forces in a rotating fluid.

Field Lines: See magnetic lines of force.

Force-Free Field: A magnetic field that exerts no Lorentz force on a conductor; i.e., $\mathbf{J} \times \mathbf{B} = 0$. Thus, $\mathbf{B}$ is parallel to its curl, but (unlike the Beltrami field) the constant of proportionality can differ from one field line to another.

Hartmann Flow: Flow in a viscous, electrically conducting fluid contained between two stationary parallel planes, driven by a pressure gradient parallel to those planes, in the presence of a magnetic field applied perpendicularly to those planes. It is the MHD generalization of **Plane Poiseuille Flow.**

Hartmann Layer: A boundary layer that may exist at the edge of an MHD flow when the Hartmann number is large; its thickness is of order $(\rho\nu/\sigma B_n^2)^{1/2}$, where B_n is the component of the magnetic field perpendicular to the boundary.

Hartmann Number: A dimensionless measure, $M = (\sigma/\rho\nu)^{1/2}\mathcal{B}\mathcal{L}$, of the strength of the magnetic field.

Helicities: Pseudoscalars measuring the lack of mirror symmetry of fields and defined as the scalar products of the fields with their curls; thus, $\mathbf{V}\cdot\boldsymbol{\omega}$ is kinetic helicity; magnetic helicity is $\mathbf{A}\cdot\mathbf{B}$, where $\mathbf{A}$ is a vector potential for the magnetic field $\mathbf{B} = \nabla \times \mathbf{A}$.

Interaction Parameter: A dimensionless measure, $N = \sigma\mathcal{B}^2\mathcal{L}/\rho\mathcal{V}$, of the relative importance of Lorentz and inertial forces.

Interchange Instability: A mode of instability characterized by the local displacement of magnetic field lines without bending them significantly.

Intermittency: A characteristic of turbulence, in which vorticity and/or magnetic field are concentrated into comparatively small fractions of the fluid volume, so that measurement of these quantities at a fixed point is highly intermittent in time, being large only while the concentrations pass through that point.

Lundquist Number: A dimensionless number measuring the ratio, $S = \tau_\eta/\tau_A = \mathcal{V}_A\mathcal{L}/\eta$, of the electromagnetic diffusion time $\tau_\eta =$

$\mathcal{L}^2/\eta$ and the time scale $\tau_A = \mathcal{L}/\mathcal{V}_A$ of Alfvén wave propagation.

Mach Numbers: Dimensionless ratios of the characteristic flow speed $\mathcal{V}$ and wave speeds. The magnetic Mach number or Alfvén number, $A = \mathcal{V}/\mathcal{V}_A$, differs from the hydrodynamic Mach number, $M_a = \mathcal{V}/a$, by the replacement of the sound speed a by a characteristic **Alfvén Speed,** $\mathcal{V}_A$.

Magnetic Buoyancy: The tendency of magnetic flux tubes, in thermal and mechanical balance with their surroundings, to rise in a gravitational field because of their lower density.

Magnetic Diffusivity: The diffusivity $\eta = 1/\mu_0\sigma$ that determines the rate at which electric currents decay through the finite electrical conductivity σ of a material.

Magnetic Flux Tube: A volume composed of a bundle of magnetic lines of force. The net flux of magnetic field across any cross section, the strength of the tube, is the same for every cross section. The surface of a flux tube is a **Magnetic Surface.**

Magnetic Lines of Force: Curves that are everywhere tangential to the direction of the magnetic field **B**; also called "(magnetic) field lines".

Magnetic Pressure: The pressure $B^2/2\mu_0$ exerted on a conductor by the magnetic field **B** it contains.

Magnetic Response Time: The characteristic time $\tau_m = \rho/\sigma\mathcal{B}^2$ for fluid flow to respond to the Lorentz forces created by the currents it induces from the prevailing magnetic field.

Magnetic Surface: A surface composed of magnetic field lines; i.e., one on which the normal component B_n of the magnetic field is everywhere zero and which therefore contains its own flux. A particular form of a magnetic surface is the **Magnetic Flux Tube.**

No-Slip Condition: The statement that a viscous fluid in contact with a solid wall moves with the same velocity as that wall.

Pinch: An electrically conducting gas compressed by magnetic forces and commonly modeled in one dimension as either a sheet pinch or a **Cylindrical Pinch** and in two dimensions as a toroidal pinch.

Plane Couette Flow: Flow set up in a viscous fluid contained between two parallel planes moving relative but parallel to one another.

Plane Poiseuille Flow: Flow in a viscous fluid contained between two stationary parallel planes and maintained by a pressure gradient parallel to those planes.

Plasma: A partially or completely ionized gas that is therefore a conductor of electricity.

Prandtl Numbers: Dimensionless measures of relative diffusion rates. The magnetic Prandtl number $p_m = \nu/\eta$ differs only from the originally defined hydrodynamic Prandtl number $p_r = \nu/\kappa$ by the replacement of the thermal diffusivity κ by the **Magnetic Diffusivity** η.

Proudman–Taylor Theorem: The slow, steady motion of an inviscid rotating and nonelectrically conducting fluid is two-dimensional with respect to the axis of rotation.

Rayleigh Numbers: Dimensionless measures of the importance of buoyancy forces, such as $g\alpha\Delta T\mathcal{L}^3/\nu\kappa$, where g is the acceleration due to gravity and ΔT is the temperature contrast across the system.

Reduced Pressure: The ratio $\varpi = p/\rho$ in a fluid of constant density; in a barotropic fluid $[p = p(\rho)]$, it is $\varpi = \int dp/\rho$; in isentropic conditions ($\nabla S = 0$), it is the enthalpy H.

Reversed-Field Pinch: A **Cylindrical Pinch** in which the magnetic field parallel to the axis of the cylinder is in opposite directions at different distances from that axis.

Reynolds Numbers: Dimensionless measures of the relative importance of the advection of fields by motion and their dissipation through diffusive loss. The magnetic Reynolds number, $R = \mathcal{V}\mathcal{L}/\eta$, differs only from the originally defined hydrodynamic Reynolds number, $Re = \mathcal{V}\mathcal{L}/\nu$, by the replacement of the kinematic viscosity ν by the **Magnetic Diffusivity** η.

Rossby Number: A dimensionless number, $Ro = \mathcal{V}/2\Omega\mathcal{L}$, measuring the relative importance of inertial and Coriolis forces in a rotating fluid.

Shear Layer: A layer in which the fluid velocity tangential to it changes rapidly.

Shercliff Layer: A boundary layer that may exist at the edge of an MHD flow when the **Hartmann Number** is large and when the magnetic field has no component normal to the boundary; also known as a "side-wall layer"; its thickness is of order $(\rho\nu\mathcal{L}^2/\sigma\mathcal{B}^2)^{1/4}$.

Suydam's Criterion: A local criterion for the stability of a general screw pinch, i.e., a cylindrical pinch in which the magnetic field has components both parallel to the axis of the cylinder and zonally about it.

Tearing Mode: An instability of a magnetic configuration that is resistive, i.e., that owes its existence to finite electrical conductivity, which allows field lines to sever and reconnect in ways disallowed by Alfvén's theorem and so to release magnetic energy to power the instability.

Theta Pinch: A **Cylindrical Pinch** in which the electric current is everywhere parallel to the axis of the cylinder so that the magnetic field is everywhere zonally directed with respect to that axis.

Total Pressure: The sum of kinetic pressure p and **Magnetic Pressure.**

Z Pinch: A **Cylindrical Pinch** in which the magnetic field is everywhere parallel to the axis of the cylinder.

Works Cited

Akhiezer, A. I., Lubarski, G. J., Polovin, R. V. (1959), *Sov. Phys. JETP* **8**, 507–511.

Alpher, R. A., Hurwitz, H., Jr., Johnson, R. H., White, D. R. (1960), *Rev. Mod. Phys.* **32**, 758–768.

Aly, J. J. (1992), *Solar Phys.* **138**, 133–162.

Baker, R. S., Tessier, M. J. (1987), *Handbook of Electromagnetic Pump Technology*, Amsterdam: Elsevier.

Bernstein, I. B., Frieman, E. A., Kruskal, M. D., Kulsrud, R. M. (1958), *Proc. R. Soc. Lond.* **A244**, 17–40.

Bevir, M. K. (1973), *J. Brit. Nucl. Eng. Soc.* **12**, 455–458.

Braginsky, S. I. (1960), *Sov. Phys. JETP* **37**, 1005–1014.

Drazin, P. G., Reid, W. H., (1981), *Hydrodynamic Stability*, Cambridge, U.K.: Cambridge University Press.

Dungey, J. W. (1961), *Phys. Rev. Lett.* **6**, 47–48.

Fadeev, V. M., Kvartskhava, I. F., Komarov, N. N. (1965), *Nucl. Fus.* **5**, 202–209.

Flowers, E., Ruderman, M. A. (1977), *Astrophys. J.* **215**, 302–310.

Freidberg, J. P. (1987), *Ideal Magnetohydrodynamics*, New York: Plenum.

Germain, P. (1960), *Rev. Mod. Phys.* **32**, 951–958.

Greenspan, H. P. (1968), *The Theory of Rotating Fluids*, Cambridge, U.K.: Cambridge University Press.

Jackson, W. D. (1983) in: H. Branover, P. S. Lykoudis, A. Yakhot (Eds.), *Liquid-Metal Flows and Magnetohydrodynamics*, New York: American Institute of Aeronautics and Astronautics, pp. 193–215.

Kantrowitz, A., Petschek, H. E. (1966), in: W. B. Kunkel (Ed.), *Plasma Physics in Theory and Application*, New York: McGraw Hill, pp. 148–206.

Krause, F. (Ed.) (1993), *The Cosmic Dynamo*, Dordrecht: Kluver.

Kulikovsky, A. G. (1973), *Fluid Dynamics* **8**, 462–467.

Landau, L. D., Lifshitz, E. M. (1987), *Fluid Mechanics*, 2nd ed., Oxford: Pergamon.

Landau, L. D., Lifshitz, E. M., Pitaevskii, L. P. (1984), *Electrodynamics of Continuous Media*, 2nd ed., Oxford: Pergamon.

Lifschitz, A. E. (1989), *Magnetohydrodynamics and Spectral Theory*, Dordrecht: Kluwer.

Malang, S., Leroy, P., Casini, G. P., Mattas, R. F., Strebkov, Yu. (1991), *Fusion Engineering and Design* Vol. 16, Amsterdam: North–Holland, pp. 95–109.

Moffatt, H. K., Zaslavsky, G. M., Compte, P., Tabor, M. (Eds.) (1992), *Topological Aspects of the Dynamics of Fluids and Plasmas*, Dordrecht: Kluwer.

Moreau, R. (1990), *Magnetohydrodynamics*, Dordrecht: Kluwer.

Parker, E. N. (1979), *Cosmic Magnetic Fields*, Oxford: Clarendon Press.

Pierson, E. S., Herman, H., Petrick, M. (1983), in: H. Branover, P. S. Lykoudis, A. Yakhot, (Eds.), *Liquid-Metal Flows and Magnetohydrodynamics*, New York: American Institute of Aeronautics and Astronautics, pp. 127–137.

Priest, E. R. (1984), *Solar Magnetohydrodynamics*, Dordrecht: Reidel.

Reed, C. B., Picologlou, B. F. (1989), *Fus. Technol.* **15**, 705–715.

Richard, R. L., Walker, R. J., Sydora, R. D., Ashour-Abdalla, M. (1989), *J. Geophys. Res.* **94A**, 2471–2483.

Roberts, P. H., Soward, A. M. (1992), *Annu. Rev. Fluid Mech.* **24**, 459–512.

Shercliff, J. A. (1962), *The Theory of Electromagnetic Flow-Measurement*, Cambridge, U.K.: Cambridge University Press.

Soward, A. M. (1982), *J. Plasma Phys.* **28**, 415–443.

Stewartson, K. (1960), *J. Fluid Mech.* **8**, 82–96.

Taylor, J. B. (1986), *Rev. Mod. Phys.* **58**, 741–763.

Waechter, R. T. (1968), *Proc. Camb. Philos. Soc.* **64**, 871–894.

White, R. B. (1983), *Handbook Plasma Phys.* **1**, 611–676.

Wu, C. C. (1992), in: M. Shearer (Ed.), *Viscous Profiles and Numerical Methods for Shock Waves*, Philadelphia: SIAM, pp. 209–236.

Further Reading

Bateman, G. (1978), *MHD Instabilities*, Cambridge, MA: MIT Press.

Beer-Sheeva International Seminar on Magnetohydrodynamic Flows and Turbulence, published every three years.

Biskamp, D. (1993), *Nonlinear Magnetohydrodynamics,* Cambridge Monographs on Plasma Physics, Vol. 1, Cambridge, U.K.: Cambridge University Press.

Cross, R. C. (1988), *An Introduction to Alfvén Waves*, Philadelphia: Adam Hilger.

Fusion Technology. A publication of the American Nuclear Society and the European Nuclear Society, started in 1984 and containing many papers on MHD.

Lielpeteris, J., Moreau, R. (Eds.) (1989), *Liquid Metal Magnetohydrodynamics*, Dordrecht: Kluwer.

Magnetohydrodynamics. Translations by the Consultants Bureau, New York, of the journal of the same name published by the Latvian Academy of Sciences.

Manheimer, W. M., Lashmore–Davies, C. N. (1989). *MHD and Microinstabilities in Confined Plasmas*, New York: Adam Hilger, Part I, pp. 3–116.

Mercier, C. (1987), *Lectures in Plasma Physics: The Magnetohydrodynamic Approach to Plasma Confinement in Closed Magnetic Configurations*, 2nd ed., Luxembourg: Euratom.

Moffatt, H. K. (1978), *Magnetic Field Generation in Electrically Conducting Fluids*, Cambridge, U.K.: Cambridge University Press.

Moffatt, H. K. Proctor, M. R. E. (Eds.) (1984) *Metallurgical Applications of Magnetohydrodynamics*, London: Metals Society.

Parker, E. N. (1994), *Spontaneous Current Sheets with Applications to Stellar X-ray Emission*, Oxford, U.K.: Oxford University Press.

Polovin, R. V., Demutski, V. P. (1990), *Fundamentals of Magnetohydrodynamics*, New York: Consultants Bureau.

Priest, E. R., Hood, A. W. (1991), *Advances in Solar System Magnetohydrodynamics*, Cambridge, U.K.: Cambridge University Press.

Shercliff, J. A. (1965), *A Textbook of Magnetohydrodynamics*, Oxford: Pergamon.

Sutton, G. W., Sherman, A. (1965), *Engineering Magnetohydrodynamics*, New York: McGraw Hill.

Symposia on *Engineering Aspects of Magnetohydrodynamics* are organized by MIT. Proceedings no. 27 appeared in 1989.

Wu, C. C. (1990), *J. Geophys. Res.* **95A,** 8149–8175.

MAGNETOSPHERES OF THE EARTH AND PLANETS

W. I. Axford, *Max-Planck-Institut für Aeronomie, Katlenburg-Lindau, Germany*

	Historical Introduction	187
1.	**The Configuration of the Earth's Magnetosphere**	189
1.1	The Polar Cusp	190
1.2	The Magnetotail	190
1.3	Convection	191
1.4	The Polar Wind	191
1.5	The Exosphere	192
1.6	The Radiation Belts	192
1.7	Auroral Phenomena	192
2.	**The Dynamics of the Earth's Magnetosphere**	195
2.1	Magnetic Reconnection	195
2.2	Magnetospheric Substorms	196
2.3	Energy Budgets	197
2.4	Particle Acceleration	198
3.	**Other Magnetospheres**	199
3.1	"Magnetospheres" of Nonmagnetized Bodies	199
3.2	The Magnetosphere of Mercury	200
3.3	The Magnetosphere of Jupiter	200
3.4	The Magnetosphere of Saturn	201
3.5	The Magnetosphere of Uranus	202
3.6	The Magnetosphere of Neptune	202
3.7	The Magnetosphere of Pluto–Charon	203
	Glossary	203
	Works Cited	204
	Further Reading	205

HISTORICAL INTRODUCTION

Magnetospheric physics has its origins in the study of geomagnetic variations, the polar aurorae, and the ionosphere of the Earth and is also closely connected with solar physics—in particular, sunspots, prominences, flares, and the solar wind (e.g., Chapman, 1967). With the advent of the space era, there has been a dramatic advance in our knowledge and understanding of the subject, which is of key importance since the magnetosphere of the Earth, in particular, is the most accessible natural "laboratory" for the study of collisionless plasmas in an astrophysical context.

Transient geomagnetic variations were first discovered in 1722 by Graham, and their association with the appearance of aurora was first noted in 1741 by Celsius. In 1858, Broun discovered a tendency for geomagnetic disturbances to recur after about 4 weeks, an effect later rediscovered by Maunder who associated it with solar rotation. A short-lived brightening on the Sun was observed by Carrington and Hodgson in 1859, and this was found to have been associated with one of the greatest episodes of geomagnetic activity (i.e., a geomagnetic "storm") to be observed in the 19th century. Such solar "flares" were not seen again until the introduction of the spectroheliograph in 1910 by Hale. The more frequent observations which then became possible showed that flares are often followed within a day or so by a geomagnetic storm with a sudden commencement (SC). Hale also discovered that sunspots are associated with magnetic fields, which are a crucial component of the flare phenomenon.

Following his observations of a great aurora which occurred over northern Europe in 1716, Halley made the suggestion that it was produced by "magnetic particles" moving along geomagnetic field lines. In contrast, in the second edition of his book on aurorae (1754), de Mairan argued that the cause of the phenomenon lies on the Sun and that it is the result of the extended solar atmosphere, as indicated by the zodiacal light, interacting with the atmosphere of the Earth. De Mairan argued that, on this basis, aurorae should be

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

seen also in the southern hemisphere, as was ultimately confirmed by Cook in 1773. The auroral "zone" was drawn first by Loomis in 1860 and showed that the frequency of occurrence of aurorae is maximum in a belt which cuts across northern Scandinavia and decreases markedly in more polar regions. The fact that there is a similar auroral zone in the southern hemisphere and that both are located roughly symmetrically with respect to the geomagnetic axis was not established until much later. Eventually it was also found that auroral and associated geomagnetic disturbances tend to occur simultaneously in both hemispheres, in a manner consistent with the hypotheses of Halley and de Mairan.

The properties of the geomagnetic quiet-day solar variation (Sq), discovered in 1722 by Graham and studied in greater detail by Gauss and his successors, are such that Stewart concluded in 1882 that the associated electric currents must flow mainly in the upper atmosphere, which he inferred to be electrically conducting. The need for such a conducting layer at moderately high altitudes was demonstrated by the first long-distance radio-wave propagation experiments of Marconi, which showed that the waves were somehow refracted so that they could propagate around the curved surface of the Earth. The "ionosphere" (*q.v.*), as we now know it, was first detected directly in 1925 by the experiments of Appleton and Barnett in England and Breit and Tuve in the United States, using sounding techniques. Since that time, the ionosphere has been investigated in great detail and has been shown, in particular, to exhibit diurnal variations that are associated with the Sq geomagnetic variations and also irregular variations in the polar regions. The latter are associated with the occurrence of aurorae and polar geomagnetic disturbances.

Since the early 1950s and with the development of the space age, there has been an extraordinary growth in our understanding of all of the phenomena described above. It was shown by Storey, using low-frequency "whistler" waves generated by atmospheric lightning discharges, that the plasma environment of the Earth extends from its base in the ionosphere out to distances of at least several Earth radii and exists in a region which we now term the "plasmasphere." At still greater distances, experiments carried on spacecraft launched by the USSR and the United States around 1960 showed that interplanetary space is indeed filled with the rapidly expanding atmosphere of the sun. This had been proposed by Biermann on the basis of observations of the behavior of comet tails and termed the "solar wind" by Parker. De Mairan's hypothesis concerning the origin of the aurora therefore can be regarded as being correct, although the component of the zodiacal light to which he referred was certainly associated with interplanetary dust rather than the solar wind.

The region surrounding the Earth in which the geomagnetic field has a dominant role in controlling the dynamics of the plasma and the more energetic particles is now called the "magnetosphere," following a suggestion by Gold (1959). The magnetosphere was investigated in great detail by many early spacecrafts, notably in 1964 by the first Interplanetary Monitoring Platform (IMP 1), which provided a rather complete survey of the main features of the structure of the magnetosphere. Earlier, Van Allen (1987), with the aid of the first U.S. Earth satellites, had made the startling discovery that the magnetosphere contains zones of relatively energetic particles which were termed the "radiation belts." The so-called "inner" belt is rather stable and contains mainly protons with energies in the range 1–100 MeV. The "outer" belt contains highly variable fluxes of electrons and ions, mainly in the range 1–1000 keV, which are associated with the aurora and also with the electric currents which produce the various classes of geomagnetic storm effects. The latter include the "ring current," which causes the main-phase global depression of the field strength observed at the Earth during geomagnetic storms, and the auroral zone "electrojet" currents, associated with transient "substorms," which lead to the injection and energization of magnetospheric plasma and cause the growth of the ring current.

In the following, we provide a summary of our knowledge of the magnetosphere of the Earth and our understanding of the main processes involved in determining its behavior. Ionospheric phenomena are included where they are an essential component of the magnetosphere, and comparisons are made with the magnetospheres of other planets that have been investigated in some detail during the past 20 years.

1. THE CONFIGURATION OF THE EARTH'S MAGNETOSPHERE

The overall configuration of the Earth's magnetosphere is shown in Fig. 1. The first point is that the magnetosphere is bounded, as a consequence of the confining pressure of the solar wind. That is, the external pressure of the solar-wind plasma and magnetic field balances the internal geomagnetic field pressure at the interface (called the magnetopause). Thus, approximately,

$$nmV^2 = B^2/2\mu_0 \qquad (1)$$

where n is the number density of ions in the solar-wind plasma, m is their mean mass, V is the speed of the wind, and B is the magnetic induction just inside the magnetopause. This simple relation, which was first given by Chapman and Ferarro (1931), provides an estimate for R, the distance to the magnetopause at the subsolar point: taking $n = 5 \times 10^6$ m^{-3}, $V = 4 \times 10^5$ m/s, and $B = 2B_0(R_e/R)^3$, where $B_0 = 0.3$ gauss $= 3 \times 10^{-5}$ T is the equatorial magnetic induction at the surface of the Earth ($R = 1R_e \approx 6.4 \times 10^6$ m), we find that $R \approx 10R_e$. In fact, the various parameters involved vary considerably with time so that R may vary by as much as 50% from this value depending on conditions in the interplanetary medium.

It is possible to generalize the above result and thus produce an approximate but quite satisfactory model for the shape of the magnetopause by replacing V with $V \cos^2 \theta$, where θ is the angle between the mean solar-wind direction and the local normal to the magnetopause, and iteratively solving for the magnetic field so that it has no normal component. This is the so-called "Newtonian" approximation, which is reasonably good provided that θ is not close to $\pi/2$ and neglected components of the pressure inside and outside the magnetosphere are indeed negligible. The approximation works well because the solar wind at the orbit of the Earth is highly supersonic (the Mach number is usually of the order of 10 or more). Since the magnetosphere constitutes an obstacle in the solar-wind flow and waves with wavelengths comparable to the scale of the magnetosphere are ordinary Alfvén and magnetoacoustic waves, a bow shock with thickness

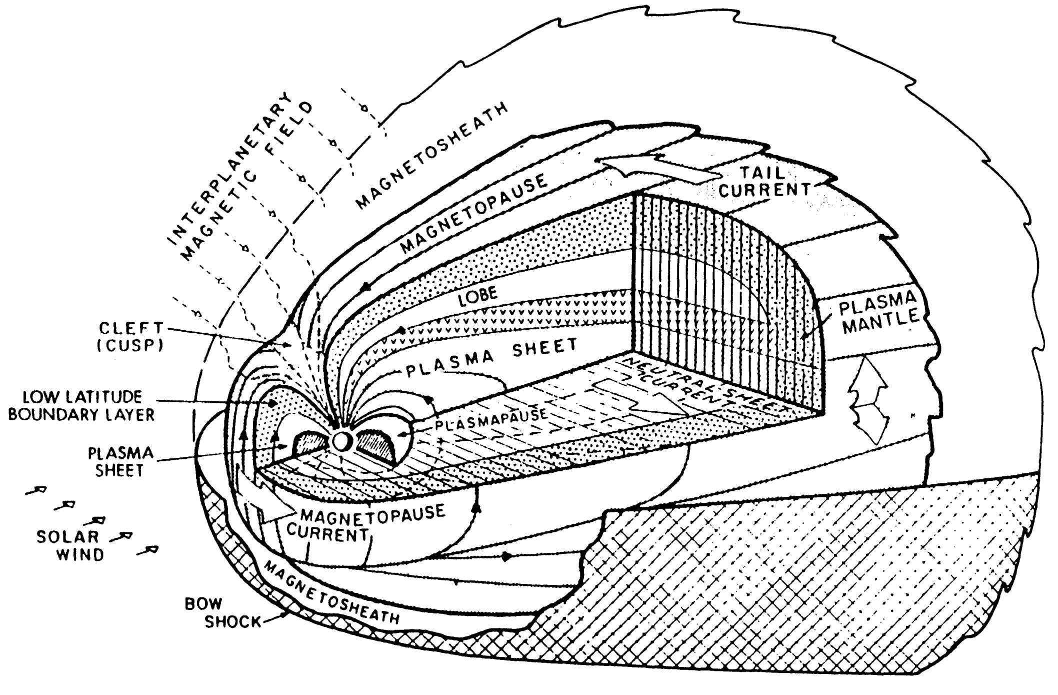
INTERPLANETARY MAGNETIC FIELD
MAGNETOSHEATH
MAGNETOPAUSE
MAGNETOPAUSE
TAIL CURRENT
LOBE
PLASMA MANTLE
PLASMA SHEET
NEUTRAL SHEET CURRENT
PLASMAPAUSE
CLEFT (CUSP)
LOW LATITUDE BOUNDARY LAYER
PLASMA SHEET
SOLAR WIND
MAGNETOPAUSE CURRENT
MAGNETOSHEATH
BOW SHOCK

FIG. 1. A three-dimensional representation of the magnetosphere of the Earth showing most of the important features described here (from Heikilla; see Meng *et al.*, 1991).

usually of the order of the ion gyroradius (100–1000 km) is formed upstream (Onsager and Thomsen, 1991). The shock plays only a secondary role in the interaction of the solar wind with the magnetosphere, but it is important in accelerating particles in much the same way that is presumed to occur in supernova remnants where the result is the production of cosmic rays.

1.1 The Polar Cusp

A feature of the idealized model of the magnetopause described above is the "polar cusp," which is connected to the day side of the Earth at geomagnetic latitudes between 65° and 75°, depending on interplanetary conditions (Hoffet and Egeland, 1984; Crooker and Burke, 1991). Ideally, the cusp is a single geomagnetic field line in each hemisphere which splits on encountering the magnetopause and covers the entire surface. This implies that a particle which becomes attached to the surface should eventually find its way to the ionosphere in one hemisphere or the other. In fact, this is an oversimplification resulting from the assumption that the geomagnetic field is completely confined within the magnetopause and the interplanetary magnetic field completely excluded. Some interconnection between these two magnetic fields at the magnetopause cannot be avoided as a consequence of the process of magnetic "reconnection" which would normally tend to occur over the forward side of the magnetopause. Thus, the polar cusp consists, in general, of a bundle of magnetic field lines in each hemisphere which intersect the Earth in small regions surrounding the idealized Chapman–Ferarro cusp positions. These are strongly confined in local time and geomagnetic latitude; the field lines connect with the interplanetary field so that there is always a small normal component of the field at the magnetopause allowing the escape or entry of plasma across the interface. This normal component plays an important role in the interaction of the solar wind with the magnetosphere, since it implies the existence of a tangential stress at the magnetopause that results in the exchange of energy and momentum as well as plasma.

1.2 The Magnetotail

A direct consequence of reconnection between the interplanetary and geomagnetic fields is the formation of an extended tail to the magnetosphere ("magnetotail") comprised of two bundles of oppositely directed field lines (Liu, 1989; Birn, 1991). These are connected to the polar caps of the Earth at latitudes greater, approximately, than the latitudes of the cusps. A large fraction of the magnetic flux in the magnetotail is linked with the interplanetary field rather than connecting the two polar cap regions of the Earth. This "openness" of the magnetic field at high latitudes permits terrestrial plasma to escape into interplanetary space and allows moderately energetic interplanetary particles (solar-flare particles and low-energy cosmic rays) direct access to the Earth's polar caps. The open field regions, called the "lobes," contain little plasma and to maintain pressure balance with the external interplanetary medium have field strengths of the order of 10 nT at the orbit of the Moon, for example. The total amount of magnetic flux contained in each tail lobe is about (5–10) × 10^{16} Mx, since the width of the magnetotail at this distance is about (30–40)R_e. This is roughly the same as the flux intersecting the polar caps of the Earth at latitudes greater than about 70–75°.

The region separating the two tail lobes is often called the "neutral sheet," since the main component of the magnetic field (tailward in the northern lobe and earthward in the southern lobe) reverses its sign. In fact, the neutral sheet is not strictly neutral but contains a small normal component that partly links the two lobes so that, in magnetically quiet times, the field lines involved are connected at both ends to the Earth. The magnetic field around the neutral sheet consists, therefore, of highly stretched dipole-like field lines containing about 10^{16} Mx. This field connects with the auroral zones of the Earth, which are at most about 5 degrees of latitude wide in each hemisphere, and contains plasma with a pressure approximately equal to the magnetic pressure in the lobes of the magnetotail. Assuming that the mean energy of the ions is roughly the same as in the solar wind, it is easily estimated that the density of these ions is about 10^4–10^5/m^3 and the electron flux (assumed >200 eV) about 10^{12}/

m^2 s, as observed. Obviously, the plasma sheet is intimately connected with the auroral phenomenon (Sec. 1.7), which requires that the electron component be accelerated further to 1–10 keV and be, to a large extent, precipitated into the upper atmosphere in the auroral zones where it produces visible radiation and considerable heating. No good explanation as to why a (small) component of the auroral particle precipitation should be highly localized, producing spectacular features such as arcs and rays, has yet been produced. Very probably, such features are a reflection of corresponding features in the outer magnetosphere and, in particular, the inner edge of the plasma sheet, which is often quite sharply defined.

1.3 Convection

The ionosphere forms the base of the magnetosphere and is a source of plasma for the outer regions as well as providing, at low altitudes (<130 km), an electrical short circuit perpendicular to the geomagnetic field which acts to discharge electric fields of external origin unless they are maintained in some way. At altitudes below about 200 km, the ionosphere is largely controlled by a balance between (solar) photoionization and local recombination (Van Zandt, 1967; Rishbeth and Garriott, 1969). At higher altitudes, the plasma may be transported over relatively large distances before recombining, provided the speed of transport is sufficiently great. This speed follows from

$$\mathbf{E} + \mathbf{V} \times \mathbf{B} = 0 \qquad (2)$$

where the electric field **E** is determined by magnetospheric processes or by tidal and other large-scale motions of the neutral atmosphere in the so-called "dynamo" region at 100–120 km altitude. These are usually called "convective" motions, and they pervade the entire magnetosphere since Eq. (2) holds good for low-energy plasma in the magnetosphere provided the time scales involved are long (Axford, 1969). A good example of convective effects on the ionosphere is the "equatorial anomaly"; this is the result of the electric fields associated with the Sq geomagnetic variations causing an upwelling of plasma near the geomagnetic equator during the day in association with a tendency of the plasma to drop under the influence of gravity. The overall effect is for the ionospheric plasma density to have maxima on either side of the equator in the afternoon and early evening (Cohen, 1967). A second example is the transport of plasma from the sunlit side of the Earth over the polar caps to the night side as a consequence of convection associated with auroral and polar cap magnetic disturbances.

1.4 The Polar Wind

In addition to convection perpendicular to the local magnetic field direction, as implied by Eq. (2), ionospheric plasma can move freely along the field lines, being influenced only by gravity. Thus, on field lines that are closed, the ionosphere may extend far out into space, being filled from either end by upwelling plasma composed mainly of the lightest species available, namely, protons, until hydrostatic equilibrium is reached. Such an equilibrium prevails in the plasmasphere discovered originally by Storey. Observations, however, show that the plasmasphere usually has an outer boundary, the "plasmapause" at about $4R_e$, beyond which the plasma density suddenly drops to levels of the order of $10^6/\text{m}^3$. Beyond the plasmasphere, on field lines that are empty of plasma or are directly open to interplanetary space (i.e., the tail lobes), the ionosphere continues to supply plasma from below, but there is no equilibrium on a given field line until it is finally filled. The lack of hydrostatic equilibrium on closed field lines is a consequence of the convection of plasma to the outer parts of the magnetosphere where it may eventually find itself on open field lines and so escape into space. The existence of the plasmapause, which was discovered by direct spacecraft measurements as well as by the whistler technique around 1960, can be taken as an indication that such convective and loss processes are indeed important in the magnetosphere.

The upwelling of ionospheric plasma described above is termed the "polar wind" (Banks, 1978). It is a supersonic flow, somewhat analogous in its characteristics to the solar wind, and should be prevalent above most of the polar ionosphere at geomagnetic latitudes greater than about 60°. The polar wind is responsible for controlling the amount of helium in the Earth's atmosphere, since it provides an efficient escape mechanism capable of balancing the rate at which fresh helium is emitted from the Earth's crust as a result of radioactive processes. As described so far, the polar wind is a thermal process associated with photoionization and heating of the plasma by photoelectrons. Stronger heating processes are available, however, notably those resulting from cyclotron heating which favor heavier species—in particular, helium, oxygen, and even molecular ions. These play a significant role in forming the plasma sheet and the outer radiation belt following modest acceleration to the ~10-keV range (Axford, 1970; Cornwall and Schulz, 1979).

1.5 The Exosphere

While helium escapes from the Earth via the polar wind, hydrogen escapes in neutral atomic form as a result of Jeans (thermal) escape; that is, the atoms are placed on escape trajectories by collisional processes at the base of the exosphere where the mean free path for collisions begins to exceed the scale height of the distribution. The exosphere is quite extensive, depending on the phase of the solar cycle, which determines the temperature at its base, and neutral hydrogen may have a significant density out as far as $10R_e$. This cloud of neutral hydrogen surrounding the Earth's exosphere can be detected from scattered solar Lyman-alpha radiation and is usually dense enough to be optically thick in this line in its inner regions. The exosphere is important, as charge exchange of energetic protons with hydrogen atoms is a significant loss mechanism for the outer radiation belt and may determine the rate of decay of the associated ring current (Williams, 1983). Charge transfer with other species also leads to changes in their charge state distribution which can now be measured.

1.6 The Radiation Belts

The outer radiation belt has its origin mainly in transient events in which the plasma sheet is pushed close to the Earth and the energetic plasma becomes trapped. The same process, which can be regarded as temporarily enhanced convection, erodes the plasmasphere and hence, at the time it occurs, places the inner edge of the newly introduced hot plasma at the plasmapause. Since the pressure of the plasma is not insignificant, the magnetic field is inflated, and the result is observable within the plasmasphere and, in particular, on the Earth, as a ring current. Initially, this ring current may be asymmetric, but after a while, it settles down and becomes reasonably symmetrical when the magnetic field depression can be calculated from

$$\mathrm{grad}P = \mathbf{j} \times \mathbf{B} \tag{3}$$

with $\mathbf{j} = (1/\mu_0)$ curl $\mathbf{B}$ and P a given (symmetrical) distribution of plasma pressure. It has been shown by Dessler, Parker, and Sckopke (see Williams, 1983) that this leads to a simple relation for the field depression ΔB at the Earth:

$$\Delta B/B_e = -2E/3E_m \tag{4}$$

where $B_e = 30{,}000$ nT is the horizontal component of the geomagnetic field at the equator, E is the total energy of the trapped particles, and $E_m = 8 \times 10^{17}$ J is the total energy in the geomagnetic field outside the Earth. In a large geomagnetic storm, ΔB may be of the order of 300 nT so that the total energy of the plasma producing the ring current effect is about 10^{16} J. Typical particle energies are 1 keV–1 MeV. The inner radiation belt is quite different in its characteristics from the outer belt. The particles are much more energetic (100 keV–10^3 MeV), the distribution of particles does not vary greatly with time, and the ions seem to be mainly protons. The main source of energetic particles appears to be the decay of neutrons produced by cosmic rays interacting with the Earth's upper atmosphere, and this is consistent with the observed predominance of protons. Nevertheless, there may be an external source, possibly initiated by energetic solar particles, such that the particles diffuse slowly inward from an injection region around $4R$, the diffusion resulting from an interaction of drift motions in the inhomogeous geomagnetic field and fluctuations in the convection fields associated with magnetospheric substorms.

1.7 Auroral Phenomena

The most spectacular aspect of magnetospheric physics is the polar aurora, which, in magnetically disturbed periods, is a highly structured and often rapidly changing fluorescence of the upper atmosphere resulting from excitation by electrons with energies in the range 1–20 keV. The most important visible emissions are those of atomic oxygen (green λ 5577, red λ 6300) and nitrogen molecules (blue λ 3914, λ 4278). In addition, there is very strong infrared emission from molecular oxygen (λ 12 700), which is heavily absorbed in the atmosphere. Protons, with energies in the range 1–100 keV, also produce weak Balmer lines, which can be detected from the ground (e.g., red λ 6563) with Doppler shifts indicating downward-moving particles, as well as Lyman lines which are absorbed or scattered into space. Most of these emissions are produced at altitudes of 100 km and higher. High-energy electrons (>50 keV) produce x rays that can be detected at balloon altitudes (20–40 km) and that lead to low-altitude ionization detectable by radiowave diagnostic techniques. (For a detailed discussion of auroral emissions, see Omholt, 1971; Bone, 1991; Meng *et al.*, 1991; and Meng and Mauk, 1991).

Most aurora is produced in zones 5–10° wide surrounding both polar caps at magnetic latitudes of about 65–70°, corresponding to the "feet" of magnetic field lines which connect to the plasma sheet in the magnetosphere. There is, however, another class of aurora which occurs at lower latitudes, namely, the midlatitude or "subauroral red arcs." These are observed during large geomagnetic storms and are spatially associated with the plasmapause. The energy input to the normal aurora ranges from, typically, 10^{-3} J/m^2 s, for subvisual quiet-time conditions, to 1 J/m^2 s for bright, active aurora seen during disturbed periods. The midlatitude arcs, which appear to be produced by low-energy (10–100 eV) electrons, are faint and involve perhaps only 10^{-3}–10^{-2} J/m^2 s. The total power associated with the production of aurora is of the order of 10^8–10^{11} J/m^2 s depending on the level of magnetic activity. In addition, the electric currents that are associated with the aurora (the electrojet) and that produce the geomagnetic variations may involve a total of as much as 10^6 A. The resulting ionospheric Joule heating power is of the same order as that of direct particle heating. The heating results in considerable atmospheric upwelling and, together with the $\mathbf{j} \times \mathbf{B}$ forces, launches atmospheric gravity waves, the effects of which can be observed at low latitudes.

The most striking features of the visible aurora are the structures that are commonly observed, namely, arcs, rays, coronae, and so on. The arcs are often found to be many thousands of kilometers in length but are only a few kilometers wide. As magnetic activity increases, the arcs tend to fold and assume a rayed structure. The folded arcs may seem to break up into isolated rays which, on passing overhead, appear as "coronae." These features may be very dynamic, moving with speeds of the order of several kilometers per second, changing in form and often pulsing in brightness with periods of the order of 10–20 s. The pattern of auroral motions corresponds roughly to that of magnetospheric convection at ionospheric levels as confirmed by radar Doppler backscatter measurements and direct measurements of drifts and electric fields in situ. There are, however, motions in auroral patterns which do not correspond straightforwardly to Eq. (2) but are instead associated with the motion of disturbances moving across geomagnetic field lines (i.e., waves and reconnection as described in Sec. 2.1 and 2.2). It should be noted that, despite their prominence as far as visible observations are concerned, auroral structures are not dominant in terms of the total energy deposited in the atmosphere. The human eye, for example, is most sensitive to the bright features but fails to register the more ubiquitous diffuse emissions that fill much of the sky in the auroral zone during periods of high activity.

There is a certain amount of auroral activity associated with the polar cusp and occurring also in the polar cap. This aurora is evidently produced on open geomagnetic field lines to a large extent and has different characteristics from that occurring in the auroral zone. In particular, it becomes bright and active in the early phases of a geomagnetic storm, suggesting that it is activated by the enhanced dayside magnetic reconnection taking place at such times (see Sec. 3.1).

Major disturbances in the aurora, the associated electrojet, and other related phenomena were formerly called "polar elementary storms" or "magnetic bays" and are now known as substorms (see Sec. 3.2). The auroral disturbance, sometimes called "breakup," has a particular pattern, which was first clearly described by Akasofu (1977) and which is shown in Fig. 2(a). It is important to note that the poleward motion of the bright arc depicted here does not correspond to convection, which, in fact, tends to have the opposite sense. A typical example of the magnetic disturbance occurring at such times is shown in Fig. 2(b).

The cause of most auroral emissions is the precipitation from the plasma sheet of moderately energetic electrons by wave scattering or anything that leads them to interact with the atmosphere. In addition, there are strong indications for acceleration of particles in the few-keV range associated with anomalous electrical resistivity affecting magnetospheric currents. These currents flow out of the ionosphere into the magnetosphere as a consequence of the stresses induced in the system by plasma pressure, interaction with the solar wind, and ionospheric convection drag. The anomalous resistivity results in the formation of plasma turbulence involving small-amplitude electrostatic waves, which provide an integrated electric potential drop of a few kilovolts and a corresponding "monoenergetic" feature in the spectra of the incoming electrons.

In general, the aurora exhibits geomagnetic conjugacy to a surprising degree of detail. There is, however, a tendency for the patterns observed in the two hemispheres to be displaced in longitude by many hundreds or even thousands of kilometers, depending on the level of activity. It is suggested that this is the direct result of magnetic reconnection in the magnetotail, since reconnection need not involve conjugate magnetic field lines, as defined from the internal field of the

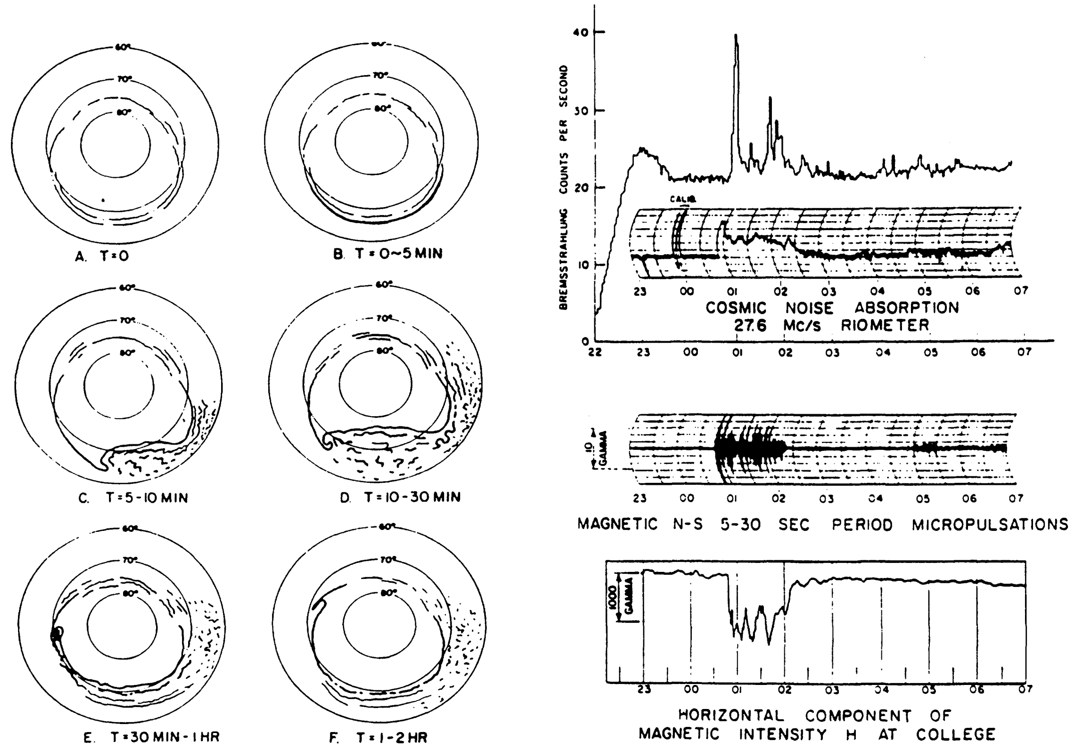

(a) (b)

FIG. 2. (a) The pattern of development of auroral arcs and activity associated with a magnetospheric substorm as seen from above the north geomagnetic pole. The rapid northward movement (C, D) does not correspond to the direction of magnetospheric convection at the time. (From Akasofu, 1977.) (b) The simultaneous occurrence of x-ray bursts, cosmic radio noise absorption, micropulsations in the geomagnetic field, and the negative magnetic disturbance associated with the auroral electrojet during a magnetospheric substorm observed in Alaska (From Akasofu, 1977.)

Earth. The component of the interplanetary field perpendicular to the geomagnetic axis skews the two lobes of the tail, thus leading to a systematic longitudinal shift of conjugacy. Small-scale nonconjugate reconnection can be relieved by the propagation of Alfvén waves, which remove the tangled component of the field as they are dissipated at low altitudes by ionospheric resistivity. Accordingly, there is a great deal of micropulsation activity associated with the aurora and detectable from ground-based observatories, directly and via fluctuations in luminosity; it may also be detected in space where the amplitudes of the waves may be very large.

2. THE DYNAMICS OF THE EARTH'S MAGNETOSPHERE

2.1 Magnetic Reconnection

The most important single process occurring in the magnetosphere is magnetic field reconnection, which is depicted schematically in Fig. 3. In general, Eq. (2) provides an accurate description of the relationship between the bulk plasma velocity and the electromagnetic field. When combined with Faraday's law, it yields

$$\frac{\partial B}{\partial t} + \operatorname{curl}(\mathbf{V} \times \mathbf{B}) = 0, \tag{5}$$

which can be shown to require that plasma elements on a common magnetic field line at one instant in time remain on a common field line thereafter and that a closed loop which moves with the plasma always encloses the same amount of magnetic flux (Sonnerup, 1979; Axford, 1984). This "frozen field" theorem is valid provided additional terms neglected in Eq. (2), such as might be associated with the finite electrical conductivity of the plasma, do not cause a breakdown of any significance. Such effects are usually negligible except in the vicinity of "neutral" points in the magnetic field where $\mathbf{V} \times \mathbf{B}$ vanishes locally and $\mathbf{E} \cdot \mathbf{B}$ is not zero. If a neutral point exists, the property of "connection" between plasma elements implied by the frozen-field theorem need not be preserved. As a consequence, magnetically connected plasma elements may become disconnected if their common magnetic field line intersects a neutral point somewhere between them, as illustrated in Fig. 3.

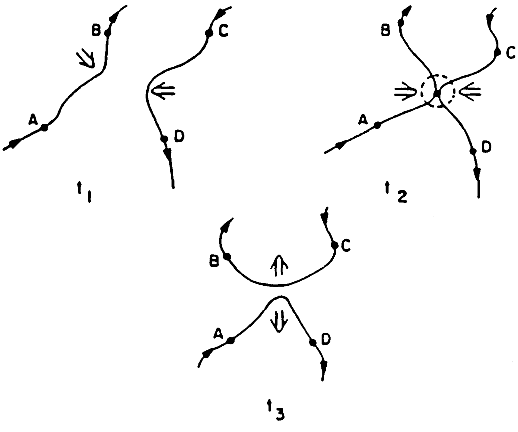

FIG. 3. At time t_1, the magnetic field line defined by plasma elements A and B approaches an oppositely directed field line defined by elements C and D. At time t_2, the two field lines touch at some point intermediate between A and B and C and D. At time t_3, the connection between A and B and between C and D is broken and the field line defined by A and D moves away from the field line defined by B and C. (Axford, 1984).

The manner in which reconnection permits linkage between the interplanetary and magnetospheric fields is shown schematically in Fig. 4(a). The numbers indicate the successive positions of a geomagnetic field line following reconnection with interplanetary field (1) and subsequently reconnection between field lines in the geomagnetic tail (6). Between positions (1) and (6), the feet of the field lines move across the polar ionosphere from the dayside to the nightside and the field lines are open ended, allowing magnetospheric plasma to escape into interplanetary space. Between positions (6) and (9), the field lines and the plasma move toward the Earth from the night side (their feet move toward the equator); between positions (9) and (10), the motion is around the Earth (out of the plane of the paper); and from (10) back to (1), the feet move poleward. The pattern of flow seen at ionospheric levels is also shown in Fig. 4(b). This pattern is consistent with what is observed on average by radar backscatter techniques and satellite-borne electric field probes and is the dominant component of magnetospheric convection in the outer magnetosphere.

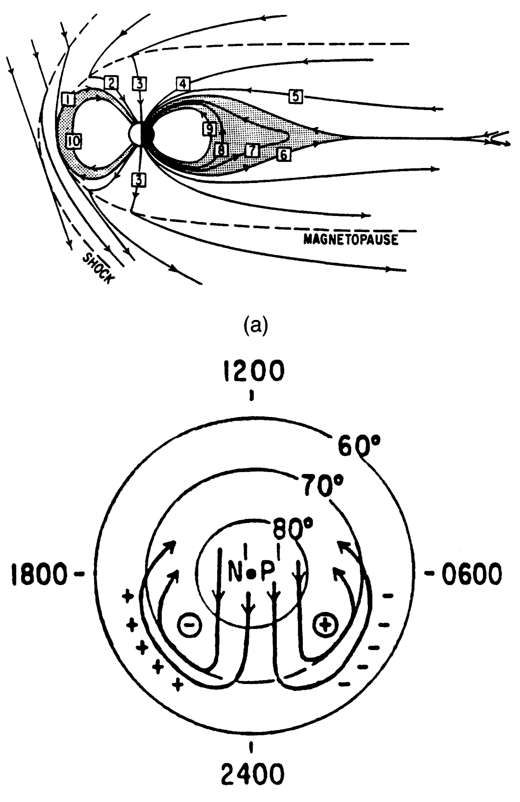

(a)

(b)

FIG. 4. (a) Steady convection as seen in the noon-midnight meridian with reconnection occurring at positions (1) and (6). Between points (9) and (10), the convection is out of the plane of the paper, going around the Earth at latitudes corresponding to the auroral zones. (b) Motion of the feet of the geomagnetic field lines at ionospheric levels, shown schematically. If the Hall conductivity is dominant, the streamlines also represent current flow lines if the direction of the arrows is reversed.

2.2 Magnetospheric Substorms

The above description of reconnection is purely kinematic. In fact, there are limits on the "rate" of reconnection, that is, on the electric field strength along the neutral points shown in Figs. 3 and 4. Depending on external conditions, this lies between 0 and $V_A B$, where V_A and B are the Alfvén speed and magnetic field strength characterizing the plasma outside the reconnection region, respectively. In effect, the rate is determined by the rate at which mass can be ejected from the reconnection region to make room for further reconnection to take place, the ejection resulting from the "catapult" effect of the magnetic stresses in the strongly curved field lines in the narrower parts of the "X." Since these stresses can be counteracted by pressure gradients in the plasma, the rate of reconnection can be controlled by the degree to which such pressure gradients build up and, indeed, can be brought to a halt if the plasma on reconnected field lines is unable to escape. On examining Fig. 4, it can be seen that, on the upstream side of the magnetosphere where the solar-wind plasma flow is determined to a first approximation by externally induced pressure gradients, reconnection may proceed on a more or less continuous basis, whereas in the magnetotail, where the inner magnetosphere may block the flow, reconnection may be stopped almost entirely until the blockage is somehow overcome. The plasma that produces the blockage is the plasma sheet, and the nonlinear instability that results in its disruption and a reconfiguration of the magnetotail corresponds to the magnetospheric substorm phenomenon (Fig. 2). This kind of instability, which has been termed a "gross resistive" or "reconfigurational" instability, also plays a role in solar flares.

The precise details of substorm initiation are not completely understood at present. However, it is clear that the whole process involves growth of the magnetotail, in both size and magnetic field strength, followed by a collapse associated with reconnection beginning in the existing plasma sheet, the injection of energetic particles into the ring current region in the inner magnetosphere, and the cutting off of a large piece of the magnetotail, which is carried away downstream where it eventually merges with the interplanetary medium (Fig. 5). As a consequence of a series of such substorms, the ring current grows in strength over a period of several hours and ultimately decays again as the substorm activity declines and the losses of ring current plasma to charge exchange with the exospheric neutral hydrogen and precipitation into the atmosphere are able to restore order (see Akasofu, 1977; Kan *et al.*, 1991; Birn, 1991).

Dayside reconnection at the magnetopause is not constrained in the same way as it is in the magnetotail, since the reconnected plasma and field can easily escape (Smith, 1991). Accordingly, periods in which the interplanetary magnetic field is strong and

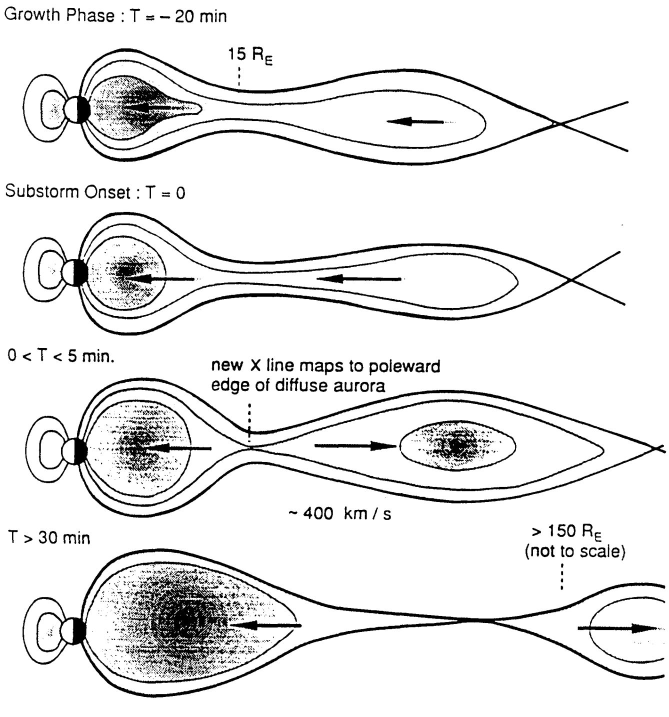

FIG. 5. Schematic diagram of the changes taking place in the plasma sheet and magnetotail during the various phases of a magnetospheric substorm. (From Kan *et al.*, 1987.) A second reconnection (X) line is formed close to the Earth, and a plasmoid is detached and moves down the magnetotail away from the Sun.

directed southward are also periods in which rapid erosion of magnetic flux from the dayside of the magnetosphere can be expected, perhaps episodically ("flux transfer events"). As a consequence of rapid flux transfer, the magnetotail grows and substorms and a geomagnetic storm must follow. The events are also associated with strong poleward convection in the high-latitude dayside ionosphere and the presence of bright and active aurora in this region.

2.3 Energy Budgets

A summary of the magnetospheric energy budget at varying levels of geomagnetic activity is given in Table 1. The average efficiency for energy transfer between the solar wind and the magnetosphere is only about 10^{-3}, but this does not include the stored magnetotail energy lost downstream. This may be much greater than that trapped in the inner regions of the magnetosphere and dissipated in the form of auroral-zone particle precipitation and Joule dissipation of ionospheric electric currents. It is interesting to note that the efficiency of capture of solar-wind plasma is quite low and may be less than the 10^{-4} quoted, since the plasma sheet and the lower-energy ring-current particles contain a significant component of ionospheric origin, notably in the form of singly ionized helium and oxygen but also, occasionally, as molecular ions such as NO^+ (Moore, 1991). Nevertheless, it seems possible that capture from the solar wind may be an important source of the atmospheric ^3He isotope, for example.

The rate of energy transfer to the magnetosphere from the solar wind, as a consequence of reconnection at the magnetopause, is of the order of $VB_tB_nA/2\mu_0$, where V is the solar-wind speed, B_t and B_n are the components of the magnetic field tangential and normal to the magnetopause, respectively, and A is the effective area of the magnetopause. Since $B_nA \approx \phi_p$, the total "open" magnetic

Table 1. Magnetospheric energy budget.

Incident solar wind flux	3×10^{12}–10^{15} J
Ring current particles	3×10^{14}–10^{16} J
Ring current injection rate	3×10^{9}–10^{12} J/s
Ring current loss rate	3×10^{9}–10^{11} J/s
Auroral particle precipitation	3×10^{9}–10^{11} J/s
Auroral Joule dissipation	3×10^{9}–10^{11} J/s
Total injection rate	10^{10}–10^{12} J/s
Efficiency	$\sim 10^{-3}$

flux intersecting the polar caps of the Earth, then, taking $V = 10^6$ m/s and $B_t = 10$ nT and assuming that the geomagnetic field above 70° latitude is open, we find that the rate of energy transfer is about 10^{13} J/s. This easily accounts for the total injection rates given in Table 1, and it is presumed that any excess is lost again to interplanetary space in the form of detached magnetized "plasmoids," following reconnection in the magnetotail.

2.4 Particle Acceleration

The violent release of electromagnetic energy associated with magnetic field reconfiguration during substorms leads to the acceleration of particles in the magnetosphere as a result of at least six distinct processes (Chang, 1986).

1. The most important acceleration process in the magnetosphere in terms of total energy conversion is associated with the convection. The process is effectively adiabatic and is a consequence of the compression and expansion of magnetic flux tubes as they move deep into and out of the magnetosphere. It can also be thought of as the result of particle drift across the equipotentials of the convection electric field, and this immediately implies that the maximum energy achievable is of the order of 100 keV/charge under conditions of strong convection. All of the plasma participates, in particular, the ions that produce the ring-current effect and the electrons that give rise to the bulk of the auroral precipitation.
2. Electric fields associated with nonsteady convection, which includes nonconservative fields, can affect particles of higher energies, especially in giving rise to radial diffusion with an associated acceleration if the diffusion is inward. The diffusion may be relatively slow, since it depends very much on the component of time-varying electric fields resonant with the particle drift period around the Earth. However, very high energies are achievable provided that losses are not important; consequently, this process may constitute a major source for the inner radiation belt.
3. The reconnection process itself can give rise to a localized but very striking acceleration, especially of ions that have large gyroradii in the weak fields near the neutral point. These are able to experience much of the available inductive voltage drop before exiting along the magnetic field lines intersecting the neutral point. When reconnection occurs on tail-lobe field lines, the voltage drop may be as large as $V_A BL$ for a short period. With $V_A = 3 \times 10^6$ m/s, $B = 20$ nT, and $L = 3 \times 10^7$ m, the voltage drop can exceed 10^6 volts, which is much greater than that associated with ordinary convection.
4. Large-amplitude waves, with their wave normals almost perpendicular to the magnetic field, are capable of producing substantial electric fields parallel to the magnetic field and can, therefore, accelerate electrons very efficiently. This may occur very close to the neutral point, where $\mathbf{E} \cdot \mathbf{B} \neq 0$ in general, and also when strong compressional waves are produced in the inner magnetosphere, as may occur when the inner magnetotail collapses toward the Earth in association with the onset of a substorm.
5. Resonant wave damping can be a very efficient heating process, but, of course, it is limited by the amount of wave energy available at suitable frequencies. A very important example is cyclotron heating of ionospheric ions in the auroral zones, which produces He^+, O^+, and even molecular ions with energies of the order of several keV. These ions escape into the outer magnetosphere where they provide a source for the ring current and are subject to further acceleration by other processes.
6. The stresses exerted by the plasma and magnetic field in the outer magnetosphere are, to some extent, taken up by the ionosphere where the plasma is held as a result of collisions with the neutral atmosphere. The communication of such stresses involves a strongly localized "shear" in the magnetic field, for example, at the boundary between open and closed field lines. This magnetic shear is associated with a field-aligned electric current which, if it is sufficiently strong, can result in plasma instabilities that produce fluctuating parallel electric fields and cyclotron waves as described above. These electric fields may be important in accelerating electrons to moderate energies. They are the result of "anomalous" elec-

trical conductivity and are not to be confused with the hypothesized "double layers," which do not appear to exist with large, stationary potential drops.

3. OTHER MAGNETOSPHERES

3.1 Magnetospheres of Nonmagnetized Bodies

All solar system bodies that have no intrinsic magnetic fields and no atmospheres have a minimum interaction with the solar wind or whatever plasma environment that may exist. The surface layers of such objects, of which the Moon is a typical example, have such a low electrical conductivity that they may be considered as being insulators through which magnetic field lines can pass unhindered. The only interactions are associated with absorption of the solar wind or other plasma, leaving a void in the wake that slowly fills in from the sides and a weak effect associated with the emission of secondary ions and neutral atoms as a result of sputtering and other processes. The objects that behave in this manner are the Earth's Moon, Phobos, Deimos, asteroids, comets at large distances from the Sun, the icy satellites of the outer planets (interacting mostly with magnetospheric plasma), and possibly the Pluto–Charon system if no atmospheres are involved.

Objects that have atmospheres but negligible intrinsic magnetic fields interact strongly with external flowing plasma and can "capture" the magnetic field of the external medium producing a kind of "induced" magnetosphere. The important components of the interaction are the ionospheres produced by collisional or photoionization and the mass loading of the more distant external plasma as a consequence of the pickup of ions from the exosphere of the body. The case of Venus (Hunten *et al.*, 1983; Luhmann, 1991) is rather special, because the ionosphere normally has enough pressure to withstand the ram pressure of the solar wind. This flows around the effective obstacle, being affected mainly by a rather strong contribution from pickup ions and the fact that the size of the obstacle is not large compared with the ion gyroradii. The interface between the topside ionosphere and the external plasma (the "ionopause") is often clearly defined, but it is liable to instabilities associated with the unfavorable curvature of the external magnetic field, especially when pushed inward as a result of a sudden increase of the external pressure; in such circumstances, magnetic field lines can be trapped in the ionosphere, forming "flux ropes" which may be relatively long lived. Mars is rather similar to Venus, but it appears that the ionosphere is not usually capable of withstanding the external pressure and hence the magnetic field lines become fully imbedded and constitute a kind of trapped or induced magnetosphere (Blamont, 1991). It is also uncertain as to whether or not Mars has a very weak (<50 nT) intrinsic field that complicates matters further. Both Venus and Mars lose material to interplanetary space as a consequence of their interactions with the solar wind, and in the case of Mars this may have played a significant role in the evolution of the atmosphere.

Comets have certain similarities to Mars and Venus but differ in that the parent bodies are negligibly small and, especially in the case of large comets such as Halley, have enormously extended atmospheres when close to the sun. As a consequence of plasma pickup, these atmospheres constitute correspondingly large effective obstacles to the solar-wind flow. At 1 a.u. from the Sun, the shock wave produced by Comet Halley was about 10^9 m from the nucleus and within 10^8 m the plasma flow was dominated by the presence of cometary ions with the magnetic field accordingly draped around the nuclear region much as it is around Venus but on a much smaller scale. The most interesting features of the interaction are the strong turbulence induced by the unstable distribution of fresh pickup ions, which in turn leads to the acceleration of some particles to several hundred keV and the formation of an "ionopause" close to the nucleus where the magnetic field is held off by the friction between the dense outflowing neutral gas and plasma contained in the field. (Ip and Axford, 1990; Johnstone, 1991; Gombosi, 1991).

The situation with regard to satellites with atmospheres, such as Triton, Io, and Titan, is rather similar to that of Venus and Mars except that the external magnetic field is much stronger and the flow need not be supersonic in every sense. Plasma lost from these bodies has an important role in supplying the magnetospheres of the host planet with heavy ions, which form a plasma torus at the orbit of the satellite concerned. Io is particularly interesting in that its atmosphere is very thin and almost entirely controlled by the vapor pressure of sulfur dioxide ice that has been deposited on the surface by volcanic activity.

3.2 The Magnetosphere of Mercury

It was something of a surprise to discover in 1974 that Mercury has a magnetic field (with a dipole moment of about 5×10^{12} T m^3, surface field ~ 400 nT), in view of the small size and slow rotation of the planet (Vilas *et al.*, 1990). The field is sufficiently strong to cause the solar wind to stand off from the surface of the planet by a few tenths of a planetary radius in most circumstances. As a consequence of its smallness and lack of a significant atmosphere, the magnetosphere is very active. The plasma in the magnetosphere must be mainly comprised of photoelectrons, sputtered ions from the surface, and such solar-wind plasma as can enter; sodium is very prominent in the spectrum of the atmosphere and is presumably sputtered from the surface. With magnetic field strengths of the order of 30–100 nT, the Alfvén speed in the magnetotail is accordingly very high, and, even with the small scales involved, it is to be expected that, for brief periods during unstable reconnection events, the voltage drops available are 10^5–10^6 V. Bursts of particles with the corresponding energies have been observed. The small size also implies that there should be rapid tail growth so that substorms recur with a time scale of a few minutes (Siscoe and Slavin, 1979).

3.3 The Magnetosphere of Jupiter

Jupiter is remarkable as far as its magnetosphere is concerned in that the planet has a large magnetic moment, large surface fields (~3.2–14.3 G, dipole field 4.3 G at the equator), rapid rotation (period 9.9 hours), a massive atmosphere and substantial ionosphere, and, most important, a satellite in the inner magnetosphere (Io) that has a partial atmosphere comprised mainly of SO_2 (Gehrels, 1976; Dessler, 1983; McNutt, 1991; Bagenall, 1992). Plasma "picked up" from the ionosphere and exosphere of Io together with the rapid rotation of the planet combine to produce a magnetosphere that is almost completely dominated by internal processes rather than by the solar wind. The net result is that a dense plasma torus is produced in the vicinity of Io's orbit, comprised mostly of ions formed from SO_2 and its daughter products and lost ultimately by being "centrifuged" into the outer magnetosphere and eventually into space (Fig. 6). This involves the magnetic field being stretched so that the plasma is contained in a disk roughly perpendicular to the axis of rotation, the field lines being almost parallel to the disk on either side and the field at higher latitudes having the character of a monopole rather than a dipole. The plasma is eventually lost to interplanetary space by magnetic reconnection and the formation of plasmoids that escape along the magnetotail away from the Sun.

This, of course, is a gross oversimplification of the situation, but it contains the most essential aspects as far as the energy and particle budgets are concerned. The solar wind plays a role mainly in shaping the more distant parts of the magnetosphere and constraining the magnetotail to the downwind direction. There is, nevertheless, evidence for sources of energetic particles other than the atmosphere of Io, namely, the ionosphere of the planet itself, since energetic H_2^+ and H_3^+ ions are observed, and also the interplanetary medium (either the solar wind or solar-flare particles), since the outer magnetosphere contains energetic ions with typical solar abundances and charge-state distributions. The inner magnetosphere is characterized by a very energetic population, presumably produced by inward diffusion resulting from time-varying convective motions, which, as far as the electron component is concerned, are sufficiently energetic to produce synchrotron radiation in the GHz range. Particles escaping from the magnetotail include intense fluxes of electrons with energies in the MeV range, which are a major contributor to the interplanetary population even near the Earth. These particles maintain a 10-hour periodicity at large distances from the planet, suggesting that their escape is controlled by rotation in some manner.

The environment of the satellite Io is especially interesting, because the plasma it supplies to the magnetosphere in the form of pickup ions has a high temperature. It is dense enough to permit reasonably rapid energy

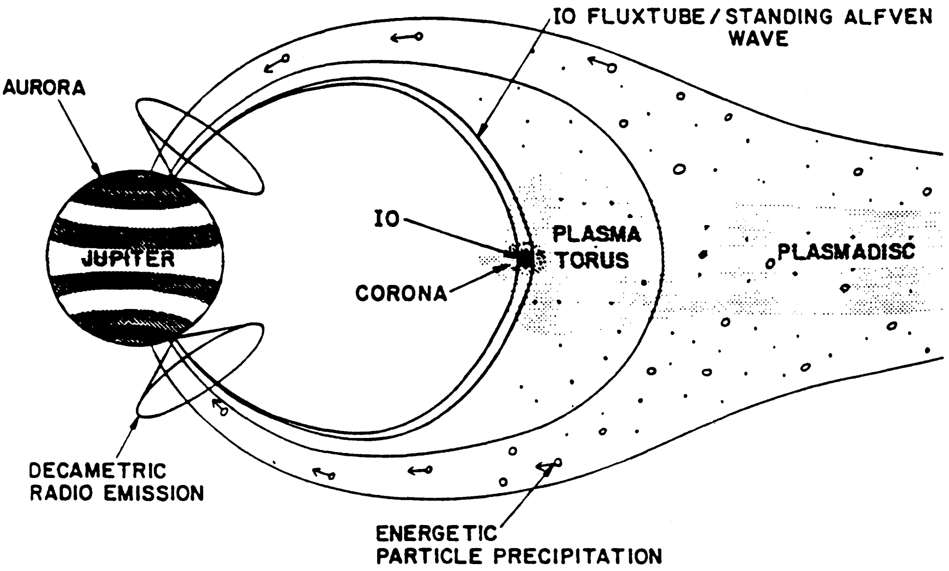
IO FLUXTUBE/STANDING ALFVEN WAVE
AURORA
JUPITER
IO
CORONA
PLASMA TORUS
PLASMADISC
DECAMETRIC RADIO EMISSION
ENERGETIC PARTICLE PRECIPITATION

FIG. 6. The connections between Io, the surrounding magnetosphere, and Jupiter's atmosphere. (From Bagenall, 1992.)

transfer to the electrons, which, in turn, lose energy by collisional excitation of S^+ and other ions. It is noteworthy that the pickup energy is about 17 eV/amu and that some 10^{28}–10^{29} ions/s must be produced in order to account for the observed emission from the plasma torus. The source of the material is obviously dominated by the thin SO_2 atmosphere of the satellite, in which ions are produced by photoionization and electron impact, and also by sputtering from the surface, which must be partially unprotected from direct impact. The surface is partly covered by SO_2 frost, but there must be sufficient rocky material to account for the relatively large amounts of potassium and sodium neutrals that are relatively easily observed in the neutral torus as a consequence of their strong resonant scattering of sunlight. It is evident that much of the ion pickup takes place near Io, because the Alfvén wave produced by the interaction, involving a current of some 10^6 A, causes strong decametric radio emission on interacting with the ionosphere of Jupiter. Strong polar auroral emissions are observed, partly associated with the plasma torus of Io but also at high latitudes, presumably associated with outer magnetospheric phenomena.

3.4 The Magnetosphere of Saturn

The magnetosphere of Saturn is much more like that of the Earth than that of Jupiter in that there is no strong source of plasma in the inner magnetosphere comparable to Io (Gehrels and Matthews, 1986; McNutt, 1991; Bagenall, 1992). The atmosphere of Titan, including its neutral torus, is potentially a strong source, but this satellite spends much of its orbit outside the magnetosphere ($20.3R_S$), or at least on its outer fringes, so that the plasma it contributes, which is mainly N_2^+ or an ion with similar mass, is largely lost into the plasma sheet in the tail. The planetary magnetic field has a dipole component that is almost exactly aligned with the rotational axis and has an equatorial field strength of 0.22 G; the minimum distance to the magnetopause is typically $(15–20)R_S$. The dipole may be displaced from the center of the planet, as the surface field strength appears to vary from 0.18 to 0.84 G. One might expect substorms to be as prevalent as they are in the Earth's magnetosphere, but the plasma contributed by Titan and its torus as well as the rapid rotation of the planet may shield the inner magnetosphere from their effects to some extent.

The inner magnetosphere of Saturn is apparently dominated by plasma sputtered from the water-ice mantles of the larger satellites, Dione and Tethys. As a result, the plasma contains H_2O^+ ions and their dissociation products, some directly sputtered and others produced from the disk-shaped cloud of neutral molecules surrounding these satellites in the equatorial plane of the planet. Interestingly, energetic molecular hydrogen ions have also been observed in Saturn's magnetosphere, suggesting, perhaps, that an ionospheric source might play a role, especially for protons.

One of the most interesting aspects of the energetic particle population in the magnetosphere of Saturn is that it is sufficiently stable for clear absorption signatures coincident with several of the satellites to have been found. These provide, in principle, an indication of the rate of radial transport of the particles concerned. Similarly, the rings provide absorption signatures in the inner magnetosphere, which suggest that the particle population is very stable indeed in this region.

3.5 The Magnetosphere of Uranus

Uranus was expected to have an unusual magnetosphere, because at the present epoch the rotational pole is tipped so that it points almost directly toward the sun. In fact, this is to a large extent compensated by the 59° tilt of the magnetic dipole relative to the axis of rotation. Thus, the magnetosphere does not wobble much more than that of the Earth, although it turns over and presents a different magnetic field direction to the external interplanetary medium with each half-rotation. The dipole equatorial field strength is almost the same as that of Saturn (0.23 G) and is similarly displaced from the center of the planet so that the surface fields vary from 0.08 to 0.96 G (McNutt, 1991; Bagenall, 1992).

The magnetosphere of Uranus is rather quiet, since there are no strong satellite sources of plasma, as at Jupiter and Saturn, and the rotation of the planet tends to keep any solar-wind–induced convection to beyond $5R_U$ as evidenced by a sudden dropout of hot plasma inside this region. The main ions are protons, which could originate both in the solar wind and in the ionosphere but, probably, mostly in the latter. The energetic particle population is not intense and is sufficiently stable to show absorption features associated with the satellites, as at Saturn.

3.6 The Magnetosphere of Neptune

Neptune's magnetosphere proved to be a surprise when first encountered, since its magnetic dipole axis is inclined by 47° from the axis of rotation of the planet, which, in turn, is inclined by 30° relative to the ecliptic (Figure 7). By chance, the polar cusp was

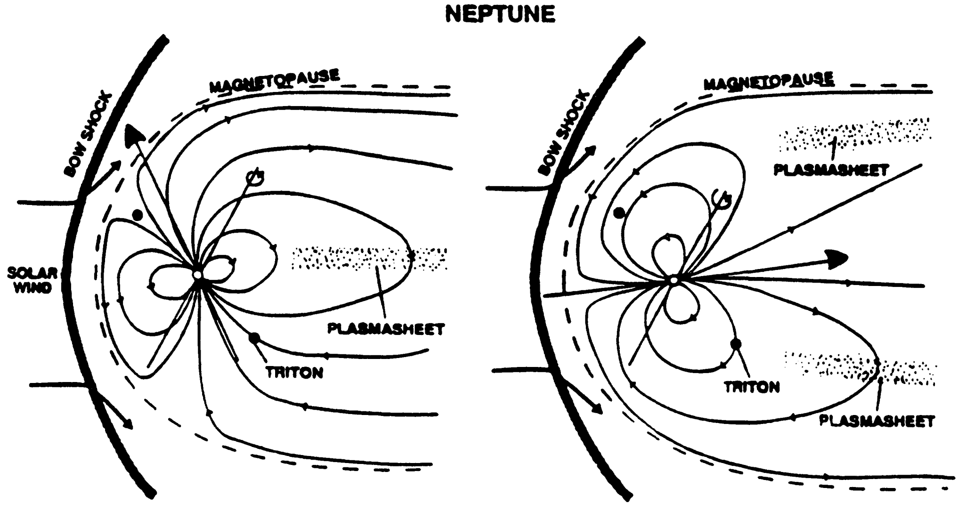

FIG. 7. The magnetosphere of Neptune at two phases of the planet's rotation separated by half a rotation period. (From Bagenall, 1992.)

pointing directly into the solar-wind direction at the time of the approach to the planet by Voyager 2, and the magnetopause was not encountered until well after the initial bow shock crossing. This situation had originally been expected at Uranus but not found. In fact, apart from differences associated with the direction of the axis of rotation, the magnetospheres of the two planets are very similar, the dipole equatorial field strength at Neptune being 0.14 G but the surface field varying from 0.1 to 0.9 G (McNutt, 1991; Bagenall, 1992). As a consequence of the tilts of the magnetic dipole and rotational axes, the magnetotail of Neptune continually changes, having, at times, an Earth-like configuration with a roughly plane plasma sheet separating the two tail lobes and, at other times, the originally expected Uranus configuration with a cylindrical plasma sheet separating the field from the sunward-directed pole (outside) from that of the night-side pole (inside).

The energetic-particle population in the magnetosphere of Neptune is about as intense as that at Uranus but, in comparison with the other planets, is of no great significance. The major satellite Triton, which has a retrograde orbit at $14.6R_{\mathrm{N}}$, well inside the magnetosphere, has an atmosphere. This atmosphere clearly contributes to the plasma background (notably N^+ ions) and also to a neutral hydrogen disk which extends in to about $8R_{\mathrm{N}}$. There must be some potential for substorms, especially in view of the forced reconfiguration of the magnetotail, as described above. However, the plasma and the rotation may well provide some protection for the inner magnetosphere from the associated convection and other disturbances. There is some evidence for auroral activity.

3.7 The Magnetosphere of Pluto–Charon

It seems unlikely that either of these very interesting bodies is large enough or has a suitable internal constitution to have an intrinsic magnetic field (McNutt, 1991). However, there are indications of the presence of a thin atmosphere, which would permit a Mars-like interaction with a very weak solar wind existing at 40–50 a.u. from the Sun and involving an induced magnetosphere with a maximum field strength of a few nT.

GLOSSARY

Alfvén Wave: Transverse magnetohydrodynamic wave.

Aurora: Emissions from the upper atmosphere (>90 km) associated with the precipitation of magnetospheric particles.

Bow Shock: Collisionless shock wave produced by the interaction of the supersonic solar wind with an obstacle such as a magnetosphere.

Electrojet: Intense electric current flowing in the ionosphere at 100–120 km altitude, especially at the geomagnetic equator and in the auroral zone.

Exosphere: Outer atmosphere of a planet where the collision mean free path of its constituents is comparable to the scale height.

Flare: Eruption on the Sun associated with regions of strong and complex magnetic field.

Geomagnetic Storm: A severe disturbance of the magnetic field, especially of the Earth, associated with multiple substorms (see below) and the buildup of a ring current (see below) which persists for several days.

Ionopause: A plasma barrier ahead of a planet capable of withstanding the pressure of the solar wind.

Ionosphere: The region of the upper atmosphere of the Earth (or other planet) in which photoionization by solar UV produces a plasma (typically in the altitude range 80–1000 km).

Magnetopause: The interface between the magnetosphere and the solar wind.

Magnetosphere: The region surrounding the Earth (or other planet) in which the intrinsic magnetic field plays a dominant role in controlling the behavior of the plasma.

Magnetotail: The comet-like tail of the magnetosphere swept downstream by the solar wind to a distance of thousands of planetary radii.

Pickup ions: Ions produced initially at rest in a moving magnetized plasma.

Plasmasphere: The region surrounding the Earth, typically contained by the shell of magnetic field lines defined by ca. 60° geomagnetic latitude, which contains dense plasma originating from the ionosphere.

Plasmapause: The outer boundary of the plasmasphere.

Plasma Torus: A torus-like region surrounding a planet and containing plasma originating from a satellite (e.g., Io at Jupiter and Titan at Saturn).

Polar Cusp: A region on the sunward side of the magnetosphere where magnetic field lines link the auroral zones and the magnetopause.

Polar Wind: A flow of plasma from the ionosphere in polar regions upward into the magnetosphere.

Radiation Belt: [Also called the Van Allen belt(s) in the case of the Earth.] Regions of energetic particles trapped in a planetary magnetosphere.

Reconnection: A term used to describe the changes in linkages of magnetic field lines in magnetized plasmas.

Ring Current: The electric current surrounding the Earth and produced as a result of the pressure of trapped energetic particles in the magnetosphere.

Solar Wind: A supersonic wind flowing radially outward from the Sun into space, which interacts with the magnetospheres of the planets.

Substorm: A short-lived disturbance, usually part of a geomagnetic storm, in which there is strong auroral electrojet activity, auroral particle precipitation, and the injection of fresh plasma into the magnetosphere from the magnetotail.

Whistlers: Audio-frequency radio waves that travel through the magnetosphere, more or less along magnetic field lines, and that have a characteristic frequency dispersion running from high to low following impulsive injection from lightning flashes.

Works Cited

Akasofu, S.-I. (1977), *Physics of Magnetospheric Substorms*, Dordrecht: D. Reidel.

Axford, W. I. (1969), in: D. J. Williams, G. D. Mead (Eds.), *Magnetospheric Physics, Res. Geophys.* **7** (1), 321–459.

Axford, W. I. (1970), in: B. M. McCormac (Ed.), *Particles and Fields in the Magnetosphere*, Dordrecht: D. Reidel, pp. 46–59.

Axford, W. I. (1984), in: E. W. Hones, Jr. (Ed.), *Magnetic Reconnection in Space and Laboratory Plasma*, Geophysical Monograph No. 30, Washington, DC: American Geophysical Union, pp. 1–8.

Bagenall, Fran (1992), *Annu. Rev. Earth Planet. Sci.* **20**, 289–328.

Banks, P. M. (1979), in: L. J. Lanzerotti, C. F. Kennel, E. N. Parker (Eds.), *Solar System Plasma Physics*, Amsterdam: North-Holland, **2**, 57–102.

Birn, J. (1991), *Rev. Geophys.* **29**, Suppl. 1049–1065.

Blamont, J. E. (Ed.) (1991), *Planet. Space Sci.* **39**, 1–394.

Bone, Neil (1991), *The Aurora: Sun-Earth Interactions*, Chichester: Ellis Horwood.

Chang, T. (Ed.) (1986), *Ion Acceleration in the Magnetosphere and Ionosphere*, Geophysical Monograph No. 38, Washington, DC: American Geophysical Union.

Chapman, S. (1967), in: S. Matsushita, W. H. Campbell (Eds.), *Physics of Geomagnetic Phenomena*, Vol. 1, New York: Academic Press, pp. 1–28.

Chapman, S., Ferraro, V. C. A. (1931), *Terrest. Magnetism Atmospheric Elec.* **36**, 77–97 and 171–186.

Cohen, R. (1967), in: S. Matsushita, W. H. Campbell (Eds.), *Physics of Geomagnetic Phenomena*, Vol. 1, New York: Academic, pp. 561–615.

Cornwall, J. M., Schulz, M. (1979), *Solar System Plasma Phys.* **3**, 165–210.

Crooker, N. U., Burke, W. J. (1991), *Rev. Geophys.* **29**, Suppl. 1017–1027.

Dessler, A. J. (Ed.) (1983), *Physics of the Jovian Magnetosphere*, Cambridge, U.K.: Cambridge University Press.

Gehrels, T. (Ed.) (1976), *Jupiter*, Tucson: University of Arizona Press.

Gehrels, T., Matthews, M. S. (Eds.) (1984), *Saturn*, Tucson: University of Arizona Press.

Gold, T. (1959), *J. Geophys. Res.* **64**, 1219–1224.

Gombosi, T. (1991), *Rev. Geophys.* **29**, Suppl. 976–984.

Hoffet, J. A., Egeland, A. (Eds.) (1984), *The Polar Cusp*, Dordrecht: D. Reidel.

Hunten, D. M., Colin, L., Donahue, T. M., Moroz, V. I. (Eds.) (1983), *Venus*, Tucson: University of Arizona Press.

Ip, W.-H., Axford, W. I. (1990), in: W. H. Huebner (Ed.), *Physics and Chemistry of Comets*, Heidelberg: Springer-Verlag, 177–233.

Johnstone, A. D. (Ed.) (1991), *Cometary Plasma Processes*, Geophysical Monograph No. 61, Washington, DC: American Geophysical Union.

Kamide, Y., Slavin, J. A. (Eds.) (1986), *Solar Wind–Magnetosphere Coupling*, Dordrecht: D. Reidel.

Kan, J. R., Potemra, T. A., Kokobun, S., Iijima, T. (1991), *Magnetospheric Substorms*, Geophysical Monograph 64, Washington, DC: American Geophysical Union.

Lanzerotti, L. J., Kennel, C. F., Parker, E. N. (Eds.) (1979), *Solar System Plasma Physics*, Vols. I–III, Amsterdam: North-Holland.

Liu, A. T. Y. (Ed.) (1987), *Magnetotail Physics*, Baltimore, MD: Johns Hopkins University Press.

Luhmann, J. G. (1991), *Rev. Geophys.* **29**, Suppl. 965–975.

McNutt, R. L., Jr. (1991), *Rev. Geophys.* **29**, Suppl. 985–979.

Meng, C.-I., Rycroft, M. J., Frank, L. A. (Eds.) (1991), *Auroral Physics*, Cambridge, U.K.: Cambridge University Press.

Meng, C.-I., Mauk, B. H. (1991), *Rev. Geophys.* **29**, Suppl. 1028–1038.

Moore, T. E. (1991), *Rev. Geophys.* **29**, Suppl. 1039–1048.

Omholt, A. (1971), *The Optical Aurora*, Heidelberg: Springer-Verlag.

Onsager, T. G., Thomsen, M. F. (1991), *Rev. Geophys.* **29**, Suppl. 998–1007.

Potemra, T. (Ed.) (1984), *Magnetospheric Currents*, Geophysical Monograph No. 28, Washington, DC: American Geophysical Union.

Rishbeth, H., Garriott, O. K. (1969), *Introduction to Ionospheric Physics*, New York: Academic Press.

Russell, C. T. (1987), in: S.I. Akasofu, Y. Kamide (Eds.), *The Solar Wind and the Earth*, Dordrecht: D. Reidel.

Siscoe, G. L., Slavin, J. A. (1979), *Rev. Geophys. Space Sci.* **17**, 1677–1693.

Smith, M. F. (1991), *Rev. Geophys.* **29**, Suppl. 1008–1016.

Sonnerup, B. U. O. (1979), in: L. J. Lanzerotti, C. F. Kennel, E. N. Parker (Eds.), *Solar System Plasma Physics*, Vol. 3, Amsterdam: North-Holland, pp. 45–108.

Van Allen, J. A. (1987), *Origins of Magnetospheric Physics*, Washington, DC: Smithsonian Institute Press.

Van Zandt, T. E. (1967), in: S. Matsushita, W. H. Campbell (Eds.), *Physics of Geomagnetic Phenomena*, Vol. 1, New York: Academic Press, pp. 509–559.

Vilas, F., Chapman, C. R., Matthews, M. S. (Eds.) (1990), *Mercury*, Tucson: University of Arizona Press.

Williams, D. J., Mead, G. D. (Eds.) (1969), *Magnetospheric Physics, Rev. Geophys.* **7**, 1–459.

Williams, D. J. (1983), in: J. G. Roederer (Ed.), *Process in Solar-Terrestrial Physics*, Dordrecht: D. Reidel, 223–236.

Further Reading

Details on current developments can be found in the proceedings of COSPAR Meetings (Advances in Space Research) and the Chapman Conferences of the American Geophysical Union (Geophysical Monographs). The books by Lanzerotti *et al.* (1979), Kamide and Slavin (1986), Russell (1987), and Liu (1987) are helpful and extensive. Most of the current literature is contained in journals, such as the *Journal of Geophysical Research, Reviews of Geophysics, Geophysical Research Letters, Planetary and Space Science, Annales Geophysicae*, and *Space Science Reviews*.

MAGNETOSTATICS

ROBERT M. DEL VECCHIO, *North American Transformer, 1200 Piper Drive, Milpitas, California, U.S.A.*

Introduction 207
1. **Units** 208
2. **Magnetic Induction B and Vector Potential** 210
2.1 Current Density 210
2.2 Biot-Savart Law 210
2.3 Ampère's Circuital Law 211
2.4 Examples of Field Calculations 211
2.4.1 Straight Wire Segment 211
2.4.2 Ideal Solenoid and Toroid 212
3. **Material Magnetism** 213
3.1 Magnetization 214
3.2 Magnetization Currents 214
3.3 Magnetic Field Intensity *H* 215
3.4 Permeability and Hysteresis 215
3.5 Fictitious Magnetic Poles 216
3.6 Magnetic Scalar Potential 217
4. **Interface Conditions on the Fields and Potentials** 218
5. **Magnetic Circuits** 219
5.1 Example 219
6. **Magnetostatic Energy** 220
6.1 Faraday's Law 220
6.2 Energy in Terms of Current Sources 220
6.3 Energy in Terms of Fields 221
7. **Flux Linkages and Inductance** 221
7.1 Energy in Terms of Inductances 222
7.2 Example 222
8. **Magnetic Forces and Torques** 222
8.1 Force and Torque on Current Distributions 222
8.1.1 Example 223
8.2 Force and Torque on Localized Current Distributions 223
8.3 Maxwell Stress Tensor 223
8.3.1 Examples 224
8.4 Force and Torque by Energy Methods 224
8.4.1 Example 226
Appendix: Mathematical Notation and Useful Theorems 226
Further Reading 227

INTRODUCTION

Magnetism developed as a science separate from the study of electricity until the discovery by H. C. Oersted in 1819 that an electric current deflects a compass needle. Prior to this, magnetism had been treated from the point of view of magnetic poles—north and south. Even though it was not possible to isolate these poles, a close approximation to an individual point pole was obtained at the ends of a long thin permanent magnet. Thus it was possible experimentally to determine a law of pole attraction and repulsion. The force between poles varied inversely as the square of the separation and directly as the product of pole strengths, in exact analogy with the force between point electric charges. Both the electrostatic and magnetostatic force laws were established by C. A. Coulomb.

Permanently magnetized materials such as lodestone were known to the ancient Greeks. These materials eventually came to be used in compasses for navigation. In 1600, with the publication of *De Magnete*, W. Gilbert put forth the view that the Earth itself was a giant magnet and that this could account for the behavior of compasses. With the development of the galvanic pile or battery by A. Volta around 1800, it became feasible to experiment with direct-current electricity. This invention made possible Oersted's discovery that currents produce magnetic fields. The laws governing the magnetic interactions between currents were formulated by

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

A. M. Ampère, J. B. Biot, and F. Savart soon after Oersted's discovery.

Ampère quickly realized that all magnetic phenomena, even permanent magnetism, could be accounted for by currents or charges in motion. Thus he postulated the existence of circulating atomic currents as the origin of material magnetism. These currents, called *Ampèrian currents*, do indeed exist according to our present understanding of atomic physics. However, the electron itself possesses an intrinsic magnetic moment that is much harder to understand in terms of charges in motion. Nevertheless, basing magnetostatics on magnetic fields produced by currents does not require the notion of isolated magnetic poles, whose existence has never been demonstrated, and thus leads to a natural and unified description of magnetism.

It should be noted that the use of fictitious magnetic poles can be justified as a mathematical device in solving a broad class of magnetostatic problems. It often leads to simpler methods of solution than would be possible from a strictly Ampèrian current viewpoint. This procedure involves no contradiction with fundamental theory or, more importantly, with experiment. However, the terminology such as the use of the word *poles* can be misleading.

Modern theories of magnetism are based on a quantum-mechanical treatment of the electron's spin and orbital magnetic moments and their interactions. Because of the mathematical difficulties involved in treating this complex many-body system, our understanding of magnetism at this level is still incomplete. However, by making suitable approximations and taking averages, a semiclassical treatment of magnetism, called micromagnetics, has emerged that retains many features of the atomic-level description. The discreteness of quantum mechanics, however, is replaced by a continuum description. Thus a continuous magnetization distribution results. The magnetization vector has a constant magnitude but can vary in direction from point to point. Quantum-mechanical interactions are approximated by interactions sensitive to the direction of the magnetization vector. In addition, the usual classical dipole–dipole interactions are present as well as interactions with external magnetic fields.

While micromagnetics calculations can also present formidable difficulties when applied to complex geometries, successful calculations have been performed for ideal or reasonably simple configurations. For instance, the characteristics of magnetic domains (regions where the magnetization vectors are aligned) and domain walls (transition regions between domains having different orientations) can be understood to a large extent from this theory. Magnetic domains were predicted by P. Weiss in 1907 and were subsequently observed experimentally.

Magnetostatics and micromagnetics play a crucial role in the magnetic recording industry, e.g., in magnetic-tape, hard- or floppy-disk, and bubble-memory technologies (see MAGNETIC STORAGE OF INFORMATION). States of remanent magnetization represent information on these devices that can be stored for long periods of time.

Magnetic fields produced by permanent magnets or electromagnets (see MAGNETS) are used to bend and focus charged particles in particle accelerators and in electron microscopes. Magnetic fields are used to store energy in inductors, to transfer electrical energy in transformers, and to transfer electrical to mechanical energy or vice versa in motors and generators.

In this article, we present a classical-field approach to magnetostatics. Vector and scalar potentials are defined as an aid to the mathematical development of the theory. Material magnetism is treated classically, which means that in materials with a domain structure the magnetization vector represents an average over many domains and therefore can vary in both magnitude and direction from point to point in the medium. The phenomenon of hysteresis in ferromagnetic materials is presented. General expressions are derived for magnetic force and energy. Magnetic flux and flux linkages are introduced together with the concept of inductance. The magnetic circuit approach is presented as a practical method of obtaining an approximate solution to certain types of problems. Throughout, useful examples are given to illustrate the theory.

1. UNITS

The development of a system of units to measure magnetic effects mirrors the history of discoveries in this field. When the two subjects of electricity and magnetism were thought to be unrelated, the force laws between point electric charges q_e and magnetic poles or charges q_m were used to define the unit electric or magnetic charge. Thus the forces in vacuum are

$$\mathbf{F}_e = k_e \frac{q_e q_e'}{R^2} \hat{\mathbf{a}}_R, \tag{1}$$

$$\mathbf{F}_m = k_m \frac{q_m q_m'}{R^2} \hat{\mathbf{a}}_R, \tag{2}$$

where R is the separation distance, $\hat{\mathbf{a}}_R$ is a unit vector pointing along R, and k_e and k_m are constants that set the units. The simplest choice is to take $k_e = k_m = 1$. This defines the charge in terms of mechanical units $(ML^3T^{-2})^{1/2}$. This also, in a natural way, sets the units for the electric and magnetic fields, which are defined at each point in space as the force that would act on a unit test charge placed at that point:

$$\mathbf{E} = \mathbf{F}_e/q_e, \tag{3}$$

$$\mathbf{B} = \mathbf{F}_m/q_m. \tag{4}$$

However, once it was realized that magnetic fields are created by currents, the question of the units of current had to be addressed. In vacuum, a current I flowing in a thin wire of infinitesimal length dl produces a magnetic field given by

$$d\mathbf{B} = k_i(I d\mathbf{l} \times \hat{\mathbf{a}}_R/R^2), \tag{5}$$

where $d\mathbf{l}$ is the directed length of the wire, R is the distance from the wire element to the field point, and $\hat{\mathbf{a}}_R$ is a unit vector directed along R from the source towards the field point (see Fig. 1). k_i is used to adjust the units. This is called the Biot-Savart law. The mathematical notation used in this paper and key theorems and definitions are indicated in the Appendix. The cross product is defined there in Eq. (A3).

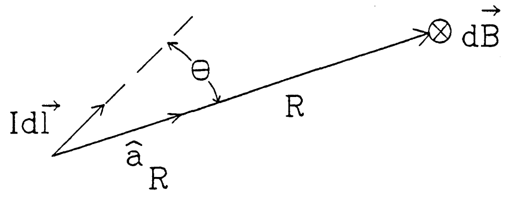

FIG. 1. Biot-Savart law illustrated.

The magnetic field given by Eq. (5), or any magnetic field **B**, acts on another current element $I'dl'$ to produce a force given by

$$d\mathbf{F} = k_f I' d\mathbf{l}' \times \mathbf{B}. \tag{6}$$

When **B** is produced by another current element, the force law between two such elements becomes

$$d\mathbf{F} = k_f k_i \frac{I'I}{R^2} d\mathbf{l}' \times (d\mathbf{l} \times \hat{\mathbf{a}}_R). \tag{7}$$

Such a force between isolated current elements cannot occur in reality but is an abstraction from experiments performed on complete circuits.

By setting $k_f k_i = 1$, Eq. (7) can be used to define a unit of current that in the cgs system is called the abampere. However, since current is the flow of electrical charge, a unit of charge, called the abcoulomb, is naturally defined as the charge accumulated by the flow of one abampere for one second. But Eq. (1) with $k_e = 1$ defines another unit of charge in the cgs system, called the statcoulomb. It was discovered experimentally that the ratio of a given charge measured in statcoulombs to the same charge measured in abcoulombs was numerically and dimensionally equal to the speed of light, $c = 2.9979 \times 10^{10}$ cm/s. The consequence of this on the system of units is that, from Eqs. (1) and (7) and the observation that the force measured in a given set of mechanical units between two given current elements or two given charges must be the same regardless of the units in which the charges or currents are expressed, we require

$$k_e/k_f k_i = c^2. \tag{8}$$

It follows from Eqs. (4) and (5) that for the magnetic field produced by a unit pole to be the same as that produced by a unit current element, we must have $k_i = k_m$.

Various systems of units can be identified with the choice of k's within the above constraints. With $k_f = k_i = k_m = 1$, $k_e = c^2$, currents in abamperes, and charges in abcoulombs, we get the electromagnetic system (cgs-emu). With $k_e = k_f = 1$, $k_i = k_m = 1/c^2$, we get the electrostatic system (cgs-esu). It was noted by C. F. Gauss that one could have a consistent set of units with charge measured in statcoulombs ($k_e = 1$) and current in statamperes while retaining Eq. (2) with $k_m = 1$, provided we set $k_i = 1/c$. In this Gaussian system, $k_i \neq k_m$ since the unit pole and current element produce different fields B. However, $1/c$ provides the required conversion factor. We also have $k_f = 1/c$ in this system.

While the Gaussian system is favored in fundamental physics research, the system of units generally used in practical engineering or laboratory work is the rationalized mks system (or mksa system since the ampere is the standard unit of current). In this system, charge is assigned a new fundamental unit called the coulomb, abbreviated C. A current of 1 C/s is called an ampere, k_e in vacuum is defined in terms of another constant ϵ_0 by $k_e = 1/4\pi\epsilon_0$. k_i in vacuum is defined in terms of a new constant μ_0 by $k_i = \mu_0/4\pi$. In Eq. (6), $k_f = 1$ so that from Eq. (8), $\epsilon_0\mu_0 = 1/c^2$. The units of ϵ_0 or μ_0 can readily be determined in the mksa system from Eqs. (1) and (7). However, these are more simply expressed in terms of derived units,

$$\mu_0 = 4\pi \times 10^{-7}\ \mathrm{H/m},$$
$$\epsilon_0 = 1/\mu_0 c^2 = 8.8542 \times 10^{-12}\ \mathrm{F/m}, \tag{9}$$

where the henry (H) is the unit of inductance (V s/A) and the farad (F) the unit of capacitance (C/V), and 1 V = 1 J/C. We will use mksa units in this paper. This is a subset of the Système International d'Unités or SI system of units. μ_0 is called the permeability of vacuum and ϵ_0 the permittivity of vacuum.

2. MAGNETIC INDUCTION B AND VECTOR POTENTIAL

The magnetic field, like the electric field, was devised to replace action at a distance. Thus one object (an electric charge or a current element) gives rise to a field throughout space that then acts on another object. Since these fields are vectors, the net field arises as the vector sum or integral of the fields produced by all the relevant objects.

2.1 Current Density

In general the current need not be confined to a thin wire, and we define a current density **J** as the current per unit area flowing through an infinitesimal directed surface element $d\mathbf{S}$ oriented in such a way that the unit normal to dS, **n**, is in the direction of **J**. Thus

$$\mathbf{J} = \frac{dI}{dS}\mathbf{n}. \tag{10}$$

In this case $Id\mathbf{l}$ generalizes to $(\mathbf{J}\cdot\mathbf{n})dSd\mathbf{l}$. But since **J**, **n**, and $d\mathbf{l}$ are collinear, this can be written as $\mathbf{J}dV$, where dV is the volume element. Occasionally there is a need for a surface current element. This is defined as the current per unit length flowing perpendicular to an infinitesimal element of length dL along a surface:

$$\mathbf{K} = \frac{dI}{dL}\hat{\mathbf{a}}_K, \tag{11}$$

where $\hat{\mathbf{a}}_K$ is normal to the direction of $d\mathbf{L}$ in the direction of **K**. Thus $Id\mathbf{l}$ becomes $(\mathbf{K}\cdot\hat{\mathbf{a}}_K)dLd\mathbf{l}$. But since **K**, $\hat{\mathbf{a}}_K$, and $d\mathbf{l}$ are collinear, this becomes $\mathbf{K}dS$, where dS is the element of surface area.

Magnetostatics deals with stationary magnetic fields produced by steady currents. It also assumes that any electric charge present is not changing with time. Therefore the net current flowing into any closed surface is zero, since otherwise charge would accumulate. This implies by the divergence theorem, Eq. (A4), that

$$\nabla\cdot\mathbf{J} = 0. \tag{12}$$

2.2 Biot-Savart Law

In terms of current density, Ampere's force law, Eq. (6), can be expressed more generally as

$$d\mathbf{F} = \mathbf{J}\times\mathbf{B}\,dV, \tag{13}$$

where $\mathbf{J}\times\mathbf{B}$ is a force density. The Biot-Savart law, Eq. (5), now becomes, in integral form,

$$\mathbf{B} = \frac{\mu_0}{4\pi}\int\frac{\mathbf{J}\times\hat{\mathbf{a}}_R}{R^2}dV', \tag{14}$$

where $R = |\mathbf{r} - \mathbf{r}'|$, **r** being a vector from the origin to the field point and **r**′ one to the source point. Primes are used to distinguish source variables whenever necessary. The units of B are teslas (T). This field is called the magnetic flux density or the magnetic induction. It is often simply referred to as the B field. It is considered the fundamental magnetic field because it produces the force on a current element.

In Eq. (14), R involves both the source and field coordinates, as does $\hat{\mathbf{a}}_R$ since $\hat{\mathbf{a}}_R = \mathbf{R}/R$. $\mathbf{J}$ and dV' are evaluated at the source point. Let ∇ denote taking the gradient with respect to the field position and ∇' with respect to the source position. Then

$$\nabla'(1/R) = -\nabla(1/R) = \hat{\mathbf{a}}_R/R^2. \tag{15}$$

Using this, we can rewrite Eq. (14):

$$\begin{aligned}\mathbf{B} &= -\frac{\mu_0}{4\pi}\int \mathbf{J} \times \nabla\left(\frac{1}{R}\right)dV' \\ &= \nabla \times \frac{\mu_0}{4\pi}\int \frac{\mathbf{J}}{R}\,dV'.\end{aligned} \tag{16}$$

The second equality follows from Eq. (A15) and the fact that ∇ operates on the field position. Thus $\mathbf{B}$ has been expressed as

$$\mathbf{B} = \nabla \times A, \tag{17}$$

with $\mathbf{A}$, called the vector potential, given by

$$\mathbf{A} = \frac{\mu_0}{4\pi}\int \frac{\mathbf{J}dV'}{R} \quad \text{or} \quad \mathbf{A} = \frac{\mu_0}{4\pi}\int \frac{Id\mathbf{l}'}{R}. \tag{18}$$

The two forms for $\mathbf{A}$ are appropriate to a distributed current source and one localized to a thin wire, respectively.

We should note that since the curl of a gradient is identically zero, Eq. (A10), an arbitrary gradient can be added to $\mathbf{A}$ without changing the value of B. This arbitrariness is exploited in calculations by imposing additional requirements on $\mathbf{A}$, e.g.,

$$\nabla \cdot \mathbf{A} = 0. \tag{19}$$

That this condition can be imposed consistently follows from Helmholtz's theorem, which states that a vector field is determined by specifying its curl and divergence. [The curl of $\mathbf{A}$ is given in Eq. (17).] Equation (19) is called the Coulomb gauge. (Changing $\mathbf{A}$ without changing $\mathbf{B}$ is called a gauge transformation.) Also, since the divergence of a curl vanishes identically, Eq. (A9), we see from Eq. (17) that

$$\nabla \cdot \mathbf{B} = 0. \tag{20}$$

This equation recalls the experimental fact that isolated magnetic poles do not exist.

2.3 Ampère's Circuital Law

We obtain another important relationship by taking the curl of $\mathbf{B}$. Using Eq. (A11) and the Coulomb gauge, we obtain

$$\nabla \times \mathbf{B} = -\nabla^2 \mathbf{A} = -\frac{\mu_0}{4\pi}\int \mathbf{J}\nabla^2\left(\frac{1}{R}\right)dV'. \tag{21}$$

Here ∇^2 has been taken inside the integral, where it can operate only on $1/R$ since this is the only factor depending on the field point. If we use spherical coordinates, we find that $\nabla^2(1/R) = 0$ except at the origin where it is undefined. However, by applying the divergence theorem to a small volume surrounding the origin, we find that $\int \nabla^2(1/R)\,dV = -4\pi$ so that Eq. (21) becomes

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J}. \tag{22}$$

Applying Stokes's theorem to this, we obtain

$$\oint_{\mathscr{C}} \mathbf{B} \cdot d\mathbf{l} = \mu_0 \int_S \mathbf{J} \cdot \mathbf{n}\, dS = \mu_0 I, \tag{23}$$

where I is the total current through the surface S, which is bounded by the closed path $\mathscr{C}$. This is Ampère's circuital law and is very useful when applied to problems that have sufficient symmetry.

2.4 Examples of Field Calculations

2.4.1 Straight Wire Segment We calculate the field of a straight wire segment carrying current I as shown in Fig. 2(a). This cannot exist as a separate entity but could be part of a circuit made up of connected wire segments. We use Eq. (5) to obtain the field $\mathbf{B}$ at some point a perpendicular distance h from the wire or its straight-line extension. The current flows upward in the direction of the positive x axis so that $dl = dx$. We have, without loss of generality, shown the wire

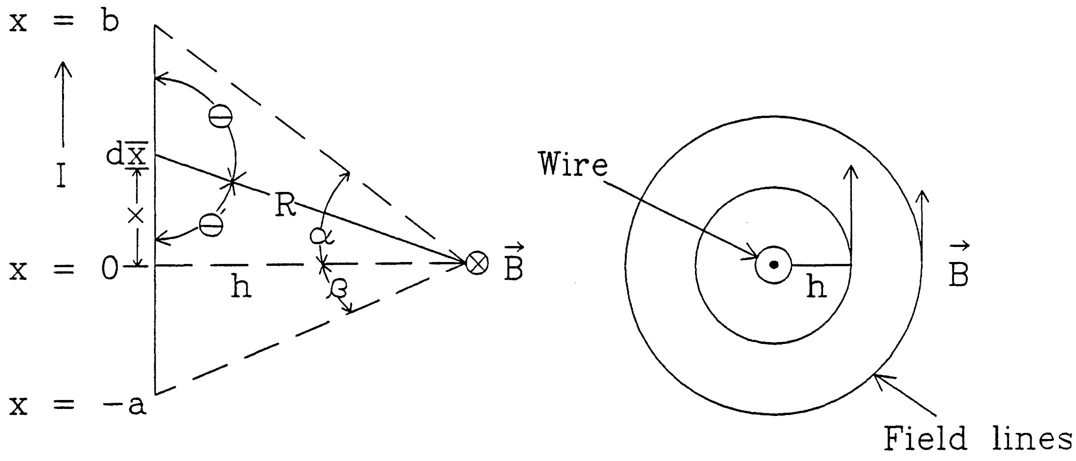

(a) (b)

FIG. 2. (a) Geometric parameters used in calculating the field *B* produced by a straight wire segment carrying current *I*. (b) Pattern of field lines surrounding the wire.

segment and field point in the plane of the paper in Fig. 2(a). We see that $d\mathbf{l} \times \hat{\mathbf{a}}_R$ points into the plane of the paper for any position along the wire so that the vector sum reduces to a scalar sum. In fact $|d\mathbf{l} \times \hat{\mathbf{a}}_R| = dx \sin\theta = dx \sin\theta' = hdx/R$ from the figure. Thus

$$\begin{aligned} B &= \frac{\mu_0 Ih}{4\pi}\int_{-a}^{b}\frac{dx}{R^3} \\ &= \frac{\mu_0 Ih}{4\pi}\int_{-a}^{b}\frac{dx}{(x^2+h^2)^{3/2}} \\ &= \frac{\mu_0 Ih}{4\pi}\left[\frac{x}{h^2(x^2+h^2)^{1/2}}\right]_{-a}^{b} \\ &= \frac{\mu_0 I}{4\pi h}\left[\frac{b}{(b^2+h^2)^{1/2}}+\frac{a}{(a^2+h^2)^{1/2}}\right] \\ &= \frac{\mu_0 I}{4\pi h}(\sin\beta+\sin\alpha). \end{aligned} \tag{24}$$

The last form is particularly simple. Because the selected field position could have been anywhere along a circle of radius *h* centered on the wire or its straight-line extension without affecting the calculation, we see that the field directions are tangent to such circles everywhere in space as shown in Fig. 2(b). The sense of the field is easily found by pointing the thumb of the right hand in the direction of current flow. Then the fingers will curl in the field direction. If we let the wire become infinitely long so that $\alpha = \beta = \pi/2$, we obtain $B = \mu_0 I/2\pi h$. There is sufficient symmetry in this case to use Ampère's circuital law directly to find *B*.

2.4.2 Ideal Solenoid and Toroid The magnetic field *B* inside an infinitely long ideal solenoid or an ideal toroid can be found by exploiting the symmetry and using Ampère's circuital law, Eq. (23) (see Fig. 3). The solenoid and toroid are assumed to be wrapped with many closely spaced turns of fine wire. From symmetry considerations, only an axial field can exist inside the solenoid. (There

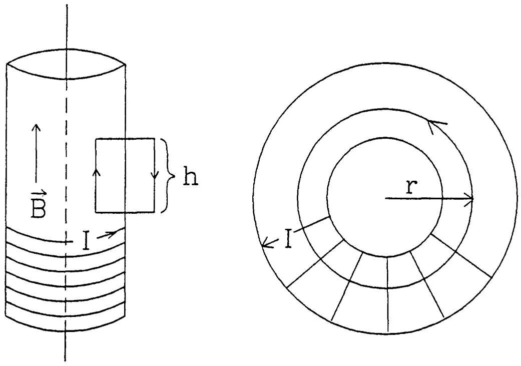

(a) (b)

FIG. 3. Using Ampère's circuital law to find the field *B* inside (a) an ideal solenoid and (b) a toroid.

is a small azimuthal B component outside the solenoid due to the slight pitch of the coil.) The solid path in the figure has a nonvanishing contribution only along the inside vertical path, and by the circuital law $Bh = \mu_0 nhI$, where n is the number of turns per unit length and I the current in each turn. Therefore the field inside the solenoid is

$$B = \mu_0 nl \tag{25}$$

and is independent of radial distance from the axis.

For the toroid in Fig. 3(b), symmetry requires that B tangential to the circular path of radius r shown be constant along the path. This path loops the entire toroidal current of NI, where N is the total number of turns and I the current in each turn. Therefore, by the circuital law, $2\pi r B_t = \mu_0 NI$, so that

$$B_t = \mu_0 NI/2\pi r. \tag{26}$$

Thus this tangential component is a function of r.

3. MATERIAL MAGNETISM

Magnetic fields that arise from the presence of matter have their origin in localized atomic currents. To study this type of magnetism, we therefore begin with the study of the magnetic fields generated by localized current sources at distances far from the sources. This is done most conveniently by considering the vector potential, Eq. (18), which we write more explicitly in terms of source (primed) and field coordinates as

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int \frac{\mathbf{J}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\,dV'. \tag{27}$$

Since $r' \ll r$, we may approximate the denominator:

$$\begin{aligned}|\mathbf{r} - \mathbf{r}'| &\cong \sqrt{r^2 - 2\mathbf{r}\cdot\mathbf{r}'}\\ &= r\sqrt{1 - 2\mathbf{r}\cdot\mathbf{r}'/r^2}\\ &\cong r(1 - \mathbf{r}\cdot\mathbf{r}'/r^2)\end{aligned} \tag{28}$$

so that Eq. (27) becomes approximately

$$A \cong \frac{\mu_0}{4\pi r}\int \mathbf{J}(\mathbf{r}')\left(1 + \frac{\mathbf{r}\cdot\mathbf{r}'}{r^2}\right)dV'. \tag{29}$$

Because the current source is localized, $\int \mathbf{J}dV' = 0$. (If the current were confined to a thin wire, then $\mathbf{J}\,dV = I\,d\mathbf{l}$ and $I \oint d\mathbf{l} = 0$ since the current loop is closed. Because $\nabla\cdot\mathbf{J} = 0$, tubes of current can be defined that approximate thin wire circuits.) Therefore, to lowest order,

$$A \cong \frac{\mu_0}{4\pi r^3}\int \mathbf{J}(\mathbf{r}')(\mathbf{r}\cdot\mathbf{r}')dV'. \tag{30}$$

Using Eq. (A13), we can write

$$\mathbf{J}(\mathbf{r}\cdot\mathbf{r}') = (\mathbf{r}\cdot\mathbf{J})\mathbf{r}' - \mathbf{r}\times(\mathbf{r}'\times\mathbf{J}). \tag{31}$$

Choose a Cartesian coordinate system with $\mathbf{r}$ along the x axis (ith coordinate) and consider the jth component of the left-hand side of Eq. (31):

$$\begin{aligned}rJ_j r_i' &= r[\nabla'\cdot(r_j'\mathbf{J})r_i']\\ &= r[\nabla'\cdot(r_i'r_j'\mathbf{J}) - r_j'\mathbf{J}\cdot\nabla' r_i']\\ &= r[\nabla'\cdot(r_i'r_j'\mathbf{J}) - J_i r_j'].\end{aligned} \tag{32}$$

The first equality in Eq. (32) follows from Eq. (A14), together with $\nabla'\cdot\mathbf{J} = 0$ and $\nabla'(r_j') = \mathbf{j}$, the unit vector along the jth Cartesian coordinate. The second equality follows directly from Eq. (A14) and the third equality from $\nabla'(r_i') = \mathbf{i}$, the unit vector along the ith coordinate direction. Now the volume integral over the divergence produces a surface integral by the divergence theorem, and this vanishes if the surface is taken outside the current distribution. Therefore, omitting the integral signs for convenience,

$$[\mathbf{J}(\mathbf{r}\cdot\mathbf{r}')]_j = rJ_j r_i' = -rJ_i r_j' = -[(\mathbf{r}\cdot\mathbf{J})\mathbf{r}']_j \tag{33}$$

so that the first term on the right-hand side of Eq. (31) is the negative of the left-hand side. Hence

$$\int \mathbf{J}(\mathbf{r}\cdot\mathbf{r}')dV' = -\tfrac{1}{2}\int \mathbf{r}\times(\mathbf{r}'\times\mathbf{J})dV' \tag{34}$$

so that Eq. (30) may be written

$$\mathbf{A} = \frac{\mu_0}{4\pi r^3}\left[\frac{1}{2}\int \mathbf{r}'\times\mathbf{J}\,dV'\right]\times\mathbf{r}. \tag{35}$$

The term in brackets is called the magnetic moment of the localized current distribution $\mathbf{m}$:

$$\mathbf{m} = \tfrac{1}{2}\int \mathbf{r}' \times \mathbf{J}\, dV' \quad \text{or}$$
$$\mathbf{m} = I\,\tfrac{1}{2}\oint \mathbf{r}' \times d\mathbf{l}'. \tag{36}$$

The latter expression applies to a thin wire loop. In the case of a plane loop, the magnetic moment equals the current times the loop area, $\mathbf{m} = IS\mathbf{n}$, where $\mathbf{n}$ is the unit normal to the surface, as visualized in Fig. 4. The field $\mathbf{B}$ can be found by means of Eq. (17), remembering that $\mathbf{m}$ is constant with respect to the curl operation. We find, using vector identities found in the Appendix,

$$\begin{aligned}\mathbf{B} &= \nabla \times \left(\frac{\mu_0}{4\pi}\frac{\mathbf{m} \times \hat{\mathbf{a}}_r}{r^2}\right)\\ &= (\mu_0/4\pi r^3)[3(\mathbf{m}\cdot\hat{\mathbf{a}}_r)\hat{\mathbf{a}}_r - \mathbf{m}].\end{aligned} \tag{37}$$

This is seen to be the field of a dipole, analogous to that of an electric dipole of dipole moment $\mathbf{p}$, except for the constants.

3.1 Magnetization

The above derivation shows that, to first order, the field far from a localized current distribution is that of a magnetic dipole. If the current distribution is associated with an atom or molecule, then the dipole approximation would be good a few atomic diameters away, which is small on a macroscopic scale. On an atomic scale, we are not usually interested in field variations but are more interested in averages. We therefore define an average magnetic moment per unit volume, called the magnetization $\mathbf{M}$, by averaging over the atomic moments $\mathbf{m}_{\text{at}}$:

$$\mathbf{M} = \sum_{\Delta V} \frac{\mathbf{m}_{\text{at}}}{\Delta V}. \tag{38}$$

Here the volume ΔV is large enough to include a significant number of atoms but small on a macroscopic scale, so that we can consider $\mathbf{M}$ to vary smoothly with position, except possibly at boundaries.

The vector potential associated with this magnetization distribution is given by

$$\mathbf{A} = \frac{\mu_0}{4\pi}\int \frac{\mathbf{M}(\mathbf{r}') \times \hat{\mathbf{a}}_R}{R^2}\, dV', \tag{39}$$

where $R = |\mathbf{r} - \mathbf{r}'|$. Using Eqs. (15) and (A15), Eq. (39) becomes

$$\begin{aligned}\mathbf{A} &= \frac{\mu_0}{4\pi}\int \mathbf{M}(\mathbf{r}') \times \nabla'\left(\frac{1}{R}\right) dV'\\ &= \frac{\mu_0}{4\pi}\int \left[\frac{\nabla' \times \mathbf{M}}{R} - \nabla' \times \left(\frac{\mathbf{M}}{R}\right)\right] dV'.\end{aligned} \tag{40}$$

Using Eq. (A7) on the last term, we get

$$\mathbf{A} = \frac{\mu_0}{4\pi}\left(\int \frac{\nabla' \times \mathbf{M}}{R}\, dV' + \int \frac{\mathbf{M} \times \mathbf{n}}{R}\, dS'\right). \tag{41}$$

3.2 Magnetization Currents

Comparing Eq. (41) with the vector potential derived for a current density, we see that a magnetization distribution gives rise to an equivalent volume current density given by

$$\mathbf{J}_m = \nabla \times \mathbf{M} \tag{42}$$

and a surface current density

$$\mathbf{K}_m = \mathbf{M} \times \mathbf{n}. \tag{43}$$

The surface current expressions can be made plausible by means of Fig. 5. We see a cross section of a long cylinder having a uniform magnetization distribution. Away from the surface, the currents from neighboring dipoles are in opposite directions so that their fields cancel. However, at the surface there is an unbalanced current that appears to flow completely around the cylinder. If we approximate the dipoles as wire loops carrying

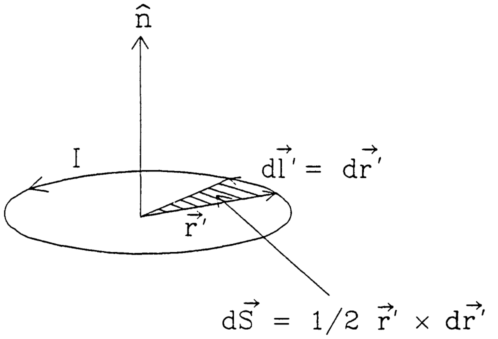

FIG. 4. Illustration of area formula for a plane wire loop.

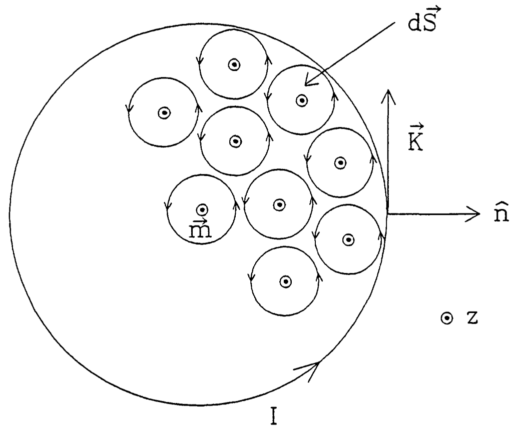

FIG. 5. Illustration of the equivalence of a uniform magnetization distribution with a surface current.

current I and of area dS, their magnetic moment is the current times the loop area. However, their magnetic moment can also be obtained from the magnetization distribution, i.e., $\mathbf{m} = I\, d\mathbf{S} = \mathbf{M}\, dV = \mathbf{M}\, dS\, dz$. Since $d\mathbf{S}$ and $\mathbf{M}$ point in the same direction, the surface current is given in magnitude by $I = M\, dz$ so that the surface current density or current per unit distance in the z direction is just M in magnitude. Its direction is tangential to the surface, in the direction of $\mathbf{K}_m$ in Eq. (43). Thus both the magnitude and direction are given correctly by Eq. (43). A similar plausibility argument can be given to show the reasonableness of Eq. (42).

3.3 Magnetic Field Intensity H

Since a surface current density can be regarded as a limiting case of a volume current density, we can consider the second term in Eq. (41) as a special case of the first term. It can be approximated by means of a rapidly changing magnetization over a short distance or dealt with by means of boundary conditions. Thus the vector potential, when both free currents and magnetizable matter are present, is given by

$$\mathbf{A} = \frac{\mu_0}{4\pi} \int \frac{(\mathbf{J} + \nabla' \times \mathbf{M})}{R}\, dV'. \tag{44}$$

If we repeat the steps leading up to Eq. (22), we obtain, using Eq. (44) for $\mathbf{A}$,

$$\nabla \times \mathbf{B} = \mu_0(\mathbf{J} + \nabla \times \mathbf{M}). \tag{45}$$

$\mathbf{J}$ is the free current density and $\nabla \times \mathbf{M}$ the current density associated with magnetized matter. Rewriting Eq. (45),

$$\nabla \times (\mathbf{B}/\mu_0 - \mathbf{M}) = \mathbf{J}, \tag{46}$$

suggests the definition of a new field, called the magnetic field intensity H or simply the H field,

$$\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}. \tag{47}$$

The units of H and M are amperes/meter or ampere-turns/meter in the mksa system. Thus when magnetized matter is present Eq. (22) is replaced by

$$\nabla \times \mathbf{H} = \mathbf{J} \tag{48}$$

and Ampère's circuital law becomes

$$\oint \mathbf{H} \cdot d\mathbf{l} = I. \tag{49}$$

When $\mathbf{M} = \mathbf{0}$, $\mathbf{B} = \mu_0\mathbf{H}$ so that the two fields differ by a constant factor.

3.4 Permeability and Hysteresis

For many materials, $\mathbf{M}$ is proportional to $\mathbf{H}$, $\mathbf{M} = \chi\mathbf{H}$, where χ is called the magnetic susceptibility. In this case, $\mathbf{B} = \mu_0(1 + \chi)\mathbf{H} = \mu_0\mu_r\mathbf{H} = \mu\mathbf{H}$, where $\mu_r = 1 + \chi$ is the relative permeability and μ is called the permeability. For most materials χ is close to zero, and so μ_r is close to one. Materials with $\chi < 0$ are called diamagnetic while materials with $\chi > 0$ but not much greater than 1 are called paramagnetic. Materials with very large χ, often for which a linear relation between $\mathbf{M}$ and $\mathbf{H}$ is no longer satisfied, are called ferromagnetic. For these latter materials, M or B may not even be a single-valued function of H but may depend on the magnetic history of the material, a phenomenon called hysteresis.

Figure 6(a) shows such a hysteresis process. The material starts out in a demagnetized state at the origin of the B-H plane. We are assuming that B, H, and M are collinear here so that only signed magnitudes are involved. In this configuration, the magnetic moments are, on average, randomly distributed so that they vectorially sum to zero. As H increases, B follows the initial curve. Eventually all the magnetic moments are

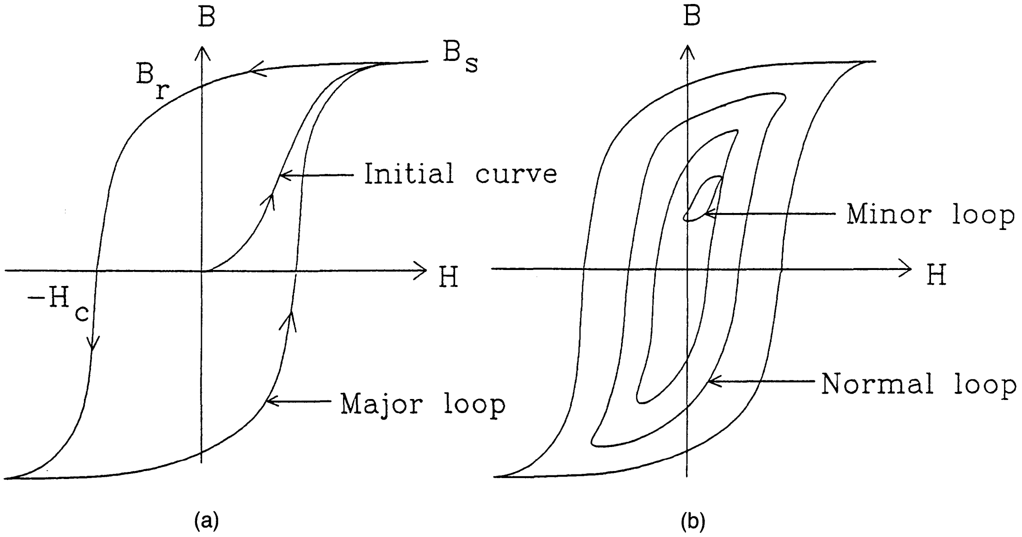

(a) (b)

FIG. 6. (a) Parameters of the major hysteresis loop for a ferromagnetic material. (b) Normal and minor loops.

aligned so that M reaches a maximum value called the saturation magnetization, M_s. The field B associated with this value is called the saturation induction, B_s. Any further increase in B beyond this point is due entirely to an increase in H. As H is lowered, B does not retrace the initial curve but rather follows the upper curve in the figure. When H drops to zero, there is still some induction present, called the remanent induction, B_r. As H is reversed, the magnetization and induction drop further until eventually B reaches zero at a point where $H = -H_c$, where H_c is called the coercivity. If H continues to drop, i.e., continues to increase in a reversed direction, eventually all the dipoles align in the opposite direction at reverse saturation. If H in then increased, the lower curve in Fig. 6(a) is followed. Over many cycles, the outer curve or hysteresis loop will be retraced. While this process occurs for many ferromagnetic materials, in general it is even more complicated because B, H, and M are not necessarily collinear.

Figure 6(b) shows some of the other processes that can occur for collinear hysteresis. If H reverses before saturation is reached, smaller loops, called normal loops, can be traced. Reversing the field in the process of generating a normal loop can lead to the generation of incremental or minor loops, etc. For these nonlinear processes, the ratio of B to H is called the permeability, although other permeabilities, such as the differential permeability, which is the slope of the B-H curve, can be defined.

3.5 Fictitious Magnetic Poles

Many problems in magnetostatics can be simplified mathematically by the use of fictitious magnetic poles, associated with H and not with the more fundamental field B. From the definition equation (47), together with the divergence-free property of B, we see that

$$\nabla \cdot \mathbf{H} = -\nabla \cdot \mathbf{M}. \tag{50}$$

This looks like an electrostatic equation with the electric displacement D corresponding to H and the charge density corresponding to $-\nabla \cdot \mathbf{M}$; i.e.,

$$\rho_m = -\nabla \cdot \mathbf{M} \tag{51}$$

defines an equivalent magnetic charge or pole density. Integrating Eq. (51) over a pillbox-shaped volume partially inside and partially outside a magnetic material, as shown in Fig. 7, and applying the divergence theorem lead to the concept of a surface pole density given by

$$\sigma_m = \mathbf{M} \cdot \mathbf{n}, \tag{52}$$

where $\mathbf{n}$ is the outward normal to the sur-

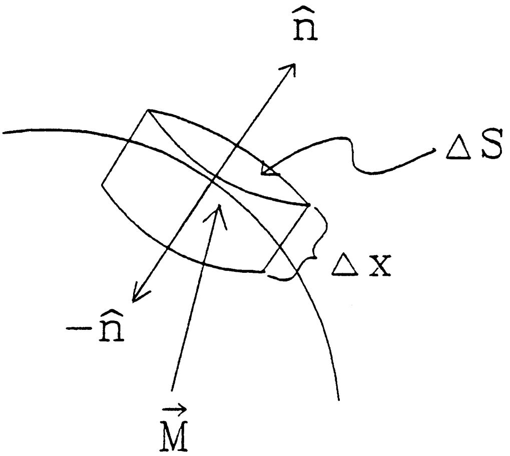

FIG. 7. Surface pole density calculation illustrated.

face. For convenience, as was the case with surface magnetization currents, it is sometimes easier to deal with surface poles as a limiting case of a rapidly varying magnetization over a small distance near the surface of the material or by means of boundary conditions.

From the preceding, it can be seen that these fictitious poles can be thought of as producing the field **H**, just as electric charge produces the field **D**. This point of view is contrasted with the magnetization current point of view in Fig. 8 for an ideal uniformly magnetized cylinder. In Fig. 8(a), the magnetic charge resides on the two ends of the cylinder and the lines of H emanate outwardly from the positive pole and inwardly towards the negative pole. In contrast, B arises from peripheral surface currents, the same as for a tightly wound solenoid, so that the lines of B are continuous, maintaining the same direction through the end surfaces. On the outside of the magnet, B and H are related by the scale factor μ_0. The surface poles tend to demagnetize the magnet. The magnet is operating in the upper left-hand quadrant of the B-H plane of Fig. 6(a).

3.6 Magnetic Scalar Potential

When there are no free currents present or when a region of space excluding such currents and excluding any closed path that could link such currents can be identified, then $\nabla \times \mathbf{H} = 0$ from Eq. (48). This enables us to define a magnetic scalar potential U_m by

$$\mathbf{H} = -\nabla U_m \tag{53}$$

since the curl of a gradient is identically zero. Substituting into Eq. (50), we obtain

$$\nabla^2 U_m = -\rho_m, \tag{54}$$

which is Poisson's equation. If the charge density (volume and surface) is known throughout space, then the solution to Eq. (54)

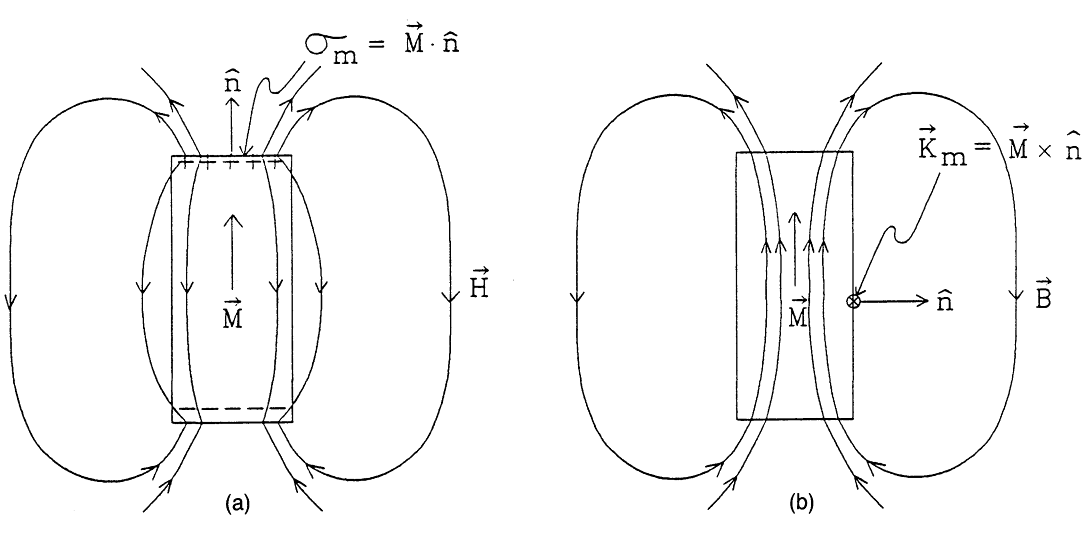
$\sigma_m = \vec{M} \cdot \hat{n}$
$\hat{n}$
$\vec{M}$
$\vec{H}$
(a)
$\vec{K}_m = \vec{M} \times \hat{n}$
$\vec{M}$
$\hat{n}$
$\vec{B}$
(b)

FIG. 8. Contrasting behaviors of fields B and H for a uniformly magnetized cylinder.

is given by

$$U_m = \frac{1}{4\pi}\left(\int \frac{-\nabla'\cdot\mathbf{M}}{R}\,dV' + \int \frac{\mathbf{M}\cdot\mathbf{n}}{R}\,dS'\right), \tag{55}$$

where R is the distance between field and source points. This solution follows from the property of the operator ∇^2 used in the derivation of Eq. (22). Equation (54) can also be solved as a boundary-value problem. In fact all the apparatus of electrostatic potential theory can be brought to bear in this context. Using Eqs. (A14) and (A15), Eq. (55) can be rewritten as

$$\begin{aligned} U_m &= \frac{1}{4\pi}\int \mathbf{M}\cdot\nabla'\left(\frac{1}{R}\right)dV' \\ &= \frac{1}{4\pi}\int \frac{\mathbf{M}\cdot\hat{\mathbf{a}}_R}{R^2}\,dV'. \end{aligned} \tag{56}$$

We can obtain H in terms of the magnetization from Eqs. (53), together with (55) and (15):

$$\begin{aligned} \mathbf{H} = \frac{1}{4\pi}\Bigg[\int \frac{(-\nabla'\cdot\mathbf{M})\hat{\mathbf{a}}_R}{R^2}\,dV' \\ + \int \frac{(\mathbf{M}\cdot\mathbf{n})\hat{\mathbf{a}}_R}{R^2}\,dS'\Bigg] \end{aligned} \tag{57}$$

since the gradient in Eq. (53) is only with respect to the field point.

4. INTERFACE CONDITIONS ON THE FIELDS AND POTENTIALS

In crossing an interface between two materials having different magnetic properties, e.g., different permeabilities, some conditions are placed on the fields by the laws discussed so far. A condition on **B** can be found by integrating Eq. (20) over a pillbox-shaped volume as shown in Fig. 9(a), applying the divergence theorem, and letting the cylinder sides go to zero. We find

$$\mathbf{B}_1\cdot\mathbf{n} = \mathbf{B}_2\cdot\mathbf{n}, \tag{58}$$

i.e., the normal component of **B** is continuous across an interface. An interface condition on **H** can be found by applying Eq. (49) to the integration path shown in Fig. 9(b). By letting the sides normal to the interface approach zero, we obtain

$$\mathbf{n}\times(\mathbf{H}_2 - \mathbf{H}_1) = \mathbf{K}, \tag{59}$$

where **K** is a possible surface current. In the absence of such a surface current, the tangential component of **H** is continuous across an interface.

The tangential **B** and normal **H** components are not, in general, continuous across an interface. In fact, if materials 1 and 2 have permeabilities μ_1 and μ_2, respectively, then Eq. (58) can be used to establish a condition on the normal **H** components,

$$\mu_1\mathbf{H}_1\cdot\mathbf{n} = \mu_2\mathbf{H}_2\cdot\mathbf{n}, \tag{60}$$

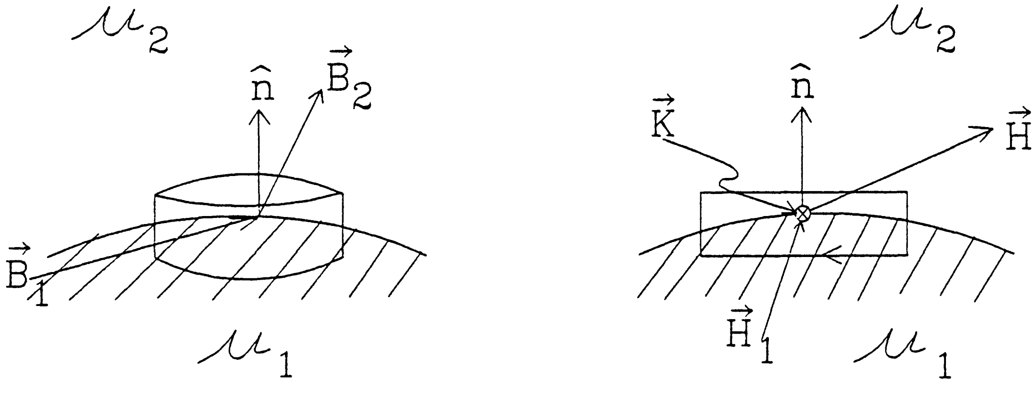

(a) (b)

FIG. 9. Interface conditions on the fields *B* and *H*.

while Eq. (59) implies, for the tangential **B** components,

$$\mathbf{n} \times (\mathbf{B}_2/\mu_2 - \mathbf{B}_1/\mu_1) = \mathbf{K}. \tag{61}$$

Thus these components are not continuous if the permeabilities are dissimilar, even in the absence of free surface currents. A closer examination shows that the origin of these discontinuities is in the material magnetization currents and surface pole densities.

The vector potential **A** is continuous, including across interfaces. Since **B** is obtained from **A** by differentiation, a discontinuity in **A** could lead to an infinite value of **B**. However, the derivatives of **A** need not to be continuous. It can be shown that across an interface separating materials of permeability μ_1 and μ_2,

$$\frac{1}{\mu_2}\frac{\partial \mathbf{A}_2}{\partial n} - \frac{1}{\mu_1}\frac{\partial \mathbf{A}_1}{\partial n} = -\mathbf{K}, \tag{62}$$

where $\partial/\partial n = \mathbf{n}\cdot\nabla$ is the normal derivative.

The magnetic scalar potential is continuous, including across interfaces. This is necessary to keep H finite and to keep the potential single valued. An interface boundary condition can be obtained from Eq. (58):

$$\mu_1 \frac{\partial U_1}{\partial n} = \mu_2 \frac{\partial U_2}{\partial n}. \tag{63}$$

5. MAGNETIC CIRCUITS

A magnetic circuit approach is often useful in obtaining approximate solutions and physical insight into magnetics problems. The possibility of doing this follows from the mathematical similarity between the equations governing electric circuits and their magnetic counterparts. The basic equations governing electric circuits, from which Kirchhoff's laws follow, are

$$\nabla\cdot\mathbf{J} = 0, \quad \oint \mathbf{E}\cdot d\mathbf{l} = \text{emf}. \tag{64}$$

The first equation implies that the algebraic sum of currents into a node is zero and the second says that the voltage drop around a circuit equals the applied electromotive force (emf). The parallel magnetic equations are

$$\nabla\cdot\mathbf{B} = 0, \quad \oint \mathbf{H}\cdot d\mathbf{l} = I. \tag{65}$$

Here **B** corresponds to **J**, **H** to **E**, and I to emf. The analogy is made even closer when we compare Ohm's law

$$\mathbf{J} = \sigma\mathbf{E} \tag{66}$$

to its magnetic parallel

$$\mathbf{B} = \mu\mathbf{H}. \tag{67}$$

The analogy between these last two equations, when applied to materials comprising actual circuits, is not exact. This is because, whereas the electrical conductivity of a conductor is many orders of magnitude higher than that of the surrounding air or insulation, the permeability of ferromagnetic materials is typically only a few orders or magnitude higher. The consequence of this is that the electric current remains confined to the electric circuit whereas the magnetic field has a greater tendency to leak beyond the confines of the circuit.

Corresponding to the electric current, $I = \int\mathbf{J}\cdot d\mathbf{S}$, we have the magnetic flux Φ,

$$\Phi = \int \mathbf{B}\cdot d\mathbf{S}. \tag{68}$$

In the mksa system, the units of flux are T m^2 or webers. If we have a conductor of uniform cross-sectional area S carrying a constant current density $J = I/S$ and having a length l over which the electric field has a constant value of $E = V/l$, where V is the voltage drop, then substituting in Eq. (66), we obtain $V = I(l/\sigma S)$, where the quantity in parentheses is recognized as the electrical resistance. Similarly, for a ferromagnetic bar having the same dimensions and uniformity, the magnetic parallel is

$$\int \mathbf{H}\cdot d\mathbf{l} = \text{mmf} = \Phi(l/\mu S), \tag{69}$$

where the quantity in parentheses is called the reluctance of the bar. The left-hand side is referred to as the magnetomotive force or mmf. In a complete circuit, it is just the total current looped.

5.1 Example

Consider the simple magnet with air gap shown in Fig. 10. Assume a uniform cross-sectional area S and no fringing. Then the flux Φ around the circuit is constant and the mmf

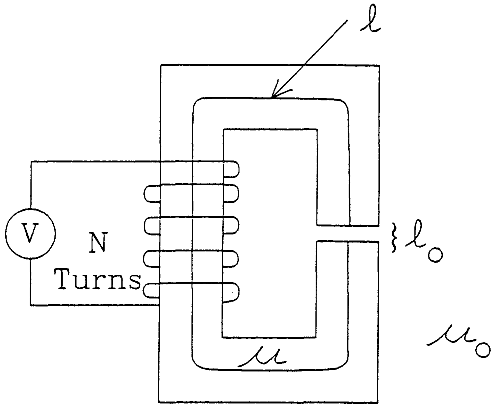

FIG. 10. An example of a magnetic circuit.

is NI since there are N turns, each turn carrying current I. The reluctance of the iron section is $l/\mu S$ and that of the air gap $l_0/\mu_0 S$. Since these reluctances are in series, we have

$$\begin{aligned} NI &= \Phi\left(\frac{l}{\mu S} + \frac{l_0}{\mu_0 S}\right) \\ &= \Phi\frac{l}{\mu S}\left(1 + \mu_r\frac{l_0}{l}\right). \end{aligned} \tag{70}$$

Thus the effective reluctance with air gap can be much larger than the reluctance without an air gap since $\mu_r l_0/l$ is often much greater than 1. This means that, for a given mmf, the flux and hence the flux density are generally much lower with an air gap than without one.

6. MAGNETOSTATIC ENERGY

6.1 Faraday's Law

Although not strictly within the field of magnetostatics, we need to state Faraday's law in order to derive expressions for the magnetostatic energy. This law, discovered in 1831 by M. Faraday, states that the voltage change or emf induced around a closed path is equal to the time rate of change of flux enclosed by the path. The sign of the emf is such that if the path were made of wire, the current would flow in a direction such as to oppose the flux change. Mathematically,

$$V = \oint \mathbf{E}\cdot d\mathbf{l} = \frac{d\Phi}{dt} = -\frac{d}{dt}\int \mathbf{B}\cdot d\mathbf{S}. \tag{71}$$

The positive direction of **B** is given by the direction a right-hand screw would advance if turned in the direction of the path integration. Stokes's theorem can be used to convert the above formula into a differential form of the law,

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}. \tag{72}$$

Thus the electric field is no longer conservative when a changing magnetic field is present.

6.2 Energy in Terms of Current Sources

To obtain the energy content associated with a collection of fixed circuits carrying steady currents with or without the presence of magnetizable matter, we assume that the currents are increased gradually in such a manner that $\nabla\cdot\mathbf{J} = 0$ holds. We also assume that there are no dissipative losses for simplicity. Then, because of the induced emf's in the circuits by Faraday's law, the batteries must supply equal and opposite emf's to keep the currents quasistationary. Thus, at any point in the process of increasing the currents, the incremental energy required to increase the currents further is given by

$$dW = -\sum_i I_i V_i dt = \sum_i I_i d\Phi_i. \tag{73}$$

The sum is over all the current loops.

An alternative expression for the flux can be obtained from Eqs. (17) and (A5):

$$\Phi = \int (\nabla \times \mathbf{A})\cdot d\mathbf{S} = \oint \mathbf{A}\cdot d\mathbf{l}. \tag{74}$$

Thus for fixed current loops, Eq. (73) can be written

$$dW = \sum_i I_i \oint \delta\mathbf{A}\cdot d\mathbf{l}_i = \sum \oint \delta\mathbf{A}\cdot I_i d\mathbf{l}_i. \tag{75}$$

We use δ here to distinguish it from the integration differential. This can be easily generalized to the case where the wire has a finite cross section by subdividing the wire into current tubes. Then each tube carries a current ΔI_i and $\Delta I_i d\mathbf{l}_i = \mathbf{J}_i dV_i$ so that Eq. (75) becomes

$$dW = \sum_i \int \delta\mathbf{A}\cdot\mathbf{J}_i dV_i = \int \delta\mathbf{A}\cdot\mathbf{J}\, dV. \tag{76}$$

The last integral above is over all space since **J** is nonzero only at the wire locations. In case **A** increases linearly with **J**, Eq. (76) can be integrated to yield

$$W = \tfrac{1}{2}\int \mathbf{A}\cdot\mathbf{J}\, dV. \tag{77}$$

6.3 Energy in Terms of Fields

We can transform Eq. (76) by means of Eqs. (48) and (A16) to obtain

$$\begin{aligned} dW &= \int [\mathbf{H}\cdot(\nabla\times\delta\mathbf{A}) - \nabla\cdot(\delta\mathbf{A}\times\mathbf{H})]\, dV \\ &= \int \mathbf{H}\cdot\delta\mathbf{B}\, dV. \end{aligned} \tag{78}$$

The last form follows on transforming the volume integral to a surface integral, using the divergence theorem, and letting the surface approach infinity, where the integral will vanish because the sources are localized. If **B** increases linearly with **H**, we obtain

$$W = \tfrac{1}{2}\int \mathbf{H}\cdot\mathbf{B}\, dV \tag{79}$$

but, in general,

$$W = \int dV \int_0^{\mathbf{B}} \mathbf{H}\cdot d\mathbf{B}, \tag{80}$$

so that $\int \mathbf{H}\cdot d\mathbf{B}$ is a magnetic energy per unit volume.

The expression for W in Eq. (77) shows that the magnetic energy is associated with the sources, while Eq. (80) indicates that it resides in the fields. Both points of view are valid, and one form or the other may be more convenient in specific calculations. If the magnetic fields are taken around a hysteresis loop as shown in Fig. 6(a), then the energy per unit volume is seen to equal the area of the loop. This magnetic energy is dissipated on each traversal of the loop. In the absence of hysteresis and other losses, the magnetic energy associated with a given state of the sources or fields is unique and may be regarded as a stored energy.

7. FLUX LINKAGES AND INDUCTANCE

If N wire loops in series link the same flux Φ, then the emf induced in the series circuit is given by

$$V = -N\frac{d\Phi}{dt} = -\frac{d\lambda}{dt}, \tag{81}$$

where $\lambda = N\Phi$ is the flux linkage. More generally, if N_1 wires link flux Φ_1, N_2 wires link flux Φ_2, etc., and all the wires are in series, then $\lambda = N_1\Phi_1 + N_2\Phi_2 + \ldots$. The flux linkage is the relevant concept when circuit emf's are desired.

For a finite cross-section wire, Eqs. (73) and (76) imply that

$$VI\, dt = -dW = -\int \delta\mathbf{A}\cdot\mathbf{J}\, dV, \tag{82}$$

where the integral is over the wire. Thus

$$\begin{aligned} V &= -\frac{1}{I}\int \frac{d\mathbf{A}}{dt}\cdot\mathbf{J}\, dV \\ &= -\int \frac{d\mathbf{A}}{dt}\cdot\left(\frac{\mathbf{J}}{I}\right) dV \\ &= -\frac{d}{dt}\left(\frac{1}{I}\int \mathbf{A}\cdot\mathbf{J}\, dV\right). \end{aligned} \tag{83}$$

In going to the last equality we have assumed that $\mathbf{J}/I$ is independent of time. Thus comparing Eq. (81) with (83), we obtain a general expression for the flux linkage for a finite cross-section wire:

$$\lambda = \frac{1}{I}\int \mathbf{A}\cdot\mathbf{J}\, dV. \tag{84}$$

Here **A** is the vector potential due to all sources and the integral is over the volume of the wire. For a thin wire circuit, this reduces to $\lambda = \int \mathbf{A}\cdot d\mathbf{l}$. In general, for a collection of windings, Eq. (73) is replaced by

$$dW = \sum_i I_i d\lambda_i, \tag{85}$$

where $d\lambda_i = (1/I)\int \delta\mathbf{A}\cdot\mathbf{J}_i dV_i$.

Often we are interested in the details of how different circuits are coupled to each other, and so we distinguish the flux linkages produced by one circuit from that of another. Then we write $d\lambda_i = \Sigma_j d\lambda_{ij}$, where $d\lambda_{ij}$ is the differential flux produced by circuit j and linking circuit i. Since for rigid circuits, a change in λ_{ij} is, to a good approximation, only due to a change in the current in circuit j, we have

$$d\lambda_i = \sum_j \frac{d\lambda_{ij}}{dI_j} dI_j = \sum_j M_{ij} dI_j, \qquad (86)$$

where we have defined a (coefficient of) mutual inductance M_{ij}. The coefficient M_{ii} is called the self-inductance (coefficient) and is generally denoted by L_i. These inductances have units of webers/ampere or henrys. The sign is determined by the chosen direction of positive current flow in the two circuits. Substituting Eq. (86) into Eq. (81) yields

$$V_i = -\sum_j M_{ij} \frac{dI_j}{dt} = -L_i \frac{dI_i}{dt} - \sum_{j \neq i} M_{ij} \frac{dI_j}{dt}. \qquad (87)$$

This equation gives only the emf in circuit i associated with the magnetic fields and does not include resistances, capacitors, etc.

7.1 Energy in Terms of Inductances

Substituting Eq. (86) into (85), the differential magnetic energy can be expressed as

$$dW = \sum_i \sum_j M_{ij} I_i dI_j. \qquad (88)$$

Before we can integrate this, we need to look more closely at the M_{ij}. From Eqs. (84) and (86),

$$M_{ij} = \frac{1}{I_i} \int \frac{d\mathbf{A}_j}{dI_j} \cdot \mathbf{J}_i dV_i. \qquad (89)$$

If we assume that $\mathbf{J}/I$ depends only on position, then using Eq. (18) for $\mathbf{A}$ gives

$$\mathbf{A} = I\left(\frac{\mu_0}{4\pi I} \int \frac{\mathbf{J}\, dV}{R}\right), \qquad (90)$$

where the quantity in parentheses is independent of current. Therefore $d\mathbf{A}/dI = \mathbf{A}/I$ and Eq. (89) becomes

$$\begin{aligned} M_{ij} &= \frac{1}{I_i I_j} \int \mathbf{A}_j \cdot \mathbf{J}_i dV_i \\ &= \frac{\mu_0}{4\pi I_i I_j} \int \int \frac{\mathbf{J}_i \cdot \mathbf{J}_j}{R_{ij}} dV_i dV_j, \end{aligned} \qquad (91)$$

where R_{ij} is the distance between the positions of the two current densities. It can be seen here that $M_{ij} = M_{ji}$. This is a general result for linear systems. If the space between the circuits were filled with a material of linear permeability μ, Eq. (91) would hold with μ_0 replaced by μ.

Assuming these properties for M_{ij}, Eq. (88) becomes

$$\begin{aligned} dW &= \sum_i L_i I_i dI_i + \sum_{i<j} M_{ij}(I_i dI_j + I_j dI_i) \\ &= \tfrac{1}{2} \sum_i L_i d(I_i)^2 + \sum_{i<j} M_{ij} d(I_i I_j), \end{aligned} \qquad (92)$$

so that integration produces

$$W = \tfrac{1}{2} \sum_i L_i I_i^2 + \sum_{i<j} M_{ij} I_i I_j. \qquad (93)$$

7.2 Example

We calculate the self-inductance of the torus of Fig. 3(b). Assume that it is uniformly wound with N turns of thin wire carrying current I and that it is filled with magnetic material of constant permeability μ. From Eq. (26), with μ replacing μ_0, we find approximately

$$\Phi = B_t S = \mu NIS/2\pi r_c, \qquad (94)$$

where S is the cross-sectional area and r_c is the distance from the axis to the center of the circular cross-sectional area. Hence the self-inductance is

$$L = \frac{d\lambda}{dI} = \frac{\lambda}{I} = \frac{N\Phi}{I} = \frac{\mu N^2 S}{2\pi r_c}, \qquad (95)$$

and we see that L is a geometric quantity apart from the factor μ. The approximations used here amount to assuming that the radius of the small circle is much less than r_c and that there are many fine wire turns, uniformly wound.

8. MAGNETIC FORCES AND TORQUES

8.1 Force and Torque on Current Distributions

The magnetic force on a current distribution is obtained by integrating Eq. (13):

$$\mathbf{F} = \int \mathbf{J} \times \mathbf{B}\, dV \quad \text{or} \quad \mathbf{F} = \int I d\mathbf{l} \times \mathbf{B}, \qquad (96)$$

where the second equation refers to filamentary conductors. Similarly, the magnetic torque about some point O is given by

$$\mathbf{T} = \int \mathbf{r} \times (\mathbf{J} \times \mathbf{B})\, dV \quad \text{or}$$
$$\mathbf{T} = \int \mathbf{r} \times (I\, d\mathbf{l} \times \mathbf{B}), \tag{97}$$

where $\mathbf{r}$ is measured from the point O.

8.1.1 Example The force between two infinitely long, thin, parallel wires a distance h apart and carrying currents I_1 and I_2 is found from Eq. (96) and the previously obtained value of B produced by a single long wire, $B_1 = \mu_0 I_1/2\pi h$, where 1 labels the first wire. From Fig. 2(b), we see that this field is everywhere perpendicular to the second wire and constant along its length so that the force per unit length is given by $F/L = \mu_0 I_1 I_2/2\pi h$. The sense of the cross product is such that if the currents are flowing in the same direction the force is attractive; otherwise it is repulsive.

8.2 Force and Torque on Localized Current Distributions

We assume that the localized current distribution is in an external magnetic field that varies slowly over the distribution. Then the field can be written

$$\mathbf{B}(\mathbf{r}) = \mathbf{B}(0) + (\mathbf{r}\cdot\nabla)\mathbf{B}(0) + \ldots, \tag{98}$$

where the origin is taken inside or near the current distribution, and in the second term on the right hand side the evaluation at $\mathbf{r} = 0$ is performed after the gradient operation. Substituting into Eq. (96), we obtain approximately

$$\mathbf{F} = \left(\int \mathbf{J}\, dV'\right) \times \mathbf{B}(0) + \int \mathbf{J} \times (\mathbf{r}'\cdot\nabla)\mathbf{B}(0)\, dV'. \tag{99}$$

The first term vanishes because the currents are localized and $\nabla\cdot\mathbf{J} = 0$. The gradient operator only operates on $\mathbf{B}$ in the second integral. After a series of manipulations similar to those leading to Eq. (35), the force can be transformed to the form

$$\mathbf{F} = (\mathbf{m}\cdot\nabla)\mathbf{B}, \tag{100}$$

where $\mathbf{m}$ is the magnetic dipole moment given by Eq. (36). Thus a magnetic dipole experiences a force if the field has a gradient.

The torque on a localized current distribution is found by substituting Eq. (98) into Eq. (97). Here the constant term in Eq. (98) contributes, so that to lowest order

$$\mathbf{T} = \int \mathbf{r}' \times [\mathbf{J} \times \mathbf{B}(0)]\, dV'. \tag{101}$$

This likewise can be transformed, by means of identities and similar procedures to those used to derive Eq. (35), to the form

$$\mathbf{T} = \mathbf{m} \times \mathbf{B}. \tag{102}$$

Thus a magnetic moment experiences a torque in a uniform magnetic field. An example is a compass needle in the Earth's magnetic field.

8.3 Maxwell Stress Tensor

In general, when both free currents and magnetization are present, we expect the force to be given by

$$\mathbf{F} = \int (\mathbf{J} + \nabla \times \mathbf{M}) \times \mathbf{B}\, dV = \frac{1}{\mu_0}\int (\nabla \times \mathbf{B}) \times \mathbf{B}\, dV, \tag{103}$$

where the second equality follows from Eqs. (47) and (48). The integral is over the volume of the object on which the force is acting. By means of Eq. (A17) the above integrand can be written

$$(\nabla \times \mathbf{B}) \times \mathbf{B} = (\mathbf{B}\cdot\nabla)\mathbf{B} - \tfrac{1}{2}\nabla B^2. \tag{104}$$

By adding a term $\mathbf{B}(\nabla\cdot\mathbf{B})$, which is zero, to the right-hand side, Eq. (104) can be written as the divergence of a tensor so that Eq. (103) becomes

$$\mathbf{F} = \int \nabla\cdot\mathsf{T}\, dV, \tag{105}$$

where $\mathsf{T} = (1/\mu_0)[\mathbf{BB} - (\tfrac{1}{2})\mathsf{I}B^2]$ is called the Maxwell stress tensor. We are using dyadic notation, where I, the unit dyadic, is given in Cartesian coordinates by $\mathsf{I} = \mathbf{ii} + \mathbf{jj} + \mathbf{kk}$, where $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ are the Cartesian unit vectors. The tensor components are given by $T_{ij} =$

$\mathbf{i}\cdot\mathbf{T}\cdot\mathbf{j}$, so that, in component form

$$T_{ij} = (1/\mu_0)(B_iB_j - \tfrac{1}{2}\,\delta_{ij}B^2), \tag{106}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

Inserting Eq. (104) into (103) and making use of Eqs. (A6) and (A8), we can express **F** as a surface integral:

$$\mathbf{F} = \frac{1}{\mu_0}\oint [\mathbf{B}(\mathbf{B}\cdot\mathbf{n}) - \tfrac{1}{2}B^2\mathbf{n}]dS. \tag{107}$$

To avoid ambiguity, the integral should be taken over a surface slightly outside the object. If we decompose **B** into a normal and tangential component at the surface, $\mathbf{B} = B_n\mathbf{n} + B_t\mathbf{t}$, the last equation becomes

$$\mathbf{F} = \frac{1}{\mu_0}\oint [\tfrac{1}{2}(B_n^2 - B_t^2)\mathbf{n} + B_nB_t\mathbf{t}]dS. \tag{108}$$

8.3.1 Examples From Eq. (108), we see that when B is normal to a surface, so that $B_t = 0$, the surface force density is $\mathbf{F}/S = (B_n^2/2\mu_0)\mathbf{n}$ so that the force is directed along the outward normal to the surface. This means the object is being pulled on, i.e., the force is attractive. An illustration of this is shown in Fig. 11 where an iron slab is near a lifting magnet. The field B is nearly normal to the upper surface of the bar under the magnet poles and nearly vanishes along the other surfaces. Since the force does not depend on the sign of B_n, both poles of the magnet exert the same force on the magnet. The force is upward or attractive since this is the direction of the outward normal to the top surface surrounding the slab.

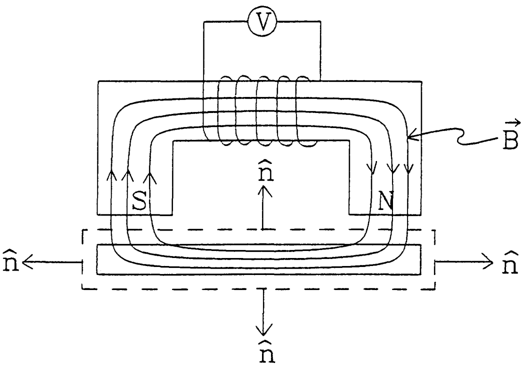

FIG. 11. An electromagnet exerting a force on a ferromagnetic slab.

When the magnetic field is tangential to a surface so that $B_n = 0$, then the surface force density is $\mathbf{F}/S = (-B_t^2/2\mu_0)\mathbf{n}$ so that it is directed along the inward normal to the surface, regardless of the sign of B_t. A long solenoid illustrates this situation since B is tangential along the inner surface but zero outside the solenoid. If we draw an imaginary surface around the solenoid wires, the outward normal to the inner part of the surface points toward the axis. Therefore the surface force pushes outward on the inner surface of the wires.

8.4 Force and Torque by Energy Methods

Since we are interested in static or nearly static magnetic forces, we assume that displacements occur so slowly that no kinetic energy is produced. We also ignore losses and capacitive energy. Then the principle of conservation of energy says that for a collection of circuits, with or without the presence of magnetizable material, the energy supplied by the sources equals the increase in magnetic energy plus the mechanical work done by the forces. The energy supplied by the sources is given by Eq. (85), and so the energy principle states

$$\sum_i I_i d\lambda_i = dW_m + F_x dx. \tag{109}$$

We have isolated a single force acting in a specified direction x. To find torques, simply substitute $T_\theta d\theta$ for the force term. Since we are ignoring hysteresis, the magnetic energy W_m is a state function of the independent variables of the system. If we choose the independent variables to be the λ_i and x, so that $W_m = W_m(\{\lambda_i\},x)$, where $\{\lambda_i\}$ means the set $\lambda_1,\lambda_2,\ldots$, then

$$dW_m = \sum_i \left(\frac{\partial W_m}{\partial \lambda_i}\right)_{x,\{\lambda_{j\neq i}\}} d\lambda_i + \left(\frac{\partial W_m}{\partial x}\right)_{\{\lambda_i\}} dx, \tag{110}$$

where the subscript notation indicates the variables held constant in the partial differentiation. Substituting into Eq. (109), we obtain

$$\sum_i \left[I_i - \left(\frac{\partial W_m}{\partial \lambda_i}\right)_{x,\{\lambda_{j\neq i}\}} \right] d\lambda_i - \left[\left(\frac{\partial W_m}{\partial x}\right)_{\{\lambda_i\}} + F_x \right] dx = 0. \tag{111}$$

Since the λ_i and x are independent, all the coefficients of the $d\lambda_i$ and dx above must vanish; in particular,

$$F_x = -\left(\frac{\partial W_m}{\partial x}\right)_{\{\lambda_i\}}. \tag{112}$$

Thus, at constant flux linkages, the force is the negative gradient of the magnetic energy.

If the I_i and x are taken as independent variables so that $W_m = W_m(\{I_i\},x)$, then define a magnetic coenergy, $W'_m(\{I_i\},x)$, by

$$W'_m = \sum_i I_i \lambda_i - W_m. \tag{113}$$

Then

$$dW'_m = \sum_i I_i d\lambda_i + \sum_i \lambda_i dI_i - dW_m \tag{114}$$

with

$$dW'_m = \sum_i \left(\frac{\partial W'_m}{\partial I_i}\right)_{x,\{I_{j\neq i}\}} dI_i + \left(\frac{\partial W'_m}{\partial x}\right)_{\{I_j\}} dx. \tag{115}$$

Substituting Eqs. (114) and (115) into (109), we obtain

$$\sum_i \left[\lambda_i - \left(\frac{\partial W'_m}{\partial I_i}\right)_{x,\{I_{j\neq i}\}} \right] dI_i + \left[F_x - \left(\frac{\partial W'_m}{\partial x}\right)_{\{I_i\}} \right] dx = 0. \tag{116}$$

By the independence of the I_i and x, we have

$$F_x = \left(\frac{\partial W'_m}{\partial x}\right)_{\{I_i\}}. \tag{117}$$

Thus the force at constant current is given by the positive gradient of the magnetic coenergy.

The force calculated by means of the magnetic energy or coenergy is the same force. Mathematical convenience dictates which of these methods is used to obtain the force. For linear systems, both methods are equivalent

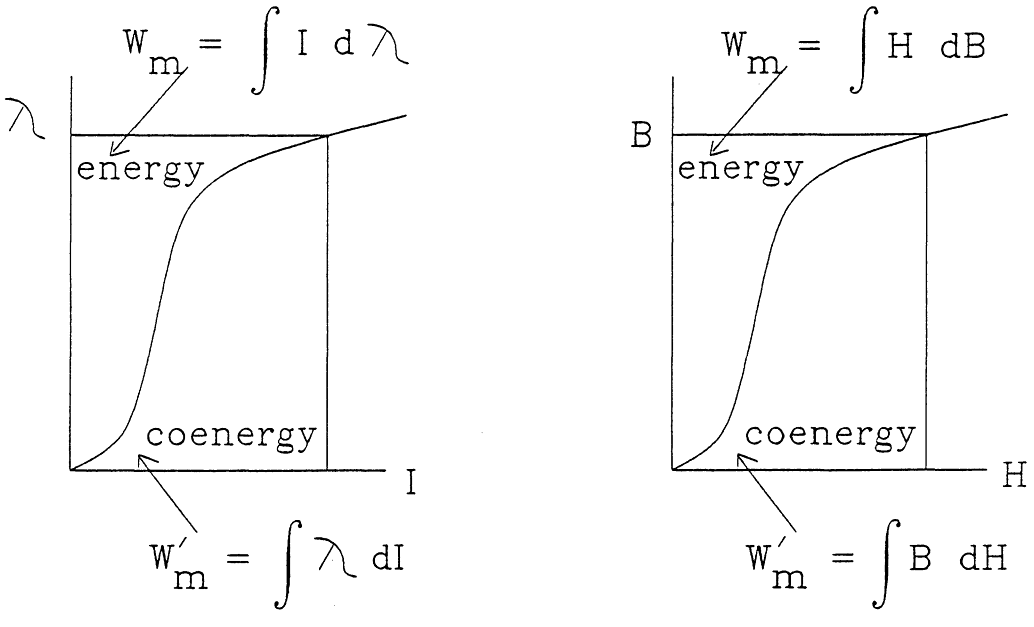

(a)

(b)

FIG. 12. Graphical illustration of energy and coenergy in terms of (a) λ–I and (b) b–H variables.

since, in this case, the energy and coenergy are equal. For a single circuit, the energy and coenergy are shown graphically in Fig. 12. The magnetic energy can be obtained from the current–flux relationship as shown in Fig. 12(a) or via the fields as shown in Fig. 12(b). When these relationships are linear, the equivalence of the energy and coenergy is apparent.

8.4.1 Example As an illustration of these ideas, consider a system of coupled linear inductors with magnetic energy given by Eq. (93). If the forces between inductors i and j is desired and x is taken to be their relative separation, then only M_{ij} depends on x, so that

$$F_x = \left(\frac{\partial W'_m}{\partial x}\right)_{\{I_i\}} = I_i I_j \frac{\partial M_{ij}}{\partial x} \tag{118}$$

since the magnetic energy and coenergy are equal.

APPENDIX: MATHEMATICAL NOTATION AND USEFUL THEOREMS

Vector or scalar fields are assumed to have the required continuity or differentiability to make the mathematical manipulations performed on them valid. In the following **A**, **B**, **C** are general vector fields; Φ is a general scalar field; V is a volume and dV a volume differential; S is a surface and dS a surface differential; **n** is a surface unit outward normal vector; $d\mathbf{S} = \mathbf{n}\, dS$ is a vectorial surface differential; $\boldsymbol{l}$ is a path length and $d\mathbf{l}$ is a vectorial differential path length; $\oint$ denotes an integral over a closed path or closed surface; $\hat{\mathbf{a}}$ is a general unit vector; **i**, **j**, and **k** are unit vectors pointing in the x, y, and z directions, respectively, in a Cartesian coordinate system; ∇ is the gradient operator, which in Cartesian coordinates is given by

$$\nabla = \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}. \tag{A1}$$

Dot product:

$$\mathbf{A}\cdot\mathbf{B} = |\mathbf{A}|\,|\mathbf{B}|\cos\theta \tag{A2}$$

where θ is the angle between the two vectors.

Cross product:

$$\mathbf{A}\times\mathbf{B} = |\mathbf{A}|\,|\mathbf{B}|\sin\theta\mathbf{n}, \tag{A3}$$

where θ is the angle measured from **A** to **B** and **n** is a unit vector perpendicular to **A** and **B** and directed in the sense that a right-hand screw would advance in being turned from **A** to **B**.

Divergence theorem:

$$\int_V \nabla\cdot\mathbf{A}\, dV = \oint_S \mathbf{A}\cdot\mathbf{n}dS, \tag{A4}$$

where S is the closed surface enclosing volume V and **n** is its outward normal.

Stokes's theorem:

$$\int_S (\nabla\times\mathbf{A})\cdot\mathbf{n}\, dS = \oint_{\mathscr{C}} \mathbf{A}\cdot d\mathbf{l}, \tag{A5}$$

where $\mathscr{C}$ is the closed path enclosing surface S. If a right-hand screw is turned in the direction of $d\mathbf{l}$, it will advance in the direction of **n**.

Other theorems:

$$\int \nabla\Phi\, dV = \oint \Phi\mathbf{n}\, dS, \tag{A6}$$

$$\int \nabla\times\mathbf{A}\, dV = \oint \mathbf{n}\times\mathbf{A}\, dS, \tag{A7}$$

$$\int [(\mathbf{A}\cdot\nabla)\mathbf{B} + \mathbf{B}(\nabla\cdot\mathbf{A})]dV = \oint (\mathbf{A}\cdot\mathbf{n})\mathbf{B}\, dS, \tag{A8}$$

$$\nabla\cdot(\nabla\times\mathbf{A}) = 0, \tag{A9}$$

$$\nabla\times\nabla\Phi = \mathbf{0}, \tag{A10}$$

$$\nabla\times(\nabla\times\mathbf{A}) = \nabla(\nabla\cdot\mathbf{A}) - \nabla^2\mathbf{A}, \tag{A11}$$

$$\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C}) = \mathbf{C}\cdot(\mathbf{A}\times\mathbf{B}) = \mathbf{B}\cdot(\mathbf{C}\times\mathbf{A}), \tag{A12}$$

$$\mathbf{A}\times(\mathbf{B}\times\mathbf{C}) = (\mathbf{A}\cdot\mathbf{C})\mathbf{B} - (\mathbf{A}\cdot\mathbf{B})\mathbf{C}, \tag{A13}$$

$$\nabla\cdot(\Phi\mathbf{A}) = \mathbf{A}\cdot\nabla\Phi + \Phi\nabla\cdot\mathbf{A}, \tag{A14}$$

$$\nabla\times(\Phi\mathbf{A}) = \nabla\Phi\times\mathbf{A} + \Phi\nabla\times\mathbf{A}, \tag{A15}$$

$$\nabla\cdot(\mathbf{A}\times\mathbf{B}) = \mathbf{B}\cdot(\nabla\times\mathbf{A}) - \mathbf{A}\cdot(\nabla\times\mathbf{B}), \tag{A16}$$

$$\nabla(\mathbf{A}\cdot\mathbf{B}) = (\mathbf{A}\cdot\nabla)\mathbf{B} + (\mathbf{B}\cdot\nabla)\mathbf{A} + \mathbf{A} \times (\nabla\times\mathbf{B}) + \mathbf{B}\times(\nabla\times\mathbf{A}), \qquad \text{(A17)}$$

$$\nabla\times(\mathbf{A}\times\mathbf{B}) = \mathbf{A}(\nabla\cdot\mathbf{B}) - \mathbf{B}(\nabla\cdot\mathbf{A}) + (\mathbf{B}\cdot\nabla)\mathbf{A} - (\mathbf{A}\cdot\nabla)\mathbf{B}. \qquad \text{(A18)}$$

Further Reading

Brown, W. F., Jr. (1962), *Magnetostatic Principles in Ferromagnetism*, Amsterdam: North-Holland.

Brown, W. F., Jr. (1963), *Micromagnetics*, New York: Interscience.

Durney, C. H., Johnson, C. C. (1969), *Introduction to Modern Electromagnetics*, New York: McGraw-Hill.

Griffiths, D. J. (1981), *Introduction to Electrodynamics*, Englewood Cliffs, NJ: Prentice-Hall.

Grover, F. W. (1973), *Inductance Calculations*, New York: Dover (special edition prepared for the Instrument Society of America).

Jackson, J. D. (1962), *Classical Electrodynamics*, New York: Wiley.

Landau, L. D., Lifshitz, E. M. (1984), *Electrodynamics of Continuous Media*, 2nd ed. New York: Pergamon (revised and enlarged by E. M. Lifshitz and L. P. Pitaevskii).

Lorrain, P., Corson, D. R. (1970), *Electromagnetic Fields and Waves*, San Francisco: Freeman.

Maxwell, J. C. (1891), *Treatise on Electricity and Magnetism*, 3rd ed. 2 vols., Oxford: Clarendon [reprinted (1954), New York: Dover.]

Mocn, F. C. (1984), *Magneto-Solid Mechanics*, New York: Wiley.

Panofsky, W. K. H., Phillips, M. (1955), *Classical Electricity and Magnetism*, Reading, MA: Addison-Wesley.

Paris, D. T., Hurd, F. K. (1969), *Basic Electromagnetic Theory*, New York: McGraw-Hill.

Peck, E. R. (1953), *Electricity and Magnetism*, New York: McGraw-Hill.

Plonsey, R., Collin, R. E. (1961), *Principles and Applications of Electromagnetic Fields*, New York: McGraw-Hill.

Pugh, E. M., Pugh, E. W. (1962), *Principles of Electricity and Magnetism*, Reading, MA: Addison-Wesley.

Reitz, J. R., Milford, F. J. (1967), *Foundations of Electromagnetic Theory*, 2nd ed., Reading, MA: Addison-Wesley.

Rogers, W. E. (1954), *Introduction to Electric Fields*, New York: McGraw-Hill.

Scott, W. T. (1959), *The Physics of Electricity and Magnetism*, New York: Wiley.

Seely, S. (1958), *Introduction to Electromagnetic Fields*, New York: McGraw-Hill.

Smythe, W. R. (1968), *Static and Dynamic Electricity*, New York: McGraw-Hill.

Stratton, J. A. (1941), *Electromagnetic Theory*, New York: McGraw-Hill.

Whittaker, E. T. (1951, 1953), *A History of the Theories of Aether and Electricity*, 2 vols., London: Nelson.

MAGNETOSTRICTIVE MATERIALS AND DEVICES

J. B. Restorff, *Naval Surface Warfare Center, Dahlgren Division, White Oak Laboratory, Silver Spring, Maryland, U.S.A.*

Introduction 229
1. **Effects** 230
1.1 Linear Magnetostriction 230
1.2 ΔE Effect 231
1.3 Magnetovolume Effect 232
1.4 Wertheim Effect 232
1.5 Wiedemann Effect 232
1.6 Garshelis Effect 233
1.7 Magnetoacoustic Effect 233
2. **Materials** 233
2.1 Linear Magnetostriction Theory 233
2.1.1 Magnetostriction at Magnetic Saturation 233
2.1.2 Linearized Magnetostriction below Saturation 233
2.2 Magnetic Anisotropy 234
2.3 Important Material Properties 234
2.4 Useful Materials 235
2.4.1 Crystalline Materials 235
2.4.2 Amorphous Materials 235
3. **Devices** 235
3.1 Sensor Applications 235
3.1.1 Inventory Control 236
3.1.2 Torque Sensors 236
3.1.3 Pressure Sensors 237
3.1.4 Anisotropy Sensors 237
3.1.5 Villari Bridges 237
3.1.6 Magnetostrictive/Piezoelectric Combinations 237
3.1.7 Position Sensing 238
3.1.8 High-Sensitivity Strain Gauges 238
3.1.9 Accelerometers 238
3.1.10 Engine-Knock Sensors 239
3.1.11 Nondestructive Evaluation 239
3.1.12 Magnetometers 239
3.1.13 Potential Applications 239
3.2 Actuator Applications 239
3.2.1 Sonar Transmitters 240
3.2.2 Active Vibration Control 240
3.2.3 Deformable Mirrors 240
3.2.4 Magnetostrictive/Piezoelectric Combinations 240
3.2.5 Geological Tomography 241
3.2.6 Potential Applications 241
3.2.6.1 Driven-Tooth Bone-Conduction Hearing Aid 241
3.2.6.2 Smart (Self-Sensing) Actuators 241
3.2.6.3 Magnetostrictive Motors 241
3.3 Other Applications 242
3.3.1 Variable Delay Line 242
3.3.2 Passive Damping 242
3.3.3 Current Threshold Sensors 242
3.3.4 Record/Playback Heads 242
3.3.5 Vibratory Stress Relief 243
Glossary 243
Works Cited 243
Further Reading 244

INTRODUCTION

When some materials are placed in a varying magnetic field (which alters the material's magnetic state), a change in the physical dimensions of these materials occurs. This effect is called *magnetostriction*, and it is important for device applications because it represents an avenue for converting magnetic energy into physical motion for actuator applications. In addition, the magnetostrictive effect is reversible; changes in a material's applied stress generate magnetic energy, i.e., the material can be used as a sensor. The physical mechanism that couples the material's magnetic state to its physical dimension is the change in orientation of spatially asymmetric magnetic ions. Figure 1 shows a mechanical analog of this. In (a), there is no magnetic field and the asymmetric ions are turned to their "short" direction by the weight of the mass. In (b), a magnetic field

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

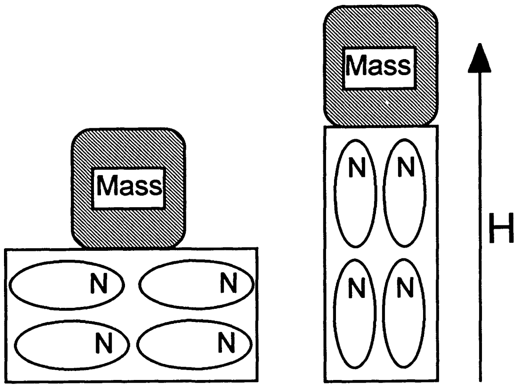

(a) (b)

FIG. 1. A schematic showing how the rotation of asymmetric atomic ions changes the length of a material. H is the applied field. (a) No magnetic field, (b) magnetic field.

is applied and pulls the ions toward their long direction.

Historically, magnetostriction was discovered by Joule in 1847 when he measured the change in length of an iron rod when it was magnetized. During World War II, nickel was used in the magnetostrictive sonar devices on warships. In the 1960s, measurements on Tb and Dy revealed magnetostrictive strains of nearly 1%. Unfortunately, these strains occur only at low temperatures. Attempts to obtain large magnetostrictions at room temperature culminated in the synthesis of $TbFe_2$ (Terfenol) compounds in the early 1970s. Dy was later added to the recipe to improve performance at low applied fields, and the result is now commercially available as Terfenol-D ($Tb_{0.27}Dy_{0.73}Fe_{1.95}$).

Magnetostrictive materials have applications as both sensors and actuators. The most important traits differ, however, depending upon the application. The crucial property for sensors is sensitivity, i.e., how much of the input energy can be converted into output energy. Transverse-field–annealed iron-based amorphous ribbons offer the highest sensitivity for sensor applications. For actuators, the magnitude of the strain and the available force are the most important properties.

This article first describes various magnetostrictive effects in Sec. 1, then looks at the commonly used materials and their properties in Sec. 2, and finally discusses current and planned magnetostrictive devices in Sec. 3. Electromagnetic units and symbols always offer an opportunity for confusion. In this article, we follow IEEE Standard 319-1990 (IEEE, 1990), which is based on the SI system of units.

1. EFFECTS

There is an entire menagerie of magnetostrictive effects. Almost all of them can be used in either sensor or actuator devices. The most commonly used effect is linear magnetostriction.

1.1 Linear Magnetostriction

A magnetic material is composed of domains, small regions in which all of the atomic moments are aligned. In an unmagnetized state, the domains point in directions that minimize the total energy. If a large enough magnetic field is applied, the minimum-energy configuration is one in which all of the magnetic moments line up in the same direction as the applied field; under these conditions, the material is said to be *saturated.* In order to get some idea of the size of the magnetostriction effect, let us consider one important type of magnetostriction. The sample is placed in a saturating magnetic field whose direction is rotating. As the direction of the field changes, the atomic moments rotate to keep themselves aligned with the applied field. The rotation of the moments changes the sample dimensions according to the formula $\Delta l/l = \frac{3}{2}\lambda_s(\cos^2\theta - \frac{1}{3})$, where $\Delta l/l$ is the fractional change in length, λ_s is the saturation magnetostriction coefficient, and θ is the angle between the field and the measurement direction. Values for some materials of interest are shown in Table 1. Positive values of λ_s indicate that the material lengthens when the applied magnetic field is increased; negative values indicate that the material shortens. The change in sample dimensions in the direction of the applied magnetic field is called Joule magnetostriction. The inverse effect, a change in the magnetic state of a material when applied stress is changed, is called the Villari effect.

In actual use, a geometry like that shown in Fig. 2 is normally used. A rod of the material is wrapped with a coil. As the coil is energized, the material changes its dimen-

Table 1. Saturation magnetostrictions.

Material	$10^6\lambda_s$	Temperature (K)	Material	$10^6\lambda_s$	Temperature (K)
Fe	−9[a]	300	$TbCo_3$	65[f]	300
Ni	−35[a]	300	$SmFe_2$	−1560[f]	300
Co	−60[a]	300	$TbFe_2$	1753[f]	300
60% Co-40% Fe	68[b]	300	Fe_3O_4	40[g]	300
60% Ni-40% Fe	25[c]	300	Metglas 2605SC	40[h]	300
Tb	3000[d]	77	Terfenol-D	1600[i]	300
Dy	6000[e]	77	$Tb_{0.6}Dy_{0.4}$	6000[j]	77

---

[a] Bozorth (1951).
[b] Weil and Reichel (1954).
[c] Masiyama (1931).
[d] Rhyne and Legvold (1965).
[e] Clark *et al.* (1965).
[f] Clark (1974).
[g] Smit and Wijn (1959).
[h] Gibbs *et al.* (1988), transverse field annealed $Fe_{81}B_{13.5}Si_{3.5}C_2$ amorphous ribbons.
[i] Clark (1980).
[j] Clark *et al.* (1992).

sions. Figure 3 shows the magnetostriction of a Terfenol-D rod vs applied field at three different compressive stresses. The magnetization M is intimately linked with the magnetostriction, since both are caused by the rotation of the magnetic moments. Figure 4 shows the magnetization corresponding to the magnetostriction in Fig. 3. Note that the strain does not depend on the direction of the applied field.

1.2 ΔE Effect

In addition to the conventional stress-strain elasticity arising from atomic forces, magnetostrictive materials can also change their length when stress is applied by rotating their magnetic moments. This is known as the ΔE effect, where $\Delta E = (E_s - E_0)/E_0$, where E_0 is the Young's modulus at high field and E_s is the minimum value of the modulus. The compression of the material due to moment rotation is in addition to the nonmagnetic elastic compression, so that the material becomes softer when the moments are free to move. Figure 5 shows the ΔE effect in $Tb_{0.67}Dy_{0.33}$. The modulus decreases by about a factor of 7 as the magnetic field is varied. At high applied magnetic fields, the moments are virtually locked in place and the

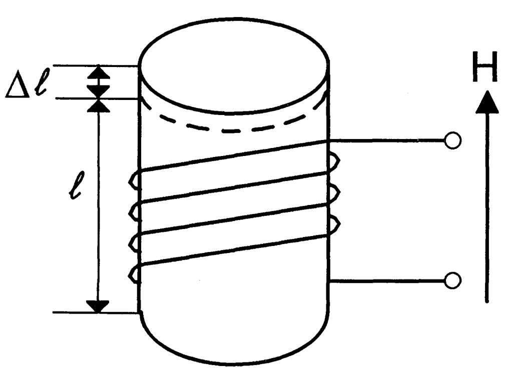

FIG. 2. Joule magnetostriction. An applied axial field H causes the rod to change its length by Δl; the magnetostriction is given by $\Delta l/l$.

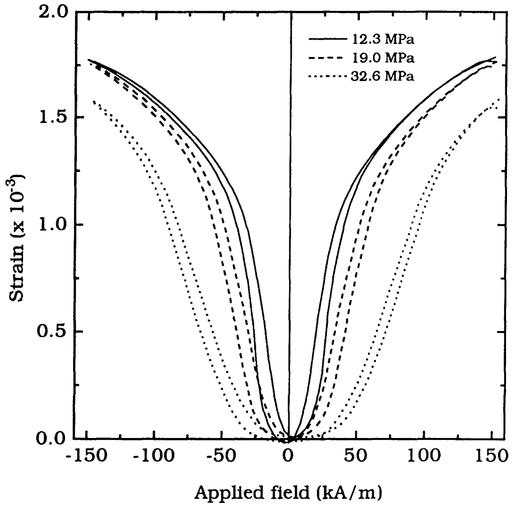

FIG. 3. Strain of a Terfenol-D rod as a function of applied field at various applied stresses.

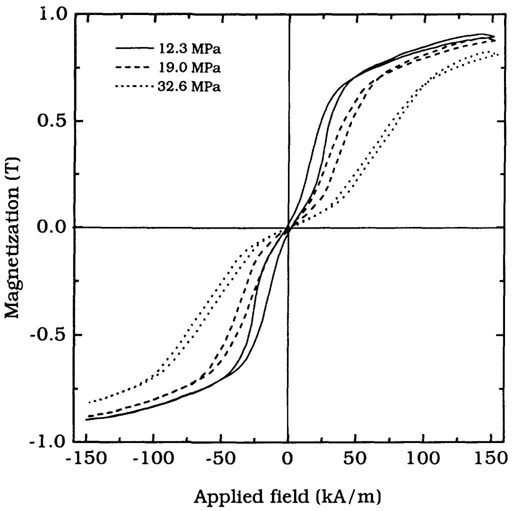

FIG. 4. Magnetization of a Terfenol-D rod as a function of applied field at various stresses. In this plot, *M* has been multiplied by μ_0 and plotted in units of teslas. Technically, this is called the intrinsic magnetic flux density *J*.

modulus becomes larger. The ΔE effect is about 0.06 in Ni, 2 in Terfenol-D, and 10 in transverse-field–annealed amorphous Fe_{81}-$B_{13.5}Si_{3.5}C_2$ ribbons.

1.3 Magnetovolume Effect

The magnetovolume effect is a change in volume caused by a change in the applied magnetic field which in turn alters the magnetization. The largest magnetovolume effects occur near the Curie temperature where $\Delta M/\Delta T$ is large. In cubic and isotropic materials, the change in volume caused by rotation of the magnetization is zero. The magnetovolume effect is also called the Barrett effect, and the inverse, a change in magnetic state caused by a change in the volume, is called the Nagaoka–Hondo effect. In the $Ni_{0.36}Fe_{0.64}$ alloy Invar, the magnetovolume effect compensates the usual thermal expansion, giving a material with a nearly zero thermal expansion coefficient at room temperature (Bozorth, 1951).

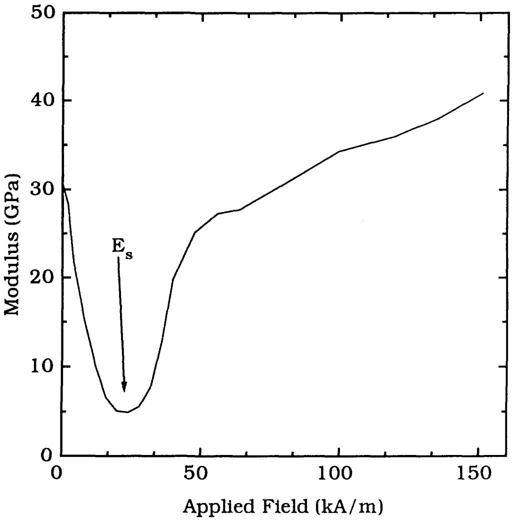

FIG. 5. ΔE effect in a $Tb_{0.67}Dy_{0.33}$ rod at 77 K.

1.4 Wertheim Effect

In a ferromagnetic rod placed in an axial magnetic field, the moments are predominantly lined up axially. When the rod is twisted, some of the moments become helically oriented, and a transient voltage is generated across the ends of the rod. This is called the Wertheim effect (see Fig. 6). In the inverse Wertheim effect, a voltage applied to the rod ends causes a twist in the rod.

1.5 Wiedemann Effect

A current carrying ferromagnetic wire has the preponderance of its moments lined up in the circumferential direction, parallel to the magnetic field. When an axial magnetic field is applied, some of the moments align themselves in a helical fashion, twisting the wire. This is called the Wiedemann effect (see Fig. 7). The inverse Wiedemann effect is the

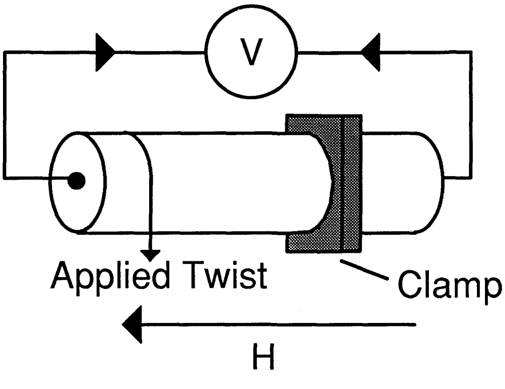

FIG. 6. Wertheim effect.

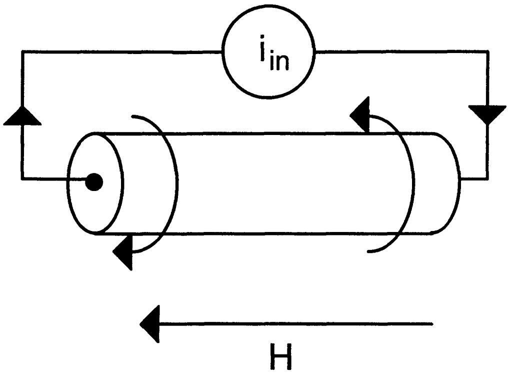

FIG. 7. Wiedemann effect.

change in axial magnetization of a current carrying wire when it is twisted.

1.6 Garshelis Effect

The magnetization of a spring formed of ferromagnetic wire changes with the length of the spring; this is called the Garshelis effect. Depending on whether the magnetization is due to currents in the wire or is a remanent magnetization, the Garshelis effect can be an extrapolation of the Wertheim or inverse Wiedemann effect.

1.7 Magnetoacoustic Effect

The change in dimensions of a magnetostrictive material can generate acoustic energy. An example of this effect is hum that electrical transformers give off or a "click" that is sometimes heard when a field is applied to a magnetostrictive material. For this reason, transformer core materials are usually designed with the least possible magnetostriction.

2. MATERIALS

2.1 Linear Magnetostriction Theory

2.1.1 Magnetostriction at Magnetic Saturation In general, the effect of a saturating magnetic field on a magnetostrictive material is given by a fourth-rank tensor which couples strain to field. In the most general case, it takes 15 constants to describe the magnetostriction in a material. However, for the important cases of cubic and hexagonal symmetry, only two and four constants, respectively, are needed. The following equation shows the saturation magnetostriction for cubic crystals (neglecting the volume change):

$$\Delta l/l = \tfrac{3}{2}\lambda_{100}(\alpha_x^2\beta_x^2 + \alpha_y^2\beta_y^2 + \alpha_z^2\beta_z^2 - \tfrac{1}{3}) + 3\lambda_{111}(\alpha_x\alpha_y\beta_x\beta_y + \alpha_y\alpha_z\beta_y\beta_z + \alpha_z\alpha_x\beta_z\beta_x), \tag{1}$$

where λ_{100} and λ_{111} are the saturation magnetostrictions measured along the (100) and (111) directions, respectively, and the α's and β's are the direction cosines of the applied field and the measurement direction with respect to the crystal axes (Clark, 1980).

2.1.2 Linearized Magnetostriction below Saturation Many devices operate at fields much less than saturation where the magnetostrictive response is quasilinear, even though in general, the magnetic properties depend on the applied field H. Also, in practice, the applied field and the strain are often in the same direction. In this linear approximation, the applied magnetic field H and the strain S are related by the *effective piezomagnetic constant d*. The relationship is given by

$$d = \frac{\partial S}{\partial H}. \tag{2}$$

This relationship describes the simplest idea of a change in magnetic field causing a change in a material's length. In a similar fashion, the relationship between a change in the magnetic flux density B and a change in H is given by

$$\mu = \frac{\partial B}{\partial H}, \tag{3}$$

where μ is the usual permeability. However, the changes in S and B also depend on the changes in applied stress. The complete linearized difference equations of state for S and B are given by

$$\begin{aligned} \Delta S &= s^H \Delta T + d\Delta H, \\ \Delta B &= d\Delta T + \mu^T \Delta H, \end{aligned} \tag{4}$$

where T is the applied stress, s^H is elastic compliance at constant H, d is the effective piezomagnetic constant, and μ^T is the permeability at constant stress. For this one-dimensional, linear case, it is useful to define a figure of merit from the fundamental properties described above. This figure of merit is called the coupling factor and is denoted by k, with

$$k^2 = d^2/s^H \mu^T. \tag{5}$$

The coupling factor is a measure of the conversion efficiency of magnetic energy into mechanical energy, and vice versa.

2.2 Magnetic Anisotropy

In some magnetostrictive crystalline materials, it is energetically favorable for the magnetization vector to lie along certain crystallographic directions. This is called the *magnetic anisotropy energy* and has units of J/m^3. If a smoothly varying magnetic field is applied to a material with a sufficiently high anisotropy, the resultant variation in magnetostriction is not smooth: "jumping" occurs. An example of such a material is shown in Figs. 8(a) and 8(b). This figure shows a $\mathrm{Tb}_x\mathrm{Dy}_{1-x}$ alloy that has substantial magnetic anisotropy. Note the jumping and nonjumping regions of the curve. The magnetization is also shown, since it displays the "jumps" more clearly. This jumping occurs when the moment direction abruptly enters or leaves a low-energy direction. Jumping is frequently an undesirable trait for device applications, and elements with opposite anisotropies are often alloyed together in an attempt to reduce or eliminate the magnetic anisotropy energy. Figures 8(c) and 8(d) show a different $\mathrm{Tb}_x\mathrm{Dy}_{1-x}$ alloy with nearly zero anisotropy; the magnetostriction is much smoother.

2.3 Important Material Properties

In this section, we discuss some of the material properties that are desirable from a device standpoint.

The effective piezomagnetic constant d represents the rate of change in length with applied field. The "d constant" should be large.

Magnetization M is the internal magnetic moment per unit volume of the sample along a given direction. Useful magnetostriction occurs either by rotation of the internal moment or by motion of the domain walls. The saturation magnetization M_s is the maximum value of M. A large M gives a large "handle" to rotate the moments using an applied magnetic field; in addition, the magnetic energy available for conversion to mechanical energy is given by $\mathbf{M}_s \cdot \mathbf{H}$.

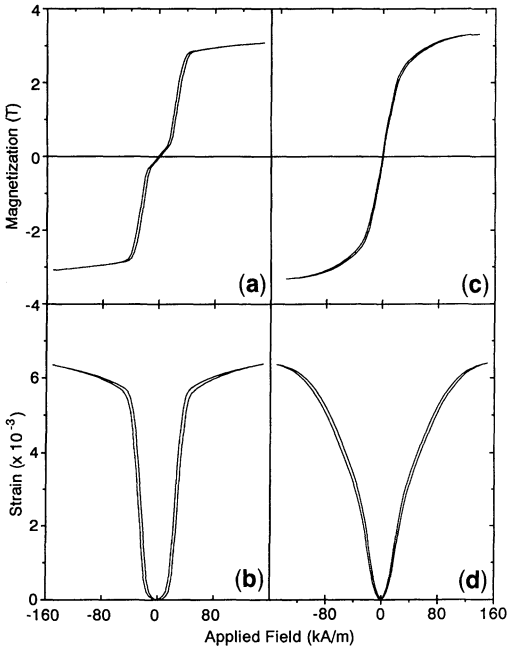

FIG. 8. (a),(b) Magnetization and strain of a $\mathrm{Tb}_{0.67}\mathrm{Dy}_{0.33}$ rod with large anisotropy. (c),(d) The same measurements on $\mathrm{Tb}_{0.6}\mathrm{Dy}_{0.4}$, which has almost zero anisotropy. Note the sharp jumps in (a) and (b). See Fig. 4 for a note on the units.

Permeability μ is the rate of change in B with applied field. It should be large when the object is free to change its dimensions.

Magnetic anisotropy K is a measure of the energy required to move the moment away from its preferred crystallographic direction. For most applications, K would ideally be zero. The magnetic anisotropy provides a restoring force when no external stress is applied.

Elastic compliance s describes the stiffness of the material and has units of Pa^{-1}. The compliance is the inverse of the modulus. A material with a high compliance is very "soft." In order to generate large amounts of work, a large force is desired. Therefore, the material should have low compliance; i.e., it should be "stiff" in order to support the large forces.

In saturation magnetostriction λ_s, the saturation strain, is the strain measured at full extension. This number should be as large as possible in actuators; it is not as critical in sensor applications.

The coupling factor k is a measure of the conversion efficiency between mechanical and magnetic energies. If a very sensitive device is desired, it is critical that this number be large, i.e., close to 1.

2.4 Useful Materials

2.4.1 Crystalline Materials Nickel, with a magnetostriction of -35×10^{-6}, was the first material used for magnetostrictive devices. It was used for sonar transmitters in World War II. At the same time, the Japanese, who did not have access to large supplies of nickel, developed the $Al_{0.13}Fe_{0.87}$ alloy Alfer. Alfer has a positive magnetostriction of about 40×10^{-6}.

Terfenol-D ($Tb_{0.27}Dy_{0.73}Fe_{1.95}$) is the principal room-temperature magnetostrictive material now in use. Its large magnetostriction (about 0.15% for polycrystalline material) and commercial availability make it an attractive material. The combination of Tb and Dy leads to near-zero anisotropy at room temperature. It is available in rods of up to nearly 4-cm (1.5-in.) diameter and 20-cm (8-in.) length. It is normally operated with compressive stresses of between 10 and 40 MPa.

Tb, Dy, and their alloys have the largest known magnetostriction; however, this is achieved only at cryogenic temperatures. The advent of high-temperature superconductors has rekindled interest in TbDy, since the combination of superconducting coils and materials with extraordinary magnetostrictions is attractive. The alloy $Tb_{0.6}Dy_{0.4}$ has near-zero anisotropy at liquid nitrogen temperature (77 K) where these superconductors operate. The magnetostriction is huge, about 0.6 percent, and operation is possible with stresses as high as 20 MPa (Clark *et al.*, 1992). Use of this material is in the prototype stage.

Structural steels have a small but usable magnetostriction for some applications—for example, torque sensors. In addition, various iron alloys, especially alloys containing nickel, have been developed. The nickel content can be varied to obtain the best possible tradeoff between magnetostriction and other necessary properties. The magnetostrictive properties of these alloys varies widely with composition and crystal orientation. For example, with a composition of 40% Ni, the [100] direction has a negative magnetostriction of about 7×10^{-6} and the [110] and [111] directions have positive magnetostrictions around 23×10^{-6}. The material is nearly isotropic at 60% Ni with a positive magnetostriction of 20×10^{-6}, while at 85% Ni, the magnetostriction is almost zero (Bozorth, 1951).

2.4.2 Amorphous Materials Amorphous magnetic materials are made by rapidly quenching the melt. These materials can be made in the form of ribbons (commercial ribbons are typically about 13 cm wide and 13–50 μm thick) and wires (diameter about 125 μm). The ribbons are made by spraying the liquid material onto a cooled metal wheel; cooling rates in the region of millions of degrees/second prevent the formation of crystals. Wires are made by injecting the molten metal into a much cooler liquid. In the case of the wires, cooling rates are somewhat less than those for ribbons. Typical magnetostrictive materials consist of iron and boron, with other materials such as silicon and carbon added as needed. The ribbon materials are designed to be used in transformer cores; however, after proper annealing in a magnetic field (Modzelewski *et al.*, 1981), they become excellent sensor materials, with extremely high coupling coefficients, i.e., almost complete conversion between mechanical and magnetic energy. In contrast to their high coupling coefficients, amorphous materials have low magnetostrictions (0.004%) and saturate at low magnetic fields. Similar properties can be obtained in the wire materials by the same annealing process.

3. DEVICES

3.1 Sensor Applications

In this section, we use the term sensor in a broad sense, which includes any application that measures a physical phenomenon. When sensitivity is of prime importance, sensor materials need to have a very high coupling coefficient. In the case of transverse-field–annealed amorphous ribbons, k can be as high as 0.95. Magnetostrictions (40×10^{-6}) and saturation magnetic fields (80 A/m) are low. This means that, although the materials are very sensitive, they can easily be saturated if even moderate strains or fairly low stray magnetic fields are present. Furthermore, they are mechanically fragile. In applications requiring less sensitivity, ordinary steels or alloys specially formulated for high magnetostriction provide enough sensitivity and are strong enough to carry mechanical loads. For example, the magnetostriction in a steel shaft can be used to measure the applied torque.

Many sensor designs use a measurement of the active material's permeability to derive the output signal. This measurement is accomplished by using a drive coil to generate a magnetic field. The resulting magnetic flux (which includes the contribution from the material's magnetization) is measured with a pickup coil. A common refinement of this technique is to use multiple coils in a bridge arrangement. The two coils can either measure materials that are stressed in opposite directions (Fig. 9) or measure stressed and unstressed sensors (Fig. 10). This helps reduce problems such as thermal drift.

Magnetostrictive sensors have a wide range of applications. All that is needed is that the desired phenomenon can generate a force that results in a change in length. For example, a pressure gauge can be made by attaching a magnetostrictive material to a metal diaphragm. As the pressure changes, the diaphragm bends and strains the sensor. One important advantage of sensors based on magnetostrictive "strain gauges" is that the measurement can be made on a noncontact basis; it is not necessary to attach leads to moving objects in order to measure the strain. Magnetostrictive sensors can also be used to measure magnetic field. As the magnetic field changes, the material changes length. This length change can be measured by a variety of means, such as strain gauges.

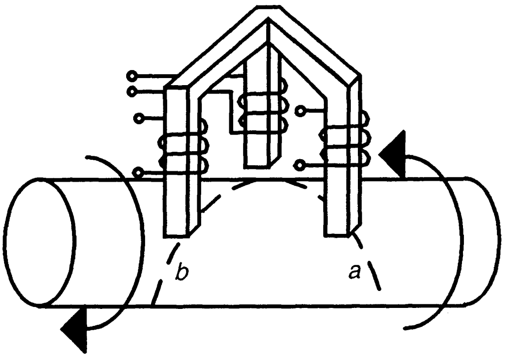

FIG. 9. Differential torque measurements on a shaft.

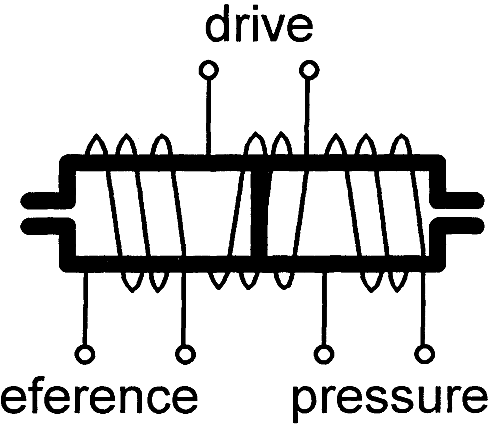

FIG. 10. Differential measurements on a pressure sensor.

3.1.1 Inventory Control Magnetostrictive ribbons are used for inventory control (antitheft) systems. A short section of ribbon is attached to the item that is to be protected, and a monitoring station is placed at the store exit. The monitoring station generates an ac magnetic field. When the ribbon encounters the magnetic field, its length begins to oscillate. If the applied frequency matches the mechanical resonant frequency of the ribbon, energy is absorbed. This energy loss is detected and an alarm sounds when a tagged item is carried out of the store. Other designs use ribbon shapes that generate harmonics or subharmonics of the applied frequency. The presence of the additional frequency is used to trigger an alarm. The ribbon can be deactivated by permanently magnetizing a piece of steel or other permanently magnetizable material that is colocated with the ribbon. This magnetically saturates the ribbon which then no longer responds to the ac field.

3.1.2 Torque Sensors Consider a shaft that has a torque applied (see Fig. 9). The material along line *a* is compressed and that along line *b* expanded. If the shaft is magnetostrictive or has a magnetostrictive material attached to it, the magnetic properties along *a* and *b* will change. These properties are usually measured in a differential fashion, as shown in the figure. An alternative method is to measure the magnetization change directly with the use of Hall sensors.

3.1.3 Pressure Sensors Magnetostrictive pressure sensors measure the change in magnetic properties of a magnetostrictive material when it is deformed by applied pressure. To compensate for environmental changes, such as temperature, a differential measurement technique is generally used. A schematic of such a sensor is shown in Fig. 10. The pressure sensor consists of two chambers: one is exposed to the pressure to be measured and the other is exposed to a reference pressure or the atmosphere. Magnetostrictive material either forms the chamber or is attached to it. A drive coil excites both chambers of the sensor, and the signal from the two pickup coils measures the permeability of the sensor material. An alternative approach is to measure the deflection of a metal diaphragm in a similar fashion.

3.1.4 Anisotropy Sensors Strain anisotropies in a magnetostrictive material are converted into magnetic anisotropies; therefore, a measurement of magnetic anisotropies reveals the anisotropy in the strain. For example, when metal is rolled to reduce its thickness, the metal expands in the other two dimensions. Improper roller settings can cause the metal to expand more in one direction than another, which leads to either wavy edges or buckling in the center. If the metal is magnetostrictive, an anisotropy sensor can be used to make a noncontact measurement of the strain in each direction and adjust the roller settings. Figure 11 shows a typical bridge arrangement used to measure magnetic anisotropy (Burke, 1986).

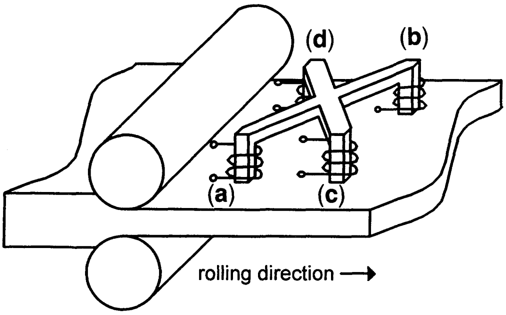

FIG. 11. A bridge arrangement for measuring strain anisotropy. The drive field is applied to legs (c) and (d), and the difference is measured across (a) and (b).

3.1.5 Villari Bridges Villari bridges are force transducers sold by ABB under the name Pressductor®. They consist of two coils of wire oriented to produce zero mutual inductance. The coils are wound through a block of laminated steel (Fig. 12) and force is applied to the top and bottom. The force causes a decrease in the permeability of the steel in the direction of the applied force, which breaks the symmetry of the device and allows an energy transfer to the pickup coils. These devices have been used in a variety of applications, such as rolling-mill pressure sensing. One interesting use is in metal rolling. A sensing roller with a series of force sensors along its length measures the pressure applied by the rolled metal as it passes over. If rolling conditions are not correct, ripples appear in the sheet metal and the pressure at different points across the roller varies. Corrections can then be applied to the rolling machinery.

3.1.6 Magnetostrictive/Piezoelectric Combinations A combination of piezoelectric and magnetostrictive materials can be used to measure magnetic field. The two materials

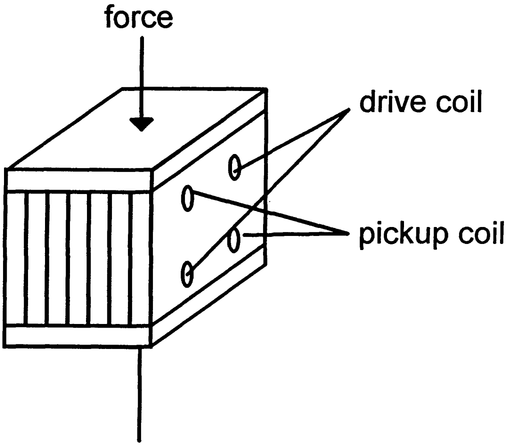

FIG. 12. Villari bridge force sensor. A drive coil is wound through two of the holes, and the pickup coil is wound through the other two. The pickup is rotated 90° with respect to the drive coil.

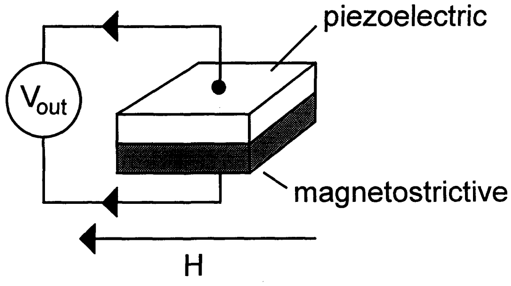

FIG. 13. Magnetostrictive/piezoelectric combination.

are bonded together (Fig. 13), and when a magnetic field is applied to the magnetostrictive material, it changes its length. This strains the piezoelectric material which then generates a voltage proportional to the magnetic field.

3.1.7 Position Sensing Position sensing can be accomplished with a rather clever device that uses a magnetostrictive material as an acoustic waveguide. The waveguide is placed inside a hollow tube (for environmental protection), and a doughnut-shaped sensor containing permanent magnets rides on the outside of the tube (Fig. 14). For liquid level sensing, the sensor floats on the liquid to be measured; for general position sensing, the sensor and the tube are connected to the moving parts. The magnet interacts with the magnetostrictive waveguide and changes its physical properties where the sensor is. In one version, an acoustic pulse is sent down the waveguide. When the pulse hits the region of the waveguide changed by the magnet, which in this case produces a different modulus (ΔE effect), the discontinuity partially reflects the pulse back to the sensor head. The far end of the waveguide is embedded in a damper material to prevent unwanted reflections. An alternative method sends a current pulse down the waveguide where the cylindrical field from the current pulse interacts with the axial field from the sensor magnets to form a local helical magnetic field. The Wiedemann effect generates a helical twist in the wire, which travels back to the sensor head. In both designs, a measurement of the transit time of the pulse gives the position of the sensor. At least one design also uses magnetostrictive tapes to measure the occurrence of the arriving pulse.

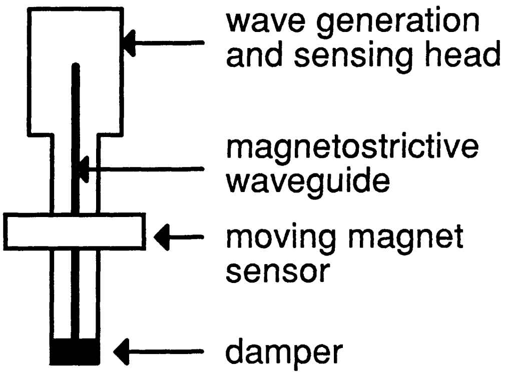

FIG. 14. Position sensor.

Adjustable limit switches using a compliant magnetostrictive sensor element are available. The element is excited at the power line frequency, and the output of the sensor depends on the strain produced by the plunger position. A threshold sensor sets the on-off position; travel up to 1.25 mm is available.

3.1.8 High-Sensitivity Strain Gauges In many of the sensors discussed in this article, the magnetostrictive material is acting as a strain gauge, measuring the strain induced by the physical parameter of interest. The transverse-annealed magnetostrictive ribbons and wires make very sensitive strain gauges (Wun-Fogle *et al.*, 1987). If a gauge factor F is defined as the fractional change in the measured parameter (permeability or resistance) divided by the change in strain, we find that conventional wire and foil resistive strain gauges have F equal to about 2, semiconductor gauges have F equal to about 250, and ribbons have F equal to about 200 000 (with μ as the parameter). The rotating magnetic state also affects current flow in the materials; annealed ribbons with resistance as the measured parameter have an F equal to about 60 (Restorff *et al.*, 1991). One problem encountered with this extraordinarily high-sensitivity strain gauge is that normal thermal expansion can saturate the sensor. This problem is solved by mounting the material with an extremely viscous fluid; however, this in turn limits operations to ac measurements.

3.1.9 Accelerometers Prototype accelerometers based on a cantilevered beam that bends under acceleration using a magnetostrictive pickup have been constructed. Magnetostrictive material is attached to the top and bottom of the beam (when the beam bends, one is under compression and the other, tension), and a differential measurement is made of the induced stress.

3.1.10 Engine-Knock Sensors A sensor for detecting engine knock has been devised (Anderson *et al.*, 1984). It consists of a resonant "C"-shaped structure with a magnetostrictive element across the open ends of the C. A permanent magnet bias is applied, and a pickup coil senses the vibration.

3.1.11 Nondestructive Evaluation There are several techniques involving magnetostriction that are useful for nondestructive testing of materials. The first of these techniques involves the measurement of sound velocity either in or, using surface waves, on a material (Kwun, 1985). The sound velocity is affected by both strain and applied magnetic field (ΔE effect). A magnetic bias is applied to the material, and an ultrasonic generator and detector measures the sound velocity. From the velocity-vs-bias information, the strain can be deduced. Another technique measures the *amplitude* of ultrasonic waves generated in a material as a function of applied bias field (Alers and Thompson, 1977). The amplitude and bias information also reveals strains in the material. A third technique using sound is the utilization of magnetic Barkhausen noise. When an ac magnetic field is applied to a material, the internal magnetization does not change smoothly on a microscopic scale. Each domain abruptly changes direction when it becomes energetically favorable. Since the material is magnetostrictive, the dimensions also change abruptly, which generates a small sound. The sum of the many weak sounds is called Barkhausen noise. The amount of Barkhausen noise emitted depends upon the strain, so that noise measurements reveal information about the material strain. In the third technique, the permeability depends upon the strain, and so direct measurements of this property also give strain information.

3.1.12 Magnetometers When exposed to a magnetic field, a magnetostrictive material by definition changes its length. This converts the problem of measuring magnetic field into one of measuring length. Many methods for measuring length have been used to make a magnetometer: laser interferometric, capacitance, conventional strain gauge, and piezoelectric. In addition, the magnetic properties can be measured directly.

3.1.13 Potential Applications Much of the future use of magnetostrictive materials depends upon the application of new materials to conventional measurement schemes. These new materials, such as transverse-field–annealed amorphous ribbons or wires, and new alloys with giant magnetostrictions, will extend the sensitivity of current sensors. For example, Giallorenzi (1981) calculates that optical-fiber interferometers coated with a magnetostrictive material will exploit the most sensitive aspect of the interferometer and should achieve magnetic field measurement capabilities of 10^{-9}–10^{-10} A/m.

3.2 Actuator Applications

The purpose of an actuator is to convert an electrical signal into mechanical motion. Many types of actuators are needed, from micropositioning actuators that require precise movements and little force to high-power actuators that can deliver up to 45 000 N (10 000 lb) or more of force. Magnetostrictive actuators have advantages over other types of actuators in applications that need low-frequency operation and/or low voltages, since they can easily be driven by conventional low-impedance amplifiers at low frequency. The low impedance of a magnetostrictive actuator means that driving voltages can be low. This can simplify designs, for example, in medical applications.

It is possible to make a very simple actuator by taking a nickel or Terfenol rod and wrapping a coil around it. Such an actuator will work, but it will be far from optimum. A *prestress* is needed on the rod to make sure that the moments are aligned properly. A rod without prestress has the moments aligned randomly and will only achieve about half of its maximum magnetostriction. This is because some of the moments are already aligned with the rod axis and will not be affected by an applied field. Another technique that is often used is *permanent magnet biasing*. Because the magnetostriction is independent of the sign of the magnetic field, a magnetostrictive actuator driven by an ac signal moves at twice the drive frequency and the motion is only in one direction. Both of these problems can be corrected and a quasi-linear operating region achieved by placing a static magnetic field on the material to bias the material half-way up the magnetostriction curve (see Fig. 15). This can be accomplished either by adding a dc current to the drive coil or by placing permanent magnets into the magnetic path. Permanent-magnet biasing has the advantage of considerable power savings. Finally, careful design of the magnetic circuit is required, especially in high-power devices, in order to achieve high efficiency.

There are two basic high-power actuator designs in use today. The first, shown in Fig. 16(a), is a simple "one-legged" axially symmetric design. The second [Fig. 16(b)] is a "two-legged" design, usually used for high-power devices.

3.2.1 Sonar Transmitters The first widespread use of sonar transmitters was during World War II and used nickel magnetostrictive elements. Sonar transmitters are perhaps an ideal application for magnetostrictive transducers, since low frequencies and high forces are needed. One commercial family of sonar transmitters has a usable frequency range of 50 Hz–5 kHz and an output of 206 dB (relative to 1 μPa at 1 m) at the transmitter's resonant frequencies. Uses include submarine detection, mine sweeping, seismic studies, and communication links.

3.2.2 Active Vibration Control Magnetostrictive actuators have been used as the active elements in active vibration control applications. The high-force, low-frequency response of the actuators makes them attractive in some applications.

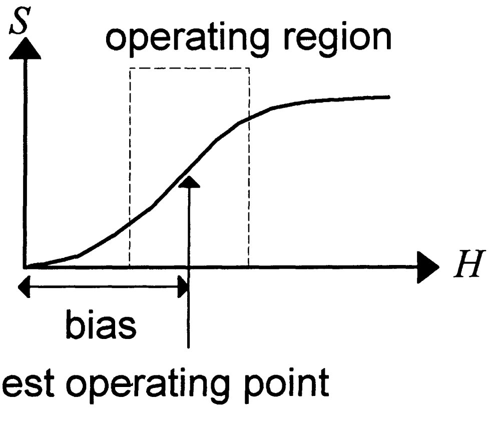

FIG. 15. Permanent-magnet bias used to pick a linear operating region.

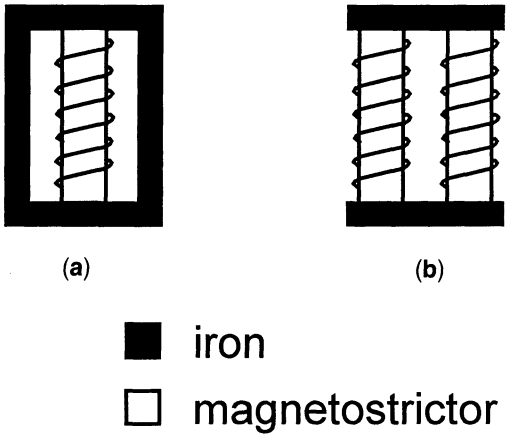

FIG. 16. Typical high-power actuator configurations; (a) single-legged design and (b) double-legged design.

3.2.3 Deformable Mirrors A deformable mirror is one whose shape is changed by a group of actuators. There are several reasons for doing this. Large mirrors (such as those used in astronomy) must be very thick in order to maintain their shape against gravitational and mounting forces. A deformable mirror can be made much thinner and use the actuators in concert with a real-time control system to maintain the correct figure. Another use is to warp the mirror actively to correct for atmospheric turbulence. The same high-force, low-frequency response that makes magnetostrictive actuators useful for vibration control makes them attractive for use in deformable mirrors. One problem that must be considered in this application is that of removing the heat generated by the driving coils, since the actuators would be placed fairly close together.

3.2.4 Magnetostrictive/Piezoelectric Combinations Sonar transducers using magnetostrictive and piezoelectric elements have been built (Butler *et al.*, 1990) (see Fig. 17). Since magnetostrictive elements present an inductive load to the driving amplifier and piezoelectric elements present a capacitive load to the driver, at resonance, the driver sees a purely resistive load. This greatly simplifies amplifier design. The motion of the two

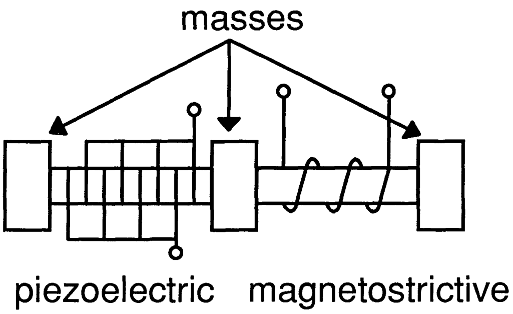

FIG. 17. Piezoelectric/magnetostrictive sonar transmitter.

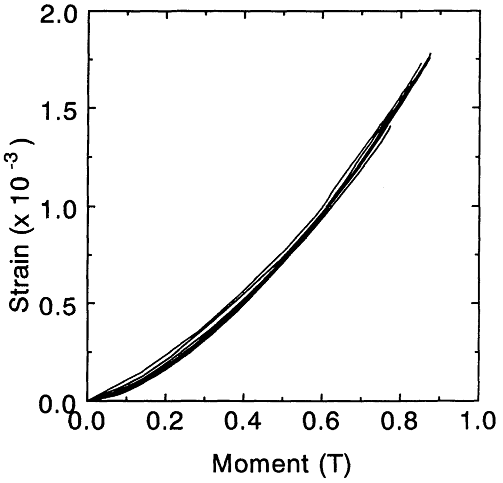

FIG. 18. Moment vs strain for a Terfenol-D rod at various prestresses.

elements is 90° out of phase, and by proper operation, the transmitter in Fig. 17 produces output with a front-to-back ratio of 10 dB.

3.2.5 Geological Tomography Terfenol-D sound sources are used in geological tomography and other oil-well logging applications. An actuator is lowered into the well with one end clamped to the well casing and the other attached to a mass. The resulting shaking induces seismic waves; sensors in the same or other wells are used to study the surrounding geology.

3.2.6 Potential Applications

3.2.6.1 Driven-Tooth Bone-Conduction Hearing Aid Bone-conduction hearing aids can be used by some people with hearing ailments, because the sound travels around the damaged part of the ear. Since the teeth are solidly connected to the jaw bone, exciting the teeth at an audio frequency provides another method of injecting audio signals directly into the inner ear. Magnetostrictive drivers for exciting teeth using a tiny rod as the active element have been built. Magnetostrictive devices are ideal for an "in-mouth" location, because their low impedance allows voltage throughout the hearing aid to be kept low.

3.2.6.2 Smart (Self-Sensing) Actuators The intrinsic magnetization M and the strain in a magnetostrictive material are closely related and almost completely independent of the hysteresis and applied stress. Figure 18 shows a curve of magnetization vs strain for a Terfenol-D rod for several different applied stresses from 5.5 to 40 MPa. This simple relationship allows a measurement of M to be converted into a length; i.e., the rod is "self-sensing." Self-sensing actuators allow a reduced parts count in the device.

3.2.6.3 Magnetostrictive Motors There are two main areas where magnetostrictive motors look promising: linear motion and rotary motion. Since the length change of a 6-cm Terfenol-D rod is on the order of 0.1 mm, direct coupling between the object and magnetostrictive rod is not useful. Instead, the motors are "inch worm" designs, whereby many individual rod movements can be added together to obtain the desired motion. Some designs use multiple sets of magnetostrictive elements, with some of the elements clamping and unclamping the load shaft and other elements moving the shaft. If the clamping and pushing motions are sequenced properly, useful motion can be achieved. A magnetostrictive rotary motor has achieved a torque of 12 N m with 800-μrad steps (Vranish *et al.*, 1991).

An unusual type of linear motor, which utilizes the fact that Joule magnetostriction in Terfenol-D is a constant-volume process, is the Kiesewetter motor. The constant-volume constraint means that the diameter of a magnetostrictive rod decreases slightly when the rod lengthens. In the Kiesewetter motor, a Terfenol rod is placed inside a tube. The diameter of the rod is chosen so that it just fits inside the tube when the rod is extended by an applied magnetic field. When the field is removed, the rod's diameter expands and clamps the rod in place. The rod is moved back and forth by moving a magnetic field along the length of the rod. Where the field is applied, the rod shrinks and lengthens. Figure 19 shows the first step of the process. A version of the motor which has been reported (Roth, 1992) has a useful travel of 200 mm and a speed of 20 mm/s. It also has the advantage of holding its position when no field is applied. The reference also contains a more complete description of the motor's operation.

3.3 Other Applications

The applications in this section do not fit neatly into either the sensor or actuator category and cover a variety of topics, such as variable delay lines and passive damping.

3.3.1 Variable Delay Line The modulus of a magnetostrictive material, and hence the speed of sound, vary with applied magnetic field. If a piece of magnetostrictive material has a transmitter on one end and a receiver on the other end, the transit time can be varied by varying a magnetic field (Hongu, 1979). Simple coils can be used for both the transmitter and the receiver (Fig. 20). A permanent-magnet bias is required to prevent frequency doubling by the magnetostrictive material.

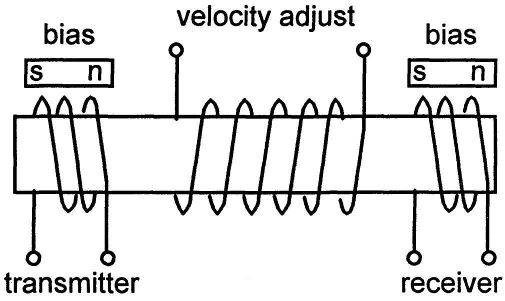

FIG. 20. Variable delay line. The magnetic field varies the modulus and hence the sound speed in the rod.

3.3.2 Passive Damping When a vibratory stress is applied to a magnetostrictive material, the magnetic state of the material constantly varies. An energy loss occurs because of the magnetic hysteresis caused by irreversible domain wall motion. Several alloys have been designed to maximize the hysteresis loss and are used to provide passive damping. One of these alloys is Nivco-10 ($Co_{0.72}Ni_{0.22}Ti_{0.02}Al_{0.005}$). The damping properties of this material are independent of frequency over most of the normal range of engineering concern and increase linearly with applied stress up to 120 MPa, where the ratio of the damping energy to total strain energy is about 0.5 (Goodman, 1988). Because the damping is a magnetic effect, it disappears above the Curie temperature or when the alloy is placed in a saturating magnetic field.

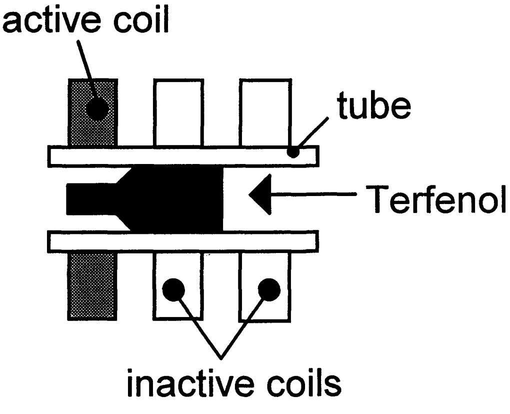

FIG. 19. Kiesewetter motor in the first stage of operation.

3.3.3 Current Threshold Sensors Garshelis (1979) describes a current threshold sensor (Fig. 21) for ac current. The device uses helical anisotropy in a magnetostrictive cylinder as the sensing principle. Below the current threshold, the magnetic moment makes only small variations. When the current exceeds the threshold, the magnetostrictive energy exceeds the anisotropy energy and a large shift in the magnetization occurs. The amount of helical anisotropy, and thus the current threshold, can be varied by altering the twist on the cylinder.

3.3.4 Record/Playback Heads A record/playback head which uses an acoustic wave in a substrate to alter the magnetic properties in an overlying magnetostrictive film has been designed for use with magnetic tape (Ballinger *et al.*, 1986). The head is aligned parallel to the width of the tape. An acoustic wave launched in the substrate generates a magnetically active region in the

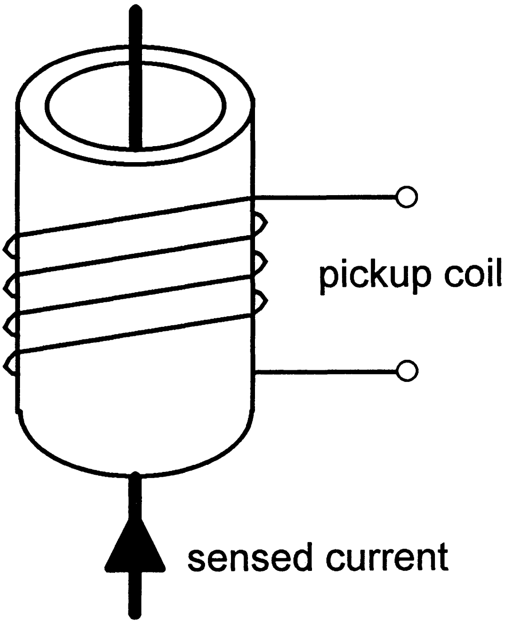

FIG. 21. Current threshold sensor. At currents above the threshold, the magnetization reverses abruptly, generating a large signal in the pickup coil.

magnetostrictive film. A pulse in an adjacent conductor generates a magnetic signal on the tape. This allows a single head to produce many channels across the width of the tape.

3.3.5 Vibratory Stress Relief It is not clear whether or not to call vibratory stress relief an "application" of magnetostrictive materials. Machine tools, such as drills, tool bits, etc., that are constructed of magnetostrictive materials (this includes carbides and steels, among others) are placed in a strong ac field generated at power line frequencies. This induces magnetostrictive vibrations in the material, which in turn relieve internal strains. Tool life extensions of 20–200% are claimed (*Tooling & Accessories*, 1987).

GLOSSARY

Actuator: a device that converts electrical (magnetic) energy to mechanical energy.

Curie Temperature: The temperature above which the thermal energy overwhelms the magnetic energy and the material loses all its ferromagnetic properties.

Domain: a small region in a material where all the atomic moments point in the same direction. A normal material is made up of many small domains. If the material is unmagnetized the domains point in random directions; if magnetized, they point in the same direction.

Magnetization: the sum of all of the atomic moments per unit volume pointing in the measurement direction.

Sensor: a device which converts mechanical energy to electrical (magnetic) energy.

Works Cited

Alers, George A., Thompson, Robert B. (1977), U. S. Patent No. 4,048,847.

Anderson, Phillip M., III, Raskin, Donald, Reich, Ronald K. (1984), U. S. Patent No. 4,463,610.

Ballinger, Dale O., Chynoweth, William R., Wade, Gerald J. (1986), U. S. Patent No. 4,604,669.

Bozorth, Richard M. (1951), *Ferromagnetism*, Princeton, NJ: D. Van Nostrand, p. 448.

Burke, Harry E. (1986), *Handbook of Magnetic Phenomena*, New York: Van Nostrand Reinhold Co., Chap. 16.

Butler, J. L., Butler, S. C., Clark, A. E. (1990), *J. Acoust. Soc. Am.* **88**, 7–11.

Clark, A. E. (1980), in: E. P. Wohlfarth, (Ed.), *Ferromagnetic Materials*, Vol. 1, Amsterdam: North-Holland Publishing Co., p. 531

Clark, A. E., DeSavage, B., Bozorth, R. (1965), *Phys. Rev.* **138**, A216–A224.

Clark, A. E., Belsen, H. S., Tamagawa, N., Callen, E. (1974), in: *Proceedings of the International Conference on Magnetism, ICM-73, Moscow, September 1973*, Moscow: Nauka, Vol. 4, p. 335.

Clark, A. E., Wun-Fogle, M., Restorff, J. B., Lindberg, Jan F. (1992), *IEEE Trans. Magn.* **28**, 3156–3158.

Garshelis, I. J. (1979), *IEEE Trans. Magn.* **15**, 1809–1811.

Giallorenzi, T. G. (1981), *Optics and Laser Technology* **3** (2), 73.

Gibbs, M. R. J., Squires, P. T., Ford, P. J., Brugel, D. (1988), *IEEE Trans. Magn.* **24**, 1764–1766.

Goodman, L. E. (1988), in: Cyril M. Harris (Ed.), *Shock and Vibration Handbook*, New York: McGraw-Hill, pp. 36–17.

Hongu, Masayuki (1979), U. S. Patent No. 4,165,498.

IEEE (1990), *IEEE Standard on Magnetostrictive Materials: Piezomagnetic Nomenclature*, IEEE Standard 319–1990, New York: IEEE.

Kwun, Hegeon, Teller, Cecil M., II (1985), U.S. Patent No. 4,497,209.

Masiyama, Y. (1931), *Sci. Rep. Tohoku Univ.* **20**, 574.

Mason, Warren P. (1958), *Physical Acoustics and* *the Properties of Solids*, Princeton, NJ: D. Van Nostrand, Appendix.

Modzelewski, C., Savage, H. T., Kabacoff, L. T., Clark, A. E. (1981), *IEEE Trans. Magn.* **17,** 2837–2839.

Restorff, J. B., Wun-Fogle, M., Hathaway, K. B., Clark, A. E. (1991), *J. Appl. Phys.* **69,** 4668–4673.

Rhyne, J., Legvold, S. (1965), *Phys. Rev.* **138,** A507–A514.

Roth, R. C. (1992), in: H. Borgmann, K. Lenz (Eds.), *Actuators 92: Proceedings of the 3rd International Conference on New Actuators, Bremen, Germany,* Berlin: VDI/VDE-Technologiezentrum Informationstechnik GmbH, p. 138.

Smit, J. W., Wijn, H. P. J. (1959), *Ferrites*, New York: Wiley, p. 169.

Tooling & Accessories (1987), August, p. 65.

Vranish, J. M., Naik, D. P., Restorff, J. B., Teter, J. P. (1991), *IEEE Trans. Magn.* **27,** 5355–5357.

Weil, L., Reichel, K. (1954), *J. Phys. Radium* **15,** 725.

Wun-Fogle, M., Savage, H. T., Clark, A. E. (1987), *Sensors and Actuators* **12,** 323–331.

Further Reading

Cullen, James R., Clark, Arthur E., Hathaway, Kristl B. (1993), "Magnetostrictive Materials," in: K. H. J. Buschow (Ed.), *Materials Science and Technology*, Vol. 3B, Part II, Weinheim, Germany: VCH, p. 529.

de Lacheisserie, Etienne du Trémolet (1993), *Magnetostriction, Theory and Applications of Magnetoelasticity*, Boca Raton, FL: CRC Press.

Hathaway, Kristl B., Clark, Arthur E. (1993), *MRS Bull.* **18,** 34–41.

MAGNETS

FRITZ HERLACH, *Katholieke Universiteit Leuven, Leuven, Belgium*

HARRY JONES, *Clarendon Laboratory, University of Oxford, Oxford, England*

Introduction 245
1. **Electromagnets** 246
1.1 Magnets with Iron Cores 246
1.2 Solenoid Magnets 247
1.2.1 Water-Cooled High-Power Magnets 247
1.2.2 Superconducting Magnets 248
1.2.3 Hybrid Magnets 250
2. **Permanent Magnets** 251
2.1 Applications 252
2.1.1 Direct Use of the Magnetic Force 252
2.1.2 Applications Based on Faraday's Law 253
2.1.3 Lorentz Forces on Electrical Currents 253
2.1.4 Lorentz Forces on Free Electrons 253
3. **Pulsed Magnets** 254
3.1 Submegagauss Fields (Nondestructive) 254
3.2 Multimegagauss Fields (Destructive) 254
3.2.1 Capacitor-Driven Single-Turn Coils 254
3.2.2 Magnetic Flux Compression 255
4. **Magnets for Special Applications** 257
4.1 Magnetic Separation 257
4.2 Magnetic Levitation and Propulsion 257
Glossary 257
Works Cited 259
Further Reading 259

INTRODUCTION

Magnetic minerals (lodestone) were discovered in ancient Greece in the region of Magnesia; by tradition, this is ascribed to Thales of Miletus (Asimov, 1982). Magnets have been surrounded by mystery ever since, because they represent an action at distance that can be directly felt. The use of the compass is documented in China as early as 121 A.D. (on land only). In the West, Peregrinus (born 1240) is credited with giving it the form with the pivot and the rose (compass card) that made it a practical instrument for navigation. This had far-reaching consequences, as it opened the oceans for exploration. Around 1600, Gilbert demonstrated with experiments on a model that the Earth is a spherical magnet. Shortly after the discovery by Oersted in 1819 that an electric current generates a magnetic field, Ampère invented the solenoid electromagnet and guessed correctly that all magnetism is due to circulating electric currents. Sturgeon combined the solenoid with a soft iron core to create the first laboratory electromagnet, whose field was much stronger than that of all permanent magnets known at the time. This gave rise to the development that culminated in the huge electromagnet of the French Académie des Sciences at Bellevue near Paris (in a way, the first "national magnet laboratory") with a total weight exceeding 100 tons. More importantly, iron-core electromagnets have spawned the huge electrical industry by providing the basis for the development of efficient motors and generators.

The foremost pioneer of magnet design in recent times was Bitter (Erber and Fowler, 1969). He designed magnets with high homogeneity as well as a most efficient iron-core magnet; in 1936, he designed and built high-power water-cooled magnets as they are still used today in the big national magnet laboratories. A review of the historical development and of the design principles of these magnets is given by de Klerk (1965).

After the discovery of superconductivity by Kamerlingh Onnes in 1911, it took more than

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

four decades until Matthias discovered type II superconductivity (which is not destroyed by a moderate magnetic field) in niobium-tin that made it possible to build usable superconducting magnets. These magnets can now compete with all but the most powerful water-cooled magnets, and they have been used in conjunction with such magnets to generate the highest continuous magnetic fields in so-called "hybrid" magnets. Still higher fields up into the megagauss range have been generated by pulsed magnets, pioneered by Kapitza (ter Haar, 1964) in the late 1920s. Fields in excess of 100 T can only be generated for very short time intervals (0.1–10 μs) at the expense of a violent destruction of the coil.

Permanent magnets are most suitable for practical applications. Their widespread use was long hampered by their bulk and relatively low field. This has been drastically changed by recent developments in materials technology, which provided a variety of permanent magnets with exceptional properties. Among others, these have made possible the miniaturization of high-performance electric motors, which are now widely used in many technical appliances.

Many aspects of magnet design are discussed in the series of International Conferences on Magnet Technology that was inspired by the extensive use of sophisticated magnets in particle accelerators (e.g., Dawson, 1992). The range of fields obtained with different types of magnet is shown in Fig. 1.

1. ELECTROMAGNETS

1.1 Magnets with Iron Cores

By analogy to an electric circuit where the "electromotive force" $\oint \mathbf{E}\cdot d\mathbf{l}$ drives electrical current through resistors, an iron-core electromagnet can be described as a magnetic circuit where the "magnetomotive force" $\oint \mathbf{H}\cdot d\mathbf{l}$ drives magnetic flux $\Phi = \int \mathbf{B}\cdot d\mathbf{S}$ through materials with different permeabilities μ:

$$I = \oint \mathbf{H}\cdot d\mathbf{l} = \Phi \sum\left(\frac{l}{\mu\mu_0 S}\right) = R\Phi. \tag{1}$$

Here, I (NI in case of a solenoid with N turns) is the total current passing through a surface bounded by the integration path, $\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}) = \mu\mu_0\mathbf{H}$ is the magnetic induction, $\mathbf{M}$ is the magnetization, and l is the length of each segment with relative permeability μ and cross section S; R is called the (total) magnetic reluctance. Another important concept is the demagnetizing field and associated demagnetizing factor. This is best illustrated with the example of a toroidal ring of magnetic material with permeability μ, total circumference c, and an air gap d, encircled by a total current of NI ampere-turns. The magnetic field in the gap is then given by

$$H = \frac{NI}{(\mu - 1)d + c}; \tag{2}$$

i.e., the integration path in Ampère's law is increased by the amount $(\mu - 1)d$, which results in a substantial decrease of the field. This can formally be represented by a "demagnetizing field" $H_d = (d/c)M$, which is opposed to the applied field; d/c is called the demagnetizing factor and can be calculated for symmetrical shapes of the magnetic ma-

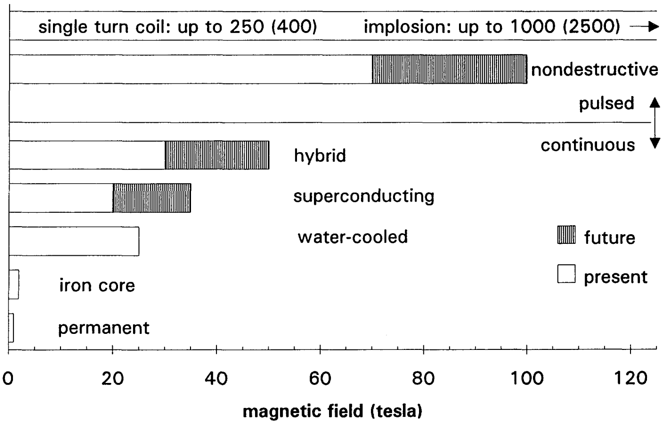

FIG. 1. State of the art and perspectives of generating strong magnetic fields. Estimates for the future are, of course, speculative (and optimistic).

terial such as a sphere where it is equal to $\frac{1}{3}$. Of course this simple type of calculation can only give rough (but useful) estimates, because the $M(H)$ relation is highly nonlinear in a ferromagnet.

From Eqs. (1) and (2), the design of iron-core electromagnets appears fairly straightforward; however, unexpected effects occur due to the strong nonlinearity in the $B(H)$ relation. Magnet performance is drastically reduced when the magnetization is driven into saturation in some part of the magnetic circuit (de Klerk, 1965).

Apart from their general use in motors and generators, iron-core magnets are still widely used in technical applications, e.g., for lifting weights and in actuators and relays, although the latter are now mostly replaced by semiconductor switching devices. In scientific research, iron-core magnets are gradually being phased out in favor of superconducting magnets. A large-scale application is in high-energy particle accelerators, where flexibility in shaping the magnetic field for focusing and beam transport is required. Now there is a trend toward superconducting magnets for these applications as well.

1.2 Solenoid Magnets

1.2.1 Water-Cooled High-Power Magnets

The need for high magnetic fields for research purposes has led to the establishment of several dedicated high–magnetic-field facilities worldwide. In these institutions, continuous (i.e., dc) fields in the range 10–25 T are routinely generated. The basic tool is the water-cooled electromagnet and large dc power supplies of several MW rating which can be either rotary machines, as at Clarendon Laboratory, Oxford University, Oxford, United Kingdom, Francis Bitter National Magnet Laboratory, Cambridge, Massachusetts, U.S.A., and International Laboratory of High Magnetic Fields and Low Temperatures, Wroclaw, Poland, or transformer–rectifier sets used elsewhere. Details of most of the world's facilities can be found in Herlach and Franse (1989), with an update in Miura (1994).

As iron-cored magnets become intolerably large for fields of only a few tesla, it becomes more efficient to dispense with the iron altogether and use air-cored copper-wound solenoids. This is, however, at the expense of power P. The field in an air-cored solenoid is given by

$$B_{\max} = \mu_0 G(\alpha,\beta)(P\lambda/\rho a_1)^{1/2}. \tag{3}$$

Here, $G(\alpha,\beta)$ is the Fabry function, which depends on the current density distribution $j(r,z)$ in the solenoid, where r is the radius and z is the axial displacement from the centroid of the solenoid. The parameters are $\alpha = a_2/a_1$ (outer/inner radius), $\beta = b/a_1$ (b = half length), and λ = packing factor (conductor volume/total volume). Different distributions $j(r, z)$ are described along with their relevant functions $G(\alpha,\beta)$ and values in Montgomery (1969) for each type. One of the most efficient, practical designs is that pioneered by Bitter (Erber and Fowler, 1969) in the early 1930s. The magnet (Fig. 2) consists of a stack of annular copper plates, each of which has a radial split, resulting in a $j \propto 1/r$ distribution. Series electrical connections between plates are made by pressure contact between two sectors adjacent to the split. The remaining area is insulated by insertion of a suitable material, such as polyimide film. Bitter magnets are the most commonly used in the major magnet installations. A more recent development is the polyhelix design. First proposed by Wood and developed by Carden, by Hudson and Jones, and by Schneider-Muntau, this construction comprises a set of single-layer helical windings nested coaxially and joined electrically in series/parallel groups [see references in Herlach and Franse, (1989) and in the Magnet Technology conferences]. This allows the current density to be tailored in any of the layers by adjusting the conductor cross section. Above about 20 T, the problem of electromagnetic forces is such that copper alone is not strong enough. One way to overcome this is to adjust cross section such that the hoop stress given by

$$\sigma = B(r)\,jr \tag{4}$$

nowhere exceeds the proof stress of copper.

Because of the massive power dissipation in these solenoids (of order 10 MW or more), they have to be actively cooled. A cooling fluid has to be pumped through the windings, and therefore, cooling channels have to be incorporated in the structure of the magnet. This has an impact on the space factor λ. Ideally,

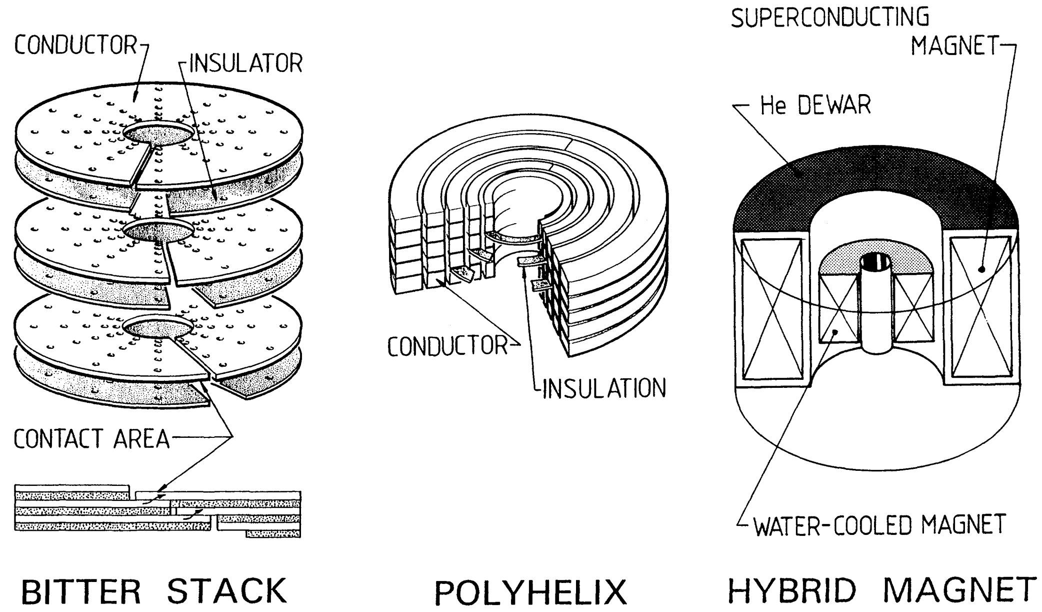

FIG. 2. Schematic drawings of the different types of water-cooled high-power magnets. For the Bitter magnet, only the plates with holes for the coolant are shown. Alternating copper and insulation, these are clamped together axially in a stack to form a solid spiral.

the cooling channels need to sustain high-velocity turbulent flow, and in practice, this is achieved by distributing many narrow channels throughout the winding. This is relatively easy to achieve in polyhelix magnets by separating layers with spacers through which the coolant can flow axially. In the case of Bitter magnets, cooling is achieved by punching a precise array of holes in each plate and insulator and aligning them very carefully in the stack to give continuous axial cooling channels, although there is a variant in which channels are etched in the faces of the plates to allow radial cooling. Water is by far the best coolant, but as the presence of electrical insulation on the cooled surface of the windings would severely impair heat transfer, the water must have very low conductivity and hence be deionized. Deionized water is adequate for the low-voltage (500 V or less), high-current (of order 10 kA) systems installed in most magnet laboratories. Flow rates need to be very high, perhaps several hundred $m^3\ h^{-1}$, so that large pressure drops result, and therefore substantial pumps are required. A typical cooling system then consists of a closed circuit of reservoir, pumps, pipework, magnet, water cooler, and deionizer. Coolers for the water differ from system to system but may, for example, comprise a heat exchanger with a secondary water circuit which dumps the heat into a nearby river as at the French-German "Hochfeld-Magnetlabor" at Grenoble (run by CNRS and the Max Planck Institute), and the Francis Bitter Laboratory, or use a forced-draft air cooler as at Oxford. Some systems incorporate a refrigeration unit as at Katholieke Universiteit, Nijmegen, The Netherlands, and at the new U.S. National High Magnetic Field Laboratory (Tallahassee, Florida). The thermal capacity of the reservoir can be used to some extent, but this places an upper limit on the time for sustained running at high power, especially in hot weather. All systems require an on-site deionizer to keep the water conductivity sufficiently low.

1.2.2 Superconducting Magnets Superconducting magnets are now very common, especially in research laboratories, and are an everyday tool for many experiments. Although there is a significant industrial effort worldwide in the manufacture of superconductors and superconducting magnets, industrial applications of these magnets, with the possible exception of nuclear magnetic resonance (NMR) spectroscopy, are developing rather slowly. This is largely because although modern magnets and their associated cryogenics are relatively reliable, they do require some specialist knowledge and resources, e.g., liquid helium, to operate at a high utilization rate and economically. In the world of medicine, however, magnetic resonance imaging (MRI) is becoming a routine tool for noninvasive diagnosis, sustained by a sizable industry. Great advances have been made here using superconducting magnets as an economic and practical means to generate fields over the large volumes necessary to accommodate the human body.

Conductors that are stable against magnetothermal disturbances have been available commercially for winding superconducting magnets operating in liquid helium for several decades. They consist of fine (≈ 10 μm) filaments of type II superconductor embedded in a copper, or copper-bearing, matrix. In practice, only two materials, NbTi and Nb_3Sn, are routinely available. The former is a ductile alloy that is easily coprocessed with copper and is therefore relatively inexpensive. Its limitation is that its upper critical field B_{c2} is of order 10 T, and thus it is only suitable for lower field applications. Nb_3Sn is a more complex, brittle intermetallic compound and is normally formed into wire by codrawing Nb filaments and a Sn-bearing copper matrix, for example, the "bronze route." The Nb_3Sn is formed by promoting solid diffusion of the Sn into the Nb by heat treating the composite to a temperature of order 700°C for a period of time. Often this heat treatment is applied after the solenoid has been wound, i.e., the "wind and react" method. This is because excess mechanical strain severely degrades the superconducting properties of the material and strains introduced in the winding process can be high. The advantage of Nb_3Sn is its high B_{c2}, and it can be used in magnets operating a little in excess of 20 T below the lambda point of liquid helium (2.2 K). An essential feature of practical superconductors is good flux-pinning properties. Flux-pinning centers prevent the dissipative movement of flux due to the Lorentz force resulting from the interaction of magnetic field and the transport current. In NbTi, these centers are mainly precipitates that form during a final anneal at the sites of defects introduced by the cold work of wire drawing, and in Nb_3Sn, these are the grain boundaries.

The field produced at the centroid of a superconducting solenoid is given by

$$B_0 = \mu_0 F(\alpha, \beta)\, j \lambda a_1, \tag{5}$$

where j is the current density. $F(\alpha,\beta)$ is a geometry function that depends on the shape of the coil. Before selecting his parameters, such as current and shape of the coil, the designer must have knowledge of the performance of the conductor he proposes to use. This is normally in the form of a curve that shows the critical current (the current at which the superconductor becomes normal) against magnetic field as measured on an approximately 1-m length of superconductor and is known as the "short sample curve." Thus, having determined the load line (tesla/A), the maximum possible current and field (i.e., where the load line intersects the critical current) can be seen. In fact, it is not the centroid field that is important here but the value of the peak field which, for a simple solenoid, is at the inner radius on the midplane. For some geometries, typically short fat coils, the ratio of peak field to centroid field can be high (up to 2), and so this must be taken into account. As the field falls off approximately linearly with radius, it is possible to grade the conductor, i.e., increase current density by decreasing conductor size. This can be done by nesting several concentric windings of different size wires that change at the appropriate $B(r)$ value. Also, conductor material can be changed at the appropriate point, and so a typical high-field magnet will have Nb_3Sn inner sections and NbTi windings on the outside.

Potentially a superconducting magnet can suffer from several magneto/mechanical-thermal instabilities. One such is flux jumping. This is eliminated by the fine subdivision of filaments in the copper matrix described above. In the liquid-helium temperature regime, specific heat capacities of materials are very low, so that even the slightest mechanical movement (e.g., of a single loose turn) can result in a temperature rise that exceeds the critical temperature of the superconductor; hence, a "quench" (the whole magnet going normal) can be initiated. Mechanical movement is inhibited by impregnating the winding with a medium, such as epoxy resin or wax. Impregnation itself can be the cause of the problem of training, whereby the strains created by differential contraction when the magnet is first cooled down to 4.2 K get released with the concomitant generation of heat, causing a quench at progressively higher fields until the design field is reached. When a superconducting magnet quenches, the magnetic energy is converted into heat in the windings, and considerable quantities of liquid helium are rapidly boiled off from the cryostat that contains the magnet. Energies stored are typically a few kJ for small magnets but can be many MJ for large windings such as those used, e.g., for high-energy particle detectors and hybrid magnets. Studies are under way for storing electrical energy in large toroidal windings as a buffer reservoir for utilities.

During a quench, high voltages can be developed within the windings. Protection circuitry, consisting of diodes and resistors, is arranged in parallel to the subdivided sections of the magnet so that the energy is dissipated outside the magnet and voltages are limited. One feature of superconducting magnets is that they can be operated in persistent mode using a superconducting switch. This consists of a length of superconducting wire with a high-resistivity matrix, such as CuNi, connected across the terminals. This link can be driven normal with a heater while the magnet is being energized, and at the desired field, the heater is turned off. The power supply can then be dispensed with, as the current is flowing in the windings and the link persistently. This is particularly useful in NMR magnets where very high temporal stability is required.

There are possibilities that superconducting magnets may soon be built for fields well in excess of 20 T. The Chevrel-phase materials, such as $PbMo_6S_8$, have very high B_{c2}'s (up to 60 T) so that, in principle, they could be used for very high-field magnets. So far, attempts to produce practical conductors with sufficient current densities have failed, but work is still going on at several research centers. Probably more promising are the so-called high-temperature superconductors (HTS). Discovered by Bednorz and Müller in 1986, these are superconducting oxides with high critical temperatures (T_c), such as YBaCuO ($T_c \approx 90$ K), BiSrCaCuO ($T_c \approx 105$ K), and TlBaCaCuO ($T_c \approx 125$ K). All of these are superconducting in liquid N_2 (77 K). However, at liquid helium temperatures, these materials have B_{c2}'s of order 100 T and so they could feature in very high-field magnets. It will be some time before long lengths of practical high–current-density conductors will be available, but many programs of work are in train worldwide and results so far are encouraging. At higher temperatures, there is some doubt whether these materials can exhibit the necessary flux-pinning strengths for high-field applications, but for low-field work, such as power transmission and perhaps some NMR magnet systems, superconductivity in liquid N_2 is a very real prospect.

For an extensive treatment of superconducting magnets, see Wilson (1983); for high-temperature superconductors, see Bednorz and Müller (1990).

1.2.3 Hybrid Magnets The highest continuous fields available in the world today are generated using hybrid magnets. Hybrid refers to a system in which a superconducting magnet surrounds a high-power, water-cooled solenoid. The principle is simple and was first proposed by Wood and Montgomery who remarked that, as the field strength of superconducting magnets was intrinsically limited, they could be used in the outer, low-field region of an arrangement whereby the field of a conventional resistive magnet was supplemented by the superconducting section. In practice, the engineering problems are profound. First, the superconducting section has to have a room-temperature bore of up to 500 mm; therefore, necessarily, the superconducting magnet must be very big and its cryostat relatively complicated. The associated cryogenics are very large scale, and often a dedicated helium liquefier is needed. The inner, resistive section has to be compact in order to minimize the bore of the superconducting section, yet at the same time, it must be robustly engineered, as the forces on the windings are severe. Another problem is that the magnetic centers of the two magnets must be accurately aligned or a large (mostly axial) force would displace the inner magnet; in the case of asymmetric failure, this force has to be catered for. The polyhelix design is well suited to hybrid magnets as, in the case of asymmetric failure, the out-of-balance force is restricted to one layer of windings, but Bitter magnets are used or combinations of both.

Inevitably, hybrids are expensive installations and are operated in only a few centers. The first operating hybrid was at Oxford and gave 20 T. At the Francis Bitter National Magnet Laboratory, Grenoble Hochfeld-Magnetlabor, Magnetenlaboratorium at Katholieke Universiteit Nijmegen, and Tohoku University, Japan, there are hybrids generating at least 30 T. At the National Research Institute for Metals, Tsukuba, Japan, a 40-T system is under construction, and at the U.S. National High Magnetic Field Laboratory, a 45-T magnet is planned.

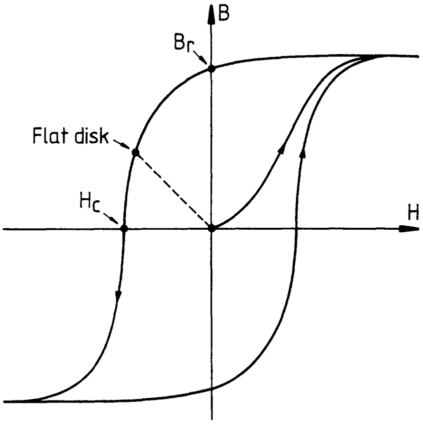

FIG. 3. Schematic of a magnetic hysteresis loop. The surface inside the loop is the energy transformed into heat in one cycle. Also shown are the "virgin" curve of the first magnetization and the line $M = -H$ related to the demagnetizing field in a thin sheet. For a ferromagnet, the values of B and $\mu_0 M$ are nearly identical.

2. PERMANENT MAGNETS

A permanent magnet (Parker, 1990) is an artifact made from a material that can be permanently magnetized when exposed to a magnetic field. The useful field from a permanent magnet can only be accessed outside the material, and therefore, a typical applications magnet comprises a shape such as a ring with an air gap. The field in the air gap can be approximated by

$$\mu_0 H_a^2 = H_c B_r V_m / V_a, \tag{6}$$

where H_c is the coercive field needed to reduce the magnetization of the material to zero, B_r is the remanent magnetic induction (see Fig. 3), V_m is the volume of the magnetic material, and V_a is the volume of the air gap. This argument applies strictly for a rectangular hysteresis loop, but for a real loop with rounded corners, it is the maximum value of the product $H \times B$ that enters on the right-hand side of Eq. (6). This is known as the energy product and is a figure of merit for permanent-magnet materials (Table 1). Between 1880 and 1990, the energy product has increased exponentially from 2 to 400 kJ/m^3 as new materials were developed. In the last 10 years, progress has been such that permanent magnets are now a \$10^9/a industry.

There are basically five families of materials; they are, in order of increasing merit, steels, Alnico, ferrites, samarium-cobalt alloys, and the rare earth-iron-boron compounds. A promising new class of materials now under development comes from the intermetallic nitrides (Wallace and Huang, 1992).

The energy product is a general measure for the magnetic field that can be sustained outside the magnet. For a specific magnet geometry, the outside field is related to the demagnetizing factor. This is the same as discussed with the example of an electromagnet in the shape of a ring with an air gap [Eq. (2)]. Instead of an outer winding carrying current, the permanent magnetization is now sustaining the magnetic field. The demagnetizing factor can be calculated exactly

Table 1. Range of basic data for different classes of permanent magnets.

Material	B_r (T)	H_c (A/m)	$BH_{\max}$ (kJ/m^3)
Steel (3% Cr)	0.98	4	1.8
Steel (35% Co)	0.9	20	7.5
Alnico	0.71	44	14
	1.24	50	44
	1.06	120	73
ferrites	0.23	145	8.4
	0.34	275	22
	0.41	225	32
RE-Co alloys[a]	0.75	560	110
	0.95	720	175
RE-TM-B alloys[a]	1.02	730	190
	1.15	875	240
	1.2	835	280

---

[a] RE represents a rare earth, and TM a transition metal. In most commercial materials, mixtures of different elements are used.

only for a few simple geometries; this is 0 for a long thin needle, $\frac{1}{3}$ for a sphere, and 1 for a thin flat plate. The last case is illustrated in Fig. 3, where a line with slope 1 corresponding to the relation $M = -H$ for a flat disk is drawn to intersect with the hysteresis loop. For a long needle, the corresponding line is the M axis, and for a sphere, it is the line $M = -3H$.

A great step forward in permanent magnet technology was the development of materials that retain their permanence over a long time scale and in the presence of mechanical shock and variations of external fields. This is of great importance for practical applications. The permanence of a magnet is characterized by the coercive field. This depends on several factors and mechanisms. To minimize the total energy, a magnetic material subdivides into domains with random orientations of the magnetization (called Weiss domains). The size of the domains, in the range 10^{-6}–10^{-2} cm^3, is determined by the ratio between the energy in the magnetic field and the energy in the domain walls. If the material is subjected to an external magnetic field, the domain walls are moved to increase the size of those domains with the magnetization oriented parallel to the applied field. The energy needed to move the domain walls results in magnetic hysteresis (Fig. 3). Usually, the magnetization changes in discrete steps, called Barkhausen jumps. A large coercive field is obtained by introducing obstacles (e.g., lattice faults) that impede the movement of the domain walls. Another factor that enhances the coercive field is anisotropy, which results in a favorable direction of the magnetization. This is mainly due to the crystal structure (including the crystalline electric field, spin-orbit interactions, etc.), but it can be enhanced by the shape of the magnet. Anisotropy is most pronounced for ceramic magnetic materials that consist of an agglomeration of many small particles. The ideal situation is given by needle-shaped particles in which the easy direction of the magnetization is parallel to the needles and with the needles oriented in the direction of the global magnetization. Another effect that can result in a strong coercive field is the single-domain particle. Below a certain size of a magnetic particle ($\approx$10 nm for Fe up to 1 μm for some other materials), it is energetically favorable to be homogeneously magnetized in a single domain. An interesting example of this is given by so-called "magnetotactic" bacteria, which synthesize a string of magnetic particles sized just below this critical limit. The particles thus get spontaneously magnetized and serve as simple direct means of guidance for the bacterium by orienting it along the lines of the earth's magnetic field. A propulsion mechanism (flagellum) at one end of the bacterium then propels it along the field lines (Maret and Dransfeld, 1985).

The practical realization of magnetic materials with the desired properties is still largely an empirical process for which the theoretical arguments can only provide a rough guidance. Research and technology are in rapid development, and the latest results are extensively covered by the annual conferences on Magnetism and Magnetic Materials (e.g., Yelon *et al.*, 1993) and INTERMAG (e.g., Lord, 1992).

2.1 Applications

The dramatic improvement of permanent-magnet materials has resulted in a surge of practical applications. Following Parker (1990), these can be classified in four major areas depending on the physical principles involved.

2.1.1 Direct Use of the Magnetic Force

The simplest of these and probably the earliest application is the compass. Attraction/repulsion of the poles of a magnetized needle interacting with the Earth's magnetic field creates a torque that aligns the needle north-south, and this is the basis of the fundamental navigation aid. Other applications related to the compass are polarity indicators and devices that detect ferromagnetic objects.

There are a variety of applications that exploit mechanical force, ranging from simple novelties to major industrial uses such as magnetic clutches. Unlike the electromagnet, a permanent magnet cannot be simply turned on and off. Modification of the magnetic field is achieved by the movement of ferromagnetic plates with appropriate slits that either short-circuit the magnetic flux or guide it to the intended destination. A well-known example is the magnetic clamp used to fix objects to a steel table, which can be released by the easy turn of a lever.

Mechanical force applications depend on whether a simple holding operation is required, i.e., the magnetic device is in contact with a ferromagnetic material on which it is exerting a force, or whether the force is exerted over a distance. In the latter case, the design of the magnet depends on whether the pull is to be maximum at a large distance, and the design is adjusted accordingly. In some applications, a compromise between the two is required. A notable example is in the case of magnetic drum separators where sufficient force is required to be imparted to a ferromagnetic particle so that it is translated over a distance but at the same time a strong holding force is required to make the particle stick to the drum. A related application is the magnetic conveyor belt. Frictionless bearings for small loads can be realized with permanent magnets as well.

2.1.2 Applications Based on Faraday's Law Permanent magnets are widely used to convert electrical into mechanical energy or vice versa. One of the earliest applications was the dc dynamo, which can be used as a generator or a motor. Alternators, in which permanent magnets on the rotor induce a sinewave voltage on the stator windings, are an important applications area. Generally, permanent magnets are used for small motors and generators, while the large machines incorporate electromagnets. However, with the development of permanent-magnet materials such as $SmCo_5$ and NdFeB, use of permanent magnets for the excitation of large alternators is attractive in some applications where high power density is important; permanent magnets can be superior to wound excitation in these cases. The dynamic microphone employs permanent magnets in the field of which a coil moves as a result of the pressure of sound.

Eddy-current devices are a special case of Faraday's-law applications. In these, induced voltages in a conducting element sustain eddy currents that interact with the field of permanent magnets and thus give rise to a mechanical force. This principle is used in torque transmission equipment and damping devices. A common application is the tachometer, where a permanent magnet rotates in front of a conducting disc. The interaction with eddy currents induced in the plate results in a drag that is proportional to the velocity; this can be measured by the torque exerted on the disc.

2.1.3 Lorentz Forces on Electrical Currents This category is based on the principle that is the most basic in converting electrical energy to mechanical energy. It relies on the interaction of current and magnetic flux and is essentially applied in electric motors, although one of the earliest applications of the effect is in loudspeakers. Basically there are four families of electric motors: dc, ac, hybrid, and linear, although subdivisions of these have led to at least 50 different types. In principle, all of these can employ permanent magnets in either the stator or the rotor. The present-day choice of permanent magnetic materials with greatly enhanced properties allows improved motor design in all areas, especially miniaturization. Examples are the very tiny motors used in cameras and clocks.

For linear devices, a modern volume application is in linear force devices in computer peripherals, e.g., printers and disc drives. The high volumetric efficiency of permanent magnets has made a major contribution to the reduction of size, response time, and thermal losses. The increase of digital electronics has led to considerable use of the stepper motor, which is a discrete positioning device (see MACHINES, ELECTRICAL). It uses the field of a permanent-magnet rotor that interacts with the stator poles, of which there can be a great number, and therefore very precise control can be achieved.

The familiar moving-coil or moving-magnet measuring instruments (see ELECTRICAL MEASUREMENTS) require permanent magnets with excellent long-term stability. Another use of permanent magnets in electronics is in reed relays.

2.1.4 Lorentz Forces on Free Electrons A free electron interacting with a perpendicular magnetic field will execute a circular path (cyclotron orbit) of radius

$$r = mv/eB, \tag{7}$$

where m is the mass of the electron, v is its velocity, and e is the electronic charge. This principle, which applies, of course, to other particles, is exploited in many devices and instruments such as magnetrons, traveling-wave tubes, and mass spectrometers. There are also applications of permanent magnets in particle accelerators. A good example is the "wiggler" magnet in the free-electron laser, which consists of a large number of permanent magnets arranged to generate a magnetic field that undulates periodically in space. It appears feasible to use permanent magnets even for MRI.

3. PULSED MAGNETS

3.1 Submegagauss Fields (Nondestructive)

For generating the highest possible fields, the limitations imposed by electrical power and associated cooling are overcome by using pulsed magnets. The pulse duration is proportional to the energy dissipated in the pulse; this energy is limited by the heat capacity of the coil. For copper precooled by liquid nitrogen, this is of the order 100 kJ/kg. A typical value is 50 T in a 20-mm bore for 10 ms at 200 kJ. Most pulsed-field coils are powered by a capacitor discharge (typical voltage 5–20 kV); the resulting wave form is a damped sine wave. To extend the pulse duration, a short circuit can be applied across the coil at peak field (usually by means of diodes). This is called a "crowbar"; it results in an exponential decay with the time constant of the coil. For large magnets with a controllable wave form, power from a flywheel-driven generator or from the mains is used. The peak field is determined by the mechanical strength of the coil. At 50 T, the Maxwell stress $BH/2$ is 1 GPa, which corresponds to the yield strength of a fairly strong steel. Special alloys and fiber composites have an ultimate tensile strength up to 4 GPa. For a given coil construction and materials, the limiting peak field depends on the ratio α between the outer and inner diameter; an approximate scaling law is given by

$$B = g\sqrt{2\mu_0\sigma_h}, \qquad (8)$$

where σ_h is the hoop stress. For constant current density, the geometrical factor g is

$$g = \sqrt{2}(1 - 1/\alpha), \quad \alpha \geq 2, \qquad (9)$$

and for a current distribution optimized for constant stress throughout the coil, it is

$$g = \sqrt{\ln\alpha}. \qquad (10)$$

The optimized current density peaks at the inner and outer radii and therefore results in stronger heating, which reduces the pulse duration. A good alternative is a spatial distribution of reinforcements that results in constant stress throughout the coil. As high mechanical strength and good electrical conductivity are mutually exclusive, it is necessary to develop structurally optimized coils using a combination of different materials. Recent advances in materials technology have resulted in different materials with very high mechanical strength and reasonably good conductivity (Foner, 1986), notably copper wires with microfilaments of niobium or silver. The highest field obtained so far without coil destruction is about 70 T in a 12-mm bore; it is expected to approach the 100-T limit within the next decade or two. Latest results may be found in the proceedings of the conference "Frontiers in High Magnetic Fields" (Miura, 1994); an earlier survey of pulsed field facilities is contained in Herlach and Franse (1989).

3.2 Multimegagauss Fields (Destructive)

Above 100 T, coils are inevitably blown up by the magnetic forces. The magnetic field drives a shock wave into the material; the pulse duration is given by the time interval until the expanding coil is distended so much that the field/current ratio is severely reduced. The strength of a megagauss field and the associated natural pulse duration can be directly related to the particle speed of the magnetically driven shock wave in the conductor material. For copper, this is 4.3 km/s at 1000 T and 1.7 km/s at 500 T. Different methods have been developed to supply adequate power and energy and to cope with the coil destruction. This is comprehensively covered by the series of conferences dedicated to megagauss fields; the most recent of these was held at Albuquerque in 1992 (Cowan and Spielman, 1993); for a review, see Herlach (1985).

3.2.1 Capacitor-Driven Single-Turn Coils The most simple and straightforward method is the fast capacitor discharge into a small single-turn coil. Typically, 150 T can be obtained in a volume of 1 cm^3 for a pulse duration of 2 μs with a 100-kJ capacitor discharge, and 250 T in 0.25 cm^3. The internal inductance of the capacitor bank must be of the order 10 nH, and the current is of the order 2 MA. Although the coil destruction is quite violent, the sample space inside the coil is protected by the magnetic field to the extent that even cryogenic equipment or photographic emulsions placed inside the coil will usually survive the field pulse. Typical field and current wave forms are shown in Fig. 4, and a flash x-ray photograph of an exploding coil is illustrated in Fig. 5. It is remarkable that, at peak field, the coil structure is already breaking up violently, but this is not reflected in the field and current wave forms, which are surprisingly smooth and well suited for experimentation. This leads to the assumption that the current is taken over by the plasma formed by the electrical explosion of the copper, which is visible as a bright fireball that emerges before peak field is reached. This type of magnet is well suited for solid state experiments at cryogenic temperatures.

3.2.2 Magnetic Flux Compression

Higher fields require a strong concentration of energy; at 500 T, the energy density is 100 kJ/cm^3, corresponding to a pressure of 1 Mb = 100 GPa. This can be achieved by magnetic flux compression in a high-speed implosion; the speed must be slightly larger than the particle speed in the shock wave to allow for the diffusion of the magnetic field into the conducting wall that is compressing the field. To compress magnetic flux, an initial flux is injected into a metal cylinder (by diffusion or through a slit) which is then compressed in a time interval much shorter than the L/R time constant which governs the flux diffusion out of the cylinder. Originally, high explosives were used to drive the cylinders; very regular and reproducible implosions can be achieved with precision explosives. Fields in the range 200–2000 T were obtained at several laboratories (Los Alamos, New Mexico, U.S.A.; Kurchatov, Moscow, Russia; EURATOM, Frascati, Italy; CEA, Paris, France; and Foulness, U.K.). An alternative and promising method is the electromagnetically driven implosion. While the implosion speed obtained with chemical explosives is limited by the finite speed of sound in the detonation products, there is no such fundamental limitation in an electromagnetic implosion. So far, fields well in excess of 500 T have been obtained at the Institute for Solid State Physics, University of Tokyo, using a 4-MJ capacitor bank. The pulse duration of the peak field in all implosion systems is a fraction of a microsecond. A typical current and field wave form is shown in Fig. 6.

Another method, developed at Frascati and now used for experiments at Los Alamos, is the compression of magnetic flux from a fairly simple bellows-shaped structure into a massive single-turn coil. The peak field is limited

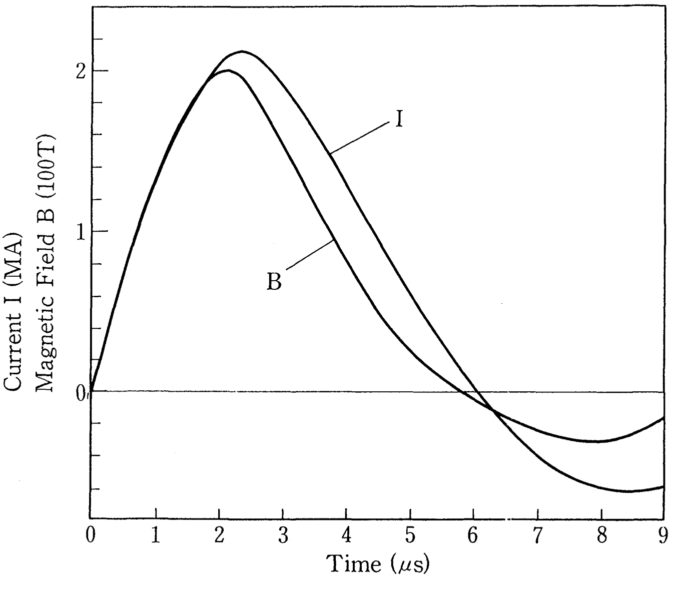

FIG. 4. Current and field wave forms from an exploding single-turn coil. The field maximum and decrease occur before those of the current, indicating the onset of expansion and destruction. The wave form is remarkably smooth and even shows the negative swing of a slightly undercritically damped *C-L-R* discharge, in spite of the violent destruction illustrated in Fig. 5. (Courtesy of N. Miura, ISSP, Tokyo.)

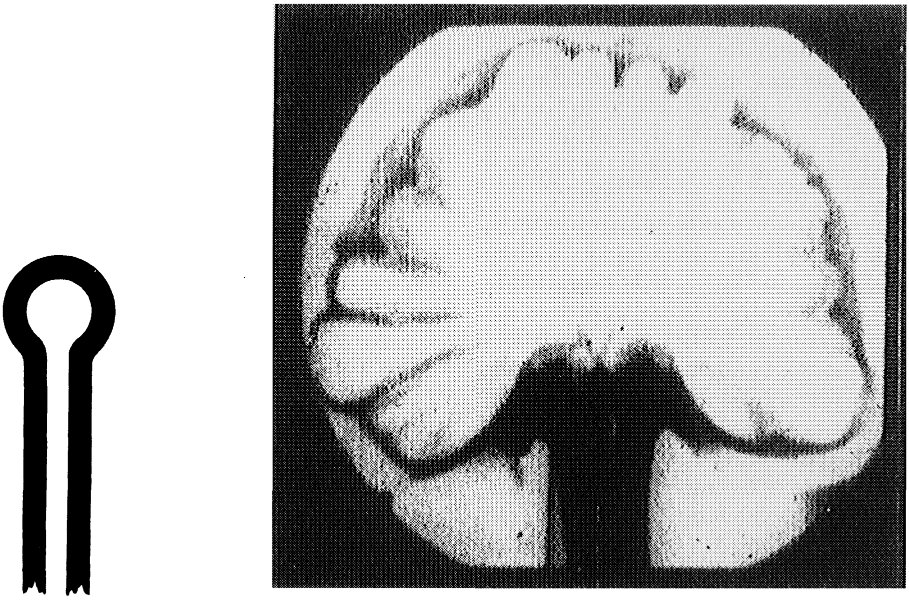

FIG. 5. Flash x-ray photograph of an exploding single-turn coil, taken at the end of a 180-T field pulse. The profile of the coil before the pulse is shown to the left. The coil, with inner diameter 8 mm, was made from a copper plate with 3-mm thickness. (Courtesy of N. Miura, ISSP, Tokyo.)

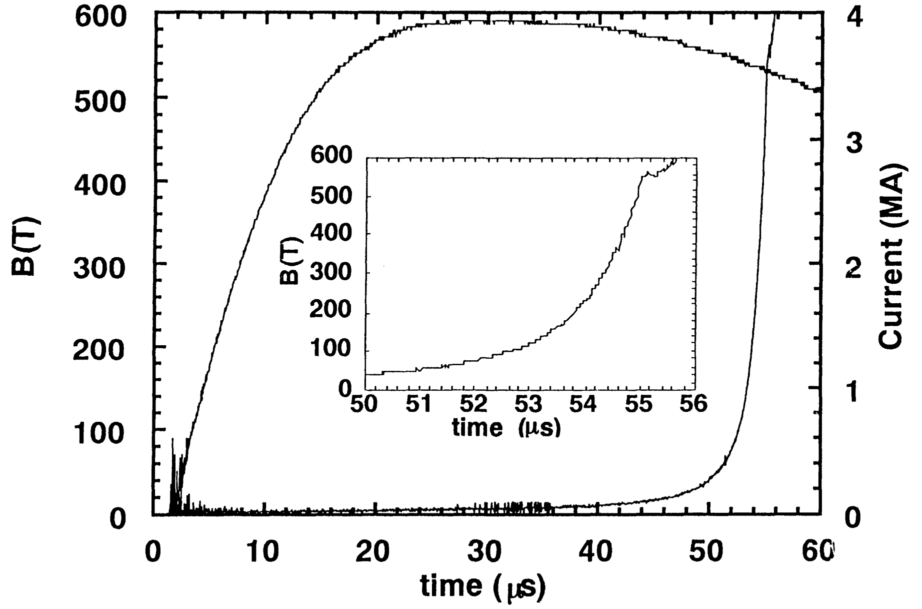

FIG. 6. Current and field wave forms from an electromagnetic implosion. The field wave form is typical for an implosion experiment, with a long flat initial part followed by a rapid increase and probe destruction shortly after the peak. This last part of the pulse is enlarged in the inset; the steps are due to the resolution of the digital recorder. (Courtesy of N. Miura, ISSP, Tokyo.)

around 200 T, but the pulse duration is longer than with the other methods, up to 10 μs.

4. MAGNETS FOR SPECIAL APPLICATIONS

4.1 Magnetic Separation

The separation of ferromagnetic from nonferromagnetic materials is straightforward and does not impose special requirements on the magnets other than that they must withstand a rough environment. There is a large variety of configurations depending on the material to be processed. More sophisticated methods are among the recurrent topics at the magnet technology conferences (e.g., Dawson, 1992).

For the separation of weakly magnetic (e.g., paramagnetic) materials, several methods have been developed. The most widely used of these is high-gradient separation, where a matrix of thin ferromagnetic wires is placed into a magnetic field of the order 1 T that magnetizes the wires perpendicular to their axis. This produces very strong field gradients in the vicinity of the wires, which attract small particles containing weak magnetic dipoles. These particles, usually suspended in a liquid for transport, stick to the wires and must be flushed out after accumulation; this can be a reciprocating or continuous process (e.g., with the matrix on a disk or a moving belt). Iron-core electromagnets are adequate and perform well in this industrial environment; large superconducting magnets have been used with variable success in large installations. These are good candidates for "open-gradient" separators where dry mixtures are processed by gravity feed into an open region where a magnetic field gradient exists.

Other methods are magnetic-field induced flocculation and enhanced sedimentation, and magnetically controlled sedimentation of particles suspended in a ferrofluid.

4.2 Magnetic Levitation and Propulsion

One application of magnets is in the area of mass transport and uses both magnetic levitation and propulsion (see TRANSPORT BY MAGNETIC LEVITATION). The advantage of levitating vehicles is that frictional losses associated with wheeled transport and considerable noise pollution are eliminated. At Birmingham airport in the United Kingdom, a fully commercial operational "people mover" system is installed. This uses conventional electromagnets to levitate the cars by attraction to a ferromagnetic rail. Japanese National Railways have developed a model Maglev train which has been operating on a 7-km test track at Myazaki prefecture for some 10 years, and a much longer track for commercial operation is under consideration for the line between Tokyo and Osaka. This system uses onboard superconducting magnets running in persistent mode. These magnets induce eddy currents in copper loops in the track beneath when the train is propelled by a linear motor (wheels are needed for very low speeds and when at rest). The magnetic field produced by the eddy currents opposes that of the superconducting magnets, and at a certain speed, the force generated exceeds the train's weight and it levitates. Speeds in excess of 500 km/h have been achieved.

With high-temperature superconductors being developed, there is renewed interest in an alternative approach, suggested by Bevir, in which the intrinsic diamagnetism and induced currents in bulk superconducting material can be made to react against a ferromagnetic track (Homer *et al.*, 1977). Also, the long-term possibility of superconducting magnets operating in liquid nitrogen may improve the efficiency of levitated transport using such magnets by reducing refrigeration requirements.

Again, in Japan, a novel form of marine transport has been developed that also uses superconducting magnets. The magnets are long dipoles through the air gap of which sea water is allowed to flow fore-to-aft. Because of the salinity of the sea water, electric currents can be passed orthogonal to the magnetic field and the direction of flow; the magnetohydrodynamic effect results in a force on the water, which is thus expelled from the stern of the ship with consequent forward momentum imparted to the vessel. A successful test ship has been launched.

GLOSSARY

Coercive Field: The magnetic field needed to reduce the magnetization of the fully magnetized material to zero.

Critical Current: The current in a superconductor at a given magnetic field and temperature above which it begins to revert to the normal state. For practical purposes, the definition is arbitrary and usually stated as that current which generates a given electric field along the conductor, typically 10 μVm^{-1}.

Critical Temperature: The temperature above which a superconductor becomes normal.

Cryostat: Alternatively, a Dewar vessel or "Dewar." A vessel (often constructed of stainless steel), usually insulated by vacuum enclosed by a double wall, for holding cryogenic fluids, e.g., liquid helium. Shielding against room-temperature radiation is usually incorporated.

Curie Temperature: The temperature above which a ferromagnet or ferrimagnet becomes paramagnetic.

Ferrimagnet: Similar to a ferromagnet but with magnetic dipoles of different strength oriented in opposite directions, thus also exhibiting spontaneous magnetization.

Ferrite: A material containing iron oxides that is ferrimagnetic, of practical importance because of its low electrical conductivity. There is a variety of ceramic compositions including soft and hard magnetic materials.

Ferromagnet: A material that exhibits spontaneous magnetization (with all dipoles in a domain parallel) below a certain temperature, called the Curie temperature.

Flux Jumping: The phenomenon whereby magnetic flux in a superconductor changes suddenly resulting in a temperature rise.

Flux Pinning: The mechanism whereby flux vortices threading a type II superconductor in the mixed state are prevented from movement under the Lorentz force.

Hard Magnetic Material: A ferromagnet or ferrimagnet with a high coercive field.

Lambda Point: The temperature (2.186 K) at which liquid helium becomes superfluid and has a thermal conductivity some 10^6 times that above the lambda point.

Lorentz Force: The force exerted by an electric field **E** and and a magnetic field **B** on a particle with charge q and velocity $\mathbf{v}$; this is $q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$. For a current I flowing through an infinitesimal length of conductor $d\mathbf{l}$, it is $I\, d\mathbf{l} \times \mathbf{B}$.

Magnetic Dipole Moment: Electric current circulating in a loop represents a dipole moment equal to current times surface of the loop. The magnetic field of a dipole falls off with the third power of the distance.

Magnetic Field, *H*: The measure of the magnetic excitation due to electric currents flowing in conductors; the unit is A/m.

Magnetic Flux, *Φ*: The surface integral of the magnetic induction. The voltage induced in a closed loop encircling the flux is equal to the time derivative of the flux (Faraday's law). The unit is weber = V s.

Magnetic Induction, *B*: The magnetic force field, which is also the source of induced electric fields (Faraday's law); the unit is T (tesla) = V s/m^2 (magnetic flux density).

Magnetic Permeability in Vacuum, μ_0: The ratio B/H in vacuum, which is equal to $4\pi \times 10^{-7}$ V s/A m by definition of the unit ampere.

Magnetic Permeability, Relative, μ: The ratio $B/\mu_0 H$ in a material. For a ferromagnet, this is a function of the magnetic field and history; it is usually taken as the tangent to the $B(H)$ curve.

Magnetic Resonance Imaging (MRI): A specialized form of NMR mostly used in the medical diagnosis of living tissue, in particular, the human body.

Magnetization, *M*: The measure of the magnetic excitation due to magnetic dipoles in a material; the unit is A/m = dipole moment per volume.

Maxwell Stress: The stress (force/surface) $BH/2$ associated with the magnetic field. This is positive (pushing) perpendicular to the field lines and negative (pulling) along the field lines. The Maxwell stress is equal to the energy density of the magnetic field.

Nuclear Magnetic Resonance (NMR): An effect whereby radio frequency radiation interacts with quantized nuclear magnetic moments in a magnetic field. For most applications, the field must be highly homogenous. The effect can be used for the study of chemical compounds.

Paramagnet: A material containing permanent magnetic dipoles that are not strongly interacting with each other. The relative permeability is slightly larger than 1, by an amount of order 10^{-4}. An exception is a ferromagnet above the Curie temperature, which is also considered paramagnetic.

Remanent Induction: The magnetic induction of a fully magnetized material after the magnetizing field is reduced to zero. This is proportional to the spontaneous magnetization.

Soft Magnetic Material: A ferromagnet or ferrimagnet with a small coercive field.

Upper Critical Magnetic Field, B_{c2}: The magnetic field that delineates the upper limit of the mixed state of a type II superconductor at a given temperature and current density.

Works Cited

Asimov, I. (1982), *Asimov's Biographical Encyclopedia af Science and Technology,* New York: Doubleday.

Bednorz, J. J., Müller, K. A. (Eds.) (1990), *Earlier and Recent Aspects of Superconductivity,* Berlin: Springer.

Cowan, M., Spielman, R. B. (Eds.) (1993), *Proceedings of the Sixth International Conference on Megagauss Magnetic Field Generation and Pulsed Power Applications, Nov. 8–11, 1992, Albuquerque, NM,* New York: Nova Science Publishers.

de Klerk, D. (1965), *The Construction of High-Field Electromagnets,* Newport Pagnell: Newport Instruments; also published in *Nederlands Tijdschrift voor Natuurkunde* **26**, 1, 65, 345; **27**, 353.

Dawson, A. (Ed.) (1992), "Proceedings of the 12th International Conference on Magnet Technology, Leningrad, June 23–28, 1991," *IEEE Trans. Magn.* **MAG-28**, 1 and 96–966.

Erber, T., Fowler, C. M. (Eds.) (1969), *Francis Bitter, Selected Papers and Commentaries,* Cambridge, MA: MIT Press.

Foner, S. (1986), *Appl. Phys. Lett.* **49**, 982–983.

Herlach, F. (Ed.) (1985), *Strong and Ultrastrong Magnetic Fields,* Topics in Applied Physics Vol. 57, Berlin: Springer.

Herlach, F., Franse, J. J. M. (Eds.) (1989), "Proceedings of the International Symposium on High Field Magnetism including a Review of High-Field Facilities Worldwide, Leuven, Belgium, July 20–23, 1988," *Physica* **B155**, 1–443.

Homer, G. J., Randle, T. C., Walters, C. R., Wilson, M. N., Bevir, M. K. (1977), *J. Phys. D* **10**, 879–886.

Lord, W. (Ed.) (1992), "Proceedings of the 1992 INTERMAG Conference, St. Louis, MO, April 13–16, 1992," *IEEE Trans. Magn.* **MAG-28**, 2097–3335.

Maret, G., Dransfeld, K. (1985), in: F. Herlach (Ed.), *Strong and Ultrastrong Magnetic Fields,* Topics in Applied Physics Vol. 57, Berlin: Springer.

Miura, N. (Ed.) (1994), "Proceedings of the Todai Symposium 'Frontiers in High Magnetic Fields', Tokyo, Nov. 10–12, 1993," *Physica B* (to be published.)

Miura, N., Herlach, F. (1985), in: F. Herlach (Ed.), *Strong and Ultrastrong Magnetic Fields,* Topics in Applied Physics Vol. 57, Berlin: Springer.

Montgomery, D. B. (1969), *Solenoid Magnet Design,* New York: Wiley-Interscience.

Oberteuffer, J. A. (1974), *IEEE Trans. Magn.* **MAG-10**, 223–238.

Parker, R. J. (1990), *Permanent Magnetism,* New York: Wiley Interscience.

ter Haar, D. (1964) *The Collected Papers of P. Kapitza,* Oxford: Pergamon Press.

Wallace, W. E., Huang, M. Q. (1992), "Magnetism of Intermetallic Nitrides: A Review," *IEEE Trans. Magn.* **28**, 2312–2315, and following papers.

Wilson, M. N. (1983), *Superconducting Magnets,* Oxford: Clarendon Press.

Yelon, W. B., Dahlberg, E. D., Nix, J. L., Pinkerton, F. E., Victoria, R. H. (Eds.) (1993), "Proceedings of the 37th Annual Conference on Magnetism and Magnetic Materials, Houston, TX, Dec. 1–4, 1992," *J. Appl. Phys.* **73**, 5309–7023.

Further Reading

Books and Articles

Brechna, H. (1973), *Superconducting Magnet Systems,* Berlin: Springer.

Buschow, K.H.J. (Ed.) (1991), *Handbook of Magnetic Materials (formerly Ferromagnetic Materials),* Vol. 6, Amsterdam: North-Holland.

Buschow, K. H. J., Wohlfarth E. P. (Eds.) (1990), *Ferromagnetic Materials,* Vol. 5, Amsterdam: North-Holland.

Craik, D. J. (Ed.) (1975), *Magnetic Oxides (Part 1 and 2),* London: John Wiley & Sons.

Cyrot, M., Pavuna D. (1992), *Introduction to Superconductivity and High T_c Materials,* Singapore: World Scientific.

Doss, J. D. (1989), *Engineer's Guide to High Temperature Superconductivity,* New York: Wiley Interscience.

Ekin, J. W. (1983), *Superconductors,* in R. P. Reed, A. F. Clark (Eds.), *Materials at Low Temperatures,* Metals Park, OH: ASM, p. 645–513.

Fletcher, D. (1991), "Fine Particle High Gradient Magnetic Entrapment," *IEEE Trans. Magn.* **MAG-27**, No. 4.

Foner, S., Schwartz, B. B. (Eds.) (1981), *Superconducting Materials Science, Metallurgy, Fabrication and Applications,* New York: Plenum Press.

Heck, C. (1974), *Magnetic Materials and their Applications,* London: Butterworths & Co.

Knoepfel, H. E. (1970), *Pulsed High Magnetic Fields,* Amsterdam: North-Holland.

Kolm, H., Oberteuffer, J., Kelland, D. (1975), "High Gradient Magnetic Separation," *Sci. Am.* **233** (5), 46–54.

Kossowki, R., Raveau, B., Wohlleben D., Stamatis K. P. (1992), *Physics and Materials Science of High Temperature Superconductors, II,* NATO Advanced Study Institutes, Series E: Applied Sciences, Vol. 209, Dordrecht: Kluwer.

Kroon, D. J. (1968), *Laboratory Magnets*, Eindhoven: Philips Technical Library.

Nesbitt, E. A., Wernick, J. H. (1973), *Rare Earth Permanent Magnets*, New York: Academic Press.

Parkinson, D. H., Mulhall, B. E. (1967), *The Generation of High Magnetic Fields*, New York: Plenum Press.

Wohlfarth, E. P. (Ed.) (1980), *Ferromagnetic Materials*, Vols. 1 and 2, Amsterdam: North-Holland.

Wohlfarth, E. P. (Ed.) (1982), *Ferromagnetic Materials*, Vol. 3, Amsterdam: North-Holland.

Wohlfarth, E. P., Buschow, K. H. J. (Eds.) (1988), *Ferromagnetic Materials*, Vol. 4, Amsterdam: North-Holland.

Zijlstra, H. (1967), in: E. P. Wohlfarth (Ed.), *Selected Topics in Solid State Physics*, Vols. 1 and 2, Amsterdam: North-Holland.

Conferences

Blums, E., Gogorov, V. V., Charles, S. W., Rosensweig, R. E. (Eds.) (1990), "Proceedings of the 5th International Conference on Magnetic Fluids, Riga, Sept. 18–22, 1989," *J. Magn. Magn. Mater.* **85**.

Marinucci, C., Weymuth, P. (Eds.) (1985), *Proceedings of the 9th International Conference on Magnet Technology, Zürich, Sept. 9–13, 1985*, Villigen, Switzerland: SIN.

Kobe, S., Roth, S. (Eds.) (1991), "Proceedings of 'EMMA', European Magnetic Materials and Applications Conference, Dresden, Apr. 16–19, 1991", *J. Magn. Magn. Mater.* **101**.

Palmer, S. B. (Ed.) (1992), "Proceedings of the International Conference on Magnetism 1991, Edinburgh, Sept. 2–6," *J. Magn. Magn. Mater.* **104–107**.

Sekiguchi, T., Shimamoto, S. (Eds.) (1990), *Proceedings of the 11th International Conference on Magnet Technology, Tsukuba, Aug. 28–Sept. 1, 1989*, London: Elsevier Applied Science.

Williams, J. E. C. (chairman) (1988), "Proceedings of the 10th International Conference on Magnet Technology, Boston, MA, Sept. 21–25, 1987," *IEEE Trans. Magn.* **MAG-24**, 691–1598.

MAINTENANCE AND CALIBRATION OF LABORATORY INSTRUMENTATION

TIMOTHY J. MCINTYRE, *Instrumentation and Controls Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A.*

Introduction 261
1. **Calibration** 261
2. **Elements of Calibration** 262
2.1 Calibration Standards 262
2.2 Precision Measurement 263
2.3 Measurement Environment 264
2.4 Calibration Errors 264
2.5 Calibration Results 265
3. **Calibration Standards** 265
3.1 Dimensional Measurements 265
3.2 Mechanical Measurements 267
3.3 Thermodynamic Measurements 268
3.4 Optical Radiation Measurements 269
3.5 Ionizing Radiation Measurements 270
3.6 Electromagnetic Measurements 270
3.7 Time and Frequency Measurements 271
4. **Performing Calibrations** 272
4.1 Calibration Procedures 272
4.2 Calibration Setup 273
4.3 Making Measurements 273
4.4 Analyzing the Results 273
5. **Instrument Maintenance** 274
5.1 Standards Calibration 274
5.2 Calibration Intervals 274
5.3 Storage and Care of Standards 274
6. **Future Trends** 275
6.1 Modular and Miniature Instrumentation 275
6.2 Self-Calibrating Instrumentation 275
6.3 Transport Standards 275
Glossary 276
Works Cited 276
Further Reading 276

INTRODUCTION

Calibration and maintenance of laboratory instrumentation are important functional aspects of any manufacturing or research organization. Since the quality of any product, whether it is intellectual material or some manufactured good, relies on the level of conformance to specifications, one's ability to meet desired goals depends on the means for doing so. Measuring and test equipment is used throughout the manufacturing and development process, and hence its performance, or its ability to quantify the performance of a process, depends on a prescribed and regimented measurement assurance program. Calibration and maintenance organizations or individual researchers can provide the measurement assurance necessary to maintain, continually, high standards of productivity and customer satisfaction.

1. CALIBRATION

Calibration can be defined as "the set of operations which establish, under specified conditions, the relationship between values indicated by a measuring instrument or measuring system, or values represented by a material measure, and the corresponding known values of a measurand" (ISO/IEC, 1990). The philosophy of *exactly* what constitutes a calibration has been widely discussed. However, the common thread that pervades most discussions is that a comparison to a standard or reference instrument or some intrinsic physical constant must be made to certify the system or device under test (DUT). The reference instrument or intrinsic value establishes the measurand; understanding the relationship between the DUT indication and the known measurand constitutes the calibration. Of course, the characterization of a device or system involves a

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

myriad of operations including measurements, calculations, judgments, and finally presentation of the information in a meaningful way to the end user.

Overview: In order for any calibration process to be successful, many factors need to be considered. For large manufacturing or research organizations, it is very important to have an operationally structured calibration program—one that is organized to handle the flow of instruments, maintenance of calibration standards, concise and up-to-date documentation, and tracking of standards and DUTs to ensure that recall intervals and inventories are current and adequate. The calibration laboratory environment is a critical component in realizing measurement precision. Ergonomics facilitates efficient and safe measurement practice, while the simple lack of an organized and clean setup may prevent measurement accuracy. The proper standards are necessary to accommodate required accuracy ratios and standards hierarchy. Finally, measurement technique and knowledge or understanding of measurement circumstance can lend significant added value to the calibration.

For the individual researcher, who is concerned with validating a measurement process or quantifying measurement uncertainties, calibration is a critical component. Questions such as How do I calibrate and maintain the instrumentation being used on my project? or What standards do I need in order for my data to have traceability to standards of measurement? or How do I determine the error budget of my calibration process, and how does it relate to the uncertainty requirements of my experiments? need to be answered. The answers to these and many other technical concerns lie in the understanding of the measurement process, calibration methodology, and potential error sources and how to limit or control these error sources.

2. ELEMENTS OF CALIBRATION

In order to characterize measuring and test instrumentation properly, the calibration standard(s) and the measurement techniques employed must mimic the real operating environment. For example, calibrating a resistor in a well-maintained laboratory environment does not necessarily indicate resistor performance in a manufacturing application where the device may be exposed to high moisture levels, contaminating fluids like oils, or thermal variations vastly different from the laboratory environment. Therefore, the proper choice of calibration standards, measurement techniques, and environmental controls will optimize one's understanding of instrumental performance and how it might affect a measurement or research process.

2.1 Calibration Standards

The nature of calibration and the necessity for measurement traceability have led to the establishment of extensive hierarchies of standards laboratories and standards instrumentation. National standards laboratories are at the apex of these hierarchies, providing ultimate reference for every standards laboratory. At the field instrumentation level, the requirements imposed on calibration instruments in a manufacturing or maintenance environment are much less demanding than those imposed on a reference-level instrument or system customarily found in a metrology laboratory. Working standards are at the lowest stratum of the calibration instrumentation hierarchy. These instruments are utilized in daily, routine calibration to check and characterize measuring instrumentation used in research, development, manufacturing, maintenance, or general operation of facilities and equipment. These devices are generally multifunctional source or sense instruments which operate equally effectively as transportable or benchtop standards, or individual artifacts like resistors, thermometers, pressure gauges, or gauge blocks. At the next level are transfer standards; these are laboratory instruments that are used to calibrate working standards. Transfer standards are periodically calibrated by laboratory or reference standards and are unique in that they provide the short-term measurement precision to convey the accuracy of the reference-level standard to the working standard with negligible additional uncertainty. Typically, the reference standards are not available for routine calibration work. Rather, to protect such top-level standards from deterioration or misuse, the transfer-standard class instrument is implemented. The reference standards are the primary instruments or artifacts of most standards laboratories. These devices are carefully maintained and controlled to assure that performance levels remain sufficient to provide accuracy ratios for lower-level standards. Reference standards are highly specialized devices, which are expensive to recertify if proper care is not provided. All standards laboratories depend heavily on these standards to sustain traceability to international standards of measurement. Reference standards provide a direct link to fundamental standards or internationally recognized standards through interlaboratory comparison.

Fundamental standards are those quantities of the *Système International* (SI) whose fundamental physical nature has distinguished them as well-defined, naturally occurring processes. The triple point of water [measured in terms of the SI unit for temperature known as the kelvin (K); one kelvin is defined as the 1/273.16 of the thermodynamic temperature of the triple point of water] and the cesium-beam time standard (measured in the base SI unit of seconds, where 1 s is defined as the duration of 9 192 631 770 periods of radiation emitted by the transition between two hyperfine levels of the ground state of the cesium-133 atom) are two fundamental standards to which many measurements are traced. The platinum-iridium mass artifact (which defines the SI base unit of the kilogram) is a *de facto* standard to which mass measurements as well as many physical and electrical measurements trace their accuracy. There are, of course, other base physical units on which primary reference calibration systems rely. The set of base units includes mass, time, temperature, electric current [ampere (A)], amount of substance [mole (mol)], and luminous intensity [candela (cd)]. The length unit [meter (m)] is now a derived unit relating the distance that light travels in vacuum in a given number of cesium oscillations. Other derived units are force, voltage, resistance, and pressure, whose accuracy and ultimate traceability are derived from the fundamental standards, the kilogram artifact, and base units.

2.2 Precision Measurement

Calibration implies precision and accurate measurement. By virtue of the fact that comparison of a test instrument with a standard necessitates transference of measurement accuracy, minimizing uncertainty and maximizing accuracy require attention to both system configuration and system performance.

System configuration involves the use of standards and measurement system architecture as they relate to maximizing calibration performance. Depending on the accuracy and precision needed to fulfill measurement goals, one may choose to employ standards further up the hierarchy. In doing so, measurement techniques may need to be altered, greater care must be exercised, and time and cost may be somewhat compromised. Therefore, clearly defined goals and objectives for a test or calibration are important for optimizing system performance (e.g., reference standards are not necessary for a 1% accuracy level calibration).

System performance impacts the overall quality of the measurement product as much as system configuration, but performance characteristics are in some sense intensive quantities of the overall system, whereas configuration and choice of the appropriate standards are extensive quantities. Accuracy and uncertainty of a measurement system can be influenced by many things: environmental factors, calibration errors, data acquisition errors, and data reduction errors. Clearly, environmental factors can be controlled by proper construction of the facility and careful operational methodology. Calibration and data contributions can be minimized by employing statistical process control algorithms.

If the output data are to yield a meaningful description of the measurement process, the data must form what is called a random sequence. In other words, the process must be in a state of statistical control. First, note that it is meaningless to speak of the accuracy of an instrument as an isolated device. One must always consider the instrument plus its environment and method of use (i.e., the instrument and its inputs). This aggregate system constitutes the measurement process. Every instrument has practically an infinite number of inputs—that is, causes that can conceivably affect the output, however small. Such effects as atmospheric pressure, temperature, and humidity are among the obvious factors, but there are a number of others if one is willing to concede that some only have a very small influence. In defining the calibration procedure for a specific instrument, one specifies that particular inputs be held constant within certain limits. These inputs, it is hoped, are those that contribute the largest components to the overall error. The other influences are left completely uncontrolled, and through the use of process control charts, it can be shown that each contributes only a very small amount of error and that their aggregate influence on the instrument output will be of a random nature. If this is indeed the case, then the measurement process is said to be in statistical control. Lack of control is often obvious if one takes repeated measurements and plots the results. Any statistically significant trends in the data will indicate that an uncontrolled input to the measurement process has contributed more than a negligible amount to the overall error.

2.3 Measurement Environment

The measurement or calibration environment can be vastly different depending on the level of accuracy required and the circumstances under which the calibration takes place. In any case, environmental conditions can introduce gross errors into the measurement process if not controlled or at least understood. Sensitive electrical measurements are influenced by thermal gradients (air currents in the room or gradients that manifest themselves as emfs or contact potentials because of the thermocouple effect between dissimilar metal contacts) and changes in relative humidity. Pressure measurements can also be affected by thermal gradients since standard pressure devices often derive pressures from force-per-unit-area–type configurations. Hence, environmental controls should be implemented to sustain temperature and relative humidity (and sometimes particulate contamination) levels commensurate with the calibration objectives. For laboratory-type instrumentation, this generally equates to a nominal temperature of 22.4 to 23.6 °C, with variations no greater than 0.6 °C in a 1-h period, and relative humidity controlled between 35 and 55%. For occasions where these levels of control are not possible or practical, environmental conditions should be noted, and instrumental uncertainty calculations should include any significant contributions from these effects.

2.4 Calibration Errors

All measurements have errors. The errors arise from indicated differences between the measuring device being calibrated and the true value being supplied by the calibration standard. The total error of any measurement has two components: a fixed bias error β and a random precision error ϵ_k, such that

$$\delta_k = \beta + \epsilon_k.$$

Since all measurement systems are subject to both bias and precision errors, it follows that an accurate measurement is one that has both small bias and small precision errors. These bias and precision errors are commonly referred to as calibration errors, but other error sources such as data acquisition errors and data reduction errors loom as potential pitfalls to achieving calibration objectives.

Bias (fixed) errors are systematic errors that remain constant for the duration of the test. In repeated measurements employing the same measurement technique and at ambient conditions, each measurement has the same bias error. The bias, however, is usually an accumulation from several individual bias errors. Bias errors include those that are known and can be calibrated out, those that are negligible and are ignored, and those that are unknown and are estimated. The estimated component must be taken into account when performing the uncertainty analysis. In most cases, the bias error is equally likely to be positive or negative. Hence, it is not known if the bias error is plus or minus, and so the bias limit must reflect this as $\pm B$. The bias limit B is estimated as the upper limit of the bias error β. The bias error is difficult to estimate, and so special tests must often be devised to provide bias information. The following tests are useful for obtaining bias information:

1. interlaboratory or interfacility comparisons, which make it possible to obtain the distribution of bias errors between laboratories;
2. comparison of standards with test instruments in the actual test environment;

3. comparison of independent measurements that depend on different principles or have been independently calibrated;
4. when a bias is known to result from a particular cause, calibrations in which the cause is allowed to vary throughout its full range to determine the range of the bias.

In the event that there is no known cause or source of data for bias, an estimate must be made based on judgment or experience. Instrument manufacturers or other references may provide some information.

Precision (random) errors are observed in repeated measurements, which do not and are not expected to agree exactly because of numerous error sources. The standard deviation is a measure of the precision error. For a normal (bell-shaped) distribution, the interval $\mu \pm 2\sigma$ will include approximately 95% of the total scatter in the measurements where μ is the true average of the population and σ is the true standard deviation. A large standard deviation means large scatter in the measurements. The statistic S is calculated to estimate the standard deviation. S is known as the precision index and is defined as

$$S = \sum_{k=1}^{N} (X_k - \bar{X})/(N - 1),$$

where N is the number of measurements made and $\bar{X}$ is the average value of the measurements X_k, given by

$$\bar{X} = \frac{1}{N}\sum_{k=1}^{N} X_k.$$

The effect of precision error is reduced by averaging several measurements and using the average instead of any individual measurement. Thus,

$$S_k = S/\sqrt{N}.$$

2.5 Calibration Results

The results of the calibration or test become an important historical reference document. Calibration reports are useful for predicting instrument maintenance requirements, tracking error budgets, adjusting calibration intervals, and satisfying auditory requirements, and generally as a reference for system accuracy and uncertainty. However, for the calibration results to be truly useful, several important pieces of information must be included such as

1. the date of calibration;
2. reference to the calibration procedure employed;
3. the environmental conditions present at the time of calibration;
4. the calibration standards employed;
5. indication of whether adjustments were required to bring the instrument into specifications;
6. calibration data in the "as found" condition and in the "as left" condition if the instrument was adjusted;
7. indication of the calibration errors; and
8. information about the individual or organization who performed the calibration.

Each of these items is critical to maintaining an instrument's performance history and legitimizing traceability to higher-level standards.

3. CALIBRATION STANDARDS

Calibration standards exist in a wide variety of measurement disciplines. They are employed to provide an accurate and precise measurand relative to which a calibration is made. The accuracy and precision that can be realized depend on the measurement discipline of concern and also at what level in the calibration hierarchy the measurement is being done. At the national level, calibration standards have been segregated in terms of physical discipline as shown in Table 1. These measurement disciplines have well-characterized, stable, and statistically controlled calibration systems for which traceability to the national level is available. However, in a research environment it is often necessary to develop specialized measurement systems, comprised of one or more working standards, to relate less common physical processes or unique physical observables to known measurement parameters.

3.1 Dimensional Measurements

Until recently, all dimensional measurements were tied to the krypton-86 funda-

Table 1. Listing of calibration disciplines.

- Dimensional measurements
  - Length
  - Texture
  - Volume/density
- Mechanical measurements
  - Flow/airspeed
  - Mass
  - Force/acceleration
  - Acoustic pressure/emission
- Thermodynamic measurements
  - Pressure
  - Temperature
  - Humidity
- Optical radiation measurements:
  - Luminous intensity
  - Spectral characteristics
  - Laser power and energy
- Ionizing radiation measurements
  - Radioactive sources
  - Dosimetry
- Electromagnetic measurements
  - Resistance
  - Impedance
  - Voltage
  - Current
  - Power and energy, low frequency
  - Microwave power, frequency, phase
  - Electromagnetic field strength
  - Pulse waveform
- Time and Frequency measurements
  - Time base
  - Frequency
  - Phase

mental meter standard. Now the meter is traced to the cesium fundamental time standard. There are innumerable secondary and tertiary standards in this area of calibration. Perhaps the most common and well-known length standard is the gauge block. Gauge blocks are generally useful for performing calibration of measuring and indexing equipment with accuracy requirements in the 1 m to 1 mm range. Since the dynamic range of gauge block standards is fairly limited, many interferometry techniques have become more practicable in recent years, offering calibration services from greater than 100 m to less than 1 μm with accuracies in excess of 1 part in 10^7 at the helium-neon wavelength. Many artifact standards exist in the size and length realm, from polystyrene spherical particulates of a few micrometers to straight edges and cylinders up to 2 m in length. Surveying and gauging tapes, leveling rods, sieves, and thread, gear, diameter, angle, and line standards are just a few of the many dimensional artifact standards that support precision tolerance manufacturing.

At the transfer or working level, the computer numerically controlled coordinate measuring machine (CMM) has become the workhorse for part and process validation and certification. These machines rely heavily on reference artifacts for their calibration and are used to certify low-level artifacts that would be employed in a manufacturing process. The most accurate CMMs today, once certified by an artifact and sufficiently error mapped, can provide calibration measurement services over a large operating volume (1 m^3) with accuracies of less than a micrometer. The CMM is a very versatile machine. Gears, threads, straight edges, plug or gauge diameter, ball bearings, and step standards are just a sampling of the geometric dimensional artifact standards that can be calibrated with this technique.

In the ultraprecision regime, many of the calibrations performed are really at the developmental stage of maturity. Less conventional interferometry techniques like Fabry-Pérot interferometry or x-ray interferometry can be used to calibrate with extreme precision. These techniques are usually traceable directly to the cesium time standard or rely on fundamental constants of nature (silicon lattice constant) in order to achieve the accuracy levels needed to calibrate more conventional heterodyne or phase-shifting interferometers, capacitance gauges, and linear variable differential transformers.

The aspect of dimensional calibration that deals with characteristics and properties of surfaces is called surface texture metrology. There are three general categories of surface texture calibrations: those that employ specimens with regular geometric profiles (periodic sinusoids or gratings, etc.), roughness specimens (random texture), and step-height specimens. Information pertaining to surface rms or average roughness, average wavelength, autocorrelation function, or power spectral density function is usually sought. Stylus-type profilometers, laser profilometers, or interferometers are commonly used to certify texture artifacts as secondary or interlaboratory standards.

Volume standards and graduated scales can be calibrated by gravimetric techniques through which direct comparison to mass standards takes place. Secondary or transfer standards are typically well-characterized volumes. Once a volume is determined via the gravimetric techniques with fluids of known densities, the predetermined volume can then be used to certify materials of unknown density via the gravimetric technique.

3.2 Mechanical Measurements

Mechanical calibration systems all derive their accuracy from the fundamental standards of mass and time. Whether flow rate, flow speed, force, or acceleration is of interest, each depends directly on or is related to mass or, in the case of a speed or rate, depends on time.

Two general approaches are used to measure flow. One involves the use of some simple volume flow meter, some means for measuring density, and a computer to calculate mass flow rate. The other, more basic, approach is to find flow metering concepts that are inherently sensitive to mass flow rate (see MEASUREMENT AND INSTRUMENTATION OF FLOW). There are several sensors or transmitters available with signal output that depends on an attenuation of signal strength as a function of fluid density. These sensors can be instrumented on volume flow devices to measure density concurrently and thus provide mass flow information. However, these secondary-level techniques depend on gravimetric methods for higher-level calibration, where a given amount of fluid is allowed to accumulate in a holding tank and the change in tank mass is monitored per unit time. Hence, this technique relies on accurate mass reference standards as well as a precision time base to achieve sufficient accuracy to satisfy current secondary and working standards accuracy ratio requirements.

Studies of fluid-flow phenomena attempt to determine the magnitude and direction of the fluid velocity vector at a point within the medium. In practice, measuring instrumentation can only sample the average conditions over a small area or volume. The pitot-static tube is used as a reference standard in wind tunnel facilities to determine flow velocity, except when flow velocities are not sufficient to produce large enough differential pressures in the tube. At lower flows, laser velocimetry is employed. Air density in the wind tunnel must also be determined from pressure, temperature, and humidity data to assure that measurement errors are minimized. Hot-wire and hot-film anemometers are common working standards as well as spinning-rotor and Venturi-tube gauges. Laminar flow elements and nozzle/orifice type gauges are also in widespread use.

Mass is considered a base unit, and the artifact standard cylinder of platinum-iridium, called the International Kilogram, is kept in a vault in Sèvres, France. Other masses, such as national standards, are compared with this standard via the equal-arm balance technique. At the 1-kg level, measurement precision of a few parts in 10^9 is possible (see MASS AND DENSITY, MEASUREMENT OF). Secondary mass standards generally take the form of single weights or weight sets, which are then used to calibrate working standard weight sets. These weights are then used to certify secondary or working standard scales or digital electronic balances. Standard weight sets are widely used in other branches of calibration and metrology. Dead-weight pressure testers and some force and acceleration measuring equipment rely on mass calibration, as they are derived quantities.

Acceleration is not a base unit but rather is derived from length and time. Length is now derived from the fundamental time standard; hence, acceleration is traceable only to the cesium clock. Acceleration of gravity g is a convenient reference standard, which can be determined with an accuracy of about 1 part in 10^6 by measuring the period and effective length of a simple pendulum. It can also be determined by measuring the rate of change of the speed of a freely falling body. However, in practical terms acceleration or vibration is measured by electrically monitoring the displacement or frequency and amplitude of a displacement of a known mass. Calibration of such a device involves characterizing the internal spring forces or determining the sensitivity and frequency response of the displacement sensor. The instrumental transfer function, which for a mechanically coupled sensor amounts to determining the ratio of the electrical output to the mechanical input, describes the device performance in terms of the frequency and magnitude of the vibration exciter.

Force is defined by Newton's second law, $F = MA$; thus, a standard for force depends on standards for mass and acceleration. When the gravitational constant g has been determined for a particular locality (pendulum technique), the gravitational force (weight) on accurately known standard masses may be computed to establish a force standard. This is the basis for the so-called "dead-weight" calibration of force-measuring systems. Dead-weight calibration can be employed to certify secondary standard load cells or proving rings. Several other techniques are recognized as having sufficient accuracy for field calibration: analytical balance against the gravitational force on a known mass, acceleration measurement of a known mass to which the unknown force is applied, balancing against magnetic forces created by current flowing through a coiled wire, or conversion of the force to a fluid pressure via a piston gauge or similar transducer. Finally, once force measurement capability is established, torque standards can be realized by sufficiently defining the length of a moment arm to which a known force is applied.

Measurement of air-borne and water-borne sound is generally performed using microphones or hydrophones, which are basically pressure-measuring devices (see MEASUREMENTS, ACOUSTICAL). Sonic measurement involves sensing the magnitude of the fluctuating component of pressure in some fluid medium. The rms value of the fluctuating component of pressure is employed because most sounds are random signals rather than periodic functions. Sound-level meters, microphones, or pistonphones are calibrated by comparison with standard microphones or by assessing the open-circuit voltage response to pressures applied uniformly over the diaphragm of the sensor.

3.3 Thermodynamic Measurements

The principal thermodynamic quantities of interest to a calibration laboratory are pressure, temperature, humidity, and heat flux. These quantities are ultimately traceable to the triple point of water and the mass standard, but an array of reference standards is utilized in routine calibration.

Pressure is not a base unit but rather is derived from force and area, which in turn are derived from mass and time. The basic calibration standards for pressure ranging from medium vacuum up to several hundred thousand pounds per square inch are the mercury manometer and the dead-weight piston gauge. For pressures ranging further down in vacuum, the McLeod gauge is considered the standard. Hot-cathode or radioactive-ionization gauges as well as molecular drag gauges are useful for exploring the high-vacuum regime, but each requires a spot check against the McLeod gauge and characterization of instrumental linearity to provide useful high-vacuum measurements. Since pressure can easily be transduced to force by allowing it to act on a well-characterized area, the basic method for measuring force applies to pressure except for the high-vacuum regime.

Temperature can be considered to be a measure of the *level of thermal energy* of a system given by the average of the random kinetic energy of all constituents of the system. The zeroth law of thermodynamics also represents a useful concept. Thus, if temperature is thought of as the level of thermal energy, it can be considered the driving force for heat flow just as voltage is for electrical current flow. Fixed points of temperature are used to establish a reproducible means of generating a range of temperatures (see TEMPERATURE MEASUREMENT AND CONTROL). Calibration is subsequently performed by comparing unknown temperatures to standard temperatures through the use of a thermometer. The most common thermometer employed as a calibration standard is the platinum resistance thermometer (PRT). Generally, the PRT is allowed to come into equilibrium with some number of temperature-fixed points, and its accuracy and linearity are noted. The calibration process amounts to measuring the resistance of the PRT once it has come into equilibrium with the thermal reservoir. Resistance measurements can be made directly using an ac bridge network or by exciting the PRT with a low-level dc current (1 mA or less) and detecting the voltage drop across the resistance element. The latter technique involves measuring the voltage drop with a precision digital multimeter and employing Ohm's law to determine the resistance values. Once a set of resistance values is measured, temperatures on the ITS-90 temperature scale are expressed in terms of the ratio $W(T_{90}) = R(T_{90})/R(273.16\ \mathrm{K})$ where $R(T_{90})$ is the resistance at temperature according to the defining fixed points of the ITS-90 temperature scale, and $R(273.16\ \mathrm{K})$ is the resistance at the triple point of water (Mangum and Furakawa, 1990). The PRT is perhaps the most useful secondary standard because of its inherent stability and precision; it is, however, fragile and cumbersome to work with, and so for applications requiring a robust temperature indicator, thermistors are commonly employed. PRTs provide temperature reference from 13.8 K [hydrogen triple point; from 0.65 to 5.0 K, helium vapor-pressure thermometry is employed, and from 3.0 to 24.5561 K (triple point of neon), helium constant-volume gas thermometry is used] up to 1234.9 K (the silver freezing point). The working standard temperature devices commonly employed for routine calibration are the thermocouple or thermistor. These devices are much more hardy, and while sacrificing accuracy and stability, they provide sufficient reference for maintenance and process instrumentation. At high temperatures, optical methods are used to assess temperature. Above the silver freezing point, the International Temperature Scale is defined in terms of Planck's radiation law. Temperature is defined by the relation

$$\frac{L_\lambda(T_{90})}{L_\lambda(T_{90}(X))} = \frac{e^{[c_2/\lambda T_{90}(X)]} - 1}{e^{[c_2/\lambda T_{90}]} - 1}$$

where $L_\lambda(T_{90})$ and $L_\lambda(T_{90}(X))$ are the spectral concentrations of the radiance of a blackbody at wavelength λ (in vacuum) at T_{90} and at $T_{90}(X)$, respectively. $T_{90}(X)$ refers to the temperature at a fixed point such as the silver, gold, or copper freezing point, and $c_2 = 0.014388$ mK. The values of temperature are obtained by observing the ratios of spectral concentrations, or the radiance, of a blackbody. Usually, optical pyrometers or photoelectric pyrometers are used to determine the ratio of the radiance of a source of unknown temperature to that of a reference source. The reference source in this case might be a calibrated filament lamp.

Knowledge of the amount of moisture in the environment is important for many manufacturing operations. Hence, calibration of hygrometers, dew-point sensors, and other humidity-sensing instrumentation has become ubiquitous. The national standard in this discipline is the gravimetric hygrometer, which has traceability to the fundamental standard of mass. However, more commonly used measuring and test equipment relies on determination of the dew-point temperature from observations of first condensation on a mirrored surface. Other devices are in use that employ psychrometric charts and dual temperature measurement of one dry- and one wet-bulb thermometer. In any case, a calibration typically requires characterization of instrumental response to moisture-laden environments from 0 to 100% relative humidity and in some cases down to a few parts per million moisture.

3.4 Optical Radiation Measurements

Calibration and testing of optical radiation instrumentation pertains to measurement of light—more specifically the measurement of energy or power in the ultraviolet, visible, or infrared spectrum. Radiometric measurements are among the most difficult to make. The radiant power varies with position, direction, wavelength, polarization, and time, and the responsivity of most radiometers also varies with these parameters. In general, calibration and testing take place in two separate scientific disciplines: radiometry and photometry.

In radiometry, the physical quantity measured is the radiant power incident on an aperture or the portion of a surface. From this measurement and the geometry of the setup, several radiant power densities can be calculated. These include the radiant power incident per unit area (irradiance) and the radiant power incident or exiting per unit projected area and unit solid angle (radiance). In addition, two quantities typically used to characterize sources are the intensity of a source, which is the radiant power emitted by a source in a given direction per unit solid angle, and the total flux, which is the radiant power emitted by the source in all directions.

In photometry, the physical quantities of interest are similar to those in radiometry except that they are spectrally weighted by a function that approximates the spectral response of the human eye. The radiant power detected by such a "standard eye" is called the luminous flux. The standard unit of luminous flux is the lumen, and it is defined in terms of the radiant power emitted by a blackbody at the temperature of melting platinum. More specifically, a lumen is the luminous flux emitted per steradian in a perpendicular direction by 1/600 000 m^2 of a blackbody at the temperature of melting platinum. Finally, in terms of the SI base unit of luminous intensity, the candela is equal to 1 lm/sr.

Optical radiation measurements reduce to quantifying the radiant power or flux in terms of geometrical parameters such as distance from the source or emitter, and solid angle. Hence, radiometry and photometry are traced back to fundamental physics via temperature and time standards, where the temperature standards are used to measure the caloric heating of a detector due to incident flux, and the time standards are used to define geometrical quantities such as distance or angle via the definition of the meter in terms of the cesium clock.

Secondary reference standards in radiometry range from tungsten filament to arc-discharge lamps, depending on the spectral band of interest. Other working standards include gas lamps, detectors, blackbody ovens, and x-ray sources for the high-energy regime. Finally, radiometric calibration and testing are performed on window, filter, mirror, and baffle materials for transmittance and reflectance properties.

The most accurate laser power and energy calibration is fulfilled by standard calorimetry techniques. Detector calibration is also done using a beam splitter to direct equal amounts of energy from a standard laser source to a well-characterized detector and the detector under test.

3.5 Ionizing Radiation Measurements

Calibration of ionizing-radiation instrumentation falls into four general categories: electromagnetic radiation detection (x-rays, gamma rays), beta-radiation detection, alpha-radiation detection, and neutron detection (see RADIATION DETECTORS). Depending on the type of radiation and the source geometry, particular measurements most accurately describe the response of an instrument with that of a standardized source. For example, exposure rate (roentgens per hour) or flux density of electromagnetic radiation (from x-ray machine or accelerators), at a known distance from the source, should be measured. Alpha-radiation detectors are standardized in terms of activity, or activity per unit area, of the source with the reference geometry defined as either 2π or 4π. Beta-radiation detectors are characterized in terms of air or soft-tissue absorbed dose rate at the surface or at a specified distance from the source. And neutron detection instruments are standardized by the number of neutrons emitted per unit time and the average or effective energy of the neutron. In addition to the four main types of ionizing radiation, heavy particles and very high-energy particles that are generated by atomic beam collisions are also of interest. However, these sources of ionizing radiation require either indirect detection via detection of by-products (neutrons or photons that result from collisions between heavy particles in shielding or detector material nuclei) or proportional voltage or current detection via crystal scintillation or fluorescence.

Instruments are typically calibrated as transfer standards. Comparison of the instrumental response to the radiation field of an unknown source to that of a national or derived standard source should be performed to characterize the detector over the specified energy range of use. Periodic performance reassessments using check standards are then done to confirm continued satisfactory operation of the instrument.

A full range of photon-emitting radionuclides is available for electromagnetic-radiation detector calibration. Alpha, beta, and neutron radiation sources are also available as reference standard sources. It is critical that sources that provide a radiation field similar to that which is to be sampled in regular operation be used to perform the calibration. The radiation background at the calibration facility should be well characterized to assure that measurements are accurate and reliable. The environmental conditions should be noted, and instrumental response should be characterized as a function of pressure, temperature, and humidity, as many types of detectors are sensitive to changes in the environment.

3.6 Electromagnetic Measurements

The field of electromagnetic technology is extremely diverse, and so only a cursory glance at the multitude of disciplines and calibration standards will be given here. In general, exploitation of electromagnetic waves, either propagating through transmission lines or through free space, is the basis for one of the two subgroups of electromagnetic measurements. This subcategory includes measurements that focus on electromagnetic fields and antennas and the requirements as they relate to transmission and communication. The other subgroup of electromagnetic measurements deals with the more conventional quantities like resistance, voltage, current, charge, and power.

Electromagnetic field measurements occupy an extremely large spectrum of frequencies (10^6 to 10^{15} Hz). For practical purposes, two of the most important characteristics of the electromagnetic spectrum are the transparency of the atmosphere and the amount of background noise and its variation with frequency. Transmission efficiencies are high for frequencies below 1 GHz, and so most communication and microwave technologies have concentrated in this regime. At lower frequencies, below a few hundred megahertz, components are cheap, antennas are omnidirectional, and hence accurate measurements of components or characteristics of systems are rarely called for. Greater accuracy requirements exist for microwave systems. Frequency and power levels of microwave communication systems are enforced through regulatory decree, and so transmitter and receiver characterization is common. Antenna calibration is also prevalent, with particular emphasis on radiated power and polarization in the far field. With current trends moving toward digital communications and time-division multiplexing, it is increasingly important to characterize instrumental wave forms in the time domain. Hence, oscilloscope and spectrum/network analyzer calibration is an important adjunct activity to communication system calibration. Electromagnetic field-strength measurement is also of particular interest relative to electromagnetic interference (EMI). Depending on the nature of the fields present, EMI can be both hazardous to human health and compromising to proprietary information. Finally, a few other parameters of interest are impedance, attenuation, phase, noise, and voltage.

The second area of emphasis involves standards and methodology that pertain to measuring electrical quantities such as capacitance, current, voltage, resistance, charge, inductance, power, and energy in the frequency spectrum from dc to 10 MHz. These parameters are measured in terms of the so-called legal units, those embodied by the national physical standards or defined by them as a function of physical constants (see ELECTROMAGNETIC QUANTITIES, MEASUREMENT OF BASIC). These legal units are related to the SI system, as practical, via the seven base units from which all other physical parameters are derived. In practice, though, a set of four basic equations is used to define all six basic electrical quantities: $E = IR$ (Ohm's law), $Q = IT$ (Q is charge and t is time), $C = Q/E$ (C is capacitance, and $L = E(dI/dt)^{-1}$ (L is inductance). The other two quantities (energy and power) are realized by requiring the unit of electrical energy to be equal to the unit of mechanical energy and devising absolute electromechanical experiments involving force and length. A descriptive diagram of the derivation of electrical units is shown in Fig. 1.

3.7 Time and Frequency Measurements

Time is one of the base units of the SI measurement system. However, time interval and frequency, which is the dimensional reciprocal of time and is the derived unit, are the most commonly measured parameters. The time base unit is the second and it is currently defined as "the duration of 9,192,631,770 periods of the radiation corresponding to the transition between hyperfine levels of the ground state of the cesium-133 atom" (Blair and Morgan, 1972). All time and frequency measurements are traceable to this fundamental standard.

Time and frequency calibration involves a wide variety of tests and measurements with today's sophisticated instrumentation. Instruments such as the universal time-interval counter offer functionality in time measurement, pulse width, ratio of rise to fall time, frequency, period, phase, and event counts per time interval. The precision and accuracy of each of these functions, however, rely almost entirely on the quality of the time base to which each is referenced. Since cesium atomic-beam standards are expensive, many standards laboratories depend on long-range, low-frequency air-wave transmissions of a reference frequency to standardize their time base measurements. The National Institute of Standards and Technology provides radio linkage to the national reference frequency through the WWV and WWVH programs. The LORAN-C navigation system or Global Positioning System (GPS) can realize time-base accuracy with a simple slave oscillator and a trigger generator. The LORAN and GPS system's local trigger generator bears a relationship to a master cesium oscillator, which provides time base accuracies of 1 μs.

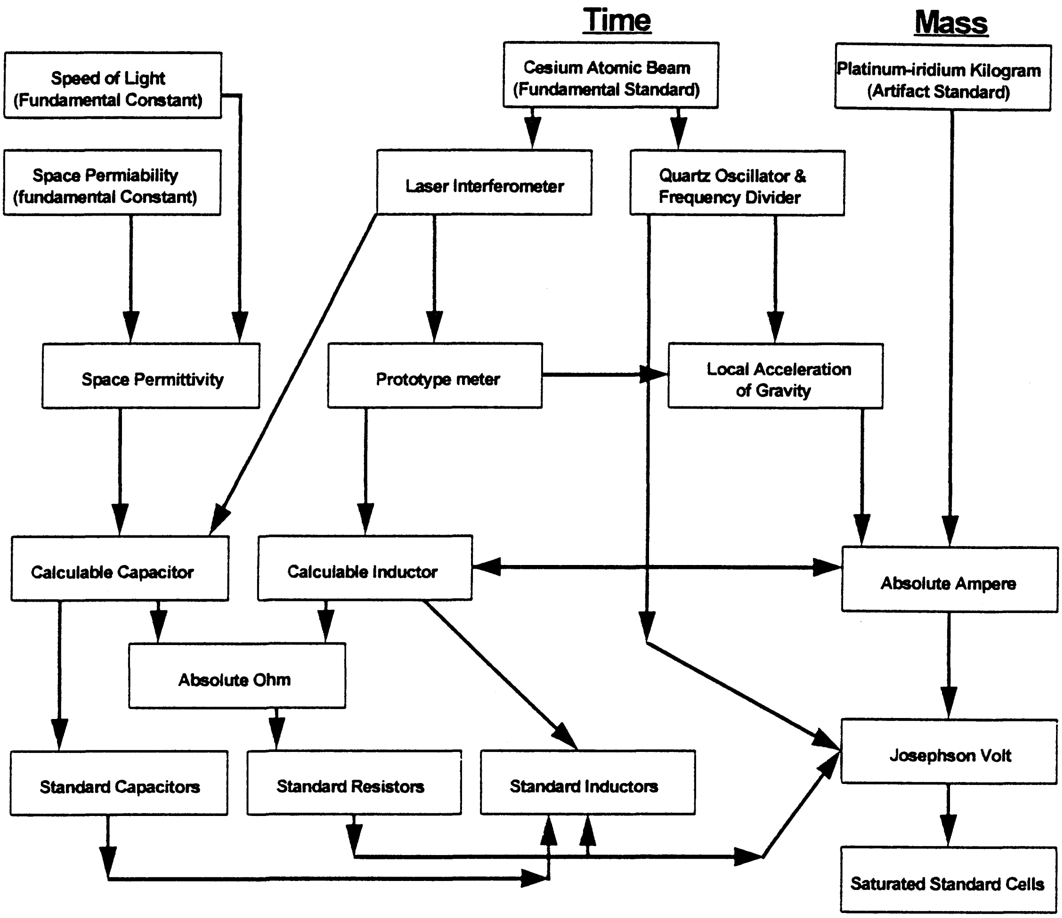
Time
Mass
Speed of Light (Fundamental Constant)
Cesium Atomic Beam (Fundamental Standard)
Platinum-iridium Kilogram (Artifact Standard)
Space Permiability (fundamental Constant)
Laser Interferometer
Quartz Oscillator & Frequency Divider
Space Permittivity
Prototype meter
Local Acceleration of Gravity
Calculable Capacitor
Calculable Inductor
Absolute Ampere
Absolute Ohm
Josephson Volt
Standard Capacitors
Standard Resistors
Standard Inductors
Saturated Standard Cells

FIG. 1. The standards hierarchy used to derive the basic electrical units.

With an accurate time base, merely a simple electronic comparator/linear amplifier can provide frequency, phase, and period measurements with part per billion resolution. Hence, most calibration facilities rely on remote links to a cesium reference and thus use that link as an external reference input to bench-top counters, synthesizers, or analyzers.

4. PERFORMING CALIBRATIONS

In any measurement process, the ultimate objective is predictability: being able to make probability statements about past and future output from the process. If this predictability is not present, then the process will yield conclusions that are only temporal and local in nature, and that will lack the generality typical of scientific experimentation. To achieve such predictability, a measurement process must be "in control." This will be the case when the output from the process behaves like random sampling from some fixed distribution with fixed location and fixed variation.

4.1 Calibration Procedures

Calibrations are meant to be an accurate assessment of the relationship between a measuring device and a measurement standard. Hence, it is necessary to perform these measurements according to a set procedure so that measurements may be repeated at some later time. Calibration procedures should clearly define the objectives of the measurements to be made and how to realize those objectives. For example, "Certify the accuracy of the digital multimeter on the voltage function, over a measurement range of 0 to 1000 V, in the direct-current mode" might be a measurement objective. Once the measurement objectives are clearly defined, the procedure should identify the proper standards necessary to fulfill these objectives, how the standards/device under test should be configured (block diagram), what measurement techniques are applicable, how the data are acquired and analyzed, and a standard format for data presentation.

4.2 Calibration Setup

An integral part of any effective calibration procedure is a circuit (mechanical if the measurements involve mechanical stimuli like pressure or temperature) block diagram. The block diagram is often useful to facilitate efficient setup and to begin identifying potential error sources or measurement problems. The diagram should include things like instrument serial numbers so that standards are clearly identified. Critical connections should be clearly marked so that instrument damage is avoided. Integrity of connections (electrical or mechanical), thermal perturbations induced by making connections, electromagnetic interference, and any other potentially destabilizing contributions should be assessed and minimized where practical.

4.3 Making Measurements

With the setup assembled, taking data may involve turning dials and writing down numbers from a display, or it may be as sophisticated as a fully automated, computer-controlled data acquisition system. In either case, the data retrieval aspect of the procedure must be carried out with great care to minimize impact on data quality. Potential acquisition error sources are excitation errors, signal conditioning errors, recording device errors, transducer errors, probe errors, electrical connection errors, environmental effects, and data processing errors (averaging, filtering, etc.). All of these potential error sources must be considered when analyzing the data acquired by the measurement process. Eliminating or minimizing as many as possible by clever experimental design and careful measurement execution is recommended.

4.4 Analyzing the Results

In Sec. 2.4, calibration errors were defined in terms of the bias limit and the precision index. Once all elemental error sources are identified as bias limits and precision indices, they are combined to obtain the total bias limit and precision index (ANSI/ASME, 1985). Now all the additional information required to make an estimate of the measurement uncertainty is an estimate of the number of degrees of freedom of the precision index (see Glossary for definition of degrees of freedom).

For simplicity, a single number U is often needed to express a reasonable limit of error. The single number, which is some combination of bias and precision, must have a simple interpretation and be useful without complex explanation. For example, the interval $\bar{X} \pm U$ represents the bandwidth within which the true value is expected to lie. It is, however, impossible to associate the uncertainty with a rigorous confidence level because the bias limit is often based on judgment or past experience, while precision is calculated using statistics.

Although rigorous confidence levels are not available, two uncertainty intervals are widely used with coverages analogous to 95 and 99% confidence levels. Coverage indicates how often an interval estimate is expected to contain the true value (ANSI/ASME, 1985). Ninety-five percent uncertainty intervals provide 95% coverage of the true value. Two uncertainty intervals which provide these coverages are given by

$$U_{\mathrm{ADD}} = B + tS_{\bar{X}},$$

called the additive uncertainty, which provides ≈99% coverage, and

$$U_{\mathrm{RSS}} = [B^2 + (tS_{\bar{X}})^2]^{1/2},$$

called the root sum square uncertainty, which provides ≈95% coverage. Here B is the bias limit, $S_{\bar{X}}$ is the precision index of the mean, and t is the 95th percentile point for the two-tailed Student t distribution. Monte Carlo simulation of the uncertainty intervals has shown that for various bias error distributions, bias limits, and precision indices,

1. U_{ADD} provides approximately 99% coverage, and U_{RSS} provides approximately 95%

coverage, when neither bias errors or precision errors are negligible compared to one another; and
2. if either the bias or precision error is negligible compared to the other, both U_{ADD} and U_{RSS} provide 95% coverage.

5. INSTRUMENT MAINTENANCE

Maintenance is generally divided into three main categories: corrective, preventive, and predictive. The latter two must be the focus of any maintenance effort, as otherwise the system will tend toward substantial amounts of corrective maintenance. Corrective maintenance is essentially a reactive management technique that waits for instrument failure before any maintenance action is taken. This approach is also the most expensive because the loss of availability of equipment resources results in a loss of productivity. Preventive maintenance programs are all time driven; maintenance tasks are based on elapsed time or hours of operation. The mean time to failure or "bathtub" curve (Mobley, 1990) indicates that a system has a high probability for early failure, due to installation problems, followed by a relatively low failure probability period for an extended time, and that finally the failure probability increases sharply with elapsed time. Hence, scheduled maintenance is based on a mean-time-to-failure statistic. Of course, for calibration equipment the mean time to failure is generally when the instrument reaches its specification limit. Therefore, calibration intervals are adjusted such that the device under test or the calibration standard is still within specification 95 or 99% of the time at recalibration. Predictive maintenance refers to a management philosophy that is premised on monitoring or periodically intercomparing instruments or standards so as to be able to predict when an instrument will be out of its specification limits. This is a very useful management philosophy because it reduces corrective maintenance on lower-level standards or test equipment that results from standards being found out of tolerance.

5.1 Standards Calibration

Calibration of standards generally requires significantly greater attention to potential sources of error than does general-purpose measuring and test equipment. Calibration standards are relied upon for the most accurate of measurements and thus are very sensitive. Environmental or mechanical stress can compromise a measurement standard. For example, slight thermal or mechanical shock to a platinum resistance thermometer will shift its calibration curve. For this reason, the transfer standard was invented.

Calibration of the highest level standards is performed by national laboratories or other internationally accredited calibration organizations. These calibrations sometimes involve complex and rigorous characterization of instrumental performance (repeatability, linearity, etc.). Calibration procedures often invoke first principles of physical science, use state-of-the-art equipment and methodology, and are expensive and time consuming. In any case, environmental conditions and human interaction are strictly controlled to minimize uncertainty, and all but the intrinsic perturbations to the system are characterized to the greatest degree possible.

5.2 Calibration Intervals

Calibration intervals typically range from 6 to 60 months. Standards, however, are generally calibrated on an annual or biennial basis. The calibration interval can, in principle, be any length of time, but the idea is to optimize the usefulness of standards and measurement equipment by keeping the interval as long as possible without exceeding the target for out-of-tolerance occurrence at recalibration. The easiest way to determine optimum intervals is to let the calibration history for an instrument determine it. Choosing an annual recall interval is a safe starting place unless the instrumentation is to operate in a volatile environment. Calibration history will then tell whether it is prudent to lengthen or shorten intervals. As a rule of thumb, out-of-tolerance at recalibration twice in a row indicates that the interval is too long.

5.3 Storage and Care of Standards

Measuring and test equipment should be stored in a clean, dry place with due consideration for environmental conditions. Equipment should be protected, locked if necessary, while remaining readily available for use. During extended storage intervals, instruments should be clearly marked as to their calibration status. Instruments and equipment with multiple users should be logged and tracked to provide a history of instrument usage. Storage compatibility issues should also be addressed (e.g., corrosive chemicals should not be stored with instrumentation). The objective is to assure proper performance of precision measuring and test equipment, and so a thorough and handy usage history is necessary to minimize unwarranted recalibration.

6. FUTURE TRENDS

Enabling technologies such as modular and miniature instrumentation and graphical user interface (GUI) software are changing the face of measurement and test equipment. Historically, calibration standards were stand-alone instruments that occupied copious amounts of laboratory bench space and required highly controlled environmental conditions. Now, miniature, self-calibrating, and highly transportable/reconfigurable instrumentation is becoming reality.

6.1 Modular and Miniature Instrumentation

Through the development of the VXI (VME instrument interface) and MMS (Modular Measurement System) architectures, whole instruments can now be placed on a single circuit board. As technology progresses, these instruments will continue to shrink until digital multimeters, oscilloscopes, and other test and calibration equipment can be placed on a single chip. Micromachining and microlithography technologies are making possible the development of solid-state sensors for physical measurements such as pressure, temperature, acceleration, and flow as well. While manufacturing advancements provide the opportunity to package more and more powerful measurement systems in a smaller space, the communication and interfacing issues need to keep pace. Ultimately, what can be envisaged is the GUI depicting a virtual instrument—whatever instrument(s) is (are) needed for the particular calibration. These virtual instruments are created by using only the necessary assets from a core set of miniature source and sense instruments. Then, through multiplexing the core assets, multiple calibration setups can be realized via virtual instruments. These systems leverage size, speed, and extreme flexibility to actualize a multifunctional, working standard calibration system.

Finally, through the developments in micromachining and nanotechnology, solid-state microsensors will provide the means for implementing a philosophical change in calibration methodology. Transient calibration techniques can be developed where parameters such as pressure or temperature are sampled without long dwell times to allow the system to come into equilibrium. These techniques are, of course, only useful for certain applications, but their applicability may pervade low-level mechanical, thermodynamic, and electrical calibration of the future.

6.2 Self-Calibrating Instrumentation

The widespread availability of microprocessor technology, the advances in software, and the advances in imbedded memory have enabled instruments to become vastly more capable than those of a generation ago. The next generation of instruments will employ an even higher level of intelligence through the use of technologies such as fuzzy logic and expert systems. These technologies will facilitate instrument self-diagnosis and self-calibration through continuous monitoring of internal responses to self stimuli. Instruments will be able to respond to drifting of internal standard artifacts such as resistors or Zener diodes and either correct themselves or warn the operator that the calibration standard is approaching or out of the specified operating range.

6.3 Transport Standards

While the virtual calibration set is highly useful, it will continue to rely on reference standards to achieve traceability to international standards. An emerging technology that will facilitate accurate and efficient recertification of such a diverse array of miniature instruments is the multifunction transportable reference standard (MTRS). The MTRS will present a group of cardinal point reference artifacts (resistors, Zener diodes, dividers, solid-state strain-gauge pressure sensors, etc.) to the virtual instrument calibration set and through cross-checks and statistical techniques maintain the measurement accuracy. Spot checks can be done periodically and quickly at high accuracies since the MTRS can transfer directly from national secondary or primary standards through some sort of measurement assurance program.

GLOSSARY

Accuracy: Degree of conformity of a measure to a standard or a true value as determined by calibration.

Calibration: The process of comparing the response of some instrument to a standard instrument over some measurement range.

Coverage: The frequency with which an interval estimate of a parameter may be expected to contain the true value.

Degrees of Freedom: A sample of N values is said to have N degrees of freedom, and the first statistic calculated from those N values is also said to have N degrees of freedom. The degrees of freedom are reduced by one for each subsequent statistic calculated from the set of N values.

Error Limit: The limit of the random error based on a set of repeated measurements.

Fundamental Standard: Of or relating to a well-defined, naturally occurring process or one that is premised on basic physical principles. There are two fundamental standards: the cesium atomic beam and the triple point of water (the platinum-iridium kilogram is an artifact standard). Fundamental constants are often mistaken for fundamental standards; the speed of light, proton mass, and electron charge are all fundamental constants as defined by the SI system of units, but they are not fundamental standards because their accurate measure depends on circumstances which preclude sufficiently accurate agreement with the SI unit unless related through the fundamental standards of time and temperature or the kilogram artifact.

Laboratory Standard: An instrument that is calibrated periodically against a primary standard.

Metrology: The science that deals with methodology and systems of precision measurement.

Primary Standard: A fundamental physical standard or one maintained at the National Institute of Standards and Technology or another globally recognized standards laboratory.

Resolution: The smallest measurable input change an instrument is capable of displaying.

Student t Value: Value used to estimate the precision error limit for a given confidence level, obtained from tables with degrees of freedom for a given confidence level.

Traceability: The ability to trace the calibration of a measuring device through a chain of calibrations to a primary standard. Traceability does not guarantee accuracy; it only indicates documented calibration hierarchy.

Transfer Standard: A laboratory instrument that is used to calibrate working standards and that is periodically calibrated against the laboratory standard.

Uncertainty: The estimated error limit of a measurement or result for a given coverage.

Working Standard: An instrument that is calibrated in a laboratory against a transfer standard or laboratory standard and is used as a standard for calibrating measuring, test, or process instrumentation.

Works Cited

ANSI/ASME (1985), PTC 19.1–1985, *Instruments and Apparatus, Measurement Uncertainty Part 1*, New York: American National Standards Institute, pp. 9–12.

Blair, B. E., Morgan, A. H. (1972), *Frequency and Time*, NBS Special Publication 300, Washington, D.C.: U.S. GPO, Vol. 5.

ISO/IEC (1990), *General Requirements for the Competence of Calibration and Testing Laboratories*, ISO/IEC Guide 25: 1990 (E).

Mangum, B. W., Furakawa, G. T. (1990), *Guidelines for Realizing the International Temperature Scale of 1990 (ITS-90)*, NIST Technical Note No. 1265, Washington, D.C.: U.S. GPO.

Mobley, R. Keith (1990), *An Introduction to Predictive Maintenance*, New York: Van Nostrand Reinhold.

Further Reading

ANSI/ASME (1985), PTC 19.1–1985, *Instruments and Apparatus, Measurement Uncertainty. Part*

1, New York: American National Standards Institute.

Doebelin, Ernest O. (1983), *Measurement Systems, Applications and Design*, New York: McGraw-Hill.

Gregory, B. A. (1981), *An Introduction to Electrical Instrumentation and Measurement Systems*, 2nd ed., New York: Wiley, Halsted Press.

Helfrick, Albert D., Cooper, William D. (1990), *Modern Electronic Instrumentation and Measurement Techniques*, Englewood Cliffs, NJ: Prentice-Hall.

Johnstone, Joseph J., Ward, Kenneth G. (1981), *How to Manage Maintenance*, American Management Associations, Extension Institute.

Kamper, Robert A. (1977), *The National Electromagnetic Measurement System*, National Bureau of Standards NBSIR 75–936, Washington, D.C.: U.S. GPO.

Mangum B. W., Furukawa, G. T. (1990), *Guidelines for Realizing the International Temperature Scale of 1990 (ITS-90)*, National Institute of Standards and Technology Technical Note No. 1265, Washington, D.C.: U.S. GPO.

McCoubrey, A. O. (1985), *Precision Measurement and Calibration: Electricity*, National Bureau of Standards Special Publication No. 705, Washington, D.C.: U.S. GPO.

Mobley, R. Keith (1990), *An Introduction to Predictive Maintenance*, New York: Van Nostrand Reinhold.

Nakajima, S. (1988), *Total Productive Maintenance*, Cambridge, MA: Productivity Press.

MANUFACTURING ENGINEERING

G. CHRYSSOLOURIS AND D. HARDT, *Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*

	Introduction	279
1.	**Manufacturing Processes**	282
1.1	Primary Forming Processes	283
1.2	Deforming Processes	284
1.3	Removing Processes	285
1.4	Joining Processes	286
1.5	Material Properties Modification Processes	286
1.6	Process Attribute Comparison	287
2.	**Manufacturing Equipment**	287
2.1	Machine Tool Types	289
2.2	Machine Tool Design and Analysis	290
2.3	Machine Tool Control and Automation	290
2.4	CAD/CAM	291
2.5	Process Control	292
2.6	Sensing Techniques and Devices	292
3.	**Manufacturing Systems**	293
3.1	Design of Manufacturing Systems	294
3.2	Operation of Manufacturing Systems	296
3.3	Areas for Further Research	301
	Glossary	302
	Works Cited	303
	Further Reading	304

INTRODUCTION

Manufacturing, defined as the transformation of materials and information into goods for the satisfaction of human needs, is one of the primary wealth-generating activities for any nation (Chryssolouris, 1991). Manufacturing can be thought of as a system in which product design is the initial stage, and the delivery of finished products to the market is the final output. Since the field of manufacturing integrates many disciplines in engineering and management, it is useful to divide it in a way that facilitates the identification of issues and allows a scientific approach to the problems encountered. Manufacturing can be subdivided into the areas of manufacturing *processes*, which alter the form, shape, and/or physical properties of a given material; manufacturing *equipment*, used to perform manufacturing processes; and manufacturing *systems*, which are the combinations of manufacturing equipment and humans bound by a common material and information flow. Design and manufacturing interfaces should also be considered, since manufacturing always begins with a design such as a drawing, sketch, CAD file, or other way of communicating the features and characteristics of the product or subassembly that is to be manufactured.

Decisions regarding the design and operation of manufacturing systems require technical understanding and expertise, as well as the ability to satisfy certain business objectives. Thus, a combination of engineering and management disciplines is required to provide a decision-making framework for manufacturing. In general, there are four classes of manufacturing attributes that are considered when making manufacturing decisions: cost, time, quality, and flexibility (Fig. 1). In the context of manufacturing decisions, cost refers to the economic resources required to produce a product; time refers to the time required to produce a product; quality refers to the extent to which the product satisfies the customers' needs and expectations; and flexibility refers to the ease with which the manufacturing process can respond to changes in customer requirements, raw materials, etc. The four attributes manifest themselves, depending on the particular problem, in specific objectives, goals, and criteria.

A manufacturing decision is basically the selection of values for certain decision vari-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

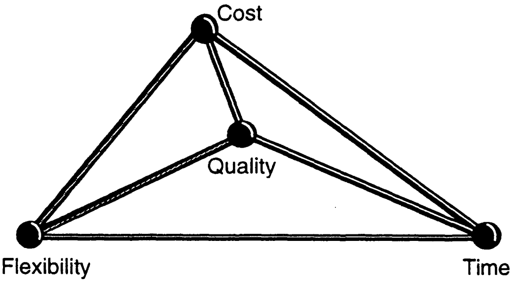

FIG. 1. The manufacturing tetrahedron.

ables related to either the design or the operation of a manufacturing process, machine, or system. Such a decision-making process is based on performance requirements, which specify the values of the relevant manufacturing attributes. In this light, the decision-making process can be regarded as a mapping from desired attribute values onto corresponding decision variable values. In order for this mapping to be scientifically founded, it must be based on *technoeconomical models*. This mapping essentially constitutes a modeling process wherein the levels of manufacturing attributes are related to the levels of different decision variables. For example, the value of an economic attribute for a turning process is determined from a technoeconomical model based on engineering analyses of the tool wear behavior (Metcut Research Associates, 1980).

The tetrahedral arrangement of manufacturing attributes emphasizes the interrelationship of these attributes. It is not possible simultaneously to optimize cost, time, flexibility, and quality. Rather, the overall outcome of a manufacturing decision is governed by trade-offs between the different manufacturing attributes. Assessment of these trade-offs requires a means of quantitatively evaluating each attribute. The more precisely defined these are, the easier it is to trade them off and make decisions. In industry to date, most quantitative measures in manufacturing refer to cost- and time-related attributes. Consequently, the trade-off between cost and production rate has been extensively treated. Thus, although cost- and time-related attributes remain extremely important for any manufacturing decision, emphasis must be placed on finding quantitative definitions for quality and flexibility, so that appropriate trade-offs among the four different classes of manufacturing attributes can be established in order to achieve a comprehensive treatment of manufacturing problems.

Costs related to manufacturing encompass a number of different factors, which can be broadly classified into several categories. Equipment and facility costs include the cost of equipment necessary for the operation of manufacturing processes, the facilities used to house the equipment, the factory infrastructure, etc. Materials costs include the raw materials for producing the product, along with tools and auxiliary materials for the system, such as coolants and lubricants. Direct labor costs are the costs needed for operating the equipment and facilities. Energy costs are required for the performance of the different processes. In some manufacturing industries, this cost may be negligible compared to other factors, while in others it contributes significantly to the financial burden of the manufacturing system. Maintenance and training costs include the labor, spare parts, etc., that are needed to maintain the equipment, facilities, and systems, as well as the training necessary to accommodate new equipment and technology. Overhead costs are the costs that are not directly attributable to the operation of the manufacturing system, but support its infrastructure. The cost of capital (or the cost to attain the required capital) may not be readily met within the manufacturing firm and therefore must be borrowed under specific terms. This classification provides a general framework of how cost issues can be addressed in the manufacturing environment by establishing a systematic way to measure the cost performance of different solutions.

In manufacturing systems, time attributes refer to both how quickly a manufacturing system can respond to changes in design, volume demand, etc. (discussed below) and how quickly a product can be produced by the system—usually expressed as the *production rate* of the system. Production rate affects, directly or indirectly, all other types of attributes. Higher production rates typically result in lower cost but possibly lower quality as well. Furthermore, to achieve high production rates it is often necessary to resort to automation, which may have an impact on the flexibility of the system. An example of such automation is a transfer line, which because of its fully automated part transfer mechanism and machining operations provides production rates on the order of one piece every 20 seconds, even for very complex parts, but is very inflexible.

For many decades, cost and production rate were the most important performance criteria in manufacturing, and manufacturers relied on dedicated mass-production systems to achieve economies of scale. However, as living standards improve it is increasingly evident that the era of mass production is being replaced by the era of market niches. The key to creating products that can meet the demands of a diversified customer base is a short development cycle yielding low-cost, high-quality goods in sufficient quantity to meet demand. This makes flexibility an increasingly important attribute of manufacturing. Flexibility, however, cannot be properly considered in the decision-making process if it is not properly defined in a quantitative fashion. The quantification of flexibility has been the focus of academic work, but industrial applications have been few. Academic research has focused on one-of-a-kind or small–lot-size production systems such as those found in the aerospace industry, but not on mass production systems. This is because the flexibility debate has concentrated on the ability of a manufacturing system to produce a spectrum of products quickly and economically. However, other aspects of flexibility (Browne *et al.*, 1984) are equally important such as *routing flexibility*, which is the ability to handle breakdowns and to continue producing a given set of part types; *volume flexibility*, which is the ability to operate profitably at different production volumes; and *operation flexibility*, which is the ability to interchange ordering of several operations for each part type.

A generic measure of flexibility that is nonetheless relatively easy to apply to realistic manufacturing situations is based on the premise that the flexibility of a manufacturing system is determined by its sensitivity to change. The lower the sensitivity, the higher the flexibility. High flexibility or low sensitivity to change brings three principal advantages to a manufacturing system. It is convenient to think of these advantages as arising from the various types of flexibility, which can be summarized in three main categories. Product flexibility enables a manufacturing system to make a variety of part types with the same equipment. Over the short term, this means that the system has the capability of economically using small lot sizes to adapt to changing demands for various products. Over the long term, this means that the system's equipment can be used across multiple product life cycles, increasing investment efficiency. Operation flexibility refers to the ability to produce a set of products using different machines, materials, operations, and sequences of operations. It results from the flexibility of individual processes and machines, the flexibility of product designs, and the flexibility of the structure of the manufacturing system itself. It provides breakdown tolerance, which is the ability to maintain a sufficient production level even if machines break down or humans are absent. This affects mass production in particular, where a large number of identical or very similar parts must be produced, and production quantity is often the most important indicator of manufacturing success. Capacity flexibility allows a manufacturing system to vary the production volumes of different products to accommodate changes in demand while remaining profitable. It reflects the ability of the manufacturing system to contract or expand easily. It has traditionally been seen as critical for make-to-order systems but is also very important in mass production, especially of high-value products like automobiles.

The quality of a product, broadly related to customer satisfaction, is often difficult to define in quantitative terms, since customer satisfaction depends not only on the actual features of a product, but also on its feasibility, maintainability, and a host of other factors that are often subjective and thus difficult to quantify. However, customer satisfaction can be traced to two major factors at the origin of a product: its design and manufacture. In manufacturing, quality typically refers to how well the production process meets design specifications related to the different features and properties of a product. The quality of a product is an aggregate of the quality of individual features and properties, which can be broadly divided into two major groups: geometric characteristics, and the physical and/or chemical properties of the materials that constitute the product.

Quality, pertaining to manufacturing systems, has many different aspects. When identical parts or products are mass produced, the repeatability of the manufacturing process becomes important, so that every item produced conforms to predefined specifications. This aspect of quality has been the subject of extensive studies in the area of statistical quality control. However, as manufacturing systems move toward low-volume or even one-of-a-kind production, the tools based on statistical methods may have to be adjusted, since the population of items that is needed for deriving statistically meaningful measures for quality may not be available.

Another aspect of quality concerns the life cycle of the product. If, after its release, a product is immediately unacceptable to the market, its design and overall features may be at fault more than its manufacturing. On the other hand, manufacturing quality becomes apparent after using a product for a period of time, through the wear behavior of the product's components.

Since manufacturing quality reflects the meeting by the manufacturing process of established expectations within tolerances, it is important to note that such tolerances may be overestimated or underestimated. An overestimation of tolerances leads to unnecessary cost during product manufacture and unnecessary pressure on the manufacturing system, while an underestimation of tolerances may lead to a malfunctioning product.

Measuring quality is critical for manufacturing, since it reflects the performance of the production process as a whole and facilitates the establishment of trade-offs between quality and other manufacturing attributes. Quality is measured from the most aggregate level, in terms of acceptance or rejection of a product, down to the elementary characteristics of a component. In general, the more aggregate the quality consideration, the less quantifiable it is and the more it relies on customer reports and subjective judgment. More elementary characteristics, for example dimensions, and physical properties, such as hardness and strength, are easier to quantify and therefore provide an easier measuring task for testing and inspection, which are usually carried out during the production process or immediately thereafter.

1. MANUFACTURING PROCESSES

A manufacturing process is defined as the use of one or more physical mechanisms to transform a material's shape or form. Manufacturing processes can be further divided into *discrete parts* processes and *continuous* processes. The metal-working industry, where many single items are produced, uses discrete-parts manufacturing. Chemical processing, used in, for example, the film- or fiber-making industries, often uses continuous processing.

Manufacturing processes can be classified into five broad categories. Forming or primary forming processes are processes in which an original shape is created from a molten or gaseous state or from solid particles of an undefined shape. During primary forming processes, cohesion is normally created among particles. Deforming processes are processes that convert the original shape of a solid to another shape without changing its mass or material composition. During deforming processes, cohesion is maintained among particles. Removing processes are processes in which material removal occurs during the process itself; cohesion among particles is destroyed. Joining processes are processes which unite individual workpieces to make subassemblies or final products. These include additive processes, such as filling and impregnating of workpieces; cohesion among particles is increased. Material properties modification processes are processes which purposely change the material properties of a workpiece in order to achieve desirable characteristics without changing its shape.

The selection of which process to apply to a particular material is influenced by a number of factors which affect cost, production rate, flexibility, and part quality. Probably the two most important factors are the lot size of the parts to be made and the physical properties of the material (e.g., the melting point). Small lot sizes require flexible processes, such as material removal processes, which are able to accommodate different geometric features, etc. Large lot sizes allow the use of primary forming or deforming processes; since tooling cost for such processes is high, a large lot size is required to amortize the cost. Metals have relatively high melting temperatures and therefore are usually processed in solid form using material removal and deforming processes. Polymers and composites have much lower melting points, allowing the use of primary forming processes, where the material is often in liquid state; however, secondary operations, usually grinding, are often needed to achieve the required dimensional accuracy and surface quality. Because these materials often contain abrasive fibers or fillers, they often exhibit abrasive behavior when conventionally machined. Ceramics are usually brittle, making them difficult to process in solid form with conventional machining techniques. With ceramics, primary forming processes are used to create the basic shape of the workpiece, and secondary operations (usually machining) are used to create the final shape and surface quality.

1.1 Primary Forming Processes

Primary forming processes produce a net shape either through solidification of a liquid or a vapor or through cohesion of solid particles. Forming processes can be used effectively on all major classes of engineering materials. Among the primary forming processes, casting has the most applications in manufacturing. Casting describes a family of processes in which molten material is introduced into a mold and then solidified. Metal casting processes are used when large parts, possibly having intricate shapes and internal cavities, are produced with mechanical properties such as high strength, brittleness, or hardness, which make these parts difficult to produce through other methods. Casting processes are limited in terms of surface quality, porosity, and, consequently, the strength of the parts produced. Generally, casting processes are used in production with relatively large lot sizes so that the high capital cost required to establish a foundry can be justified. It should be noted, however, that some casting processes are not capital- and investment-intensive.

Primary forming processes have found wide use in polymers and composite materials. Almost all processing techniques for these materials belong to the primary forming category and have a counterpart in processing metals and ceramics. In general, plastics can be molded, cast, formed, machined, and joined; and because they can be processed at much lower temperatures than metals, primary forming processes are very attractive for making components of plastics and composites.

The manufacturing process for plastics and composites often begins with raw materials in the form of pellets or powders, which are usually melted before the actual forming process begins. Plastic raw materials are also available as sheets, plates, rods, and tubing. Often, plastics used in producing composite materials are initially in liquid form. In addition, thermosetting plastics can be formed into their final shapes through chemical reactions as in reaction injection molding, thermoset molding, and thermosetting composites.

The applicability of forming processes is often constrained by the high cost and relative inflexibility of tooling. Forming dies and molds cost on the order of $10 000 to $100 000 and often require weeks to months of lead time to design and produce. However, the cost of the die can be distributed over the cost of producing a batch of parts; therefore, primary forming processes are economical when the lot size is large (on the order of 1000 parts and up). Also, forming processes often do not require highly skilled workers to operate since the tool and workpiece motions do not need to be programed, and so labor costs can be reduced relative to other processes.

Forming processes can achieve high production rates, with cycle times from seconds to minutes depending on part thickness. In general, casting and molding processes that use permanent dies machined from metals have lower cycle times than processes that use expendable foam, ceramic, or wax molds. However, expendable molds can produce intricate shapes and internal part features not available with permanent molds. Setup times for forming are on the order of eight hours, which makes casting and forming prohibitive for low-volume production.

Parts produced through permanent mold or die casting can achieve surface finishes of 1 to 2 μm R_a (roughness average), although typical cast parts require secondary removal processes such as grinding or EDM (electric discharge machining) to improve surface finish. Sand-casting processes can produce parts with up to 25 μm R_a surface roughness. Besides surface roughness, porosity in the cast or molded part due to entrainment of gas bubbles in the flow of molten material, and dimensional accuracy, are problems associated with parts produced by primary forming. The part quality can be influenced by process parameters such as the die temperature, cooling time, and cooling rate, as well as design of die or mold features such as sprues and conduits for molten material flow and mold cooling lines.

The flexibility of forming processes is limited by the kinematics of forming processes, which are constrained such that only one part geometry can be produced for a die geometry and the part geometry cannot be changed through workpiece-tool motions. However, forming processes have the potential to produce parts with very intricate geometric features, especially internal features, and workpiece thicknesses that can vary from under 1 mm to over 100 mm.

1.2 Deforming Processes

Deforming processes have a variety of applications in the industrial world. Components for automobiles, industrial equipment, and machine tools as well as hand tools such as hammers, pliers, and screwdrivers are some industrial applications of deforming processes. Containers such as metal boxes and cans, and fasteners such as screws, nuts, bolts, and rivets are other components created using deforming processes. The most commonly used materials in deforming processes are steel, carbon alloyed steel, stainless steel, and heat-resistant steel; nonferrous heavy and light alloys, such as aluminum, zinc, and copper; and titanium and heat-resistant nickel alloys. The significance of deforming processes can be best understood by considering that a medium-sized automobile has at least 150 pounds of steel in cold-formed parts.

Generally speaking, deforming processes used in the metal-working industry can be divided into *bulk* or *massive deforming* methods (Lange, 1985), such as extrusion, upsetting, or forging, and *sheet metal forming* techniques, such as deep drawing and bending. Irrespective of the type of deforming method, the stresses required for deformation are quite high, ranging between 50 and 2500 N/mm^2 [7–360 ksi (kilopounds per square inch)] depending on the method and the material. Because the entire workpiece, or a major part thereof, is deformed, the loads which result from a deforming process also are very high; for example, in a forging press loads may be up to 750 MN, or 85 kts. Comparing this with the heavier equipment for material removal processes, such as large planing machines, the load is a few orders of magnitude lower (20 kN or 2.3 tons) than the load which occurs in deforming processes.

Industrially applied deforming processes can be classified in a number of ways. One way to classify these processes is to examine the temperature present while the forming occurs. Following this criterion, one can distinguish between hot forming processes, which occur after the workpiece material has been heated, and cold forming processes, which occur without additional heating. Stress is another classification criterion referring to the amount of loading on the material when the deforming process occurs. The purpose of these classifications is to facilitate the training of people for the different processes, to provide uniform, reliable, speedy communication regarding issues of deforming processes, and also to provide some precise definitions of the different processes for use in process planning.

Deforming processes generally have relatively high tooling costs, similar to forming processes, because of the complicated die geometries required. Deforming methods are directly related to machining and metal-removing methods because the tools used for metal deformation are usually produced by machining processes, as a result of the small lot size (one-of-a-kind in most cases) and high precision required. Because the costs of tooling (up to $250 000 per die) and machinery (up to $200 000 per machine) are high, deforming processes are usually applied in lot sizes large enough to justify economically the high cost of the tools and machinery required. Since a deforming tool usually has a life of several million parts, the high tooling cost can be distributed over a large number of parts, so that the total manufacturing cost per part will be low for large batch sizes. Finally, highly skilled workers are usually not needed to operate deforming processes, and so labor costs are also relatively low compared with other manufacturing processes.

For large lot sizes, deforming processes have production rates up to 5000 parts per hour. Similar to primary forming processes, deforming dies require setup times up to several hours, making deforming processes impractical for small–lot-size production.

The range of surface finishes achievable through deforming processes is similar to that of primary forming processes, with surface roughness down to 0.8 μm R_a for extrusion and cold rolling. However, hot forming processes such as hot rolling, where the workpiece is preheated to an annealing temperature, generally yield much higher surface roughness (up to 50 μm R_a) and will require secondary finishing operations. One benefit of parts produced through deformation is that the deformation process produces work hardening in the workpiece, which increases the mechanical strength of the part. However, excessive material deformation may lead to crack and overlap formation in the workpiece.

Deforming processes have a relatively low degree of flexibility compared with other manufacturing processes, since the kinematics of deforming machines are constrained by either motion, force, or energy. The geometry of the part is governed solely by the tool geometry, as in the case of primary forming. However, since deforming dies must move relative to the workpiece, the geometric features that are producible are limited.

1.3 Removing Processes

Mechanical material-removal processes are the backbone of industrial manufacturing practice. These processes provide a great deal of flexibility, since the geometry of the part is defined by the shape of the tool and the kinematics of the tool and workpiece. The material-removal mechanism is a very important aspect of removing processes. There are four types of material-removal mechanisms: mechanical, in which mechanical stresses are induced by a tool to overcome the strength of the material; thermal, in which the thermal energy provided by a heat source melts or vaporizes the volume of material to be removed; electrochemical, in which the electrochemical reactions induced by an electrical field destroy the atomic bonds of the material to be removed; and chemical, in which chemical reactions are employed to destroy the atomic bonds of the material to be removed.

Removing processes typically have greater flexibility in terms of the lot size, workpiece geometry and shape, and part quality achievable compared with other manufacturing processes. The machine and tooling costs associated with material-removal processes are low compared with primary forming and deforming processes. Machines typically cost from \$10 000 for manual equipment to \$80 000 for CNC (computer numerically controlled) equipment, while tooling cost ranges from \$1 to \$100 per tool. However, the skill level involved for programing or manually setting the tool and workpiece kinematics is relatively high, and so labor costs for operating material-removal processes are correspondingly high. Therefore, machining processes are better suited for low- to medium-volume production, with lot sizes ranging from one to several hundred parts.

The production rates for machined parts are much lower than for casting or deforming processes, since the tool is required to make multiple passes over the workpiece surface to produce the final shape. The material-removal rate is dependent on the surface quality desired, the workpiece material, the cutting-tool material, and the cutting fluid used. For steels, the material-removal rates can range from less than 1 mm^3/min for a grinding process up to the order of 10 cm^3/min for a milling process. However, time required for workpiece setup and fixturing (on the order of several minutes to one hour) is generally much less than that for primary forming and deforming processes.

Surface quality and surface technology are clearly very important aspects of material-removal processes. Surface *texture* can be described in terms of surface roughness, macro effects such as imperfections, and geometric considerations such as tolerances. On the other hand, surface *integrity* refers more to microstructured effects such as microcracks and residual stresses. Surface effects are caused both by the process itself and by the workpiece material properties. These effects have a direct influence on the mechanical characteristics of the workpiece and eventually on the reliability of the component.

Material-removal processes are among the most flexible of manufacturing processes. Since the geometry of the finished part is defined by the geometry and the kinematics of the tool and workpiece, material-removal processes can produce parts with a wide range of sizes, shapes, and surface quality. Milling, turning, and grinding span a large segment of achievable surface quality, exhibiting their flexibility and popularity in industrial practice. The surface generation mechanism in milling and turning explains the great flexibility of these processes in terms of creating a variety of surface qualities through the selection of process parameters. Another reason for the flexibility of material-removal processes is the number of degrees of freedom incorporated into material-removal machines. Current numerically controlled (NC) machining centers have up to nine translational and rotational degrees of freedom for a given tool-workpiece combination. Even in drilling processes, which are limited to one degree of freedom, a range of hole depths can be produced without changing tools.

1.4 Joining Processes

Joining includes a number of processes such as welding, brazing, soldering, adhesive bonding, and mechanical joining (Kalpakjian, 1989; Shey, 1987; Schwartz, 1979). Each of these is an integral part of modern manufacturing operations, as most products are impossible to make as a single piece. Indeed, joining processes offer the ability to manufacture a product as individual components and then combine them into a single product, which may be easier and less expensive than manufacturing the whole product at once. Many products must be taken apart for maintenance, repair, or service, so it is beneficial to join the components so they can be easily dismantled and reassembled. Joining processes also allow the inclusion of features and properties in a particular product which may differ from the majority of the components used in the product.

The capital and tooling costs associated with most joining operations are relatively low compared with other processes, since most joining equipment is inexpensive. A typical welding machine ranges in price from \$5000 to \$50 000. However, the skill level of operators is high, especially in operations where parts with complex geometries are joined; therefore the labor costs associated with joining processes can be high. Other costs such as tooling (welding nozzles) and material (adhesive, filler, and operating gas) are relatively low.

Joining processes are very labor-intensive, especially for adhesive bonding or joining parts with complex geometries. Most joining processes require preprocessing of the joining surfaces to minimize surface roughness, and a period of time after joining for curing of the bond or cooling of the weld. These factors result in low production rates for joining processes compared with forming, deforming, or removing processes.

The quality of joints is limited by the lack of effective methods for nonintrusive joint evaluation. In welding, the weld pool formation is accompanied by the propagation of a heat-affected zone, in which the microstructure of the material is changed by heating. Other types of defects common in joints include porosity, entrapment of contaminants in the joint, incomplete fusion or penetration, crack formation, surface damage, and residual stresses. However, effective use of joining techniques can produce joints with mechanical strengths that exceed those of their joining members.

Joining processes also have a high degree of flexibility in part geometry and lot size. Since most welding tools have several degrees of freedom, a wide range of sizes and shapes of parts that may be joined can be accommodated.

1.5 Material Properties Modification Processes

The manufacturing processes introduced above are primarily concerned with changing the shape of a workpiece. However, since all manufacturing processes involve interactions between many phenomena, including mechanical, thermal, acoustic, and electromagnetic, there are invariably secondary effects of the processes, including alteration of material structure and composition, which are unintentional results. In this section, the processes introduced have the intended purpose of changing the physical properties of the material while maintaining the basic geometry of the part.

The practicality of many processes often depends on the physical properties of the workpiece materials. Heat and surface treatment methods can substantially influence the properties of these materials. These treatments are particularly useful with metals. Heat treatment permits the modification of the mechanical properties of metals, within certain limits, and is usually used to create suitable conditions for subsequent manufacturing operations and for the final use of the component.

The capital and operating costs of heat and surface treatment equipment can be high, with modern furnaces costing on the order of several hundred thousand dollars. Computer control on modern furnaces reduces the labor cost by minimizing the need for skilled human control of process parameters. The fuel or electrical consumption on furnaces is relatively high compared with other types of machine tools.

Since heat treatment processes require elevation of workpiece temperature, a significant amount of processing time may be required to heat the workpiece, maintain an elevated temperature, and cool the workpiece to room temperature. Also, the startup time for conventional furnaces may take several hours. As a result, heat treatment processes can constitute a significant portion of the total production time of a part. However, the improvement in mechanical properties of heat-treated parts may eliminate the need for using a more expensive material.

Surface and heat treatment processes are typically used to improve the quality of the part through material hardening (up to 65 Rockwell hardness C), relief of residual stresses, and coating of the workpiece with a material which has favorable properties. The treated region usually extends to 0.05–1 mm below the workpiece surface. Incorrect heat treatment of any metal alloy can lead to defects such as the formation of a coarse grain, surface burning, or surface decarborization. Coarse grains may be caused by overheating or by being held too long at the annealing temperature. Surface or skin decarborization occurs in steels as a result of the diffusion of carbon into the atmosphere because of excessively long annealing times. Often the burn structure is a defect which cannot be overcome. Heat treatment processes carried out at too high a temperature may lead to irreversible oxidation of the grain boundaries, which weakens their material cohesion, decreasing strength and rendering the material unusable.

Surface and heat treatment processes can accept a wide range of part geometries and sizes for processing, since there is no contact between a tool and the workpiece. However, in most conventional processes, the size of the part acceptable for processing is limited to the size of the oven or furnace used for treatment. Some state-of-the-art processes such as laser surface treatment eliminate this restriction by directing a heat source such as a laser beam at the workpiece and scanning the heat source along the workpiece surface.

1.6 Process Attribute Comparison

The applicability of the criteria that form the manufacturing tetrahedron discussed earlier to each class of manufacturing processes is shown in Table 1. The comparison illustrates the nature of the tradeoff between the manufacturing attributes for any manufacturing process. Selection of an appropriate process often depends on a balance of all four criteria based on the requirements for making a desirable product in an efficient and profitable manner. For example, if the primary concern is production rate, then primary forming and deforming processes should be considered. However, if many different part types with small batch sizes are to be produced in the same facility, then removing and joining processes are more appropriate. Finally, if the quality of the part is of prime concern, material removal and material property modification processes should be incorporated as secondary operations.

2. MANUFACTURING EQUIPMENT

Machines and manufacturing equipment are basically the "embodiment" of processes, while machine tools are the "building blocks" of manufacturing systems. The term "manufacturing equipment" covers a large variety of devices and machines, ranging from hand tools to complex automated machining centers, encountered in modern production facilities. This section emphasizes machine tools because they may be the most important elements of manufacturing systems, particularly in discrete parts manufacturing. Manufacturing equipment, and particularly machine tools, have been used in industrial production for over 100 years; the evolution during this period has been quite remarkable. Indeed, considering the example of the lathe (Fig. 2), one can notice that in the first decade of this century, a typical lathe was powered by a motor positioned outside of the machine itself. This configuration forced the machine to remain fixed at the site of installation and prevented easy relocation. In the 1920s, the motor became an integral part

Table 1. Characteristics of the classes of manufacturing processes.

	Forming	Deforming	Removing	Joining	Modifying
Cost	High tooling/ low labor cost	High tooling/ low labor cost	Medium tooling/ high labor cost	Low capital/ high labor cost	Medium to high capital/low labor cost
Production rate	High	High	Medium (milling) to low (grinding)	Medium (welding) to low (adhesives)	Low
Quality	Medium to low	Medium to low	Medium to high	Medium to low	High
Flexibility	Low	Low	High	High	Medium to high

of the machine, which allowed an easy relocation of the machine in case of a factory reorganization and also provided a stiffer machine that could achieve a higher production rate. In the 1960s, machine tools were produced with completely enclosed frame to obtain greater stiffness, and often included template and stylus for the purpose of copying a complex shape onto the workpiece. The lathes of the 1970s and 1980s became numerically controlled, with the ability to be programmed with the help of microcomputers (Figs. 3 and 4). There is a synergism between manufacturing processes and the equipment that performs them. Developments in manufacturing processes and materials have substantially influenced the evolution of machines. The introduction, for example, of new cutting-tool materials continuously increased the cutting speeds attainable in machining operations. Thus, machines had to be properly designed in order to accommodate the need for higher cutting speeds and production rates.

Manufacturing equipment, particularly machine tools, is the critical link between intermediate products formed from raw materials and finished discrete parts and components. These components are then assembled either into end products or into yet

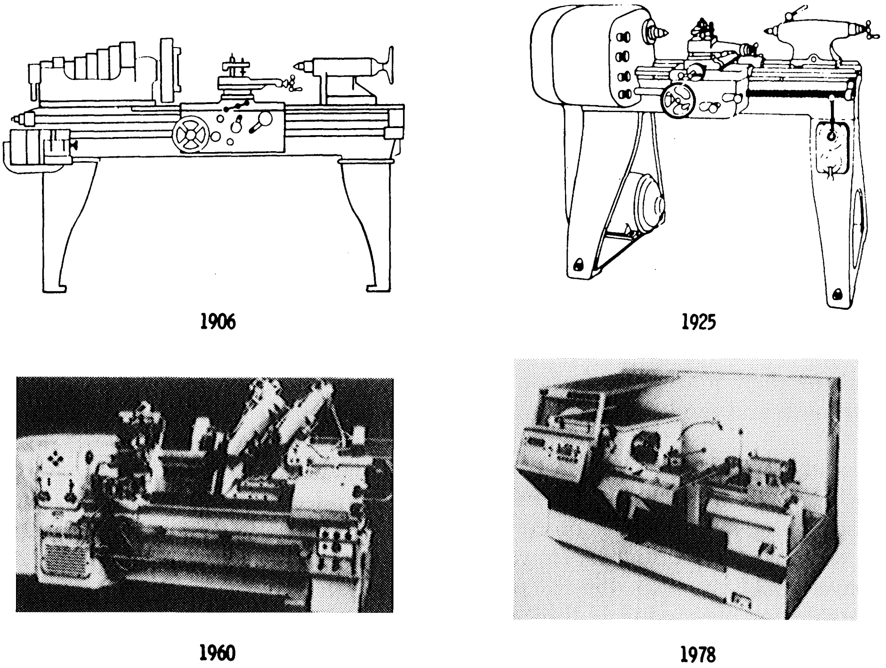

1906

1925

1960

1978

FIG. 2. The evolution of the lathe (from Weck, 1984).

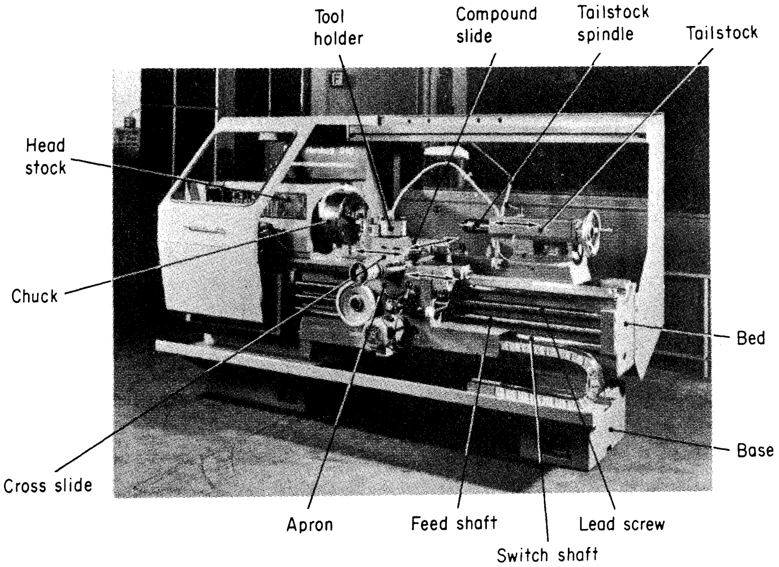

FIG. 3. Sample lathe (from Weck, 1984).

other machines that will in turn make end products. Thus, virtually every manufactured product is built either directly or indirectly by machine tools (M.I.T., 1989). It is this position at the heart of the manufacturing infrastructure that makes manufacturing equipment and the machine tool industry so critical to a nation's manufacturing competitiveness and productivity. Furthermore, exporting has made the machine tool industry a global one; while less than 30% of world production of machine tools was exported 20 years ago, by 1976 exports were nearly 43% of production, and now they are close to half.

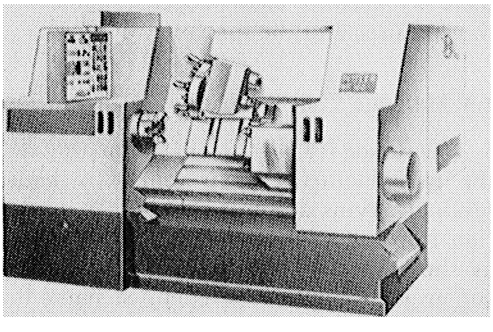

FIG. 4. Numerically controlled automatic lathe with drum turret (from Weck, 1984).

2.1 Machine Tool Types

In general, a machine tool can be defined as a nonportable machine, with an integral power source, that causes the relative motion of a tool and a workpiece in order to produce a predetermined geometric form or shape. This "narrow" definition primarily reflects the two major classes of machines: deforming (or forming) and removing (or cutting) machines. However, machines used for primary forming processes such as casting, and joining processes such as welding or assembly, can also be regarded as machine tools.

In order for machine tools to be successfully integrated into the production process, a number of criteria must be met. In modern manufacturing systems, quality is a major concern and may be the primary driving force in machine tool design, development, acquisition, and operation. Quality in machine tools falls into two categories: accuracy in geometric and kinematic terms, under static, dynamic, or thermal loading; and reliability, or the behavior of the machine over an extended period of time. Related to these characteristics are maintainability, work-space accessibility, and safety. Modern machine tools are expected to operate with accuracy on the order of microns in a wide range of loading conditions, some of which may exceed several tons. This accuracy, combined with the high velocity with which the machine tool elements must move in order to achieve required production rates, presents a formidable design and manufacturing challenge.

Cost is also an important factor, since machines typically represent a high capital investment which can reach into the order of billions of dollars for mass production systems typically in the automotive industry. Often, both cost and quality of the machine are determined by its degree of automation. Over-automated machines may be unreliable, while machines that are under-automated may have low production rates. Furthermore, the degree of automation often determines the degree of flexibility. Highly automated machines or systems such as transfer lines provide a very high production rate but limit product and operational flexibility. Thus, automation of machine tools should strike a balance between cost, quality (particularly reliability), and flexibility. In general, product flexibility is high among metal removing machines and relates to the "envelope" of the machine, namely the maximum and minimum dimensions of the parts that the machine can handle, and the accuracy range of the different motion axes of the machine.

Machines usually need to satisfy safety and environmental requirements. For example, machine tools that operate at high noise levels may be required to be encapsulated as a safety and environmental measure. Safety-related issues have the highest priority and can influence the design of a machine tool considerably. However, safety considerations often have a detrimental effect on the efficiency or productivity of a machine tool.

2.2 Machine Tool Design and Analysis

Machine tool design, in terms of overall structure and kinematics, is dictated primarily by the process to be performed by the machine. The design is also substantially influenced by the range of workpiece sizes, the required accuracy, operation mode (manual, semimanual, fully automated, etc.), maintenance procedures, cost, and a host of other factors related to the requirements and attitudes of the machine end user. Since machine tools are complex electromechanical structures which include many interconnected devices, their design involves a number of disciplines ranging from applied mechanics to ergonomics. This unique level of complexity, coupled with strict requirements in terms of kinematic accuracy, static and dynamic behavior, etc., causes the machine tool design process to rely heavily on empirical knowledge and expertise. Engineering analysis is typically confined to the identification of trends and the creation of a framework for machine design, instead of delivering exact solutions for the structure of the machine, the size of its components, etc. In general, stand-alone machines are designed to accommodate a variety of conditions in terms of workpiece size, etc., while dedicated machines such as transfer lines are designed for reliability rather than flexibility.

Examining the historical development of machine tool designs and their slow evolution, one must consider two factors: first, machines require a large initial capital investment, and second, the production and ultimately the revenues of the end user depend on the machine's proper functionality. An experimental, unproven machine design usually involves a high risk of not achieving the required production quotas, which may explain the conservative attitude of many machine tool builders throughout the world in terms of introducing radically new ideas in their machine tool designs. On the other hand, the lack of an innovative approach to machine tool design may lead to a loss of competitive advantage. The correct attitude in machine tool building is to strike a balance between new ideas for the future and well-proven designs of the past. Furthermore, since machine tool design substantially influences the production capabilities of the end user, there is a need for close cooperation between the end user and the machine tool builder.

2.3 Machine Tool Control and Automation

One reason for introducing automation into the manufacturing environment is higher productivity as a result of a higher production rate. In mass production systems such as those used in the automotive and consumer-goods industries, complex parts that would take a few hours to manufacture in manually or semi-automated systems can be made in seconds by automated systems such as transfer lines. Such "hard-wired" automation is designed to reduce material handling time, or the time the part spends between operations. In the metalworking industry, parts often spend only 5–15% of their time in added-value operations, while the rest is spent in handling and waiting. Automation may considerably increase the production rate of the system, but it also increases the required capital investment and substantially reduces the product, operation, and volume flexibility of the system.

Another motive for introducing automation is more consistent quality. This objective involves eliminating human errors and inconsistencies, which result in quality deviations. Manually performed manufacturing processes are influenced by the skills and abilities of the human operator, and so the quality of the finished parts also depends on the skill and ability level of the operator. Implementing automation for the purpose of consistent quality may reduce the flexibility of the system, however, since humans are one of the most flexible factors in the production process. Finally, better work content motivates the proliferation of automation. Repetitive or hazardous tasks can be automated to relieve humans. This "humanistic" automation also streamlines the production process, making it more efficient.

Since the introduction of automation provides higher productivity while, in general, reducing flexibility, considerable research and development have been devoted to the area of "flexible automation," which provides high production rates and consistent quality, while being flexible enough to accommodate a variety of tasks. Two results of these efforts are the establishment of Flexible Manufacturing Systems (FMS) and the introduction of robots that can be programed to perform different tasks.

2.4 CAD/CAM

The basis of any manufacturing effort is the design drawing. Painstakingly prepared drafts and blueprints are used in every step of a product's inception, design, and construction. But like the old-fashioned office, drafting is paperbound and is very labor intensive. Unlike the electronic data in a computer, paper drawings cannot be manipulated, processed, and quickly analyzed.

CAD can be most simply described as "using a computer in the design process." In the design process, a computer can be used in both the representation and the analysis steps. The application of CAD for representation is not limited to drafting. Three-dimensional wire-frame modeling, boundary representation (B-rep) modeling, and solid modeling are presentation methods available to CAD users. To aid in engineering analysis, there are packages that perform kinematic simulation and finite-element modeling for different engineering analyses. CAD systems frequently consist of a collection of many application modules under a common database and graphics editor.

A CAD system is a tool used to automate design work. Instead of using a pencil and paper, the designer uses a sophisticated computer graphics display. A light pen is used to draw on the monitor screen or to extract lines, curves, and shapes from the computer's memory. Solid shapes called from a computer library of geometric primitives can be assembled to form complex shapes. A CAD system can eliminate the tasks of searching through blueprint libraries, transcribing what has been drawn before, or redrawing existing figures or part assemblies. Using simple commands, a designer can display a cutaway of a part, a closeup view, or an exploded view of all the parts in an assembly. Using a CAD system, the complete geometry of a part can be visualized without the necessity of viewing different layers of a drawing. Any product or part that was drafted on paper can be drafted using a CAD system. Some CAD systems are more complex than others, and the design capabilities are limited by the machine itself. However, even the most limited CAD system typically offers significant productivity gains over traditional manual methods of drawing.

A CAD/CAM system is the application of a computer to the manufacturing process, in addition to its use in the design process. Similarly to CAD, CAM is a tool used to automate manufacturing. The manufacturing of a workpiece can be aided with computers at any stage of manufacturing, from the process planning to the system planning. Ideally, a CAM system can work in conjunction with a CAD system to allow a direct link (via computers) from the desk of a designer to the workbench of a shop floor. Current problems

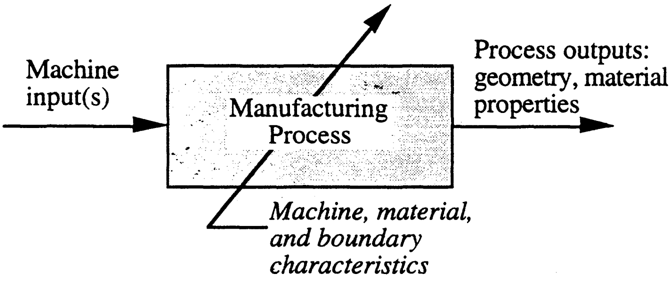
Machine input(s)
Manufacturing Process
Process outputs: geometry, material properties
Machine, material, and boundary characteristics

FIG. 5. Block diagram of a manufacturing process.

with CAD/CAM systems involve the interface between the two systems. Different CAD systems have different methods of representing dimensions, and translation from CAD to CAM is difficult and often requires a human interface. This problem will diminish as improved methods of translating geometric representation for CAM systems develop and as CAD systems become more standardized.

2.5 Process Control

From a control perspective, a manufacturing process is easily modeled as an input-output (I/O) function as shown in Fig. 5. Notice here that this block includes both the actual material transformation process and the equipment supplying and modulating the transformation power. The outputs from any manufacturing process are by definition the geometry and properties of the product. The input is that set of equipment variables that have a clear and deterministic causal effect on the output, and the input-output relationship is parametrized by the characteristics of the workpiece material, the working environment or boundaries, and the basic characteristics of the equipment in use. The *raison d'être* for process control is the rejection or elimination of disturbances that typically arise from parameter variations; primary among these are the uncertainty and variation in the material properties. Uncertainties also arise from the inability to model precisely, in sufficient detail, the physics of most three-dimensional manufacturing processes.

From this process block, a generic manufacturing process control-block diagram can be drawn as in Fig. 6. The essential features of this diagram are the dominant time constant t_p of the process and the delay characteristics T_d of the feedback. The time constant represents either the process dynamics or the cycle time of the process for completion of a single product. The measurement delay is used to delineate various forms of process control. These are summarized in Table 2. While there are much more detailed differences among these various approaches (see, for example, Box and Kramer, 1990, where a detailed comparison of statistical process control and feedback control is presented), Table 2 illustrates that all processes are subject to some type of feedback control, either automatic or human operated, and even "no control" involves feedback in the form of customer complaints or warranty recalls.

Clearly, therefore, the choice of control method to be used involves factors such as the volume of parts (since low volumes cannot tolerate low control bandwidths, for example), the flexibility of the equipment, and the overall sensitivity of the process.

2.6 Sensing Techniques and Devices

The overall goal of automation for manufacturing processes necessitates the estab-

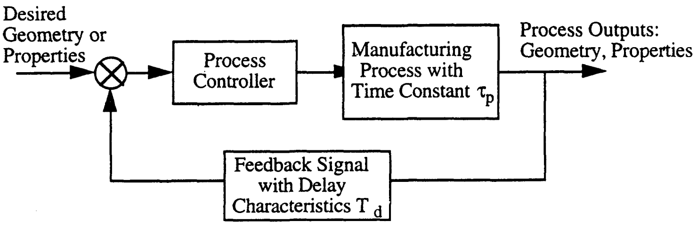
Desired Geometry or Properties
Process Controller
Manufacturing Process with Time Constant τ_p
Process Outputs: Geometry, Properties
Feedback Signal with Delay Characteristics T_d

FIG. 6. Canonical process control block diagram.

Table 2. Families of control methods based on decreasing bandwidths.

Control method bandwidth	Control method
$T_d < t_p$	Real-time or in-process control
$T_d = t_p$	Iterative control
$T_d \sim (10\text{–}100)t_p$	Sampling and statistical process control (SPC)
$T_d \sim 1000t_p$	Empirical modeling and process optimization
$T_d \sim$ months	Warranty recall

lishment of control loops by providing some feedback information about the state of the manufacturing process. This function is carried out by human supervision in conventional process control. With increasing automation, effective sensor systems are needed in order to achieve the automatic control, supervision, and error recovery tasks required to realize the full potential of intelligent manufacturing. Process sensors can be divided into three categories according to their function: tool monitoring, workpiece monitoring, and machine performance monitoring.

In the conventional case, most of the sensing functions are fulfilled by human perception and evaluation. The push toward untended machining and "lights-out" factories requires high levels of automatic machine monitoring. A sensor system must reach at least the level of effectiveness of the human sensing capabilities in order to replace a human operator. Sensors having sensing and processing capabilities equal to or exceeding those of humans can be called "intelligent" sensors.

3. MANUFACTURING SYSTEMS

The manufacture of products in the modern industrial world requires the combined and coordinated efforts of people, machinery, and equipment. Thus, a manufacturing system can be defined as a combination of humans, machinery, and equipment that are bound by a common material and information flow. The materials input to a manufacturing system are raw materials and energy. Information is also input to a manufacturing system, in the form of customer demand for the system's products. The outputs of a manufacturing system can likewise be divided into materials, such as finished goods and scrap, and information, such as measures of system performance.

In general, manufacturing system design can be conceptualized as the mapping from the performance requirements of a manufacturing system, as expressed by values of certain performance measures, onto suitable values of decision variables, which describe the physical design or the manner of operation of the manufacturing system (Fig. 7).

A performance measure is a variable whose value quantifies an aspect of the performance of a manufacturing system. Performance measures are either *benefit* measures (the higher the better) or *cost* measures (the lower the better). Like manufacturing attributes, they can be divided into the four categories time, quality, cost, and flexibility. In general, a number of performance measures will be relevant for a given manufacturing system. However, these will differ from one manufacturing system to another.

Given performance requirements, the manufacturing system designer must describe a suitable system design. This design can be captured numerically by specifying the values of an appropriate collection of *decision variables*. An example of a decision variable is the number of each type of machine in a manufacturing system.

Designing manufacturing systems (mapping performance measures onto decision variables) is a difficult task because:

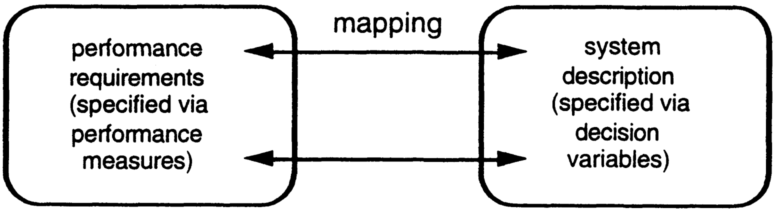
mapping
performance requirements (specified via performance measures)
system description (specified via decision variables)

FIG. 7. The manufacturing system design problem.

- Manufacturing systems are large and have many interacting components.
- Manufacturing systems are dynamic.
- Manufacturing systems are open systems, which influence and are influenced by their environment.
- The relationships between performance measures and decision variables usually cannot be expressed analytically. Well-behaved functions do not apply.
- Data may be difficult to measure in a harsh processing environment.
- There are usually multiple performance requirements for a manufacturing system, and these may conflict.

Assuming that these difficulties can be surmounted, the next task is the construction of the actual manufacturing system. As this process is very capital intensive, it places a premium on the cost performance measures of the manufacturing system. No matter how worthy its performance in other aspects, a manufacturing system will never be constructed if it is not shown to be financially viable. The potential benefits of a well-designed production system include (Choobineh, 1986) low work-in-process; inventories; increased production rate; few fixtures, jigs, and tooling; improved product quality; reduction of floor space used; and reduced labor costs.

3.1 Design of Manufacturing Systems

In industrial practice, there are five general approaches to structuring the processing area: the job shop, project shop, cellular system, flow line, and continuous system. In a job shop, machines with the same or similar material processing capabilities are grouped together. The lathes form a turning work center, the milling machines form a milling work center, and so forth. The machines are usually general-purpose machines that can accommodate a large variety of part types. In this structure, the part or lot of parts moves through the system by visiting the different work centers according to the part's process plan (Fig. 8). Material handling must be very flexible in order to accommodate many different part types, and so it is usually done by manually controlled implements such as forklifts and handcarts. Within each work center, a number of machines can be used for a particular operation. This is advantageous for a number of reasons:

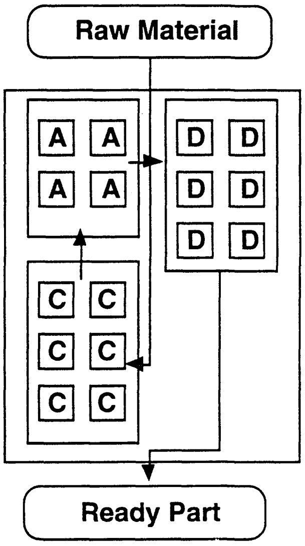

Machines/Resources are grouped according to the process they perform
Letters refer to individual machines/resources

FIG. 8. Schematic of a job shop.

1. Each operation can be assigned to a machine which yields the best quality or the best production rate;
2. machines can be evenly loaded; and
3. machine breakdowns can be accommodated easily.

These advantages are particularly evident when the job shop's work load includes a large variety of different parts with different process sequences. However, fully exploiting the flexibility of a job shop requires making and implementing complex decisions in real time. For example, when a machine becomes available after finishing an operation, which of the possibly many eligible operations should it perform next? Furthermore, lack of decision-making coordination among the work centers can impede the smooth flow of parts from work center to work center. This results in parts spending a long time in the job shop and creates high inventory levels.

In a project shop (Fig. 9), a product's position remains fixed during manufacturing because of its size or weight. Materials, people, and equipment are brought to the product as needed. Facilities organized as prod-

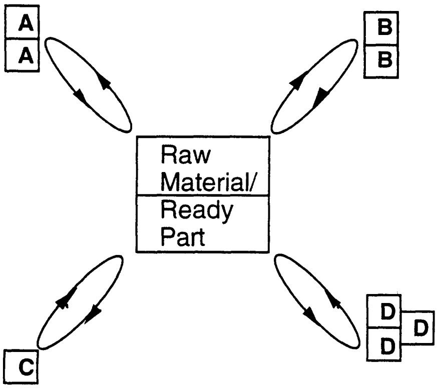

Machines/Resources are brought to and removed from stationary part as required

Letters refer to individual machines/resources

FIG. 9. Schematic of a project shop.

uct shops can be found in the aircraft and shipbuilding industries and in bridge and building construction.

In manufacturing systems organized according to the cellular plan (Fig. 10), the equipment or machinery is grouped according to the process combinations that occur in families of parts. Each cell contains machines that can produce a certain family of parts. The material flow within the cell may differ for different parts of a part family. Intracellular material flow can be performed either automatically or manually. By confining the path of each part to an individual cell within the manufacturing system (when possible), cellular systems enable each cell to be considered independently for detailed scheduling purposes, thus greatly simplifying this type of decision. Transportation times and the times spent in the system for the different parts are reduced compared to those of the job shop.

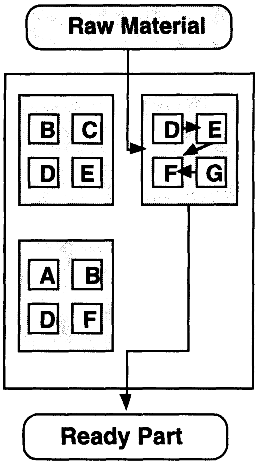

Machines/Resources are grouped according to the processes required for part families

Letters refer to individual machines/resources

FIG. 10. Schematic of a cellular manufacturing system.

The fourth way of structuring a manufacturing system is to create a flow line in which the machines and other equipment are ordered according to the process sequences of the parts to be manufactured (Fig. 11). A typical example is the transfer line, which is often used in the automotive industry. A transfer line consists of a sequence of machines each of which is dedicated to a particular operation or a particular part. Only one part type is produced at a time. Setup for the production of a new part type, if possible at all, takes hours or even days. The machines are linked by automated material-handling devices such as conveyors. The structure of a transfer line presupposes that there is a well-defined, rigid process sequence for the different parts, and that the lot size of each part is high enough to guarantee that the capacity of the equipment will be fully exploited and not wasted on setup.

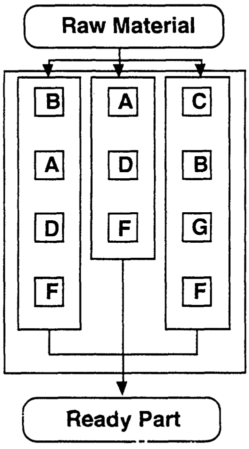

Machines/Resources are grouped in lines according to the operation sequence of one or more part types

Letters refer to individual machines/resources

FIG. 11. Schematic of a flow line.

In the actual manufacturing world, these standard system structures often occur in combinations, or with slight changes. The choice of a structure depends on the design of the parts to be manufactured, the lot sizes of the parts, and market factors such as the required responsiveness to market changes. In general, job shops and project shops are most suitable for small–lot-size production, flow lines are most suitable for large–lot-size production, and cellular systems are most suitable for production of lots of intermediate size. Job shops are most suitable for the low-volume production of multiple part types that are very dissimilar. They possess general-purpose machines and flexible, manual material handling, which are ideal for this situation. Cellular systems are most suitable for the manufacture, in low-to-medium production volumes and lot sizes, of part types with enough similarity to be clustered into part families. Flow lines are best suited to the high-volume, high–lot-size production of a single part type or a few very similar part types. This is a consequence of dedicated machines and material-handling equipment.

Another substantial part of a manufacturing company's production facilities is the assembly system. Assembly systems can be categorized according to the motion of parts and workplaces. Stationary part systems are usually employed for large assemblies, such as airplanes, which are difficult to move around. Moving part systems can be divided into stationary workplace systems, in which parts are brought to stationary workplaces, and moving workplace systems, in which the workplaces move along with the parts.

The fundamental activity in design is decision making: The design of a manufacturing system is the process of deciding the values of the decision variables of the manufacturing system, where the decision variable might be, for example, the number of lathes. Methods and tools which are useful for making design decisions fall into three broad categories: operations research, artificial intelligence (*q.v.*) and simulation. The divisions between these categories are fuzzy. For example, simulation is often considered to be an operations research tool, while the mathematical programs of operations research are solved using search methods studied in artificial intelligence. The categories are an organization convenience; in truth, they overlap to a considerable extent.

3.2 Operation of Manufacturing Systems

The operation of a manufacturing system is the complex task of planning the material and information flows in the system. Proper material flow is what enables a manufacturing system to produce products on time and in sufficient quantity. It is a direct consequence of the system's information flows: *Command* information from human planners or from planning software prescribes the material flow in the system, while *sensory* information about the status of the system's resources is used to decide on the appropriate commands. The fundamental activity in the operation of a manufacturing system is thus determining the commands that prescribe the material flow in the system.

Command information flows can be organized into a bilevel hierarchy. High-level commands dictate the flow of materials *into* the manufacturing system. Therefore, they determine the system's workload—namely, how many of each type of part to manufacture in each time period. The time periods involved are typically long, on the order of days or weeks. For this reason, the decision-making activity of generating high-level commands is called *long-term planning*. Planning is concerned with determining the *aggregate timing of production* (when to produce and what quantities to produce).

Low-level commands dictate the flow of materials *within* and *out of* a manufacturing system. They determine which production resources are to be assigned to each operation on each part and when each operation is to take place. The role of lower-level commands is to resolve contention for the resources of a manufacturing system. This contention occurs among the parts released into the system as a consequence of long-term planning decisions. Since low-level commands control individual operations, they must be generated much more frequently than high-level commands. The time between commands is typically on the order of seconds or minutes. For this reason, the decision-making activity of generating lower-level commands is called *short-term dispatching*. Dispatching is concerned with the *detailed* *assignment of operations to production resources.*

Long-term planning and short-term dispatching are collectively labeled, in broad terms, as "production scheduling." We use the specific terms "(long-term) planning" and "(short-term) dispatching" when possible for greater precision.

Academic research in production scheduling has been based primarily on operations research techniques (see OPERATIONS RESEARCH) (more specifically, mathematical programming). Recently, many research efforts have employed artificial intelligence techniques, especially rule-based systems (see ARTIFICIAL INTELLIGENCE). Unfortunately, the use of academic results in industry has been minimal; many academic approaches make assumptions that do not reflect reality. Actual manufacturing environments are extremely variable, and usually cannot be rigidly classified into one of the classical scheduling models often assumed in academic research. In perhaps no other field is the dichotomy between academic research and actual practice more pronounced.

The academic literature contains an abundance of theoretical work on a number of classical scheduling problems. The entities to be scheduled are typically referred to as *jobs*. Usually, each job corresponds to an individual part. The individual production operations that make up a job are referred to as *tasks*. Classical scheduling problems can be categorized on the basis of requirements generation, processing complexity, scheduling criteria, and scheduling environment. *Requirements* are generated either directly by customers' orders or indirectly by inventory replenishment decisions. This distinction is often made in terms of an *open shop* versus a *closed shop*. In an open shop, all jobs are by customer request and no inventory is stocked; production scheduling becomes a *sequencing problem* in which tasks corresponding to the requested jobs are sequenced at each facility. In a closed shop, all customer requests are serviced from inventory and jobs are generally a result of inventory replenishment decisions. For a closed shop, production scheduling involves lot-sizing decisions associated with the inventory replenishment process in addition to sequencing. Depending on how the requirements are generated, the specification of the requirements may be termed as *deterministic* or *stochastic*. For example, in an open shop the processing time for each task of a job may be deterministic if it is known, or may be stochastic if it is a random variable with a specified probability distribution. Similarly, in a closed shop the customer demand process, which drives the inventory replenishment decisions, may be assumed to be stochastic or deterministic.

Processing complexity is concerned with the number of tasks associated with each job and is often classified as one-stage, one processor (facility); one-stage, parallel processors (facilities); multistage, flow shop; and multistage, job shop. In the one-stage, one processor problem, which is also known as the one-machine problem, all jobs consist of one task which must be performed on the one production facility. The one-stage, parallel processors problem is similar to the one-machine problem except that the jobs, which still contain a single task each, may be performed on any of the parallel processors. For the multistage problem, each job requires processing at a set of distinct facilities, where there is a strict precedence ordering of the tasks for a particular job. The flow-shop problem assumes that all jobs are to be processed on the same set of facilities with an identical precedence ordering of the tasks. In the job-shop problem, there are no restrictions on the task sequence for a job, and alternative routings for a job may be allowed. The $n/m/A/B$ notation is often used to represent processing complexity, where

n is the number of jobs. Each job is assumed to consist of a number of tasks.
m is the number of resources (machines).
A describes the flow pattern or discipline within the system. When $m = 1$, A is left blank. A may be
  F for the flow-shop case, i.e., the machine order for all jobs is the same.
  P for the permutation flow-shop case. Here not only is the machine order the same for all jobs, but the job order is the same for each machine. Thus, a schedule is completely specified by a single permutation of the numbers $1, 2, 3, \ldots, n$, giving the order in which the jobs are processed on each and every machine.
  G for the general job-shop case where there are no restrictions on the form of the technological constraints (constraints

on the order in which the machines may process the jobs or the order in which the jobs may visit the machines).

B describes the performance measure by which the schedule is to be evaluated.

As an example, *n/2/F/(total production time)* represents the *n*-job, two-machine, flow-shop problem where the aim is to minimize the total production time of the *n* jobs.

Scheduling criteria are measures by which schedules are to be evaluated, and may be classified broadly into schedule costs and schedule performance measures. Schedule costs include fixed costs associated with production setup or changeover, variable production and overtime costs, inventory holding costs, shortage costs for not meeting deadlines or for stocking out, expediting costs for implementing "rush" jobs, and costs for generating the schedule and monitoring the progress of the schedule. Common measures for schedule performance are the utilization level of the production resources, the percentage of late jobs, the average or maximum tardiness for a set of jobs, and the average or maximum flow time for a set of jobs.

Tardiness is the positive part of the difference between a job's actual completion time and its desired completion time. The flow time for a job is the difference between the completion time of the job and the time at which the job was released to the manufacturing system. In most production environments, the schedule evaluation is based on a mixture of both cost and performance criteria.

The *scheduling environment* deals with assumptions about the certainty of information regarding future jobs and can be classified as being static or dynamic. In a static environment, the problem contains a finite set of fully specified jobs; no additional jobs will be added to this set, and none of the jobs in this set will be altered. In a dynamic environment, the problem contains not only an initial set of known jobs but also additional jobs which arrive over future time periods. The scheduling problem in most production environments is stochastic and dynamic; however, most models for scheduling problems are deterministic and static (Graves, 1981).

Scheduling methods that either give some guarantee as to the insensitivity of the schedule to future disruptions or explicitly reflect the uncertain nature of the available information are the objective of control theory (Rodammer, 1988). Control theory attempts to command material flow in a way that is robust with respect to disruptions such as machine failures, operator absences, material shortages, and surges in demand (Akella *et al.*, 1984; Gershwin *et al.*, 1986), and with respect to the absence or inaccuracy of status information (Abraham *et al.*, 1985). Control theory views scheduling as a dynamic activity (Fig. 12) where the production scheduling problem is really one of understanding how to reschedule (Rodammer, 1988). The system under control operates on a sequence of inputs $u(t)$, yielding a sequence of outputs $y(t)$. The outputs at any time are a function of the state of the system $x(t)$, where the state $x(t)$ of the system is a function of the inputs $u(t)$. A general statement of the control problem is to determine a means to guide the system state (and thus the system output) through time according to some trajectory that satisfies the constraints of the system (e.g., maximum number of working hours per week) and that simultaneously satisfies some set of performance criteria.

The inputs to a manufacturing system include production orders, raw material and labor, disturbance inputs, and controlled inputs. Production orders specify the quantities of various jobs to be processed and the dates at which these jobs are due. Disturbance inputs include machine failures and labor outages, over which the scheduler has little influence. Controlled inputs include scheduling, maintenance, and overtime decisions, which the scheduler can regulate within bounds. The state of a manufacturing system defines the levels for all completed and partially completed jobs, the status of all machines (whether active, idle, or under repair), the availability of labor, and the inventories for all materials. Outputs of the manufacturing system may be defined as any portion of the state—for example, the inventory levels of all jobs ready for shipment on a specified date (Fig. 12).

Since the formulation of the production scheduling problem using control theory is relatively recent, adequate tools for solving it are still the subject of continuing research (Rodammer, 1988; Ho, 1984; Suri and Dille, 1984). Artificial intelligence (AI) techniques, in particular rule-based systems, have been

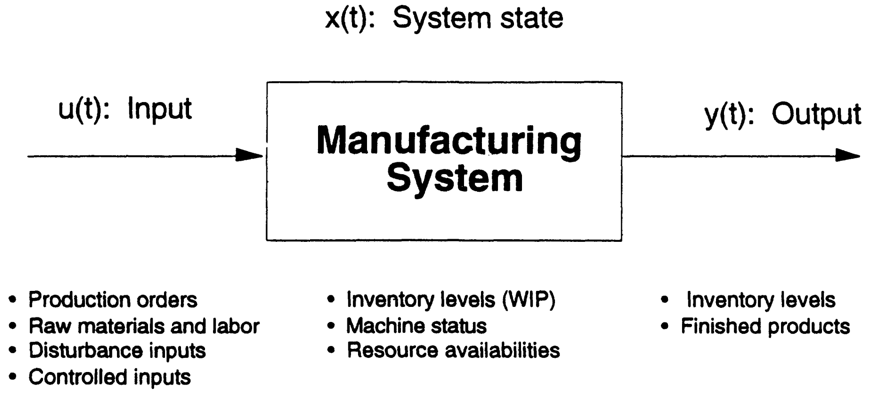

FIG. 12. A Manufacturing system from a control theory perspective (from Rodammer, 1988).

a recent focus of academic research in production scheduling. Rule-based systems seek to capture generic scheduling rules that are applicable to a wide range of situations. In general, these rules must be tailored to meet specific system requirements (Bensana *et al.*, 1988). Knowledge in a rule-based system can be classified into static knowledge (or data) and dynamic knowledge (or solution methods). Static knowledge includes all information about the manufacturing system itself (e.g., the number and types of machines) and the production objectives (e.g., the part types to be produced, along with their processing sequences, quantities, due dates, etc.). Dynamic knowledge describes available expertise about how to derive feasible schedules and consists of theoretical expertise, which refers to operations research techniques that deal with the management of time and resources; empirical expertise, which consists of heuristic dispatch rules; and practical dedicated expertise, which is provided by experienced shop floor supervisors who are aware of the constraints that have to be taken into account when preparing the schedule. In a rule base, these types of expertise are captured in the form of rules. They are invoked by higher-level rules that select a particular type of expertise on the basis of the status of the manufacturing system and the production scheduling problem to be solved.

In industry, it is customary to divide the operation of a manufacturing organization into three major levels (Fig. 13): strategic planning, operations planning, and detailed planning and execution (Anthony, 1965). At the strategic level, few decisions are made, but each decision takes a long time, and its impact is felt throughout the organization. At the detailed planning and execution level, many decisions are made, each requiring a much shorter time. Although the impact of each decision is local in time and in place, the great number of decisions, taken together, can have a significant impact on the performance of the organization. The characteristics of decisions at the operational planning level lie between those of the other two levels.

Strategic planning is performed by the top management of the organization. Several basic questions are asked and answered at this level, including "What is our business?" and "How will we compete?" At this level, the decision is made as to which manufacturing system structure (flow line, cellular system, job shop, project shop, or continuous system) is most consistent with the organization's overall strategy. Furthermore, a timetable is established for the major activities of the organization, such as the entry into a new market or the acquisition of additional manufacturing capacity. Finally, performance measures are established for the organization, such as return on investment, market share, earnings, growth levels, and so forth.

The resulting overall corporate strategy is used to guide the development of functional plans, primarily the marketing plan, the financial plan, and the manufacturing plan. The marketing plan specifies how the organization will compete in the marketplace. Included in this plan are forecasts of dollar sales and unit sales, and descriptions of distribution channels. Similarly, the financial plan

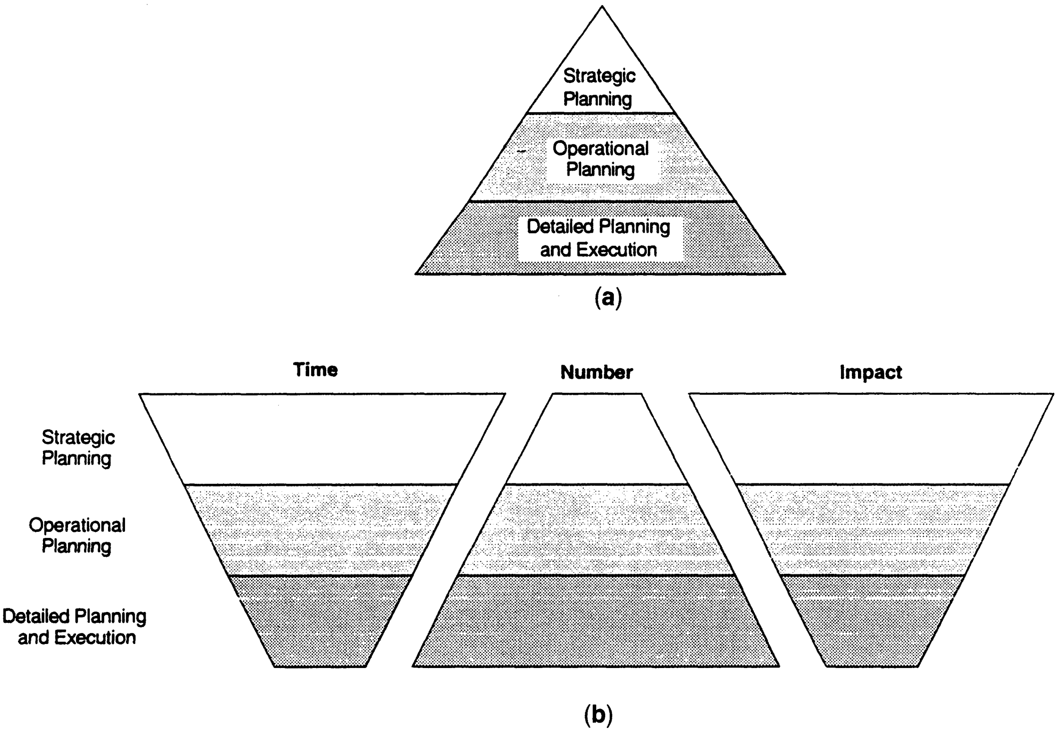

FIG. 13. (a) The levels of manufacturing operation and (b) the characteristics of decisions made at each level.

specifies the role that finance will play in the overall corporate strategy. Included are plans for cash flow and capital expenditures. Finally, the manufacturing plan specifies what manufacturing resources will be needed, how the resources will be deployed, and how much is to be produced. The output component of the manufacturing plan is stated in aggregate terms, such as annual or quarterly units of production, for each major product line (Berry *et al.*, 1979). This plan also sets guidelines and policies within which manufacturing is expected to perform. Ideally, any conflicts in strategies and policies among functions will be solved at the strategic level. Once acceptable strategic functional plans have been developed, the next step is to develop a more detailed operations plan.

In operational planning, the objectives of the strategic manufacturing plan are converted into more detailed and specific plans. One of the most important of these plans is the master production schedule (MPS). The MPS is expressed in specific product configurations, quantities, and dates and is frequently the key link between top management and the manufacturing facility. It is also a key link between manufacturing and marketing and serves to balance the demands of the marketplace against the constraints of the manufacturing facility.

The master production schedule is one of the inputs to material requirements planning (MRP). The other inputs to MRP include, for each type of final product, the existing inventory level, bill of materials, and the production time or order lead time for each individual component. A bill of materials lists the subassemblies of a final product and the individual components that constitute each subassembly. The output of MRP is a plan that describes the quantity of each individual component or subassembly to order or to begin producing in each time interval. Since the generation of the material plan requires a vast amount of data processing when many part types with possibly many layers of subassemblies are involved, it is usually handled by computer programs called MRP systems. MRP systems calculate material requirements on the basis of infinite capacity, meaning that the production and assembly times assumed for individual components and subassemblies do not incorporate waiting times due to resource contention.

Neither of the planning activities introduced thus far, master production scheduling and material requirements planning, incorporates any consideration of whether the manufacturing system has enough capacity to achieve the demanded levels of production. This is the role of capacity planning. Capacity planning provides projections of the capacity needs implied by the material plan, so that timely actions can be taken to balance the capacity needs with available capacity. In general, four variables may be adjusted to balance required and available capacity: inventory levels, the number of shifts worked, the quantity of resources used for production, and the amount of orders accepted for production. The first three options adjust the effective available capacity of the manufacturing system to fit the required capacity. The last option results in the modification of the master production schedule, which in turn results in a more feasible material plan. It adjusts the "required" capacity to meet the available capacity.

The third and final level of planning is the detailed planning and execution level. The shop-floor control system is the main operational system at this level. The shop-floor control system is responsible for detailed planning, such as determining the priority sequence of orders in a queue, and for monitoring the execution of the detailed plans, such as capacity use at the machine center (i.e. input/output control). The shop-floor control system is also responsible for collecting data from the shop floor and directing a flow of information back up to the operations planning level. After interpretation and aggregation, this information will pass from the operational level to the strategic level (Fig. 14).

Master production scheduling is a form of aggregate planning. Typically, it considers only finished products, not the individual components which make up those products. The objective of master production scheduling is to develop an aggregate production schedule that will meet seasonal requirements for finished products and at the same time minimize the incremental costs incurred. It is important to identify and measure these costs so that alternative master production schedules may be evaluated on the basis of total cost. Some costs that may be relevant are payroll costs; costs of overtime, second shifts, and subcontracting; costs of hiring and laying off workers; costs of excess inventory and backlog; and costs of production rate changes.

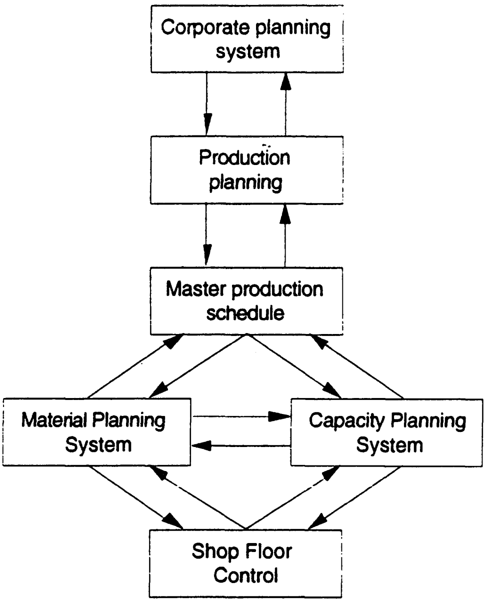

FIG. 14. Major components of the manufacturing planning system (from Melnyk, 1985).

The costs that are included in the total cost model should vary with changes in the decision variables. In a MPS, these decision variables typically denote the size of the work force to employ, the number of overtime hours to schedule, and how many of each finished product to produce during each time interval of some planning horizon. However, the behavior of cost with respect to changes in these decision variables is not necessarily easy to quantify. Often, approximations are made by assuming the costs to be linear or quadratic functions of the appropriate decision variables. This permits the use of optimization techniques such as linear programming.

3.3 Areas for Further Research

In manufacturing, the demand to develop more efficient manufacturing systems and processes will always exist, and currently there are numerous topics on which today's researchers are spending their efforts. Much of the effort is toward the advancement of computer-integrated manufacturing (CIM), especially in control, because as manufacturing systems become larger and as the processes become more complex, efficient control is possible only with the integration of computers. Also, many advancements in the field of artificial intelligence (AI) propagate and help to advance manufacturing, since AI tools are becoming more commonplace for industrial applications. Rapid prototyping, with the concept of "overnight" development of a product from the draftboard to the physical prototype, is another current research topic. A topic of high priority in research is the implementation of environmentally conscious manufacturing. As society slowly comes to the realization that the world has limited resources (and as governments are starting to enact environmental legislation), there is a need in industry to make current manufacturing practices more environmentally friendly. Today's researchers in manufacturing are making advancements in all these areas to satisfy the needs of today and tomorrow.

GLOSSARY

Brazing: Any of several joining processes similar to welding in which a high-temperature molten filler metal is distributed in a joint by capillary action. Note that melting of the base metal does not occur.

Casting: A process where a molten substance is poured into and solidifies inside a mold that is the negative image of the desired part shape.

Cellular System: A manufacturing configuration in which several different machine tools are grouped together into units called cells.

Computer-Aided Design (CAD): A process where computers are used to assist in the design of parts, often by manipulating three-dimensional computer descriptions of the part shape.

Computer-Aided Manufacturing (CAM): The use of computers to assist in manufacturing as in computerized process planning.

Computer Numerically Controlled (CNC): Employing a self-contained control system using a dedicated computer in conjunction with stored digital instructions.

Continuous System: A manufacturing system in which the product is continuous, rather than discrete. Often used in the chemical or textile industry.

Deforming Process: A manufacturing process where the shape of a part is altered while cohesion among the constituent particles is maintained.

Die: A generic term to describe tooling used to cut or form material.

Electric-Discharge Machining (EDM): A material-removal process that employs high-voltage electric discharges to ablate material.

Flexible Manufacturing System (FMS): An automated manufacturing system containing several different machine tools capable of making many different part geometries at various production volumes.

Flow Line: A manufacturing configuration in which the part is transported through a linear arrangement of machine tools that perform the manufacturing operations.

Forming: A manufacturing process in which the cohesion among the particles is increased so as to create a new part, e.g., casting.

Hardness: In manufacturing, defined as a material's resistance to mechanical indentation and thus closely correlated to the material yield stress. Units of measurement depend on the scale employed; Rockwell B, Rockwell C, and Vickers scale are commonly used.

Job: An individual part to be manufactured.

Job Shop: A manufacturing configuration in which several diverse machine tools are grouped together such that they may efficiently manufacture many different low-volume parts.

Joining Process: A manufacturing process in which two or more parts are brought together to create a new part, e.g., welding.

Lathe: A machine tool used for turning.

Manufacturing: The transformation of materials and information into goods for the satisfaction of human needs.

Material Requirements Planning (MRP): A plan (usually implemented through computer software) that describes the material requirements for the production of a part or a set of parts.

Master Production Schedule (MPS): A master plan for the production of a part or a set of parts. It serves as a link between management and the manufacturing facility.

Milling: A material-removal process in which a rotating cutting tool is applied to a part in order to create a desired shape.

Mold: The negative image of a desired part shape, used for casting operations.

Numerical Control (NC): A technique for controlling actions of a machine tool and similar equipment with a digital controller and numerical data.

Process Control: A method for ensuring that a manufacturing process produces useful parts by adjusting process parameters so that process outputs meet target values.

Project Shop: A manufacturing configuration in which the part remains stationary while materials and equipment are brought to the part in order to perform manufacturing operations.

Roughness Average (R_a): The arithmetic average of the absolute values of the measured profile height deviations taken within the sampling length and measured from the centerline.

Removal Process: A manufacturing process in which material is removed from the part in order to create a specified geometry, e.g., turning.

Resource: A machine capable of performing a manufacturing **task**.

Scheduling: The process of assigning tasks to manufacturing **resources**.

Simulation: In manufacturing, the representation of a manufacturing system by computers.

Soldering: A joining process in which the filler metal melts at a low temperature and is distributed through the joint by surface tension and capillary action. Note that melting of the base metal does not occur.

Task: An individual manufacturing operation; several tasks comprise a **job**.

Turning: A removal process in which a workpiece is rotated against a tool to form a desirable rotationally symmetric shape.

Welding: A joining process in which heat or pressure or filler materials are employed to achieve cohesion between two parts.

Works Cited

Abraham, C., Dietrich, B., Graves, S., Maxwell, W. (1985), *A Research Agenda for Models to Plan and Schedule Manufacturing Systems*, Cambridge, MA: Massachusetts Institute of Technology, Alfred P. Sloan School of Management.

Akella, R., Choong, Y., Gershwin, S. B. (1984), *IEEE Trans. Components Hybrid Manufact. Technol.* **CHMT-7** (3), 225–238.

Anthony, R. N. (1965), *Planning and Control Systems: A Framework for Analysis*, Boston: Harvard University Graduate School of Business Administration.

Bensana, F., Bel, G., Dubois, D. (1988), *Int. J. Prod. Res.* **26**, 795–819.

Berry, W. L., Vollman, T. E., Whybark, D. C. (1979), *Master Production Scheduling: Principles and Practice*, Washington, DC: American Production and Inventory Control.

Box, G., Kramer, T (1990), *Statistical Process Control and Automatic Process Control—A Discussion Report #42*, Center for Quality and Productivity Improvement, College of Engineering, University of Wisconsin, Madison, WI.

Browne, J., Rathmill, K., Sethi, S. P., Stecke, K. E. (1984), *FMS Magazine* April, 114–117.

Choobineh, F. (1986), in: Choobineh, F., Suri, R. (Eds.), *Flexible Manufacturing Systems Current Issues and Models*, Atlanta: Industrial Engineering and Management Press, pp. 269–281.

Chryssolouris, G. (1991), *Manufacturing Systems: Theory and Practice*, New York: Springer-Verlag.

Gershwin, S. B., Hildebrant, R. R., Suri, R., Mitter, S. K. (1986), *IEEE Control Syst.* **6** (2), 3–15.

Graves, S. C. (1981), *Operations Res.* **29**, 646–675.

Ho, Y. C. (1984), in: *Proceedings of the First ORSA/TIMS Special Interest Conference on Flexible Manufacturing Systems*, pp. 372–378.

Kalpakjian, S. (1989), *Manufacturing Engineering and Technology*, Reading, MA: Addison-Wesley.

Lange, K. (1985), *Handbook of Metal Forming*, New York: McGraw-Hill.

M.I.T. (1989), *The Working Papers of the M.I.T. Commission on Industrial Productivity*, Cambridge, MA: M.I.T. Press.

Melnyk, S. A., Carter, P. L., Dilts, D. M., Lyth, D. M. (1985), *Shop Floor Control*, Homewood, IL: Dow Jones–Irwin.

Metcut Research Associates (1980), *Machining Data Handbook, 3rd ed., Vol. 2*, Cincinnati, OH: Metcut Research Associates, Inc.

Rodammer, F. A. (1988), *IEEE Trans. Syst. Man Cybernet.* **18** (6), 841–851.

Schwartz, M. M. (1979), *Metal Joining Manual*, New York: McGraw-Hill.

Shey, J. (1987), *Introduction to Manufacturing Processes*, New York: McGraw-Hill.

Suri, R., Dille, J. W. (1984), in: *Proceedings of the First ORSA/TIMS Special Interest Conference on Flexible Manufacturing Systems*, pp. 379–384.

Weck, M. (1984), *Handbook of Machine Tools*, New York: John Wiley and Sons.

Further Reading

Chryssolouris, G. (1991), *Manufacturing Systems: Theory and Practice*, New York: Springer-Verlag.

French, S. (1982), *Sequencing and Scheduling: An Introduction to the Mathematics of the Job Shop*, Chichester, England: Ellis Horwood Ltd.

Kalpakjian, S. (1984), *Manufacturing Processes for Engineering Materials*, Reading, MA: Addison-Wesley.

Law, A. M., Kelton, W. D. (1991), *Stimulation Modeling and Analysis*, New York: McGraw-Hill.

Niebel, B., Draper, A., Wysk, R. (1989), *Modern Manufacturing Process Engineering*, New York: McGraw-Hill.

Shey, J. (1987), *Introduction to Manufacturing Processes*, New York: McGraw-Hill.

Society of Manufacturing Engineers (1976), *Tool and Manufacturing Engineers Handbook*, New York: McGraw-Hill.

Weck, M. (1984), *Handbook of Machine Tools*, New York: John Wiley and Sons.

MANY-BODY EFFECTS

See COLLECTIVE PHENOMENA IN SOLIDS

MARINE GEOPHYSICS

JOHN C. MUTTER, SUZANNE CARBOTTE, AND JAMES COCHRAN, *Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York*

Introduction 305
1. **Methodologies and Instrumentation** 307
1.1 Acoustic Imaging 307
1.1.1 Bathymetric Systems 307
1.1.2 Side-Scan Sonar 309
1.2 Seismic Methods 310
1.2.1 Crustal Velocity Distribution from Seismic Kinematic Data 312
1.2.2 Crustal Velocity Distribution from Seismic Wave Forms ... 315
1.2.3 Seismic Imaging Using the Scattered Wave Field 318
1.3 Potential Fields 319
1.3.1 Studies of the Gravity Field 319
1.3.2 Magnetic Field Studies 322
2. **Marine Geophysical Methods and the Study of Midocean Ridges** 323
2.1 The Shifting Paradigm I: Controls on the Form of Midocean Ridges 324
2.2 The Shifting Paradigm II: Magma Chambers at Midocean Ridges 327
Glossary 330
Works Cited 332
Further Reading 333

INTRODUCTION

The term *marine geophysics* applies to a branch of the geosciences in which a group of physically based techniques is applied to study the floor of the oceans and the subsurface layers beneath. Marine geophysical techniques *per se* are tools of the scientists; the objective of marine geophysical research is to apply these tools to the furthering of our understanding of the nature and evolution of the ocean basins of our planet. Marine geophysical techniques fall broadly into three groups:

1. high-frequency acoustics for mapping the seafloor;
2. low-frequency acoustics (seismic) to map structure beneath the seafloor; and
3. potential-field studies to map mass distribution and magnetic properties in the earth.

The three approaches provide different but complementary information that can be used by scientists to build an understanding of processes that shape ocean basins. The high-frequency acoustic methods give a measure of the shape of the seafloor and its acoustic response. Different geological features give different strengths of return, and thus the pattern of returned energy can be used as a proxy for a geological map of the seafloor. Low-frequency acoustic methods provide an image of the subsurface, allowing scientists to map the internal components of the crust. Here the sound propagation speed and scattered response of internal layers are used as proxy for geological structure. Potential-field methods can be used to derive estimates of the distribution of anomalous masses and the varying magnetization of deep rock layers. Acoustic methods rely on analyzing returns from sound energy transmitted into the Earth and hence are limited in depth to the depth of sound penetration. While information from potential-field methods is typically of lower resolution than that available from acoustic methods, these data can be used to deduce information about structure at much greater depth than is available from acoustic methods because gravity and magnetic fields are the integrated response to mass and magnetization distributions at all depths in the Earth.

Marine geophysical research has no central theme but rather has been applied to a

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

very wide variety of problems specific to understanding the oceans. These range from the development of sediment accretion in zones of plate convergence, through the history of seamount construction, to rifting that leads to the rupturing of continents to form passive continental margins. One research area that has generated considerable interest in the last several years is in developing an understanding of the nature and evolution of midocean-ridge spreading centers. This topic area has seen extensive application of all three major classes of geophysical techniques, and we use this example here as a useful basis for discussion of marine geophysical methods and their applications.

The globe-encircling filigree of midocean ridges (Fig. 1) is the locus for generation of new oceanic lithosphere by the process of seafloor spreading. As part of this process, molten magma rises to the surface from the deep mantle, freezes, and forms the basaltic oceanic crust. The molten magma is generally thought to have been stored for some period of time in a shallow-level magma reservoir, or chamber, because the chemistry of volcanic rocks recovered from spreading centers is not consistent with magma having been in equilibrium with the parent mantle. The nature of the magmatic system is one problem that has drawn special attention, as it has become increasingly recognized that magma is delivered to the crust in a highly nonuniform way. Along the axis of a ridge, there exist regions that are much more volcanically active than others. The mechanism by which melt is focused in specific regions along the ridge is not well understood, and marine geophysical approaches, combined with geochemical studies of rocks recovered from the axes of midocean ridges, have proven particularly insightful in addressing this problem.

In the following, we present a discussion of tools and methodologies most commonly used in marine geophysical studies. The list is not intended to be inclusive of all marine geophysical tools, to avoid producing a simple listing of techniques. Following that, we describe some uses of these methodologies, using the example of midocean-ridge spreading systems as an area of current research focus. Again, we do not mean to suggest that this is the only, or the most important, focus

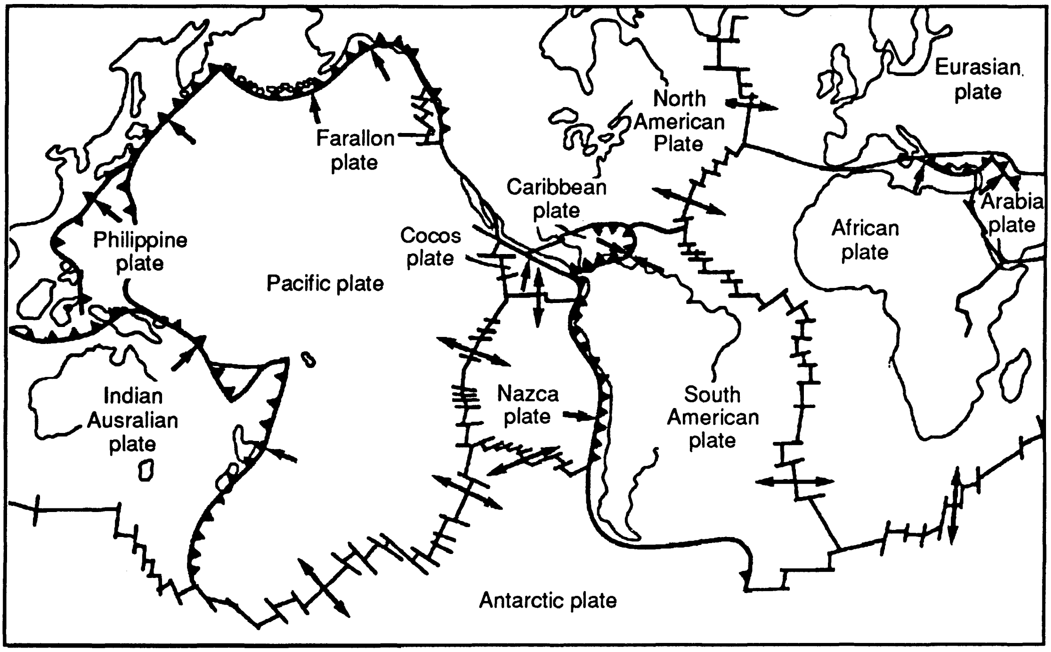

FIG. 1. Global distribution of lithospheric plate boundaries. Regions where new lithosphere is being created at midocean-ridge centers of seafloor spreading are shown with double arrows. The orientation of the arrow indicates the direction of relative motion of the lithosphere created at the spreading centers.

of research, only that it provides a convenient example.

1. METHODOLOGIES AND INSTRUMENTATION

1.1 Acoustic Imaging

Seafloor imaging involves the collection of two classes of acoustic data, bathymetric and side-scan sonar. Both data types are obtained using systems comprised of the same fundamental components: a transmitter that emits sound toward the seafloor, a receiver that detects reflected acoustic energy, and a timing device that records the time between transmitted and received signal. Both kinds of data are obtained using high-frequency sources that achieve minimal sub-bottom penetration. Bathymetric systems measure the time elapsed between transmitted and received acoustic pulses. Using the velocity of sound in seawater, travel times are converted to seafloor depths. Side-scan sonar systems measure the amplitude of energy backscattered from the seafloor, providing proxy information on bottom lithology and structure.

1.1.1 Bathymetric Systems Acoustic methods were first developed for measuring depth to the ocean bottom in the late 1920s. These early systems employed a single wide beam of sound insonifying an area of the sea floor with a "footprint" several kilometers in diameter (Fig. 2, upper illustration). This large area resulted because the beam sent into the ocean was not a narrow pencil of sound but had a conelike shape—the deeper the seafloor, the greater the footprint area. Any single depth point from such a system therefore represented an average depth in the insonified area, and the beamwidth fundamentally limited resolution of seafloor features. During the next 40 years, narrower beam systems were developed, providing improved resolution, but depth sounding remained limited to profiles of a series of single-point measurements (taken to be the center of the beam) beneath the ship.

In the 1970s, the first so-called multibeam systems became available. Instead of sending out a single beam of sound energy, these systems propagate sound from an array of transmitting devices on the hull of a vessel, typically arranged across the direction of travel of the vessel. With these instruments, a swath of seafloor transverse to the ship's track is insonified (Fig. 2, lower panel). The width of current multibeam swaths is 90° to 120°, so that the swath width on the seafloor is more than twice the water depth. This leads to efficient collection of virtually continuous high-resolution bathymetry by a survey strategy akin to mowing a lawn. The first multibeam system widely used for deep sea research was the 54° beamwidth Seabeam system, built by General Instruments Corporation, Westwood, Massachusetts (Renard and Allenou, 1979). Multibeam systems have become the standard tool for mapping the seafloor and have led to an explosion in our knowledge of its shape and properties.

Data obtained from multibeam systems are governed by insonification frequency, number of beams, beamwidth, and ping repetition rate. Operating frequency and beamwidth determine the system resolution, the number of beams and beamwidth control across-track data density and swath width, and the ping repetition rate governs the along-track data density. Choice of operating frequency, typically 12 to 16 kHz for deep-water systems and 30 to 200 kHz for shallow-water systems, is constrained by the absorption of high-frequency acoustic energy by seawater. The aperture or beam width, typically 1° to 3°, determines the size of the seafloor footprint insonified, and this area will vary with distance from nadir and with water depth. The density of information along the survey track is controlled by ping rate—the interval between successive acoustic transmissions. Present systems cannot process data for which more than one transmission is in the water column at the same time, and so ping rate is controlled by water depth. Thus, the data density along track is controlled by water depth and speed of the ship. Typically, then, the along- and across-track data densities are quite different, and data are usually averaged to roughly equalize densities.

Multibeam surveys typically involve a strategy to obtain essentially 100% coverage of the seafloor in an area of interest. Lines will be surveyed along an azimuth perpendicular to the dominant trend of seafloor features, with a few lines run parallel, or at another angle. Because swath width on the

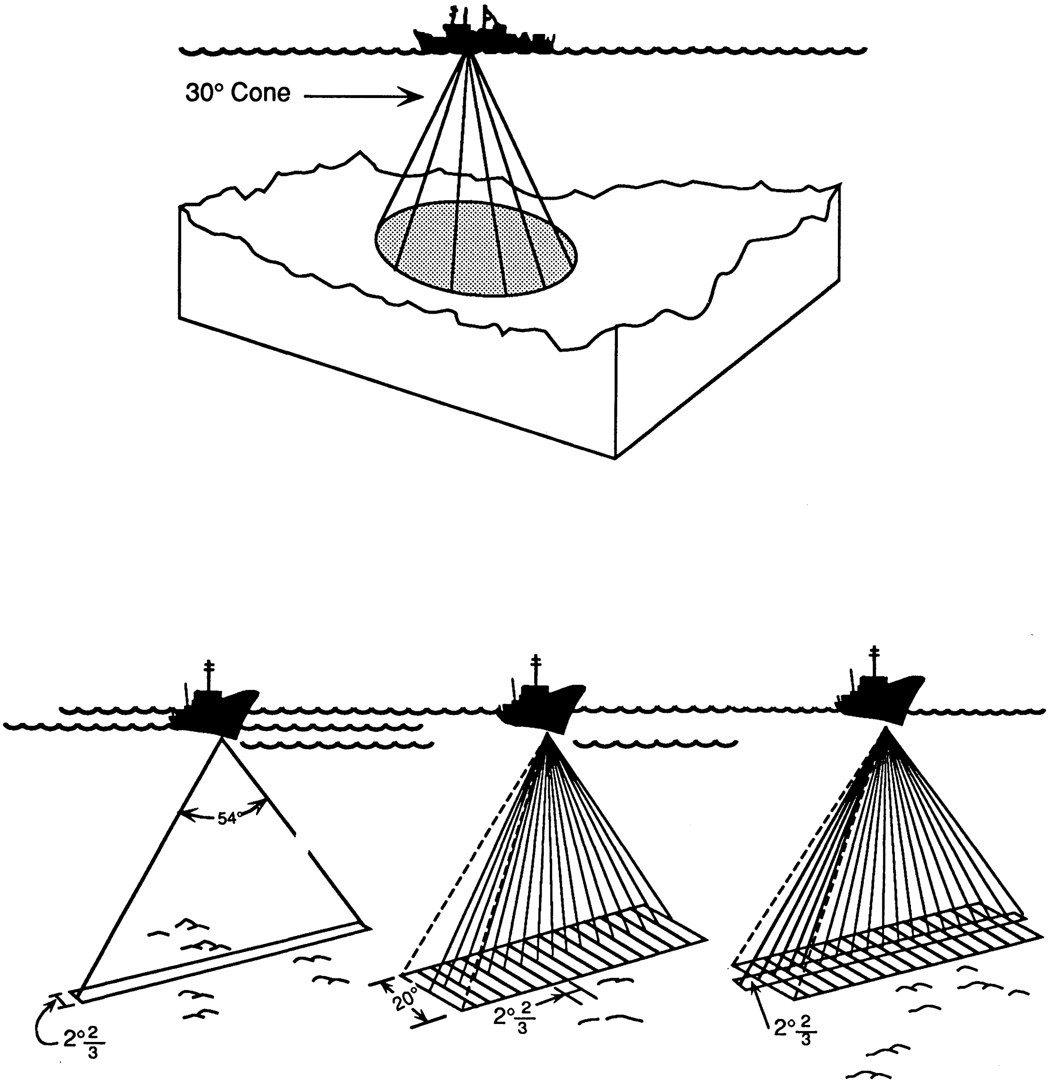

FIG. 2. The shape of the seafloor first became known in some detail by the use of single-beam echo sounders (upper illustration) which propagated a cone-shaped beam of sound from a single transmitter/receiver on the hull of a vessel and timed the return. Converting the time to the depth of the seafloor only required knowledge of the speed of propagation of sound in water, which is near to 1500 m/s in most oceans. The modern generation of bottom mapping tools are multibeam systems (lower illustration) which propagate a swath of sound energy into the water from an array on the vessel's hull. This achieves a rapid mapping of a wide region of the seafloor, so complete coverage can be obtained by a survey vessel tracking back and forth as if mowing a lawn. Figures after Renard and Allenou (1979) and Johnson and Helferty (1990).

seafloor is controlled by water depth (the greater the depth, the wider the swath), lines run at a uniform spacing will sometimes have overlapping swaths and sometimes will include gaps in coverage. These effects have to be accounted for by editing (or interpolation) in subsequent data processing. Navigation is usually achieved by Global Positioning System (GPS) satellites that effectively provide a constellation of celestial references giving very accurate position information on a virtually continuous basis. Data can be displayed as plots of contoured swath data, presented as contoured maps or as three-dimensional images such as the map of the East Pacific Rise (Plate 1).

1.1.2 Side-Scan Sonar Side-scan sonar systems recover information on seafloor properties, rather than its shape, by measuring the strength of returned acoustic energy as a function of time after an acoustic pulse is transmitted. The returned energy includes both backscatter (diffracted) and specular energy reflected off seafloor slopes. Backscatter results from the diffraction of sound due to seafloor roughness that is comparable in size to the wavelength of the acoustic pulse. Because rougher surfaces will diffract sound more efficiently than smooth surfaces, side-scan images can be used to distinguish different types of seafloor terrains (e.g., volcanic versus sedimented) (Fig. 3) In addition, side-scan imagery forms an excellent database for studying tectonic features such as faults where unsedimented scarps produce strong linear acoustic returns and shadows.

Side-scan sonar systems were first used to map geological features of the seafloor in the late 1950s, primarily on continental shelves. Use of these systems in deep ocean research did not begin until the late 1960s to early 1970s with development of instruments such as the long-range GLORIA side-scan sonar developed by the Institute of Ocean Sciences, U.K., and the Deep-tow system of the Scripps Institute of Oceanography.

Side-scan sonar systems are typically housed within a vehicle or "fish" towed be-

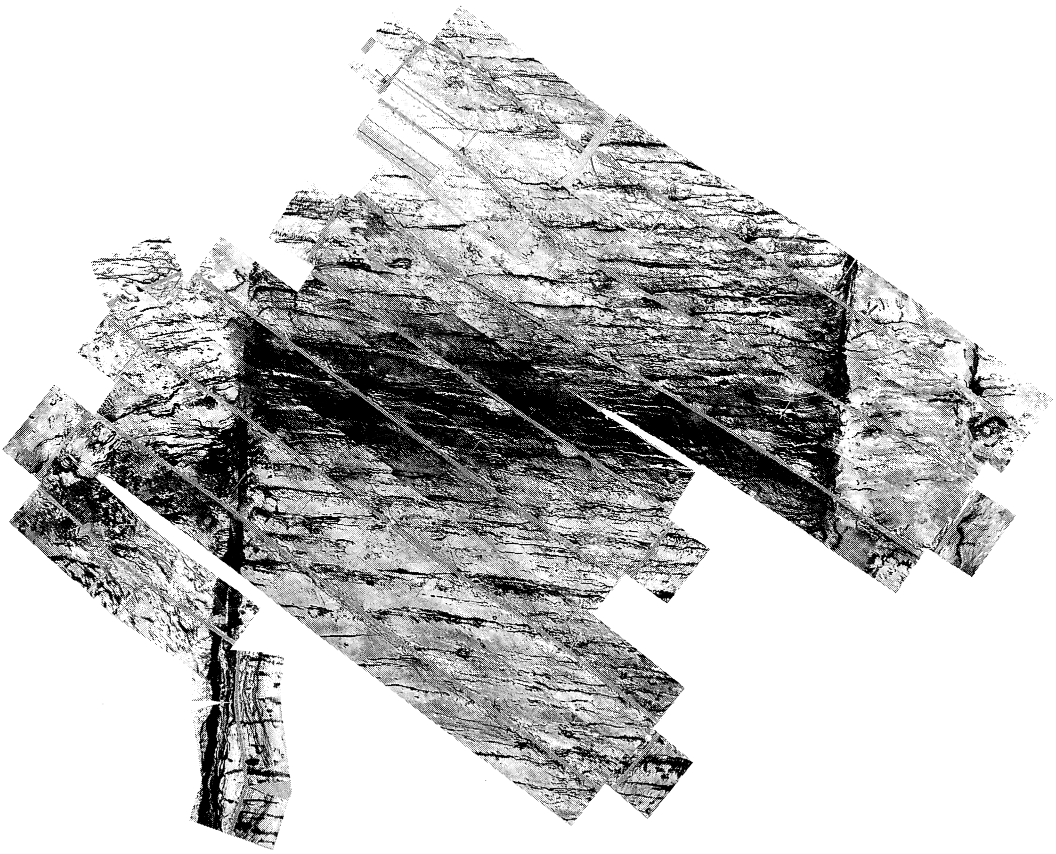

FIG. 3. A typical image of the seafloor made by a side-scan sonar system does not reveal the shape of the bottom (newer systems are able to recover bathymetric information also), but allows the texture to be revealed in detail. Dark and light areas correspond to areas of differing reflectivity. Strong, dark linear features correspond to faults that disturb the seafloor. The mid-ocean ridge is the dark region in the center of the map. Other dark regions are believed to correspond to regions of relatively young lava flows that have been recently extruded onto the seafloor. To first order, the dark, more reflective areas are thought to be formed more recently than the less reflective white areas, which may have a thin sediment cover. Tracks run by the ship follow grey strips of no data within the center of each swath. Data courtesy of K. C. Macdonald.

hind the ship (Fig. 4). A coaxial cable connects the fish to the ship and is used to transmit power and control signals down to the fish, and side-scan and fish attitude data are sent back. Towing the system, rather than having a hull-mounted design, decouples the instrument from ship's motion and reduces ship-generated noise. It also permits the instrument to be kept below the thermocline where water temperatures and hence velocities vary rapidly with depth. Both deep-towed (within 200 m of ocean bottom) and shallow-towed (within 200 m of the sea surface) systems are used. As with multibeam systems, the attenuation of high frequencies in seawater limits surface-towed systems to frequencies less than 30 kHz, whereas operating frequencies of 30 to 150 kHz are typical for deep-towed systems. Wider swaths are insonified with shallow-towed systems and at faster towing speeds, but lower-resolution data is obtained. For example, the SeaMARC II system, built by International Submarine Technology, Washington, and the Hawaii Institute of Geophysics, is towed at depths of ~100 m and at speeds up to 10 knots. It operates at 11 to 12 kHz with a total swath width of 10 km. In contrast, the Scripps's Deep-tow system towed at heights of 150 m off bottom gives a swath width of only 500 to 750 m. The operating frequency of this instrument is 110 kHz, and tow speeds are 1 to 2 knots.

The side-scan system is constructed of parallel linear transducer arrays that scan to either side of the tow fish. The same transducers are used for both transmission and receiving. Slightly different insonification frequencies are used to port and starboard to facilitate identification and rejection of returned acoustic energy from the opposite side of the fish, known as cross-talk. The transducer array emits a fan-shaped acoustic beam that is narrow in the along-track direction (1° to 2°) but wide in the across-track direction (over 50°), much like a multi-beam bathymetric system. Resolutions in both along- and across-track directions are determined largely by this spreading of the emitted beam (Johnson and Helferty, 1990).

A number of corrections must be applied to side-scan data before a meaningful image is produced. These include corrections for sound absorption in seawater and geometric spreading, both of which result in a progressive loss of acoustic energy for returns from greater distances from the fish. To compensate, a time-varying gain is applied, producing a constant average signal level across the swath. A slant-range correction is then needed to convert two-way travel time to horizontal distance from the ship's track to the acoustic target. This correction is based on the assumption that the seafloor is flat and horizontal, and this can give rise to considerable errors in positioning acoustic targets if seafloor relief is actually significant. Recent developments in side-scan sonar design have made possible the use of bathymetric information to place data correctly in space. The SeaMARC II and MR1 systems measure the phase difference between incoming acoustic energy recorded at a pair of transducer arrays mounted on each side of the fish. This phase difference is a measure of the angle of incoming acoustic energy and in conjunction with travel times can be used to obtain bathymetry (Blackinton *et al.*, 1983). At present, these bathymetric data are used during post-processing of select areas to locate side-scan data correctly (Reed and Hussong, 1989).

Image enhancement and analysis techniques similar to those used in digital processing of optical and satellite imagery can also be used in analysis of acoustic data. These include contrast and gain adjustments and filtering to emphasize specific features of interest. Some application of pattern recognition techniques or texture analysis has been made in side-scan processing and will likely see increased development and application in the future (Reed and Hussong, 1989). A wide range of visualization capabilities are presently being developed that will soon be applicable to side-scan imaging. For instance, the three-dimensional bathymetric image (Plate 1) can be "painted" with backscatter strength data to combine information on shape and reflectivity of the seafloor to aid marine geologists in interpreting these data types.

1.2 Seismic Methods

The central thesis underlying the use of seismic methods is that detectable properties of seismic waves are perturbed in their passage through the crust by changes in crustal parameters of interest. Analysis of perturbations in the seismic wave field should then yield information on these parameters. Two

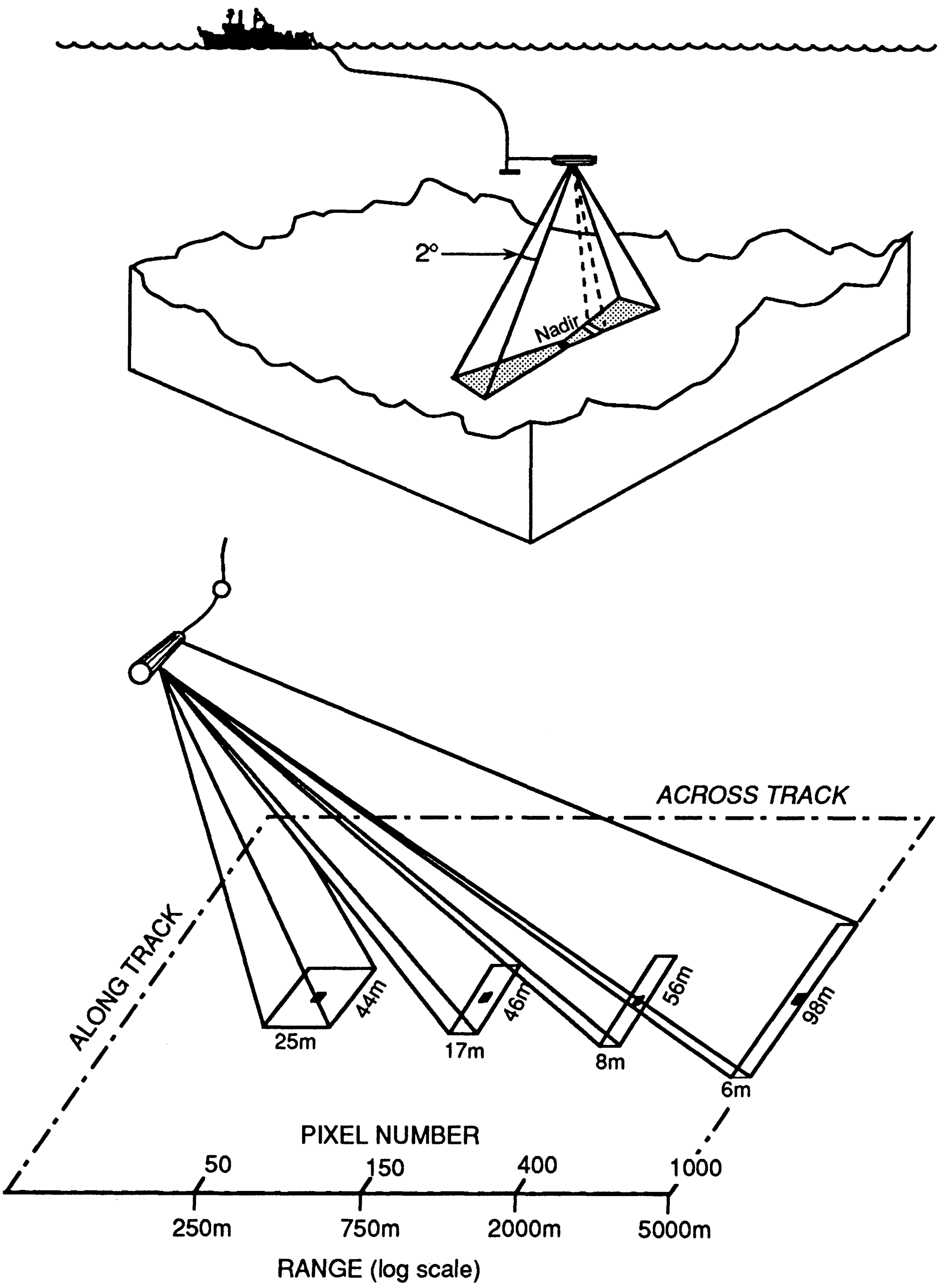
2°
Nadir
ACROSS TRACK
ALONG TRACK
44m
46m
56m
98m
25m
17m
8m
6m
PIXEL NUMBER
50
150
400
1000
250m
750m
2000m
5000m
RANGE (log scale)

FIG. 4. The technology for creating a side-scan image typically involves the use of a towed "fish," which emits an acoustic signal from an array of transducers. Side-scan images measure the strength of the returned energy from the seafloor rather the shape of the seafloor, but some information on the seafloor depth is required to place the side-scan information (here shown as the data pixel) on the seafloor. Note that the "footprint" of individual beams changes with distance from the fish. Different side-scan fish are towed near the surface or near the seafloor to achieve very high resolution. The latter are referred to as deep-tow instruments and must be towed very slowly (less than 2 knots) to ensure that they stay near the bottom. Figure from Johnson and Helferty (1990).

different techniques are commonly employed. In the first, properties of the seismic wave field are used to infer the distribution of seismic-wave propagation speed in the subsurface. In the second, the direct response of the earth to seismic energy is used to create an image of structure. The first approach derives a quantitative estimate of a particular physical parameter. The second achieves a visualization of the Earth from which quantitative information is not readily available. In both cases, the information recovered is a *model* of Earth structure that is much simpler than we believe the real Earth to be. The specific model parametrization and resolution achieved depend on experimental design, data type, and analysis techniques.

1.2.1 Crustal Velocity Distribution from Seismic Kinematic Data To observe the seismic wave field, recording instruments are placed either on the seafloor [ocean bottom seismometers (OBS)] or deployed near the sea surface (sonobuoys or towed hydrophone arrays), and a near-surface seismic energy source (airguns or explosives) is discharged at numerous locations away from the receiver(s) [Figs. 5(a) and 5(b)]. Less common types of experiment involve seafloor shots and receivers [Fig. 5(c)], or surface shots fired to borehole receivers [Fig. 5(d)]. When towed hydrophone arrays are to be employed, both source and receiver can be moved simultaneously to create an expanding spread profile (ESP) [Fig. 5(e); Stoffa and Buhl, 1979]. Expanding spread profiles acquired with large-volume, tuned, multi-airgun source arrays and multichannel hydrophone receiver arrays can achieve an essentially continuous wave-field sampling over source–receiver offsets of up to 100 km. In each configuration, the objective is to record the seismic wavefield at many different distances from the origin point of a seismic source. Changes in seismic travel times and wave form with increasing range from the source can be analyzed to recover information on crustal velocities.

Techniques for estimating one-dimensional $V(z)$ structure rely on a commonly used theoretical abstraction of elastic-wave propagation in terms of wave fronts and their outwardly directed normals, called rays, such as those drawn schematically in Fig. 5. Seismic rays relate closely to classical descriptions of light in terms of geometric optics. If velocity continuously increases with depth, rays that originate from the source nearly vertically will penetrate more deeply than those originating at shallower angles. The depth to which a ray penetrates is described as its turning-point depth. Figure 6 shows rays traced through a structure typical of the crest of the East Pacific Rise spreading center. This includes an interval where velocity decreases with depth because of the presumed presence of a body of near-liquid magma. Rays that penetrate into this region do not have turning points because they bend toward the vertical.

Estimates of $V(z)$ can be derived from two approaches. In *forward modeling,* a trial model function is prescribed, and kinematic variables such as travel times are calculated for that model. Predicted values are compared with observations, and by iterative trial and error modification of the model and recalculation of the kinematic variables, observed and calculated versions can be brought into approximate coincidence. The process can be made interactive by use of modern workstations, and model velocity functions that generate travel-time predictions that visually satisfy the data quite well can be established very quickly. A second approach is to solve directly for the parameters of the unknown earth model using travel-time information as input. The basic principle involved is that the ray parameter (the inverse of an arrival's horizontal component of velocity) of any ray characterizes that ray throughout its entire path. This means that the ray parameter measured at the surface of the earth is the inverse of the velocity of the seismic wave in the earth at the ray's turning-point depth. A velocity model can be constructed from turning-point depths of a suite of ray parameters measured from seismic arrivals.

Construction of a $V(z)$ model by travel-time techniques gives information with variable resolution in depth. An instructive way to visualize travel-time resolution limitations is by turning-point density (Fig. 6). Rays traced through the Vera *et al.* (1990) model of the East Pacific Rise axis show many ray turning points within a small depth interval in some regions, with associated rays returning to the surface in a dense pattern. Turning-point density is sparse in other parts of the model. Since estimating $V(z)$ involves calculating turning-point depths, intervals that include

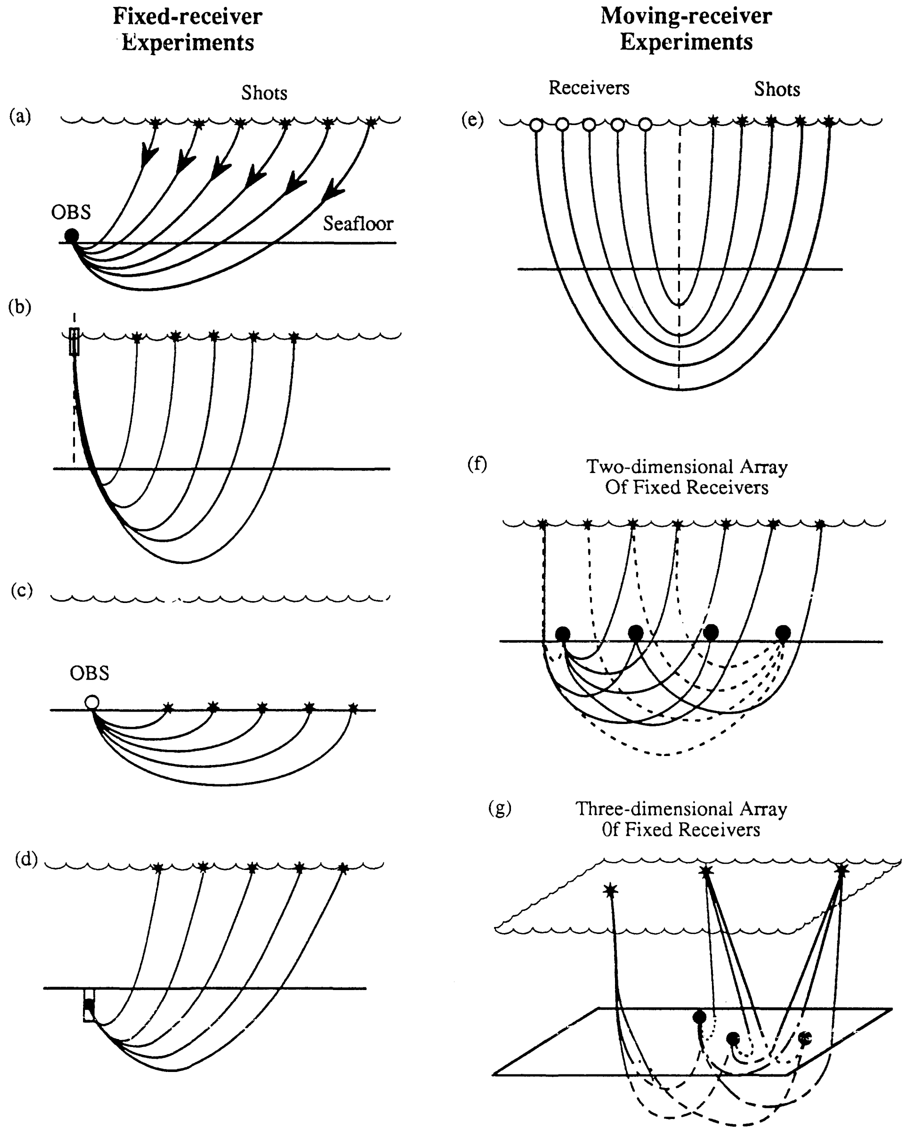
Fixed-receiver Experiments
Moving-receiver Experiments
(a)
Shots
OBS
Seafloor
(b)
(c)
OBS
(d)
(e)
Receivers
Shots
(f)
Two-dimensional Array Of Fixed Receivers
(g)
Three-dimensional Array Of Fixed Receivers

FIG. 5. Marine seismic refraction experiments typically involve the collection of seismic arrivals that have traveled from a source (shown here as the star symbol) to either a single receiver or several receivers. In each case, the seismic wavefield is recorded at numerous distances from the source. The passage of seismic energy through the subsurface can be represented by lines that trace the path of propagation of a seismic "ray"—a convenient descriptive construction that is used to visualize the paths taken by seismic energy perpendicular to the wave fronts of seismic energy. The experimental configurations shown in (a) through (e) can be used to provide one-dimensional estimates of the velocity structure of the crust. Arrays of instruments are used to determine two- and three-dimensional estimates of velocity structure.

many turning rays will be better resolved than those with fewer turning rays. Regions of steep velocity gradient show densely spaced turning rays and act to focus energy to the surface, while low-velocity gradients have few turning rays and defocus energy. Because rays do not turn in low-velocity zones, no information on such an interval is directly available there. Low-velocity zones are an extreme example of the important general rule of thumb regarding model resolution—no arrivals, no resolution. Magma chambers and the adjacent regions of solid rock at elevated temperature are expected to exhibit lower

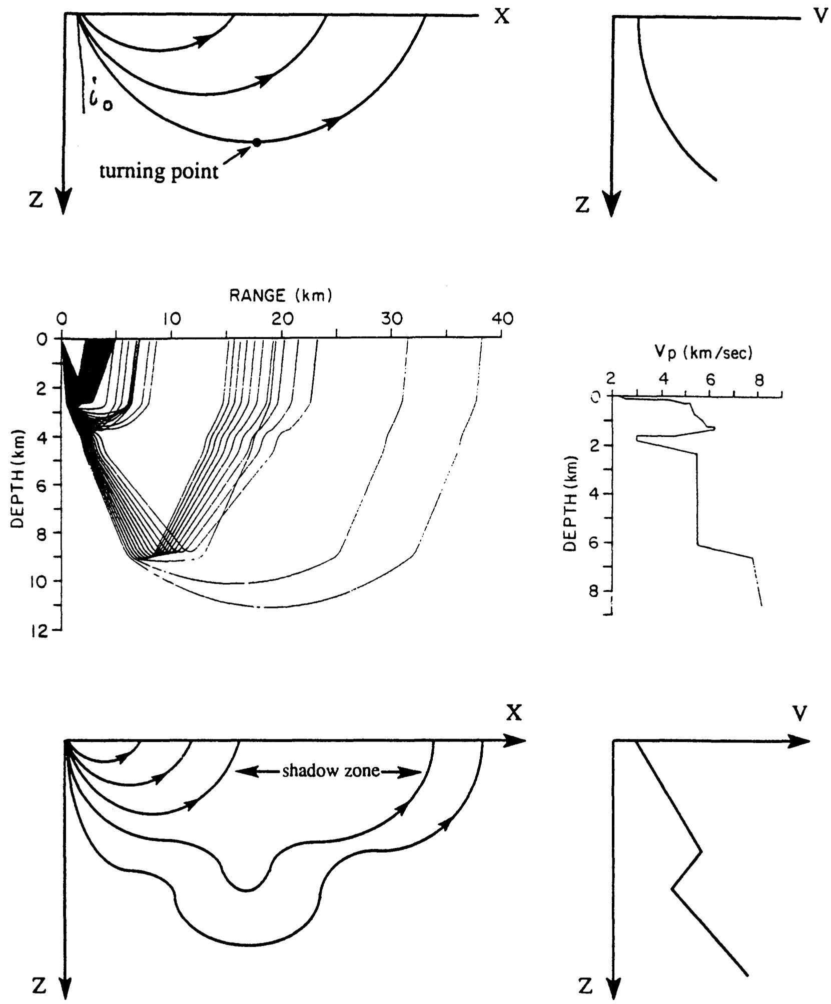

FIG. 6. When seismic energy penetrates into the earth, it is typically bent back toward the surface and travels in a path that is roughly like a semicircle. There is a turning point where the energy begins to travel back up to the surface. These turning-point depths can be estimated by measuring properties of the recorded wavefield and a profile of seismic velocity as a function of depth (middle illustration, right-hand side). If the crust includes a region such as a liquid magma body where velocity decreases, seismic rays are bent down into the crust and do not begin turning back toward the surface until they again enter an area of increasing velocity. These low-velocity zones present considerable interpretational difficulties for seismic data analysis.

velocities than the surrounding rocks. Hence, travel-time inversion alone is incapable of recovering information on one of the most fundamental and important features of mid-ocean ridges. We rely on methods that use different aspects of the seismic wave field, such as described below, to define these features.

Estimates of two- and three-dimensional velocity structure, including low-velocity regions, can be derived from observations using spatially distributed arrays of instruments and sources [Figs. 5(f) and 5(g)] and the application of tomographic techniques of analysis. Sources are fired across a distributed array of seafloor instruments so that many rays pass through a particular crustal volume of interest (Fig. 7, top panel). The analysis does not calculate depths to a particular part of a ray path, but estimates a velocity distribution (Fig. 7, bottom panel) consistent with total travel times from many source–receiver combinations. In a tomographic analysis, an initial, very slowly vary-

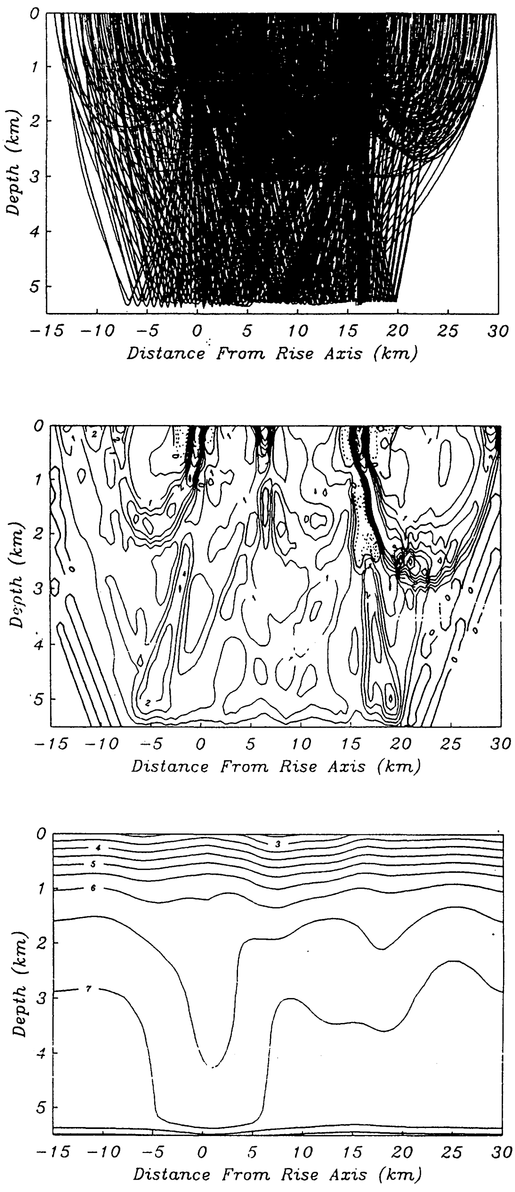

FIG. 7. Tomography uses the travel times of a suite of seismic rays that cross one another (top panel) from recordings of surface shots made by several instruments, typically deployed on the seafloor. The analysis leads to an estimate of the two-dimensional crustal velocity field as in the bottom panel (or three-dimensional if a spatially distributed array of recording instruments is deployed). Note the broadly defined region of lower velocity in the central part of the figure. This corresponds to a region of magma and hot rock associated with the zone of crustal formation beneath the ridge axis. Resolution of the velocity structure depends on the density of crossing rays and is illustrated in the middle panel. Illustrations are from Caress *et al.* (1992).

ing two-dimensional $V(z,x)$ model is specified, rays are traced through the model from known source locations to known receiver locations, and associated travel times are calculated (Fig. 7, top panel). The objective is to modify the velocity model iteratively until observed and calculated travel times come into acceptable agreement. This is essentially the forward modeling approach, but applied by a process that assesses the extent of misfit and modifies the model until the misfit is reduced to an acceptable amount.

Resolution can be calculated from the response of synthetic impulsive velocity perturbations. Spikes in the velocity field are inserted, and a partial derivative matrix associated with the preferred model is used to try to recover the spike, yielding a model δv (Fig. 7, middle panel) in which the information density, as represented by the contour density, is a reasonable proxy for resolution. This can be done over the entire model space to give an estimate of model resolution throughout the space of interest (Caress *et al.*, 1992). What is immediately apparent is a repetition of the result stated earlier—resolution is controlled by ray density. In a tomographic analysis, resolution relies on rays crossing from different source–receiver configurations. There is often a simple trade-off between velocity and dimension—small structures with large velocity perturbations have an almost identical effect to larger bodies with smaller velocity anomalies. However, since tomography uses transmitted rays, information on low-velocity zones *can* be directly recovered.

1.2.2 Crustal Velocity Distribution from Seismic Wave Forms

Ray-based methods used in travel-time analyses embody substantial simplification to propagation phenomena that inherently restricts information gained from travel-time analysis. The seismic signal is treated as being of infinite frequency and wave fronts as discontinuity surfaces. Practical seismic sources emit complex signals with frequency content usually ranging from around 4 to 60 Hz, with peak at 10 to 15 Hz. Amplitude, phase, and frequency content of these signals are modified in their passage through the crust by properties that affect travel times by irresolvably small amounts, or not at all. Wave-form studies can therefore add information not available from seismic travel times or can improve on the resolution available from travel-time methods.

Wave-form information is most commonly employed by iterative forward modeling using the same approach as that employed for forward modeling of travel-time data, except that the "complete" seismic response is calculated. Several methods are available, and a useful basic discussion of them is given by Mooney (1983). The most extensively used class of synthetic seismogram methods is transform methods, in which the elastic wave equation is doubly transformed (space and time) to yield an ordinary differential equation that is solved for displacement potentials in terms of p (horizontal slowness), z (depth), and ω (radial frequency) by standard approaches (Aki and Richards, 1980). A synthetic seismogram is generated by double inverse transforming this solution. Many different synthetic seismograms can be created depending on how the transforms and inverse transforms are taken. The most commonly used are reflectivity (real slowness integration followed by frequency; Muller, 1985) and WKBJ (frequency integration followed by real slowness integration; Chapman, 1985). The most powerful synthetic seismogram computation uses the finite-difference (FD) method (Kelly *et al.*, 1976; Stephen, 1983). In this approach, the wave equation is approximated by a low-order discrete-difference equation and applied to a medium that is subdivided (gridded) into numerous very small elements in which the differencing approximation holds valid. Complex structure can be readily included by assigning different properties to elements of a finely subdivided model. Thus, models can be made two- or three-dimensional, and anisotropy can be modeled. The great penalty of FD methods is the enormous computation time required, because the medium must be gridded on a scale that is very small compared to a seismic wavelength, and the wave equation solved at a correspondingly large number of points.

Once the synthetic response of the model has been computed, a comparison is made to observed data, misfits are noted, and model adjustments are made until a satisfactory fit is obtained. A typical comparison of observed and synthetic seismograms using the reflectivity method is shown in Fig. 8 (Vera *et al.*, 1990).

Wave-form analysis provides a route to deriving estimates of seismic attenuation, a parameter that cannot be obtained from travel-time data. Anelastic attenuation describes the degree to which seismic energy is absorbed, primarily through internal friction that causes dissipation of the energy into tiny amounts of heat. Attenuation factors in common crustal rocks vary by a much greater range than exhibited by either p-wave or s-wave velocities. This is because attenuation is more sensitive than the bulk elastic properties to factors such as porosity (including the shape of pore spaces), presence of fluids, and temperature. Attenuative properties of crustal rocks near their solidus probably range through an order of magnitude, perhaps five times the range of the elastic property changes in the same region. Attenuation is the energy loss per cycle (through anelastic processes) relative to peak energy at a given frequency (Aki and Richards, 1980) and must be estimated from changes in seismic wave forms. One approach—the spectral ratio method (see, for example, Jacobson and Lewis, 1990)—involves comparison of seismic pulse spectra after propagation through the earth with the same pulse before propagation. Wilcock *et al.* (1992) extended this approach by using a spectral method to estimate attenuation of wave forms, which they map into the distribution of crustal attenuation of the East Pacific Rise axis using a velocity model and ray paths derived from tomographic analysis of seismic travel times (Fig. 9).

Like travel-time analysis, wave-form analysis can be treated as an inverse problem. Wave-form inversion involves an iterative search for a model parametrized in terms of p- and s-wave velocities and density that best fits observed travel times and wave forms of observed seismic data. In one approach, a starting model is chosen that adequately describes the long-wavelength or background parameters, perhaps using a travel-time inverse model. The associated synthetic seismogram is calculated and compared to an observed seismogram. A quantitative measure of misfit between observed and calculated wave-form data is made for the entire seismogram, and used to calculate how to perturb the model structure in such a way that a new synthetic seismogram, calculated

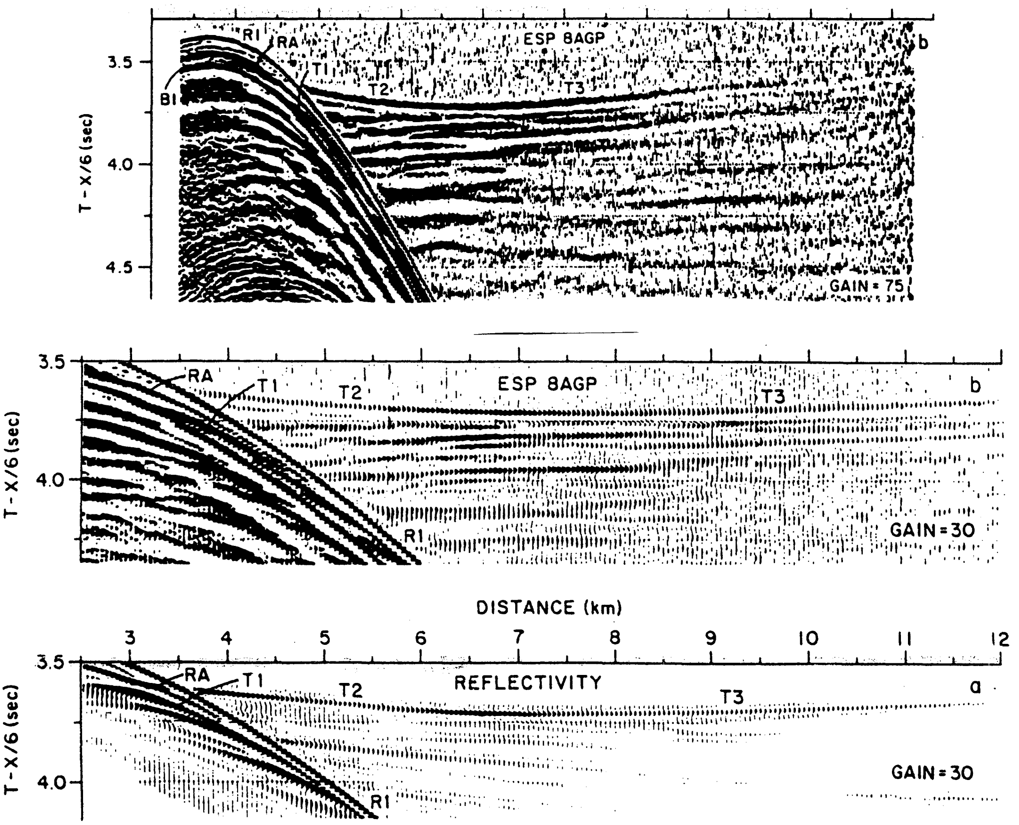

FIG. 8. A typical ESP seismogram recorded on the East Pacific Rise is shown in the top panel, and an analysis using reflectivity synthetic seismogram methods is shown in the lower two panels. The middle panel is a detail of the observed data, and the bottom panel shows the seismogram calculated for the structure shown in Fig. 6 (middle panel, right-hand side). Note that most of the important features of the observed data have been simulated in the calculated synthetic data. Illustrations are from Vera *et al.* (1990).

for the perturbed model, will produce an improved fit to observations. Least-squares criteria for misfit can be used provided the initial model is suitably close to the final model. As in travel-time tomography, it is necessary to calculate a large matrix of partial derivatives of the synthetic seismogram with respect to model parameters. In wave-form inversion, this is a huge task since it involves computation of the full synthetic response. Use of wave-form inversion for study of the oceanic crust has largely been limited to use of the refracted wave field calculated by the WKBJ technique (Carey and Chapman, 1988; Shaw and Orcutt, 1985).

A recent advance in model-space search schemes involves the application of methods borrowed from life sciences. The genetic algorithm (GA) is an approach that works on groups of models rather than a single model and proceeds in a very different way from traditional parameter-estimating schemes. Model parameters are represented in binary form, and the search begins with a random population of models called chromosome by analogy to a biological form. The correlation between synthetic data generated from these models and observations is calculated as a measure of the "fitness" of each model. Then the processes of selection, crossover, and mutation are performed on the models in sequence. Selection involves copying models in proportion to their so-called fitness. The selected pairs of chromosomes (models) then exchange bit strings at some crossover site within the model string. The population then experiences mutation whereby a bit is randomly swapped in a chromosome's bit string. Following these three steps, the models are compared to the old generation of models and

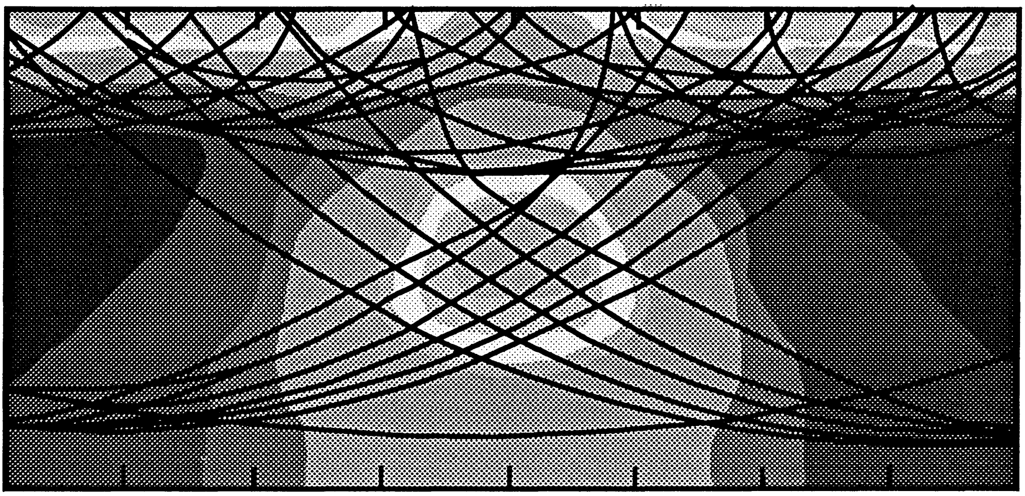

FIG. 9. Seismic waveforms can be used to derive estimates of the attenuation properties of the crust. This illustration from Wilcock *et al.* (1992) shows how the region of magma and hot rock can be identified using the extent to which rocks attenuate seismic energy. The most lightly shaded areas are the most strongly attenuating. These data, together with the results of tomographic studies that give the velocity distribution beneath the ridge, are clarifying the structure beneath fast-spreading centers and providing insight into the process of seafloor spreading.

accepted or rejected based on the same fitness measure. The procedure repeats until all models are almost equally fit.

1.2.3 Seismic Imaging Using the Scattered Wave Field The basic principle in this approach is that seismic energy propagating at near-vertical take-off angles will, on encountering changes in elastic properties on a spatial scale that is small relative to the seismic pulse length, be scattered by them and returned to the surface. This scattered response can be used to construct a visualization of short-wavelength structure often called a reflection image. Energy sources and hydrophone arrays (streamers) for reflection experiments are the same as those used to create refraction seismograms, but both are towed from the same vessel. The source is fired at regular intervals corresponding to fixed along-track increments, usually less than 50 m. Hydrophone arrays of up to 6 km in length, containing as many as 1000 individual hydrophones, are used, together with source arrays of 60 or more airguns firing total volumes of more than 10 000 in.^{3}. The experiment produces an essentially continuous insonification of the subsurface because the receiving array moves only a small fraction of its length between shots.

Each shot fired during an experiment creates a seismogram with a fixed source geometry and a suite of regularly spaced traces, one for each hydrophone group in the streamer—usually 60 to 240 traces. To produce a reflection image from field recordings, traces are regrouped ("gathered"), corrected for varying source–receiver distance, and summed ("stacked"). Gathering is done so that summed arrivals have come from approximately the same reflection points. Each regrouped seismogram is a small common-midpoint experiment in which maximum range is the streamer length. Time corrections to gathered seismograms are made to align traces, which are then summed to enhance signal-to-noise characteristics. The correction attempts to bring arrivals observed at all ranges to the travel time they would have were they recorded at an offset of zero—the correction appropriate for situations where the subseafloor structure is nearly horizontal is called a normal move-out (NMO) correction, and a different type of correction, known as dip move-out (DMO), is applied if structure is complex or steeply dipping. In both cases, the correction is equivalent to a transformation to zero offset. In theory, summation of traces after trace alignment produces signal-to-noise enhancement by a factor equal to the square root of the number of traces summed.

Reflection seismograms obviously have an appearance that matches our expectation of what a profile of the Earth's interior might look like, and really qualify to be described as an *image* (Plate 2). It is not a direct representation of subsurface structure, but rather its acoustic response. An image is produced by reversing propagation effects by downward wave-field projection or extrapolation ("migration"). Several wave-field extrapolation methods have been developed, all of which involve several stages of data manipulation, and all are very demanding of computer time (Yilmaz, 1987). The velocity field must be known very well (and may not be). Most extrapolation methods involve some form of approximation to the full wave equation to make computations tractable. Most, for instance, allow only mild lateral variations. Extrapolation methods are similar to synthetic seismogram methods. Kirchoff integration methods operate in the time domain and are the simplest methods. A popular transform method operates on data in the frequency-wavenumber (FK) domain but is strictly applicable only when no lateral velocity variations are present. Finite-difference migrations can account for mild lateral variations (Claerbout, 1985). To perform extrapolation in a completely correct way requires use of the full wave equation, must operate on pre-stack data, and must account for refraction of rays across boundaries. The FK and FD extrapolates are the most commonly used in studies of oceanic crust.

1.3 Potential Fields

1.3.1 Studies of the Gravity Field Variations in the strength of the earth's gravitational field provide information on geologic structures associated with inhomogeneities in the distribution of mass within the Earth. The magnitude of the gravitational acceleration at the Earth's surface is about 980 cm/s^2, but varies both because of the shape of the Earth (a measurement taken at the equator is further from the Earth's center of mass than is a measurement taken at the pole) and because of local inhomogeneities in the Earth's density structure. The standard unit of gravitational measurements is named the Gal in honor of Galileo and has a value of 1 cm/s^2. However, a more useful unit for geophysical applications is the milliGal (mGal). Gravity at the earth's surface is then slightly less than 10^6 mGal. Gravity anomalies resulting from geological structures are generally less than about 50 mGal but can be as large as a few hundred mGal.

Gravity measurements made during geophysical surveys, whether on land or at sea, are almost invariably relative, rather than absolute. Determination of the difference in gravity between several locations is much simpler, faster, less expensive, and, until very recently, more accurate than actually determining the value of gravity at each location. A survey therefore needs to be "tied in" by including measurements at locations where the absolute value of the gravity field is known. For marine surveys, this is done in port prior to and following the survey.

Gravity surveys at sea are made difficult because the vessel is subject to large accelerations caused by wave motion. A moderate-sized ship, such as commonly used in marine surveys, can be subjected to accelerations of up to 100 000 mGal, or 10% of the earth's gravity. Since gravimeters cannot distinguish accelerations due to gravity from those due to the ship's motion, it is necessary to suppress or remove these environmental accelerations to isolate the much smaller variations in gravity resulting from geologic structure. This is commonly done by passing the gravity measurements through a low-pass filter. Gravity variations resulting from geological structures are commonly a kilometer or more across. Since a ship traveling at 10 knots (a common cruising speed during surveys) travels about 300 m in a minute, any gravity anomaly resulting from geological structures will have a period of several minutes. On the other hand, ship motions typically have a characteristic period of the order of 10 to 30 seconds, and average to zero (since the ship remains on the sea surface). Thus, by passing the gravity signal through a low-pass filter that suppresses signals at periods shorter than 2 to 3 minutes, vertical ship motions can be effectively removed. Marine gravity measurements can be obtained with an accuracy of about 1 mGal, while gravity measurements on land can be made with an accuracy of about 0.01 mGal.

There are two basic type of meters presently in use, beam gravimeters and forced-feedback accelerometers. A beam gravimeter

[Fig. 10(a)] consists of a beam, one end of which is free to rotate in a vertical plane. The force of gravity on the free end of the beam is balanced by a zero-length spring, i.e., one for which the relationship between elongation and applied force acts as if the spring had zero length when unloaded. For such a spring, a small change in gravity produces a large deflection of the beam (Harrison, 1960; Worzel and Harrison, 1963). The sensor for a forced-feedback accelerometer gravimeter [Figure 10(b)] is an inertial-navigation grade accelerometer. The sensor consists of a proof mass wrapped in a coil and constrained to move along the vertical axis between two permanent magnets. The mass is maintained at its null position by a balance between gravity and the electromagnetic force induced by a current through the coil. Changes in the position of the coil are detected by a feedback loop, which regulates the current in the coil to drive the proof mass back to its null position. The current through the coil, averaged over 1-s intervals, is monitored to determine changes in the vertical acceleration. Both types of gravimeter record accelerations along their vertical axis. They must therefore be kept level to detect changes in the Earth's gravity field correctly, and this is accomplished by mounting the gravimeter on a gyrostabilized platform.

To make gravity measurements suitable for interpretation, effects not related to geological structure have to be removed. Gravity changes with latitude because the Earth is not a perfect sphere and because it is rotating. The latter effect enters because the net force experienced by a body at the surface of the Earth is the sum of a downward "gravitational" force due to the earth's mass and an upward "centrifugal" force due to the rotation of the earth. The centrifugal acceleration is dependent on latitude because a body's distance from the Earth's spin axis and therefore its velocity vary with latitude. The motion of the ship also alters the centrifugal acceleration. This is referred to as the Eötvös effect (after Hungarian geophysicist Baron Roland von Eötvös). It can be as large as 90 mGal for normal ship speeds, much larger than the magnitude of most geological anomalies in deep-water areas.

Gravity variations are defined in terms of an anomaly, being the difference between the measured and the expected value of gravity at a location. The difference between a gravity measurement at sea level and the expected gravity at that latitude, after making

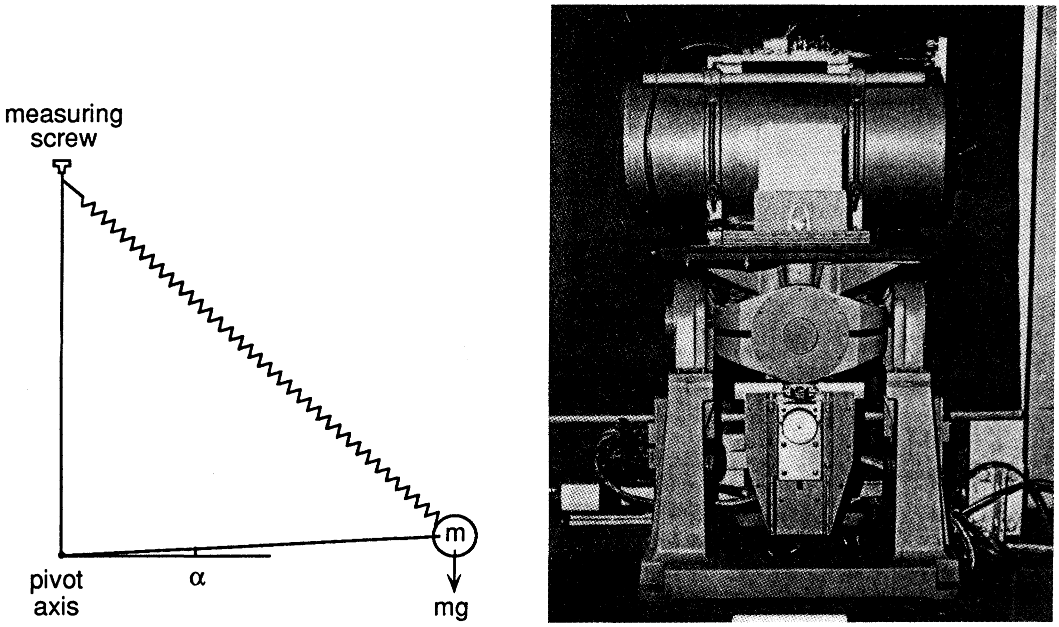

(a)

FIG. 10. (a) Basic internal components of the two major types of gravimeters. (a) The spring and balanced beam of a beam gravimeter.

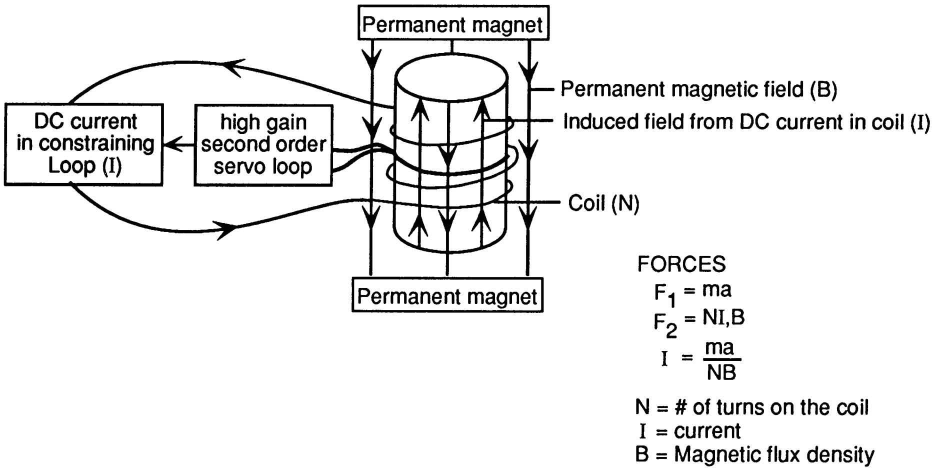
Permanent magnet
DC current in constraining Loop (I)
high gain second order servo loop
Permanent magnetic field (B)
Induced field from DC current in coil (I)
Coil (N)
Permanent magnet
FORCES
F1 = ma
F2 = NI,B
I = ma/NB
N = # of turns on the coil
I = current
B = Magnetic flux density

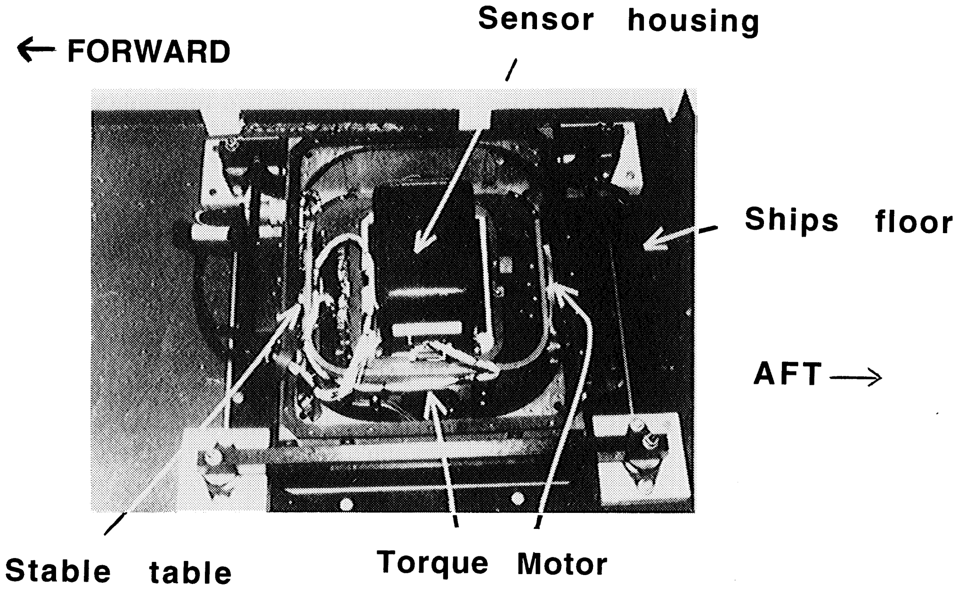
← FORWARD
Sensor housing
Ships floor
AFT →
Stable table
Torque Motor

(b)

FIG. 10. (b) The components of a forced-feedback accelerometer.

an Eötvös correction, is referred to as a "free-air anomaly". This is the most commonly used anomaly for marine studies. Gridded bathymetric data sets from swath mapping surveys have also made possible the use of a "mantle Bouguer anomaly" (Prince and Forsyth, 1988). Calculation of this anomaly involves correction for the gravity effect of bathymetric relief (water-crust interface) plus the effect of the crust-mantle interface achieved by assumption of a constant-thickness crust (usually 6 km). Examples of use of these types of anomalies are discussed below (see Plate 4).

Marine gravity surveys are generally carried out with lines run perpendicular to the expected trend of the feature being surveyed. Until recently, interpretation has usually assumed that the structure being investigated is two-dimensional. The gravity anomalies resulting from two-dimensional model bodies can be easily calculated (Bott, 1960; Talwani *et al.*, 1959) and interpretations made based on forward modeling procedures much like those described for seismic methods. The application of multibeam swath mapping systems since the early 1980s has resulted in surveys which include complete (or nearly complete) bathymetric coverage of sizable areas. This has allowed three-dimensional variations in structure to be explicitly included in gravity modeling. These are usually calculated with Fourier transform techniques (Parker, 1973).

1.3.2 Magnetic Field Studies The Earth's external magnetic field is very near to that expected from a bar magnet with its poles roughly corresponding to the earth's rotational poles. This was recognized in the sixteenth century by Gilbert, who wrote in his treatise *De Magnete* that "*magnus magnes ipse est globas terrestis*" (the Earthly globe itself is a great magnet). The approximate coincidence of rotational and magnetic poles has been used for navigation for hundreds of years, since it implies that a magnetized compass needle will point toward the pole. This dipolar nature of the field means that its direction relative to the earth's surface varies with position on the earth. The field is oriented in a vertical direction at the poles (down at the north pole, up at the south pole) and is horizontal at the equator. The field is specified by the "inclination" or "dip" (the angle with the horizontal), the "declination" (the angle between the horizontal component and true north), and magnitude of the total field. The SI (mks) unit for magnetic intensity is the tesla, which is defined as 1 Wb/m^2 (or as 10^4 G to use a more familiar unit). A more useful unit for discussing the earth's magnetic field is the nanotesla (nT), often called the gamma (γ). The earth's magnetic field varies from about 30 000 to 60 000 nT. Magnetic anomalies resulting from geological structures are commonly a few hundred nT, but may be as large as several thousand nT.

The magnetic field arises from electrical currents within the core. The interaction of these currents with fluid motions in the liquid outer core (caused by thermal convection) and the rotation of the Earth can result in a feedback mechanism or dynamo, so that the Earth's magnetic field is self-sustaining (Backus, 1958; Bullard and Gellman, 1954; Elsasser, 1956; Herzenberg, 1958; Parker, 1955). Magnetohydrodynamic models suggest that the earth's field consists of a very strong (perhaps 100 to 500 Oe) internal toroidal field confined to the core and a much weaker external poloidal field, which results from distortions in the toroidal field and is the magnetic field observed at the surface.

As in the case of gravity-field studies, magnetic-field values are treated as anomalies with respect to a reference field. For magnetic studies, an empirical reference field is determined by a spherical-harmonic analysis of magnetic measurements at numerous fixed locations around the world. These International Geomagnetic Reference Fields are determined at a particular time and are referred to by the epoch for which they are valid, for example IGRF 1985. Since the secular variation is reasonably smooth, predictive models are used to interpolate between the times represented by them and to extrapolate from the most recent IGRF to the present. In addition, the anomalous field must take account of secular changes due to two forms of variations in the external field phenomenon. One is the diurnal or daily variation associated with the pattern of ionospheric currents following the sun. These are generally smoothly varying and reasonably small (30 to 50 nT). The second type of variation is the much higher-amplitude, shorter-period, irregular disturbances known as magnetic storms that frequently occur at times of vigorous sunspot activity and are accompanied by spectacular auroras and disrupted radio communications. It is frequently impossible to collect useful magnetic data during magnetic storms.

As was the case with gravity studies, the advent of swath map surveying in the 1980s has led to the collection of densely sampled grids of magnetic data. With the assumption that the magnetization is in a single direction (or 180° from that direction), these grids of data can be inverted to give the distribution of magnetization of the crustal rocks (Macdonald *et al.*, 1980; Parker and Huestis, 1974).

For most marine studies, the magnetic-field strength only is measured, and the instrument used is a proton precession magnetometer, which uses nuclear magnetic resonance to determine field strength. Protons making up the nuclei of hydrogen atoms in water possess a magnetic moment that causes them to try to align along any external magnetic field. Since the Earth's field is fairly weak, this tendency is more than overcome by thermal agitation, which constantly disorients the protons. The magnetometer basically consists of a bottle of water with a coil around it. A large current is passed through the coil to create a strong magnetic field, several hundred times as large as the Earth's field, and this aligns proton spin axes in the same direction. When the current is stopped, the applied field is removed, and the protons sense only the Earth's field. They will try to reorient themselves along the Earth's field, and, for a short period of time, until thermal agitation disorients them, the protons will coherently precess around the Earth's field. The precession frequency is directly proportional to the external magnetic field strength and can be measured by the induced current. Proton precession magnetometers therefore determine the intensity of the total magnetic field.

Magnetic anomalies arise because the Earth's field can *induce* a magnetic moment in magnetic materials by causing the spins of free electrons to align along the magnetic-field direction. In addition, many rocks contain ferromagnetic minerals that can acquire a *permanent* or *remnant* magnetization when they originally form. The rock magnetization, whether induced or permanent, produces a magnetic field. The anomaly associated with local structure results from the interaction of the Earth's field with the field generated by the magnetized body. Hence, even when a body is magnetized along the direction of the Earth's field, the associated anomaly will vary with location on the Earth. The total intensity magnetic anomaly over a spherical body magnetized in the direction of the Earth's field will be positive and symmetric at the magnetic poles, while at the equator, the anomaly will be symmetric and of the same magnitude but have negative amplitude (Fig. 11). At intermediate latitudes, the anomaly will be asymmetric or skewed. For the same reasons, the anomaly depends on the body's orientation. A north-south profile perpendicular to and across the center of a linear feature that is oriented in an east-west direction near the equator will show a large negative anomaly. If the body is rotated 90°, so that it is oriented in a north-south direction, an east-west profile perpendicular to and across its center will show no magnetic anomaly.

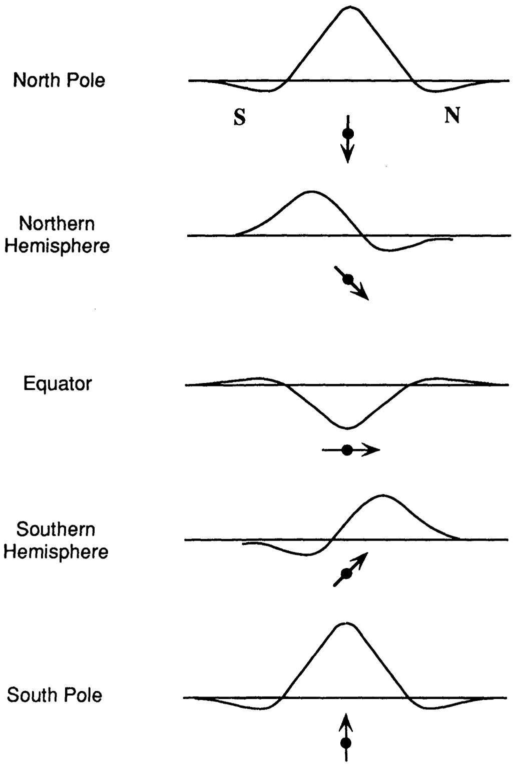

FIG. 11. The magnetic response of a buried object with magnetic properties varies according to its position on the Earth because the anomaly produced results from the interaction of the present field with the magnetic field of the body that may have been acquired at a very different time when the Earth's field was oriented differently from today's field. Even if the body is magnetized in the direction of the present field, the anomaly produced varies with latitude.

2. MARINE GEOPHYSICAL METHODS AND THE STUDY OF MIDOCEAN RIDGES

The midocean-ridge systems of the world's oceans are the locus of creation of new seafloor through a complex system of processes that involves melting of underlying mantle to produce magma, and transport of magma through a low-porosity mantle to the surface, where it freezes to form oceanic crust—a layer of igneous rock about 6 km thick that floors virtually all the world's ocean basins. The rate of crustal production, or spreading rate, is measured from the pattern of magnetic anomalies. This use of marine magnetic anomalies to establish the age of the seafloor and hence the rate at which it was created is axiomatic to the modern era of understanding of the ocean basins and their mechanism of formation. Magnetic anomalies can be used to establish the age of the seafloor because the Earth's magnetic field periodically reverses its polarity. Igneous rocks acquire a permanent remanent magnetization in the same direction as the Earth's field at the time they cooled. Seafloor formed when the magnetic field is reversed will acquire a magnetization that is 180° different from that acquired when the field is in its present, "normal" direction. Reversals during the past 45 Ma have occurred irregularly, averaging roughly 200 000-yr intervals. The irregularity of the reversals creates a distinctive pattern to a sequence of anomalies (Fig. 12) that can be correlated between areas of the same age and, once the dates of the reversals are known, can be used to assign ages to the seafloor (Heirtzler *et al.*, 1968; Cande and Kent, 1992). Three-dimensional magnetic surveys have been carried out in a number of settings. Macdonald *et al.* (1980) and Sempere *et al.* (1987) studied the transition from normally to reversely magnetized crust. They found that the transition is sharp, linear, and quite narrow. This implies that reversal of the field takes place over a relatively short period of time, and that the zone in which new crust is created at intermediate and fast-spreading ridges is narrow and spatially stable.

Oceanic crustal thickness is only 20% of that of the continents. This was not recognized until the 1950s, when the first seismic experiments were carried out in the deep oceans. Furthermore, these early seismic experiments found that oceanic crustal thickness was almost invariant throughout the world's oceans. While it has more recently been shown that thickness varies more than the resolution of earlier experiments allowed us to appreciate, the variation is still small and is not well understood. One of the thrusts of marine geophysical research at present involves attempts to understand how oceanic crust is formed and the controls on crustal generation processes. It has recently become apparent, for instance, that the rate at which crust is produced by seafloor spreading at ridges profoundly affects *some* aspects of the crust-generating process but has little effect on others. We will use this research area as a vehicle for describing one way in which the marine geophysical measurements described above are used as a tool for research.

2.1 The Shifting Paradigm I: Controls on the Form of Midocean Ridges

The existence of an extensive ridge system in the world's ocean basins was first recognized in the early part of the century when ships engaged in stringing telephone cables across the Atlantic found systematic evidence for significant shallowing in the middle regions of the ocean. By the 1970s, through compilation of many widely spaced crossings of ocean basins made by ships carrying single-beam bathymetric systems, it had been discovered that ridges like the Mid-Atlantic Ridge were present in essentially all the oceans, though not always located as centrally as the Mid-Atlantic Ridge. With the parallel recognition that ridges were the site of lithospheric generation, there emerged a conceptual picture of the ridge system as a quasilinear zone of generation of new lithosphere, save for a few places where very large offsets, fracture zones, were present.

The modern era of exploration of the ridges can be thought of as beginning when multibeam swath mapping systems were first deployed to provide systematic maps of the axial regions. Aspects of the ridges' shapes that had been hinted at from the earlier data became clearer, and their importance became profoundly understood. One of the most revealing findings was that the axial region of ridges spreading at fast rates (around 10 cm/yr) displays a distinct topographic high, while those spreading at rates around 2 cm/yr or less exhibit a distinct topographic depression (Plate 3). The nature and cause of this great difference appear to relate to the changing balance of forces that combine at the midocean ridge. Another finding is that the ridges are far from quasilinear features, but instead are finely divided into *spreading segments*—individual cells of the ridge that are effectively isolated from their neighbors. The offsets associated with fracture zones were recognized to be part of a multimember family of offsets that exhibit remarkably different form in different ocean basins.

The creation of new lithosphere is funda-

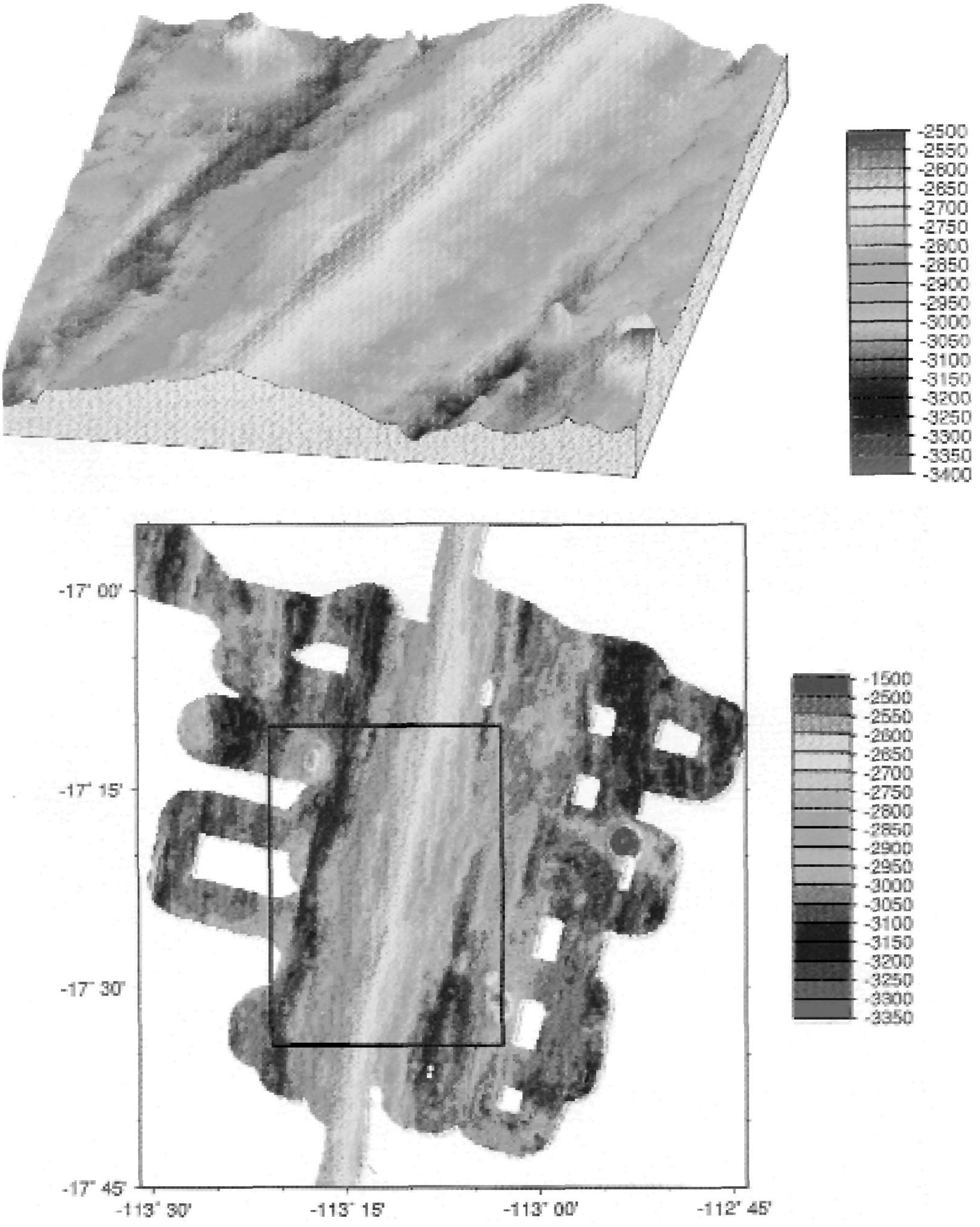
-2500
-2550
-2600
-2650
-2700
-2750
-2800
-2850
-2900
-2950
-3000
-3050
-3100
-3150
-3200
-3250
-3300
-3350
-3400
-17° 00'
-17° 15'
-17° 30'
-17° 45'
-113° 30'
-113° 15'
-113° 00'
-112° 45'
-1500
-2500
-2550
-2600
-2650
-2700
-2750
-2800
-2850
-2900
-2950
-3000
-3050
-3100
-3150
-3200
-3250
-3300
-3350

Vol. 9 – PLATE 1. Typical renderings of the shape of the seafloor include three-dimensional representations of the midocean ridge such as the one shown in the upper panel, a region at the East Pacific Rise. The ridge is the elevated area in the center of the diagram. A seamount is seen on the far left of the map. Note that the general trend of the seafloor topography forms a typical linear pattern approximately parallel to the trend of the ridge axis. This map was produced using the Hydrosweep swath mapping system of Atlas Electronic. The lower panel is a plan view contour representation. The area of its three-dimensional figure is shown in the rectangle.

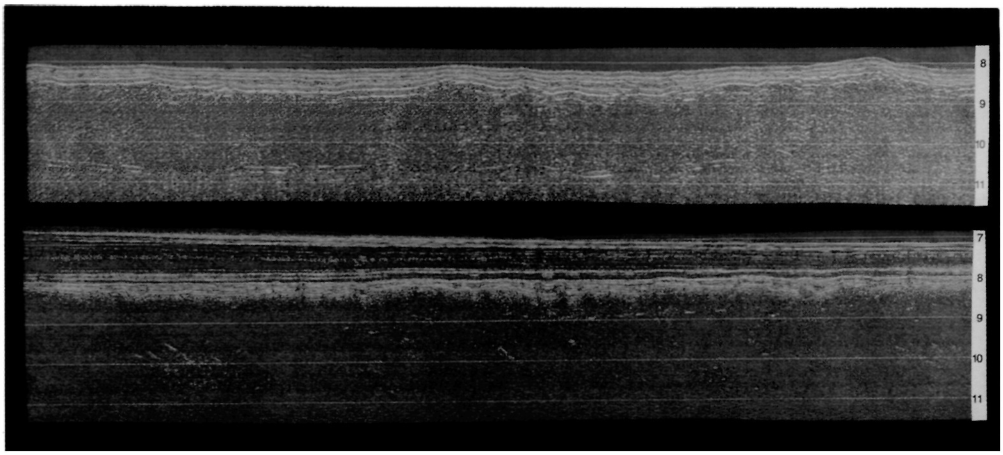

Vol. 9 – PLATE 2. Seismic images of crust produced by seafloor spreading at ridges creating crust at very different spreading rates. In the upper panel, crust produced at the fast-spreading East Pacific Rise shows almost no internal structure but has a distinct base that is the Moho. By contrast, the slowly-spreading Mid-Atlantic Ridge creates crust that is densely structured with dipping features within, but has a poorly resolved base. These differences relate to the differing balance between magmatic production and tectonic stresses in the two settings. These differences are also apparent in the topography of the ridges (Plate 3).

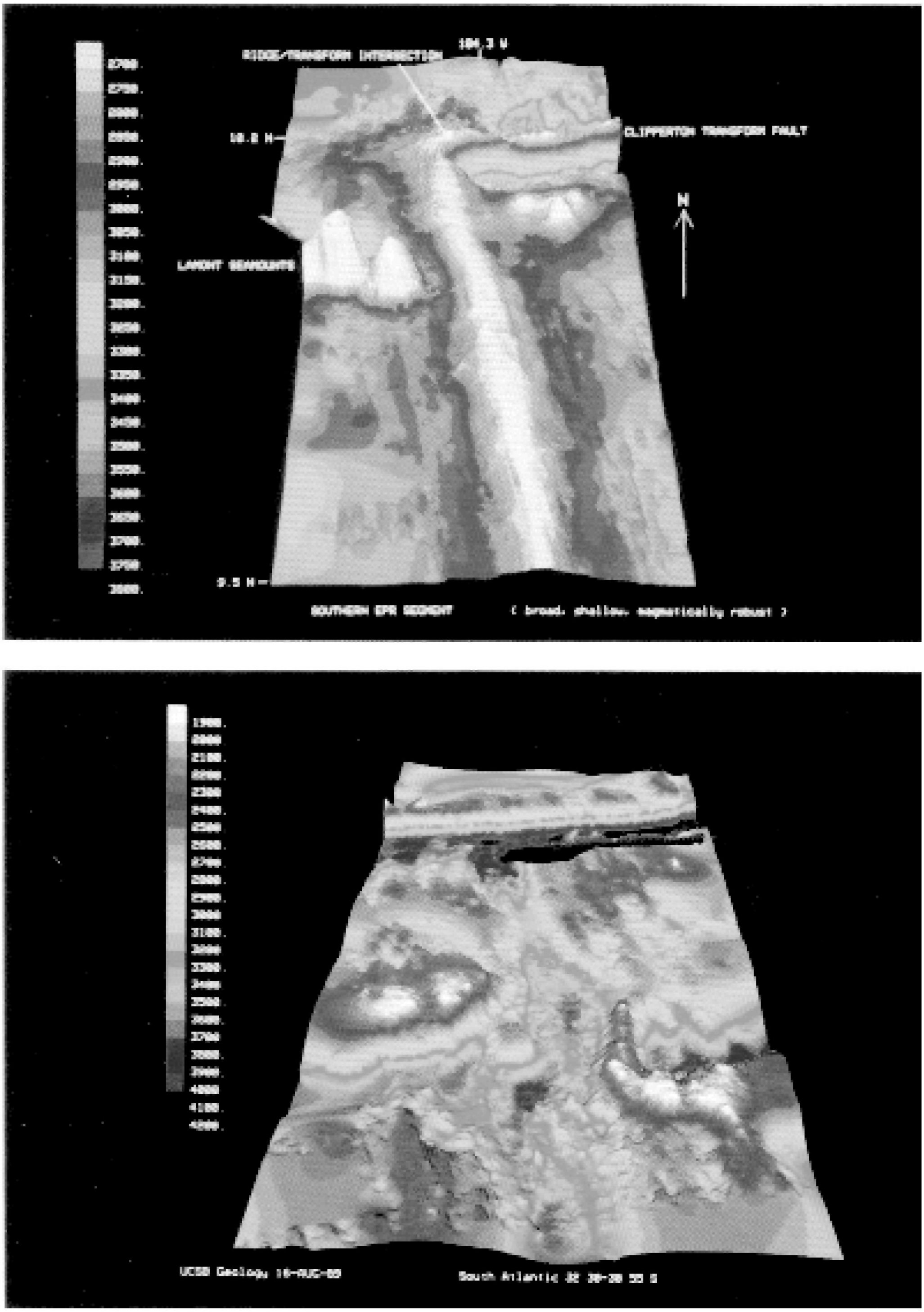
2700.
2750.
2800.
2850.
2900.
2950.
3000.
3050.
3100.
3150.
3200.
3250.
3300.
3350.
3400.
3450.
3500.
3550.
3600.
3650.
3700.
3750.
3800.
RIDGE/TRANSFORM INTERSECTION
104.3 W
10.2 N
CLIPPERTON TRANSFORM FAULT
N
LAMONT SEAMOUNTS
9.5 N
SOUTHERN EPR SEGMENT
(broad, shallow, magmatically robust)
1900.
2000
2100.
2200.
2300
2400.
2500
2600.
2700
2800.
2900.
3000
3100.
3200
3300.
3400
3500.
3600.
3700
3800.
3900
4000
4100.
4200.
UCSD Geology 10-AUG-89
South Atlantic 32 30-28 55 S

Vol. 9 – PLATE 3. Multibeam bathymetric maps rendered in the three dimensions display the remarkable differences between the axial regions of fast-spreading (upper panel) and slowly-spreading (lower panel) ridges (illustrations are from Macdonald et al., 1988, and Macdonald and Fox, 1990). In each illustration, the ridge intersects a major transform. Note that the axial region is characterized by a depression for slowly spreading ridges and a high for fast-spreading ridges.

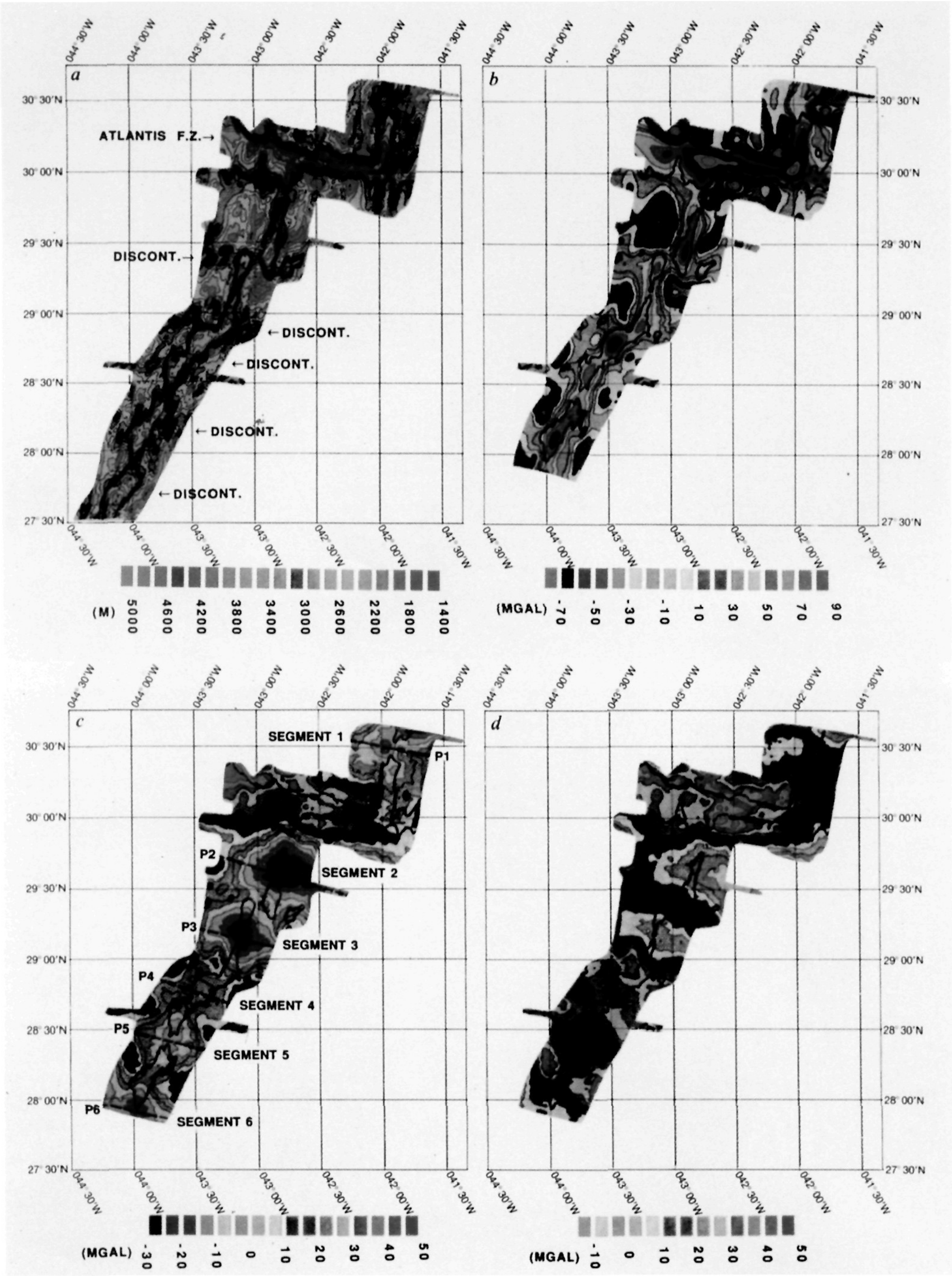
a
ATLANTIS F.Z.→
DISCONT.→
←DISCONT.
←DISCONT.
←DISCONT.
←DISCONT.
(M)
5000
4600
4200
3800
3400
3000
2600
2200
1800
1400
b
(MGAL)
-70
-50
-30
-10
10
30
50
70
90
c
SEGMENT 1
SEGMENT 2
SEGMENT 3
SEGMENT 4
SEGMENT 5
SEGMENT 6
P1
P2
P3
P4
P5
P6
(MGAL)
-30
-20
-10
0
10
20
30
40
50
d
(MGAL)
-10
0
10
20
30
40
50
30° 30'N
30° 00'N
29° 30'N
29° 00'N
28° 30'N
28° 00'N
27° 30'N
044° 30'W
044° 00'W
043° 30'W
043° 00'W
042° 30'W
042° 00'W
041° 30'W

Vol. 9 – PLATE 4. The gravity response of a slowly-spreading ridge such as the Mid-Atlantic Ridge (upper left) can be expressed as a free air anomaly (upper right), but when expressed as a mantle Bouguer anomaly (see text for description), a remarkable circular anomaly pattern is revealed (lower left). These are believed to be the expression of zones of upwelling mantle containing a high percentage of magma. When the large-scale thermal structure of the ridge is subtracted, the circular anomalies become elongate perpendicular to the ridge, perhaps marking zones of a thick and a thin crust produce near to, and away from, upwelling centers. Illustrations from Lin *et al.* (1990).

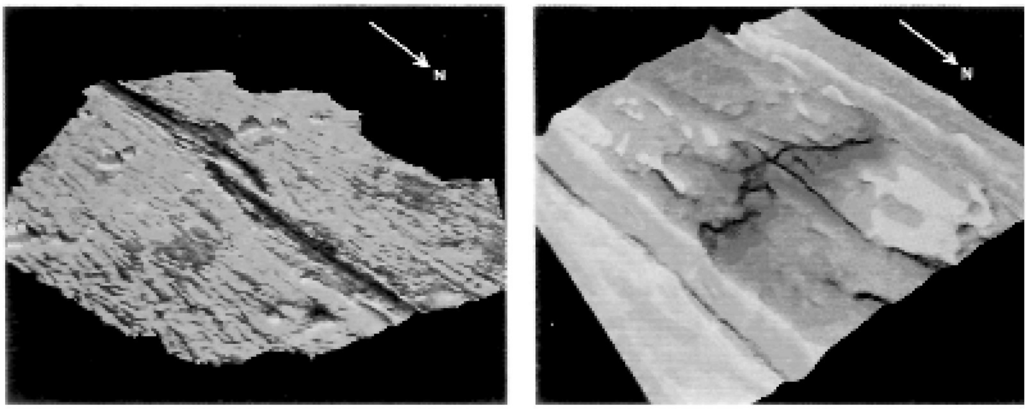

Vol. 9 – PLATE 5. A typical feature of fast-spreading ridges is the overlapping spreading center (OSC). This is a region where two portions of the ridge run in parallel for a short distance (left-hand image). These features migrate along the ridge axis, as one of the parallel limbs grows at the expense of the other. Strong magnetic anomalies associated with crustal formation at the dying ridge tip form a roughly "V"-shaped pattern that tracks the past motion of the OSC. Illustration from Macdonald and Fox (1990).

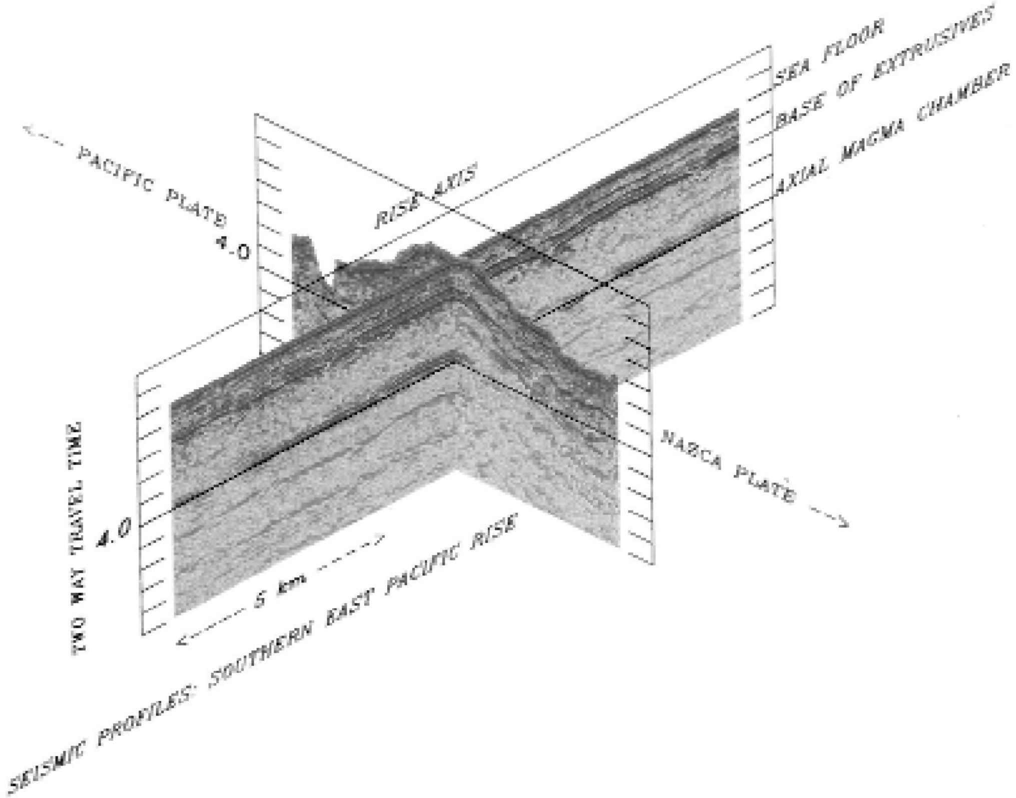
SEA FLOOR
BASE OF EXTRUSIVES
AXIAL MAGMA CHAMBER
PACIFIC PLATE
RISE AXIS
4.0
TWO WAY TRAVEL TIME
4.0
NAZCA PLATE
5 km
SEISMIC PROFILES: SOUTHERN EAST PACIFIC RISE

Vol. 9 – PLATE 6. Perspective rendering of two seismic images of the axial region of the East Pacific Rise (Detrick *et al.*, 1993). The longer line runs along the very summit of the ridge crest. One feature of interest in these data is the axial magma chamber (AMC) reflection that appears as a tiny bright event beneath the ridge in Plate 5. In the profile obtained along the summit of the ridge, we see that the AMC, although small in cross section, displays a surprising continuity along the length of the ridge.

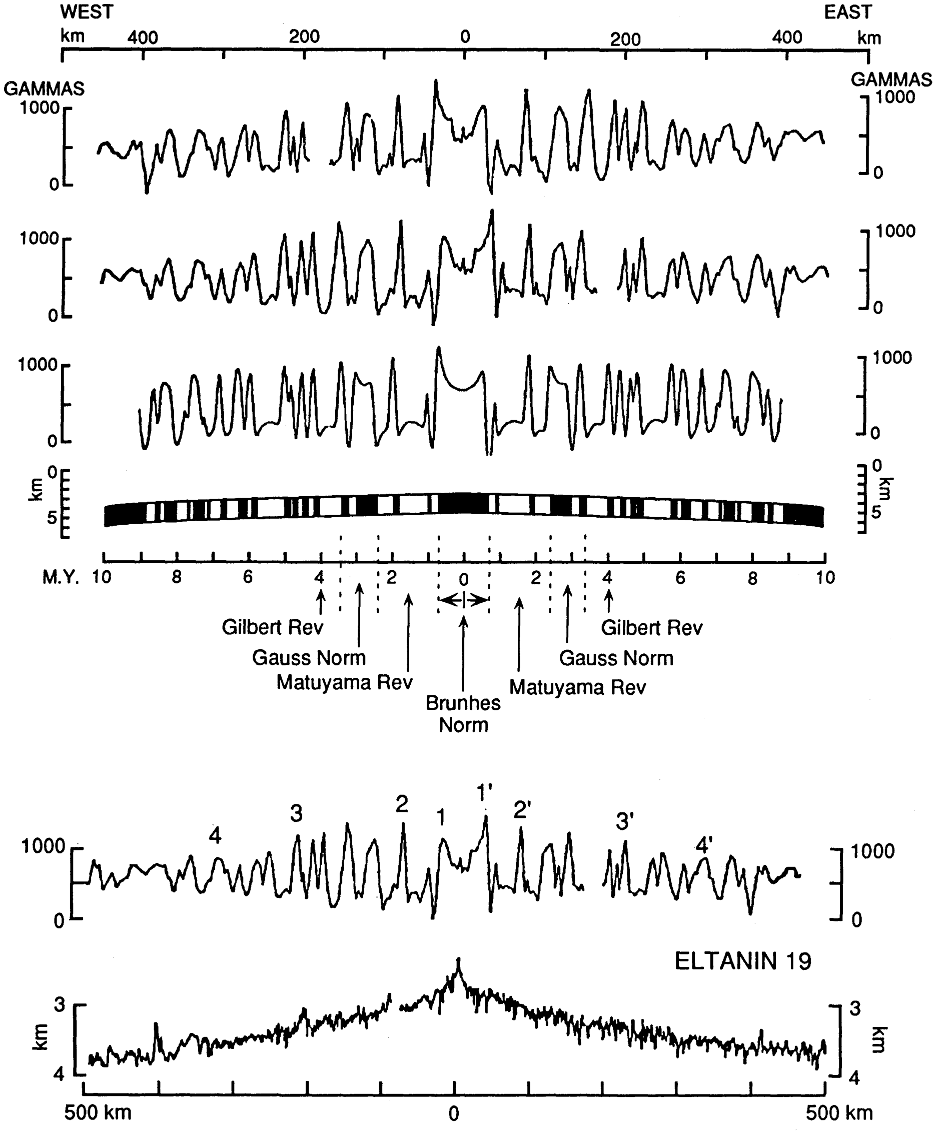

FIG. 12. The pattern of magnetic anomalies recorded over a midocean ridge shows a symmetric pattern that is the response of rocks magnetized in alternating directions (black and white regions of the magnetization model beneath), as the Earth's magnetic field has reverses in polarity with time. These anomalies can be dated and used to establish the history of seafloor spreading.

mentally a magmatically driven process. As plates separate, the underlying mantle rises upward, and a small fraction melts by decompression. The melt separates from the matrix of the still solid mantle and moves toward the surface at a faster rate than the mantle itself. The melt is not completely free to move because the mantle does not contain free pathways, and the partly trapped melt causes buoyant forces to be generated that assist the mantle in its upward motion. This buoyancy-driven flow is critically important because its form and strength depend on the temperature field of the spreading region, which in turn depends on spreading rate—the environment of a slowly spreading ridge is cooler on average than that of a fast-spreading ridge. Buoyancy-driven flow is expected to be stronger and more three-dimensional at slowly spreading ridges than at fast-spreading ridges. The role of mantle flow may therefore be critical to understanding the differences between fast- and slowly spreading ridges.

Another important difference is that slowly spreading ridges are associated with significant earthquake activity, while fast-spreading ridges have only very minor activity. This means that mechanical stresses associated with creation of new lithosphere at slow rates are greater than those associated with spreading at fast rates. Why should that be? Again the reason can be related to the temperature-field differences of the two systems. Fast-spreading centers are hot, and hot material, subjected to stresses, yields plastically without breaking. Slowly spreading ridges are relatively cold, and the rocks are brittle by comparison with the softer, more pliable hot rocks of fast-spreading regions, and, when subject to stress, they will often break. When rocks break, they create faults in the crust, and the rock-breaking process gives rise to earthquakes. It is believed that this accounts for much of the difference in axial topography of the ridges.

The axial depression of slowly spreading centers is created as the response of the crust to mechanical stresses that are trying to pull the crust apart. This stretching causes faulting as the crust extends laterally. The depression is a rift valley, much like those that underlie the lakes in East Africa where the African continent is also subject to extension. The great difference here is that the crust at a spreading center is being created and deformed at essentially the same time. Most investigators believe that the process is episodic—crust is created by injection of magma at irregular and largely unknown intervals. As the new magma body cools, it becomes increasingly rigid and able to sustain mechanical stresses. During cooling, the axial region is continuously subject to extensional stresses, and so the crust will deform initially by plastic flow and then by faulting as the crust becomes more rigid. Thus, the degree to which the crust experiences faulting is largely governed by the periodicity of magmatism. It is widely accepted that magmatism is almost continuous at fast-spreading ridges, and, because of this, the crust of the axial regions remains hot and pliable there. At slowly spreading centers, considerable time probably elapses between magmatic pulses, and this allows mechanical deformation to create the faulted rift valley that characterizes these features.

At fast-spreading centers, the axial region is characterized by a distinct topographic elevation that rises above the regional level of the seafloor. Side-scan sonar imagery in the axial region allows investigators to map volcanic flows of different ages using the property that different-aged material will have a different extent of sediment cover, and this will affect the backscatter properties differently (Fig. 3). The same data show the pattern of faults and fissures as linear bands of intense reflectivity. In conjunction with other work, side-scan studies have shown that the zone of active volcanism and faulting at the crest of the axial ridge is surprisingly narrow, only 5 to 10 km wide (Bicknell *et al.*, 1988; Macdonald and Luyendyk, 1977). In fact, most of the volcanism occurs in a depression first described as an axial summit graben and now more typically described as the axial summit caldera (ASC). This feature is only a few hundred meters to perhaps a kilometer wide (Fig. 13). This observation has proven to be a major puzzle to investigators. Melt is generated at depths in the mantle of about 60 km, over a region perhaps a hundred or so kilometers in breadth. That melt emerges at the crest of a ridge over a region that is so small relative to the zone of melt generation suggests that a mechanism exists that can efficiently focus melt into the region of the ridge axis. A considerable amount of theoretical modeling has been carried out in an effort to explain this process, but the models require the input of unknown parameters such as the viscosity and permeability of the mantle. Each different parametrization leads to a different theoretical pattern of flow. To investigate the problem, a large-scale experiment is planned for a very fast-spreading region of the southern East Pacific Rise, in which seismic and electrical methods will be used to map the pattern of flow. Electrical methods have not been widely employed in studies of the oceanic lithosphere, but they hold great promise since they are able to sense the presence of melt more directly than seismic methods because the differences in electrical conductivity of liquid and solid phases of the same material are much larger than the corresponding differences in seismic wave speed for these same materials.

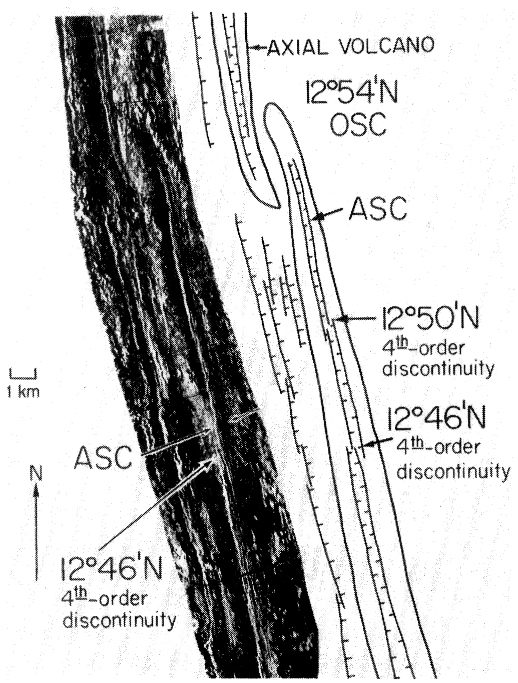

FIG. 13. Side-scan sonar image of the axial region of a fast-spreading ridge and its interpretation. Images like these show the narrow region only a few hundred meters wide called the axial summit caldera (ASC in the illustration) within which most recent volcanic eruptions are confined. Data from Macdonald *et al.* (1993).

The migration of melt to the surface probably does not occur in a sheetlike fashion, but instead the melt rises in a series of relatively isolated plumes. It is now widely held that this focused upwelling is largely responsible for the segmentation of ridge morphology that swath mapping surveys have made apparent. One particularly powerful line of evidence comes from gravity observations (Lin *et al.*, 1990). When mantle Bouguer anomalies are calculated for regions of the slowly spreading Mid-Atlantic Ridge, they reveal remarkably circular patterns that can be associated with the morphological segments (Plate 4). These patterns identify low gravity values, suggesting that they are associated with deficiencies in mass beneath the centers of spreading segments. Because melt is less dense than the parent mantle, the gravity lows may well represent areas of melt concentration in upwelling plumes.

While the inferred plumelike rise of magma to the surface beneath the centers of spreading segments reveals fundamental structure of the deep mantle beneath, the present behavior of deep-mantle processes beneath the boundaries between segments may tell us about the time history of magmatic processes. Swath mapping has shown that there is a rich variety of discontinuities between individual spreading segments, and these take different forms for ridges spreading at different rates. At fast-spreading centers, these include regions where portions of the spreading system actually run in parallel for a short distance [Plate 5(a)]. These overlapping spreading centers (OSC) are very common on the East Pacific Rise, where they have been extensively mapped. The combination of detailed swath mapping and magnetic surveys shows that OSCs are not stable structures, but move along the ridge axis—apparently one limb of the OSC typically dominates over the other and lengthens while the other shortens. Evidence for this comes in the form of an apparent wake of deep seafloor left behind by the motion of the OSC, together with a striking pattern of magnetic anomalies. In Plate 5(b), we see that the ridge parallel lineations generated by the seafloor spreading process are overprinted in the central part of the map by a band of magnetic highs that is almost orthogonal to the trend of the ridge axis. This high magnetization is believed to result from chemical processes associated with the cooling of magma in a region that is relatively starved of magma, such as might occur on the dying limb of an OSC. The observation that the band of magnetic highs is not quite orthogonal to the ridge but forms a "V" shape suggests that the location of the position of the dying limb has moved in time. The overall motion is tracked by the shape of the wake and is to the south, but irregularities in the V show that there may have been an initial step to the north. This implies that one limb may dominate for a period of time and give way to the other.

2.2 The Shifting Paradigm II: Magma Chambers at Midocean Ridges

In that they appear as near proxies to geological cross sections, seismic reflection images have become a commonly used and extremely powerful tool in our quest for understanding the physical processes associated with seafloor spreading. A typical reflection image of the fast-spreading region of the East Pacific Rise south of the Garrett Fracture Zone (Fig. 14) reveals the structure beneath the elevated spreading axis that is well described from bathymetric mapping (see Sec. 2.1). The most distinct feature of the subsurface is a layer that approximately parallels the seafloor except immediately under the axis, where it shallows considerably. This reflection event lies at the boundary between rocks with relatively low velocity and a higher-velocity layer as determined from analysis of common-midpoint gathers that form part of the processing steps for creating the image, and from travel-time and wave-form analysis of expanding spread profile refraction data obtained in the same area. The boundary between slow and fast layers appears to be extremely sharp and has led to considerable speculation over the responsible geological mechanism. The problem is that the rock units themselves cannot vary very significantly. The rocks are all derived from the melting of mantle, so that we do not anticipate extreme variation in rock type—all the rocks that form the crust are basically basaltic in composition. What then can give rise to such an extreme change in the near surface?

Most investigators believe that the change relates to the mechanical state of the layers and not to rock type. At the ridge axis, lava is extruded to the surface, where it forms a highly fractured layer. Lavas are erupted at temperatures in excess of 1000 °C into water that is at effectively 0 °C, and the result is that they cool very rapidly and are usually very highly fractured. Rocks that solidify at deeper levels in the crust cool slowly and are less fractured. Because of this, the seismic velocity of rock layers near the surface does not represent the velocity of the rock itself but a combination of the rock and the material filling the cracks and fractures. In fresh lava at the surface, the cracks are filled with water, and the seismic velocity measured from the surface is much less than that of the extruded basalts themselves. The slow surface velocities are therefore thought to be indicative of the high degree of fracturing of the extruded sequences. The strong reflection is widely thought to mark the base of a layer of extruded lava, beneath which lie relatively unfractured intruded rocks that never came into direct contact with seawater. The boundary is then regarded as the mechanical interface between highly fractured and less fractured rocks. The very distinct nature of the boundary is surprising, however, since it might be expected that the degree of fracturing would diminish progressively with depth,

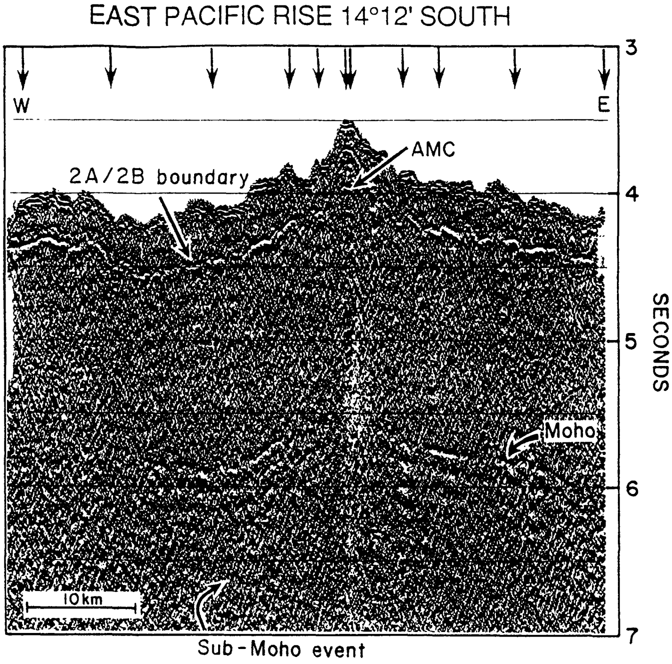

FIG. 14. Seismic reflection image across the crestal region of the East Pacific Rise, where the spreading rate is as fast as is known from any current ridge. Features revealed in the image are discussed in the text.

rather than change abruptly. Furthermore, the rapid deepening of the boundary down the flanks of the ridge axis has no agreed-upon explanation. One explanation is that it may be caused by off-axis extrusion of lava. There is some evidence in side-scan images that lava units flow down the flank of the axial high, but this explanation seems to contradict the evidence that *most* volcanism is confined to the narrow axial summit caldera. Another explanation is that the mechanical boundary may move down in the crust with time as a result of the effects of hydrothermal circulation that might drive a cracking front down into the crust. This mechanism, too, has not been substantiated.

At deeper levels, beneath the strong event in the image, we see the base of the crust marked by the Mohorovičić discontinuity, or Moho, and deeper still are indistinct features that may represent structure within the mantle. The reflection image also detects a tiny bright signal from just below the rise axis. The strength and spectral properties of the reflection suggest that it is the response of a body with much lower velocity than its surroundings and may therefore mark the top of a liquid magma body, the axial magma chamber (AMC). This feature is much smaller than any structure resolved by travel time or attenuation tomography (Fig. 7 and 9), and this testifies to the resolution potential of the reflected wavefield. The magma body, resolved by reflection images, is so small, in fact, that the term "chamber" seems barely applicable. It is typically not more than 1-km wide, and the thickness of the melt layer (judged from modeling refraction seismic data, and the spectral characteristics of the AMC reflection itself) may be less than 100 m.

Despite being very small and probably very thin, the AMC is surprisingly continuous along the axis of the rise. Plate 6, from Detrick *et al.* (1993), shows intersecting crossings of the southern East Pacific Rise on which the AMC is clearly visible. Although it is only 1 to 2 km wide at most, the AMC shows along-axis continuity for several tens of kilometers. The seismic reflection images therefore lead to the suggestion that the AMC might better be regarded as a high-level sill-like structure than as a chamber *sensu stricto.*

No reflection is observed from the base of the AMC or from its sides. Thus, although the reflection image gives a high-resolution image of one part of the structure, it fails to provide information on the larger structure surrounding the AMC. Because the reflection method images the response to relatively sharp changes in density and/or wave speed, what may be implied is that the lens of liquid magma has sides and a base that are too gradational to appear in the reflected response of the crust. Results of travel time and attenuation tomography reveal a region of reduced velocities that is much broader than that represented by the reflection image of the AMC, and this may imply that the AMC reflection arises from a thin pool of liquid magma that lies atop a broader zone containing traces of melt or perhaps solid rock in which the temperature has been raised by proximity to the region of melt.

If the pool of liquid magma is indeed very small, the crust cannot form by the simple process of solidification of magma since this would only account for a comparatively small amount of material. The crust above the magma chamber is formed by dike injection and lava eruption, but the crust beneath requires a different process. Recent models of oceanic crustal formation put forward by Phipps Morgan and Chen (1993), and similar models of Quick and Denlinger (1993), offer an explanation. These models use numerical techniques to calculate the temperature, crustal flow, and strain fields for models of spreading centers in which the crust is produced by a magma lens that is arrested in its upward descent when it reaches a freezing horizon controlled by hydrothermal cooling. The lens will reach shallowest levels in crust produced at fast spreading rates and be absent in crust forming at slow rates, consistent with the seismic reflection observations of magma-chamber characteristics. Figure 15 compares the calculations for three spreading rates. One critical prediction of this model is the great difference in crustal strain patterns predicted for the different spreading rates. The overall geometry of the strain field is similar in that we observe greater strain in the lower crust with elongated strain ellipses dipping toward the axis, flattening out as the Moho is approached. In fast-spreading crust, the crustal strain is very large, while at slow rates the strain is moderate.

If we were to somehow introduce strain markers into the crustal sections, they would

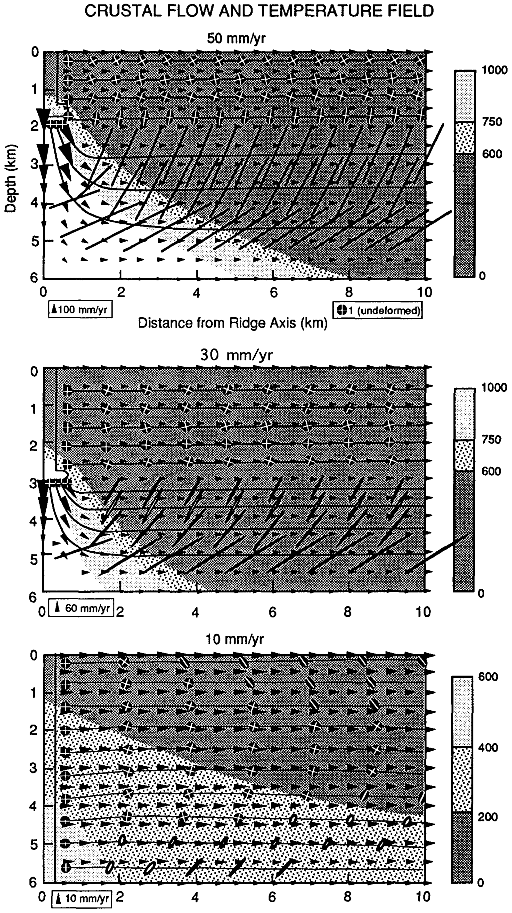

FIG. 15. Temperature, crustal flow, and strain fields calculated by Phipps Morgan and Chen (1993) using numerical simulation techniques to account for the creation of oceanic crust from a small magma body located within the crust. The crust above the magma chamber is produced by dike injection and volcanic eruption, while that below forms from the magma body and flows into the shape defined by the strain markers. These trace the shape of crustal layering and show a strong dependence on spreading rate (fastest is shown at the top, slowest at the bottom). Some of the differences in the seismic images of fast- and slowly-spreading crust (Plate 2) may be attributable to this mechanism of crustal accretion.

be deformed into the "stretched ellipses" shown in Fig. 15. In an actual ridge setting, this means that solidifying lower crustal rocks experience greater ductile strain at fast-spreading ridges than at slowly spreading ridges. Thus, crustal layering would be "strung out" to a much greater extent at fast-spreading centers than at slowly spreading centers and hence produce thinner layering. This may provide a key to understanding some of the differences in reflectivity of slow- and fast-spread crust (Plate 2). When layering is present of the type believed to constitute the features in the crust of the western North Atlantic, the response of the layers depends very strongly on its thickness. Obviously, a pathological case is established when the layer has zero thickness and hence cannot be resolved regardless of the wavelength of the seismic source. As layer thickness increases, its detectability does not simply improve as the thickness increases. In fact, a tuning phenomenon occurs in which reflections from the top and bottom of a thin layer can interfere constructively and produce a combined reflection that greatly exceeds the strength of the reflected response of either interface forming the layer. It is not unreasonable to assert that, given the noisy data field in which these images are created, crustal layering may *only* be detectable if it is developed at thicknesses near to the tuning thickness. One possible explanation for the differences in crustal reflectivity between slow-spread and fast-spread crust is not that the layering is absent in fast-spreading crust, but that the strain associated with crustal generation at a fast-spreading ridge is so large that any layering that develops is strung out so thinly that it becomes greatly less than the tuning thickness. The layering may be present, but seismically undetectable.

GLOSSARY

Airgun: Device that uses the discharge of compressed air to create a source of acoustic energy for marine seismic experiments.

AMC (Axial Magma Chamber): A strong shallow feature recognized in seismic reflection images of the axis of fast-spreading mid-ocean ridges such as the East Pacific Rise. It is believed to represent a thin pool of liquid magma that is the reservoir for lava eruptions at the ridge axis.

ASC (Axial Summit Caldera): A narrow, shallow, nearly linear depression in the summit of fast-spreading midocean ridges. Recognized in high-resolution sonar images of the seafloor, it is thought to represent a type of caldera collapse structure.

Bathymetry: The shape of the seafloor defined from the results of echo sounding experiments. Can be represented as profiles or contours or in perspective renderings.

DMO (Dip Move-Out): A seismic-image enhancement process used in conjunction with others to create clear images of the subsurface in areas of complex structure.

Eötvös Effect: The change in apparent value of gravitational acceleration caused by the motion of the platform on which the measurement is being made.

ESP (Expanded Spread Profile): A technique for acquiring wide-angle marine seismic data using two vessels, in which the midpoint position between the vessels is maintained fixed as the vessels separate, one discharging a seismic source while the other receives the seismic transmissions.

Free-Air Anomaly: The simplest means of expressing the gravitational effect of density heterogeneity in the earth. Obtained by removing a standard reference field from observed values reduced to the same plane as the reference.

Gather: A term originating in the oil exploration industry to describe the reorganization of seismic data from the geometry in which they were recorded to one more suitable for image processing.

Genetic Algorithm: A procedure for estimating unknown parameters from a suite of measured quantities related to the unknowns. Based on testing a population of models defined by test parameter strings by processes of selection, crossover, and mutation that are analogous to those of biological systems. Currently becoming used in seismic inversion.

GPS (Global Positioning System): A navigation system that uses a constellation of military satellites with accurately known positions as the basis for highly accurate, continuous position information.

Gravimeter: Device for measuring the strength of the Earth's gravitational field. Those used at sea rely on the principle of a balance beam or feedback accelerometer.

Induced Magnetization: The magnetic field of a rock body that is caused by the interaction of the Earth's field with magnetic minerals within the rock.

Lithosphere: The outer shell of the Earth (including crust and upper mantle) that deforms in a relatively rigid manner. The mosaic of plates that move across the face of the Earth is formed of lithosphere.

Magnetometer: Device for measuring the strength of the Earth's magnetic field. Most marine instruments use the forced precession of protons in the Earth's field as the basis for the measurement.

MBA (Mantle Bouguer Anomaly): The anomalous gravity field that remains after removing the effect of the shape of the seafloor and an estimate of the shape of the crust–mantle interface.

Midocean Ridge: Elongate, mountainous structure that occurs in most ocean basins and is the locus for creation of new lithosphere by seafloor spreading.

Migration: A term used in seismic-reflection image enhancement that refers to a process by which image geometry is positioned to portray structure in the earth more accurately.

Moho: A shorthand term for the Mohorovičić discontinuity that is the seismically defined interface beneath which propagation velocity exceeds 8.0 km/s, normally taken to be that in the mantle. Moho is therefore usually considered to represent the crust–mantle interface.

Multibeam Echo Sounder: Device for measuring the shape of the seafloor using sound transmitted from an array of transducers on the hull of a vessel. These send a wide swath of sound into the ocean that insonifies a swath of the seafloor, allowing for the efficient mapping of large areas of the seafloor.

NMO (Normal Move-Out): The process of correcting seismic data recorded on an extended array of receivers for the effect of transmission times to different parts of the array.

OBS (Ocean Bottom Seismometer): Device for recording seismic information either from artificial sources such as airguns or explosives, or from earthquakes. They typically comprise an external sensor attached to a pressurized instrument package that rests on the seafloor during recording.

OSC (Overlapping Spreading Center): Feature of fast-spreading midocean ridges in which two parts of the ridge run nearly parallel for a distance of as much as 50 km.

Ray: A construction used in seismic data analysis to describe the passage of seismic energy through the earth.

Remnant Magnetization: Magnetic property of a rock acquired at the time of its formation that it holds independent of the present magnetic field.

Seafloor Spreading: The process by which new lithosphere is created at midocean ridges through the upwelling of mantle and its partial melting and cooling.

Side-Scan Sonar: Device for measuring the strength of the return of acoustic energy from the seafloor in a broad swath. Strong returns and shadows provide a proxy for structure and geological formations.

Spreading Segment: A morphological portion of a midocean ridge, typically centered around an elevated portion of the axial depth, that is believed to be centered over an isolated, plumelike region of mantle upwelling. Segments are thought to be independent units of the ridge, communicating little with adjacent parts of the ridge.

Stacking: Term used in seismic-reflection image processing to describe the summation of signals from numerous parts of a recording array following NMO or DMO correction. Typically precedes migration in data processing.

Swath Mapping: Used in acoustic studies of the seafloor to describe the insonification of a wide region of the seafloor by arrays of transducers mounted either on the hull of a vessel or on a towed instrument package.

Synthetic Seismogram: The calculated seismic response of a specified Earth structure. Used as a guide to the interpretation of observed seismic data and as a basic for the inversion of seismic observations.

Tomography: A method of analysis of seismic data that applies to data obtained from arrays of instruments recording wide-angle information. The method creates a two- or three-dimensional estimate of the structure by optimizing travel times to all instruments from all the shots fired during the experiment. Resolution is related to the number of ray paths that cross one another.

Works Cited

Aki, K., Richards, P. G. (1980), *Quantitative Seismology Theory and Methods,* San Francisco: W. H. Freeman.

Backus, G. E. (1958), "A class of self sustaining dissipative spherical dynamos," *Ann. Phys. (New York)* **XX**, 372–447.

Bicknell, J. D., Sempere, J.-C., Macdonald, K. C., Fox, P. J. (1988), "Tectonics of a fast spreading center: A Deep-Tow and Seabeam survey at EPR 19°30′S," *Marine Geophys. Res.* **9**, 25–46.

Blackinton, J. G., Hussong, D. M., Kosalos, F. (1983), "First results from a combination of side-scan sonar and seafloor mapping system (SeaMARC II), Houston, Texas," in: *Offshore Conference Proceedings,* OTC **4478**, pp. 307–311.

Bott, M. H. P. (1960), "The use of rapid digital computing methods for direct gravity interpretation of sedimentary basins," *Geophys. J. R. Astron. Soc.* **3**, 63–67.

Bullard, E. C., Gellman, H. (1954), "Homogeneous dynamos and terrestrial magnetism," *Philos. Trans. R. Soc.* **A247**, 213–278.

Cande, S. C., Kent, D. V. (1992), "A new geomagnetic polarity time scale for the Late Cretaceous and Cenozoic," *J. Geophys. Res.* **97**, 13917–13951.

Caress, D. W., Burnett, M. S., Orcutt, J. A. (1992), "Tomographic image of the axial low velocity zone at 12°50′N on the East Pacific Rise," *J. Geophys. Res.* **97**, 9243–9264.

Carey, P. W., Chapman, C. H. (1988), "Automated 1D waveform inversion of marine seismic refraction data," *Geophys. J.* **93**, 527–546.

Chapman, C. H. (1985), "Ray theory and its extensions: WKBJ and Maslov seismograms," *J. Geophys.* **58**, 27–43.

Claerbout, J. F. (1985), *Imaging the Earth's Interior,* Oxford: Blackwell.

Detrick, R. S., Harding, A. J., Kent, G. M., Orcutt, J. A., Mutter, J. C., Buhl, P. (1993), "Seismic structure of the Southern East Pacific Rise," *Science* **259**, 499–503.

Elsasser, W. M. (1956), "Hydromagnetic dynamo theory," *Rev. Mod. Phys.* **28**, 135–163.

Harrison, J. C. (1960), "The measurement of gravity at sea," in: S. C. Rucorn (Ed.), *Methods and Techniques in Modern Geophysics,* New York: Wiley-Interscience, pp. 211–229.

Heirtzler, J. R., Dickson, G. O., Herron, E. M., Pitman, W. C., III, LePichon, X. (1968), "Marine magnetic anomalies, geomagnetic field reversals and motions of the ocean floor and continents," *J. Geophys. Res.* **73**, 2119–2136.

Herzenberg, A. (1958), "Geomagnetic dynamos," *Philos. Trans. R. Soc.* **A250**, 543–585.

Jacobson, R. S., Lewis, B. T. R. (1990), "The first direct measurements of upper oceanic crustal compressional wave attenuation," *J. Geophys. Res.* **95**, 17417–17429.

Johnson, H. P., Helferty, M. (1990), "The geological interpretation of side-scan sonar," *Rev. Geophys.* **28**, 357–380.

Kelly, K., Ward, R. W., Treitel, S. (1976), "Synthetic seismograms—a finite difference approach," *Geophysics* **41**, 2–27.

Lin, J., Purdy, G. M., Shouten, H., Sempere, J. C., Zervas, C. (1990), "Evidence from gravity data for focused magmatic accretion along the Mid-Atlantic Ridge," *Nature* **344**, 627–632.

Macdonald, K. C., Fox, P. J. (1990), "The mid-ocean ridge," *Sci. Am.* **262** (6), 72–79.

Macdonald, K. C., Luyendyk, B. P. (1977), "Deep-tow studies of the structure of the Mid-Atlantic Ridge crest near 37°N," *Geol. Soc. Am. Bull.* **88**, 621–636.

Macdonald, K. C., Miller, S. P., Huestis, S. P., Spiess, F. N. (1980), "Three-dimensional modelling of a magnetic reversal boundary from inversion of deep-tow measurements," *J. Geophys. Res.* **85**, 3670–3680.

Macdonald, K. C., Fox, P. J., Perram, L. J., Eisen, M. F., Haymon, R. M., Miller, S. P., Carbotte, S. M., Cormier, M.-H., Shor, A. N. (1988), "A new view of the mid-ocean ridge from the behavior of ridge-axis discontinuities," *Nature* **335**, 217–225.

Mooney, H. M. (1983), "Synthetic seismograms for body waves: an overview," *First Break*, December, 9–20.

Muller, G. (1985), "The reflectivity method: a tutoral," *J. Geophys.* **58**, 153–174.

Parker, E. N. (1955), "Hydromagnetic dynamo models," *Astrophys. J.* **122**, 295–313.

Parker, L. R. (1973), "The rapid calculation of potential anomalies," *Geophys. J. R. Soc.* **31**, 447–455.

Parker, R. L., Huestis, S. P. (1974), "The inversion of magnetic anomalies in the presence of topography," *J. Geophys. Res.* **79**, 1587–1593.

Phipps Morgan, J., Chen, Y. John (1993), "The genesis of Oceanic crust: Magma injection, hydrothermal circulation, and crustal flow," *J. Geophys. Res.* **98**, 6283–6297.

Prince, R. A., Forsyth, D. W. (1988), "Horizontal extent of anomalously thin crust near the Vema Fracture Zone from the three-dimensional analysis of gravity anomalies," *J. Geophys. Res.* **93**, 8051–8063.

Quick, J. E., Denlinger, R. P. (1993), "Ductile deformation and the origin of layered gabbro in ophiolites," *J. Geophys. Res.* **98**, 14015–14027.

Reed, T. B., Hussong, D. (1989), "Digital image processing techniques for enhancement and classification of SeaMARC II side scan sonar imagery," *J. Geophys. Res.* **94**, 7469–7490.

Renard, V., Allenou, J.-P. (1979), "Sea Beam, multibeam echo-sounding in 'Jean Charcot': Description, evaluation and first results," *Int. Hydrogr. Rev.* **56**, 35–67.

Sempere, J. C., Macdonald, K. C., Miller, S. P. (1987), "Detailed study of the Brunhes/Matuyama reversal boundary on the East Pacific Rise at 19°30′S: Implications for crustal emplacement processes at an ultra fast spreading center," *Mar. Geophys. Res.* **9**, 1–23.

Shaw, P. R., Orcutt, J. A. (1985), "Waveform inversion of seismic refraction data and application to young Pacific crust," *J. Geophys. Res.* **82**, 375–414.

Stephen, R. A. (1983), "A comparison of finite difference and reflectivity seismograms for marine models," *Geophys. J.* **72**, 39–57.

Stoffa, P. L., Buhl, P. (1979), "Two-ship multichannel seismic experiments for deep crustal studies: Expanded spread and constant offset profiles," *J. Geophys. Res.* **84**, 7645–7659.

Talwani, M., Worzel, J. L., Landisman, M. (1959), "Rapid gravity computations for two-dimensional bodies with application to the Mendocino submarine fracture zones," *J. Geophys. Res.* **64**, 49–59.

Vera, E. E., Mutter, J. C., Buhl, P., Orcutt, J. A., Harding, J., Kappus, M. E., Detrick, R. S., Brocher, T. M. (1990), "The structure of 0- to 0.2 m.y.-old oceanic crust at 9°23′N on the East Pacific Rise from expanded spread profiles," *J. Geophys. Res.* **95**, 15529–15556.

Wilcock, W. S. D., Solomon, S. C., Purdy, G. M., Toomey, D. R. (1992), "The seismic attenuation structure of the East Pacific Rise," *Science* **258**, 1470–1473.

Worzel, J. L., Harrison, J. C. (1963), "Gravity at sea," in: N. M. Hill (Ed.), *The Sea*, New York: Wiley-Interscience, pp. 134–174.

Yilmaz, O. (1987), *Seismic data processing*, Investigations in Geophysics No. 2, Tulsa: Society of Exploration Geophysicists.

Further Reading

Texts that are suitable for obtaining an appropriate theoretical foundation in marine geophysical methods include:

Claerbout, J. F. (1985), *Imaging the Earth's Interior*, New York: Blackwell Scientific Publications, p. 398.

Menke, W., Abbott, D. (1990), *Geophysical Theory*, New York: Columbia University Press, p. 458.

Robinson, E. A., Treitel, S. (1980), *Geophysical Signal Analysis*, Englewood Cliffs, NJ: Prentice-Hall, p. 466.

Telford, W. M., Geldart, L. P., Sheriff, R. E., Kays, D. A. (1976), *Applied Geophysics*, Cambridge, U.K.: Cambridge University Press, p. 860.

A useful general text that includes descriptions of plate tectonic processes is:

Brown, C. G., Mussett, A. E. (1981), *The Inaccessible Earth*, London: Allen and Unwin, p. 235.

A recent general discussion of topics in marine geophysics appears in: *Anon.* (1992/3), "Marine geology and geophysics," *Oceanus* **35** (4), 104.

An introductory description of seismic reflection imaging appears in:

Mutter, J. C. (1986), "Seismic images of plate boundaries," *Sci. Am.* **254** (2), 66–75.

A very fine account of the history of the development of global ocean sciences by one of the founding investigators is given in:

Menard, H. W. (1986). *The Ocean of Truth, a Personal History of Global Tectonics*, Princeton, NJ: Princeton University Press, p. 535.

5
Measurement of Mass and Density

Richard S. Davis

5.1 **Introduction** **139**
5.2 **Mass** **139**
5.2.1 A Few General Principles 139
5.2.2 SI Unit, Relation to Atomic Mass 140
5.2.3 Mass Balances (General) 142
5.2.4 Need for a Buoyancy Correction 142
5.2.5 Specification of Secondary Standards 143
5.2.6 Analytical Balances 145
5.2.6.1 Design Considerations 145
5.2.6.2 Servocontrol and Automation 145
5.2.6.3 Force 146
5.2.7 Microbalances 147
5.2.7.1 Analytical Type, Adapted to Microweighing 147
5.2.7.2 Vibrating Transducers 147
5.2.8 Astronomical and Subatomic Mass 147
5.3 **Density** **148**
5.3.1 Measurements Based on Buoyancy 148
5.3.1.1 Hydrostatic Weighing 148
5.3.1.2 Flotation Methods 150
5.3.1.3 Hydrometry 150
5.3.2 Measurements Based on Mass and Volume 150
5.3.2.1 Liquid Pycnometers 151
5.3.2.2 Vibrating-Tube Densimeters 151
5.3.3 Miscellaneous Techniques 152
5.3.4 Reference Standards 153
Glossary 153
References 155
Further Reading 155

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

5.1
Introduction

Measurements of mass and density have had commercial importance since antiquity. With the rise of modern physics, beginning with Newton, mass (or density) became one of the fundamental quantities used to describe the natural world. In the eighteenth century, Lomonosov in Russia and Lavoisier in France demonstrated that mass is conserved in chemical reactions. Analytical balances soon became the one piece of equipment common to all chemistry laboratories. Balances have remained essential tools of physics and chemistry even though our understanding of mass has changed greatly since the time of Lavoisier.

The first section of this chapter is devoted to mass measurements. After some basic principles, the unit of mass in the International System of Units (SI) is defined, and research toward a redefinition in terms of physical constants is mentioned. Practical considerations in the specification of secondary mass standards are introduced. Balances, the essential tools that allow comparison of one mass with another, are then discussed. We conclude the section with the link between the laboratory standard of mass and the mass of astronomical bodies or subatomic particles.

The second section, on density measurements, reviews the most widely used techniques for the accurate determination of this quantity.

5.2
Mass

5.2.1
A Few General Principles

A parameter called *mass* appears in fundamental equations of physics. Instead of defining this quantity, we take the view that the equations themselves show how the mass ratio of any two entities can be measured.

Newton recognized that the quantity now called *mass* fulfilled two roles in his physics: as the proportionality constant between force and acceleration (inertial mass) and as a parameter in describing the force of gravitation (gravitational mass). We may further distinguish two types of gravitational mass: active mass, responsible for generating the gravitational field, and passive mass, influenced by the field. In Newtonian physics, the two types of gravitational mass are identical, but

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

their experimentally observed equivalence to inertial mass appears fortuitous.

The postulate that the ratio of inertial mass to passive gravitational mass is the same for all objects is now known as the *weak equivalence principle or the universality of free fall* (see, e.g., Misner, Thorne, and Wheeler, 1973). The principle may safely be assumed for the practical measurements described later.

According to special relativity, the rest mass of an object is not simply the total rest mass of all its parts. The rest mass also includes a contribution E/c^2, where E is the sum of all internal potential and kinetic energies and c is the speed of light. Thus, for example, the Coulomb energy within an atomic nucleus contributes to its rest mass (e.g., about 0.4% in the case of gold). It is the total mass energy that is conserved, and there is no mass measurement that can distinguish between the individual components of rest masses and energies.

In general relativity, mass plays the role of generating the curvature of space-time.

5.2.2
SI Unit, Relation to Atomic Mass

The SI provides a practical framework for the precise measurement of physical quantities. The kilogram is among the base units of the SI and is now unique in that its definition remains tied to an artifact (BIPM, 2006):

> *The kilogram is the unit of mass: it is equal to the mass of the international prototype of the kilogram.* 3rd General Conference of Weights and Measures (CGPM), 1901.

The international prototype, the artifact referred to, is a cylinder made of an alloy of 90% platinum and 10% iridium. Since its fabrication in the 1880s, it has been stored and used at the International Bureau of Weights and Measures (BIPM) in Sèvres, France, according to provisions of the Convention du Mètre (BIPM, 2006). The national laboratories of most industrialized countries are equipped with 1-kg prototypes that are copies of the defining artifact. The mass in kilograms of these national prototypes is traceable through calibration at the BIPM to the international prototype. It is then the task of the national laboratories to calibrate standards whose masses are multiples or submultiples of 1 kg for the requirements of science, industry, and commerce.

In essence, the mass m (in kilograms) of any object is the ratio m/m_0, where m_0 is the mass of the international prototype. Operationally, the desired ratio is achieved by a chain of calibrations:

$$\left(\frac{m}{m_n}\right)\left(\frac{m_n}{m_{n-1}}\right)\cdots\left(\frac{m_i}{m_j}\right)\cdots\left(\frac{m_1}{m_0}\right) \tag{5.1}$$

where intermediate ratios involve secondary standards. Every intermediate ratio m_i/m_j in the chain represents a measurement (or series of measurements) and thus contributes to the uncertainty of the final result. Intermediate ratios involving artifacts whose mass ratio is not nominally unity or whose densities are not nominally equal pose particular problems, as discussed later (Sections 5.2.3 and 5.2.4).

In the case of other base units of the SI, such as the meter, it became clear long ago that the original artifact standard was noticeably unstable when compared with a ‘‘natural’’ standard, such as the wavelength of light emitted in certain atomic transitions. An analogous situation has yet to be confirmed as regards the kilogram. However, the precision with which 1-kg artifacts can be compared

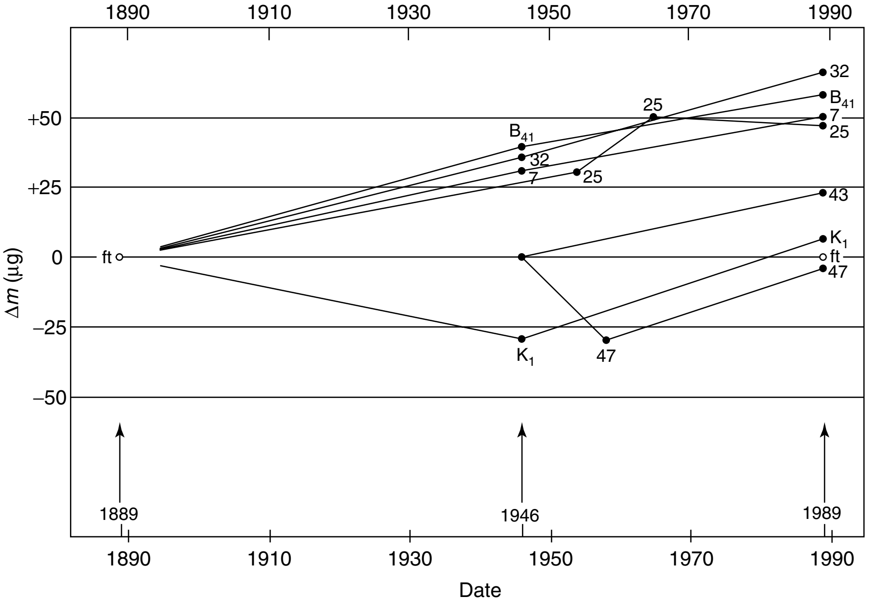

Fig. 5.1 The change in mass of seven similar artifacts since their first calibration, always with respect to the mass of the international prototype of the kilogram. The mass of the latter is defined to be 1 kg. The artifacts numbered 43 and 47 were first measured in 1946; the others in 1889. Courtesy of the Bureau International des Poids et Mesures.

is about 1 μg (1×10^{-9}), and one can easily see that a group of Pt/Ir kilogram artifacts made at the turn of the nineteenth century is slowly drifting with respect to the international prototype (Figure 5.1; Girard, 1994). It is conceivable, though unsubstantiated, that the entire group is drifting with respect to the constants of nature (Quinn, 1991). Several schemes are being pursued with a goal of checking the stability of the international prototype with respect to physical constants (Mohr, Taylor, and Newell, 2008). As yet, none of these schemes has achieved a precision sufficient to interpret the drifts shown in Figure 5.1. However, redefinition of the kilogram in terms of a physical constant, such as Planck constant, *h*, would already benefit electrical metrology and some areas of physics (Mills *et al.*, 2006). Redefinition of the kilogram, if it occurs, would maintain historical continuity.

Atomic masses can be measured to high precision relative to the mass of carbon-12 (Audi, Wapstra, and Thibault, 2003). The SI mole is the unit of "amount of substance" linking macroscopic and atomic mass: 1 mol of unbound atoms of carbon-12 at rest in their ground state has by definition a mass of exactly 0.012 kg. Each such atom has a mass of exactly 12 atomic mass units (u). The number of entities in 1 mol of any specified substance is the Avogadro constant N_A(6.022 141 79(30) $\times 10^{23}$ mol^{-1} (Mohr, Taylor, and Newell, 2008)), numerically equivalent to the number of atomic mass units in 0.001 kg. Hence, a measurement of the Avogadro constant to sufficient accuracy (and repeated at regular intervals) would provide a practical means of realizing the kilogram in terms of the mass of a single atom of carbon-12. Since the value of the molar Planck constant, $N_A h$, is already well known, redefinition of the kilogram could also be made in terms of h (Mills *et al.* 2006; Mohr, Taylor, and Newell, 2008).

5.2.3
Mass Balances (General)

The mass ratios in Equation (5.1) may be determined to high accuracy by the use of analytical balances. In principle, these devices are transducers whose output can be related to the difference in mass between two objects. For the smallest uncertainty, the two objects are nominally equal and the range of the transducer (or on-scale range of the balance, to use more traditional terminology) is a small fraction of the mass of each object.

The balance capacity puts an upper limit on the nominal mass of each of the objects no matter how small the difference between them. The use of such a transducer requires a companion set of mass standards that can be assembled to approximate the mass of the unknown.

Modern analytical balances have on-scale ranges equal to their capacity. This is a major convenience. If the scale is linear, it can be calibrated with a single mass standard. However, demonstrating linearity requires additional standards, so that even these balances achieve lowest uncertainty when used to compare the mass of nominally equal objects.

The existence of narrow-range mass transducers is sufficient to create standards whose mass is any multiple or submultiple of 1 kg. For instance, it is possible, using a sensitive balance, to demonstrate that 10 artifacts are approximately equal in mass and that the 10 taken as a group have a mass approximately equal to 1 kg. Thus, the mass of each of the 10 artifacts is nominally 100 g. Corrections to the nominal mass can be calculated by taking the balance readings into account. The uncertainty in each correction is ultimately determined by the precision of the balances that were used (see, e.g., Prowse and Anderson, 1974).

5.2.4
Need for a Buoyancy Correction

Weighings carried out in air (or any other fluid) will require correction for buoyancy. If comparison weighing is done in air of density ρ_a, then the unknown object acts as if its effective weight were $m(\text{eff})g$:

$$m(\text{eff}) = m - \rho_a V \tag{5.2}$$

$$= m(1 - \rho_a/\rho) \tag{5.3}$$

where V is the volume of the object, ρ is its density, and g is the gravitational acceleration in the laboratory. For simplicity, assume the weighing is done on a linear balance whose scale range equals its capacity. The balance output is routinely calibrated weighing a standard of known mass m_s, with nominal value near the capacity. The calibration procedure adjusts the scale output to read (m_s), where {} indicates the numerical value of the quantity within the brackets. However, the balance has actually responded not to m_s but to the weight $m_s(\text{eff})g$

$$m_s(\text{eff}) = m_s(1 - \rho_a/\rho_s) \tag{5.4}$$

where g is the local gravitational acceleration and ρ_s is the density of the standard. Consequently, m is related to its balance reading $\{y\}$ by

$$m = \{y\}\,\frac{(1 - \rho_a/\rho_s)}{(1 - \rho_a/\rho)} \tag{5.5}$$

The quantity $\{y\}$ is often referred to as the *apparent mass of the unknown object*. The local gravitational acceleration is the same for the standard and unknown and thus does not appear in Equation (5.5) as long as the centers of mass of the two objects are at the same elevation (the relative gradient of g is about $-3 \times 10^{-9}\ \text{cm}^{-1}$ at the surface of the earth). The buoyancy terms are more problematic, especially when one is weighing materials much less dense than the standard. Since ρ_s is typically $8000\,\text{kg m}^{-3}$ and ρ_a is about $1.2\,\text{kg m}^{-3}$ in laboratories at sea level, the correction for air buoyancy is, for example, roughly 0.1% for water.

The air density in a given laboratory whose temperature is controlled to within a few degrees will be mainly a function of ambient air pressure and thus will not vary beyond 5% of its average value. Air pressure is a function of elevation so that the mean air density in laboratories at 1600 m above sea level is about $1.0\,\text{kg m}^{-3}$. For lower uncertainty, the density of air can be computed from an equation of state whose input parameters are temperature, barometric pressure, relative humidity (or dew point), and atomic fraction of carbon dioxide (Picard *et al.*, 2008). A simplified version of this equation (OIML, 2004a) is adequate to achieve an uncertainty within 0.01% in air density:

$$\rho_a = \frac{0.34848p - 0.009(hr) \times \exp(0.061t)}{273.15 + t} \tag{5.6}$$

with ρ_a in kilograms per cubic meter, and where the air in the balance case has the following measured parameters: barometric pressure p (hectopascal or millibar), temperature t (degree Celsius), and relative humidity hr (percent).

5.2.5
Specification of Secondary Standards

The SI unit of mass is disseminated by means of secondary standards. Analytical mass standards are readily available commercially in denominations from 1 mg to 50 kg. Summations up to 400 tons of much larger artifacts may be needed to calibrate force transducers such as strain-gauge load cells. Because of the difficulty in their manipulation, large masses are most easily calibrated at the point of use (Bray, Barbato, and Levi, 1990). In recent years, there has been a growing need for standards with mass smaller than 1 mg. Production of such standards, previously left to the ingenuity of the individual user, is now being taken up by a number of national metrology institutes.

Mass standards used in legal metrology, named *weights*, are manufactured to specifications set by each country, usually in accordance with Recommendation R111-1 of the International Organization of Legal Metrology (OIML, 2004a). Thus, the mass of artifacts of a particular class are specified to be within certain maximum permissible errors, as shown in Table 5.1. If the user needs to know the mass to better accuracy, then each artifact must undergo calibration so that individual deviations from the nominal mass can be known.

The need for buoyancy corrections to weighings in air has led to conventions for the construction of weights. OIML, specifies the following convention (OIML, 2004b).

1. An artifact of mass m_s and density ρ_s has a conventional mass m_c defined as

$$m_c = m_s \frac{(1 - 1.2/\rho_s)}{(1 - 1.2/8000)} \tag{5.7}$$

It is the conventional mass that is referred to in Table 5.1 and in typical

Tab. 5.1 Maximum permissible errors for three best accuracy classes of weights, as recommended by the OIML.

Nominal value	Class E$_1$	Class E$_2$	Class F$_1$
50 kg	25 mg	80 mg	250 mg
20 kg	10	30	100
10 kg	5.0	16	50
5 kg	2.5	8.0	25
2 kg	1.0	3.0	10
1 kg	0.5	1.6	5
500 g	0.25	0.8	2.5
200 g	0.10	0.3	1.0
100 g	0.05	0.16	0.5
50 g	0.03	0.10	0.3
20 g	0.025	0.08	0.25
10 g	0.020	0.06	0.20
5 g	0.016	0.05	0.16
2 g	0.012	0.04	0.12
1 g	0.010	0.03	0.10
500 mg	0.008	0.025	0.08
200 mg	0.006	0.020	0.06
100 mg	0.005	0.016	0.05
50 mg	0.004	0.012	0.04
20 mg	0.003	0.010	0.03
10 mg	0.003	0.008	0.025
5 mg	0.003	0.006	0.020
2 mg	0.003	0.006	0.020
1 mg	0.003	0.006	0.020

reports of calibration, although the mass m_s may be given as well. To limit confusion, m_s is sometimes be referred to as the *true mass* or *vacuum mass* of the standard if it is reported along with the conventional mass. The user is normally expected to treat the artifact standard as if its mass were m_c and its density were 8000 kg m^{-3}. From Equations (5.4), (5.5), and (5.8), this fiction does not propagate as error provided $\rho_a = 1.2\,\mathrm{kg\,m^{-3}}$. In general, however, the use of conventional mass does introduce an error e given as

$$e \approx m_s(1.2 - \rho_a)\left(\frac{1}{8000} - \frac{1}{\rho_s}\right) \qquad (5.8)$$

Recommendation R111-1, therefore, contains a second specification.

2. The density ρ of an artifact standard must be such that a deviation in ρ_a of 10% from the assumed value of 1.2 kg m^{-3} will lead to an error that does not exceed one-fourth the absolute value of the maximum permissible error given in Table 5.1.

The introduction of conventional mass and density is intended to simplify the data analysis of routine weighing near sea level. The convention must, of course, be abandoned if its use will lead to serious error, for instance, if mass standards are used in vacuum.

The magnetic properties of secondary mass standards are sometimes of interest to experimentalists. Austenitic alloys of stainless steel and plated brass (where it meets the density requirements) are the most often used materials. The best grades of stainless steel have relative permeabilities less than 1.004. Common grades of brass and 18/8 stainless steel are much worse.

5.2.6
Analytical Balances

5.2.6.1 Design Considerations

The design of analytical balances has evolved significantly due to the perfection of electromagnetic servocontrol. The basic mechanical configuration is, however, traceable to ancient origins: two pans suspended symmetrically from the ends of a horizontal beam. The operating principle, shown schematically in Figure 5.2, is that a mass imbalance between the two pans creates a net torque on the beam so that its equilibrium angle is no longer zero. The sensitivity S is

$$s \equiv \frac{d\phi}{dm_2} = \frac{L}{m_1 l + (m_2 + m_3)h} \tag{5.9}$$

which will depend on the total load if the three pivots are not aligned. The most important design features of this type of balance are reviewed by Quinn (1992) and Speake (1987).

The so-called top-loading balances are designed with the pans above the beam. In order that an object produces the same torque no matter where it is placed, the pan must be guided so that it does not rotate with the beam. This can be accomplished by using twin beams in the form of a parallelogram. For large masses, the pan may be additionally equipped with a self-leveling mechanism. This design, though allowing unimpeded access to the pan, did not become practical for high-precision balances until the advent of servocontrol.

5.2.6.2 Servocontrol and Automation

The convenience and practical precision of balances are improved by automation and servocontrol. One can, in fact,

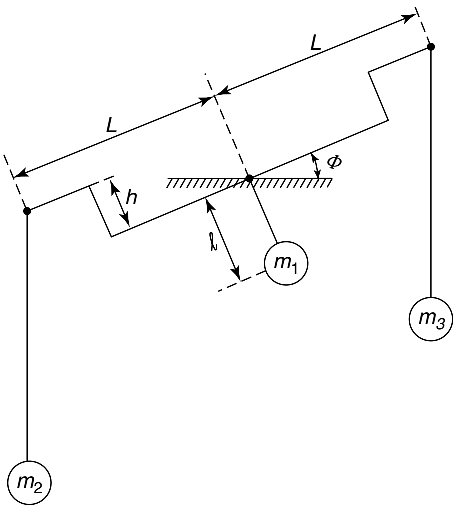

Fig. 5.2 Schematic of an equal-arm balance without servocontrol. The three pivot points are represented by solid circles. The mass of the beam is m_1 and is centered a distance ℓ below the central pivot. The loads on each pan are represented by m_2 and m_3.

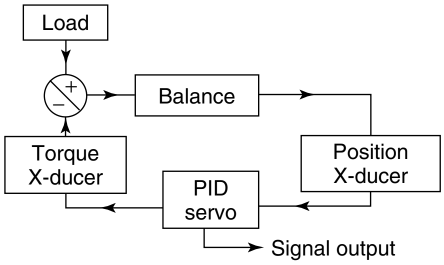

Fig. 5.3 Typical arrangement of a servocontrolled balance. The load, which would cause an unserved balance beam to rotate, is compensated by an electromagnetic torque transducer.

consider electronic balances as servocontrolled, modified versions of the balances presented above (Schoonover, 1982). Figure 5.3 shows a typical control loop where the balance is seen to be a transducer that converts a torque imbalance to an angular deflection that is itself converted to a voltage. An amplifier acting through an electromagnetic transducer supplies feedback to maintain the beam at constant angle. To optimize accuracy and transient response, the feedback signal is typically a linear combination of a force proportional to the angular deflection, its time integral, and its time differential (PID feedback).

There are several practical benefits of servocontrol. The rotation of the beam can be greatly restricted so that relatively robust leaf springs (also known as *flexure strips*) can replace delicate knives at the pivot points. The beam can be maintained precisely horizontal, thereby rendering the balance immune to spurious horizontal forces. The sensitivity of the unservoed balance can be adjusted until a tilt of the balance case causes the beam to tilt through exactly the same angle. A balance adjusted to this "autostatic" condition is relatively immune to common sources of mechanical noise (Speake, 1987). Finally, the on-scale, linear range of the balance can be greatly increased. Quinn (1992) describes a servocontrolled balance with a precision of order 0.05 μg for loads up to 2 kg. Speake (1987) derives fundamental theoretical limitations to balance precision. Commercial balances are now available that can compare 1-kg mass standards to a precision of 0.1 μg in air or vacuum.

Commercial balance manufacturers include a variety of convenience features in servocontrolled balances. The most common of these are automatic or semiautomatic scale calibration and suppression of the zero drift. Depending on the sophistication of the balance and its accessories, complete weighing operations can be fully automated.

5.2.6.3 **Force**

Analytical balances cannot distinguish the gravitational force on a mass from other vertical forces. In the case of air buoyancy, corrections can be applied and error avoided. Other forces must be made negligible. Of these, the principal concerns are moving air and electrostatic and magnetic interactions.

Thermal effects are minimized by keeping the objects to be weighed in good thermal equilibrium with the interior of the balance. Otherwise, a convection cell can be created that may result in serious systematic errors (Gläser, 1999).

Electrostatic problems often occur when one is weighing good insulators such as fused silica or poly(tetrafluoroethylene). Equipping the balance pan with a Faraday cage of stable mass can eliminate such problems, as can exposing the insulator to ionized air.

The weighing to small uncertainty of permanent magnets or even ferrous materials is often troublesome. To elucidate a problem, the object in question can be weighed in various orientations on the balance pan to see whether the results are independent of such maneuvers. If not, the object can be suspended well below the pan of the balance in hopes that the unwanted interaction is between the object and a part of the balance itself.

That analytical balances are blind to the origin of the force that they sense can be turned to advantage. Thus, laboratory balances are commonly used in measuring phenomena such as surface tension and magnetic susceptibility. Since the pan of a servocontrolled balance is maintained at fixed position with respect to the laboratory, measurement of forces that vary strongly with position is convenient with these balances.

5.2.7
Microbalances

5.2.7.1 Analytical Type, Adapted to Microweighing

Although modern analytical balances can have standard deviations of less than 1 μg, microbalances have traditionally been used for different types of measurements than those required of normal analytical balances. Typically, microbalances have relatively small capacity but precisions of 1 μg or better. They will operate in vacuum or controlled atmospheres and are often used to follow the evolution in mass of a sample that is undergoing some physical or chemical change.

5.2.7.2 Vibrating Transducers

Some commercial microbalances measure inertial mass. Typically this is accomplished by relating the change in the frequency of a mechanical resonance to the change in mass of the resonant system. In special cases, it is possible to design transducers that measure inertial mass to picogram sensitivity. The most widely known example is the quartz-crystal microbalance (Ward and Buttry, 1990; Lu and Czanderna, 1984). In this case, the piezoelectric effect is exploited in order to excite the shear resonance in a plate of AT-cut quartz. The resonant frequency f_0, usually about 10 MHz, decreases by an amount f that is proportional to the mass per unit area m/A loading the free surface. The mass sensitivity of such a device is approximated by a relation first derived by Sauerbray (1959):

$$\Delta m/A = -4.42 \times 10^6 \Delta f/f_0^2 \qquad (5.10)$$

with $\Delta m/A$ in kilograms per meter square, for AT-cut quartz.

The field of MEMSs (micro-electromechanical systems) and NEMSs (nano-electromechanical systems) is burgeoning, and devices with attogram (10^{-18} g) precision have been reported.

5.2.8
Astronomical and Subatomic Mass

The product Gm_E, where m_E is the mass of the Earth, is approximately equal to gr^2, where r is the Earth's radius and G is the gravitational constant. Cavendish's measurement of G was popularly known as *weighing the Earth* in recognition of the role of G in relating planetary and other astronomical masses to laboratory standards. More generally, a small satellite in an elliptic orbit of semimajor axis a and angular frequency ω about a mass m will, in the appropriate limits, obey Newtonian mechanics:

$$m = \frac{a^3\omega^2}{G} \tag{5.11}$$

Thus, for example, the mass in kilograms of the Sun can be known. The currently recommended value of G (Mohr, Taylor, and Newell, 2008) is $6.674\,28(67) \times 10^{-11}\ \mathrm{m^3\ kg^{-1}\ s^{-2}}$.

It is difficult to accurately determine the mass of celestial objects having no observable satellites. Consider the asteroids, for example. It is sometimes possible to infer the mass of the largest asteroids from observations of gravitationally perturbed neighbors. Otherwise, even more indirect methods must be used: the volume of the asteroid is estimated from its brightness (assuming a value for albedo), and the mass is then estimated by assuming a value for the asteroid's density.

The mass of subatomic particles may be found by applying the conservation laws of energy and momentum. The acceleration in electromagnetic fields of either the particle itself or its decay products may be measured, as well as the energy and momentum of emitted photons. By the nature of these measurements, the mass is most directly reported in electron volts (eV), an energy unit that is outside the SI. To convert electronvolts to kilograms, it is necessary to multiply by a factor equal to the numerical value of the elementary charge, e, divided by c^2 (Mohr, Taylor, and Newell, 2008).

5.3
Density

The density ρ of an infinitesimal volume dV of material of mass dm is

$$\rho = \frac{dm}{dV} \tag{5.12}$$

In the SI, density is a derived quantity with units of kilograms per cubic meter. Depending on the application, it may be convenient to measure the density of the unknown relative to a known density (most usually that of water). Alternatively, one may measure directly the mass and volume of the unknown. The latter technique is, of course, the only method of determining a reference density directly in terms of SI units.

5.3.1
Measurements Based on Buoyancy

Buoyancy methods rely on Archimedes' principle, just as do the corrections for air buoyancy described in Section 5.2.4. Immersion of the sample in a liquid results in a buoyancy "correction" so large that the density of the sample may be thus determined with respect to the density of the immersion liquid.

The special case of neutral buoyancy forms the basis of the flotation methods discussed in Sections 5.3.1.2 and 5.3.1.3.

5.3.1.1 Hydrostatic Weighing

Two methods of hydrostatic weighing are shown schematically in Figure 5.4. Method 1 is the classic technique capable of relative uncertainty below 10^{-6} in favorable cases. Method 2 suffers from a greater minimum uncertainty but is more convenient when one is using a servocontrolled top-loading balance.

Let x_1 be the ratio of the balance reading when the sample is placed on the pan in air to the reading when immersed in the liquid bath; ρ_w and ρ_a the density of the water bath and the ambient air; and ρ the unknown density to be determined. It

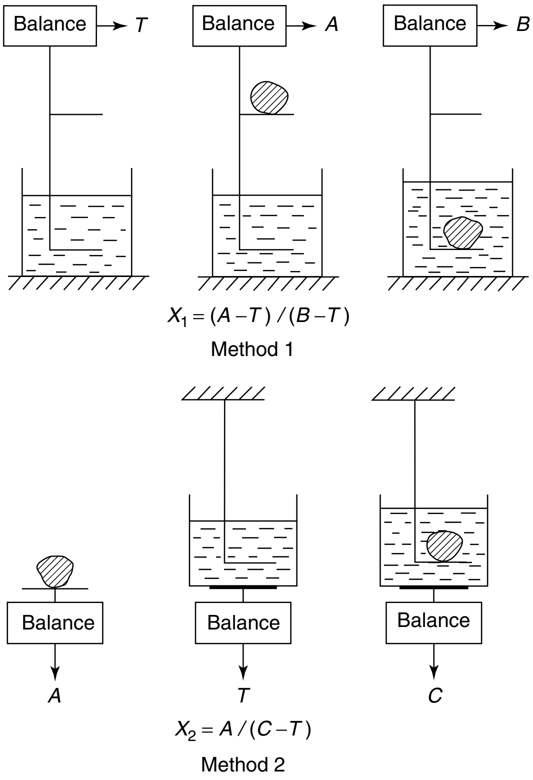

Fig. 5.4 Two methods for determining the density of a solid sample relative to an immersion liquid. The dimensionless quantities x_1 and x_2 are used in Equations (5.13) and (5.14).

is then easy to show that

$$\rho = \frac{x_1\rho_w - \rho_a}{x_1 - 1} \text{ (Method 1)} \qquad (5.13)$$

For Method 2, let x_2 be the ratio of the balance reading when the sample is placed on the pan in air to the difference in readings upon immersion. Then

$$\rho = x_2(\rho_w - \rho_a) + \rho_a \text{ (Method 2)} \qquad (5.14)$$

Note that x_1 and x_2 are dimensionless ratios so that the balances used need not be calibrated to any particular unit of mass (although their linearity must be assured). This feature is often cited as an advantage to using the same balance for both the air and water weighings. The two methods will only be sensitive if the balance is precise compared to $\rho_w V$. In addition to the obvious requirements of knowing ρ_w and ρ_a, achieving accurate results depends on careful attention to the suspension wire and prevention of gas bubbles adhering to the sample. The report of Bowman and Schoonover (1967) still serves as a useful introduction to these subjects and a paper by Fujii (2006) shows the present state of advancement. The method can also be used to determine the density of the surrounding liquid if the density of the solid is known. Techniques to measure the density of mass standards of various nominal values are described in OIML (2004a).

The suspension wire must be eliminated when determining the density of fluids far from ambient conditions (under high pressure, for instance). In such cases, the solid can take the form of a magnet that is servocontrolled to a position within the liquid by means of external coils (see the review by Wagner and Kleinrahm, 2004 and included references). Analogs of both Methods 1 and 2 shown above are practical.

It may happen that water is not suitable as an immersion liquid for a given sample. In this case, a different liquid is used and its density at the time of the measurements determined from published tables, either by performance of an auxiliary experiment or by means of a solid density reference standard (Bowman *et al.*, 1974). There are obvious advantages if the reference standard has approximately the same mass and density as the unknown.

5.3.1.2 Flotation Methods

As density is an intensive quantity, it should be possible to determine very small samples to high accuracy. Flotation methods accomplish this feat. When a sample of volume V whose density ρ is to be determined is immersed in a liquid of density ρ_L, the vertical force on it will be $(\rho - \rho_L)Vg$. The sample will, in general, rise or sink depending on whether this quantity is negative or positive. If the value of ρ_L can be adjusted so that the sample does neither (neutral buoyancy), then $\rho = \rho_L$ and the problem then becomes that of determining the density of a relatively large volume of liquid.

The method is most easily realized with transparent liquids. Since the sample and liquid generally have different thermal expansion and compressibility, equilibrium may be achieved through temperature or pressure changes as well as by adjusting the composition of a binary liquid.

When the surrounding liquid approaches the density of the sample, it becomes increasingly difficult to determine whether the sample will ultimately rise or fall. Greater sensitivity can be achieved by the use of a centrifuge rather than gravity to produce the forces. Alternatively, one can create a uniform density gradient in a column of liquid. Samples introduced in the column will, unattended, eventually come to rest at their position of neutral buoyancy.

Flotation methods are useful in quantifying small differences in density between two objects of the same material. Objects of different sizes and shapes can be compared in this way (Bettin and Toth, 2006).

5.3.1.3 Hydrometry

Hydrometers have been used as a relatively convenient method to find the density of liquids, especially when only small differences among samples are expected. The hydrometer consists of a graduated stem with a weighted volume at the bottom. The hydrometer floats in the test liquid, with the base of its stem in the liquid and the top of the stem in the air. Depending on the desired measurand, the stem is calibrated in units of density, relative density, or a specialized quantity related to density.

The review paper by Lorefice and Malengo (2006) describes how these instruments are calibrated and includes an extensive bibliography.

5.3.2
Measurements Based on Mass and Volume

If a sample of unknown density ρ has the same volume as a known standard of density ρ_s, then the ratio of masses m/m_s of the unknown and standard must equal ρ/ρ_s, the ratio of their densities. Although the volume of the unknown and the standard must be the same, it need not be known. Thus, mass determinations constitute the only precise measurements needed. Special containers, known as *pycnometers (or pyknometers)*, are designed to hold a reproducible volume of liquid.

5.3.2.1 **Liquid Pycnometers**

A variety of liquid pycnometers have been designed over the years and marketed by laboratory supply companies. Although all these have design advantages specific to given applications, it may be sufficient to use a simple hypodermic syringe as a volume standard. In this case, the syringe is first weighed empty. If V_i is the internal volume of the syringe, V_e its external volume, m its mass, ρ_a the density of ambient air, and y_1 the balance reading, then

$$y_1 = \rho_a(V_i - V_e) + m \tag{5.15}$$

The syringe is then filled with the standard liquid of density ρ_L, usually water, and weighed again:

$$y_2 = \rho_L V_i - p_a V_e + m \tag{5.16}$$

Finally, it is filled with the sample of unknown density ρ and weighed for the third time:

$$y_3 = \rho V_i - \rho_a V_e + m \tag{5.17}$$

The desired density is then determined from the known densities of the liquid standard and the ambient air:

$$\rho = (\rho_L - \rho_a)\frac{y_3 - y_1}{y_2 - y_1} + \rho_a \tag{5.18}$$

This example suggests many of the potential problems of pycnometry: the interior volume must be reproducible to the needed precision. The chamber must be carefully cleaned between fillings. The pycnometer itself must be a suitable object for weighing. If all measurements cannot be carried out at the same temperature, then the thermal expansion of the pycnometer must be known. Liquids must be introduced without entraining air bubbles (ISO, 1999).

Some commercially available pycnometers are designed for density measurements of solids, often in the form of grains or powders. In this case, the first observation is the mass of the solid sample itself. The third observation is the mass of the pycnometer with its interior volume containing the solid sample plus enough of the standard liquid to fill the interior volume completely. The density of the solid is then determined. This technique will only work well if atmospheric gases occluded by the solid can be eliminated.

Pycnometers were traditionally designed for use with a chemist's analytical balance with a capacity of about 200 g. Thus, their interior volume is typically 25–100 ml.

The so-called gas pycnometers are available commercially for the determination of solid volumes. They are based on the fact that, at constant temperature, the pressure of a fixed number of atoms of a noble gas varies inversely with the volume to which the gas is confined. In a gas pycnometer, introduction of a sample decreases the original volume by that of the sample. The measurements can be automated to an extent. The method is especially suited to the determination of powders or samples with open pores, since noble gases penetrate better than liquids.

5.3.2.2 **Vibrating-Tube Densimeters**

Pycnometry, as developed by Gay-Lussac and refined by later investigators, has been superseded in many laboratories by the vibrating-tube densimeter. First developed by Kratky, Leopold, and Stabinger

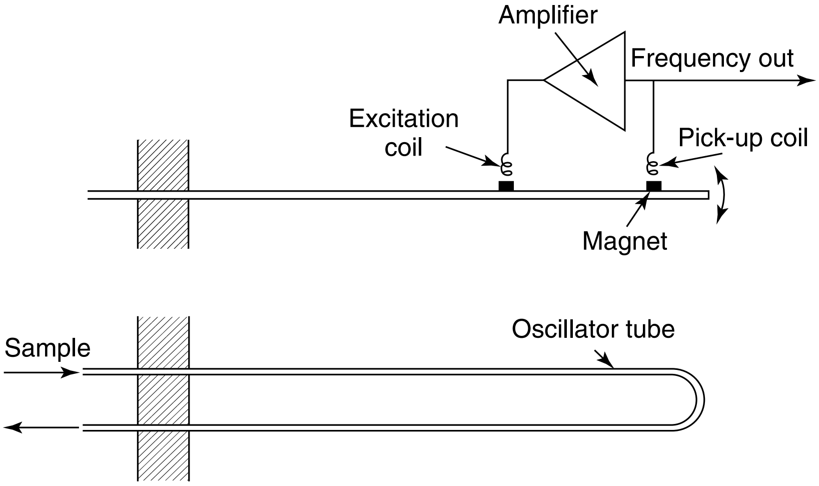

Fig. 5.5 Two views showing the principle of a vibrating-tube densimeter. The natural frequency of the hollow oscillator tube will depend on the density of the sample that fills it. Based on a sketch provided by H. Stabinger, used with permission.

(1969), this device differs from the traditional liquid pycnometer in only one fundamental respect, namely, that the inertial mass of the material within the volume chamber is sensed rather than the gravitational mass. This one conceptual change leads to many experimental advantages.

To realize the device, the liquid container takes the form of a hollow tube bent into the form of a "U" and anchored at each end (Figure 5.5). Like a tuning fork, the tube will have a mechanical resonance whose frequency depends on elastic properties, geometry, and mass. In particular, the natural period τ of the tube is proportional to $m^{1/2}$, the square root of its mass. The latter quantity is equal to the mass of the tube itself plus ρV_e, the mass of any fluid filling the vibrating region of the tube. In practice, V_e is about 1 ml, that is, much smaller than the volume of typical pycnometers. On the basis of this simple model,

$$\rho = A\tau^2 - B \tag{5.19}$$

The parameters A and B are constants, and thus the device must be calibrated at a minimum of two points over its range. Dry nitrogen and water are frequently used as standards.

The highest relative precision obtainable with such devices is of order 1×10^{-6}. Deviations from Equation (8.19) are noticeable in highly precise data, and so more careful calibration is required to achieve comparable accuracy. Devices based on the same principle of operation have been constructed for temperatures to 400°C and pressures to 40 MPa. Practical considerations are given in the ISO (1998) standard.

5.3.3
Miscellaneous Techniques

The techniques discussed above are those most generally used for density determinations. Nevertheless, myriad other methods have been developed to meet special needs.

Where volume or density appears in a physical equation, it has probably formed the basis for a determinative method. Numerous examples of traditional methods are reviewed by Davis and Koch (1992).

5.3.4
Reference Standards

Distilled or deionized tap water has a density that is known to a relative uncertainty of about 1×10^{-5} at 0.1 MPa and between 0 and 40°C and is the single most useful reference standard. Uncertainty may be improved by about a factor of 3 by correcting for or eliminating atmospheric gases dissolved in the water and by taking account of the isotopic variability of individual water samples.

A particular isotopic mix of water is designated Vienna Standard Mean Ocean Water (V-SMOW), small samples of which can be obtained from the International Atomic Energy Commission in Vienna for use in determining the relative isotopic abundances of other water samples of interest. A table of water of density suitable for typical metrological needs is given by Tanaka *et al.* (2001). A much more general formulation of the thermodynamic properties of V-SMOW has been published by Wagner and Pruß (2002).

Good use is sometimes made of the fact that the density of water has a maximum near 4°C. Problems that might be caused by convection are greatly reduced near this temperature since the density is insensitive to small changes in temperature.

The density of selected mercury samples was determined with respect to standards of mass and length by Cook (1961), who achieved a relative uncertainty of about 1×10^{-6} at 0.1 MPa and 20°C. In the succeeding years, these samples have played a key role in precision manometry, gas thermometry, and measurements of fundamental physical constants.

Glossary

Apparent Mass: The reading, uncorrected for air buoyancy, of a perfectly calibrated balance. The ratio of mass to apparent mass is a function of the ambient air density, the density of the object in question, and the density of the standards used to calibrate the balance (Section 5.2.4).

Atomic Mass Unit (Dalton): The unit mass, u, in a system where the mass of an atom of carbon-12 is taken to be exactly 12 u. One mole of carbon-12 atoms has a mass of exactly 0.012 kg.

Avogadro Constant: The number of entities in 1 mol of substance and hence the number of atomic mass units in 0.001 kg.

Balance: Originally, an ancient lever device used to demonstrate that two objects have the same (gravitational) mass. The term is now used more generally to designate almost any transducer configured to make mass measurements.

Buoyancy Correction: The correction applied to balance readings in order to eliminate the component due to air buoyancy.

Cook's Mercury (Informal): Mercury stock whose density was measured by A. H. Cook. The stock has been used to advantage in several areas of precision measurement.

Conventional Mass: At a reference temperature of 20°C, the mass of material of density 8000 kg m^{-3} that will balance an object of interest, usually a secondary standard of mass, in air of density 1.2 kg m^{-3} (Section 5.2.4).

Dalton: See atomic mass unit.

Effective Mass: The mass of an object less the product of its volume and the density of the ambient fluid, usually air (Section 5.2.4).

Gas Pycnometer: A device used to find the volume of porous solids and powders through application of the gas laws.

Gradient Column: A column of liquid whose density decreases uniformly from the bottom to the top. The densities of a sample or set of samples can be determined from their equilibrium heights in the column.

Gravitational Mass: The mass that appears in Newtonian equations for gravitational force.

Hydrostatic Weighing: A traditional method of measuring the density of a solid by weighing it both in air and submerged in a liquid, usually water (Section 5.3.1.1).

Inertial Mass: The mass that appears in Newton's second law and in special relativity.

International Prototype of the Kilogram: The artifact that defines the unit of mass in the SI.

Maximum Permissible Error: The maximum allowed absolute difference between the conventional mass of a weight (2) and its nominal mass.

Microbalance: A balance of microgram or better resolution.

Neutral Buoyancy: The condition in which an object at equilibrium neither rises nor sinks in a surrounding fluid. It is assumed that the system is subject to a uniform acceleration, usually gravitational but sometimes centripetal.

Nominal Mass: The intended mass of a standard weight (2).

Pycnometer (Pyknometer): A vessel with precisely reproducible interior volume, traditionally used to determine the relative densities of liquids.

Quartz-Crystal Microbalance: A device that measures mass per unit area in terms of the oscillation frequency of a quartz crystal (Section 5.2.4).

Relative Density: A dimensionless quantity representing the density ratio of a given sample to a standard of reference. Unless otherwise stated, the standard is assumed to be distilled water at 20°C. Relative density is the modern name for specific gravity.

True Mass: See vacuum mass.

Vacuum Mass: Another term for mass. The term vacuum mass or true mass is sometimes used to distinguish mass from conventional mass in reports of calibration.

Vibrating-Tube Densimeter: A device that measures the relative densities of fluids by a mechanical resonance technique (Section 5.3.1.2).

Vienna Standard Mean Ocean Water (V-SMOW): Water having a certain, specified isotopic composition. Distilled tap water generally has a density about 0.003 kg m^{-3} less than V-SMOW.

Weak Equivalence Principle: A basic principle of physics whose consequence for mass measurements is as follows: equivalent results are obtained independent of whether the measuring device responds to the inertial or the gravitational properties of mass.

Weighing: The process of determining mass by means of a balance.

Weight (1): The product of a mass and the local gravitational acceleration. The term *standard weight* is used to signify weight at a conventional acceleration, usually 9.80665 m s^{-2}.

Weight (2): An artifact manufactured to embody a specified nominal mass.

References

Audi, G., Wapstra, A.H., and Thibault, C. (**2003**) *Nucl. Phys. A*, **729**, 129–336, 337–676.

Bettin, H. and Toth, H. (**2006**) *Meas. Sci. Technol.*, **17**, 2567–2573.

BIPM (**2006**) *The International System of Units (SI)*, 8th edn, Bureau International des Poids et Mesures, Sèvres, *http://www.bipm.org/en/si/*.

Bowman, H.A. and Schoonover, R.M. (**1967**) *J. Res. Natl. Bur. Stand.*, **71C**, 179–198.

Bray, A., Barbato, G., and Levi, R. (**1990**) *Theory and Practice of Force Measurement*, Academic, London.

Cook, A.H. (**1961**) *Philos. Trans. R. Soc. Lond. A*, **254**, 125–154.

Davis, R.S. and Koch, W.F. (**1992**) Chapter 1, in *Physical Methods of Chemistry*, 2nd edn, vol. **6** (B.W. Rossiter and R.C. Baetzold), John Wiley & Sons, Inc., New York.

Fujii, K. (**2006**) *Meas. Sci. Technol.*, **17**, 2551–2559.

Girard, G. (**1994**) *Metrologia*, **31**, 317–336.

Gläser, M. (**1999**) *Metrologia*, **36**, 183–197.

http://physics.nist.gov/cuu/Constants/RevModPhys_80_000633acc.pdf.

ISO (**1998**) *Oscillation-type density meters – Part 1*, 15212-1:1998, International Organization for Standardization.

ISO (**1999**) *Laboratory glassware – Pyknometers*, 3507:1999, International Organization for Standardization.

Kratky, O., Leopold, H., and Stabinger, H. (**1969**) *Z. Angew. Phys.*, **4**, 273–277.

Lorefice, S. and Malengo, A. (**2006**) *Meas. Sci. Technol.*, 2560–2566.

Lu, C. and Czanderna, A.W. (eds) (**1984**) *Applications of Piezoelectric Quartz Crystal Microbalances*, Elsevier, Amsterdam.

Mills, I.M., Mohr, P.J., Quinn, T.J., Taylor, B.N., and Williams, E.R. (**2006**) *Metrologia*, **43**, 227–246.

Misner, C.A., Thorne, K.S., and Wheeler, J.A. (**1973**) *Gravitation*, Freeman, San Francisco.

Mohr, P.J., Taylor, B.N., and Newell, D.B. (**2008**) *Rev. Mod. Phys.*, **80**, 633–730.

OIML (**2004a**) International Recommendation R111-1 (E): International Organization of Legal Metrology. *http://www.oiml.org/publications/*.

OIML (**2004b**) International Document D28 (E): International Organization of Legal Metrology. *http://www.oiml.org/publications/*.

Picard, A., Davis, R.S., Gläser, M., and Fujii, K. (**2008**) *Metrologia*, **45**, 149–155.

Prowse, D.B. and Anderson, A.R. (**1974**) *Metrologia*, **10**, 123–128.

Quinn, T.J. (**1991**) *IEEE Trans. Instrum. Meas.*, **40**, 81–85.

Quinn, T.J. (**1992**) *Meas. Sci. Technol.*, **3**, 141–159.

Sauerbray, G. (**1959**) *Z. Phys.*, **155**, 206–222.

Schoonover, R.M. (**1982**) *Anal. Chem.*, **54**, 973A–980A.

Speake, C.C. (**1987**) *Proc. R. Soc. Lond. A*, **414A**, 333–358.

Tanaka, M., Girard, G., Davis, R., Peuto, A., and Bignell, N. (**2001**) *Metrologia*, **38**, 301–309.

Wagner, W. and Kleinrahm, R. (**2004**) *Metrologia*, **41**, S24–S39.

Wagner, W. and Pruß, A. (**2002**) *J. Phys. Chem. Ref. Data*, **31**, 387–535.

Ward, M.D. and Buttry, D.A. (**1990**) *Science*, **249**, 1000–1007.

Further Reading

(a) Kochsiek, M. and Gläser, M. (eds) (**2000**) *Recent Monographs on Mass Metrology are: Comprehensive Mass Metrology*, Wiley-VCH Verlag GmbH; (b) Jones, F.E. and Schoonover, R.M. (eds) (**2002**) *Handbook of Mass Measurement*, CRC Press.

Experimental progress in linking the SI unit of mass to physical constants are reported at the Conference on Precision Electromagnetic Measurements, held biannually. Proceedings are published in a special issue of the IEEE Transactions on Instrumentation and Measurement.

A brief introduction to mass measurements below the microgram level can be found in the chapter by Davis, R.S. (**2007**) in *Metrology and Fundamental Constants* (eds T.W. Hänsch, S. Leschiutta, and A.J. Wallard), IOS Press, pp. 473–497.

MATERIALS DEGRADATION

R. G. KELLY, *Department of Materials Science, University of Virginia, Charlottesville, Virginia, U.S.A.*

J. A. BARDWELL, *Institute of Microstructural Sciences, National Research Council, Ottawa, Ontario, Canada*

Introduction 349
1. **Physical and Technical Principles** 350
1.1 Thermodynamic Driving Force 350
1.2 Kinetics 352
1.3 Types of Degradation 353
1.4 Environmental Variables 355
1.5 Prevention 355
2. **Methods and Equipment** 356
2.1 Measurement of Thermodynamics 356
2.2 Measurement of Kinetics 357
2.3 Measurement of the Loss of Mechanical Properties 360
3. **Specific Examples for Different Classes of Materials** 360
3.1 Metals and Alloys 360
3.2 Polymers 361
3.3 Ceramics 363
Glossary 363
Works Cited 363
Further Reading 364

INTRODUCTION

The degradation of materials, more commonly referred to as corrosion, has a huge impact on modern society. Studies have shown that corrosion annually costs a nation on the order of (4–5)% of its gross national product (Payer *et al.*, 1980). Thus, for small nations, the absolute cost is small, while for industrialized nations, the cost can be staggering. For example, the U.S. gross national product for 1988 was approximately $5 trillion. Thus, the cost of corrosion to U.S. society was approximately $200 billion, or of the same order as the annual U.S. federal debt. The replacement, maintenance, and repair of chemical processing equipment, autos, ships, aircraft, and other facilties because of corrosion constitute a large fraction of this cost. While the costs of corrosion can be large, many of them are preventable. It has been estimated that $\frac{1}{3}$ of the costs of corrosion could be prevented by the proper application of present-day corrosion control measures.

Corrosion manifests itself in many ways, from the commonplace to the tragic. The replacement of automobile mufflers, while irritating to the auto owner, is usually only a minor inconvenience. However, the collapse of the Pt. Pleasant (KY) bridge in 1967 due to corrosion-related cracking of a support cost the lives of over 30 people. Corrosion diminishes the usable lifetime of components and also limits the advancements of technology. For example, in the case of advanced jet engines, the limiting technology is that of materials capable of prolonged exposure to high-temperature exhaust gases without the loss of mechanical integrity.

While the formal study of corrosion has been ongoing for over 200 years, there are still many unanswered questions. In fact, with each new material that is produced, new manifestations of degradation are observed. For example, stainless steels were hailed in the 1930s as the solution to corrosion problems. While these alloys of iron with more than 11% Cr are indeed "stainless" upon exposure to atmospheric conditions, they can undergo highly localized, high-rate attack in fairly mild environments such as seawater. Alloying can have a major effect on this sensitivity. The addition of as little as 1 at. % molybdenum increases the corrosion resistance substantially, by an as yet unknown mechanism. As another example, the new high-T_c superconductors are prone to degradation under atmospheric conditions, result-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ing in the loss of their superconducting properties. With the drive to higher component densities in the electronics industry, amounts of corrosion previously considered completely insignificant are now limiting. For example, the loss of 1 pg (10^{-12} g) of Al is sufficient to cause an interconnect failure for micron technology. These and many other applications provide a strong technological driving force for understanding corrosion processes.

In addition to the immense technological importance of materials degradation processes, they also are of great scientific interest. In part, this is because of the inherent complexity of the subject. Most degradation occurs at the interfaces between a material and its environment. It has been said that interfaces are the invention of the devil. While it is true that interfacial processes are much more complex than bulk, their satanic origin has never been verified. However, understanding interfacial behavior is critical to developing a complete view of physical processes. The interaction of a large number of variables also makes corrosion a fascinating field of scientific study. Identifying the controlling factors, and then designing experiments in which they are isolated, is a constant challenge. Theoretically, the symmetry-breaking effect of the interface provides a fruitful field for study.

At the same time, reactions that result in materials degradation provide us with vehicles to understand the fundamental nature of solid-liquid, solid-solid, and solid-gas interfaces. While our understanding of bulk effects is far from complete, the areas of surface chemistry and surface physics are two of the most challenging and fascinating research areas today. New tools are constantly being applied that provide new insights into the origins of the reactions. Because of the coupled nature of many of the phenomena, these processes are a rich area of dynamic behavior, including chaotic phenomena (Hudson and Bassett, 1991).

This article provides an overview of the phenomena involved in the degradation of materials. Because of space limitations and the extreme breadth of the field, only brief descriptions of the many important processes can be given here. However, the interested reader can find a wealth of in-depth information in the references given under Further Reading. In the first section, the physical principles of the degradation process will be presented with an emphasis on the basic thermodynamics and kinetics, the different types of degradation, and the effects of environmental variables on the processes. The second section will introduce the methods by which the relevant thermodynamic and kinetic parameters are measured. The final section consists of examples of material degradation for metals, polymers, and ceramics.

1. PHYSICAL AND TECHNICAL PRINCIPLES

1.1 Thermodynamic Driving Force

Thermodynamics dictates that materials have the tendency to reach their lowest possible free energy. A particular chemical reaction is thermodynamically allowed if the change in free energy, ΔG, for the reaction is negative. Three types of information are available in the literature in order to predict the thermodynamic stability of materials. First, the values of ΔG for possible chemical reactions can be examined directly. Some examples of the free energy of formation of common degradation products are given in Table 1; many more can be found in the literature (Goldsmith *et al.*, 1961). For most metals and alloys, the oxide, hydroxide, or sulfide is thermodynamically favored over the pure metal, as expressed by a negative free energy of formation of these corrosion products. This means that given the availability of oxygen, water, sulfides, etc. in the environment, most metals are not thermodynamically stable and should be entirely converted into the corresponding reaction products.

The free energy is $\Delta G = \Delta H - T\Delta S$, where ΔH is the enthalpy change and ΔS is the en-

Table 1. Standard free energy of formation of some degradation products.

Compound	ΔG (kJ mol^{-1})	Compound	ΔG (kJ mol^{-1})
Al_2O_3	−1576	$Pt(OH)_2$	−285
Cr_2O_3	−1047	NiO	−216
SiO_2	−805	CuO	−127
Fe_2O_3	−741	Ag_2S	−40.3
$ZnCO_3$	−731	Au_2O_3	163

tropy change in the reaction. The value of ΔS may be thought of as the change in randomness of the system. Thus for strongly ordered systems, such as fiber-reinforced composites and some kinds of polymers, a degradation of the material would result in a decrease in free energy and thus would be expected on thermodynamic grounds.

A second method of predicting thermodynamic stability is to examine the electromotive series. When a metal is immersed in a solution, a potential difference, called the single-electrode potential or half-cell potential, is developed at the metal–solution interface. In order to measure this potential, it is necessary to include another electrode in the circuit. This means that the potential cannot be measured on an absolute scale but must be measured relative to a reference electrode. The standard reference electrode is the hydrogen electrode, assigned a potential of zero. An abbreviated list of half-cell potentials is given in Table 2; very extensive lists are given in the literature (Parsons, 1959; Weast, 1991). Many degradation reactions are electrochemical in nature; that is, they consist of two coupled electrochemical half reactions. If a given degradation reaction has been divided into its two half reactions, the reaction with the higher value of E^0 would be expected to proceed in the reverse direction. For example, if an iron pipe is connected to a copper wire, the electromotive series can be used to predict that a potential difference will be established between the two metals, and current will flow, resulting in corrosion of the Fe according to the overall reaction

$$Fe + Cu^{2+} \rightarrow Fe^{2+} + Cu. \qquad (1)$$

Similarly, the corrosion reaction

$$Fe + \tfrac{1}{2}O_2 + H_2O \rightarrow Fe(OH)_2 \qquad (2)$$

can be separated into the two half reactions

$$\tfrac{1}{2}O_2 + H_2O + 2e^- \rightarrow 2OH^-, \qquad (3)$$

$$Fe \rightarrow Fe^{2+} + 2e^-, \qquad (4)$$

followed by the chemical reaction

$$Fe^{2+} + 2OH^- \rightarrow Fe(OH)_2, \qquad (5)$$

and would be expected to proceed in the direction written on the basis of the electromotive series.

Table 2. Some standard electrode reduction potentials.

Half-cell reaction	E^0 (V vs SHE)
$Li^+ + e^- = Li$	−3.045
$Al_3^+ + 3e^- = Al$	−1.706
$SiO_2 + 4H^+ + 4e^- = Si + H_2O$	−0.84
$2H_2O + 2e^- = H_2 + 2OH^-$	−0.8277
$Zn^{2+} + 2e^- = Zn$	−0.7628
$Cr^{3+} + 3e^- = Cr$	−0.74
$S + 2e^- = S^{2-}$	−0.508
$Fe^{2+} + 2e^- = Fe$	−0.409
$Fe^{3+} + 3e^- = Fe$	−0.036
$2H^+ + 2e^- = H_2$	0
$Cu^{2+} + 2e^- = Cu$	0.3402
$O_2 + 2H_2O + 4e^- = 4OH^-$	0.401
$Fe^{3+} + e^- = Fe^{2+}$	0.770
$Au^+ + e^- = Au$	1.68

Half reactions with more negative E^0 are referred to as "active" or "anodic" with respect to half reactions with a more positive value of E^0, which are termed "noble" or "cathodic." The value of E^0 is related to ΔG through the equation

$$\Delta G = -nFE^0, \qquad (6)$$

where F is the Faraday.

Thermodynamic information is also available in a Pourbaix diagram (Pourbaix, 1974). Pourbaix diagrams are a special type of phase diagram, where the potential is plotted on the vertical axis and the pH on the horizontal. They are constructed from consideration of chemical reaction equilibria, e.g.

$$Ni^{2+} + H_2O = NiO + 2H^+,$$
$$\log[Ni^{2+}] = 12.18 - 2pH, \qquad (7)$$

and electrochemical reaction equilibria, e.g.

$$Ni = Ni^{2+} + 2e^-,$$
$$E = -0.250 + 0.0295 \log[Ni^{2+}]. \qquad (8)$$

The diagrams can be used to predict the stability of elements for a large variety of conditions. Figure 1 illustrates the example of Ni in water. The two dashed lines show the equilibrium conditions for oxygen evolution and hydrogen evolution from water, as indicated on the diagram. Thus, water is ther-

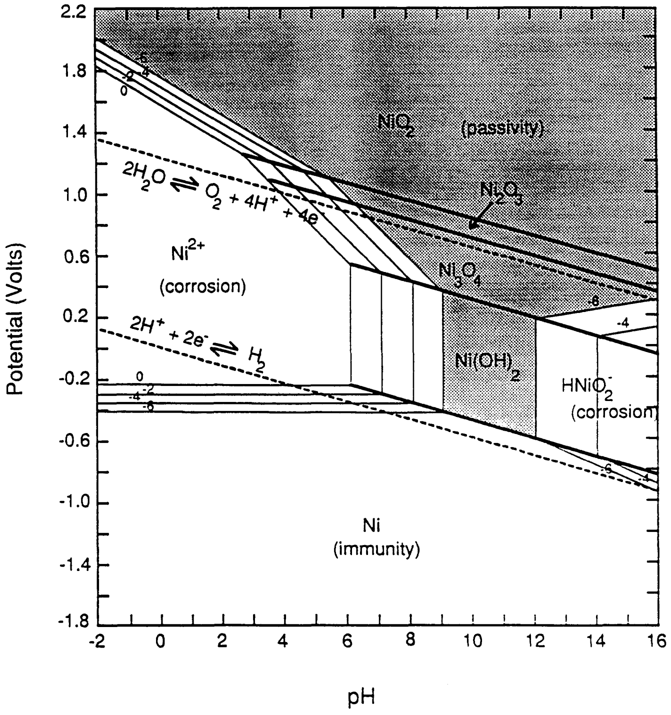

FIG. 1. A Pourbaix diagram for the system water/nickel. The potential is given relative to the hydrogen electrode. Regions of stability of various materials are shown. A thick solid line shows the equilibrium between two solid species. A thin solid line shows the equilibrium between a solid and a dissolved species, with the logarithm of the concentration of the dissolved species shown as a small number. The dashed lines indicate the equilibria for the oxygen evolution and hydrogen evolution from water. The shaded area indicates the region of passivity for nickel for a solution concentration of 10^{-6} M.

modynamically stable in the region between the two dashed lines. The thick lines indicate equilibria between two solid species; the narrow lines indicate equilibria between one dissolved and one solid species, with the small numbers giving the logarithm of the concentration of the dissolved species corresponding to each line. Thus, the diagram may be divided into regions of corrosion (dissolved species thermodynamically stable), immunity (metal thermodynamically stable), and passivity (shaded area, solid species other than metal thermodynamically stable). Thus, it may be seen that Ni would be expected to dissolve at low and high *p*H's, and exhibit passivity at intermediate *p*H values in the region where water is thermodynamically stable.

1.2 Kinetics

Although thermodynamics is very useful in predicting whether or not a given reaction is possible, it says nothing about the kinetics of a degradation reaction. For example, one would predict from Table 1 that chromium would be much more unstable than iron. However, under most circumstances, chromium is much less reactive than iron because of the kinetics of the dissolution of the corrosion products. From an engineering standpoint, the kinetics of degradation reactions are fundamentally important.

Since degradation is normally an environmentally induced process, the kinetics of any degradation process are critically dependent on the exact environmental conditions. For many systems, no general rule can be given for predicting degradation rates; however, for metals in aqueous environments, the kinetics are well understood, and thus this system will be discussed in some detail later.

As discussed above, many degradation reactions are electrochemical in nature; that is, they can be divided into two half reactions, each of which involves the transfer of electrons. In the absence of an externally applied potential, the currents for each half reaction are the same, and equal the so-called corrosion current density, which is proportional to the corrosion rate. The open-circuit corrosion potential is determined by the intersection of the lines describing the rates of the two reactions. Kinetically, the two half reactions can be examined individually by the application of an external potential (or alternatively by application of an externally applied current). This allows the construction of a mixed potential diagram, known as an *Evans diagram* for the case of corrosion reactions, as shown in Fig. 2. For a single electrode, the theory of electrode kinetics predicts that (Bockris and Drazic, 1972)

$$\eta = b \log(i/i_0), \tag{9}$$

where η is the overpotential, i.e., the difference between the reversible potential and the actual potential; b is the Tafel slope; i is the measured (or applied) current density; and i_0 is the exchange current density for the reaction. The Tafel slope is the slope of the straight-line portion of a polarization curve, usually occurring at more than 50 mV from the open-circuit potential, when presented in a semilogarithmic plot in terms of voltage per logarithmic cycle of current density (commonly referred to as volts per decade). This equation is analogous to the Arrhenius equation

$$\text{rate} = A \exp(-E_a/RT) \tag{10}$$

where A is the pre-exponential factor and E_a is the activation energy. As in the kinetics of thermally activated processes, electrochemical kinetics are plotted on semilogarithmic axes.

Thus, for $\eta = 0$, the potential is the reversible potential, and the extrapolated current is equal to the exchange current for the reaction. For each of the half reactions of interest, this point is indicated on Fig. 2. The open-circuit potential and current for the corroding system are determined by the intersection of the two lines of the half reactions (the slope of each being the Tafel slope for that half reaction). Thus, the corrosion rate is established by the individual kinetics of the two half reactions.

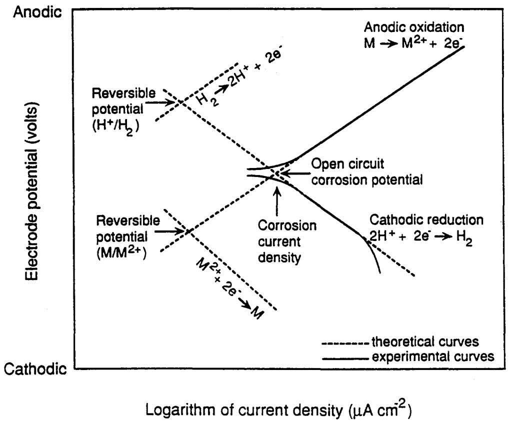
Anodic
Anodic oxidation $M \rightarrow M^{2+} + 2e^-$
$H_2 \rightarrow 2H^+ + 2e^-$
Reversible potential (H^+/H_2)
Open circuit corrosion potential
Reversible potential (M/M^{2+})
Corrosion current density
Cathodic reduction $2H^+ + 2e^- \rightarrow H_2$
$M^{2+} + 2e^- \rightarrow M$
theoretical curves
experimental curves
Cathodic
Electrode potential (volts)
Logarithm of current density (µA cm^{-2})

FIG. 2. A diagram showing the kinetic equations governing an electrochemical corrosion reaction for the metal M.

The diagram shown in Fig. 2 is directly applicable to the case of general corrosion, where the entire metal surface can be considered as the anode for the metal oxidation reaction and the cathode for the hydrogen evolution reaction. In this case, it is appropriate to plot the logarithm of the current density on the horizontal axis, and the diagram may be used to predict the degradation rate. In the case of localized corrosion, the anode and cathode will be of different sizes. In this case, the total current should be plotted. In addition, the relevant conditions will not be those in the bulk solution, but those within the confines of the local area of interest.

In conclusion, Evans diagrams are useful for predicting the kinetics of degradation reactions under a wide variety of conditions (Narayan, 1983).

1.3 Types of Degradation

Degradation may, for convenience, be divided into eight types, illustrated schematically in Fig. 3. While the mechanisms can be very different, the classification is, in general, based on the appearance of the attack. The types are as follows:

1. General, or uniform, attack. This is a very common form of corrosion and can result from exposure of materials to the atmosphere or to solutions which are aggressive toward the material. This form of degradation can be relatively benign since its rate can often be predicted. Examples are rusting of iron and steel, and tarnishing of brass, silver, and aluminum. Often the oxidation products of the degradation reaction must be removed to ensure that the underlying surface is generally corroded (Parsons, 1982; Narayan, 1983).
2. Pitting attack. The material remains passive and stable over most of the surface but at other points is rapidly degraded,

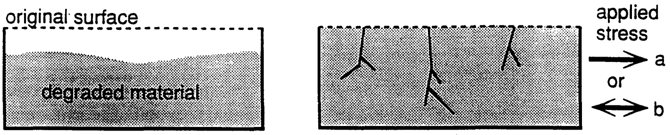

1. General attack

5. a. Stress corrosion cracking
b. Corrosion fatigue

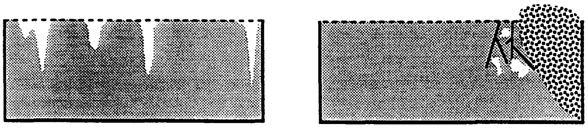

2. Pitting attack

6. Intergranular corrosion (at a weld)

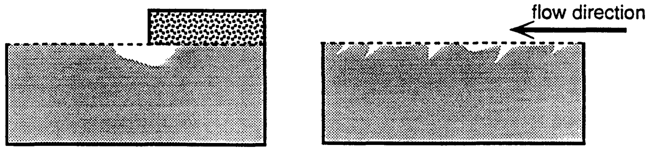

3. Crevice corrosion

7. Corrosion - erosion

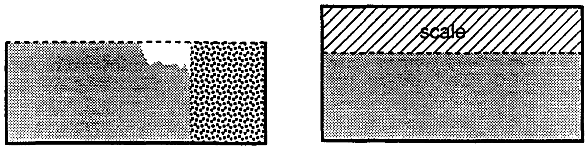

4. Galvanic corrosion

8. High temperature attack

FIG. 3. Schematic pictures of various types of degradation phenomena.

resulting in a very dangerous form of attack. Visually, it may be difficult to identify pitting corrosion until complete perforation of the material has occurred (Szlarska-Smialowska, 1986).

3. Crevice corrosion. When confined spaces or crevices are present in a structure, localized corrosion may occur in the occluded area. The crevices may be caused by geometric arrangements or merely by surface deposits of corrosion products (Butler and Ison, 1966).
4. Galvanic corrosion. Galvanic corrosion occurs when there is electrical coupling of two dissimilar metals that are placed in an electrolyte, resulting in corrosion of the less noble metal. The rate of galvanic corrosion is directly proportional to the ratio of the exposed areas of the cathodic and anodic metals (Narayan, 1983).
5. Stress corrosion cracking and corrosion fatigue. Stress corrosion cracking or corrosion fatigue results from the combined action of a tensile or periodic stress, respectively, and a corrosive environment on a material. For both types of degradation, the rate is accelerated over the case where either the aggressive environment or the stress is present alone. Unless a complete failure of the material has occurred by fracture, it can be difficult to identify these dangerous forms of degradation without microscopic examination (Sedricks, 1979; Narayan, 1983).
6. Intergranular corrosion. The grain boundaries of an alloy are usually physically and chemically different from the interior of the grain. Thus, they may be more susceptible to attack, resulting in a deep and rapid penetration of the metal and subsequent failure. Often welding or improper heat treatment of the metal can re-

sult in intergranular attack (Sedricks, 1979; Uhlig, 1948).

7. Erosion-corrosion. An increase in the corrosion rate may result from the relative motion of a metal surface and the aggressive environment surrounding it. Such attack is usually characterized by bright surfaces, free from corrosion products, but with distinct, localized pits or grooves oriented parallel with respect to the direction of solution flow (Butler and Ison, 1966). Often this attack results from the constant stripping of a protective film.
8. High-temperature attack. Since most metals and alloys react with air at ambient temperatures, it is expected that the rate of the reaction will increase, sometimes to a dangerous level, at elevated temperatures. The surface becomes covered with a scale, which may consist of oxides, sulfides, nitrides, etc. The damage caused by the attack depends on whether the scale is protective or nonprotective, and may often be predicted by the Pilling-Bedworth ratio, which is defined as the volume of scale formed from a unit volume of metal (Benard, 1964).

1.4 Environmental Variables

Some environments are more corrosive than others. High levels of acidity (or basicity in some cases), large concentrations of aggressive anions, such as the halides, and high levels of dissolved oxygen are conducive to degradation processes. However, since the different types of degradation are influenced in different ways by the environment, this discussion will follow the general types of degradation as shown in Fig. 3.

1. General corrosion is a result of exposure of metals to aggressive solutions or to the atmosphere. Thus, the rate will be determined by the aggressiveness of the solution in the former case and by the composition of the ambient atmosphere in the latter case. The variables include temperature, rainfall, wind conditions, and the concentration of corrosive pollutants such as ammonia, hydrogen sulfide, sulfur dioxide, etc.
2. Pitting, which results from a local disruption of the passive film, is usually caused by appreciable concentrations of chloride, bromide, or thiosulfate ions.
3. Crevice corrosion is most detrimental when the crevice is wide enough to permit the entry of the corrosive fluid but sufficiently narrow to maintain a stagnant zone. It may be attributed to lack of oxygen, change in acidity, buildup of a corrosive species, or depletion of a beneficial species within the crevice.
4. Galvanic corrosion. Environmental factors usually do not influence the reaction rate to any large extent, provided that the electrolyte has sufficient electrical conductivity.
5. Stress-corrosion cracking (SCC) occurs in specific environments, which depend on the metal alloy involved. However, it is difficult to predict SCC for a specific system without experimental data. Corrosion fatigue depends on the composition, temperature, *p*H, and oxygen content of the environment.
6. Intergranular corrosion can occur in most metal alloys and is dependent on the composition of the solution. However, the most important factors determining the susceptibility to intergranular corrosion are the composition and microstructure of the alloy.
7. Erosion-corrosion depends on the relative motion between the material surface and the corrosive environment. The nature of the movement and the corrosiveness of the environment are the main variables. However, factors such as the concentration of suspended impurities in the environment are also important.
8. High-temperature attack occurs in the absence of a liquid at elevated temperatures. The nature of the scale depends on the composition of the environment; the presence of sulfides or chlorides may result in nonprotective scales under conditions where an environment containing only oxygen would result in a protective scale.

1.5 Prevention

The detrimental effects of material degradation could be largely prevented through the proper application of existing knowledge, specifically by the following procedures:

1. A proper choice of materials for a given application. In this case, the mechanical strength, cost, and other properties must

be weighed against the material's resistance to degradation.

2. Proper geometric design of the material. In general, simpler forms, without crevices, should be used. Residual moisture should be avoided, as should unnecessary mechanical stress and galvanic coupling that will lead to degradation.
3. Elimination of corrosive materials from the environment. This includes lowering the humidity, the oxygen level, and the concentration of aggressive ions, reducing the acidity, reducing pollution levels, etc.
4. Protection of the chosen material by using inhibitors, coatings, cathodic or anodic protection. Inhibitors are materials added to the solution or to the vapor phase. They interact with the surface of materials, reducing the rate of one or both of the half reactions. Coatings act to form a physical barrier between the material and the aggressive environment. Cathodic protection, which relies on adjusting the potential of the metal into the immunity region of a Pourbaix diagram (see Fig. 1), may be accomplished through the sacrificial-anode method (Morgan, 1987), or by the impressed-current method, where an external power supply is used. Anodic protection relies on passivating the material, that is, forcing the potential into the passive region as shown in a Pourbaix diagram.

2. METHODS AND EQUIPMENT

2.1 Measurement of Thermodynamics

In the previous section, the fundamental thermodynamic and kinetic parameters were introduced. However, in order to use this knowledge, we must make measurements of the important parameters, or quantities that are directly related to them. In this section, the approach used for such measurements will be presented. The section will be divided into two major subsections dealing with the measurement methods and equipment for thermodynamic measurements and those used to determine reaction kinetics. Within each subsection, specific techniques for measurements in solutions and gaseous atmospheres will be presented.

Thermodynamic measurements are vital since they allow the problem to be divided in two: those reactions that can occur, and those that cannot. By quantifying this information, the driving force for a reaction can be determined. As discussed above, however, an understanding of the thermodynamics of a materials degradation phenomenon is a necessary, but not sufficient, state of understanding. Nonetheless, careful measurements of the important thermodynamic parameters provide a starting point for understanding corrosion.

In order to make electrochemical thermodynamic measurements in solution, a reference electrode is required. A reference electrode consists of a material–electrolyte interface that is in thermodynamic equilibrium. Thus, the conditions at this interface do not change with time and can be used as a guidepost for changes that occur at other interfaces. An extremely common reference electrode is the saturated calomel electrode, which consists of mercury in contact with Hg_2Cl_2 in a solution saturated with KCl. This interface is kept in equilibrium by containing it and its surrounding solution in a vessel with a porous plug on one end (see Fig. 4).

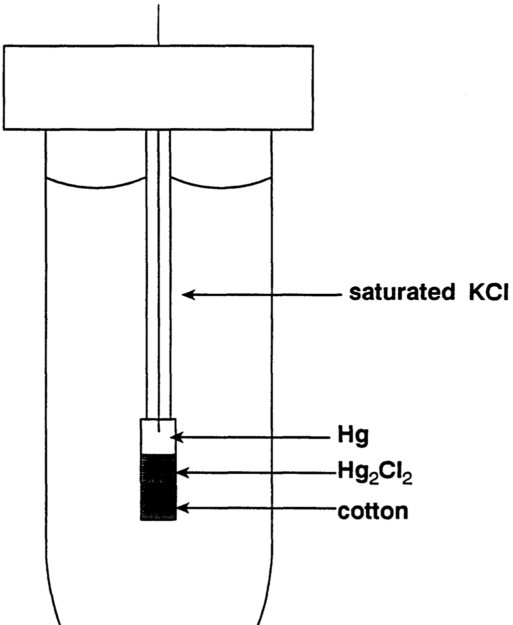

FIG. 4. Schematic of a saturated calomel reference electrode showing the component parts. The entire electrode is immersed in solution, allowing the potentials of other interfaces to be compared against it.

The porous plug allows ionic contact between the reference electrode and the test environment but prevents bulk movement of solution either into or out of the reference electrode chamber, preserving its equilibrium condition. A variety of other reference electrodes is used, as shown in Table 3. Any potentials measured with these electrodes can be referred to the standard hydrogen electrode (SHE) discussed earlier. However, these reference electrodes are far easier to construct and operate than the SHE, which requires a constant flow of pure hydrogren gas at 1 atm pressure.

Often the conditions for equilibrium measurements are extremely difficult to achieve. However, we will see that the ability to measure electrochemical potentials is critical in kinetic measurements as well. In this way, the driving force for a reaction can be related to the reaction rate. While some measurements of thermodynamic quantities can be made electrochemically, the vast majority of free-energy measurements for chemical reactions are made via calorimetry (Atkins, 1982). However, potential measurements are widely used for the determination of solubility products as well as for the determination of ionic concentrations, including H^+ detected by *p*H meters.

Measurements of thermodynamic properties of interfacial reactions when the environment is gaseous are inherently more difficult. Equilibrium conditions are more difficult to establish, especially at elevated temperatures. In most cases, the partial pressure of gases is the critical parameter to determine. In most instances, oxygen is the gas of greatest interest. Ellingham diagrams use an approach similar to the Pourbaix diagrams mentioned previously to illustrate graphically the most thermodynamically stable phases as a function of temperature and oxygen partial pressure (Lou and Heuer, 1990).

Table 3. Common reference electrodes.

Type	Voltage vs SHE (*V*)
Saturated calomel ($Hg_2Cl_2 + 2e^- = 2Hg + 2Cl^-$)	+0.241
Mercury–mercurous sulfate [$HgSO_4 + 2e^- = 2Hg + (SO_4)^{2-}$]	+0.615
Silver–silver chloride ($AgCl + e^- = Ag + Cl^-$)	+0.222
Copper–copper sulfate [$CuSO_4 + 2e^- = Cu + (SO_4)^{2-}$]	+0.318

2.2 Measurement of Kinetics

While thermodynamic considerations are a necessary first step in understanding materials degradation, they are by no means sufficient unto themselves. The fact that all materials (except gold) are *thermodynamically* unstable under ambient conditions does not prevent their successful use technologically. It is the kinetics of their reversion to their most thermodynamically stable form that is limiting. This section will provide an overview of three aspects of kinetic measurements: in solution, in gaseous atmospheres, and as a loss of mechanical properties.

The quantity to be measured that characterizes the kinetics of corrosion depends directly on the type of degradation considered. In solution, the forms of corrosion can be divided into two groupings: uniform material loss and localized attack. These two types of degradation require different approaches to the measurement of their kinetics. Uniform material loss is the simplest form of corrosion and the easiest to measure quantitatively. The most common engineering approach to measuring uniform corrosion rate is the use of coupon exposure tests in which small (e.g., 50 × 150 cm^2) panels of the material of interest are immersed in the solution of interest for a set period of time. Afterward, they are cleaned, and the loss of material is determined by the loss of specimen weight. Standard practices and methods have been developed for such tests (Jones, 1992). Usually, inspection of the surface morphology is performed with both optical and scanning electron microscopy. Surface analysis with x-ray photoelectron spectroscopy, Auger-electron spectroscopy, or secondary-ion mass spectroscopy may also be used. These postexposure tests can provide information on both the mechanism and rate of corrosion attack.

However, the main difficulty involved with exposure tests is the time required for measurable weight loss to occur. In most cases, the time constraints on materials selection decisions require answers in shorter times. It is for this reason that electrochemical methods have become increasingly popular for determining corrosion rate. The electrochemical nature of metallic corrosion is apparent from the earlier sections of this article. The utility of electrochemical measurements for the estimation of corrosion rates relies on the relation between the electrochemical potential and the reaction rate, measured as a current density. The potential has been discussed and relates to the driving force for the reaction to occur. For an oxidation reaction, an (algebraic) increase in the potential increases the driving force in the same way that an increase in temperature increases the driving force for a chemical reaction. The rate of the reaction is exactly the flux of the species:

$$i = nFJ. \qquad (11)$$

This relation allows extremely accurate measurements to be made. Currents in the μA range can be routinely measured. A simple calculation shows that 1 μA cm^{-2} corresponds to a flux of 10^{12} atoms/s. A surface contains approximately 10^{15} sites. Thus, at a current of 1 μA cm^{-2}, it would take approximately 1000 s to remove a monolayer of material. Thus, the sensitivity of electrochemical measurements makes them an attractive addition to longer-term weight-loss experiments. Exposure tests of the materials of interest to the environments cannot be ignored, as these are the most relevant measure of performance and can exhibit rates and types of degradation that were not expected.

Just as one cannot measure the absolute potential across a single interface, one can only measure the net current. At the corrosion potential, the cathodic sites are using all of the electrons liberated during the anodic process. No current needs to be applied by an external measurement device in this case. In fact, no current could be measured by an external device since all of the charge is circulating locally on the same surface, as shown in Fig. 5. No electrons would move through an external connection. However, in order to determine the corrosion rate, we need to disturb the surface and measure the applied current necessary to do so.

The device used to perturb the surface is a potentiostat. It operates like an ideal operational amplifier, as shown in Fig. 6. The interface of interest (called the working electrode) is connected to one of the inputs. The reference electrode is connected to the other. Since the potentiostat has a very high (ideally, an infinite) impedance between these electrodes, no current passes between them, and the reference electrode is maintained in thermodynamic equilibrium. In order to polarize the interface to a set potential vs the reference electrode, a voltage offset (V_{set}) is imposed between the working and reference electrodes (in this case, by a battery and a variable resistor). The potentiostat reacts to this by increasing the voltage at its output in order to return the voltage difference between its inputs to zero. This output is connected to the counter electrode, an inert electrode immersed in the same solution. This causes a current to pass between the counter electrode and the working electrode. The current changes the potential across the working electrode–solution interface until it reaches the desired value. In this way, applied current–potential curves can be generated, as shown by the solid lines in Fig. 2. If both the anodic and cathodic reactions obey Tafel kinetics [i.e., the current–potential behavior follows Eq. (9)], an extrapolation of the linear portions provides an estimate of the corrosion rate (i_{corr}). This current density can be converted to a penetration rate by standard formulas (Jones, 1992). These proce-

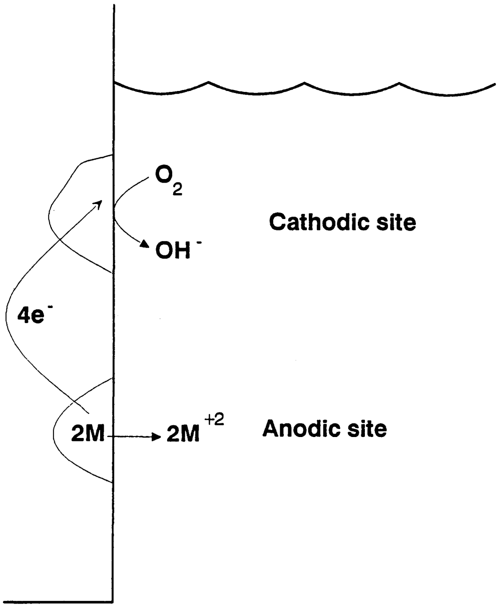

FIG. 5. Schematic of one cathodic and one anodic site communicating on an immersed surface at its free corrosion potential. The electrons liberated during the oxidation of the metal are consumed by the reduction of oxygen.

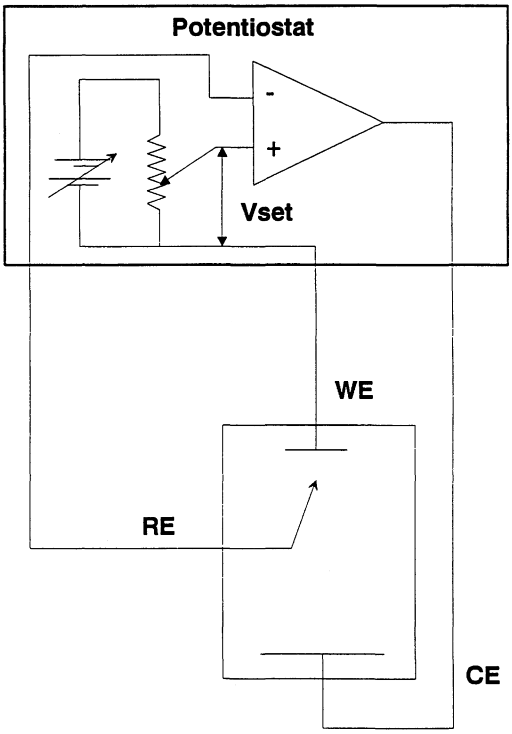

FIG. 6. Schematic of circuit and connections for a potentiostat. The working electrode (WE) is the interface of interest, while the reference electrode (RE) and counter electrode (CE) are used to control the potential of the WE at the desired level (V_{set}).

dures are easily automated and allow a large number of alloys and environmental variables to be studied.

Another method to measure corrosion rate relies on the fact that the equations describing the coupled electrochemical reactions can be linearized in the potential range near the corrosion potential. Stern and Geary (1957) showed that the corrosion rate can be directly related to the slope of the plot of applied current vs potential, as shown in Fig. 7. This approach forms the basis for the polarization-resistance method of corrosion-rate determination. Many commercial corrosion rate monitors make use of this approach.

Localized attack is a much more difficult type of degradation to analyze quantitatively. Its apparent stochastic nature and the simultaneous existence of areas of low and high corrosion rate on the same surface make it theoretically difficult to address. Its dependence on local geometry (as in crevice corrosion) adds an additional level of complexity. Besides exposures to determine the amount or probability of localized attack, electrochemical measurements have been widely used to determine the potential dependence of both the initiation and propagation phases as a function of alloy content and environmental conditions. The application of these measurements to corrosion engineering estimates of performance is at this time controversial. There is growing acceptance that carefully made measurements can be applied to specific field conditions. One area of current interest is the application of electrochemical monitoring for localized corrosion. While localized corrosion rates are not determined, conditions conducive to the development of localized corrosion can be monitored.

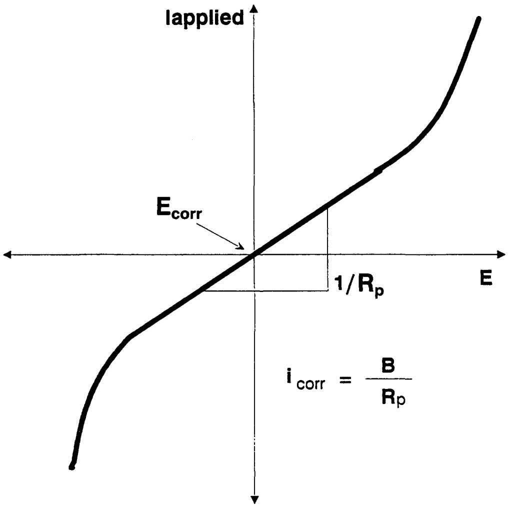

FIG. 7. Schematic plot of applied current vs potential for a polarization resistance measurement. The slope of the line at the corrosion potential is directly related to the corrosion rate.

Atmospheric corrosion kinetics have traditionally been determined by long-term exposures of coupons to the relevant conditions. For example, coupons have been exposed to industrial, urban, and rural atmospheres for up to 50 yr. Similarly, corrosion of materials in hot gases, such as in aircraft engines (exhausts), has been most extensively studied with the exposure approach. In recent years, the application of microbalance techniques has provided additional information.

The quartz-crystal microbalance (see MASS

AND DENSITY, MEASUREMENT OF) is an adaptation of the technology used in the monitoring of thin-film deposition. A thin, *AT*-cut quartz crystal is used as the substrate for the deposition thin film of the material of interest. It is then exposed to the atmospheric conditions of interest, and the resonant frequency of the crystal is followed. Since its initial resonant frequency is in the MHz region and can be measured to an accuracy of less than 1 Hz, changes in mass of less than 10 ng cm^{-2} can be observed (Deakin and Buttry, 1989). These changes are nearly equivalent in sensitivity to the electrochemical current measurements discussed above and open a new avenue to questions concerning atmospheric degradation mechanisms. This approach has been used to show quantitatively that Ni exposed to slightly polluted air corrodes at a much higher rate when both SO_2 and NO_2 are present than when either one is present alone (Zakipour *et al.*, 1986).

2.3 Measurement of the Loss of Mechanical Properties

One of the most important ways in which localized corrosion manifests itself in failures is through the loss of mechanical properties. Stress-corrosion cracking (SCC) and hydrogen embrittlement (HE) result in the loss of mechanical integrity of a wide variety of structures. The failure of the Pt. Pleasant bridge discussed above occurred as a result of hydrogen-embrittlement cracking of a support member. While a great deal of work is focused on the prevention of the conditions under which SCC and HE can occur, measurements of crack growth rate are also important to allow damage tolerance of structures to be evaluated properly. A variety of test methods is available for measuring the susceptibility of materials to SCC and HE, and many have been standardized. The tests fall under two broad categories: those that test susceptibility to both initiation and propagation, and those that quantify the propagation stage. In many cases, alloys are screened for susceptibility by exposing stressed samples to either the service environment or a simulated environment made more aggressive by increasing some environmental variable (e.g., temperature, acidity, etc.). For these tests of both initiation and propagation susceptibility, specimens are usually in the form of C rings or U bends. After a defined exposure time, the samples are removed and checked for signs of cracking. Because of the ill-defined mechanics of such a specimen, very little information of use to designers results. However, such testing is very popular for screening and ranking alloys.

More quantitative information is available from dynamic testing. Two types of approaches are used. In slow strain-rate testing (SSRT), a smooth sample of material is dynamically strained while exposed to the environment of interest. This straining is usually performed at a low rate (e.g., 10^{-5}/s). The increasing strain is necessary to fracture the protective films on the materials repeatedly and allow any crack initiation that will occur to take place. The strain is applied at a slow rate in order to provide the time necessary for the environment to interact with the material. The measured mechanical properties (ultimate tensile strength, strain to failure, etc.) and the fracture surface morphology are compared to those produced in tests performed in an inert environment such as high vacuum or oil. Fracture mechanics testing uses specimens with precracks, often initiated by fatigue. These methods use well-defined specimen geometries to relate the measured cracking to the stress state at the crack tip. Crack growth rates on the order of 10^{-5} cm/s can be measured. However, the equipment and specimen costs are much higher than the SSRT method.

3. SPECIFIC EXAMPLES FOR DIFFERENT CLASSES OF MATERIALS

3.1 Metals and Alloys

In this section, several common and important types of degradation are discussed. Much more information is available elsewhere (Uhlig, 1948; Fontana and Greene, 1967; Parsons, 1982).

Iron and low-alloy steels are well known to be susceptible to general corrosion in the atmosphere or under damp conditions, giving rise to "rusting." The composition of ordinary, red-brown rust is $Fe(OH)_3$, although layers of $Fe_3O_4 \cdot nH_2O$ and $Fe(OH)_2$ may also be present. Rusting results from the concerted action of water and dissolved oxygen on the metallic iron. It is accelerated by an increase in temperature and by appreciable concentrations of halides and acids. A special case occurs under anaerobic conditions when sulfate-reducing bacteria are present. This phenomenon is known as microbiologically induced corrosion.

Stainless steels, which are iron alloys containing at least 11% chromium, are generally resistant to atmospheric corrosion; however, they are susceptible to many other important forms of degradation, including pitting, SCC, and intergranular attack. In this section, a brief review of some of these effects will be given.

The superior corrosion resistance of stainless steels is due to the formation of a stable, highly protective, passive film. The composition of this film is still debated in the literature but is usually thought to consist of a mixture of chromium and iron oxides and hydroxides, enriched in chromium with respect to the bulk metal (Marcus and Olefjord, 1988). Many different stainless steel alloys have been developed for different applications. For example, carbon, molybdenum, titanium, aluminum, and copper are added for strength, nickel and molybdenum for corrosion resistance, sulfur and selenium for machinability, and nickel for formability and toughness (Sedricks, 1979).

Pitting attack results from a local disruption of passivity. The mechanism of the initiation of pitting is hotly debated in the literature (Mansfeld, 1987; Szlarska-Smialowska, 1986); however, for stainless steels, significant evidence exists to suggest that the sites of pit initiation are associated with defects in the film due to the presence of impurities in the alloy (Sedricks, 1979). The propagation of pits is somewhat better understood and is thought to result from a localized acidification within the confines of the pit, leading to a local increase in the dissolution rate and thus autocatalytic growth of the pit (Fontana and Greene, 1967).

Stress-corrosion cracking occurs in the presence of a tensile stress and a specific corrodant. It results in a failure of the steel at much lower stress level than in the absence of the corrodant. Again, the mechanism of this form of degradation is still not well understood. It has been variously proposed that the crack propagates by local dissolution of the metal at the crack tip enhanced by the plastic deformation of the metal, or that atomic hydrogen (produced by the reduction of water or H^+ during the corrosion) accumulates within the metal in the crack tip region, lowering the cohesive strength, or by combinations of these two mechanisms. The specific corrodants that can produce SCC in stainless steels include chloride, sodium hydroxide, and sulfides.

Intergranular attack usually occurs as a result of the segregation of alloying elements or precipitation of electrochemically active phases at the grain boundaries. The segregation often results from heat treating, aging, or welding, though equilibrium segregation can also occur. Mechanistically, either the grain boundary can be actively dissolving itself while the remainder of the material remains passive, or the precipitated phase can be actively dissolving.

The atmospheric corrosion of copper and bronze can be considered a beneficial degradation reaction, since it results in a decorative and protective coating for the metal. In contrast, silver corrodes to produce an unsightly tarnish (Uhlig, 1948; Evans, 1963). The green patina formed on copper roofs consists of a basic copper sulfate [$CuSO_4 \cdot 3Cu(OH)_2$], except near a source of chloride, when it also contains copper oxychloride [$CuCl_2 \cdot 3Cu(OH)_2$]. Sulfur-containing pollutants are necessary for the development of the patina. The patina formed on bronzes also depends on the atmosphere. In industrial areas, greenish-black and black films are formed, while bright green films and reddish-brown films are formed near the ocean and in dry rural areas, respectively. Bronze is also subject to an interesting periodic tarnishing, with alternate layers of red cuprite (Cu_2O) and green malachite [$CuCO_3 \cdot Cu(OH)_2$] formed on the surface. This phenomenon is due to a fluctuating supply of CO_2 in the environment. The tarnish on silver is again due to sulfur pollution and is not observed unless traces of H_2S are present in the atmosphere.

3.2 Polymers

Unlike for most materials, the degradation of polymers can be either detrimental or beneficial. For polymeric materials that are used in structural components (such as advanced aircraft skin), the design life is usually such that degradation is viewed as a problem. For applications in which the object is used for only a short time (such as a beverage container), rapid degradation of the polymer afterward would be advantageous, especially for easy recycling. The degradation of polymeric materials usually results from chemical reactions that break bonds, rather than electrochemical oxidation reactions as is usually the case for metallic materials exposed to solution. Because of the huge number of polymers, generalities are difficult, but all polymers show some sensitivity to heat, radiation, oxygen, and solvents. The degradation reactions damage the properties of the polymers, including their mechanical, optical, and electrical properties. Common examples of degradation include the chalking of automobile paint, embrittlement of rubber hoses, cracking and peeling of vinyl tops on automobiles, the fading of fabrics and paints, and the failure of electrical wire insulation.

Bond breaking results in a change in both the average molecular weight and its distribution, as well as a change in the chemical character of the polymer molecules. Since the molecular weight characteristics of polymers determine many of their properties, uncontrolled changes in these characteristics usually lead to a loss in performance. Degradation reactions include the cleavage of main chains, cross-linking, and the cleavage of side groups. These reactions occur via free-radical pathways, which allow both chain scission and cross-linking to occur. These three examples are shown in Fig. 8. In addition to the reduction in molecular weight caused by these processes, the products of the degradation reaction can be deleterious to the polymer's properties as well, making the process essentially autocatalytic. Chemical oxidation often occurs by exposure of the material to a combination of heat, oxygen, and uv radiation. Impurities in the polymer, which are a remnant of the manufacturing process, often act as catalysts for these reactions. Oxygen has an accelerating effect since it is reactive toward both the polymer and the degradation intermediates, leading to the formation of more free radicals.

Additives to polymers are included to stabilize them against uv radiation. In general, these antioxidants act as radical scavengers, reacting with the free radicals or the agents that form them before the polymer can be further degraded. Thus, they do not prevent damage altogether but simply lessen its rate to acceptable levels. For example, carbon black acts by absorbing uv radiation before it penetrates the bulk of the polymer. Thus, black polyethylene (PE) lasts longer when exposed to sunlight than does clear PE (Kelen, 1983).

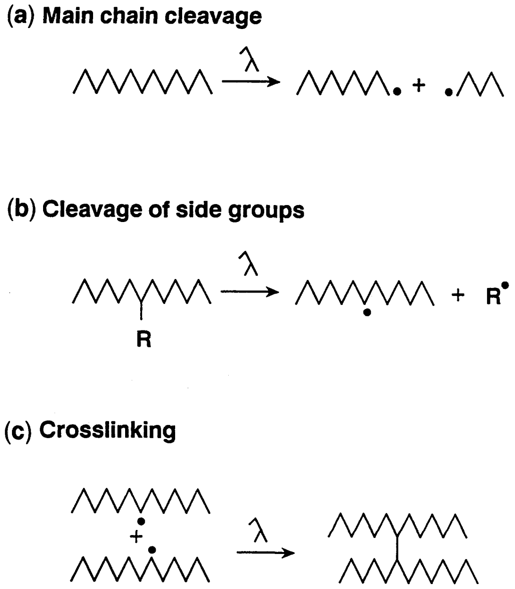

FIG. 8. Schematic showing three types of polymer degradation reactions resulting from the absorption of radiation (λ). Similar reactions can occur as a result of other active agents such as O_2.

Depolymerization of linear polymers is essentially the reverse of the reaction that forms them. The repeat units of the molecule (sometimes referred to as the "mers") start to come apart, starting with the end groups of each chain. This process is usually initiated by heat, but radiation can induce it as well if the end groups are able to absorb at the wavelength of interest. Polymers are also susceptible to a type of degradation analogous to the SCC of metallic materials.

With the recent increase in the desire for recyclable products, polymer degradation has been studied in terms of methods to generate reusable raw materials from the wide variety of plastic packaging currently used. The technological difficulties are significant, in part because of the expense of collecting and transporting the large volumes (but small weights) of polymeric materials. While it is necessary to remove the product labels from the plastic packages, the difficulties involved make it economically unattractive given current costs of landfills. In essence, it is far cheaper to throw these materials away than it is to recycle them. Incineration is typically not an option, as many commercial plastics have combustion products that are harmful or extremely corrosive. However, as landfill costs increase, the economics of plastic recycling are bound to change, and many of the technological difficulties will be overcome. Present applications for biodegradable polymers include mulch fillers and seedling holders used during tree planting. Caprolactone polyester has been found to dissolve completely in soil after 12 months, supporting the seedling tree while it takes root (Potts, 1978).

3.3 Ceramics

While ceramics are often used in situations where corrosion of metals and alloys prohibits their use, these materials are also susceptible to degradation. In most cases, the degradation rates for ceramics are much lower than for metals and alloys, even under elevated temperature conditions. However, there are a number of applications in which ceramic degradation is an important factor. One of the most important applications of ceramics is as a refractory material in the containment of molten salts, slags, and metals. For example, magnesium oxide impregnated with carbonaceous material is used to line vessels for steelmaking in basic oxygen furnaces. During the casting operation, the MgO reacts with the FeO in the slag to form a solid solution with a low melting point (compared to the operation temperature). This leads to dissolution of the refractory.

Nonoxide ceramics (such as SiC and Si_3N_4) are of interest for automotive engine and gas turbine applications because of their maintenance of strength at high temperature as well as their excellent wear resistance. However, these materials can be attacked by oxygen, hydrocarbons, NO_x, and sulfur contamination in fuel. Si_3N_4 densified with additions of Al_2O_3 and Y_2O_3 is of interest in coal gasifiers. However, it is found to degrade rapidly under conditions of high temperature (greater than 1200 °C) and in the presence of sulfur species (Oliveira *et al.*, 1990). The formation of silica-rich protective scale is thought to be important. A variant of SCC can also be observed under application conditions (Weisskopf and Morgenthaler, 1990).

Other applications of ceramics in which degradation can be a problem are hot corrosion (termed *fluxing*) during contact with molten ionic condensates such as encountered in the effluent of coal combustion plants, aqueous corrosion of ceramic containers in the containment of nuclear wastes, and the preferential leaching of stabilizers such as MgO, CaO, and Y_2O_3 from stabilized zirconia.

GLOSSARY

Many definitions of terms related to corrosion can be found in ASTM Standard G-15 (ASTM, 1992). Some of these are listed below.

Anode: The electrode in an electrochemical cell at which oxidation is the principal reaction. Electrons flow away from the anode in an external circuit.

Cathode: The electrode in an electrochemical cell at which reduction is the principal reaction. Electrons flow toward the cathode in an external circuit.

Exchange-Current Density: The rate of charge transfer per unit area when an electrode reaches a dynamic equilibrium at its reversible potential in a solution; that is, when the rate of anodic charge transfer (oxidation) balances the rate of cathodic charge transfer (reduction).

Overpotential: The change in potential of an electrode from its equilibrium or steady-state value when a current is applied.

Passive: The state of the metal surface characterized by low corrosion rates in a potential region that is strongly oxidizing for the metal.

Tafel Slope: The slope of the straight-line portion of a polarization curve, usually occurring at more than 50 mV from the open-circuit potential, when presented in a semilogarithmic plot in terms of voltage per logarithmic cycle of current density (commonly referred to as volts per decade).

Works Cited

ASTM (1992), *Annual Book of ASTM Standards, Vol. 3.02, Erosion and Wear, Metal Corrosion*, Philadelphia: American Society for Testing and Materials.

Atkins, P. W. (1982), *Physical Chemistry*, 2nd ed., San Francisco: Freeman.

Benard, J. (1964), *Oxydation des Metaux*, Paris: Gautheri-Villars.

Bockris, J. O'M., Drazic, D. M. (1972), *Electrochemical Science*, London: Taylor & Francis.

Butler, G., Ison, H. C. K. (1966), *Corrosion and its Prevention in Waters*, New York: Reinhold.

Deakin, M. R., Buttry, D. A. (1989), *Anal. Chem.* **61**, 1147A–1154A.

During, E. D. D. (Ed.) (1988), *Corrosion Atlas, A Collection of Illustrated Case Histories*, Amsterdam: Elsevier.

Evans, U. R. (1963), *An Introduction to Metallic Corrosion*, 2nd ed., London: Edward Arnold.

Fontana, M. G., Greene, N. D. (1967), *Corrosion Engineering*, New York: McGraw-Hill.

Goldsmith, A., Waterman, T. E., Hirschhorn, H. J. (Eds.) (1961), *Handbook of Thermophysical Properties of Solid Materials*, rev. ed., New York: Macmillan.

Hudson, J. L., Bassett, M. R. (1991), *Rev. Chem. Eng.* **7**, 111.

Jones, D. A. (1992), *Corrosion*, New York: Macmillan.

Kelen, T. (1983), *Polymer Degradation*, New York: Van Nostrand Reinhold.

Lou, V. L. K., Heuer, A. H. (1990), in: R. J. Fordham (Ed.), *High Temperature Corrosion in Technical Ceramics*, New York: Elsevier, p. 33.

Mansfeld, F. (Ed.) (1987), *Corrosion Mechanisms*, New York: Marcel Dekker.

Morgan, J. (1987), *Cathodic Protection*, 2nd ed., Houston, TX: National Association of Corrosion Engineers.

Marcus, P., Olefjord, I. (1988), *Surf. Interfac. Anal.* **11**, 569–576.

Narayan, R. (1983), *An Introduction to Metallic Corrosion*, New Delhi: Oxford and IBH.

Oliveira, F. C., Edwards, R. A. H., Fordham, R. J., Riley, F. L. (1990), in: R. J. Fordham (Ed.), *High Temperature Corrosion in Technical Ceramics*, New York: Elsevier, p. 53.

Parsons, R. (1959), *Handbook of Electrochemical Constants*, London: Butterworths.

Parsons, R. N. (Ed.) (1982), *Corrosion Processes*, London: Applied Science.

Payer, J. H., Boyd, W. K., Dippold, D. B., Fisher, W. H. (1980), *Mater. Perform.* **11**, 1.

Potts, J. E. (1978), in: H. H. G. Jellinek (Ed.), *Aspects of Degradation and Stabilization of Polymers*, New York: Elsevier, p. 617.

Pourbaix, M. (1974), *Atlas of Electrochemical Equilibria in Aqueous Solutions*, Houston, TX: National Association of Corrosion Engineers.

Sedricks, A.J. (1979), *Corrosion of Stainless Steels*, New York: Wiley.

Stern, M., Geary, A. L. (1957), *J. Electrochem. Soc.* **104**, 56.

Szlarska-Smialowska, Z. (1986), *Pitting Corrosion of Metals*, Houston, TX: National Association of Corrosion Engineers.

Uhlig, H. H. (1948), *The Corrosion Handbook*, New York: Wiley.

Weast, R.C. (Ed.), (1991), *Handbook of Chemistry and Physics*, Cleveland, OH: CRC Press.

Weisskopf, K.-L., Morgenthaler, K.D. (1990), in: R. J. Fordham (Ed.), *High Temperature Corrosion in Technical Ceramics*, New York: Elsevier, p. 3.

Zakipour, S., Leygraf, C., Portnoff, G. (1986), *J. Electrochem. Soc.* **133**, 873–876.

Further Reading

American Society for Metals (1987), *ASM Handbook*, Vol. 13, *Corrosion*, 9th ed., Metals Park, OH: ASM International.

Baboian, R. (Ed.) (1976), *Electrochemical Techniques in Corrosion Engineering*, Houston, TX: National Association of Corrosion Engineers.

Fordham, R. J. (Ed.) (1990), *High Temperature Corrosion in Technical Ceramics*, New York: Elsevier.

Jellinek, H. H. G. (Ed.) (1978), *Aspects of Degradation and Stabilization of Polymers*, New York: Elsevier.

Jones, D. A. (1992), *Corrosion*, New York: Macmillan.

Kelen, T. (1983), *Polymer Degradation*, New York: Van Nostrand Reinhold.

Shrier, L. L. (Ed.) (1979), *Corrosion*, London: Butterworths.

MATERIALS DEFORMATION

See REVERSIBLE AND IRREVERSIBLE DEFORMATION OF MATERIALS

MATERIALS PREPARATION: SOLIDS

David C. Paine, *Division of Engineering, Brown University, Providence, Rhode Island, U.S.A.*

	Introduction	365
1.	**Vapor-State Processes**	366
1.1	Physical Vapor Deposition	366
1.1.1	Common Themes	366
1.1.2	Evaporation	367
1.1.3	Sputtering	370
1.2	Chemical Vapor Deposition	372
1.3	Ion Implantation	374
2.	**Liquid-State Processing**	375
2.1	Casting Methods	377
2.1.1	Sand Casting	378
2.1.2	Investment (Lost-Wax) Casting	380
2.1.3	Permanent-Mold Castings	380
2.2	Oriented or Single-Crystal Growth	381
2.2.1	General Single-Crystal Growth Considerations	382
2.2.2	Structural Components: Bridgman Technique	382
2.2.3	Float-Zone (Zone-Refining) Method	383
2.2.4	Semiconductor and Optical Materials	383
2.3	Rapid Solidification	385
2.4	Electrochemical Processes	385
3.	**Solid-State Processing**	386
3.1	Powder Processing	386
3.1.1	Powder Compaction	387
3.1.2	Sintering	388
3.2	Mechanical Forming	390
3.2.1	Bulk Forming	391
3.2.2	Sheet Forming	392
4.	**Polymer Processes**	393
4.1	Polymeric Materials	394
4.2	Polymer Forming	395
	Glossary	397
	Works Cited	397
	Further Reading	398

INTRODUCTION

The word *material* for the purposes of this article refers to the physical matter from which useful objects are engineered. The preparation of materials begins (at least for the purposes of this discussion) after a mineral or other natural resource has been converted into a raw material. The process of forming a raw material through the refinement of mineral or organic precursors is not discussed, since these processes vary greatly depending on the final application and the degree of chemical change necessary to realize the final product. To illustrate this point, consider the production of a metal (such as Fe or Cu) from the mineral state (usually either a sulfide or an oxide). The product of the mining process is essentially rock (gangue) with a relatively low concentration (0.1–1 wt %) of mineral in the form of metal oxides and sulfides. This material must be mechanically ground to a fine powder to allow separation of the useful minerals from the surrounding rock. The mineral is concentrated by mechanical separation (based on differences in density and surface properties and on chemical dissolution) of the mineral from the waste rock. The concentrate is then chemically treated [typically at high temperature in a *blast* (for Fe) or *smelting* (for Cu) furnace] to reduce the oxide (e.g., FeO) or sulfide mineral (e.g., CuS) to an impure elemental state. The material of interest then undergoes purification or refining to eliminate undesirable impurities (e.g., S, P, Si and C in Fe), and other constituents are added (alloyed) to modify the physical properties of the material. For metals, the process of refinement can include dissolution of the impure metal in an electrolyte followed by electrochemical plating (e.g., electrorefining of Cu) or chemical treatment of a molten metal to incorporate impurities selectively into a oxide slag (e.g., CaO/O_2 treatment of Fe to form steel). In contrast, many useful engineering materials such as wood receive relatively little preparation other than drying and milling before they find engineering application. In this dis-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

cussion, the preparation of materials that require only crude chemical preparation steps before being shaped into a useful form (e.g., bricks, wood, stone, and natural fibers) is neglected.

This article addresses, in general terms, materials preparation steps between those in which a raw material is produced through a refining or mineral processing technology and the final shaping process (such as machining or lithographic patterning). The emphasis is on processing raw materials into a near-final chemical and local-microstructural state immediately prior to the assembly or shaping of a final material into a useful device, structure, or component. These materials preparation steps will be subdivided into three principal categories based on the primary matter transport mechanism that participates in the preparation of the final solid. As such, the main headings will include vapor-state, liquid-state, and solid-state processing. A fourth category, polymer processes, is added since the preparation of a polymer begins with gaseous precursors (e.g., ethane) and passes through either or both liquid and solid states during the course of its preparation for engineering application. Representative examples of common process(es) in each of these major headings are provided; no effort has been made to survey exhaustively all materials preparation technologies since new methods are constantly being developed. In many cases, detailed information on a particular materials preparation process may be found under a more specific heading.

1. VAPOR-STATE PROCESSES

In this section, a selection of materials preparation processes that employ the vapor phase as the primary material transport mechanism are discussed. After deposition *via* any one of several vapor-state methods (some of which are discussed below), the material is typically thin (10 nm to 1 mm) relative to the substrate dimensions. All of the methods selected for this section are used in the fabrication of integrated-circuit devices in which a variety of conducting, semiconducting, and insulating materials are sequentially deposited. These methods form the basis for all modern microelectronic technologies ranging from integrated circuits to solid-state flat panel displays to modern magnetic- and ferromagnetic-based memory devices. Vapor-phase deposition is also widely used for the deposition of decorative and abrasion- and wear-resistant coatings. All vapor-deposition techniques, by definition, take place under controlled atmosphere conditions onto a suitable substrate. In evaporative, sputter, and chemical vapor-deposition processes, the material is deposited onto a substrate. In ion implantation, the material is deposited into the near-surface region of the substrate.

1.1 Physical Vapor Deposition

1.1.1 Common Themes The two principal methods of physical vapor deposition (PVD) are sputtering and evaporation. These methods are used to deposit thin films of material on a substrate to a thickness that is typically less than 1 μm and, less commonly, up to several millimeters. Thicker films may be removed after deposition to become free-standing foils. In all PVD processes, the coating material is passed from a source to the substrate *via* the vapor phase. The interaction of the source molecules with the atmosphere as they travel to the substrate must be avoided, and for this reason, the deposition chamber is usually evacuated by vacuum pumping. The chamber pressure must be low enough during deposition to prevent molecules leaving the source from colliding (and therefore reacting) with residual gas molecules (e.g., O_2) before landing on the substrate. The pressure necessary to achieve this aim is actually quite modest. At atmospheric pressure (1.013×10^5 Pa), the molecular density at 25 °C is about $2.5 \times 10^{25}/\mathrm{m}^3$, while at 10^{-3} atm (100 Pa), the vapor-phase density is three orders of magnitude lower, $2.5 \times 10^{22}/\mathrm{m}^3$. Attaining this vacuum level is easily achieved with a modern rotary vane pump. An important parameter is the mean free path λ, which is defined in the kinetic theory of gases as the average distance traveled by a gas atom between collisions with another gas atom. The mean free path decreases with decreasing pressure since the spacing between molecules increases. For example, the average distance traveled by a N_2 molecule at 25 °C can be calculated from its molecular diameter, $\delta = 0.368$ nm, the temperature (i.e., kinetic energy of the molecule), and the pressure by the following relation:

$$\lambda = \frac{kT}{\pi\delta^2 P}, \tag{1}$$

where k is Boltzmann's constant, P is the gas pressure, and δ is the molecular diameter. Using appropriate units, the mean free path at 1 atm (1.013×10^5 Pa) can be calculated to be 9.5×10^{-8} m, while at 100 Pa, it is 10 mm.

The second (and more important) reason for establishing a decreased pressure of reactive species in the deposition chamber is associated with the impingement of nonfilm species (i.e., impurities) on the growth surface during deposition. If these non-film species are incorporated into the growing layer, then the composition of the deposited material will be seriously contaminated by the impurities in the deposition system. In reactive deposition processes, a gaseous impurity is purposely introduced into the vacuum system, as is widely done in the deposition of oxides (e.g., indium oxide using O_2 and an in source) from metal sources. The surface impingement rate (I) can be calculated from the kinetic theory of gases and is given, approximately, by

$$I = \frac{2^{0.5}P}{(\pi mkT)^{0.5}} \tag{2}$$

where P is the pressure in Pa, m is the molecular weight, and T is the temperature in Kelvins.

On the basis of the above considerations, it is possible to see that when layer-by-layer atomic control of the composition of a deposited layer is important, then low pressures are required ($10^{-9} < P < 10^{-7}$ Pa is typical). Consider the case where an evaporation process is being performed in a chamber that was initially evacuated to 1.3×10^{-4} Pa. The deposition rate is 1 monolayer/s, and we will assume for simplicity that the deposited atoms are 0.3-nm-diam hard spheres that are packed in a square array. With this information, it is easy to calculate the composition of the film in terms of the ratio of impurities (i.e., air molecules) to film species. The 1-monolayer/s rate corresponds to a deposition of 1×10^{15} atoms/cm^2 s. The surface impingement rate from impurities left over after evacuation of the deposition chamber is about 7.6×10^{16} atoms/cm^2 s. Note that the ratio of impurity atoms to film atoms exceeds unity. Actually, this gives an exaggerated (<1 at. % is more typical at this background pressure) view of the importance of a low background pressure since, in reality, not all atoms that impinge on the surface stick. The ratio of the number that stick to the total impingement rate (known as the sticking coefficient) is typically much lower than unity and is dependent on the incident species and changes with the composition of the deposited material.

The microstructure of a deposited film depends on the materials being deposited, the substrate, the temperature of the substrate, and the energy of the incident material flux. The resultant film may be amorphous, fine- or coarse-grained polycrystalline, or single-crystal. The film may contain either compressive or tensile stresses depending on the thermal expansion difference between the substrate and the deposited film and, particularly in sputter deposition processes, is dependent on the background pressure. In the case of a single-crystal substrate and the deposition of a material with like crystal structure but different natural lattice parameter, a single-crystal layer with a commensurate or coherent interface may grow in a process known as epitaxy. In this case, the difference in lattice parameter is accommodated either entirely elastically (if the mismatch is small) or by a combination of elastic strain and the presence of strain-relieving interfacial dislocations.

1.1.2 Evaporation In evaporative processes, the deposit is formed by heating a source material so that its equilibrium vapor pressure is locally raised. This causes the evaporation or sublimation of the desired species into the vapor phase that passes through the apparatus and condenses on the substrate, which is at a much lower temperature. A schematic view of this process is shown in Fig. 1. The source material is heated by any one of a wide variety of methods including simple resistance heating, induction, or electron-beam or laser heating of the source material. The advantage of this method of thin-film preparation is that extremely high-purity materials can be formed since the process takes place in vacuum (typically less than

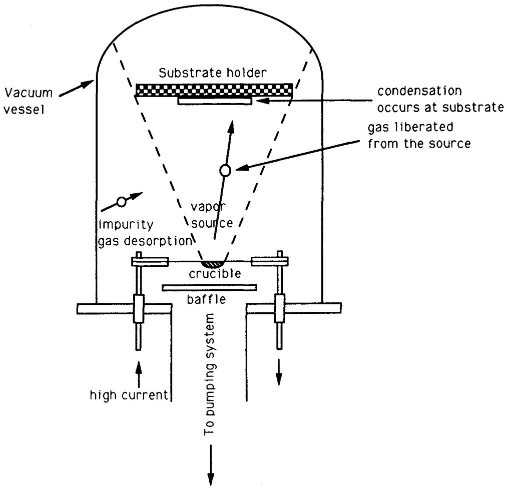

FIG. 1. Schematic view of a evaporative deposition process. The source is heated to a temperature sufficient to raise the equilibrium vapor pressure of the source material. The entire assembly is contained in a vacuum chamber.

1.3×10^{-4} Pa), and high deposition rates (relative to the surface impingement rate) can be achieved. This can also be a cost-effective thin-film deposition choice since relatively simple heating sources can be used. For example, plastic wrapping material used to protect moisture-sensitive food or electronic products is made by depositing a few tens of nanometers of Al on the surface of a polyethelyene or polypropylene polymer sheet. This is done by passing the polymer sheet (100s of meters long and up to 10-m wide) past a set of Al evaporation sources arranged in a line so that the entire width of the plastic sheet is coated as it passes by the evaporation source. In this common industrial procedure, the rolls of thin plastic sheet are placed in a large vacuum chamber, which can accommodate both the feed and take-up rolls of plastic.

The transport of species in the gas phase occurs under molecular flow conditions at low pressures (below 1.3×10^{-3} Pa). Molecular flow implies that the molecules in the gas do not interact with each other because the mean free path of the gas species is much larger than the distance between the substrate and the source. For this reason, typical evaporative deposition processes only allows the substrate areas that are in the line of sight of the source to be coated. This means that shadowing and poor step coverage are typical of evaporative processes. To alleviate this problem, inert gases are sometimes introduced so as to modify the conditions in the chamber from molecular to viscous (cooperative flow involving interaction between molecules) flow. The problem can also be reduced by rotating the object that is to be coated in a way that allows all parts of it to be exposed to the source.

Heating the source material causes an increase in the local equilibrium vapor pressure of the evaporated species. A pressure gradient in the source molecules leads to gas-phase transport to cooler surfaces where condensation occurs. The rate at which atoms pass into vacuum from a heated source is given by the Hertz-Knudsen equation

$$W = \frac{\alpha_v(P^{\mathrm{eq}} - P)}{(2\pi mkT)^{0.5}}, \tag{3}$$

where W is the rate of evaporation per unit area, P^{eq} is the equilibrium vapor pressure of the source material, P is the actual pressure in the chamber, and α_v is an emission or evaporation constant. For a pure source materials with an atomically clean surface, α_v is unity; if the surface is contaminated, then it may be three orders of magnitude lower. Consequently, a low background pressure P is more important for maximizing the emission constant than for providing a diffusion gradient. Note that the equilibrium vapor pressure is extremely sensitive to temperature since $P^{\mathrm{eq}} = \exp(-\Delta G^0/kT)$ where ΔG^0 is the Gibbs free energy for either sublimation or boiling. The equilibrium vapor pressures for a variety of elements and materials are shown in Fig. 2. In general, in order to establish a reasonable deposition rate at the substrate, an equilibrium vapor pressure of about 1×10^{-2} Torr is required. For some materials such as Al, this pressure requires that the metal be melted (T = 1220 °C at $P^{\mathrm{eq}} = 10^{-2}$ Torr while T_{melting} = 659 °C), while for others such as Cr, sufficient pressure is achieved by heating the solid below its melting point (T = 1400 °C at $P^{\mathrm{eq}} = 10^{-2}$ Torr while T_{melting} = 1900 °C). Relatively few compounds decompose into the vapor congruently (i.e., composition of the gas is not identical to the composition of the compound). Materials such as CdSe develop an equilibrium vapor pressure of 10^{-2} Torr at 660 °C where the melting

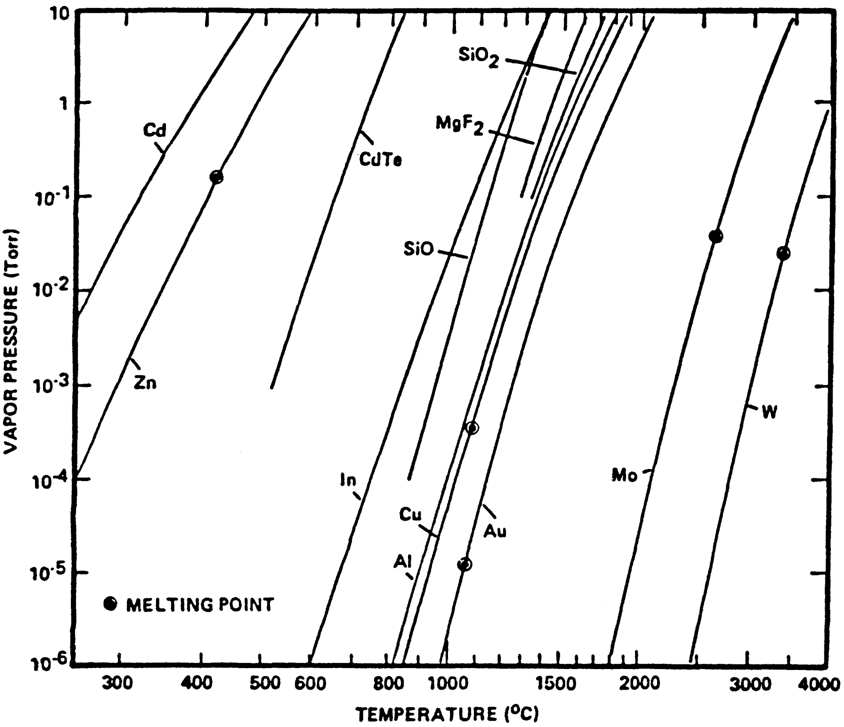

FIG. 2. Equilibrium vapor pressure of selected elements and compounds.

point of CdSe is 1250 °C, although the gas that evolved from the surface is not congruent since it is composed primarily of Se_2 and Cd. An important electronic oxide is WO_3 that melts at 1473 °C but can be volatilized into WO_3 and $(WO_3)_3$ at 1140 °C. Note that in any evaporation technique the crucible must be carefully selected since the support material must not react with the heated source material.

Conventional evaporation techniques do not allow the direct deposition of multicomponent alloys. Multicomponent films require the use of separate sources for each component, since they will typically have different equilibrium vapor pressures and, hence, different evaporation rates. Different evaporation rates, in turn, lead to an enrichment in the source crucible of the slower evaporating species and, hence, can result in compositional nonuniformities in the thickness of the deposit. In addition, great care must be taken to ensure compositional uniformity laterally across the substrate since the material flux decreases with increasing distance from the source. The deposition of compounds such as oxides and nitrides usually involves the introduction of a gaseous species into the deposition chamber. For example, the evaporative deposition of In_2O_3 from In metal involves heating the relatively low–vapor-pressure In species to volatilize it into an oxygen-containing atmosphere.

More sophisticated methods are also available for depositing multicomponent materials. One example is the use of a multiple hearth system in which a separate source (or hearth) for each component allows the temperature of each source to be adjusted in a way that sets the vapor pressures, and hence the evaporation rate, of the individual species. The equilibrium vapor pressure can be set over the source material independently from the pressure in the deposition chamber by limiting the orifice through which the evaporating species can escape. This is the principle employed in a Knudsen cell where vapor-solid (or -liquid) equilibrium is established inside the cell independent of the pressure in the deposition chamber, and at a particular temperature, the rate at which the vapor species pass into the vacuum chamber is controlled by the size of the orifice and the temperature of the cell. *Molecular beam epitaxy* (MBE) (*q.v.*) is a specialized, ultrahigh-vacuum technique that typically uses multiple hearths and Knudsen cells. Another approach is known as flash evaporation and avoids the problem of differential evaporation rates by completely and rapidly vaporizing the source material with a single high-power pulse. In this way, the composition of the source material and that of the deposited film are sure to be the same. This approach has the disadvantage of requiring exact control over the amount of material in the crucible, since it is all evaporated and hence defines the final thickness of the film. A variation on this approach is known as *pulsed laser deposition* or *laser ablation* where a small portion of the source material is evaporated by a high-power laser pulse that results in the ablation of the small region of the target that is heated by the focussed laser.

1.1.3 Sputtering In sputter deposition processes, atoms are stripped from the surface of a cathodically biased solid source material (also known as the target), which is placed in close proximity to an energetic plasma containing positively charged inert gas atoms. The plasma is generated using either a dc or an rf field and back filling the vacuum chamber with an inert gas (usually Ar). The positively charged ions (typically Ar^+) are accelerated toward the target by the bias (50–500 V), and secondary momentum transfer processes result in the ejection of atoms from the near-surface region of the target. The general setup for sputter deposition is shown in Fig. 3. The energy of the sputtered target atoms (50–500 eV) is much higher than the thermal energy ($kT < 0.1$ eV) possessed by an evaporated molecule. This allows the formation of crystalline deposits at lower substrate temperatures than would be needed for evaporative deposition of the same material. Further, the deposition rate is, to some degree, controllable, and is typically much higher than can be achieved by evaporation methods. In addition, sputter deposition techniques are available for the deposition of essentially any metal or insulator, compound or alloy.

The mechanism by which target atoms are sputtered into the vapor phase involves a two-step momentum-transfer process. This two-step process is required because a single incident-ion/target-atom collision cannot impart momentum to the target that is opposite to the incident ion momentum. Instead, a collisional cascade must occur where a primary knock-on atom reflects off a lower plane of atoms to collide (from below) with a second target atom in the top monolayer. The process is shown schematically in Fig. 4. The choice of Ar to form the plasma is mandated by simple conservation-of-momentum considerations since it is obvious that the greatest exchange of momentum occurs between particles of equal mass. Argon, with mass 40, is suitable for many sputter-deposited materials. Heavier inert gases such as Kr or Xe have advantages for some applications (with larger atomic mass targets) but are, in general, too rare and expensive for typical applications. If the mass of the incident ion is larger than that of the target material, then the incident ion will not be back-reflected by a single collision. If the mass of the incident ion is less than that of the target, then, in the case where there is a direct impact, the reflected ion may possess significant energy. Although ionized Ar ions are accelerated toward the target, a large fraction of them are neutralized immediately prior to impact by secondary electrons that are emitted from the surface of the target. As a consequence, the reflected particles are often uncharged and possess significant kinetic energy, which is imparted to the growing film in a process

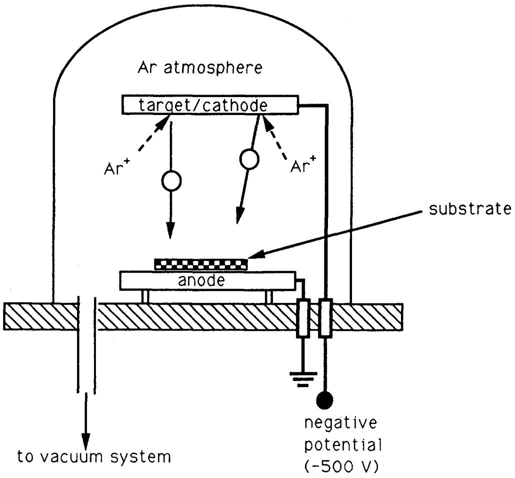

FIG. 3. Schematic view of a dc diode sputtering system.

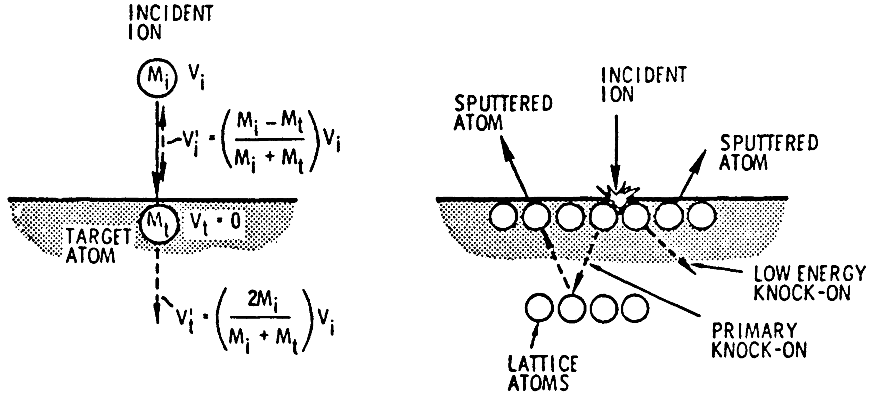

FIG. 4. Schematic view of the momentum transfer processes that take place in the target during sputtering. V_i is the ion velocity, V_t is the target atom velocity, and the prime denotes velocities after the collision (Thornton, 1982).

known as ion-peening. This has a significant effect on the microstructure and physical properties (mechanical strain, resistivity, etc.) of the film and results in the entrainment of Ar in the deposited layer.

The binding energy of a surface atom in the source material is typically in the range 5–10 eV, while the energy needed to displace a bulk atom from its lattice site is about 25 eV. The kinetic energy of an atom that has been sputtered from the source into the vapor phase depends on its molecular weight but typically ranges from 5 to 40 eV. The sum of these energy terms is significantly less than the 500 eV that the incident Ar^+ ions possess as a result of being accelerated from the plasma to the cathode/source by an applied bias. The excess energy is consumed by multiple collision/ejection events that produce a number of atoms ejected from the target per incident Ar^+ ion that is typically greater than unity. The ratio of target atoms ejected to incident ions is called the sputter yield. The sputter yield depends on the target species, target surface topography, bombarding species, energy of the bombarding species, and angle of incidence of the bombarding species. It does not depend upon the temperature of the target or the ionization of the incident particle immediately prior to collision. Experimentally determined sputter yields are given as a function of incident Ar energy for a number of species in Fig. 5.

The compositional uniformity of sputtered alloys and compounds in the through-thickness direction is assured if the target is sufficiently cooled to prevent heating and subsequent diffusion or vaporization of the constituents. It is also necessary that there are no reactive contaminants present in the deposition chamber. Variations in the composition of the target compared to the deposited film occur in multicomponent systems when vapor-phase transport is not the same for all constituents and when their sticking coefficients are different.

Although the sputter yields of the various elements are rarely the same, the film de-

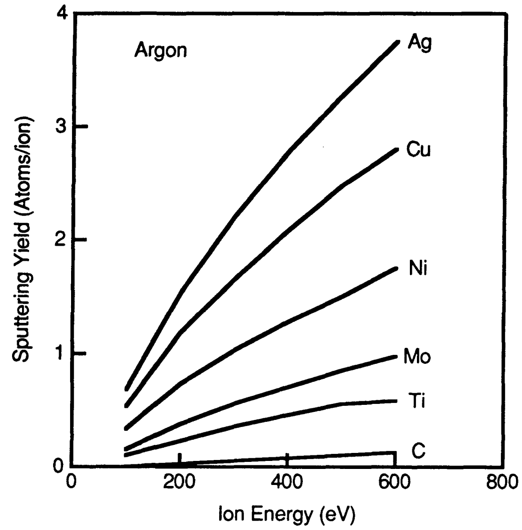

FIG. 5. Plot of sputter yield versus incident Ar^+ ion energy for various target species and normal incidence (McGuire, 1982).

posited on the substrate has, in general, the same composition as the target, provided that the constituents of the target do not diffuse or decompose and if the sticking coefficients of the various species at the substrate are the same. This is dictated by simple conservation of mass which requires that, under steady-state conditions, the composition of the flux of material leaving the target must be the same as the bulk composition of the target. The explanation can be seen by considering the following. Initially, the surface of the target has a composition that is equal to that of the bulk. After some period of time, the component in the target that has a higher sputter yield is depleted so that the composition in the near-surface region of the target (typically 3- to 10-nm thick) is enriched in the component with the lower sputter yield. The flux of a particular component leaving the surface is equal to its sputter yield multiplied by its surface concentration. Those elements that have a lower sputter yield become increasingly enriched at the surface until the product of the sputter yield and the steady-state surface concentration of a particular component is equal to its concentration in the bulk. The effective sputter yield of a constituent in an alloy or compound is not, in general, the same as that of the constituent by itself. This is due to different binding energies and to different atomic masses involved in the atomic collision sequence within the alloy or compound.

The composition of the material flux leaving the target is, at steady state, the same as the composition of the bulk target. The stoichiometry of the deposited film, however, will not necessarily be the same as the target if the film contains a volatile component. This is due to the volatilization of one of the species at the substrate surface as a result of a low sticking coefficient. Examples include O-, N-, and H-based compounds which are typically deficient in these elements when sputter deposited. The solution is found by adding a make-up gas containing the volatile species in a process known as reactive sputtering. This technique is similar to reactive evaporation and has the advantage that many complex compounds may be formed from easily manufactured targets and the introduction of a suitable gas. This technique differs from reactive evaporation in that, in addition to the reactions that occur at the substrate, reactions also occur on the target surface and in the gas phase at high reactant-gas working pressures.

In all sputter deposition processes, a target with appropriate composition must be fabricated. For metal targets, vacuum casting (see below) is typically used, while hot pressing and slip casting are used for the formation of high–melting-point metal and ceramic sputter targets. An alternative approach is to produce a mosaic target that consists of tiles of two or more elements. The ratio of the area of tiles of one kind to tiles of another provides compositional control.

1.2 Chemical Vapor Deposition

This technique is based on the use of chemical reactions to deposit solid films or coatings from gaseous precursors on heated substrates. The gas-phase reactions that result in the deposition of a solid include thermal decomposition and reduction of fluorides, chlorides, bromides, iodides, hydrides, organometallics, and hydrocarbons. These reactions are used to deposit all types of materials including metals, alloys, intermetallics, oxides, carbides, and nitrides. All CVD (chemical vapor deposition) (*q.v.*) processes take place in an enclosed CVD reactor vessel at either atmospheric or low-pressure conditions. The details of each process are extremely involved and required detailed descriptions of the thermodynamics, kinetics, and gas-flow dynamics to model fully. The CVD system includes a vessel, often made of quartz, which is used to contain the reactive gases and the substrate, a gas delivery system, and a source of heat. Some methods use a simple resistively heated tube furnace and are called hot-wall processes, while others use radiant (IR lamp heating) or induction heating and are called cold-wall systems. In the cold-wall systems, the substrate sits on a susceptor (often made of graphite) that absorbs radiant energy from the heat source.

Four examples of atmospheric-pressure CVD reactor vessel configurations are shown in Fig. 6 and include vertical tube, horizontal tube, turret, and barrel reactor types. Gases are brought into the vessel in a controlled ratio at a fixed flow rate. The reaction occurs in the hot zone of the vessel and the resultant effluent gases are passed out of the vessel. The design of a CVD system is determined in part

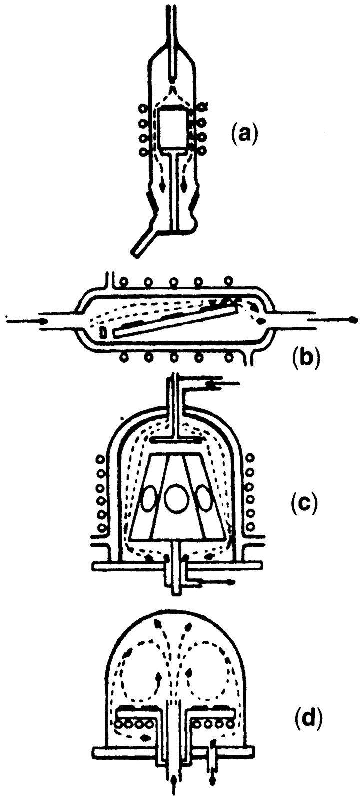

FIG. 6. Four examples of CVD reactor configurations: (a) vertical tube; (b) horizontal tube; (c) turret; and (d) barrel type.

by the physical state of the starting materials, which can be gases, volatile liquids, or volatile solids at room temperature. The use of liquids and solid sources complicate CVD reactor design, since these materials must be vaporized at the source and must remain in the vapor state as they are transported to the deposition zone.

A technologically important CVD process that is widely used (in one form or another) is the deposition of silicon from silane at either atmospheric or low pressure. The reaction can proceed *via* pyrolysis (thermal decomposition) from silane alone but is most often accompanied by a carrier gas. The reaction is written as

$$SiH_{4(g)} \leftrightarrow Si_{(solid)} + 2H_{2(g)}, \tag{4}$$

and the deposition occurs by simply passing $SiH_{4(g)}$ (diluted with a carrier gas such as H_2 or N_2) over a heated substrate.

Most useful CVD systems are of the open variety where there is a constant flow of fresh reactant gas over the substrate. There are two principal types of open reactor systems that are categorized on the basis of the total pressure in the chamber during deposition. The atmospheric system has the advantage of being relatively simple without the need for elaborate vacuum pumping systems but must be carefully designed with careful consideration given to the dynamics of the gas flow. In atmospheric deposition systems, the growth rate is typically limited by the transport of reactants to the wafer surface. Thus, the positioning of the substrate with respect to the gas flow is important. This is not the case for low-pressure CVD systems (LPCVD), since at relatively modest pressures, say 100 Pa, the gas-phase diffusivity increases by 1000-fold since the diffusivity is inversely proportional to pressure. This allows the wafers in a LPCVD deposition process to be stacked relatively close to one another without regard to the gas-flow conditions. Thus, some of the benefits associated with the increased cost of an LPCVD system are larger loads and better uniformity.

The kinetics of the CVD process can be modeled by the series of sequential steps shown schematically in Fig. 7. The steps are

1. transport of reactants to the deposition region;
2. transport of reactants by diffusion from the main gas stream through a stagnant boundary layer to the growth surface;
3. adsorption of reactants on the growing surface;
4. surface reaction: chemical decomposition, surface migration, and site incorporation;
5. desorption of by-products;
6. transport of by-products through the stagnant boundary layer and back into the main gas stream; and finally,
7. transport of by-products away from the deposition region.

The steps listed above can be further subdivided to accommodate various additional reaction kinetic models. Whichever step is the slowest will determine the overall rate of the deposition process and is known as the "rate-limiting" step. When a process is limited by steps as in **3–5** listed above, it is known as surface-reaction-rate limited while all other steps represent mass-transport limiting steps.

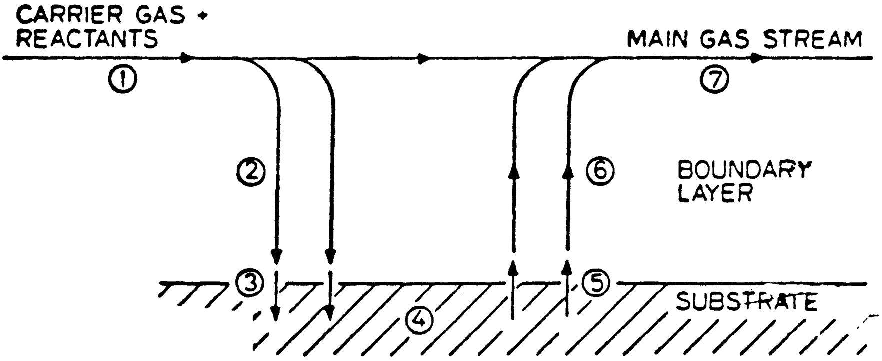

FIG. 7. Schematic view of the CVD deposition mechanism showing seven sequential steps, any one of which may be rate limiting.

To illustrate the notion of the rate-limiting step, consider the case where the input reactants to the CVD process are relatively low, the gas velocity (i.e., flow rate) is low, and the temperature (and hence the surface reaction rate) is high. Under these conditions, the initial mass transport step (step **1** above) will control the rate of the deposition process. In most cases, mass transport conditions will eventually dominate as the temperature of the reaction is increased since the activation energy for surface reaction mechanisms is larger than the activation energy for mass transport processes. Thus, at higher reaction temperatures, surface reactions proceed rapidly and mass transport becomes rate limiting. If, on the other hand, the concentration of the reactants and the gas velocity is high but the reaction temperature is low, surface kinetics usually take over as the controlling step. In general, low temperatures, and low pressures where gas-phase diffusivity is large (as in LPCVD), tend to favor surface-reaction rate control.

1.3 Ion Implantation

Ion implantation is extensively used for the introduction of impurities into the near-surface region of a single-crystal electronic material such as Si. It is also now being used in the modification of surface properties of structural and magnetic materials (see ION-BEAM MODIFICATION OF MATERIALS). Ion implantation involves the introduction of atoms into a substrate with sufficient energy to penetrate beyond the surface. To introduce specific elemental impurities into the near-surface region of a substrate, the species of interest are accelerated through a high accelerating potential, pass through a mass spectrometer that selects only those ions with a characteristic energy, and are directed into the substrate. Atoms from the lattice of the target substrate are displaced from their lattice positions through momentum transfer processes between the energetic incident impurity ions and the atoms of the substrate. Under conditions that depend on the substrate and ion species, implant dose, energy, and substrate temperature, an amorphous region can be created by this process. For example, the surface of a Si wafer can be amorphized by implanting 50- to 200-keV $^{28}Si^+$ at a dose greater than $10^{15}/cm^2$ at room temperature or below. At 200 keV, this process amorphizes the Si surface to a depth of about 380 nm, where the peak of the implant dose (R_p) is located at a depth of 266 nm, and the projected deviation (ΔR_p) is about 78 nm. Immediately below the amorphous-crystalline interface, a narrow region of crystalline material exists that has suffered lattice damage due to ions that, although they are near the end of their range, still possess sufficient energy to cause lattice displacements. This damage and the amorphous region can be fully recrystallized by a relatively low-temperature annealing procedure. If the implantation is performed at elevated temperature, or if the ion dose is relatively low, then the displaced substrate atoms can refind their lattice positions during the implantation process.

2. LIQUID-STATE PROCESSING

Preparation of solid materials from the liquid phase is widely applied to metal and semiconductor materials and includes such technologically important processes as ingot casting, continuous casting, foundry processes, single-crystal growth, and fusion welding. This approach to the shaping of metals into useful objects has been in use since 4000 B.C. and continues to this day to be the most versatile and flexible process available for forming complex shapes out of metal. In fact, nearly all metals are cast at some point in their processing sequence. An example is the process of steel making (see STEEL). To make steel (i.e., Fe-C alloys), iron ore [hematite (Fe_2O_3), limonite ($Fe_2O_3 \cdot xH_2O$), or magnetite (Fe_3O_4)] is chemically reduced to form impure iron metal. The reduction is performed in a blast furnace that produces an iron metal (known as pig iron) that contains about 7 wt % impurities such as C, S, Mn, S, and P. The formation of steel from pig iron involves the refining of the metal to remove undesirable impurities such as S and P, and the purposeful introduction of alloying elements to improve the physical properties of the metal. Finally, the steel is continuously cast into billets or ingots, which are used for further (solid-state) forming processes.

The solidification process requires that the liquid be cooled to below its equilibrium melting temperature T_m enough to overcome kinetic barriers to the formation of solid. The magnitude of the driving force necessary to create a small volume of solid (called a nucleus) in a liquid is given by nucleation theory, which balances the decrease in the Gibbs free energy (G_v) of the system due to the formation of solid against the energy increase due to the formation of a solid–liquid interface (γ_{SL}). The interfacial energy contribution represents a barrier to the formation of the solid phase from the liquid, which must be overcome by providing additional driving force through supercooling (cooling the liquid below its natural melting point). The degree of supercooling required to nucleate a solid in the liquid depends mostly upon the availability, and effectiveness, of heterogeneous nucleation sites. These sites reduce the surface energy cost when the solid forms from the liquid and can be purposely introduced in a process known as seeding.

For homogeneous nucleation (in the absence of heterogeneous sites) of spherical particles with isotropic solid–liquid surface energy, the degree of undercooling (and hence the amount of driving force) is readily calculated by minimizing the free energy of a particle of radius r, which is found by summing the volume free-energy decrease (volume × ΔG_v) and the surface free-energy increase (surface area × γ_{SL}):

$$\Delta G_r = \tfrac{4}{3}\pi r^3 \Delta G_v + 4\pi r^2 \gamma_{SL}, \tag{5}$$

where ΔG_r is the total free energy of a particle of radius r formed from a liquid that is undercooled. The effect of undercooling on the volume free energy can be seen from

$$\Delta G_v = (G_L - G_S) = \Delta H_f \Delta T / T_m, \tag{6}$$

where ΔH_f is the latent heat of fusion, $\Delta T = T_m - T$ is the undercooling, and T_m is the melting temperature. Note that the interfacial term (which is positive) increases as r^2, while the volume term (which below T_m is negative) increases as r^3. This implies that there is a critical particle radius r^* above which a particle is stable (i.e., will tend to increase in size) and below which it is not (i.e., would remelt into the liquid if such a sized particle were to form in the melt at a given undercooling). This critical radius is easily found by differentiating the above expression and setting the result equal to zero:

$$r^* = 2\gamma_{SL}/\Delta G_v, \tag{7}$$

which is given in terms of undercooling by

$$r^* = 2\gamma_{SL} T_m / \Delta H_f \Delta T \tag{8}$$

where γ_{SL} is the solid/liquid free energy.

Expression (7) reveals that the critical nucleus radius decreases with increasing undercooling. The larger the critical radius, the lower the probability that a cluster of this size will form as a result of random thermal motion of the atoms of the liquid. The liquid is thought of as containing particles that are constantly changing their positions and in which transient clusters form and disappear that briefly have the crystalline arrangement. The maximum-sized cluster that is likely to form in a given volume of liquid increases with decreasing temperature since the solid arrangement of atoms is more favored as the temperature is lowered. As the undercooling increases, the critical particle size decreases and the maximum-sized cluster that is likely to form in a liquid of a fixed volume in reasonable time increases. The formation of clusters of atoms with dimension r in the liquid occurs with a probability that is exponentially dependent on the energy of the particle (ΔG_r) so that the number of clusters of size n_r in a volume containing n_0 atoms is given by

$$n_r = n_0 \exp(-\Delta G_r/kT). \tag{9}$$

As the undercooling increases (i.e., T decreases), the maximum-sized cluster that has a reasonable likelihood of forming in the volume of liquid increases while, at the same time, the critical radius r^* decreases with increasing undercooling. This is shown in Fig. 8, which shows the critical particle size r^* and the maximum cluster size $r_{\max}$ that has reasonable probability of being found in a given volume of liquid as a function of undercooling. At ΔT_N, the critical radius equals the size of the largest probable cluster found in the volume and rapid nucleation takes place.

The undercooling predicted by the homogeneous nucleation theory for typical metals can be hundreds of degrees Celsius, but this behavior is only seen under carefully controlled experimental conditions. In reality, the solid does not typically nucleate homogeneously in the melt but at energetically more favorable heterogeneous nucleation sites.

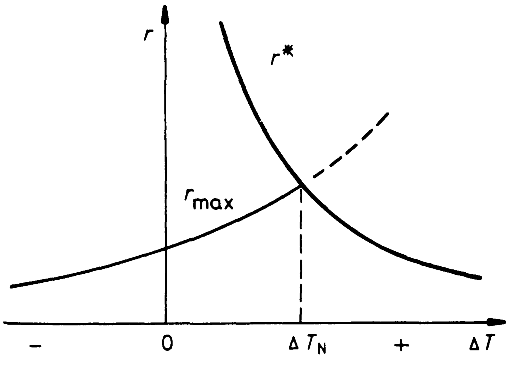

FIG. 8. Schematic plot showing the effect of undercooling (ΔT) on critical radius size (r^*) and the maximum cluster size likely to be found in a fixed volume of liquid (Porter and Easterling, 1987).

After nucleation has occurred, the rate at which the solidification proceeds and solid–liquid interface planarity are controlled by the diffusion of heat or impurities away from the solid–liquid interface. The heat that is present at the solid–liquid interface is a result of the latent heat that is released during solidification and the heat that is thermally conducted toward the solid–liquid interface from the melt. In most systems undergoing solidification, there are impurities and alloying elements that must be rearranged so that the difference in composition between solid and liquid in thermodynamic equilibrium can be accommodated. This compositional difference is given for a two-component system by the schematic binary phase diagram shown in Fig. 9. Consider an imaginary binary system of A and B species. Above the melting point of A, T_m^A, the two species are completely miscible and exist as a single liquid condensed phase. If a liquid of composition X_0 is cooled, the first solid will appear at $T = T_1$. The compositions of the liquid and solid phases that coexist at this temperature are different: the liquid has composition X_0 while that of the solid that is in thermodynamic equilibrium with this liquid is kX_0. As the temperature of the system is lowered further, say to $T = T_2$, both the compositions of the liquid and the solid change as defined in Fig. 9 by the solidus line for the solid and the liquidus line for the liquid. Thus, these lines give the equilibrium compositions of the coexisting liquid and solid phases as a function of temperature. The compositional difference between solid and liquid that coexist at equilibrium results in both macroscopic and microscopic chemical inhomogeneities because solid-state diffusion rates are too slow to establish equilibrium in the bulk of the crystal.

The final microstructure of a solid formed by solidification depends on heat flow, impurity diffusion, convective mixing in the remaining liquid, and the imposed cooling rate. Detailed description of these processes is beyond the scope of this article; however, there are three situations that will be considered briefly below. First, typical casting processes have a relatively slow cooling rate ($<10^2$ °C/s) and shallow temperature gradients, which result in both microscopic and macroscopic compositional gradients. Relatively local

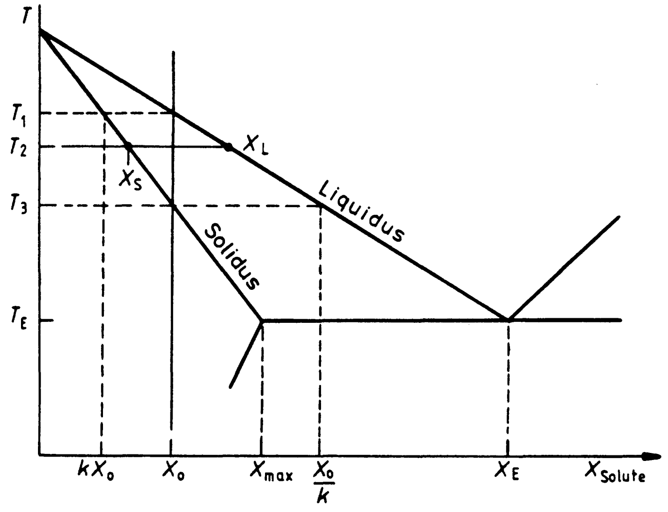

FIG. 9. Binary phase equilibrium diagram (temperature versus composition) showing the equilibrium concentration of solute when solid (X_s) and liquid (X_l) coexist at a fixed temperature (Porter and Easterling, 1987).

compositional gradients (over 10 to 100 μm) appear in the final microstructure of a cast object as *dendrites*. Dendrites are rodlike and armlike features that project into the solidifying liquid as shown in Fig. 10. The cooling rate and rate of diffusion and convectional mixing in the liquid determine the dimensions of these features, which can range from <1 to 100s of microns in size. Unfortunately, solid-state diffusion alone is usually an impractical method for homogenizing the compositional variations in cast alloys.

A second scenario is discussed below in which crystallographically textured and single-crystal materials are grown. This microstructure results when temperature gradients at the solid–liquid interface are steep and a single-crystal seed is employed. In this case, the liquid is not cooled below T_m except at the solid–liquid interface where the amount of undercooling is extremely small. Finally, extremely rapid cooling (10^3 to 10^6 °C/s) dramatically decreases compositional segregation and is used for the synthesis of powders with uniform composition. Even faster cooling (10^6 to 10^{10} °C/s) of some alloys results in the formation of amorphous metals. Metallic alloys that can be cooled so that they are solid but retain the liquid amorphous state usually have a temperature-composition phase diagram that is characterized by an alloy solidification temperature that is much lower than either of the pure components (known as a deep eutectic phase diagram).

2.1 Casting Methods

The basis for all casting processes is the feeding of molten metal into a cavity of the required shape (a mold) followed by cooling to produce a solid object. In most casting methods, the mold walls have a large number of heterogeneous nucleation sites and solidification proceeds close to (but below) the melting point by heat transfer through the mold walls to the exterior. The resulting microstructure is often described as columnar and may contain large compositional variations. These compositional variations may be neglected, or they may be eliminated by further processing. For example, the solid-state process of forging can be applied to cast ingots to breakup these compositional inhomogeneities.

The various casting processes (and there are many) differ principally in the way in which the mold is formed. In some cases, a nonreusable mold is created for each casting and is subsequently broken up to remove the cast object. The two most common forms of casting processes are called *sand casting* and *investment casting* (*lost wax*) and are described below. In other cases, a permanent mold is used repeatedly for a succession of casts and the casting is removed after each pouring without damage to the mold. A common example is called *die casting*. In all cases, some provision for feeding metal into all parts of the mold is required. All alloy castings processes produce materials with microscopic and macroscopic compositional variations that can affect the mechanical strength and performance of the cast object.

2.1.1 Sand Casting To form an object using the sand-casting approach a *pattern* of the object to be cast is manufactured using an easy-to-work-with medium such as a soft wood or plastic. The pattern is shaped into an exact model of the desired object. A mold is then made by packing sand around the pattern. The mold is typically made in two parts, an upper section, which is known as the *cope*, and a lower part, which is known as the *drag*. The two parts together contain the full mold, and the container holding the sand is called the *flask*. Where the cope and drag meet is called the *part line*. If the final form is to have a hollow section, separate mold parts known as *cores* are made and placed inside the cavity left by the pattern. A pathway is created to feed the liquid to all parts of the mold cavity. This pathway is called the *gating* system and consists of *sprues*, *runners*, and *risers* which, respectively, feed the molten metal, distribute the metal laterally, and provide liquid to replace the volume lost because of the decrease in molar volume of the solid compared to the liquid. A schematic view of the mold along with representations of the various parts of the system is shown in Fig. 11. After the liquid is completely solidified, the rough casting is removed from the mold, any cores that were included are knocked out, and the risers, sprues, and runners are cut off.

During the solidification process, the direction of motion of the solid–liquid growth front is controlled by carefully designing the mold so that the direction of heat flow leaves

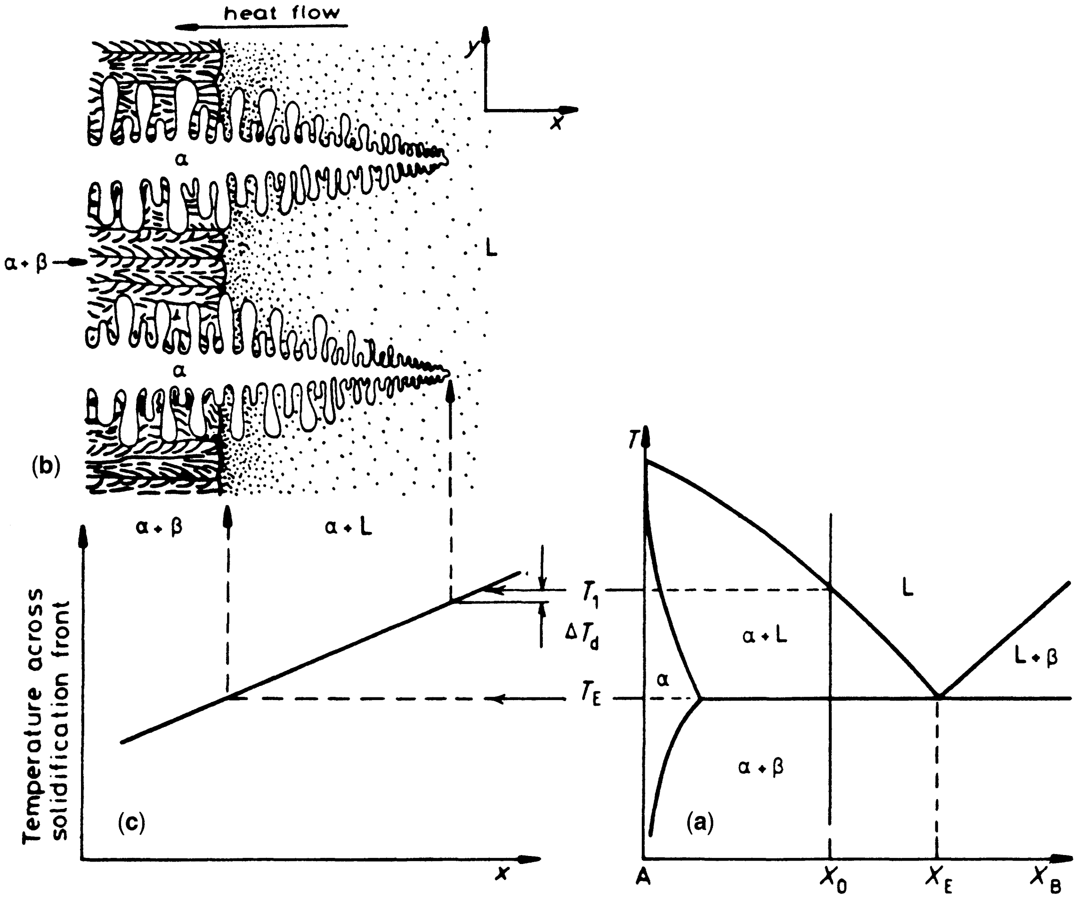

FIG. 10. Schematic view showing the formation of dendrites in a eutectic alloy of composition X_0 (hypoeutectic composition) (Porter and Easterling, 1987).

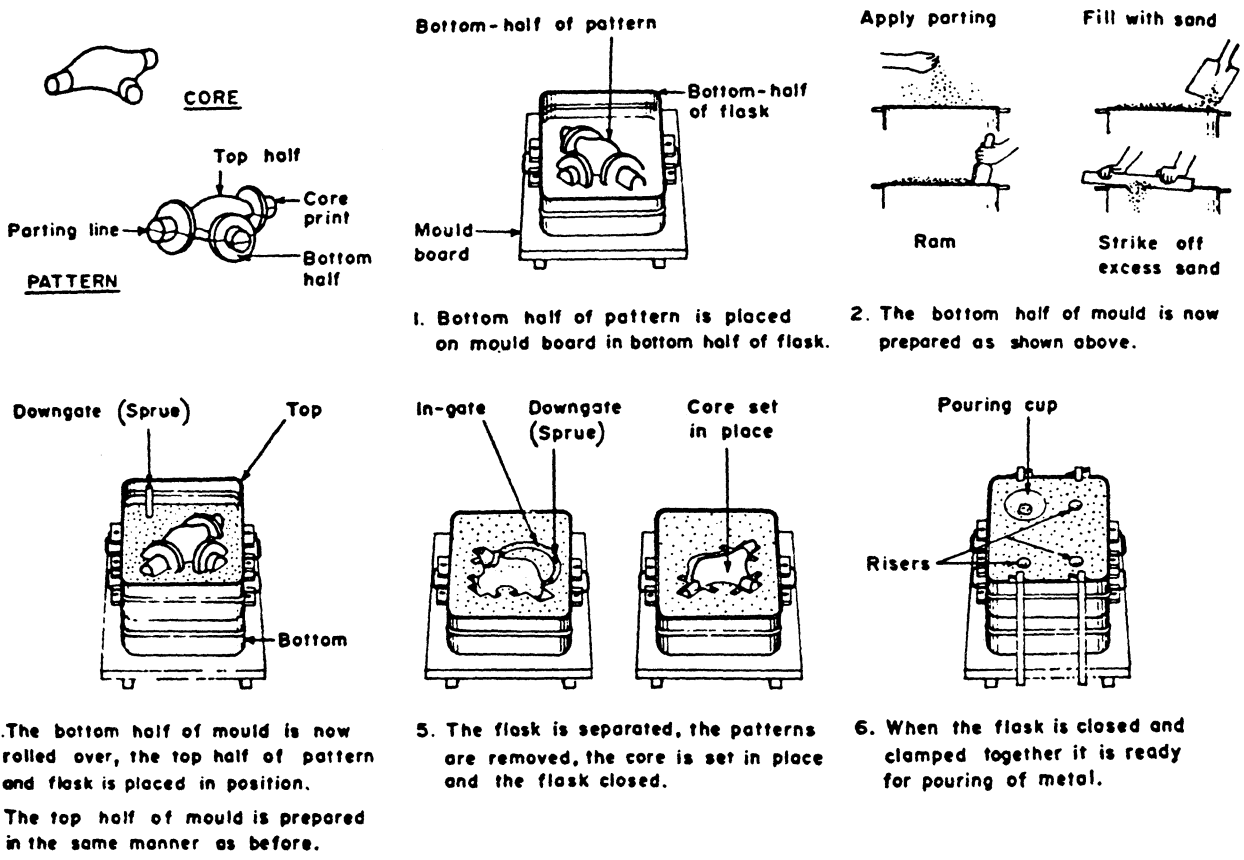
CORE
Top half
Core print
Parting line
Bottom half
PATTERN
Bottom-half of pattern
Bottom-half of flask
Mould board
1. Bottom half of pattern is placed on mould board in bottom half of flask.
Apply parting
Fill with sand
Ram
Strike off excess sand
2. The bottom half of mould is now prepared as shown above.
Downgate (Sprue)
Top
Bottom
3. The bottom half of mould is now rolled over, the top half of pattern and flask is placed in position.
4. The top half of mould is prepared in the same manner as before.
In-gate
Downgate (Sprue)
Core set in place
5. The flask is separated, the patterns are removed, the core is set in place and the flask closed.
Pouring cup
Risers
6. When the flask is closed and clamped together it is ready for pouring of metal.

FIG. 11. Sequence of steps in the formation of a mold in the sand-casting process (Davies, 1973).

no region of liquid that is isolated from the riser. Metal inserts called *chills* are sometimes included in the mold to promote heat extraction and solidification in some parts of the mold, while insulating materials are used to delay the onset of freezing in regions of the mold that would leave other sections isolated from the risers. The careful design of the mold with respect to heat flow is essential for the making of pore-free castings. Corners, for example, are hot spots, which commonly cause casting defects that result from slower heat transfer at internal corners compared to straight runs.

The advantages of sand casting are that it is cheap, it is simple, it is very flexible from a design viewpoint, and it can be used to manufacture objects over a wide range of size scales from about 10 g to 1 ton. Its disadvantages include poor surface finish and low dimensional tolerances, making thin sections and intricate shapes impossible to produce.

2.1.2 Investment (Lost-Wax) Casting Investment (also known as lost-wax) casting is a technique that has been used for centuries for the fabrication of decorative objects. Today it is used when the as-cast form must have fine dimensional tolerances. An example of an industrial application of this method is in the fabrication of some types of aircraft turbine blades. The process for creating a lost-wax mold and casting an object is described briefly in what follows.

An expendable pattern is manufactured using wax. If the pattern is to be used repeatedly, it is often economical to fabricate a mold that can be repeatedly used for casting the wax pattern in a process similar to die casting (described in Sec. 2.1.3). The wax pattern in its final shape is coated with a refractory ceramic slurry in a process known as investing. The ceramic mold is fabricated by repeatedly dipping the pattern into a ceramic slurry (often plaster), allowing it to set between dips. A fine-particle slurry is used for the first dipping in order to give a smooth surface to the cast object. A coarser slurry is used in later dippings in order to strengthen the mold. Finally, the ceramic shell is dried and fired in a furnace to melt and/or volatilize the wax pattern so that a cavity is left in the ceramic mold. The molten metal is poured into the mold and, after solidification, the ceramic shell is broken up to recover the casting. Steps in the production of a ceramic shell investment casting are shown in Fig. 12. As the description implies, this process requires a large investment in the preparation of the mold and pattern, which limits the use of this process to large–value-added objects like jewelry or aircraft parts.

The expense of the mold-making process can be reduced on a cost-per-cast-unit basis by creating many different objects (or many copies of the same object) in wax and then attaching these smaller features *via* runners to a central sprue. In this way, a single investment will yield many castings. Preformed cores can be used if hollows are needed as part of the cast structure. For example, cooling channels are created in turbine blades by incorporating into the wax pattern silica rods, which are chemically dissolved using HF acid after the blade is cast.

2.1.3 Permanent-Mold Castings Processes that make use of metal molds and simple hydrostatic pressure to introduce the liquid metal are called permanent-mold castings. The mold must be designed in two or more parts that free the casting when separated. Processes that use metal molds are limited to low–melting-point alloys such as those that are based on Zn, Al, Mg, and Cu (leaded yellow brass). The molds are economical since they can be used for up to 100 000 castings. Consequently, these permanent-mold processes are commonly used for the fabrication of metal parts for almost every commercial application such as automobile parts (e.g., carburetor) and camera bodies.

When molten metal is forced under pressure into a steel mold (known in this process as a die) the process is known as die casting. As shown in Fig. 13, the molten metal enters the die under pressure when a plug or valve is released. The rapid introduction of the metal into the mold combined with preheating of the die assembly helps prevent nonuniform heat flow and the consequent generation of casting defects. Injection of the liquid also allows much thinner-walled castings to be formed. For example, cast Al can be produced with a 0.8-mm minimum wall thickness by die casting. Once solidified, the die is separated at the part line and the object is forced out of the mold with pusher rods as is shown in Fig. 13.

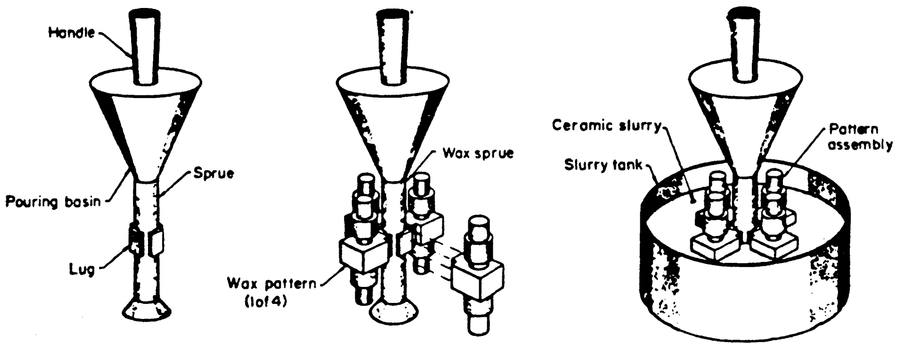
Handle
Sprue
Pouring basin
Lug
Wax sprue
Wax pattern (1of4)
Ceramic slurry
Slurry tank
Pattern assembly

(a) Wax sprue pattern

(b) Pattern assembly (wax patterns attached to wax sprue)

(c) Pattern assembly dipped in ceramic slurry

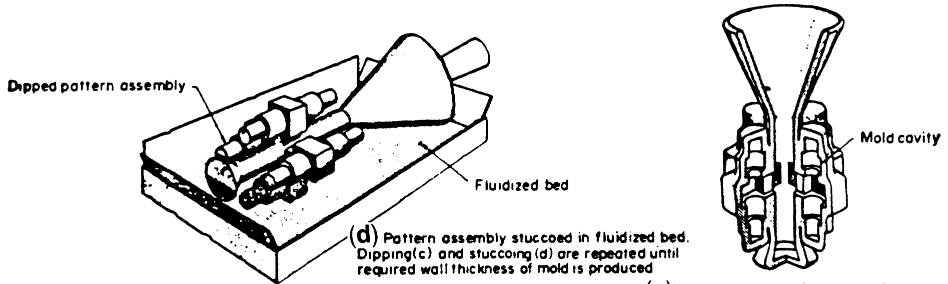
Dipped pattern assembly
Fluidized bed
Mold cavity

(d) Pattern assembly stuccoed in fluidized bed. Dipping(c) and stuccoing(d) are repeated until required wall thickness of mold is produced

(e) Completed mold after wax pattern has been melted out (mold shown in pouring position)

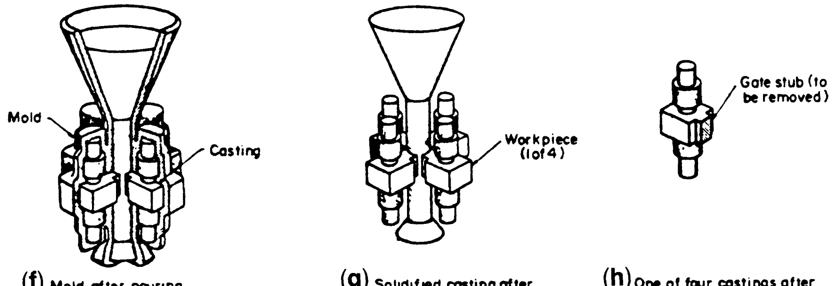
Mold
Casting
Workpiece (1of4)
Gate stub (to be removed)

(f) Mold after pouring

(g) Solidified casting after mold has been broken away

(h) One of four castings after removal from sprue

FIG. 12. Sequence of steps in the production of an investment mold and casting (Davies, 1973).

2.2 Oriented or Single-Crystal Growth

The principal driving force for the development of single-crystal growth techniques (see CRYSTAL GROWTH) continues to be the demand for solid-state electronic and optoelectronic materials. The largest commercial demand is for single-crystal Si for use in integrated-circuit devices. Compound semiconductors (most notably GaAs) make up about 2% of current semiconductor single-crystal growth and are used for high-speed, laser, and detector device applications. In the case of electronic device substrates, wafers are typically cut from cylindrical boules grown by the Czochralski method. Single-crystal (or strongly oriented) growth is also used when high-temperature creep (time-dependent permanent deformation of a loaded structure) resistance is desired. An example of this is high-performance turbine blades, which operate at temperatures greater than half the melting point of the material on the Kelvin scale. Since grain boundaries provide rapid diffusional paths, their elimination drastically decreases the creep rate of materials. This effect does not require perfect

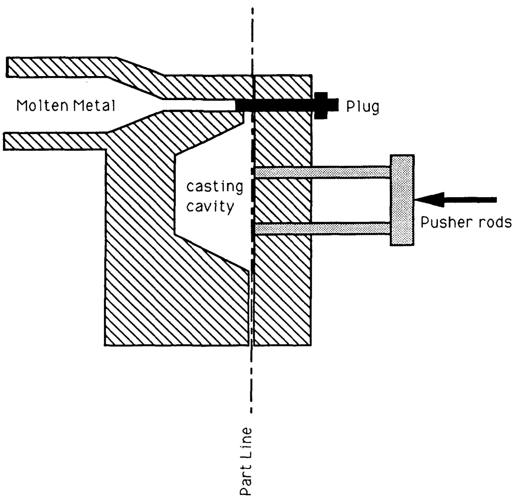

FIG. 13. Cross-sectional view of a die casting. When the plug is removed, the molten metal is forced under pressure into the steel mold. After solidification, the die is parted at the part line and pusher rods remove the casting.

single crystals and, in many applications, highly textured or directional growth is adequate for the reduction of creep.

2.2.1 General Single-Crystal Growth Considerations Most single-crystal growth processes begin with an oriented seed crystal (i.e., a crystal that contains only one crystallographic orientation) that is appropriately oriented and provides the template for the arrangement of atoms during controlled crystallization of the melt. The temperature of the molten material is carefully controlled so that undercooling, and spontaneous nucleation of randomly oriented crystals, is avoided. All methods are based on the simple principle that the extraction of the latent heat of solidification must be achieved without allowing the melt to supercool enough for random nucleation of solid to occur in the melt away from the seed crystal. This is usually done by extracting heat by conduction through the solidified single crystal. In practice, a heat sink is provided at the seed end of the crystal, and a heat source is provided at the melt so that the heat flow away from the interface through the solid is exactly balanced by the heat flow from the liquid plus the latent heat of solidification that is generated at the interface.

Ultrahigh-purity feed material is required for crystals grown for electronic applications. In the case of Si, the refining process proceeds as follows. First quartz (SiO_2) is placed along with carbon in the form of coal or wood in a submerged electrode arc furnace. Several reactions take place under these conditions, including

$$SiC_{solid} + SiO_{2(solid)} = Si_{(solid)} + SiO_{(gas)} + CO_{(gas)}. \tag{10}$$

This process produces metallurgical-grade Si, which is used as the feedstock for further chemical-refining processes. The first step is the reaction of Si powder with HCl gas to form trichlorosilane ($SiHCl_3$) *via* the following:

$$Si_{solid} + 3HCl_{gas} = SiHCl_{3gas} + H_{2gas}. \tag{11}$$

At room temperature, the $SiHCl_3$ is a liquid and can be purified by the fractional distillation method. The distilled $SiHCl_3$ is then reacted with H_2 *via*

$$2SiHCl_{3gas} + 3H_{2gas} = 2Si_{solid} + 6HCl_{gas}, \tag{12}$$

and the solid thus produced becomes the feedstock for the growth of single-crystal Si.

2.2.2 Structural Components: Bridgman Technique Single-crystal growth for the preparation of structural components (or highly textured, directionally solidified crystals) require that the as-grown object be close to the final shape. The most common example of the application of the techniques used to form highly textured or single crystals of arbitrary shape is in the fabrication of high-performance turbine blades. These objects are grown from high-performance, Ni-based alloys using the Bridgman technique in which a boat (or crucible) is created with the appropriate shape and is filled with molten metal. The boat is slowly translated out of the hot zone of the furnace so that solidification begins at a seed crystal. If a particular orientation is not required, it is possible to form a single crystal by using a boat with a pinched end. If the pinched end is slowly passed through the temperature gradient, a single crystal may nucleate in the heterogeneous site provided. This single crystal then acts as a seed crystal for the rest of the solidification process. A schematic view of the Bridgman technique is provided in Fig. 14.

2.2.3 Float-Zone (Zone-Refining) Method

In this technique, a rod-shaped sample is passed through a localized heat source (e.g., a rf coil or a focussed electron beam), which melts a section of the rod. The setup is shown schematically in Fig. 15. The molten zone is held in place between the solid region above and below by the surface tension of the liquid. The rod is slowly lowered through the heat source so that melting occurs on one side of the molten zone while solidification occurs on the trailing side. This method can be used to produce large grained, textured, or single-crystal metals but is primarily used for the creation of extremely pure materials by zone refining.

The principle of zone refining can be seen by again referring to the temperature–composition phase diagram presented in Fig. 9. The thermodynamic equilibrium composition of the binary liquid in equilibrium with the solid at a particular temperature is given by the tie line that connects the liquidus and solidus curves as shown in the figure. Note that the composition of the liquid is richer than that of the solid which means that under equilibrium conditions, the solid that freezes from a liquid with composition x_l will contain a lower concentration x_s than the starting liquid. For this reason, the solidifying solid rejects impurities into the molten zone, which tend to remain in the liquid as the molten zone moves up the length of the rod. Only the very last solid to form contains a higher concentration of impurities than that of the initial rod. This end of the rod may be cut off and the process repeated numerous times to decrease further the impurity concentration in the rod.

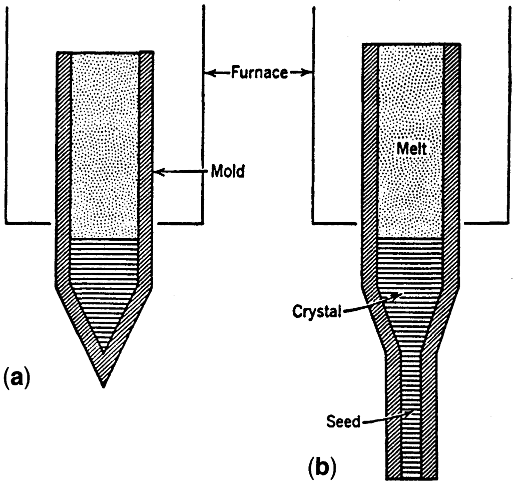

FIG. 14. Schematic view of the Bridgman technique for oriented crystal growth. (a) The pinched end and (b) the seed methods (Flemings, 1974).

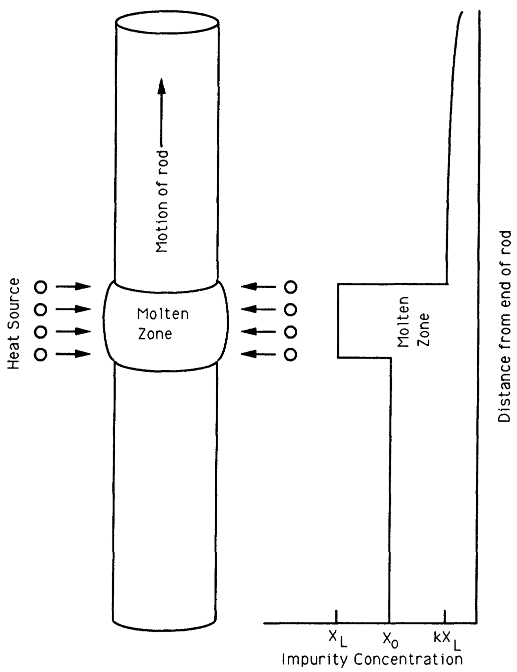

FIG. 15. A schematic view of float-zone crystal growth and refinement. In this operation, the rod is slowly lowered through a localized heat source that melts a region of the rod. The molten zone is supported by the surface tension of the liquid.

2.2.4 Semiconductor and Optical Materials

The vast majority of single-crystal material grown for the semiconductor industry is grown by the Czochralski crystal-pulling method. The setup is shown schematically in Fig. 16. A seed crystal of suitable crystallographic orientation is attached to a *puller* and is dipped into a melt that is contained in a *crucible*. The seed is rotated (to ensure uniform interfacial temperature and composition) and slowly withdrawn from the liquid so that the rate of crystal motion equals the rate of crystal growth. The rate of crystal growth, in turn, is determined by the rate at which the latent heat of solidification is removed through the crystal, seed, and puller.

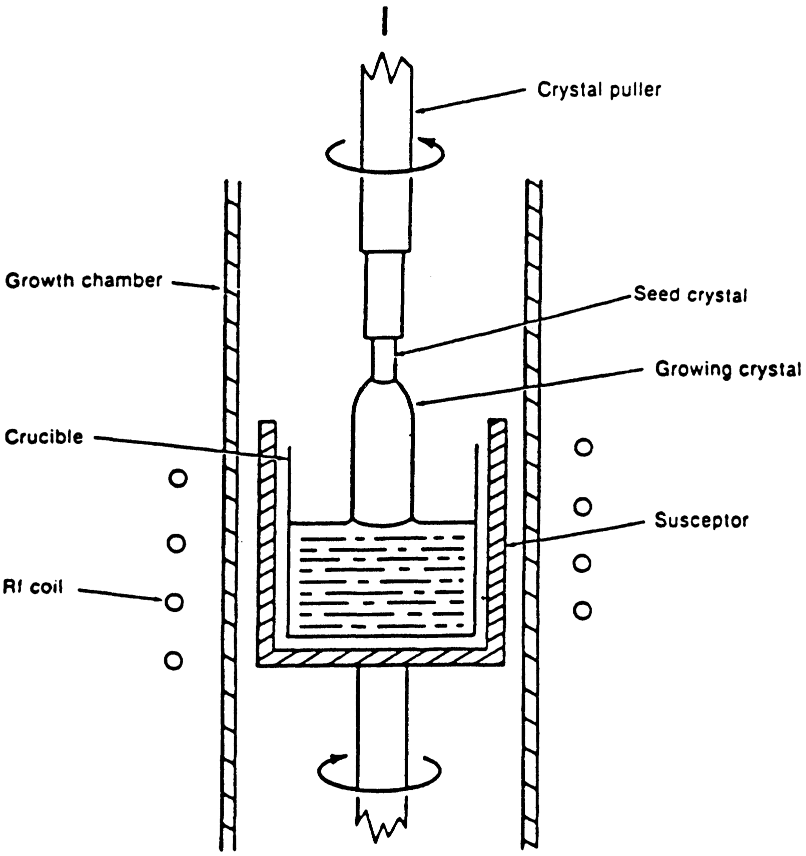

FIG. 16. The Czochralski method for the growth of single crystals from the melt.

The solid single crystal that is drawn from the melt is called a *boule.*

The latent heat of solidification must be carried away from the solid–liquid interface *via* the grown crystal. The temperature gradient down the length of the crystal is more or less fixed. Since heat flux is given by $q = -kdT/dx$ and since the temperature gradient (dT/dx) and the thermal conductivity (k) are fixed, so too is the heat flux (Q). This means that the latent heat evolved by the solidification process must equal Q at steady state. If it is greater, the amount of solidifying material decreases; if it is less, the amount increases so that steady state is reached. Consequently, the crystal growth rate for a boule of fixed diameter is limited by the rate at which latent heat is carried away from the solid–liquid interface.

The composition of the as-grown semiconductor crystal must be carefully controlled. For example, the composition of the crystal can be affected by contamination and, in compound crystal growth, by the preferential loss of a more volatile component. Furthermore, wafers that are cut from different parts of the boule must have similar dopant concentrations to allow reproducible characteristics in electronic devices made from them. Contamination other than that from the initial feed material comes from the atmosphere and/or the crucible. While it is easy to use a vacuum or inert gas to control the atmosphere above the melt, it is necessary (except in float-zone growth) to contain the melt in a crucible, which can contaminate the melt. Preferential evaporation of one or more species from the melt can be suppressed by applying an over-pressure or by providing an encapsulant. In the case of GaAs crystal growth by the Czochralski method, As tends to evaporate from the melt leaving it rich in Ga and deficient in As. This is eliminated by providing an atmosphere of As gas above the melt at a pressure that is selected so that the vapor and melt are in thermodynamic equilibrium. Compositional nonuniformities also can result from the rejection of solute at the solid–liquid interface. As the crystal is drawn from the melt, the composition of the melt changes as a result of the buildup of rejected solute. When dopants (*n* or *p* type) are added to the melt, their concentration is a function of position in the boule, with the highest dopant concentration in the material that was the last to be pulled from the melt.

A crystal grown carefully from the melt by the Czochralski method should be free of crystallographic defects other than point defects, and the number of line defects (e.g., dislocations, which are linear discontinuities in the periodic crystal lattice) should be minimal. There are two possible mechanisms for the introduction of dislocations that must be carefully monitored during single-crystal growth. First, they can propagate from the seed crystal, and second, they can be introduced as a result of large temperature gradients in the boule during growth or thermal shock to the boule after growth. It is possible to eliminate dislocations that propagate from the seed by causing the diameter of the crystal to decrease as it is drawn from the melt so that dislocations can run easily to the nearby outer surfaces where they terminate. This process of necking down the crystal diameter by increasing the pull rate is done shortly after the crystal growth begins. Since energy considerations demand that dislocations minimize their length in a crystal, there is a tendency for the dislocations that propagate from the seed to be drawn to the surface. Thus, if the crystal is made narrow enough, the dislocations that start in the seed will be grown out of the crystal.

2.3 Rapid Solidification

When the rate at which a liquid is cooled is slow relative to the rate of diffusional and convective mixing in the liquid, then the solid–liquid interface is relatively close to thermodynamic equilibrium, and the solid product is a near-equilibrium phase. As a consequence, as a multicomponent liquid is cooled through the liquidus line, there is a separation of the components in the system since the equilibrium concentration of impurity in the solid is not typically the same as that in the liquid (as shown previously in Fig. 9). Consequently, when solidification occurs on a time scale that allows equilibrium to be maintained at the solid–liquid interface, say at a rate of 1 K/s, macroscopic segregation of impurities is observed in the final microstructure. If the rate of cooling is increased so that the melt is brought into a highly supercooled state (i.e., $T < T_m$), then the microstructure of the solid is dominated by nucleation of individually oriented solid particles rather than by their growth into the liquid.

In one application of rapid solidification, small-particle-diameter powders are produced from a metal alloy melt with a high degree of chemical homogeneity. For compositional uniformity, it is necessary that the particles formed by the rapid cooling process be small relative to typical segregation distances. For example, the synthesis of 10- to 100-μm diameter powders of a particular alloy ensures that if these particles are fused together through the process of sintering (see Sec. 3.1.2), the scale of the compositional segregation remains the same as the scale of the particles.

Cooling rates of 10^3 to 10^5 K/s can be achieved by gas atomization processes in which an inert gas of N_2 or He is blown into a stream of molten metal. If ultrasonic methods are used, particle diameters as small as 20–30 μm may be achieved. More rapid cooling may be obtained by directing of a stream of molten metal onto a spinning, water-cooled, copper wheel to form 100-μm-thick amorphous metal tapes with a cooling rate of 10^5 to 10^9 °C/s. These amorphous metals find application in transformer cores and magnetic recording devices.

2.4 Electrochemical Processes

Surfaces can be coated by the process of electroplating, which is similar in concept to electrowinning of metals during the refining process. Copper is electroplated on large cathode substrates for the synthesis of high-purity Cu, which is then drawn into wire for electrical applications. Coatings of Cr, Ni, Cu, Zn, Cd, Sn, and Mo are deposited to thicknesses up to 0.25 mm for corrosion prevention, wear protection, and decorative purposes. The process is performed in an electrolyte that contains the metal ions of interest. Sulfates and nitrates are commonly dissolved in appropriate acidic baths to form the electrolyte. A potential is applied between the part to be plated (the cathode, negative potential) with respect to an anode placed in the electrolyte. The concentration of plating ions, the bath temperature, and the applied voltage must be carefully controlled in order to optimize the microstructure (planarity, porosity, adhesion to the substrate) of the deposit. Deposit uniformity is dependent on maintaining a uniform current density over the entire cathodic region to be plated. The basis for the deposition can be described using Faraday's law, which says that the amount of electrochemical reaction occurring at an electrode totals 1 g equivalent for each 96 484 A·s of electricity that passes through the electrode, where 1 g equivalent equals the atomic weight in grams of the deposited species divided by the charge on the ion. One consequence of Faraday's law is that overvoltage (i.e., a voltage larger than that needed to cause the reaction to proceed) at the electrodes must be avoided, since the application of a potential that is greater than is necessary results in increased creation of H_2 at the cathode. This has two negative effects: First, it causes the incorporation of H into the deposited metal, which in some metals results in embrittlement. Second, the creation of H_2 means that the amount of material that is being deposited is decreased since Faraday's law applies to the total electrochemical reaction (both the plating reaction and the hydrogen reaction).

3. SOLID-STATE PROCESSING

Solid-state processing generally can be subdivided into processes in which the solid undergoes plastic deformation (i.e., metal forming), materials removal (i.e., machining), or solid-state consolidation and densification (i.e., powder processing). In this section, the processes of material removal are not discussed since these methods do not generally affect the bulk microstructure of the final product. Plastic deformation, on the other hand, usually is applied to metals and results in profound changes in the microstructure and mechanical properties of the deformed material. Solids that cannot be plastically deformed at reasonable temperatures are typically prepared from powders of the appropriate composition in powder consolidation and sintering processes. These processes take advantage of high temperatures to cause mass transfer of material *via* diffusion that allows the fusing of particles and filling in of interparticle voids. In this section, powder processing and mechanical forming of materials are discussed in a materials preparation context.

3.1 Powder Processing

Powder processing is an important technology for the manufacture of useful objects from ceramic materials that, in general, do not exhibit ductility and, hence, cannot be deformed by conventional metal stamping, forging, or rolling processes. Powder-processing methods are also widely applied to high–melting-point metals (W, 3400 °C; Ta, 2996 °C; Mo, 2620 °C) and highly reactive metals (e.g., Ti and Zr) and in the preparation of alloys without the macroscopic segregation that results when casting methods are used. Other applications of powder-processing methods include the formation of porous structures for the fabrication of filters, and porous surface coatings for forming a permanent bond between natural bone and prosthetic devices (e.g., hip replacements) *via* in-growth of bone into the pores of the coating. In powder processing, a nearly fully dense object is produced from loose ceramic or metal powder by compaction of the material in a die followed by sintering (heating) the compact at high temperatures. The compaction step establishes the basic shape of the object while the sintering step densifies the material through diffusional transport of material into voids.

Traditionally, powders of naturally occurring ceramic materials in the form of clay (hydrated aluminosilicates) have been used directly for the preparation of ceramic materials. These naturally occurring minerals are layered structures with strong primary bonding within the layers but weak van der Waals bonding between layers. As a result, the addition of water to these materials makes them pliable and easily formed. Nonlayered ceramic materials are not as readily formed and require powder-processing methods and careful precompaction powder preparation. Most simple oxide powders, such as Al_2O_3, ZrO_2, ThO_2, BeO, MgO, and $MgAl_2O_4$, are formed as a product during the refinement of the metal from metal-bearing ores. For example, Al_2O_3 is formed from bauxite (a mixture of aluminum hydroxide and iron hydroxide) in the Bayer Al refinement process. More complex ceramic materials such as $PbFe_{12}O_{19}$, $ZnFe_2O_4$, and $Y_6Fe_{10}O_{24}$, and ferroelectric ceramics such as $BaTiO_3$ are formed by reaction between simpler oxides at high temperature. Nitrides such as Si_3N_4, carbides such as SiC, and borides such as TiB_2 have become interesting for high-temperature structural applications. An example of the production of one of these powders is the *Acheson* process for the synthesis of SiC powder. In this process, silica and carbon are piled around a central graphite electrode. Current is passed through the electrode to produce temperatures in excess of 2000 °C, which allows the solid state reaction $SiO_2 + 3C = SiC + 2CO$ to proceed. The central mass of SiC powder is then cleaned with HF acid to remove remaining silica and is placed in a rolling mill to be processed into powder.

Metal and alloy powders with a mean particle size of 25–100 μm are formed by atomization of the liquid metal in a stream of inert gas (as described briefly in Sec. 2.3). Some metals, and all ceramic materials, can easily be pulverized by grinding in ball mills to produce powders that have the peak of the particle size in the range 1–25 μm. Finer powders may be formed using more sophisticated powder synthesis, milling, and particle separation and sizing methods.

3.1.1 Powder Compaction The first step in every powder process is compaction of the powder to form a "green body." Compaction methods include simple die compaction where pressure is uniaxially applied using a press, isostatic compaction in which the powder is placed in a flexible mold and then hydrostatically compressed using a fluid (liquid or inert gas), or slip casting in which compaction is achieved through capillary action of a liquid. By applying pressure to a powder, the particles are forced to rearrange themselves in a way that tends to lock or key the particles together mechanically. This provides strength to the green (unheated) compact sufficient to preserve the shape imparted by the die. If the powder is a ductile metal, the particles will plastically deform during compaction and bonding will occur between particles. The final density of a cold-pressed powder and the postsintering tensile strength of the structure are dependent upon the amount of pressure applied during compaction and on the powder used.

In many powder compaction processes, organic or inorganic binders (e.g., kaolinite in the form of common clay) are added to the powder to provide additional strength to the green body. A plasticizer is also added to the powder to improve the pliability of the binder. Other additives include lubricants, which reduce the friction between particles and between the compact and the walls of the die.

Slip casting allows compaction of powders without the direct application of pressure through capillary interaction between the porous walls of the mold and a liquid slurry. In this process, the particles of the powder are suspended in a suitable liquid to form a slurry called a slip. The slip is then cast into a mold made of plaster, which contains pores that draw the liquid out of the slurry by capillary action. There are many forms of slip casting, including simple casting into a single- or multiple-piece mold and drain casting. Drain casting is a commercially important process that is used for fabricating common porcelain sinks and many other products. The method is shown schematically in Fig. 17. The slip is poured into the mold, and nearest the wall of the mold, the water is drawn into the pores of the plaster mold. The dried region of the slip consists of a relatively close-packed compact that grows as more water is wicked into the mold. When a suitable compact thickness is established, the excess liquid is drained from the mold leaving behind a powder compact shell that can then be sintered.

Uniaxial pressing involves the compaction of powder into a die by applying pressure in a single axial direction with a mechanical or hydraulic press and a suitably designed die (see Fig. 18). The powder feed may be either dry (less than 4% moisture) or, if a clay binder is used, 10 to 15% moisture. Nonuniform stress distributions in the compact during loading result from friction at the die wall and nonuniform particulate flow before and after loading (known as rebound), which may result in the formation of cracks and density inhomogeneities. Isostatic pressing allows greater flexibility in the selection of the final shape of the solid. It also results in improved

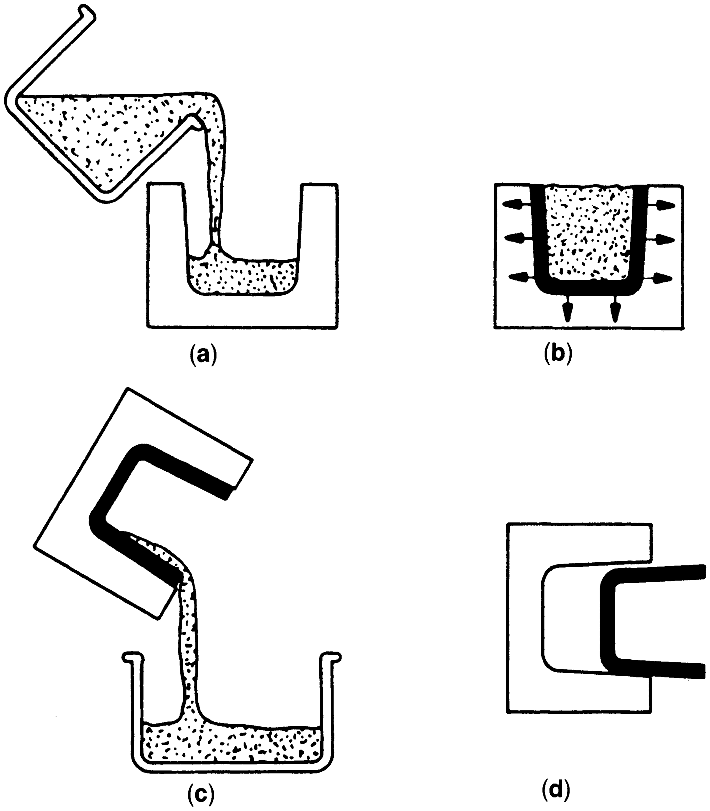

FIG. 17. The drain-casting approach to slip casting. (a) The slip is poured into a plaster mold, (b) the pores of the mold draw water out of the slip near the mold wall, (c) the excess liquid is drained off, and (d) the compacted form is removed after partial drying (Richerson, 1992).

compaction uniformity. In these processes, the powder is loaded in hydrostatically (equally loaded in all directions) by sealing the powder in a flexible bag and applying pressure using a liquid or a gas. This results in improved compaction uniformity.

Hot pressing (and hot isostatic pressing) combine the processes of compaction and sintering (described below). The application of pressure at the sintering temperature accelerates the compaction process by increasing the driving force for densification by mass transfer. These effects are described below.

3.1.2 Sintering A powder compact that is heated to below the melting point ($T/T_m > 0.75$) of the solid undergoes densification by a process known as sintering. The driving force for densification during sintering is the reduction of internal surface area. The amount of surface area obviously scales inversely with the size of the particles that make up the powder. Consequently, coarse powders densify more slowly than do fine powders. Upon heating, a variety of diffusion-based mass transfer processes (initially dominated by surface, then grain-boundary, then bulk diffusion) result in the fusion of particles and the filling in of the pores between particles. Initially, the porosity will be fully interconnected in a three-dimensional network, and the primary change in microstructure will be the smoothing of the particle surfaces and the fusion of connecting particles. After moderate times at temperature, the particles become fused and begin to form a three-dimensional network of solid material consisting of individually oriented grains, with pores present mostly at the grain boundaries. Grain growth occurs concurrent with the reduction in pore size and the increase in pore spacing. When pores become trapped within individual grains, they become difficult to remove by further annealing, because grain boundaries provide a rapid diffusional path for vacancies away from the void. Consequently, after long times at temperature, the density may not change with time, but entrapped voids make the actual density less than that of the fully dense structure.

The mechanisms of material transport that

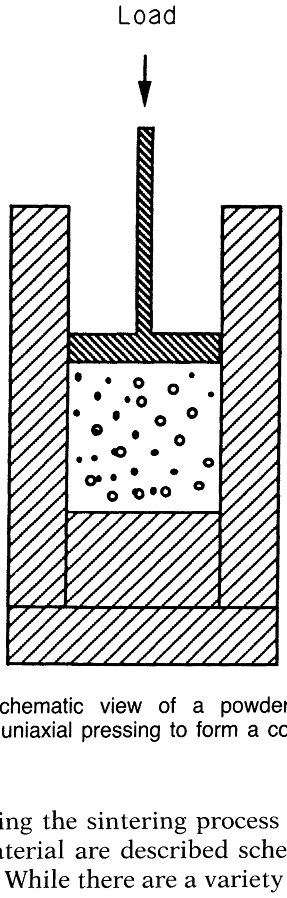

FIG. 18. Schematic view of a powder in a die undergoing uniaxial pressing to form a compact.

occur during the sintering process of a crystalline material are described schematically in Fig. 19. While there are a variety of matter transport mechanisms that can operate during sintering, there are only three that directly contribute to densification by removing material from between grains and depositing it on a void surface. Densification requires material transport paths that cause grains in the structure to move closer together. These three mechanisms are grain-boundary diffusion, volume diffusion, and dislocation motion. The transport mechanisms that do not directly result in densification, and actually compete against the densification processes, are evaporation/condensation at and to pore surfaces, surface diffusion (which results in the smoothing of the particle surfaces), and grain coarsening by volume diffusion (i.e., movement of material from one grain to another rather than to a pore surface).

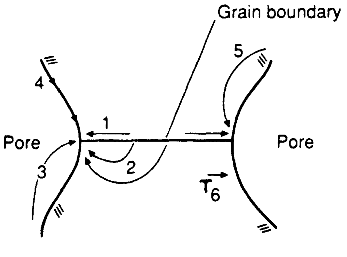

Mechanism	Densifying	Non-densifying
1. Grain boundary diffusion	X	
2. Volume diffusion	X	
3. Evaporation/condensation		X
4. Surface diffusion		X
5. Coarsening by volume diffusion		X
6. Dislocation (T) creep	X	

FIG. 19. Densification occurs by material transport to the pores of the compact. Six mechanisms for densification are shown schematically (Evans and Jonghe, 1991).

At some time *t*, there will be grains surrounding a pore as shown in Fig. 20. The atom flow can be calculated if the diffusivity of the atoms of the material is known. Empirically, it has been seen that the rate at which a compacted specimen undergoes densification (as measured by the reduction in length of the compact) is given by $\Delta L/L_0 = Kt^m$, where ΔL is the change in length, L_0 is the initial length of the compact, *t* is the time, *m* is a constant that lies between 0.3 and 0.5, and *K* is a factor that is constant in time but depends inversely on the cube of the particle size. As the compact gets to be within 85% of the full density, the rate decreases from that which is predicted by this expression since the pores become completely closed, and, as more of them become isolated within individual grains (i.e., with no grain-boundary intersection with the pore), the rate of densification goes to zero. Consequently, it is necessary to use additional driving force—say, by the application of pressure, as in hot pressing—to produce fully dense materials.

The compaction of a powder and subsequent annealing of the compact at between 0.5 and 0.8 of the melting point is sufficient to produce a free-standing object from a powder. Annealing for long times at these temperatures will cause a 1-μm powder to produce a structure that has no interconnected porosity and a grain size of 10–100 μm. If one is not concerned about the mechanical strength of the object produced, then this may be sufficient. Otherwise, additional steps, such as the application of a repressing step before additional sintering, may be necessary. More rapid densification and more fully dense materials may be formed if multicomponent materials are sintered at temperatures above

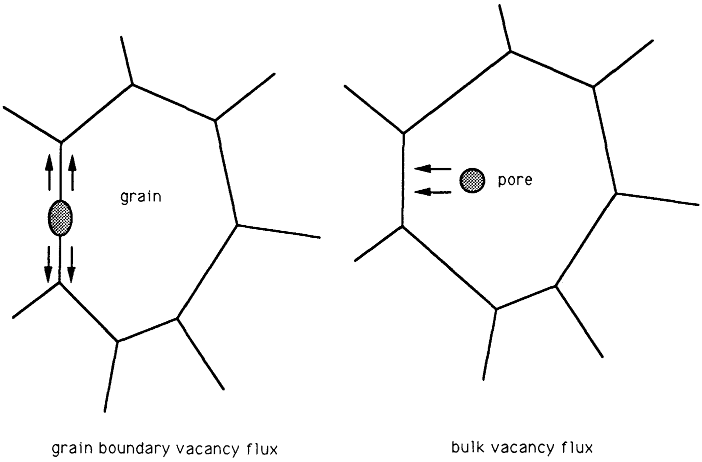

FIG. 20. Schematic view of pores present in and out of grain boundaries.

the melting point of one of the components. This results in the formation of a liquid that wets the grain boundaries of the material and provides a more rapid diffusion path for densification. This is routinely done in ceramic and cemented carbide (for hard-tool materials) sintering processes and is known as liquid-phase sintering.

3.2 Mechanical Forming

Mechanical forming is a process that is applied to those materials that can tolerate significant amounts of plastic deformation. Plastic deformation occurs in polymers by the flow of the molecules that make up the plastic and in metals by the motion of dislocations. A dislocation is a line defect that represents a boundary between atomic planes that have been displaced relative to one another in response to external forces. Plastic deformation of metals can be appreciated from a plot of engineering stress (force per unit initial cross-sectional area) versus strain elongation, $(L_f - L_0)/L_0$, as shown schematically in Fig. 21 for a bar that is loaded in tension. The initial deformation is elastic (i.e., fully recoverable), but above a critical stress (called the yield stress) the strain induced is no longer linearly proportional to the applied load, and it is not recoverable. The shape of the stress-strain curve depends on the material (both composition and microstructure) that is being loaded.

A feature of the plastic deformation behavior of crystalline materials is that the yield strength decreases at higher temperatures, and hence the energy necessary to perform mechanical shaping by plastic deformation also decreases. Consequently, most initial metal-forming operations (known as finishing operations) are performed at elevated temperatures. Ceramics cannot sustain any plastic deformation at reasonable temperatures and hence must be formed by powder methods. Hot-working processes are defined as those in which the deformation takes place under conditions of temperature and strain rate such that recovery processes take place

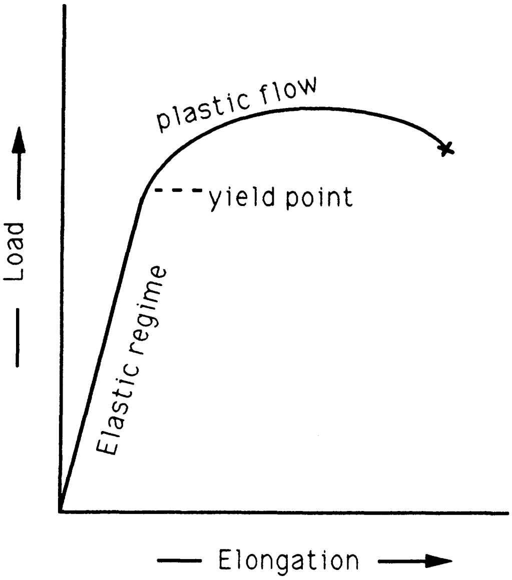

FIG. 21. Schematic illustration of load versus elongation for a metal.

simultaneously with the deformation. Recovery processes leave the grains of the metal relatively large and equiaxed and free of dislocations. Deformation at low temperatures (at $T \ll 0.3T_m$) is called cold-working and results in dramatic changes in microstructure and mechanical response of a material.

The mechanisms of plastic deformation are discussed elsewhere (e.g., Dieter, 1986); only a few essential concepts are discussed here. For instance, the process of strain hardening can be seen in Fig. 21 which shows that the stress necessary to continue deforming a material past its initial yield point increases with increasing strain. The load at which a material initially yields, and its strain-hardening behavior, is easily determined under uniaxial conditions by a uniaxial tensile test. The loads applied during metal forming are, however, rarely simply uniaxial but instead include a combination of stress components. Consequently, a criterion for yielding is required. Metals deform by crystallographic slip (*via* dislocations) in specific slip planes and in specific directions. In polycrystalline materials where the bulk is considered to be isotropic and homogeneous and in which volume is conserved, a criterion for yielding can be developed based on distortion–energy considerations and is called the von Mises yield criterion:

$$(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2 = C, \tag{13}$$

where C is a constant and $\sigma_{1;2;3}$ are the principal stresses. Principal stresses are found by rotating the orthonormal coordinate system with respect to the applied stresses so that the stress state can be described in terms of just three axial stress components that act normal to the three planes defined by the new axes (i.e., no shear components). With tension conventionally noted as positive and compression negative, it is easy to show that for uniaxial tensile loading where σ_1 is positive and $\sigma_2 = \sigma_3 = 0$, yielding will occur when σ_1 = yield strength (σ_{ys}) as defined for pure tension in Fig. 21 and $C = 2\sigma_{ys}^2$. For pure torsional loading, $\sigma_1 = -\sigma_3$ and $\sigma_2 = 0$, and yielding occurs when the applied torsional stress exceeds the shear yield strength ($\sigma_{ys\ shear}$) and the von Mises criteria is $C = 6(\sigma_{ys\ shear})^2$.

3.2.1 Bulk Forming Operations that cause shape changes in a plastically deformable material through the application of compressive stresses to an ingot are referred to in this article as bulk-forming methods. In these processes, there is a large degree of contact between the tool and the work piece (ingot). Flat rolling is used to generate plates, sheet, and foil by the reduction of the plate thickness and, although the forces are applied to the surface of the work piece, the metal flows at right angles to the applied load. These operations are performed in a rolling mill, which consists of cylindrical rolls that are set a fixed distance apart. When the width-to-thickness ratio of the original piece is large (>10:1), then the piece is constrained to deform only in the length direction. Constraint in both the width and thickness directions results in the development of stresses only in these directions and not in the unconstrained length direction. The geometry of the rolling process for a simple one-roll pair is shown in Fig. 22. Instead of using rolls with parallel surfaces, it is also possible to use rolls with complex shapes so as to prepare I-beams, rails, and other sections.

Metal that can be made to flow can also be formed by extrusion, which is a process in which the metal is forced (by application of compressive pressure) to flow plastically through a restricted orifice (known as a die) of arbitrary shape. The pressure is applied by a piston that, along with the chamber that surrounds the metal, applies compressive stresses to the work piece. The basic layout for the extrusion process is shown in Fig. 23. A related process is known as drawing, which is the method used to make wire or rod from bulk metal. In this process, the metal is pulled through the die so as to develop tensile stresses in the material after it passes through the die. In drawing processes, it is necessary to remain in the region of the stress-strain curve where homogeneous deformation occurs. If the reduction in cross section is too large, the tensile stresses that are required to pull the metal through the die cause necking (i.e., localized reduction in cross section) and fracture. By selecting a sequence of dies with an ever-decreasing diameter, extremely large reductions in cross section can be achieved, and, consequently, extremely long lengths of wire or rod can be formed.

Forging is a method that is roughly described as the mechanical working of a metal into a useful shape by hammering or press-

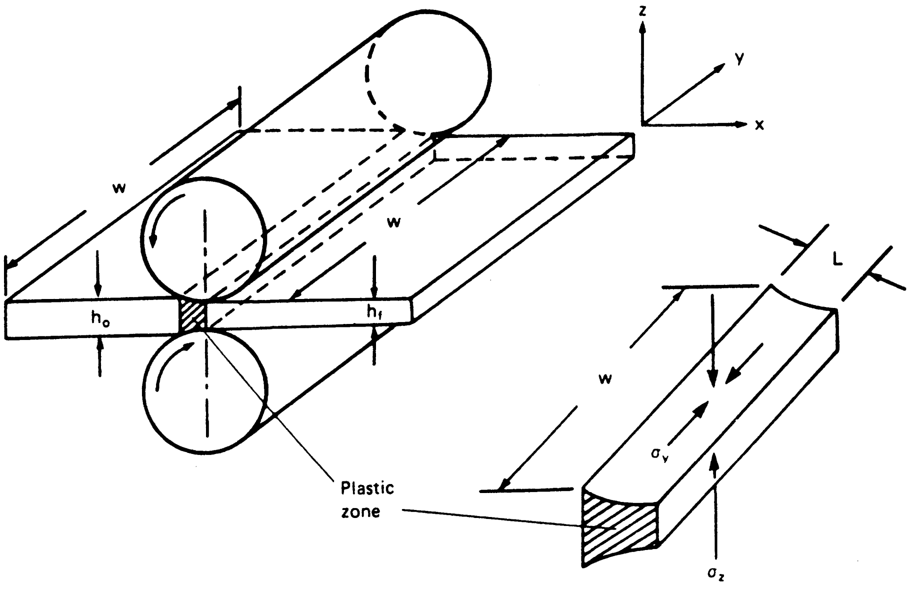
z
y
x
w
w
h_o
h_f
L
w
σ_y
Plastic zone
σ_z

FIG. 22. The geometry of a rolling mill and the resulting deformation.

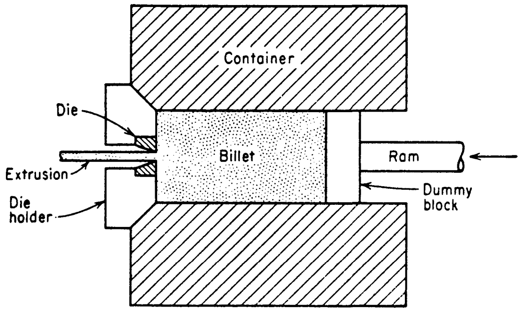
Container
Die
Extrusion
Die holder
Billet
Ram
Dummy block

(a) Direct extrusion

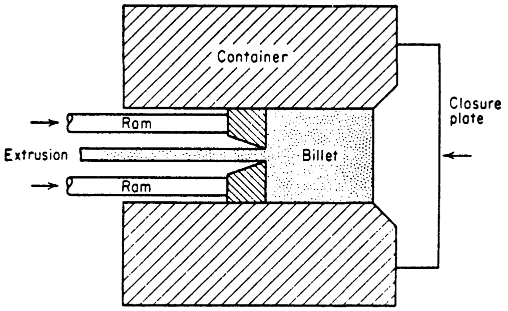
Container
Ram
Extrusion
Ram
Billet
Closure plate

(b) Indirect extrusion

FIG. 23. Two types of metal extrusion arrangements: (a) direct and (b) indirect extrusion (Dieter, 1961).

ing. This method is widely used to form very large pieces from cast steel ingots (hot forging). Steel-forging operations take place at elevated temperature in equipment that is designed to apply blows to the sample in one of two ways. A forging hammer (or drop hammer) is set up to repeatedly deliver rapid impacts to the metal. A forging press subjects the metal to slow-speed compressive force. In both cases, direct compression is applied by driving two patterned dies toward each other. An example of items that are fabricated by forging include drive shafts for large stationary engines and turbines. On a smaller scale, it is also used to form heads on nails and to produce the relief pattern observed on coins. An example of a forging machine is shown in Fig. 24. Forging also is very important for reducing the dendritic microstructure of an as-cast ingot into a finer, more equiaxed and less segregated structure with superior mechanical and corrosion properties.

3.2.2 Sheet Forming The forming of sheet is distinguished from the broad category

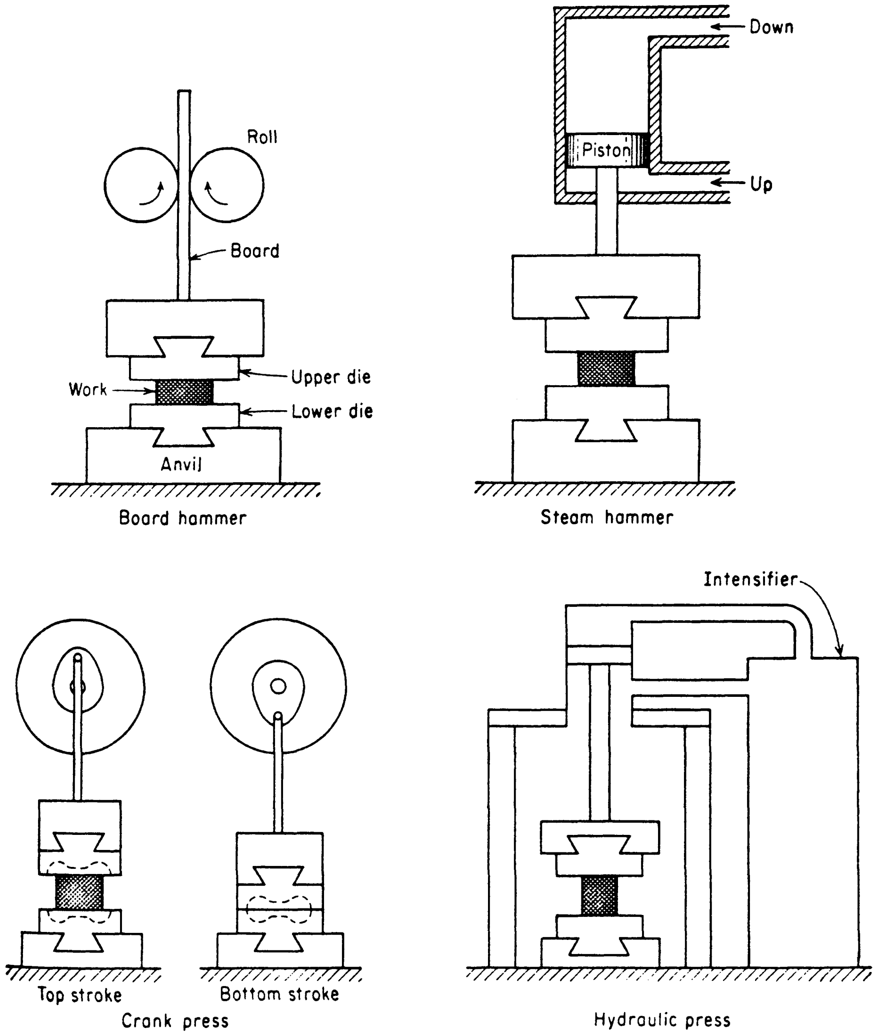
Roll
Board
Upper die
Work
Lower die
Anvil
Board hammer
Down
Piston
Up
Steam hammer
Top stroke
Bottom stroke
Crank press
Intensifier
Hydraulic press

FIG. 24. Forging equipment shown schematically.

known as bulk forming by the application of forces that result in inhomogeneous stress states. For example, the simplest sheet-forming processes imaginable are those that involve bending to form simple objects (e.g., brackets, metal boxes, cabinets) by taking a piece of sheet metal and applying bending moments. In this case, the top surface of the material experiences tension while the bottom surface experiences compression, and hence, the piece experiences inhomogeneous stress during deformation. More complex sheet-forming processes include deep drawing, which is used to form the familiar aluminum beverage can. The formation of a beverage can begins with a round slug of Al that is clamped in a hold-down device over a die. Pressure is then applied to the central region of the slug using a punch to cause the material to flow (at room temperature) into the die. The flange of the blank is subjected to compressive forces due to the hold-down device, while the wall of the cup experiences tensile stresses. The process is shown schematically in Fig. 25.

4. POLYMER PROCESSES

The creation of a plastic product or structure requires two principal steps: the first is the chemical manufacture of long-chain car-

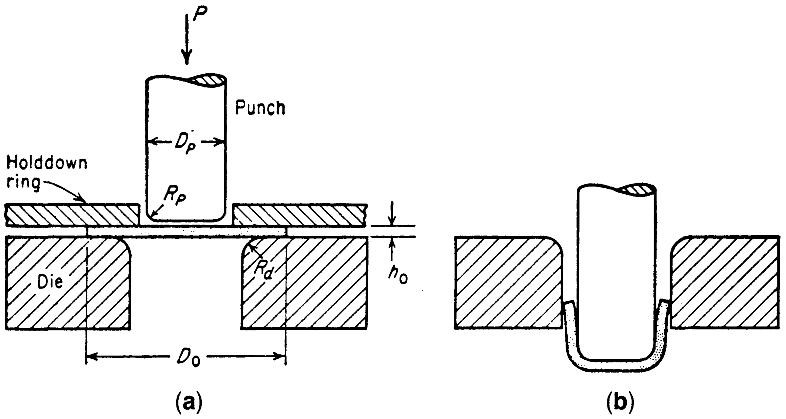

FIG. 25. Deep drawing of a cylindrical cup or can showing the material before and after drawing.

bon-based polymeric molecules, and the second is the arrangement of these molecules into a near-final solid shape. The shaping of polymeric materials depends to a large extent on the class of polymer to which the material belongs. The three classes are thermoplastics, thermosets, and elastomers. Thermoplastics flow when heated and therefore can be formed by any one of a variety of flow processes (such as extrusion) that were described earlier for mechanical forming of metals. The only difference between metal and thermoplastic forming is that thermoplastics flow at much lower stresses at close to room temperature. Thermosetting polymers are formed from polymeric precursor molecules that undergo cross-linking when subjected to heat or to a catalyst. As a consequence, thermoset materials must be in their near-final shape before they undergo the cross-linking process, during which they transform irreversibly from a liquid to a solid state. Thermosets are widely used as the matrix in polymer-based composites while thermoplastics, owing to the ease with which they are formed, are the basic material from which most low-cost packaging and consumer goods are formed.

4.1 Polymeric Materials

Most polymers (or plastics) are based on some form of synthetic organic resin that is a long-chain compound molecule containing carbon as the central element. The carbon atoms in each molecule are covalently bonded together, and the molecule is saturated with hydrogen and subunits known as radicals, which typically contain nitrogen, oxygen, and chlorine. The size of the basic molecule that forms the polymer, known as a polymer macromolecule, depends on the number of atoms in the carbon backbone. This quantity, known as the degree of polymerization, has a profound affect on the properties of the material. Polymer macromolecules are formed by the bonding together of much smaller subunits called monomers. For example, the thermoplastic polymer known as polyethylene is formed by the linear addition of activated ethylene monomer units:

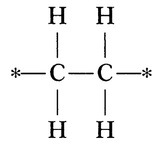

$$\begin{array}{ccccccc} & & H & & H & & \\ & & | & & | & & \\ * & - & C & - & C & - & * \\ & & | & & | & & \\ & & H & & H & & \end{array} \qquad (14)$$

where the * represents an activated site. Hundreds of these ethylene monomer units add together in what is known as an addition reaction to form the long (>500 monomer units) chain macromolecule. There is no by-product formed during an addition reaction.

The properties of these materials greatly depend upon the size of the polymer macromolecules and the interaction and arrangement of the molecules in the solid. For instance, as base units are added, the stable form of the molecule at room temperature changes from a simple gas, to a liquid, to a solid. Polymer molecules can contain up to thousands of monomers. The polymer macromolecule can be made up of a single type of monomer as is the case with polyethylene, or it can be made in a process known as copolymerization from two or more monomers.

Thermoplastic materials are composed of chemically separate long molecules that are held together by secondary chemical bonds. While the primary covalent bonds of the backbone of each molecule remain strong upon heating, the secondary ones are weakened at elevated temperature and ultimately dissociate. Thermoplastics, therefore, can flow at elevated temperatures because the macromolecules of the polymer are not bonded to each other and can respond to an applied stress by sliding past one another.

When the molecules of a polymer are randomly arranged, the material is called an amorphous (or glassy) polymer. Examples of amorphous polymers include polystyrene, polymethyl methacrylate, and polycarbonate. Other plastics are called crystalline because their molecules are locally aligned in a pseudoregular array to form structures called spherulites. The spherulites are surrounded by amorphous material but are interconnected by small fibers of crystalline material. Examples of crystalline polymers include polyethylene, isostatic polypropylene, acetal, and nylon. Crystalline polymers tend to have the strongest secondary bonds and hence are able to withstand higher temperatures than amorphous polymers. The arrangement of the radical groups on the long-chain molecule affects the crystallization behavior of some polymer materials. For example, if the radical groups are arranged randomly along the length of the molecule, then the packing of the macromolecules is less efficient and the polymer is less likely to crystallize. The increasingly organized arrangement of radical groups is referred to as the tacticity of a polymer. The higher the tacticity, the more likely the molecules will arrange themselves in a crystalline form. Tacticity also affects the efficiency with which molecules can be packed together. In general, lower tacticity results in increased distance between molecules, weaker secondary bonds, and lower softening temperature.

A second class of plastics is known as thermosetting. Plastics in this class are composed of macromolecules that are linked together in three dimensions (as opposed to the linearly bonded carbon backbone in thermoplastics) by primary covalent bonds. An example of a thermosetting plastic is phenolic resin, which is formed by the reaction between phenol and formaldehyde. Thermosetting plastics do not soften upon heating and remain intact until they thermally decompose into smaller molecules. The thermosetting plastics are generally stronger, harder, and more resistant to heat and chemical attack, but they do not flow under loads and tend to be brittle.

A third class of polymeric materials is known as elastomers. These consist of macromolecules that are cross-linked at widely separated points into a three-dimensional network. Examples of elastomers include natural rubber, isoprene, polybutadiene, and neoprene. Elastomers can be stretched to at least twice their original length, and they preserve their ability to recover their original shape after loading over a wide range of temperatures (typically −20 to 200 °C). The elastic behavior of elastomers is permanently lost when, through exposure to heat, light, or chemical attack, the network becomes increasingly cross-linked.

4.2 Polymer Forming

Plastic objects are formed by a variety of methods, most of which are a simple application of the methods that were discussed for

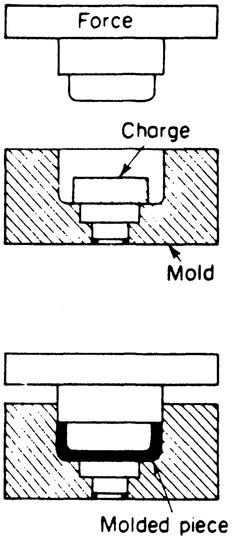

FIG. 26. Compression molding setup for forming thermoplastics (Ludema, 1987).

metal forming. These include extrusion, casting, and injection molding. Many of these processes have been discussed in the previous sections (e.g., injection molding), and only the additional details relevant to plastic forming will be discussed here. Thermoplastics are formed by mechanically forcing them to flow (typically above their softening point). Most of the thermoplastics that are commonly used are formed at temperatures between 120 and 250 °C. Thermosets are not plastically deformable and hence must be formed from the liquid state by casting or otherwise applying the liquid in nearly its final shape before the cross-linking process proceeds to the final formation of a solid.

Thermoplastics can be produced in any cross-sectional shape by extrusion. In plastic extrusion equipment, the feed takes the form of small plastic pellets that are heated before the application of a force that extrudes the plastic through a die. Unlike the extrusion of metal, thermoplastics can be formed at 120 to 250 °C with relatively light loads that may be applied using a screw-type auger feed mechanism. Compression molding involves the squeezing of a set amount of polymer in a cavity with a punch as shown in Fig. 26. In this process, the plastic is made to flow by the application of pressure to fill the space between an appropriately shaped piston and a mold. Blow molding is used for the fabrication of plastic bottles and is shown schematically in Fig. 27. In this process, a tube (called a parison) of thermoplastic material is heated and placed into a mold. One end of the tube is pinched closed, and gas is used to pressurize the inside of the tube so that it bulges out against the mold wall.

Direct casting can be performed with thermosets that, in their liquid state, can be cast in relatively soft molds made of rubber,

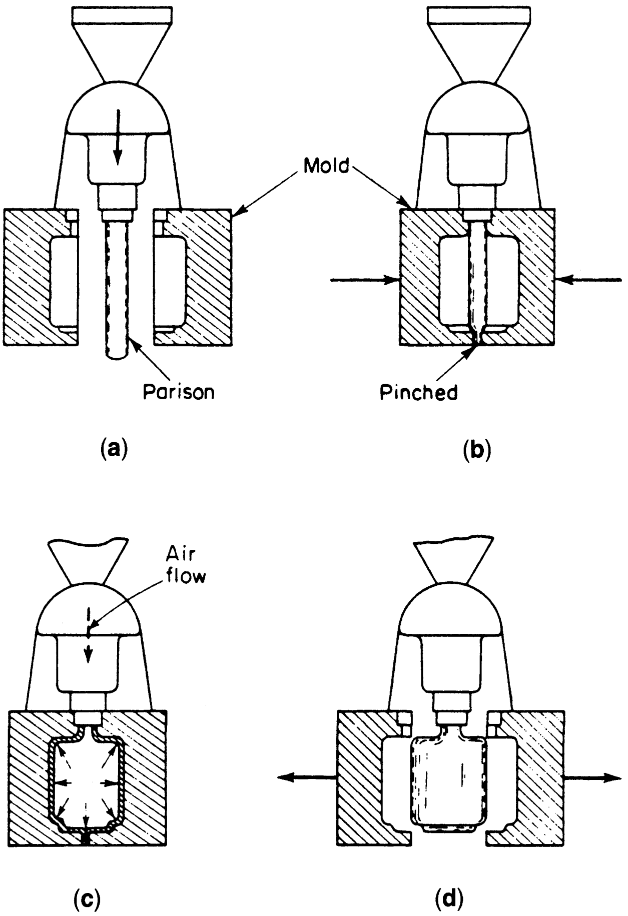

FIG. 27. Blow molding for the formation of plastic bottles from thermoplastic (Ludema, 1987).

wood, plaster, etc. A catalyst is added to the resin to speed up polymerization. Thermosets are commonly used in the formation of polymer matrix composites. Composites are materials that consist of a soft matrix material (i.e., the polymer) that is reinforced by fibers of a much stiffer and more brittle material. These two-part materials are tougher (resistant to failure by crack propagation) than either of the constituents and are stronger and stiffer than the matrix polymer. Composites may be formed by using a thermosetting polymer matrix (epoxy, for example) and mixing in fibers made from glass, carbon, or Kevlar (aramid). In all of these cases, the resulting composite material is much tougher (i.e., resistant to failure by crack propagation) than either the matrix epoxy or the reinforcing material. The mechanical properties of composites are very sensitive to the orientation of the applied load with respect to the alignment of the fibers.

GLOSSARY

Addition Reaction: The process by which bifunctional monomer units combine sequentially to form a linear polymer macromolecule.

Burgers Vector: A vector that describes the magnitude and direction of the lattice distortion associated with a dislocation.

Ceramic: A compound of metallic and nonmetallic elements that are bonded together *via* ionic bonds.

Crystalline: A solid state of matter in which the atoms of the solid are distributed in a repeating three-dimensional array.

Dislocation: A linear crystalline defect that marks the center of a localized distortion of the crystal lattice. Plastic deformation occurs by the movement of these defects.

Drawing: In metal forming, a technique used to fabricate metal wire and tubing. Deformation is accomplished by pulling the material through a die by means of a tensile force.

Elastic Deformation: Nonpermanent (i.e., recoverable) response of a material to an applied load.

Epitaxy: The coordinated arrangement of one set of atoms on a crystalline template formed by a single crystal of either the same or a different material.

Grain: An individual crystal in a polycrystalline metal or ceramic.

Grain Boundary: The interface in a polycrystalline material between two grains that have different crystallographic orientations.

Gibbs Free Energy (ΔG): A thermodynamic quantity that is a function of both the enthalpy and entropy of a system. Gibbs free energy is a minimum at equilibrium.

Lattice: A regular geometrical arrangement of points in space such that every point has identical surroundings.

Lattice Parameters: The combination of unit cell edge lengths and interaxial angles that defines the unit cell geometry.

Liquidus Line: A line on a binary phase diagram that defines a boundary between the single-phase liquid and the liquid-plus-solid regions. This line indicates the temperature at which solid will first appear when a particular alloy composition is cooled from the single-phase liquid regime.

Macromolecule: Usually a polymer molecule that consists of hundreds or thousands of atoms.

Monomer: The group of atoms that make up the repeating unit that forms a polymer macromolecule.

Plasma: A gas of positive ions and free electrons in approximately equal numbers.

Polymer: A solid, usually carbon-based, compound of high atomic weight that is composed of smaller repeating monomer units.

Plastic deformation: Change in shape, due to the application of a load, which is not recovered when the load is removed.

Solidus Line: A boundary on a phase diagram that defines the point at which solidification is complete (i.e., a single-phase solid regime is entered) when an alloy is cooled.

Strain, Engineering: The change in length of a specimen loaded uniaxially divided by the original length.

Stress, Engineering: The instantaneous load applied to a specimen divided by its cross-sectional area before any deformation.

Works Cited

Davies, G. J. (1973), *Solidification and Casting*, New York: John Wiley and Sons.

Dieter, George Ellwood (1986), *Mechanical Metallurgy*, 3rd Ed, New York: McGraw-Hill.

Evans, J. W., De Jonghe, L. C. (1991), *The Pro-*

duction of Inorganic Materials, 1st ed., New York: Macmillan.

Flemings, M. C. (1974), *Solidification Processing*, New York: McGraw-Hill.

Ludema, K. C., Caddell, R. M., Atkins, A. G. (1987), *Manufacturing Engineering: Economics and Processes*, Englewood Cliffs, NJ: Prentice-Hall.

McGuire, G. E. *Semiconductor Materials and Process Technology Handbook*, Park Ridge, NJ: Noyes Publications.

Porter, D. A., Easterling, K. E. (1987), *Phase Transformations in Metals and Alloys*, London: Van Nostrand Reinhold Co. Ltd.

Richerson, D. W. (1992), "Modern Ceramic Engineering," in: Peter Hilton (Ed.), *Engineered Materials*, New York: Marcel Dekker, Inc.

Further Reading

Ashby, M. F. (1992), *Materials Selection in Mechanical Design*, Oxford: Pergamon Press.

Barrett, C. R., Nix, W. D., Tetelman, A. S. (1973), *The Principles of Engineering Materials*, Englewood Cliffs, NJ: Prentice-Hall.

Davies, G. J. (1973), *Solidification and Casting*, New York: Wiley.

Dieter, George Ellwood (1986), *Mechanical Metallurgy*, 3rd Ed, New York: McGraw-Hill.

Doyle, L. E., Keyser, C. A., Leach, J. L., Schrader, G. F., Singer, M. B. (1985), *Manufacturing Processes and Materials for Engineers*, 3rd ed., Englewood Cliffs, NJ: Prentice-Hall.

Evans, J. W., De Jonghe, L. C. (1991), *The Production of Inorganic Materials*, 1st ed., New York: Macmillan.

Flemings, M. C. (1974), *Solidification Processing*, New York: McGraw-Hill.

Grant, N. J. (1981), "Powder and Particulate Production of Metallic Alloys," in: G. Y. Chin (Ed.), *Advances in Powder Technology*, Metals Park, OH: American Society for Metals, pp. 1–20.

Kingery, W. D. (Ed.) (1963), *Ceramic Fabrication Processes*, Cambridge, MA: MIT Press.

Ludema, K. C., Caddell, R. M., Atkins, A. G. (1987), *Manufacturing Engineering: Economics and Processes*, Englewood Cliffs, NJ: Prentice-Hall.

McGuire, G. E. (1988), *Semiconductor Materials and Process Technology Handbook*, Park Ridge, NJ: Noyes Publications.

O'Hanlon, J. F. (1986), *A Users Guide to Vacuum Technology*, New York: Wiley.

Porter, D. A., Easterling, K. E. (1987), *Phase Transformations in Metals and Alloys*, London: Van Nostrand Reinhold Co. Ltd.

Ray, H. S. Ghosh, A. (1991), *Principles of Extractive Metallurgy*, 2nd ed., New Delhi, India: John Wiley and Sons.

Reed, J. S. (1988), *Introduction to the Principles of Ceramic Processing*, New York: Wiley.

Richerson, D. W. (1992), "Modern Ceramic Engineering," in: Peter Hilton (Ed.), *Engineered Materials*, 2nd ed., Vol. 1, New York: Marcel Dekker, Inc.

Sze, S. M., (Ed) (1983), *VLSI Technology*, New York: McGraw-Hill.

Thornton, J. A. (1988), "Physical Vapor Deposition," in: G. E. McGuire (Ed.), *Semiconductor Materials and Process Technology*, Park Ridge, NJ: Noyes Publications, pp. 378–399.

Young, J. F. (1985), *Materials and Processes, Part B*, New York: Marcel Dekker Inc.

MATERIALS TREATMENT

James E. Krzanowski, *Mechanical Engineering Department, University of New Hampshire, Durham, New Hampshire, U.S.A.*

Introduction 399
1. **Materials Treatments for Modification of Mechanical Properties** 400
1.1 Deformation Treatment of Metals and Alloys 400
1.2 Thermal Treatment of Materials 401
1.2.1 Process Equipment for Thermal Treatments 401
1.2.2 Annealing Heat Treatments 402
1.2.3 Heat Treatment of Ferrous Alloys 403
1.2.3.1 Classifications of Steels 403
1.2.3.2 Transformations in Steels 403
1.2.3.3 Annealing, Normalizing, and Spheroidizing 404
1.2.3.4 Quench and Temper Heat Treatments 405
1.2.4 Precipitation-Hardening Heat Treatments 406
2. **Materials Treatments for Modification of Electrical Properties** 408
2.1 Effect of Materials Treatments on the Conductivity of Metals 408
2.2 Thermal Treatments for Semiconductor Materials 409
2.2.1 Annealing Treatments 409
2.2.2 Gettering Treatments 409
3. **Surface Modification of Materials** 410
3.1 Thermal Treatments for Surface Modification 410
3.2 Diffusion Processes 410
3.2.1 Carburizing 410
3.2.2 Nitriding 411
3.3 Electrochemical and Vacuum Coating Methods 412
3.4 Surface Treatments to Strengthen Glass 413
Acknowledgments 414
Glossary 414
Works Cited 415
Further Reading 415

INTRODUCTION

The development of methods that can be used to alter the microstructure of materials to obtain useful materials properties is one of the major accomplishments of materials science and engineering. The development of these methods has had a profound impact on our technological society and is thus of great importance to scientists and engineers.

A discussion of methods of *materials treatment* is presented in this article. A materials treatment is defined as a process that changes the physical properties of a material. The scope of this subject is enormous, and therefore in this article, only the most common materials treatments will be examined. Furthermore, this article will concentrate on those treatments likely to be applied by the end user of the material rather than the manufacturer.

The article is organized into three major sections. The first section deals with materials treatments that modify the mechanical properties of materials. These treatments include deformation and thermal treatments, and examples are presented for metals and alloys. The second major section deals with treatments that affect electrical properties of materials. The effects of cold working and annealing on the electrical conductivity of metals are first discussed, followed by a discussion of treatments for semiconductor materials. The third section deals with surface modification of materials. This section includes discussions of thermal methods, diffusion processes, and coating methods. The section concludes with a discussion of methods used to modify the surface of glasses and thereby improve bulk mechanical properties. Since the scope of this article is quite broad,

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

the discussion of each treatment is limited. However, the interested reader can find further details for many of the treatments introduced here elsewhere within this Encyclopedia. Specific references are given in each major section.

1. MATERIALS TREATMENTS FOR MODIFICATION OF MECHANICAL PROPERTIES

1.1 Deformation Treatment of Metals and Alloys

The process of deformation treatment of a metal or alloy is accomplished by applying forces to the metal that result in permanent changes to its macroscopic shape, physical properties, and microstructure. When a material undergoes plastic deformation, dislocations are generated and begin to propagate on slip planes oriented to have the highest resolved shear stress. As the applied stress is increased, further plastic deformation occurs and the density of dislocations (in terms of the total dislocation line length per unit volume) begins to increase. At sufficiently high plastic strains, dislocation tangles and cells form and become barriers to further propagation of slip dislocations, requiring yet higher applied stresses to continue plastic deformation. The phenomenon by which deformation strengthens a material is known as *work hardening* or *strain hardening*.

Cold rolling is a common process for deformation treatment of metals and alloys. In the cold-rolling process, the metal is reduced in thickness as it passes between two rollers. The amount of deformation is gauged by the percentage reduction in cross-sectional area or thickness of the plate which is being rolled. This is also known as the percent cold work, which is often useful in describing the effects of cold working on the properties of metals and alloys. Since the stresses and strains that develop during cold rolling are very complex, a detailed description is beyond the scope of this article. A discussion of stresses developed in metal-working processes can be found in Mielnick (1991).

Cold rolling can have a dramatic effect on the properties of the metal. The effect of cold rolling on the mechanical properties of copper is shown in Fig. 1. As a result of work hardening, the tensile and yield strengths increase with the percent cold work. The percentage elongation decreases with increasing cold work, and the maximum amount of cold working allowed in this example is 90%. Further cold working can cause cracking and tearing. At this point, the dislocation density has reached a very high level, limiting further plastic deformation. Additional effects that can result from cold rolling are the formation of a grain structure that is elongated in the direction of rolling, formation of crystallographic or sheet texture, elongation of inclusions, and the development of residual stresses. Elongated inclusions are known as stringers and can act as nucleation sites for cracks and reduce the fracture toughness of a material.

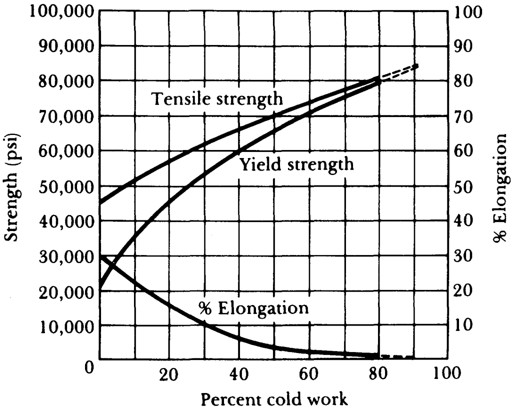

FIG. 1. The effect of cold work on the tensile strength, yield strength, and percentage of elongation of copper. Note the reduction in percentage of elongation at large amounts of cold work (Askeland 1989).

The extent of cold rolling in copper alloys is denoted by a temper designation. For alloys which are cold worked only, the H*xx* designation is used. For example, H02 refers to half-hard or 20.7% reduction in thickness, while H10 is extra spring (68.6% reduction). The maximum is H14, which represents an 80.3% reduction and is referred to as super spring. Copper of commercial purity that has been annealed to a coarse grain size has a yield strength of 69 MPa and a percentage elongation of 55, while commercial-purity copper cold worked to the H10 condition has a yield strength of 370 MPa and a percentage elongation of 4. Aluminum and its alloys are also commonly treated by deformation processes. For aluminum, H1*x* temper designations are used, where H12 is quarter-hard and H18 is full-hard, obtained with approximately 75% reduction. For aluminum of 99.60% minimum purity, the yield strength in the annealed condition is 28 MPa with 43% elongation. After cold working to H14, the yield strength is 91 MPa (12% elongation), and at H18, the yield strength is 126 MPa (6% elongation). Similarly, an aluminum alloy containing 1.2% Mn and 0.4% Mg has a yield strength of 56 MPa (25% elongation) in the annealed condition and 230 MPa (4% elongation) in the H18 condition. Clearly, cold working dramatically increases the yield strength, although a concomitant reduction in the percentage of elongation is also realized. For a more extensive compilation of data on cold working, the reader is referred to the ASM Handbook (1979).

Other metalworking processes, such as wire drawing, extrusion, forging, and stamping, can cause similar microstructural and property changes. In many of these processes (as well as rolling), the metal is worked while it is hot (above $\frac{1}{2}\,T_m$, where T_m is the melting temperature in K) rather than cold. In this case, recrystallization of the metal may occur during or immediately after working and a fine strain-free grain structure is obtained. Properties of recrystallized metals are discussed in the Sec. 1.2.2.

1.2 Thermal Treatment of Materials

The treatment of materials by thermal processes is probably the oldest and most common method for treating materials. Thermal treatments are carried out to change the properties of a material without causing significant changes in its shape. In this section, methods for heat treating bulk materials are presented. Before describing specific heat treatments, the equipment commonly used for thermal treatments is described. Heat treatment of ferrous alloys is then considered, followed by a discussion of precipitation-hardening heat treatments.

1.2.1 Process Equipment for Thermal Treatments To carry out thermal treatment of materials properly, the type of furnace and furnace atmosphere used must be given careful consideration. Furnaces may be classified in two general categories: continuous furnaces and furnaces used for batch operations. *Continuous furnaces* have an automatic loading or conveying system to feed the workload and are used in large production applications, such as the continuous annealing of sheet steel. Continuous furnaces can also be used for individual components; roller hearth, belt conveyor, and pusher furnaces are used for continuous processing of individual parts. *Batch furnaces* are used for smaller volume applications and by commercial heat treaters. Considerably more labor is required for loading and unloading of the parts being heat treated in batch operations, but these systems provide much more flexibility.

Furnaces generally consist of a steel shell with insulated walls and loading door and a heating system. Furnaces may be heated by a number of different methods and fuels, including electrical resistance, gas-fired, oil-fired, and radiant tubes (electrically heated or oil-fired). Selection of a furnace requires consideration of many factors, such as the thermal cycle requirements, type and quantity of material to be processed, initial equipment cost, operating/fuel costs, maintenance costs, quenching facilities, furnace controls and atmospheres, and environmental concerns.

The furnace atmosphere during the heat-treatment cycle often is carefully controlled to prevent or minimize oxidation or other surface alteration of the material being treated. The main objective is to exclude oxygen from the gas mixture, which usually consists of nitrogen as the main ingredient and various percentages of carbon monoxide, carbon dioxide, hydrogen, and methane (or other gaseous hydrocarbon). Reactive metals and their alloys are often heat treated in inert gas atmospheres of argon or helium. If oxidation is not a concern, a simple and economical air atmosphere may be suitable.

Molten *salt baths* are often used in the thermal treatment because of their rapid heat-transfer characteristics. Isothermal annealing treatments of steel, described in Sec. 1.2.3.2, are often carried out in salt-bath furnaces. A particular advantage of salt-bath furnaces is the ability to maintain a uniform and closely controlled temperature.

In recent years, *vacuum furnaces* are being increasingly used for a number of thermal processes, including heat treating, sintering of powder metallurgy components, brazing, stress relieving, and diffusion bonding of metal fabricated structures. Almost all vacuum furnaces are now of the cold-wall type, in which the outside wall or vacuum chamber is water cooled. A number of thin, light-weight radiation shields protect the outside shell from the high temperatures of the central working zone. The furnace is heated with electric resistance elements of molybdenum sheet, graphite, tungsten, or tantalum. Vacuum furnaces are usually equipped with facilities for inert-gas backfill on component quenching. They also may have interlock systems and arrangements for oil quenching the furnace load.

Fluidized-bed equipment is also finding increased usage. In these systems, various gases are introduced into the bottom of the vessel and cause the small ceramic particles comprising the bed to act like a fluid. An important advantage of fluidized-bed furnaces is the efficient heat-transfer characteristics, which approach that of salt baths.

In carrying out thermal treatments, safety factors must be carefully considered because of the high temperatures involved and the possibility of fires and explosions of furnace gases. For a thorough review of heat-treatment methods, the reader is referred to the ASM Handbook (1979).

1.2.2 Annealing Heat Treatments The process of *annealing* involves heating a material to a temperature sufficient to negate at least some of the effects of cold working. After annealing at low temperatures, reductions in the residual stress and electrical resistivities are realized, while annealing at higher temperatures results in a reduction of the flow stress of a metal.

Process Anneal: During cold working, the ductility rapidly decreases as the percent cold work is increased. The *process anneal* heat treatment is used to negate effects of cold working. After a process anneal is carried out, a fine-grained microstructure is obtained, which can have a higher ductility and lower yield strength than the cold-worked state.

Stress-Relief Anneal: Residual stresses can develop in metals as a result of cold working or machining processes, or as a result of nonuniform cooling after casting, welding, or heat treating. In order to eliminate residual stresses and avoid possible distortion of the part, a *stress-relief anneal* is applied. Effects of cold work are not completely negated during this process, since the temperature is relatively low compared to the process anneal. However, there may be some dislocation recovery, as described in the following section.

Recovery, Recrystallization, and Grain Growth: After a metal has been cold worked, a deformation microstructure remains. This excess stored energy provides the driving force for the processes of recovery and recrystallization.

Recovery and recrystallization can both lead to softening of the cold-worked metal. However, the mechanisms and kinetic features of these processes are distinctly different. Recovery starts at low temperatures, and the extent of recovery increases with time and temperature. Recovery is a continuous process where the defect concentration is gradually reduced with increasing time and temperature. The dislocation microstructure begins to recover at temperatures above $0.3T_m$ (T_m = melting temperature in K), and as recovery proceeds, the electrical resistivity is substantially reduced through both vacancy and dislocation annihilation.

In contrast, recrystallization proceeds by the nucleation and growth of equiaxed, strain-free grains. Any region of the sample either is recrystallized or is still in the cold-worked (or partially recovered) state. Recrystallization starts slowly, increases to a maximum rate, and then approaches completion at a slow rate. A useful parameter in recrystallization treatments is the *recrystallization temperature*, which is taken as the temperature required to achieve complete recrystallization in a given period of time, usually 1 h. The purity of a metal can significantly affect the recrystallization temperature. For example, the addition of 0.01 at.% silver to copper will raise the recrystallization temperature by 80°C.

Once a cold-worked structure has been recrystallized, the dislocation density and the strain energy associated with the remaining dislocations are greatly reduced. However, the recrystallized metal will still contain excess energy due to the interfacial energy of the grain boundaries, which provides the driving force for *grain growth*. The increase in grain size that results from grain growth is generally detrimental to the mechanical properties, since both strength and ductility decrease with increasing grain size. Grain growth can be limited by proper control of the annealing temperature and time. In addition, the rate of grain growth is strongly dependent on the level of impurities in the metal. For example, grain growth in high-purity aluminum will proceed at a much higher rate than in commercial purity aluminum. Grain-boundary migration rates can be reduced by impurities either through an impurity-drag effect (Cahn, 1962) or by inclusions or second-phase particles pinning migrating grain boundaries. In addition, sample size may limit the ultimate grain size. For example, in thin metal films, the ultimate grain size is often about the same dimension as the film thickness.

For further details on recovery, recrystallization, and grain growth, the reader is referred to the article RECRYSTALLIZATION, DYNAMIC, as well as the text by Reed-Hill and Abbaschian (1992).

1.2.3 Heat Treatment of Ferrous Alloys

Heat treatments of ferrous alloys are undertaken to modify their mechanical properties by altering the microstructure of the alloy. These treatments may be in-process to facilitate further processing or a final treatment of the finished component. A brief summary of the various classes of ferrous alloys is presented followed by a discussion of the most frequently used heat treatments.

1.2.3.1 Classifications of Steels. *Plain-carbon steels* are alloys of iron with carbon contents of up to about 1 wt%, along with some manganese and silicon [usually about (0.3–0.5)%]. The binary iron-carbon phase diagram, shown in Fig. 2, forms the basis for the heat treatment of most ferrous alloys. The equilibrium phases in this system are α-ferrite (bcc crystal structure), γ or austenite (fcc), δ-ferrite (also bcc but present only at high temperatures), and the compound Fe_3C, also known as cementite or iron carbide. The temperature of 727°C, at which the *eutectoid* reaction ($\gamma \rightarrow \alpha + Fe_3C$) occurs on cooling in plain-carbon steels, is referred to as the Ae_1 temperature (e specifies equilibrium conditions). For steels with less than 0.77% carbon (hypoeutectoid), only austenite is present above a temperature denoted Ae_3, while between Ae_1 and Ae_3, austenite and ferrite are stable. Similarly, for steels with greater than 0.77% carbon (hypereutectoid), the temperature above which only austenite is stable is denoted A_{cm}, while between Ae_1 and A_{cm}, both austenite and cementite are stable. The composition of 0.77% carbon is referred to as the eutectoid composition.

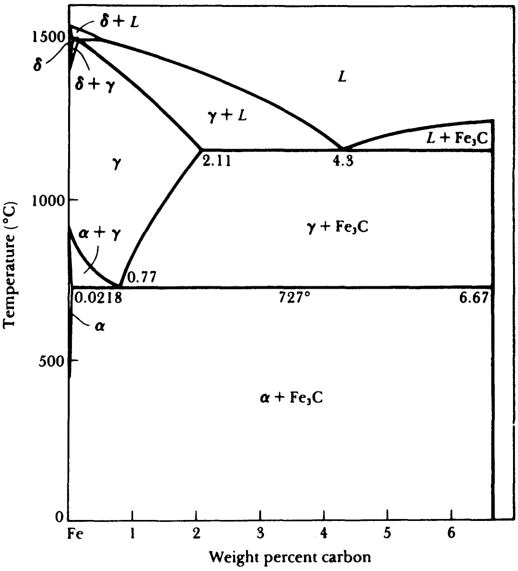

FIG. 2. The Fe-Fe_3C phase diagram (Askeland, 1989).

Alloy steels contain up to a few weight percent of additional alloying elements, such as nickel, chromium, molybdenum, vanadium, tungsten, silicon, and sometimes copper. *High-strength low-alloy* (HSLA) or *microalloyed* steels contain small amounts of vanadium, titanium, or niobium, which are all strong carbide-forming elements. For steels that contain significant amounts of alloying elements, the Ae_1, Ae_3, and A_{cm} temperatures and the eutectoid composition will be different from the plain-carbon steel values. *Tool steels* are alloys with medium and high carbon levels, often highly alloyed, designed with specific properties, such as high hardness, hot hardness, and wear resistance. They often contain chromium, molybdenum, tungsten, and/or vanadium. *Stainless steels*, used for their corrosion or heat resistance, contain at least 11% chromium to develop the necessary passive surface film. Other alloy additions are then made to develop the various grades that are classified according to their microstructure types: austenitic, ferritic, martensitic, and precipitation hardening.

1.2.3.2 Transformations in Steels. The rate of cooling from the austenite region greatly affects the structure of the resulting product. With slow cooling, a carbon steel of the eutectoid composition transforms to *pearlite,* a lamellar microstructure of α-ferrite and Fe_3C. However, at more rapid cooling rates, steel transforms to *bainite,* which is a nonlamellar microstructure consisting of α and Fe_3C, whereas quenching a steel can produce *martensite,* which is a metastable phase with a body-centered tetragonal structure. For a given steel, the martensite structure will have the highest strength and hardness, followed by bainite and then pearlite.

Both the carbon content and alloying elements affect the transformation characteristics of steels. Isothermal transformation (IT) and continuous-cooling transformation (CCT) diagrams are useful in predicting the equilibrium phase distribution of isothermally annealed steels. An IT diagram for a eutectoid steel is shown in Fig. 3. The log(time) axis measures the time of isothermal heating which directly follows quenching from the austenite region. The left C-shaped curve indicates the start of transformation, while the right-most C curve indicates the completion of transformation. The midpoint of the "nose" of the C curve defines the type of transformation product: holding the steel at temperatures above the nose results in pearlite formation, while holding at temperatures below the nose [but above martensite start (M_s) temperature] gives bainite. Quenching directly to below the M_s temperature causes transformation to martensite; the amount of martensite formed depends on the difference between the quench temperature and M_s. For practical applications, CCT diagrams are more useful. These diagrams describe the decomposition products formed as a function of cooling rate during continuous cooling.

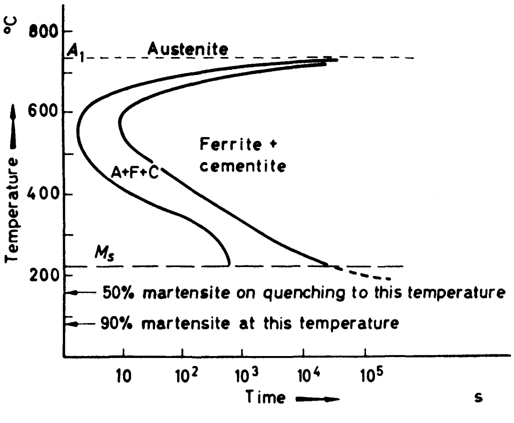

FIG. 3. Isothermal transformation diagram for eutectoid steel (0.77% carbon) (Smallman, 1985).

1.2.3.3 Annealing, Normalizing, and Spheroidizing. Annealing is undertaken to soften the steel, usually to facilitate an intermediate processing step such as machining or cold forming. Annealing of plain-carbon and alloy steels can be accomplished by heating to a specified temperature, holding for a period of time (usually about 1 h/in. thickness of the specimen) in the austenite region, and slowly cooling to room temperature. For hypoeutectoid steels, the annealing temperature is about 30°C above the Ae_3 temperature. Very slow cooling will give the structure a full anneal and a microstructure of coarse pearlite. A temperature of 30°C above the Ae_1 temperature is used to anneal hypereutectoid steels. If the steel were instead heated to a temperature above Ae_{cm}, where only austenite is stable, continuous cementite films would form on cooling between the Ae_{cm} and Ae_1 temperatures and cause embrittlement of the steel. Heating to 30°C above the Ae_1 instead permits the formation of rounded (spheroidal) cementite particles, which are less detrimental to the toughness.

Normalizing of hypereutectoid and hypoeutectoid steels is carried out by first heating to about 55°C above the Ae_3 or Ae_{cm} temperature as appropriate to transform the steel to austenite. The steel is then air cooled to room temperature to produce a ferrite-pearlite or pearlite microstructure. The more rapid cooling rate used in this process results in a finer interlamellar spacing in the pearlite than obtained after a full anneal and slow cool. During cooling of hypereutectoid steels, cementite films do not usually form because of the more rapid cooling rate. Normalizing is often carried out to refine the austenite grain size present from prior hot-working operations and is usually given before the final quench and temper operation.

The effects of normalizing and annealing on the mechanical properties of plain-carbon steels are shown in Fig. 4. As expected, the annealed steel has lower tensile and yield strengths in comparison with a normalized steel of the same carbon content. However, the impact energies and percentage elongations of the annealed steels are greater than those obtained in normalized steels.

The process of *spheroidizing* is used pri-

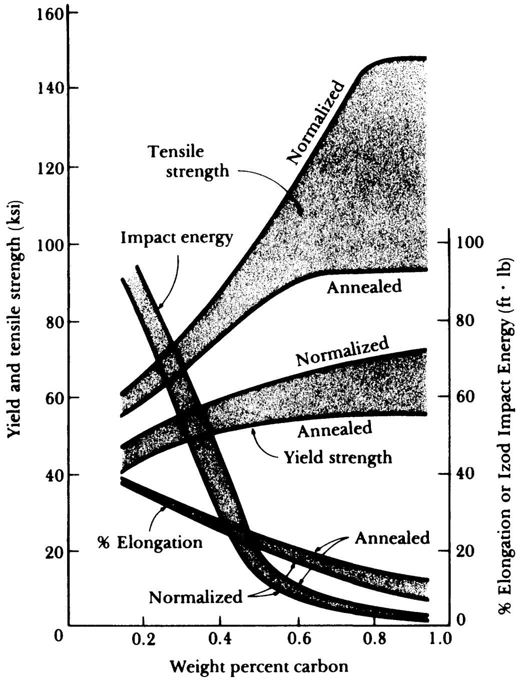

FIG. 4. Effect of carbon content on mechanical properties of plain-carbon steel (Askeland, 1989).

marily to improve machinability of hypereutectoid steels. This heat treatment is carried out by heating to about 30°C below the Ae_1 temperature. The heating time is relatively long, during which the cementite originally present in pearlite or as proeutectiod cementite (cementite which, on cooling from above A_{cm}, forms before the eutectoid temperature is reached) transforms into large spherical particles known as spheroidite. The steel is softened considerably by this process and easily machined. After machining, the steel may again be heat treated to higher hardness and strength levels.

1.2.3.4 Quench and Temper Heat Treatments. *Austenitizing* is the first step in all quench and temper heat treatments. In this process, the steel is heated to the austenite range and held at a particular temperature, usually about 25–50°C above the critical temperature for the particular grade of carbon or alloy steel.

After austenitizing, the steel is quenched to produce a martensitic structure. In order to produce a martensitic structure, the M_s temperature must be reached before the steel transforms to other decomposition products such as bainite or pearlite. In addition, during quenching of the steel, the surface will cool faster than the center, leading to a variation in cooling rate across the cross section of the steel. The hardenability of a steel describes the ease with which martensite can be formed on cooling. The hardenability of a steel is strongly dependent on the amount and type of alloying elements present. For example, a highly alloyed steel may be through-hardened by simply air cooling, whereas a plain-carbon steel of low carbon content may only form a thin hardened surface layer even after the most severe (rapid) quenching procedure.

A number of different quench media are employed to cool the steel at the required rate. The choice of quenching medium is dictated by the hardenability of the steel. The quenching media may range from iced brine through water (sometimes with polymer additions) and oil to air. The amount of agitation or circulation of the quenching medium also affects the quench rate. If the quench rate is too rapid, quench cracks can form, whereas if it is too slow, transformation products other than martensite may form; proper selection of quenching media is therefore very important. The martensite structure formed on quenching is a body-centered tetragonal form of bcc iron whose hardness depends on the carbon content of the steel. Hardness values approaching about Rockwell C67 can be attained in the higher carbon steels.

Tempering treatments are undertaken to improve the toughness of as-quenched martensite. Martensite is a very hard, brittle phase, and tempering is required to produce a part with the required strength and toughness levels; no steel is used in the as-quenched condition. Microstructural changes that occur during tempering include precipitation of ϵ-carbide or Fe_3C, transformation of retained austenite to ferrite and carbide, and precipitation of alloy carbides. The exact transformations that occur in a given temperature range are classified by tempering stages (Speich, 1969; Speich and Leslie, 1972). While tempering generally improves toughness, it is important to note that tempering certain alloy steels in the range 200–400°C can result in a loss in toughness. This phenomenon is known as *tempered martensite embrittlement.*

In the *martempering*, the steel is first austenitized in the usual fashion and is then quenched into a salt bath at a temperature just above M_s. The steel is held at that temperature for a period of time to obtain a uniform temperature across the part but is quenched before transformation to bainite can begin. During the quench, transformation to martensite occurs and the martensitic structure is then tempered as necessary. The martempering treatment minimizes distortion and cracking tendencies. As in most quench and temper steel products, it is important that the alloy has sufficient hardenability to prevent transformation to other austenite decomposition products during the final quench. *Austempering* is another special heat treatment applied to higher carbon steels and certain cast irons to reduce cracking and distortion problems. The alloy is austenitized in the normal fashion and again quenched into a salt bath at a temperature above M_s. However, in austempering the steel is held at this temperature until it fully transforms to bainite. The process significantly reduces distortion and the resulting product has a higher toughness than tempered martensite at the same hardness level. Also of advantage is that in austempering treatments no tempering is required, although this may be partially negated by the longer time required for the transformation to be completed.

1.2.4 Precipitation-Hardening Heat Treatments Heat treatments that result in the precipitation of a fine dispersion of a second phase can be used to strengthen an alloy, since precipitates can impede the motion of slip dislocations. The presence of a second phase alone will not be sufficient to achieve significant hardening, but the second-phase particles must be fine and uniformly dispersed to be effective barriers to dislocation motion. The precipitation-hardening heat treatment is carried out in three major steps: solutionizing, quenching, and age hardening.

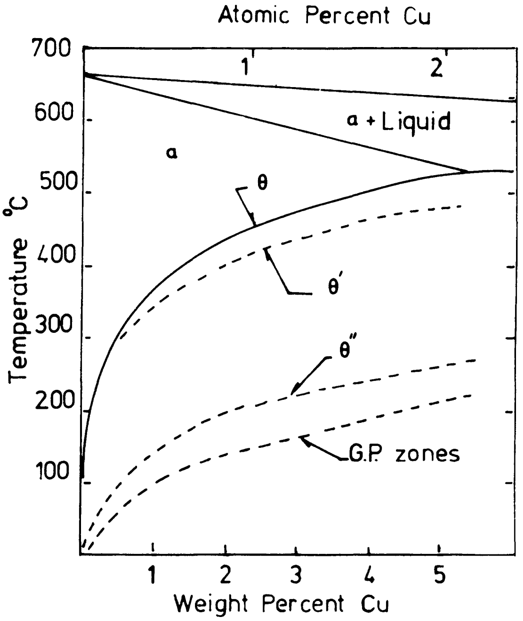

FIG. 5. Aluminum-rich portion of Al-Cu phase diagram, showing the equilibrium phase regions and the solvus lines for the metastable phases (Lorimer, 1978; reproduced with permission).

An alloy that is precipitation hardenable must contain a solute that has limited solubility in the solvent or matrix material. For example, consider an Al–4 wt% Cu alloy, with the relevant phase diagram shown in Fig. 5. The single-phase solid solution of Cu in Al (α) is stable between temperatures of 500 and 575°C. Below this temperature, the θ ($CuAl_2$) and α phases coexist in thermodynamic equilibrium. (The line indicating the temperatures below which the θ phase can form is known as the θ-phase solvus line. At further undercoolings, metastable, intermediate phases can form, each with its own solvus line, as shown in Fig. 5.) The solubility of Cu in Al decreases with temperature, and the equilibrium amount of the θ (or intermediate) phase increases with the degree of undercooling below the solvus temperature. The decrease in the solubility of the solute with decreasing temperature is an essential feature of the precipitation-hardening alloy systems.

The first step in the precipitation-hardening process is the solutionizing treatment. This is carried out by heating to a temperature above the equilibrium solvus. However, the temperature must be below the temperature at which the liquid phase will form (known as the solidus temperature). During solutionizing, most second phases will dissolve, leaving primarily the solid-solution phase. If the alloy were then very slowly cooled to room temperature, the equilibrium second phase would form, which in the Al-Cu system would be the θ phase. The second phase formed in this manner would be a relatively coarse dispersion which would not have a significant effect on the yield strength. However, if the alloy is instead rapidly quenched from the solutionizing temperature, precipitation of the second phase is avoided and a metastable, supersaturated solid solution is retained at room temperature.

The third step in precipitation hardening is the aging process. This is carried out by heating the alloy to a temperature below the solvus. The hardness will increase with time, as precipitation occurs, reach a maximum or peak hardness, and then start to decline. This final decrease in hardness is due to the coarsening of the precipitate and is known as overaging. An example of yield strength vs aging time for aluminum alloy 6061 (alloyed with 1% Mg, 0.6% Si, 0.2% Cr, and 0.3% Cu) is shown in Fig. 6 for several aging temperatures. Note that the peak hardness obtained is higher for lower temperatures but is only achieved after a longer aging time.

Because of the commercial importance of precipitation hardening, the process has been widely studied. The mechanisms by which precipitates impede dislocation motion are very complex and have been recently reviewed by Ardell (1985). The precipitation mechanism itself is also complex in nature and is discussed in detail in the text by Porter and Easterling (1981). One aspect of the precipitation process relevant to precipitation-hardening treatments is the precipitation of a sequence of metastable phases prior to the formation of the stable phase. For example, as the Al–4%Cu alloy is aged, it may first form coherent clusters of solute atoms 1 or 2 atomic diameters thick, which are known as Guinier–Preston (GP) zones. The formation of GP zones requires aging at a temperature below the GP-zone solvus, as shown in Fig. 5. Other metastable precipitates, such as the θ'' or θ' phases, can form on aging if aged below the relevant solvus lines. As aging proceeds, a sequence of metastable precipitates appears and eventually the equilibrium θ phase is obtained. A list of common precipitation-hardening systems and their precipitation sequences is shown in Table 1. The relation between strength and precipitate characteristics is complex, but in general, the metastable phases are more effective in strengthening an alloy. Therefore, aging at temperatures close to the equilibrium solvus will not be very effective in strengthening the alloy. Further examples of the effects of precipitation hardening on the mechanical properties of several aluminum alloys are shown in Table 2. Aluminum alloys that are age hardenable include both wrought and cast types. Several copper alloys are also precipitation hardenable; beryllium is a particularly effective alloying element in age hardening of copper alloys.

Steels that contain sufficient molybdenum or tungsten can also be strengthened by

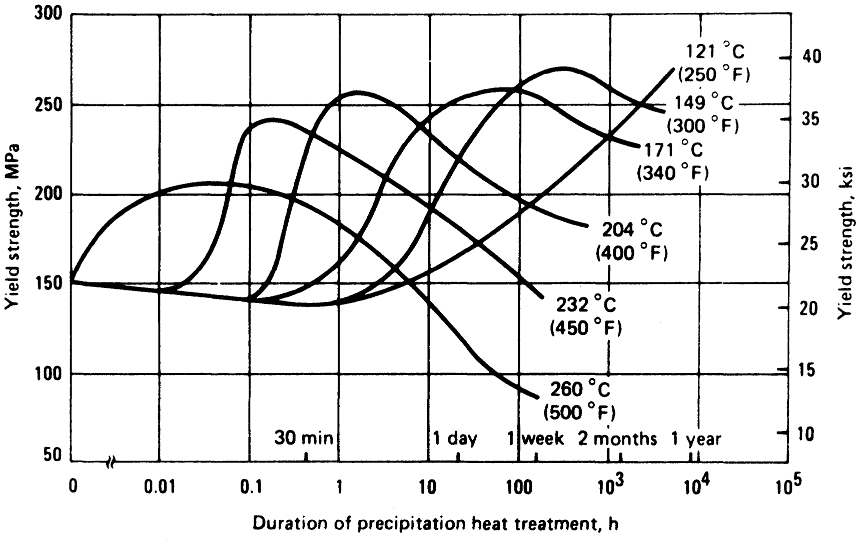

FIG. 6. Age hardening curves for a 6061 aluminum alloy (ASM, 1979, Vol. 2, p. 42; reproduced with permission).

Table 1. Precipitation sequences in age-hardening alloys.

Base metal	Solutes	Precipitation sequence
Al	Cu	GP Zones → θ'' → θ' → θ ($CuAl_2$)
Al	Ag	GP Zones → γ' (plates) → γ (Ag_2Al)
Al	Cu, Mg	GP Zones → S' (laths) → S ($CuMgAl_2$)
Al	Mg, Si	GP Zones → β' → β (Mg_2Si)
Al	Mg, Zn	GP Zones → η' (plates) → η ($MgZn_2$)
Cu	Be	GP Zones → γ' → γ (CuBe)
Fe	C	ϵ-carbide → Fe_3C
Fe	C, Mo, Cr	Mo_2C → M_3C, M_6C or $M_{23}C_6$ (M = Mo, Cr)

precipitation hardening if tempered near 550°C. A fine dispersion of Mo_2C or W_2C forms during this process, which is known as *secondary hardening*. Other steels that are strengthened by precipitation hardening are the maraging steels and certain types of stainless steels. One class of nickel-based alloys, known as superalloys, is also strengthened by precipitate formation. Further details on these alloy systems can be found in the text by Smith (1981).

2. MATERIALS TREATMENTS FOR MODIFICATION OF ELECTRICAL PROPERTIES

Materials treatments can be applied to alter the electrical properties of materials. In addition, treatments applied for other purposes can affect electrical properties. In this section, examples in both subject areas are discussed.

2.1 Effect of Materials Treatments on the Conductivity of Metals

The cold-work and annealing processes affect the electrical conductivity of a metal. As a metal is cold worked, defects are generated and conductivity is reduced; as the defects are annihilated during annealing, the conductivity is restored. In general, the conductivity of a metal is improved by annealing out defects and can also be significantly improved by removing or reducing impurities.

The resistivity of metals as a function of temperature can be described in the following manner: at low temperatures (near 0 K), the resistivity has a constant value, but as the temperature is increased, the resistivity increases linearly with temperature. This behavior is due to the nature of collision of electrons with lattice imperfections. These imperfections are generally of two types: lattice vibrations, or phonons, and collision with static impurities. Static impurities can be foreign atoms or crystal defects. The effect of static impurities on resistivity is related to the average time between collisions, which decreases with increasing impurity concentration and scattering cross section of the impurity. The impurity contribution to the total resistivity is termed the residual resistivity and is independent of temperature, whereas the temperature-dependent lattice vibration contribution is termed the ideal resistivity. At sufficiently high temperatures, the residual resistivity becomes negligible in comparison to the ideal resistivity.

Table 2. Typical mechanical properties of age-hardenable aluminum alloys.

Alloying elements (wt %)	Condition[a]	Tensile strength (MPa)	Yield strength (MPa)	Percent elongation
6.3 Cu-0.3 Mn	Annealed	175	70	20
	Aged	420	294	10
4.4 Cu-1.5 Mg0.6 Mn	Annealed	189	77	20
	Aged	483	400	10
1 Mg-0.6 Si-0.2 Cr-0.27 Cu	Annealed	126	56	25
	Aged	315	280	12

---

[a] Values for the aged condition refer to "T6" temper, obtained by solutionizing and artificial aging.

The effect of alloying and deformation on the conductivity of copper is shown in Table 3. For high-purity annealed copper, the conductivity is 101% IACS. Deformation reduces this value to 98% IACS, as a result of interaction of electrons with dislocations. A more substantial reduction in conductivity is realized by alloying, as shown in the Cu–35%Zn alloy, because of scattering from Zn solute atoms. Copper dispersion strengthened with 0.7% Al_2O_3 also demonstrates reduced conductivity, but in this case, the Al_2O_3 particles do not alloy with the copper, leaving the copper in a higher purity condition and therefore retaining a substantial amount of the conductivity of pure copper.

2.2 Thermal Treatments for Semiconductor Materials

The presence of both lattice defects and impurities can be detrimental to the performance of semiconductor devices, and annealing and gettering processes have been developed to reduce defect and impurity concentrations. Dislocations in semiconductors represent a row of dangling electrons. The dislocation core becomes negatively charged and repels electrons, forming a space-charge region with a positive charge about the dislocation. This causes the conduction electrons to follow a more winding path, decreasing mobility and increasing resistivity. In addition, dislocations can act as carrier recombination and generation centers. Dislocations can also getter metallic impurities, allowing metal precipitates to nucleate at dislocations and provide an electrically conductive path, which, if threaded through a *p-n* junction, will short the junction and render the junction unusable.

Dislocations can be present as a result of the crystal growth process. The process of ion implantation, used to dope semiconductors, may create defects or disordered regions. When an energetic ion enters the silicon, it can displace silicon atoms. If the dose is sufficient, an amorphous layer will form. The amorphous layer may be continuous up to the surface, or a buried amorphous layer can form, depending on the mass of the ion, the dose, and the substrate temperature at the time of implantation. The amorphous nature of the implanted region is undesirable, since location of the dopant atoms at substitutional sites in a crystalline lattice is often required to obtain the desired electrical properties.

Table 3. Conductivity of copper and copper alloys.

Alloy	Conductivity, percent IACS
Pure annealed copper	101
Pure copper deformed 80%	98
Cu-35% Zn	28
Cu-0.7% Al_2O_3	85

2.2.1 Annealing Treatments Annealing treatments are used to reduce defect concentrations and to make the dopants electrically active. Annealing can be done using a variety of heat sources, including furnaces, lasers, electron beams, plasmas, and incoherent light sources.

Furnace annealing is the most conventional process for annealing. The effect of furnace annealing is to transform an amorphous layer to a crystalline one and, in some cases, to redistribute the dopant atoms. During annealing, the amorphous layer crystallizes by regrowth from the amorphous/crystalline interface by the process of solid-phase epitaxy. Dislocation loops and networks are formed in the regrown layer. The sharpness of the interface can have a dramatic effect on the defect density of the regrown layer. Sharp interfaces tend to regrow layers with lower defect densities than diffuse ones.

The processes of pulsed or rapid thermal annealing are used to anneal amorphous regions in ion-implanted semiconductors. As noted, rapid thermal annealing techniques can employ lasers, electron beams, or incoherent light sources such as arc lamps or graphite heaters. Typical heating times are 1–100 s. Rapid annealing has several advantages over conventional annealing, including more rapid reductions in defect concentrations and the ability to reduce defect concentrations to lower levels. Another advantage is reducing the extent to which other undesirable metallurgical reactions occur. For example, pit formation may occur in silicon that is in contact with aluminum metallization. This problem can be alleviated by rapid thermal processing.

2.2.2 Gettering Treatments The purity of silicon wafers for semiconductor applications can be improved by the process of *gettering*. In *extrinsic* gettering, dislocations are introduced to the unused back side of the waver by abrasion, ion implantation, or laser heating. Oxygen as well as metallic impurities will be attracted to the dislocations, since they can reduce dislocation strain energy. Higher-purity silicon is obtained at the top wafer surface where active devices will be placed. The process of *intrinsic* gettering is used to remove oxygen impurities from active device areas in silicon. This is accomplished by causing the precipitation of SiO_2 particles. In a three-step annealing process described by Wolf and Taub (1986), the wafer is first heated to a high temperature to permit out-diffusion of oxygen from the surface, followed by a low-temperature anneal in which SiO_x precipitates form, after which the wafer is heated to a higher temperature at which growth of SiO_2 precipitates occurs. In practice, both extrinsic and intrinsic gettering processes may be used to reduce effective impurity concentrations in silicon wafers.

3. SURFACE MODIFICATION OF MATERIALS

Surface treatments of materials are undertaken to obtain surface or near-surface properties which are in some way different from the bulk. Reasons for using surface treatments include improving wear resistance or corrosion resistance, altering friction properties, changing the dimensions of a material, or achieving certain decorative or aesthetic goals. For glasses, surface treatments are used to alter the surface properties in such a way that bulk mechanical strength is improved.

Surface treatments can be roughly classified by the manner in which they change the surface. One class of surface treatments primarily changes the structure of the surface, and this class includes *flame* or *induction* hardening. A second class alters both the composition and structure of the surface, such as *carburizing*. A third class adds material on top of an existing surface, with little or no change to the structure or composition of the existing surface. This class includes coating methods, such as *electroplating, ion plating*, or *physical vapor deposition*. In this section, surface treatments are discussed in subsections corresponding to the method used to alter surface properties. A separate subsection is used to discuss special treatments for surface modification of glass.

3.2 Thermal Treatments for Surface Modification

Thermal methods for surface treatment are applied primarily to steels with carbon contents above 0.4%. The objective of a thermal surface treatment is to change the structure of the surface to a metastable but harder phase, typically martensite. The core material retains its original microstructure, which has the desired level of strength and toughness. An additional advantage of these treatments is that some can be applied selectively to appropriate but limited areas of a surface.

The process of *flame hardening* is no doubt one of the oldest methods of surface treatment of steels. In this process, a flame created using an appropriate fuel gas is directed at the surface of the steel, raising its temperature and transforming it to austenite. The heat source is then removed and the surface is quickly quenched with water or oil. The surface transforms to martensite with a high hardness level. An alternative source of heat input is obtained using induction heating, leading to the process known as *induction hardening*. In this process, the part is placed in a coil, usually made of copper tubing, and an alternating current flows through the copper while cooling water flows through the inside of the tube. This current induces a current in the work piece and results in heating of the piece. The part heats very rapidly and is then quenched in the appropriate quenching medium. The limitation of this process is that the piece must be able to fit within the coil. Other methods can be used to heat the surface of a steel and achieve surface hardening, including *laser* and *electron-beam hardening*. Since electron-beam hardening must be carried out in vacuum, gas quenching is used instead of liquid quenching.

3.2 Diffusion Processes

3.2.1 Carburizing Carburizing processes increase the carbon content near the surface of a steel by utilizing a diffusion process. In this process, the carbon content in the surface of the part is increased by one of a variety of methods. At elevated temperatures, carbon diffuses into the part, creating a varying carbon content high at the surface and decreasing for some distance toward the interior. This region of carbon content variation is termed the *case*. Low-carbon (0.2%) steels are the materials most commonly carburized, where the carbon content is raised to about 0.9% in the surface region. The consequence of this increase is a possible 200% increase in both strength and hardness in the case. High yield strengths can be developed via heat treatment to resist high operating stresses, as in cases produced on gear teeth, while high hardness can accommodate severe wear conditions. Heat treatment of carburized steels usually begins immediately following carburization; in the nitriding process, discussed below, the heat treatment occurs prior to nitriding.

The governing process variables in carburizing are time, temperature, and surrounding environment. Gas carburizing and the other carburizing methods, pack carburizing, liquid carburizing, vacuum (low gas pressure) carburizing, and plasma (ion) carburizing, employ a variety of methods to create various carburizing environments. By far the most common environment is a gas composed of a reacted mixture of air and methane (CH_4). This gas is principally nitrogen, hydrogen, and carbon monoxide (CO), with small amounts of carbon dioxide (CO_2) and water; the common term for this gas is endogas. Close control of gas composition in the furnace is vital. This control is accomplished by addition of small amounts of CH_4 or air to the furnace atmosphere; CH_4 increases the carburizing effect while air decreases it. Endogas contains nitrogen (diatomic) but no reaction between that form of nitrogen and the work occurs. At any fixed temperature, the gas composition in the furnace determines the carbon content in the surface of the part, i.e., determines the boundary condition. Temperatures commonly are in the range 875–950°C; 925°C is most common. At this temperature, a case of 0.25 mm develops in 1 hour and case thickness increases with the square root of time. The effect of temperature on the rate of diffusion is substantial; a 10°C change in temperature causes the rate to change by 13%.

Pack carburizing utilizes an environment called a carburizing compound that consists of granules of some form of carbon such as coke along with an active compound that is some mixture of alkali or alkaline earth carbonates; $BaCO_3$ is common. The work is packed into a steel container along with the carburizing compound and sealed to exclude air. At the carburizing temperature (about 900°C), carbon and carbonates react to form principally CO. Hence, the boundary condition is similar to that in gas carburizing except that there is no control of the boundary condition; the carbon level is the saturation value at the pack carburizing temperature.

Additional carburizing methods are *liquid carburizing*, which is carried out in a molten salt bath at about 900°C; *plasma carburizing*, which utilizes a glow discharge in a low-pressure (1 to 25 torr) hydrocarbon gas; and *vacuum carburizing*, which involves carburizing at pressures less than atmospheric.

3.2.2 Nitriding Nitriding processes increase the nitrogen content by utilizing the diffusion mechanism. In most respects, nitriding involves the same procedures and phenomena as found in carburizing and only the essential differences will be outlined herein.

The nitriding environment can be created by a variety of methods, but the most common environment is formed from dissociated ammonia (NH_3). Nascent (monatomic) nitrogen forms and this establishes the boundary condition. A wide variety of steels are the materials most commonly nitrided; steels containing aluminum, chromium, vanadium, tungsten, and molybdenum are the most compatible with nitriding, because these elements form nitrides during processing. A brittle, usually undesirable, iron nitride layer often forms during nitriding; this is called the white layer. Usually the extent of the white layer can be avoided or minimized by proper combination of temperature and degree of ammonia dissociation. Other nitrides form in the case; these are usually desirable because wear resistance is significantly improved. Nitriding of steels is carried out in the 475–575°C range. Heat treatment is carried out prior to nitriding at temperatures below 475°C. In *liquid nitriding*, the nitriding environment is established using molten salt mixtures of cyanides, cyanates, and sometimes carbonates or chlorides. *Plasma nitriding* involves the same phenomena as plasma carburizing in that NH_3 is fragmented by a glow discharge.

Carburizing and nitriding processes can be carried out simultaneously. The environment is usually a mixture of gases arising from an initial mixture of NH_3 and hydrocarbons. Some carburizing processes are simply modified by NH_3 additions. Processes such as carbonitriding and nitrocarburizing are used to develop unique properties.

3.3 Electrochemical and Vacuum Coating Methods

Electrochemical and vacuum methods are used primarily to add material to a surface, that is, to apply a coating. These methods often change the dimensions of a piece, more significantly with electrochemical methods, and are sometimes intentionally used for this purpose. Coatings are widely used to improve wear and corrosion resistance.

The process of electroplating is carried out by placing in a solution, called the electrolyte, two metals that are connected to an electrical power supply. Atoms of one metal can be dissolved to form positive ions in solution, although in most cases, an electrolyte with the desired ions already in solution is used (in which case the anode reaction need not be dissolution). The ions can migrate from the anode to the cathode where they are reduced and adhere to the cathode. The flux of deposited ions is usually supplied by diffusion in the boundary layer, with migration being negligible. Typical plating rates are on the order of 25 μm/h. A wide variety of metals can be electroplated; chromium, nickel, copper, iron, silver, gold, palladium, zinc, and tin are commonly plated, as well as alloys such as brass and bronze. Plating can be done on most metallic substrates, although metal substrates that form adherent oxides, such as aluminum, stainless steel, or chromium, can be difficult to plate.

Electroless plating is used to obtain metal coatings without the use of an external power supply. In this process, the ions to be plated are reduced while simultaneous anodic oxidation of hydrogen ions occurs, providing the electrons for reduction of the metal. Nickel, copper, and gold can be plated using this method. Deposits of electroless nickel can be very hard; hardnesses of 65 HRc can be obtained after heat treatment of the plated coating. Because of this property, electroless nickel coatings can provide good wear resistance.

Coatings can also be formed by converting the base metal to a metal compound. *Conversion coating* processes are used for this purpose. *Phosphate* coatings convert the metal to a phosphate compound. These coatings are produced by dipping or spraying the part with a solution of dilute phosphoric acid (normally less than 1% concentration) which has been heated to below the boiling point. The process is usually applied to steels. Coatings of manganese phosphate are often used to break in machine parts that may be slightly mismatched. *Chromate* coatings are typically used on zinc or aluminum to protect these metals in harsh or corrosive environments. In this process, the part is immersed in a dilute aqueous solution of chromic acid or chromium salts. Chromium atoms then form an amorphous layer on the metal part. A conversion coating type process specific to aluminum is known as *anodizing*. Aluminum naturally forms a hard and tenacious layer of aluminum oxide, but the anodizing process allows production of thicker layers than would normally be found after exposure to air. Layer thicknesses range from 2.5 to over 50 μm. Thinner coatings are used for decorative purposes, while thicker coatings can provide corrosion and wear resistance. Anodizing is done in an electrochemical cell. Lead is normally used as the cathode, Al is the anode, and the electrolyte contains a dilute solution of acids such as sulfuric acid, chromic acid, or phosphoric acid. The aluminum is converted to aluminum oxide by an electrochemical conversion process.

Vacuum coating or physical vapor deposition methods are also used to deposit coatings on substrate materials. A wide variety of vacuum coating methods have been developed, including *thermal evaporation*, *e-beam evaporation*, *sputtering*, and *ion plating*. No attempt will be made here to provide detailed descriptions of these processes, but the interested reader is referred to Maisel and Glang (1970) as well as the article THIN FILM DEPOSITION in this Encyclopedia. In comparing vacuum coatings with electrochemical coatings, it can be noted that vacuum coatings tend to be thinner than electrochemical coatings as a result of lower deposition rates and that equipment costs for vacuum coating are greater than that for electrochemical coating. However, vacuum coating methods can be used to produce films of both metals and nonmetals, and alloy or layered films are often more easily produced by vacuum methods.

For completeness, two other methods of surface modification will be mentioned here. The first is *chemical vapor deposition* [see CVD (CHEMICAL VAPOR DEPOSITION)], in which a coating is produced by reaction between gaseous species. Deposited compounds include SiO_2, Si_3N_4, TiB_2, SiC, and TiC. The second method is *ion implantation* (see ION-BEAM MODIFICATION OF MATERIALS), in which energetic ions are directed toward a substrate and implanted near the substrate surface. This method is often used to dope semiconductor materials. For further details on these methods, the reader is referred to the appropriate article of this Encyclopedia.

3.4 Surface Treatments to Strengthen Glass

Glass is an amorphous solid material. On cooling glass from the liquid state, a progressive increase in the viscosity is observed until the glass becomes rigid below the *glass transition temperature* T_g. Because of their noncrystalline structure, glasses resemble undercooled liquids. Most glass contains large amounts of silica or SiO_2; fused silica contains at least 99.5% SiO_2. Other glasses contain lesser amounts of SiO_2 but contain other oxides such as Na_2O, PbO, Al_2O_3, CaO, B_2O_3, and MgO. One of the most common glasses is "soda-lime" glass, which contains SiO_2, Na_2O, CaO, Al_2O_3, and MgO.

Glasses can suffer from low fracture toughness, particularly at low temperatures. While the theoretical cohesive strength of glass is very high (18 000 MPa for silica glass), design strengths of glasses are 500–1000 times lower. The low engineering strength is due to the presence of surface cracks or flaws. Stresses at the tips of these cracks can be many times the applied stress, greatly reducing the tensile strength. The stress at a crack tip is proportional to the crack depth and inversely proportional to the crack tip radius.

Thermal stresses develop in glasses through the difference in cooling rates between the interior regions and the surface of the glass. These stresses can be reduced by very slow cooling from the working temperature. They can also be reduced by heating to the annealing point of a the glass (about 500°C for soda-lime glass) and slow cooling. Annealing at high temperatures causes the glass to become strain free more quickly, but very slow cooling must be employed to avoid residual stress development on cooling. Annealing at lower temperatures requires longer times for the glass to become strain free, but more rapid cooling can be used.

Differential cooling rates in a specimen are used to improve the strength of glass in the *thermal tempering* process. To temper a piece of glass it is first heated to above T_g. During cooling, jets of air or oil are directed at the surfaces of the glass. The surface becomes rigid before the interior, but on further cooling, the interior continues to contract. As a result, compressive stresses develop at the glass surface. Upon loading, the magnitude of an applied tensile stress must first exceed the magnitude of the residual compressive stress before crack propagation and fracture are likely to occur.

The method of *fire polishing* has been used to obtain very high strengths in glass. In this method, a filament of glass is drawn out through a flame. Fire-polished rods of fused silica have demonstrated strength levels of up to 14 000 MPa at temperatures of -196°C. However, such strength levels are short-lived because the specimens are very susceptible to surface damage or corrosion. Another method for removing surface flaws it to etch a glass in a hydrofluoric acid solution. Even if the flaws are not completely removed, the crack tip radii can be increased, leading to lower stress concentrations.

Another chemical method for improving glass strength is *ion exchange* (Doremus, 1973). In this method, a large ion in the glass is exchanged for a smaller one, causing compressive stresses to develop at the surface. This process may also change the thermal expansion coefficient of the surface, which can enhance development of residual compressive stresses at the surface. The effects of relative ion size, changes in thermal expansion coefficients, and thermal stresses must all be considered in determining the effectiveness of a particular ion exchange process. Another possible result of ion exchange is surface crystallization. If the crystals have a lower thermal expansion coefficient than the bulk, then compressive surface stresses develop on cooling. However, in most cases, surface crystallization causes tensile stresses to develop on cooling and weakens the glass.

ACKNOWLEDGMENTS

The author gratefully acknowledges contributions from Dr. Martin G.H. Wells, U.S. Army Research Laboratory, Materials Directorate, Watertown, MA, for contributing to the section *Heat Treatment of Ferrous Alloys,* and from Prof. Samuel J. Hruska, Purdue University, West Lafayette, IN, for the section *Diffusion Processes.*

GLOSSARY

Austenite: An equilibrium phase in the Fe-Fe_3C system which is a solid solution of carbon in fcc iron. The definition of austenite is often extended to include the fcc phase in alloy-steel systems.

Bainite: A microstructure in steels that is made up of α-ferrite and iron carbide. In one form of bainite, the structure consists of very fine nonlamellar cementite particles in an α-ferrite matrix.

Cementite: A name used to identify the Fe_3C compound that forms in steels. The name derives from the hardness of the cementite phase.

Dislocation: A line defect in a crystal. The motion of dislocations results in permanent or plastic deformation of the crystal.

Eutectoid Reaction: A phase transformation in which a single solid phase is transformed to two different solid phases on cooling.

Ferrite: An equilibrium phase in the Fe-Fe_3C system that is a solid solution of carbon in bcc iron. Two types of ferrite are present in the system, a low-temperature ferrite, denoted α-ferrite, and a high temperature ferrite, denoted δ-ferrite.

Grain: A portion of a solid material that consists of only one crystal.

Grain Boundary: A planar defect at the boundary between two grains of different orientations.

Guinier–Preston (GP) Zones: Coherent clusters of solute atoms one or two atomic diameters thick, which can form during the early stages of solid-state precipitation.

Hardness: The ability of a material to resist indentation. A variety of hardness tests have been established, and results are given with reference to the test used. For example, for testing steel specimens, the Rockwell "C" hardness test is used, in which a diamond indentor is utilized and a 150-kg load applied; for nonferrous alloys, the "B" scale is used (using a $\frac{1}{16}$-in. steel ball indentor with a 100-kg weight).

Martensite: A metastable phase formed in steels, which has bct structure and a high hardness. It forms almost instantaneously by a diffusionless transformation if a steel is quenched at a sufficiently rapid rate to a sufficiently low temperature.

Pearlite: A microstructure formed in steels, which consists of α-ferrite and cementite in a lamellar arrangement.

Recrystallization Temperature: The temperature required to achieve recrystallization of a material in a fixed time, usually 1 h.

Sheet Texture: Texture refers to the alignment of crystal directions or planes in a polycrystalline structure relative to the geometry of the part. This alignment is often found in rolled sheet or plate, where it is known as sheet texture.

Solid Solution: An alloy that consists of two or more elements but is single phase.

Solidus: The temperature below which an alloy is completely in the solid state.

Solutionize: A heat-treatment process used to obtain a homogeneous, solid-solution single-phase structure in an alloy. During solutionizing, all second phases are completely dissolved.

Solvus: The temperature below which a particular equilibrium or metastable phase can form by precipitation.

Tempered Martensite Embrittlement: A phenomenon in which, after tempering of a martensitic steel, a loss in toughness is realized.

Toughness: The ability of a material to resist failure by crack propagation. It is also the amount of energy the material can absorb before fracturing.

Yield Strength: The stress level at which a material begins to deform plastically. Below the yield strength, only elastic deformation occurs.

Works Cited

Ardell, A. J. (1985), *Metall. Trans.* **16A**, 2131–2165.

American Society for Metals (1979), *Metals Handbook*, 9th ed., Metals Park, OH: ASM.

Askeland, D. R. (1989), *The Science and Engineering of Materials*, Boston: PWS-Kent, pp. 187, 331, and 362.

Cahn, J. W. (1962), *Acta Metall.* **10**, 789–798.

Doremus, R. H. (1973), *Glass Science*, New York: Wiley, Ch. 17.

Lorimer, G. W. (1978), "Precipitation in Aluminum Alloys," in: K. C. Russell, H. I. Aaronson (Eds.), *Precipitation Processes in Solids*, Warrendale, PA: TMS, p. 88.

Mielnik, Edward M. (1991), *Metalworking Science and Engineering*, New York: McGraw-Hill, pp. 137–138.

Maisel, L. I., Glang, R. (Eds.) (1970), *Handbook of Thin Film Technology*, New York: McGraw-Hill.

Porter, D. A., Easterling, K. E. (1981), *Phase Transformations in Metals and Alloys*, Berkshire, England: Van Nostrand Reinhold, Ch. 5.

Reed-Hill, R. E., Abbaschian, R. (1992), *Physical Metallurgy Principles*, 3rd ed., New York: Van Nostrand, Ch. 16.

Smallman, R. E. (1985), *Modern Physical Metallurgy*, 4th ed., London: Butterworths, p. 418.

Smith, W. F. (1981), *Structure and Properties of Engineering Alloys*, New York: McGraw-Hill.

Speich, G. R. (1969), *Trans. AIME* **245**, 2553–2564.

Speich, G. R., Leslie, W. C. (1972), *Metall. Trans.* **3**, 1043–1054.

Wolf, S., Taub, R. N. (1986), *Silicon Processing for VLSI Era, Vol. 1, Processing Technology*, Sunset Beach, CA: Lattice Press, Ch. 2.

Further Reading

Askeland, D. R. (1989), *The Science and Engineering of Materials*, Boston: PWS-Kent.

Shackleford, J. F. (1992), *Introduction to Materials Science for Engineers*, New York: MacMillan.

Meyers, M. A., Chawla, K. K. (1984), *Mechanical Metallurgy, Principles and Applications*, Englewood Cliffs, NJ: Prentice-Hall.

Reed-Hill, R. E., Abbaschian, R. (1992), *Physical Metallurgy Principles*, 3rd ed., Boston: PWS-Kent.

Smallman, R. E. (1985), *Modern Physical Metallurgy*, 4th ed., London: Butterworths.

Doremus, R. H. (1973), *Glass Science*, New York: Wiley.

ASM International (1979, 1980), *ASM Metals Handbook*, 9th and 10th Eds., Metals Park, OH: ASM.

MATHEMATICAL METHODS

See ALGEBRAIC METHODS; ANALYTIC METHODS; FOURIER AND OTHER MATHEMATICAL TRANSFORMS; GEOMETRICAL METHODS; GREEN'S FUNCTIONS; GROUP THEORY; MATHEMATICAL MODELLING; MONTE CARLO METHODS; PERTURBATION METHODS; QUANTUM LOGIC; SEMANTIC AND OTHER LOGIC METHODS; SPECIAL FUNCTIONS; THEORY AND APPLICATIONS OF FUZZY LOGIC; VARIATIONAL METHODS

MATHEMATICAL MODELING

KENNETH HARTT, *Physics Department, University of Rhode Island, Kingston, Rhode Island, U.S.A.*

	Introduction	417
1.	**About Models: Types and Trends**	418
1.1	Chaos as a New Paradigm in Science	419
1.2	Complex Systems as a Focus of Mathematical Modeling	419
1.3	Computer Simulation Experiments as a Third Branch of Science	419
2.	**Digital Computing Developments**	420
2.1	Hardware	420
2.2	Software	420
3.	**Selected Models and Modeling Tools**	421
3.1	Fractals	421
3.2	Nonlinear Dynamics	423
3.2.1	Deterministic Models	423
3.2.2	Stochastic Models	423
3.2.3	Formal and Experimental Dynamics	423
3.2.4	Time-Delay Embedding	424
3.2.5	Cellular Automata	424
3.3	Rational Functions	424
3.4	Monte Carlo Methods	425
3.5	Numerical Modeling	426
3.5.1	Chaotic Systems	427
3.5.2	Finite Elements	428
3.5.3	General Circulation Models	428
3.5.4	Data-Assimilation Models	429
3.6	Wavelets	429
3.7	Systems Analysis	430
3.7.1	Control Theory	431
3.7.2	Self-Organization	432
4.	**Information Processing Models**	432
4.1	Principal-Component Analysis	432
4.2	Neural Networks	433
4.3	Time-Series Analysis	434
4.3.1	Signal Processing	435
4.4	Statistical Inference	436
4.5	Information Theory	436
5.	**Applications**	437
5.1	El Niño	437
5.2	Chaotic Ocean Heat Transport	439
5.3	Controlling Chaos	440
5.4	Solitons	440
	Glossary	441
	Works Cited	442
	Further Reading	444

INTRODUCTION

A mathematical model is a set of equations and algorithms, together with a collection of interpretive rules, designed to represent something. Mathematical models are invariably approximations. They include physical laws that, like Newtonian mechanics, the special and general theories of relativity, or quantum mechanics, are guiding precepts at the moment, within their specified limits. They also include exponential growth models that are Band-Aids, meant to be replaced. Mathematical models selectively magnify some parts of a system while ignoring or severely approximating other parts. The art of modeling is to distinguish essential from superfluous detail and to be able to "consider a spherical cow" (Harte, 1988). To learn of the goals, philosophy, and methods of mathematical modeling within any field of physics, one can do nothing better than read the article in this Encyclopedia pertaining to the field. The purpose here is to sample the breadth of emerging tools and techniques.

Change through computer power is the dominant feature in the world of mathematical modeling. During the brief history of electronic computation, barriers to formal mathematics eroded slowly. In 1976, the four-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

color problem was solved: It was proved that only four colors are needed for maps to assure that no two countries sharing a common border have the same color (Appel and Haken, 1978). This is a combinatoric problem, solved by making use of a computer in essential ways. Since then, computer-assisted researchers have won acceptance as not being children of a lesser God.

Although analytic methods and rigorous mathematics theorems continue to be pillars, they are enhanced by modern-day computers, sometimes through the use of symbolic computer software. Computer technology at present can give us commercially available teraflop speeds (10^{12} floating point operations per second), and will perhaps give 10^{15} flops in ten years (Pool, 1992). Massively parallel processing already exists. Every two years produces workstations and personal computers using more sophisticated graphics and software, making the previous generation obsolete. A researcher must carefully choose the appropriate hardware and software system, because of overhead costs in financial resources (measurable in units of $10 000) and learning time (units of 6 months). The reward is the power to explore new models of greater complexity and old models to greater depth and to share rapidly advancing computer visualization.

Just as a treatise on mathematical modeling 180 years ago would have been remiss to omit mention of Fourier analysis, so would it be remiss today to omit mention of wavelets, chaos, fractals, nonlinear dynamics, numerical modeling, rational functions, symbolic computation, time series, Monte Carlo and inverse methods, systems analysis, and principal-component analysis. In their present form, they all are children of the computer age, with a scope extending beyond the context of their inception. We concentrate on these and add statistical inference, information theory, cellular automata, self-organization, and data assimilation, which are comparable in breadth of interest. We also mention software development.

Our discussion is illustrative rather than comprehensive and is not rigorously parallel or hierarchical. The recent treatise on mathematical modeling with which this article most overlaps is the two-volume work of Casti (1992). Each work contains things the other does not. Whereas Casti tends to emphasize techniques more than the models, and the journal *Applied Mathematical Modelling* emphasizes models more than techniques, we attempt to strike a median between the two. Our physicist's viewpoint relegates—perhaps unfairly—numerical modeling to less than 10% of the whole. Consequently, some important numerical analysis developments such as splines, singular integrals, and large matrix inversion are given little or no space. Whether one uses commercial numerical software systems or does one's own numerical programming, we advise caution and careful testing. We refer the reader to the article NUMERICAL METHODS for more detailed advice. By mathematicians' standards, this work is decidedly qualitative. We omit or relegate to the incidental almost all the standard fare of a first course on mathematical methods of physics, such as tensor analysis, finite- and infinite-dimensional linear spaces, distribution theory, Fourier and Hilbert transforms, special functions, methods of solving integral and differential equations, and group theory, for which excellent textbooks exist (Hassani, 1991; Arfken, 1985). What remains are topics we believe practicing scientific modelers can ignore only at their own peril!

1. ABOUT MODELS: TYPES AND TRENDS

We follow the classification of climate modelers (Schneider and Dickinson, 1974) in identifying two types of models: *mechanistic models*, which emphasize mechanisms; and *simulation models*, which introduce explicitly as many relevant degrees of freedom as possible. Most models are a mixture of the two. Consider, for example, models designed to simulate circulations in the atmosphere and the oceans. They are called general circulation models (GCM's). The GCM's do not possess a grid scale sufficiently fine to permit simulation of clouds. However, insight gained from cloud simulation models within which environments of clouds are mechanistically described has led to a better mechanistic description of cloudiness within GCM's. The interplay between mechanistic and simulation models also sheds light on feedbacks.

A mechanistic description of a subphenomenon within a larger simulational system is called *parametrization*. In GCM's, both cloudiness and precipitation are parametrized. The more mechanistic a model, the more directly observable are its variables. For example, pressure as an operationally defined variable in thermodynamics, a mechanistic model, contrasts with pressure as an end product of a lengthy computation in kinetic theory, a simulation model. The ideal gas law is an often-used parametrization. The validations of parametrizations within GCM's, particularly for climatic regimes other than the present and, hence, only indirectly accessible to observation, have been perhaps the major issue in their development and are crucial in the study of global warming.

Models are further classified according to their relation to experimental observations. *Direct models* make predictions for comparison with experiment, using given parameters. *Inverse models* make predictions of parameters of direct models, using experimental data. *Data-assimilation models* make direct predictions of the near future or more complete descriptions of the present, while using inverse methods for the purpose of updating and improving parameter values. Models that show temporal evolution are called *dynamic models*. Dynamic models are further classified as *deterministic* if the rule by which the next successive state in time computed is unique or *stochastic* if that rule can lead to more than one next state, with some probability.

We comment on two environmental systems: El Niño and the Southern Oscillation, in which various models are discussed; and North Atlantic Ocean circulation, where mechanistic models reveal a possibility of low-dimensional chaos. Also, we touch on two major successes: controlling chaos and soliton propagation in optical fibers. Three dominant motifs recur throughout the present discussion. They are centered on chaos, complex systems, and computer simulation, as we now explain.

1.1 Chaos as a New Paradigm in Science

Chaos is the apparently random behavior seen in some deterministic nonlinear dynamical systems. Noisy processes once thought to require stochastic modeling may be chaotic (Tufillaro *et al.*, 1992). Even in planetary motions, chaos has a long-time destabilizing effect. Chaotic systems, intrinsically classical, have a poorly understood relationship with quantum mechanics. Therefore, chaos is sometimes advanced as a fundamentally new point of view, a new paradigm (Ford and Martica, 1992).

Distinguished from randomness by measures of order, chaotic evolution always has a fractal orbital structure, and there is always sensitivity to initial conditions (SIC). Both conservative and dissipative systems can be chaotic. Dissipative chaotic systems evolve to a low-dimensional asymptotic orbit, called an *attractor*. The attractor is reached from initial conditions in a region of phase space called the *basin of attraction*. Since lower dimensions are easier to work with, the existence of an attractor helps analyze dissipative chaos. There is information loss, measurable in bits per time step. Weather, with its long-range unpredictability, is probably chaotic.

1.2 Complex Systems as a Focus of Mathematical Modeling

As prime examples of complex systems, GCM's are heavily parametrized. Systems analysis, emphasizing the relationship among the components of a system, is relevant here; mechanistic models of complex systems use systems concepts of feedback and control. In other developments, phase-transition models for aggregate matter are used in the theory of strongly interacting elementary particles, self-organization, and chaos. Adaptive learning systems such as neural networks, themselves models of complex systems, are used heuristically for model building (Zurada, 1992).

1.3 Computer Simulation Experiments as a Third Branch of Science

Simulation of complex or chaotic systems generates large data sets. Complex systems may be deterministic, such as with molecular dynamics (Yonezawa, 1993), stochastic, as in Monte Carlo simulations (Creutz, 1992), or a mixture of the two. Interesting, complicated data are also produced from simulation of a chaotic one-dimensional nonlinear driven oscillator. Investigations using such computer-generated data have become a dominant effort in scientific work and are sometimes referred to as *experimental theoretical physics* (ETP). Improved digital computer power will make possible ETP simulations of aerodynamic flows and many important systems, which can also include hypothetical structures such as exotic chemical compounds not yet seen in the laboratory.

2. DIGITAL COMPUTING DEVELOPMENTS

Trends in digital computing today are

1. the explosion of computer power, both serial and parallel;
2. a rapid increase in connectivity ("networking" and "clustering") making standalone computing obsolete;
3. a rapid increase in specialized software; and
4. the broadening scope of software systems.

2.1 Hardware

The order of magnitude of current workstation speed is 100 megaflops. This resembles supercomputer speed of just ten years ago. It is slower by many orders of magnitude than current linear supercomputers, now in the gigaflop range; however, through parallel links of the slower workstation processors, competitive parallel computers are obtained. Massively parallel computer designs are now at a teraflop level. Pressure to increase computer speeds continues from industrial, defense, environmental, medical, and academic demands.

Massively parallel computer designs currently couple 10^3 to 10^4 individual processors with their own individual memories (Pool, 1992). In a large class of designs, each processor is only interconnected with its nearest neighbors. Current major efforts are devoted to solving the problem of communications among processors and the problem of programming for their efficient use. Modeling of some, but not all, complex systems lends itself to massively parallel computing. Modern supercomputers, such as those produced by Cray, excel because of their sequential power, even though some can possess up to 16 parallel processors. A sequential problem run on the current generation of massively parallel computers will go slowly, because their individual processors are slow by supercomputer standards. There are diverse concepts of parallel processors. They in turn can be modeled in broad categories such as nonshared memory, shared memory, and Boolean circuits (Lakshmivarahan and Dhall, 1990). It is useful for mathematical modelers whose work challenges modern computer capabilities to become familiar with computer architecture at this level. Consequently, modelers collaborate with computer manufacturer in-house scientists (Butler *et al.*, 1993). Highly connected parallel computers being built show promise for problems that are not essentially sequential and that do contain synchronous parts.

2.2 Software

The past 30 years have seen the development of numerous higher-level programming languages and the increased use of software systems. FORTRAN, long a standard bearer for numerical programming, has progressed to FORTRAN 90 with parallel processing capabilities. Many other supported programming languages are available. Programming languages are accompanied by increasingly sophisticated debugging and structuring software.

Commercial software systems exist for practically every purpose and are heavily advertised in computer journals. The journal *Computers in Physics* also has sections on research and education, visualization, techniques and technology, and refereed articles. Technical emphasis in universities is on learning to use a progression of specialized software systems, often starting with spreadsheets and including computer algebra. MATLAB is a commercial numerical software system for interactive use. Accessible graphics, facile handling of matrix problems, some limited symbol manipulation, and continual broadening of scope are features of MATLAB. For numerical work, Numerical Recipes and NAG (Numerical Algorithms Group) are well-known commercial groups that also continue to diversify. Several commercial software systems exist for neural networks. In the public domain are many specialized software packages such as FERRET, used by geophysicists to do interactive processing of large data sets.

Historically, computer algebra has been developed for analytic problems of celestial mechanics, general relativity, and quantum mechanics, saving months to years over pencil and paper. The oldest well-known symbolic software system, REDUCE, is open and accessible to independent developers. MATHEMATICA has strong advocates, and its graphics have been extolled in recent reviews. MAPLE continues its development through an academic consortium, and like MATHEMATICA has strong graphics, a users' group, and a newsletter. Many find the academic commitment to MAPLE, through Ph.D. dissertations and faculty interest at the University of Waterloo and Eidgenössische Technische Hochschule, Zurich, to be appealing. Users can affect work on MAPLE through direct electronic mail correspondence with faculty. These symbolic systems and others, including AXIOM, DERIVE, and MACSYMA, have recently been discussed (Fitch, 1993).

Computer algebra is playing an increasing role in the study of ordinary and partial differential equations (Tournier, 1991). This is because symmetries (characterized through Lie algebras and Lie groups) and analytic transformations among classes of differential equations (such as Bäcklund and auto-Bäcklund transforms) are sometimes far more easily explored with computer algebra than by hand.

Another modeling enhancement is telecommunications. Electronic mail and networking, remote and local, are currently estimated to serve more than 10^7 individual workers. For many researchers, telecommunications provides ready access to remote installations offering vastly more powerful machines when needed, with more versatile and more appropriate software. Collaborations of groups spread across the globe can be carried out on just one computer, in some cases with almost instant response to comments and queries.

3. SELECTED MODELS AND MODELING TOOLS

3.1 Fractals

A "fractal" is an infinite set of points described in terms of a noninteger scaling parameter. This common thread encompasses an increasing variety. One broad class, *finite-size fractals*, was first analyzed by Mandelbrot, who originated the modern study of fractals, coining the word fractal (Mandelbrot, 1983). Some finite-size fractals are described in terms of a self-affine scaling behavior: if one looks at them through a magnifier, the images will have clear similarities to the originals, although there may be distortions in different directions. If there are no distortions at all, these fractals are called self-similar. The Hausdorff dimension is the nonintegral scaling parameter for many finite-size fractals. *Fat fractals*, an exception, have integral Hausdorff dimension, though they still possess a nonintegral scaling parameter (Vicsek, 1992). *Growth fractals* form another major class of fractals. They are not self-affine except in some statistical sense. Prototypes are plane dendritic growths (Vicsek, 1992). A union of fractals with differing scaling behaviors is called a *multifractal*. It is easy to construct fractals (including multifractals), using either deterministic or stochastic rules. It is harder to identify naturally occurring fractal structures.

Fractals are ubiquitous in spatial structures and are even associated with modern art. A chaotic system always possesses a fractal orbit. For example, fractal structures have been observed in preturbulent flow (Brandstater and Swinney, 1987). Fully developed turbulence has been shown to have a well-defined multifractal spatial structure (Sreenivasan, 1991), which is not yet fully understood. The detection and characterization of fractals anywhere in nature is an important step in mathematical modeling. Models of fractal growth are being studied intensively (Vicsek, 1992). In Secs. 3.2.1, 3.5.1, 4.4, and 5.3, fractal structure is intrinsic to the discussions. It is therefore relevant to understand how, in a simpler context, noninteger scaling is inferred.

Here we introduce a fractal dimension D which is a standard approximation to the Hausdorff dimension. Dimension D is defined so that, intuitively, D is an integer for nonfractal sets. Consider first a line segment of length L. The number $N(l)$ of small segments of length l needed to cover the length L is Ll^{-1}. This is true for progressively smaller l. The negative of the exponent of l is the scaling parameter, D; the fractal dimension D is unity for this nonfractal set. Now consider a square of side L. The number $N(l)$ of small squares of side l needed to cover the square of size L^2 is L^2l^{-2}. Then again, $N(l) \sim l^{-D}$, and the scaling parameter D is the integer 2 for this nonfractal set. This relation can also be written

$$D = -\lim_{l\to 0} \ln N(l)/\ln(l). \tag{1}$$

We now construct a fractal, Smale's horseshoe map, perhaps the simplest exemplar of a chaotic dynamical system (Smale, 1967). We define an infinite sequence of iterations by first drawing a rectangle of width W and height H. Each iteration occurs in two steps. The first step, stretching, is to draw another rectangle of width $2W$ and height reduced by a factor $1/2\delta$ where δ is some fixed number greater than 1—say, 2. The last step of the iteration, folding, is to bend the rectangle over into the shape of a horseshoe of width W again, whose sides do not touch. The new area inside the boundary of the stretched and folded rectangle is, of course, less than the old. If we draw a vertical line through the new shape, its intersection with the inner area is in two pieces, each of height $H(1/2\delta) = H/4$. When the stretching and folding process is iterated exactly the same way k times, a layered structure develops whose inner area makes $2k$ intersections with a vertical line, each of height $H(1/2\delta)^k$. We now use Eq. (1) to obtain $D = (\ln 2)/(\ln 2\delta) = 0.5$. The limiting internal area is zero. The fractal dimension of the whole layered figure is the sum of the vertical and horizontal fractal dimensions, or $1 + (\ln 2)/(\ln 2\delta) = 1.5$.

Shown in Fig. 1 are the first few iterations that in the limit $k \to \infty$ define two other fractals. They are the Cantor middle-thirds set in (a) finite-size and (b) growth versions. We compute the fractal dimension in (a) as before. Case (a) corresponds to a chaotic orbit of the logistic map (see Sec. 3.5.1). After k iterations, there are 2^k pieces, each of length 3^{-k}, leading to a value $D = (\ln 2/\ln 3) = 0.6309 \ldots$. Figure 1(a) is converted to a fat fractal by removing the middle $(1/3^k)$th piece rather than the middle third. The fractal dimension D for this fat fractal equals unity.

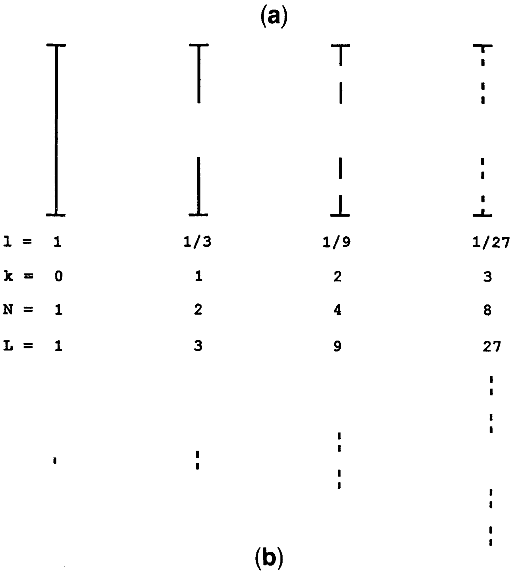

FIG. 1. Construction of a Cantor middle-thirds fractal, for two cases: (**a**) finite-size and (**b**) growth fractals. Fractals are defined iteratively in the limit where $k \to \infty$. In both cases, the fractal dimension D equals ln(2)/ln(3) (see text for details). Figs. 1 and 3 were quickly drawn using "paint" facility on a DEC-5000 workstation.

For growth fractals [case (b)], D is calculated differently. Consider a growth fractal of maximum linear extent L. Here, there is some minimum number $N(L)$ of covering segments of *fixed* edge length l. As $k \to \infty$, we have $L \to \infty$, and D is defined through the relation $N(L) \sim L^D$. The growth rule is iterated three times in Fig. 1(b). The result for $k \to \infty$ is the same as before, $D = (\ln 2)/(\ln 3)$. The union of all iterations in Fig. 1(b) also has the same fractal dimension. In nature, the infinite limit never occurs, but scale invariance can be found where L or l varies over many powers of ten (Mandelbrot, 1983).

It is simple to generalize the conceptual process for finding D to a higher-dimensional space, where covering balls (or hypercubes) of volume $V \propto l^d$ are used, with d representing the usual dimension of the underlying space, frequently as large as 10. However, when d exceeds 2, a method of computing D based upon its definition, known as a box-counting method, must be replaced by a more computation-efficient method (Schuster, 1988). Finally, multifractals are observed by using scaling parameters sensitive to the density distribution within a fractal. The examples given here are not multifractal because the distribution of points within the fractals is uniform. Detecting nonuniformity can be accomplished through introduction of generalized dimensions, D_q, defined in terms of the qth moments of this density distribution. The values of q for which D_q is measured can be $q = 0, \pm 1, \pm 2, \ldots, \pm \infty$, and D_0 is identified with the previously defined D. For multifractals, D_q decreases rather than remaining constant as q increases. The art of measuring multifractals is discussed by Sreenivasan (1991). Major efforts to refine the experimental determination of all these nonintegral scaling parameters continue (Abraham *et al.*, 1990).

3.2 Nonlinear Dynamics

3.2.1 Deterministic Models Dynamics is the study of how systems change in time. If time changes in discrete intervals, the evolution is described as an iterative map, or simply, a map. A map can be written

$$\mathbf{X}_{M+1} = \mathbf{F}(\mathbf{X}_M,\mu). \tag{2}$$

Here, $\mathbf{X}$ can lie in an N-dimensional vector space, $\mathbf{F}(\mathbf{X})$ is a vector-valued function in that space, and μ is a set of parameters. The space is called the *phase space*. If the time changes continuously, the evolution is referred to as a flow. The dynamical equations for a flow can be written

$$\frac{d}{dt}\mathbf{X} = \mathbf{F}(\mathbf{X},\mu). \tag{3}$$

Here, $\mathbf{X}$, $\mathbf{F}(\mathbf{X},\mu)$, and μ are defined as before, except that now $\mathbf{X}$ depends upon continuous time. Solutions of both kinds of systems are referred to as orbits. Flows are numerically approximated as maps. Hence, the connection between these two versions of dynamics is important.

Nonlinear dynamics is a rich field for modelers. In controlled experiments, a change of control parameters sometimes leads to topologically distinguishable states. The nonlinear dynamics theory of such transitions, called bifurcations, is well developed (Schuster, 1988). Bifurcations in lasers can lead to chaotic states, and circuits have been designed that bifurcate into chaos.

3.2.2 Stochastic Models What have just been defined are deterministic dynamical systems. Other systems, referred to as stochastic, have functions describing time evolution, such as $\mathbf{F}$, that are random variables, wholly or partially. A class of stochastic dynamical systems of considerable interest (Abraham *et al.*, 1990) are the kth order Markov processes, defined in the discrete case by

$$P(\mathbf{X}_L|\mathbf{X}_1,\ldots,\mathbf{X}_{L-1}) = P(\mathbf{X}_L|\mathbf{X}_{L-k},\ldots,\mathbf{X}_{L-1}), \tag{4}$$

representing the conditional probability of observing $\mathbf{X}_L$ as the Lth state in the sequence $\mathbf{X}_1, \ldots, \mathbf{X}_L$. When $k = 1$, this is simply called a Markov process. Stochastic noise is usually mixed in with chaos.

Analysis of chaotic data attempts to distinguish between stochastic noise and true chaotic variability. Stochastic resonance also can occur, in which noise plays an essential role (Moss, 1991). In stochastic resonance, transitions occur among different stable states of a deterministic nonlinear dynamical system driven by a periodic function plus noise. Stochastic resonance has been observed in lasers, and provides one conceptual model for climatic change. Nonlinear dynamics—deterministic, stochastic, or both together—is the framework for modeling of weather and climate.

3.2.3 Formal and Experimental Dynamics The literature on nonlinear dynamics is blossoming. In describing this work, we use the word "formal" to denote an analytic mathematical approach. Formal work is sometimes concerned with questions pertaining to effects of finite numerical precision; however, any results reached in a formal approach are characterized by a logical precision that meets the standards of pure mathematics. We employ the broadened meaning of the word "experimental" to include experiments with computer-generated data. Monographs tend to contain a mixture of formal and experimental work (Schuster, 1988). Lorenz's computation of chaotic behavior in 1963, discussed below, is a premier example of experimental dynamical theory. Incisive formal results buttress the field nonlinear dynamics. They include those of Sharkovsky in 1964, discussed here in Sec. 3.5.1, and Gardner *et al.* in 1967, discussed here in Sec. 5.4. Also, in 1967 Smale investigated classes of smooth maps and developed an important exemplar, the horseshoe map (Tufillaro *et al.*, 1992). Formal, too, was the 1971 proposal by Ruelle and Takens of a new route to chaotic turbulence *via* a few bifurcations, seen experimentally in 1975 (Lichtenberg and Lieberman, 1992).

The formal and the experimental are combined in shadowing theorems. A shadowing theorem seeks to answer an important question in chaotic dynamics. All chaotic systems show SIC, seen in divergence of nearby orbits and loss of memory of initial conditions when finite precision calculations are made. Then can any finite precision calculation approximate an asymptotic orbit? Formal computer-assisted answers in the affirmative have been found in special cases—for example, for the Hénon and the Ikeda maps—through the introduction of machine-independent procedures, both in defining truncation and in finding overlapping regions, i.e., a shadow, of a large succession of mapped neighborhoods (Hammel *et al.*, 1988).

3.2.4 Time-Delay Embedding Time-delay embedding is a technique for reconstructing an orbit in a nonlinear dynamical system from a time series of a single scalar observable $y(t)$. If d is the topological dimension of the manifold in which the orbit lies, there is now considerable evidence that the vector whose components are time-delayed values of $y(t)$, namely,

$$\mathbf{X}(t) = (y(t), y(t + \tau), y(t + 2\tau), \ldots, y(t + \{d_E - 1\}\tau))^{\mathrm{T}} \tag{5}$$

(where T is matrix transposition), often adequately mimics the topological and metric behavior of that orbit. If so, the orbit is said to be embedded in the reconstructed finite-dimensional space. This is an indispensable insight: Without it, one would be required to identify and measure all the observables relevant to that orbit for an adequate description. Time-delay embedding was articulated ca 1980 independently by physicists and a mathematician (Tufillaro *et al.*, 1992). In the absence of an infinite set of noise-free data, it becomes important to optimize the delay parameter τ and to find a practical embedding dimension d_E smaller than the value $2d + 1$, which has been conditionally proved to be sufficient. Information theory has been of practical value in finding τ (Fraser and Swinney, 1986), while numerical experiments show that a good d_E is often smaller than $2d + 1$. Time-delay embedding is a major experimental tool of chaos research.

3.2.5 Cellular Automata Cellular automaton (CA) models are simplified dynamical systems having discrete spatial and temporal degrees of freedom (Wolfram, 1986). The spatially distinct points, called "cells," can be interpreted to represent arrays filling different dimensions, according to how they are made to interact with their defined "neighborhoods." The CA models have seen use in describing excitable media, both biological and chemical, lattice-gas automata and turbulence, and deterministic evolution of Ising lattices. In three-dimensional excitable media, CA models of propagating scroll-shaped waves are two orders of magnitude faster than partial differential equation (PDE) models (Gerhardt *et al.*, 1991); however, the validation of CA models is relatively more problematic. Although straightforward computationally, allowing rapid checks of various hypotheses, CAs need development of their formal foundations. For example, formal criteria for determining conserved quantities are not generally known (Hattori and Takesue, 1991).

3.3 Rational Functions

A rational function is any ratio of two polynomials, written $P_L(x)/Q_M(x)$, where the degrees of the polynomials are L and M. Taking a function $f(x)$ to be rational is a potent working assumption. Use can often be made of the meromorphic extension of $f(x)$ to the complex plane. Differing from a smoothing technique like a spline, the rational function *Ansatz* is a global statement about the nature of the analyticity as well as the asymptotic nature of $f(x)$. Rational approximants are successfully applied in solving integral and differential equations, but still, much about them remains to be understood (McInnes and Marshall, 1993).

High-degree rational functions sometimes require ultrahigh-precision arithmetic to distinguish between noncanceling poles and zeros. Hence, restriction to a rational function is seminumerical modeling in that analytic structure is maintained, while ultimately numerical evaluation is needed. Sometimes the function to be approximated is entire, having no singularities in the finite complex plane, but, as Padé discovered in his doctoral dissertation in 1892, even the exponential can be represented as a rational function in some domains. Baker rediscovered the power of Padé approximants in 1960 while solving for critical exponents in the theory of phase transitions. Recognition of the power of this tool continues to grow. Even functions with cuts in the complex plane are sometimes representable as rational functions, with arrays of poles and zeros tracing the cuts.

As reintroduced by Baker (Baker and Graves-Morris, 1981), Padé approximants are easily obtained by solving linear equations or linear recursion relations that lead to numerical values of the $\{a_i\}$ and the $\{b_i\}$ in the expression

$$f(z) = \frac{a_0 + a_1 z + \cdots + a_L z^L}{b_0 + b_1 z + \cdots + b_M z^M} + O(z^{L+M+1}) \tag{6}$$

when the Taylor series expansion

$$f(z) = c_0 + c_1 z + c_2 z^2 + \cdots \tag{7}$$

is known. The Baker convention is to make the additional constraint $b_0 = 1$.

One extension of rational-function approximants useful for data modeling and inverse models is the "statistical Padé approximant" (SPA). The SPA has the same form as the PA but is obtained by fitting numerical data by a linear iterative method (Yidana and Hartt, 1988) that frequently converges well even in the presence of significant noise. The SPA has been used for fitting nuclear-scattering data, for inverse-scattering problems, for direct solutions of the Schrödinger equation, and for finding analytic solutions to Riemann boundary-value problems.

Rational functions are fully described by enumeration of their zeros and poles and a leading numerical coefficient and are highly accessible to symbolic computer software and iteration dynamics studies, including chaos. Rational-function iterations form nonlinear dynamical systems through which Fatou, Julia, and Mandelbrot sets are definable (Mandelbrot, 1983) and which occur in fractal image compression methods. Finally, rational functions play a central role in linear circuit theory and in signal processing. Because of their computability and usefulness, they have an interdisciplinary following and are the subject of frequent conferences.

3.4 Monte Carlo Methods

The Monte Carlo (MC) method (*q.v.*) employs algorithms that use random numbers for making convergent statistical estimates. These estimates can be solutions of deterministic problems, evaluations of high-dimensional integrals, or averages over random processes. This method vastly increases the range of problems that can be dealt with theoretically. Although analytic approaches are continually being proposed for complex systems, there are many problems for which no known analytic procedure yet works. Massively parallel computer architecture is useful in MC calculations.

Historically, the MC approach to a modeling problem has had three aspects:

1. generation of random numbers;
2. development of algorithms to use random numbers for solving deterministic problems; and
3. development of models for the use of the other two aspects.

Currently, there is activity in all three. Even the first has its hazards. All methods using computers to generate number sequences involve finite algorithms, which must fail some test for randomness, if for no other reason than that computers are finite-state systems. Such computer-generated sequences are therefore called "pseudorandom numbers." All finite tests for randomness are incomplete (Knuth, 1981), and even the concept of randomness is debated (Lichtenberg and Lieberman, 1992). Linear congruence methods, which perform multiplications modulo a properly chosen large number, receive heavy use in the constructions of uniformly distributed sequences in [0,1]. Recent calculations solving problems with known solutions show some methods to be better than others (Ferrenberg *et al.*, 1992). It behooves MC users to validate their methods separately for all algorithms.

For MC modeling, algorithms are needed that generate a pseudorandom number sequence according to an arbitrary probability distribution $p(x)$, given a pseudorandom sequence $\{y_i\}$ distributed in [0,1]. This is sometimes done analytically in one dimension

(Gould and Tobochnik, 1988), by direct inversion of $P(x) = \int_{-\infty}^{x} dxp(x)$, since $dP(x) = dy = p(x)dx$. Direct inversion generalizes to a small dimension m and a rapid MC approximation of $\int d^m xF(x_1, \ldots, x_m)$ only if $F(x_1, \ldots, x_m)$ resembles a tractable probability density function $p(x_1, \ldots, x_m)$. Efficient numerical inversion methods are necessary for higher dimensions.

The powerful Metropolis-Hastings (MH) procedure (Creutz, 1992) was first used 40 years ago. The MH introduces an artificial first-order Markov process and is known as a dynamic MC method. Like direct inversion, MH produces "importance sampling" through which a nonuniform distribution emphasizes regions where the integrand is large (Binder, 1987). It converges by selecting a long string of steps through successive configurations of the random variables, in accordance with general constraints such as the detailed-balance condition. The steps are highly correlated. If an uncorrelated sequence of N configurations is used, as in the simplest MC calculations, estimated errors diminish notoriously slowly, by $N^{-1/2}$. As the sequence becomes more correlated, convergence is slower, because smaller regions tend to be sampled by a fixed number of configurations. The evaluation of when convergence has occurred must consider these correlations.

Models are often formulated from the start for use of the MC method. Subjects of separate monographs are the MC approach to boundary value problems (Sabelfeld, 1991), atmospheric radiative transfer (Marchuk *et al.*, 1980), small quantum mechanical systems (Kalos and Whitlock, 1986), quantum fields (Creutz, 1992), and statistical physics (Binder, 1987). Each subfield has its own set of computational algorithms. We illustrate current activities with two recent accomplishments.

The quantum mechanical ground-state energy E_g of a few-body system satisfies the inequality $E_g \leq E_t$, where E_t is the expectation value of the energy operator using a trial wave function $\boldsymbol{\Psi}_t(\mathbf{X},\mu)$. The function $\boldsymbol{\Psi}_t(\mathbf{X},\mu)$ is an approximate solution of the time-independent Schrödinger equation. For an n-particle system, $\mathbf{X}$ has up to $4n$ space and spin components, and μ is a parameter vector whose value determines an analytic $\boldsymbol{\Psi}_t$. For electrons in a molecule, $\boldsymbol{\Psi}_t$ can contain 10^2 parameters. The parameters are not directly associated with the electronic degrees of freedom. Rather, they give the relative weights of analytically defined Slater determinants and the shapes of state-dependent correlation functions. Evaluation of E_t requires importance sampling of large numbers of values of $\mathbf{X}$, referred to as configurations. This whole process must be repeated many times with a minimization algorithm. Even supercomputer power can be too slow to reach convergence. However, the variance of the local energy mean over all configurations of an exact solution is zero. In practice, that variance is adequately approximated using far fewer configurations (by a factor of 10^3) than needed for energy evaluation. Little accuracy is lost in using the same configurations for successive steps of the parameter vector μ (Umrigar, 1989), and a positive benefit accrues through noise reduction among successive values of the variance, which must be compared. Immense time savings occur from minimizing the variance with infrequent calculations of the energy. As a result, molecular-structure problems can sometimes be done on workstations.

Recent MC calculations of eight particle masses were good to 6% (Butler *et al.*, 1993). They took one year on a special-purpose massively parallel computer, the IBM GF11 computer, where integrals on a space-time lattice were evaluated. Successively larger lattices up to $30 \times 32^2 \times 40$ in size approximated the limit of zero lattice spacing and infinite volume, requiring 10^{17} floating point operations. This result is a breakthrough for computer modeling, for the QCD (quantum chromodynamics) particle model, and for parallel processing. A similar program, the QCD Teraflops Project, is underway. It will use a $128^3 \times 256$ lattice on a platform derived from the Thinking Machines Corporation's CM5 massively parallel, highly connected computer.

3.5 Numerical Modeling

A numerical model is a set of computational algorithms associated with a mathematical model. In order to be complete, all but the simplest mathematical models must contain a numerical model. Just a few examples illustrate this need. The logistic equation,

$$x_{N+1} = 4\lambda x_N(1 - x_N), \tag{8}$$

where x is real and lies in the interval [0,1], provides the prime textbook example of the period-doubling approach to chaos, as λ increases from 0. Feigenbaum's discovery in 1978 of universality in the bifurcation sequence of this map has been a cornerstone in the development of chaos theory (Schuster, 1988). Yet this iterative equation is also an approximate numerical scheme to solve the differential equation, $dx/dt = ax - bx^2$, with a and b suitably chosen, which has only analytic, nonchaotic solutions. However, the logistic map also approximates a delay-differential equation, which can have chaotic solutions. Delay-difference and delay-differential equations are frequently used in biological modeling. Clearly, studies of their global behavior are important.

Discretization is the approximation of a continuum by points with some minimum spacing, as can be carried out formally to obtain the logistic equation. Alternatively, *spectral models* maintain a partial continuum description in introducing expansions of continuous functions of some space-time degrees of freedom. The expansions may use infinite complete orthonormal sets. The approximation in using them lies in keeping only a finite number of terms. As a byproduct, such truncation reduces the spectral width of the observables, a form of filtering—important both in spectral and discrete methods (Ghil and Malanotte-Rizzoli, 1991; O'Brien, 1986).

Smoothing is a process by which numerical data are either replaced or interpolated by means of other data that display a minimum of short-range fluctuations. Smoothing is standard experimental practice as a preliminary to applying the finite Fourier transform (see Sec. 4.3). This is sometimes accomplished by averaging over a finite subset of nearby data. The data can be results of experimental measurements or results of computations. The smoothing assumptions are that

1. local data fluctuations represent an imperfection best eliminated; or
2. analyticity requirements need to be enforced (as opposed to requiring a specific functional form).

In chaotic systems with sparse data, smoothing is sometimes unwarranted and leads to the inference of spurious low fractal dimensions (Grassberger, 1986). Therefore, the issue as to whether to smooth needs to be settled first. From a systems point of view, smoothing is a special kind of filtering (see Sec. 3.7).

If smoothing is warranted, *splines* represent a computer intensive, yet conceptually simple, approach. The spline process introduces a basis set of functions in one or more dimensions that possess smoothness often defined in terms of existence of higher-order derivatives. The data are partitioned, and the spline basis is fitted within each separate partition so as to satisfy interpartition smoothness criteria (Schumaker, 1980). A cubic spline, in which the cubic polynomial is the basis function, has long been a common option available to modelers (Press *et al.*, 1987). Yet cubic interpolation continues to be refined (Huynh, 1993). Spline smoothing also provides a foundation for nonparametric regression analysis, which in some contexts is an alternative to autoregressive/moving average (ARMA) (Eubank, 1988; also see Sec. 4.3).

3.5.1 Chaotic Systems An approach sometimes used in numerical modeling of chaotic systems is to create a kind of coarse-graining of the possible states associated with a mathematical model. Each coarse-grained state is assigned a different letter, and the totality of letters is called the alphabet. This approach is referred to as symbolic dynamics. There are other valid ways than that mentioned to create an alphabet, such as, for example, a symbolic classification of the unstable periodic orbits. Here, the impact of the Sharkovsky theorem (Sharkovsky, 1964) becomes evident, as it introduces symbolic dynamics to make fundamental topological statements suggesting a certain set of preconditions for chaos (coexistence of infinite numbers of kinds of periodic solutions). It holds for *all* one-dimensional continuous, unimodal maps on the interval [0,1]. The continuous function $f(x)$ which maps the interval [0,1] into itself is termed unimodal if $f(0) = f(1) = 0$ and $f(x)$ possesses exactly one critical point. Sharkovsky's theorem does not necessarily imply a positive Liapunov exponent (Schuster, 1988). It establishes the following ordering of the natural numbers:

$$\begin{aligned}&3 \succ 5 \succ 7 \succ \cdots \succ 2\cdot 3 \succ 2\cdot 5 \cdots 2^2\cdot 3\\ &\quad \succ 2^2\cdot 5 \succ \cdots \succ 2^3\cdot 3 \succ 2^3\cdot 5\\ &\quad \succ \cdots\cdots \succ 2^3 \succ 2^2 \succ 2 \succ 1,\end{aligned}$$

where the symbol $>$ is used for "precedes." If $f(x)$ has a cycle of period p then there exist points x' in $[0,1]$ that are members of cycles of period q, for all q that satisfy $p > q$. In particular, if one numerically determines the presence of a cycle of period three, then the map $f(x)$ contains cycles of all integral periods. The search goes on for similar insights using symbolic dynamics in higher dimensions (Tufillaro *et al.*, 1992).

The Lorenz equations, which ushered in the modern era of chaos theory (Lorenz, 1963), are themselves a drastically truncated spectral model of a dissipative system undergoing Rayleigh-Bénard convection. Rayleigh-Bénard convection is essentially the motion of a closed fluid system in response to heating from the bottom. It is known experimentally to exhibit chaotic behavior in regions of its parameter space. Yet the Lorenz equations do not properly describe the real physical system when it is in its chaotic state: this is a case where a model takes on a life of its own, independently of its original motivation. In particular, identification of chaotic behavior as characterized by one-dimensional maps, both through computation (Lorenz, 1963) and experiment in similar systems (Brandstater and Swinney, 1987), reveals how low dimensionality of the fractal attractor provides a simplifying insight into complex phenomena. The Lorenz equations, a set of three coupled first-order differential equations, require further numerical modeling (discretization). There is continued interest in the dependence of chaotic solutions upon numerical algorithms.

The local evolution of chaotic systems is extremely sensitive to discretization errors; however, the global properties, such as the Liapunov exponents $\{\lambda_i\}$, the Kolmogoroff-Sinai entropy K_{KS}, and the generalized fractal dimensions $\{D_q\}$, are not (Schuster, 1988; see also Secs. 3.1 and 4.5). All chaotic states have fractal orbits. Therefore, determining that at least one measure of dimensionality of the orbit, such as the Hausdorff dimension, is nonintegral and hence fractal is suggestive that chaos is present. For the logistic map, a wide range of values of λ gives chaos (Schuster, 1988). The asymptotic chaotic orbit for the value $\lambda = \frac{9}{8}$ is just the Cantor middle-thirds set discussed in Sec. 3.1.

The maximum Liapunov exponent λ_0 is the real litmus test for chaos, because only the inequality $\lambda_0 > 0$ yields exponential divergence of nearby orbital points, assuring SIC. Smale's horseshoe map discussed in Sec. 3.1 has two Liapunov exponents. They are both computed from a relation derivable from the general definition (Schuster, 1988) in this special case: $\Delta x' = \exp(\lambda \Delta t)\Delta x$. We take $\Delta t = 1$ for one iteration step, and $\Delta x'$ and Δx are the new and the old distances between nearby points after one iteration measured

1. along the direction of stretching in the asymptotic chaotic orbit and
2. perpendicular to that direction.

The Liapunov exponents are $\lambda_0 = \ln 2$ and $\lambda_1 = \ln(1/2\delta)$. *Dissipative chaos* corresponds to the present case in which the sum of Liapunov exponents is negative, which implies that an elementary volume element in the space of the dynamical system evolves to zero. *Conservative chaos*, in which the sum of all the Liapunov exponents is zero, denotes systems in which the size of a volume element evolves unchanged. Conservative chaos occurs in Hamiltonian systems and is heavily investigated for its relevance to the foundations of classical and quantum mechanics (Ford and Mantica, 1992).

The 1980s saw successful numerical modeling, much of it using time-delay embedding (Abraham *et al.*, 1990; Schuster, 1988). Improvement is sought in the study of experimental data. A typical exercise is to generate small, noisy data sets from known chaotic attractors and to try to recover their chaotic properties. There are procedures to identify spurious Liapunov exponents generated by time-delay embedding (Parlitz, 1992).

3.5.2 Finite Elements Finite elements is a form of numerical analysis devoted to continuum systems such as solid structures and electromagnetic fields (Cook *et al.*, 1989), in which continua are discretized into highly symmetric elements. For example, these might be triangles, tetrahedra, or hexahedra. Algorithms are then developed for interpolating between nodes. Finite elements is also applied in simple quantum mechanical systems of low symmetry for which analytic solutions do not exist (Shertzer, 1989). Large generalized eigenvalue problems sometimes result, with variational bounds.

3.5.3 General Circulation Models The GCM's coupling the atmosphere, ocean, land, biosphere, and cryosphere on global scales are faced with Herculean tasks. This is because space-time discretization must be so coarse-grained. Consider the global atmosphere. Even 10^6 cells, counting 20 vertical levels, make a sparse approximation to a continuum and corresponds to somewhere in a mid-meso-scale size, far too large to simulate the physics of clouds, precipitation, turbulence, or local storms or eddies. Therefore, the dynamical and thermodynamical laws must be averaged. If the horizontal and vertical scales are reduced by a factor of 100 and if the order of 10 quantities are calculated at each time, a microscale description of the atmosphere could result in which a teraflop computer would require 10 s to simulate one time step. Within such a discretization, eddies might be resolved, but clouds, precipitation, and turbulent dissipation still need to be parametrized. For oceanic phenomena, characteristic time scales are orders of magnitude larger and length scales orders of magnitude smaller.

Decrease in GCM computation times has been achieved with hybrid spectral and discretization methods (Peixoto and Oort, 1992). The horizontal fluxes are expressed in terms of a truncated spherical-harmonics expansion; the vertical and the time coordinates are discretized. Spherical harmonics make a complete set on a sphere. Figure 2 shows a perturbation of a sphere by a higher-order spherical harmonic. It was produced by a few simple MAPLE graphics commands. When discretizing, modelers work within required upper bounds to products of $\Delta t \Delta x$, the space and time intervals, and employ filters to remove fluctuations corresponding to unwanted sound and gravity waves (O'Brien, 1986). Different numerical modeling schemes are regularly compared.

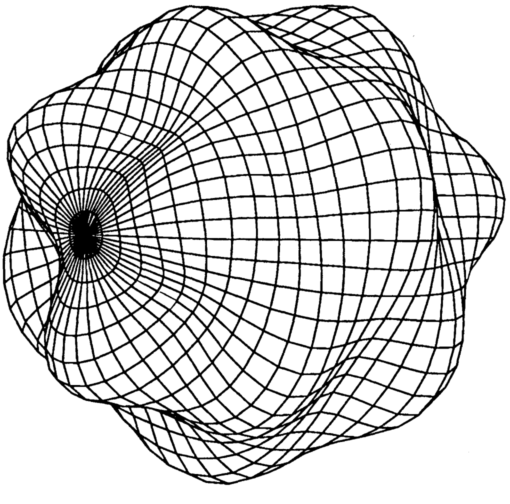

FIG. 2. Three-dimensional plot of a small perturbation of a sphere by the surface spherical harmonic Y_7^3. Requiring a few short commands in MAPLE, this illustrates computer visualization in modeling.

3.5.4 Data-Assimilation Models The goal of several-day weather prediction has been achieved through supercomputer calculations of atmospheric circulation fields on fine grids, combined with updates, i.e., corrections, from measurements. Sophisticated assimilation models have been developed for atmospheric and oceanographic purposes (Ghil and Malanotte-Rizzoli, 1991). Assimilation models are needed because of lack of knowledge of a complete set of initial conditions, inaccuracies in the numerical models, and atmospheric chaos. Atmospheric sampling and modeling are good enough for the National Meteorological Center (NMC) to create ensembles of computer runs that, combined, make possible predictions and estimates of reliability. In contrast, experimental sampling of oceanic temperature, salinity, and velocity fields is poor and will continue to be so in the foreseeable future. Its ill-posed nature makes oceanic circulation the greater challenge. Many techniques are used, including two discussed in this article—singular-value decomposition and Kalman filters (Ghil and Malanotte-Rizzoli, 1991; Carter, 1989).

3.6 Wavelets

A wavelet is an integral transform designed to perform a localized time-frequency analysis. It provides a microscope with which to look at details of a function's spectrum within a small time interval centered about some arbitrary time. Taking its early motivation from geophysical applications, the wavelet has undergone robust development mathematically and in application (Daubechies, 1992). The wavelet transform provides a useful tool to study signals that, like speech, impart information that evolves in time. There is a discrete wavelet transform, and it is important, but we shall concentrate here on salient defining characteristics of the continuous transform. Wavelets and windowed Fourier transforms are sometimes similar. Remarkably, the slight generalization of the windowed Fourier transform in the wavelet transform can make a vast difference.

Consider a one-dimensional Fourier transform:

$$g(\omega) = \int dt\, e^{-i\omega t} f(t). \tag{9}$$

The Fourier transform takes one from the time (t) domain to the frequency domain (ω is the angular frequency). The windowed Fourier transform replaces $f(t)e^{-i\omega t}$ with

$$f_{\text{win}}(t)e^{-i\omega t} = f(t)h(t - s)e^{-i\omega t}, \tag{10}$$

to produce $g_{\text{win}}(\omega)$. Here, h is a simple function, such as a Gaussian, centered around $t - s = 0$, giving a smooth cutoff for large values of its argument. This transformation does not yet provide the microscope—just the localization. In the wavelet transform, the function $he^{-i\omega t}$ is replaced by a function Ψ, referred to as the "mother wavelet," which there is considerable leeway to define:

$$|a|^{-1/2}\Psi\left(\frac{t-b}{a}\right) \tag{11}$$

subject to the condition

$$\int dt \Psi(t) = 0. \tag{12}$$

For acceptable Ψ, a serves as a microscope parameter and b a localization parameter. As a decreases, the wavelet transform $g_{\text{wav}}(\omega)$ zooms in to a decreasing width of times, centered around b. The function Ψ is reminiscent of, but not equivalent to, the spectral window introduced in the Fourier transform (Priestley, 1988).

The form $\Psi(t) = (1 - t^2)\exp(-t^2/2)$, sometimes called the Mexican hat function, is a typical choice for Ψ. Significant progress has been made in developing orthonormal bases for the mother wavelet. A mother wavelet and associated orthonormal bases of compact support have also seen considerable attention. In analogy with Fourier analysis, the wavelet transform can often be inverted to recover the original function, which can thereby be formally expressed in terms of the wavelet basis (Daubechies, 1992).

Wavelets promote systematic investigations of data. Turbulence is an application of wavelet transforms, because theories of turbulence point to self-similarity, which wavelets elegantly characterize (Vergassola and Frisch, 1991). Wavelets also provide a tool for studying multifractals.

3.7 Systems Analysis

A system is a process or a structure or both combined. In *systems analysis* (SA), the relationships among a system's parts are studied. SA is concerned with explanation and prediction, when applied to natural systems, and achieving a goal, when applied to constructed systems. Natural systems include climate, living organisms, neural networks, a flow into a pile of grains of sand, and much more. Constructed systems include manufacturing processes, servomechanisms, circuit networks, communication devices, and artificial neural networks. Social hierarchies could be considered either natural or constructed. The same concepts from SA can be universally applied. In natural systems, *synergetics* is a SA that seeks extensions of physical principles required for explanation of self-organization. In climate and weather, SA is used for mechanistic modeling and for understanding simulation models. Life processes such as self-replication, evolution, and cell specification are another endeavor of SA. SA is a framework for mechanistic modeling.

Frequently encountered concepts are feedback, control, filter, and optimality. Unfortunately, working definitions vary. The concepts are precise in engineering. In signal processing, a system is any process that transforms signals, a signal being a function of one or more independent variables. Linear systems, feedbacks, filters, and control theory are standard in engineering curricula. Unless formal, SA heavily employs flow diagrams. An overview of applications and modern research, including the theory of nonlinear systems, is in a Kalman *Festschrift* (Antoulas, 1991).

Filter is a name for a dynamical system that performs a task such as removing some unwanted variance, such as noise, in a signal. As viewed in time-series analysis (Priestley, 1988), a linear and time-invariant system which relates input U_t to output V_t

through the equation

$$V_t = \sum_{s=0}^{\infty} a_s U_{t-s} \tag{13}$$

is a filter. Filters can be more generally defined and are not necessarily linear. A Kalman filter, in its simpler form, is a linear finite-state process that optimally estimates the state of a linear dynamical system in the presence of noise and deterministic control input. Optimality is defined as minimization of variance. The Kalman filter becomes a unique, recursive algorithm and can be adapted to smoothing, prediction, or control. Unlike earlier filters, the Kalman filter works for nonstationary time series. The Kalman filter has been extended now for use with nonlinear systems. A textbook example is to estimate the orbit of a satellite (Chui and Chen, 1987).

3.7.1 Control Theory Viewed dynamically, a system is describable in terms of an n-dimensional state vector, $\mathbf{X}(t,\boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ can be an m-dimensional control vector. Much of control theory is a subfield of nonlinear dynamics and can be stated in terms of either maps or flows. When the system is nonlinear, bifurcations can sometimes occur and be relevant. One problem in control theory is to find an optimal control vector $\boldsymbol{\Theta}(t)$ such that some state $\mathbf{Z}(t)$ is reached. Optimality can be defined in terms of minimization of cost or of time taken to reach $\mathbf{Z}(t)$, within defined constraints on $\boldsymbol{\Theta}(t)$.

The concept of a *closed-loop* or *feedback control* arises simply in a linear stochastic system as represented by a vector flow equation

$$\frac{d}{dt}\mathbf{X}(t,\boldsymbol{\Theta}) = F(t)\mathbf{X}(t,\boldsymbol{\Theta}) + G(t)\mathbf{w}(t) + L(t)\boldsymbol{\Theta}(t) \tag{14}$$

where $\mathbf{w}(t)$ is a noise process, and $F(t)$, $G(t)$, and $L(t)$ are gain operators. In particular, $L(t)$ contains the influence of the control vector $\boldsymbol{\Theta}(t)$ upon the state vector $\mathbf{X}(t,\boldsymbol{\Theta})$. If $\boldsymbol{\Theta}(t)$ is a function of time alone, then this is *open-loop control*, while if $\boldsymbol{\Theta}(t)$ is explicitly a function of $\mathbf{X}(t,\boldsymbol{\Theta})$ also, then this is called a closed-loop or feedback control.

Stringent necessary conditions for attainability, reachability within a fixed interval, and controllability are more easily obtained for linear than nonlinear systems. Such conditions, sometimes for open-loop control, can be useful to mathematical modeling. Consider, for example, the growth of several competing species controlled by a single essential growth-limiting nutrient. There is a well-known principle of competitive exclusion: if the rate of input and washout are constant, then at most one species survives. It turns out that extinctions do not always have to occur. Input and washout rates can be allowed to vary with time. Under rather general conditions, a nonlinear control-theoretic mathematical model establishes necessary and sufficient conditions for admissible controls [i.e., an admissible nutrient concentration $s(t)$] such that all n species can coexist, and it gives insight into producing controlled selective extinctions (Rao and Roxin, 1990).

Atmospheric and oceanographic modeling often focuses on natural feedback loops and can use a SA approach. In unraveling the impact of atmospheric CO_2 doubling, for example, modelers are concerned about many feedback loops; the radiation–cloudiness–temperature feedback, for example, has only recently been understood to be globally stabilizing (Levi *et al.*, 1992). A more explicit SA model was for large-scale climate oscillations in the global glacier-ocean-atmosphere system (Sergin, 1980). Schlesinger (1989) seeks to unify the climate modelers' nomenclature by employing standard electrical engineering definitions of feedback. Figure 3 shows a block diagram for the climate system adapted and redrawn from his work. In it, the "feedback factor" of the climate system is $f = GF$, where G is the gain, F is a measure of the feedback, and the input forcing, caused for example by increase of CO_2, is ΔQ. When no feedback is present, $J = \Delta Q$ and the surface temperature response is just $\Delta T = GJ = G\Delta Q$. When there is feedback, the output J is increased by $\Delta J = F\Delta T$ to become $J = \Delta Q + F\Delta T$. Then

$$\Delta T = \frac{G}{1 - f}\Delta Q. \tag{15}$$

A positive feedback factor destabilizes, while a negative factor diminishes the response (the case $f \geq 1$ is excluded). Numerical simulation

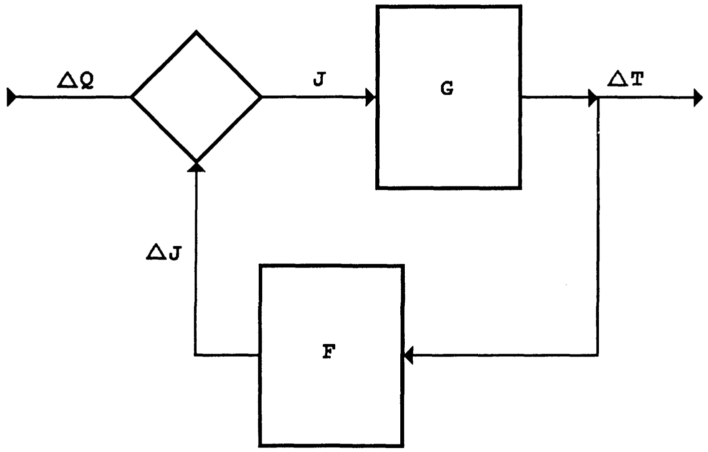

FIG. 3. Block diagram for natural climate feedback control. In the absence of feedback ($F = 0$), an amount of input forcing heat ΔQ produces an output J, which multiplied by the gain G equals the temperature change response ΔT. In the presence of feedback ($F \neq 0$), the output is changed by $\Delta J = F\Delta T$, causing ΔT to be multiplied by $1/(1 - f)$, with a feedback factor $f = GF$.

experiments and ingenious physical experimentation are ingredients in SA modeling like this. Low-dimensional climate models viewed as spatial averages, with few competing feedbacks, have also produced insights (Peixoto and Oort, 1992).

3.7.2 Self-Organization Self-organization is modeled in many ways. These include specializations and extensions of known physical models such as thermodynamics and dynamical systems theory, as well as control theory. Neural networks have been employed. Further, a relatively new field, synergetics, pioneered by H. Haken, is dedicated to the study of structures arising within far-from-equilibrium systems that manifest self-organization (Yates, 1987). Synergetics has enunciated a slaving principle, supported by various examples both from animate and inanimate phenomena. Essentially, slaving occurs when long-lasting quantities serve as order parameters in transient states and dominate over short-lasting quantities. For example, when pump power in a laser goes from weak to strong, the laser's fluctuating dipoles become coherently enslaved by the external field. At just enough pumping, a bifurcation occurs, symmetry is broken, and the motion of each dipole is enslaved by the external electric field. In such a manner, it is posited that enslaving creates a hierarchic structure that reduces the number of degrees of freedom of complex systems. Synergetics attempts to describe the development and operation of living structures in this way.

Fractal structures abound in large systems. The suggestion of a fractal structure is an inverse power-law (or $1/f$) behavior of a temporal or spatial power spectrum $S(f)$, taken to mean $S(f) \propto f^{-\alpha}$, with $\alpha > 0$ (West and Shlesinger, 1990). Some models predict $1/f$ phenomena associated with critical states toward which large systems naturally evolve. Such a self-organized criticality stems from work initiated by Bak and his collaborators (Bak *et al.*, 1987). Using CAs to simulate grain-by-grain formation of a sandpile, they reached a critical state. In the critical state, temporal self-similarity (yielding flicker noise) arises, and spatial self-similarity is also observed in the comparable probabilities of large and small avalanches. Self-organized critical states are only weakly chaotic, neighboring points showing power-law rather than exponential divergence. This modeling has stimulated theoretical and experimental work; a notable success is a power-law prediction for energies and spatial distributions of earthquakes (Lichtenberg and Lieberman, 1992).

4. INFORMATION PROCESSING MODELS

4.1 Principal-Component Analysis

Principal-component analysis (PCA) is a method of analyzing the variance within a data stream for the purpose of finding a reduced set of variables. Ideally, these variables should be

1. statistically uncorrelated and
2. responsible for most of the variance.

Property **1** is automatically achieved, but property **2** is just a possible outcome. In analogy with Fourier series, an orthonormal basis set is obtained, but with PCA, the basis is empirically defined. After the advent of the digital computer, outstanding successes in meteorology and oceanography (Preisendorfer, 1988) and signal processing have marked the history of this 90-year-old method. The PCA has many applications, sometimes used as a low-pass filter in conjunction with statistical regression. Another major use of PCA is in chaos research, to help separate extraneous noise from chaotic variability.

A standard first step is to construct a symmetric covariance matrix of space-time events. We use bold face for matrices but not their components, superscript T for transpose, and matrix notation for vectors. Consider a time series of p-dimensional vectors $\mathbf{x}_i$ of some observable scalar field, such as temperatures at p different places, measured at n times. Assume that these are all zero-mean measurements, i.e., they are all given relative to their mean values. The sample covariance matrix $\mathbf{S}$ for the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ has as its (j,k)th element

$$\frac{1}{(n-1)} \sum_{i=1}^{n} x_{ij} x_{ik}, \tag{16}$$

where x_{ij} is the jth component of the ith vector. This is a symmetric matrix diagonalizable by a matrix whose columns are its orthonormal eigenvectors. Here we adopt geophysical terminology and refer to an eigenvector $\mathbf{a}_r$ as an empirical orthogonal function (EOF). In statistics literature, it is referred to as a principal component. The matrix $\mathbf{S}$ is non-negative; the eigenvalue l_r for each nontrivial solution is positive and is the variance associated with its eigenvector, $\mathbf{SA}_r = l_r\mathbf{A}_r$. This eigenvalue problem satisfies a variance maximization principle. Hence, the first EOF, having the largest eigenvalue, is the normalized linear combination of the p variables with maximum variance. The second eigenvector is the combination of largest variance such that it is orthogonal to the first, and so on. Each of the n data vectors is expandable in terms of the p EOF's. In geophysical terminology, the time-dependent expansion coefficients of the n data vectors are the principal components (PCs). Criteria for dropping terms associated with small variances, called selection rules, are developed in the context of the applications. In one of the first computer-assisted PCA analyses, E. N. Lorenz at MIT made prediction studies in 1956 of a 500-mbar-height anomaly field over North America, using five years of data obtained at 64 stations. The first nine eigenvectors obtained accounted for 88% of the variance for a one-day prediction (Preisendorfer, 1988).

Singular-value decomposition (SVD) is a different but equivalent formulation of this eigenvalue problem, based on an SVD theorem showing that an arbitrary $n \times p$ matrix is expressible in the form $\mathbf{X} = \mathbf{ULA}^{\mathrm{T}}$. Here, $\mathbf{L}$ is an $r \times r$ diagonal matrix, where r is the rank of $\mathbf{X}$, and $\mathbf{U}$ and $\mathbf{A}$ are matrices with orthonormal columns. In a PCA application, the elements of $\mathbf{L}$ are proportional to $l_r^{1/2}$, where l_r is the rth eigenvalue of $\mathbf{S}$. SVD has an algorithm for finding EOFs.

PCA is also adaptable to data vectors $\{\mathbf{x}_i\}$ having components that represent different kinds of physical quantities, such as pressures and temperatures. A PCA is then usually defined in terms of a correlation matrix, formed by dividing each element x_{ij} of the n data vectors by $\sigma_{jj}^{1/2}$, where σ_{jj} is the variance among the jth components, the variance being taken over the columns in x_{ij} defined by fixed values of j.

4.2 Neural Networks

A *neural network* (*q.v.*) is a system containing a highly interconnected set of elements inspired by characteristics of the brain. In one simple formulation, the elements (called nodes or neurons) fire (i.e., are placed in a binary "up" state of +1 as opposed to 0) when the summed weighted inputs S, both external and from all other elements, reach some threshold condition, $S \geq T$. A node thereby receives multiple inputs but produces a single output. A typical nonlinear feedback to a node is the sigmoid function $1/[1 + \exp(-Gz)]$, with G the gain parameter and z the sum of binary inputs from other nodes, each input taking values ± 1. Then $+1$ acts as an excitation and -1 an inhibition. Generally, a neural network as a whole accepts multiple inputs and delivers multiple outputs. There is a wealth of significantly differing neural networks using these simple ideas, along with feedforward, feedback, nonlinear dynamics, and hidden layers. A hidden layer is a collection of nodes not directly connected to input or output. The weights can generally be made to change in the training mode of operation. *Back-propagation* is the name for algorithms by which output errors are sent back to a hidden layer to process changes of the weights (Zurada, 1992).

Neural networks have had a slow beginning from their earliest inceptions 50 years ago, which antedated Rosenblatt's feedforward "perceptron" of 1962. But now the effort devoted to neural network research and implementation has become staggering. The reason is partly that neural networks can be trained to perform easily and quickly valuable pattern recognition tasks. Such tasks as optical character recognition (OCR), financial forecasting (through function estimation), and manufacturing process control are among present major uses of neural networks (Hammerstrom, 1993). Neural networks are parallel structures, and recently developed parallel integrated circuit devices have increased performance speeds by factors of 10^3. Training times tend to be long, on the order of months, but the trained neural network can often perform in real time. A drawback is that although the nonlinearity and feedback structures of neural networks mirror complexities of real systems, their results, even when correct, are difficult to explain.

Some implementations have a capacity for unsupervised learning. Some of the Hopfield type, with symmetric connections, achieve associative memory, and neural networks over a wide range of complexity can become chaotic. Neural networks are used in meteorology, where their pattern recognition and predictive skills are competitive with older methods (Elsner and Tsonis, 1993).

4.3 Time-Series Analysis

A *time series* is a record of the values of any set of fluctuating quantities measured at different times. In the analysis of time series, regularities are sought, sometimes for making predictions. If there is more than one fluctuating quantity, the time series is called *multivariate*; otherwise, it is called *univariate* or scalar. A time series can be continuous or discrete. Time-series analysis is used in information theory, dynamical systems theory, systems analysis, and data assimilation.

Recent work is considerably beyond the pioneering 1956 success of Lorenz, for using PCA to uncover dynamical processes in multivariate time series (Vautard *et al.*, 1992) and to forecast. A related El Niño forecast is discussed in Sec. 5.1 (Keppenne and Ghil, 1992). The analysis of noise is often crucial: for example, whether determinism in paleoclimate variability could be construed from the present climatic time series data hangs in the balance of the inference of noise (Vautard and Ghil, 1989). Retreating from criteria for chaotic determinism, one can settle for less and test for the existence of nonlinearity in short, noisy series (Theiler *et al.*, 1992). Testing for nonlinearity uses advanced statistical inference techniques, to be discussed. In spoken communication, where the temporal behavior of the sound pulse is vital, signal processing has made use of SVD and PCA applied to the discretized data stream.

The modeling of time series is generally begun with the definition of a *strict white noise* process, denoted by $\{e_t\}$. For a stationary (i.e., exhibiting time-independent statistical behavior) strict white noise process, the mean μ and variance σ^2 are independent of time t. In any case, the autocovariance function, $\mathrm{cov}\{e_t,e_{t'}\}$, is zero for all $t \neq t'$. Here we specialize our discussion to the covariance function for a univariate continuous random variable X_t, given by

$$\mathrm{cov}\{X_t,X_{t'}\} = E[(X_t - \mu)(X_{t'} - \mu)]. \tag{17}$$

The expectation, or mean, is just $E[X_t] = \mu$, a constant whenever X_t is stationary.

Modeling in terms of a random variable such as X_t implies the existence of an associated probability distribution p_t, which is time dependent unless X_t is stationary. If X_t can take a continuous range of values, $x \in [a,b]$, then p_t is called a *probability density function*. One then defines the *expectation E* of X_t at time t by

$$E[X_t] = \mu_t = \int_a^b x p_t(x) dx. \tag{18}$$

Finally, the *variance* of X_t is $\sigma^2 = \mathrm{cov}\{X_t,X_t\}$. Such statistical quantities are estimated from the data stream. A computational method is often a basis for choosing a form of time-series analysis.

Although for brevity we explicitly refer to scalar time series, the extensions to the multivariate case are generally necessary and straightforward. Even scalar series from a dynamical system become multivariate through time-delay embedding. Consider now a discretized time series $\{X_t\}$ with t taking on only integer values. A model for this series is an equation

$$F\{\ldots X_{t-2}, X_{t-1}, X_t, X_{t+1}, X_{t+2}, \ldots\} = e_t, \tag{19}$$

where e_t is a zero-mean strict white noise process and $F\{\cdot\}$ is the model function. Finding the model function is equivalent to reducing the unexplained part of the series to a strict white noise process. Well-known linear model functions include AR (autoregressive), MA (moving average), and ARMA (autoregressive/moving average). The ARMA(k,l) model includes the AR(k) and MA(l) models as special cases and is expressible as a linear operator: $X_t = G(B)e_t$, where $G(u)$ is the rational function $G(u) = \beta_l(u)/\alpha_k(u)$, and u is replaced by the linear back-shift operator, B: $Be_t = e_{t-1}$. Also, $\beta_l(z) = 1 + b_1 z + \cdots + b_l z^l$ and $\alpha_k(u) = 1 + a_1 u + \cdots + a_k u^k$. This process can be pictured as representing the action of a linear filter. The estimation of the unknown parameters $k, l, a_1, \ldots, a_k, b_1, \ldots, b_t$ is a model-fitting problem. For convergence with this model, it is necessary that the time series be stationary; also, consistency requires $G(u)$ to have no poles for $|u| \leq 1$. An ARMA model works only if the system is stationary. Optimal simulation and parameter estimation both for stationary and nonstationary time series can be done through modeling of filters.

An explicit representation of nonlinear series arises from inverting and expanding the function F. Just up to the first nonlinear term, the resulting discrete-time Volterra series is

$$X_t = \mu + \sum_{i=0}^{\infty} h_i e_{t-i} + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} h_{ij} e_{t-i} e_{t-j} + \cdots. \tag{20}$$

A way to look for nonlinearity is to compute polyspectra. Polyspectra are multiple Fourier transforms of third- and higher-order moments of the series about its mean. If $\{X_t\}$ is a Gaussian process, the bispectrum and all higher-order polyspectra vanish (Priestly, 1988).

4.3.1 Signal Processing With digital technology come new and powerful modeling techniques following quantization. A signal is quantized by sampling it at a time interval T, the sampling period, and rounding to a certain fixed number of significant figures. The quantized signal (and also the unquantized one) can be subjected to *spectral analysis*, which is the determination of its frequency content. The Fourier transform, in either its discrete or its windowed version, permits useful analysis in the frequency domain. Finding the finite Fourier transform is a common step in spectral analysis. It can now be performed in real time by use of the fast Fourier-transform (FFT) algorithm proposed by Cooley and Tukey in 1965 (see FOURIER AND OTHER MATHEMATICAL TRANSFORMS). The power spectrum $S(f)$, which is the square of the magnitude of the Fourier transform, provides the most accessible vital characterization of a signal. For a chaotic system, the windowed power spectrum must be continuous and not just contain a large, finite collection of peaks. Quasiperiodicity and complicated periodicity can be identified but not usually distinguished in the power spectrum. The Fourier transform creates a frequency-domain representation sometimes used in ARMA, other parameter-estimation algorithms, and SVD (Kumaresan, 1993). Meteorological and oceanographic PCA can also be profitably recast through Fourier transforms of the time or the azimuth variables, to focus upon wave structures (Bernardet *et al.*, 1990).

Discrete-time analysis uses the z transform, a generalization of the Fourier transform. The z transform of a signal expressible as a sum of exponentials can be written as a rational function in the variable $z = \exp(i\omega)$. This allows a "pole–zero" approach to signal analysis and the use of non-numerical, symbol manipulation modeling of signals. In the context of digital signal processing, it is easy to understand various pitfalls of spectral analysis such as *aliasing*, the misidentification of power spectrum frequencies because of low sampling rates (Oppenheim and Schafer, 1989).

4.4 Statistical Inference

Statistical inference in the computer age has dramatically changed (Efron and Tibshirani, 1991) and is developing rapidly. Powerful formal results such as the central limit theorem, which establishes asymptotic normal distributions (Priestly, 1982), are more useful than ever, because digital computers help reach the asymptotic region in a statistical analysis. More important is that computer experiments make it possible to learn some of the limitations of statistical inference methods. Estimation errors are more easily evaluated from finite sampling of infinite populations, noise, and finite sampling of finite populations. With the use of current computer-intensive data processing techniques, the original data can be

1. sampled through partitioning or through successive computer-assisted selection models;
2. altered in definable ways to generate surrogate data; or
3. reorganized along paths linking common characteristics.

The jackknife and bootstrap methods exemplify **1**, randomizing phases of Fourier-transformed data exemplify **2**, and the CART (classification and regression trees) method exemplifies **3**. Given small and imperfect historical data sets of natural phenomena such as sunspot activity, global temperatures, or various proxy data for global climate (Peixoto and Oort, 1992), processing techniques give new meaning to calculations of fractal dimensions (Vautard and Ghil, 1989; Theiler *et al.*, 1992) in searches for underlying nonlinear dynamics and possibly chaos. Theories of processing of samples of finite populations are developed along similar lines by Särndal *et al.* (1992).

Simplicity serves as a guiding principle in model development (Harte, 1988). The Bayesian approach to scientific inference, though controversial, gives support to simplicity. The Bayesian approach asserts that probability can be associated with plausibility in evaluation of theories. In particular, a theory that gives the higher probability to what is actually observed is the more likely to be correct. Bayes's theorem, proved by the English clergyman in 1761, is suggestive of such an approach (Jefferys and Berger, 1992): Let $P(X|Y)$ be the conditional probability that hypothesis X is true, given information Y. Then Bayes's theorem states

$$P(H_i|D\&I) = \frac{P(D|H_i\&I)P(H_i|I)}{P(D|I)}. \tag{21}$$

This theorem relates the contingent probability of the truth of hypothesis H_i to the prior information I and to the new data D. In this way, Bayesian inference supports simplicity: Consider an alternative, more complicated hypothesis H_j. If the new data are more sharply predicted by H_i than by H_j, then this can increase $P(H_i|D\&I)$. But by their nature, simpler theories tend to have fewer possible outcomes over which to spread the probability, tending to make sharper predictions.

Less controversial is another plausible inference method, the method of maximum likelihood, which also has an analogy with Bayes's theorem. Maximum likelihood is used in the same contexts as least-squares estimation: Assume that the probability density for a measured vector $\mathbf{X}$ is $f(\mathbf{X},\boldsymbol{\Theta})$, where the parameter vector $\boldsymbol{\Theta}$ is to be estimated from the data. One reinterprets $f(\mathbf{X},\boldsymbol{\Theta})$ to signify the likelihood for a particular choice of $\boldsymbol{\Theta}$, given the measurement $\mathbf{X}$. The parameter estimate is reduced to finding that parameter vector $\hat{\boldsymbol{\Theta}}$ that maximizes f or $\ln(f)$. Under general conditions, the estimator $\hat{\boldsymbol{\Theta}}$ can be proved to have a normal distribution asymptotically when the number n of measurements gets large (Priestly, 1982). With computing power available, the details of the large-n limit become more tractable, increasing the significance of such formal results.

4.5 Information Theory

Information theory was developed as a mathematical model of communication by Shannon in 1948. It considers transmission rates in the presence of noise (Gallager, 1968). Defined by analogy with physical entropy, the information H is given by

$$H(Q) = -\sum_i p_i \log_2(p_i), \tag{22}$$

where p_i is the probability of a possible message i. Here, Q represents the space of all possible messages, and $H(Q)$ is the information gained, on the average, from measuring which message occurs. All messages are treated as statistically independent, and $\Sigma_i p_i = 1$. A continuum formulation is also often used. An important quantity in communication theory is *mutual information*, $I(Q,S)$, defined in the space of ordered pairs of outcomes from Q and S, and given by

$$I(Q,S) = H(Q) + H(S) - H(S,Q). \qquad (23)$$

Given a measurement in S, $I(Q,S)$ gives the number of bits of information, on the average, that can be predicted about a subsequent measurement in Q. Here, I can represent a rate with which information crosses a channel in the presence of noise, computable from the conditional distribution function $P_{q|s}(q_i,s_j)$, where $q_i \in Q$ and $s_j \in S$. Shannon's channel capacity theorem employs $I(Q,S)$.

Information is an important concept for dynamical systems. So are extensions of information, following work in the late 1950s of Kolmogoroff and Sinai (Schuster, 1988). The Kolmogoroff-Sinai (or metric) entropy K_{KS} and an infinite sequence of generalized entropies are computable from time series, where K_{KS} is a measure of the rate at which information is generated in a map or a flow. Chaotic systems constantly generate information; consequently, the information existing at any one moment is insufficient to make long-range predictions. Hence, the long-range unpredictability of the weather (viewed as a deterministic chaotic system) is explained in information-theoretic terms. A random sequence has an infinite K_{KS}, because the same finite amount of information ΔH is generated at successive times, no matter how small the time interval Δt, and hence $dH/dt = \infty$. Mutual information is also definable for a map or flow, giving the number of bits of information, on the average, that can be predicted about the measurement of the state at time t from having measured the state at a time $t - T$. When time-delay embedding is used, the delay time τ taken as the first minimum of the mutual information of the vectorized series is more optimal than a τ chosen as the first zero of the correlation function (Fraser and Swinney, 1986).

Ergodic theory is related to information theory. By the ergodic property, the time average of an observable equals its phase-space average. Only when a dynamical system is ergodic are most of the procedures described in this article for analyzing dynamical systems valid. Formal theorems exist using K_{KS} to compare ergodic systems (Schuster, 1988; Lichtenberg and Lieberman, 1992).

5. APPLICATIONS

5.1 El Niño

El Niño is an anomalous persistent warming of the eastern equatorial Pacific that occurs irregularly, generally every two to five years. Invariably, this sea-surface temperature (SST) anomaly is accompanied by a reversal of the usual low atmospheric pressure in the western equatorial Pacific and high pressure in the eastern equatorial Pacific. One measure of this is the southern oscillation index (SOI). The SOI is a scalar time series that can be constructed as a normalized and smoothed sea-level pressure difference between Darwin, Australia, and Tahiti (Keppenne and Ghil, 1992).

A mechanism for El Niño seems clear: positive feedback between the equatorial easterly trade winds and the SST gradient. When the pressures reverse, the mean wind forcing of the Pacific surface waters toward the west stops, and as a consequence, the warmer surface waters accumulate off the western South American coast. Upwelling of cold eastern Pacific water stops. This promotes a convergence of warm moist air, cloudiness, precipitation, and low-surface atmospheric pressure. This positive feedback works in the other direction, as well, to produce a cold SST episode (La Niña). The two coupled phenomena are referred to as ENSO (El Niño–Southern Oscillation). It is ENSO that is modeled (Philander, 1990). Predicting ENSO would have practical importance to the South American fishing industry, as ENSO episodes are destructive to west coastal fish abundance. Various globally distributed large-scale atmospheric and oceanic phenomena are correlated with ENSO, referred to as "teleconnections."

Modeling ENSO poses a major challenge. So far, mechanistic models can reproduce the phenomenon qualitatively, without necessarily being reliable predictors either of the occurrence or the structure of episodes. The ENSO occurs in a large variety of ways with respect to frequency, intensity, duration, and spatial distribution (Philander, 1990). The extent to which GCM's will be able to describe and/or predict ENSO and its teleconnections is an open question. To what extent tropically based mechanisms are sufficient to trigger ENSO episodes is an important theoretical question. On a more technical level, the role played by the annual weather cycle and the relationship of the quasibiennial oscillation (QBO) to ENSO, if any, require clarification. The QBO is a tropical lower stratospheric fluctuation of zonal winds of approximately a 2-yr period.

One simple mechanistic model of ENSO is Vallis' one-dimensional, two-point description, the equations of which resemble the Lorenz system. In his model, a mean Pacific equatorial wind creates a mean westward ocean flow, with feedbacks. The equations are

$$\frac{du}{dt} = B(T_e - T_w)/2\Delta x - C(u - u^*); \tag{24}$$

$$\frac{dT_w}{dt} = \frac{u}{2\Delta x}(\bar{T} - T_e) - A(T_w - T^*); \tag{25}$$

$$\frac{dT_e}{dt} = \frac{u}{2\Delta x}(T_w - \bar{T}) - A(T_e - T^*). \tag{26}$$

Here, T_e and T_w are east and west upper-ocean temperatures separated at a distance Δx, $\bar{T}$ is the deep-ocean constant temperature, and u is eastward oceanic current. The terms A, B, C, u^*, and T^* are constants. The mechanism is simply that an eastward (westward) wind field driven in part by a temperature difference $T_e > T_w$ ($T_e < T_w$) would produce an oceanic downwelling (upwelling) in the east and upwelling (downwelling) in the west. The equations conserve mass. Low-dimensional chaos is produced, as can be inferred by evaluation of the random-appearing model data.

Vallis has extended this mechanism to a one-dimensional continuum and integrated the resulting equations, making the chaos disappear and fluctuations be modulated (Vallis, 1988). One-dimensional models allow easy exploration of ideas. With stochastic forcing, the Vallis continuum model can produce stochastic resonance oscillations. When natural seasonal variability is introduced, ENSO appears as the occasional enlargement of an annual signal. In the spirit of Lorenz's original modeling effort, the prediction itself of chaos is interesting. Because of the paucity of long-term, frequently sampled time series for ENSO, chaos is difficult to corroborate experimentally.

A recent model has used simplified equatorial ocean and atmosphere dynamics, computed on an equatorial grid. In turn, Münnich *et al.* (1991) have constructed an interpretive mechanistic model of that model. Included is oceanic thermal reservoir feedback, expressed as a nonlinear relation between the depth of the eastern Pacific ocean thermocline, h_e, and the amplitude A of the zonal forcing wind stress τ. The thermocline depth is the mean depth of the thin oceanic layer where the temperature changes rapidly between the warm surface and the cold deeper region. Hence, h_e is a measure of the thermal energy of the western equatorial Pacific ocean. When $A(h_e)$ is expressed as a nonlinear function, bifurcations occur as the strength parameter of the function $A(h_e)$ increases. Successively more realistic parameterizations, starting with a simple cubic nonlinearity

$$A(h_e) = \kappa(h_e - h_e^3), \tag{27}$$

have all produced three bifurcations, the last being into a nonperiodic regime as the strength parameter κ is increased. This pattern of bifurcations suggests that a chaotic state has been reached *via* the Ruelle-Takens-Newhouse route (Schuster, 1988). For their model of ENSO, Münnich *et al.* conclude that stochastic forcing is not needed, various mechanisms for chaos are possible within their simplified framework, and annual forcing enhances ENSO episodes.

Keppenne and Ghil (1992) have done a PCA of a time series of the SOI. A temporal PCA is referred to as a singular spectral analysis (SSA). By keeping just the first four principal components in a reconstruction of the time series, they filter out higher-frequency oscillations. This smooths the reconstructed series, filtering some possible noise perturbations of a deterministic system. The agreement between the 50-yr time series (smoothed by taking a 5-month running mean) and the recombined first four PCs is remarkable, with El Niño and La Niña episodes identified with all the maxima and minima of greatest amplitude. Yet these four principal components only account for 21.7% of the SOI variance! When Keppenne and Ghil combine the two high-frequency components and the two low-frequency components separately, they obtain smooth oscillatory behavior amenable to autoregressive linear prediction. The 36-month predictions agree with El Niño and La Niña events over the last ten years, and suggest a La Niña event in the winter of 1993–1994. This prediction uses no SST data at all and a minimal procedure—time-delay embedding and temporal PCA analysis—for filtering out external noise.

5.2 Chaotic Ocean Heat Transport

Climatic data show fluctuations on all time scales. In the long paleoclimatological time scales, the fluctuations are partly governed by stochastic mixing in data storaging and accessing processes (Crowley and North, 1991). However, climate also shows a significant variability associated with Earth's orbital variations on these same time scales. On decadal and interdecadal time scales, some observed fluctuations may be associated with the important poleward ocean transport of heat driven by thermohaline circulation (THC) (Covey, 1991). THC is the largely vertical oceanic motion caused by differing densities. It is an Archimedes-principle buoyancy that lifts warmer and less saline water. In the present climatic regime, THC produces an equator-to-poles flow of warm surface water and a deep cold-water return flow.

Modeling searches for a significant climatic impact of THC. Just as with ENSO, a major problem is to model the air-sea interaction. Here we discuss models where the ocean physics (momentum, heat, and salinity transport) is discretized while boundary conditions describe the atmospheric effects. Models differ on the amount of natural variability and stability. Recent numerical experiments with oceanic GCM's (OGCM's) have produced a phenomenon known as the polar halocline catastrophe. This is a blocking of the polar downwelling component of oceanic circulation, to alter dramatically oceanic salinity and heat fluxes.

A recent series of numerical OGCM experiments (Weaver *et al.*, 1993) has parametrized the freshwater flux forcing of THC in an OGCM computed with a coarse grid and other features allowing for rapid numerical modeling. One of these features is use of a linearized numerical model for horizontal momentum transport. Subsystem time steps are allowed to differ, and they range between 2 h and 5 d. The ocean system is a flat-bottomed basin 60° wide, with latitude ranges from 0° to between 64°N and 72°N. Different runs of this OGCM employ 15 to 33 vertical levels and horizontal grids ranging from 2° × 2° to 4° × 3.75°. This "spherical cow" is not quite complete: boundary conditions are needed, both to "wind up" the system, i.e., to start it up and put it into some climatologically reasonable steady state, and to maintain a reasonable air-sea interaction. A temporally constant wind field and wind-stress forcing are set up. The air-sea interaction involves only negligible transport of salts between air and sea. However, the freshwater flux, $P-E$, where P is precipitation and E is evaporation (in zonally averaged units of $\mathrm{m \cdot yr^{-1}}$) clearly affects salinity (measured in grams of salts per kilogram of seawater). During the windup, the salinity and heat fluxes are subjected to restoring damping laws:

$$Q_T = C_T(T_O - T_A), \tag{28}$$

$$Q_S = C_S(S_O - S_A), \tag{29}$$

where Q_T and Q_S are heat and salinity fluxes, C_T and C_S are positive constants, while T_O and T_A are temperatures and S_O and S_A are salinities of ocean and atmosphere. The forcing temperature and salinity fields T_A and S_A are prescribed from data.

After windups to experimental zonally averaged salinity and temperature fields, the systems are cut loose from the restoring salinity forcing, but are subjected to a temporally constant value of $P - E$, inferred from the model in its wound-up steady state. This is called mixed boundary conditions, because the thermal forcing is still maintained. Grounds for mixed boundary conditions are the absence of any physical basis for forcing through salinity damping.

Here is where chaos can set in. If the salinity boundary condition used in spin-ups equals the zonally averaged salinity field, a $P - E$ minimum occurs at 54°N, and an unstable downwelling takes place at that point, a reverse circulation cell occurring at high latitudes. An irregular flush takes place on the average every eight years, restoring to normal downwelling at the Pole. A stochastic forcing component changes the frequency and intensity of the flushes. To what extent the reported results here are artifacts is a matter of ongoing debate. Again, in the spirit of the Lorenz equations, what is important here is perhaps not this model's validation and its detailed predictions so much as the possibility of chaos or stochastic resonance it presents.

5.3 Controlling Chaos

Controlling chaos is a process of forcing a chaotic system to reach and remain in a state or a sequence of states. Control theory uses the high leverage of chaotic behavior to advantage: SIC is exploited both to rapidly target a state and to remain there. The formal insight that chaotic orbits consist of unstable periodic orbits and their limit points can be used here. A full range of technical tools such as embedding and numerical determination of the stability matrix (with or without knowledge of the underlying dynamics) has already been employed since some practical methods of controlling chaos were introduced in 1990 by Ott, Grebogi, and Yorke (Auerbach *et al.*, 1992). Technological applications have been rapid, and to date they include controlling the following chaotic systems: vibrations of a magnetoelastic ribbon, lasers operating at ultrahigh intensities, diode circuits, and cardiac arrhythmia (Garfinkel *et al.*, 1992). An associated message-encoding algorithm has been developed that appears robust in the presence of noise and advantageous for information transmission (Hayes *et al.*, 1993).

5.4 Solitons

Solitons are solutions of nonlinear, dispersive wave equations with particle-like properties. They are spatially localized and do not change shape either when propagating or after colliding with other solitons. Solitons can only be solutions to equations for which the effects of dispersion and nonlinearity somehow cancel. Theoretical interest has been high since demonstration of their mathematical properties in a numerical computer experiment by Zabusky and Kruskal in 1965, although a physical soliton had been observed over a century before (in 1834) as a wave in a Scottish canal. Solitons of widely differing topological characteristics arise in many mathematical systems, including nuclear interactions, elementary particles, and plasmas. Now, engineering interest is high, too. Solitonic pulses on optical fibers may soon become the mechanism for transoceanic communications, at rates exceeding 10^{10} bits per second (10 Gb/s) (Haus, 1993).

The history of solitons reveals interplay among different approaches to modeling. In 1895, Kortweg and deVries succeeded in deriving a shallow-water equation for small-amplitude gravity waves. Their (KdV) equation takes the form

$$u_t - 6uu_x + u_{xxx} = 0. \tag{30}$$

Here the subscripts refer to time and space differentiation. Dissipation effects were ignored in the derivation. Analytic solutions were known in 1895, but the general stability and behavior of multiple-soliton solutions were first obtained numerically in 1965. Then, with another breakthrough, Gardner *et al.* in 1967 discovered powerful formal solutions by reducing the problem to inverse scattering in a linear Schrödinger-type equation. The developments of *inverse scattering theory* (IST), originally motivated by nuclear physics, saw their first major application in this soliton context. Nonlinear equations, if cast into IST form, are shown to be integrable—as is the one-dimensional nonlinear Schrödinger equation, relevant to optical fibers.

There is no general recipe for determining whether a nonlinear equation is reducible to IST (Drazin and Johnson, 1990). However, a determined effort is underway to develop better tests, and with them, appropriate reductions to IST (Ames and Rogers, 1992). One key would be to find sufficiently simple equivalent characterizations of complete integrability (Ablowitz and Segur, 1981). This can be succinctly expressed in the language of Hamiltonian dynamics. Qualitatively, completely integrable systems of differential equations when cast into the Hamiltonian form must have a sufficient number of properly constrained functionals whose Poisson brackets with each other vanish. Not all initial conditions lead to solitonic solutions. If a solitonic solution occurs, the IST generates the soliton or solitons through a process equivalent to generating action-angle variables. Such action-angle variables contain constants of the motion required to exist from complete integrability. Some of the relationships among some possible alternative approaches to determining complete integrability are discussed by Ablowitz and Segur (1981), and newer formulations and updates are given by Ames and Rogers (1992). For more details, the reader is referred to the article NONLINEAR SYSTEMS.

The IST is concerned with finding a linear Schrödinger-type equation, given the data in the form of asymptotic behavior of its solution. What remains to be determined is a potential energy operator. The integral equations for that potential, the Marchenko equations, are linear; hence, IST simplifies the solution of the original nonlinear equation. The data include scattering phase shifts (or in one dimension, reflection coefficients), and bound-state poles and residues. Solitonic solutions are found after setting the reflection coefficient equal to zero. Relevant to optical fiber design, the nonzero reflection-coefficient case is algebraically challenging, even when data are provided in terms of rational functions (Sabatier, 1988). The IST work should benefit from computer algebra. As yet, only numerical methods are used to study solitonic stability in the presence of dissipation.

GLOSSARY

Cellular Automaton: Dynamical system temporally and spatially discrete (in cells).

Chaotic System: Deterministic dynamical system that exhibits noisy, random-appearing behavior. Technically, must show sensitivity to initial conditions (SIC).

Deterministic Dynamical System: Dynamical system having a rule that specifies its state at the next instant of time uniquely, given the state at any specific time.

Discretization: The approximation of a continuum by a discrete set of points or a discrete set of elements. These elements can be finite physical segments, or they can be sets of functions. Invariably performed when there are space-time dimensions to be modeled.

Dynamical System: Mathematical model of time evolution.

El Niño: Sporadic anomalous warming of the waters of the eastern equatorial Pacific.

Filter: General term for a dynamical system that performs a task such as tuning or removing some unwanted variance, such as noise, in a signal.

Finite Elements: A form of discretization of a continuum that treats the system as composed of small, highly symmetric, interacting parts, such as polyhedra.

Flow: Name for continuous time evolution of a dynamical system.

General Circulation Model (GCM): Parametrized simulation model of the circulation within Earth's oceans, the global atmosphere, or both combined.

Importance Sampling: A method for generating a sequence of configurations for performing Monte Carlo (MC) calculations, in which successive configurations are correlated.

Inverse Scattering Theory (IST): A formulation of scattering theory, the solution of which is a potential operator in a Schrödinger-type equation, given sufficient data.

La Niña: Sporadic anomalous cooling of the waters of the eastern equatorial Pacific.

Map (Or Iterative Map): Name for discrete-time evolution of a dynamical system.

Mechanistic Model: Mathematical model that attempts to describe the processes, or mechanisms, within a system that are relevant to the problem at hand.

Markov Process of Order *k*: Dynamical process that is accurately described in terms of transition probabilities between states, or configurations, of the system, $P(\mathbf{X} \rightarrow \mathbf{X}')$, that are functions only of the last k states. When $k = 1$, it is simply referred to as Markov process.

Monte Carlo Method: Computational method of great scope based upon the use of random-number sequences and stochastic algorithms.

Neural Network: A system possessing elements, called nodes or neurons, with large numbers of interconnections. Patterned after real neural systems and capable of learning to perform complex tasks such as pattern-matching.

Padé Approximant: Rational approximation matching a function's partial Taylor series.

Paleoclimatology: The study of climate on time scales that can be observed using historical and geological data ranging, approximately, from 10^2 to 10^8 years.

Parametrization: Introduction of a mechanistic description of an important subsystem within a simulation model.

Power Spectrum: The magnitude squared of a Fourier transform. An important tool for distinguishing periodicity and quasiperiodicity from noise and $1/f$ phenomena.

Pseudorandom Numbers: A sequence of numbers generated by a computer algorithm, which passes a set of tests for randomness.

Rational Function: Function expressed as a ratio of two polynomials.

Self-Organized Criticality: State toward which some large systems naturally evolve, characterized by spatial and temporal self-similarity. Weakly chaotic.

Simulation Model: Model containing as many of the known relevant degrees of freedom of the system as possible. In the ideal limiting case, there is no parametrization.

Soliton: A particle-like solution to a nonlinear, dispersive wave equation. It is localized and retains its shape while propagating or upon colliding with another soliton.

Statistical Padé Approximant: Rational function that is fitted to numerical data by minimizing errors in some way.

Stochastic Dynamical System: Dynamical system having a rule that specifies its state at the next instant of time only as one of several possible states with some probability.

Synergetics: An approach to the study of large, complex systems, which seeks explanations in terms of the slaving principle, bifurcations, order parameters, and physical principles.

Time-Delay Embedding: Technique used to reconstruct a vector orbit in a nonlinear dynamical system, using time-delayed values of a scalar observable $y(t)$.

Works Cited

Abraham, N., Albano, A. M., Passamante, A., Rapp, P. E. (Eds.) (1990), *Measures of Complexity and Chaos*, New York: Plenum.

Ablowitz, M. J., Segur, H. (1981), *Solitons and the Inverse Scattering Transform*, Philadelphia: SIAM.

Ames, W. F., Rogers, C. (Eds.) (1992), *Nonlinear Equations in the Applied Sciences*, New York: Academic Press.

Antoulas, A. A. (Ed.) (1991), *Mathematical System Theory*, Berlin: Springer-Verlag.

Appel, K., Haken, W. (1978), in: L. A. Steen (Ed.), *Mathematics Today*, New York: Springer-Verlag.

Arfken, G. (1985), *Mathematical Methods for Physicists*, 3rd ed., Orlando: Academic Press.

Auerbach, D., Grebogi, C., Ott, E., Yorke, J. A. (1992), *Phys. Rev. Lett.* **69**, 3479–3482.

Bak, P., Tang, C., Wiesenfeld, K. (1987), *Phys. Rev. Lett.* **59**, 381–384.

Baker, G. A., Jr., Graves-Morris, P. (1981), *Padé Approximants Part I & Part II*, Reading, MA: Addison-Wesley.

Bernardet, P., Butet, A., Déqué, M., Ghil, M., Pfeffer, R. L. (1990), *J. Atmos. Sci.* **47**, 3023–3043.

Binder, K. (Ed.) (1987), *Applications of the Monte Carlo Method in Statistical Physics*, 2nd ed., Berlin: Springer-Verlag.

Brandstater, A., Swinney, H. L. (1987), *Phys. Rev. A* **35**, 2207–2220.

Butler, F., Chen, H., Sexton, J., Vaccarino, A., Weingarten, D. (1993), *Phys. Rev. Lett.* **70**, 2849–2852.

Carter, E. F. (1989), *Dynamics of Atmospheres and Oceans* **13**, 335–348.

Casti, J. L. (1992), *Reality Rules: I & II, Picturing The World In Mathematics*, New York: Wiley-Interscience.

Chui, C. K., Chen, G. (1987), *Kalman Filtering*, Berlin: Springer-Verlag.

Cook, R. D., Malkus, D. S., Plesha, M. E. (1989), *Concepts and Applications of Finite Element Analysis*, New York: Wiley.

Covey, C. (1991), *Nature* **353**, 796–797.

Creutz, M. (1992) (Ed.), *Quantum Fields on the Computer*, Singapore: World Scientific.

Crowley, T. J., North, G. R. (1991), *Paleoclimatology*, Oxford: Oxford University Press.

Daubechies, I. (1992), *Ten Lectures on Wavelets*, Philadelphia: Society for Industrial and Applied Mathematics.

Drazin, P. G., Johnson, R. S. (1990), *Solitons: An Introduction*, Cambridge, U.K.: Cambridge University Press.

Efron, B., Tibshirani, R. (1991), *Science* **253**, 390–395.

Elsner, J. B., Tsonis, A. A. (1993), *Bull. Am. Meteor. Soc.* **74**, 243.

Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.

Ferrenberg, A. M., Landau, D. P., Wong, Y. J. (1992), *Phys. Rev. Lett.* **69**, 3382–3384.

Fitch, J. (1993), *Phys. World* **6**, 48–52.

Ford, J., Mantica, G. (1992), *Am. J. Phys.* **60**, 1086–1097.

Fraser, A. M., Swinney, H. L. (1986), *Phys. Rev. A* **33**, 1134–1140.

Gallager, R. G. (1968), *Information Theory and Reliable Communication*, New York: Wiley.

Garfinkel, A., Spano, M. L., Ditto, W. L., Weiss, J. N. (1992), *Science* **257**, 1230–1235.

Gerhardt, M., Schuster, H., Tyson, J. J. (1991), *Phys. D* **50**, 189–206.

Ghil, M., Malanotte-Rizzoli, P. (1991), in: R. Dmowska, B. Saltzman (Eds.), *Advances in Geophysics*, vol. 33, San Diego: Academic Press, pp. 141–266.

Gould, H., Tobochnik, J. (1988), *Computer Simulation Methods, Part I and Part II*, Reading, MA: Addison-Wesley.

Grassberger, P. (1986), *Nature* **323**, 609–612.

Hammel, S. M., Yorke, J. A., Grebogi, C. (1988), *Bull. Am. Math. Soc.* **19**, 465–469.

Hammerstrom, D. (1993), *IEEE Spectrum* **June**, 26–32.

Harte, J. (1988), *Consider a Spherical Cow*, Mill Valley, CA: University Science Books.

Hassani, S. (1991), *Foundations of Mathematical Physics*, Boston: Allyn and Bacon.

Hayes, S., Grebogi, C., Ott, E. (1993), *Phys. Rev. Lett.* **70**, 3031–3034.

Hattori, T., Takesue, S. (1991), *Phys. D* 295–322.

Haus, H. A. (1993), *IEEE Spectrum* **March**, 48–53.

Huynh, H. T. (1993), *SIAM J. Numer. Anal.* **30**, 57–102.

Jefferys, W. H., Berger, J. O. (1992), *Am. Sci.* **80**, 64–72.

Kalos, M. H., Whitlock, P. A. (1986), *Monte Carlo Methods*, Vol. 1, New York: Wiley-Interscience.

Keppenne, C. L., Ghil, M. (1992), *J. Geophys. Res.* **97**, 20449–20454.

Knuth, D. E. (1981), *The Art of Computer Programming*, vol. 2, Reading, MA: Addison-Wesley.

Kumaresan, R. (1993), in: S. K. Mitra, J. F. Kaiser (Eds.), *Handbook for Digital Signal Processing*, New York: Wiley, Chap. 16.

Lakshmivarahan, S., Dhall, S. K. (1990), *Analysis and Design of Parallel Algorithms: Arithmetic and Matrix Problems*, New York: McGraw Hill.

Levi, B. G., Hafemeister, D., Scribner, R. (Eds.) (1992), *Global Warming: Physics and Facts*, New York: American Institute of Physics.

Lichtenberg, A. J., Lieberman, M. A. (1992), *Regular and Chaotic Dynamics*, 2nd ed., New York: Springer-Verlag.

Lorenz, E. N. (1963), *J. Atmos. Sci.* **20**, 130–141.

Mandelbrot, B. B. (1983), *The Fractal Geometry of Nature*, New York: Freeman.

Marchuk, G. I., Mikhailov, G. A., Nazarliev, M. A., Darbinjan, R. A., Kargin, B. A., Elepov, B. S. (1980), *The Monte Carlo Methods in Atmospheric Optics*, Berlin: Springer-Verlag.

McInnes, A. W., Marshall, T. H. (1993), *J. Approx. Theory* **75**, 85–106.

Moss, F. (1991), *Ber. Bunsenges. Phys. Chem.* **95**, 303–311.

Münnich, M., Cane, M. A., Zebiak, S. E. (1991), *J. Atmos. Sci.* **47**, 1562–1577.

O'Brien, J. J. (Ed.) (1986), *Advanced Physical Oceanographic Numerical Modeling*, Dordrecht: Reidel.

Oppenheim, A. V., Schafer, R. W. (1989), *Discrete-Time Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall.

Parlitz, U. (1992), *Int. J. Bifurcation and Chaos* **2**, 155–165.

Peixoto, J. P., Oort, A. H. (1992), *Physics of Climate*, New York: American Institute of Physics.

Philander, S. G. (1990), *El Niño, La Niña, and the Southern Oscillation*, San Diego: Academic.

Pool, R. (1992), *Science* **256**, 44–62.

Preisendorfer, R. W. (1988), *Principal Component Analysis in Meteorology and Oceanography*, Amsterdam: Elsevier.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vettering, W. T. (1987), *Numerical Recipes*, Cambridge, U.K.: Cambridge University Press.

Priestly, M. B. (1982), *Spectral Analysis and Time Series*, 2 vols., London: Academic Press.

Priestly, M. B. (1988), *Non-Linear and Non-Stationary Time Series Analysis*, San Diego: Academic.

Rao, N. S., Roxin, E. O. (1990), *SIAM J. Appl. Math.* **50**, 853–864.

Sabatier, P. C. (Ed.) (1988), *Some Topics on Inverse Problems*, Singapore: World Scientific.

Sabelfeld, K. K. (1991), *Monte Carlo Methods in Boundary Value Problems*, Berlin: Springer-Verlag.

Särndal, C.-E., Swensson, B., Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Schlesinger, M. E. (1989), in A. Berger, R. E. Dickinson, J. W. Kidson (Eds.), *Understanding Climate Change*, Washington, DC: American Geophysical Union, pp. 177–187.

Schneider, S. H., Dickinson, R. E. (1974), *Rev. Geophys. Space Phys.* **12**, 447–493.

Schumaker, L. L. (1980), *Spline Functions: Basic Theory*, New York: Wiley-Interscience.

Schuster, H. G. (1988), *Deterministic Chaos*, 2nd ed., Weinheim, Germany: VCH.

Sergin, V. Ya. (1980), *Science* **209**, 1477–1482.

Sharkovsky, A. N. (1964), *Ukr. Math. Z.* **16**, 61–72.

Shertzer, J. (1989), *Phys. Rev. A* **39**, 3833–3835.

Smale, S. (1967), *Bull. Am. Math. Soc.* **73**, 747–817.

Sreenivasan, K. R. (1991), *Annu. Rev. Fluid Mech.* **23**, 539–600.

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer, J. D. (1992), *Physica D* **58**, 77–94.

Tournier, E. (Ed.) (1991), *Computer Algebra and Differential Equations*, New York: Academic Press.

Tufillaro, N. B., Abbott, T., Reilly, J. (1992), *An Experimental Approach to Nonlinear Dynamics and Chaos*, Reading, MA: Addison-Wesley.

Umrigar, C. J. (1989), *Int. J. Quantum Chem. Symp.* **23**, 217–230.

Vallis, G. K. (1988), *J. Geophys. Res.* **93**, 13979–13991.

Vautard, R., Ghil, M. (1989), *Physica D* **35**, 395–424.

Vautard, R., Yiou, P., Ghil, M. (1992), *Physica D* **58**, 95–126.

Vergassola, M., Frisch, U. (1991), *Physica D* **54**, 58–64.

Vicsek, T. (1992), *Fractal Growth Phenomena*, Singapore: World Scientific.

Weaver, A. J., Marotzke, J., Cummins, P. F., Sarachik, E. S. (1993), *J. Phys. Oceanogr.* **23**, 39–60.

West, B. J., Shlesinger, M. (1990), *Am. Sci.* **78**, 40–45.

Wolfram, S. (Ed.) (1986), *Theory and Applications of Cellular Automata*, Singapore: World Scientific.

Yidana, P. V. A., Hartt, K. (1988), *J. Comp. Phys.* **78**, 481–492.

Yonezawa, F. (1993), *Science* **260**, 635–640.

Yates, F. E. (1987), *Self-Organizing Systems*, New York: Plenum Press.

Zurada, J. M. (1992), *Artificial Neural Systems*, St. Paul: West.

Further Reading

Borchers, R. R. (Ed.) (1990), *Computers in Physics—Special Issue: Modeling the Environment*, vol. 3 (4).

Casti, J. L. (1992), *Reality Rules: I and II*, New York: Wiley.

Gould, H., Tobochnik, J. (1988), *Computer Simulation Methods, Parts 1 and 2*, Reading, MA: Addison-Wesley.

Schroeder, M. (1991), *Fractals, Chaos, Power Laws*, New York: Freeman.

MBE

See MOLECULAR BEAM EPITAXY

6

Measurement and Instrumentation of Flow

Francis R. Ruppel and James E. Hardy

6.1 Introduction 159
6.2 Classification, Principle of Operation, Applications, and Advantages and Disadvantages of Flowmeters 160
6.2.1 Differential Producing 160
6.2.2 Volumetric 166
6.2.2.1 Vortex Shedding 166
6.2.2.2 Magnetic 169
6.2.2.3 Turbine 171
6.2.3 Positive Displacement 172
6.2.4 Drag 173
6.2.5 Variable Area 174
6.2.6 Ultrasonic 176
6.2.7 Mass Flow 178
6.2.8 Open Channel 180
6.2.9 Anemometers 181
6.2.10 MEMS Flowmeters 182
6.2.11 Summary of Flowmeters 183
6.3 Flow Measurement Uncertainties 183
6.3.1 Installation Effects 185
6.3.2 Fluid Properties 185
6.3.3 Velocity Profiles 186
6.4 Calibration Methods 187
6.4.1 Time-Collection Techniques 187
6.4.1.1 Monitored Volume Methods 187
6.4.1.2 Monitored Weight Methods 188
6.4.2 Transfer Standards 188
6.4.3 Metered Volume Methods 189
Glossary 190
References 191
Further Reading 193

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

6.1
Introduction

The measurement of flow is one of the most common requirements for processes in research and industry. In the development of new processes or products, the flow rate of the various ingredients is needed to assess the impact of mixture ratios on product or process quality. In industry, once a particular process has been ''set up,'' the recipe must be controllable and repeatable, which often requires the measurement and control of flow. In energy usage and custody transfer, the accurate measurement of flow is necessary to track usage trends and customer energy consumption for billing purposes. Instrumenting pipelines and vessels with flow sensors is often required to ensure process, equipment, or personnel safety or to monitor for leaks or infiltration.

Because flow is so commonly measured, a wide variety of flowmeter types based on a variety of measurement principles are available. A universal flowmeter does not exist. Each type has its own advantages and limitations. Careful consideration of the measurement criteria must be given to each application before deciding what type of flowmeter should be used.

As in all areas of instrumentation, advances in microelectronics have significantly improved flowmeter accuracy, ease of use, maintainability, and cost of ownership. ''Smart'' flow sensors are becoming more popular, extending the range and accuracy of flowmetering. The predominant use of flowmeters is in the measurement of single-phase fluids in closed systems (pipes or ducts) and in open channels. In this article, the use of flowmeters for these two areas is discussed. The measurement of the flow of solids or multiphase flow is not considered here. The basic principle of operation of various flowmeters is discussed along with some applications and major advantages and limitations. A few flow velocity sensors are also described. Finally, some parameters that affect flow measurement uncertainty as well as calibration techniques that may quantify some of the uncertainties are briefly outlined. This article is intended to give an overview of flow instrumentation but is not all inclusive. Most popular meters as well as a few recent innovations are discussed.

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

6.2
Classification, Principle of Operation, Applications, and Advantages and Disadvantages of Flowmeters

6.2.1
Differential Producing

Bernoulli's equation for a nonviscous fluid shows that the pressure P, fluid density ρ, elevation z, and velocity v at two points are related as follows, where g_c = gravitational constant:

$$\frac{v_1^2}{2} + g_c z_1 + \frac{P_1}{\rho_1} = \frac{v_2^2}{2} + g_c z_2 + \frac{P_2}{\rho_2} \quad (6.1)$$

This equation is the basis for all differential-producing flow meters. In different hardware configurations, a differential-producing flow meter uses available measurements of differential pressure at a minimum and may measure temperature and static pressure as well for compressible fluids, where the density is not constant.

The first hardware configuration to consider is the Pitot tube as shown in Figure 6.1. In this configuration, there are two pressure measurements, the impact pressure of the fluid stream (also referred to as *velocity pressure*), indicated as P_2. The second measurement is the static pressure at the back side of the insertion tube, indicated as P_1. When measuring gas flow rate, the pressure difference between the two measurement points for Pitot tubes is usually a draft-range difference, meaning only a few centimeters of water column. Therefore, the density at the two points can be considered to be the same. For liquid measurements, the fluid is considered to be incompressible, so the density at the two points is the same. Also, because the elevation at the two measurement points are the same, and the velocity at the impact point is zero, Equation (6.1) can be simplified and rearranged as shown in Equation (6.2)

$$v_1 = \sqrt{\frac{2(P_2 - P_1)}{\rho}} \quad (6.2)$$

The velocity of the fluid is related to the differential pressure between the pressure taps. The velocity can be converted to volumetric flow rate q_V by using the cross-sectional area A of the pipe.

$$q_V = vA \quad (6.3)$$

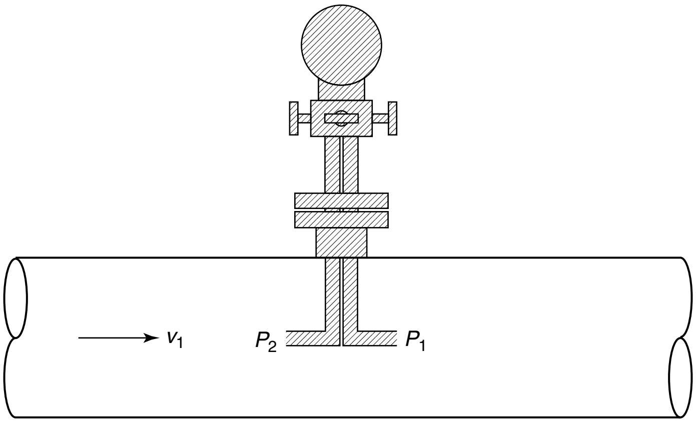

Fig. 6.1 Pitot tube flowmeter with fluid velocity v_1, impact pressure P_2, and static pressure P_1.

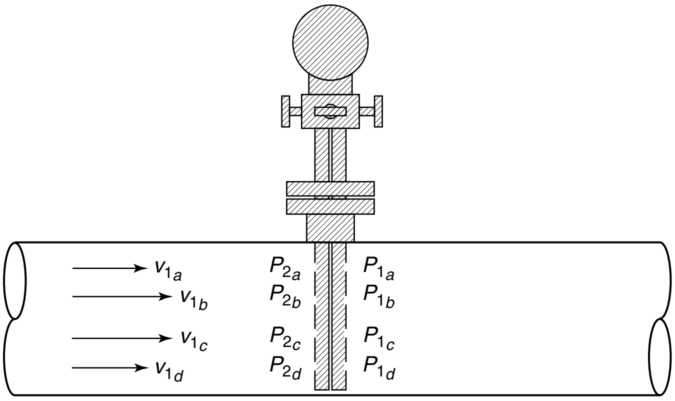

Fig. 6.2 Averaging Pitot flowmeter with fluid velocity v_1, impact pressure P_2, and static pressure P_1. Subscripts refer to points along the cross section of pipe.

Because the velocity pressure measurement is measured at a single point, it is sensitive to the velocity flow profile of the fluid. For example, the velocity at the middle of the pipe is not necessarily the same as the velocity at one-third of the distance into the pipe. Because of this, the inaccuracy is probably not better than ±5%. Another configuration of the Pitot tube is the averaging Pitot tube, as shown in Figure 6.2. This type of flow meter has several openings in each of the upstream and downstream tubes that measure velocity pressure and static pressure. The overall effect is that a single pressure measurement is taken in each tube, which represents the averaged pressure measurement. Some vendors with advanced, aerodynamic designs for the measuring tubes claim inaccuracies on the order of ±1% for averaging Pitot flow meters. Both designs should be used in applications with Reynolds number of 20 000 or more and have turndown ratios on the order of 4 : 1. A disadvantage of Pitot-type flowmeters versus other differential-producing flowmeters to be discussed in this section is that the developed differential pressure is a function only of the fluid parameters and diameter of the pipe line. For gas applications, it is common that very low differential pressure is produced, making for a challenging measurement. In contrast, for an orifice plate for example, it is always possible to reduce the diameter of the orifice to produce more differential pressure.

The second hardware configuration to consider is the orifice plate. A typical cross-sectional view of flow through an orifice and the corresponding pressure distribution are shown in Figure 6.3. The fluid stream contracts as it passes through a constriction. The velocity at each point is assumed to be inversely proportional to the cross-sectional area of flow at that point (not necessarily the same as the cross-sectional area of the pipe or orifice). The flow area contracts to a minimum at the vena contracta, the point of smallest flow cross section. The kinetic energy of the stream increases as it contracts, and the potential energy (pressure) decreases. Therefore, a pressure difference is created and can be measured. The relationship is more complicated with compressible flow because factors that affect the internal energy of the fluid, such as the density, compressibility, and

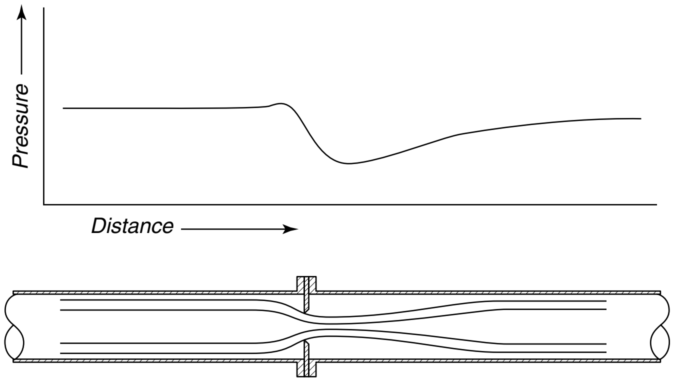

Fig. 6.3 Cross-sectional view of flow through an orifice.

isentropic exponent, have to be taken into account. However, the net result is the same: by measuring differential pressure across an engineered flow restriction, flow rate can be determined as a function of the square root of differential pressure when the cross-sectional flow area of each point is known.

The velocity at two points is related to the cross-sectional area as follows, where A_1 and A_2 are the cross-sectional areas at each point:

$$v_2 = v_1 \frac{A_1}{A_2} \tag{6.4}$$

By rearranging Equation (6.1) and using the relationship in Equation (6.4), the relationship of flow velocity for a constant density situation is given as follows:

$$v_1 = \sqrt{\frac{2}{1-\left(\frac{A_1}{A_2}\right)^2}}\sqrt{\frac{\Delta P}{\rho}} \tag{6.5}$$

By knowing the measurements of the pipe and orifice cross-sectional area and the density, the flow velocity is a function of the measured differential pressure across the orifice. For incompressible fluids, Equation (6.5) represents the actual flow fairly well, except that the cross-sectional area in the equation is of the fluid stream, not of the pipe and the orifice. Many empirical relations have been made to relate flow rate using dimensions of the pipe and orifice to actual flow rate. The relations are a function of where the pressure measuring taps are located. A term called the *discharge coefficient* C is added to the equation. The *discharge coefficient* is defined as the ratio of actual flow rate to theoretical flow rate. The discharge coefficient corrects the theoretical equation for velocity profile, actual energy loss between pressure taps for viscous fluids, and for the particular pressure tap location. ISO 5167 defines the discharge coefficient as a function of beta ratio β (the ratio of orifice diameter to pipe internal diameter D), Reynolds number Re, distance from upstream pressure tap to upstream face of the plate l_1, and distance from downstream pressure tap to downstream face of the plate l'_2 as follows (ISO, 2003):

$$\begin{aligned} C = {} & 0.5959 + 0.0312\beta^{2.1} - 0.184\beta^8 \\ & + 0.0029\beta^{2.5}\left(\frac{10^6}{Re}\right)^{0.75} \\ & + 0.090\frac{l_1}{D}\frac{\beta^4}{(1-\beta^4)} - 0.033\frac{l'_2}{D}\beta^3 \end{aligned} \tag{6.6}$$

For compressible flow, Equation (6.5) has to be modified to account for the density at each measuring point. The density at the upstream measuring tap can be calculated from the measurements of pressure and temperature at that point. It is necessary to determine the density at the downstream measuring point.

The expansion of the gas is a constant entropy process. A gas expansion factor Y_1 can be derived to determine density at the downstream point, where x_1 is the differential pressure divided by the upstream pressure, and k is the isentropic exponent (ISO, 2003).

$$Y_1 = 1 - (0.41 + 0.35\beta^4)\frac{x_1}{k} \tag{6.7}$$

The flow equation with the discharge coefficient and gas expansion factor is then expressed as shown in the following equation, where d is the orifice diameter, ρ_{f1} is the upstream fluid density, and the mass flow rate q_m is used for convenience.

$$q_m = \frac{\pi}{4}\frac{CY_1 d^2}{\sqrt{1-\beta^4}}\sqrt{2\Delta P \rho_{f_1}} \tag{6.8}$$

Therefore, it becomes necessary to iteratively solve Equation (6.8), because the downstream pressure and downstream density are related to each other, and the equation cannot be solved outright. If Equation (6.8) is rearranged to solve for the differential pressure, the following equation results.

$$\Delta P = \frac{8q_m^2(1-\beta^4)}{(\pi CY_1 d^2)^2 \rho_{f_1}} \tag{6.9}$$

From Equation (6.9), an invariant is defined as follows:

$$Inv = \frac{8q_m^2(1-\beta^4)}{(\pi C d^2)^2 \rho_{f_1}} \tag{6.10}$$

The differential pressure is then as follows:

$$X = \Delta P = \frac{Inv}{Y_1^2} \tag{6.11}$$

A residual δ is defined in the following equation:

$$\delta = X - \frac{Inv}{Y_1^2} \tag{6.12}$$

The iteration calculation is as follows, where X_n is the value of the calculation at the current iteration, X_{n-1} is the value of the calculation at the last iteration, and X_{n-2} is the value of the calculation two iterations before:

$$X_n = X_{n-1} - \delta_{n-1}\frac{X_{n-1} - X_{n-2}}{\delta_{n-1} - \delta_{n-2}} \tag{6.13}$$

For the first calculation, it is necessary to "seed" the calculation with estimated values for the past calculations. A typical starting seed for the gas expansion factor is 1. For each iterative calculation, X_n, the residual is calculated. The iteration continues until the residual is extremely small. The iteration will need to be seeded for the first two iterations, because past values of the residual are not available. The coefficient is strongly dependent on the Reynolds number *Re* (see Section 6.14).

As we see in Section 6.3.3, the determination of flow rate by differential pressure measurement is heavily influenced by the flow profile. This is because skewed or swirling flow causes differential pressure errors because of the different velocity vector at the point of measurement as compared to the average velocity in the measurement plane. We have also seen that empirical equations are necessary for differential pressure flowmeters because the vena contracta of the fluid flow does not occur at a plane in the fluid flow that has a corresponding physical dimensional measurement. In other words, the vena contracta does not occur in the plane of the orifice.

A relatively new type of differential-producing flow meter reduces the effects of the two parameters that are discussed in the last paragraph by shaping the flow profile in a controllable way. The cone-type flow meter is shown in Figure 6.4. By shaping the flow profile around the cone, the downstream differential pressure measurement is made in a known annular cross-sectional area. This is not possible with an orifice plate, because it is not possible to place a pressure tap in the exact same plane of the orifice. Also, by ‘‘shaping’’ the flow by mixing the low and high velocities to the downstream portion of the cone, a more consistent flow profile is produced (Ifft and Mikkelsen, 1993). Similar to the effect in orifice plates, the consistency of the shaped flow profile is a function of the meter’s beta ratio (see Section 6.3.3). Comparison testing against other conventional flowmeters has shown the superior accuracy of the cone-type meters (Lloyd, Guthrie, and Peters, 2002; Dahlstrom, 1994). Independent tests have also shown the meter’s ability to deal with swirling flow (Shen, Bosio, and Larsen, 1995). The cone-type flowmeter has an advertised inaccuracy of ±0.5% of flow rate, a turndown ratio of 10 : 1, straight line requirements between 0–3 pipe diameters upstream and 0–1 pipe diameters downstream, and can measure fluids with a minimum Reynolds number of 10 000. Because the manufacturing tolerances are not as good as, say a simple orifice, it is advised to have each meter calibrated in the factory before being placed in service.

Perhaps, the most important hydraulic parameter to flow metrology is the dimensionless *Re*. It is the ratio of inertia forces to viscous forces in a flow system as described by

$$Re = \frac{\rho v D}{\mu} \tag{6.14}$$

where ρ = fluid density, v = fluid velocity, D = characteristic dimension of the flow (often called *hydraulic diameter*), μ = absolute viscosity of fluid.

Dissimilar flow systems can be compared by the use of the *Re*. The flow condition in a system can be described as laminar, transitional, or turbulent. The magnitude of the *Re* indicates the type of flow condition present. For pipeline flows, laminar flow exists if the *Re* is less than 2000. Transitional flow occurs for the *Re* range of 2000–10 000, and turbulent flow exists for *Re* over 10 000. The advantage of using the nondimensional *Re* is that the above ranges hold for any gas, vapor, or liquid. Most flowmeters function in a particular flow condition; thus, the

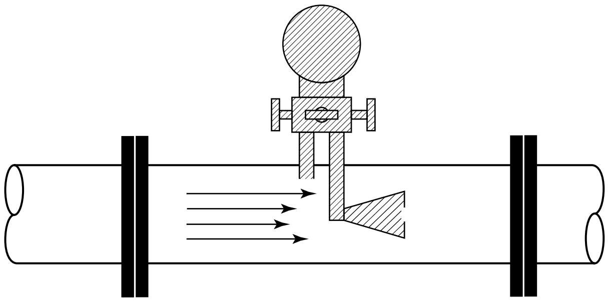

Fig. 6.4 Cone-type flowmeter.

Re should be calculated with the appropriate characteristic dimension for each case. When thermal properties are important, the Prandtl number is also a dimensionless number that can be very useful in characterizing dissimilar flow systems.

By far, the most predominant flowmeter in existence by number, the differential-producing-type flowmeter is popular because the technology of pressure and differential pressure sensing is quite advanced in comparison to other measurements and because there are many empirical relationships to account for the difference between theory and practice. The concentric orifice has been the workhorse in industrial applications. Other differential producers include the integral orifice, the venturi tube, and the simple pipe elbow with pressure taps upstream and downstream. The integral orifice meter combines in one assembly an orifice and pressure taps along with a honed pipe section in which the inner surface has been deburred and smoothed. It is also common to include the differential pressure transmitter with the assembly. The venturi-tube meter uses a gradual constriction and expansion of the flow stream instead of an orifice plate. Its advantages over the orifice plate are that the downstream pressure recovers closer to the upstream pressure, thereby causing less energy loss, and it is also less susceptible to erosion when used with solids-bearing fluids. The pipe elbow meter is very simple but not very accurate.

Highly sophisticated computer programs have been developed for the orifice meter to accurately determine flow on the basis of differential pressure, flowing temperature, absolute pressure, dimensions of the orifice and pipeline, thermal expansion factor of orifice material, tap location, and fluid viscosity. For the custody transfer of valuable fluids, these types of calculations are carried out to several decimal points – more resolution than can possibly be justified. However, on a contractual basis, these are the accepted calculations regardless of the inaccuracy. Many configurations exist for the orifice meter. Pressure tap locations vary in their distance upstream and downstream of the orifice plate but usually fall into one of several standard configurations. The most common orifice plate type is square-edged concentric, but quadrant- and conical-edged eccentric and segmental orifices exist as well.

A big disadvantage of differential-producing devices has been that because of the square-root relationship between flow and differential pressure, an error in differential pressure, especially at the low end of the range, becomes a magnified error in the flow reading. For example, in Figure 6.5, a 2% error in differential pressure at 60% of full-scale differential pressure causes a 1.3% of full-scale error in flow. However, the same differential pressure error at 3% of full-scale differential pressure causes a 5% of full-scale error in flow. The term *turndown ratio* (or *rangeability*) refers to the maximum feasible reading of a flow device divided by its minimum feasible reading. Differential producers, in the past, were typically limited to a turndown ratio of three. With improved accuracy and repeatability of modern pressure measurement devices, and on-line determination of the Reynolds number and gas expansion factor as discussed in Section 6.3.2, the turndown ratio is claimed to have been increased to 10. Inaccuracies vary from 0.25 to 2% of full scale. Flow ranges vary from very high for industrial applications to somewhat low for research and development applications. However, the flow must remain turbulent (the Reynolds number $>10\,000$).

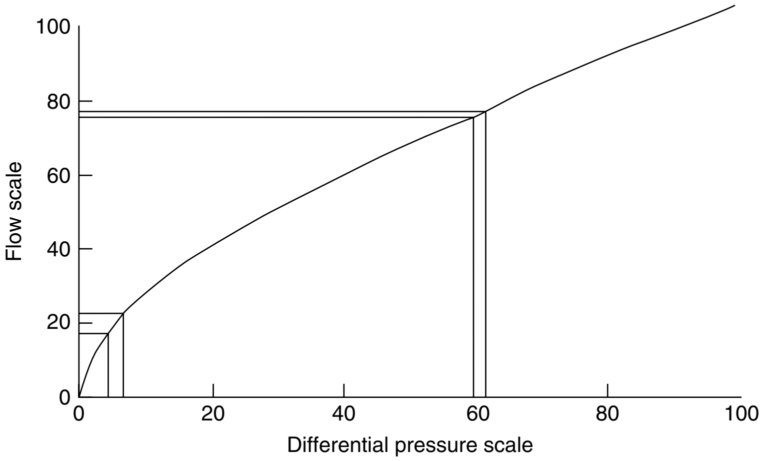

Fig. 6.5 Typical flow versus differential pressure curve for an orifice, arbitrary flow, and differential pressure units.

Straight-pipe requirements, which ensure a symmetric velocity profile through the meter (see Section 6.3.3), vary depending on the type of flow obstruction and the beta ratio. The range of straight-pipe requirements usually falls within 5–50 pipe diameters upstream and 2–4 pipe diameters downstream of a pitot tube or orifice, but much less for a cone-type meter.

6.2.2
Volumetric

For this class of meters, volumetric flow rate q_V is generally inferred by the fluid flow velocity v multiplied by the meter cross-sectional area A (see Equation (6.3)). Because the meter reading is actually based on fluid velocity, care must be taken to ensure that the entire meter cross-sectional area is filled with fluid; otherwise, errors will occur in the volumetric flow rate.

6.2.2.1 **Vortex Shedding**

The principle of the vortex-shedding flowmeter, as shown in Figure 6.6, is that a blunt, nonstreamlined object perpendicular to the direction of flow sheds vortices alternately from its sides when flow attains a high enough Reynolds number. The frequency that vortices are formed on alternate sides of a bluff body exposed to flow is proportional to the velocity of the flow around the bluff body. Vortices roll into

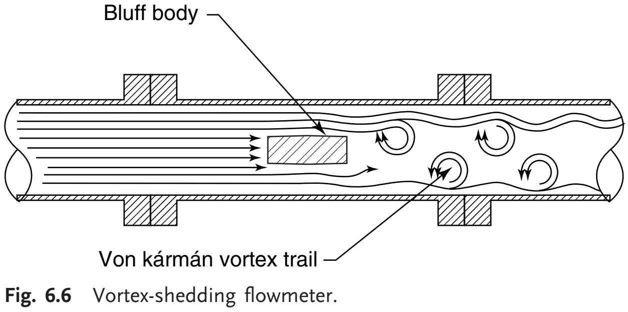

Fig. 6.6 Vortex-shedding flowmeter.

low-pressure areas behind the bluff body. As the vortex is shed on one side of the bluff body, fluid velocity increases, which causes a pressure decrease on that side. On the opposite side of the bluff body, the velocity decreases and the pressure increases – there is a net pressure change across the bluff body. The velocity and pressure change at the same frequency as the vortex-shedding frequency. The Strouhal number Sr is a dimensionless number that relates shedding frequency f, a characteristic dimension of the bluff body l, and the fluid velocity v,

$$Sr = \frac{fl}{v} \tag{6.15}$$

The vortices downstream of the bluff body are known as the *Von Kármán vortex trail*. The Strouhal number is essentially independent of the Reynolds number and process fluid for Reynolds numbers greater than 20 000. Below that, the relationship between frequency and flow velocity is nonlinear. At Reynolds numbers below 4000 or 5000, the stability of the vortex-shedding degrades. Vortex-shedding flowmeters are not generally applicable for Reynolds numbers less than 5000 (Mattar, 2004).

The relationship between shedding frequency and flow rate is a function of only the dimensions of the bluff body – Vortex shedding is not a function of fluid properties. Expansion of the bluff body dimensions can cause errors. Installation requirements to meet upstream straight-pipe requirements range from 10–20 diameters upstream and 5 diameters downstream to 25–45 diameters upstream and 8 diameters downstream. Fluids with viscosities lower than 8 cP do not work well. The flow signal will have a dropout below the minimum Reynolds number. Also, because the sensing methods are passive, a minimum amount of energy must also be available in the fluid. This is expressed as the minimum velocity constraint. Usually 30–60 cm s^{-1} for liquids, but greatly varies for gases (Spitzer, 2005). The combination of the minimum Reynolds number and the minimum velocity constraint causes the flow reading to fall to zero; therefore, low flows cannot be detected. However, this can also be an advantage as compared with the orifice meter. With the orifice meter, small differential pressure errors can falsely indicate flow at near zero differential pressure. In applications where flow is being totalized, this error can lead to large totalization errors over long durations of low or no flow when there is an error in the differential pressure reading. For the vortex-shedding flowmeter, the flow signal simply falls to zero, and no flow totalization occurs. If the meter is correctly sized for the minimum anticipated flow, having the flow signal drop to zero for flow rates below the minimum will cause no substantial errors in flow totalization. On the negative side, if the flow signal is used in a control loop, dropout can cause control loop instability. For example, consider the case where a control valve is being used to control flow rate. When the flow signal drops out, the controller interprets this as zero flow and opens the control valve farther. When it surpasses the dropout point, the flow meter will read flow again. However, at this time, the flow is higher than needed, so the controller will close the valve farther. The opening and closing of the control valve will cause a condition known as *limit cycling*. The meter can be sensitive to vibration. In the past, passive techniques with fixed range band pass filters have been used to overcome vibration effects. Recent efforts are using digital signal processing to dynamically adjust the band pass filters to overcome the vibration effects (Ostling and Oki, 2001).

For calibration, a disadvantage is that the meter cannot be verified in-line. For maintenance, the flowmeter is fixed in line and cannot be repaired without opening up the pipeline (as opposed to differential pressure meters that use transmitters that are mounted remotely from the line). However, some designs allow replacement of piezo sensors while the meter is in-line. On the plus side, a "sealed" meter can be an advantage to reduce fugitive emissions. There are redundant configurations available to increase the safety integrity level of the measurement (Mattar, 2004). This is important when the measurement is critical to the safety of personnel or the integrity of the process.

A typical vortex-shedding flowmeter consists of a pipe section, a bluff body in the center of the pipe section, a means to detect the vortex frequency, and power and signal conditioning. The measurement can be made in different ways depending on what is to be measured:

- an oscillating disc measured by a magnetic sensor;
- the pivot of the shedder measured by a piezo element;
- alternating pressure;
- fluid flow through a passage in the bluff body that alternates back and forth, as measured by a temperature sensor;
- the twist of a torque tube measured by a piezo element; and
- downstream vortices measured ultrasonically.

The meter has an inaccuracy of around 0.5–1% of flow rate for liquids, and 1.5–2% of flow rate for gases. Turndown ratios of up to 40 : 1 are claimed by manufacturers. A word of caution is in order for these claims, though – they are stated relative to the maximum flow possible for the size of pipe in which the flowmeter is installed. This is almost always higher than the maximum design flow expected in an application, because pipe sizing criteria is usually conservative with regard to the pipe diameter. The net effect is that turndown for an application, defined as maximum design flow rate divided by minimum flow rate readable by the flow meter is much lower than 40 : 1. A common practice to improve turndown has been to decrease the pipe diameter for the meter run only. New vortex flow meters are available in reduced body sizes, that is, they are connected to the large diameter pipe, but have decreased cross-sectional area within the meter to increase velocity and improve turndown. However, data on the line-size reduction and expansion at the meter coupled with not having straight runs before and after reduced bore meters is not mature yet. The meter can be used for liquids, vapors, and gases; however, to obtain the minimum Reynolds number, gas flows must be at relatively high pressure. It is quite common to use these meters to measure steam flow rate. Care must be taken to avoid cavitation of liquid flow by keeping the pressure at the vortices above the vapor pressure of the fluid. The performance can be affected by the thermal expansion of the bluff body. In applications that are susceptible to condensation, the bluff body should not be on the bottom of the line (Spitzer, 2005).

Some claims have been made that pressure and temperature compensation is not required for compressible flow. However, this would assume that an output in actual volumetric flow or velocity is desired. In practice, most users are interested in volumetric flow that is expressed in "standard" (US) or "normal" (Europe) flow units, which are actually mass flow rates. There are configurations available with pressure and temperature sensors that allow converting to these flow units (Mattar, 2004).

6.2.2.2 Magnetic

For conductive fluids, the magnetic flowmeter is a very reliable flow device. The meter is based upon Faraday's law. Voltage V will be induced in a conductor that is moving through a magnetic field proportional to the length of the conductor l (the diameter of the pipe in this case), the velocity of the conductor v, and the strength of the magnetic field B,

$$V = vBl \tag{6.16}$$

The induced voltage is perpendicular to both the fluid flow direction and the magnetic field. By knowing the area of the flowing cross section, a volumetric flow is calculated once the velocity is known as shown in Equation (6.3).

A typical magnetic flowmeter is shown in Figure 6.7. Electromagnets surrounding the flow tube generate a magnetic field within it. Process fluid passing through the magnetic field acts as a moving conductor, inducing voltage in the fluid. Flush-mounted electrodes inside the primary tube pick up the voltage. A nonconducting pipe-liner section is required. The liner is typically made from a polymer, rubber, or ceramic. The meter must be electrically bonded to earth ground to prevent stray currents from flowing through the meter and causing a zero shift. For nonconductive pipe, grounding rings are required to be installed upstream and downstream of the meter when the conductivity of the fluid is less than 100 μs cm^{-1}; otherwise, grounding electrodes can be used. The grounding rings may have to be made of exotic materials when exposed to acidic or corrosive fluids and can be expensive. Grounding straps are acceptable for conductive, unlined pipe. Buildups of conductive solids on the electrodes may

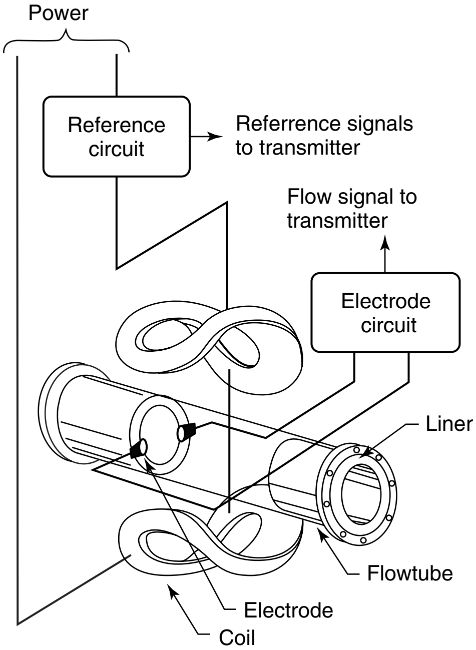

Fig. 6.7 Magnetic flowmeter.

form, which decrease the cross-sectional flow area, and therefore cause an error in the inferred volumetric flow rate. Almost all manufacturers of magnetic flow meters today incorporate means for optional electrode cleaning to overcome coating problems. The coating can be sensed by measuring the relaxation time of a pulse across the electrodes and ground (Walker, 2001).

Typically, the minimum conductivity of the fluid must be $1\,\mu\mathrm{s\,cm}^{-1}$. The temperature range of the fluid should be taken into account because the fluid conductivity is a function of temperature. The measured flow signal is not a function of the conductivity so long as the conductivity is above the minimum threshold level. Because this type flowmeter is relatively insensitive to velocity-profile effects, only ~3 pipe diameters of straight piping is required before the meter (Rose and Vass, 1993). Inaccuracies vary from ~0.2 to 1% of flow rate. Turndown ratio varies from ~10 : 1 to 30 : 1. Pipe sizes from 0.1- to 244-cm diameter can be accommodated. Two types of magnetic-field excitation are commonly used: ac and pulsed dc. The ac excitation methods were very common in the past, but pulsed dc excitation methods are mostly used now. The trade-offs between the two methods are that ac methods have good signal-to-noise ratios and are more stable in two-phase applications such as liquids with either entrained gas or high suspended-solids content. The dc-pulsed methods have a stable zero, are not as susceptible to coatings on the electrodes, are smaller and lighter, and can be made to be intrinsically safe in areas that are exposed to explosive gases and dusts because of their lower power. Steady dc is not used to avoid polarization of ions in the fluid. It is typical for an ac meter to have a zero drift of as much as 1% per month. To correct the drift, it is necessary to fill the meter with fluid, stop flow, and zero the meter. The pulsed dc units have typical drifts of only 0.1% per month. Meters powered with ac have a signal that includes components of flow and noise. Dc-pulsed meters can measure the noise during periods when the coil is not energized and subtract this from the signal during energized pulses to obtain better accuracy. The ac meter would require a nonflow zero calibration to achieve the same effect. Pulsed dc meters do not have as good a time response as ac because of time required for the pulsed technique. Two-wire meters are available that use energy from the 4 to 20 mA loop. However, due to the limited amount of energy, low- flow measurement is less accurate. Also the pulsing frequency is low, so response time is correspondingly slow. Meters designed to measure pipelines that are only partially full are available with built-in level measurement. The flow tube is generally made of stainless steel that does not affect the magnetic field, as would carbon steel. Applications are limited in vacuum service due to the possibility of the liner separating. Magnetic flowmeters are good choices for slurries because of their obstructionless design. Electrodes should be oriented in a horizontal plane to avoid being uncovered by gas bubbles (Spitzer, 2005). A rule of thumb for flow velocity through the meter is $60\text{–}600\,\mathrm{cm\,s}^{-1}$ for normal service, $90\text{–}300\,\mathrm{cm\,s}^{-1}$ for abrasive slurry, and $150\text{–}450\,\mathrm{cm\,s}^{-1}$ for nonabrasive slurry. Flow can be laminar, transitional, or turbulent. This meter can even measure bidirectional flow. Noncontact electrodes are becoming popular. Meters with noncontact probes can measure fluids down to $0.05\,\mu\mathrm{s\,cm}^{-1}$ but are very expensive. The obstructionless design of the noncontact probes is good for slurries and corrosives. Organic liquids, in general, do not have high enough conductivity to work.

6.2.2.3 **Turbine**

A turbine flowmeter consists of a bladed rotor suspended axially in a pipe section in the direction of flow such that flow through the pipe causes the rotor to rotate. Speed of rotation of the rotor, which is correlated to flow velocity, is measured by a pickup outside the pipe section. For a turbine blade with angle θ with respect to the pipe axis, as shown in Figure 6.8, the ideal turbine rotational velocity ω is related to the flowing velocity v and the rotor radius r by resolving vectors as

$$\omega = \frac{v}{r}\tan\theta \tag{6.17}$$

Therefore, by combining Equation (6.3) with Equation (6.17), the volumetric flow rate q is given by

$$q = rA\frac{\omega}{\tan\theta} = K\omega \tag{6.18}$$

K is the called the *K factor*. Although the K factor is derived theoretically here, it is used to characterize the meter on the basis of meter calibration tests (Sydenham, 1999). It provides a flow output based on the measured rotational speed of the turbine.

The K-factor curve is valid only above a minimum threshold in flow. Then it typically shows a peak, as shown in Figure 6.9, before settling out at a fairly constant value. The K factor that is used in practice is usually a constant over the whole operating range of the meter, because the range of actual K-factor values from a high at the top of the peak to the low at higher flow rates does not vary from an average K value more than would cause the meter to exceed its tolerance specification (Lipták, 2003). However, with more and more microprocessor-based instruments being built and benefiting from smaller component size and surface-mount technology, it is now possible to store the K-factors in electrically erasable, programmable, read-only

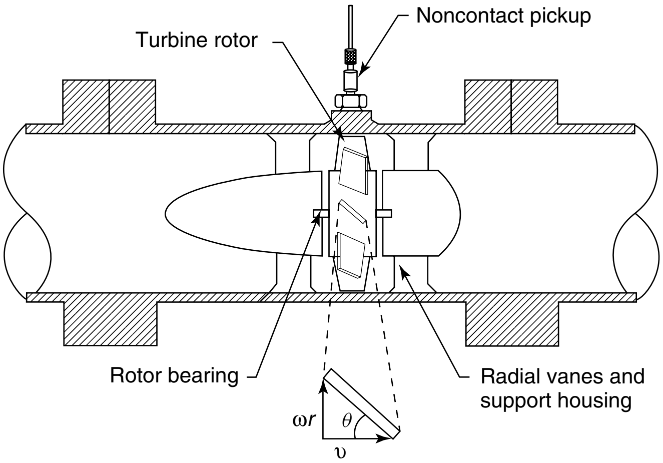

Fig. 6.8 Turbine flowmeter arrangement with turbine rotational velocity ω, rotor radius r, turbine blade with angle θ with respect to the pipe axis, and fluid velocity v.

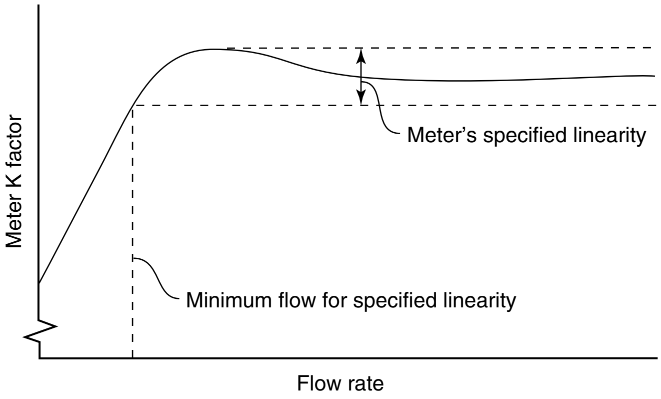

Fig. 6.9 Typical *K*-factor curve for a turbine flowmeter.

memory chips within the manufacturer's signal conditioning units (McCoy, 1992).

The relationship between fluid velocity and rotor speed is linear over a wide range of flow conditions. Retarding forces cause the rotor to spin at a speed less than theoretical. This effect is called *slip*. Factors influencing the *K* factor are changes in rotor-blade angle (due to collisions with solids), swirl and velocity-profile distortion, rotor-bearing friction, viscous drag, and fluid density. These factors can combine to cause errors from a few tenths of a percent to 15% (Miller, 1996).

Two types of noncontact pickups are used: reluctance and inductance. These types of meters can be used for liquids, vapors, and gases. Low-cost units have inaccuracies of ~1% of flow rate with a 10 : 1 turndown ratio; high-cost units have inaccuracies of 0.1–0.5% of flow rate with turndown of 10 : 1 to 50 : 1. The volumetric flow is measured at the operating temperature. It may be necessary to correct for temperature effects (Lipták, 2003).

Primary uses include flow totalization (by counting pulses from the pickups) and inventory control. Frequency-to-analog signal conditioning must be used to convert to standard analog signals. Sizes range from 1.25 to 30 cm. End connections include threaded, male flared tube, flanged, and plain for slip-on couplings (McMillan and Considine, 1999).

Changes in fluid viscosity affect the rotor speed-to-flow velocity relationship. The use of "universal viscosity curves" that plot flowmeter sensitivity versus Reynolds number allow the combined effects of velocity, density, and absolute viscosity to be taken into account (Miller, 1996). Turbine flowmeters can compensate for viscosity by knowing the fluid's viscosity versus temperature curve. The temperature of the fluid is measured and is fed into a microprocessor-based signal conditioner (Madison, 1995).

6.2.3
Positive Displacement

When all applications of flowmeters are considered, positive-displacement meters by far outnumber all others because of the almost exclusive use of positive-displacement meters for household gas and water metering and for the metering of motor fuels. The positive-displacement meter, in these applications, is not actually a flowmeter but a volume meter. The fundamental positive-displacement meter has no time base; but for industrial applications where a flow rate is desired to be monitored, the positive-displacement meter can be outfitted to provide flow readings. Otherwise, the receiving instrument can be made to compare volume readings over time to provide flow readings.

Positive-displacement meters work on the principle that successive chambers within the meter are filled with the metered fluid. As one chamber is emptied, the successive chamber is filled. Each chamber holds a fixed volume of fluid. The filling and emptying is translated into a rotary motion. The speed of rotation is converted to pulses. The pulses are sensed optically or magnetically. The pulses can be totalized, or the frequency can be measured for flow rate. Any gas will be treated as "volume" and thus cause errors in liquid flow rate indications that contain gas bubbles. Because positive-displacement meters measure volume, the pressure and temperature must be taken into account to arrive at a mass exchange value. There are no Reynolds numbers or straight-pipe constraints. Viscosity changes cause an error (which is function of temperature). There can be slippage at low viscosity. Liquids that do not lubricate the rotor bearings can cause errors at low flow rates. The following are the different types available (Spitzer, 2005):

- *Helical gear*: These are usually used for extremely viscous service, up to 300 000 cP. They have an inaccuracy of ~0.4% of rate. This style causes relatively little differential pressure for high viscosity fluids as compared to other types.
- *Nutating disc*: These are often used for water service where inaccuracy (~2% of rate) is not a primary concern. Their turndown ratio is on the order of 5 : 1 to 10 : 1.
- *Oscillating piston*: These are used for viscous liquid service where turndown is not a concern. They have 0.5% of rate inaccuracy, turndown of 3 : 1 to 4 : 1, and can work with a viscosity range of 0.2–10 000 cP.
- *Oval gear (rotor)*: These are used for viscous service where other types of flow meters have Reynolds number constraints. They are available with an inaccuracy of 0.25–1% of rate, turndown up to 20 : 1 for intermittent service, but as low as 3 : 1 for continuous service. The viscosity range is between 0.2 and 500 000 cP.
- *Piston*: These are used in low-flow viscous liquid service where other types of meters would have Reynolds number constraints. They have inaccuracy of 0.5–1% of rate. The turndown can be as high as 100 : 1, and the viscosity range is 0.5–10 000 cP.
- *Rotary*: These are used for liquid service where accuracy is important but turndown is not. They are available with 0.1–0.2% of flow rate inaccuracy, turndown of 5 : 1 to 10 : 1, and viscosity ranges of 1–25 000 cP.

6.2.4
Drag

The drag force F_D exerted on an object by a fluid is a function of the object's projected area A_p, the drag coefficient C_D, and the density of the fluid ρ:

$$F_D = \frac{1}{2} A_P C_D \rho v^2 \qquad (6.19)$$

The drag coefficient is a function of the fluid's Reynolds number and a shape factor associated with the object. For disks with

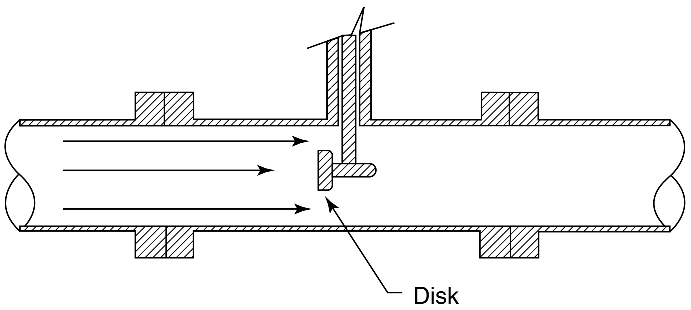

Fig. 6.10 Drag flowmeter.

the flat side perpendicular to flow, the drag coefficient is nearly a constant for Reynolds numbers higher than ~4000.

Flowmeters use the principle by exposing a disk to flow as shown in Figure 6.10. The force exerted on the disk is measured by force balance or strain gauge methods. By taking the square root of the force signal, a signal proportional to flow velocity is obtained. Volumetric flow rate is inferred from the velocity and the cross-sectional area of the pipe line as shown in Equation (6.3).

The use of a drag meter is good for liquids, gases, steam, and slurries. Inaccuracy usually lies between 0.5 and 2% of full scale. Absolute inaccuracy increases as flow decreases because of the square-root relationship (similar to the orifice meter). Turndown ratios of 10 : 1 to 15 : 1 can be obtained. The drag force depends on the geometry of the target. Erosion of or buildups on the disk will cause errors. The meter requires about the same straight-pipe lengths as orifices. The meter is usually used for reliability, speed of response, repeatability, and usable turndown ratio. It is not known for good accuracy. This design is economically attractive for large pipe sizes using insertion-style designs, because flanges or pipe wafers, that can be very expensive in large diameters, are not required. Pipe vibration can affect the measurement (Spitzer, 2005).

6.2.5
Variable Area

The variable-area flowmeter, or rotameter, is a device for measuring flow that maintains a relatively constant differential pressure across the meter by increasing the annular area around a float within the meter as the flow rate increases. As shown in Figure 6.11, the rotameter is oriented vertically. The float is situated inside a tapered-glass or -plastic tube. In the more common case, where the float is denser than the metered fluid, flow enters at the bottom of the tapered tube and exits at the top. It is common to have a needle valve in either the inlet or the outlet flow tube to control the flow rate. The buoyant forces, F_A, and upward hydraulic forces, F_D, of the fluid are balanced by the weight of the float, F_G.

$$F_G = F_A + F_D \tag{6.20}$$

The weight of the float is given by the following, where V_S is the volume of the float and ρ_S is the density of the float.

$$F_G = V_S \rho_S g_c \tag{6.21}$$

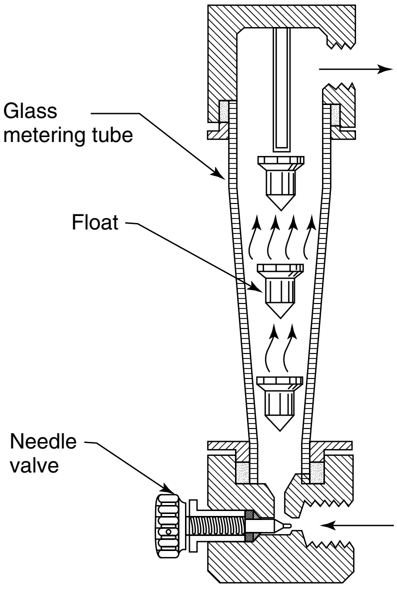

Fig. 6.11 Variable-area flowmeter.

The buoyancy of the float is given by the following, where ρ_m is the density of the fluid:

$$F_A = V_S \rho_m g_c \tag{6.22}$$

As we saw in the previous section, the upward hydraulic drag forces are represented in Equation (6.19). For these purposes, we will use ρ_m in the equation, for the density of the fluid. The volumetric flow rate through the meter is represented as in Equation (6.3), where A_r, the annular surface area between the float and the tube wall is substituted into the equation.

By combining Equations (6.3) and 9.19–22, we obtain a volumetric flow rate as follows:

$$q_V = A_r \sqrt{\frac{2g_c V_S(\rho_S - \rho_m)}{C_D A_P \rho_m}} \tag{6.23}$$

As the tube taper is a linear function of height, the annular area A_r is directly proportional to height. The other parameters are constant, and as we see from this equation, the volumetric flow rate is directly proportional to the height of the float.

The float is positioned in the tapered tube in proportion to fluid flow rate and annular area. As the float rises with increasing flow rate, the annular area around the float increases. As the annular area increases, the differential pressure decreases. The glass or plastic tube is marked with gradations for flow. When the rotameter is used properly, the flow rate is determined by noting the level of the float and the corresponding flow gradation on the tube.

The rotameter must be used with the design fluid and the operating parameters, or else a correction factor must be applied to the reading obtained. The variable-area flowmeter typically has no interface into data acquisition and control systems, although some manufacturers do offer an electrical interface, especially when metal flow tubes are used. In this case, the float position is measured magnetically or electrically. Output types are available with pneumatic, electric, or pulse output.

Variable-area flowmeters measure liquids with viscosities less than 3 cP, gases, and steam. Inaccuracies vary from 0.5 to 5% of flow rate, with a turndown ratio of ~10 : 1. The meter is usually limited to less than 8-cm pipe diameters, tends to be somewhat self-cleaning, and is relatively insensitive to piping configuration (no straight-pipe requirements).

Some rotameters use a spring that counteracts the fluid forces. It is not necessary for these to be oriented vertically. Ball floats are almost exclusively used for purge flow meters. For liquids, the minimum Reynolds number is 250–950 for low flows and 4400 for high flows. Gas flows have no minimum Reynolds number requirement, but must have high enough density to raise the float (Spitzer, 2005).

6.2.6
Ultrasonic

Ultrasonic flowmeters and flow sensors work on one of the two basic principles – Doppler shift or time of flight. The Doppler-shift method functions by transmitting an ultrasonic signal of known frequency into a flow stream (Figure 6.12). The ultrasonic beam reflects off particles or other discontinuities in the stream. The difference in frequency (Doppler shift) between the original beam and the reflected signal is related to the flow velocity of the reflecting medium. The argument made is that the small reflecting particles travel at the same velocity as the transporting fluid. Most commercially available Doppler-type flowmeters require particles or air bubbles with a minimum size of 25 – 30 μm and a concentration of at least 25 ppm. The ultrasonic transceivers (transmitter and receiver) can clamp onto the outside of a pipe; for better accuracy, they can be mounted through the pipe wall and flush with the inside wall (wetted transducer type). These devices can be used on liquids, aerated liquids, and slurries. The turndown ratio is at least 10 : 1 with an inaccuracy of ±1 – ±5% of full scale depending on whether clamp-on or wetted transducers are used and whether *in situ* calibration is possible. The time-of-flight or transit time meter operates by pulsing an ultrasonic beam across a pipe in the direction of flow and then in the reverse direction (Figure 6.13). The difference between the travel times in the two directions is related to the velocity of the flow. This relationship is described by the following equations:

$$V_{AB} = C_0 + V_m \cos\varphi \qquad (6.24)$$

$$V_{BA} = C_0 - V_m \cos\varphi \qquad (6.25)$$

$$t_{AB} = \frac{L}{C_0 + V_m \cos\varphi} \qquad (6.26)$$

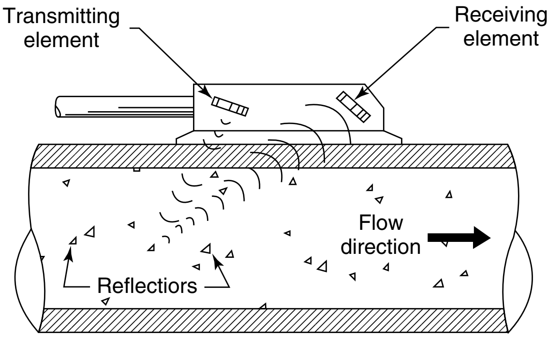

Fig. 6.12 Doppler-type ultrasonic flowmeter.

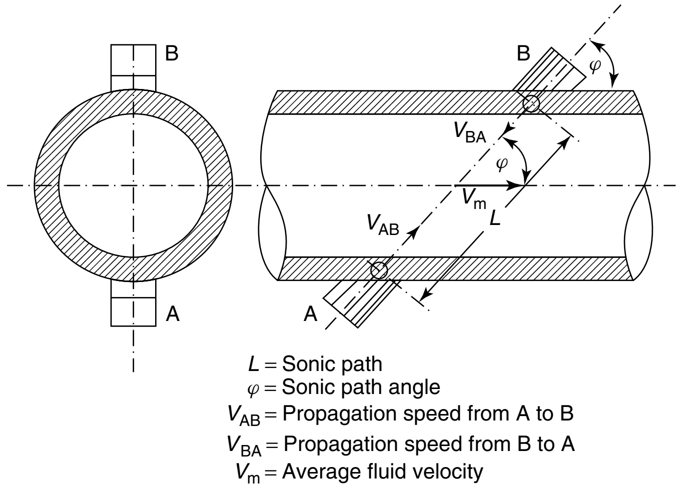
B
B
A
A
φ
φ
V_{BA}
V_{AB}
V_m
L
L = Sonic path
φ = Sonic path angle
V_{AB} = Propagation speed from A to B
V_{BA} = Propagation speed from B to A
V_m = Average fluid velocity

Fig. 6.13 Time-of-flight ultrasonic flowmeter.

$$t_{BA} = \frac{L}{C_0 - V_m \cos\varphi} \tag{6.27}$$

where L = sonic path length, φ = sonic path angle, C_0 = speed of sound for the fluid medium, V_m = average fluid velocity, V_{AB} = propagation speed from A to B, and V_{BA} = propagation speed from B to A. The parameters t_{AB} and t_{BA} are the measured ultrasonic wave flight times from points A and B, L and φ are known constants, and C_0 and the desired value V_m can be calculated. This method requires a clean fluid, and transducers can be of the wetted or clamp-on type. The turndown ratio is 40 : 1, and the meter can measure flow velocities down to $0.03\ \mathrm{m\,s^{-1}}$. Inaccuracies of $\pm 0.5 - \pm 2\%$ of flow are obtainable. Both ultrasonic methods can be nonintrusive, have no moving parts, and induce no pressure loss. Although these meters are typically used on liquids, they have been extended to gas flows as well as to solid-liquid and gas-liquid mixed flow (Rudroff, 2005; Magness, 2002; Soo, 1999; Witte, 2002).

An area of growth for ultrasonic flowmeters is in custody transfer measurements. To increase the accuracy (or lower the uncertainty of the ultrasonic flowmeter), multipath (multiple beams) systems have been developed. Typically, three to five beam paths are used at a single location in the transit time mode (Figure 6.14). The multiple paths help in reducing uncertainties that are caused by asymmetric velocity and temperature profiles, turbulence, and other flow disturbances. Uncertainties have been lowered to ±0.1% or less (Cousins, 2002; Basrawi, 2004).

A new ultrasonic method combines the Doppler and time-of-flight technique to expand the flow conditions over which the meter will operate accurately. An ultrasonic-Doppler velocity profile (UVP) technique (Mori et al., 2002) determines the velocity profile in small "channels" across a pipe section and then integrates those velocity profiles to yield the flow rate. The time-of-flight technique is used in its typical fashion. When one of the methods does not function well (too many bubbles or particles or the fluid is too

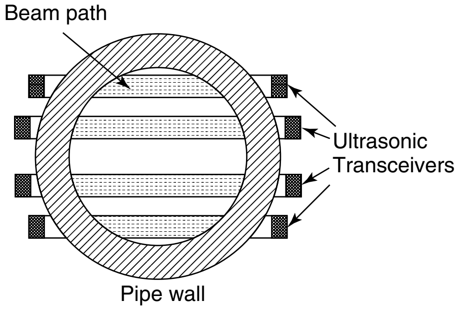

Fig. 6.14 Multibeam ultrasonic flow meter.

clean), the hybrid flow meter switches techniques to obtain a quality signal and flow measurement. The usable range and types of flow conditions that it can handle are substantially extended. (Yamamoto, Yao, and Kshiro, 2004).

A nice summary of the history of ultrasonic flowmeters is presented in a report (Lynnworth and Liu, 2006).

6.2.7
Mass Flow

True mass flowmeters are usually based on one of the two principles of operation. The Coriolis type is based on the Coriolis force that is generated when a section of pipe is vibrated while flow is passing through it (Figure 6.15). This measurement technique is essentially the application of Newton's second law of motion, where F = force, m = mass, and a = acceleration.

$$F = ma \tag{6.28}$$

The acceleration is known (from the tube vibrator), and the force is measured by the instrument itself. The amplitude of the generated force is a function of the acceleration and the fluid velocity and its mass, or the mass rate of flow. The major advantages of these meters are direct mass flow measurement, good turndown ratio (better than 10 : 1 and up to 100 : 1), small inaccuracy (±0.5% of rate), independence of fluid properties, measurement of either gas or liquid, and a requirement for little straight piping. Disadvantages include sensitivity to equipment-induced pipe vibrations, pressure loss, and the expense of the meter. Recent advances include lower cost designs, straight piping for lower pressure loss, and digital electronic signal processing designs to lower the response time from seconds to tenths of a second (Henry, Clark, and Mattar, 2004). This lower response time allows Coriolis flowmeters to work with very short batches or with pulsating flow from pumps. New flowmeters have been developed to work in

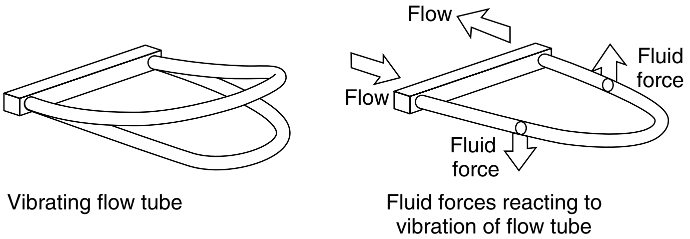

Fig. 6.15 Operating principle of a Coriolis force flowmeter.

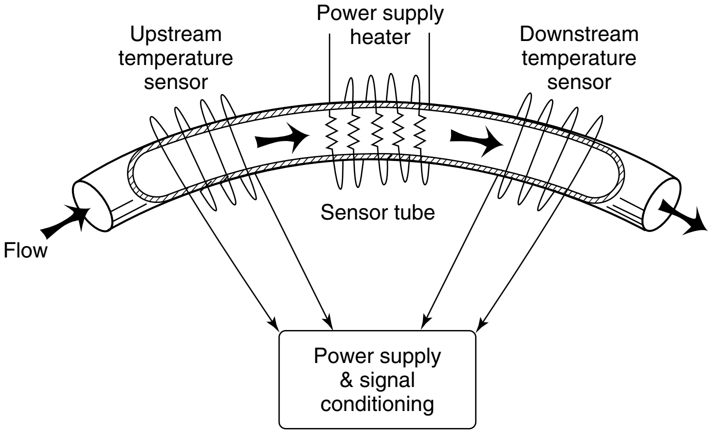
Upstream temperature sensor
Power supply heater
Downstream temperature sensor
Sensor tube
Flow
Power supply & signal conditioning

Fig. 6.16 Operating principle of a thermal mass flowmeter.

aerated liquids and continuous two-phase flow by using more than one flow tube driver and redesigning the flow tube geometry (Henry, Yeung, and Mattar, 2004). A Coriolis flowmeter was developed to handle higher flows, temperatures, and pressures. This design, an Omega Tube, combines a half circle measuring tube, torsion bar oscillation system, and a media feed tube section. This flowmeter is commercially available and is described in Hettrich and Daly (2004)).

The second mass flowmeter type is based on thermal mass effects (Figure 6.16). Typically, a temperature sensor is heated to a temperature above that of the fluid stream. The amount of cooling that the sensor undergoes is a function of the mass velocity of the fluid stream. Advantages are that the meter measures mass flow directly, has low pressure loss, has the ability to measure low velocities (Viswanathan et al., 2001), has a >10 : 1 turndown ratio, and has small inaccuracy (down to ±0.5% of full scale in some models). The disadvantages are sensitivity to fluid properties (especially thermal conductivity and capacity) and its potential for undetected calibration shifts caused by vapor condensation or coatings deposited on the sensing element.

Over the last 10 years, a significant amount of research has taken place to better understand the sensitivity of thermal mass flowmeters to fluid properties. This body of research has included modeling and improved calibration methods. Models were created to better understand the changes in linearity and accuracy (Hardy, 1999; Han, 2005; Kim, 2007) due to fluid property changes, geometry, and transients. Calibration techniques have also improved, which led to lower uncertainties (Hardy, 1999; Tison, 1996). The industry standard was to calibrate a thermal mass flowmeter on a benign gas and then use a simple thermal-physics model to ''transfer'' that calibration to a hazardous, corrosive, or toxic gas. This was essentially a linear transform. Rigorous testing in the early to mid-1990s showed that the correction factor should be a function (nonlinear) and errors using the factor could be as much as 70% (Hinkle, 1991). Gravimetric calibration devices were developed to handle toxic and corrosive gases; calibration functions were derived using experimental data and improved modeling to significantly reduce the "transfer" calibration errors (Remenyik, 1995).

6.2.8
Open Channel

Open channel refers to any conduit that allows a liquid to flow with a free surface such as a sewer line or a river. Open-channel flow measurement is often necessary at water and sewage treatment plants, waste water sites, and many other environmentally sensitive areas. This flow measurement is most often made by installing a specially designed device into the flow stream. The most commonly used insertion devices are weirs and flumes. Both devices are based on the principle that the flow in the open channel is a function of the water head (or level) that is created by the weir or flume. Weirs operate on the principle that a weir inserted into a stream will create a fluid head behind (upstream of) it. This head is a function of the flow velocity over the weir. Weirs have large turndown ratios, 20 : 1 to 60 : 1, and ±2 to ±3% full-scale inaccuracies, but they can trap entrained matter, which can lead to measurement inaccuracies. Flumes are specially designed structures that are similar to venturis in closed pipes. The flow through a flume can be related to the head at a specific location. Flumes have large turndown ratios, 40 : 1, have inaccuracies of ±2 to ±3% full scale, and are self-cleaning. Both devices use various methods to measure the head, including bubblers, differential pressure sensors, and ultrasonic instrumentation. Bubblers measure the pressure required to release a single air bubble from the end of a submerged tube. The required pressure is related to the head of liquid above the tube end. As the level of liquid changes above the tube, the required pressure changes and is measured and converted to a head or level measurement. A differential pressure sensor (submerged in the fluid) measures the pressure difference between the fluid surface and the bottom of the weir or flume. The pressure difference is directly related to the gravity head. By knowing the density of the fluid, then the fluid level can be determined. The ultrasonic head or level detector uses a time-of-flight technique to measure the distance from the ultrasonic sensor head to the fluid surface (the ultrasonic sensor is mounted above the water surface). As the level changes (the surface rises or falls), the ultrasonic sensor detects the change and converts it to a head measurement.

Open-channel flow can also be measured using an area-velocity method. This method is used when the channel geometry is known and a weir and flume is not practical. Using the channel geometry and typically a fluid level measurement, the flow area can be calculated. Multiplying the flow area by the average flow velocity results in the flow rate value. The average flow velocity is normally measured by one of the several techniques. These techniques include two ultrasonic methods (Doppler and Time-of-Flight), electromagnetic, radar, and pitot tubes. As described earlier, Doppler systems bounce acoustic signals off particle in the flow and the time-of-flight method send ultrasonic signals upstream and downstream into the flow. Electromagnetic flow measurement is based on Faraday's principle that a moving conductor (the fluid) passing through a magnetic field generates a current proportional to the velocity of the conductor. The fluid is the moving conductor in this case. Radar essentially uses a Doppler technique but bounces the signal off the moving fluid surface instead of submerged particles to determine the fluid velocity. Pitot tubes are differential pressure measuring devices that translate the pressure into a flow velocity. Typically, they are used as a point measurement; however, multiple port Pitot tube are available that measure an average pressure and thus, an average flow velocity. A fluid level or height measurement is also required (to determine the flow area) and the methods described in the previous paragraphs can be used with the average velocity techniques to derive the average flow rate. The area-velocity method works best in open pipe or conduit flow because the geometry is very well known. This method has been using in flowing streams with higher uncertainties but can be much more practical than a flume or weir in this application (see also: Gonzalez and Chanson, 2007; Sturm, 2001).

6.2.9
Anemometers

There are two major types of anemometers; one for measuring wind speed and one for measuring velocities at a point in a flow field.

The most widely used wind anemometer is the cup type. The cup type has three or four cups mounted on horizontal arms that are attached to a centered vertical shaft. The arms are equidistant apart, radially. The wind flows past the cups causing rotation. The rotation is proportional to the wind velocity. Other types of velocity anemometers include the windmill type where the axis of rotation is vertical and sonic anemometers. The windmill type usually includes a propeller and a tail on the same axis. These devices often measure the wind direction as well. The sonic instrument uses ultrasonic waves to measure wind speed and direction. These devices do not have any moving parts, which can result in higher reliability.

Various instruments have been developed to measure the velocity at a single point in the flow field. The two most commonly used in the research arena are the hot-wire or film anemometer and the laser Doppler velocimeter (LDV), sometimes called the *laser Doppler anemometer* (LDA). The hot-wire or film-sensor operation is based on the fact that the heat transfer from a heated probe depends on the properties of the flow past it. Two methods are used to determine the cooling effect. One method attempts to keep the sensor at a constant temperature by controlling the power to it. The power can then be related to the flow velocity. The second method is to keep the power to the hot film constant and monitor the probe's temperature changes, which can be related to flow velocity. The hot film and wire probes can be made very small to make flow measurements in boundary layers and small channels. Gases, liquids, and conducting metals are all acceptable media. Their time response is very fast. A schematic diagram of a hot-wire sensor is shown in Figure 6.17.

An LDV operates by splitting a laser beam into two equal-intensity parallel beams. A focusing lens crosses the two beams at its focal point, thus forming a measuring point or probe volume. At this location, the beam wavefronts interfere to form alternate regions of high- and low-intensity light "fringes." As a particle passes perpendicularly through the fringes, it results in variations in the intensity of the scattered light (Figure 6.18). The frequency of these variations can be measured, and by knowing the fringe spacing from geometry and lens characteristics, the particle velocity can be determined. The probe volume can be made very small for essentially point measurements. A

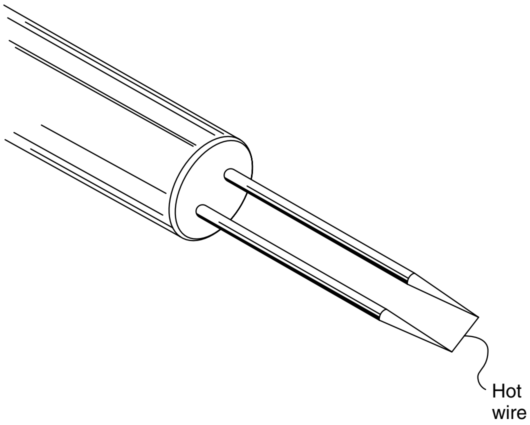

Fig. 6.17 Schematic diagram of a hot-wire anemometer.

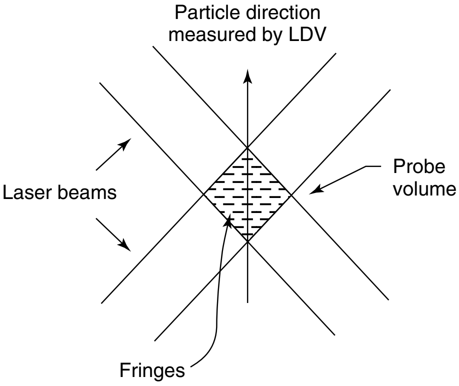

Fig. 6.18 Schematic diagram of a probe volume created by a laser Doppler velocimeter system.

major advantage of the LDV is that it is nonintrusive, but it does require particles to be present in the flow stream. A typical layout of a system is schematically shown in Figure 6.19. LDVs are available that can simultaneously measure the flow velocities in all the three dimensions. This option is often used in measuring the main velocity and orthogonal velocities to obtain an indication of turbulent intensities (see also: Durst et al., 1981; Zimmel et al., 2002).

6.2.10
MEMS Flowmeters

Over the last 30 years, MEMS (micro electro-mechanical systems) sensor technology has gone from a research interest to commercialized products. The drivers for these advancements have been the hope for very low-cost sensors when produced in large quantities and for the small footprint. An additional benefit is the low power consumption of MEMS sensors. To date, trade-offs have included moderate performance and relatively low temperature operation. Because of the large market potential, automotive applications have received the most attention, at least initially. The earliest flow sensors were designed and fabricated in miniature versions of thermal mass flowmeters. These MEMS devices demonstrated higher flow sensitivity, a faster time response, and lower power requirements than their larger counterparts. The largest application for

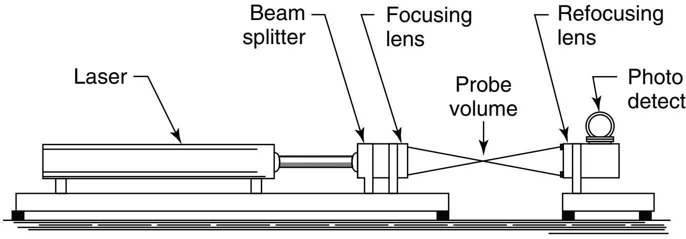

Fig. 6.19 Schematic diagram of a laser Doppler velocimeter system.

the thermal-based MEMS flow sensors is for heating, ventilation, and air conditioning systems. MEMS sensors integrated with signal conditioning electronics have a footprint of less than 25 mm × 25 mm. The thermal MEMS flow sensors typically include a heater and temperature sensing elements in a bridge circuit. The gas flows directly over the sensing elements and the heater. The electronics are designed to produce a voltage proportional to the flow rate. Some units have dual temperature elements on each side of the heater to measure flow direction. These sensors can be sensitive to flows as low as 0.1 sccm. Recent developments include a dust "proof" design for insertion in ventilation systems and inexpensive, disposable dust filters that mount immediately upstream of the flow meter.

A variety of MEMS flow sensors are becoming available and new designs are being developed. A MEMS-based vortex flow sensor (Pedersen and Andersen, 2003) that operates by measuring the vortex-shedding frequency behind a bluff body in a flow stream has been created. A MEMS-based pressure sensitive detector is located in the fluid stream downstream of a bluff body. Owing to the high sensitivity of the MEMS pressure membrane, very low flow rates can be sensed, lower than usually possible with vortex-shedding sensors (Reynolds number as low as 2000 instead of ~10 000). A coriolis mass flow sensor has also been developed with a sealed MEMS chip (Sparks et al., 2003). At the heart of the flow sensor is a U-shaped, vacuum-sealed resonating silicon microtube. The tube is driven into resonance and its motion is sensed capacitively using metal electrodes under the tube. In addition to measuring mass flow rate, chemical concentration, fluid density, and temperature can also be monitored.

As with the coriolis MEMS sensor, other multisensor MEMS chips are being developed (Xu et al., 2005). These chips can measure multiple parameters simultaneously. Parameters such as shear stress, temperature, and pressure have been combined into one chip. Flow rate measurements have been demonstrated using the pressure drop measurement and temperature values with thermal anemometry principles.

6.2.11
Summary of Flowmeters

Table 6.1 summarizes the major attributes of the flowmeters described in this article.

6.3
Flow Measurement Uncertainties

It is practically impossible to place a flowmeter in service and expect to achieve the manufacturer's performance

Tab. 6.1 Summary of flowmeter attributes (FS, full scale; D, pipe diameter).

Flowmeter	Turndown	Accuracy	Piping requirements	Reynolds number	Major advantage	Major disadvantage
Pitot tube	4 : 1	± 5% FS, ± 1% FS averaging Pitot design	10D up/5D down	>20 000	Low permanent pressure drop, low cost	Small differential pressure for gases, low accuracy, and turndown
Orifice	4 : 1	± 1–2% FS	10D up/5D down	>10 000	Low cost, lots of empirical data	Low turndown
Cone type	10 : 1	± 0.5% rate	0-3D up/0-1D down	>10 000	Minimal flow conditioning	Each meter has unique calibration
Vortex shedding	10 : 1 to 20 : 1	± 1% rate	10D up/5D down	>10 000	Accuracy, no moving parts	Signal dropout at low flows, viscosity dependent
Magnetic	10 : 1 to 30 : 1	± 0.5–1% rate	3D up/none down	No limit	Accuracy, turndown, and minimal flow conditioning	Cost, fluid must be conductive
Turbine	10 : 1	± 1% rate	10D up/5D down	>4000	Accuracy	Bearings, viscosity dependent
Positive displacement	10 : 1	± 0.5–1% FS	None	None	Rugged	Viscosity dependent
Variable area	10 : 1	± 0.5–1% rate	None	Not highly viscous fluids	Low cost	More difficult to provide with electrical interface, vertically mounted
Drag	10 : 1 to 15 : 1	± 0.5–2% FS	10D up/5D down	>4000	Handles dirty fluids	Erosion of drag target
Ultrasonic Doppler	10 : 1	± 2% FS	10D up/5D down	No limit	Noninvasive	Needs particles in fluid
Ultrasonic transit time	40 : 1	± 1% rate	10D up/5D down	No limit	Accuracy, turndown	Need clean fluid
Coriolis	>10 : 1	± 0.2% rate	None	No limit	Accuracy	Cost, size
Thermal mass flow	10 : 1	± 1% FS	None	No limit	Very low flows	Cost

specifications without paying attention to application-specific details. The end user faces the responsibilities of knowing the effects of flow measurement uncertainties for each application and ensuring that the flowmeter installation is engineered to mitigate or negate these uncertainties.

6.3.1
Installation Effects

The geometry of many flowmeter elements is critical to accurately measure the flow. Factors that could change the dimensions of the flowmeter element should be addressed before the meter is installed. Thermal expansion of the element and erosion/corrosion from solids-bearing flow or as a result of cavitation or liquid impingement should be addressed (Spitzer, 2005). The potential for coating an element should also be evaluated.

The flow transmitter electronics should also be taken into account when installation effects are addressed. Temperature and relative-humidity (RH) cycles of flow transmitters need to be considered for the actual installation location. The effect of exposing the flow transmitter to direct sunlight may, in some cases, exceed the transmitter's temperature specification. Many flow transmitter electronics are rated for noncondensing RH service. At least, one end-user company tests its flowmeters at 100% RH because, even if ambient conditions are not 100% RH, if the electronics are not sealed, leakage into the electronics during the day at warmer temperatures will cause condensation of moisture when the temperature drops below the dew point at night (DeBoom, 1992).

For volumetric flow measurements inferred from the inside cross-sectional area of the pipe, an accurate measurement of the inside diameter should be taken. For example, the nominal inside cross-sectional area of a Schedule 40, 6-in.-diameter (15.2-cm-diameter) pipe is allowed to vary as much as 4.5% according to American Society of Testing Materials guidelines (Kopp, 1989). The diameter error translates to a squared error in flow for differential pressure flow meters.

In the past, differential pressure flow meters used impulse tubes to transfer the pressure measurements from the flow element (e.g., orifice) to the flow transmitter. Recently, new flow transmitter and manifold designs have allowed direct mounting of the flow transmitter to the line, which eliminates the impulse tubing. Impulse tubing is susceptible to the following error sources (Strom and Livelli, 2002).

- *Plugging*: This can be caused by a precipitate left by steam flashing, a low spot in the line where condensation has occurred (gas flow), accumulation of solids, and freezing (liquid flow).
- *Leakage*: The number of potential leaks increases with increased number of connections (tube fittings, manifolds, seal pots, blowdown valve connection, and so on).
- *Hydrostatic head uncertainty*: The upper and lower side of dP measurement are not at same level, or not along the same axis of pipe line.

6.3.2
Fluid Properties

Several fluid properties may have significant effects on the performance of a flowmeter or a sensor. Fluid density, compressibility, vapor pressure, and viscosity are important to fluids in motion.

In addition, thermal properties such as gas constant, specific heat capacities, and thermal conductivity are important in heat transfer. The thermodynamic state of a fluid, determined by its pressure and temperature, affects most of the above-mentioned properties. Other parameters of importance, which are combinations of fluid properties, include the Prandtl number (involves viscosity, specific heat, and thermal conductivity), thermal diffusivity (involves thermal conductivity, density, and specific heat), and the Reynolds number (inertial to viscous force ratio and an indication of laminar or turbulent flow).

Because the fluid properties affect the flow instrumentation's performance and usually its calibration, one must be careful if the flowmeter experiences fluid properties that are different from the ones at calibration or if the fluid properties are not constant. Errors in the output of the flowmeter may occur, increasing the uncertainties obtained. In some cases, these errors can be very high (easily greater than 20%). Depending on the particular flowmeter's sensitivity to fluid properties, conditions that are known to be different from the calibration state or that are time-varying may require monitoring. Corrections to the output of the flowmeter can then be made, if possible, to reduce property-induced errors.

In the past, the most sophisticated method to correct for property-induced flow rate effects over just a plain differential pressure measurement was to measure temperature and pressure and to perform density compensation on the flow signal. As can be seen by Equation (6.8), there are other variables in the flow equation. Modern flow transmitters (smart transmitters) are able to continuously compensate for the following (Rowe, 2004; Strom and Livelli, 2002):

- changes in fluid density with specific algorithms optimized for fluid type (i.e., liquid, gas, or vapor);
- dimensional changes in primary device and adjacent piping due to pressure and/or temperature;
- changes of primary device discharge coefficient with Reynolds number;
- changes in saturation temperature for saturated vapor (e.g., steam, ethylene);
- velocity of approach factor; and
- gas expansion factor.

6.3.3
Velocity Profiles

Fully developed velocity profiles for laminar flow are parabolic and symmetric about the pipe axis. Fully developed velocity profiles for turbulent flow are symmetric about the pipe axis and are almost constant from the pipe axis outward, except near the pipe walls where a sharp gradient occurs. The velocity profile through a flowmeter should be taken into account when the meter is sensitive to this effect. Asymmetric flow profiles should be avoided. Swirling-type flow velocities are caused by upstream pipe elbows. Jets can be caused by high-velocity, smaller diameter streams that enter larger diameter streams upstream. Straight-pipe lengths upstream of the flowmeter are the best way to avoid asymmetric flow profiles. Flow-straightening vanes, of either the plate type or the honeycomb type, should be avoided because they inhibit lateral mixing of the fluid, which is required to reduce the asymmetric profile. Downstream straight-pipe lengths should be designed as well. Other considerations include pipe ovality, smoothness of the pipe wall, concentricity of the pipe and

Tab. 6.2 Minimum straight-pipe diameters for orifices, venturis, and nozzles with respect to diameter ratio [a].

Diameter ratio, beta	Single 90° elbow	Two 90° elbows, same plane	Two 90° elbows, out of plane	Globe valve fully open	Downstream
0.2	10	14	34	18	4
0.4	14	18	36	20	6
0.6	18	26	48	26	7
0.8	46	50	80	44	8

[a] Length for straight pipe diameters measured in same units as pipe diameter. Last column is required straight pipe downstream of flow element.

meter, and gaskets projecting into the fluid stream. Flangeless wafer-type meters may not provide adequate centering (Kopp, 1989).

Table 6.2 is an excerpt from ISO 5167-1 regarding the minimum pipe diameters of straight line that are required for different geometries for orifice plates, nozzles, and venturi tubes. Similar tables are available for other flow devices. Notice that the diameter ratio of the flow element has a large impact on the straight line required. This is because with small diameter ratios, the flow element itself plays a large role in shaping the velocity profile through the meter. At large diameter ratio diameters, the velocity profile is not affected much by the flow meter and is more susceptible to the effects of nonideal geometries (ISO, 2003).

6.4
Calibration Methods

Calibration with the meter in place is recommended over having it calibrated in a laboratory because installation effects are taken into account. By the same token, it is always better to calibrate with the same fluid that will be monitored by the meter. Pressure and temperature ranges also should be the same because they affect both the fluid and the meter (Sydenham, 1999). Calibration methods have evolved throughout the years, and this section provides an insight into the various methods.

6.4.1
Time-Collection Techniques

Time-collection techniques involve diverting a fluid in series with the meter being calibrated into either a volume- or weight-measuring device and recording time and either volume or weight data points. The techniques are probably among the oldest methods of performing a flowmeter calibration, but modern data acquisition equipment will greatly improve the results from these tests if it can be used. This calibration technique allows traceability to primary standards: mass or volume and time.

6.4.1.1 Monitored Volume Methods

Time versus volume techniques consist of measuring the volume of flow through a flowmeter over a given time duration. For example, in liquid flows, the flow could be diverted to a tank. The volume of the tank would be measured before and after the test. The volume change divided by the time duration provides the flow reference. For gases, a ''bell prover'' can be used. An inverted bell fills with gas during the calibration test. As the bell fills, it rises. The amount that the bell rises provides an indication of the volume of gas that has flowed through the flowmeter. Other gas calibration devices include the bubble meter, where a soap-film bubble rises in a graduated cylinder relative to flow; and the piston meter, where a sealed piston is used rather than a soap-film bubble (Wight, 1994).

The disadvantage of time versus volume methods is that the volume devices have not, in general, been outfitted for automatic data collection. Only beginning and ending data points are available, and flow was assumed to be steady during the entire test. The time versus volume method is acceptable if a volumetric flowmeter is being tested; but if a mass flowmeter is being tested, the density of the fluid must also be determined and used to calculate true mass flow.

6.4.1.2 Monitored Weight Methods

Time versus weight methods are similar to time versus volume methods, except that the reference variable is weight instead of volume. Many weigh-cell meters have the advantage that they have electronic interfaces for data output and respond fast enough to provide dynamic data during a calibration test. The advantage here is that the weight data can be examined for periods of unsteady flow during the test. The periods of unsteady flow can be dropped from the analysis when the calibration comparison is made.

Time versus weight methods are good if a mass flowmeter is being tested; but if a volumetric flowmeter is being tested, the density of the fluid must be determined and used to calculate true volumetric flow.

Gravimetric calibration devices are weight-method- based calibrators. In the mid-1990s, one such device was developed to handle toxic and corrosive gases for low-flow thermal mass flow meters. This gravimetric calibrator is schematically shown in Figure 6.20. The 20-l vessel was designed to be slightly heavier than neutral buoyancy so its mass is essentially tared out; thus, only the mass of the gas is weighed by a precision load cell. This technique allows the incoming gas mass to be accurately measured, even in small amounts (just a few grams of gas need to be collected). The vessel was passivated to handle corrosive gases and after completing a calibration run, the vessel was vented to a toxic gas incinerator. Flow rates to several hundred cubic centimeters a second could be run and accuracies to ±0.1% could be obtained in just a few minutes. A total calibration (10 data points) could be completed in a couple of hours. Standard gravimetric calibration techniques typically take one to two days to complete because large quantities of the gas must be collected to allow for an accurate measurement (the mass of the collection device is much larger than to the mass of the gas). To reduce the measurement uncertainty, the mass of the gas collected needs to be a reasonable percentage of the mass of the collection device, thus the long calibration times.

6.4.2
Transfer Standards

Generally, it is not practical or economical to use a primary standard for calibrating a flowmeter. The primary standard may be too expensive to build and maintain. Routine commercial practice is to have high-quality flowmeters calibrated against a primary standard (and have this done periodically like once in a year) and

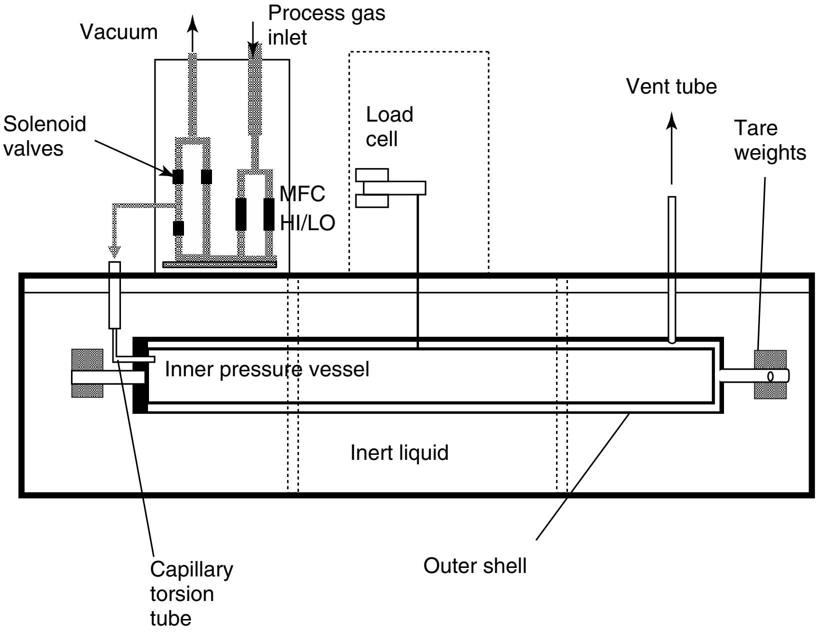

Fig. 6.20 Gravimetric calibrator schematic for mass flow controller (MFC).

then use this high-quality flowmeter as a secondary standard. The secondary or transfer standard is usually based on fundamental operating principles that conform to theoretical predictions very well. The secondary standard must be stable and predictable for long periods of time (a year or more). The transfer of the secondary standard to a flowmeter that needs calibrating is accomplished by installing the secondary standard and the device requiring calibration in series in a flow loop. The secondary reference method of calibration consists of allowing the fluid to flow through the flowmeter under calibration in series with a standard meter. The readings of the two meters are compared, and the meter under test is calibrated as necessary on the basis of the standard meter readings. Care must be taken not to allow flow obstructions caused by the upstream meter to affect the velocity profile entering the other meter. Temperature and pressure changes between the two meter locations must be accounted for as well. The most common secondary standard flowmeters are the venturi meter, critical nozzles, squared-edged orifices, and laminar flow elements.

6.4.3
Metered Volume Methods

Metered volume methods are the opposite of monitored volume methods. Metered volume methods of calibration consist of engineering a system that will meter a precise volume of fluid through a flowmeter in a measured time duration, rather than measuring the volume produced by another means. One system consists of a pipe displacer or ''pig'' that is driven through a honed section of pipe. The pig is timed from the time it passes a beginning point of the honed pipe until it passes the ending point. The pig travels between the points at a constant speed. If the diameter and length of the honed pipe and the transit duration are known, volumetric flow rate can be calculated and compared with a flowmeter reading in series with the honed section. The test duration of this type calibrator is very short. Another system uses a bidirectional U-tube system where a four-way valve switches flow at the ends of the U-tube to ensure a continuous metered flow. The flow in the U-tube switches direction when the pig reaches either of the end points. Detectors are mounted near the ends of the U-tube to sense the passing of the pig. The system can run indefinitely, as compared to the once-through pig method (Paton, 2005).

Glossary

Accuracy: The closeness of agreement between an observed value and the true value.

Boundary Layer: A very thin layer of fluid in the neighborhood of a body where friction plays an essential part in the hydrodynamics. Outside this thin layer, friction can be neglected. The boundary layer also controls the heat-transfer characteristics from a surface to the surrounding fluid.

Calibration: The procedure of comparing a particular instrument with either a primary standard or a secondary standard with a higher accuracy than the instrument to be calibrated. This process allows the inaccuracy of the instrument to be quantitized against a known, accepted reference.

Compressibility: The measure of the change of volume of a fluid under the action of external forces. The compressibility of liquids is typically very small, and, thus, liquids are considered to be incompressible. Gases do exhibit significant volume changes and are considered as compressible fluids.

Drift: An undesired change in the output–input relationship over a period of time.

Flow Totalization: The summation of the flow sensed by an instrument as a function of time. Flow totalization is used in product custody transfer and to determine home and business water and energy usage.

Flow Transmitter: A device that responds to a measured variable by means of a flow element, and converts it to a standardized transmission signal that is a function only of the measured variable.

***In situ* Calibration:** Calibration of an instrument carried out with the instrument located in its measurement or field position. The instrument is calibrated *in situ* – in place – rather than removing the device and calibrating it in a laboratory.

Invariant: A mathematical term that does not change under a set of transformations.

Isentropic: Taking place with no entropy production, that is, at constant entropy.

Laminar: The state of a fluid in motion where the velocity is free of macroscopic fluctuations or mixing.

Liquid Impingement: A process by which a material is eroded by liquid droplets, contained in a liquid/vapor stream, striking the material at high velocities.

Pressure Tap: The physical entry into a system from which the system pressure may be measured. The ‘‘tap’’ is usually a hole through the system boundary from which the system pressure may be sensed.

Repeatability: The closeness of agreement among a number of consecutive measurements of the output for the same value of the input under the same operating conditions, approaching from the same direction. It is usually measured as a non-repeatability in percent of span.

Residual: The difference between the current iteration value and the value calculated during the previous iteration step.

Signal-to-Noise Ratio: The ratio of the electrical magnitude of the desired measured quantity (signal) to the undesired component of that magnitude (noise). The undesired component comes from flow turbulence, environmental factors, and electrical interferences.

Smart Flow Sensor: Flow sensor that contains a microprocessor that allows the sensor to make corrections for changing environmental conditions or flow range or calibration changes. These changes may allow the flow sensor to improve its accuracy or extend its rangeability.

Turbulent: The state of a fluid in motion where the velocity is subject to macroscopic fluctuations or mixing. These velocity fluctuations are superimposed on the mean motion.

Uncertainty: The range within which the true value of the measured quantity is expected to be; an indication of the variability associated with a measured value.

Velocity Profile: The distribution of the magnitude of the velocity vector in a flow stream. In a closed system, the velocity is the highest at the centerline and drops to zero at the system's fixed boundary or wall.

Viscosity: The constant of proportionality between the applied shear stress and the rate of deformation or to the velocity gradient normal to the velocity of the fluid. It is the measure of the fluid's resistance to flow.

References

Basrawi, Y.F. (**2004**) Ultrasonic Flow Metering Devices and Their Applications to Royalty and Custody Transfer Measurements, ISA AUTOMATION WEST, 2004.

Cousins, T. (**2002**) Proving of Multi-Path Liquid Ultrasonic Flowmeters, 20th NSFMW.

Dahlstrom, M.J. (**1994**) V-cone meter-gas measurement for the real world. North Sea Flow Measurement Workshop, Peebles, Scotland.

DeBoom, R.J. (**1992**) Rating a Flowmeter's Field Performance, InTech, July.

Durst, F., Melling, A., and Whitelaw, J. (**1981**) *Principles and Practice of Laser-Doppler Anemometry*, Academic Press, London.

Gonzalez, C. and Chanson, H. (**2007**) Experimental measurements of velocity and pressure distribution on a large broad-crested weir. *Flow Meas. Instrum.*, **18** (3–4), 107–113.

Han, I.Y. (**2005**) Study of the transient characteristics of the sensor tube of a thermal flow meter. *Int. J. Heat Transfer*, **48** (13), 2583–2592.

Hardy, J.E. *et al.* (**1999**) *Flow Measurement Methods and Applications*, John Wiley & Sons, Inc., pp. 135–203.

Henry, M., Clark, C., and Mattar, W. (**2004**) The Dynamic Response of Coriolis Mass Flow Meters: Theory and Application, ISA, Houston, October 2004.

Henry, M., Yeung, H., and Mattar, W. (**2004**) How a Coriolis Mass Flow Meter Can Operate in Two-Phase Flow, ISA, Houston, October 2004.

Hettrich, U. and Daly, J. (**2004**) Omega Tube Design Extends the Usability of Coriolis Technology in Mass Flowmetering Applications, ISA, Houston, October 2004.

Hinkle, L. (**1991**) MFC Accuracy: is simple gas correction enough? Semiconductor International, March.

Ifft, S.A. and Mikkelsen, E.M. (**1993**) Pipe elbow effects on the V-Cone flowmeter. Presented at ASME Fluids Engineering Conference 1993, Washington, DC.

ISO (**2003**) 5167-1. *Measurement of Fluid Flow by Means of Pressure Differential Devices*, International Organization for Standards, Genève.

Kim, S.J. (**2007**) Study of the steady-state characteristics of the sensor tube of a thermal flow meter. *Int. J. Heat Transf.*, **50** (5–6), 1206–1211.

Kopp, J.G. (**1989**) How to Get the Accuracy You Expect from Flowmeters, I&CS, September.

Lipták, B.G. (**2003**) *Instrument Engineers' Handbook*, Process Measurement and Analysis, vol. **1**, CRC Press, Boca Raton.

Lloyd, K.E., Gutherie, B.D., and Peters, R.J.W. (**2002**) A flowmeter calibration facility developed at The University of Iowa to evaluate custody transfer steam flowmeters with a cone differential pressure meter used as the metering standard. ISA 2002 Technical Conference, October 21-24, 2002, Chicago.

Lynnworth, L.C. and Liu, Y. (**2006**) Ultrasonic flowmeters: Half-century progress report, 1955-2005. *Ultrasonics*, **44** (Suppl 1), 1371–1378.

Madison, R.M. (**1995**) Modern electronics meet turbine meters, Measurements & Control, April.

Magness, M. (**2002**) Ultrasonic Flow Measurement of Natural Gas Flow, ISA 2002, Chicago, October 2002.

Mattar, W. (**2004**) Vortex shedding flowmeters – a technology finally comes of age. Presented at the ISA 2004, Houston, Texas, October 5-7, 2004.

McCoy, M. (**1992**) Sensor data integration for improved flow data accuracy, Sensors, February.

McMillan, G.K. and Considine, D.M. (**1999**) *Process/Industrial Instruments and Controls Handbook*, 5th edn, McGraw-Hill, New York.

Miller, R.W. (**1996**) *Flow Measurement Engineering Handbook*, McGraw-Hill, New York.

Mori, M. *et al.* (**2002**) Development of a novel flow metering system using ultrasonic velocity profile measurement. *Exp. Fluids*, **32** (2), 153–160.

Ostling, H. and Oki, S. (**2001**) Spectral Signal Processing (SSP) enhances vibration immunity for vortex flowmeters. ISA TECH/EXPO Technology Update Conference Proceedings, Houston, Texas, 416, pp. 411–417.

Paton, R. (**2005**) Calibration and standards in flow measurement, in *Handbook of Measuring System Design* (P.H.Sydenham), John Wiley & Sons, Inc., New York.

Pedersen, N. and Andersen, P.E. (**2003**) A MEMS-based vortex flow sensor for aggressive media. *Proc. IEEE Sens.*, **2** (1), 320–325.

Remenyik, C.J. (**1995**) An instrument for gravimetric calibration of flow devices with corrosive gases. Proceedings of the 41st International Instrumentation Symposium, May 7-11, 1995, Aurora.

Rose, C. and Vass, G. (**1993**) Obstructionless magenetic flowmeters. Sensors, March.

Rowe, J.D. (**2004**) Advanced sensor technology key to new multivariable transmitter. Presented at ISA Expo 2004, October 5-7, 2004, Houston.

Rudroff, D. (**2005**) Ultrasonic meters ready for gas distribution market. *Pipeline Gas J.*, **232** (7), 24–26.

Shen, J.S., Bosio, J., and Larsen, S. (**1995**) A performance study of a V-Cone meter in swirling flow. 13th North Sea Flow Measurement Workshop, October 24-26, 1995, Lillehammcr.

Soo, S. (ed.) (**1999**) *Instrumentation of Fluid-Particle Flow*, Noyes Publications, pp. 172–196.

Sparks, D. *et al.* (**2003**) A portable MEMS coriolis mass flow sensor. *Proc. IEEE Sens.*, **2** (1), 337–339.

Spitzer, D.W. (**2005**) Industrial Flow Measurement, ISA, Research Triangle Park.

Strom, G.R. and Livelli, G. (**2002**) A New Generation of DP Measurement Instrumentation Eliminates Traditional Limitations, Intech, August.

Sturm, T. (**2001**) *Open Channel Hydraulics*, McGraw-Hill, Boston.

Sydenham, P.H. (**1999**) *Handbook of Measurement Science*, John Wiley & Sons, Inc., New York.

Tison, S.A. (**1996**) A critical evaluation of thermal mass flow meters. *J. Vac. Sci. Technol. A Vacuum, Surf. Films*, **14** (4), 2582–2591.

Viswanathan, M. *et al.* (**2001**) Development, modeling and certain investigations on thermal mass flow meters. *Flow Meas. Instrum.*, **12** (5–5), 353–360.

Walker, J.R. (**2001**) Diagnostics advance in electromagnetic flow metering. Presented at ISA 2001, September 10-12, 2001, Houston.

Wight, D.W. (**1994**) *Fundamentals of Air Sampling*, CRC Press, Boca Raton.

Witte, J.N. (**2002**) Ultrasonic gas meters from flow lab to field: a case study. American Gas Association Operations Conference Chicago, Illinois.

Yamamoto, T., Yao, H., and Kshiro, M. (**2004**) Advanced Hybrid Ultrasoinc Flow meter Utilizing Pulsed-Doppler Method and Transit Time Method, ISA, Houston, October 2004.

Xu, Y. *et al*. (**2005**) A MEMS multi-sensor chip for gas flow sensing. *Sens. Actuators A Phys.*, **121** (1), 253–261.

Zimmel, M. *et al*. (**2002**) Comparison of velocity measurements by high temperature anemometer and laser-Doppler anemometer with results of CFD-simulation. *Therm. Sci.*, **6** (1), 3–13.

Further Reading

Stauss, T. (ed.) (**2004**) *Flow Handbook,* Schaub Medien AG, Reinach.

Sydenham, P.H. (ed.) (**2005**) *Handbook of Measuring System Design*, John Wiley & Sons, Inc., New York.

Webster, W.C. (ed.) (**1999**) *Measurement, Instrumentation, and Sensors Handbook*, CRC Press, Boca Raton.

Yoder, J. (**2001**) Flow meter shootout part II, Control, February.

MEASUREMENT OF MAGNETIC PROPERTIES AND QUANTITIES

S. Foner, *Francis Bitter National Magnet Laboratory and Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*

Introduction 463
1. **Classical Macroscopic Magnetization Measurements** 464
1.1 Flux Detection 465
1.2 Force Detection 465
1.3 Mechanical Torque Detection 466
2. **Applications to Magnetic Measurements** 466
2.1 Flux Measurement 466
2.1.1 dc 466
2.1.2 External Variables 467
2.1.3 ac Susceptibility 467
2.1.4 Vibrating-Sample Magnetometer 468
2.1.5 Superconducting Quantum Interference Devices (SQUIDs) 469
2.1.6 Pulsed Magnetic Fields 470
2.2 Force Measurement 471
2.2.1 dc Gradients 471
2.2.2 ac Gradients 472
2.3 Mechanical Torque Measurement 472
3. **Measurements of Magnetic Properties of Practical Materials** 474
3.1 Magnetization 474
3.2 High Permeability 475
3.3 Permanent Magnets 476
3.4 Losses 476
3.5 Anisotropy 447
3.5.1 Magnetization Measurements 477
3.5.2 Mechanical Torque Measurements 478
3.5.3 Ferromagnetic Resonance 479
3.6 Transition Temperatures 479
3.7 Magnetostriction 480
4. **Microscopic Measurements** 481
5. **Observation of Magnetic Domains** 484
6. **Measurement of Magnetic Fields** 484
Acknowledgment 486
Glossary 486
Works Cited 488
Further Reading 489

INTRODUCTION

Until recently, most magnetic measurements required the construction of the instrument, and this led to proliferation of approaches to suit specific needs; it sometimes appears there are more ways to measure magnetic quantities than there are magneticians. New modifications continue to be developed to address specific applications. In the last 2 decades, commercial instruments have become available that embody some of the most popular and versatile methods. The availability of relatively inexpensive microcomputers that can be dedicated to the instrument makes it possible to remove much of the tedium of data taking and analysis. When coupled to associated computer-controlled magnets and temperature-controlled environments, these instruments become very useful stand-alone automated systems (sometimes at the expense of making it difficult to modify these instruments for specialized needs). The purpose of this article is to summarize some of the general principles of magnetic measurements and their applications. It is not possible here to discuss the wide variety of embodiments of these principles; see, for example, reviews by Dwight (1967), Humphrey (1967), Foner (1967, 1981), and Flanders and Graham (1993) and the books by Zilstra (1967). Here, most of the emphasis will be on those techniques in wide use and/or adopted in commercial instruments. The section on measurements of properties addresses mainly practical (ferromagnetic or

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ferrimagnetic) magnetic materials, but the principles are applicable more generally to all magnetic measurements.

The accepted magnetic units are the SI units, although Gaussian units, shown here in brackets, continue to appear in the literature. These are based on the relations

$$B = \mu_0(H + M) \quad [B = H + 4\pi M], \tag{1}$$

where B is the magnetic flux density in tesla (T) [$=10^4$ gauss (G)], $\mu_0 = 4\pi \times 10^{-7}$ H/A (henry/ampere), H is the magnetic field strength in A/m [$=4\pi/10^3$ oersted (Oe)], M is the magnetization per unit volume in A/m [$=10^3$ emu/cm^3], and mechanical torque is given in N·m [$=10^{-7}$ dyn-cm]. Often it is easier to measure the mass than the volume, and the mass (unlike the volume) does not depend on external variables such as temperature, pressure, field, etc., so that σ, the mass magnetization (A·m^2/kg) [emu/g] appears in the literature. $M=\rho\sigma$, where ρ is the density (kg/m^3) [g/cm^3]. The magnetic moment is given in A·m^2 [$=10^3$ emu]. The magnetic flux over a given area is $\Phi = Ba$ in webers (Wb) [$=10^8$ G·cm^2], where a is the area.

Most practical magnetic materials are strongly magnetic. Thus, when a field H_0 is applied to the material, the internal field H is less than H_0 because the magnetic polarization induced on the surface of the sample produces a demagnetization field. The properties of the material depend on H, which can be markedly different from H_0, so that a measurement of $M(H_0)$ must be corrected for the demagnetization field. The demagnetization field $\mathbf{H}_d$ can be written as $\mathbf{H}_d = -\mathbf{N}\cdot\mathbf{M}$, where $\mathbf{N}$ is the demagnetization factor. $\mathbf{N}$ is a tensor which, in general, is difficult to determine except when the magnetization is uniform. This is the case for a homogeneous ellipsoid of revolution in a uniform applied field. The sum of the N_i along the three principal directions is 1 [4π]. An isotropic homogeneous sphere has $N = \frac{1}{3}$ [$4\pi/3$] along three orthogonal directions. Approximations of $\mathbf{N}$ can be made in the limiting cases of degenerate ellipsoids such as a very thin plate (where the thickness is much less than the other dimensions, and $N = 1$ perpendicular to the plate and 0 parallel to the plate) and a very long thin circular rod (where the diameter is much less than the length, and $N = 0.5$ [2π] in the two directions perpendicular to the axis and 0 along the axis). In nonellipsoidal bodies, the magnetization is generally not uniform except at fields well above magnetic saturation; as the internal field increases, N changes and large errors can be made in deriving $M(H)$. Furthermore, the inhomogeneity of M depends on the history of the field change, because the magnetization of a ferromagnet shows hysteresis below magnetic saturation. Distinctions are made in the literature between a magnetometric N (N_m), which is the average N for small samples, and fluxmetric N (N_f), which refers to the N at the midplane of a long bar where a short pickup coil detects the flux. Calculations of N for circular cylindrical-shaped samples are reviewed by Chen *et al.* (1991). For accurate determination of $M(H)$ in ferromagnetic materials, a well-defined N is necessary, or the effects of the boundaries must be eliminated. For very weakly magnetic materials, the demagnetization fields can be small and easily approximated, so that negligible errors are made.

Usually, there are improvements in sensitivity if the measuring technique senses a periodic variation of the signal so that many measurements are taken under the same conditions over a short time. Such ac methods are usually processed by using a phase-sensitive detector, which rectifies the ac signal with a gate furnished by the ac source and which is adjusted (locked) to the phase of the detected signal. This device acts as a narrow-band filter, discriminating against signals having the wrong frequency or varying phase with respect to the ac source. Generally, the phase detector is combined with ac amplification prior to detection and postdetection dc amplification and further filtering; the combination is called a lockin amplifier or a lockin for brevity. This device is used for processing the data from almost all the ac instruments discussed here. Recent trends in lockin amplifiers are toward replacement of various analog components by digital units. Lockin amplifiers are readily available commercially and permit convenient processing of ac signals.

1. CLASSICAL MACROSCOPIC MAGNETIZATION MEASUREMENTS

For convenience, classical magnetic measurements are defined here as those that measure the change of flux distribution introduced by a sample, the force on a sample produced by a nonuniform field, or the mechanical torque on an anisotropic sample in a uniform field. These techniques measure the macroscopic properties of the material.

1.1 Flux Detection

A magnetic material (for instance a permanent magnet) produces magnetic fields or changes the magnetic flux density in its neighborhood. This magnetic field change can be detected by magnetic induction with pickup coils or other field-sensitive detectors in the vicinity of the sample. Suppose that a sample is positioned in one of two series-opposing pickup coils [Fig. 1(a)] with area-turns n_1a_1 and a_2a_2. The change in Φ for a change in applied field is given by

$$\delta\Phi/\mu_0 = \Delta_1\delta H + \Delta_2\delta M, \tag{2}$$

where $\Delta_1 = n_1a_1k_{H1} - n_2a_2k_{H2}$ and $\Delta_2 = n_1a_1k_1 - n_2a_2k_2$; the k_{Hj}'s are geometric factors that depend on the effective areas of the coils exposed to H_0, and the k_j include geometric factors that relate to the effective flux linkage of the sample to the coils. If the sample position is fixed, accurate measurement of $M(H)$ often depends on how closely Δ_1 approaches zero. However, if H_0 is held constant and the sample is moved between the centers of coils 1 and 2, the Δ_1 term becomes negligible (assuming no coil motion) and $\delta\Phi/\mu_0 = 2\Delta_2M$. This illustrates one of the central benefits of combining the motion of the sample with flux measurements—the pickup coils sense the entire field volume; however, when the sample is moved, only the differential change in the flux distribution due to the sample motion is sensed.

1.2 Force Detection

A force on a sample is produced by a fixed or controlled gradient field, usually constant over the sample, and the force is detected by a sensor attached to the sample. The force is given by

$$F = \left(\frac{v}{2}\right)\frac{\partial}{\partial z}(MH - \chi_0H^2), \tag{3}$$

where the force F is along the z direction, v is the sample volume, and χ_0 is the susceptibility of the medium surrounding the sam-

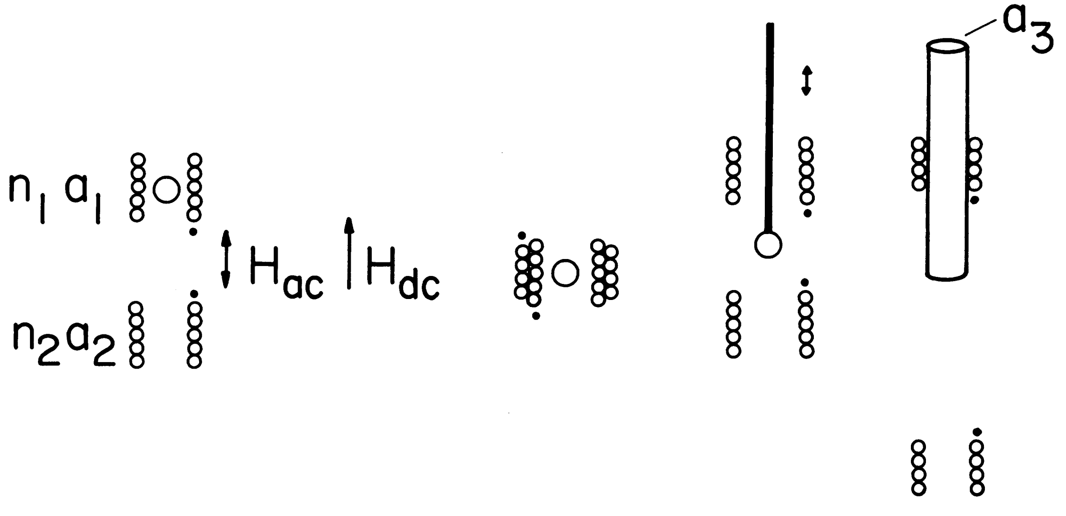

FIG. 1. Cross sections of different pairs of pickup coil configurations for flux detection discussed in the text. Each pair has coils of equal area-turns and can be connected in series opposition. The ac and/or dc fields are applied parallel to the coil axes. The positions for samples for maximum flux detection are shown in (a), (b), and (d); (c) shows a support rod which vibrates a sample parallel to the dc field (called the axial configuration for a vibrating sample magnetometer) appropriately positioned between the detection coils.

ple. The assumptions are that M is isotropic, the sample is sufficiently small, the applied field is not influenced appreciably, and the sample displacement is negligible. If the susceptibility of the sample, χ, is independent of H, then $M = \chi H$ and F is proportional to $H\partial H/\partial z$ so that both H and the gradient must be known to determine χ. However, if M is field independent (the sample is magnetically saturated), then only $\partial H/\partial z$ need be known. Highly anisotropic samples can give anomalous results (Wolf, 1957), because F also involves components of the field gradients along the x and y directions, which are always present and neglected here.

1.3 Mechanical Torque Detection

A magnetic material that is anisotropic as a result of magnetic or shape anisotropy will have a mechanical torque exerted on it if placed in a uniform magnetic field that is not along its preferred direction (easy axis). The torque is given by $T = -dE/d\theta$, where E is the anisotropy energy and θ is the angle between the field and minimum-energy direction (easy axis). For a single crystal, $E(\theta)$ has the functional dependence of the symmetry of the crystal structure. Extremely small torques can be detected, so that even small anisotropies in paramagnetic or diamagnetic materials can be measured. The torsion device measures a differential magnetic property, although it can be used to measure the total moment in special configurations, e.g., the torsion device can be used to measure forces on a sample in an inhomogeneous field. It should be noted that here a distinction is made between "mechanical torque" measurements and measurements of the vector components of the magnetization, which yield equivalent information and are often also called torque measurements. Both approaches are discussed below.

2. APPLICATIONS TO MAGNETIC MEASUREMENTS

2.1 Flux Measurement

2.1.1 dc With the coil arrangement in Fig. 1(a) discussed in Sec. 1.1, the integral of the change in voltage due to a change in Φ is $\int E dt = -\int d\Phi$, which can be measured by an integrating digital voltmeter or an electronic integrator circuit. If the series-opposing coils are well balanced, the output of the integrator is proportional to the change in M and can be plotted directly as a function of H_0 or of stepped changes of H_0 in order to determine $M(H_0)$. The integration time constant of the integration circuit must be sufficiently long compared to the time of measurement in order to avoid significant errors. One serious limitation is the inherent drift that occurs in dc integrators due to thermoelectric and other voltage drifts at the input. To calculate $M(H)$ from the $M(H_0)$ data, the demagnetization factor $\mathbf{N}$ must be known (cf. discussion in the introduction) and the output must be either calibrated by comparing with a known sample of the same dimensions at saturation or calculated knowing the coil and sample geometry. If the sample is sufficiently small compared to the coil dimensions, the sample may be treated as if it were a dipole.

One form of a moving-sample measurement is the simple extraction method in which the sample is rapidly removed from a pickup coil at fixed H_0 and the flux change corresponding to $\Delta_2\delta M$ in Eq. (2) is measured by flux integration. The maximum signal (and minimum sensitivity to sample position) is obtained for the sample starting initially centered in the pickup coil. Mutually perpendicular pickup coils transverse to the field can be added to sense any component of magnetization that is not parallel to the applied field. Highest sensitivity, as with all pickup coil arrangements, is achieved by maximizing the coupling or filling factor of the coils, consistent with experimental constraints. The nature of the extraction method restricts it to point-by-point measurements, but with computer automation, data logging becomes routine.

Although the signal is less sensitive to sample position when the sample is at the center of the pickup coil, for high-resolution experiments centering is still a problem. Very careful pickup coil design can reduce this sensitivity to position, but one still must include effects of thermal expansion of the supporting rod and forces on the sample and support rod, among other effects. A solution is to move the sample past the center of the pickup coil (recognized by the maximum of the integrated flux) to determine the magnetization. This technique was used in a very low-frequency vibrating-sample magnetometer, where the sample cycled beyond each of the centers of the pickup coils and the integrated flux was detected with a peak detector, allowing high-resolution differential sensitivity (Foner and McNiff, 1968). When the pickup coil has few turns, such as in a SQUID magnetometer (discussed in Sec. 2.1.5), it is very difficult to position the sample correctly. Either peak detection of the integrated flux as discussed above or analysis of the flux profile vs sample position is used for accurate data analysis.

2.1.2 External Variables If the magnetic properties can be affected by external variables such as light, heat, etc., with response times short compared to the integration time, the differential changes in M can be measured with high resolution. For instance, at low temperatures, the specific heat is small so that the very small changes of M versus T can be measured in ferromagnetic materials if the temperature is changed by a pulse of heat and the differential magnetic flux change measured. This technique was used by Pugh and Argyle (1962) to measure the very small temperature-dependent changes in magnetization of nickel with very high resolution.

2.1.3 ac Susceptibility If the applied field is modulated with a sinusoidal wave form, the (differential) ac susceptibility dM/dH_0 is measured. In order to balance out the large H_0(ac) background, series-opposing coil arrangement such as that in Fig. 1(a) can be used or, as an alternative, concentric coils, as in Fig. 1(b), can be employed where again $n_1a_1 = n_2a_2$ but $\Delta_2\delta M$ is proportional to $n_2 - n_1$. Generally, a further fine adjustment is needed in order to cancel the applied ac field to high order, and this can be accomplished with an additional coil of small area-turns (not shown in Fig. 1) placed in the field region. A portion of the ac signal of the additional coil, taken from a resistance divider across that coil, is added for compensation to that of the larger balanced-coil pair with the appropriate polarity. When the sample is placed in one of the detection coils, there is an in-phase component, proportional to the sample M, and a 90 deg out-of-phase component reflecting any losses in the sample, which includes magnetic losses and eddy-current losses; the apparent ac susceptibility is $dM/dH_0=\chi_a=\chi'_a + i\chi''_a$. Corrections for N_m yield the internal susceptibility χ. Measurement of the imaginary component as a function of frequency allows observation of relaxation phenomena as well as the conductivity of the sample. Unless the sample is sufficiently small and/or the conductivity low, the ac field will be nonuniform in the sample and the effective values of N may be quite different than assumed. One solution is to reduce the sample size so that it is small compared to the skin depth at the operating frequency. The voltage detected is proportional to the frequency and the number of turns of the pickup coil so that a large number of turns improves the signal-to-noise ratio at low frequencies but limits the high-frequency response because of capacitive coupling between turns. Usually a lockin amplifier is used for phase-sensitive detection and amplification of both ac components. Very important for ac susceptometry is accurate setting of the phase, which can be affected by changes in circuit parameters of the pickup coils or other neighboring elements near them as a function of temperature, field, frequency, etc. As shown in Figs. 1(a) and 1(b), both ac and dc fields can be applied in order to measure the field dependence of the ac susceptibility. Just as for flux integration discussed earlier, the background signal due to an unbalance of the series-opposing coils can be minimized at a fixed dc field by moving the sample between the centers of the two coils for each point or by removing the sample entirely from the pickup coil. This moving-sample technique results in a point-by-point measurement. In order to take advantage of the full range of the lockin, the background ac field must be well compensated in the detection coils. Typical ac susceptibility instruments operate from a few Hz to a few kHz, although specialized arrangements can be used at much higher frequencies. A moment sensitivity of 5×10^{-10} A · m^2 [5×10^{-7} emu] was achieved by Goldfarb *et al.* (1992). The amplitude of H_0(ac) is limited by power dissipation in the ac magnet. Although ac susceptibility measurements are often carried out in no dc field and with relatively small ac amplitudes (less than 10^{-3} T), instruments are available with options for high fields produced by superconducting solenoids. One commercial ac susceptometer uses a sample extraction method and a fast digital integrator to adapt an ac susceptometer to dc operation, with a sensitivity of 5×10^{-8} A·m^2 (Krause, 1992).

2.1.4 Vibrating-Sample Magnetometer

The sample can be moved in a periodic manner to obtain an ac flux change due to the periodic motion of the field distortion produced by the sample. This device, called a vibrating-sample magnetometer (VSM), has become a widely used instrument for measurement of magnetic properties of materials (Foner, 1959) and is an ASTM standard. The voltage detected by a suitable pickup coil arrangement is given by $E = G2\pi fA\sigma \cos 2\pi ft$, where G depends on the pickup coil geometry, f is the frequency of vibration, A is the amplitude of a sinusoidal vibration, and σ is the magnetic moment (usually assumed to be a dipole when the sample is small compared to the pickup coils). Usually an electromechanical transducer imparts a periodic linear motion to a supporting rod attached to the sample. An amplitude of 250 μm peak to peak is adequate to obtain high sensitivity when a lockin amplifier is used for signal processing. When a feedback signal is used to balance the ac signal E at the pickup coils, the VSM is insensitive to changes in f and A; otherwise, both must be stabilized. The simplest pickup coil arrangement for the VSM is shown in Fig. 1(c), where the applied dc field is parallel to the sample vibration axis (as for a superconducting solenoid), because in this case, the detection coils have the axial symmetry of the sample dipole. The moving dipole generates an ac quadrupole field. If the sample is at the center of one pickup coil, as for the ac susceptometer above [Fig. 1(a)], the signal at f is exactly zero; however, displacing the sample from this position yields a signal, but E is then strongly dependent on position. The appropriate sample position is symmetrically between the two coils, as shown in Fig. 1(c), where E is position independent to first order. By adjusting the distance between the coils, one can obtain either the maximum in E or an E which is least sensitive to sample position with a slight ($\cong$25%) reduction of E. Calculations can be done in closed form for this axial case, where the applied field and sample vibration direction are parallel, as in Fig. 1(c), but not for the transverse cases, where the vibration direction and applied field are perpendicular to each other, as shown in Fig. 2. For a moment oriented transverse to the vibration axis, generally the case for an electromagnet, less obvious coil arrangements are required, as shown in Fig. 2. Generally, these coil arrangements are pairs or a multiplicity of pairs of coils (there are an infinite number of possible ones) that can be used to detect the ac signal, and combinations can be used to measure the moment in any direction. The

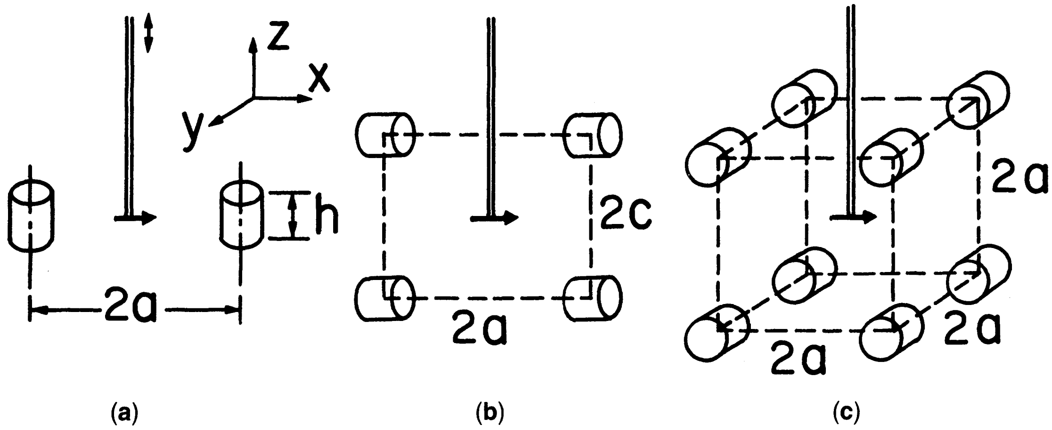

FIG. 2. Examples of transverse detection coil configurations for a vibrating-sample magnetometer, where the sample vibration along z is perpendicular to the dipole moment of the sample, along x. The axes of the coils are directed along z, x, and y for (a), (b), and (c), respectively. The appropriate position of the sample is shown for each case. [After Zieba and Foner (1982), courtesy of American Institute of Physics.]

appropriate coil arrangements follow from symmetry arguments. Figure 2(a) shows an arrangement of two coils oriented with their axes perpendicular to the applied field (Foner, 1959); Fig. 2(b) shows a four-coil arrangement with axes parallel to the applied field (Mallinson, 1966). This arrangement also is commonly used in commercial VSMs. The arrangement of eight coils in Fig. 2(c) (Bowden, 1972) intercepts flux with components along the x, y, and z directions. Combinations of these coils can be connected to detect these components of σ in the x, y, and z directions so that the result is a vector magnetometer. Similarly, an arrangement can be used where all the coils in Fig. 2(c) are rotated by 90 deg. Bernards *et al.* (1990) show configurations of eight coils parallel to the field and Bernards and Cramer (1993) show 12 coils parallel to the field. In all cases, the output is insensitive to sample position (to first order) by symmetry when the sample is centered symmetrically between the coils as indicated. Detection coil geometries, sensitivity functions, and sample geometry effects are discussed by Zieba and Foner (1982) and Pacyna and Ruebenauer (1984). VSMs have been adapted to operate in conventional iron-core electromagnets, superconducting magnets, high-power water-cooled magnets, and hybrid dc magnets up to 30 T. Commercial instruments quote a sensitivity of 5×10^{-8} to 10^{-9} A · m^2 [5×10^{-5} to 10^{-6} emu], and a small-sample sensitivity of 10^{-12} A · m^2 has been reported (Foner, 1975), where a small sample is coupled efficiently to small detection coils. The VSM also can measure ac susceptibility at very low frequencies (10^{-3}–10^{-1} Hz) (Guyot *et al.*, 1980). An ac modulation field is superimposed on the vibrating sample at frequency f_0 much less than the vibration frequency and the VSM time constant to obtain the output of the VSM at f_0.

2.1.5 Superconducting Quantum Interference Devices (SQUIDs) Recently, superconducting quantum interference devices (SQUIDs) (*q.v.*) have been incorporated in magnetic material flux-measuring devices. A SQUID can be considered as a very sensitive flux detector operating at low temperatures (4.2 K), which has an extremely low input impedance and a wide frequency response from dc to beyond the GHz range. The devices have been applied to a wide variety of applications, most involving very low-field applications rather than measurement of magnetic materials. Reviews of SQUID principles of operation are given by Clarke (1977) and Webb (1972). An example of a dc system adapted for very low frequency ac field modulation (Sager *et al.*, 1993) is shown in Fig. 3 and illustrates the basic features of dc SQUID magnetometers as well. The circuit loop, including the detection coils and L_1 and L_2, is superconducting. Consider first the dc operation, where the sample is translated in the detection coils. The change in flux at the detection coils produces a current which flows through L_1, and the field generated by L_1 is sensed by the SQUID. Any change in the operating point of the SQUID is sensed by the rf circuit and phase detector, which filters the modulated rf signal and sends back a dc current, coupled through L_2 in order to buck out the current in the detection coils. In this way, the SQUID is maintained at a fixed operating point (well within one flux quantum). The system operates as a null detector, and the (ac or dc) voltage at R_{fb} is a measure of the flux change at the pickup coils. The pickup coil is composed of two carefully balanced, series-opposed coils. In order to reduce sensitivity to background gradient fields, an additional set of series-opposing coils is added. Note that the SQUID here senses ϕ rather than $d\phi/dt$, so that the output is independent of f down to dc. Efficient coupling of the flux to the SQUID, which has a low inductance, requires a low-inductance pickup loop (few turns). If the sample is stationary in the detection coil and the ac drive coil is turned on, the ac signal at this frequency passes through the phase detector (which filters out the higher modulation frequency) and appears across R_{fb} and L_2. The ac voltage across R_{fb} is proportional to the ac flux change and is fed to a lockin amplifier. The lockin amplifier output gives the real and imaginary ac components of the χ_a as for the conventional ac susceptibility measurements but is independent of frequency from 0.1 Hz to 1 kHz. The same arrangement operates as a dc device when no field modulation is applied. In both cases, the sample is extracted or moved between the two pickup loops to discriminate between the flux change due to the sample and background effects. The shielded box (dashed line in Fig. 3) with the SQUID also is kept at 4.2 K, and the SQUID and superconducting leads must be

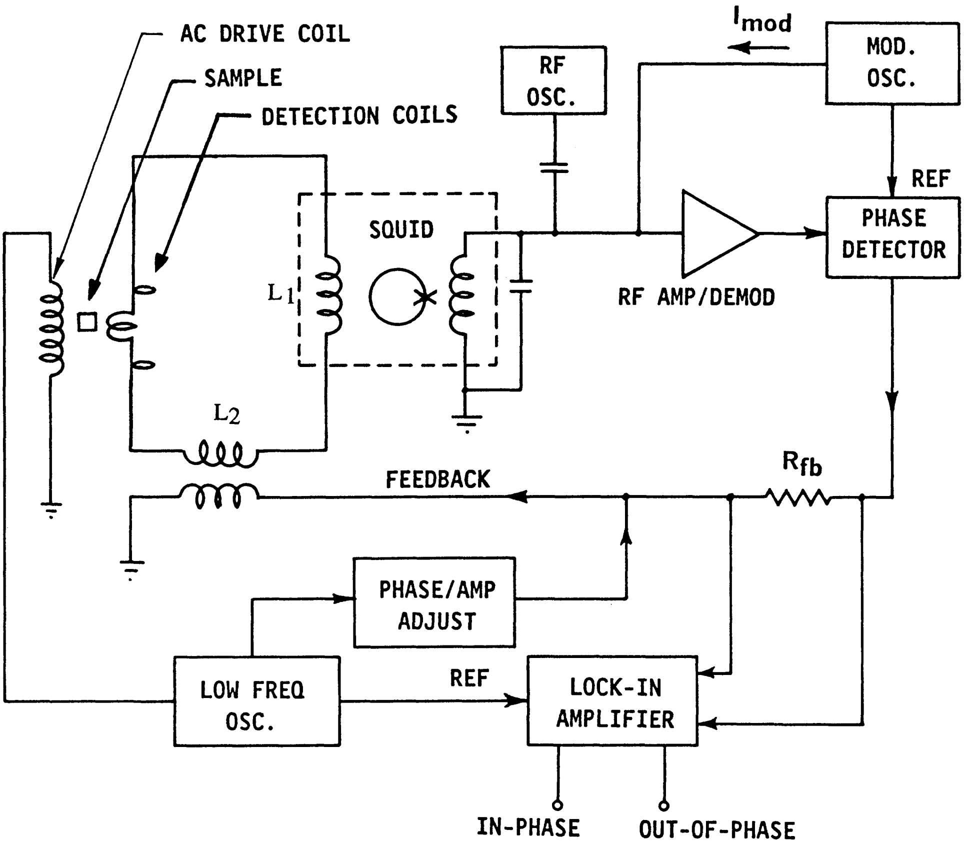

FIG. 3. Block diagram showing elements of dc flux detection SQUID magnetometer adapted for ac susceptibility measurements extended to very low frequencies. The circuit containing L_1, L_2, and the detection coils is superconducting. The operation is summarized in the text. [After Sager *et al.* (1993, ©1992 IEEE), courtesy Institute of Electrical and Electronic Engineers.]

shielded from influences of any stray magnetic fields. Such SQUID magnetometers have been manufactured for operation using superconducting solenoids. Commercial dc SQUID systems quote sensitivities of 2×10^{-11} $\mathrm{A \cdot m^2}$ [2×10^{-8} emu].

2.1.6 Pulsed Magnetic Fields Small-amplitude pulsed magnetic fields are used widely to encode digital magnetic storage media (see MAGNETIC STORAGE OF INFORMATION). These fields must overcome the coercive field of the medium, generally well below 0.1 T, but must operate at high repetition rates and generate the fields over a well-defined, extremely small volume. The flux change of the recording medium is read by the same encoding magnet or other flux sensor. Although the magnetization of the medium is large, the volume detected is very small. Ameliorating the measurement of the magnetic properties is the fact that a precise measurement of the moment is not required; only the sign of the moment is required (with high reliability).

A second application of pulsed magnetic fields is in the charging of permanent magnets. What is required is the application of a field well above the saturation of the permanent-magnet. This can be accomplished in some cases by using an electromagnet. The required field for large or complex permanent-magnet structures can be generated inexpensively by a pulsed magnet. A small (often portable) capacitor bank is used to store the energy and discharges a large current through a special fixture or a winding placed around the piece to be magnetized. Fields of 1.5–2 T are adequate for most applications, and these are generated by commercial units.

Evaluating the magnetic properties of materials often can be done in iron-core magnets (which reach 2 T in small air gaps) or in more complex superconducting magnets, which are readily available in various bores to 9 T and in smaller bores up to 20 T. Beyond this field range, central user facilities are available. The highest continuous (dc) field generated to date is 34.2 T in a 33-mm room-temperature bore and 37.2 T furnished by holmium pole pieces at 4.2 K or lower in a 2-mm gap at the Francis Bitter National Magnet Laboratory at MIT. This field is generated in a hybrid magnet, which is a combination of a high-power water-cooled magnet surrounded by a large superconducting magnet. A 40-T hybrid is under construction at the National Research Institute for Metals, Tsukuba, Japan, and a 45-T hybrid is under construction jointly by MIT, Cambridge, Massachusetts, and Florida State University, Tallahassee, Florida. Many conventional measurement techniques can be readily adapted to these high dc field environments.

Fields above available dc fields are generated inexpensively on a pulsed basis, and for the highest fields, pulsed fields are the only technologically feasible approach. At this time, nondestructive pulsed fields of 70 T have been produced in volumes of about 10^{-6} m^3 [1 cm^3] for pulse durations of about 10 ms. Pulsed fields for durations of a few μs are produced in destructive coils at fields above 100 T. The field volumes of pulsed-field magnets are generally smaller and less homogeneous than for dc fields for technical reasons. Examples of dc and pulsed central facilities are summarized in Herlach and Franse (1989), pp. 3–125, and Foner (1990), and new facilities are being developed worldwide.

Most practical properties of magnetic materials can be measured in fields below 2 T. Studies of new materials often require much higher fields in order to address fundamental mechanisms. Magnetic measurements are more restricted in pulsed fields, because thermal equilibrium is more difficult to achieve in the sample as a result of eddy currents, hysteresis losses, and other magnetothermal effects. The thermal effects are more significant at low temperatures, because the heat capacity of materials becomes quite small. Inductive measurements in pulsed fields are made in carefully balanced pickup coils, which can be designed to cancel the large transient background field to high order. Sensitivity is limited by the background residual field unbalance and by the short time available to average the data. On the other hand, practical magnetic materials have large magnetizations so that they can be measured more readily than weakly magnetic materials. Phenomena that are nonlinear in B are more readily detected, because that induced voltage is then not proportional to the background signal of the applied field so that these differential measurements discriminate against the remaining background signal more effectively (see, for example, Foner *et al.*, 1989). Of course, as a further step, the sample can be removed from the pickup coils and the experiment repeated in order to subtract the detected unbalanced background pulsed field signal from the data, as one does for the various inductive measurement techniques discussed earlier. The effects of the inductive background pickup also can be reduced or eliminated by using a noninductive sensing technique—for example, if the magnetic phenomena can be measured by optical techniques using the Faraday effect or Kerr effect (cf. Sec. 5). The pulsed field must vary slowly enough to minimize induced eddy currents and the resultant inhomogeneous field in highly conducting samples. Here, the alternative is to reduce the sample dimensions. Another is to employ very long–time-constant pulsed, or quasistatic, fields with pulse times approaching 0.1 s or more, which is intermediate in cost and complexity between the long-pulse (5–20 ms) and dc regimes. Quasistatic fields of 40 T have been produced for more than 20 years (Roeland *et al.*, 1989) for a variety of basic high-field magnetic measurements. Pulsed fields present a challenge, but within their limitations, many basic studies of magnetic materials are possible. Studies of magnetic materials with pulsed fields ranging from a few T to well over 100 T as well as with dc fields often are reported at specialized high field conferences, see, for example, Herlach and Franse (1989), pp. 199–319, and deBoer *et al.* (1992).

2.2 Force Measurement

2.2.1 dc Gradients The force on a sample is proportional to $H dH/dz$, or $M_s dH/dz$ when the magnetization reaches saturation.

The force method has a very long history and is often referred to as the Faraday method. To obtain the gradient field in an electromagnet, the pole pieces may be shaped to give an appropriate dH/dz. Large values of dH/dz (400 kA/m^2) are obtained; however, the gradient changes as the field is changed, because the permeability of the pole pieces changes with field. An alternative is to install independently controlled gradient coils at the center of the field in the air gap (Lewis, 1971, who quotes a Faraday balance with a sensitivity of 10^{-9}–10^{-10} A·m^2 [10^{-6}–10^{-7} emu] using $dH/dz = 3.85$ kA/m [485 Oe/cm]). The dH/dz is limited by the power dissipation of the gradient coils. The current in the gradient coils can be reversed giving twice the force, and they can be driven in a periodic manner. Higher fields (and gradients) can be generated in superconducting solenoids. The sample is usually positioned for maximum force (at a saddle point of the field) in order to minimize effects of changes in sample position. The force produced by the gradient field can be measured by an electromagnetic balance or any number of other force transducers. The sample is unstable to lateral motion, but this motion can be reduced by adding a hang-down mass, which is attached below the sample by a fine fiber to produce a lateral restoring force. Stewart (1969) analyzed the lateral instabilities and limits of forces on a sample in the air gap of conventional iron electromagnet arrangements and later for those in superconducting magnets (Stewart, 1975). He also points out that some care is necessary in the design of the superconducting magnet and that the gradients of both the magnet and the gradient coils must be considered. The force due to the independently controlled gradient coil can be measured by turning it off or reversing the current in it. Hill (1968) analyzed the limitations on sample size and position effects for the Faraday method. In principle, the sample is in a nonuniform field so that H must be averaged over the sample volume, and both H and dH/dz also must be measured in order to determine M. Despite some limitations, the force method has been used for many years quite effectively.

2.2.2 ac Gradients When a dc measurement can be supplanted by an ac method, there are opportunities for increased sensitivity and resolution. Zilstra (1970) measured the properties of microscopic single-domain magnetic particles by attaching them to a low-mass vibrating reed, excited at mechanical resonance to increase the amplitude of deflection by an ac field gradient, and observing the deflection of the reed under a microscope. In an improvement, Roos *et al.* (1990) used a piezoelectric force-to-voltage transducer whose output was fed to a lockin amplifier, increasing the sensitivity (by a factor of 100) to 10^{-13} A·m^2 for small samples. Various modifications have been reported recently (Flanders, 1990, 1988; Richter *et al.*, 1988). An example of a vertical-force device developed by Flanders (1990) is shown in Fig. 4. The ac gradient coils produce a gradient field directed along z so that the sample couples a vertical force to the support fiber. This vertical force produces a bending moment on the piezoelectric (bimorph) transducer, which in turn generates an ac voltage corresponding to the (small) displacement of the sample. In order to increase sensitivity, the frequency of the ac gradient field is adjusted for mechanical resonance of the structure. The structure is decoupled mechanically from the supports by the rubber pad and bands. The transducer output is fed to a lockin amplifier for data processing. The supports for small samples are very delicate at high sensitivity; larger fibers are used for larger magnetic moments. Because the sample and small addenda change the resonant frequency, the device must be tuned for each measurement, and the suspension is sensitive to inadvertent damping. Larger samples generally require more robust reed structures, which reduce the sensitivity, and it is difficult to adapt the device for measurements at low temperatures. A commercial ac gradient system quotes a sensitivity of 10^{-11} A·m^2 for small samples.

2.3 Mechanical Torque Measurement

Torsion devices have a long history of development, can be extremely sensitive, are employed in many areas other than magnetism, and are commonly used to determine anisotropy energy in magnetic materials. An extensive recent review of torsion devices and various applications is given by Gillies and Ritter (1993). A detailed discussion of an-

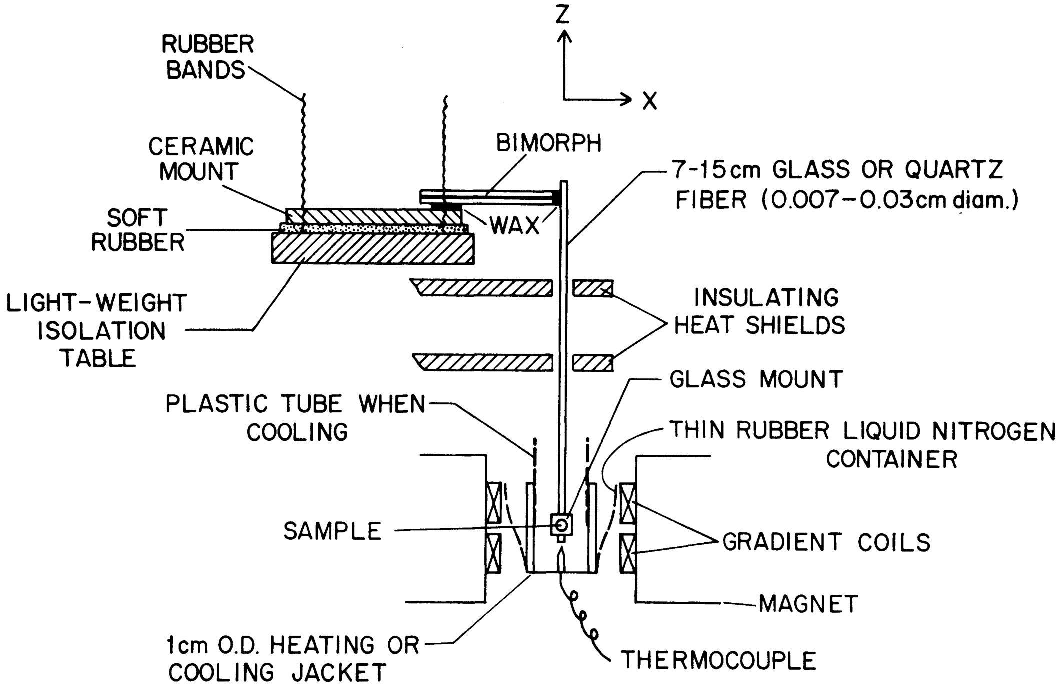
Z
X
RUBBER BANDS
CERAMIC MOUNT
BIMORPH
7-15cm GLASS OR QUARTZ FIBER (0.007-0.03cm diam.)
SOFT RUBBER
WAX
LIGHT-WEIGHT ISOLATION TABLE
INSULATING HEAT SHIELDS
GLASS MOUNT
PLASTIC TUBE WHEN COOLING
THIN RUBBER LIQUID NITROGEN CONTAINER
SAMPLE
GRADIENT COILS
MAGNET
1cm O.D. HEATING OR COOLING JACKET
THERMOCOUPLE

FIG. 4. Arrangement for vertical-force alternating-gradient magnetometer. A vertical force (along *z*) is applied to the sample by the ac gradient coils and transmitted to the piezoelectric element (bimorph). The structure operates at mechanical resonance. The dc field is applied along *x*. A smaller-diameter sample support rod is used for smaller samples. [After Flanders (1990), courtesy of the American Institute of Physics.]

isotropy measurements and a comparison of torsion balances for magnetic measurements are given by Pearson (1979). The sample assembly is supported by a torsion element, and the sample is positioned in the uniform magnetic field usually furnished by an electromagnet. The magnet or torsion system is rotated, and the torque acting on the sample is counteracted by a transducer, for example, an electrobalance, in order to maintain the original sample orientation. The torque, which is proportional to the feedback signal, is recorded as a function of rotation angle. A classic instrument described by Penoyer (1959) and shown in Fig. 5 has the basic elements of a torquemeter. The entire weight of the suspension is supported by the bearing G, and lateral motion of the assembly is restricted by the air bearing. A rotation of the magnet produces a torque on the sample, and rotation of the mirror causes an unbalance of currents in the phototubes, resulting in a current in the compensating coil at the top of Fig. 5, which produces a compensating torque in the permanent magnet field. This instrument has a sensitivity of 10^{-5}–10^{-2} N m [10^2–10^5 dyn cm], useful for rapid measurement of a wide variety of magnetic materials. Higher sensitivity can be achieved by replacing G by a supporting torsion fiber and eliminating the air bearing, with the resulting greater sensitivity to physical instabilities. The sample and its mounting must be axially symmetric with the torsion fiber in order to minimize instrumental anisotropies. If the sample is not well centered in the field, there is a transverse force that will displace it. It is difficult to vary the temperatures over a wide range. At low fields, some orientations are unstable unless the torsion restoring element is properly chosen, i.e., increased in strength which reduces sensitivity. A wide range of sensitivity is required to cover conventional ferromagnetic bulk samples as well as thin films. Until recently, torsion balances were not available commercially for conventional measurements.

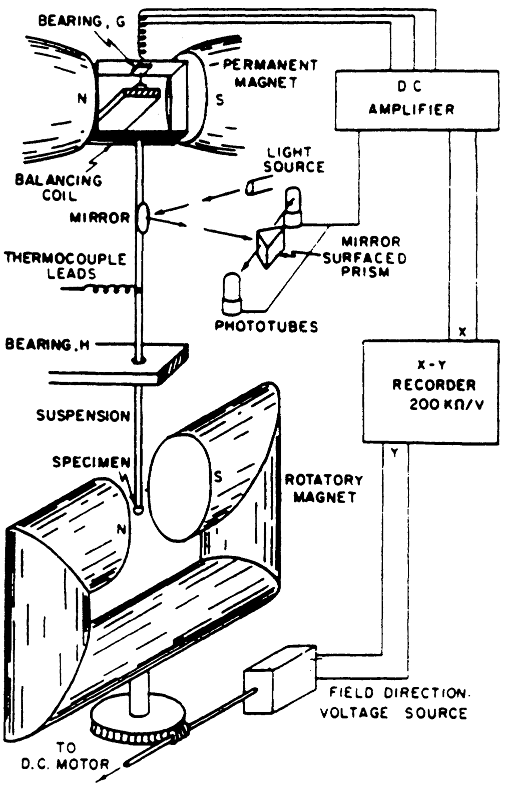

FIG. 5. A mechanical torque balance developed by Penoyer. Torque on the anisotropic sample is counteracted by the torque between the current in the balancing coil and permanent magnet at the top of the figure. Alignment of the rotating elements is maintained by an air bearing, H, and sapphire bearing, G. [After Penoyer (1959), reprinted with permission, the American Institute of Physics.]

3. MEASUREMENTS OF MAGNETIC PROPERTIES OF PRACTICAL MATERIALS

Most practical magnetic materials are ferromagnetic or ferrimagnetic, called strongly magnetic, and present special problems in determining the technical parameters of interest in applications. The quantities of interest for practical materials depend on the application. High-frequency coils and filters used in communications require low loss and high permeability at a very low field and operate in the demagnetized, reversible region of the initial magnetization curve. Relays require materials with a low coercive field H_c, high saturation magnetization M_s, and high reproducibility. Radio-frequency and power transformers operate on hysteresis loops below saturation (inner loops) in order to have low loss and high permeability. Magnetic memory materials require high remanence B_r and square hysteresis loops. Examples of methods to determine these parameters are discussed below.

3.1 Magnetization

The demagnetization field makes it difficult to determine $M(H)$ at low fields. Ideally, an ellipsoidal sample should be used. However, M can be measured with good accuracy for a long bar with a length much greater than the diameter, as shown in the top section of Fig. 1(d). The flux detected by a short pickup coil with area a_1 that fits closely at the center of the bar is given by $\Phi/n_1 = Ma_3 - N_fMa_1 + \mu_0H_0a_1$, where a_3 is the area of the bar, N_f is the fluxmetric demagnetization factor, and H_0 is the uniform applied field. The flux is detected at the central region of the bar where the M is uniform. $B(H_0)$ is determined by integrating each $\Delta\Phi$ for a series of independently measured changes in field $\Delta(H_0)$ or continuously as the field is swept. By connecting a series-opposing coil to the top coil with $n_2a_2 = n_1a_1$ [Fig. 1(d)], far from the sample but in the applied uniform field H_0, H_0 may be canceled, so that in this case, $M(H)$ is measured. Because N_f depends on the susceptibility of the sample, detailed corrections are required where N_f changes, although microcomputers can alleviate this chore. However, the important parameters, B_r or M_r, M_s, and H_c can be determined accurately here, because corrections for H_d are small near saturation and well defined when μ is large near H_c. Errors of several percent may occur at the intermediate field range, e.g., when evaluating the maximum energy product in the second quadrant $(BH)_{\max}$, if corrections for N_f are not made. This measurement is useful for permanent magnets. When the hysteresis loop is measured in the second quadrant, the intrinsic coercive field H_{ci} is obtained at $M(H) = 0$; it is always greater than H_c obtained at $B = \mu_0(H + M) = 0$ (and much greater for modern permanent-magnet materials). Although H_c and H_{ci} are obtained in the second quadrant, they are generally quoted as positive quantities.

Any of the flux measurement systems discussed above that employ periodic sample motion, including the VSM, SQUID, etc., can be used to measure $M(H_0)$. The sensitivity of these devices is quite adequate for the smallest samples. If the shape is an ellipsoid, preferably a sphere, or a thin plate or thin long rod so that N_m, the magnetometric demagnetization factor, is known, accurate corrections can be made. For a spherical sample or one small compared to the distance to the detection coils, the sample can be assumed to be a dipole to high approximation. Measurement of σ, the magnetization per gram, is usually made because it is convenient to weigh the sample, and conversion to M is made if the density is known or measured. High-resolution measurements of $\sigma(H_0, T, \text{etc.})$ can be made routinely, and when controlled by a microcomputer, corrections for N_m can be programmed so that $\sigma(H, T, \text{etc.})$ is obtained. M_s, B_r, and H_c can be determined with high accuracy. Unless the shape of the sample is a spheroid, M is not uniform and corrections for $N_m(H)$ must be made at intermediate fields.

Although it is possible to calculate the calibration constants of the moving-sample type of instrument, it is more practical (and generally more accurate) to use a known standard for calibration. Very often a small nickel sample is measured above magnetic saturation to determine the σ scale, because nickel saturates at a relative low field so that calibration can be done in conventional electromagnets. It is stable in the atmosphere, has a large σ_s that is not strongly temperature dependent (increasing by 6.4% from 300 to 4.2 K and by 0.3% from 77 to 4.2 K) , is inexpensive, and can be obtained readily at high purity. Assuming that the instrument is linear, only a single point is needed to calibrate it at σ_s. Magnetic saturation of the nickel sample is chosen so that the calibration is independent of field or N. Using a calibration sample also corrects to first order for image effects that may be induced by the sample in neighboring highly permeable materials, e.g., Fe pole pieces or a superconducting magnet (Zieba and Foner, 1983; Zieba, 1993). (The images distort the free-space sample flux and can cause small changes in the flux detected. Any *field-dependent* image effect is expected to be a second-order effect.) Of course any well-defined test sample may be used for calibration with added complications. Use of a paramagnetic material requires measuring both the field and the temperature so that corrections can be made for the temperature-dependent susceptibility. Measurement at low temperatures, in high fields, or too near to transition temperatures requires corrections for the field dependence of χ as well.

If the relative permeability $\mu_r = B/\mu_0 H$ of a material is high ($\mu_r > 100$), the apparent permeability is $\mu_a = B/\mu_0 H_0 = 1/N$ and is independent of the material. This special circumstance may appear to be uninteresting, but if the N_m is known accurately, it can be used to calibrate these instruments because the initial slope of the magnetization curve is $M/H_0 = 1/N$. A measure of H_0 at one point calibrates the system (independent of the material) to the accuracy that **N** is known.

3.2 High Permeability

Generally, the apparent permeability μ_a is measured and the relative permeability is deduced. The relation between the two depends strongly on N as follows from $\mu_0 H = \mu_0 H_0 - N(B - \mu_0 H)$, and substituting $\mu_r = B/\mu_0 H$, we obtain $\mu_a = B/\mu_0 H_0 = \mu_r/[1 + N(\mu_r - 1)]$. It follows that a very high relative permeability μ_r is difficult to determine by the above methods, because the demagnetization corrections must be very small or known with extremely high precision. Measurement of a large μ_r requires eliminating the demagnetization fields. This can be accomplished by using a sample made of a closed ring so that no free poles are induced by the applied field, as shown in Fig. 6. The field is generated by a current in a coil closely and uniformly wound around the ring sample, and B is sensed by a short overwound coil on a section of the ring. This arrangement is convenient for measurement of the hysteresis of a toroidal sample. If an ac source is used, the

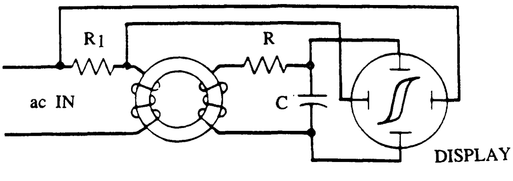

FIG. 6. Simplified arrangement of a hysteresigraph. A toroidal sample or a closed magnetic circuit is used in order to avoid demagnetization fields of air gaps. An ac or slowly varying current is applied to the primary winding, and the output of the secondary is integrated to obtain a measure of B vs H.

hysteresis can be displayed on an oscilloscope as indicated, where the voltage across R_1 is a measure of the current (and applied field) and the voltage induced in the secondary winding is integrated by the *RC* network to produce the ordinate signal proportional to B. The output can be displayed on a recorder for slow variations of the field or processed with a computer. The circuit shown in Fig. 6, with only a few turns (or a single turn) for excitation, is useful for rapid testing of cores at line frequency but lower accuracy because H in the ring is not uniform. These are the basic elements of a hysteresigraph. The magnetic circuit is replaced by a soft iron yoke or magnet with movable pole pieces, and the sample is clamped between the pole pieces in order to minimize any air gaps. The pole pieces are made with a much larger cross section than the sample cross section so that the soft iron is far from saturation during the measurement. B is measured with a tightly wound coil around the center of the sample as before, and H is measured with a small coil or other sensor as close to the surface of the sample as possible to measure the tangential component of H, which is continuous at the surface of the sample. There are many arrangements that have been developed to accommodate different shapes of samples (Bozorth, 1951). With a suitable electronic integrator, hysteresis can be measured at low or high frequencies and the data processed.

Often it is not possible to form the sample in a ring, in which case another means must be found to form a closed circuit with no air gaps. The effect of a small air gap introduced in the ring can radically change μ_a. In this case, $\mu_a = \mu_r/[l + \mu_r l_g/l]$ where l_g is the length of the air gap and l is the length of the magnetic ring. The air gap becomes more significant for large μ_r; for $\mu_r = 10^5$ and $l = 0.1$ m, a 10% error in μ_a occurs for $l_g = 10$ μm. Practical materials used for transformers are made in thin laminated strips, and it is not convenient to make toroidal samples for measurements, but these can be arranged in a closed ring. The air gap is minimized by clamping strips of the highly permeable material with large overlaps to make a closed magnetic circuit with negligible effective air gap.

3.3 Permanent Magnets

Permanent-magnet materials are often manufactured as relatively short sections. The important parameters are obtained from measurement of the second quadrant of the hysteresis loop. The demagnetization field may be eliminated by polishing the end surfaces flat and parallel so that they make good contact with large-area poles of a soft iron yoke used to generate the field. The distance between the poles is adjustable so that they can contact the permanent-magnet material with a negligible air gap. A coil closely wound around the sample measures the change of flux in the circuit produced by a current in the primary winding, which is wound around another leg of the yoke.

3.4 Losses

The area of the hysteresis (*BH*) loop represents the energy loss W_h due to irreversible processes. This energy can be determined from magnetization measurements by integrating over the area enclosed by the loop. The hysteresis areas, and the losses, increase with increasing excursions of the field until magnetic saturation is reached, at which point the saturation hysteresis loop is obtained. The hysteresis area is dependent on the rate at which the field is swept; the hysteresis loss is frequency dependent, generally increasing with increasing frequency.

Because the hysteresis loss is sensitive to various experimental parameters, a standardized arrangement developed by Epstein and illustrated in Fig. 7(a) is adopted as an ASTM standard for power transformer materials tests (Bozorth, 1951, p. 853; Cullity, 1972). An Epstein frame is assembled with four equal legs having standard dimensions and composed of fixed-length laminations of the test material. The strips overlap at the corners and are clamped together to eliminate demagnetization effects. Each leg has a primary winding overwound by a secondary winding. All the primary windings are connected in series as are the secondaries [see Fig. 7(a)]. The power loss is measured at no load, with the secondary output voltage and the primary current both fed to a watt meter, W in Fig. 7(b). The primary current furnishes the field for the watt meter. Measurements are made at line frequency (50 or 60 Hz) and fields of 1 and 1.5 T. Losses are tabulated for materials in W/kg [W/lb].

If a ferromagnetic material is *rotated* in a magnetic field, losses are observed, which increase to a maximum well below saturation

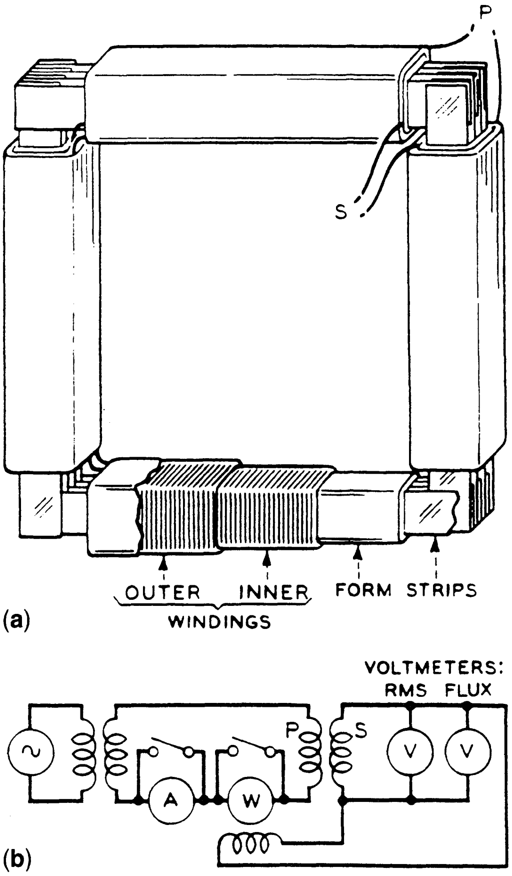

FIG. 7. (*a*) Arrangement of Epstein frame for measuring losses in a sample formed of strips of ferromagnetic material. The four primary windings are connected in series, as are the four secondaries. The input and output connections are shown in the upper right corner. (b) Circuit arrangement for measuring the power loss in the material. The primary current and the secondary voltage at no load are fed to the wattmeter W. [After Bozorth (1951), courtesy AT&T Bell Laboratories.]

and then reduce to zero. These losses are called rotational hysteresis losses and occur when the magnetization vector does not remain parallel to $\mathbf{H}_0$, so that at intermediate fields the magnetization jumps from one stable position to another. This irreversible process produces heat and is related to the anisotropic properties of the crystallites. The work required to rotate a single crystal over one cycle (2π) is W_R, the rotational hysteresis loss, and can be determined from integration of the area under the torque curve in the irreversible region. Above saturation, the integral under the curve over a full cycle is zero, because the potential energy stored in the crystal during part of the cycle is returned in a reversible manner and there is no loss.

3.5 Anisotropy

The magnetization in a single crystal has preferred directions of minimum energy, called the *easy axes*, which reflect the crystal symmetry. Rotation of the magnetization away from the easy axis stores reversible potential energy E in the crystal. A measure of this energy can be expressed as a power series of direction cosines reflecting the symmetry of the crystal, and the coefficients are called the anisotropy constants, which describe the magnetocrystalline anisotropy. These coefficients are tabulated for various crystal symmetries (Pearson, 1979). For magnetically ordered materials, the anisotropy varies with temperature, decreasing rapidly as T_c is approached. Measurement of the anisotropy involves determining these constants by magnetization, torque, or magnetic resonance measurements.

At times, it is convenient to express the anisotropy constants in terms of a fictitious anisotropy field H_A along a preferred axis of a crystal. This is an effective field that acts on M_s for small deviations from the preferred orientation given by $M_sH_A=d^2E/d\theta^2$, the curvature of the anisotropy energy surface. For a uniaxial crystal, $H_A = 2K_1/M_s$. In general, H_A involves various combinations of the anisotropy constants and is tabulated for various symmetries.

3.5.1 Magnetization Measurements

There are several ways to use magnetization measurements to measure magnetocrystalline anisotropy with conventional magnetometers, e.g., VSMs, SQUIDs, etc. These magnetometers permit convenient control of the sample temperature and applied field. In one method, the areas between $M(H)$ along the easy axis (which saturates in very low field) and $M(H)$ to saturation along the other principal axes of a single crystal are measured. These are then related to combinations of the anisotropy constants. This method is limited by the precision of the $M(H)$ data for small anisotropy, because the results are then based on small differences of large numbers. If the sample is a sphere or a flat circular disk with the principal axes in the plane of the disk, no demagnetization corrections are needed; otherwise, demagnetization corrections must be made. If the field is not applied along a principal axis, in principle, an infinite field is required to reach complete saturation, because then an effective field component perpendicular to the applied field is always present.

A special case is uniaxial anisotropy. If the crystal is oriented so that $\mathbf{H}_0$ is perpendicular to the easy axis, then $\mathbf{M}(H)$ along the field direction varies linearly with H (with a slope $M/H = M_s^2/2K$) up to saturation, which occurs at $2K/M_s = H_A$ and then remains constant at M_s. This method for determining K also is useful when the uniaxial anisotropy is very large so that the available field does not reach saturation.

The $M(H)$ data of a polycrystalline sample also can be used to determine the anisotropy if the crystallites are randomly oriented. If the field is sufficiently high, all the domain walls are removed and the magnetization process involves the rotation of the magnetization in the presence of anisotropic interactions. The approach to saturation can be written as $M = M_s\ (1-a/H^2+\ldots)$, where M_s is the saturation magnetization and a involves the anisotropy constants. Extrapolation of the high-field linear portion of a plot of M vs $1/H^2$ to the M axis yields M_s, and the slope of the linear portion yields a.

Usually the VSM (or other flux-measuring device) is arranged to detect the component of $\mathbf{M}$ along the field direction. When $\mathbf{M}$ is not aligned parallel to $\mathbf{H}_0$, the full magnetization is not detected. However, the total magnitude and direction of the magnetization can be measured, as discussed in Sec. 2.1.4., with pickup coils arranged to measure the components of $\mathbf{M}$ both perpendicular and parallel to the field. In an electromagnet, where the field is transverse to the vibration direction, a single-crystal sample is rotated about the vibration axis to obtain $M_\perp$ and $M_\parallel$ as functions of angle, or the field is swept to obtain $M(H_0)$ for the two directions. The torque can be constructed from the data, since it is proportional to $M_\perp H_0$. Very small equivalent torques can be measured with such a VSM, because the components of $\mathbf{M}$ can be measured with high resolution. (In principle, the same statements hold for other flux detection instruments if suitable detection coils are arrayed to sense the vector components of flux and if sufficient sensitivity is at hand, e.g., for SQUID magnetometers.) Demagnetization corrections must be made unless the crystal has the same N for all orientations. In recent literature, these arrangements are called vector VSMs or torque VSMs and have been applied to measuring the anisotropy of thin ferromagnetic films; see, for example, Bernards *et al.* (1990) and Richter and Hibst (1991).

The anisotropy also can be measured by a rotating-field magnetometer (RFM) described by Hagedorn (1967). A rotating field is generated by two mutually orthogonal driving coils fed by 500-Hz currents which are phase shifted by 90 deg. An ac signal is detected by well-balanced, closely coupled pickup coils surrounding the sample and fed to a lockin amplifier. The Fourier components of the output of the lockin are used to determine the anisotropy constants. Both the sample and pickup coils are stationary during the measurement, thus avoiding delicate supporting mountings or positioning problems encountered by mechanical torque techniques. The instrument can detect a moment of 10^{-10} A·m^2 [10^{-7} emu] which corresponds to an equivalent torque of one of the most sensitive mechanical torque instruments, 10^{-13} N·m [10^{-6} dyn cm], developed by Humphrey and Johnson (1964). This RFM was used to measure magnetic properties of ferromagnetic thin films but was limited to ac fields of 5×10^{-3} T.

A much earlier analog of the RFM is a rotating-sample magnetometer (RSM), where a single crystal disk is closely coupled to pickup coils and rotated at a constant frequency f_0 about the disk axis which is perpendicular to $\mathbf{H}_0$. The pickup coils are arranged to sense $\mathbf{M}$ perpendicular to $\mathbf{H}_0$, and the output voltage is fed to a lockin amplifier. Analysis of the various detected harmonics is then used to determine the anisotropy constants. Flanders (1970) lists these harmonics and describes a recent rotating-sample magnetometer with numerous applications. Advantages of the RSM over the RFM are that the RSM is quite simple and operates at much higher fields.

3.5.2 Mechanical Torque Measurements

Measurement of the anisotropy with a mechanical torquemeter is a standard method. The torque T on a crystal that is rotated in an infinite magnetic field, i.e., for $\mathbf{M}$ always parallel to the applied field, is $T = dE/d\theta = \Sigma a_n \sin n\theta$, where E is the anisotropy energy and θ is the angle between the easy axis and the magnetization. Measurement of the torque as a function of rotation perpendicular to $\mathbf{H}_0$ yields a torque curve, which is used to determine the various coefficients. For the simple case of uniaxial anisotropy, $E = K_1 \sin^2\theta$ and $dE/d\theta = K_1 \sin 2\theta$, which is a pure sine wave. The applied field is not infinite, so that **M** does not always remain parallel to $\mathbf{H}_0$ as the sample is rotated; a distorted, or sheared, torque curve results, because **M** lags or leads **H**. The $M_\perp$ contributes an additional torque on the crystal given by $HM_\perp = HM_s \sin\phi$, where ϕ is the angle between **H** and **M**, which is used to correct the data for each θ. As the field is increased, $\phi \rightarrow 0$ and the torque curve becomes more symmetric.

3.5.3 Ferromagnetic Resonance The anisotropy constants can also be determined by ferromagnetic resonance measurements at microwave frequencies. The strongly coupled electron spin system precesses about the applied field at a frequency $\omega = \gamma H_{\text{eff}}$, where $\gamma = ge/2m_ec$ and H_{eff} involves H_0, H_d, and H_a. γ is the gyromagnetic ratio, $g = 2.00$ for a free electron, m_e is the electron mass, and c is the velocity of light. Solutions of the resonance equations of motion when an rf field is applied perpendicular to the applied field give relations between fundamental constants and the effective fields acting on the spin system. These have been tabulated for various cases by Artman (1956). Usually a single-crystal sphere is mounted in a microwave cavity and rotated with the principal axes of the crystal in the plane of $\mathbf{H}_0$. As H_0 is swept, the absorption of the rf power shows a sharp maximum when the rf frequency equals the precession frequency of the spin system. The width of the resonance line is a measure of losses of the spin system, and the field at which resonance occurs is a sharp measure of the fields acting on the sample (both applied and internal). A plot of resonance field vs angle is fitted to the calculated curves expected for a given crystal symmetry in order to derive the anisotropy constants. The demagnetization fields for a sphere do not appear in the solution of the resonance equations but do for other sample geometries and can shift the resonance frequency substantially. Materials of practical interest for microwave applications are ferrimagnetic and insulating and have sharp resonance lines so that the accuracy of evaluation of the anisotropy constants is high. The anisotropy constants determined by microwave resonance may differ from that of the "static" measurements described in Secs. 3.5.1 and 3.5.2 if the relaxation time of the magnetization is frequency dependent.

3.6 Transition Temperatures

Many procedures are used to measure the transition temperature T_c, called the Curie temperature, for ferromagnetic materials. Precise predictions of T_c from first-principles calculations have not been possible. However, it is possible to predict the temperature dependence of the magnetization in the vicinity of T_c to high accuracy with modern theories of phase transitions. For practical purposes, the exact value of the transition temperature is not important but an operational and reproducible measurement technique is useful for comparing materials. Several more common methods are outlined below. There is a large field dependence of $\sigma(T)$ near T_c, so that measurements should be made at low fields.

One method is to measure $\sigma(T)$ at a temperature below T_c in a field high enough to align the magnetic domains. Assuming this value of σ is the spontaneous magnetization, a linear extrapolation of σ^2 vs T to the T axis yields T_c.

Another method measures the low-field ac susceptibility as a function of T. The anisotropy usually decreases very rapidly as T_c is approached, and a sharp maximum of χ is observed very close to T_c. A low-amplitude ac field is applied, and the peak in the susceptibility is used as a definition of T_c.

A related and convenient method uses the fact that, just below T_c, the susceptibility is very high so that $\sigma(H_0)$ at low fields is constant and depends only on the value of the demagnetization factor N, but above T_c, the susceptibility falls rapidly. Plots of $\sigma(H_0)$ vs T at different values of H_0 show a constant $\sigma(T)$ below T_c and a sharp "kink" or knee as T_c is approached. A quadratic extrapolation of the kink point vs T yields T_c. Arrott (1971) examined very precise data near T_c on a single-crystal nickel sphere and showed from model calculations that the kink-point extrapolation from very low fields (1.8 mT) leads to a small overestimate of T_c ($\leqslant 1$ K).

A more complex procedure uses the classical expansion near T_c for a ferromagnetic phase transition given by $H = a\sigma + b\sigma^3$, where $H = H_0 - N\sigma$ and includes any other contributions to the effective field and a [which varies as $(T - T_c)/T_c$) and b are temperature-dependent parameters. T_c is determined by a plot of H/σ vs σ^2 at various fixed temperatures at low fields, called an Arrott–Kouvel–Belov plot. The straight line which passes through the origin at zero field is defined as T_c. The assumptions are that H is small, yet much larger than all other internal fields and that σ varies as $H^{1/3}$. A detailed comparison of this method with the kink-point method for a dilute alloy (Praddaude and Foner, 1976) found the kink-point method more satisfactory. Modern theories of phase transitions show that the $\frac{1}{3}$ exponent is not correct , and detailed fits to determine T_c require careful fitting to a theory (Arrott, 1971). A good fit to theory for nickel was obtained by Arrott and Noakes (1971) involving a five-parameter fit.

The magnetic susceptibility of ordered magnetic materials is often described by the Curie–Weiss law where $\chi = C/(T \pm \theta)$, where C is the Curie constant and θ is the paramagnetic Curie–Weiss temperature, different from the Curie temperature T_c. Extrapolation of the linear portion of χ^{-1} vs T from the high-temperature paramagnetic region to the T axis gives θ, which occurs at $T > 0$ for ferromagnets and $T < 0$ for antiferromagnets. Usually the maximum in χ vs T is defined as the transition temperature T_N (Néel temperature) in antiferromagnets. From a practical viewpoint, this definition is useful for comparing materials, although modern theories of phase transitions show that the critical point depends on the symmetry of the magnetic structure. Accurate determinations of θ require that the extrapolations be made from data well above the ordering temperature ($10T_c$). Because χ decreases rapidly with increasing temperature, high sensitivity is required and χ^{-1} must be replaced by $(\chi + c_1)^{-1}$, where c_1 corrects for other contributions to χ such as the crystal lattice, which is usually diamagnetic, etc. It should be noted that a maximum in $\chi(T)$ does not guarantee that a material is antiferromagnetically ordered—there are other mechanisms which produce maxima in $\chi(T)$.

3.7 Magnetostriction

When a magnetic body is magnetized, there is a change in dimensions, called magnetostriction (see MAGNETOSTRICTIVE MATERIALS AND DEVICES). It is usually assumed that the magnetostriction is measured under constant external stress. In general, the magnetostriction is influenced by internal stresses, sample geometry, and temperature, among other effects. The volume magnetostriction produces an anomaly in the thermal expansion when the material passes through T_c. A linear magnetostriction occurs when the magnetic domains change orientation. If the sample is magnetized to saturation starting from the demagnetized state, where the domains are randomly oriented, the relative change in length dl/l is called the saturation magnetostriction, where l is the length of the sample. For *ferromagnets* dl/l ranges from 10^{-6} to 10^{-2}. The relative volume magnetostriction dv/v, observed as a function of field above saturation, is called forced magnetostriction. There is a form effect that depends on the sample shape and occurs because the stress distribution is not the same in the demagnetized and magnetized states. This effect is difficult to calculate but can be made small by care in choosing a sample shape (for instance, a very thin disk with thickness much smaller than the diameter). A recent review of magnetostriction is given by Lee (1979).

The standard measurement of magnetostriction employs resistance strain gauges within their elastic range, so that the relative resistance change $dr/r = kdl/l$, where k depends on the gauge characteristics and l is the length of the sample. The gauges are often made of thin wire or thin films which are dominantly oriented in one direction. The conductors are sandwiched between thin nonconducting films of paper or plastic and are glued to the surface of the sample being measured. The gauge resistances range from about 100 to 200 Ω. The resistance changes are measured in a bridge circuit and generally balanced by a gauge in a second arm of the bridge and in the close vicinity of the first (to minimize the effects of the temperature dependence of the resistivity) and with the same orientation, which balances out the changes in magnetoresistance of the gauge. Thermal expansion of the sample is large compared to *dl*/*l* so that the temperature must be controlled to 0.01 K. If dc excitation of the bridge is used, care must be used to avoid thermoelectric voltages. An ac bridge avoids this problem, but care must be taken because the gauge is in close contact with the sample and the inductance of the gauge varies as the permeability of the sample changes. The use of resistance strain gauges makes the measurement of the strain relatively easy, but correct interpretation of the data is difficult because of the numerous contributions which can affect the derived magnetostriction. These issues are discussed by Lee (1979).

The *dl*/*l* in a single crystal is measured by fabricating a thin circular disk and gluing the gauge so that it is oriented along one of the principal directions. The disk is rotated with $\mathbf{H}_0$ in the plane of the disk. Because it is difficult to achieve a demagnetized state, *dl*/*l* is measured at saturation as a function of angle. At infinite field, *dl*/*l* follows a perfect sine wave, but it is sheared at finite field because the magnetization does not remain aligned along the applied field, and corrections are needed. The maximum in *dl*/*l* is related to the magnetostriction constants for various crystallographic orientations and has been tabulated for many symmetries. Measurements also can be made in an isotropic polycrystalline thin circular disk with the gauge parallel and perpendicular to $\mathbf{H}_0$, and the magnetostriction constants can then be deduced for different crystal symmetries.

A high-sensitivity method of measuring the change in length uses the change of capacitance C between a reference conducting plate and a conducting plate or film attached to the sample with the reference plate positioned close to it. C varies as $1/t$, where t is the distance between the reference conducting plate and that attached to the sample and is made as small as possible consistent with experimental limitations in order to increase C (Fawcett, 1970). Using a sensitive ac capacitance bridge, *dl*/*l* as small as 10^{-10} can be detected.

4. MICROSCOPIC MEASUREMENTS

The microscopic arrangements of the spin distributions and their interactions may be conjectured from macroscopic measurements, but the microscopic spin distributions can be determined only by microscopic measurements. A few of the more common methods are included below.

Electron paramagnetic resonance (EPR) (*q.v.*), also referred to as electron spin resonance (ESR), is a high-resolution method for measuring the energy-level spacing of electron spins. Often the properties of a dilute system of paramagnetic ions can be related to the properties of the ordered magnetic systems. The energy levels are varied by applying a magnetic field, and when the level spacing corresponds to the energy of the radiation, a sharp absorption of the incident radiation is detected. Usually EPR is carried out from the microwave up to millimeter wavelength region. For a free electron in 1 T the spin resonance occurs at 28 GHz. EPR measurements can be extremely sensitive; when the linewidth of the resonance is small, 10^8 spins can be detected. The energy levels are affected by crystalline electric fields, which are determined by analysis of the EPR as a function of crystallographic orientation, field, and temperature. The concentration of a particular spin species can be obtained from the intensity of the absorption, and the linewidth is a measure of scattering effects. As with many of the instruments discussed here, the technology was developed for basic research and now commercial instruments are available for routine analysis. EPR has been studied for 5 decades, and there is an extensive literature. Resonance phenomena are reviewed by Ingram (1979), and instrumentation for EPR is discussed at length by Poole (1967).

The same techniques used for EPR are used for ferromagnetic resonance to measure the microwave properties of ordered magnetic systems. The magnetic materials of importance for microwave application must have low conductivity. These are generally ferrimagnetic oxides. An electron spin in a magnetic field precesses about the applied field, and the direction of rotation depends on the direction of the applied field. If a precessing electron spin is exposed to a circularly polarized electromagnetic wave rotating in the same sense at its resonance frequency, it will absorb the radiation, but this is not true for the reverse polarization. This gyrotropic behavior of the electron spin makes it possible to discriminate selectively against wave propagation in opposite directions. The magnetic material is placed in a waveguide structure with an associated (usually very compact) magnet which magnetizes the material and also establishes the resonance frequency of operation. These devices find applications from below 1 to over 150 GHz. The important properties of the magnetic materials for these microwave applications include the saturation magnetization and the width of the resonance line. Measurements of the linewidth, ac (microwave) susceptibility, and other basic ac properties are carried out using waveguide components rather than lumped circuit elements. Detailed discussions on microwave devices are given in Lax and Button (1962) and by Heck (1974).

A particularly powerful tool for obtaining microscopic information about magnetic materials is the Mössbauer effect (ME) (*q.v.*), which uses resonant absorption of γ radiation from a limited number of suitable nuclei in solids showing the ME. Fortunately, one of the most intense absorbers is the nuclide ^{57}Fe, which occurs naturally at about 2% in Fe; Fe is abundant in nature and occurs in many important magnetic and biological materials. Typically, a source or emitter is ^{57}Co which decays to ^{57}Fe, with the final decay to the ground state emitting 14.4-keV recoilless γ radiation having a remarkably narrow linewidth ($\Delta E/E = 10^{-12}$). The decay scheme is illustrated in Fig. 8(a). This extremely sharply defined frequency of the radiation allows observation of the local fields acting on the ME nucleus. The Mössbauer spectrum is observed by moving the source of γ radiation at low constant velocity or acceleration and counting the 14.4-keV γ rays transmitted through the absorber, or vice versa. A very simplified experimental arrangement is shown in Fig. 8(b). When the Doppler-shifted radiation of the moving source exactly matches that of the absorber, there is a strong absorption.

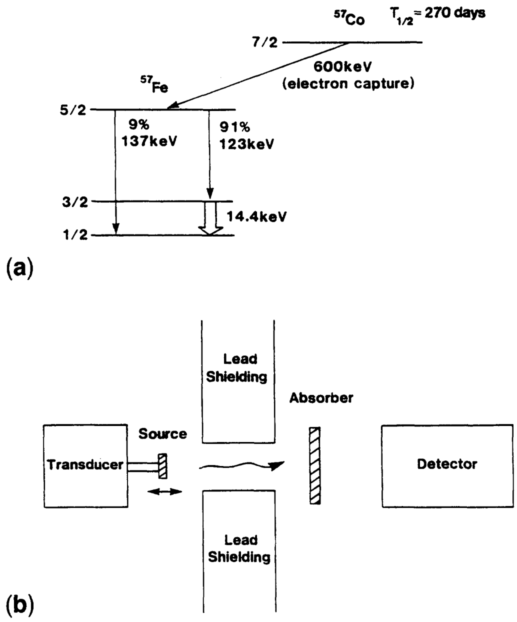

FIG. 8. (a) Decay scheme of ^{57}Co, not to scale. The frequency of the 14.4-keV γ ray is extremely sharp and used for ^{57}Fe Mössbauer spectroscopy. (b) A simplified arrangement for Mössbauer spectroscopy. A Doppler shift in frequency of the source is produced by moving the source at a low constant velocity or acceleration.

The interactions at the nucleus include the isomer shift δ, which reflects the electron charge density at the absorbing nucleus as compared to the source and varies with the host material, the quadrupole splitting Δ, which depends on the electric field gradients and reflects the coordination symmetry of the host and the magnetic interactions from local fields H_{int} such as the hyperfine and applied fields. The last produce a six-line spectrum for Fe. This splitting is shown schematically in Fig. 9. All the changes in splittings are shown on a very expanded scale compared to the 14.4-keV E_γ. The sum of the interactions δ, Δ, and H_{int} is in the range of 10^{-6} eV, whereas E_γ is greater than 10^5 eV. The center section of Fig. 9(a) illustrates the increase in energy δ due to the isomer shift; Fig. 9(b) shows the additional splitting Δ introduced by the quadrupole splitting; and Fig. 9(c) shows the additional splitting produced by H_{int}, the magnetic hyperfine field. The arrows indicate the allowed transitions. On the right in Fig. 9, the corresponding absorption spectra are shown as functions of source velocity. The hyperfine field in Fe is −33.0 T at 300 K; this corresponds to absorption extending to a ±5.3-mm/s Doppler shift of the 14.4-keV radiation. (ME data are often plotted in terms of velocity in mm/s, and the isomer shift data are reported most often with respect to Fe.)

The details of the ME spectrum identify the particular magnetic host and can be observed in quite small quantities of material

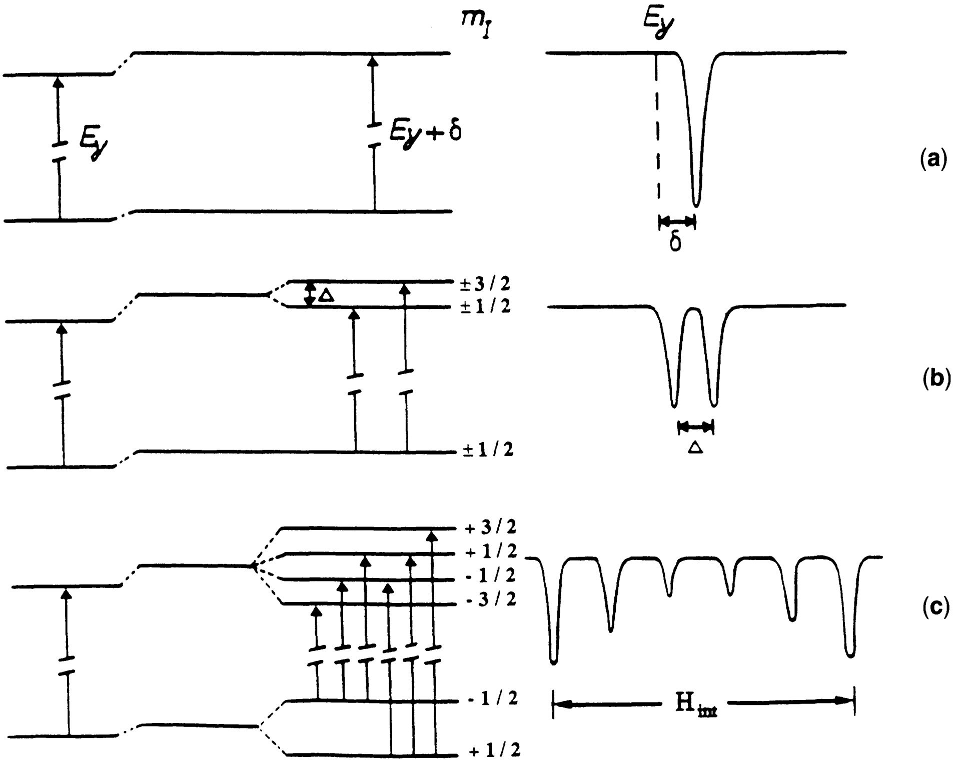

FIG. 9. Pictorial representation of the development of the Mössbauer absorption spectrum of ^{57}Fe. The 14.4-keV E_γ energy levels without interactions are shown at the left. The central section shows the change in energy levels as each interaction is turned on, and the corresponding absorption spectra are shown at the right: (a) adding the isomer shift δ; (b) adding the quadrupole splitting Δ; and (c) adding the internal hyperfine field H_{int}, which corresponds to ±5.3 mm/s of the source. All the interactions are shown on an expanded scale: $\delta + \Delta + H_{int}$ is less than 10^{-11} E_γ. (Courtesy of G. Papaefpthymiou, Francis Bitter National Magnet Laboratory, MIT.)

(10^{14} atoms). By applying a magnetic field, the spin structure can often be identified from the way the spectrum changes. Furthermore, the lifetime of the 14.4-keV state is about 10^{-7} s, so that dynamic effects of the magnetization can be observed, such as the fluctuations of the magnetization of superparamagnetic materials. If the local field fluctuates much faster than 10^{-7} s, the spectrum collapses to a two-line spectrum. The intensity of the absorption decreases at high temperature. Data are collected over time so that each spectrum may take a few hours. For low concentrations or limited γ-ray aperture, longer counting times may be required, during which time the field and temperature must be controlled. The measurements require avoiding vibrations, which broaden the spectrum. Interpretation of the spectrum involves detailed computer analysis. There are many commercial instruments available. Macroscopic magnetization measurements combined with ME data taken as a function of applied field form a powerful combination for studying the magnetic properties of materials. A ME data index lists all the reported ME data so that particular materials can be readily identified on the basis of their spectra. The ME effect has been employed for many fundamental physics problems and applied to many areas besides magnetism. Conferences are held frequently for practitioners.

The most powerful tool for elucidating the microscopic magnetic structure of materials is neutron scattering, because neutrons have a spin and no charge. Thus they can penetrate the medium, and the spin acts as a magnetic probe. It is by means of neutron scattering that the various complex ordered magnetic spin configurations in magnetic materials are determined. Neutrons also are used to determine the spin density distribution in the solid, the magnetic excitations, and the phonon excitations. Facilities exist at various locations throughout the world for such research but are not routinely used for characterizing practical magnetic materials.

5. OBSERVATION OF MAGNETIC DOMAINS

Magnetic recording media are used for most electronic information storage and represent a very large and growing market worldwide. As the storage density increases, extremely small regions must be sampled in order to determine the properties of these materials. The domain structures of thin films and recording media are often observed by optical means. The classic method of observing domain structures was developed by Bitter. To observe Bitter patterns of the domains, fine powdered magnetic particles are distributed over the surface of a ferromagnetic material. These particles are attracted to the local gradient fields produced by domain walls at the surface, and the localized powder distribution is observed under a microscope. Such decorations continue to be used but are not convenient for routine examinations of a large number of samples.

Optical phenomena have extremely rapid response and are used to examine magnetic materials nondestructively; the two techniques used are the Faraday and Kerr effects. The Faraday effect is the rotation of polarized light passing through a medium and is given by $R = kMl$, where k is the Verdet constant, l is the path length in the medium, and M is the magnetization; the rotation of the plane of linearly polarized light is reversed if $\mathbf{M}$ is reversed. Magnetic domain structures can be observed in optically transparent magnetic materials using the Faraday effect. The rotation of the plane of polarized light reflected from a magnetized surface is called the Kerr effect; rotation takes place in the penetration depth of the metal. Nearly perpendicular reflection is used if the magnetization is normal to the surface. Usually the magnetization is aligned in the plane of the surface because of demagnetization effects. In order to observe domains in this case, the polarized light is reflected at a small angle with respect to the surface. By illuminating the surface of the medium with polarized light and observing the reflected beam through an analyzer, the domain structure can be examined. If a narrow beam of light is used and a magnetic field is applied, the hysteresis of selected regions of the sample can be measured and compared. This approach is particularly useful for surveying magnetic recording media without disturbing the surface. It should be noted that the surface layer may be quite different from the bulk, and unless carefully handled, the surface may reflect strains introduced inadvertently. Although the Kerr rotation may be only a few hundredths of a degree for full saturation, this change in rotation is readily detected with high resolution. The response of the Kerr effect is proportional to M, not dM/dt, so that the response is linear from dc to high frequencies. If needed, the polarized incident beam can be amplitude or angularly modulated and the ac signal fed to a lockin to increase sensitivity. Kerr effect instruments are now available commercially.

Other imaging techniques are Lorentz microscopy, where the electron beam of a scanning electron microscope is deflected by the local field variations at the surface, and transmission electron microscopy, where the electron beam is deflected by magnetic interactions passing through very thin films. Such techniques allow observations of domain structures with high resolution. Recently, derivatives of the scanning tunneling microscope, such as the scanning atomic force and spin-polarized tunneling microscopes, are being developed that should permit examination of magnetic features on an atomic scale.

6. MEASUREMENT OF MAGNETIC FIELDS

The most common devices for field measurement are briefly mentioned below. The classical measurement of the magnetic field uses a pickup coil, calibrated from known dimensions or from measurement of the flux change of a known field. There are some applications where such an arrangement is the most convenient and logical choice. The methods discussed below are often available as commercial instruments. A pickup coil measures the average field over its volume. If a pickup coil is rotated or oscillated perpendicular to the steady applied field H_0, the ac output is proportional to $NAfH_0$, where NA is the area-turns and f the frequency. The rectified or phase-detected voltage furnishes a dc signal proportional to H_0. These devices can be used over a wide range of field.

Nuclear magnetic resonance (NMR) is similar to EPR, but here one observes the resonant precession of a nuclear spin moment in an applied field. The resonant frequency is much lower than for an electron in the same applied field, because the mass of the nucleus replaces the mass of the electron in the gyromagnetic ratio γ; the resonant frequency of a proton is 42.6 MHz at 1 T. The nuclear resonance can be extremely sharp; the full linewidth at half-maximum of protons (hydrogen in water) is about 0.1 Hz. The usual commercial NMR probes for routine field measurement use water or deuterium, but the linewidth is broadened by addition of some paramagnetic impurities in the liquid for easier observation. NMR is used for high-accuracy field calibration, measurement, or field stabilization but requires a very uniform and stable field over the probe volume, which can be 10^{-6} m^3 [1 cm^3] or more.

A very useful, general-purpose, field probe employs the Hall effect. If a current is passed through a rectangular slab conductor placed normal to an applied field, a Hall voltage E is observed perpendicular to both the field and current and is given by $E = RIH_0/t$, where I is the total current, t is the thickness of the slab, and R is the Hall constant, which varies as $1/n$ where n is the carrier density. Semiconductors give substantial Hall voltages, can be chosen for good linearity over a range of 30 T, can be fabricated as thin devices with cross sections of a square millimeter or less, are available commercially, and can be used at low temperatures (Rubin and Sample, 1980). Thermoelectric effects at the contacts are eliminated by using an ac excitation. Hall devices are used in many cases for measuring field and can be used down to 1 μT [10 mG]. As with pickup coils, the Hall probe is sensitive to angular misalignment with respect to the field. The Hall sensor can be placed against the surface of a material to measure the fields in close proximity. Hall devices are commonly used to measure the field of electromagnets without sacrificing space and as sensors for devices which involve magnetic fields, such as current measuring devices. In the latter case, the Hall element is inserted in a small air gap of a toroidal magnetic ring which surrounds the current conductor. The field around the current conductor generates an H in the toroid, and the field in the gap is thus a measure of that current. The Hall element also has been used as a multiplier element, because it involves the vector product of the current and field; one signal furnishes the field and the other furnishes the current, and the product is proportional to the Hall voltage.

Magnetoresistance can also be used as a sensitive probe. Unlike inductive sensors, the magnetoresistance depends directly on B, not dB/dt. The magnetoresistance is an even function of field and does not distinguish the sign of the field, and magnetoresistance is often a nonlinear function of magnetic field. However, because of their great simplicity, ease of manufacture in extremely small dimensions by conventional patterning techniques, wide frequency response, and low power requirements, magnetoresistance sensors are promising for computer applications.

A common device for measuring very low fields is a fluxgate magnetometer. It has a long history and many descriptive names, including saturable-core, saturated-core-reactor, saturable inductor, and peaking strip. Reviews of this instrument are given by Primdahl (1979) and Gordon and Brown (1972). It is a reliable, simple, rugged, sensitive, and directional device which has been improved over 60 years and appears in many forms. It is based on the strong nonlinearity of a highly permeable magnetic material to which a small ac field is applied. In its simplest form, an excitation coil surrounds a high-permeability strip, which is long compared to its cross section in order to reduce the demagnetization field. The strip should have a small coercive field and large apparent permeability. When a small, constant-amplitude, ac excitation current at frequency f is applied to the coil, the ac field sweeps through the hysteresis loop to saturation. The induced voltage in a pickup coil wound around the central region of the strip furnishes the output signal to be processed. The output voltage is very nonlinear, reaching a maximum output voltage each time the excitation field reaches H_c (where dB/dt is maximum) and reaching very small values when the ac field is not near H_c. When no dc field is present, two equal-amplitude and polarity pulses are observed at $+H_c$ for one half-cycle of the ac excitation and two equal-amplitude pulses (but of opposite polarity to those at $+H_c$) are observed at $-H_c$ for the other half-cycle. Analysis of the wave forms is complex, but one of the central features is that the output has a second-harmonic ($2f$) component when a dc field is applied. This signal is phase detected with a lockin and gives a measure of the dc field. The change in dc output of the lockin also can be used to feed back a dc current to produce a field at the sensor to cancel out the external dc field and restore the balance. The fed-back current is then a measure of the field. Fluxgate magnetometers routinely achieve 10^{-10} T/Hz$^{1/2}$, and special space applications units have achieved a sensitivity approaching 3×10^{-12} T/Hz$^{1/2}$ (Behannon *et al.*, 1977). A common unit in the literature for very low field studies is the gamma = 10^{-9} T [10^{-5} G].

ACKNOWLEDGMENT

The Francis Bitter National Magnet Laboratory is supported by the National Science Foundation.

GLOSSARY

Anisotropy Energy: The energy necessary to rotate the magnetization from a preferred or easy direction in a crystal to a hard direction.

Anisotropy, Magnetic: A measure of the preference of the direction of the magnetization in a material due to the crystalline symmetry or the shape of the sample.

Antiferromagnet: A magnetically ordered material, where the elementary spins in a domain are aligned antiparallel to each other and the magnetic moments of the spins in the two directions are equal.

Apparent Permeability, μ_a: The permeability measured in an applied field H_0, i.e., $B/\mu_0 H_0$. When there is no demagnetization field, $\mu_a = \mu$.

Apparent Susceptibility, χ_a: The susceptibility measured in an applied field H_0, i.e., M/H_0. When there is no demagnetization field, $\chi_a = \chi$.

Coercive Field, H_c: The field measured in the second quadrant of the saturation hysteresis loop at the point where $B = 0$, stated as a positive value of field.

Coercive Field, Intrinsic, H_{ci}: The field measured in the second quadrant of the saturation hysteresis loop at the point where $M = 0$, stated as a positive value of field.

Curie Temperature, T_c: The ordering temperature of a ferromagnetic or ferrimagnetic material.

Curie–Weiss Temperature, θ: The ordering temperature in the Curie–Weiss law, obtained by extrapolation from the high-temperature paramagnetic region. Sometimes called the paramagnetic Curie temperature; it is different from T_c and from T_N.

Demagnetization Factor, Fluxmetric, N_f: The demagnetization factor at the midplane of long bar, where the flux is detected with a short pickup coil.

Demagnetization Factor, Magnetometric, N_m: The average demagnetization factor over the sample.

Demagnetization Factor, N: A quantity which depends on the geometry of a sample and is clearly defined for an ellipsoidal sample, for which H_d is uniform.

Demagnetization Field, $H_d = -NM$: An internal magnetic field produced by poles induced on the surface of a material.

Electron Paramagnetic Resonance (EPR): Same as electron spin resonance.

Electron Spin Resonance (ESR): The resonant change in the angle at which an electronic spin precesses in a magnetic field.

Energy Product, Maximum $(BH)_{\max}$: The maximum product of B and H obtained in the second quadrant of a saturation hysteresis loop.

Faraday Balance: A device for measuring the force on a sample placed in a gradient of an applied magnetic field.

Faraday Effect: The rotation of plane polarized light on passing through a magnetized material.

Ferrimagnet: A magnetically ordered material, where the elementary spins in a domain are aligned antiparallel to each other but the magnetic moments of the spins in the two directions are unequal.

Ferromagnet: A magnetically ordered material, where all the elementary spins in a domain are aligned parallel to each other.

Hall Effect: A voltage produced between the two edges of a conductor carrying a current in a plane perpendicular to an applied field.

Hyperfine Field: Field acting on the nucleus produced by the electron charge distribution.

Hysteresigraph: A device which is used to measure and display the magnetic hysteresis of a material.

Hysteresis Loop: The plot of B vs H or M vs H as H is increased from zero to a maximum value in a positive direction, decreased to 0, and repeated in the negative direction to complete the cycle.

Hysteresis Loop, Inner: A hysteresis loop where H does not saturate the material.

Hysteresis Loop, Minor: The hysteresis observed when H is increased a small amount and returns to its original value.

Hysteresis Loop, Saturation: A hysteresis loop where H is increased until magnetic saturation is achieved. This is the maximum hysteresis.

Hysteresis Loss, Rotational, W_r: The loss developed during rotation of a sample in a magnetic field through 360 deg.

Hysteresis, Magnetic: A measure of the nonreversible magnetic behavior of a material.

Hysteresis, Rotational: The irreversible hysteresis when a sample is rotated in a magnetic field.

Isomer Shift, δ: Shift in nuclear magnetic resonance frequency produced by the difference of the electron charge density acting on the nucleus, compared to the reference nucleus.

Kerr Effect: The rotation of plane polarized light on being reflected from the surface of a magnetized material.

Lockin Amplifier: A combination of a phase-sensitive detector with ac preamplification and dc postamplification and filtering.

Lorentz Microscopy: Microscopy based on the deflection of an electron beam of a scanning electron microscope by fields from magnetic domains.

Magnetometer, Alternating-Force (AFM): A force-detection magnetometer where an ac gradient produces a force on a sample.

Magnetometer, Alternating-Gradient: An alternating-force magnetometer.

Magnetometer, Fluxgate: A sensitive low-field magnetometer which uses a magnetic core made of a high-permeability magnetic material.

Magnetometer, Mechanical Torque: Measures the mechanical torque on an anisotropic sample placed in a uniform magnetic field.

Magnetometer, Rotating-Field (RFM): A flux-detection magnetometer where the field is rotated and the sample remains stationary.

Magnetometer, Rotating-Sample (RSM): A flux-detection magnetometer where the sample is rotated.

Magnetometer, SQUID: A flux-detection magnetometer where the field is detected with a SQUID.

Magnetometer, Vibrating-Reed: A variety of alternating-force magnetometer.

Magnetometer, Vibrating-Sample (VSM): A flux-detection magnetometer where the sample is vibrated.

Magnetometer: An instrument that measures magnetic properties, usually magnetic fields.

Magnetoresistance: The change in resistance with a magnetic field.

Magnetostriction: The change in dimensions of a sample due to magnetic interactions.

Magnetostriction, Forced: The small magnetostriction observed in a magnetic field above magnetic saturation of a material.

Mössbauer Effect: The recoilless emission of γ radiation from a nucleus bound in a crystal. It has an extremely sharply defined frequency.

Néel Temperature, T_N: The ordering temperature of an antiferromagnetic material.

Nuclear Magnetic Resonance (NMR): The resonant absorption of energy associated with change of the angle of precession of the nuclear moment in a magnetic field.

Permanent Magnet: A magnetic material which remains magnetized in the ambient magnetic fields. Usually applied to materials with large remanence and large coercive fields.

Permeability, Initial: The permeability at low field in a magnetic material initially in the demagnetized state.

Permeability, μ: $B/\mu_0 H$.

Permeability, Relative, μ_r: μ/μ_0.

Permeameter: An instrument that measures the permeability of a magnetic material.

Phase-Sensitive Detector: A device that rectifies a periodic voltage with a fixed phase with respect to that voltage.

Piezoelectric Bimorph: A piezoelectric device that generates a voltage when bent slightly.

Pulsed-Field Magnet, Destructive: A magnet that is destroyed during the pulse generation.

Pulsed-Field Magnet, Nondestructive: A magnet that can be pulsed without failing during the pulse.

Pulsed-Field Magnet, Quasistatic: A magnet that is pulsed for a time of 0.1 s or longer, often with a controlled wave form to produce a flat top. It requires a large energy source. It is also called semicontinuous pulsed field.

Quadrupole Interaction, Δ: The interaction between electric field gradients and the electric quadrupole moment of the nucleus.

Remanence or Remanent Field, B_r: The B in a hysteresis loop when H reaches zero in the second quadrant.

SQUID: Acronym for superconducting quantum interference device. A very sensitive detector of magnetic field.

Susceptibility, χ: M/H.

Torque VSM: A vibrating-sample magnetometer with sensors arranged to detect the components of **M** both parallel and perpendicular to the applied magnetic field.

Vector VSM: Same as torque VSM.

Works Cited

Arrott, A. Noakes, J. E. (1971), *J. Appl. Phys.* **42**, 1288–1289.

Arrott, A. (1971), *J. Appl. Phys.* **42**, 1282–1283 .

Artman, J. O. (1956), *Proc. IRE* **44**, 1284–1293.

Behannon, K. W., Acuña, M. H., Burlaga, L. F., Lepping, R. P., Ness, N. F., Neubauer, F. M. (1977), *Space Sci. Rev.* **21**, 235–266.

Bernards, J. P. C., van Engelen, G. J. P., Schrauwen, C. P. G., Cramer, H. A. J., Luitjens, S. B. (1990), *IEEE Trans. Magn.* **26**, 216–218.

Bernards, J. P. C., Cramer, H. A. J. (1993), *J. Magn. Magn. Mater.* **123**, 141–146.

Bowden, G. J. (1972), *J. Phys. E* **5**, 1115–1119.

Bozorth, R. M. (1951), *Ferromagnetism*, New York: D. Van Nostrand Co. Inc.

Chen, D.-X., Brug, J. A., Goldfarb, R. B. (1991), *IEEE Trans. Magn.* **27**, 3601–3619.

Clarke, J. (1977), in: B. B. Schwartz, S. Foner, (Eds.), *Superconducting Applications: SQUIDs and Machines*, New York: Plenum, pp. 67–124.

Cullity, B. D. (1972), *Introduction to Magnetic Materials*, Reading, MA: Addison-Wesley Publishing Co., Inc., p. 499.

deBoer, F. J., deChâtel, P. F., Franse, J. J. M. (Eds.) (1992), Proceedings of the 3rd International Symposium on Research on High Magnetic Fields, *Physica* **B177**, 115–398.

Dwight, K. (1967), *J. Appl. Phys.* **38**, 1505–1509.

Fawcett, E. (1970), *Phys. Rev. B* **2**, 1604–1613.

Flanders, P. J. (1970), *Rev. Sci. Instrum.* **41**, 697–710.

Flanders, P. J. (1988), *J. Appl. Phys.* **63**, 3940–3945.

Flanders, P. J. (1990), *Rev. Sci. Instrum.* **61**, 839–847.

Flanders, P. J., Graham, C. D. (1993), *Rep. Prog. Phys.* **56**, 431–492.

Foner, S. (1959), *Rev. Sci. Instrum.* **30**, 548–557.

Foner, S. (1967), *J. Appl. Phys.* **38**, 1510–1519.

Foner, S. (1975), *Rev. Sci. Instrum.* **46**, 1425–1426.

Foner, S. (1981), *IEEE Trans. Magn.* **MAG-17**, 3358–3363.

Foner, S. (1990), *J. Appl. Phys.* **67**, 5687–5692.

Foner, S., McNiff, E. J., Jr. (1968), *Rev. Sci. Instrum.* **39**, 171–179.

Foner, S., Shapira, Y., Heiman, D., Becla, P., Kershaw, R., Dwight, K., Wold, A. (1989), *Phys. Rev. B* **39**, 11793–11799.

Gillies, G. T., Ritter, R. C. (1993), *Rev. Sci. Instrum.* **64**, 283–309.

Goldfarb, R. B., Lelenthal, M., Thompson , C. A. (1992), in: R. A. Hein, T. L. Francavilla, D. H. Liebenberg (Eds.), *Magnetic Susceptibility of Superconductors and Other Spin Systems*, New York: Plenum Press, pp. 49–80.

Gordon, D. I., Brown, R. E. (1972), *IEEE Trans. Magn.* **MAG-8**, 76–82.

Guyot, M., Foner, S., Hasanain, S. K., Guertin, R. P., Westerholt, K. (1980), *Phys. Lett.* **77A**, 339–341.

Hagedorn, F. B. (1967), *Rev. Sci. Instrum.* **38**, 591–597.

Heck, C. (1974), *Magnetic Materials and Their Applications*, New York: Crane, Russak and Co., pp. 523–579.

Herlach, F., Franse, J. J. M. (Eds.) (1989), Proceedings of the 2nd International Symposium on High Field Magnetism, *Physica* **B155**, 3–125 and 199–319.

Hill, G. J. (1968), *J. Phys. E* **2**, 52–54.

Humphrey, F. B. (1967), *J. Appl. Phys.* **38**, 1520–1527.

Humphrey, F. B., Johnson, A. R. (1963), *Rev. Sci. Instrum.* **34**, 348–358.

Ingram, D. J. E. (1979), in: G. F. M. Kalvius, R.

S. Tebble (Eds.), *Experimental Magnetism*, Vol. 1, New York: John Wiley & Sons, pp. 268–335.

Krause, J. E. (1992), *IEEE Trans. Magn.* **28**, 3066–3071.

Lax, B., Button, K. J. (1962), *Microwave Ferrites and Ferrimagnetics*, New York: McGraw-Hill Book Co. Inc.

Lee, E. W. (1979), in: G. F. M. Kalvius, R. S. Tebble (Eds.), *Experimental Magnetism*, Vol. 1, New York: John Wiley & Sons, pp. 226–266.

Lewis, R. T. (1971), New York: *Rev. Sci. Instrum.* **42**, 31–34.

Mallinson, J. (1966), *J. Appl. Phys.* **37**, 2514–2515.

Pacyna, A. W., Ruebenbauer, K. (1984), *J. Phys. E* **17**, 141–154.

Pearson, R. F. (1979), in: G. F. M. Kalvius, R. S. Tebble, (Eds.), *Experimental Magnetism*, Vol. 1, New York: John Wiley & Sons, pp. 138–223.

Penoyer, R. F. (1959), *Rev. Sci. Instrum.* **30**, 711–714.

Poole, C. P. (1967), *Electron Spin Resonance: A Comprehensive Treatment on Experimental Techniques*, London: Interscience.

Praddaude, H. C., Foner, S. (1976), *Solid State Commun.* **20**, 1117–1119.

Primdahl, F. (1979) *J. Phys. E* **12**, 241–253.

Pugh, E. M., Argyle, B. E. (1962), *J. Appl. Phys.* **33**, 1178–1186.

Richter, H. J., Hempel, K. A., Pfeiffer, J. (1988), *Rev. Sci. Instrum.* **59**, 1388–1393.

Richter, H. J., Hibst, H. (1991), *J. Appl. Phys.* **70**, 5512–5516.

Roeland, L. W., Gersdorf, R., Mattens, W. C. M. (1989), *Physica* **B155**, 58–60.

Roos, W., Hempel, K. A., Voigt, C., Dederichs, H., Schippen, R. (1990), *Rev. Sci. Instrum.* **51**, 612–613.

Rubin, L. G., Sample, H. H. (1980), in: C. L. Chien, C. R. Westgate (Eds.) *The Hall Effect and Its Applications*, New York: Plenum Publishing Co., pp. 463–479.

Sager, R. E., Hibbs, A. D., Kumar, S. (1993), *IEEE Trans. Magn.* **28**, 3072–3077.

Stewart, A. M. (1969), *J. Phys. E* **2**, 851–854.

Stewart, A. M. (1975), *J. Phys. E* **8**, 55–59.

Webb, W. W. (1972), *IEEE Trans. Mag.* **MAG-8**, 51–65.

Wolf, W. P. (1957), *J. Appl. Phys.* **28**, 780–781.

Zieba, A., Foner, S. (1982), *Rev. Sci. Instrum.* **53**, 1344–1354.

Zieba, A., Foner, S. (1983), *Rev. Sci. Instrum.* **54**, 137–145.

Zieba, A. (1993), *Rev. Sci. Instrum.* **64**, 3357–3375.

Zilstra, H. (1967), in: E. P. Wohlfarth (Ed.), *Experimental Methods of Magnetism*, Vols. 1 and 2, New York: John Wiley & Sons, Inc.

Zilstra, H. (1970), *Rev. Sci. Instrum.* **41**, 1241–1243.

Further Reading

Berkowitz, A. E., Kneller, E., (Eds.) (1969), *Magnetism and Metallurgy*, Vol. 1, New York: Academic Press. Reviews magnetic measurements and measurement techniques with an interest in materials.

Bozorth, R. M. (1951), *Ferromagnetism*, New York: D. Van Nostrand Co. Inc. A classic treatise on magnetism with an extensive discussion of phenomena and data on magnetic materials available at that time. This book has been recently reprinted and is available from IEEE Press, Piscataway, NJ 08855–1331.

Cranshaw, T. E., Dale, B. W., Longworth, G. O., Johnson, C. E. (1985), *Mössbauer Spectroscopy and its Applications*, London: Cambridge University Press. A small monogram; a good introduction to a variety of Mössbauer applications.

Cullity, B. D. (1972), *Introduction to Magnetic Materials*, Reading, MA: Addison-Wesley Publishing Co. Inc. An extensive text written with an appreciation of applications.

Evetts, J. (Ed.) (1992), *Concise Encyclopedia of Magnetism and Superconductivity*, Oxford: Pergamon Press. Short up-to-date articles covering materials aspects and properties of superconductors and magnetic materials.

Flanders, P. J., Graham, C. D. (1993), *Rep. Prog. Phys.* **56**, 431–492. A very recent review of dc and low-frequency measurement techniques.

Greenwood, H. H., Gibb, T. C. (1971), *Mössbauer Spectroscopy*, London: Chapman and Hall Ltd. Excellent detailed reference. See also *Mössbauer Effect Reference and Data Journal*, available from The University of North Carolina in Asheville, Asheville, NC 28804–3288.

Heck, C. (1974), *Magnetic Materials and their Applications*, New York: Crane, Russak and Co. Inc. A detailed reference book which addresses classes of magnetic materials based on their applications.

Ingram, D. J. E. (1979), in: G. F. M. Kalvius, R. S. Tebble, (Eds.), *Experimental Magnetism*, Vol. 1, New York: John Wiley & Sons, pp. 268–335. An introduction to resonance phenomena.

Jiles, J. (1991), *Introduction to Magnetism and Magnetic Materials*, London: Chapman & Hall. A very readable introductory text.

Kalvius, G. F. M., Tebble, R. S. (Eds.) (1979), *Experimental Magnetism*, Vol. 1, New York: John Wiley & Sons. Contains review articles on experimental magnetism.

Lee, E. W. (1979), in: G. F. M. Kalvius, R. S. Tebble, (Eds.), *Experimental Magnetism*, Vol. 1, New York: John Wiley & Sons, pp. 226–266. A good review of magnetostriction and problems in interpreting strain measurements.

Pearson, R. F. (1979), in: G. F. M. Kalvius, R. S. Tebble, (Eds.), *Experimental Magnetism*, Vol. 1, New York: John Wiley & Sons, pp. 138–223. A comprehensive review of anisotropy measurements.

Poole, C. J. (1967), *Electron Spin Resonance: A Comprehensive Treatment on Experimental Techniques*, London: Interscience.

Schwartz, B. B., Foner, S. (Eds.) (1977) *Superconducting Applications: SQUIDs and Machines*, New York: Plenum Press. A series of tutorial articles on SQUIDs and their applications.

Wohlfarth, E. P. (Ed.), *Ferromagnetic Materials* Vols. 1 and 2 (1980), 3 (1982); Wohlfarth, E. P., Buschow, K. H. J. (Eds.) (1988), Vol. 4, Amsterdam: North-Holland Publishing Co. A comprehensive coverage of ferromagnetic materials.

Zilstra, H. (1967), in: E. P. Wohlfarth (Ed.), *Experimental Methods of Magnetism*, Vols. 1 and 2, New York: John Wiley & Sons, Inc. An excellent reference for in-depth descriptions of various magnetic measurement techniques with lucid explanations.

Meetings of the Magnetism and Magnetic Materials Conference take place every year in the United States, and the proceedings are published regularly in *J. Appl. Phys.* in approximately April or May each year. These are more directed toward basic magnetism. Meetings of the Intermag Conferences also take place each year but do so in countries other than the United States. These are more directed toward applications and are published each year in *IEEE Trans. Magn.* about 6 months later. Both conferences discuss new developments in magnetic materials and their evaluation.

MEASUREMENT

See ELECTRICAL INSTRUMENTATION; ELECTROMAGNETIC QUANTITIES, BASIC, MEASUREMENT OF; LABORATORY INSTRUMENTATION; MEASUREMENT AND INSTRUMENTATION OF FLOW; MAINTENANCE AND CALIBRATION OF LABORATORY INSTRUMENTATION; MASS AND DENSITY, MEASUREMENT OF; MEASUREMENT, ACOUSTICAL; MEASUREMENT OF MAGNETIC PROPERTIES AND QUANTITIES; SPACE, MEASUREMENT OF; TEMPERATURE CONTROL, TECHNIQUES AND INSTRUMENTATION; TIME AND FREQUENCY, MEASUREMENT AND CONTROL OF; TRANSDUCERS AND OTHER MECHANICAL INSTRUMENTATION; VACUUM TECHNIQUES AND INSTRUMENTATION

MEASUREMENT STANDARDS

See CALIBRATION AND MAINTENANCE OF STANDARDS

MEASUREMENTS, ACOUSTICAL

ALLAN J. ZUCKERWAR, *NASA Langley Research Center, Hampton, Virginia, U.S.A.*

Introduction 491
1. **Measurement of Sound Speed and Attenuation** 491
1.1 Introductory Concepts 491
1.2 Measurement of Sound Speed 492
1.2.1 Cylindrical Resonator 492
1.2.2 Spherical Resonator 492
1.3 Measurement of Sound Attenuation 493
2. **Measurement of Sound Pressure** 494
2.1 Introduction to Measurement Microphones 494
2.1.1 Construction and Principle of Operation 495
2.1.2 Microphone Specifications ... 496
2.1.3 Directional Properties 497
2.1.4 Microphone Calibration 498
2.1.5 Microphone Performance in Harsh Environments 499
2.2 Infrasonic Measurements 500
2.3 Background Noise in Microphones 500
3. **Measurement of Sound Intensity and Power** 501
3.1 Sound Intensity 501
3.2 Acoustic Enclosures 501
3.3 Sound Power 503
4. **Measurement of Acoustic Particle Velocity** 503
5. **Measurement of Acoustic Impedance** 504
5.1 Introductory Concepts 504
5.2 Impedance Tube 504
5.3 Direct Measurement of Sound Pressure and Volume Velocity 505
5.4 Free-Field Methods 506
6. **Filters in Acoustics** 506
6.1 Introductory Concepts 506
6.2 Spectral Selection 506
6.3 Spectral Analysis 507
6.4 Spectral Shaping 507
6.5 Filter-Effect Error 507
7. **Spectrum Analyzers** 508
7.1 Introduction 508
7.1.1 Digital Filter 509
7.1.2 Fast Fourier Transform 509
7.2 Special Features 510
8. **Sound-Level Meters** 510
9. **Magnetic Tape Recorders** ... 511
9.1 Characteristics 511
9.1.1 Direct Record/Reproduce ... 511
9.1.2 FM Record/Reproduce 512
9.2 Digital Tape Recording 512
9.3 PC Adaptables 512
10. **Computers in Acoustics** 512
Glossary 513
Works Cited 514
Further Reading 514

INTRODUCTION

The first part of this article is concerned with the measurement of the basic properties of sound in fluids (primarily gases), focusing on acoustical measurands, methods available for their measurement, advantages and disadvantages, specifications, and calibration techniques. The second part describes instrumentation for processing acoustical data, their role in acoustical measurement systems, and features available to the user.

1. MEASUREMENT OF SOUND SPEED AND ATTENUATION

1.1 Introductory Concepts

Sound is a physical phenomenon identified with the propagation of a mechanical disturbance through a medium. The *speed of sound* is defined as the distance traversed per unit time by a given point on the disturbance, provided the disturbance does not change its shape (as would be the case in an ideal medium). This definition applies as well

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

to standing waves, since the latter can be resolved into traveling waves propagating in opposite directions. A reduction in amplitude with propagation time or distance is called the *attenuation,* which may be the result of spreading, scattering, or absorption (direct conversion to heat). The same apparatus can be used to measure both sound speed and attenuation; often both quantities are measured together.

1.2 Measurement of Sound Speed

The measurement of sound speed can be accomplished by time-domain methods, based on the relationship $c = d/t$, where t is the propagation time over the distance d between two measurement stations; or by frequency-domain methods, based on $c = f\lambda$, the product of the frequency and wavelength of a harmonic wave. The latter methods are mostly used to measure sound speed in gases because a generally high attenuation limits practical propagation distances, and the associated dispersion and wave-form distortion inhibit good spatial or time resolution. Two apparatus commonly used for the measurement are the cylindrical resonator and the spherical resonator.

1.2.1 Cylindrical Resonator The *cylindrical resonator* is a long, hollow cylinder capped by two parallel end plates. It employs a transmitter and receiver, either separately or as a single unit and usually located in the end plates. Many types of electroacoustic transducers are used: piezoelectric, electrostatic, electrodynamic, etc. In the "fixed-path interferometer," the excitation frequency is tuned successively to several axial resonances, and c is determined from

$$c = 2Lf_n/n, \qquad n = 1, 2, \ldots, \tag{1}$$

where L is the length of the tube and f_n the nth resonance frequency. In the "variable-path interferometer," the excitation frequency f remains fixed, but the tube length is varied to establish the resonance condition

$$c = 2fL_n/n, \tag{2}$$

where L_n is the tube length at the nth axial resonance.

The advantages of the cylindrical geometry are modal purity and ease of fabrication, especially for measurements at high pressures. The disadvantages are the large thermoviscous losses at the walls, requiring a correction to the measured resonance frequencies, and the possible excitation of higher (nonaxial) modes. The following correction for wall losses relates the ideal f_n [Eq. (1)] to the measured $f_{n\text{-meas}}$ resonance frequencies:

$$f_n = f_{n\text{-meas}}(1 + \chi/2\sqrt{f_{n\text{-meas}}}), \tag{3}$$

$$\chi = \frac{1}{R}\left(\frac{\eta}{\pi\rho}\right)^{1/2} \times \left[1 + (\gamma - 1)\left(1 + \frac{2R}{L}\right)\left(\frac{\kappa}{\eta C_p}\right)^{1/2}\right], \tag{4}$$

where η is the viscosity coefficient of the fluid; ρ is the density; γ is the specific heat ratio; κ is the thermal conductivity; C_{p} is the specific heat at constant pressure; and R and L are the radius and length of the tube (Roesler, 1966). For air at 20 °C, 1 atm, and R and L in m, Eq. (4) becomes, in units of $\mathrm{Hz}^{1/2}$,

$$\chi = (2.182 \times 10^{-3}/R) \times [1 + 0.5186(1 + 2R/L)]. \tag{5}$$

If the exciter is not a rigid piston normal to the tube axis, then nonaxial modes may be excited above their cutoff frequencies, i.e., $f_c = gc/2R$, where $g = 0.5861$ and 1.2187 for the lowest modes having diametrical and circular nodes, respectively.

1.2.2 Spherical Resonator The excitation of a radial wave in a *spherical resonator* has an important advantage: Because the particle velocity is normal to the walls, the wall losses are typically 10:1 lower than those of a cylindrical resonator at a comparable frequency. The sound speed is related to the radial mode frequencies by

$$c = 2\pi Rf_n/\nu_n, \tag{6}$$

where R is the radius of the sphere and $\nu_n = 4.493, 7.725, 10.904, \ldots$ for $n = 1, 2, 3, \ldots$. Equation (3) again gives the wall loss correction, but now Eq. (4) becomes

$$\chi = \frac{\gamma - 1}{R}\sqrt{\frac{\kappa}{\pi\rho C_p}}, \tag{7}$$

in which there is no viscous contribution (Moldover *et al.*, 1986). For air at 20 °C, 1 atm, and R in m, Eq. (7) becomes, in units of $\mathrm{Hz}^{1/2}$,

$$\chi = 9.96 \times 10^{-4}/R. \tag{8}$$

The disadvantages of the spherical resonator are the difficulty in fabricating a spherical shell with precise internal dimensions and the practical limitation on shell radius, which limits the lowest attainable measurement frequency. Because geometric imperfections affect c only to second order, the spherical resonator is capable of yielding measurements of great precision.

1.3 Measurement of Sound Attenuation

There are many ways to designate sound attenuation, most associated with a particular experimental technique. The "quality factor" Q is the most fundamental measure because it is based on energy considerations. If a medium is subjected to periodic acoustic excitation, then

$$Q = \frac{2\pi \times (\text{maximum energy stored})}{\text{energy dissipated per cycle}}. \tag{9}$$

Although Q is ill-suited to direct measurement, it may be determined through its relationship to other measures of attenuation shown in Table 1.

When the wavelength is much smaller than the experimental enclosure, the "attenuation constant" α is the appropriate measure. A popular method of determining α is the pulse-echo method, illustrated in Fig. 1(a). A transmitter T launches a tone burst, a packet of harmonic waves of frequency f, in a cylindrical resonator. In the sequence of reflections from the end plates, a receiver R at a fixed station measures the amplitude of the packet, which attenuates by a factor $\exp(-\alpha x)$ over a propagation distance x. A suitable pulse width must be chosen by the observer: Too great a width will decrease spatial resolution, but too small a width will decrease frequency resolution.

Table 1. Measures of sound attenuation.

Measure	Units	Relationship to Q
Attenuation constant	Np/m	$\alpha = \pi/\lambda Q$
Reciprocal time constant	Np/s	$\beta = \pi f/Q$
Logarithmic decrement	Np/wavelength	$\Lambda = \pi/Q$
Resonant halfwidth	Hz	$\Delta f = f_0/Q$
Tangent of phase angle	···	$\tan\delta = 1/Q$

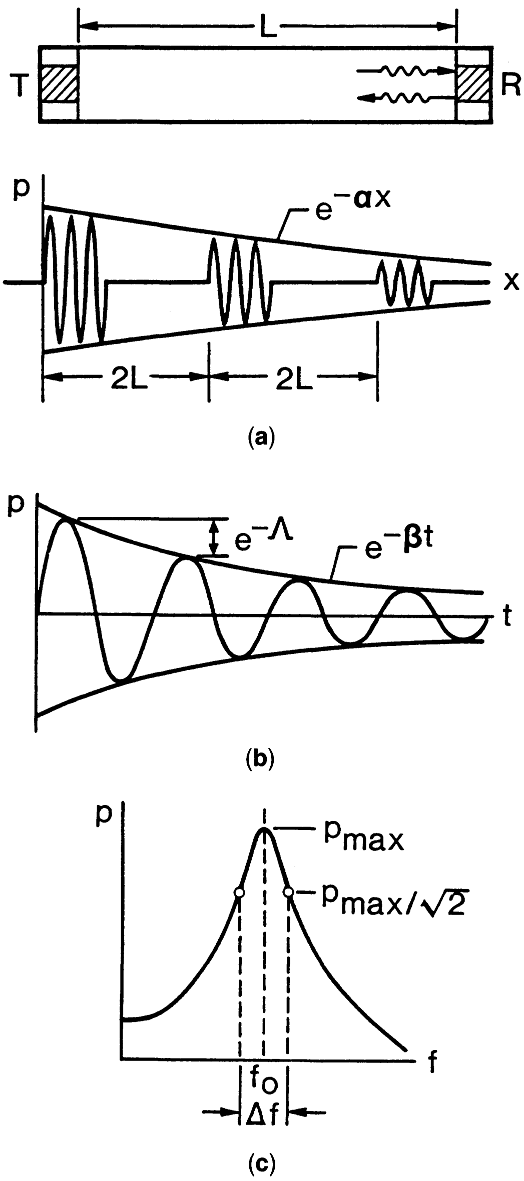

FIG. 1. Measurement of sound attenuation by (a) the pulse-echo, (b) the free-decay, and (c) the resonant halfwidth methods.

The attenuation can also be described in terms of a "time constant" (or its reciprocal β), as is common in reverberation measurements. In a typical experiment, white noise is used to excite a large number of reverberation-chamber modes, which can be observed individually on a fixed receiver through narrow-band filtering. When the excitation is removed, an amplitude decay of the form $\exp(-\beta t)$ is observed on each filter. A typical decay envelope is shown in Fig. 1(b).

The next three measures pertain to the resonant technique, whereby a natural mode of a resonator is excited, usually an axial mode of a cylinder or a radial mode of a sphere. When the excitation is removed, the sound pressure amplitude of the mode decays freely, as in the reverberation technique. The "logarithmic decrement" Λ is the natural logarithm of the amplitude ratio of two successive peaks, shown in Fig. 1(b):

$$\Lambda = \ln(p_n/p_{n+1}). \tag{10}$$

Usually Λ is averaged over a range of the free decay, typically 10–40 dB. This method is most suitable for media with high Q, say $Q > 10$; in media with extremely high Q ($>10^6$), this is the only feasible method for measuring sound attenuation.

In a steady-state experiment, the attenuation can be determined by the "halfwidth" of the resonance curve, shown in Fig. 1(c). The sound-pressure amplitude is measured as the frequency is incremented or swept through the resonance frequency f_0. The halfwidth Δf is defined as the frequency interval between the two points where $p = p_{\max}/\sqrt{2}$, that is, where the amplitude is 3 dB down from the peak $p_{\max}$. This method cannot be used when the halfwidth is too sharp, because of the natural drift of f_0, or when the halfwidth is too broad, as a result of poor peak definition. Furthermore, any losses inherent in the transmitter itself will contribute to the halfwidth. With an efficient transmitter, this method can be used effectively over a range of Q from about two to several hundred.

If the medium is very lossy ($Q < 2$), the most effective measure of attenuation is the "loss tangent" $\tan\delta$. Such low Q's are found in some polymeric liquids, seldom in gases. Because of the high loss, the sample is made the lossy element of a composite resonator, in which the independent measurements of force and displacement yield the phase angle δ.

If sound absorption measurements are conducted to study a physical process in a fluid, then the wall loss must be subtracted from the total absorption. In terms of Q, the wall loss is given by

$$Q^{-1} = \chi/\sqrt{f}, \tag{11}$$

where χ is given by Eq. (4) or (7), depending on the resonator.

2. MEASUREMENT OF SOUND PRESSURE

2.1 Introduction to Measurement Microphones

A sound wave propagating through a medium produces deviations in pressure and density about their mean or static values. The deviation in pressure is called the *acoustic* or *sound pressure*, which has *Système International* (SI) units of pascal (Pa = N/m^2). Because of the vast range of amplitude covered in acoustic measurements, the sound pressure is conveniently represented on a logarithmic scale as the "sound pressure level" (SPL). The SPL unit is the decibel (dB) defined as

$$\mathrm{SPL(dB)} = 20 \log(p/p_0), \tag{12}$$

in which p is the rms sound-pressure amplitude and p_0 the reference pressure of 20×10^{-6} Pa. The SPL's of some familiar sound sources and environments are listed in Table 2.

A device that detects a sound pressure in a gas and provides a proportional electrical output voltage is a *microphone*. Functionally, microphones fall into two categories: entertainment or broadcasting microphones and measurement microphones (see ACOUSTICAL INSTRUMENTATION). Entertainment microphones, comprising mainly electret, ribbon, and moving-coil microphones, conform to the requirements of speech and music and have preferred directionality. Measurement microphones, on the other hand, may have capabilities that extend well beyond these requirements, both in frequency response and in dynamic range, and are for the most part

Table 2. Representative sound pressure levels.

Source or environment	Level (dB re 20 μPa)
Concentrated sources: re 1 m	
Four-jet airliner	155
Auto horn, loud	115
Power lawnmower	100
Conversation	60
Whisper	20
Diffuse environments	
Concert hall, loud orchestra	105
Street corner, average traffic	95
Business office	60
Library	40
Threshold levels	
Of pain	130
Of hearing impairment, continuous exposure	90
Of hearing	0

omnidirectional. Nevertheless, the distinguishing mark of a measurement microphone is consistent performance in the face of prolonged service, exposure to various environmental conditions, and the passage of time. Recalibration often is repeatable to within tenths of a dB—a feature not required in entertainment microphones.

2.1.1 Construction and Principle of Operation The common types of measurement microphone are air condenser, electret condenser, ceramic, and piezoresistive microphones.

The construction of the *air condenser microphone* is shown in Fig. 2(a) (Bruel & Kjaer, 1982). An incident sound pressure *p* excites

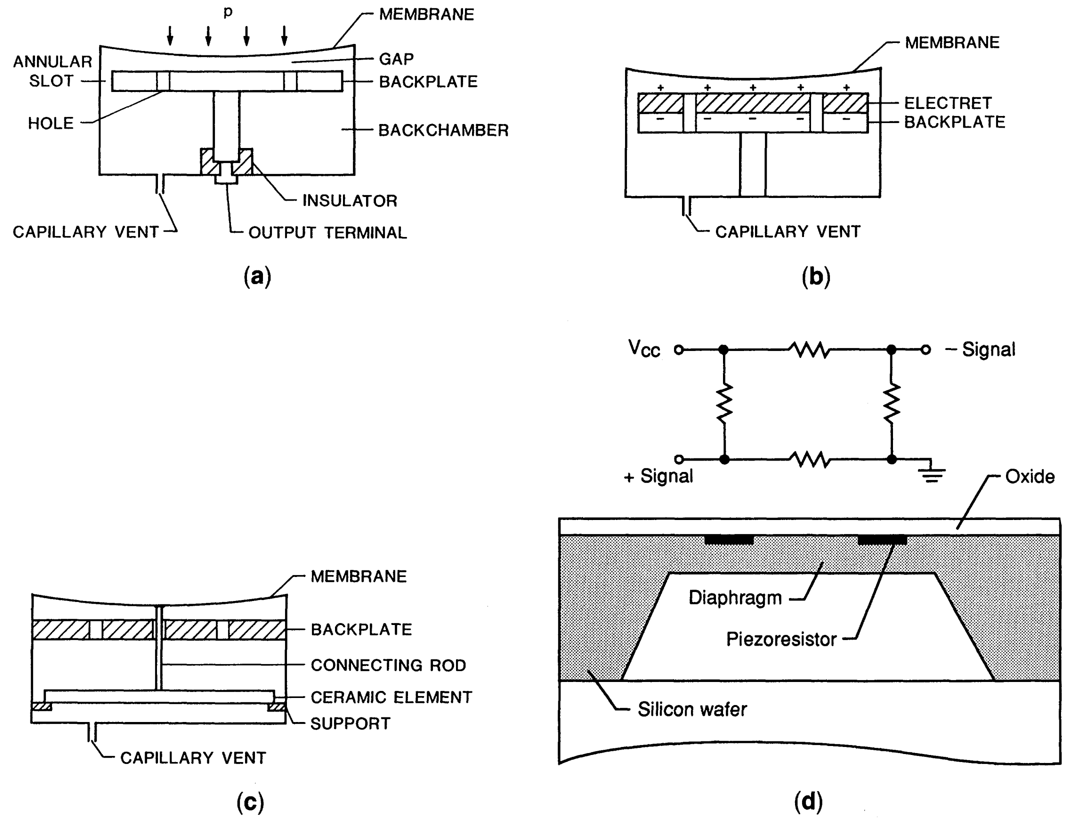
p
MEMBRANE
ANNULAR SLOT
GAP
BACKPLATE
HOLE
BACKCHAMBER
INSULATOR
CAPILLARY VENT
OUTPUT TERMINAL
(a)
MEMBRANE
ELECTRET
BACKPLATE
CAPILLARY VENT
(b)
MEMBRANE
BACKPLATE
CONNECTING ROD
CERAMIC ELEMENT
SUPPORT
CAPILLARY VENT
(c)
Vcc
− Signal
+ Signal
Oxide
Diaphragm
Piezoresistor
Silicon wafer
(d)

FIG. 2. Construction of measurement microphones: (a) air condenser, (b) electret condenser, (c) ceramic, and (d) piezoresistive.

motion of a stretched membrane, changing the capacitance between the membrane and backplate. When a fixed charge is applied on the plates of the membrane-backplate capacitor, then motion of the membrane changes the voltage between the plates proportionally. The charge is maintained by a high voltage, called a "polarization voltage," connected to the plates through a large resistor. A capillary vent provides for equalization of the static pressure on both sides of the membrane. The condenser microphone is characterized by low background noise, thus providing a very low threshold detectability, and exceptional stability, since its operation is based on a purely geometric effect. Disadvantages include a limited low-frequency response as a result of both the finite charging time of the capacitor and the finite pressure equalization time of the capillary vent; and an extremely high output impedance, which requires a preamplifier (i.e., follower circuit) in close proximity to the microphone cartridge.

An *electret* material possesses a permanent electrical dipole moment (Sessler and West, 1973). When used in a condenser microphone, it provides the polarization voltage between the membrane and backplate in place of an external supply voltage. Electret materials are generally high-resistivity polymers, a prime example being PTFE (Teflon™). They are fabricated by heating a film of the material almost to its melting point and subjecting it to an intense electric field. When the material is cooled to room temperature, the net dipole moment is "frozen in." A typical construction is shown in Fig. 2(b), in which the electret is preferably bonded to the backplate rather than the membrane. The acoustic performance of the "backelectret" microphone is not much different from that of the air condenser microphone. The major advantage is the absence of a high voltage power supply and the extensive filtering needed to obtain a low noise floor—features greatly enhancing its miniaturization potential. A disadvantage is a slow loss of surface charge with time, a process which accelerates with increasing temperature.

A *ceramic microphone* uses the piezoelectric effect, the generation of surface charge density as the result of an applied stress (Hansen, 1969). Traditionally, the sensing element is constructed of a piezoelectric ceramic, such as lead zirconate titanate (PZT). The construction of a ceramic microphone is illustrated in Fig. 2(c). The sound pressure on the membrane is transmitted to the ceramic element through a connecting rod. The sole purpose of the backplate is to provide the required mechanical damping of the membrane. Ceramic microphones are characterized by ruggedness, low cost, and simple electronics. However, the recent trend has been to replace them with low-cost electrets.

Piezoresistivity is the dependence of electrical resistivity on mechanical stress or strain (Kim and Wise, 1983). The sensing element of a piezoresistive microphone is a thin, micromachined, silicon wafer, on which dopant is diffused or implanted to form the resistors of a Wheatstone bridge, as shown in Fig. 2(d). Acoustic excitation generates a time-varying stress in the resistors and a proportional voltage at the output of the bridge. Advantages include small size and low output impedance, making possible remote control electronics and thus installation in limited confines, as well as response to static pressure. Suited for high acoustic intensities, it is a favorite for wind tunnel and aerospace testing. A disadvantage is the temperature sensitivity of the resistors, requiring rather sophisticated compensation techniques.

2.1.2 Microphone Specifications A microphone user examining a specification sheet will usually find the items listed in Table 3. It is instructive to examine their meanings and implications. In addition, there is usually a variety of specifications regarding environmental conditions (see Sec. 2.1.5).

The *nominal size* refers to the outside diameter of the cartridge, which is slightly larger than the active membrane diameter. The *open-circuit sensitivity* is the output voltage per unit sound pressure without the loading of the preamplifier input impedance. Sometimes this specification is given in dB relative to a sensitivity of 1 V/Pa. For example, an open-circuit sensitivity of −40 dB is equivalent to 10 mV/Pa.

In a strict sense, the membrane *resonance frequency* is the fundamental resonance frequency of the membrane in vacuum. In practical terms, it is the frequency at which the membrane displacement in air lags the applied pressure by 90°, which is different from the vacuum resonance frequency because of damping and loading by the air layer. At the

Table 3. Representative microphone specifications.

Microphone type	Air condenser		Electret	Ceramic	Piezoresistive
Nominal size (in.)	1	$\frac{1}{2}$	$\frac{1}{2}$	1	0.170
Open-circuit sensitivity (mV/Pa)	50	15	10	10	0.03
Resonance frequency (Hz)	8000	25 000	...	...	45 000
Frequency range, −2 dB (Hz)	2–7000	4–20 000	4–20 000	2–12 000	0–10 000
Dynamic range (dB)	15–145	25–160	30–145	25–150	100–190
Capacitance (pF)	60	20	30	400	...
Equivalent air volume (cm^3)	0.15	0.01	0.015	0.5	...

upper cutoff frequency, the microphone sensitivity falls 2–3 dB on the high-frequency side of the damped membrane response. The specification is shown for pressure microphones and is different for free-field microphones (see Sec. 2.1.3). The *lower cutoff frequency* is discussed in Sec. 2.2.

The *dynamic range* is the range of sound-pressure amplitudes over which the microphone operates linearly, usually specified in dB. The upper limit is determined by total harmonic distortion, typically taken at 3% or 4%, and is usually of electrical origin. For a given sound pressure, the harmonic distortion is less for small microphones because of a lower output voltage. The lower limit of the dynamic range is determined by the *background noise* (see Sec. 2.3).

The *capacitance* is the membrane-backplate capacitance with the polarization voltage applied (for condenser microphones) or the piezoelectric element capacitance (for ceramic microphones). This is an important parameter in the determination of sensitivity and low-frequency response. A high capacitance helps reduce the background noise.

Below resonance, the membrane motion is controlled by its tension; i.e., its acoustic impedance is essentially a compliant reactance. The *equivalent air volume* is the volume of an enclosure that would present the same acoustic compliance as the membrane. If the microphone cartridge is inserted into a coupler as for calibration purposes (see Sec. 2.1.4), then the equivalent volume must be added to the volume of the coupler.

2.1.3 Directional Properties A microphone will not disturb a sound field if its dimensions are much smaller than the wavelength of the incident sound. For this reason, the preamplifier (or adapter) is made as compact as possible, having a diameter not exceeding that of the microphone cartridge. Since the pressure distribution is uniform over the membrane area, the response of the microphone to this type of excitation is called the "pressure response." An example is shown in Fig. 3. In this case, the mechanical damping of the microphone is designed for maximum flatness, corresponding to a mechanical quality factor of $Q \approx 0.7$, and the microphone is called a "pressure microphone." Most calibration methods provide a uniform pressure distribution and thus yield the pressure response.

As the wavelength in a free sound field approaches the dimensions of the microphone, reflection and refraction cause considerable changes in the pressure distribution about the microphone. Figure 4 shows the mean pressure over the membrane surface vs frequency for different angles of incidence. The response of the microphone under this condition is called the "free-field response." At normal incidence ($\theta = 0°$), the free-field effects are greatest; at grazing incidence ($\theta = 90°$), the pressure distribution is about the same as for the uniform pressure condition. The random-incidence response curve can be regarded as the mean response when incidence from all directions is equally probable. A set of curves as shown in Fig. 4 can be used to correct the free-field response to an equivalent pressure response. A "free-field microphone" is intentionally made overdamped to yield the flattest frequency response for nor-

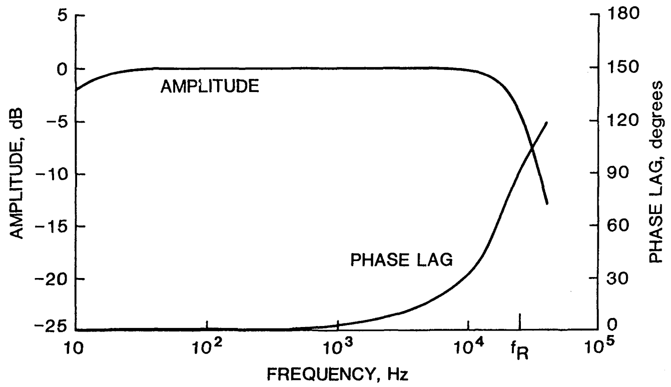

FIG. 3. Pressure response of a $\frac{1}{2}$-in. air condenser microphone. At the resonance frequency f_R, the membrane displacement lags the incident pressure by 90°.

mal incidence. As a result, the upper cutoff frequency far exceeds that of the pressure response, sometimes by as much as a factor of 2. The upper cutoff frequencies shown in Table 3 are for pressure microphones.

2.1.4 Microphone Calibration The three most widely used methods for microphone calibration are the electrostatic actuator, the pistonphone, and the reciprocity procedure.

The *electrostatic actuator* is a flat metallic electrode positioned at a nominal distance d from the microphone membrane. A voltage applied between the electrode and membrane produces a uniform electrostatic pressure on the membrane. If an ac voltage v_a is superimposed upon a high dc polarization voltage V_a, then the applied ac voltage and resulting electrostatic pressure have the same frequency. The membrane responds to the electrostatic pressure as it would to a sound pressure. The electrode is slotted to relieve acoustic loading by the air between the electrode and membrane. The electrostatic pressure exciting the membrane is

$$p = \epsilon_0 V_a v_a / d^2, \qquad (13)$$

where ϵ_0 is the dielectric permittivity of air. With typical values V_a = 800 V, v_a = 30 V rms, and d = 0.0005 m, the rms pressure is 0.85 Pa, or 93 dB. The method is excellent for obtaining frequency response but is not suitable for absolute calibration because of a twofold uncertainty in the distance d. First, the slots in the electrode necessitate a theoretically derived correction, and second, the polarization voltage shifts the equilibrium position of the membrane toward the actuator electrode.

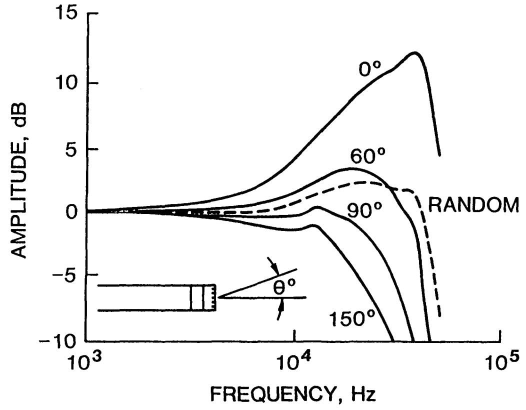

FIG. 4. Free-field response of a $\frac{1}{2}$-in. condenser microphone. (Redrawn with permission, courtesy of Bruel & Kjaer Instruments.)

The essential parts of a *pistonphone* are a coupler, into which the microphone cartridge is inserted and sealed, usually with an O ring, and a vibrating piston of known displacement. The piston may be driven by a cam having a sinusoidal contour, generating a pure tone, or by a crankshaft, which in addition produces considerable second harmonic distortion. The frequency is controlled through the speed of the cam or crankshaft. The pressure generated in the coupler is

$$p = \gamma P_0 A s / V, \qquad (14)$$

where P_0 is the static pressure, A the piston area, s the rms stroke, and V volume of the coupler (including the equivalent volume of the microphone). With typical values of γ = 1.4, $P_0 = 10^5$ Pa (1 atm), $A = 10^{-5}$ m^2, s = 1.4 × 10^{-4} m, and $V = 2 \times 10^{-5}$ m^3, the rms pressure is 9.8 Pa, or 114 dB SPL. At audio frequencies, a precision of approximately 0.2 dB is attainable. At low frequencies, a correction is needed for nonadiabatic compression. In the form of a portable, battery-operated device, the pistonphone is ideally suited for quick calibration of microphones in the field. Two limitations are fixed amplitude and relatively low operating frequency (several hundred Hz maximum). Devices using moving-coil drivers (without servo control) are not true pistonphones, for the generated volume velocity depends on the acoustic impedance of the load.

The *reciprocity procedure* (Wong and Embleton, 1979) is based on the following electromechanical principle. When a reversible transducer is operated as a receiver, the ratio of open-circuit output voltage v to applied acoustic pressure p will equal some constant A. Then when it is operated as a transmitter, the ratio of the generated volume velocity U to the input current I will equal the same constant A, if the radiation impedance seen by the transmitter is small. According to this procedure, three transducers are placed pairwise in an acoustic coupler—a transmitter T, a reversible transducer R, and the test microphone M. In step 1, transmitter T generates a sound pressure p_{T} to excite transducer R as receiver, resulting in an open-circuit voltage $v_{\mathrm{R}} = A_{\mathrm{R}} p_{\mathrm{T}}$. In step 2, test microphone M replaces transducer R to yield $v_{\mathrm{M}} = A_{\mathrm{M}} p_{\mathrm{T}}$. These lead to the relationship

$$A_{\mathrm{M}} = A_{\mathrm{R}} v_{\mathrm{M}} / v_{\mathrm{R}}. \tag{15}$$

In step 3, transducer R, as transmitter, excites test microphone M, resulting in $v_{\mathrm{M'}} = A_{\mathrm{M}} p_{\mathrm{R}}$. However, the above stated reciprocity property, $U_{\mathrm{R}} = A_{\mathrm{R}} I_{\mathrm{R}}$, and the known acoustic impedance of the coupler, $Z_{\mathrm{C}} = p_{\mathrm{R}}/U_{\mathrm{R}}$, lead to the relationship

$$A_{\mathrm{R}} = \frac{U_{\mathrm{R}}}{I_{\mathrm{R}}} = \frac{p_{\mathrm{R}}}{Z_{\mathrm{C}} I_{\mathrm{R}}} = \frac{v_{\mathrm{M'}}}{A_{\mathrm{M}} Z_{\mathrm{C}} I_{\mathrm{R}}}. \tag{16}$$

Substitution of Eq. (16) into (15) yields

$$A_{\mathrm{M}} = (v_{\mathrm{M}} v_{\mathrm{M'}} / Z_{\mathrm{C}} v_{\mathrm{R}} I_{\mathrm{R}})^{1/2}. \tag{17}$$

Thus, the microphone sensitivity depends only on electrical quantities, which are measurable to high precision, and a readily determinable acoustic impedance. The reciprocity method, the most precise of all known methods, can achieve absolute precisions of the order of hundredths of a dB.

2.1.5 Microphone Performance in Harsh Environments An increase in ambient temperature has three primary effects on condenser microphone parameters: a decrease in membrane tension, an increase in membrane-backplate gap, and an increase in air viscosity. The first two have compensating effects on the midband sensitivity. The last increases the membrane damping and is important only near resonance. As a result, the midband sensitivity of condenser microphones generally has a small temperature coefficient, typically <0.01 dB/°C over an interval from −50 to +60 °C. The primary effect for the ceramic microphone is the temperature dependence of the piezoelectric modulus; for the piezoresistive microphone, the temperature dependence of both the resistivity and the piezoresistive coefficient.

The effect of increasing pressure upon microphones is mainly an increase in stiffness of the air layer behind the membrane. Since most microphones are designed to subordinate this to the membrane stiffness, a change in ambient pressure has little effect upon midband sensitivity but has a strong influence on the membrane damping. For a condenser microphone, the pressure coefficient of the midband sensitivity is typically -10^{-5} dB/Pa.

Humidity is detrimental to condenser microphone performance primarily when it condenses and short-circuits the membrane to the backplate. A dehumidifier, containing a desiccant and inserted between the preamplifier and cartridge, has proved successful in keeping the interior of the cartridge dry. High relative humidities apparently do not affect the surface charge of an electret significantly. The immunity of the ceramic microphone from harmful effects of humidity is one reason for its popularity for many years.

The *vibration sensitivity* of a microphone depends on the direction in which the vibration is applied and is maximum when this direction is normal to the membrane surface. In this case, the equivalent acoustic pressure is

$$p = \sigma_M a, \tag{18}$$

where σ_M is the membrane surface density and

a the acceleration of the applied vibration. For a 5×10^{-6}-m-thick nickel membrane, having a surface density $\sigma_M = 0.0445\ \text{kg/m}^2$, an applied acceleration $a = 9.8\ \text{m/s}^2$ (1 g) will produce an equivalent sound pressure of 0.43 Pa or 87 dB.

In outdoor measurements, the ambient wind will generate considerable noise in a microphone and may disturb the intended measurement. A possible countermeasure is to install a windscreen, constructed of a fibrous material, either self-supporting or supported on a wire frame, about the microphone. In principle, the material displays a high flow resistance to the quasistatic pressure of the wind but low resistance to acoustic pressures. The windscreen is most effective at high frequencies and in a wind direction normal to the membrane. Overall, in moderate winds (<30 km/h) a windscreen may reduce the wind noise over the audio band by as much as 10–20 dB.

Often it is necessary to make sound-pressure measurements in extremely hostile or inaccessible locations, for example, in jet engine exhausts or in the ear canal. Such measurements can be realized with the aid of a *probe tube*—a long, thin, hard-walled tube, having one end positioned in the acoustic test environment and the other coupled to a microphone. Ideally, the sound pressure at the microphone will be the same as the test pressure at the probe tip. This will be approximately true at long wavelengths or when the impedance of a damping material placed in one end matches the characteristic impedance of the tube. Otherwise, the natural tube resonances will cause an undulating frequency response which requires a correction.

2.2 Infrasonic Measurements

Measurements of *infrasound* (below 20 Hz), a frequency region which includes sonic booms, thunder, earthquakes, rocket launches, vocalizations of large animals, etc., require special modifications of a microphone system. There are two reasons for the low-frequency rolloff of a condenser microphone using a polarization voltage: an electrical rolloff because of the finite charging time of the membrane-backplate capacitor, and a mechanical rolloff as a result of the finite time for pressure equalization through the capillary vent. Both the electrical and mechanical cutoff frequencies can be controlled through the choice of design components and are usually set equal to one another.

For infrasonic measurements, the mechanical rolloff can be eliminated by sealing the capillary vent and the electrical rolloff through the use of a microphone carrier system. Here the time-varying capacitance of a condenser microphone is used to produce amplitude modulation of a carrier voltage generated by an electrical oscillator. Such a system responds to static changes in microphone capacitance, in other words to frequencies down to dc. A feedback arrangement can be applied to compensate for slow variations in ambient pressure, but this eliminates the low end of the frequency response corresponding to the compensation response time.

2.3 Background Noise in Microphones

The *background noise* of a microphone is defined as the rms output voltage in the absence of acoustical excitation. It is one of the most important specifications because it determines the minimum resolvable sound pressure (Tarnow, 1987). Since acoustic applications generally require a signal-to-noise (S/N) ratio of at least 1:1 (0 dB), the lower limit of the microphone dynamic range is usually specified as equal to, or 5 dB above, the background noise level.

The two most important sources of microphone noise originate from

1. viscous motion of the membrane (mechanical Johnson noise), and
2. electronic motion in the preamplifier components (electrical Johnson, shot, and $1/f$ noise).

The availability of high-quality field-effect transistors has reduced the preamplifier noise to a secondary role. Today it is possible to produce 1-in. condenser microphone systems having a noise floor near the threshold of human hearing (0 dB) over the audible range of frequencies. Interestingly, the membrane displacement of a 1-in. microphone with a 200-V polarization voltage, responding to an SPL of 0 dB, is 10^{-13} m. Systematic studies of noise in other types of microphones have not received the thoroughness accorded the condenser microphone.

3. MEASUREMENT OF SOUND INTENSITY AND POWER

3.1 Sound Intensity

Sound intensity is the sound energy passing through a unit normal area per unit time (Gade, 1982). It is a vector quantity having units of watts per square meter (W/m^2). This definition can be expressed in terms of the following fundamental acoustic parameters:

$$\bar{\mathbf{I}} = \overline{p(t)\mathbf{u}(t)}, \tag{19}$$

where the pressure and particle velocity are time-varying quantities and the overbar denotes a time average. For harmonic waves,

$$\bar{\mathbf{I}} = \mathrm{Re}\overline{p\mathbf{u}^*} = \overline{p\mathbf{u}}\cos\theta, \tag{20}$$

where Re denotes the real part, the asterisk a complex conjugate, and θ the temporal phase angle between $\bar{p}$ and $\bar{\mathbf{u}}$. For illustration, let us apply Eq. (20) to two simple examples. For a plane wave, $\bar{u} = \bar{p}/\rho c$, $\theta = 0$, and $\bar{I} = \bar{p}^2/\rho c$, where ρ is the air density and c the sound speed. For a standing wave, $\bar{u} = \bar{p}/\rho c$, $\theta = \pi/2$, and $\bar{\mathbf{I}} = 0$. To avoid the difficult measurement of sound particle velocity, we invoke Newton's second law,

$$\mathbf{u} = \rho^{-1}\int\nabla p dt, \tag{21}$$

where ∇p is the pressure gradient. Substituting Eq. (21) into Eq. (19) yields an expression in p alone:

$$\bar{\mathbf{I}} = -\overline{(p/\rho)\int\nabla p dt}. \tag{22}$$

In a practical measurement system, two microphones are aligned coaxially face to face, or in a coplanar arrangement, and separated by a fixed distance Δr in the sound field (Fig. 5). As long as $\Delta r \ll \lambda$, the wavelength, then Eq. (22) can be approximated as follows:

$$I_r = -\frac{1}{\rho}\left\langle\frac{(p_A + p_B)}{2}\int\frac{(p_A - p_B)}{\Delta r}dt\right\rangle. \tag{23}$$

Note that p is represented by the mean value of p_A and p_B. Following the measurement of the sound pressures p_A and p_B, the sound intensity is evaluated by the signal processing operations of summing, integrating, multiplying, and time averaging. It is good practice to repeat a measurement with the microphones switched in order to cancel the effect of phase differences between channels. Equation (23) yields the intensity component along the microphone axis. For sound propagation in a direction α relative to this axis, I_r must be divided by cosα. In a coplanar arrangement, the reference axis passes through both microphone axes at right angles.

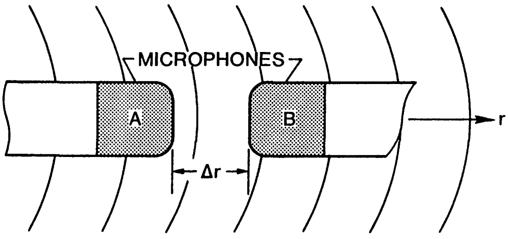

FIG. 5. Measurement of sound intensity with a two-microphone arrangement. The microphones can also be aligned with coplanar membranes.

A second measurement method is based on the fact that the time-averaging operation indicated in Eq. (22) is closely related to the cross-power spectrum G_{AB} between p_A and p_B. Formal analysis yields

$$I_r = -(\omega\rho\Delta r)^{-1}\,\mathrm{Im}(G_{AB}), \tag{24}$$

where Im denotes the imaginary part. If the microphone signals are applied to a correlator or digitally to a computer, Eq. (24) can be used to compute the intensity.

Applications of sound-intensity measurement include sound-power measurement (Sec. 3.3) and intensity mapping to locate sound sources and sinks.

3.2 Acoustic Enclosures

Before discussing sound-power measurement, we shall find it helpful to discuss four basic types of acoustic enclosure: anechoic, reverberant, resonant, and Helmholtz. These are depicted in Figs. 6(a)–6(d).

An *anechoic chamber* has no reflections (echoes) from its walls and thus simulates free-field conditions [Fig. 6(a)]. The anechoic condition is realized by lining the walls of the chamber with wedges made of a porous, absorptive material such as rock wool, glass wool, or foam. Typical wedge dimensions are 20–60 cm at the base and a wedge angle of

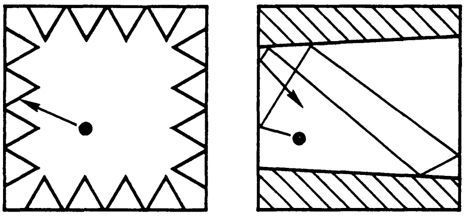

(a) ANECHOIC (b) REVERBERANT

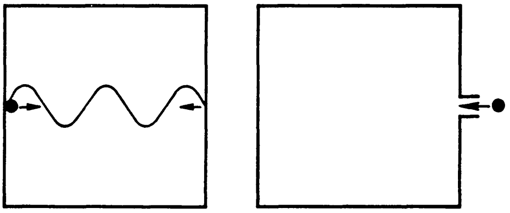

(c) RESONANT (d) HELMHOLZ

FIG. 6. Acoustic enclosures. (a) Anechoic: Waves from the sound source (dark circle) are absorbed without reflection. (b) Reverberant: Waves experience multiple reflections to establish a uniform sound energy density. (c) Resonant: Reflected waves reinforce waves generated from the source to establish a standing wave. (d) Helmholtz: At long wavelengths, the chamber behaves as a compliance and its small opening as an acoustic mass.

10–15°. Since the floor must also be lined, an enclosure usually has a wire mesh just above the wedge tips to support personnel and equipment. The free-field approximation is truest at high frequencies. The low-frequency limit is determined by the criterion

$$h/\lambda = (1 + R_w)/(1 - R_w), \tag{25}$$

where h is the appropriate chamber dimension and R_w the reflection coefficient of the wedge along its axis. Using typical values $R_w = 0.1$ and $h = 10$ m, we find $\lambda < 8.33$ m, corresponding to a minimum operating frequency of 40 Hz. The background noise level of a good anechoic chamber may lie near 0 dB SPL over most of the entire audio range. However, the mark of quality is how well the sound pressure from a point source adheres to the $1/r$ spherical spreading law, which is perturbed by reflections from the walls. Within confines no closer than about a meter from the walls, verification of the law to within 1 dB is a reasonable design goal.

A *reverberation chamber* is used to produce ideally a spatially uniform sound-energy density. The walls are hard and highly reflective, such that a sound ray emitted from a source will experience multiple reflections in haphazard fashion, as shown in Fig. 6(b), and will eventually fill the room. When the source is turned on, the energy density builds up until dissipation balances the sound power emitted by the source. The resulting sound field is called a diffuse field, independent of location or direction, a situation difficult to achieve in practice. Hard reflecting objects and moving reflectors enhance the diffuseness. The volume V of the room should exceed $3\lambda^3$ for octave analysis and $9\lambda^3$ for third-octave analysis (see Sec. 6.4), where λ is the largest wavelength. Recommended ratios for room dimensions are $1:2^{1/3}:4^{1/3}$. The source should be located at least $\lambda/4$ from the walls, and microphones should be at locations removed from known peaks and valleys in sound pressure. The reverberation chamber is widely used for measurement of sound absorption of materials and sound power emission of sources.

An enclosure is said to be *resonant* if a reflected wave returns to the source, in the direction from which it was emitted, after progressing an integral number of wavelengths. The returning waves reinforce the emitted waves to produce a standing wave pattern, which grows in amplitude until dissipation equals emission. In Fig. 6(c), a plane wave is reflected between parallel walls. The resonance frequency for this one-dimensional case is

$$f_R = c/\lambda = nc/2h, \tag{26}$$

where h is the distance between the walls and n an integer. In two and three dimensions, the standing wave patterns can become quite complex. Resonance conditions also exist for bound cylindrical and spherical waves.

A *Helmholtz resonator* is the acoustic analog of a mass-spring system. Two acoustic elements are coupled together: a chamber and an opening in the form of an orifice or tube [see Fig. 6(d)]. At wavelengths much greater than the element dimensions, the air in the opening moves as a unit and thus behaves as an acoustic mass; the chamber behaves as a compliance. The resonance frequency of the mass-compliance system is

$$f_R = (c/2\pi)\sqrt{S/lV}, \tag{27}$$

where S is the cross-sectional area and l the effective length of the opening, and V is the volume of the chamber. This frequency is much lower than the natural frequencies of the chamber alone. At one time, Helmholtz resonators were used to determine the pitch of musical tones.

3.3 Sound Power

Sound power (ANSI, 1979) is the integrated normal intensity over a surface S enclosing an acoustic source in units of watts:

$$W = \iint_S \bar{\mathbf{I}} \cdot d\mathbf{S}. \tag{28}$$

In the absence of absorption, the sound power is independent of distance from the source. Three methods for measuring sound power are free field, diffuse field, and reference source.

The *free-field* method assumes that the sound field is far field, spherically spreading, and in an anechoic environment. The sound source must be located either outdoors or in an anechoic chamber in order to approximate free-field conditions. Microphone measurement stations are located on an imaginary sphere (source suspended) or hemisphere (source grounded), having a radius r several times greater than the longest wavelength. The sound power is determined from the mean value p_m^2 of the squares of the sound pressure measurements at all the microphone stations:

$$W = (p_m^2/\rho c)(4\pi r^2)F, \tag{29}$$

in which the directivity factor F has the value 1 if the source is suspended or 0.5 if the source is grounded on a perfectly reflecting surface. At 20 °C, 1 atm, $\rho c = 415\ \mathrm{N \cdot s/m^3}$ in air. For example, if $p_m = 20\ \mu\mathrm{Pa}$, $4\pi r^2 = 1\ \mathrm{m}^2$, $F = 1$, then $W = 10^{-12}$ W.

If the sound intensity, using microphone pairs, is measured instead of the sound pressure alone, the requirements on far-field, spherically spreading sound propagation can be relaxed.

The *diffuse-field* method is based on the fact that the sound-energy density is uniform throughout a reverberant enclosure. If the sound source is placed in a good reverberation chamber, then the energy density adjusts to a steady-state condition, whereby the sound power emitted by the source balances the power dissipated in the chamber. The experimental arrangement is similar to that used in free-field measurements, but now the sound power becomes

$$W = p_m^2 R/4\rho c. \tag{30}$$

Here R is the reverberation constant of the chamber and can be determined from the measurement of reverberation time, whereby a controlled sound source is excited and switched off and the free decay time measured.

Often the enclosure about a sound source is neither anechoic nor reverberant, and accommodation to acoustic requirements is not possible—for example, if the source is a heavy machine. In such a case, the *reference source* method may prove useful. Here an identical set of measurements is taken both for the test source and for a reference source of known sound power. Then

$$W = W_r p_m^2 / p_{mr}^2, \tag{31}$$

where W_r and p_{mr}^2 are the sound power and mean square sound pressure for the reference source.

4. MEASUREMENT OF ACOUSTIC PARTICLE VELOCITY

The *acoustic particle velocity* **u** refers to the particle motion, back and forth about its equilibrium position, imparted by a passing sound wave. This difficult measurement is generally avoided if the acoustic particle velocity can be related simply to the sound pressure. Otherwise, one approach is to exploit the relationship between particle velocity and sound-pressure gradient given in Eq. (21).

Measurement of a *sound-pressure gradient* can be achieved by any of the principles used to measure sound pressure alone. A highly successful device is the foil-electret "gradient microphone" (Sessler and West, 1969) illustrated in Fig. 7. It is assumed that the microphone is sufficiently small not to perturb the sound field, requiring $\Delta l < \lambda/2$, and that the sound pressures on either side of the

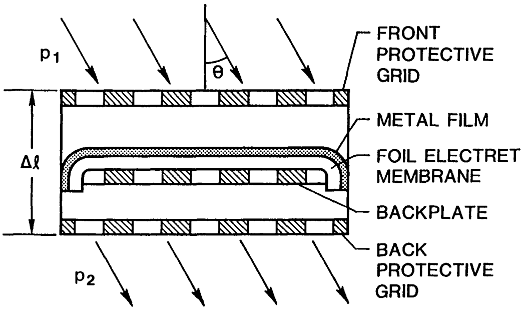

FIG. 7. Foil-electret gradient microphone, which can be used to measure acoustic particle velocity.

membrane are the same as those at the protective grids, namely p_1 and p_2.

A metallized electret foil is stretched tightly across, and in contact with, the backplate. Thus, the response is compliance-controlled, although a mass-controlled design is feasible. An air gap exists only because of irregularities in the backplate surface. Holes in the backplate and in both protective grids permit the sound wave to gain access to both sides of the membrane. The membrane responds to the net normal component of sound pressure:

$$p_i = |-\nabla p \Delta l' \cos\theta|,$$

which is related to the sound particle velocity u through Eq. (21). For a harmonic wave of frequency f,

$$p_i = 2\pi f u \rho_0 \Delta l' \cos\theta, \qquad (32)$$

where ρ_0 is the static density of air and $\Delta l'$ the effective grid spacing (a little longer than Δl). Measurement of p_i thus determines u since all other quantities in Eq. (32) are known. Measurement of the response at constant amplitude but increasing frequency is a good check as to whether the microphone is truly responding to pressure gradient.

The *Rayleigh disk* is based on the relationship between the particle velocity and the torque exerted on a suspended disk by a sound wave. The *hot-wire anemometer* is not suitable for purely acoustic excitation but has been used successfully for acoustic excitation superimposed on a mean flow.

5. MEASUREMENT OF ACOUSTIC IMPEDANCE

5.1 Introductory Concepts

The relationship between sound pressure and acoustic flow velocity plays a central role in the analysis of acoustic devices, such as mufflers and musical instruments, and in the determination of a sound field in the presence of a boundary. Quantitatively this relationship is described by one of three types of impedance used in acoustics. The *acoustic impedance* $Z = p/U$, the ratio of sound pressure to volume velocity, is used in analyzing acoustic circuits where devices are represented by equivalent lumped elements. It has units of $\mathrm{N \cdot s/m^5} = \mathrm{Rayl/m^2}$ (see below). The "volume velocity" is related to the particle velocity by $U = uS$, where S is the area crossed by the fluid.

The interaction between a sound wave and a boundary depends on the *specific acoustic impedance* of the boundary relative to that of the propagation medium. The specific acoustic impedance is the ratio of sound pressure to acoustic particle velocity $z = p/u$, which for a plane wave is $z = \rho c$ in units of $\mathrm{N \cdot s/m^3}$ = Rayl. It is related to Z through $z = ZS$.

The *mechanical* or *radiation impedance*, the ratio of force to particle velocity, is of interest in systems containing both discrete and continuous components but is not discussed here.

Methods of measuring acoustic impedance fall into three broad categories:

1. impedance-tube and waveguide methods in general;
2. direct measurement of sound pressure and particle velocity; and
3. free-field methods.

5.2 Impedance Tube

The test specimen is located at one end of a rigid *impedance tube*, and a transmitter at the other end [Fig. 8(a)]. It is important to distinguish between materials of "local reaction" and those of "extended reaction." In the former, the behavior of one point on the surface depends only on excitation at that point and not on events taking place elsewhere in the material. In the latter, acoustic excitation at a point on the surface generates

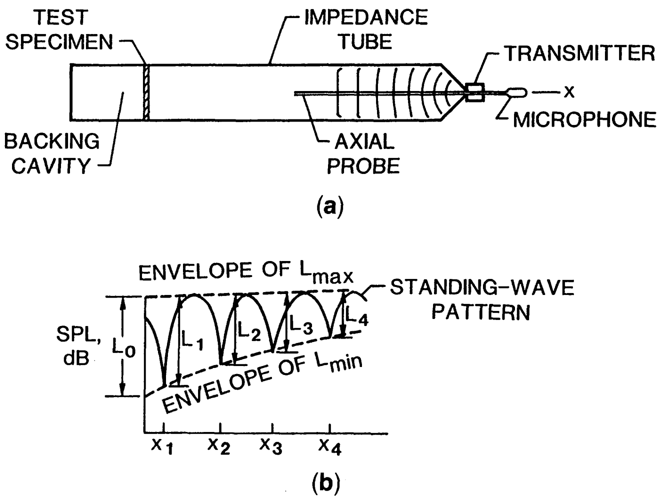

FIG. 8. Measurement of acoustic impedance with an impedance tube. (a) Impedance tube and (b) standing-wave pattern and its envelopes.

waves that propagate laterally throughout the material. Generally, a material is locally reacting if normal acoustic penetration does not exceed a wavelength.

For a locally reacting material, a thin test specimen (thickness $\leqslant \lambda/4$) is backed by a $\lambda/4$ air gap sandwiched between the specimen and a massive reflector (ASTM, 1992). For a material of extended reaction, a specimen of approximate thickness $\lambda/4$ is backed by a massive reflector directly against its surface.

The transmitter is tuned to establish a standing wave pattern, which is probed by a microphone located either within the tube or at the end of a probe tube. The observer slides the probe along the impedance tube axis and records the standing wave pattern $L(x)$ in dB (Fig. 8b). Here $L(x) = 20 \log[p(x)/p_0]$, where the reference pressure p_0 is immaterial. The impedance is evaluated from the pressure standing-wave ratio L_0 at the specimen surface, which cannot be measured directly but is computed from a best fit to the trend of $L_{\max}$ and $L_{\min}$ shown in the figure. After computing the antilog of the pressure standing wave ratio K_0 and related quantities,

$$K_0 = 10^{L_0/20}, \tag{33}$$

$$\phi = \left(\frac{x_1}{x_2 - x_1} - \tfrac{1}{2}\right) \times 360^\circ, \tag{34}$$

$$M = \tfrac{1}{2}(K_0 + K_0^{-1}), \tag{35}$$

and

$$N = \tfrac{1}{2}(K_0 - K_0^{-1}), \tag{36}$$

we determine the real (z') and imaginary (z'') parts of the specific acoustic impedance relative to that of air, ρc:

$$z'/\rho c = 1/(M - N\cos\phi), \tag{37}$$

$$z''/\rho c = N\sin\phi/(M - N\cos\phi). \tag{38}$$

This method is capable of yielding measurements of high precision, dependent upon a precise determination of distance x. A disadvantage is the large number of measurements needed to record the standing wave.

5.3 Direct Measurement of Sound Pressure and Volume Velocity

The difficulty with the *direct measurement* method lies in the measurement of the volume velocity (see Sec. 4). The most common method of attacking the latter problem is to control the volume velocity at the transmitter. This can be achieved in several ways:

1. by mounting a displacement sensor or accelerometer on the driver (Salava, 1974);
2. by using a dual driver, directing one side to the test region and the other to a known impedance Z_k, and using $U = p/Z_k$ (Singh and Schary, 1978); or
3. by using a cam to excite a driving piston

so that the generated volume velocity will be independent of the acoustic load (Zuckerwar, 1983).

The first two methods rely on the integrity of the velocity measurement technique; the third is limited to relatively low frequencies.

5.4 Free-Field Methods

A transmitter sends an incident wave at an angle ψ toward the test specimen, from which it is reflected, also at an angle ψ, toward a receiver (Fig. 9). The specific acoustic impedance is evaluated from the reflection coefficient $R_p = p_r/p_i$:

$$\frac{z}{\rho c} = \frac{1}{\sin\phi}\left(\frac{1+R_p}{1-R_p}\right). \tag{39}$$

A variety of techniques, both transient and steady state, has been devised to determine the three wave components p_r, p_i, and p_d. One steady-state method uses three separate measurements at each frequency (Jones and Stradulinsky, 1977):

1. with the specimen in place, yielding $p_1 = p_r + p_d$;
2. with the specimen replaced by a reflector of high impedance, yielding $p_2 = p_r' + p_d \approx p_i + p_d$; and
3. with the reflector removed, yielding $p_3 = p_d$ alone. Thus,

$$R_p = \frac{p_r}{p_i} = \frac{p_1 - p_3}{p_2 - p_3}. \tag{40}$$

Free-field methods are used for testing materials at short wavelengths and are popular for outdoor measurements of the earth's ground surface.

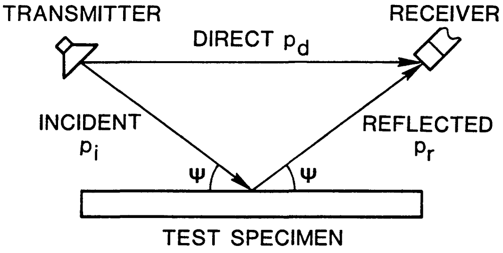

FIG. 9. Measurement of acoustic impedance by a free-field method.

6. FILTERS IN ACOUSTICS

6.1 Introductory Concepts

The representation of an acoustic time history in the domain of an integral (or discrete) transform has two advantages. First, it transforms an integro-differential (or difference) equation into a more tractable algebraic equation. Second, it often separates relevant signal from irrelevant signal and random noise. The two most common transforms used in acoustics are the Fourier transform for continuous time histories and the z transform for discrete (sampled) time histories. The Fourier transform represents a time history $f(t)$ in the frequency domain:

$$F(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\exp(-j\omega t)\,dt, \tag{41}$$

with $\omega = 2\pi f$. The z transform represents the sampled values $f(nT_s)$ in the z domain:

$$F(z) = \sum_{n=0}^{\infty} f(nT_s)z^{-n}, \tag{42}$$

where T_s is the sample interval and n the sample number. The representation of a time history in the transformed domain is called a "spectrum." We shall be concerned with the frequency spectrum.

Filters fulfill three major functions in acoustics: spectral selection, analysis, and shaping. It is assumed that the reader is familiar with the general characteristics of filters and with filter terminology (Randall and Upton, 1978; see also Sec. 8 of the article ELECTRONIC CIRCUITS).

6.2 Spectral Selection

We shall present two examples of *spectral selection*. The first is antialiasing. In sampled systems, a low-pass filter is essential to suppress all frequency components above the highest signal frequency, which must not exceed half the sampling frequency f_s to avoid "aliasing," that is, the appearance of components of frequency f_s–f in the observed spectrum. This is a consequence of the Nyquist sampling theorem. The maximum frequency for which the spectrum is uncorrupted by aliasing is called the "Nyquist frequency." The second example is S/N improvement. The observed signal is often a pure tone, for which narrow-band filtering will produce a considerable improvement in S/N ratio. If the noise is "white," that is, has a uniform spectral power density, then a reduction in bandwidth from B_W to B_N improves the S/N ratio by $10 \log(B_W/B_N)$ dB.

6.3 Spectral Analysis

The role of the filter in *spectral analysis* is to permit observation of a narrow portion of a wideband spectrum. The selected band is specified by a center frequency f_0 and a bandwidth B, defined as the frequency interval about f_0 where the output/input ratio remains within 3 dB of that at the center. The bandwidth of the filter may be constant (i.e., independent of f_0) or a constant percentage of f_0.

The constant-bandwidth filter is advantageous in cases where the measured spectrum is rich in detail over a limited frequency range, for example, where a series of harmonics appears as the result of nonlinear distortion or where a number of sharp resonances are generated from a complex sound source. The constant-percentage bandwidth filter is more appropriate in cases where the measured spectrum encompasses a large number of decades, say two or more; where the source is unstable, constantly shifting its prominent frequencies; or where the power transmitted over a band of frequencies is of interest, as in noise control engineering.

Popular choices for the constant-percentage bandwidth are the octave (factor of 2), $\frac{1}{3}$, $\frac{1}{6}$, $\frac{1}{12}$, and $\frac{1}{24}$ octave. The bandwidth of a $\frac{1}{3}$-octave filter, for example, is $2^{1/6}f_0 - 2^{-1/6}f_0 = 0.231f_0$. The $\frac{1}{3}$-octave filter, in fact; is the most widely used in acoustic spectral analysis for reasons rooted in human auditory response. The preferred $\frac{1}{3}$-octave center frequencies are 1.6, 2, 2.5, 3.15, 4, 5, 6.3, 8, and 10 (×10 to any integral power) Hz. The audible spectrum, 20 Hz–20 kHz, encompasses 31 $\frac{1}{3}$-octave bands.

6.4 Spectral Shaping

The perceived loudness of a tone of constant amplitude is a strong function of frequency and amplitude. Many acoustic instruments feature not only a linear response, but also a weighted response, which conforms to the frequency response of the human ear. The function of a weighting filter is to *shape* an acoustic spectrum to match the response of the ear. Three standard frequency-response curves A, B, and C conform to equal-loudness curves at 40, 70, and 100 phons, respectively. A "phon" is a unit of loudness, usually specified in dB; it is the same as the SPL at 1 kHz but differs at most other frequencies. The D weight is used for applications involving aircraft noise measurement. The filter response curves for the A, B, C, and D weightings are shown in Fig. 10.

6.5 Filter-Effect Error

Sound propagating through the Earth's atmosphere suffers an absorption varying approximately as the square of the frequency. Consequently, the spectrum at a source will be greatly distorted at a distant receiver. If the receiver contains a set of finite-bandwidth filters, then the incident sound pressure is greater at the lower band edge of a filter than at the upper band edge, as illustrated in Fig. 11. Since a reconstruction of the source spectrum based on measurements at the receiver requires evaluation of the total range attenuation, the question arises as to which frequency within a given band is most appropriate for the evaluation of the absorption constant. A representation of the filter by its center frequency f_0 is not fair, because it does not properly weight the domi-

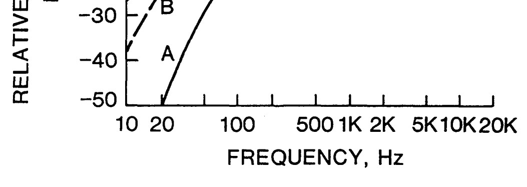

FIG. 10. Response curves of A, B, C, and D weighting filters.

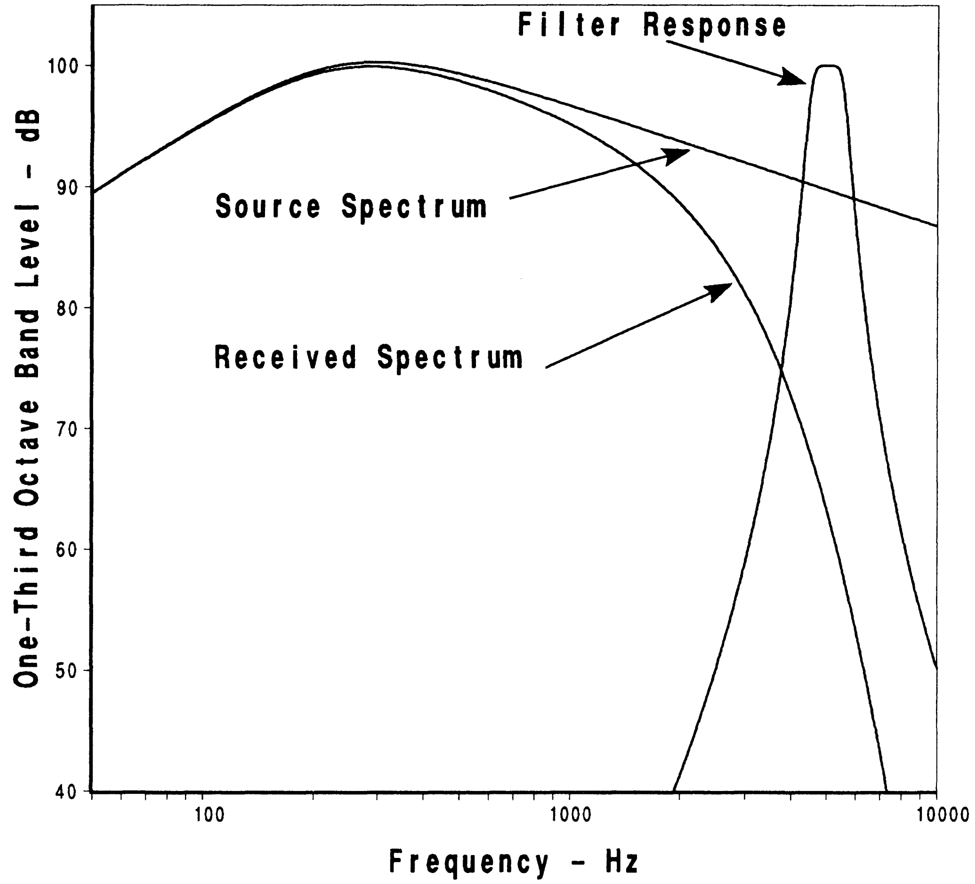

FIG. 11. Source spectrum (arbitrary), received spectrum (after attenuation by the atmosphere), and response of a selected $\frac{1}{3}$-octave narrow band filter. The received spectrum falls so steeply across the filter band that it is erroneous to represent the filter by its center frequency.

nating transmission near the lower band edge, and leads to an error in the reconstruction of the source spectrum known as the *filter-effect error*. The proper frequency to represent the band is that for which the received power, if concentrated at that frequency, would equal the integrated power over the band. For a constant-percentage, third-order Butterworth bandpass filter, the proper frequency f_B to represent the band is given by

$$f_B = f_0\{1 + [1 - 2\alpha(f_0)L]B^2/8\}, \tag{43}$$

where $\alpha(f_0)$ is the sound absorption at the band center, L the range between the source and receiver, and B the normalized bandwidth of the filter; e.g., $B = 0.231$ for a $\frac{1}{3}$-octave filter. Since in reality atmospheric absorption varies not quite as f^2, Eq. (43) is found to be more accurate if the factor 8 in the denominator is replaced by 10. For a pure tone, the frequency of the tone, of course, should be used to evaluate the absorption.

7. SPECTRUM ANALYZERS

7.1 Introduction

A *spectrum analyzer* enables an observer to view the frequency spectrum of an acoustic time history on an output device such as a monitor, recorder, or printer. A real-time analyzer produces a complete, continuously updated spectrum without interruption. The first real-time analyzers were analog in nature, based on either of two principles:

1. time compression, which used a frequency transformation to speed up processing time; or
2. a parallel bank of analog filters and detectors.

The advent of VLSI (very large-scale integration) in the semiconductor industry made the all-digital real-time analyzer a reality, offering competitive cost and enhanced stability, linearity, and flexibility.

The spectrum analyzer performs the basic functions of preamplification, analog filtering, detection, analog-to-digital (A/D) conversion, logic control, computation, and output presentation (Thane, 1979). Digital real-time analyzers operate on either of two principles: the digital filter or the fast Fourier transform (FFT).

7.1.1 Digital Filter The transfer function of a two-pole analog filter is written

$$H(s) = \frac{(s + r_1)(s + r_2)}{(s + p_1)(s + p_2)}, \tag{44}$$

where s is the Laplace operator, $r_{1,2}$ the zeros, and $p_{1,2}$ the poles. The filter characteristics—gain and cutoff frequencies—are fixed and can be changed only by changing the components making up the filter. The frequency response can be found by replacing s by $j\omega$.

The *digital filter* accepts samples $f(nT_s)$ of the time history from an A/D converter, where T_s is the sample interval, and yields an output in the form of a sequence of numbers. The transfer function is represented in the z domain as

$$H(z) = \frac{A_0 + A_1 z^{-1} + A_2 z^{-2}}{1 - B_1 z^{-1} - B_2 z^{-2}}, \tag{45}$$

where $z^{-1} = \exp(-sT_s)$ is called the unit delay operator, since multiplication by z^{-1} is equivalent to delaying the sequence by one sample number. Synthesis of $H(z)$ requires a system that performs the basic operations of multiplying, summing, and delaying. Noteworthy is the fact that once the filter characteristics are set by choice of coefficients A_0–B_2, the frequency response parameters (center frequency f_0 and bandwidth B for a bandpass filter) are controlled by the sample rate $f_s = 1/T_s$. For example, the doubling of f_s doubles f_0 and B. Thus, the digital filter is a constant-percentage-bandwidth filter and is appropriate for those applications where such is required. Typically the filters are six-pole Butterworth or Chebycheff filters of $\frac{1}{3}$-octave bandwidth. Several two-pole filters can be cascaded to produce filters of higher poles, or the data can be recirculated through the same filter several times.

7.1.2 Fast Fourier Transform The discrete Fourier transform (DFT) is the digital version of Eq. (41):

$$F(k) = \frac{1}{N} \sum_{n=0}^{N-1} f(n) \exp(-j2\pi kn/N), \tag{46}$$

where $f(n)$ is the value of the nth time sample, k the frequency component number, and N the number of time samples (block size). The time resolution depends on the time window $\Delta t = T/N$, and the frequency resolution depends on the sampling frequency $\Delta f = f_s/2N$. Obviously, the filter is a constant-bandwidth filter and again is suited to the appropriate applications. In contrast to the digital filtering technique, the data throughput is not continuous but is segmented into data blocks. Thus, for real-time analysis the analyzer must be capable of processing one block of data while simultaneously acquiring a new block.

The FFT exploits the symmetry properties of the DFT to reduce the number of computations. The DFT requires N^2 multiplications to transform a data block of N samples from the time domain to the frequency domain; the FFT requires only $N \log_2 N$ multiplications. For a block size of $N = 1024$ samples, the reduction is over a factor of 100.

The DFT of Eq. (46) differs from the continuous Fourier transform of Eq. (41) in three ways, each presenting a data processing problem:

1. the frequency limitation to less than $f_s/2$;
2. the finite block size, which convolves the window spectrum with the signal spectrum and causes a leakage of power of a spectral component to its neighbors; and
3. the "picket fence" effect, which prevents the observation of a discrete spectrum between samples.

Examples of acoustic signals necessitating analysis in real time are signals in the form of a sequence of transients, as speech; aircraft flyover noise, as required by the Federal Aviation Administration; and measurements where the analyzer is an element in a control loop. For other types of signals, such as stationary or quasistationary signals, or transients shorter than the time window, the time history can be stored and analyzed at a later time.

7.2 Special Features

Many digitally based spectrum analyzers provide additional special features such as a zoom mode, transfer function, auto and cross correlation, auto and cross spectra, and coherence function.

A *zoom* capability permits an increase in spectral density about a selected band center. For a given block size, both the frequency resolution and width of the band become narrower. Two applications of the zoom mode are

1. to permit observation of the fine detail about a spectral line; and
2. to improve the signal-to-noise (S/N) ratio of a sharp spectral line, because of the reduced noise power accompanying reduced frequency resolution.

The remaining features pertain to a dual-channel analyzer, which permits comparison of data at one test station with those at another. Suppose a noise or vibration source produces responses at two stations $f_1(t)$ and $f_2(t)$ having Fourier transforms $F_1(f)$ and $F_2(f)$. The *transfer function* is defined as

$$H_{12}(f) = F_2(f)/F_1(f). \tag{47}$$

Consequently, measurement of $F_1(f)$ permits determination of $F_2(f)$. The *cross correlation* is expressed in terms of a time delay τ:

$$g_{12}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_0^T f_1(t) f_2(t+\tau) = dt. \tag{48}$$

The Fourier transform of this function is the *cross-power spectral density*:

$$G_{12}(f) = \int_{-\infty}^{\infty} g_{12}(\tau) \exp(-j2\pi f\tau)\, d\tau. \tag{49}$$

If $f_1(t)$ and $f_2(t)$ are the same signal, say at station 1, then Eqs. (48) and (49) yield $g_{11}(\tau)$ and $G_{11}(f)$, the *autocorrelation function* and the *spectral power-density function*. Two important acoustic applications of the cross functions are time-delay estimation and transfer-function determination. First, if a signal propagates between stations 1 and 2, the function $g_{12}(\tau)$ shows a peak when τ equals the time delay between the stations. Second, the spectral densities are related to the transfer function through

$$H_{12}(f) = G_{12}(f)/G_{11}(f). \tag{50}$$

Equation (50) permits determination of $H_{12}(f)$ while the source is operating in its natural condition. The *coherence function* is defined as

$$C_{12}(f) = |G_{12}(f)|^2/G_{11}(f)G_{22}(f), \tag{51}$$

which has a value between 0 and 1. This function serves as a criterion as to whether the signals received at stations 1 and 2 have the same cause. It should have a reasonably high value even in measurement systems subject to noise and random events.

8. SOUND-LEVEL METERS

A *sound-level meter* is a compact portable instrument, usually battery-operated, for measuring SPL at a selected location. Figure 12 shows the basic operations of the instrument. The microphone signal is preamplified (attenuated), weighted, again amplified (attenuated), detected, and displayed on an analog meter. The detector is a square-law detector followed by an averaging (mean or rms) network. There are a variety of additional features such as calibration, overload indication, and external connectors for filters and output signal.

The directional response of the microphone affects the accuracy of the measurement. In a free field, corrections are based on curves such as those in Fig. 4 if the angle of incidence is known. In a diffuse field, the random response curve must be relied upon: The smaller the microphone, the more accurate are the results.

Two switch selections available to the user are "weighting" and "time constant." The weighting networks are linear (unweighted),

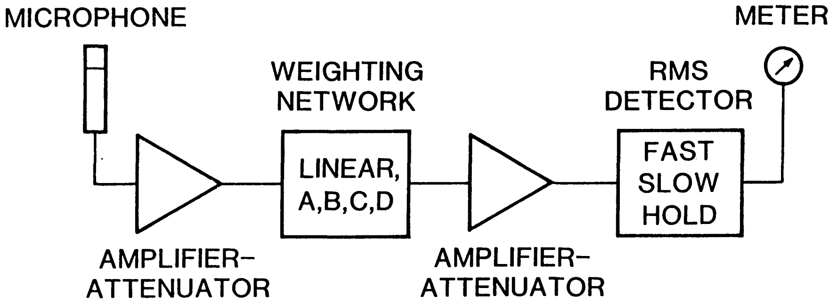

FIG. 12. Schematic showing the basic functions of a sound-level meter.

A, B, C, and sometimes D. For stationary or quasistationary signals, a "fast" or "slow" time constant is used. The fast response follows time-varying sound pressures more closely at the expense of accuracy; the slow response offers a higher confidence level for the rms sound pressure measurement. Impulsive signals present something of a problem. Current standards specify a time constant of 35 ms, in an attempt to simulate the response of the human ear, plus the capability of storing the peak or rms value of the applied signal. To prevent saturation resulting from high peak amplitudes, the detector circuit must be capable of sustaining a "crest factor," i.e., ratio of peak to rms signal, of at least 5.

9. MAGNETIC TAPE RECORDERS

9.1 Characteristics

The *magnetic tape recorder* is a widely used instrument for storing a permanent record of an acoustic time history (Finkel, 1975). Figure 13, curve a, shows a typical response of a magnetic tape recorder. The output voltage is plotted against signal frequency f, which is equal to the tape speed v divided by the record wavelength λ. For a harmonic wave, the record wavelength is the distance along the tape needed to record one complete cycle. At low frequencies, the output voltage is proportional to the time rate of change of magnetic flux passing across the head gap and is thus proportional to frequency. At high frequencies, as the wavelength approaches the size of the gap l the response falls toward zero. The minimum recordable wavelength is reciprocal to the packing density, a property of the tape. A typical packing density of 10^4 cycles/in. gives $\lambda_{\min} = 10^{-4}$ in., a common gap size. Then at a tape speed of $v = 120$ in./s (ips), $f_{\max} = 120/10^{-4} = 1.2$ MHz. Standard tape widths are $\frac{1}{2}$ in., containing 7 tracks (recording channels), and 1 in., containing 14 tracks. Standard tape speeds vary from $1\frac{7}{8}$ to 120 ips in the "intermediate" band and from $3\frac{3}{4}$ to 120 ips in the "wide" band. The two operational modes of analog tape recording are direct and FM.

9.1.1 Direct Record/Reproduce In the *direct* mode, the conditioned signal is applied directly to the head winding. Electronic equalization compensates for the poor frequency response, resulting in the improved response curve b of Fig. 13. A typical dynamic range is 32 dB, extending from the noise floor (mostly generated in the tape) to

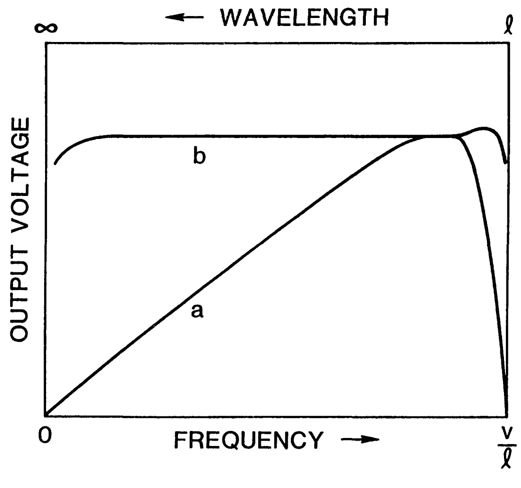

FIG. 13. Frequency response of a magnetic tape recorder (a) without and (b) with equalization.

the 1% amplitude-distortion level. The frequency response depends on tape speed and band selection. The performance specifications of an instrumentation recorder may seem inferior to those of an audio recorder, because equalization in the latter is shaped to conform to the requirements of speech and music, thus eliminating much of the noise.

9.1.2 FM Record/Reproduce In the *FM* mode, the signal is frequency modulated before being applied to the head winding. Two advantages are a low-frequency response to dc and a 10- to 15-dB improvement in dynamic range (at the expense of bandwidth). The bandwidth is limited by the maximum excursion from the carrier frequency, namely ±40%, sufficient to include most of the sidebands. Most of the noise is generated in the discriminator. In the recording of infrasonic signals, variations in tape speed, known as "flutter" ($f > 10$ Hz) and "wow" ($f < 10$ Hz), contribute to the recorded frequency spectrum.

9.2 Digital Tape Recording

A *digital tape recorder* records and reproduces data in digital format. An immediate advantage, from the acoustic point of view, is a substantial improvement in dynamic range, which is determined by the number of bits in the A/D converter. Most digital recording is IBM compatible, using $\frac{1}{2}$-inch tape with seven or nine tracks in an NRZ-1 format (nonreturn to zero-change of flux for one recording). The seven-track tape has a recording density of 200, 556, or 800 bits/in. (bpi) and a character size of six parallel bits plus one bit for a parity check; the nine-track tape has 800 or 1600 bpi with an eight-bit (7 + 1) character size. The data rate is the bit density times the tape speed—for example, 200 bpi × 120 ips = 24 000 bits/s. This is set on a clock generator and must be compatible with the rate at which the interfaced computer can transfer data. Data are organized in records, conforming to the buffer capacity of the computer, and the records are separated by gaps. A 2400-ft reel of magnetic tape with a packing density of 800 bpi can store 2400 × 12 in. × 800 bpi = 23 megabits per track (less the gap space), representing less than 10% of the analog data that can be stored in the same space.

9.3 PC Adaptables

The magnetic tape recorder is traditionally the weak link in acoustical measurement systems. Aside from the severe limitation on dynamic range and the inconvenience of long retrieval times, tape reliability is often suspect as a result of degradation from repeated use and copying, oxide separation, and "bleeding." The magnetic tape heads require frequent, often daily, maintenance. Digital tape remedies most of these ills, but is used in conjunction with very expensive equipment. The explosive technological development of *personal computers* (PCs) and associated components such as A/D converters and hard drives offers an attractive alternative.

Let us compare, for example, the storage capacities of a standard $10\frac{1}{2}$-inch reel containing 3600 ft of $\frac{1}{2}$-inch, seven-channel magnetic tape, and a hard drive having 1.0 Gbyte of memory. Let us apply a harmonic signal at 20 kHz. If the magnetic tape is driven at 30 ips, the recording time is 24 min, and the number of signal cycles stored is $24 \times (60 \text{ s}) \times (20 \times 10^4 \text{ cycles/s}) \times (7 \text{ channels}) = 2 \times 10^8$ cycles. For storage on the hard drive, if we sample the signal at 2.5 times the Nyquist frequency, then the number of cycles stored is $(1.0 \times 10^9 \text{ bytes})/[(2 \text{ bytes/word}) \times (2.5 \text{ words/cycle})] = 2 \times 10^8$ cycles—the identical result. The latter choice offers reliable storage, the convenience of immediate data processing, and practically maintenance-free operation. The data can later be transferred to a backup cassette tape.

10. COMPUTERS IN ACOUSTICS

The integration of a digital computer into an acoustic measurement system offers many practical advantages in addition to improved specifications regarding dynamic range, data storage density, flexibility, and cost effectiveness. Many acoustic measurements require inordinately complex evaluation procedures. The capability of performing an on-line evaluation during a test provides the user with an immediate readout of the evaluated data; this may aid in the making of decisions regarding further data acquisition and the choice of test parameters. The decision-making procedure can even be automated. The digital data can readily be telecommuni-

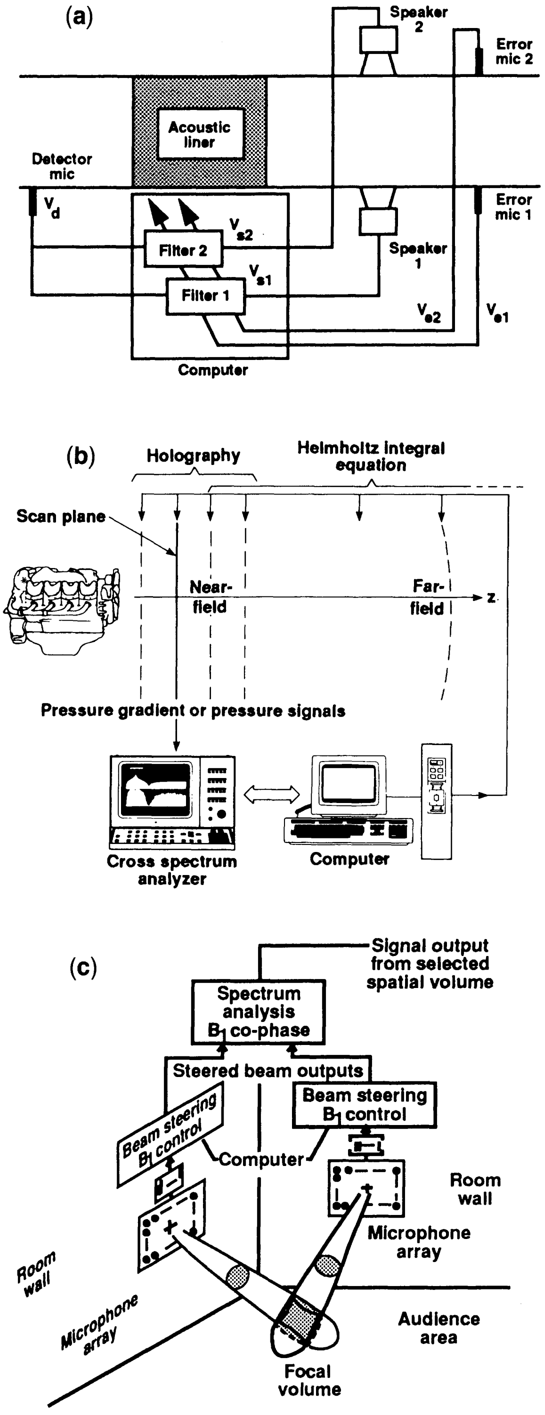
(a)
Speaker 2
Error mic 2
Acoustic liner
Detector mic
V_d
V_s2
Filter 2
V_s1
Filter 1
Error mic 1
Speaker 1
V_e2
V_e1
Computer
(b)
Holography
Helmholtz integral equation
Scan plane
Near-field
Far-field
z
Pressure gradient or pressure signals
Cross spectrum analyzer
Computer
(c)
Signal output from selected spatial volume
Spectrum analysis B1 co-phase
Steered beam outputs
Beam steering B1 control
Beam steering B1 control
Computer
Room wall
Microphone array
Room wall
Microphone array
Audience area
Focal volume

FIG. 14. Measurement systems using a digital computer. (a) Active noise cancellation; (b) spatial transformation of sound fields; and (c) computer-steered microphone arrays. (Courtesy of NASA, B & K Instruments, and *Journal of the Acoustical Society of America*.)

cated over ordinary telephone lines. Most digital systems accommodate a great variety of peripheral equipment.

Figure 14 shows examples of a computer integrated into an acoustical measurement system.

1. *Active noise cancellation* (Fig. 14a). The computer implements real-time digital filters 1 and 2, which serve as adaptive controllers to produce the required responses of noise-cancelling speakers 1 and 2 (Nelson and Elliott, 1992).
2. *Spatial transformation of sound fields* (Fig. 14b). A cross-spectrum analyzer yields a cross-spectral representation of a sound field, based on acoustical measurements over a selected scan plane; then a computer predicts the near field from the scan data using near-field acoustic holography and the far field from the Helmholtz integral equation (Ginn and Hald, 1989).
3. *Computer-steered microphone arrays* (Fig. 14c). In a large room, like an auditorium or conference hall, the computer introduces a preprogrammed time delay into each microphone of a rectangular array, thus steering the array to the direction of high selectivity; coordinating more than one array, it controls the location from which the received sound is especially sensitive (Flanagan *et al.*, 1985).

ACKNOWLEDGMENT

The author gratefully acknowledges permission to reproduce material from "Acoustical Measurement," *Encyclopedia of Physical Science and Technology*, Vol. 1, 2nd ed. (Academic Press, 1992).

GLOSSARY

Anechoic: Having no reflections or echoes.

Background Noise: The rms output of an acoustic device in the absence of acoustical excitation.

Coupler: Small leak-tight enclosure into which acoustic devices are inserted for the purpose of calibration, measurement, or testing.

Dynamic Range: The range of amplitudes over which a device operates linearly, limited by background noise at low amplitudes and harmonic distortion at high amplitudes.

Diffuse Field: Region of uniform acoustic energy density.

Free Field: Region where sound propagation is unaffected by boundaries.

Harmonic: Pertaining to a pure tone, that is, a sinusoidal wave at a single frequency; also, an integral multiple of a fundamental tone.

Infrasound: Sound at frequencies below the limit of human hearing, nominally 20 Hz.

Reverberant: Highly reflecting, as usually associated with a diffuse sound field.

Works Cited

ASTM (1992), *Annual Book of ASTM Standards*, Vol. 04.06, Philadelphia: ASTM, pp. 100–110.

ANSI (1979), *ANSI Standard S1.36-1979 (R1985)*, New York: American National Standards Institute.

Bruel & Kjaer (1982), *Condenser Microphone Data Book*, Denmark: Naerum Offset.

Finkel, J. (1975), *Computer Aided Experimentation*, New York: Wiley.

Flanagan, J. L., Johnston, J. D., Zahn, R., Elko, G. W. (1985), *J. Acoust. Soc. Am.* **78**, 1508–1518.

Gade, S. (1982), *Bruel & Kjaer Tech. Rev.* **3-1982**, 3–25.

Ginn, K. B., Hald, J. (1989), *Bruel & Kjaer Tech. Rev.* **2-1989**, 1–27.

Hansen, K. S. (1969), *Bruel & Kjaer Tech. Rev.* **1-1969**, 19–25.

Jones, H. W., Stradulinsky, D. C. (1977), *J. Acoust. Soc. Am.* **61**, 1089–1091.

Kim, S. C., Wise, K. D. (1983), *IEEE Trans. Electron Devices* **ED-30**, 802–810.

Moldover, M. R., Mehl, J. B., Greenspan, M. (1986), *J. Acoust. Soc. Am.* **79**, 253–272.

Nelson, P. A., Elliott, S. J. (1992), *Active Control of Sound*, New York: Academic Press.

Randall, R. B., Upton, R. (1978), *Bruel & Kjaer Tech. Rev.* **1-1978**, 3–25.

Roesler, H. (1966), *Acustica* **17**, 73–78.

Salava, T. (1974), *J. Audio Eng. Soc.* **22**, 146–153.

Sessler, G. M., West, J. E. (1969), *J. Acoust. Soc. Am.* **46**, 1081–1086.

Sessler, G. M., West, J. E. (1973), *J. Acoust. Soc. Am.* **53**, 1589–1600.

Singh, R., Schary, M. (1978), *J. Acoust. Soc. Am.* **64**, 995–1003.

Tarnow, V. (1987), *J. Acoust. Soc. Am.* **82**, 379–381.

Thane, N. (1979), *Bruel & Kjaer Tech. Rev.* **1-1979**, 3–25.

Wong, G. S. K., Embleton, T. F. W. (1979), *J. Acoust. Soc. Am.* **66**, 1275–1280.

Zuckerwar, A. J. (1983), *J. Acoust. Soc. Am.* **73**, 2180–2186.

Further Reading

Beranek, L. I. (1954), *Acoustics*, York, PA: McGraw-Hill.

Hassall, J. R., Zavari, K. (1979), *Acoustic Noise Measurements*, 4th ed., Denmark: Bruel & Kjaer Instruments.

Kundert, W. A. (1978), *Sound and Vibration* **12**, 10–23.

Olsen, H. F. (1973), *Acoustical Engineering*, New York: van Nostrand.

Petersen, A. P. G., Gross, E. E., Jr. (1972), *Handbook of Noise Measurements*, 7th ed., Concord, MA: General Radio Company.

MECHANICAL AND ELASTIC WAVES IN SOLIDS

KUNIHIKO IWASAKI, *Advanced Materials Research Laboratories, Nisshin Steel Co., Ltd., Chiba, Japan*

Introduction ... 515
1. **Elasticity in a Continuum Elastic Body** ... 516
1.1 Stress and Strain Tensors ... 516
1.2 Generalized Hooke's Law and Elastic Constants ... 516
2. **Propagation of Waves** ... 518
2.1 Mechanical or Elastic Waves in General ... 518
2.2 Acoustic Waves ... 520
2.3 Ultrasonic Waves ... 520
2.4 Surface Elastic Waves (Rayleigh Waves) ... 520
2.5 Seismic Waves ... 520
2.6 Magnetoelastic Waves ... 521
3. **Lattice Vibrations** ... 521
3.1 Crystal with Monatomic Basis ... 521
3.2 Distribution of Normal Modes ... 522
3.3 Crystal with Two Unlike Atoms per Primitive Cell ... 522
4. **Anelasticity** ... 523
5. **Applications to Nondestructive Evaluation of Defects in Solids** ... 524
5.1 Internal Friction Method ... 525
5.1.1 Peaks Resulting from Point Defects ... 525
5.1.2 Peaks Resulting from Dislocations ... 525
5.1.3 Peaks Resulting from Plane Defects ... 526
5.2 Ultrasonic Method ... 526
5.3 Photoacoustic Method ... 527
5.4 Acoustic Emission Method ... 528
Glossary ... 528
Works Cited ... 528
Further Reading ... 529

INTRODUCTION

In the theory of elasticity, a body is assumed to be in equilibrium under the action of applied forces. This is to be contrasted to the case of rigid dynamics, where a force applied to any one point on a body produces a linear acceleration of the whole body and an angular acceleration about its center of gravity. To describe elasticity, it is necessary to define stress and strain in a continuum body and to introduce elastic constants relating them.

In an extended isotropic body, there usually propagate two types of mechanical or elastic waves, longitudinal (dilatational) and transverse (distortional) waves. In addition to these, elastic waves of another type, Rayleigh waves, propagate along the surface of a solid. They spread two-dimensionally along the surface and travel a long distance, decaying very rapidly, however, in the depth direction. They play an important role in seismic phenomena.

In crystalline solids, the periodicity of atomic arrangements must be taken into account. This becomes important when the wavelength is comparable with the lattice constants. The periodic nature of a crystal causes behavior for elastic waves similar to that of Bragg reflection of x rays.

As real solids are not perfectly elastic, deviation from Hooke's law plays a remarkable role in some cases. At small strain levels, for example, anelastic phenomena such as internal friction and elastic after-effect are observed. At high strain levels, the strain is not elastic any more but plastic.

If defects exist in solids, they act as sources of disturbance to elastic waves. Then it is possible to detect them nondestructively by measuring the changes in wave characteristics. This principle has been used practically in a variety of ways.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. ELASTICITY IN A CONTINUUM ELASTIC BODY

In this section, stress and strain tensors and several elastic constants relating them are introduced for a continuum body (Timoshenko and Goodier, 1970; Terasawa, 1984).

1.1 Stress and Strain Tensors

Consider a small tetrahedron of volume V enclosed by x, y, and z planes and a diagonal plane perpendicular to a unit vector **n** with components (n_x,n_y,n_z) as shown in Fig. 1. A stress acting in the x direction across a face dS_x is designated as $\mathbf{T}_x$, similar notations being used for the others. By considering the equilibrium of forces acting on the tetrahedron,

$$\mathbf{T}_n dS_n = \mathbf{T}_x dS_x + \mathbf{T}_y dS_y + \mathbf{T}_z dS_z \qquad (1)$$

is obtained, where the effect of gravity is neglected. Minus signs on $\mathbf{T}_x$, $\mathbf{T}_y$, and $\mathbf{T}_z$ appear in Fig. 1 because the surface normal is assumed to be positive in the outward direction, which is opposite to those of the x, y, and z axes. Using the relations $dS_x = n_x dS_n$, $dS_y = n_y dS_n$, and $dS_z = n_z dS_n$, we get the following equation from Eq. (1):

$$\mathbf{T}_n = n_x \mathbf{T}_x + n_y \mathbf{T}_y + n_z \mathbf{T}_z. \qquad (2)$$

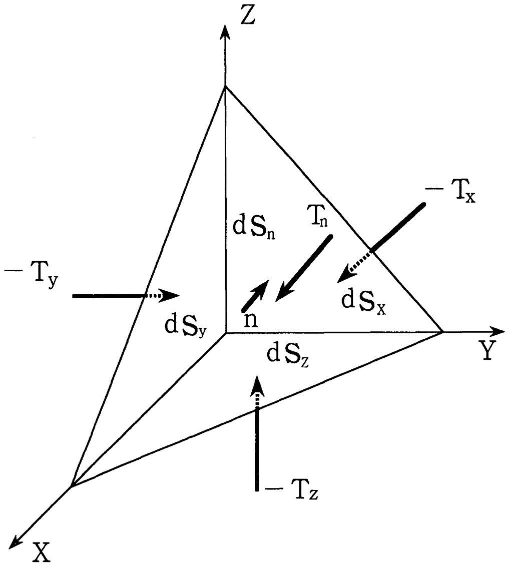

FIG. 1. Forces acting on the surfaces of a tetrahedron.

Equation (2) is rewritten in matrix form as

$$\begin{bmatrix} T_{nx} \\ T_{ny} \\ T_{nz} \end{bmatrix} = \begin{bmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{bmatrix} \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix}, \qquad (3)$$

where T_{xy}, for example, indicates a force acting on the x plane in the y direction. The first bracketed array on the right-hand side is called the *stress tensor*. T_{ij} corresponds to tension or compression in case $i = j$ and shear in case $i \neq j$. As $T_{ij} = T_{ji}$ must hold in order to keep the body nonrotating (consider the balance of moments), the stress tensor is symmetrical and has only six independent components.

Suppose a point $P(x,y,z)$ in a body is displaced to $P'(x + u, y + v, z + w)$, where u, v, and w are the components of a displacement vector. The *tensile strain* in the x direction e_{xx} and the *shear strain* in the xy plane e_{xy}, for example, are given by

$$e_{xx} = \frac{\partial u}{\partial x}, \qquad (4a)$$

$$e_{xy} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right), \qquad (4b)$$

respectively. Then the strain at an arbitrary point is given by the following *strain tensor* with six independent components:

$$\begin{bmatrix} \dfrac{\partial u}{\partial x} & \dfrac{1}{2}\left(\dfrac{\partial v}{\partial x} + \dfrac{\partial u}{\partial y}\right) & \dfrac{1}{2}\left(\dfrac{\partial w}{\partial x} + \dfrac{\partial u}{\partial z}\right) \\ \dfrac{1}{2}\left(\dfrac{\partial u}{\partial y} + \dfrac{\partial v}{\partial x}\right) & \dfrac{\partial v}{\partial y} & \dfrac{1}{2}\left(\dfrac{\partial w}{\partial y} + \dfrac{\partial v}{\partial z}\right) \\ \dfrac{1}{2}\left(\dfrac{\partial u}{\partial z} + \dfrac{\partial w}{\partial x}\right) & \dfrac{1}{2}\left(\dfrac{\partial v}{\partial z} + \dfrac{\partial w}{\partial y}\right) & \dfrac{\partial w}{\partial z} \end{bmatrix}, \qquad (5)$$

where diagonal and off-diagonal components represent tensile and shear strains, respectively.

1.2 Generalized Hooke's Law and Elastic Constants

In an elastic body, stress and strain at any place and time are linearly related with each other. If the strain is zero, the stress is also zero. Almost all materials follow this rule below a certain critical stress or strain level called the *elastic limit*. In general, the relation between stress T and strain e is described by a generalized Hooke's law with elastic constants C_{ij},

$$T_i = \sum_{j=1}^{6} C_{ij} e_j, \qquad i = 1, 2, \ldots, 6. \tag{6}$$

To make the matter simpler, stress and strain are described by a one-suffix system in Eq. (6) instead of the two-suffix system given before. The relations between the two systems are as follows:

$$\begin{aligned} &T_1 = T_{11}, && T_2 = T_{22}, && T_3 = T_{33}, \\ &T_4 = T_{23}, && T_5 = T_{13}, && T_6 = T_{12}, \\ &e_1 = e_{11}, && e_2 = e_{22}, && e_3 = e_{33}, \\ &e_4 = 2e_{23}, && e_5 = 2e_{13} && e_6 = 2e_{12}. \end{aligned} \tag{7}$$

The number of elastic constants, 36, is reduced to 21 by using the relation $C_{ij} = C_{ji}$, which comes from the law of conservation of energy. In case of an isotropic body, the number can be further reduced to only two. A variety of ways of expressing two-constant stress-strain relations are used in the literature. With the use of Lamé's constants λ and μ, for example, strain is divided into dilatational and shear components [see also Eq. (18) below], and the stress-strain relation is given by

$$T_{ij} = \lambda \epsilon \delta_{ij} + 2\mu e_{ij}, \tag{8}$$

where ϵ is dilatation given by $e_{xx} + e_{yy} + e_{zz}$ and δ_{ij} is Kronecker's delta (Condon, 1958). By using Eq. (8), Eq. (6) is rewritten in the following practical forms;

$$T_{xx} = (2\mu + \lambda)\frac{\partial u}{\partial x} + \lambda \frac{\partial v}{\partial y} + \lambda \frac{\partial w}{\partial z}, \tag{9a}$$

$$T_{yy} = \lambda \frac{\partial u}{\partial x} + (2\mu + \lambda)\frac{\partial v}{\partial y} + \lambda \frac{\partial w}{\partial z}, \tag{9b}$$

$$T_{zz} = \lambda \frac{\partial u}{\partial x} + \lambda \frac{\partial v}{\partial y} + (2\mu + \lambda)\frac{\partial w}{\partial z}, \tag{9c}$$

$$T_{xy} = T_{yx} = \mu\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right), \tag{9d}$$

$$T_{yz} = T_{zy} = \mu\left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y}\right), \tag{9e}$$

$$T_{zx} = T_{xz} = \mu\left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z}\right). \tag{9f}$$

Now let us introduce practically used fundamental elastic constants by considering three simple cases shown in Fig. 2. In case of elongation of a rod as shown in Fig. 2(a), all the stresses except T_{xx} are zero. Then we obtain $\partial v/\partial y = \partial w/\partial z$ from Eqs. (9b) and (9c). *Young's modulus* is defined as the ratio of tensile stress T_{xx} to tensile strain e_{xx},

$$E = T_{xx}/e_{xx}. \tag{10}$$

Poisson's ratio σ is defined as the ratio of the contraction in the y or z direction to the elongation in the x direction and is given by

$$\sigma = -\frac{\partial v/\partial y}{\partial u/\partial x} = -\frac{\partial w/\partial z}{\partial u/\partial x} = \frac{\lambda}{2(\lambda + \mu)}. \tag{11}$$

It takes a value between 0 and $\frac{1}{2}$.

With the use of Young's modulus E and Poisson's ratio σ, *Lamé's constants* are given by

$$\begin{aligned} \lambda &= E\sigma/(1 + \sigma)(1 - 2\sigma), \\ \mu &= E/2(1 + \sigma). \end{aligned} \tag{12}$$

When a cube of volume V and edge length a is compressed by Δa in each edge direction under isostatic pressure p [Fig. 2(b)], we get

$$T_{xx} = T_{yy} = T_{zz} = -p, \tag{13}$$

$$\Delta V/V \cong 3\Delta a/a, \tag{14}$$

where Δa is assumed to be very small. From Eqs. (9), (13), and (14) we obtain the following two equations:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} = \frac{\partial w}{\partial z} = -\frac{\Delta a}{a} = -\frac{\Delta V}{3V}, \tag{15}$$

$$p = (\lambda + 2\mu/3)\Delta V/V = K\Delta V/V, \tag{16}$$

where K is called the *bulk modulus* and is related to E and σ by

$$K = E/3(1 - 2\sigma). \tag{17}$$

In case of shear deformation shown in Fig. 2(c), the following relation is obtained from Eq. (9d):

$$\tau = \mu\theta = G\theta, \tag{18}$$

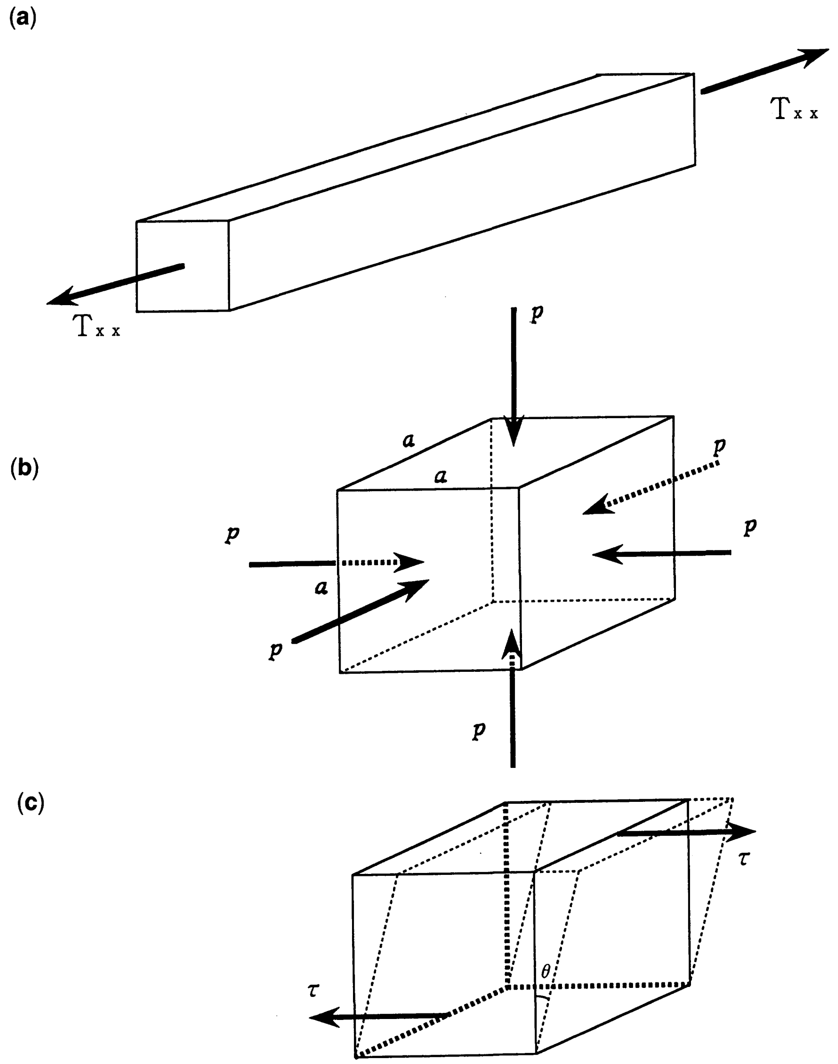

FIG. 2. Examples of simple deformation modes: (a) elongation, (b) isostatic compression, and (c) shearing.

where τ and θ are shear stress and deformation angle, respectively. The proportionality constant G, which is equal to Lamé's constant μ, is called the *rigidity* or *shear modulus*.

2. PROPAGATION OF WAVES

The general behavior of propagation of elastic waves is given first and several typical examples of elastic waves follow (Crawford, 1968; Graff, 1973).

2.1 Mechanical or Elastic Waves in General

When a small element of a continuum elastic body starts to vibrate, the neighboring elements follow its motion. By this process, the vibration extends through the body to form mechanical or elastic waves. In this text, we draw no distinction between the terms mechanical and elastic waves.

Let us first consider a simple case where a long, thin rod is set into vibration by applying a force along its length (x direction), as shown in Fig. 3. As only a tensile stress T_{xx} is assumed to act on a small volume V across a cross-sectional area of S, the equation of motion for V is given by

$$\rho S \Delta x \frac{\partial^2 u}{\partial t^2} = [T_{xx}(x + \Delta x, t) - T_{xx}(x,t)]S. \tag{19}$$

In the limit where Δx approaches zero, we obtain

$$\rho \frac{\partial^2 u}{\partial t^2} = \frac{\partial T_{xx}}{\partial x}. \tag{20}$$

By putting Eq. (9a) into Eq. (20) in consideration of the definition of Young's modulus [Eq. (10)], we get a *wave equation,*

$$\rho \frac{\partial^2 u}{\partial t^2} = E \frac{\partial^2 u}{\partial x^2}, \tag{21}$$

which shows that the vibration in each small volume of the rod propagates in the x direction as a longitudinal elastic wave with a velocity $v = (E/\rho)^{1/2}$.

The above-mentioned one-dimensional example is easily extended to three-dimensional general cases. By using Eq. (9), we obtain the following equations for an isotropic elastic body:

$$\rho \frac{\partial^2 u}{\partial t^2} = \mu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right) + (\lambda + \mu)\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right), \tag{22a}$$

$$\rho \frac{\partial^2 y}{\partial t^2} = \mu\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2}\right) + (\lambda + \mu)\frac{\partial}{\partial y}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right), \tag{22b}$$

$$\rho \frac{\partial^2 w}{\partial t^2} = \mu\left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 w}{\partial z^2}\right) + (\lambda + \mu)\frac{\partial}{\partial z}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right). \tag{22c}$$

Equation (22) is a generalized form of Eq. (21), and the behavior of elastic waves is determined by this equation.

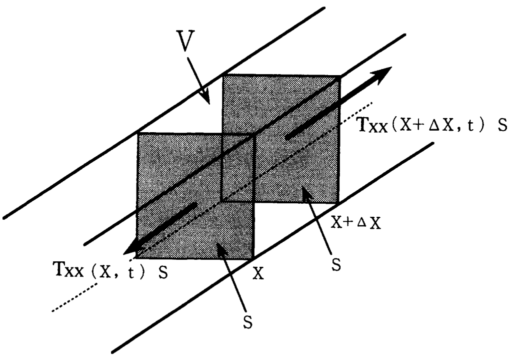

FIG. 3. Small volume element in a long, thin rod as a source of elastic-wave production.

As a simple example, we consider a case where the displacements depend only on x and t to reduce Eq. (22) to

$$\rho \frac{\partial^2 u}{\partial t^2} = (2\mu + \lambda)\frac{\partial^2 u}{\partial x^2}, \tag{23a}$$

$$\rho \frac{\partial^2 y}{\partial t^2} = \mu \frac{\partial^2 v}{\partial x^2}, \tag{23b}$$

$$\rho \frac{\partial^2 w}{\partial t^2} = \mu \frac{\partial^2 w}{\partial x^2}. \tag{23c}$$

General solutions of Eq. (23) can be written as

$$u(x,y,z,t) = f_1(x - v_l t) + g_1(x + v_l t), \tag{24a}$$

$$v(x,y,z,t) = f_2(x - v_t t) + g_2(x + v_t t), \tag{24b}$$

$$w(x,y,z,t) = f_3(x - v_t t) + g_3(x + v_t t). \tag{24c}$$

There exist three independent plane waves traveling in the x direction. One is a longitudinal wave with a velocity

$$v_l = [(2\mu + \lambda)/\rho]^{1/2} = [E(1 - \sigma)/(1 + \sigma)(1 - 2\sigma)\rho]^{1/2}, \tag{25}$$

and the other two are transverse waves with velocity

$$v_t = (\mu/\rho)^{1/2} = [E/2(1 + \sigma)\rho]^{1/2} = v_l[(1 - 2\sigma)/2(1 - \sigma)]^{1/2}. \tag{26}$$

The former is caused by T_{xx} (tension or compression) and the latter two by T_{yx} (shear).

Suppose one of f in Eq. (24) takes a local maximum or a minimum $f(a)$ at a. The plane S given by the equation $x - vt = a$ forms a plane perpendicular to the x axis (plane wave). In general, a plane like this that consists of points of equal phase is called *wave front*. The wave front of f propagates in the positive direction with velocity v. As for g, it behaves in a way exactly similar to f, but the propagation direction is opposite to that of f.

The examples given above are very simple cases. Actual waves behave in a very complicated manner. This is because a real body is anisotropic and contains interfaces at which longitudinal and transverse waves are mutually interchangeable.

2.2 Acoustic Waves

In a narrow sense, an acoustic wave means an audible *sound wave* that travels in the air. In a wider sense, however, it includes all the elastic waves regardless of audibility and traveling media. In the case of solids, it is therefore a part of elastic waves, so no special description is given here (for further details see ACOUSTICS, LINEAR).

2.3 Ultrasonic Waves

Sound waves of frequencies higher than 20 kHz are usually called ultrasonic waves. They are not special, in the sense that they obey the general rules of acoustics. Because of their high frequency and the possibility of getting large amplitude, however, there appear interesting phenomena that cannot be observed in usual sound waves. One of the typical characteristics is the possibility of attaining extraordinarily high acceleration. This, for example, causes cavitation and sonoluminescence and excites chemical, physical, and mechanical reactions. Apart from these nonlinear effects, high convergence and directionality are also typical of ultrasonic waves, which can be applied in various kinds of measurements. A few examples will be given in Sec. 5.2, but for further details see ULTRASONICS.

2.4 Surface Elastic Waves (Rayleigh Waves)

Surface elastic waves were theoretically predicted to propagate along the surface of a semi-infinitely extending elastic body (Rayleigh, 1945). They have relatively low frequency, and their amplitude decreases rapidly with the depth from the surface. However, they can spread two-dimensionally and travel a long distance along the surface.

The equation of motion is given by

$$\frac{\partial^2 u}{\partial t^2} - v^2 \Delta u = 0. \tag{27}$$

where u is an arbitrary component of either of the displacement vectors $\mathbf{u}_l$ (longitudinal) or $\mathbf{u}_t$ (transverse), v is the corresponding velocity, and Δ is the Laplacian. Let us assume the plane of propagation as the xy plane and measure the depth in the negative z direction. If the propagation direction is taken in the x direction, the solution of Eq. (27) takes the form of

$$u = \exp[i(kx - \omega t)]\, f(z), \tag{28}$$

where k is a wave vector and ω is an angular frequency. By putting this equation into Eq. (27), we obtain

$$\frac{d^2 f}{dz^2} = \left(k^2 - \frac{\omega^2}{v^2}\right) f. \tag{29}$$

Discarding meaningless solutions, we obtain the following as a possible solution for Rayleigh waves:

$$u = A \exp[i(kx - \omega t)] \times \exp[(k^2 - \omega^2/v^2)^{1/2} z], \tag{30}$$

where A is a constant. A practical displacement vector $\mathbf{u}$ is given by the sum of $\mathbf{u}_l$ and $\mathbf{u}_t$. Equation (30) shows that the waves propagate only along the surface, decaying in the negative z direction.

In general, the velocity depends on density, elastic constants, and Poisson's ratio. In the passage of Rayleigh waves, each element in the body makes an elliptical motion in the plane parallel to the propagation direction and perpendicular to the surface, where the ratio of the vertical amplitude to the horizontal amplitude depends on Poisson's ratio.

2.5 Seismic Waves

Seismic waves consist of elastic waves emitted spherically from a hypocenter and the secondary waves produced at the earth's surface and at inner boundaries of elastic discontinuity. Two kinds of elastic waves, *P* wave and *S* wave, propagate in the earth's crust. *P* and *S* designate primary and secondary, respectively, in the order of arrival of the seismic waves. The former is a longitudinal wave accompanied by volume changes, while the latter is a transverse wave accompanied by distortions.

Along the earth's surface, there also travel Rayleigh waves, Love waves, *G* waves, and so on. A Love wave is an *S*-wave-type surface wave and is known to appear when a semi-infinite body is covered with a surface layer of which the wave velocity is lower than that of the body. A *G* wave is a special kind of Love wave, which has a long wavelength and can travel a long distance without decaying. The vibrational periods of these waves decrease in the order *G* wave, Love wave, Rayleigh wave, *S* wave, and *P* wave, while the velocities decrease in the order *P* wave, *S* wave, *G* wave, Rayleigh wave, and Love wave (Allaby, 1988).

2.6 Magnetoelastic Waves

A ferromagnetic material consists of an aggregate of magnetic domains, each of which is magnetized to the saturation intensity in a particular direction. The magnetization of a material does not involve a change in the saturation magnetization but a change in the domain pattern so as to produce a net magnetic moment (Brown, 1958).

A magnetic domain usually has a strain in the direction of magnetization. Consequently, if the strength or the direction of an external magnetic field is changed, there appears a net strain as a result of redistribution of magnetic domains. This is called *magnetostriction* (see MAGNETOSTRICTIVE MATERIALS AND DEVICES).

Conversely, if an external stress is applied to a ferromagnetic material, the stress tends to orient the magnetization in such directions as to aid the deformation in favor of the stress. This magnetostrictive deformation is superposed on the ordinary elastic deformation to make the observed elastic modulus smaller than it would be otherwise. A large magnetic field, on the other hand, prevents the reorientation of magnetization to raise the elastic modulus.

Elastic waves under the influence of magnetoelastic coupling are called *magnetoelastic waves*. It is possible to excite elastic waves in ferromagnetic materials by applying an external magnetic field. On the other hand, elastic waves excited by any means in ferromagnetic materials inevitably suffer effects of magnetoelastic coupling during their propagation, which usually leads to a decrease in vibrational amplitude.

3. LATTICE VIBRATIONS

The description given above for a continuum body is applied to crystalline solids here (Kittel, 1986). We treat the short-wavelength range where the wavelength is comparable to the lattice constants. The periodic nature of a crystal causes for elastic waves behavior similar to that of Bragg reflection of x rays. Elastic waves are generally quantized as phonons (for details see PHONONS IN CRYSTAL LATTICES) and play a very important role in solid state physics.

3.1 Crystal with Monatomic Basis

As a simple example, we first consider a crystal with one atom in the primitive cell. We describe the displacement of a crystallographic plane [imagine, for example, low-index planes such as (100), (110), and (111) in a cubic lattice] from its equilibrium position with a single coordinate u_n, where all the atoms on the plane are assumed to move together. This assumption reduces the problem to one dimension.

The equation of motion for the *n*th plane is given by

$$\frac{Md^2u_n}{dt^2} = C(u_{n+1} - u_n) - C(u_n - u_{n-1}), \qquad (31)$$

where only nearest-neighbor interaction is taken into account. C is a force constant, which may take different values for longitudinal and for transverse waves. M corresponds to the mass of an atom if C is regarded as defined for one atom of the plane. We look for a traveling-wave solution of the form

$$u_n = A \exp[i(\omega t + kna)], \qquad (32)$$

where A, ω, k, and a are amplitude, angular frequency, wave vector, and spacing between neighboring planes, respectively. This function can be a solution of Eq. (31) if

$$\omega = \pm(4C/M)^{1/2} \sin(ka/2). \tag{33}$$

This relation between ω and k is called a *dispersion relation.* ω takes a maximum value $\omega_m = (4C/M)^{1/2}$ at $k = k_m = \pm\pi/a$.

It is to be noted that larger values of k merely reproduce motions already described by values of k within the limits $\pm k_m$. The region between $\pm k_m$ is called the first *Brillouin zone.* At the limiting frequency ω_m, Eq. (32) does not represent a traveling wave but a standing wave. This situation corresponds exactly to that for Bragg reflection of x rays.

3.2 Distribution of Normal Modes

Here we limit the number of planes to $N + 1$. Two edge planes located at $n = 0$ and $n = N$ are assumed to be held fixed. The normal modes, which are constructed by taking linear combinations of Eq. (32), are of the form

$$u_{k,n} = C_k \exp(i\omega_k t) \sin(kna). \tag{34}$$

C_k is an arbitrary complex number which determines the amplitude and phase of the motion. k is restricted to the following values by the boundary conditions:

$$k = \pi/L, 2\pi/L, 3\pi/L, \ldots, N\pi/L, \tag{35}$$

which means there are $N - 1$ allowed values (*eigenvalues*) of k. This number is equal to the number of planes allowed to move. Each allowed value of k is associated with a solution (*eigenfunction*) of the form of Eq. (34).

The number of states $C(\omega)d\omega$ in $d\omega$ at ω is given by

$$C(\omega)d\omega = \frac{L}{\pi}\frac{dk}{d\omega}\,d\omega, \tag{36}$$

where $dk/d\omega$ is given as

$$\frac{dk}{d\omega} = \frac{2}{a(\omega_m^2 - \omega^2)^{1/2}}. \tag{37}$$

It is to be noted that this equation has a singularity at $\pm\omega_m$.

3.3 Crystal with Two Unlike Atoms per Primitive Cell

We now consider a crystal with two kinds of planes, each of which is occupied by atoms of mass M or m. They are assumed to appear alternately with a spacing of a. Imagine, for example, a set of (100) planes in a CsCl-type crystal. The planes M are located at odd-numbered positions, $2n - 1, 2n + 1, \ldots$, while the planes m are at even-numbered positions, $2n, 2n + 2, \ldots$. On the assumption of nearest-neighbor interactions, the equations of motion are given by

$$\frac{Md^2u_{2n+1}}{dt^2} = C(u_{2n+2} + u_{2n} - 2u_{2n+1}), \tag{38a}$$

$$\frac{md^2u_{2n}}{dt^2} = C(u_{2n+1} + u_{2n-1} - 2u_{2n}). \tag{38b}$$

We look for solutions of the form

$$u_{2n+1} = A \exp\{i[\omega t + (2n + 1)ka]\}, \tag{39a}$$

$$u_{2n} = B \exp[i(\omega t + 2nka)]. \tag{39b}$$

These equations have nontrivial solutions only when

$$\omega^2 = C(1/M + 1/m) \pm C[(1/M + 1/m)^2 - 4 \sin^2(ka)/Mm]^{1/2}. \tag{40}$$

For small k, Eq. (40) becomes

$$\omega^2 = 2C\left(\frac{1}{M} + \frac{1}{m}\right) \quad \text{or} \quad \frac{2Ck^2a^2}{(M + m)}. \tag{41}$$

At $k = \pi/2a$, Eq. (41) takes the value of $2C/M$ or $2C/m$, respectively.

The dispersion relation given by Eq. (40) is shown in Fig. 4 for the case $M > m$. It is seen that it has two branches, one called the *acoustical branch* and the other the *optical branch.* In the former, the planes M and m move in phase with each other, while in the latter, they move in opposite directions.

There exists no wavelike solution for frequencies between $(2C/M)^{1/2}$ and $(2C/m)^{1/2}$; this produces a *forbidden band,* where the propagation constant k becomes complex; that is, the wave is damped.

In general, if there exist p atoms in the primitive cell, there appear $3p$ branches in

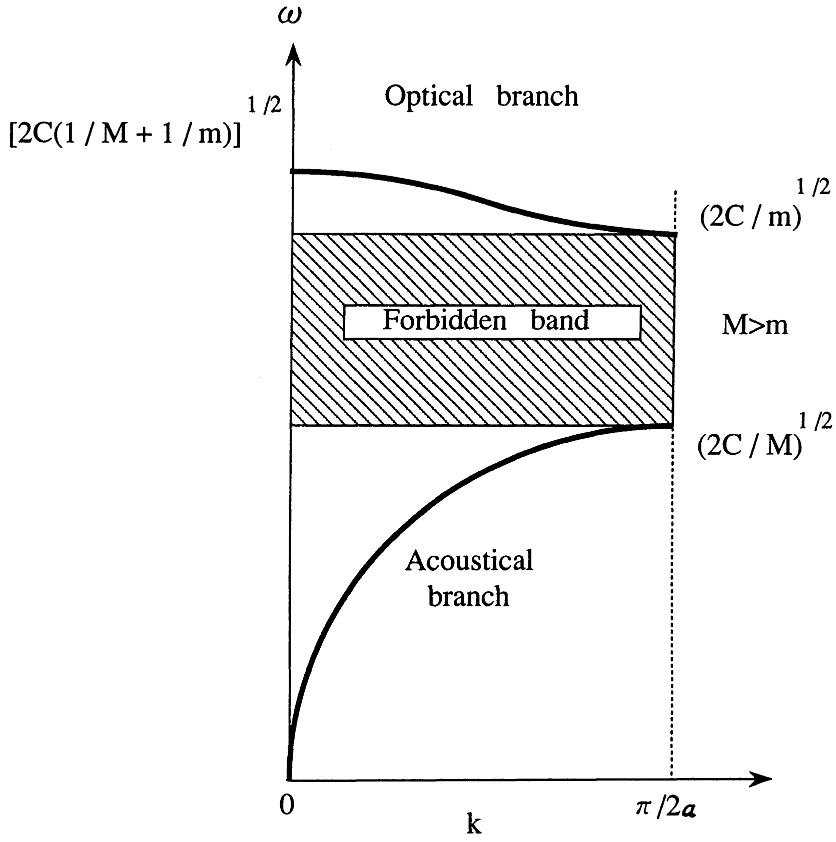

FIG. 4. Dispersion relation for a diatomic crystal.

the dispersion relation, 3 acoustical and $3p - 3$ optical.

4. ANELASTICITY

A formal theory of anelasticity begins as a departure from Hooke's law to include the time t as a variable in the stress-strain relation (Zener, 1952; Berry and Nowick, 1966; Nowick and Berry, 1972; de Batist, 1972). Here the linearity between stress T and strain e at each instant of time and the attainment of a unique equilibrium value of strain for a given stress after an infinite lapse of time are assumed. We neglect the tensorial character of stress and strain to restrict the discussion to one dimension.

One of the simplest generalizations of Hooke's law on these assumptions is found in the *standard linear solid*, of which the stress-strain relation is modified as

$$T + \tau_1 \frac{dT}{dt} = M\left(e + \tau_2 \frac{de}{dt}\right). \tag{42}$$

M is the elastic modulus, τ_1 the relaxation time for constant strain, and τ_2 the relaxation time for constant stress. In anelasticity, the notation M is generally used instead of E.

If, for example, both e and de/dt are zero, the solution of Eq. (42) takes the form

$$T(t) = T_0 \exp(-t/\tau_1), \tag{43}$$

where T_0 is the stress at $t = 0$. This shows that the stress relaxes to zero with the relaxation time τ_1.

Now suppose a constant stress T_0 is given at $t = 0$. In this case, the strain at time t is given by

$$e(t) = T_0/M + (e_0 - T_0/M) \exp(-t/\tau_2), \tag{44}$$

where e_0 is the strain at $t = 0$. After an infinite lapse of time, e reaches T_0/M. These kinds of relaxation phenomena are called *elastic after-effect*.

The *unrelaxed modulus* M_u is given by $T_0/e(0)$, which corresponds to instantaneous stress and strain, while the *relaxed modulus* M_r is given by $T_0/e(\infty)$, which corresponds to the state where a complete relaxation has occurred. They are mutually related by

$$M_u\tau_1 = M_r\tau_2. \tag{45}$$

If a cyclic stress, $T = T_0 \exp(i\omega t)$, is applied, the corresponding strain can be expressed as

$$e = e_0 \exp(i\omega t - \phi), \tag{46}$$

where ϕ is a *phase lag*. For such stress and strain, we get the following relation from Eq. (42):

$$(1 + i\omega\tau_1)T = M(1 + i\omega\tau_2)e. \tag{47}$$

Here we define a complex modulus M_c as

$$M_c = M(1 + i\omega\tau_2)/(1 + i\omega\tau_1). \tag{48}$$

The real part of M_c is called *dynamic modulus* M_d and is given by

$$M_d = M_u - (M_u - M_r)/[1 + (\omega\tau)^2], \tag{49}$$

where $\tau = (\tau_1\tau_2)^{1/2}$. It is to be noted that the resonance frequency is determined by M_d.

The tangent of the phase lag ϕ is given by the ratio of the imaginary part to the real part of M_c. ϕ is a measure of internal friction and is given by

$$\phi \cong \tan(\phi) = \frac{M_u - M_r}{(M_u M_r)^{1/2}} \frac{\omega\tau}{1 + (\omega\tau)^2}, \tag{50}$$

where ϕ is assumed to be very small.

With the use of Eqs. (49) and (50) and a new parameter, *relaxation strength* defined as $\Delta_M = (M_u - M_r)/(M_u M_r)^{1/2}$, the internal friction ϕ and the modulus defect $\delta M/M$ are given by

$$\frac{\delta M}{M} = \frac{M_u - M_d}{(M_u M_r)^{1/2}} = \frac{\Delta_M}{1 + (\omega\tau)^2}, \tag{51}$$

$$\phi = \Delta_M \omega\tau/[1 + (\omega\tau)^2]. \tag{52}$$

They are of the well-known Debye type, originally used to describe dielectric relaxation. As shown in Fig. 5, when plotted against $\log(\omega\tau)$, $\delta M/M$ changes stepwise being antisymmetric about the point $(0, \Delta_M/2)$, while 2ϕ gives rise to a symmetric peak centered at $\log(\omega\tau) = 0$. The peak width at half-maximum, $\delta \log(\omega\tau)$, is equal to 1.144.

In thermally activated processes, the relaxation time is generally described by

$$1/\tau = (1/\tau_0) \exp(-\Delta H/kT), \tag{53}$$

where $1/\tau_0$ is the frequency factor, ΔH the activation energy, k Boltzmann's constant, and T the absolute temperature. As is clear from Eq. (53), $\ln(\omega\tau)$ and $1/T$ are linearly connected with each other. This means that anelastic behavior such as shown in Fig. 5 is detected either by changing frequency at constant temperature or by changing temperature at constant frequency.

5. APPLICATIONS TO NONDESTRUCTIVE EVALUATION OF DEFECTS IN SOLIDS

If there exist irregularities, discontinuities, or defects in solids, the characteristics of elastic waves are modified there. It is then

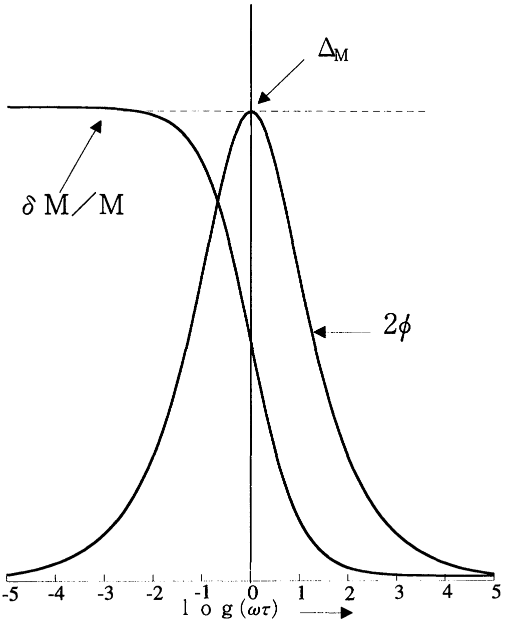

FIG. 5. Plots of modulus defect $\delta M/M$ and internal friction ϕ vs log ($\omega\tau$).

possible to make nondestructive evaluation (NDE) of solids by measuring such modifications. A few examples of NDE are shown below.

5.1 Internal Friction Method

By measuring internal friction and elastic modulus as functions of temperature or frequency, several kinds of lattice defects (irregularity of atomic arrangements) can be detected very sensitively. Internal friction usually reveals itself as a peak and elastic modulus as a step, as shown in Fig. 5. In principle, all the lattice defects that have lower symmetry than that of the lattice are detectable. Some typical examples of internal friction peaks are given below.

5.1.1 Peaks Resulting from Point Defects A *Snoek peak* (Snoek, 1939) is observed in body-centered cubic lattices containing a small amount of interstitial impurities such as C in Fe. Suppose an impurity atom takes one of the three possible interstitial positions shown with black dots in Fig. 6. If the x, y, or z position is occupied by an interstitial impurity atom, the lattice expands in the x, y, or z direction, respectively. When a tensile stress is applied in the z direction, for example, the z position is energetically favorable relative to the x and y positions, while the situation is reversed in case of compressive stress. Then if a cyclic stress is applied, impurity atoms jump between energetically favorable positions. This process is also regarded as a rotation of an elastic dipole.

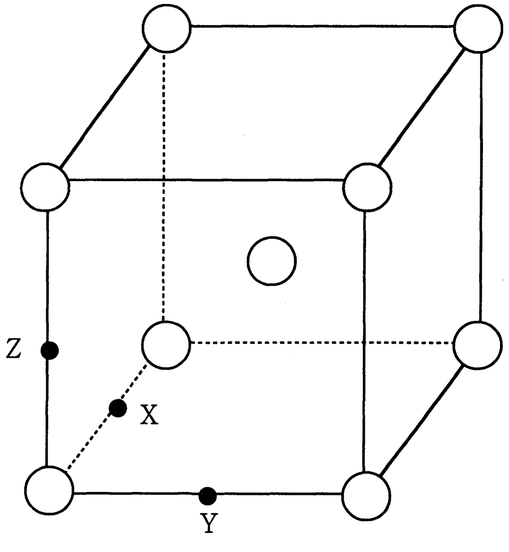

FIG. 6. Atomic model for Snoek internal-friction peak.

If the vibrational frequency of cyclic stress is very low or the temperature is very high, the elastic dipole rotates completely in phase with the stress without any time delay. This extreme case corresponds to a perfectly relaxed state and there appears no loop in the stress-strain diagram to produce no energy loss (the energy loss is proportional to the area of the loop). On the other hand, if the frequency is very high or the temperature is very low, the dipole cannot follow the stress at all (perfectly unrelaxed state). In this case, only perfectly elastic strain appears, and there is no energy loss, for the same reason as given above. Only in the intermediate frequency or temperature range does the dipole move behind the stress, figuring a loop in the stress-strain diagram. An internal friction peak and a step-like change in elastic modulus are then observed, if measured as a function of frequency or temperature.

This is a thermally activated process with the help of external stress. By measuring the Snoek peak, the concentration of interstitial impurities in solid solution is estimated separate from that in segregates.

When two substitutional impurity atoms form a pair, a similar elastic dipole is formed as given above if there is a difference in atomic sizes between the solute and the solvent atoms. An internal friction peak arising from this kind of atomic configuration is called a *Zener peak* (Zener, 1943). This, however, is not confined to pairs but applies to any atomic configurations of solute atoms that give rise to elastic anisotropy.

5.1.2 Peaks Resulting from Dislocations A dislocation lying in a Peierls valley (Peierls, 1940) jumps to an adjacent one by bowing out if an appropriate external stress is applied. This thermally activated motion gives rise to a *Bordoni peak* (Bordoni, 1949). Basic information on dislocation sliding is obtained by analyzing the characteristics of this peak.

Dislocations are very often pinned by point defects such as vacancies, interstitials, and impurities. Such dislocations drag the pinning points with them or break away from them, which gives rise to Hasiguti peaks

(Hasiguti *et al.*, 1962). The mechanism of interaction between dislocations and point defects is studied by measuring these peaks.

5.1.3 Peaks Resulting from Plane Defects Viscous motion of two adjacent grains along a grain boundary between them is known to give rise to a Kê peak (Kê, 1947) or grain-boundary peak. Since grain boundaries are very complicated lattice defects, detailed mechanisms of this peak are not known yet. Other boundaries, such as twin boundary and antiphase boundary, are also expected to reveal similar peaks.

5.2 Ultrasonic Method

This method is very widely used both in industrial and in scientific fields (Krautkramer, 1977). When an ultrasonic beam passes from a material 1 to a material 2 perpendicularly through an interface between them, the reflectivity R is given by

$$R = (Z_2 - Z_1)/(Z_2 + Z_1), \qquad (54)$$

where Z_i is the acoustic impedance of material i, given by the product of density and velocity. According to this principle, any defects that change the acoustic impedance can be detected by measuring the intensity of reflected waves. With the decrease in the wavelength of ultrasonic waves, the resolving power increases while the penetration depth decreases.

In order to detect small defects of μm order, a *scanning acoustic microscope* (SAM) (see MICROSCOPY ACOUSTIC) is used, where the ultrasonic beam is focused onto a small spot with an ultrasonic lens as shown in Fig. 7 (Quate *et al.*, 1979). There are two types of SAM, reflection type and transmission type.

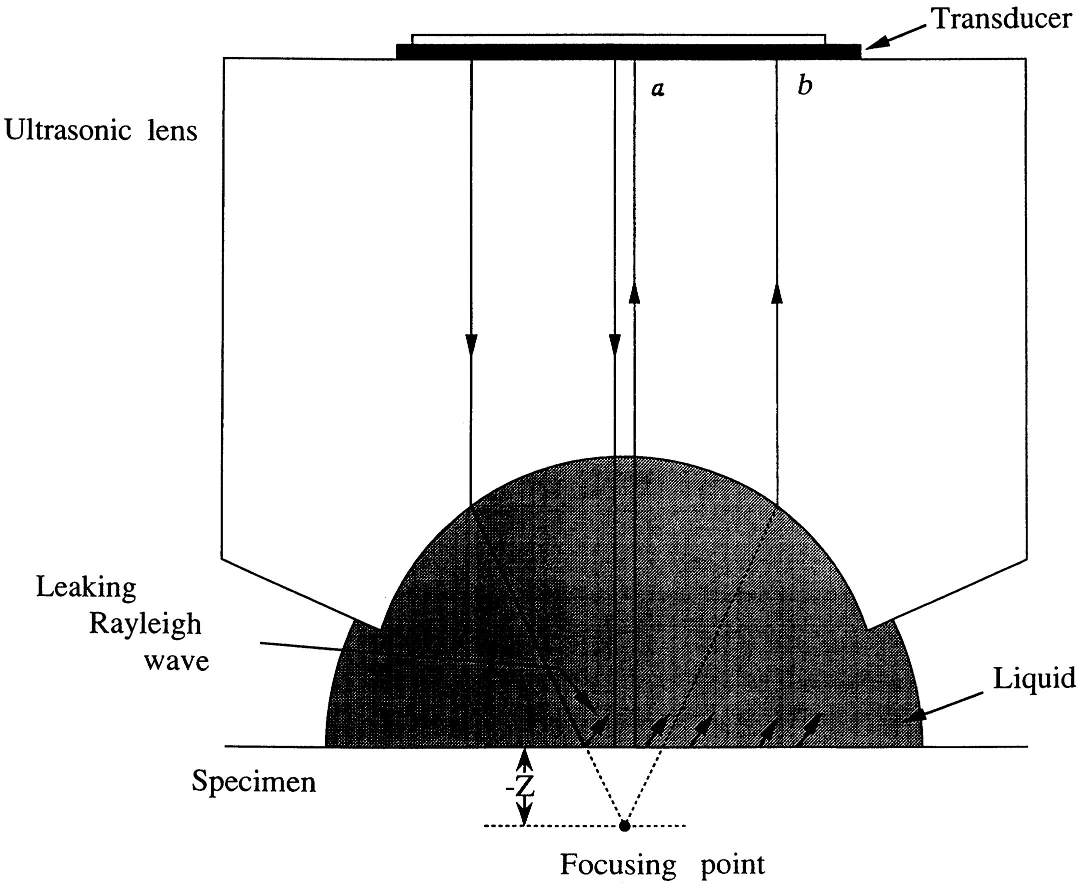

FIG. 7. Schematic illustration of an ultrasonic lens and a formation of a $V(z)$ curve.

The former is used more widely than the latter because the resolving power can be heightened by increasing frequency without any restriction on the specimen thickness.

It is also possible to evaluate the elastic properties of small surface areas in the latter method. If the intensity of a reflected beam V is measured as a function of distance z between the lens and the specimen surface, a curve called the $V(z)$ curve is obtained. This curve has periodic local maxima and minima, which arise from the interference between a directly reflected wave (*a*) and a leaking Rayleigh wave (*b*) as shown in Fig. 7. The velocity of the Rayleigh wave is given by

$$v_0/[1 - (1 - v_0/2f\Delta z)^2]^{1/2}, \tag{55}$$

where v_0 is the velocity of elastic waves in an acoustic coupler (a liquid, usually water, put between the lens and the specimen), f the ultrasonic frequency, and Δz the period of the $V(z)$ curve. Then by measuring Δz, the velocity or elastic constants of a small surface region can be determined.

5.3 Photoacoustic Method

Photoacoustic or photothermal effects have been used for NDE rather recently (Rosencwaig, 1977). If a specimen surface is illuminated with an amplitude-modulated laser beam as shown in Fig. 8, a certain part of beam energy is absorbed in the specimen and converted into thermal energy to produce thermal waves of the same frequency as that of the beam modulation. The thermal waves then produce elastic (mechanical) waves in the specimen and acoustic (pressure) waves in the surrounding atmosphere. As is clear from the explanation of the conversion processes given above, the photoacoustic signals reflect optical, thermal, and elastic properties of the illuminated material at the same time. This method can therefore detect wider varieties of defects than ultrasonic methods, which reflect only elastic properties.

During one period of illumination, the thermal wave travels one *thermal diffusion length* δ given by

$$\delta = (\kappa/\pi\rho c f)^{1/2} = (a/\pi f)^{1/2}, \tag{56}$$

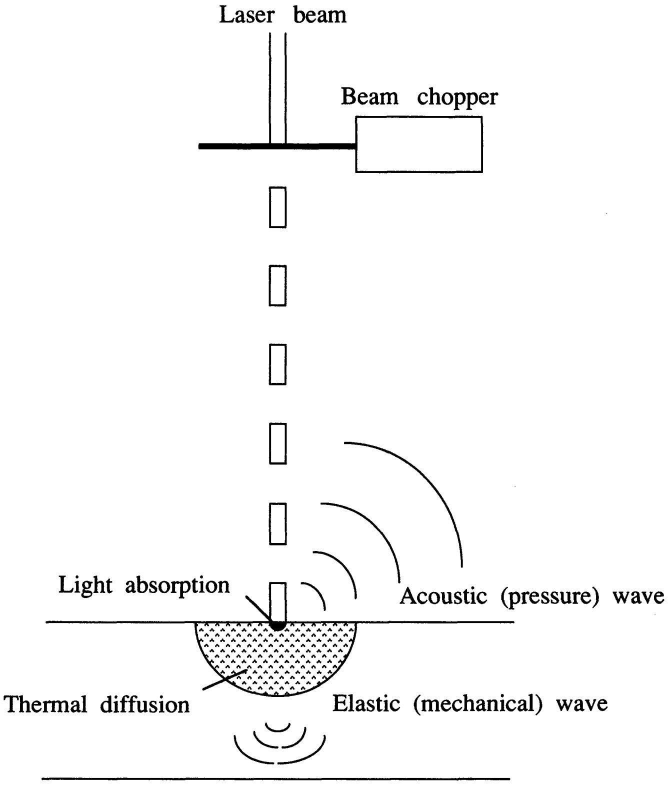

FIG. 8. Principle of the photoacoustic method.

where k, ρ, c, f, and a are thermal conductivity, density, specific heat, modulation frequency, and thermal diffusivity, respectively. Penetration depth is determined by δ and the resolving power by δ and the spot size of the laser beam.

If the illuminated position is raster-scanned in the specimen surface, an image map of defects in the subsurface region is obtained. This procedure is called *photoacoustic microscopy*. It should be noted that a three-dimensional image is also obtained by combining this scanning with a variation in modulation frequency which changes penetration depth, as given by Eq. (56).

5.4 Acoustic Emission Method

Elastic or stress waves are emitted as a result of a rapid release of strain energy caused by a structural alteration in solids. This phenomenon is called acoustic emission (AE). By detecting AE signals at three different points, the location of the signal source is determined. A remote identification of emission source mechanisms and associated structural alterations can also be made by a sophisticated analysis of the signals. There are varieties of emission sources ranging from atomic-size lattice defects to macroscopic-size earthquakes (Guild *et al.*, 1980). This method has been widely used in industrial fields to detect a small structural change that takes place before a material fractures. It should be noted here, however, that AE is not a nondestructive method in a strict sense because the signal comes from occurrence of precursory damage.

GLOSSARY

Acoustic Emission: Emission of elastic waves as a result of rapid strain energy release caused by structural alteration.

Break Away: Parting of a dislocation from pinning point defects.

Dispersion Relation: A relation between angular frequency and wave vector.

Elastic Limit: A limit of stress or strain below which Hooke's law holds.

Elastic After-Effect: An anelastic phenomenon in which stress and strain relax with time.

Internal Friction: Conversion of mechanical vibrational energy into thermal energy, which leads to a decrease in vibrational amplitude.

Lamé's Constants: Two elastic constants introduced by Lamé to describe isotropic elasticity.

Magnetostriction: Straining of ferromagnetic materials because of magnetization.

Photoacoustic Effect: Production of acoustic waves by illumination with amplitude-modulated light.

Standard Linear Solid: One of the simplest model solids for describing anelasticity.

Wave Front: A surface consisting of points of equal phase.

Works Cited

Allaby, A., Allaby, M. (Eds.) (1988), *The Concise Oxford Dictionary of Earth Sciences*, Glasgow: Oxford University Press.

de Batist, R. (1972), *Internal Friction of Structural Defects in Crystalline Solids*, Amsterdam: North-Holland.

Berry, B. S., Nowick, A. S. (1966), in: W. P. Mason (Ed.), *Physical Acoustics*, Vol. III-A, New York: Academic, p. 1.

Bordoni, P. G. (1949), *Ricerca Sci.* **4**, 851–862.

Brown, W. F. (1958), in: E. U. Condon, H. Odishaw (Eds.), *Handbook of Physics*, New York: McGraw-Hill, p. **4**-126.

Condon, E. U. (1958), in: E. U. Condon, H. Odishaw (Eds.), *Handbook of Physics*, New York: McGraw-Hill, p. **3**-3.

Crawford, F. S., Jr. (1968), *Waves*, Berkeley Physics Course, Vol. 3, New York: McGraw-Hill.

Graff, K. F. (1973), *Wave Motion in Elastic Solids*, New York: Dover.

Guild, F. J., Phillips, M. G., Harris, B. (1980), *Non-Destr. Test. Int.* **13**, 209–218.

Hasiguti, R. R., Igata, N., Kamoshita, G. (1962), *Acta Metall.* **10**, 442–447.

Kê, T. S. (1947), *Phys. Rev.* **71**, 533–546.

Kittel, C. (1986), *Introduction to Solid State Physics*, New York: Wiley.

Krautkramer, J., Krautkramer, H. (1977), *Ultrasonic Testing of Materials*, Berlin: Springer.

Nowick, A. S., Berry, B. S. (1972), *Anelastic Relaxations in Crystalline Solids*, New York: Academic.

Peierls, R. E. (1940), *Proc. Phys. Soc.* **52**, 34–37.

Quate, C. F., Atalar, A., Wickramasinghe, H. K. (1979), *Proc. IEEE* **8**, 1092–1114.

Lord Rayleigh (1945), *The Theory of Sound*, New York: Dover.

Rosencwaig, A. (1977), *Rev. Sci. Instrum.* **48**, 1133–1137.

Snoek, J. L. (1939), *Physica* **6**, 591–592.

Terasawa, T. (1984), *Vibrations and Waves*, Tokyo: Iwanami.

Timoshenko, S. P., Geedier, J. N. (1970), *Theory of Elasticity*, New York: McGraw-Hill.

Zener, C. (1943), *Trans. AIME* **152**, 122–126.

Zener, C. (1952), *Elasticity and Anelasticity of Metals*, Chicago: University of Chicago Press.

Further Reading

Bland, D. R. (1988), *Wave Theory and Applications*, Glasgow: Oxford University Press.

Bray, D. E., McBride, D. (Eds.) (1991), *Nondestructive Testing Techniques*, New York: Wiley.

Cochran, W. (1973), *Dynamics of Atoms in Crystals*, New York: Russak.

Kinsler, L. E., Frey, A. R. (1982), *Fundamentals of Acoustics*, New York: Wiley.

Kolsky, H. (1963), *Stress Waves in Solids*, New York: Dover.

Laing, D. (1991), *The Earth System: Introduction to Earth Science*, New York: Wm C. Brown.

Landau, L. D., Lifshitz, E. M. (1986), *Theory of Elasticity*, New York: Pergamon.

Love, A. E. H. (1944), *A Treatise on the Mathematical Theory of Elasticity*, New York: Dover.

Mandelis, A. (1991), *Principles and Perspectives of Photothermal and Photoacoustic Phenomena*, New York: Elsevier.

Parker, S. P. (Ed.) (1992), *McGraw-Hill Encyclopedia of Science and Technology*, New York: McGraw-Hill.

Steidel, R. F., Jr. (1989), *An Introduction to Mechanical Vibrations*, New York: Wiley.

Yamaguchi, K., Aoki, K., Kishi, T. (1986), *Progress in Acoustic Emission III*, Tokyo: The Japan Society for Non-destructive Inspection.

MECHANICAL INSTRUMENTATION

See TRANSDUCERS AND OTHER MECHANICAL INSTRUMENTATION

MECHANICAL PROPERTIES OF LIQUIDS

N. H. March, *Department of Theoretical Chemistry, University of Oxford, Oxford, England*

Introduction ... 531
1. Mechanical Models of Melting ... 532
1.1 Viscous Flow in Liquids ... 532
2. Macroscopic and Microscopic Measurements ... 532
2.1 Radiation Scattering ... 533
3. Mechanisms of Transport ... 533
3.1 Compressional (or Bulk) Viscosity: Mechanism and Measurement ... 534
4. Shear and Bulk Viscosities of Dense, Hard-Sphere Liquid ... 534
4.1 Shear Viscosity ... 534
4.2 Bulk Viscosity ... 535
5. Viscosities Correlated with Other Transport Properties ... 536
5.1 Connection between Shear Viscosity and Thermal Conductivity ... 536
5.2 Correlation between Diffusion and Shear Viscosity: Stokes's Law ... 536
5.3 Generalization of the Stokes–Einstein Relation to Embrace Bulk Viscosity ... 537
6. Dynamical Structure Factor and Time-Dependent Correlation Functions ... 538
6.1 Frequency Spectrum (or Spectral Function) $g(\omega)$ ... 538
6.2 Cusp in Frequency Spectrum at Zero Frequency Related to Shear Viscosity ... 539
6.3 Green–Kubo Formula for Viscosities: Application to Liquid Metals at Melting ... 539
6.4 Computer Simulation of Viscosities in Lennard-Jones fluids ... 540
7. Non-Newtonian Fluids ... 541
7.1 Viscoelasticity ... 541
7.1.1 Examples of Properties of Polyethylene Oxide ... 542
7.2 Some Theoretical Concepts for Viscoelastic Liquids ... 542
7.2.1 More Complex Types of Motion and Equations of State of Liquid ... 543
8. Interfacial Transport and Rheology ... 543
Glossary ... 543
Works Cited ... 544
Further Reading ... 544

INTRODUCTION

In this article, a focal point is the viscosities of liquids. The shear viscosity, defined as shear stress divided by rate of shear, is the most familiar quantity, but attention is also paid to bulk viscosity (see especially Secs. 3–5). For the moment, we merely note that shear viscosity η is about the effect of changing shape without changing volume, while in any discussion of the internal force acting on a volume element of liquid, there is another contribution involving the bulk viscosity η_B, which accounts for the effect of internal friction due to a change in volume. It proves of considerable interest to relate these two quantities, η and η_B, to dynamical aspects of liquid structure. These generalize the so-called static structure factor $S(q)$, which is directly related to the pair distribution function $g(\tau)$ by Fourier transform. These quantities, $S(q)$ and $g(\tau)$, measure quantitatively the short-range order in a liquid. The long-wavelength limit (q tends to zero) of $S(q)$ is related to the isothermal compressibility of the liquid, K_T, by $S(0) = \rho k_B T K_T$ where ρ is the number density of the liquid while k_B is Boltzmann's constant. Another compressibility (adiabatic) comes up in dealing with the velocity of sound c in liquids, the ratio of compressibilities being related to the ratio of specific heats by thermodynamics. As already indicated, a ba-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

sic connection of structure with viscosities results from a dynamical generalization of the static structure factor $S(q)$, carried out in Sec. 6. Some use is made of models, especially in establishing approximate relations between viscosity and other transport coefficients, e.g., shear viscosity to mass transport reflected in the self-diffusion coefficient D, and also to thermal conductivity (see Sec. 5.1). However, the way quantitative work can be done on, e.g., Lennard-Jones liquids by computer simulation is referred to also in Sec. 6.

1. MECHANICAL MODELS OF MELTING

Instead of expressing the equilibrium between a solid and its liquid in classical thermodynamic language, which inevitably involves both phases, various attempts have been made to characterize melting on the basis of a model in which the crystal becomes mechanically "unstable" above the melting temperature T_m. As an immediate example, such a "one-phase" model of melting is to assert that a liquid differs from a crystal in having zero resistance to shear. Thus, if the shear modulus of a crystal decreases with rising temperature, this could lead to a temperature dependence of the form suggested in early work of Sutherland, namely proportional to $(1 - T/T_m)^2$. In this model, the shear modulus evidently tends to zero at the melting temperature T_m. This, or indeed any similar approach to a theory of a "mechanical melting point" T_{mech}, suffers from the drawback of other one-phase theories of melting such as Lindemann's law. Such one-phase theories predict a limit to the stability of the solid regardless of the phase into which the solid transforms on fusion.

A further, and new experimental, objection to mechanical theories of melting is that, in those cases where the difficulties of measuring true elastic properties of crystals at high temperatures have, in fact, been overcome, the trend of the shear modulus with T does not lend support to a shear-instability limit to the crystalline phase, but suggests that T_m is usually less than T_{mech}.

1.1 Viscous Flow in Liquids

Other objections to mechanical theories of melting arise from basic theoretical considerations about actual viscous flow in liquids. These require, somewhat loosely speaking, a specification of relaxation times for each of the structurally possible processes in the liquid that can promote momentum transfer and hence lead to viscosity. For stresses of sufficiently short duration, less than the shortest relaxation time, liquids would in fact behave in the same way as crystals. As the time of application of shearing stress is prolonged, processes with sufficiently small relaxation periods can operate to relieve the stress and thus to permit flow of the liquid.

Non-Newtonian viscous flow (see also Sec. 7) in liquids can also complicate mechanical concepts about melting in other ways; for example, some fluids show much lower resistance to flow once they have been set in motion. Again, when a fluid–glass transition is observed and traversed, distinctive mechanical behavior is observed. Such transport properties can throw valuable light on mechanisms of fusion, particularly with certain classes of fluids such as network melts.

Table 1 shows the mechanical shear temperature T_{mech} and the thermodynamic melting temperature T_m for alkali halides. It can be seen that for these materials at ordinary pressures, T_{mech} always lies above T_m, but they parallel one another in the gross sense that they rise and fall together. However, in a little more detail, the ranking among cations sometimes differs; e.g., T_{mech} is lower for LiF than for RbF, but T_m is the other way around.

2. MACROSCOPIC AND MICROSCOPIC MEASUREMENTS

An important class of macroscopic measurements on liquids comprises measurements of transport coefficients, for which a variety of experimental methods is available. The shear viscosity, a major focus of the present article, can be determined from the observed damping of torsional oscillations or from capillary flow experiments. There is an intimate connection, demonstrated long ago by Stokes, between shear viscosity, involving transport of momentum, and the coefficient of self-diffusion D, corresponding to the transport of mass. Thus, it is relevant here to note that a direct method of determining the coefficient of self-diffusion involves the use of radioactive tracers (see DIFFUSION AND IONIC CONDUCTION IN LIQUIDS), which, in fact, places

Table 1. Mechanical [shear (T_{mech})] and thermodynamic (T_m) temperature for alkali halides in °C; after Ubbelohde (1978).

	F		Cl		Br		I	
Cation	T_{mech}	T_m	T_{mech}	T_m	T_{mech}	T_m	T_{mech}	T_m
Li	914	842	655	614	572	547	499	450
Na	1122	988	932	801	861	755	772	651
K	1022	846	931	776	890	730	822	686
Rb	959	775	890	715	886	682	838	642

it in the category of microscopic measurements. In favorable cases, the diffusion coefficient can also be obtained from nuclear magnetic resonance experiments. We return in Sec. 5.2 to the relation between shear viscosity and self-diffusion.

2.1 Radiation Scattering

From the standpoint of microscopic theories of liquids, of central importance are radiation-scattering experiments. One can add to the technique of inelastic scattering of thermal neutrons (see NEUTRON SCATTERING) the technique of x-ray diffraction (*q.v.*) and that of laser light scattering. It is of interest to note that the special feature of x-ray scattering experiments is the fact that the energy of the incident radiation is much greater than the thermal energies of the molecules, and the scattering process is essentially elastic. While this makes it impossible to study time-dependent processes by this particular technique, valuable information can be obtained concerning the static structure factor $S(q)$ of the liquid. A pioneering experiment of this kind that can be cited here is the determination of $S(q)$ for liquid Na and K metals near freezing by Greenfield *et al.* (1971). In contrast, for the case of thermal neutrons the energy of the incident particle is comparable with $k_{\mathrm{B}}T$, and the scattering cross section can therefore be measured as a function of energy transfer as well as momentum transfer. In this way, therefore, it is possible to gain information on wave-vector– and frequency-dependent fluctuations in liquids over a range of wavelengths comparable with the interatomic spacing.

Light-scattering techniques yield somewhat similar results to the neutron-scattering measurements referred to above, but the accessible range of momentum transfer limits the method to the study of fluctuations of wavelength of order 10^{-5} cm, bordering the hydrodynamic regime (see also Sec. 6). Such experiments are therefore particularly useful in obtaining information on critical phenomena.

3. MECHANISMS OF TRANSPORT

It is well established that in a liquid the transport of momentum (and energy), as opposed to the transport of matter, occurs by two independent mechanisms, namely the bodily movement of individual molecules through space and the action of intermolecular forces at a distance. Of these two processes, the former is the predominant mechanism at low densities, while the latter is the important mechanism at the high densities with which we are concerned in this article. The transport properties of monatomic fluids at low densities are well understood through the work of Chapman and Enskog (see, for example, Chapman and Cowling, 1939). The dense-liquid transport problem is less well understood at the time of writing. This is because in a liquid where the intermolecular potential varies continuously with distance, the collisional momentum (or energy) transport arises from a distortion of the radial distribution function $g(r)$ (Irving and Kirkwood, 1950), but at the time of writing no entirely satisfactory approach has been found to allow the calculation of this distortion.

Below, we exploit the simplicity of a model liquid of dense hard spheres to calculate the transport of momentum, reflecting the shear viscosity η, and the transport of energy, yielding the thermal conductivity λ. However, before turning to that, we need to introduce a further transport property, somewhat less well known, namely the bulk viscosity.

3.1 Compressional (or Bulk) Viscosity: Mechanism and Measurement

Consider first the viscous attenuation of a plane wave in a liquid of shear viscosity η. As the wave travels, there is relative motion of adjacent layers of the liquid, and this results in the creation of viscous forces that act against the pressure of the wave. Energy is taken out of the wave to overcome these viscous forces, a progressive extraction of energy that results in a corresponding progressive decrease in the intensity. When the product of angular frequency ω and relaxation time is very much less than unity, which is usually the case, then the acoustic absorption coefficient, α_1 say, due to shear viscosity can be shown to take the form (see, e.g., Temperley and Trevena, 1978)

$$\alpha_1 = 2\omega^2\eta/3dc^3, \tag{1}$$

where d is the mass density of the liquid while c is the velocity of sound.

We can now modify the above formula to take account of structural relaxation. This is done by introducing a compressional viscosity, η' say, through rewriting Eq. (1) as

$$\alpha_1 = 2\omega^2(\eta + \eta')/3dc^3. \tag{2}$$

Thus, in principle, one can extract the compressional viscosity η' from measurements of sound-wave attenuation, plus shear viscosity. Of course, in practice, the information that can be extracted depends on error bars on α_1 and η from experiment, and the amount that η' contributes in Eq. (2) relative to the shear viscosity η.

But it is a fact that the interpretation of such acoustic absorption measurements requires, beyond the density and velocity of sound already appearing in Eq. (2), accurate knowledge, transcending that equation, of heat capacity and also of other transport coefficients, and in particular thermal conductivity. It may be some time yet, therefore, before one can avoid considerable uncertainties in the final "experimental" values of compressional viscosity η'.

4. SHEAR AND BULK VISCOSITIES OF DENSE, HARD-SPHERE LIQUID

We have noted above that, strictly, to calculate the transport of momentum, and hence the shear viscosity η, one must be able to calculate the distortion of the pair function $g(r)$, and that this remains as a largely unsolved problem. However (see Collins and Raffel, 1954; Longuet-Higgins and Pople, 1956), in a liquid consisting solely of rigid molecules the singular nature of the intermolecular potential permits a finite flux of momentum (or energy) even when the radial distribution function is momentarily isotropic, as it is in equilibrium. It should not, however, be assumed that in a hard-sphere liquid under shear the pair distribution function really is isotropic—in fact, the velocity gradient will instantly distort it—but nevertheless by assuming an isotropic distribution function one can make a better than order-of-magnitude estimate of the transport coefficients (Longuet-Higgins and Pople, 1956).

The earliest study of the transport properties of a dense, hard-sphere liquid again goes back to Enskog (see Chapman and Cowling, 1939). Indeed, Enskog developed a modified form of the Boltzmann equation for the single-particle distribution function to apply at high densities. Subsequently, Collins and Raffel (1954) used the free-volume theory of liquids to obtain an expression for the shear viscosity of a hard-sphere liquid that is closely related to the one derived below, following the approach of Longuet-Higgins and Pople (1956). These latter workers demonstrated that it is possible to calculate the collisional contributions to both bulk and shear viscosities in a direct manner by making just two assumptions:

1. that the spatial pair distribution function depends only on the temperature and density and not on the rate of strain (or temperature gradient); and
2. that the velocity distribution function of a single particle is Maxwellian with a mean equal to the local hydrodynamic velocity, and a spread determined by the local temperature.

4.1 Shear Viscosity

Let us consider a situation in which the fluid is subjected to a steady shear represented by

$$\frac{\partial v_x}{\partial y} = -\frac{p'}{M}, \tag{3}$$

all other components of the rate-of-strain tensor being zero. If M denotes the mass of a sphere, p' represents a momentum rather than a velocity gradient. One can now proceed to evaluate the total x momentum transferred through unit distance in the y direction in a very short unit of time. This momentum transfer will be due almost exclusively to collisions of pairs of spheres that are already close together at the beginning of the short period of time. Let us denote such a pair of spheres by 1,2, and let the direction cosines of the unit vector $\mathbf{l}$ along the line of centers $\mathbf{r}_{12}$ be l_x, l_y, and l_z. If the components of the momenta of the two spheres in the direction $\mathbf{r}_{12}$ are p_1, p_2, then their relative momentum of approach is

$$p = p_1 - p_2. \tag{4}$$

Now if $p > 0$, the spheres will soon collide, and when this happens an amount of momentum $p\mathbf{l}$ will be transferred from one to the other. The x component of this momentum is pl_x, and the y component of the vector $\mathbf{r}_{12}$ is $2al_y$, where a is the radius of a sphere. Moreover, the rate of approach of the two spheres is p/M. Therefore, the total flux of x momentum in the y direction arising from collisions is

$$P^c_{xy} = \oint \int_0^\infty h(\omega,p) \frac{2ap^2 l_x l_y}{n} \, d\omega dp. \tag{5}$$

Here $h(\omega,p)d\omega$ represents the number of pairs per unit volume within unit small distance of contact having relative momentum p and such that the vector $\mathbf{l}$ lies in the solid angle $d\omega$. The first integration is over all possible orientations, while the second is over the positive range of p.

To this point, the argument is exact. But to proceed further it is, of course, necessary to insert a form for h. To obtain such a result, Longuet-Higgins and Pople assume a form

$$h(\omega,p) = (g_0/4\pi)\phi(\omega,p), \tag{6}$$

where g_0 is the equilibrium number of pairs per unit volume within unit (small) distance of contact and $\phi(\omega,p)$ is the probability that spheres 1,2 have relative momentum p in the direction $\mathbf{r}_{12}$. This form (6) has subsumed in it the assumption already referred to that the spatial pair distribution function is independent of the rate of strain. Next, it can be assumed without loss of generality that the fluid velocity is zero at the point of collision of the two spheres. Then the average momentum of the first sphere is $ap'l_y$ in the x direction, and that of the second sphere is $-ap'l_y$. The mean values of p_1,p_2 are therefore $ap'l_xl_y, ap'l_xl_y$, respectively, and one can write down their distributions ϕ_1 and ϕ_2 readily, since the momentum distribution is assumed to be locally Maxwellian. The distribution of p needed in Eq. (6) is therefore

$$\phi(\omega,p) = (4\pi M k_B T)^{-1/2} \exp\left(-\frac{(p - 2ap'l_xl_y)^2}{4Mk_BT}\right). \tag{7}$$

This can be used in Eq. (5) to calculate the momentum flux. In evaluating the integral, it is a valid procedure to expand the integrand to only the first power of p', since velocity gradients are assumed small. The result is (Longuet-Higgins and Pople, 1956)

$$P^c_{xy} = \frac{8a^2p'}{15} g_0 \left(\frac{k_BT}{\pi M}\right)^{1/2}, \tag{8}$$

where g_0 is the value of the radial distribution function $g(r)$ at contact. This can be expressed precisely in terms of the pressure P and the number density N/V to yield for $P^c_{xy}/(p'/M)$, which is simply the shear viscosity η,

$$\eta = \frac{4a}{5}\left(\frac{Mk_BT}{\pi}\right)^{1/2}\left(\frac{P}{k_BT} - \frac{N}{V}\right). \tag{9}$$

Before considering this result for the shear velocity further, let us outline the way that the corresponding result for the bulk viscosity can be calculated.

4.2 Bulk Viscosity

To calculate the bulk viscosity, η_B say, one imagines the fluid to be contracting uniformly at a rate determined by the velocity gradients

$$\frac{\partial v_x}{\partial x} = \frac{\partial v_y}{\partial y} = \frac{\partial v_z}{\partial z} = -\frac{p_b'}{M}, \tag{10}$$

say. Considering the fluid velocity to be zero at the point of collision between spheres 1 and 2, the distributions of p_1 and p_2 are now found to be

$$\phi_1(p_1) = (2\pi M k_B T)^{-1/2} \times \exp\left(-\frac{(p_1 - ap_b')^2}{2Mk_BT}\right), \tag{11}$$

while $\phi_2(p_2)$ is given by simply replacing the quantity $p_1 - ap_b'$ in Eq. (11) by $p_2 + ap_b'$. Now the flux of z momentum in the z direction is obtained by substituting z for x and y in Eq. (7). Using the same form as before for the function $h(\omega, p)$, the quantity corresponding to Eq. (8), namely P_{zz}^c, leads to a sum of two terms, one of which is independent of the rate of strain and represents the collisional contribution to the pressure, while the other, which is proportional to the rate of strain, must be interpreted as giving rise to a bulk viscosity. The coefficient of bulk viscosity η_B is, in fact, the pressure decrease divided by the divergence of the rate of strain, which in the present case is just $-3p_b'/M$. The bulk viscosity that results is

$$\eta_B = \frac{4a}{3}\left(\frac{Mk_BT}{\pi}\right)^{1/2}\left(\frac{P}{k_BT} - \frac{N}{V}\right). \tag{12}$$

This is seen by comparison with Eq. (9) for the shear viscosity of a hard-sphere liquid to show that the bulk viscosity η_B is $\frac{5}{3}$ of the shear viscosity η.

5. VISCOSITIES CORRELATED WITH OTHER TRANSPORT PROPERTIES

In this section, we are concerned with correlations that connect viscosities with other transport coefficients, the most celebrated of these being the so-called Stokes–Einstein relation (see Hansen and McDonald, 1976). A useful starting point is again afforded by the hard-sphere results for dense liquids, set out in the previous section.

5.1 Connection between Shear Viscosity and Thermal Conductivity

By quite similar arguments to those set out in Sec. 4 for shear and bulk viscosities, Longuet-Higgins and Pople also calculated the thermal conductivity λ for a dense, hard-sphere liquid, with the result

$$\lambda = 2ak_B\left(\frac{k_BT}{\pi M}\right)^{1/2}\left(\frac{P}{k_BT} - \frac{N}{V}\right). \tag{13}$$

Combining this with the result [Eq. (9)] for the shear viscosity η, one obtains

$$\lambda/\eta = 5k_B/2M, \tag{14}$$

where M is the atomic mass.

As an immediate test of the correlation (14), Fig. 1 shows a plot of experimental data for argon. There is evidently a substantial variation of the ratio η/λ with temperature at constant pressure, the constant result [Eq. (14)] being also shown in the figure for comparison. We return in Sec. 6 to a treatment of transport in a Lennard-Jones fluidlike argon, though numerical computation will then be needed. Ascough *et al.* (1988) have made a somewhat wider study of the prediction that η/λ is constant, and the interested reader is referred to their work.

5.2 Correlation between Diffusion and Shear Viscosity: Stokes's Law

We referred in Sec. 3 to common elements in the mechanisms of transport of momentum and of energy; these are reflected, of course, for hard spheres in the correlation (14). While the mechanism is different for mass transport, Longuet-Higgins and Pople (1956) have also calculated D for a dense, hard-

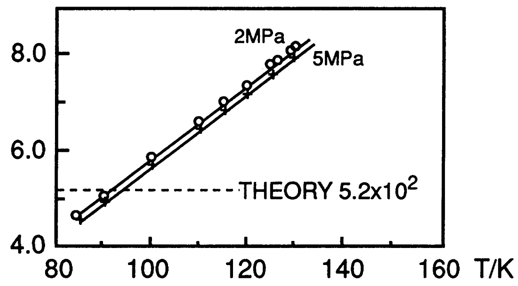

FIG. 1. Ratio of thermal conductivity λ to shear viscosity η for liquid argon. The prediction of Eq. (14) for a hard-sphere liquid is also shown. Evidently, variation of λ/η by about a factor of 2 actually occurs over the temperature range plotted.

sphere liquid as

$$D = \frac{a}{2}\left(\frac{\pi k_B T}{M}\right)^{1/2}\left(\frac{PV}{RT} - 1\right)^{-1}. \tag{15}$$

It is then immediately seen from Eqs. (9) and (15) that

$$D\eta = (2a^2/5)(Nk_B T/V). \tag{16}$$

This equation for the product $D\eta$ has the form of the Stokes–Einstein relation, namely $D\eta \sim k_B T/l$ where l is a characteristic length. Evidently $l \sim (V/N)/a^2$ in the hard-sphere result [Eq. (16)].

5.3 Generalization of the Stokes–Einstein Relation to Embrace Bulk Viscosity

In this section, we consider the generalization proposed by Zwanzig (1983) of the Stokes–Einstein relation to embrace bulk viscosity. It is based, in essence, on a "cage" model of an atom in a liquid. In such a model, one imagines an atom on which attention is focused being confined within a limited volume (cage) for an interval of time during which it performs vibrations within the cage. Eventually, by stochastic processes, the cage "opens up" and the atom "hops." Let us follow Zwanzig in subsuming these ideas into an explicit, albeit somewhat oversimplified, model. We shall then see that a generalization of Eq. (16) follows.

Let us consider, then, the motion of the atom performing vibrations about the center of the cage. The time dependence of a single normal-mode contribution has the form $\cos\omega t$, until a cell jump interrupts this motion. Normal modes in different subvolumes are interrupted at different times. Zwanzig accounts for this by introducing a factor of the form $\exp(-t/\tau)$, the waiting-time distribution for cell jumps that destroy the coherence of the oscillations in any subvolume V^*. The result is to yield a form for the diffusion constant D given by

$$D = \frac{k_B T}{3MN}\sum_{\omega}\frac{\tau}{1 + \omega^2\tau^2}, \tag{17}$$

where the sum is over all the $3N$ normal-mode frequencies.

In the absence of much detailed information about the actual frequency distribution (see below), Zwanzig now utilizes a Debye spectrum, treating longitudinal and transverse modes separately, each with its own Debye cutoff q_0:

$$\omega_l(q) = v_l q \qquad \omega_t(q) = v_t q, \tag{18}$$

for $0 < q < q_0$, v_l and v_t being longitudinal and transverse velocities of sound, respectively. The cutoff q_0 is chosen such that there are N modes in each branch of the spectrum, so that

$$(4\pi/3)(q_0/2\pi)^3 = N/V. \tag{19}$$

When the sum over frequencies in a single branch of the spectrum is replaced by an integral over q, one arrives, with a minor additional approximation, at the Zwanzig expression

$$D = \frac{k_B T}{3\pi}\left(\frac{3N}{4\pi V}\right)^{1/3}\left(\frac{1}{dv_l^2\tau} + \frac{2}{dv_t^2\tau}\right). \tag{20}$$

One next notes in Eq. (20) that dv_l^2 is an elastic bulk modulus, whereas dv_t^2 is a corresponding shear modulus. As Zwanzig notes, one expects that $dv_l^2\tau$ and $dv_t^2\tau$ are actually the longitudinal and shear viscosities η_l and η, respectively. Then the self-diffusion coefficient becomes, from Eq. (20),

$$D = \left(\frac{k_B T}{3\pi}\right)\left(\frac{3N}{4\pi V}\right)^{1/3}\left(\frac{1}{\eta_l} + \frac{2}{\eta}\right). \tag{21}$$

Equation (21) can now be rewritten in a form similar to that of the Stokes–Einstein relation (16), by defining the volume per atom $\Omega = V/N$. The quantity $\Omega^{1/3}$ is like a "length per particle" and takes over the role of the characteristic length l below Eq. (16). Then one can write, following Zwanzig,

$$\left(\frac{D\eta}{k_B T}\right)\Omega^{1/3} = 0.0658\left(2 + \frac{\eta}{\eta_l}\right) = C', \tag{22}$$

where η_l is the longitudinal viscosity, related to $\frac{4}{3}\eta + \eta_B$. Although the actual value of C', as defined by this equation, clearly depends on the ratio of the shear to the longitudinal viscosity, which as noted above is not known quantitatively in many liquids at the time of writing, Zwanzig (1983) notes that C' as defined can only vary between 0.13 and 0.18. Treating C' actually as a constant, Eq. (22) has precisely the form obtained by Brown and March (1968), discussed in Sec. 6.3 below, for the situation at the melting point of metals: see also Andrade (1934).

Zwanzig also cites examples of molecular liquids that support the form of Eq. (22), but we shall not go into further details here. Rather, we turn next to establishing an intimate link between shear and bulk viscosities on the one hand and neutron scattering on the other. This will be accomplished via the so-called dynamical structure factor $S(q,\omega)$, which we introduce immediately below.

6. DYNAMICAL STRUCTURE FACTOR AND TIME-DEPENDENT CORRELATION FUNCTIONS

Short-range order, so essential in distinguishing the dense liquid state from both the gaseous and the solid phases of matter, is described quantitatively by the pair function $g(r)$ of the liquid. This is accessible to measurement via the liquid structure factor $S(q)$ referred to in Sec. 2, which in turn is related to the Fourier transform of $g(r) - 1$. Let us now effect a frequency generalization of this static structure factor $S(q)$ such that, for classical liquids to which we confine ourselves throughout this account (March and Tosi, 1991),

$$S(q) = \int_{-\infty}^{\infty} S(q,\omega)d\omega. \tag{23}$$

Following Van Hove (1954), we note that $S(q,\omega)$ is now accessible to inelastic neutron scattering (Sec. 2). Essentially, $S(q,\omega)$ is a measure of the probability that a neutron incident on the liquid will transfer momentum $\hbar q$ and energy $\hbar\omega$ to the liquid.

The main object below is to show that, in the hydrodynamic (long-wavelength; long-time or low-frequency) limit, the dynamical structure factor $S(q,\omega)$ is related quantitatively to a combination of shear and bulk viscosities. To motivate the (initially surprising) form of such a relation, it will prove convenient to return to the self-diffusion coefficient D, connected intimately to viscosity through Eq. (21) or (22).

Then, as Van Hove especially emphasized, as well as $S(q,\omega)$ in a classical liquid, one can define a quantity $S_s(q,\omega)$ for the motion of a single (s) atom. In turn, this is observable via incoherent neutron scattering (Egelstaff, 1967). The double Fourier transform of this function ($\mathbf{q} \rightarrow \mathbf{r}$; $\omega \rightarrow t$), namely $G_s(\mathbf{r},t)$, represents the meanderings of a "labeled" atom in time. If we have the "tagged" atom at the origin $\mathbf{r} = 0$ at $t = 0$, then in the hydrodynamic limit referred to above, $G_s(\mathbf{r},t)$ is characterized by the diffusion equation

$$\nabla^2 G_s(\mathbf{r},t) = \frac{1}{D}\frac{\partial G_s(\mathbf{r},t)}{\partial t}. \tag{24}$$

With the initial condition specified above, namely $G_s(\mathbf{r},t = 0) = \delta(\mathbf{r})$, one readily finds the solution of Eq. (24) to be

$$G_s(\mathbf{r},t) = \frac{1}{(4\pi Dt)^{3/2}}\exp\left(-\frac{r^2}{4Dt}\right). \tag{25}$$

Fourier transformation to get $S_s(q,\omega)$ yields the result

$$S_s(q,\omega) = Dq^2/\pi[\omega^2 + (Dq^2)^2]. \tag{26}$$

6.1 Frequency Spectrum (or Spectral Function) $g(\omega)$

It is customary to define the frequency spectrum of the liquid by

$$g(\omega) = \omega^2 \lim_{q\rightarrow 0}[S_s(q,\omega)/q^2]. \tag{27}$$

While $g(\omega)$ is of considerable interest in its own right, the aim presently is to relate D to $S_s(q,\omega)$ in the hydrodynamic limit. From Eqs. (26) and (27), it is readily seen that one can obtain the self-diffusion coefficient D by either of the following procedures:

$$D = \pi \lim_{\omega\rightarrow 0} \omega^2 \lim_{q\rightarrow 0}\left(\frac{S_s(q,\omega)}{q^2}\right), \tag{28}$$

which is a so-called Green–Kubo transport formula, or

$$g(\omega = 0) = D/\pi. \tag{29}$$

The form of $g(\omega)$ in dense liquids is shown schematically in Fig. 2.

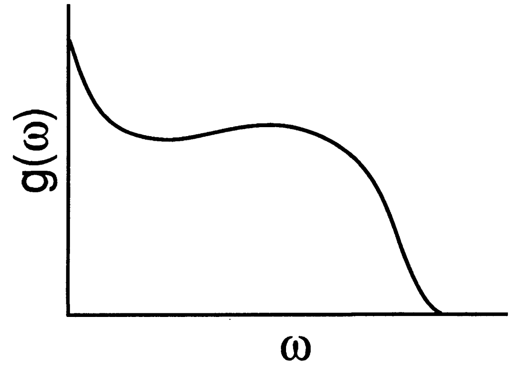

FIG. 2. Schematic form of frequency spectrum (or spectral function) $g(\omega)$ as defined in Eq. (27). The cusp depicted is made quantitative in Eqs. (33) and (34). The value of $g(\omega)$ at $\omega = 0$ is D/π, with D the coefficient of self-diffusion.

6.2 Cusp in Frequency Spectrum at Zero Frequency Related to Shear Viscosity

It can be seen from Fig. 2 that $g(\omega)$ has a cusp at the origin, and the object of the following is to clarify its origin and to relate its quantitative form to the shear viscosity η.

An equivalent form of D to Eq. (28), going back essentially to Einstein, is obtained by considering the mean square displacement $\langle r^2\rangle$ of an atom at long time t. This can be calculated directly from Eq. (25) and leads to the result that $\langle r^2\rangle$ of an atom at long time t tends to $6Dt$. It is instructive to rewrite this result in terms of velocity v rather than displacement. One has evidently

$$\mathbf{r} = \int_0^t \mathbf{v}(s)ds, \tag{30}$$

and forming the average of r^2 from this equation, one obtains, after some manipulation (see, for example, March, 1990a).

$$D = \lim_{t\to\infty} \frac{r^2}{6t} = \tfrac{1}{3}\int_0^\infty \langle \mathbf{v}(0)\cdot\mathbf{v}(t)\rangle dt. \tag{31}$$

Here we have introduced the so-called velocity autocorrelation function (for more on time-dependent correlations, see Sec. 6.4).

From hydrodynamics, and in particular the Navier–Stokes equation (see, for instance, Edwards *et al.*, 1991), it can be shown (Ernst *et al.*, 1970) that the velocity autocorrelation function falls off at large t with a so-called long-time tail proportional to $t^{-3/2}$. In turn, the Fourier transform of the velocity autocorrelation, related to the frequency spectrum $g(\omega)$ by

$$g(\omega) = \frac{k_B T}{M\pi}\int_0^\infty \frac{\langle \mathbf{v}(0)\cdot\mathbf{v}(t)\rangle}{v(0)^2}\cos\omega t\, dt, \tag{32}$$

was shown by Gaskell and March (1970) to have the small-ω expansion

$$g(\omega) = D/\pi + a_1\omega^{1/2} + O(\omega), \tag{33}$$

where

$$a_1 = -(2\pi)^{1/2} \times \frac{2}{3\rho}\left[4\pi\left(D + \frac{\eta}{\rho M}\right)\right]^{-3/2}\frac{k_B T}{M\pi}, \tag{34}$$

with M the atomic mass and ρ the atomic number density N/V. This result for a_1, the coefficient of the term in $\omega^{1/2}$ in the expansion (33), makes quantitative the form of the cusp depicted in Fig. 2. It is clear that although $g(\omega)$ was introduced through the study of the meanderings of a single atom, the cusp not only contains the expected dependence on the coefficient of self-diffusion D but also carries knowledge of the shear viscosity η.

6.3 Green–Kubo Formula for Viscosities: Application to Liquid Metals at Melting

There is an analogous formula to Eq. (28) involving $S(q,\omega)$ rather than just the self-contribution $S_s(q,\omega)$. This takes the form (see, for example, Egelstaff, 1967)

$$\frac{4\eta}{3} + \eta_B = \frac{\pi M^2}{k_B T}\lim_{\omega\to 0}\omega^4 \lim_{q\to 0}\frac{S(q,\omega)}{q^4}. \tag{35}$$

By arguments involving the assumption of a Debye frequency, Brown and March (1968) obtained an approximate formula for liquid metals at the melting point. Specializing to the shear viscosity, this takes the form

$$\eta/T_m^{1/2}M^{1/2}\rho^{2/3} = \text{const.}, \tag{36}$$

where T_m is the melting temperature. Choosing the constant in accord with the early work of Andrade (1934), there is excellent agreement between Eq. (36) and experiment, as shown in Table 2. There is close connection here with the work of Zwanzig embodied in Eq. (22) when D is obtained with similar approximations (Brown and March, 1968; March, 1990b).

6.4 Computer Simulation of Viscosities in Lennard-Jones Fluids

While the Green–Kubo formulas for D and for the combination of viscosities in Eq. (35) are formally exact, applications using analytical methods, to date, are all invoking subsequent approximations; e.g., to reach a result like Eq. (36). In deriving this result, a fairly well-defined Debye frequency has been assumed at the higher "edge" (see Fig. 2) of the frequency spectrum $g(\omega)$. This assumption is best for simple liquid metals with relatively soft cores and with long-range pair potentials. It is less well suited to treat argon and similar Lennard-Jones liquids. Therefore in this section we refer to computer experiments on such liquids, which are conveniently carried out via time-dependent correlation functions, as set out below. Heyes (1986) has given a comprehensive account of such computer simulation of transport in Lennard-Jones liquids, such as argon (see Fig. 1) and krypton.

Heyes uses results of nonequilibrium molecular dynamics (MD) as the basis for his discussion. His prime concern was with the transport properties of simple liquids approaching the normal liquid density and beyond toward the glassy state. Heyes argues that to make the most effective use of computer resources, an intermolecular potential should be employed whose general shape represents that appropriate to many simple molecules. Using the transport properties of such a chosen "reference" system, which he specifically takes to be a Lennard-Jones 6–12 pair potential

$$\phi(r) = A/r^{12} - B/r^{6}, \tag{37}$$

one can then attempt to predict those corresponding to many real molecules by substituting in parameters specific to that molecule (see, for example, Tham and Gubbins, 1969; Boon *et al.*, 1967; de Bock *et al.*, 1967). This is then, essentially, invoking the corresponding states principle. [This principle has also been applied usefully to liquid metals near melting (see also Sec. 6.3), as reviewed by Abbasi and Blairs (1993).]

Below, we merely summarize the formulas used to calculate the shear viscosity η and bulk viscosity η_B. Because the above discussion of Longuet-Higgins and Pople for hard spheres related shear viscosity and thermal conductivity, we also give the formula for this third transport coefficient. These Navier–Stokes transport coefficients are related to equilibrium fluctuations by the Green–Kubo-type formulas already referred to, in a somewhat different form (see also Berezhkovsky *et al.*, 1984; Levesque *et al.*, 1973).

As the simplest example, the self-diffusion constant D can be written as the time integral over a velocity autocorrelation function, as already discussed [see Eq. (31)]. Turning to our main focus, the shear viscosity as a time integral involves the pressure tensor (see Croxton, 1974). Explicitly, with V the MD cell volume,

$$\eta = \frac{V}{9k_B T}\int_0^\infty dt\langle p_{xy}(t)p_{xy}(0)\rangle, \tag{38}$$

where p_{xy} is the xy off-diagonal element of the pressure tensor. If p denotes $\frac{1}{3}(p_{xx} + p_{xy} + p_{zz})$, then the formula for bulk viscosity η_B is

$$\eta_B = \frac{V}{9k_B T}\int_0^\infty dt\langle[p(t) - \langle p\rangle] \times [p(0) - \langle p\rangle]\rangle. \tag{39}$$

Finally, the thermal conductivity λ is, in terms of heat flux J,

$$\lambda = \frac{V}{3k_B T^2}\int_0^\infty dt\langle \mathbf{J}(t)\cdot\mathbf{J}(0)\rangle. \tag{40}$$

Table 2. Shear viscosities of liquid metals near melting temperature in 10^{-2} poise; after Brown and March (1968).

	Li	Na	K	Rb	Cs	Cu	Ag	Au	In	Sn
Experiment	0.60	0.69	0.54	0.67	0.69	4.1	3.9	5.4	1.9	2.1
Theory Eq. (36)	0.56	0.62	0.50	0.62	0.66	4.2	4.1	5.8	2.0	2.1

These are the formulas that Heyes employs in an extensive molecular dynamics simulation for Lennard-Jones fluids. The ratio λ/η, in such a computer experiment, will now vary with temperature, at constant pressure, and thereby will reflect the behavior shown in Fig. 1.

7. NON-NEWTONIAN FLUIDS

So far in this article, we have been concerned solely with liquids that obey Newton's law of constant viscosity. For such a Newtonian liquid, the shear viscosity η is a constant at a given temperature and pressure and is independent of the rate of shear. The relation satisfied by such a liquid is

$$\eta = (\text{shear stress})/(\text{rate of shear}). \tag{41}$$

Provided the flow is laminar, the plot of shear stress against the rate of shear is a straight line whose slope is the constant shear viscosity (see Fig. 3, curve 1). Water is a Newtonian liquid, and so are many pure single-phase liquids and solutions of substances of low molecular weights. In such liquids, viscous dissipation can be regarded as the result of collisions between fairly small molecules (cf. Sec. 3).

There are, however, other types of liquids—the so-called non-Newtonian liquids, which instead of having a constant coefficient of shear viscosity have one that varies with the rate of shear. The possible types of behavior of one class of non-Newtonian fluids are shown schematically in Fig. 3. Curve 1 is the usual Newtonian behavior, the slope being the shear viscosity of this liquid. Curve 2, in contrast, has no linear portion, and so the "effective" value of η will vary as one moves along this curve. It is therefore only meaningful to quote a value of η for a particular rate of shear. Liquids of this type thin upon shearing; that is, the effective η decreases as the shear rate increases. Examples of such "shear-thinning" liquids are dilute solutions of high polymers, some polymer melts like rubbers and cellulose, and suspensions such as paints. In rough terms, such shear-thinning behavior in, say, a dilute high-polymer solution can be viewed as resulting from the increasing alignment of the long polymer molecules along the streamlines as the rate of shear increases. The effective viscosity becomes less as this degree of alignment increases; when the molecular major axes eventually coincide with the streamlines, one can expect thereafter that the effective viscosity reaches a steady value.

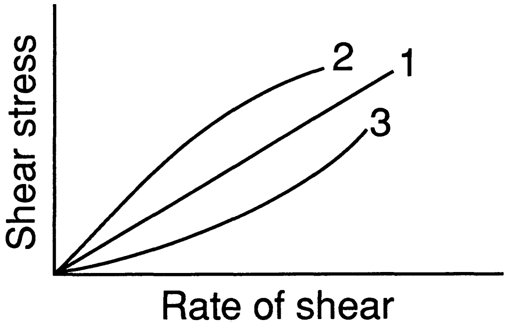

FIG. 3. Schematic plot of shear stress vs rate of shear. Curve 1: Typical Newtonian liquid. Curves 2 and 3: Different types of non-Newtonian liquids [e.g., curve 3 corresponds to dilatant (shear-thickening) type]

Curve 3 shows the opposite situation, namely the behavior of a dilatant or shear-thickening liquid. An example is that given by Collyer (1973), in which 66.7 g of wheat starch are mixed with 100 cm^3 of water. It does not seem easy to construct a molecular mechanism to interpret this shear-thickening behavior at the time of writing.

Other examples of non-Newtonian fluids exist, corresponding to materials with a yield value; these appear to behave as a mixture of a solid and a liquid, and we refer the reader to Temperley and Trevena (1978) for a brief discussion. The properties of non-Newtonian liquids are of considerable importance, for example, in the food industry.

7.1 Viscoelasticity

There is a class of liquids that possesses elastic as well as viscous properties. Such a "viscoelastic" material is mainly liquidlike in its behavior, while having some of the characteristics of an elastic solid. For example, if it is sheared, the material is able to store elastic energy just as a solid can.

Numerous solutions and melts of high polymers with very long molecules (macromolecules) are viscoelastic. Their viscous behavior is related to the degree of alignment of the macromolecules with the streamlines and depends on the bonds between these molecules. Their elastic behavior is tied up with the internal structure of the macromolecules. Each of these has a number of rigid links, and each link can take up various orientations with respect to its neighbors. This enables each molecule to assume a large number of shapes as the material is deformed.

7.1.1 Example of Properties of Solutions of Polyethylene Oxide Following Trevena (1968), let us summarize some of the effects that characterize viscoelastic liquids. Trevena focuses on solutions of polyethylene oxide because they are easily prepared. Properties include the following:

1. Elastic recovery. This property is readily observed by pouring the liquid from one beaker into another. If the pouring is stopped, for a brief period, and the vertical column of liquid is snipped sharply with a pair of scissors, the upper part of the column "recovers" quickly by recoiling suddenly into the upper beaker.
2. Weissenberg effect. The so-called Weissenberg effect occurs when a vertical cylindrical rod is immersed in a viscoelastic liquid and is then rotated rapidly. It is then observed that the liquid climbs up the rotating rod for quite some distance. Collyer (1973) has demonstrated this effect particularly impressively by "sandwiching" a layer of viscoelastic liquid between two Newtonian liquids. When a glass rod is rotated in this three-layered arrangement, the viscoelastic liquid climbs up and down the rod into both Newtonian liquids. The streamlines in the viscoelastic liquid around the rotating rod are horizontal concentric circles whose centers lie on the axis of rotation of the rod. It is also known (see, for example, Trevena, 1968) that in a viscoelastic liquid the stresses along the streamlines and those normal to the streamlines—in this case, the vertical direction—are not the same. This difference between the stresses causes the viscoelastic liquid to climb up and down the rod.
3. Die-swell effect. Another effect caused by the difference between these stresses is the die-swell effect, which occurs when a viscoelastic liquid flows downward through a tube of circular cross section. In this instance, stresses are set up normal to the streamlines, which are parallel to the axis of the tube. When the liquid emerges at the lower end of the tube, these normal stresses are relieved by a substantial increase in diameter of the emerging stream. This increase in diameter can be two or three times the diameter of the tube and is termed a die-swell. The effect is of importance in manufacturing manmade fibers and plastics.
4. Kaye effect. Kaye (1963) found that, if a certain solution of polyisobutylene in dekalin were poured into a dish, then when the falling stream was sufficiently thin the stream would bounce off the liquid in the dish every few seconds and rise in an arc before falling back into the dish. This so-called Kaye effect has been studied subsequently by Collyer and Fisher (1976; see also Temperley and Trevena, 1978). No convincing interpretation seems available at the time of writing.

7.2 Some Theoretical Concepts for Viscoelastic Liquids

It is useful in developing a theoretical treatment of the behavior of viscoelastic liquids to generalize the concept of shear modulus to become a complex quantity, G^* say. Trevena illustrates this as follows. Consider a viscoelastic liquid being subjected to a small-amplitude sinusoidal stress in a coaxial cylinder viscometer. This consists of a hollow outer cylinder and a suspended inner cylinder with the liquid occupying the gap between. If the outer cylinder is rotated sinusoidally through a small angle $\phi = \phi_0 \sin\omega t$, the resulting movement of the liquid will cause the inner cylinder to oscillate with the same angular frequency ω. There will, however, be a phase difference, δ say, between the oscillations of the two cylinders, so that $\theta = \theta_0 \sin(\omega t + \delta)$ represents the movement of the inner cylinder. If the space between the cylinders were to be occupied instead by a perfectly elastic solid, the two cylinders would oscillate in phase, corresponding to $\delta = 0$. In the other extreme, if the space were to be filled with a Newtonian (that is, purely viscous) liquid, then it is found that $\delta = \pi/2$ (see, for example, Trevena, 1975).

Returning to the intermediate case $0 < \delta < \pi/2$ corresponding to a viscoelastic liquid in the viscometer, the phase angle δ is clearly a measure of the elasticity of the liquid. In the course of each swing, a part of the energy is stored elastically with the increase of displacement and is recovered in the counterswing as this displacement decreases. The rest of the energy is lost as heat because of the viscous forces, and this is not recovered.

The real part G' of the elastic modulus

$$G^* = G' + iG'' \qquad (42)$$

is a function of the energy stored in the swing, and the imaginary part of G^*, the "viscous" modulus G'', is a function of the energy lost as heat. The phase angle δ above is then given by

$$\delta = \tan^{-1}(G''/G'). \qquad (43)$$

As already implicit in the discussion above, for a purely elastic solid $G'' = 0$ and $\delta = \pi/2$. Since, on dimensional grounds, shear modulus = viscosity/time, one can write

$$G' = \eta'\omega, \qquad (44)$$

etc., and the result

$$\eta^* = \eta' + i\eta'' \qquad (45)$$

also holds, where η^* is the so-called complex dynamic viscosity.

7.2.1 More Complex Types of Motion and Equations of State of Liquid

The above experiment is concerned with one of the simplest types of motion of a viscoelastic liquid. When a liquid is subjected to more complicated types of motion, it becomes important to formulate equations of state for the liquid. Such equations have to satisfy two principles:

1. a "memory" principle, namely that the behavior of an element of the liquid depends only on its previous deformation history and not on the state of neighboring elements; and
2. an "invariance" principle, namely that the behavior of an element of the material does not depend on the motion of the liquid as a whole through space.

These principles **1** and **2** can be used to consider the invariant forms of the equations of state for viscoelastic liquids that apply under all conditions of motion. We refer the reader to the account of Trevena (1968) for further details and other references.

8. INTERFACIAL TRANSPORT AND RHEOLOGY

The physics of the interfacial region separating two fluids is an area of considerable current interest at the time of writing (see SURFACES AND INTERFACES OF FLUIDS, STRUCTURE OF). The area embraces soft condensed matter, and in particular complex fluids such as liquid–liquid emulsions and gas-liquid foams. Interfacial rheology is concerned with the theory, measurement, and application of interfacial hydrodynamics. One can start by adopting the classical macroscopic view of fluid interfaces as idealized two-dimensional singular surfaces. The adsorption of molecular or macromolecular surfactants imparts intrinsic rheological properties to the interface (see RHEOLOGY), such as interfacial shear and dilatational viscosities (which are two-dimensional counterparts of the three-dimensional viscosities treated above) and Gibbs elasticity, which indicates the change in interfacial tension with area. Gradients in surfactant concentration and temperature cause interfacial tension gradients that produce such phenomena as the "tears" of strong wine.

As to experiment, light-scattering techniques have acquired considerable importance (see, e.g., Dorshow and Turkevich, 1993), as has the deep-channel viscometer for measuring interfacial shear viscosity. This is defined (cf. Fig. 3 for bulk liquids) as the ratio of interfacial shear stress to rate of shear.

As part of the above general area, one can cite important applications including Rayleigh and Bénard instabilities, interfacial turbulence, thin–liquid-film hydrodynamics and stability, and the rheology and stability of foams and emulsions. The understanding of the metastability of complex fluids, involving the use of Derjaguin–Landau–Verwey–Overbeek theory, is important to pure scientists and technologists alike. The reader wishing to go further should consult Edwards *et al.* (1991).

GLOSSARY

Acoustic Velocity, *c*: Speed of propagation of sound in liquid.

Compressibility: Ratio of fractional volume change to hydrostatic pressure, $K = V^{-1}\partial V/\partial P$. Can be either isothermal (K_T) or adiabatic.

Diffusion Coefficient, *D*: Ratio between the flux of a substance brought about by a concentration gradient and the value of the gradient: $J = D\partial C/\partial s$.

Elastic Moduli: Coefficients of proportionality between imposed stress and resultant strain. They include the *bulk modulus*, which is the reciprocal of the **Compressibility**, and the *shear modulus*, defined for shearing stresses and strains. All have units of pressure.

Inelastic Neutron Scattering: Scattering of neutrons primarily by nuclei, producing a change in energy of the neutron. It is a principal means of measurement of the dynamical structure factor $S(q,\omega)$ in liquids.

Thermal Conductivity, λ: Coefficient of proportionality between heat flux and temperature gradient, $\partial Q/\partial t = \lambda\partial T/\partial s$.

Viscosity: Ratio between stress and rate of strain in a fluid. Classified as bulk, η_B; compressional, η'; or shear, η.

Works Cited

Abbasi, M. H., Blairs, S. (1993), *Phys. Chem. Liq.* **25**, 201.

Andrade, E. N. da C. (1934), *Philos. Mag.* **17**, 497–511.

Ascough, J. A., Chapman, R. G., March, N. H. (1988), *Phys. Chem. Liq.* **18**, 253–267.

Berezhkovsky, L. M., Drozdov, A. N., Zitserman, A. N., Logar'kov, A. N., Trigger, S. A. (1984), *J. Phys. F.* **14**, 2315–2321.

Boon, J. P., Legros, J. C., Thomas, G. (1967), *Physica* **33**, 547–557.

Brown, R. C., March, N. H. (1968), *Phys. Chem. Liq.* **1**, 141–158.

Chapman, S., Cowling, T. G. (1939), *The Mathematical Theory of Non-Uniform Gases*, Cambridge, U.K.: Cambridge University Press.

Collins, F. C., Raffel, H. (1954), *J. Chem. Phys.* **22**, 1728–1733.

Collyer, A. A. (1973), *Phys. Ed.* **8**, 111–116 and 333–338.

Collyer, A. A., Fisher, P. J. (1976), *Nature* **261**, 682–683.

Croxton, C. A. (1974), *Liquid State Physics*, Cambridge, U.K.: Cambridge University Press.

de Bock, A., Grevendonk, W., Awouters, H. (1967), *Physica* **34**, 49–52.

Dorshow, R. B., Turkevich, L. A. (1993), *J. Chem. Phys.* **98**, 8349–8350.

Edwards, D. A., Brenner, H., Wasan D. T. (Eds.) (1991), *Interfacial Transport Processes and Rheology*, Boston: Butterworths-Heinemann.

Egelstaff, P. A. (1967), *An Introduction to the Liquid State*, New York: Academic.

Ernst, M. H., Hauge, E. H., van Leeuwen, J. M. (1970), *Phys. Rev. Lett.* **25**, 1254–1256.

Gaskell, T., March, N. H. (1970), *Phys. Lett.* **33A**, 460–461.

Greenfield, A. J., Wellendorf, J., Wiser, N. (1971), *Phys. Rev. A* **4**, 1607–1616.

Hansen, J. P., McDonald, I. R. (1976), *Theory of Simple Liquids*, New York: Academic.

Heyes, D. M. (1986), *Can. J. Phys.* **64**, 773–781.

Irving, J. H., Kirkwood, J. G. (1950), *J. Chem. Phys.* **18**, 817–829.

Kaye, A. (1963), *Nature* **197**, 1001–1002.

Levesque, D., Verlet, L., Kürkijarvi, J. (1973), *Phys. Rev. A* **7**, 1690–1700.

Longuet-Higgins, H. C., Pople, J. A. (1956), *J. Chem. Phys.* **25**, 884–889.

March, N. H. (1990a), *Liquid Metals: Concepts and Theory*, Cambridge, U.K.: Cambridge University Press.

March, N. H. (1990b), *Chemical Physics of Liquids*, New York: Gordon & Breach.

March, N. H., Tosi, M. P. (1991), *Atomic Dynamics in Liquids*, New York: Dover.

Temperley, H. N. V., Trevena, D. H. (1978), *Liquids and Their Properties*, Chichester: Ellis Horwood.

Tham, M. J., Gubbins, K. E. (1969), *Ind. Eng. Chem. Fundam.* **8**, 792–795.

Trevena, D. H. (1968), *Contemp. Phys.* **9**, 537–548.

Trevena, D. H. (1969), *Contemp. Phys.* **10**, 601–614.

Trevena, D. H. (1975), *The Liquid Phase*, London: Wykeham.

Ubbelohde, A. R. (1978), *The Molten State of Matter*, New York: Wiley.

Van Hove, L. (1954), *Phys. Rev.* **95**, 249–262 and 1374–1384.

Zwanzig, R. (1983), *J. Chem. Phys.* **79**, 4507–4508; see also March, N. H. (1984), *ibid* **80**, 5345.

Further Reading

Edwards, D. A., Brenner, H., Wasan, D. T. (Eds.) (1991), *Interfacial Transport Processes and Rheology*, Boston: Butterworths-Heinemann.

Egelstaff, P. A. (1992), *An Introduction to the Liquid State*, 2nd ed., Oxford: Oxford University Press.

Hansen, R. S., Ahmad, J. (1971), *Prog. Surf. Membrane Sci.* **4**, 1.

Heyes, D. M. (1993), *J. Non-Newt. Fl. Mech.* **46**, 1.

Jiang, T. S., Chen, J. D., Slattery, J. C. (1983), *J. Colloid Interface Sci.* **96**, 7–19.

Levich, V. G. (1962), *Physicochemical Hydrodynamics*, Englewood Cliffs, NJ: Prentice-Hall.

Marshall, W., Lovesey, S. W. (1971), *Theory of Thermal Neutron Scattering*, Oxford: Clarendon. See especially for viscoelastic theory.

MECHANICAL PROPERTIES OF SOLIDS

ALAN COTTRELL, *Department of Materials Science, University of Cambridge, Cambridge, United Kingdom*

Introduction 545
1. **Infinitesimal Elastic Deformation** 545
1.1 Stress and Strain 545
1.2 Atomic Basis of Elastic Constants 547
1.3 Harmonic Oscillations 548
1.4 Thermal Vibrations 549
1.5 Deviations from Linear Elasticity 549
2. **Mechanical Failure** 551
2.1 Limiting Strengths of a Solid 551
2.2 Intrinsically Ductile and Brittle Crystals 552
2.3 Dislocations 552
2.4 Yield Strength 554
2.5 Work Hardening and Fatigue 555
2.6 Creep 557
2.7 Fracture Mechanics 557
2.8 Plasticity 558
Glossary 559
Further Reading 560

INTRODUCTION

Solids are defined by mechanical properties. They resist changes of both shape and volume. The distinction between a solid and a fluid is, however, a purely conventional one, set at a viscosity ($\simeq 10^{14}$ Pa s) at which a familiar stress, e.g., one atmosphere, would in, e.g., one year produce significant deformation.

There is no structural definition. The crystalline state is often regarded as the solid state, but hot crystals flow, i.e., creep, and cold glass does not.

The mechanical properties of solids divide naturally into two parts: elastic properties and mechanical failure. Elastic properties are provoked by weak applied forces that displace atoms slightly from neighbors. On removal of these forces the atoms return to base. This is elastic deformation. Larger forces, applied generally to the whole solid body or locally at stress concentrators such as crack tips, can break the interatomic bonds. Mechanical failure then occurs by atoms either sliding over one another, as in ductile solids, or separating, as in brittle ones.

The first section of this article outlines the principles of infinitesimal elastic deformation, Hooke's law, elastic constants, their relation to atomic structure and cohesive properties, and deviations from linear elasticity. The inclusion of inertial forces extends the theory to harmonic oscillations, normal modes, and thermal properties.

The treatment of mechanical failure is necessarily atomistic, since the ratio of surface energy to elastic constant, which will be seen later to be two basic underlying properties, is an atomic length. The two failure modes, sliding or parting, lead to the dualities of shear and fracture strengths, of ductile and brittle solids, and of dislocation and crack defects. The rich, complex field of engineering mechanical properties—yielding, work hardening, tensile strength, fatigue, creep, fracture toughness, plasticity—then comes into view.

1. INFINITESIMAL ELASTIC DEFORMATION

1.1 Stress and Strain

It is convenient, initially, to model a solid by an elastic continuum. A second useful concept is the dividing surface. The contiguous body is imagined to be two continuous parts, continuously joined at a common dividing surface across which each exerts con-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

tact forces on the other. Body forces such as gravitation and inertia are not excluded, but become of secondary importance when the body is thus divided into infinitesimal elements, in each of which the body forces, proportional to volume, are small compared with the contact forces, proportional to surface area.

The contact force per unit area is the stress on the dividing surface. When obliquely directed, it can be resolved into a normal component, tensile or compressive, and a shear component along the surface. Since both this oblique direction and the surface orientation are vectors, stress is a second-order tensor with nine components, σ_{ij} where $i, j = 1, 2, 3$. Static equilibrium requires that

$$\sigma_{ij} = \sigma_{ji}, \tag{1}$$

so that the tensor is symmetric, with six independent components. It can be split into a hydrostatic part

$$p = -\tfrac{1}{3}(\sigma_{11} + \sigma_{22} + \sigma_{33}), \tag{2}$$

where p = pressure, and a deviatoric part, $\sigma_{ij} + p\delta_{ij}$, where $\delta_{ij} = 1$ when $i = j$ and $\delta_{ij} = 0$ otherwise. It can also be described in terms of principal stresses, i.e., an orientation of orthogonal axes can always be found in which the local stresses act only normally, not obliquely.

The deformation of a body displaces the coordinates of any point of it from x_i ($i = 1,2,3$), initially, to $x_i + u_i$ afterwards. These displacements generally deform any sufficiently small cube in the body, initially with edges of unit length, into a parallelepiped of edge lengths $1 + \epsilon_{11}$, $1 + \epsilon_{22}$, and $1 + \epsilon_{33}$, and change the angles between these edges by the amounts $-2\epsilon_{12}$, $-2\epsilon_{23}$, and $-2\epsilon_{31}$, respectively. These are the components of the strain tensor, which is second order and symmetric,

$$\epsilon_{ij} = \epsilon_{ji}, \tag{3}$$

and related to the deformation by

$$\epsilon_{ij} = \tfrac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right). \tag{4}$$

It is representable by principal strains, in axes oriented so that $\epsilon_{ij} = 0$ when $i \neq j$, and has dilatational,

$$\Theta = \Delta V/V = \epsilon_{11} + \epsilon_{22} + \epsilon_{33}, \tag{5}$$

and deviatoric, $\epsilon_{ij} - \frac{1}{3}\Theta\delta_{ij}$, parts.

Resistance to elastic deformation implies that strain increases the energy of the system. Because of symmetry about $\epsilon = 0$ this strain energy increases as ϵ^2 for infinitesimal deformation. The work done by displacement of the applied forces, as the body deforms, supplies this energy. For infinitesimal deformation the stress is proportional to strain, i.e., Hooke's law is obeyed:

$$\sigma_{ij} = \sum_{p,q} c_{ijpq}\epsilon_{pq}, \tag{6}$$

where c_{ijpq} are the elastic constants of the material. Effects of crystal symmetry and structural isotropy greatly reduce the number of independent elastic constants. Crystals of cubic symmetry have only three. In a simplified notation that allows the four-index form of c to be reduced to a two-index one, these are

$$\begin{aligned} &c_{11}\ (= c_{22} = c_{33}),\\ &c_{12}\ (= c_{23} = c_{31}),\\ &c_{44}\ (= c_{55} = c_{66}), \end{aligned} \tag{7}$$

where $\sigma_{11} \rightarrow \sigma_1$, $\sigma_{23} \rightarrow \sigma_4$, $\epsilon_{11} \rightarrow \epsilon_1$, $\epsilon_{23} + \epsilon_{32} \rightarrow \gamma_4$, etc., and γ is the angle of shear. These imply three independent deformation modes: hydrostatic dilatation, resisted by the bulk modulus of elasticity K,

$$K = \tfrac{1}{3}(c_{11} + 2c_{12}), \tag{8}$$

and two differently oriented shear modes,

$$\mu_0 = c_{44} \tag{9}$$

and

$$\mu_1 = \tfrac{1}{2}(c_{11} - c_{12}). \tag{10}$$

The material is elastically isotropic when these shear moduli of elasticity are equal,

$$\mu_0 = \mu_1 = \mu. \tag{11}$$

Other elastic constants are in this case conveniently defined. Young's modulus E relates uniaxial tension to extension,

$$E = \sigma_1/\epsilon_1, \tag{12}$$

and Poisson's ratio ν relates lateral contraction to extension,

$$\nu = -\epsilon_2/\epsilon_1. \tag{13}$$

Relations exist between K, μ, E, and ν; e.g.,

$$\begin{aligned} E &= 3\mu[1 + (\mu/3K)]^{-1} \\ &\simeq 3\mu[1 - (\mu/3K)], \end{aligned} \tag{14}$$

which shows that E depends mainly on μ.

For further treatment of stress and strain, see Bilby *et al.* (1985), Cottrell (1964), or Zener (1948).

1.2 Atomic Basis of Elastic Constants

The elastic constants are only slightly related to properties such as vaporization energy and boiling point, because they depend sensitively on the hardness of atoms, i.e., on how sharply the energy rises when these are compressed. The bulk modulus of an ionic crystal such as NaCl provides a simple example. The interaction energy of two singly charged ions, at distance r, is approximately

$$\phi(r) = \pm e^2/r + B/r^s, \tag{15}$$

where the $\pm e^2/r$ electrostatic interactions (with $e = 4.8 \times 10^{-10}$ esu; equations in this section are in cgs units) refer to like and unlike ionic charges and s ($\simeq 9$) in the short-range repulsion, B/r^s, indicates the hardness of the atoms. The sum over all interactions gives the energy per nearest-neighbor bond as

$$\phi = -\frac{Ae^2}{6r}\left[1 - \frac{1}{s}\left(\frac{r_0}{r}\right)^{s-1}\right], \tag{16}$$

where r now refers to the nearest-neighbor spacing, with equilibrium value r_0, and the dimensionless constant $A = 1.7476$. The bond energy can be written as

$$\phi = \phi(r_0) + \left(\frac{d\phi}{dr}\right)dr + \tfrac{1}{2}\left(\frac{d^2\phi}{dr^2}\right)d^2r + \ldots, \tag{17}$$

giving the elastic force in the bond as

$$f = \left(\frac{d^2\phi}{dr^2}\right)dr \tag{18}$$

and hence an elastic constant

$$\left(\frac{d^2\phi}{dr^2}\right)_0 = \frac{4Ae^2}{3r_0^3}. \tag{19}$$

From the number of bonds through a cube face of the crystal, of unit area, the bulk modulus is then

$$K = (s - 1)Ae^2/18r_0^4, \tag{20}$$

so that K increases with the hardness s. For NaCl this simple expression gives $K = 29$ GPa, whereas the observed value is 24.3.

In this example the forces act along the lines between the centers of atoms. Cauchy proved that, for this case, $c_{12} = c_{44}$. Generally, this relation fails. For example, in metals much of the bonding is provided by the free-electron gas, the energy of which is sensitive to changes of volume but not of shape.

For assemblies, such as crystals, of closely packed atoms each of which interacts strongly with all its close neighbors, the elastic constants are usually of order 100 GPa. In contrast, the soft elasticity ($\simeq$10 MPa) of rubber and elastomers is due to their chainlike molecular structures, long spaghettilike threads that wriggle across one another. This soft elasticity is due, not to straining of bonds, but to thermal motion. Starting from its fully extended state of length Nl a long chain of N links, each of length l, gains a configurational entropy $k \ln N$, where k is Boltzmann's constant, when one of its links inverts and reduces its length to $(N - 2)l$. Generalizing, when the length is $L = nl$ the configurational entropy can be shown to be

$$S(L) = C - k(L^2/2Nl^2), \tag{21}$$

where C is a constant. The force f to hold the chain at this length is then

$$f = -T\frac{\partial S}{\partial L} = \frac{kT}{Nl^2}L, \tag{22}$$

which gives Hooke's law with an elastic constant that increases linearly with temperature.

In polymers, chemical cross-linking at the intersections of long-chain molecules increases the elastic stiffness until, in the limit of dense cross-linking, as in vulcanite, the properties of a rigid solid are approached. At the other extreme, gels have a sparse framework of cross-linked molecules that gives them a soft elasticity, and an infilling of liquidlike material that gives them viscoelasticity.

For further discussion of the atomic basis of elastic constants, see Cottrell (1964), Kelly and MacMillan (1986), Ward (1983), or Zener (1948).

1.3 Harmonic Oscillations

The transition to elastodynamics, by adding inertia to the forces, introduces the mass density ρ of the material. For infinitesimal Hookeian deformation of a continuum, the equation

$$\left(\frac{d^2u}{dx^2}\right) = \frac{1}{c^2}\frac{d^2u}{dt^2} \tag{23}$$

has simple sinusoidal wave solutions for the displacement u. These are, of course, sound waves. The sonic speed c depends on the vibrational mode, longitudinal or transverse, and the mechanical constraint. For longitudinal waves in a rod stretched in uniaxial tension, c is given by

$$c_0 = (E/\rho)^{1/2}, \tag{24}$$

where E is Young's modulus. For the corresponding wave in a thick body, where the geometry constrains the lateral displacements, the value is

$$c_1 = (c_{11}/\rho)^{1/2}. \tag{25}$$

Transverse or shear waves, at constant volume, have a speed

$$c_2 = (\mu/\rho)^{1/2} \tag{26}$$

in an isotropic medium. Finally, at a free surface, waves in the surface layers have the value

$$c_3 = f(\nu)(\mu/\rho)^{1/2}, \tag{27}$$

where $f(\nu)$, typically about 0.93, is a function of Poisson's ratio. These slow surface waves are important seismologically since, spreading out only two dimensionally, they remain strong far from epicenters.

To deal with short wavelengths, atomistic theory is essential and the discreteness of the atomic structure restricts the wavelengths to certain values. Let a straight chain of N equidistant and equal atoms, each of mass M, oscillate, longitudinally only, so that atom n has displacement u_n (Fig. 1). Assume only nearest-neighbor forces, with a spring constant α. The force on atom n is then

$$\begin{aligned} f_n = \alpha[(u_{n+1} - u_n) \\ - (u_n - u_{n-1})] = M\ddot{u}_n. \end{aligned} \tag{28}$$

The N solutions of these N interlinked equations are found by representing the motion as N independent normal modes, in each of which there is one single oscillation frequency, ω, for the entire chain, i.e.,

$$u_n(t) = u_n^0 e^{-i\omega t}. \tag{29}$$

The equation of motion of atom n in this particular mode is

$$-M\omega^2 u_n = \alpha(u_{n+1} + u_{n-1} - 2u_n). \tag{30}$$

This is unchanged by the substitution $n \to n + 1$ and so has the solution

$$u_{n+1} = u_n e^{ika} \tag{31}$$

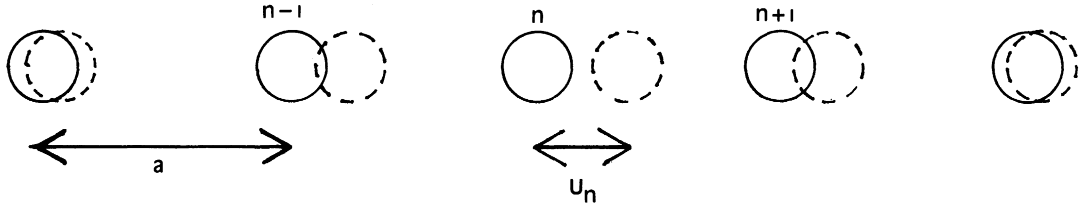

FIG. 1. Longitudinal displacements in a line of atoms.

and thus

$$u_n^0 = Ae^{ikna}, \tag{32}$$

where A is a constant and k is the wave number, i.e., 2π/(wavelength). Equation (30) is then

$$\begin{aligned} -M\omega^2 &= \alpha(e^{ika} + e^{-ika} - 2) \\ &= -4\alpha \sin^2(ka/2), \end{aligned} \tag{33}$$

so that

$$\omega = \pm(4\alpha/M)^{1/2} \sin(ka/2). \tag{34}$$

This dispersion relation reveals that there is a maximum frequency

$$\omega_m = 2(\alpha/M)^{1/2}, \tag{35}$$

typically 10^{13} s^{-1}, at the smallest wavelength of twice the atomic spacing.

In the long-wavelength limit, $k \to 0$, the dispersion relation reduces to

$$\omega = (\alpha a^2/M)^{1/2}k, \tag{36}$$

and in this nondispersive limit, where

$$\omega \propto k, \tag{37}$$

the wave speed becomes

$$c = \omega/k = (\alpha a^2/M)^{1/2}. \tag{38}$$

This, with $\rho = M/a$, reproduces the continuum result with elastic constant αa.

1.4 Thermal Vibrations

The vibrational spectrum is the number of modes of harmonic oscillation in a given range of frequency or wave number. For the linear chain of N atoms (Fig. 1), disregarding free-end effects, the N modes are distributed one for each value of

$$k = 2\pi l/Na, \tag{39}$$

where l takes all integral values between $\pm N/2$. In three dimensions the number of different modes increases mainly as ω^2, but with strong deviations and peaks at high frequencies.

Quantum mechanics shows that each mode is allowed vibrational energy E only at the levels

$$E = (n + \tfrac{1}{2})\hbar\omega, \tag{40}$$

where the integer n gives the number of energy quanta in the mode, additional to the zero-point energy $\frac{1}{2}\hbar\omega$, and $2\pi\hbar$ is Planck's constant. These quanta of elastic vibrational energy, called phonons, behave like "particles" of zero rest mass.

In thermal equilibrium at temperature T a mode of frequency ω has on average an energy

$$\epsilon = \hbar\omega/(e^{\hbar\omega/kT} - 1), \tag{41}$$

and the sum of this over all modes gives the total vibrational energy, the variation of which with temperature is the vibrational specific heat. Despite the form of Eq. (41) there is some absorption of thermal energy at very low temperatures because the long-wavelength modes have low-energy phonons.

For additional discussion of harmonic oscillations in lattices, see Cottrell (1964), Kittel (1986), or Kolsky (1953).

1.5 Deviations from Linear Elasticity

At large strains the higher-order terms in Eq. (17) introduce significant nonlinearity. In addition to this nonlinear elasticity there may be some physically different nonelastic deformation. In strong solids in tension or shear, the latter, heralding the approach of mechanical failure, usually appears at elastic strains below 0.01, where nonlinear elastic deformation is still barely perceptible. Very thin fibers and whisker crystals can, however, often withstand tensile strains of up to 0.1, and the deviation from Hooke's law is then quite plain (Fig. 2). Positive deviations can also be detected in highly compressed solids, e.g., in explosions (Fig. 2) (Walsh and Christian, 1955). These deviations cause later parts of a strong compression wave to travel faster than earlier ones, so that the wave sharpens up into a shock front as it advances.

Even small deviations from elastic linearity are physically important. In a strictly linear system the waves are completely transparent to one another. Thermal oscillations from a heated end of a bar could then travel along the bar at the speed of sound. In prac-

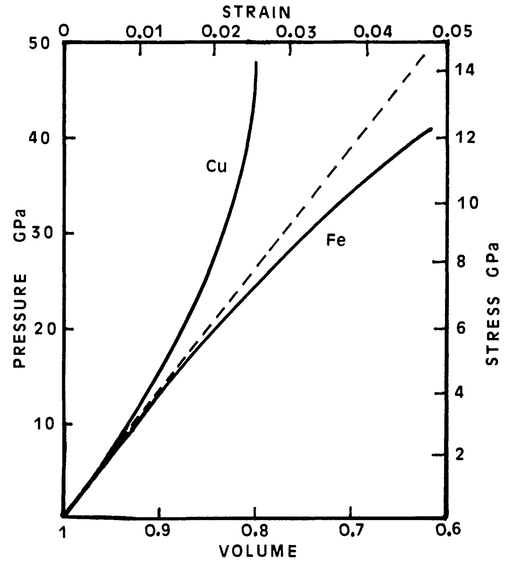

FIG. 2. Stress-strain curve of an iron whisker, in tension, and pressure-volume curve of highly compressed copper, both showing deviations from Hooke's law.

tice, however, the anharmonic terms in the law of interatomic force cause the phonons to collide and scatter each other. The passage of a phonon down the bar is then a chancy, zigzag motion, described macroscopically by Fourier's law of heat conduction. Sound waves themselves do not suffer this scattering, except insignificantly, because their wavelengths are too long. On the atomic scale a macroscopic wave appears as an almost uniform, constant deformation. The atoms in this wave vibrate very much as they would in a statically stressed solid and interact with it only weakly, through effects such as thermal expansion.

Elastic anharmonicity is responsible for thermal expansion. The softening of elastic constants in expanded material reduces $\hbar\omega$ and thus allows more quanta in, for the same total energy. The extra configurational entropy of these favors this expanded state, leading to a linear coefficient of thermal expansion,

$$\alpha = \gamma C_V/3V_0K, \tag{42}$$

where C_V is the specific heat at constant volume, V_0 is the volume of unit mass, K is the bulk modulus, and γ is Grüneisen's constant,

$$\gamma = -d(\ln \omega)/d(\ln V) \simeq 2(\pm 1), \tag{43}$$

which measures the hardness of the atoms.

Various anelastic processes can dampen mechanical vibrations in solids, an effect that is important acoustically. Church bells are made from hard, brittle alloys to suppress damping, and organ pipes from soft lead alloy to damp out resonances. Soundproof wall coverings consist of loosely adherent materials, the flaky particles of which chafe one another and so dissipate reverberation energy. These anelastic properties are described in many ways. Viscoelastic solids can be specified by Maxwell's time of relaxation

$$\tau = \eta/\mu, \tag{44}$$

where η is the viscosity and μ the shear modulus; and by the ratio of unrelaxed and relaxed elastic moduli, as measured, respectively, under impulsive and prolonged loads. A general indicator of internal friction is the damping capacity $\Delta E/E$, i.e., the energy ΔE dissipated in one cycle of vibration with vibrational energy E. Various other indicators are also used; e.g., the logarithmic decrement δ of amplitude decay, the phase lag γ between strain and stress, in sinusoidal oscillations, and the resonance factor

$$Q^{-1} = (\omega_1 - \omega_2)/\omega_0, \tag{45}$$

where ω_1 and ω_2 are the frequencies—on either side of the resonant one ω_0—at which the amplitude is 0.7 of the peak. They are related by

$$\Delta E/E = 2\delta = 2\pi \tan \gamma = 2\pi Q^{-1}. \tag{46}$$

Viscoelastic processes occur obviously in gels and hot glass, where atoms and molecules in liquidlike regions of the structure can slip, individually, past one another. In densely packed crystals these individual atomic rearrangements are largely suppressed, except at high temperatures where there is slow, viscous creep due to the formation of vacant atomic sites on one face and their migration to and disappearance on another face. This can also happen in polycrystals, with the grain boundaries (*q.v.*) acting as sources and sinks of vacancies. Also in polycrystals, the breakdown of crystalline order at grain boundaries enables viscous sliding to take place along them at high temperatures. In fine-grained polycrystals this can lead to superplasticity, the (inappropriate) name for a hot-glass-like behavior of the material at low strain rates and high temperatures. Some alloys, e.g., titanium, are industrially shaped superplastically.

For further discussion of deviations from linear elasticity see Cottrell (1964), Ward (1983), or Zener (1948).

2. MECHANICAL FAILURE

2.1 Limiting Strengths of a Solid

Under large tensile forces the cohesion of the atoms may fail by the atoms pulling apart, at an ideal tensile strength σ_t, or by sliding across one another, at an ideal shear strength σ_s. Atomic theory shows that interatomic cohesive forces peak at a tensile strain of about 0.2. With some elastic softening at these high strains, this implies that

$$\sigma_t \simeq 0.1E, \tag{47}$$

where E is Young's modulus.

Sliding in a crystal is very different from the individual atomic shuffles of fluid flow. The regimentation of the crystal lattice enforces a remarkable collective movement—plastic slip or glide, or sometimes mechanical twinning—in which immense numbers of atoms slide in a highly coordinated way (Fig. 3). Let the imaginary dividing surface of Sec. 1.1 now lie between a pair of neighboring close-packed crystallographic planes. The translational symmetry of the crystal implies that, if the half-crystal on one side of this surface slides over the other by one lattice unit in the direction of slip (usually a close-packed direction), crystalline perfection and full cohesion are then perfectly restored. All properties in fact are periodic functions of this slip displacement vector, including the contact shear stress between the faces of this slip plane. Its simplest form is

$$\sigma = \frac{b}{a}\frac{\mu}{2\pi}\sin\left(\frac{2\pi u}{b}\right) \tag{48}$$

for a slip displacement u between planes of spacing a, shear modulus μ, and repeat distance b in the slip direction (Fig. 3). It gives Hooke's law, $\sigma = \mu(u/a)$, when u is small; and it repeats with period b. The ideal shear strength is then

$$\sigma_s = \frac{b}{a}\frac{\mu}{2\pi} \simeq 0.15\mu, \tag{49}$$

at a shear strain of about 0.25. In practice, the hardness of atoms usually causes shear failure at smaller strains. Furthermore, the sliding atoms may find intermediate positions of local stability, corresponding to twin configurations or stacking faults at smaller values of b. As a result, in soft metal crystals, such as Al or Cu, σ_s is reduced to about 0.04μ, whereas in directionally bonded ones such as diamond and silicon it is about 0.1μ.

In tensile fracture the creation of new sur-

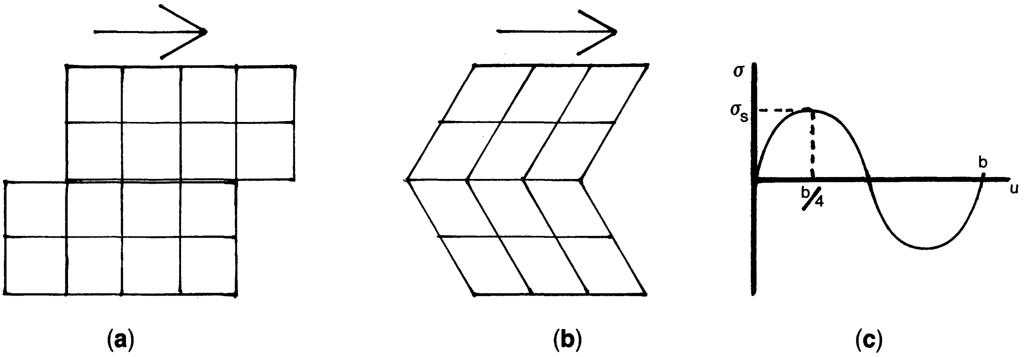

FIG. 3. (a) Crystallographic slip, (b) deformation twinning, and (c) sinusoidal model of interatomic forces between faces of slip plane.

faces, by separation of atoms, involves the surface energy γ. Figure 4 shows a simple model tensile law of force, approximated by a half sine wave of height σ_t and wavelength λ, the area under which is 2γ. This law,

$$\sigma = \sigma_t \sin(2\pi u/\lambda), \tag{50}$$

together with Hooke's law, $\sigma = E(u/a)$ at small u, gives

$$\sigma_t \simeq (E\gamma/a)^{1/2} \simeq 0.2E, \tag{51}$$

with, typically, $20\gamma = Ea$. Again, a lower value, e.g., $0.1E$, is expected in practice.

For additional discussion of limiting strengths of a solid, see Cottrell (1953, 1964), Gordon (1976), Kelly and MacMillan (1986), or Nabarro (1967).

2.2 Intrinsically Ductile and Brittle Crystals

Crystalline solids divide according to their low-temperature properties into intrinsically brittle ones, e.g., diamond and Si, and intrinsically ductile ones, e.g., Al and Cu. At the tip of a sharp crack, the locally concentrated force develops a large hydrostatic component so that the peak tensile stress there is typically some six times the shear stress. Assuming $\sigma_s = 0.04\mu$ and $\sigma_t \simeq 0.1K$ (for $\nu = \frac{1}{3}$), these local stresses bring the atoms equally to the point of sliding or parting when $\mu/K \simeq 0.4$. Pugh (1954) noted that the value of μ/K generally distinguishes the two classes of cubic crystals. Metals of cubic structure in fact are usually intrinsically ductile when $\mu/K <$ 0.4 and brittle when $\mu/K > 0.5$ (Cottrell, 1989). These two elastic constants are the simplest indicators of the resistance of the atoms to sliding or parting, respectively.

For further discussion see Cottrell (1964), Kelly and MacMillan (1986), or Knott (1973).

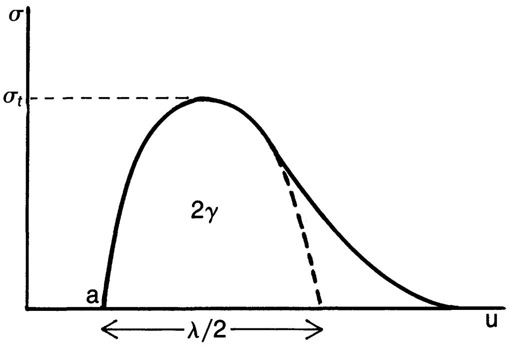

FIG. 4. Model interatomic force between layers separated by tensile stress.

2.3 Dislocations

At a crack tip, parting failure simply enlarges the crack, but sliding failure introduces a new feature, the spread of slip along glide planes from the tip to distant parts of the crystal. The slip "fronts" that herald the arrival of successive units of slip displacement and that emerge from the tip to advance along the glide plane are slip dislocations. The translation vector **b** of a unit slip displacement is the Burgers vector of its associated dislocation.

In practice, crystal dislocations exist naturally as original defects of the as-grown crystal (see POINT DEFECTS AND OTHER IMPERFECTIONS IN CRYSTALS). Moreover, when they move in slip planes they often multiply. Real crystals thus usually contain natural populations of dislocations. Only a few salient features of dislocation theory can be summarized here:

1. A dislocation is a line defect, e.g., it is the boundary between different areas in a slip plane. The line can in principle follow an arbitrary path, although crystal structure and line tension (see below) exert some limits on this. In edge, screw, or mixed dislocations the line is, respectively, perpendicular, parallel, or oblique to the Burgers vector. An edge dislocation has an extra half-plane of atoms, of thickness b, squeezed into one half of the crystal, terminating at the line and usually taken to be perpendicular to **b** (Fig. 5). A screw dislocation forms the axis of a helical structure of crystal planes.
2. The Burgers vector is an invariant. Dislocation lines can link up to form networks, not necessarily confined to slip planes, but at the triple nodes where they meet the relation

$$\mathbf{b}_1 + \mathbf{b}_2 = \mathbf{b}_3 \tag{52}$$

must be met. Links in slip planes can sometimes act as Frank-Read sources,

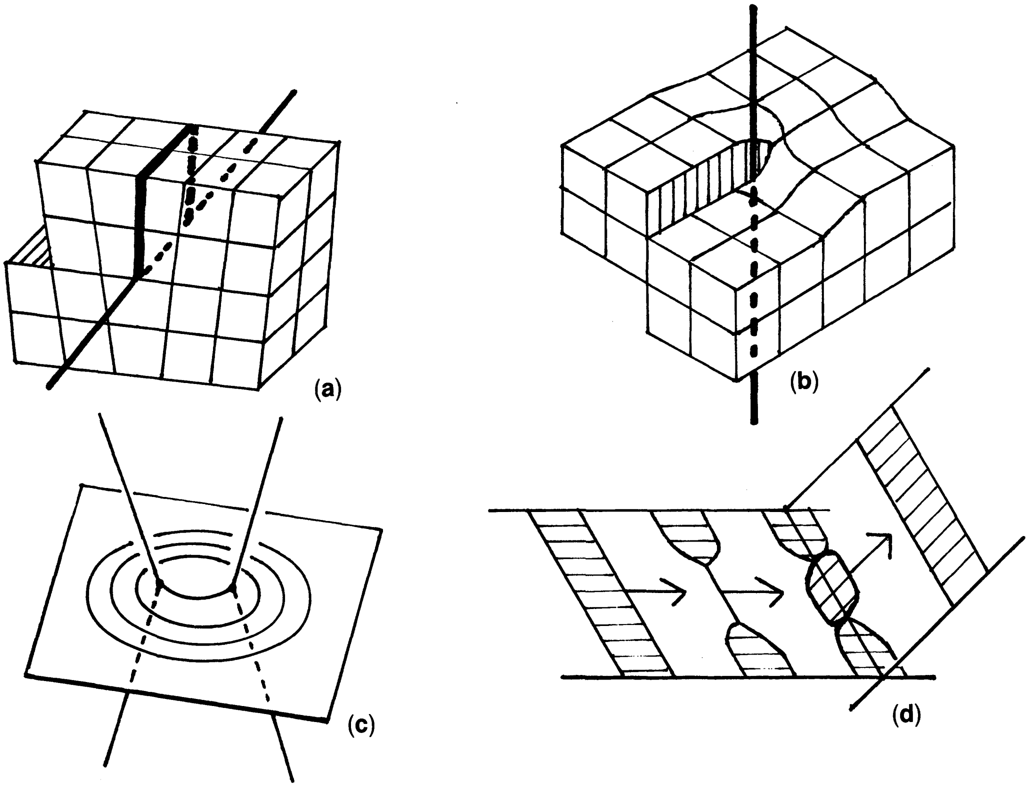

FIG. 5. Dislocations: (a) edge and (b) screw, showing slip steps cross-hatched; (c) rings expanding from a Frank-Read source in a network; (d) stages in the cross-slip of a dissociated screw dislocation, showing stacking fault cross-hatched.

able to emit many successive concentric rings of dislocation lines, so producing intense slip bands (Fig. 5).

3. Two dislocation lines may unite to form a third, and a dislocation of large Burgers vector may dissociate into two (or more) others. In all such dislocation reactions Eq. (52) is obeyed.
4. The displacement b produced by the movement of a dislocation enables an applied shear stress σ, acting on the glide plane in the direction of b, to do work, equivalent to the dislocation being driven by a force on unit length of its line,

$$F = \sigma b. \tag{53}$$

5. Soft crystals, e.g., pure Al and graphite, have mobile slip dislocations, movable by forces F equivalent to $\sigma \simeq 10^{-5}\mu$. Mobile dislocations are usually wide; i.e., the transition zone between the unslipped and slipped areas of the glide plane is several times larger than b. In this gradual transition the local atomic forces are approximately self-balancing, so that the net Peierls-Nabarro stress, needed to overcome them and thus move the dislocation, is almost zero, a most remarkable physical property. By contrast, in crystals with directed bonds and hence with large resistance to elastic shear, e.g., diamond and Si, the dislocations are less than $2b$ wide and the Peierls-Nabarro stress approaches σ_s. Crystals are thus intrinsically soft or hard, according to their (low-temperature) Peierls-Nabarro stresses. This is also approximately the division into intrinsically ductile and brittle crystals.
6. A screw dislocation can in principle glide on any dividing surface that contains its Burgers vector, and thus can cross-slip between different glide planes that intersect along its line. An edge can glide only in the particular surface that contains

both its line and Burgers vector. This surface can, however, be changed by climb, in which lattice vacancies are absorbed or emitted at small steps, i.e., jogs, along the line of the edge. Cross-slip and climb are usually thermally activated processes and so occur mainly at high temperatures.

7. A dislocation is perfect when its **b** is a perfect translational vector of the crystal. Imperfect or partial dislocations leave monolayers of stacking fault or twin configuration in their wake. They usually exist as pairs, the lines of which enclose ribbons of stacking faults (Fig. 5). A perfect dislocation will, wherever possible, dissociate into such a pair, which can be regarded as an extended core of the dislocation. Dislocation reactions [Eq. (52)] sometimes produce cores that extend out of the glide plane; the resultant dislocation "lock" is then immobile, at low temperature.
8. A dislocation is a center of elastic strain. A concentric circle around a screw dislocation is deformed into a helix of pitch b. In an isotropic medium this deformation is evenly spread around the circle, giving an elastic strain $b/2\pi r$ and strain energy density $\frac{1}{2}\mu(b/2\pi r)^2$ at a radius r. Integrated throughout a cylinder of outer radius r_1 and length l, the total elastic energy is

$$U = (\mu b^2 l/4\pi)\ln(r_1/r_0), \tag{54}$$

where the elastic field is cut off at an inner radius r_0 ($\simeq b$), inside which an atomistic estimate must be made. This core energy is usually much smaller than the elastic energy outside r_0. When its right-hand side is divided by $1 - \nu$, Eq. (54) applies also to edge dislocations.

9. Many effects result from this. Because $U \propto b^2$, a dislocation with large b tends to dissociate into two or more smaller ones [subject to Eq. (52)] A free surface, because it defines a cutoff value of r_1, is attractive to a dislocation. Similarly, two dislocations with opposite Burgers vectors attract, because their strain fields mutually cancel at large distances. To minimize l a dislocation link tends to take the shortest distance between two points, like a one-dimensional soap "film," with a line tension of order

$$\tau \simeq \tfrac{1}{2}\mu b^2. \tag{55}$$

10. The inserted half-plane provides a dipolar component in the strain field of an edge dislocation, with compression on the half-plane side and tension on the other. As a result, the dislocation is particularly attractive to oversized or undersized impurity atoms, which can migrate to it and pin it, so producing various aging, precipitation, and high upper yield-point effects, important, for example, in structural steel containing small amounts of carbon and nitrogen.
11. This dipolar field also causes parallel edges, with the same b, to align above one another, in polygonization walls (Fig. 6), which form the boundaries between slightly misoriented mosaic crystallites in the crystal.
12. Their long-range strain fields enable dislocations to be seen in the electron microscope from variations in Bragg reflectivity.

For additional discussion of dislocations see Bilby *et al.* (1985), Cottrell (1953, 1964), or Nabarro (1967).

2.4 Yield Strength

A crystal yields plastically when its dislocations begin to glide and multiply. In soft crystals the yield strength is the stress σ to drive dislocations through or around various obstacles. Many alloys are deliberately hardened by a fine dispersion of precipitated particles, of average spacing l. Orowan's formula,

$$\sigma \simeq \mu b/l, \tag{56}$$

for the stress to bypass such obstacles by bending and squeezing a dislocation between them (Fig. 7) shows how the yield strength increases with fineness of dispersion.

A grain boundary is a strong obstacle but can be attacked by a pileup of many dislocations, approaching it along the same glide plane, the members of which push their leader into the obstacle. An analysis of this, similar to crack theory (see below), leads to the Hall-Petch formula for the yield strength of a

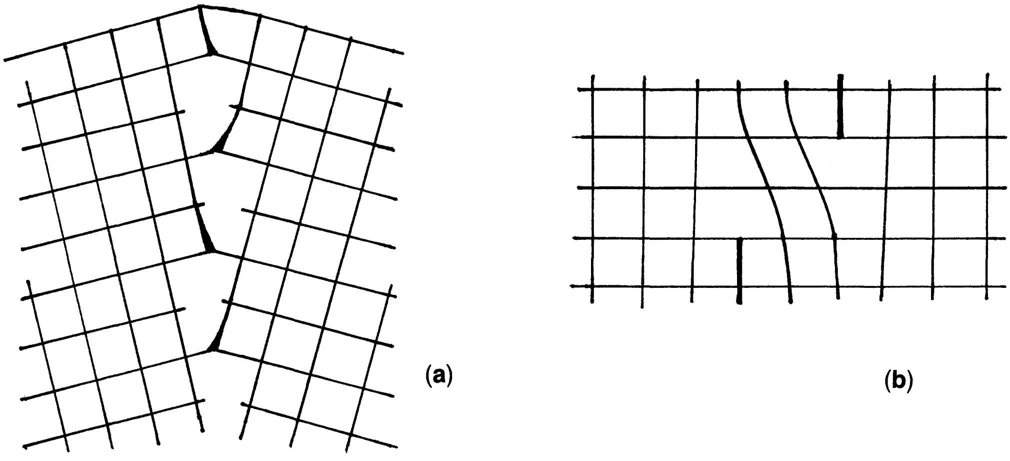

FIG. 6. Edge dislocations: (a) polygonization wall of like dislocations, (b) dipole pair of unlike ones.

polycrystal of mean grain diameter d,

$$\sigma = \sigma_i + k_y d^{-1/2}, \tag{57}$$

where σ_i is a "friction" stress due to obstacles within the crystals and k_y represents the grain-boundary obstruction. Because of orientational differences between neighboring grains, plastic deformation in a polycrystal is "turbulent," involving at least five different orientations of glide planes and directions, in each crystal, to preserve material continuity at the grain boundaries.

For additional discussion of yield strength see Cottrell (1953, 1964), Gordon (1976), Kelly and MacMillan (1986), Nabarro (1967), or Thomason (1990).

2.5 Work Hardening and Fatigue

Dislocations moving in opposite directions or in different glide systems can obstruct one another. The multiplication of dislocation numbers by intense cold working, through this effect, strongly increases the yield strength, up to about $10^{-2}\mu$ after plastic strains of order unity. The oscillatory deformation of fatigue cycling, with a typical plastic strain of 10^{-4}, reversed and repeated some 10^6 times, takes the material on, beyond ini-

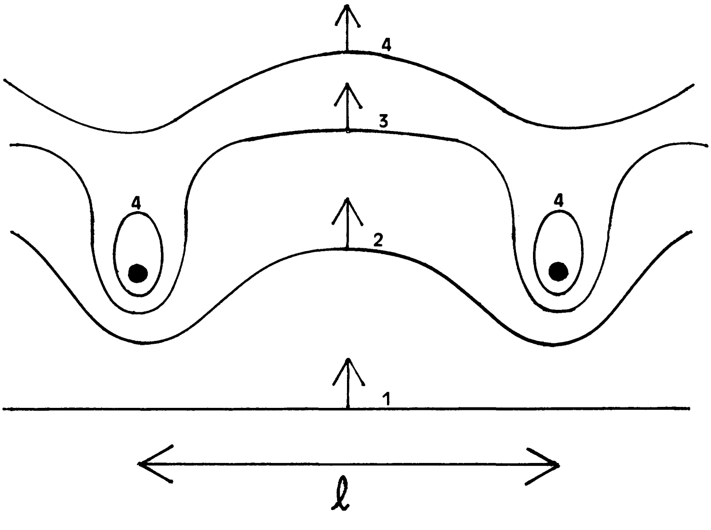

FIG. 7. Stages 1–4 in the bypassing of obstacles by a moving dislocation line.

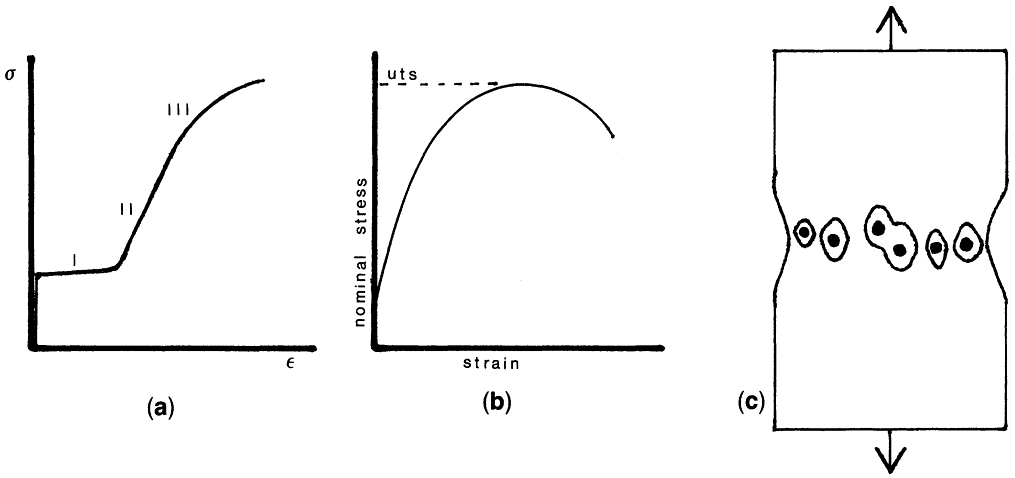

FIG. 8. Work hardening and tensile fracture: (a) typical stress-strain curve of a single crystal; (b) polycrystalline tensile curve, with ultimate tensile stress (UTS) at a nominal stress given by the maximum load divided by the original cross-sectional area; (c) tensile necking and internal dimpling fracture.

tial work hardening, to a further stage where localized work softening begins, starting with the formation of persistent slip bands, i.e., work-softened bands on which the plastic deformation concentrates, and ending with fatigue fracture.

These complex processes involve huge numbers of dislocations. They have been unraveled by electron microscopy and then interpreted by dislocation theory. A stretched crystal of, e.g., pure copper usually work hardens in three stages:

1. easy glide in which slip is confined mainly to one set of primary, parallel planes, with little hardening;
2. turbulent glide in which other, secondary, slip systems become active, with many dislocations moving on intersecting paths, leading to traffic jams, dislocation locks, forests of intersecting lines, dipole pairs, and dislocation walls across the primary slip planes (Fig. 6), and thus rapid work hardening;
3. a stage where the onset of work softening or dynamic recovery, brought about by cross-slip and consequential mutual annihilation of opposite screw dislocations at high stresses, leads to a reduced rate of work hardening.

In tensile deformation, this culminates in tensile instability beyond which the hardening is no longer able to compensate for the loss of cross-sectional area due to the plastic extension (at constant volume). Beyond this ultimate tensile strength all further plastic strain concentrates in a tensile neck that, ideally, narrows all the way down to a point or chisel-edge rupture. In practice failure usually occurs by internal necking in which plastic holes form around foreign particles, expand, and coalesce, so producing a characteristic dimpled or "fibrous" ductile fracture (Fig. 8).

Strain concentration is sometimes opposed by plastic constraint. A narrow ductile ligament at the center of a deeply notched tensile bar (Fig. 9), for example, can yield only on slip lines extending to distant sides or curving around to the free sides of the grooves.

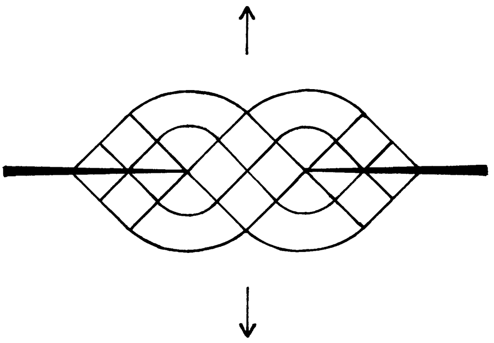

FIG. 9. Curved slip lines for constrained plastic yielding in a narrow ligament between deep notches.

The extra length of these slip paths in both cases thus, in effect, raises the apparent width of the ligament. This is usually expressed instead as an increase in effective yield strength, by a plastic constraint factor, approximately 3 in the example of Fig. 9. The very severe plastic constraint experienced at the tip of a sharp crack, in an elastic-plastic solid, is usually described in terms of a local concentration of hydrostatic tension in this region.

The accumulation of huge plastic strains, by many small oscillations, can take constant-strain-amplitude fatigue stressing well into stage **3**. If the stress amplitude falls below this level, the accumulated work hardening eventually suppresses the plastic component of the strain cycle. The material is then below its fatigue limit and has an indefinitely long fatigue life. But, in stage **3**, work softening can cause the plastic strain to concentrate on persistent slip bands; and on these, at the surface, cracks eventually form. The fatigue life, i.e., number N of cycles to fracture, is given by the Coffin-Manson rule,

$$N^{\alpha}\epsilon = \text{const}, \tag{58}$$

where ϵ is the plastic strain amplitude and $\alpha \simeq 0.5$.

The work softening in the persistent slip bands is brought about by the alignment of the dislocations, there, into walls, roughly perpendicular to the slip direction, leaving long, clear, soft regions in between. Locally high stresses, produced where the bands meet the free surface, then open up fissures that gradually spread back, down the bands, to become fatigue cracks (Brown, 1991). Chemical action from the environment can greatly speed up this crack growth, as in corrosion fatigue; otherwise the cracks go through an intermediate, dormant phase of very slow growth before finally spreading rapidly through a process of alternate plastic stretching and folding at the crack tip.

An obvious practical consequence is that fatigue resistance can be increased by hardening the surface, e.g., through alloying or hammering (peening), to suppress plastic deformation. This is effective when it does not introduce stress-concentrating irregularities into the surface.

For further treatment of work hardening and fatigue, see Bilby *et al.* (1985), Cottrell (1953), or Nabarro (1967).

2.6 Creep

High temperatures promote cross-slip and climb, suppressing work hardening and so allowing plastic flow, or creep, to occur under prolonged constant stress, usually according to a relation such as

$$\epsilon(t) = \epsilon_0 + \beta t^{1/3} + \kappa t + \gamma t^3, \tag{59}$$

where the terms involving B, κ, and γ respectively represent transient, steady-state, and accelerating stages of creep. At high temperatures, climb enables opposite edge dislocations to converge and mutually annihilate, and those of the same sign to polygonize, so that the plastically distorted crystal structure clarifies into a mosaic of soft crystallites. The approach to this state occurs during the transient stage of creep, and the self-maintenance of the state then ensures a long steady-state stage, terminated only by incipient fracture processes, which herald the onset of the accelerating stage. Most creep fractures are brought about by viscous sliding along grain boundaries, which opens up cracks at boundary junctions and also at foreign particles. At very high temperatures, vacancies can coalesce at boundary particles to form holes, leading to failure by creep cavitation. A drastic practical method of avoiding such creep fractures is to eliminate grain boundaries, as in the single-crystal turbine blades used in advanced jet engines.

For more on creep, see Cottrell (1953, 1964), Nabarro (1967), or Ward (1983).

2.7 Fracture Mechanics

Structural steel, sharply notched, usually fractures in a macroscopically brittle manner below about 250 K, and in a ductile one above about 300 K (Fig. 10). The practical importance of this ductile-brittle transition (DBT) has led to the engineering discipline of fracture mechanics. A microscopically brittle crack grows according to Griffith's theory (1920). As it spreads, its surface area increases and the surface energy for this, 2γ per unit area of crack, must come from work done by the applied stress σ. A crack of length $2c$ thus cannot grow unless the Griffith stress

$$\sigma = (2\gamma E/\pi c)^{1/2} \tag{60}$$

is exceeded. Window glass typically breaks at $\sigma \simeq 10^{-3}E$ as a result of scratches about 1

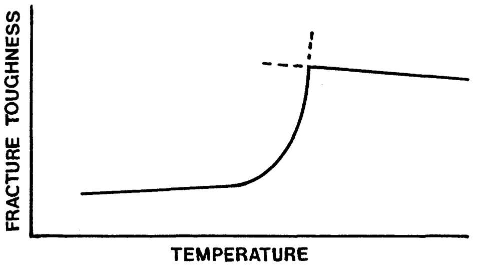

FIG. 10. Typical toughness–temperature transition for structural steel, showing the lower shelf and transition regions, where the fracture is macroscopically brittle (cumulative), and the upper shell region, macroscopically ductile (noncumulative). The fracture toughness in the upper shelf region is usually a nominal value, e.g., that at which a small, standard amount of ductile crack extension occurs from the tip of a preexisting sharp notch.

μm deep. The corresponding weakening of cold steel requires cracks of order 10^{-2} m, because its fracture, while macroscopically brittle, involves local plastic work along the fracture path. Griffith's equation still applies if γ is interpreted, not purely as surface energy γ_s, but as including this plastic work of fracture γ_p, i.e., $\gamma = \gamma_s + \gamma_p$. It then becomes the Griffith-Orowan-Irwin equation, usually written in the form

$$\sigma = K/(\pi c)^{1/2}, \tag{61}$$

where K is the fracture toughness of the material.

This theory, linear elastic fracture mechanics (LEFM), fails when the plastically deforming zones spread far from the crack tip. Various nonlinear modifications are available (*J* integral, crack-opening displacement, and strip-yielding methods). The theory is used in engineering design, but two limitations should be noted:

1. The equations assume a given microscopic fracture mode. The theory does not identify this, apart from providing a clue from the measured value of γ. The value of γ can vary greatly, for example, because of nuclear irradiation hardening, stress-corrosion cracking, and grain-boundary embrittlement by segregated impurities, and especially with temperature.
2. The equations are applicable only to macroscopically brittle fractures in which the microscopic processes of failure are localized in a small process zone, near the plane of fracture, and their crack-opening effect continues cumulatively as this zone moves forward with the advancing crack. By contrast, in more ductile material the process zones, which spread above and below the crack tip, are noncumulative, i.e., additional plastic deformation has to be pushed in, to spread them further, for each increment of crack growth. Eventually they demand a driving stress that reaches the general yield stress. This mode of stable crack growth under rising stress does not satisfy Eq. (60). The brittle-ductile transition is thus one from a cumulative to a noncumulative mode of failure, from a regime where fracture mechanics is valid to one where it is not. Fracture mechanics is further treated in Cottrell (1964), Gordon (1976), Kelly and MacMillan (1986), Knott (1973), Patterson (1978), and Thomason (1990).

2.8 Plasticity

For problems of metal working a formal theory of plasticity has been developed. The complexity and nonlinearity of the underlying processes have necessitated many simplifying assumptions. The theory assumes a model material with the following characteristics:

1. It is homogeneous, isotropic, continuous, and elastically rigid.
2. Plastic deformation occurs when the applied stresses satisfy a certain yield criterion. There is no volume change, no work hardening, no strain-rate or creep effects, and no reversal on unloading.
3. Yielding occurs when the stress state deviates from hydrostatic by an amount usually defined by the von Mises yield criterion,

$$(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2 = \text{const}, \tag{62}$$

where σ_1, σ_2, and σ_3 are the principal stresses.
4. The theory is usually developed for two-dimensional plane strain states by constructing slip-line fields, i.e., patterns of dividing surfaces along which the yield criterion is satisfied. These are not crys-

tallographic slip planes and can follow curved paths, this curvature being related to variations of hydrostatic stress. For example, in Fig. 9 the hydrostatic stress, zero at the notch faces, rises along the curving slip lines to a maximum in the central ligament. Its value there enhances the tensile yield stress by the plastic constraint factor.

The neglect of elastic deformation limits the utility of this rigid-plastic theory. For example, in the process zone of a crack tip, it is usually essential to take account of both elastic and plastic deformation (as in strip-yielding theories). The assumptions **2**, while good for metals, do not usually apply to materials such as rocks and soils, where the yield stress varies with hydrostatic pressure. A separate theory, soil mechanics, is available for these.

Additional discussion of plasticity may be found in Cottrell (1964), Hill (1950), Knott (1973), or Thomason (1990).

GLOSSARY

Burgers Vector: The vector that gives the slip displacement associated with a dislocation.

Climb: The motion of an edge dislocation out of its slip surface, produced by absorption or emission of vacancies.

Cross-Slip: Change in the plane of slip of a screw dislocation.

Crystallite: A small subcrystal, within a crystal, slightly misoriented relative to its neighbors.

Dimples: Plastic holes that form around foreign inclusions in a ductile fracture.

Dislocations: Line defects in a crystal that mark the boundaries of differently slipped areas of internal surfaces of slip. Dislocations can be of the following types: edge, screw, or mixed, according as their lines are respectively perpendicular, parallel, or oblique to their Burgers vectors; and perfect, imperfect, or partial, according to whether their Burgers vectors are or are not perfect lattice vectors of the crystal.

Dispersion Relation: The relation between frequency and wavelength in a vibrating system.

Fibrous Fracture: The appearance of a ductile fracture brought about by the coalescence of dimples.

Frank-Read Source: A link in a network of dislocation lines, which is geometrically able to generate dislocations.

Grain and Grain Boundary: A small, usually irregular shaped, crystal in a polycrystalline aggregate. Grain boundaries are the noncrystalline interfaces where grains meet.

Lock: A dislocation that is structurally incapable of gliding.

Necking: A local narrowing of the cross section, as a result of plastic elongation.

Normal Modes: Representation of the vibrations of a many-body system as a superposition of harmonic oscillations in each of which all elements of the system vibrate with the same frequency.

Persistent Slip Bands: Bands of locally intense slip, due to work softening and leading to fatigue failure. They are formed during repeated cyclic straining.

Phonon: A quantum of mechanical vibration.

Plastic Constraint: Where plastic deformation is prevented from occurring freely, due to local geometric restriction, as at the tip of a crack.

Polycrystal: A coherent aggregate of crystal grains.

Polygonization: The grouping of parallel lines of like edge dislocations, one above another, to form a dislocation wall, i.e., a narrow boundary between two slightly misoriented crystallites.

Process Zone: The zone ahead of a crack tip in which inelastic processes occur and prepare the material for fracture.

Slip Line: A line of maximum shear stress, used in a continuum representation of plastic deformation.

Stacking Fault: An irregularity in the sequence of stacking of successive atomic layers in a crystal.

Strip-Yielding Model: A method for analyzing elastic-plastic deformation in the process zone ahead of a crack tip.

Ultimate Tensile Stress: The maximum load supported by a solid, strained plastically in tension, divided by the original cross-section area.

Vacancy: An unoccupied atomic site in a crystal.

Yield Criterion: Condition for the onset of plastic yielding in a solid subjected to a combination of principal stresses.

Further Reading

Bilby, B. A., Miller, K. J., Willis, J. R. (Eds.) (1985), *Fundamentals of Deformation and Fracture*, Cambridge: Cambridge University Press.

Brown, L. M. (1991), *Metall. Trans. A* **22A**, 1693–1708.

Cottrell, A. H. (1953), *Dislocations and Plastic Flow in Crystals*, Oxford: Clarendon.

Cottrell, A. H. (1964), *The Mechanical Properties of Matter*, New York: Wiley.

Cottrell, A. H. (1989), *Mater. Sci. Tech.* **5**, 1165–1167.

Gordon, J. E. (1976), *The New Science of Strong Materials*, 2nd ed., London: Penguin.

Griffith, A. A. (1920), *Philos. Trans. R. Soc. London A* **221**, 163–198.

Hill, R. (1950), *The Mathematical Theory of Plasticity*, Oxford: Clarendon.

Kelly, A., MacMillan, N. H. (1986), *Strong Solids*, 3rd ed., Oxford: Clarendon.

Kittel, C. (1986), *Introduction to Solid State Physics*, 6th ed., New York: Wiley.

Knott, J. F. (1973), *Fundamentals of Fracture Mechanics*, London: Butterworths.

Kolsky, H. (1953), *Stress Waves in Solids*, Oxford: Clarendon.

Nabarro, F. R. N. (1967), *Theory of Crystal Dislocations*, Oxford: Clarendon.

Patterson, M. S. (1978), *Experimental Rock Deformation—The Brittle Field*, Berlin: Springer.

Pugh, S. F. (1954), *Philos. Mag.* **45**, 823–843.

Thomason, P. F. (1990), *Ductile Fracture of Metals*, Oxford: Pergamon.

Walsh, J. M., Christian, R. M. (1955), *Phys. Rev.* **97**, 1544–1556.

Ward, I. M. (1983), *Mechanical Properties of Solid Polymers*, 2nd ed., New York: Wiley.

Zener, C. (1948), *Elasticity and Anelasticity of Metals*, Chicago: University of Chicago Press.

MECHANICAL VIBRATION AND DAMPING

Sabih I. Hayek, *Penn State University, University Park, Pennsylvania, U.S.A.*

	Introduction	561
1.	**Single Degree-of-Freedom Systems**	562
1.1	Simple Oscillator—Mass-Spring System	562
1.2	Simple Damped Oscillator	562
1.3	Measure of Viscous Damping	564
1.4	Measure of Coulomb (Dry) Damping	564
1.5	Measure of Hysteresis (Structural) Damping	564
1.6	Forced Vibration	565
1.7	Forced Vibration of a Damped Oscillator	567
1.8	Vibration Transmission and Isolation	568
1.9	*Q* Factor and Forced Response	568
1.10	Impedance and Admittance	569
1.11	Transient Response	569
1.12	Isolation of Equipment from Foundation Vibration	570
1.13	Equivalent Simple Oscillator Systems	570
2.	**Multiple Degree-of-Freedom Systems**	571
2.1	Two Degree-of-Freedom System	572
2.2	Forced Vibration of a Two-DOF System	573
2.3	Tuned Dynamic Vibration Absorber	574
2.4	Vibration of Multiple Degree-of-Freedom Systems	575
3.	**Vibration of Continuous One-Dimensional Systems**	575
3.1	Vibration of Stretched Strings	575
3.2	Longitudinal Vibration of Rods	577
3.3	Torsional Vibration of Rods	578
3.4	Flexural Vibration of Beams	579
4.	**Vibration of Continuous Two-Dimensional Systems**	580
4.1	Vibration of Stretched Membranes	580
4.2	Vibration of Plates	582
4.3	Vibration of Shells	582
5.	**Practical Structural Damping Applications**	583
5.1	Surface Damping Treatment	585
5.2	Constrained-Layer Damping	585
5.3	Vibration Isolation Mounts	587
5.3.1	Elastomeric Vibration Mounts	587
5.3.2	Metal Springs	587
5.3.3	Plastic Springs	588
5.3.4	Air Springs	588
	Glossary	588
	Works Cited	589
	Further Reading	590

INTRODUCTION

The subject of mechanical vibration encompasses the field of periodic motion of mechanical systems with masses and elastic stiffnesses. Vibration of such systems may result in destructive phenomena such as resonating bridges, buildings, airplanes, nuclear reactors, and machines, leading to structural fatigue or operational failures. Excessive vibrations of cutting tools and lathes impede quality control. Beneficial results of vibration include sound generation in loudspeakers and bowed musical instruments, soil compacting, ultrasonic welding and cleaning, and vibration-related tomography and diagnostics. In this article, the subject is introduced by discussion of simple single and multiple degree-of-freedom systems, and then more advanced work is introduced for continuous one- and two-dimensional vibrators.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. SINGLE DEGREE-OF-FREEDOM SYSTEMS

1.1 Simple Oscillator—Mass-Spring System

The simplest single degree-of-freedom system is the mass-spring oscillator. Consider a mass m attached to a linear spring of stiffness k [Fig. 1(a)]. At equilibrium, the spring is stretched by a length equal to $x_g = mg/k$. Using $x(t)$ to denote the time-dependent displacement from the equilibrium position and invoking Newton's second law gives

$$m\frac{d^2x}{dt^2} = mg - k(x + x_g) + F(t),$$
$$m\frac{d^2x}{dt^2} + kx = F(t). \tag{1}$$

The initial conditions on the displacement and the velocity are $x(0) = X_0$ and $dx(0)/dt = \dot{x}(0) = V_0$, respectively. For force-free vibration, $F(t) = 0$, and the solution of Eq. (1) is given by

$$x(t) = A\cos 2\pi f_0 t + B\sin 2\pi f_0 t, \tag{2}$$

where $f_0 = (m/k)^{1/2}$ is the natural (resonance) frequency. The displacement $x(t)$ in Eq. (2) represents periodic motion, whose period is $\tau = 1/f_0$. For the initial conditions given, the unknown constants are found to be $A = X_0$ and $B = V_0/2\pi f_0$. The displacement $x(t)$ can be written in the form

$$x(t) = X\sin(2\pi f_0 t + \phi), \tag{3}$$

where $X = [A^2 + B^2]^{1/2}$ is the amplitude and $\tan\phi = A/B$, ϕ being the phase angle. A plot of the displacement $x(t)$ in Eq. (3), shown in Fig. 2(a), exhibits a periodic motion with the maximum value being the amplitude X. The instantaneous velocity and acceleration dx/dt and d^2x/dt^2 have 90° and 180° phase angles ahead of the displacement $x(t)$.

Another representation, using complex variables, may be employed to obtain similar results. This can be done by assuming a complex displacement $\tilde{x}$:

$$\tilde{x}(t) = Xe^{i(2\pi f_0 t + \phi)}. \tag{4}$$

Thus, using the imaginary part of $\tilde{x}(t)$, one can show that

$$x(t) = \operatorname{Im}\tilde{x}(t) = X\sin(2\pi f_0 t + \phi).$$

1.2 Simple Damped Oscillator

If the system has energy absorption through viscosity, such as shown in Fig. 1(b), the viscous damping force is proportional to the velocity $\dot{x}(t)$, such that the equation governing the motion of the mass becomes

$$m\frac{d^2x}{dt^2} + c_0\frac{dx}{dt} + kx = F(t). \tag{5}$$

For free vibration, $F(t) = 0$, the solution of Eq. (5) is given by

$$x(t) = e^{-2\pi f_0 \zeta t}[A\cos(2\pi\sqrt{1 - \zeta^2} f_0 t) + B\sin(2\pi\sqrt{1 - \zeta^2} f_0 t)], \tag{6}$$

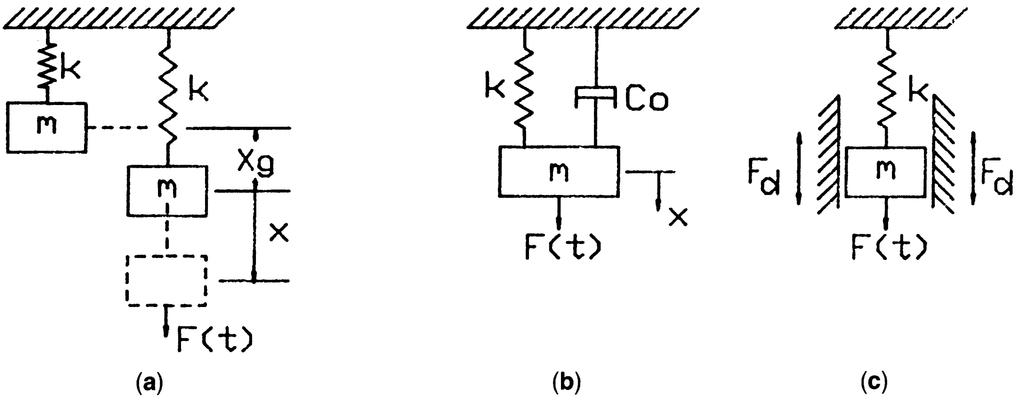

FIG. 1. Simple oscillator: (a) undamped and (b), (c) damped.

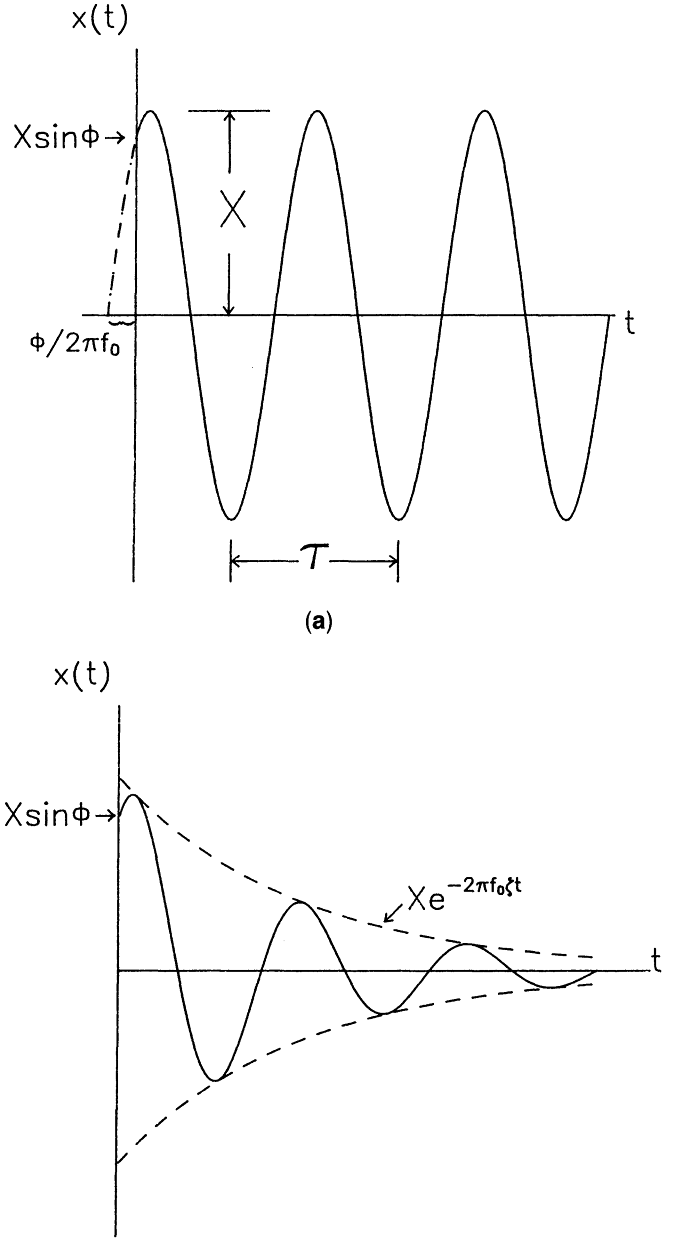

(a)

(b)

FIG. 2. Response of simple oscillator vs time: (a) undamped and (b) damped.

where $\zeta = c_0/2\sqrt{km}$ is the *damping ratio*. Here again, for the given initial conditions, the constants A and B are given by $A = X_0$ and $B = (\zeta X_0 + V_0/2\pi f_0)/\sqrt{1 - \zeta^2}$, so that the displacement can be written as

$$x(t) = Xe^{-2\pi f_0 \zeta t} \sin(2\pi\sqrt{1 - \zeta^2} f_0 t + \phi). \qquad (7)$$

The displacement $x(t)$ is represented by a decaying oscillatory motion [see Fig. 2(b)] whose damped frequency is $f_d = \sqrt{1 - \zeta^2} f_0$. The motion $x(t)$ is decaying at an exponential rate proportional to the natural frequency f_0 and the damping ratio ζ. The maximum amplitude is X, which decays at the same exponential rate [see Fig. 2(b)]. When $\zeta = 1$, then the damping coefficient $c_0 = 2\sqrt{km} = c_c$ is called the *critical damping*. For $\zeta \geq 1$, the response $x(t)$ becomes a purely decaying nonoscillating motion.

1.3 Measure of Viscous Damping

A measure of the damping is the damping ratio ζ. The damping ratio is of the order of 10^{-4} for metals to 10^{-2} for rubber.

Another measure of damping that can be obtained from the direct-displacement response measurement of $x(t)$ is the *logarithmic decrement* δ. If one calls the first peak x_0, and successive peaks $x_1, x_2, \ldots$, one can show that

$$x_n = x_0 e^{-n\delta} \quad \text{or} \quad \delta = (1/n)\ln(x_0/x_n)$$

where

$$\delta = 2\pi\zeta/\sqrt{1-\zeta^2} \approx 2\pi\zeta \quad \text{for} \quad \zeta \ll 1. \tag{8}$$

Another measure of the damping is the energy dissipated ΔU over a complete cycle as a fraction of the maximum potential energy, $U_{\max}$. For a harmonic motion, $x(t) = X \sin 2\pi f_0 t$, the energy dissipated by the viscous damper, ΔU per unit cycle, can be calculated as

$$\begin{aligned}\Delta U &= \int_0^\tau c_0\left(\frac{dx}{dt}\right)^2 dt \\ &= 2\pi^2 f_0 c_0 X^2 = 2\pi\zeta k X^2.\end{aligned} \tag{9}$$

The maximum potential energy per cycle $U_{\max}$ is given by

$$\begin{aligned}U_{\max} &= [\tfrac{1}{2}kx^2]_{\max} = \tfrac{1}{2}kX^2 \\ \Delta U/U_{\max} &= 4\pi\zeta \approx 2\delta.\end{aligned} \tag{10}$$

A comprehensive discussion of types of damping materials and their mechanical properties is given by Lazan (1968), Snowdon (1968), and Jones (1967).

1.4 Measure of Coulomb (Dry) Damping

Coulomb damping is generated by the dry friction of a moving mass m over a rough surface (see Plunkett, 1981). The frictional force is always opposite to the direction of motion. The equation of motion for a mass with dry friction [Fig. 1(c)] and free of force $[F(t) = 0]$ is

$$m\frac{d^2x}{dt^2} = -kx + F_d \quad \text{for motion downward,}$$

$$m\frac{d^2x}{dt^2} = -kx - F_d \quad \text{for motion upward.}$$

Since the frictional force is a function of the direction of the velocity and not the position of the mass, the solution must be calculated for each half period. Starting with an initial condition $x(0) = X_0$ and $\dot{x}(0) = 0$, the solution for the first half period is given by

$$x(t) = (X_0 - F_d/k)\cos 2\pi f_0 t + F_d/k, \quad 0 \le t \le \tau/2.$$

The solution for the second half period is given by

$$x(t) = (3F_d/k - X_0)\cos 2\pi f_0(t - \tau/2) - F_d/k, \quad \tau/2 \le t \le \tau.$$

The peaks at each half period are given by

$$|x(0)| = X_0, \quad |x(\tau/2)| = X_0 - 2F_d/k,$$
$$|x(\tau)| = X_0 - 4F_d/k.$$

The decay is clearly linear, i.e., a decay of $4F_d/k$ per one cycle. The total energy loss ΔU per cycle is calculated as $\Delta U = 4F_d X$. The equivalent viscous damping coefficient ζ_e may be evaluated by equating Eq. (8) to the energy loss for dry friction, i.e.,

$$\Delta U = 2\pi\zeta_e k X^2 = 4F_d X.$$

The equivalent viscous damping coefficient by dry friction becomes

$$\zeta_e = 2F_d/\pi k X. \tag{11}$$

1.5 Measure of Hysteresis (Structural) Damping

Hysteresis or structural damping represents the energy dissipated in the internal structure of a solid being cyclically stressed. The energy dissipated is proportional to the square of the stress in the solid or, equivalently, the square of the displacement. Let energy dissipation per unit cycle ΔU be given by

$$\Delta U = \pi k\eta X^2,$$

where η is a solid damping coefficient. Thus, the loss per quarter-cycle is equal to $0.25\pi k\eta X^2$. Starting with X_1 being the displacement at the start of the cycle with the potential energy being $\frac{1}{2}kX_1^2$, then the energy at the middle of the cycle, when $X = X_2$, is

$$\tfrac{1}{2}kX_1^2 - \tfrac{1}{4}\eta\pi kX_1^2 - \tfrac{1}{4}\eta\pi kX_2^2 = \tfrac{1}{2}kX_2^2,$$

so that rearranging the terms, one obtains the ratio

$$\frac{X_1}{X_2} = \sqrt{\frac{2 + \pi\eta}{2 - \pi\eta}}.$$

For the next half-cycle, with X_3 being the displacement at the end of the cycle, one obtains

$$\frac{X_2}{X_3} = \sqrt{\frac{2 + \pi\eta}{2 - \pi\eta}}.$$

Hence, the ratio of the displacements at the beginning and the end of the cycle is

$$\begin{aligned}\frac{X_1}{X_3} &= \frac{2 + \pi\eta}{2 - \pi\eta} = 1 + \frac{2\pi\eta}{2 - \pi\eta}\\ &\approx 1 + \pi\eta \quad \text{when} \quad \eta \ll 1.\end{aligned}$$

This gives for the equivalent logarithmic decrement δ_e for $\eta \ll 1$

$$\begin{aligned}\delta_e &= \ln\left(\frac{X_1}{X_3}\right) = \ln\left(1 + \frac{2\pi\eta}{2 - \pi\eta}\right)\\ &\cong \frac{2\pi\eta}{2 - \pi\eta} \cong \pi\eta.\end{aligned} \tag{12}$$

Thus, the equivalent damping ratio δ_e becomes

$$2\pi\zeta_e \approx \delta_e \approx \pi\eta,$$

or

$$\zeta_e \simeq \eta/2 \quad \text{and} \quad c_e \simeq \eta k/2\pi f_0. \tag{13}$$

Equation (13) indicates that the equivalent viscous damping of solid damping decreases with frequency. This equivalence can be employed to define a complex spring (stiffness) constant. Rewriting Eq. (5) as

$$\begin{aligned}m\frac{d^2x}{dt^2} + c_e\frac{dx}{dt} + kx\\ = m\frac{d^2x}{dt^2} + (i2\pi c_e f_0 + k)x = 0,\end{aligned}$$

$$m\frac{d^2x}{dt^2} + k\left(1 + i\frac{c_e 2\pi f_0}{k}\right)x = 0,$$

one can define a complex stiffness constant

$$\hat{k} = k\left(1 + ic_e\frac{2\pi f_0}{k}\right) = k(1 + i\eta). \tag{14}$$

The constant $\hat{k}$ is called the *complex stiffness*, and η is the *structural damping factor*. The complex stiffness can be used instead of the stiffness k to include the effect of structural damping. A practical guide to the choice of elastomers for damping treatment is given by Frye (1988), Lazan (1967), Alfrey (1948), Snowdon (1968) and Whitby (1954).

1.6 Forced Vibration

The response of the simple oscillator resulting from the force $F(t)$, with $F(t)$ being a periodic force, is called *forced vibration*. Starting with a system initially at equilibrium position and at rest, i.e., $X_0 = V_0 = 0$, with the periodic force described by

$$F(t) = F_0 \sin(2\pi f t) \tag{15}$$

where f is the frequency of the periodic force, then the response of the mass to the forced excitation is obtained as

$$x(t) = X_s\frac{f_0^2}{f_0^2 - f^2}\sin(2\pi f t), \tag{16}$$

where $X_s = F_0/k$ is the equivalent static displacement of the mass due to a static force F_0. The amplitude of the forced vibration is

$$X = X_s\frac{f_0^2}{f_0^2 - f^2} = \frac{X_s}{1 - r^2}, \tag{17}$$

where the frequency ratio $r = f/f_0$.

The amplification factor $|X/X_s|$ is plotted vs the frequency ratio r in Fig. 3. It can be

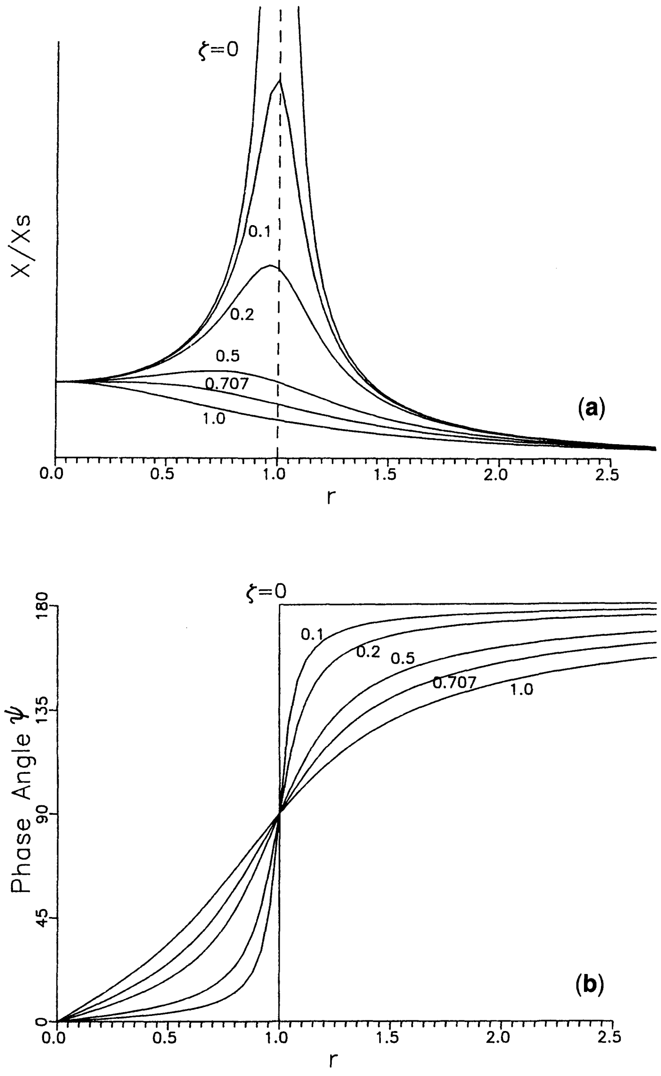

FIG. 3. Forced response of undamped and damped simple oscillator vs frequency: (a) magnitude and (b) phase angle for various damping coefficients.

seen that where $r \ll 1$, $|X/X_s| = 1$, which means that the dynamic response is almost equal to the static one. When $r \cong 1$, the response becomes unbounded. This is the state of vibration called *resonance*, i.e., when the frequency of the excitation force f coincides with the natural frequency f_0. For frequencies much higher than the natural frequency, $f \gg f_0$, the forced response diminishes to zero as $1/r^2$. Since the response would not become instantly unbound, it is instructive to obtain the complete solution for the oscillator when excited from rest by a harmonic force. For $x(0) = 0$, $\dot{x}(0) = 0$, and $F(t) = F_0 \sin 2\pi f t$, the solution is given by

$$x(t) = \frac{X_s}{1 - r^2} (\sin 2\pi f t - r \sin 2\pi f_0 t). \tag{18}$$

When the forcing frequency f is close to the natural frequency f_0, i.e., $r \approx 1$, the approximate solution for the response $x(t)$ becomes

$$x(t) \approx -\frac{X_s}{1 - r} \cos 2\pi f_0 t \sin[\pi(f_0 - f)t], \tag{19}$$

which represents an oscillating solution with time, with a frequency f having an envelope whose frequency is given by the difference frequency $f_0 - f$ as can be seen in Fig. 4. This phenomenon is called the *beating effect* with a beat frequency $f_0 - f$. The beating phenomenon also occurs when the excitation force has a time function with two close frequencies $f_1 \approx f_2$. The response exhibits an oscillating function in time with a frequency equal to the average of the two frequencies $(f_1 + f_2)/2$ and an envelope whose frequency is $\frac{1}{2}$ of the difference frequency, i.e., $(f_1 - f_2)/2$.

At resonance, i.e., when $f = f_0$, the limiting expression of the response in Eq. (19) becomes

$$x(t) = -\tfrac{1}{2}X_s(2\pi f_0 t)\cos 2\pi f_0 t, \qquad (20)$$

which shows that at resonance, the response increases linearly without limit with time.

1.7 Forced Vibration of a Damped Oscillator

If a time-harmonic force is applied to a simple oscillator with viscous damping, the steady-state response of a mass initially at rest is given by

$$x(t) = X\sin(2\pi ft - \psi), \qquad (21)$$

where the amplitude X and the phase angle ψ are respectively given by

$$X = X_s/[(1 - r^2)^2 + 4r^2\zeta^2]^{1/2}, \qquad (22)$$

and

$$\tan\psi = 2\zeta r/(1 - r^2). \qquad (23)$$

The magnification factor X/X_s is plotted in Fig. 3 vs the frequency ratio r as a function of the damping ratio ζ. It can be seen that the peak response occurs at

$$r_{\max} = \sqrt{1 - 2\zeta^2} < 1 \quad \text{for} \quad \zeta < \sqrt{2}, \qquad (24)$$

where the maximum value of the magnification factor is

$$X_{\max}/X_s = 1/2\zeta\sqrt{1 - \zeta^2}. \qquad (25)$$

If the damping ratio ζ is small when com-

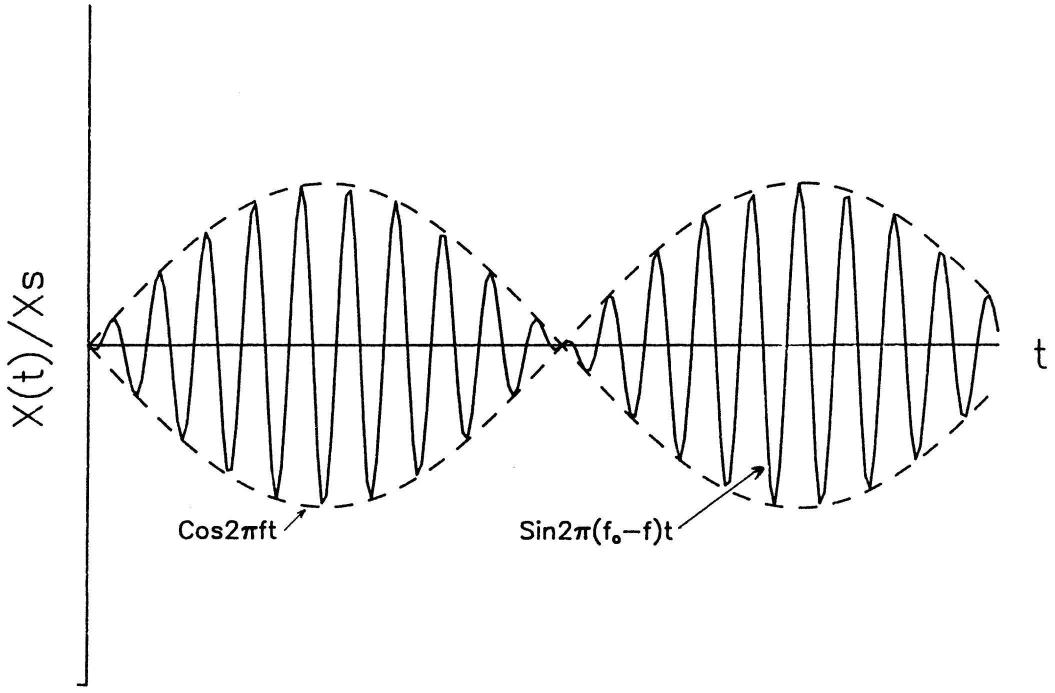

FIG. 4. Beating effect of forced response vs time.

pared to unity, the peak amplification factor $X_{\max}/X_s \approx 1/2\zeta$, i.e., the amplification factor is inversely proportional to the damping ratio ζ. The phase angle ψ increases from 0 to 90 deg as the frequency ratio increases from $r = 0$ to resonance at $r = 1$, and then to 180 deg as the ratio approaches $r \to \infty$ (see Fig. 3). The sharpness of the plot of the phase angle near resonance increases with decreasing damping ζ. Practical guides to the design of vibration isolation dampers are given by Racca (1988) and Crede (1951).

1.8 Vibration Transmission and Isolation

The total force transmitted to the supporting structure, F_t, as a result of the applied harmonic force F_0 can be calculated from the amplitude of the transmitted force resulting from the spring and dashpot:

$$F_t = [(kX)^2 + (2\pi cfX)^2]^{1/2}$$
$$= F_0\left[\frac{1 + 4\zeta^2 r^2}{(1 - r^2)^2 + 4\zeta^2 r^2}\right]^{1/2}. \tag{26}$$

The *force transmissibility* TR $= F_t/F_0$ is shown in Fig. 5 vs r. It can be seen that the transmissibility TR exceeds unity for $r < \sqrt{2}$ and decreases monotonically as $r \to \infty$. Thus, one can minimize excessive force transmission to the foundation by keeping the natural frequency f_0 below 70% of the forcing frequency f.

1.9 *Q* Factor and Forced Response

It was shown that the peak response is inversely proportional to the damping coefficient. If one measures the frequency ratios r_1 and r_2 from Fig. 6, corresponding to $X = X_{\max}/\sqrt{2}$, then the values of r at those points are

$$r_2^2 = 1 - 2\zeta^2 + 2\zeta\sqrt{1 - \zeta^2}$$
$$r_1^2 = 1 - 2\zeta^2 - 2\zeta\sqrt{1 - \zeta^2}. \tag{27}$$

If the damping is small, $\zeta \ll 1$, then those ratios can be approximated by

$$r_2^2 \approx 1 + 2\zeta \qquad r_1^2 \approx 1 - 2\zeta,$$

so that $r_2 - r_1 = 2\zeta = 1/Q$

The *quality factor Q* is defined as

$$Q = \frac{1}{2\zeta} = \frac{1}{r_2 - r_1} = \frac{f_0}{f_2 - f_1}. \tag{28}$$

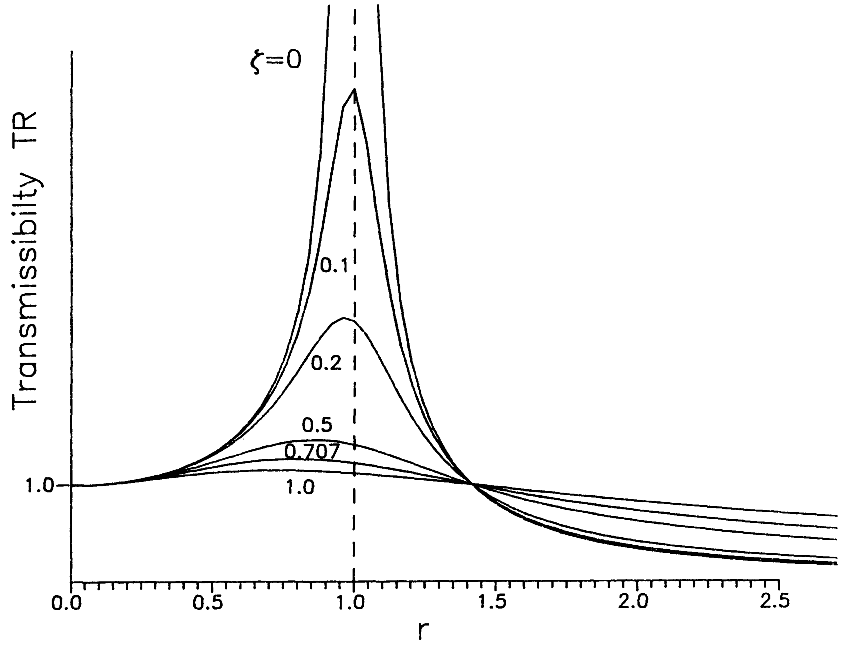

FIG. 5. Force transmissibility for a simple oscillator vs frequency for various damping coefficients.

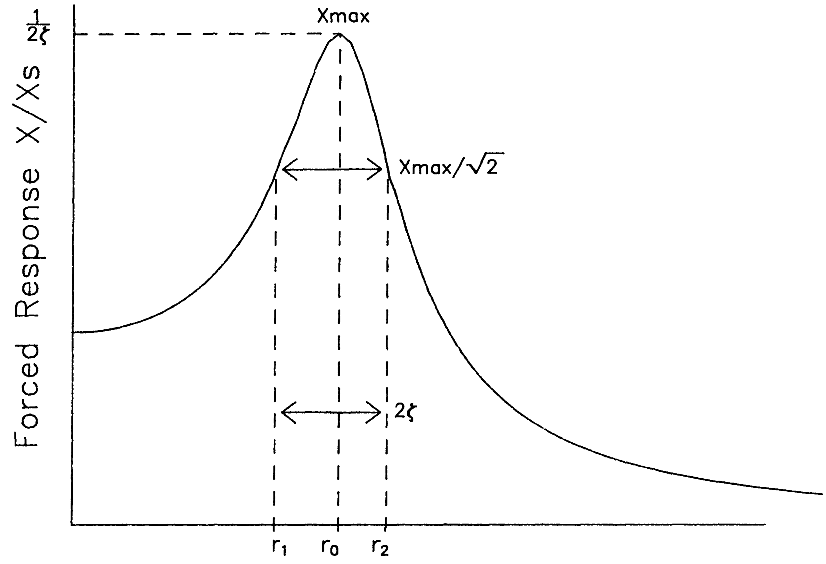

FIG. 6. Measure of the Q factor for a damped oscillator.

The damping ratio ζ or the quality factor Q can thus be extracted from the measurement of the frequency response of the forced vibration of a simple oscillator.

1.10 Impedance and Admittance

The forced response of a simple oscillator depends on its resistance to motion. The resistance is characterized by a quantity called *impedance,* denoted by Z, defined as the ratio of the applied force to the velocity:

$$Z = m\omega_0\left\{2\zeta + i\left(r - \frac{1}{r}\right)\right\}, \tag{29}$$

where $\omega_0 = 2\pi f_0$; the real part (active) represents a measure of energy loss by viscous absorption, while the imaginary part (reactive) has a positive portion resulting from the mass inertia $m2\pi f$ and a negative portion resulting from the elastic restoring force $k/2\pi f$. A plot of the real, imaginary, and absolute values of the impedance is shown in Fig. 7. In the plot of the absolute value of the impedance, there is a minimum at resonance $r = 1$, whose value is proportional to ζ. The inverse of the impedance, known as the *admittance, Y,* is defined by

$$Y = 1/Z, \tag{30}$$

which measures the velocity per unit applied force. In the plot of the absolute value of the admittance, there is a maximum at resonance, which is proportional to $1/\zeta$ or the Q factor (see Fig. 7).

1.11 Transient Response

In most situations, the oscillator is forced to vibration by nonharmonic forces. The time response of the oscillator is termed *transient response.* Let $F(t)$ represent the time-dependent force applied to the oscillator. For an oscillator that was initially at rest, $x(0) = 0$ and $\dot{x}(0) = 0$, the response $x(t)$, obtained as a solution of Eq. (1), can be calculated as

$$\begin{aligned} x(t) &= \frac{1}{2\pi m f_0}\int_0^t F(t-\eta)\sin(2\pi f_0\eta)d\eta \\ &= \frac{1}{2\pi m f_0}\int_0^t F(\eta)\sin 2\pi f_0(t-\eta)d\eta. \end{aligned} \tag{31}$$

For a force applied for a short time, whose impulse is $\hat{F}$,

$$\begin{aligned} F(t) &= \hat{F}/\epsilon, \qquad 0 \le t \le \epsilon, \\ &= 0, \qquad t > \epsilon; \\ x(t) &= \frac{\hat{F}}{2\pi^2 m f_0^2}\sin\pi f_0(2t-\epsilon)\frac{\sin\pi f_0\epsilon}{\epsilon}. \end{aligned} \tag{32}$$

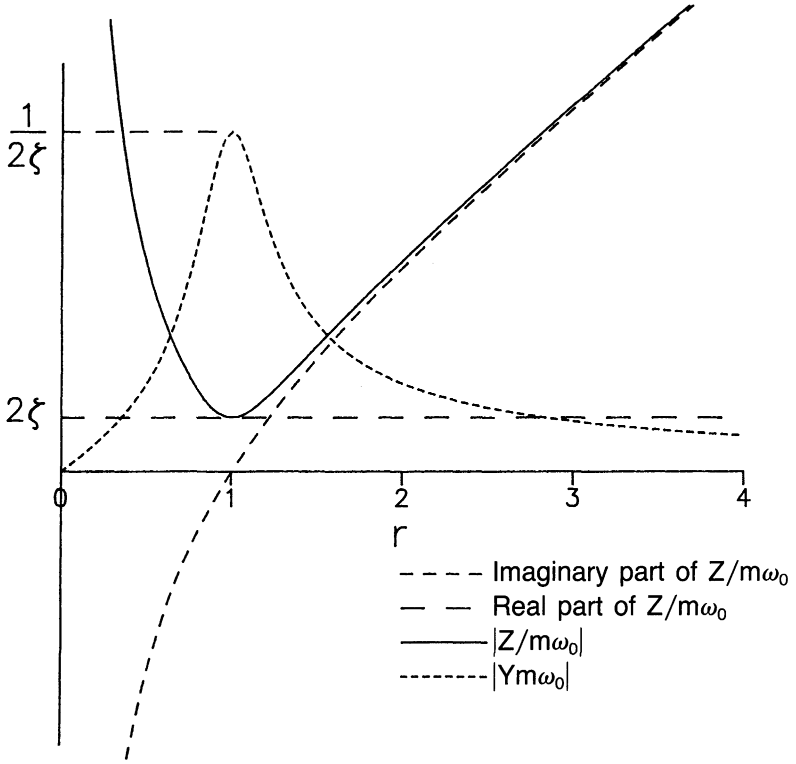

FIG. 7. Impedance and admittance of a damped simple oscillator.

For an impulsive force with $\epsilon \to 0$

$$x(t) = \hat{F}h(t), \tag{33}$$

where $h(t)$ is the impulse response due to a unit impulse, given by

$$h(t) = \frac{1}{2\pi m f_0} \sin 2\pi f_0 t. \tag{34}$$

A similar expression for the unit impulse response for a damped oscillator $h_d(t)$ is given by

$$h_d(t) = \frac{1}{2\pi m f_0\sqrt{1-\zeta^2}} e^{-2\pi f_0 \zeta t} \sin 2\pi f_0 t\sqrt{1-\zeta^2}. \tag{35}$$

The response for a force with an arbitrary time history $F(t)$ is then given by

$$x(t) = \int_0^t F(t-\eta)h(\eta)d\eta = \int_0^t F(\eta)h(t-\eta)d\eta. \tag{36}$$

1.12 Isolation of Equipment from Foundation Vibration

To shield equipment from the vibration transmitted through its supporting structure (foundation), a cushioning material is placed between the foundation and the equipment. If the foundation has a displacement $u = U_0 \sin\omega t$, then the displacement amplitude X_0 of equipment having a mass m and supported by cushioning material with stiffness k and viscous damping coefficient c is given by the same ratio as given by Eq. (26), i.e., $X_0/U_0 = F_t/F_0$. Proper choice of cushioning materials is outlined in the work of Hatae (1988) and standardized in ASTM standards (1964a, b).

1.13 Equivalent Simple Oscillator Systems

The simple oscillator consisting of a mass and a linear spring, discussed in the preceding sections, is an idealization of practical single degree-of-freedom systems (see Fig. 8). The equivalent stiffness for several such systems is given below.

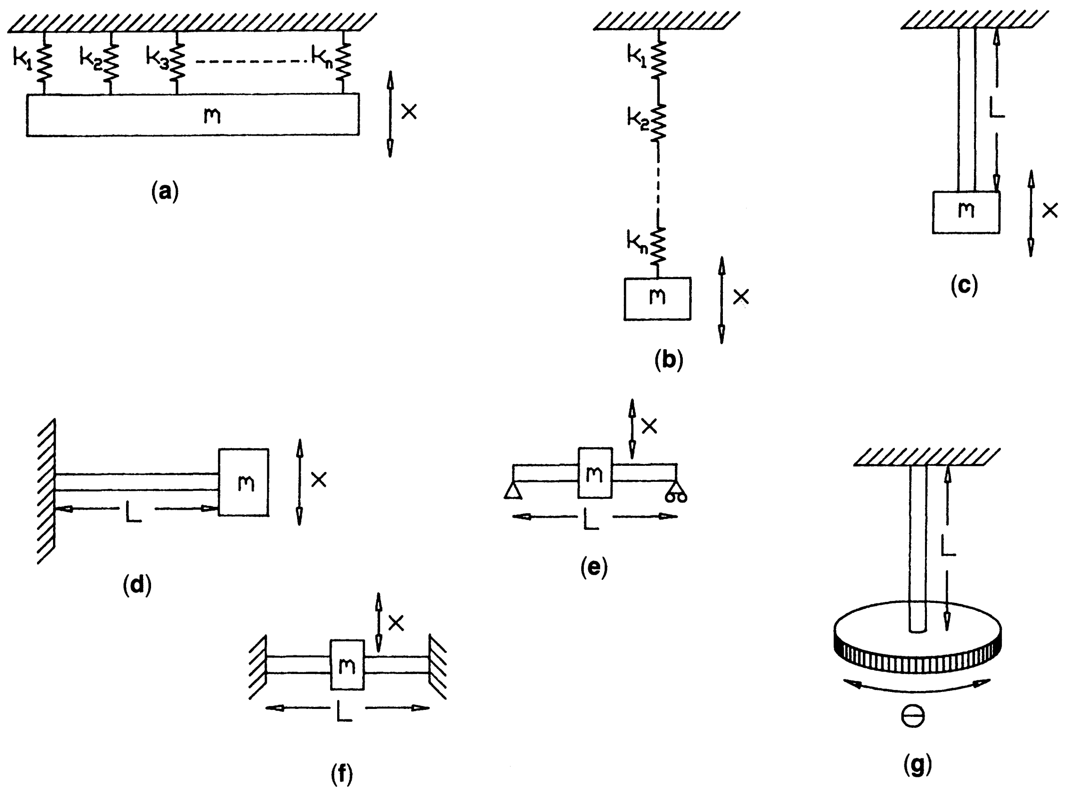

FIG. 8. Equivalent simple oscillator systems.

a. Springs in parallel, Fig. 8(a):

$$k_e = k_1 + k_2 + \ldots + k_n.$$

b. Springs in series, Fig. 8(b):

$$k_e = 1/(1/k_1 + 1/k_2 + 1/k_3 + \ldots + 1/k_n).$$

c. A bar with end-mass, Fig. 8(c):

$$k_e = EA/L,$$

where E = Young's modulus, A is the cross-sectional area, and L is the length.

d. A clamped-free beam, Fig. 8(d):

$$k_e = 3EI/L^3,$$

where I = moment of inertia of the cross section.

e. Simply supported beam, mass at center, Fig. 8(e):

$$k_e = 48EI/L^3.$$

f. Clamped-clamped beam, mass at center, Fig. 8(f):

$$k_e = 192EI/L^3.$$

g. Torsional vibration of bar with mass at end with the angle θ measuring the motion, Fig. 8(g):

$$J\ddot{\theta} + k_e\theta = 0,$$

where J is the rotational moment of inertia of the mass and $k_e = GJ/L$ where G is the shear modulus.

2. MULTIPLE DEGREE-OF-FREEDOM SYSTEMS

The foregoing discussion in Sec. 1 concentrated on the response of a simple oscillator, or a single degree of freedom (SOF). In this section, double degree-of-freedom (DOF) systems and multiple degree-of-freedom systems (MOF) will be discussed.

2.1 Two Degree-of-Freedom System

Consider a coupled two-oscillator system in series consisting of a spring k_1 and mass m_1 with another spring k_2 attached to another mass m_2 [Fig. 9(a)]. Letting x_1 and x_2 represent the displacements of m_1 and m_2, respectively, then the equations of motion are

$$m_1\ddot{x}_1 + (k_1 + k_2)x_1 - k_2x_2 = F_1,$$
$$m_2\ddot{x}_2 - k_2x_1 + k_2x_2 = F_2. \tag{37}$$

For a force-free vibration, $F_1 = F_2 = 0$, the displacement of a freely vibrating system can be assumed to be

$$x_1 = A_1 \sin(\omega t + \phi),$$
$$x_2 = A_2 \sin(\omega t + \phi). \tag{38}$$

with $\omega = 2\pi f$, and substituting into Eq. (37) one obtains

$$\begin{bmatrix} \dfrac{k_1 + k_2}{m_1} - \omega^2 & -\dfrac{k_2}{m_1} \\ -\dfrac{k_2}{m_2} & \dfrac{k_2}{m_2} - \omega^2 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{39}$$

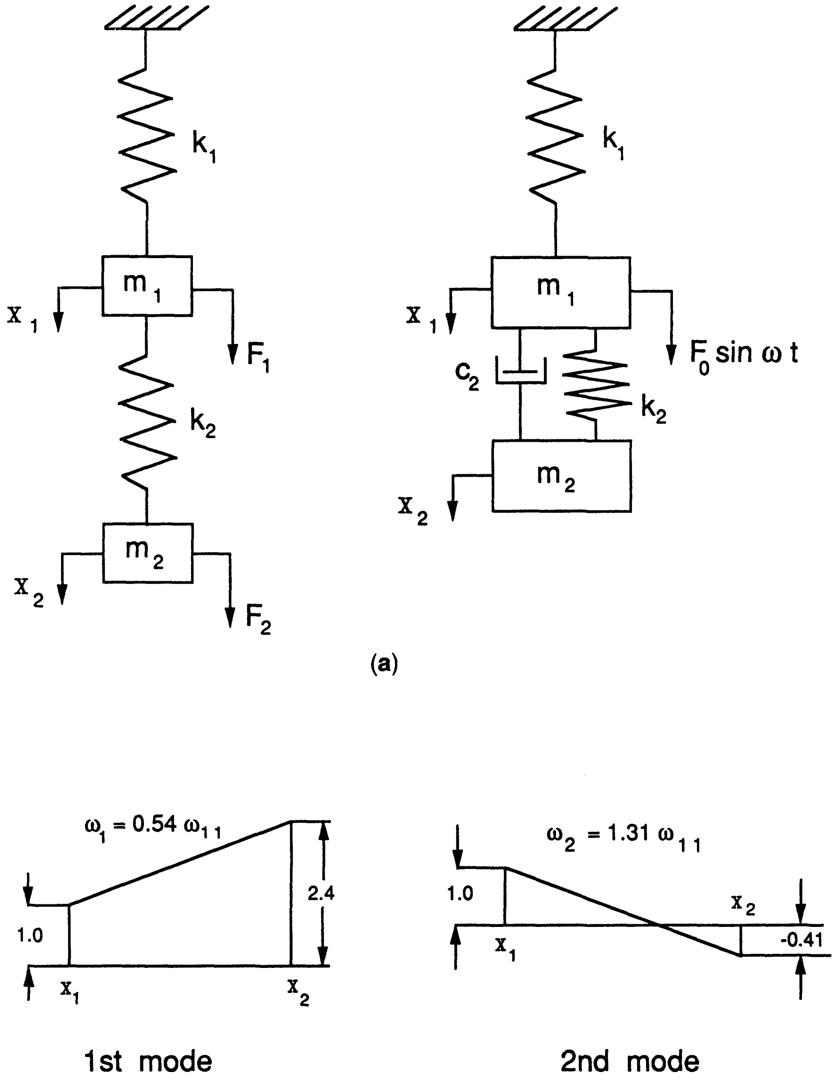
k1
m1
X1
F1
k2
m2
X2
F2
k1
m1
X1
F0 sin ω t
C2
k2
m2
X2
ω1 = 0.54 ω11
1.0
2.4
X1
X2
1st mode
ω2 = 1.31 ω11
1.0
X1
X2
-0.41
2nd mode

(a)

(b)

FIG. 9. (a) Two degree-of-freedom system and (b) modes of vibration.

For $A_1 \neq 0$ and $A_2 \neq 0$, the determinant of the matrix must vanish, giving

$$\left(\frac{\omega}{\omega_{11}}\right)^4 - \left(1 + \frac{k_2}{k_1} + \frac{\omega_{22}^2}{\omega_{11}^2}\right)\left(\frac{\omega}{\omega_{11}}\right)^2 + \frac{\omega_{22}^2}{\omega_{11}^2} = 0, \quad (40)$$

where $\omega_{11}^2 = k_1/m_1$ and $\omega_{22}^2 = k_2/m_2$ are the uncoupled resonances of the first and the second oscillators, respectively. There are two roots of Eq. (40) representing two resonance frequencies of the system, ω_1^2 and ω_2^2 ($\omega_1^2 < \omega_2^2$), for the coupled oscillators. It can be shown that ω_1 is smaller than the lesser of ω_{11} and ω_{22} and that ω_2 is larger than the larger of ω_{11} and ω_{22}. For example, if $m_2 = m_1$ and $k_2 = 0.5k_1$, then $\omega_{22} = 0.71\omega_{11}$, and the two resonance frequencies are $\omega_1 = 0.54\omega_{11} < \omega_{22}$ and $\omega_2 = 1.31\omega_{11} > \omega_{11}$. The relative motion of the masses m_1 and m_2 can be calculated from Eq. (39) for each of the two resonance frequencies. From Eq. (39), the ratio of the amplitudes is given by the expression

$$A_2/A_1 = \omega_{22}^2/(\omega_{22}^2 - \omega^2), \quad (41)$$

so that for $\omega = \omega_1$ and $\omega = \omega_2$

$$(A_2/A_1)_{\omega_1} = \omega_{22}^2/(\omega_{22}^2 - \omega_1^2),$$
$$(A_2/A_1)_{\omega_2} = \omega_{22}^2/(\omega_{22}^2 - \omega_1^2).$$

For the example given above,

$$(A_2/A_1)_{\omega_1} = +2.40, \quad \omega_1 = 0.54\omega_{11},$$
$$(A_2/A_1)_{\omega_2} = -0.41, \quad \omega_2 = 1.31\omega_{11}.$$

This shows that for vibration at the lower resonance frequency ω_1, the two masses are in phase, as shown in Fig. 9(b). The two masses are out of phase when vibrating at the second higher resonance frequency, as shown in Fig. 9(b). The existence of the two resonance frequencies of a two degree-of-freedom system introduces the concept of *natural frequencies* and *mode shapes*. For each resonance frequency, there exists a vibration pattern that corresponds uniquely to that resonance, in such a way that if one were to give the system an initial displacement that corresponds to one of the mode shapes and then releases the masses from rest, they would vibrate at the corresponding natural frequency.

2.2 Forced Vibration of a Two-DOF System

The response of a two degree-of-freedom system when the mass m_1 is under the influence of a harmonic force $F_1 = F_0 \sin\omega t$ and $F_2 = 0$ can be obtained by solving Eq. (37) with solution Eq. (38) which has the same frequency as the input force F_1:

$$|A_1/X_{1s}| = |b^2 - r^2|/D_0, \quad (42)$$

$$|A_2/X_{1s}| = b^2/D_0, \quad (43)$$

where

$$D_0 = |r^4 - r^2(b^2 + ab^2 + 1) + b^2|,$$
$$X_{1s} = F_0/k_1, r = \omega/\omega_{11}, b = \omega_{22}/\omega_{11} \text{ and}$$
$$a = m_2/m_1.$$

A typical plot of the response A_1/X_{1s} is given in Fig. 10 for a system with no damping, $\zeta = 0$.

The forced response amplitude A_1 of the mass m_1 vibrating under the influence of a harmonic force F_0 can be reduced or eliminated if one chooses the natural uncoupled frequency of the second mass ω_{22} to be the same as or close to the forcing frequency ω, i.e., $\omega_{22} = \omega$. This means that $A_1/X_{1s} = 0$ and consequently $A_2/X_{1s} \rightarrow \infty$, indicating that the elimination of the vibration response at the mass m_1 entails excessive vibration of the second mass m_2. The second oscillator, whose characteristics are judiciously chosen to eliminate the vibration of the mass m_1 being excited to vibration, is called the *tuned dynamic absorber*.

If one adds a viscous damper to the second oscillator described by c_2 [see Fig. 9(b)], then the displacements in Eqs. (42) and (43) become

$$\left|\frac{A_1}{X_{1s}}\right| = \frac{\{(r^2 - b^2)^2 + (2\zeta rb)^2\}^{1/2}}{D_d}, \quad (44)$$

$$\left|\frac{A_2}{X_{1s}}\right| = \frac{\{b^4 + (2\zeta rb)^2\}^{1/2}}{D_d}, \quad (45)$$

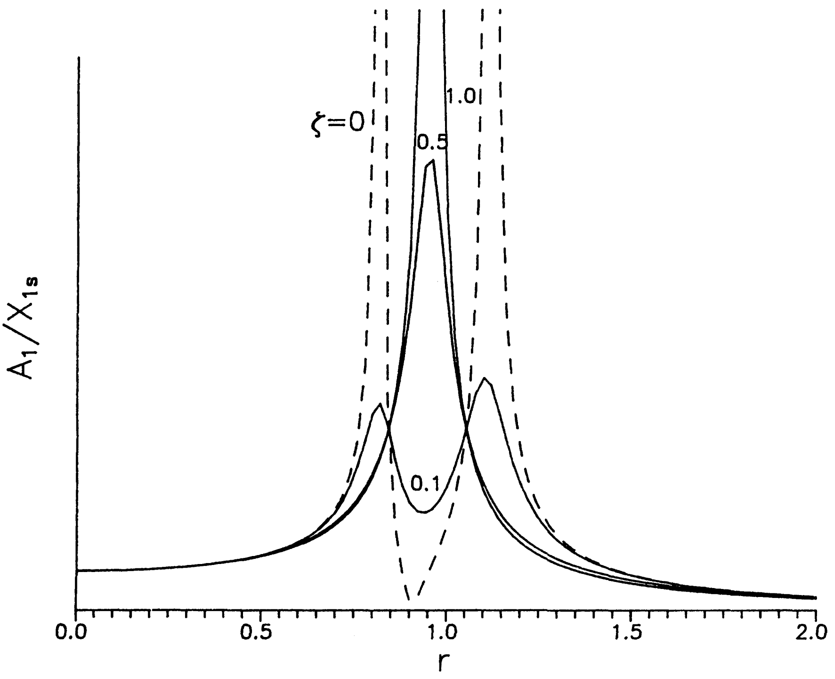

FIG. 10. Response of m_1 vs frequency for a two-DOF system for various damping coefficients.

where

$$D_d = \{(r^2 + ar^2 - 1)^2(2\zeta rb)^2 + [ab^2r^2 - (r^2 - 1)(r^2 - b^2)]^2\}^{1/2},$$

and $\zeta = c_2/2\sqrt{k_2m_2}$. A plot of the response A_1/X_{1s} vs the frequency ratio r as a function of the damping ratio ζ is shown in Fig. 10, for $a = 0.1$, $b = 1$, with $\zeta = 0, 0.1, 0.5$, and ∞. It should be noted that all four curves pass through two points r_1 and r_2, $r_1 < r_2$, and that in general there are two peaks for each value of ζ.

2.3 Tuned Dynamic Vibration Absorber

To minimize the vibration of an oscillator of mass m_1 and stiffness k_1, one can add a tuned second oscillator, whose mass, stiffness, and damping are given by m_2, k_2, and c. To start with, the mass and stiffness m_1 and k_1 are fixed. One can minimize A_1 if one picks $r = b$, i.e., $\omega/\omega_{11} = \omega_{22}/\omega_{11}$ or $\omega_{22} = \omega$. This means that the vibration absorber minimizes the response A_1 only when the forcing frequency is close to the natural frequency of the absorber alone, i.e., a narrow frequency band near $\omega = \omega_{22}$. In order to optimize the vibration absorber so that the response A_1 is minimized over a broader frequency range, one needs to make sure that both peaks of A_1 near r_1 and r_2 are approximately equal. An optimum design requires that $b = b_0$ and $\zeta = \zeta_0$ given by

$$b_0 = \frac{1}{1+a} \quad \text{and}$$
$$\zeta_0^2 = \frac{ab}{8}\left(3 \pm \sqrt{\frac{a}{a+2}}\right), \tag{46}$$

where the + sign is for the r_1 and the − sign for r_2, $r_1 < r_2$. An average value of ζ_0^2 of the two possible values is then given by

$$\zeta_{av}^2 = 3ab/8. \tag{47}$$

Using these results for $a = 0.1$, the optimum value of $b_0 = 0.909$ and $\zeta_{av} = 0.185$, while Eq. (46) gives $\zeta_1 = 0.191$ and $\zeta_2 = 0.178$. The response curve is shown in Fig. 11. More detailed treatments for the analysis and design of different dynamic absorber systems are given by Reed (1988) and Snowdon (1968).

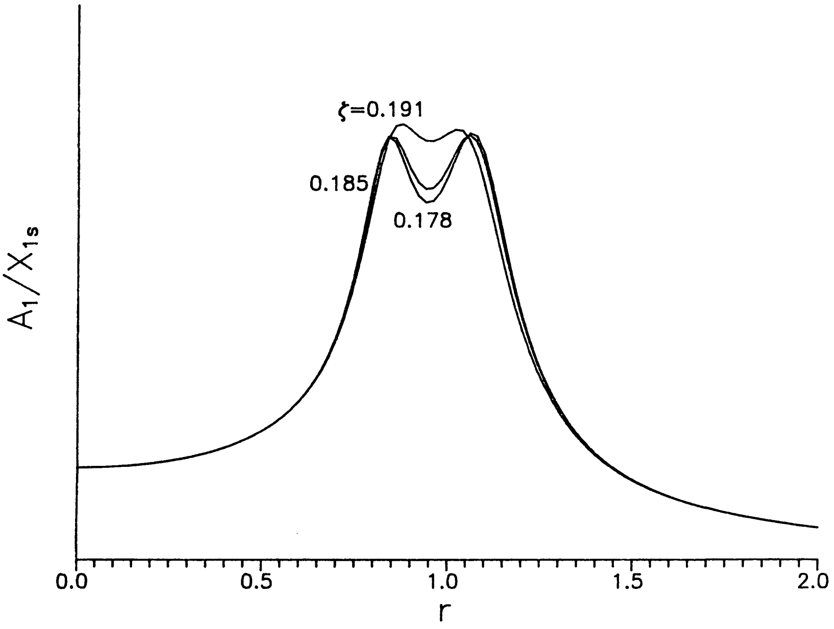

FIG. 11. Response of m_1 with an optimum design of a tuned dynamic vibration absorber vs frequency.

2.4 Vibration of Multiple Degree-of-Freedom Systems

For an n degree-of-freedom system, the two DOF can be generalized in matrix form

$$[\mathbf{M}][\ddot{\mathbf{x}}] + [\mathbf{K}][\mathbf{x}] = [\mathbf{F}], \tag{48}$$

where [**M**] and [**K**] are the $n \times n$ mass and stiffness matrices, respectively, [**x**] is the displacement vector for the n masses, and [**F**] is the force vector acting on the n masses. For free vibration of the system [**F**] = 0, and the displacement vector may be represented by

$$[\mathbf{x}] = [\mathbf{A}]e^{i\omega t}, \tag{49}$$

where [**A**] is the amplitude vector for the n masses. Substitution into Eq. (48) gives

$$\{[\mathbf{K}] - \omega^2[\mathbf{M}]\}[\mathbf{A}] = 0. \tag{50}$$

For a nontrivial solution vector [**A**], we must have the determinant

$$\mathrm{Det}|[\mathbf{K}] - \omega^2[\mathbf{M}]| = 0. \tag{51}$$

The determinant can be shown to lead to a polynomial of degree ω^{2n}, which means that there will be n roots ω_i^2, $i = 1, 2, \ldots, n$. Substitution of each root ω_i^2 into the $n \times n$ system in Eq. (50) results in the computation of the corresponding vector for the normalized amplitude vector $[\mathbf{A}]_i$. The roots ω_i are the ith natural frequencies (called *eigenvalues*) of the system and $[\mathbf{A}]_i$ is the corresponding vector mode shapes (called *eigenvectors*).

3. VIBRATION OF CONTINUOUS ONE-DIMENSIONAL SYSTEMS

In the preceding sections, the vibration of discrete n degree-of-freedom systems was developed in terms of n eigenvalues and eigenvectors. In this section, the vibrations of distributed one-dimensional continuous mechanical systems, such as strings, bars, and beams, are explored.

3.1 Vibration of Stretched Strings

Stretched strings represent the vibrating part of bowed musical instruments, which excites the instrument's surfaces to vibrate and produces music. Consider a flexible string of mass per unit length ρ and stretched by a tension T. Let $y(x,t)$ be the displacement of the string from the equilibrium position at a point x as a function of time t (see Fig. 12). The partial differential equation of motion

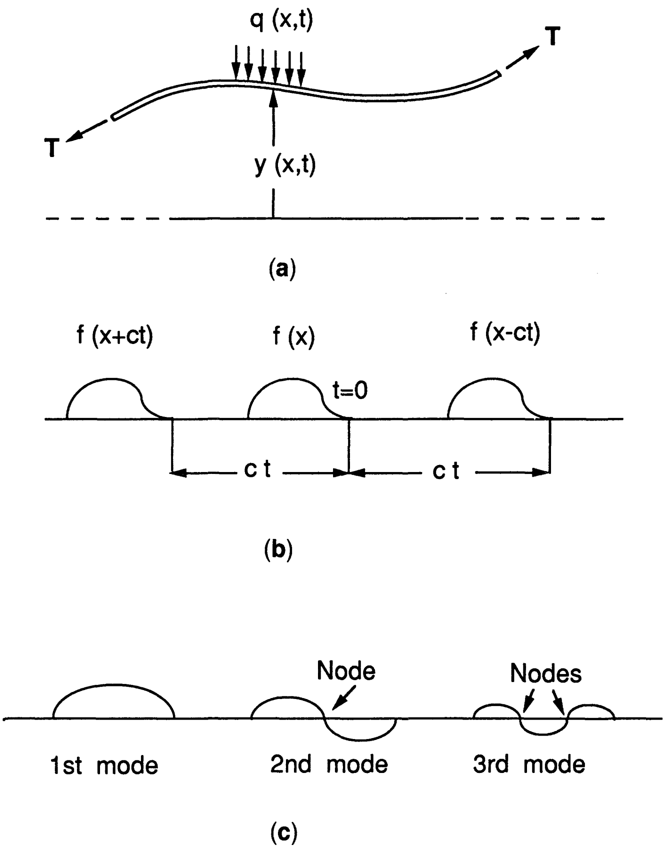

FIG. 12. Vibration of a stretched string: (a) schematic diagram, (b) propagating waves, and (c) first three modes of vibration.

governing the displacement $y(x,t)$ for a string under the influence of a distributed transverse load $q(x,t)$ is given by

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 y}{\partial t^2} - \frac{q(x,t)}{T}, \tag{52}$$

where $c = \sqrt{T/\rho}$ is known as the sound speed of propagation in the string. If the distributed load $q(x,t)$ is harmonic in time, i.e., $q(x,t) = Q(x)e^{i\omega t}$, then one can assume that the response $y(x,t) = Y(x)e^{i\omega t}$, leading to the following equation of motion:

$$\frac{d^2 Y}{dx^2} + k^2 Y = -\frac{Q(x)}{T}, \tag{53}$$

where $k = \omega/c$ is known as the wave number.

To show that Eq. (52) represents a wave motion, let the independent variables u and v replace x and t:

$$u = x + ct,$$
$$v = x - ct,$$

so that Eq. (52) transforms to $\partial^2 y/\partial u \partial v = 0$, which has a solution

$$y = f(u) + g(v) = f(x + ct) + g(x - ct).$$

It can be shown that $g(x - ct)$ represents the motion of the string that started as $g(x)$ at $t = 0$ and moves a distance x in a time interval t, so that $c = x/t$ represents the speed at which $g(x - ct)$ moves in the positive x direction. Similarly, $f(x + ct)$ represents a motion originally $f(x)$ at $t = 0$ and moving in the negative x direction at a speed of c (see Fig. 12). The vibration of a stretched string can be studied through the equation of motion of a force-free string. For a string of a finite length L stretched between two fixed pegs, the solution of Eq. (53) is

$$Y(x) = A \sin kx + B \cos kx. \tag{54}$$

The solutions are periodic in the distance x. It is seen that the displacement is periodic in a distance $\lambda = 2\pi/k = c/f$. The distance λ is called the *wavelength* of the wave and $k = \omega/c$ is the number of full waves per period, hence known as the *wave number*.

For a string fixed at both ends, $Y(0) = 0$ and $Y(L) = 0$, then it can be shown that

$$B = 0 \quad \text{and} \quad A \sin kL = 0. \tag{55}$$

If $A = 0$, then one obtains the trivial solution. However, for a nontrivial solution, $A \neq 0$, and $\sin kL = 0$ or

$$k_n = n\pi/L = \omega_n/c \quad \text{and}$$
$$\omega_n = n\pi c/L, \qquad n = 1, 2, \ldots.$$

The frequencies $f_n = \omega_n/2\pi$ are the natural frequencies of the stretched string:

$$f_n = \frac{nc}{2L} = \frac{n}{2L}\sqrt{\frac{T}{\rho}}, \qquad n = 1, 2, \ldots. \tag{56}$$

There is an infinite number of natural frequencies, equally spaced in frequency, each representing a different musical pitch. It can be seen that the natural frequencies (pitch) can be increased by increasing the tension T or decreasing the length of string L. The mode shape corresponding to each frequency ω_n is given by

$$Y_n(x) = A_n \sin(n\pi x/L), \qquad n = 1, 2, \ldots. \tag{57}$$

The mode for $n = 1$ is called the *fundamental mode*, corresponding to the pitch $f_1 = c/2L$.

If the stretched string is excited to vibration by a harmonic point force $P_0 e^{i\omega t}$ located at x_0, where P_0 is the amplitude of the force and ω is the excitation frequency, then one can show that the response $Y(x)$ is a superposition of all the possible modes of vibration $Y_n(x)$ of Eq. (57). Explicitly stated,

$$Y(x) = \sum_{n=1}^{\infty} E_n \sin(n\pi x/L), \tag{58}$$

where E_n represents the modal weighting of the response corresponding to the nth mode, given by

$$E_n = \frac{2P_0 \sin(n\pi x_0/L)}{M(\omega^2 - \omega_n^2)}, \tag{59}$$

where $M = \rho\text{L}$ is the total mass of the string.

The response is inversely proportional to the mass of the string. When the excitation frequency is at or near one of the natural (resonance) frequencies, the string's response becomes very large. However, all mechanical systems have some built-in energy absorption. Assuming that some damping is present in the string, one can substitute $\bar{\omega}_n^2 = \omega_n^2(1 + i\eta)$, where η is the distributed damping along the string, in the expression in Eq. (59). Thus, even when the forcing frequency ω is equal to one of the resonance frequencies ω_n, the denominator $\omega^2 - \bar{\omega}_n^2 \rightarrow i\eta\omega_n^2$ does not vanish. At any one of the resonance frequencies, the response is inversely proportional to the damping coefficient η. The normalized response of the string measured at the point of excitation $x = x_0$ vs ω/ω_1 is plotted in Fig. 13.

3.2 Longitudinal Vibration of Rods

Elastic bars and rods occur in many structural systems, such as isolation rubber mounts, machinery support foundation frames, and columns in buildings, and are frequently used as machinery components. An elastic rod experiences longitudinal vibration where the motion of a cross section is aligned with the axis of the rod (see Fig. 14). Assuming that the displacement of plane cross section stays plane and the elastic strains are small compared to unity, the partial differential equation governing the displacement

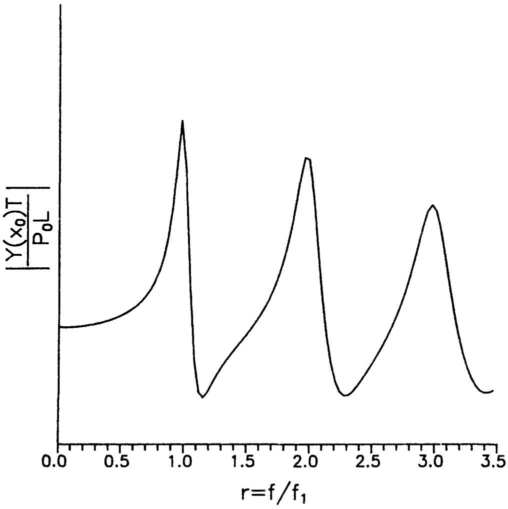

FIG. 13. Forced response of a damped stretched string by a harmonic point force vs frequency.

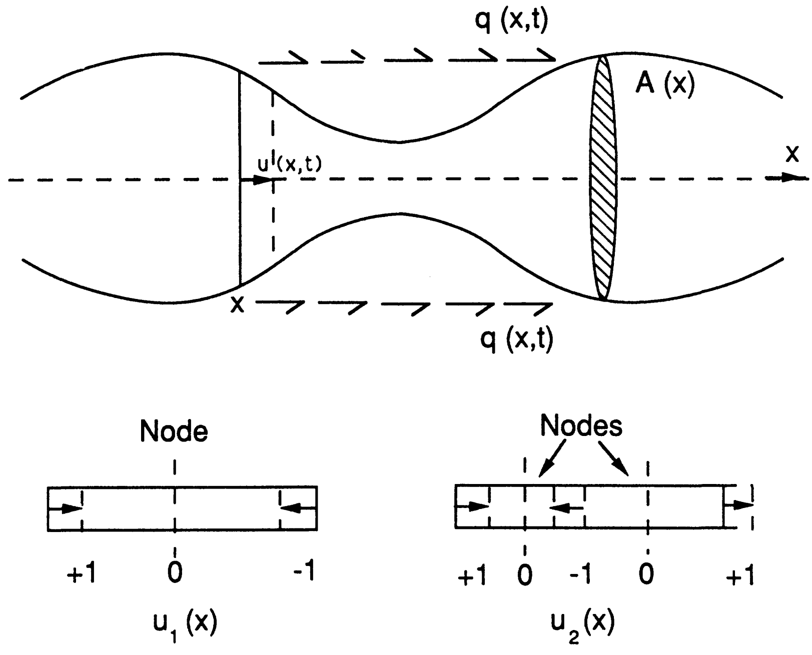

FIG. 14. Longitudinal vibration of bars: (a) schematic diagram and (b) first two modes of vibration.

$y(x,t)$ of a rod whose cross-sectional area is $A(x)$ is given by

$$\frac{\partial}{\partial x}\left(EA\frac{\partial u}{\partial x}\right) = \rho A\frac{\partial^2 u}{\partial t^2} - q(x,t), \tag{60}$$

where E is the Young's modulus, ρ is the mass density of the material, and q is the longitudinal distributed force applied along the bar length (see Fig. 14). If the cross section and the Young's modulus are constants, then Eq. (60) simplifies to

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} - \frac{q(x,t)}{EA}, \tag{61}$$

where $c = \sqrt{E/\rho}$ is the longitudinal (bar) sound speed. The possible boundary conditions can be one of three types: (a) fixed end, $u = 0$; (b) free end, $\partial u/\partial x = 0$; or (c) spring loaded end, $EA\ \partial u/\partial x \pm \gamma u = 0$ with γ the spring constant. An unloaded free bar whose length is L and satisfying boundary condition (b) above has natural frequencies and mode shapes as

$$u_n(x) = \cos(n\pi x/L),$$
$$n = 1, 2, 3, \ldots, \qquad f_n = nc/2L.$$

The mode shapes derived above exhibit displacement of the cross section at both ends of the bar, as shown in Fig. 14.

3.3 Torsional Vibration of Rods

Torsional vibration of elastic rods occurs when the bar undergoes rotational motion about an axis along its length, present in many structural systems such as shafts in machinery, auto engine cams, power transmission systems, and torsion bar suspension systems, to name a few. For a rod with a circular cross section, whose polar cross-sectional moment of inertia is $J(x)$, the equation of motion of the rotation $\vartheta(x,t)$ about the axis of the bar is

$$\frac{\partial}{\partial x}\left(JG\frac{\partial \vartheta}{\partial x}\right) = \rho J\frac{\partial^2 \vartheta}{\partial t^2} - \mu(x,t), \tag{62}$$

where G and ρ are the shear modulus and density of the rod material, respectively, and $\mu(x,t)$ is the distributed moment along the bar. The boundary conditions on each end can be one of three types: (a) fixed end, $\vartheta = 0$; (b) free end, $\partial\vartheta/\partial x = 0$; or (c) elastically supported by a helical spring, $JG\ \partial\vartheta/\partial x \pm \alpha\vartheta = 0$, where α is the helical spring constant. If G

and J are constants, the equation reduces to the standard wave equation:

$$\frac{\partial^2\vartheta}{\partial x^2} = \frac{1}{c_s^2}\frac{\partial^2\vartheta}{\partial t^2} - \frac{\mu(x,t)}{JG}, \qquad (63)$$

where $c_s = \sqrt{G/\rho}$ is the shear velocity. A practical guide to analysis of torsional vibration of machinery shafts and power transmission systems is given by Wilson (1942) and Nesturides (1958).

3.4 Flexural Vibration of Beams

Elastic beams are common structural components, such as auto suspension systems and structural frames of vehicles, ships, and buildings. Thin elastic beams can vibrate when bent so that the motion is transverse to their axes (see Fig. 15). The bending of an elastic beam produces restoring moments and shears for the beam cross section. A beam having a cross-sectional area $A(x)$ and moment of inertia $I(x)$ and undergoing transverse displacement $w(x,t)$ obeys the Bernoulli-Euler equation of motion

$$\frac{\partial^2}{\partial x^2}\left(EI\frac{\partial^2 w}{\partial x^2}\right) + \rho A\frac{\partial^2 w}{\partial t^2} = q(x,t), \qquad (64)$$

where E and ρ are the Young's modulus and density of the beam material, respectively, and $q(x,t)$ is the applied transverse distributed force on the beam. If E, I, ρ, and A are constants, then Eq. (64) becomes

$$\frac{\partial^4 w}{\partial x^4} + \alpha^4\frac{\partial^2 w}{\partial t^2} = \frac{q(x,t)}{EI}, \qquad (65)$$

where $\alpha^4 = \rho A/EI$. In this formulation, the moment and shear at the cross section are

$$M = EI\frac{\partial^2 w}{\partial x^2} \quad \text{and} \quad V = EI\frac{\partial^3 w}{\partial x^3}$$

For time harmonic force $q(x,t) = Q(x)e^{i\omega t}$, the motion of the beam is described by $w(x,t) = y(x)e^{i\omega t}$ that satisfies the equation

$$\frac{d^4 y}{dx^4} + \beta^4 y = \frac{Q(x)}{EI}, \qquad (66)$$

where

$$\beta^4 = \rho A\omega^2/EI = \alpha^4\omega^2.$$

The wave number k_b, wavelength λ_b, and sound speed c_b for a beam are defined as

$$k_b = \beta = \alpha\sqrt{f/2\pi}, \quad \text{or} \quad f = 2\pi k_b^2/\alpha^2,$$

$$\lambda_b = 2\pi/\beta = (2\pi)^{3/2}/\alpha\sqrt{f},$$

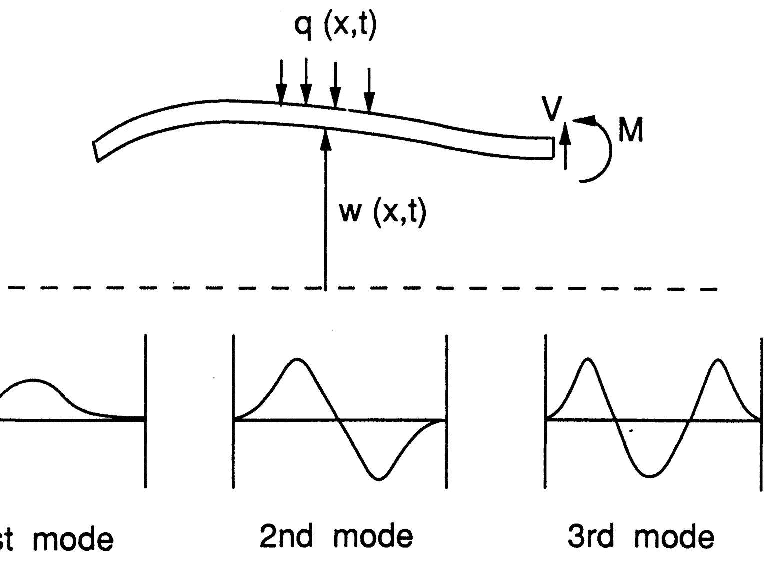

FIG. 15. Flexural vibration of beams: (a) schematic diagram and (b) first three modes of vibration of a clamped beam.

and

$$c_b = 2\pi f/\beta = [(2\pi)^{3/2}/\alpha]\sqrt{f}. \tag{67}$$

This means that the wavelength decreases slowly with frequency. The sound speed is a function of the beam geometrical parameters and increases with frequency, which is not the case for the other continuous media considered earlier. However, since flexural vibration of beams is due to the shear stresses induced in the beam, the maximum sound speed attainable should be the shear velocity of the material of the beam. Since the sound speed increases without limit in Eq. (67) for a Bernoulli-Euler model, the theory has a limited value for high frequencies. The Bernoulli-Euler model is valid up to frequencies where the wavelength λ_b is at least five times the beam thickness. A medium whose frequency spectrum (wavenumber vs frequency) is not linear is called a *dispersive medium*.

The boundary conditions can be prescribed by a pair of conditions at each end, one of nine possible pairs. A few of those are: (a) fixed, $w = 0$ and $\partial w/\partial x = 0$; (b) simply supported, $w = 0$ and $\partial^2 w/\partial x^2 = 0$; and (c) free, $\partial^2 w/\partial x^2 = 0$ and $\partial^3 w/\partial x^3 = 0$.

The general solution for the free flexural vibration of a beam is

$$y(x) = A\sin\beta x + B\cos\beta x + C\sinh\beta x + D\cosh\beta x, \tag{68}$$

which involves circular and hyperbolic functions, indicating that the mode shapes of a transversely vibrating beam are not sinusoidal in form.

For a finite beam of length L that is fixed at both ends, the natural frequencies are the roots of a transcendental equation:

$$\cosh\mu_n \cos\mu_n = 1, \tag{69}$$

where $\mu_n = \beta L$. The first three roots corresponding to the first three modes are $\mu_1 = 4.730$, $\mu_2 = 7.853$, $\mu_3 = 10.996$, and $\mu_n \to (n + \frac{1}{2})\pi$ for n large. The resonance frequencies are related to these roots by

$$f_n = (1/2\pi)\mu_n^2/\alpha^2 L^2, \tag{70}$$

so that the frequency spacing between the resonance frequencies is no longer constant as was the case for strings and bars. In fact, the spacing increases with mode number n. The mode shapes are given by

$$y_n(x) = \sin\left(\mu_n \frac{x}{L}\right) - \sinh\left(\mu_n \frac{x}{L}\right) + \xi_n\left[\cos\left(\mu_n \frac{x}{L}\right) - \cosh\left(\mu_n \frac{x}{L}\right)\right].$$

$$\xi_n = -\frac{\sinh\mu_n - \sin\mu_n}{\cosh\mu_n - \cos\mu_n}. \tag{71}$$

The first three mode shapes are shown in Fig. 15 and are not sinusoidal in form.

4. VIBRATION OF CONTINUOUS TWO-DIMENSIONAL SYSTEMS

In this section, the vibration of two-dimensional continuous mechanical systems, such as membranes, plates, and shells, is explored.

4.1 Vibration of Stretched Membranes

Stretched membranes represent the vibrating component of many ancient and modern musical instruments, such as drums and tambourines. A stretched membrane is a thin elastic sheet whose surface density is ρ, stretched on its boundary by a uniform tensile stress S per unit length of the boundary. If the membrane is disturbed from the equilibrium position by a transverse displacement $w(x,y,t)$, the tension on boundary provides the elastic restoring force. The transverse displacement obeys the two-dimensional wave equation

$$\nabla^2 w = \frac{1}{c^2}\frac{\partial^2 w}{\partial t^2} - \frac{q(x,y,t)}{S}, \tag{72}$$

where $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the Laplacian partial differential operator, $c = \sqrt{S/\rho}$ is the sound speed, and $q(x,y,t)$ is the externally applied distributed load transverse to the membrane (see Fig. 16). The vibration of a stretched membrane is the two-dimensional analog of a stretched string. If the membrane is excited to vibration by a time-harmonic surface load $q(x,y,t) = Q(x,y)e^{i\omega t}$, then the displacement is $w(x,y,t) = W(x,y)e^{i\omega t}$ such that it

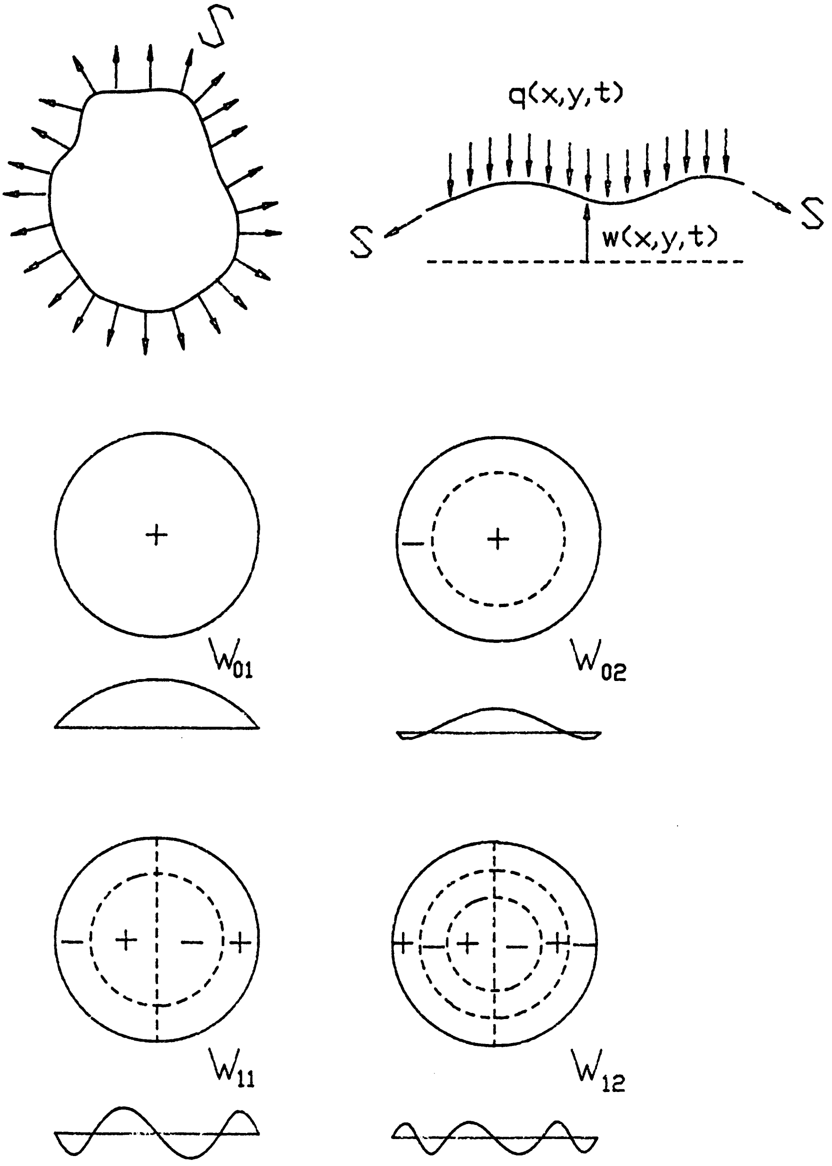

FIG. 16. Vibration of a stretched membranes: (a) schematic diagram and (b) four modes of vibration of a circular membrane.

satisfies the Helmholtz equation

$$\nabla^2 W + k^2 W = -Q(x,y)/S, \tag{73}$$

where $k = \omega/c$ is the wavenumber.

Consider a circular stretched membrane whose radius is a and that is vibrating free of surface loads. The modes of vibration can be shown to be two-dimensional in space in terms of the radial distance r and the angle θ, in the form

$$W(r,\theta) = J_n(k_{nm}r/a)\begin{cases}\sin n\theta \\ \cos n\theta\end{cases}$$
$$n = 0, 1, 2, \ldots, \qquad m = 1, 2, 3, \ldots, \tag{74}$$

where J_n is the Bessel function of order n and k_{nm} are the roots of the transcendental equation

$$J_n(k_{nm}) = 0.$$

The natural (resonance) frequencies f_{nm} are given in terms of k_{nm}:

$$f_{nm} = (1/2\pi)ck_{nm}/a, \tag{75}$$

where the first two roots of the first two Bessel functions are

$$k_{01} = 2.40, \quad k_{02} = 5.52,$$
$$k_{11} = 3.83, \quad k_{12} = 7.02.$$

The mode shapes corresponding to these frequencies are shown in Fig. 16.

4.2 Vibration of Plates

The vibration of elastic plates represents a wide class of structures including steel drums, violins and violas, cymbals, floors, walls, truck sidewalls, glass windows and doors, etc. Elastic plates undergoing deformation resist motion by counter shear forces and moments. In essence, the plate is the two-dimensional analog of a beam in flexure. Consider a uniform elastic plate whose thickness is h; then its transverse displacement $w(x,y,t)$ satisfies the following equation:

$$D\nabla^2\nabla^2 w + \rho h \frac{\partial^2 w}{\partial t^2} = q(x,y,t), \tag{76}$$

where $D = Eh^3/12(1 - \nu^2)$ is the flexural stiffness of the plate; E, ν, and ρ are the Young's modulus, Poisson's ratio, and density, respectively; q is the distributed surface load; and ∇^2 is the Laplacian operator. If the surface load excitation is harmonic in time, $q = Q(x,y)e^{i\omega t}$, then the response is $w = W(x,y)e^{i\omega t}$ satisfying

$$\nabla^2\nabla^2 W - \beta^4 W = Q(x,y)/D, \tag{77}$$

where $\beta^4 = \rho h\omega^2/D$.

The boundary conditions are similar to those of beams. For a simply supported plate, $W = 0$ and $\partial^2 W/\partial n^2 = 0$ on the boundary where n is the normal to the boundary. Consider a rectangular plate, with dimensions a and b, which is simply supported on all sides. For the free vibration of this plate, the mode shapes (see Fig. 17) and corresponding natural frequencies are

$$W_{nm} = \sin\frac{n\pi}{a}x \sin\frac{m\pi}{b}y, \qquad n, m = 1, 2, 3, \ldots,$$
$$f_{nm} = \frac{\pi}{2}\left(\frac{D}{\rho h}\right)^{1/2}\left[\frac{n^2}{a^2} + \frac{m^2}{b^2}\right]. \tag{78}$$

As was the case with the flexural vibration of beams, the natural frequencies increase nonlinearly with mode number (n,m).

4.3 Vibration of Shells

The vibration of elastic shells represents a wide class of structures such as oil tanks, water towers, container trucks, etc. Elastic shells undergoing deformation resist motion through extensional, torsional, and flexural stresses. An elastic thin shell is a curved thin plate, which can deform in the tangential and normal directions. Consider a thin cylindrical shell whose thickness is h and radius is R. The displacements of the middle surface are u, v, and w in the axial, tangential (circumferential), and normal directions, respectively (see Fig. 18). The displacements are functions of the cylindrical coordinates z and θ. The equations of motion governing the time harmonic displacements u, v, and w can be written as

$$\begin{aligned} L_{11}u + L_{12}v + L_{13}w &= -\omega^2(R^2/c_p^2)u, \\ L_{21}u + L_{22}v + L_{23}w &= -\omega^2(R^2/c_p^2)v, \\ L_{31}u + L_{32}v + L_{33}w &= \omega^2(R^2/c_p^2)w, \end{aligned} \tag{79}$$

where ω is the frequency and c_p is the low-frequency phase velocity for compressional waves in a plate, defined as

$$c_p^2 = \frac{E}{\rho(1 - \nu^2)},$$

and D is the normalized extensional rigidity of the shell, defined as

$$D = \frac{Eh}{1 - \nu^2} = \rho c_p^2 h;$$

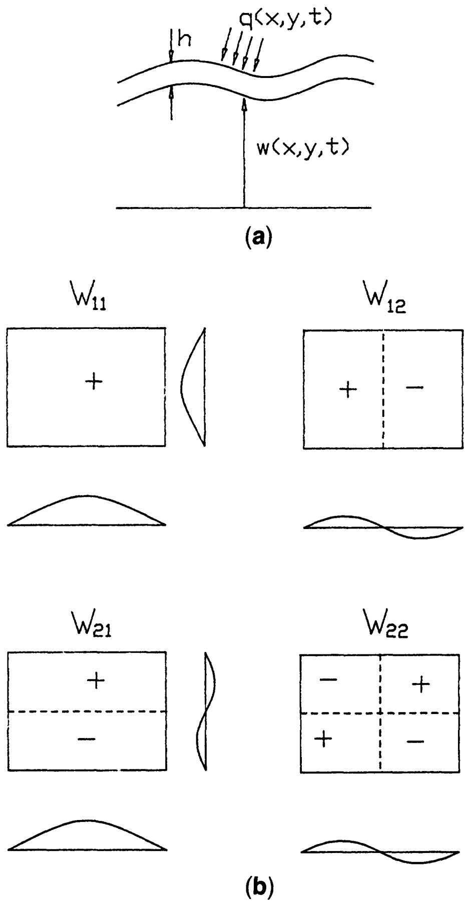

FIG. 17. Flexural vibration of plates: (a) schematic diagram and (b) four modes of vibration of a simply supported rectangular plate.

ν is Poisson's ratio, and the differential operators are given as

$$\begin{aligned}
L_{11} &= \frac{\partial^2}{\partial x^2} + \left[\frac{1-\nu}{2}\right][1+\beta]\frac{\partial^2}{\partial \theta^2},\\
L_{12} &= \left[\frac{1+\nu}{2}\right]\frac{\partial^2}{\partial \theta \partial x},\\
L_{13} &= \nu\frac{\partial}{\partial x} + \left[\frac{1-\nu}{2}\right]\beta\frac{\partial^3}{\partial x \partial \theta^2} - \beta\frac{\partial^3}{\partial x^3},\\
L_{22} &= \left[\frac{1-\nu}{2}\right][1+3\beta]\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial \theta^2},\\
L_{23} &= \frac{\partial}{\partial \theta} - \left[\frac{3-\nu}{2}\right]\beta\frac{\partial^3}{\partial \theta \partial x^2},\\
L_{33} &= 1 + \beta\left[\frac{\partial^4}{\partial x^4} + 2\frac{\partial^4}{\partial \theta^2 \partial x^2} + \frac{\partial^4}{\partial \theta^4} + 2\frac{\partial^2}{\partial \theta^2} + 1\right],\\
L_{21} &= L_{12}, \quad L_{31} = L_{13}, \quad L_{32} = L_{23}, \qquad (80)
\end{aligned}$$

where

$$\beta = h^2/12R^2, \text{ and } \quad x = z/R.$$

Consider a finite shell whose length is L and that is simply supported at both ends. The mode shapes of the displacements for the (mn)th mode are

$$\begin{aligned}
u_{mn}(x,\theta) &= U_{mn}\cos(m\pi R/L)x\cos n\theta\\
m &= 1, 2, \ldots,\\
v_{mn}(x,\theta) &= V_{mn}\sin(m\pi R/L)x\sin n\theta\\
n &= 0, 1, 2, \ldots,\\
w_{mn}(x,\theta) &= W_{mn}\sin(m\pi R/L)x\cos n\theta. \qquad (81)
\end{aligned}$$

Substitution of the mode shapes in Eq. (81) into Eq. (79) will result in a cubic equation for ω^2 which results in three natural frequencies for each mode number (m,n), i.e., $\omega_{mn}^{(1)}$, $\omega_{mn}^{(2)}$; and $\omega_{mn}^{(3)}$. The displacement amplitude ratios U_{mn}/W_{mn} and V_{mn}/W_{mn} will have distinct values for each of the three frequencies of the (m,n)th mode. These distinct modes represent coupled motion that is primarily longitudinal, torsional, or flexural. A plot of the normalized frequency Ω vs the normalized wavenumber ζ_m,

$$\Omega = \omega R/c_p, \qquad \xi_m = m\pi R/L,$$

for $n = 1$ and 2 is shown in Fig. 19 for the three types of modes.

5. PRACTICAL STRUCTURAL DAMPING APPLICATIONS

In the preceding sections, the importance of damping in reducing vibration levels has been clearly demonstrated. Damping of simple or multiple degree-of-freedom systems at the point(s) of the location of the mass(es) can be achieved by the introduction of damping at those point(s) with viscous dashpot dampers, such as automobile suspension systems. However, to reduce the vibration levels in continuous (distributed) structural systems, one must develop damping treatments that are effective over part of or the entire structure. A thorough treatment of this topic is covered by Jones (1988), Nashif *et al.* (1985), Hanson and Hampel (1984), and Ross *et al.* (1959).

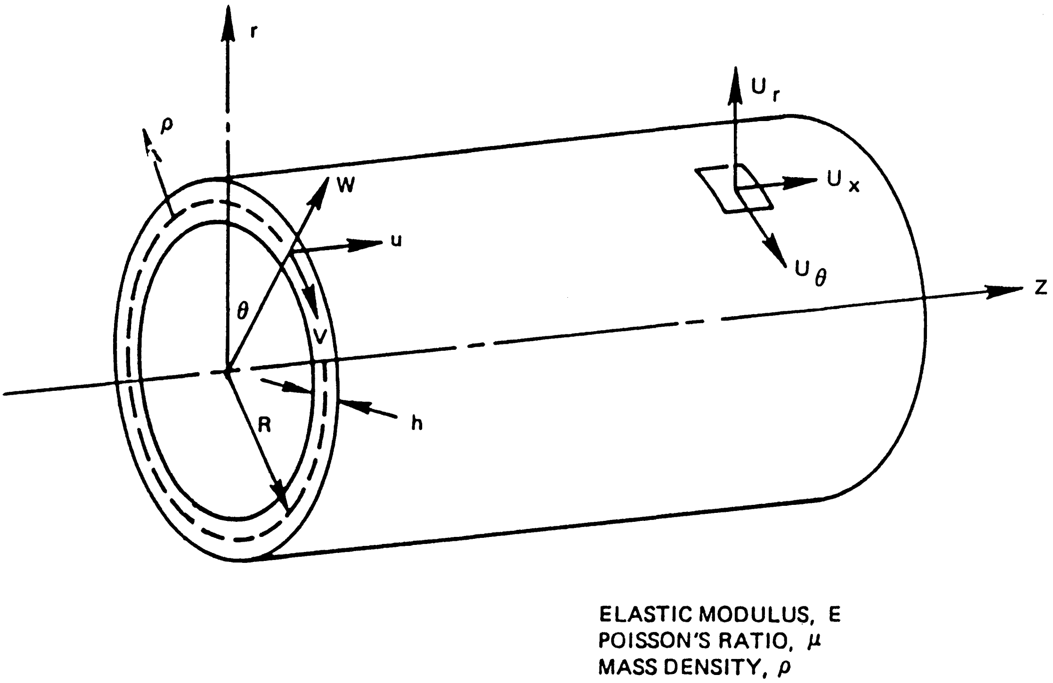
r
ρ
W
u
θ
V
R
h
U_r
U_x
U_θ
Z
ELASTIC MODULUS, E
POISSON'S RATIO, μ
MASS DENSITY, ρ

FIG. 18. Schematic diagram of displacements *u*, *v*, and *w* of a cylindrical shell.

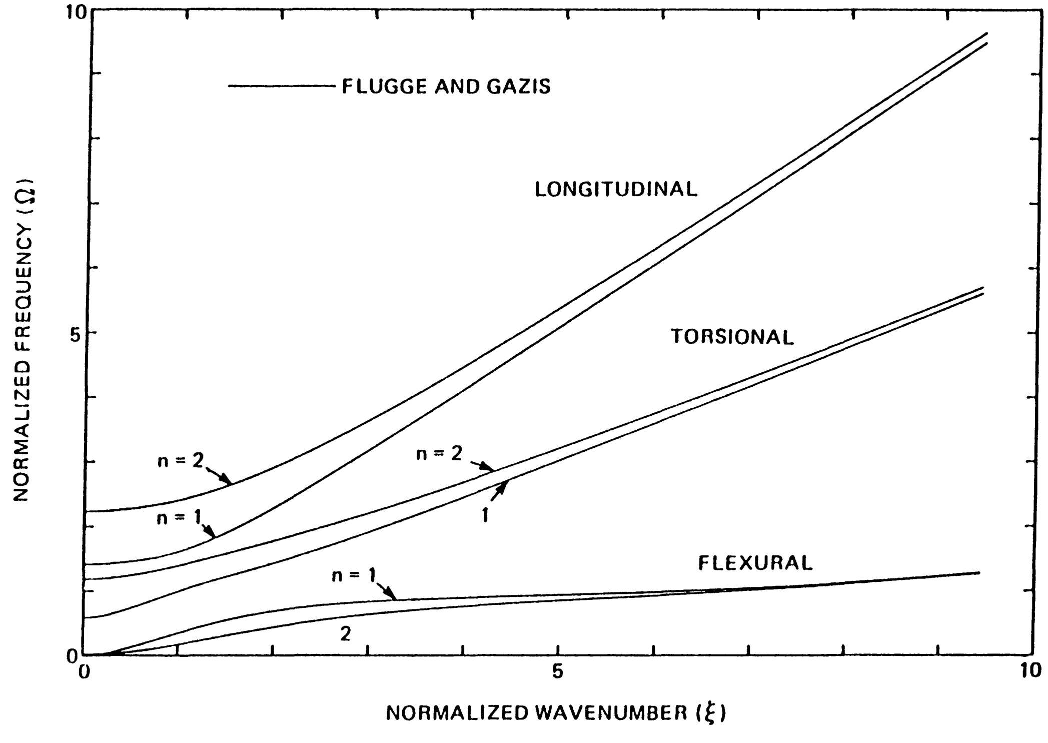

FIG. 19. Frequency Ω vs wavenumber ξ_m for a simply supported cylindrical shell.

5.1 Surface Damping Treatment

In this method of damping, a viscoelastic material with high loss factor is cemented to the surface of the structure. Thus, when the structure deforms in extension or flexure, the surface damping layer will also extend or flex. Since the loss factor for such materials is constant per unit volume of the material, the damping increases with thickness of the layer and the total energy of deformation of the structure.

Consider a plate whose thickness is h_1, Young's modulus E_1, with a viscoelastic layer whose thickness is h_2 and complex Young's modulus $E_2^* = E_2(1 + i\eta_2)$ [see Fig. 20(a)]. The Young's modulus E^* and the damping coefficient η of the composite two-layered structure treated as one material are

$$E^* = E_1(1 + i\eta),$$

where

$$\eta \cong \frac{\eta_2 E_2}{E_1}\frac{h_2^3}{h_1^3}\left(3\frac{h_1^2}{h_2^2} + 6\frac{h_1}{h_2} + 4\right), \tag{82}$$

which is valid when the ratio $E_2/E_1 \ll 1$.

It should be noted that the damping of the viscoelastic layer is generated by the extension of the layer. To increase the extension of the damping layer when the plate undergoes flexure, the damping layer should be separated from the plate as far as practically possible. This is accomplished by placing a spacer layer, thickness h_3, that is lightweight and rigid, such as aluminum honeycomb [see Fig. 20(b)]. The damping of the new composite then becomes

$$\eta \cong \frac{\eta_2 E_2}{E_1}\frac{h_2^3}{h_1^3}\left(1 + 12\frac{\bar{h}^2}{h_2^2}\right), \tag{83}$$

where

$$\bar{h} = \tfrac{1}{2}(h_1 + h_2) + h_3.$$

5.2 Constrained-Layer Damping

Unlike the energy absorption of a viscoelastic layer surface treatment through extensional deformation, constrained-layer damping is achieved through shearing of the layer. This is accomplished by cementing a thin elastic constraining layer on top of a viscoelastic layer, which is cemented to the plate [see Fig. 20(c)]. Because the constraining layer is stiffer than the viscoelastic layer it is cemented to, the latter undergoes shear deformation as much as extension. Thus, the shear deformation in the viscoelastic layer becomes the dominating portion of the loss of energy in the layer. The loss factor of composite η becomes (see Ungar, 1971)

$$\eta = \frac{\eta_2 YX}{1 + X(2 + Y) + (1 + Y)(1 + \eta_2^2)X^2}, \tag{84}$$

where

$$Y = \frac{E_1h_1^3 + E_3h_3^3}{12\bar{h}^2}\left(\frac{1}{E_1h_1} + \frac{1}{E_3h_3}\right), \tag{85}$$

and

$$X = \frac{G_2\lambda^2}{4\pi^2 h_2}\left(\frac{1}{E_1h_1} + \frac{1}{E_3h_3}\right), \tag{86}$$

with

$$\bar{h} = \tfrac{1}{2}(h_1 + h_3) + h_2.$$

G_2 is the real part of the shear modulus, and λ is the flexural wavelength of the composite. The dependence of the loss factor η on X and Y shows that one can maximize it by finding an optimum combination of X and Y given by

$$X_{\text{opt}} = 1/\sqrt{(1 + Y)(1 + \eta_2^2)}, \tag{87}$$

resulting in maximum value of η

$$\eta_{\max} = \eta_2 Y/(2 + Y + 2/X_{\text{opt}}). \tag{88}$$

The wavelength of the composite depends on the stiffness and density of its components,

$$\lambda^2 = (2\pi/f)\sqrt{\bar{D}/\bar{m}}, \tag{89}$$

where

$$\bar{D} = (D_1 + D_3)\left(1 + \frac{XY}{1 + X}\right), \tag{90}$$

$D_i = E_ih_i^3/12(1 - \nu_i^2)$ is the stiffness of the ith layer, and $\bar{m} = \rho_1h_1 + \rho_2h_2 + \rho_3h_3$, where ρ_i

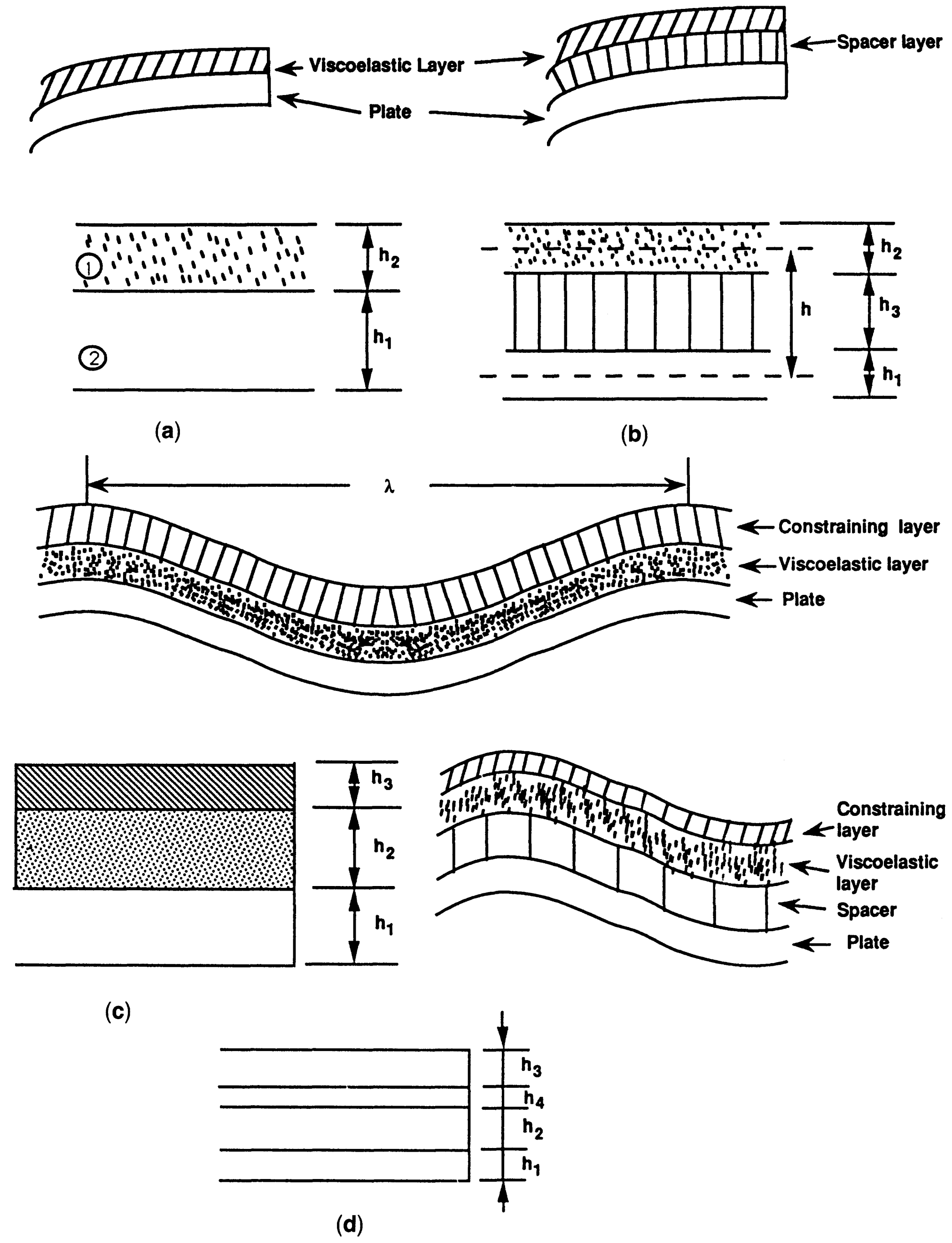
Viscoelastic Layer
Plate
Spacer layer
①
②
h2
h1
(a)
h
h2
h3
h1
(b)
λ
Constraining layer
Viscoelastic layer
Plate
h3
h2
h1
Constraining layer
Viscoelastic layer
Spacer
Plate
(c)
h3
h4
h2
h1
(d)

FIG. 20. Viscoelastic layer damping treatments.

is the density of the ith layer. The frequency at the optimum value is given by

$$f_{\max} = \frac{1}{2\pi}\frac{G_2\sqrt{1+\eta_2^2}}{E_3 h_3 h_2}\sqrt{\frac{\bar{D}}{\bar{m}}}. \tag{91}$$

To design a damping treatment for a plate, choose the largest possible value of Y, Eq. (85), and a viscoelastic material with a high shear damping coefficient h_2. This leads to an optimum choice of the viscoelastic layer thickness h_2 by use of Eq. (87).

One can increase the energy loss in the viscoelastic layer by introducing a stiff spacer layer, such as aluminum honeycomb. The only change in the above analysis is to increase $\bar{h}$ by the spacer thickness, h_4 [see Fig. 20(d)], i.e.,

$$\bar{h} = \tfrac{1}{2}(h_1 + h_3) + h_2 + h_4$$

in the expression for Y.

5.3 Vibration Isolation Mounts

To control excessive vibration or to minimize vibration load transmission of vibrating machinery, engines, pumps, sensitive equipment, etc., one needs to put vibration isolators between the mechanical source of vibration and the component or structure that needs to be isolated. There are various types of vibration isolators: elastomeric, metal, plastic, and air springs.

5.3.1 Elastomeric Vibration Mounts Elastomeric mounts are made of natural or synthetic rubber and rubberlike materials. These are characterized by efficient energy storage capacity and can be molded or shaped into a variety of forms. They can be stretched more than 300% and can sustain stresses that exceed 0.15 Pa. All elastomeric materials are inherently damped where the damping ratio increases with the frequency of the vibration of the structure being supported as well as with the amplitude of the shear strain. This property is used best when the structure or machinery is operated near resonance, i.e., when the energy absorption capacity of the elastomer is highest.

Because of the creeping behavior of elastomers under static loading, elastomeric mounts should not be subjected to large static strains over long periods of time. A prudent design load should limit the static strain to 30% in shear and 10% in compression. Continuous application of dynamic cyclic loads can fatigue elastomeric mounts. This is aggravated if sharp metal parts are supported directly by the mounts. Thus, metal washers and snubbers that come into contact with the mount should be large enough to minimize cutting of their edges into the elastomer. To minimize fatigue failure, a static precompression load should be applied so that the dynamic cyclic load does not cross zero.

Elastomeric materials include natural rubber, neoprene, nitrile, buna S, and silicone rubbers. Each material has advantages and limitations based on the temperature and the presence of hydrocarbons as well as tear and abrasion resistance.

Elastomeric mounts can be used in various forms based on the type of loading they encounter.

1. Compression load: This type of load compresses and shortens the isolator. Under compression, the elastomer load-deflection curve becomes nonlinear; i.e., the stiffness increases as the elastomer is compressed further. This stiffening effect can be used to limit dynamic deflections under shock loads, but the energy storage capacity is lower than in tension.
2. Tension load: This type of load elongates an isolator under tension. The load-deflection curve becomes nonlinear with decreasing stiffness as the elastomer is stretched. The softening effect under increased stretch would store more energy in the isolator and hence enhance its damping characteristics. In general, tension loading is not recommended because of frequent failure of metal-to-rubber bonds.
3. Shear loading: Shear loading is defined as load such that, when applied to an elastomer, adjacent sections slide parallel but in opposite directions to each other. Its load-deflection curve is linear, and so it can serve as an ideal all-purpose isolation mount. It has an efficient energy storage capacity, and it is the most suitable mount for shock loading.
4. Torsional loading: Torsional loading is defined as a load that, when applied to an elastomer, causes adjacent sections to rotate parallel but in opposite directions to each other. This load-deflection curve is almost linear, and so it also has an efficient energy storage capacity.

Combinations of loading are frequently encountered in practice. This is achieved by supporting the structure being isolated by mounts that are set at an angle, resulting in compression and shear loadings.

5.3.2 Metal Springs Metal springs come in a variety of types—coil, Belleville, leaf, and cantilever. For ferrous materials, the modulus of elasticity is $E = 4370$ Pa and the shear modulus $G = 1670$ Pa. For stainless steel, $E = 4070$ Pa and $G = 1450$ Pa. A practical guide to the choice of metal springs for any application is detailed below.

1. Coil springs: Coil springs are helical springs made of bar stock wire coiled in a continuous loop. Those used in tension load have a tightly coiled shape, and those used in compression have a gap between the coil turns. Coil springs have minimal damping and hence transmit vibration to foundations. However, they are low cost and efficient and are used to support engines and machinery for vibration and shock isolation. For tension loads, coil springs come in a variety of types depending on the shapes of the spring ends. A spring end may have a half or full loop, eye, hook, swivel, or swivel bolt. The latter two are ideal for repetitive loads, because the end turns do not have stress concentrations. For compression springs under heavy static loads, some metal housing may be necessary to prevent lateral motion.

   The choice of a metal spring depends on the static loading being applied, which directly influences the static deflection of the spring. In general, the static deflection is proportional to the applied load, the mean diameter of the spring, and the total number of active coils, and inversely proportional to the wire diameter. For design and choice of the appropriate spring, see SAE J795a (1973).
2. Belleville springs: These springs are cone-shaped disc springs. They are suitable for large loads and small deflections. Their load vs deflection characteristics are functions of the ratio of the height of the cone to its thickness. These springs can be stacked in parallel, which increases their load-carrying capacity, or in series, which increases the deflection for the same load. For choice and design of Belleville springs, see SAE J798 (1971).
3. Leaf springs: Leaf springs are stacked and used as a structural member in flexure. They are used to support heavy structures and vehicles. For design data, see SAE J788a (1970).
4. Cantilever springs: These are made from bar stock or strips and used to support machinery in a cantilever mode.

5.3.3 Plastic Springs Springs made from resilient plastics have similar characteristics to elastomeric and metal springs. The resilient components are made from resilient thermoplastic elastomers, and the structural parts are made from rigid thermoplastic materials. Plastic springs are low cost and uniform in dynamic performance but can operate only in the temperature range −40 to +40°C.

5.3.4 Air Springs An air spring employs gas as a resilient element in a sealed pressure vessel with provision for filling or releasing gas. The energy storage capacity of air springs is higher than mechanical springs per unit weight. Use of air springs can be advantageous for use in support of static load while maintaining low dynamic stiffness for vibration isolation. The stiffness of an air spring is directly proportional to the internal pressure and cross-sectional area and inversely proportional to the volume of the air.

GLOSSARY

Amplitude: The maximum value of the response of a sinusoidal motion.

Coulomb Damping: Damping induced by dry friction between the simple oscillator and an adjacent surface.

Critical Damping (c_c): The level of viscous damping that reduces the motion of a system to exponential decay without oscillation.

Damping: Dissipation of mechanical energy in a system.

Damping Ratio (ζ): The ratio of the damping coefficient of a system c to critical damping c_c, $\zeta = c/c_c$.

Force Transmissibility: The force transferred to a supporting structure when a mechanical system is excited to vibration by harmonic mechanical forces.

Free Vibration: The oscillatory motion of a system that is disturbed from its equilibrium position without being forced to motion by external mechanical forces.

Frequency (f): The frequency of a periodic motion is the reciprocal of the period of the motion and is measured by the number of periods per second known as hertz (Hz).

Harmonic: A harmonic motion is a sinusoidal excitation whose frequency is an integer multiple of another related frequency.

Logarithmic Decrement (δ): The rate of decay of the peak transverse vibration of a simple oscillator.

Mode of Vibration: The state of vibration displacement of a mechanical system that occurs at one of the natural frequencies of a system.

Multiple Degree-of-Freedom System: A system consisting of several single degree-of-freedom systems coupled mechanically.

Natural Frequency: The frequency at which a mechanical system will undergo oscillatory motion if disturbed from equilibrium.

Oscillation: The variation of a function of time with time when the value of the function increases and decreases periodically about a mean value.

Period: The period of a function of time is the shortest time between values where the function repeats itself.

Phase: The fraction of a period by which the time response has advanced, measured in terms of angular units.

Quality Factor (Q): A measure of the sharpness of the frequency response of a damped mechanical system at the natural frequencies.

Resonance: A state of high-level response of a structure when it is excited by a force whose frequency coincides with one of its natural frequencies.

Single Degree-of-Freedom System: A system whose motion is describable by a single coordinate, typified by a mass, elastic restoring medium, and a damping mechanism.

Steady-State Vibration: The state of vibration where the motion is periodic in time.

Transient Vibration: A state of a vibrating system where the motion is not periodic.

Tuned Dynamic Absorber: A damped simple oscillator attached to a mechanical system to absorb the vibrational energy of the system.

Vibration Isolation: Method of reducing the transmission of forces to a supporting structure undergoing forced excitation by harmonic mechanical forces.

Viscous Damping: Rate of absorption of mechanical energy in a simple oscillator through dissipation as a result of the viscosity of the medium.

Wavelength: The spatial period of a wave traveling in a medium, where the displacement of the medium repeats itself.

Wave Number: Reciprocal of the wavelength indicating the number of wavelengths per unit length of the medium.

Works Cited

Alfrey, T., Jr. (1948), *Mechanical Behavior of High Polymers*, New York: Interscience Publishers.

ASTM (1964a), ASTM Method D1372-64, *Package Cushioning Materials*.

ASTM (1964b), ASTM Method D1596-64, *Shock Absorbing Characteristics of Packaging Cushioning Materials*.

Crede, C. E. (1951), *Vibration and Shock Isolation*, New York: John Wiley & Sons.

Frye, W. A. (1988), in: C. M. Harris (Ed.), *Shock and Vibration Handbook*, 3rd ed., New York: McGraw-Hill, Chap. 35, "Properties of Rubber."

Hanson, W. J., Hampel, G. A. (1984), *J. Sound Vib.* **18** (22).

Hatae, M. T. (1988), in: C. M. Harris (Ed.), *Shock and Vibration Handbook*, 3rd ed., New York: McGraw-Hill, Chap. 41, "Packaging Design."

Jones, D. I. G. (1967), *Shock Vib. Bull.* **36** (4), 9.

Jones, D. I. G. (1988), in C. M. Harris (Ed.), *Shock and Vibration Handbook*, 3rd ed., New York: McGraw-Hill, Chap. 37, "Application of Damping Treatments."

Lazan, B. J. (1968), *Damping of Materials and Members in Structural Mechanics*, New York: Pergamon Press.

Nashif, A. D., Jones, D. I. G., Henderson, J. P. (1985), *Vibration Damping*, New York: John Wiley & Sons.

Nesturides, E. J. (1958), *A Handbook of Torsional Vibration*, Cambridge, U.K.: Cambridge University Press.

Plunkett, R. (1981), *Friction Damping*, ASME publication AMD Vol. 38 (4), p. 65.

Racca, R. H. (1988), in: C. M. Harris (Ed.), *Shock and Vibration Handbook*, 3rd ed., New York: McGraw-Hill, Chap. 32, "Types and Characteristics of Vibration Isolators."

Reed, F. E. (1988), in: C. M. Harris (Ed.), *Shock and Vibration Handbook*, 3rd ed., New York: McGraw-Hill, Chap. 6, "Dynamic Vibration Absorbers and Auxiliary Mass Dampers."

Ross, D., Ungar, E. E., Kerwin, E. M. (1959), *Proc. Colloquium on Structural Damping*, New York: ASME, p. 49.

SAE J795a (1973), *Design and Application of Helical and Sprial Springs*, Warrendale, PA: Society of Automotive Engineers.

SAE J798 (1971), *Manual on Design and Application of Coned Disk Springs or Belleville Springs*, Warrendale, PA: Society of Automotive Engineers.

SAE J788a (1970), *Manual on Design and Appli-*

cation of Leaf Springs, Warrendale, PA: Society of Automotive Engineers.

Snowdon, J. C. (1968), *Vibration and Shock in Damped Mechanical Systems*, New York: John Wiley & Sons.

Ungar, E. E. (1971), in: L. L. Beranek (Ed.), *Noise and Vibration Control*, New York: McGraw-Hill, Chap. 14.

Whitby, G. S. (Ed.) (1954), *Synthetic Rubber*, New York: John Wiley & Sons.

Wilson, W. K. (1942), *Practical Solutions of Torsional Vibration Problems*, New York: John Wiley & Sons.

Further Reading

Cremer, L., Heckl, M., Ungar, E. E. (1972), *Structure-Borne Sound*, New York: Springer-Verlag.

Den Hartog, J. P. (1956), *Mechanical Vibration*, 4th ed., New York: McGraw-Hill.

Harris, C. M., Crede, C. E. (1961), *Shock and Vibration Handbook*, New York: McGraw-Hill.

Kalnins, A., C. L. Dym (Eds.) (1976), "Vibrations: Beams, Plates and Shells," in *Benchmark Papers in Acoustics*, Stroudsburg, PA: Dowden, Hutchinson and Ross/Halsted Press, Vol. 8.

Meirovitch, L. (1986), *Elements of Vibration Analysis*, 2nd ed., New York: McGraw-Hill.

Steidel, R. F., Jr. (1988), *An Introduction to Mechanical Vibrations*, 3rd ed., New York: John Wiley & Sons.

Thomson, W. T. (1988), *Theory of Vibrations with Applications*, 3rd ed., Englewood Cliffs, NJ: Prentice-Hall.

Timoshenko, S., Young, D. H., Weaver, W., Jr. (1974), *Vibration Problems in Engineering*, New York: John Wiley & Sons.

Ungar, E. E. (1971), in: L. L. Beranek (Ed.), *Noise and Vibration Control*, New York: McGraw-Hill, Chap. 14.

MECHANICS, CLASSICAL

Julian L. Davis, *Cranbury, New Jersey, U.S.A.*

Introduction 1
1. **Particle Mechanics—a System of Particles** 2
1.1 Newton's Laws of Motion 2
1.2 Angular Momentum 3
1.3 Work and Energy 3
1.4 Conservation Laws 4
1.5 Generalized Coordinates 5
1.6 Constraints 5
1.7 The Principle of Virtual Work and D'Alembert's Principle 7
1.7.1 Transformation to Generalized Coordinates 8
1.8 Lagrange's Form of the Equations of Motion 8
2. **Rigid-Body Mechanics** 9
2.1 Rotation of a Body about a Fixed Point 9
2.1.1 Eulerian Angles 10
2.2 Moment of Inertia 11
2.3 Rotational Kinetic Energy 12
2.4 Euler's Equations of Motion for Rigid-Body Rotation 12
2.4.1 Torque-Free Motion of the Earth 12
2.5 Motion of a Top 13
2.5.1 Stability of a Sleeping Top 15
3. **Variational Principles and the Calculus of Variations** 16
3.1 Lagrange's Equations from Hamilton's Variational Principle 18
3.1.1 Dealing with Constraints 19
3.2 Nonconservative Systems 20
3.2.1 Rayleigh's Dissipation Function 21
4. **Hamilton's Equations of Motion** 21
4.1 Cyclic Coordinates 23
5. **Canonical Equations of Transformation** 23
5.1 Hamilton-Jacobi Theory 24
5.1.1 Hamilton-Jacobi Equations 24
5.1.2 Hamilton's Characteristic Function 25
5.1.3 Keplerian Motion of an Electron about a Nucleus 26
6. **Principle of Least Action** 27
7. **Action-Angle Variables and Phase Integrals** 29
Glossary 31
Works Cited 32
Further Reading 32

> Science is built on facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones a house. . . . Scientists believe that there is a hierarchy of facts and that a judicious selection can be made. One has only to open one's eyes to see that the triumphs of industry, which have enriched so many practical men, would never have seen the light if only these practical men had existed, and if they had not been preceded by disinterested fools who died poor, who never thought of the useful, and yet had a guide that was not their own caprice. What fools did, as Mach has said, was to save their successors the trouble of thinking. . . . It is necessary to think for those who do not like thinking, and as they are many, each one of our thoughts must be useful in as many circumstances as possible. For this reason, the more general a law is, the greater its value.
>
> Henri Poincaré (1902)

INTRODUCTION

This article is an advanced treatment of classical mechanics, sometimes called *analytical dynamics*. The material presupposes an elementary knowledge of mechanics such as that given in an undergraduate course. Some of the references in the list of Further Read-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ing give a full discussion of these aspects of mechanics. Particularly recommended is Synge and Griffith (1959).

Analytical dynamics is described by Whittaker (1974) as "that branch of knowledge in which the motions of material bodies, considered as due to the mutual interactions of the bodies, are discussed by the aid of mathematical analysis." In this spirit, mathematical analysis is used and developed within the context of the material. Mathematics is not used for its own sake, but as a powerful tool to codify and explain complex physical phenomena arising in mechanics. The material to be covered is "classical" in the sense that it will not include the special and general theories of relativity. Neither will it discuss quantum mechanics. These topics are adequately discussed in other articles in this Encyclopedia (RELATIVITY, SPECIAL; GRAVITATION AND GENERAL RELATIVITY; QUANTUM MECHANICS).

1. PARTICLE MECHANICS—A SYSTEM OF PARTICLES

Consider a system of n particles acted on by external forces and internal forces. Let $\mathbf{F}_i$ be the external force acting on the ith particle and $\mathbf{f}_{ij}$ be the internal force acting on the ith particle due to the jth particle. (Clearly, $\mathbf{f}_{ii}$ is zero.) Let $\mathbf{p}_i = m_i d\mathbf{r}_i/dt$, where m_i is the mass of the ith particle, be the linear momentum of the ith particle. Then Newton's second law yields the following equations of motion, one for each of the n particles:

$$\sum_j (\mathbf{f}_{ij} + \mathbf{F}_i) = \frac{d\mathbf{p}_i}{dt} = \dot{\mathbf{p}}_i, \tag{1}$$

where i and j range from 1 to n. Note that there are n equations, in each of which we sum all the interactions due to the other $n - 1$ particles. All the forces are assumed to obey Newton's third law.

1.1 Newton's Laws of Motion

We now obtain equations of motion in terms of the position of each particle. Consider an inertial reference system, defined effectively as one relative to which a particle's motion is determined only by the real forces exerted on it. A frame of reference fixed relative to the sun and having a fixed orientation relative to the fixed stars forms an excellent approximation to an inertial system; a frame of reference fixed on the earth's surface does not, as there are apparent forces (centrifugal and Coriolis) that affect the motion of a particle relative to such a frame. Let $\mathbf{r}_i$ be the radius vector of the ith particle. Then the linear velocity of the ith particle is $\mathbf{v}_i = d\mathbf{r}_i/dt$. We now sum Eqs. (1) over i (over all the particles). The equations of motion (1) become

$$\frac{d^2}{dt^2}\sum_i m_i\mathbf{r}_i = \sum_i \mathbf{F}_i + \sum_{i,j(i\neq j)} \mathbf{f}_{ij}. \tag{2}$$

The first sum on the right is the total external force acting on the system of n particles, while the second sum vanishes—the terms cancel by pairs—because we have assumed Newton's third law, $\mathbf{f}_{ij} = -\mathbf{f}_{ji}$. We may simplify these equations of motion by introducing a vector $\mathbf{R}$ which is the mass-weighted average of the radii vectores of the particles. This means

$$\mathbf{R} = \sum m_i \frac{\mathbf{r}_i}{M}, \quad \text{where} \quad M = \sum m_i. \tag{3}$$

The vector $\mathbf{R}$ defines the *center of mass* or *center of gravity* of the system. Then Eq. (2) becomes

$$M\frac{d^2\mathbf{R}}{dt^2} = \sum_i \mathbf{F}_i \equiv \mathbf{F}, \tag{4}$$

where $\mathbf{F}$ is the sum (resultant) of all the external forces acting on the system. We see that the internal forces acting on each particle have no effect on the system. It is clear that the center of mass moves as if the resultant external force were acting on the entire mass M concentrated at the center of gravity. Examples occur in rocket and jet propulsion as well as in many other situations. For example, the motion of the fragments of an exploding shell is such that the center of mass of the fragments travels as if the shell were still in a single piece. Note that we must neglect air resistance, since the drag force due to air resistance depends on the shape and speed of the object. In this analysis we assume that the mass is independent of time.

Let **P** be the total linear momentum of the object. We then obtain

$$\mathbf{P} = \sum m_i \frac{d\mathbf{r}_i}{dt} = M\frac{d\mathbf{R}}{dt}. \tag{5}$$

Taking the time derivative of Eq. (5), we obtain the vector equation of motion for our system of particles:

$$\frac{d\mathbf{P}}{dt} \equiv \dot{\mathbf{P}} = \mathbf{F}. \tag{6}$$

1.2 Angular Momentum

We define the *angular momentum* of a particle as $\mathbf{L}_i = \mathbf{r}_i \times \mathbf{p}_i$. The total angular momentum of the system of particles is then

$$\mathbf{L} = \sum_i \mathbf{L}_i = \sum_i \mathbf{r}_i \times \mathbf{p}_i = \sum_i \mathbf{r}_i \times m_i \frac{d\mathbf{r}_i}{dt}. \tag{7}$$

Now set $\mathbf{r}_i = \mathbf{R} + \boldsymbol{\rho}_i$; it is seen that $\boldsymbol{\rho}_i$ is the radius vector of the ith particle relative to the center of mass. If we substitute this into Eq. (7), we obtain

$$\begin{aligned}\mathbf{L} &= \sum_i (\mathbf{R} + \boldsymbol{\rho}_i) \times m_i \frac{d}{dt}(\mathbf{R} + \boldsymbol{\rho}_i)\\ &= \sum_i (\mathbf{R} \times m_i\dot{\mathbf{R}} + \mathbf{R} \times m_i\dot{\boldsymbol{\rho}}_i\\ &\quad + \boldsymbol{\rho}_i \times m_i\dot{\mathbf{R}} + \boldsymbol{\rho}_i \times m_i\dot{\boldsymbol{\rho}}).\end{aligned} \tag{8}$$

But since $\boldsymbol{\rho}_i = \mathbf{R} - \mathbf{r}_i$, then

$$\begin{aligned}\sum_i m_i\boldsymbol{\rho}_i &= \sum_i m_i\mathbf{R} - \sum_i m_i\mathbf{r}_i\\ &= M\mathbf{R} - M\mathbf{R} \equiv 0,\end{aligned} \tag{9}$$

and so the second and third terms in the last form of Eq. (8) vanish. If we examine the forms of the remaining two terms, we have the important result that *the angular momentum of a system of particles about an arbitrary origin is the sum of the angular momentum about their center of mass and the angular momentum of the system, considered as a single particle concentrated at the center of mass, about the arbitrary origin.*

Next consider a force **F** acting at a point whose radius vector from some arbitrary origin is **R**. We define the moment of the force about the origin, or the *torque* about the origin, as **R** × **F**. Returning to our system of particles, we can find the total torque **N** acting on it as

$$\begin{aligned}\mathbf{N} &= \sum_i \mathbf{r}_i \times \left(\mathbf{F}_i + \sum_j \mathbf{f}_{ij}\right)\\ &= \sum_i \mathbf{r}_i \times \mathbf{F}_i + \sum_{i,j} \mathbf{r}_i \times \mathbf{f}_{ij}.\end{aligned} \tag{10}$$

Now for every term $\mathbf{r}_i \times \mathbf{f}_{ij}$ in the double sum, there is a term $\mathbf{r}_j \times \mathbf{f}_{ji}$. We assume, as before, that the internal forces are equal and opposite, so that $\mathbf{f}_{ij} = -\mathbf{f}_{ji}$, and combine the two terms as $(\mathbf{r}_i - \mathbf{r}_j) \times \mathbf{f}_{ij}$. We now further assume that the internal forces are not only equal and opposite, but *central* forces, that is, that the forces between any two particles act along the line joining them. Then the terms in the double sum cancel by pairs, and we are left with

$$\mathbf{N} = \sum_i \mathbf{r}_i \times \mathbf{F}_i: \tag{11}$$

the total torque on a system whose internal forces are central forces obeying the law of action and reaction is due only to the external forces.

Let us now find the rate of change of angular momentum. We have

$$\begin{aligned}\frac{d\mathbf{L}}{dt} &= \frac{d}{dt}\sum_i \mathbf{r}_i \times \mathbf{p}_i\\ &= \sum_i (\mathbf{r}_i \times \dot{\mathbf{p}}_i + \dot{\mathbf{r}}_i \times \mathbf{p}_i).\end{aligned} \tag{12}$$

The last term vanishes, since $\mathbf{p}_i = m_i\mathbf{v}_i$ is parallel to $\dot{\mathbf{r}}_i$. In the remaining sum, we recognize that $d\mathbf{p}_i/dt$ is just the total force acting on the ith particle, and so the ith term in the sum is the torque on the ith particle, and the sum is the total torque on the system. We are left with

$$\frac{d\mathbf{L}}{dt} = \mathbf{N}: \tag{13}$$

the rate of change of the angular momentum of a system of particles is equal to the total torque exerted on it—which, as we have seen, is the torque due to the external forces alone.

1.3 Work and Energy

Consider a constant force **F** acting on an object that moves along a straight path segment described by the vector **S**. We define the *work* done by the force as $W = \mathbf{F} \cdot \mathbf{S}$. We can easily generalize this to a force varying in magnitude. We take $\mathbf{n}$ as the unit vector along $\mathbf{S}$, so that $\mathbf{S} = S\mathbf{n}$. We let ds be an element of length along $\mathbf{S}$. Then we write $W = \int \mathbf{F} \cdot \mathbf{n} ds$. Note that in either case, it is only the component of $\mathbf{F}$ along $\mathbf{S}$ that is significant. Note also that there is no consideration of time in the discussion. No matter how long the motion takes, the work done is the same. Moreover, as long as we can relate a time variation in $\mathbf{F}$ to a position variation along $\mathbf{S}$, the time variation of $\mathbf{F}$ is also immaterial. The further generalization to a curvilinear path is straightforward, involving only the concept of a line integral. The expression is

$$W = \oint \mathbf{F} \cdot d\mathbf{r}. \tag{14}$$

In all of these cases, it is possible for the work done to be negative. This is interpreted as (positive) work being done *by* the system on the agent exerting the force.

The concept of energy involves some arbitrariness. The energy of a system is always defined in terms of some reference state. More precisely, we define the *difference* in energy between two states of a system as the work that is done to change the system from the initial state to the final (that is, if positive work is done to change from the initial state to the final, the system has greater energy in the final state than in the initial).

Let us calculate the work done in giving a body of mass m a velocity $\mathbf{v}$. We have

$$\begin{aligned} W &= \int \mathbf{F} \cdot d\mathbf{r} = \int m \frac{d\mathbf{v}}{dt} \cdot d\mathbf{r} \\ &= \int m d\mathbf{v} \cdot \frac{d\mathbf{r}}{dt} = m \int \mathbf{v} \cdot d\mathbf{v} = \tfrac{1}{2} mv^2. \end{aligned} \tag{15}$$

Since this is the energy due to motion, it is called *kinetic energy*.

1.4 Conservation Laws

We have seen [Eq. (5)] that the rate of change of the total momentum of a system is equal to the sum of the external forces acting on it, provided that the internal forces obey the law of action and reaction. (It must be pointed out that the forces between moving charged particles do not always obey that law.) We immediately deduce the *law of conservation of linear momentum*: If the external forces on a system vanish, the linear momentum is constant.

Likewise, we have seen [Eq. (13)] that the rate of change of angular momentum is equal to the external torque exerted on the system, provided the internal forces are central and obey the law of action and reaction. This leads to the *law of conservation of angular momentum*: If the external torques on a system are zero, the angular momentum is constant.

The case of energy is a bit more complicated. Equation (15) tells us that the work done on a body is equal to the increase of kinetic energy. If we denote kinetic energy by T, we have

$$W_{12} = T_2 - T_1. \tag{16}$$

Now suppose that the force varies with position in such a way that the work done is independent of the shape of the path between the two end points. This means that W_{12} must be expressible in terms of a function that depends only on the end points. If we reduce the path to a differential, then we must have

$$dW = \mathbf{F} \cdot d\mathbf{r} = -dV, \tag{17}$$

where the minus sign is chosen for future convenience. Such a force field is called *conservative*. From the last two forms of this relationship, we find that the component of $\mathbf{F}$ along $\mathbf{s}$, F_s, is given by

$$F_s = -\frac{\partial V}{\partial s}. \tag{18}$$

This means that

$$\mathbf{F} = -\nabla V. \tag{19}$$

Clearly, the function V is arbitrary to within an additional constant, which is in harmony with our discussion of energy as being determined only as a difference. The function V is called the *potential energy*.

Combining Eqs. (16) and (19) gives, after some rearrangement,

$$T_1 + V_1 = T_2 + V_2: \tag{20}$$

In a conservative force field, the sum of the kinetic and potential energy is constant. This is what is usually meant by the conservation of energy.

There are many interesting problems in particle mechanics: the two-body problem in a gravitational field, the theory of vibration of a system of coupled particles, the oscillations of a pendulum, etc. We leave it to the reader to supplement the foregoing concepts with books on elementary mechanics such as Synge and Griffith (1959).

1.5 Generalized Coordinates

We take as our mathematical model of a dynamical system a discrete system of particles characterized by a finite number of *degrees of freedom.* The number of degrees of freedom is defined as the number of independent coordinates necessary to describe the positions of all the particles of the system. A system composed of N particles has $3N$ degrees of freedom in three-dimensional space. A rigid body has six degrees of freedom: three to define the position of its center of mass and three angles to define the orientation of the body (see Sec. 2.1.1). These independent coordinates are called *generalized coordinates* and are expressed as a set $\mathbf{q} = (q_1,q_2,\ldots,q_n)$, where $n = 3N$ for our set of N particles. If a system of N particles has k constraints (see Sec. 1.6), then the number of generalized coordinates is $3N - k$. It is important to note that these generalized coordinates need not have the significance of Cartesian coordinates, nor even of space coordinates of any kind. We discuss later situations in which some of the generalized coordinates are momenta or even energies.

Corresponding to the generalized coordinates are generalized velocities. The set of generalized velocities is $\dot{\mathbf{q}} = (\dot{q}_1,\dot{q}_2,\ldots,\dot{q}_n)$, and the generalized momentum is given by the set $\mathbf{p} = (\dot{p}_1,\dot{p}_2,\ldots,\dot{p}_n)$, $p_i = m_i\dot{q}_i$, $i = 1, 2, \ldots$

Suppose a system of particles is described by n generalized coordinates. Then the instantaneous position or configuration of the system is described by these n coordinates in an n-dimensional hyperspace called *configuration space*. This means that at each instant of time, the system is represented by a point in configuration space. As the time parameter increases, the position of the entire system changes so that the point traces a curve called the path of motion of the system. If, as previously assumed, the system consists of N particles, and there are $n = 3N$ degrees of freedom, the configuration space has $3N$ dimensions. A trajectory in configuration space does not have any resemblance to the path in actual space—not necessarily even for a single particle moving in three dimensions so that its configuration space also has three dimensions, except when the three generalized coordinates happen to coincide with the three Cartesian coordinates.

A more useful concept is *phase space*. This is a hyperspace consisting of n generalized coordinates and n generalized momenta for a system with n degrees of freedom. A given set of initial conditions (given by prescribing the initial generalized positions and momenta of all the particles) defines the configuration of the dynamical system at $t = 0$. As time increases, the motion or trajectory of the system unfolds so that it is completely described by its generalized coordinates and momenta. Phase space is essential in Hamilton's description of a dynamical system (see Sec. 4).

1.6 Constraints

A *constraint*, roughly speaking, is a restriction on the motion of the particles constituting the system. It is important to understand the nature of constraints. If a system has n coordinates, then the number k of constraints must satisfy $0 < k < n$; for if $k = 0$ then the system is free, and if $k = n$ then the system is not dynamic. We shall now classify constraints and give examples.

For a single particle in three-dimensional space moving on a surface, there is one constraint, since the equation of the surface on which the particle moves ties up one degree of freedom. This is most easily seen if the surface is taken to be one of the coordinate planes, since then the equation of the surface fixes one coordinate. In either this or the more general case, the constraint is called *holonomic*, which means that it is defined by an equation connecting the coordinates. A rigid body consisting of N particles is another simple example of a system having holonomic coordinates. Since there is no relative motion between the particles, the constraints are given by

$$(\mathbf{r}_i - \mathbf{r}_j)^2 = C_{ij}, \quad i \neq j, \quad i, j = 1, 2, \ldots, n, \tag{21}$$

where the C_{ij} are constants.

In general, constraints are holonomic if they can be expressed as equations connecting the coordinates of the particles (and possibly time). The general form of the equations for N particles is

$$f(r_1, r_2, \ldots, r_N) = 0. \tag{22}$$

A *nonholonomic* constraint is one that cannot be expressed by an algebraic equation. In some cases, it can be expressed as a relationship between the derivatives (or differentials) of the generalized coordinates, and possibly time (if it is *rheonomic*—see below). For example, a vertical disk rolls on an inclined plane with a frictional force that prevents slipping (see Fig. 1). The constraint is given by a simple relationship

$$ds = r d\theta \tag{23}$$

between the differential of the arc length and the differential angle of rotation. A nonholonomic constraint may also be expressed by an inequality. For example, a particle is placed on top of a large sphere of radius R and given an initial velocity. The particle and sphere are assumed to be in a gravitational field, so that the particle will slide down the surface of the sphere for a time and then fall off. The condition for falling off is that $r > R$ where r is the magnitude of the radius vector of the particle as measured from the center of the sphere. The constraint is that the radius vector must satisfy the inequality $r \geq R$.

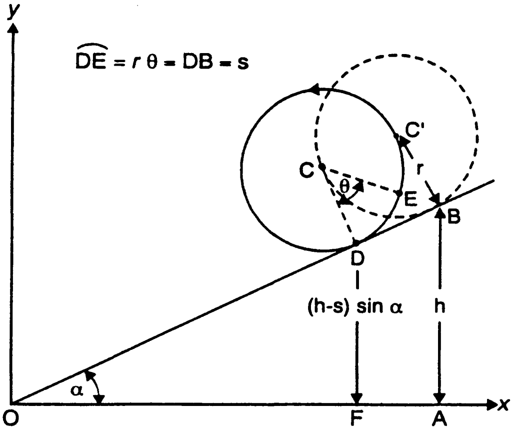

FIG. 1. Hoop rolling down an inclined plane.

Constraints are also classified as *scleronomic*, or time independent, and *rheonomic*, or time dependent. The examples given so far involve scleronomic constraints since surfaces are assumed to be stationary. If a bead slides along a moving wire, then the constraint is rheonomic.

Suppose we have a holonomic rheonomic constraint whose equation is

$$f(q_1, q_2, \ldots, q_n, t) = 0 \tag{24}$$

for a system with n degrees of freedom. This means that the generalized coordinate set $\mathbf{q}$ has n independent components. Since $df = 0$, this is equivalent to the differential equation

$$\sum_j \frac{\partial f}{\partial q_j} dq_j + \frac{\partial f}{\partial t} dt = 0. \tag{25}$$

If we set

$$\frac{\partial f}{\partial q_j} = a_{ij}, \quad \frac{\partial f}{\partial t} = a_{it}, \tag{26}$$

we get

$$\sum_j a_{ij} dq_j + a_{it} dt = 0 \tag{27}$$

for the ith constraint. If we have k constraints then there are k equations, for $i = 1, 2, \ldots, k$. These constraint equations give relations connecting the differentials dq and dt. They may also apply to nonholonomic constraints. If the time is constant during the displacements, then $dt = 0$. Therefore virtual displacements (see Sec. 1.7) must satisfy constraints of the form

$$\sum_j a_{ij} \delta q_j = 0, \quad i = 1, 2, \ldots, k, \tag{28}$$

where the symbol δ means a differential change keeping time constant.

Regardless of the nature of the constraints, if our system of n degrees of freedom has k constraints then the number of generalized coordinates is $n - k$.

1.7 The Principle of Virtual Work and D'Alembert's Principle

We next introduce the concept of *virtual displacements*. A virtual displacement is an arbitrary infinitesimal displacement of the configuration of a system of particles from its original state, consistent with the forces and constraints imposed on the system at a given instant. The important point is that *time does not change during a virtual displacement.* This is in contrast to an actual displacement, where time changes by an amount dt. Let $\delta\mathbf{r}_i$ be a virtual displacement of the ith particle. Note that the radius vector $\mathbf{r}$ is not a set of generalized coordinates, since it is subject to constraints. The actual displacement of the ith particle is $d\mathbf{r}_i$. Figure 2 shows plots of virtual and actual displacements of n particles in the (x,y) plane. The particles in the equilibrium configuration at t_0 are on the (fictitious) curve C_E. They are transformed to the curve C_V under a virtual displacement, and to the curve C_A under an actual displacement. The position of the ith particle in equilibrium is given by $\mathbf{P}_i$. Upon undergoing a virtual displacement $\delta\mathbf{r}_i$, it is transformed to the point $\mathbf{Q}_i$. If the particle at $\mathbf{P}_i$ undergoes an actual displacement $d\mathbf{r}_i$, it is transformed to the point $\mathbf{R}_i$. Clearly, these transformations hold for all the N particles. The particles on C_V are at t_0 and are consistent with the constraints acting on them. The particles on C_A are at time $t_0 + dt$.

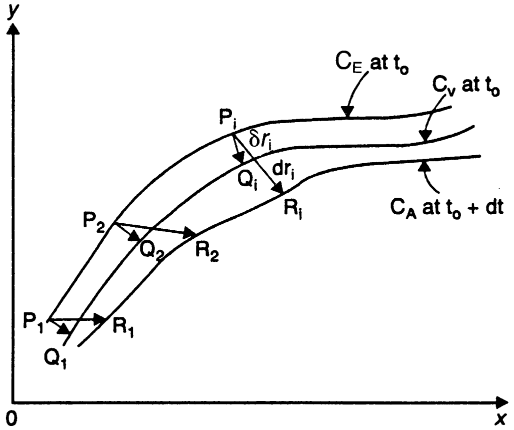

FIG. 2. Virtual and actual displacements of *n* particles in the (*x*,*y*) plane.

Virtual work is defined as the work done on a dynamical system due to the external forces acting on it when the system undergoes a virtual displacement (time independent) compatible with the constraints. We again consider a system of N particles in equilibrium. Then the resultant force on each particle vanishes so that $\mathbf{F}_i = 0$, $i = 1, 2, \ldots, N$. Let δW_i be the virtual work done on the ith particle by the force $\mathbf{F}_i$ producing a virtual displacement $\delta\mathbf{r}_i$. Then $\delta W_i = \mathbf{F}_i \cdot \delta\mathbf{r}_i$, and the condition that the system be in equilibrium is that the virtual work due to the resultant force $\mathbf{F}$ on the system will vanish:

$$\sum_{i=1}^{N} \mathbf{F}_i \cdot \delta\mathbf{r}_i = 0. \tag{29}$$

Let the resultant force on the ith particle be given by

$$\mathbf{F}_i = \mathbf{F}_i' + \mathbf{f}_i, \tag{30}$$

where $\mathbf{F}_i'$ is the applied force and $\mathbf{f}_i$ is the constraining force on the ith particle. If we restrict ourselves to a system where the virtual work done by the constraining forces is zero, then Eq. (30) reduces to the condition

$$\sum_{i} \mathbf{F}_i' \cdot \delta\mathbf{r}_i = 0. \tag{31}$$

This states that the condition for a system of n particles to be in equilibrium is that the virtual work of the applied forces on the system must vanish. This will not be true for a system subject to frictional or other dissipative forces. Virtual work must be distinguished from the *actual* work dW done by the forces acting during the time dt.

Since the principle of virtual work [Eq. (30)] assumes that the system is in equilibrium, we do not have a dynamical system. The question then arises: How can the principle of virtual work be manipulated to include dynamical systems? The answer was given by D'Alembert. He ingeniously wrote the equations of motion in the form

$$\mathbf{F}_i - \dot{\mathbf{p}}_i = \mathbf{0}, \quad i = 1, 2, \ldots, N. \tag{32}$$

This is a vector equation of motion for the ith particle. Note that the right-hand side of this equation is the zero force vector. The observation that the difference between the external and the inertial force on the ith particle is equal to the zero force has the advantage of treating the inertial force $\dot{\mathbf{p}}_i$ in the same manner as the external force $\mathbf{F}_i$. By this technique, D'Alembert essentially reduced a problem in dynamics into one in which the principle of virtual work can be applied.

Another way of stating the principle of virtual work is the following: A system of N particles in equilibrium under the action of external forces $\mathbf{F}$ and the "reversed inertial force" $-\dot{\mathbf{p}}_i$ has zero virtual work done on it. This is expressed by the equation

$$\sum_{i=1}^{N} (\mathbf{F}_i - \dot{\mathbf{p}}_i) \cdot \delta\mathbf{r}_i = 0. \tag{33}$$

Since we restrict ourselves to conservative systems, $\mathbf{F}$ is the applied force. Equation (33) is called D'Alembert's principle.

1.7.1 Transformation to Generalized Coordinates In carrying out our program of transforming the coordinate system from $\mathbf{r}$ to $\mathbf{q}$, we seek a coordinate transformation of the form

$$\mathbf{r}_i = \mathbf{r}_i(q_1, q_2, \ldots, q_{3N-k}), \quad i = 1, 2, \ldots, n \tag{34}$$

for a system of N particles subject to k constraints. Each vector $\mathbf{r}_i$ is to be obtained as a function of all the scalar q_i's. We now express the velocity of the ith particle $\mathbf{v}_i = d\mathbf{r}_i/dt \equiv \dot{\mathbf{r}}$ as a linear combination of the generalized velocity components $\dot{q}$. Expanding by the chain rule gives

$$\mathbf{v}_i = \sum_{j=1}^{3N-k} \frac{\partial \mathbf{r}_i}{\partial q_j} \dot{q}_j + \frac{\partial \mathbf{r}_i}{\partial t}, \quad i = 1, 2, \ldots, n. \tag{35}$$

Note that $\mathbf{v}$ is a vector while $\dot{q}$ is the jth component of the set $\dot{\mathbf{q}}$. It is clear that in each of the n equations of (35), we sum over all the scalar values of q. In a similar manner, the virtual displacement (vector) of the ith particle is expressed as a linear combination of the generalized virtual displacements (scalars):

$$\delta\mathbf{r}_i = \sum_{j=1}^{3N-k} \frac{\partial \mathbf{r}_i}{\partial q_j} \delta q_j. \tag{36}$$

In terms of the generalized virtual coordinates, the virtual work of the system becomes

$$\begin{aligned} \delta W &= \sum_i \mathbf{F}_i \cdot \delta\mathbf{r}_i = \sum_{i,j} \mathbf{F}_i \cdot \frac{\partial r_i}{\partial q_j} \delta q_j \\ &= \sum_i Q_i \delta q_i. \end{aligned} \tag{37}$$

Here Q_j is the *generalized force* on the jth particle, defined as

$$Q_j = \sum_i \mathbf{F}_i \cdot \frac{\partial \mathbf{r}_i}{\partial q_j}. \tag{38}$$

1.8 Lagrange's Form of the Equations of Motion

Having obtained the virtual work, generalized forces, etc., in terms of generalized coordinates, we are now in a position to derive Lagrange's form of the equations of motion, one of the cornerstones of analytical dynamics. We derive them from D'Alembert's principle, expressed in the form

$$\sum_i \dot{\mathbf{p}}_i \cdot \delta\mathbf{r}_i = \sum_j Q_j \delta q_j. \tag{39}$$

We perform the following series of calculations:

$$\begin{aligned} \sum_i \dot{\mathbf{p}}_i \cdot \delta\mathbf{r}_i &= \sum_i m_i \ddot{\mathbf{r}}_i \cdot \delta\mathbf{r}_i \\ &= \sum_i m_i \ddot{\mathbf{r}}_i \cdot \frac{\partial \mathbf{r}_i}{\partial q_j} \delta q_j \\ &= \sum_i \left[\frac{d}{dt}\left(m_i \dot{\mathbf{r}}_i \cdot \frac{\partial \mathbf{r}_i}{\partial q_i} \right) \right. \\ &\qquad \left. - m_i \dot{\mathbf{r}}_i \cdot \frac{d}{dt}\left(\frac{\partial \dot{\mathbf{r}}_i}{\partial q_j} \right) \right] \delta q_j. \end{aligned} \tag{40}$$

By interchanging the order of the operators d/dt and $\partial/\partial q_i$, we obtain the following two relationships:

$$\frac{d}{dt}\frac{\partial \mathbf{r}_i}{\partial q_j} = \frac{\partial \mathbf{v}_i}{\partial q_j}, \quad \frac{\partial \mathbf{r}_i}{\partial q_j} = \frac{\partial \mathbf{v}_i}{\partial \dot{q}_j}. \tag{41}$$

Using these expressions and the definition of the kinetic energy of the system, we obtain

$$\begin{aligned}\sum_i \dot{\mathbf{p}}_i \delta\mathbf{r}_i &= \sum_i \left[\frac{d}{dt}\left(m_i\mathbf{v}_i\cdot\frac{\partial\mathbf{v}_i}{\partial\dot{q}_j}\right) - m_i\mathbf{v}_i\cdot\frac{\partial\mathbf{v}_i}{\partial q_j}\right]\\ &= \sum_j \left[\frac{d}{dt}\left(\frac{\partial}{\partial\dot{q}_j}\right)\left(\sum_i \tfrac{1}{2}m_i v_i^2\right)\right.\\ &\qquad \left. - \frac{\partial}{\partial q_j}\sum_i \tfrac{1}{2}m_i v_i^2\right]\delta q_j\\ &= \sum_j \left[\frac{d}{dt}\left(\frac{\partial T}{\partial\dot{q}_j}\right) - \frac{\partial T}{\partial q_j}\right]\delta q_j. \end{aligned} \qquad (42)$$

From this expression we get

$$\sum_j \left[\frac{d}{dt}\left(\frac{\partial T}{\partial\dot{q}_j}\right) - \frac{\partial T}{\partial q_j} - Q_j\right]\delta q_j = 0. \qquad (43)$$

Since the virtual displacements are independent, the coefficient of each one must vanish. This yields

$$\frac{d}{dt}\left(\frac{\partial T}{\partial\dot{q}_j}\right) - \frac{\partial T}{\partial q_j} = Q_j, \quad j = 1, 2, \ldots, n. \qquad (44)$$

These are Lagrange's equations for T and the Q_j's in terms of the generalized coordinates and velocities.

In the case of a conservative system, the resultant force $\mathbf{F}_i$ on each particle is derivable from the potential energy V, so that

$$\mathbf{F}_i = -\nabla_i V, \qquad (45)$$

where ∇_i is the gradient operator on the ith particle. The generalized force on the jth particle becomes

$$Q_j = \sum_i \mathbf{F}_i\cdot\frac{\partial\mathbf{r}_i}{\partial q_j} = -\sum_i \nabla_i V\cdot\frac{\partial\mathbf{r}_i}{\partial q_j} = -\frac{\partial V}{\partial q_j}. \qquad (46)$$

Substituting this result into Eq. (44) and rearranging slightly, we have

$$\frac{d}{dt}\left(\frac{\partial T}{\partial\dot{q}_j}\right) - \frac{\partial}{\partial q_j}(T - V) = 0, \quad j = 1, 2, \ldots, n. \qquad (47)$$

Now define the Lagrangian L as $T - V$. Since V is independent of the $\dot{q}_j$, it is true that $\partial T/\partial q_j = \partial L/\partial q_j$, and Eq. (47) becomes

$$\frac{d}{dt}\left(\frac{\partial L}{\partial\dot{q}_j}\right) - \frac{\partial L}{\partial q_j} = 0, \quad j = 1, 2, \ldots, n. \qquad (48)$$

These are Lagrange's form of the equations of motion. They can be seen as generalizations of Newton's equations as given by his second law, if we make the following interpretation: The term $\partial L/\partial\dot{q}_i$ represents the generalized momentum of the ith particle, and $\partial L/\partial q_i$ represents the "generalized force" derivable from the potential V.

2. RIGID-BODY MECHANICS

We define a *rigid body* as a continuous and bounded distribution of particles having the property that the distance between any two of the particles is invariant with time. The most general motion of a rigid body is a combination of a pure rotation about an arbitrary axis and the pure translational motion of the center of mass of the body with respect to an inertial coordinate system. This is known as *Chasles' theorem*. Rather than a detailed proof, we present the following intuitive reasoning: Under rotation, one point in the body is fixed, and the rotational motion is defined in terms of the three degrees of freedom, usually described in terms of the Euler angles (see Sec. 2.1.1). Allowing the one point to move (removing the constraint) introduces translational motion represented by three more degrees of freedom. This gives the most general motion of the rigid body.

Since the motion of the center of mass of a rigid body is the same as if all the mass were concentrated at that point and all the external forces applied there, the translational motion of the center of mass of a rigid body is determined by the same methods used in earlier sections to discuss particle motion; we will therefore drop it from consideration and treat only the rotation of a body having one point fixed.

2.1 Rotation of a Body about a Fixed Point

To describe the general motion of a rigid body we need (as reference) a space-fixed coordinate system. In general, this will be a three-dimensional inertial system. We also need a coordinate system fixed in the body.

Figure 3 shows these two coordinate systems in terms of Cartesian coordinates; for convenience, we take them to have a common origin fixed in the body. We use a symmetric notation to relate to matrix manipulation. We denote the space-fixed system by $\mathbf{x} = (x_1,x_2,x_3)$ and the body-fixed system by $\mathbf{x}' = (x'_1, x'_2,x'_3)$.

We further take the common origin of the two systems to be the fixed point of the body. We then determine the rotation of the body by rotating the body-fixed coordinates through an arbitrary angle θ with respect to the space-fixed coordinates. Let $(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3)$ be unit vectors along the (x_1,x_2,x_3) axes, and let $(\mathbf{e}'_1,\mathbf{e}'_2,\mathbf{e}'_3)$ be the unit vectors in the body-fixed system. Figure 3 shows the situation. It is convenient to introduce direction cosines describing the primed coordinate system with respect to the unprimed coordinate systems. There are nine direction cosines of the form l_{ij} defined by

$$l_{ij} = \cos(\mathbf{e}'_i,\mathbf{e}_j) = \mathbf{e}'_i \cdot \mathbf{e}_j, \quad i,j = 1, 2, 3. \tag{49}$$

l_{ij} is the direction cosine of the ith body-fixed axis with respect to the jth space-fixed axis. Let $\mathbf{r}$ be the radius vector from the common origin to an arbitrary point P in the body with respect to the unprimed axes, and let $\mathbf{r}'$ be the radius vector to the same point with respect to the primed axes. A rotation of the body is represented by the rotation of the vector $\mathbf{r}$ into $\mathbf{r}'$ by the matrix transformation

$$\mathbf{r}' = \mathbf{A}\mathbf{r}, \tag{50}$$

where $\mathbf{r}$ and $\mathbf{r}'$ are now regarded as column vectors and the rotation matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{pmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{pmatrix}. \tag{51}$$

Equation (50) is thus three equations, of which the ith ($i = 1,2,3$) is

$$x'_i = \sum_{j=1}^{3} l_{ij}x_j. \tag{52}$$

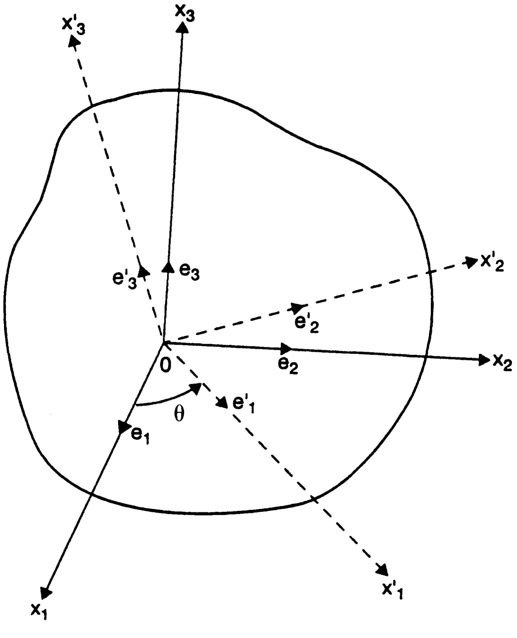

FIG. 3. Body-fixed coordinates rotated about a fixed point relative to space-fixed coordinates.

The matrix $\mathbf{A}$ has the property that the scalar product of any column, considered as a column vector, and any other column is zero, and the scalar product of any column with itself is 1; similar statements hold for the rows. This is expressed mathematically as

$$\sum_k l_{ik}l_{jk} = \delta_{ij}, \quad \sum_k l_{ki}l_{kj} = \delta_{ij}, \tag{53}$$

where the symbol δ_{ij}, known as the *Kronecker delta*, equals 0 if $i \neq j$ and 1 if $i = j$. Such a matrix is called *orthonormal*. From these properties of $\mathbf{A}$, it is easily seen that

$$\mathbf{A}^{-1} = \mathbf{A}^{\mathrm{T}}, \tag{54}$$

where $\mathbf{A}^{-1}$ is the inverse matrix and $\mathbf{A}^{\mathrm{T}}$ is the transpose matrix. Premultiplying both sides of Eq. (50) by $\mathbf{A}^{-1}$, we obtain the inverse mapping $\mathbf{r}' \to \mathbf{r}$, given by

$$\mathbf{A}^{-1}\mathbf{r}' = \mathbf{A}^{\mathrm{T}}\mathbf{r}' = \mathbf{r}. \tag{55}$$

2.1.1 Eulerian Angles

Since the direction cosines are not linearly independent, we cannot use $\mathbf{A}$ to define the generalized coordinates associated with a rotating body. We need a set of three independent angles to define an arbitrary rotation. Euler attacked this problem by defining such a set of angles, which are aptly called the *Eulerian angles of rotation* and which represent three independent modes of rotation. Figure 4 shows the space-fixed axes (x,y,z) with z vertical and (x,y) in the horizontal plane. The body-fixed axes (x',y',z') are also shown. Here z' makes the angle θ with z. (x',y') lie in the tilted plane that cuts the horizontal plane at the line ON, called the *line of nodes*. Let us assume the body to be a symmetrical top spinning about the z' axis, its axis of symmetry. Varying only ψ means that the body spins about this axis, and so ψ is called the angle of spin. Varying only θ rotates the body about the line of nodes; this mode of rotation is called *nutation*, and θ is called the angle of inclination. Varying only ϕ causes the body to rotate about the z axis, the mode called *precession*; ϕ is called the azimuth angle. The Eulerian angles are clearly independent, since each angle describes a rotation about an axis independent of the others. They are used in Sec. 2.4 as generalized coordinates in Lagrange's equations to develop Euler's equations of motion.

2.2 Moment of Inertia

Let us return to consideration of the rigid body as a system of discrete particles. Suppose that the body moves with one point, taken as the origin, fixed so that it rotates. We have defined the angular momentum as

$$\mathbf{L} = \sum_i m_i \mathbf{r}_i \times \mathbf{v}_i, \tag{56}$$

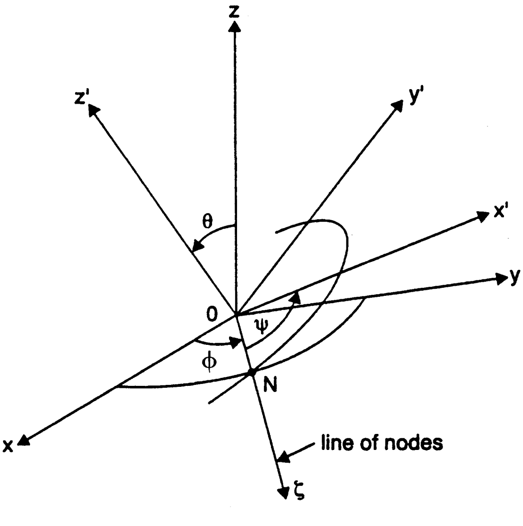

FIG. 4. Euler's angles for a symmetric body.

where $\mathbf{r}_i$ and $\mathbf{v}_i$ are the radius vector and velocity relative to the origin. In rotational motion, the body possesses *angular* velocity $\boldsymbol{\omega}$, which is related to the linear velocity of each particle by

$$\mathbf{v}_i = \boldsymbol{\omega} \times \mathbf{r}_i. \tag{57}$$

If we substitute this into Eq. (56) and expand the vector triple product, we have

$$\mathbf{L} = \sum_i m_i[\boldsymbol{\omega} r_i^2 - \mathbf{r}_i(\mathbf{r}_i \cdot \boldsymbol{\omega})]. \tag{58}$$

This shows that $\mathbf{L}$ is related to $\boldsymbol{\omega}$ by a linear transformation. We can see this more clearly by expanding Eq. (58). The x component of $\mathbf{L}$ is

$$\begin{aligned} L_x &= \omega_x \sum_i m_i(y_i^2 + z_i^2) + \omega_y \sum_i m_i x_i y_i \\ &\quad + \omega_z \sum_i m_i x_i z_i \\ &= I_{xx}\omega_x + I_{xy}\omega_y + I_{xz}\omega_z. \end{aligned} \tag{59}$$

The other two components of $\mathbf{L}$ can be obtained by cyclic permutation. The matrix equation showing $\mathbf{L}$ as a linear combination of the components of $\boldsymbol{\omega}$ is

$$\begin{pmatrix} L_x \\ L_y \\ L_z \end{pmatrix} = \begin{pmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{pmatrix} \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix} = \mathbf{L} = \mathbf{I}\boldsymbol{\omega}. \tag{60}$$

The matrix $\mathbf{I}$ represents the *inertia tensor*. The diagonal elements are known as the *moments of inertia* about the x, y, and z axes, respectively; the off-diagonal elements are the *products of inertia*. $\mathbf{I}$ is a second-rank symmetric tensor.

We now consider a rigid body composed of a continuous distribution of mass. Let $dm = \rho d\tau$ be the element of mass in a differential element of volume $d\tau$, where ρ is the (continuous) density of the body. We now generalize the sums in Eq. (58) to integrals over the volume occupied by the body to obtain the elements of the inertial tensor for a continuous mass distribution. The result can be expressed concisely as

$$I_{ij} = \int_V (r^2\delta_{ij} - x_i x_j)\rho d\tau. \tag{61}$$

2.3 Rotational Kinetic Energy

We now show that the kinetic energy of rotation of a rigid body depends on the inertia tensor **I** and is a quadratic form in ω. For simplicity, the body is considered to be a system of discrete particles. We use the following facts: The velocity of the ith particle is $\mathbf{v}_i = \boldsymbol{\omega} \times \mathbf{r}_i$, and $\mathbf{L} = \mathbf{I}\boldsymbol{\omega}$. The kinetic energy T becomes

$$\begin{aligned} T &= \tfrac{1}{2}\Sigma m_i \mathbf{v}_i \cdot (\boldsymbol{\omega} \times \mathbf{r}_i) = \tfrac{1}{2}\boldsymbol{\omega} \cdot \Sigma m_i(\mathbf{r}_i \times \mathbf{v}_i) \\ &= \tfrac{1}{2}\boldsymbol{\omega} \cdot \mathbf{L} = \tfrac{1}{2}\boldsymbol{\omega} \cdot \mathbf{I}\boldsymbol{\omega}. \end{aligned} \tag{62}$$

It is clear that T is a quadratic form in $(\omega_x,\omega_y,\omega_z)$ since $\boldsymbol{\omega} \cdot \mathbf{I}\boldsymbol{\omega}$ is the scalar product of two vectors. We show this quadratic form by expanding the last form of Eq. (62) to get

$$\begin{aligned} 2T = I_{xx}\omega_x^2 + I_{yy}\omega_y^2 + I_{zz}\omega_x^2 + 2I_{xy}\omega_x\omega_y \\ + 2I_{yz}\omega_y\omega_z + 2I_{zx}\omega_z\omega_x. \end{aligned} \tag{63}$$

A standard result of linear algebra (Birkhoff and Maclane, 1977; Noble and Daniels, 1977) assures us that since **I** is symmetric, a coordinate system can always be found in which it is diagonal; that is, $I_{xy} = I_{yz} = I_{zx} = 0$. The axes of this coordinate system are called *principal axes* of the body, and the transformation to this system from the original one is a principal-axis transformation.

2.4 Euler's Equations of Motion for Rigid-Body Rotation

For a rigid body undergoing rotation with one point fixed, we perform a principal-axis transformation. The moments of inertia along the principal axes are (I_1,I_2,I_3), and the angular velocity has components $(\omega_1,\omega_2,\omega_3)$ along these axes. The fundamental vector equation of motion for a rotating body is

$$\frac{d\mathbf{L}}{dt} = \mathbf{N}, \tag{64}$$

This equation only holds for an inertial system of coordinates. To obtain an equation valid for a rotating system, we use the following relationship between the time derivative operator in an inertial system and that in a rotating system:

$$\left.\frac{d}{dt}\right|_{\text{space}} = \left.\frac{d}{dt}\right|_{\text{body}} + \boldsymbol{\omega}\times. \tag{65}$$

Operating on **L** with this operator, we get

$$\left.\frac{d\mathbf{L}}{dt}\right|_{\text{space}} = \left.\frac{d\mathbf{L}}{dt}\right|_{\text{body}} + \boldsymbol{\omega} \times \mathbf{L}. \tag{66}$$

The x component of this equation (along the principal axis) is

$$\frac{dL_1}{dt} + \omega_2 L_3 - \omega_3 L_2 = N_1. \tag{67}$$

However, the components of **L** along the principal axes are proportional to the corresponding components of $\boldsymbol{\omega}$. We therefore get

$$I_1\dot{\omega}_1 - \omega_2\omega_3(I_2 - I_3) = N_1. \tag{68}$$

This is the component of Euler's equations of motion along the principal x axis. The other two equations are obtained by cyclic permutation of the indices. Note the coupling between the components of $\boldsymbol{\omega}$. This has the consequence that even if the motion starts with ω_x, say, equal to zero, it will not remain so if $I_2 \neq I_3$.

2.4.1 Torque-Free Motion of the Earth

As an example of the use of Euler's equations, we discuss the rotation of the earth about its axis of symmetry, neglecting torques due to the sun or moon. The origin is chosen at the center of mass. Because the earth is an oblate spheroid, the principal moments of inertia are (I_1,I_1,I_3), where I_3 is taken along the axis of symmetry (north–south poles). Euler's equations become

$$\begin{aligned} &I_1\dot{\omega}_1 + \omega_2\omega_3(I_3 - I_1) = 0, \\ &I_1\dot{\omega}_2 + \omega_3\omega_1(I_1 - I_3) = 0, \\ &I_3\dot{\omega}_3 = 0. \end{aligned} \tag{69}$$

Integrating the last of these equations tells us that the component ω_3 of angular velocity along the axis of symmetry is constant. If we let $\alpha = \omega_3(I_3 - I_1)/I_1$, differentiate the first equation, and substitute from the second for $\dot{\omega}$, the solution becomes $\omega_1 = a \cos(\alpha t + \phi)$ where a is the amplitude and ϕ the phase angle. These constants are arbitrary since they depend on the earth's position and velocity at the arbitrary time from which t is counted. The resultant angular velocity becomes $\omega = (a^2 + \omega_3^2)^{1/2}$, which is constant. Figure 5 shows a circle of radius a centered at the origin O in the equatorial plane of the earth. ω_3 points

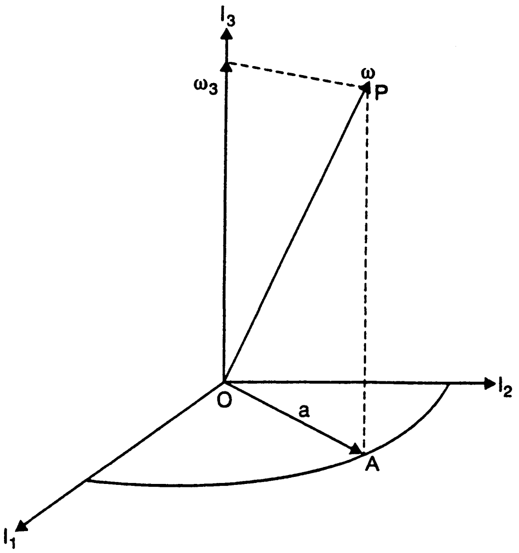

FIG. 5. Precession of the angular velocity about the axis of symmetry in force-free motion of the earth as a rotating rigid body.

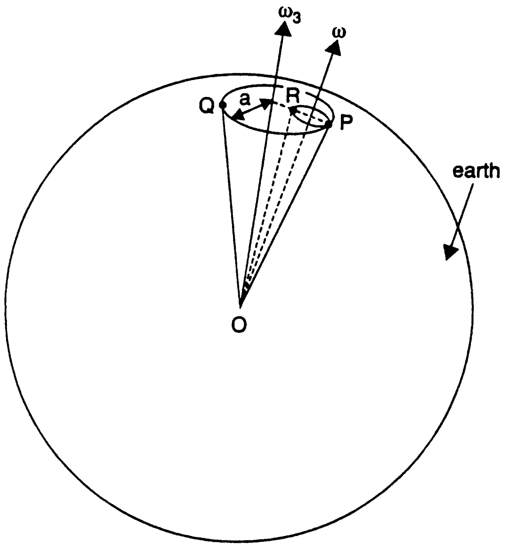

FIG. 6. Space- and body-fixed cones for torque-free rotation of the earth. The cone *OPQ* fixed in the earth rolls on the cone *OPR* whose instantaneous axis of rotation ω rotates about ω_3.

from the origin to the north pole (the axis of symmetry). From any point *A* on the circle, we draw a vertical vector *AP* of length equal to ω_3. The vector *OP* is then the vector $\boldsymbol{\omega}$ at a particular time. Clearly, *P* rotates about the axis of symmetry in a circular path. Figure 6 shows the earth as an oblate spheroid. A cone *OPQ* is drawn in the earth with the earth's axis of symmetry as its axis and a circle of radius *a* on the surface centered at the north pole. This cone is clearly body fixed. The figure also shows a cone *OPR* whose axis is $\boldsymbol{\omega}$ rotating with respect to the body-fixed cone as $\boldsymbol{\omega}$ rotates about the origin. In addition, *P* rotates on the circle of radius *a*. For this reason, $\boldsymbol{\omega}$ is called the axis of rotation.

Since $\omega_3 = 2\pi/\text{day}$, the period of rotation of $\boldsymbol{\omega}$ is given by $\tau = 2\pi/\alpha = (2\pi/\omega_3)I_1/(I_3 - I_1)$, which turns out to be about 300 days. It is called the Euler period, but it is not observed. However, a deeper analysis, taking into account the earth's imperfect rigidity, gives a period of 427 days called the Chandler period.

2.5 Motion of a Top

An analysis of the spinning top is important not only for its academic importance in the study of complicated rotational motion, but in its technical applications to gyroscopic motion used in navigational instruments. Top theory was also very important in contributing to military research on spinning shells during World War II.

We now discuss the motion of a heavy symmetrical top with one point (the origin) fixed, from an elementary viewpoint. Frictional forces are not considered. Referring to Fig. 7, z' is the axis of the top, which will *spin* about its axis (varying ψ, angle of spin), *nutate* (varying θ, angle of inclination), and *precess* (varying ϕ, the azimuth angle) about the fixed point *O*. *O'* is the center of mass of the top, a distance *l* from the origin. The weight of the top is *Mg*, acting at *O'* to produce the torque **N**, which acts along the line of nodes *ON*: $\mathbf{N} = \boldsymbol{\rho} \times M\mathbf{g}$, where $\boldsymbol{\rho}$ is the vector *AB* normal to the top axis. There is no component of **N** along the vertical axis z or the axis of the top z'. Therefore, the components of angular momentum along these axes must be constant in time.

We now derive Lagrange's equations for the rotating top. I_3 is the longitudinal component of the moment of inertia (along the axis of symmetry of the top), and $I_2 = I_1$ are the equal transverse moments of inertia. In terms of Euler's angles, the kinetic energy

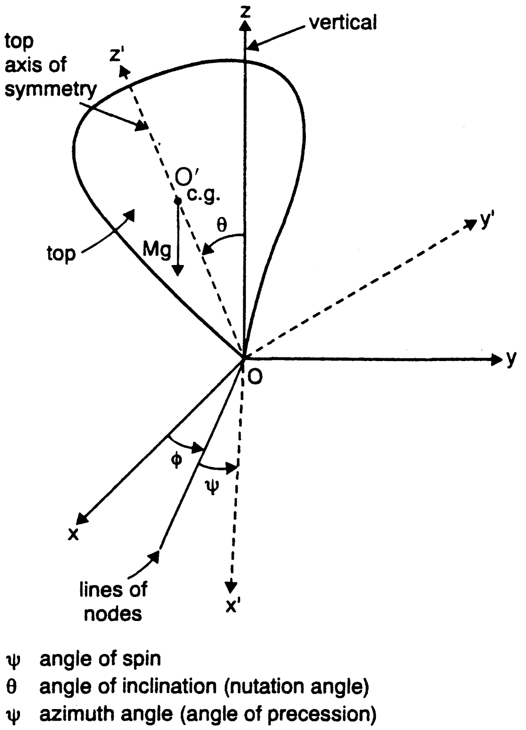

FIG. 7. Symmetrical top showing Euler's angles.

becomes

$$T = \tfrac{1}{2}I_1(\dot\theta^2 + \dot\phi^2 \sin^2 \theta) + \tfrac{1}{2}I_3(\dot\psi + \dot\phi \cos\theta)^2. \tag{70}$$

The potential energy is $V = Mgl \cos\theta$. The Lagrangian becomes

$$\begin{aligned} L &= \tfrac{1}{2}I_1(\dot\theta^2 + \dot\phi^2 \sin^2\theta) \\ &\quad + \tfrac{1}{2}I_3(\dot\psi + \dot\phi \cos\theta)^2 - Mgl \cos\theta. \end{aligned} \tag{71}$$

Since ϕ and ψ do not appear in the Lagrangian, they are called *cyclic* coordinates, and the corresponding generalized momenta p_ϕ and p_ψ are constant in time, as already shown by more elementary considerations. The two first integrals of the motion are then obtained as

$$\begin{aligned} p_\phi &= \frac{\partial L}{\partial \dot\phi} = (I_1 \sin^2\theta + I_3 \cos^2\theta)\dot\phi + I_3\dot\psi \cos\theta \\ &\equiv I_1 a, \end{aligned} \tag{72}$$

and

$$p_\psi = \frac{\partial L}{\partial \dot\psi} = I_3(\dot\psi + \dot\phi \cos\theta) = I_3\omega_3 \equiv I_1 b. \tag{73}$$

(We have expressed the two constants as multiples a and b of I_1.) An additional integral is furnished by invoking the conservation of energy:

$$\begin{aligned} E &= T + V \\ &= \tfrac{1}{2}I_1(\dot\theta^2 + \dot\phi^2 \sin^2\theta) + \tfrac{1}{2}I_3\omega_{z'}^2 + Mgl \cos\theta. \end{aligned} \tag{74}$$

By manipulating Eqs. (72) and (73), we eliminate $\dot\psi$ and obtain

$$\dot\phi = \frac{a - b \cos\theta}{\sin^2\theta}. \tag{75}$$

Substituting Eq. (75) into Eq. (74) results in

$$\dot\psi = \frac{I_1 b}{I_3} - \cos\theta \frac{a - b \cos\theta}{\sin^2\theta}. \tag{76}$$

Equations (75) and (76) are then used, along with the energy equation, to obtain a differential equation involving θ. We now introduce the constants

$$E' = E - \tfrac{1}{2}I_3\omega_z^2, \quad \alpha = \frac{2E'}{I_1}, \quad \beta = \frac{2Mgl}{I_1}. \tag{77}$$

If we now make the change of variable $x = \cos\theta$ and use Eqs. (79), after a little manipulation the energy equation (74) becomes

$$\dot x^2 = (1 - x^2)(\alpha - \beta x) - (a - bx)^2. \tag{78}$$

Solving for t, we obtain the following quadrature in the form of an elliptic integral:

$$t = \int_{x(0)}^{x(t)} \frac{dx}{[(1 - x^2)(\alpha - \beta x) - (a - bx)^2]^{1/2}}. \tag{79}$$

Using Eqs. (74) and (75), we may then also obtain ϕ and ψ in terms of quadratures.

However, it is not necessary to evaluate these integrals to obtain the nature of the nutational motion. We study the right-hand side of Eq. (78), which we denote by $f(x)$. This is a cubic polynomial in $x = \cos\theta$. The roots of this polynomial furnish the turning angles, those values of θ for which $\dot\theta$ changes sign. Figure 8 shows a rough plot of $f(x)$ vs x. The roots are (x_1,x_2,x_3), where x_3 does not correspond to a real angle. The nutational motion

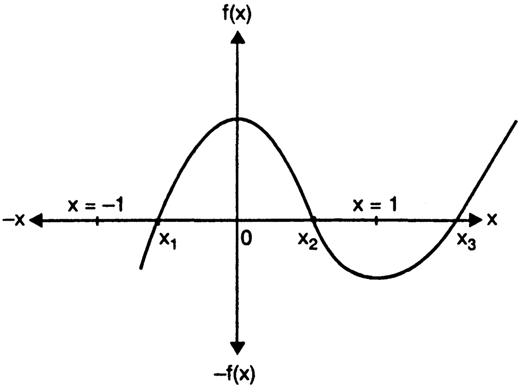

FIG. 8. Plot of $f(x)$ vs x showing location of turning angles of θ for nutational motion of top.

of the top is expressed in Fig. 9, where the motion of the top axis (called the *locus*) is traced on a unit sphere whose center is the fixed point of the top. The curves lie between the latitudes θ_1 and θ_2. On these latitudes, $\dot{\theta} = 0$. The shapes of the curves are determined by the initial conditions, which can be expressed in terms of the value $y = a/b$ (the root of $a - bx$). Figure 9(a) shows the locus curve tangent to the latitudes; Fig. 9(b) shows the curve with loops; and Fig. 9(c) shows the transition curve between Figs. 9(a) and 9(b).

If the initial conditions are such that $\dot{\phi}$ always has the same sign when $\theta_1 \leq \theta \leq \theta_2$, then there are no loops, and Fig. 9(a) obtains. According to Eq. (75), these initial conditions must satisfy the condition that $y > x_2$. The axis of the top then precesses about the vertical axis since ϕ increases in one direction or the other. If $x_1 < y < x_2$, the direction of the precessional motion will be different on the two latitudes, and the locus curve will have loops as shown in Fig. 9(b). This means that there are two values of $\dot{\phi}$, giving two different directions of precession, for each value of θ in a range $\theta_1 < \theta < \theta_2$. The average rate of precession will not vanish, however, so that there is always a net precession in one direction or the other.

Finally, for the case where a/b coincides with one of the roots of $f(x)$, then $\dot{\theta} = \dot{\phi} = 0$ on the latitude representing that root. This is shown in Fig. 9(c), where cusps touch the latitude $\theta = \theta_2$. One set of initial conditions yielding this case is $\theta = \theta_0$, $\dot{\theta} = \dot{\phi} = 0$ at $t = 0$. The initial angle θ_0 is the root that corresponds to the upper latitude in Fig. 9(c). Note that the initial value of E' is $Mgl \cos\theta_0$, and that the terms involving $\dot{\theta}$ and $\dot{\phi}$ are always positive. As $\dot{\theta}$ and $\dot{\phi}$ begin to differ from their initially zero values, the potential energy must decrease in order to conserve energy. This means that θ must increase from its minimum value at $\theta_0 = \theta_2$. When the top is released in this manner, it continues to precess, and it falls until $\theta = \theta_1$, the other latitude. Then it rises until $\theta = \theta_2$, and the process repeats, as shown in Fig. 9(c).

2.5.1 Stability of a Sleeping Top A top is said to be "sleeping" if its spin axis coincides with its axis of symmetry and is in the vertical direction. Let $s = \dot{\psi}$. The variable s is called the spin. If s is large enough, then the spin axis becomes vertical for a small disturbance (the top will perform small oscillations about the vertical). If s is smaller than a critical value, then the oscillations

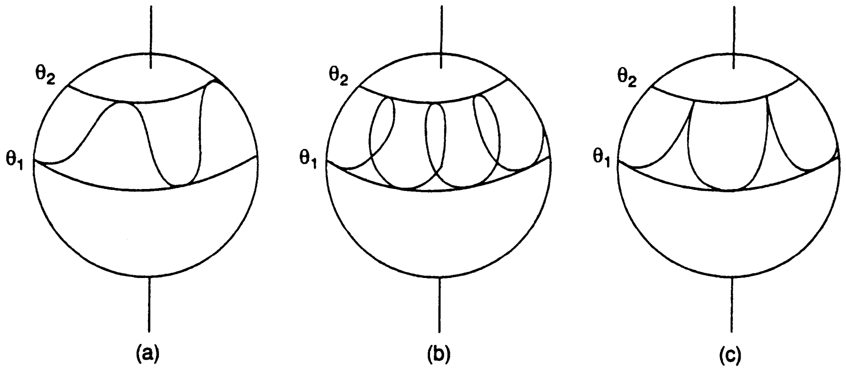

FIG. 9. Possible nutational modes of top, shown by locus curves traced by axis of top on a unit sphere: (a) tangent to latitudes, (b) loops, (c) transition.

about the spin axis will increase so that the top will wobble and eventually fall down. What is this critical value of s? The answer is obtained by study of the cubic $f(x)$. For a sleeping top, we have $\theta = \dot{\theta} = 0$. The cubic becomes

$$f(x) = (1 - x)^2\left[\frac{2Mgl}{I_1}(1 + x) - \left(\frac{I_3}{I_1}\right)^2 s^2\right], \tag{80}$$

since $a = b$ and $\alpha = \beta$. Here $x = 1$ is a double root of $f(x)$. The other root,

$$x = I_3^2 s^2/2MglI_1 - 1, \tag{81}$$

is either greater than unity (x_3) or less than unity (x_1). The case of a root greater than unity is physically unattainable, since x is the cosine of an angle. In that case, then, the only possible motion is that which makes $x = 1$, which corresponds to $\theta = 0$, or the axis vertical. The condition for this is that s be large enough to make the root greater than 1. The critical value is the solution of

$$4Mgl/I_1 = (I_3/I_1)^2 s^2, \tag{82}$$

or

$$|s| = (4I_1Mgl)^{1/2}/I_3. \tag{83}$$

The condition for stability of a sleeping top is that $|s|$ equal or exceed this value. If the condition is not fulfilled, the top if disturbed from the vertical position will fall to the inclination of θ_1 where $\cos\theta_1 = x_1$.

3. VARIATIONAL PRINCIPLES AND THE CALCULUS OF VARIATIONS

Many important results in physics are obtained from what are known as variational principles. Such principles state, in general, that the actual course of physical events is such as to render some quantity an extremum—either a minimum or a maximum. Perhaps the best known of these is Fermat's principle, which applies to optics. But there are equally important principles in mechanics. To appreciate them, we must first review some results in the calculus of variations.

We consider an elementary situation in the differential calculus where we have a differentiable function $I = I(y_i(x), y_i'(x))$, where the prime denotes derivative with respect to x, that we want to *extremize*, that is, to find the values of the y_i that yield an extremum—either a maximum or a minimum—of the function. In the calculus of variations, this concept of extremizing a function is generalized to that of extremizing a *functional*, which is roughly defined as a "function of a function." Specifically, a functional is a definite integral whose integrand is one of a class of functions each of which produces a particular value of the integral. The basic problem in the calculus of variations is to find the function that extremizes the integral (for fixed limits of integration, for simplicity). In mechanics, as well as in optics, the extremum is usually a minimum rather than a maximum. It turns out that each of the admissible functions defines a path joining the end points of the integral. The techniques of the calculus of variations allow us to find that path that minimizes the integral. We set

$$I = \int_{x_1}^{x_2} f(y_i(x), y_i'(x))dx. \tag{84}$$

We now perform variations δy_i and $\delta y_i'$, subject to the conditions that all the variations vanish at the end points of the integral. The condition that I be an extremum is that the resulting $\delta I = 0$. This means

$$\delta I = 0 = \int_{x_1}^{x_2} \sum_i \left(\frac{\partial I}{\partial y_i}\delta y_i + \frac{\partial I}{\partial y_i'}\delta y_i'\right)dx. \tag{85}$$

We present a geometric or physical interpretation of the variations of a function and functional, which relates to the principle of virtual work (Sec. 1.7). Figure 10 shows a minimizing curve $y = y(x)$ in the (x,y) plane, passing through the fixed end points A and B. In mechanics, y represents the position of a particle at a time equal to x. The minimizing curve is the trajectory of the particle, or the time history of its motion. Let $y = \bar{y}(x)$ be a neighboring curve passing through the same end points. δy is the change in y from the minimizing curve *for a fixed* x. [Note that this infinitesimal change in y is different from $dy(x)$, which represents an infinitesimal change in y *along* the minimizing curve.] Let $\delta I(y)$ be the corresponding variation in the

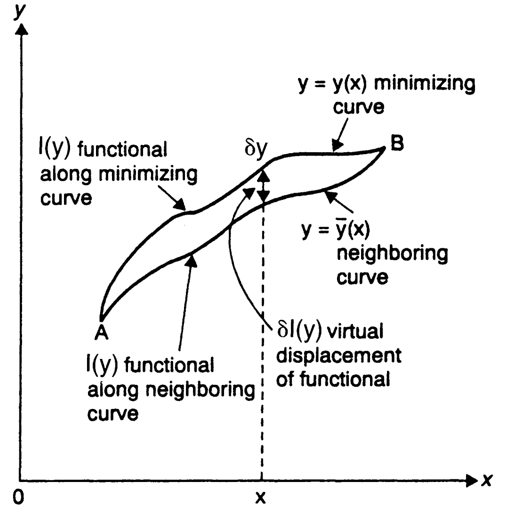
y
y = y(x) minimizing curve
B
I(y) functional along minimizing curve
δy
y = ȳ(x) neighboring curve
A
δI(y) virtual displacement of functional
I(y) functional along neighboring curve
0
x
x

FIG. 10. Minimizing and neighboring curves in (x,y) plane showing virtual displacement of coordinates.

functional $I(y)$. Clearly, δI is the variation in the functional from the minimizing to the neighboring curve for a fixed time.

Given this interpretation, it is easy to show that

$$\delta y_i' = \frac{d}{dx}\delta y_i. \tag{86}$$

We use this in the last term of the integrand in Eq. (85) and integrate by parts:

$$\begin{aligned}\int_{x_1}^{x_2}\sum_i \frac{\partial I}{\partial y_i'}\delta y_i' dx &= \int_{x_1}^{x_2}\sum_i \frac{\partial I}{\partial y_i'}\frac{d\delta y_i}{dx}dx \\ &= \sum_i \frac{\partial I}{\partial y_i'}\delta y_i\Bigg|_{x_1}^{x_2} \\ &\quad - \int_{x_1}^{x_2}\sum_i \frac{d}{dx}\frac{\partial I}{\partial y_i'}\delta y_i dx. \end{aligned}\tag{87}$$

Since the variations were specified to vanish at the end points, the first sum in this result vanishes, and Eq. (85) becomes

$$\delta I = \int_{x_1}^{x_2}\sum_i\left(\frac{\partial I}{\partial y_i} - \frac{d}{dx}\frac{\partial I}{\partial y_i'}\right)\delta y_i dx. \tag{88}$$

But since the variations δy_i are arbitrary (except for vanishing at the end points), the only way the integral can be guaranteed to vanish is for

$$\frac{\partial I}{\partial y_i} - \frac{d}{dx}\frac{\partial I}{\partial y_i'} = 0. \tag{89}$$

This set of equations, known as *Euler's equations*, is our key result. They constitute a necessary condition for the existence of an extremum of I.

Note that we say *necessary* condition. This means that if an extremum exists, then Euler's equations must be satisfied. But that does not guarantee an extremum. We give an example that illustrates Fermat's principle but shows that an extremum does not exist. This principle, enunciated by the French mathematician Pierre Fermat (1601–1655), deals with light rays in a medium of variable refractive index so that the paths of the rays are not necessarily straight lines. The statement of the principle is this: Of all the possible paths taken by a light ray in going from A to B (fixed points), the actual path taken by the ray is the one that takes the least time. Consider Fig. 11, which shows an optically transparent (x,y) plane embedded in a medium of constant index of refraction. Let A be a fixed point on the x axis such that the length of the line segment OA is L. Suppose a light beam starts at point O, travels along the y

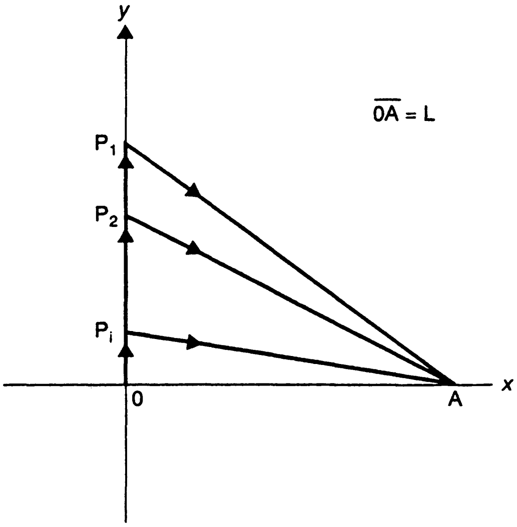

FIG. 11. Example of Fermat's principle showing the nonexistence of an extremum in a two-dimensional optically transparent medium of constant refractive index.

axis a variable distance to the point P_1, then goes along the straight line from P_1 to A. Each trajectory is a curve given by $y = y_i(x)$ for $i = 1, 2, \ldots$. Every admissible path is composed of the broken line segments $y_i = OP_i + P_iA$. According to Fermat's principle, the time taken to traverse each path is a minimum for that path. Therefore, the functional $I(y_i)$ represents the time taken to traverse each broken line. Clearly, the length of each path is greater than L by the amount OP_i. In the limit, $P_i \to P_\infty = 0$. The minimum distance is the line segment $OA = L$. But OA is *not an admissible path*, since the light beam must start at O *in the direction of the y axis* and then travel on a line of sight to A. As P_i tends to O, the length of each admissible path gets closer to L. However, *there is no minimal path*, since we obtained an infinite number of admissible paths as $P_i \to O$. We have thus demonstrated that an extremum does not exist for at least one particular functional.

This brief treatment of the calculus of variations will enhance our understanding of Hamilton's variational principle, which supplies the foundation for Hamilton's equations of motion and an alternative (and generalizable) basis for Lagrange's equations of motion, the two cornerstones of classical mechanics. Again, we emphasize that the key equation is Euler's equation. For a system of N particles in one dimension, there are N Euler equations, one for each particle.

3.1 Lagrange's Equations from Hamilton's Variational Principle

Both Lagrange's and Hamilton's formulations of mechanics involve energy considerations. In these formulations, the mechanical properties of our system are completely described by its kinetic energy T and its potential energy V. In general, we assume that T is a quadratic function of the generalized velocities, possibly also dependent on the generalized coordinates, with the following form:

$$T = \sum_{i,j} P_{ij}(q_1,q_2,\ldots,q_n,t)\dot{q}_i\dot{q}_j, \qquad (90)$$

where the P_{ij} are prescribed functions of the set $\mathbf{q}$. Equation (90) allows T to depend arbitrarily on $\mathbf{q}$. For the time being, we consider only conservative systems so that $T + V = E$, where the total energy E is constant.

The solution of Lagrange's equations of motion yields the generalized coordinates and velocities as functions of time, which completely describes the trajectory of each particle of the system in phase space. On the other hand, as we will see, the solution of Hamilton's equations of motion yields the generalized coordinates *and momenta* as functions of time, which describes the system. However, there is a conceptual difference between the two phase spaces. Lagrange's equations are second-order differential equations, thus requiring two sets of initial conditions (generalized coordinates and velocities) to describe the trajectory of the system. On the other hand, Hamilton's equations consist of pairs of first-order differential equations, each pair requiring a knowledge of initial generalized coordinates and momenta. This is reflected in the meaning of the trajectories in the two spaces.

Hamilton applied the variational principle to the Lagrangian $L(\mathbf{q}(t),\dot{\mathbf{q}}(t),t)$. His principle can be stated as follows: Between any two instants of time, t_A and t_B, the actual motion of a dynamical system of particles, defined by their trajectories $q = q(t)$ in configuration space, proceeds in such a way as to make the functional

$$I(\mathbf{q},\dot{\mathbf{q}}) = \int_{t_A}^{t_B} L(\mathbf{q}(t),\dot{\mathbf{q}}(t),t)dt \qquad (91)$$

a minimum.

We relate Eq. (91) to the functional defined by Eq. (84) and the corresponding Euler equations given by Eq. (89) as follows: the term x corresponds to t, y corresponds to q, dy/dx corresponds to dq/dt, and F corresponds to L. Therefore, L must satisfy Euler's equation and we obtain

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} = 0, \quad i = 1, 2, \ldots, n. \qquad (92)$$

As seen by reference to Sec. 1.8, these are Lagrange's equations of motion. Because of the way they have been derived here, they are sometimes called the Euler-Lagrange equations.

As an example of the use of Lagrange's equations, we consider a body of mass m in a central force field. The potential is $V = -k/r$, where k is proportional to the gravitational constant. The kinetic energy is $T = \frac{1}{2}m\dot{r}^2$. This gives

$$L = \left(\frac{m}{2}\right)\dot{r}^2 + \frac{k}{r}, \quad \frac{\partial L}{\partial \dot{r}} = m\dot{r} = p, \quad \frac{\partial L}{\partial r} = -\frac{k}{r^2}. \tag{93}$$

Using Lagrange's equation (88), we get

$$m\ddot{r} = -k/r^2 = f, \tag{94}$$

where f is the inverse-square force.

As another example, we take a system of N uncoupled harmonic oscillators in three-space. The ith particle has mass m_i, and all the particles have the spring constant k. Since the particles are noninteracting, there are no constraints, so that we may use Cartesian coordinates for the generalized coordinates. Therefore

$$q = (x_1, y_1, z_1, \ldots, x_N, y_N, z_N),$$
$$\dot{q} = (u_1, v_1, w_1, \ldots, u_N, v_N, w_N),$$

where

$$\dot{x}_i = u_i, \quad \dot{y}_i = v_i, \quad \dot{z}_i = w_i, \quad i = 1, 2, \ldots, N.$$

The Lagrangian becomes

$$L = \tfrac{1}{2}\sum_{i=1}^{i=N} m_i[u_i^2 + v_i^2 + w_i^2] - \tfrac{1}{2}k\sum_{i=1}^{i=N}[x_i^2 + y_i^2 + z_i^2]. \tag{95}$$

Differentiating L with respect to the velocity components yields the momentum components for the ith particle:

$$\frac{\partial L}{\partial u_i} = m_i u_i = (p_i)_x, \quad \frac{\partial L}{\partial v_i} = m_i v_i = (p_i)_y,$$
$$\frac{\partial L}{\partial w_i} = m_i w_i = (p_i)_z. \tag{96}$$

The spring force is given by

$$\frac{\partial L}{\partial x_i} = -kx_i, \quad \frac{\partial L}{\partial y_i} = -ky_i, \quad \frac{\partial L}{\partial z_i} = -kz_i, \tag{97}$$

which are the equations of motion for our system of uncoupled harmonic oscillators.

This has been an integral approach to Lagrange's equations, in the sense that it uses energy considerations in the variational technique to compare admissible paths with the actual path in configuration space. This is to be contrasted with the differential approach used in deriving Lagrange's equations from D'Alembert's principle, where we consider the instantaneous state of the system and superimpose small virtual displacements from this state.

3.1.1 Dealing with Constraints As we noted in Sec. 1.6, the presence of constraints means that the virtual displacements are not independent. In that case, the derivation of Lagrange's equations from Hamilton's principle fails, since that derivation depended on the independence of the virtual displacements (variations of the generalized coordinates). Lagrange found a way to deal with this problem. Note that each of the constraint conditions given by Eq. (28) equates a certain sum to zero. We can therefore add or subtract any multiple of each equation from any other quantity without effect. Specifically, we multiply each of the equations (28) by a corresponding constant λ_i, giving

$$\lambda_i \sum_j a_{ij}\delta q_j = 0, \quad i = 1, 2, \ldots, k. \tag{98}$$

We subtract the sum of these terms from the integrand in Eq. (88):

$$\delta I = 0 = \int_{x_1}^{x_2} \sum_i \left(\frac{\partial I}{\partial y_i} - \frac{d}{dx}\frac{\partial I}{\partial y_i'} - \sum_{j=1}^{k} \lambda_j q_{ji}\right) \times \delta q_i dx. \tag{99}$$

Of course, this does not affect the fact that the δq_i's are not independent. Only $n - k$ of them may be chosen independently. However, the values of the λ_i's have not yet been determined. We may therefore define them by setting the integrand in Eq. (99) equal to zero for the last k δq_i's. That is,

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} - \sum_{j=1}^{k} \lambda_j a_{ji} = 0,$$
$$i = n - k + 1, n - k + 2, \ldots, n. \tag{100}$$

These are the equations of motion for those q_i's that are related by the constraints. They form a set of k linear equations for the λ_j's. Having determined them from these equations, we can now set the integrand of Eq. (99) equal to zero for the remaining $n - k$ independent δq_i's. Combining these results yields the complete set of n equations for a nonholonomic system,

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_j} - \frac{\partial L}{\partial q_j} = \sum_{i=1}^{k} \lambda_i a_{ij}, \quad j = 1, 2, \ldots, n. \tag{101}$$

The n equations (101) combined with the k equations (100) yield a complete solution in that they allow us to calculate the n q_i's and the k λ_i's.

The right-hand side of Eq. (101) is the force applied to the jth particle by the constraints and is called a *generalized force*. If we removed the k constraints on the jth particle and applied external forces equal to the constraining forces, the system would not change.

As an easy example, we take the case of a hoop rolling without slipping down an inclined ramp. Figure 1 shows a profile of the hoop, whose mass is m and whose radius is r, rolling in the (x,y) plane down the ramp OB inclined at an angle α to the horizontal. The hoop starts at the top of the ramp tangent to the incline at B, at height h, and rolls down the incline. At some time t, the hoop is tangent to the incline at D, having rotated about its horizontal axis through the angle θ. Its center has translated a distance s parallel to the incline, having dropped through a vertical distance $(h - s)\sin\alpha$. Let the generalized coordinates be $q_1 = s$, $q_2 = \theta$. There is one constraint equation; it is given by Eq. (23), which we repeat:

$$ds - r\,d\theta = 0. \tag{102}$$

This is a specific case of the constraints given in differential form [Eq. (27)] with $i = 1$, $j = 1, 2$, $a_{11} = 1$, and $a_{12} = r$. The Lagrangian is

$$L = T - V = \tfrac{1}{2}m\dot{s}^2 + \tfrac{1}{2}r^2\dot{\theta}^2 - mg(h - s)\sin\alpha. \tag{103}$$

The first term represents the kinetic energy of the pure translational motion of the center of mass; the second term represents the kinetic energy of the pure rotational motion about the horizontal axis through the center of mass. Lagrange's equations for the two degrees of freedom are

$$\begin{aligned} &m\ddot{s} - mg\sin\alpha = \lambda, \\ &mr^2\ddot{\theta} - r\lambda. \end{aligned} \tag{104}$$

The constraint equation is

$$r\dot{\theta} = \dot{s}. \tag{105}$$

Differentiating with respect to t gives

$$r\ddot{\theta} = \ddot{s}. \tag{106}$$

From these equations we obtain

$$\begin{aligned} \lambda &= -\tfrac{1}{2}mg\sin\alpha, \\ \ddot{\theta} &= -\frac{1}{2r}g\sin\alpha. \end{aligned} \tag{107}$$

The equation for λ is the friction force of constraint. It is directed up the inclined plane and is $\frac{1}{2}$ the projection of the weight of the hoop normal to the incline.

3.2 Nonconservative Systems

We now consider a system of N particles with n degrees of freedom subject to nonconservative forces such as those due to friction and viscosity. These forces dissipate heat, so that the law of conservation of energy is not valid. This means that $\mathbf{F}$ is not derivable from a potential V that depends only on $\mathbf{q}$. Then $\mathbf{Q}$ cannot be obtained from Eq. (46). But all is not lost! Viscous forces depend on the velocity as well as position. As examples, the slow motion of a rough sphere in a viscous fluid is proportional to its velocity; the motion of an airplane or projectile in air depends on the square of its velocity (because of air drag). Therefore, we replace V by a velocity-dependent potential U that depends on the generalized velocities as well as coordinates, so that $U = U(\dot{q},q)$. The potential U is sometimes called the *generalized potential*. The Lagrangian is now given by

$$L = T - U. \tag{108}$$

We now set

$$T(\dot{q},q) = L(\dot{q},q) + U(\dot{q},q), \tag{109}$$

and insert this expression in Eq. (44) to obtain

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = Q_i + \frac{\partial U}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial U}{\partial \dot{q}_i}\right), \quad i = 1, 2, \ldots, n. \tag{110}$$

If we set the right-hand side of this equation equal to zero, we obtain Lagrange's equations. Then the generalized force for the ith particle becomes

$$Q_i = -\frac{\partial U}{\partial q_i} + \frac{d}{dt}\left(\frac{\partial U}{\partial \dot{q}_i}\right). \tag{111}$$

For the special case of a conservative force system, this reduces to Eq. (46).

3.2.1 Rayleigh's Dissipation Function For our system of particles in a nonconservative force field, we restrict ourselves to frictional forces obeying the simple law

$$\mathbf{f}_i = -\mathbf{k}\cdot\mathbf{v}_i, \tag{112}$$

where $\mathbf{f}_i$ is the friction force and $\mathbf{v}_i$ is the velocity of the ith particle. The friction coefficient $\mathbf{k}$ is a tensor, which we take to be diagonal with nonzero components k_x, k_y, k_z, and to be the same for each particle. Lord Rayleigh (1878) introduced a scalar quadratic function of the velocity called the *Rayleigh dissipation function* $\mathcal{F}$, which is defined by

$$\mathcal{F} = \tfrac{1}{2}\sum_{i=1}^{n} (k_x u_i^2 + k_y v_i^2 + k_z w_i^2). \tag{113}$$

From this definition we obtain

$$\mathbf{f} = -\nabla_{\mathrm{v}}\,\mathcal{F}, \quad \text{where } \nabla_{\mathrm{v}} = \mathbf{i}\frac{\partial}{\partial u} + \mathbf{j}\frac{\partial}{\partial v} + \mathbf{k}\frac{\partial}{\partial w}, \tag{114}$$

where $(\mathbf{i},\mathbf{j},\mathbf{k})$ are the unit vectors in the (x,y,z) directions. Note that the gradient operator is with respect to the velocity. We give a physical interpretation of $\mathcal{F}$ by considering a differential amount of work dW_{f} done by the system against the friction force $\mathbf{f}$. This is given by

$$\begin{aligned} dW_{\mathrm{f}} &= -\mathbf{f}\cdot d\mathbf{r} = -\mathbf{f}\cdot\mathbf{v}dt \\ &= (k_x u^2 + k_y v^2 + k_z w^2)dt. \end{aligned} \tag{115}$$

This means that the Rayleigh dissipation function is *half* the rate of energy dissipated by friction. The jth component of the generalized force due to friction is

$$Q_j = \sum_i \mathbf{f}_i\cdot\frac{\partial \mathbf{r}}{\partial q_j} = -\sum_i \nabla_{\mathrm{v}}\,\mathcal{F}\cdot\frac{\partial \dot{\mathbf{r}}}{\partial \dot{q}_j} = -\frac{\partial \mathcal{F}}{\partial \dot{q}}. \tag{116}$$

Lagrange's equations now become

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_j} - \frac{\partial L}{\partial q_j} = -\frac{\partial \mathcal{F}}{\partial \dot{q}_j}, \quad j = 1, 2, \ldots, n. \tag{117}$$

We see from these equations that the two scalars L and $\mathcal{F}$ must be prescribed in order to solve for the motion of a system subject to a viscous force.

4. HAMILTON'S EQUATIONS OF MOTION

Recall that Lagrange's formulation of the equations of motion is a description of a dynamical situation in terms of trajectories defined by the generalized coordinates and velocities as functions of time. Hamilton's alternative formulation of the equations of motion is based on introducing a generalized p_i called the *canonical* or *conjugate momentum* (each p_i is conjugate to the corresponding q_i). Recall that the term $\partial L/\partial \dot{q}_i$ is the momentum of the ith particle. It is only natural that Hamilton would define the conjugate momentum as

$$p_i = \frac{\partial L}{\partial \dot{q}_i}. \tag{118}$$

Again, considering a conservative system, Lagrange's equation for the ith particle is transformed into

$$\frac{d}{dt}p_i \equiv \dot{p}_i = -\frac{\partial V}{\partial q_i}, \tag{119}$$

which is a first-order system of equations for p_i and V. But this form of Lagrange's equations was not used by Hamilton, since he wanted a system of equations of motion where the independent variables are $(\mathbf{p},\mathbf{q})$. To this end, he defined a function, aptly called the Hamiltonian, as follows:

$$H = H(\mathbf{p},\mathbf{q}) = T + V, \tag{120}$$

where H is the total energy of a conservative system and is a function of $(\mathbf{p},\mathbf{q})$. Expanding the time derivative of L gives

$$\begin{aligned}\frac{dL}{dt} &= \sum_i \frac{\partial L}{\partial \dot{q}_i}\frac{d\dot{q}_i}{dt} + \sum_i \left(\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i}\right)\dot{q}_i \\ &= \sum_i \frac{d}{dt}\left(\dot{q}_i \frac{\partial L}{\partial \dot{q}_i}\right).\end{aligned} \tag{121}$$

Integrating this equation gives

$$\sum_i \dot{q}_i \frac{\partial L}{\partial \dot{q}_i} - L = \text{const.} = H, \tag{122}$$

where we provisionally set the constant of integration equal to the Hamiltonian. From this we obtain

$$H = \sum_i \dot{q}_i p_i - L, \tag{123}$$

which is an important equation, provided we can show that this constant of integration is indeed the total energy of a conservative system. We show this by first using the definition of L and recognizing that V depends only on $\mathbf{q}$. This gives

$$p_i = \frac{\partial T}{\partial \dot{q}_i}. \tag{124}$$

We also have

$$\sum_i \dot{q}_i p_i = \sum_i \dot{q}_i \frac{\partial T}{\partial \dot{q}_i} = 2T, \tag{125}$$

since T is a homogeneous quadratic function of $\dot{q}_i$. From these expressions, we obtain

$$H = 2T - L = 2T - (T - V) = T + V = E. \tag{126}$$

We now derive Hamilton's equations, which consist of a pair of first-order partial differential equations in $(\mathbf{p},\mathbf{q},t)$. To this end, we expand dH and use the above relations to obtain

$$dH = \sum_i \frac{\partial H}{\partial p_i} dp_i + \sum_i \frac{\partial H}{\partial q_i} dq_i + \frac{\partial H}{\partial t} dt. \tag{127}$$

From Eq. (123) we obtain another expansion for dH:

$$\begin{aligned}dH = &\sum_i \dot{q}_i dp_i + \sum_i p_i d\dot{q}_i - \sum_i \frac{\partial L}{\partial \dot{q}_i} d\dot{q}_i \\ &- \sum_i \frac{\partial L}{\partial q_i} dq_i - \frac{\partial L}{\partial t} dt.\end{aligned} \tag{128}$$

The coefficients of the $d\dot{q}_i$'s are zero because of the definition of the p_i's. From Lagrange's equations we obtain

$$\frac{\partial L}{\partial q_i} = \dot{p}_i. \tag{129}$$

Equation (128) then reduces to the simple form

$$dH = \sum_i \dot{q}_i dp_i - \sum_i \dot{p}_i dq_i - \frac{\partial L}{\partial t} dt. \tag{130}$$

Equating coefficients of like terms in Eqs. (127) and (130) yields

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q_i}, \quad i = 1, 2, \ldots, n; \tag{131}$$

$$-\frac{\partial L}{\partial t} = \frac{\partial H}{\partial t}. \tag{132}$$

Equations (131) are *Hamilton's canonical equations of motion*. They constitute a system of $2n$ first-order partial differential equations for $H(\mathbf{p},\mathbf{q},t)$ that replaces the Lagrange system of n second-order equations for $L(\mathbf{q},\dot{\mathbf{q}},t)$. Equation (132) relates L to H for the time-varying case.

As an example, we again take the system of n uncoupled simple harmonic oscillators, so that $\mathbf{q}$ is interpreted as a set of normal coordinates. Let the spring constant k be the same for each particle. The Hamiltonian becomes

$$\begin{aligned}H = T + V &= \frac{1}{2m}\sum_i p_i^2 + \frac{k}{2}\sum_i q_i^2 \\ &= \tfrac{1}{2}\sum \mathbf{p}\cdot\frac{\mathbf{p}}{m_i} + \tfrac{1}{2}k\sum \mathbf{q}\cdot\mathbf{q} = H(\mathbf{p},\mathbf{q},t).\end{aligned} \tag{133}$$

From this we obtain

$$\frac{\partial H}{\partial q_i} = kq_i, \quad \frac{\partial H}{\partial p_i} = \frac{p_i}{m_i}. \tag{134}$$

Applying Hamilton's equations, we get

$$\dot{p}_i = -kq_i, \quad \dot{q}_i = p_i/m_i. \tag{135}$$

The second equation is merely a rearrangement of the definition of the momentum of the *i*th particle.

4.1 Cyclic Coordinates

If the Lagrangian of a system does not contain a particular generalized coordinate q_i (although it may contain the corresponding generalized velocity), then that q_i is said to be a *cyclic coordinate*. Clearly, we have

$$\frac{\partial L}{\partial q_i} = 0, \tag{136}$$

so that the Lagrange equation of motion for the *i*th particle becomes

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i} = \dot{p}_i = 0. \tag{137}$$

This expresses the important result

$$p_i = \text{const.} \quad \text{for } q_i \text{ cyclic,} \tag{138}$$

which can be put in the form of the following conservation theorem: *The generalized momentum conjugate to a cyclic coordinate is conserved.*

Since the momentum conjugate to a cyclic coordinate is constant, it follows from one of Hamilton's equations that H does not contain the coordinate. This is expressed in the following principle: *A cyclic coordinate will not appear in the Hamiltonian. Conversely, if a generalized coordinate does not occur in the Hamiltonian, then the conjugate momentum is conserved.*

5. CANONICAL EQUATIONS OF TRANSFORMATION

The *canonical transformations* that we shall discuss are those transformations of coordinates and momenta that conserve Hamilton's canonical equations of motion (preserve their form). Recall that $(\mathbf{q},\mathbf{p})$ are the generalized position and momentum vectors in phase space (which was introduced in Sec. 1.5). Let $(\mathbf{Q},\mathbf{P})$ be the corresponding transformed vectors. The transformation equations from the old to the new vectors have the following form:

$$\begin{aligned} &\mathbf{Q} = \mathbf{Q}(\mathbf{q},\mathbf{p}), \\ &\quad \text{or } Q_i = Q_i(q_1,q_2,\ldots,q_n,p_1,p_2,\ldots,p_n), \\ &\mathbf{P} = \mathbf{P}(\mathbf{q},\mathbf{p}), \\ &\quad \text{or } P_i = P_i(q_1,q_2,\ldots,q_n,p_1,p_2,\ldots,p_n). \end{aligned} \tag{139}$$

This is the general form of the transformation equations from original to transformed phase space. We seek the properties of such a transformation that makes Hamilton's equations invariant.

Let $H = H(\mathbf{p},\mathbf{q},t)$ be a general Hamiltonian. Let the transformed Hamiltonian be designated as K. Then $K = K(\mathbf{P},\mathbf{Q},t)$. ($t$ transforms into itself.) Hamilton's canonical equations of motion in the transformed space are assumed to be given by

$$\dot{Q}_i = \frac{\partial \mathbf{K}}{\partial P_i}, \quad \dot{P}_i = -\frac{\partial K}{\partial Q_i}. \tag{140}$$

Those transformation equations (139) that satisfy Eqs. (140) are said to be *canonical* transformations. Some authors call these *contact* transformations, because neighboring curves are transformed into neighboring curves.

The key to finding the appropriate set of canonical transformations is to start with Hamilton's variational principle, where the integrand in the functional $I = \int_{t_1}^{t_2} L\,dt$ is given explicitly in terms of the Hamiltonian. This is easily done since

$$L = \sum_i p_i\dot{q}_i - H. \tag{141}$$

Therefore, Hamilton's principle may be put in the form

$$\delta \int_{t_1}^{t_2} \left[\sum_i p_i\dot{q}_i - H(\mathbf{q},\mathbf{p},t)\right] dt = 0. \tag{142}$$

Under a canonical transformation, Hamilton's principle becomes

$$\delta \int_{t_1}^{t_2} \left[\sum_i P_i\dot{Q}_i - K(\mathbf{Q},\mathbf{P},t) \right] dt = 0. \tag{143}$$

If we set the difference between the two integrands equal to the derivative of an arbitrary function of time $F(t)$, we get

$$\delta \int_{t_1}^{t_2} \left\{ \left[\sum p_i\dot{q}_i - H \right] - \left[\sum P_i\dot{Q}_i - K \right] - \frac{dF(t)}{dt} \right\} dt = 0. \tag{144}$$

(Obviously, dF is an exact differential.) From this equation, it is clear that $F(t)$ is an extremum with respect to the fixed end points (t_1,t_2), so that

$$\delta \int_{t_1}^{t_2} \frac{dF(t)}{dt}\, dt = \delta F(t_2) - \delta F(t_1) = 0. \tag{145}$$

Knowing F as a function of one of the following sets of coordinates—$(\mathbf{q},\mathbf{Q})$, $(\mathbf{q},\mathbf{P})$, $(\mathbf{p},\mathbf{Q})$, or $(\mathbf{p},\mathbf{P})$—allows us to calculate the transformation equations given by Eq. (139). For this reason, F is called a *generating function*, because it "generates" the transformation equations. In addition to time, for a system with n degrees of freedom there are a total of $2n$ generalized coordinates and $2n$ generalized momenta (old and new together). The transformation equations (139) consist of $2n$ equations, which yield $2n$ independent relationships between these variables. Here F is usually taken to have one of the following forms: $F_1(\mathbf{q},\mathbf{Q},t)$, $F_2(\mathbf{q},\mathbf{P},t)$, $F_3(\mathbf{p},\mathbf{Q},t)$, or $F_4(\mathbf{p},\mathbf{P},t)$; but many other forms are possible.

Suppose $F = F_1(\mathbf{q},\mathbf{Q},t)$. Recall that the difference of the integrands in Eqs. (142) and (143) must be a perfect differential. Therefore, we set

$$\sum_i p_i dq_i - H dt - \sum_i P_i dQ_i + K dt = dF_1(q,Q,t) = \sum \frac{\partial F_1}{\partial q_i} dq_i + \sum \frac{\partial F_1}{\partial Q_i} dQ_i + \frac{\partial F_1}{\partial t} dt. \tag{146}$$

Since $(\mathbf{q},\mathbf{Q},t)$ are independent variables, the coefficients of each dq_i, each dQ_i, and dt must vanish. This yields

$$p_i = \frac{\partial F_1}{\partial q_i}, \quad P_i = \frac{\partial F_1}{\partial Q_i}, \quad K - H = \frac{\partial F_1}{\partial t}. \tag{147}$$

The first two sets of these equations can be solved for $\mathbf{Q} = \mathbf{Q}(\mathbf{q},\mathbf{p})$ and $\mathbf{P} = \mathbf{P}(\mathbf{q},\mathbf{p})$. The third equation allows us to calculate the new Hamiltonian $K = K(\mathbf{Q},\mathbf{P},t)$ in terms of the old $H = H(\mathbf{q},\mathbf{p},t)$ and the transformation equations.

5.1 Hamilton-Jacobi Theory

There is a general procedure that allows us to calculate the appropriate generating function without using this *ad hoc* method. It involves manipulating the Hamiltonian into the form of a partial differential equation and thereby deriving the generating function (specifically, its spatial derivatives) as a solution. This technique is called the Hamilton-Jacobi theory.

5.1.1 Hamilton-Jacobi Equations The first point to make in investigating this theory is that we need a transformation from $(\mathbf{q},\mathbf{p})$ space to $(\mathbf{Q},\mathbf{P})$ space that will simplify Hamilton's equations in the transformed coordinates. The key is the use of cyclic coordinates. We saw (Sec. 4.1) that if the Hamiltonian does not contain a particular q_i, then the conjugate momentum is conserved. In particular, if the transformed Hamiltonian is independent of $\mathbf{Q}$, then all the transformed momenta are conserved, and we can set the set of momenta $\mathbf{P} = \boldsymbol{\beta} = \mathbf{P}(0)$, the initial value of the set. We can go even further by making the set $\mathbf{Q} = \boldsymbol{\alpha}$ where $\boldsymbol{\alpha}$ may be taken as the initial value also. Then we consider the transformation

$$\mathbf{q} = \mathbf{q}(\boldsymbol{\alpha},\boldsymbol{\beta},t), \quad \mathbf{p} = \mathbf{p}(\boldsymbol{\alpha},\boldsymbol{\beta},t), \tag{148}$$

where the original system is transformed into the system of initial coordinates and momenta (the initial conditions). The transformed Hamiltonian K is constant.

To derive the Hamilton-Jacobi equations, we refer to Eq. (144). Let $F = F(\mathbf{q},\mathbf{P},t)$. We again consider the integrand for a system of n degrees of freedom and add the exact differential $\Sigma P_i dQ_i + \Sigma Q_i dP_i$ to the differential dF, which is expanded in terms of $d\mathbf{q}$, $d\mathbf{P}$, $d\mathbf{Q}$, and dt. (At this stage, we consider **P** and **Q** as variable sets; later, we shall set them equal to the initial position and momentum sets.) We now use the condition that K = const., and we get

$$\begin{aligned}&\sum p_i dq_i - \sum P_i dQ_i - Hdt\\ &\quad - \sum \frac{\partial F}{\partial q_i} dq_i - \sum \frac{\partial F}{\partial P_i} dP_i - \frac{\partial F}{\partial t} dt\\ &\quad + \sum P_i dQ_i + \sum Q_i dP_i = 0. \end{aligned} \tag{149}$$

Equating the coefficients of $d\mathbf{q}$, $d\mathbf{Q}$, $d\mathbf{P}$, and dt to zero yields

$$p_i = \frac{\partial F}{\partial q_i}, \quad Q_i = \frac{\partial F}{\partial P_i}, \tag{150}$$

$$H + \frac{\partial F}{\partial t} = 0. \tag{151}$$

Recall that $H = H(\mathbf{q},\mathbf{p},t)$. Equation (151) then becomes a partial differential equation for F when the expression for the components of $\mathbf{p}$ in the first equation of (150) is substituted into H. We get (in expanded form)

$$H\left(q_1,\ldots,q_n,\frac{\partial F}{\partial q_1},\ldots,\frac{\partial F}{\partial q_n},t\right) + \frac{\partial F}{\partial t} = 0. \tag{152}$$

Equation (152) is called the *Hamilton-Jacobi* (H-J) equation. It is a first-order partial differential equation in the n variables $(q_1,\ldots,q_n,t)$ for the solution F. For a given dynamical problem, it is clear that H is a prescribed function of the arguments shown. It is customary to denote the generating function F by S, which is called *Hamilton's principal function*. Note that S itself does not appear explicitly in the H-J equation, but only its partial derivatives with respect to the components of $\mathbf{q}$. Therefore, S plus an additive constant is also a solution of the H-J equation. There must be $n + 1$ constants of integration. We may select n of those to be the components of the initial momentum set.

The additive constant may then be taken as the total energy E. The complete solution of the H-J equation is given by

$$S = S(q_1,\ldots,q_n,\beta_1,\ldots,\beta_n,t), \quad \beta_i = P_i. \tag{153}$$

Suppose we have a complete solution for a given dynamical problem. We rewrite Eq. (150) as

$$p_i = \frac{\partial S(q_1,\ldots,q_n,\beta_1,\ldots,\beta_n,t)}{\partial q_i}. \tag{154}$$

Note that S is also a known function of the initial momentum set. Therefore,

$$\begin{aligned}&Q_i = \alpha_i = \left.\frac{\partial S(q_1,\ldots,q_n,\beta_1,\ldots,\beta_n,t)}{\partial \beta_i}\right|_{t=0},\\ &\alpha_i = q_i(0), \quad \beta_i = p_i(0).\end{aligned} \tag{155}$$

We can perform an inverse transformation on Eq. (139) and obtain our final results:

$$\begin{aligned}&q_i = q_i(\alpha_1,\ldots,\alpha_n,\beta_1,\ldots,\beta_n,t),\\ &\ i = 1,\ldots,n;\end{aligned} \tag{156}$$

$$p_i = p_i(\alpha_1,\ldots,\alpha_n,\beta_1,\ldots,\beta_n,t). \tag{157}$$

Combining these equations and using t as a parameter, we get the trajectory of the solution in phase space. This is a $2n$-dimensional space where the trajectory starts at (α,β) and unwinds as t increases from $t = 0$. The trajectory thus gives the complete time history of the dynamical system for a given set of initial conditions. Thus, Hamilton's principal function S yields a solution to a dynamical problem. In addition, S is the generating function that generates a canonical transformation to constant coordinates and momenta. Instead of constructing *ad hoc* generating functions, the problem is transformed into that of solving a first-order partial differential equation for $\mathbf{q} = \mathbf{q}(t)$.

5.1.2 Hamilton's Characteristic Function

If H is not an explicit function of t, then the H-J equation is

$$\frac{\partial S}{\partial t} + H\left(\mathbf{q},\frac{\partial S}{\partial \mathbf{q}}\right) = 0 \tag{158}$$

(where $\partial S/\partial \mathbf{q}$ stands for the n spatial derivatives of S). We may separate the t-dependent part of S by setting

$$S(\mathbf{q},t) = W(\mathbf{q},\boldsymbol{\beta}) - Et, \tag{159}$$

where E is the total energy. The t-independent function is called *Hamilton's character-*

istic function. It satisfies the H-J equation

$$H\left(\mathbf{q}, \frac{\partial W}{\partial \mathbf{q}}\right) = E. \tag{160}$$

Here W generates a canonical transformation (independent of t) in which all the transformed coordinates are cyclic (the initial coordinates and momenta).

If W can be put in the form

$$W = \sum_i W_i(q_i,\beta_1,\ldots,\beta_n), \tag{161}$$

then the H-J equation can be split into n ordinary differential equations of the form

$$H_i\left(q_i, \frac{\partial W_i}{\partial q_i}, \beta_1,\ldots,\beta_n\right) = e_i, \quad i = 1, 2, \ldots, n, \tag{162}$$

where the e_i are several constants. (For a discussion of separation of variables, see Landau and Lifshitz, 1976.) Each of these n equations is a first-order ordinary differential equation and can clearly be reduced to quadratures. We need only solve for $\partial W_i/\partial q_i$ and then integrate with respect to q_i.

5.1.3 Keplerian Motion of an Electron about a Nucleus As an example of the use of the H-J equation, we consider the motion of an electron of charge $-e$ and mass m about an atomic nucleus of charge Ze and mass M. Here Z is the atomic number (the number of electrons in the neutral atom). We use the approximation that the nucleus remains stationary since $m \ll M$. In polar coordinates (r,θ), the kinetic energy of the electron is

$$T = \tfrac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2), \tag{163}$$

and the potential energy is

$$V = -e^2Z/r. \tag{164}$$

From Eq. (163) we obtain the components of the generalized momentum in polar coordinates:

$$p_r = m\dot{r}, \quad p_\theta = mr^2\dot{\theta}^2. \tag{165}$$

Therefore, the Hamiltonian is

$$H = \frac{1}{2m}\left(p_r^2 + \frac{1}{r^2}p_\theta^2\right) - \frac{e^2Z}{r}. \tag{166}$$

We then obtain the H-J equation

$$\frac{1}{2m}\left[\left(\frac{\partial S}{\partial r}\right)^2 + \frac{1}{r^2}\left(\frac{\partial S}{\partial \theta}\right)^2\right] - \frac{e^2Z}{r} = E, \tag{167}$$

where E is the total energy. This is a case where the variables can be separated:

$$\frac{\partial S}{\partial \theta} = \left[2mEr^2 + 2me^2Zr - r^2\left(\frac{\partial S}{\partial r}\right)^2\right]^{1/2} = \beta, \tag{168}$$

where β is a constant. From this we obtain the two subsidiary equations

$$\frac{\partial S}{\partial \theta} = \beta \tag{169}$$

and

$$\frac{\partial S}{\partial r} = \left(2mE + 2\frac{me^2Z}{r} - \frac{\beta^2}{r^2}\right)^{1/2}. \tag{170}$$

The first equation tells us that the angular momentum p_θ is constant. Indeed, θ is a cyclic coordinate, since it does not appear explicitly in the H-J equation; therefore the conjugate momentum p_θ is conserved. Since

$$dS = \frac{\partial S}{\partial \theta}d\theta + \frac{\partial S}{\partial r}dr,$$

we can determine S by integrating Eqs. (169) and (170) and adding. This gives

$$S = \beta\theta + \int\left(2mE + 2me^2\frac{Z}{r} - \frac{\beta^2}{r^2}\right)^{1/2} dr. \tag{171}$$

To calculate the orbit, we differentiate Eq. (171) with respect to β and obtain

$$\theta - \beta\int\frac{dr}{r(-\beta^2 + 2me^2Zr + 2mEr^2)^{1/2}} = \gamma. \tag{172}$$

Using any table of integrals, we evaluate the integral and solve for the reciprocal of the radius vector, thus obtaining

$$\frac{1}{r} = \frac{me^2Z}{\beta^2} + \left(\frac{m^2e^4Z^2}{\beta^4} + \frac{2mE}{\beta^2}\right)^{1/2}\cos(\theta - \phi), \tag{173}$$

where ϕ is a constant phase angle. Equation (173) is the equation of a conic section. It is shown in any text on analytic geometry that the eccentricity ϵ is

$$\epsilon = \left(1 + \frac{2E\beta^2}{me^4Z^2}\right)^{1/2}. \tag{174}$$

This shows at once that the orbit is elliptic, parabolic, or hyperbolic according to whether E is negative, zero, or positive. An electron rotating about a nucleus is only stable if E is negative, for only in this case can the orbit be closed. If we set $E = -W$, then the semimajor axis of the ellipse is

$$a = e^2Z/W. \tag{175}$$

This tells us that all elliptic orbits having the same energy have the same major axis irrespective of their eccentricities.

6. PRINCIPLE OF LEAST ACTION

In Hamilton's variational principle (for a conservative system), the neighboring paths of an extremum might not conserve energy, meaning that $T + V =$ const. only along the dynamical path (extremum). There exists another variational principle, the *principle of least action*. Its purpose is to rectify this weakness in Hamilton's principle.

The principle of least action is like Hamilton's principle in that it compares a dynamical path with a neighboring path that has the same fixed end points. Unlike Hamilton's principle, the principle of least action allows the virtual displacements to depend on time. This allows us to impose the constraint that if H is conserved on the actual path, it is also conserved on the varied path. Because of this constraint, corresponding points on the dynamical path and a neighboring path are not chosen as points reached at the same time, so that a varied path is not necessarily traversed in the same time as the dynamic path. We show that the functional that is made stationary and fulfills the criterion of energy conservation on neighboring paths is the *action integral*, which we now describe.

The action integral A is defined by

$$A = \int_{t_1}^{t_2} \sum_j p_j\dot{q}_j dt. \tag{176}$$

Now, a virtual displacement does not correspond to an actual physical displacement, as was pointed out in Sec. 1.7. In the latter type of displacement, H is conserved, while in a virtual displacement it is not conserved. If we wish to consider a new type of displacement, which moves to a neighboring trajectory but still conserves H, it is clear that it must involve a change in time. Thus, the transit times of points along varying paths obtained with this new variation need not be constant, and the end points may differ in time also although the variations of the q_i at the end point remain zero.

We define this new type of variation Δq_i by tagging each trajectory with a parameter α. Since each curve has the parameter t associated with each point, t must be a function of α. Thus, we must have $q_j = q_j(\alpha,t)$. We now expand δq_j and obtain

$$\Delta q_j \rightarrow d\alpha\left(\frac{dq_j}{d\alpha}\right) = d\alpha\left(\frac{\partial q_j}{\partial\alpha}\right) + d\alpha\left(\frac{\partial q_j}{\partial t}\right)\left(\frac{dt}{d\alpha}\right). \tag{177}$$

The first term is just δq_j, while the last is Δt since t depends only on itself, and thus the change in t is just a Δ variation. Thus Eq. (177) becomes

$$\Delta q_j = \delta q_j + \dot{q}_j\Delta t. \tag{178}$$

Since this relation between the two difference operators holds for any function $f(\mathbf{q},t)$, we have the more general relation

$$\Delta f(\mathbf{q},t) = \delta f + \dot{f}\Delta t = \sum_j \left(\frac{\partial f}{\partial q_j}\delta q_j\right) + \left[\sum_j \left(\frac{\partial f}{\partial \dot{q}_j}\delta\dot{q}_j + \frac{\partial f}{\partial t}\right)\right]\Delta t. \tag{179}$$

The first sum represents the expansion of the change in f due to a virtual displacement; the second, the expansion of the time derivative of f.

With this mathematical apparatus at our disposal, we can now prove that $\Delta A = 0$, which tells us that the action integral is an extremum along a dynamical path. Applying the operator $\varDelta$ to the action integral gives

$$\begin{aligned}\varDelta A &= \varDelta \int_{t_1}^{t_2} (L + H)dt \\ &= \varDelta \int_{t_1}^{t_2} Ldt + H(\varDelta t_2 - \varDelta t_1). \qquad (180)\end{aligned}$$

If we now represent the integral $\int Ldt$ as $I(t)$, then the definite integral in Eq. (180) is $I(t_2) - I(t_1)$, and its variation is $\varDelta I(t_2) - \varDelta I(t_2)$, and we have

$$\begin{aligned}\varDelta \int_{t_1}^{t_2} Ldt &= \varDelta I(t_2) - \varDelta I(t_1) \\ &= \delta I(t_2) - \delta I(t_1) + \dot{I}(t_2)\varDelta t_2 - \dot{I}(t_1)\varDelta t_1 \\ &= \delta I + L\varDelta t|_{t_1}^{t_2}. \qquad (181)\end{aligned}$$

Since we have not required the δq_j to vanish at the end points of the integral, we cannot use Hamilton's principle to assert that δI vanishes. Instead, we must calculate δI by expanding δL and then using Lagrange's equations and Eq. (178). We have

$$\begin{aligned}\delta \int_{t_1}^{t_2} Ldt &= \int \delta Ldt \\ &= \int \sum_j \left(\frac{\partial L}{\partial q_j}\delta q_j + \frac{\partial L}{\partial \dot{q}_j}\frac{d}{dt}\delta q_j\right)dt \\ &= \sum \int \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_j}\delta q_j\right)dt \\ &= \sum \int \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_j}\varDelta q_j - \frac{\partial L}{\partial \dot{q}_j}\dot{q}_j\varDelta t\right)dt \\ &= \sum \left(\frac{\partial L}{\partial \dot{q}_j}\varDelta q_j - \frac{\partial L}{\partial \dot{q}_j}\dot{q}_j\varDelta t\right)\Bigg|_{t_1}^{t_2} \\ &= -\sum p_j\dot{q}_j\varDelta t|_{t_1}^{t_2}, \qquad (182)\end{aligned}$$

since $\partial L/\partial \dot{q}_j = p_j$. (Note that the Δq_j's vanish at the end points, but Δt does not.) Combining the above results yields

$$\varDelta A = \left(-\sum_j p_j\dot{q}_j + L + H\right)\varDelta t|_{t_1}^{t_2} = 0. \qquad (183)$$

This completes the proof that the action integral is an extremum.

Example: Consider a single particle of mass m, total energy E, and kinetic energy T moving through a force field with potential energy $V(x,y,z)$. Clearly, the force field is conservative. The action integral becomes

$$\begin{aligned}A &= \int_{t_1}^{t_2} 2Tdt = \int_A^B [2m(E - V)]^{1/2}ds \\ &= \int_A^B p(E,x,y,z)ds, \qquad (184)\end{aligned}$$

where ds is the element of arc length along a path in three-dimensional space whose end points are (A,B), and p is the momentum of the particle, which clearly depends on space and the total energy. The principle of least action tells us that

$$\varDelta \int_A^B p(E,x,y,z)ds = 0 \qquad (185)$$

on a dynamical path.

The action integral is most important in H-J theory. Consider Hamilton's characteristic function $W(\mathbf{q},\boldsymbol{\beta})$, which we recall is the steady-state part of Hamilton's principal function $S(\mathbf{q},\beta,t)$. For the steady-state case, Eq. (154) becomes

$$\frac{\partial W(q_1,\ldots,q_n,\beta_1,\ldots,\beta_n)}{\partial q_i} = p_i. \qquad (186)$$

The total time derivative of W is

$$\frac{dW}{dt} = \sum_j \frac{\partial W}{\partial q_j}\dot{q}_j = \sum_j p_j\dot{q}_j. \qquad (187)$$

Integrating this from t_1 to t_2 yields

$$W = \int_{t_1}^{t_2} \sum_j p_j\dot{q}_jdt. \qquad (188)$$

Comparing this equation with Eq. (176), we see that Hamilton's characteristic function W is indeed the action integral. Therefore, with respect to the H-J theory, the principle of least action tells us that W is an extremum on a dynamical path. This demonstrates the importance of the principle of least action in H-J theory. H-J theory invokes the more realistic condition that all paths neighboring the dynamical path satisfy the conservation of energy, rather than Hamilton's principle where energy is not conserved along neighboring paths.

7. ACTION-ANGLE VARIABLES AND PHASE INTEGRALS

Dynamical systems undergoing periodic motion are most important in classical mechanics. Action-angle variables and phase integrals are introduced into H-J theory to handle such problems. Before treating these topics, we first briefly discuss periodic motion.

Essentially, two types of periodic motion occur in dynamics. We illustrate them by one-dimensional models in phase space.

1. A harmonic oscillator. Both q and p are periodic functions of time with the same frequency. Since q and p return to their starting values after one period, it is clear that the trajectory in phase space is closed.
2. Rotation. Here q is not periodic, but when it is increased by some constant value the configuration of the system remains unchanged. Suppose q is the angle of rotation about the axis of rotation of a rigid body. Increasing q by 2π radians produces no change in the system configuration. On the other hand, p is periodic. The trajectory in phase space is not closed.

To illustrate both types of periodic motion in a one-dimensional model, we consider the simple pendulum oscillating with no friction in the vertical plane. Let q be the angle θ of deflection from the vertical and p_θ be the conjugate momentum. The conservation of energy tells us that $E = p_\theta^2/2mL^2 - mg\cos\theta$, where L is the length of the pendulum suspension, m its mass, E the total energy, and g the acceleration of gravity. Solving this equation for p_θ gives

$$p_\theta = [2mL^2(E + mgL\cos\theta)]^{1/2}. \qquad (189)$$

If the motion is periodic, then one period extends between two consecutive zeros of p_θ. For a zero, we must have

$$\cos\bar{\theta} = -E/mgL. \qquad (190)$$

There are three possibilities:

1. If $E < mgL$, then the motion is of the vibration type where the pendulum oscillates between $\bar{\theta}$ and $-\bar{\theta}$.
2. If $E > mgL$, all values of θ are possible, so that the angle can increase without limit, producing periodic motion of the rotation type. In this case, the total energy is great enough to swing the pendulum through $\theta = \pi$ so that it will continue to rotate.
3. If $E = mgL$, we have a transitional case between **1** and **2**.

Figure 12 shows the three types of trajectories in phase space for the simple pendulum. Curve (1) is periodic of the vibration type. Curve (2) is periodic of the rotation type. Curve (3) is the transitional curve.

We now turn to the more general case of systems with more than one degree of freedom. We restrict the discussion to those periodic problems where Hamilton's characteristic function W is separable in at least one set of canonical variables. We repeat Eq. (186):

$$p_j = \frac{\partial W_j(q_j,\beta_1,\ldots,\beta_n)}{\partial q_j}. \qquad (191)$$

The solution of this equation is

$$p_j = p_j(q_j,\beta_1,\ldots,\beta_n). \qquad (192)$$

We see that Eq. (192) is the projected trajectory in phase space of the system point in the (q_j,p_j) plane. Periodic motion will occur if Eq. (192) describes either a closed orbit or a periodic function of q_j. Note that it is not necessary for each coordinate and momentum pair to have the same frequency. For the example of a three-dimensional oscillator with different components of the spring force con-

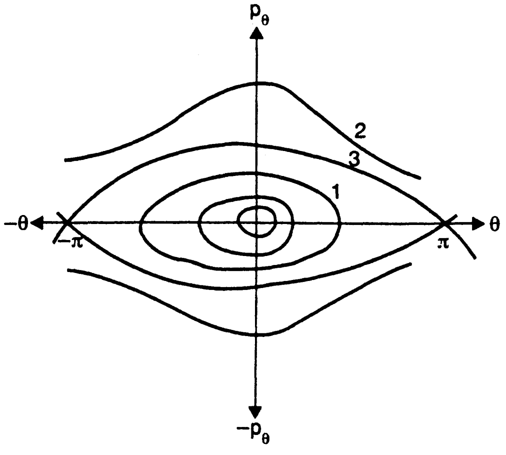

FIG. 12. Types of trajectories in phase space for a simple pendulum.

stant (different frequencies), the complete motion is not necessarily simply periodic. Indeed, if the separate frequencies are not rational fractions of each other, the phase trajectory will be an open Lissajous figure. Such a trajectory is called *conditionally periodic*.

We now discuss the *action variables* and the corresponding *phase integrals*. For a system with n degrees of freedom, we have n action variables $J_1, J_2, \ldots, J_n$ defined by

$$J_i = \oint p_i dq_i. \tag{193}$$

J_i is the ith action variable, and the contour integral on the right-hand side is the corresponding phase integral. The integration is carried out over a complete period of oscillation or rotation of q_i, as the case may be. Note the resemblance of the phase integral to the action integral given by Eq. (176) for the ith component. Using Eq. (186) we get

$$J_i = \oint \frac{\partial W(q_i,\beta_1,\ldots,\beta_n)}{\partial q_i} dq_i. \tag{194}$$

It is clear that each action variable J_i is a function only of the n constants of integration appearing in the solution of the H-J equation. Note that these constants are the n components of the constant momentum, which may be taken as the initial value of the momentum. Since the generalized coordinates and momenta are linearly independent, it follows that the n action variables form n independent functions of the β_i's and hence are suitable as a new set of constant momentum components. This means that we can express the constant momentum components as functions of the components of the action variable. We may thus write W as

$$W = W(q_1,\ldots,q_n;J_1,\ldots,J_n), \tag{195}$$

while the Hamiltonian appears only as a function of the action variables.

The *angle variable* w_i conjugate to J_i is a generalized coordinate defined by

$$w_i = \frac{\partial W_i}{\partial J_i} \tag{196}$$

If we now turn to the first equation in (131), we see that q_i is analogous to w_i (since it is a generalized coordinate), and p_i is analogous to J_i. Therefore, the equations of motion in terms of the action and angle variables are

$$\dot{w}_i = \frac{\partial H(J_1,\ldots,J_n)}{\partial J_i} = \nu_i(J_i,\ldots,J_n), \tag{197}$$

where the ν_i's are constant functions of the action variables. We immediately solve Eq. (197) and obtain

$$w_i = \nu_i t + \alpha_i, \tag{198}$$

where α_i is the constant of integration. We can always calculate q_i as a function of t, ν_i, and α_i from the above equations. But the real value of this analysis lies in the physical interpretation of the ν_i's. We show that each ν_i is the frequency associated with the periodic motion of the corresponding q_i.

To show this, we consider the change Δw_i in the angle variable w_i when the coordinate q_i goes through a complete cycle of vibration or rotation. Let δw_i be the infinitesimal change as a result of an increment in an arbitrary q_j; we write

$$\Delta w_i = \oint \delta w_i, \quad \text{where } \delta w_i = \frac{\partial w_i}{\partial q_j} dq_j. \tag{199}$$

We now use Eq. (196) and perform the following calculation:

$$\begin{aligned}\Delta w_i &= \oint \delta w_i = \oint \frac{\partial w_i}{\partial q_j} dq_j = \oint \frac{\partial^2 W}{\partial q_j \partial J_i} dq_j \\ &= \frac{\partial}{\partial J_i} \oint \frac{\partial W}{\partial q_j} dq_j = \frac{\partial}{\partial J_i} \oint p_j dq_j = \frac{\partial J_j}{\partial J_i} = \delta_{ij}.\end{aligned} \tag{200}$$

This tells us that w_i changes by a unit amount when q_i goes through a complete cycle, but is unaffected by a similar change in q_j. Let τ_i be the period associated with q_i. It follows from the above that $\nu_i = 1/\tau_i$. This is what we set out to prove: that ν_i is the period associated with the periodic variation of q_i.

To illustrate the use of the action-angle variables and the phase integrals, we discuss in detail the one-dimensional simple harmonic oscillator. There is only one action variable and one phase integral. They are obtained by constructing the Hamiltonian and the H-J equation and solving for W. The Hamiltonian is

$$H = p^2/2m + kq^2/2, \tag{201}$$

where k is the spring constant. The H-J equation is

$$\frac{1}{2m}\left(\frac{\partial W}{\partial q}\right)^2 + \frac{kq^2}{2} = E. \tag{202}$$

Solving for W, we get

$$W = (mk)^{1/2}\int(2E/k - q^2)^{1/2}dq. \tag{203}$$

The action integral or phase variable is

$$\begin{aligned} J &= \oint p\,dq = \oint \frac{\partial W(q,E)}{\partial q}\,dq \\ &= (2m)^{1/2}\oint\left(\frac{2E}{k} - q^2\right)^{1/2} dq \\ &= 2E\left(\frac{m}{k}\right)^{1/2}\int_0^{2\pi}\cos^2\theta\,d\theta = 2\pi E\left(\frac{m}{k}\right)^{1/2}, \end{aligned} \tag{204}$$

where we used the substitution $q = (2E/k)^{1/2}$ $\sin\theta$. (Note that we integrated over a complete cycle of q.) From this we get

$$E = H = \frac{J}{2\pi}\left(\frac{k}{m}\right)^{1/2}, \tag{205}$$

and the frequency of oscillation of the oscillator is

$$\frac{\partial H}{\partial J} = \nu = \frac{1}{2\pi}\left(\frac{k}{m}\right) = \frac{\omega}{2\pi}, \tag{206}$$

where ω is the angular frequency.

GLOSSARY

Angle of Inclination: Angle between fixed vertical axis and axis of symmetry of top.

Angle of Spin: Angle between x' axis of top and line of nodes.

Azimuth Angle: Angle between fixed x axis and line of nodes.

Configuration Space: $2n$-dimensional space whose axes are n generalized coordinates and n generalized velocities.

Constraints: Conditions imposed on coordinates, making them not independent of one another.

Cyclic Coordinate: A generalized coordinate not appearing in the Lagrangian. Its conjugate momentum is conserved.

D'Alembert's Principle: A method of describing equations of motion by treating the negative of the time derivative of momentum as a force, so that the method of virtual work may be applied.

Eulerian Angles: Three independent angles that define the orientation of a rotating body with one point fixed. See **Angle of Inclination, Angle of Spin**, and **Azimuth Angle**.

Euler's Equation: A partial differential equation obtained as a necessary condition for existence of an extremum of a functional. When used in Hamilton's principle, Euler's equation becomes Lagrange's equation of motion.

Functional: A definite integral involving a class of integrands; a function whose value depends on which of a set of functions is chosen as its argument.

Generalized Coordinates: A set of independent variables that define the position of a dynamical system.

Generalized Momentum: A dynamical variable associated with a particular generalized coordinate in the same fashion as a component of linear momentum is associated with the corresponding Cartesian coordinate.

Generalized Velocity: The time derivative of a generalized coordinate.

Hamiltonian: A function of the generalized coordinates and momenta, defined as the sum of the products of the generalized velocities by the corresponding generalized momenta, minus the Lagrangian. For a conservative system, it is equal to the sum of the kinetic and potential energies.

Hamilton's Principle: The statement that the actual path of a dynamical system between two fixed points in configuration space is that which minimizes the time integral of the Lagrangian between the two points.

Lagrangian: A function of the generalized coordinates and velocities equal to the difference between the kinetic and potential energies; sometimes called the *kinetic potential.*

Line of Nodes: In the motion of a top, the intersection of the horizontal fixed plane and the plane normal to the axis of symmetry of the top. Also, in treating the motion of a planet, the line of intersection of the plane of the orbit with the equatorial plane.

Nutation: Motion of a top characterized by time variation of the angle of inclination.

Phase Space: $2n$-dimensional space whose axes are the n generalized coordinates and the n generalized momenta.

Precession: Motion of a top characterized by time variation of the azimuth angle.

Principal Axes of Inertia: A set of mutually perpendicular directions in a rigid body with respect to which the products of inertia vanish so that the inertia tensor is diagonal.

Virtual Displacement: An infinitesimal variation of coordinates compatible with any constraints that may be acting, at constant time.

Virtual Work: Work done on a dynamical system in the course of a virtual displacement.

Works Cited

Birkhoff, G., Maclane, S. (1977), *A Survey of Modern Algebra*, 4th ed., New York: Macmillan.

Landau, L. D., Lifshitz, E. M. (1976), *Course of Theoretical Physics*, Vol. 1, *Mechanics*, 3rd ed., London: Pergamon.

Noble, B., Daniels, J. W. (1977), *Applied Linear Algebra*, 2nd ed., Englewood Cliffs, NJ: Prentice-Hall.

Poincaré, Henri (1902), *Science et l'Hypothèse*, Paris: Flammarion; translation (1952), New York: Dover, Chaps. 6 and 7.

Lord Rayleigh (1878), *The Theory of Sound*, New York: Macmillan; reprint (1945), New York: Dover, Vol. 1, Chap. 4.

Synge, John L., Griffith, Byron A. (1959), *Principles of Mechanics*, 3rd ed., New York: McGraw-Hill.

Whittaker, E. T. (1937), *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, 4th ed., London: Macmillan; reprint (1974), New York: Dover.

Further Reading

Corben, H. C., Stehle, Philip (1974), *Classical Mechanics*, 3rd ed., New York: Wiley.

Davis, Julian L. (1990), *Wave Propagation in Electromagnetic Media*, Berlin: Springer-Verlag. Appendix to Chap. 3 on similarity and orthogonal transformations; Chap. 5 on canonical transformations and Hamilton-Jacobi theory.

Davis, Julian L. (1988), *Wave Propagation in Solids and Fluids*, Berlin: Springer-Verlag, Chap. 9. Variational methods, Hamilton's principle, Lagrange's and Hamilton's equations, etc.

Feynman, R. P., Leighton, R. B., Sands, M. (1963), *The Feynman Lectures on Physics*, Vol. 1, Reading, MA: Addison-Wesley.

Goldstein, Herbert (1950), *Classical Mechanics*, 1st ed., (1980), 2nd ed., Reading, MA: Addison-Wesley.

Klein, Felix (1897), *The Mathematical Theory of the Top*, New York: Scribners; reprint (1967) in: *Congruences of Sets and Other Monographs*, New York: Chelsea.

Lanczos, Cornelius (1970), *The Variational Principles of Mechanics*, 4th ed., Toronto: University of Toronto Press.

Landau, L. D., Lifshitz, E. M. (1976), *Course of Theoretical Physics*, Vol. 1, *Mechanics*, 3rd ed., London: Pergamon.

Meirovitch, Leonard (1970), *Methods of Analytical Dynamics*, New York: McGraw-Hill.

Synge, John L., Griffith, Byron A. (1959), *Principles of Mechanics*, 3rd ed., New York: McGraw-Hill.

Whittaker, E. T. (1937), *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, 4th ed., London: Macmillan; reprint (1974), New York: Dover.

MEMORIES

Celia E. Yeack-Scranton*, *IBM Corporation, San Jose, California, U.S.A.*

Introduction 61
1. **General Principles and Brief History** 62
2. **Physical Phenomena Associated with Memory Devices** 64
2.1 Magnetism 64
2.2 Magneto-optics and Optical Phenomena 65
2.3 Charge Storage 65
2.4 Ferroelectricity 66
3. **Design, Function, and Operation** 67
3.1 Magnetic Devices 67
3.1.1 Core Memories 67
3.1.2 Rotating-Disk Storage 68
3.1.3 Tape Devices 70
3.1.4 Bubbles 71
3.2 Magneto-optic and Optical Storage 72
3.3 Charge Storage 73
3.3.1 Static Random-Access Memory 74
3.3.2 Dynamic Random-Access Memory 74
3.3.3 Read-Only Memory 75
3.3.4 Programmable Read-Only Memory 75
3.3.5 Electrically Erasable PROM 76
3.4 Ferroelectric Storage 77
4. **Applications** 78
Glossary 79
Works Cited 79
Further Reading 79

INTRODUCTION

Elements of modern computing were envisioned by Charles Babbage in his Analytical Engine of 1868, but it was the invention and practical implementation of punched-card data-processing machines by Hollerith and the use of these machines in the 1890 U.S. census that began a steady progression of invention and development toward modern computer memories.

A key challenge in computing technology is in data memory. Memory must provide accurate, long-term retrieval of data at access speeds compatible with the computing device. As the speed, accessibility, and uses of computing have increased, so have the types, uses, and sheer volume of memory. The economic significance has also mushroomed, amounting to a $52 billion industry in 1990 in magnetic storage, the largest single component. A brief overview of the general features of memory and its evolution as a synergy between invention and commercial development is given in Sec. 1.

Memories continue to become faster, denser, less expensive, lower in power consumption, and more rugged. Yet the underlying physical phenomena used in today's practical memory and storage devices are remarkably the same as in the 1950s and 1960s, when the first modern electronic computers with program storage and significant reliance on memory devices were introduced. Magnetic and semiconductor technologies remain dominant, with emerging devices based on optics, magneto-optics, and ferroelectrics. These phenomena are reviewed in Sec. 2.

The economic success of computing has supported a rapid pace of invention and development of many memory technologies. Historically, each new device has solved a particular technical problem and/or made possible a new application area in computing. Continued utility of the device has depended on the technological extendability and the rate of improvement of cost and function relative to that of other technologies. The design, function, and operation of a selection of devices are described in Sec. 3. Magnetic core memories are included for historical signifi-

---

*Deceased.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

cance. Tape, disk, and dynamic random-access memory (DRAM) are economically the most significant. Magneto-optic devices, electrically erasable programmable read-only memory (EEPROM), and ferroelectric memories offer potentially significant opportunities for the future. Static random-access memory (SRAM), read-only memory (ROM), programmable read-only memory (PROM), and bubble devices fill other markets.

The ideal memory device would simultaneously be the fastest, least expensive, most dense, lowest in power consumption, longest lived, and most robust. In practice, however, optimization is achieved at the system level via a computer storage hierarchy. Here, small, fast, expensive memories are used close to the (fast) processor. Additional levels of memory are configured in tiers away from the processor, each level having an increasing number of bytes at decreasing cost and performance. Those in which access time to any bit is independent of its physical location are termed "random access," and are generally referred to as "memory." Those that store data in serial blocks, requiring much longer access times, are generally termed "storage." Typical hierarchies and new applications made possible by the current trends in cost, size, and power reduction will be discussed in Sec. 4.

1. GENERAL PRINCIPLES AND BRIEF HISTORY

Tabulation for large businesses and government was the dominant use of the early electromechanical computers used in the first half of the 20th century. In these systems, a few numbers were physically stored in mechanical relays and, later, dials. Storage and retrieval times were limited by the motion of these large mechanical structures, and this rudimentary memory was used primarily for looking up constants, for instance, in insurance tables as the computation proceeded.

By the late 1940s, a recognized need in the fledgling computer industry was for large-capacity data storage. Focused projects, notably to provide a "defense calculator," fostered a series of inventions and developments leading to the introduction of the first modern computers, such as the IBM type 701 system, in 1952. Both the memory types and the mix of memory devices in this computer bear similarities to those in use today. Electrostatic devices, later supplanted by magnetic cores and then by semiconductor memories, provided fast-access memory that worked in synchrony with the computing unit. The capacity of 2048 words of 36 bits each was accomplished by use of 72 cathode ray tubes. Since the tubes could be operated in parallel, all the bits from a word were read or written simultaneously, giving a word every few microseconds. This random-access memory was "dynamic," as it needed to be rewritten, or "refreshed," every 12 ms since the charge stored in the tubes dissipated in that time. Thus, when the power was turned off, information stored in the electrostatic storage unit was lost and, hence, was termed "volatile."

The IBM 650 magnetic drum memory in the 701 computer provided a slower but more cost-effective and nonvolatile storage. Here, a specially designed electromagnetic "head" (described further in Secs. 2.1 and 3.1) was used to write individually magnetized "bits" or regions onto a magnetic thin-film surface coated onto a rotating cylindrical drum. By the inverse process, the head could be used to sense the presence of bits, or read information, since a voltage was produced as the bit pattern passed under the head. The head was held at a specific spacing above the drum, since abrasive wear from contact would quickly destroy both the head and drum. Rotation allowed writing around one circumference, creating a "track" of data. Arranging many heads side by side parallel to the drum axis allowed the creation of multiple tracks, and five were read or written in parallel to increase the word speed. The 10-cm (4-in.) diameter, 40-cm (16-in.) long drum was rotated at 12 500 rpm to provide an average access (half-rotation) time of 2.4 ms to the first bit of a desired set of data. Four drums storing a total of 8192 36-bit words were arranged in a unit. They worked in synchrony with the main electrostatic computer memory, with some clever designs to minimize the effect of the long access time.

The function of the IBM 701 computer was augmented greatly with the introduction of the RAMAC, a magnetic disk file, in 1956. Here, the rotating medium was a stack of disks, and a single head pair was moved from track to track and from disk to disk to read and write all of the data surfaces, one head for the upward-facing disk surfaces and the other head for downward-facing surfaces. This allowed up to 5 megabytes (MB) of information to be stored, but with a maximum retrieval time of 0.9 s due to the mechanical motion required to insert the head into the disk stack at the desired disk and to move to the correct track. The use of a single head pair minimized cost, but because of the enormous performance difference, drum memory was reserved for frequently used information while larger volumes of data were stored on the RAMAC.

The memory in Remington Rand's UNIVAC computer, announced in 1952, was based on a technology in which a sequence of electrical pulses was converted to a sequence of ultrasonic pulses, launched down a delay line consisting of a cylinder filled with mercury, and reconverted to the electrical waveform at the other end. Since sound travels about 10 000 times more slowly than electrons, much computation was accomplished in the processor during the acoustic propagation time, and the results from that computation were used to update the reconverted electrical pulse stream. After the update, the revised electrical pulse sequence was converted to ultrasonic pulses again, and the process continued indefinitely. While attractive in its simplicity, it was difficult to scale this memory type to larger capacities, and this memory was supplanted within a few years by ones using magnetic ferrite cores.

Magnetic core memories (described in Sec. 3.1.1) were initially introduced as a buffer between the input/output devices (magnetic tape and punch cards) and the main computer. However, since their random-access capability eliminated the access delay associated with drum memory, since their reliability exceeded that of the electrostatic storage tubes, and since they were more scalable to smaller dimensions (and thus larger storage capacities) than the acoustic delay-line memories, they quickly supplanted all three and persisted as a dominant choice for main memory until the mid-1970s. At that time, further shrinking of the individual cores to increase device capacity became less cost effective than the fabricating of integrated-circuit semiconductor memories.

Magnetic thin-film memories were developed in the 1960s, using the same principles as core memories, but the structures were created in a thin-film device, facilitating miniaturization by eliminating discrete wires and cores. While used in some products, they were supplanted quickly by the rapidly evolving semiconductor memory devices.

Semiconductors emerged as early as the 1960s as a basis for memories, but early devices relied upon assembling many discrete transistors and components in arrays and were less reliable than other choices. In 1959, however, integrated circuits were first fabricated. In integrated circuits, many transistors and other components are built on the surface of a silicon wafer and interconnected with thin films of metal. This integration made possible an ever decreasing cost per bit by steadily decreasing the size of each cell with improvements in design and processing. Higher reliability and speeds accompanied this, and semiconductor memory began displacing magnetic core memories in the early 1970s. Static random-access memory, which requires a number of transistors to store each bit, was later replaced with DRAM, which uses only a single transistor per storage bit but requires constant refreshing of the information. Semiconductors remain the highest performance memories available, with a clear scalability to much finer dimensions, faster speeds, and lower costs for at least one more decade.

In addition to storage and memory devices in which data are both read and written by the user, a class of memory devices is designed for reading fixed data. Read-only memory (ROM), first proposed in the 1960s, was initially a device in which data were fixed at the time of manufacturing by the presence or absence of transistors at each bit cell. Later devices allowed the user to "burn in" connections, creating a desired bit pattern. More modern devices store charge in such a way that it is very hard to remove, allowing some rewriting (up to 10^6 cycles). Of these, Flash EEPROMs are gaining rapid acceptance, although slow write and erase times and limited cyclability leave room for improvement.

Flash EEPROM memories offer nonvolatile random access storage with read/write capability (although writing is limited). This avoids the time delays associated with most magnetic storage but offers nonvolatility, unlike conventional DRAM and SRAM. Other technologies, such as ferroelectric random-access memory (RAM), achieve nonvolatility

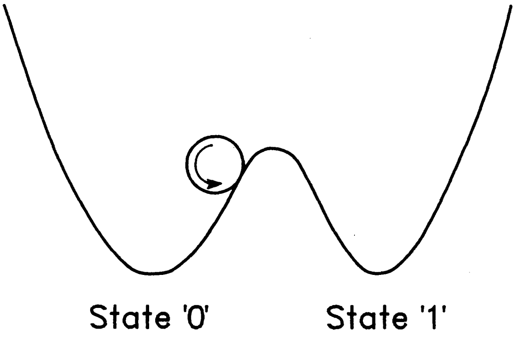

FIG. 1. Simple illustration of the principles of memories using a ball rolling between two wells.

by use of different materials, and lower-power conventional DRAM designs facilitate the use of batteries to achieve nonvolatility. Cost, performance, and application determine which nonvolatile technology is used, and increased use of portable, high-performance computers makes this a direction of growth.

2. PHYSICAL PHENOMENA ASSOCIATED WITH MEMORY DEVICES

This section describes the phenomena that are key to the different memory technologies. Each fulfills basic requirements for useful devices. In particular, the system must exhibit bistability, or two stable states, so that 1's and 0's can be represented. A simple example, Fig. 1, can be used to illustrate this and other attributes. Here, a ball can roll into either of two potential wells. To move from one well to the next requires sufficient energy to roll over the hill in the middle. A smaller hill will require less switching energy, but it must not be so low that a ball may accidentally roll from one state to the other. Switching must be as fast as possible, and it is desirable that neither the hill nor the balls wear out with time or use. The state of the ball should be read without moving it and with minimal expended energy, although in some memory devices, the read process can be likened to plucking the ball from the hole, then putting it back.

2.1 Magnetism

Magnetic storage is easily visualized by imagining a string of permanent magnets in which individual magnets can be flipped to give a desired pattern of north and south poles, as shown in Fig. 2. The orientation of each magnet can be viewed as the "1" or "0" state described above. Controlling and sensing the orientation of these magnets are the analogs of writing and reading.

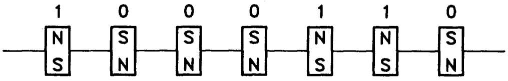

FIG. 2. A string of magnets oriented to represent a string of 1's and 0's, analogous to data storage on magnetic media.

Practical magnetic devices use continuous materials in which the magnetization (direction of the internal magnetic flux) can be locally flipped within small regions of the material. The most common materials used are ferromagnetic or ferrimagnetic, whose properties are illustrated in the hysteresis loop in Fig. 3. Initially, without an applied field *H*, the spontaneous magnetization direction, *M*, is comprised of tiny domains or particles whose directions tend to cancel each other, giving rise to a net $M = 0$ (*a* in Fig. 3). However, with a large enough applied field, these domains line up in the direction of the field (*b* in Fig. 3). Then, even after the field is removed, a remanent magnetic flux density M_r remains (*c* in Fig. 3). This remanence is analogous to one orientation of the magnets of Fig. 2 and determines the strength of the readback signal for magnetic recording. The

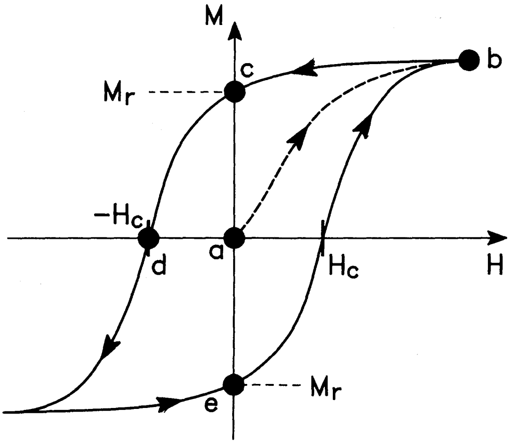

FIG. 3. Hysteresis loop showing the relationship between the applied field *H* and the resulting internal field *M* for a ferromagnetic or ferrimagnetic material.

magnetization can be reduced to zero by applying a sufficiently large field in the opposite direction, $-H_c$ (*d* in Fig. 3). An even larger opposing field causes a reversal of this magnetization and an opposite sense of remanence, $-M_r$ (*e* in Fig. 3), analogous to the oppositely oriented magnets of Fig. 2. H_c is called the coercive force or coercivity.

A key advantage of magnetic storage is that it is nonvolatile. That is, the magnetization direction is permanent until reset, even in the absence of power. Further, the magnetization direction can be flipped as many times as desired without wearing out the material, giving it infinite cyclability. Magnetic switching times are fundamentally limited only by the gyromagnetic resonance, which is in the gigahertz range. Current device designs are still an order of magnitude away from this. The minimum stable domain size, determined by the product of media magnetization and coercive force, has continued to decrease with better materials. Thus, given continued engineering advances, continued magnetic recording density improvements are possible for at least ten years.

2.2 Magneto-optic and Optical Phenomena

Magneto-optic storage relies on creating tiny magnetized regions, in a thin film, which interact with an optical beam to alter its reflection. The magneto-optic Kerr effect is commonly utilized, and the physics of this phenomenon is illustrated in Fig. 4. A thin film of magneto-optic material is used as the storage medium, with the magnetization oriented vertically up or down with respect to the surface of the film. To write information onto the thin film, a laser pulse is focused onto a tiny spot in the presence of a weak external magnetic field opposing the magnetization direction of the overall film. The laser locally heats the material very quickly above a critical temperature (T_c) at which the external magnetic field aligns the magnetization to this reverse direction, creating a stored bit. As in magnetic recording, the recording surface is moved under the head to write or read a sequence of bits. Reading is accomplished by sensing a change in polarization of a reflected low-powered laser beam as the stored bit passes under it. Erasing is done in exactly the same way as writing except with a reversed external field. The areal recording density for magneto-optic storage is very high since the optical beam used to create it can be focused to about a micrometer.

Another important optical storage mechanism for read-only applications is to use optical contrast between featured and unfeatured regions of a smooth disk. The pattern of features represents the bits. The simplest way of creating these features is to stamp pits in a plastic disk with a master. This mechanism is the basis for consumer compact-disk (CD) players and fixed data storage in CD-ROM (CD read-only memory). Another mechanism relies upon using the laser heat pulse to create spots on the disk with material structures different than those of the surrounding material, for instance, a different crystal phase. This is the basis for "write-once" optical storage.

For all optical storage phenomena, the potential number of stored bits per area increases as the wavelength is decreased, since this tightens the beam focus. Several generations of this fundamental type of improvement are expected before radical changes are necessary.

2.3 Charge Storage

Charge storage is easily visualized by imagining a planar grid of wires with a box (cell) comprised of a charge bucket and a valve

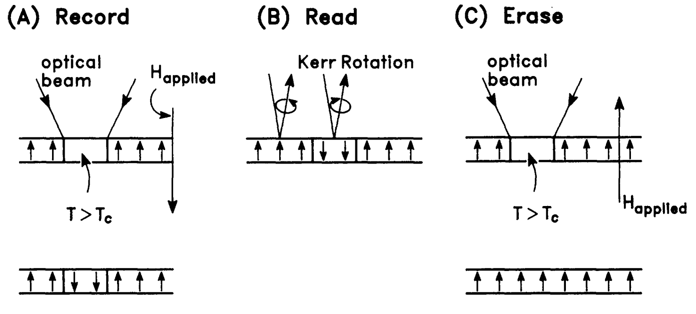

FIG. 4. Magneto-optic recording processes.

at each intersection, as sketched in Fig. 5. The valves for all the buckets in a row are turned on in unison through a single control (the word line), and charge can then be added to the desired buckets through the appropriate vertical (bit) lines, creating the desired charge storage pattern. Sensing is done by switching on the row, then measuring the presence or absence of charge in each charge bucket as a 1 or a 0 at the respective vertical line. This matrix scheme allows minimal space to be used for addressing, decoding, and routing data.

Early charge storage was accomplished by means of cathode ray tubes, but these devices were supplanted by semiconductors, as silicon-integrated circuits. While many attributes of today's devices vary, the underlying principles remain the same. Bit cells are arranged in a grid as described above, where the rows are word lines and the columns are bits lines. In each cell, transistors are used as valves, and charge can be put on a capacitor as a voltage. Sensing is done by measuring the stored charge as a voltage or as a current when the voltage is drained off.

Most semiconductor memory cells are based on metal-oxide-semiconductor field-effect transistors (MOSFETs), in which a conducting channel of electrons or holes is switched on or off by an electric field applied from the gate. In the *n*-channel MOSFET sketched in Fig. 6, a *p*-type semiconductor is implanted to form *n*-doped regions. The conductivity of a semiconductor depends on the doping and the applied field, exhibiting a very nonlinear increase in conductivity above a well-defined "threshold" voltage. When a voltage above the threshold is applied to the gate, a current of conducting electrons flows through this channel, between the source and the drain. This provides an electrically controlled switch. Most FET-based memory devices use a complementary metal-oxide-semiconductor (CMOS) in which *n*-channel and *p*-channel devices are fabricated on the same silicon substrate.

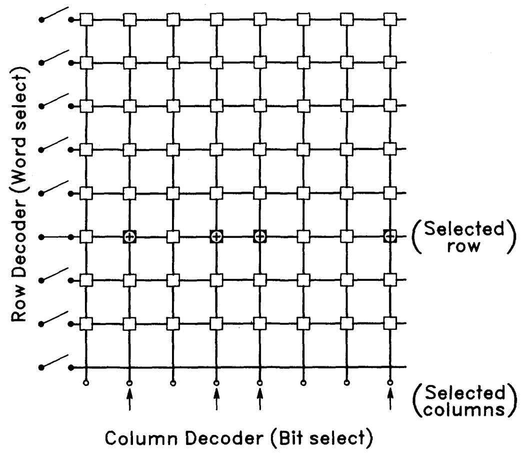

FIG. 5. Charge bucket array analogy illustrating the principles of random access semiconductor memory devices.

The number and arrangement of the transistors, choice of capacitor, and wiring determine the differences between the memory cells discussed in Sec. 3. Important functional differences are described briefly here as an introduction. "Static memory" is that in which the state of each cell is permanent until it is reset. "Dynamic memory" requires "refreshing" or rewriting the state of the cell periodically. Generally, both static and dynamic memory are "volatile"; that is, the state of each cell is lost when the power is turned off. Thus, "nonvolatile" memory is a type of memory in which the state of each cell is retained even in the absence of power. "Destructive readout" is a method of sensing the state of the cell in which the state is lost as a consequence. Hence, the cell is generally "restored" or rewritten after the bit is sensed. For all of these memory cells, important metrics are cell area (since this translates directly to cost), speed, and power consumed per bit cell.

2.4 Ferroelectricity

The use of a ferroelectric material as the insulator in a thin-film capacitor of a semiconductor memory cell provides the valuable attribute of nonvolatility in semiconductor memory. This arises from the unique properties of ferroelectrics, which exhibit a nonlinear dielectric behavior and retain a residual polarization when an external electric field is applied and then removed. When the field is reversed, spontaneous polarization in the opposite direction occurs. These two stable polarization states allow creation of an electrically controlled bistable charge state on the capacitor that is retained without any power for at least a year, hence providing nonvolatile memory.

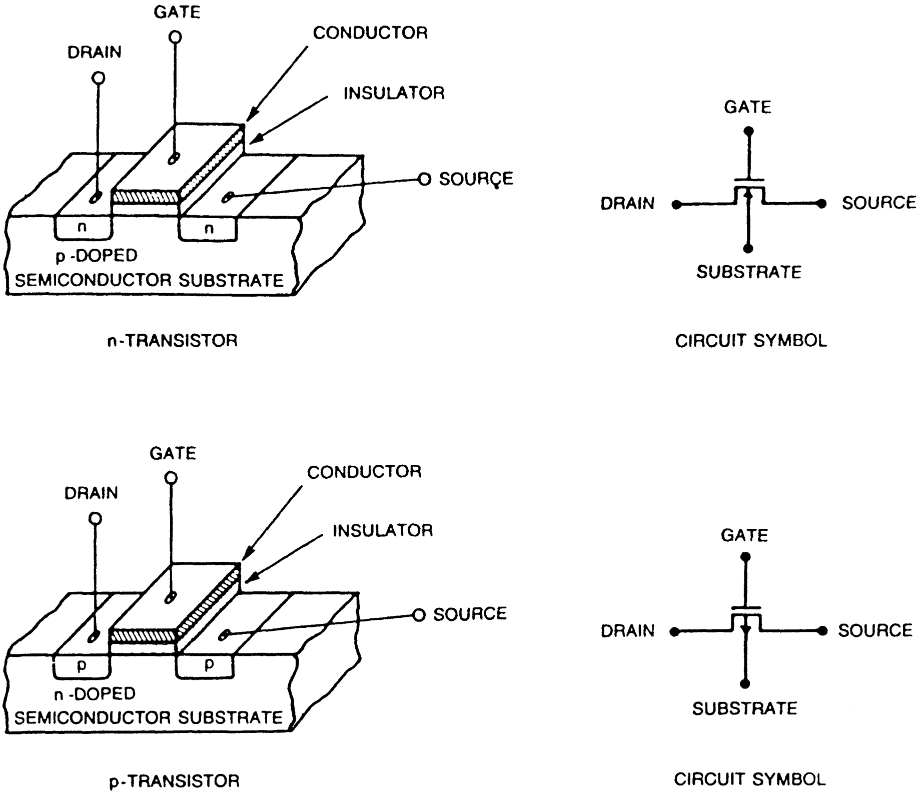

n-TRANSISTOR

CIRCUIT SYMBOL

p-TRANSISTOR

CIRCUIT SYMBOL

FIG. 6. MOSFET transistor and its circuit representation. (Weste and Eshraghian, 1985, © 1985 by AT&T Bell Laboratories, Inc. Reprinted by permission of Addison-Wesley Publishing Company.)

The intrinsic bistability is seen in a plot of polarization vs applied field, which gives a hysteresis loop analogous to those of ferromagnets (see Fig. 3). This similarity is the basis for the use of "ferro" in the name "ferroelectric." However, in ferromagnets, opposite magnetic polarizations are produced by flipping electron spin states, while in ferroelectrics, opposite electric polarization states are produced by altering the crystal structure. One important consequence of this is that the number of times the electric polarization can be flipped (cyclability) is much less than that for magnetic spins.

The readout of a ferroelectric material can be accomplished by applying a field that exceeds the coercive electric field. The two different polarization states give different responses, and the states can thus be distinguished. However, the memory cell then has to be rewritten to restore the original information. The limited cyclability (about a million cycles) and the somewhat slower switching speed compared to semiconductors are both serious constraints in some applications.

3. DESIGN, FUNCTION, AND OPERATION

3.1 Magnetic Devices

3.1.1 Core Memories Magnetic core memories are comprised of arrays of magnetic toroids, or cores. A magnetic core is a ring of ferrimagnetic material that can be magnetically switched from one circumferential magnetization direction to the opposite one. Switching is accomplished by applying a field in the desired direction to oppose the internal magnetization, as described in Sec. 2.1. The field is produced by applying a current to a wire passing through the core, as sketched in Fig. 7. Opposite currents give opposite fields.

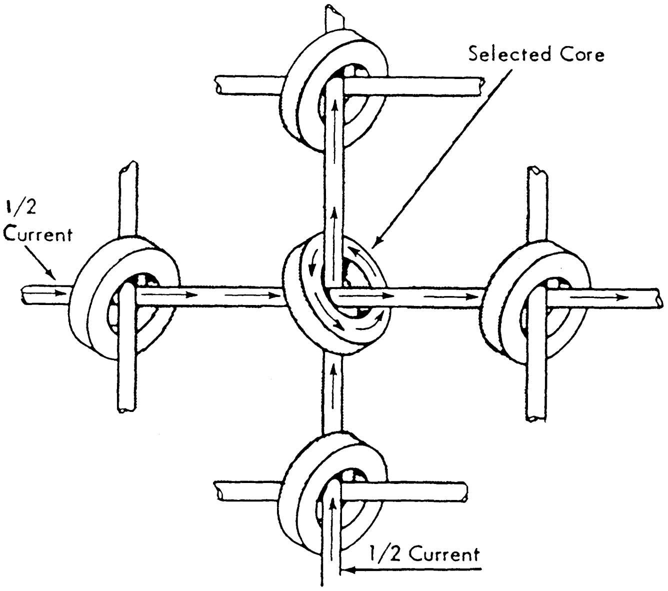

FIG. 7. A small section of a magnetic core memory, illustrating the principles of coincident-current selection. (IBM News, 10 November 1967)

Devices use the abrupt threshold for switching apparent in the hysteresis loop shown in Sec. 2.1. In particular, one-half of the coercive field is not sufficient to switch the core. Thus, to make a device, cores are arranged in a matrix so that there is a core at each intersection of two wires, one from the horizontal (*i*) and one from the vertical (*j*) direction. To selectively write a core at the (*i,j*) position, half of the required current is applied to each of the *i*th horizontal and *j*th vertical wires. Only at the (*i,j*)th core is there sufficient current to switch the core to a new state.

The matrix arrangement of cores allows random access to any desired bit (core). The switching speed of each bit is limited mainly by the electrical resonance of the core in the read/write circuit, typically in the megahertz range, giving about 1-μs cycle time per bit. Many bits are accessed in parallel for faster overall memory speeds. Readout is destructive, requiring a rewrite operation for continued storage.

Ferrite core memories were cost competitive in the 1960s with cycle times of about 1 μs and a cost of about $10 000/MB. However, assembling core memories required individually wiring each magnetic core onto a wire matrix, one core per stored bit, and was practically accomplished when memory sizes were only up to a few thousand bytes. The difficulties of fabricating these ceramic cores smaller than about 1 mm, materials frequency limits, and the challenge in wiring the tiny cores all limited continued cost-performance improvements relative to integrated semiconductor memories. The latter became the dominant random access memory choice by the mid 1970s.

3.1.2 Rotating-Disk Storage In rotating-disk storage, writing and reading information is accomplished with a magnetic recording head, which is conceptually described as a wire-wound toroid with a small slice removed to create a gap in the magnetic material, as shown in Fig. 8. Head operation is as follows. A current representing the input signal is applied to the coil. The magnetic field thus created goes across the gap but also induces a fringing field in the region below the head gap. This field magnetizes small regions of the moving magnetic medium, creating the desired bit pattern. These tiny magnetic poles in the material in turn cause magnetic flux to fringe out of the disk, which can be sensed by the same head as an induced voltage.

Early magnetic recording heads used machined magnetic cores similar to the one shown in Fig. 8. These were laminated onto a slider designed to float above the disk on a tiny cushion of air to prevent wear from rubbing against the disk. In 1979, thin-film heads were introduced in which the core was re-

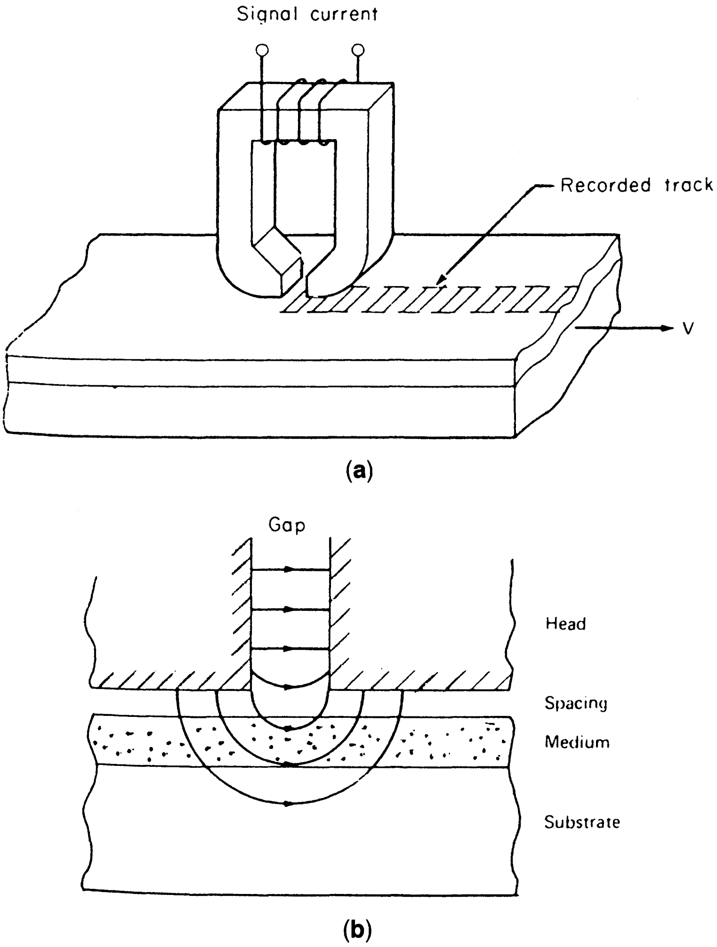

FIG. 8. Illustration of the recording process using a ring recording head. (a) Three-dimensional view; (b) cross section showing the magnetic field from the gap. (Mee and Daniel, 1987, © 1987 by McGraw-Hill, Inc. Reprinted by permission of C. D. Mee.)

placed with thin-film plated and patterned magnetic layers, the wires were replaced with patterned thin-film conductors, and the gap was replaced with a thin-film deposited structure. Many thin-film head sliders are created at once on the same wafer by means of multiple patterning and deposition steps in a manner analogous to semiconductor processing.

Further head improvements have been achieved by using heads separately optimized for reading and writing, fabricated in the same thin-film structure. The most sensitive read head used in the industry today is based on a magnetoresistive film, which has a different resistance for different directions of the magnetic field from the disk. This allows scaling to finer dimensions and higher-frequency operation, and hence higher bit densities.

An important attribute of rotating storage and tape applications is that information is stored on an unfeatured medium, within which tiny regions of material are individually switched. This means that while it may be very complex and costly to create the head that writes and reads the information, the head can be used to write many bits onto a relatively simple structure, generally lowering the overall device cost. Using this head, one track of information is created or read at a time, and the head is physically moved to write or read another track. In hard-disk storage, there are generally tens of thousands of bits per centimeter along the track direction and thousands of tracks per centimeter. A typical disk drive is sketched in Fig. 9. Note that much of the volume and power dissipation comes from the mechanical components required for disk rotation and track accessing.

Fundamental to the continued density in-

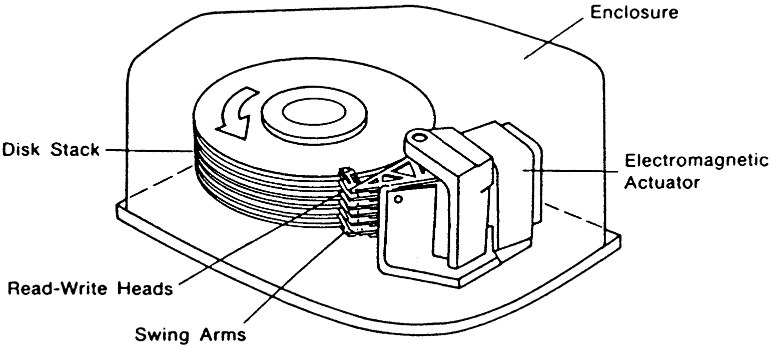

FIG. 9. Rigid disk drive. (Mee and Daniel, 1988, © 1988 by McGraw-Hill, Inc. Reprinted by permission of C. D. Mee.)

crease in magnetic storage is the ability to scale to finer dimensions. The near-field (fringe field) operation of the heads requires that the head-to-disk separation is decreased in proportion to the density increase, which in turn requires smoother, more durable surfaces to withstand the inevitable occasional contacts. Thin-film head dimensions must also be reduced. Materials and circuits must be improved to reduce magnetic and electronic noise. Finally, mechanical tracking accuracy must be increased. Nonetheless, continued extendability is expected for magnetic-recording hard-disk drives. Densities up to 300 Mb (megabits) per square centimeter, or 2 Gb (gigabits) per square inch, have been demonstrated. Product densities are at about 30 Mb/cm^2, increasing at 60% per year with no fundamental limits anticipated through the year 2000. Approximate cost and performance figures are $1/MB, access times of 12 ms, and data rates of 9 MB/s.

Disks for magnetic recording are created by coating a thin magnetic layer onto a very smooth substrate. Substrate materials are predominantly a highly polished aluminum-magnesium alloy, although in recent years glass and ceramics have been used. The first disk coatings were a mixture of iron oxide and alumina particles coated as a slurry onto the substrate, then hardened by baking to a layer several micrometers thick. The alumina particles provided durability so that the recording head slider could land onto the disk without damaging the soft iron oxide. Requirements for higher recording densities include smoother surfaces (for closer head media separation and lower noise) and thinner magnetic films (for better resolution) with higher magnetization and lower coercivities (to maintain high signal-to-noise ratios). This led to the development of thin-film deposited magnetic layers, ranging from 25 nm to 200 nm thick. Generally, high wear resistance is achieved by coating the magnetic film with a thin (25 nm) layer of carbon. Most of today's disk files use thin-film disks.

A small proportion of disk files are floppy drives, so-called because the disk itself is a flexible material coated with a thin magnetic layer. Most of the same principles apply, except that these disks are designed to be removable and interchanged with any other drive. Mechanical alignment of the disk hub to the spindle, the warping of the softer disk material with temperature and humidity, and the differences from one head to another all make the densities for these devices much less than those for hard-disk drives. Capacities of a few megabytes per disk are typical, compared with up to hundreds of megabytes for a similar-sized rigid disk, but they represent an important option for transporting programs and data.

3.1.3 Tape Devices

Tape devices, like most magnetic recording devices, operate by using a magnetic head to read and write a magnetic medium that is moving with respect to the head. Unlike disk drives, the recording medium is a flexible magnetically coated tape wound in a removable cartridge. Further, the heads are generally run in contact with this flexible medium, which is open to the environment, requiring robust heads tolerant of debris and wear at the magnetic pole tips. However, because the cartridges are removable, the heads can be cleaned periodically and tapes can be copied if signs of degradation are detected. Since the cost of tape cartridges is very low and many cartridges are used with one read/write unit, tape devices offer a very inexpensive storage option, as low as $0.01/MB.

Two basic head types shown in Fig. 10 are widely used today. The simplest structure is a linear, stationary head array in which a string of head elements creates an equal number of tracks along a tape simply by moving the tape across it. In rotary-head recording, the tape is moved and the head is rotated at an angle on a cylinder to produce diagonal tracks. The latter allows a much higher head-to-tape velocity (about 1000 cm/s), higher track densities (about 1000 tracks/cm), and a very rapid search mode for desired data blocks. High velocity allows high data rates (up to 20 MB/s), and high track densities allow greater tape capacities (typically greater than 1 GB for a small cassette and up to 200 GB in larger ones).

The magnetic tape is critical to performance and has been improved substantially in the last decade. Early types consisted of iron oxide particles coated onto a polyethylene tape, giving coercivities of about 300 Oe. Cobalt surface treatments and, most recently, metal iron particle (MP) tapes offer improvements of up to 1450 Oe coercivity. Smoother surface finishes and greater durability have also contributed to steady recording improvements.

The chief disadvantage of tape devices is that access is serial. In the worst case, the entire length of the tape would have to be searched in order to get to a desired data block, requiring at least several seconds. High winding speed is required for short access times. Improved tape transport design is important for tracking accuracy at high tape velocities and for thinner tapes (from 25 μm in 1980 to as low as 7 μm in small rotary cassettes). Better mechanical design as well as improved wear characteristics due to carbon coatings on the back of the tape have facilitated continued tape volumetric density increases and faster data searches.

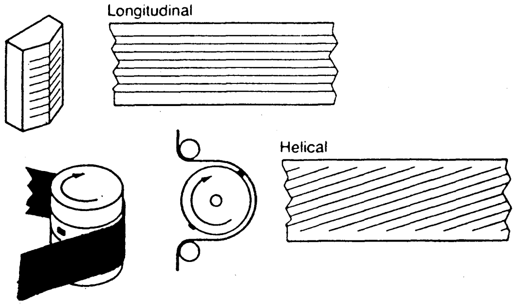

FIG. 10. Two modes of tape recording: fixed-head longitudinal and rotary-head helical (Mallinson, 1990, © 1990 IEEE).

3.1.4 Bubbles A magnetic bubble is a self-contained, cylindrical magnetic domain that is polarized in a direction opposite to the magnetization of the surrounding thin magnetic film, typically a rare-earth iron garnet deposited on a nonmagnetic substrate. A peculiar property of this film is that any domain that you put into this material spontaneously relaxes to a circular domain of a standard size that can be easily moved around, counted, or destroyed. Further, if you cut a bubble in half you get two full-size bubbles. Utilizing these attributes, sequences of bubbles (in which 1's and 0's are represented by the presence or absence of bubbles) are used to store information in the film, providing a sequential block storage device without the limitation of using mechanically moving parts.

Physically, the device, shown in Fig. 11, consists of a substrate coated with a thin garnet film (which contains the bubbles), a patterned conductor layer, and a patterned magnetic Permalloy layer, all separated by layers of silicon dioxide to provide electrical and physical isolation and optimum magnetic coupling (between the garnet and Permalloy layers). Mechanical and environmental protection is achieved with an overcoat of silicon dioxide, and electrical contact pads allow electrical control and input and output of the bit patterns. The final device, like a silicon chip, is a multipin package mounted onto a circuit board.

The bulk of the area of a magnetic bubble memory is dedicated to physically storing strings of bubbles (each about a micron in diameter), but an important fraction of the chip is dedicated to operating and controlling the device. Four basic functions are required in order to operate a magnetic bubble memory: propagation (access), generation (write), detection (read), and annihilation (read and erase). For some architectures, a method of

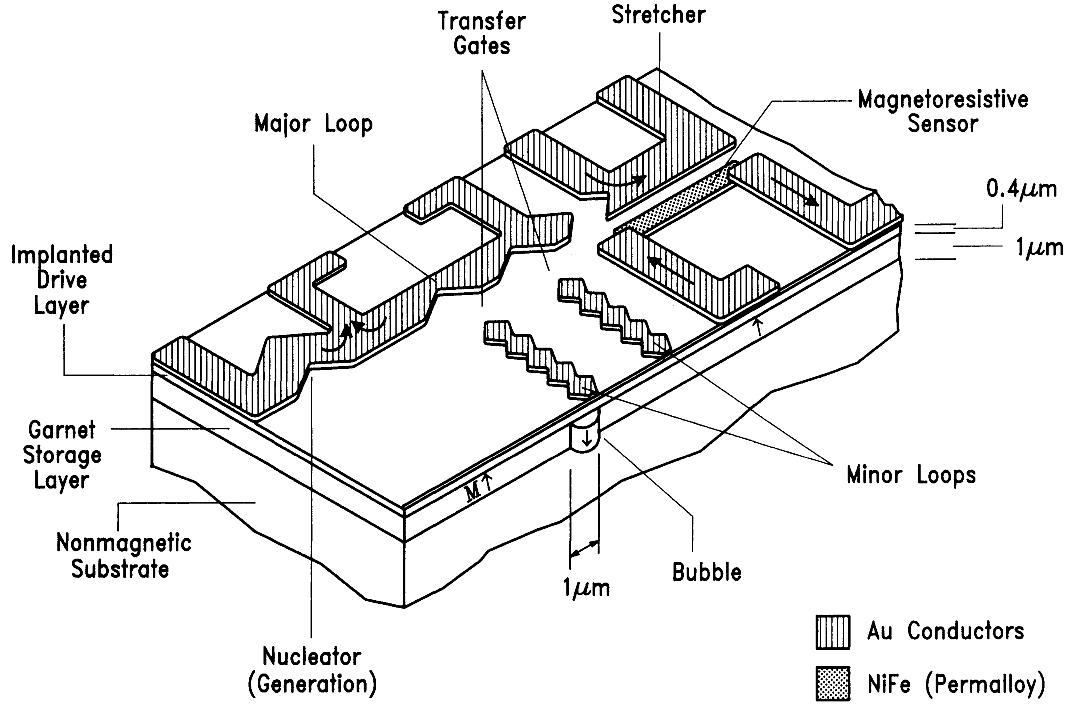

FIG. 11. Schematic diagram of one type of magnetic bubble memory device, showing major and minor loops for bubble storage and access, a nucleator for bubble generation, and a bubble stretcher and magnetoresistive sensor for bubble detection. (After Lin *et al.*, 1979.)

replicating bubble patterns is also necessary. These functions are accomplished through electrical control, using special patterning of the layers and materials properties (such as magnetoresistance, a change in resistance of a magnetic metal film depending on the magnetization).

Bubble memories are organized to provide many small sequences (minor loops) of bubbles accessed and controlled by one master sequence (major loop). Thus, when a desired piece of data is required; only the minor loop containing that data needs to be accessed. Bubble memories are used in calculators and other small-volume, nonvolatile memory applications. Difficulties in scaling bubbles to sizes below about 1 μm, materials challenges, bubble propagation speed limits, the rapid development of low-power and nonvolatile semiconductor devices, and the continued development of very small, high-performance magnetic storage devices all have limited the application area for magnetic bubble devices.

3.2 Magneto-optic and Optical Storage

Optical storage devices are in many ways similar to magnetic-disk files in that they have a specially designed head to write bits onto a rotating medium. Further, rotation provides circumferential tracks, and moving the head from track to track is accomplished with an electromechanical actuator. Common important attributes are linear (circumferential) bit density, data rate, access times to the first bit of information in the string of data, and the number of tracks per disk. Thus, the engineering requirements and progress described in Sec. 3.1.2 generally apply to this technology.

There are several key distinctions between magnetic and optical storage. First, the information is stored in a magneto-optic thin film. Second, transparent substrates are generally used, such as plastic and glass. Another distinction is that a focused laser beam is used to read and write information. This gives circular bits and thus comparable linear and track densities, requiring much greater tracking accuracy than disk drives. Also, very small, low-cost, reliable lasers and small, integrated focusing optics are essential. Today's optical storage heads generally require somewhat longer access times (due to the higher mass components), operate at a slightly lower data rate (constrained by laser power), and have larger head structures (influencing the number of disks possible in a stack) than magnetic-disk files. However, continued mass reduction and integration are expected.

An advantage of using a focused beam system, rather than operating in the near field as in magnetic recording, is that scaling to higher areal densities can be accomplished without bringing the optical head physically closer to the disk surface. Instead, the head-to-disk spacing is set by the optical focal length, generally many microns and much greater than the bit cell size. This substantially reduces concerns about head-disk interface wear and obviates the need for an air bearing to maintain accurately a small spacing between the head and disk. The larger spacing also allows the use of stamped grooves in the disk for simple tracking schemes and makes it possible to embed the information layer for optical storage into the disk below protective layers for robustness against handling damage and debris. These are among the properties that make magneto-optic and optical storage suited for very high-density removable-media storage devices.

Magneto-optic storage offers rewrite capability, infinite cyclability, and very fast reading and writing. However, erasing requires some time to achieve thermal equilibrium. This means that rewrite is done in two steps by first erasing and then writing the new information, requiring at least 2 disk revolutions. Also, a low-resolution magnet must be near the bit cell. Because this magnet takes space, it increases the height of the storage device relative to magnetic drives.

CD-ROM type storage allows low-cost media replication since the physical deformations that give rise to the optical contrast can be created by stamping with a master. However, like its (more costly) semiconductor analog described in Sec. 3.3.3, it is useful only for fixed data and program storage since it cannot be altered.

Phase-change storage has a limited cyclability and hence is used for write-once storage applications, analogous to (but much lower in cost than) PROM described in Sec. 3.3.4. Devices introduced in the mid-1980s offered 2 GB of storage per disk at 400 000 bits per square centimeter, 0.4-MB/s data rate, and 100-ms access time.

Continued extendability of optical and magneto-optical storage relies on reducing the bit size by focusing the laser beam to a tighter spot, generally by reducing the optical wavelength. In addition, direct overwrite techniques currently under research, which allow simultaneous erase and rewrite, will increase rewrite speed. Several generations of this type of improvement are expected before significant breakthroughs are required.

3.3 Charge Storage

Random-access memory is memory in which the access time is independent of the physical location of the data. It is a characteristic of all the device classes described below and offers significantly reduced access times compared to serial-access memory such as disk files, tape, or magnetic bubbles. Further, charge storage bits are accessed electrically, avoiding delays associated with the mechanical motion in disk files and tape and the physical motion in bubbles.

Fundamental to these devices is the use of a matrix of identical memory cells addressed by word and bit lines. Each cell contains one (or more) transistors arranged in a circuit, as discussed in Sec. 2.3. The CMOS offers the smallest area and power per bit and is the basis for most of today's devices.

Also fundamental to these devices is wafer fabrication in which many identical memory chips are created at the same time on a single silicon wafer. The silicon wafer (usually 20-cm diameter) is processed by superimposing several layers of conducting, insulating, and semiconducting materials. Each layer is patterned by means of masks and photolithographic processes to create the desired features. The wafer is then physically sliced up into several hundred identical chips. Each chip holds several million identical storage bit cells, connected together with thin-film wiring in an addressable matrix. Each bit cell is comprised of one to several transistors and possibly other integrated components.

Smaller bit cell sizes mean that more bits can be stored on a chip, generally lowering cost per bit. Thus, enormous effort is aimed at the physics and engineering of reducing transistor and cell areas. For instance, as cell sizes shrink, mask dimensions must be reduced to include linewidths well below a micron. Linewidths as low as 0.25 μm appear feasible with use of short-wavelength light and clever mask design. Mask dimensions substantially below 1 μm can be envisioned using x rays for patterning, and research is aimed at this approach. Also, the cell itself can be made smaller by orienting large-area capacitors vertically with respect to the rest of the films.

In addition to reducing bit cell area, desired directions of device improvement include lower power dissipation per bit, increasing speeds (faster bit rates, shorter access times, and shorter cycle times), and improving reliability. Nondestructive readout (NDRO) and nonvolatility are also important goals. No one device excels in all these parameters, giving rise to the variety of structures described below.

3.3.1 Static Random-Access Memory

Static RAM exhibits the property that the state of the cell is permanent unless rewritten, as long as the device power remains on. Hence it is static, but volatile. Static RAM is high speed but requires six transistors and resistors per cell, resulting in a large cell area and, thus, more cost per bit than many other RAM options.

A schematic of an SRAM cell is shown in Fig. 12. In this flip-flop circuit, when one cross-coupled MOSFET transistor is on, the other is off. Appropriate voltage inputs to the word and bit lines determine which transistor is turned on, making this bistable circuit into an addressable memory cell.

Bipolar SRAM is the oldest integrated memory circuit technology and dominated the 1960s and early 1970s, but was outpaced first by NMOS (*n*-channel MOS) and then by CMOS as the bipolar cell area and power consumption limited chip capacities and kept the cost high. The fastest RAM memories today are CMOS SRAM, with an 8-ns access time. However, the large cell size relative to DRAM (typically 4×) limits the use of CMOS SRAM to applications in which the faster cycle times [typically (5 to 10)×] are so important that the cost premium is justified.

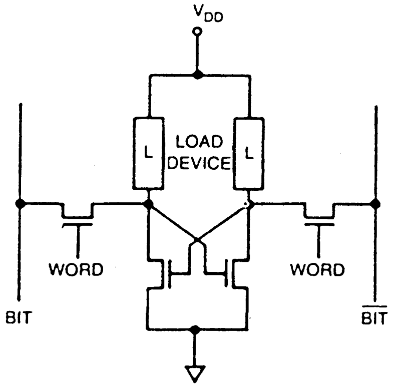

FIG. 12. Static RAM cell, based on a flip-flop circuit. The load device may be resistors or transistors and is usually transistors to minimize cell area. (Weste and Eshraghian, 1985, © 1985 by AT&T Bell Laboratories, Inc. Reprinted by permission of Addison-Wesley Publishing Company.)

3.3.2 Dynamic Random-Access Memory

Dynamic RAM exhibits the property that the memory cell must be refreshed, or rewritten, every few milliseconds to restore the charge that has leaked from the cell's capacitor. While this requires added complexity in the chip's control circuitry and adds to the access and cycle times, it also makes possible the use of very compact bit cells. This in turn gives larger capacity chips at a lower cost per bit.

A DRAM cell schematic is shown in Fig. 13. Information is stored as a charge across a single capacitor. The storage capacitance must be kept low enough to minimize the area occupied on the wafer. However, the smaller the capacitance, the more quickly the cell discharges, requiring refreshing. Thus, de-

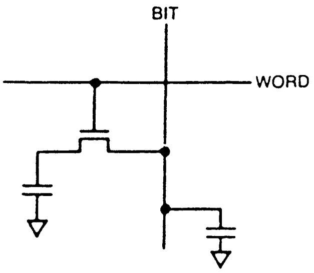

FIG. 13. Dynamic RAM cell using a single transistor. The separate capacitor shown represents the parasitic capacitances that must be included for proper device design. (Weste and Eshraghian, 1985, © 1985 by AT&T Bell Laboratories, Inc. Reprinted by permission of Addison-Wesley Publishing Company.)

sign efforts are focused on maximizing the storage capacitance within the area limits of the cell. Current designs require rewriting every 64 ms, giving a 98 to 99% availability of the cell for reading and writing.

The very small cell size, low cost, and good performance make DRAM the dominant random-access memory choice today. For instance, typical 4-Mb chips have access speeds of about 60 ns, and 16-Mb chips have access speeds of about 50 ns, at prices ranging from $15 to $50 per MB. Technology demonstrations give a high level of confidence that the historical improvement of on average 4× density increase every three years since 1965 (with associated cost reduction and speed improvement) will continue for several generations, although there is some slowdown in the rate of improvement.

3.3.3 Read-Only Memory Read-only memory differs fundamentally from other semiconductor memories in that its stored information is fixed, unalterable, and nonvolatile. A simple form of ROM is shown in Fig. 14. The state of each (*i,j*) position is fixed at the time of fabrication by the presence or absence of a transistor at that position. Practically, this means that ROM cells may be fabricated with only one transistor per bit of storage, although any updates to the content require another unique chip.

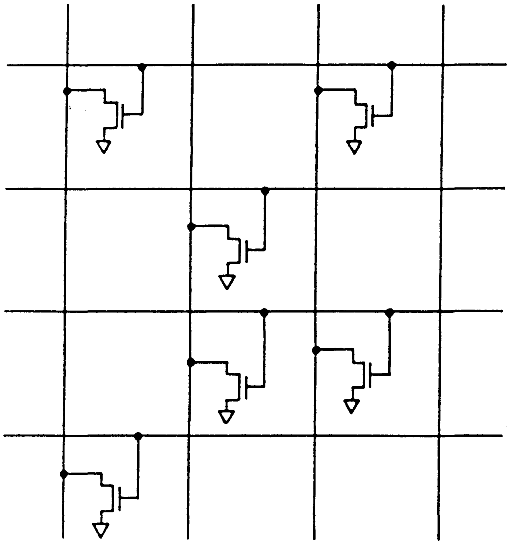

FIG. 14. ROM layout showing 1's and 0's represented by the presence or absence of an active transistor at matrix positions, requiring unique patterns for each distinct memory device. (Weste and Eshraghian, 1985, © 1985 by AT&T Bell Laboratories, Inc. Reprinted by permission of Addison-Wesley Publishing Company.)

ROM devices are fabricated like most semiconductor memories with standard materials and processing technology, except that unique personalization is required in one or more masks during fabrication to create the desired unique bit pattern. Techniques used to program the masks include contact programming, presence or absence of a transistor, or a special doping to turn the desired transistors permanently on or off.

ROM is used for fixed-program storage and offers a convenient way to "boot up" or initialize computer systems. It is very rugged, since the program is a part of the device itself. Its primary disadvantage is that any change, however small, requires an entirely new ROM to be designed and produced.

3.3.4 Programmable Read-Only Memory Programmable read-only memory offers fixed, nonvolatile storage that is set by the user as a one-time operation. This eliminates the need for personalization as the device is fabricated but requires special programming processes at the user level. This also requires that each (*i,j*) memory will be potentially either a 1 or a 0, making the area per unit of storage somewhat larger.

Programming the PROM is called "burning" since physically the process is to open fuse links at locations where a binary 1 is desired, leaving all 0's intact. Burning the fuses is accomplished by applying a high-energy pulse to the desired addresses of the 1's. This high-energy pulse alters the properties of a transistor memory cell so as to create the 1 state. PROMs are somewhat more expensive than ROMs but give useful added flexibility for the user.

Erasable PROM (EPROM) overcomes the primary disadvantage associated with ROM and PROM; that is, it can be erased and rewritten up to thousands of times by the user. To program an EPROM, the user selects the desired address, places the desired bit pattern of 1's and 0's on the data line, and pulses the PROGRAM pin. Charge is physically stored on the floating gate of a MOSFET transistor, described in more detail in Sec. 3.3.5. To erase an EPROM, the trapped charges are removed by exposing the chip to ultraviolet light through a transparent quartz window, exciting holes and electrons. This allows the trapped charge to escape over the potential barrier, erasing all stored data on the chip at once. While the erasability is a large advantage over ROM and PROM, an important limitation is that the chip itself must be physically removed from the system for erasing and reprogramming.

3.3.5 Electrically Erasable PROM An important class of EPROMs are those that are electrically erasable. The low cost and ease of erasability of this nonvolatile RAM offer significant advantages for program and data storage.

The floating gate memory cell, the basis for EEPROMs and EPROMs, contains an FET whose gate is electrically floating. Charge on the floating gate modulates the charge in the channel below the gate. The state of the cell is sensed by measuring the channel conductivity, typically in about 80 ns, and inferring from this the state of the charge on the floating gate.

A typical structure is shown in Fig. 15. To write, charges are injected from the silicon through the insulator onto the floating gate. This is generally done by putting a large bias voltage, V_D, between the source and drain, then switching the transistor on by use of the control gate (labeled V_G in Fig. 15). The large bias causes a flow of very energetic electrons between the source and drain in the channel underneath the floating gate. About 1 of every 1000 of these "hot" electrons is injected onto the floating gate and is trapped there, taking about 10 to 100 μs to establish enough charge for device use. This net charge on the floating gate, with proper design, can be retained for many years.

The charge on the floating gate, once established, can be removed via a tunneling mechanism, thus erasing the data. A large electric field is applied between the source and control gate, causing the electrons trapped on the floating gate to "tunnel" back into the silicon. The thinner the oxide, the less is the voltage required for this process, since voltage is the product of electric field and thickness. Typical oxide thicknesses are 10 nm. Enriching the oxide with silicon or using nitride layers also helps reduce the required erase field. Erase times are typically 5 to 20 ms, and thus it is a very attractive feature of this technology that an entire block or an entire chip may be erased simultaneously. EEPROMs with this block-erase characteristic are called Flash EEPROMs.

The large fields and energetic electrons required for the write and erase operations in Flash cause, over repeated cycles, a change in the insulator properties. Charges and traps build up in the insulators and at surfaces, resulting in inefficient injection and tunneling. Details of the cell design, materials, and operation mode all affect the endurance, but typical devices last only up to 10^6 cycles. EEPROMs are configured in RAMs in the same manner as described in Sec. 3.3.

The cell size of Flash EEPROMs can be very small, since there is no need for a separate capacitor. With a 0.7-μm minimum feature size, 16-Mb chips costing about \$20/MB have been shipped. The random-access capability allows access times of 15 μs, 1000 times faster than magnetic direct-access storage device and more than 100 times slower than DRAM, giving Flash EEPROMs an intermediate performance. The electrical erasability makes possible field updates of chips used in pro-

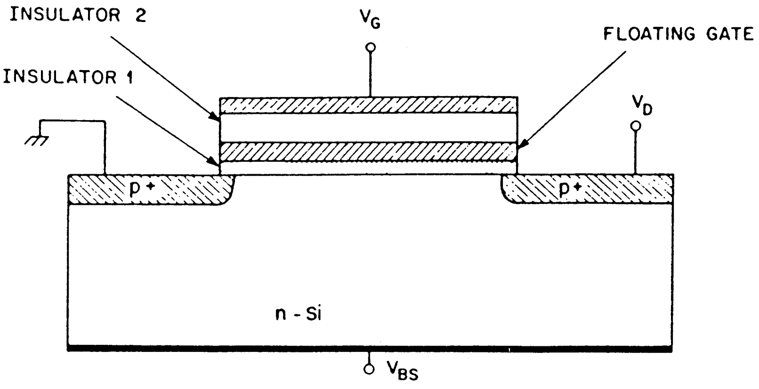

FIG. 15. EEPROM structure illustrating the floating gate that controls injection and removal of trapped charge for nonvolatile storage. (Sze, 1981, copyright © 1981 by John Wiley and Sons, Inc. Reprinted by permission of John Wiley and Sons, Inc.)

gram storage applications. These attributes make Flash EEPROMs very useful for small-form-factor, very robust, modest storage-volume applications such as laptop and palmtop portable computers.

3.4 Ferroelectric Storage

Ferroelectric RAM is fabricated with a material such as lead zirconate titanate (PZT) as the dielectric in the capacitor of an otherwise conventional DRAM-integrated memory cell, as shown in Fig. 16. Since the polarization state of a ferroelectric is permanent until reset, even in the absence of power, this provides nonvolatile random-access storage in a compact (single transistor) cell structure.

Physically, the device is created by standard semiconductor processes to create CMOS transistors. The ferroelectric thin-film capacitors are built on top of each transistor. The polarization state of this capacitor is used as nonvolatile storage in a manner analogous to the charge state of conventional capacitors in DRAM.

Commercial devices became available around 1990, although the concepts of ferroelectric storage have been known since the 1950s. Significant engineering and materials challenges were overcome to make practical devices, and densities are still very low compared to DRAM. For instance, only recently was it possible to fabricate ferroelectric thin films of sufficient quality for use in these devices, allowing lower-voltage operation and the high level of integration necessary for acceptable costs. Film material uniformity, process temperatures, thickness, cyclability, and electrode material compatibility are some of the specific challenges. The cyclability of about 10^6 still limits performance and con-

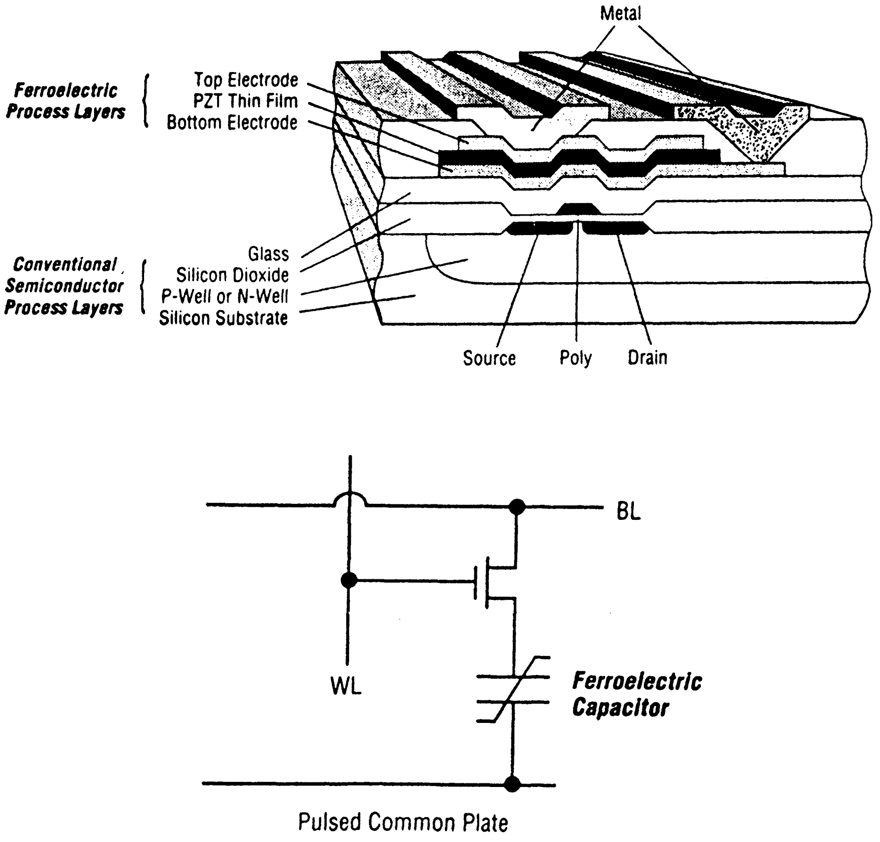
Metal
Ferroelectric Process Layers
Top Electrode
PZT Thin Film
Bottom Electrode
Conventional Semiconductor Process Layers
Glass
Silicon Dioxide
P-Well or N-Well
Silicon Substrate
Source
Poly
Drain
BL
WL
Ferroelectric Capacitor
Pulsed Common Plate

FIG. 16. Ferroelectric memory cell showing (a) the ferroelectric capacitor on conventional DRAM to provide nonvolatile storage and (b) the circuit representation of this structure. (Illustration © Ramtron International Corporation, 1993.)

strains the applications, but the demand for truly nonvolatile storage continues to motivate further development.

4. APPLICATIONS

In today's computers, many memory and storage technologies are utilized, organized into a hierarchy as shown in Fig. 17. The goal of using a hierarchy is to optimize system-level cost performance. This is largely because the fastest technologies are typically the costliest, with slower alternatives available at orders of magnitude less cost per bit. Judicious use of high-performance, high-cost memory fed by increasing amounts of lower-cost, lower-performance technologies can give optimum cost performance. In addition, nonvolatility and low power per bit become increasingly important for large volumes of data, and these have usually been characteristics of lower-cost storage devices.

In particular, at the top level in the hierarchy are the processor's internal registers. These are very fast since they reside on the processor chip itself, but chip area severely limits their number. At the second level is a memory running in synchrony with the processor, called cache. The speed requirement necessitates SRAM, and, thus, the size is constrained by the high cost of this memory. Information most likely to be used by the processor is kept in cache, while the rest is stored in lower levels of the hierarchy. The third level is main memory, generally comprised of DRAM, which is about ten times slower than SRAM but five to ten times less costly. Programs and data are kept in this memory. Through the third level in the hierarchy, memory, that is, random-access technologies, are used.

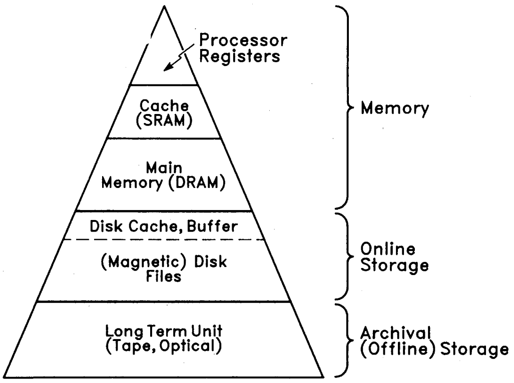

FIG. 17. Computer memory hierarchy.

The fourth level is mass storage, generally disk files. These are serial block-storage devices, connected to the RAM by special channels that optimize availability of the data in the presence of mechanical access delays. Typically, tens of milliseconds are required for access, but this provides "on-line" nonvolatile storage at a cost of $1 to $5 per MB. Even larger volumes of data, system backup information, and archival files are stored "off-line," on tapes (at costs as low as $1/MB) or on optical storage disks that have to be physically mounted onto the drive when information is requested. This can amount to petabytes in some applications and is often physically stored elsewhere as insurance against disaster.

The very large performance gap between semiconductor and magnetic memory requires a sophisticated link between the two to hold blocks of data temporarily and schedule requests for information flow between main memory and magnetic storage. Special memories are used to facilitate this process, known as buffers and cache. Because it is important to retain this information if power is lost, this is an important application for battery-backed DRAM and the emerging nonvolatile RAM technologies discussed in Sec. 3.3.

New solutions are also in sight for the performance gap between on-line and off-line storage. An area of growth is "near-line" storage, that is, storage of large volumes of seldom used data in "juke box" type tape or optical files for access in a few seconds. This is likely to be of increasing significance as high-volume data storage applications of video, high-definition television, and computer-aided design continue to grow.

Steady cost and size reductions of memory technologies have had a profound influence on the numbers and types of memories available. No longer is computer memory confined to specially protected facilities servicing large mainframe computing devices, so-called "glass houses." In fact, the fastest growth areas are in small computers and the mobile computer market, necessitating compact, robust, low-power, nonvolatile storage. Flash EEPROM and small, highly shock-tolerant disk files are particularly attractive here. While the cost per bit of semiconductor storage has historically been orders of magnitude greater than that of magnetic storage, that gap has steadily narrowed, and for the small volumes of storage (a few megabytes) required for many applications, the cost of Flash EEPROM is acceptable.

Trade-offs in cost and performance and new uses of computing have played a key role in the diversification of today's memory technologies. With future technology improvements and as the volumes of memory continue to grow, a rich mix of alternatives, complementary and competitive, will continue to evolve.

GLOSSARY

Areal Density: The number of data bits that can be stored per unit area on a thin-film storage medium.

Coercive Field, Coercivity: The reverse applied magnetic or electric field at which the material's magnetization or polarization, respectively, can be reduced to zero.

Crystal Phase: An organization of atoms into a specific geometric structure. Often the same atoms organize in a different structure (phase) at different temperatures or under different process conditions.

Cyclability: The number of times a memory can be erased and rewritten before it can no longer be used.

Dielectric: An electrically nonconducting material.

Domains: Regions within a material in which electric polarization or magnetic spins align in the same direction.

Ferrite Cores: Rings of ceramic magnetic material.

Ferroelectric: A material exhibiting an electric polarization even in the absence of an applied electric field.

Ferromagnetic: A material exhibiting a magnetic moment even in the absence of an applied magnetic field.

Holes: Vacant states in an otherwise filled electron band in a crystal, which can be represented as a positive charge.

Hysteresis: A property of a material that exhibits different behavior for increasing and decreasing applied fields.

Magnetic Bubble: A cylindrical domain in a thin magnetic film that is magnetized oppositely and vertically with respect to the overall film.

Magneto-optic: A coupling between magnetic and optical properties of a material.

Magnetoresistance: A change in electrical resistance of a material depending on the applied magnetic field.

Masks: Patterned glass or metal plates used to create desired features on a thin-film fabricated structure. Light travels through transparent sections and is absorbed or reflected by opaque sections.

Media: The material in which data bits are physically stored.

Near Field: A region of a wave very near the source compared to a wavelength.

Photolithography: The process by which images are transferred onto a structure. A photosensitive layer called photoresist is coated onto the surface. Masks are used to expose selected regions of the resist and then developed so that the desired pattern of resist remains. Layers are deposited or removed using this photoresist template.

Remanence: The residual magnetization of a material after the applied field is removed.

Substrate: A base material onto which thin films are deposited and patterned appropriately to make devices.

Works Cited

Lin, Y. S., Almasi, G. S., Keefe, G. E., Pugh, E. W. (1979), "Self-Aligned Contiguous-Disk Chip Using 1 μm Bubbles and Charged Wall Functions," *IEEE Trans. Magn.* **MAG 15**, 1642–1647.

Mallinson, J. C. (1990), "Achievements in Rotary Head Magnetic Recording," *Proc. IEEE* **78**, 1004–1016.

Mee, C. D., Daniel, E. D. (1987), *Magnetic Recording, Volume I: Technology*, New York: McGraw-Hill, p. 5.

Mee, C. D., Daniel, E. D. (1988), *Magnetic Recording, Volume II: Computer Data Storage*, New York: McGraw-Hill, p. 6.

Sze, S. M. (1981), *Physics of Semiconductor Devices*, 2nd ed., New York: Wiley, p. 496.

Weste, N., Eshraghian, K. (1985), *Principles of CMOS VLSI Design*, Reading, MA: Addison-Wesley, pp. 6, 349, 353, 355.

Further Reading

Basche, C. J., Lyle, R. J., Palmer, J. H., Pugh, E. W. (1986), *IBM's Early Computers*, Cambridge, MA: The MIT Press.

Bobeck, A. H., Scovil, H. E. D. (1971), *Sci. Am.*, **224** (6), p. 78–90.

Luecke, G., Mize, J. P., Carr, W. N. (1973), *Semiconductor Memory Design and Application*, London: McGraw-Hill.

Mallinson, J. C. (1987), *The Foundations of Magnetic Recording*, New York: Academic.

Mee, C. D., Daniel, E. D. (1987), *Magnetic Recording*, New York: McGraw-Hill, Vols. I, II, and III.

Sze, S. M. (1981), *Physics of Semiconductor Devices*, 2nd ed., New York: Wiley.

Watson, J. K. (1980), *Applications of Magnetism*, New York: Wiley.

Weste, N., Eshraghian, K. (1985), *Principles of CMOS VLSI Design*, Reading, MA: Addison-Wesley.

MESOSCOPIC SYSTEMS

YOSHIMASA MURAYAMA, *Advanced Research Laboratory, Hitachi, Ltd., Hatoyama, Saitama, Japan*

Introduction 81
1. **Background** 82
1.1 Quantum Effects 82
1.2 History 82
2. **Quantum vs Classical Physics** 83
2.1 Wave-Particle Duality 83
2.2 Coherence 85
2.3 Quantization 86
3. **Mesoscopic Systems** 86
3.1 3D Quantum Systems 86
3.2 2D Quantum Systems 87
3.3 1D Quantum Systems 88
3.4 0D Quantum Systems 89
4. **Physical Properties** 89
4.1 Optical Properties 89
4.1.1 Single/Multiple-Quantum Wells 89
4.1.2 Superlattice 90
4.1.3 Selection Rules 91
4.2 Transport Properties 92
4.2.1 Transport Parallel/Perpendicular to QW 92
4.2.1.1 Parallel Conduction–Metal-Oxide Semiconductor (MOS) 92
4.2.1.2 Perpendicular Conduction—Resonant Tunneling Device 92
4.2.2 Negative Magnetoresistance 94
4.2.3 Aharonov–Bohm Effect 95
4.2.4 Landauer–Büttiker Formula of Conductance 96
4.2.5 Universal Conductance Fluctuation 97
4.2.6 Quantized Conductance 97
4.2.7 Interference between Edge Currents 98
4.2.8 Integral Quantum Hall Effect 99
4.2.9 Fractional Quantum Hall Effect 100
4.2.10 Ballistic Transport 101
4.2.11 Coulomb Blockade 101
4.3 Magnetic Properties 102
4.3.1 Fine Particles 102
4.3.2 Magnetic Thin Films 103
4.3.3 Superlattice of Magnetic Thin Films 104
4.4 Properties of Macroscopic Quantum States 104
4.4.1 Superconducting Wire 105
4.4.2 Andreev Reflection 106
5. **Future Prospects** 107
5.1 Physics 107
5.2 Applications 108
Acknowledgments 109
Glossary 109
Works Cited 110
Further Reading 111

INTRODUCTION

Throughout the history of electronics, we have pursued smaller cellular devices in, e.g., integrated circuits, which means better performance per cost. Most people on the engineering side vaguely conjectured that these tiny devices should be ruled by quantum physics. However, they hardly recognized that these devices form a bridge between a bulk system governed by classical physics and a system governed by quantum physics, and consequently, they open up a new front of physics. This new physical world, neither classical nor quantum, is named *mesoscopic* physics, after the Greek word μεσος meaning "middle." The mesoscopic world has many partially coherent phenomena, which seem to be successfully described by introducing *decoherence* mechanisms into quantum mechanics.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. BACKGROUND

1.1 Quantum Effects

During the revolutionary two decades of the discovery of quantum mechanics, de Broglie proposed the ingenious idea that massive particles are also waves (see QUANTUM MECHANICS). This idea was soon experimentally confirmed using electron beams that can really interfere. As far as an electron in a vacuum is concerned, such a free particle behaving like a free wave can be simulated as an actual wave like a light wave, i.e., $e^{-i\omega t+ikx}$, if its frequency ω and wave vector $\mathbf{k}$ are identified with $p^2/2\hbar m$ and $p/\hbar$ ($\hbar$ is the Planck constant/2π; p, momentum). This is true when the *coherence length* of the electron beam is long enough. Otherwise, it is common to describe the beam as a *wave packet* with a specific coherence length, in other words, a span in real space.

On the other hand, in crystals, the energy-momentum relation is not as simple as in a vacuum, since a series of periodic potentials strongly modulates the plane-wave-like nature. In particular, electron wave states specified by energies around $\hbar^2\pi^2/2ma^2$ (a is the lattice constant) interfere destructively and, consequently, become extinguishing waves. This phenomenon is essentially the same as Bragg reflection, and the energies for which waves extinguish are known to be within a certain energy gap. In ordinary solids the separation between successive bands does not tend to zero at the Brillouin-zone (BZ) boundary, and the minimum separation is called the energy gap (see ELECTRON STRUCTURE OF SOLIDS).

For this kind of complicated dispersion between energy and momentum, a renormalized mass can be introduced, which is ordinarily called the *effective mass m*. Electronics engineers, especially semiconductor device engineers, frequently view electrons as billiard balls with this effective mass. This is because the concept of effective mass has been very successfully applied to analyzing the properties of solid states in those devices. Here, the momentum $\mathbf{p} = m\mathbf{v}$ should be replaced by $\hbar\mathbf{k}$, since $\mathbf{k}$ describes a state better than $\mathbf{v}$.

If we use the effective mass m instead of the bare mass m_0 of the elementary particle, most properties look as if electrons are classical particles with their modified mass. In fact, the Molecular Dynamics simulation of devices was based on this transinterpretation.

However, it has been well recognized that elementary quantum processes, such as photon absorption and electron-phonon scattering, should be treated by calculating the scattering cross section σ or the scattering time τ, which means that each process occurs not *deterministically* but *probabilistically* following quantum mechanical rules. Engineers used to describe electronic properties in terms of classical parameters such as effective mass and σ or τ, to avoid recourse to quantum mechanics. Thus, in engineering, quantum effects have been laid aside so far, at least apparently. However, in the world of electronics, there began to appear quantum devices such as *superlattices* (SLs), *quantum wells* (QWs), and resonant tunneling devices (RTD), since the beginning of the 1970s. These are recognized as explicitly showing quantum effects.

In optics, to our surprise, the most quantumlike device, the laser, can be described mostly in terms of a classical monochromatic light source. A pure quantum state behaves like a classical wave. This is because a laser has extremely long coherence length and is thus similar to a classical plane wave. In this sense, quantum effects in optics seem to appear only in nonlinear effects or fluctuations in a vacuum. A similar situation occurs for macroscopic quantum effects such as in superfluids and superconductors. Inside superconductors, the coherence length covers the whole sample so that their states are conventionally described by a macrowave. The easiest example to illustrate quantum effects is size quantization; i.e., intrinsic Bloch states must be requantized according to additionally imposed boundary conditions, because the size of concern is small enough to allow *ad hoc* quantization in ordinary mesoscopic systems. Thus, there appear new quantum effects in addition to the above-mentioned implicit quantum effects. These new quantum effects are concerned in the present mesoscopic problems.

1.2 History

When discussing quantum effects in electronic devices, we might use the term "quantum electronics." However, for historical reasons, quantum electronics has usually been used to mean electronics using lasers and related optical devices. Thus, quantum electronics can no longer be used for purely electronic quantum devices.

To avoid confusion, the present author proposed the name "quantum microelectronics" (Murayama, 1984). However, a better name, "mesoscopic physics," became popular after it was proposed by van Kampen (1976). This terminology describes a broader scope of physics.

Currently, mesoscopic physics has attracted much attention, since such small electronic devices have been widely fabricated, opening a new physical world, not only in electronics and optics but in other fields where the importance of novel physical phenomena has been more and more recognized.

2. QUANTUM VS CLASSICAL PHYSICS

2.1 Wave-Particle Duality

Let us first consider what scattering process means in quantum mechanics. Consider a wide-spread plane wave of light before the physical process of scattering takes place. Then, imagine an atom of dimensions on the order of several angstrom that can absorb the light wave with a wavelength on the order of 5000 Å. A quantum absorption process occurs, as is easily imagined, only on the portion of the amplitude of the light wave that has the same spatial extension as the atom. At the moment of absorption, the light wave must behave like a particle, say, a photon, since the absorption process occurs on an all-or-nothing basis. This nature is a manifestation of the *particle-wave duality* of a quantum mechanical entity. The process occurs *probabilistically,* depending on the amplitude of the light wave distributed over the atomic scale. The cross section is proportional to the square of the product of these terms: the extension of the atom and the amplitude of the original light wave. When this probability is sufficiently small, most photons pass through without being absorbed.

This fact should be interpreted as follows. The entity to be absorbed is *not* a part of the light wave, but an entire photon. In other words, absorption occurs only when the photon behaves as if it were atom-sized. Here, the notion of a wave is nothing but a wave representing a distribution of the probability amplitude of the existence of photons. The probability distribution is wavelike, as it can be specified by its amplitude and phase.

From an early stage in the discovery of quantum mechanics, it has been recognized that the rules governing quantum mechanics, like the Schrödinger equation, are linear and can be *superposed.* Superposition means that two quantum mechanical states (or waves), defined by ψ_1 and ψ_2, must have a physical, not mathematical, meaning superposed as $\psi_1 + \psi_2$. Actually, there should exist a superposer, which can exert a physical action on both states so as to superpose them. For example, in electron interferometry with an Aharonov–Bohm geometry, two particle waves of an electron hit a fluorescence screen, where they are superposed by interacting with atoms inside the screen, which works simultaneously as a detector.

According to the experiment cited in Fig. 1, the interference pattern is the pileup of the records of individual fluorescence spots. Each spot shows an elementary interaction between a superposed electron and a fluorescent atom. Thus, as long as the quantum states of electrons emitted from a source are well-defined, i.e., they have a similar energy within a very narrow energy spectrum and the source is pointlike, then the quantum behavior of all the electrons is also well-defined, although it is never defined where on the screen each electron jumps to, but only the probability of spot distribution is determined by the squared wave function. This eventually causes the well-known sinusoidal interference pattern. The same sort of experiment was also performed using photons (Tsuchiya *et al.*, 1985).

Waves interfere only when both the waves are *coherent*. If both are *incoherent,* their superposition can never give rise to interference, even if they reach a small region simultaneously, where some superposition mechanism should work.

Next, we will see that it is not *size* that differentiates a macroscopic system from a microscopic one. According to the correspondence principle by Niels Bohr, any microscopic state has a corresponding entity in classical physics, especially states with large quantum numbers (*correspondence principle*). Take a hydrogen atom, for instance; the states with sufficiently large principal quantum

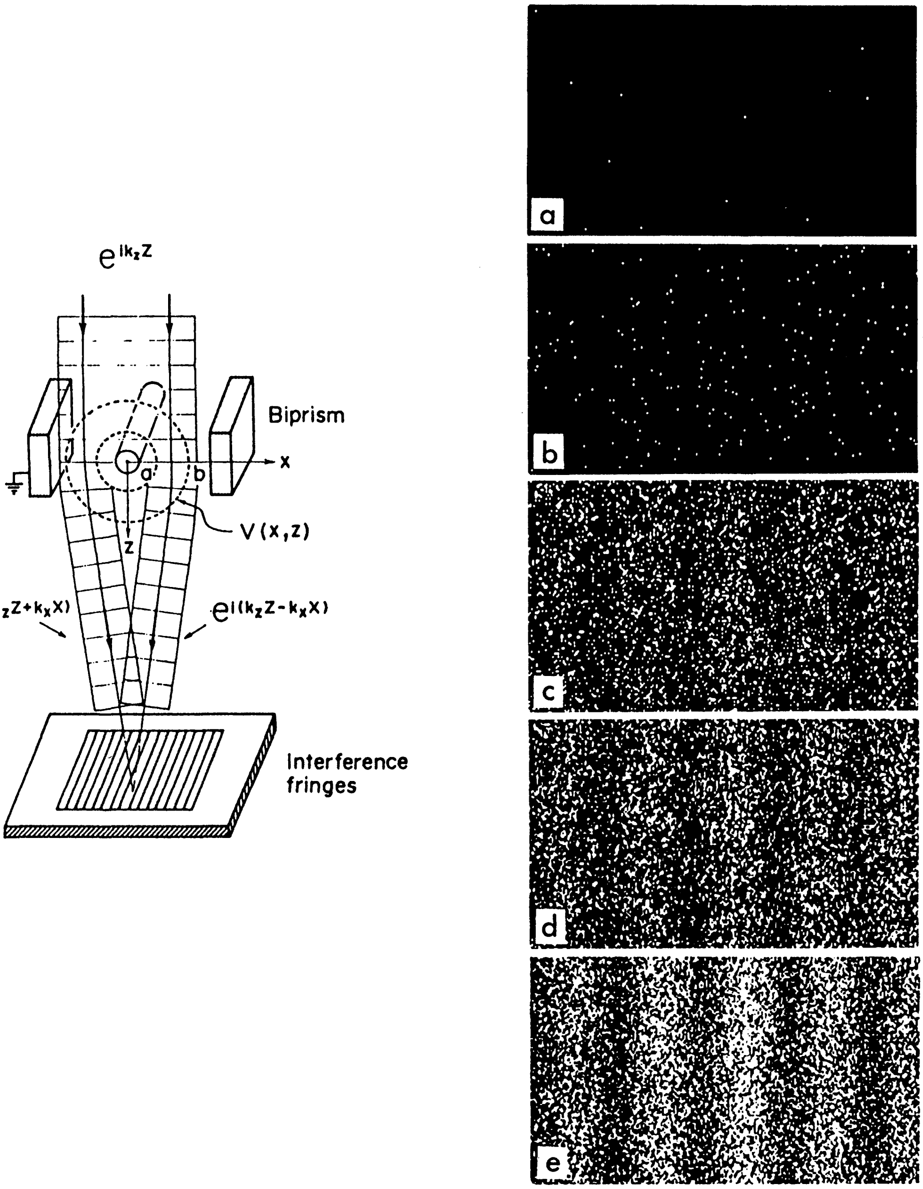

FIG. 1. Superposition of divided electron plane waves on a fluorescence plate, where an interference fringe is formed. Building of the interference pattern is shown progressively for number of electrons equal to (a) 10, (b) 100, (c) 3000, (d) 20 000, and (e) 70 000 (after Tonomura *et al.*, 1989).

numbers n are close to each other and look as if they were continuously distributed. For example, it is known that so-called Rydberg states in hydrogen atoms have orbitals larger than $\sim 1\ \mu\text{m}$ for $n \gtrsim 100$ (orbital radius $\propto n^2$).

In this context, it can be said that the boundary between macroscopic and microscopic world is not concerned with *size*. It is *coherence*. It is well established by now that the classical system is the limiting case of a quantum system when the number of degrees of freedom and/or the quantum number $\to \infty$. The former is intrinsically included in the latter but appears to be a different world because of, e.g., innumerable degrees of freedom. Then, what brings a microscopic system into a macroscopic one? It is not system size, as we have shown. It is what makes a *coherent* state *incoherent*. Any *decohering* mechanisms are thus the subjects of interest.

2.2 Coherence

Figure 2 shows the concept of a localized state in solids, after Bergmann (1983). This kind of localization is known as Anderson localization (see ELECTRON STATES, LOCALIZED), named after the first paper on the topic by Anderson (1958). In solids with disorders, an electron is scattered *elastically* many times by these disorders and eventually returns to its original position. When a wave is superposed with itself after an itinerary, it must have a peak intensity at a specific site on the path where the phase relation is *constructive* between itself and the adjoint after itinerary. This resembles an interference experiment in an electron microscope (Fig. 1). This state is well defined in the sense that it is stable and long lasting, unless it obtains enough energy to be activated from a localized to a delocalized state. The activation is only possible through *inelastic scattering* by optic phonons, photons, plasmons and so on.

Historically, it had been *a priori* presumed that, even in very small devices, we cannot directly observe coherent phenomena, since there are lots of elastic scatterers in solids and the product $k_F l_{el}$ rarely exceeds unity (l_{el} is the mean free path under elastic scatterings; k_F, the Fermi momentum/$\hbar$). However, this presumption was proved not to be the case. As is suggested in Fig. 2, elastic scatterings do not reduce coherence. Elastic scattering causes only a phase change.

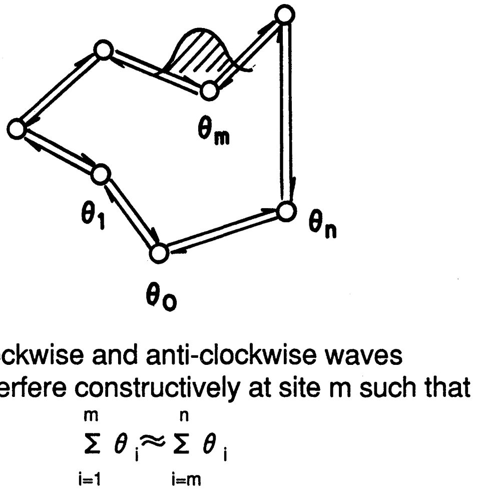
θ_m
θ_1
θ_n
θ_0
clockwise and anti-clockwise waves
interfere constructively at site m such that
$\sum_{i=1}^{m} \theta_i \approx \sum_{i=m}^{n} \theta_i$

FIG. 2. Conceptual description of the Anderson localized state. Each node represents a defect, causing an elastic scattering (after Bergmann, 1983).

On the other hand, *inelastic* scatterings change the energy from the initial value, which reduces the coherence. This is *decoherence*. After physicists realized this fact, many phenomena investigated so far in small electronic devices were recognized to show at least a small fraction of coherence, especially at lower temperatures with few inelastic scattering events. Thus, the above-mentioned criterion should be replaced by $k_F l_{in} \gtrsim 1$, although $k_F l_{el} \ll 1$, where l_{in} means the mean free path under inelastic scatterings, which ordinarily equals the phase-breaking mean free path l_ϕ. In the time domain, the scattering time for inelastic scatterings, τ_{in} or τ_ϕ, is ordinarily used, the latter of which is called the phase-breaking time.

Understanding this is very important from the viewpoint of quantum measurement theory. It is *a priori* assumed that, in order to acquire information from a detection, the detection must be an inelastic process. This is also wrong. If we can somehow determine a phase shift using a detector, then this is a kind of measurement. To detect a phase change, we need lots of similar events, not a single event, because of the uncertainty between phase and the number of events of concern. In fact, we cannot deduce anything from a single event on the fluorescence screen shown in Fig. 1, because a small number of spots can never make an interference pattern, and consequently, no information about phase is obtained from such a sparsely spotted pattern. Elastic scattering is actually a limiting case of inelastic scattering with an infinitesimal energy transfer.

A well-known difficulty in quantum measurement theory is in how a microscopic system evolves into a macroscopic one at the moment of detection. Thus, unless an inelastic process exists, the detection itself is elastic and the change caused by the process is merely a phase shift, and hence, the whole system remains in a quantum mechanical microscopic state, as was discussed by Wigner (1963). Even in this case, we can detect the change in phase by means of an ensemble average of similar events (Machida and Namiki, 1980; Murayama, 1990).

2.3 Quantization

In a periodic lattice, an extended wave is quantized so that the wave vector **k** is a good quantum number according to the Bloch theorem. Here, the minimum unit of periodicity is the lattice constant **a**. In order to form an extended wave over the whole crystal, every atomic orbital should be coherently superposed. The disturbance that hinders this superposition is, as was discussed before, inelastic scatterings. Inelastic and quasielastic scatterings such as optic- and acoustic-phonon interaction reduce the periodicity and, hence, invalidate the quantization condition claimed by the Bloch theorem. Ordinarily, a well-defined **k** value is thought to smear out according to this disturbance on the lattice points. Let us express the displacement of lattice points by $x_j = x_{j0} + \delta x_j$ (x_{j0}: regular lattice points at the jth site); then

$$\langle e^{ikx_j}\rangle_{av} \approx \exp(ikx_{j0})\exp(-\tfrac{1}{2}k^2\langle\delta x_j\rangle^2),$$

the second exponential of which is known as the Debye–Waller factor. In order that a state is well-quantized within a certain space (e.g., a box), it is necessary that $\frac{1}{2}k^2\langle\delta x_j\rangle^2 \ll 1$, if the decoherence mechanism is assumed to be electron-phonon interaction.

Again consider a box with waves inside. To obtain well-quantized states in the box, the waves on the two sides must be coherent; in other words, the information on the left extreme must be conveyed accurately to the right. This is a necessary condition for *size quantization.* The state in a solid is composed of Bloch waves with the lattice periodicity, the envelope of which is further modified according to the size quantization specified by the boundary condition of the box of concern. The latter modification according to the boundary condition of the box is called an *envelope function.* Ordinarily, the de Broglie wavelength of the envelope function is much longer than that of the Bloch wave on the order of lattice spacing. Accordingly, the two quantizations are dealt with separately; one is for the Bloch quantization, and the other is for the size quantization.

3. MESOSCOPIC SYSTEMS

In the following sections, we discuss n-dimensional quantization ($n = 0,1,2,3$). Since real space has three dimensions, we have three degrees of freedom to quantize.

3.1 3D Quantum Systems

In an ideal 3D system, all Bloch waves are extended all over the system. This means that it is still a microscopic system, however large it is, so long as there are no decoherence mechanisms.

Besides these ordinary systems, it is worth mentioning that localized states due to disorder in three dimensions behave similarly to mesoscopic systems. Take a hypercube of d dimensions with a volume L^d. As shown in Fig. 2, localization is expected as a result of the disorder. Disorder tends to decrease the extension of electron waves from delocalized to localized states. This occurs because of the change in the boundary conditions or, in other words, because the translational symmetry is broken.

According to Thouless (1977), each localized state has a scale length ξ, over which the state extends. If $L \lesssim \xi$, the system seems ideally microscopic, where all states are extended all over the hypercube.

The following discussion makes use of the scaling theory due to Abrahams, Anderson, Licciardello, and Ramakrishnan (AALR: Abrahams *et al.*, 1979). The size dependence of conductance is illustrated in Fig. 3. When we discuss the conductance G in this system, it is natural to assume that the normalized conductance

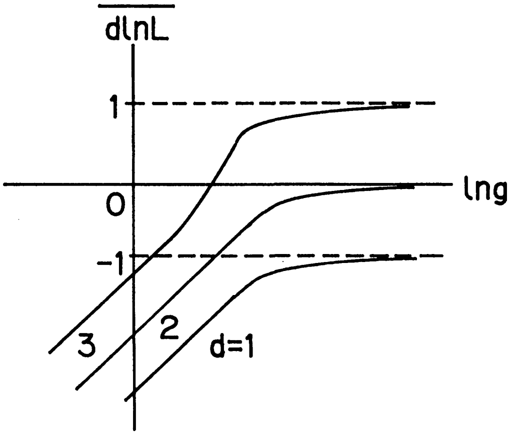

FIG. 3. Schematic diagram of $d\ln g/d\ln L$ vs $\ln g$, where $g = (2\hbar/e^2)G$, for d = 0, 1, 2 (after Abrahams *et al.*, 1979).

$$g = (2\hbar/e^2)G = (2\hbar/e^2)\sigma L^{d-2} \tag{1}$$

is proportional to $e^{-L/\xi}$ (σ is the conductivity); i.e., $g(L) = g_a e^{-L/\xi}$, or $d\ln g/d\ln L \approx \ln(g/g_a)$ for small g. Here, g_a is the conductivity that an extended wave shows. For large g, $g(L)$ is asymptotically given by the solution of the equation

$$\frac{d\ln g}{d\ln L} \sim d - 2 - \pi^2/g + \ldots \tag{2}$$

Thus, the asymptotes for $d = 1$, 2, and 3 are -1, 0, and 1, respectively. In three dimensions, this scaling implies $G(L) = \sigma L$, which increases with the size of the system L. The case for two dimensions is typical; $d\ln g/d\ln L$ is always negative and $g(L) \to 0$ for $L \to \infty$; $g(L)$ shows a crossover from logarithmic (for large g: *weakly localized regime*) to exponential localization (for small g: *strongly localized regime*). In the weakly localized regime, $g(L) \sim \text{const} - \pi^{-2} \ln L$, where L is very often replaced by l_ϕ or $\sqrt{D\tau_\phi}$ (D is the diffusion constant). Since τ_ϕ is proportional to T^{-p} (T, temperature; p, a number of order unity) according to the perturbational calculation of the electron-phonon and electron-electron interaction, it is seen that conductance has a logarithmic dependence on T. This reflects the fact that localized states can diffuse when inelastic scattering supplies them with the necessary amount of energy to delocalize.

The strongly localized regime is for $l_\phi \gg \xi$. In the opposite case, i.e., $l_\phi \lesssim \xi$, the conduction process is classified as follows:

1. for $L > l_\phi$, classical hopping conduction occurs; and
2. for $L \lesssim l_\phi$, electron waves are extended and coherent.

In order to discuss conductance in this system, Thouless (1977) introduced a dimensionless number called the "Thouless number," $g(L) = v(L)/w(L)$, where $v(L)$ is a typical energy, which may be expressed in terms of $\hbar/\tau_D$ (τ_D is the specific time for diffusion, defined by $D = L^2/\tau_D$), and $w(L)$ is the characteristic separation of levels at the Fermi energy, given by $1/N(\epsilon_F)L^d$ for a density of states $N(\epsilon_F)$. The equivalence of this Thouless number $g(L)$ to the above-mentioned $g(L) = (2\hbar/e^2)\sigma L^{d-2}$ can be easily proved from the generalized Einstein relation $\sigma = e^2 D N(\epsilon_F)$. It is interesting to note that localized states can convey an electric current via hopping through the most probable percolation path. Hopping itself is possible by the assistance of, e.g., phonon absorption. Following Mott (1974), dc conductivity is given by $\sigma \sim \exp[-(T_0/T)^{d+1}]$, where d is the number of dimensions in the system of concern (variable-range hopping).

Lastly, note that, in ordinary band theory in 3D, $N(\epsilon)L^3$ is given by $(2m/\hbar^2)^{3/2}\,(L^3/2\pi^2) \times \sqrt{\epsilon}$; i.e., $N(\epsilon) \propto \sqrt{\epsilon}$ is typical for three dimensions.

3.2 2D Quantum Systems

In a very thin slab of material, it is easily imagined that the Bloch waves are confined inside it with extinguishing wave functions outward from both the surfaces. This is an analog of the Fabry–Pérot resonator. In condensed matter, the waves can be confined, not by a vacuum but by an energy gap. Within a QW with barriers of height V_0 and width w, bound energy levels are created at $E_n = \hbar^2 k_n^2/2m$, where k_n is the solution of

$$k_n w = n\pi - 2\sin^{-1}\frac{\hbar k_n}{\sqrt{2mV_0}}.$$

Except in the case of an infinite V_0, the number of bound levels within the QW is finite.

In the following, the most well-known example of GaAs-AlAs heterostructures is numerically solved. The difference in electron affinity in the GaAs and $Al_{0.3}Ga_{0.7}As$ system is about 240 meV, and this causes a step-like change in the conduction-band bottom at the interface. Likewise, if the sum of the electron affinity and the energy gap is different in the well from that in the barrier, the system has a QW for holes. When the conduction bands of the well and barrier materials are represented by specific effective masses of 0.0665 and 0.5 times m_0, respectively, the solution of the Schrödinger equation is as shown in Fig. 4 for a well width of 50 Å. The lowest bound level within the QW is at -199.5 meV, whereas the second level is at -2.1 meV, very near but below the barrier conduction band of AlGaAs. The probabilities of these waves $|\Psi(x)|^2$ are also plotted.

In such 2D systems, the degrees of freedom parallel to the well still survive, and consequently, electrons move almost freely

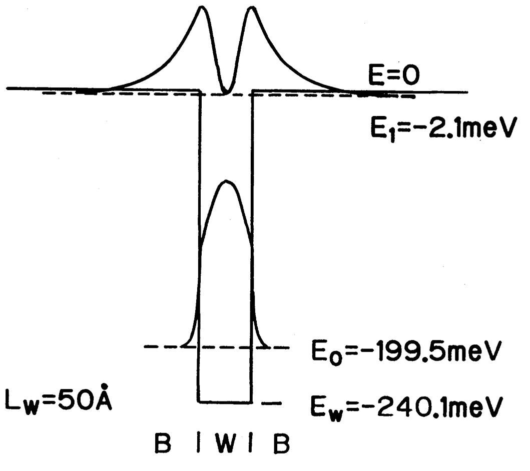

FIG. 4. Calculated $|\Psi(x)|^2$ for the ground and the excited level within a QW. The QW assumes the parameters of GaAs and $Al_{0.3}Ga_{0.7}As$ for the well and the barrier, respectively.

in that direction. Thus, 2D quantum systems can be realized by 1D confinement. If several *single-quantum wells* (SQW) are almost independent because of spaces between them, they are called *multiple-quantum wells* (MQW). Whether they are coupled or not is determined by simple quantum mechanics (see Electron States in Zero-, One-, and Two-Dimensional Structures). Any textbook teaches that the wave function within a QW leaks out to the neighbors separated by barriers of height V_0 and width $d - w$ with a probability of $\exp\{-2\kappa(d - w)\}$, where $\kappa = \sqrt{2m(V_0 - E)/\hbar^2}$.

When all QW's are closely coupled to each other, especially when they are ordered periodically with a period d, they are called an SL. An SL gives doubly periodic Bloch-like quantized states with two periods of a and d ($d \gg a$; d is ordinarily 15–100 Å; a is the lattice constant). The overlaid periodicity gives the system an additional BZ boundary in one dimension and minigaps on it. If an appropriate strength of electric field is applied perpendicular to the well, SLs can either completely reflect according to the Bragg reflection at the minigaps, or conduct incident electrons across the minibands via interminiband tunneling.

The previous section stated that the 2D systems are more likely to localize electrons, following AALR theory, than 3D systems. This is because any possible percolative paths are more strongly inhibited in 2D or, in other words, there are fewer redundancies in reaching a specific point from some starting point via a 2D percolative path.

In 2D, the density of states (DOS) is given by $N(\epsilon_F)L^2 = mL^2/\pi\hbar^2$, which is constant and independent of energy. If there is a finite number of quantized levels, the overall DOS is a sum of steps with offsets corresponding to quantized energies E_n.

3.3 1D Quantum Systems

One-dimensional quantized systems are more easily localized under the existence of disorder. They are called *quantum wires.* In them, two degrees of freedom are frozen, and only one survives, parallel to the wire.

The DOS in 1D systems is given by $N(\epsilon)L = \pi^{-1}(m/2\hbar^2)^{1/2}\, L/\sqrt{\epsilon - E_n}$, where E_n is the quantized energy. This DOS resembles that of 3D electrons under a magnetic field. In 3D, the degree of freedom parallel to the magnetic field is never affected by applying a field, whereas those perpendicular to the field are Landau quantized, and motions in 2D are reduced to the center-of-mass motion and the motion of relative coordinate. Although the center-of-mass coordinate gives a quantum number other than the Landau quantum number n, the quantized energy is incidentally independent of the center-of-coordinate quantum number or, in other words, degenerate with respect to this quantum number, so long as an electric field does not coexist. In fact, the DOS is proportional to $1/\sqrt{\epsilon - \hbar\omega_c(n + \frac{1}{2})}$, with $\omega_c = eB/m$.

In the case of intrinsically 2D systems, a magnetic field perpendicular to the system quantizes the energy into a zero-dimensional, dispersionless scheme. However, by applying an electric field within the 2D system, it is easy to lift the degeneracy with respect to the center-of-mass motion, which makes the system look like a 1D system within the 2D plane. Thus, it is easy to understand in the Hall bar geometry that the edge currents along both the sample edges parallel to the fed-in current still play an important role in experiments such as those on the quantized Hall effect (QHE), as will be discussed later.

3.4 0D Quantum Systems

These systems are called *quantum boxes* or *quantum dots*. The energy scheme is a series of levels, just like in clusters in solid states. Every state is confined within the box. In disks larger than ~100 Å, for example, the states may be approximately described by an envelope function and highly oscillatory Bloch-like states with modulation on the order of the lattice constant a. The envelope function is specified by the boundary conditions and is thus disk-shaped.

The DOS in the quantum boxes is obtained by simply counting those discrete levels existing between ϵ and $\epsilon + d\epsilon$. Ordinarily, these disks are with difficulty accessed electrically through leads. Accordingly, high-frequency properties as in optics and, sometimes, diamagnetic properties due to Landau motion are more suited for observing a zero-dimensional quantized system.

4. PHYSICAL PROPERTIES

4.1 Optical Properties

4.1.1 Single/Multiple-Quantum Wells Within semiconductors, shallow donors and acceptors, as well as excitons, can be described by hydrogenic wave functions. In order to discuss excitons, the mass should be taken to be a reduced value μ of the effective masses of an electron and a hole. Hence, the exciton binding energy E_n is $-\mathfrak{R}/n^2$, $n = 1, 2, \ldots$, where $\mathfrak{R} = m_0 e^4/2\epsilon_0^2\hbar^2$ is the effective Rydberg energy, $13.6(\mu/m_0)/\epsilon_0^2$ eV. The term ϵ_0 is the static dielectric constant. This is still the case within a QW, if the width is large enough. For a narrow QW, one degree of freedom of the spherical orbital is squeezed. An extremely thin QW will look like a 2D hydrogen atom. The binding energy in this limit is known to be given by $E_n = -\mathfrak{R}/(n - \frac{1}{2})^2$, $n = 1, 2, \ldots$. If both the lowest binding energies in 3D and 2D are compared, it is obvious that $4E_1(\text{3D}) = E_1(\text{2D})$. Correspondingly, the radial extension of the s orbital of a 2D exciton is half of that in 3D.

The fact that the binding energy of an exciton in a QW is larger than that in 3D has an important meaning for applications. For example, when a QW is formed as an active layer in a semiconductor laser, the emission efficiency is improved, partly because the binding energy can be larger than thermal excitation energy at room temperature (RT), and the exciton is not easily excited into a pair of a free electron and a hole that can more easily recombine without emitting luminescence. The other reason is that the DOS in 2D is constant, as was stated before, and consequently, the emissive transition probability is higher than in 3D, where the DOS starts from zero and is in proportion to $\sqrt{E}$.

Figure 5 shows the dependence of the binding energy of an exciton against the well width for the GaAs-Al$_{0.4}$Ga$_{0.6}$As system. For narrow widths, the energy first increases and then decreases. The reason is that for a vanishing well width, the bound energy level of an electron comes very close to the conduction-band bottom of the barrier material, and consequently, the wave function is no longer confined within the well, just as was shown for the second excited level E_1 in Fig. 4. This is also the case for holes (see EXCITONS). The

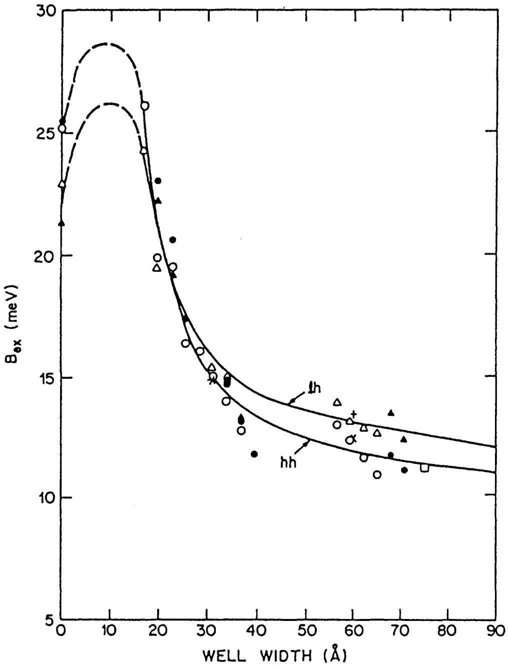

FIG. 5. Exciton binding energy within QW's vs well width. The labels *lh* and *hh* designate the exciton binding energy formed between, respectively, an electron and a light hole, and an electron and a heavy hole. Since the exciton belongs to the bound state within the QW, which has an energy inversely proportional to the effective mass, an *lh* exciton has a larger binding energy (after Nelson *et al.*, 1987).

excitonic state consisting of these extended orbitals is never confined within the well, and therefore, the binding energy of the exciton approaches again the value for 3D, lower than the value for 2D.

Figure 6 shows the photoluminescence (PL) as a function of the well width (see PHOTOLUMINESCENCE). The PL photon energy corresponds to the difference between the lowest bound-state energies for electrons and holes minus the exciton binding energy. The binding energy is given in Fig. 5. The limiting case for a vanishing well width in this figure must be equal to the energy gap E_g minus the exciton binding energy, both in 3D bulk.

Another interesting phenomenon is seen in the efficiency of PL. When MQW samples are fabricated with barriers too thick to couple with each other, it is easy to observe a number of PL spectra, under a single-shot excitation, from wells with various widths. From a well of a certain thickness, PL of a specific wavelength comes out, and its intensity is proportional to the mth power of the excitation intensity, $I_{\mathrm{PL}} \propto I_{\mathrm{ex}}^m$. This power index depends on the well width, as is shown in Fig. 7. The measurement was done by Mishima *et al.* (1986). As already stated, when the well width w increases from a very small value, the lowest bound level lowers from the conduction-band bottom of the barrier material. Around $w \approx 4.7$ nm for $\mathrm{Al_{0.3}Ga_{0.7}As/GaAs}$ MQW, a second bound level appears. This must look as shown in Fig. 4. Note that PL always comes from the recombination of an exciton belonging to the lowest-electron and highest-hole bound states. When the highest electron level comes near the barrier conduction band, i.e., the *resonance condition* holds, photoexcited electrons can easily be freed from within the well. On the other hand, when there is a considerable energy depth between the conduction-band bottom of the barrier and the highest level within the well, i.e., the *off-resonance condition* holds, photoexcited carriers are tightly trapped within the well and contribute significantly to the PL. This situation reflects on the PL efficiency, which depends on the well width in the manner shown in Fig. 7. For $m \sim 2$, the efficiency is low, whereas for $m \sim 1$ it is high. This periodicity was accurately interpreted on the basis of the argument above (Murayama, 1986).

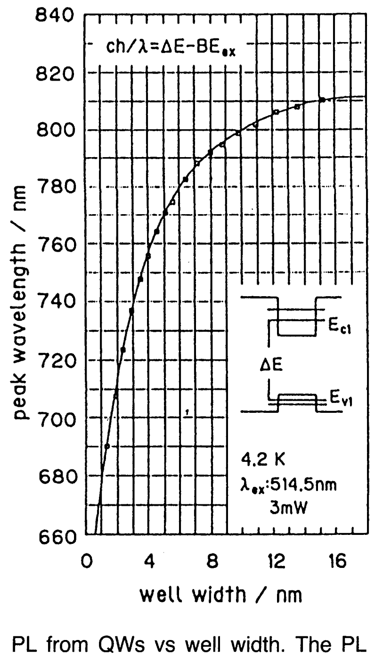

FIG. 6. PL from QWs vs well width. The PL energy is equal to the separation between the lowest electron and the highest hole level minus the exciton binding energy (after Murayama, 1992).

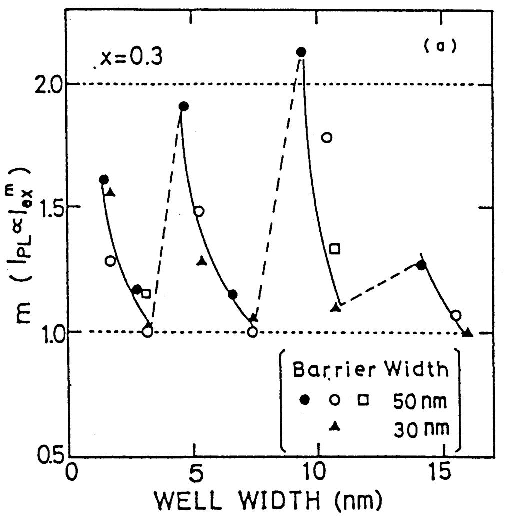

FIG. 7. Power index m: $I_{\mathrm{PL}} \propto I_{\mathrm{ex}}^m$ vs well width. $m = 2$ and $m = 1$ correspond to low-efficiency resonance and high-efficiency off-resonance, respectively (after Mishima *et al.*, 1986).

4.1.2 Superlattice In an SL, there are minibands and minigaps (see HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR), as shown in Fig. 8(a). When an electric field

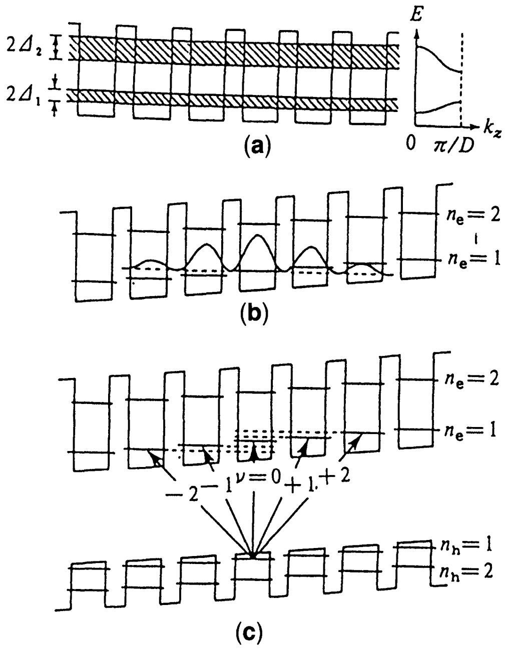

FIG. 8. (a) Minibands in an SL. (b) An electron state that has the largest amplitude at the center of the SL, when an electric field is applied. (c) Suggested optical excitation of an electron in the valence band at the center of the SL to several electron states. The largest transition probability occurs between the electron and hole within the same well (after Nakayama *et al.*, (1991).

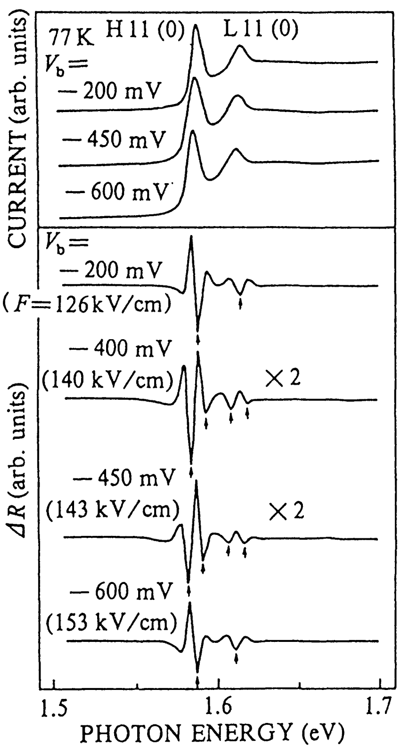

FIG. 9. Spectra of photocurrent and reflectance change vs photon energy as a function of applied electric potential V_b (after Nakayama *et al.*, 1991).

is applied perpendicularly to this device, the potential of each QW varies as a function of the well position, as schematically shown in Fig. 8(b). So, we can expect a series of optical transitions between a hole level, e.g., in the SL center, and electron levels with offset energies according to the Stark effect. Figure 9 plots the photocurrent spectrum and the variation in reflectance spectrum with electric field against the photon energy for various fields. The strongest signal is obtained for both the electron and hole within the same well, and weaker ones are for the transitions to the electron levels in neighboring wells.

It is seen in Fig. 9 that resonant tunneling occurs around $V_b \approx -(400\text{–}450)$ mV between the $n_e = 1$ miniband for $\nu = 0$ and the $n_e = 2$ miniband for $\nu = -2$ QW, causing very weak reflectance changes ΔR. When the resonance condition holds, both the $n_e = 1$ and 2 minibands are strongly mixed so that they lose specific spike structures according to the QW features.

4.1.3 Selection Rules Silicon is the best material to manufacture inexpensive transport devices in mass fabrication at the present time. However, it is never applied to light-emitting devices, because it is known to perform only indirect optical transition, and hence, the emitting efficiency is extremely low. Using SLs with an overlaid periodicity of n times the original lattice constant, it is recognized that the BZ in the bulk state can be folded into $1/n$ of the original reciprocal lattice vector in one dimension. For example, by making an SL by alternately growing Si and SiGe alloy, the conduction-band minimum of Si may come close to the Γ point in the BZ. Does such Si with a folded BZ show a higher light-emitting efficiency? This is not necessarily so, because the transition matrix element remains the same after the folding, although the k-selection rule can change from indirect to direct transition.

Regarding the optical properties of Si, recent PL studies in porous Si (Canham, 1990) are worth noting from the viewpoint of application of mesoscopic systems. When Si wafers are immersed in HF acid and anodized, the surface is mesoscopically processed and optical transition is no longer indirect as in bulk. So far, cluster and quantum wire models have been proposed, where Si dangling bonds on the surface of the mesoscopic systems must be terminated by H and/or O atoms.

4.2 Transport Properties

4.2.1 Transport Parallel/Perpendicular to QW

As described in the preceding sections, single/multiple QWs are specified by their size-quantized energy levels in the direction perpendicular to the wells. Transport properties are first of all classified into (1) transport parallel to the wells and (2) transport perpendicular to the wells.

4.2.1.1 Parallel Conduction—Metal-Oxide Semiconductor (MOS). Conduction within the well planes is the same as that within thin slabs. Ordinarily, these are nearly 2D systems. Sometimes there is a heterostructure between the well and the barrier material, which causes a sharp potential dip and makes the electronic system perfectly 2D. This case resembles the Si-SiO$_2$ interface in an MOS device. The 2D electron gas (2DEG) sheet is formed in an MOS by the built-in electric field generated between SiO$_2$ and Si so as to equalize the chemical potentials. (see SI-SIO$_2$ INTERFACE, ELECTRONIC PROPERTIES OF; TRANSISTORS, FIELD EFFECT).

Let us explain the band structure at the interface in terms of the built-in field F. Bulk n-type Si is known to have four degenerate valleys, each situated around one of the minimum points: $(\pm 0.85,0,0)\pi/a$, and the equivalents. Let us assume that an MOS is formed on the (001) plane (Fig. 10). Then the energy band in the [001] valley is given by

$$E_n(\mathbf{k})^{[100]} = E_n^l + \hbar^2 k_y^2/2m_t + \hbar^2 k_z^2/2m_t,$$

with

$$E_n^l = (3\pi\hbar eF/2\sqrt{2m_l})^{2/3}\,(n + 3/4)^{2/3}.$$

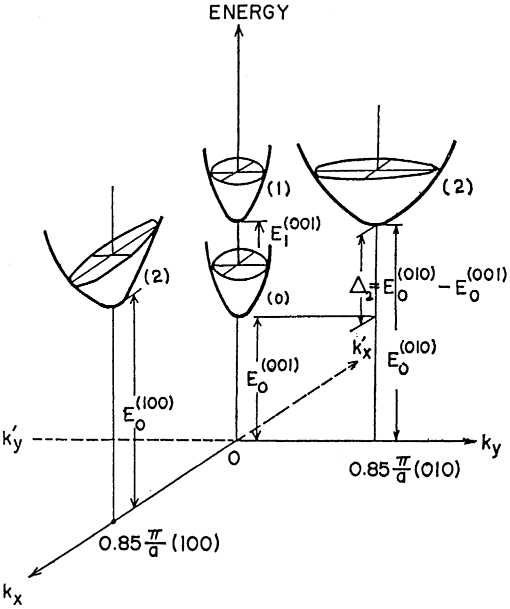

FIG. 10. Suggested valley structure quantized by a built-in electric field at the (100)-Si-SiO$_2$ interface (after Murayama *et al.*, 1972).

This analytical result is obtained in the triangular electric potential, which is good for a sufficiently strong field (Ando *et al.*, 1982). For the valleys in the directions of [100] and [010], the quantized energies E_n^t are given by replacing m_l by m_t in the above equation; m_l and m_t are the effective masses in the longitudinal and transverse directions in the [001] valley, respectively. The term $m_l \approx 1.58m_0$ and $m_t \approx 0.081m_0$, and therefore, $E_n^t \gg E_n^l$. Thus, in a (001)-MOS device, all four valleys [±100] and [0 ±10] are degenerate, but the other two valleys, [00 ±1], are not and have a lower energy. For the [100] valley, the parabolic band mass is m_t, whereas, in the other valleys, the parabolic masses are m_t and m_l. Consequently, higher bands have a larger mass and the lower band has a smaller mass. This difference in masses can cause a phenomenon similar to that called the Gunn effect in bulk GaAs, i.e., a differential negative resistance (DNR) device. Both the experimental work (Katayama *et al.*) and the theoretical analysis (Murayama *et al.*) were done in 1972.

4.2.1.2 Perpendicular Conduction—Resonant Tunneling Device. A QW is formed between two barrier materials and is specified by a size-quantized energy level. If electrons flow perpendicular to the well, the requirements of energy and momentum conservation limit the allowable tunneling between the energy levels outside the barrier and those inside the well. This relation is schematically described in Fig. 11(a). When an electric field is applied across the QW, most of the potential drop occurs across the barrier materials because of their high resistivity. First, for an electric field, tunneling is allowed at the Fermi level satisfying the energy-momentum conservation condition (see RESONANCE TUNNELING). For a higher field, tunneling may no longer be allowed. This increase in electric field sometimes means a decrease in current, i.e., a DNR (Tsu and Esaki, 1973). An example of observations is shown in Fig. 11(b). This DNR characteristic is specific to RTDs.

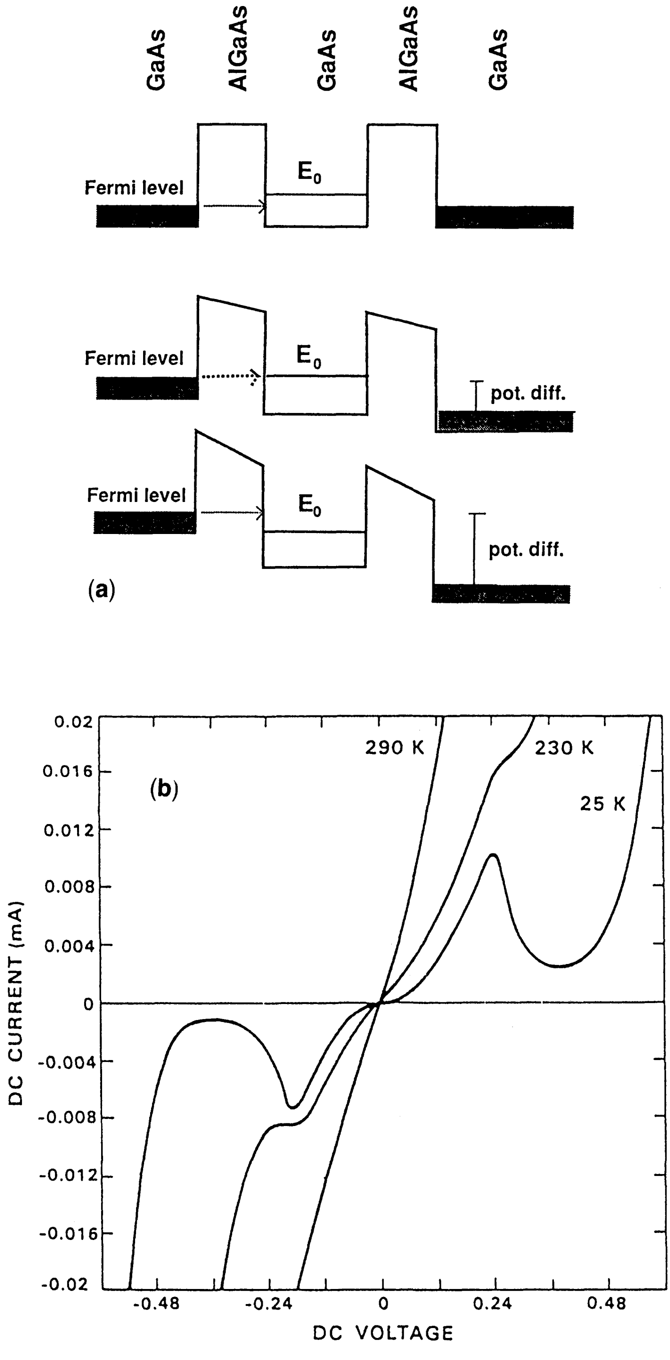

FIG. 11. (a) Suggested mechanism of an RTD. (b) Observed dc current vs dc voltage showing a negative differential resistance (after Sollner *et al.*, 1983, reprinted with the permission of Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA).

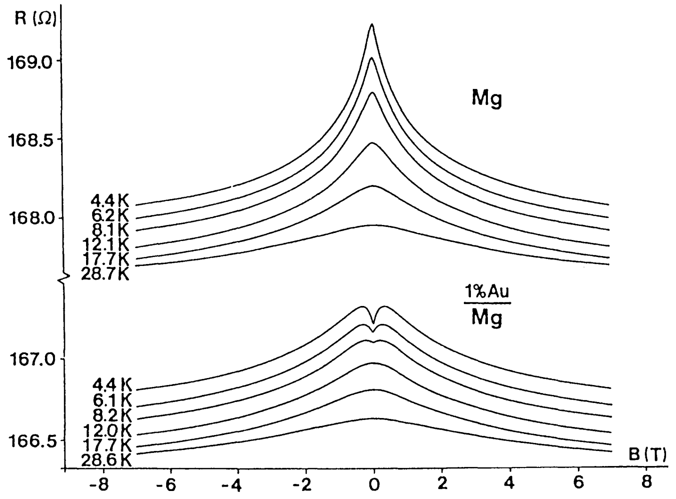

FIG. 12. Observed negative magnetoresistance of Mg thin films uncovered (upper data) and covered (lower data) with Au, as a function of temperature (after Bergmann, 1983).

4.2.2 Negative Magnetoresistance Ordinary nonmagnetic materials show a slight increase in transverse resistance under an applied magnetic field. The increase is proportional to B^2. This is natural because of symmetry; i.e., the system should behave symmetrically for B and $-B$. Since no classical transport theory explains this dependence on B (the van Leeuwen theorem), the magnetoresistance must be attributed to a quantum effect.

In a 2DEG in a very thin Mg film, magnetoresistance was measured by Bergmann (1983), as shown in Fig. 12. The resistance decreases with increasing magnetic field B. This sample seems to be in a weakly localized regime with slight disorder. In Anderson localization, coherence causes an electron to localize around a specific site on the quasi-1D path shown in Fig. 2. Let us simplify Fig. 2 by assuming a homogeneous phase distribution on the ring in Fig. 13. Here, the total phase change Θ is composed of the accumulated phase shift on each scattering site, $\Sigma_i \theta_i$, and of kL, where k is the wave vector and L the perimeter. Now, Θ is divided into $c\Theta$ and $(1-c)\Theta$. The wave is assumed to start from point O and the superposition of the waves is considered at point A. The analytical expression is shown in Fig. 13. The Aharonov–Bohm phase is denoted by $\chi = 2\pi BS/\Phi_0$ with the enclosed area S and the quantum of flux $\Phi_0 = h/e$.

When a magnetic field B is applied to this system, the most probable existence of the wave is at a site different from that in the absence of B, since

$$|\Psi(\mathrm{A})|^2 = |\psi_c + \psi_{1-c}|^2 \propto \{1 + \cos[\chi - \Theta(1 - 2c)]\}. \qquad (3)$$

It is obvious that the most probable constructive interference occurs on the antipodal point A, i.e., $c = \frac{1}{2}$, in such a simplified situation as $B = 0$.

Another effect of a magnetic field on mesoscopic systems is the change in the scale length under a strong field; i.e., the system size L should be replaced by a magnetic length $L_B = \sqrt{\hbar/eB}$, when $L \gtrsim L_B$. Thus, $g(L_B)$ in Eq.

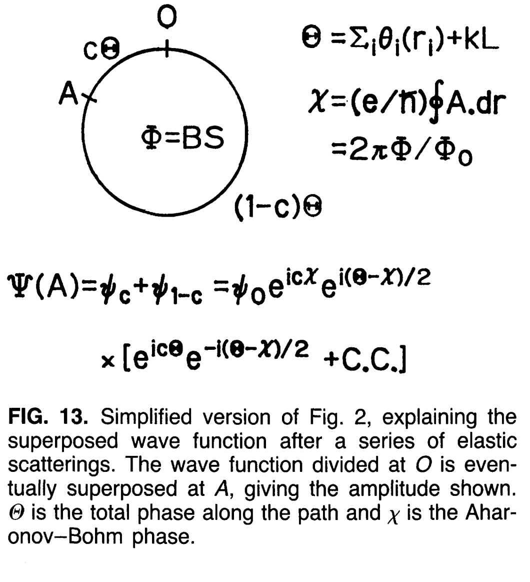
O
cΘ
A
Φ=BS
(1-c)Θ
Θ =Σ_iθ_i(r_i)+kL
χ=(e/ħ)∮A.dr
=2πΦ/Φ_0
Ψ(A)=ψ_c+ψ_{1-c} =ψ_0 e^{icχ} e^{i(Θ-χ)/2}
× [e^{icΘ} e^{-i(Θ-χ)/2} +C.C.]

FIG. 13. Simplified version of Fig. 2, explaining the superposed wave function after a series of elastic scatterings. The wave function divided at O is eventually superposed at A, giving the amplitude shown. Θ is the total phase along the path and χ is the Aharonov–Bohm phase.

(2) gives the dependence on the field in the form $\Delta g(L_B) \sim -\pi^{-2} \ln L_B$, which increases with B (see Section 3.1).

The situation differs when there is strong spin-orbit coupling. In Fig. 12, the lower part shows the data for Mg film covered with Au. There is strong spin-orbit coupling between Mg and Au, and the characteristic time is the spin-orbit scattering time τ_{SO} rather than the ordinary inelastic scattering time, especially at lower temperatures. This τ_{SO} is of an elastic nature and is almost independent of temperature. It affects the magnetoresistance positively because the spin-orbit interaction breaks the time-reversal symmetry. All these experiments are very accurately interpreted by the phenomenological theory by Hikami *et al.* (1980) in terms of a few parameters τ_{el}, τ_{in}, τ_{SO}, and B.

The same kind of negative magnetoresistance is also observed in a pure 2DEG system in the channel of MOS devices, as shown by Kawaguchi *et al.* (1978).

4.2.3 Aharonov–Bohm Effect Figures 2 and 13 showed a conceptual description of Anderson localization, where the point was that after multiple elastic scatterings the wave comes back to itself, and the original and the returned waves can be superposed constructively at a certain site or destructively at other sites. When the superposition is constructive, the wave can have a larger amplitude than the average.

In 1981, Sharvin and Sharvin (1981) prepared a quartz fiber with a diameter of ~2 μm covered with vacuum-evaporated Mg thin film and measured the transport properties of the sample with two leads separated from each other in the direction of the fiber under a parallel magnetic field (Fig. 14). This experiment was suggested by a preceding theory of Al'tshuler *et al.* (1981). This sample simulates the geometry shown in Fig. 13, although scattering occurs against defects, not within a connected but within a disconnected region. On the equation shown in the figure, assume $c = 0$ and $c = 1$, which mean clockwise and anticlockwise travel of the wave, respectively. Since for $c = 0$,

$$\Psi(c = 0) = \psi_0(1 + e^{i(\Theta-\chi)}), \tag{4}$$

while for $c = 1$,

$$\Psi(c = 1) = \psi_0(e^{i(\Theta+\chi)} + 1), \tag{5}$$

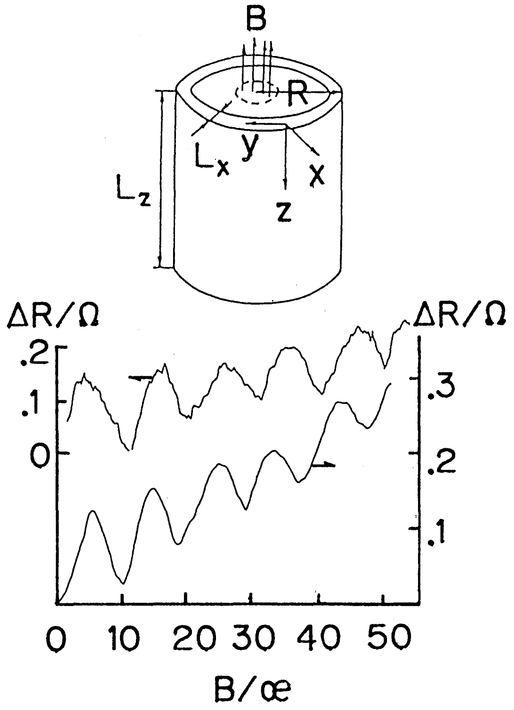

FIG. 14. Observed Aharonov–Bohm effect on disordered thin Mg hollow cylinder under a parallel magnetic field. The data on two samples are shown, scaled on the left and right ordinates as shown by the arrows (after Sharvin and Sharvin, 1981).

the superposed wave at point O for both directions gives its probability of existence to be proportional to $1 + 2\cos\Theta \cos\chi + \cos^2\chi$. This is peculiar in that a periodicity is expected for a magnetic field change of $S\Delta B = \Phi_0/2$, as actually happened in the experiment (Sharvin and Sharvin, 1981).

Another conventional Aharonov–Bohm effect in solid state was observed using a small metallic ring (Webb *et al.*, 1985, 1987; Washburn *et al.*, 1985) (Fig. 15). In this case, the effect is understood analogously to the electron-beam interference experiment. The period is, in contrast to the Sharvins' experiment, $\Delta B = \Phi_0/S$. The difference comes from whether the travel of the waves occurs along the fiber axis or does so in the plane merely in one turn, when the orbital is localized within the plane perpendicular to the direction of B. After the first observation of periodic variation of conductance through a small ring, other experiments improved the periodic behavior by using samples with a higher aspect ratio (the diameter versus the width

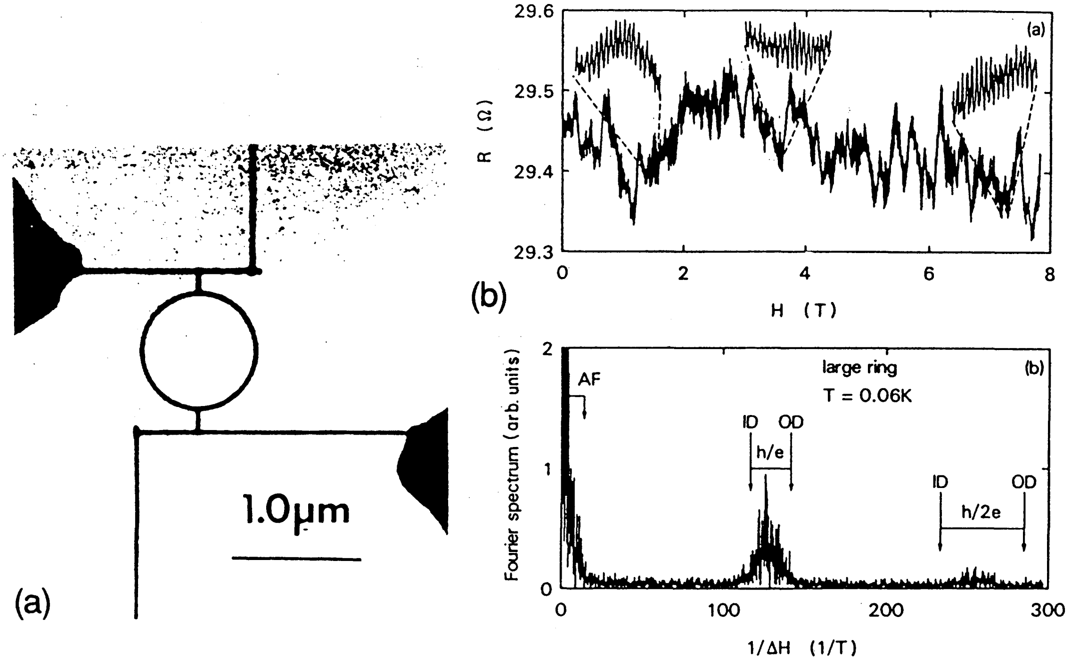

FIG. 15. (a) Photo of a small Au ring sample. (b) Measured variation in resistance vs magnetic field *H*, and its Fourier-transformed spectrum vs 1/*H* (after Washburn *et al.*, 1985).

of the ring). This improvement results from the fact that the magnetic field can penetrate the arm of the ring itself, when the arm is very wide. An enhanced aspect ratio gives fewer channels to be averaged over. If there are many paths inside the ring arm, the Aharonov–Bohm effect occurs for each path, and the conductance must eventually be averaged over all the paths.

The latter Aharonov–Bohm effect with a period Φ_0/S is sensitive to the ensemble average. This fact is obvious from the strong dependence on the aspect ratio of the sample. In contrast, the former effect, with the period of $\Phi_0/2S$, is not sensitive to the ensemble average.

4.2.4 Landauer–Büttiker Formula of Conductance Before discussing the universal conductance fluctuation, it seems helpful to derive first an intuitional formula for conductance that is specifically applicable to mesoscopic systems. This is called the Landauer formula (Landauer, 1957), after the inventor's name, and sometimes Landauer–Büttiker formula (Büttiker *et al.*, 1985), attaching the name of the researcher who extended the formula to a more general case (see Electron States in Zero-, One-, and Two-Dimensional Structures).

Let us follow Imry's discussion (1986). As shown in Fig. 16, the electron incident from the left channel *i* tunnels into the right channel *j* with a probability $\Sigma_j f_L T_{ij}$. There are numerous channels in an ordinary case. The $f_{L,R}$ is the Fermi distribution function in the left (L) or right (R) lead acting as a reservoir. Similarly, the electron incident from the right channel *i* is reflected into the right channel *j* with a probability $\Sigma_j f_R R_{ij}$. The sum of these probabilities minus the equilibrium distribution gives the current transported on

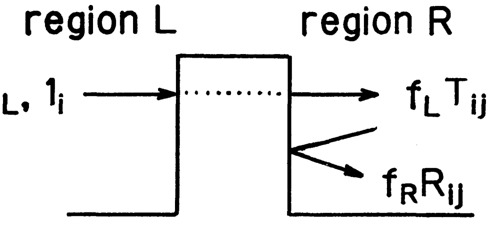

FIG. 16. Definition of channel *i* and its probability of occupation f_i around regions R and L with a tiny barrier in between. *T* and *R* are transmission and reflection, respectively.

channel i from left to right, i.e.,

$$I_i = \int dEev_iN_i(E)\sum_j (f_LT_{ij} + f_RR_{ij} - f_R) \quad (6)$$

$$\sim [e(\mu_L - \mu_R)/\pi\hbar] \int dE(-df/dE)$$
$$\times \Sigma_jT_{ij}(E), \quad (7)$$

where the density of states in 1D is $N_i(E) = 1/\pi\hbar v_i$ (v_i is velocity) and $R + T = 1$. μ_L (μ_R) is the chemical potential in region L (R). Since the conductance G is given by $\Sigma_iI_i/[(\mu_L - \mu_R)/e]$, it is easy to obtain the Landauer formula $G = (e^2/\pi\hbar)\Sigma_{ij}T_{ij}$. Very often $\Sigma_{ij}T_{ij}$ is written as $\mathrm{tr}tt^\dagger$, using the transmission probability amplitude t: $T_{ij} = t_{ij}t_{ji}^\dagger$. This formula holds when electrons are fed directly from both the "classical" reservoirs. Classical means that electrons are completely incoherent within the reservoir because of frequent inelastic scattering. On the other hand, with two "ideal" leads between the sample and the reservoirs, the formula is given by Landauer's original expression: $G = (e^2/\pi\hbar)T/R$.

4.2.5 Universal Conductance Fluctuation In Fig. 13, it was shown that a localized state is defined on an idealized quasi-1D ring with a distributed phase on the perimeter. It is important that the state thus defined is localized around α; i.e., Ψ_α is very sensitive to parameters such as magnetic field B and chemical potential μ: $\Psi_\alpha(B,\mu, \ldots)$. Thus, in what manner the state is localized appears in its dependence on B, μ, and so on. This phenomenon was called a magnetofingerprint by Stone (1987), since the system behaves always specifically as a function of B.

The system of concern intrinsically has disorder. It is characterized by the mode in which such disorder is distributed. In mesoscopic systems, these disorder-induced fluctuations appear in, e.g., conductance, and are never averaged out. This is called *conductance fluctuation* specific to the system of concern. When the system is made larger and larger, the fluctuation disappears, as can be imagined easily, and approaches a macroscopic limit.

In contrast, when the size is reduced, the fluctuation is enhanced. Let us consider the conductance G; G is proportional to $(2e^2/h) \times \int dt\,\langle\Psi_\alpha|v(t)|\Psi_\beta\rangle\langle\Psi_\beta|v(0)|\Psi_\alpha\rangle$, which is the sum over all possible pairs $\alpha\beta$. The term v is the velocity operator. The deviation in G from the mean value as a function of B, μ, and so on should be given by a higher-order term: $\langle\delta G(B,\mu, \ldots)\delta G(B + \Delta B,\mu + \Delta\mu, \ldots)\rangle$ which is the correlation among four localized states. As suggested above, this fluctuation approaches a *universal conductance fluctuation* (UCF) given by

$$\Delta G = \frac{e^2}{h} \times \begin{pmatrix} 0.729 \\ 0.862 \\ 1.088 \end{pmatrix} \quad \text{for} \quad \begin{pmatrix} \text{1D} \\ \text{2D} \\ \text{3D} \end{pmatrix} \quad (8)$$

after Lee *et al.* (1987). These three fractional numbers are only numerically calculated.

Experimentally, Webb *et al.* (1987) (Fig. 17) measured the voltage fluctuation as a function of the length of thin gold wires. As shown in the inset, the variance decreases in proportion to $L^{1/2}$ and approaches a constant value equal to the UCF. For a sufficiently small system $L \ll \xi$ (ξ is coherence length), the voltage fluctuation should disappear as suggested in the inset, because coherent conduction means "lossless" conduction.

4.2.6 Quantized Conductance In the arrangement shown on the left or right half in Fig. 18(a), electrons are emitted through an orifice formed by potential barriers on both sides and a through-channel with a lower po-

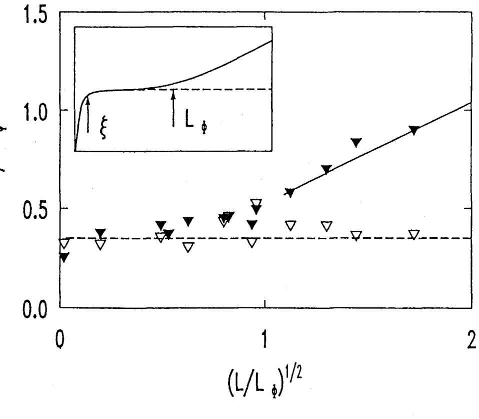

FIG. 17. Variance in voltage measured on short thin Au quantum wires. For a vanishing length a UCF appears. The inset explains schematically a theoretical variation of *V* vs *L*. The solid and open symbols represent, respectively, symmetric and antisymmetric contributions against the exchange of current and voltage leads. Only the symmetric data should be compared with the theory (after Webb *et al.*, 1987).

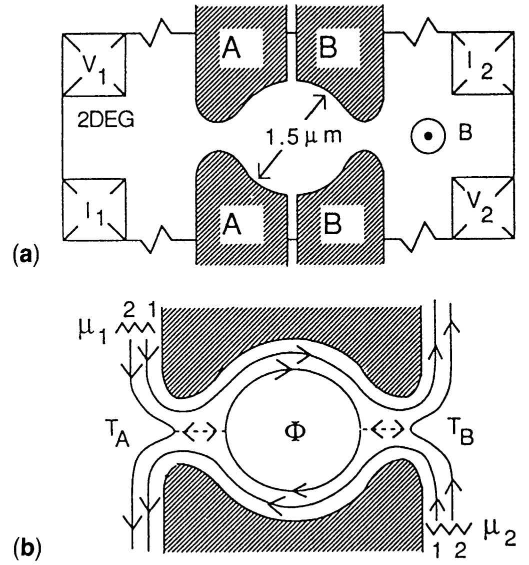

FIG. 18. Circuit within which coherent current channels are controlled by applying a voltage to the gates. With both gates A and B switched, closed-current channels are formed, and for either A or B switched, an orifice is formed (after van Wees *et al.*, 1989).

tential in between. This kind of orifice is called a *quantum point contact* (QPC). Figure 18(b) as a whole shows a circuit with two QPC's, between which there is the room to accommodate intereference between edge currents. When there is a single QPC, conductance is known to show a series of quantized values with plateaus on each step, as shown in Fig. 19.

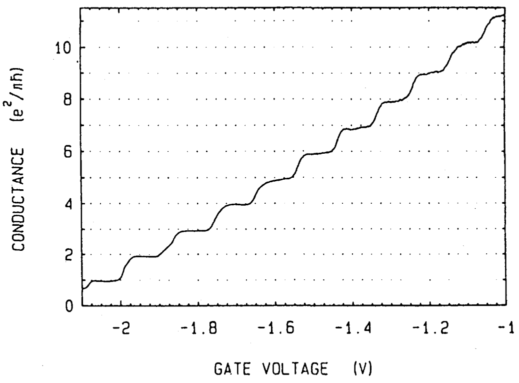

FIG. 19. Conductance quantized in units of $e^2/\pi\hbar$ observed on an orifice, which allows a number of channels determined by the gate voltage to pass (after van Wees *et al.*, 1988).

This behavior is understandable if the transport is attributed to a bundle of 1D channels, as the Landauer formula says. Let the orifice be sufficiently wide and low in its potential well. Then most channels run through the orifice without being blocked by the barrier potential. Otherwise, some may be blocked. One can count the number of channels running through the orifice. Since each channel carries a conductance of $e^2/\pi\hbar$, the total aggregate conductance may manifest a pileup of integral multiples of $e^2/\pi\hbar$, the number of which increases from unity to some integer, when the voltage increases. This is because the applied voltage lowers or widens the potential at the orifice.

In the limit where no potential channel remains wide or shallow enough to accomodate a single 1D channel, current is only transported by thermal excitation over or quantum tunneling through the potential barrier at the orifice.

4.2.7 Interference between Edge Currents Figure 20, curve A, shows observed conductance as a function of magnetic field

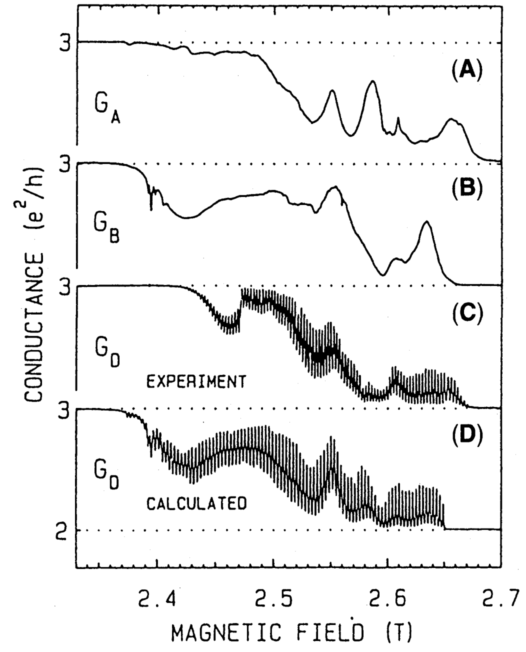

FIG. 20. Observed conductance as a function of magnetic field, when a single orifice (A) at *A* or (B) at *B* is formed and (C) when a closed-channel circuit is formed. (D) A simulation result (after van Wees *et al.*, 1989).

when the barrier potential B does not exist, and hence, the circuit appears to have a single QPC A. Figure 20, curve B, is for a similar case with a single QPC B. If the barrier potentials are switched on, a circular quantum dot is formed by the two QPC's and the circumferences of the barriers. Figure 20, curve C, depicts the observed oscillation of conductance, with a period in B given by $2\pi S\Delta B/\Phi_0 = 2\pi$, where S is the area of the circle shown in Fig. 18(b).

Here it is instructive to compare classical and quantum conditions for geometrical resonance of electrons under a magnetic field. It is known that, by application of the correspondence principle, a cyclotron motion is specified by the condition $E = \frac{1}{2}mr_c^2\omega_c^2 = (n + \frac{1}{2})\hbar\omega_c$. Thus, the classical cyclotron orbit radius is known to be $r_c = \sqrt{\hbar/eB}$ according to the correspondence principle. On the other hand, quantum mechanics tells us that the phase of an electron circulating under a magnetic field is $\pi l^2 2\pi B/\Phi_0$, which results in $l = \sqrt{2}r_c$ if the acquired phase equals 2π. Thus, the motion of a classical particle under a magnetic field can be exactly reinterpreted quantum mechanically as a change in the Aharonov–Bohm phase, resulting from the vector potential generating the field.

4.2.8 Integral Quantum Hall Effect

As stated before, conductance in 2D can be independent of the system geometry as $G \propto L^0$, and accordingly, conductance is equal to conductivity. This means that G is given by constants like e, $\hbar$, and m, independently of the sample size. This condition actually occurs in the Hall geometry in the 2DEG in a MOS under a magnetic field.

Let the width and length of a 2D channel be W and L in the y and x directions, respectively. Then

$$I_x/W = \sigma_{xx}V_x/L + \sigma_{xy}V_y/W,$$
$$I_y/L = \sigma_{yx}V_x/L + \sigma_{yy}V_y/W. \tag{9}$$

For a vanishing I_y and $V_H = V_y$ ($\sigma_{xy} = -\sigma_{yx}$) (ordinarily this is the case for the Hall effect),

$$BR_H = \frac{V_H}{I_x} = \frac{\sigma_{xy}}{\sigma_{xx}\sigma_{yy} + \sigma_{xy}^2}. \tag{10}$$

This means that for the observed voltage V_H and current I_x, the ratio is given by a quantity including universal constants and, at least, τ. When $\omega_c\tau \gg 1$, $\sigma_{xx} = \sigma_0/[1 + (\omega_c\tau)^2] \rightarrow 0$ and $\sigma_{xy} = -ne/B$ (Fig. 21). The density of Landau states in 2D is shown in Fig. 22(a) as a function of energy. Without scattering, they must be sharp levels with a degeneracy of $1/2\pi l^2$. Here, l is the magnetic length defined by $l = \sqrt{\hbar/eB}$. If the Landau sub-bands are filled up to i, the density of electrons $n = i/2\pi l^2$ gives $BR_H \sim 1/\sigma_{xy} = -h/ie^2$. Thus, the observed ratio V_H/I_x must be equal to a universal constant. This is why the effect is called the integral QHE, because the Hall constant times B is quantized to a universal constant divided by an integer. This only interprets that the observed BR_H value should reach $-h/ie^2$ at the point where $n = i/2\pi l^2$. To determine the integer n experimentally, we need the plateaus around the n on the observed curves.

Every experiment is performed as a function of the gate voltage V_G, or the density of carriers n. When the DOS is replotted as a function of n, the variation should read as in Fig. 22(b). However, it is known that electrons are very easily localized in 2D. Thus, each Landau subband has localized states besides the central delocalized narrow band.

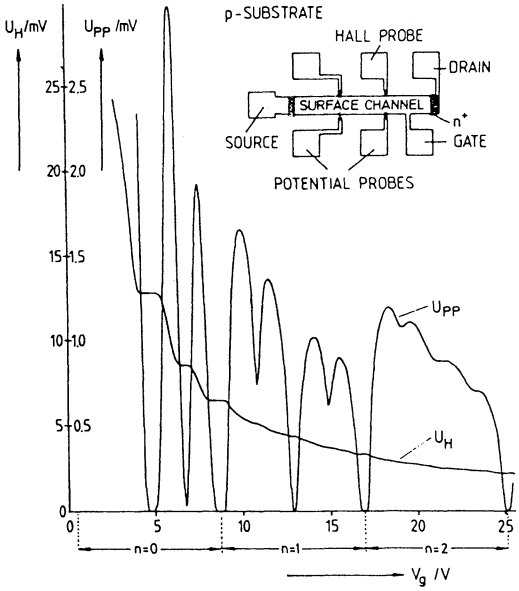

FIG. 21. Observed Hall voltage U_H (V_H in text) between the Hall probes (see the inset) and voltage U_{PP} (V_x) between the source and drain under a constant applied current as functions of gate voltage V_g on n-channel MOS Hall bar sample. U_H shows a number of plateaus, i.e., integral QHE (after von Klitzing *et al.*, 1980).

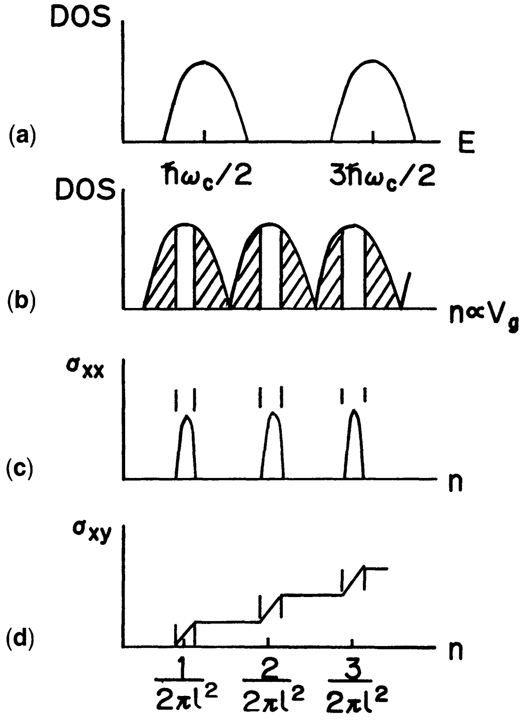

FIG. 22. Conceptual sketch of 2D Landau levels (a) as a function of energy and (b) as a function of carrier density *n*. When localization occurs on both sides of 2D Landau levels, σ_{xy} is quantized as shown in (d), where σ_{xx} vanishes (c).

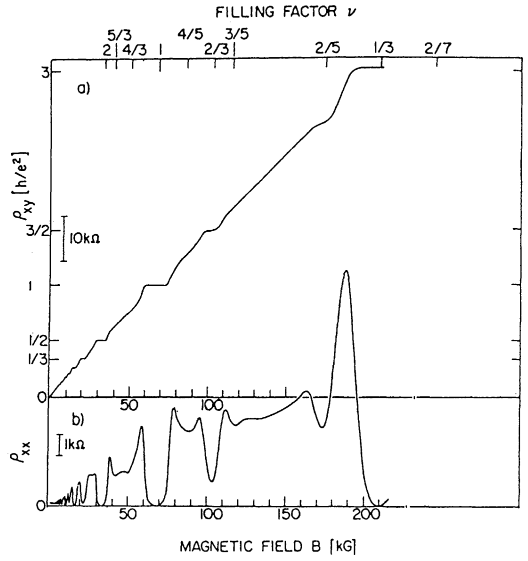

FIG. 23. Observed ρ_{xy} and ρ_{xx} as a function of magnetic field *B*. A number of plateaus appear for fractional filling factors shown on the upper axis, i.e., fractional QHE (after Störmer *et al.*, 1983).

Unless there is localization, the Hall-effect experiment would never have plateaus in the Hall coefficient as a function of *n* or V_G. The existence of localization was never thought of beforehand in experiments, but was only deduced from research on the integral QHE.

4.2.9 Fractional Quantum Hall Effect In the integral QHE, plateaus appear for parameters *n* that let the Fermi energy stand between neighboring Landau subbands. That is, all the subbands below the quantum number *N* are occupied, but those above are unoccupied. Historically, the integral QHE was discovered in a 2DEG at the Si-SiO$_2$ interface.

On the other hand, it is well established that the heterostructure consisting of GaAs and AlGaAs gives an excellent quality 2DEG. It shows a very high mobility up to ~1000 m^2/V s and a mean free path of several microns. This kind of sample often shows very flat plateaus at magnetic fields, as seen in Fig. 23 (Störmer *et al.*, 1983), for which the carrier densities are such that a Landau subband is occupied fractionally to $\nu = p/q$ of the total electrons of $1/2\pi l^2$ per subband and unoccupied beyond. The variable p ($\neq q$) is an integer and $q = 3, 5, 7, \ldots$, an odd number, because of the Fermi statistical requirement.

This effect is called "fractional" QHE. Experimentally, integral QHE is known to occur in dirty samples, which seem to ease localization. On the other hand, fractional QHE is only observed in clean samples with a high mobility. Fractional QHE shows plateaus within the same Landau subband as in integral QHE. To simplify the explanation, let us limit ourselves to the $q = 3$ case. To interpret the effect, Laughlin (1983) introduced the possibility of correlated three-particle states.

The 2D Landau electron gas is degenerate, as already mentioned, with a degeneracy of $1/2\pi l^2$, which comes from the fact that the eigenenergy is not explicitly dependent on the quantum number, the center coordinate of cyclotron motion. However, because of the correlation energy between Landau electrons, they are mutually repulsive and are likely to form, e.g., a triangular lattice, just as in the Wigner crystal. The ground three-body state has threefold symmetry due to its correlation energy, and its occupation is $\nu = \frac{1}{3} - \delta$ (δ: infinitesimal). The Laughlin state is a trial function to perform variation on and is constructed so as to maximize mutual repulsion. More accurate calculations show that the excited state with an occupation of $\frac{1}{3} + \delta$ has a tiny Coulomb gap over the ground state energy with an occupation of $\frac{1}{3} - \delta$. This Coulomb gap plays the same role as the gap between Landau subbands in the integral case. Thus, plateaus should be observed at around $\frac{1}{3}$ occupation as a function of B. The plateau at $\nu = \frac{2}{3}$ is easily understood from the electron-hole symmetry within a single Landau subband.

4.2.10 Ballistic Transport Through a sufficiently short path, e.g., in a GaAs-AlGaAs heterostructure with high mobility, electrons can be transported without being scattered. In old vacuum tubes, this was the case since there is ordinarily no scattering in a vacuum. In this case, electrons are accelerated up to a velocity limited by the accumulated space charge, as given by the Richardson formula. In the solid state, a similar situation can occur, especially in high-quality interfaces such as in GaAs-AlGaAs (see HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR). Figure 24 shows the variation in resistance as a function of the magnetic field when a current flows from lead k to lead l while the voltage is measured between leads m and n. The measured resistance is denoted by $R_{kl,mn}$.

Even if the current path is bent, there is no resistance due to the bend, so long as the generated channels consist of eigenmodes of the potential problem. However, the path might be narrowed around the bends, and consequently, there may be considerable blocking and mixing between channels. Or, there may also be a different disturbance at the corners, which must cause a resistance on the electrons curving around the corners. This additional resistance may be called the *bend resistance*. In the absence of a magnetic field, a straight path gives a lower resistance, whereas under a magnetic field paths are more natural with bends, and hence, the difference in resistance is reduced.

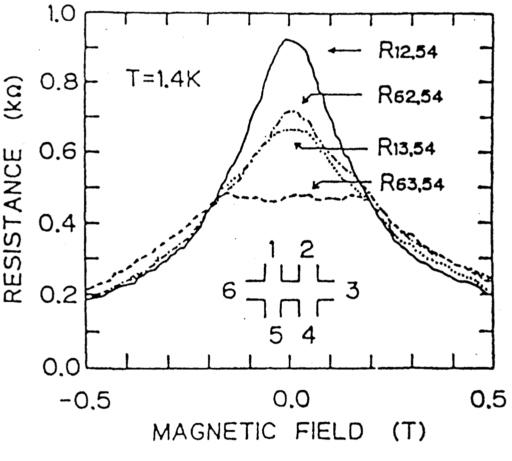

FIG. 24. Observed magnetoresistance $R_{kl,mn}$ on a sample with a current fed between leads k and l and a voltage measured between leads m and n. For a straight current path, the zero-field resistance is minimum. For a higher magnetic field, the resistance is almost independent of the path, since the current path tends to be curved (after Takagaki *et al.*, 1989).

4.2.11 Coulomb Blockade In the band theory of condensed matter physics, it is known that we can neglect most of the correlation energy between electrons, so long as their orbitals are extended sufficiently. The correlation energy is nothing but Coulomb repulsion energy. However, if electron orbitals are localized, this is not the case. They feel mutual repulsion. Throughout this section, we follow the line described in Averin and Likharev (1991).

In quantum dots, orbitals are limited by the boundary, so that Coulomb repulsion is not negligible in these dots. Let us consider a tunnel junction [Fig. 25(a)]. One electron tun-

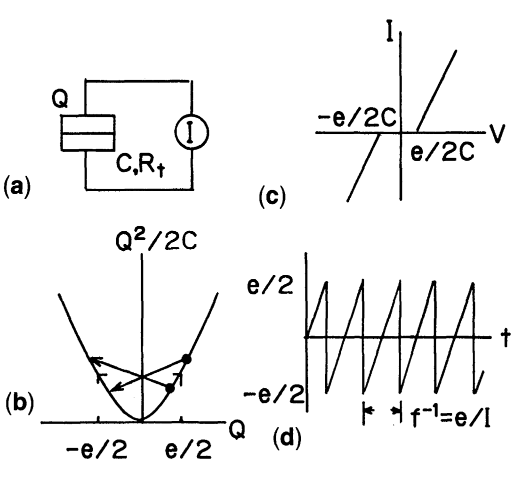

FIG. 25. (a) Diagram of the Coulomb blockade circuit. (b) Schematic interpretation of Coulomb blockade. (c) Expected I-V characteristic. (d) Single-electron tunneling oscillation with time.

nels into the opposite electrode. Then another electron may or may not tunnel in the same way because of the repulsive force. Figure 25(b) shows this concept. The parabolic curve is the electrostatic energy, $\mathscr{E} = Q^2/2C$, where Q is the charge of concern and C is the capacity of the junction. If there is a charge larger than $e/2$ and one electron tunnels out, then $\mathscr{E}$ decreases. Otherwise, $\mathscr{E}$ increases. Thus, even under an application of a voltage V where $-e/2C < V < e/2C$, the junction does not allow one electron to tunnel. Since the current I is given by dQ/dt, I is suppressed for the voltage mentioned above. This phenomenon is called *Coulomb blockade* [Fig. 25(c)].

In a circuit with a constant current I, Q is expected first to increase from 0 to $e/2$. Then Q must be suddenly reduced to $Q - e/2$ when it reaches $Q + e/2$, and starts to increase again. This is expressed by the equations $(Q_g - ne)^2/2(C + C_g)$ for the electrostatic potential energy and $V \propto (Q_g - ne)$ for the voltage between the capacitor and the tunnel junction. Thus, the charge Q oscillates as shown in Fig. 25(d). This is called *single-electron tunneling* (SET) oscillation.

Figure 26(a) shows a circuit consisting of a SET junction and an ordinary capacitor. From the discussion above, it is obvious that the electrostatic energy $\mathscr{E}$ varies as in Fig. 26(b) as a function of the number of tunneling electrons, and the inner potential drop V varies as $Q_g - ne$ when increasing $Q_g = C_gU$, where U is the external potential.

One example of experimental SET oscillation is shown in Fig. 27 (Lafarge *et al.*, 1991). A tiny tunnel junction is formed using Al, and a tiny electrometer was also formed on the same chip to detect the oscillating current I (Fig. 28).

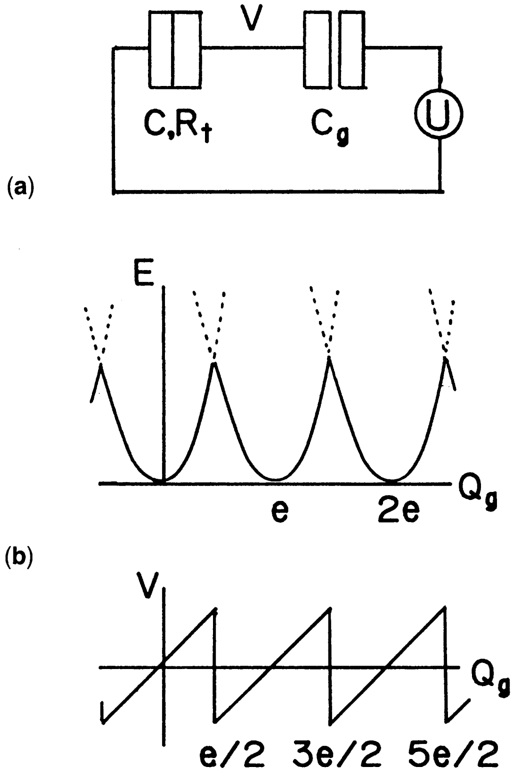

FIG. 26. (a) Circuit consisting of a conventional capacitor C and a single-electron tunnel junction C_g. (b) Electrostatic potential on the tunnel junction as a function of tunneling charge Q_g. The inner potential V varies in proportion to $Q_g - ne$, as shown below.

4.3 Magnetic Properties

4.3.1 Fine Particles Among mesoscopic systems, 0D systems such as quantum dots and clusters are peculiar in that it is hard to measure the dc transport properties, and the only detectable properties are ac effects, i.e., responses to high frequencies including light beams. Besides optical responses, magnetic properties are also detectable.

In 1962, Kubo discussed the thermodynamic properties of metallic fine particles such as their heat capacity and paramagnetic susceptibility. Since, in fine particles, the typical energy difference between neighboring levels can be larger than ambient temperature, thermal excitation is sometimes suppressed at an ambient temperature much more than that in the bulk, continuously excitable system. This reduces the heat capacity and increases the paramagnetic susceptibility. According to this theory, the number of spins in the particles affects the magnetic properties strongly, depending on whether it is even or odd. When it is odd, the particles must behave like magnetic particles with uncompensated spin (see MAGNETIC ORDERING IN SOLIDS).

Almost at the same time, Néel (1961) considered a similar problem. Let fine particles be composed of an odd number of antiferromagnetic sublattices. The magnetic moment on the surface sublattice rotates to parallel to the external magnetic field and the

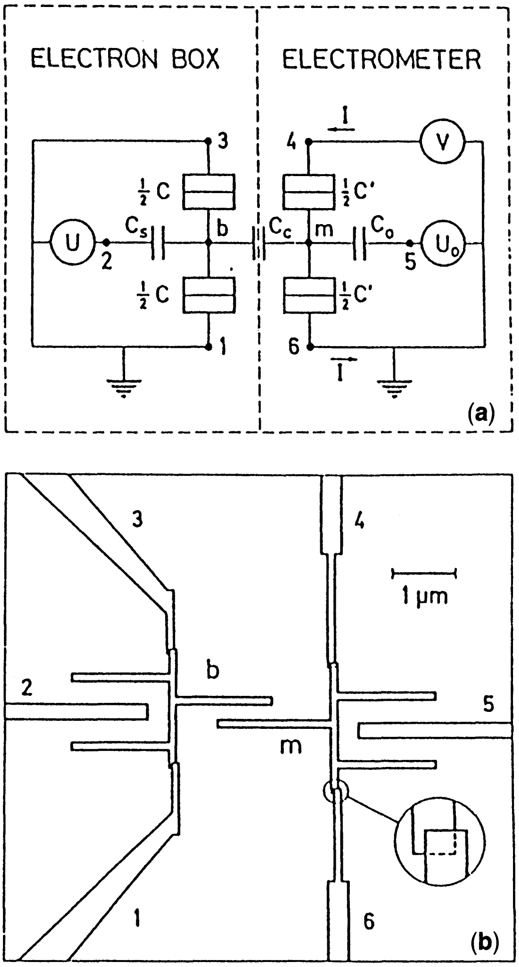

FIG. 27. Circuit used for SET oscillation experiment and its physical layout (after Lafarge *et al.*, 1991).

moments inside are sequentially affected by the neighboring spin orientation. In this manner, antiferromagnetic particles show higher magnetic susceptibility, because they behave like a spiral spin state with a remanent magnetic moment. He called this particle state *superantiferromagnetism.*

When particles have an uncompensated magnetic moment, e.g., in single-domain ferromagnetic particles, they look like particles with a giant magnetic moment. Since the particles are not ordered when they are in a nonmagnetic medium, the whole system does not ordinarily show ferromagnetism. If the particles are small enough, ~100 Å, then they behave as a paramagnetic medium under thermal agitation. This state is called *superparamagnetism.*

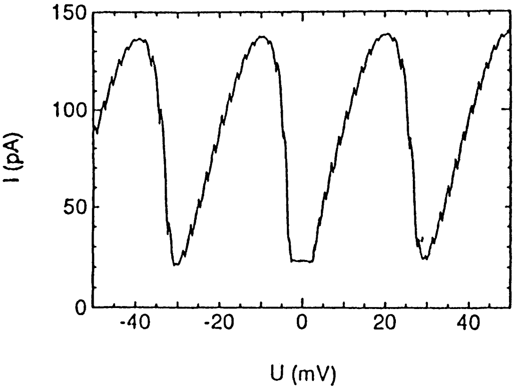

FIG. 28. Observed oscillatory current vs externally applied potential U (after Lafarge *et al.*, 1991).

It is well known that a γ ray is emitted without recoil in bulk radioactive materials, a phenomenon called the *Mössbauer effect.* However, such is not the case in a mesoscopically small-sized particle. The recoil energy must be compensated for by the motion of the particle as a whole. So, when it is sufficiently small and behaves without coupling to its environment, the recoil energy must be appreciable and manifest itself eventually as a shift of the emitted γ ray. This fact was pointed out by Murayama (1966a).

4.3.2 Magnetic Thin Films Two-dimensional magnetic systems show very interesting phenomena. The first example is the stripe domain in Permalloy thin films (see PLATES AND FILMS, MAGNETIC). The domain structure is schematically shown in Fig. 29 and was discovered by Spain (1963) and by Saito *et al.*

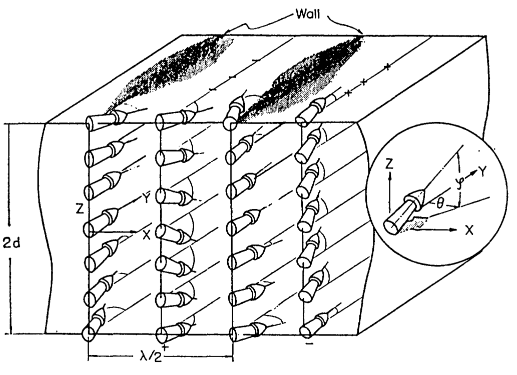

FIG. 29. Micromagnetic configuration of spins inside stripe domains (after Murayama, 1966b).

(1964), independently. Theoretically, Kaczér *et al.* (1963) proposed the simplest model, where the spin tilting from the plane varies sinusoidally as a function of the coordinate in the direction perpendicular to the spin. Murayama (1966b) showed there is a *micromagnetic* state with a lower energy, as shown in Fig. 29, where spins form a circulating Landau-type closure domain in the cross section of the film perpendicular to the mean direction of spins. In this domain structure, the thickness and stripe width are on a comparable scale, a few times 1000 Å.

In 1967, Bobeck discovered that a serpentine stripe domain in the absence of a magnetic field changes into magnetic cylindrical (bubble) domains in the presence of a perpendicular dc field in some orthoferrites. These bubbles can be driven by applying an ac magnetic field, from beneath one magnetized segment on the surface to beneath another. The domains have been developed as a BORAM (block-oriented random access memory) just like a CCD (charge-coupled device) in an MOS structure.

4.3.3 Superlattice of Magnetic Thin Films Magnetic thin films sometimes show unusual effects when they are formed in an SL. Figure 30 shows *giant magnetoresistance* on a [(Fe 30 Å)/(Cr 9 Å)]$_{60}$ SL at 4.2 K (Baibich *et al.*, 1988). Ordinary bulk material has a magnetoresistance of at most several percent, whereas in this SL it amounts to several tens of percent. From the application point of view, this effect is worth development.

4.4 Properties of Macroscopic Quantum States

Three typical examples of macroscopic quantum states are lasers, superfluids, and superconductors. Here, we are most interested in superconductivity (SC) (see SUPERCONDUCTIVITY, LOW-TEMPERATURE).

It is well established that the SC state can be described in terms of a macrowave, which is a coherent wave extending from one edge of the sample to the other. This is almost always the case even if the sample is macro-

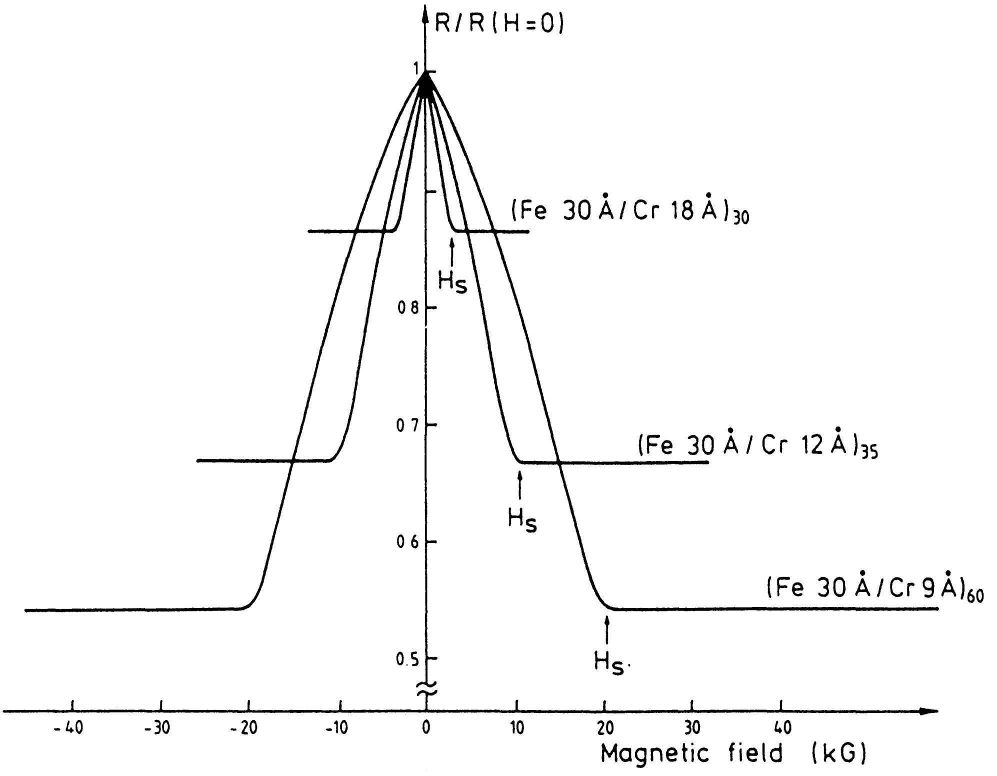

FIG. 30. Observed giant magnetoresistance on superstructured samples with Fe and Cr (after Baibich *et al.*, 1988).

scopically large, so long as it is superconducting. This means that the large sample behaves just like a tiny microscopic system from the point of view of its physical properties. The mesoscopic systems we are discussing are limited to those where the physical properties are between the microscopic and macroscopic scales. In SC, this kind of degraded coherence can occur, e.g., when the system is thin enough.

An ideal SC is made in bulk samples where there is no decay in the current between two leads or in a persistent ring current. This kind of current can flow persistently only without damping. However, there is always a slight damping mechanism even in the SC state. In high-T_c superconductors, SC is specified by 2D features. In 2D systems, it is known that thermally excited fluxon-antifluxon pairs interact with the current fed to measure the voltage drop through the sample. Thus, the *I-V* characteristic is subject to a power law, according to the Kosterlitz-Thouless (KT) (1973) modes intrinsic to 2D systems. This means that a voltage drop of exactly zero is expected only for a vanishing measuring current.

Besides the damping owing to the KT mechanism, superconducting diamagnetism caused by a Meissner current can change as a result of thermal excitation beyond barriers that separate a potential minimum with N from neighboring minima with $N \pm 1$ fluxons (Fig. 31). The case of persistent current through a ring circuit is schematically depicted in the figure. This quantized state of fluxons is known as the London quantization. This quantization is nothing other than the case for the Aharonov–Bohm phase equal to 2π, i.e., $2\pi\Phi/\Phi_0^* = 2n\pi$, or $\Phi = n\Phi_0^*$ with $\Phi_0^* = h/2e$ for Cooper pairs. This is because the macrowave must continue to itself without any phase shift (mod 2π).

4.4.1 Superconducting Wire Let us consider a 1D wire system. Figure 32 shows the variation in the phase of the macrowave along the wire when a bias current I is applied, according to the fact that the energy $E(\phi)$ is given by $E_J \cos\phi - I\Phi$. The terms ϕ and Φ are related to each other by the gauge $\phi = 2\pi\Phi/\Phi_0^*$. The variable E_J is known as the energy stored in a Josephson junction. The horizontal axis of the figure is phase ϕ and incidentally corresponds to the coordinate along the wire, since the phase is continuously rounded in one sense through the wire in regular, homogeneous states of the sample. When one fluxon jumps from one potential minimum to the neighboring minimum thermally or via resonant tunneling, the total winding number through the wire decreases. This means that a fluxon crosses the wire perpendicularly.

Thus, in wires so thin that fluxons can cross anywhere (the width $w \lesssim \lambda_L$, the London penetration length), the wires always show a voltage drop or, in other words, are resistive. Ordinarily, the resistance depends on the temperature, but, at low enough temperatures, thermally activated phase changes can no longer occur so that only tunneling takes place. The crossover is at a temperature specific to the sample (Fig. 33: Giordano, 1988). The theoretical justification of the crossover

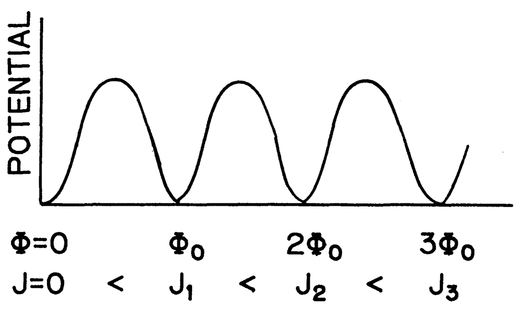

FIG. 31. Schematic interpretation of potential variation on a superconductive persistent-current-carrying ring circuit. The horizontal axis denotes a phase corresponding to a hypothetical superconducting state with a fractional number of fluxons. The fluxon-carrying state is equal to a state with a persistent current.

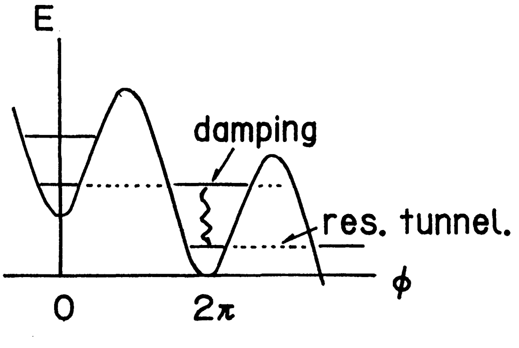

FIG. 32. "Washboard" potential of a current-biased superconducting single wire vs phase of the macrowave. Phase slippage occurs via thermal excitation and quantum tunneling.

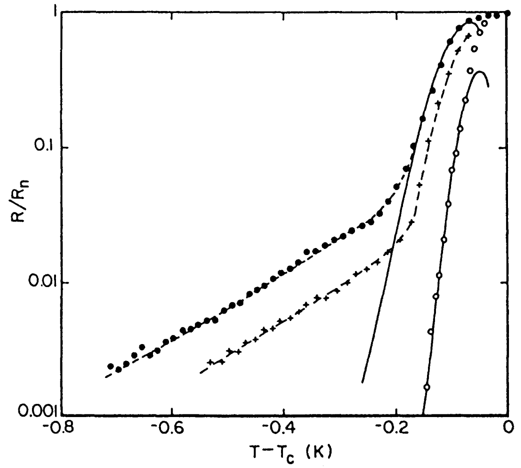

FIG. 33. Temperature vs dc resistance of thin superconducting wires with diameters 410 Å (•), 505 Å (+), and 720 Å (○) (after Giordano, 1988).

from thermal activation to tunneling was given by Saito and Murayama (1989).

It is also interesting to observe the same phenonemon in a superconductor as in an SET in its normal state. This time it is not the tunneling of *single normal electrons,* but *quasiparticles* in the superconductor. Cooper pairs may show a similar phenomenon when the Josephson junction has a small enough area.

4.4.2 Andreev Reflection Superconductivity is known to result from pairs of electrons that behave like bosons. According to the BCS theory, the most probable pair is an electron with momentum $\hbar\mathbf{k}$ and a hole with $\hbar(-\mathbf{k})$. Figure 34(a) shows schematically the case where an electron incident from the normal-metal side has an energy below the SC gap. Since the energy is not high enough for the electron to become a quasiparticle in the superconductor, it must be reflected but as a hole with an energy $\epsilon_h = 2\epsilon_F - \epsilon_{el}$. This kind of reflection was first discussed by Andreev (1964). Note that the generated hole has a momentum of $-\mathbf{k}$ and, hence, traces the same path as the electron but in the opposite direction. This is shown in Fig. 34(c), in comparison with the ordinary normal reflector [Fig. 34(b)].

In an experiment on a mesoscopic scale, Nishino (1990) observed a geometrical resonance between an Nb injector and an Nb reflector through Si. If the distance d between them is such that $\sqrt{\Delta^2 + (nhv_F/2d)^2} \sim eV$ (V

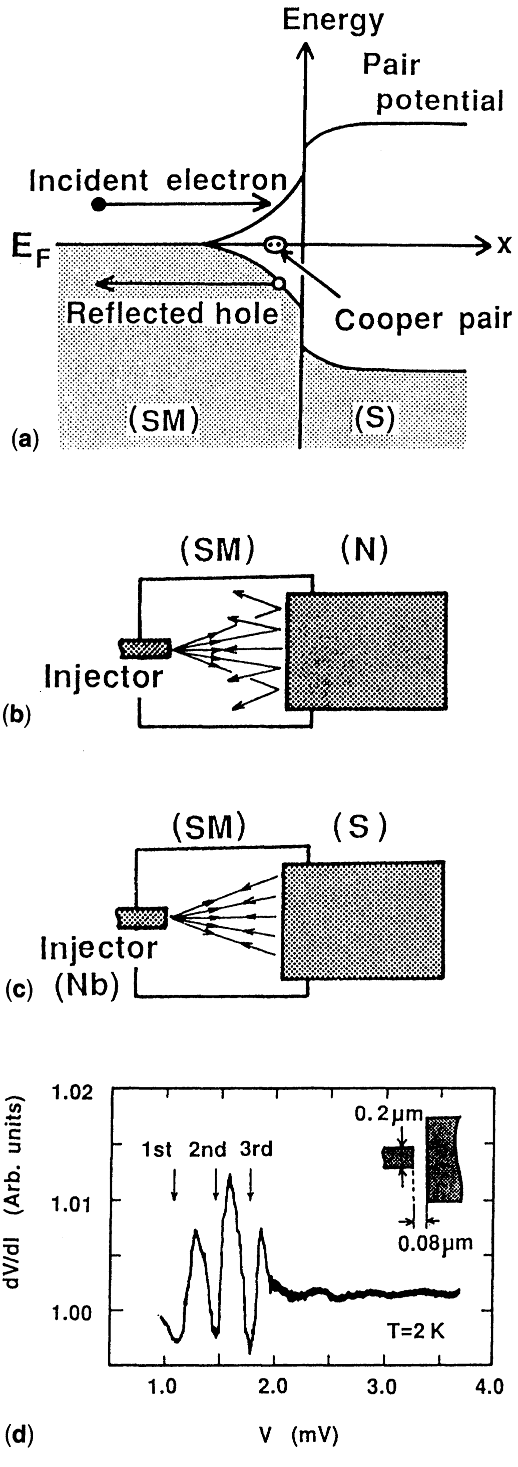

FIG. 34. (a) Conceptual explanation of the Andreev reflection. Expected paths on reflection for specular scattering (b) in the normal state and (c) for Andreev reflection. (d) Observed differential resistance vs voltage showing a few geometrical resonances (after Nishino *et al.*, 1990).

is applied voltage) with an integer n, a resistance minimum appears. These data are plotted in Fig. 34(d), which clearly shows first-, second-, and third-order geometric resonances.

5. FUTURE PROSPECTS

5.1 Physics

All the phenomena described so far are related to physics. Future prospects with respect to applications are given below. The most interesting matter in mesoscopic physics is that an appreciable degree of coherence is lost there. Regarding microscopic systems such as a hydrogen atom in its excited state, it may decay to the ground state with a specific decay time. The ground state itself can be excited by interacting with its environment such as photons, phonons, and other particles. These all reduce coherence. In fact, hydrogen atom beams do not show an interference pattern in a bad vacuum where they suffer from frequent scatterings. Thus, even an atomic-level microscopic entity may not be called "microscopic" in the exact sense in a certain environment.

On the other hand, a macrowave within superconductors may show long-lasting macroscopic coherence, even if it is macroscopically large. However, the environment always has pair-breaking (dephasing) mechanisms causing SC to be broken. There, the collectiveness of the Bose condensate and the stiffness of the macrophase prevent such dephasing. Whether we should consider a system to be microscopic or macroscopic depends on the time scale over which we observe it. The time scale typical for a mesoscopic system ranges from a very tiny scale to a considerably large one, depending what we are discussing. The only clear point is that the temporal and spatial variations are respectively specific to the system; in this sense, it becomes rather difficult to expect *a priori* the validity of ergodicity in such mesoscopic systems (Stone, 1987). The specificity in the sample of concern plays an important role in mesoscopic systems (e.g., *magnetofingerprint* in Section 4.2.5) and, accordingly, the time-averaged physical properties can deviate from the average over an ensemble of similar replicas.

Another interesting matter, seen in the examples listed so far, is that many quantities are observed in terms of universal constants, such as resistance in $\pi\hbar/e^2$, current in e times frequency f, and voltage $fh/2e$ in the ac Josephson effect. Actually, it was possible to determine the fine-structure constant, $\alpha = (\mu_0 c/2)(e^2/2h)$, where μ_0 is the permeability of the vacuum, in a mesoscopic experiment, e.g., the integral QHE.

The new physics in mesoscopic systems is Anderson localization owing to disorder. Localized states stand at the opposite position to band states. In tiny systems, localized states often appear, especially at low temperatures. It is known that low-dimensional systems ($d \leq 2$) localize electrons easily.

When electrons are localized, not in the sense of Anderson but according to the mesoscopic boundary conditions, the correlation energy between the two outweighs the difference in one-particle energy. This effect appears as the Coulomb blockade. When the correlation is large, the condensed-matter physical world can be very different from the conventional one based on band energy states.

Localization is likely to occur for a narrow band. A narrow band has a larger DOS. In condensed-matter physics, the DOS is an important concept. For a larger DOS, every physical property is enhanced. For SC to occur, a large DOS is necessary. This is the case also in a high-T_c SC (Murayama and Nakajima, 1991). Sometimes, SC competes against localization; localized electrons can obviously never show SC (Haviland *et al.*, 1989). Typical experimental evidence of competition between localization and SC in Nb thin films is shown in Fig. 35 (Nishida *et al.*, 1991).

Again, it should be emphasized that for elastic scattering any electronic state only undergoes a phase shift, which has no relation to damping or dephasing. It only shifts its original state to another (unitarily transformed) state. Physically, the difference is trivial. On the other hand, for inelastic scattering, the energy of the electron is lost to other degrees of freedom, which means damping on the electron. The degrees of freedom work frequently as a heat reservoir. This causes dephasing, and the coherence cannot survive this scattering. This degradation in coherence is a crucial concept in mesoscopic physics.

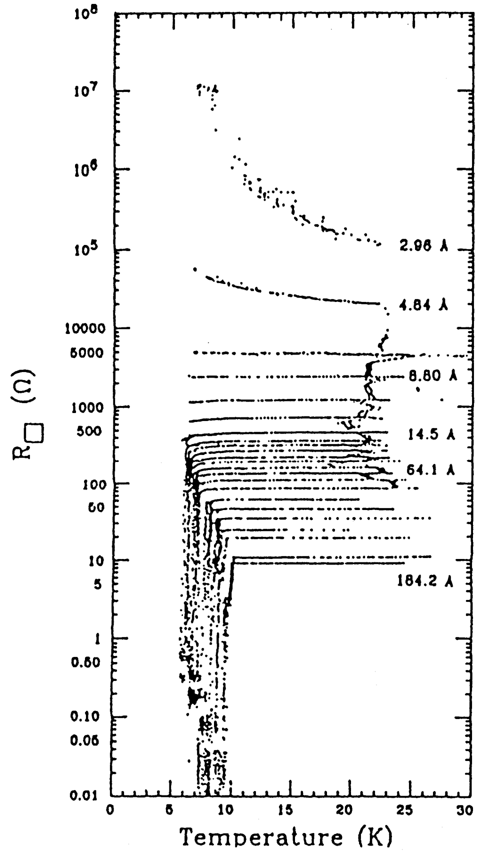

FIG. 35. Observed variation in resistance vs temperature as a function of sample thickness. For thinner samples, localization occurs, and for thicker ones, they show a transition to SC (after Nishida *et al.*, 1991).

5.2 Applications

The use of 2D structures is already conventional in semiconductor devices such as semiconductor lasers, light-emitting diodes, photo diodes/transistors, solar cells, and photoconducting membranes in image pickup tubes. In transport devices, the 2D confinement technology is applied to MOS transistors, high-electron mobility transistors (HEMT's), and RTD. All these devices aim at confining carriers, which works well for radiative recombination in optical applications. That is, radiative recombination must be over before irradiative recombination mechanisms work in order to improve the efficiency of luminescence. For semiconductor lasers, this confinement of the states is so effective in double heterostructures that continuous oscillation has been achieved at RT. Even in this device, the other two degrees of freedom in the active layer are still free. These degrees of freedom can also be frozen by implementing disorder within the layer so as to cause Anderson localization (Murayama *et al.*, 1988).

One of the most remarkable recent achievements in optical-device applications of mesoscopic systems is seen in microcavity lasers. There, an electromagnetic field is confined inside a mesoscopically small cavity, where the boundary condition affects the quantum mechanical features of the field. Even the vacuum state is controlled in such a microcavity. This topic is discussed separately in the article QUANTUM OPTICS. Several related studies are found in Ezawa and Murayama (1993).

The idea of photon equivalents of various electronic states in solids was proposed by Yablonovitch (1987). He described how, first of all, a photonic band can be formed by preparing a periodic structure composed of different electromagnetic potential layers, i.e., of different refractive indexes, for photons. Theoretically, forming such a superstructure yields all kinds of specific electronic states with photonic equivalents: the band energy states (*E-k* curves) with appropriate band gaps as well as the Anderson localization, the quantum well, the superlattice and so. The spontaneous emission rate in bulk materials is now controlled by giving the appropriate boundary conditions to the electromagnetic field. This is also a quantum optical issue.

For magnetic applications, we began with magnetic thin-film memory devices. However, this type of memory device was short-lived and was soon replaced by the magnetic core memory. Then the bubble domain memory appeared, which was actually utilized for a short while. The recent discovery of giant magnetoresistance looks promising for application, since magnetic recording heads are becoming smaller and smaller down to a thin-film structure. This is also on the mesoscopic scale. In magnetic applications, metallic superlattices are formed nonepitaxially, in a different way from in semiconductor mesoscopic systems. Nonepitaxial growth, if it produces good devices, is highly preferable for industrial applications.

For device applications, it is crucial that the effects described so far work well at RT, or at least at 77 K. The above-mentioned optical devices are used at RT. SC devices such as Josephson devices and SQUIDs will be limited to low temperatures, even if high-T_c SC's can be developed. In order to apply physical phenomena to actual devices, the energy difference between adjacent energy levels must exceed the ambient thermal energy. This means that the higher the difference, the easier the application. In this respect, a large difference in the correlation between two electrons in the phenomenon of the Coulomb blockade is highly useful. For example, a 10-Å distance means an energy difference of 1.5 eV, unless any screening exists.

Localized orbitals cause a larger correlation energy. In the future, devices on the other side of the edge opposite to the conventional band energy states may be developed. In a high-T_c SC substance like $Bi_2Sr_2Ca_{1-x}Y_xCu_2O_{8+\delta}$, holes are known to be in a localized state for $x \gtrsim 0.4$ at RT. If we inject excess carriers into this substance, they are expected to delocalize holes and change it to an SC phase. This results from the easily localizable 2D layer of CuO_2 and the fact that the band responsible for SC is sufficiently narrow. Thus, on one hand, the narrowness of bands allows many interesting phenomena due to the high DOS, and, on the other hand, localization, charge-density waves, and so on compete against the SC. This competition must be investigated in more detail for applications of mesoscopic physics to materialize.

As the last example of mesoscopically sized devices, we may mention atomic-scaled devices, which can be produced only by scanning tunneling microscope (STM) based manipulation of atoms on a certain substrate. This technology has very recently been introduced and is promising, in principle; however, it involves a number of basic problems to be resolved before becoming practical.

ACKNOWLEDGMENTS

The author is very grateful to Dr. K. Nakazato, Dr. B. J. van Wees, Dr. M. Büttiker, Dr. A. Tonomura, Dr. Y. A. Ono, and Professor T. Ando for their valuable suggestions in preparing this article. Above all, encouragement from Professor M. Tanaka in undertaking this job is highly appreciated.

GLOSSARY

Coherence: A necessary condition for two waves to interfere. The property is seen in an interferometry experiment. Coherence only occurs for two split waves that originally belonged to a single state of a fermion. Two *incoherent* waves cannot interfere. Under inelastic scattering, two waves partially lose their energy, which causes originally coherent states to degrade into incoherent ones. This process is called *decoherence*.

Localization: A concept opposite to the extended state described by the conventional band theory. In disordered solids, the Bloch theorem does not apply, and electronic states are localized on a scale much larger than the atomic size but smaller than the macroscopic sample size. Often it is called Anderson localization after the first author of the investigation.

Macrowave: When a single state is occupied by many identical waves, the ensemble of the waves acquires a definite phase, since the number-phase uncertainty allows it. This phenomenon is known to be a macroscopic quantum state. There are three typical examples: one is superconductivity, where fermions are condensed into a single ground state (Bose–Einstein condensation) by forming bosonlike Cooper pairs; the other two are superfluidity and lasers. Here, an enormous number of bosons occupy the ground state. All these states are described by macrowaves.

Quantum Well: When two materials with different electron affinities χ are epitaxially grown in series as a barrier material with a smaller χ_B, a well material with a larger χ_W, and then again a barrier material, the intermediate material forms a potential well where electrons are confined. For holes, the quantity χ plus the band gap E_g plays the same role as χ for electron confinement. A quantum well is a 2D quantized system.

Quantum Dot: A 0D quantized system, where all the degrees of freedom are lost by confining electrons into a small particle.

Quantum Wire: A 1D quantized system, where two degrees of freedom are lost by confining electrons into a thin wire.

Superlattice: Periodically repeated quantum wells form a man-made periodic system, which has a miniband structure and minigaps, following the Bloch theorem. In order to have a significant coupling between quantum wells, the barrier materials in between must be thin enough for electrons to tunnel.

Wave Packet: An analog of a classical particle with quantum mechanical properties. It is not a rigid but a soft particle in the sense that it is a superposition of eigenstates and obeys a time-dependent Schrödinger equation. The inverse of half its extension in k space is called the *coherence length*.

Works Cited

Abrahams, E., Anderson, P. W., Licciardello, D. C., Ramakrishnan, T. V. (1979), *Phys. Rev. Lett.* **42**, 673–676.

Al'tshuler, B. L., Aronov, A. G., Spivak, B. Z. (1981), *JETP Lett.* **33**, 94–97.

Anderson, P. W. (1958), *Phys. Rev.* **109**, 1492–1505.

Ando, T., Fowler, A. B., Stern, F. (1982), *Rev. Mod. Phys.* **54**, 437–672.

Andreev, A. F. (1964), *J. Exptl. Theoret. Phys.* **46**, 1823–1828.

Averin, D. V., Likharev, K. K. (1991), in: B. L. Al'tshuler, P. A. Lee, R. A. Webb (Eds.), *Mesoscopic Phenomena in Solids*, Amsterdam: North-Holland, pp. 173–272.

Baibich, M. N., Broto, J. M., Fert, A., Nguyen Van Dau, F., Petroff, F. (1988), *Phys. Rev. Lett.* **61**, 2472–2475.

Bergmann, G. (1983), *Phys. Rev. B* **28**, 2914–2920.

Bobeck, A. H. (1967), *Bell Syst. Tech. J.* **46**, 1901–1925.

Büttiker, M., Imry, Y., Landauer, R., Pinhas, S. (1985), *Phys. Rev. B* **31**, 6207–6215.

Canham, L. T. (1990), *Appl. Phys. Lett.* **57**, 1046–1048.

Giordano, N. (1988), *Phys. Rev. Lett.* **61**, 2137–2140.

Haviland, D. B., Liu, Y., Goldman, A. M. (1989), *Phys. Rev. Lett.* **62**, 2180–2183.

Hikami, S., Larkin, A. I., Nagaoka, Y. (1980), *Prog. Theor. Phys.* **63**, 707–710.

Imry, Y. (1986), in: G. Grinstein, G. Mazenko (Eds.), *Directions in Condensed Matter Physics*, Singapore: World Scientific, pp. 101–163.

Kaczér, J., Zelený, M., Sůda, P. (1963), *Czech. J. Phys.* **13**, 579–585.

Katayama, Y., Yoshida, I., Kotera, N., Komatsubara, K. F. (1972), *Appl. Phys. Lett.* **20**, 31–33.

Kawaguchi, Y., Kitahara, H., Kawaji, S. (1978), *Surf. Sci.* **73**, 520–527.

Kosterlitz, J. M., Thouless, D. J. (1973), *J. Phys. C: Solid State Phys.* **6**, 1181–1203.

Kubo, R. (1962), *J. Phys. Soc. Jpn.* **17**, 975–986.

Lafarge, P., Pothier, H., Williams, E. R., Esteve, D., Urbina, C., Devoret, M. H. (1991), *Z. Phys. B Cond. Matter* **85**, 327–332.

Landauer, R. (1957), *IBM J. Res. Dev.* **1**, 223–231.

Laughlin, R. B. (1983), *Phys. Rev. Lett.* **50**, 1395–1398.

Lee, P. A., Stone, A. D., Fukuyama, H. (1987), *Phys. Rev. B* **35**, 1039–1070.

Machida, S., Namiki, M. (1980), *Prog. Theor. Phys.* **63**, 1457–1473, 1833–1847.

Mishima, T., Kasai, J., Morioka, M., Sawada, T., Murayama, Y., Katayama, Y., Shiraki, Y. (1986), in: M. Fujimoto (Ed.), *Gallium Arsenide and Related Compounds*, Institute of Physics Conference Proceedings No. 79, Bristol: Institute of Physics, Chap. 8, pp. 445–450.

Mott, N. F. (1974), *Metal-Insulator Transition*, London: Taylor & Francis.

Murayama, Y. (1966a), *Phys. Lett.* **23**, 332–334.

Murayama, Y. (1966b), *J. Phys. Soc. Jpn.* **11**, 2253–2266.

Murayama, Y., Kamigaki, Y., Yamada, E. (1972), in: Proceedings of the 3rd Conference on Solid State Devices, Tokyo, 1971, *Suppl. Oyo Buturi*, **41**, 133–140.

Murayama, Y. (1984), *Buturi*, **39**, 485 (in Japanese).

Murayama, Y. (1986), *Phys. Rev. B* **34**, 2500–2507.

Murayama, Y., Takeda, Y., Nakamura, M., Shiraki, Y., Katayama, Y., Chinone, N. (1988), "Semiconductor Laser," U. S. Patent No. 4,759,030 (19 July 1988).

Murayama, Y. (1990), *Found. Phys. Lett.* **3**, 103–127.

Murayama, Y., Nakajima, S. (1991), *J. Phys. Soc. Jpn.* **60**, 4265–4279.

Murayama, Y. (1992), *Surf. Sci.* **263**, 604–608.

Nakayama, M., Tanaka, I., Nishimura, H., Kawashima, K., Fujiwara, K. (1991), *Phys. Rev. B* **44**, 5935–5938.

Néel, L. (1961), *Comptes Rendus*, Séance du 26 Juin, 4075–4080; Séance 3 Juillet, 9–12; Séance 10 Juillet, 203–208.

Nelson, D. F., Miller, R. C., Tu, C. W., Sputz, S. K. (1987), *Phys. Rev. B* **36**, 8063–8070.

Nishida, N., Okuma, S., Asamitsu, A. (1991), *Physica* **B169**, 487–488.

Nishino, T., Hatano, M., Hasegawa, H., Murai, F., Kure, T., Yamada, E., Kawabe, U. (1990), in: S. Kobayashi *et al.* (Eds.), *Proc. 3rd Int. Symp. Foundations of Quantum Mechanics*, Tokyo: The Physical Society of Japan, pp. 263–269.

Saito, N., Fujiwara, H., Sugita, Y. (1964), *J. Phys. Soc. Jpn.* **19**, 421–424, 1116–1125.

Saito, S., Murayama, Y. (1989), *Physica* **139**, 85–90.

Sharvin, D. Yu., Sharvin, Yu. V. (1981), *JETP Lett.* **34**, 272–275.

Sollner, T. C. L. G., Goodhue, W. D., Tannenwald, P. E., Parker, C. D., Peck, D. D. (1983), *Appl. Phys. Lett.* **43**, 588–590.

Spain, R. J. (1963), *Appl. Phys. Lett.* **3**, 208–209.

Stone, A. D. (1987), in: M. Namiki *et al.* (Eds.) *Proc. 2nd Int. Symp. Foundations of Quantum Mechanics,* Tokyo: The Physical Society of Japan, pp. 207–217.

Störmer, H. L., Chang, A., Tsui, D. C., Hwang, J. C. M., Gossard, A. C., Wiegmann, W. (1983), *Phys. Rev. Lett.* **50**, 1953–1956.

Takagaki, Y., Wataka, F., Takaoka, S., Gamo, K., Murase, K., Namba, S. (1989), *Jpn. J. Appl. Phys.* **28**, 2188–2192.

Thouless, D. J. (1977), *Phys. Rev. Lett.* **39**, 1167–1169.

Tonomura, A., Endo, J., Matsuda, T., Kawasaki, T. (1989), *Am. J. Phys.* **57**, 117–120.

Tsu, R., Esaki, L. (1973), *Appl. Phys. Lett.* **22**, 562–564.

Tsuchiya, Y., Inuzuka, E., Kurono, T., Hosoda, M. (1985), in: B. L. Morgan (Ed.), *Advances in Electronics and Electron Physics,* Vol. 64A, New York: Academic, pp. 21–31.

van Kampen, N. G. (1976), in: L. Pál, P. Szépfalus, (Eds.), *Statistical Physics, Proc. of Int. Conf.,* 25–29 August 1975, Budapest, Amsterdam: North-Holland.

van Wees, B. J., van Houten, H., Beenakker, C. W. J., Williamson, J. G., Kouwenhoven, L. P., van der Marel, D., Foxon, C. T. (1988), *Phys. Rev. Lett.* **60**, 848–850.

van Wees, B. J., Kouwenhoven, L. P., Harmans, C. J. P. M., Williamson, J. G., Timmering, C. E., Broekaart, M. E. I., Foxon, C. T., Harris, J. J. (1989), *Phys. Rev. Lett.* **62**, 2523–2526.

von Klitzing, K., Dorda, G., Pepper, M. (1980), *Phys. Rev. Lett.* **45**, 494–497.

Washburn, S., Umbach, C. P., Laibowitz, R. B., Webb, R. A. (1985) *Phys. Rev. B* **32**, 4789–4792.

Webb, R. A., Washburn, S., Umbach, C. P., Laibowitz, R. B. (1985), *Phys. Rev. Lett.* **54**, 2696–2699.

Webb, R. A., Washburn, S., Benoit, A. D., Umbach, C. P., Laibowitz, R. B. (1987), in: M. Namiki *et al.* (Eds.), *Proc. 2nd Int. Symp. Foundations of Quantum Mechanics,* Tokyo: The Physical Society of Japan, pp. 193–206.

Wigner, E. P. (1963), *Am. J. Phys.* **31**, 6–15.

Yablonovitch, E. (1987), *Phys. Rev. Lett.* **58**, 2059–2062.

Further Reading

Al'tshuler, B. L., Lee, P. A., Webb, R. A. (Eds.) (1991), *Mesoscopic Phenomena in Solids,* Amsterdam: North-Holland.

Bergmann, G. (1984), *Phys. Rep.* **107**, 1–58.

Ezawa, H., Murayama, Y. (Eds.) (1993), *Quantum Control and Measurement,* Amsterdam: North-Holland.

Grinstein, G., Mazenko, G. (Eds.) (1986), *Directions in Condensed Matter Physics,* Singapore: World Scientific.

James, T. W. (Ed.) (1985), *Characterization and Behavior of Materials with Submicron Dimension,* Singapore: World Scientific.

Kamefuchi, S., Ezawa, H., Murayama, Y., Namiki, M., Nomura, S., Ohnuki, Y., Yajima, T. (Eds.) (1984), *Proc. Int. Symp. Foundations of Quantum Mechanics,* Tokyo: The Physical Society of Japan.

Kobayashi, S., Ezawa, H., Murayama, Y., Nomura, S. (Eds.) (1990), *Proc. 3rd Int. Symp. Foundations of Quantum Mechanics,* Tokyo: The Physical Society of Japan.

Nagaoka, Y. (Ed.) (1985), in: *Anderson Localization, Prog. Theor. Phys.* Suppl. No. 84, Kyoto: Research Institute for Fundamental Physics and The Physical Society of Japan.

Namiki, M., Ohnuki, Y., Murayama, Y., Nomura, S. (Eds.) (1987), *Proc. 2nd Int. Symp. Foundations of Quantum Mechanics,* Tokyo: The Physical Society of Japan.

Ovchinnikov, A. A., Ukrainskii, I. I. (Eds.) (1991), *Electron Correlation Effects in Low-Dimensional Conductors and Superconductors,* Berlin: Springer.

Prange, R. E., Girvin, S. M. (Eds.) (1987), *The Quantum Hall Effect,* New York: Springer.

Reed, M. A., Kirk, W. P. (Eds.) (1989), *Proc. Int. Symp. Nanostructure Physics and Fabrication,* New York: Academic Press.

Thouless, D. J. (1978), *Les Houches, Ecole d'Eté de Physique Théorique, Session XXXI,* Amsterdam: North-Holland.

METAL-INSULATOR TRANSITIONS

T. Maurice Rice, *Eidgenössische Technische Hochschule, Zürich, Switzerland*

Introduction 113
1. **Metal-Insulator Transitions in Ordered Systems** 114
1.1 Band-Crossing Transitions 114
1.2 Mott Transitions 115
1.3 Metal-Insulator Transitions due to Lattice Distortion 117
1.4 Metal-Insulator Transitions due to Charge Ordering 117
1.5 Metal-Insulator Transitions in Lower-Dimensional Systems 118
2. **Metal-Insulator Transitions in Disordered Systems** 118
2.1 Uncompensated Doped Semiconductors 118
2.2 Compensated Doped Semiconductors 123
2.3 Doping of Mott Insulators 123
2.4 Metal-Insulator Transitions in Expanded Liquid Metals 123
3. **Superconductor-Insulator Transitions** 124
3.1 The $BaBi_{1-x}Pb_xO_3$ and $Ba_{1-y}K_yBiO_3$ Systems 124
3.2 High-T_c Superconductors 124
4. **Applications** 125
Glossary 125
Works Cited 126
Further Reading 127

INTRODUCTION

Most solids can be classified simply as metallic or insulating depending on the size of their electrical conductivity. This difference becomes more pronounced as the temperature is lowered, and in this chapter a metal will be defined as a material whose conductivity extrapolates to a finite value at zero temperature and an insulator as a material whose conductivity vanishes when extrapolated to zero temperature. In a small number of cases metal-insulator transitions can be triggered by changes in temperature or by the application of pressure. More generally, an insulator can be made metallic by changing the stoichiometry or by doping, i.e., by adding a small concentration of foreign elements. Such metal-insulator transitions lead to dramatic changes of many orders of magnitude in the electrical conductivity.

The explanation of the difference between metallic and insulating solids was one of the earliest successes of the quantum theory of solids. Bloch energy bands, which are the energy levels of a single electron moving in the periodic potential of the crystal, can be exactly filled in a stoichiometric material with an even number of electrons per unit cell. This leads to a finite energy gap between the ground state and all excited states. Such a material will have insulating behavior. On the other hand, if the Bloch bands are partially filled, there is no energy gap and a current-carrying state can have an energy arbitrarily close to the ground state. This can happen with an even number of electrons per unit cell if Bloch bands overlap at the Fermi (i.e., the highest occupied) energy. It must be the case if the number of electrons per unit cell is odd, and then the single-electron theory predicts metallic behavior.

It was soon recognized that exceptions to this rule occurred. In particular de Boer and Verwey (1937) raised the case of NiO, which is an insulator despite the fact that the Ni-derived 3*d* bands are partially filled. The explanation was quickly provided in a remark by Peierls, who pointed out that the Coulomb interaction between electrons could destroy metallic behavior. This idea was extended and formalized by Mott (1949, 1956, 1961). In Mott insulators, electrons are localized to individual atoms by the Coulomb interaction, and a finite energy is required to move an electron from one atom to another. The localized electrons have magnetic moments from spin, and

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

possibly orbital, degrees of freedom, and these moments undergo an ordering transition as the temperature is lowered. This magnetic behavior distinguishes Mott insulators from band insulators. There is no (or at least only a very small) energy gap in the spin degrees of freedom, but there is a large energy gap for the charge degrees of freedom, and it is this charge gap that is responsible for the insulating behavior.

It is a curious, but poorly understood, fact that a transition between metallic and insulating behavior is rarely seen for stoichiometric materials even under the application of pressure. This is true although on general grounds at high enough pressure the broadening of the energy bands will inevitably lead to overlapping of bands and so to metallic behavior for all materials. However, the energy gaps in Mott insulators are large, and consequently the pressures required to reach the metallic state are large.

The nature of the metal-insulator transition was discussed in early articles by Landau and Zel'dovich (1943) and Mott (1949), who stressed the important difference in the effective Coulomb interaction between charge carriers in insulators, where it remains long range, and in metals, where it is short range because of screening by the other carriers. Metallic screening, however, requires a finite density of carriers, and so they argued that a continuous metal-to-insulator transition with arbitrarily small values of the conductivity at zero temperature is impossible, at least in an ordered material.

Changes in the stoichiometry through doping immediately alter the conductivity of an insulator and can lead to a transition to metallic behavior. The randomness that is an inevitable by-product of the doping procedure changes the character of the transition, so that it is necessary to distinguish between the relatively rare cases of metal-insulator transitions in ordered stoichiometric crystals and the much more common cases of transitions caused by doping away from stoichiometry with the inevitable randomness that is introduced in the doping process.

1. METAL-INSULATOR TRANSITIONS IN ORDERED SYSTEMS

1.1 Band-Crossing Transitions

Conceptually the simplest kind of metal-insulator transition is that due to band crossing leading to an overlap at the Fermi energy. The width of a Bloch band in a solid is a sensitive function of the atomic separation and therefore of the pressure. If the energy gap between the occupied and unoccupied energy bands can be closed under pressure, then a transition from insulator to metal or, more precisely, from semiconductor to semimetal will ensue. Assuming this transition were to be continuous, then as the bands cross a small number of electrons would be transferred from the occupied valence bands to the unoccupied conduction bands; but there is a Coulombic attraction between the electrons in the conduction band and the holes left behind in the valence band. A single electron-hole pair in a semiconductor will always bind to form a neutral exciton as a result of the long-range nature of the Coulombic attraction. So to create a conducting state, a finite density of electrons and holes is required that are able to screen the Coulomb force and so form an electron-hole plasma. In practice a suitable system that displays such a semiconductor-semimetal transition under pressure has not been found. Instead, if pressure is applied to a typical semiconductor such as Si or Ge, the insulator-to-metal transition coincides with a crystallographic transition between two crystalline phases with different symmetries and is not in principle different from a standard crystallographic phase transition. Of interest are the cases where the metal-insulator transition takes place without a change of lattice symmetry or between phases with closely related symmetries.

The best example of such a band-crossing transition within a single crystallographic phase is provided by the divalent fcc metals Ca, Sr, and Yb. Actually in this case the transition is from metal to insulator under pressure due to an increase of the direct band gap even as the overall energy bands broaden under pressure. The best studied example is Yb, which has the lowest transition pressure, ~10 kbar (McWhan *et al.*, 1969). The metal-insulator transition is found to be continuous rather than discontinuous or first order as discussed above. It is believed that this difference arises because of a very small energy scale and therefore a great sensitivity to impurities, crystalline defects, etc., which mask the true nature of the transition in a pure system.

Nonetheless there is an elegant way to explore the band-crossing transition by examining the phase diagram of an electron-hole fluid. By optical pumping of an indirect semiconductor such as Si or Ge, an electron-hole fluid with a long lifetime can be created and its thermodynamic behavior studied (for reviews see Hensel *et al.*, 1977; Rice, 1977; and Jeffries and Keldysh, 1984). A metallic self-bound electron-hole liquid is found to be the lowest energy state, lower than exciton states. The relationship to the semiconductor-semimetal transition is clear when one considers that this transition occurs by varying the energy to create an electron-hole pair, i.e., the chemical potential of the electron-hole fluid. The electron-hole fluid has a first-order liquid-gas transition below a critical temperature. Assuming a linear relationship between the chemical potential and an external variable such as pressure, then this implies that a first-order metal-insulator transition below the critical temperature (see Fig. 1) occurs in a band-crossing transition. Note that the energy scale is typically small (~10^{-3} eV) because of dielectric screening of the Coulomb interaction due to higher-energy interband transitions.

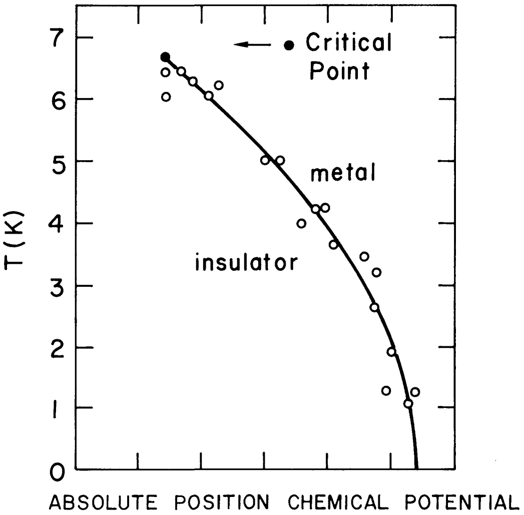

FIG. 1. The phase diagram of the electron-hole fluid in optically pumped Ge as determined by Thomas *et al.* (1974). A metallic phase is stable for values of chemical potential larger than a critical value below a critical temperature. In the analogous band-crossing transition this transition to a metallic phase corresponds to a discontinuous transition between an insulating phase, with a finite energy gap between conduction and valence energy bands, and a metallic phase, with finite overlap between these bands.

1.2 Mott Transitions

In the simplest view, a Mott transition arises because of the competition between the intra-atomic Coulomb interaction U, which opposes charge fluctuation, and the bandwidth W, which represents the kinetic energy that can be gained by allowing charged excitations to propagate through the crystal. The criterion separating metallic from insulating behavior then is determined by the comparison $U \lesssim W$ or $U \gtrsim W$. Since U is an intra-atomic Coulomb interaction, it should not be sensitive to pressure, unlike bandwidth, W. Therefore, in principle it should be possible to convert all Mott insulators into metals by applying enough pressure. In practice this is not so easy since Mott insulators generally have large energy gaps for charge excitations; e.g., NiO, the classic Mott insulator, has an energy gap ≈5 eV. Thus very large pressures are required, which, in turn, prevents detailed experimental study.

Vanadium sesquioxide, V_2O_3, is the best known exception. At atmospheric pressure, it displays a first-order metal-insulator transition (Föex, 1946) (Fig. 2) upon cooling at a temperature $T_{MI} \approx 150$ K with a resistivity increase by a factor ~10^7. The low-temperature phase is insulating and antiferromag-

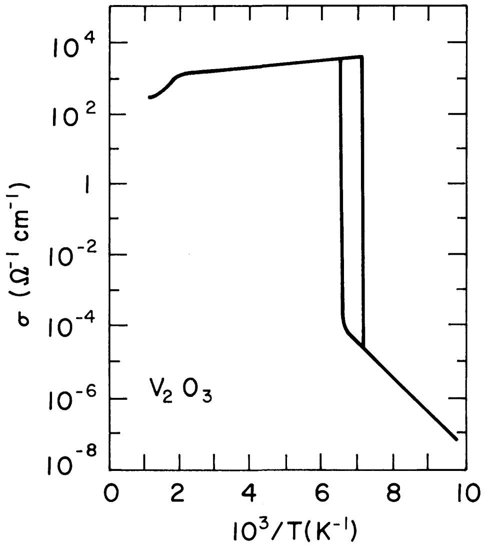

FIG. 2. The conductivity of V_2O_3 showing the metal-insulator transition found by Föex (1946) at 150 K. In addition, there is an anomaly in the conductivity at temperatures ≈600 K as discussed in the text.

netically ordered with an ordered moment ~$1.2\mu_B$/(V atom). In addition at high temperature, ~600 K, a broad anomalous increase in the resistivity is observed.

McWhan and coworkers (1971), by a judicious use of doping and pressure, established the full phase diagram of the V_2O_3 system. Substitution of Cr for V was found to act as a negative pressure that can be counteracted by applying real pressure, and this allows the complete phase diagram to be established (see Fig. 3). The key feature is an almost vertical line separating two phases with the same corundum crystal structure, a higher-temperature insulating phase and a lower-temperature metallic phase. In the insulating phases the electrons are localized to form V^{III} oxidation states, whereas the oxidation state fluctuates through charge fluctuations in the metal. This change does not require a change of symmetry so that the metal-insulator transition line ends at a critical point, which at atmospheric pressure is ≈400 K. The anomalous increase in the resistivity in pure V_2O_3 at ~600 K (see Fig. 2) can now be seen as a crossover at temperatures above the critical point. A key feature of the localization of the electrons is that the associated spin degrees of freedom remain free, leading to an increase of entropy, so that now the transition is from metal to insulator with increasing temperature. Only the charge degrees of freedom are frozen out by localization. The spins of the V^{III} oxidation state are coupled by superexchange processes and order antiferromagnetically as the temperature is lowered. The existence of this magnetic ordering transition is an important distinction from the band-crossing case discussed above. It leads to a distortion of the metal-insulator phase boundary due to the additional gain in free energy of the insulator caused by magnetic ordering. The metal-insulator transition remains first order, in agreement with the general considerations discussed above, although the actual size of the discontinuity may well be influenced by coupling with the lattice.

The metallic phase of V_2O_3 shows clear signs of the nearby localization transition. The effective degeneracy or Fermi temperature of the metal is greatly reduced near the metal-insulator boundary. This shows up in an enhancement by roughly an order of magnitude in the Pauli spin susceptibility and linear specific-heat coefficient and very clearly in the resistivity versus temperature. The latter (Fig. 4) displays a T^2 law at low temperatures, characteristic of Fermi liquids. The term Fermi liquid is used to describe an electronic state that is only minimally modified by the Coulomb interaction between the electrons

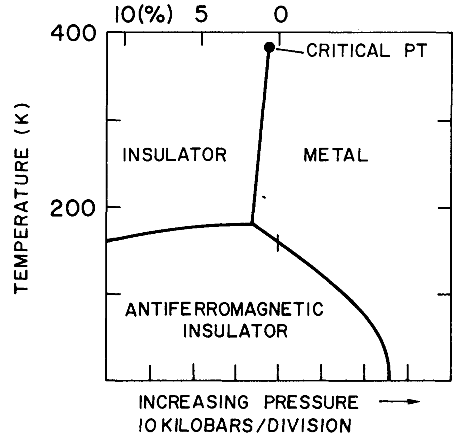

FIG. 3. Phase diagram of the metal-insulator transition in V_2O_3 and $(V_{1-x}Cr_x)_2O_3$ as a function of temperature T and pressure P as determined by McWhan *et al.* (1971).

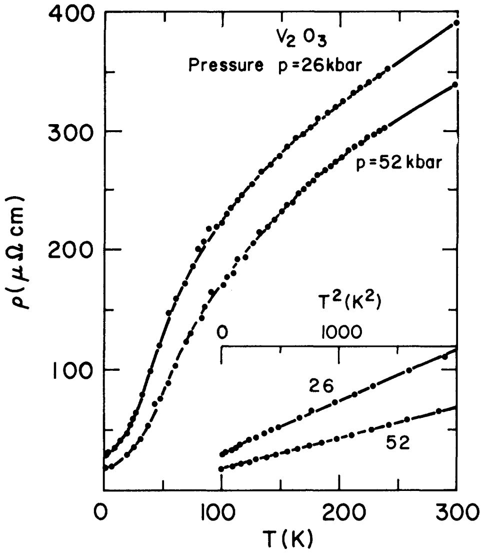

FIG. 4. The resistivity of metallic V_2O_3 under pressure as a function of temperature showing the enhanced T^2 dependence as the critical boundary to the insulating phase is approached (McWhan and Rice, 1969).

from the single-electron description in terms of partially filled energy bands. The coefficient of the T^2 term is unusually large, however, and it is a strong function of pressure and increases as the metal-insulator boundary is approached. This in turn leads to a reduction in the temperature that characterizes the crossover to coherent motion of the electrons through the periodic lattice that is a key property of a Fermi liquid (see Brinkman and Rice, 1970). The behavior of the metallic state near the Mott transition contrasts with the metallic state near a band-crossing transition that has a small overlap of the valence and conduction bands. In this latter case characteristic metallic properties such as the Pauli spin susceptibility and the linear specific-heat coefficient at low temperatures vanish as the metal-insulator transition is approached, whereas in the approach to a Mott transition they are enhanced. This difference reflects the difference in the insulating state—a band insulator has an energy gap for both spin and charge excitations, whereas in a Mott insulator the gap is only for charge excitations, and the spin excitations develop a low energy scale.

1.3 Metal-Insulator Transitions due to Lattice Distortion

In vanadium dioxide, VO_2, a strong first-order metal-insulator transition with a rise of roughly a factor of 10^5 in the resistivity at a temperature of 340 K was found by Morin (1959) (see Fig. 5). In contrast to the case of V_2O_3, the crystal structure changes at the transition from the tetragonal rutile phase, characterized by uniform chains of V atoms, in the metallic phase, to a monoclinic insulating phase with close V-V pairs. These have a separation of only 2.65 Å compared to a regular 2.88-Å separation in the high-temperature phase. The oxidation state V^{IV} has an electronic configuration $3d^1$ and a spin $s = \frac{1}{2}$ and so favors the formation of singlet pairs and a nonmagnetic insulating phase.

The large displacements of the V atoms that accompany the metal-insulator transition show that in this system the electron-lattice interaction plays an important role in stabilizing the insulating phase, and, in contrast to the V_2O_3 system, the phase transition is not sensitive to pressure. Nonetheless there are indications that electron localization plays a role here, too. In this regard there is the curious fact that an intermediate insulating phase can be stabilized by substitution of Cr or, more clearly, by applying uniaxial pressure (see Fig. 6). In this intermediate phase one-half of the V chains in the rutile structure pair, but the other half undergo a zigzag distortion, remaining unpaired. They form instead a chain of localized V^{IV} ions with spins $s = \frac{1}{2}$.

There are several Ti oxides that also have metal-insulator transitions due to the formation of singlet pairs of Ti^{III} ions. In the case of titanium sesquioxide, Ti_2O_3, the corundum structure contains Ti-Ti pairs, so that a continuous transition without a change of lattice symmetry is possible between a low-temperature insulating phase with close Ti-Ti pairs and a high-temperature metallic phase with an increased separation of the Ti-Ti pairs. In Ti_2O_3, this gradual transition with a change of a factor of 10 in the resistivity takes place around 500 K (Abrahams, 1963).

1.4 Metal-Insulator Transition due to Charge Ordering

The previous discussion was limited to cases where the average valence is an integer and all sites are equivalent. However, metal-insulator transitions also occur in cases when the average valence is a fractional number, i.e., in materials containing a mixture of at least two different oxidation states. In such systems, insulating behavior occurs when the intersite Coulomb repulsion, and possibly also the electron-lattice interaction, stabilizes a charge-ordering transition. The insulating phase has an ordered superstructure formed from ions with different oxidation states.

The Magneli phases, which are intermediate between the dioxides and sesquioxides with a chemical formulae T_nO_{2n-1} (T = Ti or V), generally undergo charge-ordering transitions. Particularly interesting is the case of Ti_4O_7, which has equal concentrations of Ti^{III} and Ti^{IV} oxidation states with $3d^1$ and $3d^0$ electronic configurations, respectively. The charge-ordering transition occurs in two stages at T = 150 and 130 K with a narrow intermediate range (~20 K wide) of reduced conductivity that is interpreted as a fluid state of singlet Ti^{III}-Ti^{III} pairs (Lakkis *et al.*, 1976). At lower temperatures ($T < 130$ K) charge or-

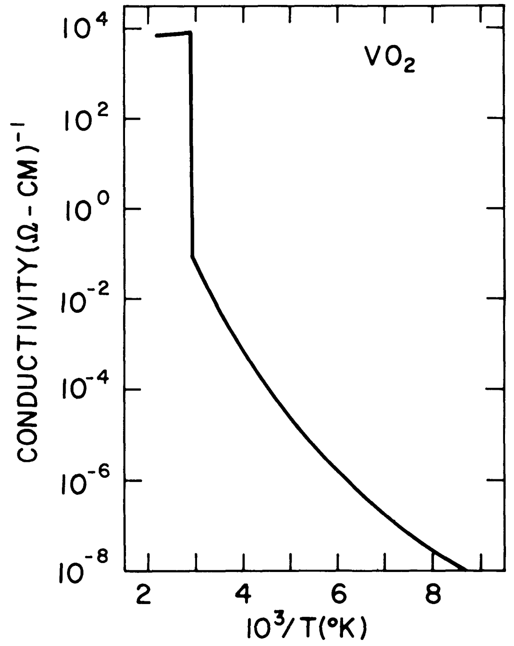

FIG. 5. The resistivity of VO$_2$ as a function of the temperature showing the metal-insulator transition at 340 K (from Ladd and Paul, 1969).

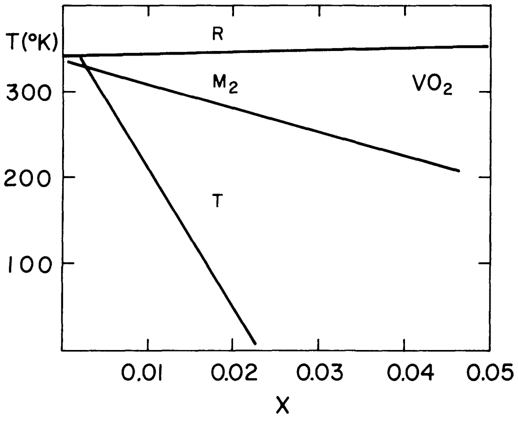

FIG. 6. The phase diagram of the V$_{1-x}$Cr$_x$O$_2$ system (Pouget *et al.*, 1974). Only the high-temperature rutile (*R*) phase is metallic. The insulating monoclinic phases *M*$_1$ and *M*$_2$ have either all V-V pairs (*M*$_1$) or one-half the V sites in pairs and the other half in localized V^{IV} chains (*M*$_2$). *T* is a triclinic insulating phase intermediate between *M*$_1$ and *M*$_2$.

dering of the Ti^{III}-Ti^{III} pairs occurs and the conductivity is further reduced.

The Verwey (1939) transition in magnetite (Fe$_3$O$_4$) is another example of a metal-insulator transition driven by charge ordering. In this case the rise in resistivity is modest (a factor of 10^2) although the transition temperature is relatively low (119 K), and the charge ordering is between Fe^{II} and Fe^{III} oxidation states on one of the sublattices in the structure.

1.5 Metal-Insulator Transitions in Lower-Dimensional Systems

The case of a one-dimensional metal is special. One-dimensional metals occur when electrons are constrained to move along chains of atoms or molecules. Peierls (1954) showed that they are inherently unstable toward the formation of charge-density-wave (CDW) states, which are characterized by a periodic charge modulation with a periodicity related to the electronic density on the chains. Such a CDW introduces a gap in the energy spectrum for charge and spin excitations. A motion of the CDW (i.e., the modulated charge) relative to the background lattice, which would contribute to the electrical conductivity, is generally not possible at small electric fields because of various effects that lock the two periodic systems to each other. As a result the CDW state is insulating and the Peierls transition is also a metal-insulator transition.

Several examples of Peierls transitions are known, and in particular the cases of the organic conductor TTF-TCNQ (tetrathiafulvalene tetracyanoquinodimethane) and the inorganic chain compound K$_{0.3}$MoO$_3$ are well studied. In both cases the metal-insulator transition is continuous at temperatures of 60 and 200 K, respectively.

2. METAL-INSULATOR TRANSITIONS IN DISORDERED SYSTEMS

2.1 Uncompensated Doped Semiconductors

When dopants with a different nuclear charge from that of the host are introduced into a semiconductor, shallow donor or acceptor levels may result. Such systems can be well described by a hydrogenic model.

Well-known examples are the donor systems Si:P and Ge:Sb and the acceptor systems Si:B and Ge:Ga. In the hydrogenic model the Coulomb interaction between the charges is reduced by the dielectric constant κ of the semiconductor. The electron or hole effective mass m^*, determined by the curvature of the conduction or valence bands, enters the kinetic energy. As a result, the hydrogenic donor or acceptor orbits can be very large and have characteristic radii that are many lattice constants, justifying the hydrogenic model.

In the dilute limit, the uncompensated donor (acceptor) system is a form of Mott insulator with the electrons (holes) localized on individual donors (acceptors) to form an array of well-separated hydrogenic atoms. As the dopant density n increases, so does the overlap of the hydrogenic atoms, and a Mott transition to a metallic state ensues. In Fig. 7 a series of curves for the resistivity ρ versus $1/T$ at various acceptor concentrations for the system Ge:Ga illustrate the extreme sensitivity of ρ to the concentration. For acceptor concentrations $n \lesssim 10^{17}$ cm^{-3}, $\rho(T)$ continues to rise down to the lowest temperatures, $\approx$1 K in the figure. But for values of $n \gtrsim 10^{17}$ cm^{-3}, $\rho(T)$ extrapolates to a finite value as $T \rightarrow 0$ K. In the clearly insulating regime with $n \ll 10^{17}$ cm^{-3} the resistivity has an exponential temperature dependence with an activation energy ϵ_i ($i = 1, 2,$ or 3) depending on the concentration and temperature regime. This difference in activation energy is believed to reflect different processes that control the resistivity, but we will not discuss this further. Rather we wish to stress the point that an activation energy in the resistivity reflects an energy gap for charge excitations. The charge carriers (electrons or holes) freeze out at low temperatures on to hydrogenic donor or acceptor orbits. This process shows up also in an exponential rise in the Hall coefficient R_H, which is inversely proportional to the effective number of carriers n^* ($R_H \propto 1/n^*$). Thermal excitation across the energy gap between localized orbits and ex-

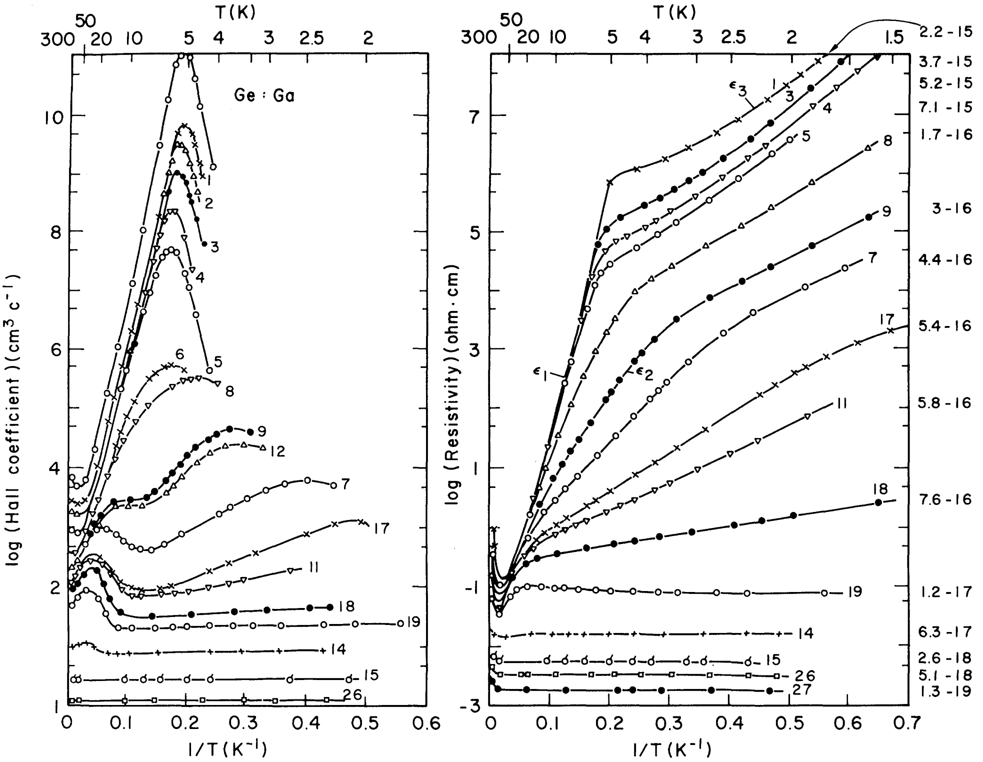

FIG. 7. The Hall coefficient and resistivity of Ga-doped Ge samples as functions of temperature for different Ga concentrations, as determined by Fritzsche (1955, 1978). The concentrations of the Ga acceptors are shown using the notation $X\text{-}Y \Rightarrow X \times 10^Y$ cm^{-3}.

tended states is then the source of the exponential form for n^*.

Using scaled hydrogenic values for the intrasite Coulomb repulsion and the band-width for charged excitations, Mott (1949, 1961) proposed a formula for the critical density n_c separating metallic and insulating behavior:

$$n_c^{1/3} a_{\mathrm{H}}^* \simeq 0.25, \qquad (1)$$

where a_{H}^* is a scaled Bohr radius ($a_{\mathrm{H}}^* = \hbar^2\kappa/m^*e^2$). This critical density depends on the material parameters κ and m^*. In a wide variety of semiconductors one finds that the critical dopant density agrees well with this criterion (see Fig. 8).

In contrast to the ordered systems, the metal-insulator transition here appears as a true critical point only at zero temperature. At finite temperatures a smooth crossover is observed with arbitrarily small values of the conductivity (see Fig. 9), rather than the discontinuous transition with a jump in conductivity discussed in Sec. 2. An important physical difference here is the strong disorder present in the doped semiconductors. Since the dopants generally are introduced at high temperatures, the spatial correlations between the dopant positions are on the scale of a lattice constant. This scale is much smaller than the size of the scaled hydrogenic orbits of the donor or acceptor wave functions. As a result, the disorder in the doped semiconductors is very strong and represents an important physical difference from the ordered systems discussed earlier.

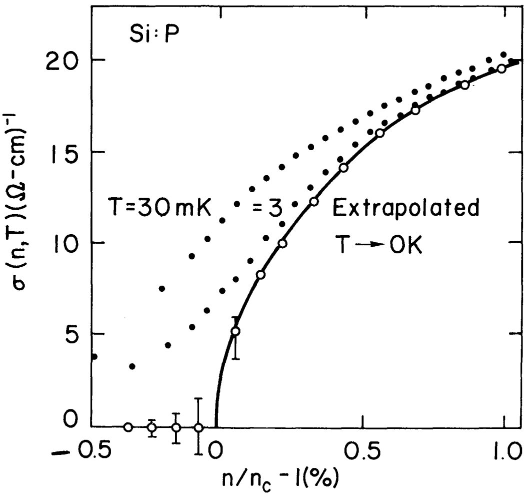

FIG. 9. The dc conductivity of Si:P as it is tuned through the metal-insulator transition by applying an external stress to vary n_c. The solid circles are data at temperatures of 30 and 3 mK. The open circles are data extrapolated to T = 0 K (Thomas *et al.*, 1983).

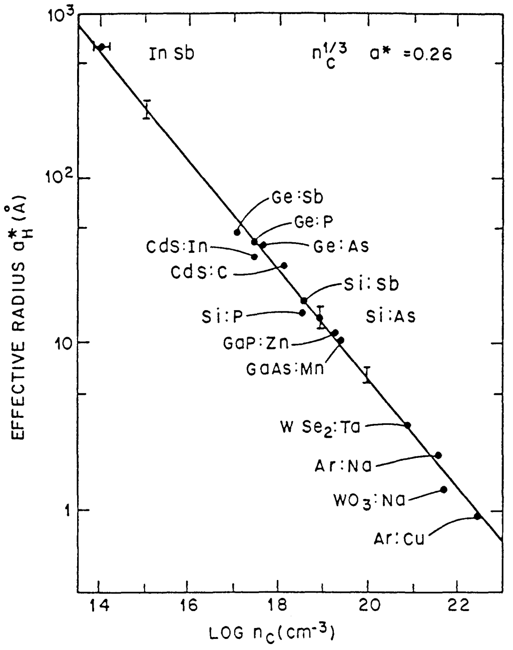

FIG. 8. The Mott criterion, Eq. (1), for the critical concentration dividing metallic ($n > n_c$) from insulating ($n < n_c$) behavior in a variety of systems. Data compiled by Edwards and Sienko (1978).

The nature of the metal-insulator transition in doped semiconductors has been the subject of many experimental investigations over the years. A particularly detailed set of measurements has been reported recently by Thomas *et al.* (1983) on the classic donor system Si:P. They extended the temperature range of the measurements down to 3 mK. Further they used an applied stress S to tune continuously across the critical region. In this way they could obtain much more precise measurements of the critical behavior. On the metallic side they extrapolated their results down to zero temperature using a form $\sigma(T) = \sigma(0)(1 + aT^{1/2})$ and obtained the critical behavior of $\sigma(0,n,S)$.

The extrapolated values fit well to a critical form

$$\sigma(0,n,S) = \sigma_0\{[n/n_c(S)] - 1\}^{\mu}. \qquad (2)$$

The prefactor σ_0 and, especially, the critical density n_c are functions of the external stress S, and it is this sensitivity that allows one to tune through the metal-insulator transition. Theoretical interest has focused on the fact that the conductivity varies continuously to zero and on the value of the exponent μ, which characterizes the critical region. Paalenen *et al.* (1983) report a value of $\mu = 0.5$ for experiments on Si:P (see Fig. 10).

On the insulating side for values of the density n/n_c, the approach to the metallic state is characterized by increasing values of the low-frequency limit of the dielectric polarizability χ. Note that low frequency in this context means that the corresponding energy is below the characteristic energies of the electronic transitions associated with the donors or acceptors. These characteristic energies decrease as the metallic state is approached and the value of χ increases. In fact, Herzfeld (1927), in a very early work on the criterion for metallic behavior, used this idea that a breakdown of an insulator occurs through a divergence of the low-frequency dielectric polarizability. The Herzfeld criterion separating metallic and insulating, when applied to a system of hydrogenic atoms, is very similar to Mott criterion, Eq. (1)—actually the difference is only that the constant on the right-hand side is 50% higher. In a metal the value of the low-frequency and long-wavelength polarizability is not well defined and depends on the ratio of the frequency to the wave vector as both go to zero in this limit. In an insulator the limiting value is well defined but can become arbitrarily large as $n \rightarrow n_c$. A power-law divergence in $\chi(n)$ with exponent ζ is observed,

$$\chi = \chi_0[(n_c/n) - 1]^{-\zeta}. \qquad (3)$$

Experimentally, a value $\zeta = 2\mu = 1.0$ is found for the exponent of the dielectric constant (see Fig. 10).

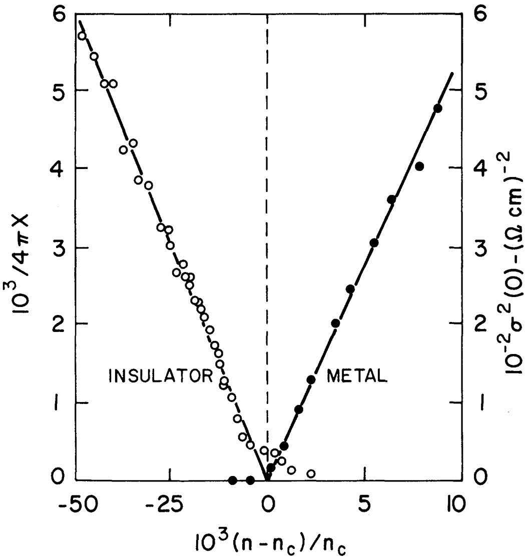

FIG. 10. The inverse donor polarizability and the square of the conductivity plotted against the deviation of the concentration from the critical value n_c, which is tuned by applying an external stress. The linear variation gives a value $\mu = \frac{1}{2}$ for critical exponent (Paalenen *et al.*, 1983).

As discussed previously, at a Mott transition in an ordered system, a change in the magnetic properties from paramagnetic metal to antiferromagnetic insulator is observed. Similarly on the metallic side, i.e., at large doping $n \gg n_c$, there is a paramagnetic Pauli spin susceptibility χ_s that is temperature independent at low temperature. On the dilute or insulating side the susceptibility is found to be strongly temperature dependent with a power-law divergence of $\chi_s(T) \propto c/T^a$. Typical values are $a \sim 1$. Such a behavior was explained by Bhatt and Lee (1982) as a successive singlet formation by donor pairs as the temperature is lowered. The question of the detailed evolution of the magnetic properties between these two limiting behaviors and the relationship to the evolution of the electrical conductivity remains open.

There are two important ingredients in the theoretical description of uncompensated doped semiconductors. One is the role of the intrasite and intersite Coulomb interactions, and the other that of the randomness inherent in the doping process. As discussed above, the former alone would cause a finite-temperature critical point and a discontinuous transition at lower temperature. However, experimentally only a zero-temperature transition is observed. This points to an essential role for the random disorder in the system.

Indeed, it was the properties of doped semiconductors that first stimulated Anderson (1958) to study the quantum-mechanical problem of a particle moving in a random potential. In a classic article he showed that a transition occurs from extended to localized states as the strength of the disorder increases. This in turn implies a zero-temperature transition from a metallic state with finite conductivity and with extended states at the Fermi energy to an insulating state with zero conductivity and localized states at the Fermi energy. The next important development was the introduction of a scaling hypothesis for the conductance as a function of the characteristic length by Abrahams *et al.* (1979). These authors calculated the scaling function perturbatively in the weak-disorder limit and showed that it was a sensitive function of the dimension of the system under consideration and that two is the critical dimension. In a lower dimension, such as a one-dimensional chain of atoms, the conductance decreases exponentially with length and reflects the fact that in a random chain all states are localized. In the special case of planar motion or two dimensions, the initial corrections are logarithmic and give rise to a set of effects known as weak localization effects, e.g., a reduction in the conductivity that is logarithmic in the temperature at low temperatures and that is very sensitive to an applied magnetic field. These effects have been observed in thin films, and Bergmann (1984) has given an elegant physical interpretation of weak localization as an interference effect between time-reversal-related pairs of paths that an electron can take in a two-dimensional film.

Returning to three-dimensional samples, in this case a transition is predicted from metallic and extended to insulating and localized with increasing disorder. The characteristic length ξ that enters the scaling theory determines the dc conductivity σ, through a relationship of the form

$$\sigma = c^1 e^2/\hbar\xi, \tag{4}$$

where c^1 is a numerical constant. This characteristic length ξ has a power-law divergence as the localization or Anderson transition is approached, $\xi \propto (n/n_c - 1)^{-\mu}$. This divergence causes critical behavior in the conductivity, of the form (2). On the insulating side, the characteristic length ξ is the localization length, which determines the extent of the localized wave functions, and this diverges also at the critical point. Assuming that on shorter length scales the static wave-vector–dependent polarizability has a metallic form, the divergence in ξ leads to a divergence in χ of the form

$$\chi = \chi_0(n_c/n - 1)^{-2\mu}. \tag{5}$$

Such effects are indeed observed as discussed previously.

The one-component-scaling theory of Abrahams *et al.* (1979) predicts a value $\mu = 1$, in contrast to experiment on uncompensated Si:P quoted earlier, which give a value of $\mu = \frac{1}{2}$. It is plausible to ascribe this discrepancy to the Coulomb interaction, which may cause the transition region to be more first-order-like, or in other words to have a smaller value of the critical exponent, μ. A number of attempts have been made to extend the scaling theory to include electron-electron interactions. The initial corrections in the weak-disorder limit, e.g., to weak-localization effects, are well understood, but it has proved difficult to extend the theory to describe the region of the critical point. In particular, key questions concerning the role of possible magnetic instabilities remain to be resolved. At present there is no generally accepted theory for the transition region and the value of the critical exponent (see Lee and Ramakrishnan, 1985).

A key idea taken over from the study of other phase transitions is that the critical exponents such as μ and ζ should have universal values that should be independent of the material system. This universality is now under question as a result of a recent series of elegant experiments by Katsumoto (1988) on the system $Al_{0.3}Ga_{8.7}As$ doped with Si. He used a special feature of this system, that Si donors can form either deep (i.e., strongly bound) states or shallow (i.e., hydrogenic) states, and that the deep states can be converted to long-lived shallow states by shining light on the material. Katsumoto used this technique to tune continuously through the critical region, and under the reasonable assumption that deep states are inert, then the metal-insulator transition takes place among the shallow states. By varying the total light flux, the effective donor concentration can be continuously changed and tuned through the metal-insulator transition. However, Katsumoto finds values $\mu = 1$ and $\zeta \sim 2$ in agreement with the theoretical estimates in the Anderson limit (see Fig. 11). At present the origin of the discrepancy between the experiments on the two systems remains unclear.

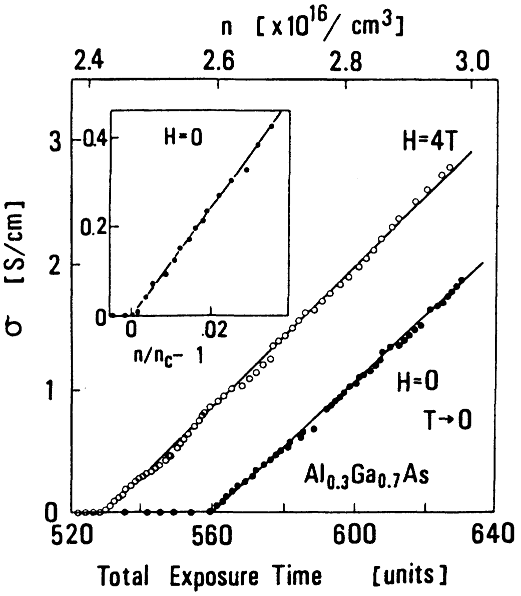

FIG. 11. The extrapolated zero-temperature conductivity as a function of the concentration of shallow Si donors, *n*, in the photoexcited $Al_{0.3}Ga_{0.7}As$:Si system (Katsumoto, 1988). Inset: the linear dependence of σ on $n/n_c - 1$ close to the critical concentration n_c. This linear dependence remains unaltered in a magnetic field $H = 4$ T.

2.2 Compensated Doped Semiconductors

A doped semiconductor is said to be compensated if both donors and acceptors are added. In this case the majority dopants of one valence are compensated by minority dopants so that average valence change is reduced. Consider the case of Si:P discussed previously. The material without acceptors is said to be uncompensated and has one electron carrier added per P donor. If now B acceptors are added, the average valence change is reduced and electrons are trapped at the B sites. As a result, the number of electron carriers is now less than one per P dopant. In this case an insulating state in the dilute limit occurs as a consequence not of the intrasite Coulomb interaction but of the randomness and possibly of the intersite Coulomb interactions. Note that the random character of the potential is strengthened by the trapped charges introduced at compensating impurity sites. As a result, compensated doped semiconductors are closer to the Anderson limit, and the description in terms of a single particle in a random potential should work better. Experimentally, values $\mu = 1$ are generally found, in agreement with this hypothesis.

Metal-semiconductor transitions can also be obtained by forming random alloys of metals and semiconductors. Examples of such systems are the amorphous alloys Nb_xSi_{1-x} (Hertel *et al.* 1983) and Au_xSi_{1-x}. In both cases the metal-insulator transition occurs as a zero-temperature critical point, and critical exponents $\mu \approx 1$ are measured. In these systems the difference between the electronic structure of the two components is very large, and presumably this determines the character of the metal-insulator transition and leads to randomness as the dominant effect.

2.3 Doping of Mott Insulators

Generally it is considerably more difficult to achieve a transition to a metallic state by doping a Mott insulator. For example, attempts to dope NiO, the best studied Mott insulator, by substituting Li atoms do not lead to a conducting metallic state. Rather the carriers seem to polarize the lattice strongly and thereby form polaron states with a small radius that have a low mobility. Such small polarons are easily localized and insulating behavior follows.

The system $La_{1-x}Sr_xVO_3$ is an exception. The parent compound $LaVO_3$ is an antiferromagnetically ordered insulator with a Néel temperature $T_N \simeq 120$ K and an ordered moment of 1.2 μ_B/(V atom). Substitution of Sr for La atoms decreases the resistivity and also the magnetic susceptibility. At concentrations $x \gtrsim 0.22$, the resistivity is temperature independent and indicates that a transition to a metallic state has taken place (Dougier and Hagenmuller, 1975). Detailed studies of the low-temperature behavior are not available.

The high-T_c superconducting state is obtained by doping the parent antiferromagnetic cuprates, e.g., La_2CuO_4, and this will be discussed below.

2.4 Metal-Insulator Transitions in Expanded Liquid Metals

In one of the earliest articles to consider the nature of the metal-insulator transition, Landau and Zel'dovich (1943) speculated that it could appear as a separate critical point from the critical point of the liquid-gas phase transition in metals. In general the study of the liquid-gas critical points of metals is technically difficult, but in the case of metals with low melting points such as Hg or Cs the critical point is more easily accessible. A number of studies of the conductivity and other properties of fluid Hg and Cs have been made [for a recent review see Freyland and Hensel (1985)]. Only one critical point of the liquid-gas type is observed, and the conductivity evolves continuously. In the case of Hg, which is divalent, the basic metal-insulator transition should be of the band-crossing type, whereas in Cs, which is monovalent, it is of the Mott type. Apparently in both cases the metal-insulator critical temperature lies below the liquid-gas phase boundary and in the physically inaccessible two-phase region. In the case of mercury the evolution from metallic to nonmetallic on the liquid-gas phase boundary occurs continuously around a density ~7.5 g cm^{-3}, which is above the density of the critical point (4.9 g cm^{-3}).

3. SUPERCONDUCTOR-INSULATOR TRANSITIONS

3.1 The $BaBi_{1-x}Pb_xO_3$ and $Ba_{1-y}K_yBiO_3$ Systems

Interest in direct transitions from insulating to superconducting metallic states has risen since the recent discoveries of high-temperature superconductivity in several such cases. One of these systems starts with the parent compound $BaBiO_3$, which would be a metal in the cubic perovskite structure due to a half-filled band derived from antibonding Bi(6*s*)-O(2*p*) hybridized states. In fact it is an insulator with a band gap ≈2 eV due to a lattice distortion in which the Bi^{IV} oxidation states disproportionate into Bi^{III} and Bi^{V} oxidation states coupled to breathing-mode distortions of the O lattice (Cox and Sleight, 1979). This distortion can be thought of as a commensurate charge-density wave (CDW) that splits the band and causes the energy gap and insulating character.

The metallic and superconducting states are reached by substituting either Pb atoms on the Bi site or K atoms on the Ba site. In both cases, the insulating character continues to be observed up to a substantial critical concentration. In the former case $x_c \sim 0.65$ while the latter case $y_c \sim 0.35$. X-ray studies in both cases do not show coherent CDW distortions in these concentration ranges, but in both cases optical studies show that the CDW-derived energy gap has not closed. Therefore, it is plausible to ascribe the insulating character to local CDW distortions that persist up to the critical concentration, at least. The transition to the metallic state occurs as the energy gap closes, although in the normal phase the conductivity is quite low. Nonetheless these poor metals are good superconductors with relatively high values of T_c [13 and 30 K respectively; see Sleight *et al.* (1975) and Mattheiss *et al.* (1988)].

Detailed studies of the insulator-superconductor transition have been made in the case of the earlier compound, $BaBi_{1-x}Pb_xO_3$, with values of $x \sim x_c$. At the optimum concentration for superconductivity ($x \approx 0.75$) the metallic state is only poorly conducting with large (~1 mΩ cm) and temperature-independent resistivity. But at smaller values of x, as shown in Fig. 12, a spectacular insulator-superconductor transition is observed. The normal phase shows a resistance that rises rapidly as the temperature is lowered until the superconductivity is reached, where it drops precipitously. This system is the clearest example of a direct insulator-superconductor transition.

3.2 High-T_c Cuprate Superconductors

In the case of the high-T_c cuprate superconductors, again a direct transition from insulating to superconducting is observed as the oxidation state of the Cu ions is changed away from Cu^{II} without any intermediate normal metallic phase. In this case the parent insulating phases are antiferromagnetic with $s = \frac{1}{2}$ moments on the Cu^{II} sites and so fall in the general class of Mott insulators. The materials La_2CuO_4 and $YBa_2Cu_3O_6$ are well-studied examples with optical energy gaps ≈2 eV. The oxidation state of the Cu is increased by substituting Sr or Ba for La in the former system or by adding O in the latter case. This introduces low-spin ($s = 0$) Cu^{III} oxidation states, which are mobile in the background of Cu^{II} sites through the exchange of an electron and very effectively destroy the long-range magnetic order. In Fig. 13, the evolution of the resistivity in the $La_{2-x}Sr_xCuO_4$

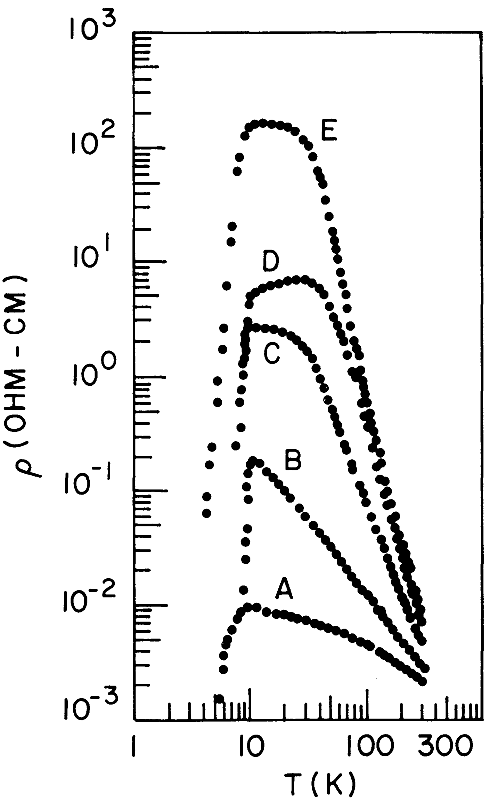

FIG. 12. The variation of the resistivity through the insulator-superconductor transitions in $BaBi_{1-x}Pb_xO_3$ samples with Pb concentrations x = 0.65 (*A*), 0.62 (*B*), 0.61 (*C*), and 0.59 (*D, E*) measured by Takagi *et al.* (1985).

system is displayed as recently reported by Takagi *et al.* (1992). Again we see a direct evolution from insulating behavior with a divergent resistivity at the lowest temperatures to a superconducting behavior with zero resistivity below a critical temperature. No normal metallic state, characterized by a finite value of the zero-temperature resistivity, is observed as the concentration of Cu^{III} states is increased. The magnetic properties evolve from antiferromagnetic to paramagnetic, but the detailed relationship to the insulator-superconductor transition continues to be studied and to be sensitive to the system.

At present, the nature of the normal phase at temperatures above the transition to superconductivity and the microscopic cause of the superconductivity continue to be actively debated.

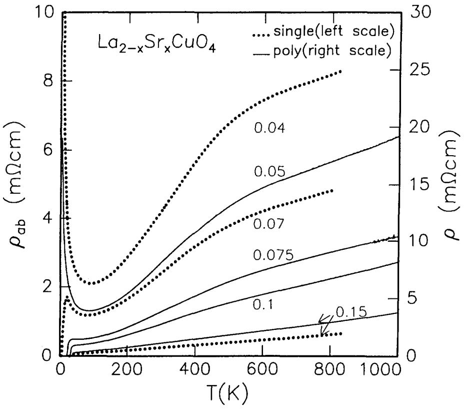

FIG. 13. The resistivity vs temperature for the high-temperature superconductor $La_{2-x}Sr_xCuO_4$ in a series of polycrystal and single-crystal samples with varying Sr concentration (Takagi *et al.*, 1992). Note that no samples show normal metallic behavior (i.e., finite resistivity) at low temperatures so that the zero-temperature transition is directly from an insulator to a superconductor with increasing values of x.

4. APPLICATIONS

The obvious potential applications for metal-insulator transitions lies in their use as switches. In particular, the vanadium dioxide system (VO_2) with a critical temperature of 340 K or 67°C could be interesting for a number of uses. However, devices built on these principles have not been popular. The relatively small change in resistance (only a factor of 10^5), the slow switching time, and the difficulty of repeated cycling of a sample through the transition are the most important drawbacks.

GLOSSARY

Acceptor: A dopant (or foreign) element with a smaller ionic charge than that of the host, which accepts (or binds) an electron from the conduction band into a state in the lower part of the energy gap.

Band-Crossing Transition: A transition between electronic states without overlap between the conduction and valence energy bands to a state in which these bands overlap in energy.

Band Insulator: An insulator caused by a gap in energy between the conduction and valence energy bands.

Charge-Density Wave (CDW): A static modulation of the charge density at atomic sites accompanied by a periodic lattice distortion.

Charge Ordering: An ordered configuration of oxidation states of different valences.

Conduction Band: The lowest unoccupied electron energy band.

Critical Concentration (n_c): The concentration of dopants that separates metallic ($n > n_c$) from insulating ($n < n_c$) behavior.

Donor: A dopant (or foreign) element with a larger ionic charge than that of the host, which donates (or liberates) an electron into a state near to the lowest energy in the conduction band.

Electron-Hole Fluids: A fluid of electrons in the conduction band and holes (i.e., unoccupied states) in the valence band that is analogous to a fluid of electrons and positrons.

Energy Bands: The quantum-mechanical states of a single electron moving in the periodic potential of the crystal have energies in allowed bands separated by gaps or forbidden zones.

Extended States: Quantum-mechanical states of a single electron in a random potential that extend throughout the whole sample.

Fermi Liquid: A quantum-mechanical state which is only minimally modified by Coulomb interaction from the single-electron description in terms of a partially occupied energy band.

Insulator: A material whose electrical conductivity, extrapolated to zero temperature, vanishes.

Localized States: Quantum-mechanical states of a single electron in a strongly random potential that are localized in space so that their amplitude is concentrated near a single point in the sample.

Metal: A material whose electrical conductivity, extrapolated to zero temperature, does not vanish.

Mott Insulator: A material that in the band description should be a metal but nonetheless is an insulator because of localization of electrons on individual atoms by the Coulomb interaction.

Mott Transition: The transition between the Mott insulating and metallic states.

Oxidation State: The formal valence of an ion in an oxide obtained by assigning formal charge of −2 to the oxygen ions. It determines the symmetry properties of the highest occupied electronic states.

Semiconductor: A band insulator with a relatively small energy gap between valence and conduction bands.

Semimetal: A metal due to small overlap between valence and conduction bands.

Superconductor: A metal whose conductivity is infinite below a critical temperature.

Weak Localization: A set of phenomena occurring in thin films, characterized by weak (logarithmic) decrease of the conductivity to zero at low temperatures.

Works Cited

Abrahams, E., Anderson, P. W., Licciardello, D. W., Ramakrishnan, T. V. (1979), *Phys. Rev. Lett.* **42,** 673–676.

Abrahams, S. C. (1963), *Phys. Rev.* **130,** 2230–2237.

Anderson, P. W. (1958), *Phys. Rev.* **109,** 1492–1505.

Bergmann, G. (1984), *Phys. Rep.* **107,** 1–58.

Bhatt, R. N., Lee, P. A. (1982), *Phys. Rev. Lett.* **48,** 344–347.

Brinkman, W. F., Rice, T. M. (1970), *Phys. Rev. B* **2,** 4302–4304.

Cox, D. E., Sleight, A. W. (1979), *Acta Crystallogr.* **B35,** 1–10.

de Boer, J. H., Verwey, E. J. W. (1937), *Proc. Phys. Soc.* **A49,** 59–71.

Dougier, P., Hagenmuller, P. (1975), *J. Solid State Chem.* **15,** 158–166.

Edwards, P. P., Sienko, M. J. (1978), *Phys. Rev. B* **17,** 2575–2579.

Föex, M. (1946), *C. R. Acad. Sci. (Paris)* **223,** 1126–1128.

Freyland, W., Hensel, F. (1985), in: P. P. Edwards, C. N. R. Rao (Eds.), *The Metallic and Non-Metallic States of Matter,* London: Taylor & Francis, p. 93.

Fritzsche, H. (1955), *Phys. Res.* **99,** 406–419.

Fritzsche, H. (1978), in: L. R. Friedman, D. P. Tunstall (Eds.), *The Metal-Nonmetal Transition in Disordered Systems,* Scottish Universities Summer School in Physics, Edinburgh: SUSSP Publications, p. 193.

Jeffries, C. D., Keldysh, L. V. (1984) (Ed.), *Electron-Hole Droplets in Semiconductors,* Amsterdam: North-Holland.

Hensel, J. C., Phillips, T. G., Thomas, G. A. (1977), in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics,* Vol. 32, New York: Academic, pp. 87–314.

Hertel, G. *et al.* (1983), *Phys. Rev. Lett.* **50,** 743–746.

Herzfeld, K. F. (1927), *Phys. Rev.* **29,** 701–705.

Katsumato, S. (1988), in: T. Ando, H. Fukuyama (Eds.), *Anderson Localization*, Springer Proceedings in Physics Vol. 28, New York: Springer, pp. 45–52.

Ladd, L. A., Paul, W. (1969) *Solid State Commun.* **7**, 425–428.

Lakkis, S., Schlenker, C., Chakraverty, B. K., Buder, R., Marezio, M. (1976), *Phys. Rev. B* **14**, 1429–1440.

Landau, L. D., Zel'dovich, G. (1943), *Acta Phys. Chem. (USSR)* **18**, 194. [See D. ter Haar (Ed.), *Collected Papers of L. D. Landau*, New York: Gordon and Breach.]

Lee, P. A., Ramakrishnan, T. V. (1985), *Rev. Mod. Phys.* **57**, 287–337.

Mattheiss, L. F., Gyorgy, E. M., Johnson, S. W. (1988), *Phys. Rev. B* **37**, 3745–3746.

McWhan, D. B., Rice, T. M. (1969), *Phys. Rev. Lett.* **22**, 887–890.

McWhan, D. B., Rice, T. M., Schmidt, P. H. (1969), *Phys. Rev.* **177**, 1063–1071.

McWhan, D. B. *et al.* (1971), *Phys. Rev. Lett.* **27**, 941–943.

Morin, F. J. (1959), *Phys. Rev. Lett.* **3**, 34–36.

Mott, N. F. (1949), *Proc. Phys. Soc.* **A62**, 416–422.

Mott, N. F. (1956), *Can. J. Phys.* **34**, 1356–1368.

Mott, N. F. (1961), *Philos. Mag.* **6**, 287–309.

Paalenen, M. A., Rosenbaum, T. F., Thomas, G. A., Bhatt, R. N. (1983), *Phys. Rev. Lett.* **51**, 1896–1899.

Peierls, R. E. (1954), *Quantum Theory of Solids*, New York: Oxford University.

Pouget, J. P. *et al.*, (1974), *Phys. Rev. B* **10**, 1801–1815.

Rice, T. M. (1977), in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, Vol. 32, New York: Academic, pp. 1–86.

Sleight, A. W., Gillson, J. L., Bierstedt, P. E. (1975), *Solid State Commun.* **17**, 27–28.

Takagi, H., Naito, M., Uchida, S., Kitazawa, K., Tanaka, S., Katsui, A. (1985), *Solid State Commun.* **55**, 1019–1022.

Takagi, H. *et al.* (1992), *Phys. Rev. Lett.* **69**, 2975–2978.

Thomas, G. A., Rice, T. M., Hensel, J. C. (1974), in: M. H. Pilkuhn (Ed.), *Proceedings of the 12th International Conference on Physics and Semiconductors*, Stuttgart: Teubner, p. 105–109.

Thomas, G. A., Paalenen, M. A., Rosenbaum, T. F. (1983), *Phys. Rev. B* **27**, 3897–3900.

Verwey, E. J. W. (1939), *Nature* **144**, 327–328.

Further Reading

The most comprehensive account of the field is in Mott, N. F. (1990), *Metal-Insulator Transitions*, 2nd ed., London: Taylor & Francis; further details on most of the topics mentioned here can be found there. Other good sources are Edwards, P. P., Rao, C. N. R. (Eds.) (1985), *The Metallic and Non-Metallic States of Matter*, London: Taylor & Francis; Efros, A. L., Pollak, M. (Eds.) (1985), *Electron-Electron Interactions in Disordered Systems*, Amsterdam: North-Holland.

Conferences and summer schools devoted to this field, and especially to the study of disordered semiconductors, are organized fairly often, although there is no regular conference series devoted to this subject. The proceedings of these meetings also give valuable overviews of the subject. Recent examples include Friedman, L. R., Tunstall, D. P. (Eds.) (1978), *The Metal-Nonmetal Transitions in Disordered Systems*, Scottish Universities Summer School in Physics, St. Andrews, Edinburgh: SUSSP Publications; "Impurity Bands in Semiconductors," *Philos. Mag.* **B42**, 723–1003; Landsberg, P. T. (Ed.) (1985), "Transition in Semiconductors," *Solid State Electron.* **28**, 3–216; Kramer, B., Bergmann, G., Bruynseraede, Y. (Eds.) (1985), *Localization, Interaction and Transport Phenomena*, Springer Series in Solid State Science, Berlin: Springer; Ando, T., Fukuyama, H. (Eds.) (1988), *Anderson Localization*, Springer Proceedings in Physics Vol. 28, Berlin: Springer; Benedict, K. A., Chalkov, J. T. (Eds.) (1990), *Localization 1990*, Institute of Physics Conference Series No. 108: Bristol: The Institute of Physics.

A recent review of the metal-insulator transition in Si:P has been written by Löhneysen, H. v. (1990), *Festkörperprobleme* **30**, 95–111.

METALLIC GLASSES

T. Egami, *Department of Materials Science and Engineering, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.*

Introduction 129
1. **Atomic Structure of Metallic Glasses** 130
1.1 Method of Structural Characterization 130
1.2 Atomic Pair-Distribution Function 130
1.3 Structural Models 131
2. **Thermodynamics** 131
2.1 Structure of Liquids and the Glass Transition 131
2.2 Kinetics of the Liquid-to-Crystal Transition 132
2.3 Glass-Forming Composition 132
2.4 Crystallization and Structural Relaxation 133
3. **Methods of Producing Metallic Glasses** 134
3.1 Rapid Quenching Methods 134
3.2 Solid-State Amorphization 135
3.3 Physical and Chemical Deposition 135
4. **Properties** 135
4.1 Mechanical Properties 135
4.2 Magnetic Properties 136
4.3 Electrical Properties 137
4.4 Chemical Properties 138
5. **Applications** 138
5.1 Magnetic Applications 138
5.2 Other Applications 139
6. **Summary** 139
Works Cited 139
Further Reading 139

INTRODUCTION

In our daily life, the word glasses implies oxide glasses used as window panes and kitchenware. Scientifically, however, the glassy state is defined as an amorphous or vitreous solidlike state of matter having an aperiodic and disordered atomic structure (see Glasses), in contrast to the crystalline state, which is characterized by a periodic atomic structure (see Crystalline State). Thus, glasses are not limited to oxide glasses, and include polymers, other chalcogenides than oxides, and the most recent addition to the family of glasses, metallic glasses. Some of the characteristics associated with oxide glasses are not necessarily found in other glasses. For instance, oxide glasses are usually transparent and brittle, but metallic glasses show a shiny metallic appearance and are ductile as produced. But they share certain properties intrinsic to glasses, such as the presence of a glass transition and structural relaxation behavior. The nature and properties of metallic glasses are better understood by recognizing which properties depend more strongly upon chemical composition, and which properties reflect more sensitively the glassy nature of the atomic structure. This combination of properties found in crystalline metallic solids and those seen in oxide glasses makes metallic glasses scientifically interesting and uniquely useful for various applications. For instance, iron- or cobalt-based metallic glasses are beginning to be widely used as soft magnetic materials in power transformers, transducers, and sensors, because of their high magnetic permeability and resistance to mechanical deformation. High specific mechanical strength coupled with appreciable ductility of aluminum-based metallic glasses makes them attractive for application in aviation. Some chromium-containing glasses show remarkable corrosion resistance and are considered for various applications in hostile environments.

Metallic glasses became known only recently, because they do not form naturally, and it usually takes deliberate measures, such as rapid cooling of the melt or vapor depo-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

sition, to obtain them. Because of the need for rapid extraction of heat during cooling, the dimensions of metallic glasses have been limited for some time to thin films, ribbons, and sheets. However, as a result of the recent discovery of certain compositions that yield metallic glasses even with the cooling rate commonly used for bulk casting, bulk metallic glasses are beginning to become available. Metallic glasses became commercially available in the early 1970s, and intense research activities ensued in the following ten years or so. In the last ten years, there has been marked progress in commercial application of metallic glasses.

1. ATOMIC STRUCTURE OF METALLIC GLASSES

1.1 Method of Structural Characterization

Diffraction measurements of x rays, electrons, and neutrons are commonly used in characterizing the atomic structure of crystalline solids. These methods are also basically effective in characterizing metallic glasses; however, there are some complications, among them the fact that it is not easy to differentiate glasses from microcrystalline solids by diffraction methods alone. When a glasslike substance was first obtained by vapor deposition in the 1930s, researchers thought it was likely to be microcrystalline. Interest in vapor-deposited films resumed after World War II, in the 1950s, and investigators started to explore their physical properties. Since then, there have been long, and often heated, discussions regarding whether metallic glasses are truly glassy or merely microcrystalline. The glassy nature of the metallic glass was confirmed only after many detailed structural studies involving atomic structural modeling and corroboration by thermodynamic evidence.

The intensity of x rays diffracted from a polycrystalline solid as a function of diffraction vector, **q** (= $4\pi \sin\theta/\lambda$, where θ is the diffraction angle and λ is the wavelength of the x ray), shows a number of sharp Bragg peaks. On the other hand, the x-ray intensity diffracted from a glass, including a metallic glass, is continuous with several broad peaks, as shown in Fig. 1. The electron-diffraction pattern from a crystal consists of many Bragg spots, but the electron-diffraction pattern from a glass has several weak diffuse rings. The problem is that, as the size of the crystalline grain becomes small, the Bragg diffraction peak width increases, and it becomes virtually impossible to distinguish true glasses from polycrystalline aggregate with the grain size below 15 Å by diffraction, without a very detailed analysis. The diffracted intensity from a metallic glass is similar to that from a liquid metal, except that the peaks are slightly sharper. This is very strong evidence that the atomic structure of a metallic glass is similar to that of a liquid. As we will discuss below, the structure of a liquid is basically modeled by a dense random packed structure, and the metallic glasses are also described by this model.

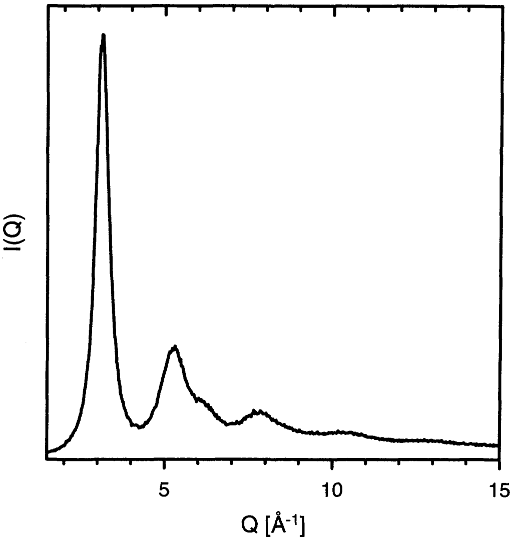

FIG. 1. Intensity of x rays scattered from a metallic glass, as a function of diffraction vector **q**.

1.2 Atomic Pair-Distribution Function

The continuous diffraction intensity, such as the one shown in Fig. 1, is difficult to interpret as is, in contrast to the diffraction pattern from a crystal, which can readily be interpreted in terms of crystal lattice symmetry and lattice parameters. Therefore, it is customary to process the data to extract the atomic interference function, apply the Fourier transformation, and obtain the atomic pair-density, or distribution, function (PDF), as shown in Fig. 2. The PDF, $\rho(r)$, describes

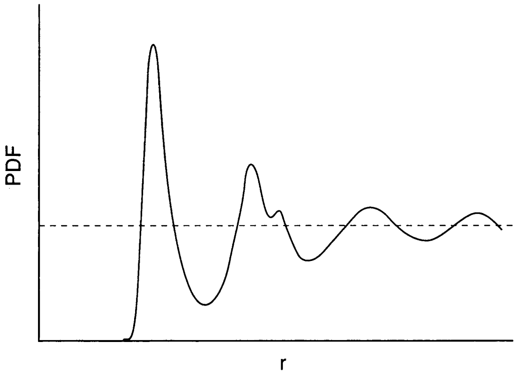

FIG. 2. Atomic pair-distribution functions of a metallic glass, obtained by Fourier transforming the normalized x-ray or neutron-scattering intensity.

the probability of finding an atom at a distance *r*, centering on another atom. The PDF of a metallic glass is characterized by a relatively sharp first peak, which corresponds to the nearest-neighbor atoms, and oscillations that become weaker as the distance increases. By integrating the PDF times $4\pi r^2$ over the first peak, one obtains the number of nearest neighbors, which is usually about 13.

1.3 Structural Models

Since the PDF gives only one-dimensional information, it is necessary to build a three-dimensional model to interpret the PDF further. The first successful models of the liquid structure were proposed simultaneously by Bernal and Scott in 1960, and were named the dense random packing model. The physical model that Bernal made consisted of many steel balls randomly packed in a bag, glued together with wax. The structure they created was clearly distinct from a microcrystalline aggregate, with very few well ordered regions. By analyzing this model, Bernal noticed that local clusters rarely found in crystals, such as an icosahedron (Fig. 3), are abundantly found in this model. Since an icosahedron is characterized by fivefold symmetry, it is not compatible with crystalline periodicity, except in complex multielement compounds. Earlier, in 1952, Frank had already pointed out that, in isolation, an icosahedron is the most stable form of a cluster made of 13 atoms, and that the presence of such clusters stabilizes the liquid structure.

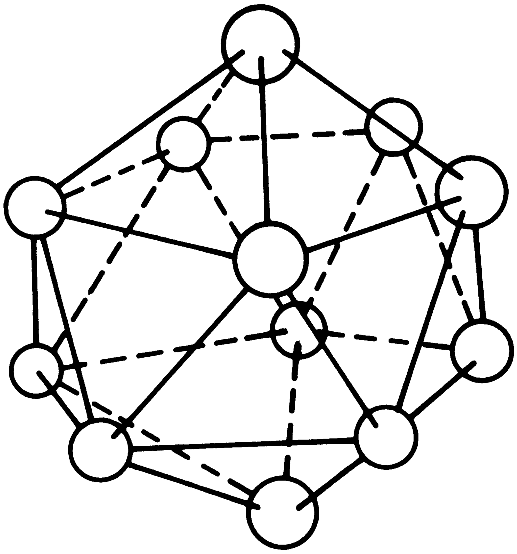

FIG. 3. Icosahedral cluster of atoms.

A solid with the symmetry of the icosahedron was later discovered (1984), and was named a quasicrystal (see QUASICRYSTALS). The atomic structure of a quasicrystal is completely ordered,with two periodicities that are incommensurate to each other, while the structure in a glassy state is highly disordered, with no periodicity. In liquids and glasses, with the constraint of periodicity removed, local structures that are usually not allowed in a crystal but are locally more stable can be found. A glass is formed with the principle of minimizing the local energy when the kinetic constraint does not allow a wide phase space (in statistical mechanics) to be probed by the system, while a crystal is formed by globally minimizing the total energy when the atomic mobility is high, and the system is allowed to look for the free-energy minimum in the entire phase space.

2. THERMODYNAMICS

2.1 Structure of Liquids and the Glass Transition

When a crystalline solid is heated up, the amplitude of lattice vibration is increased, but the average structure itself is unchanged and remains periodic except for slight thermal expansion. In contrast, the topology of the structure of liquid defined by the connectivity of the nearest-neighbor atoms depends upon temperature, so that a change in temperature results in rearrangements of atomic bonds, involving cutting and reforming of bonds. These rearrangements cannot be made instantaneously, so that a finite relaxation time is required for a liquid to assume the equilibrium structure when the temperature is changed abruptly. This structural relaxation time is proportional to the viscosity, and, thus, it is inversely related to the diffusivity. It is a strong function of temperature, and it rapidly increases when a liquid is deeply supercooled below the melting temperature. At a particular temperature, usually 40% to 70% of the melting temperature, the viscosity becomes so high that the system starts to behave like a solid. This temperature is the glass transition temperature T_g, and is usually defined by the viscosity reaching 10^{13} P. The glass transition, however, is not a phase transition, but is merely a thermal arrest phenomenon. If a liquid is cooled at a constant cooling rate, initially the structure of the liquid continuously keeps changing to the equilibrium structure at each temperature. However, the relaxation time increases quickly as the temperature is lowered, and eventually the structure becomes unable to follow the changing temperature, resulting in the thermal arrest phenomenon.

2.2 Kinetics of the Liquid-to-Crystal Transition

It appears that below the melting temperature, the crystalline state is always the state with the lowest energy, and the glassy state is metastable, although a theorem stating this has not been proven. Thus, when a liquid is slowly cooled, it transforms to a crystalline state. However, the boundary between liquid and crystal is well defined, having the thickness of a few atomic distances, and has an energy higher than either liquid or solid. Therefore, in order for a crystalline particle to form in a liquid, this interfacial energy barrier has to be overcome; consequently, it takes a finite amount of time before a crystalline particle is formed, even below the freezing point. An example of the time-temperature-transformation (T-T-T) diagram describing this transformation kinetics is shown in Fig. 4. Here the alloy is in the liquid state at $t = 0$, after the melt is cooled to temperature T infinitely fast. The C curve indicates the boundary between the liquid and crystalline states. On the left of the C curve, the material remains a supercooled liquid. If cooling is fast enough, it is possible to suppress the nucleation of crystalline particles altogether, through reducing atomic motion by bringing the liquid well below its glass transition temperature, and to form a solid-like state without a liquid-to-crystal transition. The solid thus produced is a glass. A metallic glass of Au-Si was first obtained by rapid quenching of the melt by Duwez and his students in 1960 (Klement *et al.*, 1960). They then discovered that various alloys of transition metals and metalloids can be quenched into a glass. It is interesting to note that the principle of rapid cooling discussed above applies in general. In other words, if one could cool the liquid fast enough, any material could be made glass. Turnbull (1969) was the first to recognize this intriguing fact, and established the theoretical basis for the existence of a metallic glass.

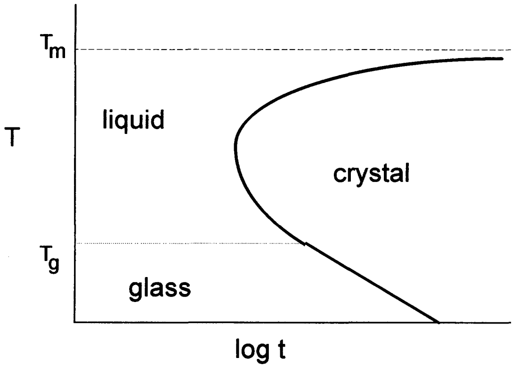

FIG. 4. T-T-T (Time-Temperature-Transformation) diagram for liquid-to-crystal transformation. The horizontal axis is the time a liquid takes to transform itself into a crystal, on a logarithmic scale. By circumventing the nose of the *C* curve by rapid cooling, glass can be obtained.

2.3 Glass-Forming Composition

While, in principle, any material can be made into a glass, the critical cooling rate necessary to form a glass varies greatly from one system to another. For a pure metal, the critical rate is estimated to be over 10^{12} K/s, a rate virtually impossible to achieve except by vapor deposition. Also, a glassy pure metal is so unstable that it crystallizes well below room temperature. Thus, only alloys over certain composition ranges can form a glass that is stable at room temperature by rapid cooling from the melt with the cooling rate of the order of up to 10^6 K/s. Various factors have been considered to influence the glass-forming ability, but the size difference among the constituent atoms seems to be the most dominant factor. The glass-forming composition range is also characterized by the eutectic phase diagram. Near the eutectic composition, liquids are relatively more stable, and thus glasses are more stable as well.

The relative stability of glass can best be understood by considering the so-called T_0 line in the phase diagram (Fig. 5). The T_0 line defines the composition range within which a liquid is the phase with the minimum free energy, if the phase separation is suppressed by rapid cooling. Figure 6 shows the free-energy curves for two solid phases and a liquid, defining the T_0 composition. Note that the free energy is the lowest for a liquid in the composition range from x_1 to x_2, provided the material remains a single phase. If phase separation is allowed by diffusion, the mixture of α and β phases has a lower energy than a liquid. By bringing the liquid within the T_0 lines below the glass transition temperature, a glass can be obtained with a cooling rate high enough to suppress diffusion and phase separation. Typical glass forming compositions are transition-metal–metalloid alloys, such as $TM_{100-x}M_x$, where TM is an alloy of Fe, Ni, Co, and Cr, M is a metalloid element, such as B, Si, C, and P, or their mixture, and $15 < x < 25$; transition-metal alloys, such as $Zr_{100-x}Cu_x$, where $40 < x < 75$; and aluminum–transition-metal alloys, such as $Al_{100-x}Fe_yCe_{x-y}$, where $9 < x < 15$. The exact range of glass-forming composition depends on the details of the production method, but, since the critical cooling rate very strongly depends upon composition, the glass-forming ranges determined by various researchers are remarkably similar. Figure 7 shows one of these composition ranges for the Fe-B-Si alloy system. Glass forming is easiest at the center of the range shown. The glass transition temperature for these compositions weakly depends on the composition, and is about 300–450 °C.

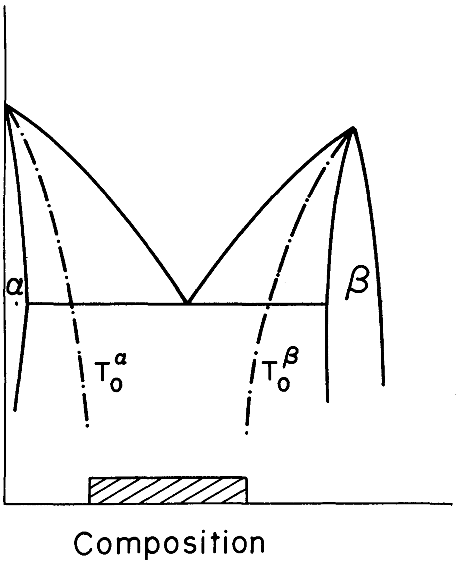

FIG. 5. A binary phase diagram showing the T_0 line (chained curves). The α phase is a solid solution of *B* atoms in the *A* phase, and the β phase is a compound of *A* and *B*. Hatched range indicates glass-forming composition.

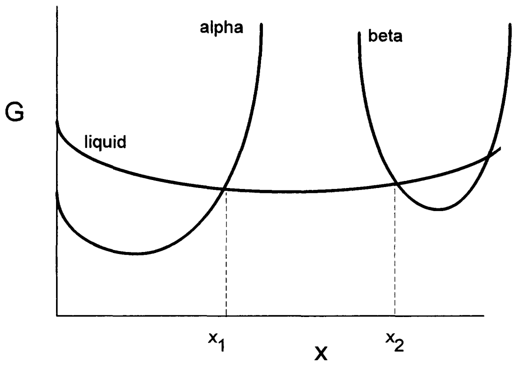

FIG. 6. Free-energy diagram of a binary system. The figure shows the free energy *G* of various phases (α, liquid, β) vs composition *x* at temperature *T*. Between x_1 and x_2, liquid (glass) is the most stable phase, except for the mixture of the α and β phases. The loci of x_1 and x_2 at various temperatures define the T_0^α and T_0^β lines.

2.4 Crystallization and Structural Relaxation

Once a glass is formed, it is stable well below T_g. However, if the temperature is raised up to the vicinity of T_g, the atomic mobility becomes high enough to nucleate crystalline particles, and the glass crystallizes. Since this

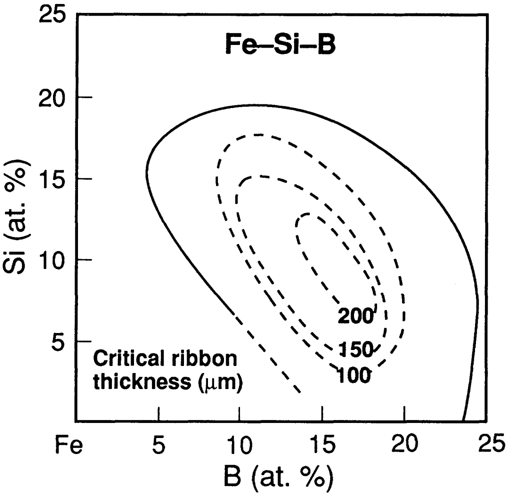

FIG. 7. Glass-forming range and crystallization temperature for Fe-B-Si alloys (inside the solid curve, after Waseda *et al.*, 1990). Numbers denote the maximum thickness (in microns) of the ribbon that can be cast as metallic glass.

is just a liquid-to-crystal transition taking place at a low temperature, the kinetics is basically the same as that for the crystallization near the freezing (solidification) temperature. The crystallization temperature T_x is defined usually by the volume fraction of the crystalline portion reaching 0.5 for a constant heating rate. However, since crystallization is a kinetic phenomenon, T_x depends upon the heating rate. The temperature dependence of T_x is Arrhenian, with the apparent activation energy ranging from 2 to 5 eV. The value of T_x shows moderate composition dependence.

A glass produced by rapid cooling retains the structure that it had at a temperature, called the fictive temperature, where thermal arrest has happened during cooling. Therefore, the higher the cooling rate, the higher the fictive temperature. When a glass with a high fictive temperature is annealed, or held at an elevated temperature, the structure gradually reverts to a more stable structure, and its fictive temperature approaches the annealing temperature. This structural relaxation occurs even well below the glass transition temperature, and is accompanied by various changes in the properties, including magnetic as well as mechanical properties. Incidentally, the structural relaxation is not an incipient crystallization. It results in a more stable glass that is structurally totally distinct from the crystalline state.

It is useful to recognize the distinction between topological short-range ordering and compositional short-range ordering during structural relaxation. The topology of the structure described by its fictive temperature cannot attain equilibrium, except for the immediate vicinity of T_g. Thus, additional annealing treatments are always additive. However, the compositional short-range order among similar atoms, for instance, among the transition-metal elements, can reach an equilibrium in a relatively short time, and can exhibit reversible behavior.

3. METHODS OF PRODUCING METALLIC GLASSES

3.1 Rapid Quenching Methods

For most glass-forming compositions, a cooling rate of the order of 10^6 K/s is required to form metallic glasses from the melt, suppressing crystalline nucleation. In order to achieve such a high cooling rate, various methods have been developed to extract heat from the melt by contact to a chill block. In each case, since heat has to be extracted in a very short time, the dimension of the solid is limited to a small one, typically below 50 μm. The method most widely used and most successful is the melt-extraction technique shown in Fig. 8. In this method, the molten liquid metal alloy is pushed out of a nozzle by gas

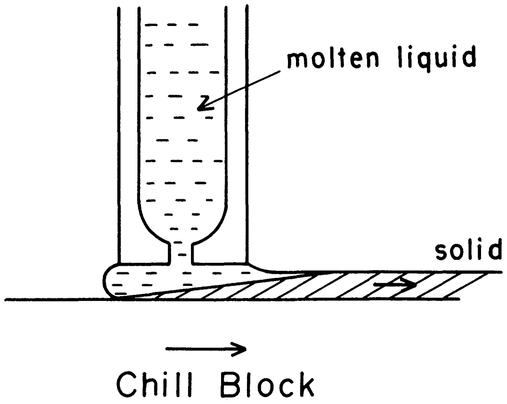

FIG. 8. Melt-extraction method. The chill block is usually a copper wheel rotating with a surface speed over 30 m/s.

pressure, forms a layer between the nozzle and the chill block, a rotating metallic disk, and is continuously extracted in the form of a ribbon by the rotating disk. A ribbon about 1 cm wide, 30 μm thick can be easily produced by this method in the laboratory. By fabricating a wide, flat nozzle, much wider sheets, up to half a foot wide, can be produced in industrial plants. Recently, certain ranges of composition, such as $(Ti,Zr)_{55}(Ni,Cu)_{25}Be_{20}$, were found to have a very low critical cooling rate comparable to those of oxide glasses, and, therefore, not to require special rapid quenching to form a glass (Peker and Johnson, 1993). They can be cooled into a bulk glass by regular casting.

3.2 Solid-State Amorphization

Metallic glasses can also be obtained without going through a liquid phase. For instance, alloys subjected to radiation damage by particles such as electrons and ions can be amorphized. Most interestingly, a two-phase crystalline material annealed at a moderate temperature can achieve amorphization through interdiffusion. An example is the multilayered thin film of Ni-Zr, which becomes amorphous when annealed at 350 °C for several hours. Furthermore, mechanical alloying can also achieve amorphization. By mixing small particles of two kinds of crystalline phase using the ball-milling machine for a long time, it is possible to force the two components to become atomistically mixed, and, if the composition is right, to become vitreous. In each case, the purpose of solid-state amorphization is to achieve a single-phase material with the composition between the T_0 compositions shown in Fig. 5. Since the most stable phase in this composition range is a glass, amorphization occurs. By this technique, bulk metallic glasses that cannot be obtained by rapid cooling methods can be produced.

3.3 Physical and Chemical Deposition

The method first used in producing metallic glasses was vapor deposition in vacuum at low temperatures. By depositing atoms one by one onto a cold substrate, a very high quenching rate can be achieved, because an atom is cooled in a time scale comparable to that of the lattice vibration (10^{-12} s). Vapor of metallic alloys can be produced by various methods, such as evaporation, ion sputtering, and electron-beam heating. Since the deposition rate is usually of the order of 10^{-2}–$\sim 10^{+1}$ nm/s, only thin films ranging from 10 nm to 10 μm can be easily produced by this method, but thicker materials can be deposited by special high-rate deposition systems.

Electrodeposition (plating) or electroless chemical deposition sometimes can produce amorphous alloys, for instance, of transition metal and phosphorus. However, it is difficult to control the homogeneity of the product, and the films produced often contain hydrogen. Thus, electroplating is not the most preferred method of producing metallic glasses.

4. PROPERTIES

4.1 Mechanical Properties

What makes the metallic glasses so distinct from oxide glasses is their ductility. When a stress is applied, metallic glasses can yield and plastically deform. Oxide glasses have no ductility, so that small cracks grow fast when a stress is applied, resulting in brittle behavior. Metallic glasses are ductile, so that the stress concentration at the tip of the crack is diffused by local deformation, stopping the propagation of the crack. Thus, they show tensile strength comparable to the strongest steel (piano wire), ranging from 3000 to over 4000 MPa for transition-metal–metalloid glasses. Al-based alloys, such as $Al_{90}Ce_3Fe_7$, have a lower strength, about 1000 MPa, but show the highest specific strength (strength divided by the density) of all known metallic materials, reaching 30 $(kg/mm^2)/(g/cm^3)$ and above (He *et al.*, 1988). However, particularly in the case of iron-based alloys, much of the ductility is often lost when they are annealed and structural relaxation takes place. Also, unlike crystalline alloys, they show no work hardening, so that, as soon as they start to yield, they fail, often catastrophically in the case when tensile stress is applied. Therefore, special care has to be exercised in using these glasses in applications requiring strength. Deformation occurs uniformly when the temperature is close to the glass transition temperature; however, at lower temperatures, deformation is localized to form shear bands.

Metallic glasses show a fair amount of internal friction and sound attenuation above about 200 °C. However, the room-temperature sound attenuation is surprisingly small, making them attractive for certain applications involving ultrasonic propagation.

4.2 Magnetic Properties

Metallic glasses containing a substantial amount of Fe and Co are usually ferromagnetic. Most of the basic magnetic properties, such as the saturation magnetization and the Curie temperature, depend primarily upon composition and, thus, are similar in the crystalline and amorphous states. For instance, the magnetic moment of transition metals and their alloys is determined by the difference in the spin density in the majority-spin (for instance, up-spin) electron band and that in the minority-spin (for instance, down-spin) electron band. Since the majority-spin band of cobalt is full, and the magnetic moment depends only upon the number of holes in the minority-spin band (this case is called strong magnetism), the magnetic moment of cobalt depends only on composition, and is thus independent of the structure. On the other hand, both the majority- and minority-spin bands are not full in iron (this case is called weak magnetism), so that the magnetic moment of iron does depend somewhat on the structure, varying when the structure changes from crystalline to amorphous. For the transition-metal–metalloid (TM-M) glasses, if the composition of TM is changed while keeping the composition of M fixed, the magnetic moment is a linear function of the *d*-electron density, as shown in Fig. 9. Such a curve for TM crystalline alloys, shown by the dotted curve in the figure, is known as the Slater-Pauling curve. The curve for the TM-M glasses is slightly shifted because of the effect of metalloid elements.

Secondary magnetic properties, such as magnetic anisotropy and magnetostriction (change of dimension due to magnetization), reflect the atomic structure a little more sensitively. A glass in an ideal state should be isotropic, but real glasses are not quite isotropic because of directional solidification or mechanical deformation. Thus, real ferromagnetic metallic glasses usually have magnetic anisotropy. In particular, sputter-deposited amorphous-iron–rare-earth alloys, such as $Fe_{75}Tb_{25}$, show strong anisotropy as deposited. If the deposition condition is right, the anisotropy can be strong enough to keep the magnetization perpendicular to the film, even though magnetostatic energy prefers it to be parallel to the plane of the film. This anisotropy is critical in the use of these alloys as magnetic recording media. Also, magnetic anisotropy can be induced by annealing either in a magnetic field or under stress. The microscopic origin of this anisotropy has been

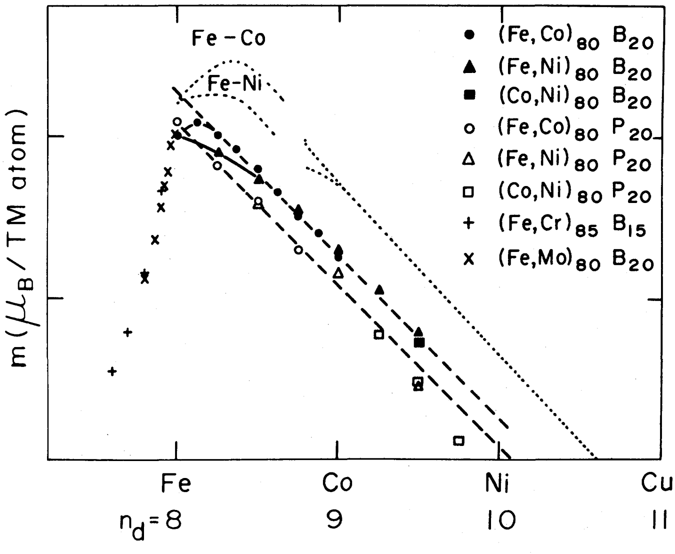

FIG. 9. Dependence of the saturation magnetic moment on the nominal *d*-electron concentration in the transition-metal–metalloid metallic glasses (after Egami, 1984).

the subject of numerous studies. The topology of the structure and compositional anisotropic short range ordering are considered to be responsible.

Low-field magnetic properties, such as permeability, depend more on the microstructure at the length scale equal to the domain-wall thickness than on the atomic structure. Metallic glasses are random and inhomogeneous at an atomic scale, but at the scale of domain-wall thickness (~100 nm), they are much more homogeneous than crystalline alloys, since they do not have any grain boundaries or lattice dislocations. Thus, ferromagnetic metallic glasses usually show high magnetic permeability and are excellent soft magnetic materials. In particular, magnetostriction is zero in certain cobalt-based alloys, making them ideal for soft magnetic applications. An example of the composition dependence of magnetic permeability around the zero-magnetostriction composition is shown in Fig. 10.

However, effects of structural relaxation have to be taken into account in using metallic glasses. Metallic glasses as cast are unstable, and their properties, particularly the low-field properties, are strongly affected by subsequent heat treatments. Thus, it is customary to anneal them beforehand, often in a magnetic field, in order to stabilize their properties. In designing the annealing treatment, a detailed knowledge of the structural relaxation phenomena is required. Many of the magnetic properties, such as the Curie temperature and magnetic anisotropy, are affected by the compositional short range ordering. Thus, their changes during annealing are fast, and they can attain equilibrium during annealing.

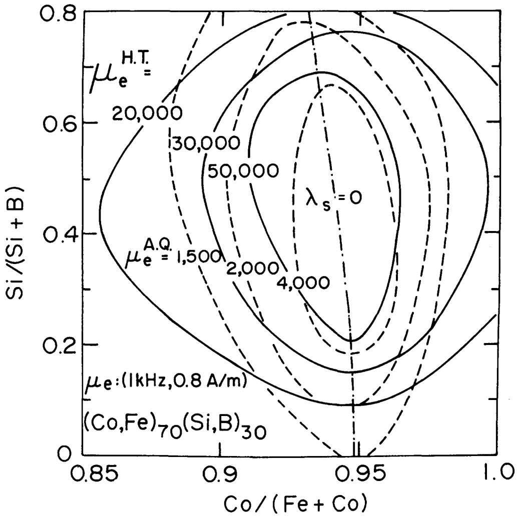

FIG. 10. Composition dependence of magnetic permeability μ (= dB/dH, where B is the magnetic flux density and H is the magnetic field) of $(Co,Fe)_{70}(B,Si)_{30}$ alloys, as quenched (μ_e^{AQ}, dashed curve) or heat treated at 450 °C and water quenched (μ_e^{HT}, solid curve) (after Masumoto and Egami, 1981). The dot-dashed line indicates the zero-magnetostriction composition.

Not all amorphous films show high magnetic permeability. Sputter-deposited thin films of transition-metal–rare-earth alloys, such as $(Fe,Co)_{75}Tb_{25}$, show relatively high coercivity, of the order of 1 kOe, because of high local anisotropy of Tb and the surface effect. Together with the high magnetic anisotropy mentioned above, which keeps the magnetization perpendicular to the film, this high coercivity stabilizes small circular ferromagnetic domains called bubble domains. These bubble domains are used in the magneto-optical data storage application of these films.

4.3 Electrical Properties

In crystalline solids, because of periodicity in the atomic structure, only the electrons with specific wave vectors are scattered (Bragg's law). Consequently, at $T = 0$ K, the electrical resistivity of an ideal metal goes to zero, and only the deviations from the perfect periodicity, lattice vibrations, produce resistance. In metallic glasses, Bragg's law does not apply because of their disordered atomic structure. Thus, all electrons are scattered, resulting in a very short mean free path of only a few atomic distances. Consequently, the resistivity is high even at $T = 0$ K, and is usually of the order of 100 μΩ-cm. Since the resistivity is already high at 0 K, temperature has little effect on resistivity. It can either weakly increase or weakly decrease, depending upon the conduction electron concentration.

Metallic glasses containing $4d$ transition metals such as Zr, Mo, and Nb are superconducting at low temperatures. Their superconducting transition temperature is usually slightly lower than the crystalline counterpart, but the opposite can be true when the transition temperature of the crystalline solid is low. On the other hand, the critical magnetic field is usually higher for superconducting metallic glasses than for crystals. Superconducting metallic glasses provide an interesting proof that periodicity, which is assumed in the conventional theories of superconductivity, is not a prerequisite for superconductivity.

4.4 Chemical Properties

The homogeneity of the structure of metallic glasses at a mesoscopic level, mentioned above, is also important for surface chemical properties. Metallic glasses containing significant amounts (usually more than 5%) of chromium form very uniform surface oxide layers that inhibit corrosion. The presence of phosphorus, in addition to chromium, is also known to be helpful. Thus, Cr-containing metallic glasses show excellent corrosion resistance. They also show interesting catalytic activities and are promising as catalysts, as well.

5. APPLICATIONS

5.1 Magnetic Applications

As has been mentioned several times above, metallic glasses containing Fe or Co are used in several applications, including uses in electrical power-distribution transformers, in high-frequency electrical power supplies, in sensors (for instance, in antitheft devices), in magneto-optical data storage media, and as transient materials for permanent magnets.

Wide ribbons of rapidly cooled Fe-B-Si alloys, such as $Fe_{78}B_{13}Si_9$, are now beginning to be used as core materials in many electrical power-distribution transformers. Since their magnetic permeability is high, their magnetic core loss is small, resulting in considerable energy saving. Currently, the power loss due to the distribution transformer with crystalline silicon-iron core amounts to about 1% of the total power generated, or about 4×10^9 dollars a year in the United States. Replacing these conventional distribution transformers with amorphous-core transformers will reduce this loss to a half, resulting in enormous savings. In order to choose the best alloy composition for this application, several factors have to be considered. One wants to maximize the magnetic flux density and magnetic permeability, as well as thermal stability, but these requirements often conflict with each other. For instance, the magnetization increases as the iron content is increased. However, if the iron content is too high, magnetic permeability and thermal stability start to degrade. Furthermore, economics also interfere in the decision. When the cost of energy is high, as it was right after the start of the Iran-Iraq war in 1983, it makes sense to place priority on reducing the core loss rather than the flux density. However, if the energy cost is not so high, it is better to choose a material with a higher flux density and minimize the size of the transformer to reduce the cost of the transformer. Currently, the preferred alloy system is Fe-B-Si alloys with Fe 78–80 at. %. Magnetic metallic glasses are useful not only in transformers, but also in motors, for the same reason.

High magnetic permeability of cobalt-based zero-magnetostrictive metallic glasses makes them attractive for applications as magnetic shielding materials and magnetic sensors, such as the recording heads for tape recorders. Since the electrical resistivity of metallic glasses is relatively high, and it is easy to produce thin foils of metallic glasses, cobalt-based metallic glasses retain high magnetic permeability up to high frequencies (100 kHz or more), and they are excellent for application in high frequency power supplies. They are used in switching-mode transformers in which ac or dc voltage is changed into high-frequency ac voltage, transformed, and switched back to the original frequency. They are also used in antitheft devices in retail stores, to prevent shoplifting. Their high-frequency response, particularly at very high harmonics because of their square *B-H* loop, makes them excellent tagging materials. High magnetic permeability coupled with reasonable magnetostriction makes some metallic glasses attractive as transducer materials. They are used in various force sensors, for instance, in engine and power trains in automobiles.

Sputter-deposited transition-metal–rare-earth amorphous films are used as magneto-optical data storage media. Small bubble domains of about 1 μm in diameter are produced by locally heating the film by a semiconductor laser. Domains are read by the rotation of the polarization of light reflected from the film (Kerr effect). By this method, a recording density as high as the compact disk can be attained, and yet, unlike the conventional aluminum-based compact disks, the recording can be erased and overwritten. Already several manufacturers are commercially producing optical data storage systems (CD-RAM) using amorphous films.

Metallic glasses are also used in fabricating a crystalline permanent magnet, $Fe_{14}Nd_2B$. Because of a complex phase diagram, it is not easy to produce this compound, which shows excellent magnetic induction and coercivity. Thus, in a technique widely used now, the alloy is first cast as a ribbon in an amorphous state and then annealed into a crystalline permanent magnet.

5.2 Other Applications

The possibility of using metallic glasses in mechanical applications as structural materials was recognized already in the 1960s. While magnetic applications increased rather quickly, mechanical applications remained merely a possibility for some time, mainly because of the price, dimensions, and structural relaxation. Functional materials, such as magnetic materials, can be reasonably costly, but structural materials need to be inexpensive. Also, the thickness of the sheet, usually of the order of 30 μm, severely limited its use. An additional problem is the loss of ductility upon structural relaxation in iron-based glasses.

The recent discovery of glass compositions with very low critical cooling rates for vitrification can change this situation dramatically. With these compositions, a bulk glass can be cast easily and inexpensively. By machining these bulk glasses, complex tools, such as gears, can be produced. It is most likely that they will find major markets in the near future.

Metallic glasses are found in other applications, for instance, as brazing alloys. Brazing alloys, usually in the form of foils, are used in joining two metallic blocks upon heating that melts the brazing alloy. Since metallic glasses are ductile and can be cut into various forms, they are convenient in joining pieces with complex shapes. When the joining is done, the glasses are crystallized. Mainly nickel-based alloys are used for this purpose.

6. SUMMARY

While metallic glasses belong to the family of glasses, their properties are only in limited cases similar to those of oxide glasses; in some other cases, they are closer to those of crystalline alloys, and in other cases, they resemble neither. This combination of properties makes the metallic glass an interesting scientific subject of study, and offers unique opportunities for industrial and commercial applications. While the structure of metallic glasses is disordered at an atomistic scale, at a scale slightly larger than the atomistic scale, the structure is more homogeneous than crystalline solids because of the absence of grain boundaries and lattice dislocations. This gives rise to interesting mechanical as well as low-field magnetic properties, which are attractive to applications. Metallic glasses are beginning to be used widely as soft-magnetic materials, and their use as structural materials is likely to grow. They are clearly one of the advanced materials that are important in technology today and tomorrow.

Works Cited

Egami, T. (1984), *Rep. Prog. Phys.* **47**, 1601–1725.

Frank, F. C. (1952), *Proc. R. Soc. London, Sec. A* **215**, 43–46.

He, Y., Poon, S. J., Shiflet, G. J. (1988), *Science* **241**, 1640–1642.

Klement, K., Willens, R. H., Duwez, P. (1960), *Nature* **187**, 869–870.

Masumoto, T., Egami, T. (1981), *Mater. Sci. Eng.* **48**, 147–165.

Peker, A., Johnson, W. L. (1993), *Appl. Phys. Lett.* **63**, 2342–2344.

Turnbull, D. (1969), *Contemporary Phys.* **10**, 471–488.

Waseda, Y., Ueno, S., Hagiwara, M., Aust, K. T. (1990), *Prog. Mater. Sci.* **34**, 149–280.

Further Reading

Beck, H., Güntherodt, H.-J. (Eds.) (1981, 1983), *Glassy Metals I, II*, Berlin: Springer-Verlag.

Liebermann, H. H. (Ed.) (1993), *Rapidly Solidified Alloys*, New York: Marcel Dekker.

Luborsky, F. E. (Ed.) (1983), *Amorphous Metallic Alloys*, London: Butterworths.

METALLURGY, PHYSICAL

ROBERT W. CAHN, *Department of Materials Science and Metallurgy, Cambridge University, Cambridge, United Kingdom*

Introduction 141
1. **Phase Structure of Metals and Alloys** 142
1.1 Lattice Structure 142
1.2 Phases 143
2. **Equilibrium Diagrams** 143
2.1 The Silver-Magnesium System 144
2.2 The Sodium-Bismuth System 145
2.3 The Copper-Gold System 146
2.4 The Iron-Carbon System 147
2.5 Ternary Alloys 150
2.6 Calculation of Phase Diagrams 150
3. **Instrumentation for Characterization of Alloys** 150
3.1 X-Ray Diffraction 150
3.2 Optical Microscopy 151
3.3 Electron Microscopy 152
3.4 Determination of Local Composition 154
4. **Microstructure and Properties** 154
4.1 Development of Microstructures 154
4.2 Mechanical Properties of Metals and Alloys 156
4.3 Mechanically Induced Phase Transformations 157
5. **Computer Modeling** 159
Acknowledgments 160
Glossary 160
Works Cited 161
Further Reading 161

INTRODUCTION

"Physical metallurgy" is a term coined in 1914 by Walter Rosenhain, of the National Physical Laboratory near London, to denote the application of physical principles and experimental techniques to the understanding, processing, and improvement of alloys. Among the physical experimental techniques applied to metals and alloys, microscopy is crucial. Metallography, an older discipline concerned largely with the microscopic examination of alloys in various stages of heat treatment, is now part of physical metallurgy. The methods and approach of physical metallurgy have in recent years been extended to ceramics also.

The four basic activities of physical metallurgists are

1. the study of equilibrium between phases as a function of composition and temperature;
2. the study of the geometry and development of microstructures;
3. the relation of useful properties, particularly mechanical properties, to the phases present and to their proportions and microstructural disposition; and
4. the rate and the mechanism by which a microstructure and its constituent phases change when the temperature of an alloy is changed.

The central concept of physical metallurgy is that of equilibrium among phases in an alloy; this is expressed graphically by means of equilibrium diagrams (see also PHASE EQUILIBRIA). The mechanism and kinetics of the transition of a phase into one or more other phases, with different crystal structures, when the temperature is changed, is a central study of physical metallurgy. However, alloys are frequently trapped in nonequilibrium, "metastable" states, for instance by rapid cooling from a high temperature. Examination of the transition from a metastable to a stable (equilibrium) state is also an important part of physical metallurgy (see also PHASE EQUILIBRIA).

Metals and alloys are always full of metastable defects that govern their behavior un-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

der load, aggressive environments, or irradiation. The interpretation of such behavior, as well as the examination of the geometry and dynamics of defects as a topic in its own right, forms a third constituent of physical metallurgy.

Finally, but by no means least, physical metallurgists accord a place of honor to the study of microstructure, that is to say, the geometrical and crystallographic disposition of the individual crystals, or grains, of a single phase, and of the distinct phases in a multiphase alloy. The physical and mechanical properties of an aggregate of grains and phases depend in subtle ways on the scale and morphology of the constituents of microstructure; to pick just one variable, grain size affects the strength, ductility, and toughness of any metal or alloy (see also MECHANICAL PROPERTIES OF SOLIDS).

Since World War II, the variety and subtlety of the physical techniques used to examine metals and alloys have been vastly extended. As a result, physical metallurgy has been revolutionized.

1. PHASE STRUCTURE OF METALS AND ALLOYS

Metals and alloys are generally crystalline, though they usually lack the crystalline faces familiar on minerals. The crystal structure of metals and alloys determines their properties, in particular that of anisotropy: the variation of properties along different directions in a crystal. Thus, properties such as elastic stiffness, magnetic permeability, and thermal expansion can differ along different crystal axes. Anisotropy is often disguised in a metallic object because such an object consists of many grains (see Fig. 1) whose crystal axes, randomly oriented, tend to average out the anisotropy in the object as a whole. If, however, the crystal axes have a tendency to line up in a preferred direction, the object is said to have a preferred orientation, or texture, and many properties then differ, for instance, along the length or width of a rolled sheet.

1.1 Lattice Structure

Most elementary metals crystallize with very simple structures, such as cubic or hexagonal close packing, or the body-centered cubic structure in which atoms reside at the corners and center of a cubic unit cell (Fig. 2). Tin, gallium, and uranium are examples of the few metals that have more complex crystal structures with tetragonal or orthorhombic unit cells (see also CRYSTALLOGRAPHY).

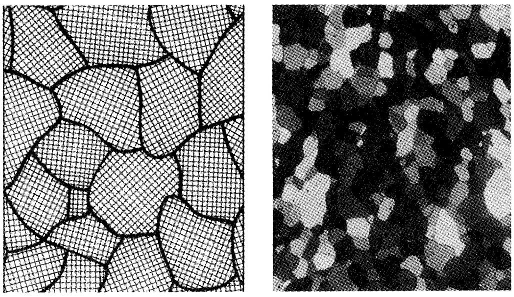

FIG. 1. Polycrystalline metal, consisting of a jumble of tiny crystal grains (typically ≈0.1 mm across). Left, diagrammatic; right, an optical micrograph of a polished and oxidized copper surface, viewed by polarized light (courtesy R. M. J. Cotterill).

When one metal is mixed intimately with another on the atomic level, e.g., by melting together or by mechanical alloying, either the second metal dissolves in the first or the metals form an intermetallic compound. In the first case, a solid solution is created in which the atoms are distributed at random and the mean spacing between them varies, but the crystal structure does not change. An intermetallic compound usually has a crystal structure distinct from that of either constituent metal, and its atoms are usually distributed in an ordered manner.

Intermetallic compounds mostly have simple formulas such as CuZn, Ni_3Al, or

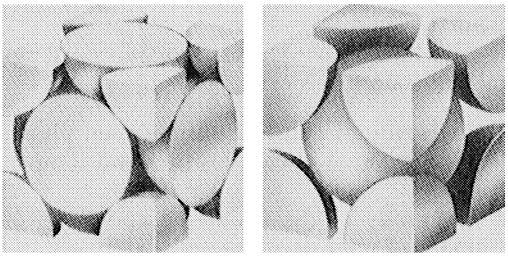

FIG. 2. Simple metallic crystal structures: left, a face-centered cubic, or cubic close-packed structure (representing Al, Cu, Au, etc.); right, a body-centered cubic structure (representing Na, Mo, the low-temperature form of iron, etc.) (courtesy R. M. J. Cotterill).

Cu_3Sn. Some are rigorously limited to their ideal compositions; many others, however, have a considerable range of possible compositions, or homogeneity range. In the early days of physical metallurgy, when the nature of chemical bonding was as yet mysterious, the two types of intermetallic compound were known as daltonides and berthollides, terms referring back to two early chemists: John Dalton, the 18th-century proponent of the atomic basis of chemical combination, and Claude Berthollet, who was skeptical of this concept. The bonding of metals with each other is now much better understood, and the terms are falling into desuetude. (The International Union of Pure and Applied Chemistry has now recommended that these antiquarian terms no longer be used!) A berthollide of nonideal composition is in effect a solid solution based on a compound of ideal composition; a solid solution based on a pure metal is known as a terminal solid solution.

1.2 Phases

An alloy may contain regions, in juxtaposition, of two or more crystal structures, each of a different composition. Typically, one region will consist of a terminal solid solution, and others will contain one or more intermetallic compounds. The occupants of such regions are called *phases*. A particular phase has its own characteristic crystal structure, which defines it. The composition of a phase can vary according to the composition of the alloy as a whole, but in equilibrium the composition, in a particular alloy, will be uniform throughout that phase. In a partly or wholly molten alloy, the liquid region is also a distinct phase, and so is the gas above the alloy itself. A phase can be regarded as the central entity in physical metallurgy.

The number, lattice structures, composition ranges, and relative proportions of the phases of an alloy, studied as functions of the temperature and overall makeup of the alloy, are the basis of phase or equilibrium diagrams. To these we turn next.

2. EQUILIBRIUM DIAGRAMS

The notion of a stable combination of phases for a material of specified composition at a given temperature and pressure goes back to the work of the American physicist Josiah Willard Gibbs (1839–1903). Gibbs recognized that, in principle, even a pure element might adopt a number of different crystal structures. Of these, the structure with the lowest free energy would be that adopted in equilibrium. The free energy, or Gibbs free energy, G, is based upon two subsidiary concepts: internal energy and entropy.

The internal energy H of a gram of metal can be thought of as the sum of the energies of all the bonds between neighboring pairs of atoms in that gram. The internal energy depends upon crystal structure because both the number of atom pairs and the distances between neighboring atoms will vary from structure to structure. (By convention, the most firmly bonded collection of atoms has the lowest internal energy.)

However, the structure with the lowest internal energy will not necessarily be adopted by the metal at all temperatures. This is because the thermal vibrations of the atoms will affect the strength of bonding: the higher the temperature, the greater the vibration and the weaker the bonding. The magnitude of these vibrations is expressed by a quantity termed the entropy S, most conveniently regarded as a measure of disorder. The free energy G is given by the relation $G = H - TS$, where T is the temperature on the absolute or Kelvin scale. In a solid solution (see above), the entropy is affected not only by the thermal vibrations of the atoms but also by the number of different ways in which dissimilar atoms can be distributed in the crystal structure. This number, which is an aspect of disorder, becomes greater as the solid solution becomes more concentrated.

The relation $G = H - TS$, combined with Gibbs's conclusion that the lowest possible value of G determines which crystal structure is stable, indicates that internal energy and entropy are in permanent competition in physical metallurgy. An equilibrium diagram indicates how this competition works out for two or more metals in various proportions at various temperatures. The principles underlying this idea will be made clear by considering the equilibrium diagram for each of three very different alloy systems.

A metastable phase (a concept introduced in the Introduction) is one that has a free energy higher than the phase or phase combination that would be stable for a specific alloy composition and temperature but has been created by some process of rapid or forceful change such as quenching from a high temperature: The metastable phase is then trapped because at ambient temperature, the atoms are incapable of changing places as would be necessary to change the crystal structure and local composition to that (or those) of the state(s) that would be in equilibrium. Metastability is a central concept in understanding the genesis of commonly found alloy structures.

The concepts briefly touched on here are dealt with in more detail in the articles on THERMODYNAMICS, EQUILIBRIUM and THERMODYNAMICS, NONEQUILIBRIUM, and particularly in PHASE EQUILIBRIA.

2.1 The Silver-Magnesium System

The equilibrium diagram for this combination of metals is shown in the lower part of Fig. 3. The vertical scale is temperature, and the horizontal scale is atom fraction of magnesium in the silver-magnesium mixture. The diagram shows five single-phase fields: at high temperatures, liquid; at lower temperatures, the α field, a range of terminal solid solutions based on the cubic close-packed silver crystal structure; the γ field, based on the hexagonal close-packed magnesium crystal structure; the β′ field, a [berthollide] intermetallic compound centered on the formula AgMg with a simple cubic structure of Ag atoms at cube corners and Mg atoms at cube centers; and the ϵ phase, a narrower berthollide field based on $AgMg_3$, the structure of which has a more complicated hexagonal unit cell. These single-phase fields are separated by two-phase fields in which the adjacent phases coexist, at compositions defined by the boundaries on either side and varying with temperature, and in relative proportions depending on the point within the field that defines the temperature and the overall composition. [Thus, point *P*, in the β′ + ϵ field, defines an alloy with 0.6 atom fraction of Mg, equilibrated at 600 K (327 °C).] An alloy such as this illustrates the fact that overall alloy composition is quite distinct from the composition of the equilibrium phases where two or more of these exist.

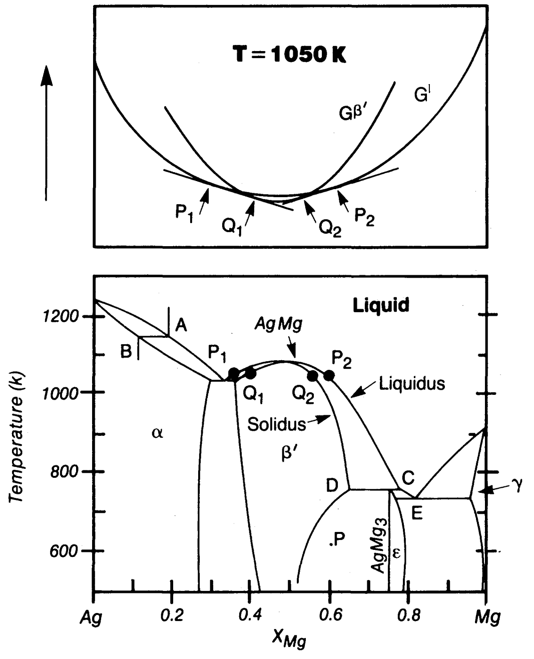

FIG. 3. Silver-magnesium equilibrium diagram, with corresponding free-energy diagram for 1050 K (777 °C) above. (*G* represents the Gibbs free energy.)

An alloy melt usually cannot freeze without a change in composition. Thus, when a melt of composition A reaches the liquidus, it will initially freeze as a solid of composition B, defined by the solidus. This process alters the composition of the remaining melt, enriching it in magnesium, so that it freezes at successively lower temperatures, producing a solid of steadily changing composition, with an onionlike configuration. Such an alloy is said to be "cored," or "microsegregated"; it cannot be in true equilibrium until migration of the atoms in the hot solid has evened out the compositional gradient so that the entire solid returns to the uniform composition A. If, however, the frozen solid is promptly cooled to ambient temperature, the atoms are unable to move, and the structure remains permanently cored. This is a simple example of a metastable structure, in the sense that the material is not in equilibrium and is yet, for purely kinetic reasons, unable to reach that state.

The first intermetallic compound, at its ideal composition, AgMg, freezes "congruently," without change of composition—that is to say, the liquidus and solidus coincide at this alloy composition. This often serves to define the ideal composition of a berthollide. The ϵ phase, however, freezes out by a more complex process termed peritectic solidification, which involves a reaction between a liquid of composition C and a solid of composition D.

Finally, at the right-hand side of the diagram, we have a "eutectic" solidification process. In this, an alloy of overall composition E freezes to form an intimate two-phase mixture of ϵ phase and Mg-rich terminal solid solution, a so-called eutectic mixture, solidifying at a constant eutectic temperature. If the alloy as a whole contains a little more or less Mg than does composition E, then one or the other of the two phases mentioned above freezes out first until the residual liquid achieves composition E, and the eutectic mixture then forms. The result is a curious two-phase mixture consisting (in the case of magnesium excess) of regions of coarse magnesium-rich solid solution interspersed with regions of the fine eutectic mixture, so that the same phase exists in two different microstructural forms.

The upper diagram of Fig. 3 shows the estimated variation of the Gibbs free energy G, at a constant temperature of 1050 K (777 °C), with phase composition for the liquid and β' phases. For a given overall alloy composition, the stable phase structure is always that corresponding to the lower value of G, except in those composition ranges where a common tangent to the two curves lies below both curves, such as the regions P_1Q_1 and P_2Q_2. Here, a two-phase mixture will have a lower total Gibbs free energy than if the alloy consisted exclusively of one or the other phase. All other portions of the equilibrium diagram can be likewise interpreted.

2.2 The Sodium-Bismuth System

This system (Fig. 4) exemplifies well the circumstances that lead to the appearance of extreme daltonide phases, of vanishingly small homogeneity ranges. Two such phases exist in the sodium-bismuth system: Na_3Bi and NaBi. Very narrow (daltonide) phase fields are often referred to as line compounds, because the single-phase fields have to be drawn as single lines in the equilibrium diagram. There is also a eutectic, in which both constituent phases, NaBi and Bi, have vanishingly small mutual solubilities.

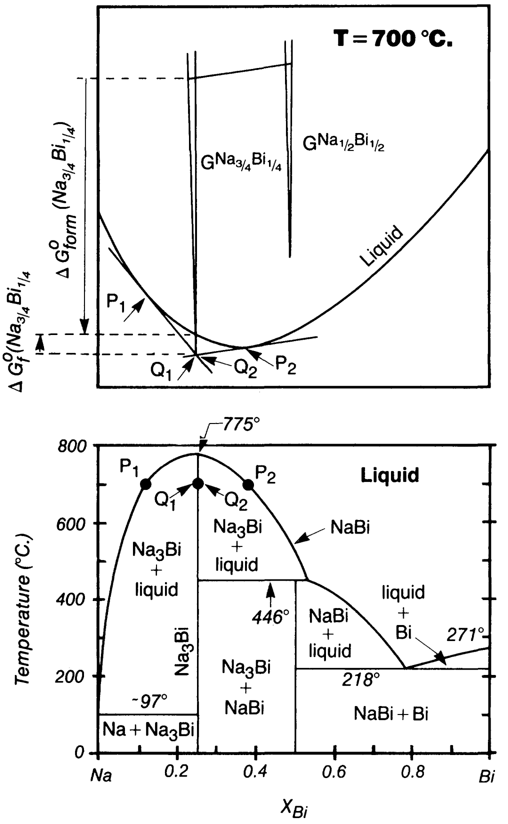

FIG. 4. Sodium-bismuth equilibrium diagram, with corresponding free-energy diagram for 700 °C above. (G represents the Gibbs free energy.)

The upper diagram shows the variation of the Gibbs free energy for a line compound at a fairly high temperature. Very small deviations from the ideal composition lead to very sharp increases in the Gibbs free energy. This happens because such phases have extremely strong bonds between the neighboring Na and Bi atoms but weak bonds between pairs of identical atoms. The structure of compounds of the sodium-bismuth type is such that there are no, or very few, atoms of the same kind as next neighbors. When there is a small excess of either Na or Bi, then additional Na-Na or Bi-Bi neighbor pairs necessarily exist. This gives rise to a sharp increase of the internal energy H, with but a small change in entropy S (because there are so few "wrong" atoms present); hence, at a fixed temperature, the free energy G rises sharply. In the upper diagram, the contact points of the tangents referring to the phase Na_3Bi at the points Q_1 and Q_2 are so close that the homogeneity range of the phase is of negligible width.

2.3 The Copper-Gold System

This system, often regarded as a metallurgical classic, illustrates a further consequence of the arguments presented in the preceding paragraph.

Figure 5 shows a section of the Cu-Au equilibrium diagram in simplified form. Cu_3Au is a berthollide compound, the crystal structure of which is shown in Fig. 6(a). When this compound in its equilibrium state is slowly heated, it undergoes (at a critical temperature T_c = 390 °C) an abrupt transformation to a state in which the copper and gold atoms are distributed at random on the sites of a face-centered cubic structure. The transformation is accordingly called an order-disorder transition. On further heating, the disordered form remains in equilibrium until it reaches the melting temperature. If the compound is of a nonstoichiometric or nonideal composition, for example $Cu_{78}Au_{22}$, then T_c falls to a lower value, ≈350 °C, and a narrow region exists in which the ordered and disordered forms of the compound, of slightly different compositions, coexist in equilibrium.

An ordered form such as that of Cu_3Au exists because it has as many unlike (Cu-Au) nearest neighbors in its crystal structure as are compatible with the formula. Such neighbors have relatively stronger bonds, implying a lower overall internal energy H of the ordered form.

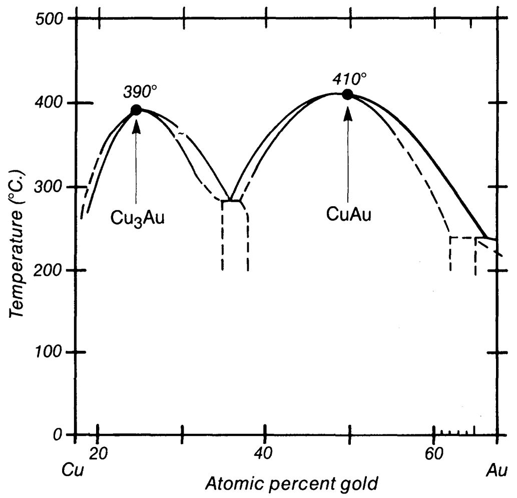

FIG. 5. Copper-gold equilibrium diagram (partial, simplified). The dashed lines represent phase boundaries that have not been located with certainty.

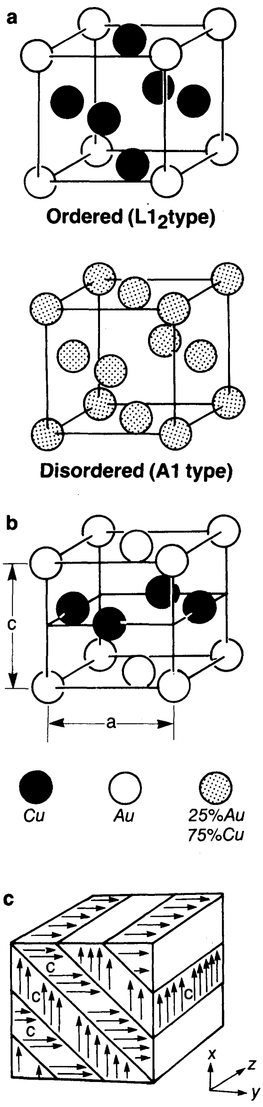

FIG. 6. Crystal structures of (a) ordered and disordered Cu_3Au and (b) ordered CuAu. (c) Schematic of an array of CuAu domains (two out of three possible orientations) in what was originally a single disordered crystal.

An order-disorder transition occurs because the TS in the expression $G = H - TS$ rises with rising temperature until it becomes advantageous—in terms of free energy—for the alloy to assume disorder, which entails a greatly enhanced entropy. In this disorder, a large number of "like" atom pairs (Cu-Cu and Au-Au) come into existence. These pairs have relatively weak bonds, and *H* rises. This rise is more than compensated by the sharp rise in configurational entropy resulting from the random distribution of the Cu and Au atoms.

The disappearance of ferromagnetism when a piece of iron is heated above its Curie temperature is an analogous phenomenon. The individual atomic magnets in the iron behave much as do the Cu and Au atom pairs in Cu_3Au (see also MAGNETIC ORDERING IN SOLIDS).

To explain why the transition can be abrupt (as is the case for Cu_3Au) requires the techniques of statistical mechanics, the theory of the collective behavior of large aggregates of atoms (see STATISTICAL MECHANICS, CLASSICAL). When the critical temperature is reached, a kind of avalanche takes place: The more atoms disorder, the easier it becomes for the remainder to follow. Statistical mechanics can make sense of structural transitions, given information about relative bond strengths; it cannot interpret the fact that Cu-Au bonds are stronger than Cu-Cu and Au-Au bonds. To do that, it is necessary to apply the theory of quantum mechanics. Both classical statistical mechanics and quantum mechanics (*q.v.*) are important in the study of modern physical metallurgy.

When more gold is added to the alloy to reach the composition CuAu, another order-disorder transition is observed. The ordered form of this berthollide compound is shown in Fig. 6(b). Ordered CuAu has a tetragonal structure consisting of successive sheets of Cu only and Au only. Counting shows that this structure has more unlike nearest neighbors than does the disordered (face-centered cubic) structure. This kind of order-disorder transition, associated with a change of symmetry and unit cell dimensions, leads to interesting complications. The disordered cubic form has three crystallographically equivalent cube axes at right angles to each other. Any of these axes can turn into the unique tetragonal *c* axis, so that a single disordered crystal can turn into an array of so-called domains, each of which has a *c* axis oriented differently from its immediate neighbors; this is shown schematically in Fig. 6(c). Since the *c* axis of an ordered domain is shorter than the cube edge of a disordered unit cell, ordering will impose localized strains on the alloy. One consequence is that the configuration of domains [Fig. 6(c)] is optimized so as to minimize the total elastic strain energy, because differently oriented domains relieve each other's elastic strains to some degree. The other consequence is that an external stress will modify the preferred pattern of domains and in the extreme case will ensure that an entire crystal consists of a single domain. The role of stress in determining structure has increasingly been recognized in recent research in physical metallurgy.

Alloys near the CuAu composition, and containing small amounts of other noble metals, are used for dental crowns and bridges. When quenched from a high temperature, such an alloy is metastably disordered and in this state is soft. When heated at 200 to 300 °C, the alloy orders and assumes a complex domain structure that is much harder. In this way, dental crowns can be strengthened.

Further discussion of the categories of binary phase diagrams and their relationship to the compositional and temperature dependence of the Gibbs free energy will be found in the article PHASE EQUILIBRIA.

2.4 The Iron-Carbon System

The iron-carbon system is of crucial industrial importance. Figure 7 shows the accepted version of the phase diagram for this system. The diagram, in which the metastable phase Fe_3C, cementite, appears, is appropriate to the study of steels. In cast irons, which contain large carbon concentrations, carbon forms as the most stable phase, that is, elemental carbon (graphite). In steels, carbon forms so sluggishly that cementite, which forms easily, is found instead. Figure 7 is thus an example of a metastable phase diagram.

The essential point to grasp concerning the heat treatment of simple (that is, iron-carbon) steels is that they have two different crystal structures, depending on temperature. Just below the freezing temperature, iron and dilute alloys are body-centered cubic (δ phase). At lower temperatures, the steel transforms to austenite (γ phase), which is face-centered cubic. At still lower temperatures, the form reverts to body-centered cubic, here called α phase. (The transition from body-centered cubic to face-centered cubic to body-centered cubic again is unique to iron.) Beta (β) iron disappeared from the scientific

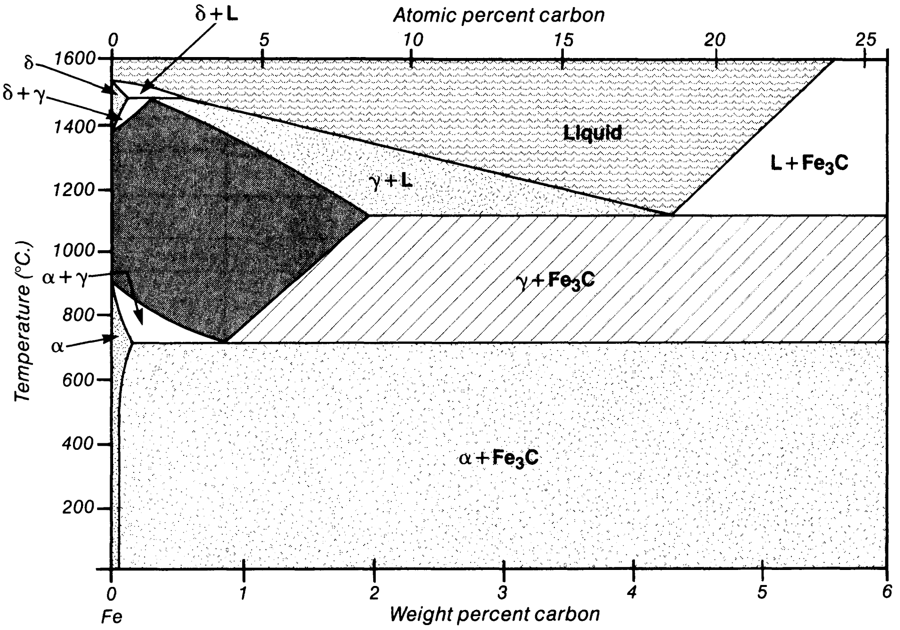
δ+L
Atomic percent carbon
0
5
10
15
20
25
δ
1600
δ+γ
1400
Liquid
L+Fe₃C
γ+L
1200
1000
α+γ
γ+Fe₃C
800
α
600
Temperature (°C.)
400
α+Fe₃C
200
0
1
2
3
4
5
6
Fe
Weight percent carbon

FIG. 7. Iron-carbon equilibrium diagram (omitting equilibria involving graphite).

vocabulary when it was recognized that the phase it was intended to describe entails no crystallographic but only a magnetic transition. At low temperatures, α is ferromagnetic, but it becomes paramagnetic at higher temperatures.

When a plain carbon steel containing 3.6 atoms of carbon per 100 atoms of alloy (3.6%) is slowly cooled, it undergoes a eutectoid reaction. This reaction is akin to a eutectic reaction (see above) but takes place in the solid state. The product is a mixture called pearlite, consisting of nearly pure body-centered iron (called ferrite) and cementite, Fe_3C. Figure 8(a) shows a polished and etched section of pearlite under the microscope. If, however, the austenite is quenched rapidly in water, pearlite has no time to form, and a product called martensite forms instead. This transformation does not appear in the phase diagram, because martensite is metastable in relation to pearlite and forms by a shear process requiring no migration of carbon atoms, in contrast to pearlite. (Indeed, it cannot be displayed in a conventional phase diagram, because the fraction of the alloy transformed into martensite varies continuously with temperature below a critical value, called M_S.)

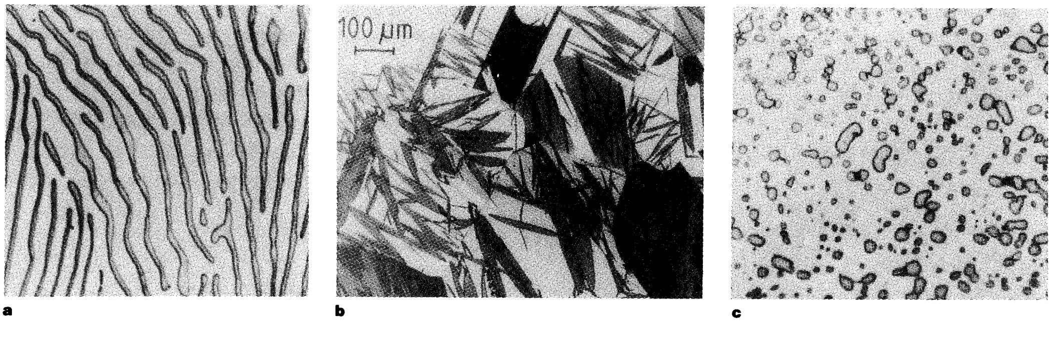

FIG. 8. Optical micrographs of (a) pearlite, (b) martensite, and (c) a distribution of cementite particles in ferrite (courtesy E. Hombogen).

Figure 8(b) is a micrograph of a steel that has been partly transformed into martensite. Each "needle" is the product of a single shear process.

Martensite is tetragonal, with a severely distorted form of body-centered cubic crystal structure, making it extremely hard and brittle. When martensite is heated to moderate temperature, it begins to decompose into an intimate mixture of iron and cementite, becoming progressively softer and more deformable [Fig. 8(c)]. This is the age-old process of tempering. A high-quality knife or sword is shaped while the steel is soft, then heat treated to quench the cutting edge very fast and the thicker part more slowly, and finally mildly tempered. The result is an edge that is very hard and a backing that is less martensitic and thus less brittle. Many actual steels are the result of far more complex heat treatment than the simple Fe-C steel just discussed and also contain other alloying elements such as chromium, vanadium, niobium, and molybdenum. These additions can enhance the tendency to form martensite, so that a thicker section can still be effectively hardened. Each steel has characteristic time-temperature transformation curves (Fig. 9 shows an example) so that an engineer knows how thick a section of that steel can be effectively hardened.

Until recently, rolled steel sheet, as used for the manufacture of automobile bodies, could not be heat treated because the huge coils of sheet had far too much thermal inertia to permit rapid changes of temperature. In continuous annealing, the sheet is cold rolled and then passed through a furnace and a forced cooling device before coiling. The temperature-time schedule is regulated by computer control of every stage of the process. In this way, lightly alloyed steel sheet

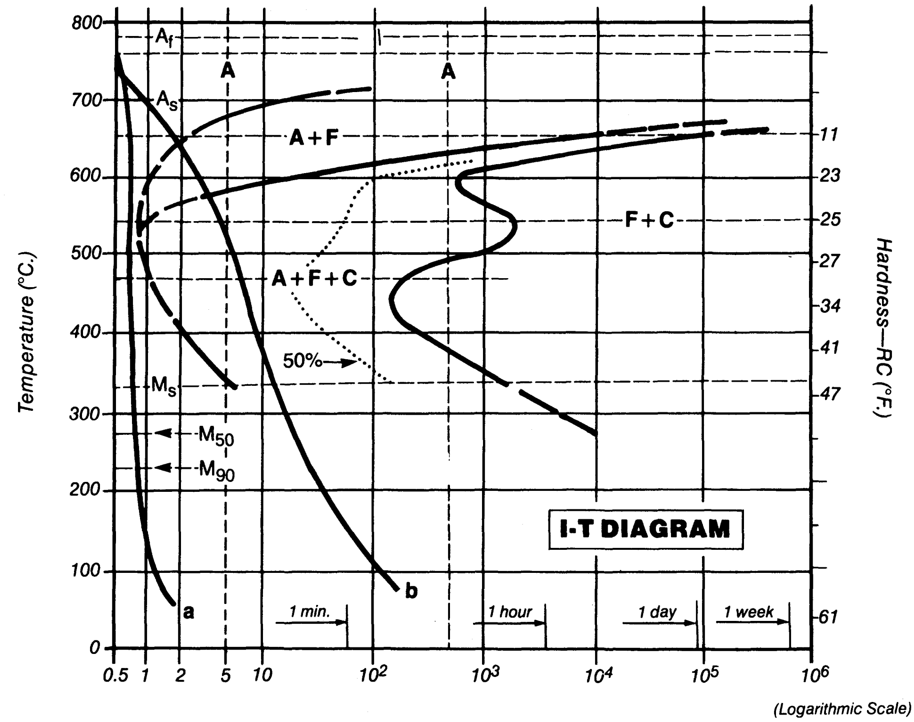

FIG. 9. Time-temperature transformation curves for isothermal heat treatment of an initially quenched steel containing 0.4 wt% C and 1.0 wt% Mn. A = austenite (solution of carbon in face-centered cubic iron); F = ferrite (solid solution of carbon in body-centered cubic iron); C = cementite, Fe_3C. M_s is the temperature at which martensite starts to form on cooling; M_{90} indicates 90% completion of the transformation. Curve **a** represents the temperature-vs-time curve characteristic of a water quench, which generates martensite only; curve **b** represents the cooling rate for an oil quench (slower cooling), which generates a mixture of martensite, ferrite, and cementite. The other curves represent the start, 50% stage, and termination of the isothermal transformation of austenite to yield ferrite and cementite.

can be manufactured to standards of strength and uniformity that were previously unattainable. The result is an automobile body that is light and strong. Heat treatment of steels, and indeed that of nonferrous alloys, is further discussed in MATERIALS TREATMENT; see also STEEL.

2.5 Ternary Alloys

Many important phase diagrams refer to ternary systems—systems comprising three elements. The Fe-Ni-Cr alloy system is a prominent instance; it is the key to understanding stainless steels. Ternary diagrams involve three-phase, two-phase, and single-phase fields. They are thus quite complex and require three-dimensional graphical representation by means of a prism on an equilateral triangular base. The axis of the prism represents temperature, and each point on the base represents a unique ternary composition. Phase boundaries are displayed in two-dimensional sections. Quaternary systems are even more complex and are not often examined. They are represented by means of a regular tetrahedron in which each point represents a unique composition. All such points refer to a single temperature.

2.6 Calculation of Phase Diagrams

The time and effort involved in the accurate determination of equilibrium diagrams, especially ternary systems, have been greatly reduced by the development of theoretical methods such as CALPHAD (CALculation of PHAse Diagrams). Thermochemical quantities such as heats of formation and activities of solid solutions can be measured experimentally and then used to predict the location of phase boundaries. In particular, from measurements for three binary systems, phase boundaries in the corresponding ternary system can be calculated. This activity is a flourishing subspecialty in physical metallurgy.

3. INSTRUMENTATION FOR CHARACTERIZATION OF ALLOYS

Nothing has yet been said about the experimental methods used to obtain the information discussed in the preceding pages. We now turn to the role of metallurgical instrumentation in determining the nature of crystal structure, the position of phase boundaries in equilibrium diagrams (which involves both temperature and composition measurement), the particulars of order-disorder transitions, and the kinetics of phase transformation in alloy steels at various temperatures. For convenience, the methods of attack can be divided into three categories:

1. means of determining crystal structure;
2. means of observing microstructure; and
3. means of determining local variations in composition.

3.1 X-Ray Diffraction

Until 1912, no method existed for the determination of even the simplest crystal structures. In that year, Max von Laue (German), who was interested in establishing the nature of the recently discovered x rays, sought to diffract x rays from crystals, which he estimated from indirect evidence to have interatomic dimensions of the same order of magnitude as the postulated wavelength of x rays. He argued that a periodic three-dimensional structure would scatter (diffract) a beam of x rays strongly in particular directions, much as a diffraction grating scatters a beam of light. There is a simple mathematical relationship between the directions of incident and diffracted beams and the periodic spacing of the diffraction grating or crystal lattice planes (Fig. 10). Von Laue's experiment was successful, and the wave nature of x rays was proved. (At the time, he could not know that, soon after, quantum wave mechanics and the experiments of Arthur Compton would show that the particle model for x rays is just as valid as the wave model; both models are correct.)

Von Laue's lead was followed up at once by the English physicists William Henry and William Lawrence Bragg, father and son, who determined the simple structures of materials such as rock salt and copper. The Braggs were the first to realize that it is possible in principle to locate the geometric positions of atoms in a crystal from measurements of the directions and intensities of the scattered beams. Soon after, the technique was used by metallurgists to determine the structures not only of elementary metals and solid solu-

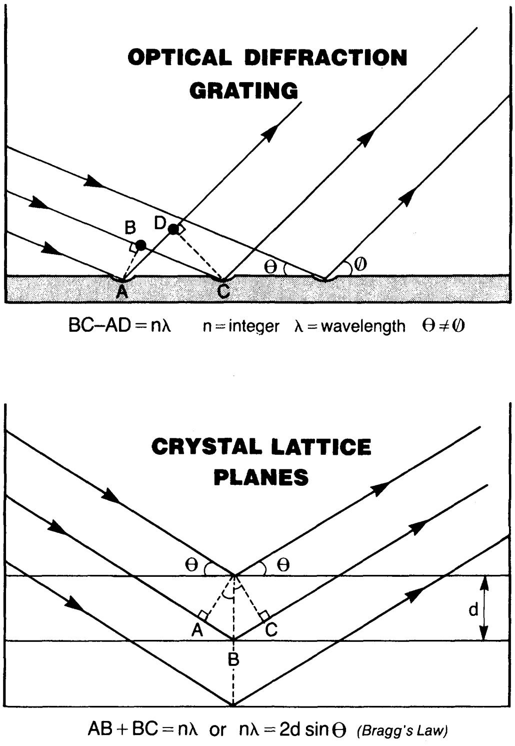
OPTICAL DIFFRACTION GRATING
B
D
A
C
θ
Ø
BC–AD = nλ n = integer λ = wavelength θ ≠ Ø
CRYSTAL LATTICE PLANES
θ
θ
A
C
B
d
AB + BC = nλ or nλ = 2d sinθ (Bragg's Law)

FIG. 10. X-ray diffraction from an optical grating and from a family of lattice planes in a crystal. For the lattice planes but not for the uniperiodic grating, the angles of incidence and diffraction must be equal. In both instances, successive diffracted rays are an exact integral (n) number of wavelengths (λ) out of step.

tions, but also of numerous intermetallic compounds. By now, thousands and thousands of such individual structures are known, many of extreme complexity. X-ray diffraction was the most important single experimental innovation of the 20th century in the field of physical metallurgy (see X-RAY DIFFRACTION).

Two types of x-ray technique developed. In the powder method, an assembly of millions of minute, randomly oriented crystallites is exposed to a beam of single wavelength. In the single-crystal method, a small monocrystal is rotated or oscillated while being bathed by a single-wavelength x-ray beam or is exposed motionless to a multiwavelength beam. In either case, diffracted beams flash out in specific preferred directions. Single-crystal methods are used to determine crystal structure and also to identify the orientations of single metal crystals used for research. The powder method is used for such purposes as identifying unknown compounds, determining phase boundaries in phase diagrams, and searching for nonrandom (preferred) orientation of the crystal grains in metal sheets. It is also used to determine internal strains within the grains of a polycrystalline metal object by using the spacing of lattice planes as a built-in gauge. One of the more sophisticated applications of the method is to the measurement of the relationship in crystal orientation between a parent phase in an alloy and a phase formed from it by partial transformation during heat treatment. (Solid-state chemists call this sort of relationship topotaxy.)

Diffraction experiments were originally done with special cameras, using photographic film. Diffractometers, in which the diffracted beams are located and measured by means of devices such as Geiger counters or photomultipliers, are preferred nowadays.

When progressively increasing amounts of a solute are incorporated in solid solution in a metal (or in a berthollide intermetallic compound), the spacing of the lattice planes progressively alters according to the relative sizes of the two atom species. When the solubility limit is reached, the change of spacing ceases, and the position of the phase boundary of, say, a terminal solid solution at a particular temperature can then be accurately determined. X-ray diffraction is widely used in physical metallurgy for this rather simple type of measurement.

3.2 Optical Microscopy

The light microscope was first applied systematically to the study of metals and alloys by Henry C. Sorby, working between 1847 and 1887. Ever since, seeing has been believing for the physical metallurgist, and a substantial subsidiary science of metallography has grown from the use of the light microscope.

To examine an alloy under the light microscope, a section is cut and mechanically or chemically polished to a very fine finish. The sample is then etched in acid or other reagent, so that different phases can be distinguished. Extensive training and skill are required to produce a section of optimum appearance. Figure 8 illustrates alternative microstructures of a simple carbon steel. Figure 8(a) shows pearlite; Figure 8(b) the shear-generated structure of martensite; and Fig. 8(c) the fine distribution of cementite in tempered martensite. These micrographs show clearly that very distinct microstructures can be created in one and the same alloy by different heat treatments.

A major use of the light microscope is to define phase fields in equilibrium diagrams. An alloy is water quenched from a high temperature, and a section is then polished, etched, and examined to identify the phase or phases in the microstructure. This identification may require the use of other instruments as well, such as the electron probe microanalyzer described below.

3.3 Electron Microscopy

Over the last several decades, the light microscope has been supplemented by a pair of instruments that permit much finer detail to be seen. The resolving power of a light microscope is limited by the wavelength of visible light, so that features separated by less than 10^{-3} mm (1 μm) cannot be properly resolved. The transmission electron microscope can improve this resolution by more than a thousandfold, allowing, in the best instruments, features about 0.2 nm (2×10^{-7} mm) apart to be resolved. The image is either recorded on a photographic plate or projected onto a fluorescent screen for immediate viewing.

The specimen is in the form of a foil thin enough (about 0.1 μm) to allow the electrons to pass through. The resolving power of the microscope should ideally approximate the electron wavelength, $\approx$0.004 nm. If this were possible, fine detail within individual atoms ($\approx$0.2 nm diameter) should be resolvable. In practice, however, the resolving power is limited by aberrations intrinsic to the construction of the instrument. Accordingly, the resolution achieved is less than the ideal by a factor of about 50.

The formation of the image entails diffraction of the electrons from crystal planes in the specimen. Crystallographic information can be obtained by manipulating the diffracted beam in various ways. Thus, it is possible to check the crystal structure, and the spatial orientation of the crystal axes, of the phases in the foil, as well as the spatial disposition of the phases. An electron microscope combines the best features of a light microscope and an x-ray diffraction camera and in addition has a greatly superior ability to resolve fine detail.

Figure 11 shows two microstructures as seen in the transmission electron microscope. Figure 11(a) shows the structure of an "age-hardened" aluminum-copper alloy. The individual "zones" consist of regions in which copper atoms are locally enriched along specific crystal planes of the aluminum-rich solid solution that has been abruptly cooled from

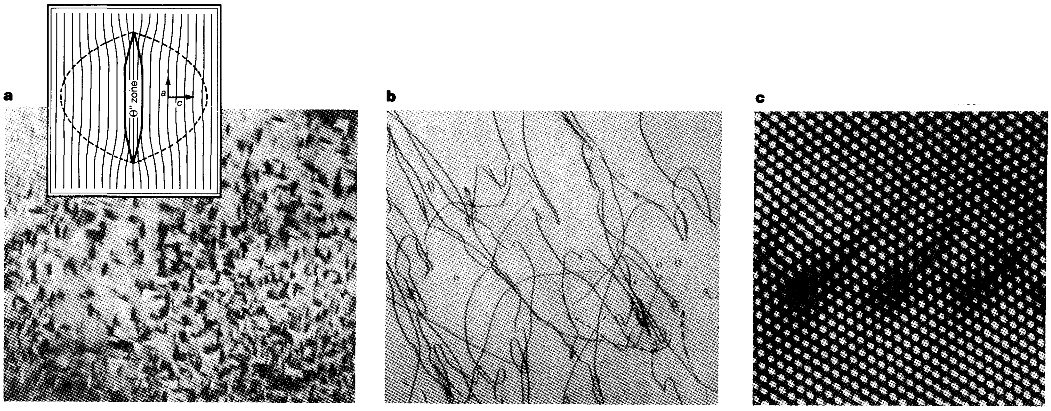

FIG. 11. Transmission electron micrographs: (a) Localized zones enriched in copper in a supersaturated (quenched) and subsequently aged aluminum alloy containing 4 wt% Cu; the inset shows the distortion of lattice planes near a copper-enriched zone (courtesy J. Nutting). (b) A network of dislocations in a magnesium oxide crystal (courtesy Y. Moriyoshi). (c) A high-resolution micrograph of two adjacent grains, lightly misoriented, after computer processing to enhance the contrast; the magnification is approximately ten million times (courtesy W. Krakow).

a high temperature at which all the copper is in solid solution. The solid solution is unstable because the compound $CuAl_2$ should be present in equilibrium at room temperature. But migration of atoms is too slow for $CuAl_2$ to be deposited, and the micrograph shows instead the attainment of a halfway stage: The diffuse areas appearing around the zones represent regions of intense strain, occasioned by the fact that copper atoms are smaller than aluminum atoms. Strain alters the spacing of the lattice planes, and local differences in spacing are revealed by the electron beam as local variations in diffraction intensity. The localized strain, in turn, hardens the alloy; this is the basis of age-hardening, used to strengthen an aircraft structure after it has been shaped and assembled.

Figure 11(b) shows a thin ceramic section containing dislocations. These defects, which are present in all metals and alloys as well as in ceramics, move through a crystal under applied stress and cause the crystal to change shape. The capacity of an alloy to be shaped and to resist mechanical shock depends entirely on these crystal defects, and physical metallurgists spend much time in examining the geometry of dislocations. (More about these entities below.) There is no space in this article to describe the geometry of dislocations; particulars can be found in the article POINT DEFECTS AND OTHER IMPERFECTIONS IN CRYSTALS.

The multiplication and interaction of dislocations causes work-hardening, that is, the increase of resistance to plastic deformation when a metal is extensively deformed. The lines of dislocations show up in the electron microscope because the lattice is distorted in the immediate vicinity of the defect, and the electron beam is scattered less strongly when it passes close to such lines.

Figure 11(c) shows a high-resolution image of an ultrathin crystal foil, which was made by an electron microscope of particularly precise construction. The image shows the positions of stacks of atoms, viewed parallel to the incident electron beam. To a first approximation, one can say that such an image shows the positions of individual atoms in the crystal structure of the specimen. Images of this kind are widely used in examining the fine structure of the boundaries between crystal grains [one such is seen in Fig. 11(c)] and the interfaces between two phases. Grain boundaries in polycrystalline assemblies (the normal condition of industrial metals and alloys) have become particularly important in understanding mechanical and electrical properties and are nowadays the subject of intense study: Electron microscopy has made it possible to demonstrate that the fine structure, and hence the properties, of grain boundaries depend in a complex way on the exact orientation relationship between two adjacent grains. For instance, for certain orientation relationships, dislocations transmit with particular readiness from one grain to its neighbor, and the prevalence of such grain boundaries enhances plastic deformability. The statistical control of grain-boundary crystallography by appropriate deformation and heat treatment has been termed grain-boundary engineering.

The scanning electron microscope is a type of electron microscope used to examine the surface of a solid. Here, an electron beam is made to scan a surface continuously, and scattered electrons are picked up by a particle counter. The reading of the counter is continuously amplified and fed into a television tube that forms an image in synchrony with the beam. It is this image that is observed. Figure 12 shows such an image for a directionally solidified eutectic. This kind of alloy consists of rods of one phase regularly disposed in a matrix of another and is useful

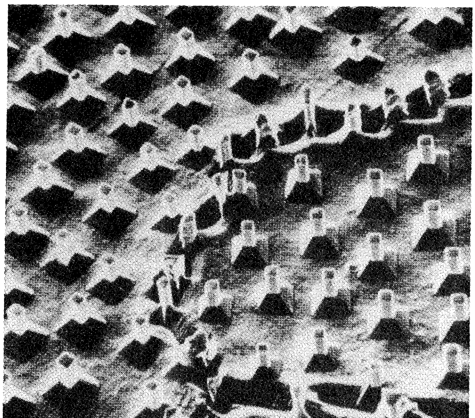

FIG. 12. Directionally frozen eutectic microstructure consisting of tantalum carbide needles in a nickel-chromium matrix. The micrograph was taken using a scanning electron microscope. (Courtesy M. J. Henry.)

for resisting stress at high temperatures. Scanning electron micrographs are particularly useful for revealing perspective detail, at a resolution and depth of focus considerably higher than is possible with a light microscope, which is also used to examine the surface of a solid.

For extensive further details concerning both types of electron microscopy, see ELECTRON MICROSCOPY.

3.4 Determination of Local Composition

Undergraduate students of metallurgy once studied chemical analysis, and most metallurgical research laboratories employed professional analytical chemists. Today, physical metallurgists have a range of instruments that not only analyze an alloy automatically but also can reveal variations in composition from point to point in a microstructure.

All such instruments use the principles of physics to obtain chemical information. They include secondary ion mass spectrometers, x-ray photoelectron spectrometers, Auger spectrometers, and electron microprobe analyzers (EMPAs). The EMPA has become an indispensable tool of every research metallurgist. It is briefly described here. Further details can be found in ELECTRON MICROSCOPY.

In an EMPA, an electron beam is focused on a point, on the order of 1 μm diameter, on the surface of a polished specimen. Two kinds of information result. The backscattered electrons form an image as in a scanning electron microscope, allowing different phases to be distinguished. Also, when an electron beam is slowed by the alloy it generates a burst of x rays, of wavelengths characteristic of the particular metals present. The x-ray wavelengths are measured. The electron beam can be scanned over the surface, and the x rays emitted at a particular electron wavelength can be fed into a picture tube, so that an image of the point-to-point variation of the concentration of the relevant metal can be displayed on the screen.

The EMPA has become an indispensable research tool because of its ability to localize, measure, and identify impurities and phases within a microstructure. Indeed, the EMPA, the x-ray diffractometer, the transmission electron microscope, and the scanning electron microscope have, together, transformed the study of alloys since World War II.

The techniques outlined in Secs. 3.1–3.4 are further discussed in the context of characterization in general in MATERIALS ANALYSIS AND CHARACTERIZATION.

4. MICROSTRUCTURE AND PROPERTIES

An equilibrium diagram specifies only the compositions, crystal structures, and atomic or mass fractions of two or more phases existing in mutual equilibrium. The shapes, sizes, and mutual separation of the regions occupied by the distinct phases are not specified by the Gibbs free energy equation or by the equilibrium diagram depending on this equation. The study of these variables constitutes the study of microstructures.

The quantitative study of microstructures entails the measurement of variables such as the mean diameter of a dispersed phase (for example, the rods in the eutectic of Fig. 12); the form of such a phase (spheres, cylinders, laths, sheets); the mean separation between units of a dispersed phase; the volume fractions of two phases; and the mean free path—the average distance a point moving in a straight line must go between successive intersections with a dispersed phase. All such measurements must be treated statistically and a mean value and a variance determined, since the variables are not constant for any particular microstructure. The science of deducing this information from statistical sample measurements on two-dimensional sections is termed quantitative metallography, or stereology. Computerized image analyzers are used to assemble the statistical array of measurements and to calculate the numerical information from the raw data. Thus, an image analyzer can rapidly determine the distribution of mean diameters of the crystal grains shown in Fig. 1, or the mean spacing and thickness of the cementite plates in the pearlite structure of Fig. 8(a).

4.1 Development of Microstructures

No microstructure is ever wholly stable. When a polycrystalline assembly of grains such as that of Fig. 1 is heated, some grains grow larger by consuming others. Again, a distribution of dispersed particles such as that seen in Fig. 8(c) will progressively coarsen. The control of coarsening is especially important for the optimization of mechanical behavior; thus, the dispersion of phases in a superalloy used in an aero engine must remain fine after long service at high temperatures. This kind of coarsening is also known as Ostwald ripening, after the German physical chemist Wilhelm Ostwald (1853–1932) who first examined it. An example will help to make the point clear. Suppose we have two spheres of phase *A* disposed in a matrix of phase *B*. Suppose further that atoms of the predominant constituent *x* of phase *A* can diffuse (migrate) through the *B* matrix. The local solubility of *x* in *B* will be greater near the smaller sphere, because its sharper curvature exposes the *x* atoms at the interface more freely to the matrix. On balance, therefore, the smaller sphere will dissolve, and the *x* atoms will diffuse along a composition gradient through the matrix and deposit on the more gradually curved interface of the larger sphere. (This is essentially what happens when a large raindrop in a cloud grows at the expense of a smaller one—except that in this case, water molecules evaporate and migrate singly through the air.) The mathematical laws of the process are well understood. To slow down or even eliminate the process, *x* should have a very low solubility in *B*, and the rate of migration of *x* atoms should be low. Furthermore, the atomic fit between the two phases *A* and *B* at their interface should be as nearly perfect as possible. This last feature stabilizes superalloys, as used in the construction of aero engines, against Ostwald ripening.

Ostwald ripening, as discussed above, is an important illustration of the kind of microstructural study that physical metallurgists engage in. In this particular case, the proportions of the two phases *A* and *B* do not change during the process. A more common case is a phase transformation. Thus, Fig. 9 expresses the kinetics of the decomposition of an austenitic steel structure into other phases, and the physical metallurgist's task is to measure these kinetics experimentally and to interpret them in terms of atomic diffusion and events at the interface between the old and the new phases. Indeed, as remarked at the outset, the study of phase transformations is a central concern of physical metallurgy.

Another form of microstructural feature, which originally provided early evidence of the existence of dislocation lines in crystals, is the growth spiral on a crystal surface. Such spirals were first observed in the early 1950s; an example, an optical micrograph of silicon carbide, is seen in Fig. 13. When a particular type of dislocation named a screw dislocation (see POINT DEFECTS AND OTHER IMPERFECTIONS IN CRYSTALS) emerges at a free crystal surface, then an "inextinguishable" step, typically one unit cell in height, is formed at the surface. Such steps are essential for a crystal face to grow, because itinerant atoms are easily captured by a growing crystal at such a step (see CRYSTAL GROWTH). Under certain circumstances, the step height can be larger, as in Fig. 13, and it is then easily captured even in an optical micrograph with its limited resolution. This kind of micrograph caused enormous interest and provoked much

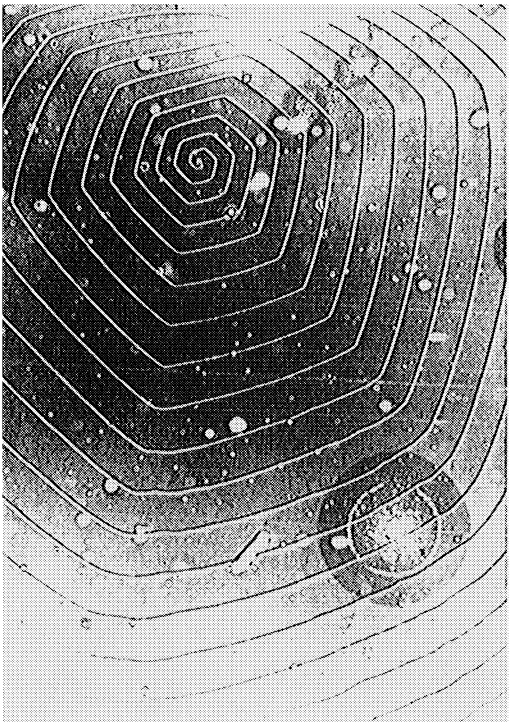

FIG. 13. A growth spiral on a silicon carbide crystal, originating from the point of emergence of a screw dislocation (courtesy S. Amelinckx).

research when dislocations were new and disputed entities.

4.2 Mechanical Properties of Metals and Alloys

The mechanical behavior of solid materials in general, including metals, is discussed in MECHANICAL PROPERTIES OF SOLIDS. The physical metallurgist needs to make sense of the ways in which changes of microstructure modify properties such as ductility and strength. One way in which this has been done is to turn a wire or bar of an alloy into a single large crystal grain—by slow solidification from one end, for example. If the crystal is then deformed, it is found that plastic deformation is concentrated in separate packets, or glide planes, as seen in Fig. 14, so that the crystal looks like a pack of cards being shuffled. In general, concentration of deformation along a single plane as seen in the figure, instead of dispersion among a large number of parallel planes close together (one thick "card" instead of many thin cards, as it were) makes for limited plastic deformability.

The ductility of a metal or alloy has been found to be intimately linked with the number of alternative, differently oriented lattice planes on which such deformation takes place. Thus, metals with high (cubic) crystal symmetry, where many crystallographically equivalent but differently oriented planes exist, are more likely to be highly ductile. The rate at which the crystal hardens as it is progressively deformed is determined by the specific ways in which the dislocation lines [see Fig. 11(b)] multiply and interact with each other. To make sense of this, the investigator combines experiments with single crystals under tension with examination of thin slices of the deformed crystal in a transmission electron microscope.

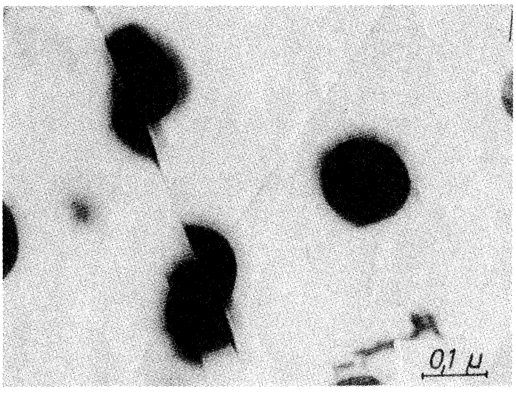

FIG. 14. Precipitated particles in an alloy containing iron, nickel, and titanium. A slip plane arising from plastic deformation is seen shearing the particles. (Transmission electron micrograph, courtesy E. Hombogen.)

Figure 14 shows particles being sheared by the passage of a succession of dislocations on the same plane (though the dislocations themselves are not seen imaged in this micrograph). Depending on the composition and strength of bonding of such particles, such shearing action may entail too large an energy penalty and thus does not happen at all readily. Then particles are not sheared at all but repel dislocation lines; especially if the particle spacing has a critical value, this leads to extensive strengthening of the alloy. Dislocation lines being repelled from fine particles in a high-strength aluminum alloy are seen in Fig. 15. Smaller particles can "pin" a dislocation line; to unpin it, the dislocation line has to cut right through the pinning particles, and that requires much energy and therefore a high stress. An example of dislocation lines pinned at several points by fine particles is seen in Fig. 16, again in an aluminum alloy.

When a metal has been intensely deformed, it becomes filled with a tangle of dislocation lines and loses its capacity to accept further deformation; that is, its ductility becomes exhausted. To restore the ductility, it is necessary to anneal the metal; that is, it must be heated so that a new population of crystal grains consumes the damaged set. This process is known as recrystallization, and its study is a major branch of physical metallurgy (see RECRYSTALLIZATION, DYNAMIC; RECRYSTALLIZATION, STATIC). One aspect of importance is the rate at which the new grain population replaces the old. Another is the statistical tendency for the new grains to line up with their crystal axes in particular orientations in space; the result is termed a preferred orientation, or texture, and can be characterized by x-ray diffraction. Certain preferred orientations are desirable because they enhance the ability of the annealed metal to accept subsequent deformation. Thus, steels to be pressed into shapes for automobile bod-

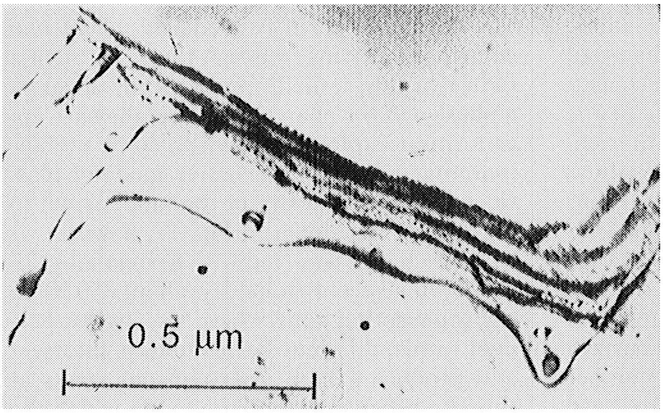

FIG. 15. Dislocations held up by precipitated particles in an aluminum alloy (transmission electron micrograph, courtesy R. A. Ricks).

ies, or steel sheets for the magnetic core of a transformer, are required to have a pronounced preferred orientation.

Another deformation mechanism that is important in some metals, for instance zinc, titanium, and uranium, is deformation twinning. Here, successive crystal planes shear through equal, crystallographically determinate vectors, generating thereby a new lamella of the same phase but in a distinct orientation.

4.3 Mechanically Induced Phase Transformations

Intense mechanical deformation generates a pronounced preferred orientation or deformation texture, and indeed this is a precondition of the formation of annealing textures as just mentioned. But quite apart from this, the application of an external stress can have strange consequences.

It can happen that a phase can be impelled to change into a crystallographically quite distinct form by mechanical stress alone, without necessarily any change in temperature. This has become industrially important in the family of shape-memory alloys, one of which is nearly equiatomic NiTi. These alloys are unstable against martensitic transformations induced by stress. Since martensitic transformation entails a mechanical shear, such alloys can change shape substantially by phase transformation alone, though they are entirely brittle in the sense that dislocations cannot move in them. A shape-memory

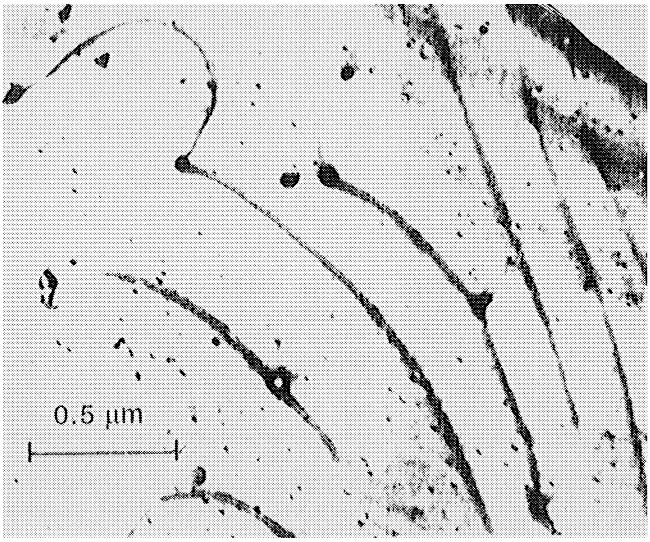

FIG. 16. Dislocations pinned by particles in an aluminum alloy (transmission electron micrograph, courtesy R. A. Ricks).

alloy has the striking property that if it is distorted by an applied stress from its stable shape, it remains metastably in the distorted shape, but on gentle heating, the martensitically generated phase changes back to the original phase, and in the process, the material returns to its pristine shape. Such alloys find extensive use, both in mass applications such as thermostat valves in automobile radiators and in more exotic uses such as wire aerials for satellites, crumpled into a small space before takeoff and deployed into the proper shape by passage of a heating current once the satellite is in orbit. A large number of physical metallurgists have recently been studying this type of alloy to understand just what determines the details of the memory and the force against which the shape-memory effect can successfully operate.

Some very surprising examples of stress-induced phase changes have been discovered recently. Thus, silicon, the most studied of any material, an element of cubic symmetry, has been found to be unstable against a concentrated stress in a limited temperature range and found to transform into a crystal form of hexagonal symmetry when loaded with a sharp diamond. Germanium, an element similar to silicon, behaves in the same way. The hexagonal form is produced where two deformation twins intersect (Fig. 17). Such a twin is a thin lamella produced by shear in a phase *without* change of phase; they are apt to form at explosive speed. In some metals, twinning offers an alternative mode of plastic deformation to slip of the kind shown in Fig. 14. What is seen in Fig. 17, however, is a phase change from cubic to hexagonal symmetry, in a specimen held at about 350 °C, in the intersecting zone where two twins, parallel to two different lattice planes, intersect and produce a high instantaneous stress at this site because of the speed with which the second twin lamella crashes into the first-formed one. This high stress generates the unexpected hexagonal phase.

This recently prepared (1992) micrograph indicates that physical metallurgists can still expect surprises even in their relatively mature science! The fact that Fig. 17 refers to germanium and Fig. 13 referred to silicon carbide, neither of which is a metal, also shows that no branch of materials science can nowadays be hermetically sealed off from its neighboring fields. Most of the journals carrying papers on physical metallurgy now also

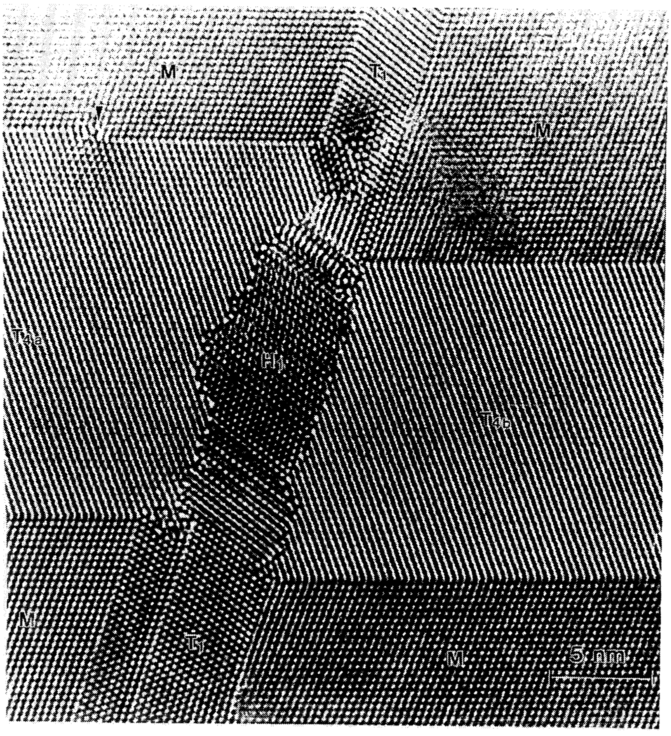

FIG. 17. Stress-induced phase transformation in germanium held at ≈350 °C and deformed under a diamond indentation. M = matrix; T_1, T_4 = primary deformation twins; H = hexagonal polymorph generated where the primary twins intersect. Lattice planes are clearly visible. (High-resolution transmission electron micrograph.) (From Xiao and Pirouz, 1992, courtesy P. Pirouz.)

invite studies of nonmetallic materials that can cast light on the behavior of metallic systems.

5. COMPUTER MODELING

In the last few years, the understanding of the various phenomena that together make up the science of physical metallurgy has been much enhanced by modeling, or simulation, of the processes by means of high-speed computers. This is particularly beneficial when the analytic theoretical expressions that describe processes such as recrystallization, Ostwald ripening, texture formation, or melting, to give just a few examples, are too complex to solve readily, or to solve at all, or alternatively when analytical expressions are not attainable and one is restricted to numerical computations. When it further becomes desirable to carry out simulations for many different values of one or more parameters that enter into the equations, then computers become indispensable.

A typical example of a series of simulated micrographs, here used to map the structural evolution of a coherent particle (i.e., one whose interface with the matrix is of particularly low energy) in the presence of an internal stress in the system, is shown in Fig. 18. This kind of splitting is an alternative response of a microstructure, distinct from the shape-memory response outlined above. The application of computer modeling to problems thrown up by physical metallurgy has recently reached the point where journals entirely dedicated to this technique have been launched (see also Physics Applications of Computers—Modeling).

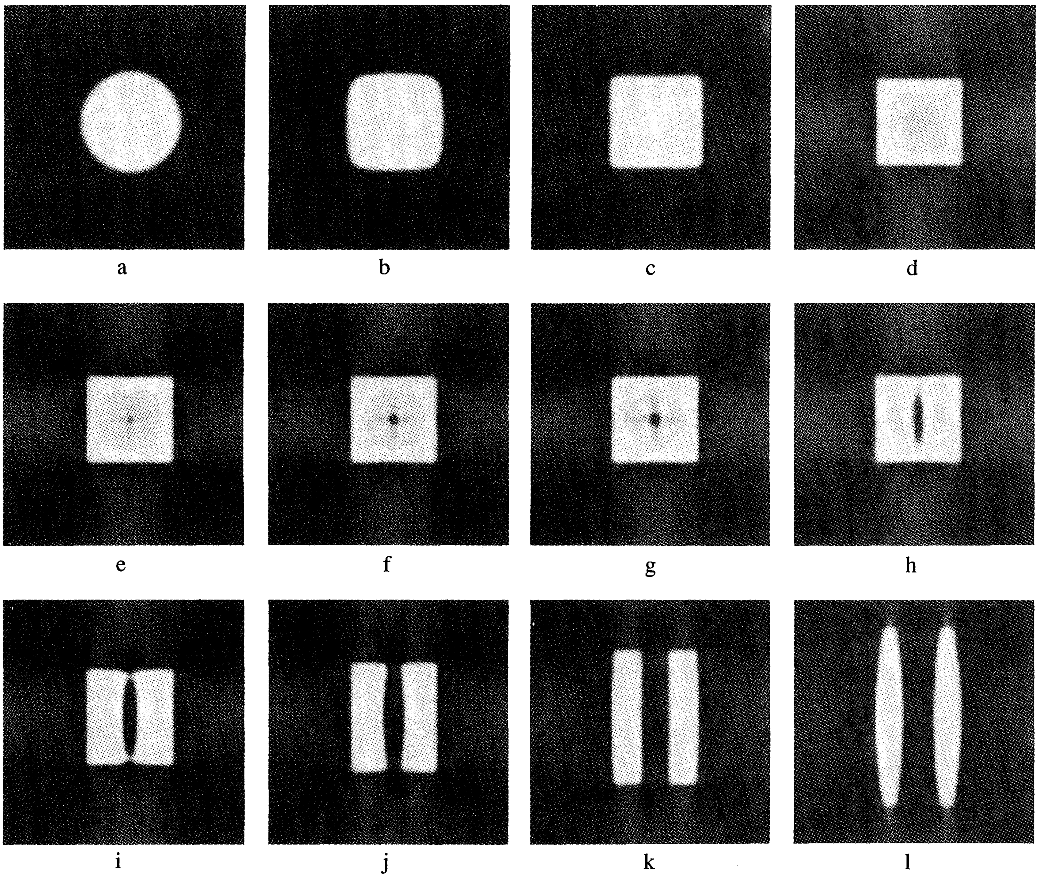

FIG. 18. Computer-simulated sequence of shape relaxation and splitting of a coherently precipitated particle during stress-induced coarsening (from Y. Wang *et al.* 1993, courtesy A. G. Khachaturyan).

ACKNOWLEDGMENTS

A shorter version of this article was published in *Collier's Encyclopedia*, and I am most grateful to Dr. Quigley of Macmillan in New York for his permission to print the expanded version here and to use the versions of some of the illustrations as first printed in *Collier's Encyclopedia*.

I am also deeply indebted to the following colleagues for the provision of original micrographs: Dr. M. J. Henry, Prof. E. Hornbogen, Prof. A. G. Khachaturyan, Dr. W. Krakow, Dr. Y. Moriyoshi, Prof. J. Nutting, Dr. P. Pirouz, and Dr. R. A. Ricks.

GLOSSARY

Cubic Crystal Structure: A structure that is assembled by stacking unit cells in the form of cubes. Such crystals have a high symmetry, which implies that many properties are isotropic.

Diffusion: The process that makes possible atomic transport within a solid. The most common form of this process involves the migration of vacancies (sites where an atom is missing), which entails a counterflow of atoms. Phase transformations requiring a local change of composition can only take place if the temperature is high enough for diffusion to take place at an appreciable rate.

Dislocation: A crystal defect in the form of a line, at which the stacking regularity of a crystal is interrupted. Plastic deformation of crystals requires that dislocations be free to move, under the action of a stress, along a lattice plane, termed the "glide plane" or "slip plane." The direction of displacement of one-half of the crystal relative to the other is termed the "slip direction" and is parallel to a unique vector that characterises any dislocation, the "Burgers vector." During extensive plastic deformation, the density of dislocations is increased by several orders of magnitude by a process of geometrical multiplication.

Eutectic Transformation: A form of solidification in which a homogeneous liquid phase solidifies, at constant temperature, into an intimate mixture of two phases of different compositions (and usually of different crystal structures).

Eutectoid Transformation: Analogous to an eutectic transformation, but here the phase of origin is solid.

Equilibrium Diagram: A graphical representation, with composition as abscissa and temperature as ordinate, of the domains of stability of different phases in an alloy system. "Phase diagram" is an alternative name. Binary diagrams refer to alloys with two components, ternary diagrams to those with three components.

Hexagonal Crystal Structure: A structure that is assembled by stacking unit cells in the form of right regular hexagonal prisms. Such crystals have a lower symmetry than cubic ones and accordingly are more anisotropic in their properties.

Martensitic Transformation: A phase transformation involving the coordinated shear displacement of successive planes of atoms, without any diffusional change of composition. Because no diffusion is involved, the process can be very rapid and can take place at low (even subambient) temperatures.

Metastability: A state (of an alloy, in the present context) in which one or more phases have crystal structures or compositions, or both, that are not in accordance with the equilibrium diagram. To qualify as metastable, such a state must be "frozen in," in the sense that diffusional processes are too slow for the system to evolve at a measurable rate in the direction of the equilibrium configuration. (Term coined by Ostwald in 1893.)

Microstructure: A broad term, denoting the geometrical disposition of the constituent phases of an alloy, whether in equilibrium or in a metastable state. The concept is distinct from that of "crystal structure," which refers to the disposition of atoms within the unit cell of a phase.

Ostwald Ripening: A process during which a population of particles in a biphase alloy progressively coarsens, through the preferential growth of the larger particles at the expense of the smaller. The overall volume fraction of the particles and their crystal structure do not alter during Ostwald ripening.

Pearlite: The product of a eutectoid transformation in a steel, produced by relatively slow cooling. The phase of origin is austenite (a face-centered cubic structure), and the product is an intimate lamellar mixture of almost pure body-centered iron and cementite (Fe_3C), with a complex orthorhombic crystal structure.

Phase: A region in microstructure that is uniform with respect to crystal structure (or can be amorphous, liquid, or gaseous) and, when in equilibrium, is also uniform in composition.

Phase Diagram: See **Equilibrium Diagram**.

Preferred Orientation: The condition of a polycrystalline metal or alloy in which the individual crystal grains tend statistically toward an ideal orientation (which would be equivalent to a single crystal). Such a condition can be produced either by intense plastic deformation or by annealing following such deformation. An alternative name is "texture."

Recrystallization: The process that leads to softening of a work-hardened metal or alloy during high-temperature annealing. The original population of crystal grains, full of dislocations, is replaced by a new population of undamaged grains.

Solid Solution, Terminal: A crystal with two or more components, generated by dissolving solute(s) in a pure metal. Such a solid solution necessarily has the same crystal structure as the pure "parent" element.

Texture: See **Preferred Orientation**.

Work Hardening: The progressive increase in hardness, or strength, of a deformable metal or alloy, resulting from plastic deformation.

Works Cited

Wang, Y. *et al.* (1993), *Acta Metall. Mater.* **41**, 279.

Xiao, S.-Q., Pirouz, P. (1992), *J. Mater. Res.* **7**, 1406.

Further Reading

Numerous books on physical metallurgy and its broader modern variant, materials science, are available, some suitable for the general reader, others more appropriate for the specialist. The first on the list is a general reference work very suitable for the nonspecialist.

Cotterill, Rodney (1985), *Cambridge Guide to the Material World*, Cambridge, U.K.: Cambridge University Press. A new edition is in preparation.

The next two titles assume that the reader has some background in the physical sciences.

Hume-Rothery, W., Smallman, R. E., Haworth, C. W. (1969), *The Structure of Metals and Alloys*, London: Institute of Metals.

Weaire, D. L. W., Windsor, C. G. (Eds.) (1987), *Solid State Science*, Bristol, U.K.: Hilger.

Next is a well-tried exposition that has been used for many years as an introductory text leading on to an intermediate level. A recent edition is cited.

Reed-Hill, Robert E., Abbaschian, R. (1992), *Physical Metallurgy Principles*, Boston: PWS-Kent Publishers.

The last three books are for more advanced readers, and each includes a good deal of theory.

Cahn, Robert, Haasen, Peter (1983), *Physical Metallurgy*, 3rd ed., Amsterdam: Elsevier-North-Holland; (1995), 4th ed.

Haasen, Peter (1986), *Physical Metallurgy*, 2nd ed., Cambridge, U.K.: Cambridge University Press.

Porter, D.A., Easterling, K.E. (1981), *Phase Transformations in Metals and Alloys*, New York: Van Nostrand Reinhold.

METALS

See ALUMINUM; COPPER; PRECIOUS METALS; STEEL

METALS AND ALLOYS, CONDUCTIVITY IN

P. L. ROSSITER, *Department of Materials Engineering, Monash University, Clayton, Victoria, Australia*

J. BASS, *Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan, U.S.A., and Max-Planck-Institut für Festkörperforschung, Hochfeld Magnetlabor, Grenoble, France*

Introduction 164
1. **Electronic Transport Properties** 164
1.1 Background 164
1.2 Electrical Resistance *R* and Resistivity ρ 165
1.3 Thermopower *S* 165
1.4 Magnetoresistance *MR* and Hall Coefficient R_H 166
1.5 General Transport Equations and Coefficients 167
2. **The Boltzmann Transport Equation** 167
2.1 Background 167
2.2 Linearized Boltzmann Equation 168
2.3 Relaxation-Time Approximation 168
2.3.1 Electrical Conductivity 169
2.3.2 Thermoelectric Effects 169
2.3.3 Electronic Thermal Conductivity 169
2.3.4 Hall Effect 170
2.3.5 Magnetoresistance 170
3. **Scattering Mechanisms** 171
4. **Electrons in a Magnetic Field** 171
4.1 Cyclotron Orbits and $\omega_c\tau$ 171
4.2 High-Field Forms of *MR* and R_H 172
5. **Electrical Resistivity ρ** 172
5.1 Pure Metals 173
5.1.1 Static Displacements 173
5.1.2 Thermal Effects 176
5.2 Alloys 176
5.3 Magnetic and Nearly Magnetic Metals and Alloys 179
5.4 Liquid and Amorphous Alloys 180
5.5 ac Effects 182
6. **Magnetoresistance (*MR*), Transverse (T) and Longitudinal (L)** 182
6.1 Kohler's Rule 182
6.2 Low Magnetic Fields ($\omega_c\tau \ll 1$) 183
6.3 High Magnetic Fields ($\omega_c\tau \gg 1$) 183
6.4 Additional Examples 183
6.4.1 Deviations from the Ideal 183
6.4.2 Quantum Oscillations 183
6.4.3 Size Effects 184
6.4.4 Ferromagnets 184
7. **Hall Effect, R_H** 184
7.1 Free Electrons 184
7.2 Beyond Free Electrons 186
7.2.1 Low Magnetic Fields 187
7.2.2 High Magnetic Fields 187
7.3 Additional Examples Including Magnetic Metals 187
8. **Thermopower *S*** 187
8.1 The Temperature Dependence of *S* in Pure Metals 187
8.2 Rough Derivations of S_d and S_g 188
8.3 More Detailed Analysis of S_d 190
8.4 Additional Examples 192
8.5 Magnetic-Field Effects 194
Glossary 194
Works Cited 195
Further Reading 196

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

The electrical transport properties of metals and alloys have been extensively studied for over 100 years, both in order to understand the basic phenomena and to provide a tool for monitoring other effects such as phase transformations, defect levels, and electron states and mobilities. The conductivities of solids cover a range of about 10^{30} from the generally highly conducting metals and alloys at one extreme to insulators at the other. Applications based upon specific conductivity characteristics include electrical power distribution systems, stable electrical resistances, heating elements, and transducers to measure temperature and strain, for example. Similarly, Hall-effect devices find wide application in magnetic field measurement and various rotation and displacement transducers, while the thermopower determines the operating characteristics of thermocouples, and thermoelectric cooling is quite commonplace.

All of the properties of interest here—the conductivity (or resistivity), thermopower, magnetoresistance, and Hall coefficient—are determined by the rate of scattering of conduction electrons from initial to final states on the Fermi surface. This is described in an approximate way by the Boltzmann transport equation and the simplified idea of a relaxation time. Scattering mechanisms are briefly discussed leading to a statement of the scattering problem in terms of the nature of the atomic potentials (form factor) and their distribution in space (structure factor).

Specific results for the resistivity, thermopower, magnetoresistance, and Hall coefficient are given for a range of conducting materials under various temperature and field conditions.

1. ELECTRONIC TRANSPORT PROPERTIES

1.1 Background

In 1900, Drude attempted to explain the transport properties of metals by assuming that the electrons behaved as classical particles moving freely through the metal but occasionally bouncing off positive ions, with a mean time between collision (or relaxation time) of τ. While this model led to a seemingly plausible explanation of the electrical resistivity, it failed in other areas such as accounting for the relatively small electronic specific heat observed in metals at room temperature. Following the realization that electrons must obey the Pauli exclusion principle, Sommerfeld replaced the classical Maxwell–Boltzmann distribution of electron energies in the Drude model with the Fermi–Dirac distribution. This theory, known as the "free-electron theory," accounted for a wide variety of metallic properties but could still shed no light on the actual scattering process. Nor could it account for many other observations such as positive Hall constants, the temperature dependence of conductivity, or, indeed, why not all elements are good conductors.

These matters were resolved by properly taking into account the effects of the potential field generated by the ions. In a crystalline solid, this produces diffraction effects leading to the notion of energy bands, with scattering of the conduction electrons (and hence finite conductivity) produced by deviations from perfect periodicity of the lattice potential. It turns out that in many cases the potential seen by the electrons is only weak, and so much of the subsequent work falls under the title "nearly-free-electron theory." This theory forms the basis for the discussion that follows.

We wish to examine how the electrons in metals transport charge and heat under the influence of an applied electric field **E** and/or a temperature gradient $\nabla_r T = \Delta T/\Delta x$. Here, ΔT is a temperature difference applied over a small distance Δx, and boldface indicates a vector.

In zero magnetic field, the quantities of main interest are the conductivity σ (or the electrical resistivity, $\rho = 1/\sigma$) and the thermopower S. When a magnetic field **B** is applied, we shall focus upon the magnetoresistance $MR = [\rho(\mathbf{B}) - \rho(\mathbf{B} = 0)]/\rho(\mathbf{B} = 0)$, and the Hall coefficient R_{H}. The third of the (generally) independent transport coefficients, the thermal conductivity κ, is discussed only briefly, in terms of the Wiedemann-Franz Law, which relates κ to ρ at low and high temperatures. The reader interested in further information about κ is referred to the articles HEAT TRANSFER and THERMAL CONDUCTION IN SOLIDS.

In this section, we first briefly describe how each of these four quantities is measured, and then discuss the general linear transport equations relating the current density of charge $\mathbf{j}$ and the current density of heat $\mathbf{j}_q$ to $\mathbf{E}$ and $\nabla_r T$. We then outline the microscopic theory of electronic transport in metals and subsequently examine a variety of experimental data.

1.2 Electrical Resistance *R* and Resistivity *ρ*

The electrical resistance R of a metallic sample is measured by injecting a known current I_x through a wire [Fig. 1(a)] or foil [Fig. 1(b)] having cross-sectional area $A = \pi r^2$ or Wt, respectively, where r is the wire radius and W and t the foil width and thickness. From Ohm's law,

$$R = V_x/I_x, \tag{1}$$

where V_x is the emf (or voltage) generated over length L, and R is an extensive quantity that doubles if L is doubled or A is halved. Arrangements for liquid metals and alloys generally follow these configurations, with the additional requirement of some form of containment. The intrinsic property of the metal is ρ, related to R by

$$\rho = R(A/L), \tag{2}$$

provided that the material is homogeneous and that $\mathbf{j}_x = I_x/A$ is uniform across the sample area. The resistivity ρ is discussed in Sec. 5.

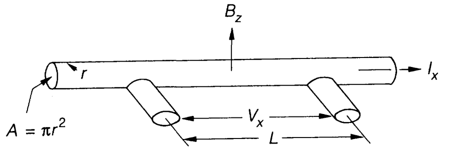

Wire

(a)

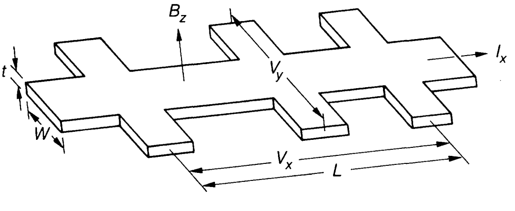

Foil

(b)

FIG. 1. Specimen configuration for measuring the electrical resistance of (a) a wire and (b) a foil. Arrangement shown in (b) also allows determination of the Hall voltage V_y.

1.3 Thermopower *S*

There are three experimental thermoelectric phenomena with three corresponding coefficients: the Thomson effect (coefficient μ), the Peltier effect (coefficient Π), and the Seebeck effect (thermopower coefficient S). Only the Thomson effect can be measured on a single metal. It was discovered in 1854 by William Thomson (Lord Kelvin) and involves the reversible generation of heat when current flows in a conductor while a temperature gradient is present (i.e., the production of heat changes to absorption if the direction of either the current or the temperature—but not both at once—is reversed). In the Peltier effect, discovered by French watchmaker Charles Peltier in 1834, heat is reversibly generated at the junction between two different metals when a current flows through the junction. The most widely known of the thermoelectric effects is the Seebeck effect, discovered by a Dutchman, Thomas John Seebeck. In 1821, he observed that a magnetic needle deflected when the two junctions of a nearby closed electrical circuit consisting of two different solid or liquid conductors were held at different temperatures; we now know that this deflection was caused by an electrical current that the temperature difference caused to flow in the circuit. The fundamental thermoelectric property of the conductors comprising the circuit is the voltage (or electromotive force) produced by the circuit when one junction is electrically disconnected (open circuit), since the current in the fully connected circuit depends upon the electrical resistivities of the circuit elements. Such an open-circuit configuration is called a thermocouple (Fig. 2). Thermocouples find wide use in temperature measurement and regulation. Connecting a number of thermocouples in series, and heating one common junction (e.g., with a small piece of radioactive material) of the resulting "thermopile," produces thermoelectric "bat-

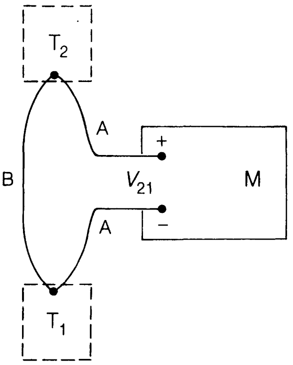

FIG. 2. Diagram of apparatus usually used for measuring thermoelectric (Seebeck) voltage V_{21}. M is an instrument for measuring potential.

teries" or "generators" that find use in space probes and remote places where long life, small size, absence of moving parts, and insensitivity to environmental changes are needed.

The three thermoelectric coefficients are connected by the Kelvin relations:

$$S_1 = \int_0^T (\mu_1/T')dT', \tag{3a}$$

$$\mu_1 = T(\partial S_1/\partial T), \tag{3b}$$

$$\Pi_{12} = T(S_2 - S_1), \tag{3c}$$

so that complete knowledge of any one yields complete knowledge of all three; S is usually the easiest to measure.

When a temperature difference $\Delta T = T_H - T_L$ is applied across a thermocouple, the resulting thermoelectric voltage is

$$\begin{aligned} V_{21} &= \int_{T_L}^{T_H} [S_2(T) - S_1(T)]dT \\ &= \int_{T_L}^{T_H} S_2(T)dT - \int_{T_L}^{T_H} S_1(T)dT, \end{aligned} \tag{4}$$

where S_1 and S_2 are the thermopowers of metals 1 and 2, respectively. If both metals are homogeneous, the integrals depend only upon the materials and the end temperatures T_H and T_L, and not upon how the temperature varies along each arm of the couple. If ΔT is small, then

$$S_2 - S_1 = V_{21}/\Delta T. \tag{5}$$

Equation 5 shows that S_1 for a sample with unknown properties can only be determined if S_2, the thermopower of a reference material, is known. At very low temperatures, the reference material can be a superconductor, for which $S = 0$ in its superconducting state. "High-temperature" ceramic superconductors can be used up to around 90 K. At higher temperatures, Pb has been calibrated as a reference standard up to 350 K. Above 350 K, there is no universal standard, although Au, Cu, and Pt have been used.

The thermopower S consists of two very different parts: S_d, the electron diffusion component, involves a differential flow of electrons from one end of the metal to the other under the influence of a temperature gradient, with the phonons in the metal in thermal equilibrium; S_g, the phonon-drag component, results from the "dragging along" of electrons by the nonequilibrium phonon flow from the hot end of the metal to the cool end that is caused by the temperature gradient. In pure metals, S_g often dominates S from a few kelvins to above 100 K. The quantities S_d and S_g are discussed further in Sec. 8.

1.4 Magnetoresistance *MR* and Hall Coefficient R_H

In 1879, while studying the force exerted on a current carrying metal by an applied magnetic field, and looking for a change in resistance with increasing field (*MR*), the Englishman E. H. Hall discovered that applying such a field perpendicular to the direction of current flow gave rise to a voltage in the direction perpendicular to both the current and the field. This "Hall effect" is characterized by a coefficient R_H; *MR* and R_H can be measured together on the foil sample shown in Fig. 1(b). A current I_x is passed through the sample from left to right, a field $\mathbf{B}_z$ is applied normal to the sample, and the voltages V_x and V_y are measured. The direction of $\mathbf{B}$ is then reversed, and the two voltages measured again. The *MR* is defined to be an even function of $\mathbf{B}$ and is determined from the average of the values of V_x for the two field directions. The term R_H is defined to be an odd function of **B**, and it is determined from the difference between the two values of V_y. This reversing procedure is one of the steps necessary to minimize systematic errors in MR and R_H. Errors can result if care is not taken to ensure spatial uniformity of **B**, good alignment of the sample in the field, and a uniform current flow through the sample.

Since $V_x = E_xL$ and $I_x = j_xWt$, we have $\rho(\mathbf{B}) = V_xWt/I_xL$ for each of the two terms to be averaged in determining the MR. Since $V_y = E_yW$, each term for R_H must have the form $R_H = V_yt/BI_x$; R_H is defined with B divided out, because in many cases V_y is proportional to B.

The terms MR and R_H are discussed in Secs. 6 and 7, respectively.

1.5 General Transport Equations and Coefficients

The formal transport equations for metals provide linear relations among four vector quantities: **E**, $\nabla_r T$, the electrical current density **j**, and the heat current density $\mathbf{j}_q = \mathbf{I}_q/A$, where $\mathbf{I}_q$ is the total heat current. These equations can be written in two different forms.

The first form defines the directly measured quantities. The vectors **E** and $\mathbf{j}_q$ are related to the vectors **j** and $\nabla_r T$ by tensor coefficients (also designated by bold symbols):

$$\mathbf{E} = \boldsymbol{\rho}(\mathbf{B}) \cdot \mathbf{j} + \mathbf{S}(\mathbf{B}) \cdot \nabla_r T, \tag{6a}$$

$$\mathbf{j}_q = \boldsymbol{\Pi}(\mathbf{B}) \cdot \mathbf{j} - \boldsymbol{\kappa}(\mathbf{B}) \cdot \nabla_r T, \tag{6b}$$

where the coefficients are generally functions of **B**. The thermal conductivity $\boldsymbol{\kappa}$ is found by applying $\mathbf{j}_q$ and measuring $\nabla_r T$ while holding $\mathbf{j} = 0$. Similarly, $\boldsymbol{\rho}$ is determined by applying **j** and measuring **E** while holding $\nabla_r T = 0$, and **S** is determined by applying $\nabla_r T$ and measuring **E** while holding $\mathbf{j} = 0$. From the Onsager relation (Ziman 1960, 1972), $\boldsymbol{\Pi}(\mathbf{B}) = T\mathbf{S}(-\mathbf{B})$, and there are only three independent tensors in Eqs. (6): $\boldsymbol{\rho}$, $\boldsymbol{\kappa}$ and **S**.

The second pair of equations defines the calculated quantities. The terms **E** and $\nabla_r T$ are treated as "forces" and **j** and $\mathbf{j}_q$ as derived "flows." These equations have the advantage of being directly comparable with the Boltzmann transport equation which we discuss in Sec. 2:

$$\mathbf{j} = \boldsymbol{\sigma}(\mathbf{B}) \cdot \mathbf{E} + \boldsymbol{\gamma}(\mathbf{B}) \cdot \nabla_r T, \tag{7a}$$

$$\mathbf{j}_q = \boldsymbol{\gamma}(-\mathbf{B})T \cdot \mathbf{E} - \boldsymbol{\kappa}'(\mathbf{B}) \cdot \nabla_r T, \tag{7b}$$

For $\mathbf{B} = 0$, Eqs. (6) and (7) reduce to scalars in cubic metals. Comparing Eq. (7) and Eq. (6) in scalar form shows that $\kappa' = \kappa - S^2/\sigma$ and $\gamma = -S\sigma$. In metals, S^2/σ is nearly always negligibly small compared to κ. The differences between Eqs. (6) and (7) are of little importance when the coefficients are simply scalars, since converting from, e.g., ρ to σ simply involves scalar inversion. When the full tensors must be retained, however, the conversion from one form to another involves matrix inversion. Writing $\boldsymbol{\rho}$ as a tensor with elements ρ_{ij} where $i,j = x,y,z$, the definitions of the MR and R_H given above yield

$$MR = \frac{\frac{1}{2}[\rho_{xx}(B) + \rho_{xx}(-B)] - \rho(B = 0)}{\rho(B = 0)}, \tag{8a}$$

$$R_H = [\rho_{yx}(B) - \rho_{xy}(-B)]/2B. \tag{8b}$$

2. THE BOLTZMANN TRANSPORT EQUATION

2.1 Background

If an electric field is applied to a conductor, all electrons not in completely filled bands are displaced at a uniform rate in **k** space. Scattering by defects tends to restore the electrons to their equilibrium distribution. The steady-state distribution is thus determined by a dynamical balance between the acceleration of the electrons in the applied field and their scattering by the lattice. A distribution function $f(\mathbf{k},\mathbf{r},t)$ may be used to describe the location of the state of an electron in six-dimensional $(\mathbf{k},\mathbf{r})$ phase space. It is defined so that $(4\pi^3)^{-1}f(\mathbf{k},\mathbf{r},t)\, d^3k d^3r$ is the number of electrons that lie in an element d^3r of real space and d^3k of wave-vector space at time t, such that $f = 1$ if all states are filled or $f = 0$ if all are empty. If the electron gas is at equilibrium at some temperature, f is simply the Fermi–Dirac distribution function, designated f_0.

In the presence of applied electric and magnetic fields **E** and **B**, but without scattering, any particular state will move through phase space according to semiclassical equations, but its occupancy will not change (i.e., if it was initially filled it will remain so). However, if scattering occurs, an electron can discontinuously change its momentum from some initial state **k** to a final state **k**′, and the time derivative of f is given by

$$\frac{df}{dt} = \left.\frac{\partial f}{\partial t}\right|_{\text{scatt}}, \tag{9}$$

where the last term is the change in f due to that scattering process.

Expanding the left-hand side in terms of its partial derivatives gives

$$\frac{\partial f}{\partial t} + \frac{\partial f}{\partial \mathbf{r}} \cdot \frac{d\mathbf{r}}{dt} + \frac{\partial f}{\partial \mathbf{k}} \cdot \frac{d\mathbf{k}}{dt} = \left.\frac{\partial f}{\partial t}\right|_{\text{scatt}} \tag{10}$$

Since

$$\hbar \frac{d\mathbf{k}}{dt} = \mathbf{F}, \tag{11}$$

where **F** is any applied force, Eq. (10) may be written as

$$\frac{\partial f}{\partial t} + \mathbf{v}(\mathbf{k}) \cdot \frac{\partial f}{\partial \mathbf{r}} + \frac{e}{\hbar} [\mathbf{E} + \mathbf{v}(\mathbf{k}) \times \mathbf{B}] \cdot \frac{\partial f}{\partial \mathbf{k}} = \left.\frac{\partial f}{\partial t}\right|_{\text{scatt}} \tag{12}$$

where $\mathbf{v}(\mathbf{k})$ is the velocity of electrons in state **k** and e the electronic charge, which is one form of the celebrated Boltzmann equation. The solution of this equation for a specific system will give the distribution function in the presence of external fields or temperature gradients. Such a function determines the electrical and thermal conductivities and all thermoelectric effects, including their dependence upon magnetic fields. However, finding this function is not trivial. The difficulty lies in the complexity of the scattering term, which involves the rates of transition from all states to some particular state, **k**′, and these in turn depend upon the occupation numbers of these states. This term will thus generally involve an integral over all values of **k**′ with the distribution function itself appearing in the integrand, leading to computational difficulties.

2.2 Linearized Boltzmann Equation

Under steady-state conditions, the deviation in f from the equilibrium electron distribution is usually only small. By expressing f explicitly in terms of this deviation $g(\mathbf{k},\mathbf{r},t)$,

$$f(\mathbf{k},\mathbf{r},t) = f_0(\mathbf{k},\mathbf{r}) + g(\mathbf{k},\mathbf{r},t), \tag{13}$$

one may substitute f_0 for f on the left-hand side of Eq. (12) and retain only the lowest power of $f - f_0$ that does not vanish in the scattering term. This procedure leads to the linearized Boltzmann equation:

$$\mathbf{v}(\mathbf{k}) \frac{\partial f_0}{\partial t} \cdot \nabla_r T + e \frac{\partial f_0}{\partial \epsilon_\mathbf{k}} \mathbf{v}(\mathbf{k}) \cdot \mathbf{E} = \left.\frac{\partial f_\mathbf{k}}{\partial t}\right|_{\text{scatt}} - \mathbf{v}(\mathbf{k}) \cdot \frac{\partial g_\mathbf{k}}{\partial \mathbf{r}} - \frac{e}{\hbar} [\mathbf{v}(\mathbf{k}) \times \mathbf{B}] \cdot \frac{\partial g_\mathbf{k}}{\partial \mathbf{k}}, \tag{14}$$

where $\epsilon_\mathbf{k}$ is the energy of an electron of wave vector **k** (see e.g., Ziman, 1972, p. 213), and the subscript **k** has been added to f and g to indicate their explicit dependence upon **k**.

This linearized form yields the observed linear response of the electron and thermal fluxes to electric and magnetic fields and thermal gradients. Equation (14) is the starting point for many derivations of the electrical and thermal conductivities. However, its solution still requires evaluation of the scattering term. If $P_{\mathbf{k}\mathbf{k}'}$ is defined as the probability of transition from some state **k** (assumed to be filled) to some other state **k**′ (assumed to be empty), the scattering term may be written as

$$\left.\frac{\partial f_\mathbf{k}}{\partial t}\right|_{\text{scatt}} = \frac{\Omega}{2\pi^3} \int (g_{\mathbf{k}'} - g_\mathbf{k}) P_{\mathbf{k}\mathbf{k}'} d^3k, \tag{15}$$

where Ω is the specimen volume.

An evaluation of Eq. (15) then requires determination of the transition probability, often with the assistance of first-order perturbation theory.

2.3 Relaxation-Time Approximation

Instead of evaluating the transition probability $P_{\mathbf{k}\mathbf{k}'}$ directly, one may make the phenomenological assumption that the distribution will relax exponentially in time if the field is switched off. Equation (15) then simply becomes

$$\left.\frac{\partial f_{\mathbf{k}}}{\partial t}\right|_{\text{scatt}} = -\frac{f_{\mathbf{k}} - f_0}{\tau_{\mathbf{k}}} = -\frac{g_{\mathbf{k}}}{\tau_{\mathbf{k}}}, \tag{16}$$

where $\tau_{\mathbf{k}}$ is again the relaxation time, which may now vary with the magnitude of **k** (but not its direction). This approximation is rigorous if the scattering is purely elastic (i.e., $|\mathbf{k}| = |\mathbf{k}'|$) and the Fermi surface (FS) is spherical (Ashcroft and Mermin, 1976). In most situations, the assumption of isotropic scattering over a spherical FS is not valid, and more general solutions must be found. Nevertheless, the approximation leads to considerable simplification and serves to illustrate the evaluation of the transport coefficients from the linearized Boltzmann Eq. (14).

2.3.1 Electrical Conductivity In the presence of a uniform electric field, but no magnetic field or thermal gradient, Eqs. (14) and (16) give

$$e\frac{\partial f_0}{\partial \epsilon_{\mathbf{k}}}\mathbf{v}(\mathbf{k})\cdot\mathbf{E} = \frac{-g_{\mathbf{k}}}{\tau_{\mathbf{k}}}. \tag{17}$$

The current density $\mathbf{j}(\mathbf{r},t)$ is obtained by summing the contributions $e\mathbf{v}(\mathbf{k})$ for all electrons:

$$\begin{aligned}\mathbf{j}(\mathbf{r},t) &= \frac{e}{4\pi^3}\int \mathbf{v}(\mathbf{k})f(\mathbf{k},\mathbf{r},t)d^3k \\ &= \frac{e}{4\pi^3}\int \mathbf{v}(\mathbf{k})g_{\mathbf{k}}d^3k \end{aligned} \tag{18}$$

[from Eq. (13)]. Substitution of $g_{\mathbf{k}}$ from Eq. (17) ultimately leads to

$$\mathbf{j}(\mathbf{r},t) = \frac{e^2}{4\pi^3\hbar}\int \tau_{\mathbf{k}}\mathbf{v}(\mathbf{k})[\mathbf{v}(\mathbf{k})\cdot\mathbf{E}]\frac{dS}{|\mathbf{v}(\mathbf{k})|}, \tag{19}$$

where the integration over a volume $d\mathbf{k}$ of **k** space has been transformed into an integration over an element of area of the FS dS.

A comparison of Eq. (19) with Eq. (7a) ($\boldsymbol{\nabla}_rT = 0$) leads to an expression for components of the conductivity tensor:

$$\sigma_{ij} = \frac{e^2\tau_{\mathbf{k}}}{4\pi^3\hbar}\int\frac{v_i(\mathbf{k})v_j(\mathbf{k})dS}{|\mathbf{v}(\mathbf{k})|}. \tag{20}$$

The tensor is symmetric and can be reduced to diagonal form giving the three principal conducting coefficients parallel to the principal axes of the crystal. In cubic crystals, the tensor becomes a scalar, and, for free electrons, Eq. (20) becomes the simple Drude expression

$$\sigma = 1/\rho = n_e e^2\tau/m, \tag{21}$$

where n_e is the electron density and m the electron mass.

2.3.2 Thermoelectric Effects If the temperature gradient is nonzero, Eqs. (14) and (16) give

$$\mathbf{v}(\mathbf{k})\cdot\frac{\partial f_0}{\partial T}\boldsymbol{\nabla}_rT + e\frac{\partial f_0}{\partial \epsilon_{\mathbf{k}}}\mathbf{v}(\mathbf{k})\cdot\mathbf{E} = \frac{-g_{\mathbf{k}}}{\tau_{\mathbf{k}}}. \tag{22}$$

This may be substituted into Eq. (18) to give the electric current, which is now Eq. (19) plus an additional term

$$\frac{e^2}{4\pi^3\hbar}\int \tau_{\mathbf{k}}\mathbf{v}(\mathbf{k})\left[\mathbf{v}(\mathbf{k})\frac{\partial f_0}{\partial T}\cdot\boldsymbol{\nabla}_rT\right]d^3k, \tag{23}$$

i.e., a temperature gradient $\boldsymbol{\nabla}_rT$ alone produces an electric current through the thermoelectric effect. Comparison with Eq. (7a) (with $\mathbf{E} = 0$) allows determination of γ.

2.3.3 Electronic Thermal Conductivity An additional effect of a temperature gradient is to cause a flow of heat. The heat current density $\mathbf{j}_q$ is then

$$\mathbf{j}_q = \frac{1}{4\pi^3}\int(\epsilon_{\mathbf{k}} - \mu)\mathbf{v}(\mathbf{k})g_{\mathbf{k}}d^3k, \tag{24}$$

since the heat current is just the electron flux times the difference between the kinetic energy per electron $\epsilon_{\mathbf{k}}$ and the chemical potential μ of the electron system.

By analogy with the previous derivations, one obtains the electronic thermal conductivity (see e.g., Ziman, 1972). In cubic materials,

$$\kappa = \frac{\pi^2}{3}\frac{n_e\tau}{m}k_B^2T, \tag{25}$$

where k_B is the Boltzmann constant.

Comparison with Eq. (21) allows definition of the Wiedemann-Franz ratio

$$\frac{\kappa}{\sigma} = \frac{\pi^2}{3}\frac{k_B^2 T}{e^2}. \tag{26}$$

Because it was derived from the relaxation-time approximation, Eq. (26) is valid only at very low or very high temperatures (see e.g., Ziman, 1972).

2.3.4 Hall Effect In the presence of both electric and magnetic fields, Eq. (14) becomes

$$e\frac{\partial f_0}{\partial \epsilon_{\mathbf{k}}}\mathbf{v}(\mathbf{k})\cdot\mathbf{E} = \frac{-g_{\mathbf{k}}}{\tau_{\mathbf{k}}} - \frac{e}{\hbar}[\mathbf{v}(\mathbf{k})\times\mathbf{B}]\cdot\frac{\partial g_{\mathbf{k}}}{\partial \mathbf{k}}. \tag{27}$$

This equation cannot be solved directly for $g_{\mathbf{k}}$. However, by using a trial function of the same form as Eq. (17),

$$g_{\mathbf{k}} = -e\tau\frac{\partial f_0}{\partial \epsilon_{\mathbf{k}}}\mathbf{v}(\mathbf{k})\cdot\mathbf{A}, \tag{28}$$

equation (27) becomes

$$\mathbf{v}(\mathbf{k})\times\mathbf{E} = \mathbf{v}(\mathbf{k})\times\mathbf{A} + \frac{e\tau}{m}[\mathbf{v}(\mathbf{k})\times\mathbf{B}]\cdot\mathbf{A}, \tag{29}$$

which has a solution

$$\mathbf{E} = \mathbf{A} + (e\tau/m)(\mathbf{B}\times\mathbf{A}). \tag{30}$$

By analogy with Eq. (7a)

$$\mathbf{j} = \sigma\mathbf{A}, \tag{31}$$

and Eq. (30) becomes

$$\mathbf{E} = \frac{1}{\sigma}\mathbf{j} + \frac{e\tau}{m\sigma}\mathbf{B}\times\mathbf{j}. \tag{32}$$

This equation shows that there are two components to the field: the normal longitudinal field parallel to **j** and, if **B** is perpendicular to **j**, a transverse field known as the Hall field, which may be written as

$$E_{\mathrm{H}} = \frac{1}{n_e e}|\mathbf{B}||\mathbf{j}| = R_{\mathrm{H}}|\mathbf{B}||\mathbf{j}|. \tag{33}$$

In the relaxation-time approximation, it has the same sign as the charge carrier:

$$R_{\mathrm{H}} = 1/n_e e. \tag{34}$$

A more general analysis of R_{H} is given in Sec. 7.

2.3.5 Magnetoresistance The relaxation-time approximation does not predict any dependence of conductivity upon the magnetic field since the longitudinal electric field (i.e., that component of **E** parallel to **j**) is given simply by $\mathbf{E}_{\mathrm{L}} = \sigma^{-1}\mathbf{j}$ [see Eq. (32)]. However, if there is more than one band of carriers, or if the scattering is not isotropic, there does appear a component of the longitudinal electric field that depends upon **B**. This is just the magnetoresistance introduced in Sec. 1 and discussed further in Sec. 6.

The detailed behavior of the conduction electrons in a magnetic field is further complicated by the fact that the Lorentz force $\mathbf{v}(\mathbf{k}) \times \mathbf{B}$ causes them to orbit around the FS as shown in Fig. 3. Different FS's have very different topologies. These topologies are most important at high fields, where they lead to

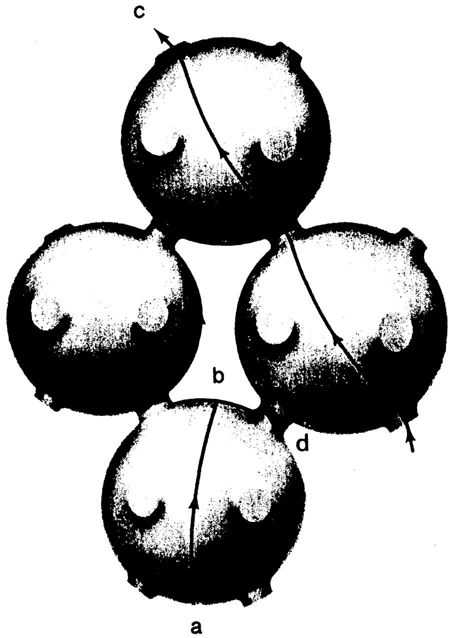

FIG. 3. The FS of Cu. (c) Open orbits, (a) and (d) closed electron-like orbits, and (b) holelike orbits for electrons in a uniform magnetic field (from Ashcroft and Mermin, 1976).

different field dependences of both R_H and MR as discussed in later sections.

3. SCATTERING MECHANISMS

A perfect, infinite lattice would produce no obstacle to the passage of an electric current. Only deviations from a perfect periodicity of the lattice potential cause the scattering of electrons from initial to final states, leading to a reduced conductivity. These deviations may be separated into those that do not have any intrinsic dependence upon thermally induced atomic displacements ("static" atomic or magnetic disorder) and those directly resulting from the thermal motion of atoms (thermally induced disorder).

One form of static disorder in solids results from atomic displacements such as occur around point and extended defects, from interstitials, or from atomic size effects in alloys containing atoms with different atomic radii. Such disturbances are characterized by the total atomic displacements produced.

Liquid metals and amorphous solid metals may be regarded similarly, with the static disruption to the atomic arrangement so severe that there is no long-range translational periodicity, although there will still be some local sense of atomic positional order ranging over the first few neighbors (see AMORPHOUS MATERIALS, STRUCTURE OF; LIQUIDS, SIMPLE, STRUCTURE OF; METALS AND ALLOYS, STRUCTURE OF). In alloys there is an additional disruption to the periodicity of the lattice potential as a result of imperfect spatial ordering of the different atomic potentials. (A structure with perfect long-range atomic order would again have no resistivity.)

Grain boundaries, free surfaces, and interphase and antidomain boundaries also disrupt the periodicity of the potential and so will cause electron scattering. Furthermore, if the scale of a decomposition (e.g., grain size, precipitation, spinodal decomposition, and antiphase domains), exceeds the conduction-electron mean free path $l = v_F\tau$ (where v_F is the velocity of the electron at the FS), it is necessary to regard each region and the boundary as separate elements and to develop models consisting of arrays of these elements to represent the microstructure adequately.

The sensitivity of the scattering process to the scale of the structure or microstructure presents a significant complication. In principle, the Boltzmann equation should be solved within each characteristic region to obtain a "local" mean free path $l_{(\mathrm{loc})}$. However, the scattering term in the Boltzmann equation depends upon the scattering matrix, which in turn depends upon the size of region $\sim l_{(\mathrm{loc})}$ over which the potential is averaged. Thus, a complete solution to the problem strictly requires a self-consistent solution to the equation.

At any realistic temperature, thermal excitations will cause the atoms to be displaced from their equilibrium positions. As far as the conduction electrons are concerned, the adiabatic approximation holds, and the atoms may be considered to be fixed or "frozen" into this distorted configuration. The electron scattering is thus determined by the size of the displacements and hence by the temperature.

Conduction electrons are also scattered by magnetic moments so that any deviation from perfect periodicity in an array of these moments will also contribute to the resistivity. The magnetic structure may be described in terms of spin-spin correlation parameters, in direct analogy to the atomic order parameters described above, or alternatively by the use of spin waves.

4. ELECTRONS IN A MAGNETIC FIELD

4.1 Cyclotron Orbits and $\omega_c\tau$

A fundamental quantity for characterizing electronic transport in the presence of **B** is the dimensionless parameter $l/r_c = \omega_c\tau$, which specifies the average number of times an electron traverses a given cyclotron orbit between scattering events. Here r_c is the cyclotron radius, and ω_c is the cyclotron frequency given by (Abrikosov, 1972; Ashcroft and Mermin, 1976)

$$\omega_c = \frac{h}{2\pi}\frac{\partial A}{\partial \epsilon}, \tag{35}$$

where h is Planck's constant and $\partial A/\partial \epsilon$ is the energy derivative of the cross-sectional area of the cyclotron orbit in **k** space. We must distinguish four different regimes:

1. Low magnetic fields, $\omega_c\tau \ll 1$. In this limit, electrons traverse only a small fraction of a cyclotron orbit before being scattered. Their behavior is determined primarily by the scattering processes and by the properties of the FS in the local vicinity of their wave vectors before and after scattering.
2. Intermediate magnetic fields, $\omega_c\tau \approx 1$. Here, electrons traverse about one full orbit between scattering events. Both scattering processes and surface topology are important, making this usually the most difficult region of **B** in which to analyze data.
3. High magnetic fields, $\omega_c\tau \gg 1$. In this limit, electrons traverse their cyclotron orbits many times between scattering events. The field dependences of both *MR* and R_H (i.e., whether *MR* is proportional to B^0, B^1, or B^2, and whether R_H is proportional to B^0 or B^{-2}) are determined solely by the topological structure of the FS in the plane in **k** space perpendicular to **B**, with details of scattering affecting only the magnitudes of the transport properties.

At low temperatures and high magnetic fields, quantum oscillations appear in transport properties because of Landau quantization of the electronic energy levels. These occur in a fourth limit:

4. The Landau level-quantization limit: $\omega_c\tau > k_BT$, where T is the absolute temperature.

4.2 High-Field Forms of *MR* and R_H

We start by considering metals that in the free-electron approximation have odd numbers of conduction electrons per atom (i.e., one or three), and thus odd numbers per primitive unit cell (which for these metals contains only one atom). Since each Brillouin zone (BZ) can accommodate two electrons/atom, whatever changes in the FS occur when the number of conduction electrons per unit cell N_e increases, the numbers of electrons and holes per atom in such metals must remain unequal; e.g., for Al, $N_h - N_e = 1$. Metals for which $N_e \neq N_h$ are called "uncompensated." In contrast, metals or alloys that in the free-electron approximation have an even number of conduction electrons per primitive unit cell must have $N_e = N_h$. Such materials are called "compensated." The alternative high-field forms of *MR* and R_H for different FS topologies are listed in Table 1 and illustrated in Fig. 4. Note the generally different behaviors for compensated and uncompensated metals. For further details, see Abrikosov (1972, 1988).

5. ELECTRICAL RESISTIVITY

All of the scattering mechanisms discussed previously produce one or both of two effects:

1. a change in the nature of the total scattering potential $W(\mathbf{r})$ at the "defect" (whether a foreign atom or vacancy) and
2. a disruption to the spatial periodicity of the lattice potential.

If the disturbances are not too large, first-order perturbation theory allows a fairly straightforward evaluation of the scattering associated with the defect. Equation (20) then leads to the expression (see e.g., Rossiter, 1987)

$$\rho = \frac{3\pi m^2 \Omega_0 N}{4\hbar^3 e^2 k_F^6} \int_0^{2k_F} |\langle \mathbf{k}+\mathbf{q}|W(\mathbf{r})|\mathbf{k}\rangle|^2 q^3\,dq, \tag{36}$$

where N and Ω_0 are the number and volume of unit cells, k_F the magnitude of the Fermi wave vector (see ELECTRON STRUCTURE OF SOLIDS), the integrand $\langle \mathbf{k}+\mathbf{q}|W(\mathbf{r})|\mathbf{k}\rangle$ represents the matrix elements of the total scattering potential $W(\mathbf{r})$, and the integration is over the magnitude of the scattering wave vector **q** defined by

$$\mathbf{q} = \mathbf{k} - \mathbf{k}'. \tag{37}$$

The scattering potential may be factorized into two terms:

$$\langle \mathbf{k}+\mathbf{q}|W(\mathbf{r})|\mathbf{k}\rangle = \mathfrak{S}(\mathbf{q})\langle \mathbf{k}+\mathbf{q}|w(\mathbf{r})|\mathbf{k}\rangle, \tag{38}$$

where $\mathfrak{S}(\mathbf{q})$ is the structure factor

$$\mathfrak{S}(\mathbf{q}) = \frac{1}{N}\sum_{\mathbf{r}} \exp(-i\mathbf{q}\cdot\mathbf{r}) \tag{39}$$

and depends only upon atom positions, and $\langle \mathbf{k}+\mathbf{q}|w(\mathbf{r})|\mathbf{k}\rangle$ [often written simply as $w(\mathbf{q})$] is the form factor that is determined entirely by the atomic potential $w(\mathbf{r})$ associated with site **r**.

Table 1. Types of behavior of the Hall and transverse magnetoresistance effects in metals (from Hurd, 1972).

Type	Behavior in the high-field condition		Nature of orbits in planes normal to the applied field direction; state of compensation
	Magnetoresistance[a]	Hall coefficient	
Single-crystal sample			
1	$\Delta\rho/\rho(0)$ saturates with H	$R_H \propto (n_e - n_h)^{-1}$	All closed; $N_e \neq N_h$
2	$\Delta\rho/\rho(0)$ tends to H^2 variation (transverse)	R_H tends to a constant[b]	All closed; $N_e = N_h$
	$\Delta\rho/\rho(0)$ saturates with H (longitudinal)		
3	$\Delta\rho/\rho(0)$ saturates with H	$R_H \propto (n_e - n_h)^{-1}$	Negligible number of open orbits; $N_e \neq N_h$
4	$\Delta\rho/\rho(0)$ tends to $H^2 \sin^2\theta$[c] variation	R_H tends to a constant[b]	Open orbits in one direction only
5	$\Delta\rho/\rho(0)$ saturates with H	$R_H \propto H^{-2}$	Open orbits in more than one direction
Polycrystalline sample			
6	$\Delta\rho/\rho(0) \propto H$	R_H tends to a constant[b]	All types if the cyrstallites have random orientations
	Behavior in the low-field condition		
Single-crystal or polycrystalline sample			
7	$\Delta\rho/\rho(0) \propto H^2$	R_H tends to a constant[d]	It is irrelevant what type of orbit predominates

[a] Transverse magnetoresistance except as indicated for type 2.

[b] The Hall coefficient is not related to the effective number of carriers in any simple manner.

[c] Note that $\Delta\rho/\rho(0)$ tends to saturation when θ, the angle between the applied electric field and the axis of the open orbit in real space, is zero.

[d] The Hall coefficient depends upon the anistropy of the dominant electron scattering process and the electron's velocity and effective mass at each point on the FS.

[e] N_e and N_h are, respectively, the number of electrons and holes per unit cell of the Bravais lattice; n_e and n_h are, respectively, the density of electrons and holes in real space.

It must be emphasized that Eq. (36) and the simple separation of Eq. (38) are only strictly valid for metals with a simple electronic structure (i.e., nearly-free-electron-like bands) and isotropic scattering over a spherical FS. These requirements are not satisfied in many metals and alloys, particularly those involving transition metals, where more complex methods are needed.

If the scattering mechanisms are independent and the scattering isotropic, the contributions to the total resistivity are simply additive. In an alloy, this leads to Matthiessen's rule

$$\rho_{\mathrm{TOT}} = \rho_0 + \rho_p(T), \tag{40}$$

where all of the static mechanisms are collected into the "residual" resistivity ρ_0, and $\rho_p(T)$ is the temperature-dependent phonon-scattering term. The assumption of independent scattering is only partly true and there is a large body of work devoted to studying deviations from this rule (see e.g., Bass, 1972).

5.1 Pure Metals

5.1.1 Static Displacements A vacancy at some site $\mathbf{r}'$ removes the average lattice potential $\bar{w}(\mathbf{r}')$ from that site (neglecting the effect of charge redistribution) and causes lattice distortions around that site leading to a change in the structure factor $\mathfrak{S}(\mathbf{q})$. The former produces a positive contribution to the resistivity that is reduced somewhat by lattice relaxation. Values for the resistivity due to 1 at. % of vacancies in some pure metals are given in Table 2. Since a typical equilibrium vacancy density is only about $1 \times 10^{13}/\mathrm{cm}^3$ ($\sim 10^{-9}$ at. %) at 300 °C and $2.7 \times 10^{16}/\mathrm{cm}^3$ ($\sim 10^{-6}$ at. %) at 600 °C, the total contribution of the equilibrium vacancies to ρ_0 is very small.

The resistivity contribution due to dislocations may be considered as due to the potential discontinuity at the core and the effects of atomic displacements in the surrounding strain field, with the former probably dominant (see e.g., Brown, 1977). Some

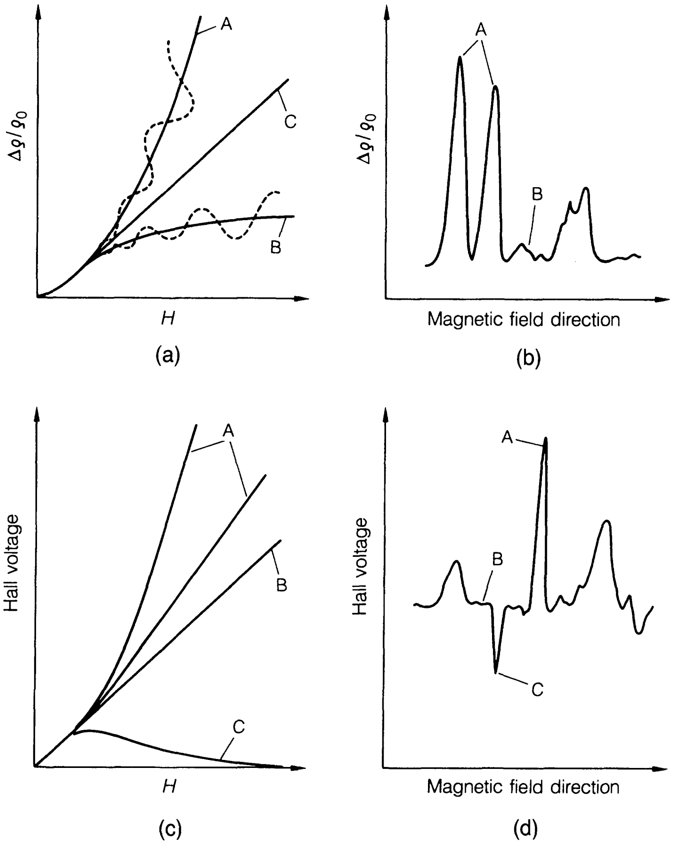

FIG. 4. Types of field dependence observed experimentally for the transverse magnetoresistance and the Hall voltage (from Hurd, 1972). (a), (b) *A* illustrates a quadratic (i.e., B^2) high-field variation in $\Delta\rho$, *B* illustrates saturation in $\Delta\rho$, and *C* illustrates the linear (i.e., *B*) variation in $\Delta\rho$, which is generally observed in polycrystalline samples. Quantum oscillations [dashed curves in (a)] may be superimposed upon behaviors *A* and *C* for single-crystal samples. (c), (d) *A* indicates high-field linear variations of different magnitudes; *B* illustrates a linear variation extending all the way down to $B = 0$; and *C* illustrates a B^{-1} high-field variation (from Hurd, 1972).

data are compared with the prediction of a resonant scattering model due to Brown in Table 3. Since the dislocation density in a heavily cold-worked metal is of order $10^{12}/\text{cm}^2$, the maximum dislocation resistivity is ~0.1 $\mu\Omega$ cm.

The resistivity due to grain boundaries has been considered using a number of models, including an array of scattering potentials or an array of dislocations. Some results for a number of metals are given in Table 4, along with calculated values based simply upon the assumption that the grain boundary is represented by an array of dislocations and that these act as independent scatterers.

Table 2. Electrical resistivity due to 1 at. % of vacancies in the metals indicated.

Metal	ρ_0 (μΩ cm)/at. %	References
K	1.9–2.1	Benedek and Baratoff (1971)
Sn	4.4	Sun and Ohring (1976)
Al	1.0–3.3	Benedek and Baratoff (1971)
Cu	1.0–3.0	Simmons and Balluffi (1960)
Ag	1.3	Dugdale (1977)
Au	1.5	Dugdale (1977)

Mechanisms proposed for the scattering of conduction electrons at free surfaces, which become important in thin foils or fine wires, include microscopic surface roughness and localized surface charges (see e.g., Ziman, 1960; Greene, 1964). Such approaches often lead to the notion of a specularity parameter *p* that describes the fraction of electrons that are scattered in a specular (i.e., nondiffuse) manner. The resistivity or conductivity of a foil or wire can then be written in the form

$$\rho_0(d,T)/\rho_\infty = 1 + A(p,l_\infty,d)\, l_\infty/d, \qquad (41)$$

where the parameter $A(p,l_\infty,d)$ depends upon the model used, d is the foil or wire thickness, l_∞ is the bulk mean free path, and ρ_∞ and σ_∞ are the corresponding bulk values of the

Table 3. Dislocation specific resistivities per unit dislocation density N (measured in lines/cm^2). These figures represent averages over all dislocation orientations. (From Brown, 1977, 1982.)

Metal	Temp. (K)	ρ_0/N 10^{-13} $\mu\Omega$ cm^3	Theory (Brown, 1977)
K	4.2	4	8
Cu	4.2	1.6 ± 0.2	1.3
Ag	4.2	1.9	1.9
Au	4.2	2.6	1.9
Be	80	34	28
Cd	80	24	25
Al	4.2	1.7 ± 0.3	1.8
Zr	80	100	40
Ti	4.2	100	29
Pb	80	1.1	4.2
Bi	1.3	2×10^5	1.7×10^5
Mo	4.2	5.8	3.7
W	4.2	7.5	7.4
Pt	80	9	4
Fe	80	10 ± 4	1.9
Ni	80	10	1.1
Rh	80	32	1

resistivity and conductivity. The dependence of A upon d/l_∞ for some different models is discussed in other literature (Rossiter, 1987). Note that this size-effect scattering. Eq. (41), represents an explicit deviation from Matthiessen's rule since it depends explicitly upon T.

A similar form of analysis may also be applied to grain boundaries where the grain-boundary resistivity may be expressed in the form of Eq. (41) where d is now the mean grain diameter. The corresponding variation of $A(p,l_\infty,d)$ with d/l_∞ is shown in Fig. 5. This size effect is demonstrated in Fig. 6, which shows the normalized grain-boundary resistivity as

Table 4. Grain-boundary specific resistivities in terms of area of boundary per unit volume S (measured in cm^2/cm^3) (from Brown, 1982).

Metal	ρ_0/S 10^{-12} Ω cm^2	Temp. (K)	Theory (Brown, 1977) (10^{-12} Ω cm^2)
Cu	2.4	4.2	2.7
Au	3.5	300	3.9
Cd	17	4.2	32
Al	1.1–2.4	4.2	2.6
Pd	0.3, 1.3	5–20	1.4
Bi	6.9×10^4	77	17×10^4
W	20	77	22
Fe	80–160	4.2	33
Ni	140	77	17

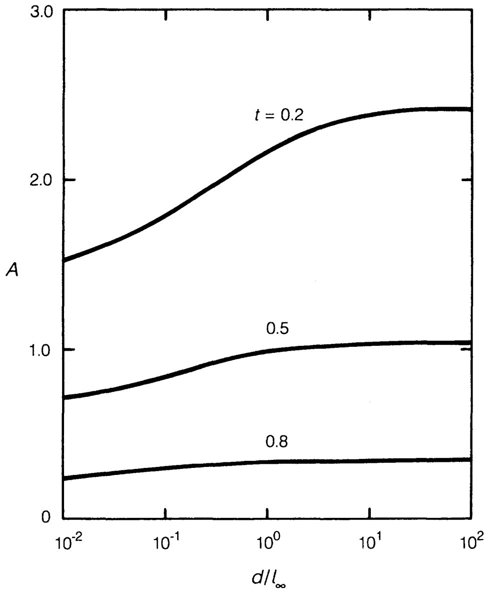

FIG. 5. The parameter A as a function of d/l_∞ from the grain-boundary scattering model for various values of the transmission coefficient t (from Rossiter, 1987).

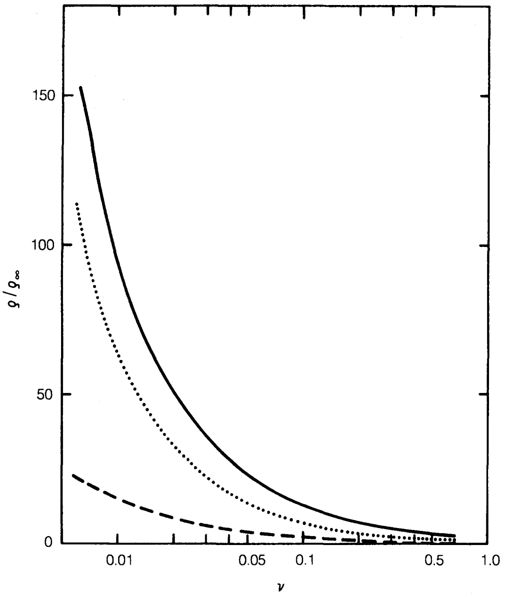

FIG. 6. The normalized grain-boundary resistivities as functions of the grain-size–dependent parameter ν. Solid line, $\rho_g(\nu)/\rho_\infty$; dotted line, $\rho_\perp(\nu)/\rho_\infty$; dashed line, $\rho_\parallel(\nu)/\rho_\infty$ where $\rho_\perp$ and $\rho_\parallel$ are the resistivities perpendicular and parallel to the current flow; ρ_g is the average resistivity for a three-dimensional array of boundaries, and ρ_∞ is the resistivity for infinite boundary spacing (from Rossiter, 1987).

a function of the grain-size–dependent parameter $\nu = d/l_\infty ln(1/t)$, where t is an empirical specular transmission coefficient.

5.1.2 Thermal Effects The effects of thermal scattering may be incorporated into the structure factor by allowing the atom to be displaced away from the lattice site $\mathbf{r}_i$ by a distance δ_i, obtained as a sum over all phonon wave vectors and modes. Such information is available either from a solution of the dynamical matrix based upon known force constants or from experimentally determined phonon dispersion curves (see PHONONS IN CRYSTAL LATTICES).

An alternative approach is to write the structure factor directly in terms of the dynamical matrix, allowing $\rho_p(T)$ to be determined directly from known force constants. Once again, for all but the simple metals, it is necessary to go beyond the first-order perturbation-theory approach.

Some experimental results for Cu are shown in Fig. 7 and indicate that reasonable agreement between theory and experiment is now possible. A compilation of results for a number of pure metals is given in Bass (1983).

A simpler though less satisfactory approach is to use the Debye (or Einstein) model to evaluate the thermal displacements, leading to the general prediction of $\rho_p(T) \propto T^5$ at low temperatures ($T < \theta_D$) and $\rho_p(T) \propto T$ at higher temperatures, in general agreement with experiment for free-electron-type metals, although the predicted magnitudes of the resistivity are usually not correct. Nevertheless, this approach leads to a useful phenomenological expression due to Bloch (1930), Grüneisen (1933), and Wilson (1937):

$$\frac{\rho_p(T)}{\rho(\theta_D)} = 4.225\left(\frac{T}{\theta_D}\right)^5 \int_0^{\theta_D/T} \frac{z^5 dz}{(e^z - 1)(1 - e^{-z})}, \tag{42}$$

where $\rho(\theta_D)$ is the resistivity at the Debye temperature. In practice, a better fit between experiment and theory is obtained by allowing θ_D in Eq. (42) to vary slightly from the actual Debye temperature (as obtained from specific-heat measurements, for example). The corrected characteristic "resistivity" temperatures, usually denoted θ_R, for some metals are shown in Table 5 and the results plotted in Fig. 8. Some other results are given in Meaden (1965).

At room temperature, $\rho_p(T) \sim 1\ \mu\Omega$ cm which is at least one order of magnitude larger than the contribution from lattice defects, even in severely cold-worked materials.

5.2 Alloys

Substitutional and interstitial impurities lead to changes in the form factor as well as the structure factor through the local lattice

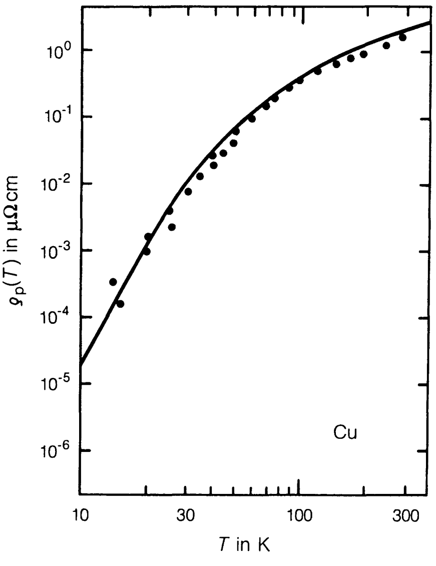

FIG. 7. The temperature dependence of the resistivity of Cu derived from the form factors of Borchi and de Gennaro (1970) (solid line). The solid circles represent a collection of experimental data (from Rossiter, 1987).

Table 5. Some representative high-temperature values for θ_R (valid for $\theta_R/3$ to θ_R) compared with corresponding values for θ_D (from Meaden, 1965).

Metal	θ_D (K)	θ_R (K)
Be	1000	1240
Na	160	195
Mg	325	340
Al	385	395
K	100	110
Ca	225	~145
Ti	355	342
Cr	450	485
Cu	320	320
Zn	245	175
Pd	300	270
Ag	220	200
Cd	165	130
W	315	333
Pt	225	240
Au	185	200
Pb	88	~100

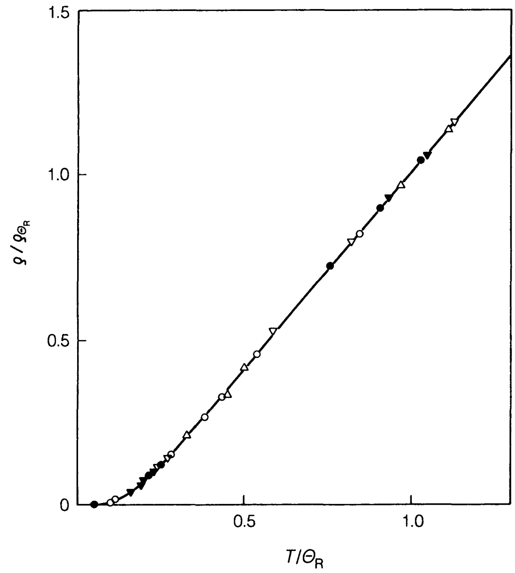

FIG. 8. The temperature dependence of the resistivity ρ/ρ_θ plotted against T/θ_R. ●, Li, $\theta_R = 363$ K; ○, Na, $\theta_R = 202$ K; ▽ Cu, $\theta_R = 333$ K; △ Au, $\theta_R = 175$ K; ▼ Pb, $\theta_R = 86$ K (from Olsen, 1962). These values of θ_R differ from those shown in Table 5 as the data have been fitted by Eq. (42) over a range of lower temperatures.

distortions. In dilute alloys, one finds that the residual resistivity contribution varies linearly with impurity concentration,

$$\rho_0 \propto C_i, \tag{43}$$

where C_i is the concentration of the impurity atoms. This expression may be obtained directly from Eq. (36) since, if the effects of lattice strain are neglected,

$$\rho_0 = \frac{3\pi m^2 \Omega_0}{4\hbar^3 e^2 k_F^6} C_i \int_0^{2k_F} |w_i(\mathbf{q}) - \bar{w}(\mathbf{q})|^2 q^3 dq \tag{44}$$

where $w_i(\mathbf{q})$ and $\bar{w}(\mathbf{q})$ are the form factors for the impurity and host, respectively. This linear relationship is well obeyed up to a few atomic percent; see, e.g., Bass (1983, 1985).

From Eq. (44), one expects a variation of ρ_0 with the square of the magnitude of the scattering potential. Substitution of form factors for some simple model potentials (such as screened Coulomb or empty core potentials) leads to the Linde-Norbury rule, i.e., the impurity resistivity should depend upon the square of the difference in valency Δz between the solute and host,

$$\rho_0 \propto \Delta z^2. \tag{45}$$

Such simple behavior (Fig. 9) is generally observed only in free-electron-like metals.

In more concentrated binary alloys, the form factor of interest turns out to be the difference in the form factors of the alloy components, in which case

$$\rho_0 = \frac{3\pi m^2 \Omega_0}{4\hbar^3 e^2 k_F^6} N \int_0^{2k_F} |\mathfrak{S}^d(\mathbf{q})|^2 \times |w_A(\mathbf{q}) - w_B(\mathbf{q})|^2 q^3 dq, \tag{46}$$

where $\mathfrak{S}^d(\mathbf{q})$ is the structure factor that describes the spatial distribution of the A and B atoms. Since the form factor is squared in Eq. (46), a random alloy containing 1 at. % of A in B should lead to the same resistivity as 1% B in A. This is known as Mott's rule.

Short- and long-range atomic correlations in concentrated alloys directly affect the structure factor, leading to expressions of the form

$$\rho_0 = \frac{3\pi m^2 \Omega_0}{4\hbar^3 e^2 k_F^6} C_A C_B \int_0^{2k_F} \sum_i c_i \alpha_i \frac{\sin(qr_i)}{qr_i} \times |w_A(\mathbf{q}) - w_B(\mathbf{q})|^2 q^3 dq, \tag{47}$$

where c_i is the number of atoms in the ith

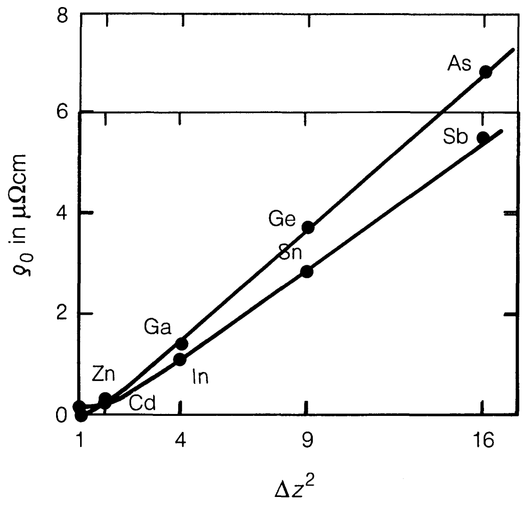

FIG. 9. The change in resistivity for 1 at. % impurity in Cu as a function of $|\Delta z|^2$ (from Rossiter, 1987; using the data of Linde, 1932).

coordination shell of radius r_i, and α_i is the Warren-Cowley short-range order parameter (Warren, 1969). In a random alloy, there is no order; $\alpha_0 = 1$ and $\alpha_i = 0$ for all $i > 0$. Equation (47) then gives Nordheim's rule:

$$\rho_0 \propto C_A C_B. \tag{48}$$

In concentrated alloys containing transition metals, it is necessary to take into account the composition dependence of the band structure and also to allow for the fact that the charge carriers may occupy (and scatter between) a number of energy bands. Examples of data for Pd-Au and Pd-Ag alloys are given in Fig. 10.

The presence of short-range atomic ordering or clustering may lead to either an increase or a decrease in the resistivity from the random-alloy value. However, when the size of the ordered or clustered region becomes comparable with l, any further increase in the degree of atomic correlation (whether clustering or ordering) will cause the resistivity to decrease since this represents a decrease in the degree of disorder on a scale of l. This form of behavior is typified in the changes in resistivity that occur during Guinier-Preston zone formation and growth (see e.g., Porter and Easterling, 1981), as shown in Fig. 11. As the zones evolve, scattering can become quite anisotropic over the Fermi surface (see Sec. 2.3), necessitating a more detailed calculation (Rossiter, 1987).

In the case of long-range order, Eq. (47) leads to

$$\rho_0(s) = \rho_0(0)(1 - s^2), \tag{49}$$

where $\rho_0(s)$ is the residual resistivity of an alloy with Bragg-Williams long-range order parameter s (Warren, 1969). Since complete long-range ordering ($s = 1$) is only possible at stoichiometric compositions, one obtains a composition dependence of the residual resistivity of fully ordered specimens as shown in Fig. 12.

If a material separates into two phases, it is necessary to assume some model of the microstructure in order to calculate the resistivity. Some results for a two phase mixture of Mg_2Pb–Pb are shown in Fig. 13.

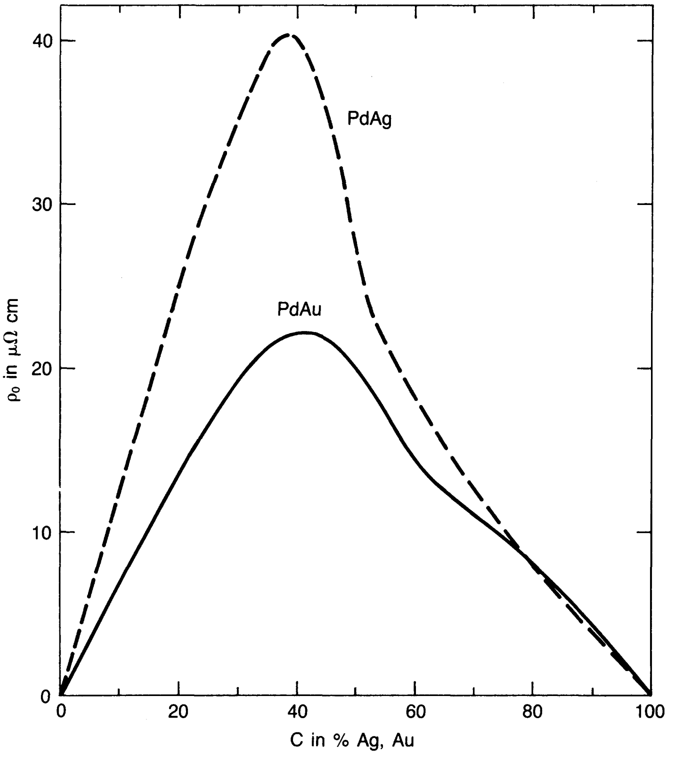

FIG. 10. Composition dependence of the residual resistivity of Pd-Au (solid line) and Pd-Ag (dashed line) (from data of Coles and Taylor, 1962).

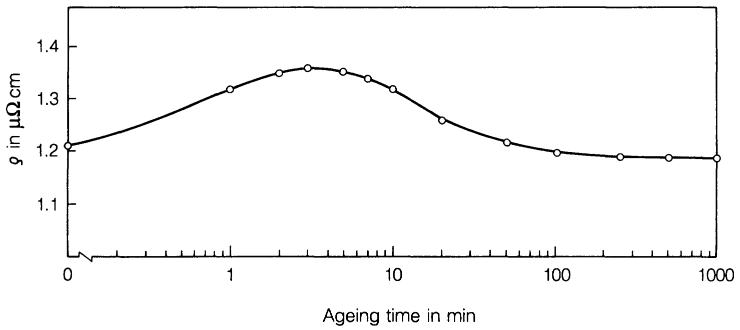

FIG. 11. Electrical resistivity measured at 4.2 K for Al–4.9 at. % Zn as a function of aging time (aging temperature 20 °C) (from Osamura et al., 1973).

5.3 Magnetic and Nearly Magnetic Metals and Alloys

Scattering from disordered magnetic moments in the transition metals is complicated by their decidedly non-free-electron band structure and the associated interband scattering effects. The conduction electrons may be classified into two groups having spin directions either parallel (spin "up," ↑) or antiparallel (spin "down," ↓) to the local magnetization. Elastic scattering of the electrons will not change the population of the subbands, but inelastic (spin-flip) scattering will result in a transfer of momentum between the two groups of electrons.

In the simplest approximation, it is assumed that the conduction band contains nearly free electrons. The current is then carried by the spin-up and spin-down bands in parallel. Solution of the Boltzman equation

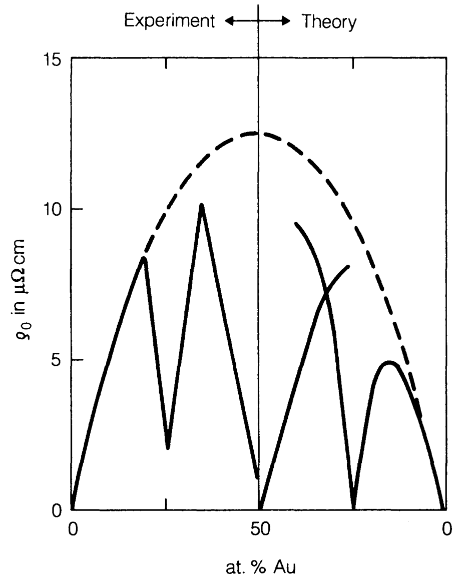

FIG. 12. Variation of resistivity with composition for long-range–ordered Cu-Au alloys. The experimental results of Johansson and Linde (1936) are shown in the full curve on the left, whereas the theoretical curve is shown on the right-hand side. The resistivity of a disordered alloy is shown as the dashed curve. The theoretical curve is scaled so that the disordered resistivities are the same at Cu-50%Au. The experimental data have been reduced by 1.7 $\mu\Omega$ cm to give $\rho = 0$ for pure Cu (from Rossiter, 1987).

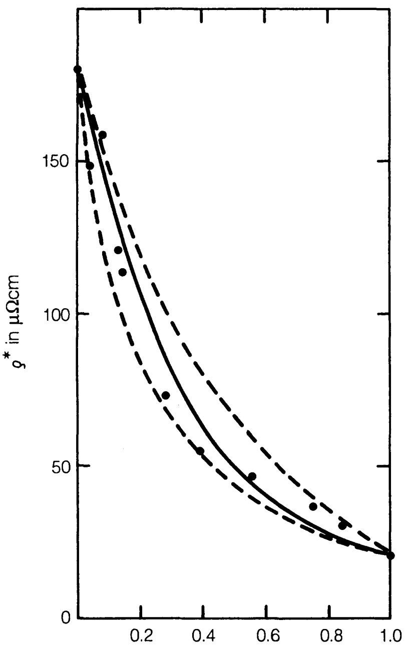

FIG. 13. Resistivity of two-phase Mg_2Pb–Pb mixture as a function of volume fraction of Pb: dashed lines, upper and lower bounds; solid line, effective-medium theory for random spheres; points, experimental data (from Rossiter, 1987, using the data of Hashin and Shtrikman, 1962).

for each band then leads to

$$\rho_{\mathrm{TOT}} = \frac{\rho_{\uparrow}\rho_{\downarrow} + \rho_{\uparrow\downarrow}(\rho_{\uparrow} + \rho_{\downarrow})}{\rho_{\uparrow} + \rho_{\downarrow} + 4\rho_{\uparrow\downarrow}}, \tag{50}$$

where $\rho_{\uparrow\downarrow}$ describes the contributions of the spin-flip processes and each resistivity $\rho_{\uparrow}, \rho_{\downarrow}$ is given by the sum of residual and phonon-scattering contributions:

$$\rho_{\uparrow} = \rho_{0\uparrow} + \rho_{p\uparrow}(T), \tag{51}$$

and similarly for $\rho_{\downarrow}$ (i.e., Matthiessen's rule is assumed valid within each sub-band). Note that the spin-disorder resistivity is contained in $\rho_{0\uparrow}$ and $\rho_{0\downarrow}$ and that different sub-band resistivities will lead to a dependence of ρ_{TOT} upon the direction of the current in relation to the direction of magnetization.

The phonon contribution to the resistivities is complicated by band-structure effects. At low temperatures, the excitations within the spin system are spin waves (magnons), which lead to inelastic scattering. This mechanism introduces a temperature-dependent resistivity $\rho_{p\uparrow\downarrow}(T)$ that must freeze out as $T \rightarrow 0$ K. Other possible contributions to $\rho_{\uparrow\downarrow}$ include up-electron–down-electron interactions and spin-orbit interactions, although these are expected to be negligible in all but very dilute alloys. Thus, as $T \rightarrow 0$, all temperature-dependent parts also go to zero and

$$\rho_0 = \frac{\rho_{0\uparrow}\rho_{0\downarrow}}{\rho_{0\uparrow} + \rho_{0\downarrow}}. \tag{52}$$

Note that even though each of the sub-bands is assumed to obey Matthiessen's rule, the total resistivity does not.

At high temperatures, there will be complete spin mixing and

$$\rho_{\mathrm{TOT}} = 1/4[\rho_{0\uparrow} + \rho_{0\downarrow} + \rho_{p\uparrow}(T) + \rho_{p\downarrow}(T)]. \tag{53}$$

By varying temperature and composition, it is possible to separate out the different contributions. Some results for Ni are shown in Table 6.

As in the case of long-range atomic ordering, long-range antiferromagnetic ordering may introduce new energy gaps at superlattice BZ boundaries that will modify this behavior and possibly result in an increase in ρ_{TOT} just below the Néel temperature. Experimental data given in Meaden (1965) and Shröder (1983) indicate a rich variation in the types of behaviors observed.

Competing ferromagnetic and antiferromagnetic interactions can give rise to an amorphous spin structure—a spin glass. At low temperatures, a $T^{3/2}$ behavior has been attributed to scattering from long-wavelength magnon modes, damped ferromagnetic modes, and discrete excitations. However, a T^2 behavior is also sometimes observed, supposedly as a result of random spin reversals at higher temperatures. In many such systems, spin fluctuation effects are also important, leading to low-temperature resistivity minima and maxima.

5.4 Liquid and Amorphous Alloys

Liquid and amorphous alloys exhibit a wide range of behaviors (see, e.g., Mizutani, 1983; Naugle, 1984).

In many nearly-free-electron materials, the resistivity may be evaluated by substituting appropriate static or dynamical structure factors (obtained from diffraction experiments or by calculation) into Eq. (38) and then using Eq. (36). However, this diffraction approach is not valid if $l \sim$ lattice spacing, since the electron wave packets and associated wave vectors **k** are no longer clearly defined, and one would expect a localized (rather than extended) state model to be more realistic. The transition should occur where $\rho_{\mathrm{TOT}} \approx 100$–$160$ $\mu\Omega$ cm. (The same statement applies to high-

Table 6. Values of the parameters used to interpret the sub-band resistivities of Ni-based alloys (after Fert and Campbell, 1976).

Temperature (K)	$\rho_{\uparrow\downarrow}(T)$ ($\mu\Omega$ cm)	$\rho_0(T)$ ($\mu\Omega$ cm)	$\rho_{p\uparrow}(T)$ ($\mu\Omega$ cm)	$\rho_{p\downarrow}(T)$ ($\mu\Omega$ cm)
77	0.9 ± 0.3	0.32	0.38	1.9
200	5 ± 2	2.5	3	15
300	11 ± 4	5.4	6.7	27

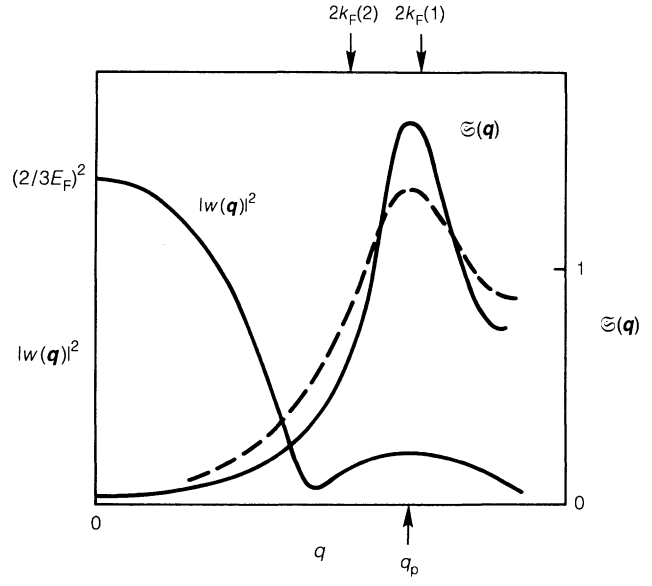

FIG. 14. Schematic representation of the factors $|w(q)|^2$ and $\mathfrak{S}(q)$ that appear in the integrand of the resistivity integral. The structure factor $\mathfrak{S}(q)$ is represented by the full line at the temperature T_1, and the dashed line at $T_2 > T_1$ (from Rossiter, 1987).

resistance crystalline materials—see, e.g., Rossiter, 1987.) Nevertheless, where the model may be applied it can explain the observed positive and negative temperature coefficients of resistivity. In particular, Fig. 14 gives a schematic representation of the structure factor at two different temperatures. The resistivity given by Eq. (36) is heavily weighted by the q^3 term in favor of the $q = 2k_F$ region. Thus, for systems with $2k_F$ indicated by $2k_F(1)$ on the diagram, the resistivity will increase with increasing temperature, whereas with $2k_F(2)$, a negative $d\rho/dT$ will result. From this simple argument, it is expected that the behavior is determined by the position of $2k_F$ in relation to the peak in the structure factor at q_p. Such a correlation is in fact observed, as illustrated in Fig. 15. For the same reason, divalent metals exhibit negative $d\rho/dT$ (at constant volume) whereas others have positive $d\rho/dT$.

In amorphous metals, two competing effects appear at temperatures well below θ_D. The phonon scattering goes as $+T^2$, but the static structure factor is modified by a Debye-Waller factor $\exp(-2M)$, the importance of which depends upon the size of ρ_0. At low temperatures, $M \propto T^2$, and so there is a direct

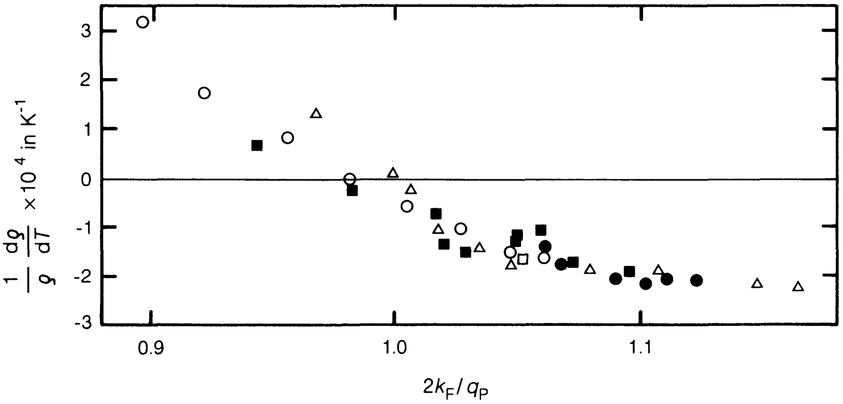

FIG. 15. $d\rho/dT$ as a function of $2k_F/q_p$ for a number of simple metallic glasses: ○: Ag-Cu-Mg; △: Ag-Cu-Al; ■: Ag-Cu-Ge; ●: Mg-Zn; □: Mg-Cu (from Rossiter, 1987, using the data of Mizutani, 1983).

competition between the $+T^2$ phonon term and $-T^2$ Debye-Waller term. If ρ_0 is small, the phonon term is dominant and gives an overall $+T^2$ dependence, but if ρ_0 is large, a $-T^2$ dependence will prevail.

The resistivity of Fe- and Ni-based ferromagnetic amorphous alloys tends to vary as $T^{3/2}$ at temperatures less than $\sim T_C/2$, where T_C is the Curie temperature, and then varies as T^2 up to T_C. The former results from spin-wave scattering, which breaks down into random spin reversals at higher temperatures leading to the T^2 variation.

Many amorphous metals exhibit resistivity minima at low temperatures. This behavior does not appear to depend upon whether the material is paramagnetic or ferromagnetic, and behavior is insensitive to composition, suggesting that it is a fundamental property of the disordered lattice rather than some specific fluctuation effect.

Because the scattering due to thermal motion of the solid lattice is already quite large at temperatures near the melting point, the change in ρ upon melting is not very large, usually less than a factor of 2 (see, e.g., Fig. 24). Liquid alloys tend to show a similar dependence upon composition to that in their disordered solid alloy counterparts.

5.5 ac Effects

The equations described previously are not significantly changed for time-varying (i.e., ac) fields for frequencies up to megahertz, except that the field inside a conductor will now decay exponentially from the surface into the conductor. A measure of the depth of penetration is the classical skin depth δ_0 given by

$$\delta_0 = c(2\pi\sigma\omega)^{-1/2} \tag{54}$$

and is valid for $\delta_0 \gg l$. In this expression, ω is the frequency of the applied electric field, and c is the speed of light. If δ_0 is less than the specimen size, the cross section of specimen-carrying current will differ from the specimen cross section (see Fig. 1), significantly complicating the determination of resistivity.

At frequencies in the microwave range, the field penetration will become dependent upon features of the FS ("anomalous" skin-depth effect), and cyclotron resonance will occur when the electric field equals the cyclotron frequency ω_c [Eq. (35)]. At still higher frequencies ($\omega\tau \sim 1$), one should replace the conductivity σ by a frequency-dependent (ac) conductivity

$$\sigma(\omega) = \frac{\sigma_0}{1 - i\omega\tau}. \tag{55}$$

Finally if $\omega\tau \gg 1$, oscillating excitation of the conduction electrons (charge density waves, or "plasmons") may occur at the plasma frequency ω_p given by

$$\omega_p^2 = 4\pi\sigma_0/\tau. \tag{56}$$

The material will become transparent to electromagnetic radiation at frequencies above ω_p.

6. MAGNETORESISTANCE (*MR*), TRANSVERSE (T-) AND LONGITUDINAL (L-)

There are two separate magnetoresistances (*MR*): transverse (T-*MR*), where the magnetic field is applied perpendicular to the current direction; and longitudinal (L-*MR*), where the field is applied parallel to the current.

6.1 Kohler's Rule

If the electrons on the spherical FS of a free-electron metal are all scattered with a single relaxation time τ, then both the T-*MR* and L-*MR* are zero as noted in Sec. 2.3. This occurs because the average force eE_y produced on an electron by the Hall voltage V_y exactly cancels the Lorenz force $e\langle v_x\rangle B_z$ acting upon the same electron. Here, B_z is the applied magnetic field, pointing in the z direction, and $\langle v_x\rangle$ is the electron's average drift velocity in the x direction. If the electrons occupy a more complex FS, with different values of the effective mass m^* and different relaxation times for electrons on different parts of the FS, then the single Hall field cannot simultaneously cancel the different Lorenz forces felt by different electrons, and nonzero T-*MR* and L-*MR* both appear.

If the electron scattering on a real FS can still be approximated by a single relaxation time, the *MR* should be a function only of the unitless quantity $\omega_c\tau$ defined in Sec. 4, and all data for a given metal should fall on a single curve when the *MR* is plotted as a function of $\omega_c\tau$. Such behavior, called Kohler's rule, is found to be experimentally valid under somewhat more general conditions than just described. Figure 16 shows how Kohler's rule applies to the T-*MR* of polycrystalline samples of In at temperatures ranging from 2 to 20 K. Similar plots of the T-*MR* and the L-*MR* of a variety of polycrystalline metals are shown in Figs. 17 and 18, respectively. The abscissa scale in these figures is proportional to $\omega_c\tau$.

6.2 Low Magnetic Fields ($\omega_c\tau \ll 1$)

At very low fields, the general theory of *MR* predicts a B^2 increase of the *MR* independent of the FS topology and whether the sample is a single crystal or polycrystalline. This form automatically satisfies the Onsager-relation requirement that the *MR* must be an even function of **B**. Examples of approximately B^2 variations are shown in the low-field regions of Figs. 16 and 17.

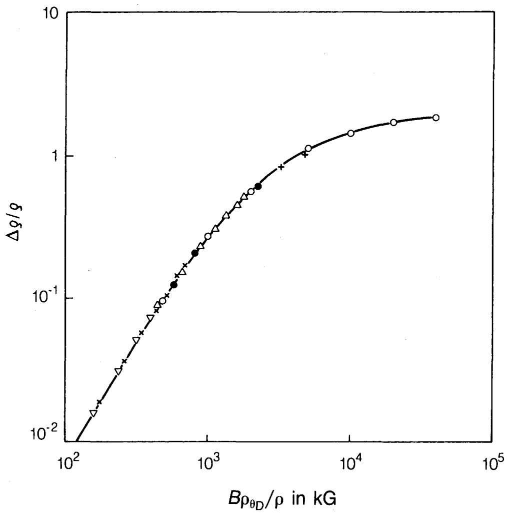

FIG. 16. Magnetoresistance in polycrystalline indium in a transverse magnetic field showing agreement with Kohler's rule. Some points affected by boundary scattering have been suppressed. ▽: In (specimen 1), 14 K, ρ/ρ_{273} = 0.024; X: In (specimen 6), 20 K, ρ/ρ_{273} = 0.023; △: In (specimen 6), 14 K, ρ/ρ_{273} = 0.0086; +: In (specimen 6), 4.2 K, ρ/ρ_{273} = 0.0012; ○: In (specimen 2), 4.2 K, ρ/ρ_{273} = 0.00007; ●: In (specimen 2), 2 K, ρ/ρ_{273} = 0.00003 (from Olsen, 1962).

6.3 High Magnetic Fields ($\omega_c\tau \gg 1$)

In single-crystal samples, the high-field L-*MR* becomes independent of *B* ("saturates") for all FS topologies and for both compensated ($N_e = N_h$) and uncompensated ($N_e \neq N_h$) metals. The T-*MR*, in contrast, can either become independent of *B* or vary as B^2, depending upon the topology of the FS and the orientation of **B** with respect to the crystallographic axes; the permitted alternatives were listed in Table 1.

For an uncompensated metal with no open orbits (see Sec. 4.3), the T-*MR* "saturates" for all directions of **B** when $\omega_c\tau \gg 1$. If the FS contains open orbits, then the T-*MR* saturates when there are no open orbits in the plane perpendicular to **B**, increases as B^2 when there is one band of open orbits in the plane perpendicular to **B**, and saturates again when there are two perpendicular open orbits in the plane perpendicular to **B**. Saturation and B^2 variations for different field directions are illustrated in Fig. 19 for the uncompensated intermetallic compound $AuGa_2$, which can be produced in a highly ordered state that permits the high-field condition to be reached.

For a compensated metal with no open orbits or one band of open orbits, the T-*MR* is proportional to B^2 independent of field orientation, except that the T-*MR* saturates when the open orbit direction in **k** space is exactly the direction of current flow in the sample. The B^2 behavior for compensated metals is illustrated in Fig. 20 for Pb.

Finally, in polycrystalline samples that contain open orbits, an approximately linear T-*MR* may result from averaging of B^0 and B^2 behaviors on different orbits.

6.4 Additional Examples

6.4.1 Deviations from the Ideal The analyses just given are rigorous for a completely homogeneous sample in which the current density is uniform across the sample area and the magnetic field is uniform along the sample length. If the current density is perturbed, for example by defects in the sample or on its surface, or the magnetic field is not uniform, the T-*MR* can become linear in *B*.

6.4.2 Quantum Oscillations At low temperatures and high magnetic fields, quantum

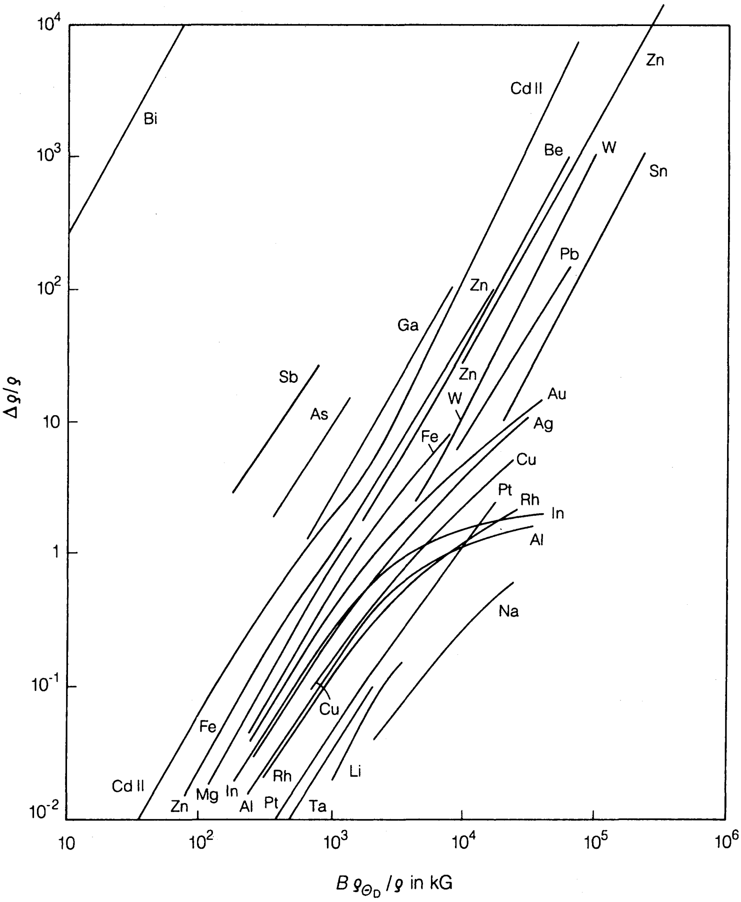

FIG. 17. The transverse magnetoresistance of polycrystalline metals in the reduced Kohler diagram. $\Delta\rho/\rho$ is plotted against $B\rho_{\theta D}/\rho$, where ρ is the resistivity in zero field, and $\rho_{\theta D}$ that at $T = \theta_D$ in zero field (from Olsen, 1962).

oscillations (called Shubnikov–de Haas oscillations) are seen in single crystals of high-purity metals, as illustrated in Fig. 21. The periods of these oscillations are determined by the cross-sectional areas of the FS perpendicular to the direction of **B**.

6.4.3 Size Effects If a transverse **B** is directed in the plane of a thin, high-purity foil, Sondheimer oscillations (Ziman, 1972) appear in the T-*MR*. The period of these oscillations is determined by the average Fermi momentum of the electrons. If **B** points along the axis of a thin wire, the resulting L-*MR* first increases slightly with increasing B and then decreases, the decrease beginning when the electron's cyclotron orbit becomes comparable to the wire radius; thereafter, increasing the field reduces the number of electrons that reach the wire surface, and thus reduces the contribution of surface scattering to ρ.

6.4.4 Ferrogmagnets The *MR* of ferromagnets can be complex and anomalous. For example, Fig. 22 shows that the initial values of the T-*MR* for Ni are negative, and those of the L-*MR* are positive. At fields high enough to align all of the magnetic domains parallel to the field, both the T-*MR* and L-*MR* decrease with increasing field, probably as a result primarily of magnetostrictive effects (see e.g., Chikazumi, 1964).

7. HALL EFFECT, R_H

7.1 Free Electrons

The free-electron model of a metal provides the simplest picture of the physical phenomenon underlying the Hall effect. The current density j_x in a foil of width W is given by $n_e e\langle v_x\rangle$. When a magnetic field B_z is applied in the negative z direction, it produces

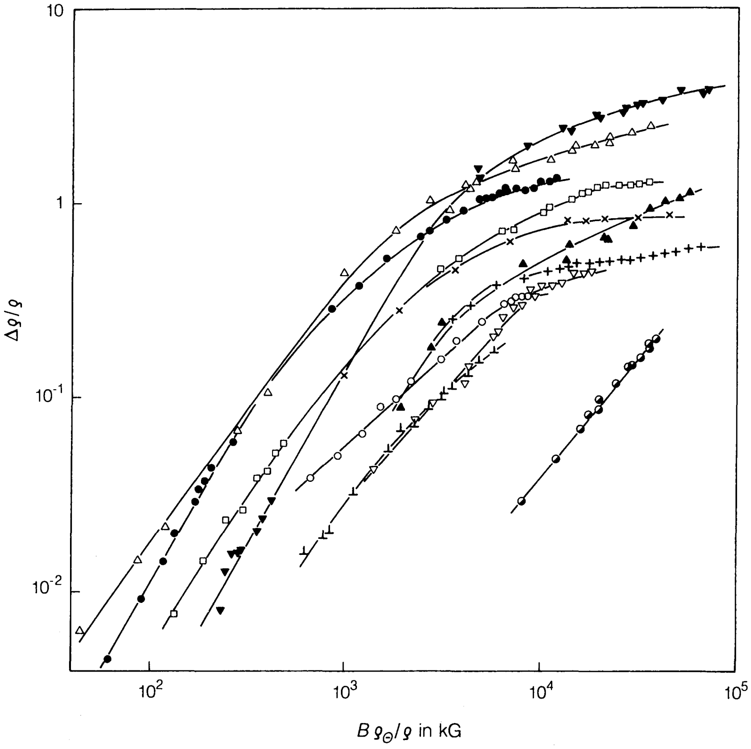

FIG. 18. The magnetoresistance of polycrystalline metals in longitudinal fields. ρ is the zero-field resistivity at the temperature of observation, and ρ_{θ_D} that at $T = \theta_D$. The measurements were made in fields up to 200 000 gauss. ◑, Li, $\theta_D = 430$ K; +, Al, $\theta_D = 410$ K; ▼ Sn, $\theta_D = 160$ K, ○ Ni, $\theta_D = 410$ K, □ Ag, $\theta_D = 220$ K, × In, $\theta_D = 100$ K; ▲ Pb, $\theta_D = 90$ K; ▽ Pt, $\theta_D = 240$ K; ● Au, $\theta_D = 320$ K; △ Zn, $\theta_D = 240$ K; ⊥ Fe, $\theta_D = 355$ K (from Olsen, 1962).

on the electron stream a force $e\langle v_x\rangle B_z$, which is directed in the y direction and causes the electrons to try to flow also in this direction. Since the sample has finite width, the electrons can only pile up on the edge of the foil, thereby generating a charge separation in the foil that leads to an electric field E_y. The electronic system comes into dynamic equilibrium when this field produces a force just large enough to balance the magnetic force,

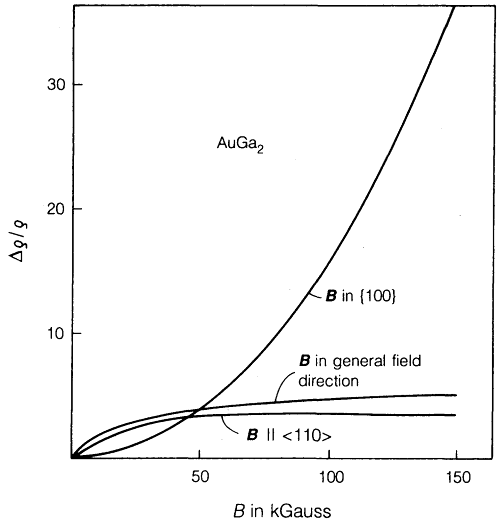

FIG. 19. $\Delta\rho/\rho$ of AuGa$_2$ versus B for various directions of **B** (from Longo et al., 1969).

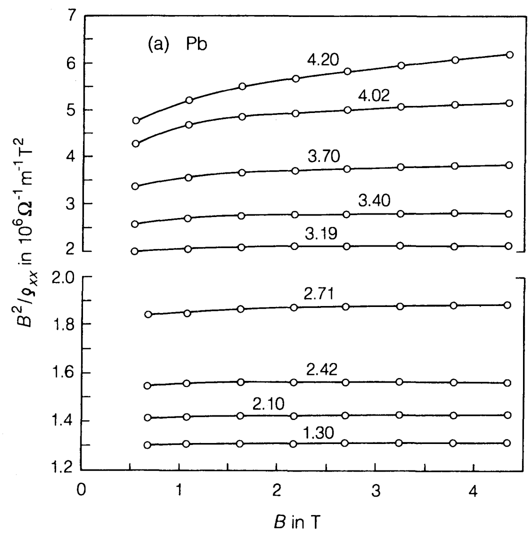

FIG. 20. The quantity B^2/ρ_{xx} plotted as a function of magnetic field B for Pb (after Fletcher, 1977).

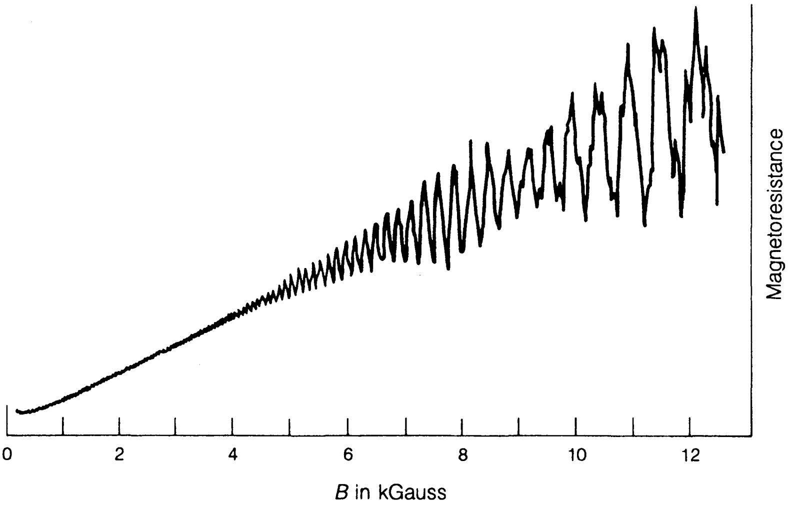

FIG. 21. Shubnikov–de Haas oscillations in the magnetoresistance of gallium versus field at 1.3 K (from Ashcroft and Mermin, 1976).

i.e.,

$$eE_y = e\langle v_x \rangle B_z. \tag{57}$$

Substituting $j_x = n_e e \langle v_x \rangle$ into the right-hand side of Eq. 57 and defining $R_H = E_y/J_x B_z$ yields

$$R_H = 1/n_e e, \tag{58}$$

as already derived under similar assumptions with the Boltzmann transport equation in Sec. 2.3. Equation (58) predicts R_H to be directly proportional to the electron charge (i.e., strictly negative) and inversely proportional to the electron density.

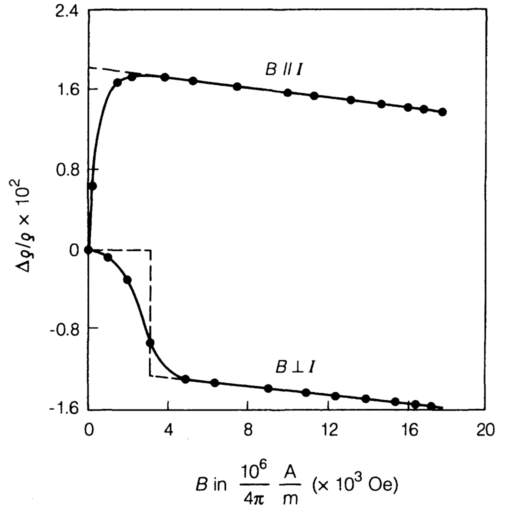

FIG. 22. Variation of Ni resistivity as a function of a magnetic field (from Chikazumi, 1964, using the data of Englert, 1932).

7.2 Beyond Free Electrons

Table 7 lists approximate normalized values of R_H at low temperatures and high magnetic fields for several uncompensated metals (i.e., $N_e \neq N_h$). Equation (58) gives the correct order of magnitude of R_H for all of these metals, but rarely gives precisely the correct value, and sometimes gives even the wrong sign. These deviations from Eq. (58)

Table 7. Hall coefficients of selected elements in moderately high fields (from Ashcroft and Mermin, 1976).

Metal	Valence	$-(R_H n_e e)^{-1}$
Li	1	0.8
Na	1	1.2
K	1	1.1
Rb	1	1.0
Cs	1	0.9
Cu	1	1.5
Ag	1	1.3
Au	1	1.5
Be	2	−0.2
Mg	2	−0.4
In	3	−0.3
Al	3	−0.3

arise because different sets of electrons in a metal have different effective masses and are scattered in very different ways, neither of which was taken into account in the free-electron model.

The Boltzmann-equation analysis given in Sec. 2 can be generalized to allow for different effective masses on different parts of the FS. It is useful to consider this expression in two extreme limits: the low-field limit ($\omega_c\tau \ll 1$) and the high-field limit ($\omega_c\tau \gg 1$).

7.2.1 Low Magnetic Fields The simplest generalization is a two-band model, with relaxation times τ_1 and τ_2, conductivities σ_1 and σ_2, and effective masses m_1^* and m_2^*, where τ and σ are positive quantities but m^* can be positive (electron-like) or negative (holelike). This model yields (Olsen, 1962)

$$R_H = e\frac{\tau_1\sigma_1/m_1^* + \tau_2\sigma_2/m_2^*}{\sigma_1^2 + \sigma_2^2}. \tag{59}$$

For a free-electron FS with single values of τ, σ, and m^*, Eq. (59) reduces to Eq. (58), since $\sigma = n_e e^2\tau/m^*$. In general, however, the sign of R_H depends on a balance between the two bands.

7.2.2 High Magnetic Fields Three different cases occur in this limit, as were listed in Table 1.

1. For an uncompensated metal ($N_e \neq N_h$) with no open orbits, R_H takes the value

$$R_H = 1/(n_e - n_h)e, \tag{60}$$

which is simply a generalization of Eq. (34) or (58) to take into account that the FS's of real metals can contain both electron-like and holelike portions. For one-electron metals such as the alkali and noble metals, where $N_e = 1$ and $N_h = 0$, Eq. (60) reduces to Eq. (58), and the numbers in Table 7 are nearly unity. For Al, in contrast, the second-zone surface contains ≈2 holes/atom, and the third-zone surface contains ≈1 electron/atom. In high magnetic fields, Eq. (60) thus predicts R_H to be $-\frac{1}{3}$ of the free-electron value. In low fields, however, electrons execute only a small fraction of a cyclotron orbit before being scattered, and thus, most of them never collide with a BZ boundary. The BZ boundaries can then be neglected, and Eq. (60) would reduce to Eq. (58) if the details of scattering were unimportant. The high-field value of R_H in Al should be well defined, but the low-field value changes as the dominant scatterer changes. This behavior is illustrated in Fig. 23.

For uncompensated metals, and compensated metals with open orbits, the situation is more complex.

2. In most cases, R_H still becomes constant at high fields, but the value of this constant depends upon details of both scattering and the FS.
3. When open orbits are present in more than one direction, R_H decreases like $1/B$ as B increases.

7.3 Additional Examples Including Magnetic Metals

The temperature dependences of R_H for metals and alloys can be quite complex. Both ρ and R_H change upon melting. For simple metals, the change in R_H (and ρ) is usually not large, as illustrated in Fig. 24 for Li.

As illustrated in Fig. 25 for a $Mg_{0.7}Zn_{0.3}$ alloy, R_H can exhibit different behaviors in crystalline, amorphous, and liquid forms of the same metal or alloy.

Although quantum oscillations of all kinds occur in R_H under the same conditions that they occur in *MR*, they are usually very small (an exception is quantum size effects in the semimetal Bi).

An additional phenomenon, called the anomalous Hall effect, has made Hall-effect measurements important for studies of metals containing large, localized magnetic moments (e.g., ferromagnets), because of an additional contribution, called the anomalous Hall effect. Figure 26 shows schematically what is expected for the Hall resistivity $\rho_H = R_H B$ as a function of B in such a metal.

8. THERMOPOWER *S*

8.1 The Temperature Dependence of *S* in Pure Metals

Figure 27 shows the thermopower S for several pure metals (for more complete data see Foiles, 1985). For each metal, the data can be divided into two components (see Sec. 1.2).

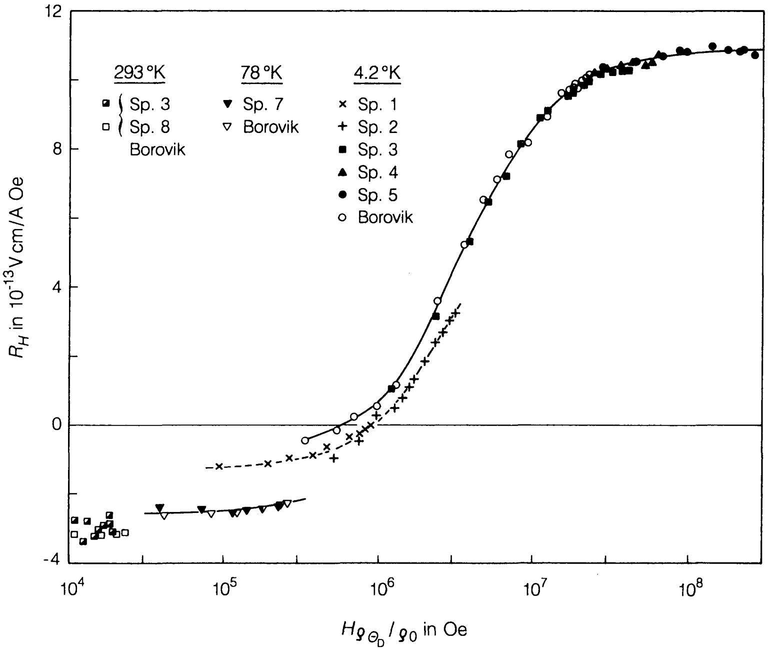

FIG. 23. Hall coefficient of Al as a function of the reduced magnetic field measured through the high-field/low-field transition. Sp1 and Sp2 refer, respectively, to Al + 0.4 at. % Zn and Al + 0.1 at. % Zn. The other specimen numbers refer to pure Al specimens of various thicknesses (from Hurd, 1972, using the data of Forsvoll and Holwech, 1964).

The first, S_d, is approximately linear in T, as indicated by the dashed lines. The second, S_g, passes through a maximum value well below room temperature.

8.2 Rough Derivations of S_d and S_g

The presence of a temperature gradient $\nabla_r T$ causes both electrons and phonons to flow from the hot end of the metal to the cold end. In each case, we define an effective "force per unit volume" associated with this flow that acts on electrons:

$$F_{\text{flow}} = \frac{dU}{dx} = \frac{\partial U}{\partial T}\frac{dT}{dx} = C\frac{dT}{dx} \tag{61}$$

where U and C are the energy density and heat capacity per unit volume, respectively.

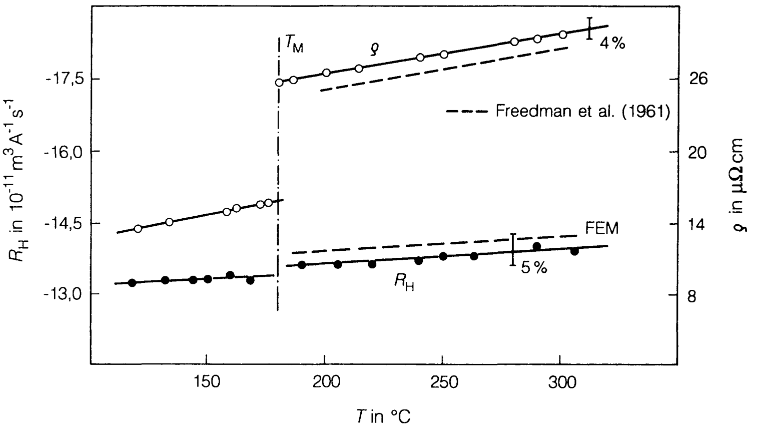

FIG. 24. Hall coefficient and electrical resistivity of solid and liquid Li. T_M = melting point, and FEM = free-electron value (from Kunzi and Guntherodt, 1980).

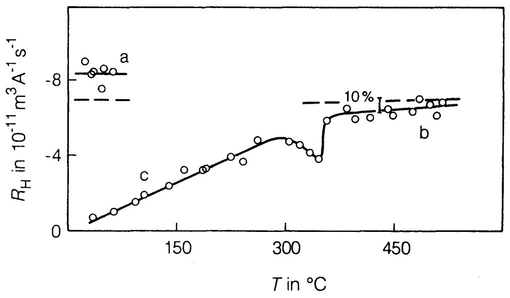

FIG. 25. Hall coefficient of (a) an amorphous, (b) a liquid, and (c) a crystalline $Mg_{70}Zn_{30}$ alloy; dashed line indicates the free-electron values for the Hall coefficient (from Kunzi and Guntherodt, 1980).

In Sec. 1, we noted that S is measured under conditions of no net current flow. This means that the electrons being "pushed" by the force of Eq. (61) must pile up at one end of the sample to produce an electric field E that opposes further electron flow. The force per unit volume due to this field acting on the electron system must be

$$F_E = n_e eE. \tag{62}$$

For electron flow alone, writing $C = C_e$ and equating the forces of Eqs. (61) and (62) gives

$$S_d = \frac{E}{dT/dx} \sim \frac{C_e}{n_e e} = \pi^2 \frac{k_B}{e} \frac{k_B T}{\epsilon_F}, \tag{63}$$

where ϵ_F is the Fermi energy. Equation (63) predicts that S_d should increase linearly with T. For a typical $\epsilon_F \approx 1$ eV, the coefficient in Eq. (63) is $\sim 10^{-8}$ V/K^2, leading to room-temperature values of $S_d \sim 1$ μV/K^2.

For phonon flow, we initially assume that the phonons collide only with electrons, thereby driving them along with their own motion. As above, we then find

$$S_g = E/(dT/dx) \sim C_g/n_e e. \tag{64}$$

Aside from a factor of $\frac{1}{3}$, this is the correct term for free electrons interacting with Debye phonons via normal scattering processes alone.

Since phonons also scatter off other phonons, impurities, etc., Eq. (64) must be multiplied by a factor α, the ratio of the probability for phonon-electron scattering (the only scattering that drives the electrons along with the phonons) to the probability for all scattering (including electron-phonon scattering). Writing the probability for phonon-electron scattering as inversely proportional to one relaxation time τ_{pe}, and the probability for all other types of scattering as inversely proportional to another, τ_{po}, gives

$$S_g \sim (C_g/n_e e)\alpha \sim (C_g/n_e e)\tau_{po}/(\tau_{po} + \tau_{pe}). \tag{65}$$

At low temperatures compared to the Debye temperature θ_D (i.e., $T \ll \theta_D$), phonon-electron scattering dominates, τ_{pe} is short compared to τ_{po}, and the ratio of the τ's cancels out. This leaves $S_g = C_g/n_e e$, so that $S_g = (12\pi^4/5)(k_B/e)(T/\theta_D)^3$. For $T > \theta_D$, C_g becomes constant at the classical value $3nk_B$, where n is the num-

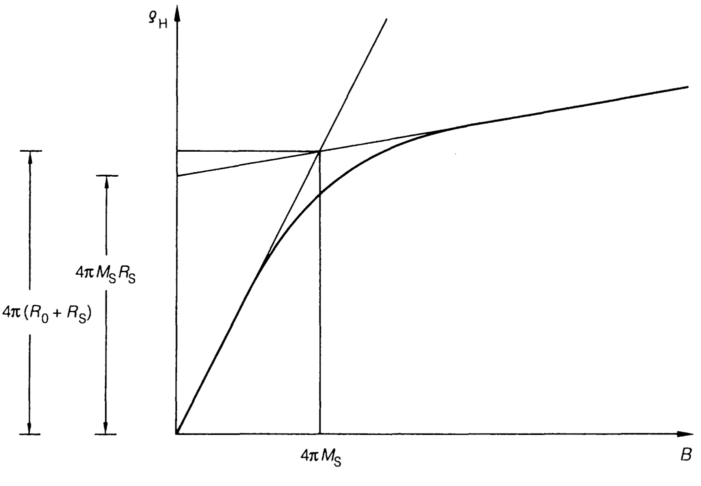

FIG. 26. Schematic behavior of the Hall resistivity ρ_H as a function of magnetic induction B in a metal showing appreciable magnetization (from Hurd, 1972). M_S is the spontaneous magnetization; R_0 and R_S are constants called, respectively, the ordinary and spontaneous Hall coefficients.

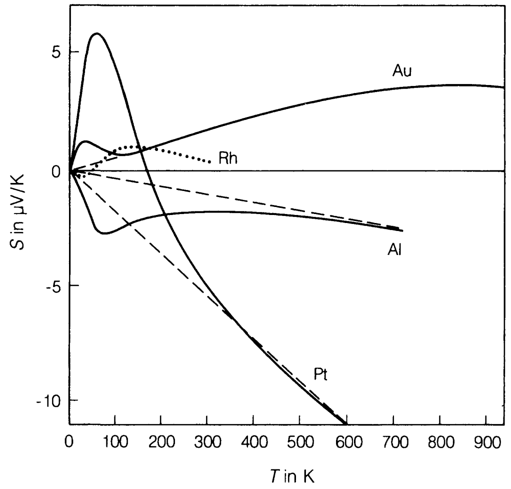

FIG. 27. The thermopowers S of the metals Au, Al, Pt (solid curves), and Rh (dotted curve) as a function of temperature. For Pt, Al, and Au, the differences between the solid curves and the dashed lines indicate the magnitude of the phonon-drag components S_g (from Bass, 1982).

ber of atoms per unit volume in the metal. τ_{pe} becomes temperature independent, whereas τ_{po} is dominated by phonon-phonon scattering which varies as T^{-1}; S_g thus decays as T^{-1}. Inserting appropriate values for τ_{pe} and τ_{po} reveals that as T increases from $T = 0$ K, the S_g of Eq. (65) first increases as T^3, then passes through a maximum value of order $\sim\mu V/K^2$ at $T \approx \theta_D/5$, and finally decreases as T^{-1}.

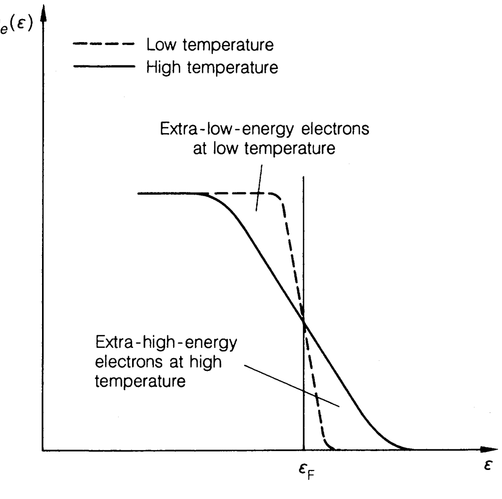
$n_e(\varepsilon)$
Low temperature
High temperature
Extra-low-energy electrons at low temperature
Extra-high-energy electrons at high temperature
ε_F
ε

FIG. 28. The variation with energy ϵ of the number of conduction electrons n_e in a metal in the vicinity of the Fermi energy ϵ_F for two different temperatures. In the interests of simplicity, a small variation of ϵ_F with temperature has been neglected in drawing this figure (from Bass, 1982).

Combining S_d from Eq. (63) with S_g from Eq. (65) yields an S containing one term linear in T with room-temperature values of order μV/K, and another which peaks at $\sim\theta_D/5$ with maximum values also of order μV/K. This combination correctly describes both the forms and magnitudes of the data of Fig. 27. The model predicts, however, that both terms will be strictly negative, whereas Fig. 27 shows that each can be either negative or positive.

This simple model has neglected

1. details of the scattering of electrons on real FSs by realistic phonons and by other scatterers;
2. the presence of "Umklapp" scattering events, which can cause the scattered electron to end up moving oppositely to the direction of the incoming phonon; and
3. contributions to S_d from higher-order scattering events and from many-body effects.

The first two items provide mechanisms for changes in the sign of S_g, and all three provide mechanisms for changes in the sign and magnitude of S_d.

8.3 More Detailed Analysis of S_d

Direct application of the Boltzmann transport equation described in Sec. 2 yields for S_d (e.g., Barnard, 1972)

$$S_d = \frac{\pi^2 k_B^2 T}{3e\epsilon_F}\left[\frac{d\ln\sigma(\epsilon)}{d\ln\epsilon}\right]_{\epsilon_F} = \frac{\pi^2 k_B^2 T}{3e\epsilon_F}\left[\frac{d\ln n_e}{d\ln\epsilon} + \frac{d\ln\tau}{d\ln\epsilon}\right]_{\epsilon_F} \tag{66}$$

where ϵ is the electron's energy, $\sigma(\epsilon)$ is the energy-dependent conductivity, and the derivatives are to be evaluated at ϵ_F. The second expression involves the single-relaxation-time approximation. Since for a spherical FS $n_e \propto \epsilon^{3/2}$, the first term in Eq. (66) is simply the generalization of Eq. (63). The second term in Eq. (66) contains the effects of scattering alluded to at the end of Sec. 8.2. The simplest way to understand the physical sources of the two terms in Eq. (66) is to examine the Fermi distributions at the

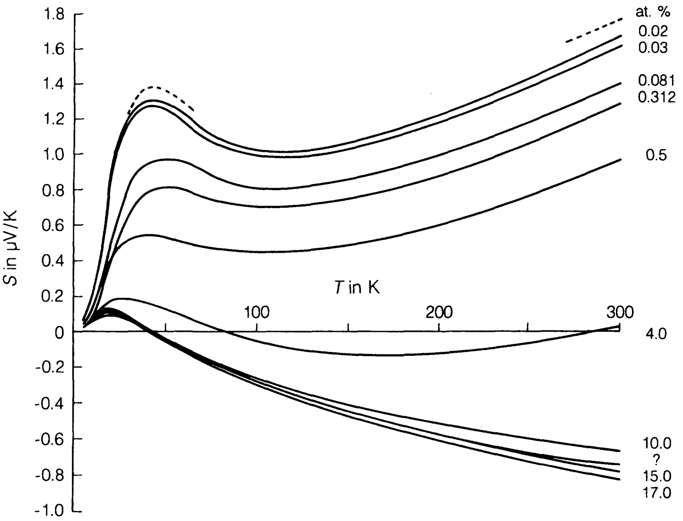

FIG. 29. Measured thermopowers as functions of temperature for the Ag-Hg system (from Craig and Crisp, 1978). The concentrations of Hg in at. % are shown adjacent to each curve.

two ends of the metallic sample of interest. If T_H and T_L are the temperatures at the high- and low-temperature ends of a bar, the resultant Fermi distributions are as shown in exaggerated fashion in Fig. 28. This diagram shows that there are more "high-energy" electrons (i.e., energy greater than ϵ_F) at T_H than at T_L, and more "low-energy" electrons (energy less than ϵ_F) at T_L than at T_H. Thus, if both n_e and τ are larger at higher electron energies, the energy derivative of each will cause electrons to flow more easily from the hot end of the metal to the cold end, thereby building up an excess of electrons at the cold end. These electrons will produce an electric field, which will grow until it just counteracts the flow due to the two "forces" just described. This is exactly what happens on a spherical FS with simple scattering and yields the normal negative S_d. If, however, the energy dependences of n_e or τ are opposite to those assumed in the foregoing, then electrons will tend to pile up at the hot end of the metal, yielding a positive contribution to S_d. This can occur, for example, for n_e, in the case where the FS contacts the BZ boundary.

Even in the free-electron model, the magnitude of the linear coefficient in S_d changes with temperature. The simplest models yield $\tau \propto \epsilon^{3/2}$ for electron-phonon scattering (which is dominant at high temperatures), but they yield $\tau \propto \epsilon^{-1/2}$ for electron-impurity scatter-

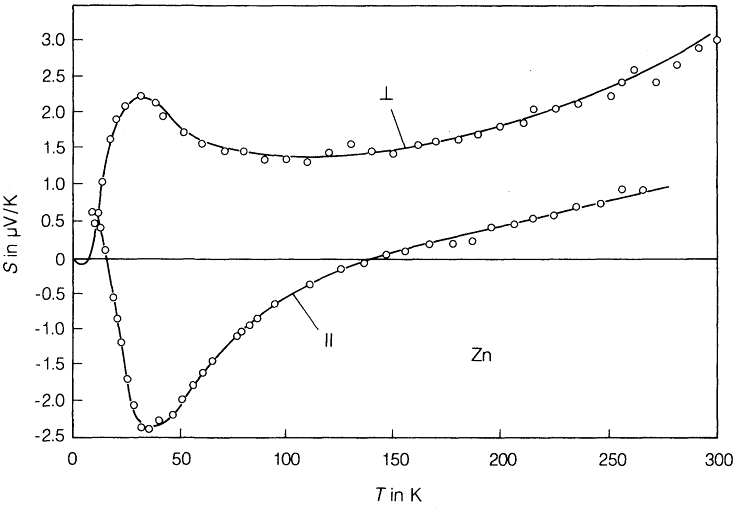

FIG. 30. The thermopower of Zn parallel (∥) and perpendicular (⊥) to the hexagonal axis (from Rowe and Schroeder, 1970).

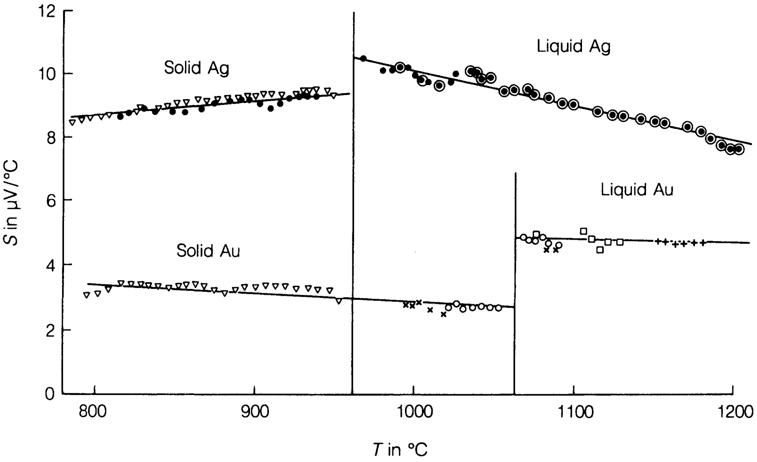

FIG. 31. The changes in the thermopowers of Au and Ag upon melting (from Howe and Enderby, 1967).

ing (which is dominant at low temperatures). Combining each of these terms with $n_e \propto \epsilon^{3/2}$, we find that the coefficient of S_d decreases by $\frac{1}{3}$ from high to low temperatures.

More generally, $\sigma(\epsilon)$ is a complex integral over the FS that depends upon how $\tau_{\mathbf{k}}$ varies from place to place, and it is not possible to separate out terms related simply to n_e and τ. A complete calculation of S_d must also include contributions from higher-order scattering events (i.e., events involving virtual scattering contributions) and corrections due to many-body effects.

Adding impurities to a metal generally reduces S_g and changes the slope of S_d, as illustrated in Fig. 29 (see also Foiles, 1985). The change in slope of S_d can be understood in terms of the Nordheim-Gorter rule:

$$\begin{aligned} S_{\mathrm{TOT}} &= (\rho_p S_p + \rho_0 S_0)/\rho_{\mathrm{TOT}} \\ &= (\rho_p/\rho_{\mathrm{TOT}})(S_p - S_0) + S_0, \end{aligned} \tag{67}$$

which is derived from Eq. (66) by assuming that the impurities represent an additional scatterer for which ρ can be treated by Matthiessen's rule (Eq. 40). In this equation, S_{TOT} is the total thermopower, S_p is the temperature-dependent thermopower of the ideally pure metal, and S_0 is the impurity thermopower. The far right-hand side of Eq. (67) predicts that a plot of S_{TOT} versus $1/\rho_{\mathrm{TOT}}$ should yield a straight line with slope $\rho(S_p - S_0)$ and ordinate-axis intercept S_0. The reductions in S_g shown in Fig. 29 are usually attributed to quenching of phonon drag due to impurity scattering, which drives the phonons back toward thermal equilibrium before they can "drag" the electrons along with them.

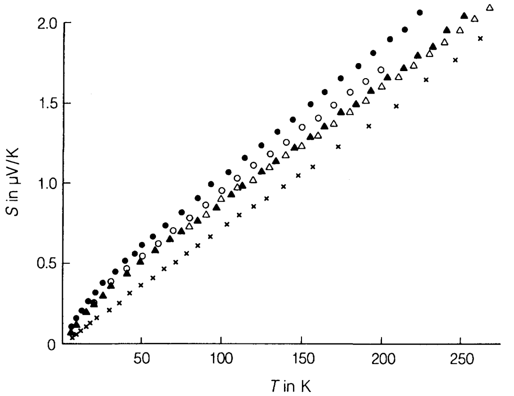

FIG. 32. Measured thermopowers for three Cu-Zr systems: ●, $Cu_{45}Zr_{55}$; ○, $Cu_{27.5}Zr_{72.5}$; △, ▲, $Cu_{60}Zr_{40}$; and × $Cu_{50}Zr_{40}Fe_{10}$ (from Gallagher and Greig, 1982).

8.4 Additional Examples

The thermopower of a single crystal of a noncubic metal can be very different along different crystallographic axes, as illustrated in Fig. 30.

Straining or thinning a metal can also produce significant changes in S.

For a liquid metal, S generally has a magnitude similar to that for the solid; the tem-

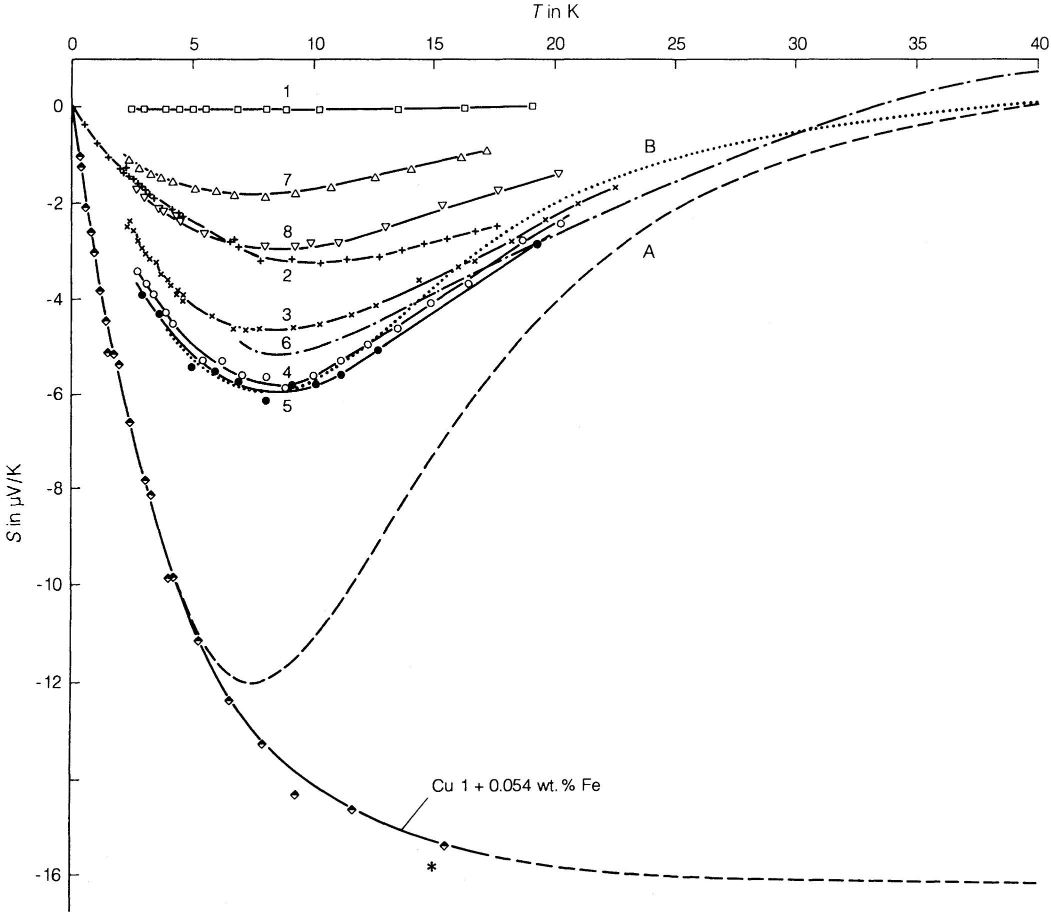

FIG. 33. The low-temperature thermopowers of various samples of copper containing very small concentrations of iron. Sample 1 is most representative of pure Cu, because the Fe is present as oxide (from Gold et al., 1960).

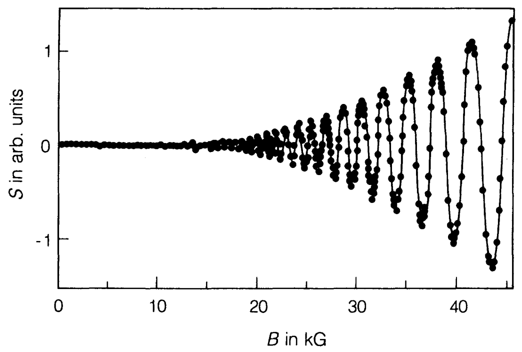

FIG. 34. Magnetic-field dependence of the thermoelectric power of Al for **B** 2° off a ⟨001⟩ direction. One unit in the vertical scale represents about 5 μV per degree (from Kesternich and Papastaikouidis, 1974).

perature-dependent component of S normally varies approximately linearly with T, and the coefficient of this linear term can have the opposite sign from that in the solid. These behaviors are illustrated in Fig. 31.

Like the resistivity (see Sec. 5.2), the thermopower undergoes characteristic changes when a ferromagnet becomes paramagnetic.

The scattering of electrons by defects and disorder in nonmagnetic amorphous metals is so large that phonon drag is completely quenched. Examples of S for several nonmagnetic amorphous metals are given in Fig. 32. The change in slope of S with increasing temperature is attributed to the disappearance with increasing temperature of many-body electron-phonon mass enhancement. (Howson and Gallagher, 1988).

When a magnetic impurity, such as Fe, is dissolved in a nonmagnetic noble metal, such as Au, a "giant" low-temperature anomaly (also called a Kondo anomaly; cf. Sec. 5.3) appears in *S*. Compare, for example, the curve of *S* for 0.054% Fe in Cu in Fig. 33 with that for nominally pure Cu (curve 1).

8.5 Magnetic-Field Effects

Application of a magnetic field to a pure metal can produce significant changes in *S*, in both the quasiclassical and quantum regimes defined in Sec. 4. Unusually large oscillations in *S* occur in single crystals of Al in the quantum regime, because of magnetic breakdown. Fig. 34 shows that at high fields these oscillations become much larger than the nonoscillatory component.

In Sec. 7, it was shown that the Hall coefficient of Al changes sign when a magnetic field is applied at low temperatures, a phenomenon attributed to a change from electronlike to holelike behavior with increasing *B*. The same behavior is displayed in *S* of dilute Al-based alloys.

Application of a magnetic field also leads to more complex thermomagnetic effects. For example, when a field **B** is present in the *z* direction, application of a temperature gradient in the *x* direction gives rise to a voltage in the *y* direction. If the sample is thermally isolated, this phenomenon is called the adiabatic Nernst-Ettingshausen (NE) effect. If the sample is immersed in a constant-temperature bath, the voltage is more nearly representative of the isothermal NE effect. Such measurements have been used to show that thermoelectric coefficients are enhanced by many-body effects.

GLOSSARY

Alloy: A general term used to describe a metal that contains two or more different components. An alloy is characterized by its composition, which defines the relative amounts of each component (usually as either an atomic or weight percentage, or as an atomic ratio). A solid alloy may be *crystalline* (with long-range spatial order), or it may be *glassy* or *amorphous* (no long-range spatial order). The term *liquid alloy* is often used to describe an alloy above its melting point.

Amorphous (or Glassy) Structures: Solid structures that have no long-range spatial periodicity (cf. **Liquids**).

Atomic Order: A description of the extent to which the distribution of atoms over a lattice is periodic or repetitive in space. Ordering may occur over a distance that is large compared to scale of interest (or coherence length of the probe used to investigate the state of order) in which case it is termed *long-range* atomic ordering, or it may occur over a shorter distance in which case it is termed local, or short-range, atomic order. The term ordering is usually reserved to describe structures with unlike nearest neighbors (e.g. ABABAB . . .) whereas clustering or precipitation is used to describe like nearest-neighbor structures (e.g. AAAABBBB . . .). Ordering is a cooperative effect that vanishes above a critical order-disorder temperature.

Brillouin Zone (BZ): Geometrical construction in wave-vector (or reciprocal) space, such that the BZ about a given reciprocal-lattice point is the region of reciprocal space closer to that point than to any other reciprocal-lattice point.

Conductance: An extrinsic property of a body that measures the rate at which charge or heat is conducted through that body. Conductance is related to conductivity by the size and shape of the body. The inverse of conductance is *resistance*.

Conductivity: An intrinsic property of a material that describes the ease with which electrical charge or heat is conducted through it. The inverse of conductivity is *resistivity*.

Electron Orbit: Path followed by an electron on the Fermi surface. Orbits traversed under the influence of a magnetic field may be "open" if the path is open-ended, or "closed" if the path traces out a closed loop.

Electron State: A state of given energy and wave vector that is available for occupation by an electron. An occupied state is a "filled state"; an empty state is a "vacant" or "unfilled" state.

Energy Band: A continuum of electron energy levels in a solid derived from the atomic energy levels of the isolated atoms. A band may overlap with an adjacent band or may be separated from it by an energy gap (also known as a band gap).

Fermi Surface (FS): A surface of constant energy in wave-vector space which, at 0 K, separates occupied electron states from empty ones.

Fermi–Dirac Distribution: A function that describes how the probability of occupation of an electron state varies with energy and temperature.

Free Electrons: Electrons that behave totally independently of any potential fields resulting from the atoms or ions in a solid, except that they are constrained to remain within the solid.

Hall Effect: Applying a magnetic field transverse to the direction of current flow in a conductor produces a "Hall" electric field transverse to both the current and the magnetic field.

Hole: A fictional "quasiparticle" concept, used to describe the motion of a vacant state that results from the collective motion of all of the filled electron states in an energy band.

Lattice: A periodic array of points in space.

Lattice Defects: Disruption of the perfect spatial periodicity of an atomic lattice caused by a missing atom (vacancy), an extra atom between lattice sites (interstitial), a missing line of atoms (dislocation), a plane of atomic misfit (grain boundary), or a free surface. Displacements of atoms away from perfectly periodic positions (due to local strain or thermal excitation) may also be classed as defects.

Magnetic Order: Same as for atomic order, except that ordering of magnetic spin directions is of concern rather than the type of atom. Parallel nearest-neighbor spin alignment is termed ferromagnetic ordering, whereas antiparallel spin alignment is termed antiferromagnetic or ferrimagnetic ordering.

Magnetoresistance: An additional contribution to the electrical resistance that results from the application of a magnetic field.

Mean Free Path: A measure of the average distance traveled by a conduction electron between scattering events.

Nearly Free Electrons: Electrons whose motion suffers a moderately small perturbation from the atomic or ionic potential fields in a solid.

Peltier Effect: The reversible absorption or emission of thermal energy at the junctions of a thermocouple when a current passes.

Relaxation Time: A measure of the average time between scattering events of a conduction electron.

Scattering Process: A process that causes an electron (or phonon) to undergo a transition from some initial state into a final state, such as interaction with another electron or with deviations in the periodicity of the lattice potential. Such deviations may result from lattice defects, from the presence of different atomic species, or from displacement of atoms away from periodic lattice sites (caused by local strain or thermal motion). The process may be *elastic* if there is no difference in energy between the initial and final states, or *inelastic* if a change in energy occurs.

Seebeck Effect: The generation of voltage by a thermocouple.

Thermocouple: Two different conducting materials connected together with one junction at a different temperature from the other.

Thermoelectric Effects: Effects involving the direct conversion of heat into electrical energy, or vice versa (see **Thermopower, Seebeck Effect, Peltier Effect, Thomson Effect**).

Thermopower (Thermoelectric Power): The voltage developed in a thermocouple per unit increase in temperature difference between the thermocouple junctions.

Thomson Effect: The reversible absorption or emission of heat produced by the passage of an electrical current through a conductor along which a temperature gradient has been imposed.

Transport Properties: Properties of a material that depend upon the transport of electrical charge or heat under the influence of external fields or temperature gradients.

Wave Vector: A vector that describes the crystal momentum of a particular mode of vibration of a lattice (i.e., a phonon) or of a particular electron state.

Works Cited

Abrikosov, A. A. (1988), *Fundamentals of the Theory of Metals*, Amsterdam: North-Holland.

Abrikosov, A. A. (1972), "Introduction to the Theory of Normal Metals," in: F. Seitz, D. Turnbull, H. Ehrenreich (Eds.), *Solid State Physics*, Supplement 12, New York: Academic Press.

Ashcroft, N., Mermin, N. D. (1976), *Solid State Physics*, New York: Holt, Rinehart and Winston.

Barnard, R. D. (1972), *Thermoelectricity in Metals and Alloys*, London: Taylor and Francis.

Bass, J. (1972), *Adv. Phys.* **21**, 431–604.

Bass, J. (1982), "Thermoelectricity," in: S. P.

Parker (Ed.), *McGraw-Hill Encyclopedia of Science and Technology*, Vol. 13, p. 295.

Bass, J. (1983), in: J. Bass, K. H. Fischer (Eds.), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, New Series, Group 3, Vol. 15a, *Electronic Transport Phenomena*, Berlin: Springer.

Bass, J. (1985), in: J. Bass, J. S. Dugdale, C. L. Foiles, A. Myers (Eds.), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, New Series, Group 3, Vol. 15b, *Electronic Transport Phenomena*, Berlin: Springer.

Benedek, R., Baratoff, A. (1971), *J. Phys. Chem. Solids* **32**, 1015–1024.

Bloch, F. (1930), *Z. Phys.* **59**, 208–214.

Borchi, G., de Gennaro, S. (1970), *Phys. Lett.* **32A**, 301–302.

Brown, R. A. (1977), *J. Phys. F. Metal Phys.* **7**, 1477–1488.

Brown, R. A. (1982), *Can. J. Phys.* **60**, 766–778.

Chikazumi, S. (1964), *Physics of Magnetism*, New York: John Wiley & Sons.

Coles, B. R. Taylor, J. C. (1962), *Proc. R. Soc. A* **267**, 139–145.

Craig, R., Crisp, S. (1978), in F. J. Blatt, P. A. Schroeder (Eds.), *Thermoelectricity in Metallic Conductors*, New York: Plenum Press. p. 51.

Dugdale, J. S. (1977), *The Electrical Properties of Metals and Alloys*, London: Arnold.

Englert, E. (1932), *Ann. Phys.* **14**, 589–613.

Fert, A., Campbell, I. A. (1976), *J. Phys. F. Met. Phys.* **6**, 849–871.

Fletcher, R. (1977), *Solid State Commun.* **21**, 1139.

Foiles, C. L. (1985), in: J. Bass, J. S. Dugdale, C. L. Foiles, A. Myers (Eds.), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, New Series, Group 3, Vol. 15b, *Electronic Transport Phenomena*, Berlin: Springer.

Forsvoll, K., Holwech, I. (1964), *Phil. Mag.* **9**, 435–450.

Gallagher, B. L., Greig, D. (1982), *J. Phys. F.* **12**, 1721–1741.

Gold, A. V., MacDonald, D. K. C., Pearson, W. B., Templeton, I. M. (1960), *Phil. Mag.* **5**, 765–786.

Greene, R. F. (1964), *Surface Sci.* **2**, 101–113.

Grüneisen, E. (1933), *Ann. Phys.* **16**, 530–540.

Hashin, Z., Shtrikman, S. (1962), *J. Appl. Phys.* **33**, 3125–3131.

Howe, R. A., Enderby, J. E. (1967), *Philos. Mag.* **16**, 467–476.

Howson, M. A., Gallagher, B. L. (1988), *Phys. Rep.* **170**, 265–324.

Hurd, C. M. (1972), *The Hall Effect in Metals and Alloys*, New York: Plenum Press.

Johansson, C. H., Linde, J. O. (1936), *Ann. Phys.* **25**, 1–48.

Kesternich, W., Papastaikouidis, C. (1974), *Phys. Stat. Sol.* **B64**, K41–K43.

Kunzi, H. U., Guntherodt, H.-J. (1980). in C. J. Chien, C. R. Westgate (Eds.), *The Hall Effect and Its Applications*, New York: Plenum Press.

Linde, J. O. (1932), *Ann. Phys. (Leipzig)* **15**, 219–248.

Longo, J. T., Schroeder, P. A., Sellmyer, D. J. (1969), *Phys. Rev.* **182**, 658–670.

Meaden, G. T. (1965), *Electrical Resistance of Metals*, London: Heywood Books.

Mizutani, U. (1983), *Prog. Mater. Sci.* **28**, 98–228.

Naugle, D. G. (1984), *J. Phys. Chem. Solids* **45**, 367–388.

Olsen, J. L. (1962), *Electronic Transport in Metals*, New York: John Wiley & Sons.

Osamura, K. O., Hiraoka, Y., Murakami, V. (1973), *Philos. Mag.* **28**, 809–825.

Porter, D. A., Easterling, K. E. (1981), *Phase Transformations in Metals and Alloys*, New York: Van Nostrand Reinhold.

Rossiter, P. L. (1987), *The Electrical Resistivity of Metals and Alloys*, Cambridge, U.K.: Cambridge University Press.

Rowe, V. A., Schroeder, P. A. (1970), *J. Phys. Chem. Solids* **31**, 1–8.

Schröder, K. (1983), *CRC Handbook of Electrical Resistivity of Binary Metallic Alloys*, Boca Raton, FL: CRC Press, p. 1.

Simmons, R. O., Balluffi, R. W. (1960), *Phys. Rev.* **117**, 62–68.

Sun, P. H., Ohring, M. (1976), *J. Appl. Phys.* **47**, 478–485.

Warren, B. E. (1969), *X-Ray Diffraction*, Reading, MA: Addison-Wesley.

Wilson, A. H. (1937), *Proc. Cambridge Philos. Soc.* **33**, 371–379.

Ziman, J. M. (1960), *Electrons and Phonons*, Cambridge: Cambridge University Press.

Ziman, J. M. (1972), *Electrons and Phonons*, 2nd ed., Oxford: Clarendon Press.

Further Reading

Abrikosov, A. A. (1988), *Fundamentals of the Theory of Metals*, Amsterdam: North-Holland Physics Publishing.

Ashcroft, N., Mermin, N. D. (1976), *Solid State Physics*, Philadelphia: Holt, Rinehart, and Winston.

Barnard, R. D. (1976), *Thermoelectricity in Metals and Alloys*, London: Taylor and Francis.

Blatt, F. J. (1968), *Physics of Electronic Conduction in Solids*, New York: McGraw-Hill.

Blatt, F. J., Schroeder, P. A. Foiles, C. L., Greig, D. (1976), *Thermoelectric Power of Metals*, New York: Plenum Press.

Chien, C. J., Westgate, C. R. (Eds.) (1980), *The Hall Effect and Its Applications*, New York: Plenum Press.

Hurd, C. M. (1972), *The Hall Effect in Metals and Alloys*, New York: Plenum Press.

Kittel, C. (1986), *Introduction to Solid State Physics*, 6th ed. New York: John Wiley & Sons.

Mott, N. F., Jones, H. (1936), *The Theory of the Properties of Metals and Alloys*, Oxford: Clarendon Press.

Rossiter, P. L. (1987), *The Electrical Resistivity of Metals and Alloys*, Cambridge: Cambridge University Press.

Ziman, J. M. (1965), *Principles of the Theory of Solids*, Cambridge: Cambridge University Press.

Ziman, J. M. (1972), *Electrons and Phonons*, 2nd ed. Cambridge: Cambridge University Press.

METALS AND ALLOYS, STRUCTURE OF

Fereshteh Ebrahimi and Michael J. Kaufman, *Materials Science and Engineering Department, University of Florida, Gainesville, Florida, U.S.A.*

Introduction 199
1. **Crystal Structure** 200
1.1 Elemental Metals 200
1.2 Solid Solutions and Superlattices 201
1.3 Short-Range Ordering and Clustering 202
1.4 Metallic Intermediate Phases 203
2. **Defects in Crystalline Structures** 204
3. **Microstructure** 205
3.1 Liquid Alloys 206
3.2 Solid Alloys 207
3.2.1 Cast Structures 207
3.2.2 Wrought Structures 207
3.2.3 Solid-State Phase Transformations 208
3.2.3.1 Precipitation 208
3.2.3.2 Eutectoid Reaction 209
3.2.3.3 Martensitic Transformation .. 209
3.3 Rapid Solidification Processing 210
3.3.1 Metallic Glasses 210
3.3.2 Nanocrystals 211
3.3.3 Quasicrystals 212
Glossary 212
Works Cited 213
Further Reading 213

INTRODUCTION

Metals are generally distinguished from nonmetals by their high electrical conductivity, which increases with decreasing temperature. Most of the elements in the periodic table are metals, with nonmetals occupying the right upper corner of the table (see Fig. 1). Metals, like other elements, may exist in gas, liquid, or solid state; this article is concerned mainly with the last state.

When metallic atoms are gathered together, the most loosely bound electrons (valence electrons) become separated from their parent atoms and wander almost freely through the entire body of the material. These free electrons are responsible for the electrical, optical, and magnetic properties and partially for the thermal conductivity of metals. The cohesion between metallic atoms is controlled mainly by the electrostatic attraction between the ionic (positively charged) cores and the free electrons, and, therefore, the energy of a group of metallic atoms gathered together is not very sensitive to the positions of ionic cores, and there are no directional atomic bonds. Consequently, metallic atoms tend to gather in close-packed structures, making it much easier to shear than to compress them. Hence, metals tend to be ductile and may be formed into complicated shapes without failure. Normally, in solid metals, the atoms arrange themselves in a periodic array known as the crystal structure. However, under nonequilibrium conditions, such as those encountered in fast cooling from the liquid state, condensation from the vapor state, or bombardment with high-energy particles, the atoms may be arranged in an irregular manner (amorphous structure) that resembles the structure of liquids.

Metals are mixed with other elements (metals and nonmetals) to form alloys, in order to improve properties such as strength and environmental stability. A system composed of two or more elements may consist of a variety of phases—a phase being defined as a structurally homogeneous arrangement of atoms. Most metals mix homogeneously in the gaseous and liquid states, in contrast to the solid state, where complete mixing is achieved at the atomic level only in a few binary (two-component) systems.

In this article, we describe the structure of metals and alloys in terms of the crystal structure—the arrangement of atoms in crystals—and the microstructure—the ar-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

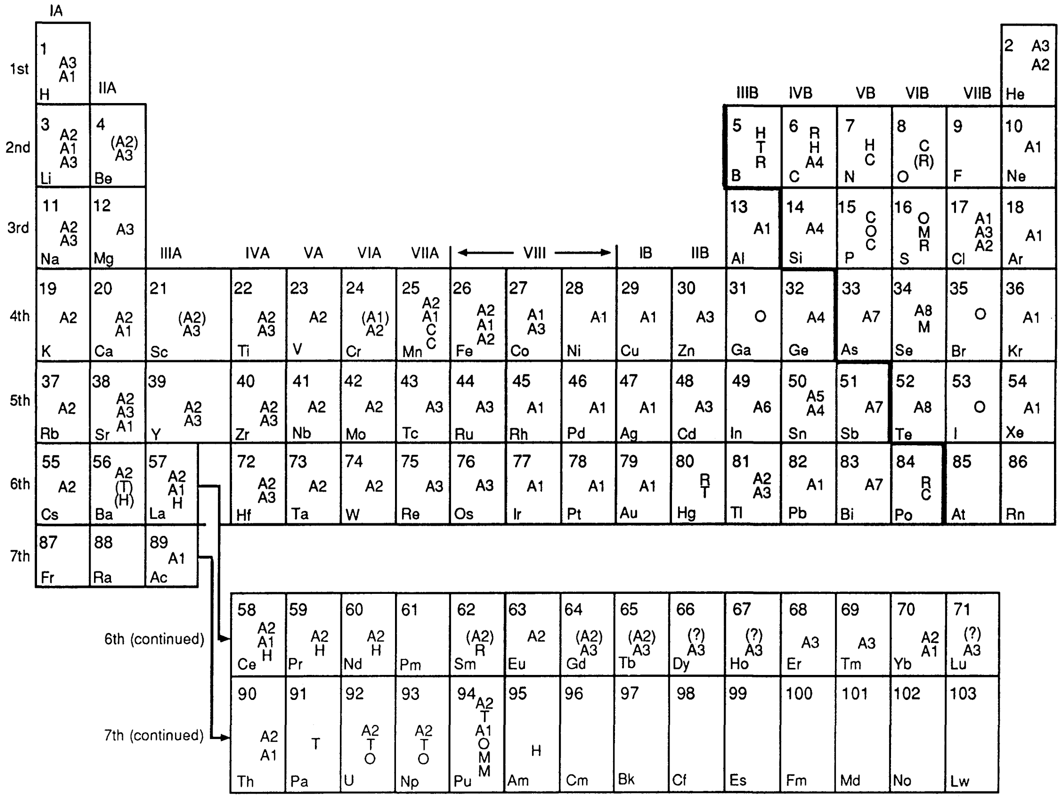

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	← VIII →			IB	IIB	IIIB	IVB	VB	VIB	VIIB	
1st	1 H A3 A1																	2 He A3 A2
2nd	3 Li A2 A1 A3	4 Be (A2) A3											5 B H T R	6 C R H A4	7 N H C	8 O C (R)	9 F	10 Ne A1
3rd	11 Na A2 A3	12 Mg A3											13 Al A1	14 Si A4	15 P C O C	16 S O M R	17 Cl A1 A3 A2	18 Ar A1
4th	19 K A2	20 Ca A2 A1	21 Sc (A2) A3	22 Ti A2 A3	23 V A2	24 Cr (A1) A2	25 Mn A2 A1 C C	26 Fe A2 A1 A2	27 Co A1 A3	28 Ni A1	29 Cu A1	30 Zn A3	31 Ga O	32 Ge A4	33 As A7	34 Se A8 M	35 Br O	36 Kr A1
5th	37 Rb A2	38 Sr A2 A3 A1	39 Y A2 A3	40 Zr A2 A3	41 Nb A2	42 Mo A2	43 Tc A3	44 Ru A3	45 Rh A1	46 Pd A1	47 Ag A1	48 Cd A3	49 In A6	50 Sn A5 A4	51 Sb A7	52 Te A8	53 I O	54 Xe A1
6th	55 Cs A2	56 Ba A2 (T) (H)	57 La A2 A1 H	72 Hf A2 A3	73 Ta A2	74 W A2	75 Re A3	76 Os A3	77 Ir A1	78 Pt A1	79 Au A1	80 Hg R T	81 Tl A2 A3	82 Pb A1	83 Bi A7	84 Po R C	85 At	86 Rn
7th	87 Fr	88 Ra	89 Ac A1															

6th (continued)	58 Ce A2 A1 H	59 Pr A2 H	60 Nd A2 H	61 Pm	62 Sm (A2) R	63 Eu A2	64 Gd (A2) A3	65 Tb (A2) A3	66 Dy (?) A3	67 Ho (?) A3	68 Er A3	69 Tm A3	70 Yb A2 A1	71 Lu (?) A3
7th (continued)	90 Th A2 A1	91 Pa T	92 U A2 T O	93 Np A2 T O	94 Pu A2 T A1 O M M	95 Am H	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lw

A1 = f.c.c.; A2 = b.c.c.; A3 = c.p.h.; A4 = diamond cubic; A5 = b.c.t.; A6 = f.c.t.; A7 = rhombohedral; A8 = trigonal; H = hexagonal (usually ABAC • • • close-packed); R = rhombohedral; O = orthorhombic; C = complex cubic; T = Tetragonal; M = monoclinic; ( ) uncertain.

FIG. 1. Periodic table of the elements (Barrett and Massalski, 1980).

rangement of crystals and phases in polycrystals; we conclude with a brief description of both amorphous and quasicrystalline structures.

1. CRYSTAL STRUCTURE

The crystal structure of metals and alloys can be identified and characterized using various diffraction techniques (x-ray, neutron, and electron). As mentioned above, atoms of elemental metals tend to be packed closely, since there is no significant interaction between their ionic cores. However, additions of alloying elements may modify the electron distribution significantly and add covalent and/or ionic characters to the forces holding the atoms together. This modification of the metallic binding may result in the ordering of the atoms and/or the formation of complex crystal structures as discussed in Secs. 1.2–1.4.

1.1 Elemental Metals

The crystal structure of metals is given in the periodic table presented in Figure 1, which clearly shows that the majority of metals crystallize in relatively close-packed structures, namely face-centered cubic (fcc), hexagonal close-packed (hcp), and the less closely packed body-centered cubic (bcc) structures. Elements of the same group tend to have similar structures; for example, the alkali metals (Group IA), Li, Na, K, Rb, and Cs, are bcc, and Group IB metals, Cu, Ag, and Au, are fcc. In the group of metals known as transition metals (Groups IV, V, VI, VII, and VIII elements), the *s* energy states are filled with electrons before the *d* energy states are filled. At the usual interatomic spacing of these metals, the atomic *d* orbitals do not overlap significantly, and a tight-binding approach based on the density-of-states distribution seems to explain the types of crystal structures along the three transition-metal rows (Cottrell, 1991).

In the case of the hcp crystals, the *c/a* ratio (*c* and *a* are the atomic lattice parameters) depends on the electronic structure. For example, both Group IVA, Ti, Zr, and Hf, and Group IIB, Zn and Cd, have the hcp structure; in the latter group, the outer electron shells are nearly filled, and the structure is looser and, hence, has a larger *c/a* ratio. The large *c/a* ratio of Zn and Cd limits the extent of plastic deformation to slip in the basal planes, and, consequently, through a lack of a large enough number of slip systems, polycrystalline forms of these metals are relatively brittle.

Since the cohesion of atoms in metals is not very sensitive to the positions of the ionic cores, metals often exist in more than one crystal structure. For example, the Group IVA elements are hcp at low temperatures and transform to the more loosely packed bcc structure, which has higher vibrational entropy, at elevated temperatures. This solid-state phase change is known as an allotropic transformation, and it may also be induced by a change in pressure. In the case of iron, the volume decrease due to a change from the ferromagnetic to the paramagnetic state overrides the temperature effect, and iron transforms from bcc to fcc with increasing temperature. On further heating, however, it transforms back to the bcc structure.

1.2 Solid Solutions and Superlattices

If the crystal structure of a metal (solvent) remains unchanged by the addition of another element (solute), then the mixture is called a solid solution. There are two types of solid solutions: substitutional and interstitial. Figure 2 shows the positions of interstitial and substitutional atoms in a lattice.

The interstitial solute atoms, since they are squeezed in between the regular lattice atoms, need to be considerably smaller than the solvent atoms and are limited to a few elements, including H, B, C, N, and O. Since interstitials strain close-packed crystals greatly, their solubility in metals is limited and depends on the crystal structure. In general, the interstitial solubility is much higher in transition metals than in other metals. Although the bcc structure is a less closely packed structure than fcc, the free volume is distributed as smaller spaces in the former structure, and, hence, bcc phases tend to have a lower solubility for interstitial atoms. The higher solubility of carbon in fcc iron compared with bcc iron is the basis for achieving a wide variety of microstructures and properties in steels (see STEEL).

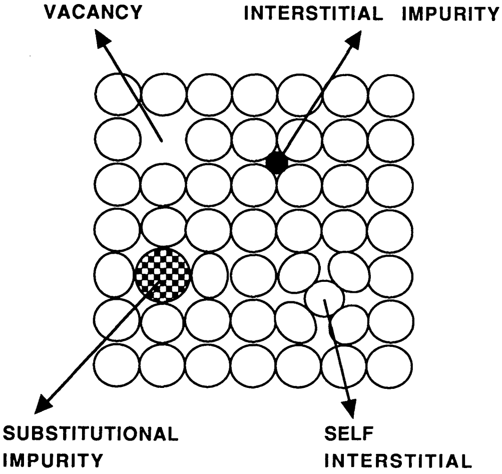

FIG. 2. Schematic showing various types of point defects in disordered solid solutions.

Metals may dissolve extensively in each other by forming substitutional solid solutions. The extent of the solubility is governed by three empirical rules, known as Hume-Rothery rules:

1. the atomic size difference must be less than 15% for extensive solubility;
2. the difference in electronegativity (or electropositivity) of the atoms must be small (otherwise they prefer to form compounds); and
3. a metal of lower valency tends to dissolve less in a metal of higher valency than vice versa.

The solubility range also depends on temperature; for example, Nb-Ta, Nb-Ti, and Cu-Au systems show complete solubility in the solid state at high temperatures. While Nb and Ta remain completely soluble at low temperatures, the bcc structure of Nb is not favored by Ti because of the polymorphism of the latter element, and a new hcp solid solution evolves at lower temperatures. In the case of the Cu-Au system, because of the attraction between copper and gold, as the temperature drops, the atoms tend to order, thereby increasing the number of copper-gold bonds and creating ordered solid solutions (superlattices). The lattice sites chosen by the Cu and Au atoms depend on the composition. For example, as shown in Fig. 3(a), at about 25% gold, Cu atoms occupy the face centers of the fcc lattice, and Au atoms go to the cube corners, creating an ordered structure known as the Cu_3Au-I or $L1_2$ (*Strukturbericht* designation) superlattice. In this structure, the copper atoms have eight nearest neighbors of similar kind and four of the opposite kind, whereas the gold atoms have twelve nearest neighbors of the opposite kind. Because of the ordering, this structure is no longer fcc, but it can be generated by repeating the small four-atom unit marked in Fig. 3(a) on a simple cubic lattice. When the composition is close to 50 at. % gold, the copper and gold atoms occupy alternate planes, as shown in Fig. 3(b). Because of the difference in the ionic radii of copper and gold, and a modification of the cohesion forces, the resulting structure is slightly tetragonal and is known as $L1_0$ or CuAu-I–type superlattice.

Since the majority of metals are fcc, bcc, or hcp, the common superlattices are derived from these structures. Table 1 presents a list of common superlattices and typical examples of each kind.

Table 1. A list of common superlattices that occur in ordered solid solutions and intermetallics.

Structure Designation	Examples
$L1_2$	Cu_3Au, Zr_3Al, Co_3V, Ni_3Fe, Ni_3Al, Zn_3Ti
DO_3	Fe_3Al, Li_3Bi, Cu_3Sb, Al_3Cu
DO_{19}	Mg_3Cd, Fe_3Sn, Co_3Mo, Ti_3Al, Co_3V
B2	CuZn, NiAl, AuCd, NiZn, LiTl
$L1_0$	CuAu, TiAl, CoPt, CuTi, FePt, NiZn

Many ordered solid solutions go through an order-disorder transition with temperature. The ordering is considered to be a first-order reaction whenever it occurs by the nucleation of highly ordered regions that grow into a disordered matrix. When the ordered regions touch each other, the atomic ordering may be in or out of phase; when it is out-of-phase, it creates a boundary, known as an antiphase boundary (APB), across which the crystal structure and crystal orientation remain unchanged, but atoms have the "wrong" nearest neighbors. Figure 4 presents schematics of the ordered B2 lattice structure containing an APB, along with actual boundaries as revealed by transmission electron microscopy (TEM).

1.3 Short-Range Ordering and Clustering

The formation of superlattices as explained above is known as long-range ordering and occurs below some critical temper-

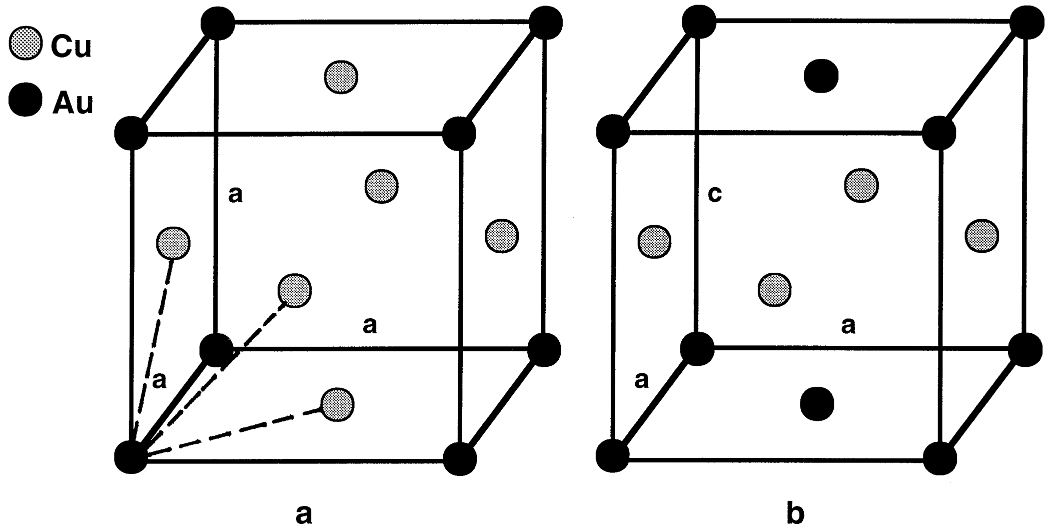

FIG. 3. The structures of (a) Cu_3Au and (b) CuAu superlattices.

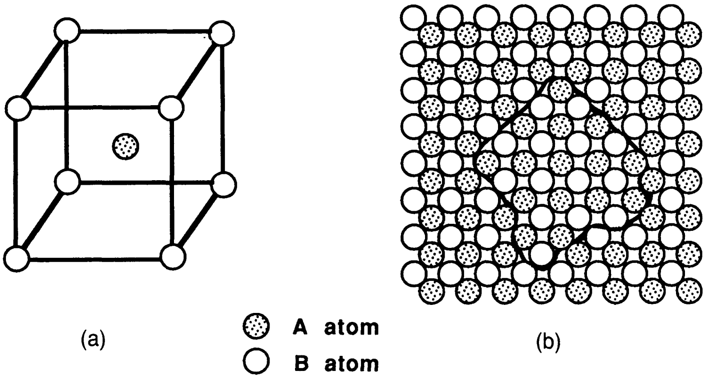

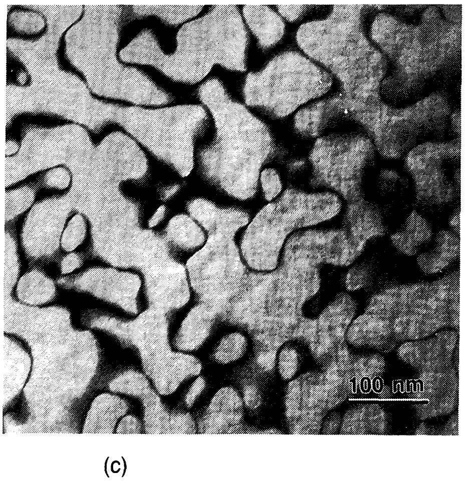

FIG. 4. (a) The B2 structure of AB compounds. (b) Schematic representation of an APB in the B2 structure. (c) TEM micrograph showing APB's (black areas) in a Nb–25Ti–42Al alloy with B2 structure.

ature. The degree of ordering in an alloy is measured as the proportion of atoms occupying the correct lattice site. Short-range ordering, in contrast to long-range ordering, is measured by the extent to which an atom is surrounded by unlike neighbors, with no reference to the lattice site. Short-range ordering has been observed in solid solutions just above the critical (long-range) order-disorder temperature and in alloys quenched from above this temperature. The opposite of short-range ordering is clustering, which occurs in solid solutions that prefer like atoms to unlike atoms as nearest neighbors. Clustering is usually observed in systems with limited solubility and as a precursor to precipitation of a second phase. Both short-range ordering and clustering can cause tremendous strengthening in alloys. For example, the enhanced clustering of copper and phosphorus in iron caused by neutron irradiation is responsible for radiation embrittlement in water reactor pressure vessels (Ebrahimi *et al.*, 1988). On account of the small size of clusters and short-range–ordered regions, field-ion microscopy (FIM) and small-angle neutron and x-ray scattering (SANS and SAXS) techniques are conventionally employed to study these phenomena. More recently, high-resolution transmission electron microscopy techniques have been successfully used to reveal such fine structures.

1.4 Metallic Intermediate Phases

Intermediate phases form at compositions beyond the solid solubility limits in systems that do not show complete mixing, even at high temperatures. These phases can be disordered and similar to solid solutions, they may become ordered upon cooling, or they may have such a strong ordering tendency that they remain ordered up to their melting point or to temperatures at which they transform to other phases. Metallic intermediate phases (intermetallics) are phases in which the atomic bonding is intermediate between metallic and covalent or metallic and ionic. In cases where the ionic or covalent component of the atomic bonding is significant, the intermediate phases behave like chemical compounds; i.e., they have a highly ordered structure with a limited or negligible solubility range. At the outer limits are "ionic compounds," such as NaCl and CaF_2, which have mainly ionic bonding and low conductivities, and "covalent compounds," such as GaAs and InP, which have mainly covalent bonding and are semiconductors.

When there is a large difference between the atomic sizes of the elements, the ordering and the limited solubility may be induced by a "size-factor" effect; examples include interstitial compounds such as carbides, nitrides, borides, and hydrides, and substitutional compounds such as Laves and sigma phases. If the size effect is the dominant factor governing the structure, then the intermediate phases composed of atoms of similar atomic-size ratios should have similar crystal structures and show significant mutual solid solubilities. Examples include complete solid solubility between the carbides of V, Ti, Nb, Ta, and Zr, and the formation of carbonitrides when both carbon and nitrogen are mixed with these elements (Barrett and Massalski, 1980).

Another group of intermediate phases involves the so-called "electron compounds," which occur at specific electron-to-atom ratios (Barrett and Massalski, 1980). Examples of this group include the α, β, γ, and ϵ phases that form in alloys based on Cu, Ag, and Au mixed with Group B metals. These phases have a relatively large solubility range and may become ordered at low temperatures.

There is a multitude of crystal structures that are formed by combining various elements, and it is beyond the scope of this article to cover them all. In general, the interplay of atomic size, electronic structure, and temperature must be considered, and easily categorized compounds are merely those in which one of these factors is dominant. Many intermetallics behave like metals i.e., they are good conductors and reflect light, while the bond directionality results in high melting points and large elastic moduli, which make them possible candidates for high-temperature structural applications. For example, aluminides such as Ni aluminides (Ni_3Al, NiAl), Ti aluminides (Ti_3Al, TiAl, $TiAl_3$), and Nb aluminides (Nb_3Al, Nb_2Al, $NbAl_3$) are being considered for high-temperature structural applications such as turbine blades and vanes in next-generation aircraft engines (Liu and Stiegler, 1984).

2. DEFECTS IN CRYSTALLINE STRUCTURES

Metal crystals, like other crystals, contain defects. Crystal defects are usually characterized as point, line, surface, and volume defects (see POINT DEFECTS AND OTHER IMPERFECTIONS IN CRYSTALS).

The primary point defects are vacancies, self-interstitial atoms, and foreign atoms (see Fig. 2). Point defects are the only thermodynamically stable defects, and may sometimes further reduce the energy of the system by coming together and forming extended defect complexes. Point defects, particularly impurities, affect the properties of metals considerably. For example, the shear yield stress of copper is increased by 45% when the purity is decreased from 99.999 to 99.98% (McLean, 1962). Likewise, the high concentration of vacancies produced by quenching aluminum from high temperatures can result in a fivefold increase of its shear yield strength (Smallman, 1985).

In ordered structures, the substitution on a lattice site of a wrong type of atom creates an "antisite" defect or substitutional disorder. In strongly ionic (or covalent) compounds, point defects should ionically (or electronically) compensate each other. For example, the absence of a positively charged ion from a lattice site should be compensated by the absence of a negatively charged ion (Schottky defect) or the existence of a positive ion in an interstitial site (Frenkel defect).

Dislocations are line defects, which are usually produced in metallic crystals by plastic deformation. Ductile polycrystalline metals and alloys are used extensively in the wrought form and may contain a dislocation density of about 10^8 cm/cm^3 in a completely annealed state and a dislocation density of up to 10^{12} cm/cm^3 in a heavily cold-worked condition. Cast structures tend to have lower dislocation densities than wrought products.

Undeformed single crystals may contain a much lower density of dislocations; in the extreme case of metallic whiskers, the dislocation density can be as low as 10^3 cm/cm^3. Usually, an increase in the dislocation density results in strengthening of metals, except when the dislocation density is initially very low, in which case the contribution of the dislocation nucleation process to deformation—a harder process than dislocation motion—results in an inverse relationship. The most widely used techniques for the observation of dislocations are TEM and etch-pitting, examples of which are shown in Fig. 5.

Surface and volume defects, similar to dislocations, are produced during the solidification and thermomechanical treatment of metals (see Materials Treatment). Examples of surface defects are grain boundaries, interfaces between phases, twin boundaries, antiphase boundaries, and stacking faults. All surface defects, except for highly disordered grain and interfacial boundaries, can be constructed by networks of dislocations. Indeed, twin boundaries, antiphase boundaries, and stacking faults can be generated by the motion of dislocations during plastic deformation. Cold working produces an increase in the number of grain and interfacial boundaries, which may act as both sources and sinks for dislocations. Voids are the most common type of volume defects. They may form as solidification defects, or develop during plastic deformation or by diffusion processes during high-temperature deformation.

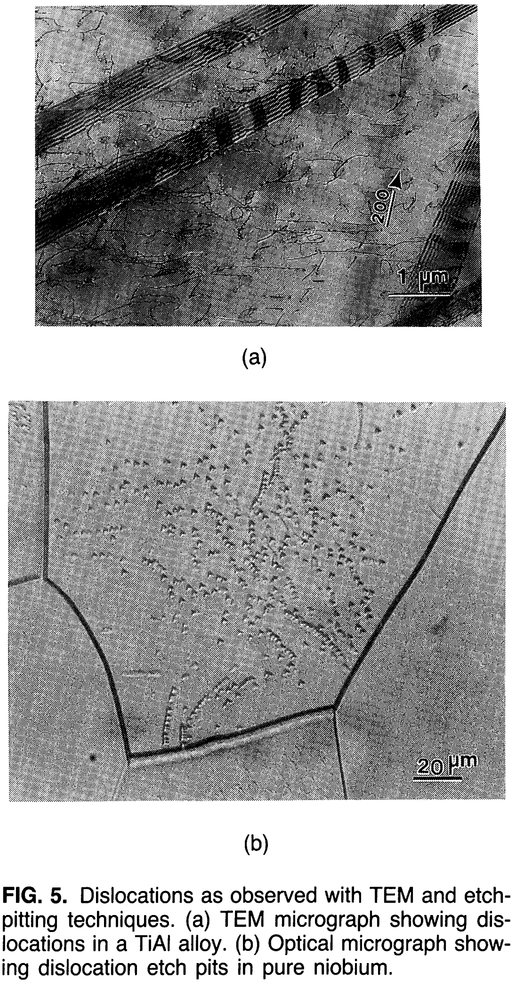

FIG. 5. Dislocations as observed with TEM and etch-pitting techniques. (a) TEM micrograph showing dislocations in a TiAl alloy. (b) Optical micrograph showing dislocation etch pits in pure niobium.

3. MICROSTRUCTURE

Microstructure is the term that materials scientists use to describe the distribution, shape, and size of phases, grains, etc. in a given material. It is important in view of the fact that the mechanical properties of any material depend strongly on microstructure. Consequently, it is necessary to understand how microstructures evolve during processing and/or application and how to use this understanding to "tailor" the microstructure and thereby achieve desirable properties.

There are numerous metallurgical processes used to manipulate the microstructure of metals and alloys. The most common processes, and those that are discussed here, include solidification (liquid to solid) and various solid-state processing schemes such as heat treating, mechanical working, and combinations of these (see Materials Treatment). Consequently, it is important to understand the structure of liquids and solids, their similarities and differences, and the factors that influence their transformations during both types of processing.

One of the most useful tools available to the materials scientist is the phase diagram (see Phase Equilibria). Phase diagrams are graphical representations of the phases that exist as a function of composition and temperature (most are published for pressures of 1 atmos) when the system is in equilibrium. The phase diagram for the two-component (or binary) Ag-Cu alloy system is shown in Fig. 6 and will be referred to in the following sections in order to describe the microstructural features of interest and how they might be

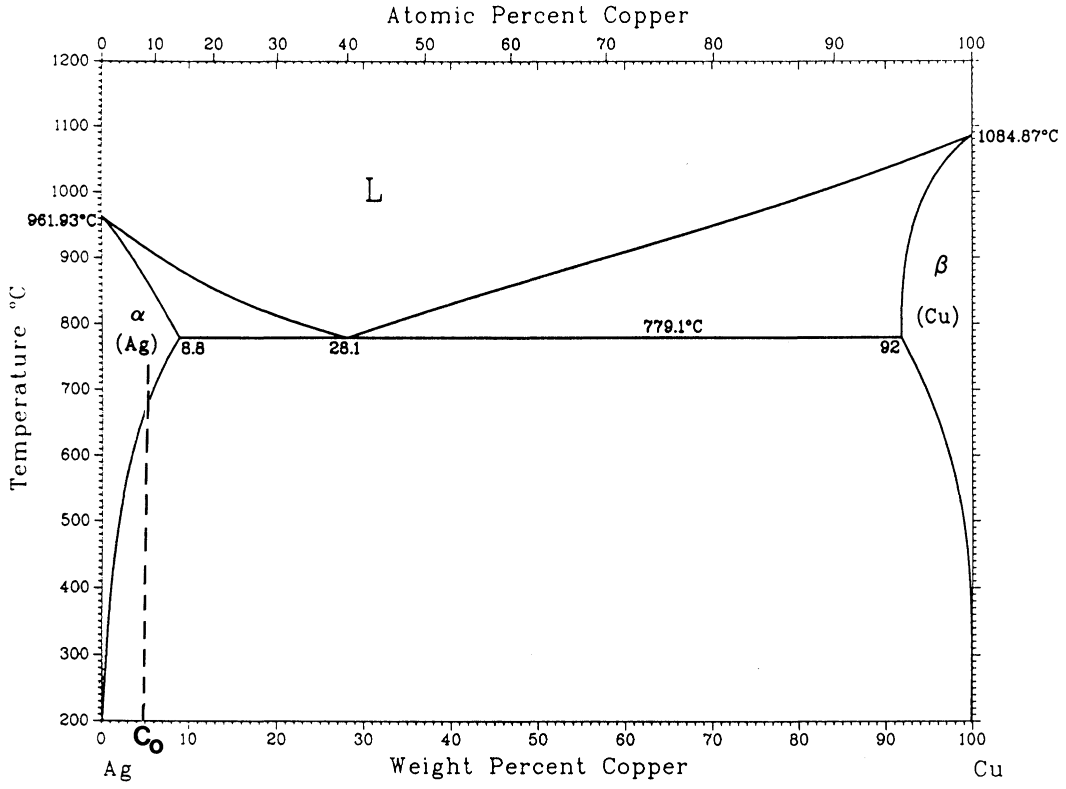

FIG. 6. Ag-Cu phase diagram (ASM Handbook, 1992).

controlled through composition and temperature variations. For simplicity, only binary systems are considered in the following discussion, although it should be recognized that more complex, multicomponent phase diagrams and alloys are frequently encountered in practice.

3.1 Liquid Alloys

Since most commercial metallic objects are formed from a liquid phase at some point during their evolution, it is important to discuss the structure of the liquid state in some detail (see Liquids, Simple, Structure of). First of all, liquid alloys of practical importance occur as single homogeneous phases. Unlike solid crystalline phases, where the arrays of atoms have long-range periodicity, or gases, where the atoms exhibit a complete lack of periodicity, liquid alloys tend to have "intermediate structures," in which the atoms display some degree of local short-range periodicity. This implies that the atoms, over short distances, are arranged in a fashion similar to that characteristic of the solid phase. However, because of the presence of a large number of structural defects, the long-range periodicity characteristic of the crystalline state is not achieved. The volume of the liquid phase is typically within a few percent of that of the solid crystalline phase: it may be larger, as in the case of close-packed solids with high coordination numbers, or smaller, as in the case of less-close-packed solids with lower coordination numbers.

Although the exact nature of the structural defects in liquids remains unknown, it is believed that these defects are responsible for many of the properties of liquid metals. For example, the diffusivity of atoms in the liquid state just above the melting point is typically several orders of magnitude larger than that in the solid state just below the melting point. This implies that the liquid is constantly changing such that the local order existing at any one position in space changes continuously with time. Because of this feature, liquids cannot support shear stresses even when they are very small in magnitude. Finally, it should be mentioned that most liquid metals tend to have similar properties, indicating that the properties are controlled primarily by the structural defects rather than the interatomic bond energies. This is in direct contrast to the properties of metallic solids, which depend strongly on the bonding and crystal structures that exist and less on structural defects.

3.2 Solid Alloys

3.2.1 Cast Structures The microstructures that develop when a metal or alloy is cooled from above to below its melting point depend on such things as composition, cooling rate, diffusivity of the various species in the liquid and solid phases, and the nature of the phase boundaries for the particular alloy under investigation. For pure metals, the structure will typically consist of columnar grains that are oriented in a direction parallel to the direction in which the heat is extracted. For alloys, the situation becomes considerably more complex. Typically, alloys will solidify in either a cellular or dendritic (treelike) fashion, with compositional variations from the center to the periphery of the cells or dendrites. These compositional differences are often undesirable and have to be removed (by some sort of homogenization treatment) prior to application. Such homogenization treatments are frequently achieved by reheating the alloys to elevated temperatures sufficient for the atoms to move (diffuse) from regions of higher concentration to regions of lower concentration. Whenever possible, this type of treatment is performed by heating the alloy of interest into a single-phase field (e.g., the α field in Fig. 6). If the elements involved are "slow diffusers," and/or the scale of the segregation is large, the times required for homogenization may become impractical. In such cases, the segregation is typically removed by mechanically working the alloy. Although there are a variety of methods for achieving this goal, these will not be discussed here, as this topic is beyond the scope of this article.

In addition to the cellular and dendritic structures described above, some alloys can be solidified into two-phase structures. The most common two-phase structures are known as eutectics and occur when two solid phases having different compositions grow "cooperatively" from a parent liquid (liq $\rightarrow \alpha + \beta$ in Fig. 6). The thicknesses and amounts of the two phases are strong functions of their compositions relative to the initial liquid and the imposed cooling rate. In some alloys, the mechanical properties can be enhanced by aligning the eutectic phases; this is achieved by cooling the liquid directionally and is known as directional solidification. An example of a directionally solidified eutectic is shown in Fig. 7.

3.2.2 Wrought Structures The microstructural features of the wrought alloys that are most frequently referred to include grain size, shape, texture, second-phase distribution, and dislocation density (or amount of

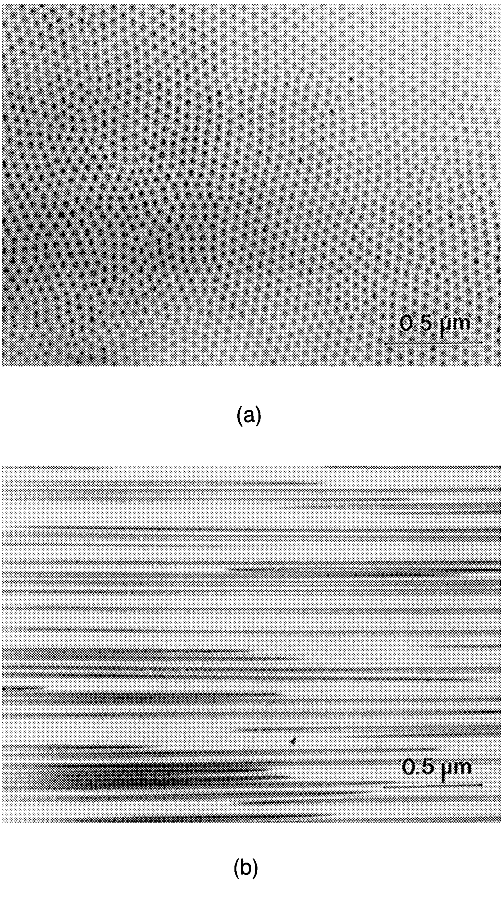

(a)

(b)

FIG. 7. Optical micrographs showing the unidirectionally solidified structure of MnSb-Sb eutectic: (a) transverse section, (b) longitudinal section (Cole, 1971).

cold work). Many of the properties of metals and alloys depend strongly on these features. Figure 8 depicts two examples of grain structures—equiaxed and elongated. As might be expected, the mechanical properties of the alloys are a strong function of the grain distribution. In addition, the properties of the structures with elongated grains are anisotropic; i.e., they vary with test direction relative to the grain boundaries. Another microstructural feature that is commonly observed in the wrought structure of fcc metals is annealing twins, which can be seen in the microstructures shown in Fig. 8.

In some instances, cold working of ductile two-phase cast structures results in very strongly elongated structures, which resemble those found in conventional fibrous or lamellar composites. Since very small (submicron) interlamellar or interfibrous spacing can be produced in heavily deformed alloys, they are frequently referred to as "nanocomposites" (see, e.g., Fig. 9).

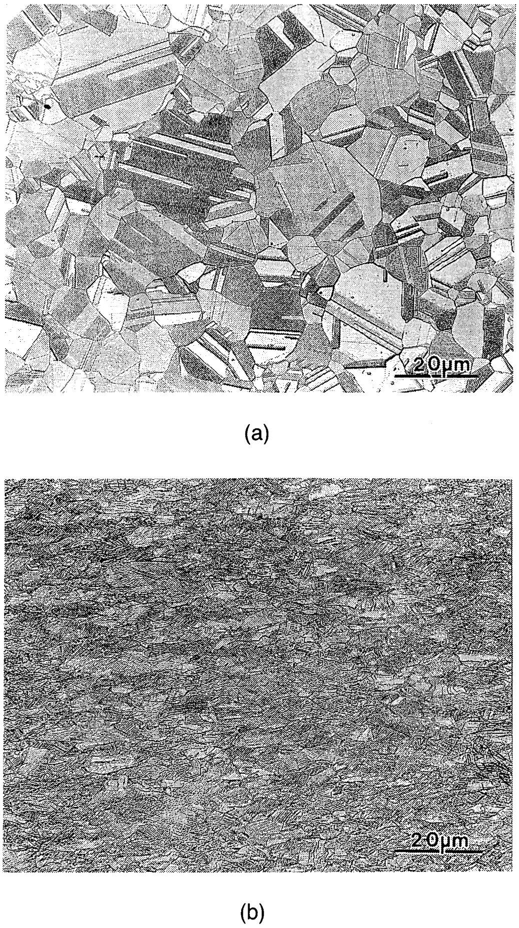

FIG. 8. Optical micrographs showing (a) equiaxed and (b) elongated grain structures in recrystallized and cold-rolled red-brass (Cu–15% Zn alloy). The features in (a) with straight boundaries are annealing twins.

3.2.3 Solid-State Phase Transformations

3.2.3.1 Precipitation. Many alloys are strengthened by second phases, which may be introduced by a variety of methods. Probably the most commonly employed method is precipitation hardening, in which the alloy of interest is heat treated in a manner that produces a fine distribution of submicron second-phase particles that are effective in inhibiting dislocation movement (see, e.g., Fig. 10). Typically, this is achieved by homogenizing the alloy of interest in a single-phase field prior to cooling and then reheating somewhere in the two-phase field in order to cause formation (i.e., precipitation) of a second phase. Referring to the phase diagram in Fig. 6, an alloy of composition C_0 might be heated into the single-phase α field, quenched to room temperature, and then reheated to an intermediate temperature in the two-phase $\alpha + \beta$ field. During this treatment, the fine β precipitates can form in the α matrix, giving rise to considerable strengthening.

Alternatively, some alloys are strengthened with insoluble dispersoids, which may be introduced by rapid solidification processing (see the following) or mechanical alloying, a process in which insoluble phases

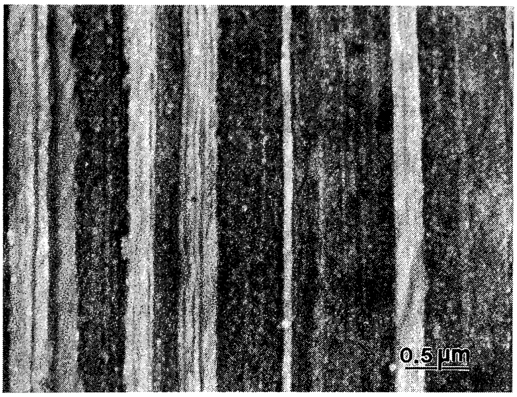

FIG. 9. SEM micrograph of a heavily worked Cu-Ag alloy. Note the fibrous structure of the Ag (light phase) in the Cu matrix. The drawing direction is vertical in the micrograph.

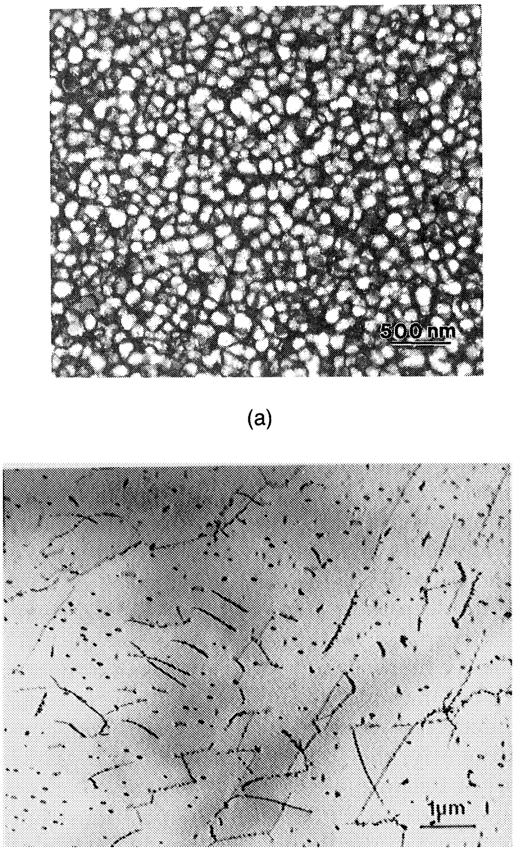

(a)

(b)

FIG. 10. (a) Dark-field transmission electron micrograph of Al_3Li precipitates (white) in an aluminum matrix (dark). (b) Bright-field transmission electron micrograph of nitride precipitates in an Fe–0.3Cu–0.7Ni alloy. Note the uniform distribution of the precipitates in (a) in contrast to the preferential precipitation on dislocations in (b).

are mixed intimately using mechanical means rather than thermal means.

3.2.3.2 Eutectoid Reaction. The equivalent of the eutectic reaction also occurs in the solid state, in which one phase transforms into two other phases of different compositions ($\alpha \rightarrow \beta + \gamma$). The classical and most common example of the eutectoid reaction occurs in steels (Fe-C–based alloys), in which the high-temperature austenite (fcc) phase transforms into the carbon-lean ferrite (bcc) phase and the carbon-rich cementite (Fe_3C) phase. As with eutectic structures, the scale of the eutectoid reaction product is a strong function of cooling rate and is finer at the higher cooling rates (Fig. 11).

3.2.3.3 Martensitic Transformation. In addition to precipitate formation during heat treatment, there are other transformations of technological significance. One of the most important is the martensitic transformation, where the transformation product is known as martensite (see MATERIALS TREATMENT). Martensitic transformations usually occur when an alloy is cooled quickly from a single-phase field at elevated temperatures into a two-phase field at lower temperatures (Fig. 12). Unlike precipitation, which requires long-range diffusion and rearrangement of the atoms in the material, martensite is said to occur in a "diffusionless" manner. Specifically, these reactions occur by movement of the at-

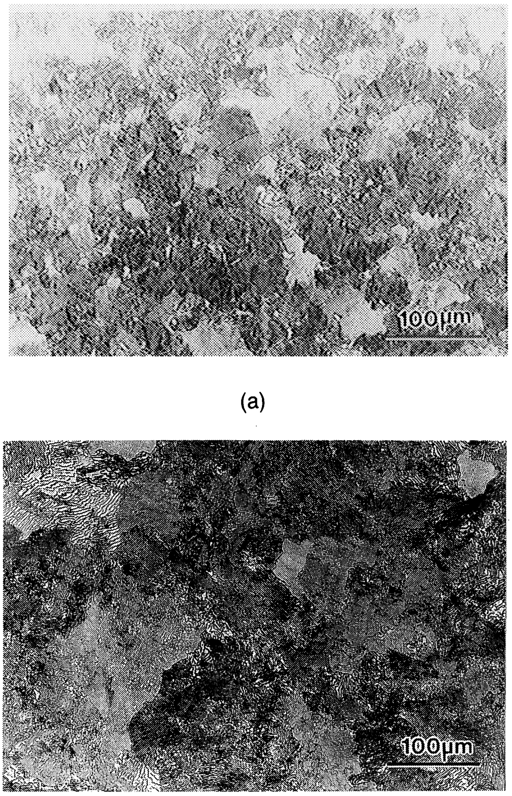

(a)

(b)

FIG. 11. Optical micrographs of a eutectoid steel after (a) air cooling and (b) furnace cooling from the high-temperature single-phase (austenite) field. Note the coarser microstructure after furnace cooling in comparison with that obtained by the faster air cooling.

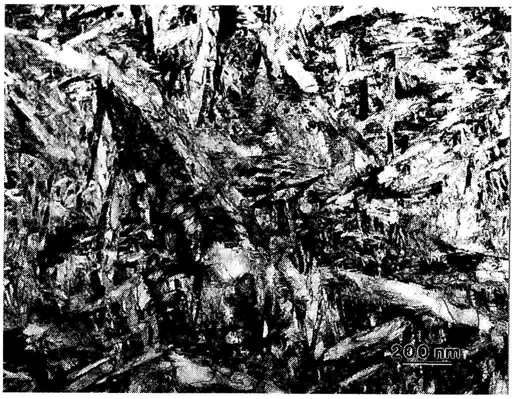

FIG. 12. TEM micrograph of martensite in a Ti–24Al–11Nb alloy after cooling from elevated temperatures. The overall complexity of this type of structure should be noted.

oms in a cooperative or regimented fashion. Furthermore, since the atoms move only short distances, martensitic transformations occur at speeds approaching the speed of sound in the material.

There are many alloys which undergo martensitic transformations—for example, Cu-Zn, Cu-Mn, Ti-Mo, Fe-Ni, and many more. The most common example of martensite occurs in steels. Specifically, when certain iron-carbon alloys (steels) are quenched rapidly from high temperatures, where the γ phase is stable, to room temperature, a body-centered tetragonal phase is formed in which the distortion from the normal α-bcc structure increases with increasing carbon concentration. The morphology of the martensite in the Fe-C system depends on the carbon content and varies from a lath to a plate shape with increasing carbon content (Marder and Krauss, 1969).

3.3 Rapid Solidification Processing

As mentioned above, there has been a considerable amount of work in recent years in the area of rapid quenching of metals and alloys from the liquid and vapor states. This has led to a wide range of metastability-related phenomena such as

1. microstructural refinement in terms of the distribution of phases and segregation during solidification;
2. extended metastable solid solubility of solute atoms in the metallic matrix phase;
3. formation of metastable crystalline phases that have crystal structures considerably different from those predicted from the equilibrium phase diagram;
4. formation of quasicrystalline phases; and
5. formation of metallic glasses.

In order to account for these phenomena, it is important to describe the structures of the metals in their various equilibrium and nonequilibrium states. This section will describe the structure of metallic glasses, nanocrystals, and quasicrystals.

3.3.1 Metallic Glasses When certain alloys are cooled rapidly below their melting points, the formation of the stable crystalline phase may not occur, leading to the formation of what is referred to as a metallic glass. This occurs primarily in those systems where the viscosity of the liquid increases rapidly with decreasing temperature. In these instances, the liquid simply reaches a sufficiently high viscosity that the atomic mobility becomes insufficient for the atoms to come together to form crystalline nuclei, and, consequently, the liquid transforms to a glassy solid. Because of their low atomic mobility, the atoms are "frozen" in the positions they occupied in the liquid, i.e., they have similar short-range periodicity.

The physical and mechanical properties of metallic glasses tend to be considerably different from those of regular crystalline materials. For example, metallic glasses sometimes exhibit high strengths coupled with high ductilities and unusual electrical and magnetic properties. Also, since metallic glasses are single phase with essentially no segregation, they tend to have considerably better corrosion resistances than their crystalline counterparts, which are typically multiphase with some compositional segregation within the microstructure.

Finally, the mechanical properties of metallic glasses vary depending on whether deformation is being performed above or below the glass transition temperature (the temperature at which the glass transforms to a supercooled liquid). For deformation above the glass transition temperature, metallic glasses tend to deform homogeneously, provided they do not crystallize, since they are essentially supercooled liquids. Below this temperature, metallic glasses tend to deform in a very inhomogeneous manner, leading to the formation of highly localized shear bands. Normally, metallic glasses exhibit little if any work hardening, so that their tensile strengths are comparable to their yield strengths.

3.3.2 Nanocrystals

A relatively new class of materials receiving considerable current attention is that referred to as nanocrystals. As the name implies, nanocrystals consist of nanometer-size crystals, which tend to be oriented randomly relative to each other as a result of the manner in which they are produced. Because of their fine structures, nanocrystals are comprised of about 50% grains that exhibit the long-range order characteristic of bulk material and 50% grain-boundary regions.

Unlike the glassy structures just described, in which there is a considerable degree of short-range order between the atoms, the grain-boundary regions in nanocrystals have essentially no order; this lack of order between the atoms gives rise to completely different physical properties (heat capacities, conductivities, etc.) compared with their crystalline and amorphous counterparts. These differences have spurred considerable interest and research into this new class of materials.

Since nanocrystals tend to have unusual properties, they are being considered for unique applications. However, it should be noted that structures with such fine grains are metastable with respect to forming larger grains, and this "coarsening" will occur, given sufficient thermal energy, leading to structures with more bulklike properties. Thus, it becomes important to realize that there are probable temperature limitations on the application of these materials. Furthermore, while descriptions of the production methods are beyond the scope of this article, it is

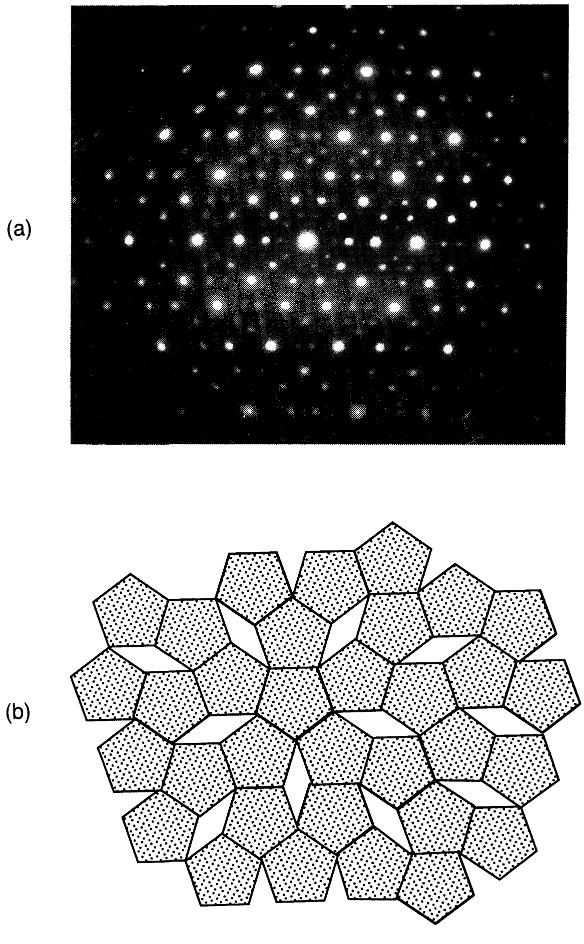

FIG. 13. (a) Electron diffraction pattern demonstrating the fivefold symmetry of the icosahedral phase that forms in Al-Mn alloys when rapidly solidified. (b) Two-dimensional illustration showing that it is impossible to fill space using objects with five-fold symmetry.

worth noting that most of the methods used to date are capable of producing very small quantities of powder only; these small quantities are then consolidated into sections of sufficient size for various physical property measurements.

3.3.3 Quasicrystals In 1984, Shechtman and co-workers from the National Institute of Science and Technology (NIST) reported the formation of a rather unique structure in binary Al-Mn alloys after rapid solidification. Specifically, electron diffraction patterns obtained from this phase (Fig. 13) appeared to display fivefold symmetry, which is unexpected because of the impossibility of filling space with objects of such symmetry [the reader can appreciate this in two dimensions by drawing pentagons and trying to arrange them side by side in a space-filling manner; see Fig. 13(b)]. Consequently, it is difficult to discuss the long-range order of this phase, as discussed above for crystalline phases; for this reason, other structural models have been developed to account for this anomaly.

Since the initial report of the icosahedral phase in Al-Mn alloys, quasicrystalline phases with icosahedral, decagonal (tenfold), and other structures have been reported in numerous other systems. While many of these can only be produced by rapid solidification processing, there are exceptions. For example, the stable T2 phase in the Al-Li-Cu ternary system has the icosahedral structure, and relatively large "crystals" of it have been reported in the shrinkage cavities of slowly cooled alloys; these have been useful for achieving a better understanding of the atomic arrangements in this phase.

As with nanocrystals, the unusual structures of quasicrystals also lead to some rather unique properties. Unfortunately, the property differences between the quasicrystals studied to date and their crystalline counterparts have been insufficient to warrant the application of these materials, and they remain a laboratory curiosity at this time. In fact, the formation of this phase in commercial Al-Li alloys is undesirable from a mechanical properties standpoint and, therefore, is usually avoided if possible. Even so, the complex structures characteristic of quasicrystals continue to generate considerable interest and curiosity among the scientific community.

GLOSSARY

Cohesive Properties: The forces and energies that hold atoms together.

Covalent Bonding: The bonding that occurs when adjacent atoms "share" electrons. Such a bond precludes the possibility of these elements being good conductors, since the electrons are not sufficiently free to conduct a current.

Dendrites: When, during the growth of crystals from liquids, the temperature falls in the liquid in advance of the liquid-solid interface, the solid may grow by branching out, with the appearance of a miniature tree. The word "dendrites" is a Greek word meaning "of a tree."

Diffusion: Thermally activated motion of atoms in matter. Diffusion may occur as a result of the existence of either a temperature or a concentration gradient.

Ductility: Ductility is a measure of the amount of plastic deformation that can be applied to a material before it breaks. Ductility is conventionally measured by tensile testing of metals.

Grain Boundary: The boundary between two grains in a polycrystalline material. For example, during casting, metals solidify by nucleation and growth of many independent grains or crystals that are oriented differently in the space. Obviously, when two crystals meet each other, their atomic arrangements do not match, and, consequently, the intersection plane, i.e., the grain boundary, will be disordered and hence distinguishable from the two crystals.

Ionic Bonding: The bonding that occurs when one element (positive ion) "transfers" electrons to another element (negative ion). The principal bonding force in ionic solids arises from the electrostatic attraction between opposite-sign ions.

Stacking Fault: The atomic arrangement of close-packed structures of metals can be defined by stacking of close-packed planes. In a fcc crystal, the stacking sequence is given by ABC ABC ABC, etc., and, in hcp structure, by AB AB AB, etc. Errors, or faults, in the stacking sequence may develop during solidification or by plastic deformation; these are known as stacking faults.

***Strukturbericht* Designation:** Books that codify all the structural work that is recorded in the literature. They cover the period from 1913, just after discovery of the use of x rays for identifying crystal structures, to the present. Initially, they were published in German, but, since World War II, their publication has been in English.

Texture: A metal that has been deformed extensively—for example, by rolling or wire drawing—may develop a preferred crystallographic orientation or texture. Texturing may also develop during solidification. In a textured structure, the grains of the metal tend to orient themselves in a particular crystallographic orientation.

Twinning: Twinning is a process by which a portion of the crystal takes up an orientation that is related to the rest of the untwinned lattice in a symmetrical fashion. The twinned portion of the crystal is a mirror image of the parent crystal. Twins may be produced by mechanical deformation (mechanical twins) or as a result of annealing following plastic deformation (annealing twins).

Wrought Structure: The microstructure of metals and alloys that have been manufactured by processes that involve plastic deformation.

Yield Stress: Yield stress is conventionally measured by tensile testing and defines the stress at which permanent deformation, i.e., plastic deformation, becomes apparent.

Works Cited

ASM Handbook (1992), *Alloy Phase Diagrams*, Vol. 3, Metals Park, OH: ASM International.

Barrett, C., Massalski, T. B. (1980), *Structure of Metals*, 3rd ed., New York: Pergamon Press, pp. 227, 261, 247.

Cole, G. S. (1971), in: T. J. Hughel and G. F. Bolling (Eds.), *Solidification*, Metals Park, OH: ASM International, p. 290.

Cottrell, A. (1991), *Introduction to the Modern Theory of Metals*, London: The Institute of Metals, p. 94.

Ebrahimi, F., Hoelzer, D. T., Venables, D., Krishnamoorthy, V. (1988), *Development of a Mechanistic Understanding of Radiation Embrittlement in Reactor Pressure Vessel Steels*, NUREG/CR-5063, U.S. Nuclear Regulatory Commission.

Liu, C. T., Stiegler, J. O. (1984), *Science* **226**, 636–642.

Marder, A. R., Krauss, G. (1969), *Trans. ASM* **62**, 2343–2357.

McLean, D. (1962), *Mechanical Properties of Metals*, New York: Wiley, p. 75.

Schechtman, D., Belch, I., Gratias, D., Cahn, J. W. (1984), *Phys. Rev. Lett.* **53**, 1951–1954.

Smallman, R. E. (1985), *Modern Physical Metallurgy*, 4th Ed., Boston: Butterworths, p. 297.

Further Reading

Barrett, C., Massalski, T. B. (1980), *Structure of Metals*, 3rd Ed., New York: Pergamon Press.

Reed-Hill, R. E., Abbaschian, R. (1992), *Physical Metallurgy Principles*, Boston: PWS-Kent Publishing Company.

METALS, LIQUID

See LIQUID METALS, STRUCTURE OF

METEOROLOGY AND CLIMATOLOGY

KEVIN HAMILTON, *Geophysical Fluid Dynamics Laboratory/NOAA, Princeton University, Princeton, New Jersey, U.S.A.*

Introduction 215
1. **Governing Equations and Some Basic Considerations** 216
1.1 The Primitive Equations 216
1.2 Geostrophy 217
1.3 Potential Temperature, Stability, and Available Potential Energy 218
2. **Observed Aspects of Large-Scale Circulation** 219
2.1 Observational Network and Satellites 219
2.2 Surface Climate 221
2.3 Zonal-Mean Climate 221
2.4 Time-Mean Circulation 224
3. **Dynamics of Climate** 226
3.1 Radiative Forcing 226
3.2 Radiative Equilibrium and Radiative-Convective Models of Climate 226
3.3 Effects of Horizontal Heat Transport 227
3.4 General Circulation Models 228
4. **Climate Variability** 229
4.1 The Southern Oscillation and El Niño 230
4.2 Interdecadal Variability 231
4.3 Anthropogenic Effects 232
5. **Dynamics of Weather** 233
5.1 Baroclinic Instability and the Formation of Extratropical Cyclones 233
5.2 Fronts 235
5.3 Tropical Cyclones and Hurricanes 235
Glossary 236
Works Cited 236
Further Reading 237

INTRODUCTION

Mankind has no doubt always had a fascination with and a practical interest in the weather. Meteorological phenomena were a major subject of speculation in the philosophical works of classical antiquity, but scientific study of the weather is generally dated from the invention of the thermometer and barometer in the 17th century. There were sporadic attempts to plot weather maps from surface observations in the 18th century. The invention of the telegraph in the 19th century opened the prospect of producing and disseminating real-time forecasts using data gathered over a large geographical area. Government-sponsored observing networks were begun in several countries in the mid- and late 19th century. The 19th century also saw important developments in basic fluid dynamics and thermodynamics, which put the study of the atmosphere on a firm basis as a problem in applied physics. In recent decades, spectacular advances have been made in both observational and theoretical studies of the atmosphere. This recent progress has been greatly facilitated by the availability of satellite platforms for atmospheric observing systems and the development of digital computers allowing approximate numerical solution of the nonlinear governing equations.

Historically, the study of the atmosphere has been divided into the disciplines of *meteorology* and *climatology*. Climatology could be defined as the study of those processes that determine the time-mean state of the atmosphere—where the mean is defined as an average over some substantial period (a year or perhaps a number of years). Meteorology then deals with the physics of the higher-frequency components of atmospheric variability. It has become increasingly obvious that this distinction is quite arbitrary and not particularly useful. The circulation in the atmosphere displays variability at all time scales, and there are important interactions among the various frequency components.

Much recent work on atmospheric moni-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

toring and modeling has been motivated by an awareness that mankind has the potential to alter significantly (if inadvertently) the global climate. Particular concerns have been raised about the increasing atmospheric levels of so-called *greenhouse gases* such as CO_2 caused by industrial and agricultural activities. Reliable predictions of the sensitivity of climate to such anthropogenic influences would be enormously valuable in formulating strategies to mitigate the social and economic consequences of global environmental change.

1. GOVERNING EQUATIONS AND SOME BASIC CONSIDERATIONS

1.1 The Primitive Equations

Below very high altitudes (say ~100 km), the molecular nature of the atmosphere can be neglected, and the usual continuum approximation of fluid mechanics may be used. The air in this region is sufficiently well mixed that species with very long chemical lifetimes (notably O_2, N_2, and Ar) always occur in the same proportions. Those species with shorter atmospheric lifetimes (notably water vapor) are observed to have concentrations that vary significantly in space and time. The chemical composition of the atmosphere is discussed in more detail in the article ATMOSPHERIC STRUCTURE.

The motions of the atmosphere can be modeled quite well by the Navier-Stokes equations of fluid mechanics formulated for a gas in a gravitational field. Thus, three scalar momentum equations, the first law of thermodynamics, an equation of state, and a continuity equation allow the determination of the three components of the wind and the temperature, pressure and density of the air. In addition, any realistic model of the atmosphere must include a continuity equation for the water vapor. To a very good approximation, the ideal gas law may be used as the equation of state for the atmosphere (noting that the exact value of the gas constant depends on the local concentration of water vapor). The usual approach in meteorology is to formulate the equations so that air motions are defined relative to the earth's surface. The wind components used are *u* in the zonal (east-west) direction, *v* in the meridional (north-south) direction, and *w* in the local vertical (see Fig. 1). The use of the rotating reference frame of the earth necessitates the inclusion of Coriolis and centrifugal forces in the governing equations. The importance of the Coriolis force on motions in the atmosphere (and ocean) is perhaps the most significant feature that distinguishes meteorology (and oceanography) from most other branches of applied fluid mechanics.

The relevant governing equations can be written as

$$\frac{D\mathbf{v}}{Dt} = -\frac{1}{\rho}\nabla p + \mathbf{g} - 2\boldsymbol{\Omega} \times \mathbf{v} + \Omega^2\mathbf{r}, \tag{1}$$

$$C_p\frac{DT}{Dt} = \frac{1}{\rho}\frac{Dp}{Dt} + Q, \tag{2}$$

$$\frac{D\rho}{Dt} = -\rho\nabla \cdot \mathbf{v}, \tag{3}$$

$$p = \rho RT, \tag{4}$$

where **v** is the three-dimensional wind vector defined relative to the rotating frame of the earth, *p* is the pressure, *ρ* is the density, **Ω** is

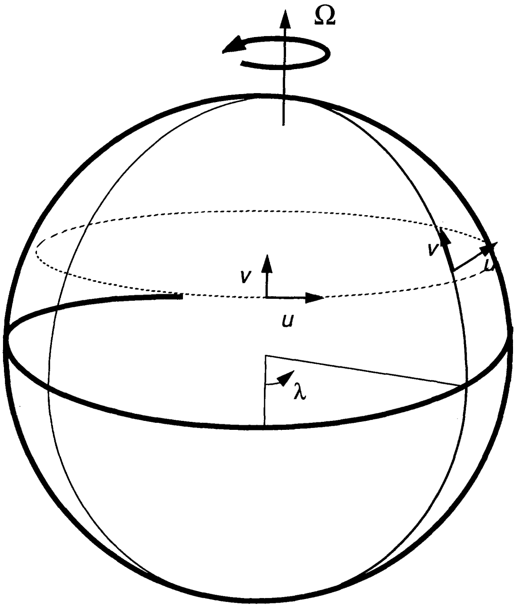

FIG. 1. An illustration of the usual notation used by meteorologists to describe the horizontal wind field. In addition to the horizontal components *u* and *v*, a component *w* is defined as parallel to the local vertical.

the angular velocity vector, Ω is the magnitude of the earth's angular velocity, $\mathbf{g}$ is the vector acceleration of gravity, T is the temperature, $\mathbf{r}$ is the position vector normal to (and pointing outward from) the earth's rotation axis, C_p is the specific heat of air at constant pressure, Q is the diabatic heating, and R is the gas constant. The operator D/Dt is the material (or substantial) derivative reflecting the rate of change following air parcel motions. The effects of molecular viscosity and heat conduction have been omitted in writing Eqs. (1)–(4) but can easily be included.

One very important simplification of the full Navier-Stokes equations results from the smallness of large-scale vertical air velocities in the atmosphere. Generally, w is of the order of 0.1 m s^{-1} or less, and vertical air-parcel accelerations are many orders of magnitude smaller than the gravitational acceleration g. Thus to an extremely good approximation, the vertical momentum equation reduces to hydrostatic balance:

$$\frac{\partial p}{\partial z} = -g\rho = -\frac{gp}{RT}, \tag{5}$$

where z is the local vertical (usually defined relative to sea level). Some minor simplifications to the other governing equations (based on the facts that the atmospheric depths of interest are much less than the earth's radius and that typical vertical velocities are small) are also generally made. Notable among these is the neglect of the part of the Coriolis force associated with the component of the earth's angular velocity vector that lies in the local tangent plane to the earth's surface. The Navier-Stokes equations on a rotating sphere with hydrostatic balance (and these other shallow-atmosphere approximations) are referred to as the *primitive* equations. The primitive equations are generally regarded as the appropriate starting point for even the most comprehensive models of the large-scale atmospheric circulation. Nonhydrostatic equations are normally required only to understand quite small-scale features in the atmosphere (such as intense thunderstorms).

One useful feature of the hydrostatic assumption is that it implies a one-to-one relation between pressure and geometric height. It is generally convenient to work with pressure as the independent vertical coordinate and height as the dependent variable to be computed. Thus, most weather maps are plotted on constant-pressure or *isobaric* surfaces. It is easy to show that the horizontal gradients of height along an isobaric surface are linearly related to the pressure gradients on a constant-height surface. Thus, high- (low-) pressure centers appear on isobaric maps as areas of high (low) height.

The hydrostatic equation can be integrated to relate the pressure at any height to the pressure at some reference height and the temperature in the intervening layer, i.e.,

$$p(z) = p(z^*)\exp\left(-\int_{z^*}^{z} \frac{1}{H}\,dz\right), \tag{6}$$

where the pressure scale height $H(z) = RT(z)/g$. For typical values of atmospheric temperature, H is about 7 km. The global-mean atmospheric pressure at sea level is about 1.013×10^5 Pa (= 1.013 bar). Day-to-day fluctuations of surface pressure are typically 1–2% of the mean. Thus, it is very convenient in meteorology to measure pressure in terms of millibars (mb). The 500-mb level (which marks the height below which roughly half the atmospheric mass resides) is typically found at an altitude between 5 and 6 km.

1.2 Geostrophy

An approximation to the horizontal momentum equations can sometimes be useful in describing atmospheric circulation. In particular, the dominant components of the momentum balance are usually the pressure-gradient term and the Coriolis force. When these terms are exactly equal, the winds and pressure fields are said to be in *geostrophic balance,* i.e.,

$$2\Omega \sin\phi u = \frac{-1}{a\rho}\left(\frac{\partial p}{\partial \phi}\right)_z = -\frac{g}{a}\left(\frac{\partial z}{\partial \phi}\right)_p \tag{7}$$

in the meridional direction and

$$\begin{aligned} 2\Omega \sin\phi v &= \frac{1}{\rho a \cos\phi}\left(\frac{\partial p}{\partial \lambda}\right)_z \\ &= \frac{g}{a \cos\phi}\left(\frac{\partial z}{\partial \lambda}\right)_p \end{aligned} \tag{8}$$

in the zonal direction. Here ϕ is the latitude, λ is the longitude, and a is the radius of the earth. Only the component of the Coriolis force associated with the projection of the earth's angular velocity vector in the local vertical is considered (hence the factor $2\Omega \sin\phi$). The pressure gradient at constant height has been related to the height gradient at constant pressure by use of the hydrostatic relation. Under geostrophic balance, the horizontal wind and horizontal pressure gradients are perpendicular. This accounts for the tendency apparent on any weather map for the winds to blow around isolated high- and low-pressure regions. Of course, geostrophic balance is only a fairly rough approximation to the full momentum equations (errors typically ~10%). This relation breaks down at low latitudes where $2\Omega \sin\phi$ is small. There are also strong deviations from geostrophic balance in the lowest kilometer or so of the atmosphere where stresses from turbulence affect the large-scale horizontal momentum balance. As a result of these turbulent stresses, the wind near the ground generally has a significant component from high pressure toward low pressure. Thus, the surface level winds around a low- (high-) pressure center tend to spiral inward (outward).

The geostrophic and hydrostatic equations may be combined to produce the *thermal wind relation,* which connects the vertical derivative of horizontal wind to the horizontal temperature gradient. In particular, for the zonal wind component,

$$\frac{\partial u}{\partial \ln p} = \frac{R}{2\Omega a \sin\phi} \frac{\partial T}{\partial \phi}. \tag{9}$$

When we note that geometric height is very roughly proportional to $-\ln p$, the left-hand side is a measure of the strength of the vertical shear in the zonal wind. Since in the lower atmosphere the temperature normally decreases toward the pole (i.e., $\partial T/\partial \phi < 0$ in the Northern Hemisphere), one usually finds that the westerly wind intensifies with height.

The geostrophic approximation to the horizontal momentum equations eliminates the time derivative term, and so this relation has no predictive power. There are approximations to the primitive equations that use geostrophic balance to first order but consider the effects of deviations from geostrophy on the time tendency of the pressure and wind fields. These simplified prognostic equations are known as the *quasigeostrophic* system. The quasigeostrophic equations were the basis of early attempts at numerical weather forecasting, and they are still widely used in theoretical studies of atmospheric behavior in the extratropics (Phillips, 1963).

1.3 Potential Temperature, Stability, and Available Potential Energy

The potential temperature θ of air is defined as the temperature it would attain if compressed (or expanded) adiabatically to some standard pressure p^* (usually taken to be 1000 mb). Using the first law of thermodynamics and the equation of state, it is easy to show that $\theta = T(p^*/p)^{R/C_p}$. The potential temperature is obviously conserved during an adiabatic process.

Consider a horizontally homogeneous column of the atmosphere in which $\partial\theta/\partial z > 0$. If a small parcel of air in this column is displaced upward, it will adjust rapidly to the lower ambient pressure at its new level. If the process is rapid enough, it will be adiabatic, and the potential temperature of the parcel will be unchanged (and hence lower than the ambient θ). This means that the parcel temperature will be lower (and the parcel density will be higher) than in the ambient atmosphere. Buoyancy forces will then push the parcel down toward its initial position. Thus, this atmospheric column is regarded as stable to vertical displacements (a situation called *convective stability* by meteorologists). The stronger the positive stratification of θ, the more resistant the atmosphere is to vertical mixing of air parcels. Similar arguments can be made to demonstrate that if $\partial\theta/\partial z < 0$, adiabatic vertical air displacements are reinforced by buoyancy forces (i.e., *convective instability*) and that this should lead to strong vertical mixing within the air column. The large-scale mean temperature structure of the atmosphere is stably stratified, but local regions of convective instability appear each day scattered over the globe.

Figure 2(a) shows a schematic meridional section through a hemisphere of the atmosphere marked with lines of equal θ. Note that $\theta_1 < \theta_2 < \theta_3 < \theta_4 < \theta_5$ so that each atmospheric column is convectively stable. The

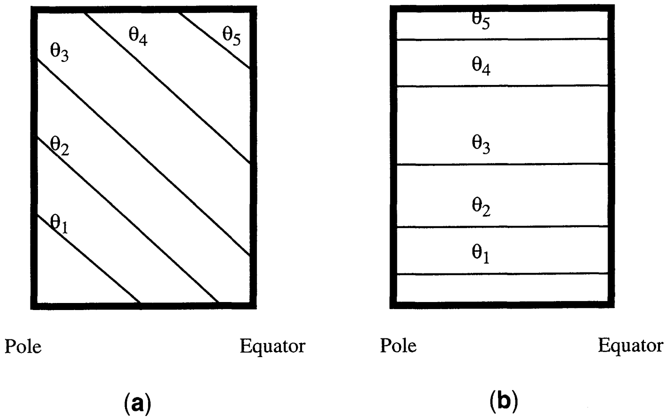

FIG. 2. Schematic meridional section through the atmosphere in one hemisphere. Contours of constant potential temperature are shown. (a) A realistic state for the atmosphere; (b) the lowest energy reference state.

meridional gradient of θ shown in this figure simply reflects the fact that on any horizontal pressure surface, the temperatures are lower at the pole than at the equator. Figure 2(b) shows a hypothetical situation in which the air has been rearranged adiabatically so that the lowest-θ air is at the bottom. This represents the lowest gravitational energy state (consistent with hydrostatic balance) that can be obtained from 2(a) through adiabatic air motions. The difference between the gravitational potential energy in 2(a) and 2(b) is referred to as the *available potential energy* (APE) of the atmosphere in state (a). The APE obviously increases with increasing horizontal temperature gradient. Atmospheric wind systems are ultimately driven by the equator-to-pole gradient in radiative heating that produces APE. Atmospheric motions then grow by converting the APE to kinetic energy of the winds. The strength of the observed atmospheric flow reflects the resulting balance between radiative generation of APE and its release by motions that bring low-θ air downward and equatorward and high-θ air upward and poleward.

This simple picture needs to be modified when the effects of moisture and latent-heat release are included. In particular, the upward displacement of an air parcel that is saturated with water vapor will not conserve θ. As the parcel rises, it will cool less rapidly than a dry (or unsaturated) air parcel because of the release of latent heat as water condenses out. Discussions of the thermodynamics of moist air can be found in any standard meteorological textbook (e.g., Wallace and Hobbs, 1977).

2. OBSERVED ASPECTS OF LARGE-SCALE CIRCULATION

2.1 Observational Network and Satellites

Our detailed knowledge of the atmospheric circulation and its variations comes largely from analysis of the routine measurements taken every day for the purposes of operational weather forecasting. One component of this observational data set consists of measurements of wind speed and direction, barometric pressure, precipitation, air temperature, and humidity at three-hour intervals at several thousand surface stations (and on commercial ships) throughout the world. Another very important component of the data archive is the collected measurements from *radiosondes* borne aloft by unmanned balloons. These radiosondes are instrument packages that transmit values of the pressure, temperature, and humidity at specified intervals during the balloon ascent. By tracking the balloon visually or with radar, the horizontal wind speed and direction can also be determined as a function of height. These radiosonde ascents are typically made once or twice a day at several hundred stations worldwide. Figure 3 shows the distribution of the 1093 stations where upper-air balloon measurements were performed for some period during 1963–73. On any given day, roughly 600 stations actually reported observations. Weather balloon measurements are normally limited to heights below about 30 km.

Balloon measurements are critical for exploring the three-dimensional structure of the

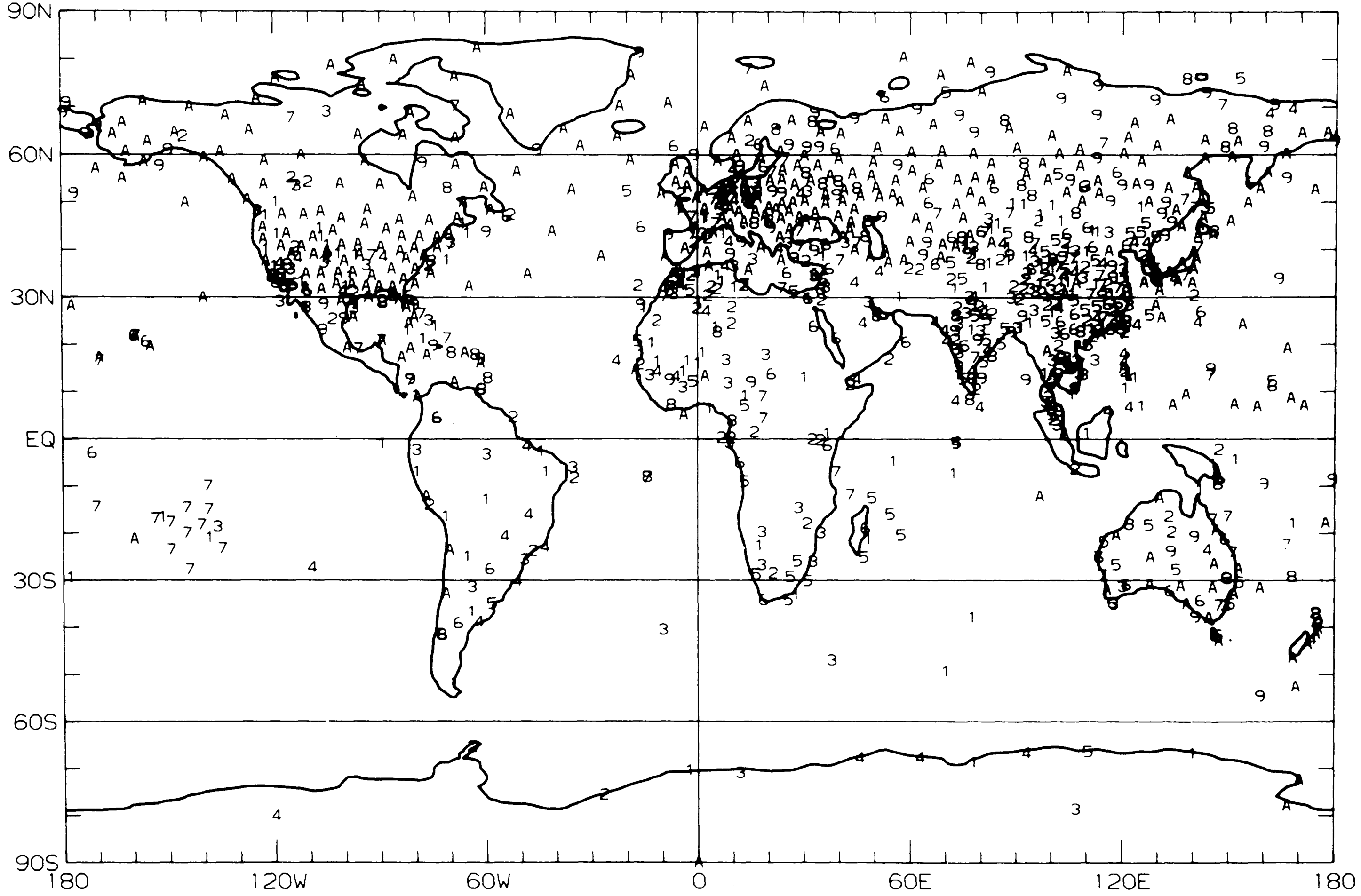
90N
60N
30N
EQ
30S
60S
90S
180
120W
60W
0
60E
120E
180

FIG. 3. The 1093 stations for which balloon-borne radiosonde data were available at 500 mb for at least one January between 1964 and 1973. The number at each station location shows the number of years for which fairly complete January data records exist ("A" indicates that data is available from all ten years). From Peixoto and Oort (1992).

atmospheric circulation, but they have obvious limitations in terms of geographical coverage (particularly over the oceans). The gaps in station coverage can be filled in to some extent by meteorological satellites. *Geostationary satellites* provide fairly continuous observations of the cloud field over the tropics and midlatitudes. Figure 4 shows a typical image from a geostationary satellite. *Polar orbiting satellites* in fairly low orbits carry radiometers that measure outgoing radiation at the top of the atmosphere. By performing these measurements at several wavelengths, information concerning the vertical structure of temperature and humidity can be extracted (see ATMOSPHERIC PHYSICS for more details). These satellites provide information along a sun-synchronous satellite trajectory as the earth rotates beneath it. Thus, global coverage is achieved roughly twice each day.

2.2 Surface Climate

The observed long-term mean of the surface air temperature in January is shown in Fig. 5. This figure displays a feature common to many climatological statistics, namely that the zonal gradients are generally weaker than the meridional gradients. This tendency reflects the fact that the daily-average solar radiative flux incident at the top of the atmosphere is symmetric about the rotation axis. The deviations from axial symmetry in long-term mean climate data such as that shown in Fig. 5 are caused by the inhomogeneity of the earth's surface (i.e., land-sea contrast and topography).

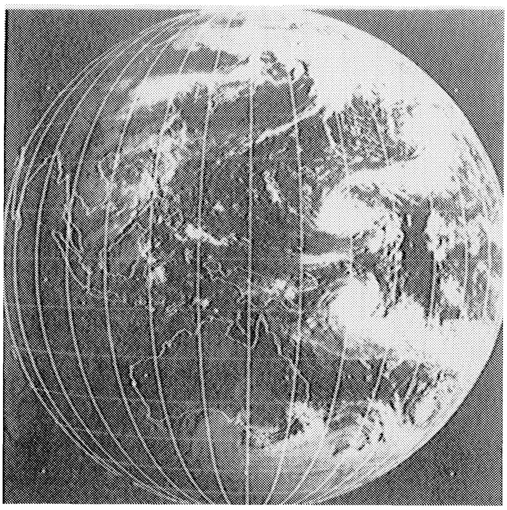

FIG. 4. A visible-light image of the earth taken on 18 May 1986 from a satellite in geostationary orbit above 160°E. The continental outlines and latitude-longitude grid have been superposed on the actual image. Reproduced from Philander (1991), courtesy of Taiwan Central Weather Bureau.

Figure 6 shows the long-term mean of the horizontal wind obtained from surface observations for December–February. Noteworthy here is the predominance of westerly winds (i.e., winds blowing from the west) in the midlatitudes and easterly winds in the tropics (the so-called *trade winds*). Beyond this some significant deviations from zonal symmetry are evident. In particular, there is a net tendency for surface flow off the continent onto the oceans in the winter hemisphere and a reverse flow in summer. In midlatitudes, this effect generally makes a fairly small perturbation to the dominant zonal flows, but in the tropics this *monsoon* circulation can dominate the wind field. In particular, the flow over the Indian Ocean is oriented outward from the Asian land mass in winter (winter monsoon) and completely reversed in summer (summer monsoon). The wind field in the 40°S–60°S band is particularly zonal, presumably reflecting the virtual absence of east-west inhomogeneity in the lower boundary at these latitudes.

Also shown in Fig. 6 are contours of the height of the 1000-mb surface (very nearly proportional to the sea-level pressure field) averaged for the same period. The tendency for the winds to blow parallel to the constant-height surfaces is quite evident in the extratropical regions. In the tropics, the component of flow from high to low pressure is more pronounced.

2.3 Zonal-Mean Climate

Given the tendency for climate elements to display a very approximate axial symmetry, it is useful to examine the zonally averaged circulation to obtain a rough picture of the state of the atmosphere. In the usual terminology, the circulation can be divided into the *zonal-mean* (i.e., the average around each latitude circle) and the *eddy* (the deviation from the zonal-mean) components. Figure 7 shows the height-latitude section of the long-term average of the zonal-mean tem-

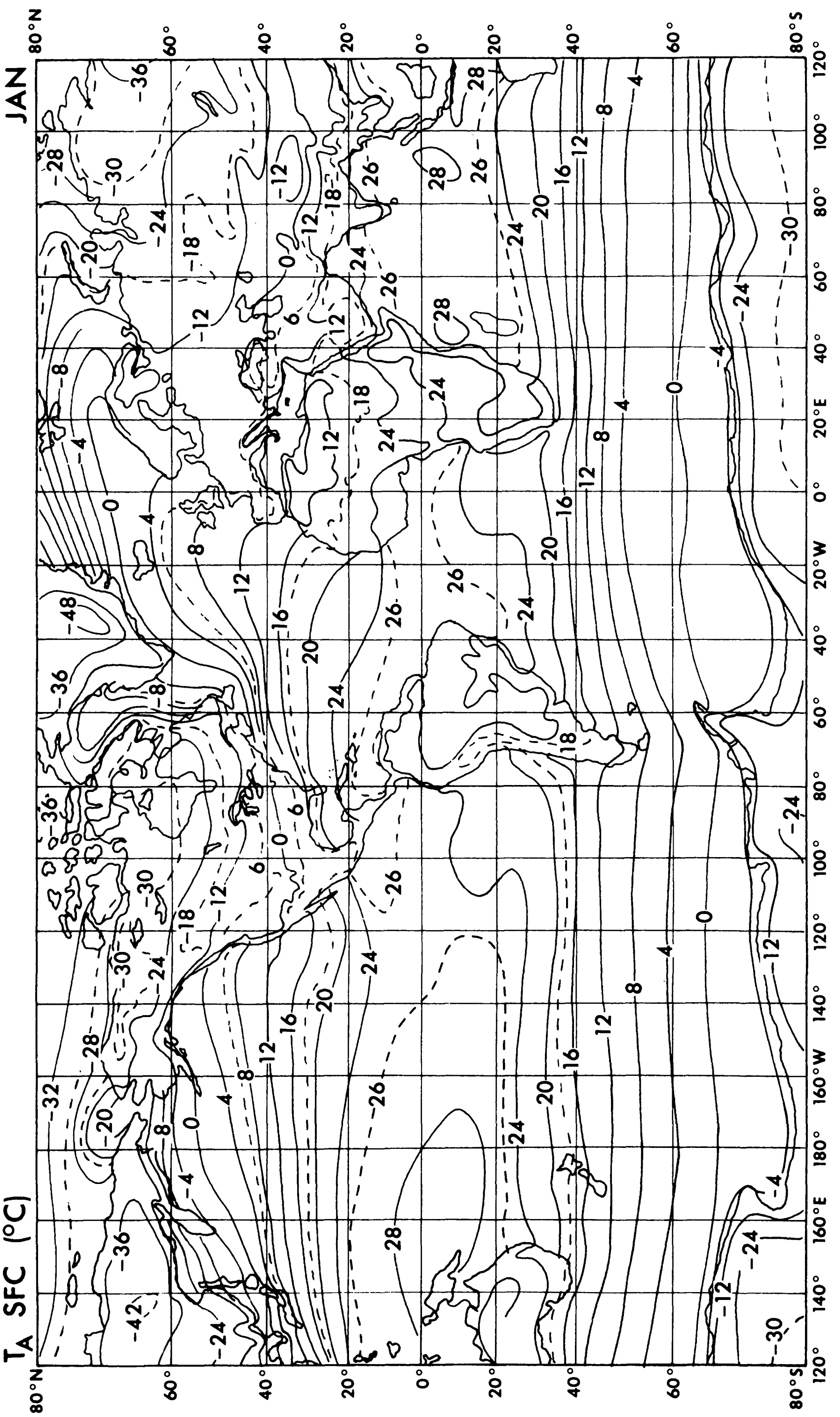
T_A SFC (°C)
JAN
80°N
60°
40°
20°
0°
20°
40°
60°
80°S
120°
140°
160°E
180°
160°W
140°
120°
100°
80°
60°
40°
20°W
0°
20°E
40°
60°
80°
100°
120°

FIG. 5. The long-term mean (°C) of the observed surface air temperature in January. From Peixoto and Oort (1992).

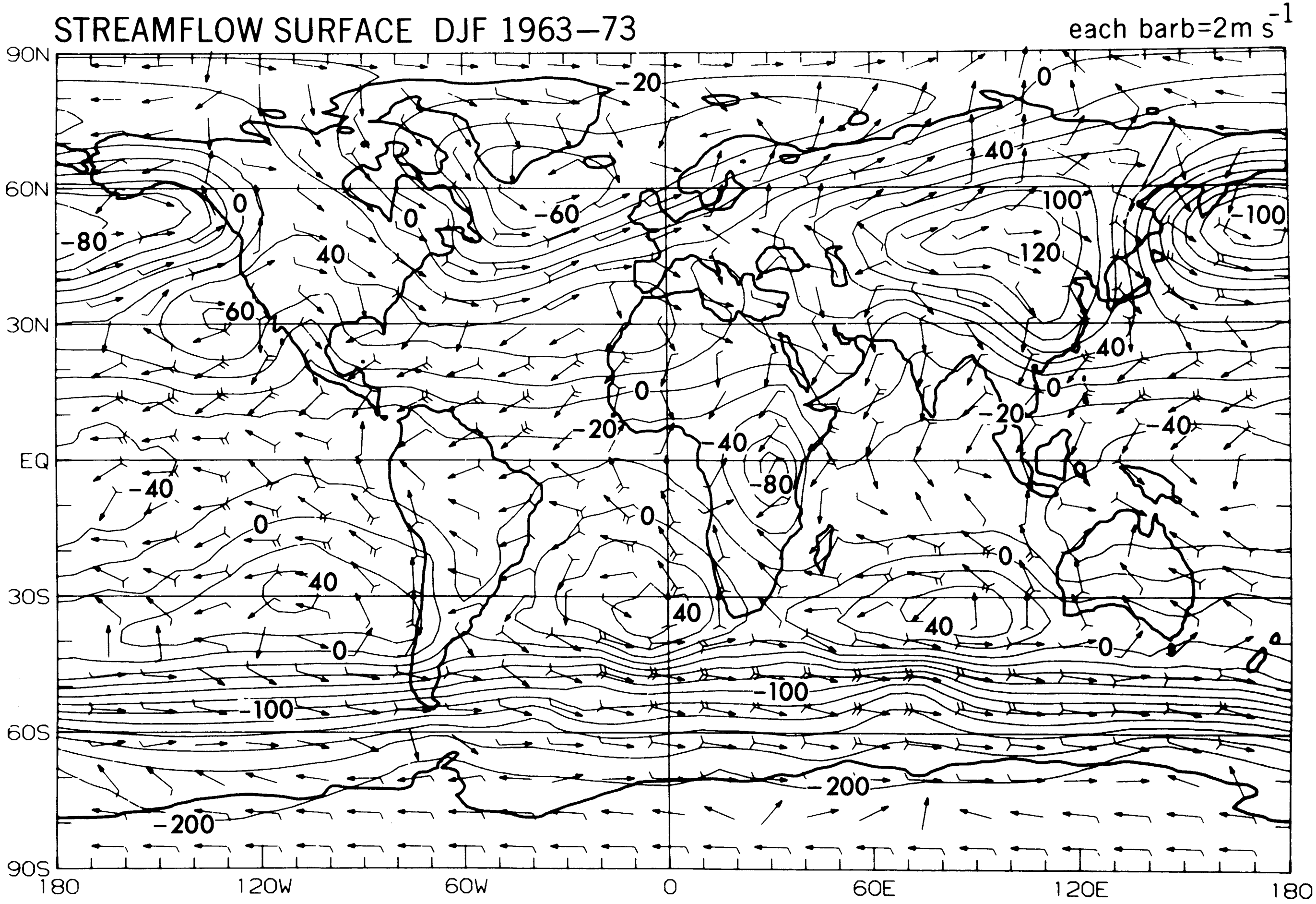
STREAMFLOW SURFACE DJF 1963–73
each barb=2m s-1
90N
60N
30N
EQ
30S
60S
90S
180
120W
60W
0
60E
120E
180
-20
0
40
100
-100
120
-80
0
40
-60
60
0
20
-40
-80
0
40
40
0
-100
-100
-200
-200

FIG. 6. Long-term mean surface-flow climatology for the December–February period. The arrows show the direction of the horizontal wind measured about 2 m above the surface. The wind speed is represented by the number of barbs on each arrow (each barb represents 2 m s^{-1}). The contours show the height above sea level of the 1000-mb surface relative to the global mean value of 113 m. From Peixoto and Oort (1992).

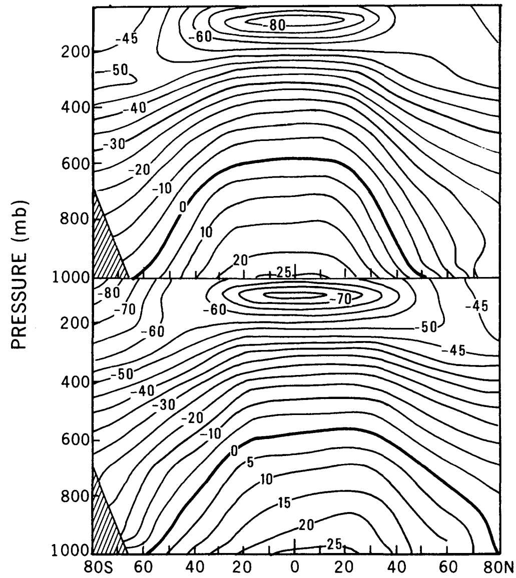

FIG. 7. Long-term mean of the zonally averaged temperature in the December–February (top) and June–August (bottom) periods. Contour labels in °C. Adapted from Peixoto and Oort (1992).

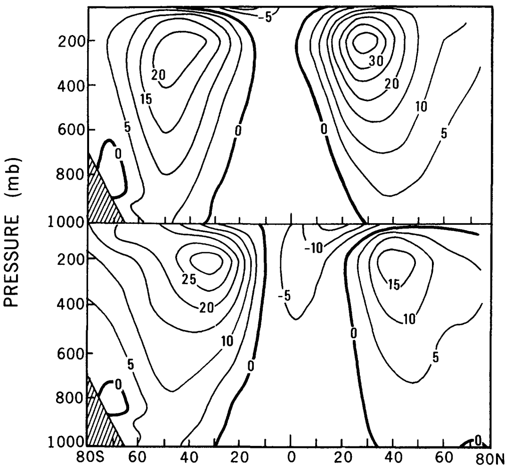

FIG. 8. Long-term mean of the zonally averaged zonal wind in the December–February (top) and June–August (bottom) periods. Contour labels in m s^{-1}. Positive values denote westerly winds (i.e., winds blowing from the west). Adapted from Peixoto and Oort (1992).

perature up to 100 mb (about 18 km altitude) in December–February (top) and June–August (bottom). This figure clearly shows two regions with different vertical temperature structures. The *troposphere,* where the temperature decreases with height, extends from the ground to the minimum-temperature level called the *tropopause.* The tropopause is at about 300 mb (~8 km) near the poles and reaches to about 120 mb (~17 km) at the equator. Above this region is the *stratosphere* where temperature increases with height. The stratosphere extends up to about 0.5 mbar (~50 km). See ATMOSPHERIC STRUCTURE for more details on the temperature profile of the upper atmosphere. In the troposphere, the temperatures at any level are generally warmer near the equator than in the polar regions. This meridional temperature gradient is significantly stronger in the winter hemisphere than in the summer. This difference reflects the stronger pole-to-equator contrast in incident solar radiation in the winter. Remarkably, the meridional temperature gradient is reversed (so that the pole is actually warmer than the equator) in the winter lower stratosphere (e.g., near 150 mb in Fig. 7).

Figure 8 shows the comparable height-latitude sections for the long-term mean zonally-averaged zonal wind. In both seasons and both hemispheres, the most prominent feature in this figure is the core of strong westerlies peaking in the midlatitudes near 200 mb. Each of these wind maxima is the time-averaged and zonally-averaged manifestation of a *subtropical jet stream.* Even in this average sense, the winds in these jets are quite strong, and at a particular time and place the jet stream level winds may exceed 100 m s^{-1}. The fact that there is westerly shear (i.e., u increases with height) up to 200 mb is consistent with thermal wind balance between the zonal wind and the meridional gradient of temperature field in Fig. 7.

2.4 Time-Mean Circulation

The eddy component of the atmospheric circulation has particularly strong day-to-day variability associated with the development, decay, and propagation of weather disturbances. However (as seen above in the surface fields shown in Figs. 5 and 6), there are important eddy components in the long-term mean of the circulation as well. The deviations from the zonal average seen in long-term mean fields are referred to as the *stationary eddies* or *stationary waves.*

The contours in Fig. 9 show the long-term mean November–March 500-mb height field

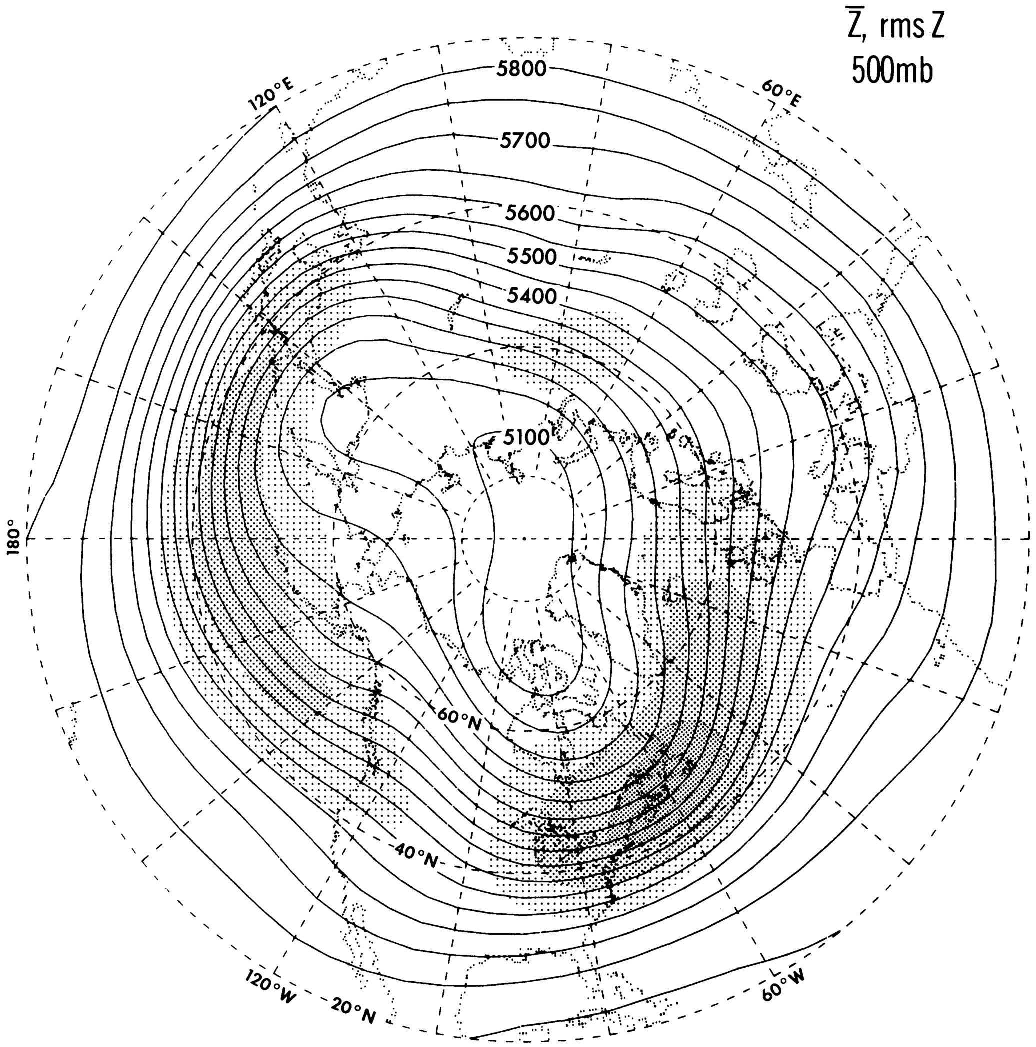

FIG. 9. Long-term mean for the November–March period of the height of the 500 mb surface over the Northern Hemisphere. Contour labels in m. The shading denotes regions of high day-to-day transient fluctuations in the 500-mb height. In particular, the light, medium, and heavy shading correspond to regions where the standard deviation of the 500-mb height time series, filtered to retain only periods between 2.5 and 6 days, exceeds 50, 60, and 70 m, respectively. From Lau (1988), courtesy of the American Meteorological Society.

over the Northern Hemisphere. The most evident feature is the decrease of height with latitude. This is geostrophically consistent with the strong mean westerlies at this height (Fig. 8). Superimposed on this zonally symmetric variation is a longitude-dependent component that seems to be dominated by a wave-number 2 pattern at high latitudes and a wave-number 3 pattern in midlatitudes. In particular, the meridional gradient of the geopotential (and hence the zonal wind) is enhanced locally over the eastern North Pacific and the eastern North Atlantic regions. This is the mid-tropospheric manifestation of the stationary wave pattern seen in the surface pressure field in Fig. 6. The stationary waves play an important role in controlling the day-to-day weather. In particular, regional-scale transient weather systems tend to be strongest in particular geographical regions called *storm tracks*. The intensity of the transient disturbances is shown by the shading in Fig. 9. The time series of the 500-mb height at each location was bandpass filtered to retain only fluctuations with periods between 2.5 and 6 days. Then the standard deviation of the filtered time series was computed. The three levels of shading in Fig. 9 show where this standard deviation exceeds 50, 60, or 70 m. It is apparent that the two principal Northern Hemisphere storm tracks are oriented along a rough west-southwest/east-northeast axis in the North Atlantic and in the North Pacific. These correspond closely to the regions of enhanced meridional height gradients seen in the time-mean pattern in Fig. 9.

3. DYNAMICS OF CLIMATE

3.1 Radiative Forcing

On an annual-mean, global-mean basis, 30% of the solar radiation incident at the top of the earth's atmosphere is reflected elastically back into space, 19% is absorbed by various atmospheric constituents, and 51% is absorbed by the earth's surface. The direct atmospheric absorption of solar radiation (particularly by O_2 and O_3) is crucial for maintaining the temperature structure of the stratosphere and higher regions. This is discussed in more detail in the articles ATMOSPHERIC PHYSICS and ATMOSPHERIC STRUCTURE.

For the troposphere, the dominant radiative effects are the solar heating of the ground and the exchange of infrared (IR) radiation between the ground and the overlying atmosphere. The IR opacity of the atmosphere is provided primarily by water vapor, clouds, CO_2, and O_3. The downward flux of IR radiation from the atmosphere helps to heat the surface in a manner described in the article ATMOSPHERIC PHYSICS. This phenomenon is sometimes called the *greenhouse effect,* and it will be enhanced (and hence surface temperatures are expected to rise) if the IR opacity of the atmosphere increases (e.g., by anthropogenic CO_2 emissions). The detailed mathematical treatment of both solar and IR radiative transfer in the atmosphere is reviewed in Liou (1980).

3.2 Radiative Equilibrium and Radiative-Convective Models of Climate

The simplest model of climate is one in which the solar heating at each point in the atmosphere is balanced by the IR cooling, a situation called *radiative equilibrium.* Actually to implement such a model for a single column of the atmosphere, one needs to specify the distribution of all radiatively active constituents as well as the reflectivity of the surface and cloud layers. A detailed computer code that solves a finite-difference version of the governing radiative transfer equation can be used to calculate the radiative heating rates for any given vertical temperature structure. Then an iterative algorithm is applied to determine the atmospheric $T(z)$ (as well as the earth's surface temperature) that leads to exactly zero heating rates. The solid curve in Fig. 10 shows the result of such a calculation using a very detailed radiative model and employing absorber distributions, solar zenith angle, etc. appropriate for 35° latitude at equinox. The result is quite unrealistic with temperatures of more than 60 °C predicted for the ground and a drop-off of more than 100 °C in the lowest kilometer of the atmosphere. In reality, the large-scale mean temperature is observed to fall off at a fairly constant rate of about $-6.5\ ^\circ\mathrm{C\ km}^{-1}$ in the troposphere (see Fig. 7), and actual midlatitude surface temperatures are closer to 20 °C.

The failure of the purely radiative model to simulate a realistic structure for the troposphere suggests that the effects of air motions are of critical importance in determining atmospheric temperature. The radiative equilibrium model can be modified in a simple way to take into account the first-order effects of dynamics. The pure radiative-model prediction has temperature decreasing so fast that the result is convectively unstable. The real tropospheric $T(z)$ is determined by a balance between the effects of radiation trying to heat the near-surface layer, and the effects of vertical mixing produced by convective instability in transporting warm (i.e., high-θ) air upward and cold (low-θ) air downward. Convective instability can homogenize θ in the vertical in a few hours, while the time scale for radiative processes to establish temperature gradients in the lower atmosphere is of the order of weeks. Thus, the troposphere might be expected to be in a condition of nearly uniform θ (which can be shown to be equivalent to a uniform temperature decrease $\partial T/\partial z = -9.8\ ^\circ\mathrm{C\ km}^{-1}$). This simple prediction must be modified to account for the fact that most vertical mixing in the

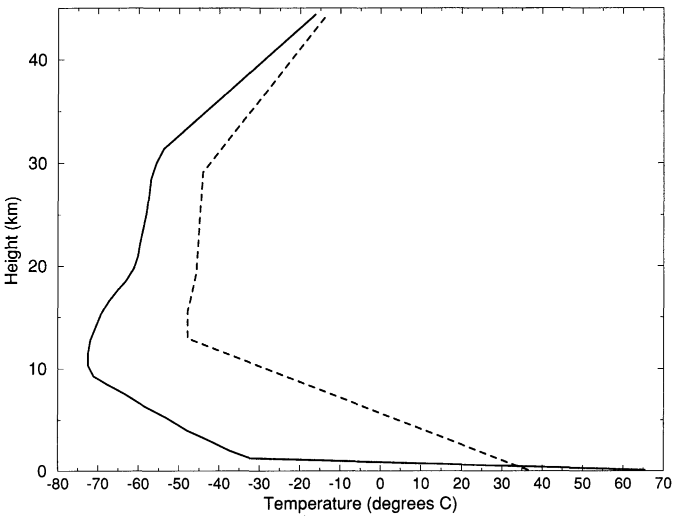

FIG. 10. The predicted vertical temperature structure $T(z)$ for a radiative equilibrium model (solid curve) and a radiative-convective equilibrium model (dashed curve). Results appropriate for cloudless skies at 35° latitude at equinox. Adapted from Manabe and Wetherald (1967).

troposphere occurs in clouds, and the associated latent-heat release actually maintains a positive stratification in the $\theta(z)$ profile. In *radiative-convective equilibrium* (RCE) climate models, the details of the maintenance of the tropospheric stratification are ignored, and a simple constraint is imposed so that enough heat is mixed upward to ensure that the predicted $\partial T/\partial z$ is never steeper than the observed value of -6.5 °C km^{-1}. The dashed curve in Fig. 10 shows the result of adding this fix to the radiative model. The result is very encouraging in that more reasonable surface temperatures are predicted. The temperatures are still too warm in this particular simulation because the reflectivity of clouds was not considered (when this effect is included, a very realistic surface temperature is obtained; see Manabe and Wetherald, 1967). The application of RCE models to climate studies is reviewed in Ramanathan and Coakley (1978).

3.3 Effects of Horizontal Heat Transport

While the RCE model can produce reasonable predictions of the global-mean vertical temperature structure, it fails in simulating the full range of seasonal and latitudinal variation of climate. As an extreme example, in the polar night region a straightforward application of the RCE model would result in a predicted temperature of 0 K . The actual temperature structure in the atmosphere is very strongly influenced by the dynamical transport of heat from low latitudes to the polar regions. Figure 11 summarizes the terrestrial heat budget using the climatological radiative fluxes at the top of the atmosphere (measured from satellite platforms). The top panel gives the annual mean of the absorbed solar flux (i.e., the incident downward flux minus upward reflected shortwave flux) as a function of latitude. Naturally, this quantity peaks in the tropics and drops rapidly with latitude. The middle panel shows the annual mean outgoing IR radiation at the top of the atmosphere. This also peaks in low latitudes, but the equator-to-pole contrast is much weaker than for the absorbed solar component. The bottom panel gives the difference between the curves in the upper panels, thus representing the net radiative heating of the atmosphere/ocean column (the area-weighted global mean of this curve is zero). This shows that the tropical region is heated by radiation, while the regions poleward of about 40° latitude are cooled. This net heating and cooling must be balanced by dynamical transports in the atmosphere and ocean that act to move heat poleward. More detailed analysis of atmospheric data suggests that in the extratropics this transport is accomplished largely by the day-to-day weather disturbances, which act to mix cold polar air equatorward and warm air poleward. The dynamics of this process will be examined in Sec. 5.1. By contrast, in low latitudes most of the atmospheric heat transport is in the form of latent heat of condensation (i.e., moist

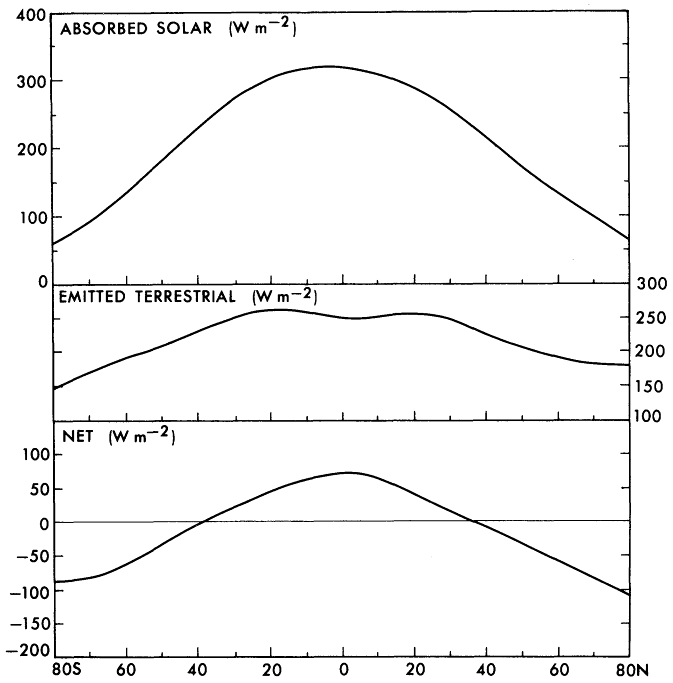

FIG. 11. Components of the long-term mean radiation budget as measured at the top of the atmosphere. Top panel, the net incident downward solar radiative flux. Middle panel, the outgoing terrestrial IR flux to space. Bottom panel, the difference between these quantities. Adapted from Peixoto and Oort (1992).

air is transported poleward where it condenses and releases latent heat), and the poleward transport of heat within the ocean is significant. The prediction of realistic meridional heat transports is a critical test of the adequacy of a climate model.

3.4 General Circulation Models

The most comprehensive numerical models for the climate system are referred to as general circulation models (GCM's). At the minimum, a GCM involves a time-dependent simulation of the full three-dimensional wind and temperature field in the global atmosphere. The model climate is then found as a mean of a very long numerical integration (e.g., several years of model time) with the numerical code.

GCM's now exist with varying degrees of sophistication. All involve some numerical approximation to solve the basic primitive equations (and the continuity equation for atmospheric water vapor). This is normally done with centered second-order finite differences in time. In *grid point* models, the horizontal derivatives are approximated by finite differences on a specified three-dimensional grid. In *spectral* models, the dependent variables are written as linear combinations of some finite set of spatial basis functions, and so the original differential equations are replaced by a large coupled set of first-order ordinary differential equations (in time) for the coefficients of each of the basis functions. In practice, spectral GCM's have used finite differences in the vertical and a spectral representation in terms of spherical harmonics in the horizontal (e.g., Gordon and Stern, 1982).

At present, the typical resolution of a model employed in climate studies (i.e., one that can be readily integrated for many years of model time) might be ~400 km in the horizontal and ~2 km in the vertical (with a total of ~15 levels). Such a GCM would be integrated with a time step that would be in the range of ~5–30 minutes (depending on the details of the model formulation).

In addition to the dynamical core of the model designed to solve explicitly for those aspects of the atmospheric flow resolvable on the numerical grid, a GCM needs to include other processes. In particular, the effects of motions with scales smaller than those resolvable by the model must be incorporated. These effects have to be parametrized; i.e., the contributions of the subgrid-scale flow must be expressed somehow in terms of the resolved variables. The usual approach to this problem is to assume that unresolved scales act as a dissipation of the resolved-scale flow. A special case of such parametrization is the treatment of clouds (which in the real atmosphere have scales much smaller than resolved in current GCM's). The degree of cloudiness, the vertical mixing, and the latent-heat release in any vertical column of model grid points are generally specified as functions of the convective stability as well as the humidity field on the resolved grid. The model precipitation is determined as part of this subgrid-scale vertical-mixing parametrization. The effects of radiative transfer are also included by solving the radiative transfer equations in each column of grid points (using the instantaneous temperatures as well as the distribution of water vapor and cloudiness predicted by the model).

A GCM also requires the effects of interactions between the atmosphere and the underlying surface to be included. Heat, moisture, and momentum are all transferred between the atmosphere and the underlying ocean or land. The atmospheric model needs to know the surface temperature and, over land, the soil wetness (as well as a measure of the roughness of the surface). These quantities can be either specified (usually using climatological data) or calculated using very simple models of the land surface and ocean. A more ambitious approach is actually to couple the atmospheric GCM to an analogous three-dimensional, time-dependent simulation model for the global ocean. In such a coupled model, the ocean provides surface temperature to the atmospheric model, which in turn provides the heat flux, water flux (through precipitation and evaporation), and surface wind stresses that drive the ocean circulation.

A fairly complete description of a typical finite-difference GCM is found in Holloway and Manabe (1971), and a description of a typical spectral GCM is given in Gordon and Stern (1982). While the development of more sophisticated GCM's is an area of active research, these models have already produced impressively realistic results. The tropospheric zonal-mean temperature and wind structure, the strength and position of the stationary waves and storm tracks, and even the climatological distribution of precipitation can now be reasonably well simulated.

4. CLIMATE VARIABILITY

The interannual variability of seasonal-mean temperature and precipitation at a particular location is a matter of common experience. To some extent, this simply reflects the sampling of day-to-day weather variability. This sampling effect can be reduced by considering area-averaged quantities (and not surprisingly, the variance in the time series of an area-averaged variable decreases as the area included in the mean is enlarged). However, even at global scales there is clear evidence for systematic interannual variability in climate. Detailed understanding of the nature of this variability would have great practical value. In particular, recent developments in modeling have raised the possibility of producing useful forecasts of seasonal-mean climate anomalies several months in the future. Detection and modeling of possible anthropogenic effects on climate also require an understanding of the natural "background" climate variability.

The current understanding of the basic nature of interannual climate variability is closely connected to the concept of *atmospheric predictability*. If two parallel runs of a sophisticated three-dimensional GCM (with fixed surface boundary conditions) are conducted, each starting with slightly different initial conditions, the results are always found to diverge very strongly after a period of a few weeks. The period it takes for the two integrations to diverge is called the *theoretical limit of predictability*. After this time limit, the model has essentially "lost" memory of its initial conditions (behavior which has analogs in many other kinds of turbulent fluid systems). This somewhat vague definition can be made more precise, and other means have also been used to estimate the predictability limit (Lorenz, 1969). For the present purposes, the limited persistence of memory in the atmosphere suggests that interannual fluctuations must critically depend on the interactions of the atmosphere with the ocean or land surfaces (which act as the long-term memory for the coupled system). While the land surfaces can play a significant role (through storage of soil moisture and changes in vegetation), it is the oceans with their enormous capacity to store and transport heat that are thought to be most important in the generation of interannual variability in the climate system.

4.1 The Southern Oscillation and El Niño

One quite systematic component of large-scale interannual variability in the atmosphere has its roots in the tropical Pacific Ocean. The *Southern Oscillation* (SO) involves the shifting of atmospheric mass (and hence surface pressure) between the eastern and western tropical Pacific regions. Figure 6 shows the long-term mean pattern with high surface pressure in the equatorial eastern Pacific and low pressure in the equatorial western Pacific. The mean winds near the equator then blow from east to west. At one extreme of the Southern Oscillation, this climatological pattern is intensified, while at the other extreme both the zonal pressure gradient and wind direction in the equatorial Pacific reverse. The period of oscillation varies between about 2 and 6 years. The phase of the oscillation has traditionally been measured by the Southern Oscillation Index (SOI) defined as the surface pressure measured at Tahiti (18°S, 149°W) minus that measured at Darwin (12°S, 131°E). The SOI varies by about 4 mb between extremes of the SO.

The variations in the atmosphere during the SO are accompanied by profound changes in the ocean. Figure 12 shows the ocean surface temperature in November 1982 (near the low SOI extreme) and in November 1983 (near the high SOI extreme). The eastern and central Pacific surface waters are as much as 4 °C warmer at the low SOI extreme. The warm (low SOI) phase of the Southern Oscillation has received particular study. The warming along the west coast of Southern America (and related changes in ocean currents and fish populations) have been known to fishermen for centuries. The local people called this phenomenon *El Niño,* and in the last decade the term *El Niño/Southern Oscillation* (*ENSO*) *event* has become standard in the scientific literature. The ENSO events have a strong phase locking with the annual cycle, so that the peak warming in the eastern Pacific normally occurs in the December–February period. A detailed description of the typical evolution of oceanic and atmospheric con-

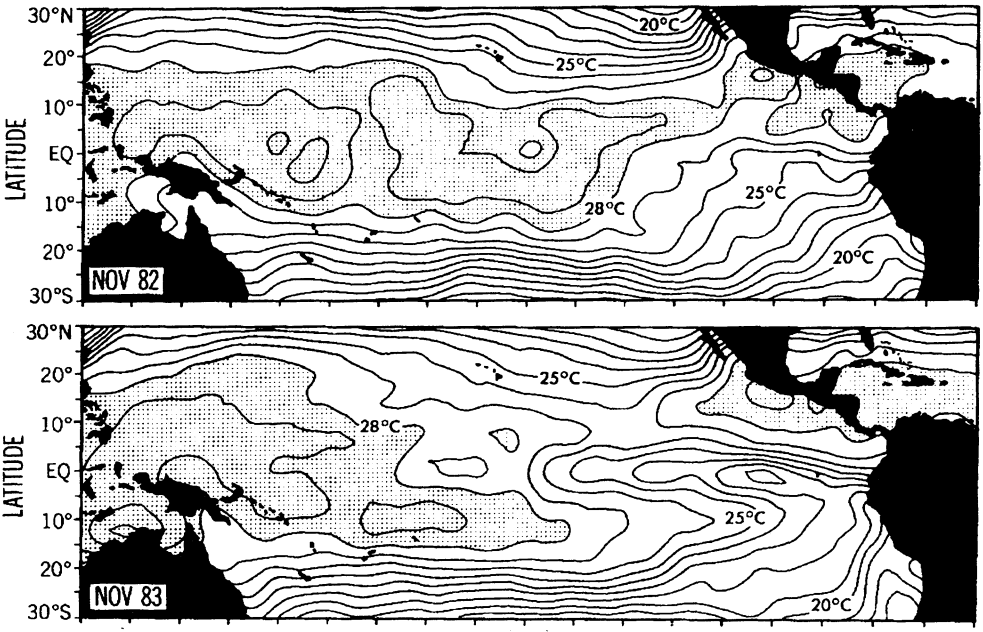

FIG. 12. Ocean surface temperature in the tropical Pacific measured at two extremes of the Southern Oscillation: (top) November 1982 and (bottom) November 1983. Regions where the ocean temperature exceeds 28 °C are shaded. From Philander (1991).

ditions during an ENSO event is given by Rasmusson and Carpenter (1982).

Some aspects of the dynamics of the SO are now well understood, notably the manner in which the atmosphere and ocean interact to excite the extreme warm and cold phases. The warming of the ocean surface in the eastern Pacific during ENSO heats the overlying atmosphere, generating upward motion and consequent low-level convergence of air. This convergence is accomplished by the anomalous westerly winds along the equatorial Pacific seen during El Niño (and a strengthening of the low-level easterly winds over the South American continent). The wind changes affect the stresses acting on the ocean surface. It turns out that westerly wind stresses near the equator inhibit the upwelling of cold deep ocean water, thus reinforcing the ocean surface warming. Analogous arguments can account for the excitation of the cold extreme of the SO. The mechanism by which the atmosphere-ocean system switches from one phase to the other is still not completely clear, however. This presumably involves the internal dynamics of the ocean in the Pacific basin (and this oceanic circulation must largely determine the dominant 2–6-yr time scale for the SO). The SO and the ENSO phenomenon are reviewed by Philander (1991).

While primarily a tropical phenomenon, the SO has detectable effects on the climate in the extratropics as well (e.g., Horel and Wallace, 1981). The clearest signal is found in Northern Hemisphere winter over the eastern North Pacific and North America. Figure 13 presents a schematic picture of the perturbation to the seasonal mean circulation of the troposphere induced at the peak warming of an ENSO event in the tropical Pacific. The effect is a wave train of alternating high- and low-pressure regions propagating along a nearly great-circle trajectory. This basically reinforces the long-term mean standing-wave pattern, most notably in the intensification of the low-pressure region seen in winter over the eastern North Pacific (see Fig. 6). SO effects can account for as much as 25% of the variance seen in the interannual time series of winter-mean surface pressure at some midlatitude locations.

4.2 Interdecadal Variability

Figure 14 shows observational estimates of the time series of yearly-mean surface air temperature averaged over the Northern and

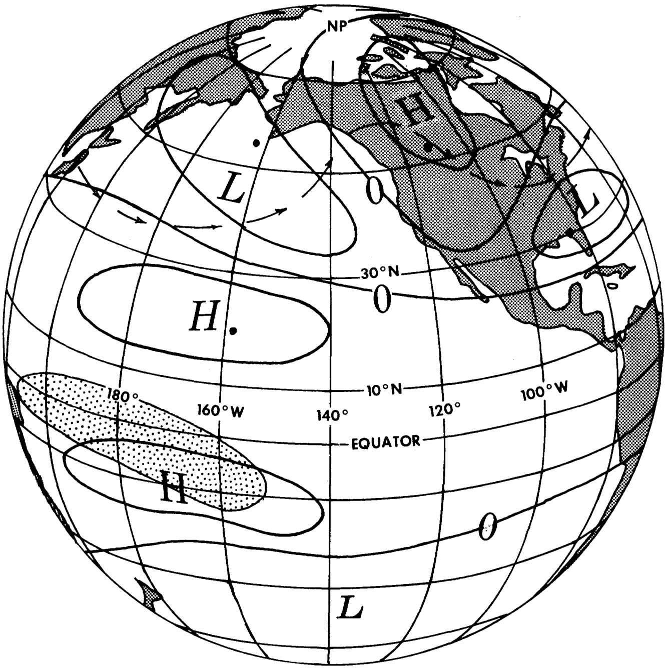

FIG. 13. Schematic diagram showing the perturbation to the usual seasonal mean pressure patterns forced by the ocean warming associated with an ENSO event. The shaded region in the equatorial central Pacific shows the region of most anomalous precipitation during ENSO. The alternating pattern of high- and low-pressure regions extending into the Northern Hemisphere extratropics describes the observed effect at all levels in the troposphere. From Horel and Wallace (1981), courtesy of the American Meteorological Society.

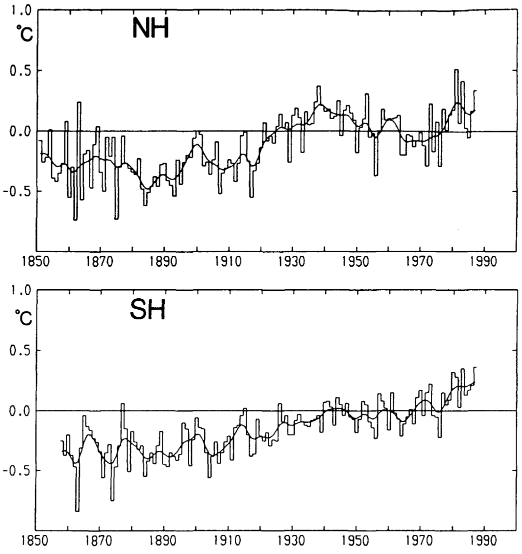

FIG. 14. Time series of yearly mean surface air temperature averaged over (top) the Northern Hemisphere and (bottom) the Southern Hemisphere. In each case, only departures from the long-term mean are shown. The smooth curves are Gaussian-weighted ten-year running means. From Peixoto and Oort (1992).

Southern Hemispheres (the long-term mean in each case has been removed). See Jones *et al.* (1986a, b) for a discussion of the issues involved in the construction of such a time series. This figure reveals a great deal of noisy year-to-year variability (some may be connected with the Southern Oscillation and some with volcanic activity, and some may simply represent sampling error). In addition to this, however, there seems to be significant longer-period variability. The long-term warming trend seen in both hemispheres may be due to anthropogenic modification of atmospheric composition (see Sec. 4.4 below), but the variations on scales of a few decades are thought to be natural fluctuations of the coupled atmosphere-ocean system. Recent research suggests that this may be largely connected with the rate of overturning of the waters in the North Atlantic Ocean. Periods of high overturning rate are times when more cold deep ocean water rises to the surface, leading to colder global air temperature. During periods of weak oceanic overturning, the atmosphere is more insulated from the cold deep ocean, resulting in warmer atmospheric conditions. Since much of the water in the global deep ocean originates from sinking surface waters north of Iceland, this variability appears most strongly in the high-latitude North Atlantic. Levitus (1989) presents detailed observations of substantial variability in oceanic conditions in the North Atlantic during two periods separated by 15 years.

In a quite recent development, rather realistic interdecadal variability has been found in a very long simulation performed with a coupled atmosphere-ocean GCM (Delworth *et al.*, 1993). The modeling of interdecadal climate variations is now an area of extremely active research.

4.3 Anthropogenic Effects

Anthropogenic increases in the IR opacity of the atmosphere will result in increased radiative forcing of the surface. Modeling the effects of this increased radiative driving on the actual global and regional climate is now a major activity in the environmental sciences. The most important driver for this anthropogenic climate change is the increase of CO_2 concentration resulting from burning of fossil fuels and deforestation (particularly in the tropical rain forests). Atmospheric CO_2 concentration has been measured since the 19th century and has been monitored very carefully since the 1950s. Preindustrial values are estimated to be ~280 ppmv (parts per million by volume), while present-day values are close to 360 ppmv and are increasing at almost 2 ppmv per year. Without significant governmental regulation, the CO_2 increase is likely to proceed at an accelerating rate, and doubling of the preindustrial concentrations could well occur before the end of the 21st century. This likely scenario has led modelers to consider a standard benchmark experiment which involves simulating climate with something close to the preindustrial value and then repeating the simulation after doubling the CO_2 concentration. Recent research has shown that trends in other trace constituents (e.g., CH_4, N_2O, and chlorofluorocarbons) may also significantly affect the IR opacity of the atmosphere. It has been estimated that anthropogenic changes in these trace constituents have contributed about half as much as the CO_2 rise itself to the increase in radiative forcing of surface climate since preindustrial times.

The earliest useful predictions of anthropogenic climate change were made with a simple one-dimensional radiative-convective model by Manabe and Wetherald (1967). They found that the global-mean surface temperature should increase by about 3 °C in response to a doubling of CO_2. Later work has focused on using comprehensive GCM's for such studies. A number of different general circulation models have been applied to the problem of calculating the "equilibrium" response of climate to doubling CO_2. In these experiments, the upper layer of the ocean is included solely as a heat reservoir, and the preindustrial and double CO_2 cases are both integrated for a period sufficiently long to obtain stable climate statistics. The report IPCC (1990) summarizes 22 such calculations from groups all over the world. All predict a global-mean warming of the surface of between 2 and 5.5 °C (with most predictions between 3 and 4.5 °C). There is also a consistent prediction that the warming will be most intense in the high-latitude regions. This is largely the result of the feedback of snow cover (and sea ice) with temperature. As the climate warms, the snow line (or ice pack) is pushed poleward, reducing the reflectivity of the surface. Beyond this basic feature, however, there remains considerable disagreement among models concerning the predicted regional patterns of climate change.

Some recent studies have used coupled atmosphere–ocean GCM's in an attempt to model the transient evolution of climate expected over the next century. Such studies typically begin from preindustrial conditions and then slowly increase the CO_2 in a ~100–200-yr integration. The results from these experiments differ considerably from those obtained in the equilibrium experiments. When the full effect of the ocean is included, the predicted surface climate warming is considerably slower, since the ocean acts as a huge heat reservoir. Changes in ocean circulation induced by the changes in atmospheric conditions also play an important role in the coupled-model simulations. In particular, the ocean response tends to reduce greatly the warming over the next century predicted for the high-latitude Southern Hemisphere (IPCC, 1990).

The warming found in even the least sensitive doubled-CO_2 experiments would be easily detectable. More troublesome is the question of whether the long-term global warming trends actually observed over the last century (see Fig. 14) can be attributed to anthropogenic effects. At least one reputable research group now regards the observed temperature record as confirming model predictions of anthropogenic warming (Hansen *et al.*, 1991), but at present there is no consensus on this issue.

5. DYNAMICS OF WEATHER

5.1 Baroclinic Instability and the Formation of Extratropical Cyclones

The day-to-day weather fluctuations in midlatitudes are largely associated with the west-to-east passage of cyclones (low-pressure centers) and anticyclones (high-pressure centers). The scale and intensity of these systems vary considerably, but as a general rule one can identify a train of roughly six cyclones alternating with anticyclones circling the midlatitude zone in each hemisphere, resulting in a dominant scale for a cyclone/anticyclone pair of roughly 5000 km. The mean speed of propagation is $\sim 10–15 \text{ m s}^{-1}$, suggesting a period of ~5 days for the passage of the cyclone and anticyclone. Power spectra computed from observed tropospheric pressure in midlatitudes generally show a broad peak centered near periods of 5 days. The actual day-to-day evolution of the weather is complicated by the simultaneous propagation, development, or decay of the cyclones and anticyclones. Growth and decay of individual systems generally occur on time scales of a few days.

From a practical standpoint, the low-pressure systems merit more attention than the high-pressure systems, since cyclones generally develop more intensely and are associated with more severe weather than anticyclones. A schematic picture of the structure of a developing cyclone is shown in Fig. 15. At midtropospheric levels, the cyclone is manifested as an equatorward dip (sometimes called a *trough*) in the prevailing westerly wind. Near the ground, the perturbation to the zonal mean is greater, and closed height contours around the central low pressure are observed. There is also a westward tilt of the cyclone with altitude; i.e., the midtropospheric trough is located west of the surface low. This phase tilt is a very char-

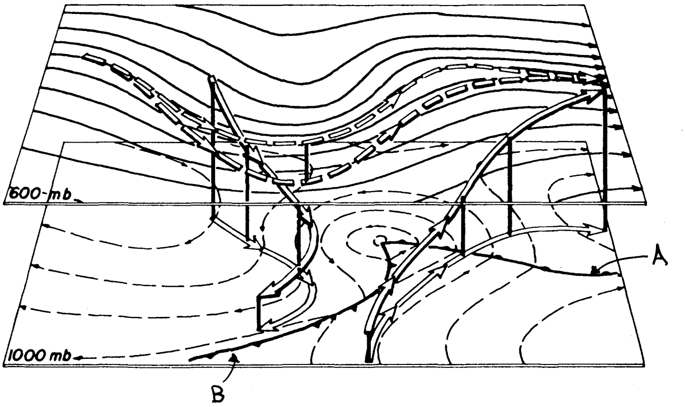

FIG. 15. Schematic depiction of a developing midlatitude cyclone. In the Northern Hemisphere, the storm is being viewed from the south (so that east is to the viewer's right). The dashed curves on the lower plane represent the 1000-mb heights (very nearly proportional to sea-level pressure), while the solid curves on the upper plane are contours of the heights of the 600-mb surface. The continuous heavy arrows represent the three-dimensional trajectories of air parcels, one originating at midtroposphere north of the storm center and one originating near the ground south of the storm center (the same heavy arrows are also used to depict the projection of the trajectories onto the 1000-mb surface). The dashed heavy arrows show trajectories of parcels originating at a midtroposphere level west of the storm center. The solid curves on the lower plane marked "A" and "B" are the warm and cold fronts associated with the cyclone. From Palmen and Newton (1969).

acteristic feature of intensifying weather systems in midlatitudes, and it is easy to show that the westward shift with increasing height implies that the net heat transport by the geostrophic wind will be poleward. The air parcel motions depicted in Fig. 15 illustrate the actual details of heat transport. Air parcels in the cold air north and west of the cyclone center are brought downward and southward, while air parcels in the warm southern sector travel upward and northward. Since small-scale moist convection and precipitation are normally set off by rising motion (leading to adiabatic cooling and then condensation of water vapor), the worst weather normally occurs during the passage of the warm sector (followed by fair weather in the cold sector).

The dynamics of the day-to-day weather systems in the midlatitudes can be largely understood as a manifestation of the phenomenon of *baroclinic instability.* Looking once more at Fig. 2(a), one might imagine that the release of APE would be a straightforward matter. In a small-scale laboratory experiment, one would simply expect the cold (low-θ) air to flood down from the left and displace the high-θ air near the bottom in a simple sloshing motion. On the rotating earth, however, the pressure gradients associated with the meridional temperature gradients in Fig. 2(a) can simply be balanced by the Coriolis force on the zonal component of the wind, and the situation depicted in this figure can be a stationary state of the system. However, for realistic values of the temperature gradient, the situation in Fig. 2(a) is indeed stationary but also unstable to perturbations with wavelike structures in the zonal direction. Thus, the release of APE (and poleward net heat transport) occurs in the cyclones and anticyclones that develop.

The mathematical theory of baroclinic instability was first presented by Charney (1947). The usual approach is to linearize the governing equations about a zonal-mean state. Then the normal-mode solutions (each with a single zonal wave number) are computed. Some of these normal modes will represent neutral traveling-wave solutions, while others will be associated with exponential growth (as well as zonal propagation). For realistic meridional temperature gradients (and hence realistic mean zonal flow shear), the linear theory predicts that the most unstable modes should have zonal wavelengths of several thousand kilometers and have propagation speeds ~ 10 m s^{-1}. The *e*-folding times for growth are typically ~ 2 days. The theoretical results have other realistic features, such as the westward phase tilt with height and the intensification of the pressure perturbations near the ground. The predicted growth rates increase with increasing meridional temperature gradient, and so it is easy to explain why large-scale cyclones and anticyclones are more intense in winter than summer, and why the storm tracks (see Sec. 2.4) tend to occur in specific geographic areas where the shear in the time-mean zonal wind is strongest.

The linear theory of baroclinic instability provides a direct explanation only for the initial growth phase of midlatitude weather disturbances. More complicated nonlinear theories have been advanced to model the entire life cycle of cyclones and anticyclones (see Pedlosky, 1979, for a review).

5.2 Fronts

One very characteristic aspect of the weather in the extratropics is the concentration of gradients in very narrow regions called *fronts.* The passage of a very intense front over some locality can result in temperature changes of more than 20 °C and a complete shift in wind direction in less than an hour. The typical situation is depicted in Fig. 15. The developing cyclone produces a poleward penetration of a wedge of warm air. The temperature gradients at the edges of this wedge intensify into the two fronts. The zone marked A in Fig. 15 leads the warm air as the cyclone moves eastward and is called a warm front, while the one marked B is a cold front. Typically, the cold front will move faster than the warm front. If the cold front actually overtakes the warm front, an *occlusion* is said to occur. An occluded cyclone obviously has little APE to feed on, and thus occlusion is a sign that the winds in a cyclone are about to diminish.

The basic reason for the prominence of fronts in the atmospheric flow may be understood as follows. A blob of passive tracer in any complicated flow tends to get stretched into long filaments by the action of the deformation component of the flow field. Dynamically active quantities such as the temperature may be more resistant to this process. However, in a rotating system like the atmosphere, it is easy to sustain a long narrow zone of temperature gradient. The temperature gradients produce pressure gradients normal to the front, but these pressure forces are balanced by the Coriolis force acting on the wind component along the front. Thus, the pressure gradient provides no resistance to further stretching in the along-front direction. The result is a narrow zone of strong cross-front temperature gradient accompanied by strong vertical shear in the along-front component of the wind. The classical mathematical theory for the formation of atmospheric fronts was presented by Hoskins and Bretherton (1972).

The concentration of horizontal temperature gradients in frontal regions makes them potential sites for development of small-scale baroclinically unstable waves. A theory for the formation of frontal cyclones is discussed by Moore and Peltier (1987).

5.3 Tropical Cyclones and Hurricanes

Meteorology in the tropics differs from that in the extratropics for three principal reasons. First, the factor $2\Omega \sin\phi$ becomes small, thus rendering the geostrophic approximation less accurate. The meridional gradient of radiative forcing decreases rapidly near the equator (see Fig. 11), and thus baroclinic instability is a much less important excitation for the eddy motions in the tropics than in the extratropics. Since the saturation concentration of water vapor rises very rapidly with temperature, the importance of latent-heat release associated with phase changes of atmospheric water is a much more important aspect of the dynamics of the tropics than of the colder, higher-latitude atmosphere.

These differences in the underlying meteorology affect the day-to-day weather systems seen in the tropics. The most frequently observed organized weather systems at low latitudes are the *tropical easterly waves,* weak (~1–2 mb) low-pressure systems with zonal scales ~2000–3000 km, which propagate in the direction of the easterly flow that prevails at low levels (see Fig. 6). In the Atlantic, such waves typically originate over Africa and intensify as they propagate across the ocean in ~10 days. These waves are invariably accompanied by regions of enhanced rainfall. The main energy source for tropical easterly waves during their oceanic propagation is the latent-heat release associated with the precipitation. Detailed observations of the structure of Atlantic tropical easterly waves is given by Reed *et al.* (1977). There are analogous disturbances over the Pacific Ocean.

Some waves in the tropical ocean regions develop further into tropical cyclones. These are storms in which the low-pressure center is strong enough to form an isolated region of closed contours in surface isobaric maps. When the surface winds in a tropical cyclone exceed 33 m s^{-1}, it is referred to as a *hurricane* (sometimes called a *typhoon* if it occurs in the Pacific region). Hurricanes differ greatly in size and intensity (radii from 100 to 1000 km and maximum pressure perturbations from 20 to 100 mb), but all display a rough circular symmetry.

The development of tropical cyclones and hurricanes can be thought of as a kind of instability. The surface winds spiraling into the low pressure bring moist air, which rises near the center, releasing the latent heat that ultimately drives this storm-scale convection cell. The hurricane is then a heat engine with a source of fuel that increases as the engine works faster. The explosive growth phase of a hurricane can last from 12 hours to a couple of days. The hurricane then generally persists for a few days until it reaches land or heads northward over colder water. Either of these contingencies leads to a reduction in the latent-heat energy source for the storm, and the hurricane will then generally dissipate in one or two days.

The instability process that leads to hurricane formation depends on having sufficiently warm ocean surface waters. This imposes both geographical and seasonal constraints on hurricane formation. Almost all North Atlantic hurricanes and strong tropical cyclones form in the late summer or early fall, when ocean surface temperatures north of the equator are highest. In addition, almost all Atlantic hurricane formation is restricted to the western half of the Atlantic basin where the ocean is warmest. Over the Pacific Ocean, there are also specific seasons and geographical regions where hurricane formation is favored. Recently, there has been some speculation that one consequence of enhanced greenhouse warming of climate will be an increase in the number or intensity of severe tropical cyclones.

GLOSSARY

Anthropogenic: Induced by the activities of man.

Convection: Used by meteorologists to denote overturning fluid motions (this use differs from some other branches of fluid mechanics).

Greenhouse Gases: Trace constituents that contribute to the IR opacity of the atmosphere.

Greenhouse Effect: Blanketing effect of the downward emission of IR radiation from the atmosphere, leading to a warming of the earth's surface climate.

Radiative Equilibrium: A hypothetical state in which the net radiative heating at each point in the atmosphere (and the earth's surface) is identically zero.

Spectral Method: A numerical technique for the approximate solution of time-dependent differential equations in which the dependent variables are written as finite sums over a set of spatial basis functions (e.g., spherical harmonics in a global model).

Stratosphere: The layer of the atmosphere above the troposphere, characterized by increasing temperature with height.

Trace Constituent: Chemical species present in the atmosphere in very small concentrations.

Tropopause: The level between the troposphere and stratosphere.

Troposphere: The lowest layer of the atmosphere, characterized by decreasing temperatures with height.

Wind Shear: Spatial gradient of the wind.

Works Cited

Charney, J. G. (1947), *J. Meteorol.* **4**, 135–162.

Delworth, T., Manabe, S., Stouffer, R. (1993), *J. Climate* **6**, 1993–2011.

Gordon, C. T., Stern, W. F. (1982), *Mon. Weath. Rev.* **110**, 625–644.

Hansen, J., Johnson, D., Lacis, A., Lebedeff, A., Rind, D., Russel, G. (1991), *Science* **213**, 957–966.

Holloway, J. L., Manabe, S. (1971), *Mon. Weath. Rev.* **99**, 335–370.

Horel, J. D., Wallace, J. M. (1981), *Mon. Weath. Rev.* **109**, 813–829.

Hoskins, B. J., Bretherton, F. P. (1972), *J. Atmos. Sci.* **29**, 11–37.

IPCC (1990), *Climate Change—The IPCC Scientific Assessment*, Intergovernmental Panel on Climate Change, Cambridge, U.K.: Cambridge University Press.

Jones, P. D., Raper, S. C. B., Bradley, R. S., Diaz, H. F., Kelly, P. M., Wigley, T. M. L. (1986a), *J. Climate Appl. Meteorol.* **25**, 161–179.

Jones, P. D., Raper, S. C. B., Bradley, R. S., Diaz, H. F., Kelly, P. M., Wigley, T. M. L. (1986b), *J. Climate Appl. Meteorol.* **25**, 1213–1230.

Lau, N.-C. (1988), *J. Atmos. Sci.* **45**, 2718–2743.

Levitus, S. (1989), *J. Geophys. Res.* **94**, 16125–16131.

Liou, K. N. (1980), *Atmospheric Radiation,* New York: Academic Press.

Lorenz, E. N. (1969), *J. Atmos. Sci.* **26**, 636–646.

Manabe, S., Wetherald, R. T. (1967), *J. Atmos. Sci.* **24**, 241–259.

Moore, G. W. K., Peltier, W. R. (1987), *J. Atmos. Sci.* **44**, 384–409.

Palmen, E., Newton, C. W. (1969), *Atmospheric Circulation Systems,* New York: Academic Press.

Pedlosky, J. (1979), *Geophysical Fluid Dynamics,* New York: Springer-Verlag.

Peixoto, J. P., Oort, A. H. (1992), *Physics of Climate,* New York: American Institute of Physics.

Philander, S. G. H. (1991), *El Niño, La Niña, and the Southern Oscillation,* New York: Academic Press.

Phillips, N. A. (1963), *Rev. Geophys.* **1**, 123–176.

Ramanathan, V., Coakley, J. D. (1978), *Rev. Geophys. Space Phys.* **16**, 465–489.

Rasmusson, E. M., Carpenter, T. (1982), *Mon. Wealth. Rev.* **110**, 354–384.

Reed, R. J., Norquist, D. C., Recker, E. E. (1977), *Mon. Weath. Rev.* **105**, 317–333.

Wallace, J. M., Hobbs, P. V. (1977), *Atmospheric Science,* New York: Academic Press.

Further Reading

Andrews, D., Holton, J. R., Leovy, C. B. (1987), *Middle Atmosphere Dynamics,* New York: Academic Press.

Houghton, J. T. (1986), *Physics of Atmospheres,* Cambridge, U.K.: Cambridge University Press.

Riehl, H. (1979), *Climate and Weather in the Tropics,* New York: Academic Press.

METROLOGY

WOLFGANG WÖGER AND SIGMAR GERMAN, *Physikalisch-Technische Bundesanstalt, Braunschweig, Germany*

Introduction 239
1. **Basic Concepts in Metrology** 242
1.1 Quantity, Value, and Unit 242
1.2 Measurement 242
1.3 Value of the Measurand 243
1.4 Units and Fundamental Constants 243
2. **Results of Measurement** 244
2.1 Error of Measurement 244
2.2 Uncertainty of Measurement 247
2.3 Accuracy of Measurement 248
3. **General Characteristics of Measuring Instruments** 249
Glossary 250
Works Cited 250
Further Reading 251

INTRODUCTION

Metrology usually is described as the field of knowledge concerned with measurement. However, this rather general description can be slightly misleading, as it does not sufficiently emphasize that any metrological consideration and operation ultimately focus on the notions of the accuracy and uniformity of measurement. In fact, as metrology concentrates on these two aspects of measurement, it does not encompass the bulk of experimental physics, but rather can be viewed as a branch of applied physics. As such, it provides the scientific basis of measurement with special regard to the accuracy of measurement as well as to methods and means to ensure uniformity of measurements. Accuracy and uniformity of measurement contribute to the growth of scientific knowledge, bring about reliability and confidence, support the quality of products, protect the consumer, and add to the capability of economy to successfully compete in the markets. Therefore, metrology is important to all fields of science, technology, and commercial or social transactions.

In practice, metrological tools and operations can be expensive. This is particularly true for measurements on the highest possible level of accuracy such as measurements of the fundamental constants of physics. They usually are performed in national metrology laboratories or academic institutions. But even at somewhat lower levels of accuracy, such as in manufacturing industries, adequate metrology may still cause significant cost. Nonetheless, appropriate on-line measurements of critical parameters of a manufacturing process, acknowledged by many as introducing "non-value-added" expense, can often be shown to be ultimately less expensive than the application of only poor production metrology (see e.g. Kudva and Potter, 1992). In fact, the present level of development of some technologies, such as microelectronics, would not have been possible without paralleling advances in production metrology.

Metrology may be divided into three partially overlapping fields:

1. Scientific metrology, dealing with the theory of systems of units; the theory of measurement; methods of evaluation of results of measurement and uncertainty of measurement (numerical expression of the accuracy of measurement); terminology; realization of the units on the highest possible accuracy level; and very accurate measurements of, for instance, fundamental constants, or measurements to test or compare physical theories.
2. Applied scientific metrology, dealing with all components of practical importance that influence the accuracy and uniformity of measurements. Such components are the closeness between the actual ref-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

erence quantity used in measurement and the defined unit the reference quantity is supposed to represent; the properties of the measuring instruments; and the choice of the measurement procedure. Thus, applied metrology includes the representation of the units or of their multiples and submultiples by standards arranged in steps of uncertainty, the maintenance of the standards, the dissemination of the units by comparison of standards; and calibration of measuring instruments.

3. Legal metrology, dealing with the inspection examination of properties of measuring instruments intended to be used in the public interest. The properties must meet legally specified requirements (mostly in the form of specified maximum permissible errors of the instrument). Regulations are set up, for instance, for instruments used in consumer protection or in medicine, or for instruments monitoring environmental pollution. Legal metrology is not intimately connected to scientific metrology and its direct applications.

Historically, up to the last few decades, metrology was perceived by much of the scientific community as a science of auxiliary and custodial character, not contributing essential elements to scientific progress. This view simply disregards important scientific discoveries by metrologists such as the establishment of the laws of blackbody radiation in 1900. After development of a method to realize accurately the concept of the blackbody, it was the subsequent accurate measurement of its radiation's spectral energy distribution that enabled Planck to initiate quantum physics.

But undoubtedly the tasks of metrology for a long time were predominantly understood as to provide, maintain, and disseminate a set of units—in particular, of length and mass—in order to support uniformity of measurements in commercial and social transactions. But the units of length and mass in former times were only locally defined by arbitrary man-made artifacts, and these were reproduced as accurately as possible to yield reliable secondary material measures valid in the geographical region for which the units were defined. In Europe, for example, this local restriction often made trade difficult even between different cities of the same country, as the cities usually insisted on having their own units of length and mass.

It was only in 1875 that a first step to national and international uniformity was made. In this year, because of the apparent growing need for common standards and accurate measurements in the emerging industrial society, the Convention du Mètre (agreement on the metric system) was signed by 17 countries, including the U.S.A. A decimal system of weights and measures was established and the agreed-on units, the prototypes, of length and mass were decided to be kept in Paris. This event can be viewed as the birth of modern metrology. Additionally, to propagate and further to develop the metric system, the Conférence Générale des Poids et Mesures (CGPM), the Bureau International des Poids et Mesures (BIPM), and the Comité International des Poids et Mesures (CIPM) were founded at that time (for details see UNITS).

In the sequel, national metrological institutions were installed that, on the basis of the above-mentioned agreement, were in charge of scientific metrological research and of the national uniformity of measurement in technology and industry by realizing, maintaining, and disseminating national standards in concordance with the internationally agreed-on units. The first of such institutes were the Physikalisch-Technische Reichsanstalt (now Physikalisch-Technische Bundesanstalt, PTB), founded in 1887 in Germany, the National Physical Laboratory, NPL, founded in 1900 in England, and the National Bureau of Standards (now National Institute of Standards and Technology, NIST), founded in 1901 in the U.S.A.

In the course of time, progress in measurement showed that particularly the accuracy of a length measurement was nearly exclusively determined by the accuracy with which the actual reference quantity in measurement could be compared with the prototype in Paris. For example, whereas the lattice parameter of a crystal could very accurately be measured in terms of a specific x-ray wavelength, on expressing the lattice parameter in the length unit meter the experimental accuracy immediately was lost.

It was only when improved interferometric measurement techniques suggested redefining the meter as a multiple of the wavelength of a characteristic electromagnetic radiation of krypton atoms (^{86}Kr) that the accuracy in length measurements was increased. The multiple was chosen in such a way that for all practical purposes there was no change to the original prototype definition of the meter (concept of the "continuity of the units").

The unit of length with its above new definition was one member of the set of seven base units of the coherent International System of Units (SI) that in 1960 replaced all of the various earlier unit systems (see UNITS). The definitions of these base units, however, should not be viewed as fixed once and for all within the SI. For example, according to the needs of even higher accuracy and taking into account the principle of continuity, a redefinition of the length unit was agreed on in 1983, tracing it to the definition of the unit of time, the second, which can be realized with a higher accuracy than any other of the base units. Also, in 1979, the 1967 definition of the unit of luminous intensity, the candela, was changed (for details, see UNITS).

The least satisfying definition in the SI is that of the unit of mass, the kilogram, which since 1889 is given by an artifact maintained in Paris at the BIPM. Any change with time of the mass of the prototype results in the change of the unit. Great efforts are being made in modern metrology to replace this anachronism by a very accurate determination of the Avogadro constant (see CONSTANTS, FUNDAMENTAL).

However, whereas the increasing sophistication of measuring instruments and techniques allow for improvement of the utmost accuracy of measurement by appropriate definition of the SI units, the corresponding primary or even secondary standards as well as the methods of unit transfer may differ markedly from their prototype predecessors. An example is the standard of the meter, represented by the wavelength of a specific visible line of the He-Ne laser. The line is linked experimentally to the frequency (in the infrared) of the radiation corresponding to a well-defined transition of the cesium atom, i.e., it is derived from the radiation used in defining the second. This standard meets scientific needs and is known to be of a high accuracy, but it is far from the much more practical line scale, which allows for meter comparisons by a simple physical operation. Elaborate interferometry must be used to compare the wavelength with a line scale accurately.

Therefore, quite a large part of metrological work in the national institutions consists of providing a hierarchy of standards of decreasing accuracy. The chain starts from the first transfers from the unit and, as in the case of the line scale, ends in an appropriate physical embodiment of the unit to be used in practice. The hierarchy of standards is of special importance in quality management of measuring equipment and standards. Here, for calibration purposes measurement standards must have the property of "traceability," which means that the accuracy of the standard in use must be traceable to that of a corresponding national or international standard through a documented unbroken chain of comparison (see CALIBRATION AND MAINTENANCE OF TEST AND MEASURING EQUIPMENT).

The basic idea underlying the SI is to define, whenever possible, a unit not by an arbitrary artifact, but by a property of a prescribed specific natural phenomenon that can be realized very accurately at any place and time. The impact of this concept on metrology resulted in a change of its role and scope in science. Modern metrology is acknowledged to be a universal discipline intimately linked to nearly all frontier developments of the physical sciences and of technology.

To treat so vast a topic as metrology and its wide range of practical applications in detail is impossible in a brief space. This article thus concentrates on basic and general metrological aspects that are related to applied physics. For additional information on, for instance, methods and algorithms to evaluate measurements, on the system of fundamental constants, or on calibration and maintenance of standards and measuring instruments, the reader is referred to the corresponding articles in the Encyclopedia.

In the first section, basic terms in metrology are discussed. The important concept of a system of units completely based on fundamental constants of nature is introduced. The second section is devoted to the result and accuracy of measurement. Quantitative measures of the accuracy such as the uncertainty of measurement, as well as other related terms, are described. The article closes with a list of selected basic and general metrological characteristics of measuring instruments.

1. BASIC CONCEPTS IN METROLOGY

In the first two subsections, a series of definitions and statements is presented concerning some of the most important terms related to measurement. Section 1.3 deals with the problem of the value of a measurand defined prior to measurement. In Section 1.4, the concept of completely basing a system of units on the fundamental constants of nature is introduced as a major topic of scientific metrology.

1.1 Quantity, Value, and Unit

The qualitative and quantitative description of an attribute of a physical phenomenon (body, process, state) is accomplished by the introduction of a (physical) "quantity." The magnitude of a quantity is its "value."

To establish an appropriate physical phenomenon under prescribed physical conditions means to establish a definite value of a quantity. Obviously, there are arbitrarily many ways to establish the same value of a quantity. Thus, a quantity cannot be specified by a value but only by a complete description of the physical context fixing the value of the quantity of interest. In reality, a description of a quantity in most cases only approximates the ideal complete description, the approximate description being regarded as sufficient for the purposes under consideration. For example, to talk of the resistance of a given specimen of wire at a specified temperature is only an approximate description of the quantity, as its value is known to depend also, at least, on the existing magnetic field. In metrology, the definition of a quantity plays a role when evaluating experimental information.

A quantity can be compared quantitatively with an arbitrary reference quantity of the same kind in order to express its value relative to that of the reference quantity.

The value of a quantity usually is written as the value of the considered reference quantity multiplied by a number called "numerical value." The value of the reference quantity is the "unit" considered in the comparison. Its numerical value conventionally is accepted to be equal to 1. Depending on the chosen unit, the value of a quantity may be expressed in more than one way.

1.2 Measurement

The set of operations having the object of comparing quantitatively a quantity with a chosen unit (of measurement) is called "measurement." In metrology, a measurement always includes a set of experimental operations but, in general, also consists of theoretical considerations. (There are more general interpretations of the term measurement, where the comparison need not be performed experimentally, and the result may consist of a symbol.)

A quantity subject to measurement is the "measurand." It need not be directly measured. Its value may be derived from those of directly measured quantities by known physical or mathematically defined relationships. The measurement of a measurand always is based on a chosen principle of measurement and a method of measurement, both influencing accuracy and precision of the measurement.

The aim of measurement need not be the value of the measurand expressed by the chosen unit of measurement, but may also be just a statement about whether the value is larger or smaller than some multiple of the unit.

Any result of a comparison need not be observed immediately. It may be automatically processed further in a measuring system.

The evaluation of information obtained from measurement, whether performed by hand or by an integrated computer, is considered to be the final part of measurement.

When the same measurand is measured repeatedly, the measurements should be performed under fixed conditions in such a way that the observed data can be regarded as independent of one another. Otherwise, the evaluation of the obtained data to infer the value of the measurand will prove to be difficult. If additionally the conditions chosen to be the same during repetition allow for the assumption of constant systematic errors (see Sec. 2), the measurement commonly is said to be performed under "repeatability conditions." In this case, the fluctuation of the observed data stems solely from uncontrollable influences of the measuring instrument. Repeatability conditions consist at least of the same observer, the same measuring procedure, the same measuring instrument, and the same influence quantities, which are quantities other than the measurand that affect the result of measurement.

1.3 Value of the Measurand

The measurand is the quantity to be measured. In the simplest case, the measurand may be defined to be the quantity that is directly recorded by the measuring instrument under the actual physical conditions existing at the time and place of measurement. Although these conditions never can be known completely, they uniquely determine the value of the actual measurand. The instrument provides direct but incomplete information about that value.

However, a comparison process always is characterized by an interaction energy between the observed actual measurand and the measuring instrument. Therefore, the value of the measurand observed is not identical with that of the corresponding undisturbed measurand. The latter value must be inferred from knowledge of the disturbing process. It is an essential part of the evaluation of given experimental information to set up a model for the relationship between the value of the actual measurand and that of the undisturbed measurand if, as usual, the latter is the value of interest.

In practice, often, the measurand is specified in advance of measurement by a list of prescribed physical states and conditions. But this description of the measurand may not be complete and, thus, may not fix the measurand's value uniquely. Rather, the given conditions defining the measurand will allow for a set of values consistent with the conditions, where each value out of the set is considered to fit the purpose of measurement. It is meaningless in case of a so-defined measurand to talk of "the" value of the measurand before measurement.

But if the quantity realized for measurement is known to be fully consistent with the (incomplete) definition of the measurand, its value is defined to be the value of the measurand. For example, if the measurand is the resistance of a given wire for any temperature between 20 and 30 °C, then realizing a temperature within this interval will yield a value of the resistance not in contradiction to the defined measurand. The value of the realized quantity can be inferred from the information gained by measurement.

If the definition of the measurand is regarded as complete, i.e., is regarded as fixing the value of the measurand, then, as a rule, the quantity realized for measurement only approximates the measurand. It is the realized quantity that will be directly subject to measurement. To infer the value of the measurand from the gained information, one not only needs a model of measurement of the realized quantity, but also a model of the relationship between measurand and realized quantity.

1.4 Units and Fundamental Constants

Although the choice of the unit of measurement (see Sec. 1.1) in principle is quite arbitrary, to achieve uniformity in measurement, the same unit for all measurements of quantities of the same kind should be used.

A convenient definition of a unit consists of an appropriately chosen phenomenon under prescribed physical conditions such that

1. the value of the quantity of interest, i.e., the unit, does not change in space and time; and
2. any experimental reproduction of the phenomenon and the conditions is possible very accurately at any place and time (also see UNITS).

If the result of a measurement is to be stated in terms of a well-defined unit, the measurement cannot be more accurate than the definition of the unit was realized. Thus, it is a primary metrological task to realize such a definition as accurately as possible and to find ways to disseminate the unit for practical use.

Fundamental constants are of particular interest to modern metrology because of the idea of establishing a system of units entirely based on fundamental constants of nature. The constants to our best knowledge satisfy requirement **1**. Additionally, it is well known that there are some constants like the magnetic flux quantum $h/2e$ (h is Planck constant, e, the elementary charge) that are connected to phenomena satisfying requirement **2**.

In particular, the magnetic flux quantum in the ac Josephson effect (see Constants, Fundamental) relates a voltage to a frequency. Whereas the voltage, and thus the value of the flux quantum, can be reproduced with a relative uncertainty of a few parts in 10^{13}, the value of the flux quantum in SI units is only known with a relative uncertainty that is six orders of magnitude larger. The reason is the poorly known relationship of the as-maintained unit of voltage to the SI volt.

The general idea to improve this situation is to fix the numerical value of the flux quantum to introduce via the ac Josephson effect a new unit of voltage without the deficiencies of the existing SI unit. The procedure could be repeated with other phenomena such as the quantum Hall effect where the von Klitzing constant h/e^2 can be used to establish a new ohm (see Constants, Fundamental). Similarly, the recently discovered single-electron tunneling effects (SET) (Geerligs *et al.*, 1990) could lead via the elementary charge e to a new unit of electric current closing the "electrical triangle" between fundamental constants and the electrical units.

However, because of the demand of continuity of units and the fact that most of the constants are interrelated, the choice of units based on phenomena the theory of which contains fundamental constants cannot be arbitrary. Such a system of units can only be found on the basis of a very accurately known consistent set of experimentally determined values of fundamental constants expressed in the adopted SI units. At present, a set satisfying these needs does not exist. It is a major goal of scientific metrology to improve this situation.

For more details on units and fundamental constants, see Constants, Fundamental and Units. The first of these articles to some extent treats the problem of the unit of mass to be based on the Avogadro constant.

2. RESULTS OF MEASUREMENT

As stressed in the Introduction, the accuracy of measurement is an important element of all metrological considerations. In this section, an overview is given of the metrological concepts related to accuracy.

A measuring instrument used to determine the value of a measurand yields an output from which the "measured value" can be extracted (often the output directly is the measured value). The measured value is equal to the product of some numerical value and the unit of measurement. It must be assumed to deviate unpredictably from the unknown value of the measurand, but it is associated with the measurand as it contains information about the measurand's value.

The same is true for the "uncorrected result of measurement" (see below). It is a function of the measured values obtained from repeated measurements of the measurand under the same conditions. The uncorrected result of measurement serves as a starting point for stating a best estimate of the unknown value of the measurand. It is this best estimate that finally is stated as the "result of measurement."

In establishing the result of measurement, apart from measured values and the uncorrected result of measurement, additional knowledge on the properties of influence quantities and measuring instruments must be taken into consideration. Moreover, if the measurand is to be determined indirectly from measurement of directly measured quantities, known physical relationships between the quantities must be used to find the result of measurement.

In the context of the evaluation of experimental data, the unknown value of the measurand commonly is called the "true value" of the measurand. For a given purpose, it often is convenient to replace the unknown true value of the measurand by a known "conventional true value," regarded as sufficiently close to the true value for the difference to be insignificant for the given purpose. For instance, for the purpose of calibration, an appropriate experimentally determined value of a measurand may be substituted for the true value by convention (see Calibration and Maintenance of Test and Measuring Equipment).

2.1 Error of Measurement

The result of measurement is an estimate of the true value of the measurand and presumably will deviate from it (if it happens not to deviate from the true value, the occurrence of this event is unknown). The deviation of the result x_{res} of measurement from the true value x_{tr} is called "error of (the result of) measurement."

The error of measurement

$$e = x_{\mathrm{res}} - x_{\mathrm{tr}} \tag{1}$$

originates from at least one of the following causes:

1. The actually measured quantity is not exactly equal to the measurand. In principle, this always happens when the measuring instrument is coupled to the measurand. Also, when the measurand is prescribed and has to be realized experimentally, the realization does not exactly agree with the prescribed measurand.
2. The actual unit of measurement is not exactly equal to the unit intended to be used (usually the SI unit).
3. The comparison of the value of the actually measured quantity with the actual unit of measurement is not exact.
4. The values of the actual influence quantities do not agree completely with those the influence quantities are intended to have.
5. The result of measurement is calculated from directly measured quantities on the basis of a relationship that does not correspond to the correct physical relationship between the quantities.

The estimation of the true value from the known data of measurement most often is accomplished by application of the methods of estimation developed in statistics. These methods can be applied to a series of observed indicated values, obtained independently of one another from repeated measurements under the same conditions. This, in particular, means that a new measurement must only be made if the experiment as a whole is in the same state as it was immediately before taking the first measurement of the series.

The basic ingredient for construction of an estimate x_{res} for the unknown true value x_{tr} of a measurand is the measured value. Taking n measurements, it will be denoted by x_i $(i = 1, \ldots, n)$.

Under repeatability conditions, the unknown error $x_i - x_{\mathrm{tr}}$ of the measured value can be regarded as the sum of two components, which differ drastically in their behavior when the measurement is repeated. Whereas the so-called "systematic error" e_s for any i can be assumed to be the same, the "random error" e_{ri} fluctuates at random from measured value to measured value;

$$x_i - x_{\mathrm{tr}} = e_i = e_{ri} + e_s \tag{2}$$

The systematic error mainly is caused by items **1**, **2**, **4**, and **5** listed above, whereas the random error may be traced to causes like the uncontrollable influences of the measuring instruments, uncontrollable changes of the values of the influence quantities, unpredictable changes of the value of the measurand, or varying influences of the observer (e.g., in reading the measured values).

The systematic error is constant for all measured values of a measurement series of size n, provided the chosen set of experimental conditions can be assumed to be constant during repetition. It is only in this sense that the error is "systematic," because this constant error occurs at random immediately before the first measured value is observed. The systematic error is not observable and cannot be extracted from the measured values of the measurement series. If error source **5** above has to be taken into account, it introduces an unobservable systematic error, which is not of a random character. Usually two limits enclosing that error can be stated. Then, on the basis of this knowledge, each of the values between the limits is assumed to have the same probability to be equal to the actual systematic error.

In metrology, the random error is understood to satisfy the demand that it averages out, when one takes the arithmetic mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3}$$

of a very large number (ideally, an infinite number) of measured values obtained under the same conditions. This demand is an essential part of the model of the measured value [Eq. (2)].

The scatter of the observed measured values within a measurement series of size n usually is characterized by the standard deviation

$$s = \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2}, \tag{4}$$

which is independent of the systematic error in a series of measurements under repeatability conditions.

The standard deviation of the measured values is a numerical measure of the closeness of the agreement between the measured values and, thus, of the "precision" of the measurement performed. The use of the qualitative term precision for "accuracy" should be avoided, as the latter term is understood as the closeness of agreement between the result of measurement and the true value of the measurand.

The constant value approached by $\bar{x}$ in the limit $n \to \infty$ is the expectation value of the measured values, denoted by μ. (The probability of $\bar{x}$ deviating from μ vanishes for $n \to \infty$). The expectation value in principle is observable but in practice always is unknown. It can be estimated by $\bar{x}$ with finite n. The arithmetic mean $\bar{x}$ is the uncorrected result of measurement. It is an estimate of the expectation value.

The expectation value of the measured value is equal to the true value if there is no systematic error. For, inserting Eq. (2) into Eq. (3), letting $n \to \infty$, and using the fact that the average random error in this limit approaches zero, i.e., the expectation value of the random error is zero, yields

$$\mu = x_{\mathrm{tr}} + e_s. \qquad (5)$$

In Eq. (5), it is assumed that for all measured values the systematic error is the same, i.e., it is possible to completely control the set of constant experimental conditions even in an infinite measurement series.

Expressed in terms of μ, it follows that

$$e_s = \mu - x_{\mathrm{tr}}. \qquad (6)$$

Using this in Eq. (2), one finds the random error

$$e_{ri} = x_i - \mu. \qquad (7)$$

The measured values x_i $(i = 1, \ldots, n)$ are grouped around their expectation value in such a way that the average deviation from μ approaches zero for a large number of measured values.

The random error e_{ri} $(i = 1, \ldots, n)$ is unknown. If something is known at all about the error e_i [Eq. (2)], this can only refer to e_s. Apart from exceptional cases, the systematic error is not known completely. Therefore, it commonly is decomposed into an estimate of e_s, called the "known systematic error" e_{sk}, and the "unknown systematic error" e_{su};

$$e_s = e_{\mathrm{sk}} + e_{\mathrm{su}}. \qquad (8)$$

In practice, the knowledge of the systematic error often merely consists of a set of conceivable systematic errors enclosed between two limits a and b $(a < b)$. It is assumed that a value outside this interval will not occur in the measurement. As an estimate of e_s, the known systematic error e_{sk} is chosen as the average of all conceivable systematic errors for the measurement under consideration. Therefore, it is an element of that set and mostly is situated in the middle of the interval. The unknown systematic error e_{su} is the unknown deviation of the actual systematic error e_s from e_{sk}. As e_s can be any of the conceivable errors, the average of e_{su} vanishes.

The expression

$$x_i^c = x_i - e_{\mathrm{sk}} \qquad (9)$$

is the "corrected measured value." The value $K = -e_{\mathrm{sk}}$, which compensates for the known systematic error when algebraically added to the measured value, is called the "correction."

Figure 1 schematically shows the connection of a measured value to the various kinds of errors introduced above.

The corrected measured values are the basis of the statement of the result of measurement. In fact, under repeatability conditions, the result of measurement is

$$x_{\mathrm{res}} = \overline{x^c} = \frac{1}{n}\sum_{i=1}^{n} x_i^c. \qquad (10)$$

It can be taken as the best estimate of x_{tr} on the basis of the knowledge about the measurements performed; x_{res} is a statement about one particular measurement series under the same conditions.

If, for the systematic error, limits a and b were stated inappropriately, e_{sk} is not a valid estimate of e_s. This unknown fact can only be revealed by comparison of x_{res} with one or more other results of measurement for the same measurand.

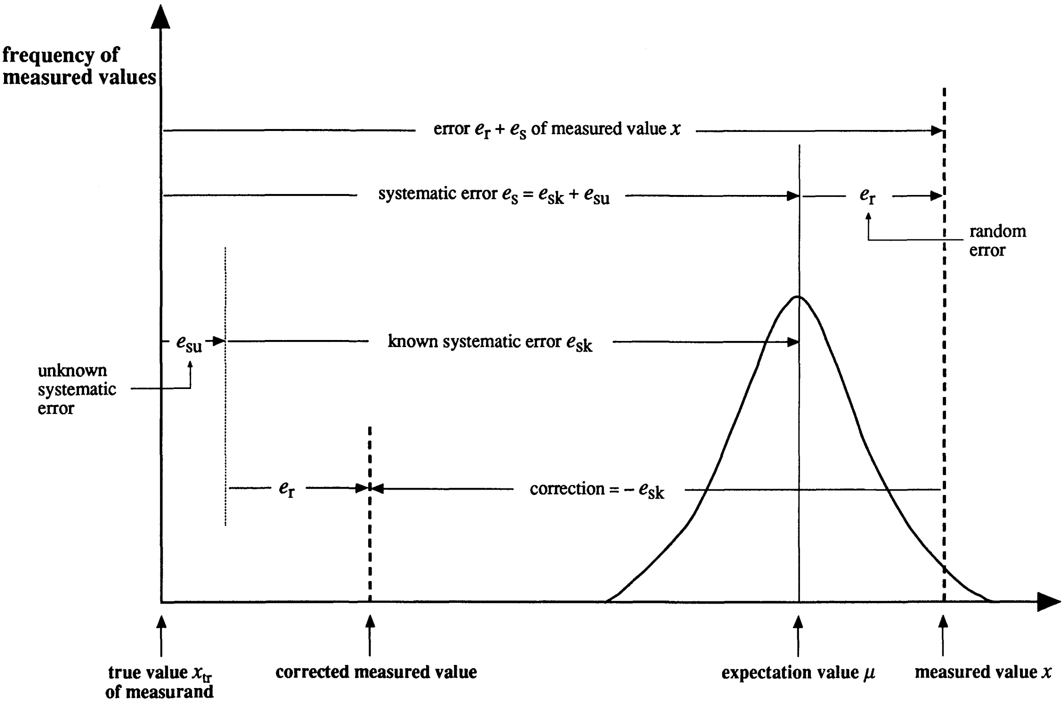
frequency of measured values
error $e_r + e_s$ of measured value x
systematic error $e_s = e_{sk} + e_{su}$
e_r
random error
e_{su}
known systematic error e_{sk}
unknown systematic error
e_r
correction = $-e_{sk}$
true value x_{tr} of measurand
corrected measured value
expectation value μ
measured value x

FIG. 1. Schematic representation of the relations between measured value, corrected measured value, expectation value, and true value of the measurand for measurements under repeatability conditions. According to a distribution, the measured values group around the expectation value, which deviates by the systematic error from the true value of the measurand. The corrected measured value results from addition of the correction to the measured value and deviates from the true value by the sum of random and unknown systematic error. The representation is valid also when one substitutes the uncorrected result of measurement for the measured value and the result of measurement for the corrected measured value (the distribution of uncorrected results of measurement then replaces that of the measured values).

Inserting Eqs. (2) and (8) into the right-hand side of Eq. (10), one has

$$x_{\text{res}} - x_{\text{tr}} = \overline{e_r} + e_{\text{su}} \qquad (11)$$

The error of measurement can be reduced by increasing n, the number of measurements under the same conditions, because then the average random error gets smaller and smaller. But e_{su}, which is constant during repetition of measurement, can only be reduced by reducing $b - a$, the width of the interval containing all conceivable systematic errors.

2.2 Uncertainty of Measurement

After measurement, the knowledge of the value of the measurand has increased compared to the situation before measurement. But this knowledge still is incomplete. There always remains a set of values, consisting of reasonable estimates of the true value of the measurand, so that any of these values can be attributed to the measurand without contradicting the (incomplete) information obtained from the experiment. Thus, the set represents uncertainty in the knowledge of the measurand's value after measurement.

The actual result of measurement is an element of this set. But its statement alone obviously cannot be regarded as a complete description of what one has learned from the measurement. It must be taken into consideration that any of the other values of the set could also have occurred as the result of measurement. The complete description consists of the result of measurement and a numerical characterization of the uncertainty in the knowledge of the true value.

From all that one knows from and about the measurement (directly measured values, influence quantities, properties of measuring instruments, relations between quantities, etc.), a typical such set or interval I_{est} of estimates can be found out of the set of all possible estimates of the true value of the measurand (for methods see STATISTICAL DATA AND ERROR ANALYSIS). It is centered around the actual result x_{res} of measurement and, to repeat, contains those values considered to be reasonably attributable to the measurand on the basis of the measurement.

The spread of the estimates in I_{est} on either side of x_{res} is given by the numerical value of an appropriately chosen parameter. This numerical parameter value is called the "uncertainty of measurement," the "uncertainty of measurand," or simply "uncertainty," and will here be denoted by u.

In stating the uncertainty, one takes account not only of the typical spread of the observations but also of the spread of values of the total systematic error that could reasonably occur in the measurement performed. Both spreads influence the spread of the reasonable estimates. Thus, the uncertainty of measurement decreases for increasing information about the true value.

As the uncertainty of measurement is determined by the knowledge of the properties of the measurement and nothing else, it is to be distinguished from the spread of values characteristic of a confidence interval with given confidence level (see below).

It is meaningless to talk about the "uncertainty of the result of measurement," as this result is one of the few things certainly known after measurement. Additionally, this wording supports the misleading identification of the error of the result of measurement with the uncertainty of measurement. The two are distinct concepts. In fact, although there always is a nonvanishing uncertainty of measurement, the error of the result of measurement unknowingly can be zero.

The interval I_{est} with limits $x_{\text{res}} - u$ and $x_{\text{res}} + u$ cannot be regarded as necessarily containing the true value of the measurand. The degree of confidence or the probability attached to the statement to be correct that the actually quoted I_{est} contains the true value depends on the uncertainty u calculated on the basis of the chosen parameter that measures the spread of estimates around the result of measurement. Consequently, if two such parameters are chosen, one of which on the basis of the same experimental information consistently yields larger uncertainty than the other, the corresponding interval of estimates will be preferred in situations where a high degree of confidence is asked for. Parameters of this kind are recommended for commercial, industrial, regulatory, and other applications (e.g., concerning health and safety).

But in scientific metrology and most of applied scientific metrology the parameter is chosen to yield the uncertainty as the standard deviation of the result of measurement [this standard deviation is not equal to s in Eq. (4)], in which case the uncertainty is called "standard uncertainty." The standard uncertainty does not satisfy the demand of an interval with a high level of confidence.

The main reasons for this choice are that in science one frequently compares different measurements of the same measurand or often uses a previous result of a measurement of some measurand to evaluate the measurement of a different measurand. But if there are differences in two or more measurements of the same quantity, they should not be masked by stating too large an uncertainty. Only if such differences are clearly detectable can one learn how to remove them by improved measurements. Additionally, it is only the standard uncertainty of measurement of a measurand for which an algorithm exists to propagate the uncertainty consistently to compute the final standard uncertainty of measurement of a related measurand.

The algorithm, along with least-squares adjustment, is dealt with in the article STATISTICAL DATA AND ERROR ANALYSIS (also see Martin, 1971; Press *et al.*, 1989). In its most general form, it allows us, for instance, to calculate several measurands of interest and their uncertainties on the basis of their known relationships to a set of other quantities simultaneously measured. This kind of evaluation task gets more and more important because of the increasing trend to complex measuring systems, where integrated computers facilitate the calculations.

2.3 Accuracy of Measurement

The "accuracy of measurement" is the closeness of the agreement between the result of a measurement and the true value of the measurand. It should be distinguished from the precision of measurement, as the accuracy implicitly contains statements on systematic errors. Therefore, a precise measurement can be very far from being accurate.

The uncertainty u is a quantitative measure of the accuracy of the measurement in the sense that smaller uncertainty means higher accuracy. Another measure for the accuracy of measurement is given by the statement of a confidence interval with accompanying confidence level. Here the confidence level $1 - \alpha$ is chosen in advance and numerically expresses the confidence that an interval of estimates to be determined contains the true value. On the basis of the measurement, one then finds a corresponding interval using an assumed distribution of estimates (mostly a normal distribution). For a symmetric distribution, this confidence interval with confidence level $1 - \alpha$ is centered around the corrected mean (result of measurement) of a measurement series of size n under repeatability conditions. Half of its width w, when stated together with confidence level $1 - \alpha$, can be used as a numerical measure of the accuracy of the measurement. For nonsymmetric distributions, the total width of the resulting interval characterizes the accuracy of measurement. For a normal distribution and if the unknown systematic error in the measurement series is negligible, w is given by the product of the standard deviation $s_{\text{res}} = s/\sqrt{n}$ [for s see Eq. (4)] of the result of measurement and the Student factor $t(n - 1,1 - \alpha)$, which depends on n and the arbitrarily chosen α (see, e.g., Martin, 1971).

It must be stressed that w cannot be stated as the uncertainty of measurement, which, by definition, is constructed independent of the confidence in the trueness of the assumption that the true value of the measurand is contained in the stated interval I_{est}. Moreover, there is no rule to propagate w.

The complete final result of a measurement consists of the result of measurement and any of the above-mentioned quantitative statements characterizing the accuracy of measurement.

3. GENERAL CHARACTERISTICS OF MEASURING INSTRUMENTS

The properties of a measuring instrument not only influence the result of measurement, but also the uncertainty of measurement. To describe numerically the properties of a measuring instrument, various metrological characteristics are of interest (ISO/IEC/OIML/BIPM, 1993). Their values depend on the conditions under which the measuring instrument is used. In practice, the properties of instruments often are investigated under rated operating conditions or reference conditions.

This section contains a list of characteristics of measuring instruments that are of general importance to metrology. The list is not complete. In particular, characteristics specific to the various types of instruments such as mass spectrometers may be found in the corresponding articles of the Encyclopedia.

The "response characteristic" is the relationship between an input signal (also called stimulus) to a measuring instrument and the corresponding output signal (also called response) under defined conditions.

The stationary-state properties of a measuring instrument are described by a response characteristic obtained by finding the relationship between different fixed values of the input quantity and the corresponding final steady values of the output quantity. The relationship usually is represented by the characteristic curve.

The dynamical behavior of a measuring instrument is characterized by varying the input signal in time in a specific way (e.g., stepwise variation, sinusoidal variation) and observing the response. The response characteristic, then, often is represented in form of a transfer function (Laplace transform of the response divided by that of the stimulus).

The "sensitivity (coefficient)" of a measuring instrument is understood as the change in its response divided by the corresponding causing change in the stimulus. The value of the coefficient may depend on the actual value of the stimulus.

The slow variation with time of a metrological characteristic of a measuring instrument is called "drift."

The "response time" of a measuring instrument is the time interval between the instant when a stimulus is subjected to a specified abrupt change and the instant when the response reaches and remains within specified limits around its final steady value.

The "error of a measuring instrument" is that contribution to the error of measurement that is caused by the measuring instrument. The two errors, of course, need not be the same. The error of a measuring instrument consists of a random and a systematic part.

In practice, often the "bias error (of a measuring instrument)" is of particular interest. It is an estimated contribution of an indicating measuring instrument to the systematic error of a measurement. The bias error A_s usually is determined as the difference between the uncorrected result of measurement (which may consist of a single value) and the conventional true value x_r of the measurand:

$$A_s = \bar{x} - x_r.$$

This estimate sufficiently describes the contribution of the instrument to the systematic error of measurement, if the absolute value of the random error of the instrument is negligible compared to the absolute value of its systematic error, and if the absolute value of the deviation of x_r from the true value of the measurand is much smaller than $|A_s|$.

The "measuring (or working) range" of a measuring instrument is that set of values of measurands for which the error of the measuring instrument is intended to lie within specified limits.

The "maximum permissible errors of a measuring instrument" are the extreme values of an error permitted by specifications, regulations, etc. for a given measuring instrument.

A class of measuring instruments meeting certain metrological requirements such that errors of the instruments are kept within specified limits is called "accuracy class." The accuracy class characterizes the properties of the instruments with respect to the error of a measuring instrument. But it must be stressed that, whereas the error of measurement obtained by an instrument of a given accuracy class does contain the error of the instrument, it is not characterized by the accuracy class.

GLOSSARY

Calibration: Set of operations that establish, under specified conditions, the relationship between values of quantities indicated by a measuring instrument, or values represented by a material measure, and the corresponding values realized by standards.

Influence Quantity: Quantity that is not the measurand but that affects the result of measurement.

Measuring Instrument: Device intended to be used to make measurements, alone or in conjunction with supplementary device(s).

Measurement Signal: Quantity that represents the measurand and that is functionally related to it.

Method of Measurement: Set of operations in measurement that is independent of the principle of measurement and appropriate to yield the measurand's value (examples: the null method, the substitutional method).

Principle of Measurement: Scientific basis of measurement consisting of a physical phenomenon or effect that uniquely relates the measurand of interest to a quantity to be directly measured. The relationship is known from theory or established empirically (example: the thermoelectric effect applied to measurement of temperature).

Rated Operating Conditions: Conditions of use for which specified metrological characteristics of a measuring instrument are intended to lie within given limits.

Reference Conditions: Conditions of use prescribed for testing the performance of a measuring instrument or for intercomparison of results of measurements.

Relative Uncertainty: Uncertainty divided by the modulus of the result of measurement.

(Measurement) Standard: Measuring instrument or material measure intended to define, realize, conserve, or reproduce a unit or one or more values of a quantity to serve as a reference.

Works Cited

Geerligs, L. J., Anderegg, V. F., Holweg, P. A. M., Mooij, J. E., Pothier, H., Esteve, D., Urbina, C., Devoret, M. H. (1990), *Phys. Rev. Lett.* **64**, 2691–2694.

ISO/IEC/OIML/BIPM (1993), *International Vocabulary of Basic and General Terms in Metrology (VIM)*, Geneva: International Organization for Standardization.

Kudva, S. M., Potter, R. W. (1992), in: M. T. Postek (Ed.), *6th Annual Conference on Integrated Circuit Metrology, Inspection and Process Control*, SPIE Proceedings Vol. 1673, Bellingham, WA: SPIE.

Martin, B. R. (1971), *Statistics for Physicists*, New York: Academic Press.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T. (1989), *Numerical Recipes (The Art of Scientific Computing)*, Cambridge, U.K.: Cambridge University Press.

Further Reading

Petley, B. W. (1985), *The Fundamental Physical Constants and the Frontier of Measurement*, Bristol: Adam Hilger.

Two different points of view concerning the expression of the uncertainty of measurement are taken in the following two papers:

Colclough, A. R. (1987), *J. Res. Nat. Bur. Stand.* **92**, 167–185.

Weise, K., Wöger, W. (1993), *Meas. Sci. Technol.* **4**, 1–11.

MICROELECTRONICS

T. Sugano, *RIKEN, Wako, Saitama, Japan*

A. Morino, *NEC Corporation, Kawasaki, Kanagawa, Japan*

Introduction 254
1. **Components of Integrated Circuits** 255
1.1 Design Targets of Integrated Circuits 255
1.2 Basic Device Structures of Integrated Circuits 255
1.2.1 Transistors 255
1.2.2 Memory Cells 256
1.2.3 Isolation 257
1.3 Chip Yield and Reliability 258
1.3.1 Chip Yield 259
1.3.2 Reliability 259
1.3.2.1 Gate-Oxide-Related Failures 260
1.3.2.2 Hot-Carrier–Induced Failures 260
1.3.2.3 Metallization- and Interconnection-Related Failures 261
1.3.2.4 Electrostatic Discharge (ESD) 261
1.3.2.5 Latchup 262
1.3.2.6 Alpha-Particle–Induced Soft Errors 262
1.4 Circuit Configurations of Integrated Circuits 262
1.4.1 Circuit Configurations for VLSIs and Beyond 262
2. **Selection of Process Technologies for Integrated Circuits** 263
2.1 Lithography 263
2.2 Etching 263
2.3 Metallization 264
2.4 Thin-Film Formation 264
2.5 Common Issues 264
3. **Design Methodology and Computer-Aided Design of Integrated Circuits** 265
3.1 Design Methodology of Integrated Circuits 265
3.2 Test CAD 265
3.3 CAD for Process, Device, and Circuit Design 266
3.3.1 Process Simulation 266
3.3.1.1 Lithography 267
3.3.1.2 Ion Implantation 267
3.3.1.3 Thermal Diffusion 267
3.3.1.4 Oxidation 267
3.3.2 Device Simulation 267
3.3.3 Circuit Simulation 268
3.4 Design Environment and Technology Standardization 268
3.4.1 CFI (CAD Framework Initiative) 268
3.4.2. EDIF (Electronic Design Interchange Format) 268
3.4.3. HDL (Hardware Description Language) 269
3.5 Future Trends 269
4. **Applications of Integrated Circuits to Electronics Systems** 269
4.1 Communication Systems 269
4.2 Computer Systems 271
4.3 Consumer Electronics 272
4.4 Human Interface 273
4.5 Future Trends 273
5. **Future Prospects and Limitations of Microminiaturization** 274
5.1 Future Prospects 274
5.1.1 Expectations of Features 274
5.1.2 Technical Issues and Solutions 274
5.2 Limitations of Microminiaturization 274
5.2.1 Physical and Practical Limitations 275
5.2.1.1 Tunneling Effects 275
5.2.1.2 Hot-Carrier Effects 275
5.2.1.3 Subthreshold Leakage Current (Threshold Voltage Scaling) 275
5.2.1.4 Punchthrough Phenomenon (Short Channel Effect) 276
5.2.1.5 Parasitic Elements 276
5.2.1.6 Power Dissipation 276
5.2.1.7 Fluctuations 276

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

5.2.2 Ultimately Scaled-Down MOSFET and Future Prospect of VLSIs 277
Glossary 278
Works Cited 278
Further Reading 279

INTRODUCTION

The need for and importance of miniaturized circuits were recognized long ago, for applications to a variety of fields, including space development and military systems. Microminiaturization, that is, the development of electronics systems with very small-dimensioned circuits, was carried out in several ways. A typical example was the assembling of microminiaturized components in a very small area on a printed wiring board or a ceramic substrate.

The term "microelectronics" usually refers to the technology of design, manufacture, and use of electronics devices with very small dimensions. An epoch-making event in the history of microelectronics is the invention of the transistor in 1947, by J. Bardeen, W. H. Brattain, and W. Shockley, all with Bell Telephone Laboratories, AT&T. The invention of the integrated circuit (IC) in 1959 by J. Kilby of Texas Instruments, and R. Noyce of Fairchild, independently, is another monumental landmark, leading to the development of a completely new world of microelectronics. Since then, the number of components integrated on a silicon chip has shown a constant increase, as shown in Fig. 1 (*IEEE ISSCC Digest*, 1970–1993; *Digest of Symposium on VLSI Circuits*, 1987–1993). Usually, the terms LSI (Large-Scale Integrated Circuits), VLSI (Very Large-Scale Integrated Circuits), and ULSI (Ultralarge-Scale Integrated Circuits) refer to integrated circuits with numbers of components equal to or larger than 10^3, 10^5, and 10^7, respectively. Based on these situations, microelectronics, in recent years, usually refers to a field of electronics that includes the integrated circuits and the area of applications that has been developed by them.

The introduction of microelectronics has had a tremendous impact on industry and human society. The first impact is the realization of electronic systems with smaller size and lighter weight. As is shown in Fig. 1, the sizes of memory and digital functions have been reduced roughly by 1/100 and 1/30, respectively, for every ten years. The corre-

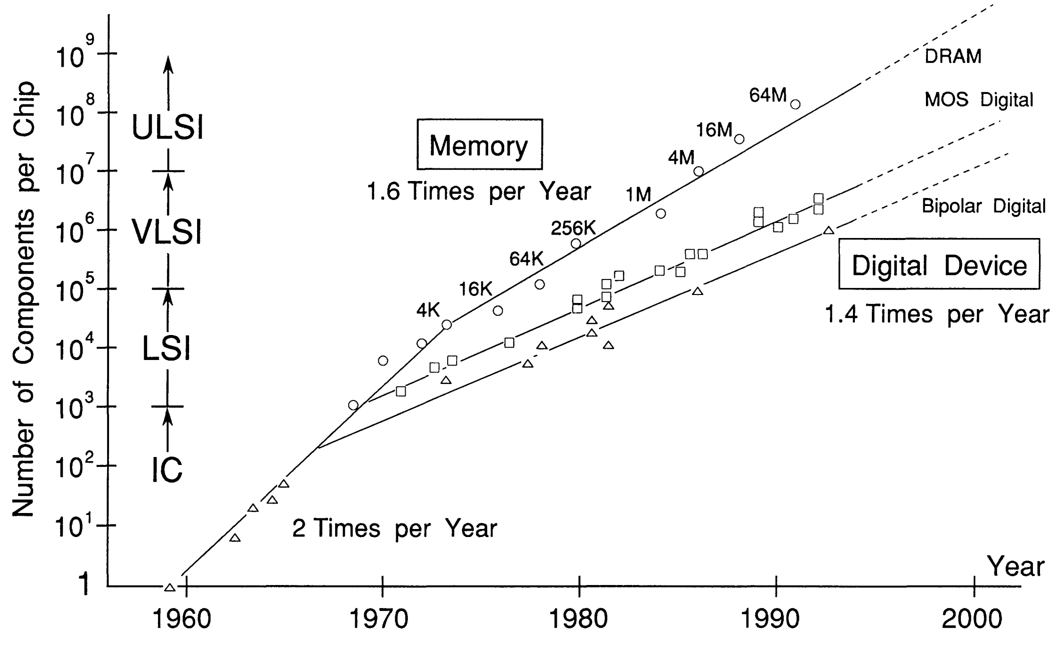

FIG. 1. Trend of the number of components integrated on a chip.

sponding weight reduction is easily understood.

The second impact is cost reduction. It results from both component and interconnections sides, as follows. As is well known, ICs are fabricated by batch processing of wafers containing large numbers of chips. The cost of each component integrated on a chip is reduced as the number of components integrated on a chip increases. Furthermore, the production cost of each chip is reduced as the production volume increases. An experimental rule, called a learning curve, says that the production cost is reduced by roughly 30% every time the accumulated production volume becomes double. Another cost reduction results from making large numbers of interconnections between components on a chip, by batch processing. On the other hand, making interconnections outside chips, on a printed wiring board, for example, requires much more time, and so leads to much higher cost.

Increase in reliability is the third impact of microelectronics. Transistors, as compared with electron tubes, have essentially higher reliability because of not employing a heated electrode. Furthermore, the drastic reduction in the number of soldered contacts leads to the excellent reliability of the electronic systems. Generally speaking, soldered contacts are the least reliable portions in electronic systems. Interconnections between components on a chip, on the other hand, are made by using thin-film metal layers, without using soldered contacts.

The final, but not the least, impact is the performance improvement. Important factors to determine the operational speed of electronic systems are the time required to charge/discharge parasitic capacitances and the time required for the signal to propagate through the signal lines. Microelectronics consisting of components with smaller size and shorter wiring length contributes to the realization of high-speed operation in terms of both of these mentioned factors.

1. COMPONENTS OF INTEGRATED CIRCUITS

1.1 Design Targets of Integrated Circuits

Typical targets of integrated circuits products include the following items:

1. functions;
2. performance, including operational speed;
3. power consumption;
4. price; and
5. timing of introduction into the market.

The importance of item **3** is now increasing, because of the increase in power consumption resulting from the increase in the number of components integrated on a chip, and also because of the growing market of portable/personal terminals.

The design of IC products starts with selecting the most appropriate basic circuit type, such as CMOS (Complementary symmetry Metal-Oxide-Semiconductor), BiCMOS (Bipolar-CMOS), and so on. The CMOS circuit is characteristically of low power consumption, whereas the BiCMOS circuit affords higher switching speed, but at larger power consumption, compared with CMOS (see CIRCUITS, DIGITAL).

Design styles of LSIs can be divided roughly into three types. The first is standard products, such as memory LSIs and microprocessors. The second refers to so-called ASIC (Application-Specific Integrated Circuit), such as gate arrays and cell-based ICs. These devices are used to integrate such functions as are specific to particular systems. Between these two types is the third one, which is called ASSP (Application-Specific Standard Product). This is the device widely used in specific fields, such as telecommunications, networks, computer terminals, and so on., Examples include the CODEC (COder+DECoder; see Sec. 4.1), and the floppy-disk controller. Generally speaking, production volume is the largest for the first type, and the smallest for the second one.

1.2 Basic Device Structures of Integrated Circuits

1.2.1 Transistors Typical examples of the basic device structures used in integrated circuits are briefly described in this section.

Fig. 2(a) shows a typical cross-sectional view of a CMOS configuration, a basic structure of CMOS integrated circuits. Here, the right-hand side corresponds to an n-channel MOS (Metal-Oxide-Semiconductor) transistor. Three parts, that is, the metal, oxide, and semiconductor, correspond to the polycrystalline silicon, silicon dioxide, and p well, or

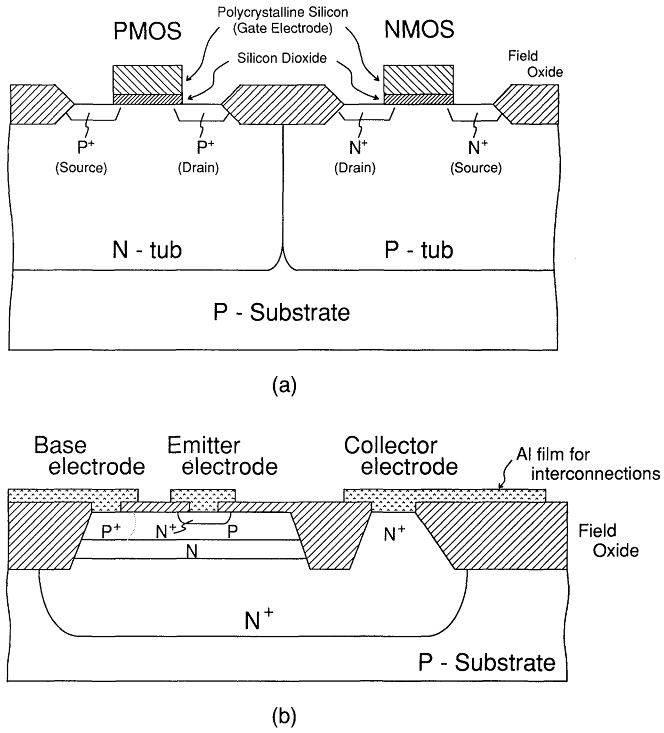

FIG. 2. Cross-sectional view of transistors. (a) CMOS device, (b) bipolar device.

"tub," respectively. The polycrystalline silicon performs as the gate electrode of a MOS transistor. Two *N*+ regions correspond to the source and drain of a transistor. For the left-hand side of the figure, which corresponds to a *p*-channel MOS transistor, a similar structure is formed. Each tub is prepared to provide an isolated island for the respective transistor. Field oxide is formed to assure electrical isolation between adjacent transistors, by eliminating the leakage current through the surface.

Shown in Fig. 2(b) is a typical cross-sectional view of a bipolar transistor. Here, the three electrodes, that is, the emitter, base, and collector, are shown. Three regions beneath the emitter electrode, *N*+, *P*, and *N*, correspond to the emitter, base, and collector regions, respectively, of an *npn* bipolar transistor. The *N*+ region under the collector region and through the collector electrode is formed to reduce the collector series resistance. Similarly, the *P*+ region beneath the base electrode is formed for the reduction of the base series resistance. These *N*+ and *P*+ regions contribute to the realization of high-speed operation of the transistor.

The most favorable feature of the bipolar transistor is its high-speed operation. The circuit type that affords the highest switching speed is ECL (Emitter Coupled Logic; see CIRCUITS, DIGITAL.) Key parameters to determine the operational speed of ECL circuits are cutoff frequency f_T, base resistance r_b, collector parasitic capacitance C_c, and collector series resistance R_c. Approaches to realization of high-performance ECL include the reduction of lateral and vertical dimensions, employment of trench isolation, and/or self-aligned transistor structures. The cutoff frequency of leading-edge transistors in practical applications is around 30 GHz.

Usually, the switching speed of a digital circuit is expressed by the propagation delay time T_{pd}, which is defined as the time interval between the output and the input signals. The smallest values of T_{pd}, for the typical digital circuits under production in large quantity at present, are as follows: ECL, 30 ps/gate; CMOS circuit, 350 ps/gate; BiCMOS circuit, 300 ps/gate.

1.2.2 Memory Cells A memory cell is one of the most significant components of memory integrated circuits. Figure 3(a) shows an equivalent circuit of a memory cell of Dy-

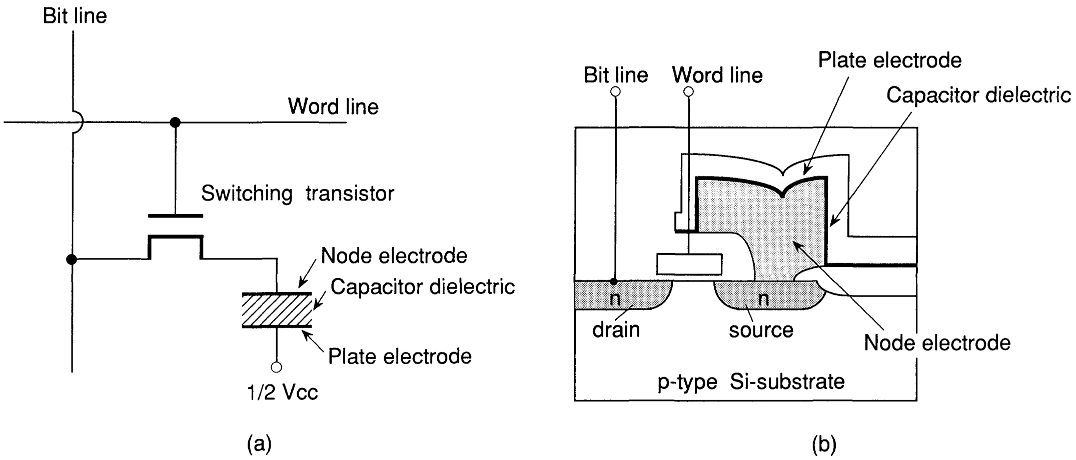

FIG. 3. (a) Equivalent circuit of a DRAM cell. (b) Cross-sectional view of a DRAM cell (stacked capacitor cell).

namic Random-Access Memory (DRAM; see CIRCUITS, DIGITAL). The cell consists of a capacitor for charge storage and a switching transistor for read/write control. The voltage setup on the bit line is written into the capacitor when the word line is made high. In the read phase, the word line is made high, and the signal charge is derived from the capacitor to the bit line.

The key point of a memory-cell design is to realize a cell area as small as possible, in order to develop a memory device with a chip area as small as possible. In fact, the cell area per bit has continuously been reduced, corresponding to the continuous increase in the bit capacity per chip. It is required, on the other hand, that the charge stored in a cell is no smaller than a critical value that is determined from the consideration of (a) immunity to soft errors (see Sec. 1.3.2.6) and (b) the minimum value that can stably be amplified by the sense amplifier. The corresponding critical capacitance is roughly 30 fF. As the reduction of the cell area per bit proceeds, keeping pace with the increase in the bit density, realization of 30-fF capacitance is becoming increasingly difficult. For high-density (4 Mbit and beyond) DRAMs, the following approaches are employed to avoid the decrease in capacitance:

1. employment of three-dimensional cell structures in order to increase the effective area of the electrodes; and
2. the use of materials with higher dielectric constant.

The first example of approach **1** described above is a stacked capacitor cell, as shown in Fig. 3(b). This is one of the most typical cell structures used in high-density DRAMs. In the cell, effective cell area is increased through the use of both the area over the switching transistors and the sidewall of the capacitor as electrodes for the capacitor. Here, the height of the pad is optimized by taking into account the target capacitance and the process complexities. Typical value of the pad height is from 0.2 to 0.5 μm.

Another typical example of approach **1** is a memory cell that employs a trench capacitor. A silicon substrate is etched to form a narrow, deep groove in the silicon, and then specific films are formed. The trenched groove is typically 3–4 μm in depth. Although this structure can realize a relatively large capacitance, the issue is how to decrease the junction leakage current due to the crystal defects induced by the digging of the silicon substrate.

As for approach **2**, silicon nitride, with the dielectric constant of roughly 7.5, is used as a capacitor material. (The dielectric constant of silicon dioxide is around 3.9.) For future DRAMs with higher bit density, applications of other materials with larger dielectric constant, including Ta_2O_5 (the dielectric constant is around 25) and $BaSrTiO_3$ (the dielectric constant is over 100), are being developed.

1.2.3 Isolation Isolation is one of the most important structures of integrated circuits. First, all of the components on a chip are required to be electrically isolated from each other. Leakage current between them should be small enough to meet specifications. Second, the parasitic capacitance associated with the isolation structure should be as small as possible. Third, the area of the isolated region is required to be as small as possible, because it has a big impact on chip size. Here, three typical examples of isolation structures will be reviewed.

The first one is the LOCOS (LOCal Oxidation of Silicon) structure, which is extensively used in almost all MOS integrated circuits. It employs a thick oxide selectively formed in the field regions, as shown in Fig. 2(a). The term field regions means those parts outside the islands in which transistors and various components are formed. The thick oxide is effective in reducing the leakage current between components, thus avoiding the electrical coupling between those components. Selective oxidation of the silicon is carried out by using a silicon nitride layer as an oxidation mask. The field oxide is partially formed under the silicon surface, resulting in a decreased step height of the field oxide. This relatively small step is important to avoid open failures of interconnects.

In order to reduce the leakage current more completely, a channel stopper, a region with high impurity concentration, is formed near or under the field oxide. The high impurity concentration avoids channel formation, thus keeping the electrical isolation more thorough.

On the other hand, as the reduction of the device feature size proceeds, reduction of the area for the isolated region is also necessary. The bird's beak, a characteristic structure formed at the boundary of the field oxide and silicon, causes difficulty for area reduction of the LOCOS structure. A typical approach to eliminate the bird's beak is the employment of a polysilicon-buffered LOCOS structure. By inserting the polysilicon layer under the silicon nitride layer, field oxide with a reduced step height can be formed.

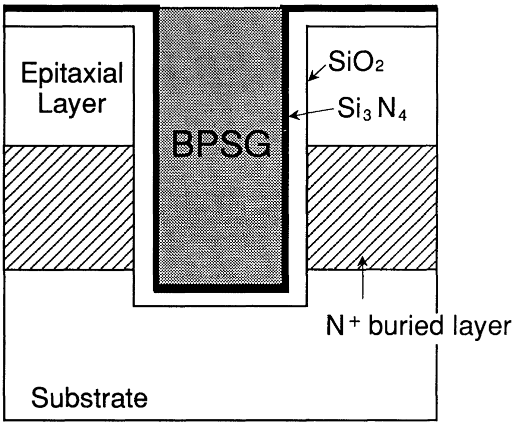

FIG. 4. An example of a cross-sectional view of a trench isolation, filled with BPSG.

The second approach is trench isolation. The structure is formed by etching a narrow groove in silicon and filling it with oxide and/or polysilicon and the like. Figure 4 shows an example of a cross-sectional view of trench isolation. The trench is filled with BPSG (Boro-Phospho-Silicate Glass). It leads to compact isolation with a width of 2 μm or smaller. Generally, the requirements for the isolation include low leakage current and low parasitic capacitance. From this viewpoint, it is found that the trench structure filled with BPSG is superior to those filled with polysilicon. The low glass-transition characteristics of BPSG induce less mechanical stress, thus resulting in low leakage current of the isolated region.

The third approach is the use of SOI (Silicon On Insulator) substrate. Starting with the silicon substrate, silicon oxide and silicon layers are formed successively. The thickness of the silicon layer ranges from 50 to 500 nm, depending on the application. Isolation is made by means of either oxidation of, or etching off, the silicon layer, up to the underlying silicon oxide layer. Because of the isolation by silicon oxide or air, the leakage current between the components can be made quite small. Furthermore, latchup-free CMOS devices can be realized, because parasitic *pnpn* switches are not formed. SOI substrates may be fabricated by SIMOX (Separation by IMplanted OXygen) or BESOI (Bonding and Etchback SOI) technology.

1.3 Chip Yield and Reliability

In the actual applications of integrated circuits to electronic systems, price and stable operations are important factors, as well as chip functions, performance, and so on. Here, chip yield and reliability issues will be reviewed.

1.3.1 Chip Yield The first factor that determines chip yield is the point defects on a chip. One of the most common causes of point defects is dust or other particles present in the environment. These particles, when they have fallen on wafers, may induce imperfect patterns, leakage current through insulating layers, and so on. This may lead to chip yield loss. For this reason, wafer processing is done in clean rooms that contain particles within a specified number. A "class 1 room," for example, refers to one that contains no more than one particle with a size of 0.15 μm in diameter (see AEROSOLS). In response to the continuous decrease in device feature size, clean rooms with tighter cleanness control will be required.

The second factor to determine chip yield is the fluctuation of process parameters, such as the thickness of oxide layers, or impurity concentration. These fluctuations are due to the variations from the specified values of the process conditions, including temperature, and gas flow in production equipment. They will lead to unevenness of the device parameters and, hence, may result in chip yield loss. The third factor is a circuit design-related issue. When the circuit performance changes sensitively in response to fluctuations of the device and process parameters, high chip yield is not expected. A circuit design that affords circuit performance insensitive to parameter changes, that is, low-sensitivity circuit design, is highly desired.

For the sake of quantitative understanding of the chip yield issue, yield modeling and cost predictions have been made, taking various related factors into account. The results are used also to get insight into the possibilities of improvements in process facilities, design of process and circuit, and so on.

The chip yield is also an important factor that determines the maximum chip size. That is, larger chip size leads to lower yield and, hence, higher cost. Given the cost target of a chip, the maximum chip size is determined from the chip yield considerations based on the process technology level at the time. In fact, the maximum chip size has increased at the rate of roughly 1.14 times per year, or a factor of 15 during the 20 years from 1970 to 1990 (*IEEE ISSCC Digest*, 1970–1993; *Digest of Symposium on VLSI Circuits*, 1987–1993).

1.3.2 Reliability Fabricated devices should exhibit their functions and performance throughout their intended life. Generally, the term *reliability* is defined as the probability that a device will afford required functions under specified conditions for a stated period of time.

Failure rates for typical integrated circuits are classified into three regions, as shown in Fig. 5. After introduction into practical usage, the devices will encounter a relatively high failure rate, so-called early failures. Generally, this type of failure results from manufacturing defects. In the steady-state region, device failure occurs accidentally. The final stage is called the wearout region. In most ICs, there are no wearout mechanisms observed, as opposed to electron tubes. However, some time-dependent failure modes, such as electromigration or oxide breakdown, are

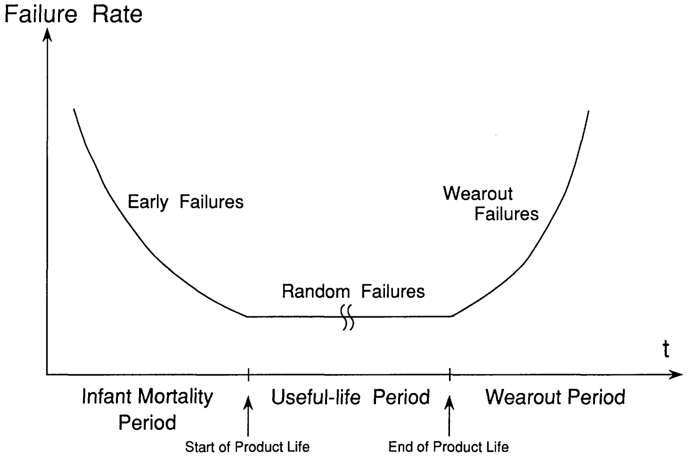

FIG. 5. Failure rate vs time of typical integrated circuits.

observed, especially in devices with poor reliability.

Many types of effort have been devoted to avoiding device failures after shipping the products. A typical approach to reject early failures is a screening, which is carried out by the burn-in process. In this process, the products are operated under accelerated conditions, such as increased temperature or increased voltage, for some period of time. During this process, most of the devices that are subject to early failure actually fail.

With regard to the steady-state failure shown in Fig. 5, a common approach is to estimate the failure rate of a device lot itself. The actual procedure is to pick out a specified number of devices from each lot, and subject them to accelerated conditions for some period of time. These conditions are set up to simulate those the devices will actually encounter in the market. Based on comparison of the measured failure rate with the target value, the decision is made whether each lot is able to be shipped out or not. As for the wearout failure shown in Fig. 5, identification of the possible failure mechanisms, and modifications of the process parameters or design, are carried out.

Several measures of failure rate have been used, including these:

1. ppm (parts per million); one ppm corresponds to one device in failure per one million devices. This measure is frequently used to express the reliability level of shipped devices, especially in the early failure region. The present market requires less than 50-ppm level for usual applications, and less than 10-ppm level for severe environment, such as automobile applications and so on.
2. FIT (Failure Unit); one FIT refers to one failure per 10^9 device hours of use of the products.
3. MTTF (Mean Time To Failure); this measure expresses the average lifetime of the device before it goes wrong.

From the viewpoint of reliability improvement, identification and elimination of failure mechanisms are quite important. It is especially true for devices with ever-increasing density and performance. For CMOS LSIs, the following typical failure mechanisms have been identified.

1.3.2.1 Gate-Oxide–Related Failures. The voltage applied to the gate oxide, or the current flow through it, is known to induce time-dependent breakdown of the oxide. This mode is TDDB (Time-Dependent Dielectric Breakdown), or simply wearout. A higher temperature leads to a higher failure rate. The mechanism of oxide breakdown is considered to be destruction of a Si-O bond by either a locally increased electric field or thermal runaway triggered by an increased current flow. The possible candidates for their origins are sodium contamination, electron traps, and hole traps in the oxide. Sodium ions segregated at the interface induce a lowering of the Si–SiO_2 barrier height, resulting in an increase in the localized gate current. The sodium-ion model gives a good explanation of the temperature dependence of TDDB. The electron-trap or hole-trap model explains the increased internal electric field by charge traps. These models are attempts to explain the mechanisms of intrinsic breakdown. The TDDB test is a tool to estimate the quality of the oxidation process and the silicon substrate, and also to screen a weak spot of gate oxide. Gate-oxide failure will be decreased by making the electric field applied to the oxide film no higher than the critical value.

1.3.2.2 Hot-Carrier–Induced Failures. Hot-carrier–induced failure refers to a change of the electrical characteristics of metal-oxide-semiconductor field-effect transistors (MOSFETs), such as threshold voltage or ON current, due to the injection of high-energy (hot) carriers into the gate oxide. Generally, hot electrons, as compared with hot holes, give more serious impacts on the device failure. So, the following descriptions will be limited to hot-electron–induced failure only.

With the continuous decrease in the transistor feature size, the lateral electric field between the drain and source electrodes has constantly increased, under the constant power-supply voltage of 5 V. The resulting high electric field accelerates electrons in the channel region, and also secondary generated electrons. As a result, part of the accelerated carriers gain the energy to flow into the gate oxide over the Si–SiO_2 barrier. Injected electrons introduce charge traps and interface states, through interactions between the injected electrons and Si-OH or Si-Si bonding. The major part of the degradation mechanisms in the gate oxide, however, is not fully identified.

Several models have been proposed to describe the population of hot electrons. Two typical models are a "lucky-electron" model and an electron-gas model. The lucky-electron model treats the electrons on the basis of the probability concept: Each electron has a specific probability to be accelerated to an energy higher than the barrier energy ϕ. The probability is $\exp(-\phi/\lambda E)$, where λ is the mean free path of electrons and E is the local electric-field strength. The electron-gas model, on the other hand, assumes a nondisplaced Maxwellian for the mass of the electron. The ratio of the number of electrons that have energy higher than ϕ is given by $\exp(-\phi/kT_e)$, where k is Boltzmann's constant and T_e is the electron temperature.

The common approach for this mode of failure is to employ a device structure that avoids hot-electron effects by lowering the electric field near the drain region. Examples include LDD (Lightly Doped Drain) and DDD (Double Diffused Drain), where some region is introduced with reduced electric-field strength. By use of these structures, devices with submicron feature size can operate under 5-V supply. The structure, on the other hand, requires precise design of transistor parameters, to avoid both the parasitic series resistance and degradation of the transistor drivability.

When the device feature size is around 0.5 μm or smaller, however, a 5-V power supply cannot be employed, to avoid hot-electron–induced failure. For devices with 0.5-μm feature size, for example, a power-supply voltage around 3 V is employed. Further reduction in the power supply and/or the introduction of new transistor structures will be necessary, depending on further reduction of the feature size of devices.

1.3.2.3 Metallization- and Interconnection-Related Failures. There are two typical modes for metallization- and interconnection-related failures. The first is electromigration. It is a failure resulting from material transport in metals, under enhanced and directional mobility of atoms caused by the electric field. The electric field is produced either by the applied voltage or by the current flow itself. The most typical example is electromigration of aluminum. Aluminum atoms are accumulated in the direction of electron flow, resulting in a discontinuity of the interconnect, which leads to cessation of the device operation. This type of failure is one of the most serious failure modes in integrated circuits.

The aluminum atoms that are upstream in terms of electron flow from a vacancy have a higher probability of occupying a vacancy site, and thus inducing electromigration. In polycrystalline aluminum films, the electromigration occurs mainly along grain boundaries, where aluminum atoms are bound less tightly than in lattice positions. An equation for MTTF of aluminum interconnects, called Black's formula, has been empirically derived as follows:

$$t_{50} = AJ^{-n} \exp(Q_0/kT),$$

where t_{50} is median time to failure, A is a process- and material-dependent constant, J is the current density, n is a current-dependent constant, Q_0 is the activation energy for the diffusion of aluminum atoms, k is Boltzmann's constant, and T is the absolute temperature. Typical approaches to suppress electromigration are design of the integrated circuits so as to limit the value of J within the critical value, addition of Cu to Al, and formation of structured Al layers by employing transition metals.

The second failure mode is stress-induced migration. For aluminum lines with a linewidth of roughly 1 μm or smaller, the failure becomes more serious. When Al interconnects covered with passivation layers are subject to IC fabrication process (400–500 °C), or high-temperature (150–250 °C) storage, voids are induced. This results in an open-circuit failure. Possible mechanisms include compressive intrinsic stress in passivation films, tensile stress in the Al interconnects due to the difference in the thermal-expansion coefficients between Al and CVD films, and deformation of passivation films. Countermeasures for this failure include addition of Cu or Ti to Al, or employment of a layer of Al laminated with Ti, W, or TiN.

1.3.2.4 Electrostatic Discharge (ESD). In the handling of devices, voltages higher than the breakdown voltage of the gate oxide may be applied on the circuits, thereby inducing device failure. Typical examples of induction of very high voltages include a person walking near ICs, and removal of an IC from its plastic packaging material. To avoid this type of failure, input and output circuits are designed so as to provide paths for the discharge current, thus preventing induction of excessive voltage across the gate oxide of the device.

1.3.2.5 Latchup. In CMOS devices, there are many source and drain regions close to adjacent tubs. These structures may be causes of parasitic bipolar transistor action, and, thus, may lead to parasitic *pnpn* switches. When the firing condition is met, parasitic *pnpn* switches are turned on, and the chip turns into a very low-impedance state. In this state, the device is not controlled by the input signals. The phenomenon is reversible. That is, the device will come back to its old state if the power supply is turned off. This failure mode will be avoided by the design of the distance between the impurity regions comprising the parasitic bipolar transistors large enough to avoid the firing of the *pnpn* switches, employment of a trench isolation structure, and/or employment of epitaxial wafers. For the final case, the substrate for the epitaxial layer plays a role as a stable equipotential plane, thus avoiding the parasitic bipolar actions.

1.3.2.6 Alpha-Particle–Induced Soft Errors. Alpha particles emitted by radioactive elements contained in IC packaging materials cause soft errors in the ICs. The penetration of alpha particles into the silicon causes the generation of electron-hole pairs. The generated carriers may lead to the loss of charge stored in dynamic circuits, including the memory cells of a dynamic memory, or dynamic shift registers. The term soft error refers to a failure that does not lead to a physical defect of the device. Approaches to avoid this type of failure include making the capacitance value larger, and the coating of chips with material with a very low density of radioactive elements.

With the continuous decrease in device feature size, increase in the number of components on a chip, and also the increase in process complexities, conventional approaches for reliability assurance are becoming unsatisfactory. One novel concept for more reasonable reliability assurance is the building-in of reliability, or design for reliability (DFR). This is a typical example showing the shift of approaches from measurement and analysis, to design and synthesis, of device reliability. Deeper understanding of the physical phenomena, and development of better modeling and simulation systems, are highly required (Sze, 1988; Takeda *et al.*, 1993).

1.4 Circuit Configurations of Integrated Circuits

1.4.1 Circuit Configurations for VLSIs and Beyond The circuits to be applied to VLSIs and beyond are required to have better figures of merit, that is, higher-speed operation, lower power consumption, and smaller chip area. From this viewpoint, several circuit configurations of CMOS or BiCMOS circuits have been proposed and applied to practical usage.

The first circuit type that has the above-mentioned features is a dynamic circuit. One of the most typical examples that employ this circuit type is DRAM (Dynamic Random-Access Memory), which has always realized the largest number of components on a chip. Through the development of the DRAM chip, a number of novel circuits have been designed (Itoh, 1990). Varieties of other circuit types for better figures of merit have been proposed, including clocked CMOS and Domino Logic circuits. These circuits require high-level design techniques. In order to utilize fully the apparent advantages of these circuit types, judicious studies on the circuit operations and the design strategies are indispensable.

For further power reduction, circuit operation under low supply voltage is extensively studied. In this case, simultaneous realization of low power consumption and high-speed operation is the requirement (Chandrakasan, *et al.*, 1992).

The BiCMOS circuit has the advantage of higher operational speed compared with the CMOS circuit. This is because the bipolar devices have higher driving capability than MOS devices. The power consumption and chip area of BiCMOS circuits, on the other hand, are larger than for CMOS. In the selection of circuit type, comparison of pros and cons of the candidate circuits should be carried out (Santo, 1989).

High-speed interface circuits are becoming increasingly important. This is because the high-speed signal transmission between chips on a printed wiring board is one of the key factors to realize high-speed digital systems, as well as high-speed chips (Gunning *et al.*, 1992).

For the basic circuit types, see CIRCUITS, DIGITAL and CIRCUITS, ANALOG.

2. SELECTION OF PROCESS TECHNOLOGIES FOR INTEGRATED CIRCUITS

In chip development, selection of the best possible process technologies and fabrication facilities is of utmost importance. Generally, candidates for the selection have their own strong and weak points. Therefore, detailed comparison of the pros and cons of the candidates is indispensable for the reasonable selection of process technologies.

Important factors to be taken into account in the selection of the process technologies will be grouped into three. They are (1) high performance, (2) high productivity, and (3) high reliability. Higher productivity is important to realize lower production cost and also shorter turnaround time in chip development and production. Higher reliability is important for long and stable operation of the electronic systems, which is an important condition for customer satisfaction. Each of the above three groups consists of multiple factors. The starting point of the selection of process technologies is to list those factors in the order of importance. Shown below are some examples of discriminating criteria used in the selection of alternative process technologies. There may be, of course, different criteria, depending on the situation.

Here, factors to be considered in the selection will be reviewed, with some specific examples. As for the element process technology, see VLSI TECHNOLOGY.

2.1 Lithography

There are three approaches for the lithography process: optical, electron-beam, and x-ray systems. Factors to be taken into account in the selection of the technology are listed as follows, in the order of importance.

1. Resolution. The resolution, that is, the minimum dimension that can be defined by the technology, is the most important factor in lithography systems. At present, optical systems with various wavelengths are employed in the actual fabrication or development: *g* line (wavelength equal to 436 nm) or *i* line (365 nm) from a mercury lamp, KrF (248 nm) and ArF (193 nm) lasers. Electron-beam systems that employ an electron beam with the wavelength of 0.02–0.05 nm, depending on the energy, afford much higher resolution compared with the optical systems. As will be described later, issues of electron-beam systems include low throughput and expensive production facility. On the other hand, lithography systems using x-rays with the wavelength of 0.3–5 nm have also much higher resolution capability compared with optical systems. The issues for this type of system include difficulty in the fabrication of masks and the expensive production facility. Efforts are now being devoted to overcome the above-mentioned issues.
2. Depth of focus (DOF). In order to define patterns on silicon surfaces with unevenness of the silicon oxide generated during the fabrication processes, good focus should be assured within some range of depth. DOF refers to this focus margin. Generally, electron-beam and x-ray systems afford better DOF, compared with optical systems.
3. Field size. In order to make a pattern transfer for every mask layer at once, the field size of the system should be large enough to cover the chip size. To keep pace with the increase in chip size, increasing field size is required. The maximum field size for present optical systems is around 22 mm × 22 mm.
4. Uniformity.
5. Throughput. From the viewpoint of production cost, the throughput of the process is quite important. At present, electron-beam systems have a throughput much lower than that of optical systems.
6. Reliability. From the reliability standpoint, the radiation damage to the transistors induced by the incident electron beams is an issue for electron-beam systems.

2.2 Etching

In regard to the etching process, there are two types of approaches: wet and dry etching systems. Dry etching is further divided into several technologies: ion milling, chemical dry etching, and reactive ion etching (RIE). Factors to be taken into account in the selection of the technologies are listed as follows, in the order of importance.

1. CD (critical dimension) loss. CD loss refers to the deviation of the dimension of the most important pattern (usually the minimum dimension), after processing, from the designed one. Zero CD loss means the formation of patterns just the same as the designed ones. There are several sources leading to CD loss in the mask-making, lithography, and etching processes. Dry etching is superior to wet etching in this respect, thanks to its anisotropic nature.
2. Gate-oxide destruction. Destruction of the gate-oxide film by charge-up damage during the etching process can be a fatal problem.
3. Selectivity. One common approach to enhance the selectivity is the use of a protection layer. For example, selective etching of vias can successfully be carried out with the addition of CO gas, which forms the protection layer.
4. Uniformity.
5. Throughput. Etching rate is important to realize high throughput.
6. Particle contamination. Etching by ion milling technology, for example, has many favorable features. However, it is not employed in volume production, because of the particle contamination.

2.3 Metallization

The metallization process refers to the formation of the electrical contacts and interconnection layer(s). There are three approaches for that; evaporation, sputtering, and chemical-vapor deposition (CVD). Factors are listed as follows in a similar way.

1. Resistance. Low resistance of the interconnection layers and, also, low contact resistance are of utmost importance. The former requirement is the issue of materials selection. As for the latter requirement, employment of several metals in layers is being tried to realize good adhesion between a metal and silicon.
2. Coverage. Especially, coverage of the vias is important to realize a low contact resistance and high reliability. From this viewpoint, sputtering is better than evaporation. Evaporation is unable to realize full coverage of vias with a large aspect ratio, because of the incident-angle issue. Furthermore, the chemical-vapor deposition of W affords better coverage of vias with both a large aspect ratio and small size.
3. Adherence. Whereas the adherence of Al to silicon oxide is enough, that of W is not so good. Addition of TiW between W and the silicon oxide is very effective to improve the adherence to the oxide.
4. Immunity to electromigration. See Sec. 1.3.2.3.
5. Reproducibility. As an example, the selective chemical-vapor deposition of W has not been applied to practical usage, because of its poor reproducibility.

2.4 Thin-Film Formation

For the formation of thin films for gate dielectric and capacitor applications, oxidation and chemical-vapor deposition are the two types of approaches employed. The related factors are as follows.

1. Breakdown voltage and leakage current. From the applications point of view, high breakdown voltage and low leakage current are the most important requirements.
2. Step coverage.
3. Uniformity.
4. Particle contamination. Avoidance of particle contamination is highly desired.
5. Throughput. The rate of the oxidation or deposition is important.

2.5 Common Issues

In addition to the above-mentioned factors, there are other important factors that are common to each process technology. First, typical requirements for the production facilities include

1. price;
2. stable operation;
3. volume;
4. maintainability; and
5. extensibility.

With regard to item **3**, the recent trend is the growing introduction of single-wafer systems. It has attractive features from the viewpoint of flexibility and facility cost. Regarding item **5**, efforts are being devoted to the utilization of fabrication equipment over several device generations, after necessary modifications. This is an example of a countermeasure to the rapid increase in investment on fabrication equipment due to the decrease in device dimensions and the increase in process complexities.

Second, requirements for the process-flow-related issues include the following:

1. The number of processing steps should be as small as possible;
2. the process can integrate any type of function, such as memory, digital, analog/digital mixed circuits, and so on; and
3. there must be high controllability and high reproducibility.

Regarding item **1**, one example is the CMP (Chemical Mechanical Polishing) technology, which has a smaller number of processing steps than the etchback approach.

3. DESIGN METHODOLOGY AND COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS

With the continuing increase in the number of components integrated on a chip, increase in the complexities of transistor structures, and decrease in the feature size of transistors, design of integrated circuits (ICs) is becoming increasingly sophisticated. The demand for shorter design turn-around time (TAT), on the other hand, is increasing, because of the time-to-market competition of the new product development. From these points of view, development and introduction of both advanced design methodology and computer-aided design (CAD) technology are greatly needed.

3.1 Design Methodology of Integrated Circuits

In the typical design of integrated circuits, product specifications are successively transformed to more detailed descriptions (top-down design) (see COMPUTER-AIDED DESIGN IN ELECTRONICS). Here, each design stage makes full use of the library. The library stores the generic basic circuits such as gates or flip-flops, which are called "primitive cells," and function units, called "function macros." This library is prepared from process, device, and circuit designs, which are carried out by using simulations and refined on the basis of the results of measurements. This approach is effective for realizing shorter design TAT.

So-called design methodology refers to any strategy for improving the design efficiency to obtain shorter TAT using less engineering resources. The first is top-down design. It is also called hierarchical design, because of the transformation from abstract description to a more detailed one. The second strategy is the so-called divide and conquer design. It refers to the design of a chip after dividing it into several portions with data sizes that can easily be processed. Each portion will be designed in a short design TAT. Also, a design can be carried out in parallel, by multiple designers.

The third strategy is concurrent engineering (CE) (Rosenblatt *et al.*, 1991). In CE, many jobs that are related to the chip development, that is, marketing, design, and manufacturing, for example, are carried out simultaneously, keeping good communication among groups. By working together from the start of the device development, each group can anticipate problems and eliminate them at an early stage of the development. Compared with the so-called serial approach, where each job is carried out in cascade, this parallel approach leads to shorter development time and cheaper development cost. In this case, computer-aided design, engineering, and manufacturing technology is one of the key factors for success, in addition to a good framework.

For the CAD tools for high-level design (that is, behavior, register-transfer, and structural level design), and layout design, see COMPUTER-AIDED DESIGN IN ELECTRONICS.

3.2 Test CAD

Technology for chip testing is quite important in the production of integrated circuits. There are two serious issues associated with the increase in the number of components integrated on a chip: the increase in the time required for testing fabricated chips and the decrease in the test coverage. The term "test coverage" refers to the percentage of cases that chips in failure are exactly detected. Design style that provides means for easy chip testing is generally called "design for testability (DFT)." A typical example of DFT is a so-called scan design. It converts flip-flops to combinational circuits by using additional control circuits, thus drastically reducing the number of test patterns required for testing. As a result, a reduction in testing time and an increase in test coverage can be obtained. The technology has been widely applied to the testing of ASIC products. This approach, however, has not been applied to the custom design of chips in volume production, including microprocessors. This is because the increase in the chip area caused by the application of the scan technology is too large from the chip cost point of view. DFT technology that affords a chip area increase much less compared with the existing approach, while keeping high test coverage, is greatly needed. Such approaches, including partial scan and built-in self-test (BIST) systems, are being developed for that purpose.

Furthermore, automatic logic synthesis that takes into account the test design, called "synthesis for testability (SFT)," is a promising approach for high-quality design and short design time, by avoiding repetitive rework cycles in the chip design.

Regarding the testing of integrated circuits after assembly on the printed circuit board, boundary scan technology is widely used (Levitt, 1992).

3.3 CAD for Process, Device, and Circuit Design

Conventionally, the development of a new process is carried out by a trial-and-error approach. Test chips are fabricated under varieties of process parameters and then evaluated. By the comparison of the measured data with the target values, the optimal process conditions are determined. This approach, however, is becoming increasingly inefficient. It is because reduction in the device feature size and the employment of more complex device structures lead to an increase in the process complexities and the number of process steps, and coupling between process conditions, device behavior, and circuit performance in an increasingly intimate fashion. Hence, the conventional approach requires a great deal of development time and cost, which is becoming impractical from the time-to-market point of view in new product development.

On the other hand, thanks to both the progress of simulation technology and related CAD tools, device structure, doping profile, device behavior, and circuit performance can be simulated on computers. By employing these technologies, conventional test fabrication and performance evaluation in the process technology development can be replaced by simulations on computers. It is expected that the full use of so-called computer experiments will lead to a drastic saving of time and cost of process technology development.

The sequence of the process, device, and circuit simulation is roughly as follows. Process simulation covers the element process (see VLSI TECHNOLOGY), including lithography, ion implantation, oxidation, and so on. The results of the process simulation, such as the impurity profile, are forwarded to the device simulation. Starting with the data supplied from the process simulation, and the given device structure, the device simulator then simulates the device behavior by calculating the carrier transport and the electrostatic potentials induced by the external bias voltages. Parasitic parameters of the interconnect are also extracted by calculating the electrostatic and vector potentials. Then the circuit performance is estimated through the circuit simulation, by using the device behavior supplied by the device simulation, together with the circuit configuration.

The above-mentioned simulation consists of the modeling of each phenomenon or device, and the subsequent numerical calculations based on the model. Here, employment of a model with reasonable accuracy is of utmost importance. This is because an excessively accurate model does not necessarily give results leading to good insight into the phenomena, and a complex model generally requires too lengthy computing time, which may be impractical from the viewpoint of the time frame of the integrated circuit development.

3.3.1 Process Simulation The history of general purpose process simulators effectively started in 1977, with SUPREM and SAMPLE programs developed by Stanford University and the University of California, Berkeley, respectively. Such processes as lithography, etching, deposition, oxidation, ion implantation, diffusion, and epitaxial growth are covered by both or one of the above simulators. Since then, improvements have been carried out in many respects (Antoniadis *et al.*, 1977; Sze, 1988).

3.3.1.1 Lithography. For the lithography process, optical imaging and exposure of resist materials are successfully modeled by calculating irradiation distributions and light propagation within the layer, respectively, and are used in practice in the optimization of optical projection systems. On the other hand, development of resist and subsequent etching simulation are not well established yet, because of the lack of knowledge about the chemical and physical phenomena at the wafer surface combined with the wide variety of process and equipment variations.

3.3.1.2 Ion Implantation. Two types of modeling methods are well established:

1. Monte Carlo simulation that calculates the motion of dopants through their successive collisions with target particles, and the resulting defect distributions;
2. impurity distribution functions that are analytically characterized by their moments such as projection range and standard deviation.

These two models complement each other. Monte Carlo simulation is accurate but very CPU intensive, while the analytical distribution-functions model is less expensive, but requires the moment-parameter extraction. Both models cover the various doping species and the wide range of dose and energy conditions, including high-energy implantations, often used in modern processes. Figure 6 shows an example of an analytically simulated impurity profile, as compared with the measured one.

3.3.1.3 Thermal Diffusion. In full generality, the impurity diffusion process is simulated by calculating the time evolution of dopant concentrations governed by the continuity equations for all charged particles. Here, species can be donor, acceptor, vacancy, or interstitial. By determining the diffusion coefficient for each species, the impurity distribution is obtained with good accuracy. To take interstitial diffusion into consideration is important to simulate profiles for modern processes accurately, such as low-temperature and short-time annealing processes.

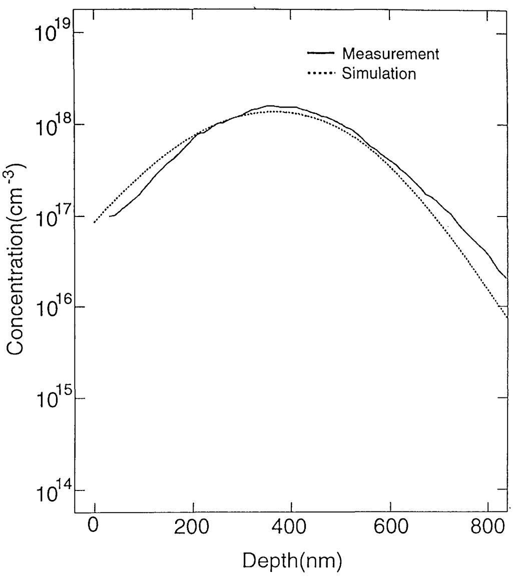

FIG. 6. An example of an analytically simulated impurity profile, as compared with the measured one. (Implanted boron profile under the condition of 120 keV, 5 × 10^{13} cm^{-2}, followed by annealing at 900 °C for 30 min.)

3.3.1.4 Oxidation. The growth of SiO_2 is successively simulated by integrating the oxygen flux diffusing through the existing SiO_2 layer. A compact equation representing the growth of the layer was derived by Deal and Grove (Sze, 1988).

3.3.2 Device Simulation

The following two simulation models are used:

1. A drift-diffusion model, which assumes that carrier distribution is in thermal equilibrium with the semiconductor lattice and that the current is a sum of the drift and diffusion components.
2. A hydrodynamic model, which further considers the carrier energy-conservation equations (Engl, 1986; Selberherr, 1984).

Model **1** predicts the electrostatic potentials and the carrier distributions fairly well. The drain current characteristics down to 0.1-μm MOS channel-length regions are predicted. Model **2** is useful, for example, to predict hot-carrier–related characteristics.

Figure 7 shows an example of simulated electrostatic potential for a MOS device.

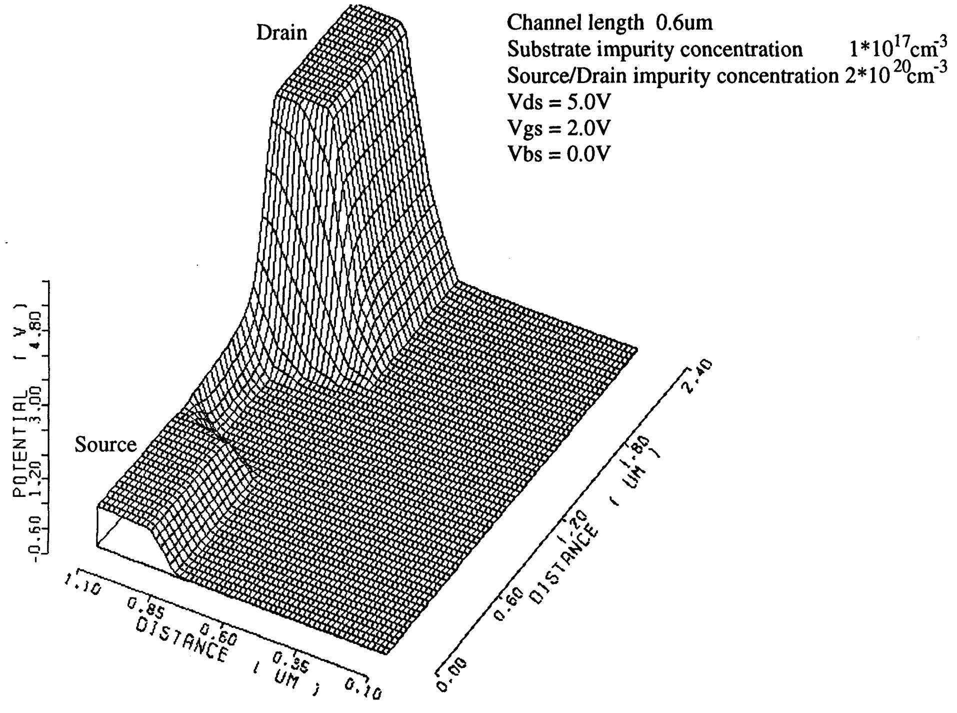

FIG. 7. An example of simulated electrostatic potential of a MOS device.

3.3.3 Circuit Simulation A set of linear equations of the entire circuit is solved to obtain circuit voltages and currents. Device characteristics are formulated in closed forms after the Poisson and the drift-diffusion equations are integrated under the appropriate assumptions. Typical models are the Gummel–Poon model for bipolar devices, and the BSIM (Berkeley SIMulation) model for MOS devices. The performance of a LSI with device technology down to 0.35-μm MOS channel length can be simulated (Nagel, 1975; Sangiovanni-Vincentelli *et al.*, 1980).

3.4 Design Environment and Technology Standardization

At present, varieties of CAD tools are supplied from many tool vendors. From the viewpoint of chip design at maximum design efficiency, designers want to choose the best possible CAD tools and combine them into an integrated design system. In this case, each CAD tool is operated on various platforms, such as engineering workstations (EWSs) or personal computers (PCs) (Comerfold and Mogal, 1993). On the other hand, ASIC vendors are delivering databases for their ASIC devices, such as gate arrays and cell-based ICs. Under these multivendor situations, in the areas of both CAD tools and IC databases, establishment of common bases for promoting mutual data exchange and smooth combination of tools is gaining increasing importance. From this viewpoint, world-wide activities on technical standardization are on the way for the realization of data interchangeability and interoperability. Described below are some examples.

3.4.1 CFI (CAD Framework Initiative) The purpose of the CFI's activity is the establishment of a common interface between individual CAD tools and the framework into which individual CAD tools should be incorporated. When the common interface is set up, individual CAD tools developed by any tool vendor can be incorporated into the framework. This opens a way to the realization of an optimal CAD system, comprised of the combination of the best possible individual CAD tools.

3.4.2 EDIF (Electronic Design Interchange Format) EDIF is one of the most typical standardized formats, for design data exchange.

3.4.3 HDL (Hardware Description Language) HDL is widely employed in describing electronic circuits, in place of the schematic expression. At present, VHDL (VHSIC Hardware Description Language) and Verilog-HDL are typical ones (Waxman *et al.*, 1989).

3.5 Future Trends

In order to realize shorter TAT for the design of chips with increasingly larger numbers of components, the following approaches will be indispensable:

1. Development of CAD tools to support higher-level design, including automatic synthesis of register-transfer–level description.
2. Establishment of an integrated design system. The final goal of the design system is a silicon compiler, which automatically translates chip specifications to mask-making data, the final level of the description.
3. Development of layout technology to produce high component-packing density. Realizing a chip size as small as possible is one of the most important requirements, from the viewpoint of the cost competitiveness of the chip.
4. Introduction of powerful platforms. Thanks to the rapid increase in the performance of microprocessors, EWSs and PCs with increasingly higher performance will be developed. Furthermore, CAD engines, dedicated hardware systems for specific design, will be developed. Timely introduction of these platforms will be more and more important for the increase in design efficiency.

4. APPLICATIONS OF INTEGRATED CIRCUITS TO ELECTRONICS SYSTEMS

Integrated circuits have been applied to almost all electronics systems, and have contributed to the creation of new functions and services with drastically reduced cost. Functions of the electronics systems are integrated in a form selected out of the following three types: standard products, ASIC, and ASSP, depending on the target specifications (see Sec. 1). This section briefly reviews such a trend and shows some typical examples.

4.1 Communication Systems

Steady progress has been made in communication systems to get closer to the goal, "Anytime, Anywhere, Access to Anyone."

The telephone system is one of the most typical systems in a communications network. At the center of the system is an exchange system, in which incoming signals are connected to a specified telephone number, via switching subsystems. Following the initial mechanical exchange systems employing relays or crossbar switches, electronic exchange systems employing electron tubes or transistors were developed to realize lower system cost. Then, based on the progress of integrated circuits and digital transmission technologies, a digital exchange system was developed in the 1960s and put into practical service in the 1970s. A digital exchange system has many favorable features, including higher transmission quality, lower system cost, and versatility in system implementations. It is a key system component of the ISDN (Integrated Services Digital Network).

One of the most important functions in digital exchange systems is the conversion of the speaker's voice signal to a Pulse-Code Modulated (PCM) signal, which is called "coding," and the reverse function, which is called "decoding." The device to perform both functions, which is called a CODEC (COder+DECoder), was first realized as a combination of transistors and passive components on a printed wiring board, then as a combination of small-scale integrated circuits and passive components on a ceramic board (that is, a hybrid integrated circuit). In accordance with the progress of integrated-circuit technology, the CODEC was first integrated in a single silicon chip in early 1980s, by use of CMOS technology. CMOS devices have the significant features of low power consumption and high packing density, which are very effective to realize a single-chip CODEC. By employing a single-chip CODEC, the size, power, and production cost of digital exchange systems were drastically reduced, which made a big contribution to the widespread penetration of the telephone network.

Regarding "Anytime, Anywhere, Access" service, the contribution of personal communications, that is, the mobile and portable telephone service, is remarkable. Especially, in cellular telephone systems, communications between handheld transceivers are made through networked base stations or "cells" (Fischetti, 1993). They are in wide use all over the world. The majority of the handsets for cellular telephones employ CMOS devices, for their features of low power consumption and operation under low supply voltages, as well as high packing density of components on a chip. The former two features are of utmost importance in battery-drive operation of portable terminals. The present battery operating time of the telephone handset is typically one to two hours. Realization of much longer service times between battery recharging is a target for the moment. Reduction of the power consumption of the chips, as well as increase of the capacity of the battery itself, is necessary for that purpose. Lowering the supply voltage and reducing the device feature size will be practical approaches for the power reduction of chips, for the time being.

Future personal communications systems are now under development, where a single phone number is assigned to each individual, no matter where the subscriber is. The system will afford "Anytime, Anywhere, Access" service in a more advanced way. One of the most important issues for that is the development of low-power, small, and light-weight personal communicators, for which the role of integrated circuits is quite significant (Heilmeier, 1992; Kobb, 1993).

Image signals, including still pictures and motion video, play quite important roles in communication systems. In just the same way as for voice signals, image signals are converted to pulse code modulated signals by a coder at the transmitting terminals, and converted back to the original signals by the decoders at the receiving ends. The amount of information inherent to the image signals, however, is quite enormous, say a thousand times as much as that for the case of voice signals. Therefore, the clock frequency required for image processing is high, roughly over 10 MHz, for example. In earlier days, these frequency ranges could be covered only by bipolar devices. Bipolar devices, however, are not appropriate to integrate on a chip enough functions for image-processing applications, because of the large chip area occupied by transistors, and large power consumption. On the other hand, MOS devices, which have the features of small chip area and low power consumption, were not fast enough to meet the speed requirements for image processing.

Thanks to the continuous reduction of the feature size of MOS transistors and the improvement in circuit technology, the clock frequency of MOS LSIs has made a steady increase. In the mid-1980s, the clock frequency of CMOS LSIs exceeded 30 MHz, which opened a way to image processing by use of CMOS LSIs. Since that time, varieties of CMOS LSIs for image-processing applications have been developed, thanks to the constant increase in the clock frequencies of CMOS LSIs. So-called multimedia, that is, image signals in addition to voice and data signals, have the possibilities to create new and broader areas of services. In this sense, the mid-1980s are considered to be the starting point of the so-called multimedia era. From this viewpoint, the goal of communication services may be expressed as "Anytime, Anywhere, Access to Anyone with Anymedia."

One of the most typical examples of multimedia services is a video teleconference, where multiple attendees in remote places join the mutual discussions, information exchange, and so on, by using image signals, voice, and data. As an extended version of the video teleconference system, desk-top video teleconference systems, operating on workstations or personal computers, are now being developed by many vendors. The systems will make possible video teleconferences among attendees sitting at their own desks, which will be a powerful tool for the promotion of a type of CSCW (Computer-Supported Cooperative Work). In the future, handheld terminals will be able to handle multimedia information, as a result of the drastic progress of CMOS technology, in terms of the large number of components on a chip and also very low power consumption.

Establishment of several technical standards for compression/decompression of images and voices, and the development of CMOS LSIs following the standards, are in progress (Yurgan, 1992).

One of the most urgent issues to realize multimedia services is the establishment of a broadband network that can transmit broadband signals in a reasonable time. A broadband integrated services digital network (abbreviated as B-ISDN) is an important infrastructure for future communication network systems (Kaplan, *et al.*, 1991). It has a transmission speed of 150 Mb per second or higher. Several devices for B-ISDN are now under development in many organizations.

In the United States, an "Information Super Highway," a next-generation high-speed information network, is under establishment.

4.2 Computer Systems

Computer systems have always provided the largest markets for semiconductor devices and also stimulated the development of leading-edge devices to realize the most advanced computer systems. Transistors were first introduced into computer systems in the 1960s, and then integrated circuits, in the 1970s. This contributed to the realization of a system with much smaller size, higher reliability, and much lower cost, compared with the system comprising electron tubes. Integrated circuits widely used at the initial stage were bipolar small-scale integrated circuits, such as TTL (Transistor-Transistor Logic), ECL gates, and flip-flops. Here, TTL had the features of low power consumption with fairly high speed at that time, and ECL showed the highest operational speed (see CIRCUITS, DIGITAL). Since then, ECL integrated circuits with an increasing number of components on a chip have been developed, making contributions to the realization of so-called main-frame computers which have leading-edge performances. Here, gate array or cell-based ICs have been mainly employed to integrate digital system functions.

On the other hand, MOS integrated circuits have many advantages, including smaller chip area and lower power consumption, compared with ECL devices, although with lower operational speed. In the early 1970s, two MOS LSIs were first disclosed, a 1-kbit DRAM in 1970 and a 4-bit microprocessor in 1971. Microprocessors are devices that integrate the central processing unit (CPU) of the computers on a chip (see COMPUTERS; COMPUTER HARDWARE; MICROPROCESSORS). After that, the number of components integrated on a chip for memory and digital devices has steadily increased, as shown in Fig. 1. These MOS LSIs have been widely applied to minicomputers and desk-top machines (that is, workstations and personal computers).

The performance of microprocessors has increased at the rate of roughly 2.7 times per every three years. Recently, RISC (Reduced Instruction Set Computer) microprocessors have given big contributions to the rapid increase in microprocessor performance. Compared with the conventional CISC (Complex Instruction Set Computer) microprocessor, the RISC microprocessor has an essential advantage of realizing a short cycle time. As a result, the performance of the desk-top machines has been approaching that of main-frame computers. As a consequence, the percentage of computing jobs that are processed on small-sized machines is increasing year by year. This trend is called "downsizing." The term refers to the reduction in both the size and price of the machines. This is a big and typical change generated by widespread introduction of microelectronics (Gelsinger *et al.*, 1989; Comerfold and Mogal, 1993).

The rapid increase in the performance of MOS LSIs is stimulating a new trend in the computer field. So-called massively parallel processors (MPP), composed of large numbers of processors connected in parallel, can afford high throughput performance as a whole. This type of machine can be applied to heavy computing jobs, including two- or three-dimensional simulations of semiconductor devices, or the simulation of the fabrication processes of integrated circuits. As a result, MPP is becoming one of the most promising approaches for recent most advanced supercomputers.

LSIs for controlling peripheral functions, such as controllers for floppy and hard disks, and display devices, are also indispensable to realize high-performance computers without overloading microprocessors. In addition, recently, displaying a large amount of data in a visual form is becoming increasingly important, for easier understanding of the data. This field of technology is called "scientific visualization." For this purpose, a high-speed graphic display controller (GDC) chip is one of the most important devices for workstations and personal computers (see COMPUTER GRAPHICS).

Historically, because of the rapid increase in the performance of workstations, the price-to-speed ratio of the computers for computer graphics (CG) applications, for example, has been reduced by a factor of roughly 1/100 during the past five years. This is one of the most important factors to accelerate the applications of CG technology to many fields, including the production of movies or pictures.

4.3 Consumer Electronics

One of the most typical products in consumer electronics is a desk-top calculator. The first model was developed in the late 1960s using bipolar ICs, including TTL. In the 1970s, 1-chip CMOS LSI with low power consumption and battery-drive capability was developed. A low-price palm-top calculator was put to practical applications by the combination of this chip and the liquid crystal display device. Since then, the palm-top calculators have always created new markets, based on the increasingly enhanced functions/performances and cost reduction.

The automobile system is another example in which the introduction of microelectronics to electronic systems has been making big impacts. Employment of semiconductor devices in automobile systems was first made in 1960, with a rectifying semiconductor diode. Following the first introduction of an *n*-type metal-oxide semiconductor (NMOS) microprocessor in the mid-1970s, CMOS microprocessors and LSIs have extensively been introduced since the mid-1980s. Microprocessors are widely used to control engine systems, the sequence of door or seat operations, and information systems such as a car navigation system. Contributions of microelectronics will further include safety and amenity improvement, and information assistance, as well as optimal engine operation control from the viewpoint of environmental protection.

A television system is also a typical consumer product to which varieties of integrated circuits have been applied, resulting in cost reduction and reliability increase. Recently, a HDTV (High-Definition Television) is under development all over the world. The system has a resolution of roughly two times as large as the present television system, in both horizontal and vertical directions. This high-resolution presentation of video signals is expected to develop new services in consumer electronics, and also other applications, including medical and industrial fields. At present, experimental broadcasting service has started in Japan. In the United States and Europe, examinations on the system specifications are being carried out. A big commercial market for HDTV systems is expected to build up late this century or in the beginning of the next century. Important requisites for promoting the buildup of HDTV service will be preparation of both low-price receiver sets and abundant software. One of the key factors for success is the development of LSI chip sets with a chip count as small as possible.

When multimedia information is expressed in a digital format, the signal can be treated seamlessly in television, computer, and communication, which leads to the creation of varieties of new services. For example, CD-ROM, a type of disk, can store 540 Mbytes of data. On the other hand, video information, when compressed according to the MPEG1 standard, has a data transfer rate of roughly 1.5 Mbit/s. Therefore, roughly 48 minutes of compressed video information is stored in the CD-ROM [$(540 \times 8/1.5)/60 = 48$ minutes] (Yurgan, 1992; Adam and Cole, 1993). This capacity, however, is not enough to store varieties of types of information, from the viewpoint of recording time and signal quality. The development of a disk with much larger capacity is greatly needed.

Figures 8(a) and 8(b) provide an insight into the storage capacity required for the typical services. Figure 8(a) shows some examples of services employing still pictures. The horizontal axis shows to what extent the original signal is compressed, in order to limit the signal bandwidth. The vertical axis corresponds to the number of sheets that can be recorded on a memory chip. The figure shows that the development of a memory chip with 64 Mbit capacity or larger will find varieties of services. Figure 8(b), on the other hand, shows typical examples of video-signal–related services and required memory capacity. Here, it is assumed that the video signal has the usual National Television System Committee (NTSC) format. In this case, to record video information on a single chip is a very tough issue, even if a high compression ratio is assumed. These applications are

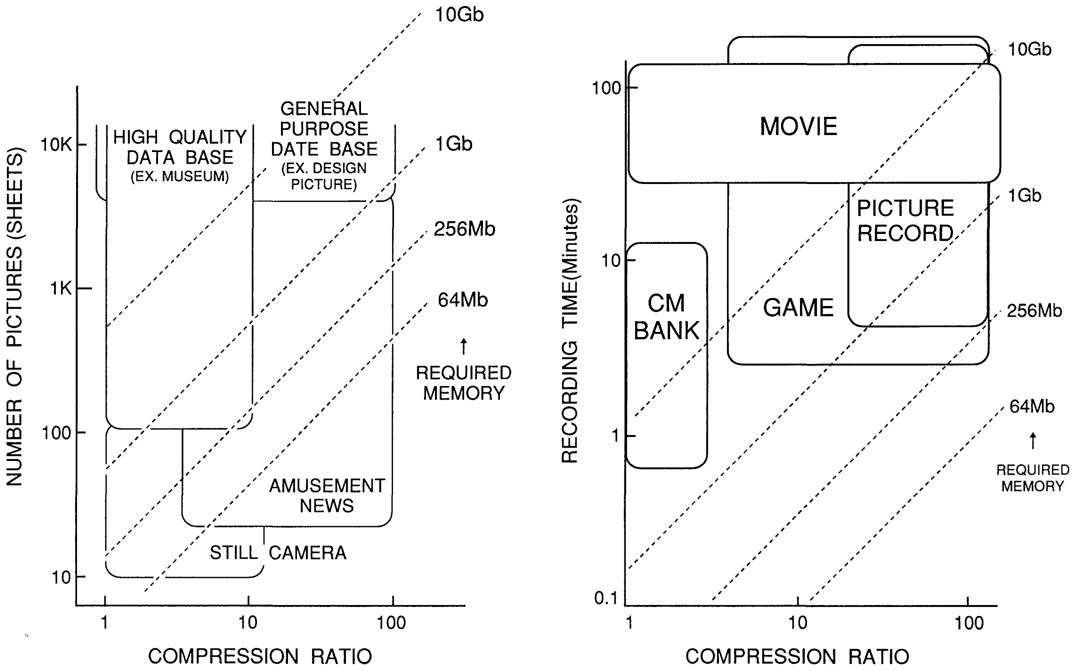

(a) (b)

FIG. 8. Typical examples of image-signal–related services and necessary memory size: (a) still pictures, (b) motion video.

waiting for the development both of the advanced semiconductor technology and efficient signal-compression algorithms. As for the recording of the HDTV signals, the required memory capacity is several times as large as that for signals with NTSC format, corresponding to the broader signal bandwidth of the original signal. This application will be a target of big technical challenge.

4.4 Human Interface

One of the most familiar forms of Human Interface (HI) with electronics systems is the user interface with computers. The development of GUI (Graphical User Interface) has contributed to better user interface (see COMPUTER GRAPHICS).

For portable computers, on the other hand, making an input of a name, address, phone number, and so on, through handwriting using a pen may be convenient compared with the use of keyboards. In order for such a system to be practical, any person's handwriting will have to be recognized accurately and speedily. This target is much harder to meet than OCR (Optical Character Recognition) systems, which recognize printed characters. At present, several recognition algorithms are being developed on 32-bit low-power microprocessors.

With regard to voice recognition, the present system can recognize very limited numbers of words spoken by persons whose spoken words are recorded beforehand. The main target of present development is the realization of a system that can recognize a large number of words spoken by anyone, accurately and speedily. Varieties of algorithms are being developed on digital signal processors, aiming at practical applications.

A voice synthesis system integrated on a chip has already been in practical application. The present and future target of chip development is the synthesis of a comfortable voice.

4.5 Future Trends

It can naturally be anticipated that boundaries among consumers, computers, and communications markets will disappear in the future, and new types of services will be generated. One typical example is an interactive television, which will afford video-on-demand service, home shopping, and so on. Factors accelerating such movement include the incorporation of digital technology and integrated circuit technology. Deregulation will favor the early introduction of those services.

Based upon the increase in the number of components on a chip, handheld terminals with varieties of multimedia functions will be available at a reasonable price. Integration of all or part of a personal computer, a cellular telephone, a television, and so on, will provide users with attractive services on portable terminals.

5. FUTURE PROSPECTS AND LIMITATIONS OF MICROMINIATURIZATION

5.1 Future Prospects

5.1.1 Expectations of Features Features of LSIs in the year 2000 will be anticipated as follows, if the trends up to now are simply extrapolated:

Number of components on a chip	
DRAM	4 Gbit
Digital LSI	50 M components
Operational frequency	1 GHz
Performance of a microprocessor	1000 MIPS
Chip size	25×25 mm^2

A complete desk-top computer, with additional functions, will be integrated on a single chip. Moreover, multimedia handheld terminals with attractive functions and performances are expected to be available, at an attractive price.

5.1.2 Technical Issues and Solutions Reduction in device feature size and increase in the number of components integrated on a chip raise the following issues:

1. Degradation of the device reliability, due to the reduction of the device feature size. Lowering the supply voltages will be effective.
2. Rapid increase in the power consumption, due to the increase in the number of components integrated on a chip. Low-power design is one of the most significant issues for the realization of high-performance, portable terminals. For this purpose, reduction of the device feature size and supply voltages, and also circuit and system designs, will be indispensable.
3. Increase in the design time, corresponding to the increase in both the complexities of the device structures and the number of components integrated on a chip. Employment of top-down design style and powerful CAD systems will be the solution. Development of the most powerful CAD system, through the selection of the CAD tools, should be carried out strategically.
4. Rapid increase of manufacturing cost. Decrease in the device feature size and accompanying increase in the complexities of the fabrication process are giving rise to exponential cost increases for equipment and facilities. It is likely that this may slow down the rate of progress of semiconductors (Moore, 1991). Furthermore, an advanced system-on-silicon chip requires the integration of varieties of functions, including digital, memory [SRAM (static RAM), EEPROM (electrically erasable PROM), or DRAM], and analog-digital mixed circuits, which tends to cause further cost increase. Efforts should be directed to the development of process technology with a reduced number of process steps, based on the overall design strategy.
5. Development of new markets and application technology.

Throughout the above activities, the necessity of effective cooperation among engineers from the system, circuit, and device/process sides is increasing.

5.2 Limitations of Microminiaturization

The great progress in Si VLSIs has been realized by MOSFET miniaturization based on scaling theory. High speed, low power, and high packing density have resulted from scaling down MOSFET dimensions by a factor of k, as well as supply voltage. The scaling theory is still considered to be a good guideline for designing future MOSFETs.

However, as a result of several physical and practical limitations, it will not be a perfect rule for very small MOSFETs. Approaches for pushing the limits are becoming more important when the design rule is less than a quarter micron.

Since the 1970s and 1980s, there have been many discussions on the limits of digital electronics (Fukuma, 1988; Meindl *et al.*, 1993). In the 1970s, "0.25 μm" seemed to be a minimum design rule, and, in the 1980s, "0.1 μm" was believed to be a limit for room-temperature operation. However, now we have an experimental 256-Mbit DRAM using 0.25-μm MOSFETs (Sugibayashi *et al.*, 1993) and a room-temperature test vehicle using ~0.1-μm MOSFETs (*IEEE IEDM Digest*, 1993). In this section, the future prospects of MOS VLSIs are reviewed, emphasizing several limitations based on modern knowledge.

5.2.1 Physical and Practical Limitations

Although scaling theory provides a guideline for reducing MOSFET size, it does not give optimum MOSFET parameters. Hence, some parameters that optimize VLSIs may be inconsistent with scaling theory. This means that, in addition to purely physical limitations, practical considerations also seriously hamper device miniaturization. Here, several limiting factors and countermeasures are discussed.

5.2.1.1 Tunneling Effects. When the gate-oxide thickness is reduced to 2–3 nm, direct tunneling between the gate and the substrate occurs. Similarly, when the channel length approaches 10 nm, tunneling between source and drain occurs. MOSFETs with leakage current due to such tunneling phenomena will not be acceptable for VLSI circuits where MOSFETs are treated as switches. As long as SiO_2 and Si are used for the gate insulator and the substrate, respectively, the minimum sizes mentioned above are the absolute limits in MOSFET miniaturization.

5.2.1.2 Hot-Carrier Effects. In order to increase the operational speed of VLSI circuits, higher supply voltage is preferable. Increasing the supply voltage, however, induces many hot carriers near the drain. These hot carriers generate states at the Si–SiO_2 interface and/or may be trapped in the gate oxide. Interface states and trapped carriers lead to long-term characteristic degradation in MOSFETs, so that the hot-carrier effect is one of the most serious limiting factors for maximum supply voltage. Note that MOSFET characteristic degradation is observed even when the supply voltage is much lower than the Si–SiO_2 barrier height or threshold voltage (energy) for interface state generation. This is because several modes of carrier scattering provide some high-energy carriers. Since electron mobility is larger than hole mobility and since the Si–SiO_2 barrier height for an electron is lower, characteristic degradation is more severe for *n*-channel MOSFETs.

According to scaling theory, the supply voltage should be scaled down by the same scaling factor, k, as that for MOSFET dimensions, to keep the inner electric field constant. However, because of nonstationary carrier transport, the carrier energy is much lower than that predicted from the local electric field. As a result, the maximum supply voltage allowed, $V_{DD,\max}$, is almost proportional to $k^{-0.5}$ even if conventional single-drain structure is used. Recent experiments have shown a maximum supply voltage of 1.5 V for 0.1-μm *n*-channel MOSFETs; therefore, 0.5 V may be expected for 0.01-μm MOSFETs.

5.2.1.3 Subthreshold Leakage Current (Threshold Voltage Scaling). Drain current in the subthreshold region is proportional to $\exp(qV_G/nkT)$. Here, V_G is gate voltage, and other symbols have conventional meanings. Since we need sufficiently low leakage current at $V_G = 0$ V for dynamic circuit operation and low standby power dissipation, the threshold voltage V_T has to be greater than 0.5 V at room temperature; i.e., V_T cannot be scaled down. In this case, however, the gate delay τ_{pd} increases when the design rule is less than 0.1 μm. Because the gate delay is proportional to $V_{DD}/(V_{DD} - V_T)^m$, the lower supply voltage required by hot-carrier reliability considerations cancels out the MOSFET miniaturization merits with respect to speed. To overcome this problem, we have to accept relatively large off-state current due to the lower threshold voltage. Although we will not be able to employ dynamic-type circuits, the standby power issue will be solved by introducing a sophisticated power-management system.

If low-temperature operation can be employed, threshold voltage can be reduced without any leakage current increase. For instance, at liquid nitrogen temperature (77 K), a threshold voltage of 0.1–0.2 V is possible. A compact cooling system is a breakthrough for realizing high-performance sub-0.1-μm CMOS VLSIs.

5.2.1.4 Punchthrough Phenomenon (Short Channel Effect). Generally the "punchthrough" of a MOS transistor refers to the state in which the depletion regions on the drain and source sides touch each other. In this case, a large current will flow through the transistor, and the device ceases normal operation. Punchthrough occurs for a large applied voltage and/or small device dimensions.

Hence, the width of the depletion-layer spread from source and drain should be scaled down to suppress the punchthrough phenomenon. Ideally, we need to increase the substrate impurity concentration by a factor of k and to decrease the effective substrate bias $V_{\text{bi}} + V_{\text{sub}}$, where V_{bi} is built-in potential and V_{sub} is the actual substrate bias. However, V_{bi} related to the Si band gap cannot be scaled down, and, therefore, the punchthrough phenomenon is extremely serious when the design rule is less than 0.1 μm. To reduce the effective substrate bias, forward biasing on the substrate with reducing operation temperature is attractive. This corresponds to virtual band-gap scaling. Another approach is to use extremely thin Si film on an insulator (SOI: Silicon On Insulator). If the Si film thickness is comparable with the inversion-layer thickness, the potential in the whole Si film is fully controlled by the gate electrode. Then, the punchthrough phenomenon is well suppressed.

5.2.1.5 Parasitic Elements. Resistances and capacitances associated with wiring are becoming dominant factors in determining the limits of VLSI operational speed. Both (a) wiring resistance per unit length and (b) contact resistance are proportional to k^2, while (c) MOSFET resistance in the ON state is constant. Thus, the sum of terms (a) and (b) will be equal to the value of (c) for the design rule around 0.1 μm. Finer design rule will make the former terms larger. New materials such as Cu and Ag are expected to reduce wiring resistance. To reduce the contact resistance, increasing effective contact area by silicidation may be more practical than reducing specific contact resistance by new materials.

On the other hand, wiring capacitance is not simply scaled down because of coupling with neighboring wires and because of longer wirings necessary in a large chip. New interlayer dielectrics with low permittivity such as organic materials will help this situation. However, chip architectures that avoid long wiring are more important.

5.2.1.6 Power Dissipation. Power dissipation for CMOS circuits is expressed by $P = V_{DD}^2 \Sigma fC$, where f is frequency and C is node capacitance. The ideal scaling does not change the power dissipation per unit area. However, in the practical case, power dissipation density increases with smaller design rules, because possibly high supply voltage is adopted for high-speed operation, and wiring capacitance is not ideally scaled down.

The maximum power dissipation of a chip, determined mainly by packaging technology, is around 30 W for an air-cooling system. Present high-performance processors have almost reached this limit. Therefore, the power dissipation is an important issue for increasing operational speed, even at present.

Several approaches may be used to overcome this difficulty, e.g., reducing wiring capacitance by sophisticated layout tools and/or interlayer dielectric material with low dielectric constant; reducing supply voltage and clock frequency with massively parallel architecture; using memory-oriented architecture; using dynamic power and clock control; and improving package.

5.2.1.7 Fluctuations. Fluctuation in device structures caused by poor process controllability or statistical randomness is one of the most serious practical limiting factors for miniaturization. For example, when the design rule is 0.1 μm (10^7–10^8 MOSFETs can be integrated), the standard deviation allowed for the effective channel length is estimated to be about 3 nm in order to realize reasonable production yield. Note that the effective channel length is determined by the controllability of various process steps, including the lithography for the polycrystalline silicon as a gate electrode, etching, source/drain implantation, annealing, and so on. At present, the standard deviation of 3 nm seems to be a very difficult target. Another example is the threshold voltage fluctuation caused by statistical fluctuation of impurity distribution. Since the total number of substrate impurities involved in each MOSFET

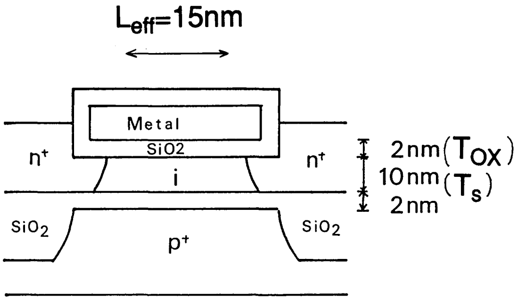

FIG. 9. Cross-sectional view of an ultimately scaled-down *n*-MOSFET.

is scaled down by k^{-1}–k^{-2}, it rapidly decreases with the decrease in the design rule. Then the distribution becomes Poisson type, and the fluctuation is proportional to the square root of the total impurity number, which results in large fluctuation in the threshold voltage. This statistical fluctuation cannot be completely controlled.

To overcome these fluctuation problems, it is essential to develop new processes with high controllability, and new devices with low process sensitivity. New circuits and architectures including fault-tolerant technique will be appreciated as well.

5.2.2 Ultimately Scaled-Down MOSFET and Future Prospect of VLSIs Among the limitations discussed in Sec. 5.2.1, the tunneling effect is the final one in determining the minimum physical size. The smallest possible value of the oxide thickness is ~2 nm, and the effective channel length is ~10 nm.

Figure 9 illustrates an ultimately small *n*-channel MOSFET. The MOSFET is based on SOI structure, and the thickness of the active Si layer is designed to be equal to that of the inversion layer. To avoid the impurity fluctuation problem, this Si layer is intrinsic type. The threshold voltage adjustment must be done by designing Si and SiO_2 thicknesses and/or by controlling a bias on the p^+ substrate. The p^+ substrate is also effective to prevent punchthrough. The maximum supply voltage will be ~0.5 V from the hot-carrier reliability consideration. This value is reasonably high even at room temperature for avoiding thermal noise problems. Figures 10(a) and 10(b) show the simulated electrical characteristics of an *n*-channel MOSFET with L_{eff} = 15 nm (L_{gate} = 20 nm), at 77 K. Sufficiently large drain current and high on/off ratio are predicted. Estimated propagation delay time is 1–2 ps.

To realize VLSIs using the MOSFETs shown in Fig. 9, many problems mentioned in Sec. 5.2.1 must be solved. Some of these problems might be overcome individually by countermeasures described in that section. However, difficulties concerned with power dissipation and impurity fluctuations are is-

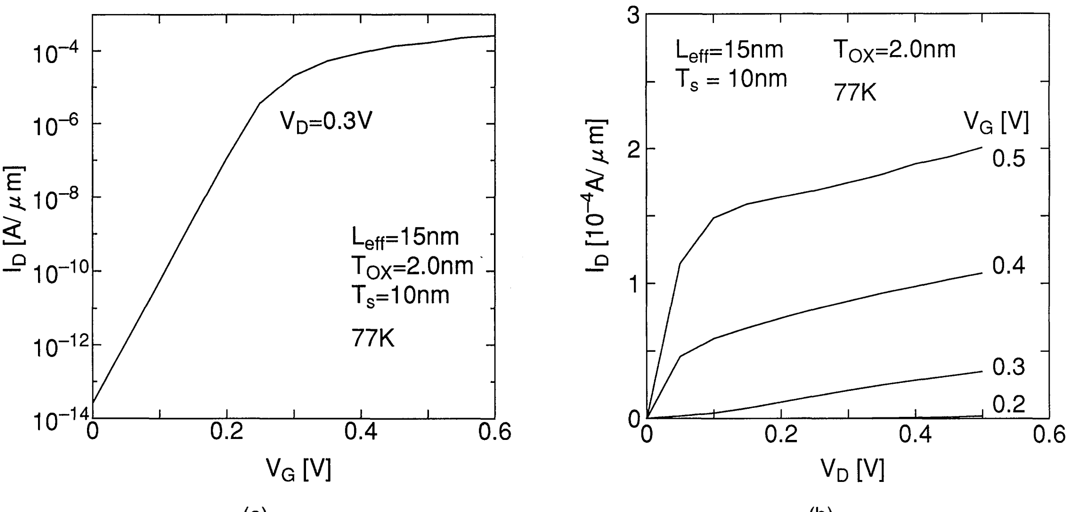

(a) (b)

FIG. 10. Simulated electrical characteristics of the *n*-MOSFET shown in Fig. 9, at 77 K: (a) drain current vs gate–source voltage, (b) drain current vs drain–source voltage.

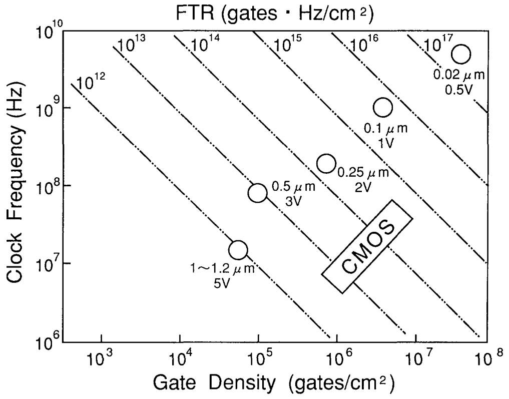

FIG. 11. Estimated clock frequency vs gate density for CMOS microprocessors, consisting of the transistors shown in Fig. 9.

sues of trade-off between performance and the practical situation, including economics. This means the necessity of comprehensive optimization from the system viewpoint. In particular, parallel processing and fault-tolerant technique will be important breakthroughs. Once the optimization is successful, progress similar to that shown in Fig. 11 might be expected. Here, a CMOS microprocessor has been taken as an example. A functional throughput rate (FTR) of 10^{17}–10^{18} may be expected with these ultimately scaled-down MOSFETs, where

$$\text{FTR} = (\text{gate density}) \times (\text{clock frequency}).$$

GLOSSARY

ASIC (Application-Specific Integrated Circuit): An IC designed for application to some specific systems or markets, as opposed to standard products, such as memories or microprocessors.

ASSP (Application-Specific Standard Product): An IC that is widely used in some specific market. Examples include a floppy-disk controller.

B-ISDN (Broadband Integrated Services Digital Network): A public end-to-end digital communications network that supports a wide range of services, such as audio, voice, video, and data, over standard interfaces.

CD-ROM: Compact-Disc ROM (Read-Only Memory), that is, a laser-encoded disc that in a typical format stores 540 Mbytes (540 × 8 Mbit) of randomly accessible video, audio, and/or text data.

Cellular Telephone: A mobile and portable radio telephone service that uses networked stations or "cells."

CISC (Complex Instruction Set Computer): A computer in which individual operations carry out elaborate data processing. It is much slower than RISC.

CMOS (Complementary symmetry MOS): A circuit type that employs both *n*-channel and *p*-channel MOS transistors. It has a favorable feature of very low power consumption.

Depth of Focus: An amount of defocus tolerance within which the image shows a quality to satisfy the specifications.

Design for Reliability (DFR): Chip design so as to realize high reliability.

Design for Testability (DFT): Chip design so as to make chip testing easier.

ESD (Electrostatic Discharge): Potentially damaging discharge into the pins of an IC from an outside source, including a person's hands.

MPEG: Moving Picture coding Experts Group. A technical group investigating international standards on the compression and decompression of video and audio signals.

Multimedia: Delivery of information that combines different content formats (motion video, still pictures, graphics, animation, audio, text, and so on).

RISC (Reduced Instruction Set Computer): A computer based on a set of simplified instructions. Instructions' simplicity affords higher throughput, compared with CISC.

Virtual Reality (VR): A virtual environment generated by computers around a person, based on information sensed by a person himself.

Works Cited

Adam, J. A., Cole, B. (1993), "Special Report on Interactive Multimedia," *IEEE Spectrum* **30** (3), 22–39.

Antoniadis, D. A., Hansen, S. E., Dutton, R. W., Gonzalez, A. G. (1977), *SUPREM 1—A program for IC process modeling and simulation*, Stanford University Stanford Electronics Labs. Technical Report No. 5019-1.

Chandrakasan, A. P., Sheng, S., Broderson, R. W. (1992), "Low-Power CMOS Digital Design," *IEEE J. Solid-State Circuits*, **SC-27**, 473–484.

Comerfold, R., Mogal, J. (1993), "Focus Report and Guide to Engineering Work stations and PCs," *IEEE Spectrum* **30** (5), 37–79.

Engl, W. L. (Ed.) (1986), *Process and Device Modeling*, Amsterdam: North-Holland, Elsevier Science Publishers B.V.

Fischetti, M. (1993), "The Cellular Phone Scare," *IEEE Spectrum* **30** (6), 43–47.

Fukuma, M. (1988), "Limitations on MOS ULSIs," *Technical Digest of Symposium on VLSI Technology*, 7–8.

Gelsinger, P. P., Gargini, P. A., Parker, G. H., Yu, A. Y. C. (1989), "Microprocessors circa 2000," *IEEE Spectrum* **26** (10), 43–47.

Gunning, B., Yuan, L., Nguyen, T., Wong, T. (1992), "A CMOS Low-Voltage-Swing Transmission-Line Transceiver," *ISSCC Digest of Technical Papers*, 58–59.

Heilmeier, G. H. (1992), "Personal Communications: Quo Vadis," *ISSCC Digest of Technical Papers (Invited)*, 24–25.

Kobb, B. Z. (1993), "Personal Wireless," *IEEE Spectrum* **30** (6), 20–25.

IEEE ISSCC Digest of Technical Papers, from 1970 to 1993, and the *Digest of Technical Papers of the Symposium on VLSI Circuits*, from 1987 to 1993.

Itoh, K. (1990), "Trends in Megabit DRAM Circuit Design," *IEEE J. Solid-State Circuits*, **SC-25**, 778–789.

Kaplan, G., Gerla, M., Sharma, R. L., Delisle, D., Pelamourgues, L. (1991), "Special Guide to Data Communications: High-Speed Networks and Interconnections," *IEEE Spectrum* **28** (8), 21–44.

Levitt, M. E. (1992), "ASIC Testing Upgraded," *IEEE Spectrum* **29** (5), 26–29.

Meindl, J. D., De, V. K., Agrawal, B. (1993), "Prospects for Gigascale Integration (GSI) Beyond 2003," *ISSCC Digest of Technical Papers* 124–125.

Moore, G. E. (1991), "Semiconductor Technology Reaches Middle Age," *IEEE Comput. Mag.*, 105.

Nagel, L. (1975), "SPICE2: A Computer Program to Simulate Semiconductor Circuits," University of California, Berkeley, ERL Memo ERLM75/520.

Rosenblatt, A., Watson, G. F., Shina, S. G., Hall, D., Reddy, R., Wood, R. T., Cleetus, K. J., Turino, J., Wheller, R., Burnett, R. W. (1991), "Special Report on Concurrent Engineering," *IEEE Spectrum* **28** (7), 22–37.

Sangiovanni-Vincentelli, A. L. (1980), "Circuit Simulation," in: P. Antognetti, D. O., Pederson, H. De Man (Eds.), *Computer Design Aids for VLSI Circuits*, Alphen aan den Rijn, The Netherlands: Sijthoff & Noordhoff, pp. 19–112.

Santo, B. (1989), "BiCMOS Circuitry: the Best of Both Worlds," *IEEE Spectrum* **26** (5), 50–53.

Selberherr, S. (1984), *Analysis and Simulation of Semiconductor Devices*, New York: Springer-Verlag.

Sugibayashi, T., *et al.* (1993) , "A 30ns 256Mb DRAM with Multi Divided Array Structure," *ISSCC Digest of Technical Papers*, 50–51.

Sze, S. M. (1988), *VLSI Technology*, 2nd Ed., New York: McGraw-Hill.

Takeda, E., *et al.* (1993), "VLSI Reliability Challenges: From Device Physics to Wafer Scale Sysems," *Proc. IEEE*, **81** (5), 653–674.

Waxman, R., Saunders, L., Carter, H. (1989), "VHDL Links Design, Test, and Maintenance," *IEEE Spectrum* **26** (5), 40–44.

Yürgan, R. K. (1992), "Digital Video," *IEEE Spectrum* **29** (3), 24–30.

Further Reading

Hennessy, J. L., Jouppi, N. P. (1991), "Computer Technology and Architecture: An Evolving Interaction," *IEEE Comput. Mag.*, 18.

Mavor, J., Jack, M., Denyer, P. (1983), *Introduction to MOS LSI Design*, London: Addison-Wesley.

Mead, C., Conway, L. (1980). *Introduction to VLSI Systems*, Reading, MA: Addison-Wesley.

Sze, S. M. (1981), *Physics of Semiconductor Devices*, 2nd Ed., New York: Wiley.

Weste, N. H. E., Eshraghian, K. (1985), *Principles of CMOS VLSI Design; A Systems Perspective*, Reading, MA: Addison-Wesley.

MICROLITHOGRAPHY

Henry I. Smith, *Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*

	Introduction	281
1.	**Basic Principles of Microlithography and Resists**	281
1.1	The Planar Process	281
1.2	Lithography, A Binary Process	283
1.3	Positive-Tone and Negative-Tone Resists	283
1.4	Approaches to Lithography with Photons, Electrons, and Ions	284
2.	**Photolithographic Shadow Printing**	285
3.	**Optical-Projection Lithography**	286
3.1	The Push to Higher Resolution	287
4.	**Scanning-Electron-Beam Lithography**	288
4.1	Writing Strategies	289
4.2	Writing Time vs Linewidth	289
4.3	Field Stitching	290
4.4	Reducing Exposure Time	291
4.5	Electron Scattering	291
5.	**X-Ray Lithography**	291
5.1	Sources	291
5.2	Masks	292
5.3	Results	293
5.4	Alignment	293
6.	**Conclusions**	294
	Glossary	294
	Works Cited	294
	Further Reading	295

INTRODUCTION

This brief review of microlithography focuses on the basic principles of the technology, and the methods which have proven most effective in forefront research and commercial applications, most notably the fabrication of microelectronic and optoelectronic devices. These methods are photolithographic shadow printing, optical-projection lithography, scanning-electron-beam lithography, and x-ray lithography.

1. BASIC PRINCIPLES OF MICROLITHOGRAPHY AND RESISTS

1.1 The Planar Process

Figure 1 illustrates a process, known as the planar fabrication process, whereby a complex patterned structure can be created in or on the surface of a substrate. As depicted, it proceeds in four stages. A substrate to be patterned is first coated with a film, commonly called a "resist," and this resist is exposed to a pattern of radiation that can be visible or ultraviolet photons, x-ray photons, electrons, ions, or even neutral atoms if they have sufficient energy. As discussed below, the chemical alteration of the resist produced by the irradiation is used to produce a relief structure or "relief image" in the resist layer. Thereafter, this relief image can be used as a mask for subsequent modification of the substrate. As depicted in Fig. 1, this modification can be

1. an etching of the substrate in unprotected areas by a chemical solution, a reactive gaseous plasma, or energetic atomic or ion bombardment;
2. a growth of material up through the interstices in the resist by a process such as electroplating;
3. a patterned doping via diffusion of impurity atoms or by ion implantation into areas unprotected by the resist coating; or
4. a patterning of deposited material through a process known as liftoff.

In the latter case, material is deposited over

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

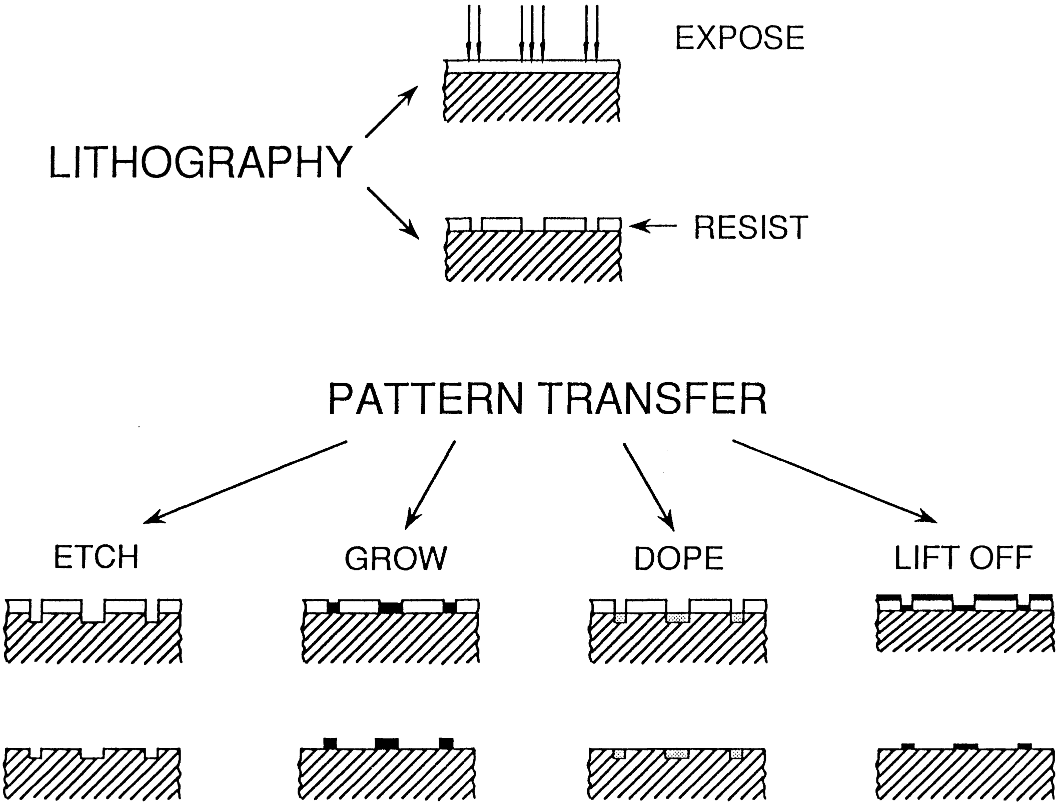

FIG. 1. Schematic of the planar fabrication process.

the entire relief image, and then, after removal by dissolution or etching of the resist, a patterned deposit is left behind.

The basic concept inherent in the planar process was actually invented well over 1000 years ago by neolithic people (Haury, 1967). The Hohokam Indians of the American Southwest had a thriving agricultural civilization which lasted from about 300 B.C. to 1400 A.D. As a resist, the Hohokam used a pitch obtained from the mesquite plant to coat the substrates, which were commonly seashells obtained in trade with native Americans living along the seacoast. The Hohokam were not able to image photons or charged particles, of course, but instead produced the relief image in the resist by mechanical means, such as scratching. A weak acidic acid solution produced from the blossoms of the saguaro cactus enabled them to etch relief images into the seashell substrates. This process, which is usually called lithography, from the Greek *lithos*, meaning stone, and *graphein*, meaning to write, was later reinvented by the Europeans and used for such things as decorating armor and, many years later, for producing plates used in the printing of graphics. It is a curious irony of technological history that the very first photographic image, by Nicephore Niepce in 1826, was actually produced in a photosensitive polymer (bitumen of Judea) and made permanent by chemical etching into a substrate, effectively duplicating the planar process of Fig. 1, not the silver-halide-based process that we now call photography. The planar process of Fig. 1 is used in a wide range of industries from printing and the production of mechanical parts to the production of the microelectronic "chips" that are the basic memory and logic components of computers.

From the point of view of etymology, the term *lithography* should refer to the entire planar process, as depicted in Fig. 1. However, today it is more common that the terms lithography, microlithography, and nanolithography refer only to the process of producing the relief image in the resist film; the subsequent stages are referred to as "pattern transfer." We will follow this usage in the rest of this article.

The printing industry also uses the term lithography for the process in which a pattern is formed in an organic material that repels water-based inks on a substrate to which the inks adhere. Pressing the substrate onto a paper sheet transfers the ink pattern onto the paper.

1.2 Lithography, A Binary Process

The term "microlithography" is used to designate the creation of structures in resist with minimum feature sizes from a few micrometers to 0.1 μm (100 nm). The terminology "nanolithography" is widely used in reference to sub-100-nm features, and we will follow this custom here.

In this article, we refer frequently to "resolution," a term used rather loosely to designate the minimum size of features (lines, or openings) or the minimum "spatial period" or "pitch" achievable with a given lithographic technique. The minimum-pitch criterion is preferred because it is directly relatable to resolution criteria, such as the Rayleigh criterion, used with optical imaging systems (i.e., telescopes, camera, microscopes). Although not all applications require the highest resolution, the following discussions will be framed around the systems and methods that yield the highest practical resolution.

It is important to emphasize that lithography is a binary process and hence is quite distinct from photography. By this we mean that the relief image in the resist is either "on" or "off"; i.e., the substrate is either protected by the resist mask or not. In photography, on the other hand, the task is to record scenes from the world around us, which involves a wide range of light intensities. As a result, a photographic process must represent "shades of gray" and, in some cases, colors. In lithography, there are no shades of gray, only "black and white," that is, "on" and "off."

1.3 Positive-Tone and Negative-Tone Resists

A positive resist is one in which the areas irradiated by photons, ions, or electrons are removed, and the unirradiated areas remain as the relief image in resist. With a negative-tone resist, the relief image corresponds to the irradiated areas. Figure 1 depicts the operation of a positive resist. A wide range of resists is in common use, and an even wider range of resists has been the subject of experimentation and research. We will discuss only a few, those which have found wide application.

Perhaps the simplest resist to discuss is polymethyl methacrylate (PMMA), a common polymer, known under various trade names as Plexiglas, Lucite, etc. The polymer PMMA, as it is commonly called, is sensitive to ionizing radiation, that is, photons with wavelengths below about 200 nm (6.2 eV), electrons, and ions. For lithography purposes, the PMMA molecular weight is usually between about 400 000 and 950 000. The effect of the ionizing radiation is to break bonds, leading to chain scission and reduction of molecular weight in irradiated areas. Thereafter, a weak solvent can be used to remove irradiated areas selectively. Some results (Hawryluk *et al.*, 1975) are shown in Fig. 2,

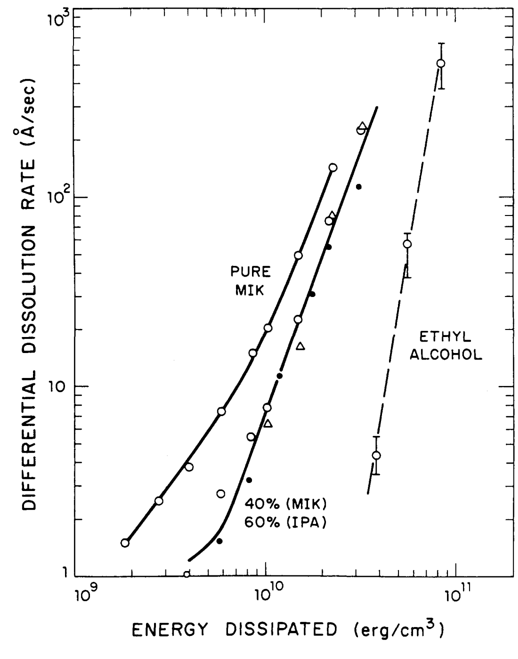

FIG. 2. Plot of the development rate of polymethyl methacrylate resist as a function of absorbed energy. For a developer solution consisting of a 40/60 mixture of methyl isobutyl ketone (MIK) and isopropyl alcohol (IPA), there is a cubic dependence, which implies a high-contrast resist.

which is a logarithmic plot. Note that beyond a certain threshold dose, the development rate (i.e., the rate of dissolution) in the 40/60 mixture has a cubic dependence. That is, a doubling of the irradiation dose leads to an increase in dissolution rate of a factor of 8. Such a strong nonlinearity is referred to as "high-contrast resist."

High-contrast resist response is critically important to the lithography process. Figure 1 depicts abrupt discontinuity in the irradiance distribution between exposed and unexposed regions. This is almost never the case in practice. Rather, the irradiance might have a distribution as depicted in Fig. 3. Because the ratio of $I_{\max}$ to $I_{\min}$ is ~3 to 2, rather than a more desirable 5 to 1 or 10 to 1, this distribution is said to have a "low contrast." The task of the resist response then is to convert this low-contrast irradiance distribution into a binary relief structure in resist. The "clipping" level depicted in Fig. 3 implies that an irradiance above this level leads to full development of the resist, whereas no resist development occurs for irradiance below this level. This implies a highly nonlinear or "high-contrast" resist response. This is certainly the case for PMMA, as illustrated in Fig. 2.

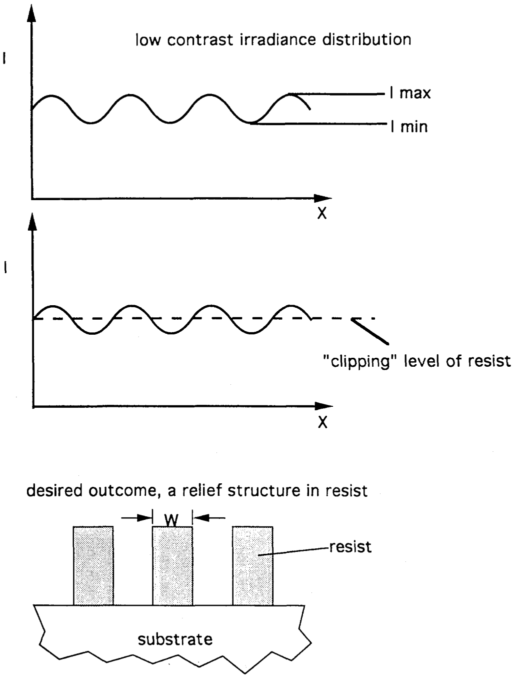
I
low contrast irradiance distribution
I max
I min
X
I
"clipping" level of resist
X
desired outcome, a relief structure in resist
W
resist
substrate

FIG. 3. Schematic illustrating the binary nature of the lithographic process. A low-contrast irradiance distribution can be converted into a binary resist structure with vertical profiles if the resist has a highly nonlinear response equivalent to "clipping" (i.e., development occurs only for irradiance above a threshold or clipping level).

An added virtue of PMMA is its resolution: Relief features as fine as 18 nm on a 36-nm pitch (centerline-to-centerline spacing) have been produced. As indicated above, PMMA is sensitive only in the deep ultraviolet (UV) and beyond. Most photolithography is done in the violet or near UV, and, for such work, a totally different resist chemistry is required.

The most common photoresists, useful in the blue and near UV, are based on a combination of the acidic polymer novolak and the photosensitive material napthaquinone diazide (known as the inhibitor). The combination is relatively insoluble in basic solutions because the inhibitor protects the novolak from attack by the base. However, with irradiation the inhibitor is transformed, allowing rapid dissolution or etching of the novolak. This family of photoresists also has high contrast, i.e., highly nonlinear response.

There are many other types of resists, the most novel being the so-called chemically amplified resists (CAR). With these, the sensitivity and resolution are controlled by a post-exposure baking which determines how far an acid group, released upon exposure, diffuses beyond the point of exposure, creating additional reactions along the way. We will not have space to discuss resists further. Suffice it to say that a resist must have a high contrast and be capable of resolving fine linewidth features.

1.4 Approaches to Lithography with Photons, Electrons, and Ions

A number of approaches can be taken to expose resist. We distinguish between two generic categories: pattern generation and pattern replication. Pattern replication is done using a reticle or mask made, in advance, by a pattern-generation instrument. Examples of pattern-generation schemes include the controlled scanning, usually under computer control, of a laser beam, an electron beam, or an ion beam over a resist-coated substrate. Holographic exposure of resist should also be considered pattern generation. Pattern replication can be done by shadow printing the reticle or mask, or by projecting its image through a lensing system, often with demagnification. Any one of the radiations—UV, x rays, electrons, or ions—can be employed to replicate a reticle or a mask. However, the structure of the mask must be quite different in each case. The following types of radiation have been used, or proposed for use, in microlithography and nanolithography instruments: UV photons (300 nm $< \lambda <$ 400 nm), deep UV (150 nm $< \lambda <$ 300 nm), very soft x rays (5 nm $< \lambda <$ 15 nm), soft x rays (0.4 nm $< \lambda <$ 5 nm), hard x rays ($\lambda <$ 0.4 nm), electrons (200 eV $< E <$ 400 keV), and ions (20 keV $< E <$ 500 keV). In most cases, these forms of radiation can be used in either the pattern-generation or the pattern-replication mode. (A notable exception: No x-ray-based pattern *generation* scheme has been proposed since x-ray beams cannot be readily focused and scanned.) Hence, one could consider perhaps 50 different types of microlithographic instruments that either exist, are under development, or have been proposed. However, most would prove to be impractical or ineffective. We will confine this exposition to only those few that have proven to be effective: photolithographic shadow printing, optical-projection lithography, scanning-electron-beam lithography, and x-ray lithography.

2. PHOTOLITHOGRAPHIC SHADOW PRINTING

Photolithographic shadow printing is the simplest type of microlithography, the most economical, and, in many cases, highly effective. The scheme is depicted in Fig. 4. When required, alignment of the mask to a pattern preexisting on the substrate is done by simultaneously viewing, with an optical microscope, complementary alignment marks on the mask and substrate, and mechanically shifting one or the other until the marks superimpose. Typically, the mask consists of the desired pattern in a metal such as chromium, or absorber such as iron oxide, on a glass or quartz plate. The mask must be made in advance by another lithographic technique, one capable of pattern generation. In shadow printing, the quality of the areal image at the resist layer depends on the magnitude of the mask-sample gap, because it is diffraction in the gap that degrades the image. If the gap is zero, i.e., "contact," the quality and resolution of the image can be extremely high since the resist would then be recording the transmitted irradiance distribution prior to any significant diffraction. In addition, the diffraction that takes place in traversing the resist is, in effect, suppressed somewhat because the resist index of refraction is typically about 1.6. As depicted in Fig. 4, mask and substrate are perfectly flat and parallel, but this is almost never the case in practice. Real substrates are not generally very flat relative to UV wavelengths. Moreover, buildup of resist occurs at substrate edges, and dust particles or other surface asperities are often present and prevent one from achieving a zero gap. Nevertheless, even with commercially available masks (which are mechanically rigid and optically flat) and typical, non-flat substrates, one can get good contact over limited areas of the substrate and thus achieve high-resolution photolithography. Many high-performance binary optical and microelectronic devices, especially microwave devices in *III-V* materials, are made in this way. However, contact photolithography is not considered a mass production technique because of damage to mask and substrate that can occur when the two are brought into hard contact.

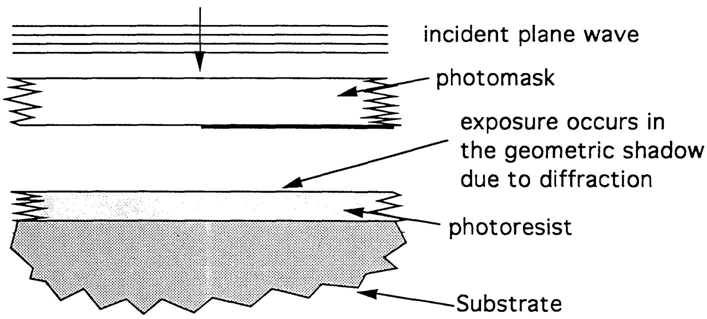

FIG. 4. Schematic of photolithographic shadow printing.

The problem of damage can be avoided through the use of flexible or conformable photomasks. These are made on thin glass, similar to that used in microscope cover glass. Damage is avoided, and contact can be achieved over a much larger fraction of the substrate. Special provisions must be made to avoid loss of contact by the nitrogen evolution that occurs during exposure of novolak-diazide types of photoresist. Extremely high-resolution photolithography is achievable with conformable photomask lithography (Smith, 1974; Everett *et al.*, 1991).

3. OPTICAL-PROJECTION LITHOGRAPHY

Figure 5 is a schematic of an optical-projection lithography (OPL) system. An image of the mask or "reticle" is cast onto a resist-coated substrate. For reasons that will become clear below, the mask pattern is depicted as a simple grating. Figure 5 depicts in-line imaging with a single refractive lens. Actually, many types of projection systems are in use. Some project the reticle image at the same scale (1-to-1 systems), others provide demagnification (typically 5× or 10×). Some use entirely refractive optics, others primarily reflective optics. Reduction refractive systems are probably the most widely used. We limit our discussion to such systems.

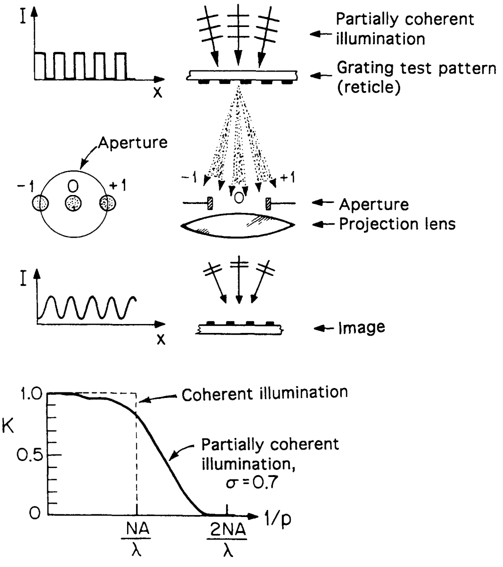
I
x
Partially coherent illumination
Grating test pattern (reticle)
Aperture
-1
0
+1
-1
0
+1
Aperture
Projection lens
I
x
Image
1.0
K
0.5
0
Coherent illumination
Partially coherent illumination, σ=0.7
NA/λ
2NA/λ
1/p

FIG. 5. Schematic of an optical-projection lithography system. The reticle is depicted as a binary transmission grating. The illumination is monochromatic, but the cone of illumination partially fills the entrance aperture, so-called partially coherent illumination. The image, having pitch or spatial period *p*, has a contrast *K* that depends on the value of *p* and the type of illumination, as given in the plot of *K* vs 1/*p*. The contrast *K* is defined as $(I_{max} - I_{min})/(I_{max} + I_{min})$.

In principle, a reduction projection system that uses refractive optics is similar to an optical microscope operated in reverse (i.e., image and object positions are reversed). In fact, optical microscopes can be, and have been, used for lithography, and with them the highest-resolution optical projection lithography has been obtained (Kawata *et al.*, 1989). It has been known since the work of Ernst Abbé in the mid-19th century that the resolution of a lens is related to its numerical aperture (NA), defined as the sine of the half angle of collection (or emission) multiplied by the local index of refraction. Specifically, the minimum resolvable pitch p is given by

$$p_{\min} = \lambda/2\mathrm{NA}, \tag{1}$$

where λ is the wavelength used. Optical microscopy gets very close to this resolution limit. During the last century, driven by the need to achieve the finest microscope resolution for medical and biological research, lenses with NAs as high as 0.95 in air and 1.4 in oil were developed. When used for lithography in a research setting, such lenses have produced minimum features of 140 nm, finer than $\frac{1}{3}$ the wavelength of the blue light used (Kawata *et al.*, 1989). Microscope lenses are not suitable for commercial lithography because of their relatively small fields of view, and distortion within the field.

The lenses used in modern optical projection systems are enormous relative to their microscope and camera forebears. Their fields of view, as measured by the number of resolvable features across them, are 10 to 100 times larger, and distortion is significantly lower. Projection lenses must be telecentric; i.e., demagnification should not change with defocus. The wavelengths used include the *G* and *I* lines of Hg (436 and 365 nm, respectively), the KrF excimer laser (248 nm), and, perhaps in the future, the ArF excimer laser (λ = 193 nm). Numerical apertures range from 0.3 to 0.6 and fields of view from 10 × 10 to

20×20 mm^2 (Okazaki, 1991). In order to achieve such enormous fields of view with low distortion and minimal aberrations, between 20 and 30 separate lens elements are required. They are assembled in a precision housing, up to 40 cm in diameter, with a total weight of over 200 kg.

The substrate "wafers" used in the production of Si integrated circuits range from 10 to 20 cm in diameter, and so, to expose an entire substrate, multiple exposures are required. Hence, the wafer must be "stepped" and realigned multiple times. For this reason, OPL systems are often called "steppers." A variety of techniques are used for aligning or positioning the reticle image on the substrate, which generally must be done to a tolerance finer than the minimum lithographic feature. The most effective of these techniques image a mask feature onto a substrate mark (which might consist of a single line, or a grating) and converge to alignment by interpreting the pattern of scattered light. The x-y position of the table upon which the wafer is held is monitored by laser interferometry, which can detect a change in stage position as small as 1 nm, although perhaps not in a production environment. Because of the high precision of table positioning, direct image alignment need be done at only a few sites on the substrate, other sites being positioned for exposure under control of the laser interferometer alone.

Because of the widespread use of OPL in Si microelectronics production, and the enormous costs and dislocations that would be caused by any major change in production technology, there is an urgent imperative to continue using OPL as the minimum feature sizes in microelectronic circuits follow their historic trend of continuous shrinkage.

3.1 The Push to Higher Resolution

The resolution of OPL can be expressed in a manner similar to Eq. (1):

$$p_{\min} = 2W_{\min} = 2k_1(\lambda/\mathrm{NA}), \qquad (2)$$

where $W_{\min}$ is the minimum feature size. The value of the parameter k_1 depends on a number of factors, including the angular distribution of the reticle illumination, the type of reticle, the image contrast one is willing to tolerate, and the required depth of focus.

The angular distribution of reticle illumination affects the image contrast at a given pitch, as illustrated in Figure 5. Traditionally, the illumination of the reticle has been uniform over about half the range of angles that could be accepted by the lens entrance pupil. This is referred to as "partially coherent illumination" with "filling factor $\sigma \simeq 0.5$." Such illumination tends to boost the image contrast for spatial periods larger than $0.8\lambda/\mathrm{NA}$ relative to the contrast one would have with incoherent illumination, which generally is used in optical microscopy. The connection between the angular distribution of illumination and contrast at a given pitch has been known since the last century. Off-axis and annular illumination was used in microscopy to boost the contrast of a specific range of spatial periods. Recently, this strategy has been investigated in OPL, where it is particularly attractive for Si integrated circuit manufacturing, which uses only a limited repertoire of shapes and sizes (Tamechika *et al.*, 1992).

A contrast K of 0.5 or higher is considered essential for production lithography. For partially coherent illumination, this corresponds to a k_1 value of about 0.4. To approach this value in production, every aspect of the OPL technology is being scrutinized, and trade-offs are made between minimum feature size and the latitude of the processing. In this engineering optimization of OPL, the reticle can also be optimized. Figure 5 assumes that only the irradiance transmission is altered. But one could also alter the relative phase of the transmitted light. In the limit of a "pure phase" mask of a grating, k_1 could be effectively reduced by the factor of 2! However, for patterns of general interest, the gain cannot be so large. The "attenuating phase-shift mask" is applicable to all pattern geometries (Lin, 1992). It provides about 10-to-1 contrast and a π phase shift at the reticle, which gives some gain in image contrast.

One can also introduce phase and amplitude filtering in the pupil plane of the projection lens (Inoue *et al.*, 1992), something that is closely related to phase-contrast microscopy introduced by Zernike in the 1930s.

In summary, OPL is a highly developed production technology and represents the forefront of optical systems engineering. It is currently in a state of intensive development in an effort to provide cost-effective commercial production of integrated circuits as minimum feature sizes are shrunk from 0.35 μm to 0.25 and 0.13 μm. Whether OPL will succeed in this before competitive technologies such as x-ray lithography is a major question of the day.

4. SCANNING-ELECTRON-BEAM LITHOGRAPHY

Electron-beam lithography (EBL) systems can take many forms, from those that scan a beam of circular cross section and sequentially fill in a pattern shape to those that project an entire pattern using a reticle mask. The most successful and widely used are the "round beam" and "shaped beam" scanning systems. The latter generally project a rectangular shape and scan it to fill in a pattern more efficiently than a round beam can. Figure 6 illustrates schematically the main features of a scanning-electron-beam lithography (SEBL) system. The similarity to a scanning electron microscope (SEM) is evident. Many research groups have, in fact, modified SEMs to serve as SEBL systems, but these have serious limitations in accuracy and speed, which make them inadequate for all commercial and many research applications.

Because a SEBL system can also function as a SEM, alignment of patterns relative to marks on the substrate (or a stage) is relatively straightforward. One merely views the alignment mark and adjusts the position of the scan field, either electronically or mechanically, so that the mark has the desired location in the scan field. Thereafter, a laser-interferometer-controlled stage moves the substrate to desired *x* and *y* locations. This viewing of alignment marks and establishment of a coordinate system are done under computer control.

The electron gun must provide a stable emission, and as high a brightness as possible, as this directly determines current density in the focal spot on the sample, which in turn determines exposure time. Thermal-field-emission sources (see ELECTRON MICROSCOPY), which achieve brightnesses approaching 10^8 A/cm^2 sr, are preferred. The lenses demagnify the source onto the sample. The objective lens usually has a relatively large bore to enable a large field to be scanned. The design of the deflection coils also influences how large a field can be scanned with minimal distortion. The role of the beam blanker is to turn the beam on and off as needed to create the desired pattern. A laser interferometer can determine stage position to better than 10 nm. If a stage position error occurs, a correction signal is fed to the electron deflector.

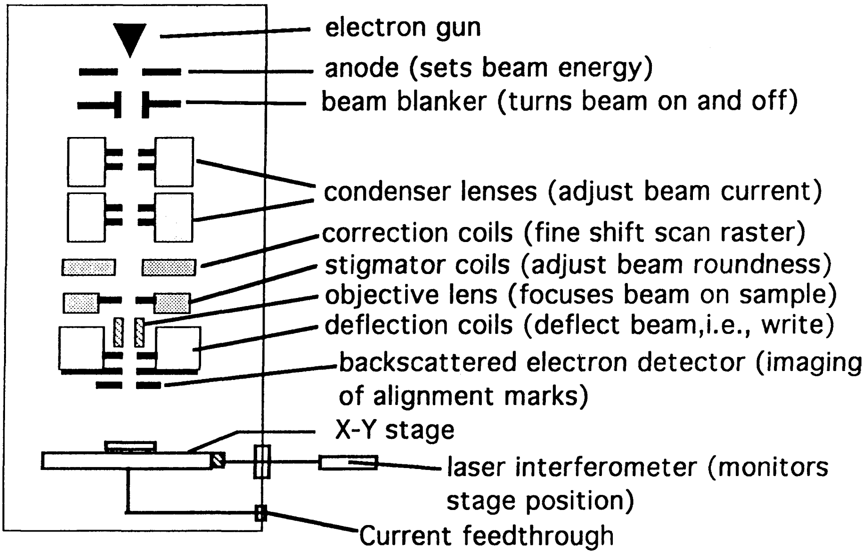
electron gun
anode (sets beam energy)
beam blanker (turns beam on and off)
condenser lenses (adjust beam current)
correction coils (fine shift scan raster)
stigmator coils (adjust beam roundness)
objective lens (focuses beam on sample)
deflection coils (deflect beam,i.e., write)
backscattered electron detector (imaging of alignment marks)
X-Y stage
laser interferometer (monitors stage position)
Current feedthrough

FIG. 6. Schematic of a scanning-electron-beam lithography system.

4.1 Writing Strategies

It is convenient to subdivide any pattern that is written into square "pixels" with edge dimension $\frac{1}{5}$ the minimum feature size of interest W, as shown in Fig. 7, and to assume that the pixel size ϵ is equal to the minimum beam diameter d_o. The writing of a pattern is done within a scan field that is almost always much smaller than the substrate that is being written on. The scanning beam addresses pixels in sequence under the control of a digital computer.

It is appropriate to describe the size of a scan field in terms of the number of pixels it encompasses. If we then assume that the written pattern elements cannot be allowed to deviate from their assigned positions by more than one pixel, for example, this fixes the maximum field size. In other words, the maximum "distortion free" field size is dictated by the requirement that the pixel grid not deviate from a perfect Cartesian grid by more than ϵ.

Two factors determine how many pixels span the maximum allowable field: the electron optics and the digital electronics. It is difficult to design an objective lens and deflector that can meet the above field size criterion at a field size that is much more than 10^4 pixels across. We might then consider a 14- to 16-bit digitally specified field to be the limit imposed by the electron optics, although this is not a truly fundamental limit. Similarly, although probably not a fundamental limit, it is difficult to construct a digital control system that can scan a 16-bit field with 1-bit accuracy at high speed. Nevertheless, we will take 65 536 pixels as the limit of the scan field size, realizing that current SEBL scan fields are generally $\frac{1}{2}$ to $\frac{1}{4}$ this size.

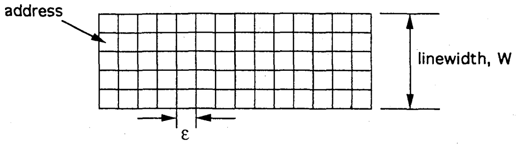

FIG. 7. The "pixel model," useful in discussing SEBL. The minimum linewidth W is subdivided into square pixels, each approximately equal in edge dimension ϵ to the beam diameter d_o. Each pixel is separately addressed by the scanning beam.

For a minimum beam diameter of 10 nm, with a corresponding minimum feature size of 50 nm, the maximum scan field size is 0.65 mm.

Within a scan field, writing can be done in a raster fashion, as in a TV, or by vector scanning. In raster scanning, all points are addressed, and the beam is blanked or unblanked as needed. In vector scanning, only the parts that make up a desired pattern are addressed. Vector scanning is more efficient from an exposure-time viewpoint but is more difficult from a control point of view, and its exposure-time advantage tends to disappear for very dense patterns.

A variant on the raster scanning strategy, which has been highly successful in SEBL systems designed for mask making, is raster scanning with the stage continually moving under laser interferometer control (Abboud *et al.*, 1992). The scanning pattern on the sample is thus serpentine. Because of the dependence on a laser interferometer for pattern integrity, the scan field (which is just a line scan along one axis) need not be very large. It is typically only 10^3 to 10^4 pixels across, which implies simpler electron optics and control electronics.

4.2 Writing Time vs Linewidth

When the pixels that constitute a pattern are addressed in a serial fashion, writing time can quickly become excessive as sample area and pattern complexity are increased, which is the trend in modern applications of microlithography. The dwell time per pixel, t_P, is given by $t_P = S/J$, where S is the resist sensitivity expressed in charge per unit area and J is current density in the beam. With thermal-field-emission sources, current densities of 1000 A/cm^2 and somewhat higher are realizable. Resist sensitivities (i.e., the amount of charge input per unit area required to achieve full resist development) of 10^{-6} C/cm^2 are realizable, giving $t_P = 10^{-9}$ s. However, this ignores the stochastic nature of the current, i.e., shot noise in the beam (Smith, 1986, 1988). If we assume that the mean number of electrons per pixel is some number (e.g., 100) in order that the probability of error in pixel exposure be below some given value, then we can determine the minimum possible writing time for a given current density. This is given in Fig. 8. The writing time per unit area

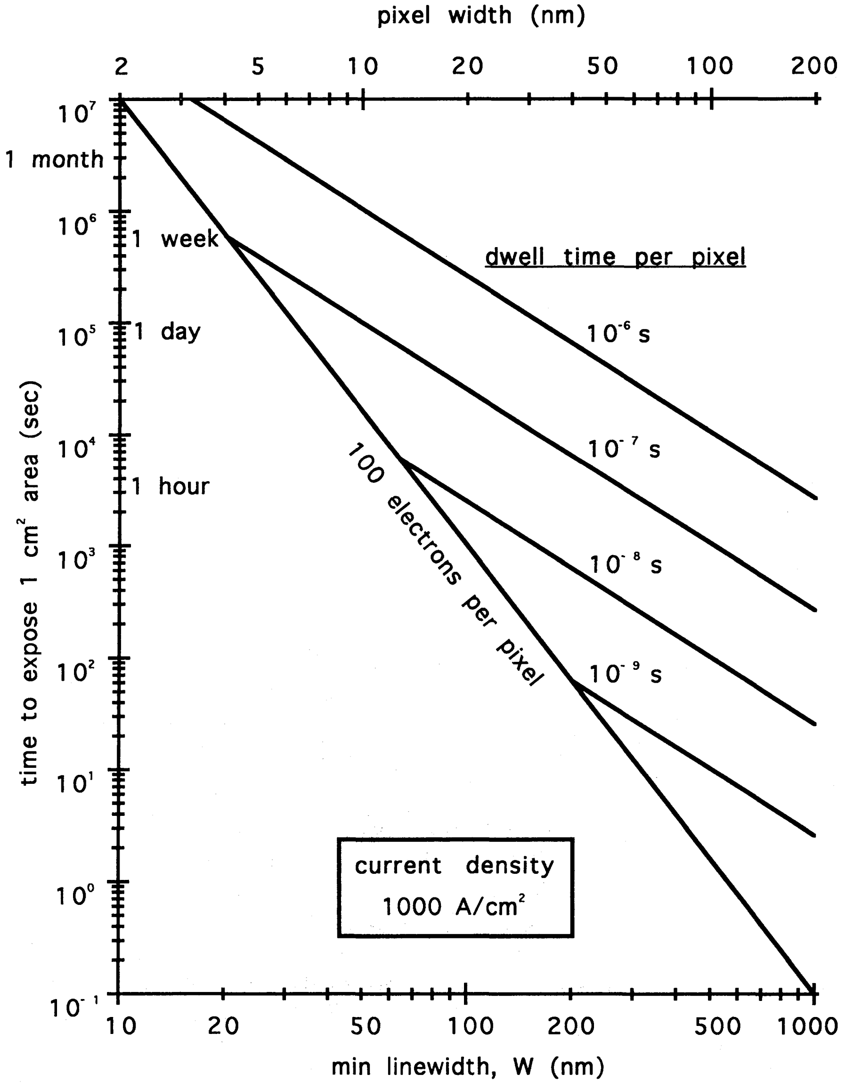

FIG. 8. Plot of *e*-beam writing time for a 1-cm^2 area as a function of minimum linewidth for various dwell times per pixel, assuming a current density of 1000 A/cm^2, and that every pixel is addressed. The line marked 100 electrons per pixel represents an assumed shot-noise limit.

is limited either by the speed with which the pixels can be addressed or by fundamental shot-noise considerations. One can, of course, make different assumptions with regard to current density or minimum numbers of electrons per pixel, but the general conclusion will still be the same: Writing time can easily get out of hand for complex patterns, and so, for industrial applications, SEBL should be considered primarily a mask-making tool.

4.3 Field Stitching

Because field size in SEBL is small compared to most sample areas of interest, fields must be stitched together. Generally, this is done via laser interferometer control of stage position. To achieve distortion free patterns, the scan field axes, sample axes, and stage axes have to be made collinear to within less than one pixel over the scan field, or other suitable corrections made.

A fundamental problem in all forms of SEBL is that they are run "open loop." That is, the electron beam can drift with respect to the stage coordinate system as a result of thermal expansion, charging, and a variety of other causes. Thus, although the laser interferometer's computer can know exactly where the stage is at any time, the position of the beam is not known precisely unless some reference is made to a fiducial mark on the stage or sample. A proposal for "closing the loop" in SEBL, known as "spatial-phase-locked SEBL," would use a fiducial grid patterned directly on the substrate in such a way as to not disturb the electron beam writing (Smith *et al.*, 1991; Ferrera *et al.*, 1993). By giving rise to a modulation of secondary (or backscattered) electrons, the grid would enable a computer to keep track of the beam position on the sample at all times, independent of influences that cause drift. To ensure accuracy and freedom from distortion, the fiducial grid is generated by holographic lithography.

4.4 Reducing Exposure Time

The most successful approach to reducing the exposure time in SEBL has been the so-called shaped-beam systems (Pfeiffer *et al.*, 1993). Although such systems allow many pixels to be projected in parallel, the equivalent pixel transfer rate does not generally go up in proportion because the current density is significantly reduced over that achievable in round-beam systems with thermal-field-emitter sources. The mutual repulsion of electrons, which causes shape distortion, limits the current density. In addition, the throughput gain with shaped-beam systems depends very much on the pattern type. For complex patterns of fine lines, the gain goes down in proportion. When all these factors are considered, the gain in throughput with shaped beams is generally about an order of magnitude, seldom two orders.

4.5 Electron Scattering

When a high-energy electron enters a material, it interacts with the resident electrons, losing energy more or less continuously through the excitation of secondary electrons, most of which have energies of only several eV. The primary electron undergoes multiple small-angle scattering events. By this mechanism, a well-defined incident beam spreads out as it passes through material. Depending on the substrate atomic number, between 10% and 40% of the incident electrons backscatter and emerge from the substrate at points that can be quite remote from the original entry point. As a result, the exposure dose at a given location depends not only on the local incident dose but also on the dose incident at all points within the electron range. This is known as cooperative exposure or "proximity effect." In principle, by knowing the electron scattering characteristics, which can be determined either experimentally or by Monte-Carlo methods (*q.v.*) (Hawryluk *et al.*, 1975), one can correct for the proximity effect, but with a lowering of the image contrast (Owen, 1990).

5. X-RAY LITHOGRAPHY

The basic principle of proximity x-ray lithography is illustrated schematically in Fig. 9. A membrane, which is x-ray transparent, carries a patterned x-ray absorber whose shadow is cast on a substrate held in close proximity. An x-ray wavelength around 1 nm, or slightly longer, provides adequate image contrast and resolution down to feature sizes of 20 nm, which is about the practical limit of the lithographic process using polymeric resists. A unique virtue of x-ray lithography is that scattering is negligible. In addition, the deleterious effects of photoelectrons released when a photon is absorbed are negligible for wavelengths longer than about 0.8 nm (Early *et al.*, 1990; Ocola and Cerrina, 1993). Hence, proximity effects are essentially absent, and high-quality lithography is possible, irrespective of pattern geometry, substrate composition, or topography.

5.1 Sources

Electron-bombardment x-ray sources are used in research but are too inefficient for commercial x-ray lithography. Both synchrotrons (Grobman, 1981) and dense plasma sources are used instead. In a synchrotron, an electron beam with an energy of 600–1000 MeV is forced by strong magnetic fields to

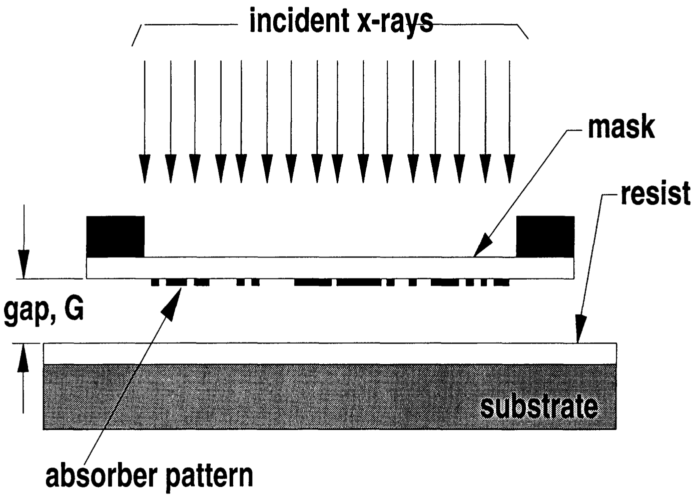
incident x-rays
mask
resist
gap, G
substrate
absorber pattern

FIG. 9. Schematic of the x-ray lithography technique. A pattern, in the form of a high–atomic-number x-ray absorber, resident on a thin, x-ray–transparent membrane, is replicated into a resist film on a substrate separated from the membrane by a gap *G*.

follow a circular or racetrack orbit. The radiation emitted when the electron trajectory is altered is most intense in the soft x-ray regime around 1 nm and is quite adequate for mass-production lithography. A synchrotron is a fixed facility that provides 10 to 20 exposure stations or "beam lines" that radiate out tangent to the electron orbit. A superconducting synchrotron can be somewhat compact, measuring only a few meters in its largest dimension.

At the present time, laser plasma sources lag behind synchrotrons in output flux per exposure station but make up in part for this by operating at somewhat longer wavelengths (e.g., 1.4 nm) where more of the incident radiation is absorbed in the resist (Hector *et al.*, 1993). Other plasma sources, such as the "plasma focus" (Prasad, 1994), offer the possibility of significantly higher flux, comparable to a synchrotron. Such sources require additional development.

5.2 Masks

A mask structure suitable for use with a wavelength of 1.3 nm is shown in Fig. 10. One's intuition, based on everyday experience, would say that such a structure is highly fragile and subject to distortion. However,

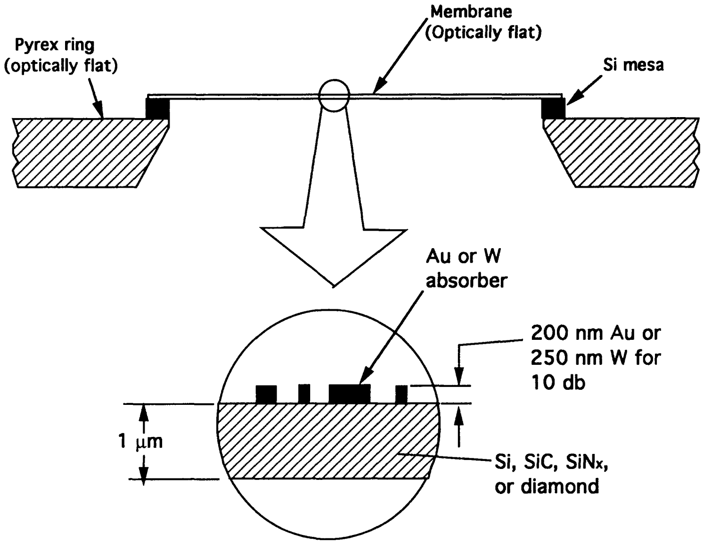
Membrane (Optically flat)
Pyrex ring (optically flat)
Si mesa
Au or W absorber
200 nm Au or 250 nm W for 10 db
1 μm
Si, SiC, SiNx, or diamond

FIG. 10. Configuration of x-ray lithography mask suitable for use at the 1.3-nm wavelength. Typically, the membrane is flatter than 0.25 μm and over 30 mm in diameter. The glass ring provides rigidity.

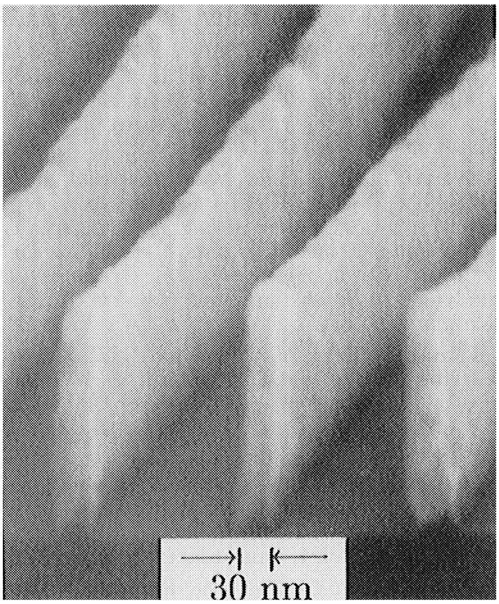

FIG. 11. Scanning electron micrograph of grating pattern exposed in PMMA resist using a 4.4-nm-wavelength x-ray and a mask-substrate gap less than 1 μm (Smith and Craighead, 1990).

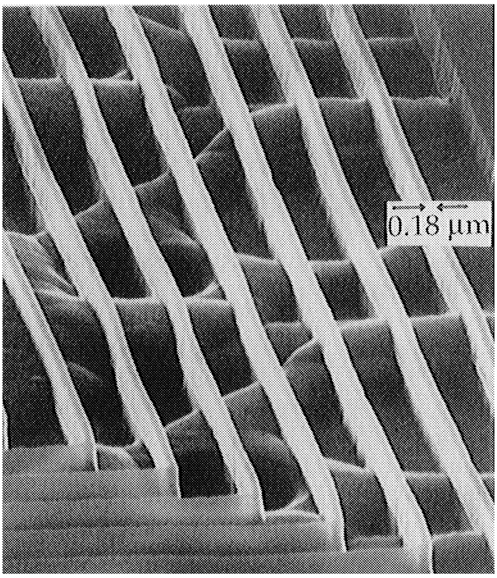

FIG. 12. Scanning electron micrograph of pattern exposed in a chemically amplified resist (PN 114) using a laser-plasma x-ray source (1.4 nm average wavelength) and a mask-substrate gap of 20 μm (courtesy of I. Plotnik).

membranes of Si, SiN_x, and SiC, 1 to 2 μm thick, have proven to be unusually robust. They are readily handled in research and manufacturing environments and are even used as vacuum windows.

Pattern distortion can arise from compressive or tensile stress in the absorber. By keeping such stress below 10 MPa, pattern distortion is well below 1 part per million for the types of patterns encountered in microelectronics.

5.3 Results

Figures 11 and 12 illustrate some results. The former was exposed at a near-zero gap between mask and substrate, and illustrates high-resolution nanolithography in thick resist. The latter was exposed at a 20-μm gap and illustrates exposure over topography. The relationship between gap G and minimum feature size W is given by

$$G = \alpha W^2/\lambda, \tag{3}$$

where λ is the x-ray wavelength and α is a parameter. Experiment and numerical simulation of diffraction effects have shown that one can readily operate at quite large gaps, corresponding to α in the range 1 to 1.5, provided the effective source size (i.e., its spatial incoherence) is sufficiently large to smooth out edge ripples that occur with spatially coherent illumination (Hector *et al.*, 1992).

5.4 Alignment

In x-ray lithography, because the mask and substrate are in close proximity, alignment of a mark on the mask to a mark on the substrate is technologically simpler than in other forms of lithography. It can be done with 10-nm precision using a variety of methods, the most sensitive of which employ optical interferometry (Ishihara *et al.*, 1989). However, overlayers grown on top of alignment marks can confuse the interferometric alignment. To avoid such confusion, it is preferable to use a range of optical wavelengths, which is not feasible with traditional interferometric methods. Recently, an on-axis interferometric scheme that uses white light was described. In conjunction with modern digital image-processing technology, high precision, together with greater insensitivity to overlayer growth, should be achievable (Moel *et al.*, 1993).

6. CONCLUSIONS

Each of the several forms of lithography has distinct advantages, disadvantages, and a domain of applicability. The commercial manufacture of high-density integrated circuits is the most publicized application of microlithography, and the one that provides the greatest financial motivation. In this sphere, many nontechnical factors, such as existing infrastructure and costs, affect the choice of lithography technique. Such considerations are largely beyond the scope of this analysis, which has focused on engineering and physical aspects.

GLOSSARY

Irradiance: A measure of the power per unit area of radiation incident on a substrate.

Microlithography: The creation of structures in resist with minimum feature sizes from a few micrometers to 100 nm.

Nanolithography: The creation of structures in resist with minimum feature sizes below 100 nm.

Photoresist: A resist film sensitive to visible or ultraviolet radiation.

Pitch: The center-to-center distance between two adjacent lines.

Resist: A film coating onto a substrate, which, upon exposure to a pattern of radiation, can be converted, via a development process, into a relief image.

Resolution: A term used rather loosely to designate the minimum size of relief structures achievable with a given lithographic technique. In microscopy, resolution usually refers to the minimum center-to-center distance between two objects where one can distinguish that there are two objects rather than one.

Spatial Period: The center-to-center distance between two adjacent lines.

Works Cited

Abboud, F., Gesley, M., Colby, D., Comendant, K., Dean, R., Eckes, W., McClure, D., Pearce-Percy, H., Prior, R., Watson, S. (1992), "Electron Beam Lithography using MEBES IV," *J. Vac. Sci. Technol. B* **6**, 2734–2742.

Early, K., Schattenburg, M. L., Smith, H. I. (1990), "Absence of Resolution Degradation in X-ray Lithography for λ from 4.5 nm to 0.83 nm," *Microelectron. Eng.* **11**, 317–321.

Everett, P. N., Delaney, W. F., Griswold, M. P. (1991), "Flexible and Rigid Masks and Alignment in Binary Optics," in: *Proceedings of the Lasers and Electro-Optics Society (Institute of Electrical and Electronics Engineers) Meeting, July 31–August 2, 1991, Newport Beach, California*, New York: IEEE, Sec. 2, pp. 19–20.

Ferrera, J., Wong, V. V., Rishton, S., Boegli, V., Anderson, E. H., Kern, D. P., Smith, H. I. (1993), "Spatial-Phase-Locked Electron-Beam Lithography: Initial Test Results," *J. Vac. Sci. Technol. B* **11**, 2342–2345.

Grobman, W. D. (1981), "Synchrotron Radiation X-ray Lithography," in: E. E. Koch, D. E. Eastman, P. Farge (Eds.), *Handbook on Synchrotron Radiation*, Vol. 1, Amsterdam: North Holland.

Haury, E. W. (1967), "The Hohokam," *Natl. Geogr. Mag.* **131** (May), 670–695.

Hawryluk, R. J., Smith, H. I., Soares, A., Hawryluk, A. M. (1975), "Energy Dissipation in a Thin Polymer Film by Electron Beam Scattering—Experiment," *J. Appl. Phys.* **46**, 2528–2537.

Hector, S. D., Schattenburg, M. L., Anderson, E. H., Chu, W., Wong, V. V., Smith, H. I. (1992), "Modeling and Experimental Verification of Illumination and Diffraction Effects on Image Quality in X-ray Lithography," *J. Vac. Sci. Technol. B* **10**, 3164–3168.

Hector, S. D., Smith, H. I., Schattenburg, M. L. (1993), "Simultaneous Optimization of Spectrum, Spatial Coherence, Gap, Feature Bias, and Absorber Thickness in Synchrotron-Based X-ray Lithography," *J. Vac. Sci. Technol. B* **11**, 2981–2985.

Inoue, S., Fujisawa, T., Tamaushi, S., Ogawa, Y., Nakase, M. (1992), "Optimization of Partially Coherent Optical System for Optical Lithography," *J. Vac. Sci. Technol. B* **6**, 3004–3007.

Ishihara, S., Kanai, M., Une, A., Suzuki, M. (1989), "A Vertical Stepper for Synchrotron X-ray Lithography," *J. Vac. Sci. Technol. B* **7**, 1652–1656.

Kawata, H., Carter, J. M., Yen, A., Smith, H. I. (1989), "Optical Projection Lithography using Lenses with Numerical Apertures Greater than Unity," *Microelectron. Eng.* **9**, 31–36.

Lin, B. J. (1992), "The Attenuated Phase-Shifting Mask," *Solid State Technol.* (January), 43–47.

Moel, A., Moon, E. E., Frankel, R., Smith, H. I. (1993), "Novel On-axis Interferometric Alignment Method with Sub-10 nm Precision," *J. Vac. Sci. Technol. B* **11**, 2191–2194.

Ocola, L. E., Cerrina, F. (1993), "Parametric Modelling Photoelectron Effects in X-ray Lithography," *J. Vac. Sci. Technol. B* **11**, 2839–2844.

Okazaki, S. (1991), "Resolution Limits of Optical Lithography," *J. Vac. Sci. Technol. B* **9**, 2829–2833.

Owen, G. (1990), "Methods for Proximity Effect Correction in Electron Lithography," *J. Vac. Sci. Technol. B* **8**, 1889–1892.

Pfeiffer, H., Davis, D. E., Enichen, W. A., Gordon, M. S., Groves, T. R., Hartley, J. G., Quickle, R. J., Rockrohr, J. D., Stickel, W., Weber, E. V. (1993), "EL-4, A New Generation Electron-Beam Lithography System," *J. Vac. Sci. Technol. B* **11**, 2332–2341.

Prasad, R. R., Krishnan, M., Mangano, J., Greene, P. A., Qi, N. (1994), "Neon Dense Plasma Focus Point X-Ray Source for ≤0.25 μm Lithography," in: D. Patterson (Ed.), *Electron-Beam, X-Ray, and Ion-Beam Submicrometer Lithographies for Manufacturing IV*, SPIE Proceedings Vol. 2194, Bellingham, WA: SPIE.

Smith, H. I. (1974), "Fabrication Techniques for Surface-Acoustic Wave and Thin-Film Optical Devices," *Proc. IEEE* **62**, 1361–1387.

Smith, H. I. (1986), "A Statistical Analysis of UV, X-ray and Charged-Particle Lithographies," *J. Vac. Sci. Technol. B* **4**, 148–153.

Smith, H. I. (1988), "A Model for Comparing Process Latitude in UV, Deep-UV and X-ray Lithography," *J. Vac. Sci. Technol. B* **6**, 346–349.

Smith, H. I., Craighead, H. G. (1990), "Nanofabrication," *Phys. Today* **43** (2), 24–30.

Smith, H. I., Hector, S. D., Schattenburg, M. L., Anderson, E. H. (1991), "A New Approach to High Fidelity E-Beam Lithography Based on an In-Situ, Global Fiducial Grid," *J. Vac. Sci. Technol. B* **9**, 2992–2995.

Tamechika, E., Matsuo, S., Komatsu, K., Takeuchi, Y., Mimura, Y., Harada, K. (1992), "Investigation of Single Sideband Optical Lithography using Oblique Incidence Illumination," *J. Vac. Sci. Technol. B* **10**, 3027–3031.

Further Reading

Brodie, I., Muray, J. J., (1992), *The Physics of Micro/Nano-Fabrication*, New York: Plenum.

Moreau, W. M., (1988), *Semiconductor Lithography*, New York: Plenum.

Thompson, L. F., Willson, C. G., Bowden, M. J., (1983), *Introduction to Microlithography*, ACS Symposium Series Vol. 219, Washington, D.C.: American Chemical Society.

MICROMECHANICAL DEVICES

See SYSTEMS TECHNOLOGY, MICROMECHANICAL

MICROPROCESSORS

Nick Tredennick, *Tredennick, Inc., Los Gatos, California, U.S.A.*

Introduction 297
1. **History** 299
1.1 Microprocessor as a Logic Device 299
1.2 Microprocessor as a CPU 301
2. **Architecture** 301
3. **Operation** 302
3.1 Reset 303
3.2 Initialization 303
3.3 Cache 304
3.4 Interrupts and Exceptions ... 304
4. **Design** 305
4.1 Execution Unit 305
4.2 Bus Interface Unit 306
4.3 Controller 306
4.4 Pipelining 307
4.5 Branch Execution Model 313
4.6 External Bus Protocol 314
5. **Technology and Manufacturing** 315
6. **Applications** 317
6.1 The Microprocessor as a CPU 317
6.2 Embedded Control 317
Glossary 318
Works Cited 318
Further Reading 319

INTRODUCTION

Figure 1 is the block diagram of a simple computer (see COMPUTERS). The simple computer consists of a central processing unit (CPU), a memory, and peripheral-interface logic connected to a common bus. The CPU is the engine for the computer. It runs the programs and controls the system. The memory holds the programs, the input data, and the results. The peripheral-interface logic is the means for getting the data and programs into the computer (from a keyboard, mouse, or disk, for example) and the means for getting the results from the computer (to a printer or display screen, for example).

The first computers, built in the early 1940s, were room sized and were built from vacuum tubes. The transistor was invented in 1947 by Bardeen, Brattain, and Shockley (Davies and Wharton, 1983). Later computer generations used transistors instead of vacuum tubes and became smaller and more reliable. After the integrated circuit, which allowed numerous transistors on a single semiconductor chip, was invented in 1958, computers got even smaller. Today, large integrated circuits sometimes contain several million transistors.

There is very little conceptual difference between this simple computer system, first built with vacuum tubes, and a computer built with a microprocessor. (There are enormous differences in physical implementation, however.) Figure 2 shows a simple microprocessor system. A microprocessor is a CPU on a single integrated circuit (sometimes called a "chip"). In the illustration of a simple microprocessor system. I have broken the memory into the random-access memory (RAM) and the read-only memory (ROM). RAM is useful because it allows rapid access to any location for reading or writing and is generally faster than ROM, but it loses any stored information when power is lost. ROM does not change or lose its contents when power is lost or when the system is restarted, and so it is useful for holding initialization programs and other programs often referenced, but not changed.

Figure 3 shows how a computer program, residing in the memory, looks to the programmer and to the computer. The programmer thinks of the computer program as being a sequential series of instructions such as add and subtract, punctuated by branches. The computer sees the program a little differently; it sees only strings of ones and zeros.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

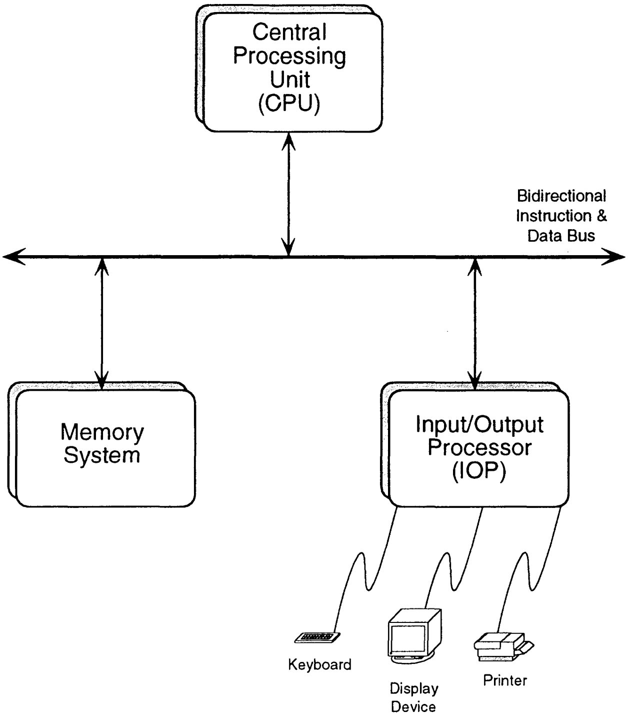
Central Processing Unit (CPU)
Bidirectional Instruction & Data Bus
Memory System
Input/Output Processor (IOP)
Keyboard
Display Device
Printer

FIG. 1. Simple computer system. The simplest computer is a central processing unit (CPU) connected on a common bus to a memory and an input/output processor (IOP). (Figure components courtesy of Joe Higham.)

Programmers work with symbols, numbers, and mnemonics so that the programs are easier to understand. The computer works with binary ones and zeros because that is all it understands. A look at the contents of a location in the memory system would reveal only ones and zeros: the symbols, numbers, and mnemonics are constructed for the convenience of the programmer. Programmers create programs of symbols, numbers, and mnemonics to be translated (by another computer program called an assembler or compiler) to ones and zeros for use by the computer.

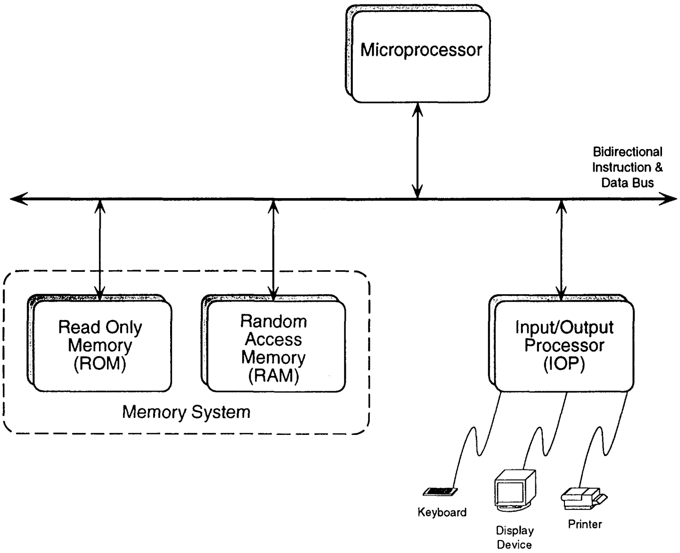
Microprocessor
Bidirectional Instruction & Data Bus
Read Only Memory (ROM)
Random Access Memory (RAM)
Memory System
Input/Output Processor (IOP)
Keyboard
Display Device
Printer

FIG. 2. A simple microprocessor system has a microprocessor, ROM, RAM, and an IOP connected to a common instruction and data bus.

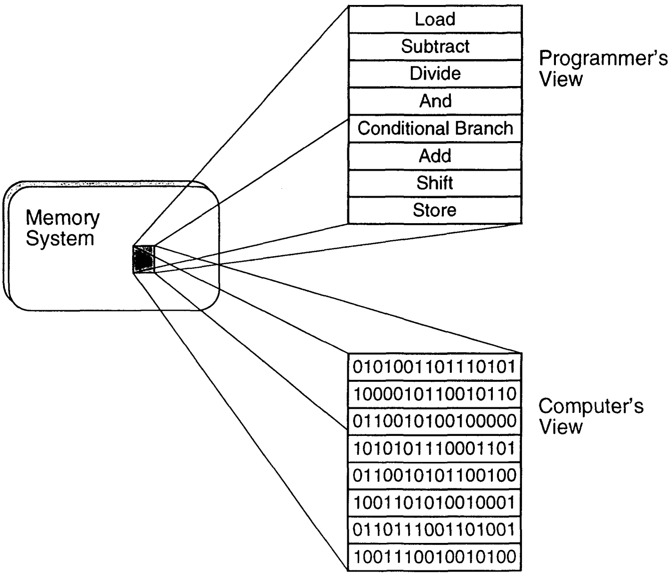

FIG. 3. Simple computer program. To the programmer, the computer program looks like a sequence of actions. To the computer, the computer program looks like a sequence of ones and zeros.

1. HISTORY

Even though the block diagrams of the simple computer system and the simple microprocessor system are strikingly similar, the microprocessor did not start out as part of a general-purpose computer. The microprocessor started as an alternative way to implement digital circuits built from standard logic modules. Before the invention of the integrated circuit by Jack Kilby at Texas Instruments in 1958 (Davies, 1983), logic circuits were built on boards (called circuit boards or logic boards) using discrete components: transistors, diodes, resistors, capacitors, and inductors (Bell *et al.*, 1978). The integrated circuit permitted multiple transistors on a single chip. Transistors, diodes, capacitors, and resistors could be built on an integrated circuit. After the invention of the integrated circuit, standard logic families were invented—the most notable being the 74xxx series. For example, the Texas Instruments SN7400 integrated circuit contained four two-input NAND gates (the output of each two-input gate is the inverse of the Boolean AND of the two inputs) in a fourteen-pin package. The logic family began with basic logic gates (NAND, NOR, AND, OR, and NOT) and grew to include more complex functions [such as register files, arithmetic-logic units (ALUs), and multiplexers]. These standard logic families placed modular logic functions in the hands of designers and led to increased complexity and variety in circuit board design. As integrated circuit densities improved, complexity and variety of the logic modules increased correspondingly. Soon, manufacturers were supporting catalogs of hundreds of different standard logic modules (see, for example, Texas Instruments, 1973).

1.1 Microprocessor as a Logic Device

The first implementation of a microprocessor was the result of a special design request from Busicom, a Japanese company, contracting with Intel for the custom design of calculator logic in 1969 (Faggin, 1992). Busicom was not content to build their calculator product using available standard logic modules and felt that they could get a more efficient implementation with custom integrated circuits; they wanted to specify the logic modules and have Intel build them. The innovative logic structure proposed by Dr. Marcian "Ted" Hoff of Intel led to the development of the MCS-4 chip set. One of the MCS-4 chips, the 4004 (see Fig. 4), was the first microprocessor. The MCS-4, with its 4-bit 4004 microprocessor, was introduced to the market by Intel in 1971. While the first microprocessors were expensive, they still found their way into some applications. Eventually, prices began to decrease and industry came to view the microprocessor as a

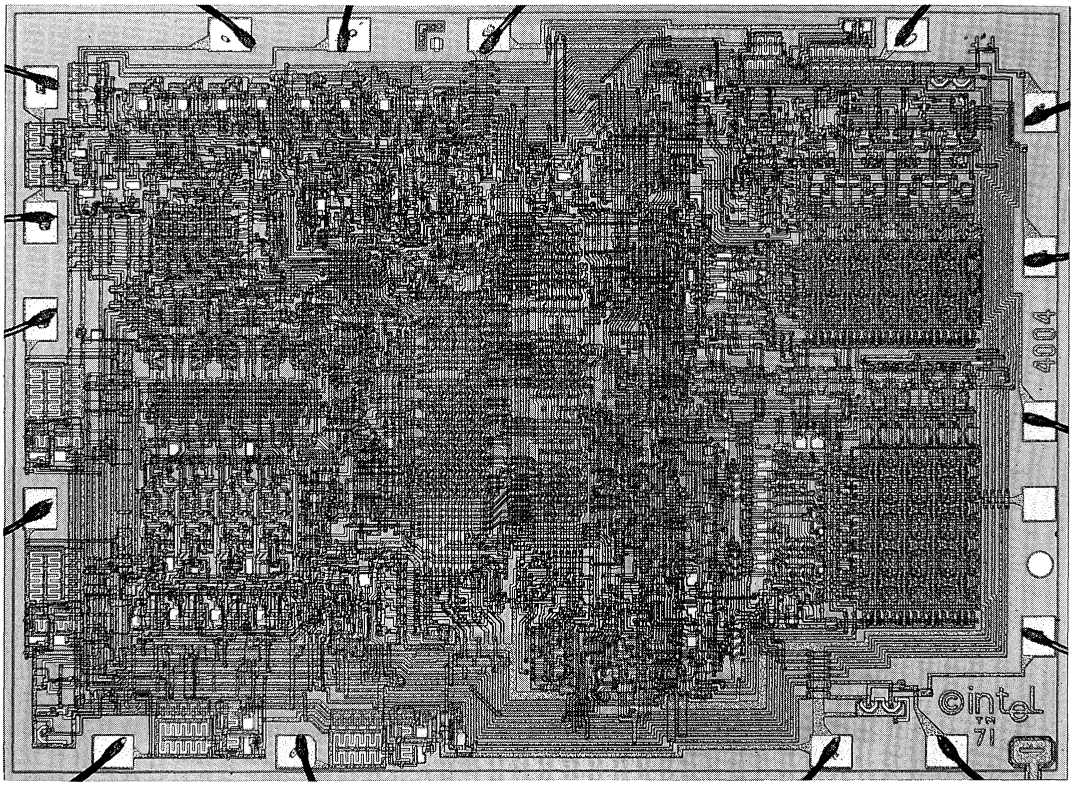

FIG. 4. Intel's 4004 introduced in 1971 was probably the world's first commercial microprocessor. (Courtesy Intel Corporation.)

good substitute for circuits built from standard logic modules. The microprocessor found application in calculators, washing machines, microwave ovens, instruments, and even computers. Rather than have a large number of standard logic modules on a circuit board, the designer could use a microprocessor with some RAM and ROM and program the logic functions. This could reduce component count, system cost, and the variety of components stocked by the circuit board manufacturer. For this to work, the microprocessor had to provide equivalent function, performance, and reliability at a lower cost. Simple bus protocols (the way the microprocessor "talks" to other chips in the circuit) allowed the microprocessor to be connected to a common bus with a variety of memory and peripheral-interface components. Market emphasis on the use of microprocessors to replace standard logic modules led to the following characteristics:

1. lowest cost design with adequate performance (efficient, optimized designs lead to smaller, cheaper chips);
2. simple, flexible bus protocols (to allow the chips to work with a variety of memory and peripheral components with differing speed requirements);
3. few package pins so that the microprocessor would fit in commonly available integrated circuit packages and, therefore, use standard sockets (most of the standard logic-module integrated circuit packages were 14 or 16 pins; an integrated circuit with 28 to 40 pins would be considered large).

These microprocessors were designed to be manufactured in very high volumes. Even today, many of these microprocessors are shipped in quantities of millions of units per year. Almost all of these microprocessors are used as alternatives to standard logic modules. Low-end microprocessors of all kinds are shipped at more than a billion units per year.

1.2 Microprocessor as a CPU

Use of the microprocessor as the central processing unit in a computer system began with the invention of the personal computer in the mid 1970s (Smith, 1981). The market for personal computers grew rapidly after IBM introduced their version of the personal computer in 1981. Just 10 years later, in 1991, about 21 million IBM-compatible personal computers were shipped worldwide. Intel Corporation makes the microprocessors used as the CPU in IBM-compatible personal computers. In 1991 Intel (and other manufacturers of 80386-compatible microprocessors) shipped over 2 million 80486 microprocessors and more than 15 million 80386 microprocessors, almost all of which ended up as CPUs in personal computer systems (Slater, 1992). During the same year, Intel also shipped more than 10 million 80286 microprocessors (also used in versions of the IBM-compatible personal computer). Probably most of these became the CPU's in personal computer systems. Motorola makes the 68000 family of microprocessors used as the CPU in Apple's popular Macintosh computers, as well as in the Amiga and Atari computer systems. In 1991, Motorola shipped over 3 million 680x0 microprocessors for use as the CPUs in computer systems (McWilliams, 1992). Motorola and Intel make the most popular microprocessors used in personal computers, but several other manufacturers share the much smaller market for microprocessors used as the CPUs in engineering work stations (a computer for engineering design that is more powerful and more expensive than a personal computer). So the market for microprocessors as a CPU may be 25 to 30 million units per year (Tredennick, 1991). It is not a large part of the total market for microprocessors, which is about 1.5 billion units per year—just about 2%. But a microprocessor used as a CPU tends to be more expensive than a microprocessor for other applications, and so the dollar value of this market is high.

The problem with the microprocessor-as-a-CPU market was its requirements: the CPU market demands performance. The CPU market is performance driven, not cost driven. The requirements of the market for microprocessors replacing standard logic modules (low cost and adequate performance) and the CPU market (maximum performance at any cost) were in direct conflict. From a marketing viewpoint, there is no contest—the logic replacement market is about 20 times the size of the CPU market. Performance of the early microprocessors was limited by the simple, leisurely bus protocols, multiplexed instruction and data buses, packages with few pins, and slow, immature technology. The instruction sets and the implementations were an excellent fit for applications where the microprocessor replaced standard logic modules.

The market in logic applications, being 98% of the market volume, was too large to be ignored, and so microprocessor manufacturers concentrated their efforts on devices to support its requirements: low cost and adequate performance. This meant that the market for microprocessors used as CPUs was not well served by available microprocessors. In the early 1980s, researchers began to pro pose microprocessor architectures specifically suited to the market for central processors (Hennessy *et al.*, 1982; Patterson and Séquin, 1980). These microprocessors increased the number of available package pins and improved the speed of memory transactions. Implementers provided separate 32-bit buses for addresses and instructions or data. Conventional microprocessors had few pins, and so pins were shared for addresses, instructions, and data. Bus protocols were sped up at the cost of requiring tightly coupled special memory systems called caches and not allowing direct connection to peripherals. These designs considered performance first. Today we have two markets for microprocessors: one market for the microprocessor replacing standard logic modules on a circuit board (98% of the volume) and one market for the microprocessor as a CPU (2% of the volume). The market for the microprocessor replacing standard logic modules (called embedded control) is characterized by high volumes and extreme price competition. The market for the microprocessor as a CPU is characterized by low volumes and competition for performance.

2. ARCHITECTURE

The "architecture" of a microprocessor refers to the view of the microprocessor as seen by a programmer (Stallings, 1990). Elements of a microprocessor's architecture are usually described in a user's manual or programmer's reference manual (much like the owner's manual for an automobile). The architecture description includes the instruction set (what the microprocessor can do), the instruction formats (what commands to the microprocessor look like), and the internal register set (storage registers available to the programmer).

Instructions are the commands to the microprocessor. Figure 5 is an example of a 32-bit instruction format for a microprocessor with 64 general-purpose internal registers (2^6, the length of the source and destination specifiers). The instruction format consists of an operation field, two source operand fields, a mode field, and a destination field. The operation field tells the microprocessor the operation (e.g., add, subtract, multiply, shift). In a modern microprocessor architecture, the source operand fields point to register locations on the microprocessor. The destination field tells the microprocessor where to place any result from the operation.

I divide instructions into five categories: arithmetic, logical, control, transfer, and branching. Arithmetic instructions include add, subtract, multiply, and divide. Logic instructions include AND, OR, and EOR (exclusive OR). Control instructions help the programmer set the internal and external context for the execution of the other instructions. Transfer instructions include the load and store instructions and more operands between the memory system and the CPU. Branching instructions change the sequential execution of the program, often in conjunction with the testing of condition codes resulting from an arithmetic or logical operation.

3. OPERATION

The microprocessor is controlled by a program stored in the memory. A program might be written in a high-level language (e.g., FORTRAN, COBOL, C^{++}, BASIC, ADA; see Computer Programming Languages). The high-level language representation is translated by a compiler into instructions that the microprocessor understands (the ones and zeros of Fig. 3). The program might also be written in an

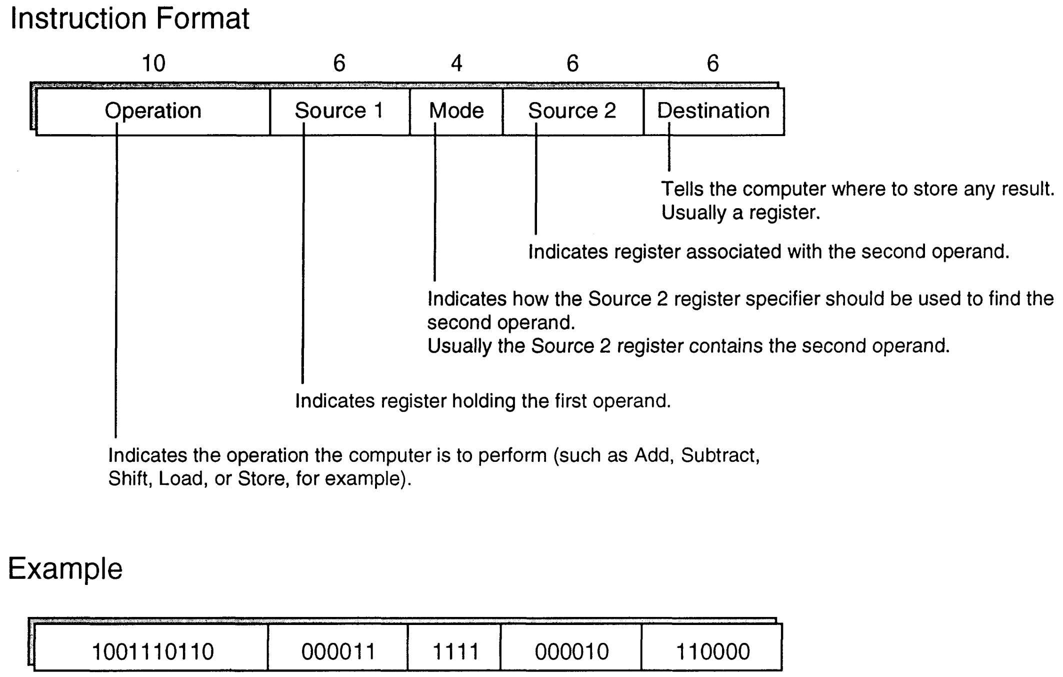
Instruction Format
10
6
4
6
6
Operation
Source 1
Mode
Source 2
Destination
Tells the computer where to store any result.
Usually a register.
Indicates register associated with the second operand.
Indicates how the Source 2 register specifier should be used to find the second operand.
Usually the Source 2 register contains the second operand.
Indicates register holding the first operand.
Indicates the operation the computer is to perform (such as Add, Subtract, Shift, Load, or Store, for example).
Example
1001110110
000011
1111
000010
110000

FIG. 5. Example format for a microprocessor instruction. The instruction tells the microprocessor where the source operands are, what operation to perform, and where to store any result.

assembly language, which is a mnemonic and symbolic representation of the microprocessor instruction set. An instruction-set assembler is then used to translate the assembly-language program into the ones and zeros that the microprocessor understands. The program and data are stored on a floppy disk or hard disk that can be accessed by the microprocessor through the peripheral-interface logic. Tiny magnetic domains on the floppy disk or hard disk are flipped in one direction to store a one and in the opposite direction to store a zero. The direction of the domain remains stable even when power is no longer applied to the device. Since the floppy disk and hard disk retain their data even when the power is turned off, they are nonvolatile storage. Floppy disks, hard disks, tapes, and other magnetic media provide nonvolatile storage for programs and data. ROM is also nonvolatile storage. RAM and the register file in the microprocessor lose their information when power is removed. Storage devices that lose information when power is removed are called volatile.

Now that the microprocessor, computer system block diagram, and instructions have been introduced, it is possible to describe operation of a simple microprocessor. (From the description of the simple microprocessor, the explanation will progress to more complex implementations.) To run a program (see Fig. 3), the microprocessor fetches an instruction from the memory system by communicating with the memory system over the bidirectional instruction and data bus (see Fig. 2). The microprocessor decodes the instruction to determine the required operation and the location of source operands. Then the microprocessor "executes" the instruction by obtaining the necessary operands, sending them to the ALU, and storing the result. For the example instruction shown in Fig. 5, the microprocessor would fetch the instruction from the memory system. Decoding the instruction would tell the microprocessor to set up the ALU for addition (as opposed to subtraction or some other operation) and that the source operands are in registers R3 and R2. The microprocessor executes the instruction by reading the contents of R3 and R2 from the register file, sending them to the ALU (which is set for addition), and storing the result in R48. The simple microprocessor updates the program counter (which is an internal pointer to the next instruction in the program) and fetches the next sequential instruction from the memory system.

3.1 Reset

When power is turned off, the microprocessor and the RAM lose their internal state information (including the contents of the program counter). When power is applied, voltage and current transients occur throughout the chip and may cause unpredictable states inside the microprocessor. A reset pin allows the user to set the microprocessor to a known internal state. The microprocessor might have an additional reset pin that it can drive to set the system in which it resides to a known internal state. Once the microprocessor is in a known internal state, it should be able to bring the rest of the computer system to a known operational state. But the RAM, where the program and data for execution normally reside, will not contain useful information (since information in the RAM was lost the last time the power was removed). The ROM attached to the microprocessor as a part of the memory system (see Fig. 2) does not lose its information content when power is lost; so, when power is applied to the system, the microprocessor goes to the ROM to get the instructions for system initialization.

3.2 Initialization

The microprocessor will reset the system and then move programs and data from the nonvolatile magnetic storage devices (hard disk, floppy disk, or tape drive) into the RAM in the memory system for execution. One of the first programs brought into the memory system for execution will be the operating system. The operating system is the program that sets the policies and procedures for executing application programs on the microprocessor. Why bother with the RAM? Why not just run the programs from the magnetic storage devices? The reason is speed. Typical RAMs have access times measured in nanoseconds (billionths of a second), while typical magnetic storage devices have access times measured in milliseconds (Rao, 1978). Even though the access time for a magnetic storage device is about a million times slower than the access time of the RAM in the memory system, transfer rates only differ by a factor of 10 or 100. A typical memory system might transfer 4 bytes every 40 ns, for example, for a transfer rate of 100 MB/s. A hard disk attached to the peripheral-interface logic in a microprocessor system usually has a transfer rate of between 1 and 10 MB/s. The efficiency of the computer system is greatly increased, therefore, by transferring blocks of programs and data from the magnetic storage device to the memory system and running the programs from the memory system.

3.3 Cache

Cache is a special form of memory. A cache memory resides between the main memory system and the CPU. In most modern microprocessors, the cache is located on the chip with the CPU and might be between 4 kB (in computer terms, k represents 1024, so that this is really 4096 bytes) and 32 kB (really 32 768 bytes). Main memory might be about 1 to 32 MB. The cache memory is much faster than the main memory system because it is smaller, located closer to the CPU, and built of more expensive technology. Cache memory exploits a characteristic of programs known as locality of reference. Instructions and data used by a computer program tend to be used again within a short time. The first time an instruction or memory operand is referenced, it is brought into the cache. Subsequent references to the same instructions or operands produce a "hit" in the cache, and the access is returned to the CPU much quicker than if the reference had to go to the main memory system. Even though the main memory system might be a thousand times the size of the cache, instructions and data might "hit" in the cache 90% of the time because of locality of reference. The cache can have a dramatic effect on the performance of the microprocessor. Suppose, for example, that it takes two clock cycles to access the cache on the microprocessor chip and four clock cycles to access the main memory system. If there is no cache, the time to execute a program is the number of accesses n in the program times the number of clock cycles, 4, in a memory system access:

$$t = 4n. \tag{1}$$

If there is a cache with a 90% "hit" ratio, then the time to execute the same program becomes

$$t = 1 \times 0.9n + 4[(1 - 0.9)n] = 1.3n. \tag{2}$$

Locality of reference helps a cache reduce execution time from $4n$ to $1.3n$. In this example, the small cache makes the CPU think the memory system is three times faster than it really is.

3.4 Interrupts and Exceptions

The microprocessor is the CPU running the system, but it cannot conveniently do everything (Money, 1990). It is simpler, for example, to have a separate processor in the keyboard watching for keystrokes. When you strike a key on your keyboard, the keyboard processor sends a message to the CPU requesting processing time for the keyboard. The form of this message is called an interrupt. The interrupt allows the CPU to process program instructions that are its primary responsibility and not have to waste time explicitly checking for numerous infrequent external activities. Interrupts are the means for devices outside the CPU to request attention and get service for infrequent (relative to the speed of CPU processing) events. Interrupts, since they are requests from outside the microprocessor, are usually asynchronous, meaning that their timing is not synchronized with the microprocessor's clock. An interrupt request may occur at any time during the processing of any instruction. Interrupt requests must, therefore, be synchronized with the microprocessor clock. This can be accomplished by use of synchronizing circuit to capture the request in a register. The microprocessor will then recognize and process the request at the next convenient interval (usually at the end of the current instruction and before beginning the following instruction).

An exception occurs when the microprocessor cannot complete the current instruction. If, for example, the microprocessor attempts to decode an instruction with an illegal bit pattern (i.e., a bit pattern not defined for the instruction set), normal processing cannot continue since no operation is defined for the bit pattern. The microprocessor will execute instead an exception sequence in an ef-

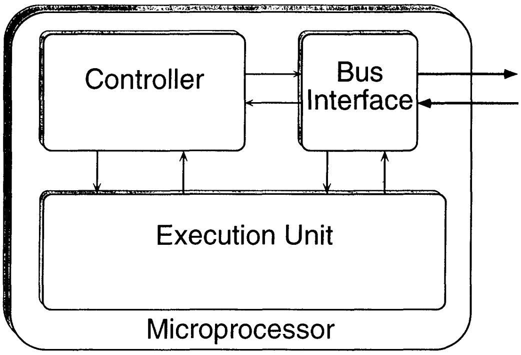

FIG. 6. The major blocks inside a simple microprocessor are the controller, the execution unit, and the bus interface.

fort to identify the problem and continue processing. Exceptions can also occur if the microprocessor cannot complete execution of an instruction because it is unable to obtain data from memory. Arithmetic exceptions such as overflow and divide by zero may also be defined.

4. DESIGN

Figure 6 shows the inside of a microprocessor. The three necessary components of the microprocessor are the bus interface, the controller, and the execution unit. The execution unit (or data path) contains the resources for computing such as the register file, program counter, and arithmetic units. The controller runs the execution unit and issues commands to the bus interface. The bus interface is responsible for transactions between the external bus and the execution unit. This block diagram is good for describing functions on the first microprocessor and it is good for describing modern microprocessors.

4.1 Execution Unit

Figure 7 shows an execution unit with a three-bus structure. The execution unit contains the resources seen by the programmer such as the programmer's register set, the program counter, and the condition code register. In addition, the execution unit contains the instruction register, the ALU, data and address temporary storage registers, an address output register, and input and output registers for data.

The internal *A* and *B* buses in the execution unit transmit operands from the register file or other source to the ALU (or other destination). The internal *C* bus transmits the result to the register file or other destination. Since the execution unit shown in Fig. 7 has two source buses and a destination bus, it is able to return the result of the previous operation while executing the current operation. The execution unit could be made cheaper by leaving out one bus (so that the registers would only have two access ports, rather than three), but then the execution of the current instruction could not be overlapped with storing the result of the previous operation.

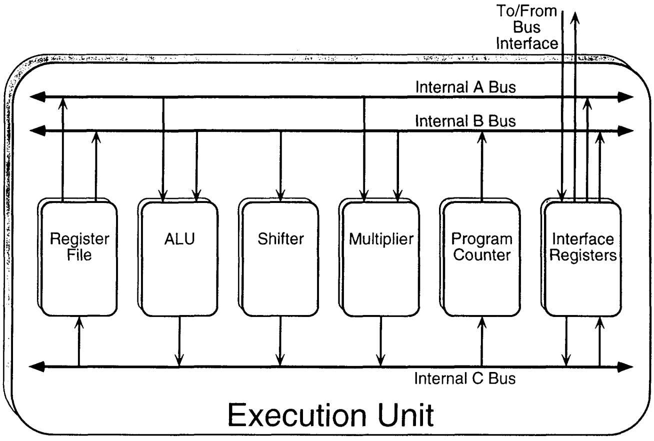

FIG. 7. The execution unit contains the resources (such as the registers and arithmetic units) that the controller manipulates to execute a program.

4.2 Bus Interface Unit

The data in and out registers and the address out register are shared with the bus interface (see Interface Registers in Fig. 7). The bus interface communicates with the external bus using commands from the controller and data and addresses from the execution unit. For a store-to-memory operation, the bus interface will send the address from the address out register and the data value from the data out register to the memory system via the external bus. To load a value from memory to a register in the microprocessor, the bus interface will send the address from the address out register to the memory system and place the value returned on the external bus in the data in register in the execution unit.

4.3 Controller

Figure 8 is the controller of a microprocessor. The controller is built of an instruction decoder, a state sequencer, and a control decoder. The instruction decoder looks at the ones and zeros in the instruction and directs the initial action of the state sequencer. The state sequencer steps the microprocessor through the states necessary to execute an instruction. The control decoder mixes cycle-by-cycle dynamic control signals from the state sequencer with static decode signals from the instruction decoder. The output of the control decoder helps the state sequencer determine the next control state and also directly controls actions in the execution unit. Input to the instruction decoder might be 16 to 32 bits. Output of the instruction decoder might be 9 to 14 bits. Output of the state sequencer could be 30 to 150 bits. Output of the control decoder might be 180 to 400 bits with 9 to 14 of that being the feedback to the state sequencer to help the state sequencer determine what needs to be done next.

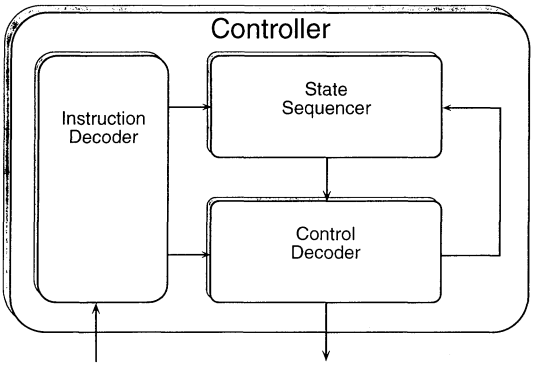

FIG. 8. The controller is the engine that runs the microprocessor. The controller decodes the instructions and sequences the microprocessor through the states necessary to execute the instruction.

The instruction decoder uses the operation field to determine what the initial action of the state sequencer should be. Values of the register fields are sent to the control decoder. The state sequencer needs to know which operation is to be accomplished, but the operations performed on a cycle-by-cycle basis do not depend on which registers hold the source operands or store the results. This is a good example of the difference between static and dynamic control information. Pointers to the source and destination registers are static control information. Dynamic control information tells the execution unit when to do what and static control information tells which resources to use.

The state sequencer is the heart of the controller. The state sequencer controls the resources of the execution unit (through the control decoder). The state sequencer determines when to fetch the next instruction, access operands, perform calculations, store results, or process interrupts. The state sequencer holds the repertoire of commands for the control decoder and options for next state actions. If this information is held in a memory array, it is called microcode. If the same information is held in a collection of logic gates [such as a programmed logic array (PLA)], it is often called random logic. In a random logic implementation, the boundary between the state sequencer and the control decoder may blur. In either case (microprogrammed or random logic), the information in the state sequencer is likely to be encoded to save space.

The control decoder translates commands from the state sequencer into actions in the execution unit. The control decoder mixes control information from the static decoder (e.g., pointers to the source and destination registers) with the dynamic information (e.g., when to read the operands and when to store a result) from the state sequencer to control the cycle-by-cycle action of the resources in the execution unit.

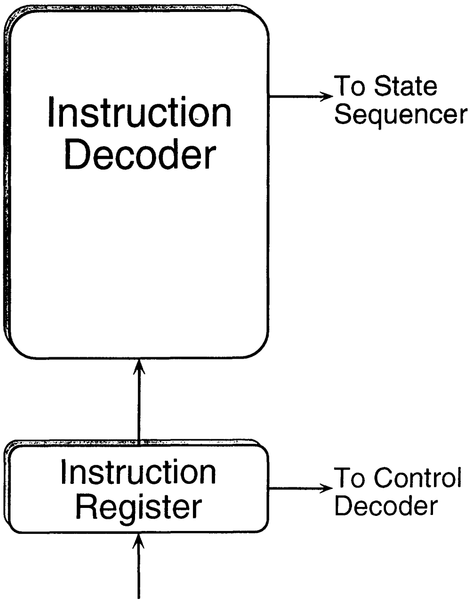

FIG. 9. Instruction register and instruction decoder in a simple microprocessor. A single register drives the instruction decoder and the static values for the control decoders.

4.4 Pipelining

The first microprocessors used only a single register to hold the instruction being executed, as is shown in Fig. 9. This is the simplest hardware to support instruction decoding, but also leads to a limited execution model. The processor first fetches the instruction into the instruction register. Once the instruction is fetched, it is decoded to determine initial state for the state sequencer. The state sequencer cycles through the proper states to execute the instruction. The state sequencer provides the dynamic control information to the control decoder and the instruction register provides the static control information (e.g., pointers to the operand source and destination register), and so it cannot be changed until after the last use of any of the static control information.

Figure 10 is the execution model for a microprocessor using the simple instruction decoder shown in Fig. 9. In Fig. 10, each row represents a single instruction and time progresses from left to right. Each instruction has to go through fetch, decode, and execute stages before the next instruction can be allowed to start. The microprocessor fetches the first instruction, then decodes the first instruction, and then executes the first instruction. When the first instruction is completed, the microprocessor fetches the second instruction, then decodes the second instruction, and so on. In the simplest microprocessors, these stages (fetch, decode, and execute) were not overlapped because the instruction decoder had only a single register. If a new instruction is fetched into the instruction register before the previous instruction completes execution, the pointer to the result register for the previous instruction would be overwritten by the new instruction. The performance bottleneck in early microprocessors is the controller (since the external bus, decoder, and execution unit are busy only every third cycle).

Figure 11 is an instruction decoder to support an instruction pipeline. This instruction decoder has separate registers for fetch, decode, and execute. This instruction decoder configuration makes the execution model of

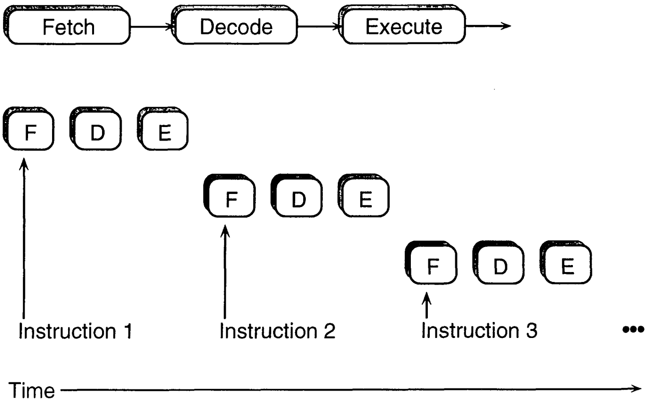

FIG. 10. In the simplest microprocessor implementation, the CPU steps sequentially through fetch, decode, and execute cycles.

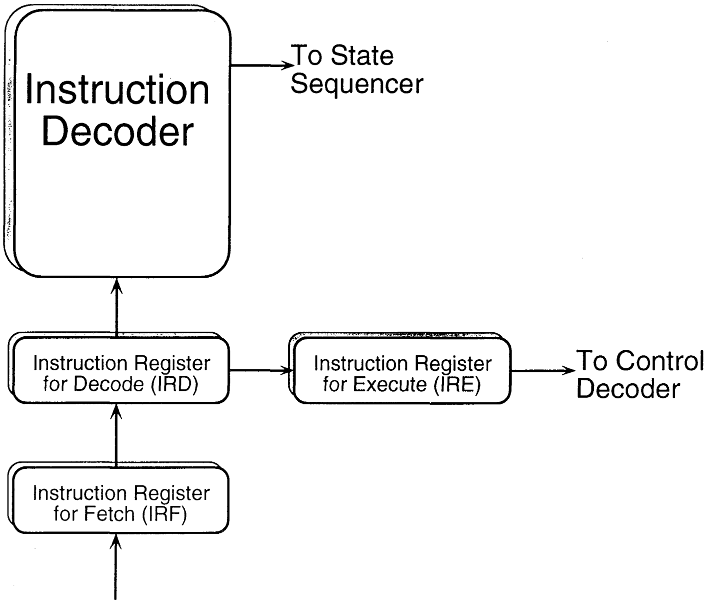
Instruction Decoder
To State Sequencer
Instruction Register for Decode (IRD)
Instruction Register for Execute (IRE)
To Control Decoder
Instruction Register for Fetch (IRF)

FIG. 11. The instruction decoder for overlapped execution has separate instruction registers for fetch, decode, and execute.

Fig. 12 possible. The IRE holds the instruction ready for execution and supplies static control information (e.g., pointers to the registers) to the control decoder. At the same time, the IRD holds the next instruction for the instruction decoder. The IRF is available to capture the next instruction. All three registers can be active at the same time.

Figure 12 shows the execution model for the microprocessor with the instruction decoder of Fig. 11. Since there are separate registers to preserve the instruction values for the decode and execute phases, fetch, decode, and execute can be overlapped. The bottleneck in the microprocessor is the memory system. One instruction is completed every cycle, but the external bus is busy on every cycle. If one of the instructions is a load or store, additional cycles will be introduced and subsequent instructions will be correspondingly delayed. In Fig. 12 each row represents an instruction, and time flows from the left of the diagram to the right. The first instruction (at the top of the diagram) is fetched. As the second instruction is fetched, the first instruction is decoded. As the third instruction is fetched, the second is decoded, and the first is executed. To determine resources required

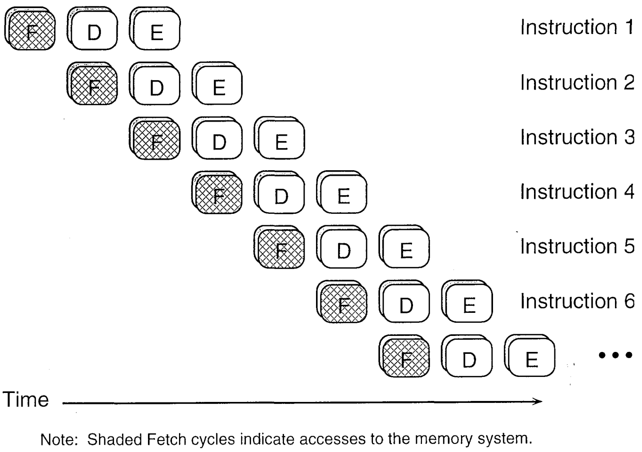
F
D
E
Instruction 1
Instruction 2
Instruction 3
Instruction 4
Instruction 5
Instruction 6
• • •
Time

Note: Shaded Fetch cycles indicate accesses to the memory system.

FIG. 12. Execution model for a simple instruction pipeline. Each instruction is represented by a fetch, decode, execute row. A new instruction enters the pipeline on each cycle.

to build this microprocessor, the designer need only look at a vertical slice in the execution model (this gives a complete sample of activity during a single cycle).

The execution model of Fig. 12 shows an instruction completing every cycle and it shows the external bus being used on every cycle. But it does not show any loads and stores. What happens if the microprocessor executes loads and stores? A load instruction will have to read an operand from the memory system. During the execute cycle for the load instruction, the operand is read from the memory system into the microprocessor. The only external bus available to the microprocessor will be used for the operand read transaction, so that the instruction fetch, which normally would occur simultaneously with the execute cycle (for the instruction after the one currently being decoded and two instructions after the one being executed) will have to be delayed. If an instruction fetch is delayed, there will be no instruction to decode in the next cycle, and so on. Bubbles are introduced in the pipeline to share the external bus between the instruction fetches and data reads and writes. If this happens very often, it will not be possible to complete an instruction each cycle. Suppose that for those same load and store instructions, execution unit resources (such as the ALU) are needed to calculate the operand address. During the cycles when the execution unit resources are being used to calculate the operand address, they are not available for the execute phase of another instruction. If the execute phase of an instruction is blocked, the decode phase and the fetch for the two instructions behind it are prevented from advancing. Another bubble is introduced in the pipeline.

Figure 13 is an early microprocessor for overlapped instruction execution. There is only a single external bus for instructions and data. The execution unit contains a dual-port register file. Overlapping fetch, decode, and execute gave good performance improvement over earlier microprocessors using sequential execution of fetch, decode, and execute, while not significantly increasing required resources.

Figure 14 shows the instruction pipeline extended to prevent delays caused by operand reads, writes, and address calculation. I have extended the execution model from fetch, decode, and execute (F, D, E, respectively) to fetch, decode, address calculate, operand read, execute, and operand write (F, D, A, R, E, W, respectively). Not every instruction uses all of these execution phases, but every instruction must pass through all of the phases.

A vertical slice through this execution model determines the resources required to support it. Supporting requirements for a vertical slice in this execution model will ensure that delays are not introduced because of a lack of resources. The fetch stage will need a single read port to the external memory system. The address calculate stage may require two ports to the register file (the contents of two registers may be added to compute an operand address). The operand read stage may require a read port to the external memory and a write port to the register file (some address calculations update the value

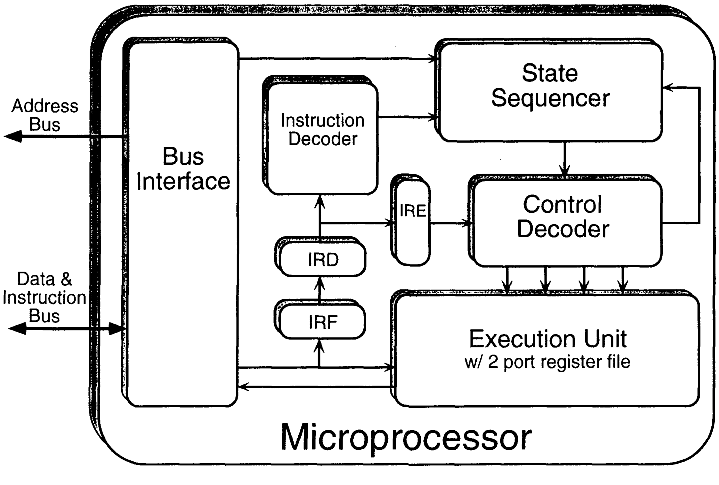

FIG. 13. Block diagram of a microprocessor with overlapped instruction execution. The instruction decoder contains three instruction registers, the register file has two read/write ports, and there is a single external bus for instructions and data.

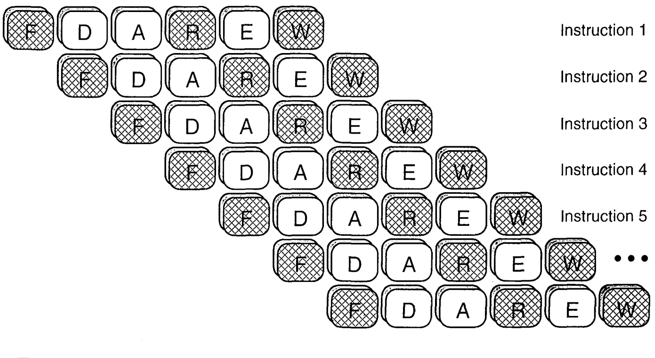
F
D
A
R
E
W
Instruction 1
Instruction 2
Instruction 3
Instruction 4
Instruction 5
• • •
Time
Note: Shaded cycles indicate accesses to the memory system (for instruction fetch, operand read, or operand write).

FIG. 14. The extended instruction pipeline has fetch, decode, address calculate, operand read, execute, and write stages. Pipelining is increased to reduce conflicts that delay instruction execution.

of a register used to calculate an operand address). The execute stage may require two read ports to the register file. The operand write stage may require a write port to the external memory system and a write port to the register file. To support the extended instruction pipeline, the microprocessor will need two read ports and one write port to the memory system. It will also need two ALUs, one for address calculation and one for instruction operation, and a six-port register file. Many of the recently introduced high-end microprocessors, such as the 80486 (see Fig. 15) and the Motorola 68040 (see Fig. 16), implement an execution model similar to the extended instruction pipeline. Early microprocessors contained a few thousand transistors. The 80486 and the Motorola 68040, which were both announced in 1989 (the 80486 shipped in 1989 and the 68040 shipped in 1991), contain about 1.2 million transistors each on a chip of approximately 1.5 cm^2 area.

In the execution model of Fig. 14, it looks as if the microprocessor has finally reached the goal of executing one instruction every cycle, even if there are loads and stores. This is not so. Instructions such as multiply and divide might tie up the execution unit for several cycles. If they do, they block everything behind them for a few cycles, too. How does the microprocessor prevent bottlenecks in the execution unit from delaying the whole pipeline? One solution advocates throwing out all the instructions taking more than a single execute cycle. One might also provide enough hardware to execute any instruction in a single cycle. A third solution, becoming more popular in the industry, is to provide multiple execution units. This solution is based on the premise that instructions taking more than one cycle will not usually occur next to each other. In other words, instructions will be distributed over the available execution units to improve performance reasonably.

With multiple execution units, it becomes possible to initiate or complete the execution of more than a single instruction each processor cycle (Diefendorff and Allen, 1992). Implementations attempting to issue multiple instructions per cycle are called superscalar. Superscalar implementations attempt to break through the barrier of one instruction per cycle as the theoretical maximum execution rate. Figure 17 is a die photo of the Motorola MC88110, which is a modern superscalar microprocessor. It employs multiple integer execution units, multiple floating-point execution units, a graphics execution unit, a load/store execution unit, and separate on-chip instruction and data caches.

The return on invested hardware is decreasing. The microprocessor improved performance by a factor of 3 initially by adding two registers to the instruction decoder to overlap instruction fetch, decode, and execute. The next improvement, the extended instruction pipeline, added considerable hardware to the execution unit and memory system and lots of pins to the microprocessor, and the improvement was only seen when operand accesses would have conflicted with instruction fetches. Since operand accesses

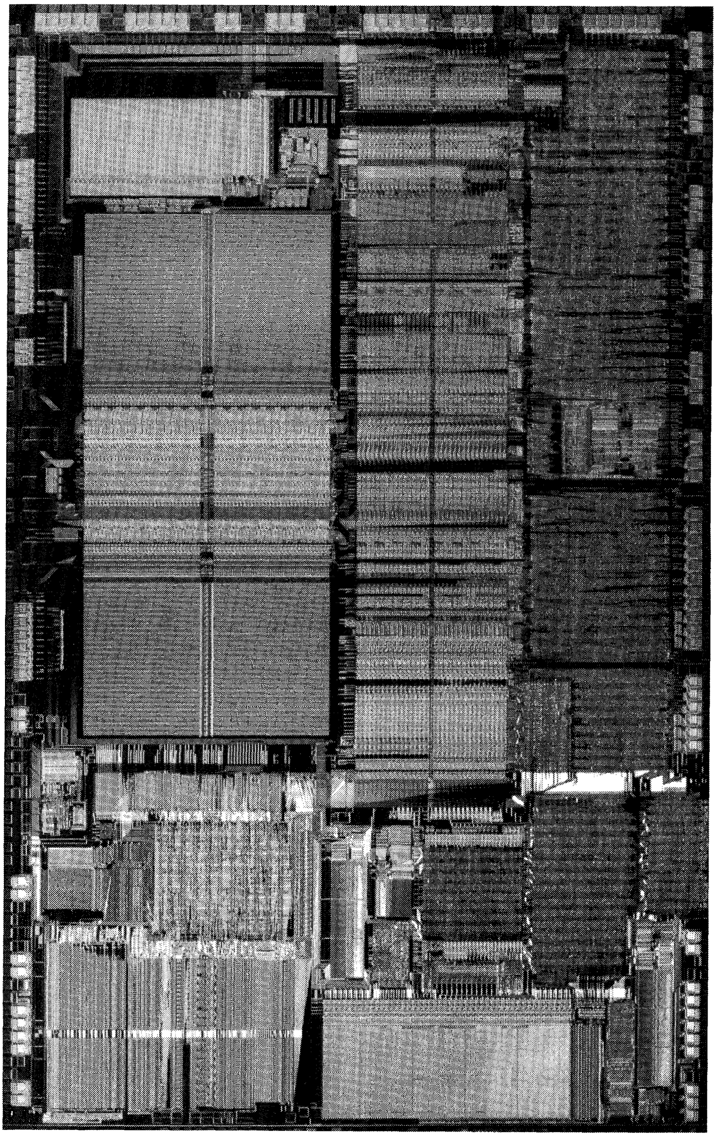

FIG. 15. The Intel i486™, containing 1.2 million transistors, is a good example of a modern microprocessor with an extended instruction pipeline. The Intel i486™ has an 8-kB combined instruction and data cache. (Courtesy Intel Corporation.)

are less than half the frequency of instruction accesses, performance is now only one and a half times the original performance. This is still a substantial improvement, but the design has added substantial complexity. Adding multiple function units and allowing the decoder to issue multiple instruction per cycle has the potential to improve performance substantially, but there is a corresponding increase in complexity. The decoder must decide which instructions are eligible to execute concurrently. If out-of-order completion of instructions is allowed, cleaning up the state of the processor after an exception may be difficult.

Hardware can also be added to compensate for delays introduced by operand or address conflicts. An operand conflict occurs when the result of the current instruction is needed by some following instruction. The pipeline is stalled until the result is available. This problem is worse if the result of the current instruction is needed to complete the address calculation for the following instruction. Adding hardware for these cases can improve performance. New microprocessors,

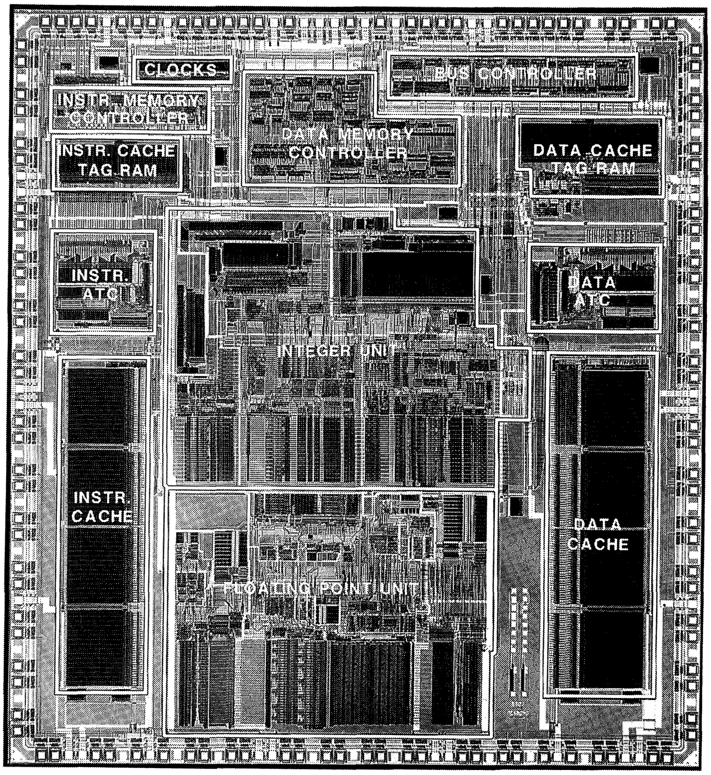
CLOCKS
BUS CONTROLLER
INSTR. MEMORY CONTROLLER
DATA MEMORY CONTROLLER
INSTR. CACHE TAG RAM
DATA CACHE TAG RAM
INSTR. ATC
DATA ATC
INTEGER UNIT
INSTR. CACHE
DATA CACHE
FLOATING POINT UNIT

FIG. 16. Motorola's MC68040 is a good example of a modern microprocessor with an extended instruction pipeline. The MC68040 has separate buses to its internal 4-kB instruction cache and internal 4-kB data cache. (Reprinted with permission of Motorola.)

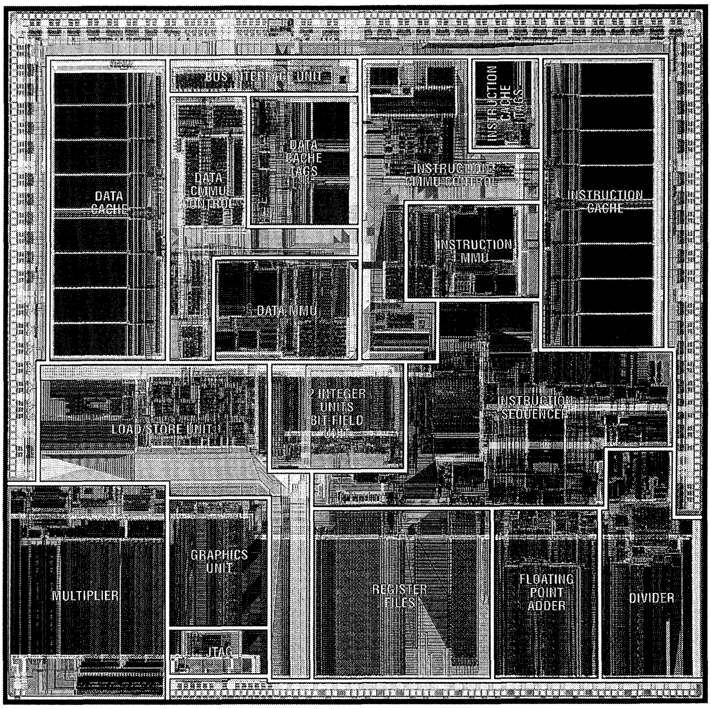
BUS INTERFACE UNIT
DATA CACHE
DATA CMMU CONTROL
DATA CACHE TAGS
INSTRUCTION CACHE TAGS
INSTRUCTION CMMU CONTROL
INSTRUCTION CACHE
INSTRUCTION MMU
DATA MMU
LOAD/STORE UNIT
2 INTEGER UNITS BIT FIELD UNIT
INSTRUCTION SEQUENCER
MULTIPLIER
GRAPHICS UNIT
JTAG
REGISTER FILES
FLOATING POINT ADDER
DIVIDER

FIG. 17. Motorola's MC88110, containing 1.2 million transistors, is one of the first superscalar microprocessors. A superscalar microprocessor is capable of issuing and completing more than a single instruction each clock cycle. (Reprinted with permission of Motorola.)

however, are still adding more and more hardware for less and less return in performance.

4.5 Branch Execution Model

The description to this point tells how a microprocessor works and reviews possible hardware improvements from a pipeline point of view. These improvements consider throughput but have not considered the effect of branches. Figure 18 is an execution model showing the effect of a branch instruction with the branch condition satisfied. A branch taken causes a significant penalty because the pipeline is blocked until the outcome of the branch condition is known. Since the outcome of a branch may not be known until the end of the execute cycle, the fetch to the branch destination instruction may not begin until the write cycle of the current instruction. A five-cycle bubble is introduced in the pipeline. If branches occur frequently, performance could suffer considerably.

But branches are frequent. Branches constitute about 20% of instructions by dynamic frequency of occurrence. Figure 19 illustrates the effect of branches on processor performance. The vertical axis is processor performance normalized so that performance is one in the absence of branch instructions. The horizontal axis shows branch frequency of occurrence. The family of curves represents the likelihood that a branch is successful. The calculations assume that instructions other than branches complete every 1.5 clock cycles on average. Branches are 5 clock cycles if taken and 1 clock cycle if not taken. The example lines show that for a branch frequency of occurrence of 20% (a branch occurs every fifth instruction) and a branch success rate of only 40%, the performance of the processor is only about 70% of its performance in the absence of branches. If the branch success rate is 80%, the performance is reduced to only 58% of performance in the absence of branches. Plenty of effort in microprocessor design goes into reducing the cost of branches.

One suggestion for reducing the cost of branches is to incorporate a delayed branch instruction. The instruction immediately after the delayed branch instruction is always

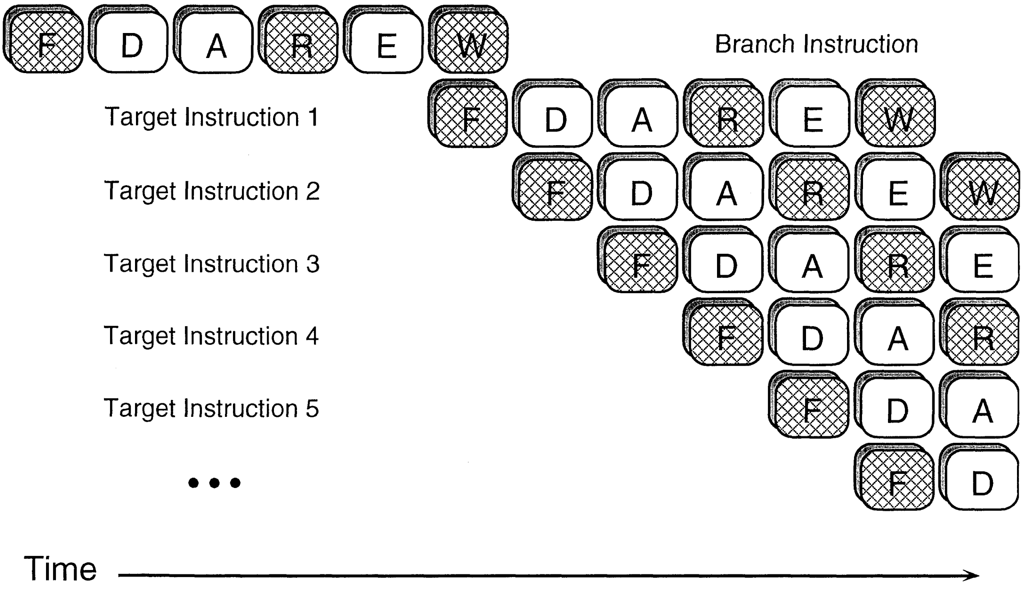

Note: Shaded cycles indicate accesses to the memory system (for instruction fetch, operand read, or operand write).

FIG. 18. Execution model showing the effect of a branch instruction with the branch condition satisfied. The top row is the branch instruction. The second row is the target of the branch. The pipeline is blocked until the outcome of the conditional branch is known.

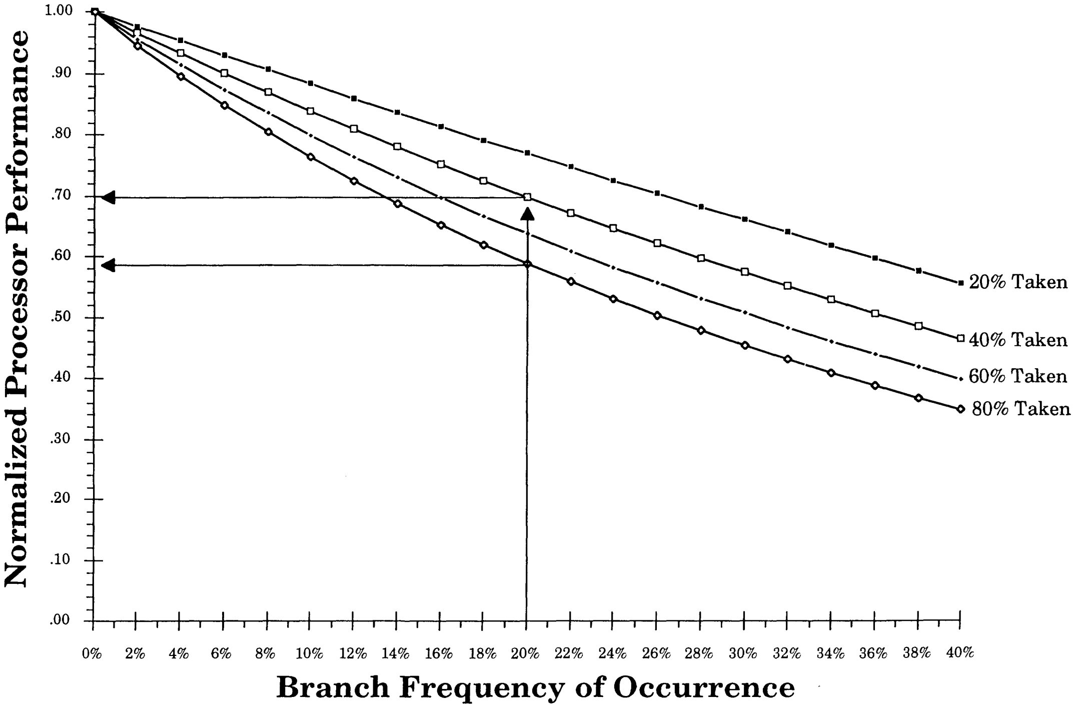

FIG. 19. Branch penalty significantly reduces processor performance. If branches occur 20% of the time and are successful (taken) 40% of the time, the processor is only 70% as efficient as it would be if there were no branches.

executed. This recovers one of the cycles lost when a branch is successful and is based on the premise that the instruction after the branch was already mostly through the pipeline when the outcome of the branch was decided, so its execution might as well finish.

The branch prediction cache, shown in Fig. 20, tracks the program counter and maintains a record of branches recently encountered. When the hardware sees a branch instruction, it checks the branch prediction cache. The branch prediction cache may have recorded the success rate for this branch instruction. If the instruction is expected to branch, the branch prediction cache provides the target address to the memory system. With enough hardware and early recognition, the entire branch penalty may be eliminated. Hardware is added to detect the branch instructions, record the branch history and branch target addresses, and recover in the event the branch does not meet its prediction.

The branch target cache is a variation of the branch prediction cache. The branch target cache is much like the branch prediction cache except that it stores the target instructions rather than the address of the target instructions. This further helps reduce branch penalty because the branch target cache can provide the target instructions to the decoder and help cover the latency in accessing the external memory system.

4.6 External Bus Protocol

Figure 21 shows how the microprocessor communicates with the memory system to obtain the program instructions (and data). The microprocessor, using the common bus connection, sends the memory system a request saying "I want to read an instruction from the following memory location," or, perhaps, "I want to write a result to the following memory location." What the microprocessor is really placing on the bus is a bunch of ones and zeros (the address) representing the location of the instruction in the memory and some tags indicating what kind of transaction (read or write, for example) the microprocessor is requesting. For an instruction read request, the memory system will

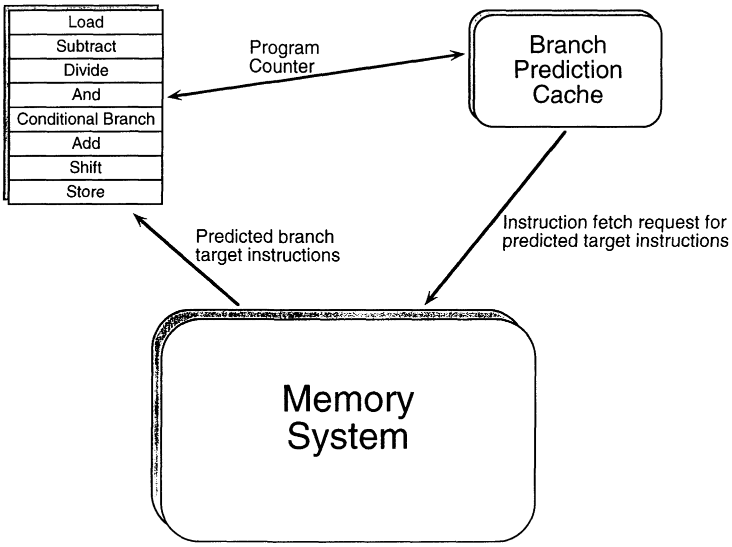
Load
Subtract
Divide
And
Conditional Branch
Add
Shift
Store
Program Counter
Branch Prediction Cache
Predicted branch target instructions
Instruction fetch request for predicted target instructions
Memory System

FIG. 20. The branch prediction cache keeps track of program branching history and attempts to fetch the target of branches expected to succeed before the pipeline stalls.

retrieve the contents of the memory location and place the value on the data and instruction bus lines along with a tag saying "Here's the instruction." Once the microprocessor has taken the instruction, it will signal the memory system "Thank you" to tell the memory system it is okay to end the current transaction. If the microprocessor initiates the transaction and waits an undetermined number of cycles for the response from the memory system, the transaction is an asynchronous handshake protocol. If the microprocessor initiates a transaction and expects a response after a fixed number of clock cycles, the transaction is a synchronous protocol. Fast microprocessors in 1992 can do about 20 million transactions with an external memory system every second. Today's computer system built with a microprocessor is significantly faster and more powerful than the giant computers of only a few years ago.

5. TECHNOLOGY AND MANUFACTURING

The original Intel 4004 microprocessor in 1971 contained about 2300 transistors (Smith, 1992), was designed with about 8-μm feature sizes (Hennessy and Jouppi, 1991), operated at 108 kHz, and came in a 16-pin package. The early packages for microprocessors were

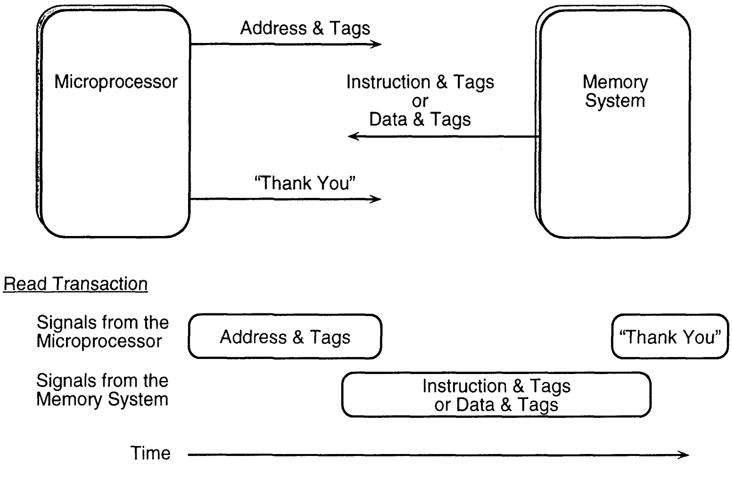
Address & Tags
Microprocessor
Instruction & Tags or Data & Tags
Memory System
"Thank You"
Read Transaction
Signals from the Microprocessor
Address & Tags
"Thank You"
Signals from the Memory System
Instruction & Tags or Data & Tags
Time

FIG. 21. A transaction between the microprocessor and the memory system using a handshake protocol.

designed to be compatible with the contemporary transistor-transistor-logic packages, so that they could be more easily designed into standard circuit boards. The most popular package for 8-bit microprocessors such as the Zilog Z80 was the 40-pin plastic package. The Motorola MC68000, announced in 1979, is a 32-bit processor with a 16-bit external interface and was introduced in a 68-pin ceramic package. In 1992, Intel introduced a version of the 80486 containing 1 200 000 transistors, designed with 0.8-μm feature sizes, operating internally at 66 MHz, in a 168-pin package. Figure 22 shows the progression of transistors on Intel microprocessors from the 4004 through the 80486 to the latest-generation microprocessor, now on the market as Pentium (data from Smith, 1992). Data points labeled with the name of the microprocessor indicate number of transistor for each microprocessor generation. The plain line is the theoretical curve for a doubling of the transistor count every two years. This curve shows the combined effects of the reduction in linewidth (allowing more transistors per unit area) and increase in manufacturable chip sizes (allowing more transistors per chip). Linewidths have decreased from about 8 μm in 1971 to between 0.8 and 0.5 μm in 1992 (Hennessy and Jouppi, 1991). As the linewidth decreases by a factor of 2, the number of transistors in a fixed area goes up by a factor of 4. Doubling the size of the chip doubles the number of transistors on the chip.

The first microprocessors were small chips about 4 mm or less on a side and were produced on small wafers about 50 mm in diameter. Recent microprocessors may be 15 mm on a side and are commonly produced on wafers 150 mm in diameter and sometimes on wafers 200 mm in diameter. Today's highest performance microprocessors may contain 3 000 000 transistors, have about 500 pins, and operate internally at frequencies in the range of 100 MHz. At these frequencies, external chips connected to the microprocessor will not be able to keep up with the speed of the microprocessor. One technique being used today to help solve this problem allows the microprocessor to operate at some multiple (2×, 3×, 4×) of the external system clock rate. In the near future, microprocessors will contain more than 10 000 000 transistors, be designed with 0.2-μm feature sizes, and operate internally above 200 MHz (5-ns cycle time).

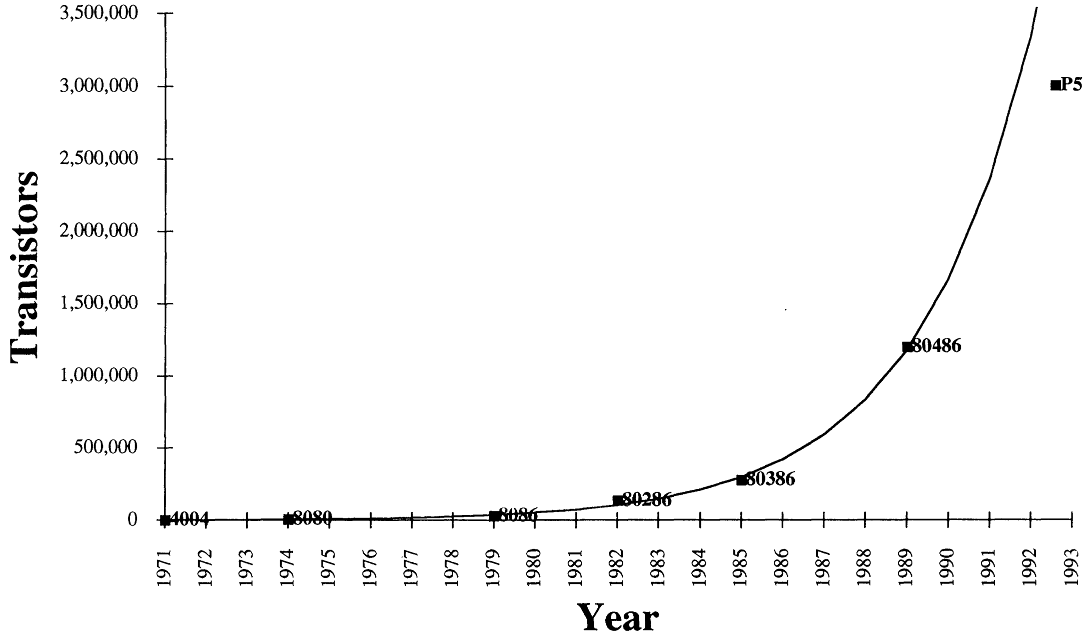

FIG. 22. The number of transistors in some Intel microprocessors from 1971 to 1993. The point labeled P5 refers to the microprocessor that has now reached the market as Pentium.

6. APPLICATIONS

6.1 The Microprocessor as a CPU

More than 21 million IBM-compatible personal computers were shipped in 1991. All of those computers used microprocessors based on the architecture of the Intel 80x86. Personal computers from Apple, Atari, and Commodore and based on the Motorola 680x0 microprocessor shipped probably between 2 and 3 million systems in 1991 (McWilliams, 1992). In addition to the personal computer market, there is a smaller market for engineering work stations. Engineering work stations are more expensive than personal computers and are based on chips from many different companies such as Intel, Motorola, MIPS, IBM, Cypress, DEC, Intergraph, Texas Instruments, and Hewlett-Packard. The total shipment of engineering work stations from all companies in 1991 was below 5×10^5 (Slater, 1992). There are also other computer systems using the microprocessor as the CPU, but their shipping volumes do not account for a significant number of microprocessor sales. The total market for the microprocessor used as a CPU in a computer system is probably below 30 million units.

Microprocessors designed for use as the CPU in a computer system may cost from below $50 to well over $1000, depending on the development cost, technology of the processor, and performance features of the design. Computer applications for microprocessors fall into two general categories: desktop computers (designed to be plugged into a permanent power source) and portable computers (designed with an internal battery for some operation independent of a permanent power source). Today, the same microprocessors serve for both applications. In the future, microprocessors designed for desktop computers will diverge from microprocessors designed for portable computers. Microprocessors designed for desktop computers will emphasize performance. The transistor budget for these microprocessors will be spent on special circuits for multiply and divide, additional arithmetic units for address calculation, on-chip instruction and data caches, and instruction and execution pipelining. Microprocessors designed for portable computers will emphasize low power and maximum function (to reduce the number of additional chips required to build a system) because portable computers often run using batteries and battery life is an important measure of perceived market value for the user. Performance is still important, but the first consideration is conserving power. The transistor budget for these microprocessors will be used to integrate as many of the computer system peripheral functions as possible, such as the serial and parallel communication ports, keyboard scanner, and memory controller. In addition, microprocessors for portable computers will have sophisticated on-chip power management circuits. Microprocessors for portable computers will become more like microprocessors for embedded control.

6.2 Embedded Control

An embedded control application is any use of a microprocessor other than as the CPU in a computer system. By contrast with the relatively small number of microprocessors used as CPUs in computer systems (fewer than 30 million per year), a relatively large number of microprocessors are used in embedded control applications. Microprocessors are used in microwave ovens, automobiles, printers, computer keyboards, calculators, games, gasoline pumps, and toys. They are everywhere. In 1991, about 1.5×10^9 microprocessors were shipped worldwide for use in embedded control applications.

Requirements of embedded-control applications make them significantly different from applications for microprocessors used as the CPU in a computer system. Most computer systems demand maximum performance from the CPU, and so the design of microprocessors for computers is driven primarily by performance. High-volume applications in embedded control are driven by cost. In these applications, the microprocessor competes to displace standard logic modules, and so the microprocessor must provide adequate performance (equal to the standard logic) at a lower cost. The requirements of these two applications (CPUs and embedded control) have split the design of microprocessors along two diverging lines. As the underlying technology improves, microprocessors designed for use as CPUs in computer systems use the additional transistors, speed, and chip area to improve the performance of the microprocessor. Microprocessors designed for embedded-control applications take advantage of improvements in chip technology to increase chip function, to reduce power consumption, and to reduce chip size. They try to squeeze onto the chip as much of the target application's system as possible. These chips, called microcontrollers or microcomputers, may incorporate most of the functions of the simple computer system from Fig. 2, including ROM, RAM, peripheral-interface logic, and even the peripheral functions themselves. In high-volume embedded-control applications, the average selling price of the microprocessor may range from under $1 to about $3.

GLOSSARY

Address Mode: Information in an instruction that says how to calculate the address of the operand(s).

ALU (Arithmetic and Logic Unit): The ALU is used for all general-purpose arithmetic.

Architecture: The view of the computer as seen by the user. Generally includes the instruction set, address modes, instruction formats, and registers visible to the computer user.

Asynchronous (Hardware): Not synchronized with the system clock. An event is asynchronous if its occurrence depends on signals other than the system clock.

Block Diagram: A drawing, consisting primarily of labeled boxes, showing the logical structure of part of a computer.

Bus Controller: The logic on a microprocessor chip controlling the protocol at the chip's pins and providing synchronized signals to the state sequencer.

Cache: A smaller, faster memory that sits between the CPU and the main memory to give the main memory the appearance of higher speed. CPUs are faster than memory, and so caches can serve a speed-matching function.

Central Processing Unit: See CPU.

Condition Code: An attribute of the result of an ALU operation. The following are fairly standard condition codes: c = carry out (this is borrow—not for subtraction; n = sign of the result (1 is negative, 0 is positive); v = overflow (significant result bits lost); z = zero result.

CPU (Central Processing Unit): The engine controlling the operation of a computer system. It contains the arithmetic and logic units, program counter, register file, and state sequencer.

Decoder: Logic that maps an input bit pattern to some new output bit pattern.

IC: Integrated circuit; a chip; literally, a "circuit in one place."

Implementation: The logical representation of a functional specification (e.g., the logic design implements the specification in the user's manual).

Interrupt: A signal that disrupts the normal sequence of events.

Microcode: The contents of a control store. Microcode is the collective term for the bit patterns controlling the actions of the execution unit and the state sequencer.

Microprocessor: A single integrated circuit containing at least the functions of a CPU in a computer system.

Multiplexer: Logic that selects a single output from several inputs.

Operand: A quantity on which an operation is to be performed.

Register: Separately addressed (from memory) storage location, at least conceptually, in the CPU.

Semiconductor: A material whose electrical conductivity is between that of an insulator and that of a conductor.

Synchronous (Hardware): Synchronized with the system clock. An event is synchronous if its occurrence depends only on a clock.

Works Cited

Bell, C. G., Mudge, J. C., McNamara, J. E. (1978), *Computer Engineering*, Bedford, MA: Digital, Chap. 1.

Davis, H., Wharton, M. (1983), *Inside the Chip*, London: Usborne.

Diefendorff, K., Allen, M. (1992), "Organization of the Motorola 88110 Superscalar RISC Microprocessor," *IEEE Micro* **12** (2), 40–63.

Faggin, F. (1992), "The Birth of the Microprocessor," *Byte* **17** (3), 145–150.

Hennessy, J. L., Jouppi, N. P., Gill, J., Baskett, F., Strong, A., Gross, T., Rowen, C., Leonard, J. (1982), in: *The MIPS Machine*, Proceedings of COMPCON, pp. 2–7, New York: IEEE.

Hennessy, J. L., Jouppi, N. P. (1991), "Computer Technology and Architecture: An Evolving Interaction," *Computer* **24** (9), 18–29.

McWilliams, G. (1992), "The Reward of Many RISCs," *Business Week* No. 3258, p. 84.

Money, S. A. (1990), *Microprocessor Data Book,* 2d ed., San Diego: Academic, Chap. 1.

Patterson, D. A., Séquin, C. H. (1980), "Design Considerations for Single-Chip Computers of the Future," *IEEE Trans. Comput.* **C-29,** 108–116.

Rao, G. V. (1978), *Microprocessors and Microcomputer Systems,* New York: Van Nostrand Reinhold, Chap. 4.

Slater, M. (1992), "Wave of High-End Processors Due," *Microproc. Rep.* **6** (2), 9–15.

Smith, B. R. (1981), *Computers,* London: Usborne.

Smith, R. (1992), "Intel to delay rollout of next-generation microprocessor," *San Jose Mercury News* 23 July, pp. 1F, 6F.

Stallings, W. (1990), *Computer Organization and Architecture: Principles of Structure and Function,* 2d ed., New York: Macmillan, Chap. 1.

Texas Instruments (1973), *The TTL Data Book for Design Engineers,* Dallas: Texas Instruments.

Tredennick, N. (1991), "MIPs and Sunset," *Microproc. Rep.* **5** (12), 12–17.

Further Reading

Anceau, F. (1986), *The Architecture of Microprocessors,* Workingham, England: Addison-Wesley.

Comerford, R. (1992), "How DEC developed Alpha," *IEEE Spectrum* **29** (7), 26–31.

Glasser, L. A., Dobberpuhl, D. W. (1985), *The Design and Analysis of VLSI Circuits,* Reading, MA: Addison-Wesley.

Gupta, A. (1987), *Advanced Microprocessors, II,* New York: IEEE.

Linderholm, O., Malloy, R., Reinhardt, A., Sheldon, J. M. (1991), "A Talk with Intel," *Byte* **16** (6), 131–140.

Mirapuri, S., Woodacre, M., Vasseghi, N. (1992), "The MIPS R4000 Processor," *IEEE Micro.* **12** (4), 10–22.

Money, S. A. (1990), *Microprocessor Data Book,* 2d ed., San Diego: Academic.

Myers, W. (1991), "The Drive to the Year 2000," *IEEE Micro.* **11** (1), 10–13, 68–74.

Oehler, R. R., Blasgen, M. W. (1991), "IBM RISC System/6000: Architecture and Performance," *IEEE Micro.* **11** (3), 14–17, 56–62.

Patterson, D. A., Hennessy, J. L. (1990), *Computer Architecture: A Quantitative Approach,* San Mateo, CA: Morgan Kaufmann.

Sachs, H. G., McGhan, H., Hanson, L. F., Brookwood, N. A. (1991), "Design and Implementation Tradeoffs in the Clipper C400 Architecture," *IEEE Micro.* **11** (3), 18–21, 74–80.

Slater, M. (1987), *Microprocessor-Based Design: A Comprehensive Guide to Effective Hardware Design,* Mountain View, CA: Mayfield.

Stallings, W. (1990), *Computer Organization and Architecture: Principles of Structure and Function,* 2nd ed., New York: Macmillan.

Stockley, C., Watts, L. (1983), *Usborne Guide to Computer Jargon,* London: Usborne.

Tatchell, J., Bennett, B. (1982), *Understanding the Micro,* London: Usborne.

Tredennick, N. (1987), *Microprocessor Logic Design,* Bedford, MA: Digital.

Williams, T. (1992), "Superscalar Sparc chips offer performance gains, compatibility," *Comput. Design,* July, 32–36.

MICROSCOPY, ACOUSTICAL

See SCANNING ACOUSTICAL MICROSCOPY

MICROSCOPY, ELECTRON

See ELECTRON MICROSCOPY

MICROSCOPY, FIELD-ION

See FIELD-ION MICROSCOPY

MICROSCOPY, OPTICAL

See OPTICAL MICROSCOPY

MICROSENSORS

SIMON MIDDELHOEK, *Department of Electrical Engineering, Delft University of Technology, Delft, The Netherlands*

Introduction ... 322
1. **General Principles** ... 322
1.1 Measurement and Control Systems ... 322
1.2 The Six Signal Domains ... 322
1.3 Sensor Technologies ... 323
2. **Microsensors for Radiant Signals** ... 324
2.1 The Electromagnetic Spectrum ... 324
2.2 Basic Physical Principles ... 324
2.2.1 Low Photon Energies ... 324
2.2.2 High Photon Energies ... 324
2.2.3 Interaction of Nuclear Particles with Silicon ... 325
2.3 Microsensors for Visible Light ... 325
2.3.1 Photoconductors ... 325
2.3.2 Photodiodes ... 326
2.3.3 Position-Sensitive Photodetectors (PSDs) ... 326
2.3.4 Charge-Coupled Devices (CCDs) ... 327
3. **Microsensors for Mechanical Signals** ... 327
3.1 Basic Physical Principles ... 327
3.2 Diffused-Diaphragm Pressure Microsensors ... 328
3.3 Capacitive Pressure Sensors ... 329
3.4 Accelerometers ... 329
4. **Microsensors for Thermal Signals** ... 330
4.1 Basic Physical Principles ... 330
4.2 Thermocouple-Based Microsensors ... 330
4.4 Silicon Diodes and Transistors as Thermal Sensors ... 333
5. **Microsensors for Magnetic Signals** ... 334
5.1 Basic Physical Principles ... 334
5.1.1 Hall Effect in Metals and Semiconductors ... 334
5.1.2 Magnetoresistance Effects ... 335
5.1.2.1 Physical Magnetoresistance Effect ... 335
5.1.2.2 Geometrical Magnetoresistance Effect ... 335
5.1.2.3 Magnetoresistivity in Magnetic Layers ... 335
5.2 Silicon Hall Plates ... 336
5.3 Magnetotransistors ... 336
5.4 Magnetic Recording Heads ... 337
6. **Microsensors for Chemical Signals** ... 337
6.1 Chemoresistors and Chemocapacitors ... 338
6.1.1 Chemoresistors ... 338
6.1.2 Chemocapacitors ... 338
6.2 MOSFET-Based Chemical Microsensors ... 338
6.2.1 Ion-Sensitive Field-Effect Transistors (ISFETs) ... 339
6.2.2 Biosensors ... 339
6.2.3 Pd-Gate MOSFET Gas Microsensors ... 340
7. **Processing of Microsensors** ... 341
7.1 Silicon Planar Technology ... 341
7.2 Micromachining Technologies ... 341
7.2.1 Etching Solutions and Gases ... 341
7.2.2 Etching Procedures and Techniques ... 342
8. **Interfaces, Smart Sensors, and Bus Systems** ... 343
[illegible]
8.2 Binary Encoded Signals and Bus Systems ... 344
Glossary ... 345
Works Cited ... 346
Further Reading ... 346

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

The rapid development of microelectronics over the last two decades has been celebrated in many places and on many occasions. There is general agreement that the phenomenal improvement in the performance/price ratio of microelectronic components has given impetus to what is commonly referred to as the "information society," in which not only materials and energy but also information and information-processing systems play a major role. A further penetration of microelectronic components into electronic and currently nonelectronic information processing systems can be foreseen. However, this penetration is being impeded by the lack of input transducers or sensors for the measurement of physical and chemical parameters whose performance/price ratio is comparable to that of microelectronic components (Middelhoek and Audet, 1989). One solution to this problem might be to apply the batch-fabrication technology used for the production of integrated circuits to the fabrication of sensors. The resulting sensors are called *microsensors*, since the technology by which they are made allows them to be quite small.

We begin by defining the role of sensors in measurement systems and then present some microsensors for a variety of physical and chemical parameters. We conclude with a discussion of the special processing sequences used to fabricate sensors and electronic circuits which interconnect sensors and microprocessors.

1. GENERAL PRINCIPLES

1.1 Measurement and Control Systems

The term "information-processing systems" encompasses a wide spectrum ranging from computers and word processors to oscilloscopes, door locks, clinical thermometers, satellites, automatic vending machines, and so on. All these systems process information in one way or another. An important subgroup is that of measurement and control systems. Like all information-processing systems, these systems can be represented by a sequence of three elements (Fig. 1). The first of these is the input transducer, more often called the *sensor*. In the input transducer, a physical or chemical quantity, such as light intensity, displacement, temperature, magnetic field, or *p*H value, is converted into an electrical or electronic signal. A great number of physical and chemical effects are available for this purpose. A sensor is distinguished by the fact that the energy carrying the signal is converted from a nonelectrical form into an electrical form.

In the second element—the signal processor—the electronic signal is, as the name implies, processed in some way. The processor can be as simple as an amplifier, but it can also consist of a microprocessor or even of the central processing unit of a large-scale computer. In contrast to sensors, in a processor the form of the energy carrying the signal is not converted into another form.

Last, in the output transducer the electrical signal is once again converted into a nonelectrical signal, which can be detected by at least one of the human senses. For example, in a display device the signal is carried by radiant energy and can be perceived by our eyes.

1.2 The Six Signal Domains

From a theoretical point of view, the process of information transduction in a sensor can be viewed as a coupling between the states of two systems (Duyn and Middelhoek, 1992). If a change of state in one system affects the state in the other system, we say that the two systems are coupled and that this coupling enables information to be exchanged between the systems. To devise a useful classification of sensors, it is necessary to examine the state variables used to describe a

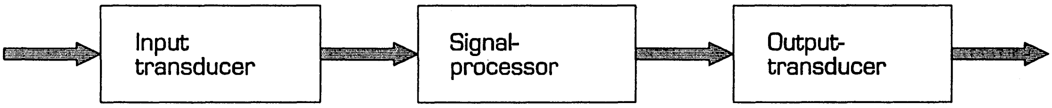

FIG. 1. Functional block diagram of a measurement system.

system. Two types of state variables exist: extensive state variables, such as mass, charge, strain, entropy, and polarization, and intensive state variables, such as electric potential, stress, temperature, and electric and magnetic fields. If we consider the products of the matching intensive and extensive state variables, we can define six energy domains and, since signals are carried by energy, also six signal domains: radiant, mechanical, thermal, electrical, magnetic, and chemical.

Using the six energy domains, we can depict the signal transduction in a sensor by the general diagram shown in Fig. 2. Each circle represents a signal or an energy domain. A sensor is represented by an arrow from a nonelectrical signal domain to the electrical domain. For instance, a photodetector (vertical arrow) will convert a radiant signal into an electrical signal.

When the various physical and chemical effects in sensors are reviewed, it appears that some effects can be used to make sensors that do not require auxiliary energy sources for the signal conversion, while others only yield useful sensors when some form of energy is supplied. The first group, referred to as *self-generating sensors,* includes, for example, the solar cell based on the photovoltaic effect and the thermocouple based on the Seebeck effect. The second group, referred to as *modulating sensors,* includes the photoconductor based on the photoelectric effect and the pressure sensor based on the piezoresistive effect.

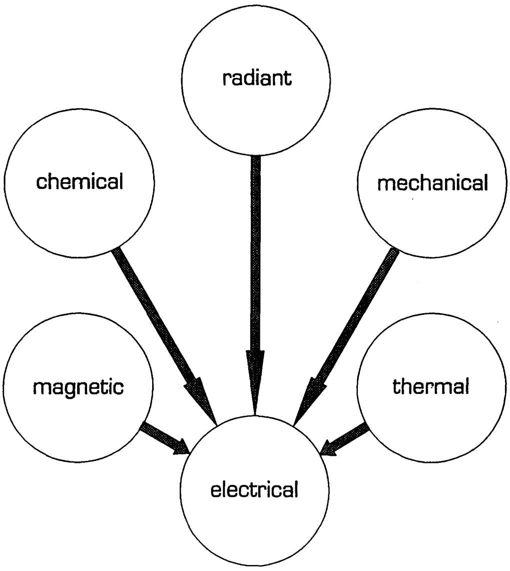

FIG. 2. Diagram indicating the five possible signal conversions in sensors.

1.3 Sensor Technologies

As already stated, the introduction of microelectronic components into traditionally nonelectrical information-processing systems could be more successful if more sensors with an acceptable performance/price ratio were available. Therefore, a global search for new and improved sensor technologies was initiated. These days, the main technologies for fabricating sensors are based on piezoelectric materials, such as quartz, ceramics, and polymers, on gas-sensitive metal oxides, on III-V and II-VI semiconductors, such as GaAs, GaP, AlSb, InSb, CdS, CdSe, ZnO, and ZnS, on thick- and thin-film materials, on optical glass fibers, and last but not least, on monocrystalline silicon. Progress in silicon planar technology has exceeded the most daring predictions. This has led to an increasing number of very sophisticated and yet cost-effective very large-scale integrated (VLSI) components, such as microprocessors and memories. Therefore, it made sense to apply this technology to sensors as well. As a result, the first silicon photodetectors were produced as early as 1960. Ten years later, silicon microsensors for the measurement of magnetic fields, pressure, and *p*H were presented. Scientific literature currently abounds with descriptions of a wide variety of silicon microsensors, many of which are commercially available. The attractiveness of using silicon technology for sensors lies in the fact that a large number of wafers, each containing hundreds of highly similar sensors, can be simultaneously processed as one batch, markedly improving the performance/price ratio. Moreover, advanced photolithographic techniques permit the fabrication of very fine patterns on a silicon substrate so that the size of silicon sensors can be very small; hence the name microsensors. Their small size makes it possible to use them for many applications, particularly in the medical field, where small size is crucial.

Silicon has a wide variety of physical and chemical effects that are suitable for the construction of sensors for many measurands. In many sensors, the silicon must be able to be machined into shapes such as thin membranes, cantilevers, and resonators. Several chemical, electrochemical, and dry plasma etching techniques have been developed to achieve this in silicon. These so-called micromachining techniques are discussed in Sec. 7. Fabricating sensors and electronic circuits with almost similar processing steps makes it possible to combine the sensors and circuits on one chip. Such sensors are called *smart sensors,* and they are discussed in the last section (Sec. 8).

2. MICROSENSORS FOR RADIANT SIGNALS

2.1 The Electromagnetic Spectrum

The radiant-signal domain involves the electromagnetic radiation of all frequencies as well as nuclear-particle radiation. Electromagnetic radiation includes microwaves, far- and near-infrared light, visible light, ultraviolet light, soft x rays, x rays, and gamma rays. The difference between the various types of electromagnetic radiation lies in their characteristic frequencies. Nuclear-particle radiation includes alpha particles, beta particles, neutrons, and minimum-ionizing particles, resulting from targets bombarded with charged particles in particle accelerators.

Thanks to research efforts dating back to the 1960s, numerous types of microsensors have been developed for both application fields. In this section, we briefly discuss the basic principles and the most important microsensors for radiant signals.

2.2 Basic Physical Principles

2.2.1 Low Photon Energies A wide range of interactions occurs between photons and silicon. Some interactions result in the conversion of the photons into an electrical charge (the photoelectric effect), while other interactions convert impinging photons into entities that have no net electric charge (phonons, excitons, and other photons). The only effects of interest here will be those involving the absorption of photons in silicon. During this absorption process, electron-hole pairs are freed. The pairs are then collected, in the process of which radiant energy is transformed into electrical energy. When the energy of an incident photon exceeds the band gap E_g of a semiconductor, the photon can be absorbed by exciting a valence electron into the conduction band. This band-to-band absorption results in the creation of an electron-hole pair. The free electron and hole then become excess carriers and are free to contribute to the conductivity of the material.

If the semiconductor is a direct band-gap material such as gallium arsenide, an incident-light photon can excite an electron directly from the valence band into the conduction band because the electron does not need to change its momentum. Thus, once the incident-photon energy equals the band-gap energy, the absorption coefficient stays virtually the same, even when the photon energy is increased further.

In order for an incident-light photon to excite a valence electron into the conduction band of an indirect band-gap semiconductor such as silicon or germanium, the electron must change its momentum. The probability of such an interaction taking place depends on the presence of a phonon with the required momentum. Because the phonon density increases as the momentum decreases, and as a relatively smaller change in momentum is required to generate electron-hole pairs of higher-energy photons, absorption increases with increasing photon energy. This effect can be put to great use in a color microsensor (Wolffenbuttel, 1987).

2.2.2 High Photon Energies Interactions of incident electromagnetic radiation at high photon energies, i.e. within the x-ray and gamma-ray regions, occur with atomic electrons. X rays and gamma rays are absorbed by three main processes, each characterized by different absorption coefficients: the photoelectric effect, the Compton effect, and pair production. The total absorption coefficient is the sum of these three. The process that takes place depends on the photon energy (Fig. 3). The photoelectric effect dominates at lower energies, from visible light up to about 0.1 MeV. If a photon interacts in a photoelectric process, it will transfer all of its energy to the interacting electron and disappear. The highest probability that this process will occur is with a *K*-shell electron. The excited electron, falling down to the conduction-band edge, will behave like a charged particle by creating other electron-hole pairs.

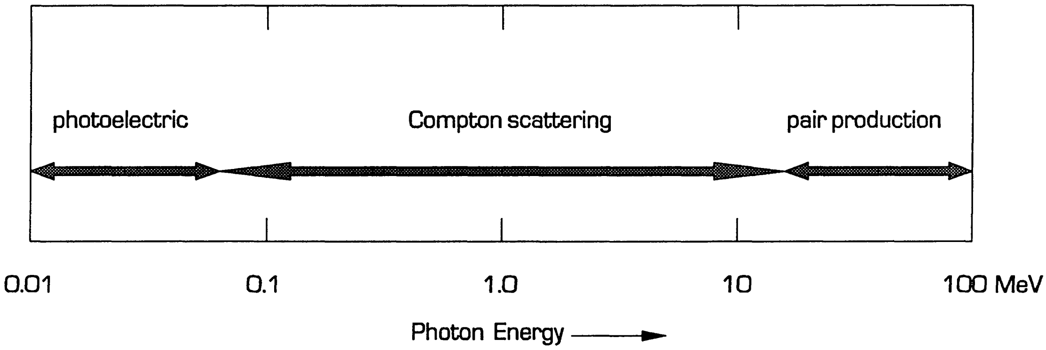

FIG. 3. Interaction processes in silicon with photons of high energy. The photoelectric effect dominates at lower energies, from visible light up to about 0.1 MeV. From 0.1 MeV up to 15 MeV the Compton process dominates, and above 15 MeV production of electron-positron pairs will dominate.

From 0.1 MeV up to 15 MeV, the Compton process dominates. In this type of interaction, only part of the photon energy is transferred to the electron. The original photon is deflected and decreased in energy. The deflected photon can then be reabsorbed by the photoelectric or Compton processes, or it can escape from the silicon.

Electron-positron pairs will start to be produced at 1.022 MeV and will dominate at energies above 15 MeV. In this interaction, the original photon will disappear while the excess energy will be imparted to the electron-positron pair. The electron and positron lose their energy through collisions and eventually come to rest. The positron and an electron from the crystal then annihilate. This interaction produces two gamma rays, each having an energy of 511 keV, which are emitted in opposite directions. The gamma rays can then be reabsorbed by the photoelectric or Compton processes, or they can escape from the silicon.

2.2.3 Interaction of Nuclear Particles with Silicon When alpha particles and other heavy charged particles enter a silicon crystal, they lose their energy through a succession of Coulomb interactions that occur between the particle's electromagnetic fields and the crystal's electrons. In each electronic interaction, a bound electron is freed and brought into the conduction band. The average ionization energy, i.e., the average energy necessary to create one electron-hole pair in silicon at room temperature, is 3.6 eV and is much smaller than the 30 eV needed by gaseous detectors. The ionization energy in silicon is constant for different types of radiation and for different energies. As the free electron in the conduction band will have an energy of only a few keV, it will lose this energy by creating secondary electron-hole pairs. The resulting ionization is distributed as a track, centered on the path of the particle. For heavy charged particles, the ionization track is comparatively straight, with only minor deflections from the original path direction, as these particles do not become deflected from their track by collisions with crystal electrons (Fig. 4).

The interactions of beta particles and other light charged particles are similar to those previously mentioned for heavy charged particles. Despite their physical differences, all high-energy nuclear particles tend for the most part to interact with silicon through Coulomb interactions, producing comparatively straight ionization tracks. They are therefore known as minimum-ionizing particles and produce approximately 85 electron-hole pairs per micron in silicon. Several silicon-based detectors for nuclear particles have been constructed. The best known of these are the silicon-microstrip detector and the silicon drift chamber. These detectors can be used to determine both the energy and the position of incident nuclear particles (Ten Kate, 1986).

Neutron detectors cannot be made with silicon as the interaction of neutrons with silicon is nil.

2.3 Microsensors for Visible Light

2.3.1 Photoconductors Photoconductors or photoresistors can be made in silicon by creating *n*-type or *p*-type areas in the usual

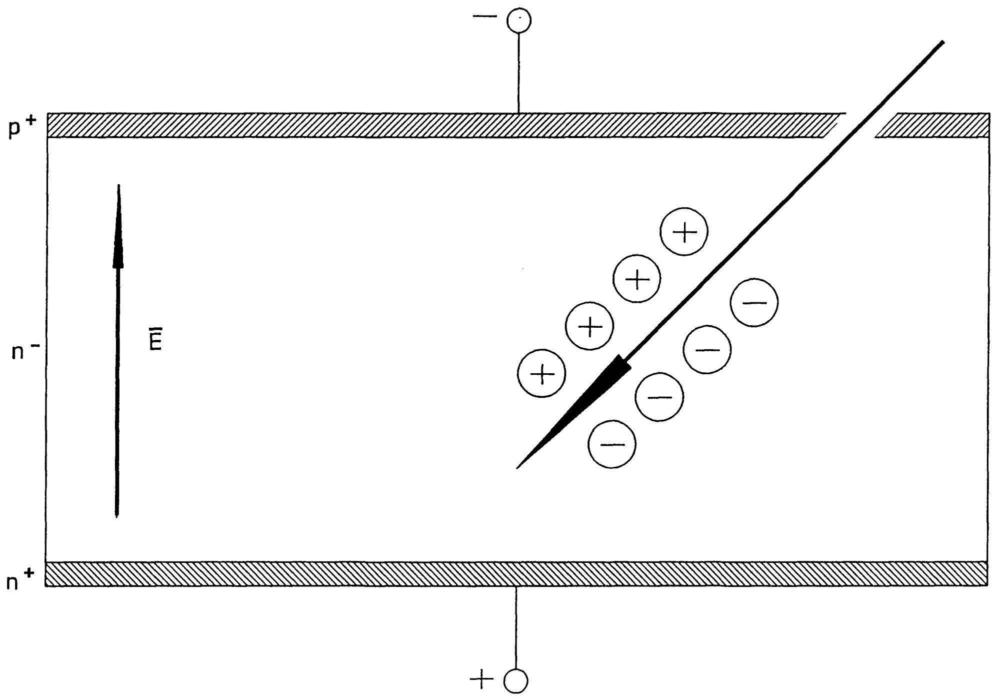

FIG. 4. Schematic of a silicon nuclear-particle detector.

way. Incident radiation creates electron-hole pairs, causing an increase in conductivity. To obtain a high sensitivity and a low switching time, the lifetime of the charge carriers must be large with respect to the transit time, i.e., the time required for the charge carriers to move between the electrodes of the photoconductor.

2.3.2 Photodiodes Silicon photodiodes can be used either without a bias voltage or with a reverse bias voltage. When a reverse bias voltage is applied, incident light creates electron-hole pairs in and around the depletion region, and these pairs are swept across the junction (Fig. 5). This creates a photocurrent, which is a measure of the light intensity. Because of the wavelength dependence of the absorption coefficient in silicon, radiation with a short wavelength (blue) is absorbed relatively close to the silicon surface, while radiation with a longer wavelength (red) penetrates deeper into the silicon. In order to collect the electrons and holes generated by the red light, the reverse bias voltage must be taken larger than for blue light. Measuring the photocurrents at three or more different bias voltages allows us to extract information about the spectrum and color of the incident light (Wolffenbuttel, 1987).

A photodiode can also make use of the photovoltaic effect. Here as well, the electric field in the depletion layer will, without a bias voltage, separate the charge carriers generated by the incident light. The *n*-type layer will become negative with respect to the *p*-type layer, and the voltage will be a measure of the incident light.

2.3.3 Position-Sensitive Photodetectors (PSDs) A special kind of photodiode is the position-sensitive photodetector. If the illumination of a *p*-*n* junction is nonuniform, i.e., produced by a light spot much smaller than the photodiode surface, a lateral photovoltage is established. By placing a pair of Ohmic contacts on opposite edges of an *n*-type layer of the junction, we can fabricate a device and with it measure the lateral photovoltage $V(x)$ as a function of the location of the light spot. $V(x)$ will vary in magnitude and polarity as the light spot is moved between the two contacts. The x coordinate of the light spot can be obtained from this voltage. If two additional contacts in the y direction are made to

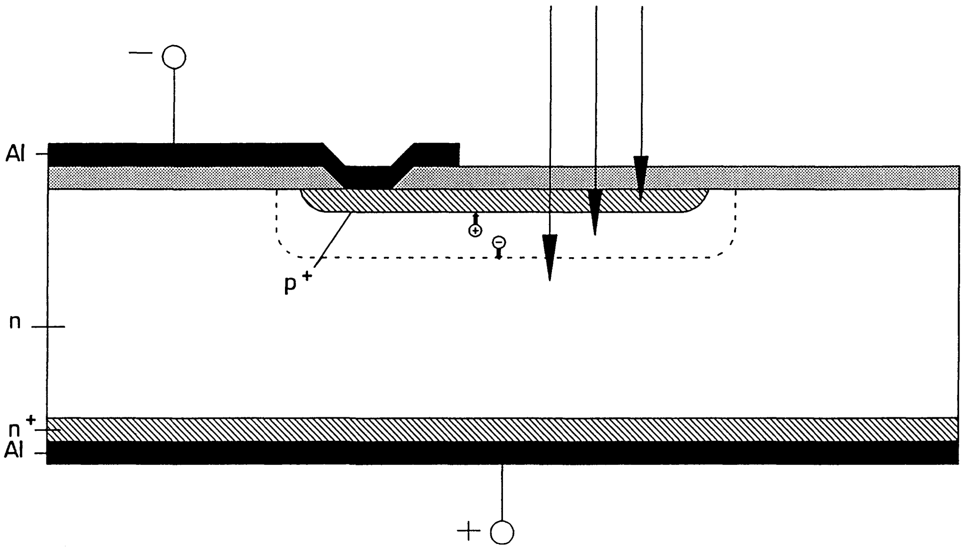

FIG. 5. Cross section of a photodiode. The depletion region is indicated with the dotted line.

the p-type layer, both the x and y coordinates can be obtained from voltages $V(x)$ and $V(y)$. Devices with a resolution of better than 1 μm have been fabricated.

2.3.4 Charge-Coupled Devices (CCDs) A very interesting and commercially valuable microsensor is the charge-coupled device (*q.v.*). CCDs are currently being applied in solid-state TV cameras. When a MOS (metal–silicon dioxide–silicon) capacitor is exposed to light, charge carriers are generated and are stored on the capacitor. Placing many such MOS capacitors side by side makes it possible to transfer this charge between the capacitors. The basic structure of a CCD consists of an array of electrodes placed on the silicon-dioxide layer. Clock voltages can be applied to these interconnected electrodes to allow for the transfer of stored charges. The CCD structures require a three-phase clock system to ensure one-directional charge movement. Three MOS capacitors are thus required for each elemental cell. A schematic of a basic three-phase CCD is shown in Fig. 6. Different configurations have been introduced to form two-dimensional CCD arrays. One approach uses a two-dimensional CCD array with an image and a storage section. During a certain time period, the image section of the CCD is exposed to the light, and charge is formed in this section. These charge packets are then transferred to the storage section and are read out one line at a time (Collet, 1986). The fabrication of CCDs is extremely difficult, as an acceptable image sensor has to contain at least 1 million functioning pixels.

3. MICROSENSORS FOR MECHANICAL SIGNALS

3.1 Basic Physical Principles

Microsensors for mechanical signals are used to measure

1. motion-related measurands, such as position, displacement, surface roughness, velocity, flow, and speed of rotation; and
2. force-related measurands, such as weight, pressure, acceleration, torque, strain, attitude, and vibration.

Most solid-state sensors for mechanical signals are based on the piezoresistive effect. The change in resistance of a metallic conductor when subjected to a mechanical strain was first reported by Lord Kelvin in 1856. As the change is caused by changes in the geometry of the resistors, it is referred to as the *geometric piezoresistive effect*. This effect has been used to produce strain gauges that are widely used in industry today. Since the piezoresistive effect in metals is rather small,

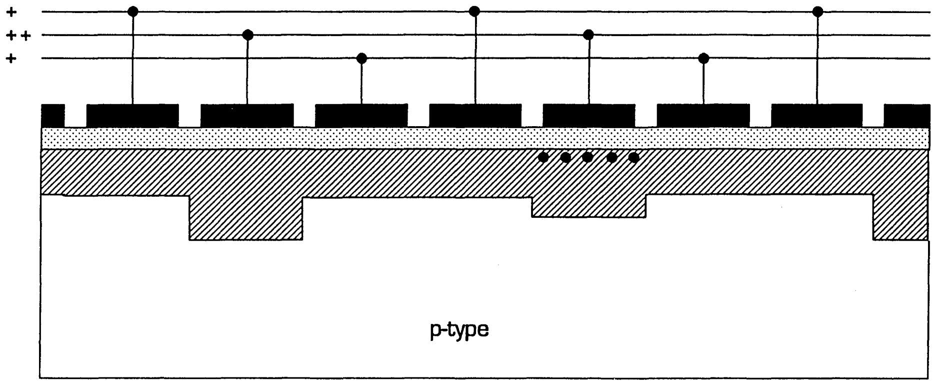

FIG. 6. Schematic of a three-phase MOS *n*-channel CCD.

researchers turned their attention to semiconductors.

Smith (1954) reported experiments indicating that the effect in silicon was about a hundred times larger than in metallic conductors. Smith showed that *p*-type silicon has a very large gauge factor of up to 175 and that *n*-type silicon has a large negative gauge factor of −135. The gauge factor of a resistance strain gauge is defined as the relative change of the resistance divided by the relative change of the length of the resistance.

To understand the piezoresistive effect in *n*-type silicon, we have to study how the band structure is affected by the application of a stress. The forbidden energy band in silicon is the result of the interaction between moving electrons and the periodic lattice. When an anisotropic stress is applied, the lattice spacing increases in one direction, whereas it decreases in the directions perpendicular to this direction. As we might expect, the interaction between the electrons and the lattice is also affected: The mobility of electrons moving in the [100] direction decreases when the crystal is compressed in this direction. The conductivity in this direction decreases as well. The relative change in the resistivity can be expressed by

$$d = dR/R = d\rho/\rho = \pi_{\parallel}\sigma_{\parallel} + \pi_{\perp}\sigma_{\perp} \qquad (1)$$

in which $\sigma_{\parallel}$ is the stress parallel to and $\sigma_{\perp}$ the stress perpendicular to the current direction. Further, $\pi_{\parallel}$ and $\pi_{\perp}$ are the parallel and perpendicular piezoresistive coefficients. These coefficients depend very much on the orientation of the elongated resistor with respect to the crystal axes and on the type of silicon. For (100) planes, $\pi_{\parallel}$ and $\pi_{\perp}$ are virtually identical and are highly angle dependent. The (111) crystal plane is sometimes favored because in this plane the coefficients for longitudinal as well as transverse stresses do not depend on the crystal direction.

3.2 Diffused-Diaphragm Pressure Microsensors

Gieles (1969), working at the Philips Research Laboratories in The Netherlands, was the first to describe a miniature silicon diffused-diaphragm pressure sensor based on the piezoresistive effect. Many research laboratories around the world then set to work on this novel sensing device, drawn by the unique features offered by silicon pressure microsensors. The general structure of a pressure microsensor is shown in Fig. 7. It consists of a housing with one or two pressure ports. At the heart of the pressure sensor is a silicon chip etched to form a thin diaphragm. The diaphragms can be fabricated by spark erosion or chemical etching techniques. The piezoresistors are diffused or ion-implanted on the upper surface of the diaphragm. When the pressures on the two sides of the diaphragm are unequal, the diaphragm will bend. This will cause a change in the piezoresistors which can easily be detected.

A wide variety of configurations and positions of the resistors on the diaphragm can be found in the literature. There are four different types of sensors: absolute, sealed, gauge, and differential pressure sensors. Absolute pressure sensors feature the right pressure

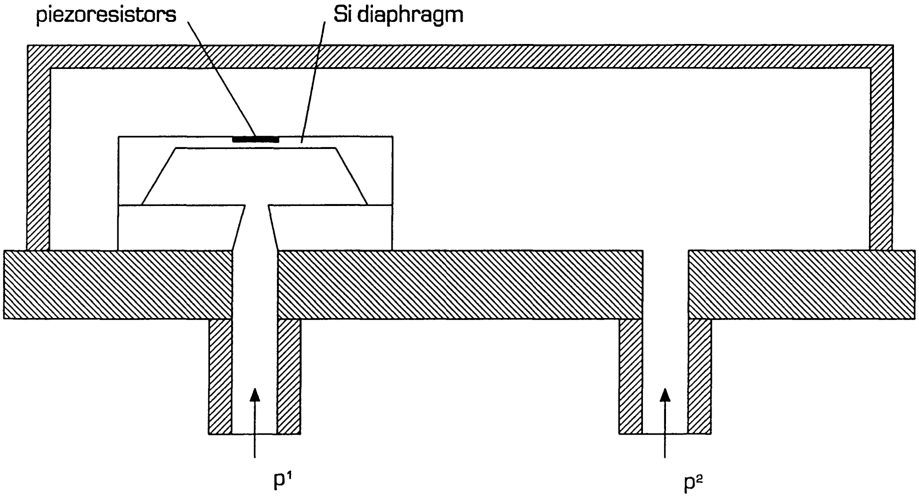

FIG. 7. Schematic drawing of a silicon-diffused diaphragm pressure microsensor.

port only, which connects to the upper surface. The sensor cavity is sealed in a vacuum and is the reference at the bottom of the diaphragm. The upper surface of the diaphragm might be covered by a silicone gel or other material to protect the surface from humidity or corrosive fluids. Differential pressure sensors have two pressure inlet ports. Pressure sensors with all kinds of operating pressure ranges can be obtained by varying the diameter and thickness of the diaphragms. While very sensitive pressure sensors with a range from 0 to 10 kPa exist, sensors with a range from 0 to 200 MPa are also available. Bridge voltages usually lie between 5 and 10 V, and the sensitivity of silicon pressure sensors varies from 10 mV/kPa for low-pressure devices to 0.001 mV/kPa for high-pressure devices.

Today, silicon pressure sensors are used in a large number of applications. They are especially successful in the medical and automotive markets. Silicon pressure sensors can be found in altimeters, blood-pressure gauges, tire-leak detectors, pushbuttons, microphones, and level detectors.

3.3 Capacitive Pressure Sensors

The pressure sensors presented in the preceding section are based on the piezoresistivity effect in silicon. Piezoresistive pressure sensors use a deformable diaphragm in which the strain-sensitive components are diffused or implanted. However, a metallized silicon diaphragm can also form an electrical capacitance with, for instance, an overlying metallized glass plate (Ko, 1986). When the diaphragm bends, because of a pressure change, this capacitance will also change. This capacitance change can be measured. The measurements show that capacitive pressure sensors are much more sensitive than piezoresistive sensors. However, in capacitive sensors the leads connecting the pressure sensor to the outside tend to show stray capacitances that are of the same order as the sensor capacitance. This makes it necessary to include electronic circuitry on the sensor chip in order to convert the capacitance change into a reliable electrical signal.

3.4 Accelerometers

With the help of micromachining, it is also possible to make silicon accelerometers. Roylance and Angell (1979) were the first to present such a device. Their accelerometer, which still forms the basis for many present-day devices, consists of a very thin 15-μm cantilever containing a diffused piezoresistor. A 200-μm-thick silicon seismic mass is formed at the end of the cantilever, and this serves as the inertial reference (Fig. 8). A 200-μm-thick silicon supporting rim surrounds the beam and the mass. When the device is accelerated, the seismic mass will change its position with respect to the surroundings and

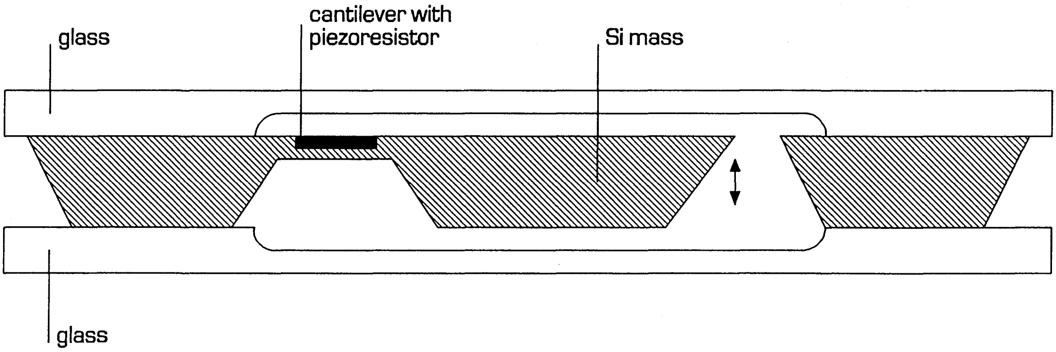

FIG. 8. Cross section of miniature piezoresistive accelerometer.

cause the cantilever to bend. The piezoresistor in the cantilever will change its value by an amount that can easily be detected. A device of this sort is capable of measuring accelerations from 0.01*g* to 100*g*.

4. MICROSENSORS FOR THERMAL SIGNALS

4.1 Basic Physical Principles

Since in most applications of conventional electronic circuits the temperature sensitivity of the components, such as diodes and transistors, is undesirable, this sensitivity has been subject to extensive research. Not unexpectedly, this accumulated knowledge has also been useful for the design of silicon-based temperature sensors, where temperature sensitivity is a desirable property. As discussed in the Introduction, sensors can be divided into self-generating and modulating sensors. The self-generating sensors require no source of power other than the signal being measured. Temperature sensors based on thermocouples belong to this category. Modulating sensors are temperature sensors based on the thermal modulation of electric currents or voltages supplied by auxiliary energy sources, for example resistors, diodes, and transistors. In the following sections, we first discuss self-generating temperature sensors based on the thermoelectric effect and then modulating sensors based on resistors, diodes, and transistors.

4.2 Thermocouple-Based Microsensors

Three thermoelectric effects are known to exist: the Seebeck effect, the Peltier effect, and the Thomson effect. The Seebeck effect is the only one of the three that can be used to design microsensors for the measurement of temperature. The Seebeck effect describes the experimental observation that when two conductors made of different materials are joined together at one point and a temperature difference is maintained between the joined and the nonjoined parts, an open-circuit voltage will develop between the nonjoined parts. For small temperature differences, the voltage is proportional to the temperature difference and is strongly dependent on the materials that are used to form the junctions. When the temperature difference between the hot junction (T_2) and the cold junction (T_1) is given by ΔT, the voltage ΔV can be expressed by

$$\Delta V = \alpha_{ab} \Delta T, \qquad (2)$$

where α_{ab} is defined as the relative Seebeck coefficient between materials A and B. Many phenomena contribute to the Seebeck effect in silicon, such as the change in the Fermi level, band-gap changes, charge-carrier-concentration gradient changes, diffusion-coefficient changes, thermodiffusion, and phonon drag. The Seebeck effect in silicon still requires further theoretical and experimental study in order to explain all observed findings.

The Seebeck coefficients can be measured by using structures as shown in Fig. 9. *p*-type strips are made by ion implantation or dif-

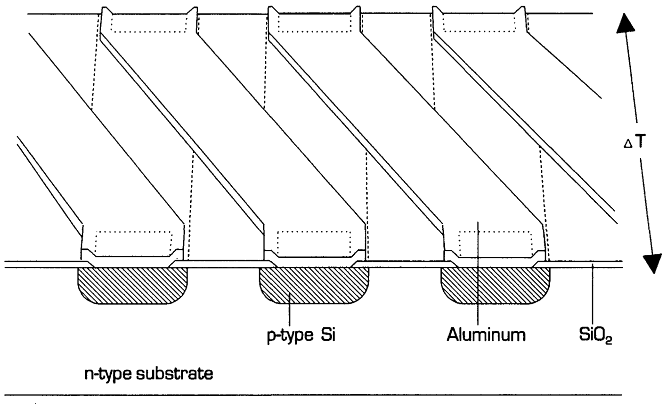

FIG. 9. Schematic drawing of a silicon/aluminum thermopile.

fusion in an *n*-type epitaxial layer. Many strips can be fabricated simultaneously and connected by means of evaporated aluminum strips, forming a thermopile. Such thermopiles have been shown to have a very high sensitivity, and Seebeck coefficients of 1 mV/K per couple can be achieved. The largest Seebeck coefficients are obtained for small impurity concentrations and thus large resistivities. Silicon thermopiles can be used for the design of quite a number of tandem sensors, such as thermal gas flow, vacuum, infrared, and acceleration sensors (van Herwaarden and Sarro, 1986). By way of example, Fig. 10 depicts a thermopile-based infrared microsensor. An infrared sensor consists of a thermopile located in a thin etched silicon beam. An infrared absorbing layer is placed at the end of the thermopile. The absorbed infrared radiation generates heat, which flows through the beam to the rim where the beam connects to the rest of the silicon chip. The output voltage produced by the thermopile is a direct measure of the infrared radiation. Infrared sensors with a responsivity of 10 V/W have been made.

4.3 Silicon Thermoresistors

The dependence of the resistivity of metals on temperature is fairly predictable, so that the temperature can be determined by measuring the resistance. Microsensors for the measurement of temperature can easily be obtained by depositing metal resistors on top of an oxide-covered silicon substrate. In general, the resistivity of a metal increases as the temperature increases. The temperature coefficient of the resistivity in silicon itself displays a more complex pattern than in

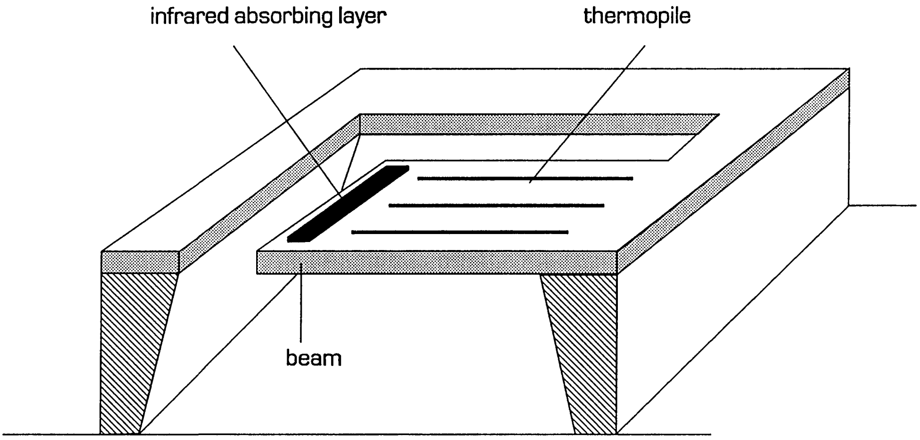

FIG. 10. Cross section of a thermopile-based infrared microsensor.

metals and depends on both the impurity concentration and the temperature. Around room temperature, the temperature coefficient of a silicon resistor is positive for low-doped silicon and constant for high-doped silicon. For temperatures above 200 °C, silicon shows a negative temperature coefficient.

Silicon thermoresistors can be made by diffusing meander-shaped *n*-type resistors in a *p*-type substrate. The *p-n* junction isolates the resistor from the substrate. In thermoresistors processed in this way, there is a large spread in nominal value and temperature coefficient, because both the *p-n* junction and the leakage current are also temperature-dependent. To obtain better silicon resistors, researchers set out to find silicon structures that are not dependent on *p-n* junctions for isolation. The so-called spreading resistance temperature sensor was the result of their efforts. A common technique for measuring the resistivity of a semiconductor wafer is to press a needle on the surface of the wafer. When the diameter of the needle is much smaller than the thickness of the wafer, the resistance *R* appears to be

$$R = \rho/2d, \tag{3}$$

in which *d* is the diameter of the needle and ρ is the resistivity of the material. Thus the needle-silicon surface combination provides us with a resistor that depends only on the resistivity of the silicon and the diameter of the needle and not on the other dimensions of the silicon crystal or on a *p-n* junction. This makes it feasible to design a thermoresistor based on this structure and on the temperature sensitivity of the resistivity. In such a thermoresistor, the needle contact is replaced by a small diffused contact. In Fig. 11, a device of this type consists of a small crystal (0.5 × 0.5 mm^2) with one large flat Ohmic contact on the bottom and a point contact with a diameter of 20 μm on the upper surface. For better contact, n^+ layers are diffused at both the bottom and the point contact so that an n^+nn^+ structure results. A spreading resistance contact can be applied to obtain an even more interesting device that can measure temperatures up to 350 °C.

In order to explain this, it is worth while to consider the product of hole and electron densities in a semiconductor, which can be written as

$$n_n n_p = n_i^2, \tag{4}$$

in which n_n is the electron density, n_p is the hole density, and n_i is the intrinsic carrier density. The variable n_i depends only on the temperature. This means that almost no holes will occur in the layers. When the temperature is increased beyond 200 °C, electrons from the valence band will be thermally excited to the conduction band, and the number of charge carriers will increase markedly so that the resistivity will start to decrease rapidly. When the large bottom contact is made pos-

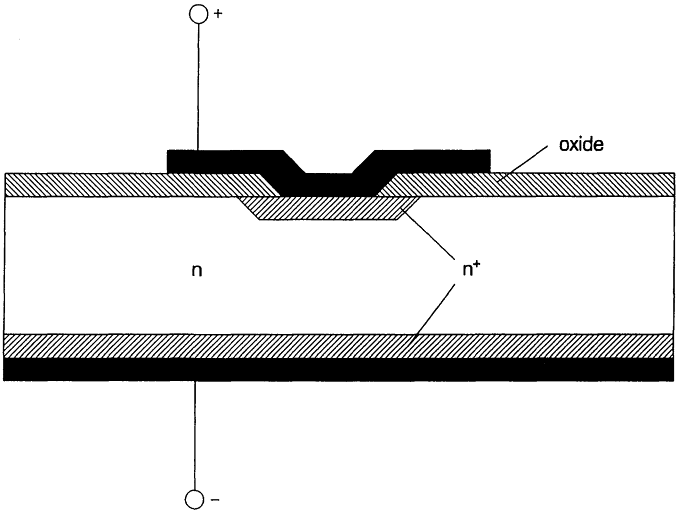

FIG. 11. Spreading resistance thermoresistor.

itive, this is indeed what is experimentally observed for any magnitude of current. However, when the spreading contact is made positive very different and interesting things start to happen: The resistivity continues to increase at higher temperatures, and the silicon shows extrinsic behavior. This unexpected feature of the spreading resistors has to do with the decreasing ability of the n^+ layer at the tiny spreading contact to supply sufficient holes to support the conduction process in the n substrate. This interesting effect has been used to good advantage to produce thermoresistors and other microsensors for temperatures up to 350 °C (Middelhoek and Audet, 1989).

4.4 Silicon Diodes and Transistors as Thermal Sensors

Diodes were already proposed many years ago as simple and low-cost devices for the measurement of temperature in the range between −50 °C and +150 °C. The characteristics of the temperature behavior of p-n junctions are well understood, and diodes are quite often used for simple temperature measurements. When a silicon diode is used as a thermal microsensor, the forward voltage is usually measured as a function of the temperature for a constant small forward current. In practice, measurements on many different diodes show temperature coefficients between 1 and 3 mV/K in the forward direction.

Not unexpectedly, the base-emitter junction of a transistor also shows a temperature sensitivity, which can be employed for temperature measurement. Transistors are generally much better temperature sensors than diodes because their bases are very thin and because they are three-port devices, which makes it much more convenient to incorporate them in electronic circuits.

The use of a transistor as a temperature sensor is usually based on the I_c-V_{be} characteristics. The emitter current I_e consists of a diffusion, a surface leakage, and a recombination component. Because of the thin base, in a transistor I_c consists mainly of the diffusion component, whereas the two other components of I_e are drained away by the base current. The expression for the dependence of V_{be} on the collector current I_c and the temperature T is

$$V_{be} = V_g(0) + (kT/q)\ln[I_c - \ln(KT^r/\eta)]. \qquad (5)$$

The values of $V_g(0)$ and the structure-dependent parameters K and r can be measured for each transistor. The factor η is the ionization factor, which is 1 above 20 K. It has been found that $V_g(0)$ varies between 1.12 and 1.19 eV and r between 3 and 5 for commercially available devices. According to Eq. (5), a transistor can be used as an electrical thermometer. However, in most cases a more sophisticated method is employed. By making use of only one transistor, the collector current can be varied between a fairly high value I_{c1} and a fairly low value I_{c2}, with corresponding base voltages V_{be1} and V_{be2} (Fig. 12). With the aid of Eq. (5) we then find for the voltage difference that

$$\Delta V_{be} = V_{be1} - V_{be2} = (kT/q)\ln(I_{c1}/I_{c2}). \qquad (6)$$

Differentiating this voltage difference with respect to temperature gives

$$\alpha = d(\Delta V_{be})/dT = (k/q)\ln(I_{c1}/I_{c2}). \qquad (7)$$

The result is interesting: The temperature coefficient α is a constant, and ΔV_{be} is exactly proportional to the absolute temperature T (Meijer, 1986). Further, both ΔV_{be} and α are

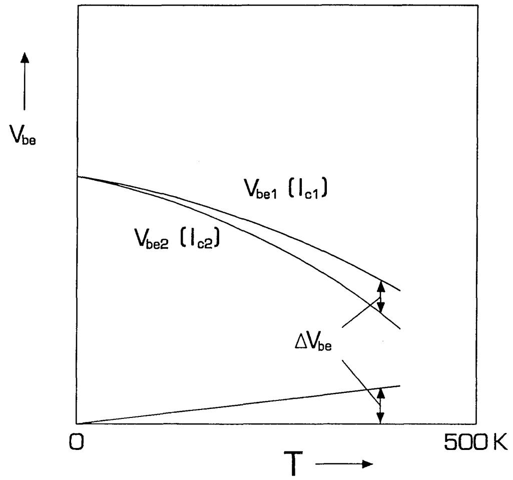

FIG. 12. Base-emitter voltage V_{be} as a function of the absolute temperature for two collector currents. The difference between the base-emitter voltages ΔV_{be} is exactly proportional to the absolute temperature.

independent of material or transistor parameters. Such sensors are known as PTAT (Proportional To Absolute Temperature) sensors. In PTAT sensors, only the current ratio I_{c1}/I_{c2} has to be known. Instead of using one transistor and two currents, it is also possible to use two neighboring transistors with unequal emitter surfaces at equal collector currents. To obtain a well-defined ratio between the emitter surfaces, it is advisable to design the large transistor in such a way that it consists of a multiple of N small transistors in parallel, where N is the desired ratio.

5. MICROSENSORS FOR MAGNETIC SIGNALS

5.1 Basic Physical Principles

In 1879 Edwin Hall, then a graduate student in physics at Johns Hopkins University, discovered, when working with gold foil, an effect that later would bear his name: the Hall effect. Hall found that when a transverse magnetic field is applied to a current-carrying conductor, an electric field results that is perpendicular to the current as well as to the magnetic field. The Hall voltage arises because, as was later shown, a charge carrier traveling in a magnetic field is subject to the so-called Lorentz force.

Since the Hall effect in semiconductors is rather high, semiconductors such as silicon, GaAs, and InSb have been studied extensively. The Hall effect can be employed for the construction of all kinds of practical devices, such as keyboards, magnetometers, magnetic recording heads, brushless electromotors, and displacement sensors.

In addition to the Hall effect, the so-called magnetoresistive effect can also be observed in metals, semiconductors, and magnetic materials. The resistivity of these materials changes when a magnetic field is applied.

As we wish to combine the microsensors for magnetic signals and electronic circuits on one chip, silicon, in particular, is of great interest as a semiconductor. Two kinds of silicon devices have been amply studied:

1. Hall plates, which are sensitive to a magnetic field perpendicular to the silicon surface; and
2. magnetotransistors, which are sensitive to a field parallel to the silicon surface.

Since the magnetoresistive effect in silicon is very small, practical silicon devices using this effect cannot be produced. In contrast, the magnetoresistive effect in magnetic layers is very large. Magnetic recording heads, consisting of silicon substrates on which magnetoresistive structures are fabricated by means of modern photolithographic techniques, are very useful for high-density recording equipment and are briefly discussed in the last section.

5.1.1 Hall Effect in Metals and Semiconductors The Hall effect in metals is very small, and only by using a very thin sample was Hall able to detect the Hall voltage. The Hall effect in semiconductors is much larger, because as we shall see later on, the Hall constant is inversely proportional to the number of charge carriers.

A Hall plate is typically a small, thin square of a metal or a semiconductor with length l, width w, and thickness t. When a current I passing through the plate is subjected to a magnetic flux density B_z perpendicular to the plane of the plate, a Hall voltage V_H occurs at the contacts, as shown in Fig. 13.

The Hall voltage V_H is proportional to the current and the magnetic field and can be expressed by (Popovic, 1991)

$$V_H = R_H I_x B_z / t, \tag{8}$$

in which R_H is the so-called Hall constant. The Hall voltage appears to be inversely proportional to the thickness of the plate t. For n-type silicon, the Hall constant is

$$R_H(\text{electrons}) = -1/nq. \tag{9}$$

The Hall constant is inversely proportional to the carrier density. This explains why the Hall effect is so much more pronounced in semiconductors than in metals. Holes also take part in the electrical conduction process in semiconductors. For p-type silicon, we find for the Hall constant

$$R_H(\text{holes}) = 1/pq \tag{10}$$

When the experimentally determined and theoretically calculated values of the Hall constant are compared, it appears that the theoretical values are a factor r smaller than the experimental values. This factor varies

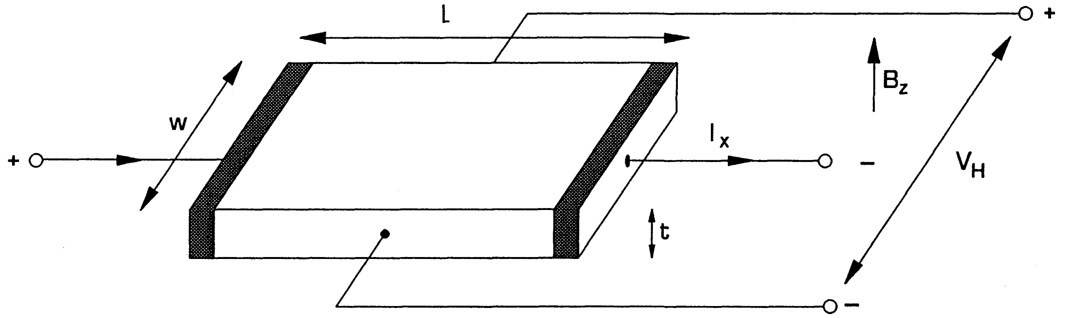

FIG. 13. Schematic representation of a Hall element.

between 1 and 2 and depends on the kind of semiconductor and the temperature. For the Hall constant, we obtain

$$R_H(\text{electrons}) = -r/nq \tag{11}$$

and

$$R_H(\text{holes}) = +r/nq. \tag{12}$$

The correction factor is necessary because in a semiconductor the charge carriers show a certain velocity distribution, and so the Lorentz force is not equal for all charge carriers.

5.1.2 Magnetoresistance Effects

5.1.2.1 Physical Magnetoresistance Effect. Not all charge carriers in a semiconductor have the same velocity. In the stationary state, this will cause some charge carriers to deflect upward and others to deflect downward. Consequently, the path of these charge carriers in crossing the Hall plate will be somewhat longer, which will slightly increase the resistance. This effect, the physical magnetoresistance effect, can be expressed as follows

$$\sigma(B) = \sigma(0)(1 - r^2\mu^2B_z^2). \tag{13}$$

The mobility μ in silicon is relatively low, and because the square of the mobility appears in the above equation, the physical magnetoresistive effect in silicon can be disregarded.

5.1.2.2 Geometrical Magnetoresistance Effect. A standard Hall plate consists of a rectangular semiconductor platelet with current contacts usually extending along the edges of the platelet in the y direction and pointlike Hall contacts halfway between the current contacts. When the length is much smaller than the width, the current conductors form a short-current path for the Hall voltage. In this case, the Hall voltage is reduced and does not fully balance the Lorentz force. As a result, the average path of the charge carriers deviates from the x direction, making the path longer and the resistance larger. This increase in resistance caused by the geometry of the Hall platelet is generally known as the geometrical resistance effect. For the resistance change, we get

$$R = R(0)(1 + \mu^2B_z^2). \tag{14}$$

In the geometrical magnetoresistive effect, the mobility again plays a very important role. Therefore, this effect is also very small in silicon and cannot be used for practical devices. The geometrical resistance effect is very large in certain high-mobility materials such as InSb. InSb might be of interest once it becomes possible to deposit this material on a silicon substrate. This would enable us to construct smart sensors consisting of magnetically sensitive InSb layers and electronic circuits on the same silicon substrate.

5.1.2.3 Magnetoresistivity in Magnetic Layers. In magnetic materials, the magnetoresistive effect is based on the interaction between the charge carriers and the magnetization of the material. This interaction is much stronger than the Lorentz force in silicon. When a small magnetic field is applied, the magnetization in suitable magnetic layers will rotate. This rotation of the magnetization affects the path of the charge carriers through the magnetic layer. Magnetic layers are characterized by a so-called uniaxial anisotropy. In the absence of a magnetic field, this anisotropy causes the magnetization to lie along the so-called easy axis. When a magnetic field is applied, the magnetization rotates in the plane of the film. When a current flows through the layer, an external magnetic field can thus rather easily change the angle between the current and the magnetization of the film, causing a change in the film's resistance.

The resistance of the magnetic film can be expressed by

$$R(\Theta) = R(0) - [R(0) - R(90)] \sin^2\Theta, \tag{15}$$

in which $R(0)$ is the resistance when the current and the magnetization are parallel, $R(90)$ is the resistance when the magnetization is perpendicular to the current direction, and Θ is the angle between current and magnetization. In most cases, NiFe or NiCo films are used for the magnetic layers, as these films can be made with a small magnetostriction and a small crystal anisotropy. For fields of the order of several mT, relative resistance changes of a few percent can be achieved by using these materials.

5.2 Silicon Hall Plates

Smart magnetic microsensors can be constructed by integrating Hall plates and readout electronic circuits on one silicon substrate. Standard silicon bipolar technology can be used to fabricate the Hall devices. Though the characteristics of Hall plates can be improved by carefully optimizing all the processing steps and impurity concentrations, this is rarely done. To obtain low-cost devices, it is much better to use the standard processing steps used for the production of integrated circuits, omitting only those steps that are not needed for Hall plates. This is illustrated in Fig. 14, showing the cross section of a standard bipolar *npn* transistor with a buried layer and a Hall plate. The Hall plate is formed by an *n*-type epitaxial layer (e.g., 10 μm, 2 Ω cm), which constitutes the collector area in an *npn* transistor. This layer is deposited on top of a *p*-type substrate. The high-concentration *n*-type diffusion that is used for the emitter and collector contacts is used to make the current and Hall contacts in the Hall plate. The *n*-type buried layer normally used to reduce collector resistance in an *npn* transistor is not needed for Hall plates. The same applies to the *p*-type base diffusion. The collector layer is quite suitable for Hall plates because it offers a low carrier concentration and because, being *n*-type, it offers the mobility of electrons, which is higher than that of holes. The reverse-biased *pn* junction isolates the Hall plate from the neighboring devices and the substrate.

5.3 Magnetotransistors

The path of the charge carriers in a transistor can be influenced by a magnetic field. This fact is used in the construction of mag-

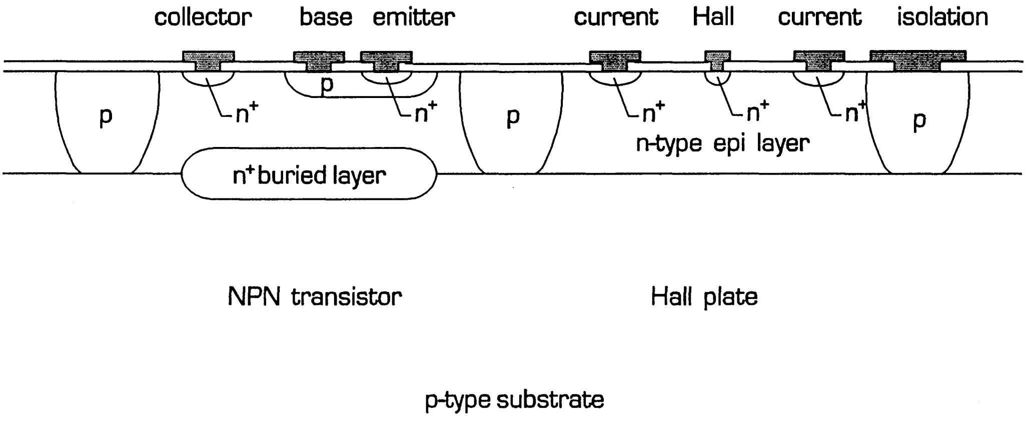

FIG. 14. Cross section of a bipolar *npn* transistor and a Hall plate.

netotransistors (Zieren and Middelhoek, 1982). Most devices consist of a single emitter and more than one collector contact to pick up the current. Figure 15 depicts a so-called vertical dual collector magnetotransistor. It consists of two *pnp* transistors with a common emitter and base. The collector area reaches beneath the emitter base structure by means of buried layers. There has to be a gap between the buried layers to avoid short-circuiting the collector contacts. Because of the rather thin base and the large surface of the base-emitter junction, the current is injected mainly in a direction normal to the substrate. A magnetic field parallel to the elongated transistor causes the electrons to deflect toward one of the collectors. The collector-current difference is proportional to the magnitude of the magnetic field. When the field is not parallel to the transistor, the current difference is proportional to the component of the field that is parallel to the transistor. For reasons of symmetry, an in-plane magnetic field perpendicular to the transistor does not cause a collector-current difference. Magnetotransistor-based devices have been developed that can simultaneously measure the magnitude and the direction in space of a magnetic field.

5.4 Magnetic Recording Heads

Over the last decade, the bit density of disk memories has increased remarkably thanks in large measure to improved read/write heads. The areal density has, at present, reached about 10^6 bits/cm^2. This increase was made possible when the inductive sensing of magnetic transitions gave way to direct field measurements by means of reading heads based on magnetoresistivity in Ni-Fe films. When a magnetoresistive head is brought into proximity to a magnetic disk or tape, the external magnetic field of the magnetic surface rotates the magnetization in the magnetic layer in the head. The resulting change in the resistance is measured and becomes an image of the stored signal on the tape or disk. The S/N (signal-to-noise) ratio of magnetoresistive heads has been shown to be superior to the S/N ratio of silicon Hall plates, and they also allow the required track and bit densities. Another advantage magnetoresistive heads have over inductive heads is that they are sensitive to a magnetic field and not to its time derivative, as would be the case in conventional inductive recording. Therefore, magnetoresistive heads are also suitable for the readout of magnetic tracks on credit cards and the like. Magnetoresistive heads consist of a combination of magnetic Ni-Fe layers and thin-film coils and are made by delicate deposition processes borrowed from semiconductor technology.

6. MICROSENSORS FOR CHEMICAL SIGNALS

Since time immemorial, mankind has taken a vivid interest in the composition of the materials used to make utensils or pre-

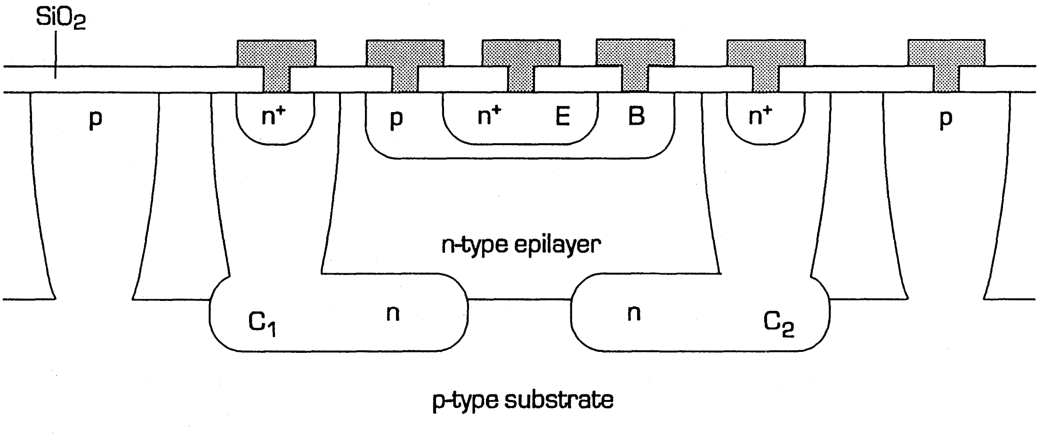

FIG. 15. Cross section of a dual-collector magnetotransistor.

pare food. Over the years, this need developed into the discipline of analytical chemistry. Today, thanks to a multitude of modern, often very expensive, analytical equipment and devices, very minute traces of practically any atom, molecule, or ion can be detected. However, analytical equipment is often bulky and can only be operated by trained personnel in suitable laboratory environments. Moreover, most measurements are time consuming. Given our increasing interest in our environment, our need for the on-line control of gas mixtures in industrial systems, our desire to monitor, for instance, oxygen in the blood during surgery, our need for reliable fire alarms, and our desire to optimize the performance of car engines, many scientists have been motivated to investigate the extent to which it is possible to make microsensors for chemical parameters. In the last few decades, many sensors for the detection of atoms, molecules, and ions have been presented in the literature. These sensors are based on materials such as metal oxides, organometallic compounds, ceramics, conducting polymers, and catalysts and are based on devices such as glass fibers, membranes, thermocouples, and ion-selective electrodes. In more recent years, silicon has also been proposed as a material for chemical microsensors.

6.1 Chemoresistors and Chemocapacitors

6.1.1 Chemoresistors Many chemical sensors depend on the measurement of the change in the conductivity or the dielectric constant of a chemical layer when subjected to a gas or an electrolyte. A very well-known metal oxide such as tin dioxide, which is an *n*-type semiconductor, changes its conductivity when heated and subjected to natural or town gas (methane). Standard metal-oxide gas sensors consist of a sintered block of the oxide that is heated to a few hundred degrees centigrade. Surrounding gases react with the oxygen on the hot surface, causing changes in the resistivity of the material. Properly doped gas sensors based on this principle and employing oxides of tin, zinc, iron, zirconium, or titanium have been shown to be sensitive to many gases. Metal-oxide gas sensors can be fabricated not only in thin- and thick-film technology but also in silicon technology (Demarne and Grisel, 1988). As is schematically shown in Fig. 16, this type of sensor consists of a silicon substrate on which a SiO_2 layer is grown thermally. A thin gold layer is deposited on top of this layer. This gold layer is patterned by the usual photolithographic methods in such a way that a meander-shaped resistance heater is obtained. Another SiO_2 layer is sputtered on top of the gold layer. This is followed by the sputtering and patterning of a tin-oxide layer to obtain a CO sensor. Last, to reduce the heat capacity of the structure, all silicon is removed from the back of the sensor by chemical etching. The resulting device is known as a chemoresistor.

6.1.2 Chemocapacitors The so-called chemocapacitor is based on the change that occurs in the dielectric constants of many organic materials when they are exposed to a gas. Interdigitated structures can be used to form a planar capacitor on top of a silicon substrate. By measuring the capacity, we can see whether the structure has been exposed to a gas. An accurate integrated dew-point humidity sensor was designed by using this type of chemocapacitor (Regtien, 1981). In this case, silicon technology makes batch production possible. Moreover, temperature-sensing elements and heating elements can easily be integrated on the same substrate. The real challenge is to find materials that are sensitive to only one chemical parameter and whose sensitivity does not change with time.

6.2 MOSFET-Based Chemical Microsensors

Microelectronics is dominated by two technologies: the bipolar technology and the metal-oxide-semiconductor field-effect transistor (MOSFET) technology. The MOSFET is based on the creation of conductive chan-

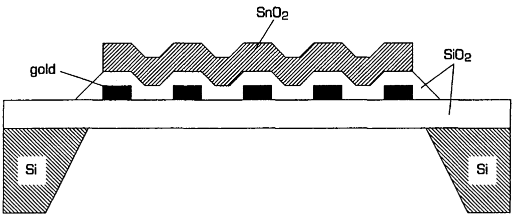

FIG. 16. Schematic representation of a silicon-based tin-oxide CO gas sensor.

nels along the surface of a silicon substrate and the modulation of the conductivity by properly applied electrical fields. MOSFET structures can also be used for the construction of chemical microsensors. A few decades ago, Bergveld (1970) proposed his so-called ion-sensitive field-effect transistor or ISFET. This is a MOSFET without a gate contact. When the ISFET is immersed in a solution containing ions, the drain current depends on the ion concentration. A few years later, Lundstrom *et al.* (1975) proposed a MOSFET with palladium as the gate metal for the detection of hydrogen. Since then, these devices have been studied extensively all over the world, with the result that many new microsensors have been found for a multitude of chemical parameters.

6.2.1 Ion-Sensitive Field-Effect Transistors (ISFETs) An ISFET, as is shown in Fig. 17, consists of a MOSFET in which the oxide layer is replaced by a combination of an oxide layer and a solution containing ions. The gate of an ISFET is positioned at a certain distance from the silicon substrate. The *p*H of an ionic solution can be determined by an ISFET. As standard *p*H-measurement cells are rather large and fragile, the invention of the ISFET in 1970 was greeted with avid interest, as it promised to allow the construction of elegant, miniature, and mass-produced *p*H-measurement cells. The first ISFETs used a thermally grown silicon oxide as the "ion-selective membrane." For calibration, the ISFET is brought into an electrolyte in which the *p*H can be varied. A frequently used electrolyte is a solution of NaCl in water. The *p*H is increased by adding NaOH or decreased by adding HCl. The threshold-voltage shift is measured as a function of the *p*H for a constant drain current and voltage.

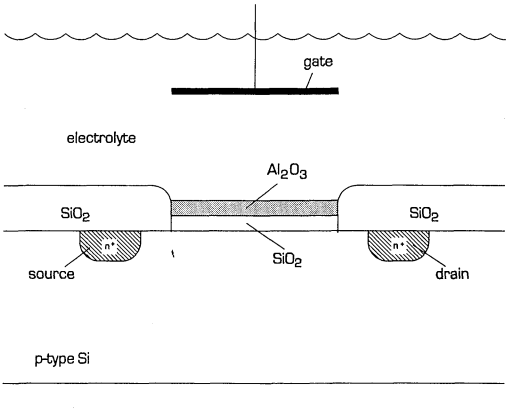

FIG. 17. Schematic cross section of an ISFET.

In the first decade following Bergveld's invention of the ISFET, the working mechanism of ISFETs was the focus of much discussion. It was initially assumed that the operation of ISFETs would be similar to that of a standard glass membrane cell and that hydrogen ions from the solution would diffuse through the thin SiO_2 layer to the oxide-silicon interface. In the case of the ISFET, they would change the interface charge and thus the threshold voltage of the transistor. Later on, it was shown that hydrogen diffusion is much too slow (seconds) to explain the experimentally observed quick response (milliseconds) of ISFETs. A new model, known as the *site binding model* in electrochemistry, was then applied to the ISFET. Today, almost all the features of ISFETs and related devices can be explained on the basis of this model. In the site-binding model, it is assumed that the *p*H response of an ISFET arises from surface reactions between the electrolyte and the oxide and not from hydrogen diffusion through the oxide. A gate oxide of SiO_2 can easily be obtained by thermal oxidation. However, other insulators such as Al_2O_3, Ta_3O_5, TiO_2, and Si_3N_5 have also been studied. It appears that Al_2O_3 is a much better material than SiO_2.

In most present-day experiments with ISFETs, SiO_2 has been replaced by a sandwich of SiO_2-Al_2O_3 layers. A disadvantage is that the technology of Al_2O_3 is much more complex, as these layers have to be made by chemical vapor deposition. ISFETs are mainly *p*H-sensitive devices. However, when the insulator is covered by a layer that allows some ion exchange to take place, the device can also be made sensitive to other ions. Such devices are known as CHEMFETs. When the gate area is covered with an enzyme, the resulting devices are known as *biosensors*.

6.2.2 Biosensors Many biological reactions must take place in the human body to keep the organism functioning. These reactions occur at a moderate temperature of 37 °C. They would be much too slow except that, fortunately, certain materials, known as *catalysts*, can increase the reaction speed. Biochemical catalysts, called *enzymes*, have complex molecular structures. Many thousands of enzymes have been identified to date. Enzymes also play an important role in the production of food, medicine, washing powder, etc. As living organisms depend on enzymes to survive, great interest has been shown in the possibility of controlling the enzyme concentration, especially when illness causes a deficiency. For example, insulin is an enzyme that plays an important role in diabetes. Diabetic patients have an excess of glucose, and this can be corrected by administering insulin. Of course, they would greatly benefit from a system that is capable of monitoring the blood glucose level and of automatically injecting the proper amounts of insulin and that is, moreover, small enough so that it can easily be carried around by the patient. Silicon-based chemical microsensors can be made very small. In addition, much of the electronic control circuitry can be placed on the sensor chip. As a result, these microsensors are understandably exciting great interest in the biomedical field.

A typical example of a biosensor is one that can measure the urea concentration in human blood, which gives an indication of the functioning of the kidneys. The urease enzyme catalyzes the reaction

$$(NH_2)_2CO + 2H_2O + H^+ \rightarrow 2NH_4^+ + HCO_3^-.$$

Since the *p*H of the solution is affected by the reaction, an ISFET can be used to measure the *p*H. Figure 18 depicts a schematic cross section of an ENFET (ENzyme FET). An ENFET is fabricated by depositing a SiO_2 and a Si_3N_4 layer on top of the *n*-channel structure of an ISFET by means of thermal oxidation and chemical vapor deposition (CVD). Next, in an elaborate process, a gel containing the urease enzyme is applied to the top of the structure. Then the ENFET is immersed in a solution that allows the urea content to be changed. The output signal of the ENFET is compared to that of a reference ISFET that lacks the enzyme cover. The result is a measure of the urea concentration (Miyahara *et al.*, 1983).

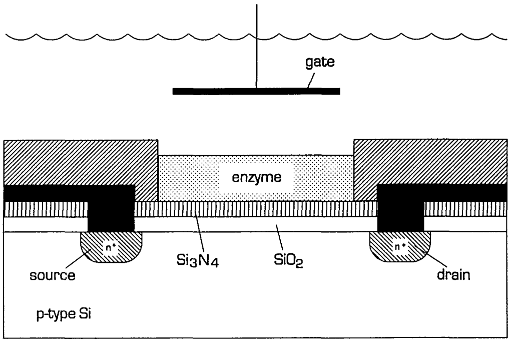

FIG. 18. Cross section of a biosensor.

6.2.3 Pd-Gate MOSFET Gas Microsensors

When the aluminum gate of a MOSFET is replaced by a palladium gate, the device's *I-V* characteristics are hydrogen dependent. Lundstrom and co-workers (Lundstrom *et al.*, 1975) were the first to report on these types of devices. The metal palladium has unique properties with respect to hydrogen. Molecular hydrogen dissociates to atomic hydrogen on the surface of the palladium gate. The atomic hydrogen diffuses through the palladium layer and is adsorbed at the palladium–silicon dioxide interface in the form of dipoles. The adsorbed hydrogen atoms thus give rise to a change in the working mechanism of the palladium. This in turn changes the threshold voltage of the MOSFET. The change can be expected to depend on the hydrogen coverage of the interface, which in turn depends on the hydrogen pressure. The shift in the threshold voltage is equal to

$$\Delta V_t = \Delta V_{\max}/[1 + 1/C(P_{H_2})^{1/2}], \qquad (16)$$

where $\Delta V_{\max}$ is the maximum observable voltage shift (about 0.5 V), C is a constant, and P_{H_2} is the partial pressure of hydrogen. In the palladium gate layer, at a certain hydrogen pressure an equilibrium occurs between the hydrogen molecules being dissociated into atoms and the process in which the atoms are recombined into hydrogen molecules. The actual dipole coverage of the interface layer is the result of the balance between these two processes. The most sensitive measurements are obtained when hydrogen is detected in an inert-gas atmosphere. Hydrogen concentrations down to 0.03 ppm can be detected.

Pd-gate devices are normally operated at temperatures from 60–150 °C, as the response time then assumes acceptable values of the order of some seconds. In a practical device, the Pd-gate MOSFET, a heating resistor, and a temperature sensor can be integrated on one silicon chip. There is still a lot of work that has to be done on the reproducibility, stability, poisoning, etc. of Pd-gate devices before a breakthrough in Pd-gate and related devices can be expected. Nevertheless, several possible applications have been studied, such as a simple hydrogen leak detector and a smoke detector.

7. PROCESSING OF MICROSENSORS

7.1 Silicon Planar Technology

Sensors can be fabricated in many technologies, including glass-fiber, metal-oxide, polymer, thin-film, thick-film, piezoelectric-material, and III-V and II-VI semiconductor technologies. However, only silicon-based sensor technologies have uniquely been able to progress by directly reaping the benefits of the expertise and skill gained in the development of the silicon integrated circuit and its fabrication technology. Silicon-based sensor technologies are of particular significance because they make it possible to produce smart sensors, in which the sensor and the signal-processing circuits are integrated on the same chip.

7.2 Micromachining Technologies

Micromachining pertains to the use of specialized fabrication techniques for the controlled, selective etching of silicon and films used in silicon-planar processing in order to obtain beams, diaphragms, valves, miniature electromotors, and many other structures. Early etching methods employed wet-chemical isotropic etchants. Anisotropic-etching technologies were later introduced to overcome the problems associated with precision, sensitivity, and temperature dependence.

There are two approaches to chemical etching: bulk micromachining and surface micromachining. In bulk micromachining, the silicon wafers are etched on both sides to achieve the desired structures. However, the most important advantage of silicon planar technology for the production of integrated circuits is that the circuits can be fully fabricated by processing only one side of a wafer. It therefore made sense to come up with methods in which the sensor and actuator structures could be obtained by processing only one surface of the wafer. The general term presently used for these methods is surface micromachining. In surface micromachining, selected materials are sequentially deposited and removed by selective etching procedures. Silicon etching has developed into an art with many followers. The achievements are very impressive and occasionally even artistic. Miniature electromotors having a diameter of less than 120 μm have attracted a great deal of attention and have shown how sophisticated structures can be made with present-day techniques (Muller, 1990).

7.2.1 Etching Solutions and Gases In wet-chemical isotropic etching solutions, usually consisting of a mixture of hydrofluoric, nitric, and acetic acids, the etch rate does not show a preference for any crystallographic orientation. Problems with isotropic etchants arise in the areas of etch control, selectability, and precision. The etch rate is agitation sensitive and temperature dependent and is limited by the transport of the etchant to the reacting surface. Narrow windows take longer to etch than wide windows because their access to fresh etchant is restricted. Anisotropic wet-chemical etchants differ from isotropic etchants in that they are orientation dependent. They are known to etch selectively the [100] and the [110] crystal orientations, leaving the [111] orientation relatively free from attack. Representative etch rates are typically 50 times slower in the [111] direction than in either the [100] or [110] direction. The use of anisotropic etchants with [110] silicon substrates results in openings with vertical sidewalls, while sidewalls set at an angle of 54.7° with respect to the surface are produced in [100] silicon. Lateral geometries and etch profiles can be controlled more precisely with anisotropic wet-chemical etchants than with isotropic ones.

The most commonly used wet-chemical anisotropic etchants include potassium hydroxide (KOH) and ethylenediamine–pyrocatechol–water (EDP). These etch solutions must be warmed to 85–115 °C to obtain reasonable etch rates, and the temperature must be kept very stable, since these etchants, like all other chemical etchants, are sensitive to temperature.

The dry etching in radio-frequency–generated plasmas of silicon, polysilicon, silicon dioxide, silicon nitride, resist, aluminum, and other films used in silicon planar technologies has become a well-accepted alternative to conventional wet-chemical techniques. Dry-etching techniques allow greater control over the etching procedure. Dry-plasma etching can be isotropic or anisotropic (i.e., respectively orientation-independent or -dependent) in nature. The directionality of the etching, the absolute etch rates, the etch-rate ratios, and the degree of radiation damage are all determined by procedural and instrumental parameters such as the composition, temperature, and flow rates of the reactant gases, the instrumental operating frequency, the pressure and the power density of the plasma, the voltage between the substrates and the plasma, and the wafer temperature.

Dry etching occurs by different mechanisms, which can mainly be broken down into ion-etching techniques and reactive-etching techniques. In ion-etching techniques, etching occurs primarily by physical means, such as ion bombardment. In reactive-etching techniques, a radio-frequency–generated plasma produces neutral atoms, neutral molecules, and radicals that react with the films to produce volatile compounds. Common gases used for silicon etching include Freons such as CF_4 and chlorine gases such as CCl_4 and Cl_2. Atomic fluorine and chlorine are produced in their respective plasmas, which react to produce the volatile compounds SiF_4 and $SiCl_4$, respectively. Silicon dioxide and silicon nitride are also etched in these plasmas.

7.2.2 Etching Procedures and Techniques

To fabricate precisely dimensioned micromechanical structures such as cantilever beams, diaphragms, and bridges, we have to be able to control the etching procedure so that it can be stopped completely at a predesignated point, thus providing dimension control on the order of 1 μm. The most common method of termination is predetermination of the etch rate followed by the timing of the etch procedure, but because of such factors as the sensitivity of the etch rate to agitation and temperature, as well as variations in the processing parameter and substrate thickness, this method of fabricating microstructures is unsatisfactory. Satisfactory etch-rate reduction methods that are currently employed include the boron etch-stop procedure and electrochemical etch (ECE) stops. In a number of applications, the boron etch-stop technique can be used when the anisotropic etchants EDP or KOH are used. When these anisotropic etchants are applied to the *n* regions of structures with p^+n junctions, the etch rate is significantly reduced at the interface between the *n* region and the heavily doped p^+ region (where the boron impurity is greater than approximately 5×10^{19} cm^{-3}). This procedure cannot be used in all applications. Considerable mechanical strain is introduced when the impurity level is so high, making it difficult to grow a high-quality epilayer on top of this p^+ layer. In addition, the high boron-doping level also prohibits the direct fabrication of microelectronics within this layer. Although anisotropic ECE etch-stop procedures are more difficult to use, they offer significant advantages in this area. The etching will terminate at the epitaxy/substrate interface of an *n*-type epitaxial layer grown on an n^+ or a *p* substrate with the standard doping level used in microelectronics fabrication.

The microelectronic or sensor devices can be fabricated in the epilayer by standard processing techniques followed by the use of the ECE etching procedure to define the microstructures. Figure 19 shows an experimental setup used for the electrochemical etching of silicon (Sarro and van Herwaarden, 1986). The hot-water control and the magnetic stirrer maintain the etchant at a selected temperature. An anode contact is made on the *n*-type epilayer, while a platinum cathode is installed within the solution. The etching of the *p* substrate begins as soon as contact is established within the etch solution between the anode and the *n*-type epilayer. The etching ends when the *n* epilayer is reached. The etching procedure is monitored by recording the cell current versus time. As the *p* substrate is being etched, the cell resistance will decrease. The cell current will simultaneously increase because the power supplied is kept constant. The termination of the etching procedure is then seen as a peak in the cell current.

The first well-established plasma-etching techniques were isotropic and nonuniform in nature, occurring primarily through the formation of volatile compounds and practiced in low-voltage barrel-type reactors. These in-

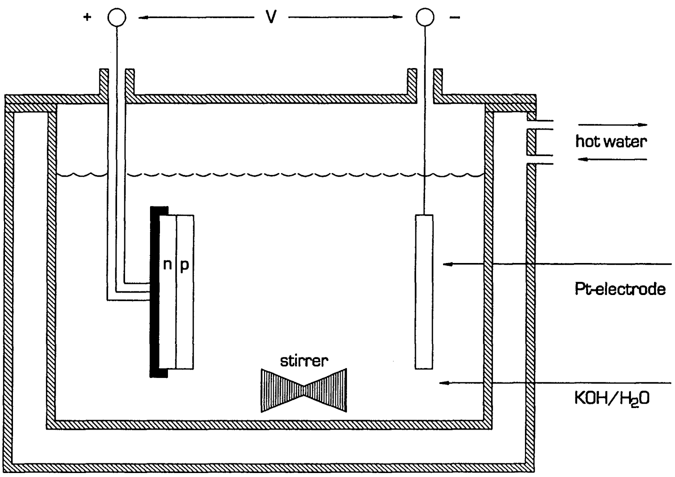

FIG. 19. Experimental setup for the electrochemical etching of silicon.

struments are still widely in use for applications not requiring a high resolution to etch the numerous films used in silicon planar technology (i.e., photoresist) with line widths on the order of a few microns. Etch ratios of 20:1 for Si:SiO_2 and 10:1 for Si_3N_4:SiO_2 are typical. Highly directional etching can be performed by reactive-ion etching in parallel-plate reactors. These systems are also primarily etched through the formation of volatile compounds, but ion bombardment also plays an important role. The reactors are similar to sputter-etching instruments. Vertical-etch rates that greatly exceed lateral-etch rates result with reactive-ion etching procedures, but the etch selectivity is reported to be poor because of the nonselective sputtering.

8. INTERFACES, SMART SENSORS, AND BUS SYSTEMS

In the past hundred years, myriads of sensors (as many as 50 000 according to a rough estimate) have been developed for the measurement of approximately 100 physical and chemical parameters. It might seem as if using silicon as the sensor material might merely add a few more sensors to the already existing mass of sensing devices. However, silicon offers something that other materials do not, namely, the possibility to integrate the sensor and part of the electronic processing circuit on one mass-produced silicon chip. These types of sensors are called *integrated smart sensors*. The on-chip integration may be the simple realization of a temperature-correction circuit, or in the future, a complete microprocessor might even be added to a sensing element. As we will elucidate below, smart sensors have an important role to play in future measurement and control systems.

Most present-day sensors are based on elements in which one of the parameters (resistivity, dielectric constant, resonance frequency, threshold voltage, etc.) shows a minute change in response to one or more measurands. This change is amplified and converted into a variety of electrical formats such as voltage, current, resistance, capacitance, frequency, and pulse duration. Moreover, the output signal will show a number of unwanted deficiencies such as offset, drift, nonlinearity, cross sensitivities, and poor interchangeability, to name a few. The multitude of output formats and the errors in the output signal make it very hard for system engineers to apply sensors unless they have an intimate knowledge of the workings of the sensing element itself.

The situation with respect to electronic components is even more favorable. Many standard components with well-defined input and output characteristics are available. For a trained system engineer, designing large systems based on these advanced components is not too difficult a task. As yet, this does not apply to sensors. Therefore, integrated smart sensors, which have a standard output and in which most of the deficiencies have been corrected by electronic circuits on the sensor chip, will have a significant impact on the development of measurement systems. In this last section, we delve into some of the issues concerning smart sensors.

8.1 Analog Signal Conversion

In most measurement and control systems, large numbers of sensors are connected to a central computer system. On the basis of the incoming information, the computer calculates relevant data about the process and, where necessary, sends data back along the bus system to actuators, such as valves or heaters, that adjust the process. One important feature in that process is the type of format the information should be in when it is sent along the bus data lines. Most physical measurands such as temperature, pressure, displacement, and so on, come to us in analog form. A simple measurement system may be based on locally amplifying the sensor signal to make it more resistant to interference and then sending it along the data lines in the form of an analog signal.

Various methods for analog signal conversion exist (Huijsing, 1986). In the simplest case, an analog signal from a sensor is electronically amplified into a voltage or a current representing the measurand. An analog output signal from a sensor can also be converted into a quasidigital signal, such as a frequency, a pulse count, a phase shift, or a pulse duration, since this type of signal is much less prone to noise and interference and can be recovered by filtering, phase-lock-loop techniques, or synchronous detection. Many sensors themselves generate a frequency output. In this case, while electronic circuitry might be needed for amplification or impedance matching, it will not be needed for the frequency-conversion step itself. One group of such sensors is based on resonators. Micromachining techniques can be applied to create suitable resonant structures, such as membranes or tuning forks, and the device can be activated electrically by means of a piezoelectric film or thermally.

Another group of sensors is based on integrating periodic geometrical structures in silicon. An example of this group is the CCD (charge-coupled device) (*q.v.*). CCDs can be used for one- and two-dimensional displacement measurements. With CCDs, displacement or position can be converted directly into number of pulses. The resolution of such devices depends almost solely on the size of the photosensitive elements of the CCDs.

Yet other sensors are based on electronic oscillators. One way to obtain a frequency output is to use ring oscillators in which the sensor element itself is the frequency-determining element. Pressure sensors based on two nine-stage MOS ring oscillators have already been achieved. A strain-sensitive ring oscillator can be obtained by making use of the fact that the drain current of a MOSFET is proportional to the mobility in the channel, which happens to be sensitive to strain. A temperature coefficient of the output signal as small as 0.008%/°C can be achieved when two ring oscillators are located in a silicon pressure diaphragm such that in one ring oscillator the stress is parallel and in the other the stress is perpendicular to the current in the channels, and the ratio of the oscillator frequencies is taken (Neumeister *et al.*, 1985).

8.2 Binary-Encoded Signals and Bus Systems

In the preceding section, we dealt with signal-conversion methods in which the resulting frequency, number of pulses, or pulse duration is a linear function of the measurand. Signal conversion can also be performed in such a way that the output contains an encoded form of the amplitude of the parameter being measured. The encoding of output signals is highly advantageous when the signal has to be transmitted or processed in some way. Binary encoding is the most frequently applied type of encoding: Since a bit can assume only two values (0 or 1), the output signal is virtually free of noise and interference.

In large measurement and control systems, communication between the central computer and the sensors and sensor subsystems scattered throughout a factory site is difficult. In all likelihood, the sensors will have been produced by several different companies. To make matters worse, the central computer will probably have been made by yet another manufacturer. To ensure that the data transfer is organized in an orderly and reliable fashion, bus systems, or networks, were designed. In the early days, bus systems were only implemented in spatially extended systems such as oil refineries, steel factories,

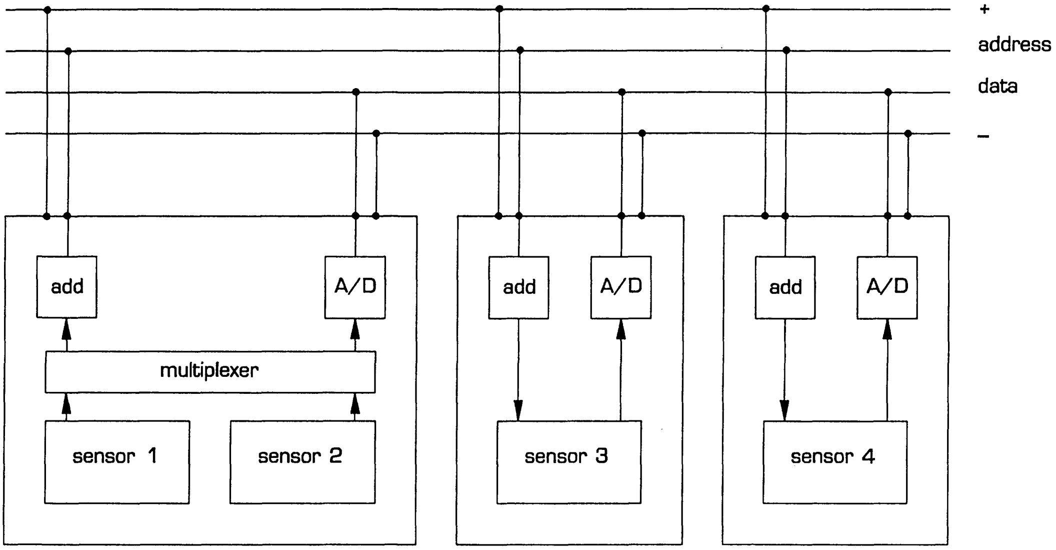

FIG. 20. Sensor bus system.

and large airplanes, but thanks to the introduction of low-cost microelectronic components, these days bus systems are being applied in increasingly smaller systems, even in consumer products. Moreover, the rise in the number of personal computers and workstations coupled to central computers has spurred the development of these data transfer systems and boosted the demand for cost-effective sensor systems.

A bus system generally consists of a central computer connected by several parallel wires to a large number of sensor systems. All sensor systems are switched in parallel, and electronic switches are required to switch the desired sensor to the data transfer lines. Many types of networks are feasible: ring structures, star networks, and a variety of trees. Figure 20 depicts a sensor subsystem with four bus lines: two for the voltage supply, one for sending addressing information, and one for the data line. When a sensor needs to be activated to send information, the address of the sensor is selected in the central computer, and that particular sensor is switched to the digital data line. The sensor subsystem may consist of a simple sensor, a signal conditioner, an analog-to-digital (A/D) converter, and an address identifier, although multiplexing techniques may also be applied so that the A/D converter and the addressing circuit are shared by several different sensors. Similarly, the data sent along the data lines may consist of simple data, such as the temperature or the pressure encoded in binary form, or other more complex information may be sent between the central computer and the sensors.

The central computer can also initiate all kinds of periodic tests and can order recalibration procedures. In addition, processing circuitry may be added to ensure that information about the fabrication date, the last calibration date, the range, or the average signal will periodically be sent to the central computer (Brignell, 1986). In the coming decade, silicon smart sensors will undoubtedly lead to many innovations in measurement and control systems. To meet the demands, it will be necessary to design new communication protocols and to process sensor signals at an even greater local level.

GLOSSARY

Anisotropic Etching: Etching method in which the use of certain etching solutions makes the etch rate depend on the crystal direction.

Band Gap: The energy states between the conduction and the valence band in a semiconductor or insulator that cannot be occupied by electrons.

Batch Fabrication: Manufacture of a number of products at the same time, so that they have very similar properties.

Biosensor: A sensor that can measure the biological reactions in a living organism.

Buried Layer: A layer that is obtained by diffusion or ion implantation prior to the growth of an epitaxial layer on the substrate.

Conduction Band: Band of allowed energy states in a semiconductor or insulator in which the lowest energy states are occupied by electrons.

Cross-Sensitivity: The sensitivity of a sensor to more than one parameter.

Dew Point: The temperature at which water vapor starts to condense.

Enzyme: A biochemical catalyst that increases the speed of biological reactions.

Exciton: A bound electron-hole pair with an energy in the band gap of a semiconductor.

***K* Shell:** Shell nearest the nucleus of an atom, containing two electrons.

Lorentz Force: The force exerted on a moving charge in a magnetic and an electric field.

Phase-Lock Loop: A system in which the frequency of a voltage-controlled oscillator can be made exactly equal (locked) to the frequency of an incoming periodic signal.

***pn* Junction:** The region between a *p*-type and an *n*-type semiconductor.

Valence Band: Band of allowed energy states in a semiconductor or insulator in which the highest energy states are not occupied by electrons.

Works Cited

Bergveld, P. (1970), *IEEE Trans. Biomed. Eng.* **19**, 70–71.

Brignell, J. (1986), in: S. Middelhoek, J. Van der Spiegel (Eds.), *Sensors and Actuators 10, State of the Art of Sensor Research and Development*, Lausanne: Elsevier Sequoia, pp. 249–261.

Collet, M. G. (1986), in: S. Middelhoek, J. Van der Spiegel (Eds.), *Sensors and Actuators 10, State of the Art of Sensor Research and Development*, Lausanne: Elsevier Sequoia, pp. 287–302.

Demarne, V., Grisel, A. (1988), *Sensors Actuators* **13**, 301–314.

Gieles, A. C. M. (1969), *Digest IEEE International Solid-State Circuits Conference*, New York: IEEE, pp. 108–109.

Huijsing, J. H. (1986), in: S. Middelhoek, J. Van der Spiegel (Eds.), *Sensors and Actuators 10, State of the Art of Sensor Research and Development*, Lausanne: Elsevier Sequoia, pp. 219–237.

Ko, W. H. (1986), in: S. Middelhoek, J. Van der Spiegel (Eds.), *Sensors and Actuators 10, State of the Art of Sensor Research and Development*, Lausanne: Elsevier Sequoia, pp. 303–320.

Lundstrom, I., Shivaraman, S., Svensson, C., Lundkvist, L. (1975), *Appl. Phys. Lett.* **26**, 55–57.

Meijer, G. C. M. (1986), in: S. Middelhoek, J. Van der Spiegel (Eds.), *Sensors and Actuators 10, State of the Art of Sensor Research and Development*, Lausanne: Elsevier Sequoia, pp. 103–125.

Middelhoek, S., Audet, S. A. (1989), *Silicon Sensors*, London: Academic Press.

Miyahara, Y., Matsu, F., Moriizumi, T., Matsuoka, H., Karube, I., Susuki, S. (1983), in: *Proceedings of the International Meeting on Chemical Sensors, Fukuoka, Japan*, Tokyo: Kodansha Ltd., pp. 501–506.

Muller, R. S. (1990), *Sensors Actuators* **A21–A23**, 1–8.

Neumeister, J., Schuster, G., von Münch, W. (1985), *Sensors Actuators* **7**, 167–176.

Popovic, R. S. (1991), *Hall Effect Devices*, Bristol: Adam Hilger.

Regtien, P. P. L. (1981), *Sensors Actuators* **2**, 85–89.

Roylance, L., Angell, J. B. (1979), *IEEE Trans. Electron Dev.* **ED-26**, 1911–1917.

Sarro, P. M., van Herwaarden, A. W. (1986), *J. Electrochem. Soc.* **133**, 1724–1729.

Smith, C. S. (1954), *Phys. Rev.* **94**, 42–49.

Ten Kate, W. R. Th. (1986), in: S. Middelhoek, J. Van der Spiegel (Eds.), *Sensors and Actuators, State of the Art of Sensor Research and Development*, Lausanne: Elsevier Sequoia, pp. 83–101.

van Duyn, D. C., Middelhoek, S. (1992), in: P. Ciureanu, S. Middelhoek (Eds.), *Thin Film Resistive Sensors*, Bristol: Adam Hilger, Chap. 1.

van Herwaarden, A. W., Sarro, P. M. (1986), in: S. Middelhoek, J. Van der Spiegel (Eds.), *Sensors and Actuators, State of the Art of Sensor Research and Development*, Lausanne: Elsevier Sequoia, pp. 321–346.

Wolffenbuttel, R.F. (1987), *IEEE J. Solid-State Circuits*, **22**, 350–356.

Zieren, V., Middelhoek, S. (1982), *Sensors and Actuators* **2**, 251–262.

Further Reading

Middelhoek, S., Audet, S. A. (1989), *Silicon Sensors*, Academic Press: London.

Göpel, W., Hesse, J., Zemel, J. N. (Eds.) (1991–1993), *Sensors—A Comprehensive Survey*, 8 vols., Weinheim: VCH.

Vol. 1, Grandke, T., Ko, W. H. (Eds.), *Fundamentals and General Aspects.*

Vol. 2, Göpel, W., Jones, T. A., Kleitz, M., Lund-

ström, J., Seiyama, T. (Eds.), *Chemical and Biochemical Sensors Part I.*

Vol. 3, Göpel, W., Jones, T. A., Kleitz, M., Lundström, J., Seiyama, T. (Eds.), *Chemical and Biochemical Sensors Part II.*

Vol. 4, Ricolfi, T., Scholz, J. (Eds.), *Thermal Sensors.*

Vol. 5, Boll, R., Overshott, K. J. (Eds.), *Magnetic Sensors.*

Vol. 6, Wagner, E., Dändliker, R., Spenner, K. (Eds.), *Optical Sensors.*

Vol. 7, De Rooij, N. F., Kloeck, B., Bau, H. H. (Eds.), *Mechanical Sensors.*

Vol. 8, *Cumulative Index and Selected Topics.*

Details on current developments can be found in the proceedings of the biennial international TRANSDUCERS conferences (1981 Boston, 1983 Delft, 1985 Philadelphia, 1987 Tokyo, 1989 Montreux, 1991 San Francisco, 1993 Yokohama, 1995 Stockholm) published in *Sensors and Actuators* or as a separate proceedings. In Europe, the proceedings of the annual EUROSENSOR conferences are published in *Sensors and Actuators.* In the U.S.A., the proceedings of the American biennial sensor conferences are published by the IEEE. In Japan, English-language proceedings of the annual conferences are published by the IEE of Japan. Many topical conferences and workshops focusing on such topics as biosensors, micromachining, chemical sensors, and the automotive applications of sensors are held every year.

MICROWAVE CIRCUITS

PETER STAECKER, *M/A-COM, Inc., Corporate Research and Development Center, Lowell, Massachusetts, U.S.A.*

Introduction 349
1. **Concepts and Definitions** 350
1.1 Boundary Conditions 352
1.2 Dissipative Media (Loss) 352
1.3 Guided Waves 353
1.4 Lumped and Distributed Circuits, Characteristic Impedance 354
1.5 Reflection Coefficient and the Smith Chart 354
2. **Propagation Media** 355
2.1 TEM Structures 356
2.2 Quasi-TEM Structures 357
2.3 Non-TEM Structures 358
2.4 Enabling Technology: High-Temperature Microwave Superconductivity 359
2.5 Future Developments 360
3. **Components** 361
3.1 Passive and Signal Control Functions 361
3.2 Signal Generation and Amplification 364
3.3 Frequency Conversion Components: Mixers and Frequency Multipliers 368
3.4 Future Developments 368
4. **Circuit Integration** 370
4.1 Microwave Integrated Circuits 370
4.2 Monolithic Microwave Integrated Circuits 371
4.3 Integrated Subassemblies: Selecting from the Menu of Technologies 374
4.4 Future Developments 376
5. **Summary** 376
Glossary 376
Works Cited 377
Further Reading 378

INTRODUCTION

Electronic circuit elements (resistors, capacitors, and inductors are elementary examples) and their terminal electrical characteristics measured in voltage and current are approximate realizations of the laws of electromagnetism; *microwave circuits* provide the framework in which these approximations are validated, modified, or replaced at high frequencies. Historically, the basis for microwaves dates to a series of papers written by Maxwell between 1855 and 1865 which established the theory of electromagnetic wave propagation, and Hertz, who in 1886 experimentally verified Maxwell's theory. Contemporary applications of microwaves developed rapidly during World War II with the formation of the Royal Radar Establishment in England, research activities in Germany and Japan, and the MIT Radiation Laboratory in the United States.

In general, electromagnetic field descriptions of circuit behavior become necessary when the wavelength (speed of light in the medium divided by the frequency) approaches the size of the circuit or component. For this discussion, the microwave region is bounded from above by the frequency of 100–500 GHz, where quasi-optical techniques apply, and given this definition, includes the millimeter-wave frequency region, which extends from 30 to 300 GHz. A lower frequency boundary is set by the somewhat fuzzy interface between analog and digital circuits, presently in the few-GHz range, but as the speed of digital logic increases, microwave circuit techniques will embrace this technology also. The following provides an overview of microwave electronics, from its basis in electromagnetic theory (Sec. 1), to propagation media used in microwave circuits (Sec. 2), to single-function components (Sec. 3), to circuit integration techniques (Sec. 4). Contemporary practices as well as projections of future activity are noted within these

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

sections as well as in the summary (Sec. 5). Applications of microwave circuits include radar, navigation, communications, medical applications and other interaction with materials (heating), remote sensing, and energy transmission. Radar (RAdio Detection And Ranging) is so commonplace it rarely is given uppercase acronym status. In 1940, on the eve of the formation of the MIT Radiation Laboratory when detection research was in its infancy, the focus was on detection of aircraft by UHF waves, but curious scientists were attracted to other moving targets. The success with which the observers detected violations of the highway speed limit of 50 miles per hour led one scientist to remark, "Don't let the cops hear about this." Collision-avoidance and object-detection systems for automobiles are current topics of intense interest in commercial radar applications.

Operating at 1.7 GHz, the Global Positioning Satellite (GPS) network broadcasts information allowing users with commercial receiving equipment to locate themselves anywhere on earth to a precision of 20–40 m. Palm-size GPS receivers are available now for about $1000 and will appear in automobiles as navigation aids in the mid-1990s. (1993 U.S. dollars are used throughout this article.)

For communications applications, bandwidth (Hz) is key to the rate of information transfer (bits/s). Communications systems are being extended to light-wave frequencies for Tbit/s network requirements of supercomputer interconnections, full-motion color graphics, and medical imaging, which may require high-resolution images to be transmitted uncompressed. Microwave signal transmission is used for terrestrial radio telephone communications. Satellite communication networks service television, telephone, and other commercial communication network needs. Cellular telephones operating at 900 MHz and 1.8 and 2.4 GHz presently serve local areas with terrestrial transmitters. Military communications users favor new frequencies in the millimeter-wave region, some as high as 60 GHz, to take advantage of an oxygen absorption band in the earth's atmosphere, which ensures privacy [Fig. 1(a)]. Commercial wireless applications (automotive communications) planned for the same frequency band will achieve greater signal densities (channels/km^2) through the same characteristic. Wireless markets have exploded in the past five years and will probably overtake military microwave applications near the turn of the century [Fig. 1(b)].

Medical uses of microwaves represent more contemporary applications of microwave technology. Microwave hyperthermia uses the heating effect of microwaves in human tissue to destroy malignant tumors. Radiometric techniques developed at the MIT Radiation Laboratory are today in use to detect temperature differentials characteristic of these same tumors. Microwaves are used to heat blood instantly in trauma cases where it is crucial to elevate body temperature to normal levels in a short time.

The microwave oven was a direct spinoff of MIT Radiation Laboratory activity, using a magnetron as the power source operating at 2450 MHz where absorption of electromagnetic energy by water is extremely efficient. The same frequency region has been proposed for use in beaming power collected by orbiting solar arrays to antenna farms on earth.

The markets supporting these applications are currently strongly military but by the year 2000 should be exceeded by the wireless data and telephony markets.

1. CONCEPTS AND DEFINITIONS

Central to the discussion of microwave circuits is the understanding of electromagnetic wave propagation. Indeed, the concepts of distributed circuits and dispersion find their basis in the wave equation, which in turn is derived from the time-dependent Maxwell's equations. Equally important notions of loss and lumped equivalent circuit elements result from the static approximations to these same equations. Maxwell's equations in differential form, with steady-state sinusoidal time dependence, are

$$\nabla \times \mathbf{E} = -j\omega\mathbf{B}, \tag{1}$$

$$\nabla \times \mathbf{H} = j\omega\mathbf{D} + \mathbf{J}, \tag{2}$$

$$\nabla \cdot \mathbf{D} = \rho, \tag{3}$$

$$\nabla \cdot \mathbf{B} = 0, \tag{4}$$

where **E** and **D** are the electric field and electric displacement intensities, **H** and **B** are the

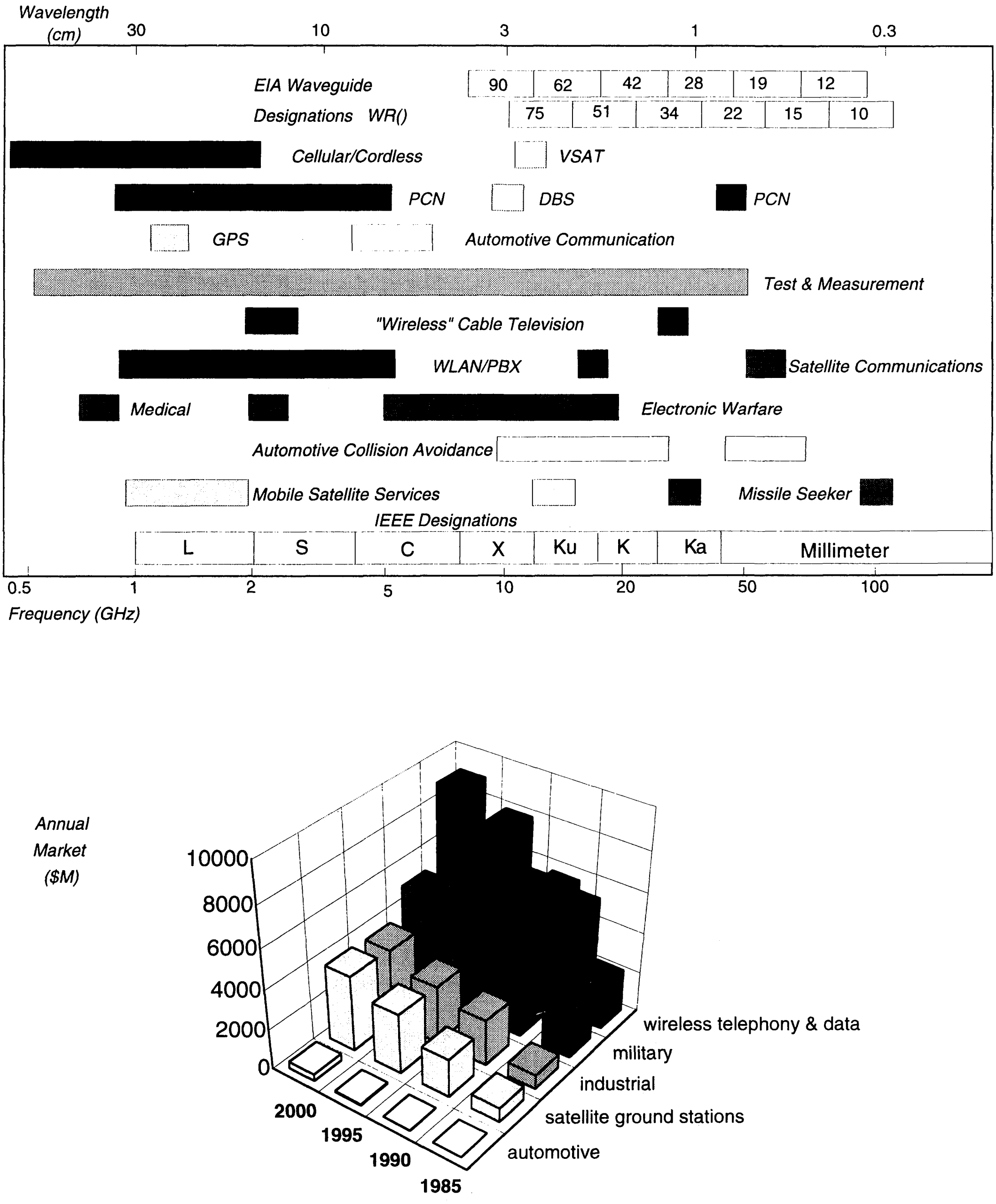

FIG. 1. (a) Microwave frequency spectrum, band designations, and applications. In this figure, the following acronyms are used: EIA: Electronics Industry Association. DBS: Direct Broadcast Satellite. GPS: Global Positioning Satellite, a system of satellites that broadcast time-coded messages to Earth, which, when received and decoded, allow latitude and longitude information to an accuracy of 30 m (commercial) and 2 m (military). IEEE: Institute of Electrical and Electronics Engineers. PBX: Private Branch Exchange, a private telecommunications exchange that includes access to a public telecommunications exchange. PCS: Personal Communications Services, a family of mobile or portable radio communications services providing services to individuals and businesses and through a variety of competing networks. PCN: Personal Communications Network, a form of PCS that offers metropolitan-area portable radio telephony competitive with cellular and public telephone networks. VSAT: Very small-aperture terminal. WLAN: Wireless Local Area Network. (b) Microwave world-wide market segments: history and projections. (Courtesy of Strategies Unlimited, Mountain View, CA.)

magnetic field and magnetic flux intensities, $\mathbf{J}$ is the conduction current, and ρ is the electric charge density. All fields are harmonic in time ($e^{j\omega t}$) and are vector quantities. The associated constitutive relations are

$$\mathbf{D} = \epsilon\mathbf{E} = \epsilon_0\epsilon_r\mathbf{E}, \tag{5}$$

$$\mathbf{B} = \mu\mathbf{H} = \mu_0\mu_r\mathbf{H}, \tag{6}$$

$$\mathbf{J} = \sigma\mathbf{E}, \tag{7}$$

where ϵ is the electrical permittivity, μ is the magnetic permeability, and σ is the electrical conductivity of the medium (all of which are, in general, tensor quantities), ϵ_0 is the permittivity of free space ($\epsilon_0 = 8.854 \times 10^{-12}$ F/m), and ϵ_r is the relative permittivity of the dielectric medium. Likewise, μ_0 is the permeability of free space ($\mu_0 = 4\pi \times 10^{-7}$ H/m), and μ_r is the relative permeability of the medium. The properties of materials are embedded in their permittivities and permeabilities; the imaginary part of ϵ introduces dielectric loss, an important consideration at high frequencies. The anisotropic nature of μ allows operation of magnetic circulators described below.

1.1 Boundary Conditions

From the integral formulation of Eq. (1), it can be shown that the tangential component of the electric field is continuous across any boundary in the medium. The same manipulation of Eq. (2) reveals that the tangential magnetic field discontinuities across any boundary in the medium require the presence of surface current on the boundary. In a similar manner, integral formulations of Eqs. (3) and (4) yield boundary conditions on field components normal to surfaces: The normal component of magnetic flux must be continuous across media interfaces, whereas discontinuities in the normal component of electric displacement are due to the presence of surface charge ρ_s at the interface. The special case of electrical conductor–vacuum interface (electrical wall) leads to the relations

$$E_t = 0, \qquad H_t = K, \tag{8}$$

$$D_n = p_s, \qquad B_n = 0, \tag{9}$$

at the metal wall (where K is surface current in A/m). Equations (8) and (9) are used extensively in the solution of wave propagation in transmission media with conducting boundaries. For solutions to problems in fiber optics, the appropriate boundary conditions are applied at dielectric interfaces, and the field solutions are more complicated, but the basic electromagnetic framework is the same.

1.2 Dissipative Media (Loss)

Solving the Helmholtz (or wave) equation [derived from Eqs. (1) and (2)]

$$\nabla^2\mathbf{E} - \gamma^2\mathbf{E} = 0 \tag{10}$$

for fields harmonic in time and space [$e^{(j\omega t-\gamma z)}$] leads to an expression for the complex propagation constant $\gamma = \alpha + j\beta$ in terms of the complex permittivity $\epsilon = \epsilon' - j\epsilon''$ and the conductivity as

$$\gamma^2 = j\omega\mu(\sigma + j\omega\epsilon) = j\omega\mu([\sigma + \omega\epsilon''] + j\omega\epsilon'). \tag{11}$$

The dielectric damping term $\omega\epsilon''$ can be combined with the conductivity σ to give the total effective electrical conductivity. The loss tangent (sometimes called dissipation factor) of a dielectric medium is defined as the ratio of the sum of the conduction current ($\sigma E + \omega\epsilon'' E$) to the displacement current $j\omega\epsilon' E$:

$$\tan\delta = \frac{\sigma + \omega\epsilon''}{\omega\epsilon'}. \tag{12}$$

Dielectric damping accounts for the absorption of microwaves in nonconductive materials but for transmission media is usually ignored over most of the microwave frequency region, i.e., $\epsilon \approx \epsilon'$. For situations in which the conduction current σE is much less than the displacement current $\omega\epsilon' E$, as might be the case in a slightly imperfect dielectric, propagation is described by

$$\gamma \approx \omega\sqrt{\mu\epsilon}\left(\frac{\tan\delta}{2} + j\right), \quad \tan\delta \ll 1; \tag{13}$$

whereas when conduction currents greatly exceed displacement currents, as would be the case in a slightly imperfect conductor, for example, the real and imaginary parts of the propagation constant are equal and are given by

$$\gamma = \sqrt{\frac{\omega\mu\sigma}{2}}(1 + j), \quad \tan\delta \gg 1. \tag{14}$$

Under these conditions, field amplitudes decay exponentially in the direction of propagation, and give meaning to the concept of *skin depth*, δ_s, the inverse of the attenuation constant α. Surface resistance ($R_s = 1/\sigma\delta_s$ ohms/square) is used to describe properties of conductive films.

The ratio of electric to magnetic field is defined as the *wave impedance*, which has dimensions of ohms [a subscript zero (η_0) is used when the propagation medium is vacuum].

$$\eta \equiv |\mathbf{E}|/|\mathbf{H}| = \sqrt{\mu/\epsilon} \tag{15}$$

and is the basis for and related to circuit-theoretical impedance concepts discussed below.

In resonant circuits, the ratio of average energy stored in the resonant circuit to the resonator energy loss per cycle is the (unloaded) quality factor Q_u. [Two additional quality measures which consider energy storage/loss in the external circuit (Q_E), or energy storage/loss in the resonator *and* external circuit (loaded Q: Q_L), are used frequently in resonator design.] This convenient descriptor of losses in both distributed and lumped circuit elements can be defined for transmission-line resonators with resonant wavelength λ_r:

$$Q_u = 2\pi \frac{\text{average stored energy}}{\text{energy loss/cycle}} = \frac{\beta}{2\alpha} = \frac{\pi}{\alpha\lambda_r}. \tag{16}$$

The quality factor is a function of frequency as well as geometry, having typical values (at 10 GHz) of 10–100 for lumped element reactances and ~1000 for distributed resonators, and can reach values greater than 10 000 for cavity resonators, and greater than 100 000 for cavities with superconducting walls. This figure of merit is very useful in the design of frequency-selective structures used in microwave filters and oscillators. When size is a consideration, low-loss dielectrics (whose dielectric constant can reduce λ_r by as much as a factor of 5) or acoustic resonators (with resonant wavelengths a factor of 10 000 less than air-filled cavities) can be used.

1.3 Guided Waves

The electromagnetic boundaries of the guiding structure define the nature of field solutions for wave propagation in the structure under consideration. The wave equation in a lossless medium may be written as

$$\frac{\partial^2 \mathbf{H}}{\partial x^2} + \frac{\partial^2 \mathbf{H}}{\partial y^2} + (\omega^2\mu\epsilon - \beta^2)\mathbf{H} = 0. \tag{17}$$

Transverse electromagnetic (TEM) fields by definition have no longitudinal component (z component, in the notation of this article). These fields exist in free space (no guiding medium) and in certain guiding structures described below. Solutions to Eq. (17) are impossible unless $\omega^2\mu\epsilon - \beta^2 = 0$ (TEM solutions therefore exist to zero frequency), which reduces the form of Eq. (17) to Laplace's equation in the transverse plane, familiar from electrostatics. As a direct consequence, the voltage between conductors defining the guiding medium can be found from the integral form of Eq. (3) as

$$\mathcal{V}_{12} = \int_1^2 \mathbf{E} \cdot d\mathbf{l}. \tag{18}$$

Likewise, from the integral form of Eq. (2), the current on the conductors can be found from

$$I = \oint_c \mathbf{H} \cdot d\mathbf{l}, \tag{19}$$

where c is the cross-sectional contour of the conductor surface. Because of the uniqueness of solutions to Laplace's equation, the voltage and current, and therefore the ratio $\mathcal{V}/I$, are unambiguous and useful for circuit representations in TEM structures.

For regions enclosed by a single conductor, solutions to Eq. (17) depend upon the exact geometry of the surrounding region. For a conducting boundary of infinite extent in the z direction, Eq. (17) can be written as three scalar equations. For the z component,

$$\frac{\partial^2 H_z}{\partial x^2} + \frac{\partial^2 H_z}{\partial y^2} + k_c^2 H_z = 0. \tag{20}$$

If the cross section of the conducting boundary is rectangular, solutions to Eq. (20) exist of the form

$$H_z = [A\cos(k_x x) + B\sin(k_x x)] \times [C\cos(k_y y) + D\sin(k_y y)], \qquad (21)$$

where k_x and k_y are both multivalued:

$$k_x^2 + k_y^2 = k_c^2 \equiv \omega^2\mu\epsilon - \beta^2 \equiv k^2 - \beta^2. \qquad (22)$$

For frequencies greater than a cutoff frequency f_c $(=k_c/2\pi\sqrt{\mu\epsilon})$, propagating modes ($\beta$ real) exist, whereas for frequencies below f_c, β is imaginary, and waves are evanescent (attenuated exponentially with distance, not propagating). The phase velocity, which is greater than the speed of light in the same medium, is given by

$$v_p \equiv \frac{\omega}{\beta} > \frac{\omega}{k} = \frac{1}{\sqrt{\mu\epsilon}}. \qquad (23)$$

This treatment defines transverse electric (TE) waves, i.e., where $H_z \neq 0$ and $E_z = 0$. The same boundary conditions allow transverse magnetic (TM) field solutions, where $E_z \neq 0$ and $H_z = 0$. In general, discrete modes of propagation [solutions to Eq. (20)] exist for each possible value of k_c allowed by the boundary conditions.

Some additional characteristics of TE (TM) waves may now be noted:

1. Because there is a longitudinal component of field, unambiguous definitions of voltage and current are not rigorously correct. These concepts are used only under restricted conditions as will be mentioned below.
2. Because $k_c \neq 0$, TE (TM) wave solutions do not exist to zero frequency but propagate for frequencies above a cutoff frequency f_c that is determined by the cross-section dimensions. In addition, the phase velocity of TE (TM) waves varies with frequency, causing individual frequency components of a wave form to change their original phase relationships with respect to each other during propagation through the guiding structure, resulting in wave form distortion. This property, called *dispersion*, is of extreme importance in pulse propagation over long distances.

1.4 Lumped and Distributed Circuits, Characteristic Impedance

Energy storage in the magnetic and electric fields of a transmission line is associated with a series inductance and shunt capacitance per unit length. Likewise, power loss in the conductors and dielectric can be modeled by a series resistance and shunt conductance per unit length. The equivalent circuit of a transmission line of differential length dz can thus be represented as an equivalent circuit whose element values are related to the material properties of the transmission medium [Fig. 2(a)]. Trading heuristic arguments for rigorous computer-based electromagnetic field solvers, equivalent circuit models can be generated for more complex structures such as spiral inductors [Fig. 2(b)].

Applying Kirchoff's voltage and current laws to the model of Fig. 2, and assuming voltage wave propagation of the form $V = V^+e^{j\omega t-\gamma z} + V^-e^{j\omega t+\gamma z}$ (and similar solutions for current waves), the lumped-circuit version of the propagation characteristics can be expressed as

$$\begin{aligned}\gamma &= \sqrt{(R + j\omega L)(G + j\omega C)}\\ &= \omega\sqrt{LC}\left\{\left[\frac{R}{L} + \frac{G}{C}\right] + j\right\},\end{aligned} \qquad (24)$$

$$Z_0 \equiv \frac{V^+}{I^+} = \frac{V^-}{I^-} = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \approx \sqrt{\frac{L}{C}}, \qquad (25)$$

where the approximations apply for small losses ($R \ll \omega L$ and $G \ll \omega C$). If the transmission line is terminated in its characteristic impedance Z_0, there will be no reflected wave.

Because the impedance concept embodies the phase and amplitude characteristics of electromagnetic wave propagation in a network description, it is an extremely important tool in all aspects of microwave circuit design. Specific applications include maximizing power-transfer or power-generating characteristics, or minimizing additive thermal noise of active devices.

1.5 Reflection Coefficient and the Smith Chart

The reflection coefficient Γ is defined as the ratio of the reflected wave to the incident wave:

$$\Gamma(z) \equiv \frac{V^-(z)}{V^+(z)} = \frac{V^-(0)}{V^+(0)}e^{-2\gamma z} \equiv \Gamma_R e^{-2\gamma z}, \qquad (26)$$

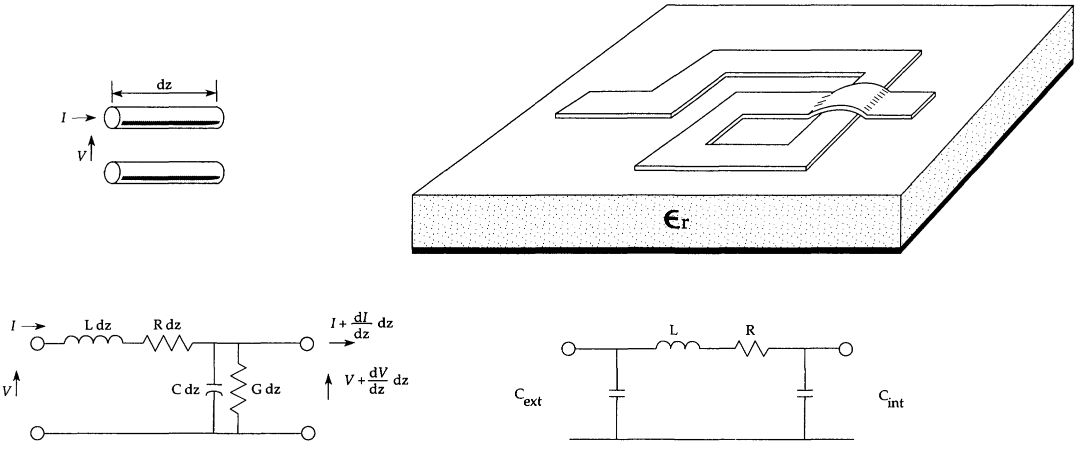

(a) (b)

FIG. 2. (a) Transmission-line equivalent circuit; and (b) printed spiral inductor and its equivalent circuit (for frequencies below resonance).

where $z = 0$ denotes the wave amplitude reference location. (These reference locations, where amplitude and phase are specified, are called reference planes and define the phase datum for each network port, since phase is a relative quantity. These locations can be coincident with the planes of electromagnetic access to the network, such as a waveguide flange or a coaxial connector, and can be analytically relocated within the network using the propagation characteristics of its transmission lines.) In this notation, a perfect open (short) circuit at $z = 0$ has a reflection coefficient of $+1$ (-1), and a transmission line terminated in its characteristic impedance has a reflection coefficient of zero.

Reflection coefficient is related to the normalized impedance Z_n by the bilateral transformation

$$\Gamma_R = (Z_n - 1)/(Z_n + 1) \tag{27}$$

(where the normalized impedance Z_n is equal to the reference impedance Z_R divided by the characteristic impedance Z_0), which maps the infinite half-plane of impedance space into the unit circle (Fig. 3). This unit circle representation is called the Smith chart (Wheeler, 1984), in recognition of Philip Smith, its inventor, and has been widely used by microwave circuit designers since its disclosure. Circuit synthesis with the Smith chart uses simple graphical constructs: From a given reference impedance, lengths of transmission line are rotations in electrical length about the center of the chart; inductive or capacitive reactance is added by moving on constant-resistance circles. Impedance to admittance conversions are executed by rotating halfway around the chart on a constant radius. Having suffered the same fate as its computational counterpart, the slide rule, the Smith chart today is rarely used as a design aid but still serves as the universal template for displaying measured and computed data as a function of impedance and frequency.

The ratio of maximum voltage magnitude to minimum voltage magnitude is called the voltage standing-wave ratio (VSWR) of the line, and is defined as

$$\mathrm{VSWR} \equiv \frac{1 + |\Gamma_R|}{1 - |\Gamma_R|} = \frac{|V(z)|_{\max}}{|V(z)|_{\min}}. \tag{28}$$

Loci of constant VSWR on the Smith chart are circles with centers at the origin of the unit circle, making infinite VSWR locus the edge of the chart, and perfect match (unity VSWR) the center of the chart.

2. PROPAGATION MEDIA

Electromagnetic wave propagation structures used for microwave and millimeter-wave components and circuitry fit into a broad

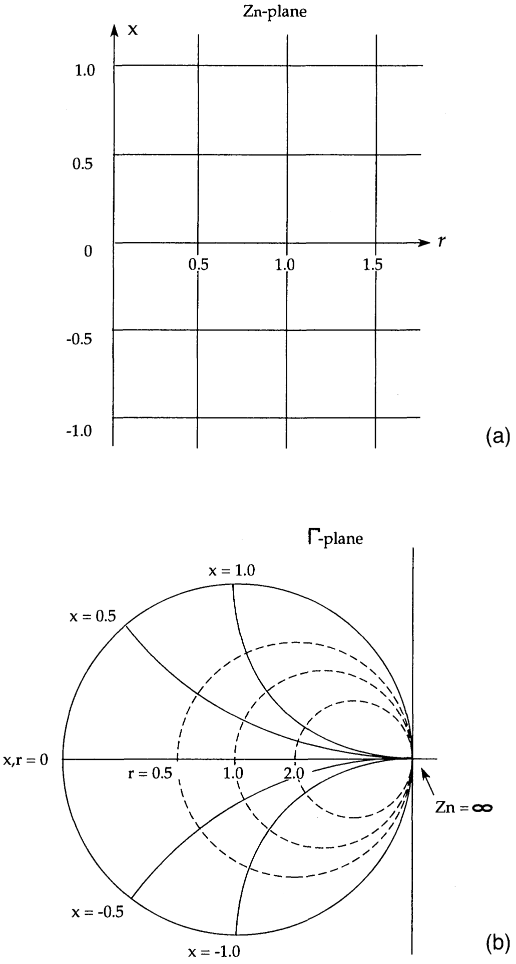

FIG. 3. (a) Mathematical representation of the reflection coefficient Γ in rectangular coordinates; (b) transformation of $\Gamma < 1$ to the unit circle (Smith chart).

collection of methods of confining fields and guiding power. The simplest physical method is to use electrically conductive pipes or tubes; indeed, metallic waveguide development prior to World War II depended upon unlikely sources of supply. The extrusions available at that time defined rectangular waveguide cross sections available to microwave practitioners and have survived in many instances to today. An example of particular note is WR 137 waveguide, whose outside dimensions are $1\frac{1}{2}$ in. by $\frac{3}{4}$ in., and which graces architectural designs as exterior stair railings and metal storm doors in the northeastern United States. Dielectrics in an air surround can also be used to confine energy (dielectric waveguide), the best example of which is single-mode optical fiber cable, with an installed base of greater than 4 Gm in the United States alone. At the circuit level, inexpensive planar guiding structures are easily defined with precise photolithographic processes on a wide range of dielectric substrates. Discrete active and passive devices can be integrated on the surface of these dielectrics, or can be monolithically integrated with waveguiding structures on semiconducting materials that serve both as low-loss substrate and as host material for active devices. Selection criteria for propagation media include producibility and cost, size, mechanical interfaces to devices and other structures, impedance, power handling, electrical loss, and finally accuracy of analytical circuit models.

2.1 TEM Structures

Transverse electromagnetic fields, by definition, have no field component in the direction of propagation. The cutoff wave number is zero (TEM waves are dispersionless), and field solutions are found from Laplace's equation in the transverse plane, familiar from electrostatics. As a direct consequence, the voltage between and current on the conductors defining the guiding medium (and therefore the impedance V/I) are unambiguous and directly applicable as circuit representations for TEM structures. Where defined, these impedance levels are functionally related to the dielectric constant and the ratio of characteristic transverse dimensions of the transmission line.

The simplest guiding medium supporting TEM waves is coaxial transmission line [Fig. 4(a)]. The frequency dependence of attenuation per unit length resulting from conductor losses depends on R_s, whereas the dielectric loss depends directly on frequency. This frequency dependence is universally valid for TEM structures and is an accurate low-frequency approximation for quasi-TEM structures. At very high frequencies (in particular, when the wavelength in the medium approaches the median circumference of the coaxial structure), higher-order (non-TEM) modes are supported by the coaxial structure. Cost and single-mode propagation at high frequencies favor smaller diameters, at the expense of conductor loss. These compromises are general and apply to many other guiding structures.

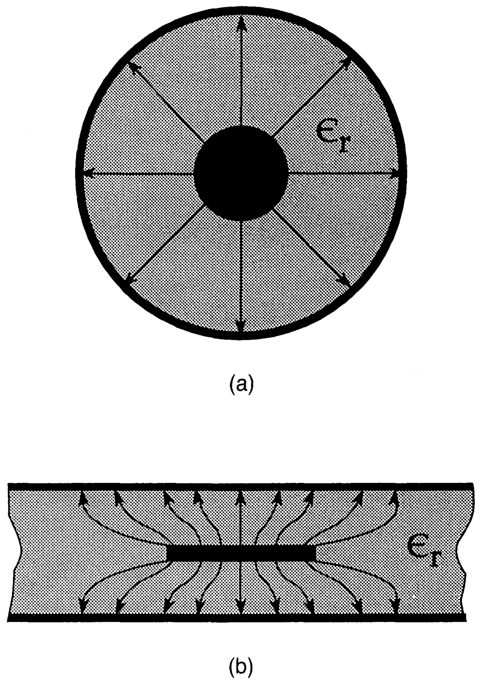

(a)

(b)

FIG. 4. TEM structures: (a) coaxial transmission line; (b) symmetric stripline.

In symmetric stripline, a conductor is embedded in a dielectric medium with ground planes located symmetrically about it [Fig. 4(b)]. Commonly used materials for stripline construction employ polytetrafluoroethylene (PTFE) in a woven glass matrix, a so-called "soft" substrate with a dielectric constant slightly larger than 2.

2.2 Quasi-TEM Structures

Quasi-TEM lines are comprised of two separate conductor electrodes embedded in an inhomogeneous dielectric. Longitudinal (small with respect to transverse) components of fields exist and lead to ambiguous definitions of impedance. An effective dielectric constant valid at nonzero frequency can be defined and computed using a number of methods. These procedures give results that converge to pure TEM field solutions as frequency approaches zero (Hoffman, 1987).

Microstrip consists of a single conductor of width *w* on a dielectric medium (substrate) of thickness *h* with a ground plane on its opposite face [Fig. 5(a)]. The electrical ground plane is also an effective system thermal ground, accommodating devices that generate or dissipate heat. Because of its ease of fabrication, microstrip is the workhorse of contemporary microwave circuit realizations. Microstrip geometry allows propagation of higher-order TE or TM modes as surface waves on a grounded dielectric slab, which exhibit exponential decay in regions outside the dielectric. These modes are minimized when the substrate thickness is chosen to be less than a quarter wavelength at the highest operating frequency.

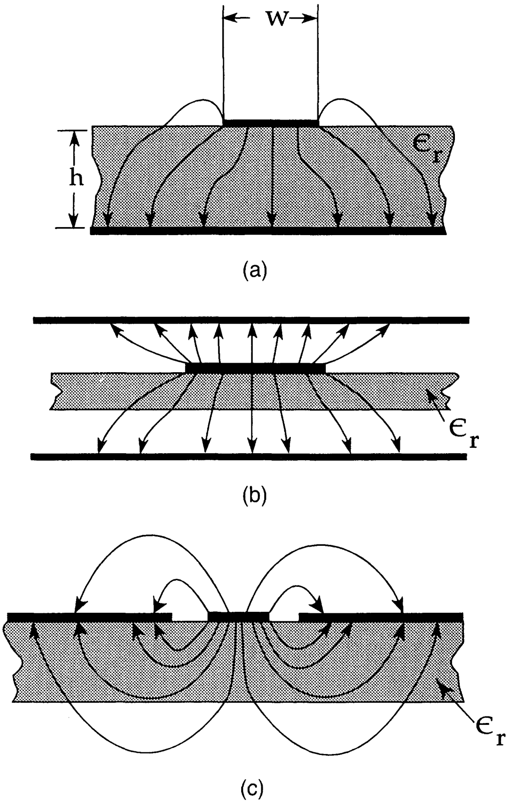

(a)

(b)

(c)

FIG. 5. Quasi-TEM structures: (a) microstrip; (b) suspended substrate stripline; (c) coplanar waveguide.

Removing the microstrip ground plane and suspending the resulting structure in the H plane of a rectangular conducting channel results in a low-loss quasi-TEM transmission medium called suspended stripline [Fig. 5(b)]. This configuration is useful as a transition from planar media to waveguide and has low losses, thus offering a planar alternative to waveguide at millimeter-wave frequencies.

Coplanar waveguide (CPW) consists of a single conductor centered in a gap between conducting ground strips on the top surface of a dielectric medium of semi-infinite extent [Fig. 5(c)]. Because both conductor electrodes are on the top surface of the substrate, CPW can be used where metallization of both sides of the substrate is difficult. In addition, the relative ease of integrating shunt and series devices in CPW offers a complementary alternative to microstrip integration techniques. Advances in the technology of microwave measurements have incorporated CPW in a probe/transition configuration in order to extend the nondestructive vector (amplitude and phase) measurement capability on planar circuits to 100 GHz (Fig. 6).

2.3 Non-TEM Structures

For regions enclosed by a single conductor, solutions to the wave equation depend upon the exact geometry of the surrounding region. For a conducting boundary of rectangular cross section and of infinite extent in the propagation direction, TE and TM waves can exist. In general, discrete modes of propagation exist for each possible value of k_c allowed by the boundary conditions. Unambiguous definitions of voltage and current do not exist, and therefore circuit definitions of impedance apply only under very restrictive conditions.

Metallic waveguide confines the power while retaining most of the low-loss properties of air. The TE and TM solutions each exist in an infinite set. Dimensions of standard rectangular guides are such that fundamental-mode operating bandwidths are less than an octave. A number of variations of rectangular guide have been devised to take advantage of its low-loss characteristics. Metallic guiding structures in the E plane of rectangular guide, usually supported by dielectric and called finline, allow a smooth transition from waveguide to printed substrates. To increase bandwidth while minimizing loss, ridge guide (Hopfer, 1955), a modification of rectangular guide, is used. Circular waveguide (which can be envisioned as a deformation of square waveguide) also has an infinite set of TE and TM modes.

Signal transmission using optical fibers offers a low-loss, low-cost alternative to circular waveguide for long-distance communications applications and has been the subject of intensive research and development since the development of the semiconductor laser in 1964. Losses as low as 0.2 dB/km have been measured at a wavelength of 1.55 μm, the wavelength of one available type of semiconductor laser frequency source. This frequency (194 THz!) hardly fits the microwave boundaries mentioned above, but since fiber-optic transmission lines are displacing present microwave transmission media and because modulation and demodulation techniques will continue to require microwave circuit methods, the distinction between microwave and light-wave techniques and vocations will continue to blur.

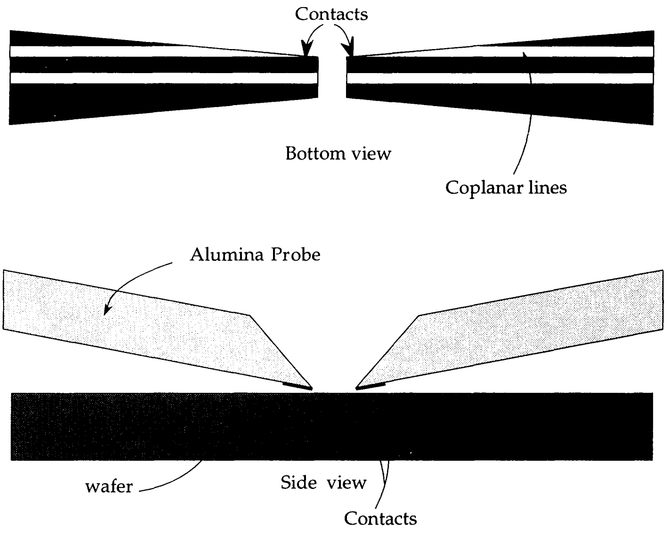

FIG. 6. Coplanar waveguide microwave probing geometry, allowing in-circuit (or on-wafer) measurements at frequencies as high as 100 GHz.

2.4 Enabling Technology: High-Temperature Microwave Superconductivity

A major driver in the development of transmission media for microwave circuits has been the desire to minimize both dielectric and conductor losses. Recent spectacular results in superconducting oxides show great promise for microwave applications. Superconductivity was first observed in 1911 by H. Kamerlingh Onnes, the first person to liquefy helium (Gorter, 1964), but applications of the effect were not realized until the early 1960s when superconducting properties of Nb_3Sn wire, and the invention of the Josephson junction, initiated the superconducting magnet and Josephson electronics. The recent intense interest in superconductivity dates to the 1986 announcement of a 30-K transition temperature in a superconducting oxide by Bednorz and Müller. In a two-year burst of activity, thin-film superconducting oxides with transition temperatures to 125 K were produced, giving rise to the description high-temperature superconductivity (HTSC). Recent activity in the mercury-based layered copper oxides has raised this temperature to 133 K (Schilling *et al.*, 1993).

Although the frequency dependence of superconductor surface resistance varies as f^2 compared to the more gradual $f^{1/2}$ variation in normal conductors, much lower absolute values of resistance can be attained by superconductors at frequencies in the microwave region (Fig. 7). At 77 K, the most-studied HTSC oxide $YBa_2Cu_3O_7$ (or YBCO) has a surface resistance of 0.2–0.6 mΩ at 10 GHz, compared to 12 mΩ for copper. This advantage increases markedly below 10 GHz and disappears at frequencies above 100 GHz. As the technology of the oxide superconductors matures, the liquid-nitrogen loss of YBCO should approach that of niobium at 4.2 K, and the liquid-hydrogen loss of YBCO should approach that of Nb_3Sn at 4.2 K.

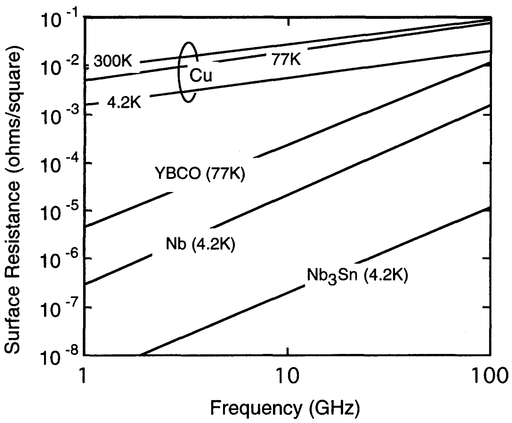

FIG. 7. Surface resistance vs frequency for selected normal conductors and superconductors.

The physics of growth techniques, physical properties, and low-temperature (4 K) performance of the HTSC oxide films are not well understood; nevertheless, performance at liquid nitrogen temperatures may be approaching practical limits. A particular item of interest is film stability; patterning of YBCO films causes increases in film resistance due to the heavy concentration of current at the edges of the conductor. This issue is of vital significance to successful realization in media like microstrip, and for devices requiring narrow line structures. Substrates for these oxides necessarily match the crystal lattice structure of the as-deposited films. The compounds $LaAlO_3$, MgO, and α-Al_2O_3 (sapphire) are all acceptable substrate materials but except for sapphire are dominated by dielectric losses in the microwave region at 77 K. Development of low-loss substrate materials, therefore, will continue to be an important part of HTSC work for future microwave applications.

Electromagnetic coupling between adjacent conductors in planar transmission media depends on the ratio of conductor spacing to dielectric thickness, and can be precisely computed for applications such as filters and directional couplers. Unintentional coupling (crosstalk) is reduced to minimal effect when the conductor spacing is more than a few dielectric thicknesses. Circuit density limitations due to crosstalk can be decreased while maintaining constant impedance by reducing the thickness of the substrate and the line width, but at the cost of increased loss. While all conductors experience increased losses with reduced conductor widths, superconductors, for the same loss and impedance level as normal conductors, allow increases of greater than two orders of magnitude in circuit density considering coupling effects.

Applications of HTSC thin films in the microwave frequency region exist where the system cost of providing the cooling necessary to keep the film cooler than the critical temperature is justified by the reduction of loss. Immediate uses of this technology exist in spacecraft signal-processing systems, where passive radiation-cooling techniques can achieve component temperatures below 100 K, and simple single-stage closed-cycle refrigeration systems can reach temperatures below 60 K.

2.5 Future Developments

Attenuation characteristics vary widely among different propagation media (Fig. 8) and therefore serve different needs of signal transmission. Although the most recent transmission-line configurations discussed above are nearly 20 years old, new life for transmission media in those applications where cooling to 77 K can be accommodated is promised by HTSC technology.

Development in substrate materials and packaging technology has helped to maintain a steady growth in this area, which has a great deal of overlap with circuit integration. Thermoset microwave materials, for example, combine the dimensional stability attributes of hard substrates with the machinable and low-cost characteristics of soft substrates. Typical costs for soft dielectric substrates are \$0.15/cm^2, compared to ceramic substrate costs of greater than \$1/cm^2, and semiconductor substrate costs of \$10/cm^2.

In the area of semiconductor materials, work will continue on integration of lightwave components such as light-emitting diodes, semiconductor lasers, and waveguide with microwave and millimeter-wave components used for modulation and frequency conversion purposes. Multilayer propagation media address the need for increased packaging density by stacking layers of signal and ground planes sandwiched between dielectrics. A popular version of this technique is low-temperature co-fired ceramic (LTCC). Thick-film conductive paste defines signal and ground traces on unfired ceramic tapes, which are stacked and fired to final configuration. Low-temperature co-fired ceramic finds use to frequencies as high as 20 GHz but is lossier and more expensive than thin-film techniques on single-layer ceramic. Organic polymers such as benzocyclobutane, an organic dielectric with low dielectric constant and low loss tangent, offer an alternative method of

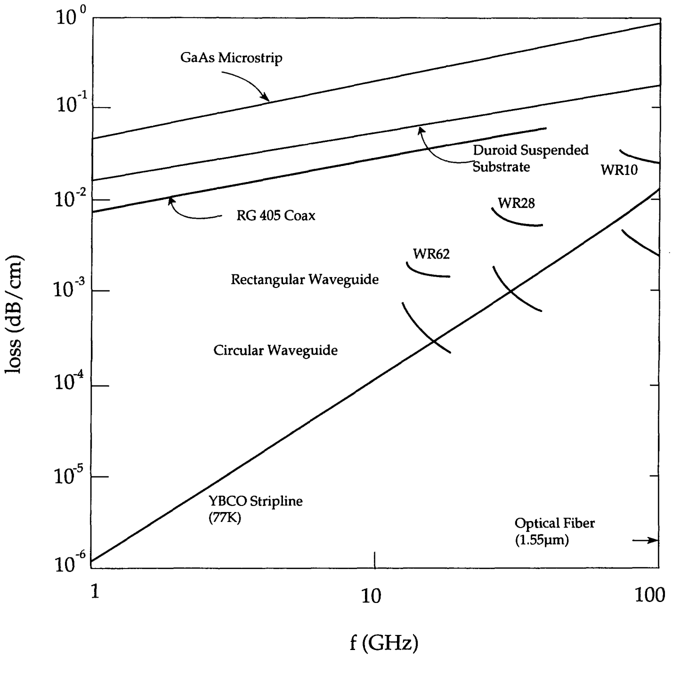

FIG. 8. Attenuation vs frequency of selected microwave and optical propagation media. Waveguides are copper. Circular guide diameters are chosen to match cutoff frequency of corresponding rectangular guide.

constructing three-dimensional propagation media using thin-film metallization techniques.

3. COMPONENTS

Single-function circuit elements used for signal control, frequency conversion, and amplification employ lumped and distributed circuits and propagation media described above with active elements such as transistors and diodes. Control components modify the amplitude or phase of the input signal, providing methods of routing the signal through a microwave subassembly. Signal generation and amplification components, classified by signal level, play a key role in both receiver and transmitter sections of microwave subsystems. Frequency conversion components include mixers and frequency multipliers. Both are used extensively in subsystems where (increasingly digital) signal *processing* is most efficiently implemented at low (baseband) frequencies, whereas signal *transmission* is at higher microwave or millimeter-wave frequencies. At these frequencies, large information content (or equivalently, bandwidth) is a small percentage of the carrier frequency, allowing manageable component design.

3.1 Passive and Signal Control Functions

Attenuators are used to control the amplitude of a microwave signal in microwave systems. One-port attenuators (terminations) eliminate reflections from transmission lines and are realized in either lumped-element (smaller, less expensive) or distributed-circuit (higher-frequency, lower-reflection coefficient) form. Mechanically variable attenuators are realized in waveguide with the controlled insertion of lossy E-plane card into the center of the guide. Variable attenuators of this type are used in precise measurements of power or control of signal level in instrumentation applications. Input- and output-matched attenuators of the pi or tee (Fig. 9) construction provide losses of less than 1 dB to greater than 20 dB, and represent a lumped-element approach to the fixed attenuator. Thin-film resistors on planar substrates or shunt discs and series rods for coaxial use are common. For waveguide, a tapered resistive card located in the E plane of the guide will provide fixed loss with very small return loss. Using *p-i-n* diodes or metal-semiconductor field-effect transistors (MESFETs) as current- or voltage-controlled (respectively) variable resistors, and the matched-tee approach as described above, inexpensive ($0.60/chip) low- to medium-power variable attenuators exhibiting extremely large bandwidths can be fabricated. Analog dc voltages control the drain–source resistance of the field-effect transistor (FET) to achieve the required attenuation.

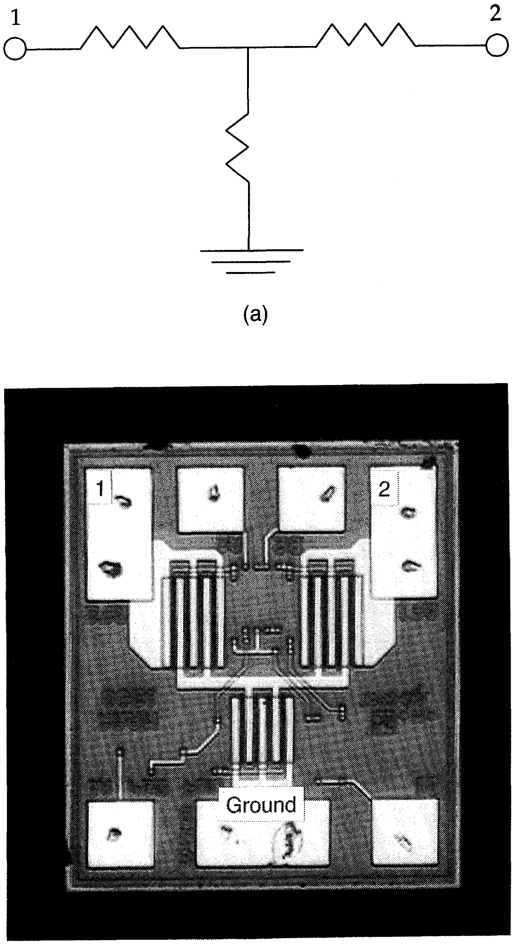

(a)

(b)

FIG. 9. (a) Matched-tee variable attenuator (schematic). (b) Matched-tee attenuator in MESFET MMIC technology; chip size: 600 × 600 μm. (Courtesy of M/A-COM, Inc., Lowell, MA.)

Microwave switches can be of mechanical construction, or can be realized with semiconductor devices (*p-i-n* diodes or FETs), or with ferrite devices. Mechanical waveguide switches have extremely low mismatches (VSWR < 1.05) and are used in instrumentation applications where switching speed, size, and cost are of secondary importance. Switching with *p-i-n* diodes or FETs offers advantages of fastest switching speeds (<10 nsec), low (*p-i-n*) or extremely low (FET) drive-current requirements, and the possibility of monolithic integration to realize higher functional integration. Both devices have a high-impedance state and a low-impedance state that depend upon bias, and can be used in the shunt or series configuration to increase on/off ratio (isolation). Although *p-i-n* diodes have been successfully used in applications to 100 GHz, and offer lower losses than FETs, low-cost monolithic integration techniques have made FET-based switches at lower microwave frequencies one of the early successes of gallium arsenide (GaAs) commercial foundry operations (Fig. 10). "Drivers" for switches transform digital-logic voltage levels to proper current waveforms for *p-i-n* diodes or proper voltage levels for FETs. These circuits are evolving from a hybrid form (using discrete components on a separate printed circuit board) to a monolithic form, either as a separate application-specific integrated-circuit (ASIC) chip or as part of the microwave monolithic circuit fabricated to support both digital and analog functions. Ferrite switches are three-port devices that are inherently single-pole, double-throw. These devices offer lower loss than the semiconductor switches and higher power handling, with the disadvantages of microsecond switching speeds, larger size, and more complicated driver circuitry. These devices require no bias in the steady state; i.e., they are latching switches. Issues of size, cost, speed, integration ability, and driver complexity are causing semiconductors to displace ferrites in all but a few custom applications.

Because of the importance of phase in any signal-processing system, phase shifters serve important roles as calibration tools, as balance or trim elements in multichannel systems, or to realize efficient power combining. They are, of course, key elements in phased-array radars, where beam steering is achieved electronically by controlling the phase relationship among thousands of individual transmit/receive modules. Continuous phase shift can be realized mechanically with "trombone"-like coaxial lines up to 20 GHz, and with waveguide configurations similar to the variable attenuator described above, where the resistive card is replaced by a low-loss dielectric card. Discrete values of phase shift can be electronically realized using FETs or *p-i-n* diodes as electronic switches, to select from one of two phase-shift states ("bits"). Analog techniques using the variable (with dc bias) reactance properties of low-loss varactor diodes are used to achieve continuous phase shift with application of dc bias. Here, too, ferrites offer alternatives to semiconductors for phase-shift applications but show ad-

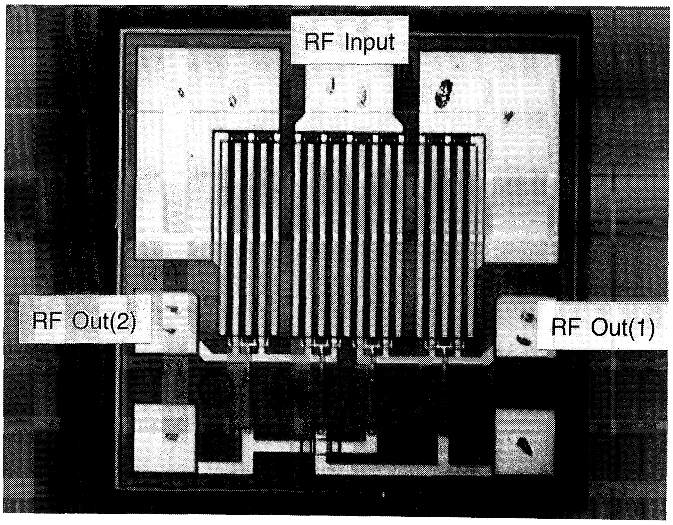

FIG. 10. Single-pole, double-throw MMIC switch in MESFET MMIC technology; chip size: 600 × 600 μm. (Courtesy of M/A-COM, Inc., Lowell, MA.)

vantage only where highest powers (tens of kilowatts) are required. Monolithic realizations of semiconductor phase shifters will operate to higher frequencies, switch faster, but handle less power than hybrid realizations.

The anisotropic permeability in ferrites allows the direction of wave propagation to be modified by the application of a dc magnetic field. This property has been exploited in the development of nonreciprocal devices, two examples of which are circulators and isolators. A junction, or Y, circulator is a three-port network in which signals (waves) may "circulate" from port 1 to 2, 2 to 3, and 3 to 1, but not in the opposite direction, while all ports are matched at their inputs. These junctions are commonly implemented in stripline (Fig. 11). With the application of an axial dc magnetic field, an electric field null is placed at port 3, and "circulation" of power from port 1 to port 2 is allowed. The direction of circulation is controlled by the polarity of the dc magnetic field. The conditions for circulation (as well as impedance matching and loss) depend on operating frequency, magnetic properties of the material, and physical dimensions of the ferrite disc. By terminating the isolated port, three-port junction circulators become two-port isolators. Z-directed magnetic fields resulting from a current-carrying wire placed within the ferrite allow operation without permanent magnets. By reversing the direction of the current, the circulation pattern is reversed, and switching behavior is achieved. Because of magnetic field hysteresis within the ferrite material, no sustaining current is required to maintain a given state. Switching speeds in the microsecond range are attained with these ferrite latching devices.

Used to define the frequency band of interest at the component or subsystem level, microwave filters are categorized as low-, band-, or high-pass, or band-reject, depending on the desired frequency selectivity. Realizations are primarily passive at microwave frequencies, although active filters, including switched-capacitor topologies with high-gain amplifiers as key elements, are used extensively at frequencies below 100 MHz, where inexpensive monolithic chips provide gain block functions. The theory of direct coupled-cavity filters has been advanced continuously since work initiated at the MIT Radiation Laboratory (Levy and Cohn, 1984). Distributed circuit theory in filters traces its origin to Richards, who established a simple relationship between lumped and distributed circuits in 1948, allowing the existing lumped-element filter theory to be applied to distributed realizations. In the intervening time, synthesis techniques have been developed for increasingly compact designs. Higher-order mode cavities and dielectric resonators have allowed further size reduction. Today, filter theory provides a valuable base for integrated designs requiring large

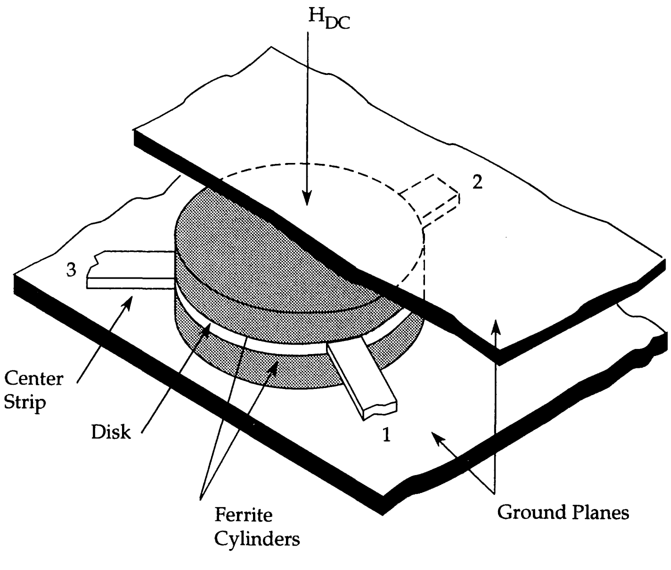

FIG. 11. Stripline junction circulator.

percentage bandwidths. Conventional component requirements for filters include limiting the noise input power to a sensitive receiver, reducing spurious signals generated by active or nonlinear components in a system, separating input and output signals in nonlinear components such as mixers and multipliers, and segmenting wide-bandwidth (octave or greater) receivers into frequency bands that can be processed by components with modest bandwidths. Microwave spectrum analyzers are instrument-grade versions of wide-band receivers, which use fixed and tunable filters to identify frequency of input signals.

Four-port junctions have special properties which have made them very useful in microwave circuits, including applications of power combining, signal duplexing, and power coupling. Important elements of these junctions were discovered by Schwinger and Dicke in the early 1940s at the MIT Radiation Laboratory using mathematical symmetry arguments (Montgomery *et al.*, 1948). These plus later synthesis techniques in planar transmission media are major milestones in the history of passive microwave circuit development. Hybrids, a special class of four-ports, are impedance matched at all ports simultaneously. [This special use of the word "hybrid" should not be confused with its other use in microwave circuits (see Sec. 4, below) to describe a mixture of integrated circuits and discrete devices on a common substrate.] A signal incident on any port is coupled to two others and isolated from the fourth. Hybrids with equal power coupling are called 3-dB hybrids or *magic tees* and have been realized in waveguide and distributed- and lumped-element forms. The magic tee in waveguide consists of an H-plane tee junction with the fourth port forming an E-plane tee [Fig. 12(a)]. When a TE_{10} mode is incident on port 1, the electric field pattern distribution is such that port 4 is completely isolated, and the power split is equal and in phase between ports 2 and 3. Likewise, for TE_{10} excitation at port 4, port 1 is isolated, and power is equally split with phase difference of 180° between ports 2 and 3. Lumped-element and planar versions of the magic tee also exist. Directional couplers are four-port hybrids in which portions of the forward and reverse traveling waves are separately coupled to two additional ports. Realizations of these components usually involve transmission lines whose relative proximity determines the coupling, but can be implemented with lumped elements as well [Fig. 12(b)]. Synthesis procedures in many media have been developed (Wadell, 1991). Special constructions have been invented for specific media. In microstrip, for example, couplings tighter than 6 dB require extremely narrow line spacing or interdigital construction techniques. Waveguide directional couplers exhibit extremely low loss and directivities as high as 50 dB, which find extensive use in instrumentation or measurement applications where amplitude and phase information of a traveling wave is desired.

3.2 Signal Generation and Amplification

Techniques for microwave signal generation address a variety of application-specific requirements. For communication, radar, and instrumentation purposes where efficient use of allocated bandwidths is important, stability (frequency change with time) and spectral purity (spectral line width and spurious signals) of frequency sources are key requirements. For industrial applications such as microwave heating, these requirements are much less demanding but are replaced by the need for high power and dc/rf power conversion efficiency. Microwave frequency generation at the fundamental frequency uses active devices with gain in a circuit which provides feedback appropriate to the conditions for oscillation. Oscillators can be tuned mechanically or electronically by perturbations of the resonant circuit. The quality factor of the resonator will determine the tuning range and stability of the oscillator, with high Q's resulting in narrow-band, low–phase-noise components. [The figure of merit used for oscillator comparison, phase noise, is caused by phase fluctuations within the oscillator, which broaden the frequency spectrum of the source. The term $\mathscr{L}_m(f_m)$ describes the rf power spectrum and is defined as the ratio of the single-sideband phase noise in a 1-Hz bandwidth f_m Hz away from the carrier frequency to the total signal power.] Quartz crystal resonators at frequencies of 5–100 MHz, with unloaded Q's greater than 10^6 at 5 MHz, followed by frequency multiplication to the microwave region, or with dielectric or cavity resonators used as the fre-

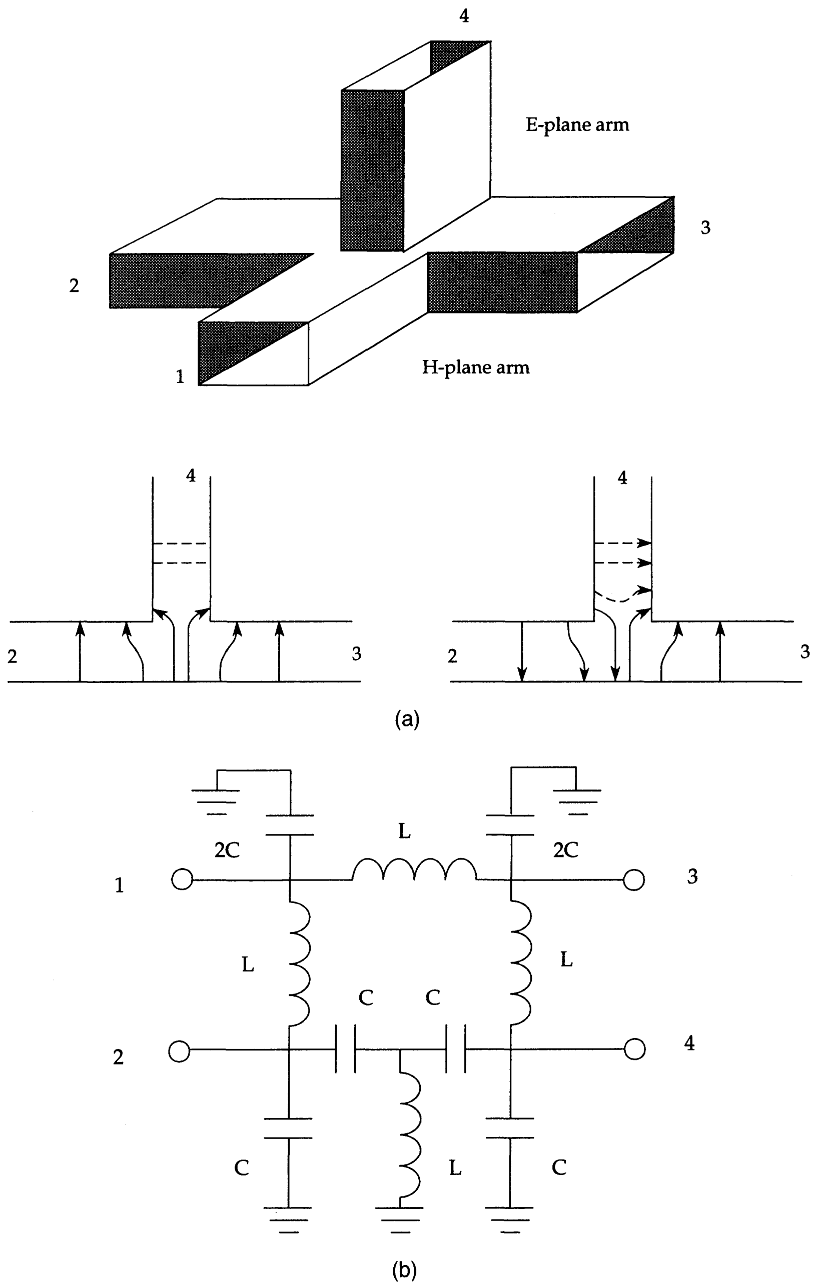

FIG. 12. (a) Waveguide magic tee; and (b) lumped element magic tee.

quency-determining element in fundamental frequency oscillators are commonly used to provide low–phase-noise sources. The active element used in the oscillator also affects the phase noise: Oscillators using Si bipolar junction transistors have better phase noise performance than FETs in otherwise identical oscillator environments (Fig. 13).

Microwave amplifiers are designed for various applications, including a low-noise region, where very weak signals at receiver inputs are amplified with a minimum of internally generated noise and spurious signals, and a high-power region, where power output and conversion efficiency are critical characteristics. Low-noise amplifiers (LNAs)

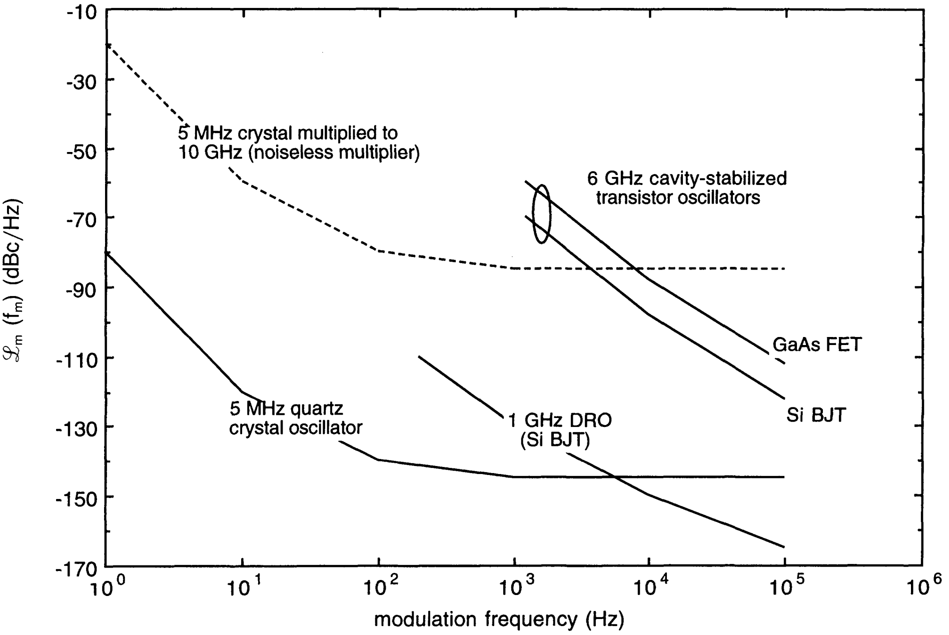

FIG. 13. Oscillator phase noise vs modulation frequency (normalized to operation at 10 GHz).

used at the input of sensitive receivers amplify signals that in some applications can be lower than background thermal noise power of −174 dBm (measured in a 1-Hz bandwidth at room temperature). An ideal (noiseless) component will add no noise to the thermal noise available at its input. Noise figure *F* is the ratio of output noise power to the product of amplifier gain by input noise, whose minimum value of 1 is being approached by a number of components in the microwave region (Fig. 14). The maser provides the ultimate in low-noise amplification, but its expense and need for cryogenic operating temperatures limit its use to the most demanding of applications such as radio astronomy. Cooled three-terminal solid-state devices offer lower cost and size advantage with a slight degradation in performance. Parametric amplifiers, developed during the early 1960s, use variable reactance (*varactor*) diodes to accomplish frequency conversion with gain and little additive noise, but are being replaced by cooled high–electron-mobility transistor (HEMT) amplifiers, which show excellent low-noise results to frequencies as high as 100 GHz. The 3-K galactic noise temperature, measured at 6 GHz by Penzias and Wilson in 1964, is one of the most profound examples of microwave remote sensing. Measurements on the galaxy made with an extremely low-noise antenna/receiver at Bell Laboratories repeatedly yielded noise powers higher than expected. Later verified as accurate readings, these noise measurements were used to support the Big-Bang theory of the birth of the universe.

Methods of amplifying large signals include vacuum-tube techniques, which are capable of megawatts of power but are expensive and large, and solid-state approaches, which are suited for lower power applications but offer the cost advantages of high-volume semiconductor manufacturing technology (Fig. 15). Vacuum-tube techniques include traveling-wave tubes, magnetrons, crossed-field amplifiers, gyrotrons, and klystrons (see VACUUM AND GASEOUS ELECTRON TUBES). Developments in this field date from the mid-1930s. These components use structures whose fabrication requires a great deal of mechanical precision (high cost), whose operation requires high-voltage power supplies (>1 kV), and whose lifetime is limited by tube "wearout" and power-supply failure. In contrast, solid-state methods of generating power in the microwave and millimeter-wave frequency region include two-terminal

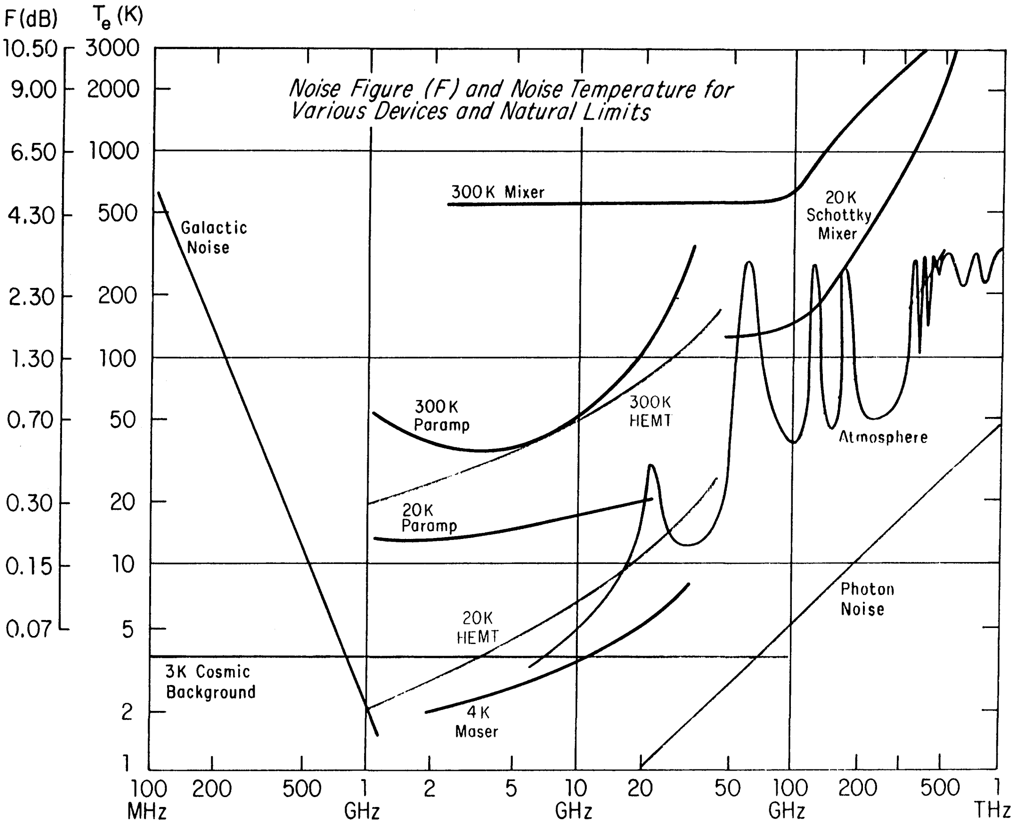

FIG. 14. Noise performance (in equivalent measures of noise figure and noise temperature) of microwave receiver components. Indicated temperatures are physical temperatures of devices or components. (Courtesy of National Radio Astronomy Observatory, Charlottesville, VA.)

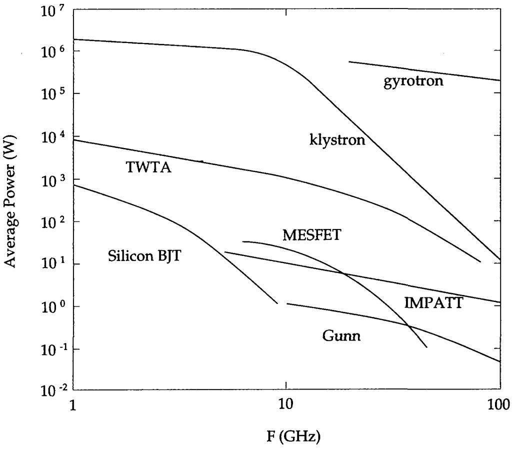

FIG. 15. Output power vs frequency for solid-state and vacuum-tube amplifiers.

devices such as IMPATT and Gunn diodes, and three-terminal devices such as MESFETs, heterojunction bipolar transistors (HBTs), and HEMTs. General features of solid-state power generation techniques include use of low-voltage power supplies (<40 V) and device lifetime predictions in the millions of hours. In the three-terminal device area, much effort is being applied in the GaAs homojunction and heterojunction areas to produce watt-level modules at frequencies as high as 35 GHz. Power densities twenty times higher than those achieved in FETs have been observed in HBTs, which also show extremely linear operation and high power-added efficiency ideal for cellular radio handset applications. In the portable telephone marketplace, "talk-time," or battery life, is crucially dependent on the efficiency of the microwave power amplifier. Advances in three-terminal solid-state power capability in the millimeter-wave frequency region during the next decade will be driven by automotive applications which are targeted for the 60–80 GHz range; already FETs, HEMTs and HBTs exhibit small-signal oscillation frequencies which satisfy these requirements.

3.3 Frequency Conversion Components: Mixers and Frequency Multipliers

The frequency mixer, a three-port device, converts an input frequency signal (f_{in}) to an output signal using a third signal, called the local oscillator (f_{LO}). Based upon device nonlinearities within the mixer, the output signal spectrum consists of all the multiplicative single-tone intermodulation products, namely $\pm mf_{\text{in}} \pm nf_{\text{LO}}$ (m and n are the positive integers, including zero). It is the task of the mixer or circuit designer to select only the output frequency or interest (usually the $m = n = 1$ product). Other intermodulation (IM) products are suppressed using filtering external to the mixer or by symmetry within the mixer. The higher-order responses (high values of m and n) are not efficiently converted within the mixer. Devices used in mixer design are chosen for low loss, low noise, and nonlinearity; Schottky diodes, FETs, and, more recently, HBTs are most popular. Silicon diodes exhibit lower spectral noise than GaAs diodes but have greater conversion loss. Mixers based on transistors as nonlinear elements are lower in performance than the best diode mixers but offer the possibility of monolithic integration with other transistor-based receiver functions. In receivers, mixers can contribute a large part of the IM distortion, which sets the large-signal limit on dynamic range, the range of signals that allow spurious-free or linear operation. Reducing the gain in front of the mixer stage to reduce the mixer IM distortion contribution will increase the system noise figure and increase the small-signal limit on dynamic range. Receiver design is a compromise between dynamic range and sensitivity, which requires careful choice and design of the individual components as well as detailed attention to the frequency plan to avoid unwanted mixing products and other spurious signals. The simplest mixer uses a single semiconductor diode. (Single) balanced mixers use two diodes and either a 90° or 180° hybrid to provide LO-rf isolation or even-order harmonic rejection. Double-balanced mixers using four diodes provide isolation among all three ports while suppressing even-order harmonics (Fig. 16). Additional balancing techniques employing increased circuit complexity can be used to reduce intermodulation terms further. Conversion losses of 3–8 dB are typical, although careful reactive termination of unwanted harmonic outputs can reduce the loss below 3 dB.

Analytically, frequency multipliers are degenerate cases of mixers (no local oscillator), with output frequency spectrum given by mf_{in}. Multipliers offer methods of converting low-frequency, ultralow-noise signal sources (produced by crystal oscillators in the 5–100 MHz frequency range) to microwave and millimeter-wave frequencies. Conversion efficiencies vary from a few percent (typical of multipliers that use resistive nonlinearities) to nearly 100% for multipliers that use reactive nonlinearities. These latter devices have been developed in high-power versions for providing watt-level power in the millimeter-wave frequency range. Spurious responses are reduced using circuit balance and filter techniques, as for mixer circuits.

3.4 Future Developments

Control functions such as switches and attenuators will increase in complexity with the addition of digital logic and digital-to-analog conversion to the analog semiconductor functions noted above. Switch and amplifier topologies using active elements, and designed as filter topologies, will allow efficient space-conservative broadband designs. Improvements in materials and electromagnetic design tools will allow continuous reductions in the size of filter functions while maintaining performance. At the same time, digital techniques will invade the lower-GHz frequencies and establish equivalent functions that can be integrated into the rest of the digital circuit functions. Thin-film acoustic resonators (Lakin *et al.*, 1993) will compete with high–dielectric-constant materials for filter applications where small size and integration with other functions are important. Ferrite circulators, isolators, and

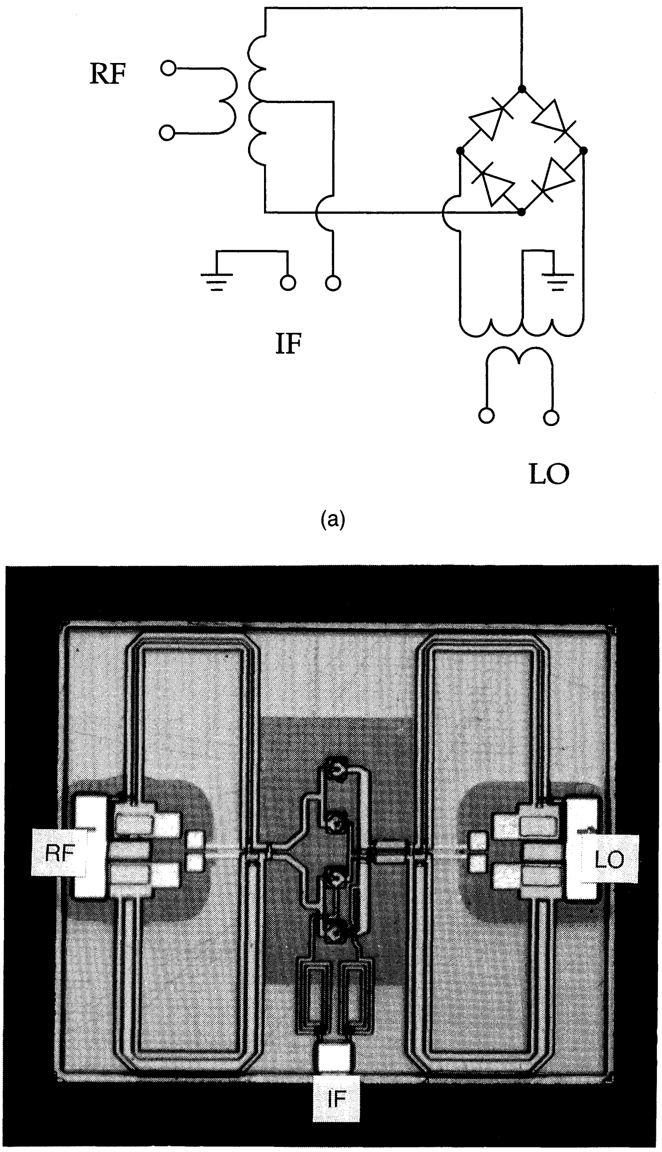

FIG. 16. (a) Double-balanced mixer schematic; (b) realization using Si monolithic Schottky diodes and glass circuit integration; chip size, 1.9 mm × 2.3 mm. (Courtesy of M/A-COM, Inc., Lowell, MA.)

switches will be replaced by semiconductor equivalents as needs for small size and low cost force alternative approaches.

With increases in the power capability of solid-state devices, most likely to occur within the three-terminal category, the area of overlap with vacuum techniques will gradually shift to higher power–frequency products. Advances in multiple-channel HEMTs and pseudomorphic HEMTs for power, extension of the HBT to millimeter-wave frequencies, and development of FET material properties allowing higher power density will allow transistors to compete with tubes at higher power levels. Quantum effects will be used to push operation to higher frequencies, and diamond will be used as a semiconductor to achieve reliable operation at extremely high temperatures (Trew *et al.*, 1991). Combining the advantages of vacuum microelectronics with modern solid-state fabrication techniques, research in field-emitter arrays (FEAs) has been recently rejuvenated. The structure of the FEA allows significant current densities from micron-sized conical semiconductor mesas with modest applied voltages, giving projections of performance competitive with solid-state devices. Projected improvements in printed antenna distribution networks, and the possibility of integration of HTSC films with the active components of monolithic microwave integrated circuit (MMIC) technology such as the gate structure of FETs or HEMTs will improve receiver noise characteristics.

4. CIRCUIT INTEGRATION

Markets for single-function components include instrumentation and moderate-quantity microwave subassemblies, where performance is more critical than size and cost. With automated assembly techniques, discrete devices are used to manufacture multifunction assemblies in large volume. As the microwave integrated circuits (MICs) and monolithic integration of multiple circuit functions improve in performance and become less expensive to prototype, these methods of integration will serve component customers who need increasingly complex functionality, over broader frequency bandwidths, in less space, and at a lower price.

4.1 Microwave Integrated Circuits

MIC technology uses planar dielectric substrates that carry printed transmission lines and other planar passive circuitry, together with discrete passive and active devices to realize complex circuit functions. The earliest attempts to construct planar transmission lines on dielectric substrates used conductive pressure-sensitive tape to define the transmission lines. By plating both sides of a soft dielectric substrate with copper and skillfully using a sharp blade, similar results could be achieved, with resolution dependent upon the expertise of the individual, but limited to tenths of millimeters. With the development of photoresist and metal thin-film technology together with the availability of smooth dielectric substrates, however, the definition and resolution of patterns have moved to the range of microns, and batch processing has become a low-cost method of reproducing circuits. Microwave integrated circuits are used where timely empirical design verification, modifications, and repairs are important. Discrete devices can be selected before integration to assure high performance, and relatively quick design realizations can be achieved (a week or so from finished photomask to circuit).

Both soft and hard substrates are popular for use. Soft substrates made of materials such as PTFE or PTFE-based material are easy to machine into intricate shapes and are available in relative dielectric constant range from 2 to 10 (Table 1). Dielectric constants at the higher end of this range are obtained by using ceramic filler in the PTFE base. Pattern resolution is of the order of 20 μm. Hard substrates can be ground or polished to surface finishes (<1 μm) that result in better pattern resolution but require diamond machining tools or laser cutting techniques. Relative dielectric constants of commonly used hard substrates vary from 4 to 40, although dielectrics with $\epsilon_r > 100$ are available. Glass offers an attractive substrate choice because of its low microwave loss, smooth surface finish, and compatability with semiconductor processing (Fig. 17). Both substrate types allow multilayer structures consisting of alternating layers of conductive transmission lines or ground planes separated by insulating dielectric. Conductive interconnects between transmission-line layers or to electrical ground planes are formed by metallic plugs or drilled and plated "via holes" through the insulating dielectric layers.

Mounting of active power devices or other dissipative elements requires attention to the thermal path from the device to thermal ground. A number of methods have been devised to handle this problem in the MIC environment. Because of its high thermal conductivity and relatively low loss tangent, and in spite of its toxicity, beryllium oxide (BeO) has been used as microwave substrate material for microwave power amplifiers as well as for device packages. Aluminum nitride (AlN), an excellent thermal conductor and good dielectric, is an environmentally acceptable alternative to BeO. Epitaxial diamond, combining the highest thermal conductivity with excellent insulator properties, is commercially available in thicknesses of 1 mm and areas of 50 cm^2, allowing substantial improvements in power handling on MICs.

Table 1. Electrical and thermal properties of materials used in microwave circuits.

Material	Bulk conductivity (S/m)	Relative dielectric constant	Loss tangent (at 10 GHz)	Thermal conductivity (W/cm-°C)	Thermal expansion coefficient (ppm/°C)
Soft dielectrics					
PTFE glass random fiber	5×10^{-12}	2.2–2.4	1×10^{-3}	1.7×10^{-4}	100
Benzocyclobutane (BCB)	1×10^{-18}	2.65	8×10^{-4}	0.007	35–60
Cross-linked polystyrene	1×10^{-14}	2.55	5×10^{-4}	0.015	70
Hard dielectrics					
Fused silica	2×10^{-17}	3.8	1×10^{-4}	0.01	0.55
7070 glass	$<1 \times 10^{-15}$	4	2×10^{-3}	0.01	3.2
CVD diamond	$<1 \times 10^{-14}$	5.6	5×10^{-4}	13–15	2.6
BeO	$<1 \times 10^{-16}$	6.6	6×10^{-3}	2.6	7.8
Low-temperature co-fired ceramic	$<1 \times 10^{-16}$	7.1	1×10^{-3}	0.2	7.9
Si_3N_4		7	3×10^{-3}	0.3	2.3
AlN	$<1 \times 10^{-15}$	8.8	4×10^{-3}	1.7	4.4–4.8
(99.5%) Al_2O_3	$<1 \times 10^{-16}$	9.6	2×10^{-4}	0.19	6.7
Titanium dioxide	1×10^{-14}	40–100	4×10^{-3}	0.06	7.5
High-temperature superconductivity materials (properties at $T = 77$ K)					
YBCO	$7 \times 10^{+9}$				10
TCCBO	$7 \times 10^{+9}$				10
α-Al_2O_3 (sapphire)	1×10^{-16}	11.5	1×10^{-6}	0.42	6
MgO	1×10^{-16}	9.9	5×10^{-4}	6	0.8
$LaAlO_3$	1×10^{-16}	24.2	1×10^{-4}	5	10
Semiconductors (properties for semi-insulating material)					
GaAs	1×10^{-7}	12.9	2×10^{-3}	0.46	6.9
Si	1×10^{-3}	11.7	1.5×10^{-2}	1.5	4.2
InP	1×10^{-7}	12.6	1×10^{-3}	0.68	4.8
Metals/alloys					
Cu	$6 \times 10^{+7}$			4	16
Al	$3 \times 10^{+7}$			2.2	24
Cu/W (20%/80%)	$3 \times 10^{+7}$			2.5	8.5
Pb/Sn (eutectic)	$7 \times 10^{+6}$			0.5	25
Ni/Co/Fe (Kovar™)	$(1–2) \times 10^{+6}$			0.17	5.5

Prices today are nearly \$200/cm^2 (about as expensive as a completely processed GaAs monolithic integrated circuit) but are projected to drop to less than \$1/cm^2 by 1997. Even with the present price structure, overall circuit assembly costs may favor diamond substrates over more complicated heat-spreading techniques for some applications.

In normal MIC assembly practice, dielectric substrates are attached to metal carriers for integration into higher-level subassemblies. Thermal coefficients of expansion (TCE) of substrate and carrier must be well matched to avoid fracture of the substrate during thermal cycling of the subassembly. While alloys of Ni/Co/Fe provide good TCE match to Si and GaAs and are relatively inexpensive, the thermal properties are relatively poor. Copper/tungsten alloys achieve good thermal match to both alumina and GaAs while maintaining excellent thermal properties.

4.2 Monolithic Microwave Integrated Circuits

Properties of semiconductors such as Si and GaAs allow the addition of active devices such as transistors and diodes to a library of process-defined passive components. The key to success of passive functional integration is the low electrical conductivity of the substrate; in this regard, GaAs significantly outperforms Si. The monolithic integration of device and circuit functions on a single substrate holds the promise of reducing or eliminating expensive skilled manual labor used for discrete die attach and wirebonding operations, and has received support of in-

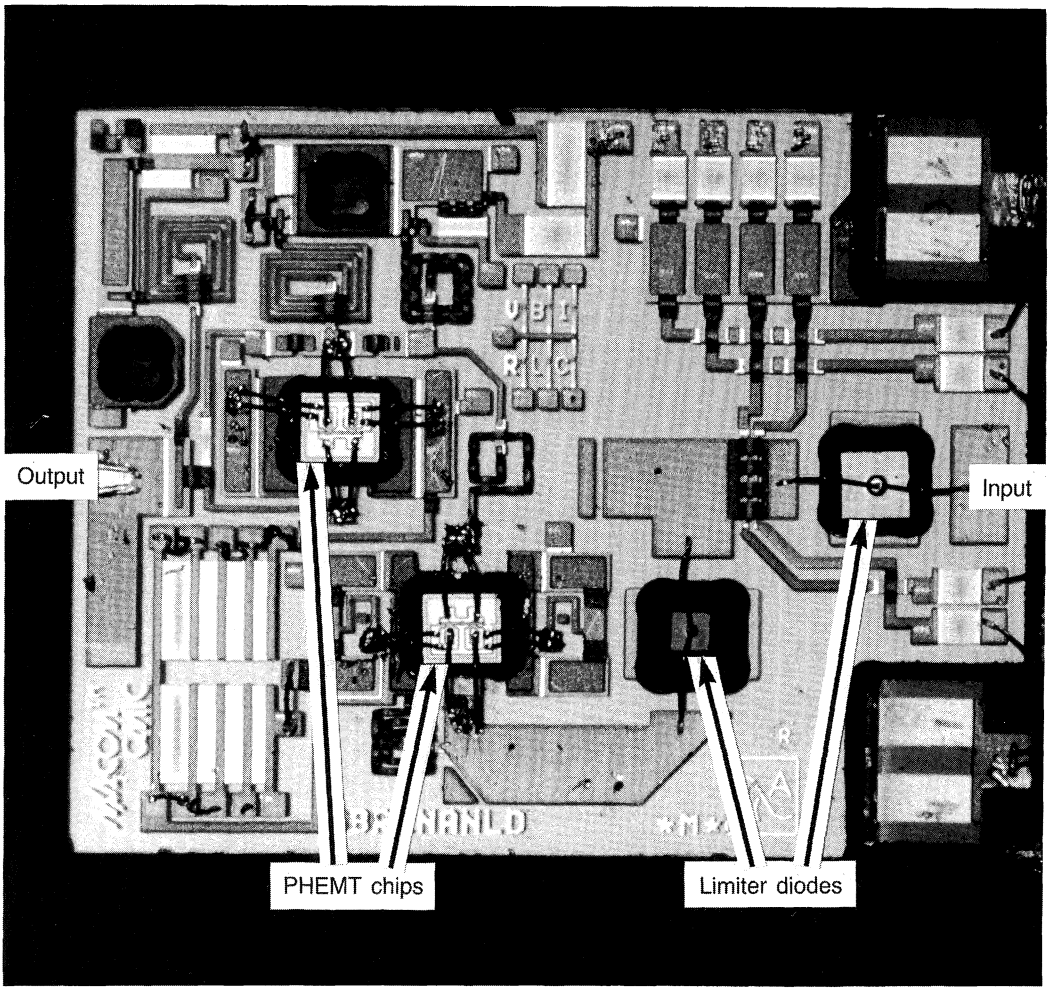

FIG. 17. 9–10 GHz limiter/low-noise amplifier MIC; component size, 3 mm × 4 mm.

dustry and government funding since the early 1980s.

The use of MMIC technology is prescribed when high volume and low unit cost are paramount, or in certain instances where circuit performance can be achieved only with monolithic fabrication. Large-scale integration and multifunction capability in a very small area are possible. The higher density of circuit elements minimizes undesirable interconnect lengths and results in higher operating bandwidths than MICs. Fewer wirebonds result in higher reliability.

Surface finish of semiconductor substrates is held to much closer tolerances than that of MIC substrates. Because active devices with feature sizes (FET gate finger "lengths") of as little as 0.1 μm must be made with high yield, surfaces must be optically smooth and free from scratches and particles. The resulting surface finish allows planar passive components such as resistors and capacitors to be defined as part of the processing of the substrate. These processing steps require more complex substrate preparation and processing but result in significant savings in assembly time, yielding cost savings in high-volume applications. The *airbridge* process photolithographically defines transmission lines (typically gold) to connect conductors separated by an air gap and an obstacle such as another transmission line. The airbridge replaces the wirebond with a precise, repeatable connection, and, because it is part of a batch manufacturing process, many subsequent labor-intensive wirebonding assembly procedures are eliminated.

Monolithic microwave integrated circuit processes are categorized according to the active device required for the bulk of the functional realization. These processes can be diode-based or transistor-based, and use either Si, GaAs, or more advanced semiconductor compounds. A process in use today for inexpensive low-noise as well as power MMIC functions is the GaAs ion-implanted process (Fig. 18). The starting (substrate) material is semi-insulating GaAs of extremely high quality, available as 3-in.- and 4-in.-diam wafers. The active channel layer impurity distribution is defined using a 100–500 keV accelerator. Subsequent wafer processing steps include the deposition and patterning of one or more thin films. Processing of the back side of the wafer can be as simple as mechanically thinning the wafer, or, in the case of a process which involves power devices, can include the formation of back-side metal vias to lower the parasitic inductance to electrical ground, as well as thermal resistance of power devices.

Monolithic device/circuit fabrication has resurrected the technique of distributed amplification (Ginzton *et al.*, 1948), developed to increase the operating bandwidth of vacuum-tube amplifiers. In this idea, transistors, each with finite input and output reactances, are absorbed into separate input and output transmission-line sections, and the resulting structure can be designed to have constant gain from very low frequencies right up to the maximum operation frequency of the individual active devices. Today's transistor technology yields gains to frequencies of 200 GHz, and monolithic distributed amplifier results have been quoted to 100 GHz. Distributed amplifier designs are used heavily in the 6–18-GHz electronic warfare applications (Fig. 19).

Substantial initial fabrication and design costs are required for MMIC development. Material-growth and processing-technology capital costs can be in the \$10 M range for even a modest operation. Approximately 20 GaAs foundries exist in the United States and Europe, offering production capabilities in MESFETs, HEMTs, and HBTs. A new MMIC realization with via holes and 0.5-μm feature size costs between \$50 K and \$100 K and requires three to six months per design iteration. Because the circuit cannot be tuned once fabricated, and because of the high cost of a single foundry run, stability and accuracy of every circuit fabrication step are imperative.

The growth of semiconductor foundries has been paralleled by the increase in speed and complexity of computer-aided design (CAD) programs, providing linear and nonlinear

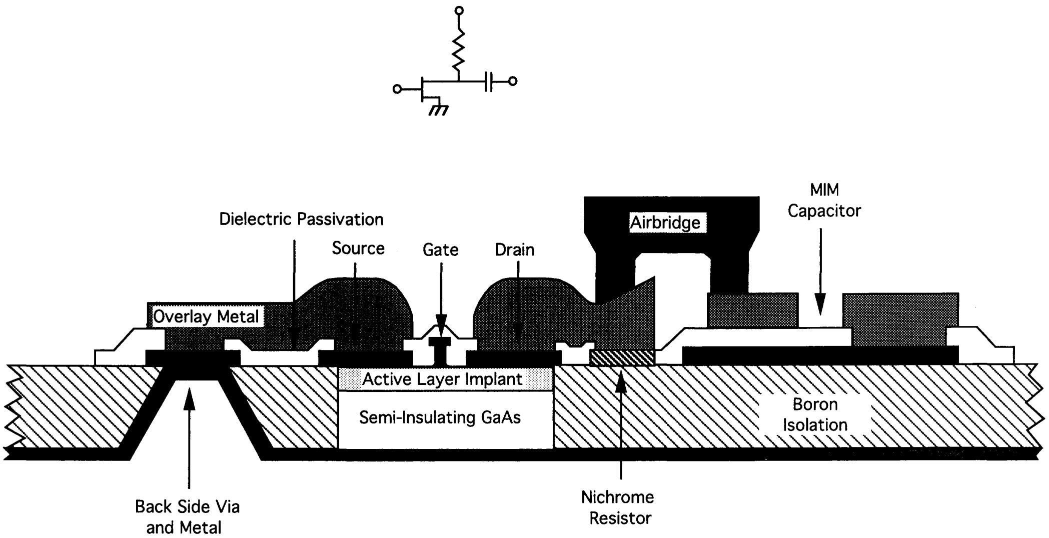

FIG. 18. Gallium arsenide MMIC ion-implanted process (cross-sectional view of processed wafer).

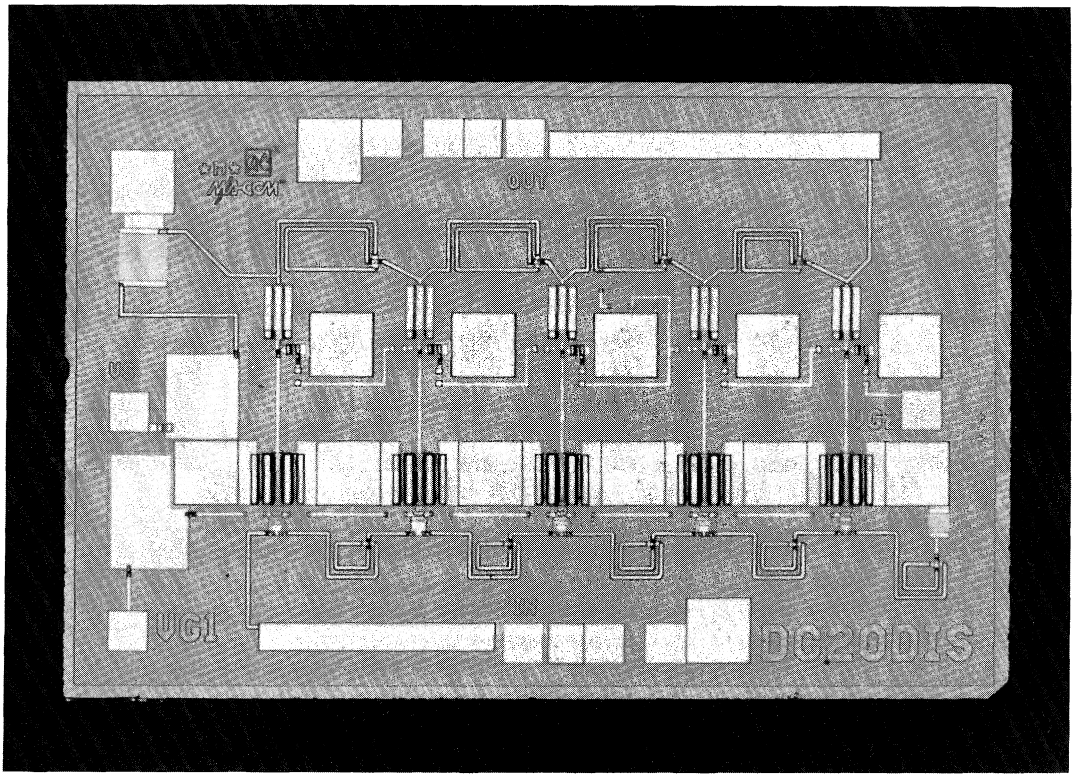

FIG. 19. 2–20 GHz MMIC distributed amplifier; chip size, 1.0 × 1.5 mm.^{2} (Courtesy of M/A-COM, Inc., Lowell, MA.)

analysis, synthesis, and layout of the MMIC circuitry (see COMPUTER-AIDED DESIGN IN ELECTRONICS). A typical "mid-range" [40 million floating point operations per second (Mflops)] workstation with these design capabilities costs approximately $50 K, although prices on hardware and software are dropping. A "high-end" personal computer (1.5 Mflops) with linear analysis/optimization software can be obtained for around 1/20 the cost of a full workstation. These improvements in processing speed allow detailed electromagnetic field simulations of complex two- and three-dimensional geometries such as transitions between different transmission media, multiline coupling, and package cavities, which have no models in present circuit libraries. As these tools grow in accuracy, circuit realization will consist of design-simulate-analyze iterations using CAD tools, followed by a single circuit-fabrication operation, instead of design-build-measure iterations, the present costly and time-consuming practice.

4.3 Integrated Subassemblies: Selecting from the Menu of Technologies

For high-volume commercial applications, MMIC technology provides a method for producing quantities of millions per year at prices in the $1–10 range. A complete direct broadcast satellite (DBS) receiver down-converter, for example, consisting of a rf amplifier, oscillator, mixer, and i.f. amplifier, occupies approximately 1 mm^2 of area, allowing inexpensive production of this circuit function.

For applications where quantities are lower (1000–20 000 parts) but performance is extremely important, and where circuit repeatability is a performance driver, higher-performance, generic single-function MMICs may be profitably used as building blocks in a larger, more labor-intensive subassembly. Multichannel wideband receivers where phase and amplitude of signals over wide bandwidths must be identical among channels are examples of multifunction assemblies where

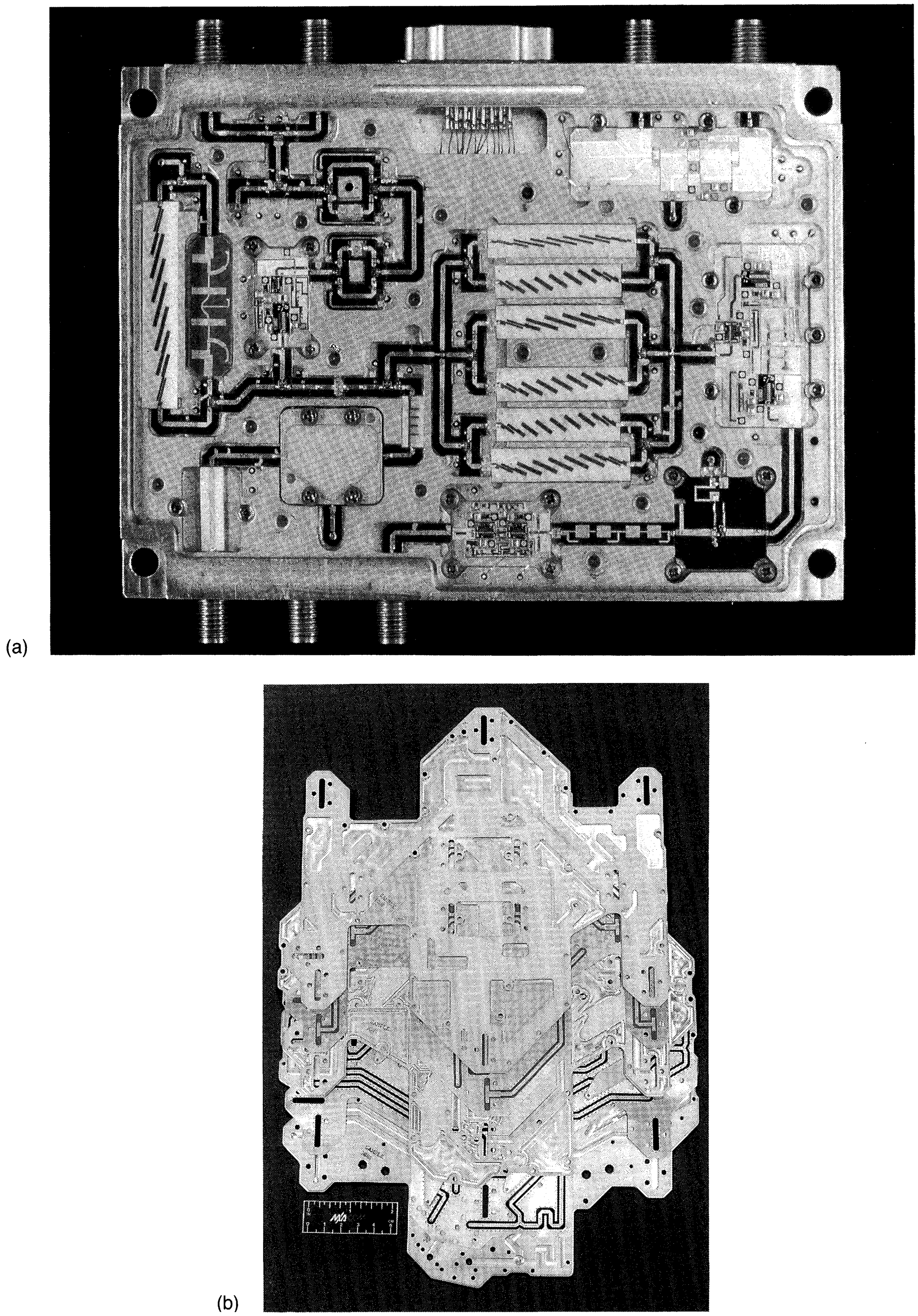

FIG. 20. Examples of microwave subassembly integration techniques: (a) Multichannel microwave receiver using combination of MIC and MMIC techniques; module size, 15 × 20 × 2.5 cm. (Courtesy of M/A-COM, Inc., Lowell, MA.) (b) X-band antenna element signal-distribution network using waveguide and suspended-stripline techniques; module size, 25 × 35 × 1 cm^3. (Courtesy of M/A-COM, Inc., Lowell, MA.)

circuit tuning is still required to achieve performance specifications; MMIC here is used for low-cost single-function components [Fig. 20(a)].

For other low-quantity applications with extremely low loss requirements, an integration method is used consisting of milled aluminum plates that combine waveguide, suspended stripline, microstrip, and other transmission media. Machining is done with fully automatic milling equipment to lower manufacturing costs [Fig. 20(b)].

4.4 Future Developments

For commercial applications, cost reduction will drive technology development for circuit integration. Monolithic microwave integrated circuit technology will require circuit size reduction, process automation, and manufacturing yield improvements to compete with mature MIC techniques. The cost requirements of a functionally complicated commercial assembly like a cellular telephone rf section will not be met by a single multifunction monolithic circuit. Microwave integrated circuit techniques, which have the cost advantage of inexpensive and low-loss substrates over MMIC technology, are easily adaptable to automated assembly techniques to chip sizes as small as 400 μm square, and thus allow selection of semiconductor chips or MMICs from many processes. This multichip module (MCM) integration philosophy has the advantage that device and circuit innovation can be transferred to production very quickly. Size reduction at the MCM level will use multilayer techniques to increase circuit density and allow process-oriented, chip-connect, multilayer hybrid techniques. Silicon will continue to compete with GaAs, with Si holding the edge on process costs (12-in. Si wafers are routinely being processed for digital memory applications, compared to 6-in. GaAs fabrication lines still in development.). Gallium arsenide, on the other hand, claims the advantage on performance, an assertion that becomes clear at frequencies above 5 GHz.

5. SUMMARY

Microwave circuits comprise concepts such as distributed effects and device–circuit interactions that are constant with time; because the circuit functions and component names are also time invariant, one might incorrectly conclude that the discipline is static. In fact, growing commercial markets for microwave functions and the need for new communications spectra at millimeter-wave frequencies are driving forces behind the remarkable changes in *implementation* and *synthesis* in microwave circuits. These changes are increasingly focused toward semiconductor device and process development, which includes the two-dimensional world of MMIC, but will be extended to the third dimension with multilayer techniques as the requirements for increased functional density accelerate. Methods of defining new semiconductor structures with (vertical) growth control at the molecular level will continue to change the manner in which performance limits in power, noise, and cost are defined. Improvements in design tools represent a technology overlay on the future of microwave circuits, as do developments in enabling technologies such as superconductivity, and the related fields of digital electronics and optoelectronics.

GLOSSARY

Characteristic Impedance: That impedance which, when terminating a transmission line, causes no reflection to result from an incoming wave.

E (H) Plane: That plane in guided-wave media, usually rectangular waveguide, which is parallel to the electric (magnetic) field lines.

FET: See MESFET.

Gate: The control electrode in a field-effect transistor. Voltage applied to the gate controls the current from source to drain electrodes.

Gunn Diode: Device operating on principle of negative bulk resistivity resulting from transfer of electrons in a high-energy (from applied electric field), low-mass (high-velocity) state to low-energy, high-mass (low-velocity) state. With proper circuit matching can oscillate or amplify to frequencies exceeding 100 GHz. Also called TEDs (transferred electron devices).

HBT: Heterojunction bipolar transistor. Extends bipolar (homo)junction transistor (BJT) operation to microwave and millimeter-wave frequencies by using an extremely low-loss base region with material whose energy band structure is different from that of emitter and collector, but whose crystal structure is similar.

HEMT: High–electron-mobility transistor. A field-effect device where the conduction path (channel) between source and drain has no impurities and therefore produces very low loss. The supply of electrons is in an adjacent material layer whose ability to contribute electrons to current is controlled by the gate voltage. Operation frequency is as high as 150 GHz. Also called MODFET (modulation-doped FET), TEGFET (two-dimensional electron gas FET).

Heterojunction: The interface between two materials of different band structure, but similar crystal structure.

Homojunction: The interface between regions of the same material with different impurity concentrations.

Hybrid Junction: A waveguide or transmission-line arrangement with four ports, which, when the ports have reflectionless terminations, has the property that energy entering any one port is transferred to two of the remaining three.

IMPATT Diode: IMPAct-ionization Transit-Time diodes use the high electric fields of reverse breakdown in semiconductor junctions to create electron or hole multiplication. As these free carriers pass through a drift region, the resulting phase shift in current causes negative resistance at the terminals of the diode. Appropriate circuit termination produces oscillation or amplification to frequencies greater than 200 GHz.

LTCC: Low-temperature co-fired ceramic. A ceramic-based manufacturing method of achieving multiple (as many as 20) layers of signal propagation in a single fabrication process.

Maser: Microwave Amplification by Stimulated Emission of Radiation, which defines the operation of a device for amplifying or generating microwave radiation by induced transitions of electrons, atoms, molecules, or ions between two energy levels.

MESFET: Metal Epitaxial Semiconductor Field-Effect Transistors (commonly denoted as FETs in this article) have gate control electrodes that are deposited directly on the semiconducting channel region, distinguished from junction FETs (JFETs) where the gate is formed by a semiconducting junction on the channel, or a MOSFET (metal-oxide field-effect transistor) where the gate metal is formed on an oxide grown on the channel. These devices are used in many microwave signal control and amplification applications.

(M)MIC: (Monolithic) Microwave Integrated Circuit.

Ohmic Contact: A region between two materials where voltage is linearly proportional to current.

On/Off Ratio: Used to describe the signal attenuation ratio in a switch from the transmission or "on" state to the "off" state.

Phase Velocity: The speed at which a constant-phase point travels in a propagating wave.

Quadrature Hybrid: A hybrid junction having the property that a wave leaving one output port is in phase quadrature (90° shifted in phase) with respect to the wave leaving the other output port.

Spectrum Analyzer: A broadband receiver that is capable of measuring amplitude and frequency of multiple signals.

T_e: (Effective) noise temperature. A figure of merit used to describe extremely low-noise operation. When normalized to 290 K, T_e is equal to the noise figure minus unity.

Transition Temperature: That temperature (also called critical temperature) which separates normal conducting regime from superconducting regime in a superconductor. Similar descriptions apply to current and magnetic field.

Varactor Diode: VARiable reACtance diodes, semiconductor junction diodes which are designed to give large (as high as 10:1) capacitance changes with an applied voltage, while maintaining low loss.

Works Cited

Ginzton, E. L., Hewlett, W. R., Jasberg, J. H., Noe, J. D. (1948), *Proc. IRE* **36**, 956–969.

Gorter, C. J. (1964), *Rev. Mod. Phys.* **36**, 1.

Hoffmann, R. K. (1987), *Handbook of Microwave Integrated Circuits*, Norwood, MA: Artech, Chap. 2.

Hopfer, S. (1955), *IRE Trans. Microwave Theory Techniques* **MTT-2**, 20–29.

Lakin, K. M., Kline, G. R., McCarron, K. T. (1993), *1993 IEEE MTT-S International Microwave Symposium Digest*, 1517–1520.

Levy, R., Cohn, S. B. (1984), *IEEE Trans. Microwave Theory Techniques* **MTT-32**, 1055–1067.

Montgomery, C. G., Dicke, R. H., Purcell, E. M. (1948), *Principles of Microwave Circuits*, MIT Rad. Lab. Series, Vol. 8, New York: McGraw-Hill, Chaps. 5, 12.

Trew, R. J., Yan, J.-B., Mook, P. M. (1991), *Proc. IEEE* **79**, 598–620.

Schilling, A., Cantoni, M., Guo, J. D. Ott, H. R. (1993) *Nature,* **363**, 56–58.

Wadell, B. C. (1991), *Transmission Line Design Handbook*, Norwood, MA: Artech. Chap. 4.

Wheeler, H. A. (1984), *IEEE Trans. Microwave Theory Techniques* **MTT-32**, 1008–1021.

Further Reading

Collin, R. E. (1992), *Foundations for Microwave Engineering*. New York: McGraw Hill. This recently revised and expanded text covers a broad scope of optical and microwave topics.

Guerlac, H. E. (1987), *Radar in World War II*, New York: American Institute of Physics/Tomash Publishers. An excellent and comprehensive review of the research activity during World War II in laboratories in England and the United States that served as the foundation for contemporary microwave theory and applications. Henry Guerlac, retained by the MIT Radiation Laboratory to document the events of that organization, was prevented from publishing the work during his lifetime.

Maas, S. A. (1993), *Microwave Mixers*, Norwood, MA: Artech.

Marcuvitz, N. (1951), *Waveguide Handbook*, MIT Radiation Laboratory Series Vol. 10, New York: McGraw-Hill. Many of the 28-volume set of books written at the close of the MIT Radiation Laboratory are classics of microwave theory. Two (Montgomery *et al.*, 1948, and Ragan *et al.*, 1948) are mentioned in Works Cited; Marcuvitz's text is still widely used as the definitive reference on waveguide models.

Matthaei, G. L., Young, L., Jones, E. M. T. (1964), *Microwave Filters, Impedance-Matching Networks and Coupling Structures*. New York: McGraw-Hill, pp. 775–842.

Rizzi, P. A. (1988), *Microwave Engineering*, Englewood Cliffs, NJ: Prentice-Hall. A good general reference for passive components, which includes electromagnetic theory while giving insight into the practical applications.

Rodrigue, G. P. (1988), *Proc. IEEE* **76**, 121–137. This invited paper summarizes the development and status of microwave ferrites and their applications.

Saad, T., special editor (1984), *IEEE Trans. Microwave Theory Techniques* **MTT-32**, 955–1263. This special (centennial) issue of the *IEEE Transactions on Microwave Theory and Techniques* describes in detail the technical contemporary milestones of microwave theory and applications.

White, J. F. (1982), *Microwave Semiconductor Engineering*, New York: Van Nostrand Reinhold. This text covers semiconductor control devices thoroughly.

MICROWAVE TRANSISTORS

See TRANSISTORS, MICROWAVE

MISSILES

See SPACE MISSILES

MODULATORS AND DEMODULATORS, ELECTRICAL

Kohji Hohkawa, *Faculty of Engineering, Kanagawa Institute of Technology, Atsugi, Kanagawa, Japan*

Introduction 379
1. Electrical Modulation Techniques 380
1.1 Amplitude Modulation (AM) 380
1.2 Angle Modulation (Frequency Modulation, FM; Phase Modulation, PM) 380
1.3 Pulse Modulation 381
1.3.1 Analog Pulse Modulation 381
1.3.2 Digital Pulse Modulation 381
1.3.3 Modulation of Bandpass Pulses 382
2. Modulators 382
2.1 Circuits for AM 382
2.2 Circuits for Angle Modulation (FM and PM) 383
2.3 Circuits for Pulse Modulation 383
3. Demodulators 384
3.1 AM Demodulators 384
3.2 Circuits for Angle Demodulation 385
3.3 Circuits for Pulse Demodulation 386
4. Devices for Modulation and Demodulation 386
5. Applications of Modulators and Demodulators 388
6. Conclusions 390
Abbreviations 390
Works Cited 391
Further Reading 391

INTRODUCTION

Today's highly developed information society is supported by a wide range of communication systems, such as telephones, television (TV) and radio broadcasting systems, citizens band radio, computer networks, and special security systems. These systems have been introduced all over the world, and use a variety of transmission media, such as coaxial cables, optical-fiber transmission lines, mobile, satellite, and microwave radio transmission links, and many special-purpose private networks.

Modulation and demodulation are among the main functions of communication systems for transmitting a large amount of information signals economically in real time. When we need to transmit and receive message signals, such as speech, video signals, or digital data, the information must be modified into a form suitable for transmission over the medium. This modification is achieved by means of a process known as *modulation*, which involves varying some parameters of a *carrier wave* (e.g., the amplitude, frequency, or phase of a sinusoidal wave). After transmission, the modulated signal is distorted by several degradation sources, such as noise, interference, and imperfection of frequency characteristics of the transmission medium. The *demodulation* process, which is the reverse of the modulation process, is to create original message signals from the distorted signal at a receiver. These are the primary applications of modulators and demodulators, although they can also be applied to other fields—for example, measurement equipment, sensing systems, and various kinds of consumer electronics.

Electrical modulation and demodulation processes can now be achieved using the high-performance signal-processing capabilities of advanced semiconductor technology. Operating frequencies of tens of gigahertz are currently available, and up to 100 GHz should be available in the future. Modulation and demodulation of signals with higher frequency can be done by means of optical technology. In most optical modulation and de-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

modulation systems, almost all the basic modulation processes, such as forming the modulation wave, extracting baseband signals, coherent signal generation, and precise timing control, are carried out electrically, except for the electrical-to-optical and the optical-to-electrical conversion processes.

This article gives an overview of the most widely used electrical modulation methods, giving examples of actual modulator and demodulator circuits and devices. Some examples of applications in communication systems are also reviewed.

1. ELECTRICAL MODULATION TECHNIQUES

1.1 Amplitude Modulation (AM)

The wave forms produced by a variety of constant-wave modulation methods are illustrated in Fig. 1. The frequency spectrum of an AM signal consists of the carrier signal at the center and two sidebands located symmetrically on both sides of the carrier. AM techniques are categorized in terms of how this spectrum is modified before transmission. DSB (double sideband), in which both sidebands are sent in addition to the carrier wave, has the advantage that it can be demodulated by simple circuitry without the need for carrier recovery, but this technique is inefficient in terms of power dissipation. This method is used for commercial AM radio broadcasts, which have one source and numerous receivers. DSBSC (double sideband, suppressed carrier), in which both sidebands are sent without a carrier wave, is suitable for point-to-point communication involving one transmitter and one receiver. SSB (single sideband), in which only one sideband is sent, has many advantages over the others in terms of power consumption, frequency bandwidth, and noise characteristics against fading. This method is suitable for long-distance transmission of voices over wire links, and is also used in marine radiotelephony. VSB (vestigial sideband) involves sending not just one sideband, but also the low-frequency components of the other sideband near the carrier wave. VSB is suitable for transmitting, for example, TV signals, which contain significant low-frequency components.

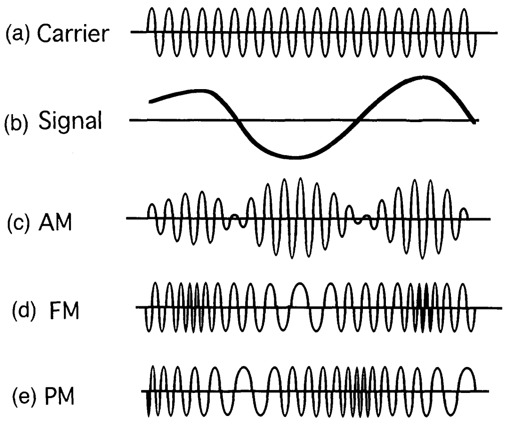

FIG. 1. Comparison of various constant-wave modulation methods. (a) Carrier signal. (b) Baseband information signal. (c) AM, in which the amplitude of a sinusoidal carrier signal is varied in accordance with the baseband information signal. (d) FM, in which the frequency of a carrier wave is modulated continuously in accordance with the baseband signal. (e) PM, in which the phase of a carrier signal is modulated. As shown in the figure, the instantaneous frequency of the FM wave varies in proportion to the baseband signal level, so it has the highest value when the baseband signal is at maximum level. On the other hand, a PM wave has the highest instantaneous frequency when the transition of the baseband signal level is at maximum because its instantaneous frequency varies in proportion to the deviation of the baseband signal levels.

1.2 Angle Modulation (Frequency Modulation, FM; Phase Modulation, PM)

In FM, the frequency of the carrier wave is modulated continuously by the baseband signal, while in PM, the phase of the carrier wave is modulated. In both methods, the amplitude of the carrier wave remains constant. The spectra of the modulated waves have the carrier frequency at the center, with an infinite number of side spectra extending from both sides of the carrier. An important feature of these techniques is that they can provide better discrimination against noise and interference than AM. However, this improvement in performance is achieved at the expense of increased transmission bandwidth. That is, angle modulation provides us with a practical means of exchanging transmission bandwidth for improved noise performance. Such a tradeoff is not possible with AM. Angle modulation is extensively used in public communication systems and is also used in special-purpose communication systems, such as VHF-FM marine radio.

1.3 Pulse Modulation

1.3.1 Analog Pulse Modulation In this system, a periodic pulse train is used as the carrier wave, and some characteristic of each pulse is varied continuously according to the sampled value of the baseband signal. Figure 2 illustrates various pulse modulation schemes. Among them, pulse-amplitude modulation (PAM) is the most efficient in terms of power and utilization, and pulse-position modulation (PPM) is more efficient than pulse-width modulation (PWM) because the latter uses long pulses that expend considerable power while bearing no additional information. Most pulse systems require synchronization of the receiver to the transmitter. A simple synchronization method used in many analog pulse modulation systems is the so-called start-stop method, in which a time marker signal is inserted in every frame interval in addition to the message signal. The marker signal is identified by making its amplitude, width, and interval exceed those of PAM, PWM, and PPM signals, respectively.

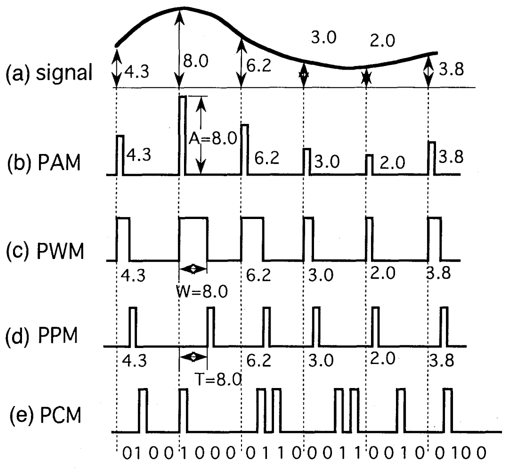

FIG. 2. Comparison of various pulse modulation methods. (a) Continuous-baseband message signal. (b) PAM, in which the amplitude of the regularly spaced rectangular pulses varies with the instantaneous sampled values of the baseband signal. (c) PWM, in which the width of pulses is varied. (d) PPM, in which the position of each pulse is varied according to the baseband signal. (e) PCM. For a comparison of digital pulse modulation technique, PCM is shown. In this technique, the baseband analog signal is converted into digital codes.

1.3.2 Digital Pulse Modulation In pulse-code modulation (PCM), the message is sampled, and the amplitude of each sample is quantized and coded into a series of binary signals, as shown in Fig. 2(e). In the quantization process, the amplitude is rounded off to the nearest integer within a finite set of allowable values, so that both time and amplitude are in discrete form. This process causes quantizing errors in the demodulated signals. Reduction of this error by increasing the number of levels also increases the bandwidth. In actual systems, this problem can be overcome by using nonuniform quantization, in which frequently used signal levels are quantized with more precision than those that are rarely used. For example, the range of voltages covered by voice signals, from the peaks of loud talk to the weak passages of weak talk, is on the order of 1000 to 1. By using a nonuniform quantizer with the feature that the step size increases as the separation from the origin of the input-output amplitude is increased, the large-end step of the quantizer can take care of possible excursions of voice signal into the large-amplitude ranges, which occur relatively frequently. In other words, the weak passages, which need more precision, are favored at the expense of loud passages. In this way, a nearly uniform percentage precision is achieved through the greater part of the amplitude range of the input signal, with the result that fewer steps are needed than would be the case if a uniform quantizer were used. PCM modulation systems are considerably more complex than analog pulse modulation systems and require increased transmission bandwidth. However, they offer the following advantages:

1. ruggedness to transmission noise;
2. efficient regeneration of the coded signal; and
3. the possibility of adopting a uniform format for different kinds of baseband signals (video, speech, and digital-computer data).

Generally, baseband signals, such as speech and video signals, have redundancy in that there exists a high correlation between adjacent samples. Differential PCM (DPCM) is an efficient coding method that reduces the bandwidth by transmitting information about the correlation between the adjacent samples, thereby removing the redundancy. Delta modulation is a one-bit version of DPCM. In this method, the baseband signal is oversampled (i.e., at a rate much higher than the Nyquist rate), and the difference between adjacent signals is quantized to only two levels, namely δ and $-\delta$, corresponding to positive and negative differences, respectively. Recently, the bandwidth required for transmitting video and speech signals by pulse coding has been reduced drastically by using complex signal-processing circuits consisting of advanced very large-scale integrated circuits (VLSIs). For example, in video signals such as TV and HDTV, bandwidth reduction by efficient coding is carried out by reducing redundancies not only between adjacent samples but also between lines, or even frames. In line with the recent development of high-speed semiconductor devices and large-scale integrated circuits, PCM has become the main technology.

1.3.3 Modulation of Bandpass Pulses For transmitting digital signals such as PCM waves or computer signals along a channel with limited bandwidth, the incoming pulses are modulated onto a sinusoidal carrier wave. The modulation process involves switching or "keying" the amplitude, frequency, or phase according to the incoming data. Based on these methods for bandpass pulses are amplitude-shift keying (ASK), frequency-shift keying (FSK), and phase-shift keying (PSK) methods, as illustrated in Fig. 3. Extending these methods, various improved modulation methods have been introduced for transmitting larger quantities of data without errors in a limited bandwidth. For example, in CPFSK (continuous-phase FSK), the duration time of the baseband pulses and the frequencies of carriers used for data 1 and 0 are chosen so as to make the modulated signal continuous at the data transition points. This is because abrupt phase changes in the modulated wave cause a large signal distortion at the transition points and spectral spreading when the wave passes through nonlinear elements. In quadrature-PSK (QPSK), the carrier takes on one of four possible phases relative to the carrier wave, such as $\pi/4$, $3\pi/4$, $5\pi/4$, or $7\pi/4$. The advantage of QPSK is that twice as much information compared to binary-PSK (BPSK) can be transmitted with the same bandwidth because the modulated signal of the QPSK is the superposition of two BPSK signals with the same bandwidth. Generally, FSK signals require a larger bandwidth than PSK signals. However, the frequency deviation between carrier signals can be reduced by using coherent detection (in which the reference carrier signal for detection has the same phase and the same frequency as the carrier signal of the modulated wave), achieved at the expense of increased receiver complexity. The so-called MSK (minimum-shift keying) technique makes it possible to transmit information with half the frequency deviation of that with conventional CPFSK (Pasupathy, 1979).

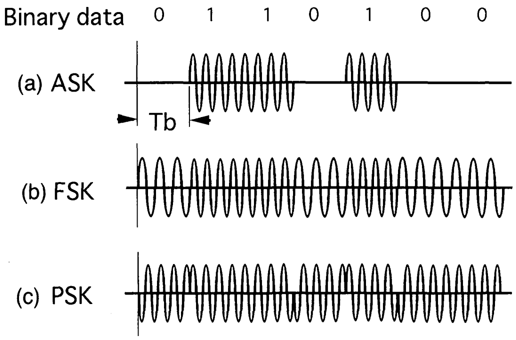

FIG. 3. Bandpass signal waveforms. (a) Amplitude-shift keying, in which the carrier signal's amplitude is modulated by the baseband pulses in the same way as that of DSBSC. (b) Frequency-shift keying, in which the frequency of the carrier is varied according to the baseband pulse. (c) Phase-shift keying, in which the phase of the carrier wave is varied according to the baseband pulse signals.

2. MODULATORS

2.1 Circuits for AM

Ordinary AM signals are modulated by means of a square-wave modulator or switching modulators. A DSBSC wave is generated by product modulators, such as balanced modulators, switching modulators, and ring modulators, as shown in Fig. 4. The balanced modulator uses two AM modulators arranged in a balanced configuration so as to suppress the carrier wave. Several types of product modulators have been implemented

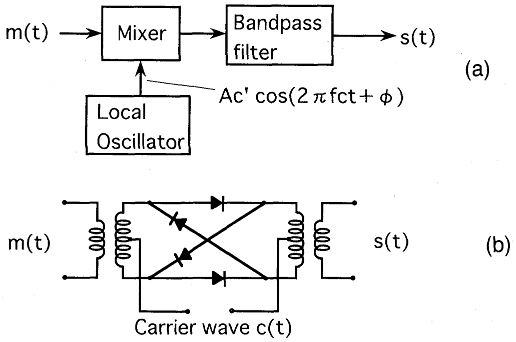

FIG. 4. AM wave modulators. (a) Switching modulator, which is one of the most useful product modulators suited for DSBSC waves. Modulation is done by multiplying the carrier signal with the baseband signal by means of a switching device, such as a diode, transistor, or mixer, and by extracting a bandpass signal with a bandpass filter. This circuit is also used for phase-shift keying technology, discussed in Sec. 2.3: If the baseband signal consists of a bipolar-pulse sequence having amplitude ±1, the modulated signal forms a bandpass-pulse-sequence signal whose amplitude is almost constant and whose phase is switched into 0 and π according to the polarity of the baseband signal. (b) Ring modulator, in which the four diodes form a ring in which they all point the same way, as shown in the figure. These diodes are controlled by a square-wave carrier that is applied by means of two center-tapped transformers. When the carrier supply is positive, the outer diodes are forward biased and have zero impedance, so that the modulator multiplies the baseband signal by +1. When the carrier supply is negative, the situation is reversed, and the modulator multiplies the baseband signal by −1.

in integrated circuits and are now in practical use. SSB and VSB waves are both generated by use of a product modulator followed by a bandpass filter that extracts the desired signal spectrum from the DSBSC wave.

2.2 Circuits for Angle Modulation (FM and PM)

A frequency modulator and a phase modulator are basically the same circuit because an instantaneous frequency is equivalent to the time differentiation of a phase; if we connect an integrator in front of a phase modulator, it operates as a frequency modulator, and if we connect a differentiator in front of a frequency modulator, it operates as a phase modulator. An indirect frequency modulator generates FM waves by employing PM. One method for generating PM waves is a vector synthesis method, in which the PM signal is generated by the vector sum of a carrier signal and an AM signal that has the same frequency as the carrier but a different phase. In a direct frequency modulator, the instantaneous frequency of the carrier is varied directly by the baseband signal by means of a device known as a voltage-controlled oscillator (VCO). One way of implementing a VCO is to use a sinusoidal oscillator with a relatively high-*Q* frequency-determining network, and to control the oscillator by symmetrical incremental variation of the reactive components. A voltage-variable capacitor consisting of a reverse-biased *p*-*n*-junction diode is commonly used for the variable-reactance component. In order to generate wideband frequency modulation with a required frequency deviation, we may use a configuration that consists of a voltage-controlled oscillator followed by a series of frequency multipliers and mixers, as shown in Fig. 5. This permits the attainment of good oscillator stability, constant proportionality between output frequency change and input voltage change, and the necessary bandwidth for wideband FM.

2.3 Circuits for Pulse Modulation

PAM waves are generated by switching the baseband signal with short-duration pulses at the Nyquist interval. PWM waves are generated by using a sawtooth sweeper and a slicer; the sum of the baseband signal and the sweep wave is sliced at a suitable signal level by the slicer. PPM waves are generated from PWM by simply using a monostable multivibrator that is designed to trigger on the

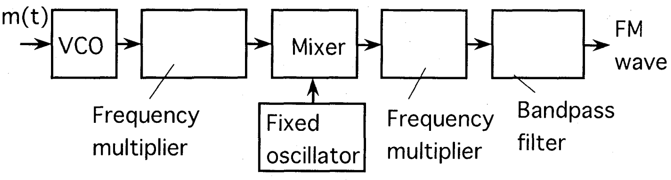

FIG. 5. Wideband FM modulator.

trailing edge of each width-modulated pulse (assuming that the time of occurrence of the trailing edge of each width-modulated pulse is varied in accordance with the message signal), as shown in Hakin (1983).

The process of PCM boils down to analog-to-digital (A/D) conversion. This involves quantizing the amplitude of a signal to a number of discrete levels by sampling the baseband analog signal at regular time intervals, and then generating a digital code consisting of 1's and 0's aligned to each time interval. Figure 6 illustrates an example of an A/D converter for PCM. Another A/D converter is obtained by replacing the sense amplifiers with parallel comparators and exclusive-OR gates. These circuits require a number of parallel amplifiers. Such circuits are complicated but have the greatest speed because all the existing outputs are capable of generating the required code simultaneously. Several types of A/D conversion circuits have been developed and are in practical use, as previously reported (Glasford, 1986). By using a counter-type converter, such as a counter-ramp A/D converter or a dual-slope counter-ramp converter, a simple circuit configuration can be achieved with one or two comparators, but such circuits require large processing time. Another type of A/D conversion extensively used employs a D/A converter in the feedback loop.

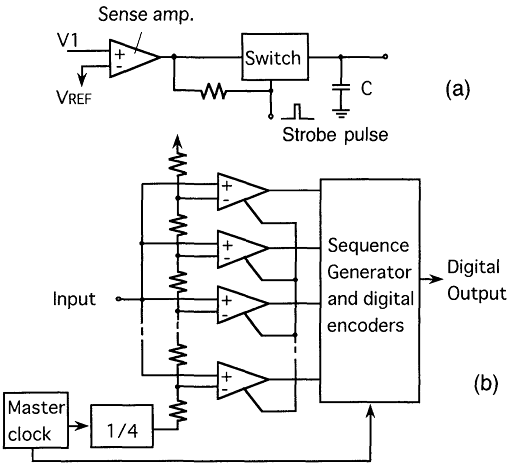

FIG. 6. PCM modulators. (a) A sample-and-hold sense amplifier. The sense amplifier outputs a high-level signal if the input signal is between the lower and upper threshold levels for a certain quantum window during the strobe pulse. (b) PCM wave modulator using it. The sense amplifier outputs a high level when the input signal is within the window of a quantized signal level. The sequence generator and encoder outputs PCM waves.

Some shift-keying techniques bear a close resemblance to DSBSC modulation, except that the input baseband signal consists of binary digits. Using a product modulator, shown in Fig. 4(a), ASK or PSK is achieved by modulating the carrier with baseband pulses of amplitude 0 and 1, or 1 and -1, respectively. On the other hand, FSK signals are produced using two ASK modulators with different carrier frequencies. The data input to one modulator are inverted before being input to the other, and the two inputs are summed together. In QPSK, the modulation is done by using two BPSK modulators, as shown in Fig. 7.

3. DEMODULATORS

3.1 AM Demodulators

A simple yet highly effective device known as an envelope detector can be used for narrow-band AM signals in which the signal bandwidth is small compared to the carrier frequencies, and the modulation depth is less than 100%. Alternatively, AM waves can be detected by a square-law detector, which uses the square-law modulator for the purpose of demodulation.

The baseband signal can be uniquely recovered from a DSBSC wave by using coherent detection or synchronous detection, as shown in Fig. 8(a). In this method, the phase error between the reference local wave and the time-varying signal carrier wave generated in the modulation process causes distortion of the demodulated signal. Therefore, the receivers must include circuitry to maintain the local oscillator in perfect synchronization, as previously mentioned. One method is to use the Costas receiver (Costas, 1965) shown in Fig. 8(b). Another method is to use a squaring loop, in which a sinusoidal wave with twice the carrier frequency, $2f_c$, is generated by a phase-locked loop that follows the $2f_c$ component produced by squaring the original DSBSC. The local wave is synchronized to the carrier wave by dividing its frequency by 2. SSB signals are also demodulated by synchronous detection. However, errors in the frequency and phase of the

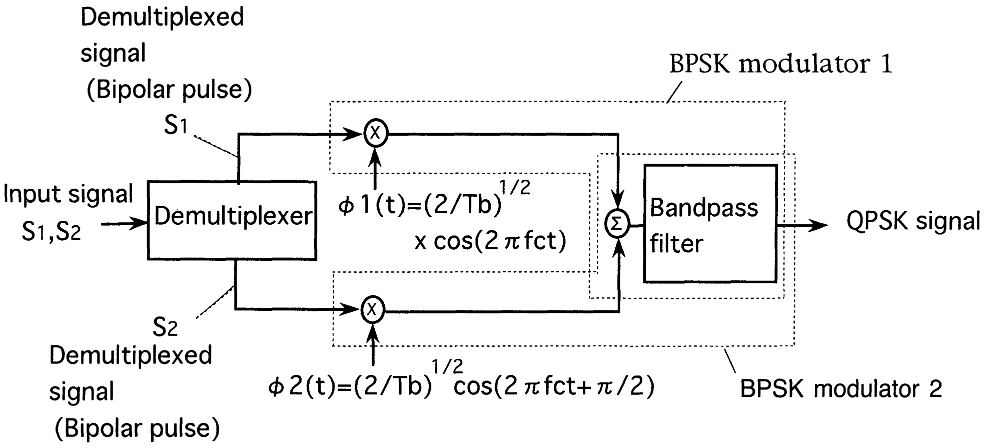
Demultiplexed signal (Bipolar pulse)
S1
Input signal S1,S2
Demultiplexer
S2
Demultiplexed signal (Bipolar pulse)
BPSK modulator 1
$\phi 1(t)=(2/Tb)^{1/2} \times \cos(2\pi fct)$
$\phi 2(t)=(2/Tb)^{1/2} \cos(2\pi fct+\pi/2)$
Bandpass filter
QPSK signal
BPSK modulator 2

FIG. 7. QPSK modulator. The binary data are converted to two-bit bipolar symbols, which are then used to phase-modulate the carrier. The modulation is done by using two BPSK modulators with a common oscillator that produces two different carrier signals with a phase shift of $\pi/2$. The two bits of each symbol are fed to the modulators and phase-modulate the carrier waves. QPSK wave is formed by combining the two BPSK waves.

local carrier cause large distortion of the demodulated waves. Perfect synchronization can be achieved either by transmitting a pilot signal with the selected sideband or by using a highly stable local oscillator.

3.2 Circuits for Angle Demodulation

The demodulation of angle-modulated signals basically involves producing a voltage whose instantaneous amplitude is directly proportional to the instantaneous frequency offset of the input signal. One common technique is to use a frequency discrimination circuit, which consists of a differentiator circuit followed by an envelope detector. Another technique is to use a phase-locked loop, as shown in Fig. 9, which is a negative-feedback system consisting of three major components: a multiplier, a loop filter, and a voltage-controlled oscillator. The circuit is designed so that the VCO control signal corresponds to the baseband signal when the VCO oscillation frequency is tracked to the

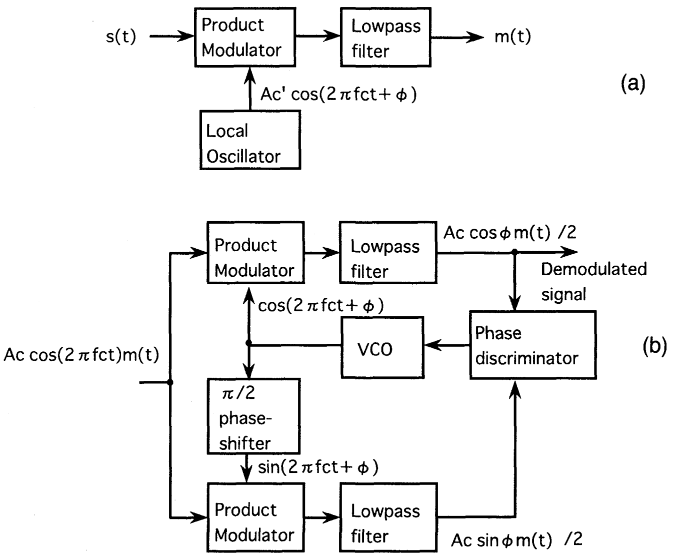
s(t)
Product Modulator
Lowpass filter
m(t)
Ac' cos(2πfct+ϕ)
Local Oscillator
(a)
Ac cos(2πfct)m(t)
Product Modulator
Lowpass filter
Ac cosϕm(t) /2
Demodulated signal
cos(2πfct+ϕ)
VCO
Phase discriminator
π/2 phase-shifter
sin(2πfct+ϕ)
Product Modulator
Lowpass filter
Ac sinϕm(t) /2
(b)

FIG. 8. AM demodulators. (a) Product modulator. (b) Costas receiver. This consists of two coherent detectors with individual local oscillators that are in phase quadrature to each other. These two detectors are coupled together to form a negative-feedback system designed in such a way as to synchronize the local oscillator with the carrier wave.

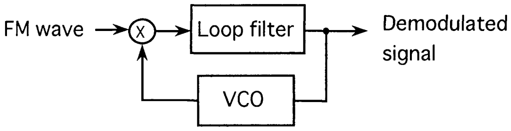

FIG. 9. FM demodulator.

incoming FM wave. As in the modulation process, a PM wave is demodulated using an FM demodulator by simply attaching an integration circuit.

3.3 Circuits for Pulse Demodulation

Since PAM and PWM signals include the baseband signal in their frequency spectrum, they can be demodulated by a simple low-pass filtering. PPM signals are demodulated by converting them to the PWM and applying a low-pass filter.

The demodulation of PCM signals involves a digital-to-analog (D/A) conversion. D/A conversion is inherently simpler and consumes less hardware than A/D conversion, and can be achieved using a resistive network, switches, and a high-gain operational amplifier used as a summing amplifier for improved stability. An example of a resistive-network D/A converter is shown in Fig. 10. In this configuration, a wide range of resistance values is required when there is a large number of bits to convert. This can cause problems in integrated implementations. However, several kinds of resistor networks, free from this problem have been introduced so far. For example, the *R*-2*R* ladder network and the inverted-ladder network D/A converters are implemented by using only two kinds of resistance values, as previously reported (Glasford, 1986). Switching is implemented in a variety of ways using various forms of clamp circuit.

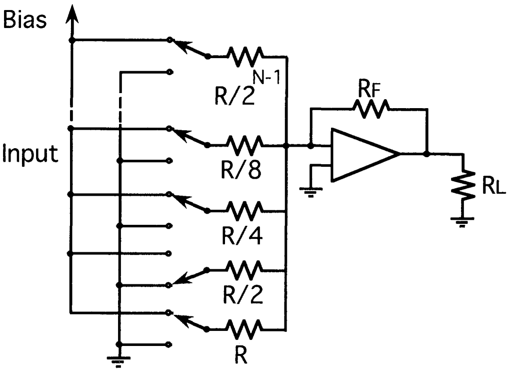

FIG. 10. D/A converter for PCM demodulation. As is shown, the resistors are weighted so as to generate signals with quantized levels. In this configuration, the signal is decoded by using the amplifier to sum up the weighted voltage levels corresponding to each bit in the digital code.

At the receiver, the input PSK signals are bandpass filtered to produce raised cosine pulses and limit the input noise. The signal is then demodulated in the same way as DSBSC signals. Most systems use coherent detection, as shown in Fig. 8(a), in which incoming shift-keyed signals are multiplied by a signal that is synchronized to the original carrier waves. The baseband binary signal is extracted by feeding the multiplied signal to a low-pass filter. A coherent BPSK uses a carrier recovery circuit, such as a squaring loop, in which there is an ambiguity of π in the phase of the reference carrier wave. To overcome this, a known binary signal is transmitted as part of the reference signal. By comparing the demodulated reference sequence with a stored version of the known sequence, the ambiguity can be resolved. The receiver extracts a clock signal of binary data, and this clock signal is used as a timing reference for recognizing the data in a decision circuit. As shown in Fig. 11, a QPSK demodulator consists of two BPSK demodulators.

Another demodulation method for PSK signals is differential detection, which detects pulses by comparing the incoming datum with that received one bit earlier. This method does not require coherent detection.

4. DEVICES FOR MODULATION AND DEMODULATION

The rapid progress of semiconductor technology over the last few decades has resulted in great advances in information-processing systems and communication systems. Communication systems are not only less expensive but also highly integrated, are capable of highly complex signal-processing capability, and operate at much higher frequencies. Even in high-frequency radio communication systems, vacuum tubes have been replaced by highly developed semiconductor elements or integrated circuits, except for

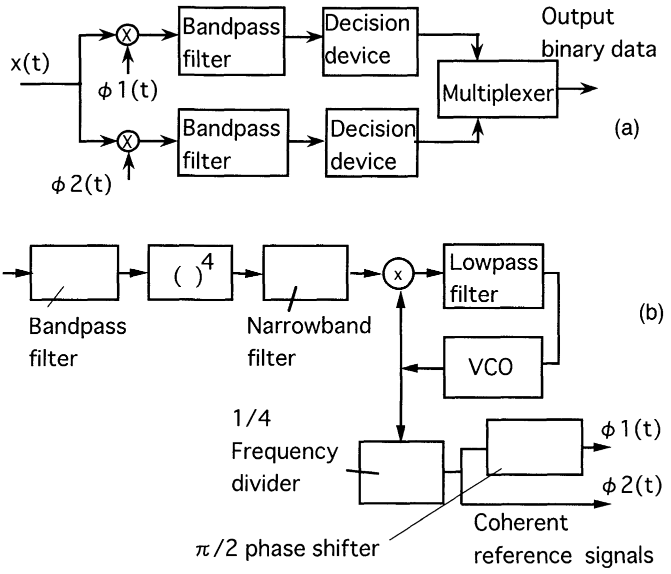

FIG. 11. QPSK demodulation circuit. (a) Demodulator. Data d_1 and d_2, which are demultiplexed in the modulator, are recovered by feeding the coherent reference carrier and $\pi/2$ phase-shifted signal to each BPSK demodulator. (b) Carrier extraction circuit. The reference carrier is synchronized using fourth-power harmonic generator, a phase-locked loop oscillator, and a 1/4 frequency divider, whose circuit structure and operation are similar to that of the square loop.

special-purpose high-frequency high-power amplifiers, such as traveling-wave tubes (TWTs) and klystrons.

As shown in various circuit examples, the main devices used for modulators and demodulators are nonlinear elements, amplifiers, filters, oscillators, delay elements, and logic devices. Many modulator and demodulator integrated circuits (ICs) have been developed and are in practical use. Examples include standard ICs, such as amplifiers, phase-locked loop oscillators, synthesizers, and switches, and many custom ICs for modulation and demodulation in radio and TV receivers, wireless telephones, and citizen's-band transceivers. For digital pulse signals, a variety of coders and decoders and A/D and D/A converters have been developed using monolithic ICs containing analog and digital devices. At microwave frequencies, various kinds of monolithic microwave ICs have been developed using GaAs technology.

Among devices required for modulation and demodulation, high-*Q* circuit elements, such as filters and stable oscillators, are the devices that are hard to implement with semiconductor technology. Filters with a high *Q* or a sharp cutoff are made with electromechanical devices, such as bulk-wave crystals and surface acoustic-wave (SAW) devices, as shown in Fig. 12. Crystal filters, such as energy-confined multiple-mode–coupled filters, are extensively used—for example, as channel filters and pilot signal filters in radio communication systems and SSB frequency-division multiplexing systems. However, their upper frequency is limited to 100 MHz. SAW filters, including resonators and delay-line type filters whose operation is the same as that of transversal filters,* have been extensively used in the frequency range from 10 MHz to several GHz. SAW filters have the advantage that they can be mass produced using advanced semiconductor fabrication technologies, enabling small, high-performance filters to be produced at a low cost. In spite of a relatively large insertion loss, SAW devices are suitable for analog processing of bandpass signals because of their transversal filter nature, in which both amplitude and phase characteristics are controlled (designed) independently. Various

---

*A transversal filter consists of a delay line with multiple weighted taps. The input signal is fed to the tapped delay line, and the output signal is obtained as a sum of the tapped signals. The delay time and weighting value of each tap are designed so as to correspond to the sampling time and sampled value, respectively, of the impulse response of the required filter. The filter can be designed directly in the time domain by simply estimating its impulse response characteristic. It presents a superior design flexibility; for example, we can design a filter with an arbitrary phase and amplitude characteristic. This filter is applied to various kinds of signal processing circuits, such as an equalizer, a matched filter, or a dispersive delay device.

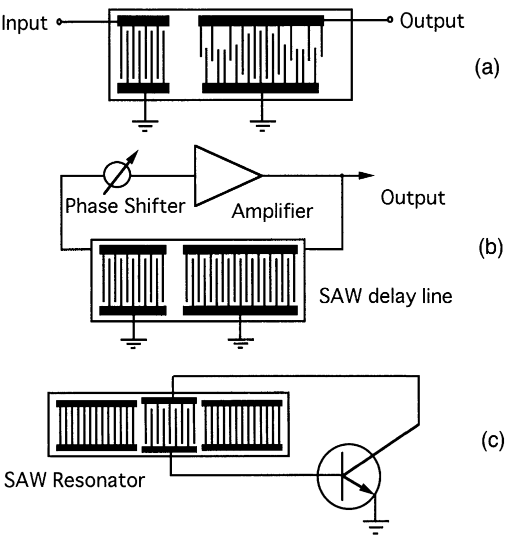

FIG. 12. SAW devices. Schematic diagram of (a) SAW transversal filter, (b) a delay line oscillator, and (c) a resonator oscillator.

types of functional circuits have been developed, including modulators and demodulators of AM, FM, and bandpass pulse signals. At the frequencies of microwaves and millimeter waves, dielectric resonators and microwave components such as microwave cavities, wave-guide filters, and strip-line filters are used.

An oscillator consists of an amplifier with a positive feedback loop containing reactance elements. Quartz oscillators, which use a steep reactance variation with frequency of the crystal resonator, are extensively used in modulators and demodulators as a source of carrier waves and reference waves to stabilize other oscillators such as VCOs by means of phase-locked loops. Quartz oscillators are stable against temperature variation and aging. For high-frequency applications, harmonics of their outputs are used. A SAW oscillator also has a relatively high performance and can be used at high frequencies. There are two types of oscillator in SAW devices, as shown in Figs. 12(b) and 12(c): resonator oscillators and delay-line oscillators. The former uses a steep reactance variation in the same way as quartz oscillators. The latter consists of a loop with an amplifier and a delay line. The delay line has a bandpass characteristic for which only one component in multiples of $1/T$ is passed with a small loss and fed back to the amplifier, where T is the total delay time of the loop. The delay-line oscillator has the advantage that the oscillation frequency can be varied over a wide range by changing the loop delay.

In the microwave frequency range, Gunn-effect and IMPATT diodes, as described in Liao (1985), with reactance elements have been used for high-power signal sources. Since these oscillators are not inherently stable, phase-locking technologies using harmonics of a stable oscillation source, such as a quartz oscillator, are required for frequency stabilization. Recently, oscillators using high-frequency GaAs transistors have also been applied to this frequency range. Extraordinarily stable oscillators, such as rubidium and cesium beam oscillators, which use the inherent transition frequency of the atom, are employed for the standard carrier signal source in various systems.

5. APPLICATIONS OF MODULATORS AND DEMODULATORS

For public communication applications, such as the telephone network, the bandwidth of the transmission medium is far greater than that required by a single user. A large number of baseband signals are multiplexed before transmission, and demodulation is carried out after demultiplexing. Frequency-division multiplexing (FDM) and time-division multiplexing (TDM) technologies, shown in Fig. 13, are used for analog waves and pulse waves, respectively. For both multiplexing methods, hierarchies are introduced to facilitate signal handling and to simplify instrumentation.

FDM has been used in many conventional communication systems—for example, coaxial cable transmission systems, microwave radio transmission systems, and personal, mobile, and satellite communication systems. The multiplexed signal is either directly transmitted after passing through bandpass filters to reject spurious signals outside the signal band, or amplified for long-distance transmission before filtering. For long-distance transmission, repeaters are used for compensating transmission loss. Two types of repeaters are available. In direct repeaters, amplification is carried out at the high frequency of the transmission band. In indi-

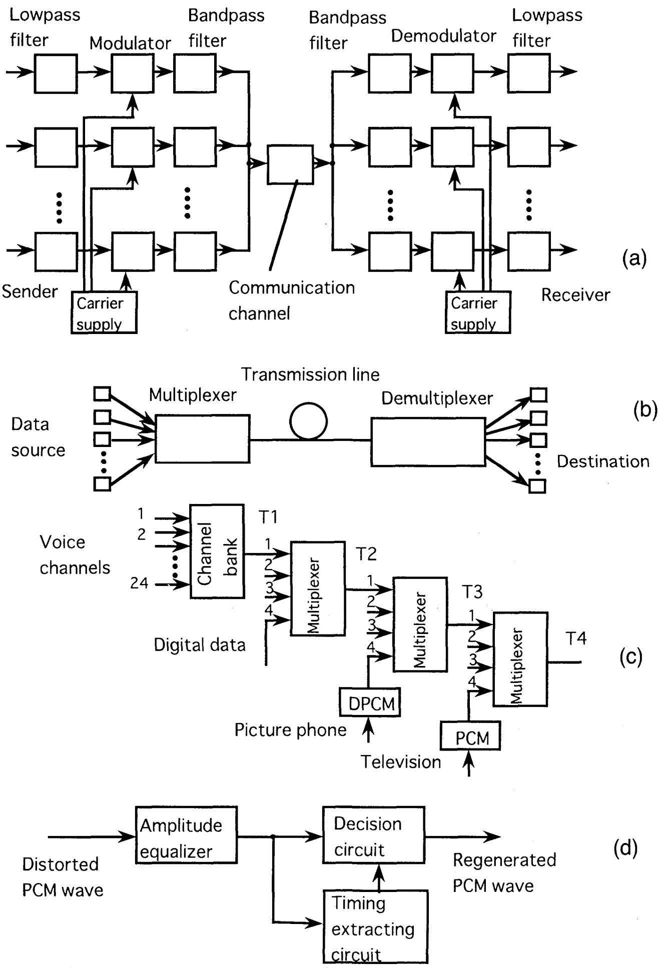
Lowpass filter
Modulator
Bandpass filter
Bandpass filter
Demodulator
Lowpass filter
Sender
Carrier supply
Communication channel
Carrier supply
Receiver
(a)
Transmission line
Multiplexer
Demultiplexer
Data source
Destination
(b)
Voice channels
1
2
24
Channel bank
T1
1
2
3
4
Multiplexer
T2
Digital data
1
2
3
4
Multiplexer
T3
DPCM
Picture phone
1
2
3
4
Multiplexer
T4
PCM
Television
(c)
Distorted PCM wave
Amplitude equalizer
Decision circuit
Timing extracting circuit
Regenerated PCM wave
(d)

FIG. 13. Block diagram of multiplexing and demultiplexing systems. (a) FDM, in which the modulated signals are arranged onto the frequency domain with suitable separation so as to reject interchannel interference. The baseband signal of each channel is directly modulated by cw analog modulation methods, such as FM and SSB, onto a predetermined frequency band with a corresponding carrier frequency. Alternatively, each baseband signal is modulated onto a fixed intermediate frequency, and each modulated signal is modulated onto the predetermined different frequency by a product modulator with a corresponding carrier frequency. In an SSB system, a pilot signal is inserted for synchronization in the multiplexed waves. The demultiplexing process is the reverse of the multiplexing process. (b) TDM, in which modulated digital pulses of each channel are arranged on the time domain. (c) An example of hierarchical structure, in which multiplexing is carried out by classifying in several groups and higher-group signals are composed by multiplexing a number of multiplexed signals in lower groups. (d) Regenerative repeater.

rect repeaters, a heterodyne amplifier is used to convert the received signal to an intermediate frequency, at which low-noise, high-performance amplifiers and filters can be used to compensate phase or delay distortions caused in the transmission lines. The amplified intermediate-frequency signals are then converted back to the transmission band and retransmitted. In microwave systems, each repeater retransmits the signals at a different frequency from that of the received signal, so as to avoid interference.

TDM technology is used for the integrated-service digital network (ISDN) and will be used for the coming broadband ISDN (BISDN), which provides various new services and functions and is gradually replacing the conventional analog system. The reliability, flexibility, and ease of implementation provided by large-scale integrated (LSI) technology are not obtainable in analog systems. In most coaxial cable systems, the data is formed into bipolar pulses, such as the alternate mark inversion (AMI) pulse format. In intensity-modulation direct-detection optical-fiber transmission systems, data are formed into nonreturn-to-zero (NRZ) pulses. In radio communication systems, such as personal, mobile, and satellite communication systems operating at frequencies from UHF to the millimeter band, and in coherent optical-fiber systems, bandpass pulse waves are used. In all systems, the data are inserted in a fixed-format container called a frame and then multiplexed. After multiplexing them into a predetermined frame format, the data are transmitted over the line. At repeaters, the distorted and attenuated data are reformed, using a clock extracted from the data, and retransmitted as shown in Fig. 13(d).

In digital systems, such as PCM systems, we need to extract a timing clock and to identify the exact data position from the sequence in order to demodulate the data correctly. After amplifying received waves to a certain level by means of an equalizing amplifier, a timing signal is extracted by feeding the signals to a timing extraction circuit consisting of, for example, a high-Q amplifier and a limiting amplifier. If the input data contain a long sequences of marks or spaces, the data lose their timing signal. In addition, the equalizer amplifier is capacitive coupled and does not pass the component. As a result, the data cannot maintain their exact signal amplitude; for example, a long sequence of high-level pulses in AMI and NRZ signals drifts into the low levels. This effect causes incorrect operation in the decision process. To overcome this problem, bit-sequence-independent (BSI) line-coding and data-scrambling technologies are introduced. BSI line coding is done by feeding extra bits into the data to increase the data transitions between marks and spaces. For example, nMIC inserts one complementary signal to the last bit of every n bits in the data sequence. Data scrambling is done by modulating the data with a pseudorandom-noise binary sequence (PRBS) with a mark-to-space ratio of 1/2 that is generated by a feedback shift register. Modulation is done by obtaining the exclusive-OR of the signal data and the PRBS data. The output signal has a mark-to-space ratio of 1/2. Demodulation is done by the same process: The data are reproduced by obtaining an exclusive-OR of the input data and the same PRBS codes.

The data position is identified by inserting a predetermined sequence of extra pulses into the transmitting data. Usually, the data are inserted into a frame with a predetermined format to each hierarchy, with a frame pulse and other service bits that control the operation of the network. Recently, several systems have been introduced with synchronous digital hierarchy network-node interface (SDH-NNI) and are now in commercial use. SDH-NNI is a worldwide standard for network interfaces and should permit efficient construction of digital networks, including BISDN, and effective network operation, administration, and maintenance, as previously reported (Asatani, 1988). The technology can be applied to coaxial cable networks, to optical-fiber networks, and even to the radio communication systems whose bandwidth is limited.

Electrical modulation and demodulation technologies are also applied to optical-fiber transmission systems. In most optical systems, almost all the basic modulation processes, such as forming the modulation wave, extracting baseband signals, coherent signal generation, and precise timing control, are carried out electrically, except for the electrical-to-optical and the optical-to-electrical conversion processes. The electrically modulated signals are converted to optical signals before transmission, and at the receiver, the received optical signal is converted back to the electrical signal.

6. CONCLUSIONS

The rapid development of semiconductor technologies, especially digital technology, has made drastic changes in signal processing and communication systems. Further development is expected in semiconductor technologies. Recent deep submicron silicon processes make it possible to fabricate ultra-large-scale integrated circuits (ULSIs) with several million gates. A small-scale Si bipolar IC operating at more than 20 Gb/s already has been reported, and circuits operating at several GHz are commercially available. Compound semiconductor devices, such as GaAs MESFETs, HEMTs, and HBTs, operate at higher frequencies, and devices operating around 10 GHz are now available. The operating frequency of GaAs devices is likely to reach 40 GHz in small-scale IC devices. Higher-frequency operation up to 100 GHz should be possible using InGaAs. These developments should increase the operating frequency range of various modulators and demodulators. These device technologies will provide high-frequency communication systems with small size and a low cost. In addition, the increasing sophistication of LSI technologies should make the modulation and demodulation of digital signals more complicated but also more efficient.

ABBREVIATIONS

A/D: Analog to digital.
AM: Amplitude modulation.
AMI: Alternate mark inversion.

ASK: Amplitude-shift keying.
BISDN: Broadband integrated-service digital network.
BPSK: Binary phase-shift keying.
BSI: Bit-sequence-independent.
CPFSK: Continuous-phase frequency-shift keying.
D/A: Digital to analog.
DPCM: Differential pulse-code modulation.
DSB: Double sideband.
DSBSC: Double-sideband suppressed-carrier.
FDM: Frequency division multiplexing.
FM: Frequency modulation.
FSK: Frequency-shift keying.
HBT: Hetero-bipolar transistor.
HDTV: High-definition television.
HEMT: High–electron-mobility transistor.
IC: Integrated circuit.
IMPATT: Impact-ionization avalanche transit time.
ISDN: Integrated-service digital network.
MESFET: Metal-semiconductor gate field-effect transistor.
NNI: Network-node interface.
NRZ: Nonreturn to zero.
PAM: Pulse-amplitude modulation.
PCM: Pulse-code modulation.
PM: Phase modulation.
PPM: Pulse-position modulation.
PRBS: Pseudorandom binary sequence.
PSK: Phase-shift keying.
PWM: Pulse-width modulation.
OPSK: Quadrature phase-shift keying.
SAW: Surface acoustic wave.
SDH: Synchronous digital hierarchy.
SSB: Single sideband.
TDM: Time-division multiplexing.
TWT: Traveling-wave tube.
ULSI: Ultralarge-scale integrated circuit.
VCO: Voltage-controlled oscillator.
VHF: Very high-frequency.
VLSI: Very large-scale integrated circuit.
VSB: Vestigial sideband.

Works Cited

Asatani, K. (1988), in: *Digital Technology Spanning the Universe*, Proceedings of the International Conference on Communication, New York: IEEE, pp. 118–124.

Costas, J. P. (1965), *Proc. IRE* **44**, 1713–1718.

Glasford, G. M. (1986) *Analog Electronic Circuits*, Englewood Cliffs, NJ: Prentice-Hall, chap. 10.

Haykin, S. (1983), *Communication Systems*, New York: Wiley, Sec. 7.7.

Liao, S. Y. (1985), *Microwave Devices and Circuits*, 2nd Ed., Englewood Cliffs, NJ: Prentice-Hall, chaps. 9 and 10.

Psupathy, S. (1979), *IEEE Commun. Soc. Mag.*, **17**, 14–22.

Further Reading

Glasford, G. M. (1986), *Analog Electronic Circuits*, Englewood Cliffs, NJ: Prentice-Hall.

Korn, I. (1985), *Digital Communications*, New York: Van Nostrand Reinhold Electrical/Computer Science and Engineering Series.

Roddy, D. (1989), *Satellite Communications*, Englewood Cliffs, NJ: Prentice-Hall.

Sinnema, M., McGovern, T. (1986), *Digital, Analog and Data Communication*, A Reston Book, Englewood Cliffs, NJ: Prentice-Hall.

MOLECULAR AND ATOMIC CLUSTERS

JOANNA M. HUNTER AND MARTIN F. JARROLD, *Department of Chemistry, Northwestern University, Evanston, Illinois, U.S.A.*

	Introduction	411
1.	**Synthesis of Clusters**	412
1.1	Molecular and Atomic Cluster Beams	412
1.1.1	Supersonic Expansions	412
1.1.2	Oven Sources	412
1.1.3	Laser Vaporization	412
1.1.4	Sputtering	413
1.2	Condensed-Phase Clusters	413
1.2.1	Colloidal Dispersions	413
1.2.2	Capped Clusters	413
1.2.3	Nanocrystalline Materials	414
2.	**Molecular Clusters**	414
2.1	Spectroscopic Methods	414
2.1.1	Electronic Spectroscopy	414
2.1.2	Vibrational Spectroscopy	415
2.1.3	Photoelectron Spectroscopy	415
2.2	Chemical Reaction Dynamics in Clusters	416
2.3	Rare-Gas Clusters	417
2.3.1	Structure	417
2.3.2	Phase Transitions	418
2.3.3	Quantum Clusters	418
3.	**Atomic Clusters**	419
3.1	Alkali-Metal Clusters	419
3.1.1	Electronic Shell Model	419
3.1.2	Collective Electronic Effects	420
3.1.3	Geometric Structure	421
3.1.4	Alkali-Halide Clusters	421
3.2	Group II Elements	422
3.3	Transition Metals	422
3.3.1	Chemical Reactivity	423
3.3.2	Magnetic Deflection	424
3.3.3	Photoionization	425
3.3.4	Collision-Induced Dissociation	426
3.3.5	Nonmetal-to-Metal Transitions	426
3.3.6	Noble Metals	427
3.4	Group III Elements	428
3.4.1	Boron	428
3.4.2	Aluminum	428
3.5	Group IV Elements	428
3.5.1	Carbon	428
3.5.2	Silicon and Germanium	429
4.	**Condensed-Phase Clusters**	430
4.1	Deposited Clusters	430
4.2	High-Resolution Imaging	430
4.3	Quantum Dots	431
4.4	Inorganic Clusters	432
4.5	Nanophase Materials	433
	Glossary	433
	Works Cited	433
	Further Reading	435

INTRODUCTION

The diverse field of molecular and atomic clusters lies at the interface of physics, chemistry, and materials science. A cluster is an aggregate of atoms or molecules, the number of which lies somewhere between the limits of a single atom and bulk matter. The physical and chemical properties of these microscopic aggregates evolve as a function of size. Clusters can be considered the building blocks of bulk matter, and a better understanding of bulk properties may be gained through their study. In addition, clusters have unique characteristics that are distinct from those of the bulk. The physical and chemical properties of clusters often do not vary smoothly with increasing size, and in some cases there are "special" clusters (containing a particular number of atoms) that have exceptional properties—enhanced stability, for example. The main topics of this article are effects due to the finite size of clusters, which can be structural, electronic, thermodynamic, or quantum in nature.

Two main categories of clusters arise from a classification of the bonding interaction between the individual components of the aggregate. We define molecular clusters as aggregates that are generally bound by less than about 0.5 eV/atom, such as by weak van der

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Waals interactions, hydrogen bonding, or ion–molecule interactions. Atomic clusters are typically more strongly bound, having ionic or covalent chemical bonds between component atoms. Although the distinction between these two types of cluster sometimes becomes blurred, it is nevertheless useful to divide the vast amount of information in this way.

Weakly bound molecular clusters are generally not stable at room temperature and pressure. Therefore, special techniques, usually employing molecular beams under high-vacuum conditions, are required to generate and study these species. These aggregates can only be studied during their transient lifetime in the vacuum chamber. Chemically bound atomic clusters are, in principle, more stable, but, because of their extreme reactivity, naked (unprotected) atomic clusters are also usually studied in an interaction-free beam environment. An alternative approach involves isolation of the clusters in an inert matrix such as solid argon at cryogenic temperatures. Passivation of the cluster surface with ligands is another way of circumventing the reactivity problem. Although attaching ligands clearly modifies the electronic and geometric structure, these passivated clusters can be generated in macroscopic quantities and manipulated in air. Both gas-phase (naked) and condensed-phase (ligated) clusters will be discussed in this article, spanning the size range from three to several thousand atoms.

1. SYNTHESIS OF CLUSTERS

There is a plethora of methods for fabricating clusters. These techniques are constantly being improved and tailored to suit the requirements of investigating a specific property. Here we will merely outline the function of the most common types of cluster beam sources, as well as some general methods for condensed-phase cluster synthesis. For further detailed descriptions of the techniques, the reader is referred to the articles cited at the end of the text.

1.1 Molecular and Atomic Cluster Beams

1.1.1 Supersonic Expansions Weakly bound molecular clusters can be generated by the adiabatic expansion of a gas from a high-pressure source through a small orifice into a region of low pressure. Cluster growth is promoted by the cooling that occurs in the expansion and takes place through low-energy monomer–cluster and cluster–cluster collisions. The nascent clusters can reach vibrational and rotational temperatures of as low as a few kelvins, and have their translational energy directed along the beam axis. Cluster-size distributions, internal temperatures, and terminal velocities are dependent on parameters such as pressure, temperature, and nozzle shape. Some experimental control over cluster-size distribution is possible, although this approach always generates a broad distribution of cluster sizes. Because of the high source-pressure requirement, pulsed valves are often used to generate large, weakly bound molecular clusters. Short pulses of gas, 10 μs to 1 ms long, eliminate the need for a large pumping capacity required to maintain low pressures in the instrument (Gentry, 1988).

1.1.2 Oven Sources While simple supersonic expansions can be used to generate many types of molecular clusters, for atomic clusters it is necessary first to volatilize the element. The most straightforward method of volatilizing elements is known as a Knudsen cell, which involves heating bulk material in an oven or crucible and then allowing the vapor to effuse into a region of low pressure. Clusters generated in this way have a broad distribution of velocities and sizes and are internally hotter than those formed by supersonic expansion. Colder clusters can be formed by coexpanding the vapor with an inert carrier gas that collisionally cools the clusters.

1.1.3 Laser Vaporization At the high temperatures necessary to vaporize many bulk materials, oven sources often encounter severe materials limitations. Refractory materials such as tungsten and carbon, for example, are difficult to vaporize by direct heating. A versatile approach that avoids the use of high temperatures is pulsed-laser vaporization, which was developed by Smalley and co-workers (Powers *et al.*, 1982). Focusing a pulsed laser onto a sample transiently heats the material to extremely high temperatures. The resulting plasma is cooled by collisions with a carrier gas and expanded through a nozzle to form an intense pulsed cluster beam containing both neutral and ionic species. Although continuous carrier-gas flows produce more reproducible beams, a pulsed carrier-gas flow is often used to reduce the pumping requirements.

1.1.4 Sputtering Direct sputtering with high-energy ions or neutral atoms is another useful method of producing atomic clusters. The impact of high–kinetic-energy (10–20 keV) particles on a solid surface removes neutral and ionic clusters directly. The intensity of the resulting clusters decreases exponentially with increasing cluster size, and, thus, this approach is generally not suitable for making large clusters. Furthermore, the clusters are internally hot and must be thermalized before subsequent study (Parent and Anderson, 1992).

1.2 Condensed-Phase Clusters

1.2.1 Colloidal Dispersions Small metal particles, of gold in particular, have been prepared in colloidal solutions since the middle of the last century (Faraday, 1857). Typically, colloidal particles can range in size from 4 to 400 nm in diameter. The most general synthesis of colloidal metal particles involves reduction of a metal salt in solution. The size and shape of the resulting colloids depend on the reducing agent, *p*H, and reagent concentration. Thus, it is possible to synthesize selectively particles of a desired average size and shape (Milligan and Morriss, 1964). Protective polymers are sometimes added to prevent coalescence, but because of their inherent instability, colloidal particles have not been isolated from solution. Incorporation of the colloidal particles into solid matrices such as polymers and glasses solves this problem. As it is clearly desirable to prepare a narrow particle size distribution, chromatographic methods such as capillary-zone electrophoresis and size-exclusion chromatography have been applied to reduce the broad size distributions found in colloids.

1.2.2 Capped Clusters An alternative to colloidal particles, which have relatively broad size distributions, is the direct chemical synthesis of ligand-stabilized (also called capped) clusters with well-defined structures. The size range of these compounds extends from nanometers down to several atoms. For example, Schmid has been particularly successful at developing methods to generate ligated metal clusters of specific size (Schmid, 1992). A series of clusters with the general formula $M_{55}L_{12}\mathrm{Cl}_x$ (M = Rh, Ru, Pt, Au; L = PR_3, AsR_3; x = 6, 20, where R = phenyl, *t*-butyl) are formed by reduction of metal salts with B_2H_6 in organic solvents. Numerous other ligated metal clusters have been crystallized, as described in recent reviews (Mingos and Wales, 1990; Schmid, 1992; Lewis, 1993).

Considerable effort has been expended to synthesize semiconductor clusters in which to study quantum confinement effects. The size control attainable here is generally not as great as in the ligand-stabilized metal clusters. One method used to generate capped II-VI semiconductor clusters (e.g., CdSe) is the inverse micelle preparation (Steigerwald *et al.*, 1988). In this procedure, surfactant molecules encapsulate and stabilize growing CdSe clusters to form micelles. When all of the reactant Cd^{2+} and Se^{2-} have been depleted, cluster growth is terminated, thus allowing some control over the cluster size. Treating the clusters with organoselenides, such as $C_6H_5Se^-$, covalently links organic caps to the Cd^{2+}-rich cluster surface. Because this chemically modified surface is hydrophobic, the clusters precipitate as stable entities, which can be collected and subsequently redispersed. Other semiconductor cluster synthesis methods also involve terminating the cluster surface with stabilizing caps. For example, $\mathrm{Cd}_{10}\mathrm{S}_4$ has been capped with thiophenolate groups, and these pyramidal $\mathrm{Cd}_{10}\mathrm{S}_4(\mathrm{SPh})_{16}{}^{4-}$ precursors can then be fused to form stable clusters containing 55 Cd and S atoms (Wang and Herron, 1991). The occurrence of surface defects in clusters prepared by these methods can be reduced by surface passivation with, for example, ammonia, hydroxide ions, amines, and ZnS. Well-defined CdS cluster arrays have also been generated in templates such as zeolites that dictate the cluster size and shape (Herron *et al.*, 1989).

Several recently developed methods of generating well-defined silicon nanocrystals are noteworthy. Efforts in this area are partially driven by the desire to understand luminescence from porous silicon. Thermolysis of SiH_4 in a gas-phase flow tube, followed by oxidation to passivate the cluster surface, generates Si clusters of controlled size (Littau *et al.*, 1993). Also, Heath (1992) has produced nanocrystalline silicon in a high-pressure and -temperature bomb by reducing $SiCl_4$ or $RSiCl_3$ (R = H or octyl) with sodium metal.

1.2.3 Nanocrystalline Materials Consolidation of small particles produces materials consisting of regions, 1–100 nm in diameter, of crystalline structure separated by grain boundaries. Bulk quantities of these nanocrystalline materials can be produced by evaporating the component materials into a vacuum chamber containing an inert gas. The small particles that grow under these conditions are collected on a liquid-nitrogen–cooled surface and then compacted at high pressures to form a nanophase material (Birringer *et al.*, 1984). A more detailed discussion can be found under NANOPHASE MATERIALS.

2. MOLECULAR CLUSTERS

Molecular clusters are weakly bound aggregates of stable molecules. (We include rare-gas atomic clusters here because of the weak interactions that exist between closed-shell atoms.) A variety of different interactions are responsible for binding molecular clusters. These include van der Waals interactions [e.g., Ar_n, $(N_2)_n$], dipole-dipole interactions [e.g., $(CO)_n$, $(NO)_n$], hydrogen bonding [e.g., $(H_2O)_n$, $(NH_3)_n$], and ion-induced dipole interactions [e.g., $(CO_2)_n^+$, Ar_n^+]. The properties of the individual (monomer) components in these systems are only slightly perturbed, and, consequently, the properties of molecular clusters can be considered a composite of the slightly perturbed properties of the monomer and the collective properties of the weakly bound aggregate. There are limited opportunities for exploiting molecular clusters in practical applications. The impetus for studying molecular clusters instead lies in probing the nature of weak interactions between molecules and atoms. These interactions are important in, for example, solvation phenomena and nucleation processes.

2.1 Spectroscopic Methods

A comparison of theoretical calculations and the results of high-resolution spectroscopic studies of cold molecular clusters permits assignment of the ground-state (minimum potential energy) structure and yields valuable information about the role of intermolecular forces. Tracking the evolution of spectroscopic observables as a function of cluster size, together with computer simulations, has led to a deeper understanding of processes such as the structural and energetic changes that occur with stepwise solvation. Most high-resolution spectroscopic techniques couple lasers with supersonic beams and mass spectroscopic detection.

2.1.1 Electronic Spectroscopy Resonance-enhanced multiphoton ionization (REMPI) is an extremely sensitive and versatile method for probing the excited states of molecules and clusters (Brutschy, 1992). Resonant absorption of the first photon to an intermediate electronic state is followed by absorption of a second photon that raises the energy of the molecule to above the ionization threshold. Scanning the excitation laser frequency and monitoring the ion signal results in a resonance-enhanced two-photon ionization (R2PI) spectrum, which is an indirect measurement of the absorption spectrum of the neutral chromophore. Both rotational and vibrational resolution are possible, providing a spectroscopic fingerprint indicative of the molecule and its environment. If the experimental conditions are optimized such that ionization does not lead to evaporation of monomers, mass analysis of the resulting photoions enables these measurements to be performed as a function of cluster size.

A variety of doped molecular clusters, $M \cdot A_n$ (A = rare-gas atom), have been thoroughly studied by R2PI spectroscopy (Whetten and Hahn, 1990). The dopant molecule is predominantly an aromatic chromophore such as benzene, carbazole, or dichloroanthracene (DCA). Solvation of a molecule, M, generally lowers the energy of the first excited electronic (S_1) state with respect to the ground (S_0) state, primarily through stabilizing dispersive interactions. As a result, the $S_0 \rightarrow S_1$ transition shifts to the red. Size-dependent spectral shifts and linewidths of the $S_0 \rightarrow S_1$ transition reflect the extent of solvation and have been interpreted by comparison to molecular-dynamics simulations of cluster structure and spectral features (Jortner, 1992). For small clusters, the spectra exhibit sharp features due to intermolecular van der Waals modes, and dramatic changes occur with each addition of a rare-gas atom. As the size increases, these sharp features give way to broad features that evolve more slowly with size. The point at which this change occurs is attributed to the completion of the first close-packed solvation shell. The number of atoms in the first solvent shell is dependent on the system. A filled shell occurs at $n = 17$ for benzene, $n = 30$ for carbazole, and $n = 36$ for DCA. With increasing cluster size, the spectral shifts and linewidths converge smoothly toward those of the molecule in the bulk matrix.

2.1.2 Vibrational Spectroscopy Vibrational spectra of neutral molecular clusters in supersonic beams can be recorded by a technique developed by Gough *et al.* (1985). The cluster beam is irradiated with photons from an infrared laser, and, if a cluster undergoes vibrational transitions that are in resonance with the laser frequency, then it will absorb energy. The associated increase in temperature is detected when the neutral clusters and dissociation fragments strike a bolometer. This approach has been used to study doped molecular clusters and provides a probe of both the structure and properties of the "host" cluster and the effects of solvation on the "guest" chromophore. Because the species involved are in their ground electronic states, interaction potentials are well understood, and accurate modeling of the system is possible. Studies of SF_6 embedded in rare-gas clusters show that the shift in the frequency of the fundamental ν_3 band of SF_6 can be correlated to the packing density of rare-gas atoms around the SF_6 (Goyal *et al.*, 1993). The shift of this band in small argon clusters is greater than that in solid argon matrices, indicating that the argon atom density around the SF_6 is higher than in the bulk. One possible explanation for this is that the rare-gas atoms are packed in an icosahedral arrangement, where the atoms are more densely packed than in fcc packing. As the cluster size increases, peaks begin to appear at frequencies identifiable as those of bulk argon. Eventually, the fcc features dominate over the icosahedral features, and this crossover is taken to be the structural phase transition. As this method involves no size selection, the spectra necessarily pertain only to an average cluster size. The transition is observed to take place gradually over a broad cluster-size range, and, in clusters of both argon and krypton, it occurs at around 1800 atoms.

Fundamental aspects of ion solvation can be probed by vibrational predissociation spectroscopy of size-selected solvated cluster ions. The interaction between Na^+ and methanol molecules has been probed by monitoring the methanol C–O ν_4 stretching mode in $Na^+(CH_3OH)_n$, $n = 3$–25 (Selegue *et al.*, 1992). Monte Carlo simulations present a picture of the structure of the solvent shell, as well as the kinetics of dissociation. Analysis of the measured spectra indicates that the first solvation shell comprises six methanol molecules. No hydrogen bonding occurs between methanol molecules in the first shell, and the structure is dominated by interactions of the Na^+ with methanol oxygen atoms. The C–O stretching frequency shifts toward that of bulk methanol for $n > 6$, indicating that the ion has less influence on additional methanol molecules, and hydrogen bonding interactions are observed. Figure 1 illustrates the geometry determined for the $n = 10$ cluster by Monte Carlo calculations. In clusters containing more than 15 methanol molecules, hydrogen bonding interactions dominate as in bulk methanol.

2.1.3 Photoelectron Spectroscopy The existence of hydrated electrons in the condensed phase has prompted investigations of electron localization and solvation in gas-phase polar molecular clusters (Barnett *et al.*, 1990). The photodetachment energy (vertical binding energy) of a solvated anion or electron may be determined by photoelectron spectroscopy (PES). Of interest is the minimum cluster size that is necessary to stabilize a bound electron. Theoretical modeling of the $(H_2O)_n^-$ system predicts that, for $11 \leq n \leq 64$, the electron is in a relatively strongly bound surface state. A gradual transition to internal solvation occurs for $32 \leq n \leq 64$. The electron occupies a diffuse, weakly bound surface state in clusters containing fewer than ten water molecules. In agreement with these predictions, measured vertical binding energies suggest that 11 water molecules are required to fully "solvate" an excess electron. In cluster anions containing fewer than 11 molecules, the electron is weakly bound. For

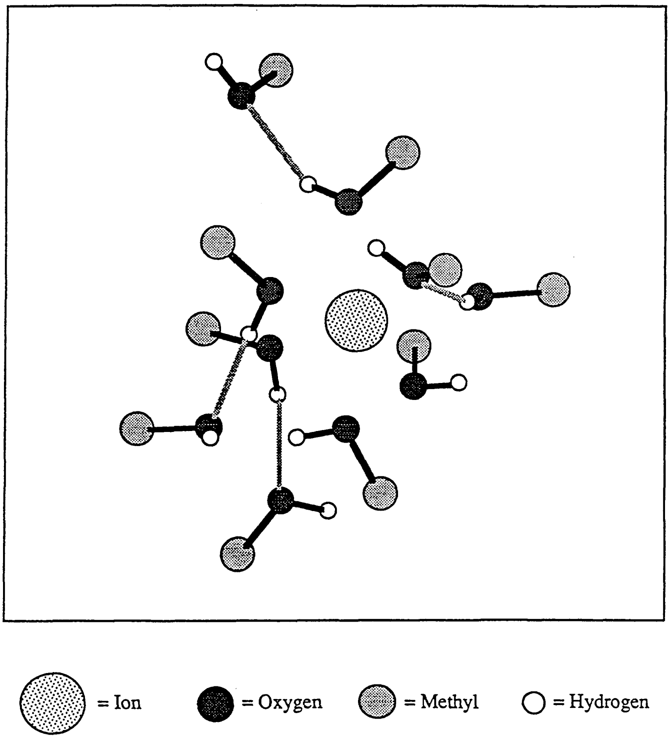

FIG. 1. Monte Carlo simulation of the configuration of $Na^+(CH_3OH)_{10}$ (Selegue *et al.*, 1992). The six methanol molecules that comprise the first solvent shell are arranged about the ion with octahedral symmetry, and hydrogen bonding to the remaining four methanols is indicated by the shaded lines.

$n \geq 11$, the electron is more strongly bound, and the binding energy increases slowly with cluster size.

The electrostatic stabilization of a solvated anion is defined as the difference between photodetachment energy and the electron affinity of the bare ion. A drop in the incremental stabilization energy with the addition of a solvent molecule suggests the existence of a complete solvation layer. Figure 2 plots the evolution of the stabilization energy of I^- as a function of the number of surrounding H_2O molecules (Markovich *et al.*, 1993). Six water molecules form the first solvation layer, as indicated by the large incremental increases in stabilization energy for the first six water molecules and the sudden drop with the addition of the seventh water molecule. The magnitude of the binding energies suggests that the I^- is located near the center of the cluster. As the cluster increases in size to $n = 34$–40, the appearance of small peaks at lower binding energies in the photoelectron spectrum has been interpreted as evidence that there is also a probability that the anion will be located near the surface of the cluster rather than in the interior.

2.2 Chemical Reaction Dynamics in Clusters

In addition to the static investigations discussed above, the fairly recent advances in ultrafast laser techniques are invaluable for probing the detailed dynamics of chemical processes occurring in solvated molecules. Following photoexcitation of the solute mol-

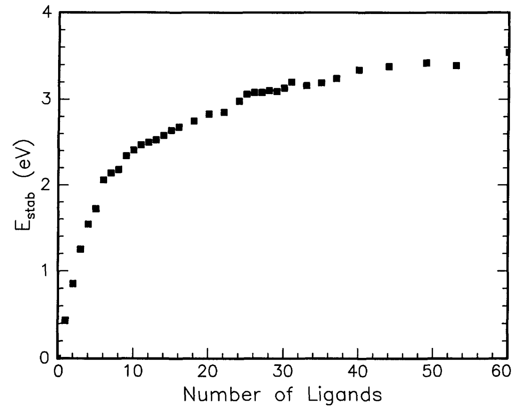

FIG. 2. Evolution of the stabilization energy of I^- in $(H_2O)_n$ as a function of the number of H_2O molecules (Markovich *et al.*, 1993).

ecule, solvent reorganization, proton or electron transfer, or photodissociation of molecules in the solvent cage can occur. These dynamical events can now be interrogated by time-resolved pump-probe spectroscopic methods.

A central issue in understanding charge-transfer processes is the number of solvent molecules that are needed to stabilize an excited-state ion pair. Excited-state proton transfer in phenol–ammonia clusters, $C_6H_5OH \cdot (NH_3)_n$, has been followed by picosecond pump-probe ionization (Syage, 1992). A potential energy diagram illustrating this process for $n = 5$ is given in Fig. 3. A pump pulse excites the phenol (PhOH), which can then undergo proton transfer to the basic "solvent" cluster (B_n) to form the excited-state ion-pair complex $PhO^{*-} \cdots H^+B_n$. If the system is probed before proton transfer occurs, then $PhOH^+B_n$ is detected, whereas ionization of the ion-pair complex results in dissociation to PhO and H^+B_n. Thus, the time evolution of the proton transfer process can be elucidated. Successive addition of NH_3 molecules lowers the energy of the ion-pair state. When this charge-transfer stabilization energy falls below the energy of the S_1 state, excited-state proton transfer will occur. In the phenol–ammonia system, the threshold for excited-state proton transfer occurs at $n = 5$.

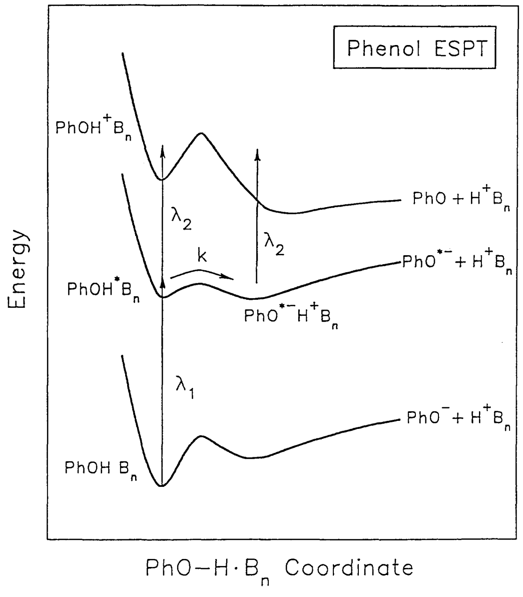

FIG. 3. Potential energy diagram corresponding to $PhOH \cdot (NH_3)_5$, where $B = NH_3$ (Syage, 1992).

Picosecond pump-probe photodissociation spectroscopy of size-selected cluster ions has been used to isolate the dynamics of the solvent cage effect at the molecular level. This is accomplished by photodissociating I_2^-, for example, and then determining the rate of $I\text{-}I^-$ recombination within a $(CO_2)_n$ solvent cage (Papanikolas *et al.*, 1993). By appropriate tuning of the probe laser frequency, these experiments are capable of providing a detailed picture of coherent $I \cdots I^-$ internuclear motion following photoexcitation. Wave-packet motion on a dissociative excited-state potential energy surface has been observed for $n = 14$–22. In this cluster-size range, the mechanism for recombination involves "reflection" of the wave packet off the surrounding solvent cage. Recombination dynamics are qualitatively different for $n \leq 12$, where attractive forces between the CO_2 molecules and the dissociating diatom dominate the recombination process.

2.3 Rare-Gas Clusters

2.3.1 Structure From a theoretical perspective, investigation of bonding interactions and geometric structure in rare-gas clusters is appealing because of the simplicity of the interatomic potential. The rare gases have completely filled valence-electronic shells and spherically symmetric electron density. Hoare and Pal performed a comprehensive search for minimum-energy structures in noble-gas clusters using Lennard-Jones pair potentials (Hoare, 1979). In small clusters, structures based on tetrahedral, pentagonal, and icosahedral symmetry are found to be more stable than those with face-centered-cubic (fcc) symmetry. The icosahedral packing scheme, first introduced by Mackay, appears to be an important packing geometry in small particles (Mackay, 1962). The higher atomic coordination numbers in these densely packed geometries make them energetically favorable over the less-compact fcc packing. The first five complete icosahedral shells contain 13, 55, 147, 309, and 561 atoms, as illustrated in Fig. 4. The fivefold symmetry prohibits extended icosahedral packing without introducing additional strain energy, and, as the cluster size increases, this noncrystal-

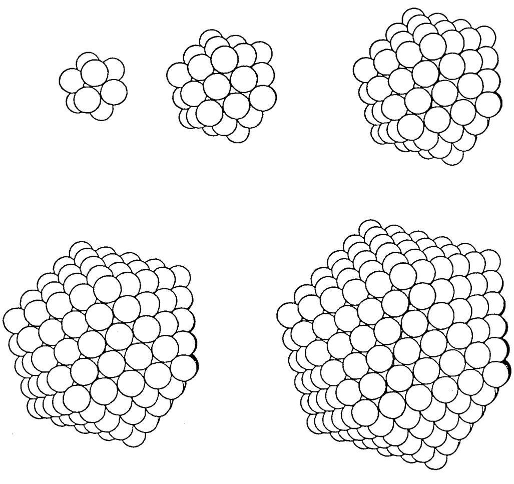

FIG. 4. The first five Mackay icosahedra contain 13, 55, 147, 309, and 561 atoms, respectively.

lographic symmetry gives way to more energetically favorable bulk fcc packing. The number of atoms required for this structural transition to occur is an important quantity because it indicates the emergence of bulk properties. Calculating the cluster size where this transition occurs is fraught with difficulty because the precise number of atoms is sensitive to the interatomic potential, and it varies from 2000 to 10 000, depending on the method of calculation.

Experimentally, electron diffraction is the only method of directly determining the structure of molecular clusters larger than the limit of direct spectroscopic methods. Although a broad cluster-size distribution is sampled, comparison of simulated and experimental diffraction patterns makes possible a structural assignment. The electron-scattering technique has been used extensively by Farges and co-workers (1986), who find that, for clusters of argon, the transition from icosahedral to bulk fcc structure lies at around 800–1000 atoms. Another structural transition occurs in smaller clusters, which adopt less symmetric "polyicosahedral" structures, clusters with interpenetrating 13-atom icosahedra. This type of structure is observed in the size range $30 < n < 60$ atoms. Further evidence for icosahedral structure in rare-gas clusters is found in electron-impact–ionization mass spectra (Miehle *et al.*, 1989). Clusters of Ar, Kr, and Xe with up to 1000 atoms exhibit peaks of local stability at $n = 147, 309$, and 561, all of which are complete icosahedral shells. Intermediate abundance maxima that correspond to partially filled shells are also observed. As discussed in Sec. 2.1.2, addition of atomic or molecular dopants to rare-gas clusters provides another indirect structural probe, although the dopant may perturb the geometry of the cluster.

2.3.2 Phase Transitions Another type of transition, that from rigid (solidlike) to nonrigid (liquidlike), has been extensively investigated. For very small clusters, the definition of "melting temperature" as applied to macroscopic systems no longer applies, and "melting" can be considered the rapid interconversion between many isomers with similar energy. Molecular-dynamics simulations indicate that this type of transition is not always abrupt, and "phase coexistence" has been observed, where both fluid and rigid phases may be in equilibrium (Berry, 1993). This is characterized by an S-shaped bend (van der Waals loop) in the microcanonical caloric curve (T vs E). Although the thermodynamic transitions that occur in small clusters do not necessarily have a solid-state analog, this phenomenon is attributed to a rigid–nonrigid transition, similar to a bulk first-order phase transition. Molecular van der Waals clusters, such as $(SF_6)_n$, $(CH_3OH)_n$, and $(H_2O)_n$, as well as rare-gas clusters, can exhibit this type of behavior. Spectroscopic evidence for a possible nonrigid–rigid phase transition has recently been observed in $(\text{benzene})_{13}$ clusters (Easter *et al.*, 1993). The environment of C_6H_6 in $(C_6H_6)(C_6D_6)_{12}$ has been probed by recording has been probed by recording R2PI spectra as a function of the distance from the nozzle of a supersonic expansion source. In this way, the time evolution of the cluster structure can be obtained. Initially, broad features in the R2PI spectrum suggest a nonrigid state. Several microseconds later, sharp spectroscopic features emerge, suggesting that the cluster has solidified.

2.3.3 Quantum Clusters Helium (and also H_2 and D_2) clusters are appropriate systems in which to investigate quantum effects. Because of the very weak interatomic interactions and relatively large zero-point energy, bulk helium (^4He and ^3He) is a quantum fluid and remains liquid even at 0 K. Consequently, clusters of helium are also expected to be liquid and thus cannot be adequately described using the classical potentials that are used to describe heavier noble-gas clusters. Differences between clusters of ^{4}He and ^{3}He are of interest because ^{4}He is a boson and ^{3}He is a fermion. Calculations predict that because of these statistical effects, all clusters of ^{4}He are bound, whereas, in the ^{3}He system, only clusters with more than about 30 atoms are bound (Stringari, 1990). Experimental information is relatively scarce, a result in part of the difficulty of the experiment and also interpretation of the results. However, efforts have been directed at unraveling their unique properties (Toennies, 1990). Because of the weak binding energies, extensive fragmentation occurs upon ionization of He clusters, making mass spectra difficult to interpret. Although He_2^+ has been observed in mass spectra, it is unclear whether the origin of this ion signal is from fragmentation of larger clusters or from ionization of He_2. However, recent experiments have verified the existence of $^4He_2$ (Luo *et al.*, 1993). There is an extremely long-range interaction between the He atoms, and the bond energy of roughly 1 mK makes this the weakest chemical bond that has been observed.

3. ATOMIC CLUSTERS

Unlike molecular clusters, atomic clusters are strongly bound entities. The chemical bonding between atoms can be either covalent, metallic, or ionic in nature. Because of this strong bonding, addition of a single atom can dramatically alter the properties of the cluster. The organization of this section is roughly along the lines of the periodic table. We begin by considering clusters of the Group I metals, which have been extensively studied because they provide simple model systems in which to examine the emergence of metallic properties.

3.1 Alkali-Metal Clusters

3.1.1 Electronic Shell Model Mass-spectral studies of alkali-metal clusters reveal the presence of particularly abundant clusters containing 8, 20, 40, 58, 92, . . . atoms. The existence of these "magic" numbers can be rationalized by the electronic shell model. According to this model, closed–electronic shell clusters are unusually stable (de Heer *et al.*, 1987). Manifestations of electronic shell structure also appear in size-dependent measurements of the ionization potential, electron affinity, binding energy, static electric polarizability, and other properties (Cohen and Knight, 1990). In the electronic shell model, the jellium approximation (see COLLECTIVE PHENOMENA IN SOLIDS) is used to replace the lattice with a uniformly charged positive background, and the valence electrons are treated as a quantized Fermi gas confined by the cluster surface. The effective one-electron eigenfunctions of a spherically symmetric potential are characterized by a main quantum number, n, and an angular momentum quantum number, l. The electronic energy levels (subshells) are $2(2l + 1)$-fold degenerate. A closed shell occurs when subshells are filled with valence electrons and there is a large energy gap to the next empty level. The ordering of the levels (and, thus, the location of the shell closings) is dependent on the potential. Periodic oscillations resulting from the filling of angular momentum shells are a common physical phenomenon. The filling of electronic angular momentum shells in atoms results in highly stable closed-shell atoms, such as the rare-gas elements, and the nuclear shell model (developed in the 1940s) accounts for particularly stable nuclei.

When the restriction of spherical symmetry is relaxed, spheroidal or ellipsoidal distortions of the jellium background potential partially lift the energy-level degeneracies (Clemenger, 1985). Some levels increase in energy and some decrease, depending on the nature of the distortion. Distortion raises the total energy of clusters with filled subshells, and so these clusters retain spherical symmetry. However, ellipsoidal or spheroidal distortions can lower the total energy of clusters with partially filled subshells, and this gives rise to additional shell structure.

A low-frequency oscillation envelops the higher-frequency shell oscillations, resulting in a "supershell" beating pattern in the electronic level density and total electronic binding energies (Nishioka *et al.*, 1990). Classically, this periodicity results from a superposition of the amplitudes associated with triangular and square valence-electronic orbits. Calculated periodic oscillations in valence-electron energy are indicative of shell and supershell structure. This is plotted in Fig. 5 for the case of spherically symmetric sodium clusters. Electronic shell closings are indicated, and the first supershell node occurs at about 1000 atoms. Agreement of the shell-model predictions with experimentally observed magic numbers provides evidence for electron delocalization. Multiphoton ionization mass spectra of Na cluster beams containing up to 22 000 atoms exhibit abundance maxima at the predicted shell closings (Martin *et al.*, 1991). In addition to electronic structure, abundance distributions in the mass spectrum in Fig. 6 suggest that a transition from filled electronic shells to atomic shells occurs at approximately 1500 atoms. At this point, geometric constraints become dominant over electronic requirements, such that the cluster stability is dictated by closing of atomic shells rather than electronic shells.

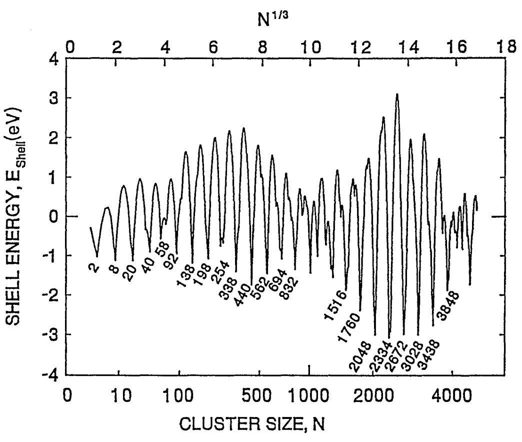

FIG. 5. Calculated valence-electron contribution to the binding energy of spherically symmetric sodium clusters (Nishioka *et al.*, 1990). Magic numbers occur at low energies. Supershell structure is also observed.

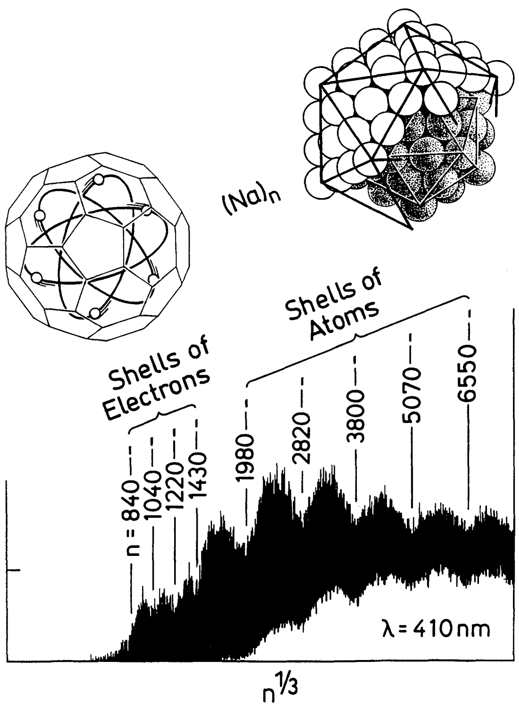

FIG. 6. Photoionization mass spectrum of sodium clusters (Martin *et al.*, 1991). Shells of electrons are observed up to n = 1430, and shells of atoms are observed in larger clusters.

3.1.2 Collective Electronic Effects There has been extensive theoretical and experimental work on the collective electronic behavior of alkali-metal cluster systems (Kresin, 1992). The classical description of the interaction of small metallic spheres with electromagnetic waves was derived by Mie, and extended to prolate and oblate ellipsoidal particles by Gans (Bohren and Huffman, 1983). The response of bulk free-electron metals to electromagnetic radiation is attributed to excitation of collective electronic motion (plasmon oscillations). Both surface and volume plasmons can be observed in small clusters, although surface effects dominate. Generally, the electronic properties of large clusters are well described by the Mie–Gans theory. However, in the smaller clusters, the finite boundaries cause a modification of the electronic oscillation that further red-shifts the resonance frequency with respect to the classical prediction. Deformations from spherical symmetry can cause the single collective electronic absorption to be split into multiple components. For example, open-shell clusters such as Na_{11}^+ (and isoelectronic Na_{10}) exhibit a two-component absorption, which agrees with that predicted for an ellipsoidally distorted cluster.

Experimentally, the collective electronic properties of free alkali-metal clusters have been determined by resonant optical absorption and cross-section measurements (by photodepletion spectroscopy). Because of the fast relaxation of the intermediate electronic excited state in metal clusters, R2PI techniques using nanosecond lasers do not lead to ionization. Under these circumstances, photodepletion provides a method for measuring absorption spectra (Morse *et al.*, 1983). Cluster-ion signals are recorded as a function of the frequency of the excitation laser. Absorption of photons results in photofragmentation and is detected as a decrease in the ion signal. If all of the photoexcited clusters dissociate before reaching the detector, then the photofragmentation cross section is equivalent to the photoabsorption cross section. Photoabsorption spectra for several Na_n are shown in Fig. 7. The single resonance in the spectrum of closed-shell Na_8 is predicted by Mie theory. Other size-dependent resonances are discussed further below.

3.1.3 Geometric Structure For small clusters, quantum-mechanical calculations that explicitly consider the positions of the nuclei give more meaningful descriptions of geometric structure than the structureless jellium approximation. Most *ab initio* quantum-chemical structure calculations have been done on lithium, the simplest metal, but most experimental work has centered on sodium (Bonačić-Koutecký *et al.*, 1991). A combination of theoretical modeling and photodepletion measurements has led to an understanding of the bonding in small alkali-metal clusters. Figure 8 depicts Hartree–Fock optimized geometries and calculated binding energies for neutral and cationic Na_n. Electronic transitions calculated for minimum-energy geometries are correlated to features in absorption spectra such as those of small Na_n in Fig. 7. In this manner, the minimum-energy ground-state geometry of Na_4 has been determined to be a planar rhombus, and Na_5 has a planar C_{2v} structure. The cluster geometries undergo a transition from planar to three dimensional at the hexamer, which has several conformers that are close in energy. In agreement with the jellium-model prediction, clusters with 8 and 20 atoms are particularly stable, close-packed structures. High-level calculations have been done on clusters with up to 14 atoms, above which *ab initio* theory becomes intractable. Semiempirical methods predict an odd-even alternation in stability that persists to greater than about 40 atoms. This effect is observed experimentally and is attributed to electron-pairing effects. Even-numbered (paired electrons) systems are more stable than those with unpaired electrons.

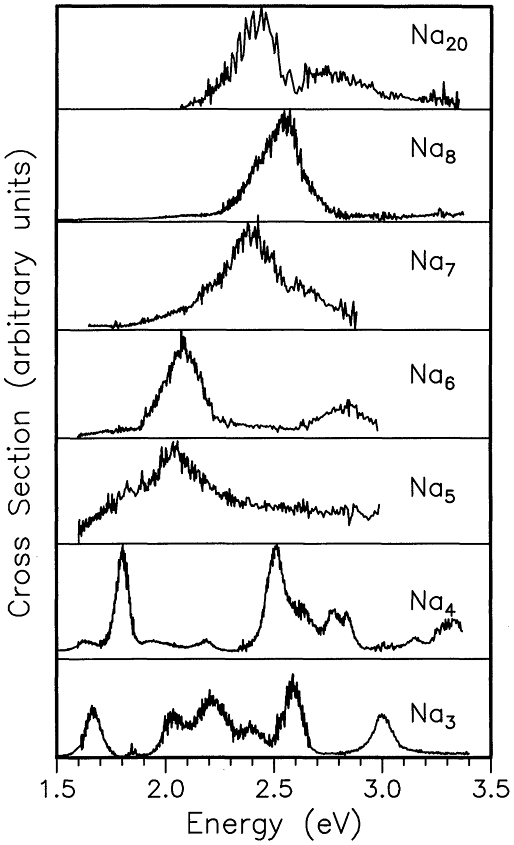

FIG. 7. Photoabsorption spectra for Na_n (n = 3–8, 20) obtained from photodepletion measurements (Wang *et al.*, 1992).

3.1.4 Alkali-Halide Clusters Alkali-halide (MX) cluster systems can be considered microscopic analogs of ionic crystals and surfaces. Simple electrostatic interaction potentials allow accurate calculations of ground-state geometries and binding energies. Of particular interest are alkali-halide cluster anions containing extra electrons that are not associated with a halogen ion. These are of the form $(MX)_n^-$, $M(MX)_n^-$, and $M_2(MX)_n^-$. Molecular-dynamics calculations predict that the excess electron could be localized either near an M^+ or in an empty X^- lattice site (as in a bulk *F*-center defect), or it could occupy a delocalized surface state. Ultraviolet photoelectron spectroscopy has been used to determine the location of the extra electron in $(KI)_n^-$ systems (Yang *et al.*, 1992). Electron binding energies for n = 3, 4, 6, and 12 indicate that the excess electron is loosely

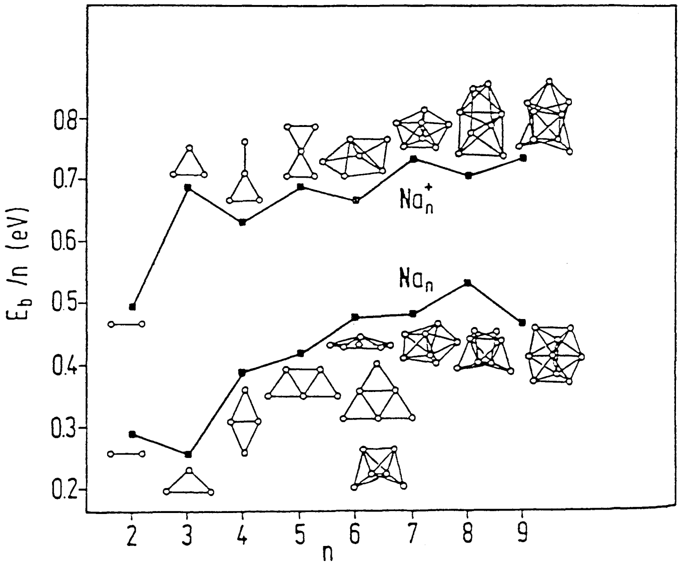

FIG. 8. Binding energy per atom calculated using configuration interaction and the corresponding Hartree–Fock optimized geometries for Na_n and Na_n^+, n = 2–9 (Bonačić-Koutecký *et al.*, 1991).

bound. Because the neutral cluster has a low electron affinity, the extra electron does not significantly change the neutral ground-state geometry and is delocalized over all of the K^+ ions. A relatively high binding energy is observed for $n = 13$. For this cluster, it has been suggested that the electron is localized at the missing corner I^- site of the 3 x 3 x 3 atom cube.

Investigations of the reactivity of alkali-halide clusters such as $(Na_nF_{n-1})^+$ with H_2O and NH_3 confirm that these clusters have structures that are chunks of the bulk lattice (Homer *et al.*, 1993). Defect structures occur in clusters with one less NaF unit than a complete cuboidal cluster. These clusters, such as n = 13 (3 x 3 x 3), 22 (3 x 3 x 5), 31 (3 x 3 x 7), and 37 (3 x 5 x 5), are significantly more reactive, and calculations indicate that NH_3 fits readily into the hole created by the missing atoms.

3.2 Group II Elements

In contrast to the Group I elements, little is known about clusters of the divalent Group II elements. To a first approximation, the $n\ s$ orbitals of an n-atom cluster result in $n/2$ bonding and $n/2$ antibonding orbitals, which are both filled by the $2n$ valence electrons. Because the resulting bond order is zero, the cluster should only be bound by weak van der Waals interactions. In contrast, chemical bonding in the bulk results from overlap of the filled s band and the empty p band. An insulator-to-metal transition is thus expected with increasing cluster size. Nonmetal–metal transitions in Group II elements are a subject of intense study, and this issue is addressed further for mercury clusters in Sec. 3.3.5.

High level *ab initio* quantum-chemical calculations of the ground-state structures of Group II clusters have proved to be somewhat difficult because of the involvement of nearly degenerate s and p levels (Bonačić-Koutecký *et al.*, 1991). The same trends are observed for both Be_n and Mg_n. For n = 3–7, calculations predict that the ground-state geometries are equilateral triangle, tetrahedron, trigonal bipyramid, octahedron, and pentagonal bipyramid. *Ab initio* molecular-dynamics simulations suggest that, for Be clusters with $7 < n < 20$, the growth pattern leads to structures that resemble hexagonal-close-packed (hcp) lattice fragments (Kawai and Weare, 1990). Experimentally, evidence for icosahedral packing has been found in mass-spectral abundance patterns of Mg clusters containing from 147 to 2869 atoms and Ba clusters containing 13 to 35 atoms (Martin *et al.*, 1991).

3.3 Transition Metals

The unfilled d shells in isolated transition-metal atoms result in a large number of low-lying excited electronic states. Consequently, transition-metal clusters, even small ones, have exceedingly high electronic state densities. This makes spectroscopic studies extremely difficult to interpret. The complex electronic structure of these systems results in their rich and varied chemical properties.

3.3.1 Chemical Reactivity Studies of the variations in the reactivity of transition-metal clusters as a function of size can reveal information about their electronic and geometric structure. Most heterogeneous catalysts consist of supported transition-metal clusters, so that understanding the chemical properties of isolated transition-metal clusters may have important practical implications. A large amount of data on the chemistry of transition-metal clusters is presently available, but only a few relevant examples will be presented below. Further references to reviews providing additional information are given at the end of this article.

The fast-flow reactor has been widely used to study the reactivity of atomic clusters. The clusters are typically formed by pulsed-laser vaporization (see Sec. 1.1.3) and are carried by the buffer gas along a channel into the flow tube, where a reagent gas is added. After a fixed reaction time (determined by the flow velocity), the clusters, reagent, and buffer gas expand into vacuum. The clusters are then photoionized and analyzed by time-of-flight mass spectrometry. Thermodynamic properties such as reaction enthalpy and entropy can be determined from equilibrium constants recorded as a function of temperature. A potential source of uncertainty is fragmentation of the products, induced by either the reaction exothermicity or the ionization process.

Many systems display a reactivity that is a strong function of cluster size. For example, in reactions of Fe_n with H_2, the rate constants for $n < 23$ vary by several orders of magnitude. For $n > 23$, only slight variations occur, and the reaction rate is close to that of bulk iron (Riley and Parks, 1987). There appears to be an approximate correlation between the chemical reactivity and the ionization potential (IP) of Fe_n. As shown in Fig. 9, Fe_n with low IPs are generally more reactive toward H_2. This has been rationalized in terms of the ease of electron transfer from the Fe_n to H_2. H_2 adsorption is dissociative, and transfer of an electron from the cluster to the lowest unoccupied molecular orbital on H_2, which is an antibonding orbital, will weaken the H–H bond, decrease the activation energy, and increase the reaction rate. The origin of these variations in reactivity is not a completely resolved issue, however. Other studies suggest that variations in both reactivity and IP may result from simultaneous changes in geometric structure. It is clearly difficult to isolate effects due to electronic structure from those due to geometric structure.

Chemical reactivity has been used as an indirect probe of the geometric structure of neutral transition-metal clusters (Parks *et al.*, 1993). The maximum number of weakly bound adsorbates, such as H_2, N_2, NH_3, or H_2O, that can physisorb onto a cluster is dependent on the number of available surface sites. Saturation coverage is therefore deter-

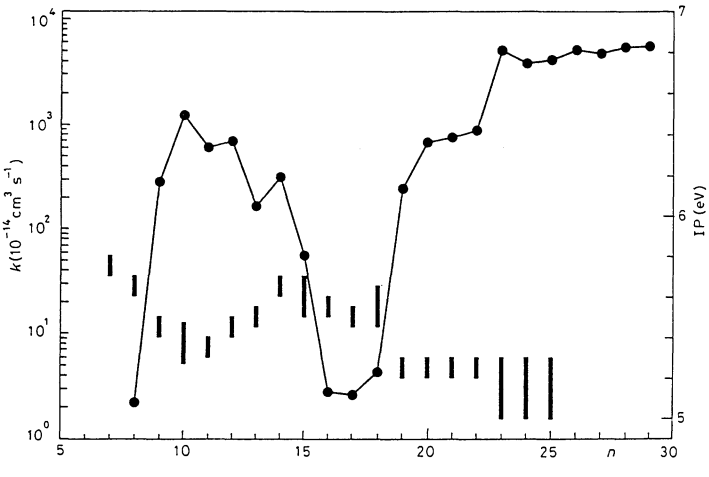

FIG. 9. Correlation of reactivity (left scale, ●) of $Fe_n + H_2$ and ionization potentials (right scale, ■) of Fe_n (Riley and Parks, 1987).

mined by the underlying geometric structure. These studies suggest polyicosahedral geometries for clusters of Fe, Co, and Ni containing 19 to 34 atoms, and icosahedral structure for Cu clusters containing more than 70 atoms. Evidence for the existence of multiple structural isomers appears in many of these systems. Because of the possibility of adsorbate-induced structural rearrangements, these geometries correspond to those of the saturated cluster, and the geometry of the bare metal cluster may be different.

To avoid complications associated with product fragmentation, investigations of cluster reactivity are ideally performed on one cluster size at a time. Mass selection necessitates a study of the cluster ion rather than the neutral species. Probes of size-selected metal-cluster ions offer a controlled method of studying reactivity (Parent and Anderson, 1992). Cluster ions of a specific mass may be selected by quadrupole mass filters, magnetic sectors, or Wien filters. The selected ions are then focused into a low-energy ion beam. Reaction cross sections, as functions of kinetic energy, can be recorded by passing the ion beam through a gas cell or ion guide containing a low pressure of reagent gas and mass analyzing the products. Reaction-rate constants, under thermal conditions, can be recorded as functions of temperature by injecting the ions into a drift tube containing a high pressure of buffer gas (Jarrold, 1991). The mobility of an ion in the drift tube is related to its geometry (von Helden *et al.*, 1991), and increasing the injection energy to transiently heat (anneal) the clusters provides a way to study structural transitions.

Fourier-transform ion cyclotron resonance (FT-ICR) is an alternative approach to probing cluster-ion chemistry. Here, the ions are trapped in a cell by magnetic and electrostatic fields, and a mass spectrum is recorded by coherently exciting the cyclotron motion of the ions and recording the decay of an image signal. Fourier transforming the resulting time-domain signal yields the mass spectrum. The power of this method lies in the ability to store ions selectively and to follow multiple reaction steps. Sequential reaction pathways have been elucidated in the reaction of Fe_4^+ with ethene (C_2H_4) (Irion and Schnabel, 1992). Up to four ethene units can react with the Fe_4^+, and each addition results in elimination of H_2. The reaction products are then probed by collision-induced dissociation. In the case of the $Fe_4(C_2H_2)_3^+$ product, the lowest-energy fragmention pathway is the one-step elimination of $(C_2H_2)_3$ (presumably benzene). The Fe_4^+ is then available to restart the reaction cycle. Thus, Fe_4^+ can effectively catalyze the formation of benzene from three ethene units. This remarkable behavior is specific to the Fe tetramer.

Naked transition-metal cluster ions bind and then dehydrogenate many hydrocarbons. Dehydrogenation occurs to a varying degree, depending on the metal and the cluster size. An example is the reaction of transition-metal clusters with benzene. For niobium cluster cations containing up to 13 atoms, the extent of dehydrogenation increases with size. Complete dehydrogenation to $Nb_nC_6^+$ can occur in the reaction of clusters containing more than four atoms, although partial dehydrogenation also occurs (Zakin *et al.*, 1989). Similar reactivity trends are observed for neutral niobium clusters, and this suggests that geometric structure is an important factor in determining reactivity. Cationic clusters of another Group V element, vanadium, are found to be less reactive than niobium. Although complete dehydrogenation is observed for $n = 5$–7, partial dehydrogenation is the major reaction channel. The difference in reactivity between Nb and V may be interpreted in terms of electronic configuration. Niobium may be more reactive because it is a d^4s^1 metal, while V has a d^3s^2 electronic configuration.

3.3.2 Magnetic Deflection Although the ferromagnetic properties of bulk transition metals are reasonably well understood, the size-dependent magnetic behavior of transition-metal clusters is still under investigation. The magnetic properties of metal clusters can be determined by Stern–Gerlach magnetic deflection experiments. The cluster magnetic moment μ_n will align with the magnetic field B, to an extent that is dependent on the counteracting effects of thermal motion. From the magnitude of the deflection of the cluster in the magnetic field, the magnetic moment and magnetization can be determined. In the case where $\mu_n B/kT \ll 1$, the magnetization M can be derived as

$$M = \mu_n^2 B/3\,kT. \qquad (1)$$

This expression predicts simple thermodynamic behavior known as superparamagnetism. Magnetic deflection experiments have recently been performed for mass-selected iron clusters containing up to 700 atoms (Billas *et al.*, 1993). The deflection of small Fe clusters is always in the high-field direction, which suggests that rapid spin relaxation occurs via coupling of spin and rotational angular momentum. The magnetic moment per atom decreases from about $3\mu_B$ for $25 < n < 130$ to near the bulk limit of about $2.2\mu_B$ for clusters containing more than about 500 atoms. The rotational temperature of iron clusters has a significant effect on the measured magnetization. When the precession frequency of the spin is similar to the rotational frequency, resonant spin–rotation coupling can occur. In this case, the approximations of Eq. (1) are no longer valid, and the magnetization of rotationally cold clusters is greatly reduced and exhibits a nonlinear dependence on the magnetic field.

3.3.3 Photoionization Trends in metal-cluster ionization potentials (IPs) may be reproduced by a classical electrostatic image-charge model (Wood, 1981). The ionization potential for a spherical metal droplet is expressed in terms of the bulk work function (WF) and the radius of the cluster, R:

$$\mathrm{IP} = \mathrm{WF} + \tfrac{3}{8}e^2/R. \tag{2}$$

Although the constant term $\frac{3}{8}$ is mathematically correct in the electrostatic limit, if a more realistic cutoff to the image potential is considered, the constant term becomes $\frac{1}{2}$ (Makov *et al.*, 1988). The limitations of this description of metal-cluster IPs stem from the fact that clusters do not have uniform valence-electron density, and, in many cases, the assumption of spherical shape is not valid. Inclusion of a term for spillout of the electron density beyond the edge of the cluster gives the semiclassical expression

$$\mathrm{IP} = \mathrm{WF} + \tfrac{1}{2}e^2/(R + a), \tag{3}$$

where a is the element-specific extent of electron spillout. Figure 10 compares measured IPs for silver clusters to the predictions of Eqs. (2) and (3) (Alameddin *et al.*, 1992). The best agreement is obtained for clusters with nearly spherical geometries, which correspond to clusters with closed electronic shells within the electronic shell-model description. The agreement also improves for larger cluster sizes. For small clusters, large size-dependent oscillations in IP result from quantum size effects, and deviations are significant. Although this model fits the IPs of simple metal clusters such as the alkalis and noble metals, other transition-metal clusters do not conform to the classical description. Clusters of transition metals such as Ni, Nb, Fe, and Co exhibit similar trends in their IPs. There is a rapid decrease in the IP for up to approximately $n = 25$, and then, for $25 < n < 100$, the IP levels off to a value that is about 0.5 eV above the bulk work function. Disagreement with the classical model suggests that the d band has a significant influence on the electronic structure of these clusters.

A delayed ionization phenomenon, attrib-

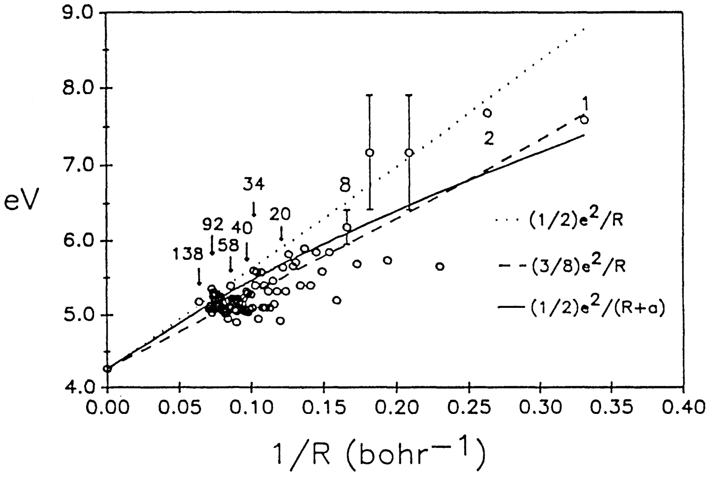

FIG. 10. Silver-cluster ionization potentials compared to the classical predictions (Alameddin *et al.*, 1992).

uted to thermionic emission, has been observed in clusters of W, Nb, and Ta (Amrein *et al.*, 1991; Leisner *et al.*, 1991). Multiphoton excitation (with photon energies less than the ionization energy) may not lead to direct ionization of metal clusters. Rapid internal conversion results in a highly vibrationally excited (hot) cluster that can cool either by evaporation of atoms or by emission of an electron. For strongly bound transition-metal clusters, the rate of dissociation of the excited cluster is slow compared to that of ionization, and delayed ionization may be observed. A statistical model has been developed by Klots (1991) to predict rate constants for electron emission from small neutral clusters. Taking into account the Coulombic attraction between the separating particles, the rate constant is

$$k(T) = \left(\frac{2k_bT}{h}\right)\left(\frac{Q_{\text{vib}}}{Q^0_{\text{vib}}}\right)\exp\left(\frac{-IP}{k_bT}\right) \times \left[\frac{2b}{a_0} + 2\left(\frac{Q_s\pi}{4}\right)^{1/2} + Q_s\right], \tag{4}$$

where

$$Q_s = 8\pi^2\mu b^2 k_b T/h^2.$$

The terms Q^0_{vib} and Q_{vib} are the vibrational partition functions before and after electron emission, μ is the reduced mass (essentially the mass of an electron), b is the hard-sphere collision radius, and a_0 is the Bohr radius. In the limit of large n, this expression converges to the bulk description of thermionic emission given by the Richardson–Dushman equation. Verification of this model requires that the ionization rate be measured as a function of cluster temperature, which has not yet been performed.

3.3.4 Collision-Induced Dissociation Dissociation energies of atomic clusters provide information about cluster stability. Collision-induced dissociation (CID) is a widely used method of determining adiabatic bond dissociation energies. These experiments are performed by colliding a beam of internally cold ions with a collision gas (such as Xe), and recording the CID cross section as a function of kinetic energy. After the finite lifetime of the collisionally excited cluster is taken into account by means of a statistical RRKM (Rice, Ramsperger, Kassel, and Marcus) model, the bond dissociation energy (BDE) of the cluster ion can be extracted from the CID threshold. Armentrout and co-workers (Su *et al.*, 1993) have measured dissociation energies for a series of transition-metal cluster ions. The CID of V_n^+, $n = 2$–20, with Xe will be discussed as an example. All V_n^+ dissociate by sequentially losing atoms except for $n = 4$, which fragments by losing the dimer. Fragmentation by loss of individual atoms is characteristic of nearly all metal-cluster ions and suggests that the cohesive energies increase relatively smoothly with cluster size. Bond energies for neutral V clusters are derived from the BDEs of the ion and IPs of the neutral. An even-odd oscillation in the BDE of both neutral and cationic clusters, $4 \leq n \leq 10$, correlates with trends in the reactivity of V_n with D_2 and N_2. Even-numbered clusters are more strongly bound and less reactive, suggesting that these are closed-shell clusters.

Cluster cohesive energies are the sum of the BDEs divided by the number of atoms. It is informative to compare the cohesive energies to the predictions of the liquid-drop model (Miedema, 1978). In this model, an increase in the surface energy (per atom) with decreasing cluster size results in a decrease in the cohesive energy. The cohesive energy of a cluster with n atoms is expressed as

$$E_c(n) = \Delta H_{\text{vap}}{}^0 - (36\pi/n)^{1/3}\gamma^0 V_a^{2/3}, \tag{5}$$

where γ^0 is the surface energy, $\Delta H_{\text{vap}}{}^0$ is the bulk heat of vaporization, and V_a is the atomic volume. A plot of E_c against $n^{1/3}$ in Fig. 11 shows that the measured cohesive energies approach the spherical-drop prediction with increasing cluster size and are in remarkably good agreement for clusters containing more than approximately 13 atoms. Similar results have been obtained for other transition metals.

3.3.5 Nonmetal-to-Metal Transitions An unresolved issue in studying clusters of metal atoms is the minimum number of atoms that is necessary for the cluster to exhibit properties of the bulk metal. Here the exact definition of "metallic" behavior is critical. Intrinsically metallic properties appear when the valence electrons are delocalized and the gap between highest occupied and lowest un-

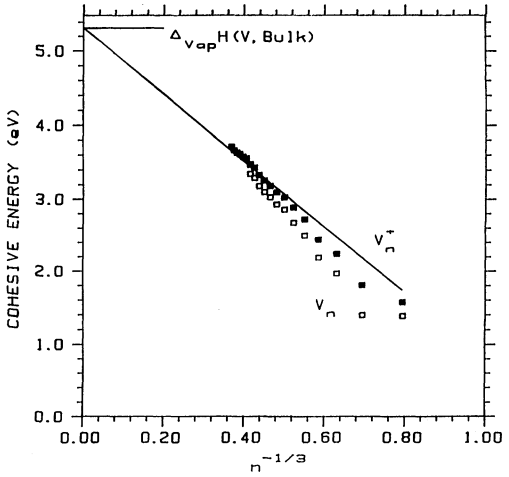

FIG. 11. Cohesive energies for neutral (□) and cationic (■) vanadium clusters (Su *et al.*, 1993). The line corresponds to cohesive energies calculated using the spherical-drop model.

occupied electronic energy levels is small relative to *kT*. Nonmetal–metal transitions in Group II clusters, briefly discussed in Sec. 3.2, have been systematically studied in mercury clusters. Because of the closed-shell atomic configuration ($5d^{10}6s^2$), small mercury clusters are expected to be bound by weak van der Waals interactions. With increasing cluster size, the 6*s* and 6*p* shells broaden until the full 6*s* shell eventually overlaps with the empty 6*p* level, delocalizing the electrons into metal-like levels. Cohesive energies and ionization energies have been calculated using a tight-binding electronic theory, and these suggest that covalent bonding begins to occur in Hg clusters containing 13–20 atoms. Calculations of the electronic band gap and density of states suggest that clusters containing approximately 20–80 atoms are semiconducting, whereas those containing more than 80 atoms are metallic. The transition from van der Waals to metallic is therefore expected to be gradual, and experimental evidence supports this prediction. Electron binding energies for Hg_n, $n < 110$, have been measured by vacuum-ultraviolet photoelectron spectroscopy (Kaiser and Rademann, 1992). In small Hg_n clusters, $n < 18$, the IP deviates significantly from that predicted by the spherical-droplet model, providing evidence for the dominance of van der Waals bonding. For clusters containing about 18–70 atoms, the IP begins to evolve gradually toward the spherical-droplet prediction. In addition, the photon-energy dependence of the photoemission spectra indicates that Hg_n, $n > 60$, already have a high density of *p* states near the Fermi energy, providing further evidence of the emergence of metallic properties in this size range.

3.3.6 Noble Metals

Because of their $d^{10}s^1$ valence-electronic structure, a description that is conceptually similar to that of the alkali metals can be applied to the noble metals, Cu, Ag, and Au. Although the metal–metal bonding in clusters of these metals is dominated by the *s*-*s* interaction, contributions of the underlying *p* and *d* electrons significantly influence the properties. Therefore, *d*-*s*-*p* hybridization and *d*-*s* correlation energy must be taken into account in theoretical studies (Bonačić-Koutecký *et al.*, 1991). Because of the large number of electrons in these systems, *ab initio* calculations have thus far been limited to $n < 9$. Calculated ground-state geometries and size-dependent variations in the binding energies are similar to those of the alkali elements.

There have been several experimental investigations of the evolution of electronic properties in noble-metal clusters (both ionic and neutral). Vertical electron affinities of the anions of copper, silver, and gold clusters have been measured by ultraviolet photoelectron spectroscopy (Taylor *et al.*, 1992). For up to 70 atoms, variations in electron affinity are found to agree with the predictions of the electronic shell model for ellipsoidally distorted clusters. (Clusters with closed electronic shells have high electron affinities.) In clusters containing only a few atoms, the *d*-*s* band separation is fixed at approximately that of the bulk, and, with increasing cluster size, the *d* band evolves smoothly toward that of the bulk. Photofragmentation studies of size-selected silver-cluster cations provide verification of the near-spherical geometric structure of closed electronic-shell clusters (Tiggesbäumker *et al.*, 1992). The resulting optical spectra show collective electronic resonances that are in qualitative agreement with the predictions of Mie theory. Also in agreement with the predictions of the electronic shell model are the ionization potentials of Cu_n, $n < 150$, and Ag_n, $n < 100$ (Knickelbein, 1992; Alameddin *et al.*, 1992).

3.4 Group III Elements

The elements of Group III (B, Al, Ga, In, and Tl) have an s^2p^1 valence configuration. The properties vary dramatically down the group. Bulk boron is a large–band-gap semiconductor, whereas bulk aluminum is a nearly ideal free-electron metal. Gallium, indium, and thallium are also metals. There is a considerable amount of information available for clusters of boron and aluminum, but relatively little is known about the clusters of the other elements in this group.

3.4.1 Boron In the solid state, many boron compounds consist of three-dimensional networks of octahedral B_6 and cuboctahedral B_{12} clusters. However, there is no evidence to suggest that these species are important building blocks of the isolated atomic clusters. Mass spectra of boron clusters generated by laser vaporization show particularly abundant clusters at 5, 10, 11, and 13 atoms. Collision-induced dissociation measurements of the bond dissociation energies indicate that B_{13}^+ is particularly stable. The bond dissociation energies rapidly approach the bulk cohesive energy, suggesting that these clusters have close-packed, highly coordinated geometries. This is consistent with quantum-chemical calculations, which suggest that small boron clusters adopt three-dimensional geometries. According to these calculations, B_5 is a trigonal bipyramid, and B_6 is a capped pentagon. Molecular-dynamics calculations for B_{13}^+ suggest that the plausible icosahedron geometry is not even a local minimum. The lowest-energy geometry of B_{13}^+ appears to be a relatively low-symmetry structure consisting of three hexagonal pyramids. The enhanced stability of this cluster seems to be related to a closed-shell electronic structure.

3.4.2 Aluminum Since bulk aluminum is almost an ideal free-electron metal, the electronic shell model might be expected to provide a reasonable description for the properties of aluminum clusters. Assuming that aluminum is trivalent, the shell closings should occur with approximately one-third of the atoms required for shell closings with the s^1 alkali metals. Evidence for shell structure of aluminum clusters has been found in dissociation-energy measurements, ionization-energy measurements, and studies of chemical reactivity. The expected shell closings are observed for clusters with 20 and 40 valence electrons (Al_7^+ and Al_{13}^-). However, for clusters with >20 atoms, there are some significant discrepancies between the predictions of the shell model and experimental results. The origin of these discrepancies has not been resolved. However, perturbations provided by the highly charged positive cores (3+) and issues related to the overlap of the s and p bands may be important. In bulk aluminum, the 3s and 3p bands are broad and overlap. These bands are well separated in the atom, and the cluster-size regime where they begin to overlap is an issue of considerable interest. Photoelectron spectra of Al_n^- suggest that the 3p band begins to overlap with the 3s band for clusters with $n > 20$.

Quantum-chemical calculations for small aluminum clusters suggest that they adopt relatively open, covalently bound geometries, where the bonding arises mainly from the p orbitals. Closed-shell Al_7^+ (20 valence electrons) has a distorted capped octahedron geometry. The significance of this geometry is that all of the atoms lie approximately the same distance from the center of mass, as if distributed on the surface of a sphere. (Spherical geometries are expected for closed–electronic-shell clusters.) The electronic structure of this cluster can be related to the shell-model prediction, even though there is a substantial gap between the s and p levels. Al_{13}^- is the next largest cluster to have to have a closed shell. Quantum-chemical calculations for this cluster suggest that the gap between the s and p levels has disappeared. The geometry is icosahedral, and the bonding is more metallic. For large, cool (room temperature) aluminum clusters, a geometric shell structure consistent with octahedral symmetry has been observed in mass spectra. However, for warmer (possibly liquid) aluminum clusters, the geometric shell structure vanishes and electronic shell structure appears to emerge (Baguenard *et al.*, 1993). Gallium and indium clusters display a similar electronic shell structure.

3.5 Group IV Elements

3.5.1 Carbon Carbon clusters have been extensively studied. Most of the interest in carbon clusters has recently been focused on the fullerenes, which are discussed in the article FULLERENES. Fullerenes [spheroidal-shell carbon clusters consisting of 12 pentagons and $(n - 20)/2$ hexagons] exist for carbon clusters containing more than around 30 atoms. Bulk quantities of C_{60} and a few other larger fullerenes have been isolated. Mobility measurements for gas-phase carbon-cluster ions have revealed the existence of a large number of different structural isomers, including a range of roughly planar monocyclic and bicyclic rings (von Helden *et al.*, 1991). The formation of these unusual geometries reflects carbon's tendency to form multiple bonds. A number of mixed metal–carbon clusters have also been studied. These include the metallofullerenes (Bethune *et al.*, 1993) and metallocarbohedrenes (Wei *et al.*, 1992). The latter are a class of M_8C_{12} clusters (M = Ti, V, Zr, Hf) that appear to be relatively stable, but they have not yet been isolated in bulk quantities.

3.5.2 Silicon and Germanium

Below carbon in the periodic table are silicon and germanium, which are also covalently bound materials at room temperature and pressure. However, unlike carbon, these elements show little tendency to form double bonds. Thus, there is no analog of the graphite geometry for these elements in their bulk states, and there is no evidence for any fullerene analogs in the clusters. Silicon, in particular, is an important material in the semiconductor industry. With the continuous push to smaller and smaller features in semiconductor devices, there is now considerable interest in how the properties of silicon change as dimensions approach atomic scales. Although much less is known about germanium clusters than silicon clusters, the preliminary indications are that, like the bulk materials, silicon and germanium clusters have similar properties (Jarrold, 1993).

Unlike bulk Si and Ge, which evaporate mainly individual atoms when heated, the atomic clusters dissociate by preferential loss of fragments such as Si_6, Si_{10}, and Ge_{10}. These fragments are particularly strongly bound (with cohesive energies close to that of the bulk), and they appear as "magic numbers" in mass spectra of the clusters generated by pulsed laser vaporization. The high stability of Si_6 and Si_{10} has been reproduced in *ab initio* quantum-chemical calculations (Raghavachari, 1990). These calculations suggest that small silicon clusters have geometries that cannot be considered as fragments of the bulk diamondlike lattice. The geometry of Si_6 is a distorted octahedron, and Si_{10} is probably a tetracapped biprism. (A tetracapped octahedron geometry appears to be slightly higher in energy.) The bulk fragments (the cyclohexane and adamantane skeletons) are considerably higher in energy because of the large number of dangling bonds. Reconstruction to the more compact and highly coordinated geometries mentioned above is driven by these dangling bonds, in much the same way that bulk silicon surfaces reconstruct. The size regime where silicon clusters adopt a bulklike geometric structure is not yet known.

The chemical properties of silicon clusters have also been extensively studied (Jarrold *et al.*, 1991). Surprisingly, even quite large clusters, such as Si_{70}, appear to be substantially less reactive than bulk silicon surfaces. The low reactivity probably reflects the low density of dangling bonds on the cluster's surface. Whether or not there exist some clusters (such as Si_{33}, Si_{39}, and Si_{45}) with particularly low reactivity is an unresolved controversy. The chemical properties of silicon clusters can be modified by annealing, which demonstrates that cluster growth may not lead to the lowest-energy geometric isomer. The relatively large number of structural isomers suggested by these studies is probably a consequence of the directed nature of the chemical bonding in these clusters.

Direct experimental information on the geometries of silicon clusters has been very difficult to obtain. Spectroscopic information has been obtained for silicon clusters containing 2–7 atoms using photoelectron spectroscopy and surface-enhanced Raman spectroscopy of matrix-isolated, size-selected clusters (Honea *et al.*, 1993). These experimental studies generally confirm the main features of calculated predictions for these smaller clusters. For clusters with more than ten atoms, there is little direct experimental information about the geometries. Mobility measurements shown in Fig. 12 suggest that clusters with 10 to ~27 atoms follow a one-dimensional growth sequence to give prolate geometries (Jarrold and Bower, 1992). At ~27 atoms, a structural transition occurs, and the clusters reconstruct to a more spherical ge-

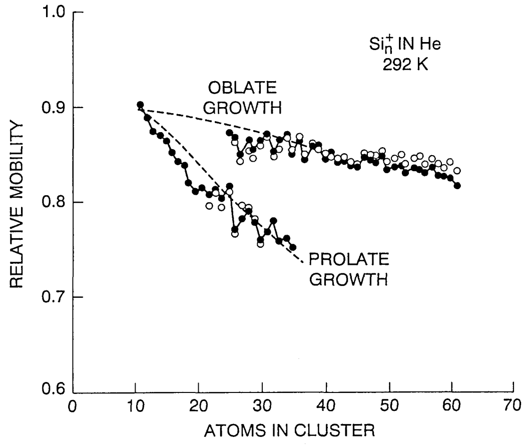

FIG. 12. The relative mobility of silicon clusters as a function of cluster size (Jarrold and Bower, 1992). Dashed lines indicate predicted oblate and prolate growth sequences.

ometry. For these larger clusters, the theoretical calculations become less reliable (and more controversial). Results obtained from molecular-dynamics simulations using the local-density approximation suggest that clusters with 30–50 atoms consist of two strongly interconnected and buckled shells (Röthlisberger *et al.*, 1994). The buckling arises because of the tendency of silicon to form sp^3 hybrids. For Si_{45}, the core consists of 7–9 atoms, and the rest of the atoms are in the surrounding shell.

4. CONDENSED-PHASE CLUSTERS

4.1 Deposited Clusters

While a well-defined cluster beam provides the best environment in which to probe the physical and chemical properties of atomic clusters, exploitation of these properties requires that the clusters be collected either on a surface or in a matrix. Deposited atomic clusters can be studied using a variety of techniques that cannot be applied to gas-phase clusters. These include x-ray photoelectron spectroscopy (XPS), extended x-ray absorption fine structure (EXAFS), electron energy-loss spectroscopy (EELS), and scanning tunneling microscopy (STM). A critical consideration is the nature of the cluster–surface and cluster–cluster interactions. The importance of cluster–surface interactions is well known in studies of the catalytic properties of supported transition-metal clusters because these interactions can modify both the electronic and geometric structure of the cluster (Poppa, 1984; Lewis, 1993). If the cluster–substrate interaction is strong, then the cluster will be strongly perturbed and anchored to the surface. Conversely, weak cluster–substrate interactions mean greater mobility of the cluster on the surface. Coalescence may occur, even at low surface coverages, when cluster–cluster interactions are stronger than cluster–substrate interactions. Many of the issues raised above remain largely unexplored, as there have been relatively few controlled experimental studies of deposited, monodisperse clusters.

There have been some efforts to generate high-quality thin films by ionized cluster beam deposition. If the kinetic energy of the cluster is raised to the point that it is a substantial fraction of the cohesive energy, the violence of the cluster–surface impact will either cause the cluster to shatter or embed it into the surface, depending on the nature of the cluster and the surface. If the cluster sticks to the surface, then the kinetic energy is initially deposited in a relatively small area, leading to locally high temperatures and possibly even local melting (Landman and Luedtke, 1992). The energy provided by the cluster impact increases surface mobility and leads to highly crystalline thin films at low substrate temperatures.

The large surface-to-volume ratio of clusters gives them unique thermodynamic properties. One example is the reduction of the melting point with decreasing particle size. Several models have been developed to account for this effect. These approaches describe the modification of the cluster chemical potential by the surface free energy. Experimentally, the melting temperature range for a distribution of small gold particle sizes has been measured by observing the loss of sharp rings in the electron diffraction pattern (Buffat and Borel, 1976). From the good agreement between the experimental data and the model (Fig. 13), it appears that the thermodynamic model provides a reasonable description of the size dependence of the melting temperature.

4.2 High-Resolution Imaging

Of critical importance in the study of metal clusters is the experimental determination of cluster geometry. Direct spectroscopic deter-

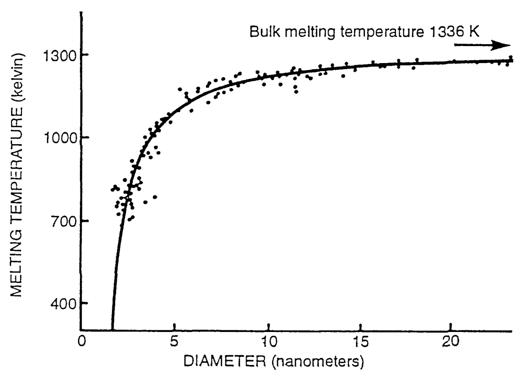

FIG. 13. Size dependence of the melting temperature of Au particles as a function of particle diameter (Buffat and Borel, 1976). The line is calculated using a thermodynamic model discussed in the text.

mination of the structure of small, unsupported metal clusters has so far only been possible for clusters with fewer than four atoms. Atomic-resolution imaging techniques such as scanning tunneling microscopy (STM), atomic force microscopy (AFM), and high-resolution (transmission) electron microscopy (HREM) have provided some insight into the geometric structure of larger clusters, even though the exact relationship between gas-phase structure and surface structure is unknown. On crystalline substrates, HREM is capable, in theory, of imaging clusters containing as few as several atoms (Spence, 1988). However, STM and AFM are better techniques for imaging the smallest clusters (down to single atoms), and additional information about the electronic states of deposited clusters may be obtained from the voltage dependence of the tunneling current (Chen, 1993).

There have been extensive studies of the structure of small metal particles by transmission electron microscopy since the discovery that fcc metals often form noncrystallographic multiply twinned particles (MTPs) (Ajayan and Marks, 1990). Decahedral, or pentagonal bipyramidal, MTPs contain five twin-related tetrahedral units, while icosahedral MTPs contain twenty such units. In HREM images of gold particles, single-crystalline and singly twinned structures are observed to predominate in smaller particles (1–2 nm diameter), while icosahedral and decahedral MTPs, as well as other highly twinned structures, are observed in larger particles. A variety of theoretical models also indicate that clusters containing fewer than a few thousand atoms do not pack as chunks of the bulk lattice. Molecular-dynamics simulations, using many-body potentials, have been performed for clusters of a number of transition metals. The relative energies of optimized decahedral, icosahedral, and fcc cuboctahedral clusters have been compared, and an element-specific cluster size for the icosahedral to fcc cuboctahedral transition is found (Uppenbrink and Wales, 1992). Although electron-microscope images do not allow determination of the exact number of atoms for this transition, the results of this model agree qualitatively with the electron-microscope observations.

Under the influence of the energetic electron beam in the electron microscope, small metal particles are often observed to undergo rapid (<1/30 s) structural transitions. This structural instability is an intriguing phenomenon because of the possibility of phase coexistence. A possible mechanism for these structural fluctuations is quasimelting, in which rapid isomerization between energetically similar geometries occurs at temperatures lower than the cluster thermodynamic melting point. Molecular-dynamics simulations suggest the coexistence of a variety of different structures with similar stabilities, which may be a reason for structural fluctuations (Uppenbrink and Wales, 1993). However, recent estimates of the temperature of gold particles in an electron microscope suggest that a mechanism based on rapid recrystallization through a highly excited, possibly molten, intermediate state (created by the high-energy electron beam) is more plausible (Ugarte, 1993).

4.3 Quantum Dots

Semiconductor clusters in the nanometer size range exhibit unique quantum-confinement effects. In the bulk semiconductor, because electrons are strongly delocalized and the effective mass is small, the electron–hole pair that results from an electronic transition from the valence band to the conduction band has an effective diameter that is relatively large, on the order of 1–20 nm (depending on the material). When the dimension of the cluster is of the same magnitude or less, this exciton becomes spatially confined (a quantum dot). This quantum confinement increases the band gap as the cluster size decreases (Brus, 1991).

The size-dependent band gap can be qualitatively described by a simple model that reduces the problem to a two-body interaction between the electron in the bottom of the conduction band and the hole that it left in the top of the valence band (Steigerwald and Brus, 1990). The electron and the hole are considered particles with effective masses m_e and m_h, respectively. The allowed electronic states are given by the solutions to the Schrödinger equation for two particles confined to a sphere of radius R:

$$E(R) = E_g + (h^2\pi^2/2R^2)(1/m_e + 1/m_h) - 1.8e^2/\epsilon R, \qquad (6)$$

where ϵ is the bulk dielectric constant. The bulk band gap, E_g, is modified by the positive kinetic energy and by the negative Coulomb attraction between the electron and the hole. More accurate electronic-structure calculations employ an empirical psuedopotential method (Ramakrishna and Friesner, 1993). Diagonalization of the Hamiltonian matrix at a set of points in the Brillouin zone results in the band structure. A comparison of experimental and calculated band gaps is shown in Fig. 14. Although the simple description of Eq. (6) is a good qualitative prediction of the increase in the energy of the lowest electronic transition with decreasing cluster size, it fails for small clusters. On the other hand, the psuedopotential model agrees surprisingly well with the measured energies, given that cluster surface effects are not taken into account.

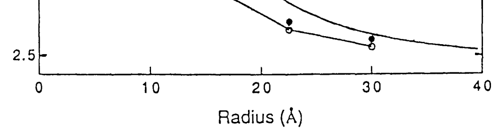

FIG. 14. Size-dependent band gaps of CdS clusters. The solid line is the prediction of the simple particle-in-a-box model given in Eq. (6), the experimental points are the filled circles, and the open circles (connected by lines) are calculated using the empirical pseudopotential method (Ramakrishna and Friesner, 1993).

Efforts to understand luminescence from porous silicon have prompted the investigation of this property in nanocrystalline silicon (Littau *et al.*, 1993). Bulk silicon is an indirect-gap semiconductor. However, relaxation of translational symmetry in finite-size silicon crystals causes the electronic transition to the lowest excited state to become weakly dipole allowed. Emission occurs at approximately 0.5 eV above the bulk Si band gap in 2-nm-diam Si particles, in agreement with the predicted quantum size effect.

4.4 Inorganic Clusters

Although the perturbing influence of the ligands prohibits direct comparison to free metal clusters of the same size, ligated clusters provide an alternative route to the investigation of monosized clusters (Schmid, 1992). One of the best-characterized ligated metal clusters is $Au_{55}(P(C_6H_5)_3)_{12}Cl_6$. EXAFS measurements show that this complex has a slightly distorted closed-shell cuboctahedral rather than an icosahedral core. The fourth and fifth closed cuboctahedral shells, containing 309 and 561 atoms, have also been investigated, for Pt and Pd complexes. As expected, the electronic states of the ligated surface atoms are distinctly different from those of the core atoms. Mössbauer data and optical-absorption spectra for the Au_{55} complex suggest that the inner Au_{13} atomic core is partly metallic in character. In ligated platinum systems, large Knight shifts, indicative of delocalized electrons, in ^{195}Pt NMR studies provide evidence that the core of the Pt_{309} complex manifests bulk behavior.

A mathematical scheme called the Hume–Rothery rule has been extended and developed to predict metal-packing arrangements in ligated "high nuclearity" metal clusters by counting electrons and atoms (Teo and Zhang, 1990). If the total nuclearity of a cluster, m, is the sum of bulk (core) and surface atoms, $m_b + m_s$, then the total number of electrons (both metal valence and ligand donated) is given by

$$N = m_b n_b + m_s n_s = (m_b + r m_s) n_b, \qquad (7)$$

where n_b and n_s are the number of electrons in bulk and surface atoms, respectively, and r is the ratio of n_s to n_b, empirically determined to be $18/16 = 1.125$. The predicted electronic requirements correlate well with the number of electrons in experimentally observed stable clusters. Two general classifications arise from a systematic analysis of cluster structures. The first is called "supraclusters," which are vertex-, edge-, or face-sharing clusters of primary clusters (tetrahedral or icosahedral units, for example). The other class is termed "polyhedral" clusters, which grow from primary units in a layer-by-layer fashion. Recently, these electron- and atom-counting rules have been applied to enumerate all possible stereoisomers of binary (two metals) icosahedral clusters (Teo *et al.*, 1993).

The structures of a wide variety of ligated clusters have been determined by x-ray diffraction (Mingos and Wales, 1990; Schmid, 1992). An intriguing indium-containing fullerene cage analog deserves mention (Sevov and Corbett, 1993). Double-hexagonal close-packed arrays of pentagon-sharing In_{74} fullerenes are stuffed with $In_{10}Ni$ clusters and Na_{39} polyhedra. The electronic structure of these $Na_{96}In_{97}Z_2$ (Z = Ni, Pd, or Pt) arrays apparently parallels that of carbon fullerenes.

4.4 Nanophase Materials

A logical extension of isolated cluster research is the assembly of nanophase structures by selective manipulation of physical and chemical properties of nanoscale aggregates. The effects of spatial confinement of electrons that results in unique cluster properties are also responsible for novel properties in nanoscale materials. This subject is mentioned only briefly here. More detailed information may be found in the article NANOPHASE MATERIALS and in several recent reviews (Siegel, 1993; Andres *et al.*, 1989).

GLOSSARY

Capped Cluster: Condensed-phase cluster that is stabilized by chemical modification (with ligands) of the cluster surface.

Doped Cluster: A pure cluster containing a foreign (guest) atom or molecule.

Jellium Approximation: An approximation in which the nuclei are smeared out into a uniform positively charged background surrounded by the nearly free valence electrons.

Molecular Beam: A stream of clusters (of atoms or molecules) in a collision-free vacuum environment.

Photodepletion Spectroscopy: A form of absorption spectroscopy obtained by monitoring a decrease (depletion) in the cluster-ion signal when the excitation laser is tuned to an electronic resonance.

Quantum Dot: A cluster in which electronic excitation is spatially confined in three dimensions by the cluster boundaries.

R2PI: Resonance-enhanced two-photon ionization, where the first photon is absorbed into an intermediate electronic state, and absorption of a second photon results in ionization.

Vertical Binding Energy: The diabatic energy difference between initial and final states.

Works Cited

Ajayan, P. M., Marks, L. D. (1990), *Phase Transitions* **24–26**, 229–258.

Alameddin, G., Hunter, J., Cameron, D., Kappes, M. M. (1992), *Chem. Phys. Lett.* **192**, 122–128.

Amrein, A., Simpson, R., Hackett, P. (1991), *J. Chem. Phys.* **95**, 1781–1800.

Andres, R. P., Averback, R. S., Brown, W. L., Brus, L. E., Goddard, W. A., III, Kaldor, A., Louie, S. G., Moscovits, M., Peercy, P. S., Riley, S. J., Siegel, R. W., Spaepen, F., Wang, Y. (1989), *J. Mater. Res.* **4**, 704–736.

Baguenard, B., Pellarin, M., Lermé, J., Vialle, J. L., Broyer, M. (1993), *J. Chem. Phys.* **100**, 754–755.

Barnett, R. N., Landman, U., Rajagopal, G., Nitzan, A. (1990), *Israel J. Chem.* **30**, 85–105.

Berry, S. D. (1993), *Chem. Rev.* **93**, 2379–2394.

Bethune, D. S., Johnson, R. D., Salem, J. R., de Vries, M. S., Yannoni, C. S. (1993), *Nature* **366**, 123–128.

Billas, I. M. L., Becker, J. A., Châtelain, A., de Heer, W. A. (1993), *Phys. Rev. Lett.* **71**, 4067–4070.

Birringer, R., Gleiter, H., Klein, H.-P., Marquardt, P. (1984), *Phys. Lett. A* **102**, 365–369.

Bohren, C. F., Huffman, D. R. (1983), *Absorption and Scattering of Light by Small Particles*, New York: Wiley.

Bonačić-Koutecký, V., Fantucci, P., Koutecký, J. (1991), *Chem. Rev.* **91**, 1035–1108.

Brus, L. E. (1991), *Appl. Phys. A* **53**, 465–474.

Brutschy, B. (1992), *Chem. Rev.* 92, 1567–1587.

Buffat, Ph., Borel, J.-P. (1976), *Phys. Rev. A* **13**, 2287–2298.

Chen, C. J. (1993), *Introduction to Scanning Tunneling Microscopy*, New York: Oxford University Press.

Clemenger, K. (1985), *Phys. Rev. B* **32**, 1359–1362.

Cohen, M. L., Knight, W. D. (1990), *Phys. Today* **43** (12), 42–50.

de Heer, W. A., Knight, W. D., Chou, M. Y., Cohen, M. L. (1987), in: H. Ehrenreich, D. Turnbull (Eds.), *Solid State Physics*, Vol. 40, New York: Academic.

Easter, D. C., Baronavski, A. P., Hawley, M. (1993), *Chem. Phys. Lett.* **206**, 329–336.

Faraday, R. (1857), *Philos. Trans. R. Soc. London* **141**, 145.

Farges, J., deFeraudy, M. F., Raoult, B., Torchet, G. (1986), *J. Chem. Phys.* **84**, 3491–3501.

Gentry, W. R. (1988), in: G. Scoles (Ed.), *Atomic and Molecular Beam Methods*, Vol 1, New York: Oxford University Press, p. 54.

Gough, T. E., Mengel, M., Rowntree, P. A., Scoles, G. (1985), *J. Chem. Phys.* **83**, 4958–4961.

Goyal, S., Schutt, D. L., Scoles, G. (1993), *Acc. Chem. Res.* **26**, 123–130.

Heath, J. R. (1992), *Science* **258**, 1131–1133.

Herron, N., Wang, Y., Eddy, M., Stucky, G., Cox, D. E., Moller, K., Bein, T. (1989), *J. Am. Chem. Soc.* **111**, 530–540.

Hoare, M. R. (1979), *Adv. Chem. Phys.* **40**, 49–135.

Homer, M. L., Livingston, F. E., Whetten, R. L. (1993), *Z. Phys. D* **26**, 201–203.

Honea, E. C., Ogura, A., Murray, C. A., Raghavachari, K., Sprenger, W. O., Jarrold, M. F., Brown, W. L. (1993), *Nature* **366**, 42–44.

Irion, M. P., Schnabel, P. (1992), *Ber. Bunsenges. Phys. Chem.* **96**, 1091–1100.

Jarrold, M. F. (1991), *J. Cluster Sci.* **2**, 137–181.

Jarrold, M. F., Ijiri, Y., Ray, U. (1991), *J. Chem. Phys.* **94**, 3607–3618.

Jarrold, M. F., Bower, J. E. (1992), *J. Chem. Phys.* **96**, 9180–9190.

Jarrold, M. F. (1993), in: C.-Y. Ng, T. Baer, I. Powis (Eds.), *Cluster Ions*, Chichester: Wiley.

Jortner, J. (1992), *Z. Phys. D* **24**, 247–275.

Kaiser, B., Rademann, K. (1992), *Phys. Rev. Lett.* **69**, 3204–3207.

Kawai, R., Weare, J. H. (1990), *Phys. Rev. Lett.* **65**, 80–83.

Klots, C. E. (1991), *Chem. Phys. Lett.* **186**, 73–76.

Knickelbein, M. B. (1992), *Chem. Phys. Lett.* **192**, 129–134.

Kresin, V. V. (1992), *Phys. Rep.* **220**, 1–52.

Landman, U., Luedtke, W. D. (1992), *Appl. Surf. Sci.* **60/61**, 1–12

Leisner, T., Athanassenas, K., Echt, O., Kandler, O., Kreisle, D., Recknagel, E. (1991), *Z. Phys. D* **20**, 127–129.

Lewis, L. N. (1993), *Chem. Rev.* **93**, 2693–2730.

Littau, K. A., Szajowski, P. J., Muller, A. J., Kortan, A. R., Brus, L. E. (1993), *J. Phys. Chem.* **97**, 1224–1230.

Luo, F., McBane, G. C., Kim, G., Giese, C. F., Gentry, W. R. (1993), *J. Chem. Phys.* **98**, 3564–3567.

Mackay, A. L. (1962), *Acta. Crystallogr.* **15**, 916–918.

Makov, G., Nitzan, A., Brus, L. E. (1988), *J. Chem. Phys.* **88**, 5076–5085.

Markovich, G., Pollack, S., Giniger, R., Cheshnovsky, O. (1993), *Z. Phys. D* **26**, 98–100.

Martin, T. P., Bergmann, T., Göhlich, H., Lange, T. (1991), *J. Phys. Chem.* **95**, 6421–6429.

Miedema, A. R. (1978), *Z. Metall.* **69**, 287–292.

Miehle, W., Kandler, O., Leisner, T., Echt, O. (1989), *J. Chem. Phys.* **91**, 5940–5952.

Milligan, W. O., Morriss, R. H. (1964), *J. Am. Chem. Soc.* **86**, 3461–3467.

Mingos, D. M. P., Wales, D. J. (1990), *Introduction to Cluster Chemistry*, Englewood Cliffs, NJ: Prentice-Hall.

Morse, M. D., Hopkins, J., Langridge-Smith, P., Smalley, R. (1983), *J. Chem. Phys.* **79**, 5316–5328.

Nishioka, H., Hansen, K., Mottelson, B. R. (1990), *Phys. Rev. B* **42**, 9377–9386.

Papanikolas, J. M., Vorsa, V., Nadal, M. E., Campagnola, P. J., Buchenau, H. K., Lineberger, W. C. (1993), *J. Chem. Phys.* **99**, 8733–8750.

Parent, D. C., Anderson, S. L. (1992), *Chem. Rev.* **92**, 1541–1565.

Parks, E. K., Zhu, L., Ho, J., Riley, S. J. (1993), *Z. Phys. D* **26**, 41–45.

Poppa, H. (1984), *Vacuum* **34**, 1081–1095.

Powers, D. E., Hansen, S. G., Geusic, M. E., Puiu, A. C., Hopkins, J. B., Dietz, T. G., Duncan, M. A., Langridge-Smith, P. R. R., Smalley, R. E. (1982), *J. Phys. Chem.* **86**, 2556–2560.

Raghavachari, K. (1990), *Phase Transitions* **24–26**, 61–90.

Ramakrishna, M. V., Friesner, R. A. (1993), *Israel J. Chem.* **33**, 3–8.

Riley, S. J., Parks, E. K. (1987), in: P. Jena, B. K. Rao, S. N. Khanna (Eds.), *Physics and Chemistry of Small Clusters*, New York: Plenum, p. 727.

Röthlisberger, U., Andreoni, W., Parrinello, M. (1994), *Phys. Rev. Lett.* **72**, 665–668.

Schmid, G. (1992), *Chem. Rev.* **92**, 1709–1727.

Selegue, T. J., Moe, N., Draves, J. A., Lisy, J. M. (1992), *J. Chem. Phys.* **96**, 7268–7278.

Sevov, S. C., Corbett, J. D. (1993), *Science* **262**, 880–883.

Siegel, R. W. (1993), *Physics Today* **46** (10), 64–68.

Spence, J. C. H. (1988), *Experimental High-Res-*

olution Electron Microscopy, Oxford: Oxford University Press, chap. 6.

Steigerwald, M. L., Alivisatos, A. P., Gibson, J. M., Harris, T. D., Kortan, R., Muller, A. J., Thayer, A. M., Duncan, T. M., Douglass, D. C., Brus, L. E. (1988), *J. Am. Chem. Soc.* **110**, 3046–3050.

Steigerwald, M. L., Brus, L. E. (1990), *Acc. Chem. Res.* **23**, 183–188.

Stringari, S. (1990), in: G. Scoles (Ed.), *The Chemical Physics of Atomic and Molecular Clusters*, Amsterdam: North-Holland, p. 199.

Su, C.-X., Hales, D. A., Armentrout, P. B. (1993), *J. Chem. Phys.* **99**, 6613–6623.

Syage, J. A. (1992), in: J. D. Simon (Ed.), *Ultrafast Spectroscopy in Chemical Systems*, Boston: Kluwer Academic.

Taylor, K. J., Pettiette-Hall, C. L., Cheshnovsky, O., Smalley, R. E. (1992), *J. Chem. Phys.* **96**, 3319–3329.

Teo, B. K., Zhang, H. (1990), *J. Cluster Sci.* **1**, 155–187.

Teo, B. K., Zhang, H., Kean, Y., Dang, H., Shi, X. (1993), *J. Chem. Phys.* **99**, 2929–2941.

Tiggesbäumker, J., Köller, L., Lutz, H. O., Meiwes-Broer, K. H. (1992), *Chem. Phys. Lett.* **190**, 42–47.

Toennies, J. P. (1990), in: G. Scoles (Ed.), *The Chemical Physics of Atomic and Molecular Clusters*, Amsterdam: North-Holland, p. 597.

Ugarte, D. (1993), *Z. Phys. D* **28**, 177–181.

Uppenbrink, J., Wales, D. J. (1992), *J. Chem. Phys.* **96**, 8520–8534.

Uppenbrink, J., Wales, D. J. (1993), *Z. Phys. D* **26**, 258–260.

von Helden, G., Hsu, M.-T., Kemper, P. R., Bowers, M. T. (1991), *J. Chem. Phys.* **95**, 3835–3837.

Wang, C. R. C., Pollack, S., Dahlseid, T. A., Koretsky, G. M., Kappes, M. M. (1992), *J. Chem. Phys.* **96**, 7931–7937.

Wang, Y., Herron, N. (1991), *J. Phys. Chem.* **95**, 525–532.

Wei, S., Guo, B. C., Purnell, J., Buzza, S., Castleman, A. W., Jr. (1992), *Science* **256**, 818–820.

Whetten, R. L., Hahn, M. Y. (1990), in: E. R. Bernstein (Ed.), *Atomic and Molecular Clusters*, Amsterdam: Elsevier, chap. 8.

Wood, D. M. (1981), *Phys. Rev. Lett.* **46**, 749.

Yang, Y. A., Bloomfield, L. A., Jin, C., Wang, L. S., Smalley, R. E. (1992), *J. Chem. Phys.* **96**, 2453–2459.

Zakin, M. R., Cox, D. M., Brickman, R. O., Kaldor, A. (1989), *J. Phys. Chem.* **93**, 6823–6827.

Further Reading

Haberland, H. (Ed.) (1994), *Clusters of Atoms and Molecules*, Berlin: Springer-Verlag.

Schmidt, R., Lutz, H. O., Dreizler, R. (Eds.) (1992), *Nuclear Physics Concepts in the Study of Atomic Cluster Physics*, Berlin: Springer-Verlag.

Scoles, G. (Ed.) (1990), *The Chemical Physics of Atomic and Molecular Clusters*, Amsterdam: North-Holland.

Scoles, G. (Ed.) (1988), *Atomic and Molecular Beam Methods*, Vol. 1, New York: Oxford University Press.

Sugano, S. (1991), *Microcluster Physics*, Berlin: Springer-Verlag.

MOLECULAR AND ATOMIC COLLISION PROCESSES WITH ION BEAMS

REINHARD BRUCH, ELIZABETH RAUSCHER, AND STEPHAN FÜLLING, *Department of Physics, University of Nevada, Reno, Nevada, U.S.A.*

DIETER SCHNEIDER, *Lawrence Livermore National Laboratory, Livermore, California, U.S.A.*

SVEN MANNERVIK AND MATS LARSSON, *Manne Siegbahn Institute of Physics, Stockholm, Sweden*

Introduction 438
1. **An Overview of Collision Processes of Molecular and Atomic Ion Beams** 438
2. **Experimental Methods Using Molecular and Atomic Ion Beams** 442
2.1 Typical Ion Accelerator Setup 442
2.2 Beam-Foil Excitation 442
2.2.1 Experimental Techniques 443
2.2.2 Lifetime Measurements 443
2.2.3 Spectroscopy of Ions 444
2.2.4 Quantum Beats 444
2.2.5 Molecular Effects 444
2.3 Laser Excitation 445
2.3.1 Atomic and Molecular Spectroscopy 445
2.3.2 Lifetime Measurements 446
2.3.3 Photofragment Spectroscopy 446
2.4 Spectroscopy and Dynamics of Molecular Ions 447
2.4.1 Experimental Techniques 447
2.4.2 Structure of Molecular Ions 448
2.4.3 Dynamics of Molecular Ions 449
2.5 Photodissociation and Half Collisions of Molecular Ions 449
2.6 Photodetachment 450
2.7 Delta-Electron and Other Secondary Emission Processes in Fast Heavy-Ion–Atom Collisions 451
2.7.1 Introduction 451
2.7.3 Double-Differential Cross Sections 452
2.8 Auger-Electron Spectroscopy Following Ion-Atom and Ion-Molecule Collisions 453
2.8.1 Forward-Angle Projectile Auger-Electron Spectroscopy 453
2.8.2 Auger-Electron Angular Distributions 454
2.9 Photon Emission Processes in Fast Projectile Collisions 454
2.9.1 Multielectron Transitions 454
2.9.2 EUV Spectroscopy 455
2.9.3 Ionization-Excitation of He 457
2.9.4 Excitation of Molecular Gases 457
2.9.5 Electron Capture to Rydberg States 458
2.10 Collisions with Slow Multiply Charged Ions 458
2.10.1 Introduction 458
2.10.2 Charge-Transfer Reactions 459
2.10.3 Electron-Beam Ion Traps: New Facilities for Collision Experiments with Heavy, Highly Charged Ions 460
2.10.4 Ion–Surface Interaction 460
2.10.5 Ion Storage Rings: A New Tool for Atomic and Molecular Collision Physics 462
3. **Theoretical Models for Atomic and Molecular Collision Processes** 464
4. **Summary** 467
Glossary 468
Works Cited 468

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

Research in the field of collision processes of molecules and atoms has developed at an explosive rate since the early 1960s (Rudd and Macek, 1974; Kessel and Fastrup, 1974; Bashkin, 1976; Crasemann, 1975; Sellin, 1978; Mott and Massey, 1987; Fano and Rau, 1988; Alonso *et al.*, 1989). The knowledge of such collision processes has been of great value in numerous applications from pure to applied physics, in particular plasma diagnostics, condensed matter physics, gas laser design, x-ray lasers, physics and technology of ion sources and accelerators, optics of charged particle beams, surface, thin film, and materials analysis, ion implantation, atmospheric collision physics and chemistry, radiation physics and biology, and astrophysics. In order to understand the physics of planetary and stellar atmospheres, including those of the Earth and the sun, detailed knowledge is required of a number of molecular and atomic collision processes. In fusion energy research, molecular and atomic collision physics (De Heer, 1982) plays a significant role in the heating, cooling, and loss mechanisms, as well as in diagnostics. Consequently, a fundamental knowledge of molecular and atomic collision processes (CIAMDA, 1987) is necessary for modeling of high-temperature plasmas (Vainshtein and Shevelko, 1986). Such collision processes occur when molecules and atoms are ionized and accelerated and allowed to interact with atomic, molecular, liquid, surface, or solid targets. The target can also comprise electrons, ions, or photons.

In any ion-atom or ion-molecule interaction, if no excitation of the target occurs, the process is called an elastic collision. Elastic atomic collisions are of primary importance at very low energies (Scoles, 1988). Inelastic collisions may take place if the collision partners come close enough together for atomic or molecular excitation or ionization to occur. As the projectile velocity is increased, first the outer-shell electrons and then the inner-shell electrons are affected and participate in the dynamical process. Inelastic collision events are characterized by the emission of interaction products, such as photons, electrons, scattered ions, or molecular fragments. The use of ion beams for the study of collision processes of molecules and atoms involves many different reaction channels. For readers not familiar with the field of ion interactions and ion-beam techniques, these topics are discussed in general terms in Sec. 1. A more detailed description of individual high-, medium-, and low-energy techniques is provided in Sec. 2, and in Sec. 3, we present briefly some basic atomic and molecular theoretical models.

1. AN OVERVIEW OF COLLISION PROCESSES OF MOLECULAR AND ATOMIC ION BEAMS

Ion-molecule and ion-atom collision processes are controlled by many parameters including ion velocity or energy, projectile and target size, composition, charge state, and mass, interaction time, etc. Useful parameter estimates and simple equations relevant for collision physics are defined in Table 1. A schematic illustration of a typical collisional process characterized by different emission products and reaction channels is shown in Fig. 1. When the projectile beam interacts with a gas, vapor, foil, or solid target, a number of observable elementary reactions occur. These processes involve, for example, photon and electron emission, recoil ions, scattered projectiles, and fragments. Typical detection systems used are optical monochromators, electron spectrometers, time-of-flight analyzers, translational energy spectrometers, lasers for molecular beam fragmentation, precision atomic spectroscopy, photodetachment of negative ions, polarization and angular distribution studies, and multiparameter coincidence techniques. One of the major goals of these collision studies is to understand the many-body aspect of the interaction dynamics, which supersedes independent-particle models. Generally, the most important forces are the Coulomb forces. In few-electron heavy ions at relativistic velocities, quantum electrodynamic (QED) effects must be taken into account. Also, relativistic effects occupy an important role in the collision dynamics and spectroscopy of electrons and photons at low velocities, if the projectile is heavy and highly charged.

It is convenient to classify collisional processes according to the projectile velocity v and the velocity v_e of the participating elec-

Table 1. Some characteristic properties used in ion-atom collision experiments.

Quantity	Symbol	Formula	Units	Typical values
Beam energy	E		MeV	10^{-2}–10^{3-}
Reduced beam energy[a,b]	ϵ	$545E/M$	eV	10^{-2}–10^{5}
		$20E/M$	a.u.	10^{-3}–10^{4}
Velocity[a,c]	v	$\sqrt{40E/M}$	a.u.	10^{-2}–10^{2}
		$13.8\sqrt{E/M}$	mm/ns	10^{-1}–10^{2}
Time of flight	t	$72\sqrt{M/E}$	ps/mm	10^{-11}–10^{-6}
Time resolution	Δt	$\Delta z/v$	s	10^{-12}–10^{-9}
Collision time[d]	T	$2b/v$	s	
Ion-beam current[e]	I		μA	10^{-6}–10^{2}
Particle rate	$\Delta N/\Delta t$	I/g	Ions/s	10^{7}–10^{15}
Gas-target density[f]	n	$10^{17}p/3$	Particles/cm^3	10^{14}–10^{17}
Mean free path length[f]	l	$3 \times 10^{-17}/(p\sigma)$	cm	
Cross section for single electron capture[g]	$\sigma_{q,q-1}$	πR_C^2	cm^2	10^{-14}–10^{-26}
Single-collision condition[f]		$n\sigma l \ll 1$		

[a] Beam energy E (MeV), projectile mass M (amu), nonrelativistic velocities.

[b] 1 a.u. = 2 Ry = 27.2 eV.

[c] Bohr velocity $v_0 = 2.18765\ 10^8$ cm/s.

[d] Impact parameter b.

[e] Ion current for singly charged ions.

[f] Gas pressure p (Torr), cross section σ (cm^2).

[g] R_C = critical distance (1 Å = 10^{-8} cm), only valid for slow collisions.

trons. The low-energy regime is characterized by $v < v_e$. In this region, the semiempirical approximation is appropriate, which treats the heavy particles classically and the electron motion quantum mechanically. Excitation cross sections increase with v and reach a maximum at about $v/v_e \approx 1$. At higher velocities, $v > v_e$, perturbation theory is used to describe scattering amplitudes. In this case, the interaction between electrons and nuclei

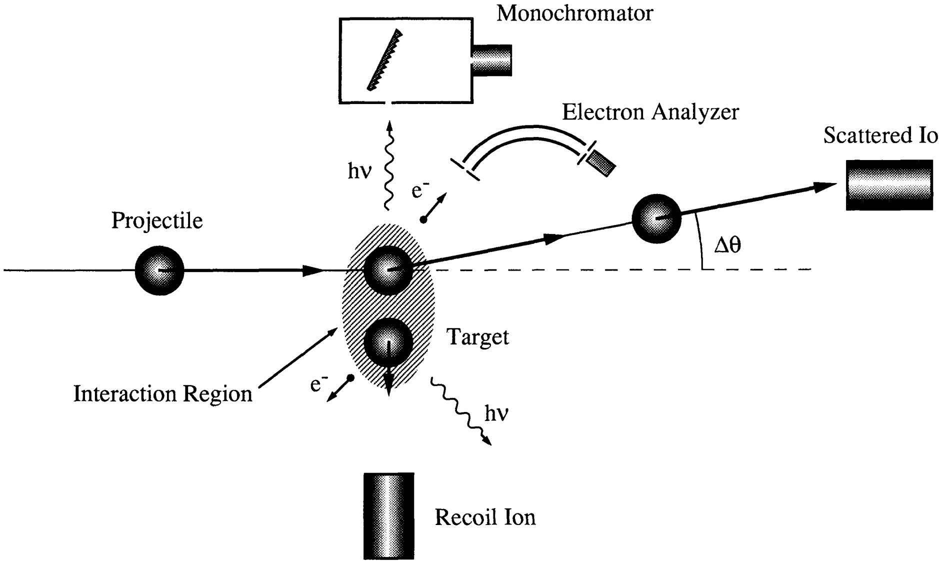

FIG. 1. View of the target region with characteristic ion-molecule and ion-atom interaction processes and reaction fragments (electrons, photons, scattered projectiles, and recoil ions) and corresponding detection systems.

Table 2. Overview of collision processes.

Collision process	Particles detected	Experimental technique
Beam-foil interaction: $A^{q+} + \text{C foil} \rightarrow A^{m+*}$ Molecular fragmentation: $AB^{+} + \text{C foil} \rightarrow A^{+} + B$	Photons (visible, UV, EUV, x rays), electrons (autoionization, Auger effect, continuum electrons), scattered ions, molecular fragments.	*Multiple collision process* (formation of highly excited states): Beam-foil spectroscopy, lifetime measurement, quantum beats, polarization, field ionization, charge state measurement, Auger-electron spectroscopy, molecular dissociation, spectroscopy of cusp electrons.
Beam-laser interaction: $A^{q+} + h\nu \rightarrow A^{q+*}$ Photodetachment: $A^{-} + h\nu \rightarrow A + e^{-}$ Photofragmentation: $AB^{+} + h\nu \rightarrow A^{+} + B$ (or $A + B^{+}$)	Fluorescence studies (infrared, visible, UV photons), electrons, atoms, ions, molecular fragments.	*Laser excitation*: Laser spectroscopy of fast beams, Doppler-free methods, laser cooling, collinear laser ion-beam method, photodetachment, photoionization, photofragmentation spectroscopy, half collisions.
Beam surface interaction: $A^{q} + \text{surface} \rightarrow A^{m+*}$ Beam-solid interaction: $A^{q} + \text{solid} \rightarrow A^{m+*}$	Photons (visible, UV, EUV, x rays), Auger electrons, scattered ions.	*Ion-surface collision* (angle of ion impact from normal to grazing incidence): X-ray and Auger-electron spectrometry, ion-beam crystallography, ion implantation, Rutherford backscattering, particle-induced x-ray emissions, channeling.
Electron capture: $A^{q+} + B \rightarrow A^{(q-1)+*} + B^{+}$ Ionization: $A^{q+} + B \rightarrow A^{q+} + B^{+} + e^{-}$ Transfer ionization: $A^{q+} + B \rightarrow A^{(q-1)+} + B^{2+} + e^{-}$ General case: $A^{q+} + B \rightarrow A^{(q-s)+} + B^{r+} + (r-s)e^{-} + \Delta E$	Photons (visible, UV, EUV, x rays), electrons (Auger decay, autoionization, continuum electrons, cusps), and scattered ions.	*Single collision process*: Charge-state, energy, and angle selection of projectile and target ions, impact-parameter–dependent coincidence measurements, photon and electron emission studies, translational spectroscopy, accel-decel method, recoil-ion spectroscopy, zero-degree electron spectroscopy, EUV and x-ray studies, dissociative charge transfer.
Few (mostly two) electron transitions: Double ionization: $A^{q+} + B \rightarrow A^{q+} + B^{2+} + 2e^{-}$ Double excitation: $A^{q+} + B \rightarrow A^{q+} + B^{**}$ Ionization plus excitation: $A^{q+} + B \rightarrow A^{q+} + B^{+*} + e^{-}$	Projectile and target photons, electrons and ions.	*Single collision process* (medium to high energy): Charge state, photon and electron emission cross section measurements, photon and electron angular distribution.

Information obtained	Theoretical model	Review articles
Multielectron effects, energy levels of highly excited states, lifetimes, casacade effects, spin-forbidden transitions, fine and hyperfine structure, Lamb shift, magnetic substate population, coherence parameters, charge-state distribution, energy-loss mechanisms, molecular properties, photon yields, population of Rydberg levels.	Many-body theories, atomic structure calculations, relativistic and QED effects, time-dependent phenomena, Auger decay, ion-solid and ion-surface theory, molecular dissociation.	Andrä (1979), Bashkin (1976), Berry (1977), Betz (1983), Cowan (1981), Gemmel (1980), Lindgren and Morrison (1982), Mohr (1989), Sellin (1978).
Accurate measurements of energy levels, fine and hyperfine splitting, Lamb shift, lifetimes, resonance widths, photodetachment and molecular fragmentation channels and energies, formation of ordered ion structures.	Precision calculations of atomic and molecular properties, excitation and decay dynamics of selectively excited states, quantum electrodynamics, quantum optics, time-dependent phenomena.	Andrä (1979), Hotop and Lineberger (1985), Letokov (1989), Müller-Dethlefs and Schlag (1991), Neugart (1987), Walther (1992).
Photon and electron spectra, ion scattering, charge state of ions, secondary-ion yield, surface modification, short-wavelength lasers, stopping power, ion-beam modification, surface composition and structure.	Deexcitation and neutralization processes near surfaces, ion scattering, secondary-electron, ion, and photon production, cascades, optical polarization, charge-exchage mechanisms, surface modifications, radiation damage.	Andrä (1989), Datz (1983), Elton (1990), Hasselkamp *et al.* (1992), Marrus (1989), Tully and White (1977).
Total and differential capture or excitation cross sections (inner- and outer-shell contributions), state-selective photon- and electron-emission cross sections, polarization, angular distributions, multiple-electron transitions, threshold effects, atomic and molecular data for fusion, astrophysics, x-ray lasers.	*Low energy*: Overbarrier model, Landau-Zener theory, post-collisional interaction, molecular orbital picture. *Intermediate and high energy*: Perturbative methods: Born, eikonal, Glauber, distorted-wave, higher-order contributions. Non-perturbative methods: Close coupling, R- and T-matrix approaches. Statistical methods. Hartree-Fock methods.	*Low energy*: Marrus (1989), Niehaus (1990), Ivanov (1988), Salzborn *et al.* (1991). *Medium to high energy*: Briggs *et al.* (1988), Cocke and Olson (1991), Dubé (1986), Fritsch and Lin (1991), Rudd and Macek (1974), Stolterfoht (1987).
Multielectron effects and electron correlation, kinematics in several-body scattering, ionization of many electrons, multiple excitation, ionization plus excitation, multiple-electron capture, transfer, and ionization, transfer and excitation, few-electron transitions, differential cross sections, Fano shape profiles.	Independent-electron approximation, expansion in projectile charge (Z_p) and velocity (v), 1st- and 2nd-order terms, higher-order terms, many-body perturbation expansion, sudden approximation, forced impulse method, coupled-channel calculations, time-dependent Hartree-Fock, binomial distributions, multiple-center scattering.	Datz (1990), McGuire (1992).

occurs during an extremely short time interval (for 1-MeV H^+, $\Delta t \approx 10^{-17}$ s), and high fields are obtained leading to inner-shell excitation and ionization. Currently, ions with energies of GeV/u (1 u = atomic mass unit) can be employed using relativistic velocities and the ability to produce few-electron or bare ions throughout the periodic table. These time-dependent, high-energy, few-body scattering problems provide a challenging theoretical endeavor.

Even though multiply charged ions dominate over neutral species in the universe, accurate experimental data and theoretical models on the structure and dynamics of multiply charged ions in low-, medium-, and high-energy collisions are limited. In particular, few-electron ions are ideal for testing relativistic many-body theories and QED effects (Wilson *et al.*, 1991). Therefore, experiments with highly charged ions are currently of great interest.

Collision processes between a wide range of atomic and molecular species have been investigated using a variety of experimental techniques, and some theoretical models have been developed. Representative collision processes that elucidate the field of molecular and atomic collisions are summarized in Table 2 along with additional information concerning the type of particles detected, experimental techniques used, experimental parameters obtained, and theoretical methods applied. Also listed in Table 2 are some review articles and books covering specific topics and categories of ion-molecule and ion-atom collisions. For more information, the reader is referred to the conference proceedings on the physics of electronic and atomic collisions (ICPEAC; see, for example, Datz, 1990) and the Proceedings of the NATO Advanced Study Institutes on Fundamental Processes of Atomic Dynamics (Briggs *et al.*, 1988). The physics of highly charged ions has recently been reviewed by Marrus (1989).

2. EXPERIMENTAL METHODS USING MOLECULAR AND ATOMIC ION BEAMS

2.1 Typical Ion Accelerator Setup

Tunability in energy, mass, charge state, current, repetition rate, etc. makes ion accelerators powerful tools for the exploration of atomic and molecular structure and collision dynamics. Several informative reviews and conference proceedings exist that cover practical aspects of ion accelerators, e.g., optics of charged particles (Wollnik, 1987), principles of charged-particle acceleration (Humphries, 1986), the physics and technology of ion sources (Brown, 1989), ion beams for material analysis (Bird and Williams, 1989), and the excellent Denton Conference proceedings on "Applications of Accelerators in Research and Industry" published periodically in *Nuclear Instruments and Methods*. A typical ion-beam collision apparatus consists of a positive- or negative-ion source and a means to accelerate the ions such as a Van de Graaff (2 or 3 MeV) or any other type of electrostatic accelerator (see Fig. 2). The ion beam is usually focused, mass analyzed, and collimated before entering the target chamber, which may contain a foil, gas, vapor, jet, surface, or solid target. Particle detectors may be employed to analyze target or projectile fragments, or a spectrometer may be used to observe the photons, electrons, and ions emitted by excited target or projectile states.

Atomic and molecular collisions at low to medium energy are conducted mostly with electrostatic machines. Recently, radio-frequency boosters have been designed, extending the range to higher energies and charge states and heavier ions. Relatively large-scale accelerators, such as the Kansas State, Oak Ridge, Brookhaven, Berkeley, and Argonne accelerators, and ion storage rings (Aarhus, GSI Darmstadt, Heidelberg, and Stockholm) as well as other large-scale accelerator facilities like ATLAS, BEVALAC, RIKEN, GANIL, UNILAC, etc. are frequently used for atomic and molecular collision studies.

2.2 Beam-Foil Excitation

The beam-foil technique was independently proposed by Kay and Bashkin in the early 1960s. The basic concept is to produce an ion beam in an accelerator and to let the ion beam pass through a thin foil. The ions in the beam will be excited and can be observed by spectroscopic methods. In particular, since the emitted photons stem from fast-moving particles, time-of-flight techniques can be used to measure decay curves and time-delayed spectra. Many review articles and books about this method have been pub-

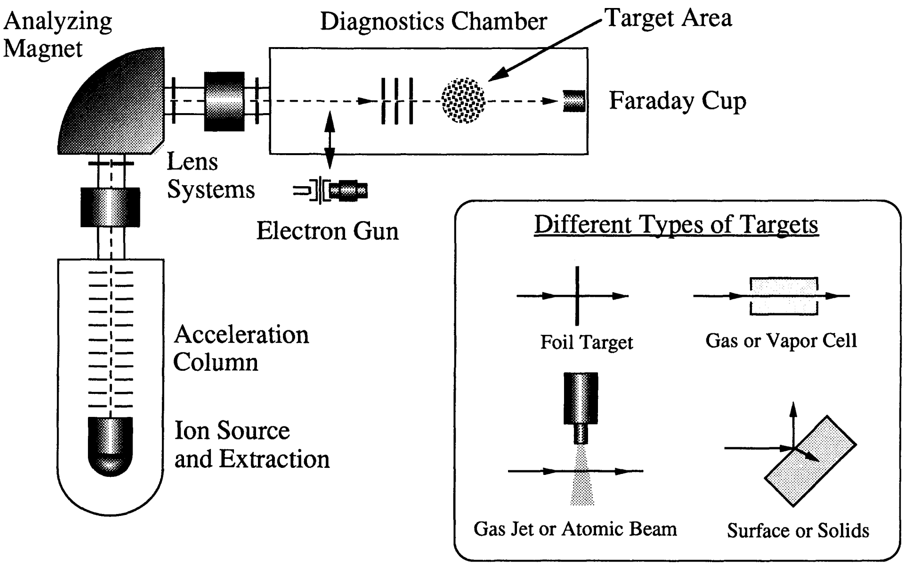

FIG. 2. Schematic layout of a low- to medium-energy ion accelerator with target chamber and different types of possible targets.

lished; see, e.g., contributions by Bashkin (1976), Berry (1977), and Andrä (1979).

2.2.1 Experimental Techniques Early beam-foil experiments were performed mainly at Van de Graaff machines or electrostatic isotope separators. These accelerators deliver typically beam energies between 50 keV and 3 MeV with beam currents of the order of 1 μA. Such beam energies correspond to velocities of a few mm/ns (see Table 1). As more powerful accelerators have been constructed, beam-foil experiments have more recently been performed at much higher energies (many GeV). Two illustrative examples are given here. For example, a 250-keV He beam will mainly leave the foil in charge state +1, while a 150-MeV U beam will have a mean charge of about +32. The excitation process is generally nonselective. It has been found that the population of the different excited Rydberg states usually follows approximately a $1/(n^*)^3$ dependence where n^* is the effective principal quantum number (Stebbings and Dunning, 1983). When the ion beam traverses the foil, it interacts with many atoms, causing energy loss and energy spread. These kinematic effects often have to be taken into account as spectroscopic measurements are evaluated. As an example, the stopping power dE/dx for 100-keV H^+ ions interacting with C foils is about 0.7 keV/(μg cm^{-2}).

In a typical experimental setup for beam-foil measurements, either the photon or electron emission spectrum is observed close to the foil, or the intensity decay of a specific spectral line is measured by increasing the distance between the excitation (foil) and the observation region (usually by moving the foil upstream in the vacuum chamber). The distance is converted to time through the simple relation $t = x/v$, where x and v are the distance and beam velocity after foil excitation, respectively. Measurements are normalized to a specific particle flux by monitoring either the total light emitted at the foil or the ion-beam current in a Faraday cup. An optical spatial resolution of the order of 0.1 mm is usually sufficient to measure atomic lifetimes. The high velocity of the ion beam generally causes large Doppler broadening and shifts. Even for a perpendicular observation, which eliminates the Doppler shift in first order, there will be a large Doppler broadening as a result of the angular spread of the monochromator. To a large extent, this can be compensated for by moving the exit slit of the monochromator ("refocusing"). This method works for photon as well as electron spectroscopy.

2.2.2 Lifetime Measurements There is a great need for experimental measurements of atomic transition probabilities. Such measurements are important for astrophysics, plasma diagnostics studies, and tests of many-body theories. The atomic lifetime (τ) associated with a specific state is closely related to the transition probabilities (A_{ki}) by

$$\tau_i = 1/\Sigma A_{ik}, \tag{1}$$

where the sum involves all lower states k to which the state i can decay. Generally, beam-foil lifetime curves are composed of more than one exponential profile as a result of cascades from higher-lying states. The differential equation for the decay process is

$$\frac{dN_i}{dt} = -\frac{N_i}{\tau} + \Sigma N_j A_{ji}, \tag{2}$$

where the last term describes the repopulation of the state i by direct cascades. When the lifetimes of the cascading transitions are very different from the primary lifetime to be measured, it is generally possible to fit the decay curve by a multiexponential function. However, if this is not the case a more careful cascade analysis is needed. The most frequently used method is the so-called ANDC method (arbitrarily normalized decay curves). In this method, the decay curves of all direct cascades are experimentally measured. Another problem to handle in the extraction of lifetimes from a decay curve is to determine the beam velocity. The beam energy and the energy spread after the passage of the foil can be measured by an energy analyzer. In favorable cases, accuracies for lifetimes of the order of 1% have been achieved by the beam-foil method.

2.2.3 Spectroscopy of Ions The beam-foil light source has additional advantages compared to other, more conventional, light sources. First, the ion beams used are isotopically pure. Second, a specific beam energy can be chosen to optimize a specific charge state. In principle, all charge states of any element are presently available with accelerators in operation today. Third, the time-resolved character of the light source can also facilitate line identification. Spectroscopic studies using the beam-foil method have been performed all over the periodic table (Bashkin and Stoner, 1975, 1978, 1981). In particular, access to heavy-ion accelerators has recently made it possible to study relativistic interactions and QED effects in H-like and He-like uranium.

It was also recognized that the beam-foil interaction process can efficiently produce exotic multiply excited states that lie in the continuum of the atomic system (i.e., above the lowest ionization limit) and may decay by autoionization. However, there exist some multiply excited states that are not allowed to decay via autoionization because of specific selection rules. These states instead decay predominantly to lower states by radiative transitions. Much experimental and theoretical work has been devoted to the study of such exotic systems (see Mannervik, 1989). Some of the most spectacular findings include the optical emission lines from multiply excited states in the negative ions of lithium and beryllium (see also Sec. 2.6). The beam-foil method is also ideally suited to the study of spin- or angular-momentum–forbidden transitions—for instance, intercombination lines. "Forbidden" means in this context that the transition is much less likely to occur. For the higher members of an isoelectronic sequence, relativistic interactions become more pronounced, and hence spin-forbidden transitions become more likely. Thus, by observing time-delayed photon or electron spectra most prompt-decaying states will be suppressed. Using the beam-foil method, such time-delayed spectra are easily obtained by measuring the photon or electron emission from the projectile beam some distance behind the foil so that a certain time has elapsed after excitation (time-of-flight method).

2.2.4 Quantum Beats Unresolved multiplet transitions can in some cases be studied via quantum beat measurements. Since the excitation process is coherent, a superposition of fine-structure or hyperfine states is produced. The decay curve for such a superposition of states can exhibit interference patterns similar to that from a double slit where the beat frequencies correspond directly to the splitting of the levels. For some light elements, it has been possible to determine fine- and hyperfine-structure splittings with an accuracy even below 1%.

2.2.5 Molecular Effects Furthermore, experiments have been performed with primary beams consisting of molecular ions. As the ions pass the foil, electrons will be stripped off, and the ion fragments will repel each other by Coulomb interaction ("Coulomb explosion"). For example, energies and angular distributions of these fragments can be measured. The electrostatic potential energy stored in such molecular ions is of the order of eV, and because of the high velocity of the ion beam transforms to higher energies in the laboratory system (typically keV). These energies can be measured with high accuracy to study internuclear distances in molecular ions. In these types of experiments, the response of the solid also has been seen; in particular, a phenomenon called the "wake effect" has been observed (Hasselkamp *et al.*, 1992).

2.3 Laser Excitation

In Sec. 2.2, excitation of fast-ion beams by thin foils was discussed. Problems with that technique concern the nonselective excitation mechanism and also substantial energy loss and the energy and angular spread resulting from multiple collisions in the foil. A method to circumvent these problems is to excite the beam with laser photons directly. With laser excitation, much higher spectral resolution can be obtained (typically 10^{-3} Å) and for lifetime measurements, because of the selective excitation process, a cascade-free decay is observed. Besides the spectroscopic applications, laser excitation of fast-ion beams can be used for other purposes. For example, excitation of an ion beam with a laser is an excellent tool to determine accurately beam velocities and velocity distributions of the ions. As a laser photon is absorbed by an ion, momentum is transferred to the projectile. If this process is repeated many times, the velocity distribution of ions can be changed (laser cooling). This method has been used to narrow velocity distributions in fast- and slow-ion beams and ion storage rings (Sec. 2.10.5). Review articles about beam-laser excitation and applications have been published by, e.g., Andrä (1979) and Neugart (1987). Laser photoionization spectroscopy and lasers in atomic, molecular, and nuclear physics are treated in the book by Letokov (1989).

2.3.1 Atomic and Molecular Spectroscopy

For excitation with a narrow-bandwidth single-mode laser, the spectroscopic resolution is limited by the ion-beam properties and geometry. The linewidth $\Delta\nu$ is determined by the relation (including first-order terms)

$$\Delta\nu = \frac{v}{c}\,\nu(\Delta\theta + \Delta\phi)\sin\theta + \frac{\Delta v}{c}\,\nu\cos\theta + \frac{v}{\pi\Delta x}\sin\theta, \tag{3}$$

where v is the ion velocity and c is the speed of light. The angle θ and the angular divergences $\Delta\theta$ and $\Delta\phi$ are illustrated in Fig. 3. From Eq. (3), it is seen that for the collinear geometry ($\theta = 180°$) all terms except the second will become zero. For a fast beam, there is an effect called "velocity compression" that reduces Δv. Thus, for a certain energy spread ΔE in the ion source, and corresponding velocity spread Δv_i, the resulting velocity spread Δv_a after acceleration is reduced, i.e., $\Delta v_a \ll \Delta v_i$. Therefore, in the longitudinal direction an accelerated beam will appear as a very cold ion ensemble. For the collinear geometry, the resulting Doppler shift

$$\nu_0 - \nu = \frac{v}{c}\,\nu_0\cos\theta \tag{4}$$

will reach its maximum. Such shifts can, however, be corrected for if the ion velocity is accurately known. If the laser resonance frequency is measured both for parallel (ν_+) and antiparallel (ν_-) laser beams, an accurate measurement of the velocity can be per-

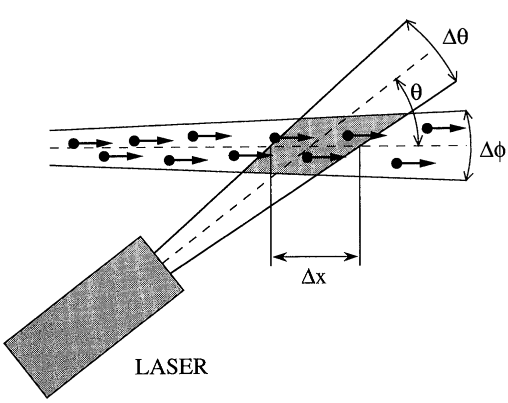

FIG. 3. Schematic illustration of typical ion beam–laser beam interaction region. $\Delta\phi$, $\Delta\theta$, and Δx are the ion-beam divergence, laser-beam divergence, and interaction length, respectively.

formed according to

$$\nu = c(\nu_+ - \nu_-)/(\nu_+ + \nu_-). \tag{5}$$

Specifically, the collinear geometry has been used to perform precision measurements of the Lamb shift for one- and two-electron systems, of hyperfine structure splitting of atoms and ions, and of vibrational–rotational structure of molecular species (Sec. 2.4). A typical experimental setup is shown in Fig. 4, where the excitation region can be shifted through a local change of the ion velocity to maximize the fluorescence signal at the detector, which usually consists of a broad-band monochromator or an interference filter in conjunction with a photomultiplier.

2.3.2 Lifetime Measurements If the excitation region is made very narrow when compared to the mean decay length of a specific projectile state, cascade-free lifetime measurements can be performed. This is usually accomplished in a crossed-beam geometry. To enhance the laser power and interaction time, intracavity excitation is often used; i.e., the ion beam crosses the laser beam many times inside the laser cavity. By rapid Doppler shifting, it is also possible to restrict the excitation region to a narrow region even for a collinear geometry.

2.3.3 Photofragment Spectroscopy Laser spectroscopy of fast molecular ions is also an important field. A very attractive method is to excite the molecular ion to a dissociative state. The fragments can then be collected with a very high efficiency, since they travel in the forward direction and can be separated by an electrostatic analyzer and collected by a particle detector (see Fig. 4). A fast beam will act as an amplifier and transform a small released energy to an energy that is relatively easy to measure. For a diatomic molecular ion, the fragment will, in the forward direction, have a kinetic energy

$$E_f = \tfrac{1}{2}E_0 \pm \sqrt{E_0 W} + \tfrac{1}{2}W, \tag{6}$$

where E_0 is the initial energy and W is the energy release in the dissociation process. For excitation in a collinear geometry, the effect of velocity compression discussed above (Sec. 2.3.1) corresponds to spectroscopy of very cold

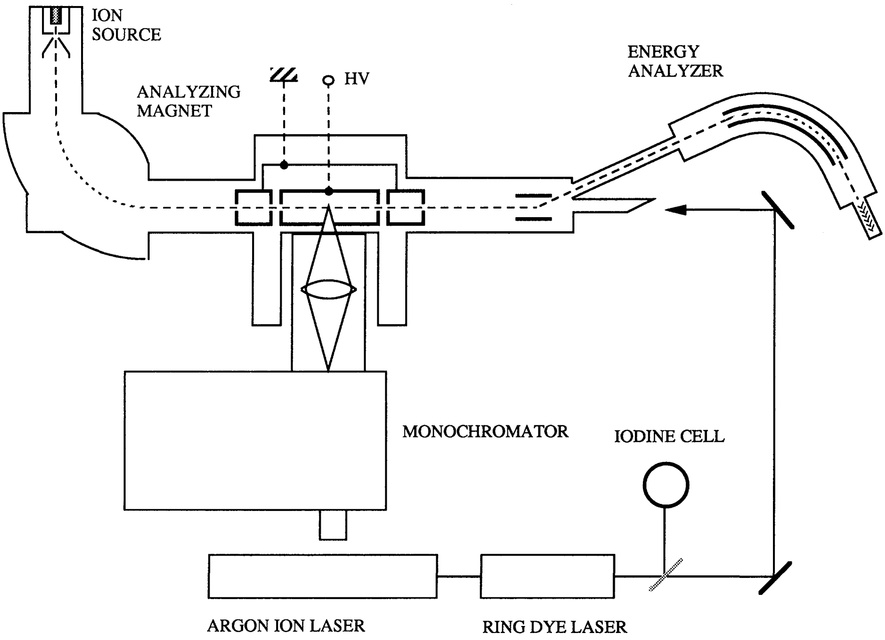

FIG. 4. Experimental setup for collinear ion beam–laser spectroscopy, lifetime measurements, and molecular-ion photofragment spectroscopy. Photoproduced fragment ions are deflected into an electrostatic energy analyzer. Moreover, fluorescence photons can be detected by an optical monochromator, equipped with a photomultiplier.

ions with small Doppler width (sub-Doppler spectroscopy). In Sec. 2.4, the application of this technique is discussed.

2.4 Spectroscopy and Dynamics of Molecular Ions

Molecular ions are major chemical agents in many crucial processes that occur in physics, chemistry, astronomy, and biology. Despite this fundamental importance, the structure and properties of neutral and charged molecules in the gas phase were for a long time poorly understood, and the rich chemistry and physics of molecular ions were the almost exclusive domain of mass spectroscopists. While mass spectroscopy can give very useful information about the existence and stability of molecular ions, it provides little information about their structures. In recent years, a number of more sensitive techniques to examine molecular structure have been developed and applied to these species. Most of these techniques involve the absorption or emission of photons or the emission of electrons and fall into the general area of spectroscopy. There has also been a rapid development in *ab initio* quantum mechanical methods during the last decades, which, in combination with the enormous enhancement in computer power, has led to a close interplay between experiment and theory. Several books and reviews that give in-depth coverage of the general area of spectroscopy and dynamics of molecular ions have been published (Miller and Bondybey, 1983; Gudeman and Saykally, 1984; Maier, 1989; Müller-Dethlefs and Schlag, 1991).

2.4.1 Experimental Techniques Molecular spectroscopy is traditionally and naturally divided into electronic, vibrational, and rotational spectroscopy. Vibrational and rotational spectroscopy are the more direct routes to molecular structure, while electronic spectroscopy also reveals information about excited electronic states. Most simple stable molecules have been studied by means of these spectroscopies. Two difficulties arise when these methods are applied to molecular ions. First, ions are generally, because of their high reactivity, present in very low concentrations. This means that highly sensitive techniques are required to detect them. Second, neutral molecules are usually present among the ions in much larger abundance. Hence, neutral absorption or emission spectra usually overwhelm those from the charged molecules.

The first obstacle discussed above can be overcome by using lasers as light sources (see Sec. 2.3) in absorption experiments. In the infrared region (1–100 μm), a variety of systems are in use, such as difference-frequency laser systems, color-center lasers, diode lasers, and CO_2 lasers. Spectroscopic transitions occurring in the infrared region are generally those between vibrational levels within a given electronic state (vibrational spectroscopy). In the visible and UV regions, dye lasers have dominated for a long time but have in recent years been replaced by solid-state lasers in some wavelength regions (most notably titanium sapphire for the 700–1100-nm region). Spectroscopic transitions occurring in the visible and UV regions are generally those between different electronic states in a molecule (electronic spectroscopy). Spectroscopy in the microwave region (~4–600 GHz), which is not laser based, has advanced by the availability of monochromatic sources. Spectroscopic transitions occurring in the microwave region are generally those between rotational levels of a given vibronic state of a molecule. Microwave spectroscopy in general fits into the conceptual framework of a classic absorption experiment, with a source, a cell, and a detector. Spectroscopy in the other wavelength regions has most often employed indirect detection schemes in order to increase the sensitivity (see below).

The second obstacle (i.e., that of higher neutral abundance) can be overcome by using ion-beam techniques (Miller and Bondybey, 1983). Figure 4 shows an example of an experimental setup for ion-beam laser spectroscopy. The ions are generally produced by electron-impact ionization/fragmentation of a neutral molecule, extracted by applying a 1–200-kV (in some cases, even as low as 10–230 eV) potential to the ion source, and are collimated into a narrow beam by a series of electrostatic lenses. The advantages with the resultant beam are that unambiguous identification of the ions in the beam can be made by the introduction of a mass analyzer (magnetic or quadrupole) in the beam line, an efficient interaction with the laser beam can be obtained by a collinear arrangement, frequency scanning may be achieved with a fixed-frequency laser line by sweeping the source potential, and, finally, sub-Doppler resolution is obtainable as a result of an effect known as kinematic compression or velocity bunching (see Sec. 2.3). The disadvantage with a fast-ion beam is that ion densities are low, typically 10^4 to 10^6 ions/cm^3. This means that the direct observation of the attenuation of the laser beam by the ion beam is very difficult, although it has been demonstrated in a few cases in the infrared region. In general, however, it is necessary to use indirect methods to detect the absorption of radiation. One such method, which is based on the occurrence of fragment ions as a result of photon absorption, will be discussed in Sec. 2.5; it suffices to mention here that other indirect schemes involve detection of charge transfer that is vibrational-state dependent, and detection of laser-induced fluorescence perpendicular to the ion beam.

Another way to enhance spectra of ions at the expense of neutrals is to pass a laser beam through a cell in which an alternating-current plasma is maintained. The molecular ions are accelerated back and forth by the alternating electric field, thus frequency modulating the absorption line through the Doppler effect. Neutral molecules are essentially unaffected by the electric field and do not exhibit observable Doppler shifts in their absorption spectra. A further advantage with this so-called velocity modulation technique is that ion densities in the plasma are much higher than in ion beams, of the order of 10^{13} ions/cm^3. The technique was developed at Berkeley around 1982–83 (Gudeman and Saykally, 1984) and has yielded information on more than 60 molecular ions (both positively and negatively charged), mainly through studies in the infrared region.

A technique that involves absorption of radiation by the neutral precursor while still revealing information on the molecular ion is photoelectron spectroscopy. This technique was developed already in the 1950s and must now be considered to be a mature field as far as stable neutral molecule precursors are concerned. Briefly, radiation in the VUV or x-ray regions is used to ionize a neutral molecule. The kinetic energy (E_k) of the electron is measured, and from the relation $E_k = h\nu - I$, where $h\nu$ is the photon energy and I is the binding energy of the electron, information on I is obtained. Compared to the techniques described previously, which allow the rotational structure to be resolved, photoelectron spectroscopy in its traditional use is a low-resolution technique (10 meV, or about 100 cm^{-1}). Recently, a technique has been developed (Müller-Dethlefs and Schlag, 1991) that allows a three–order-of-magnitude increase in resolution. The new method is based on the idea that ionization is carried out so that electrons with $E_k = 0$ are created in a field-free region. After a time delay of some μs, the photoelectrons are extracted, and the zero–kinetic-energy (ZEKE) electrons are separated from near-ZEKE electrons by a time-of-flight device. Lasers are used for the initial ionization step.

2.4.2 Structure of Molecular Ions

Structural information derived from vibrational and rotational spectroscopy involves equilibrium rotational constants and corresponding equilibrium moments of inertia, bond angles, vibrational frequencies, etc., which may in some cases lead to a construction of the ground-state potential energy surface. Electronic spectroscopy also contributes to information on the ground state; in addition, it gives information on excited electronic states.

Around 1970, about 30 diatomic ions and 6 polyatomic ions had been studied by high-resolution spectroscopic methods, where high resolution means that the rotational structure is resolved. With the exception of H_2^+, these studies had been carried out with absorption or emission techniques in the UV–visible wavelength region. Today, the number of molecular ions studied in the infrared and microwave regions exceeds 60 and 20, respectively. A large number of ions have also been studied in the UV–visible region during the last 20 years. It is not possible to discuss all these molecules here, and only some characteristic examples will be given.

The infrared absorption spectrum of the simplest triatomic positive ion, H_3^+, was discovered in 1980 by Oka (see Miller and Bondybey, 1983). The difficulties involved in the search for spectra of molecular ions are well illustrated by the fact that the existence of H_3^+ was reported already in 1912. The molecular ion H_3^+ is, because of its relative simplicity, suitable for theoretical treatments. It plays a key role in the chemistry that takes place in the interstellar molecular clouds.

One of the most fundamental processes in chemistry and biology is the proton transfer of liquid water to form a hydronium ion (H_3O^+) and a hydroxide ion (OH^-). The characterization of these ions in the gas phase was therefore a goal for high-resolution spectroscopists for a long time. It was, however, not until the development of the velocity modulation technique that high-resolution spectra of H_3O^+ and OH^- were observed.

The C_2^+ fragment ion has no chemically stable neutral precursor, is the smallest carbon cluster, and is believed to occupy an important role in many terrestrial and extraterrestrial environments. Fragment ions in general have mainly been identified by mass spectrometry. A spectrum of C_2^+ in the visible-wavelength region was recently observed and analyzed (Maier, 1989) by a combination of low-resolution absorption spectroscopy of C_2^+ frozen into a Ne-crystal matrix and high-resolution laser excitation spectroscopy.

2.4.3 Dynamics of Molecular Ions The subject of the dynamics of molecular ions is a vast field because their reaction and scattering processes can be very complex. The close examination of the individual steps of gas-phase reaction processes, i.e., elementary reactions, has resulted in the detection of a large number of short-lived molecules as reaction intermediates. Identification of these species is a prerequisite for understanding the reaction mechanism. The high-resolution spectra of reaction intermediates are very valuable in examining, in real time, distributions over quantum states of molecules taking part in reactions. A typical environment where molecular ions act as reaction intermediates is a laboratory plasma. The ions are initially produced by ionization from electron bombardment and are destroyed by recombination. If the ions are produced in a pulsed discharge, one can measure the infrared absorption signal as a function of time and thereby monitor the population of particular vibrational-rotational states of the ion. If a molecular ion recombines with an electron, a common scenario is $AB^+ + e \rightarrow A + B$, i.e., dissociative recombination. If the laser is tuned to a specific transition in AB^+, the signal will then decrease as a function of time and reflect the rate of destruction of AB^+. It should be mentioned that there are several other techniques suited for studies of dissociative recombination—for example, merged ion-electron beam and flowing afterglow techniques.

Photoionization of a neutral molecule can be regarded as a half collision between an electron and the molecular ion core. The specific features of the ZEKE method make it particularly suited to study the short-range interaction between the photoejected electron and the ion core, the angular distribution of the ejected electron, and coupling of electronic and nuclear motions. There is a clear analogy between photoionization and photodissociation; the latter process is the subject of the next section.

2.5 Photodissociation and Half Collisions of Molecular Ions

Photodissociation of neutral molecules is an important subfield of molecular reaction dynamics (i.e., the study of the molecular mechanism of elementary physical and chemical rate processes). A typical question posed in studies of the dynamics of photodissociation is: What is the fate of a molecule after it has absorbed a photon whose energy exceeds the threshold for dissociation of one or more bonds? The experimental techniques employed for these studies in general involve a neutral molecular beam, a pulsed laser for photodissociation, and a second pulsed laser for probing the internal-state distribution of one of the photofragments.

Photodissociation of molecular ions differs from that of neutral molecules both when it comes to experimental techniques and in terms of information extracted from the experiments (Miller and Bondybey, 1983; Maier, 1989). The technique used to detect photodissociation of molecular ions has already been discussed in Sec. 2.4.1 and is known as ion photofragment spectroscopy (IPS). Figure 5 shows two examples of how photon absorption may induce a bond-breaking process in a molecule. The potential curves, which are the potentials in which the nuclei move, are for a fictive diatomic molecule, but the arguments can easily be generalized to polyatomic molecules. In Fig. 5(a), it is shown how the molecule is promoted by a photon from its electronic ground state to a repulsive state. Since this repulsive state does not support any bound vibrational levels, the molecule dissociates on a time scale of 10^{-13} s. If the cross

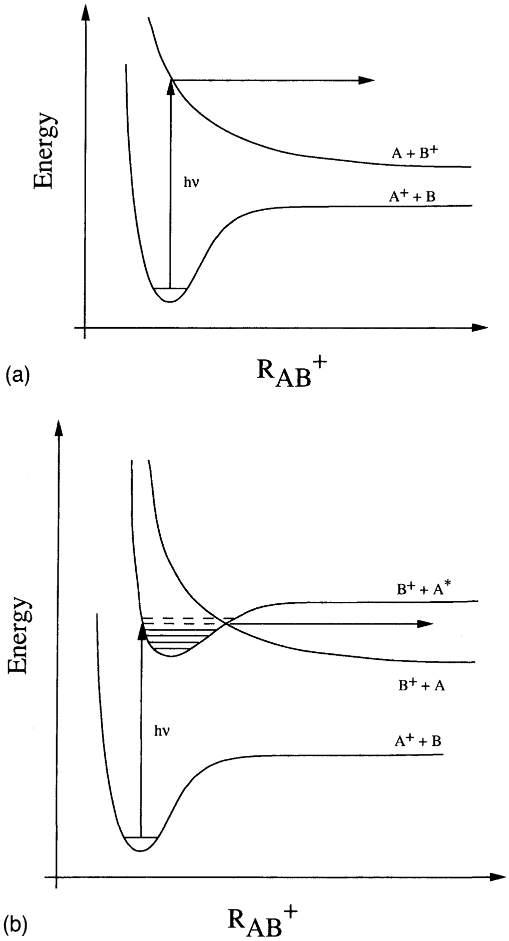

FIG. 5. Schematic potential curves for a diatomic molecule. (a) The process of direct photodissociation where the molecule is promoted by a photon from its ground to a repulsive state, which leads to a breaking of the molecular bond. (b) An indirect electronic predissociation process. The molecule is excited to a bound state that interacts with a repulsive state. This leads to a nonradiative transition to the repulsive state, and the molecule dissociates.

section for photodissociation is measured as a function of photon wavelength, there will be only a slow variation and no structure. The wavelength range over which the molecule photodissociates will depend on the initial state distribution. In Fig. 5(b), the molecule is excited into a bound state that interacts with a repulsive state. This interaction leads to a rearrangement of electronic energy from the bound state to the repulsive state, and the molecule dissociates. The rate-determining step is in this case the nonradiative transition from the bound to the repulsive state (predissociation). Photodissociation only occurs when the laser is tuned into resonance with one of the discrete lines of the transition between the two bound states. The photodissociation cross section as a function of laser wavelength will be structured. Thus, for the molecular spectroscopist IPS serves as a highly sensitive tool to study absorption spectra of molecular ions, often at a spectral resolution of the natural width of the excited state. For those interested in slow-ion–atom collisions, IPS allows the study of half collisions in a well-defined system of reference, exploring the dynamic paths that connect a precisely defined molecular system with an equally precisely defined continuum of separated states.

A photofragment spectroscopy experiment can briefly be described as follows. A highly collimated, mass-selected, fast-ion beam is irradiated coaxially by a single-frequency laser. A transition from a molecular level populated in the ion beam to a level lying above the dissociation limit is induced by tuning the laser (or Doppler tuning the ion beam) into resonance. Provided that the excited level interacts with a continuum state, dissociation will occur, and the photoinduced ionic or neutral fragments can be detected, often with nearly unity detection efficiency. The laser wavelength provides information about the transition frequency, the measured linewidth reveals the upper-level lifetime toward predissociation, and measurement of the released kinetic energy imparted to the fragments reveals the location of the excited level with respect to the dissociation limit. If the transition takes place directly to a repulsive state, measurement of the angular distribution of fragment ions gives information on what type of electronic transition is involved and the product-state distributions. In this latter case, it is necessary to irradiate the ion beam perpendicular, so that the plane of polarization of the laser beam can be varied. An impressive number of molecular ions have been investigated by means of IPS.

2.6 Photodetachment

The process whereby a negative ion loses an electron through the absorption of a photon is known as photodetachment (see Table

2). The process may be symbolized as

$$A^- + h\nu \rightarrow A + e^-, \qquad (7)$$

where the products are an atom, which may be in an excited state, and an electron with a kinetic energy sufficient to satisfy energy conservation. The interaction of a photon and a negative ion has been used to determine the binding energy of the attached electron and to study the initial and final states of the photodetachment process as well as the dynamics of the emission process.

The pioneering work in photodetachment was performed in the 1950s and early 1960s (McDaniel *et al.*, 1993). A crossed-beams apparatus, using an arc lamp as the light source, was used to measure absolute photodetachment cross sections. The advent of the laser as a light source for photodetachment studies (see Sec. 2.3), with the inherent increase in photon flux, accelerated the experimental investigation of negative ion properties. In particular, Lineberger's group at the Joint Institute for Laboratory Astrophysics investigated a large fraction of the negative-ion systems during the 1970s using single-line and tunable lasers. Most of these experiments were designed to measure electron affinities and determine relative cross sections for photodetachment. Hotop and Lineberger (1985) discuss electron affinities for elements up to $Z = 85$. The recommended electron affinities of the noble gases and group II*A* elements have negative values; i.e., stable negative ions are not formed in these groups. The formation of a stable negative ion in these groups would involve the attachment of an electron to the core of the parent atom. Calculations had supported the general belief that under such circumstances, the enhanced screening of the nuclear charge by the atomic electrons would prohibit stable negative-ion formation (Massey, 1976; Smirnov, 1982). Recently, a crossed-beams laser-photodetachment electron spectroscopy technique (LPES) was used by Pegg and coworkers (1987) to determine that Ca^- is a stable negative ion.

Photodetachment of the simplest negative ion, H^-, has been studied in detail by Halka *et al.* (1991) using a unique experimental technique. In these experiments, a relativistic H^- beam from the Los Alamos Meson Physics Facility is intersected by a fixed-frequency laser. The Doppler effect allows a systematic variation (factor of 10) of the center-of-mass photon energy, simply by adjusting the laboratory intersection angle between the two beams. This technique has been used to study doubly excited resonances of H^-. Relative cross sections for multiphoton detachment above the one-electron detachment threshold have also been measured.

2.7 Delta-Electron and Other Secondary Emission Processes in Fast Heavy-Ion–Atom Collisions

2.7.1 Introduction When an ion of high velocity interacts with a target atom, one of the possible interaction processes is the ejection of a free electron. Secondary-electron emission has been studied for a large variety of ion-atom collision systems. Here we summarize measurements involving projectiles at energies larger than 3 MeV/u and incident ions of high charge state up to uranium incident primarily on He and Ar gas targets. These recent experiments have been performed at the Lawrence Berkeley Laboratory SUPERHILAC accelerator.

For fast light-ion impact, the electron emission is primarily determined by the potential of the target atom, and the ionization is normally well described by the first Born approximation (see Sec. 3). For fast heavy-ion impact, however, it has been demonstrated that the doubly differential electron emission cross sections (DDCS) are not generally in good agreement with this approximation. Therefore, investigations of the electron emission probabilities for heavy-ion projectiles provide a sensitive measure of the dynamics of the ionization process during the collision. It has been shown that in fast collisions using highly charged projectiles, particularly at high electron emission energies, at forward and backward angles, higher-order scattering effects are important where both target and projectile fields contribute to the ionization process. In such cases, the "single-center emission" concept fails and has to be replaced by a "two-center emission" picture. In other words, the high-energy electrons are affected strongly by the target as well as the projectile potentials. Such studies and the determination of reliable absolute cross sections are relevant to plasma physics, fusion research, astrophysics, biophysical tissues (energy deposition in tissues), and basic collision theory.

2.7.2 Experimental Procedure The experimental method of electron-emission cross-section measurements has been pioneered by Rudd, Stolterfoht, and Toburen (Rudd and Macek, 1974; Briggs and Macek, 1991). For a more recent discussion, the reader is referred to the review by Toburen (1991). A characteristic setup for DDCS measurements for secondary-electron observation between a few eV and about 8 keV energies is shown in Fig. 6. In these experiments, a tightly collimated ion beam interacts with a gas target jet in the center of a magnetically shielded scattering chamber. The secondary-electron emission is observed at different observation angles using an electrostatic electron analyzer that is rotated in a plane around the scattering center.

Absolute electron-emission cross sections can be determined for the observation angles ranging from approximately 27° to 160°. The target density in the gas jet is typically of the order of 10^{14} atoms/cm^3 over a target length of about 3 mm. This low density ensures single-collision conditions. The analyzer typically has an energy resolution of a few percent [full width at half maximum (FWHM)] and a solid angle of about 4×10^{-3} steradians. The accurate target density is obtained by comparison of electron-yield measurements with the gas target jet in place and with the jet removed (homogeneous target). Furthermore, the electron detector is calibrated with electron beams at different electron energies.

2.7.3 Double-Differential Cross Sections A comparison of theoretical and experimental DDCS for 3.5-MeV/u Fe^{22+} + He is shown in Fig. 7 for $\theta = 27°$ and 155°. In this figure, the DDCS are multiplied by the electron energy for forward angles, which makes the specific differences in the cross-section behavior more visible. For example, the DDCS for Fe^{22+} + He indicate a maximum below an electron energy of 10 eV and fall off by several orders of magnitude toward the high-energy limit at electron energies of about 8 keV. At high electron energies and at forward angles, a broad characteristic peaked structure, the "binary encounter peak" (BEP), resulting from electron emission following collisions with maximum momentum transfer, dominates the cross section. The BEP energy is strongly dependent on the electron observation angle θ and is given by

$$E(\text{BEP}) = 4t_p \cos^2\theta, \qquad (8)$$

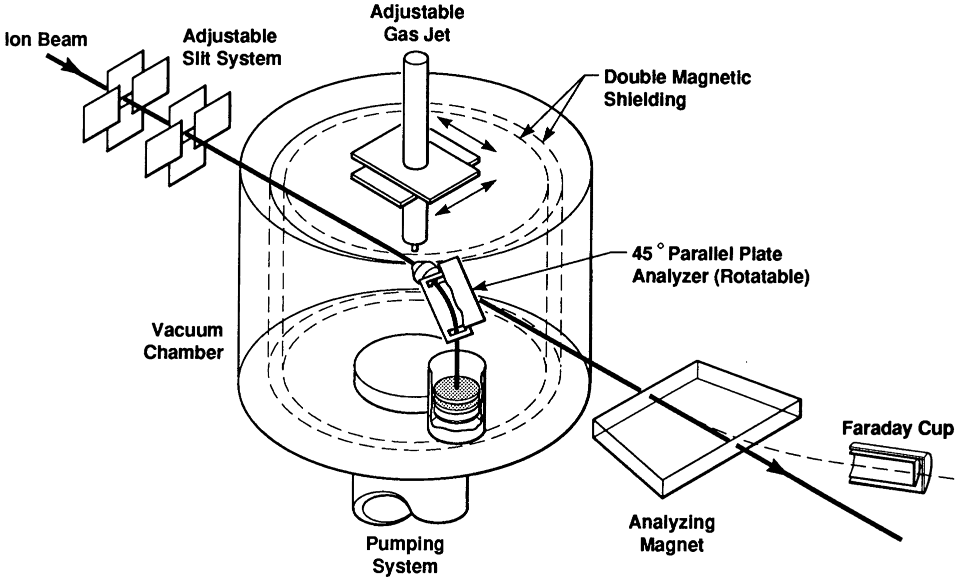

FIG. 6. Experimental setup for double-differential electron-emission cross-section measurements, consisting of ion-beam collimation system, target chamber with gas jet and electron spectrometer, and Faraday cup for charge normalization.

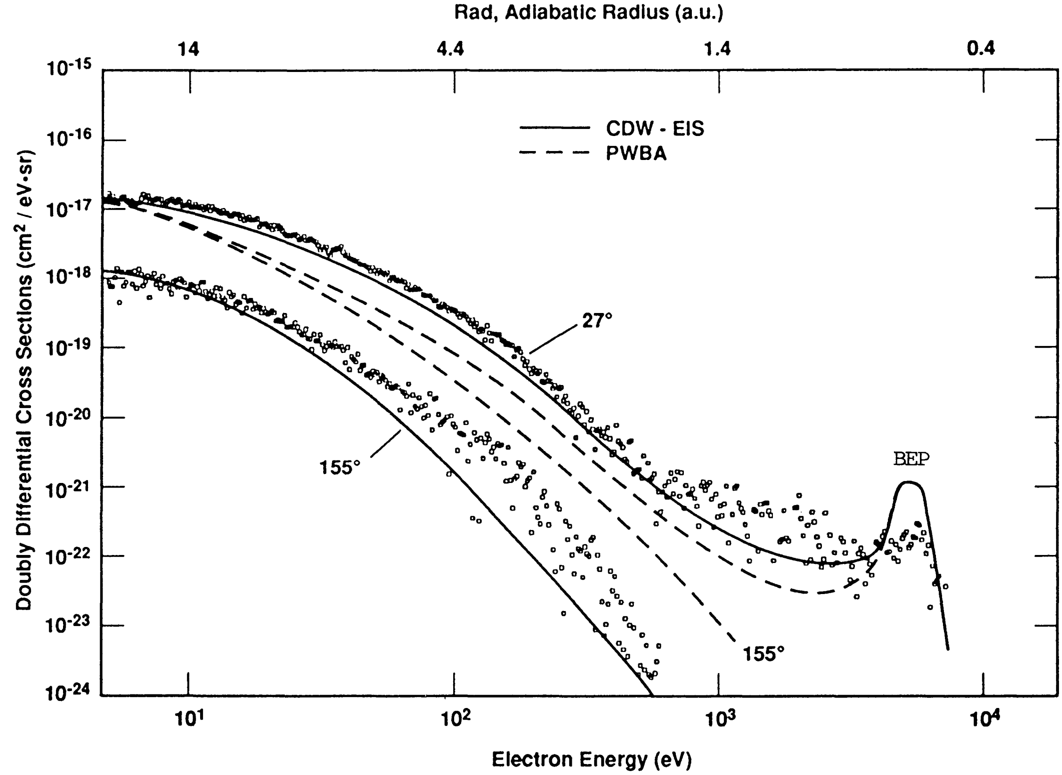

FIG. 7. Comparison of calculated double-differential electron-emission cross sections (PWBA and CDW-EIS) with experimental data for 3.5 MeV/u Fe^{22+} on He collisions. The upper scale represents the approximate impact parameter dependence. PWBA represents the plane-wave Born approximation and CDW-EIS the continuum distorted-wave eikonal initial-state calculation (Schneider *et al.*, 1992).

with

$$t_p = (m_e/M_p)E_p \tag{9}$$

the reduced projectile energy. Here m_e, M_p, and E_p are the electron mass, projectile mass, and energy, respectively. Another broader electron structure, which may occur at the reduced beam energy t_p, is the "electron loss peak" (ELP). This structure, which is not a predominant peak in the present case, is independent of the electron observation angle. As can be seen from Fig. 7, the intensity of the BEP contributions decreases with increasing observation angles. Continuum electrons centered at the projectile (ELP) and the target (BEP) may overlap depending on the projectile, the angle of observation, the projectile charge state, and target atom used. Measurements with heavier projectiles like U^{q+} show interference effects in the BEP, which have been related to the elastic differential cross section for scattering of target electrons from the impinging ion, similar to the phenomenon of rainbow scattering.

2.8 Auger-Electron Spectroscopy Following Ion-Atom and Ion-Molecule Collisions

Projectile ion beams colliding with multielectron target atoms, molecules, thin foils, or surfaces also give rise to the emission of Auger or autoionization electrons. These discrete line structures are superimposed on the continuous electron background. Auger-electron spectroscopy (*q.v.*) has become a well-established spectroscopic method (Sevier, 1972) for studying the structure and collision dynamics of atoms, molecules, and ions (Mehlhorn, 1985).

2.8.1 Forward-Angle Projectile Auger-Electron Spectroscopy The availability of intense, highly charged ion beams makes possible the production of inner-shell excitation and/or ionization at high velocities and multielectron capture processes at low velocities. Such highly excited states subsequently decay by two competing processes, radiative and radiationless via Auger decay or autoionization to one or more adjacent continuum states. The later Auger processes result in the ejection of free electrons with well-defined kinetic energies depending on the excitation energies of the initial and final ionic states. In the limit of nonrelativistic theory, the Auger decay is caused by the electron-electron interaction, whereas in the relativistic regime, spin-induced transitions also may occur. The identification and classification of such highly excited states has been one of the challenges in atomic physics over the last decade. These states are of fundamental importance for the understanding of high-temperature laboratory plasmas and electron correlation effects. They also provide important test cases to probe relativistic interactions. If Auger electrons are emitted from a moving emitter frame (projectile), they exhibit kinematic effects; i.e., Auger lines are modified in energy position, width, and intensity (Rødbro *et al.*, 1979; Stolterfoht, 1987). For example, the Auger-line energy position E in the laboratory frame is transformed to the center-of-mass (source particle) energy E' by

$$E' = E + t_p - 2\sqrt{t_p E}\cos\theta, \qquad (10)$$

where θ is the observation angle in the laboratory frame. When the ion velocity increases, the projectile electrons are strongly affected by kinematic broadening effects originating mainly from the variation of the observation angle θ and the finite angular spread $\Delta\theta$ of the electron analyzer. Such broadening effects can be reduced when the electrons are observed at forward angles, especially zero degrees.

Recently, the zero-degree electron spectroscopy method (Stolterfoht, 1987) has been used to study high-resolution Auger spectra of a variety of multiply charged ions. In these measurements, electrons emitted at 0° and 180° can be observed with energies E' as low as 0.1 eV. These low-energy line structures are centered around the electron cusp peak, which has a maximum at t_p [see Eq. (9)].

2.8.2 Auger-Electron Angular Distributions The measurement of magnetic substate specific alignment and orientation parameters in ion-atom and ion-molecule collisions can reveal more detailed information on dynamical effects that is not directly accessible via total cross-section measurements (Kleinpoppen and Williams, 1980; Andersen *et al.*, 1988). Thus, close collisions between fast-moving ions and target atoms or molecules may result in anisotropic "aligned" population of excited projectile states with inner-shell vacancies. One technique to study such anisotropies is by means of Auger-electron angular distributions (Bisgaard *et al.*, 1980). An interesting case is represented by the projectile Auger electron spectra for $\mathrm{Be}^+ + \mathrm{He}$ and $\mathrm{Be}^+ + \mathrm{CH}_4$ collisions. The Be^+ ions were obtained from the 600-keV ion accelerator of the University of Aarhus, Denmark, and directed through a He or CH_4 gas target. The ejected projectile Auger electrons were analyzed by an electrostatic parallel-plate spectrometer, which was continuously rotated between 0° and 150°. The $\mathrm{Be}^+\ (1s2s^2)^2S \rightarrow (1s^2\epsilon s)^2S$ Auger transition is isotropic and was used for normalization purposes. In order to obtain the cross sections in the projectile emitter frame, it was necessary to correct for the Doppler anisotropy. Representative results of $\mathrm{Be}^+\ 1s(2s2p\ {}^3P^\circ)^2P^\circ$ and $1s(2s2p\ {}^1P^\circ)^2P^\circ$ angular distributions are shown in Fig. 8. The anisotropy coefficient A_2 was found by fitting

$$I(\theta_0) = C(1 + A_2 P_2 \cos\theta_0) \qquad (11)$$

to the angular distribution where θ_0 is the emission angle in the projectile emitter frame. The large positive alignment parameter A_2 corresponds to an overpopulation of $M_L = 0$ magnetic substates, when compared to $M_L = \pm 1$ states. Large anisotropies have also been observed recently in $\mathrm{C}^{5+}(1s) + \mathrm{He}(1s^2) \rightarrow \mathrm{C}^{3+}(1s2l2l)^2L + \mathrm{He}^+$ double electron-capture processes at low collision velocities by Prior and co-workers at the Lawrence Berkeley Laboratory.

2.9 Photon Emission Processes in Fast Projectile Collisions

2.9.1 Multielectron Transitions Many-electron processes like double ionization, double excitation, and ionization-excitation,

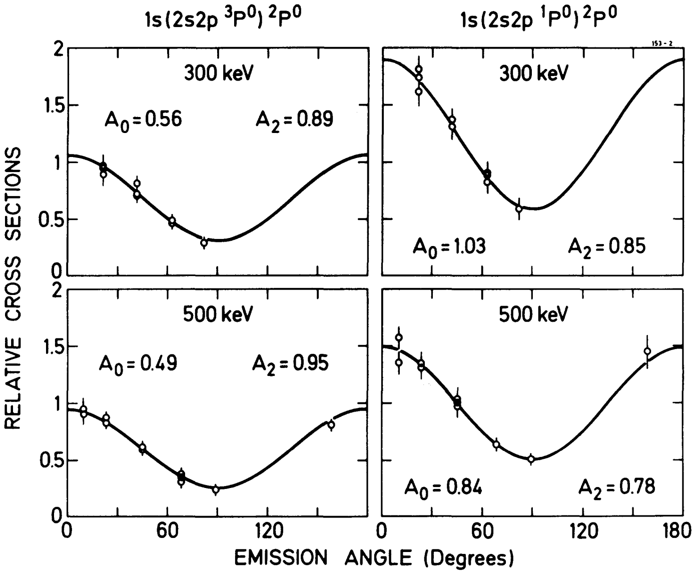

FIG. 8. Angular distributions of ejected projectile Auger electrons from Be^+ $1s(2s2p\ ^3P^o)\ ^2P^o$ and $1s(2s2p\ ^1P^o)\ ^2P^o$ core-excited states excited in $Be^+ \rightarrow He$ ion-atom collisions.

which occur in collisions of fast charged particles with few-electron targets, have received much attention recently (McGuire, 1992). Such two-electron processes are of fundamental importance to an understanding of the dynamics of few-body atomic collisions. Specifically, the measurements of absolute extreme-ultraviolet (EUV) photon and electron emission cross sections may allow development of a more comprehensive formulation of few-body atomic processes. Some important applications in which those few-electron transitions occur are plasma dynamics and astrophysical processes including solar atmospheric physics. For example, a challenging problem in astrophysics is the transformation of stored electromagnetic energy and particle emission in solar flares, which is still not well understood. Also, electron correlation occupies a fundamental role in inner-shell photoionization of atoms and molecules (Amusia, 1990). Furthermore, these many-body effects are relevant to plasma diagnostics, x-ray laser research, chemical physics, and quantum biology of simple and complex systems.

2.9.2 EUV Spectroscopy State-selective cross-section measurements from excited target atoms can be studied by using the method of high-resolution EUV spectroscopy. In these measurements, a positively or negatively charged or neutral particle beam enters a differentially pumped gas target or intercepts a gas jet. After exciting the target gas, the beam is collected in a Faraday cup for the purpose of charge normalization. A schematic view of a gas target and EUV monochromator is shown in Fig. 9. The photons emitted by the target cell are analyzed by a grazing-incidence monochromator and counted by a Channeltron electron detector. All measurements are performed under single-collision conditions. The gas pressure in the cell is accurately monitored with a capacitance manometer and kept constant with a feedback control system. Typical ion-beam currents in the target area are 1–20 μA in the energy range of 50 keV to 2 MeV. Identical excitation and detection geometries are used for electron and ion impact. The gas pressure used in these experiments typically ranges from about 0.1 to 100 mTorr.

Characteristic spectra of EUV Lyman series of He^+ are shown in Fig. 10 for both electron and proton impact at equal velocities. The Lyman series is clearly resolved up to $n = 5$. To derive excitation cross sections from measured line intensities, the following formula can be used:

$$\sigma = \frac{4\pi}{\omega}\frac{qe}{Q}\frac{I(\lambda)}{k(\lambda)BNL}, \tag{12}$$

where e is the electron charge, q the charge

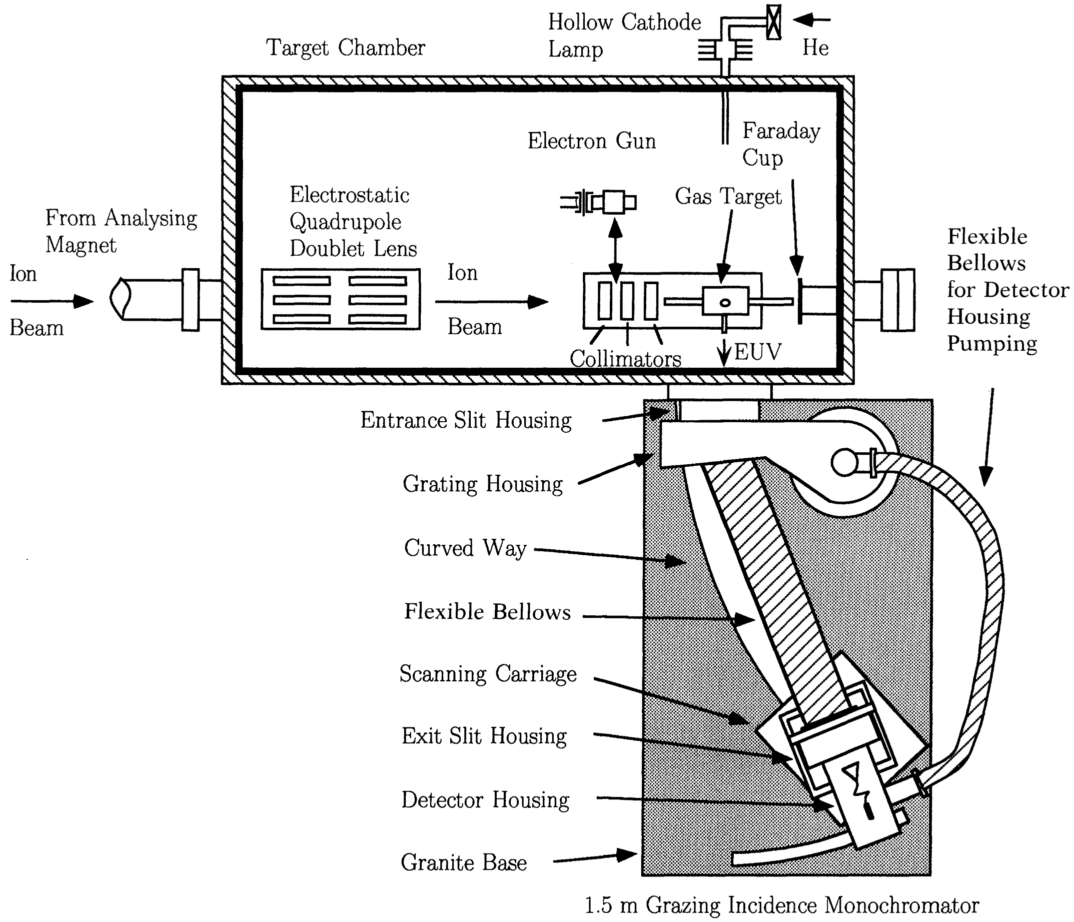
Hollow Cathode Lamp
He
Target Chamber
Electron Gun
Faraday Cup
Electrostatic Quadrupole Doublet Lens
Gas Target
From Analysing Magnet
Ion Beam
Ion Beam
Collimators
EUV
Flexible Bellows for Detector Housing Pumping
Entrance Slit Housing
Grating Housing
Curved Way
Flexible Bellows
Scanning Carriage
Exit Slit Housing
Detector Housing
Granite Base
1.5 m Grazing Incidence Monochromator

FIG. 9. Target chamber for EUV emission spectroscopy with gas cell, electron gun, and grazing-incidence monochromator.

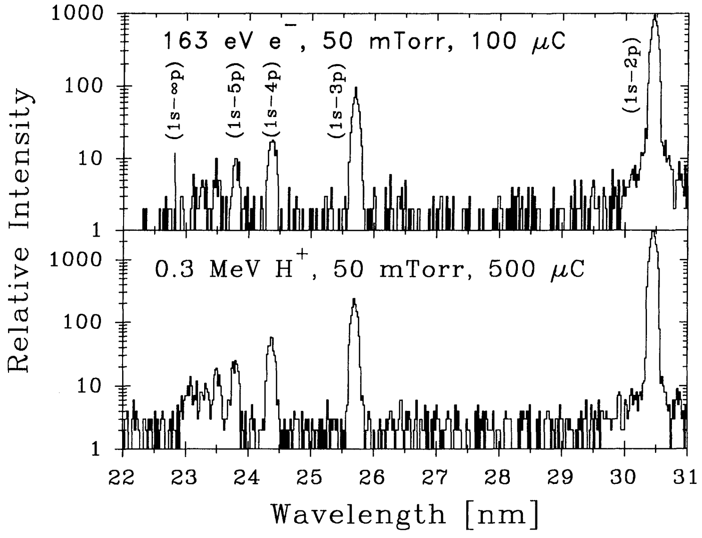

FIG. 10. EUV emission spectrum of the Lyman series of He^+ *np* → 1*s* for (top) 163-eV electron impact on He and (bottom) 300-keV proton impact on He at equal projectile velocities. The data were generated at the University of Nevada, Reno, 2-MV Van de Graaff accelerator facility.

state of the projectile, Q the integrated beam charge or charge normalization, $I(\lambda)$ the measured line intensity, $K(\lambda)$ the sensitivity of the detection system, B the branching ratio associated with a specific transition, N the target number density, ω the solid angle of acceptance for the monochromator, and L the observation length (the reaction region visible to the monochromator).

2.9.3 Ionization-Excitation of He In particular, ionization plus excitation is a fundamental process relevant to the testing of basic scattering theories. Helium is the simplest many-electron atom and therefore ideally suited for achieving a better understanding of two-electron processes. According to McGuire (1992), cross terms between first-order and second-order perturbation-theory contributions may be responsible for the observed large differences in ionization-excitation cross sections at equal velocities for protons and electrons (see Fig. 11). In this figure, absolute $He^+(2p)$ cross sections are plotted for $H^+ + He$ and $e^- + He$ as a function of the projectile velocity. Higher *np* states have also been measured (Fuelling *et al.*, 1992). The following observations can be made from the plotted cross section data: For velocities v/v_0 above 4, the $e^- + He$ cross sections lie consistently above the corresponding proton results. Two basic reaction mechanisms contribute to the observed cross sections, namely few-electron processes where the projectile interacts only once with the target, and a second Born term where the projectile interacts twice with the target in a single collision event. Such processes can be distinguished by their Z_p dependence, where Z_p is the projectile charge. Furthermore, if these first- and second-order scattering amplitudes add coherently, a Z_p^3-dependent term appears in the cross section, which varies as the square of the total amplitude. Hence, cross-section differences for electrons ($Z_p = -1$) and protons ($Z_p = +1$) at equal velocities may provide direct evidence of first- and second-order effects. Such Z_p^3 cross terms occupy also a significant role in double ionization and double excitation of He. These two-electron transitions have been studied experimentally and theoretically for double ionization of He by antiprotons ($\bar{p}$) and protons (p) as well as electron (e^-) and positron (e^+) impact on He (Andersen, 1988). Moreover, measurements have been made for double excitation of He as a function of the projectile charge, mass, and velocity (Giese *et al.*, 1990; Wang *et al.*, 1993).

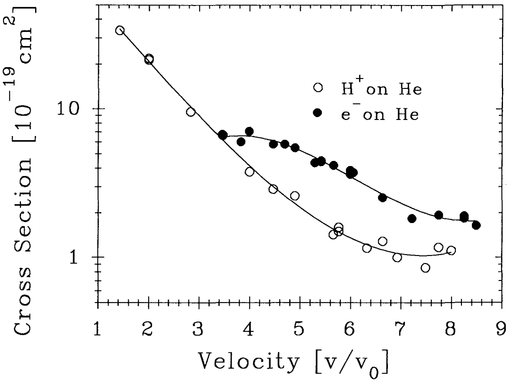

FIG. 11. Absolute ionization-excitation cross sections for electron and proton impact on He for $He^+(2p)$ states, as a function of the projectile velocity in units of v_0 (Bohr velocity).

2.9.4 Excitation of Molecular Gases Laboratory studies of ion–molecule collisions are an important aspect of ionospheric research and can add to the understanding of the processes affecting auroral phenomena. Describing the aurora borealis presents a challenging problem, the solution to which depends on the knowledge of photon emission cross sections for the excitation and ionization-excitation of atmospheric gases by the impact of energetic electrons, protons, and other positively and negatively charged species. The experimental method is identical with the one described in Sec. 2.9.2 except that the EUV grazing-incidence spectrometer is replaced, for example, by a 1-m air Czerny–Turner optical monochromator. Recently observed nitrogen and oxygen spectra in the visible range following H^+, H_2^+, and H_3^+ impact on N_2 and O_2 have revealed some spectral lines which arise from dissociation of molecules and ionization and excitation of the reaction fragments. The dominant spectral features arise from neutral and singly charged molecules. High spectral resolution in the visible or ultraviolet (UV) is needed to resolve the electronic transitions, each accompanied by additional vibrational structures and clusters of rotational components. Because of the many degrees of freedom, it is very difficult to determine accurate state-specific molecular cross sections (Bashkin *et al.*, 1991). Studies of these molecular collision processes in the laboratory can also yield important knowledge relevant to a better understanding of extra-terrestrial plasmas.

2.9.5 Electron Capture to Rydberg States The mechanism of electron capture at higher velocities, $v > v_e$, has received much attention recently. In this case, the first Born approximation is inadequate for electron capture by charged particles, and higher-order contributions are required to explain experimental cross-section data. As a prototype example, the following capture reaction is discussed here:

$$\begin{aligned} &C^{4+}(1s^2) + A \rightarrow C^{3+}(1s^2nl) + A^+ \\ &\qquad\qquad\qquad \hookrightarrow C^{3+}(1s^2n'l') + h\nu, \end{aligned} \tag{13}$$

where the target A is either H_2 or He (Bruch *et al.*, 1982; Dubé, 1986). High-resolution EUV spectroscopy of the subsequent photon emission is used to extract selective information about the final-state (nl) population produced during the capture process. The impact energies ($v \approx 2.6$ to 4.1 a.u.) were selected such that the relative velocities were large enough for perturbative charge transfer theories to be applicable but small enough for the cross sections to remain significant. A common feature of all experiments performed at those velocities is that, in contrast to similar investigations at lower velocities (Janev and Winter, 1985), where the electron is preferentially captured into one or two well-defined principal shells, several prominent lines can be identified as transitions from Rydberg states (nl) with $n \leq 10$. The cross-section measurements on s, p, and d series obtained in C^{3+} give evidence of the inadequacy of the single-scattering theory and provide a direct measure of the importance of multiple-scattering effects—for example, theories based on a distorted-wave formalism like the continuum distorted-wave approximation (CDW) (see Sec. 3). Other experimental studies (Chetioui *et al.*, 1988) have reached the same conclusions and confirm the superiority of multiple-scattering descriptions for single-electron capture mechanism over the first Born approximation.

2.10 Collisions with Slow Multiply Charged Ions

2.10.1 Introduction The prospect of new exciting physics that can be studied with highly charged heavy ions in various areas of physics has stimulated new developments of advanced types of ion sources (Marrus, 1989), new accelerator facilities (Stöckli *et al.*, 1992), ion storage rings (Datz, 1987; Schuch, 1989), and ion traps (Levine *et al.*, 1988; Church, 1988). Low-velocity, highly charged ions strongly attract electrons in their vicinity and therefore represent the most reactive species in the universe (Schneider, 1991). For example, close interaction of one such ion with a molecule can result in the Coulomb explosion of this molecule.

The possible impact regarding basic physics as well as applied physics related to highly charged ions is expressed in recent reviews and international workshops and conferences (Ivanov, 1988; Alonso *et al.*, 1989; Salzborn *et al.*, 1991; Poth, 1990). Nuclear, atomic, solid-state, and surface physics (Andrä, 1989; Tully and White, 1977), as well as x-ray laser research and space science, are realizing the potential of new-generation facilities including advances in instrumentation, detection, and laser technology. The physics of highly ionized atoms is equally important to those questions at the boundary of pure and applied science (e.g., input for plasma modeling) and to those questions having a far-reaching impact on the understanding of the fundamental laws of nature, i.e., strong relativistic interactions (Wilson *et al.*, 1991) and multielectron (QED) effects (Mohr, 1989). The most recently developed facilities that are able to produce highly charged high-Z ions (HCHZI) throughout the periodic table are electron cyclotron resonance (ECR) ion sources (Meyer and Kirkpatrick, 1991), electron-beam ion sources (EBIS) or traps (EBIT) (Schneider *et al.*, 1991), and heavy-ion storage rings (HISR).

The EBIT, in particular in conjunction with additional ion traps, can compensate partly for the lack of storage rings. A wealth of experiments have been suggested for storage rings and ion traps for HCHZI, and some of the work has been in progress over the last few years. In this context, it should be noted that this new scope of atomic and molecular physics involves rather sophisticated, highly efficient x-ray and extreme-ultraviolet (EUV) optics, and high-resolution spectrometers, including absolute intensity standards and detection systems. The resulting highly accurate measurements may serve as benchmark data both for basic theoretical calculations of atomic processes and as input for high-temperature plasma modeling codes and in other areas of applied and basic physics. In ion traps and storage rings, a major experimental effort is, for instance, to produce ion crystals in the fast beam (Schuch, 1989).

2.10.2 Charge-Transfer Reactions Electron-capture processes following low- to medium-energy ($E \leq 25$ keV/amu) ion-atom and ion-molecule collisions can be characterized by

$$A^{q+} + B \rightarrow A_{n,l}^{(q-s)+} + B^{r+} + (r-s)e, \tag{14a}$$

$$A^{q+} + B + h\nu_L \rightarrow A_{n',l'}^{(q-s)+} + B^{r+} + (r-s)e, \tag{14b}$$

where A^{q+} is a multiply charged ion of charge state q, B describes a target atom (molecule or surface), and $h\nu_L$ characterizes a possible additional laser field. Process (14a) represents a typical charge transfer plus target ionization reaction, whereas mechanism (14b) can be viewed as the "photon-assisted" counterpart of reaction (14a). If the laser frequency is tuned into resonance of the target system, reaction (14b) is considered electron capture from an excited state of the target atom (molecule). The motivation for these experiments is

1. to determine relative and absolute cross-section data relevant for plasma diagnostics, fusion research, and laser gain in the EUV and soft x-ray region, and
2. to extend knowledge and understanding of the charge-transfer process to the case of excited target atoms and quantum optics of photon-assisted interactions.

For example, x-ray laser pumping by virtue of charge-transfer processes has a great potential for multiply charged ions because of the extremely large electron-capture cross sections ($\sigma \geq 10^{-15}$ cm^2) and the state-selective nature of the transfer process.

In the past few years, single and double electron-capture processes of multiply charged ions with atoms and molecules have provided new insight into few-electron dynamical phenomena that are dominated by electrostatic interactions. For the more highly charged ions, relativistic effects need to be considered for theoretical models describing these systems. We discuss here the formation and Auger decay of metastable Fe^{15+} $(2p^5 3s3p)^4D_{7/2}$ states by means of the zero-degree electron spectroscopy technique (Sec. 2.8.1). Assuming single-collision conditions, such quartet states can be formed in two-electron–capture processes as follows:

$$\mathrm{Fe}^{17+}(2p^5) + \mathrm{He}(1s^2) \\ \rightarrow \mathrm{Fe}^{15+}(2p^5 3snl)^4L + \mathrm{He}^{2+}. \tag{15}$$

In Fig. 12, typical high-resolution projectile-electron spectra produced in 170-keV Fe^{17+} + He collisions are displayed. The most intense

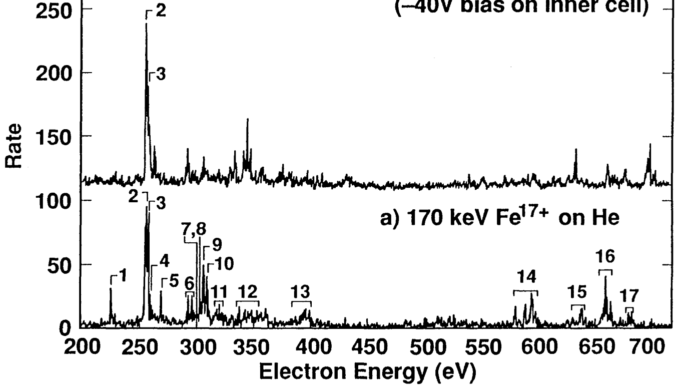

FIG. 12. (a) Auger-electron spectrum of sodium-like Fe$^{15+}(2p^5 3lnl')$ states following 170-keV Fe^{17+} + He collisions in the energy range from 200 to 700 eV. The electron energies are given in the projectile emitter frame. (b) Separation of long-lived metastable autoionizing states from prompt Coulomb autoionizing states. The delayed lines 2 and 3 are mainly due to the $(2p^5 3s3p)^4D_{7/2} \rightarrow (2p^6 \epsilon g)^2G_{7/2}$ and $(2p^5 3s3p)^4P_{5/2} \rightarrow (2p^6 \epsilon d)^2D_{5/2}$ transitions.

peaks arise from initial states associated with the $2p^5 3s3p$, $2p^5 3p3p$, $2p^5 3s3d$, and $2p^5 3p3d$ configurations. In particular, the prominent lines in the spectrum, labeled 2 and 3, are due to spin-induced Auger decay of the $(2p^5 3s3p)^4D_J$ and $(2p^5 3s3p)^4P_J$ levels. "Prompt" and "time-delayed" Auger transitions have been isolated. The direct excitation of the $^4D_{7/2}$ state involves a spin flip in Fe^{17+} + He single collisions.

2.10.3 Electron-Beam Ion Traps: New Facilities for Collision Experiments with Heavy, Highly Charged Ions A schematic diagram of EBIT, including the ion extraction system, is shown in Fig. 13. A 3-*T* axial magnetic field from two superconducting Helmholtz coils compresses an electron beam to current densities of up to 6000 A/cm^2 in the drift-tube region. Typical operating electron-beam currents are of the order of 100 mA. The axial magnetic field varies by less than 0.2% over the 2-cm length of the ion trap. The drift-tube assembly accelerates the beam electrons to their full interaction energy and provides an axial electrostatic trap for the positive ions. At present, EBIT is capable of producing electron energies of up to 30 keV. An upgrade of EBIT allows use of electron energies up to 150 keV (Super EBIT). Neutral atoms can be injected through one of the side ports of the trap. In order to inject ions, EBIT is also equipped with a MEVVA (metallic vapor vacuum arc) source that produces the initially low-charged ions up to 4, e.g., U ions. Only a fraction of these ions are trapped in the central drift-tube region, where the ions are confined axially by a potential well provided by two end drift tubes. The axial potential-well depth is typically around 100–300 eV. Radial confinement is provided by a combination of the space-charge field of the electron beam and the applied axial magnetic field. The ions undergo successive stages of ionization in the electron beam. The highest ionization stage present in the trap is determined by the electron-beam energy. The number of ions available depends on the ionization, recombination, and charge-exchange cross sections. In order to inject the ions into the trap, confine them for a given time and ionize them, and extract them, the potentials on the trap have to be varied in a fast switching mode. In the extraction phase, the ions are electrostatically deflected and focused through an einzel lens in the path between the MEVVA and the trap. The same lens provides focusing of the initially injected ions from the MEVVA into the trap. After passing the deflector, the ions are refocused via a second einzel lens and are momentum analyzed in a 90° magnet. A recent example of charge-state measurements of extracted uranium ions illustrates the capability of EBIT (see Fig. 14). These measured charge-state distributions allow detailed modeling calculations including ion cooling which is essential for heavy-ion trapping. Using such highly charged ions, new experiments in different areas of physics can be performed. Spectroscopy experiments focus on precision (one per million) measurements on highly charged few-electron ions. Photon interaction experiments will look at photoionization, photoexcitation, and lifetime measurements with lasers. Ion-ion and ion-atom collisional processes can be used to study resonant transfer and excitation, charge exchange, recoil ion production, and ion-ion collisional cross sections. Furthermore, electron-ion interactions, such as dielectronic recombination and electron-impact excitation and ionization, can be studied. All these processes are needed to model inertial confinement and other relevant plasmas, such as astrophysical sources (Burke and Eissner, 1983).

2.10.4 Ion–Surface Interaction Much effort is presently directed toward research on ion–surface interactions using very highly charged ions (Andrä, 1989). In particular, the recently developed technique (Schneider *et al.*, 1991) of extracting highly charged ions from EBIT allows one to use ions up to Th^{80+}, e.g., for ion-surface interaction studies. The goal of these studies is to provide an understanding of the neutralization dynamics of very highly charged ions as they approach the surface and penetrate into the solid. Such highly charged projectile ions carry up to several hundred keV potential energy, and x-ray emission studies are particularly suitable to illuminate the different interaction processes that lead to the transfer of this energy to the surface. The satellite intensities and energy positions of the emitted x rays and Auger electrons provide important information on the history of the projectile ions interacting with the surface and the bulk material. As illustrated in Fig. 15, highly charged slow ions capture electrons efficiently into

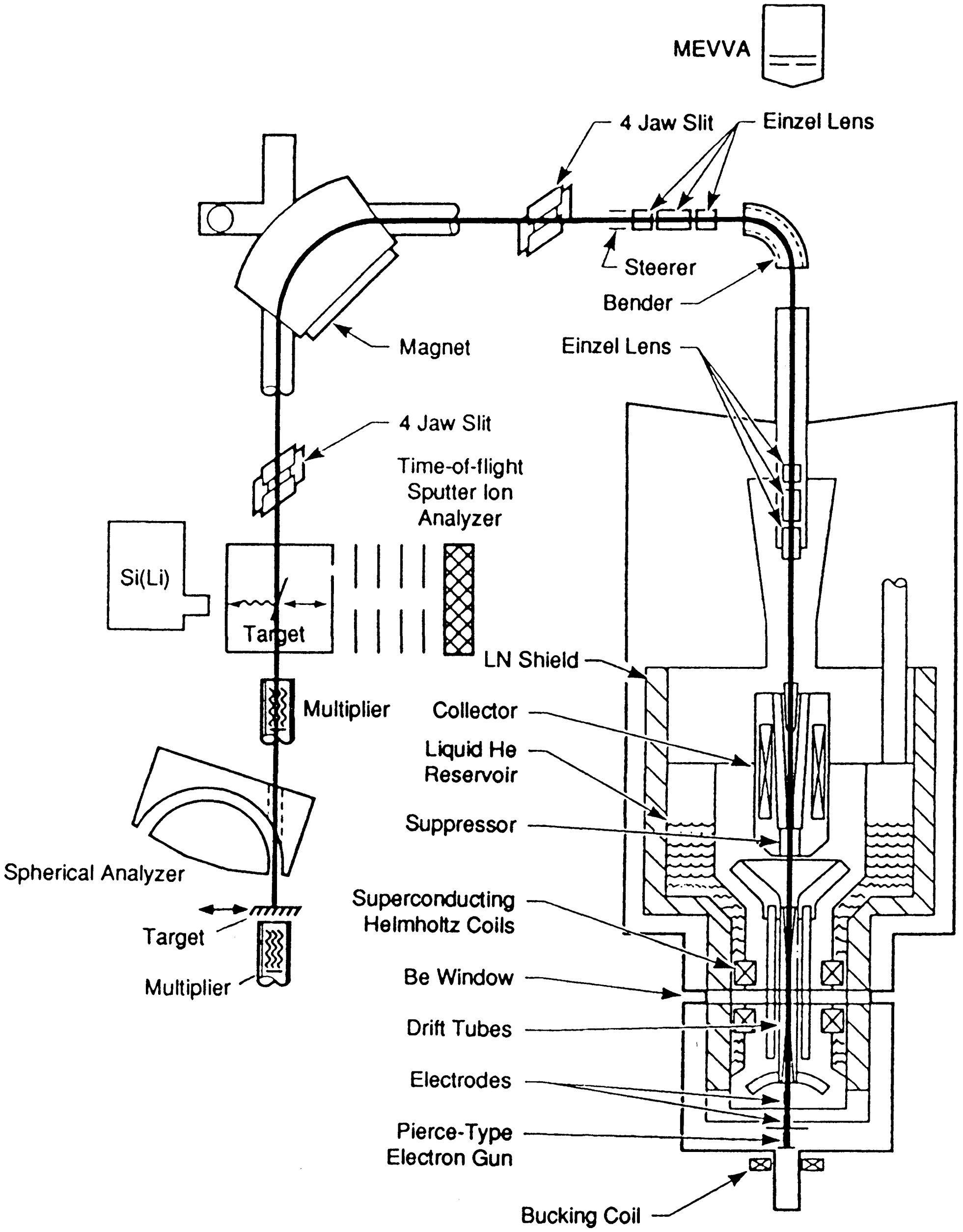
MEVVA
4 Jaw Slit
Einzel Lens
Steerer
Bender
Magnet
Einzel Lens
4 Jaw Slit
Time-of-flight
Sputter Ion
Analyzer
Si(Li)
Target
LN Shield
Multiplier
Collector
Liquid He
Reservoir
Suppressor
Spherical Analyzer
Superconducting
Helmholtz Coils
Target
Multiplier
Be Window
Drift Tubes
Electrodes
Pierce-Type
Electron Gun
Bucking Coil

FIG. 13. Electron-beam ion trap, including electron gun and extraction system with focusing elements, analyzing magnet, and target arrangements. Also shown is the experimental setup for highly charged ion-surface studies, consisting of a Si(Li) x-ray detector, time-of-flight sputter ion analyzer, and spherical electron spectrometer.

high-*nl* states at relatively large distances depending on the ionic charge (Tully and White, 1977). Thus, the ion is promoted into a multiply excited state as it approaches the surface with electrons occupying high-*nl* levels while the core is virtually empty (hollow atoms) (Briand *et al.*, 1990).

Following the classical overbarrier model (Folkerts, 1992), the neutralization process starts at a critical distance $R_c \sim 4.6q$ (Å) where q is the projectile charge (assuming a metal work function $W_\phi = 4.5$ eV). From these measurements, the occurrence of "hollow atoms" during the neutralization process has been

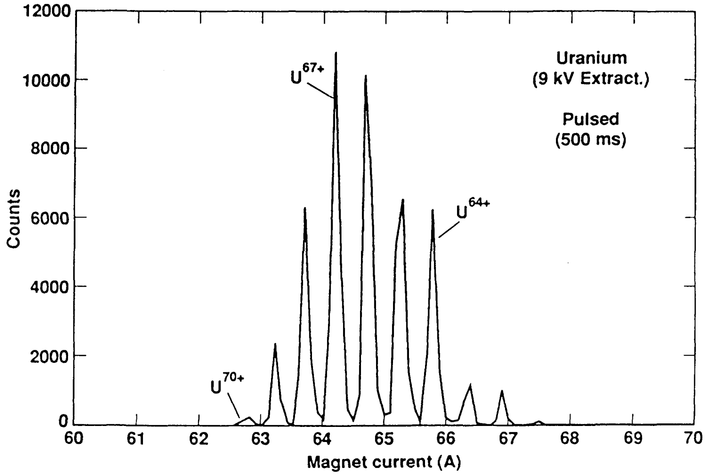

FIG. 14. Spectrum of charge analyzed uranium ions following 9-kV electron impact.

inferred. The created multiple inner-shell vacancy states provide time windows (atomic clocks) that allow study of stepwise deexcitation and evaporation of the excitation energy. In these studies, the following physical processes and parameters are of major importance (Clark *et al.*, 1993):

1. The initial electronic configuration of the approaching ion and the surface potential of the solid determine the transfer of electrons from the surface into states with high *n* quantum numbers.
2. The approach velocity and the autoionization and radiative decay rates determine the transfer of electrons to lower states or back to the solid, together with the distribution and time evolution of the population of different quantum states during neutralization.
3. Transfer of electrons into states with low quantum numbers occurs when the ion enters into the surface.

The ion–bulk interaction is more pronounced at higher approach velocities. In this case, the inner-shell vacancies have a larger probability of interpenetrating the first surface layer, and the interaction with the bulk causes a fast filling of shells with lower *n* quantum numbers with a simultaneous "peeling off" of electrons in shells with higher *n* manifolds.

Recently, doubly differential electron-emission yields following the impact of fast (3.95×10^7 cm/s) highly charged ions on Cu and Au targets have been measured (McDonald *et al.*, 1992). The electron emission is dominated by low-energy electrons (<20 eV). Specifically, it is found that the total yield, which rises to about 200 electrons per ion, is a nonlinear function of the total potential energy of the incident ion (up to about 200 keV). Ion–surface interaction studies with 1–3-keV/amu ions up to Th^{75+} have also been reported for conductor, semiconductor, and insulator materials. Microscopic surface analysis using an "atomic force microscope" revealed interesting nanoscale defects caused by single highly charged ion bombardment.

2.10.5 Ion Storage Rings: A New Tool for Atomic and Molecular Collision Physics

Heavy-ion storage rings make possible experiments with highly ionized and cooled ion beams (Datz, 1987; Schuch *et al.*, 1989). The main motivation for the design and construction of such heavy-ion storage rings is the study of fundamental dynamical processes of ions interacting with ions, atoms, molecules, electrons, and photons. One major advantage achieved in cooled storage-ring systems is the gain in luminosity and emittance as well as interaction time when compared to conventional ion experiments (see Sec. 2.1), leading to a substantial increase in sensitivity and precision in collisions with heavy, highly charged ions. These new facilities are also ideally suited for accurate measurements of atomic and molecular properties such as binding energies and relativistic and QED effects over a wide range of nuclear charges.

A simplified schematic view of the Stockholm ion storage ring is presented in Fig. 16.

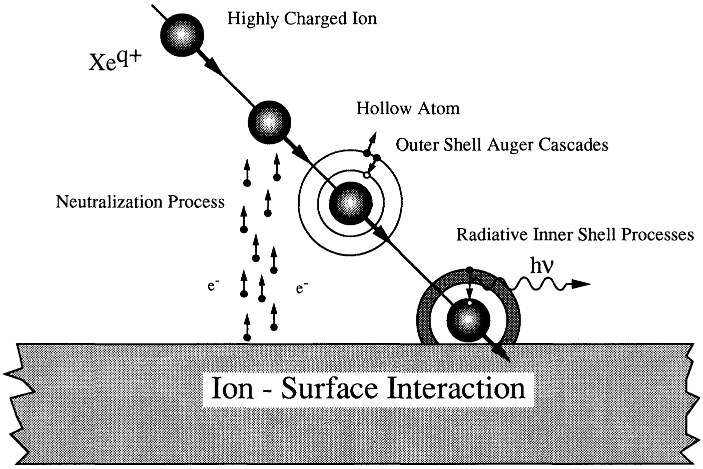
Highly Charged Ion
Xe^{q+}
Hollow Atom
Outer Shell Auger Cascades
Neutralization Process
Radiative Inner Shell Processes
hν
e^-
e^-
Ion - Surface Interaction

Ion-Surface Interaction:

- Charge Transfer to Highly Charged Ion
- Projectile X-Ray and Electron Emission
- Projectile Penetration into the Bulk
- Material Ablation and Surface Modification
- Coulomb Explosion and Fragmentation of Surface Molecules

FIG. 15. Schematic diagram illustrating the neutralization process of highly charged Xe^{q+} (q = 44 to 48) ions close to a metal surface. The figure shows the creation of a highly excited hollow atom by multiple electron capture. This superexcited state then decays via outer-shell Auger cascades and inner-shell radiative processes.

This facility is uniquely suitable for studies of the fundamental interaction of very highly charged heavy ions with other ions, electrons, or photons. Investigations also involve ground-state molecular ions. The ring consists basically of a periodic structure of magnet components, i.e., dipole magnets for bending, focusing, and defocusing the beam. Also inserted in the ring are beam-cooling and synchrotron acceleration devices. The highly charged ions are created from a cryogenic electron-beam ion source (CRYSIS) and are injected into the ring after preacceleration. Fast acceleration is necessary (~100 ms) to minimize electron-capture beam losses.

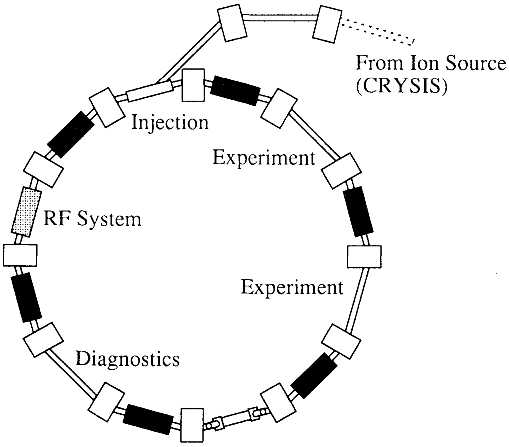

FIG. 16. Schematic layout of the Stockholm heavy-ion storage ring with rf acceleration system, main magnets, dipoles, focusing elements, electron cooler, straight experimental sections, and injection system.

The ions circulating in the storage ring have slightly different velocities. This velocity spread can be regarded as "thermal" motion. The reduction of the ion temperature in the ring is important for precision experiments. At present, there are three methods available for cooling hot ions, i.e., electron cooling, stochastic cooling, and laser cooling. Electron cooling is accomplished by the exchange of momentum through repeated Coulomb scattering between the ion beam in the ring and a low-temperature electron beam. The two beams are merged together over a distance of about 1 m with the cold electrons having the same average velocity as the ions. Through electron-ion interaction, heat is transferred from the ions to the electrons. Cooling times for 10-MeV/u particles are approximately 0.02 s for Xe^{44+} ions. The electron cooler can also be used as a target in experiments with cooled ion beams. Typical electron densities in the cooler are of the order 10^8 cm^{-3}. Important electron-ion recombination processes that can be studied include radiative recombination, laser-induced recombination, dielectronic recombination, and three-body recombination.

For singly charged ions such as Li^+, Be^+, Mg^+, etc., laser cooling also can be used in the ring to study ion-ion plasmas and crystals. Recently, Walther and co-workers have demonstrated crystallization of Mg^+ ions in a tiny table-top ring structure (Waki *et al.*, 1992). One of the ultimate goals is to observe phase transitions from ion clouds to liquids or crystals. These measurements can reveal information on the Coulomb coupling in one-dimensional (linear string, necklace) or three-dimensional helical structures. A recent simulation of 5-MeV/u Ar^{18+} ions has shown that a crystallized beam would retain its order after about four revolutions in a large storage ring. Similar results have been predicted at GSI in Darmstadt. Ion-beam crystallization of highly charged ions can be monitored by laser-induced fluorescence, Bragg scattering of laser photons, position-sensitive detection as a function of a varying deflection field, or measurement of the Schottky noise.

These processes of transforming ionic (plasma) beams to aligned crystalline phases are unique and will yield new insight into the nature of phase transitions and mechanisms of rotational and vibrational states in highly charged ionic solids. Other challenging experiments are concerned with precision measurements of atomic energy levels, fine-structure and hyperfine-structure splitting, QED effects of highly charged ions, and the masses of radioactive nuclei. Accurate mass measurements are of interest to chemists for the analysis of structure and modeling of complex molecules. These investigations are performed with ion traps (Church, 1988), in which ions can be stored by electromagnetic fields. At heavy-ion ring facilities, using ions with masses up to 100 and charge states up to 50, such experiments become feasible for the first time. Using this technique, it is planned to measure masses of heavy ions with a relative precision of $1:10^8$ using ion traps.

3. THEORETICAL MODELS FOR ATOMIC AND MOLECULAR COLLISION PROCESSES

Phenomenological and *ab initio* theoretical models have been applied to atomic and molecular systems undergoing collision, reaction, emission, and decay processes. The relative simplicity and extreme success of the Bohr atomic model for the two-body hydrogen system does not extend to three- or more-body systems. Hence, a number of theoretical models have been applied to few- and many-body systems to elucidate some of the single-particle and multiparticle collision and photon reaction mechanisms. Theoretical many-body treatments have met with some limited success. But the field of theoretical atomic and molecular physics is a dynamic one with many exciting discoveries to be made.

Although there are a number of theoretical treatments, these methods can be divided into several different classes of few- and many-body approaches. In atomic and molecular as well as nuclear systems, these approaches may involve semiempirical formulations for reaction-product parameters and relative and absolute cross-section predictions as well as fits to other useful experimental parameters. All theories involve such basic input parameters as projectile and target potentials, atomic and molecular radii or sizes, projectile charge and mass, projectile velocity, particle masses, and single-particle or multiparticle wave functions. Specifics of potential strength and form, reaction size such as channel radius, and wave functions depend on the specific system and relative velocity of the collision partners.

The usual approach is to define an operator that acts on appropriate wave functions, which are defined in terms of the mechanism of interaction such as scattering, reaction, or transition processes within the system. In Table 3, we list and summarize some basic mathematical forms of the major categories of theoretical models. In any system containing more than two bodies, we are immediately confronted with the need to use a much more complex description compared to two-body hydrogenic form.

The basic tenet of atomic and molecular processes is that they are describable by the Schrödinger wave equation. Most of these processes are nonrelativistic so that we do not require the Dirac wave equation, which is applied to more energetic processes. Some reactions are taken in the time-independent picture, and in some cases time dependence is used. Primarily, we have dealt with molecular and atomic processes such as pho-

Table 3. Theoretical models applicable to many-body physics.

Theory	Structure
1. The Born approximation in powers of the interaction potential and Green's function G.	$T = V + VG_0^+ + VG_0^+V + \ldots$ for $T^{(1)} = \langle\Psi_f\vert V\vert\Psi_i\rangle$ $T^{(2)} = \left\langle \Psi_f \middle\vert V\frac{1}{E-H+i\delta}V \middle\vert \Psi_i \right\rangle$ $V(r)G = G_0^+$ free particle, $G = \frac{1}{E-H}$ or $G = \frac{1}{4\pi r}e^{-qr}$
2. Coupled radial channels with channel radius r_c.	$\left(\frac{d^2}{dr^2} - \frac{l(l+1)}{r^2} + V_0^2\right)u_0(r) = u_{l'}(r)$ $\left(\frac{d^2}{dr^2} - \frac{l'(l'+1)}{r^2} + V_{l'}^2\right)u_{l'} = u_0(r)$
3. R matrix or reaction matrix.	$R = R_0 + R'$ for $R' = \sum\frac{\gamma_\lambda \times \gamma_\lambda}{E_\lambda - E}$ Sum over singularities $\gamma_\lambda = \left(\frac{\hbar^2}{2\mu r_c}\right)^{1/2} u_\lambda$ for $u_\lambda = G_\lambda + iF_\lambda$ $\gamma_\lambda = \frac{2\Gamma_\rho^2}{\hbar}\left(\frac{\rho}{(F_l^2 + G_l^2)r_c}\right)$ where γ_λ^2 is termed the reduced width.
4. T matrix or transition matrix.	$T = V + V\frac{1}{E-H-i\delta}V$ $R = V + R\frac{\mathrm{P}}{E-H}V$ for P = Cauchy principle value $T = R - i\pi R\delta(E-H)T$
5. S matrix or scattering matrix, interaction domain in complex plane.	$S_{if} = \lim\limits_{t\to\infty}\langle\Psi_f^-\vert\Psi_i^+\rangle$ $S_{if} = \delta_{if} - 2\pi i\delta(E_f - E_i)(H - E_f)\langle\Psi_f^-\vert\Psi_i^+\rangle$ $S_{if} = \delta_{if} - 2\pi i\delta(E_f - E_i)T_{fi}$

toexcitation and photoemission, atomic and molecular collision, scattering processes, and reactions. Most of these processes involve collision energies from a few keV to MeV energies, and below 1 eV molecular rotation-vibration excitations are involved. In this energy region, plasma collective phonon and plasmon states can occur in an ionized medium (Rauscher, 1968). At the upper end of this energy range, we have K-shell electron ionization and excitation and electron capture including L and M electron processes (Briggs and Macek, 1991).

Details on the Schrödinger or Lippmann-Schwinger equation and other quantum mechanical terminology, such as Green's functions, Born-Oppenheimer approximation, diagonalization, and transition probabilities, can be found in any good standard atomic physics textbooks such as Friedrich (1990) and Yoachain (1975).

The Born approximation summarized in Table 3 has found widespread use. For example, an expansion of the cross section for ion-atom collisions can be made in terms of Z_p^n, where Z_p represents the projectile charge (Sec. 2.9.3). The transition probabilities for a given process are determined from the transition operator or T matrix, which acts on an initial incoming unperturbed plane-wave state to produce a perturbed outgoing state. If several possible "open channel" states mix, coupling can exist between these various states. The Schrödinger equation contains channel-dependent potentials with coupled channel terms dependent on the various possible states, such as angular momentum variables, of the system beyond the channel radius, which can be taken as approximately the atomic radius. One method of solving complex equations, which can be anywhere from two to more channels depending on the possible final states of the system, is to use the Green's function method. Channel wave functions are expressed in terms of the Green's function, which is the solution to the hypergeometric equation in which the time-dependent part is taken as a delta function between the possible states. The basis wave function is the solution to corresponding homogeneous equations, and a perturbative expansion is made in terms of the appropriate Green's functions and coupling potential between the various relevant states.

The number of terms in the perturbative expansion and the identification of the specific order of Born terms with specific processes depend on the problem under consideration. This technique is the basic approach in atomic and molecular collision physics. There are dominating terms that usually occur and are identified with the major features of the interactive processes. Currently of interest is the consideration of time ordering of the perturbative expansion terms and the use of time-ordering operators. It is often useful to treat scattering problems using an equivalent integral equation in place of the Schrödinger equation,

$$[-(\hbar^2/2\mu)\nabla^2 + V(\mathbf{r})]\Psi(\mathbf{r}) = E\Psi(\mathbf{r}), \tag{16a}$$

rewritten as

$$[E + (\hbar^2/2\mu)\nabla^2]\Psi(\mathbf{r}) = V(\mathbf{r})\Psi(\mathbf{r}), \tag{16b}$$

where μ is the reduced mass. We can solve this equation using the three-particle Green's function

$$G(\mathbf{r},\mathbf{r}') = \frac{-\mu}{2\pi\hbar^2}\frac{e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')}}{|\mathbf{r}-\mathbf{r}'|}; \tag{17}$$

using the equation

$$[E + (\hbar^2/2\mu)\nabla^2]G(\mathbf{r},\mathbf{r}') = \delta(\mathbf{r}-\mathbf{r}'), \tag{18}$$

the wave function

$$\Psi(\mathbf{r}) = e^{ikz} + \int G(\mathbf{r},\mathbf{r}')V(\mathbf{r}')\Psi(\mathbf{r}')d^3r', \tag{19}$$

solves the above Schrödinger equation with the $V(\mathbf{r})\Psi(\mathbf{r})$ right-hand term. To transform from space-time dependence of the variables to equivalent momentum-energy dependence, we can use the Lippmann-Schwinger equation. Its solutions automatically fulfill the free-particle boundary conditions at large distances from the scattering event.

Using the T matrix, defined as the transition operator between initial and final states, we can describe the scattering process in an analogous manner to the time-dependent perturbation theory as a transition from the incoming plane wave to the outgoing plane waves. With the incoming "unperturbed" wave function $\Psi_i(\mathbf{r}') = e^{ikz'}$, we can define the T matrix as $T_{fi} = \langle\Psi_f|T_{op}|\Psi_i\rangle$. This assumption leads directly to the Born approximation. Once in this form, we can express the transition process, $i \rightarrow f$, in terms of an angular dependent scattering amplitude.

The scattering amplitude in the Born approximation is derived by a Fourier transformation of the potential. If the potential $V = V(\mathbf{r})$ is radially symmetric, then the time-independent Schrödinger equation can be reduced to the radial Schrödinger equation. Since the basic forces in atomic and molecular systems are Coulombic, we can employ the Born-Oppenheimer approximation where we solve the electron problem quantum mechanically and treat the nucleus classically.

In general, we treat the problem of an incoming plane wave which undergoes scattering under the influence of an appropriate potential to an outgoing spherical wave. But, for example for electron scattering, the potential corresponds to a pure Coulomb potential at large separations only, and short-range modifications resulting from electron screening need to be used. Then the incoming plane wave is taken to be a distorted wave in the distorted-wave Born approximation, DWBA, leading to an additional term in the scattering amplitude.

Other forms of wave functions such as Hartree-Fock, harmonic oscillator, or rotational wave functions can be used for various molecular systems with internal-state quantum numbers of vibration, rotation, excitation, and ionization. The cross section can be expressed in terms of the T matrix and scattering amplitude, as an expansion which yields the T- or S-matrix elements (see Table 3).

Symbolically, we represent the total cross section in terms of the T matrix as $\sigma = \tilde{T}^*T$ where $\tilde{T}$ is the transpose of T and T^* is the complex conjugate of T. The notation $T^\dagger$ is used for $\tilde{T}^*$. Furthermore, the R-matrix or reaction method is extremely useful for describing isolated states and is expressed in terms of the regular and irregular Coulomb wave functions, F_l and G_l, respectively. Both functions solve the free equation, but the irregular wave function does not vanish at $r = 0$. Their general form can be expressed asymptotically for $r \rightarrow \infty$ in terms of sine and cosine, respectively, with arguments that depend on momentum k, radius r, and Coulomb phase shift δ_l. For smaller radial dependence at lower energies, the regular and irregular Coulomb-wave functions are expressed in terms of ordinary Bessel functions. Because of the form of the dependence of the R matrix on the potential, this approach is more suitable for describing higher-energy processes. Observed experimental data for the energy and angular dependence of secondary electron emission following fast ion-atom collisions (see Sec. 2.7) can be compared with the BEA (binary encounter approximation; Garcia *et al.*, 1973), PWBA (plane wave Born approximation; Merzbacher and Lewis, 1958; Fano, 1963), and CDW-EIS (continuum distorted wave eikonal initial state) calculations. The BEA and PWBA calculations account only for the interaction between the static Coulomb fields of the projectile and target electrons. A representative illustration of the PWBA and CDW-EIS approximation is shown in Fig. 7 from the work of Schneider and co-workers. The collision system is 3.5-MeV/u Fe^{22+} on He. A number of statistical methods applied to molecular and atomic collision processes have been developed over the last two decades. One of the most successful is the n-body classical trajectory Monte Carlo method (nCTMC), which has been used extensively by Olson and co-workers (Cocke and Olson, 1991) to describe molecular and atomic collision processes.

Various semiempirical expansions can be represented by Feynman graphical techniques (Mattuck, 1992), which are frequently used in many-body collision physics. The approaches briefly described in Table 3 and other theories currently in use can assist us in analyzing the vast amount of diverse experimental collision data. By analyzing these models, we will be able to design new experiments and better understand atomic and molecular interactions. This fundamental knowledge is also of importance for applications to industrial problems.

4. SUMMARY

We have given some characteristic examples of molecular and atomic collision processes in three broad categories, namely, fragmentation, spectroscopy, and few-body collision dynamics. With each improvement in experimental methods, ion-beam facilities, and computational physics, the understanding of complex yet fundamental collision processes has advanced dramatically in the last few years. In fact, the richness of recent molecular and atomic collision research has yielded valuable insight into few-body dynamics. Nonetheless, a detailed quantum mechanical description of many-body collision processes has still to be developed. In particular, few-body problems still require extensive investigations to elucidate the interplay of electron-correlation, relativistic, and quantum electrodynamic effects in collective particle interactions as a function of the velocity, charge state, atomic number, participating electrons, and atomic and nuclear spins. Also, phenomena like multipole transitions, or hyperfine-induced mixing in multiply charged heavy ions, require precision atomic many-body calculations. The availability of new highly charged ion sources, ion-storage rings, and ultrashort-duration high-power laser systems has opened completely new avenues of research in atomic and molecular collision physics that are of vital interest for pure and applied research in the academic and industrial environment. The advances at the forefront of science and technology are closely related to the fundamental nature of matter and collisional phenomena involving atoms, molecules, and ions. Collisional processes are, for example, crucial to the understanding of auroral, magnetic, and ionospheric phenomena and also contribute to the understanding of plasmas, x-ray lasers, astrophysics, ion implantation, surface diagnostics and modification, nanostructure and subnanostructure technologies, and exotic new electron-electron, electron-ion, ion-ion, and electron-nuclear interactions.

ABBREVIATIONS

BEA: Binary encounter approximation.

BEP: Binary encounter peak.

CDW: Continuum distorted-wave approximation.

CRYSIS: Cryogenic electron-beam ion source.

CSW-EIC: Continuum distorted-wave eikonal initial-state calculations.

DDCS: Double-differential electron cross section.

EBIT: Electron-beam ion trap.

EBIS: Electron-beam ion source.

ECR: Electron cyclotron resonance.

ELP: Electron loss peak.

EUV: Extreme ultraviolet.

HCHZI: Highly charged high-*Z* ions.

HISR: Heavy-ion storage ring.

LPES: Laser photodetachment electron spectroscopy.

nCTMC: *n*-body classical trajectory Monte Carlo method.

PWBA: Plane-wave Born approximation.

GLOSSARY

Auger Effect and Autoionization: A radiationless decay of a highly excited molecular or atomic state where one electron undergoes a transition to a lower-lying state and another electron is transferred to the adjacent continuum, resulting in a free electron and the formation of a residual ion.

Born Approximation: High-energy limit of the scattering amplitude in ion-atom collisions in which the Born series converges at sufficiently high projectile velocities.

Cross Section: The cross section (σ) represents an effective area seen by the incoming projectile and has the units cm^2 per atom.

Works Cited

Alonso, J. R., Dietrich, D. D., Prior, M. H., Schneider, D. H. G. (Eds.) (1989), "Proceedings of a Workshop on Highly Charged Ions: New Physics and Advanced Techniques," *Nucl. Instrum. Methods Phys. Res.* **B43**, 265–470.

Amusia, M. Ya (1990), *Atomic Photoeffect*, New York: Plenum Press.

Andersen, L. H. (1988), in: H. B. Gilbody, W. R. Newell, F. H. Read, A. C. H. Smith (Eds.), *Electronic and Atomic Collisions*, Amsterdam: North-Holland, p. 451.

Andersen, N., Gallagher, J., Hertel, I. V. (1988), *Phys. Rep.* **165**, 1–188.

Andrä, H. J. (1979), in: W. Hanle, H. Kleinpoppen (Eds.), *Progress in Atomic Spectroscopy*, New York: Plenum Press, p. 829.

Andrä, H. J. (1989), in: Marrus, R. (Ed.), *Physics of Highly Ionized Atoms*, NATO Advanced Study Institutes, Series B, Physics, Vol 201, New York: Plenum.

Bashkin, S. (Ed.) (1976), *Beam-Foil Spectroscopy*, Topics in Current Physics No. 1, Berlin: Springer.

Bashkin, S., Träbert, E., Thiede, D. A., Sercel, P. C., Lin, P.-C., Li, M.-L., Jenkins, D. G., Shemansky, D. E., Wells, K. Bruch, R., Fülling, S., DeWitt, D. (1991), *Phys. Rev. A*43, 3553–3562.

Bashkin, S., Stoner, J. O., Jr. (1975), *Atomic Energy Levels and Grotrian Diagrams*, Vol. 1, Amsterdam: North-Holland; (1978), Vol. 2; (1981), Vol. 3.

Berry, H. G. (1977), *Rep. Prog. Phys.* **40**, 155–217.

Betz, D. (1983), in: S. Datz (Ed.), *Applied Atomic Collision Physics*, Vol. 4, Orlando, FL: Academic, pp. 1–42.

Bird, J. R., Williams, J. S. (1989), *Ion Beams for Materials Analysis*, Sydney: Academic.

Bisgaard, P., *et al.* (1980), in: H. Kleinpoppen, J. F. Williams (Eds.), *Coherence and Correlation in Atomic Collisions*, New York: Plenum.

Briand, J. P., De Billy, L., Charles, P., Essabaa, S., Briand, P., Geller, R., Desclaux, J. P., Bliman, S., Ristori, C. (1990), *Phys. Rev. Lett.* **65**, 159–162.

Briggs, J. S., Kleinpoppen, H. J., Lutz, H. O. (1988), *Fundamental Processes of Atomic Dynamics*, NATO Advanced Study Institutes, Series B, Physics, New York: Plenum.

Briggs, J. S., Macek, J. H. (1991), in: Sir David Bates, Benjamin Bederson (Eds.), *The Theory of Fast Ion-Atom Collisions*, Advances in Atomic, Molecular, and Optical Physics Vol. 28, Boston: Academic.

Brown, I. G. (Ed.) (1989), *The Physics of Ion Sources*, New York: Wiley.

Bruch, R., Dubé, L. J., Träbert, E., Heckmann, P. H., Raith, B., Brand, K. (1982), *J. Phys. B* **15**, L857–L862.

Burke, P. A., Eissner, W. (Eds.), (1983), *Atoms in Astrophysics*, New York: Plenum.

Chetioui, A., Politis, M. F., Wohrer, K., Rozet, J. P., Blumenfeld, L., Touati, A., Vernhet, D., Roncin, P., Gaboriaud, M. N., Barat, M., Laurent, H. (1988), in: H. B. Gilbody, W. R. Newell, F. H. Read, A. C. H. Smith (Eds.), *Electronic and Atomic Collisions*, Amsterdam: North-Holland, p. 309.

Church, D. A. (1988), *Physica Scr.* **T22**, 164–170.

CIAMDA (1987), *An Index to the Literature on Atomic and Molecular Collision Data Relevant to Fusion Research*, Vienna: International Atomic Energy Agency.

Clark, M. W., Schneider, D., Dewitt, D., McDonald, J. W., Bruch, R. Safronova, U. I., Tolstikhina, I. Yu., Schuch, R. (1993), *Phys. Rev. A* **47**, 3983–3997.

Cocke, C. L., Olson, R. E. (1991), *Phys. Rep.* **205**, 153–219.

Cowan, R. D. (1981), *Theory of Atomic Structure and Spectra*, Berkeley, CA: Univ. of California Press.

Crasemann, B. (Ed.) (1975), *Atomic Inner-Shell Processes*, Vols. 1 and 2, New York: Academic.

Datz, S. (Ed.) (1983), *Applied Atomic Collision Physics*, Vol. 4, *Condensed Matter*, Orlando, FL: Academic.

Datz, S. (1987), *Nucl. Instrum. Methods Phys. Res.* **B24/25**, 3–10.

Datz, S. (1990), in: A. Dalgarno, R. S. Freund, M. S. Lubell, T. B. Lucatorto (Eds.), *The Physics of Electronic and Atomic Collisions*, AIP Conference Proceedings No. 205, New York: American Institute of Physics, p. 2.

De Heer, F. J. (1982), in: C. J. Joachain, D. E. Post (Eds.), *Atomic and Molecular Physics of Controlled Thermonuclear Fusion*, New York: Plenum, pp. 269–312.

Dubé, L. (1986), in: D. C. Lorents, W. E. Meyerhof, J. R. Petersen (Eds.), *Electronic and Atomic Collisions*, Amsterdam: North-Holland, pp. 345–364.

Elton, R. C. (1990), *X-Ray Lasers*, Boston: Academic.

Fano, U. (1963), *Annu. Rev. Nucl. Sci.* **13**, 1–66.

Fano, U., Rau, A. R. P. (1988), *Atomic Collisions and Spectra*, New York: Academic.

Folkerts, L. (1992), "Charge Exchange of Multiply Charged Ions with Metal Surfaces," Ph.D. Thesis, University of Groningen.

Friedrich, H. (1990), *Theoretical Atomic Physics*, Berlin: Springer.

Fritsch, W., Lin, C. D. (1991), *Phys. Rep.* **202**, 1–97.

Fülling, S., Bruch, R., Rauscher, E. A., Weill, P. A., Träbert, E., Heckmann, P. H., McGuire, J. H. (1992), *Phys. Rev. Lett.* **68**, 3152–3155.

Garcia, J. D., Fortner, R. J., Kavanagh, T. M. (1973), *Rev. Mod. Phys.* **45**, 111–177.

Gemmel, D. S. (1980), *Chem. Rev.* **80**, 301–311.

Giese, J. P., Schulz, M., Swenson, J. H., Schöne, H., Benhenni, M., Varghese, S. L., Vane, C. R., Dittner, P. F., Shafroth, S. M., Datz, S. (1990), *Phys. Rev. A* **42**, 1231–1244.

Gudeman, C. S., Saykally, R. J. (1984), *Annu. Rev. Phys. Chem.* **35**, 387–418.

Halka, M., Bryant, H. C., Mackerrow, E. P., Miller, W., Mohagheghi, A. H., Tang, C. Y., Cohen, S., Donahue, J. B., Hsu, A., Quick, C. R., Tiee, J., Rozsa, K. (1991), *Phys. Rev. A* **44**, 6127–6129.

Hasselkamp, D., Varga, P., Groenveld, K. O., Winter, H., Kemmler, J., Rothard, H. (1992), *Particle-Induced-Electron Emission II*, Springer Tracts in Modern Physics, Vol. 123, Berlin: Springer.

Hotop, H., Lineberger, W. C. (1985), *J. Chem. Ref. Data* **14**, 731–750.

Humphries, S., Jr. (1986), *Principles of Charged Particle Acceleration*, New York: Wiley.

Ivanov, L. N. (1988), *Phys. Rep.* **164**, 315–375.

Janev, R. K., Winter, H. P. (1985), *Phys. Rep.* **117**, 265–387.

Joachain, C. J. (1975), *Quantum Collision Theory*, Amsterdam: North-Holland.

Kessel, Q. C., Fastrup, B. (1974), in: E. W. McDaniel, M. R. C. McDowell (Eds.), *Case Studies in Atomic Physics III*, Amsterdam: North-Holland, pp. 139–213.

Kleinpoppen, H., Williams, J. F. (Eds.) (1980), *Coherence and Correlation in Atomic Collisions*, New York: Plenum.

Letokov, V. S. (1989), *Lasers in Atomic, Molecular, Nuclear Physics*, Singapore: World Scientific.

Levine, M. A., Marrs, R. E., Hendersen, J. P., Knapp, D. A., Schneider, M. B. (1988), *Phys. Scr.* **T22**, 157–163.

Lindgren, I., Morrison, J. (1982), *Atomic Many-Body Theory*, Berlin: Springer.

Maier, J. D. (1989), *Ion and Cluster Ion Spectroscopy and Structure*: Amsterdam: North-Holland.

Mannervik, S. (1989), *Phys. Scr.* **40**, 28–52.

Marrus, R. (Ed.) (1989), *Physics of Highly Ionized Atoms*, NATO Advanced Study Institutes, Series B, Physics, Vol. 201, New York: Plenum.

Massey, H. (1976), *Negative Ions*, Cambridge, U.K.: Cambridge Univ. Press.

Mattuck, R. D. (1992), *A Guide to Feynman Diagrams in the Many-Body Problem*, 2nd ed., New York: Dover.

McDaniel, E. W., Mitchell, J. B. A., Rudd, M. E. (1993), *Atomic Collisions: Heavy Particle Projectiles*, New York: Wiley.

McDonald, J. W., Schneider, D., Clark, M. W., Dewitt, D. (1992), *Phys. Rev. Lett.* **68**, 2297–2300.

McGuire, J. H. (1992), *Adv. At. Mol. Opt. Phys.* **29**, 217–323.

Mehlhorn, W. (1985), in: B. Crasemann (Ed.), *Atomic Inner-Shell Physics*, New York: Plenum.

Merzbacher, E., Lewis, H. (1958), *Encyclopedia of Physics*, Vol. 34, Berlin: Springer.

Meyer, F. W., Kirkpatrick, M. I. (1991), *Proceedings of the 10th International Workshop on ECR Ion Sources*, ORNL Conf-9011136.

Miller, T. A., Bondybey, V. E. (1983), *Molecular Ions: Spectroscopy, Structure and Chemistry*, Amsterdam: North-Holland.

Mohr, P. J. (1989), in: Marrus (Ed.), *Physics of Highly Ionized Atoms*, NATO Advanced Study Institutes, Series B, Physics, Vol. 201, New York: Plenum.

Mott, N. F., Massey, H. S. (1987), *The Theory of Atomic Collisions*, 3rd ed., Vols. 1 and 2, Oxford: Clarendon.

Müller-Dethlefs, K., Schlag, E. W. (1991), *Annu. Rev. Phys. Chem.* **42**, 109–136.

Neugart, R. (1987), in: H. J. Beyer, H. Kleinpoppen (Eds.), *Progress in Atomic Spectroscopy*, Part D, New York: Plenum, p. 75.

Niehaus, A. (1990), *Phys. Rep*, **186**, 149–214.

Pegg, D. J., Thomson, J. S., Compton, R. N., Alton, G. D. (1987), *Phys. Rev. Lett.* **59**, 2267–2270.

Poth, H. (1990), *Nature* **345**, 399–405.

Rauscher, E. A. (1968), *J. Plasma Phys.* **2**, 517–541.

Rødbro, M., Bruch, R., Bisgaard, P. (1979), *J. Phys. B.* **12**, 2413–2447.

Rudd, M. E., Macek, J. H. (1974), in: E. W. McDaniel, M. R. C. McDowell (Eds.), *Case Studies in Atomic Physics III*, Amsterdam: North-Holland.

Salzborn, E., Mueller, A., Mokler, P. H. (1991), *Atomic Physics of Highly Charged Ions*, Berlin: Springer.

Scoles, G. (Ed.) (1988), *Atomic and Molecular Beam Methods*, Vol. 1, New York: Oxford Univ. Press.

Schneider, D., Clark, M. W., Penetrante, B. M., McDonald, J., DeWitt, D., Bardsley, J. N. (1991), *Phys. Rev. A* **44**, 3119–3124.

Schneider, D. H., DeWitt, D. R., Bauer, R. W., Mowat, J. R., Graham, W. G., Schlachter, A. S., Skogvall, B., Fainstein, P., Rivarola, R. D. (1992), *Phys. Rev. A* **46**, 1296–1302.

Schuch, R., Bárány, A., Danored, H., Elander, N., Mannervik, S. (1989), *Nucl. Instrum. Methods Phys. Res.* **B43**, 411–424.

Sellin, I. A. (Ed.) (1978), *Structure and Collisions of Ions and Atoms*, Topics in Current Physics, Vol. 5, New York: Springer.

Sevier, K. D. (1972), *Low Energy Electron Spectroscopy*, New York: Wiley.

Smirnov, B. M. (1982), *Negative Ions*, New York: McGraw-Hill.

Stebbings, R. F., Dunning, F. B. (1983), *Rydberg States of Atoms and Molecules*, Cambridge, U.K.: Cambridge Univ. Press.

Stöckli, M. P., Ali, R. M., Cocke, C. L., Raphaelian, M. L. A., Richard P., Tipping, T. N. (1992), *Rev. Sci. Instrum.* **63**, 2822–2824.

Stolterfoht, N. (1978), in: I. A. Sellin (Ed.), *Structure and Collisions of Ions and Atoms*, Topics in Current Physics Vol. 5, New York: Springer.

Stolterfoht, N. (1987), *Phys. Rep.* **146**, 315–424.

Toburen, L. H. (1991), *Atomic and Molecular Physics in the Gas Phase*, Physical and Chemical Mechanisms in Molecular Radiation Biology, New York: Plenum.

Tully, J. C., White, C. W. (1977), *Inelastic Ion Surface Collisions*, New York: Academic.

Vainshtein, L. A., Shevelko, V. (1986), *Structure and Properties of Ions in Hot Plasmas*, Moscow: Nauka.

Waki, I., Kassner, S., Birkl, G., Walther, H. (1992), *Phys. Rev. Lett.* **68**, 2007–2010.

Wang, H., Bruch, R., Hao, F., Fuelling, S., Xu, Z., Wang, Z., Rauscher, E. (1993), *Nucl. Instrum. Methods Phys. Res.* **B79**, 114–116.

Wilson, S., Grant, I. T., Gyorffy, B. L. (1991), *The Effects of Relativity in Atoms, Molecules, and The Solid State*, New York: Plenum.

Wollnik, H. (1987), *Optics of Charged Particle Beams*, Orlando, FL: Academic.

Further Reading

Bransden, B. H., Joachain, C. J. (1986), *Physics of Atoms and Molecules*, London: Longman.

Lin, C. D. (1993), *Review of Fundamental Processes and Applications of Atoms and Ions*, Singapore: World Scientific.

McDaniel, E. W. (1989), *Atomic Collisions*, New York: Wiley.

McDaniel, E. W., Mitchell, J. B. A., Rudd, M. E. (1993), *Atomic Collisions: Heavy Particle Projectiles*, New York: Wiley.

Svanberg, J., (1992), *Atomic and Molecular Spectroscopy*, Berlin: Springer.

MOLECULAR BEAM EPITAXY

R. J. FISCHER AND A. Y. CHO, *AT&T Bell Laboratories, Murray Hill, New Jersey, U.S.A.*

Introduction 471

1. **History and General Principles** 471

1.1 The MBE Apparatus 472

1.2 Stoichiometry and the Surface Phase Diagram 474

1.3 Flux Control of Molecular Beams 475

2. **III-V Compounds** 476

2.1 Conventional Techniques 476

2.2 *In Situ* Diagnostics and RHEED Oscillations 477

2.3 Gas-Phase Sources 479

2.4 Device Examples 481

3. **Group-IV Materials** 482

3.1 Apparatus and Growth Technique 482

3.2 Growth of SiGe 483

4. **II-VI Compounds** 483

4.1 Narrow-Band-Gap Materials 484

4.2 Wide-Band-Gap Materials ... 484

5. **Other Material Systems** 485

5.1 Epitaxial Metal Films 485

5.2 Epitaxial Silicides 485

5.3 Superconducting Materials .. 487

Glossary 488

Works Cited 488

Further Reading 489

INTRODUCTION

The term *molecular beam epitaxy* (MBE) encompasses a variety of specific techniques for deposition of thin-film electronic materials. These thin films are the basis of many electronic and photonic devices. Cost and reliability factors place stringent requirements on the purity, number of defects, and controllability with which these films are deposited. The molecular beam epitaxy process offers a high degree of controllability and reproducibility. It is for this reason that many companies in the compound-semiconductor industry are using MBE in their processes. For example, at the time of this writing, more than half of the lasers used in compact-disk players worldwide are produced by MBE.

In the conventional form, MBE is a highly controlled evaporation process. Elemental sources are heated in an ultrahigh-vacuum environment, and thus the source material is transported to a heated substrate. Epitaxy, which is a process by which a deposited film takes on a crystalline structure that is related to that of the substrate, occurs on the substrate in this process. The reason why this occurs is that the substrate, being heated, provides the impinging molecules with enough energy to be mobile on the substrate, such that these atoms or molecules can find appropriate lattice sites on which to condense.

The typical rates used for deposition are about one atomic layer per second. This view of the process allows one to see why there is such controllability—the process is building up a thin film one atomic layer at a time. This feature allows many phenomena of quantum physics to be investigated. In what follows, a more detailed description of MBE will be given, and many variants of the process will be described, and material and device examples will be given.

1. HISTORY AND GENERAL PRINCIPLES

The modern-day MBE apparatus has evolved from its earlier form, which was an experimental system designed to measure adsorption and desorption kinetics on GaAs surfaces. From these studies, a method was discovered for epitaxially depositing GaAs. In contrast, today's multiwafer MBE systems are true production tools in the semiconductor industry. Figure 1 shows the first MBE effusion-cell assembly, while Fig. 2 shows a modern production system. In this section, the

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

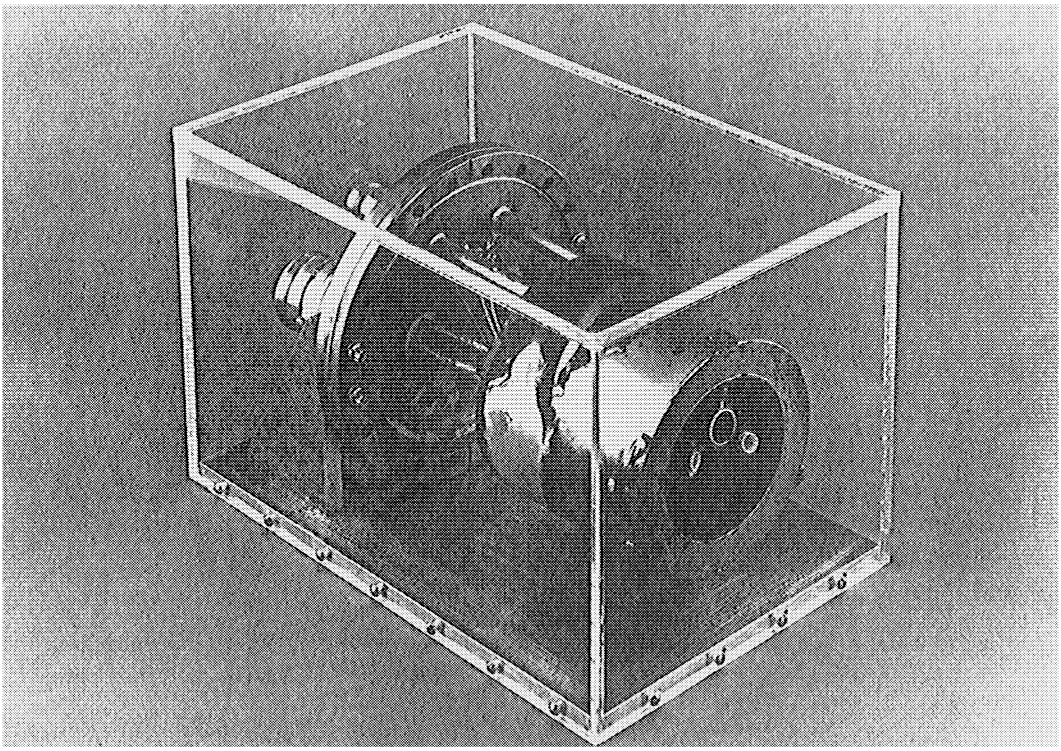

FIG. 1. Photograph of early MBE system.

hardware of MBE will be described, highlighting many of the innovations in design through the generations that have brought the technique to its present maturity.

1.1 The MBE Apparatus

Figure 3 shows a cutaway view of a modern MBE system and highlights most of the important features. MBE is an ultrahigh-vacuum technique with system base pressures of typically 10^{-11} Torr (a hundred-trillionth of our atmospheric pressure). These vacuum levels are required in order to maintain purity of the deposited film. To achieve such a vacuum, liquid-nitrogen-cooled panels are fitted to surround the entire deposition chamber. This liquid-nitrogen cooling substantially reduces the amount of water vapor present. Another feature that aids the vacuum is the use of a sample-exchange load lock. In this way, the deposition chamber need not be exposed to atmospheric pressure each time a new sample is introduced. Sample transfer mechanisms are required for this feature and on modern systems have become completely automated.

With several important differences, MBE is in essence a specialized form of vacuum evaporation. The so-called effusion cells that are aimed at the substrate (e.g., a GaAs wafer) are the sources in the technique and form the molecular beams. Figure 4 is a photograph of such an effusion cell. The cell has a crucible (usually made of pyrolytic boron nitride) in which a charge of the appropriate element is contained. For growth of GaAs, three such effusion-cell sources would usually be required; one for Ga, one for As, and one for a dopant. The crucible with its elemental charge is heated in the cell by resistive heater windings to a temperature that will produce the molecular beams by evaporation (sublimation, in the case of solid charges). In front of each effusion cell a shutter serves to block the beam when that particular element is not wanted. These shutters, which swing in and out of the beam path, allow composition and doping profiles that are atomically abrupt. A typical MBE system

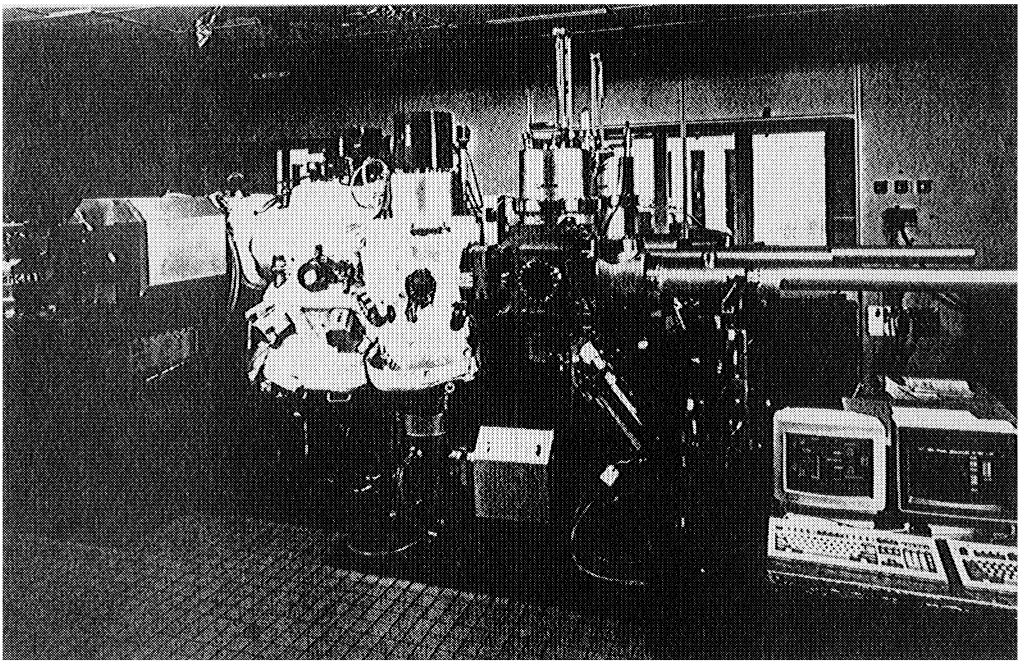

FIG. 2. Modern production MBE system. This system is fully automated and can deposit films on several wafers simultaneously. (Photo courtesy of ISA-Riber.)

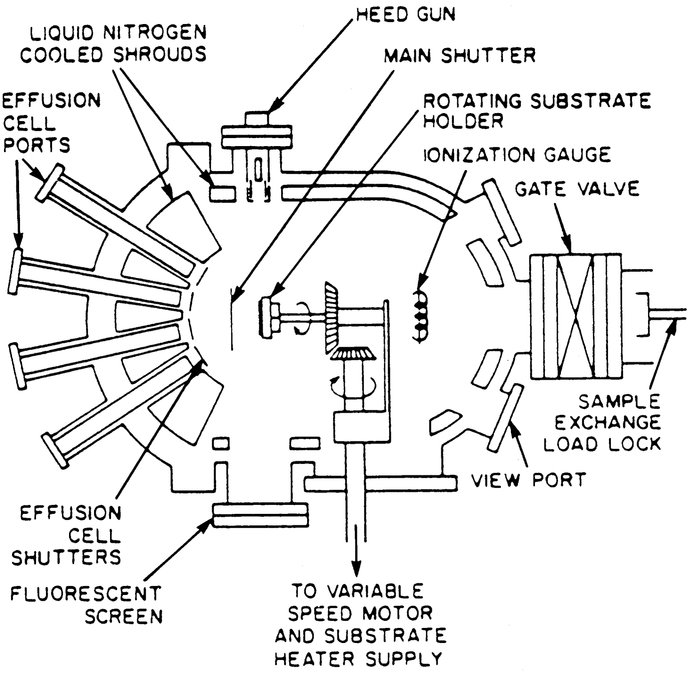
LIQUID NITROGEN COOLED SHROUDS
HEED GUN
MAIN SHUTTER
EFFUSION CELL PORTS
ROTATING SUBSTRATE HOLDER
IONIZATION GAUGE
GATE VALVE
SAMPLE EXCHANGE LOAD LOCK
VIEW PORT
EFFUSION CELL SHUTTERS
FLUORESCENT SCREEN
TO VARIABLE SPEED MOTOR AND SUBSTRATE HEATER SUPPLY

FIG. 3. Cutaway view of a modern MBE system viewed from the top. (After Cho and Cheng, 1981.)

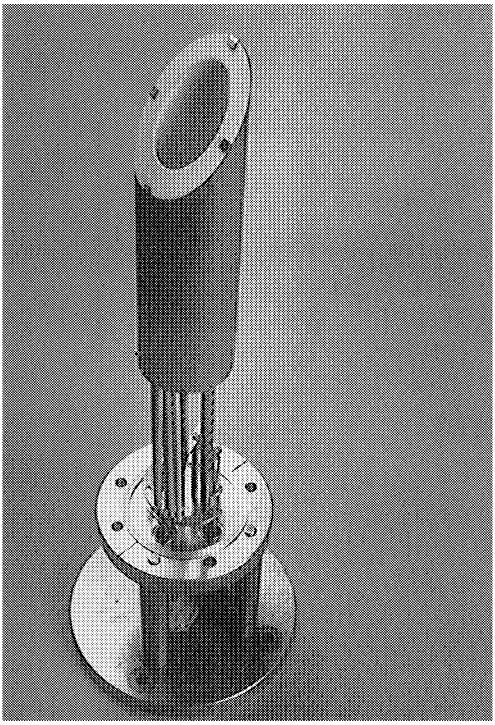

FIG. 4. Photograph of an effusion-cell source. (Photo courtesy of ISA-Riber.)

has eight effusion cells with eight shutters, and a main shutter in front of the substrate.

As outlined above, the effusion cells generate molecular beams that are directed toward the substrate. A heated substrate is required, because epitaxy is a process of depositing a crystalline thin film on a crystalline substrate such that the film has a well-defined crystalline relationship to the substrate. When the molecular beams impinge on the substrate they are absorbed (in other words, they stick to the surface but are not incorporated into the film). Since the substrate is heated, these atoms have the ability to migrate on the surface until they encounter an appropriate site to incorporate. If the substrate were not heated, epitaxy would not generally occur because the impinging atoms would not have this mobility, resulting in an amorphous film.

One of the most significant advantages of MBE is the uniformity and reproducibility that can be achieved. This surface atomic motion mentioned above happens on the scale of atomic dimensions. The rate of deposition is controlled by the intensity of the molecular beam. Therefore, the uniformity of the film depends on the uniformity of the beams. For many reasons, it is impractical to design a source that by itself produces a highly uniform beam. Therefore, a rotating sample (substrate) holder is used to average out beam nonuniformities. The design of these rotating mechanisms is again not a simple problem, and these designs have evolved over many generations. However, it is this innovation that has allowed the degree of uniformity necessary for production of optical and electronic devices (Cho and Cheng, 1981).

Since MBE was a development prompted by investigations of surface physics, many of the surface analytical methods can be used to monitor the MBE process. Such analyses are termed *in situ* diagnostics. One of the most used techniques is that of reflection high-energy electron diffraction (RHEED). An electron gun is mounted in the deposition chamber such that a beam of high-energy (approximately 10 keV) electrons hits the sample at a grazing angle. This technique can monitor the condition of the surface of a growing film and will be described in more detail in a later section. Another diagnostic commonly used in MBE is beam flux measurement. In this technique, an ionization gauge, which measures pressure, is mounted in such a way that it can be positioned in the path of the molecular beams. The pressure reading from this gauge can be converted into a molecular flux (number of molecules passing a unit area per unit time), which is directly related to the deposition rate. In this way, the deposition rate is measured without actual deposition. The pressures measured in this way are called beam equivalent pressures. The relative flux values can be measured to four decimal places if a digital voltmeter is used in conjunction with the commercial gauge readout.

1.2 Stoichiometry and the Surface Phase Diagram

In general, for a compound semiconductor such as GaAs, one of the constituent species will evaporate more readily than the other or, in other words, have a higher vapor pressure. For example, when GaAs is heated, As, which has a higher vapor pressure than Ga, will preferentially evaporate. Therefore it is not desirable in general to use a III-V compound as a source to deposit the same III-V material, because the ratio of the group-V to group III element varies as a function of the effusion-cell temperature. For high growth rates, i.e., high effusion-cell temperatures, the arsenic component in the GaAs source will quickly be depleted.

Historically, several schemes were used to achieve stoichiometric GaAs films. The first method was the "flash evaporation" technique (Richards *et al.* 1963), in which small particles of the source material were completely vaporized on a hot filament. Since the temperatures were very high, essentially all of the source compound was evaporated simultaneously, thus producing a stoichiometric beam flux. A second method was to use two crucibles situated in a way that caused large variations in beam flux across a substrate. In this way the optimum beam ratio would be achieved at some point on the wafer (Richards *et al.*, 1963). The third method, the "three-temperature technique" (Gunther, 1958), approximate that used for MBE. Again, two separate crucibles, one for Ga and one for As, were used and were held each at a different temperature. In this experiment, stoichiometry was achieved in the film because any excess As was reevaporated from the substrate. Note that this technique did not produce single-crystal epitaxial films, because of the vacuum conditions and the amorphous glass substrate. For a more complete account, the reader is referred to Cho and Arthur (1975).

As the development of MBE proceeded, pulsed-beam experiments revealed new insights into the behavior and interaction of Ga and As beams on the GaAs surface. It was found that adsorption and desorption of Ga on the (111) GaAs face obeyed a simple first-order rate equation with temperature-dependent time constants (Arthur, 1968; Cho and Arthur, 1975). In contrast, the behavior of As was found to depend on the surface coverage of Ga. Therefore, stoichiometry could be achieved by providing a significantly higher flux of As such that all Ga reacts and excess As would not stick to the surface. This insight formed the basis for epitaxial growth of III-V compounds by MBE.

While the above-mentioned conclusion is a useful concept, it is somewhat oversimplified. Changes in surface structure occur as the As-to-Ga ratio and substrate temperature are varied. In RHEED studies, various surface structures have been identified and compiled to form the so-called surface phase diagram. Figure 5 shows such a diagram for the GaAs(100) surface (Cho, 1971b, 1983). GaAs is typically grown under As-stabilized conditions because stoichiometry is more readily obtained there.

1.3 Flux Control of Molecular Beams

As already outlined, effusion-cell sources are used to generate the molecular beams used for epitaxy. Each cell contains a thermo-couple for temperature measurement and regulation. Knowing the geometry of the apparatus and the temperature, the beam flux can be calculated as follows:

$$\Gamma = p(T)AN/(2\pi MRT)^{1/2}, \qquad (1)$$

where Γ is the number of molecules leaving the effusion-cell aperture, of area A, per second (flux), $p(T)$ is the vapor pressure of the element at the cell temperature T (in kelvins), N is Avogadro's number, M is the mass, and R is the gas constant. If the substrate is positioned at a distance d from the aperture,

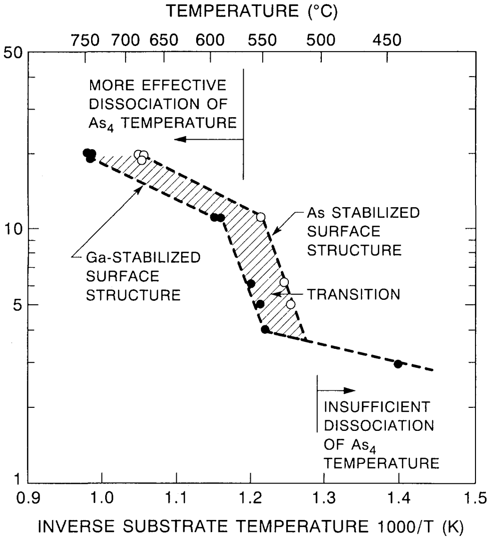

FIG. 5. Surface phase diagram for GaAs on GaAs(100). The plot shows the surface condition at a given temperature and As-to-Ga flux ratio (After Cho, 1983.)

the expression for the number of molecules striking the substrate per unit area is

$$G = 1.118 \times 10^{22} p(T) A/d^2 \times (MT)^{1/2} \text{ molecules cm}^{-2}\text{ s}^{-1}, \qquad (2)$$

where $p(T)$ is expressed in Torr.

Once the beam flux for the group-III elements is established, there is a one-to-one correspondence to the deposition rate. In III-V epitaxy, it is usually the case that alloys (or mixtures) of different III-V compounds are deposited. One important example would be the aluminum-gallium-arsenide ($Al_xGa_{1-x}As$) system. Compositional control of such ternary compounds also follows from establishment of beam flux. For example, the amount of Al or the mole fraction x in $Al_xGa_{1-x}As$ can be found as follows:

$$x = G(\text{AlAs})/[G(\text{GaAs}) + G(\text{AlAs})], \qquad (3)$$

where G(AlAs) and G(GaAs) are the respective growth rates for AlAs and GaAs and further are directly related to the beam fluxes of Al and Ga. Note that the deposition rates and composition are independent of As flux.

The flux equation above is valid for any effusion source, including dopants. However, the beam flux of a dopant source is usually many orders of magnitude lower than a constituent and therefore is usually not measurable directly. Therefore the beam flux is inferred by growing thin films of, say, GaAs with dopant sources set at various temperatures. Dopant concentrations are then measured *ex situ*. Figure 6 shows a typical plot of dopant concentration versus temperature for many dopants used for GaAs with a nominal growth rate of 1 μm/h. Si and Be are the most commonly used.

2. III-V COMPOUNDS

Molecular beam epitaxy, as outlined in the preceding section, was developed as a growth method for III-V compounds. While many of the original techniques for MBE growth are still in use today, there have also been numerous advances in the technology. In this section, the growth techniques for III-V compounds will be described.

2.1 Conventional Techniques

Given the understanding of the growth process outlined in Sec. 1, the most crucial element of epitaxial growth is that of obtaining a clean and defect-free starting surface. As outlined earlier, epitaxy is a process of depositing a crystalline material on a substrate, also crystalline, such that the structure of the deposited film has a well-defined relationship to that of the substrate. If the starting surface has defects or is contaminated, the overlayer will be inhibited from replicating the crystal surface of the substrate. Under these conditions the deposited film will be of little use.

The technique for substrate surface preparation developed early in MBE technology is still predominantly in use today. The method consists of first polishing, to achieve a nearly atomically smooth surface, and then a chemical etching step to remove surface damage. After this etching step, a thin oxide layer is formed on the surface by immersion in deionized water. This oxide layer protects the surface of the wafer until it can be loaded into the deposition system. Once ready to initiate growth, the wafer is heated to about 580 °C where the oxide layer is desorbed from the surface. When samples prepared and heated in this way are analyzed, there are no detectable surface contaminants. Figure 7 demonstrates this (Cho, 1983), showing Auger-electron spectra both before and after the heating step. Auger-electron spectroscopy (*q.v.*) is an analytical technique that can detect and is especially sensitive to elemental species on a surface.

Once the high-quality surface is obtained in this way, epitaxial growth is initiated. It is beyond the scope of this article to go into detail about specific procedures and detailed recipes used for growth of the various compounds. The power of MBE is the ability to make highly abrupt heterostructures between various compound or elemental materials. Growth of these structures requires optimization of growth parameters.

The growth of these heterostructure materials can be quite complex. Figure 8 shows a cross-sectional view of an innovative laser structure called the vertical-cavity surface-emitting laser. In this structure very precise control of layer thickness and composition is required. As can be seen, the interfaces are

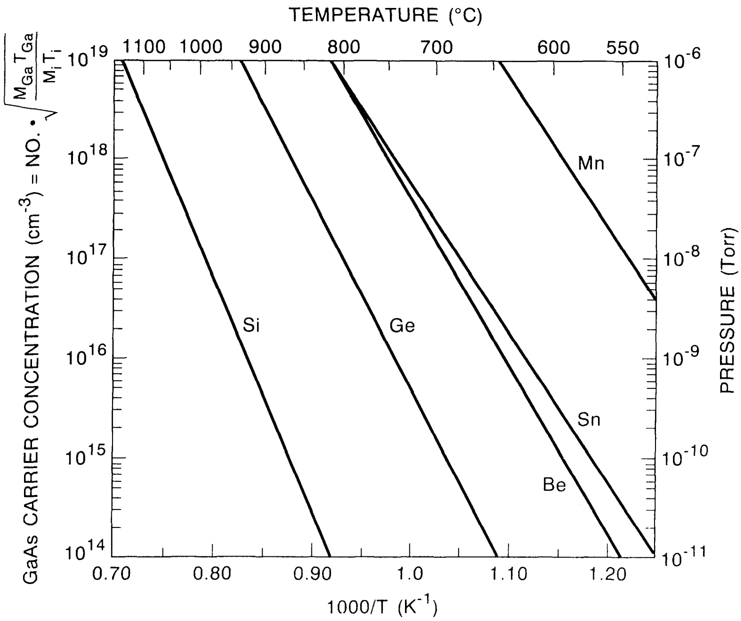

FIG. 6. Dopants used in MBE growth of GaAs. Concentration is plotted as a function of effusion-cell temperature. (After Cho, 1983.)

quite abrupt. In fact, the control offered by MBE allows precise control to the monolayer level (Cho, 1971a). The type of structure shown in Fig. 8 is called a compositional superlattice. It is also possible to modulate doping on this dimensional scale, producing what would be called a doping superlattice.

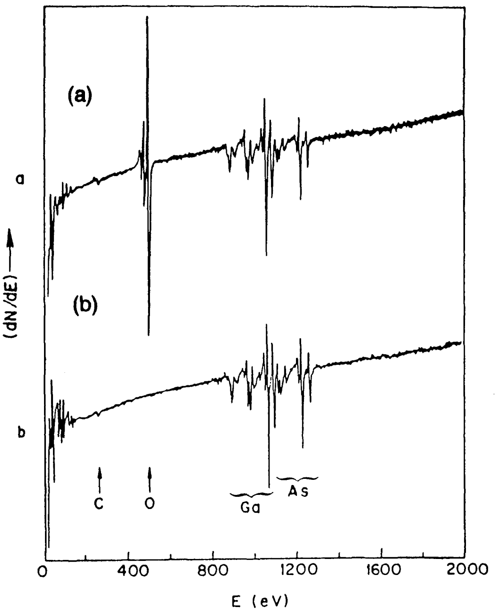

FIG. 7. Auger-electron spectra of GaAs (curve *a*) before and (curve *b*) after heating to desorb the oxide. (After Cho, 1983.)

2.2 *In Situ* Diagnostics and RHEED Oscillations

The above discussion was centered on the importance of clean surfaces. The process of MBE, as pointed out earlier, is a clean-surface technique. In this section we will discuss the methods used to examine surface cleanliness and structure. Because MBE is done in an ultrahigh-vacuum environment, many surface analytical techniques may be used. The important point is that these techniques may be used during the deposition—so called *in situ* analysis.

One of the most important tools of MBE growth is reflection high-energy electron diffraction (RHEED), which was introduced in Sec. 1.1. Figure 9 shows the geometry of the RHEED apparatus. A collimated beam of high-energy electrons is directed toward the sample at a grazing angle (usually 1° to 3°). Because the de Broglie wavelength of an electron at this energy is comparable to the atomic spacings on the surface, a diffraction pattern is formed on a fluorescent screen mounted opposite the electron source. Prior to the use of this geometry, diffraction experiments were carried out by diffracting

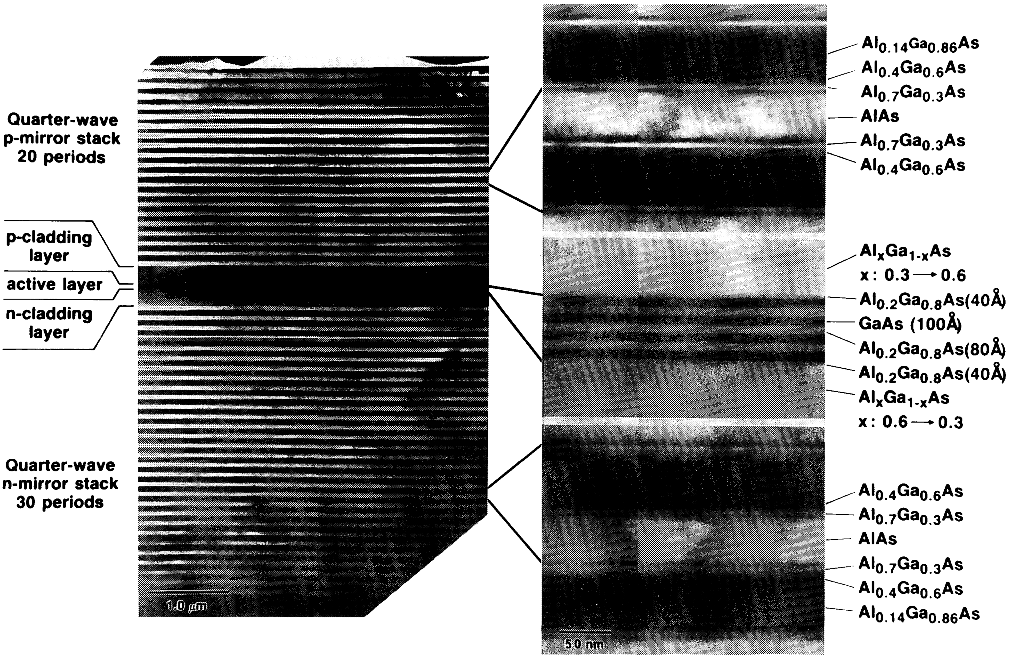
Quarter-wave p-mirror stack 20 periods
p-cladding layer
active layer
n-cladding layer
Quarter-wave n-mirror stack 30 periods
1.0 μm
50 nm
$Al_{0.14}Ga_{0.86}As$
$Al_{0.4}Ga_{0.6}As$
$Al_{0.7}Ga_{0.3}As$
AlAs
$Al_{0.7}Ga_{0.3}As$
$Al_{0.4}Ga_{0.6}As$
$Al_xGa_{1-x}As$
x: 0.3 → 0.6
$Al_{0.2}Ga_{0.8}As$(40Å)
GaAs (100Å)
$Al_{0.2}Ga_{0.8}As$(80Å)
$Al_{0.2}Ga_{0.8}As$(40Å)
$Al_xGa_{1-x}As$
x: 0.6 → 0.3
$Al_{0.4}Ga_{0.6}As$
$Al_{0.7}Ga_{0.3}As$
AlAs
$Al_{0.7}Ga_{0.3}As$
$Al_{0.4}Ga_{0.6}As$
$Al_{0.14}Ga_{0.86}As$

FIG. 8. Cross-sectional view of an MBE-grown vertical-cavity surface-emitting laser. The complexity and precision required in this structure are readily achieved by MBE.

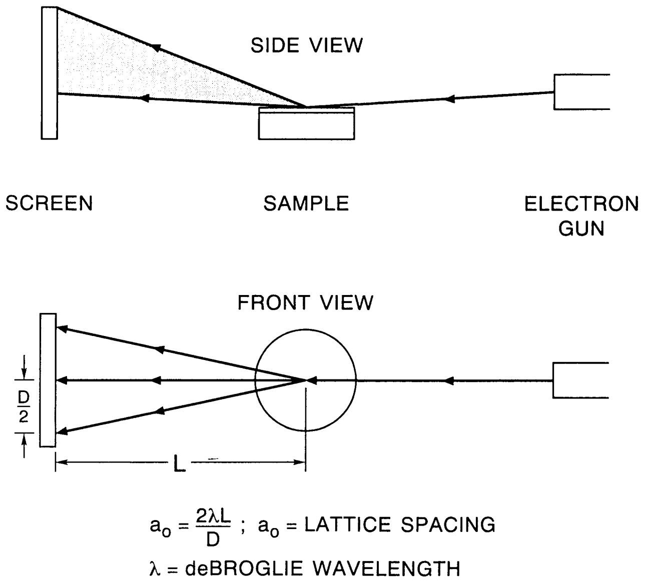
SIDE VIEW
SCREEN
SAMPLE
ELECTRON GUN
FRONT VIEW
D/2
L
$a_o = \frac{2\lambda L}{D}$; a_o = LATTICE SPACING
λ = deBROGLIE WAVELENGTH

FIG. 9. Geometry of RHEED. This configuration permits observation of diffraction patterns during deposition.

electrons to a screen situated behind the substrate [low-energy electron diffraction (LEED)]. This geometry could not be used during deposition because the sample could not face the effusion cells, and the fluorescent screens would get coated very quickly by the sources.

The diffraction pattern formed by RHEED geometry can be used to distinguish the surface condition (Cho, 1971b), whether primarily As covered or primarily Ga covered. Crystalline surfaces have unsatisfied bonds. Therefore the surface atoms rearrange themselves so as to satisfy these so-called dangling bonds. When GaAs is grown under As-stabilized conditions (surface is primarily As covered) on the ⟨100⟩ orientation, the surface atoms rearrange themselves with a periodicity (or surface structure) twice that of the underlying bulk crystal in one direction and four times that of the bulk crystal in the other. This pattern is termed the (2 × 4) reconstruction.

The RHEED technique is useful to determine the surface cleanliness. If the surface is covered by an oxide layer, the diffraction pattern either will not appear at all or will not be clear, depending on the thickness of that oxide. In practice, this is used to determine when the oxide layer is completely removed.

The ideal diffraction pattern from a perfectly flat surface consists of lines spaced in accordance with the surface atomic arrangement. In the presence of roughness, these lines become broken or spotty in appearance. This effect is shown in Fig. 10, where growth is initiated on a slightly rough surface. As growth proceeds, this roughness tends to smooth out, and the diffraction lines become uniform streaks.

Another useful feature of RHEED in MBE growth is the RHEED intensity oscillations (Wood, 1981). MBE growth proceeds in a layer-by-layer mode. When the details of how each layer is formed are examined more closely, it is found that the individual layer starts at random places on the surface (called nucleation sites) and starts to form monolayer "islands." These islands grow in size until they run into one another and coalesce. Per the preceding discussion, such island formation is a height variation and can cause interference in the diffracted beam. The intensity would be maximum at the completion of a monolayer and minimum at some intermediate (half-monolayer) point. These oscillations, then, correspond exactly with the deposition of a monolayer. This fact may be used for calibration of growth rates, as the deposition of an individual monolayer can be accurately determined (Van Hove, 1983).

2.3 Gas-Phase Sources

The preceding discussions have centered on the use of evaporation sources. A more recent development in MBE technology has been the use of gas-phase sources. These sources are similar to those used in chemical vapor deposition [see CVD (CHEMICAL VAPOR DEPOSITION)]. The important distinction is that use of these sources in MBE is done at much lower pressures and at much lower flow rates. Since pressures are kept very low, individual gas molecules do not interact, and therefore the beam nature of MBE is preserved.

There are several variations of technique for gas-phase MBE. The sources for group-III elements are typically organometallic (such as trimethyl gallium), and the group V are typically group V hybrides (such as arsine or phosphine) (Tsang, 1985). In principle, it is possible to use combinations of solid and gas-phase sources, but the two most common configurations are using either all gas-phase sources (organometallic group III and hydride group V) or standard evaporative sources for group III and hydride group V (Panish, 1980).

The initial motivation for using gas-phase sources for MBE was to grow phorphorus-containing compounds. It is difficult to use solid phosphorus evaporation sources, because the evaporation of phosphorus is not easy to control and reproduce. While recent innovations in valved cracker-cell design hold promise for obtaining this control and reproducibility in solid-evaporation P sources, it is quite simple to control the flow of a gas.

The design of a gas-phase source includes a source, such as a bubbler for organometallics or a gas cylinder for hydrides, a flow-control device, such as an electronic mass-flow controller, and an injector. The metal atom in an organometallic compound is usually not strongly bound to the organic radicals, and therefore the heat of the substrate is more than adequate to crack the molecule. Thus, the design of a organometallic injector is only re-

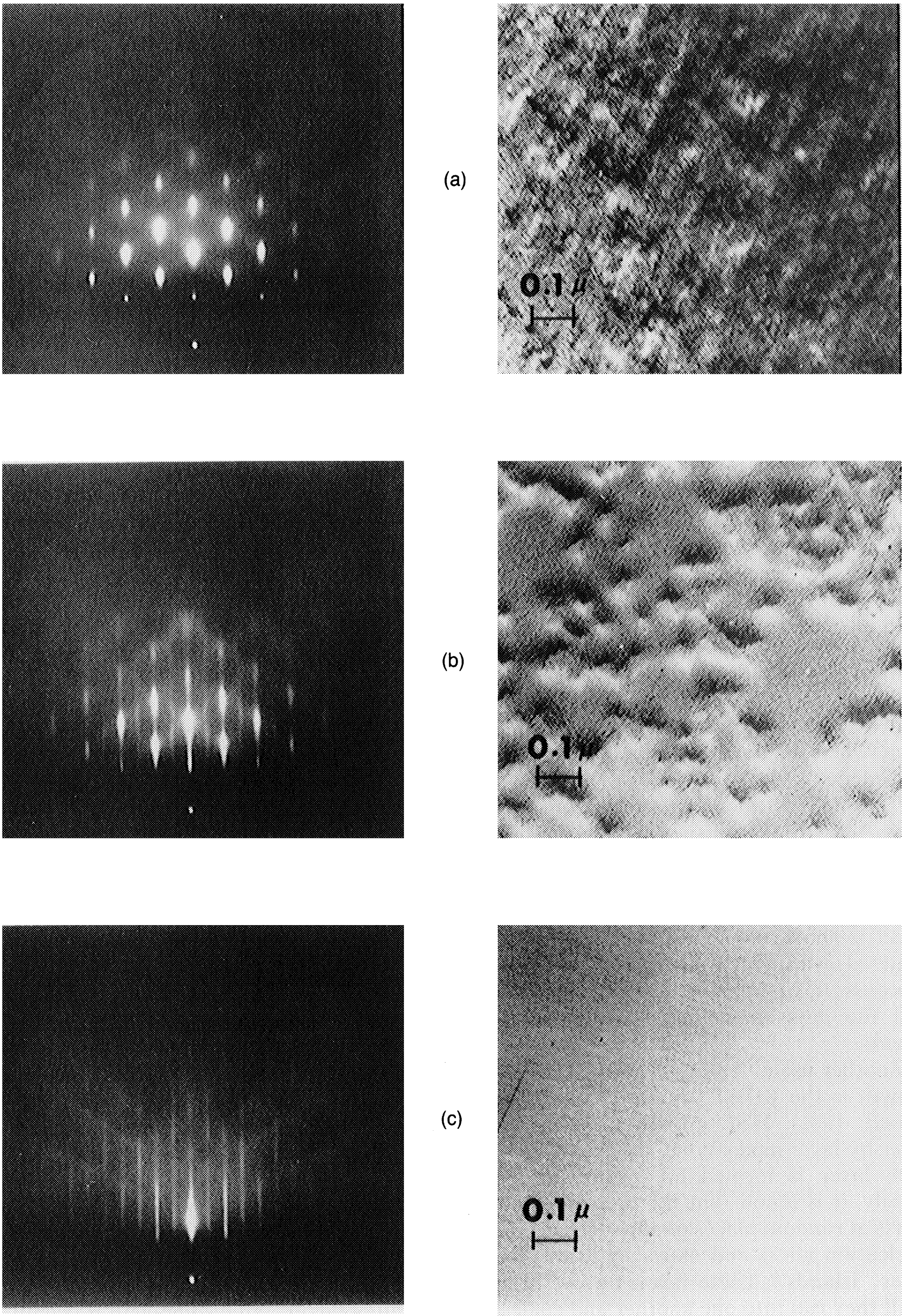

FIG. 10. RHEED pattern of a typical sequence of growth. (a) The pattern after the oxide is desorbed, (b) after deposition of 150 Å, and (c) after 1 μm of deposition. Corresponding electron micrographs are PtC replica of the surface. (After Cho 1983.)

quired to provide a uniform flux of gas molecules at the substrate. In contrast, the hydride injector must crack (thermally decompose) the hydride molecule.

While there are some differences in behavior between hydrides and solid group-V sources, aside from the controllability issue, there is little difference in use of these sources. On the other hand, metalorganic sources have some important differences from solid group-III sources. The distinction is that there is a chemical reaction that takes place on the substrate surface. This means that the species present just before condensation into the solid phase is quite different. There is evidence that the organometallic sources give much larger surface diffusion lengths, give different nucleation behavior, and, because of the different chemical kinetics, show a dependence of growth rate on substrate temperature. These gas-phase techniques have allowed MBE crystal growers to achieve high-quality InP/InGaAsP optical devices, and visible lasers using AlInP/GaInP.

2.4 Device Examples

As pointed out already, MBE has matured to the point of being a production tool in the semiconductor industry. In this section, we will discuss some examples of device structures that are in production using MBE material. While there are many new and interesting device structures made possible by MBE (the field of band-gap engineering) we will limit our discussion to only two examples.

The first example is the modulation-doped field-effect transistor (MODFET, sometimes called HEMT for high–electron-mobility transistor). This device has found a great deal of use in low-noise microwave receivers. Its low noise and high speed are attributable to the superior electron transport of the structure (Fischer and Morkoc, 1986). In a normal field-effect transistor, electrons travel through doped parts of the crystal. When a dopant is introduced into a crystal, it leaves a charged ion in the crystal lattice. This charged ion affects the electrons passing by, deflecting them through the electrostatic interaction. This effect of disturbing the electron motion results in inferior transport. The MODFET uses a heterostructure, that is, a junction between two dissimilar materials, to isolate the electrons from these ionized dopants.

Figure 11 shows an electron energy diagram of a MODFET. Doping is introduced into the material with higher band-gap energy (AlGaAs), and the free electrons transfer into the lower band-gap energy material (GaAs) to minimize their total energy. As can be seen, the electrons now travel in a part of the crystal where there is no doping and are essentially unaffected by the ionized impurities.

A metal electrode is placed on the surface to control the number of electrons available to carry current. In this way the applied voltage controls the output current of the device. From a materials point of view, it is important that the thickness and doping of the AlGaAs be highly reproducible. If there is variation in either one of these parameters, the devices will not switch on and off at the same voltage, making integration difficult, if not impossible. The other important aspect is the quality of the interface between the AlGaAs and GaAs. Any roughness or irregularity will degrade the electron transport properties. MBE has demonstrated its ability to produce highly uniform layers, in composition, thickness, and doping. The layer-by-layer nature of MBE growth also ensures a high-quality interface.

The second example of MBE as a production tool is in the AlGaAs lasers for compact-disk players. The key issues for this technology are the reproducibility in composition and thickness, and the purity of the material. The

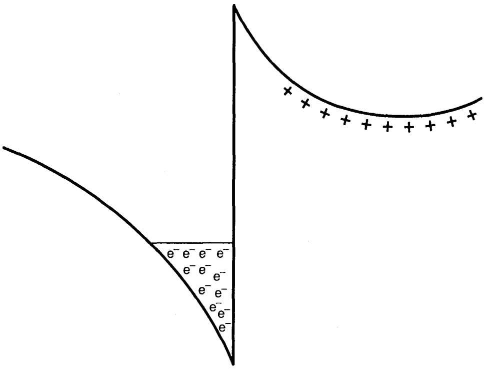

FIG. 11. Band-edge diagram of a typical MODFET. Electrons transfer from the AlGaAs into the GaAs and are separated from the ionized donors.

injection laser uses a *p-n* junction to inject electrons and holes into a semiconductor slab such that it will have optical gain. If certain types of defects are present, the electrons and holes will recombine without emitting light. If the thickness of the layer structure is not well controlled, the threshold current and emission wavelength will vary. Fortunately, MBE again provides excellent control over both these parameters.

The leading producer of compact-disk lasers, Rohm Co., Ltd., of Japan, uses a two-step MBE growth process for production of their lasers (Tanaka and Mushiage, 1991). This two-step process automatically produces a stripe geometry, which controls the optical (transverse and longitudinal) modes of the laser. Since MBE provides a high degree of reproducibility, the yields are very high in this process, which results in high-quality and low-cost lasers for many applications.

3. GROUP-IV MATERIALS

In this section, the growth of group-IV materials by MBE will be outlined. There are many important distinctions between the group-IV semiconductors and the III-V compounds. MBE as it has been described thus far relates to the III-V compound growth. Since the group-IV semiconductors are quite different, the MBE apparatus will also be quite different.

3.1 Apparatus and Growth Technique

The most technologically important group-IV semiconductor, is, of course, silicon. One of the primary differences is that the temperature required to produce an adequate flux of Si is far too high for an effusion cell of the conventional type. Therefore, electron-beam sources must be used. These sources use a beam of high-energy electrons at a relatively high current to melt a source material contained in a hearth. The walls of the hearth are cooled so that only the center portion of the source material is hot enough to evaporate, and thus serves as its own crucible. The use of electron-beam evaporation is not as stable as an effusion cell, and real-time flux monitoring must be used to control the power supplied to evaporate the source.

Another difference is that almost all evaporated doping elements surface segregate on the growing Si surface (Bean, 1981). Segration is an effect where the surface concentration is larger than that incorporated in the film. In other words, there is a buildup and tailing off of doping atoms on the surface when the doping is switched on and off, respectively. Obviously, this is an undesirable effect for devices that require sharp doping profiles. This effect also limits the maximum concentration of doping attainable, and in some cases can cause degradation of crystal quality. To get around this problem, molecular doping sources and low-energy implants are used.

The molecular species used for doping are boron oxide, B_2O_3, and boric acid, HBO_3. Effusion cells can produce a wide range of concentrations with these sources, and with good control. The use of B_2O_3, however, results in high concentrations of oxygen in the films, and only at high temperature (so high that the doping profile cannot be maintained) does the oxygen content become negligible. The use of HBO_3, on the other hand, seems to work much better, with temperatures of only about 650 °C required to eliminate oxygen. At this temperature, there is no evidence of boron movement or surface segregation.

The other method of doping is to use simultaneous low-energy ion implantation. In this process, the dopant species are ionized and accelerated to energies of 200–1000 eV. In this way, the dopant atom is buried within the film so that the surface segregation effect is reduced. Figure 12 shows the behavior of In ions co-implanted into Si at various energies as a function of temperature (Hasan, 1989). The drawback of this method is that there is the possibility of residual radiation damage. Many studies have been done, and it was found that when growth temperature is reduced from 650 to 500 °C, some features were observed that indicate incomplete annealing of the radiation damage. Other studies showed that a 2-keV implant grown at 550 °C produced little or no degradation in minority-carrier lifetime.

Growth of Si films by MBE proceeds in much the same way as for III-V MBE discussed earlier. The surface cleaning method is again a critical step. The Si oxide is a much more stable compound than the oxide on a III-V compound semiconductor. Therefore unless a special chemical process is used to

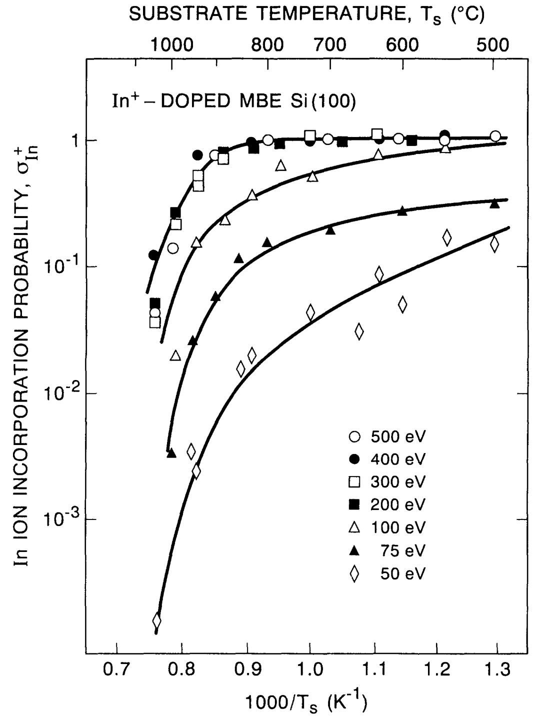

FIG. 12. Incorporation probability of In co-implanted into Si during MBE as a function of substrate temperature. The curves each correspond to a different ion energy. (After Hasan *et al.*, 1989; used by permission.)

produce a very thin oxide, it is impossible to remove the oxide thermally once loaded into the MBE chamber. This so-called "Shiraki clean" is typically used and results in an extremely thin and volatile oxide, which is sufficient to protect the substrate surface during the loading operation. With the wafer prepared in this way, it is possible to remove the oxide thermally.

Silicon technology is much more mature, and therefore particulate levels are of great concern, because of the scale of integration. Material deposited on the walls of a Si MBE chamber can be released as fine dust, and therefore many precautions are taken to reduce this effect. For example, in Si MBE, water cooling of the shrouding can be used in place of liquid nitrogen. This reduced thermal cycling greatly reduces the particle generation. Another approach is to steer particulates away from the substrate surface by means of electrostatic fields.

As was the case with III-V MBE, there are certain advantages to using gas-phase sources in Si MBE. While this concept was investigated at quite an early stage, the experiments used pumping systems that resulted in formation of SiC. More recently, improvements in pumping technology have eliminated these problems. The most obvious source for growth by Si gas-source MBE is silane, SiH_4. However, growth temperatures of greater than 700 °C are required to achieve reasonable growth rates. Furthermore, with silane as a source, the growth rate is quite sensitive to growth temperature. Most recently, the use of disilane, Si_2H_6, has increased low-temperature growth rates and is much less temperature sensitive.

3.2 Growth of SiGe

One of the greatest strengths of MBE is its ability to make highly abrupt compositional profiles. In this section, we will discuss the growth and material properties of a group-IV heterostructure system, SiGe/Si.

Germanium, while it is group-IV material as is Si, does not lattice match to Si. In other words, the atomic spacing of Ge is about 4% larger than that of Si. This mismatch makes growth of these structures difficult, because if thickness or temperatures are not within certain limits, the films will have many dislocations, which are defects that degrade the electronic performance.

One of the most sensitive tests of a material is its performance in a bipolar transistor. Optimization of SiGe growth parameters by MBE has allowed extremely high current gains of 750 to 2000 (depending on the structure), and operating frequencies of 20 GHz. This performance demonstrates the quality of SiGe films produced by MBE (Bean, 1990).

As discussed in the previous section, MODFETs, which use the properties of a heterojunction to improve electron transport, can also be done with SiGe/Si. This concept has been successfully applied in the MBE-grown SiGe/Si system. Hole mobilities as high as 18 000 cm^2/V s and electron mobilities as high as 180 000 cm^2/V s (Xie, *et al.* 1993) have been obtained in MBE-grown structures at 4.2 K.

4. II-VI COMPOUNDS

Another class of compound semiconductors are the II-VIs, which are made up from an element from column II and an element from column VI of the periodic table. The most important examples of II-VI compounds are HgTe, CdTe, ZnTe, ZnS, and ZnSe. The interest in HgTe and CdTe has been for long-wavelength infrared detectors, while in contrast, ZnTe, ZnS, and ZnSe, having wide band gaps, are interesting for visible-light emitting devices.

4.1 Narrow-Bandgap Materials

The growth of narrow-band-gap II-VI materials by MBE can be done using sources that are in principle the same as conventional ones. Usually for the growth of HgTe, a source of Hg and a source of Te are used. For the growth of CdTe, a CdTe compound source can be used, since Te is more volatile and therefore produces the excess flux required for growth of this compound. The growth of Hg-containing compounds is quite difficult in practice since at the typical growth temperatures, the sticking coefficient of Hg can be as low as 1 part in 1000. Therefore, a high flux of Hg is required to grow HgTe. This high-Hg flux depletes the source rapidly and therefore source designs which allow replenishment of Hg are required.

The first experiments with growth of HgTe/CdTe were done on ⟨111⟩-oriented CdTe substrates. It was later found that although there is a very large mismatch (14.5%) between CdTe and GaAs, HgCdTe grows well on GaAs substrates. This is a great advantage since the relatively well-developed technology of GaAs substrates can be used. It is interesting that in the growth of this material system on GaAs, the temperature used to desorb the oxide (preheating temperature) has a profound effect on the crystal orientation of the epitaxial layer. If the substrate is preheated to 480 °C or less, then CdTe grows in the ⟨100⟩ orientation. If preheated to 580 °C, then the CdTe grows in the ⟨111⟩ orientation. When the ⟨111⟩ orientation is grown, about four times less Hg flux is required compared to the ⟨100⟩ orientation. The concerns in growing this material system are interdiffusion of the HgTe/CdTe interface, and the exact electronic nature of that interface (Faurie, 1986).

4.2 Wide-Band-Gap Materials

As pointed out earlier, the wide-band-gap II-VI materials are of interest because of their application in visible-light emitters. While ZnTe emits in the green portion of the spectrum, ZnSe is in the blue. Only recently have semiconductor lasers been fabricated in this system (Haase *et al.*, 1991; Jeon *et al.*, 1991) that emit in the blue-green. While there are still many technological problems to be solved, MBE growth has been instrumental in the demonstration of these devices.

The single most problematic issue in the growth of these materials is doping. The II-VI compounds have a tendency to resist substitutional doping—the tellurides tend to be *p* type and the selenides *n* type. When these crystals are grown, introduction of impurities tends to generate compensating defects. The reason MBE has been successful in achieving this doping is that it is a nonequilibrium process.

The sources used for this system can be compound or elemental. In the growth of laser material mentioned above, elemental Zn, Se, and Cd were used, while a compound source, ZnS, was used for sulfur-containing materials. The elemental sources in II-VI growth tend to have very high vapor pressure at relatively low source temperature, making control of flux difficult. To make matters worse, very fine control of the group-II to group-VI ratio must be maintained to avoid generation of defects. Here as in earlier discussions, RHEED is an invaluable tool for determining the surface stoichiometry (whether a group-II-rich or a group-VI-rich surface). For example, a C(2 × 2) reconstruction pattern is observed for a Zn-stabilized surface, while the (2 × 1) reconstruction appears for a Se-stabilized surface. For a more complete discussion of II-VI compound growth the reader is referred to Gunshor *et al.* (1991).

The *p*-type doping of ZnSe has been historically a very difficult problem. In order to enhance the doping properties, a photoassisted technique was developed (Bicknell *et al.*, 1986) for In doping in CdTe. While dramatic improvements were made, *p*-type doping of ZnSe could not be adequately controlled. Most recently, use of nitrogen plasma sources have been successful to introduce *p*-type doping into ZnSe grown by MBE. This, coupled with the ability of MBE to produce high-quality heterostructures, has allowed the successful demonstration of blue-green semiconductor lasers in this system. Figure 13 shows a light output versus current input characteristic for this laser device.

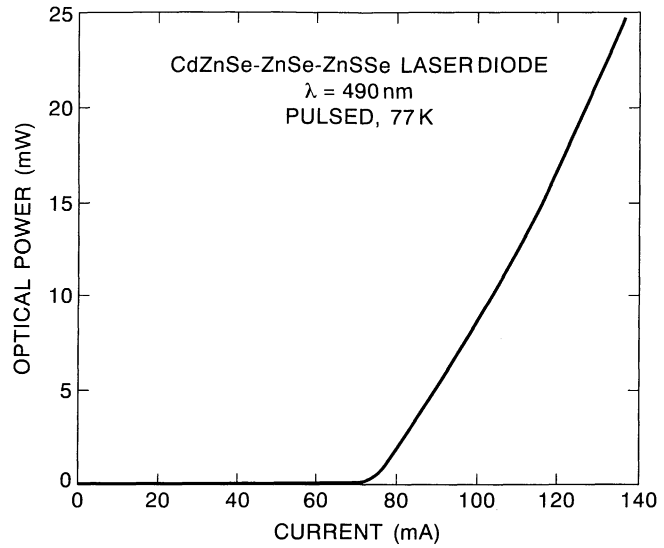

FIG. 13. Light output versus current input characteristic for a blue-green laser diode. (After Haase *et al.,* 1991; used by permission.)

5. OTHER MATERIAL SYSTEMS

In this section we will discuss some of the other material systems that have been grown using MBE or variations of MBE. Examples include epitaxial metal films, silicides, and high-temperature superconducting materials. The wide range of materials produced by MBE shows the diversity of the technique.

5.1 Epitaxial Metal Films

Metals as they occur in nature are polycrystalline in form; that is, there are small crystalline domains that are randomly oriented with respect to one another. If thin metal films are deposited under carefully controlled conditions, the metal films can be epitaxial. These epitaxial films, which uniformly have the same crystal structure and orientation, have useful properties. Because of the high degree of perfection for an epitaxial metal film interface, Schottky barriers could potentially be more reproducible and reliable, and there is the possibility of making metallic quantum wells and devices using buried metal films. Recent efforts have also included thin magnetic films for data storage applications.

Sources for metal deposition usually include both conventional effusion cells for metals with high vapor pressures, and electron-beam sources for those with low vapor pressures. While not used exclusively, semiconductor substrates are normally used, since many of the applications for these films are related to semiconductor devices.

The earliest work on depositing epitaxial metal films was Al on GaAs (Cho and Dernier, 1978) and Fe (Prinz and Krebs, 1981). More recent examples of metal films grown by MBE include NiAl on GaAs (Harbison *et al.,* 1988), single-crystal Dy (Yang *et al.,* 1988), Gd/Y superlattices (Kwo *et al.,* 1986), Dy/Y superlattices (Salamon *et al.,* 1986), and $Fe_3(Al,Si)$ on GaAs (Hong *et al.,* 1991). Figure 14 shows a sequence of RHEED patterns from deposition of $Fe_3(Al,Si)$ on GaAs. As can be seen through the quality of the patterns, highly ordered and atomically flat films were obtained.

5.2 Epitaxial Silicides

For electronic device fabrication, highly perfect interfaces between heterolayers is crucial. Higher quality and higher purity at these interfaces nearly always leads to superior electronic performance. Single-crystal epitaxial silicides on Si substrates are among the most perfect metal-semiconductor interfaces. The applications for such structures range from contacts to high-speed devices to the possibility for three-dimensional integration. Aside from the technological importance, the degree of perfection provides a model system to study the intrinsic properties of metal-semiconductor interfaces. Ex-

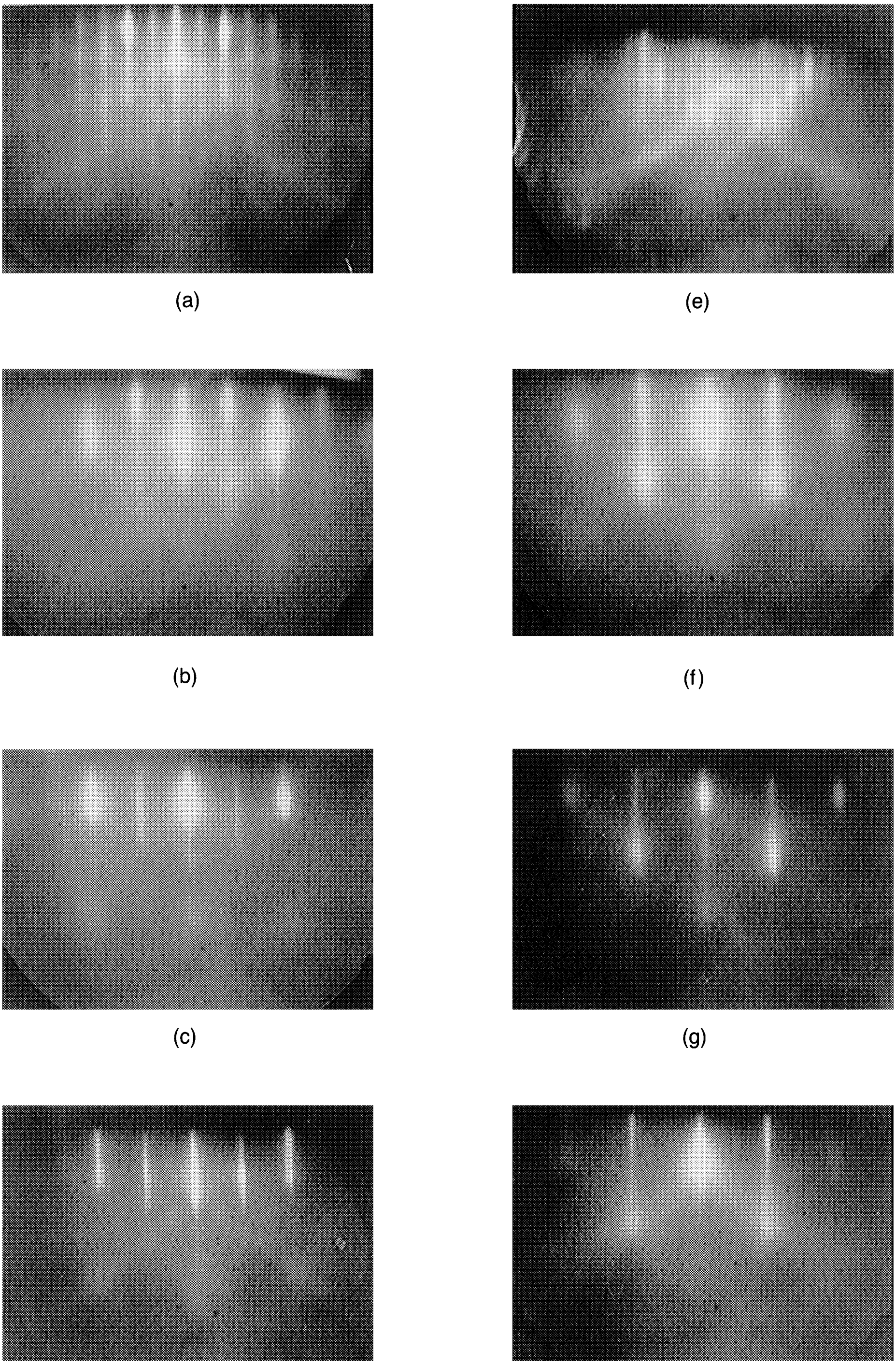

(a) (e)

(b) (f)

(c) (g)

(d) (h)

FIG. 14. RHEED patterns from the deposition of Fe(Al,Si). Patterns (a)–(d) are taken along the [011] direction, while (e)–(h) are along [0$\bar{1}$1]. (a), (e) The starting GaAs surface; (b), (f) after deposition of one monolayer of Fe$_3$Al; (c), (g) after growth of three monolayers; (d), (h) after growth of 12 monolayers. (After Hong *et al.,* 1991; used by permission.)

amples of epitaxial silicides would be $NiSi_2$ and $CoSi_2$.

Epitaxial silicides on Si have been grown by several methods, but the ability of MBE to produce clean interfaces and to maintain high purity in deposition allow it to produce the highest-quality films. Sources for this type of growth are typically electron-beam guns, because the elements have very low vapor pressures. In order to control stoichiometry, a coevaporation technique is used; that is, there are two electron-beam sources, one for the metal and the other for Si. The fluxes are adjusted so that a stoichiometric ratio is obtained.

The problems of silicide growth in MBE are in the deposition of the first monolayers. Whenever monolayers or submonolayers of silicides are deposited on Si, there is a tendency for the Si to deposit as crystalline Si on the substrate. As can be seen in Fig. 15, there is a distinct structural difference between crystalline Si and $NiSi_2$. To suppress this tendency, a method was developed called the template technique (Tung, 1992). This method is a two-step growth process in which the initial monolayers are deposited under radically different conditions than the remainder of the film. In silicides, this first nucleation process is done at very low temperatures, which promotes the formation of the silicide as opposed to the Si structure. Once the first few monolayers are deposited, there is no longer this tendency, because the Si atoms no longer see the Si substrate. At this point the temperature can be raised to the optimum for silicide growth.

Because of the difference in structure between Si and the silicide, noted above, it turns out that there are in fact two different ways that the epitaxial layer can orient itself with respect to the substrate. These two orientations are termed type A, which has the same orientation as the substrate, and type B, which is rotated from the axis of the Si. The details of the nucleation process determine which structure will predominate.

5.3 Superconducting Materials

Over the past several years, a new field of materials science and physics has emerged, that being high-temperature superconductivity. This has been brought about by the discovery of the so-called layered perovskite oxides. An example of such a material is $Y_1Ba_2Cu_3O_{7-x}$ (yttrium barium copper oxide). While the initial discovery of this material was not made by MBE, its structure is a layered one, and because of the layer-by-layer nature of MBE growth, much can be learned about the properties of these materials through the MBE growth with its associated *in situ* analysis. Furthermore, it allows one to tailor the structure on an atomic level.

Figure 16 shows the configuration of an MBE apparatus used to grow these compounds. Again a combination of electron-beam sources and effusion-cell sources is used. To provide oxygen, a separate injection tube is located in close proximity to the sample. In fact, in order to reduce the amount of oxygen required for growth, an activated oxygen source is used. The substrates used for these compounds vary and include MgO, $SrTiO_3$, and $LaAlO_3$. Because of the highly oxidizing environment, many changes in the materials of construction for the system must be made.

Use of the MBE technique has yielded very

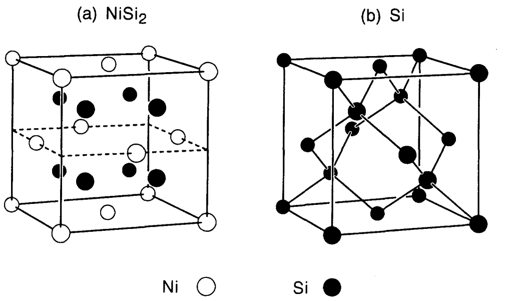

FIG. 15. Structure of NiSi and of Si. (After Tung, 1992; used by permission.)

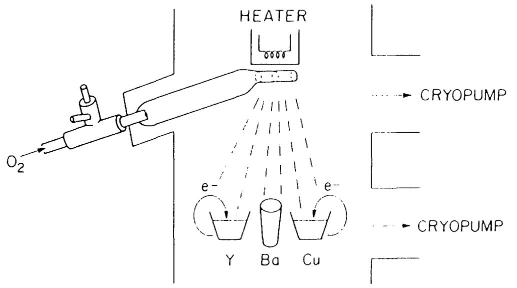

FIG. 16. Configuration of an MBE apparatus used for superconductor films. (After Kwo, 1991; used by permission).

high-quality superconducting films. Because of the controllability of the MBE process, some progress has been made in achieving tunnel junctions in this material. Josephson current has been observed (Kwo, 1991) between a high-temperature superconducting oxide and a conventional superconductor. This field has now been termed "molecular engineering of oxides."

GLOSSARY

Adsorption: Process by which an atom or molecule is bound to a surface, but is not incorporated into the underlying material.

Compound Semiconductor: A semiconductor that is composed of two or more elements, e.g., GaAs, AlGaAs.

Desorption: Process by which atoms or molecules bound to the surface are released into the vapor phase.

Dopant: An impurity added to a semiconductor in order to change its conductivity.

Epitaxy: The process of depositing one crystalline material on another such that the deposited material has a definite crystalline relationship to the substrate.

Heterostructure: A metallurgical junction between two dissimilar materials.

Mismatch: The difference, measured in percentage, between the atomic spacing in the substrate compared to that of the material being deposited.

Segregation: The tendency of some dopant species to stay at the surface when present during epitaxy.

Stoichiometry: The precise chemical ratio of one element to another in a compound semiconductor.

Substrate: A crystalline material that is usually flat and polished (in the form of a wafer), and that is used as a starting material for epitaxy.

Vapor Pressure: The tendency of a species to evaporate into the vapor form, measured in units of pressure at a given temperature.

Works Cited

Arthur, J. R. (1968), *J. Appl. Phys.* **39**, 4032–4034.

Bean, J. C. (1981), in: F. F. Y. Wang (Ed.), *Growth of Doped Silicon by Molecular Beam Epitaxy*, Amsterdam: North-Holland, Chap. 4, p. 177.

Bean, J. C. (1990), *J. Electron. Mater.* **19**, 1055–1059.

Bicknell, R. N., Giles, N. C., Schetzina, J. F. (1986), *Appl. Phys. Lett.* **49**, 1095–1097.

Cho, A. Y. (1971a), *Appl. Phys. Lett.* **19**, 467–468.

Cho, A. Y. (1971b), *J. Appl. Phys.* **42**, 2074–2081.

Cho, A. Y. (1983), *Thin Solid Films* **100**, 291–317.

Cho, A. Y., Arthur, J. R. (1975), in: J. McCaldin, G. Somorjai (Eds.), *Progress in Solid State Chemistry*, Vol. 10, Oxford: Pergamon Press, pp. 157–191.

Cho, A. Y., Cheng, K. Y. (1981), *Appl. Phys. Lett.* **38**, 360–362.

Cho, A. Y., Dernier, P. D. (1978), *J. Appl. Phys.* **49**, 3328–3332.

Faurie, J. P. (1986), *IEEE. J. Quantum Electron.* **QE-22**, 1656–1665.

Fischer, R., Morkoc, H. (1986), *IEEE Circuits Devices Mag.* **1**, 35–38.

Gunshor, R. L., Kolodziejski, L. A., Nurmikko, A. V., Otsuka, N. (1991), *Semiconductors and Semimetals*, Vol. 33, Chap. 6, p. 337.

Gunther, Von K. G. (1958), *Z. Naturforsch.* **13a**, 1081–1089.

Haase, M. A., Qui, J., DePuydt, J., Cheng, H. (1991), *Appl. Phys. Lett.* **59**, 1272–1274.

Harbison, J. P., Sands, T., Tabatabaie, N., Chan, W. K., Florez, L. T., Kerimidas, V. G. (1988), *Appl. Phys. Lett.* **53**, 1717–1719.

Hasan, M. A., Knall, J., Barnett, S. A., Sundgren, J. E., Markert, L. C., Rockett, A., J. E. Greene (1989), *J. Appl. Phys.* **65**, 172–179.

Hong, M., Chen, H. S., Kwo, J., Kortan, A. R., Mannaerts, J. P., Weir, B. E., Feldman, L. C. (1991), *J. Crystal Growth* **111**, 984–988.

Jeon, H., Ding, J., Patterson, W., Nurmikko, A. V., Xie, W., Grillo, D. C., Dobayashi, M., Gunshor, R. L. (1991), *Appl. Phys. Lett.* **59**, 3619–3621.

Kwo, J. (1991), *J. Crystal Growth* **111**, 965–972.

Kwo, J., Gyorgy, E. M., Disalvo, F. J., Horg, M., Yafet, Y., McWhan, D. B. (1986), *J. Magn. Magn. Mater.* **54–57**, 771–772.

Panish, M. B. (1980), *J. Electrochem. Soc.* **127**, 2729–2733.

Prinz, G. A., Krebs, J. J. (1981), *Appl. Phys. Lett.* **39**, 397–399.

Richards, J. L., Hart, P. B., Gallone, L. M. (1963), *J. Appl. Phys.* **34**, 3418–3420.

Salamon, M. B., Sinha, S., Rhyne, J. J., Cunningham, J. E., Erwin, R. W., Borchers, J., Flynn, C. P. (1986), *Phys. Rev. Lett.* **56**, 259–262.

Tanaka, H., Mushiage, M. (1991), *J. Crystal Growth* **111**, 1043–1046.

Tsang, W. T. (1985), *Appl. Phys. Lett.* **46**, 1086–1088.

Tung, R. (1992), in: S. Mahajan (Ed.), *Handbook on Semiconductors*, 2nd ed., Vol. 3, Amsterdam: Elsevier.

Van Hove, J. M., Lent, C. S., Pukite, P. R., Cohen, P. I. (1983), *J. Vac. Sci. Technol.* B **1**, 741–746.

Wood, C. E. C. (1981), *Surf. Sci.* **108**, L441–L443.

Xie, Y. H., Fitzgerald, E. A., Monroe, D. P., Silverman, P. J., Kortan, A. R. (1993) (unpublished).

Yang, K. Y., Homma, H., Schuller, I. K. (1988), *J. Appl. Phys.* **63**, 4066–4068.

Further Reading

For further reading there are two books devoted to various topics of molecular beam epitaxy. One is L. L. Chang and K. Ploog (Eds.), *Molecular Beam Epitaxy*, Dordrecht: Martinus Nijhoff (1985), and the other is E. H. C. Parker (Ed.), *Physics and Technology of Molecular Beam Epitaxy*, New York: Plenum (1985). Scientific articles relating to MBE work can be found in many journals, but primarily in *Journal of Applied Physics* and *Applied Physics Letters*. The proceedings of the MBE Workshop are published in the *Journal of Vacuum Science and Technology B* and the *Journal of Crystal Growth*. We, the authors, have attempted to cite review articles that cover the topics we have discussed in more depth.

MOLECULAR DYNAMICS

See SIMULATION BY MOLECULAR DYNAMICS

MOLECULAR INTERACTIONS WITH PHOTONS

See PHOTON INTERACTIONS WITH MOLECULES

MOLECULAR SPECTROSCOPY

D. A. Ramsay, *Herzberg Institute of Astrophysics, National Research Council of Canada, Ottawa, Ontario, Canada*

	Introduction	491
1.	**Rotational Spectra**	493
1.1	Diatomic Molecules	493
1.2	Polyatomic Molecules	494
1.2.1	Symmetric-Top Molecules	494
1.2.2	Asymmetric-Top Molecules	495
1.3	Determination of Dipole Moments	497
1.4	Hyperfine Structure	497
1.5	Applications to Astronomy	498
1.6	Rotational Raman Spectra	499
2.	**Vibrational Spectra**	500
2.1	Diatomic Molecules	500
2.1.1	Rotational Fine Structure	501
2.2	Polyatomic Molecules	502
2.2.1	Rotational Fine Structure	503
2.2.1.1	Linear Molecules	503
2.2.1.2	Symmetric-Top Molecules	504
2.2.1.3	Asymmetric-Top Molecules	507
2.2.1.4	Spherical-Top Molecules	507
2.3	Infrared Techniques	507
2.4	Applications of Infrared Spectroscopy	508
3.	**Electronic Spectra**	509
3.1	Classification of Electronic States	509
3.2	Intensities of Transitions	510
3.3	Franck–Condon Principle	511
3.4	Vibronic Interactions	512
3.5	Rotational Fine Structure	514
3.6	Magnetic Rotation Spectra	516
3.7	Dissociation and Predissociation	517
3.8	Linewidths and Lifetimes	517
3.9	Molecular Orbital Theory	517
3.10	Rydberg States and Ionization Potentials	518
3.11	Photoelectron Spectra and ZEKE Spectra	518
3.12	X-Ray Spectroscopy (ESCA)	519
3.13	Negative-Ion Spectroscopy	519
3.14	Laser Spectroscopy	520
3.14.1	Laser-Excited Fluorescence Spectra	520
3.14.2	Saturated Absorption Spectra	520
3.14.3	Two-Photon and Multiphoton Spectra	521
3.14.4	Nonlinear Laser Spectroscopy	522
3.15	Applications of Electronic Spectroscopy	522
4.	**Double-Resonance Spectroscopy**	522
	Works Cited	523
	Further Reading	523

INTRODUCTION

Molecular spectroscopy is primarily the study of the emission or absorption of electromagnetic radiation by molecules. The subject covers the whole range of electromagnetic radiation from radio waves to microwaves and millimeter waves, through the infrared to the visible and ultraviolet, and on to the X-ray region. The various regions overlap to some extent but may be characterized by the *wavelength* or *frequency* of the electromagnetic radiation involved (see Table 1).

Molecules exhibit various types of motion which give rise to series of *energy levels*, denoted by E_i. Transitions from one level to another may be accompanied by the emission or absorption of electromagnetic radiation whose frequency ν is given by

$$E' - E'' = h\nu \qquad (1)$$

where h is Planck's constant (6.626075×10^{-34} J s) and E' and E'' are the energies of the upper and lower levels, respectively. Frequencies of spectral lines are measured in units of hertz (cycles/s) or, more frequently, kilohertz (kHz, 10^3 s^{-1}), megahertz (MHz, 10^6 s^{-1}),

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Table 1. Regions of the electromagnetic spectrum.

Region	Wavelength (cm)	Frequency (GHz)	Wave number (cm^{-1})
Radio frequency	3×10^4–3×10^1	0.001–1	
Microwave	3×10^1–3×10^{-1}	1–100	
Millimeter	3×10^{-1}–3×10^{-2}	100–1000	3–30
Far infrared	3×10^{-2}–2.5×10^{-3}		30–400
Mid infrared	2.5×10^{-3}–2.5×10^{-4}		400–4000
Near infrared	2.5×10^{-4}–8×10^{-5}		4000–12 500
Visible	8×10^{-5}–4×10^{-5}		12 500–25 000
Near ultraviolet	4×10^{-5}–2×10^{-5}		25 000–50 000
Vacuum ultraviolet	2×10^{-5}–2×10^{-6}		50 000–500 000
X ray	2×10^{-6}–1×10^{-8}		

gigahertz (GHz, 10^9 s^{-1}), or terahertz (THz, 10^{12} s^{-1}). In the infrared, visible, and ultraviolet, it is customary to use wave number $\tilde{\nu}$, which represents the number of waves per unit distance and is the inverse of the wavelength. The unit of wave number most commonly used is cm^{-1}, which represents the number of waves per centimeter. The relation between frequency and wave-number units is

$$\nu = c\tilde{\nu}, \tag{2}$$

where c is the velocity of light.

The energy levels that molecules possess are governed by quantum theory and are characterized by a series of quantum numbers. It is not always possible to go from one energy level to another under the influence of electromagnetic radiation. The changes in quantum number that accompany an electromagnetic transition are governed by restrictions known as *selection rules*.

We will now consider the various types of energy level that molecules can occupy. Molecules in the gas phase rotate freely and occupy one of a number of discrete rotational energy levels. Transitions to neighboring levels give rise to emission or absorption of electromagnetic radiation, depending on whether the transition goes to a lower or higher energy level. The resulting radiation usually lies in the radio-frequency, microwave, or millimeter-wave region.

Molecules with N atoms have $3N - 6$ fundamental vibrational degrees of freedom ($3N - 5$ for a linear molecule; see Sec. 2.2). Transitions involving the vibrational energy levels usually lie in the infrared region of the spectrum.

Molecules are composed of nuclei and electrons. In their lowest energy state, the ground state, the electrons occupy certain orbits. Excitation of an electron to an orbit of higher energy produces an excited state of the molecule. Transitions between electronic states give rise to radiation that usually lies in the near-infrared, visible, or ultraviolet region of the spectrum.

Some molecular species, notably free radicals, have a resultant electron spin of $\frac{1}{2}h/2\pi$. Some nuclei, e.g., ^{1}H, ^{13}C, ^{19}F, ^{31}P, have a nuclear spin of $\frac{1}{2}h/2\pi$. In a magnetic field, the electron or nuclear spins can be aligned either parallel or antiparallel to the field. This produces two states of slightly different energies, the separation being proportional to the field strength. For typical laboratory fields, transitions between the two states arising from electron spin normally lie in the radio-frequency or microwave region, and the type of spectroscopy that studies these transitions is called electron spin resonance (ESR) or electron paramagnetic resonance (EPR) (see ELECTRON PARAMAGNETIC RESONANCE). The corresponding study of nuclear spins is known as nuclear magnetic resonance (NMR) and has widespread applications in chemistry and medicine (see MAGNETIC RESONANCE, NUCLEAR; MAGNETIC RESONANCE IMAGING).

The measurement of spectra involves instruments that measure the frequencies or wavelengths of transitions and the intensity of the radiation being emitted or absorbed. A variety of instruments is used, depending on the spectral region.

Spectroscopy is carried out on molecules in the gas, liquid, and solid phases. The subject is too vast for coverage in a single article; hence, the emphasis of this contribution will be on the spectroscopy of molecules in the gas phase. References will be given to related work in other areas.

1. ROTATIONAL SPECTRA

1.1 Diatomic Molecules

Let us consider a molecule composed of two atoms with masses m_1 and m_2 separated by a distance r (Fig. 1). Quantum mechanics shows that the rotational energy E_r of the molecule can have only certain discrete values given by the equation (Townes and Schawlow, 1955, Chap. 1)

$$E_r = (h^2/8\pi^2 I)J(J+1), \tag{3}$$

where J is the rotational quantum number, which can take values 0, 1, 2, ..., and I is the moment of inertia of the molecule about its center of mass:

$$I = \mu r^2, \tag{4}$$

where μ is called the reduced mass and is given by

$$1/\mu = 1/m_1 + 1/m_2. \tag{5}$$

For radiative transitions to occur, the molecule must posses an electric *dipole moment*, and the transitions are restricted by the selection rule

$$\Delta J = J' - J'' = +1. \tag{6}$$

The frequency of the transition ν is given by

$$h\nu_{J+1-J} = E_{J+1} - E_J. \tag{7}$$

Hence,

$$\begin{aligned}\nu_{J+1-J} &= (h/8\pi^2 I)\,[(J+1)(J+2) - J(J+1)]\\ &= 2B(J+1),\end{aligned} \tag{8}$$

where B is called the *rotational constant* and is given by

$$B = h/8\pi^2 I. \tag{9}$$

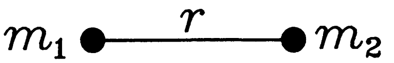

FIG. 1. Molecular parameters for a diatomic molecule.

The rotational spectrum of a diatomic molecule therefore consists of a series of equally spaced lines at frequencies $2B$, $4B$, $6B$, and so on. A diagram showing the rotational energy levels and transitions is given in Fig. 2. It is customary to plot the energy levels in units of E/hc, which are known as *term values*. The rotational term values $F(J)$ for a diatomic molecule are given by

$$F(J) = BJ(J+1) \tag{10}$$

and take the values 0, $2B$, $6B$, $12B$, $20B$ and so on. In these units,

$$B = h/8\pi^2 cI. \tag{10a}$$

Rotational spectra usually lie in the radiofrequency, microwave, or millimeter-wave regions, but, for higher values of J or for light molecules, the transitions may lie in the far-infrared region. As an example, the 1–0, 2–1, and 3–2 transitions of the carbon monoxide molecule CO lie at 115.271, 230.538, and 345.796 GHz, respectively. Transitions of this molecule have been observed up to 38–37 at 4340.138 GHz. It is found that the frequencies do not obey exactly the linear relation-

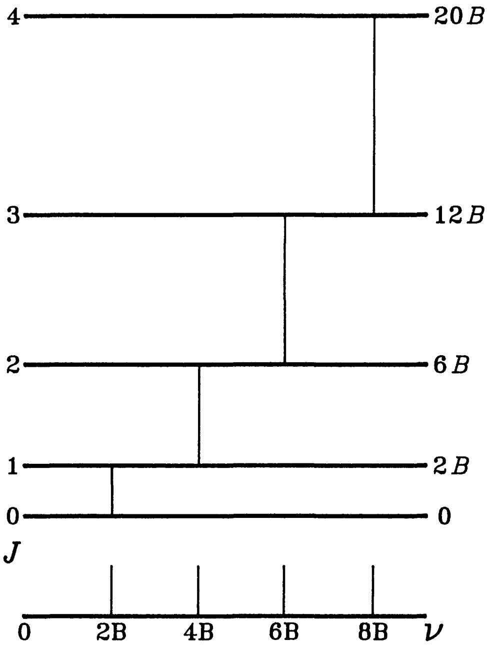

FIG. 2. Rotational term values and allowed transitions for a diatomic molecule.

ship given by Eq. (8). The reason is that the molecule stretches slightly with increasing rotation, thereby increasing its effective moment of inertia or decreasing its effective B value. Equation (10) is modified by the addition of a small correction term:

$$F(J) = BJ(J + 1) - DJ^2(J + 1)^2, \tag{11}$$

where D is known as the centrifugal distortion constant (see Sec. 2.1.1).

The frequencies in the rotational spectrum are then given by

$$\nu_{J+1—J} = 2B(J + 1) - 4D(J + 1)^3, \tag{12}$$

where D is very small compared with B. For example, for CO, $B = 57\,636$ MHz, but D is only 183.5 kHz.

Rotational constants for diatomic molecules can easily be determined with an accuracy of 1 part in 10^7. By the use of molecular beam techniques, this accuracy can be improved by another factor of 10. Provided that the atomic masses are known, accurate values for internuclear distances r can be obtained from Eqs. (4), (5), and (9). However, Planck's constant h is known only to slightly better than 1 part in 10^6.

If one or both of the atoms exist in isotopic modifications, then rotational spectroscopy can be used to obtain mass ratios. For example, by measuring the spectra of $^{127}I^{35}Cl$ and $^{127}I^{37}Cl$, the ratio of the masses $^{35}Cl/^{37}Cl$ can be determined with an accuracy of about 1 part in 10^7. Such a determination does not depend on the accuracy with which Planck's constant is known. However, the B values must be corrected for the effects of vibration (see Sec. 2.1.1).

1.2 Polyatomic Molecules

All molecules rotate freely in space about their center of mass, G. There is always an axis through G about which the moment of inertia of the molecule is a minimum. This axis is called the a axis and the least moment of inertia I_A. At right angles to this axis is another axis, c, about which the moment of inertia I_C is a maximum. A third axis, b, is chosen at right angles to the a and c axes, and the moment of inertia about the b axis is I_B. By definition,

$$I_A \leqslant I_B \leqslant I_C. \tag{13}$$

Rotational constants A, B, and C are defined so that

$$A = h/8\pi^2 cI_A, \quad B = h/8\pi^2 cI_B,$$
$$C = h/8\pi^2 cI_C. \tag{14}$$

Again, by definition,

$$A \geqslant B \geqslant C. \tag{15}$$

There are now four types of molecule to consider:

1. Spherical-top molecules for which $A = B = C$,
2. Symmetric-top molecules,
   - **a.** Prolate tops with $A \neq B = C$
   - **b.** Oblate tops with $A = B \neq C$,
3. Linear molecules with $B = C$ and $I_A = 0$,
4. Asymmetric-top molecules with $A \neq B \neq C$.

The rotational term values for spherical-top and linear molecules are the same as for diatomic molecules [Eq. (11)]. However, spherical-top molecules—e.g., CH_4—have no dipole moment and, hence, no rotational spectrum (except a very weak spectrum produced by certain vibrations or by centrifugal distortion). Linear molecules with dipole moments produce rotational spectra that obey the same formula as for diatomic molecules [see Eq. (12)].

We will now consider the other two types of rotor.

1.2.1 Symmetric-Top Molecules The rotational term values for symmetric-top molecules depend on two quantum numbers, J and K. The total angular momentum of the molecule is quantized and equals $\sqrt{J(J + 1)}h/2\pi$. Similarly, the component of the total angular momentum along the direction of the symmetric-top axis is quantized and equals $Kh/2\pi$, where K is an integer ($K \leqslant J$). The rotational term values for a prolate symmetric top are given by

$$F(J, K) = (A - B)K^2 + BJ(J + 1), \tag{16}$$

and, for an oblate symmetric top,

$$F(J, K) = (C - B)K^2 + BJ(J + 1). \tag{17}$$

Energy-level diagrams for the two types of symmetric top are given in Fig. 3. Examples of prolate tops are the methyl halides; benzene provides an example of an oblate top. It should be noted that since the component of J in the direction of the symmetric-top axis can be either positive or negative, all states with $K > 0$ are *doubly degenerate.*

Since the dipole moment of the molecule always lies along the symmetric-top axis, the selection rules are

$$\Delta K = 0, \qquad \Delta J = +1. \tag{18}$$

In the absence of centrifugal distortion, the rotational spectrum is similar to that of a diatomic molecule. However, when centrifugal terms are added, the rotational term values are given by

$$\begin{aligned} F(J, K) = (A - B)K^2 + BJ(J + 1) - D_K K^4 \\ - D_{JK}K^2J(J + 1) - D_J J^2(J + 1)^2 \end{aligned} \tag{19}$$

for a prolate top. For an oblate top, the rotational constant A should be replaced by C. The frequency of a rotational transition $(J + 1, K)$—(J, K) is

$$\nu = 2(J + 1)(B - D_{JK}K^2) - 4D_J(J + 1)^3. \tag{20}$$

The series for given J values with different values of K will be separated if the value of D_{JK} is sufficiently large.

The analysis of rotational spectra of symmetric-top molecules yields only the moments of inertia about the center of mass in a direction perpendicular to the symmetric-top axis. No information is obtained regarding the moment of inertia about the symmetric-top axis. Partial structural information, however, can be obtained if more than one isotopic species is studied.

1.2.2 Asymmetric-Top Molecules There is no simple formula for the rotational term values of an asymmetric-top molecule. For each value of J, the total angular momentum quantum number, there are $2J + 1$ rotational energy levels. The term values are obtained by diagonalization of matrices. Each level is designated by J_{K_a,K_c}, where K_a and K_c are the K quantum numbers of the levels with which the asymmetric-top level correlates in the

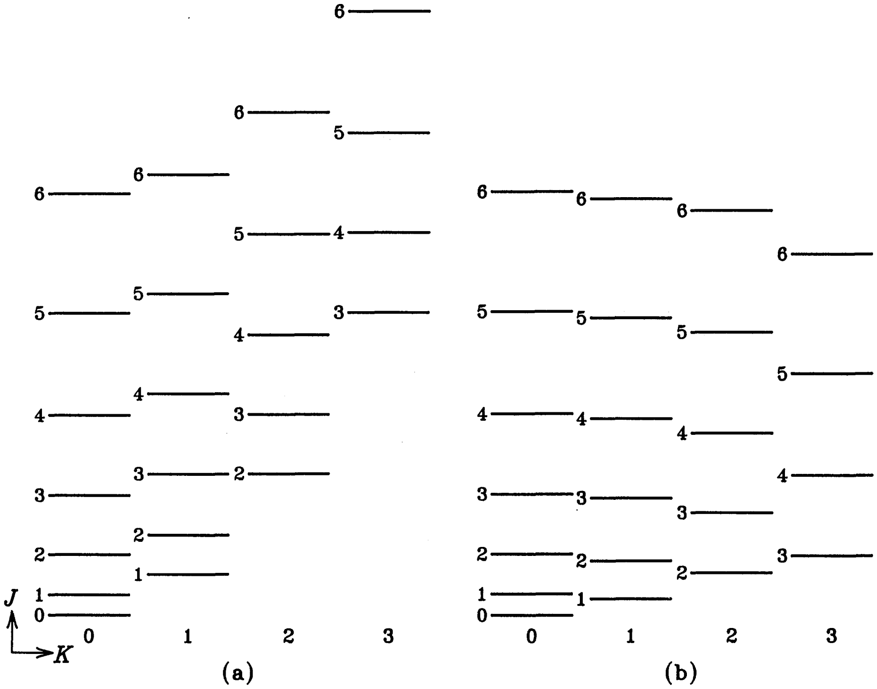

FIG. 3. Rotational term values for symmetric-top molecules: (a) prolate top with $A = 3B$, (b) oblate top with $C = \frac{1}{2}B$.

prolate and oblate symmetric-top limits, respectively. A correlation diagram is shown in Fig. 4 and is based on the principle that no levels with the same value of J can cross (the *non-crossing rule*).

Selection rules depend on the direction of the dipole moment of the molecule with respect to the inertial axes. For all transitions,

$$\Delta J = J' - J'' = 0, \pm 1. \tag{21}$$

If the dipole moment lies in the a direction,

$$\Delta K_a = 0, \pm 2, \ldots; \quad \Delta K_c = \pm 1, \pm 3, \ldots. \tag{22}$$

If the dipole moment lies in the b direction,

$$\Delta K_a = \pm 1, \pm 3, \ldots; \quad \Delta K_c = \pm 1, \pm 3, \ldots. \tag{23}$$

If the dipole moment lies in the c direction,

$$\Delta K_a = \pm 1, \pm 3, \ldots; \quad \Delta K_c = 0, \pm 2, \ldots. \tag{24}$$

If the dipole moment does not lie in the direction of an inertial axis, then more than one type of transition will be allowed, depending on the components of the dipole moment in the directions of the inertial axes.

An example of an asymmetric top is provided by the water molecule H_2O (Fig. 5). The a and b inertial axes lie in the plane of the molecule, while the c axis lies perpendicular to the plane. The dipole moment lies in the b direction; hence, the selection rules obey Eqs. (21) and (23). One of the transitions that produces absorption in the Earth's atmosphere and interfered with early radar is the 6_{16}—5_{23} transition at 22.235 GHz.

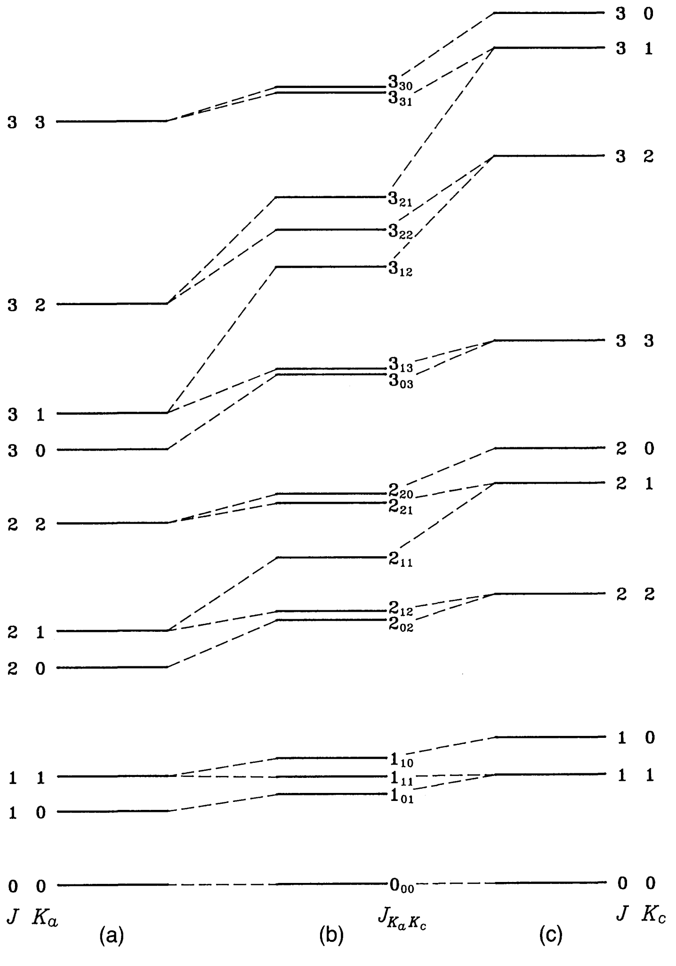

FIG. 4. Correlation diagram connecting the rotational energy levels of (a) a prolate symmetric top with $A = 2$, $B = C = 1$, (b) an asymmetric top with $A = 2$, $B = 1.5$, $C = 1$, (c) an oblate top with $A = B = 2$, $C = 1$.

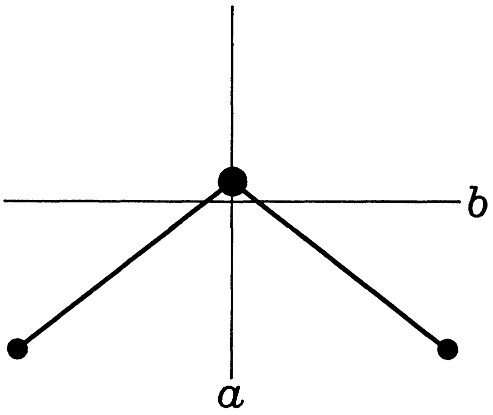

FIG. 5. Principal axes of inertia for the H_2O molecule. The c axis is perpendicular to the plane.

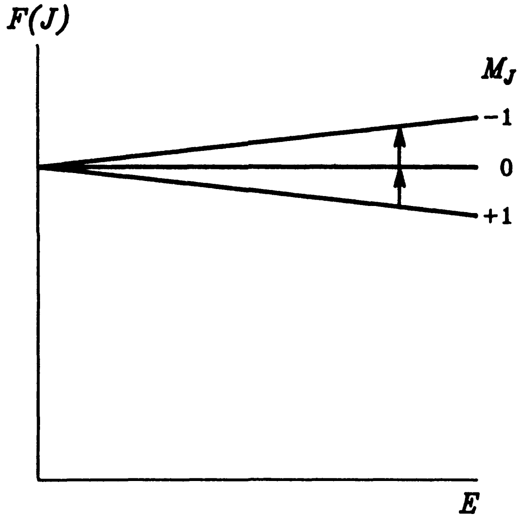

FIG. 6. Stark splittings for a $J = 1$, $K = 1$ level of a symmetric-top molecule. In perpendicular polarization, the splittings between the $M_J = 0$ level and the $M_J = \pm 1$ levels give a measure of the dipole moment of the molecule. (The M_J quantum numbers are given for $K = +1$; for the $K = -1$ component, the M_J quantum numbers should be reversed.)

1.3 Determination of Dipole Moments

The energy levels of a symmetric-top molecule, except those with $K = 0$, split into $2J + 1$ components in the presence of an electric field, a phenomenon known as the Stark effect. The components are labeled by a quantum number M_J, which represents the component of J in the direction of the electric field and takes values $J, J - 1, J - 2, \ldots, 0 \ldots, -(J - 1), -J$. The splittings ΔW are given by the formula

$$\Delta W = -\mu E M_J K/J(J + 1), \qquad (25)$$

where μ is the dipole moment of the molecule and E is the electric-field strength.

An energy-level diagram showing the effect of an electric field on a $J = 1$, $K = 1$ level is given in Fig. 6. The level splits into three components with $M_J = -1$, 0, and +1.

Selection rules for transitions depend on the direction of the electric vector of the electromagnetic radiation relative to the direction of the electric field. If the two are parallel, then

$$\Delta M_J = 0. \qquad (26)$$

If the two are perpendicular, then

$$\Delta M_J = \pm 1. \qquad (27)$$

In Fig. 6, the splittings between the level with $M_J = 0$ and the levels with $M_J = \pm 1$ give a measure of the dipole moment of the molecule.

For diatomic and linear polyatomic molecules, the splittings are much smaller and depend on the square of the electric field strength. Furthermore, levels with $+M_J$ and $-M_J$ remain degenerate; i.e., they are not split. Consequently, only $J + 1$ components are observed in an electric field. Asymmetric-top molecules behave in a similar manner except when the K-type doubling is very small; i.e., levels with the same K_a (or K_c) but different K_c (or K_a) show a very small splitting. In the latter case, the levels behave like those of a symmetric top and show a first-order Stark effect.

Dipole moments of many molecules have been measured, some with accuracies of the order of 1 part in 10^6.

1.4 Hyperfine Structure

Spectra of molecules containing atoms with nuclear spin show very small splittings called hyperfine structure. If all the nuclei in a molecule have zero nuclear spin except one with a nuclear spin I, then, in a rotational state J, the level splits into components with a resultant quantum number F which takes values $J + I, J + I - 1, \ldots, J - I$. Selection rules for transitions between hyperfine components are

$$\Delta J = 0, \pm 1; \quad \Delta F = 0, \pm 1; \quad \Delta I = 0. \tag{28}$$

The largest hyperfine splittings in molecular spectra occur if one (or more) of the atoms has an electric quadrupole moment. Of necessity, this is possible only for atoms with $I = 1$ or more.

For a linear molecule, the quadrupole energy W_Q is

$$W_Q = -eq_m Q f(I, J, F), \tag{29}$$

where e is the electronic charge, q_m is the electric-field gradient at the nucleus in the direction of the linear axis, Q is the quadrupole moment of the nucleus, and $f(I, J, F)$ is Casimir's function

$$f(I, J, F) = \frac{\frac{3}{4}C(C+1) - I(I+1)J(J+1)}{2I(2I-1)(2J-1)(2J+3)}, \tag{30}$$

where

$$C = F(F+1) - I(I+1) - J(J+1). \tag{31}$$

Since q_m is rarely known, the quantity determined experimentally is the product eqQ, which is known as the quadrupole coupling constant. Values are normally of the order of 100 MHz and may be positive or negative, but, for the ^{127}I nucleus in ICN, the large value of $eqQ = -2420$ MHz is found. An example of quadrupole hyperfine structure in the $J = 1$—0 transition of $^{16}O^{12}C^{33}S$ is shown in Fig. 7. The ^{33}S nucleus has a spin of $\frac{3}{2}$, and $eqQ = -29.2$ MHz.

The quadrupole energy for a nucleus on the axis of a symmetric-top molecule is similar to Eq. (29), but, for asymmetric-top molecules, the expressions are more complicated.

Hyperfine structure may also be observed resulting from the magnetic dipole moments of nuclei. The splittings are very small in most molecules but may become appreciable if a molecule has a resultant electron spin as in a free radical.

An important application of hyperfine structure lies in nuclear quadrupole resonance (NQR) spectroscopy. Transitions between hyperfine levels are measured directly using a radio-frequency spectrometer that usually operates over the range 2–1000 MHz. The applications are generally confined to solids (Semin *et al.*, 1975).

1.5 Applications to Astronomy

In 1968, only four molecules were known to exist in the interstellar medium. Three of these, CH, CH$^+$, and CN, had been found by optical spectroscopy between 1937 and 1941. The fourth, OH, was found by radioastron-

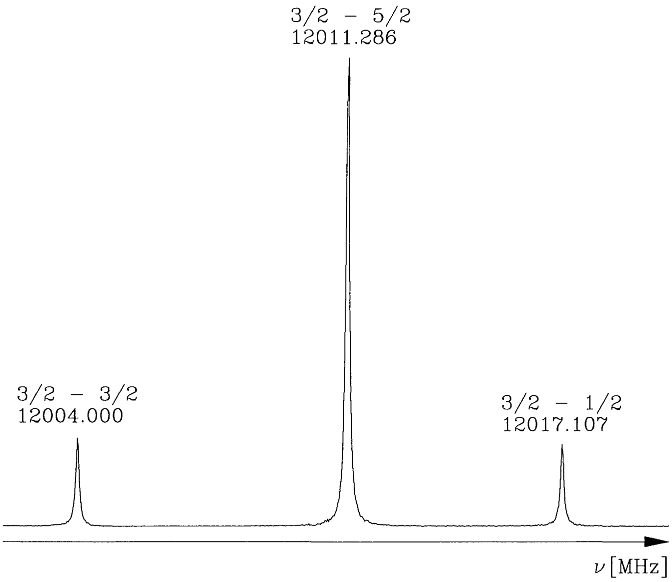

FIG. 7. Quadrupole hyperfine structure observed in the $J = 1$–0 transition of OC^{33}S, by courtesy of H. Dreizler, Universität Kiel, Germany.

omy in 1963. In 1968, interstellar NH_3 was detected by means of its microwave spectrum; this was followed in quick succession by the detection of interstellar H_2O (1968), H_2CO (1969), CO, HCN, HCO^+, HC_3N, CH_3OH, HCOOH (1970), NH_2CHO, SiO, OCS, CS, CH_3CN, HNCO, CH_3C_2H, $HCOCH_3$, and H_2CS (1971) (Snyder, 1972). Today, approximately 100 molecules have been identified in the interstellar medium by means of radioastronomy, and some of these have also been seen in the atmospheres of stars. Most of the spectra have been observed in emission, but some have been seen in absorption against the 2.7-K background radiation.

Some species, e.g., HCO^+, were found in the interstellar medium before they were detected in the laboratory. Another group of molecules that has been observed in the interstellar medium is the straight-chain cyanopolyacetylenes $H{-}(C{\equiv}C{-})_nC{\equiv}N$, where $n = 1$–5. The lower members ($n = 1, 2, 3$) had been synthesized and measured in the laboratory, but the higher members with n equal to 4 and 5 were first seen in the interstellar medium.

The conditions of low temperature and exceedingly low pressures in the interstellar medium are favorable for the observation of very sharp lines. Frequencies can be measured very accurately and usually differ slightly from laboratory frequencies on account of motions within the interstellar medium, i.e., Doppler shifts. Furthermore, microwaves, on account of their long wavelengths, can penetrate dark molecular clouds, which are opaque to visible light. These clouds are of particular interest since they are regions where new stars are formed. Consequently, microwave spectroscopy of the interstellar medium provides an important technique for probing the dynamics and structure of regions of star formation in the universe.

1.6 Rotational Raman Spectra

Molecules with a zero dipole moment, e.g., H_2, N_2, C_6H_6, have no microwave absorption. Nevertheless, a rotational Raman spectrum can be observed. This spectrum is produced by irradiating the molecules with a strong monochromatic source and recording the lines that appear on both sides of the exciting line. In the early work, a mercury lamp was usually used, but nowadays excitation is achieved by the use of lasers. The selection rules for a diatomic or linear polyatomic molecule are

$$\Delta J = 0, \pm 2 \tag{32}$$

and, for a symmetric top,

$$\Delta J = 0, \pm 1, \pm 2, \tag{33}$$

except that Eq. (32) applies if $K = 0$.

As an example, the rotational Raman spectrum of N_2 is shown in Fig. 8. A series of equidistant lines (called S branch since $\Delta J = +2$) is observed on either side of the exciting line, the series going to lower frequencies being called the Stokes branch and the series going to higher frequencies the anti-Stokes branch. The frequency shifts, neglecting centrifugal distortion, are given by

$$|\Delta\nu| = 4B(J + \tfrac{3}{2}); \tag{33a}$$

hence, the spacing between neighboring lines in the branches is $4B$, and the separation between the exciting line and the first line in each branch is equal to $6B$. The alternation in the intensities of the lines is caused by the spin statistics of the N nuclei (see Sec. 2.2.1).

The rotational Raman spectrum of nitrogen is of particular interest, since, when it was first reported, the alternation of intensities was opposite to that observed for hydrogen and led to the deduction that nuclei could not

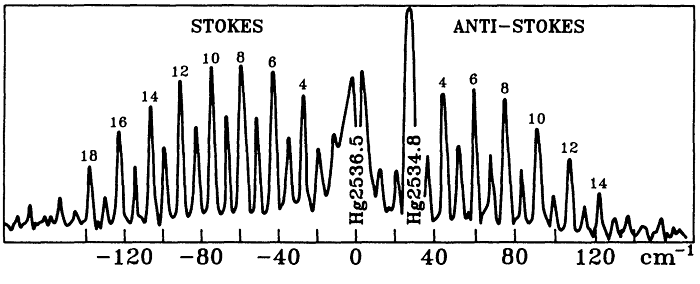

FIG. 8. Rotational Raman spectrum of N_2 (Stoicheff, 1950). The spectrum was excited by the 2536.5 = Å Hg line and shows the Stokes and anti-Stokes S branches with a 2:1 intensity alternation. There is some pressure broadening of the lines, resulting from the high pressure used (10 atm.).

be composed of protons and electrons (Heitler and Herzberg, 1929). This result preceded the discovery of the neutron.

2. VIBRATIONAL SPECTRA

2.1 Diatomic Molecules

According to quantum mechanics, the energy levels of a simple harmonic oscillator are given by (Pauling and Wilson, 1935)

$$E_v = h\nu(v + \tfrac{1}{2}), \tag{34}$$

where v is the vibrational quantum number, which takes values 0, 1, 2, ... In term values,

$$G_v = E_v/hc = \omega(v + \tfrac{1}{2}), \tag{35}$$

where $G(v)$ is the vibrational term value and ω is the vibrational frequency expressed in cm^{-1} units. The lowest energy level that the molecule can occupy has an energy $\frac{1}{2}hc\omega$, which is known as the zero-point energy and exists even at the absolute zero of temperature.

The vibrational frequency is related to the *force constant* k for the vibration by the relation

$$k = 4\pi^2\mu c^2\omega^2, \tag{36}$$

where $k(r - r_e)$ is the restoring force when the molecule is displaced from its equilibrium position and μ is the reduced mass [Eq. (5)].

In reality, diatomic molecules behave like anharmonic oscillators and dissociate into atoms when the vibrational energy is sufficiently large. A potential energy function and vibrational energy levels for an *anharmonic* oscillator are shown in Fig. 9. Equation (35) is modified by the addition of terms involving a series of anharmonic constants $\omega_e x_e$, $\omega_e y_e$, $\cdots$:

$$\begin{aligned} G(v) = {} & \omega_e(v + \tfrac{1}{2}) - \omega_e x_e(v + \tfrac{1}{2})^2 \\ & + \omega_e y_e(v + \tfrac{1}{2})^3 + \cdots . \end{aligned} \tag{37}$$

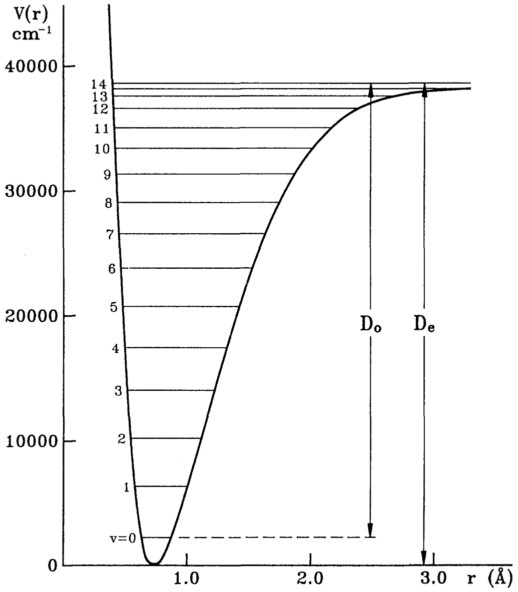

FIG. 9. Potential curve and energy levels of an anharmonic oscillator.

This form of equation does not satisfactorily explain the behavior when the molecule approaches the dissociation limit. Hence, another potential function is frequently used, the *Morse potential*, given by

$$V(r) = D_e\{1 - \exp[-\beta(r - r_e)]\}^2, \tag{38}$$

where D_e is the dissociation energy—that is, the energy difference between the minimum of the potential curve and the dissociation limit. If D_e is in cm^{-1}, then

$$\beta = \sqrt{\frac{2\pi^2 c\mu}{D_e h}}\,\omega_e. \tag{39}$$

This function has the merit that $V = 0$ when $r = r_e$ and $V = D_e$ when r becomes infinite. Note, however, that the quantity that is measured experimentally is D_0, which is the energy difference between the zeroth vibrational level and the dissociation limit.

The selection rule for the vibrational quantum number of a simple harmonic oscillator is

$$\Delta v = \pm 1. \tag{40}$$

However, molecules are not exactly harmonic oscillators; hence, transitions with $\Delta v > 1$ can also appear, though weakly. Transitions with $\Delta v = 1$ lie in the infrared region of the spectrum. Overtone spectra with $\Delta v =$

2, 3, 4, . . . are much weaker, the decrease in intensity being roughly an order of magnitude or more for each higher value of v. Nevertheless, overtone spectra may be observed in absorption using progressively longer path lengths or intracavity techniques using lasers. For example, the 6—0 band of HD has been observed in absorption in the visible region using a path length of 3 km and a pressure of 1 atm.

Vibrational transitions may also be observed in emission. The intensity in emission is proportional to the fourth power of the frequency, as compared to the first power in absorption. Consequently, $\Delta v = 2$ transitions may be observed in emission, with intensities comparable to those of $\Delta v = 1$ transitions.

The intensity of an infrared transition depends on the derivative of the dipole moment with respect to the internuclear distance. Hence, homonuclear molecules such as H_2, O_2, and N_2 do not have an infrared spectrum, although a vibrational Raman spectrum can be observed (Long, 1977). In the case of H_2, an extremely weak electric quadrupole spectrum has been observed and is common in astrophysical sources.

2.1.1 Rotational Fine Structure When a molecule vibrates, the value of r changes during the vibration. For each vibrational level, we can define a rotational constant B_v such that

$$B_v = \frac{h}{8\pi^2 c\mu}\left[\frac{1}{r^2}\right]_{\text{av}}. \tag{41}$$

The quantity B_v can be expressed as a power series in $v + \frac{1}{2}$:

$$B_v = B_e - \alpha_e(v + \tfrac{1}{2}) + \cdots \tag{42}$$

and

$$B_e = h/8\pi^2 c\mu r_e^2. \tag{43}$$

α_e is usually small compared with B_e, e.g., ~1%. The quantity B_e is the equilibrium rotational constant, i.e., the rotational constant at the minimum of the potential curve. Since r_e is independent of the atomic masses, B_e is the quantity that must be used in the determination of the ratio of isotopic masses (see Sec. 1.1).

During a vibrational transition, the rotational quantum numbers may also change. The selection rule $\Delta J = J' - J'' = \pm 1$ gives rise to an R branch ($\Delta J = +1$) and a P branch ($\Delta J = -1$). An energy-level diagram is given in Fig. 10. Note that all the transitions are numbered according to the lower-state J value.

For the R branch,

$$\begin{aligned} R(J) &= \nu_0 + F'(J+1) - F''(J) \\ &= \nu_0 + B'(J+1)(J+2) - B''J(J+1) \\ &= \nu_0 + 2B' + (3B' - B'')J \\ &\quad + (B' - B'')J^2, \end{aligned} \tag{44}$$

where ν_0 is the frequency of the band origin and $J = 0, 1, 2, \ldots.$

For the P branch,

$$\begin{aligned} P(J) &= \nu_0 + F'(J-1) - F''(J) \\ &= \nu_0 + B'(J-1)J - B''J(J+1) \\ &= \nu_0 - (B' + B'')J + (B' - B'')J^2, \end{aligned} \tag{45}$$

where $J = 1, 2, 3, \ldots.$

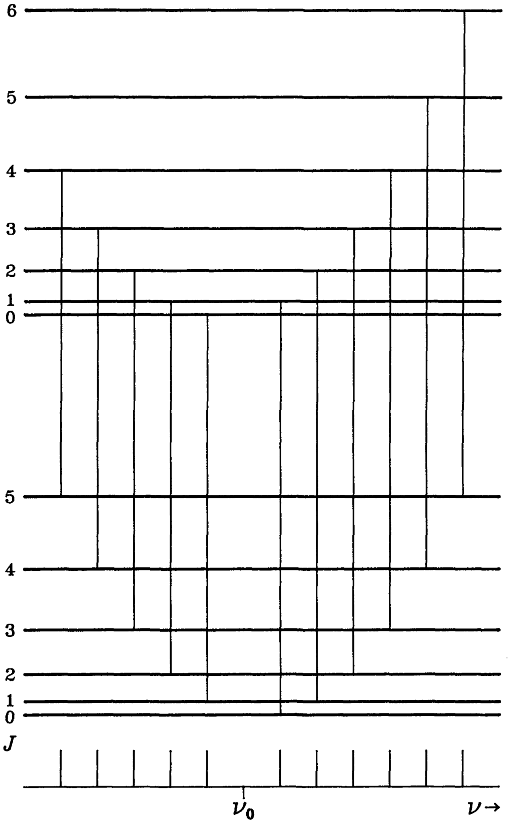

FIG. 10. Energy-level diagram and transitions for the P and R branches of a simple diatomic molecule.

Both branches can be fitted by a single equation

$$\nu = \nu_0 + (B' + B'')m + (B' - B'')m^2, \qquad (46)$$

where $m = J + 1$ for the R branch and $m = -J$ for the P branch. This equation is known as the *Fortrat parabola*. A more complete expression, including the effects of centrifugal distortion, is

$$\nu = \nu_0 + (B' + B'')m + (B' - B'' - D' + D'')m^2 - 2(D' + D'')m^3 - (D' - D'')m^4, \qquad (47)$$

where

$$D \sim 4B^3/\omega^2. \qquad (48)$$

An example is provided by the fundamental vibration-rotation band of carbon monoxide, which is shown in Fig. 11.

2.2 Polyatomic Molecules

The positions of the nuclei of a molecule with N atoms can be described by $3N$ coordinates. Three of these coordinates are required to specify the position of the center of mass and three more angular coordinates to specify the orientation of the molecule in space (only two angular coordinates are required for a linear molecule since there is no rotation of the nuclei about the internuclear axis). Consequently, $3N - 6$ coordinates are needed to specify the internal modes of vibration of the molecule ($3N - 5$ for a linear molecule).

In the simplest approximation, the vibrational motion may be considered to be the sum of a series of independent simple harmonic oscillators, and the energy levels are given by

$$G(v_1, v_2, \ldots) = \sum_{i=1}^{3N-6} \omega_i(v_i + \tfrac{1}{2}d_i), \qquad (49)$$

where G is the vibrational term value, the ω_i are the zero-order frequencies of the various modes of vibration, and d_i is the degeneracy of the ith vibration (normally 1, but may be 2 or higher).

A more complete expression that takes account of the interactions between the vibrations and of their anharmonicities is

$$G(v_1, v_2, \ldots) = \sum_i \omega_i(v_i + \tfrac{1}{2}d_i) + \sum_i \sum_{k \geq i} x_{ik}(v_i + \tfrac{1}{2}d_i)(v_k + \tfrac{1}{2}d_k) + \sum_i \sum_{k \geq i} g_{ik}l_il_k + \cdots. \qquad (50)$$

The x_{ik}'s are the anharmonic constants, and

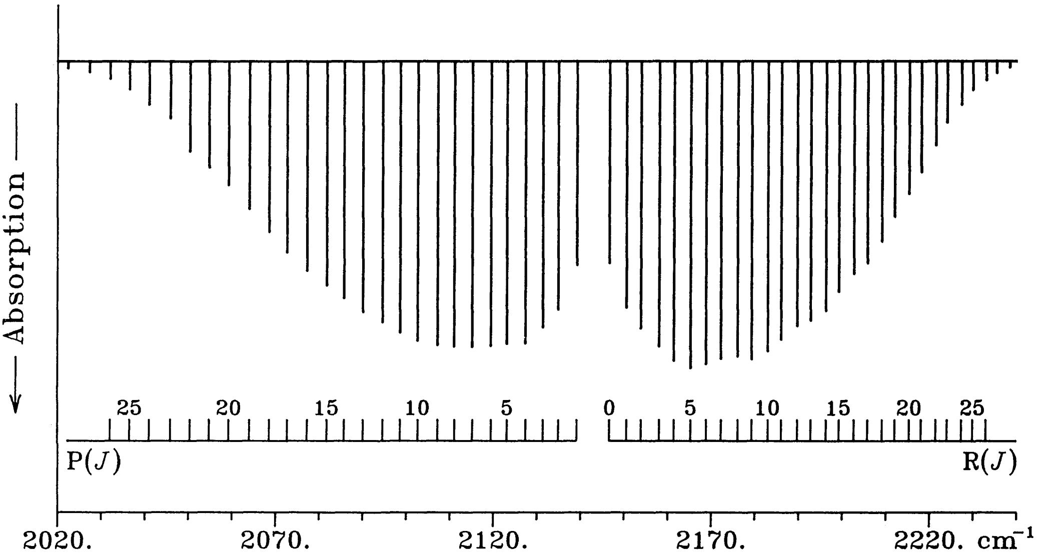

FIG. 11. The fundamental vibration–rotation band of CO in absorption (courtesy of J. W. C. Johns, National Research Council of Canada).

the last term is a small correction term that must be applied when one or more degenerate vibrations are excited. Note that the anharmonic terms are preceded by + signs, in contrast to the formula for a diatomic molecule for which − signs are customary [Eq. (37)].

As an example, let us consider the CO_2 molecule. The normal modes of vibration are ν_1, the symmetric CO stretching vibration; ν_2, the doubly degenerate bending vibration; and ν_3, the antisymmetric CO stretching vibration (Fig. 12). The two components of ν_2 can generate an angular momentum $l_2 h/2\pi$ about the linear axis depending on their relative phases. If v_2 quanta of this vibration are excited, then l_2 can take the values v_2, $v_2 - 2$, ..., 1 or 0 depending on whether v_2 is odd or even.

The vibrational term values are given by

$$\begin{aligned} G(v_1, v_2, v_3) = {} & \omega_1(v_1 + \tfrac{1}{2}) \\ & + \omega_2(v_2 + 1) + \omega_3(v_3 + \tfrac{1}{2}) \\ & + x_{11}(v_1 + \tfrac{1}{2})^2 + x_{12}(v_1 + \tfrac{1}{2}) \\ & \times (v_2 + 1) + x_{13}(v_1 + \tfrac{1}{2})(v_3 + \tfrac{1}{2}) \\ & + x_{22}(v_2 + 1)^2 + x_{23}(v_2 + 1) \\ & \times (v_3 + \tfrac{1}{2}) + x_{33}(v_3 + \tfrac{1}{2})^2 \\ & + g_{22} l_2^2. \end{aligned} \tag{51}$$

The zero-point energy is

$$\begin{aligned} G(0, 0, 0) = {} & \tfrac{1}{2}\omega_1 + \omega_2 + \tfrac{1}{2}\omega_3 + \tfrac{1}{4}x_{11} + \tfrac{1}{2}x_{12} \\ & + \tfrac{1}{4}x_{13} + x_{22} + \tfrac{1}{2}x_{23} + \tfrac{1}{4}x_{33}. \end{aligned} \tag{52}$$

The energy levels can also be expressed by a formula based on the zero-point level, viz.

$$\begin{aligned} G(v_1, v_2, v_3) = {} & \omega_1^0 v_1 + \omega_2^0 v_2 + \omega_3^0 v_3 + x_{11}^0 v_1^2 \\ & + x_{12}^0 v_1 v_2 + x_{13}^0 v_1 v_3 + x_{22}^0 v_2^2 \\ & + x_{23}^0 v_2 v_3 + x_{33}^0 v_3^2 + g_{22} l_2^2, \end{aligned} \tag{53}$$

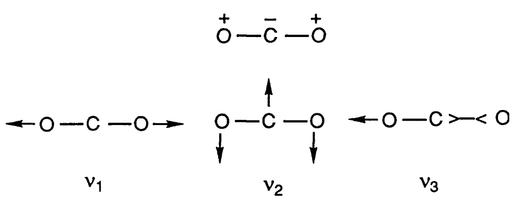

FIG. 12. Normal modes of vibration of the CO_2 molecule. The ν_2 mode is doubly degenerate.

where

$$\omega_1^0 = \omega_1 + x_{11} + x_{12} + \tfrac{1}{2}x_{13}, \tag{54}$$

$$\omega_2^0 = \omega_2 + 2x_{22} + \tfrac{1}{2}x_{12} + \tfrac{1}{2}x_{23}, \tag{55}$$

$$\omega_3^0 = \omega_3 + x_{33} + \tfrac{1}{2}x_{13} + x_{23}, \tag{56}$$

$$x_{ik}^0 = x_{ik}. \tag{57}$$

In general, for any polyatomic molecule,

$$\begin{aligned} G_0(v_1, v_2, \ldots) = {} & \sum_i \omega_i^0 v_i + \sum_i \sum_{k \geqslant i} x_{ik}^0 v_i v_k \\ & + \sum_i \sum_{k \geqslant i} g_{ik} l_i l_k + \cdots, \end{aligned} \tag{58}$$

where

$$\omega_i^0 = \omega_i + x_{ii} d_i + \tfrac{1}{2} \sum_{k \neq i} x_{ik} d_k. \tag{59}$$

The quantity that is measured experimentally and is called the fundamental vibration frequency ν is the interval between the levels $v = 1$ and $v = 0$. For the CO_2 molecule, we see from Eqs. (51) and (54) that

$$\nu_1 = \omega_1 + 2x_{11} + x_{12} + \tfrac{1}{2}x_{13} \tag{60}$$

$$= \omega_1^0 + x_{11}. \tag{61}$$

Similar expressions may be derived for ν_2 and ν_3.

2.2.1 Rotational Fine Structure

2.2.1.1 Linear Molecules. The general expression for the rotational term values of a linear molecule in a vibrational level v is

$$\begin{aligned} F_v(J) = {} & B_v[J(J + 1) - l^2] \\ & - D_v[J(J + 1) - l^2]^2 + \cdots, \end{aligned} \tag{62}$$

where l is the vibrational angular momentum about the internuclear axis, J takes the values l, $l + 1$, $l + 2$, . . . and

$$B_v = B_e - \sum_i \alpha_i(v_i + \tfrac{1}{2}d_i); \tag{63}$$

d_i is the degeneracy of the level and equals 1 if $l = 0$ and 2 if $l \neq 0$. Levels with $l = 0$, 1, 2, . . . are referred to as Σ, Π, Δ, . . . levels.

We now need to consider some symmetry properties of rotational levels. The parity of a level is *positive* or *negative* depending on whether the total wave function remains unchanged or changes sign when the coordinates of all the particles (electrons and nuclei), relative to the center of mass and measured in a space-fixed coordinate system, are replaced by their negatives. The total wave function is a product of an electronic, a vibrational, and a rotational contribution, but, for the present, we will consider the electronic contribution to be totally symmetric. For a Σ^+ state, the rotational levels with $J = 0, 1, 2, 3, \ldots$ are $+, -, +, -, \ldots$, respectively. States with $l > 0$ are doubly degenerate, and, for each value of J, there is a $+$ and a $-$ level.

For molecules with identical nuclei, there is an additional symmetry property. Levels are called *symmetric* (s) if the total wave function is unchanged when the identical nuclei are exchanged. Levels are called *antisymmetric* (a) if the total wave function changes sign when the identical nuclei are exchanged. For a Σ_g^+ state, the positive ($+$) levels are symmetric (s), and the negative levels are antisymmetric (a) (see Sec. 3.5 for further examples). If the identical nuclei have zero nuclear spin ($I = 0$), then only the s levels occur. An example is provided by the CO_2 molecule.

If $I \neq 0$, then both s and a levels occur but with different statistical weights. For a molecule with only one pair of equivalent nuclei, the ratio of the statistical weights for the s and a levels is $(I + 1)/I$ or $I/(I + 1)$ according as I is integral or half-integral, respectively.

For electric dipole transitions, *positive* levels only combine with *negative* levels, i.e.,

$$+ \leftrightarrow -, \qquad + \not\leftrightarrow +, \qquad - \not\leftrightarrow -. \tag{64}$$

Furthermore, *symmetric* levels only combine with *symmetric* levels and *antisymmetric* levels with *antisymmetric* levels. Intercombinations between *symmetric* and *antisymmetric* levels are strictly forbidden, i.e.,

$$s \leftrightarrow s, \qquad a \leftrightarrow a, \qquad s \not\leftrightarrow a. \tag{65}$$

Further selection rules for transitions are

$$\Delta l = 0, \qquad \Delta J = \pm 1 \quad \text{if } l = 0; \tag{66}$$

$$\Delta l = 0, \qquad \Delta J = 0, \pm 1 \quad \text{if } l \neq 0; \tag{67}$$

$$\Delta l = \pm 1, \qquad \Delta J = 0, \pm 1. \tag{68}$$

Σ–Σ transitions have only P and R branches, which can be expressed by Eq. (47). Π–Π and Δ–Δ transitions have, in addition, weak Q branches ($\Delta J = 0$), which are given by

$$\begin{aligned} Q(J) = \nu_0 &+ (B' - B'')[J(J + 1) - l^2] \\ &- (D' - D'')[J(J + 1) - l^2]^2. \end{aligned} \tag{69}$$

Σ–Π, Π–Σ, Π–Δ, and Δ–Π transitions have P and R branches and strong Q branches given by Eqs. (47) and (69). It should be noted that when $l = 1$, there is a small splitting of the levels, called l-type doubling, and additional terms should be added to Eqs. (47) and (69).

2.2.1.2 Symmetric-Top Molecules. The rotational term values for a symmetric-top molecule in a nondegenerate state are given by Eq. (19). If a degenerate vibration is excited, then additional Coriolis terms need to be added.

As with linear molecules, the rotational levels are either *positive* or *negative*, depending on whether the total wave function remains unchanged or changes sign when the coordinates of all the particles are replaced by their negatives. This symmetry property is important for planar molecules. For nonplanar molecules, each rotational level is doubly degenerate and consists of a $+$ and a $-$ component. The splitting is not resolved, except for a few molecules such as NH_3, which can tunnel fairly freely through the mean molecular plane.

For a molecule with an n-fold axis of symmetry, there are additional symmetry properties analogous to the *symmetric* and *antisymmetric* properties for linear molecules. If we consider a molecule with a threefold axis of symmetry, e.g., CX_3Y, then rotation of the molecule by 120° about the symmetry axis exchanges identical nuclei. In the ground state, levels with $K = 0, 3, 6, 9, \ldots$ have a larger statistical weight than levels with $K = 1, 2, 4, 5, 7, 8, \ldots$. The ratio of the statistical weights depends on the nuclear spin of X. There is a selection rule, analogous to Eq. (65), that only levels with the same overall symmetry can combine with each other.

If we consider only transitions between nondegenerate vibrational levels, two types of band structure are possible, depending on the direction of the change of dipole moment, known as the *transition moment*. If the transition moment is parallel to the symmetric-top axis, then the selection rules are

$$\Delta K = 0, \qquad \Delta J = \pm 1 \quad \text{if } K = 0; \tag{70}$$

$$\Delta K = 0, \qquad \Delta J = 0, \pm 1, \quad \text{if } K \neq 0. \tag{71}$$

A parallel band therefore is composed of a series of sub-bands similar to Σ–Σ, Π–Π, Δ–Δ, . . . bands of a linear molecule. The origins of the sub-bands are given by

$$\nu_0^{\text{sub}} = \nu_0 + [(A' - A'') - (B' - B'')]K^2. \tag{72}$$

A diagram showing the structure of a parallel band is given in Fig. 13.

If the transition moment is perpendicular to the symmetric-top axis, then

$$\Delta K = \pm 1, \qquad \Delta J = 0, \pm 1. \tag{73}$$

The band consists of a superposition of sub-bands similar to the Σ–Π, Π–Σ, Π–Δ, Δ–Π, . . . bands of a linear molecule. The origins of the sub-bands are given by

$$\begin{aligned}\nu_0^{\text{sub}} = \nu_0 &+ (A' - B') \pm 2(A' - B')K \\ &+ [(A' - B') - (A'' - B'')]K^2,\end{aligned} \tag{74}$$

where the + sign refers to $\Delta K = +1$ (with $K = 0, 1, 2, \ldots$) and the − sign to $\Delta K = -1$ (with $K = 1, 2, 3 \ldots$). A typical band structure is given in Fig. 14.

The intensity I of a line in an absorption band is given by

$$I(J, K) = C\nu S_{JK} g_{JK} \exp\left(-\frac{hcF''(J, K)}{kT}\right), \tag{75}$$

where C is a constant, S_{JK} is a Hönl–London line-strength factor (see Table 2), g_{JK} is the statistical weight of the lower state, and the last term is the Boltzmann factor for the lower state. The statistical weight factor equals $2J + 1$ if $K = 0$ and $2(2J + 1)$ if $K \neq 0$. The table also applies to linear molecules if K is replaced by l.

The line strength S_{JK} is not highly dependent on J, the statistical weight factor increases linearly with J, while the Boltzmann factor decreases exponentially with increasing J. Consequently, the intensity in a branch usually increases with J for lower J values and decreases for higher J values (see Figs. 13 and 14). It may be noted that, when $K = 0$, the intensity of the Q branch is zero.

In emission, the intensity distribution is given by

$$I(J, K) = C\nu^4 S_{JK} g_{JK} \exp\left(-\frac{hcF'(J, K)}{kT}\right) \tag{76}$$

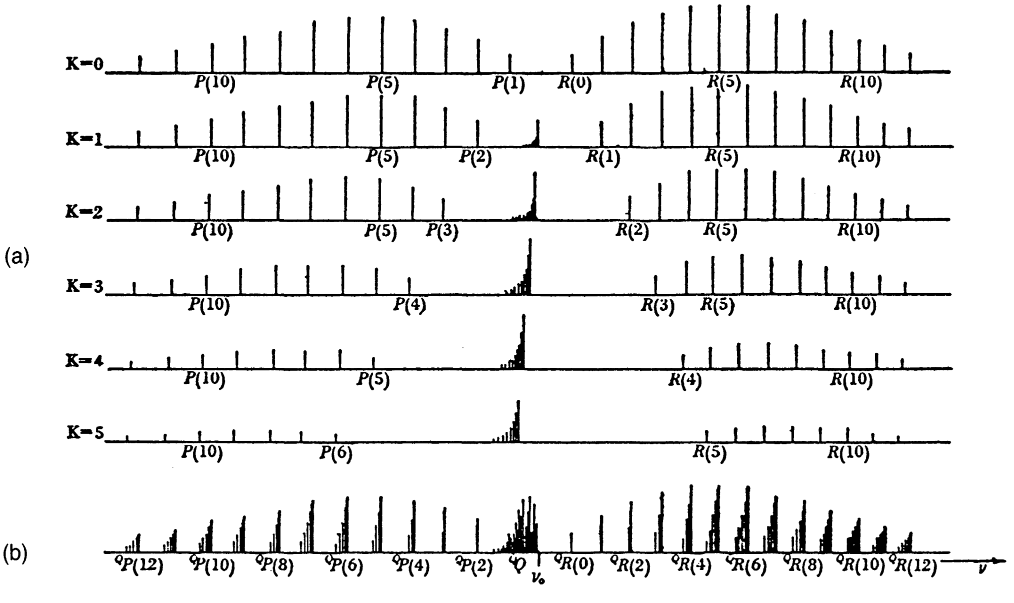

FIG. 13. Parallel band of a symmetric-top molecule: (a) the sub-band structure; and (b) the summation (from Herzberg, 1945, by courtesy of the author and publishers).

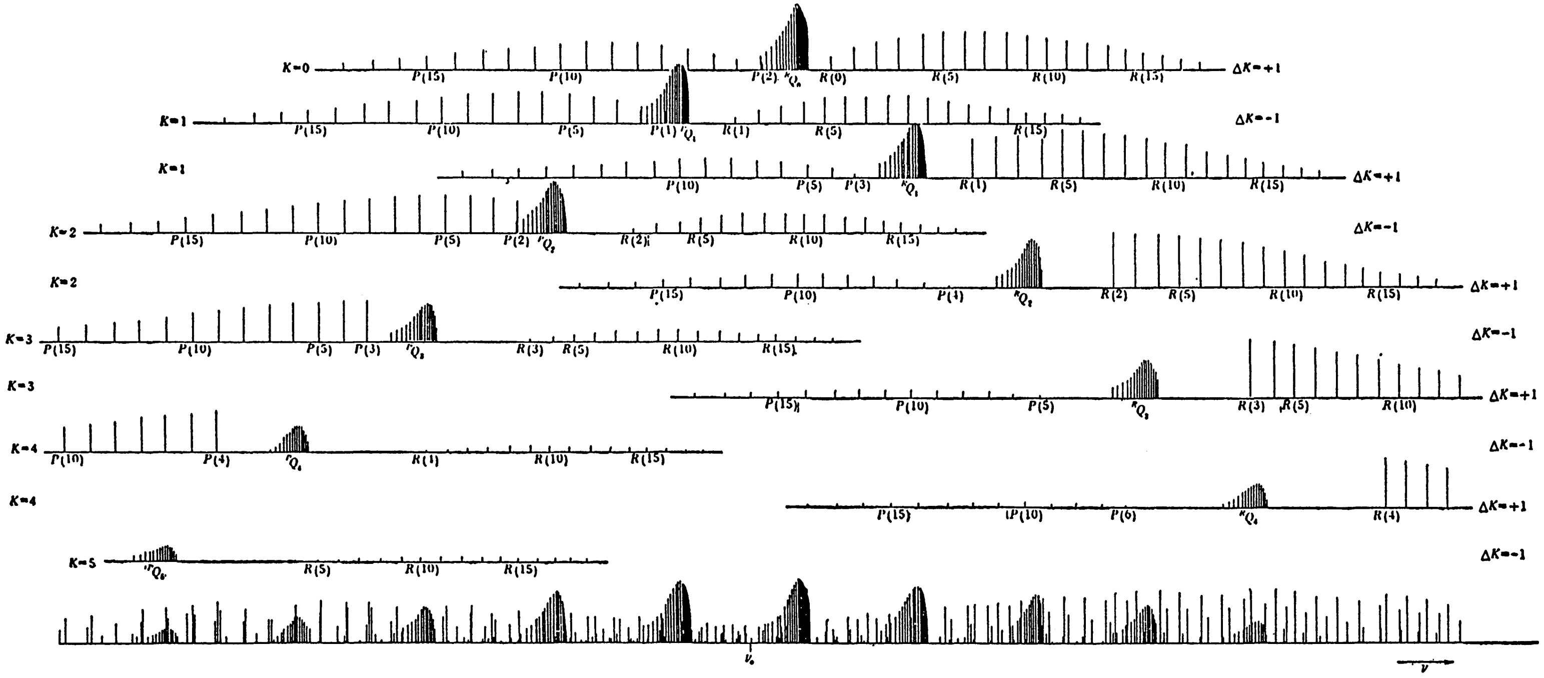

FIG. 14. Perpendicular band of a symmetric-top molecule. The sub-band structure is shown, with the summation in the bottom strip (from Herzberg, 1945, by courtesy of the author and the publishers).

Table 2. Hönl–London line strength factors. If $K = 0$ and $\Delta K = +1$, the values given have to be multiplied by 2.

	$\Delta K = +1$	$\Delta K = 0$	$\Delta K = -1$
R	$\frac{(J+2+K)(J+1+K)}{(J+1)(2J+1)}$	$\frac{(J+1)^2 - K^2}{(J+1)(2J+1)}$	$\frac{(J+2-K)(J+1-K)}{(J+1)(2J+1)}$
Q	$\frac{(J+1+K)(J-K)}{J(J+1)}$	$\frac{K^2}{J(J+1)}$	$\frac{(J+1-K)(J+K)}{J(J+1)}$
P	$\frac{(J-1-K)(J-K)}{J(2J+1)}$	$\frac{J^2 - K^2}{J(2J+1)}$	$\frac{(J-1+K)(J+K)}{J(2J+1)}$

and depends on the fourth power of the frequency and the Boltzmann factor for the upper state. The product $S_{JK}g_{JK}$ is the same as for absorption.

2.2.1.3 Asymmetric-Top Molecules. The rotational structures are very similar to those for pure rotational spectra, except that the band type depends on the direction of the change of dipole moment rather than on the dipole moment itself. There are three types of band: *a* type, *b* type, and *c* type, depending on whether the transition moment is along the axis of least, intermediate, or largest moment of inertia, respectively.

The selection rules are the same as those given in Eqs. (22), (23), and (24). If the direction of the transition moment is not along one of the inertial axes, then hybrid bands are observed with more than one band type.

The rotational structures of asymmetric-top molecules are not, in general, simple and will not be discussed further.

2.2.1.4 Spherical-Top Molecules. Spherical top molecules, e.g., CH_4, SiH_4, pose special problems on account of the high symmetry of some of the levels and splittings caused by Coriolis interactions. Since very few molecules belong to this class, the rotational fine structures will not be considered further.

2.3 Infrared Techniques

Most of the early infrared spectra were taken with prism or grating spectrometers. Nowadays, most infrared spectra are recorded with commercial Fourier-transform spectrometers. The resolution can be varied from a few cm^{-1} for low-resolution scans to 0.001 cm^{-1} for high-resolution scans. An example of a high-resolution recording is shown in Fig. 15, which illustrates the Q branch of the 588.8-cm^{-1} band of N_2O.

A technique that has been used for the detection of free-radical species is laser magnetic resonance (LMR). The species are flowed through a cell placed within the cavity of a fixed-frequency laser. The CO_2 and N_2O lasers are used in the region 900 to 1100 cm^{-1}, while CO is used at higher frequencies (1250 to 2000 cm^{-1}). The laser is tuned to one of the frequencies in the vibration-rotation band. Since these frequencies rarely coincide with a frequency of a free-radical species, the latter is tuned to resonance by the application of a magnetic field. An example of an LMR spectrum of CH_2 is shown in Fig. 16. This technique has been used for the study of many free-radical species, e.g., FO, CH_2, NH_2, HCO, and FeH. One of the advantages of this technique is that it provides a sensitive means of detection for paramagnetic species but does not detect stable diamagnetic species such as the parent and product molecules. The technique was originally developed for use in the far-infrared region of the spectrum (Evenson *et al.*, 1980).

A similar technique uses electric instead of magnetic fields and is known as laser electric resonance or laser Stark spectroscopy. The technique is sensitive to all species possessing a dipole moment and is thus less selective than laser magnetic resonance.

Other types of laser in common use are infrared diode lasers, difference-frequency lasers, and color-center lasers. Infrared diode lasers are very widely used and available for most regions of the infrared spectrum (3–30 μm). They can be tuned over limited frequency ranges, normally ±50 cm^{-1}, but the tuning characteristics frequently display dis-

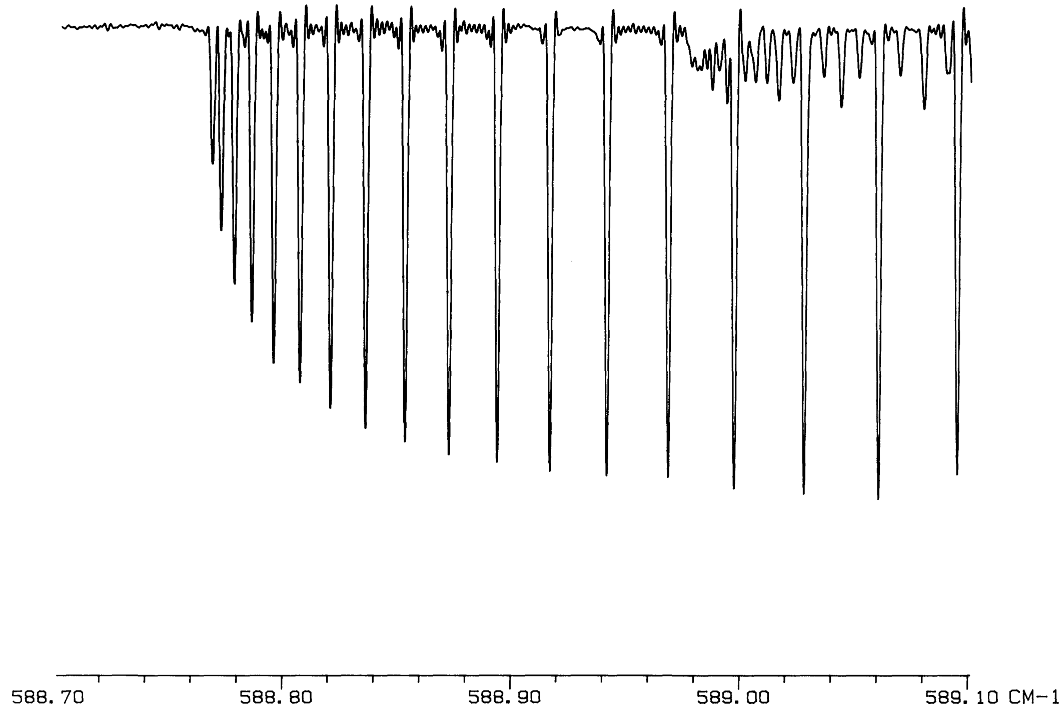

FIG. 15. *Q* branch of the 588-cm^{-1} band of N_2O recorded with a Bomem Fourier-transform spectrometer with a resolution of 0.0014 cm^{-1} (by courtesy of J. W. C. Johns, National Research Council of Canada).

continuities due to what are called "mode hops."

Difference-frequency lasers involve the mixing of two visible lasers, one fixed and the other tunable, in a nonlinear crystal such as lithium niobate. Tunable infrared radiation in the region 2000 to 4500 cm^{-1} is obtained with powers of about 10 μW. The laser has been used for the detection of molecular ions and other transient species. The first molecular ion to be detected by this method was H_3^+. An ion may be produced in an ac electric discharge so that its velocity is modulated, first in one direction and then in the other. The change in frequency caused by the Doppler effect is then sufficient so that the modulated frequency can be detected by a lock-in amplifier. Both positive and negative ions have been detected using this technique.

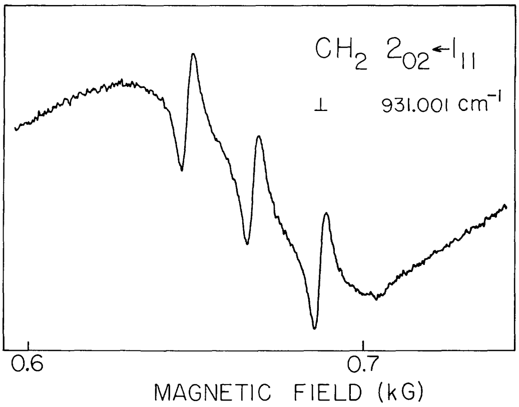

FIG. 16. Laser magnetic resonance (LMR) spectrum of CH_2 (unpublished figure from Sears *et al.*, 1982). One magnetic component only is shown. The triplet structure arises from proton hyperfine interactions.

Color-center lasers are produced by pumping alkali halide crystals with a Kr^+ laser. The lasers operate over the frequency range 4000 to 3000 cm^{-1} and have been used mainly for the detection of transient species in this region. The most notable example at present is the C_2H radical.

2.4 Applications of Infrared Spectroscopy

Infrared spectroscopy has widespread applications in many branches of chemistry. It was stated as early as the 1940s that "the in-

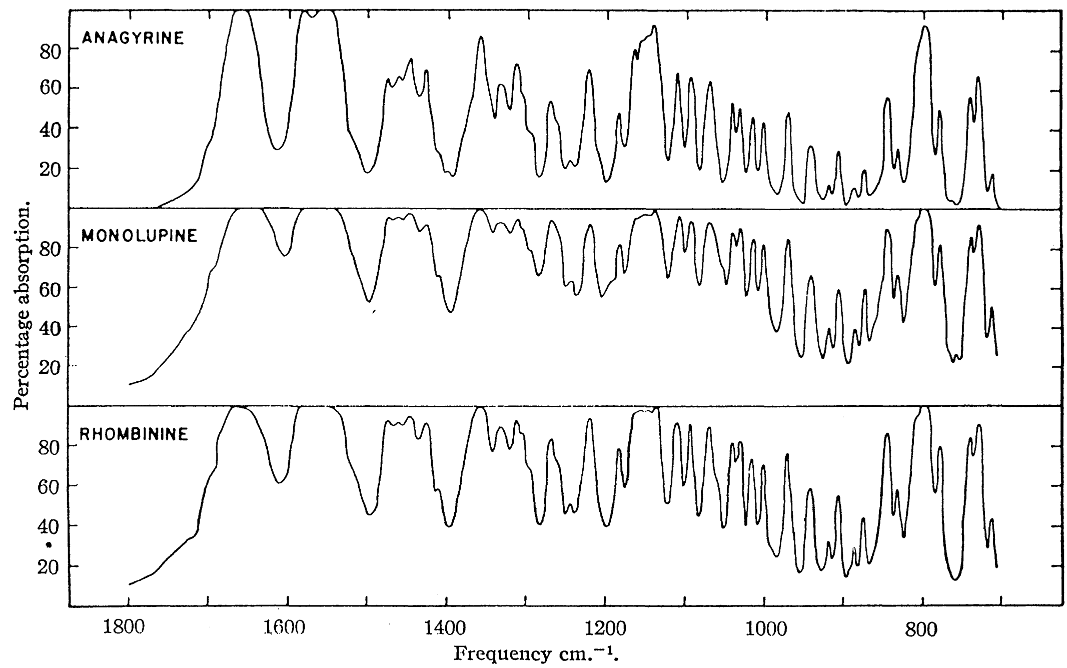

FIG. 17. Infrared spectra of anagyrine, monolupine, and rhombinine (from Marion and Ouellet, 1948, by permission of the journal).

fra-red spectrum of a chemical compound is probably the most characteristic physical property of that compound" (Sutherland and Thompson, 1945). Organic molecules containing large numbers of atoms have many fundamental bands spanning the range from about 10 to 3500 cm^{-1}, as well as overtone and combination bands. Since these bands may have widely different intensities, it is not difficult to understand why the infrared spectrum provides a unique fingerprint of the compound.

An early example involves three alkaloids, anagyrine, monolupine, and rhombinine, which were initially thought to be different compounds. Classical studies provided strong evidence that the compounds were identical, and this was convincingly confirmed by the identity of their infrared spectra (Fig. 17) (Marion and Ouellet, 1948).

Infrared spectra are frequently used for the analysis of mixtures of compounds or the detection of impurities. These analyses may use substances in the gas phase or in condensed phases. An example is provided by the detection of small amounts of impurities, e.g., SO_2 or chlorofluorocarbons in the Earth's atmosphere.

Infrared techniques are also used for the detection of molecules in astronomical sources. The presence of CO_2, NH_3, and CH_4 in planetary atmospheres has been known for a long time, and recently a spectrum of the H_3^+ ion has been observed in emission from Jupiter (Fig. 18). Molecules such as CO, CH_4, and C_2H_2 have also been found in infrared spectra of regions of star formation and circumstellar envelopes.

3. ELECTRONIC SPECTRA

3.1 Classification of Electronic States

The various electronic states of a molecule are classified according to the resultant spin $Sh/2\pi$ of all the electrons and the symmetry species of the electronic wavefunction. The *multiplicity* of an electronic state is defined as $2S + 1$. Molecules with an even number of electrons give rise to states with $S = 0$ (singlet), $S = 1$ (triplet), $S = 2$ (quintet), and so on. Species with an odd number of electrons give rise to states with $S = \frac{1}{2}$ (doublet), $S = \frac{3}{2}$ (quartet), $S = \frac{5}{2}$ (sextet), and so on.

For diatomic and linear polyatomic mol-

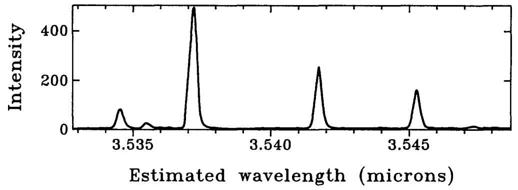

FIG. 18. Emission lines of H_3^+ observed in the spectrum of Jupiter (courtesy of T. Oka, University of Chicago).

ecules, the symmetry species of the electronic wave function depends on the component $\Lambda h/2\pi$ of the orbital angular momentum of the electrons along the axis of the molecule. For $\Lambda = 0, 1, 2, \ldots$, the states are designated as $\Sigma, \Pi, \Delta, \ldots$ (compare the usage of $s, p, d, \ldots$ in atoms).

For nonlinear molecules, the symmetry species of the electronic wave function depends on the symmetry group (point group) of the molecule and will belong to one of its symmetry types, e.g., A_1, B_2, B_{1u}, E_{2g}, ... An example will be given below.

3.2 Intensities of Transitions

The probability P of an electric dipole transition between two electronic states with wave functions ψ_e' and ψ_e'' is proportional to

$$P \propto \left|\int \psi_e'^{*} \mu \psi_e'' d\tau_e\right|^2, \qquad (77)$$

which is often written in matrix notation as

$$P \propto |\langle \psi_e' | \mu | \psi_e'' \rangle|^2, \qquad (78)$$

where the asterisk denotes the complex conjugate and μ is the dipole-moment operator with components μ_x, μ_y, μ_z such that $\mu_x = \Sigma_i e_i x_i$, $\mu_y = \Sigma_i e_i y_i$, and $\mu_z = \Sigma_i e_i z_i$. The product of the symmetry species of ψ_e', μ, and ψ_e'' must be totally symmetric; otherwise, the integral (matrix element) will vanish, in which case the transition is forbidden.

Let us consider a symmetrical AH_2 molecule that belongs to the point group C_{2v}. The axis system is given in Fig. 19 with the x axis perpendicular to the plane of the molecule. The group table is given in Table 3. I denotes the identity operator, $C_2(z)$ a rotation by 180° about the z axis, $\sigma(xz)$ a reflection in a plane perpendicular to the plane of the molecule, and $\sigma(yz)$ a reflection in the plane of the molecule. A +1 denotes that the operation leaves the wave function unchanged, while a −1 indicates that the operation changes the sign of the wave function.

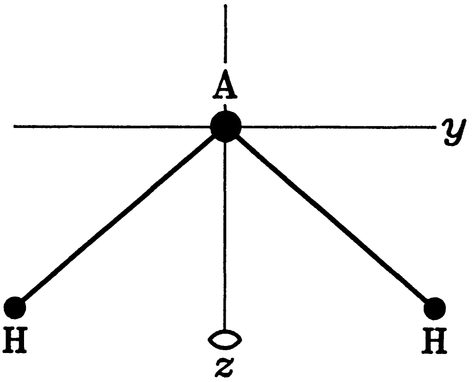

FIG. 19. Axes for an AH_2 molecule. The x axis lies perpendicular to the molecular plane.

Let us consider the symmetry species of the dipole-moment operators μ_x, μ_y, and μ_z. The term μ_x changes sign when rotated by 180° about the z axis, remains unchanged on reflection in the xz plane, and changes sign when reflected in the yz plane. It therefore belongs to the symmetry type B_1. Similarly, μ_y and μ_z have species B_2 and A_1, respectively. There is no component of the dipole-moment operator with symmetry type A_2.

Consider now a transition between two electronic states: The product of the symmetry species of ψ_e', μ, and ψ_e'' must be totally symmetric (A_1); otherwise, the integral (matrix element) will vanish. The low-lying electronic states of the NH_2 free radical have species $^2B_1$, $^2A_1$, and $^2B_2$ (see Fig. 20). The transition between the ground state ($^2B_1$) and the first excited state ($^2A_1$) is allowed with a transition moment in the x direction since $A_1 \times B_1 \times B_1 = A_1$. Similarly, the transition between the two excited states is allowed with a transition moment in the y direction since $B_2 \times B_2 \times A_1 = A_1$. The transition between the ground state and the second excited state

Table 3. Group table for a C_{2v} molecule.

C_{2v}	I	$C_2(z)$	$\sigma(xz)$	$\sigma(yz)$
A_1	+1	+1	+1	+1
A_2	+1	+1	−1	−1
B_1	+1	−1	+1	−1
B_2	+1	−1	−1	+1

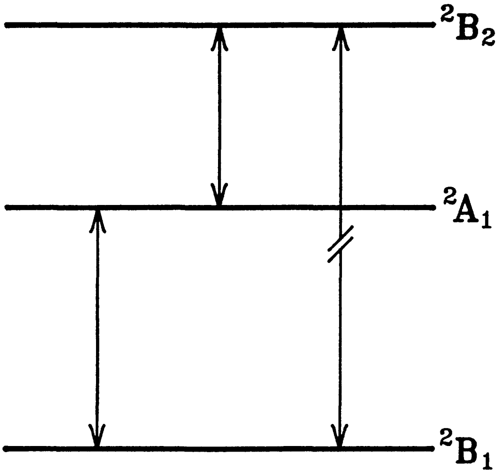

FIG. 20. Low-lying electronic states of NH_2.

is forbidden since the product of the wave functions $B_2 \times B_1 = A_2$, and there is no component of the dipole-moment operator with symmetry A_2.

3.3 Franck–Condon Principle

According to the classical model proposed by Franck, the positions and velocities of the nuclei remain virtually unchanged during an electronic transition, and the molecule then executes vibrational motion about the new equilibrium configuration. This concept was treated on a quantum-mechanical basis by Condon, who showed that the probability P of a transition between two states with total wave functions ψ' and ψ'' is given by

$$P \propto |\int \psi'^* \mu \psi'' d\tau|^2. \tag{79}$$

If the electronic and vibrational motions are completely separable, then

$$\psi = \psi_e \psi_v, \tag{80}$$

and the intensity I of a transition is given by

$$I \propto |\int \psi_e'^* \, \psi_v'^* \mu \psi_e'' \psi_v'' \, d\tau_e d\tau_v|^2. \tag{81}$$

If we separate the dipole-moment operator μ into one part that depends on the electronic coordinates and another part that depends on the nuclear coordinates, then

$$\mu = \mu_e + \mu_n, \tag{82}$$

and

$$\begin{aligned} I \propto |&\int \psi_e'^* \mu_e \psi_e'' \, d\tau_e \int \psi_v'^* \psi_v'' \, d\tau_v \\ &+ \int \psi_v'^* \mu_n \psi_v'' \, d\tau_v \int \psi_e'^* \psi_e'' \, d\tau_e|^2. \end{aligned} \tag{83}$$

But since the wave functions for different electronic states are orthogonal to each other,

$$\int \psi_e'^* \psi_e'' \, d\tau_e = 0. \tag{84}$$

Hence,

$$I \propto |\int \psi_e'^* \mu_e \psi_e'' \, d\tau_e \int \psi_v'^* \psi_v'' \, d\tau_v|^2, \tag{85}$$

or, in matrix notation,

$$I \propto |\langle \psi_e' | \mu | \psi_e'' \rangle \langle \psi_v' \psi_v'' \rangle|^2. \tag{86}$$

The first term in Eq. (86) (unsquared) is called the electronic transition moment, and the second term is the vibrational overlap integral.

It is readily seen from Eq. (86) that, for the intensity of a transition to be nonvanishing, the product of the symmetry species of ψ_v' and ψ_v'' must be totally symmetric. This condition is automatically satisfied for a diatomic molecule since ψ_v is always totally symmetric, but for a polyatomic molecule, it gives rise to selection rules originally stated by Herzberg and Teller, viz., for totally symmetric vibrations,

$$\Delta v = 0, 1, 2, \ldots; \tag{87}$$

for non-totally symmetric vibrations,

$$\Delta v = 0, 2, 4, \ldots. \tag{88}$$

A series of bands in which one vibrational quantum number changes is known as a *progression*. For totally symmetric vibrations, the length of a progression depends on the change in geometric structure accompanying the electronic transition (see HCN below). For a v'–0 transition involving a non-totally symmetric vibration,

$$\frac{I(0—0)}{\Sigma_{v'} I(v'—0)} = \frac{(v'v'')^{1/2}}{\frac{1}{2}(v' + v'')}, \tag{89}$$

where I denotes band intensity. If $v' = v''$, this ratio equals 1, and all the intensity in the progression lies in the 0—0 band. If $v' = \frac{1}{2}v''$, the above ratio equals 0.943; hence, 94.3% of

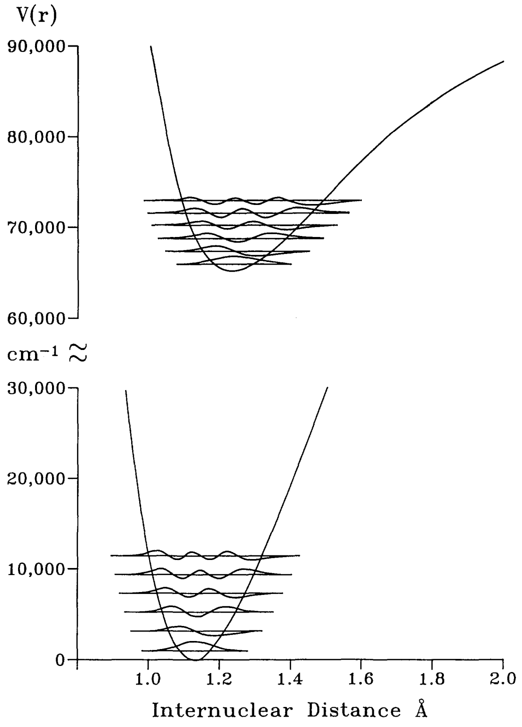

FIG. 21. Potential curves, vibrational levels, and wave functions for the $A\ ^1\Pi$–$X\ ^1\Sigma^+$ transition of the CO molecule (courtesy of R. W. Nicholls, York University, Toronto).

the intensity of the progression lies in the 0—0 band and only 5.7% in the 2—0, 4—0, . . . bands. Thus non-totally symmetric vibrations do not form strong progressions in molecular spectra. Such vibrations may give rise to *sequence* bands, i.e., bands of the type 0—0, 1—1, 2—2, . . ., particularly if the vibrational frequency is small.

An example of the vibrational levels and wave functions for two electronic states of a diatomic molecule is given in Fig. 21 for the $A\ ^1\Pi$–$X\ ^1\Sigma^+$ band system of CO. It is interesting to note that the $v = 0$ wave functions have a single maximum at the midpoint of the vibration, in contrast to classical mechanics, where the greatest probabilities are at the turning points. Furthermore, the wave function is finite, though small, outside the range of the classical turning points. The numbers of nodes in the wave functions are equal to v. A table of Franck–Condon factors $\langle\psi'_v\psi''_v\rangle^2$ is given in Table 4. In absorption, the $v'' = 0$ progression shows a single maximum at $v' = 2$; in emission, the $v' = 0$ progression also shows a single maximum at $v'' = 2$. The $v'' = 1$ progression shows two intensity maxima at $v' = 0$ and 5, while the $v' = 1$ progression shows maxima at $v'' = 0$ and 4, consistent with the fact that there are two maxima in a $v = 1$ wave function.

For a polyatomic molecule, a large change in equilibrium geometry for different electronic states is often found, especially for smaller species. For example, HCN is linear in its ground state, and the CN bond length is 1.156 Å. In the first excited state, the molecule is bent with an equilibrium angle of 125° and a CN bond length of 1.297 Å. As a result, the absorption spectrum of HCN in the ultraviolet consists of a long progression of the bending vibration in the excited state, with values of the quantum number v'_2 ranging from 0 to 10 and a maximum of intensity in the region of $v'_2 = 7$. Similar progressions are observed in which one and two quanta of the CN stretching vibration are also excited (see Fig. 22).

3.4 Vibronic Interactions

In the event that the electronic and vibrational motions are not completely separable, the system can be described by a vibronic wave function ψ_{ev} whose symmetry species is

Table 4. Franck–Condon factors for the $A^1\Pi$-$X^1\Sigma$ band system of CO (from Nicholls, 1962).

v' \ v''	0	1	2	3	4	5	6
0	1.13×10^{-1}	2.61×10^{-1}	2.85×10^{-1}	1.96×10^{-1}	9.60×10^{-2}	3.55×10^{-2}	1.03×10^{-2}
1	2.16×10^{-1}	1.55×10^{-1}	3.05×10^{-3}	7.64×10^{-2}	1.93×10^{-1}	1.86×10^{-1}	1.08×10^{-1}
2	2.30×10^{-1}	1.22×10^{-2}	9.01×10^{-2}	1.16×10^{-1}	5.08×10^{-3}	5.72×10^{-2}	1.65×10^{-1}
3	1.81×10^{-1}	2.05×10^{-2}	1.17×10^{-1}	6.45×10^{-4}	8.96×10^{-2}	8.42×10^{-2}	4.71×10^{-4}
4	1.19×10^{-1}	8.73×10^{-2}	3.44×10^{-2}	5.76×10^{-2}	6.66×10^{-2}	6.17×10^{-3}	9.80×10^{-2}
5	6.88×10^{-2}	1.23×10^{-1}	3.24×10^{-4}	9.13×10^{-2}	1.38×10^{-5}	8.24×10^{-2}	2.27×10^{-2}
6	3.67×10^{-2}	1.16×10^{-1}	3.21×10^{-2}	4.22×10^{-2}	4.29×10^{-2}	4.23×10^{-2}	2.18×10^{-2}

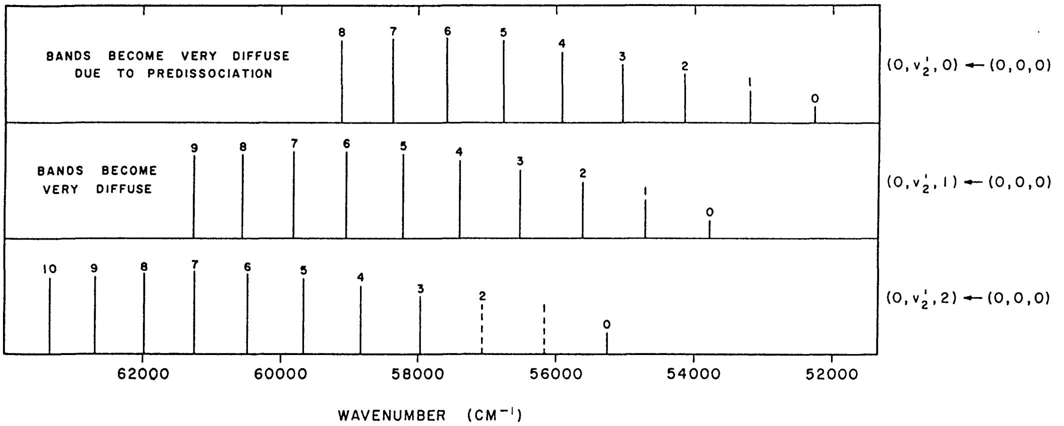

FIG. 22. Relative intensities of the bands in the ultraviolet spectrum of HCN (from Herzberg and Innes, 1957). The intensities are plotted on a logarithmic scale. The bands are plotted as separate progressions (0, v_2', v_3')–(0, 0, 0) with $v_3' = 0$, 1, and 2.

the direct product of the electronic and vibrational species.

Suppose we have an electronically forbidden transition, so that

$$\langle\psi_e'|\mu|\psi_e''\rangle = 0. \qquad (90)$$

There may be non-totally symmetric vibrations of appropriate symmetry so that

$$\langle\psi_{ev}'|\mu|\psi_{ev}''\rangle \neq 0. \qquad (91)$$

Weak bands may then be observed and are said to be vibronically induced. Classical examples are found in the spectra of formaldehyde and benzene of which the former will be considered here.

Formaldehyde, H_2CO, belongs to the point group C_{2v} and has $3 \times 4 - 6 = 6$ normal modes of vibration (see Fig. 23). Three are totally symmetric (a_1) and may be described as a symmetric C—H stretching vibration ν_1, a CO stretching vibration ν_2, and a symmetric H—C—H bending vibration ν_3. There is one vibration of the type b_1, which is the out-of-plane C—H wagging motion ν_4. The remaining two vibrations belong to the type b_2 and are the antisymmetric C—H stretching vibration ν_5 and the CH_2 rocking vibration ν_6. There are no fundamental vibrations of the type a_2. The frequencies of these vibrations in the ground ($\tilde{X}\,^1A_1$) and first excited singlet state ($\tilde{A}\,^1A_2$) are given in Table 5. The CO stretching frequency ν_2 changes considerably from the ground state to the excited state, corresponding to a lengthening of the CO bond from 1.21 Å to 1.32 Å. The most dramatic change involves ν_4. In the ground state, the molecule is planar, and the out-of-plane wagging frequency is 1167.3 cm^{-1}. In the excited state, the molecule is slightly nonplanar, and the out-of-plane frequencies are very anharmonic. Successive levels are 124.5 cm^{-1} (b_1), 542.3 cm^{-1} (a_1), 947.9 cm^{-1} (b_1), 1429.3 (a_1), . . ., where the symmetry of the vibrational level is b_1 for odd values of v_4' and a_1 for even values of v_4'.

An energy-level diagram showing some of the bands observed in the near-ultraviolet spectrum of formaldehyde is given in Fig. 24. The 0–0 band is forbidden by the electric dipole selection rules. The first band in the ab-

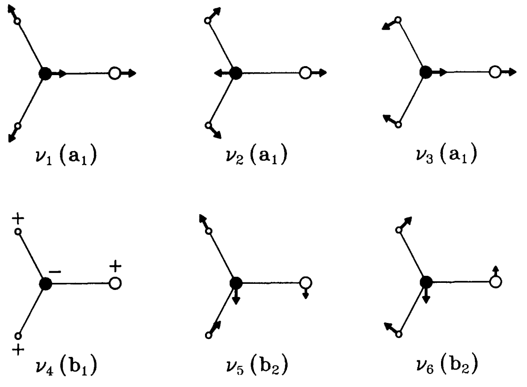

FIG. 23. Normal modes of vibration for formaldehyde, H_2CO.

Table 5. Vibrational frequencies for formaldehyde in its ground ($\tilde{X}\ {}^1A_1$) and excited ($\tilde{A}\ {}^1A_2$) states in cm^{-1} (from Clouthier and Ramsay 1983).

Vibration	$\tilde{X}\ {}^1A_1$	$\tilde{A}\ {}^1A_2$
$\nu_1(a_1)$	2782.5	2846
$\nu_2(a_1)$	1746.0	1183
$\nu_3(a_1)$	1500.2	1293.1
$\nu_4(b_1)$	1167.3	[a]
$\nu_5(b_2)$	2843.3	2968.3
$\nu_6(b_2)$	1249.1	904

---

[a] These levels are very anharmonic and take successive values of 124.5 (b_1), 542.3 (a_1), 947.9 (b_1), 1429.3 (a_1), · · ·.

sorption spectrum from the ground state is called the *A* band and terminates on the level with $v_4' = 1$. The vibronic symmetry of the latter is ${}^eA_2 \times {}^vB_1 = {}^{ev}B_2$. The transition is therefore vibronically B_2—A_1, and the transition moment lies in the plane of the molecule in a direction at right angles to the CO bond. The rotational structure is found to be *b* type, which is consistent with the above interpretation. The *A* band is the first band of a progression involving the stretching vibration of the CO bond ν_2' in the excited state. Another band similar in type to the *A* band is the *B'* band, which terminates on the level with $v_4' = 3$ in the excited state. The *B'* band is the first member of another progression involving ν_2'.

In emission, the first band from the zero level in the excited state terminates on the level with $v_4'' = 1$ in the ground state and is called the α band. This band is the first member of a progression involving the CO stretching vibration ν_2'' in the ground state. The first band in emission from the $\nu_4 = 1$ level of the excited state is the *A* band, which is weak on account of Franck–Condon considerations. The next band terminates on $v_4'' = 2$ and is called the β band. This band is also the first member of a progression involving ν_2''. It should be mentioned that the levels involving ν_4 in the ground state are coupled to levels involving ν_6 by a mechanism known as Coriolis interaction. The *b*-type bands in emission that terminate on levels involving ν_4 are accompanied by weaker *c*-type bands at lower frequencies terminating on the levels involving ν_6.

3.5 Rotational Fine Structure

Electronic bands of diatomic and simple polyatomic molecules frequently show rotational fine structure that can be analyzed to give some information on the geometrical structure of the molecule in the excited state.

For linear molecules, the expression for the

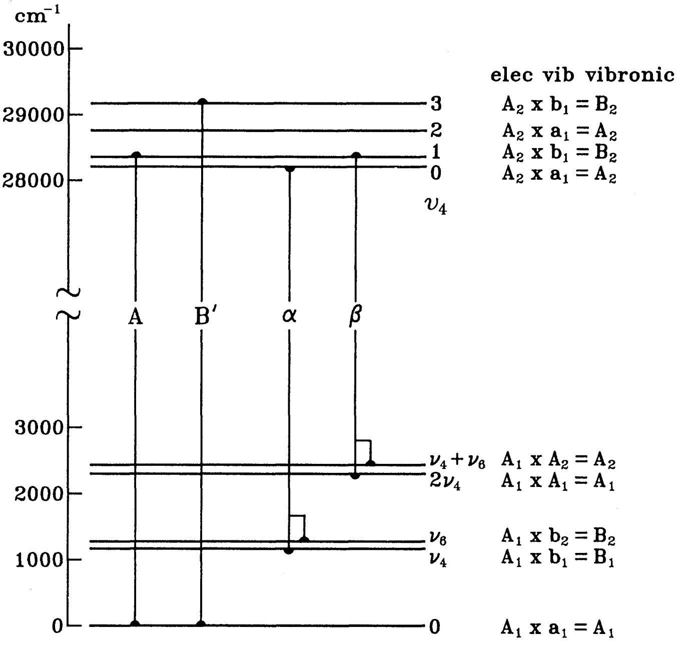

FIG. 24. Energy-level diagram showing some of the principal bands in the near-ultraviolet spectrum of formaldehyde.

rotational term values is the same as in Eq. (62), except that l is replaced by Λ for a diatomic molecule and K for a polyatomic molecule. In the latter case, there are two possible contributions to the resultant angular momentum about the internuclear axis. The first is the electronic angular momentum $\Lambda h/2\pi$, and the other is the vibrational angular momentum $lh/2\pi$. The resultant is $Kh/2\pi$, where

$$K = |\Lambda \pm l|. \tag{92}$$

We now need to consider two additional symmetry properties. If the molecule has a center of symmetry, then the *electronic* wave function either remains unchanged or changes sign when the coordinates of all the electrons are replaced by their negatives. In the former case, the state is labeled g (*gerade*) and in the latter u (*ungerade*). For a molecule in a Σ electronic state, the *electronic* wave function either remains unchanged (Σ^+ state) or changes sign (Σ^- state) when reflected in any plane through the linear axis. The symmetry properties of the rotational levels for various types of electronic state are given in Fig. 25.

The selection rules for transitions involving singlet states are the same as those for infrared bands, replacing l by Λ or K [see Eqs. (64) to (68)]. The intensity factors are the same as those given in Table 2, with K replaced by Λ or l as required. A further selection rule is

$$\Delta S = 0, \tag{93}$$

although many weak systems are known, e.g., triplet-singlet, which violate this rule.

The B values for the molecule in the ground and excited states are frequently quite different, and so a "head" may form in either the P or the R branch. It is seen that the Fortrat parabola [Eq. (46)] forms a head when $d\nu/dm = 0$ and

$$m = -(B + B'')/2(B' - B''). \tag{94}$$

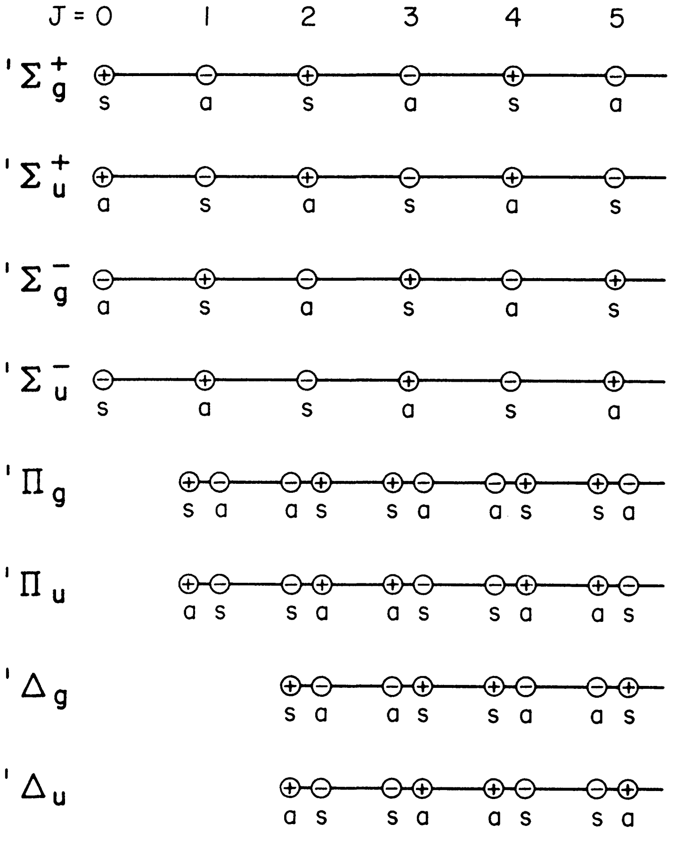

FIG. 25. Symmetry properties for the rotational levels of linear molecules. The group symbols refer to the vibronic species of the molecule. (+) denotes positive and (−) negative rotational levels; *s* stands for symmetric and *a* for antisymmetric rotational levels. For molecules which do not possess a center of symmetry, the *g* and *u*, and the *s* and *a* symbols may be deleted.

The separation between the band head and the band origin is given by

$$\nu_{\text{Head}} - \nu_0 = -(B' + B'')^2/4(B' - B''). \tag{95}$$

If $B' < B''$, m is positive, and the head is formed in the R branch. If $B' \geq B''$, m is negative, and the head is formed in the P branch. R heads are observed more frequently than P heads since bond lengths are usually larger in the excited state and $B' < B''$.

The nature of an electronic transition can frequently be deduced from the observation of the first lines in the branches. Remembering that J must always be equal to or larger than Λ or K, it is easy to deduce that the first lines in the P, Q, and R branches are $P(2)$, $Q(1)$, and $R(0)$ for a $^1\Pi$—$^1\Sigma$ transition, but $P(1)$, $Q(1)$, and $R(1)$ for a $^1\Sigma$—$^1\Pi$ transition.

For linear molecules in states other than singlet states, the nature of the coupling between the resultant electron spin and the orbital angular momentum needs to be considered. Various coupling cases were treated originally by Hund and expressions for rotational term values deduced. Additional selection rules also apply.

For symmetric-top and asymmetric-top molecules, the analysis of bands is similar to that which has been described for infrared spectra.

3.6 Magnetic Rotation Spectra

In this procedure, radiation from a continuous source, e.g., a high-pressure xenon arc, is passed through a polarizer and an absorption tube that is placed inside a solenoid. On emerging, it encounters a second polarizer, which is crossed with respect to the first one. No radiation emerges unless a molecule that produces a rotation of the plane of polarization in the presence of a magnetic field is present in the absorption tube. Weak triplet–singlet transitions can be detected in this way, whereas stronger singlet–singlet transitions are unaffected. An example of a magnetic rotation spectrum is shown in Fig. 26.

Plane-polarized light can be resolved into two circularly polarized components, one left-handed and the other right-handed. If the magnetic field introduces any optical activity, then the two components travel with different velocities when passing through the cell and, when recombined on emerging, give plane-polarized light in which the plane of polarization has been rotated. This process is called magnetic circular birefringence (MCB). If the absorption coefficients for left- and right-handed circularly polarized light are different, then the two components will emerge from the cell with different amplitudes and, on recombination, will produce elliptically polarized light. This process is called magnetic circular dichroism (MCD). The combination of MCB and MCD produces the magnetic rotation spectrum.

For theoretical interpretation, it is preferable to measure either the MCB or the MCD. The latter is experimentally simpler. Before entering the cell, the plane-polarized light is passed through a quarter-wave plate with its principal axis arranged at +45° to the plane of polarization. This produces left- or right-handed circularly polarized light. By means of a photoelastic modulator, the polarization can be modulated from left- to right-handed circular polarization at frequencies of ~50 kHz. Any magnetic circular dichroism is then detected as a modulation in the intensity of

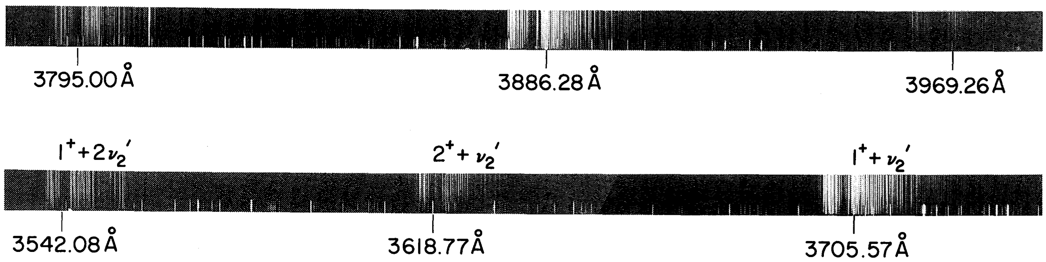

FIG. 26. Magnetic rotation spectrum of the $\tilde{a}\ ^3A_2$–$\tilde{X}\ ^1A_1$ system of formaldehyde (from Brown *et al.*, 1976, by courtesy of the authors and the journal).

the beam emerging from the cell using a detector and a lock-in amplifier.

3.7 Dissociation and Predissociation

For some molecules, vibration–rotation structure has been observed right up to a dissociation limit. If the states of the atoms at the limit are known, then an accurate value for the dissociation energy of the molecule is obtained. As an example, the $B\ {}^3\Sigma_u^- \rightarrow X\ {}^3\Sigma_g^-$ bands of oxygen, known as the Schumann–Runge bands, have been observed right up to a limit at 1750 Å. The oxygen atoms produced are in the atomic states $O({}^3P)$ + $O({}^1D)$. Subtracting the excitation energy of the $O({}^1D)$ atom, a value for the dissociation energy of the molecule in the ground state is obtained, viz., 41 260 ± 15 cm^{-1}.

More frequently, a phenomenon known as predissociation is observed in a spectrum. In absorption, the lines become broad or diffuse; in emission, a series of lines may suddenly terminate at a certain J value. The reason is that the potential curve for the excited state is crossed by a repulsive potential curve, and the molecule may dissociate along this curve rather than continue to vibrate and rotate in the initial excited state. The energy at which this phenomenon occurs gives an upper limit for the dissociation energy of the molecule.

3.8 Linewidths and Lifetimes

The relation between the width of a line and the lifetime of the excited state is given by the uncertainty relation,

$$\Delta E \Delta t \geqslant h/2\pi. \qquad (96)$$

Since $\Delta E = hc\Delta\nu$, where $\Delta\nu$ is the width of a level in cm^{-1}, then

$$\Delta\nu \geqslant 1/2\pi c\Delta t. \qquad (97)$$

Hence, the greater the mean life of a state, the smaller is its width. For strongly allowed electronic transitions, the lifetime of the excited state is of the order of 10–20 ns, and the corresponding broadening is of the order of 0.001 cm^{-1}. This is considerably less than the widths of spectral lines caused by the motions of the molecules (Doppler widths), which in the ultraviolet may be of the order of 0.05 cm^{-1} at room temperature. When an excited state is crossed by a repulsive curve, predissociation may occur during the period of one vibration, i.e., in a time of the order of 10^{-13}–10^{-14} s, for which the corresponding broadening is of the order of 50–500 cm^{-1}. Not all predissociation processes proceed this rapidly, but, even if predissociation occurs in a time of 10^{-10} s, the contribution to the line width will be 0.05 cm^{-1}, which is comparable with the Doppler width.

In larger molecules, the lifetime of an excited state may not be determined by predissociation but by rapid conversion from an excited-state level to a high vibrational level of the ground state (*internal conversion*), or to a level of a triplet state (*intersystem crossover*). Azulene provides a classic example of internal conversion, a process that has been studied using picosecond spectroscopy (Rentzepis and Struve, 1976). Formaldehyde provides another example for which fluorescence quantum yields have been measured and found to be less than unity. However, in this case, there is competition between internal conversion and predissociation (Moore, 1983).

3.9 Molecular Orbital Theory

Molecules are composed of atoms in which the electrons can occupy various orbits, 1*s*, 2*s*, 2*p*, 3*s*, 3*p*, 3*d*, 4*s*, and so on. In molecules, the electrons also occupy various orbits, called orbitals, which in simple molecular orbital theory are taken as linear combinations of the atomic orbitals. By solving the Schrödinger wave equation using computers, the energy levels of molecules are obtained (see ELECTRONIC STRUCTURE OF ATOMS AND MOLECULES). It is found that the electrons occupy certain shells as in the case of atoms. The outermost of these shells are sometimes called valence shells, and the electrons occupying them valence electrons.

The electronic structure of the ground state of a molecule is obtained by filling the orbitals with the available electrons, starting with the orbital of lowest energy. Frequently, a valence shell will not be completely filled in this manner, and excited states can be generated by promoting one or more of the valence electrons to vacancies in the valence shell. Such excited states lie usually in the range 0–5 eV, and spectra involving these states lie in the near-infrared, visible, or near-ultraviolet region. If the ground state of a molecule involves a complete valence shell, then the first excited state of the molecule will involve the excitation of an electron to a shell with a higher value of the principal quantum number n. Such states are called Rydberg states and usually occur in the vacuum-ultraviolet region of the spectrum below 2000 Å.

3.10 Rydberg States and Ionization Potentials

The simplest example of Rydberg states is provided by the hydrogen atom. The term values are given by the formula

$$T_n = I - R/n^2, \tag{98}$$

where R is the Rydberg constant (109 677.58 cm^{-1}) and is equal to the ionization potential (I = 13.59840 eV); n is the principal quantum number.

Molecules also have Rydberg states, and the term values can be represented to a good approximation by a formula

$$T_n = I - R/(n - \mu)^2, \tag{99}$$

where I is the ionization potential of the molecule and μ is known as the quantum defect. The value of μ gives some information on the nature of the Rydberg electron.

Rydberg states may be studied in absorption, and the transitions lie mainly in the vacuum ultraviolet, in the region between 2000 and 500 Å. Frequently, series are observed to the first and higher ionization limits and give values for the ionization potentials. Alternatively, transitions between Rydberg states may be studied in emission at longer wavelengths. Rydberg spectra frequently display vibrational and sometimes rotational structure; hence, to obtain accurate values for ionization potentials, it is necessary to cool the molecules in a supersonic jet.

An example of a Rydberg series is shown in Fig. 27. Members can be seen up to n = 30, and the series converges to an ionization limit, which can be determined to better than 0.1 cm^{-1}.

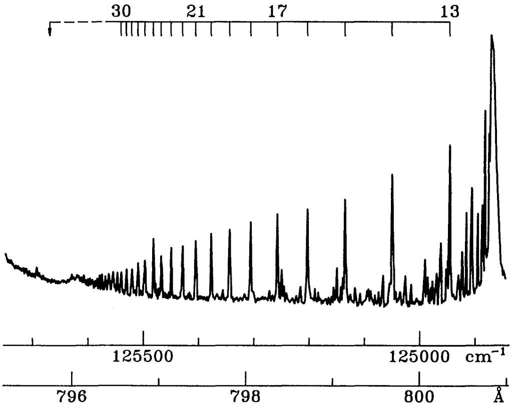

FIG. 27. Rydberg series for N_2 (courtesy of K. P. Huber, National Research Council of Canada). The sample was cooled in a supersonic jet to a temperature of ~20 K. The principal quantum numbers n are indicated.

3.11 Photoelectron Spectra and ZEKE Spectra

An alternative method of studying ionization limits of molecules is by the use of photoelectron spectroscopy. Radiation from a helium lamp at a wavelength of 584 Å has sufficient energy (21.2168 eV) to ionize all molecules, producing ions in their ground states and in excited states. The energy spectrum of the electrons emitted is measured and gives information complementary to that obtained from an optical spectrum and also gives a determination of the various ionization limits (Turner *et al.*, 1970).

A photoelectron spectrum of nitrogen is shown in Fig. 28. The peak labeled 15.57(9) corresponds to the ionization limit shown in Fig. 27. The peaks labeled 16.69(1) and 18.75(9) correspond to the second and third ionization limits, respectively. Vibrational structure is seen in all the transitions and is most pronounced for transitions to the second ionization limit. One of the limitations of this type of spectroscopy lies in the accuracy with which the electron energies can be measured, which is ~0.01 eV (~80 cm^{-1}).

A more recent development uses tunable lasers and measures only electrons with zero kinetic energy; hence the acronym ZEKE (zero electron kinetic energy) spectroscopy. The experimental technique uses zero-field ionization and delayed pulsed-field extraction. However, care must be taken to correct for

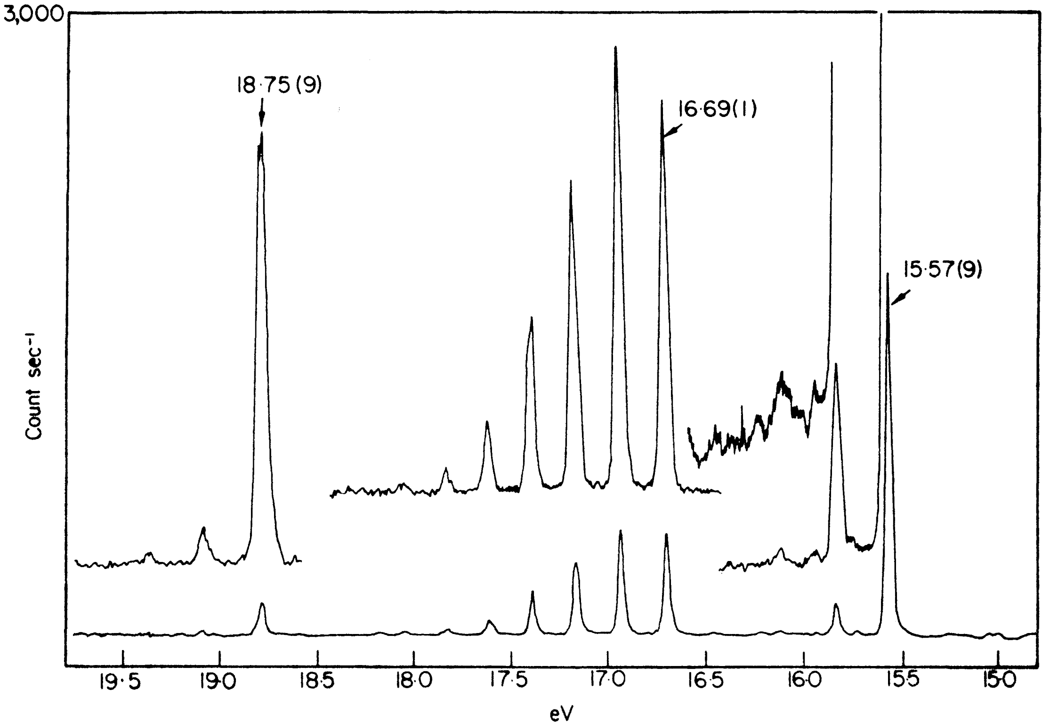

FIG. 28. Photoelectron spectrum of N_2 (from Turner *et al.*, 1970, by courtesy of the authors and publisher).

field ionization of high Rydberg states. The technique employs two-photon and multiphoton excitation, often involving near-resonant intermediate states (see Sec. 3.14.3). The technique uses supersonic jets to produce molecules at very low temperatures and is capable of giving ionization limits with an accuracy of ~0.5 cm^{-1} or better. Vibrational structure is readily seen, and, for some molecules, rotational structure has also been observed (Müller-Dethlefs and Schlag, 1991).

3.12 X-ray Spectroscopy (ESCA)

The most tightly bound orbitals in a molecule are those derived from 1*s* electrons. It is usually assumed that these are almost pure atomic orbitals. However, they are influenced slightly by molecule formation and are studied by means of x-ray spectroscopy. The shifts of lines from their positions in atoms give some information on the electronic structure of the molecule.

Most of the work has been carried out using the Mg *Kα* line as the source of excitation. The energy spectrum of the electrons emitted is measured, and linewidths of ~0.80 eV have been achieved. This corresponds almost entirely to the inherent width of the exciting radiation. The technique is very similar to photoelectron spectroscopy and was initially called x-ray photoelectron spectroscopy (XPS). It is now usually referred to as ESCA (electron spectroscopy for chemical analysis; Siegbahn *et al.*, 1969). More recently, the technique has been extended to studying molecules adsorbed on surfaces (Stohr, 1992).

3.13 Negative-Ion Spectroscopy

A technique that measures electron affinities rather than ionization potentials is negative-ion spectroscopy. A beam of negative ions is crossed with a tunable laser (see Sec. 3.14), and the beam current is monitored. When the frequency of the laser reaches the limit at which the ion dissociates into a neutral molecule and an electron, there is a sharp decrease in the beam current. Many electron affinities have been measured in this way, and, from the structure observed in the spectrum,

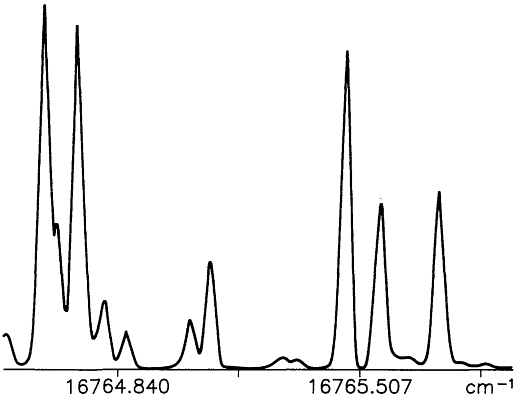

FIG. 29. Laser-excited fluorescence spectrum of thioformaldehyde, H_2CS. The spectrum covers a region of approximately 1 cm^{-1} at 16 765 cm^{-1} (courtesy of M. Barnett, National Research Council of Canada).

some deductions can be made concerning the geometry of the negative ion (Ervin and Lineberger, 1992).

3.14 Laser Spectroscopy

Many types of laser are now available. Fixed-frequency lasers, such as the argon-ion and krypton-ion lasers, can deliver powers up to 20 W at specific wavelengths. Tunable dye lasers can be operated continuously from 4000 to 8600 Å by a suitable choice of dye and optics and can yield linewidths of a few megahertz (approx. 0.0001 cm^{-1}) with powers up to 1 W in the case of rhodamine dyes. Pulsed dye lasers can be used with similar dyes giving pulses with a duration of a few nanoseconds and pulse energies of several millijoules. Frequency-doubling crystals can be used with both cw and pulsed dye lasers to produce radiation at shorter wavelengths. Other multiplication methods exist for producing tunable radiation in the vacuum ultraviolet down as far as 500 Å.

3.14.1 Laser-Excited Fluorescence Spectra One of the common applications of tunable dye lasers is to record fluorescence spectra of molecules and especially species that may be present in only small concentrations, e.g., free radicals. Since the linewidths of the laser are small compared to the Doppler widths of the lines being studied, there is no difficulty in recording high-resolution spectra. The relative intensities of the emission lines, however, may not accurately reflect the intensities of the lines in absorption, since fluorescence efficiencies vary from level to level in the excited state. The spectra obtained in this way are called laser-induced fluorescence (LIF) spectra or laser-excited fluorescence (LEF) spectra. A typical laser-excited fluorescence spectrum is shown in Fig. 29.

3.14.2 Saturated Absorption Spectra One of the great advantages of lasers lies in the narrow linewidths, providing an opportunity for carrying out sub-Doppler spectroscopy. The increase in resolution compared with Doppler-limited spectroscopy may be as high as 1000. One technique, known as saturated absorption spectroscopy, may be studied using the apparatus shown in Fig. 30. Radiation from a laser is divided into two

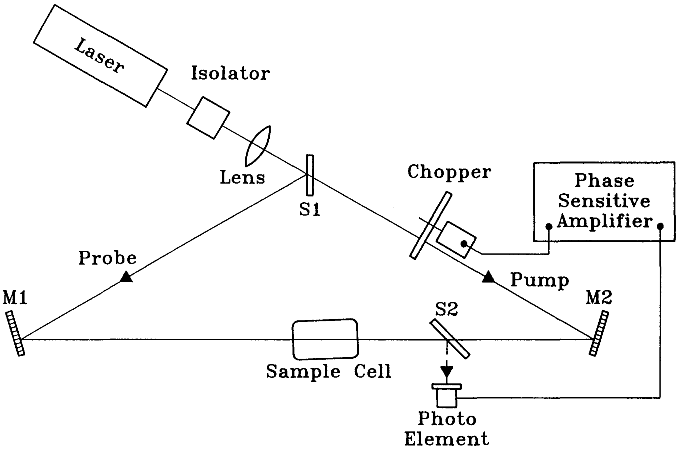

FIG. 30. Experimental arrangement for a saturated absorption experiment. The pump beam saturates the transition, and the probe beam detects the dip seen at the center of the absorption profile. The isolator prevents feedback into the laser.

beams, a stronger one known as the pump beam and a weaker one called the probe beam. The two beams pass through a sample cell in opposite directions. The pump beam excites molecules with a certain velocity v with respect to the beam, which corresponds to excitation on one side of the Doppler profile. The probe beam monitors molecules with the same velocity but in the reverse direction, which corresponds to absorption on the other side of the Doppler profile. When the frequency of the laser corresponds to the center frequency of the line, it interacts with molecules having a zero velocity with respect to the laser beam. On account of the high power of the laser, some of these molecules are transferred to an excited state by the pump beam, and the probe beam, finding a reduced population, records a reduced absorption. If the laser is scanned over the whole profile of the line, a Doppler line shape is obtained with a sharp reduction of the absorption at the line center. The latter is called an *inverse Lamb dip*, and gives an accurate measure of the line center. This technique was first used to observe the hyperfine structure of some of the iodine lines (Fig. 31). A similar technique, which monitors emission rather than absorption, is called *intermodulated fluorescence* (Sorem and Schawlow, 1972).

3.14.3 Two-Photon and Multiphoton Spectra With high-powered pulsed lasers, it is possible to observe transitions that involve the simultaneous absorption of two or more photons and probe excited states of molecules that cannot be accessed directly by one-photon spectroscopy.

The intensity I_{ik} for a two-photon transition between an initial state i and a final state k is given by

$$I_{ik} \propto \left| \sum_j \frac{\langle i|\mu|j\rangle\langle j|\mu|k\rangle}{\nu - \nu_{ij}} \right|^2, \tag{100}$$

where ν is the frequency of the photon and j represents an intermediate state. The intensity of absorption is greatly enhanced if the frequency of the photon is resonant or nearly resonant with a single-photon transition, and the latter is strongly allowed.

General two-photon selection rules are of the type

$$\Delta J = 0, \pm 1, \pm 2; \tag{101}$$

$$+ \leftrightarrow +, \qquad - \leftrightarrow -, \qquad + \not\leftrightarrow -; \tag{102}$$

$$g \leftrightarrow g, \qquad u \leftrightarrow u, \qquad g \not\leftrightarrow u. \tag{103}$$

In principle, it should be possible to access g electronically excited states from ground states with g symmetry, but, in practice, it is found that the transitions observed usually involve u vibrational levels of u electronically excited states, i.e., levels with g vibronic symmetry. Nevertheless, the spectroscopy is complementary to one-photon spectroscopy, which permits the study primarily of the totally symmetric levels in the excited state [see Eqs. (87) and (88)].

One method of detection of two-photon and

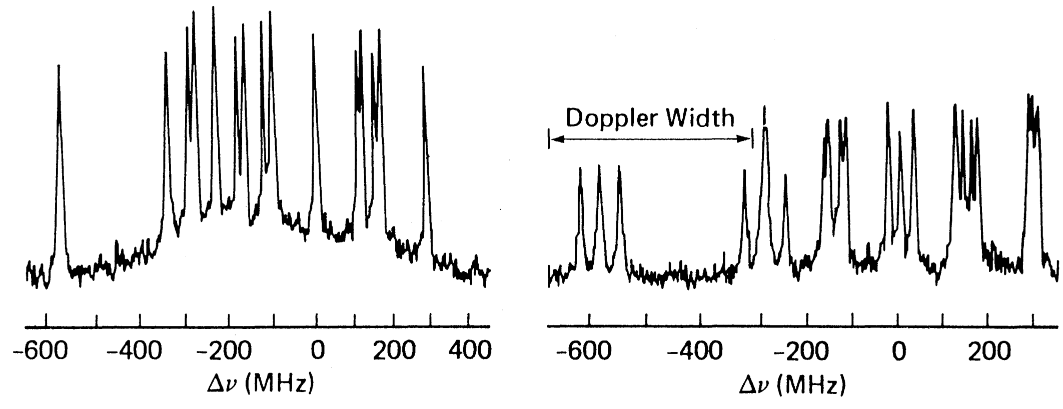

FIG. 31. Hyperfine structure of iodine, I_2, obtained by saturated absorption (from Levenson and Kano, 1988, by permission of the authors and publishers). The two lines show, respectively, 15 and 21 hyperfine components.

multiphoton spectra involves the measurement of fluorescence from the excited state. If the latter does not fluoresce, then an alternative and sensitive method involves the ionization of the molecule and the measurement of the ions. From energy considerations, this method is limited primarily to multiphoton spectra. The method may use one or, more frequently, two dye lasers. In the latter case, the process is described by the number of photons required to reach an excited state and the number of additional photons required to ionize the molecule, e.g., 3 + 1, 2 + 2, ... As with two-photon spectra, the probability of absorption is enhanced if the intermediate steps lie close to single-photon transitions, a process known as *resonance-enhanced multiphoton ionization* (REMPI).

Sub-Doppler two-photon spectra may be obtained if the experiment is carried out in an optical cavity. If a molecule is traveling with a velocity v with respect to the laser beam in one direction, it is traveling with a velocity $-v$ with respect to the reflected laser beam. The Doppler shifts cancel for two-photon absorption, and all molecules absorb two photons corresponding to the center frequency, not just those with zero velocity with respect to the laser beam. This results in greatly enhanced sensitivity (Levenson and Kano, 1988).

3.14.4 Nonlinear Laser Spectroscopy

With high-powered lasers, the strength of the electric vector E is sufficient to induce electric moments that depend not only on the first power of E but on higher powers E^2, E^3, ... These higher powers of E give rise to nonlinear optical phenomena (see OPTICS, NONLINEAR). Some of the processes that have been discussed earlier, e.g., difference-frequency laser, saturated absorption, and two-photon and multiphoton processes, are nonlinear phenomena. In addition, there are many other nonlinear optical phenomena, e.g., coherent anti-Stokes Raman spectroscopy (CARS), stimulated Raman gain spectroscopy (SRGS), Raman-induced Kerr-effect spectroscopy (RIKES), and hyper-Raman spectroscopy (Levenson and Kano, 1988).

As an example, let us consider CARS. If two lasers with frequencies ν_1 and ν_2 are incident on a sample such that the difference $\nu_1 - \nu_2$ is equal to the frequency of a Raman-active vibration of the molecule, then the scattered light contains a frequency equal to $\nu_1 + (\nu_1 - \nu_2) = 2\nu_1 - \nu_2$. By scanning one of the lasers, information on the Raman spectrum of the molecule is obtained.

3.15 Applications of Electronic Spectroscopy

The applications of electronic spectra in the laboratory are less widespread than those of infrared or other forms of spectroscopy such as NMR and ESR. Nevertheless, spectra in the near-ultraviolet region are frequently used for monitoring compounds with "chromophores," i.e., compounds with double or triple bonds such as carbonyl compounds, benzene and its derivatives, and heterocyclic compounds.

An important application of electronic spectroscopy lies in astronomy. The spectra of numerous stars have been recorded and the molecules present identified. Most of the species are diatomic, e.g., TiO and MgH, but a few polyatomic species have been found, e.g., SiC_2, C_3. The spectra of comets reveal the presence of several simple species such as CN, C_2, NH_2, and H_2O^+ that are produced by photodissociation of parent molecules under the influence of the Sun's radiation. Finally, processes in the Earth's atmosphere can be monitored by spectra observed in absorption or emission, e.g., the light of the night sky or aurorae.

4. DOUBLE-RESONANCE SPECTROSCOPY

Double-resonance spectroscopy involves the use of two sources, which may be selected from any region of the electromagnetic spectrum. The studies use klystrons or traveling-wave-tube oscillators in the microwave region and tunable lasers in the infrared or optical region. One of the aims of these studies is to assist the assignment of spectra. For example, in optical-optical double-resonance studies, one laser is used to pump a molecule to a level of an excited state for which all the quantum numbers are known—electronic, vibrational, and rotational. This level is usually monitored by means of the fluorescence it emits. The second laser is then scanned and resonances detected by a change in the fluorescence emitted. The second laser may excite the molecule to a higher excited state or stimulate emission to a lower state. In each case, the assignment of rotational quantum numbers is relatively straightforward.

Another application involves the study of collisional processes. Extensive microwave-microwave double-resonance studies showed that, when a molecule collides with another molecule, the change in the rotational quantum numbers is not random but shows a preference for the changes permitted by electric dipole selection rules (Oka, 1973). Similar results have been obtained from microwave–optical double-resonance studies (Petersen and Ramsay, 1985).

Works Cited

Brown, J. M., Buckingham, A. D., Ramsay, D. A. (1976), *Can. J. Phys.* **54**, 895–906.

Clouthier, D. J., Ramsay, D. A. (1983), *Annu. Rev. Phys. Chem.* **34**, 31–38.

Ervin, K. M., Lineberger, W. C. (1992), *Advances in Gas Phase Ion Chemistry*, Vol. 1, Greenwich, CT: JAI Press, pp. 121–166.

Evenson, K. M., Saykally, R. J., Jennings, D. A., Curl, R. F., Jr., Brown, J. M. (1980), *Chem. Biochem. Appl. Lasers* **5**, 95–138.

Heitler, W., Herzberg, G. (1929), *Naturwissenschaften* **17**, 673.

Herzberg, G. (1945), *Infrared and Raman Spectra of Polyatomic Molecules*, Princeton, NJ: Van Nostrand.

Herzberg, G., Innes, K. K. (1957), *Can. J. Phys.* **35**, 842–879.

Levenson, M. D., Kano, S. S. (1988), *Introduction to Nonlinear Laser Spectroscopy*, San Diego, CA: Academic Press.

Long, D. A. (1977), *Raman Spectroscopy*, New York: McGraw-Hill.

Marion, L., Ouellet, J. (1948), *J. Am. Chem. Soc.* **70**, 3076–3078.

Moore, C. B. (1983), *Annu. Rev. Phys. Chem.* **34**, 525–555.

Müller-Dethlefs, K., Schlag, E. W. (1991), *Annu. Rev. Phys. Chem.* **42**, 109–136.

Nicholls, R. W. (1962), *J. Quant. Spectrosc. Radiat. Transfer* **2**, 433–449.

Oka, T. (1973), *Adv. At. Mol. Phys.* **9**, 127–206.

Pauling, L., Wilson, E. B. (1935), *Introduction to Quantum Mechanics*, New York: McGraw-Hill.

Petersen, J. C., Ramsay, D. A. (1985), *Chem. Phys. Lett.* **118**, 34–37.

Rentzepis, P. M., Struve, W. S. (1976), in: D. A. Ramsay (Ed.), *MTP International Review of Science*, Physical Chemistry, Series 2, London: Butterworths, Vol. 3, pp. 263–300.

Sears, T. J., Bunker, P. R., McKellar, A. R. W. (1982), *J. Chem. Phys.* **77**, 5363–5369.

Semin, G. K., Babushkina, T. A., Yakobson, G. G. (1975), *Nuclear Quadrupole Resonance in Chemistry*, New York: Wiley.

Siegbahn, K., Nordling, C., Johansson, G., Hedman, J., Hedén, P. F., Hamrin, K., Gelius, U., Bergmark, T., Werme, L. O., Manne, R., Baer, Y. (1969), *ESCA Applied to Free Molecules*, Amsterdam: North-Holland.

Snyder, L. E. (1972), in: D. A. Ramsay (Ed.), *MTP International Review of Science*, Physical Chemistry, Series 1, London: Butterworths, Vol. 3, pp. 193–240.

Sorem, M. S., Schawlow, A. L. (1972), *Opt. Commun.* **5**, 148–151.

Stohr, J. (1992), *NEXAFS Spectroscopy*, Berlin: Springer-Verlag.

Stoicheff, B. P. (1950), thesis, University of Toronto.

Sutherland, G. B. B. M., Thompson, H. W. (1945), *Trans. Faraday Soc.* **41**, 174–179.

Townes, C. H., Schawlow, A. L. (1955), *Microwave Spectroscopy*, New York: McGraw-Hill.

Turner, D. W., Baker, C., Baker, A. D., Brundle, C. R. (1970), *Molecular Photoelectron Spectroscopy*, New York: Wiley-Interscience.

Further Reading

Bunker, P. R. (1979), *Molecular Symmetry and Spectroscopy*, New York: Academic.

Demtroder, W. (1981), *Laser Spectroscopy: Basic Concepts and Instrumentation*, Berlin: Springer-Verlag.

Gordy, W., Cook, R. L. (1984), *Microwave Molecular Spectra*, 3rd ed., New York: Wiley-Interscience.

Herzberg, G. (1945), *Infrared and Raman Spectra of Polyatomic Molecules*, Princeton, NJ: Van Nostrand.

Herzberg, G. (1950), *Spectra of Diatomic Molecules*, Princeton, NJ: Van Nostrand.

Herzberg, G. (1967), *Electronic Spectra of Polyatomic Molecules*, Princeton, NJ: Van Nostrand.

Herzberg, G. (1971), *The Spectra and Structures of Simple Free Radicals*, Ithaca, NY: Cornell University Press.

Hollas, J. M. (1982), *High Resolution Spectroscopy*, London: Butterworths.

Kroto, H. W. (1975), *Molecular Rotation Spectra*, New York: Wiley.

Papousek, D., Aliev, M. R. (1982), *Molecular Vibrational/Rotational Spectra*, Prague: Academia.

Wilson, E. B., Jr., Decius, J. C., Cross, P. C. (1955), *Molecular Vibrations*, New York: McGraw-Hill.

MOLECULES

Paul Engelking, *Department of Chemistry, University of Oregon, Eugene, Oregon, U.S.A.*

Introduction 525
1. **Nomenclature** 527
1.1 Names 527
1.2 Formulas 528
2. **Structure** 528
2.1 Bonding 530
2.1.1 Short-Range Forces 530
2.1.2 Long-Range Forces 534
2.2 Bond Angles and Lengths 535
2.3 Stereochemistry 536
2.3.1 Isomers and Conformers 536
2.3.2 Chirality 537
2.3.3 Symmetry 538
2.4 Extended Bonding Structures 539
2.4.1 Clusters 539
2.4.2 Excimers 540
2.4.3 Polymers 540
2.4.4 Crystals 541
2.5 Structure Determination 541
2.5.1 Theoretical 541
2.5.2 Experimental 542
2.5.2.1 Mass Spectrometry 542
2.5.2.2 Diffraction 542
2.5.2.3 Rotational Spectroscopy 542
2.5.2.4 Vibrational Spectroscopy 543
2.5.2.5 Magnetic Resonance 543
3. **Physical Properties** 543
3.1 Molecular Mass 544
3.2 Energy of Formation 545
3.3 Electric and Magnetic Constants 546
4. **Chemical Dynamics** 547
4.1 Separation of Nuclear and Electronic Motion 547
4.2 Conservation Rules 547
4.3 Theories of Reaction Rates 548
4.4 Reaction Steps 549
4.5 Mechanisms 549
4.5.1 Elimination 549
4.5.2 Addition 549
4.5.3 Substitution 549
4.5.4 Rearrangement 549
4.5.5 Catalysis 549
4.6 Processes 550
Glossary 550
Works Cited 551
Further Reading 552

INTRODUCTION

The study of molecules has been primarily the focus of chemistry. The subject of molecules is so broad, however, that chemistry needs considerable assistance from related areas such as materials science, solid state physics, and molecular biology.

Today, the concept of molecule adapts to encompass new structures, many only recently encountered in the laboratory. Polymer chemists now make "dynamic polymers" with unique physical properties that depend upon long chains of atoms continually forming, breaking, and re-forming networks of bonds. Physicists and physical chemists now produce "clusters" of atoms in free jet expansions that blur the distinctions between physical and chemical bonding forces. Electronic engineers align "liquid crystals," partially oriented fluids, in electric fields to produce changing, flat-panel video displays. Materials scientists construct "superlattices," containing repeating units of "nanostructure" aggregates of atoms, to give materials new physical properties. Laser spectroscopists study molecular "free radicals" and ions, so reactive that they are destroyed on contact with any container and can only be confined and manipulated by electromagnetic fields. Astronomers discover "interstellar molecules" in space, even before these molecules can be identified and studied in laboratories on earth. Biochemists manipulate and "splice" the molecules of genes, definable by replicating, even self-repairing, sequences of repeating nucleic acid units, to make new pharmaceuticals and even new organisms.

The modern concept of molecule entered

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

science in two ways: Chemists introduced the molecular concept to explain the definite combining proportions of elements in compounds; physicists introduced the molecular concept to explain the physical properties of gases. By the beginning of the 20th century, the two concepts had converged into the single unifying idea at the foundation of chemistry and the basis of our understanding of how atoms combine to form ordinary matter.

At the start of the 19th century, chemists such as Dalton began to infer the existence of elemental atoms and compound molecules from the definite combining ratios of constituent elements in chemical compounds. The concept of molecules originally explained the constant combining ratios of sodium and chlorine in table salt ($NaCl$), or of carbon and oxygen in carbon dioxide (CO_2): In the first case, each atom of sodium was imagined paired (incorrectly, as we now understand) with one atom of chlorine in individual molecules of sodium chloride; in the second case, two atoms of oxygen are attached to one atom of carbon in individual molecules of carbon dioxide. Today we have revised that image of common salt to explain the physical and chemical properties of sodium chloride, an ionic lattice solid, but the existence of individual triatomic molecules of carbon dioxide still dominates our understanding of this gas.

In the middle of the 19th century, physicists such as Maxwell introduced molecules into the kinetic theory of gases. Assemblages of hard, elastic, nearly pointlike molecules of matter explained many of the physical properties of gases such as carbon dioxide. Today a researcher using high-pressure carbon dioxide in a supercritical extraction would not explain the solvent properties of high-pressure carbon dioxide by considering carbon dioxide molecules as discrete, hard, pointlike particles, but the molecular concept underlies our understanding of even condensed phases such as liquids and solids.

Even while we currently understand that a crystalline material such as sodium chloride is more properly considered to be a lattice of Na^+ and Cl^- ions, and a solution of carbon dioxide and water contains solvated H^+, HCO_3^-, CO_3^-, and OH^- ions and H_2CO_3 molecules, the concept of molecule still would be at the base of that more complex explanation and understanding. There is no more firmly established and useful concept of the chemist or materials scientist than that of molecule.

Nevertheless, the definition given the term molecule has changed and continues to change.

The strong influence of men such as Ostwald (1900, 1922), a firm believer in a continuum theory of matter, led chemists to define molecules traditionally as a limit, in a way reminiscent of taking derivatives in calculus: A "molecule" is the smallest amount of a substance that has all the properties of that substance. Generations of students have found that definition of molecule inadequate, even as they have learned it.

We will simply define *molecules* broadly as *collections of atoms held together by bonding forces for a sufficient time that the collection can be considered as having a fixed composition*. Even so defined, the line between strong *intra*molecular chemical bonding forces and weak *inter*molecular physical interaction forces is a matter of convention and usage; the requisite time is a matter of context.

This definition of molecule is not unique, and several important variations occur. For example, while strictly speaking a molecule that becomes charged is changed to an *ion*, it is possible to think of degree of charge as an attribute and widen the definition of molecule to include charged molecules. For example, Herzberg's classic books (1945, 1950, 1966) on the spectroscopy of molecules treat ions on the same footing as molecules. Likewise, if one were to quibble with a molecular biologist, insisting that the field should more properly be named "ionic and molecular" biology because most of the molecules that are important to this field are actually ions, one would be deserving of a stern look. However, just as often, the term molecule is used specifically to mean "neutral molecule" as in the term "molecular crystal," denoting an array of neutral molecules (such as in "dry ice," solid CO_2), as opposed to "ionic crystal," denoting an array of charged ions (such as "salt," solid $NaCl$).

Molecule sometimes is restricted to denote more than one atom; other times, the definition is extended, following Maxwell, to include single atoms, such as the "molecules" of monatomic, inert, rare-gas elements He, Ne, Ar, Kr, Xe, Rn.

Different spatial arrangements of atoms are sometimes considered different molecules when their contrasts are emphasized; different spatial arrangements of atoms are sometimes considered different conformations of the same molecule when their similarities are emphasized.

Finally, while the word molecule derives from the diminutive, *molecula*, of the Latin term for mass, there is no hesitancy in using the term *macromolecule* to indicate a large number of atoms, as in a polymer or a protein. Nor do authors consider molecules as necessarily being invisibly small: Many make statements that often whole crystals, such as those of diamond containing atoms of carbon chemically bonded in a rigid network, are properly macromolecules.

As history has shown, attempts to define "molecule" uniquely must be taken with a grain of salt! In what follows, we will concentrate on a concept that derives its strength from its plasticity, its permanence from its adaptability.

1. NOMENCLATURE

The language of chemistry can confound physicists, biologists, engineers, and chemists. Knowledge of a few common names of familiar molecules and an acquaintance with the conventions of naming new molecules and compounds will help negotiating in the molecular realm, just as foreign travelers are aided by knowledge of a few indigenous words or phrases.

1.1 Names

Every molecule has at least one systematic name and often an array of common names or acronyms. It also has a sequence of chemical formulas, each of increasing detail in specifying the spatial relationships of atoms to one another.

For example, the organic molecule (formula CH_3COOH) with the *proper* IUPAC (International Union of Pure and Applied Chemists) name "ethanoic acid" has the *common* name "acetic acid." Dilute solutions of this compound with water (the common name for the hydrogen oxide molecule with the formula H_2O) are commonly known to us as "vinegar." The chemical formula may be abbreviated by an organic chemist by an *acronym* "HAc." Any, or all, of these names, formulas, or abbreviations may occur in one paper. As molecules become more exotic or more uncommon—with longer proper names—they are less likely to have "common" names and more prone to designation by acronyms.

There are two major systematic naming systems: organic and inorganic. The nomenclature of both fields is established and continuously revised by committees of the International Union of Pure and Applied Chemists. The inorganic rules are published in *Nomenclature of Organic Chemistry* (IUPAC, 1979), and the inorganic rules have been recently revised in *Nomenclature of Inorganic Chemistry, Recommendations 1990* (IUPAC, 1990). Synopses of these naming rules are contained in the *Handbook of Chemistry and Physics* (Lide, 1993).

In addition to the IUPAC nomenclature, all compounds are given a CAS name and number by the Chemical Abstracting Service. Knowing the name under which a particular compound is indexed by the editors of *Chemical Abstracts* helps locate information on that compound elsewhere. For example, Material Safety Data Sheets (MSDS's) of commercial compounds are indexed under their CAS numbers and names.

Two major compilations of data about compounds exist in the chemical encyclopedias known by their original editor's names Beilstein and Gmelin. Information on virtually any organic compound is contained in Beilstein (1918, *et seq.*) A peculiar numbering and indexing system is used here, and the novice is advised to enlist the aid of a chemical librarian when approaching this collection. A slightly more friendly indexing system is used for inorganic and organometallic compounds by Gmelin (1922, *et seq.*). These two encyclopedias offer invaluable compilations of the literature of most compounds, especially the older literature, which is not covered by the newer computerized data bases.

The easiest way to learn a new language is perhaps a crash conversational course. In addition to reading standard organic and inorganic texts, a student may make use of programed texts on nomenclature (Benfry, 1966; Banks, 1967; Tranham, 1980; Block, 1990).

Chemical etiquette would suggest that authors and speakers should not assume that their audiences know the names or formulas of any novel compounds they are discussing.

Chemical formulas, or structural drawings, should accompany the first mention of any novel compound and should be made available whenever an unfamiliar or novel compound is referred to by name.

1.2 Formulas

Molecules are given three types of chemical formulas: *empirical, molecular*, and *structural.*

An *empirical formula* lists the symbols of elements present; subscripts on each indicate the proportions in which each element's atoms are present in the molecule, usually expressed as smallest integers. For example, benzene, which consists of a hexagonal arrangement of six carbon and six hydrogen atoms, may be given the empirical formula CH (by convention the subscript "1" is dropped), while ethene, a compound of two carbon atoms and four hydrogen atoms, could be given the empirical formula CH_2.

A *molecular formula* indicates the exact numbers of atoms within each molecule by subscripts. The molecular formula of benzene is C_6H_6 to represent molecules each containing six carbon atoms and six hydrogen atoms. Ethene is represented as C_2H_4 in a molecular formula.

A *structural formula* indicates to greater or lesser degree the spatial relationship of atoms. Bonds are usually indicated. For example, benzene may be represented in one of the following structural formulas:

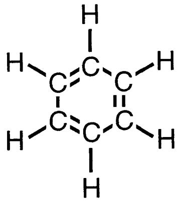
,
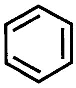
,
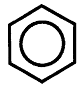
;

ethene by any one of these:

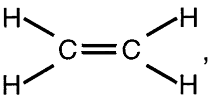
, $CH_2{=}CH_2$, =.

In the last form in each case, only the bonds between carbon atoms—the carbon *skeletal structure*—are shown. Hydrogen atoms are implied, and their number at each carbon can be inferred from the rule that each carbon has four bonds.

Groups of atoms within molecules are termed *radicals*. Separated from stable molecules, "free" radicals possess high reactivity and readiness to recombine.

If the molecule has a net charge, it is formally an *ion*, and the charge in multiples of the fundamental charge *e* is denoted by a signed integer, usually in the upper right-hand corner of its empirical or molecular formula, as in NO_3^-, SO_4^{-2}, and NH_4^+.

Sometimes it is necessary to indicate the arrangement of atoms three dimensionally with either specialized structural notation or a structural drawing. For example, while $CH_2{=}CH_2$ suffices to indicate that two hydrogens are attached to each of two carbon atoms in ethene, two structural drawings are needed to distinguish between the different *geometrical isomers cis*-1-2-dichloroethene and *trans*-1-2-dichloroethene:

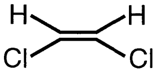
,
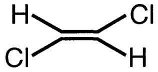
.

2. STRUCTURE

The structure of molecules is determined by the electrostatic force between charged electrons and nuclei moving quantum mechanically. The distribution of the electrons results from a delicate interplay of kinetic and potential effects: If the electrons are too distant from the nuclei, the potential energy is too large, while if the electrons are confined too tightly to the nuclei, the kinetic energy is excessive.

The overall energy operator for molecules is given by (Kotani, *et al.*, 1961; Parr 1964; Platt, 1961)

$$H = T_n + T_e + V_{nn} + V_{ne} + V_{ee} + H_{\mathrm{rel+mag}}. \tag{1}$$

Here T_e and T_n represent the kinetic energies of the electrons and nuclei, while V_{ee} and V_{nn} represent electrostatic repulsive potential energies between electrons and nuclei, respectively, and V_{ne} represents the electrostatic attractive potential energies between electrons and nuclei. Details of these terms may be found in ELECTRONIC STRUCTURE OF ATOMS AND MOLECULES.

The last term $H_{\mathrm{rel+mag}}$ represents relativistic corrections and magnetic interactions. This term is usually too small to determine the bonding within molecules made of elements from the top half of the periodic table and is often considered as a correction. (Richards *et al.*, 1981) For example, the interaction of electron spin and orbital motion in ground state OH is only about 8.7 meV, which is much smaller than the net bonding energy between the hydrogen and oxygen atoms of 4.392 eV (Huber and Herzberg, 1979).

Born and Oppenheimer (1927) originally removed nuclear motion from electronic motion in a perturbation expansion for the energy that begins with nuclear and electronic coordinates completely separated. That is, they separated the wave function into the product

$$\psi_N(R_N) \cdot \psi_e(r_e), \tag{2}$$

where R_N and r_e represent nuclear and electronic coordinates. In this treatment, corrections to the wave function enter as powers of

$$(m_e/m_N)^{1/4} \approx 0.1, \tag{3}$$

where m_e/m_N is the ratio of electron to nuclear mass. The relatively rapid convergence of this perturbation series is taken as justification of approximate separability of electronic and nuclear motion; hence, all subsequent separations of nuclear and electronic motions are termed *Born-Oppenheimer approximations*.

Our usual understanding of bonding is based upon one of these Born-Oppenheimer approximations, the *adiabatic approximation* (Born, 1931; Born and Huang, 1951). Here nuclear motion is first neglected while solutions to the quantum mechanical energy problem for only the electrons are obtained (or assumed) *at each, fixed, nuclear configuration*. This approximation starts with a parametrically separated wave function

$$\psi_N(R_N) \cdot \psi_e(r_e;R_N), \tag{4}$$

in which the electronic part $\psi_e(r_e;R_N)$ depends upon the nuclear coordinates only as slowly varying parameters. Only T_e, V_{ee}, and V_{ne} for each fixed configuration of nuclei are involved in the quantum mechanical calculation on the electrons. When the solution $\psi_N(R_N)$ to the motion of nuclei is subsequently obtained, the kinetic energy of the nuclei T_n is usually small (on the order of 10–100 meV per atom near the equilibrium geometry of typical molecules). Wave-function corrections appear in the adiabatic approximation on the order of $(m_e/m_N)^{1/2}$, an order of magnitude smaller than in the original Born-Oppenheimer scheme.

The nuclear-nuclear electrostatic interaction V_{nn} may be considered as just a classical addition to the electronic energy obtained in the adiabatic approximation. This *adiabatic potential* energy can be represented as a function of the nuclear positions; this is the function often represented in figures of molecular energy vs bond length—for example, Fig. 1. Minima on an effective, adiabatic potential energy surface are expected to correspond to experimental molecular structures.

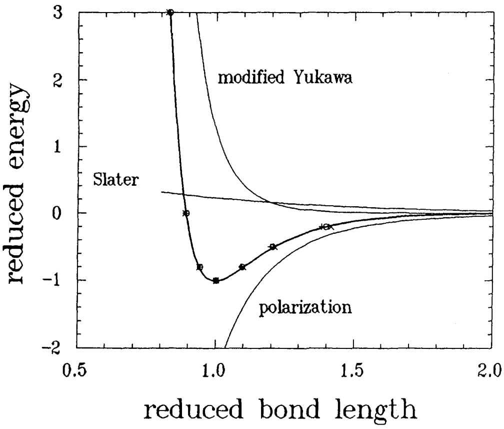

FIG. 1. Adiabatic potential energy for rare-gas atom pairs. The shape is typical of atomic interactions. Reduced units scale bond lengths and energies for each pair: Bond lengths are divided by the bond length at the minimum of the potential; energies are divided by the value of the bond energy at the minimum of the potential. For each pair, dissociation occurs at 0 energy; the minimum of the potential occurs at (1,−1) in reduced units. The curve that fits argon data includes a short-range, modified Yukawa-type repulsion term, two Slater-type terms expressing overlap of electronic functions (one attractive and the other repulsive), and a long-range r^{-6} polarization interaction; these are plotted to show their individual contributions. In addition to the experimental points for argon (o), values for helium (x) and neon (+) are plotted in their own reduced units. Except in the outer wall region, all rare-gas potentials are similar to within 1%. The potential is deep enough to form diatomic "molecules" for all but the lightest rare gas, helium. (Data compiled by Scoles, 1980.)

2.1 Bonding

Bond energy is the difference between the adiabatic potential energy corresponding to separated atoms and the adiabatic potential energy of these same atoms when combined and in their minimum energy conformation. Bond energy can be roughly factored into derived bonding "forces" that dominate at different internuclear distances.

2.1.1 Short-Range Forces The overlap of atomic electronic wave functions on different atoms gives rise to short-range forces, which can be either attractive or repulsive.

The electronic density in atoms decreases approximately exponentially with distance from the nucleus. For the outer electrons at large distances r from a charged center, the decreasing electron density becomes dominated by Slater-type terms of the form $r^x \times \exp(-ar)$ (a is a range parameter, x usually a positive integer), such as are prevalent in hydrogen-atom wave functions. Again, near the nuclei where nuclear attraction dominates the mutual electronic repulsion, the electronic density is also dominated by Slater-type terms (see ATOMS).

To lowest order, if the electronic distributions of the individual atoms are not permitted to relax, adjusting in the presence of nearby atoms, the short-range force is mostly repulsive, dominated by modified Yukawa-type effective potential terms $r^{-y} \exp(-ar)$, resulting from the r^{-1} Coulomb repulsion between nuclei as they penetrate the Slater-type $r^x \exp(-ar)$ charge shells of adjacent atoms.

When the electron distributions are allowed to relax under the influence of adjacent nuclei and electrons, the interaction may be either attractive or repulsive.

Emphasis on the diffuseness of the electronic wave function leads to the picture of an electron fluid under forces. A collection of charges in equilibrium possesses a total kinetic energy which is half of the total (net attractive) potential energy, according to the quantum virial theorem. This balance between kinetic and potential energy is not uniform throughout molecules, however. Comparison of regions of space inside molecules shows a relatively greater amount of kinetic energy for electrons near the nuclei; this condition reverses itself for distant electrons. As a result, the electrons near the nuclei act as a fluid in compression, while the electrons distant from the nuclei, near the "surface" of the molecule, act as a fluid in tension, with "surface tension" holding the molecule together.

The forces between all atoms become repulsive at very short internuclear distances, when the region between nuclei is too small for electrons to accumulate there without acquiring enormous kinetic energies; at very short range, the forces between nuclei correspond to their mutual Coulomb repulsion. This repulsion produces an inner "wall" on the adiabatic potential that is often modeled by a steep r^{-9} or r^{-12} function that diverges at the origin, or by a modified Yukawa-type $r^{-y} \exp(-ar)$ leading term, as in the case of unrelaxed, atomic electron distributions.

If the electron density increases in a region "between" nuclei, attraction results (Hirshfelder *et al.* 1954; Deb, 1973; Bamzai and Deb, 1981). Figure 2 shows the region between pairs of nuclei in which electron buildup leads to bonding.

In order for the atomic electron distributions to relax toward the intranuclear bonding region, the atomic wave functions must have sufficient flexibility and must not act as rigid shells, which only leads to repulsion. This flexibility is present when "open shell" atoms, having degenerate (or nearly degenerate) electronic wave functions, approach one another.

Emphasis on the discreteness of electrons leads to the chemist's picture of bonds. Chemical bonds are associated with pairs of electrons possessing similar spatial states, but of opposite spins. Because electrons can have only two spin states, only two electrons in the same spatial state can take advantage of the attractive region of low potential energy between nuclear charges without violating the Pauli exclusion principle. Additional electrons are excluded from having similar space and momentum coordinates; these additional electrons will anti-correlate their motion with that of a bonding pair.

Chemists have discovered empirical rules for determining the likelihood of the formation of a bond between two atoms. Their most important tool, *Lewis structures*, relies upon the observation that the rare-gas elements possess very little tendency to react chemically; these elements exist commonly as monatomic molecules, with closed *s*- and *p*-orbital shells having eight electrons. (Helium

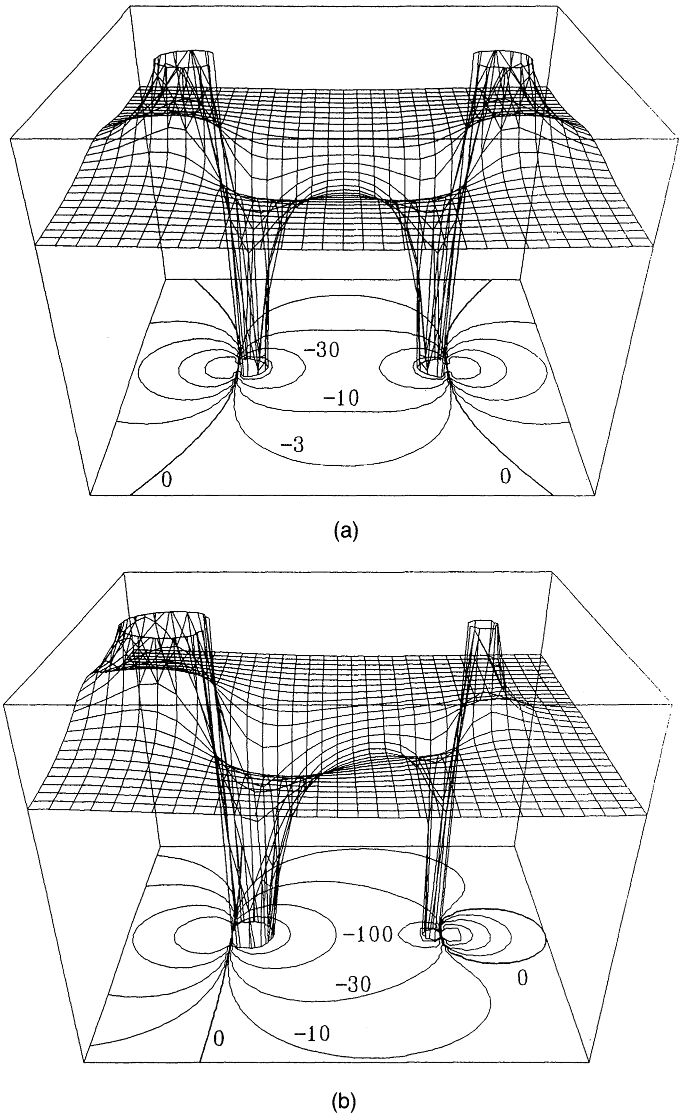

(a)

(b)

FIG. 2. Regions of bonding electron density between two atoms, shown in both contour and surface-grid plot representations. The component of bonding force between two nuclei, which is produced by an element of negative charge located variously in a plane containing the axis between nuclei, is plotted in arbitrary units for two cases: (a) homonuclear charge case showing the attractive (negative) region between two hyperbolas passing through the nuclei; (b) heteronuclear charge case corresponding to a 1:9 charge ratio, such as in HF. In the last case, the repulsive (positive) region behind the smaller charge has shrunk to a small oval, while the attractive, bonding region has expanded to nearly anywhere to the right of the larger charge. The surface plots are cut off at arbitrary positive and negative force values to aid representation; the actual force magnitudes diverge at the nuclei.

is the exception, with only two *s*-type electrons, without any *p*-type electrons.) Lewis's octet rule just generalizes this observation, predicting that stable molecules generally will have each atom surrounded by eight electrons that simulate the closure of the *s* and *p* electron shells in the rare gas atoms. To keep track of the *s*- and *p*-type electrons, Lewis invented the "electron dot" diagrams that receive his name. Basically, a dot is assigned to each *s* or *p* electron in the atoms to be combined, and then atoms are allowed to share pairs of dots between them until each atom is surrounded by eight electrons. (The major exception, hydrogen, only requires two electrons to resemble the rare gas helium, however.) For example, the Lewis structures for H, C, and O atoms are

$$\mathrm{H}\cdot,\quad \cdot\dot{\underset{\cdot}{\mathrm{C}}}\cdot,\quad :\ddot{\underset{\cdot\cdot}{\mathrm{O}}}\;.$$

The resulting Lewis structures for hydrogen, oxygen, water, and carbon dioxide molecules satisfy the octet rule by sharing electrons between atoms:

$$\mathrm{H}{:}\mathrm{H},\quad \ddot{\underset{\cdot\cdot}{\mathrm{O}}}{::}\ddot{\underset{\cdot\cdot}{\mathrm{O}}},\quad \mathrm{H}{:}\ddot{\underset{\cdot\cdot}{\mathrm{O}}}{:}\mathrm{H},\quad \ddot{\underset{\cdot\cdot}{\mathrm{O}}}{::}\mathrm{C}{::}\ddot{\underset{\cdot\cdot}{\mathrm{O}}}\;.$$

Atoms possessing partially filled *d* orbitals, such as the transition metals, or atoms possessing low-energy, empty *d* orbitals, such as sulfur or phosphorus, often violate the octet rule, but chemists have found ways of extending simple rules for electron counting to predict the likelihood of bond formation.

Atoms generally can aggregate until the resulting molecules become "closed shell," no longer electronically degenerate, at which point additional atoms no longer readily add. The exceptions to this rule usually involve atoms with partially filled *d* or *f* orbitals. Many compounds containing transition-metal, lanthanide, and actinide elements have closed *s*- and *p*-orbital shells, while the deeply buried *d*- and *f*-type orbitals only marginally participate in the bonding and remain partially open.

If the increase in bonding electron density between two nuclei comes from both atoms, and no net charge transfer occurs as evidenced by a lack of dipole moment, then the bond is termed *covalent*. If the increase in bonding electron density between two nuclei is accomplished by transfer of electrons from one atom to another, the bond is termed *ionic*. A third type, the *coordinate covalent bond* (or "Lewis acid-base" or "dative" bond), results when one atom contributes a pair of electrons to the internuclear region. Most bonds between dissimilar atoms are mixtures of these pure cases, but tending to one extreme or another. The amount of ionic character of a bond significantly influences chemical and physical properties. Covalently bonded molecules tend to remain whole when they dissolve in solutions and when they form crystals. Ionically bonded compounds typically dissociate as ions when they dissolve, and form large lattices of alternating ions when they crystalize. Typically, covalent bonding dominates organic chemistry, ionic bonding dominates inorganic chemistry.

Chemists empirically deduced the existence of multiple bonds between pairs of atoms long before multiple bonds were explained quantum mechanically. The Pauli principle requires that each bonding pair of electrons in a molecule must differ spatially from all other electron pairs. Electrons distinguish their orbitals spatially by adding extra nodes. Since a node perpendicular to the internuclear axis would exclude electrons from a large part of the attractive region between nuclei, just where electron density is needed to increase bonding, the most bond-favorable way of including this node is along the internuclear axis. Chemists denote the existence of these extra nodal surfaces along internuclear axes by the greek letters σ, π, δ, ϕ, . . ., reminiscent of the atomic series *s*, *p*, *d*, *f*, . . . to denote 0, 1, 2, 3, . . . angular nodal surfaces in atoms (see Atoms). This notation is fitting, since it is unlikely for nodes along a bond axis to arise unless the bonding atoms originally have atomic orbitals with the correct number of nodes. Thus, atoms with unpaired electrons in *s* orbitals may form σ bonds, atoms with partially filled *p* orbitals may form either σ bonds or π bonds, while electrons in *d* atomic orbitals may form σ, π, or δ bonds, as shown in Fig. 3. The requirement that nodes along bond axes arise

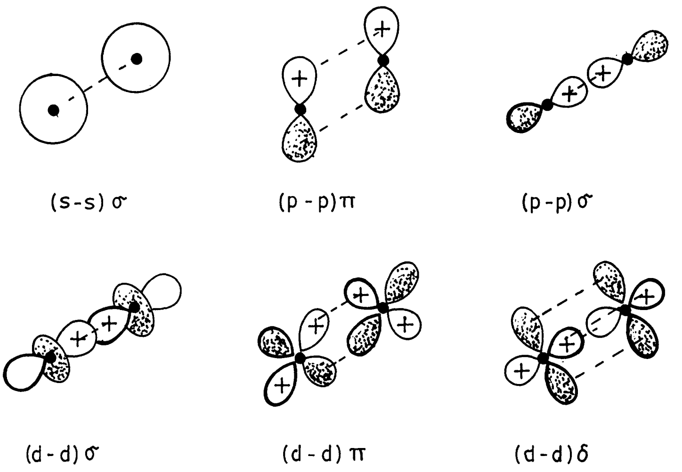

FIG. 3. The σ, π, and δ bond orbitals formed from *s*, *p*, and *d* atomic orbitals.

from the nodes originally present in the bonding atoms explains why hydrogen, with only a partially filled *s* orbital, typically bonds with single σ bonds, while oxygen, with two vacancies in a *p*-orbital shell, may form double bonds consisting of one σ and one π bond. It also explains the rarity of "quadruple" bonds between main group elements not having access to partially occupied *d*-electron shells.

For the most part, the picture of electron pairs localized in bonds between atoms gives a qualitatively correct picture of bonding. Nevertheless, electrons, being described by quantum mechanics that does not restrict electrons to certain regions of space, will "delocalize" over the whole molecule, if for no other reason than to satisfy the Pauli principle requiring that the electronic wave function be antisymmetric under exchange of electrons. The tendency of electrons to delocalize can be accounted for by the use of *molecular orbitals* extending over the whole molecule, in analogy to atomic orbitals extending over the space around atoms. This is shown schematically in Fig. 4.

The lowest-energy electronic wave function composed of molecular orbitals, each containing a maximum of two electrons and having the minimum correlation of electron motion necessary to be consistent with the Pauli principle, represents the "Hartree-Fock limit." As in atoms, this orbital picture accounts for most of the electron bonding energy in molecules. It is also fortunate that this same Hartree-Fock orbital picture can account for bonding between atoms, at least qualitatively.

Attempts to improve bonding calculations quantitatively beyond the Hartree-Fock limit quickly leave the simple molecular orbital picture behind. The farthest outpost that still associates electrons with individual electron orbitals is the Geshkenbein-Ioffe (GI) type of wave function, in which molecular orbitals are each singly occupied, and an elaborate antisymmetrization procedure insures that the overall wave function obeys the Pauli principle. The molecular orbital picture is so compelling, however, that even in cases where electronic correlation of motion is explicitly taken into account, the wave function usually is constructed by superimposing many (10^3–10^6) "configurations" composed of molecular orbital functions.

That the concept of atomic structure dominates molecular electronic structure is both an advantage and a disadvantage in understanding molecular bonding. On one hand, molecular orbitals can be made up of overlapping atomic orbital functions, and the qualitative electronic structure of a molecule often can be guessed to surprising accuracy by drawing a few atomic orbitals, estimating the shapes and relative energies of the resulting molecular orbitals, and then just

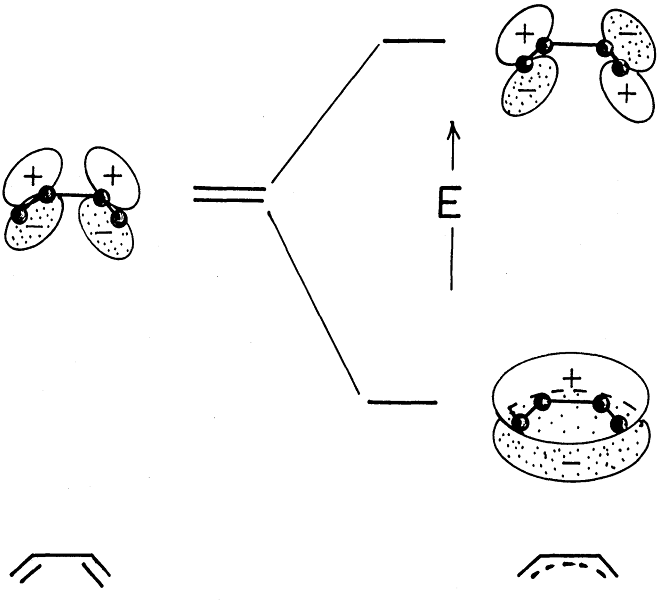

FIG. 4. Schematic of delocalization of π-bonding orbitals in *cis*-butadiene, C_4H_6. A pair of degenerate π bonds shown on the left become two delocalized molecular orbitals of different energies on the right. Delocalization occurs when it causes the overall electronic energy to decrease. One effect of delocalization is to make the otherwise single carbon-carbon bond in the center of the molecule torsionally rigid.

counting out electrons to occupy the lowest-energy orbitals. On the other hand, making quantitative predictions of bond energies is complicated by the facts that 95–99% of the electronic energy in molecules can be ascribed to the binding of electrons in atoms and that the chemical properties of molecules are predicted by perhaps the last 0.1% of the total electron binding energy. Whereas atomic orbitals provide a convenient basis for describing a molecular wave function, many linear combinations of atomic orbitals are required for the necessary precision, and 10^5–10^6 configurations of electrons occupying these orbitals is not unusual for a molecular electron wave function that gives chemically reliable adiabatic potential energies (see ELECTRONIC STRUCTURES OF ATOMS AND MOLECULES).

2.1.2 Long-Range Forces Long-range forces are dominated by multipole Coulomb effects, combined with anisotropic polarization of the electron shells of atoms. A number of these interactions are summarized in Table 1. For example, ionic crystals can be modeled to good accuracy as collections of hard, polarizable, charged ions (Stoneham and Harding, 1986). Here the forces consist of, first, the Coulomb interactions between the ions, which are attractive for arrays in which charges alternate; second, the polarizations of the atoms; and third, a short-range repulsion of the ions produced by the partial interpenetration of their electron distributions.

Between neutral atoms or molecules, an almost universal r^{-6} attraction results from the polarization of an atom by the fluctuating, instantaneous electric field of the electrons on another atom. At very long distances for which the rate at which atomic fields fluctuate exceeds the time it takes the fields to reach the other atoms, relativity modifies the form of this attraction to r^{-7}. (The wavelength of the lowest electronic transition in the isolated atoms gives a rough idea of the distance at which this occurs: 10 to 1000 picometers.) These forces induced by polarization effects appear under the names "Landau," "van der Waals," or "dispersion."

A particularly interesting and extremely important bonding force between closed shell molecules is the hydrogen bond. Partway between a strong chemical bond and a weak intermolecular interaction, the hydrogen bond is responsible for the structure of ice and much of the physical and solvating characteristics of water. It has been observed that molecules in which hydrogen is bonded to oxygen, nitrogen, or fluorine atoms have especially strong intramolecular interactions, approximately 5–10 times greater than caused by dispersion forces alone. Furthermore, these interactions are directional, whereby a hydrogen atom of one molecule attracts an O, N, or F atom of another molecule. The reason why a hydrogen bond forms between the relatively positively charged atoms of hydrogen in one molecule and the relatively negatively charged atoms of nitrogen, oxygen, or fluorine in another compound has been the subject of controversy. At first thought to be primarily an electrostatic phenomenon, it was observed that a purely multipole expansion

Table 1. Effective interatomic forces. Subscript indicates atom 1 or 2. Z_i = charge; μ_i = dipole moment; Q_i = quadrupole moment; α_i = polarizability; θ_i = angle between the axis of moment and the internuclear axis; ϕ = dihedral angle of moments; r = internuclear distance; a, x, y = parameters.

Interaction	Potential model form
	Orientation-independent potentials
Ion-ion	Z_1Z_2/r
Screened nuclear repulsion	
Modified Yukawa	$r^{-y}\exp(-ar)$
Inverse power	r^{-9}, r^{-12}
Covalent bonding	$-r^x\exp(-ar)$
Pauli repulsion	$+r^x\exp(-ar)$
Dispersion (van der Waals)	r^{-6}
Charge-induced dipole	$-Z_1^2\alpha_2/2r^4$
	Orientation-dependent potentials
Charge dipole	$Z_1\mu_2\cos\theta/r^2$
Dipole-dipole	$\mu_1\mu_2[2\cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2\cos\phi]/r^3$
Dipole-quadrupole	$3\mu_1Q_2[\cos\theta_1(3\cos^2\theta_2 - 1) - 2\sin\theta_1\sin\theta_2\cos\theta_2\cos\phi]/(4r^4)$

about molecular centers gave a poor quantitative treatment of the effect. It was also noticed that the hydrogen would locate itself in particular sites on the surface of the negative atoms. This led to speculation that quantum mechanical calculations on whole, intermolecular complexes would be needed to explain hydrogen bonds. Presently, the original idea of a particularly strong electrostatic effect appears sufficient to explain the hydrogen bond. Expansions of the electrostatic potentials about atoms, rather than about molecular centers, explain most cases of the hydrogen bond. For example, the hydrogen bond between molecules of water can be understood as the interaction of the charge distribution on oxygen with that on the hydrogen. Expansions of the atomic electrostatic potentials up to quadrupole terms on the oxygen and up to dipole terms on the hydrogen give quantitative agreement with the presently available data (Buckingham *et al.*, 1988).

The overall effective forces between neutral atoms are often represented by model potentials. Some of these forms are given in Table 2. The justification given for using any particular single form is usually mathematical convenience. Very few interatomic molecular potentials are known accurately.

More on long-range interactions and their effects can be found under CRITICAL PHENOMENA, EQUATIONS OF STATE, PHASE TRANSITIONS, PHASE EQUILIBRIA, RARE GASES, and THERMAL PROPERTIES OF GASES.

Table 2. Examples of model adiabatic potential energy functions. $\epsilon = |\text{potential minimum}|$; $\sigma = 2^{1/16}r_e$; r_e = location of potential minimum; r = internuclear distance; α, a, b, c, d = parameters.

Lennard-Jones (6–12):

$$V(r) = 4\epsilon[(\sigma/r)^{12} - (\sigma/r)^{6}]$$

Morse:

$$V(r) = \epsilon\{\exp[-2a(r - r_e)] - 2\exp[-a(r - r_e)]\}$$

Yukawa-6:

$$V(r) = \epsilon\frac{1 + \alpha r_e}{\alpha r_e - 5}\left\{\frac{6r_e}{(1 + \alpha r_e)r}\exp[\alpha(r_e - r) - r] - \frac{r_e}{r^6}\right\}$$

Buckingham:

$$V(r) = b\exp(-ar) - cr^{-6} - dr^{-8}$$

Modified Buckingham:

$$\begin{aligned} V(r) &= \frac{\epsilon}{1 - 6/a}\left\{\frac{6}{a}\exp\left[a\left(1 - \frac{r}{r_e}\right)\right] - \left(\frac{r_e}{r}\right)^6\right\} \text{ if } r \geq r' \\ &= \infty \quad \text{if } r < r' \end{aligned}$$

r' given by root of

$$\left(\frac{r'}{r_e}\right)^7 \exp\left[a\left(1 - \frac{r'}{r}\right)\right] = 1$$

2.2 Bond Angles and Lengths

Three circumstances simplify our understanding of bonding in most molecules. First, the almost hard-sphere, repulsive, short-range interaction between atoms allows us to assign almost consistent interatomic bond lengths between pairs of atoms. Second, bonding between atoms can be represented to a great degree as pairwise interactions between two atom centers. Third, bond angles at atomic centers are relatively constant for a series of molecules possessing similar elements. This leads to the "ball and stick" model that serves so well in much of chemistry. Table 3 summarizes much of this model for bonding among main-group elements.

Bond lengths and bond angles have to be adjusted to account for multiple bonds between adjacent atoms; this is readily included in ball and stick models. What is more difficult to include are the subtle variations of bond angles in response to bonding situations. A purely empirical model to describe bond-angle variations has been introduced under the name Valence Shell Electron Pair Repulsion (VSEPR), which is now taught to almost every chemistry student. This model extends the ideas of electron pairs prevalent in Lewis structures: Each atom is imagined as surrounded by pairs of electrons, both bonding and nonbonding, which "repel" one another. It gives qualitative predictions for bond angles about each atomic center in a molecule, which are surprisingly effective, considering the difficulty in justifying this model *ab initio*.

One should be careful about extending a set of bond lengths and bond angles derived for closed-shell molecules to open-shell radicals. In the latter, the electron distributions present around a bonding atom will differ considerably from those present under the

Table 3. Representative interatomic bond lengths (pm). — single, = double, ≡ triple bonds. Data from Kuchitsu *et al.* (1992) and Vilkov *et al.* (1983).

	H	C	N	O	F	Si	P	S	Cl
H	75								
C	109	—154							
		=134							
		≡120							
N	101	—148	—145						
		=127	=125						
		≡116	≡110						
O	96	—142	—145	—145					
		=121	=114	=121					
F	92	—136	—137	—157	—142				
Si	148	—187	—172	—163	—(156)	—233			
				=(150)					
P	142	—185	—165	—160	—(156)		—222		
		≡154	=154	=(150)			≡189		
			≡149						
S	134	—182		=147	—(159)		—211	—206	
		=(161–171)						=189	
Cl	128	177	176	—169	163	205	204	201	199

conditions for which these parameters are derived, and serious errors may result. For example, the methyl radical CH_3 derived from methane by removal of one hydrogen would be thought to be a carbon on a tripod of three hydrogen atoms having bond angles of approximately 109° based upon the methane (CH_4) tetrahedral geometry. In actuality, the methyl radical is planar in its ground state, with all three hydrogen atoms forming an equilateral triangle with a carbon atom in the center; the bond angles are 120°. However, the closed-shell methide anion, CH_3^-, is again nonplanar. It is a dangerous extrapolation to treat an unsatisfied, "dangling bond" on an atom as just another substituent atom or group.

For some compounds, the concept of the two-centered bond breaks down. Boron and hydrogen form networks of three-centered bonds; the expected compound borane dimerizes into the structure diborane,

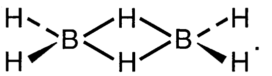

The two-electron, three-center, B—H—B bond is one of the few cases in which hydrogen is situated symmetrically between two bonding partners.

Transition-metal compounds also provide numerous examples of extended, multicenter bonds. For example, the sandwich compounds the metallocenes, of which ferrocene, $Fe(C_5H_5)_2$, is most famous, attach the two planar cyclopentadienyl rings to the central iron by a system of diffuse, eleven-center bonds.

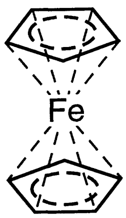

2.3 Stereochemistry

The relative permanence of the spatial configuration of atoms within molecules allows the isolation of different molecules having the same numbers of atoms of each element (i.e., the same molecular formula), but arranged differently in space.

2.3.1 Isomers and Conformers A relatively stable arrangement of atoms is termed a *conformer*; different conformers that are stable long enough to allow physical separation from one another are termed *isomers*. Rapidly interconverting conformers are termed *tautomers*. For example, the carbon skeleton of cyclohexane (C_6H_{12}) has two types of conformers, "chair" and "boat" forms, but these readily interconvert at room temperatures and cannot be isolated:

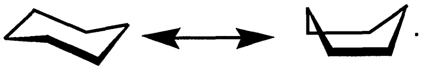

The carbon skeleton of butadiene (C_4H_6), however, is kept planar by the delocalized interaction of the electrons in the double bonds. Two *geometric isomers, cis-* and *trans*-butadiene, may be isolated, differing only in the spatial arrangements of atoms, but having the same bond topology:

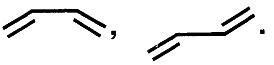

Topological isomers (or constitutional isomers) differ even more strongly from one another by having topologically different networks of bonds. Thus, bicylobutane C_4H_6 is a topological isomer of butadiene, as can be seen from the bonding topology of its carbon skeleton compared to that of butadiene:

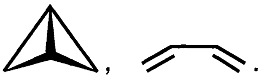

The various classes of isomers are related as shown in Fig. 5.

2.3.2 Chirality A spatial configuration of four atoms introduces the possibility of mirror-image conformers. For example, hydrogen peroxide H_2O_2 has two interconverting conformers, one left handed and the other right handed:

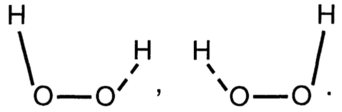

The relative rigidity of bonds about atoms such as carbon allows the isolation of mirror-image isomers termed *enantiomers* or optical isomers, based upon their ability to rotate the plane of polarization of light in opposite directions. Whenever an isomer is not rotationally superimposable upon its mirror image, it acquires the property of "handedness," and is termed *chiral.* Organic chemists are specially alert to this possibility whenever a carbon atom is bonded to four different atoms or groups, thus making it chiral. An example of this is lactic acid, which has two enantiomeric forms. Here a stereochemical convention indicates bonds receding from the plane as dotted; bonds approaching from the plane are emphasized:

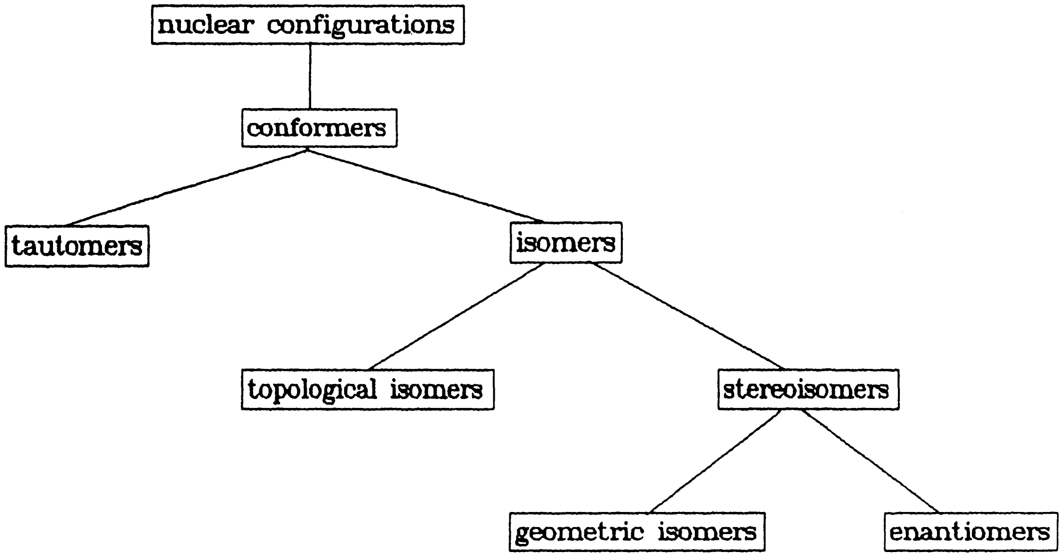

FIG. 5. Classification of nuclear conformers.

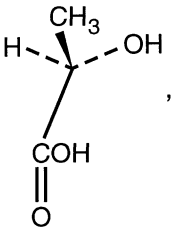

D (–)

,

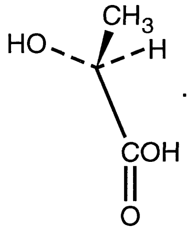

L (+)

.

Prepared by ordinary laboratory synthesis, a *racemic* mixture of both enantiomers is produced; prepared by bacterial fermentation of sucrose with bacillus *acidi levolactii*, predominantly the left-handed L-lactic acid is produced.

Racemic mixtures may be separated into enantiomers by physical, chemical, or biological techniques. Pasteur first used the fact that under the cool conditions of a 19th-century European laboratory, D- and L-tartaric acid would crystallize separately—it was just a matter of picking through crystals by hand! A seed crystal of one or the other enantiomer can be used to promote the crystallization of that form from a mixture.

Alternatively, reactions may separate enantiomers. If a chiral molecule is added to a racemic mixture, one enantiomer of a mixture may react more readily, and the reaction products will be enriched in one or another enantiomer.

An elegant physicochemical technique resolves enantiomers with a chiral chromatography column (see CHROMATOGRAPHY): A racemic mixture is made to diffuse along a chiral material such as a polymer.

One method of separating racemic mixtures of biological molecules is to find a biological species that preferentially consumes one enantiomer. Numerous optically active centers may be present in a large molecule. Life processes depend upon many molecules that possess chiral carbon centers, such as sugars and amino acids, and often preferentially consume or produce only specific enantiomers.

More information on the topic of stereochemistry will be found in CONFORMATIONAL ANALYSIS.

2.3.3 Symmetry The molecular Hamiltonian is invariant under exchange of identical particles. For molecules possessing identical nuclei, this invariance leads to classification of molecular states according to the symmetry group representations for exchange of nuclei (see GROUP THEORY).

The full spatial symmetry group of the Hamiltonian of a molecule consists of the operations of translation, rotation, electron permutation, nuclear permutation, and inversion. The first three correspond to classification of molecular states according to total linear momentum, overall angular momentum, and the Pauli principle for electrons. The last two make up a *permutation-inversion group*, according to which all molecular states may be further classified. This is an exact symmetry group of the molecule and does not depend upon a particular conformation of nuclei. It is useful for fluxional molecules or for molecules undergoing reactions. Examples of typical permutation-inversion operations are shown in Fig. 6.

If a molecule is assumed to have a fixed conformation, molecular states may be classified according to another group, the geometrical *point group* of the molecule. This group consists of all proper and improper rotations, inversions, and mirror reflections that transform the electronic wave function but which leave the positions of the nuclei invariant; examples of these operations are

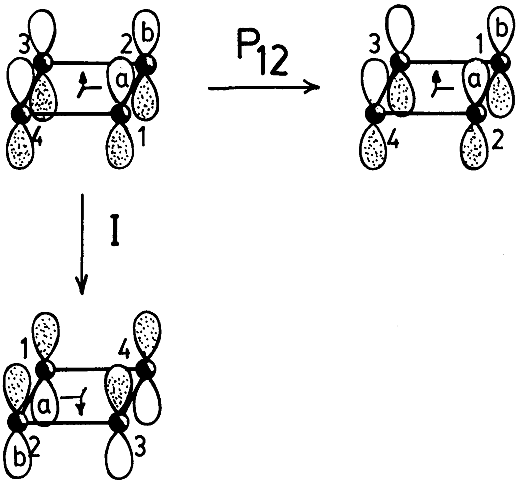

FIG. 6. Nuclear permutation-inversion group operations illustrated on four equivalent nuclei. Nuclei are numbered, while letters indicate wave functions. The whole group is generated by the operations of permuting pairs of nuclei (P_{ij}) and inverting the coordinate system (I). For four nuclei, the group contains $4! \times 2 = 48$ operations. Further description is given by Bunker (1978).

shown in Fig. 7. The molecular point group assumes that the molecule is rigid or, at least, that the vibrations of the nuclei take them only small distances from their equilibrium positions (Wilson, *et al.*, 1955).

Molecular point groups are normally used to classify molecular states in spectroscopy. Any subgroup of the permutation-inversion group may be used to classify molecular states, and in spite of being founded upon the approximation of molecular rigidity, a molecular point group acquires dynamical significance by being isomorphic to a subgroup of the larger, permutation-inversion group, which has exact symmetry. (The point group *is* a subgroup of the product of two invariance groups of the molecular Hamiltonian: the permutation-inversion group and the group of overall rotations of the molecule.) Molecular point groups have the advantages of an extensive literature, readily available character tables, and representations amenable to three-dimensional visualization. They have the disadvantage of being dynamically incomplete, not accounting for the complete permutation symmetry of the nuclei.

For fluxional molecules, interconverting between conformations, one may use (Bunker, 1978) the *molecular symmetry group,* also known as the *feasible group,* which incorporates all operations allowed under the point group, along with the "feasible" nuclear permutations accompanying interconversions between conformations.

2.4 Extended Bonding Structures

2.4.1 Clusters Extensive research currently proceeds on clusters of atoms. The bonding forces holding clusters together vary from strong electronic bonding as in clusters

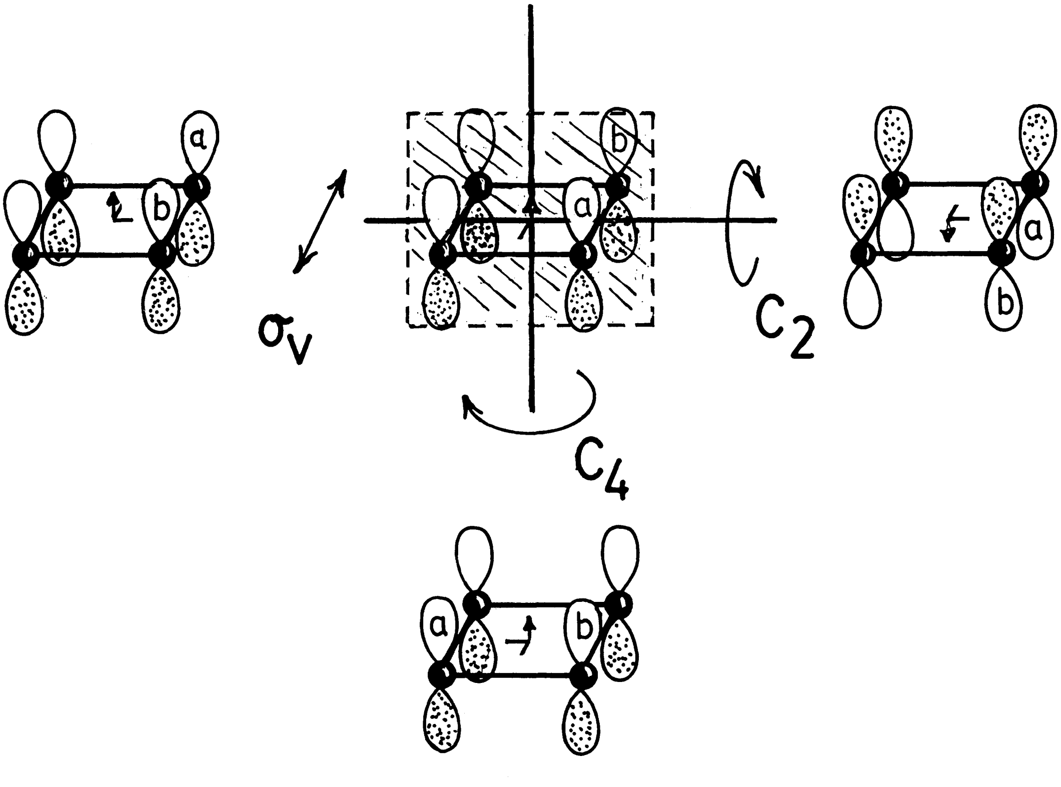

FIG. 7. Molecular point-group operations illustrated on a square arrangement of four identical atoms. All operations transform wave functions, leaving the fixed nuclear positions invariant. The whole group (in this case D_{4h}) can be generated by C_n operations (rotations about n-fold symmetry axes) and σ operations (reflections about mirror planes). For D_{4h}, there are 16 group operations, considerably fewer than for the corresponding permutation-inversion group. Further description and applications of molecular symmetry are given by Herzberg (1966).

of atoms of normally metallic elements, to weak, van der Waals bonding, as in clusters of noble-gas atoms. Work on clusters is made easier by inclusion of a charge, which permits the determination of mass with a mass spectrometer. Types of questions being explored with clusters include "How big a cluster is required to have 'metallic' behavior, as evidenced by ionization energy, electron affinity, and band structure?" Other questions focus on the statistical mechanical properties of clusters, such as "How big does a cluster have to be to resemble bulk material?" "Is a cluster crystalline or liquid?" "How tightly are the components held within clusters?" For small clusters, another question is "What is the structure of a particular cluster?"

Cluster researchers discovered that there are certain "magic numbers" in the size-distribution spectra of the clusters that they produce, generally by condensation from a cold molecular jet. Argon seems to prefer to form clusters consistent with formation of icosahedral shells. Carbon jets give relatively large amounts of C_{60}. There is good evidence that what researchers found was a strong propensity for carbon to form spherical shells of alternating hexagonal and pentangular bonding patterns, given the name "fullerene" because they resembled the geodesic domes of Buckminster Fuller (see Fig. 8). Gram quantities of this material can now be produced and treated as a chemist would treat an ordinary chemical substance.

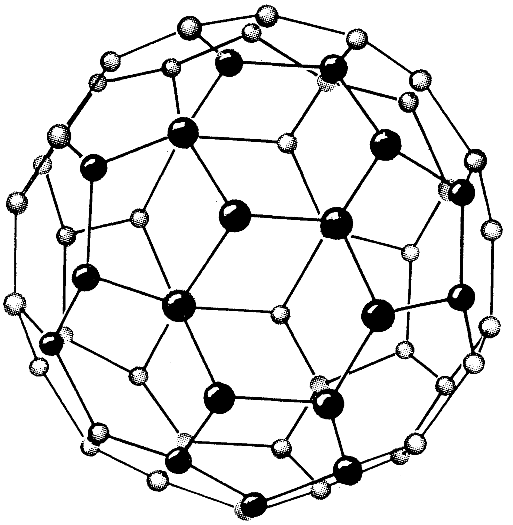

FIG. 8. Fullerene, a highly symmetrical form of carbon C_{60}.

There is some interest in producing monodisperse size distributions of clusters for the purpose of creating nanostructure materials. These could find use, for example, as catalysts in the chemical industry or as "quantum dot" structures in the electronics industry (Bawendi *et al.*, 1990).

More on this topic can be found in MOLECULAR AND ATOMIC CLUSTERS.

2.4.2 Excimers Closed-shell molecules, when excited, often increase their interactions with other molecules. For example, excited metastable rare-gas atoms form respectable chemical bonds, with binding energies on the order of an electron volt, between atoms of the same type:

$$\mathrm{Ar}^* + \mathrm{Ar} \rightarrow \mathrm{Ar}_2^*$$

This behavior of excited rare-gas atoms contrasts with their notoriously weak interatomic interactions in the ground state. Since the bonding depends upon the electronic excitation, molecules that form from otherwise distinct molecules only when one is excited, and which dissociate when the excitation relaxes, are termed *excimers*.

The increased interaction between excited molecules and other ground-state molecules is responsible for increased rates of energy transfer, increased collisional cross sections, reduced diffusion coefficients, and clustering. Excimers can serve as incipient nucleation sites. For example, a saturated or nearly saturated vapor can often be induced to form droplets by illumination with a wavelength of light that excites a constituent molecule to an energetic electronic state. Or, they serve as lasing systems in the vacuum ultraviolet, whose ground states conveniently disappear as rapidly as they are formed by stimulated emission! (See LASERS, GAS.)

2.4.3 Polymers Some molecules, designated as *monomers*, can react with many similar monomer units to form long-chain or network *polymers*. For example, ethene molecules may serve as monomers in the additive polymerization of the linear polymer polyethylene,

$$\text{—}(\mathrm{CH_2}\text{—}\mathrm{CH_2})_n\text{—}.$$

When they react, two or more different types of monomers may copolymerize. For example, the condensation polymerization of formaldehyde and resorcinol, with the elimination of water, forms a hard resin used to cement wood:

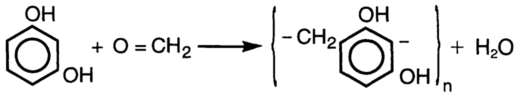

Most polymerization reactions give *polydisperse* mixtures of polymer molecules of varying size, shape, and molecular weight. The distribution of polymer mass and the frequency of branched chains strongly affect mechanical properties. Industrial production of plastics and resins attempts to control these carefully.

Some polymer samples are *monodisperse,* having all molecules the same mass (or, at least, within a narrow range of masses) and of similar structure. Monodisperse samples of common polymers are especially valuable to researchers investigating the thermodynamic and transport properties of polymeric materials. Measurements of the properties of pure monodisperse samples, and carefully prepared mixtures of these, provide the empirical parameters needed by theoretical models to predict viscosity, melting, diffusion, and other thermodynamic and transport properties of polydisperse mixtures.

Recent interest in dendritic polymers stems from the desire both to generate monodisperse polymer samples and to control chain branching. Reactions forming these carefully tailored large molecules double the molecular mass at each step by combining two polymers, rather than incrementing it by a constant amount by adding small monomers to a large, growing polymer. Because at each step the product polymer is twice the size of the starting polymers, it can be easily separated and purified before being combined with itself again in a molecular size-doubling reaction.

Many biological molecules are formally polymers composed of mixed monomer units. Proteins are chains of amino acid monomers. The structural proteins of cells contain repeating units of similar amino acids. Other cell parts, such as enzymes, consist of proteins containing careful sequences of different amino acids determined by the genetic material of the cell. The genetic material itself, DNA and RNA, consists of repeating units which encode the genetic information.

A major unsolved problem at this time is understanding how a specific sequence of amino acids affects the folding of the molecular chain in a protein, and hence determines the protein shape, and ultimately its chemical and biological function. A very delicate interplay of intermolecular and intramolecular forces, including van der Waals attraction and hydrogen bonding, is at work in the molecule and the solvent. Numerous local minima occur on the adiabatic potential surface for proteins. Direct simulations of protein folding face serious difficulties with proteins larger than about one hundred amino acids. In reality, a protein in a distorted conformation may take seconds, or minutes, to relax to its active conformation; this is a much longer time than achievable in a molecular dynamics calculation, in which time steps are often sub-femtosecond, and a nanosecond is computationally a long time! (See MOLECULAR DYNAMICS OF BIOLOGICAL MOLECULES; MONTE-CARLO METHODS.)

Three very large fields of active work are introduced in POLYMERS, STRUCTURE OF; PROTEINS AND ENZYMES; PROTEIN DYNAMICS; and NUCLEIC ACIDS.

2.4.4 Crystals Crystals of chemically bonded atoms may be considered to be the largest molecules. For example, the covalent bonding network of carbon atoms in a crystal of diamond meets all the criteria of a molecule. The ionic bonding network of ordinary sodium chloride table salt, NaCl, holds sodium and chloride ions in place with electrostatic bonding forces. Other crystals consist of molecules held together by weaker, intramolecular forces. More on crystals can be found in CRYSTALLINE STATE.

2.5 Structure Determination

2.5.1 Theoretical As noted under the topic of bonding, the stability of molecules is determined primarily by the quantum mechanical distribution of the electrons; the general theoretical background of this subject is contained in QUANTUM MECHANICS; recent advances have been summarized by Payne *et al.* (1992); specific aspects of this topic are contained in ELECTRONIC STRUCTURE OF ATOMS AND MOLECULES.

2.5.2 Experimental

2.5.2.1 Mass Spectrometry Mass spectrometry makes use of the mass-to-charge ratio of the ions produced from molecules introduced into an ionization source of a mass spectrometer (see SPECTROMETERS, MASS). If molecules are singly ionized in the source without fragmentation, they are termed *parent ions* and appear in the mass spectrum at the molecular mass, which may serve to establish the molecular weight. *Fragment* or *daughter ions* may also be formed during ionization, and the masses of these ions typically correspond to stable structural groups within the parent molecule, or to stable rearrangement products. The mass spectrum of these daughter ions may help establish the structure of unknown molecules. For example, the presence of a benzyl group $C_6H_5CH_2\cdot$ in an organic molecule can be signaled by the appearance of mass 91 in the mass spectrum.

2.5.2.2 Diffraction The most detailed information on large-molecule spatial structure comes from diffraction data on crystalline samples. Atomic spacings in molecules are on the order of 10^2 picometers, while crystal unit-cell dimensions are of the order of 10^3–10^4 picometers; the difference in the wave vectors between incoming and scattered radiation must have Fourier components in this range to satisfy the Bragg condition for coherent diffraction from multiple centers. The Bragg condition requires x rays in the 10-keV range, neutrons (or protons) in the 0.1-eV range, or electrons in the 100-eV range (possibly higher to emphasize forward scattering). The large inelastic scattering cross sections for charged particles limit electron or proton scattering to thin samples, whereas the inconveniently low energy for protons severely limits their use in diffraction. Monochromatic thermal neutrons are of limited availability; their use generally complements x-ray diffraction, the primary method of crystallographic structure determination.

X rays scatter off the electrons, providing data that strongly emphasize heavy atoms with many electrons; for large molecules, the positions of hydrogen atoms may be very difficult to determine. Neutrons scatter off nuclei; hydrogen and, even more so, deuterium scatter neutrons strongly, and thus neutron scattering provides information about the lighter atoms that is often difficult to obtain by x-ray diffraction (see X-RAY DIFFRACTION; NEUTRON DIFFRACTION).

Electron diffraction finds use in the determination of structures of small gas molecules, thin (<500 μm thick) solid samples, or even molecules on surfaces. Interatomic distances in small molecules have been determined in the gas phase by electron diffraction, and questions of conformations of moderately large molecules of a dozen atoms have been solved with electron diffraction on gases (see ELECTRON DIFFRACTION; ELECTRON SCATTERING BY ATOMS AND MOLECULES).

2.5.2.3 Rotational Spectroscopy The highest resolution techniques for determination of small-molecule structures involve measuring the moments of inertia for molecular rotation. Molecular moments of inertia may be obtained directly from the microwave spectra of molecules or indirectly from electronic, infrared, or Raman spectra of molecules (see MOLECULAR SPECTROSCOPY).

The basic equation useful here relates the observed rotational energy levels to the molecular moments of inertia. For a linear rigid molecule of moment of inertia I, possessing only rotational angular momentum J, the energy levels E_J are given by

$$E_J = BJ(J + 1), \tag{5}$$

$$B = h^2/8\pi^2 I. \tag{6}$$

Other expressions for nonlinear, vibrating, and open-shell molecules can be found elsewhere (Herzberg, 1945, 1950, 1966; Townes and Schawlow, 1955; Hirota, 1985).

In rigid molecules, at most three moments of inertia may be determined spectroscopically; for molecules in which the number of internal atomic coordinates exceeds three, individual atomic masses may be varied by isotopic substitution to change both the center of mass and the relative moments of inertia, thus providing needed additional information to determine molecular structure. In the best cases, each isotopic substitution may provide three more independent constraints on the atomic coordinates. Difficulties occur if

1. the necessary isotopes are unavailable;
2. not all moments of inertia may be measured in an experiment;

3. the system of equations that results becomes ill conditioned; or
4. the molecule is insufficiently rigid.

The last problem often occurs in molecules containing hydrogens, since ground-state vibrational motion of a light proton is roughly 10% of its bond length. On the whole, bond lengths may be determined to within a picometer for heavy atoms and five picometers for hydrogen atoms in usual cases. Vilkov *et al.* (1983) compare the different practical methods, while Kuchitsu (1992) has supervised a recent compilation of results obtained on actual molecules with these methods, combined with those from electron diffraction.

2.5.2.4 Vibrational Spectroscopy Certain groups of atoms vibrate at, or near, characteristic frequencies in molecules (Wilson *et al.*, 1955). This property can be used to identify the presence or absence of particular structural groups, such as represented in Table 4. Infrared, Raman, uv/visible, and electron spectroscopies, and even inelastic neutron scattering, have been applied to determination of molecular vibration frequencies. Summaries of data on specific molecules can be found in the compilations of Shimanouchi (1972, 1977, 1978, 1980) for stable molecules and Jacox (1984, 1988, 1990) for free radicals and ions.

2.5.2.5 Magnetic Resonance Modern organic chemistry could not exist without nuclear magnetic resonance (NMR) to provide information needed to understand the bond topology of the molecules it produces. The chemical usefulness of NMR depends upon two major physical effects: the interaction of nuclear spins to identify adjacent nuclei and the diamagnetic shielding shifts, which indicate the relative electron density near nuclei. Before NMR, only a "complete" synthesis of a molecule by firmly understood chemical reactions served to confirm molecular structure; now modern spectroscopic techniques, principally NMR, establish molecular structure by physical methods in absence of a complete chemical synthesis.

Table 4. Characteristic vibrational frequencies.

Group or motion	Wave number (cm^{-1})
C—H stretch	2900–3400
C—H bend	1300–1500
C—O—C	1050–1300
C—OH	1000–1300
$=C=O$	1000–1350
N—H	3300–3600
$—C\equiv N$	2200–2400
—S—H	2600
$=P—H$	2400
$\equiv Si—H$	2100–2300
$=S=O$	1000–1150
$=SO_2$	1150–1250
—NO	1320–1220
—O—NO	1600–1700
	1500–1600
$—NO_2$	1300–1350
—C—F	1000–1350
—C—Cl	650–800
—C—Br	200–600

To a lesser degree, where applicable, electron paramagnetic resonance has been used to determine molecular structure (see MAGNETIC RESONANCES, DETERMINATION OF STRUCTURES BY; MAGNETIC RESONANCE AND QUADRUPOLE RESONANCE, NUCLEAR; ELECTRON PARAMAGNETIC RESONANCE).

3. PHYSICAL PROPERTIES

3.1 Molecular Mass

The fundamental unit of molecular mass is the atomic mass unit u, now defined as $\frac{1}{12}$th the mass of the carbon-12 atom.

For macroscopic use, molecules are counted out in units of *moles* (the "chemist's dozen"!), which contain N_A molecular particles. Avogadro's constant, $N_A \approx 6.022 \times 10^{23}$, gives the ratio of the mass of 1 g to that of 1 u. Deslattes (1980) has reviewed how Avogadro's constant is determined. Numerically, the number of grams per mole is equivalent to the number of atomic mass units per molecule of a substance.

One of the most useful tools for identification of an unknown compound is the mass spectrometer. A revolution in the detection and identification of small amounts of chemicals has occurred with the coupling of the mass spectrometer with the capillary bore chromatography column. As mentioned earlier, one useful product of mass spectrometry is the direct determination of the molecular mass from the mass of a parent ion.

For large molecules, such as polymers or biopolymers, mass spectrometry has been less useful, since production of an unfragmented parent ion of a large, nonvolatile molecule by traditional means is problematic. The technique of "electrospray" ionization shows promise here, in which droplets of a solution of molecules are sprayed into a vacuum. Evaporation leaves the molecule of interest surrounded by only a few of the solvent molecules and solution ions; this process produces an aggregate ion having a mass only slightly greater than that of the molecule whose mass is to be determined.

Traditional methods for determining molecular mass involve one form or another of an equation of state; for a dilute gas, one may use the perfect gas equation

$$PV = NkT, \tag{7}$$

where P, V, and T are the pressure, volume, and temperature variables for N molecules (k is Boltzmann's constant). For molecules that can form a dilute gas, the molecular mass can be easily determined from the implication of this equation that 1.00 mole of gas at 1.00 atm and 0°C will occupy 22.4 L (see MASS AND DENSITY, MEASUREMENT OF; EQUATIONS OF STATE).

In cases where molecules cannot be conveniently made into gases but can be dissolved in dilute liquid solutions, van 't Hoff showed they again behave much like a perfect gas. For example, the osmotic pressure Π generated by a solution in contact with its pure solvent through a semipermeable membrane takes the perfect-gas form in the limit of low concentration:

$$\Pi V = NkT, \tag{8}$$

where V is now the volume of the solution and N is the number of molecules dissolved in the solvent.

Determining molecular weight by osmosis (*q.v.*), using the van't Hoff equation (8), or its Huggins-Flory modification (see POLYMERS, MOLECULAR STRUCTURE), can be very sensitive, even for large molecular masses (30 000 u). In fact, it can be too sensitive for small masses, and large molecules with which it is used must be free of small ions and molecules. The sensitivity of this method can be seen by comparing osmotic pressure to the vapor-pressure difference of solvent and solution in the dynamic equilibrium of the apparatus shown in Fig. 9. For equilibrium to occur, the vapor pressure of the pure solvent at the bottom of a column of vapor must equal the vapor pressure over the solution at the top of the column, plus the small pressure height of the vapor resulting from its finite density. The osmotic pressure, on the other hand, is given by the pressure height difference of a column of solvent vs a column of its vapor!

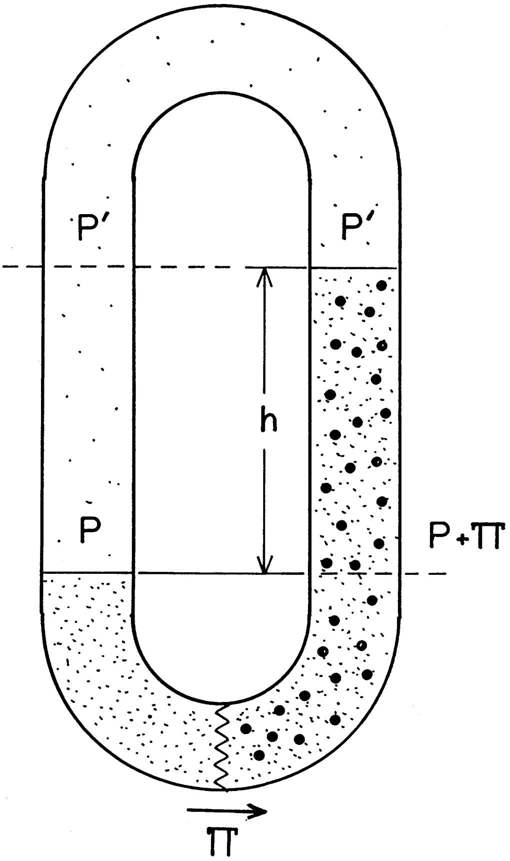

FIG. 9. Thermodynamic equivalence of a large osmotic pressure and a small vapor-pressure lowering produced by addition of a small amount of solute molecules into a pure solvent. The vapor-pressure difference *P-P′* is given by the hydrostatic head of a column of vapor equivalent in height to a column of solution held elevated by the osmotic pressure difference *Π*.

Although much smaller than osmotic pressure differences, vapor-pressure lowering of a solvent upon addition of solute molecules can be useful for determining molecular mass of small or moderate-size molecules. Thermodynamics applied to Fig. 9 gives the quantitative effect

$$\Delta P = \frac{N}{N_s}\frac{kT}{(v' - v)}. \tag{9}$$

Here a molecule of solvent occupies volumes v' and v in the vapor and liquid states, respectively; N is the number of molecules dissolved in N_s molecules of solvent.

It is often more convenient to measure the change of temperature at which a state change occurs. From Eq. (9) one obtains

$$\Delta T = \frac{N}{N_s}\frac{kT}{(s' - s)}. \tag{10}$$

Here s' and s refer to the entropy per molecule of the solvent in the two states, respectively. When additional solute molecules are present only in the liquid phase, the boiling point is elevated, while the freezing point is depressed. The changes in the freezing and boiling points of water are relatively small because of large entropy changes that occur on boiling or freezing, but even so water has a −1.85 °C freezing-point depression for 1-mol/L solutions (probably the concentration limit for applying a theory developed for "dilute" solutions).

Light scattering can be used to determine the molecular mass of dissolved molecules if their index of refraction differs from that of the solvent. Rayleigh showed that the incoherent scattering of unpolarized light from small inhomogeneities is proportional to their number density and inversely proportional to the fourth power of the wavelength λ. In the limit of low concentration c and long wavelength (much larger than the molecules of interest),

$$m = R_\theta/Kc, \tag{11}$$

where

$$K = 2\pi^2 n_s^2 (dn_s/dc)^2 \lambda^{-4}, \tag{12}$$

$$R_\theta = \frac{r^2 I(\theta)}{I_0(1 + \cos^2\theta)}, \tag{13}$$

and n_s is the index of refraction of the solution, r is the observation distance from the scattering region, and $I(\theta)$ and I_0 are the scattering intensity observed at angle θ and the incident intensity, respectively. This relationship, and others like it, find applications in polymer physics and chemistry.

Two other methods used routinely in biochemical research are electrophoresis and ultracentrifugation. Both depend upon the resistance to motion of a large molecule in a solvent and can be size and shape dependent (Noolandi, 1992).

3.2 Energy of Formation

The energies (or enthalpies) of formation of many molecules have been tabulated; tables include those of Wagman *et al.* (1982) and the JANAF tables (Chase *et al.*, 1985); also, a useful compilation is available in recent editions of the *Handbook of Chemistry and Physics* (Lide, 1993). For more specific information on methods, see THERMOCHEMISTRY.

A compilation of typical bond energies, along with actual determinations for specific molecules, is given by J. A. Kerr in the *Handbook of Chemistry and Physics* (Lide, 1993).

A useful summary of thermodynamic data of molecular ions has been compiled by Lias *et al.* (1988). Recent advances have been made by the increased precision available through electron spectroscopy (see SPECTROMETERS, ELECTRON AND ION).

The *ionization energy* E_i represents the minimum energy necessary to remove an electron from a molecule, typically in the 5–15-eV range. A related quantity, the *electron affinity* E_a, represents the energy released when an electron is attached to a molecule. These range from negative values (associated with temporary negative ions) to positive 5 eV. A recent compilation of electron affinities, the binding energies of electrons to neutral molecules, is provided by Miller in recent editions of the *Handbook of Chemistry and Physics* (Lide, 1993). Many earlier compilations of numbers obtained before the prevalence of laser photodetachment techniques contain numerous erroneous entries (see PHOTOEMISSION AND PHOTOELECTRON SPECTRA; ELECTRON SCATTERING BY ATOMS AND MOLECULES.)

The tendency of an atom within a molecule to attract electrons is referred to as *electronegativity*. Mulliken defined electronegativity of an atom as the average of its ionization energy and electron affinity. This definition of electronegativity for atoms has been extended to radicals; it has been generalized by making it a function of the degree of charge the atom or radical possesses in the molecule, or the type of atomic orbital involved. A number of other scales of electronegativity have been developed; one by Pauling (1960) uses the energy released in the formation of heteronuclear diatomics from the parent homonuclear diatomics,

$$X_2 + Y_2 \rightarrow 2XY.$$

The Pauling numerical scale approximates $(E_i + E_a)/(5.4\text{ eV})$.

Electronegativity differences are useful in predicting the amount of ionic character in a chemical bond. Electronegativity differences of 2 units (Pauling scale) between bond-forming groups produce essentially ionic bonds (Gordy *et al.*, 1953).

3.3 Electric and Magnetic Constants

Electric dipole moments are most accurately determined by the Stark effect of free molecules; magnetic dipole moments are most accurately determined by the Zeeman effect (see ZEEMAN AND STARK EFFECTS). The determination of the electric dipole moment of molecules is complicated by the fact that rotations average away dipole moments: Molecular rotational states must be parity eigenstates, which have no dipole moments inherently. For most molecules, without degenerate rotational levels, the shift of rotational energy levels in an electric field is a second-order effect. In contrast, molecules having unpaired nuclear or electronic spins, or electronic orbital angular momentum, have magnetic degeneracies of their rotational energy levels, which are split by a magnetic field in a first-order effect (Townes and Schalow, 1955; Hirota, 1985). Thus, the Zeeman effect can be large, and it has been useful recently in investigating magnetic molecules by laser magnetic resonance (LMR) in the far infrared with only fixed-frequency lasers: The molecules can be "tuned" into resonance with the laser frequency by scanning a magnetic field. For molecules in condensed phases, the electric dipole moment can be extracted from the bulk dielectric constant, if some assumptions are made about the inherent polarizability of the molecules and the statistical mechanics of the bulk. Magnetic resonance is preferred for studying paramagnetism in the bulk.

Molecular dipole moments have been extensively tabulated—for example, by Nelson *et al.* (1967).

Higher-order electric moments and polarizabilities have been discussed theoretically by Bishop (1990), who points out that except for cases where they vanish by symmetry, higher moments of both atoms and molecules can be expected to be generated by charge asymmetries on the order of electronic charges displaced by atomic radii and thus are non-negligible on the atomic scale. Except for polarizabilities, many tabulated by T. M. Miller in the *Handbook of Chemistry and Physics* (Lide, 1993), few measurements of higher electronic moments and polarizabilities have been made. Although nuclear quadrupole coupling coefficients measure the local inhomogeneous second derivative of the electric potential at the nucleus, the overall molecular quadrupole moment is relatively difficult to determine. The tabulated values of Gordy *et al.* (1953), obtained indirectly from microwave collisional line broadening, have been widely quoted. This situation is unfortunate, since, as implied in the discussion on bonding above, covalent bond formation is accompanied by the development of a quadrupole moment.

Bulk magnetic susceptibilities have been readily studied. The weak diamagnetism of closed-shell molecules is a quantum mechanical effect. That the energy associated with electronic motion increases in a magnetic field might best be understood in this picture: The magnetic field B can be locally transformed away by moving to a rotating frame and solving the electronic motion there; transforming back to the physical fixed frame imparts to the electronic motion a small amount of additional, rotational kinetic energy. Because the molecular diamagnetic susceptibility is so small, to a good approximation it can be considered to be additive up to the high concentrations associated with bulk liquids or solids. Tabulations include those of R. R. Gupta in the *Handbook of Chemistry and Physics* (Lide, 1993).

Fine and hyperfine structure constants are discussed in MOLECULAR SPECTROSCOPY.

4. CHEMICAL DYNAMICS

4.1 Separation of Nuclear and Electronic Motion

The dynamics of molecules are heavily influenced by the adiabatic potential surfaces on which the nuclei move; in most cases, the motions of nuclei can be adequately modeled with semi-classical, or even classical, equations of motion on the potential energy surfaces corresponding to the ground electronic state (Schatz, 1988). The adiabatic approximation neglects effects of the nuclear kinetic energy operator T_{nn} in calculation of the purely electronic energy. The operator T_{nn} contains derivatives with respect to nuclear coordinates; in addition to accounting for nuclear momenta, these derivatives also detect changes in the electronic wave function that depend upon nuclear coordinates. When these terms are included, the nuclei acquire additional inertia as the electrons follow the nuclei; couplings between adiabatic potential energy surfaces also result. The most serious of these "nonadiabatic" couplings between adiabatic potential energy surfaces are of "velocity" type: They introduce an interaction between adiabatic surfaces that is proportional to the nuclear velocity. To first order, the adiabatic approximation holds only if

$$E_k(R_N) - E_j(R_N) \gg \hbar \dot{R}_N \langle \psi_k(r_e,R_N) | \partial / \partial R_N | \psi_j(r_e,R_N) \rangle, \qquad (14)$$

where $E_k - E_j$ is the energy separation between adiabatic surfaces having electronic wave functions ψ_k and ψ_j that depend upon the nuclear and electronic coordinates R_N and r_e; $\dot{R}_N$ is the nuclear velocity.

Whenever adiabatic surfaces approach closely, or nuclear motions are rapid, the molecular system may "hop" from one adiabatic surface to another.

4.2 Conservation Rules

During a chemical reaction, mass (or, more properly, mass-energy) and charge are conserved. Additionally, ordinary chemical, as opposed to nuclear, reactions conserve atomic (or, more properly, nuclear) identities. A statement of a reaction that conserves these basic quantities is said to be "balanced."

Additional quantities may be conserved, or partially conserved, during a reaction. If the reaction occurs without outside fields or collisions to exert torque, angular momentum will be conserved. Electronic and spin quantum numbers are partially conserved separately during reactions, giving rise to the Wigner-Witmer propensity rules for atomic and diatomic reactions: Electronic spin is conserved, while the projection of electronic angular momentum along the axis of bond formation or destruction is conserved. For example, the energetically allowed reaction of CH with N_2 in flames to form the very stable HCN molecule plus atomic nitrogen N is relatively slow. The radical CH is commonly in its spin-$\frac{1}{2}$ ground state, both N_2 and HCN are spin 0, and the reaction must produce N atoms in a spin-$\frac{3}{2}$ state. It may be fortuitous, but this reaction requires a thermal activation energy almost exactly the amount necessary to promote CH to its lowest spin-$\frac{3}{2}$ state.

Also during a reaction, nuclear permutation symmetry must be preserved. Diatomic hydrogen, for example, consists of almost independent populations of even rotational angular momentum *para* states and odd rotational angular momentum *ortho* states, having opposite nuclear permutation symmetry. If diatomic hydrogen is eliminated from a molecule, the nuclei retain the permutation symmetry they originally had within the molecule. These statistics can modify the rates at which reactions occur, such as exhibited in the intensity alternation of the vibrational bands occurring in the photodissociation of cold ammonia. Especially severe for spin-0 nuclei, for which some rotational states may be wholly absent, this effect of nuclear permutation symmetry may account for isotopic enrichment, such as naturally occurs in ozone (O_3) in the stratosphere.

Less restrictive, but still effective, are the reaction propensity rules that result from the partial conservation of molecular orbital symmetry elements, such as node planes. For example, the Woodward-Hoffmann rules for organic reactions explain, among other things, why the Diels-Alder reaction involving six π-bonded carbons is facile,

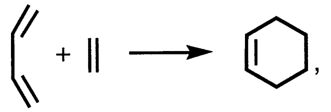

while the analogous reactions forming four and eight carbon rings are difficult:

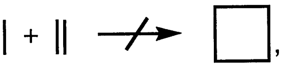

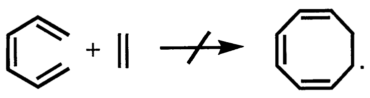

In the first case, where bonds form, the phase of the highest occupied molecular orbital of butadiene matches the phase of the lowest unoccupied molecular orbital on ethylene.

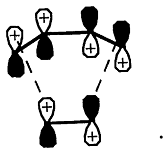

In the latter cases, where the reaction does not occur, the molecular orbital wave functions meet out of phase, inhibiting bond formation.

4.3 Theories of Reaction Rates

Reaction rate theory is discussed in Chemical Kinetics. Bimolecular processes are the subject of Molecular and Atomic Collision Processes. Aspects of dynamical modeling appear in Simulation by Molecular Dynamics, while relevant mathematical topics include Monte-Carlo Methods and Chaotic Phenomena.

Except when the adiabatic approximation fails or quantum mechanical tunneling occurs, the dynamics of reactions are dominated by classical motion on adiabatic surfaces. Over the years, modeling of chemical reactions has revealed what properties of adiabatic surfaces influence reactions. Energy barriers are most critical, acting as walls, requiring activation of the molecules undergoing reactions. Placement of these barriers is also important: An energy barrier requires energy stored in the nuclear coordinates normal to it in order to surmount it.

Integration of the equations of motion over many nuclear coordinates is a formidable task; even when done, what is required for comparison with experiments is often just a thermally averaged rate constant. Usually, the rate of a reaction step can be expressed according to

$$\text{Rate} = k(T)[X]^a[Y]^b[Z]^c \ldots, \tag{15}$$

where factors of the form $[X]$ represent the concentrations of molecules of type X entering the reaction step. The thermal rate constant is usually parameterized in an Arrhenius form,

$$k(T) = AT^q \exp(E^*/kT), \tag{16}$$

where A, q, and E^* are "constants." Roughly, E^* corresponds to the "height" of an energy barrier, while A is an "entropic" factor.

The Arrhenius form receives partial justification from statistical rate theories. Almost all of these assume that those molecular configurations corresponding to the minimum activation energy barrier are in statistical equilibrium with all the other reactant configurations. The fraction of molecules at the energy barrier is determined by statistical mechanics, or thermodynamics. These "activated" molecules in the *transition state* then proceed to leave this configuration, some passing to products, at a rate given by a "frequency factor." If statistical mechanical densities of states are used to calculate the fraction of activated molecules, one arrives at theories such as the Rice-Ramsperger-Kassels (RRK) and Rice-Ramsperger-Kassels-Marcus (RRKM) theories for gas-phase reactions. If one uses thermodynamic functions such as free energy in place of energy, including possibly solvation effects, one arrives at activated-complex theory that is useful in condensed phases. If reactions are considered in extremely viscous fluids, the frequency factor can be modified to take into account diffusion, as in Kassel's theory of diffusion-controlled reactions. All of these exhibit the basic Arrhenius form for reaction at a barrier.

Reactions that proceed without a barrier can be influenced greatly by long-range interactions and often do not show the typical Arrhenius form (Clary, 1990).

4.4 Reaction Steps

An overall molecular reaction mechanism may proceed as a sequence, or network, of fundamental reaction steps. For each step, the number of participating molecules determines its *molecularity*. *Unimolecular* reaction steps involve only one reactant molecule, in which nuclei or groups of nuclei either rearrange (*isomerize*) or separate (*dissociate*). *Bimolecular, termolecular,* and higher-order reaction steps involve first the association of two, three, or more molecules in a fundamental reaction step. More details can be found under CHEMICAL REACTIONS.

4.5 Mechanisms

4.5.1 Elimination Elimination splits a molecule into two or more parts. An example of elimination is the formation of highly reactive methylene radicals by the elimination of CO from ketene,

$CH_2CO \rightarrow CH_2 + CO.$

4.5.2 Addition The opposite of elimination, addition combines two or more molecules into a third. An example is the addition of a methylene radical into the double bond of an ethylene molecule,

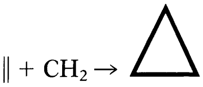

4.5.3 Substitution An elimination may occur after, or in concert with, an addition. Thus, the replacement of bromide at an optically active carbon by hydroxide may proceed through an addition of hydroxide first, then elimination of the bromide. In this case, the optically active carbon "inverts," but retains its chirality.

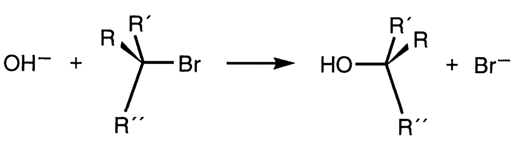

An elimination may also precede an addition in an overall substitution. Most ionic reactions in solution occur this way. In the above example of bromine replacement at a carbon, if the elimination of bromine instead occurs first, the carbon center loses its optical activity, since the hydroxide can now add to either side of the carbon, and a racemic mixture results:

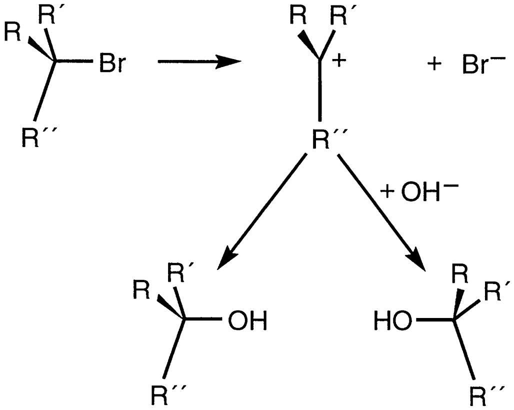

Clues like this help chemists piece together the order of steps in a reaction mechanism.

4.5.4 Rearrangement The most exotic reactions involve often surprising rearrangements of atoms from one isomer to another. Most of the "name reactions" of chemistry involve rearrangements. These are a joy to some chemists and a bane to everyone else, since they prevent chemical synthesis from being simply a matter of adding, removing, or substituting atoms or groups. An example of rearrangement is the hydrolysis of neopentylbromide (1-bromo-2,2-dimethylpropane). Here the neopentane carbon skeleton rearranges to one of isopentane:

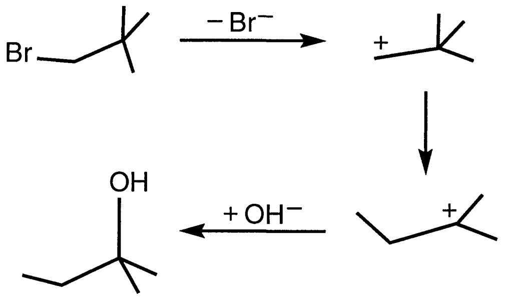

4.5.5 Catalysis A catalyst has the property that it speeds up a reaction mechanism, without itself being consumed in the overall reaction. However, a catalyst must actually participate in the fundamental reaction steps. A catalytic molecule first reacts but is subsequently regenerated. An important example of this process is the destruction of stratospheric ozone by the catalytic action of chlorine atoms in the presence of light. This reaction also illustrates the possibility that catalytic intermediates may exist, and may even be isolated, in some catalytic mechanisms:

$$\begin{array}{c} O_3 + Cl\cdot \rightarrow O_2 + \cdot OCl \\ \cdot OCl + h\nu \rightarrow O: + Cl\cdot \\ O_3 + O: \rightarrow 2O_2 \\ \hline 2O_3 + Cl\cdot + h\nu \rightarrow 3O_2 + Cl\cdot \end{array}.$$

The major effect of a catalyst on a reaction is a lowering of the activation energy necessary to cause the reaction. Because the activation energy appears in the exponent of an Arrhenius expression for a single reaction step, reduction in activation energy may greatly increase a reaction rate, making up for the additional complexity that catalysis brings to the reaction mechanism.

Most of the critical steps of biological reactions are catalyzed by enzymes. Further information may be found in CATALYSIS and PROTEINS AND ENZYMES.

4.6 Processes

More information on a number of molecular processes can be found in COMBUSTION, EXPLOSIVES, PHOTOCHEMISTRY, RADIOCHEMISTRY, LASER PHOTOCHEMISTRY, ELECTROCHEMISTRY, POLYMERS MOLECULAR STRUCTURE OF and POLYMER REACTIONS. Relevant topics include CVD (CHEMICAL VAPOR DEPOSITION) and PLASMA ETCHING.

GLOSSARY

Adiabatic Approximation: A specific method of separating nuclear and electronic motions in which the electronic motion is first solved for each set of nuclear positions; nuclear positions enter as parameters during the solution of the electronic motion, becoming dynamical variables during the subsequent solution of the nuclear motions.

Adiabatic Potential: The energy obtained by solving the electronic motion at fixed nuclear positions and adding to the electronic energy the classical nuclear repulsion energy. ("Adiabatic" does not mean "without heat" in this context; it means moving the nuclei "slowly," without inducing transitions between different resulting electronic states.)

Bimolecular: Involving two molecules at once.

Born-Oppenheimer Approximation: Separation of nuclear and electronic motion, now usually in the sense of adiabatic, rather than complete, separation of motion.

Chemical Bond: An attractive interaction between two nuclei, generally requiring more than one electron volt to break, and involving an electron pair.

Chiral: Possessing "handedness"; not superimposable on its mirror image.

Conformer: Relatively stable configuration of atoms of a specified molecular formula. Possibly transient.

Coordinate Covalent Bond: A covalent bond formed with a pair of electrons contributed by one of the two bonding atoms.

Covalent Bond: A bond formed by sharing pairs of electrons between two atoms.

Crystal: A spatially repeating array of atoms or ions.

Electronegativity: The tendency of an atom, or group of atoms, to attract electrons within a compound.

Empirical Formula: A formula denoting the relative proportions of elements in a compound.

Enantiomer: Either of a pair of isomers whose structures are mirror images of each other; a chiral molecule. Differing enantiomers are typically designated by the letters D- or L- (*dextro* or *levo*) according to their effect on the plane of polarization of light or, if known, by the actual spatial orientations of their nuclei.

Feasible Group: A subgroup of the permutation-inversion group of a molecule that includes transformations between specified "feasible" conformers and the point groups of those conformers.

Geometric Isomer: A stereoisomer that differs from another by not being superimposable on that stereoisomer or its mirror image.

Ionic Bond: Electrostatic attraction of two ions.

Isomers: Different, isolatable molecules corresponding to the same molecular formula.

Lewis Structure: A depiction of bonding in which electrons are represented by dots surrounding elemental atoms.

Mole: The amount of matter that contains the number of particles (atoms, molecules, . . .) equal to the number of atoms in 1 gram of carbon-12; an Avogadro's number $N_A \approx 6.022 \times 10^{23}$ of particles.

Molecular Formula: A formula denoting the types and numbers of atoms in a molecule.

Molecularity: The number of molecules involved at once in a reaction.

Molecular Orbital: An electron orbital extending over one or more atoms.

Monomer: A molecule that can combine with others to form a polymer.

Optical Isomer: Enantiomer. (Named for the ability to rotate the plane of polarization of light.)

Permutation-Inversion Group: The symmetry group of a molecule involving all permutations of identical nuclei and inversion.

Point Group: The symmetry group of a molecular conformer involving proper and improper rigid rotations that transform the wave function but leave the reference positions of the nuclei fixed.

Polymer: A large chain or network molecule made by combining many similar, smaller molecules (monomers).

Racemic Mixture: A mixture of pairs of enantiomers in equal proportions.

Radical: A relatively stable group of atoms, either as part of a molecule or free. A radical generally has one or more unpaired electrons.

Skeletal Structure (Diagram): A diagram showing the bonds between atoms. Names of some atoms, such as carbon, are often omitted, and other atoms, such as hydrogen, are often implied.

Structural Formula: A molecular formula indicating the relative positions of atoms.

Tautomer: A rapidly interchanging conformer.

Termolecular: Pertaining to three molecules at once.

Topological Isomer: An isomer that differs from another by having a different bond topology.

Stereoisomer: An isomer that differs from another by having the same bond topology but a different spatial arrangement of atoms. *Cis* and *trans* geometric isomers are one example; D and L enantiomers are another example.

Works Cited

Bamzai, A. S., Deb, B. M. (1981), *Rev. Mod. Phys.* **53**, 95–126.

Banks, J. (1967), *Naming Organic Compounds, a Programmed Introduction to Organic Chemistry*, Philadelphia: Saunders.

Bawendi, M. G., Steigerwald, M. L., Brus, L. E. (1990), *Annu. Rev. Phys. Chem.* **41**, 477–496.

Beilstein, F. (1908, *et seq.*) [Beilstein Institute, Frankfurt], *Beilstein's Handbuch der Organischen Chemie*, 4th ed., Berlin: Springer-Verlag [Basic Series 1908 through Fifth Suppl. 1984 and continuing].

Benfry, O. T. (1966), *The Names and Structures of Organic Compounds*, Malabar, FL: Krieger.

Bishop, D. M. (1990), *Rev. Mod. Phys.* **62**, 343–374.

Block, B. P. (1990), *Inorganic Chemical Nomenclature*, Washington D.C.: American Chemical Society.

Born, M. (1931), *Gött. Nachr. Math Phys. Kl.* 1.

Born, M., Huang, K. (1951), *Dynamical Theory of Crystal Lattices*, Oxford: Clarendon.

Born, M., Oppenheimer, R. (1927), *Ann. Phys. (Leipzig)* **84**, 457–484.

Buckingham, A. D., Fowler, P. W., Hutson, J. M. (1988), *Chem. Rev.* **88**, 963–988.

Bunker, P. R. (1978), *Molecular Symmetry and Spectroscopy*, New York: Academic Press.

Chase, M. W., Jr., Davies, C. A., Downey, J. R., Jr., Frurip, D. J., McDonald, R. A., Syverrud, A. N. (1985), *J. Phys. Chem. Ref. Data* **14**, Suppl. 1.

Clary, D. C. (1990), *Annu. Rev. Phys. Chem.* **41**, 61–90.

Deb, B. M. (1973), *Rev. Mod. Phys.* **45**, 22–43.

Deslattes, R. D. (1980), *Annu. Rev. Phys. Chem.* **31**, 435–461.

Gmelin, L. (1922, *et seq.*), *Gmelin Handbook of Inorganic Chemistry*, 8th ed., Berlin: Springer-Verlag.

Gordy, W., Smith, W. V., Trambarulo, R. F. (1953), *Microwave Spectroscopy*, New York: Wiley.

Herzberg, G. (1945), *Infrared and Raman Spectra of Polyatomic Molecules*, New York: Van Nostrand.

Herzberg, G. (1950), *Spectra of Diatomic Molecules*, New York: Van Nostrand.

Herzberg, G. (1966), *Electronic Spectra and Electronic Structure of Polyatomic Molecules*, New York: Van Nostrand.

Hirota, E. (1985), *High-Resolution Spectroscopy of Transient Molecules*, Berlin: Springer-Verlag.

Hirshfelder, J. O., Curtis, C. F., Bird, R. B. (1954), *Molecular Theory of Gases and Liquids*, New York: Wiley.

Huber, K. P., Herzberg, G. (1979), *Constants of Diatomic Molecules*, New York: Van Nostrand Reinhold.

IUPAC, (1979), *Nomenclature of Organic Chemistry*, New York: Pergamon.

IUPAC (1990), *Nomenclature of Inorganic Chemistry, Recommendations 1990*, Oxford: Blackwell Scientific Publications.

Jacox, M. E. (1984), *J. Phys. Chem. Ref. Data* **13**, 945–1068.

Jacox, M. E. (1988), *J. Phys. Chem. Ref. Data* **17**, 269–511.

Jacox, M. E. (1990), *J. Phys. Chem. Ref. Data* **19**, 1387–1546.

Kotani, M., Ohno, K., Kayama, K. (1961), "Quantum Mechanics of Electronic Structure of Simple Molecules," in S. Flügge (Ed.), *Handbuch der Physik*, Vol. XXXVII/2, Berlin: Springer-Verlag, pp. 1–172.

Kuchitsu, K. (Ed.) (1992), *Structure Data of Free Polyatomic Molecules; Landolt-Börnstein New Series II/21*, Berlin: Springer-Verlag.

Lias, S. G., Bartmess, J. E., Liebman, J. F., Holmes, J. L., Levin, R. D., Mallard, W. G. (1988), *J. Chem. Phys. Ref. Data* **17**, Suppl. 1.

Lide, D. R. (Ed.) (1993), *Handbook of Chemistry and Physics*, 74th ed., Boca Raton: CRC Press.

Nelson, R. D., Jr., Lide, D. R., Jr., Maryott, A. A. (1967), *Selected Values of Electric Dipole Moments for Molecules in the Gas Phase*, National Standard Reference Data Series, National Bureau of Standards, No. 10, Washington, DC: U.S. G.P.O.

Noolandi, J. (1992), *Annu. Rev. Phys. Chem.* **43**, 237–236.

Ostwald, W. (1900, 1922), *Grundlinien der anorganishchen Chemie* [1900; 5th ed. 1922].

Parr, R. G. (1964), *Quantum Theory of Molecular Electronic Structure*, New York: Benjamin.

Pauling, L. (1960), *The Nature of the Chemical Bond*, Ithaca, N.Y.: Cornell University Press.

Payne, M. C., Teter, M. P., Allan, M. C., Arias, T. A., Joannopoulos, J. D. (1992), *Rev. Mod. Phys.* **64**, 1045–1097.

Platt, J. R. (1961), "The Chemical Bond and the Distribution of Electrons in Molecules," in: S. Flügge (Ed.), *Handbuch der Physik*, Vol. XXXVII/2, Berlin: Springer-Verlag., pp. 173–281.

Richards, W. G., Trivedi, H. P., Cooper, D. L. (1981), *Spin-Orbit Coupling in Molecules*, Oxford: Clarendon Press.

Schatz, G. C. (1988), *Annu. Rev. Phys. Chem.* **39**, 317–340.

Scoles, G. (1980), *Annu. Rev. Phys. Chem.* **31**, 81–96.

Shimanouchi, T. (1972), *Tables of Molecular Vibrational Frequencies. Consolidated Volume 1*, National Standard Reference Data Series, National Bureau of Standards, No. 39, Washington, DC: U.S. G.P.O.

Shimanouchi, T. (1977), *J. Chem. Phys. Ref. Data* **6**, 993–1102.

Shimanouchi, T. (1978), *J. Chem. Phys. Ref. Data* **7**, 1323–1443.

Shimanouchi, T. (1980), *J. Chem. Phys. Ref. Data* **9**, 1149–1254.

Stoneham, A. M., Harding, J. H. (1986), *Annu. Rev. Phys. Chem.* **37**, 53–80.

Townes, C. H., Schawlow, A. L. (1955), *Microwave Spectroscopy*, New York: McGraw-Hill.

Tranham, J. G. (1980), *Organic Nomenclature, A Programed Introduction*, Englewood Cliffs: Prentice-Hall.

Vilkov, L. V., Mastryukov, V. S., Sadova, N. I. (1983), *Determination of the Geometrical Structure of Free Molecules*, Moscow: Mir.

Wagman, D. D., Evans, W. H., Parker, V. B., Schumm, R. H., Halow, I., Bailey, S. M., Churney, K. L., Nuttall, R. L. (1982), *J. Phys. Chem. Ref. Data* **11**, Suppl. 2.

Wilson, E. B., Jr., Decius, J. C., Cross, P. C. (1955), *Molecular Vibrations*, New York: McGraw-Hill.

Further Reading

Atkins, P. W. (1987), *Molecules*, New York: Freeman.

Bunker, P. R. (1978), *Molecular Symmetry and Spectroscopy*, New York: Academic Press.

Burdett, J. K. (1980), *Molecular Shapes*, New York: Wiley.

Cotton, F. A., Wilkinson, G. (1980), *Advanced Inorganic Chemistry*, New York: Wiley-Interscience.

Deb, B. M. (1973), *Rev. Mod. Phys.* **45**, 22–43.

Flygare, W. H. (1978), *Molecular Structure and Dynamics*, Englewood Cliffs, NJ: Prentice-Hall.

Gray, H. B., Ballhausen, C. J. (1964), *Molecular Orbital Theory*, New York: Benjamin.

Gardener, W. C., Jr. (Ed.) (1984), *Combustion Chemistry*, Berlin: Springer-Verlag.

Hirshfelder, J. O., Curtis, C. F., Bird, R. B. (1954), *Molecular Theory of Gases and Liquids*, New York: Wiley.

Johnston, H. S. (1966), *Gas Phase Reaction Rate Theory*, New York: Ronald Press.

Oxtoby, D. W., Nachtrieb, N. H., Freeman, W. A. (1990), *Chemistry: Science of Change*, Philadelphia: Saunders.

Setser, D. W. (1979), *Reactive Intermediates in the Gas Phase*, New York: Academic Press.

Steinfeld, J. I., Fransisco, J. S., Hase, W. L. (1989), *Chemical Kinetics and Dynamics*, Englewood Cliffs, NJ: Prentice-Hall.

Stryer, L. (1981), *Biochemistry*, 2nd ed., San Francisco: Freeman.

Vilkov, L. V., Mastryukov, V. S., Sadova, N. I. (1983), *Determination of the Geometrical Structure of Free Molecules*, Moscow: Mir.

Wilson, E. B., Jr., Decius, J. C., Cross, P. C. (1955), *Molecular Vibrations*, New York: McGraw-Hill.

MONTE-CARLO METHODS

K. Binder, *Institut für Physik, Johannes-Gutenberg-Universität Mainz, Mainz, Germany*

Introduction and Overview .. 568

1. **Random-Number Generation** 568
1.1 General Introduction 568
1.2 Properties that a Random-Number Generator Should Have 569
1.3 Comments about a Few Frequently Used Generators 569

2. **Simple Sampling of Probability Distributions Using Random Numbers** 570
2.1 Numerical Estimation of Known Probability Distributions 570
2.2 "Importance Sampling" versus "Simple Sampling" .. 570
2.3 Monte Carlo as a Method of Integration 570
2.4 Infinite Integration Space ... 571
2.5 Random Selection of Lattice Sites 571
2.6 The Self-Avoiding Walk Problem 571
2.7 Simple Sampling versus Biased Sampling: the Example of SAWs Continued 572

3. **Survey of Applications to Simulation of Transport Processes** 573
3.1 The "Shielding Problem" 573
3.2 Diffusion-Limited Aggregation (DLA) 573

4. **Monte-Carlo Methods in Statistical Thermodynamics: Importance Sampling** 574
4.1 The General Idea of the Metropolis Importance Sampling Method 574
4.2 Comments on the Formulation of a Monte-Carlo Algorithm 574
4.3 The Dynamic Interpretation of the Monte-Carlo Method .. 576
4.4 Monte-Carlo Study of the Dynamics of Fluctuations Near Equilibrium and of the Approach toward Equilibrium 577
4.5 The Choice of Statistical Ensembles 578

5. **Accuracy Problems: Finite-Size Problems, Dynamic Correlation of Errors, Boundary Conditions** 578
5.1 Finite-Size–Induced Rounding and Shifting of Phase Transitions 578
5.2 Different Boundary Conditions; Simulation of Surfaces and Interfaces 579
5.3 Estimation of Statistical Errors 580

6. **Quantum Monte-Carlo Techniques** 580
6.1 General Remarks 580
6.2 Path-Integral Monte-Carlo Methods 581
6.3 An Application Example: the Momentum Distribution of Fluid ^{4}He 582
6.4 A Few Qualitative Comments on Fermion Problems 582

7. **Lattice Gauge Theory** 583
7.1 Some Basic Ideas of Lattice Gauge Theory 583
7.2 A Recent Application 584

8. **Selected Comments on Applications in Classical Statistical Mechanics of Condensed-Matter Systems** .. 585
8.1 Metallurgy and Materials Science 585
8.2 Polymer Science 585
8.3 Surface Physics 587

9. **Concluding Remarks** 587

Glossary 588

Works Cited 588

Further Reading 590

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION AND OVERVIEW

Many problems in science are very complex: e.g., statistical thermodynamics considers thermal properties of matter resulting from the interplay of a large number of elementary particles. A deterministic description in terms of the equation of motion of all these particles would make no sense, and a probabilistic description is required. A probabilistic description may even be intrinsically implied by the quantum-mechanical nature of the basic processes (e.g., emission of neutrons in radioactive decay) or because the problem is incompletely characterized, only some degrees of freedom being considered explicitly while the others act as a kind of background causing random noise. While thus the concept of probability distributions is ubiquitous in physics, often it is not possible to compute these probability distribution functions analytically in explicit form, because of the complexity of the problem. For example, interactions between atoms in a fluid produce strong and nontrivial correlations between atomic positions, and, hence, it is not possible to calculate these correlations analytically.

Monte-Carlo methods now aim at a numerical estimation of probability distributions (as well as of averages, that can be calculated from them), making use of (pseudo) random numbers. By "pseudorandom numbers" one means a sequence of numbers produced on a computer with a deterministic procedure from a suitable "seed." This sequence, hence, is not truly random—see Sec. 1 for a discussion of this problem.

The outline of the present article is as follows. Since all Monte-Carlo methods heavily rely on the use of random numbers, we briefly review random-number generation in Sec. 1. In Sec. 2, we then elaborate on the discussion of "simple sampling," i.e., problems where a straightforward generation of probability distributions using random numbers is possible. Section 3 briefly mentions some applications to transport problems, such as radiation shielding, and growth phenomena, such as "diffusion-limited aggregation" (DLA).

Section 4 then considers the importance sampling methods of statistical thermodynamics, including the use of different thermodynamic ensembles. This article emphasizes applications in statistical mechanics of condensed matter, since this is the field where most activity with Monte-Carlo methods occurs.

Some more practical aspects important for the implementation of algorithms and the judgment of the tractability of simulation approaches to physical problems are then considered in Sec. 5: effects resulting from the finite size of simulation boxes, effects of choosing various boundary conditions, dynamic correlation of errors, and the application to studies of the dynamics of thermal fluctuations.

The extension to quantum-mechanical problems is mentioned in Sec. 6 and the application to elementary particle theory (lattice gauge theory) in Sec. 7. Section 8 then illustrates some of the general concepts with a variety of applications taken from condensed matter physics, while Sec. 9 contains concluding remarks.

We do not discuss problems of applied mathematics such as applications to the solution of linear operator equations (Fredholm integral equations, the Dirichlet boundary-value problem, eigenvalue problems, etc.); for a concise discussion of such problems, see, e.g., Hammersley and Handscomb (1964). Nor do we discuss simulations of chemical kinetics, such as polymerization processes (see, e.g., Bruns et al. 1981).

1. RANDOM-NUMBER GENERATION

1.1 General Introduction

The precise definition of "randomness" is a problem in itself (see, e.g., Compagner, 1991) and is outside the scope of the present article. Truly random numbers are unpredictable in advance and must be produced by an appropriate physical process such as radioactive decay. Series of such numbers have been documented but would be very inconvenient to use for Monte-Carlo simulations, and usually their total number is too limited anyhow. Thus, we do not discuss them here any further.

Pseudorandom numbers are produced in the computer by one of several simple algorithms, some of which will be discussed below, and thus are predictable, as their sequence is exactly reproducible. (This reproducibility, of course, is a desirable property, as it allows detailed checks of Monte-Carlo simulation programs.) They are thus not truly random, but they have statistical properties (nearly uniform distribution, nearly vanishing correlation coefficients, etc.) that are very similar to the statistical properties of truly random numbers. Thus, a given sequence of (pseudo)random numbers appears "random" for many practical purposes. In the following, the prefix "pseudo" will be omitted throughout.

1.2 Properties That a Random-Number Generator Should Have

What one needs are numbers that are uniformly distributed in the interval [0,1] and that are uncorrelated. By "uncorrelated" we not only mean vanishing pair correlations for arbitrary distances along the random-number sequence, but also vanishing triplet and higher correlations. No algorithm exists that satisfies these desirable requirements fully, of course; the extent to which the remaining correlations between the generated random numbers lead to erroneous results of Monte-Carlo simulations has been a matter of long-standing concern (Knuth, 1969; James, 1990); even random-number generators that have passed all common statistical tests and have been used successfully for years may fail for a new application, in particular if it involves a new type of Monte-Carlo algorithm (see, e.g., Ferrenberg *et al.*, 1992, for a recent example). Therefore, the testing of random-number generators is a field of research in itself (see, e.g., Marsaglia, 1985; Compagner and Hoogland, 1987).

A limitation due to the finite word length of computers is the *finite period*: Every generator begins after a long but finite period to produce exactly the same sequence again. For example, simple generators for 32-bit computers have a maximum period of 2^{30} ($\approx 10^9$) numbers only. This is not enough for recent high-quality applications! Of course, one can get around this problem (Knuth, 1969; James, 1990), but, at the same time, one likes the code representing the random-number generator to be "portable" (i.e., in a high-level programing language like FORTRAN usable for computers from different manufactures) and "efficient" (i.e., extremely fast so it does not unduly slow down the simulation program as a whole). Thus, inventing new random-number generators that are in certain respects a better compromise between these partially conflicting requirements is still of interest (e.g., Marsaglia *et al.*, 1990).

1.3 Comments about a Few Frequently Used Generators

Best known is the linear multiplicative algorithm (Lehmer, 1951), which produces random integers X_i recursively from the formula

$$X_i = aX_{i-1} + c \ (\text{modulo } m), \tag{1}$$

which means that m is added when the results otherwise were negative. For 32-bit computers, $m = 2^{31} - 1$ (the largest integer that can be used for that computer). The integer constants a, c, and X_0 (the starting value of the recursion, the so-called "seed") need to be appropriately chosen {e.g., $a = 16\,807$, $c = 0$, X_0 odd}. Obviously, the "randomness" of the X_i results since, after a few multiplications with a, the result would exceed m and hence is truncated, and so the leading digits of X_i are more or less random. But there are severe correlations: If d-tuples of such numbers are used to represent points in d-dimensional space having a lattice structure, they lie on a certain number of hyperplanes (Marsaglia, 1968).

Equation (1) produces random numbers between 0 and m. Converting them into real numbers and dividing by m yields random numbers in the interval [0,1], as desired.

More popular now are shift-register generators (Tausworthe, 1965; Kirkpatrick and Stoll, 1981), based on the formula

$$X_i = X_{i-p} \cdot \text{XOR} \cdot X_{i-q}, \tag{2}$$

where $\cdot$XOR$\cdot$ is the bitwise "exclusive or" operation, and the "lags" p, q have to be properly chosen [the popular "R250" (Kirkpatrick and Stoll, 1981) uses $p = 109$, $q = 250$ and thus needs 250 initializing integers]. "Good" generators based on Eq. (2) have fewer correlations between random numbers than those resulting from Eq. (1), and much larger period.

A third type of generators is based on Fibonacci series and also recommended in the literature (Knuth, 1969; Ahrens and Dieter, 1979; James, 1990). But a general recommendation is that every user of random numbers should not rely on their quality blindly, and should perform his own tests in the context of his application.

2. SIMPLE SAMPLING OF PROBABILITY DISTRIBUTIONS USING RANDOM NUMBERS

In this section, we give a few nearly trivial examples of the use of Monte-Carlo methods, which will be useful for the understanding of later sections. More material on this subject can be found in standard textbooks like Koonin (1981) and Gould and Tobochnik (1988).

2.1 Numerical Estimation of Known Probability Distributions

A known probability distribution p_i that a (discrete) state i occurs with $1 \leq i \leq n$, with $\Sigma_{i=1}^{n} p_i = 1$, is numerically realized using random numbers uniformly distributed in the interval from zero to unity: defining $P_i = \Sigma_{j=1}^{i} p_i$, we choose a state i if the random number ζ satisfies $P_{i-1} \leq \zeta < P_i$, with $P_0 = 0$. In the limit of a large number (M) of trials, the generated distribution approximates p_i, with errors of order $1/\sqrt{M}$.

Monte-Carlo methods in statistical mechanics can be viewed as an extension of this simple concept to the probability that a point $\mathbf{X}$ in phase space occurs,

$$P_{\text{eq}}(\mathbf{X}) = (1/Z)\exp\{-\mathcal{H}(\mathbf{X})/k_{\text{B}}T\},$$

k_{B} being Boltzmann's constant, T the absolute temperature, and $Z = \Sigma_{\mathbf{X}} \exp\{-\mathcal{H}(\mathbf{X})/k_{\text{B}}T\}$ the partition function, although in general neither Z nor $P_{\text{eq}}(\mathbf{X})$ can be written explicitly (as function of the variables of interest, such as T, particle number N, volume V, etc.). The term $\mathcal{H}(\mathbf{X})$ denotes the Hamiltonian of the (classical) system.

2.2 "Importance Sampling" versus "Simple Sampling"

The sampling of the Boltzmann probability $P_{\text{eq}}(\mathbf{X})$ by Monte-Carlo methods is not completely straightforward: One must not choose the points $\mathbf{X}$ in phase space completely at random, since $P_{\text{eq}}(\mathbf{X})$ is extremely sharply peaked. Thus, one needs "importance sampling" methods which generate points $\mathbf{X}$ preferably from the "important" region of space where this narrow peak occurs.

Before we treat this problem of statistical mechanics in more detail, we emphasize the more straightforward applications of "simple sampling" techniques. In the following, we list a few problems for which simple sampling is useful. Suppose one wishes to generate a configuration of a randomly mixed crystal of a given lattice structure, e.g., a binary mixture of composition A_xB_{1-x}. Again, one uses pseudorandom numbers ζ uniformly distributed in [0,1] to choose the occupancy of lattice sites $\{j\}$: If $\zeta_j < x$, the site j is taken by an A atom, and else by a B atom. Such configurations now can be used as starting point for a numerical study of the dynamical matrix, if one is interested in the phonon spectrum of mixed crystals. One can study the distribution of sizes of "clusters" formed by neighboring A atoms if one is interested in the "site percolation problem" (Stauffer, 1985), etc.

If one is interested in simulating transport processes such as diffusion, a basic approach is the generation of simple random walks. Such random walks, resulting from addition of vectors whose orientation is random, can be generated both on lattices and in the continuum. Such simulations are desirable if one wishes to consider complicated geometries or boundary conditions of the medium where the diffusion takes place. Also, it is straightforward to include competing processes (e.g., in a reactor, diffusion of neutrons in the moderator competes with loss of neutrons due to nuclear reactions, radiation going to the outside, etc., or gain due to fission events). Actually, this problem of reactor criticality (and related problems for nuclear weapons!) was the starting point for the first large-scale applications of Monte-Carlo methods by Fermi, von Neumann, Ulam, and their co-workers (Hammersley and Handscomb, 1964).

2.3 Monte Carlo as a Method of Integration

Many Monte-Carlo computations may be viewed as attempts to estimate the value of a (multiple) integral. To give the flavor of this idea, let us discuss the one-dimensional integral

$$I = \int_0^1 f(x)dx$$
$$\equiv \int_0^1 \int_0^1 g(x,y)dxdy, \qquad \text{with}$$

$$g(x,y) = \begin{cases} 0 & \text{if } f(x) < y, \\ 1 & \text{if } f(x) \geq y, \end{cases} \tag{3}$$

as an example (suppose, for simplicity, that also $0 \leq f(x) \leq 1$ for $0 \leq x \leq 1$). Then I simply is interpreted as the fraction of the unit square $0 \leq x,y \leq 1$ lying underneath the curve $y = f(x)$. Now a straightforward (though often not very efficient) Monte-Carlo estimation of Eq. (3) is the "hit or miss" method: We take n points (ζ_x,ζ_y) uniformly distributed in the unit square $0 \leq \zeta_x \leq 1$, $0 \leq \zeta_y \leq 1$. Then I is estimated by

$$\bar{g} = \frac{1}{n}\sum_{i=1}^{n} g(\zeta_{xi},\zeta_{yi}) = \frac{n^*}{n}, \tag{4}$$

n^* being the number of points for which $f(\zeta_{xi}) \geq \zeta_{yi}$. Thus, we count the fraction of points that lie underneath the curve $y = f(x)$. Of course, such Monte-Carlo integration methods are inferior to many other techniques of numerical integration, if the integration space is low dimensional, but the situation is worse for high-dimensional integration spaces: For any method using a regular grid of points for which the integrand needs to be evaluated, the number of points sampled along each coordinate is $M^{1/d}$ in d dimensions, which is small for any reasonable sample size M if d is very large.

2.1 Infinite Integration Space

Not always is the integration space limited to a bounded interval in space. For example, the ϕ^4 model of field theory considers a field variable $\phi(\mathbf{x})$, where $\mathbf{x}$ is drawn from a d-dimensional space and $\phi(\mathbf{x})$ is a real variable with distribution

$$P(\phi) \propto \exp[-\alpha(-\tfrac{1}{2}\phi^2 + \tfrac{1}{4}\phi^4)]; \qquad \alpha > 0. \tag{5a}$$

While $-\infty < \phi < +\infty$, the distribution $P'(y)$

$$P'(y) = \int_{-\infty}^{y} P(\phi)d\phi \bigg/ \int_{-\infty}^{+\infty} P(\phi)d\phi \tag{5b}$$

varies in the unit interval, $0 \leq P' \leq 1$. Hence, defining $Y = Y(P')$ as the inverse function of $P'(y)$, we can choose a random number ζ uniformly distributed between zero and one, to obtain $\phi = Y(\zeta)$ distributed according to the chosen distribution $P(\phi)$. Of course, this method works not only for the example chosen in Eq. (5) but for any distribution of interest. Often it will not be possible to obtain $Y(P')$ analytically, but then one can compute numerically a table before the start of the sampling (Heermann, 1986).

2.5 Random Selection of Lattice Sites

A problem that occurs very frequently (e.g., in solid-state physics) is that one considers a large lattice (e.g., a model of a simple cubic crystal with $N = L_x \times L_y \times L_z$ sites), and one wishes to select a lattice site (n_x,n_y,n_z) at random. This is trivially done using the integer arithmetics of standard computers, converting a uniformly distributed random number ζ_x $(0 \leq \zeta_x < 1)$ to an integer n_x with $1 \leq n_x \leq L_x$ via the statement $n_x = \text{int}(\zeta_x L_x + 1)$. This is already an example where one must be careful, however, when three successive pseudorandom numbers drawn from a random-number generator (RNG) are used for this purpose: If one uses a RNG with bad statistical qualities, the frequency with which individual sites are visited may deviate distinctly from a truly random choice. In unfavorable cases, successive pseudorandom numbers are so strongly correlated that certain lattice sites would be never visited!

2.6 The Self-Avoiding Walk Problem

As an example of the straightforward use of simple sampling techniques, we now discuss the study of self-avoiding walks (SAWs) on lattices (which may be considered as a simple model for polymer chains in good solvents; see Kremer and Binder, 1988). Suppose one considers a square or simple cubic lattice with coordination number (number of nearest neighbors) z. Then, for a random walk (RW) with N steps, we would have $Z_{RW} = z^N$

configurations, but many of these random walks intersect themselves and thus would, not be self-avoiding. For SAWs, one only expects of the order of $Z_{SAW} = \text{const.} \times N^{\gamma-1}z_{\text{eff}}^N$ configurations, where $\gamma > 1$ is a characteristic exponent (which is not known exactly for $d = 3$ dimensions), and $z_{\text{eff}} \leq z - 1$ is an effective coordination number, which also is not known exactly. But it is already obvious that an exact enumeration of all configurations would be possible for rather small N only, while most questions of interest refer to the behavior for large N; e.g., one wishes to study the end-to-end distance of the SAW,

$$\langle R^2\rangle_{\text{SAW}} = \frac{1}{Z_{\text{SAW}}}\sum_{\mathbf{X}} [\mathbf{R}(\mathbf{X})]^2, \tag{6}$$

the sum being extended over all configurations of SAWs which we denote formally as points $\mathbf{X}$ in phase space. One expects that $\langle R^2\rangle_{\text{SAW}} \propto N^{2\nu}$, where ν is another characteristic exponent. A Monte-Carlo estimation of $\langle R^2\rangle_{\text{SAW}}$ now is based on generating a sample of only $M \ll Z_{\text{SAW}}$ configurations $\mathbf{X}_l$, i.e.:

$$\overline{R^2} = \frac{1}{M}\sum_{l=1}^{M} [\mathbf{R}(\mathbf{X}_l)]^2 \approx \langle R^2\rangle_{\text{SAW}}. \tag{7}$$

If the M configurations are statistically independent, standard error analysis applies, and we expect that the relative error behaves as

$$\frac{\overline{(\delta R^2)^2}}{(\overline{R^2})^2} \approx \frac{1}{M-1}\left[\frac{\langle R^4\rangle_{\text{SAW}}}{\langle R^2\rangle_{\text{SAW}}^2} - 1\right]. \tag{8}$$

While the law of large numbers then implies that $\overline{R^2}$ is Gaussian distributed around $\langle R^2\rangle_{\text{SAW}}$ with a variance determined by Eq. (8), one should note that the variance does not decrease with increasing N. Statistical mechanics tells us that fluctuations decrease with increasing number N of degrees of freedom; i.e., one equilibrium configuration differs in its energy $E(\mathbf{x})$ from the average $\langle E\rangle$ only by an amount of order $1/\sqrt{N}$. This property is called "self-averaging." Obviously, such a property is not true for $\langle R^2\rangle_{\text{SAW}}$. This "lack of self-averaging" is easy to show already for ordinary random walks (Binder and Heermann, 1988).

2.7 Simple Sampling versus Biased Sampling: the Example of SAWs Continued

Apart from this problem, that the accuracy of the estimation of R^2 does not increase with the number of steps of the walk, it is also not easy to generate a large sample of configurations of SAWs for large N. Suppose we do this at each step by choosing one of $z - 1$ neighbors at random (eliminating from the start immediate reversals, which would violate the SAW condition). Whenever the chosen lattice site is already taken, we also would violate the SAW condition, and the attempted walk is terminated. Now the fraction of walks that will continue successfully for N steps will only be of the order of $Z_{\text{SAW}}/(z-1)^N \propto [z_{\text{eff}}/(z-1)]^N N^{\gamma-1}$, which decreases to zero exponentially $\{\propto \exp(-N\mu)$ with $\mu = \ln[(z-1)/z_{\text{eff}}]$ for large $N\}$; this failure of success in generating long SAWs is called the "attrition problem."

The obvious recipe, to select at each step only from among the lattice sites that do not violate the SAW restriction, does not give equal statistical weight for each configuration generated, of course, and so the average would not be the averaging that one needs in Eq. (6). One finds that this method would create a "bias" toward more compact configurations of the walk. But one can calculate the weights of configurations $w(\mathbf{X})$ that result in this so-called "inversely restricted sampling" (Rosenbluth and Rosenbluth, 1955), and in this way correct for the bias and estimate the SAW averages as

$$\overline{R^2} = \left\{\sum_{l=1}^{M} [w(\mathbf{X}_l)]^{-1}\right\}^{-1} \sum_{l=1}^{M} [w(\mathbf{X}_l)]^{-1}[\mathbf{R}(\mathbf{X}_l)]^2. \tag{9}$$

However, error analysis of this biased sampling is rather delicate (Kremer and Binder, 1988).

A popular alternative to overcome the above attrition problem is the "enrichment technique," founded on the principle "Hold fast to that which is good." Namely, whenever a walk attains a length that is a multiple of s steps without intersecting itself, n independent attempts to continue it (rather than a single attempt) are made. The numbers n, s are fixed, and, if we choose $n \approx \exp(\mu s)$, the numbers of walks of various lengths generated will be approximately equal. Enrichment has the advantage over inversely restricted sampling that all walks of a given length have equal weights, while the weights in Eq. (9) vary over many orders of magnitude for large N. But the disadvantage is, on the other hand, that the linear dimensions of the walks are highly correlated, since some of them have many steps in common! For these reasons, simple sampling and its extensions are useful only for a small fraction of problems in polymer science, and now importance sampling (Sec. 4) is much more used. But we emphasize that related problems are encountered for the sampling of "random surfaces" (this problem arises in the field theory of quantum gravity), in path-integral Monte-Carlo treatments of quantum problems, and in many other contexts.

3. SURVEY OF APPLICATIONS TO SIMULATION OF TRANSPORT PROCESSES

The possibilities to simulate the random motions of particles are extremely widespread. Therefore, it is difficult to comment about such problems in general. Thus, we rather prefer again to proceed by briefly discussing a few examples that illustrate the spirit of the approach.

3.1 The "Shielding Problem"

A thick shield of absorbing material is exposed to γ radiation (energetic photons), of specified distribution of energy and angle of incidence. We want to know the intensity and energy distribution of the radiation that penetrates that shield.

The level of description is here that one may generate a lot of "histories" of those particles traveling through the medium. The paths of these γ particles between scattering events are straight lines, and different γ particles do not interact with each other. A particle with energy E, instantaneously at the point $\mathbf{r}$ and traveling in the direction of the unit vector $\mathbf{w}$, continues to travel in the same direction with the same energy, until a scattering event with an atom of the medium occurs. The standard assumption is that the atoms of the medium are distributed randomly in space. Then the total probability that the particle will collide with an atom while traveling a length δs of its path is $\sigma_c(E)\delta s$, $\sigma_c(E)$ being the cross section. In a region of space where $\sigma_c(E)$ is constant, the probability that a particle travels without collision a distance s is $F_c(s) = 1 - \exp[-\sigma_c(E)s]$. If a collision occurs, it may lead to absorption or scattering, and the cross sections for these types of events are assumed to be known.

A Monte-Carlo solution now simply involves the tracking of simulated particles from collision to collision, generating the distances s that the particles travel without collision from the exponential distribution quoted above. Particles leaving a collision point are sampled from the appropriate conditional probabilities as determined from the respective differential cross sections. For increasing sampling efficiency, many obvious tricks are known. For example, one may avoid losing particles by absorption events: If the absorption probability (i.e., the conditional probability that absorption occurs given that a collision has occurred) is α, one may replace $\sigma_c(E)$ by $\sigma_c(E)(1 - \alpha)$, and allow only scattering to take place with the appropriate relative probability. Special methods for the shielding problem have been extensively developed and already have been reviewed by Hammersley and Handscomb (1964).

3.2 Diffusion-Limited Aggregation (DLA)

Diffusion-limited aggregation is a model for the irreversible formation of random aggregates by diffusion of particles, which get stuck at random positions on the already formed object if they hit its surface in the course of their diffusion [sec Vicsek (1989), Meakin (1988), and Herrmann (1986, 1992) for detailed reviews of this problem and related phenomena]. Many problems (shapes of snowflakes, size distribution of asteroids, roughness of crack surfaces, etc.) can be understood as the end product of similar random dynamical and irreversible growth processes. Diffusion-limited aggregation is just one example of them. It may be simulated by iterating the following steps: From a randomly selected position on a spherical surface of radius R_m that encloses the aggregate (that has already been grown in the previous steps, its center of gravity being in the center of the sphere), a particle of unit mass is launched to start a simple random-walk trajectory. If it touches the aggregate, it sticks irreversibly on its surface. After the particle has either stuck or moved a distance R_f from the center of the aggregate such that it is unlikely that it will hit in the future, a new particle is launched. Ideally one would like to have $R_f \rightarrow \infty$, but, in practice, $R_f = 2R_m$ is sufficient. By this irreversible aggregation of particles, one forms fractal clusters. That means that the dimension d_f characterizing the relation between the mass of the grown object and its (gyration) radius R, $M \propto \mathrm{R}^{d_f}$, is less than the dimension d of space in which the growth takes place. Again there are some tricks to make such simulations more efficient: For example, one may allow the particles to jump over larger steps when they travel in empty regions. From such studies, researchers have found that $d_f = 1.715 \pm 0.004$ for DLA in $d = 2$, while $d_f = 2.485 \pm 0.005$ in $d = 3$ (Tolman and Meakin, 1989). Such exponents as yet cannot be analytically predicted.

4. MONTE-CARLO METHODS IN STATISTICAL THERMODYNAMICS: IMPORTANCE SAMPLING

4.1 The General Idea of the Metropolis Importance Sampling Method

In the canonical ensemble (see STATISTICAL MECHANICS, CLASSICAL), the average of an observable $A(\mathbf{X})$ takes the form

$$\langle A \rangle = \frac{1}{Z} \int_{\Omega} d^k X A(\mathbf{X}) \exp\left[-\frac{\mathcal{H}(\mathbf{X})}{k_{\mathrm{B}}T} \right], \qquad (10)$$

where Z is the partition function,

$$Z = \int_{\Omega} d^k X \exp\left[-\frac{\mathcal{H}(\mathbf{X})}{k_{\mathrm{B}}T} \right], \qquad (11)$$

Ω denoting the (k-dimensional) volume of phase space $\{\mathbf{X}\}$ over which is integrated, $\mathcal{H}(\mathbf{X})$ being the (classical) Hamiltonian. For this problem, a simple sampling analog to Sec. 2 would not work: The probability distribution $p(\mathbf{X}) = (1/Z) \exp[-\mathcal{H}(\mathbf{X})/k_{\mathrm{B}}T]$ has a very sharp peak in phase space where all extensive variables $A(\mathbf{X})$ are close to their average values $\langle A \rangle$. This peak may be approximated by a Gaussian centered at $\langle A \rangle$, with a relative half-width of order $1/\sqrt{N}$ only, if we consider a system of N particles. Hence, for a practically useful method, one cannot sample the phase space uniformly, but the points $\mathbf{X}_\nu$ must be chosen preferentially from the important region of phase space, i.e., the vicinity of the peak of this probability distribution. This goal is achieved by the importance sampling method (Metropolis *et al.*, 1953): Starting from some initial configuration $\mathbf{X}_1$, one constructs a sequence of configurations $\mathbf{X}_\nu$ defined in terms of a transition probability $W(\mathbf{X}_\nu \rightarrow \mathbf{X}'_\nu)$ that rules stochastic "moves" from an old state $\mathbf{X}_\nu$ to a new state $\mathbf{X}'_\nu$, and, hence, one creates a "random walk through phase space." The idea of that method is to choose $W(\mathbf{X} \rightarrow \mathbf{X}')$ such that the probability with which a point $\mathbf{X}$ is chosen in this process converges toward the canonical probability

$$P_{\mathrm{eq}}(\mathbf{X}) = \left(\frac{1}{Z}\right) \exp\left[-\frac{\mathcal{H}(\mathbf{X})}{k_{\mathrm{B}}T} \right]$$

in the limit where the number M of states $\mathbf{X}$ generated goes to infinity. A condition sufficient to ensure this convergence is the so-called principle of detailed balance,

$$P_{\mathrm{eq}}(\mathbf{X}) W(\mathbf{X} \rightarrow \mathbf{X}') = P_{\mathrm{eq}}(\mathbf{X}') W(\mathbf{X}' - \mathbf{X}). \qquad (12)$$

For a justification that Eq. (12) actually yields this desired convergence, we refer to Hammersley and Handscomb (1964), Binder (1976), Heermann (1986), and Kalos and Whitlock (1986). In this importance sampling technique, the average Eq. (10) then is estimated in terms of a simple arithmetic average,

$$\bar{A} = \frac{1}{M - M_0} \sum_{\nu = M_0 + 1}^{M} A(\mathbf{X}_\nu). \qquad (13)$$

Here it is anticipated that it is advantageous to eliminate the residual influence of the initial configuration $\mathbf{X}_1$ by eliminating a large enough number M_0 of states from the average. [The judgment of what is "large enough" is often difficult; see Binder (1976) and Sec. 5.3 below.] It should also be pointed out that this Metropolis method can be used for sampling any distribution $P(\mathbf{X})$: One simply must choose a transition probability $W(\mathbf{X} \rightarrow \mathbf{X}')$ that satisfies a detailed balance condition with $P(\mathbf{X})$ rather than with $P_{\mathrm{eq}}(\mathbf{X})$.

4.2 Comments on the Formulation of a Monte-Carlo Algorithm

What is now meant in practice by the transition $\mathbf{X} \to \mathbf{X}'$? Again there is no general answer to this question; the choice of the process may depend both on the model under consideration and the purpose of the simulation. Since Eq. (12) implies that $W(\mathbf{X} - \mathbf{X}')/W(\mathbf{X}' \to \mathbf{X}) = \exp(-\delta\mathcal{H}/k_BT)$, $\delta\mathcal{H}$ being the energy change caused by the move from $\mathbf{X} \to \mathbf{X}'$, typically it is necessary to consider small changes of the state $\mathbf{X}$ only. Otherwise the absolute value of the energy change $|\delta\mathcal{H}|$ would be rather large, and then either $W(\mathbf{X} \to \mathbf{X}')$ or $W(\mathbf{X}' \to \mathbf{X})$ would be very small. Then it would be almost always forbidden to carry out that move, and the procedure would be poorly convergent. For example, in the lattice gas model at constant particle number, a transition $\mathbf{X} \to \mathbf{X}'$ may consist of moving one particle to a randomly chosen neighboring site. In the lattice gas at constant chemical potential, one removes (or adds) just one particle at a time, which is isomorphic to single flips in the Ising model of anisotropic magnets.

Another arbitrariness concerns the order in which the particles are selected for considering a move. Often one chooses to select them in the order of their labels (in the simulation of a fluid or lattice gas at constant particle number) or to go through the lattice in a regular typewriter-type fashion (in the case of spin models, for instance). For lattice systems, it may be convenient to use sublattices (e.g., the "checkerboard algorithm," where the white and black sublattices are updated alternatively, for the sake of an efficient "vectorization" of the program; see Landau, 1992). An alternative is to choose the lattice sites (or particle numbers) randomly. The latter procedure is somewhat more time consuming, but it is a more faithful representation of a dynamic time evolution of the model described by a master equation (see below).

It is also helpful to realize that often the transition probability $W(\mathbf{X} \to \mathbf{X}')$ can be written as a product of an "attempt frequency" times an "acceptance frequency." By clever choice of the attempt frequency, it is possible sometimes to attempt large moves and still have a high acceptance, and thus make the computations more efficient.

For spin models on lattices, such as Ising or Potts models, *XY* and Heisenberg ferromagnets, etc., algorithms have been devised where one does not update single spins in the move $\mathbf{X} \to \mathbf{X}'$, but, rather, one updates specially constructed clusters of spins (see Swendsen *et al.*, 1992, for a review). These algorithms have the merit that they reduce critical slowing down, which hampers the efficiency of Monte-Carlo simulations near second-order phase transitions. "Critical slowing down" means a dramatic increase of relaxation times at the critical point of second-order phase transitions, and these relaxation times also control statistical errors in Monte-Carlo simulations, as we shall see in Sec. 5. Since these "cluster algorithms" work for rather special models only, they will not be discussed further here. But further development of such algorithms is an important area of current research (e.g., Barkema and Marko, 1993).

There is also some arbitrariness in the choice of the transition probability $W(\mathbf{X} \to \mathbf{X}')$ itself. The original choice of Metropolis *et al.* (1953) is

$$W(\mathbf{X} \to \mathbf{X}') = \begin{cases} \exp(-\delta\mathcal{H}/k_BT) & \text{if } \delta\mathcal{H} > 0, \\ 1 & \text{otherwise.} \end{cases} \tag{14}$$

An alternative choice is the so-called "heat-bath method." There one assigns the new value α_i' of the ith local degree of freedom in the move from $\mathbf{X}$ to $\mathbf{X}'$ irrespective of what the old value α_i' was. One thereby considers the local energy $\mathcal{H}_i(\alpha_i')$ and chooses the state α_i' with probability

$$\exp[-\mathcal{H}_i(\alpha_i')/k_BT] \Big/ \sum_{\{\alpha_i''\}} \exp[-\mathcal{H}_i(\alpha_i'')/k_BT].$$

We now outline the realization of the sequence of states $\mathbf{X}$ with chosen transition probability W. At each step of the procedure, one performs a *trial move* $\alpha_i \to \alpha_i'$, computes $W(\mathbf{X} \to \mathbf{X}')$ for this trial move, and compares it with a random number η, uniformly distributed in the interval $0 < \eta < 1$. If $W < \eta$, the trial move is rejected, and the old state (with α_i) is counted once more in the average, Eq. (13). Then another trial is made. If $W > \eta$, on the other hand, the trial move is accepted, and the new configuration thus generated is taken into account in the average, Eq. (13). It serves then also as a starting point of the next step.

Since subsequent states $\mathbf{X}_\nu$ in this Markov chain differ by the coordinate α_i of one particle only (if they differ at all), they are highly correlated. Therefore, it is not straightforward to estimate the error of the average, Eq. (13). Let us assume for the moment that, after n steps, these correlations have died out. Then we may estimate the statistical error δA of the estimate $\bar{A}$ from the standard formula,

$$\overline{(\delta A)^2} = \frac{1}{m(m-1)} \sum_{\mu=\mu_0}^{m+\mu_0-1} [A(\mathbf{X}_\mu) - \bar{A}]^2, \quad m \gg 1, \tag{15}$$

where the integers μ_0, μ, m are defined by $m = (M - M_0)/n$, μ_0 labels the state $\nu = M_0 + 1$, $\mu = \mu_0 + 1$ labels the state $\nu = M_0 + 1 + n$, etc. Then also $\bar{A}$ for consistency should be taken as

$$\bar{A} = \frac{1}{m} \sum_{\mu=\mu_0}^{m+\mu_0-1} A(\mathbf{X}_\mu). \tag{16}$$

If the computational effort of carrying out the "measurement" of $A(\mathbf{X}_\mu)$ in the simulation is rather small, it is advantageous to keep taking measurements every Monte-Carlo step per degree of freedom but to construct block averages over n successive measurements, varying n until uncorrelated block averages are obtained.

4.3 The Dynamic Interpretation of the Monte-Carlo Method

It is not always easy to estimate the appropriate number of configurations M_0 after which the correlations to the initial state $\mathbf{X}_1$, which typically is a state far from equilibrium, have died out, nor is it easy to estimate the number n between steps after which correlations in equilibrium have died out. A formal answer to this problem, in terms of relaxation times of the associated master equation describing the Monte-Carlo process, is discussed in the next section. This interpretation of Monte-Carlo sampling in terms of master equations is also the basis for Monte-Carlo studies of the dynamics of fluctuations near thermal equilibrium, and is discussed now. One introduces the probability $P(\mathbf{X},t)$ that a state $\mathbf{X}$ occurs at time t. This probability then decreases by all moves $\mathbf{X} \to \mathbf{X}'$, where the system reaches a neighboring state $\mathbf{X}'$; on the other hand, inverse processes $\mathbf{X}' \to \mathbf{X}$ lead to a gain of probability. Thus, one can write down a rate equation, similar to chemical kinetics, considering the balance of all gain and loss processes:

$$\begin{aligned} \frac{d}{dt} P(\mathbf{X},t) = &-\sum_{\mathbf{X}'} W(\mathbf{X} \to \mathbf{X}') P(\mathbf{X},t) \\ &+ \sum_{\mathbf{X}'} W(\mathbf{X}' \to \mathbf{X}) P(\mathbf{X}',t). \end{aligned} \tag{17}$$

The Monte-Carlo sampling (i.e., the sequence of generated states $\mathbf{X}_1 \to \mathbf{X}_2 \to \ldots \to \mathbf{X}_\nu \to \ldots$) can hence be interpreted as a numerical realization of the master equation, Eq. (17), and then a "time" t is associated with the index ν of subsequent configurations. In a system with N particles, we may normalize the "time unit" such that N single-particle moves are attempted per unit time. This is often called a "sweep" or "1 Monte-Carlo step (MCS)."

For the thermal equilibrium distribution $P_{\text{eq}}(\mathbf{X})$, because of the detailed balance principle, Eq. (12), there is no change of probability with time, $dP(\mathbf{X},t)/dt = 0$; thus, thermal equilibrium arises as the stationary solution of the master equation, Eq. (17). Thus, it is also plausible that Markov processes described by Eq. (17) describe a relaxation that always leads toward thermal equilibrium, as desired.

Now, for a physical system (whose trajectory in phase space, according to classical statistical mechanics, follows from Newton's laws of motion), it is clear that the stochastic trajectory through phase space that is described by Eq. (17) in general has nothing to do with the actual dynamics. For example, Eq. (17) never describes any propagating waves (such as spin waves in a magnet, or sound waves in a crystal or fluid, etc.).

In spite of this observation, the dynamics of the Monte-Carlo "trajectory" described by Eq. (17) sometimes does have physical significance. In many situations, one does not wish to consider the full set of dynamical variables of the system, but rather a subset only: For instance, in an interstitial alloy where one is interested in modeling the diffusion processes, the diffusion of the interstitials may be modeled by a stochastic hopping between the available lattice sites. Since the mean time between two successive jumps is orders of magnitude larger than the time scale of atomic vibrations in the solid, the phonons can be reasonably well approximated as a heat bath, as far as the diffusion is concerned.

There are many examples where such a separation of time scales for different degrees of freedom occurs: For example, for a description of the Brownian motion of polymer chains in polymer melts, the fast bond-angle and bond-length vibrations may be treated as heat bath, etc. As a rule of thumb, any very slow relaxation phenomena (kinetics of nucleation, decay of remanent magnetization in spin glasses, growth of ordered domains in adsorbed monolayers at surfaces, etc.) can be modeled by Monte-Carlo methods. Of course, one must pay attention to building in relevant conservation laws into the model properly (e.g., in an interstitial alloy, the overall concentration of interstitials is conserved; in a spin glass, the magnetization is not conserved) and to choosing microscopically reasonable elementary steps representing the move $\mathbf{X} \to \mathbf{X}'$. The great flexibility of the Monte-Carlo method, where one can choose the level of the modeling appropriately for the model at hand and identify the degrees of freedom that one wishes to consider, as well as the type and nature of transitions between them, is a great advantage and thus allows complementary applications to more atomistically realistic simulation approaches such as the molecular dynamics (*q.v.*) method where one numerically integrates Newton's equations of motion (Heermann, 1986; Ciccotti and Hoover, 1986; Hockney and Eastwood, 1988). By a clever combination with cluster-flip algorithms, one sometimes can construct very efficient algorithms and hence span a very broad range of time scales (Barkema and Marko, 1993).

4.4 Monte-Carlo Study of the Dynamics of Fluctuations near Equilibrium and of the Approach toward Equilibrium

Accepting Eq. (17), the average in Eq. (13) then is simply interpreted as a time average along the stochastic trajectory in phase space,

$$\bar{A} = \frac{1}{t_M - t_{M_0}} \int_{t_{M_0}}^{t_M} A(t)dt, \qquad t_M = M/N, \quad t_{M_0} = M_0/N. \tag{18}$$

It is thus no surprise that, for the importance-sampling Monte-Carlo method, one needs to consider carefully the problem of ergodicity: Time averages need not agree with ensemble averages. For example, near first-order phase transitions there may be long-lived metastable states. Sometimes the considered moves do not allow one to reach all configurations (e.g., in dynamic Monte-Carlo methods for self-avoiding walks; see Kremer and Binder, 1988).

One can also define time-displaced correlation functions: $\langle A(t)B(0)\rangle$, where A, B stand symbolically for any physical observables, is estimated by

$$\overline{A(t)B(0)} = \frac{1}{t_M - t - t_{M_0}} \int_{t_{M_0}}^{t_M - t} A(t + t')B(t')dt', \qquad t_M - t > t_{M_0}. \tag{19}$$

Equation (19) refers to a situation where t_{M_0} is chosen large enough such that the system has relaxed toward equilibrium during the time t_{M_0}; then the pair correlation depends on the difference t and not the two individual times t', $t' + t$ separately.

However, it is also interesting to study the nonequilibrium relaxation process by which equilibrium is approached. In this region, $A(t) - \bar{A}$ is systematically dependent on the observation time t, and an ensemble average $\langle A(t)\rangle_T - \langle A(\infty)\rangle_T$ [$\lim_{t\to\infty} \bar{A} = \langle A\rangle_T \equiv \langle A(\infty)\rangle_T$ if the system is ergodic] is nonzero. One may define

$$\langle A(t)\rangle_T = \sum_{\{\mathbf{X}\}} P(\mathbf{X},t)A(\mathbf{X}) = \sum_{\{\mathbf{X}\}} P(\mathbf{X},0)A(\mathbf{X}(t)). \tag{20}$$

In the second step of this equation, the fact was used that the ensemble average involved is actually an average weighted by $P(\mathbf{X},0)$ over an ensemble of initial states $\mathbf{X}(t = 0)$, which then evolve as described by the master equation, Eq. (17). In practice, Eq. (20) means an average over a large number $n_{\text{run}} \gg 1$ statistically independent runs,

$$[\bar{A}(t)]_{av} = \frac{1}{n_{\text{run}}} \sum_{l=1}^{n_{\text{run}}} A(t,l), \tag{21}$$

where $A(t,l)$ is the observable A observed at time t in the lth run of this nonequilibrium Monte-Carlo averaging.

Many concepts of nonequilibrium statistical mechanics can immediately be introduced in such simulations. For instance, one can introduce arbitrary "fields" that can be switched off to study the dynamic response functions, for both linear and nonlinear response (Binder, 1984).

4.5 The Choice of Statistical Ensembles

While so far the discussion has been (implicitly) restricted to the case of the canonical ensemble (*NVT* ensemble, for the case of a fluid), it is sometimes useful to use other statistical ensembles. Particularly useful is the grand canonical ensemble (μVT), where the chemical potential μ rather than the particle number N is fixed. In addition to moves where the configuration of particles in the box relaxes, one has moves where one attempts to add or remove a particle from the box.

In the case of binary (*AB*) mixtures, a useful variation is the "semi–grand canonical" ensemble, where $\Delta_\mu = \mu_A - \mu_B$ is held fixed and moves where an A particle is converted into a B particle (or vice versa) are considered, in an otherwise identical system configuration.

The isothermal-isobaric (*NpT*) ensemble, on the other hand, fixes the pressure, and then volume changes $V \rightarrow V' = V + \Delta V$ need to be considered (rescaling properly the positions of the particles).

It also is possible to define artificial ensembles that are not in the textbooks on statistical mechanics. An example is the so-called Gaussian ensemble (which interpolates between the canonical and microcanonical ensembles and is useful for the study of first-order phase transitions, as the so-called "multicanonical ensemble"). Particularly useful is the "Gibbs ensemble," where one considers the equilibrium between two simulation boxes (one containing liquid, the other gas), which can exchange both volume ΔV and particles (ΔN), while the total volume and total particle number contained in the two boxes are held fixed. The Gibbs ensemble is widely used for the simulation of gas-fluid coexistence, avoiding interfaces (Panagiotopoulos, 1992; Levesque and Weis, 1992).

A simulation at a given state point (*NVT*) contains information not only on averages at that state point but also on neighboring state points (*NVT'*), via suitable reweighting of the energy distribution $P_N(E)$ with a factor $\exp(E/k_BT)\exp(-E/k_BT')$. Such "histogram methods" are particularly useful near critical points (Swendsen *et al.*, 1992).

5. ACCURACY PROBLEMS: FINITE-SIZE PROBLEMS, DYNAMIC CORRELATION OF ERRORS, BOUNDARY CONDITIONS

5.1 Finite-Size–Induced Rounding and Shifting of Phase Transitions

A prominent application of Monte-Carlo simulation in statistical thermodynamics and lattice theory is the study of phase transitions. Now it is well known in statistical physics that sharp phase transitions can occur in the thermodynamic limit only, $N \rightarrow \infty$. Of course, this is no practical problem in everyday life—even a small water droplet freezing into a snowflake contains about $N = 10^{18}$ H_2O molecules, and, thus, the rounding and shifting of the freezing are on a relative scale of $1/\sqrt{N} = 10^{-9}$ and thus completely negligible. But the situation is different for simulations, which often consider extremely small systems (e.g., a hypercubic d-dimensional box with linear dimensions L, $V = L^d$, and periodic boundary conditions), where only $N \sim 10^2$ to 10^4 particles are involved.

In such small systems, phase transitions are strongly rounded and shifted (Barber, 1983; Binder, 1987, 1992a; Privman, 1990). Thus, care needs to be applied when simulated systems indicate phase changes. It turns out, however, that these finite-size effects can be used as a valuable tool to infer properties of the infinite system from the finite-size behavior. As a typical example, we discuss the phase transition of an Ising ferromagnet (Fig. 1), which has a second-order phase transition at a critical temperature T_c. For $L \rightarrow \infty$, the spontaneous magnetization M_{spont} vanishes according to a power law, $M_{\text{spont}} = B(1 - T/T_c)^\beta$, B being a critical amplitude and β a critical exponent (Stanley, 1971), and the susceptibility χ and correlation length ξ diverge,

$$\chi \propto |1 - T/T_c|^{-\gamma}, \qquad \xi \propto |1 - T/T_c|^{-\nu}, \qquad (22)$$

where γ, ν are the associated critical exponents. In a finite system, ξ cannot exceed L, and, hence, these singularities are smeared out.

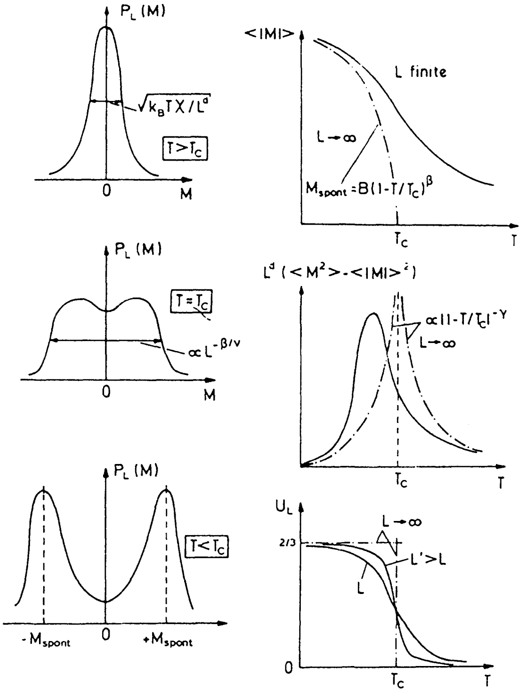

FIG. 1. Schematic evolution of the order-parameter probability distribution $P_L(M)$ from $T > T_c$ to $T < T_c$ (from above to below, left part), for an Ising ferromagnet (where M is the magnetization) in a box of volume $V = L^d$. The right part shows the corresponding temperature dependence of the mean order parameter $\langle|M|\rangle$, the susceptibility $k_BT\chi' = L^d(\langle M^2\rangle - \langle|M|\rangle^2)$, and the reduced fourth-order cumulant $U_L = 1 - \langle M^4\rangle/[3\langle M^2\rangle^2]$. Dash-dotted curves indicate the singular variation that results in the thermodynamic limit, $L \to \infty$.

Now finite-size scaling theory (Barber, 1983; Privman, 1990) implies that basically these finite-size effects can be understood from the principle that "L scales with ξ"; i.e., the order-parameter probability distribution $P_L(M)$ can be written (Binder, 1987, 1992a)

$$P_L = L^{\beta/\nu}\tilde{P}(L/\xi, ML^{\beta/\nu}), \tag{23}$$

where $\tilde{P}$ is a "scaling function." From Eq. (23), one immediately obtains the finite-size scaling relations for order parameter $\langle|M|\rangle$ and the susceptibility (defined from a fluctuation relation) by taking the moments of the distribution P_L:

$$\langle|M|\rangle = L^{-\beta/\nu}\tilde{M}(L/\xi), \tag{24}$$

$$k_BT\chi' = L^d(\langle M^2\rangle - \langle|M|\rangle^2) = L^{\gamma/\nu}\tilde{\chi}(L/\xi), \tag{25}$$

where $\tilde{M}$, $\tilde{\chi}$ are scaling functions that follow from $\tilde{P}$ in Eq. (23). At T_c where $\xi \to \infty$, we thus have $\chi' \propto L^{\gamma/\nu}$; from the variation of the peak height of χ' with system size, hence, the exponent γ/ν can be extracted.

The fourth-order cumulant U_L is a function of L/ξ only,

$$U_L \equiv 1 - \langle M^4\rangle/(3\langle M^2\rangle^2) = \tilde{U}(L/\xi). \tag{26}$$

Here $U_L \to 0$ for a Gaussian centered at M = 0, i.e., for $T > T_c$; $U_L \to \frac{2}{3}$ for the double-Gaussian distribution, i.e., for $T < T_c$; while $U_L = \tilde{U}(0)$ is a universal nontrivial constant for $T = T_c$. Cumulants for different system sizes hence intersect at T_c, and this can be used to locate T_c precisely (Binder, 1987, 1992a).

A simple discussion of finite-size effects at first-order transitions is similarly possible. There one describes the various phases that coexist at the first-order transition in terms of Gaussians if $L \gg \xi$ (note that ξ stays finite at the first-order transition). In a finite system, these phases can coexist not only right at the transition but over a finite parameter region. The weights of the respective peaks are given in terms of the free-energy difference of the various phases. From this phenomenological description, energy and order-parameter distributions and their moments can be worked out. Of course, this description applies only for long enough runs where the system jumps from one phase to the other many times, while for short runs where the systems stay in a single phase, one would observe hysteresis.

5.2 Different Boundary Conditions: Simulation of Surfaces and Interfaces

We now briefly mention the effect of various boundary conditions. Typically one uses periodic boundary conditions to study bulk properties of systems not obscured by surface effects. However, it also is possible to choose different boundary conditions to study surface effects deliberately; e.g., one may simulate thin films in a $L \times L \times D$ geometry with two free $L \times L$ surfaces and periodic boundary conditions otherwise. If the film thickness D is large enough, the two surfaces do not influence each other, and one can infer the properties of a semi-infinite system. One may choose special interactions near the free surfaces, apply surface "fields" (even if they cannot be applied in the laboratory, it may nevertheless be useful to study the response to them in the simulation), etc.

Sometimes the boundary conditions may stabilize interfaces in the system (e.g., in an Ising model for $T < T_c$ a domain wall between phases with opposite magnetization will be present, if we apply strong enough surface fields of opposite sign). Such interfaces also are often the object of study in simulation. It may be desirable to simulate interfaces without having the systems disturbed by free surfaces. In an Ising system, this may simply be done by choosing antiperiodic boundary conditions. Combining antiperiodic and staggered periodic boundary conditions, even tilted interfaces may be stabilized in the system. In all such simulations of systems containing interfaces one must keep in mind, however, that because of capillary-wave excitations, interfaces usually are very slowly relaxing objects, and often a major effort in computing time is needed to equilibrate them. A further difficulty (when one is interested in interfacial profiles) is the fact that the center of the interface is typically delocalized.

5.3 Estimation of Statistical Errors

We now return to the problem of judging the time needed for having reasonably small errors in Monte-Carlo sampling. If the subsequent configurations used were uncorrelated, we simply could use Eq. (15), but in the case of correlations we have rather

$$\begin{aligned}\langle(\delta A)^2\rangle &= \left\langle\left[\frac{1}{n}\sum_{\mu=1}^{n} A_\mu - \langle A\rangle\right]^2\right\rangle \\ &= \frac{1}{n}\left[[\langle A^2\rangle - \langle A\rangle^2 + 2\sum_{\mu=1}^{n}\left(1-\frac{\mu}{n}\right)\right. \\ &\quad \left.\times\,(\langle A_0 A_\mu\rangle - \langle A\rangle^2)\right]. \qquad (27)\end{aligned}$$

Now we remember that a time $t_\mu = \mu\delta t$ is associated with the Monte-Carlo process, δt being the time interval between two successive observations A_μ, $A_{\mu+1}$. Transforming the summation to a time integration yields

$$\begin{aligned}\langle(\delta A)^2\rangle &= \frac{1}{n}(\langle A^2\rangle - \langle A\rangle^2) \\ &\quad \times\left[1 + \frac{2}{\delta t}\int_0^{t_n}\left(1-\frac{t}{t_n}\right)\phi_A(t)dt\right], \qquad (28)\end{aligned}$$

where

$$\phi_A(t) \equiv \frac{\langle A(0)A(t)\rangle - \langle A\rangle^2}{\langle A^2\rangle - \langle A\rangle^2}.$$

Defining a relaxation time $\tau_A = \int_0^\infty dt\phi_A(t)$, one obtains for $\tau_A \ll n\delta t = \tau_{\text{obs}}$ (the observation time)

$$\begin{aligned}\langle(\delta A)^2\rangle &= \frac{1}{n}[\langle A^2\rangle - \langle A\rangle^2]\left(1 + \frac{2\tau_A}{\delta t}\right) \\ &\approx 2\left(\frac{\tau_A}{\tau_{\text{obs}}}\right)[\langle A^2\rangle - \langle A\rangle^2]. \qquad (29)\end{aligned}$$

In comparison with Eq. (15), the dynamic correlations inherent in a Monte-Carlo sampling as described by the master equation, Eq. (17), lead to an enhancement of the expected statistical error $\langle(\delta A)^2\rangle$ by a "dynamic factor" $1 + 2\tau_A/\delta t$ (sometimes also called the "statistical inefficiency").

This dynamic factor is particularly cumbersome near second-order phase transitions (τ_A diverges: critical slowing down) and near first-order phase transitions (τ_A diverges at phase coexistence, because of the large lifetime of metastable states). Thus, even if one is interested only in static quantities in a Monte-Carlo simulation, understanding the dynamics may be useful for estimating errors. Also the question of how many configurations (M_0) must be omitted at the start of the averaging for the sake of equilibrium [Eq. (18)] can be formally answered in terms of a nonlinear relaxation function

$$\phi^{(nl)}(t) = \frac{\langle A(t)\rangle_T - \langle A(\infty)\rangle_T}{\langle A(0)\rangle_T - \langle A(\infty)\rangle_T}$$

and its associated time $\tau_A^{(nl)} = \int_0^\infty \phi_A^{(nl)}(t)dt$ by the condition $t_{M_0} \gg \tau_A^{(nl)}$.

6. QUANTUM MONTE-CARLO TECHNIQUES

6.1 General Remarks

Development of Monte-Carlo techniques to study ground-state and finite-temperature properties of interacting quantum many-body systems is an active area of research (for reviews see Ceperley and Kalos, 1979; Schmidt and Kalos, 1984; Kalos, 1984; Berne and Thirumalai, 1986; Suzuki, 1986; Schmidt and Ceperley, 1992; De Raedt and von der Linden, 1992). These methods are of interest for problems such as the structure of nuclei (Carlson, 1988) and elementary particles (De Grand, 1992), superfluidity of ^{3}He and ^{4}He (Schmidt and Ceperley, 1992), high-T_c superconductivity (e.g., Frick *et al.*, 1990), magnetism (e.g., Reger and Young, 1988), surface physics (Marx *et al.*, 1993), etc. Despite this widespread interest, much of this research has the character of "work in progress" and hence cannot feature more prominently in the present article. Besides, there is not just one quantum Monte-Carlo method, but many variants exist: variational Monte Carlo (VMC), Green's-function Monte Carlo (GFMC), projector Monte Carlo (PMC), path-integral Monte Carlo (PIMC), grand canonical quantum Monte Carlo (GCMC), world-line quantum Monte Carlo (WLQMC), etc. Some of these (like VMC, GFMC) address ground-state properties, others (like PIMC) finite temperatures. Here only the PIMC technique will be briefly sketched, following Gillan and Christodoulos (1993).

6.2 Path-Integral Monte-Carlo Methods

We wish to calculate thermal averages for a quantum system and thus rewrite Eqs. (10) and (11) appropriately,

$$\langle A\rangle = \frac{1}{Z}\,\mathrm{Tr}\exp\left(-\frac{\hat{\mathcal{H}}}{k_{\mathrm{B}}T}\right)\hat{A},$$
$$Z = \mathrm{Tr}\exp\left(-\frac{\hat{\mathcal{H}}}{k_{\mathrm{B}}T}\right), \qquad (30)$$

using a notation that emphasizes the operator character of the Hamiltonian $\hat{\mathcal{H}}$ and of the quantity $\hat{A}$ associated with the variable A that we consider. For simplicity, we consider first a system of a single particle in one dimension acted on by a potential $V(x)$. Its Hamiltonian is

$$\hat{\mathcal{H}} = -\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + V(x). \qquad (31)$$

Expressing the trace in the position representation, the partition function becomes

$$Z = \int dx\langle x|\exp(-\hat{\mathcal{H}}/k_{\mathrm{B}}T)|x\rangle, \qquad (32)$$

where $|x\rangle$ is an eigenvector of the position operator. Writing $\exp(-\hat{\mathcal{H}}/k_{\mathrm{B}}T)$ formally as $[\exp(-\hat{\mathcal{H}}/k_{\mathrm{B}}TP)]^P$, where P is a positive integer, we can insert a complete set of states between the factors:

$$Z = \int dx_1 \ldots \int dx_P\langle x_1|\exp(-\hat{\mathcal{H}}/k_{\mathrm{B}}TP)|x_2\rangle$$
$$\langle x_2|\ldots|x_P\rangle\langle x_P|\exp(-\hat{\mathcal{H}}/k_{\mathrm{B}}TP)|x_1\rangle. \qquad (33)$$

For large P, it is a good approximation to ignore the fact that kinetic and potential energy do not commute. Hence, one gets

$$\langle x|\exp(-\hat{\mathcal{H}}/k_{\mathrm{B}}TP)|x'\rangle$$
$$\approx \left(\frac{k_{\mathrm{B}}TmP}{2\pi\hbar^2}\right)^{1/2}\exp\left[\frac{-k_{\mathrm{B}}TmP}{2\pi\hbar^2}(x-x')^2\right]$$
$$\times\exp\left\{\frac{-1}{2k_{\mathrm{B}}TP}[V(x)+V(x')]\right\}, \qquad (34)$$

and

$$Z \approx \left(\frac{k_{\mathrm{B}}TmP}{2\pi\hbar^2}\right)^{P/2}\int dx_1\ldots\int dx_P\exp\left\{-\frac{1}{k_{\mathrm{B}}T}\right.$$
$$\left.\times\left[\tfrac{1}{2}\sum_{s=1}^{P}\kappa(x_s-x_{s+1})^2 + P^{-1}\sum_{s=1}^{P}V(x_s)\right\}\right., \qquad (35)$$

where

$$\kappa \equiv (k_{\mathrm{B}}T/\hbar)^2 mP \qquad (36)$$

In the limit $P\to\infty$, Eq. (35) becomes exact. Apart from the prefactor, Eq. (35) is precisely the configurational partition function of a classical system of a ring polymer, consisting of P beads coupled by harmonic springs with spring constant κ. Each bead is under the action of a potential $V(x)/P$.

This approach is straightforwardly generated to a system of N interacting quantum particles—one ends up with a system of N classical cyclic "polymer" chains. As a result of this isomorphism, the Monte-Carlo method developed for simulating classical systems can be carried over to such quantum-mechanical problems, too. It is also easy to see that the system always behaves classically at high temperatures—κ gets very large, and then the cyclic chains contract essentially to a point, while at low temperatures they are spread out, representing zero-point motion. However, PIMC becomes increasingly difficult at low temperatures, since P has to be the larger the lower T: If σ is a characteristic distance over which the potential $V(x)$ changes, one must have $\hbar^2/m\sigma^2 \ll k_BTP$ in order that two neighbors along the "polymer chain" are at a distance much smaller than σ. In PIMC simulations, one empirically determines and uses that P beyond which the thermodynamic properties do not effectively change.

This approach can be generalized immediately to the density matrix $\rho(x - x') = \langle x|\exp(-\hat{\mathcal{H}}/k_BT)|x'\rangle$, while there are problems with the calculation of time-displaced correlation functions $\langle A(t)B(0)\rangle$, where t is now the true time (associated with the time evolution of states following from the Schrödinger equation, rather than the "time" of Sec. 4.3 related to the master equation).

The step leading to Eq. (34) can be viewed as a special case of the Suzuki–Trotter formula (Suzuki, 1986)

$$\exp(\hat{A} + \hat{B}) = \lim_{P\to\infty} [\exp(\hat{A}/P)\exp(\hat{B}/P)]^P, \tag{37}$$

which is also used for mapping d-dimensional quantum problems on lattices to equivalent classical problems (in $d + 1$ dimensions, because of the additional "Trotter direction" corresponding to the imaginary time direction of the path-integral).

6.3 An Application Example: the Momentum Distribution of Fluid ^{4}He

We now consider the dynamic structure factor $S(\mathbf{k},\omega)$, which is the Fourier transform of a time-displaced pair correlation function of the density at a point $\mathbf{r}_1$ at time t_1 and the density at point $\mathbf{r}_2$ at time t_2 [$\hbar\mathbf{k}$ being the momentum transfer and $\hbar\omega$ the energy transfer of an inelastic scattering experiment by which one can measure $S(\mathbf{k},\omega)$]. In the "impulse approximation," the dynamic structure factor $S(\mathbf{k},\omega)$ can be related to the Fourier transform of the single-particle density matrix $\rho_1(\mathbf{r})$, which for ^{4}He can be written in terms of the wave function $\psi(\mathbf{r})$ as $\rho_1(\mathbf{r}) = \langle\psi^+(\mathbf{r}' + \mathbf{r})\psi(\mathbf{r})\rangle$. This relation is

$$S(k,\omega) \propto J(Y) = \frac{1}{\pi}\int_0^\infty \rho_1(r)\cos(Yr)dr,$$

where $Y \equiv m(\omega - k^2/2m)/k$ (West, 1975). Since $\mathbf{S}(k,\omega)$ has been measured via neutron scattering (Sokol *et al.*, 1989), a comparison between experiment and simulation can be performed without adjustable parameters (Fig. 2). Thus, the PIMC method yields accurate data in good agreement with experiment.

The studies of ^{4}He have also yielded qualitative evidence for superfluidity (Ceperley and Pollock, 1987). For a quantitative analysis of the λ transition, a careful assessment of finite-size effects (Fig. 1) is needed since one works with very small particle numbers (of the order of 10^2 ^{4}He atoms only). This has not been possible so far.

6.4 A Few Qualitative Comments on Fermion Problems

Particles obeying Fermi–Dirac statistics (such as electrons or ^{3}He, for instance) pose particular challenges to Monte-Carlo simulation. If one tries to solve the Schrödinger equation of a many-body system by Monte-Carlo methods, one exploits its analogy with a diffusion equation (Ceperley and Kalos, 1979; Kalos, 1984). As mentioned at various places in this article, diffusion processes correspond to random walks and are hence accessible to Monte-Carlo simulation. However, while the diffusion equation (for one particle) considers the probability $P(\mathbf{r},t)$ that

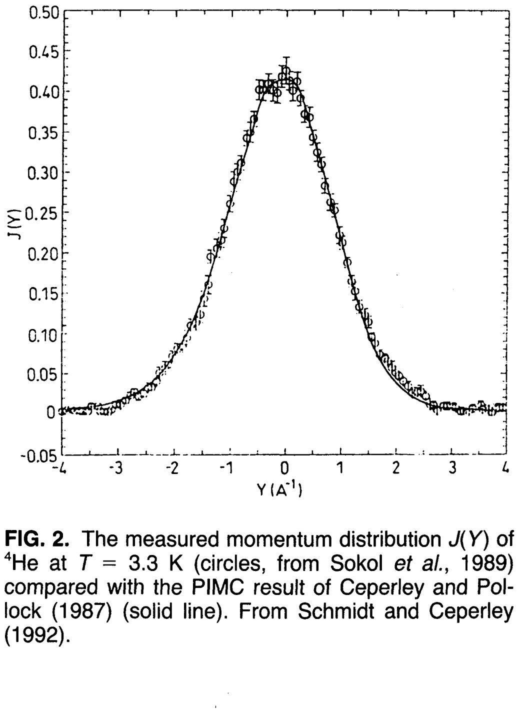

FIG. 2. The measured momentum distribution $J(Y)$ of ^{4}He at T = 3.3 K (circles, from Sokol *et al.*, 1989) compared with the PIMC result of Ceperley and Pollock (1987) (solid line). From Schmidt and Ceperley (1992).

a particle starting at time $t = 0$ at the origin has reached the position $\mathbf{r}$ at time t, the wave function ψ in the Schrödinger equation is not positive definite. This fact creates severe problems for wave functions of many-fermion systems, since these wave functions must be antisymmetric, and the "nodal surface" in configuration space (where ψ changes sign) is unknown.

Formally the difficulty of applying importance-sampling techniques to distributions $\rho(\mathbf{r})$ that are not always positive can be overcome by splitting $\rho(\mathbf{r})$ into its sign, $s = \mathrm{sign}(\rho)$, and its absolute value, $\rho = s|\rho|$, and one can use $\tilde{\rho}(\mathbf{r}) = |\rho(\mathbf{r})|/\int|\rho(\mathbf{r})|d^3r$ as a probability density for importance sampling, and absorb the sign of $\rho(\mathbf{r})$ in the quantity to be measured (e.g., De Raedt and von der Linden, 1992). Symbolically, the average of an observable A is obtained as

$$\langle A\rangle = \frac{\int d^3rA(\mathbf{r})s(\mathbf{r})\tilde{\rho}(\mathbf{r})}{\int s(\mathbf{r})\tilde{\rho}(\mathbf{r})d^3r} = \frac{\langle As\rangle_{\tilde{\rho}}}{\langle s\rangle_{\tilde{\rho}}},$$

where $\langle\ldots\rangle_{\tilde{\rho}}$ means averaging with $\tilde{\rho}$ as weight function. However, as is not unexpected, using $\tilde{\rho}$ one predominantly samples unimportant regions in phase space; therefore, in sampling the sign $\langle s\rangle_{\tilde{\rho}}$, one has large cancellations from regions where the sign is negative, and, for N degrees of freedom, one gets $\langle s\rangle_{\tilde{\rho}} \propto \exp(-\mathrm{const.} \times N)$. This difficulty is known as the "minus-sign problem" and still hampers applications to fermion problems significantly!

Sometimes it is possible to start with a trial wave function where nodes are a reasonable first approximation to the actual nodes, and, starting with the population of random walks from this fixed-node approximation given by the trial function, one now admits walks that cross this nodal surface and sample the sign as indicated above. In this way, it has been possible to estimate the exchange-correlation energy of the homogeneous electron gas (Ceperley and Alder, 1980) over a wide range of densities very well.

7. LATTICE GAUGE THEORY

Monte-Carlo simulation has become the primary tool for nonperturbative quantum chromodynamics, the field theory of quarks and hadrons and other elementary particles (e.g., Rebbi, 1984; De Grand, 1992). In this section, we first stress the basic problem, to make the analogy with the calculations of statistical mechanics clear. Then we very briefly highlight some of the results that have been obtained so far.

7.1 Some Basic Ideas of Lattice Gauge Theory

The theory of elementary particles is a field theory of gauge fields and matter fields. Choice of a lattice is useful to provide a cutoff that removes the ultraviolet divergences that would otherwise occur in these quantum field theories. The first step, hence, is the appropriate translation from the four-dimensional continuum (3 space + 1 time dimensions) to the lattice.

The generating functional (analogous to the partition function in statistical mechanics) is

$$Z = \int DAD\bar{\psi}D\psi \exp[-S_g(A,\bar{\psi},\psi)], \tag{38}$$

where A represents the gauge fields, $\bar{\psi}$ and ψ represent the (fermionic) matter field, S_g is the action of the theory (containing a coupling constant g, which corresponds to inverse temperature in statistical mechanics as $\mathrm{const.}/g^2 \rightarrow 1/k_\mathrm{B}T$), and the symbols $\int D$ stand for functional integration. The action of the gauge field itself is, using the summation convention that indices that appear twice are summed over,

$$S_G = \tfrac{1}{4}\int d^4xF^{\alpha}_{\mu\nu}(x)F^{\mu\nu}_{\alpha}(x), \tag{39}$$

$F^{\alpha}_{\mu\nu}$ being the fields that derive from the vector potential $A^{\alpha}_{\mu}(x)$. These are

$$F^{\alpha}_{\mu\nu}(x) = \partial_\mu A^{\alpha}_{\nu}(x) - \partial_\nu A^{\alpha}_{\mu}(x) + gf^{\alpha}_{\beta\gamma}A^{\beta}_{\mu}(x)A^{\gamma}_{\nu}(x), \tag{40}$$

$f^{\alpha}_{\beta\gamma}$ being the structure constants of the gauge group, and g a coupling constant.

The fundamental variables that one then introduces are elements $U_\mu(x)$ of the gauge group G, which are associated with the links of the four-dimensional lattice, connecting x and a nearest neighbor point $x + \mu$:

$$U_\mu(x) = \exp[igaT^{\alpha}A^{\alpha}_{\mu}(x)],$$
$$[U_m(x + m)]^{\dagger} = U_m(x), \tag{41}$$

where a is the lattice spacing and T^α a group generator. Here $U^\dagger$ denotes the Hermitean conjugate of U. Wilson (1974) invented a lattice action that reduces in the continuum limit to Eq. (39), namely

$$\frac{S_U}{k_B T} = \frac{1}{g^2} \sum_n \sum_{\mu > \nu} \operatorname{Re} \operatorname{Tr} U_\mu(n) \times U_\nu(n+\mu) U_\mu^\dagger(n+\nu) U_\nu^\dagger(n), \tag{42}$$

where the links in Eq. (42) form a closed contour along an elementary plaquette of the lattice.

Using Eq. (42) in Eq. (38), which amounts to the study of a "pure" gauge theory (no matter fields), we recognize that the problem is equivalent to a statistical mechanics problem (such as spins on a lattice), the difference being that now the dynamical variables are the gauge group elements $U_\mu(n)$. Thus importance-sampling Monte-Carlo algorithms can be put to work, just as in statistical mechanics.

In order to include also matter fields, one starts from a partition function of the form

$$\begin{aligned} Z &= \int DUD\bar{\psi}D\psi \exp\left\{ -\frac{S_U}{k_B T} + \sum_{i=1}^{n_f} \bar{\psi}\underline{M}\psi \right\} \\ &= \int DU(\det \underline{M})^{n_f} \exp\left(-\frac{S_U}{k_B T} \right), \end{aligned} \tag{43}$$

where we have assumed fermions with n_f degenerate "flavors." It has also been indicated that the fermion fields can be integrated out analytically, but the price is that one has to deal with the "fermion determinant" of the matrix $\underline{M}$. In principle, for any change of the U's this determinant needs to be recomputed; together with the fact that one needs to work on rather large lattices in four dimensions, in order to reproduce the continuum limit, this problem is responsible for the huge requirement of computing resources in this field.

It is clear that lattice gauge theory cannot be explained in depth on two pages—we only intend to give a vague idea of what these calculations are about to a reader who is not familiar with this subject.

7.2 A Recent Application

Among the many Monte-Carlo studies of various problems (which include problems in cosmology, like the phase transition from the quark-gluon plasma to hadronic matter in the early universe), we focus here on the problem of predicting the masses of elementary particles. Butler *et al.* (1993) have used a new massively parallel supercomputer with 480 processors ("GF11") exclusively for one year to run lattice sizes ranging from $8^3 \times 32$ to $24^3 \times 32$, $24^3 \times 36$, and $30 \times 32^2 \times 40$. Their program executes at a speed of more than 5 Gflops (Giga floating point operations per second), and the rather good statistics reached allowed a meaningful elimination of finite-size effects by an extrapolation to the infinite-volume limit. This problem is important, since the wave function of a hadron is spread out over many lattice sites.

Even with this impressive effort, several approximations are necessary:

1. The fermion determinant mentioned above is neglected (this is called "quenched approximation").
2. One cannot work at the (physically relevant) very small quark mass m_q, but rather has to take data on the various hadron masses for a range of quark masses and extrapolate these data to zero (Fig. 3).

After a double extrapolation ($m_q \to 0$, lattice spacing at fixed volume →0), one obtains mass ratios that are in very satisfactory agreement with experiment. For example, for the nucleon the mass ratio for the finite volume is $m_N/m_\rho = 1.285 \pm 0.070$,, extrapolated to infinite volume 1.219 ± 0.105, the experimen-

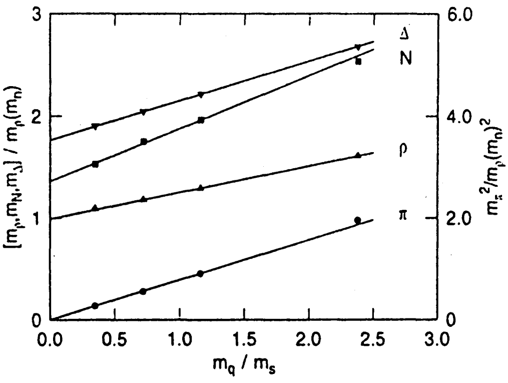

FIG. 3. For a $30 \times 32^2 \times 40$ lattice at $(k_B T)^{-1} = 6.17$, m_π^2, m_ρ, m_N, and m_Δ in units of the physical rho meson mass $m_\rho(m_n)$, as functions of the quark mass m_q in units of the strange quark mass m_s. Particles studied are pion, rho meson, nucleon, and delta baryon, respectively. From Butler *et al.* (1993).

tal value being 1.222 (all masses in units of the mass m_ρ of a rho meson).

8. SELECTED COMMENTS ON APPLICATIONS IN CLASSICAL STATISTICAL MECHANICS OF CONDENSED-MATTER SYSTEMS

In this section, we mention a few applications very briefly, just in order to give the flavor of the type of work that is done and the kind of questions that are asked and answered by Monte-Carlo simulations. More extensive reviews can be found in the literature (Binder, 1976, 1979, 1984, 1992b).

8.1 Metallurgy and Materials Science

A widespread application of Monte-Carlo simulation in this area is the study of order-disorder phenomena in alloys: One tests analytical approximations to calculate phase diagrams, such as the cluster variation (CV) method, and one tests to what extent a simple model can describe the properties of complicated materials.

An example (Fig. 4) shows the order parameter for long-range order (LRO) and short-range order (SRO, for nearest neighbors) as function of temperature for a model of Cu_3Au alloys on the fcc lattice. Here an Ising model with antiferromagnetic interactions between nearest neighbors only is studied, and the Monte-Carlo data (filled symbols and symbols with crosses) are compared to CV calculations (broken curve) and other analytical calculations (dash-dotted curve) and to experiment (open circles). The simulation shows that the analytical approximations describe the ordering of the model only qualitatively. Of course, there is no perfect agreement with the experimental data either; this is to be expected, of course, since in real alloys the interaction range is considerably larger than just extending to nearest neighbors only.

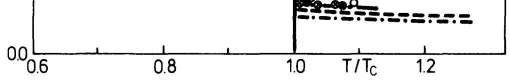

FIG. 4. Long-range order parameter (LRO) and absolute value of nearest-neighbor short-range order parameter (SRO) plotted versus temperature T (in units of the temperature T_c where the first-order transition occurs) for a nearest-neighbor model of binary alloys on the face-centered cubic lattice with A_3B structure. Open circles: experimental data for Cu_3Au; broken and dash-dotted curves: results of analytical theories. Full dots: Monte-Carlo results obtained from a simulation in the semi-grand canonical ensemble (chemical potential difference between A and B atoms treated as independent variable); circles with crosses: values obtained from a canonical ensemble simulation (concentration of B atoms fixed at 25%). From Binder *et al.* (1981).

8.2 Polymer Science

One can study phase transitions not only for models of magnets or alloys, of course, but also for complex systems such as mixtures of flexible polymers. A question heavily debated in the recent literature is the dependence of the critical temperature of unmixing of a symmetric polymer mixture (both constituents have the same degree of polymerization $N_A = N_B = N$) on chain length N. The classical Flory–Huggins theory predicted $T_c \propto N$, while a recent integral equation theory predicted $T_c \propto \sqrt{N}$ (Schweizer and Curro, 1990). This law would lead in the plot of Fig. 5 to a straight line through the origin. Obviously, the data seem to rule out this behavior, and are rather qualitatively consistent with Flory–Huggins theory (though the latter significantly overestimates the prefactor in the relation $T_c \propto N$).

Polymer physics provides examples for many scientific questions where simulations could contribute significantly to provide a better understanding. Figure 6 provides one more example (Paul *et al.*, 1991a). The problem is to provide truly microscopic evidence for the reptation concept (Doi and Edwards, 1986). This concept implies that, as a result of "entanglements" between chains in a dense melt, each chain moves snakelike along its own contour. This behavior leads to a special behavior of mean square displacements: After a characteristic time τ_e, one should see a crossover from a law $g_1(t) \equiv \langle[\mathbf{r}_i(t) - \mathbf{r}_i(0)]^2\rangle \propto t^{1/2}$ (Rouse model) to a law $g_1(t) \propto t^{1/4}$, and, at a still later time (τ_R), one should see an-

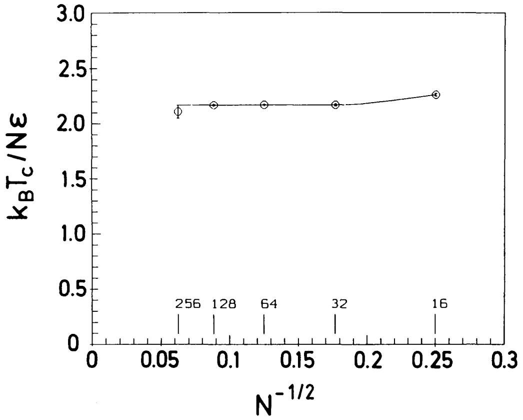

FIG. 5. Normalized critical temperature $k_BT/N\epsilon$ of a symmetric polymer mixture (N = chain length, ϵ = energy parameter describing the repulsive interaction between *A*-*B* pairs of monomers) plotted versus $N^{-1/2}$. Data are results of simulations for the bond-fluctuation model, using N in the range from $N = 16$ to $N = 256$, as indicated. The data are consistent with an asymptotic extrapolation $k_BT_c/\epsilon \approx 2.15N$, while Flory–Huggins theory (in the present units) would yield $k_BT_c/\epsilon \approx 7N$, and the integral-equation theory $k_BT_c/\epsilon \propto \sqrt{N}$. From Deutsch and Binder (1992).

other crossover to $g_1(t) \propto t^{1/2}$ again. At the same time, the center of gravity displacement should also show an intermediate regime of anomalously slow diffusion, $g_3(t) \propto t^{1/2}$. Figure 6 provides qualitative evidence for these predictions—although the effective exponents indicated do not quite have the expected values. A challenge for further theoretical explanation is the anomalous center-of-gravity diffusion in the initial Rouse regime, $g_3(t) \propto t^{0.8}$.

While this is an example where dynamic Monte-Carlo simulations are used to check theories—and pose further theoretical questions—one can also compare to experiment if one uses data in suitably normalized form. In Figure 7, the diffusion constant D of the chains is normalized by its value in the Rouse regime (limit for small N) and plotted versus N/N_e where the characteristic "entanglement chain length" N_e is extracted from τ_e shown in Figure 6 (see Paul *et al.*, 1991a, b, for details). The Monte-Carlo data presented in this scaled form agree with results from both a molecular dynamics (MD) simulation (Kremer and Grest, 1990) and experiment on polyethylene (PE) (Pearson *et al.*, 1987).

This example also shows that, for slow dif-

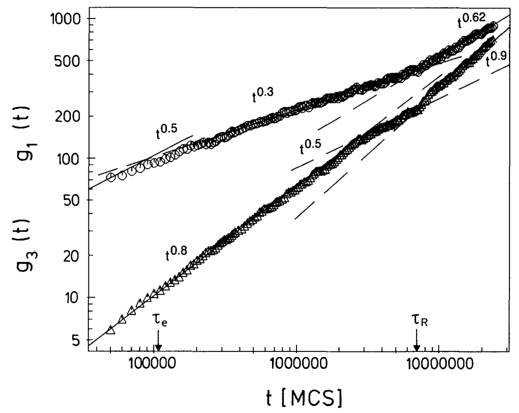

FIG. 6. Log-log plot of the mean square displacements of inner monomers [$g_1(t)$] and of the center of gravity of the chain [$g_3(t)$] versus time t (measured in units of Monte-Carlo steps, while lengths are measured in units of the lattice spacing). Straight lines show various power laws as indicated; various characteristic times are indicated by arrows (see text). Data refer to the bond-fluctuation model on the simple cubic lattice, for an athermal model of a polymer melt with chain length $N = 200$ and a volume fraction $\phi = 0.5$ of occupied lattice sites. From Paul *et al.* (1991a).

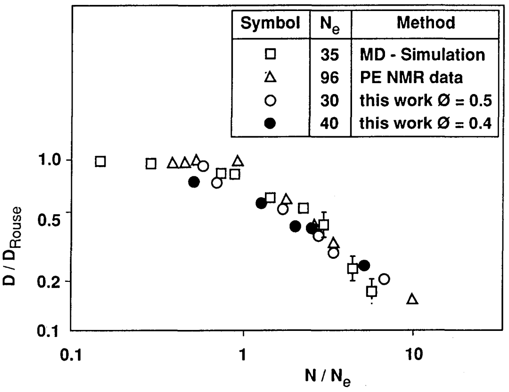

FIG. 7. Log-log plot of the self-diffusion constant D of polymer chains, normalized by the Rouse diffusivity, versus N/N_e (N_e = entanglement chain length, estimated independently and indicated in the inset). Circles: from Monte-Carlo simulations of Paul *et al.* (1991b); squares: from molecular dynamics simulations (Kremer and Grest, 1990); triangles: experimental data (Pearson *et al.*, 1987). From Paul *et al.* (1991b).

fusive motions, Monte-Carlo simulation is competitive to molecular dynamics, although it does not describe the fast atomic motions realistically.

8.3 Surface Physics

Our last example considers phenomena far from thermal equilibrium. Studying the ordering behavior of ordered superstructures, we treat the problem where initially the adatoms are adsorbed at random, and one gradually follows the formation of ordered domains out of initially disordered configurations. In a scattering experiment, one expects to see this by the gradual growth of a peak at the Bragg position $\mathbf{q}_B$. Figure 8 shows a simulation for the case of the (2 × 1) structure on the square lattice, where the Bragg position is at the Brillouin-zone boundary ($q_B = \pi$ if lengths are measured in units of the lattice spacing). Here a lattice gas with repulsive interactions between nearest and next-nearest neighbors (of equal strength) was used (Sadiq and Binder, 1984), using a single–spin-flip kinetics (if the lattice gas is translated to an Ising spin model), as is appropriate for a description of a monolayer in equilibrium with surrounding gas (the random "spin flips" then correspond to random evaporation-condensation events). Figure 9 presents evidence that these data on the kinetics of ordering satisfy a scaling hypothesis, namely

$$S(q,t) = [L(t)]^2\tilde{S}(|\mathbf{q} - \mathbf{q}_B|L(t)), \qquad (44)$$

where $\tilde{S}$ is a scaling function. This hypothesis, Eq. (44), was first proposed on the basis of simulations, and later it was established to describe experimental data as well.

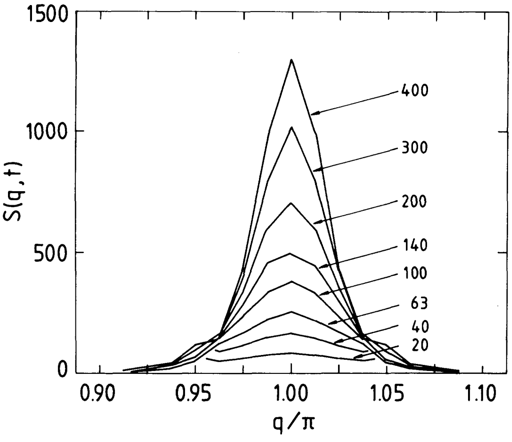

FIG. 8. Structure factor $S(q,t)$ versus wavevector q at various times t (in units MCS per lattice site), after the system was started in a completely random configuration. Temperature (measured in units of the nearest-neighbor repulsion W_{nn}) is 1.33 ($T_c \approx 2.07$ in these units), and coverage $\theta = \frac{1}{2}$. From Sadiq and Binder (1984).

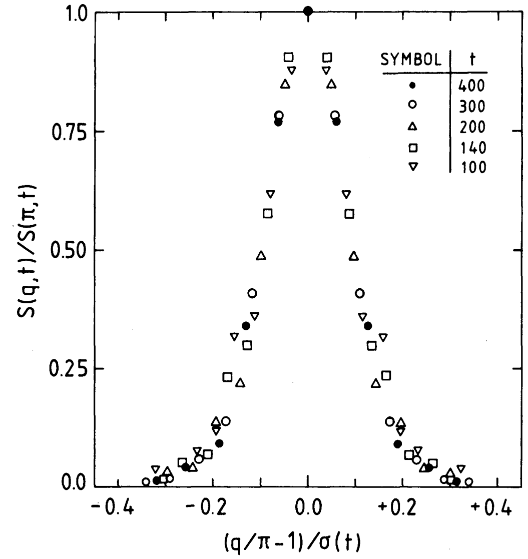

FIG. 9. Structure factor of Fig. 8 plotted in scaled form, normalizing $S(q,t)$ by its peak value $S(\pi,t)$ and normalizing $q/\pi - 1$ by the halfwidth $\sigma(t) = L^{-1}(t)$, where $L(t)$ thus defined is the characteristic domain size. From Sadiq and Binder (1984).

Of course, surface physics provides many more examples where simulations have been useful; see Binder and Landau (1989) for a review.

9. CONCLUDING REMARKS

In this article, the basic features of the most widely used numerical techniques that fall into the category of Monte-Carlo calculations were described. There is a vast literature on the subject—the author estimates the number of papers using Monte-Carlo methods in condensed-matter physics of the order of 10^4, in lattice gauge theory of the order of 10^3! Thus many important variants of algorithms could not be treated here, and interesting applications (e.g., the study of neural-network models) were completely omitted.

There also exist other techniques for the numerical simulation of complex systems, which sometimes are an alternative approach to Monte-Carlo simulation. The molecular dynamics (MD) method (numerical integration of Newton's equations) has already been mentioned in the text, and there exist combinations of both methods ("hybrid Monte Carlo," "Brownian dynamics," etc.). A combination of MD with the local-density approximation of quantum mechanics is the basis of the Car–Parrinello method.

Problems like that shown in Figs. 8 and 9 can also be formulated in terms of numerically solving appropriate differential equations, which may in turn even be discretized to cellular automata. When planning a Monte Carlo simulation, hence, some thought to the question "When which method?" should be given.

GLOSSARY

Critical Slowing Down: Divergence of the relaxation time of dynamic models of statistical mechanics at a second-order phase transition (critical point).

Detailed Balance Principle: Relation linking the transition probability for a move and the transition probability for the inverse move to the ratio of the probability for the occurrence of these states in thermal equilibrium. This condition is sufficient for a Markov process to tend toward thermal equilibrium.

Ergodicity: Property that ensures equality of statistical ensemble averages (such as the "canonic ensemble" of statistical mechanics) and time averages along the trajectory of the system through phase space.

Finite-Size Scaling: Theory that describes the finite-size–induced rounding of singularities that would occur at phase transitions in the thermodynamic limit.

Heat-Bath Method: Choice of transition probability where the probability to "draw" a trial value for a degree of freedom does not depend on its previous value.

Importance Sampling: Monte-Carlo method that chooses the states that are generated according to the desired probability distribution. For example, for statistical mechanics applications, states are chosen with weights proportional to the Boltzmann factor.

Lattice Gauge Theory: Field theory of quarks and gluons in which space and time are discretized into a four-dimensional lattice, gauge field variables being associated to the links of the lattice.

Master Equation: Rate equation describing the "time" evolution of the probability that a state occurs as a function of a "time" coordinate labeling the sequence of states (in the context of importance-sampling Monte-Carlo methods).

Molecular-Dynamics Method: Simulation method for interacting many-body systems based on numerical integration of the Newtonian equations of motion.

Monte-Carlo Step: Unit of (pseudo) time in (dynamically interpreted) importance sampling where, on the average, each degree of freedom in the system gets one chance to be changed (or "updated").

Random Number Generator (RNG): Computer subroutine to produce pseudorandom numbers that are approximately not correlated with each other and approximately uniformly distributed in the interval from zero to one. RNGs typically are strictly periodic, but the period is large enough that, for practical applications, this periodicity does not matter.

Simple Sampling: Monte-Carlo method that chooses states uniformly and at random from the available phase space.

Transition Probability: Probability that controls the move from one state to the next one in a Monte-Carlo process.

Works Cited

Ahrens, J. H., Dieter, U. (1979), *Pseudo Random Numbers*, New York: Wiley.

Barber, M. N. (1983), in: C. Domb, J. L. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 8 New York: Academic, Chap. 2.

Barkema, G. T., Marko, J. F. (1993), *Phys. Rev. Lett.* **71**, 2070–2073.

Berne, B. J., Thirumalai, D. (1986), *Annu. Rev. Phys. Chem.* **37**, 401.

Binder, K. (1976), in: C. Domb, M.S. Green (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 5b, New York: Academic, p. 1.

Binder, K. (1979) (Ed.), *Monte Carlo Methods in Statistical Physics*, Berlin: Springer.

Binder, K. (1984) (Ed.), *Applications of the Monte Carlo Method in Statistical Physics*, Berlin: Springer.

Binder, K. (1987), *Ferroelectrics* **73**, 43–67.

Binder, K. (1992a), *Annu. Rev. Phys. Chem.* **43**, 33–59.

Binder, K. (1992b) (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer.

Binder, K., Heermann, D. W. (1988), *Monte Carlo Simulation in Statistical Physics: An Introduction*, Berlin: Springer.

Binder, K., Landau, D. P. (1989), in: K. P. Lawley (Ed.), *Molecule-Surface Interaction*, New York: Wiley, pp. 91–152.

Binder, K., Lebowitz, J. L., Phani, M. K., Kalos, M. H. (1981), *Acta Metall.* **29**, 1655–1665.

Bruns, W., Motoc, I., O'Driscoll, K. F. (1981), *Monte Carlo Applications in Polymer Science*, Berlin: Springer.

Butler, F., Chen, H., Sexton, J., Vaccarino, A., Weingarten, D. (1993), *Phys. Rev. Lett.* **70**, 2849–2852.

Carlson, J. (1988), *Phys. Rev. C* **38**, 1879–1885.

Ceperley, D. M., Alder, B. J. (1980), *Phys. Rev. Lett.* **45**, 566–569.

Ceperley, D. M., Kalos, M. H. (1979), in: K. Binder (Ed.), *Monte Carlo Methods in Statistical Physics*, Berlin: Springer, pp. 145–194.

Ceperley, D. M., Pollock, E. L. (1987), *Can. J. Phys.* **65**, 1416–1420.

Ciccotti, G., Hoover, W. G. (1986) (Eds.), *Molecular Dynamics Simulation of Statistical Mechanical Systems*, Amsterdam: North-Holland.

Compagner, A. (1991), *Am. J. Phys.* **59**, 700–705.

Compagner, A., Hoogland, A. (1987), *J. Comput. Phys.* **71**, 391–428.

De Grand, T. (1992), in: H. Gausterer, C. B. Lang (Eds.), *Computational Methods in Field Theory*, Berlin: Springer, pp. 159–203.

De Raedt, H., von der Linden, W. (1992), in: K. Binder (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer, pp. 249–284.

Deutsch, H. P., Binder, K. (1992), *Macromolecules* **25**, 6214–6230.

Doi, M., Edwards, S. F. (1986), *Theory of Polymer Dynamics*, Oxford: Clarendon Press.

Ferrenberg, A. M., Landau, D. P., Wong, Y. J. (1992), *Phys. Rev. Lett.* **69**, 3382–3384.

Frick, M., Pattnaik, P. C., Morgenstern, I., Newns, D. M., von der Linden, W. (1990), *Phys. Rev. B* **42**, 2665–2668.

Gillan, M. J., Christodoulos, F. (1993), *Int. J. Mod. Phys.* **C4**, 287–297.

Gould, H., Tobochnik, J. (1988), *An Introduction to Computer Simulation Methods/Applications to Physical Systems, Parts 1 and 2*, Reading, MA: Addison-Wesley.

Hammersley, J. M., Handscomb, D. C. (1964), *Monte Carlo Methods*, London: Chapman and Hall.

Heermann, D. W. (1986), *Computer Simulation Methods in Theoretical Physics*, Berlin: Springer.

Herrmann, H. J. (1986), *Phys. Rep.* **136**, 153–227.

Herrmann, H. J. (1992), in: K. Binder (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer, Chap. 5.

Hockney, R. W., Eastwood, J. W. (1988), *Computer Simulation using Particles*, Bristol: Adam Hilger.

James, F. (1990), *Comput. Phys. Commun.* **60**, 329–344.

Kalos, M. H. (1984) (Ed.), *Monte Carlo Methods in Quantum Problems*, Dordrecht: Reidel.

Kalos, M. H., Whitlock, P. A. (1986), *Monte Carlo Methods*, Vol. 1, New York: Wiley.

Kirkpatrick, S., Stoll, E. (1981), *J. Comput. Phys.* **40**, 517–526.

Knuth, D. (1969), *The Art of Computer Programming*, Vol. 2, Reading, MA: Addison-Wesley.

Koonin, S. E. (1981), *Computational Physics*, Reading, MA: Benjamin.

Kremer, K., Binder, K. (1988), *Comput. Phys. Rep.* **7**, 259–310.

Kremer, K., Grest, G. S. (1990), *J. Chem. Phys.* **92**, 5057–5086.

Landau, D. P. (1992), in: K. Binder (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer, Chap. 2.

Lehmer, D. H. (1951), in: *Proceedings of the 2nd Symposium on Large-Scale Digital Computing Machinery*, Harvard University, Cambridge, 1951, pp. 142–145.

Levesque, D., Weis, J. J. (1992), in: K. Binder (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer, Chap. 6.

Marsaglia, G. A. (1985), in: L. Billard (Ed.), *Computer Science and Statistics: The Interface*, Amsterdam: Elsevier, Chap. 1.

Marsaglia, G. A. (1986), *Proc. Natl. Acad. Sci. U.S.A.* **61**, 25–28.

Marsaglia, G. A., Narasumhan, B., Zaman, A. (1990), *Comput. Phys. Comm.* **60**, 345–349.

Marx, D., Opitz, O., Nielaba, P., Binder, K. (1993), *Phys. Rev. Lett.* **70**, 2908–2911.

Meakin, P. (1988), in: C. Domb, J. L. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 12, New York: Academic, pp. 336–489.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. M., Teller, E. (1953), *J. Chem. Phys.* **21**, 1087–1092.

Panagiotopoulos, A. Z. (1992), *Molec. Simul.* **9**, 1–23.

Paul, W., Binder, K., Heermann, D. W., Kremer, K. (1991a), *J. Chem. Phys.* **95**, 7726–7740.

Paul, W., Binder, K., Heermann, D. W., Kremer, K. (1991b), *J. Phys. (Paris)* **111**, 37–60.

Pearson, D. S., Verstrate, G., von Meerwall, E., Schilling, F. C. (1987), *Macromolecules* **20**, 113–1139.

Privman, V. (1990) (Ed.), *Finite Size Scaling and Numerical Simulation of Statistical Systems*, Singapore: World Scientific.

Rebbi, C. (1984), in: K. Binder (Ed.), *Application* *of the Monte Carlo Method in Statistical Physics*, Berlin: Springer, p. 277.

Reger, J. D., Young, A. P. (1988), *Phys. Rev. B* **37**, 5978–5981.

Rosenbluth, M. N., Rosenbluth, A. W. (1955), *J. Chem. Phys.* **23**, 356–362.

Sadiq, A., Binder, K. (1984), *J. Stat. Phys.* **35**, 517–585.

Schmidt, K. E., Ceperley, D. M. (1992), in: K. Binder (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer, pp. 203–248.

Schmidt, K. E., Kalos, M. H. (1984), in: K. Binder (Ed.), *Applications of the Monte Carlo Method in Statistical Physics*, Berlin: Springer, pp. 125–149.

Schweizer, K. S., Curro, J. G. (1990), *Chem. Phys.* **149**, 105–127.

Sokol, P. E., Sosnick, T. R., Snow, W. M. (1989), in: R. E. Silver, P. E. Sokol (Eds.), *Momentum Distributions*, New York: Plenum Press.

Stanley, H. E. (1971), *An Introduction to Phase Transitions and Critical Phenomena*, Oxford: Oxford University Press.

Stauffer, D. (1985), *An Introduction to Percolation Theory*, London: Taylor & Francis.

Suzuki, M. (1986) (Ed.), *Quantum Monte Carlo Methods*, Berlin: Springer.

Swendsen, R. H., Wang, J. S., Ferrenberg, A. M. (1992), in: K. Binder (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer, Chap. 4.

Tausworthe, R. C. (1965), *Math. Comput.* **19**, 201–208.

Tolman, S., Meakin, P. (1989), *Phys. Rev. A* **40**, 428–437.

Vicsek, T. (1989), *Fractal Growth Phenomena*, Singapore: World Scientific.

West, G. B. (1975), *Phys. Rep.* **18C**, 263–323.

Wilson, K. (1974), *Phys. Rev. D* **10**, 2445–2453.

Further Reading

A textbook describing for the beginner how to learn to write Monte-Carlo programs and to analyze the output generated by them has been written by Binder, K., Heermann, D. W. (1988), *Monte Carlo Simulation in Statistical Physics: An Introduction*, Berlin: Springer. This book emphasizes applications of statistical mechanics such as random walks, percolation, and the Ising model.

A useful book that gives much weight to applications outside of statistical mechanics is Kalos, M. H., Whitlock, P. A. (1986), *Monte Carlo Methods, Vol. 1*, New York: Wiley.

A more general but pedagogic introduction to computer simulation is presented in Gould, H., Tobochnik, J. (1988), *An Introduction to Computer Simulation Methods/Applications to Physical Systems, Parts 1 and 2*, Reading, MA: Addison-Wesley.

A rather systematic collection of applications of Monte-Carlo studies in statistical mechanics and condensed matter physics has been compiled in a series of books edited by the author of this article: Binder, K. (1979) (Ed.), *Monte Carlo Methods in Statistical Physics*, Berlin: Springer; Binder, K. (1984) (Ed.), *Applications of the Monte Carlo Method in Statistical Physics*, Berlin: Springer; and Binder, K. (1992) (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer.

Finally we draw attention to two important areas which are only briefly covered in the present article: For quantum problems, see Suzuki, M. H. (1987) (Ed.), *Quantum Monte Carlo Methods*, Berlin: Springer; Doll, J. D., Gubernaitis, J. E. (1990) (Eds.), *Quantum Simulations*, Singapore: World Scientific. For lattice gauge theory, see Bunk, B., Mütter, K. H., Schilling, K. (1986) (Eds.), *Lattice Gauge Theory, A Challenge in Large-Scale Computing*, New York: Plenum; De Grand, T. (1992), in: H. Gausterer, C. B. Lang (Eds.), *Computational Methods in Field Theory*, Berlin: Springer, pp. 159–203; Smit, J., van Baal, P. (1993) (Eds.), *Lattice 92*, Amsterdam: North Holland.

MOSFET (METAL-OXIDE SEMICONDUCTOR FET)

See TRANSISTORS, FIELD EFFECT

MÖSSBAUER EFFECT

MICHAŁ KOPCEWICZ, *Institute of Electronic Materials Technology, Warszawa, Poland*

	Introduction	1
1.	**Theory**	2
1.1.	Emission of a Gamma Ray from a Free Nucleus	2
1.2.	Gamma Emission from the Nucleus Bound in a Solid—the Mössbauer Effect	3
1.3.	Probability of the Mössbauer Effect	4
1.4.	Shape of the Resonance Line	5
1.5.	Parameters of the Mössbauer Line	5
1.6.	Mössbauer Isotopes	6
1.7.	^{57}Fe Mössbauer Spectroscopy	6
1.8.	Other Isotopes	7
2.	**Instrumentation**	8
2.1.	Basic Principles of the Mössbauer Experiment	8
3.	**Hyperfine Interactions**	10
3.1.	Isomer Shift	10
3.2.	Electric Quadrupole Interaction	11
3.3.	Magnetic Dipole Interaction	13
3.4.	Combined Magnetic Dipole and Electric Quadrupole Interactions	14
3.5.	Information Concerning the Properties of solids from the Mössbauer Effect	15
4.	**Applications of the Mössbauer Effect in Various Fields**	16
5.	**Exotic Mössbauer Studies**	19
6.	**Prospects for the Future**	20
	Glossary	20
	Works Cited	21
	Further Reading	22

INTRODUCTION

The effect to be discussed here is the recoil-free nuclear gamma resonance commonly known as the "Mössbauer effect" in recognition of Rudolf L. Mössbauer who discovered the effect in 1958. The path to this discovery led through conventional investigations of the resonant absorption of nuclear gamma radiation. In the pre-Mössbauer period, the emission and absorption of nuclear gamma radiation were treated in a classical way. The energy and momentum conservation in the process of emission and absorption of gamma radiation was considered for a free nucleus. The forces binding the nucleus in a solid were ignored. Thus, a nuclear transition with energy E_0 was associated with the recoil of the nucleus, which decreased the energy of the emitted gamma radiation. The recoil energy E_R being large as compared with the width of the emission line, together with the shift of the emission and absorption lines due to recoil, causes that the emission and absorption lines usually do not overlap and the resonance does not occur. In the pre-Mössbauer experiments, it was possible to observe the gamma resonance effect either by compensating the recoil energy or by thermal broadening of the emission and absorption lines. In the 1950s, Mössbauer studied the conventional gamma resonance effect for ^{191}Ir and observed an unexpected increase of the resonance effect at low temperature, which was contradictory to the classic model. Mössbauer interpreted this effect in terms of the Lamb model of capture of neutrons by atoms in the crystal (Lamb, 1939). The basis of this discovery constitutes grounds for a fundamentally new concept of the recoil-free emission of nuclear gamma radiation from nuclei bound in a solid and its resonant recoil-free absorption introduced by Mössbauer in 1958 (Mössbauer, 1958a,b). The nuclide ^{191}Ir did not make this effect famous. A breakthrough was made when the recoil-free resonant gamma absorption for ^{57}Fe was observed (Schiffer and Marshall, 1959; Hanna

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

et al., 1960a). It was recognized that the unique properties of the Mössbauer effect, such as a very narrow resonance line and excellent energy resolution, allow the study of hyperfine interactions. In 1960, the isomer shift, the quadrupole interaction, and magnetic hyperfine structure in α-Fe_2O_3 (Kistner and Sunyar, 1960) and in ferromagnetic α-iron (Hanna, *et al.*, 1960b) were observed. In 1961, Mössbauer was awarded the Nobel Prize.

Since its discovery, the Mössbauer effect has been applied in many disciplines of natural science: physics, chemistry, biology, and mineralogy. Now it is one of the most sophisticated experimental methods, which allows us to measure with unrivaled accuracy the changes of energy of gamma radiation.

The Mössbauer effect and its applications have been described in detail in many books (see Further Reading). All references related to the Mössbauer studies are compiled in the Mössbauer Effect Data Index, which is published regularly (10 times a year) by the Mössbauer Effect Data Center (University of North Carolina, Asheville, NC).

In the present article, the principles, basic properties, and methodology of the Mössbauer effect are discussed. The introduction to the hyperfine interactions in solids is presented. The applications of the Mössbauer spectroscopy in solid state and nuclear physics, chemistry, biology, geology, materials science, and technology are described briefly. This article serves only as an introduction to Mössbauer spectroscopy. For more rigorous theoretical treatment and a more complete description of experimental results, the reader is referred to the books suggested for Further Reading at the end of this article.

1. THEORY

1.1 Emission of a Gamma Ray from a Free Nucleus

Let us consider the process of emission of a gamma ray from a free nucleus at rest. Let E_0 be the energy of the excited state of the nucleus, which de-excites to the ground state by emitting a gamma ray. The gamma emission event is associated with the recoil of the nucleus (to conserve momentum), and so the energy of the gamma ray is $E_\gamma = E_0 - E_R$. From energy and momentum conservation, the recoil energy of the nucleus can be calculated as

$$E_R = p^2/2M = E_\gamma^2/2Mc^2 \tag{1}$$

where M is the mass of the nucleus and p is the momentum transferred to the nucleus, equivalent to the momentum of the gamma photon. Similarly, when a gamma photon is absorbed, it loses the energy E_R because of the recoil imparted to the absorbing nucleus.

In order to find whether resonant gamma ray absorption will occur, the recoil energy must be compared to the linewidth Γ of the emission (absorption) line, which is related to the mean lifetime τ of the excited state:

$$\Gamma\tau = \hbar. \tag{2}$$

As can be seen from Eq. (1), the recoil energy depends strongly on the energy of the gamma rays. For example, for a nucleus with $M = 100$ and $E_\gamma \cong E_0 \cong 1 \times 10^4$ eV, the recoil energy $E_R \cong 5 \times 10^{-4}$ eV. However, if the energy of the gamma rays is 5×10^4 eV, then the recoil energy E_R increases already to about 1.3×10^{-2} eV. Since the typical natural linewidth Γ is of the order of 10^{-8} eV for $\tau \cong 100$ ns, the emission and absorption lines will be separated by $2E_R \gg \Gamma$ and resonance will not occur. In order to observe the resonant absorption, the emission and absorption lines must at least partially overlap (Fig. 1). Now one can easily see why nuclear gamma resonance is very unlikely ($2E_R \gg \Gamma$) and optical resonance is easy to observe. For optical transitions with photon energies of a few electronvolts, $E_R \sim 10^{-11}$ eV for $M = 100$ and $\Gamma \sim 10^{-7}$–10^{-8} eV, so that $2E_R \ll \Gamma$.

When gamma rays are emitted and absorbed by free nuclei (e.g., in a gas), which move isotropically in thermal motion, the width of the emission and absorption lines will be increased by the Doppler effect. The Doppler broadening at temperature T is

$$D \cong 2(k_B T E_R)^{1/2}, \tag{3}$$

where k_B is the Boltzmann constant. This broadening at room temperature is of the order of 10^{-3} eV, which is usually comparable to or smaller than E_R. Thus, even the Doppler broadening is in most cases insufficient to make the overlapping of the emission and absorption lines significant. A large increase

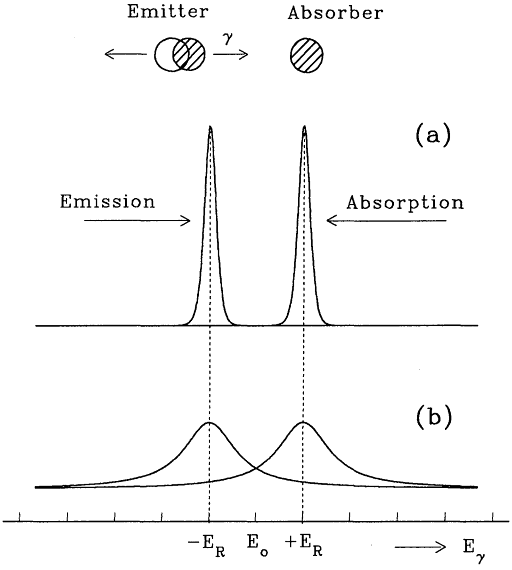

FIG. 1. Conditions for the occurrence of the nuclear gamma resonance for free nuclei: (a) case when $\Gamma < 2E_R$; (b) case when thermally broadened lines partially overlap.

of temperature will increase the overlap, but the lines become very broad, the cross section in the overlap region becomes small, and the resonance effect is of little value because of the very poor resolution.

Various methods have been used to compensate the recoil in order to make observation of the gamma resonance possible. Beside making use of the Doppler broadening, the recoil energy has been compensated by

1. the Doppler shift associated with the motion of the emitting nuclei (source) toward the absorber with a sufficiently high velocity [in the experiment by Moon (1950), the velocity of 8×10^4 cm/s was required] or
2. employing the recoil from the preceding nuclear decay or reaction.

In all these techniques, the recoil energy is *compensated*, in clear distinction to the Mössbauer effect discussed below in which the recoil energy is *eliminated* and compensation is not required.

1.2 Gamma Emission from the Nucleus Bound in a Solid—the Mössbauer Effect

In 1958, Mössbauer discovered that, when the emitting or absorbing nuclei are not free but bound in a solid, a certain fraction of such events may occur with negligible energy loss due to recoil. Mössbauer made this discovery when studying the scattering of 129-keV gamma rays from ^{191}Ir by Ir and Pt at low temperatures (Mössbauer, 1958a,b). In distinction from the existing models for the occurrence of the nuclear gamma resonance discussed above, Mössbauer observed an unexpected increase in the scattering when the temperature was decreased. The new idea of recoil-free gamma emission and absorption can be understood in terms of a simple phenomenological model. It can be shown that the mean energy transferred to the solid because of emission of a gamma quantum from a nucleus bound in the solid is equal to the recoil energy of the free nucleus E_R, Eq. (1). Three cases can be distinguished:

1. If the free-atom recoil energy E_R is larger than the binding energy of the atom in the solid (which typically is of the order of 10 eV), then the emitting atom (nucleus) will be displaced from its lattice site and the situation will be similar to emission from the free atom. Such a situation is typical for a gamma-ray energy of the order of 1 MeV or more. In such a case, the nuclear gamma resonance for lines with natural width will not occur.
2. If the free-atom recoil energy E_R is smaller than the binding energy but larger than the characteristic energy of lattice vibrations (phonon energy) defined in terms of the Debye frequency (ω_D) or Einstein frequency (ω_E), the atom will remain in its lattice position but will dissipate the recoil energy by creation of phonons. This situation is typical for gamma-ray energies of several hundred keV. In this case, the energy of the emitted gamma ray will be smaller than the resonance energy by $n\hbar\omega_D$ or $n\hbar\omega_E$, and resonance will not occur.
3. If the recoil energy E_R is smaller than the phonon energy $\hbar\omega_E$, a new effect arises, because the solid as a quantum system cannot be excited in an arbitrary way. The recoil energy is not dissipated either by displacing the nucleus from its lattice site or by heating the lattice (phonons). The nucleus behaves as if it were rigidly bound to the solid and the recoil is taken by the entire solid. Such a situation may occur for low gamma-ray energies of about 10–

150 keV. In this case, in Eq. (1), M in the denominator represents the mass of the entire solid, making E_R negligible as compared with Γ. When this occurs in the source and the absorber, then the conditions for nuclear gamma resonance are fulfilled and a large resonance for emission and absorption lines *with natural width* is observed. The recoil-free (zero-phonon) emission, absorption, and scattering of nuclear gamma radiation is called the *Mössbauer effect.*

1.3 Probability of the Mössbauer Effect

In the simple case discussed above as case **3** ($E_R \ll \hbar\omega_E$) and when we neglect multiphonon processes, the probability for recoil-free gamma transition (emission or absorption) is

$$f = 1 - E_R/\hbar\omega_\mathrm{E}. \tag{4}$$

Since

$$E_R = \frac{E_\gamma^2}{2Mc^2} = \frac{p^2}{2M} = \frac{\hbar^2\kappa^2}{2M},$$

where κ is the magnitude of the wave vector of the gamma ray and $\hbar = h/2\pi$ (h is Planck's constant), we have

$$f = 1 - \hbar^2\kappa^2/2\hbar\omega_\mathrm{E}M. \tag{5}$$

A general expression for the fraction of zero-phonon (recoil-free) processes is given by the same Debye–Waller (W) factor that is applicable to the scattering of x rays or neutrons by atoms:

$$\begin{aligned} f &= \exp(-2W) = \exp(-4\pi^2\langle x^2\rangle/\lambda^2) \\ &= \exp[-\kappa^2\langle x^2\rangle], \end{aligned} \tag{6}$$

where λ is the wavelength of the gamma quantum:

$$\kappa = 2\pi/\lambda = E/hc,$$

and $\langle x^2\rangle$ is the component of the mean square vibrational amplitude of the emitting nucleus in the direction of the gamma ray. A large probability for the Mössbauer effect (f close to 1) occurs when $\kappa^2 \langle x^2\rangle \ll 1$, i.e., when the rms displacement of the nucleus is small as compared to the wavelength λ. This is why the Mössbauer effect is not observed in gases and nonviscous liquids, where $\langle x^2\rangle$ is not limited.

Since $\langle x^2\rangle$ depends on temperature, the study of the temperature dependence of the Mössbauer effect provides useful information concerning lattice dynamics. The probability of the Mössbauer effect can be calculated in terms of the Debye model of lattice dynamics. Only $\langle x^2\rangle$ must be calculated, because κ is constant for a given gamma transition. The Debye model yields the following expression for the probability of the recoil-free fraction (see, e.g., Greenwood and Gibb, 1971):

$$\begin{aligned} f &= \exp\left[-\frac{6E_R}{k_\mathrm{B}\Theta_\mathrm{D}}\left(\frac{1}{4} + \frac{T^2}{\Theta_\mathrm{D}^2}\int_0^{\Theta_\mathrm{D}/T}\frac{x\,dx}{e^x - 1}\right)\right] \\ &= \exp[-2W], \end{aligned} \tag{7}$$

where Θ_D is the Debye temperature, k_B is the Bolzmann constant, T is the temperature, and $x = \hbar\omega/k_\mathrm{B}T$.

The recoil-free fraction f can be evaluated in the high- and low-temperature limits with respect to Θ_D. At low temperatures ($T \ll \Theta_\mathrm{D}$),

$$f = \exp\left[-\frac{E_R}{k_\mathrm{B}\Theta_\mathrm{D}}\left(\frac{3}{2} + \frac{\pi^2T^2}{\Theta_\mathrm{D}^2}\right)\right]. \tag{8}$$

At absolute zero ($T = 0$ K),

$$f = \exp[-3E_R/2k_\mathrm{B}\Theta_\mathrm{D}]. \tag{9}$$

In the high-temperature range ($T \gg \Theta_\mathrm{D}/2$),

$$f = \exp[-6E_RT/k_\mathrm{B}\Theta_\mathrm{D}^2]. \tag{10}$$

Equation (10) yields the linear dependence of the Debye–Waller factor W on T at high temperatures. This is true only when the harmonic approximation (Debye model) holds. In many experimental cases, when anharmonicity plays a significant role, a deviation from the proportionality of W to T is observed.

Concluding, the probability of the Mössbauer effect depends on three factors:

1. the free-atom recoil energy, related to the gamma transition energy [Eq. (1)];
2. the properties of the solid (i.e., the Debye temperature); and
3. the temperature of the solid (T).

A large probability is observed for low-energy gamma transitions in a solid with high Debye temperature when the measurements are performed at low temperature.

1.4 Shape of the Resonance Line

The gamma radiation emitted without recoil from the source has a Heisenberg width at half height [Eq. (2)], Γ_s, and the distribution of energies around the energy E_γ is given by the Breit–Wigner formula. The resulting shape of the line is Lorentzian:

$$w(E) = \frac{(\Gamma_s/2)^2}{(E - E_\gamma)^2 + (\Gamma_s/2)^2}. \tag{11}$$

The cross section for resonance absorption (line shape of the absorption line with width Γ_a) is also given by the Breit–Wigner formula:

$$\sigma(E) = \sigma_0 \frac{(\Gamma_a/2)^2}{(E - E_\gamma)^2 + (\Gamma_a/2)^2}. \tag{12}$$

The maximum cross section is given by

$$\sigma_0 = 2\pi\lambda\!\!\!^{-2} \frac{2I_e + 1}{2I_g + 1} \frac{1}{1 + \alpha}, \tag{13}$$

where I_e and I_g are the nuclear spins of the excited and ground states and α is the internal conversion coefficient.

In a resonance experiment, an emission line of width Γ_s is moved over an absorption line of width Γ_a. The experimentally observed shape of the resonance line is

$$\sigma_{\exp}(E) = \int_{-\infty}^{\infty} w(e)\sigma(E - e)de. \tag{14}$$

In the thin absorber approximation limit,

$$\sigma_{\exp}(E) = \sigma_0 \frac{(\Gamma_s + \Gamma_a)^2/4}{(E - E_\gamma)^2 + (\Gamma_s + \Gamma_a)^2/4}. \tag{15}$$

If $\Gamma_s \cong \Gamma_a \cong \Gamma$, where Γ is the natural linewidth, then

$$\sigma_{\exp}(E) \cong \sigma_0 \frac{\Gamma^2}{(E - E_\gamma)^2 + \Gamma^2}. \tag{16}$$

Since the theoretical line shape is well known [Eq. (16)], the Mössbauer effect allows us to measure changes of the gamma radiation energy of the order of a fraction of the natural linewidth Γ. The energy resolution, defined as

$$\Delta E/E_\gamma = \Gamma/E_\gamma, \tag{17}$$

is of the order of 10^{-12}–10^{-15} depending on the Mössbauer isotope. The Mössbauer effect provides the most accurate method for measuring the changes of energy of electromagnetic radiation. Because of this feature, the Mössbauer effect is the most powerful technique for studying hyperfine interactions (see Sec. 3).

1.5 Parameters of the Mössbauer Line

The maximum resonance absorption occurs when the energies E_γ in the source (E_γ^s) and absorber (E_γ^a) coincide. The effective E_γ value can be changed by moving the source with respect to the absorber with velocity v, i.e., by using an externally applied Doppler effect. The change of the energy E_γ due to such motion is

$$\Delta E = (v/c)\,E_\gamma. \tag{18}$$

Usually, absorption experiments are performed, and so the maximum resonance absorption corresponds to the minimum counting rate at the velocity v_r for which $E_\gamma^s = E_\gamma^a$ (Fig. 2). At any higher or lower velocity, the resonance absorption decreases until it finally vanishes at velocities far from v_r. The shape of the absorption line is Lorentzian [see Eq. (15)] with the full width at half maximum $\Gamma_r = \Gamma_s + \Gamma_a$, and the counting rate at $\pm v_\infty$ gives the nonresonant background. The nuclei in the absorber excited by resonance absorption re-emit gamma rays within $\tau \cong 10^{-7}$ s. However, when the internal conversion coefficient is high, relatively few gamma rays will be re-emitted. The re-emission process is not directional but occurs over the 4π solid angle. Thus, the number of counts in the detector corresponding to re-emitted gamma rays in a collimated, strongly directional transmission experiment is usually negligible.

Concluding, the parameters of a simple single-line Mössbauer spectrum (Fig. 2) are as follows:

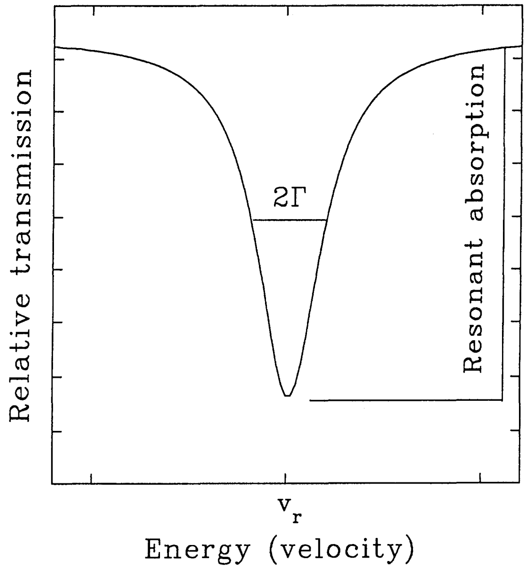

FIG. 2. Parameters of the Mössbauer transmission spectrum.

1. position of the resonance v_r;
2. full width at half maximum of the resonance line $\Gamma_r = \Gamma_s + \Gamma_a \cong 2\Gamma$;
3. depth of the resonance line, expressed as

$$D_r = \frac{\text{count rate at resonance} - \text{count rate at } \pm v_\infty}{\text{count rate at } \pm v_\infty}; \tag{19}$$

4. nonresonant background count rate at $\pm v_\infty$; and
5. area of the resonance line,

$$A_r = \tfrac{1}{2}\pi \Gamma_r D_r. \tag{20}$$

1.6 Mössbauer Isotopes

In order to observe the Mössbauer effect, a proper isotope must be found and favorable conditions for the solid matrix in which the isotope will be bound must be fulfilled. The Mössbauer nucleus should fulfill the following basic requirements:

1. low gamma transition energy (typically 10–150 keV) between the first excited and the stable ground state, and nucleus of mass $M > 50$ to minimize the recoil energy E_R;
2. mean lifetime of the excited level $\tau \cong 10^{-7}$–10^{-9} s; a longer lifetime results in a very narrow linewidth that makes the resonance difficult to observe experimentally, while a shorter lifetime results in a wide resonance line with poor resolution;
3. preferably, multipolarity of the gamma transition M1 or E2;
4. isotope production in the excited state as a product of natural radioactive decay of the parent isotope; the lifetime of the parent isotope should be reasonable, preferably of the order of 1 year or more. Mössbauer isotopes with radioactive metastable ground state also exist. The most important one is ^{129}I in which the ground state is radioactive with a half-life of 1.7×10^7 years. Other examples are ^{133}Ba (half-life of the ground state of 7.2 yr), ^{147}Pm (17.7 yr), ^{231}Pa (3.25×10^4 yr), ^{243}Am (7.95×10^3 yr); and
5. large maximum resonance cross section σ_0.

In view of the applications of the Mössbauer effect, the electric quadrupole moment and the nuclear magnetic moments of the levels involved in gamma transition should not be zero.

The solid matrix in which the isotope is bound should have a high Debye temperature Θ_D to allow observation of the Mössbauer effect with reasonable probability in a wide temperature range.

By far the best combination of the requirements listed above is found in ^{57}Fe, which is the most convenient and most popular Mössbauer isotope. However, the Mössbauer effect has been observed for 109 transitions in 90 isotopes of 47 elements.

1.7 ^{57}Fe Mössbauer Spectroscopy

Most Mössbauer investigations involve the isotope ^{57}Fe. Over 75% of all papers on experimental Mössbauer spectroscopy are concerned with the 14.4-keV transition in ^{57}Fe. It is fortunate that an element of great commercial importance and so abundant in nature, which forms a great variety of chemical compounds and alloys, has an isotope with an almost ideal combination of properties for Mössbauer spectroscopy. The gamma transition for which the Mössbauer effect is observed has very low energy (14.39 keV), and hence, the recoil energy is small (1.95×10^{-3} eV); an almost purely magnetic dipole (M1) transition (E2 admixture is negligible) occurs between the first excited state ($I_e = \frac{3}{2}^{-}$) and the ground state ($I_g = \frac{1}{2}^{-}$). The lifetime ($\tau_{1/2}$) of this level is about 100 ns, and hence, the natural linewidth is only 0.45×10^{-8} eV and produces the linewidth of the Mössbauer line of $2\Gamma = 0.194$ mm/s. The small amount of ^{57}Fe in natural iron (2.19%) is compensated by the large cross section for resonant absorption, $\sigma_0 \cong 257 \times 10^{-20}$ cm^2. The 14.4-keV transition has an internal conversion factor $\alpha \cong 10$, which allows us to use the conversion electrons for detecting the Mössbauer transition. This leads to a special technique: conversion-electron Mössbauer spectroscopy (CEMS). The conversion electrons (among which 7.3-keV ones emitted from the K shell are most abundant) emerge only from the depth equivalent to their range in a given material. Thus, the CEMS technique is very useful for surface studies, probing the depth of about 150 nm. ^{57}Fe is produced by β^+ (electron capture) decay from the parent ^{57}Co isotope whose lifetime is 270 days. ^{57}Co decay populates the second excited state in ^{57}Fe of the energy of 136 keV, which decays 90% through a cascade to the 14.4-keV level and then to the ground state, and 10% directly to the ground state. The Mössbauer effect has also been detected for the 136-keV transition, but its importance for the applications of the Mössbauer spectroscopy is negligible. The decay scheme of ^{57}Fe is shown in Fig. 3.

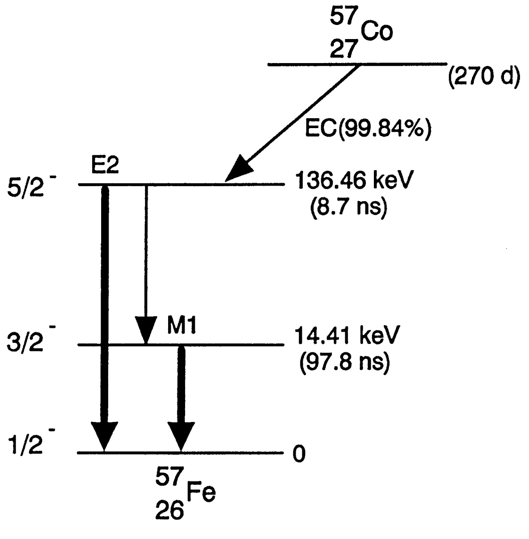

FIG. 3. Decay scheme for ^{57}Fe.

The recoil energy being small, the probability of the recoil-free emission and absorption of 14.4-keV gamma rays is large even at high temperatures (well above room temperature). Thus, the Mössbauer effect can be employed to study hyperfine interactions in a wide temperature range.

By using the Mössbauer effect, it is possible to measure changes of the energy of gamma radiation with astonishing accuracy. The energy resolution of the ^{57}Fe resonance is $\Gamma/E_\gamma \cong 3 \times 10^{-13}$.

1.8 Other Isotopes

The second most popular Mössbauer isotope is ^{119}Sn. The decay scheme and the most relevant parameters are shown in Fig. 4. The most convenient way of populating the 23.87-keV excited level used for the observation of the Mössbauer effect is to use metastable ^{119m}Sn, which can be produced by neutron capture in enriched ^{118}Sn. The metastable ^{119m}Sn has a lifetime of 250 days. The energy resolution is, however, poorer than for ^{57}Fe (the natural linewidth $2\Gamma = 0.626$ mm/s). This isotope has found fairly wide applications in chemistry. A range of tin-containing inorganic and organic compounds and alloys has been studied. It is of particular interest to study materials containing two Mössbauer isotopes, e.g., FeSn alloys, where both isotopes can be used for studying the hyperfine interactions.

The lightest nuclide for which the

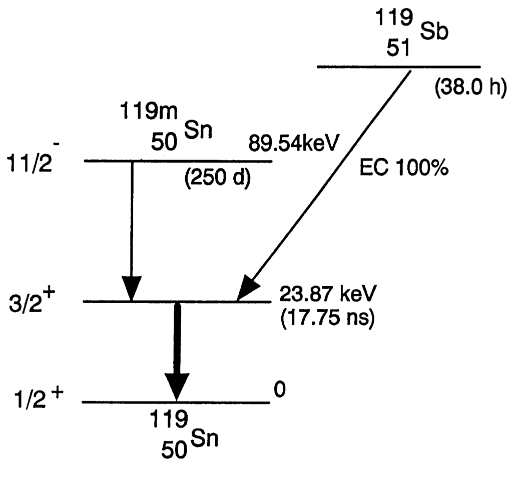

FIG. 4. Decay scheme for ^{119}Sn.

Mössbauer effect was observed is ^{40}K. However, the 29.4-keV Mössbauer level is not populated by any radioactive parent. This level is accessible only during a nuclear reaction [e.g., (d,p) or (n,γ)].

The isotope that reveals the highest energy resolution, of the order of 10^{-15}, is ^{67}Zn. The 93.26-keV excited state has a long lifetime of 9400 ns. The resulting linewidth of the Mössbauer line is 3.12×10^{-4} mm/s. Experiments with ^{67}Zn are very difficult, since the normal level of mechanical vibrations in the laboratory may easily disturb the resonance. Nevertheless, experiments with ^{67}Zn have been performed successfully in several laboratories.

Other isotopes that have found applications mostly in chemistry are ^{129}I, ^{127}I, and ^{125}Te. For the studies of magnetism of rare earth elements, Mössbauer isotopes such as ^{151}Eu, ^{153}Eu, and ^{161}Dy are available. The unique element is gadolinium, which has eight Mössbauer transitions in six isotopes. Unfortunately none of them is easy to observe; the experiments require expensive instrumentation, and even then the hyperfine effects cannot be fully resolved.

The Mössbauer isotopes of actinides (^{237}Np, ^{237}U, ^{241}Au, ^{232}Th) are interesting for measuring nuclear constants. The experiments are difficult because in most of these isotopes the Mössbauer nuclear level can be populated only in nuclear reaction or via Coulomb excitation. There is also high background due to other nuclear gamma transitions, which obscure the Mössbauer line.

An extensive list of isotopes with all the relevant data for which the Mössbauer effect was observed can be found in, e.g., Gonser (1975), Greenwood and Gibb (1971), and Shenoy and Wagner (1978).

2. INSTRUMENTATION

2.1 Basic Principles of the Mössbauer Experiment

The most typical Mössbauer experiment involves a radioactive source containing the Mössbauer isotope in an excited state and an absorber consisting of the material to be investigated that contains the same isotope in the ground state. Here, we will present the most common case of experiments with the use of the isotope ^{57}Fe. In transmission geometry, the gamma rays emitted from the source pass through the absorber in which resonant absorption occurs, and then they are detected by a suitable detector. In order to investigate the energy levels of the Mössbauer nucleus in the absorber, which may be shifted with respect to the energy levels of the same isotope in the source material or split by hyperfine interactions (see Sec. 3), it is necessary to modulate the energy of the emitted gamma radiation so that it can match its energy to the resonance. The most common method of energy modulation is Doppler modulation, consisting in a relative movement of the source with respect to the absorber. The energy change δE is given by the simple formula

$$\delta E = (v/c)\, E_\gamma, \tag{21}$$

where v is the velocity of the relative motion of the source with respect to the absorber, E_γ is the energy of the Mössbauer gamma radiation, and c is the velocity of light. For the 14.4-keV transition in ^{57}Fe, the velocity of 1 mm/s corresponds to 4.8×10^{-8} eV or 11.6 MHz. The motion of the source is usually oscillatory to provide an energy scan.

The block diagram of a simple Mössbauer spectrometer is shown in Fig. 5. The typical velocity transducer is an electromechanical device consisting of two coils fixed to a rod supported by springs and placed in a static magnetic field (double loudspeaker system). The drive signal delivered from the drive generator to one of the coils induces motion of the rod, to which the source is attached. As a result of this motion, a signal is induced in the pick-up coil that is used for correcting the shape of the drive signal via a negative-feedback loop. The shape of the drive signal is usually such that the motion occurs with constant acceleration (linear velocity scale). Sometimes a sinusoidal drive signal is used.

The detector, preamplifier, amplifier, and single-channel analyzer (SCA) form a conventional gamma spectrometer. The detector must have a sufficiently good energy resolution to ensure clear resolution of the Mössbauer gamma radiation from other gamma or x-ray lines and the background. For ^{57}Fe Mössbauer spectroscopy, usually proportional gas counters or scintillation counters do this job. All pulses from the detector, re-

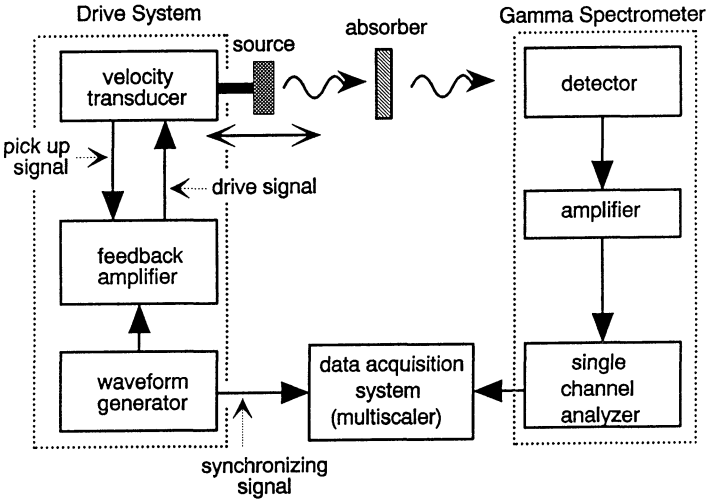

FIG. 5. Block diagram of a typical Mössbauer spectrometer.

gardless of their energy, are amplified. The SCA is used for selecting from the whole energy spectrum only the pulses corresponding to the Mössbauer gamma radiation. These pulses are delivered to the acquisition system working in the multiscaler mode (MSC). The address of the MSC is triggered by a starting pulse from the drive generator, and so the counting is synchronized with the motion of the source in such a way that each channel in the multiscaler stores the counts corresponding to a given velocity. In this way, a Mössbauer spectrum, i.e., counting rate vs velocity, is obtained.

The Mössbauer source must fulfill the following conditions. The radioactive isotope, e.g., ^{57}Co, must be embedded into a solid matrix that has a high Debye temperature to assure high probability for the recoil-free gamma emission. The preferred matrix material must be a nonmagnetic cubic metal (e.g., Rh, Cr, Pt, Pd, Cu) to avoid hyperfine splitting and ensure obtaining a single narrow emission line. The commercially available sources, such as ^{57}Co in Rh, have a probability for emission of the Mössbauer gamma radiation of about 0.75.

The absorber must be a solid (single crystal, polycrystalline powder or foil, amorphous, frozen solution, etc.) containing the resonant isotope. Its thickness should be small enough to allow transmission of gamma radiation but high enough to ensure measurable resonant absorption. The effective absorber thickness is

$$t_a = \sigma_o f_a n_a a_a d_a, \qquad (22)$$

where σ_o is the maximum absorption cross section, f_a is the recoil-free fraction for the absorber material, n_a is the number of atoms/cm^3 of the particular element, a_a is the isotopic abundance of the resonance isotope, and d_a is the physical thickness of the absorber (in centimeters). In usual experimental conditions, $t_a \cong 1$.

It is possible to carry out Mössbauer experiments in which the material studied forms the source. In such a case, a single-line absorber should be used.

The experiments are performed not only in transmission geometry (which is most common) but also in a scattering geometry. The scattering geometry is useful when a thin absorber cannot be prepared, and it allows examination of a thin surface layer of the bulk materials. In the scattering geometry, the Mössbauer effect can be monitored by detecting either the Mössbauer gamma radiation or conversion electrons (conversion electron Mössbauer spectroscopy—CEMS), or else by x rays (conversion x-ray Mössbauer spectroscopy—CXMS) associated with the gamma absorption process.

Many experiments are carried out in a wide temperature range, and so cryostats or ovens are used. Investigations of magnetic properties require in many cases measurements with external static magnetic fields produced by superconducting coils. It is also possible to apply pressure to the material investigated, but such experiments are very difficult to perform.

In the case of ^{57}Fe Mössbauer spectroscopy, it is possible to produce polarized gamma rays either from a magnetically split six-line source of ^{57}Fe in ferromagnetic α-Fe or from a paramagnetic single-crystalline source exhibiting a quadrupole-split spectrum. The Mössbauer polarimetry technique is useful for studying the magnetic properties of solids. For the details concerning this method, the reader is referred to Gonser (1975).

3. HYPERFINE INTERACTIONS

The great accuracy of the Mössbauer effect of measuring the changes of the energy of electromagnetic (nuclear gamma) radiation is the most important feature of this technique. Its importance is enhanced by the fact that the typical energy resolution for the Mössbauer effect, related to the linewidth of the resonance line, is of the order of 10^{-12}–10^{-15}; this means that for 10–100-keV gamma radiation, energy changes of the order of 10^{-8} eV can be measured. Because of this feature, the Mössbauer effect is the most powerful tool for studying the hyperfine interactions. It allows us to observe directly the shifts and splittings of the Mössbauer lines resulting from the interactions of the nucleus with electrons. The hyperfine coupling mechanisms yield information regarding electron and spin density distributions, thus providing information concerning chemical, structural, and magnetic properties of solids.

The hyperfine interaction Hamiltonian for the atom contains terms related to interactions between the nucleus and its environment (electrons):

$$\mathcal{H} = E_0 + M_1 + E_2 + \cdots, \tag{23}$$

where E_0 refers to the Coulomb interaction between the nucleus and electrons at the nuclear site, M_1 is the interaction between the nuclear magnetic dipole moment and the effective magnetic field at the nucleus (magnetic dipole hyperfine interaction), and E_2 is the interaction between the nuclear quadrupole moment and the electric field gradient at the nucleus (electric quadrupole interaction). These are the most important interactions that determine the shape of the Mössbauer spectra. The term E_0 describes the isomer shift, M_1 is the nuclear Zeeman effect, which is responsible for the magnetic hyperfine structure, and E_2 causes the quadrupole splitting. Interactions of higher order (M_3, E_4, etc.) can be neglected because their energies are by several orders of magnitude smaller than E_0, M_1, and E_2, and the electric dipole interaction (E_1) is parity forbidden.

3.1 Isomer Shift

The presence of electrons affects the energy levels of the nucleus. The nuclear levels are shifted by the interaction of the nuclear charge with the electron charge density at the nucleus. In a given nuclear state, the nucleus, considered as a uniformly charged sphere of radius R (R_g or R_e for the ground and excited states engaged in the gamma transition, respectively), interacts with the nonzero electron charge density (s electrons) within the nuclear volume. During a gamma transition, the effective nuclear volume changes, thereby affecting the nucleus-electron interaction energy. The energy of the nuclear gamma transition is shifted as compared to that for a hypothetical point nucleus of the same charge by

$$\Delta E = (2\pi/5)Ze^2|\psi(0)|^2(R_e^2 - R_g^2), \tag{24}$$

where $-e|\psi(0)|^2$ is the electron charge density at the nucleus site.

Since in the Mössbauer effect the source and absorber are involved, for which the chemical environments of the resonant nucleus can be different, the isomer shift δ (or IS) measured in the experiment is

$$\begin{aligned}\delta = \mathrm{IS} &= \Delta E_a - \Delta E_s \\ &= (2\pi/5)Ze^2[|\psi_a(0)|^2 - |\psi_s(0)|^2](R_e^2 - R_g^2) \\ &= (4\pi/5)Ze^2\mathrm{R}^2(\delta R/R)[|\psi_a(0)|^2 - |\psi_s(0)|^2],\end{aligned} \tag{25}$$

where subscripts a and s denote the absorber and the source, respectively, and $\delta R = R_e - R_g$. Figure 6(a) schematically presents the shifts of the nuclear levels causing the isomer shift in the Mössbauer spectrum, a simple example of which is shown in Fig. 6(b).

As can be seen from Eq. (25), the isomer

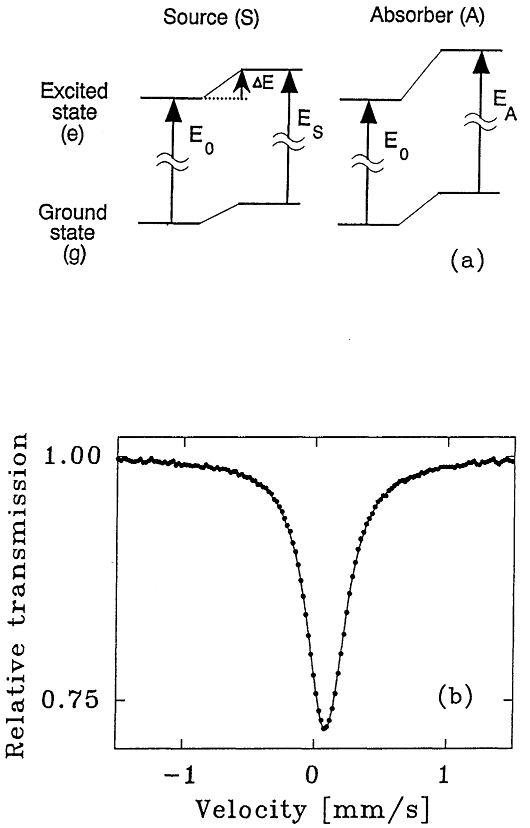

FIG. 6. The isomer shift: (a) the shift of the nuclear levels; (b) the Mössbauer spectrum of $K_4Fe(CN)_6$ enriched in ^{57}Fe revealing the isomer shift.

shift depends both on the nuclear properties (the change of the nuclear radius due to gamma transition) and on electronic properties, the latter related to the chemical properties of the material in which emitting or absorbing nuclei are contained. It is not possible to separate these contributions in a single Mössbauer experiment. Usually, the nuclear part, constant for a given gamma transition, is established first, or measured by another method, and the isomer shift is used to determine the electronic term, which allows study of the chemical properties via the changes of the electron density for various materials.

From the direction of the shift of the resonance line with increasing *s*-electron density, it is inferred that δR is negative for ^{57}Fe, which means that the ^{57}Fe nucleus is larger in the ground state than in the excited state. More typical is the opposite trend ($\delta R > 0$) as observed, e.g., for another typical Mössbauer isotope—^{119}Sn.

Despite the isomer shift being directly related to the *s*-electron density, it depends also on the *d* electrons through screening effects.

The isomer shift is measured relatively: the isomer shift of the absorber is measured with respect to the source host material. Various reference standards have been chosen for various isotopes. For the ^{57}Fe isotope, the reference point with respect to which the isomer shifts are measured is the center of the spectrum of ferromagnetic α-Fe at room temperature. For a detailed discussion of isomer shifts for various isotopes, the reader is referred to Shenoy and Wagner (1978).

The line position in the spectrum is determined not only by the isomer shift [Eq. (25)] but also by a relativistic temperature-dependent contribution (second-order Doppler shift) of the form

$$\delta E_\gamma / E_\gamma = -\langle v^2 \rangle / 2c^2, \qquad (26)$$

where $\langle v^2 \rangle$ is the mean square velocity of the resonant nucleus related to the thermal motion and can be calculated in terms of the Einstein or Debye model.

3.2 Electric Quadrupole Interaction

The electric quadrupole interaction between the quadrupole moment of the nucleus in a state with spin I and the electric field gradient (EFG) at the site of the nucleus created by the asymmetry of the electronic or lattice charge distribution is described by the Hamiltonian

$$\mathcal{H}_Q = \hat{Q}\nabla E = \frac{e^2 Qq}{4I(2I-1)} \times \left[3I_z^2 - I(I+1) + \frac{\eta}{2}(I_+^2 + I_-^2)\right], \qquad (27)$$

where $eq = V_{zz}$ and $\eta = (V_{xx} - V_{yy})/V_{zz}$ are the principal component and the asymmetry parameter of the diagonal EFG tensor, respectively, in such a reference system that $|V_{zz}| \geq |V_{xx}| \geq |V_{yy}|$ and $V_{zz} + V_{yy} + V_{xx} = 0$; thus, $0 \leq \eta \leq 1$; I_+ and I_- are the raising and lowering (shift) operators, respectively.

As a result of this interaction, the energy levels with $|Q| > 0$ ($I > \frac{1}{2}$) split into sublevels with energies given by the eigenvalues of

Eq. (27):

$$E_Q = \frac{e^2qQ}{4I(2I-1)}[3m_I^2 - I(I+1)]\left(1+\frac{\eta^2}{3}\right)^{1/2}, \tag{28}$$

where $m_I = I, I-1, \ldots, -I$ is the magnetic quantum number. The electric quadrupole interaction lifts only partially the degeneracy of the nuclear states, and the states that differ only in sign of m_I remain degenerate.

In the case of ^{57}Fe, for which the spins of the ground and excited states are $\frac{1}{2}$ and $\frac{3}{2}$ respectively, the quadrupole interaction causes the splitting of the excited state into two sublevels with $m_I = \pm\frac{3}{2}$ and $\pm\frac{1}{2}$, leaving the ground state unsplit. The Mössbauer spectrum, which corresponds to the transitions between two sets of hyperfine sublevels, splits into a two-line pattern (quadrupole-split doublet) (Fig. 7). The line intensities are determined by the angular dependence of the electric quadrupole interaction. In the case of an axially symmetric EFG ($\eta = 0$), this dependence is

$\frac{3}{2}(1 + \cos^2\Theta)$ for $\pm\frac{3}{2} \rightarrow \pm\frac{1}{2}$ transition

$1 + \frac{3}{2}\sin^2\Theta$ for $\pm\frac{1}{2} \rightarrow \pm\frac{1}{2}$ transition

where Θ is the angle between the direction of the principal axis of the EFG tensor and the propagation direction of the gamma radiation. For randomly oriented polycrystalline materials, the relative line intensities are 1:1. However, if the recoil-free fraction is anisotropic because of anisotropy of lattice vibrations (i.e., the f factor depends on the angle Θ measured with respect to the crystallographic axis of highest symmetry), then the angular dependence of the electric quadrupole interaction is modified:

$\frac{3}{2}(1 + \cos^2\Theta)f(\theta)$ for $\pm\frac{3}{2} \rightarrow \pm\frac{1}{2}$ transition;

$[1 + \frac{3}{2}\sin^2\Theta]f(\Theta)$ for $\pm\frac{1}{2} \rightarrow \pm\frac{1}{2}$ transition.

In such a case, averaging of Θ over a sphere yields, even for polycrystalline material, unequal line intensities of the quadrupole doublet. This effect is called the Goldanskii–Karyagin effect (Goldanskii, *et al.*, 1963; Karyagin, 1963).

As for the case of isomer shift, the measurements of quadrupole splitting give only the product of the nuclear parameter (Q) and the electric field gradient at the nucleus. To obtain a value of the quadrupole moment, which can serve as a valuable test of nuclear models, it is necessary to evaluate independently the EFG tensor. The nuclear parameter is, however, constant for a given nuclear level, and it is the details of the EFG, which can be derived from the Mössbauer spectrum, that are of primary interest.

The EFG is determined by a number of different contributions. The most important are the asymmetry of the electronic charge distribution of the Mössbauer atom itself (due to the partially filled electronic shells occupied by the valence electrons) and the contribution from the lattice due to the asym-

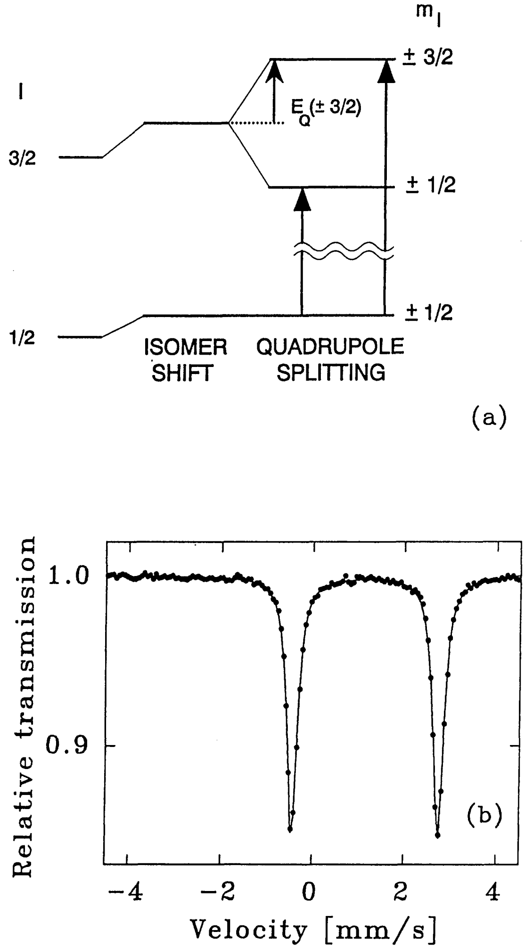

FIG. 7. The quadrupole splitting for a nucleus with $I_e = \frac{3}{2}$ and $I_g = \frac{1}{2}$ (e.g., ^{57}Fe): (a) splitting of the nuclear levels: (b) the Mössbauer spectrum of $FeSO_4 \cdot H_2O$ revealing the quadrupole splitting and isomer shift.

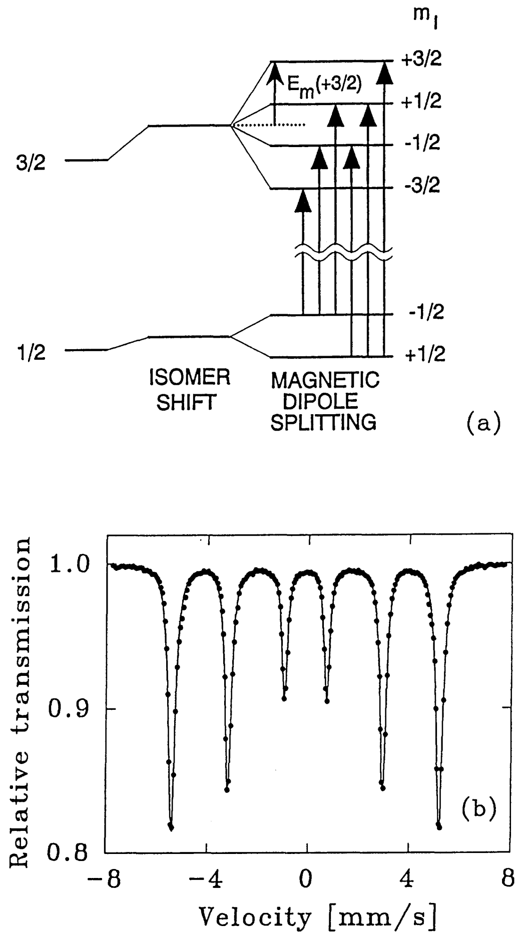

FIG. 8. The magnetic hyperfine splitting for a nucleus with $I_e = \frac{3}{2}$ and $I_g = \frac{1}{2}$ (e.g., ^{57}Fe): (a) the splitting of the nuclear levels and allowed gamma transitions ($\Delta m_I = 0, \pm 1$); (b) the Mössbauer spectrum of ferromagnetic α-Fe revealing the magnetic hyperfine splitting.

metric arrangement of the ligand atoms in a noncubic lattice.

The quadrupole splitting is completely determined by the symmetry of the bonding environment and the local structure in the vicinity of the Mössbauer atom. Thus, the quadrupole interaction is of great significance, particularly in chemistry and solid-state physics.

3.3 Magnetic Dipole Interaction

The magnetic hyperfine interaction couples the nuclear magnetic moment $\boldsymbol{\mu}$ to the magnetic field **H** acting at the site of the nucleus. The Hamiltonian is

$$\mathcal{H}_M = -\boldsymbol{\mu} \cdot \mathbf{H} = -g\mu_N \mathbf{I} \cdot \mathbf{H}, \tag{29}$$

where g is the nuclear g factor ($g = \mu/I\mu_N$) and μ_N is the nuclear Bohr magneton ($\mu_N = e\hbar/2Mc$). This interaction lifts the degeneracy of the nuclear levels and splits the nuclear state with spin I ($I > 0$) into $2I + 1$ equally spaced sublevels with energies

$$E_m = -\mu H m_I / I = -g\mu_N H m_I, \tag{30}$$

where m_I is the magnetic quantum number ($m_I = I, I - 1, \ldots, -I$).

The magnetic field H is the total magnetic field at the nuclear site, created by the atom's own electrons and any externally applied field. In the case of ^{57}Fe, the magnetic hyperfine interaction splits a single Mössbauer line into a six-line pattern (Fig. 8). Only the transitions with $\Delta m_I = 0, \pm 1$ are allowed. The relative line intensities are given by the angular dependence of each spectral line (Table 1).

From the line intensities, it is possible to determine the orientation of the magnetic field in the sample. In the general case, the line intensity ratio in the six-line spectrum is

$$3{:}\alpha{:}1{:}1{:}\alpha{:}3,$$

where

$$\alpha = \frac{4 \sin^2 \Theta_m}{1 + \cos^2 \Theta_m}.$$

The parameter α can vary from 0 to 4. The relative line intensities are 3:0:1:1:0:3 for $\Theta_m = 0$ (H parallel to the gamma-ray direction) and 3:4:1:1:4:3 for $\Theta_m = 90°$ (H per-

Table 1. Angular dependence of the allowed transitions in a Zeeman sextet of ^{57}Fe.

Transition	Δm_I	Angular dependence
$\pm\frac{3}{2} \rightarrow \pm\frac{1}{2}$	± 1	$(\frac{3}{4})(1 + \cos^2 \Theta_m)$
$\pm\frac{1}{2} \rightarrow \pm\frac{1}{2}$	0	$\sin^2 \Theta_m$
$\mp\frac{1}{2} \rightarrow \pm\frac{1}{2}$	∓ 1	$(\frac{1}{4})(1 + \cos^2 \Theta_m)$

Θ_m is the angle between the direction of the hyperfine field H and the propagation direction of the gamma rays.

pendicular to the gamma-ray direction). For randomly oriented fields, we have 3:2:1:1:2:3.

The magnetic hyperfine interaction Hamiltonian (29) contains a nuclear parameter (μ) and an atomic parameter (H), which cannot be separated experimentally. The situation is more favorable here than in the case of the isomer shift and quadrupole splitting because both nuclear states are split, and it is possible to determine in one experiment two of the three parameters that describe the interaction: μ_g, μ_e, and H. Usually, the magnetic moment of the ground state, μ_g, is known, or determined from other than Mössbauer experiments (e.g., microwave resonance, atomic beam experiments), so that μ_e and H can be found. One can also apply a sufficiently strong external magnetic field and observe the resulting splitting, whereas it is not possible to apply a sufficiently high external electric field gradient in the case of quadrupole interaction.

The magnetic hyperfine splitting in the spectra yields an effective magnetic field, at the site of the nucleus, originating from various sources that are difficult to separate. The total magnetic field experienced by the nucleus is a vector sum of the following contributions:

$$H_{\text{eff}} = H_0 - DM + \tfrac{4}{3}\pi M + H_s + H_L + H_D, \tag{31}$$

where H_0 is the magnetic field at the nucleus generated by the externally applied field (in the absence of the external field $H_0 = 0$), $-DM$ is the demagnetizing field, and $\frac{4}{3}\pi M$ is the Lorenz field. The largest contribution to H_{eff} results from the Fermi contact interaction (H_s). Its origin either is related to the intrinsic impairing of s electrons or, indirectly, results from polarization effects on the filled s orbitals due to unpaired d or f orbitals:

$$H_S = -(16/3)\pi\beta\langle\Sigma(s_\uparrow - s_\downarrow)\rangle, \tag{32}$$

where $s_\uparrow$ and $s_\downarrow$ are the s-electron spin densities at the nucleus with spin up and spin down, respectively, and β is the Bohr magneton.

The nonzero orbital magnetic moment of the atom gives rise to the field

$$H_L = -2\beta\langle 1/r^3\rangle\langle L\rangle = -2\beta\langle 1/r^3\rangle(g-2)\langle S\rangle, \tag{33}$$

where $\langle L\rangle$ and $\langle S\rangle$ are the expectation values of the orbital and spin angular momentum, respectively, and g is the Lande factor. Evidently, $H_L = 0$ if $L = 0$.

The dipolar interaction of the nucleus with the spin of the atom creates the field

$$H_D = -2\beta\langle 3\mathbf{r}(\mathbf{S}\cdot\mathbf{r})/r^5 - \mathbf{S}/r^3\rangle. \tag{34}$$

In cubic metals, in the absence of spin-orbit coupling, $H_D = 0$.

The terms H_s, H_L, and H_D can all be of the order of 1 to 10 T (10^4–10^5 G), and their sum is usually called the "internal magnetic field."

3.4 Combined Magnetic Dipole and Electric Quadrupole Interactions

In the presence of the combined hyperfine interactions, the complete Hamiltonian is

$$\begin{aligned}\mathcal{H} &= \mathcal{H}(E_0) + \mathcal{H}_Q + \mathcal{H}_M = \\ &= \mathcal{H}(E_0) - g\mu_N \hat{I}'_z H + \frac{e^2qQ}{4I(2I-1)} \\ &\quad \times [3\hat{I}_z^2 - I(I+1) + \eta(\hat{I}_x^2 - \hat{I}_y^2)],\end{aligned} \tag{35}$$

where $\mathcal{H}(E_0)$ corresponds to the isomer shift and results in a shift of all spectral lines: I'_z and $\hat{I}_z$ are defined with respect to two different axis systems related to the orientations of the magnetic field or EFG axis relative to the gamma-ray direction. There is no analytical solution of the full Hamiltonian, and so a complete numerical analysis is necessary. In general, additional normally "forbidden" transitions appear. However, usually one interaction is small compared to the other. In the case of magnetically ordered systems, the electric quadrupole interaction is much weaker than the magnetic dipole one and causes a perturbation of the magnetic interaction. In the case of paramagnetic or diamagnetic systems, the applied external magnetic field causes a perturbation of the quadrupole interaction.

The Hamiltonian (35) has simple solutions in the following cases of ^{57}Fe Mössbauer spectroscopy.

1. Axially symmetric EFG tensor with symmetry axis parallel to the magnetic field H. For $I = \frac{3}{2}$, the energies of the magnetically split nuclear levels are shifted by the same amount as compared with the pure

magnetic splitting:

$$E = -g\mu_N H m_I + (-1)^{|m_I|+1/2}\, e^2qQ/4. \tag{36}$$

2. Axially symmetric EFG tensor with symmetry axis at angle Θ with respect to H and $e^2qQ \ll \mu H$ and $I = \frac{3}{2}$:

$$E = -g\mu_N H m_I + (-1)^{|m_I|+1/2} \times \frac{e^2qQ}{4}\,\frac{3\cos^2\Theta - 1}{2}. \tag{37}$$

   The angle Θ cannot be determined from the spectrum, and so e^2qQ cannot be calculated unless the direction of H relative to the symmetry axis of the solid is measured independently.

3. Axially nonsymmetric EFG tensor with H along one of its principal axes ($I = \frac{3}{2}$):

$$E = \begin{cases} \dfrac{1}{2} g\mu_N H \pm \dfrac{e^2qQ}{4}\left[\left(1 + \dfrac{4g\mu_N H}{e^2qQ}\right)^2 + \dfrac{\eta^2}{3}\right]^{1/2}, \\ -\dfrac{1}{2} g\mu_N H \pm \dfrac{e^2qQ}{4}\left[\left(1 - \dfrac{4g\mu_N H}{e^2qQ}\right)^2 + \dfrac{\eta^2}{3}\right]^{1/2}. \end{cases} \tag{38}$$

The combined hyperfine interactions allow us to determine the sign of the quadrupole coupling, which in favorable conditions may allow the determination of the sign of the nuclear quadrupole moment.

3.5 Information Concerning the Properties of Solids from the Mössbauer Effect

The Mössbauer effect yields most information from the hyperfine interactions. As discussed above, the hyperfine interactions include nuclear and atomic parameters that are difficult to separate. However, for the most important isotopes, such as ^{57}Fe and ^{119}Sn, the nuclear parameters are well known and can be simply put into the fitting Hamiltonian. Most of the applications of the Mössbauer effect concern properties of solids, the nuclear parameters being of much less interest, since for a given gamma transition, they are constant. The isomer shift provides valuable chemical information. It reflects the changes in the electron density at the nucleus due to changes in the valence orbital population of the Mössbauer atom. The electron density is related to the type of chemical bonding, covalency effect, oxidation and reduction processes, and differences in the electronegativity of ligands coordinated to the Mössbauer atom. The study of a single compound is of little value for obtaining useful information unless data for other compounds are available for comparison. Only then is it possible to correlate isomer-shift data with the oxidation and spin state of the atom in a complex compound, the coordination between the Mössbauer atom and bonded ligands, etc. The richest data have been collected for iron compounds. The isomer shift is sensitive enough to the spin state of iron in a complex compound to allow distinguishing between high-spin and low-spin divalent and trivalent iron. The isomer-shift data alone make the distinction between low-spin iron(II) and iron(III) difficult, but with the help of quadrupole splitting [which is usually substantially larger for iron(III) compounds], they can be readily identified. High-spin iron compounds can be distinguished by isomer-shift data alone.

The quadrupole interaction provides information about the electric field gradient at the nucleus site, which can originate either from the valence electrons of the Mössbauer atom, when it is associated with asymmetry in the electronic structure, or from a nonspherical charge distribution in the ligand sphere and/or surrounding lattice with symmetry lower than cubic. Molecular orbitals can also contribute to the EFG. The effects of these contributions at the nucleus site are modified by the polarization of the core electrons of the Mössbauer atom, which may reduce or enhance the EFG. The quadrupole splitting observed in Mössbauer spectra of a given solid reflects the symmetry of the bonding environment and the local structure in the vicinity of the Mössbauer atom. Quadrupole splitting data are especially useful when combined with isomer-shift data. They yield chemical information concerning the electronic population of various orbitals, ligand structure, and structural information about the local atomic arrangement in both crystalline and amorphous solids. Quadrupole splitting is a particularly sensitive probe of short-range order in metallic glasses.

The magnetic hyperfine interaction yields information on the magnetic properties of solids, such as magnetic ordering and structure of magnetically ordered systems, the nature of the magnetic interaction (ferro- or antiferro-), and the magnitude of the magnetic moment on a particular atom. This information results from studying the magnetic hyperfine fields at the nucleus site. However, in order to compare such results with those obtained by conventional methods (e.g., magnetization measurements), phenomenological formulas have to be used to convert the information at the site of the nucleus to those at the site of the atom. Thus, the estimation of the magnetic moment of the atom is indirect. The temperature dependence of the hyperfine field is fortunately the same as that of magnetization; hence, the onset of magnetic ordering and magnetic phase transitions can be studied directly by the Mössbauer technique. Also, the shapes of the $H_{hf}(T)$ and $M(T)$ curves are identical and allow the study of deviation of experimental $H_{hf}(T)$ curves from the Brillouin function. Combined electric quadrupole and magnetic dipole interactions allow study of the direction of the magnetic field with respect to the crystallographic axes. The application of an external magnetic field to a system with no unpaired spins, which has no magnetic hyperfine field, leads to a magnetic splitting in addition to the quadrupole splitting and provides information on the geometry of the EFG at the nucleus.

From the complex spectra revealing a poorly resolved magnetic hyperfine structure of ferromagnetic amorphous alloys, it is possible to extract the distributions of hyperfine fields $P(H)$, which provide valuable information on the magnetic properties and structure of such materials.

The Mössbauer effect allows us to study not only static magnetic fields but also time-dependent relaxation effects. Every spectroscopy has a characteristic observation time of interaction of the radiation with the solid. The observation time of Mössbauer spectroscopy corresponds to the mean lifetime (τ_N) of the nuclear resonant level. For hyperfine interaction measurements, with the characteristic energy ΔE, it corresponds to the Larmor precession period $\tau_L = \hbar/\Delta E$. The times τ_N and τ_L are typically of the order of 10^{-7}–10^{-8} s. In the presence of time-dependent changes in the environment of the resonant nucleus, the experimental observation (shape of the Mössbauer spectrum and the hyperfine parameters) depends on the relative order of magnitude of the observation time (τ_N or τ_L) as compared with the residence time of the fluctuating environment, τ_R (relaxation time). When the observation time is much longer than the relaxation time (τ_N, $\tau_L \gg \tau_R$, slow relaxation), the quasistatic situation is observed. If the opposite relation occurs (τ_N, $\tau_L \ll \tau_R$, fast relaxation), a fully time-averaged situation occurs. For intermediate relaxation rates, we get complex spectra.

The time dependence can be introduced into the hyperfine interactions by using an external high-frequency magnetic field. When the frequency of such a field is higher than the Larmor frequency, the magnetic hyperfine splitting collapses and the Mössbauer spectrum consists of a single line or a quadrupole doublet instead of a Zeeman sextet. The shape of the spectra depends on the relation between the frequency of the field and the Larmor frequency. The relaxation effect induced by the radio-frequency field is fully controlled by the field frequency and intensity (Pfeiffer, 1972; Kopcewicz, 1989). The rf collapse of the magnetic hyperfine structure can be induced only in soft ferromagnetic materials.

4. APPLICATIONS OF THE MÖSSBAUER EFFECT IN VARIOUS FIELDS

The applications of the Mössbauer effect are so extensive in various branches of science and technology that only a few will be very briefly mentioned here. The interdisciplinary nature of the effect is demonstrated by applications as different as the theory of relativity, magnetism, metal physics, biology, nuclear physics, mineralogy, atmospheric physics, and even art. Most important, however, are the applications in solid state physics and chemistry. Some of the most rewarding applications of the Mössbauer spectroscopy were made in the field of magnetism (Thomas and Johnson, 1986, and, e.g., Long, 1987, and Long and Grandjean, 1989). Even though the various contributions to the effective magnetic field at the nucleus are difficult to separate experimentally and the magnetic moment associated with the atom is not necessarily proportional to the hyperfine field, valuable information on the type of magnetism, the temperature dependence of magnetization, and magnetic phase transitions was obtained. More exotic effects, such as spin-glass behavior, low-dimensional magnetism (Johnson, 1984), and superparamagnetism of ultrafine particles (Mørup, 1987), also have been studied. Magnetic properties of solids were investigated as a function of temperature (e.g., Thosar *et al.*, 1983), pressure (Williamson, 1978), and external static (e.g., Long, 1984; Long and Grandjean, 1989, 1993) and high-frequency magnetic fields (Kopcewicz, 1991b). A great variety of magnetic materials, such as alloys, intermetallic compounds, and metallic glasses, have received a lot of attention (see, e.g., Cranshaw, 1983). The Mössbauer effect greatly contributed to better understanding of Invar alloys, rare-earth magnetism (e.g., Thosar *et al.*, 1983; Long and Grandjean, 1993), and amorphous transition-metal–metalloid and metal-metal alloys (Longworth, 1987; Gonser, 1981). The thermal stability, crystallization, magnetic properties, and structure (short-range order) were extensively studied, and Mössbauer spectroscopy proved to be a powerful technique for such investigations. Recently, the possibility to form nanocrystalline phases of exceptional soft magnetic properties by controlled crystallization of some amorphous alloys was recognized, and the Mössbauer effect significantly contributed to the identification of the various nanocrystalline and microcrystalline phases formed.

The time-dependent effects readily observable by Mössbauer spectroscopy include electronic, spin, or lattice relaxation effects, the spin-flip processes in paramagnetic ions related to the spin-spin interaction with spins of the neighboring ions and spin-lattice interaction between electronic spins and phonons, valence fluctuations, after-effects of nuclear transitions, structural changes both macroscopic (diffusion) and microscopic (rotation within the molecule), and fluctuations in electronic spin orientation. The magnetic relaxation effects (time dependence of the magnitude and direction of the magnetic hyperfine field experienced by the nucleus) and the superparamagnetic relaxation in ultrafine particles were extensively studied. The fast magnetization reversal induced by an external magnetic radio-frequency field, causing a collapse of the magnetic hyperfine structure, can be observed in the Mössbauer spectra of soft crystalline and amorphous ferromagnets. The rf-collapse effect provides a unique possibility to separate the magnetic dipole and electric quadrupole interactions in the ferromagnetic state. The quadrupole splitting distributions can be extracted from the collapsed spectra, which allows a direct study of the short-range order in ferromagnetic amorphous alloys (Kopcewicz, 1991b).

The surface, thin-film, and multilayer structures were successfully studied by using CEMS and conversion x-ray Mössbauer spectroscopy (CXMS). The CEMS technique opened a wide class of applications to ion-beam modification of materials. The effects of ion implantation (especially nitrogen, carbon, boron, oxygen) into iron and steels are important for both science and technology (Williamson, 1993). Formation of new materials as a result of ion implantation or ion-beam mixing of multilayered structures, and the ion-induced transformation from crystalline to amorphous state, have been characterized by CEMS in various metallic systems (e.g., B-implanted α-Fe, ion-beam mixing of multilayers). Phenomena such as radiation damage, defects, consequences of the nuclear transformations in solids, and after-effects of radioactive decay affect the Mössbauer spectra. Ion implantation of radioactive Mössbauer isotopes into metal hosts allowed study of the migration of vacancies and interstitials, radiation damage cascades, cluster formation and dissolution, and migration of implanted impurities. The phenomena of internal oxidation and helium decoration were investigated (De Waard and Niesen, 1987). Ion implantation in semiconductors was also studied (Langouche, 1991).

The sophisticated technique of in-beam Mössbauer spectroscopy, developed already in the 1960s, makes use of Mössbauer nuclei excited directly by Coulomb excitation that simultaneously causes recoil implantation. In this way, a Mössbauer source with lifetime of the excited Mössbauer state is formed. No parent radioactive isotope is involved. Recoil implantation can be performed into any host material. In-beam Mössbauer spectroscopy yielded valuable information concerning radiation damage and the behavior of the implanted Mössbauer atoms in various materials. Pioneering studies were reviewed by Sprouce and Kalvius (1968) and Obenshain

(1968), and more recent results were presented by Sielemann (1991).

The surface magnetism of ultrathin iron films was investigated by conversion-electron and transmission Mössbauer spectroscopy. The dependence of the magnetic hyperfine field vs distance from the surface was studied by suspending one or only a few monolayers of ^{57}Fe in the ^{56}Fe layer at a well-defined depth; this yielded a good depth sensitivity. The experimentally determined hyperfine field as a function of depth was compared with theoretical predictions. Recent results are reviewed by Przybylski *et al.* (1991).

The Mössbauer effect was successfully applied for studying semiconductors (Si, GaAs) and solar cell materials (Williamson *et al.*, 1992).

When a new class of materials, such as high-T_C superconductors, appeared, the Mössbauer effect yielded useful structural information. Also, conventional superconductors have been studied (Shenoy, 1983).

Applications of the Mössbauer effect in metallurgy relate to steel research, phase formation, and phase transformation with respect to steel processing (Fujita, 1975).

The applications of the Mössbauer effect in magnetism are extensively reviewed in all books listed in Further Reading.

Chemical applications of the Mössbauer effect exploit its sensitivity to changes in electron density at the nucleus. The main fields of interest concern the electronic structure of the Mössbauer atom, oxidation states, spin states, bond properties (e.g., σ and π interaction of the Mössbauer atom with ligands and ionicity and covalency of bonds), structural properties (molecular geometry, polymerization), and dynamic processes (such as electron hopping in spinels, spin-spin and spin-lattice relaxation). The analytical potential of Mössbauer spectroscopy, especially involving the use of the isotope ^{57}Fe, was widely used for identification of various phases containing iron even in complex systems. It is possible to determine the iron content in various phases; however, the measurements of the relative concentrations are much more reliable than the determination of the iron content in absolute units. Reactions in solids, thermal- and radiation-induced decomposition, surface studies (e.g. corrosion process, adsorption, and surface reactions), and frozen solutions have been investigated. In the field of catalysis, Mössbauer spectroscopy has allowed *in situ* study of solids that catalyze gaseous reactions (see Gütlich *et al.*, 1978; Thosar *et al.*, 1983; Long, 1984; Long and Grandjean, 1987, 1989; Goldanskii and Herber, 1968; Greenwood and Gibb, 1971).

The Mössbauer effect has contributed also to nuclear physics. It is useful in the determination of nuclear parameters, especially of the excited states involved in the Mössbauer gamma transitions. Comparison of the experimentally determined parameters such as lifetime of the excited state, nuclear spin and magnetic and quadrupole moments, change of the nuclear radius, and internal conversion coefficient with the predictions of theoretical models is used for evaluation of the validity of the nuclear models. Especially interesting for nuclear physics are the Mössbauer transitions in rare-earth isotopes because of the great number of gamma transitions available, allowing the study of the effects of successive proton or neutron addition over a range of nuclei (consult Goldanskii and Herber, 1968; Greenwood and Gibb, 1971).

Mössbauer spectroscopy has contributed significantly to the better understanding of some biological systems (Johnson, 1975). As many biological molecules contain iron, the Mössbauer technique yields information on the local environment of iron in large and complex molecules in which iron atoms are at the active centers where biological changes occur. The most studied biological systems are iron proteins, such as hemoproteins (hemoglobin, myoglobin, cytochromes, catalases), iron sulfur proteins, and iron storage and transport proteins (ferritin). The structural dynamics of proteins has been investigated (Parak, *et al.* 1988). The Mössbauer effect has also found medical applications involving studies of tissue and blood samples that lead to identification of pathological factors.

In mineralogy, the Mössbauer effect has been used for studying chemical properties of various minerals, e.g., silicate minerals such as pyroxene and olivine (Bancroft, 1973). Phase identification and information concerning site population of Fe^{2+} and Fe^{3+} ions are obtained from the Mössbauer measurements. Many oxides (hematite, magnetite, goethite, etc.) and ores have been investigated. Studies of meteorites have been performed in view of their importance for reconstruction of the early history of the solar system. Even lunar samples delivered to the Earth by spacecraft were investigated by Mössbauer spectroscopy (Hafner, 1975). The Mössbauer effect has been applied for studying iron-containing atmospheric aerosol. Not only the iron concentration in the atmospheric air, but also the chemical compounds in which iron is contained and even the size of iron-containing particles of atmospheric aerosol can be evaluated. Such studies provide information concerning the sources of iron in the atmosphere, both natural and related to human activity (Kopcewicz and Kopcewicz, 1991).

The Mössbauer effect has been applied to archaeological problems. The study of ancient pottery and other archaeological artifacts provides information about the manufacturing technology and has revealed the level of technology in ancient times. The Mössbauer technique has been applied to the study of ancient bronzes and casting and firing techniques of ancient metallurgy (Zheng and Hsia, 1991).

The above list of applications is by far not complete but already shows the exceptional power of Mössbauer spectroscopy as a technique that can provide valuable information in many areas of science and technology.

5. EXOTIC MÖSSBAUER STUDIES

One of the most spectacular Mössbauer experiments performed in the early years of the "Mössbauer era" concerned the verification of the gravitational red shift predicted by the Einstein theory of relativity. Before the discovery of the Mössbauer effect, there was no experimental method that could measure the change of energy (or frequency) of electromagnetic radiation with sufficient precision to allow the measurement of the gravitational red shift in the laboratory. The relevant parameter is the high energy resolution of the Mössbauer effect ($\Gamma/E_\gamma \cong 3 \times 10^{-13}$ for ^{57}Fe or 5.2×10^{-16} for ^{67}Zn). The gravitational red shift can be considered as the change of the energy of a photon as it moves from one place to another of different gravitational potential. Let us consider a Mössbauer experiment in which a source is placed at a height h above the absorber; then the gamma photon travels a distance h through a gravitational field and interacts with it as though it had a mass of E_γ/c^2. As a result of this interaction, the energy of the photon changes by a fraction equal to

$$\delta E/E_\gamma = gh/c^2, \tag{39}$$

where g is the acceleration of the gamma photon in the gravitational field. This energy change is the gravitational red shift, which shifts the position of the Mössbauer resonance line by a factor of 1.1×10^{-16} m^{-1}. The experimental verification of this relativistic effect was performed by Pound and Rebka (1960) with the use of ^{57}Fe 14.4-keV resonance with a height separation between the source and absorber of 22.5 m. The experimentally measured shift was 0.94 ± 0.10 of that predicted. Later, the experiments gave an even more accurate value of 0.997 ± 0.0076 (Pound and Snider, 1965). In these studies, special care was taken to eliminate other than gravitational sources of the line shift, e.g., those due to the difference of the temperature at the sites of the source and absorber (second-order Doppler shift). Although the energy resolution of isotope ^{67}Zn is much better than that of ^{57}Fe, which makes the experiment with ^{67}Zn appear relatively simpler, the technical difficulties associated with the ^{67}Zn resonance prevented the use of this isotope and made the isotope ^{57}Fe a more attractive choice.

Experiments related to the red shift have also been performed in which the gravitational acceleration was replaced by a kinematic acceleration by means of a high-speed rotor with the source in the center and the absorber at the periphery (transverse Doppler effect); however, the results were not conclusive.

Another relativistic effect influencing the position of the Mössbauer resonance line is the thermal red shift (second-order Doppler shift) discussed already in Sec. 3.1, Eq. (26).

In 1984, new exotic experiments with the Mössbauer effect were performed. Synchrotron radiation was used as a source of the Mössbauer radiation (Gerdau, *et al.* 1985). The property that makes the synchrotron radiation especially attractive for Mössbauer studies is its exceptional brilliance, which exceeds the maximum brilliance of a conventional ^{57}Fe source by 1–2 orders of magnitude.

Mössbauer experiments with synchrotron radiation combine the unique properties of the Mössbauer effect, namely, its exceptional energy resolution, with the unique properties of synchrotron radiation such as outstanding brilliance, a high degree of linear polarization, and the pulsed structure of the beam. In order to use synchrotron radiation as a source of the Mössbauer radiation, the broad energy band coming from the storage ring must be monochromatized. Electronic monochromators as well as a nuclear monochromator, making use of the pure nuclear Bragg reflections that select a very narrow band corresponding to the width of the nuclear level Γ (GIAR films), were used (see, e.g., Rüffer, *et al.* 1990).

The experiments with synchrotron radiation are usually performed not on the conventional velocity (energy) scale but on a time scale. Synchrotron radiation opened new fields for Mössbauer studies, such as gamma optics and soft solid state excitations. Nuclear resonant scattering was studied in the spatially coherent decay channels in the Bragg and forward directions. Time evolution of the nuclear Bragg scattering was observed. The brilliance of the synchrotron radiation source allows experiments that are difficult in conventional Mössbauer spectroscopy, e.g., high-pressure and small-crystal investigations. The hyperfine parameters, especially the isomer shift, can be determined with much higher accuracy because of the quantum beat effect.

At present, there are several synchrotron radiation sources available for Mössbauer experiments: the European Synchrotron Radiation Facility in Grenoble (France), DESY in Hamburg (Germany), Stanford Synchrotron Radiation Laboratory in Stanford (U.S.A.), the Advanced Photon Source at Argonne National Laboratory in Argonne (U.S.A.), and the TRISTAN Accumulation Ring at KEK (Japan).

Since 1984, the technique of time-differential Mössbauer spectroscopy has developed and matured rapidly. Beginning as an exotic experiment, it is now on the way to becoming a widely used tool of Mössbauer spectroscopy.

6. PROSPECTS FOR THE FUTURE

Thirty-five years have elapsed since the discovery of the Mössbauer effect. During this time, the effect developed from an experimental curiosity to a commonly used spectroscopy that is finding wide applications in most disciplines of natural science. Nature was kind to provide a perfect isotope for Mössbauer spectroscopy—^{57}Fe, thanks to which a great variety of materials is accessible to Mössbauer studies. Mössbauer spectroscopy was quick to enter any new field that appeared in the last 35 years. Typical examples can be found in materials science. When new materials, such as amorphous alloys, high-T_C superconductors, nanocrystals, quasicrystals, and fullerenes appeared, the Mössbauer effect was immediately and successfully applied for the study of microstructure, chemical, and magnetic properties. Mössbauer spectroscopy was also combined with new experimental facilities of nuclear physics. Synchrotron radiation was used as a Mössbauer source opening new horizons. Recently, the possibility of using the Mössbauer effect to investigate the interaction of high-energy heavy ions with solids has attracted considerable attention. It seems that in the near future the Mössbauer experiments will involve combination with large installations, such as synchrotrons, cyclotrons, and other accelerators, to explore new frontiers of nuclear and solid-state physics. Mössbauer spectroscopy will certainly enter any new field of natural science that may appear in the future. The practical and technological applications, which still lag behind purely scientific research, offer wide and unexploited possibilities for future developments.

In his Nobel Laureate lecture in December 1961, Rudolf Mössbauer concluded, "We may therefore hope that this young branch of physics stands only at its threshold and that it will be developed in future, not only to extend the application of existing knowledge but to make possible new advances in the exciting world of unknown phenomena and effects." A lot has been done already. The effect has developed and matured, but Mössbauer's remark did not lose much of its novelty. Thirty-three years later Mössbauer spectroscopy has enormous and still unexploited possibilities for future applications.

GLOSSARY

Debye–Waller Factor: Originally derived in the Bragg x-ray scattering theory, describes the effect of temperature on the intensity of x-rays scattered elastically from the lattice. The Debye–Waller factor describes also the probability of the recoil-free gamma transition.

Doppler Broadening: Broadening of the lines emitted and absorbed by the atoms in a gas due to the thermal motion of atoms.

EFG: Electric field gradient.

***f* Factor:** See **Debye–Waller Factor**.

Isomer Shift: Shift of a whole Mössbauer spectrum relative to the reference standard due to the shift of the nuclear levels as a result of the interaction of nuclear charge with the electron charge density at the nucleus.

Lorentzian Line Shape: The shape of a spectral line in which the variation of intensity with energy is given by the reciprocal of a quadratic expression.

Magnetic Hyperfine Structure: Splitting of the nuclear levels due to the interaction between the nuclear magnetic moment with the effective magnetic field at the nucleus site; observed in the Mössbauer spectra as a splitting of the Mössbauer lines.

Mössbauer Effect: Recoil-free resonance absorption or scattering of nuclear gamma radiation occurring for atoms bound in solids.

Natural Linewidth: The width of the gamma emission line, related to the mean lifetime of the nuclear state, resulting from the Heisenberg uncertainty principle: $\Gamma\tau = h$.

Nuclear Zeeman Effect: See **Magnetic Hyperfine Structure.**

Quadrupole Splitting: Splitting of the nuclear levels due to the interaction of the nuclear quadrupole moment with the electric field gradient at the nucleus site; observed in the Mössbauer spectra as a splitting of the spectral lines.

Recoil Energy: The energy transferred to the nucleus due to the emission of gamma radiation from a free nucleus.

Second-Order Doppler Shift: The relativistic effect, related to thermal motion of atoms, that causes a temperature-dependent shift of the Mössbauer spectrum.

Velocity Transducer: An electromechanical device (essential part of the spectrometer) in which the source is moved with respect to the absorber (or vice versa); this provides the Doppler modulation of the Mössbauer gamma radiation and allows energy scan.

Works Cited

Bancroft, G. M. (1973), *Mössbauer Spectroscopy*, London: McGraw-Hill.

Cranshaw, T. E. (1983), in: B. V. Thosar, P. K. Iyengar, J. K. Srivastava, S. G. Bhargava (Eds.), *Advances in Mössbauer Spectroscopy*, Amsterdam: Elsevier, pp. 271–272.

De Waard, H., Niesen, L. (1987), in: G. J. Long (Ed.), *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, Vol. II, New York: Plenum, pp. 1–86.

Fujita, F. E. (1975), in: U. Gonser (Ed.), *Mössbauer Spectroscopy*, Berlin: Springer, pp. 201–236.

Gerdau, E., Rüffer, R., Winkler, H., Tolksdorf, W., Klages, C. P., Hannon, J. P. (1985), *Phys. Rev. Lett.* **54**, 835–838.

Goldanskii, V. I., Makarov, E. F., Khrapov, V. V. (1963), *Phys. Lett.* **3**, 344–346.

Goldanskii, V. I., Herber, R. H. (1968), *Chemical Applications of Mössbauer Spectroscopy*, New York: Academic Press.

Gonser, U. (1975), *Mössbauer Spectroscopy*, Berlin: Springer.

Gonser, U. (1981), *At. Energy Rev.*, Suppl. No. 1, 302–228.

Greenwood, N. N., Gibb, T. C. (1971), *Mössbauer Spectroscopy*, London: Chapman and Hall.

Gütlich, P., Link, R., Trautwein, A. (1978), *Mössbauer Spectroscopy and Transition Metal Chemistry*, Berlin: Springer.

Hafner, S. S. (1975), in: U. Gonser (Ed.), *Mössbauer Spectroscopy*, Berlin: Springer, pp. 167–200.

Hanna, S. S., Heberle, J., Littlejohn, C., Perlow, G. J., Preston, R. S., Vincent, D. H. (1960a), *Phys. Rev. Lett.* **4**, 28–29.

Hanna, S. S., Heberle, J., Littlejohn, C., Perlow, G. J., Preston, R. S., Vincent, D. H. (1960b), *Phys. Rev. Lett.* **4**, 177–180.

Johnson, C. E. (1975), in: U. Gonser (Ed.), *Mössbauer Spectroscopy*, Berlin: Springer, pp. 139–166.

Johnson, C. E. (1984), in: G. J. Long (Ed.), *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, Vol. I, New York: Plenum, pp. 619–640.

Kistner, O. C., Sunyar, A. W., (1960), *Phys. Rev. Lett.* **4**, 412–415.

Karyagin, S. V. (1963), *Dokl. Acad. Nauk SSSR*, 148, 1102–1105.

Kopcewicz, M. (1989), in: G. J. Long and F. Grandjean (Eds.), *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, Vol. III, New York: Plenum, pp. 243–287.

Kopcewicz, M. (1991), *Struct. Chem.* **2**, 313–342.

Kopcewicz, B., Kopcewicz, M. (1991), *Struct. Chem.* **2**, 303–312.

Lamb, W. E., Jr. (1939), *Phys. Rev.* **55**, 190–197.

Langouche, G. (1991), *Hyperfine Int.* **68**, 95–106.

Long, G. J. (Ed.) (1984), *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, Vol. I; (1987), Vol. II, New York: Plenum.

Long, G. J., Grandjean, F. (Eds.) (1989), *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, Vol. III, New York: Plenum.

Longworth, G. (1987), in: G. J. Long (Ed.), *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, Vol. II, New York: Plenum, pp. 289–342.

Moon, P. B. (1950), *Proc. Phys. Soc.* **63**, 1189–1196.

Mössbauer, R. L. (1958a), *Z. Phys.* **151**, 124–143.

Mössbauer, R. L. (1958b), *Naturwissenschaften* **45**, 538–539.

Mørup, S. (1987), in: G. J. Long (Ed.), *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, Vol. II, New York: Plenum, pp. 89–123.

Obenshain, F. E. (1968), in: I. J. Gruvermann (Ed.), *Mössbauer Effect Methodology*, Vol. 4, New York: Plenum, pp. 61–74.

Parak, F., Heidemeier, J., Nienhaus, G. U. (1988), *Hyperfine Int.* **40**, 147–158.

Pfeiffer, L. (1972), in: I. J. Gruvermann (Ed.), *Mössbauer Effect Methodology*, Vol. 7, New York: Plenum, pp. 263–298.

Pound, R. V., Rebka, G. A., Jr. (1960), *Phys. Rev. Lett.* **4**, 337–341.

Pound, R. V., Snider, J. L. (1985), *Phys. Rev.* **140**, B788–B803.

Przybylski, M., Korecki, J., Gradmann, U. (1991), *Appl. Phys.* **A52**, 33–47.

Rüffer, R., Gerdau, E., Grote, M., Hollatz, R., Röhlsberger, R., Rüter, H. D., Sturhahn, W. (1990), *Hyperfine Int.* **61**, 1279–1294.

Schiffer, J., Marshall, W. (1959), *Phys. Rev. Lett.* **3**, 556–557.

Sielemann, R., Yoshida, Y. (1991), *Hyperfine Int.* **68**, 119–130.

Sprouce, G . D., Kalvius, G. M. (1968),. in: I. J. Gruvermann (Ed.), *Mössbauer Effect Methodology*, Vol. 4, New York: Plenum, pp. 37–59.

Shenoy, G. K. (1983), in: B. V. Thosar, P. K. Iyengar, J. K. Srivastava, S. G. Bhargava (Eds.), *Advances in Mössbauer Spectroscopy*, Amsterdam: Elsevier, pp. 561–585.

Shenoy, G. K., Wagner, F. E. (Eds.) (1978), *Mössbauer Isomer Shifts*, Amsterdam: North Holland.

Thomas, M. F., Johnson, C. E. (1986), in: D. P. E. Dickson, F. J. Berry (Eds.), *Mössbauer Spectroscopy*, Cambridge: Cambridge University Press, pp. 143–197.

Thosar, B. V., Iyengar, P. K., Srivastava, J. K., Bhargava, S. G. (Eds.), (1983), *Advances in Mössbauer Spectroscopy*, Amsterdam: Elsevier.

Williamson, D. L. (1978), in: G. K. Shenoy, F. E. Wagner (Eds.), *Mössbauer Isomer Shifts*, Amsterdam: North-Holland, pp. 317–360.

Williamson, D. L., Niesen, L., Weyer, G., Sielemann, R., Langouche, G. (1992), in: G. Langouche (Ed.), *Hyperfine Interaction of Defects in Semiconductors*, Amsterdam: Elsevier, pp. 3–75.

Williamson, D. L. (1993), *Nucl. Instr. Methods* **B76**, 262–267.

Zheng, Y. F., Hsia, Y. F. (1991), *Hyperfine Int.* **68**, 131–142.

Further Reading

Cohen, R. L. (Ed.) (1976), *Application of Mössbauer Spectroscopy*, Vol. I, New York: Academic Press.

Cohen, R. L. (Ed.) (1980), *Application of Mössbauer Spectroscopy*, Vol. II, New York: Academic Press.

Dickson, D. P. E., Berry, F. J. (1986), *Mössbauer Spectroscopy*, Cambridge: Cambridge University Press.

Goldanskii, V. I., Herber, R. H. (1968), *Chemical Applications of Mössbauer Spectroscopy*, New York: Academic Press.

Gonser, U. (Ed.) (1975), *Mössbauer Spectroscopy*, Berlin: Springer.

Gonser, U. (Ed.) (1981), *Mössbauer Spectroscopy II, The Exotic Side of the Effect*, Berlin: Springer.

Greenwood, N. N., Gibb, T. C. (1971), *Mössbauer Spectroscopy*, London: Chapman and Hall.

Gruvermann, I. J. (Ed.) (1965–1975), *Mössbauer Effect Methodology*, Vols. 1–10, New York: Plenum.

Gütlich, P., Link, R., Trautwein, A. (1978), *Mössbauer Spectroscopy and Transition Metal Chemistry*, Berlin: Springer.

Long, G. J. (Ed.) (1984), *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, Vol. I; (1987), Vol II, New York: Plenum.

Long, G. J., Grandjean, F. (Eds.) (1989), *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, Vol. III, New York: Plenum.

Long, G. J., Grandjean, F. (Eds.) (1993), *Mössbauer Applied to Magnetism and Materials Science*, New York: Plenum.

Shenoy, G. K., Wagner, F. E. (Eds.) (1978), *Mössbauer Isomer Shifts*, Amsterdam: North-Holland.

Thosar, B. V., Iyengar, P. K., Srivastava, J. K., Bhargava, S. G. (Eds.) (1983), *Advances in Mössbauer Spectroscopy*, Amsterdam: Elsevier.

Vertes, A., Korecz, L., Burger, K. (1979), *Mössbauer Spectroscopy*, Amsterdam: Elsevier.

Werthelm, G. K. (1964), *Mössbauer Effect: Principles and Applications*, New York: Academic Press.

MRI

See MAGNETIC RESONANCE IMAGING

MUON SPIN ROTATION/RELAXATION/ RESONANCE

Jess H. Brewer, *Canadian Institute of Advanced Research and Department of Physics, University of British Columbia, Vancouver, British Columbia, Canada*

Introduction 24
1. **Basic Muon Physics** 25
1.1 Muon Production 25
1.2 Muon Decay 26
1.2.1 Asymmetry Calibration 27
1.3 The "Heavy Electron" versus the "Light Proton" 28
2. **μSR Techniques** 29
2.1 Time-Differential μSR in Transverse Field 29
2.1.1 The Basic Technique 30
2.1.1.1 Frequency 30
2.1.1.2 Asymmetry 31
2.1.1.3 Initial Phase of Precession ... 31
2.1.1.4 "Relaxation" 31
2.1.1.5 Muonium Precession 32
2.1.2 High Field 32
2.1.2.1 Fourier Transforms 32
2.1.2.2 Rotating Reference Frames .. 33
2.1.3 Paramagnetic States 34
2.2 Time-Differential μSR in Longitudinal and Zero Field 34
2.2.1 Two-Counter Asymmetry 35
2.2.1.1 The "Corrected" Asymmetry 36
2.2.2 Zero Field 36
2.2.3 "Relaxation" 36
2.2.3.1 Nuclear Dipole Relaxation ... 37
2.2.3.2 Motional Narrowing 37
2.2.3.3 Nuclear Dipole Oscillations .. 37
2.2.4 True Relaxation 38
2.2.4.1 Longitudinal-Field Muonium Relaxation 38
2.3 Time-Integral μSR 38
2.3.1 Transverse-Field I-μSRotation 38
2.3.2 Longitudinal-Field I-μSRelaxation 39
2.3.3 Avoided Level-Crossing Resonance 40
2.3.3.1 Nuclear Quadrupolar μALCR 40
2.3.3.2 "Muonated" Radicals 41
2.3.3.3 Muonium in Semiconductors 41
2.4 Muon Spin Resonance 42
2.4.1 rf I-μSResonance 42
2.4.2 Muon Spin Echoes 42
3. **μSR Applications** 43
3.1 Muonium Chemistry 43
3.1.1 Chemical Kinetics 43
3.1.2 Radicals 43
3.1.3 Molecular Structure 44
3.2 Condensed-Matter Physics ... 44
3.2.1 Hydrogen in Semiconductors 44
3.2.2 Quantum Diffusion 44
3.2.2.1 Motion of the μ^+ in Metals .. 44
3.2.2.2 Muonium Diffusion in Nonmetals 45
3.2.3 Magnetism 46
3.2.4 Superconductivity 46
3.2.4.1 Phase Diagrams 47
3.2.4.2 Magnetic Penetration Depth 47
3.2.4.3 Universal Correlations between T_c and Ω_s 48
3.2.4.4 Non-*s*-wave Pairing? 48
4. **μSR Facilities** 48
4.1 Logistics of Accelerator Experiments 48
4.2 Standard μSR Hardware 49
4.2.1 Detectors 49
4.2.2 Electronics 49
4.2.3 μSR Spectrometers 51
4.2.4 Target Vessels and Cryostats 51
4.3 μSR Software 51
5. **Speculations** 51
Works Cited 51
Further Reading 53

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

The acronym "μSR" was coined in 1974 (*μSR Newsletter*, 1974); the definition and explanation offered on that occasion are still apt.

> μSR stands for M̲uon S̲pin R̲elaxation, R̲otation, R̲esonance, R̲esearch or what have you. The intention of the mnemonic acronym is to draw attention to the analogy with NMR and ESR, the range of whose applications is well known. Any study of the interactions of the muon spin by virtue of the asymmetric decay is considered μSR, but this definition is not intended to exclude any peripherally related phenomena, especially if relevant to the use of the muon's magnetic moment as a delicate probe of matter.

Although muons were used as probes of magnetism in matter as early as 1944 (Rasetti, 1944), the essential property of the weak interaction that makes μSR possible—violation of parity (*P*) symmetry—was posited by Lee and Yang (1956) to explain anomalies in kaon decay experiments (Fitch, 1981; Cronin, 1981). Of the famous measurements (Garwin *et al.*, 1957; Friedman and Telegdi, 1957; Wu *et al.*, 1957) confirming their hypothesis (Lee and Yang, 1957), one also suggested that *P* nonconservation in $\pi \rightarrow \mu \rightarrow e$ decay might furnish a sensitive general-purpose probe of matter. The history of μSR began with that experiment (Garwin *et al.*, 1957), which used an experimental method similar to the most common and familiar of modern μSR techniques: transverse-field (TF)-μSR.

Later sections will treat TF-μSR and other μSR techniques in some detail, but it is useful to begin with a qualitative phenomenological description just to establish some terminology. A crude apparatus for TF-μ^+SR (μSR using positive muons) is pictured schematically in Fig. 1. The μ^+ arising from π^+ decay at rest is perfectly spin polarized as it enters the sample; later, it decays asymmetrically, with the decay positron emitted preferentially along the muon spin direction.

After stopping $\sim 10^6$ muons in the target sample, one obtains a time spectrum like that shown in Fig. 2 (top), which ideally has the following form:

$$N(t) = B + N_0 \exp(-t/\tau_\mu)[1 + A(t)], \qquad (1)$$

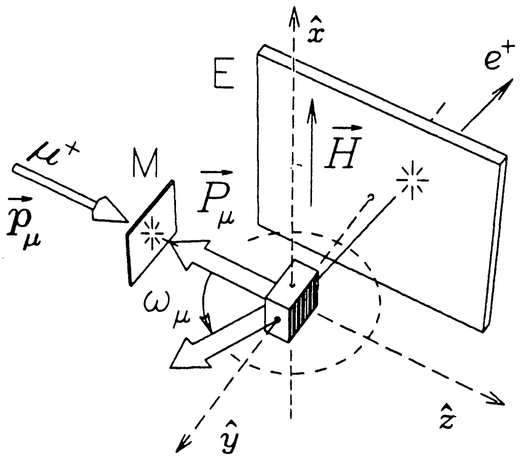

FIG. 1. A simple transverse-field (TF)-μ^+SR experiment: The μ^+ beam enters from the left with its polarization antiparallel to its momentum. A magnetic field **H** is applied vertically, causing the μ^+ spins to precess at the Larmor frequency $\omega_\mu = \gamma_\mu H$, where $\gamma_\mu / 2\pi = 0.013\,553\,42$ MHz/Oe. An incoming μ^+ triggers the *M* counter, generating a start pulse for a fast time digitizer ("clock"), and an outgoing decay positron later stops the clock with a pulse from the *E* counter; for each such event, the time interval is digitized, and the corresponding bin in a discrete time spectrum is incremented.

where N_0 is an overall normalization, B is a time-independent background, $\tau_\mu = 2.197$ μs, and $A(t)$ is the corresponding *asymmetry spectrum* shown in Fig. 2 (bottom), which can be extracted numerically from $N(t)$ as[*]

$$A(t) = \left[\frac{N(t) - B}{N_0}\right] e^{+t/\tau_\mu} - 1. \qquad (2)$$

---

[*] In order to convert $N(t)$ to $A(t)$, one must know both B and N_0; fortunately, both constants can be extracted numerically from $N(t)$ in cases where the period of μ^+ Larmor precession is a negligible fraction of the muon lifetime. Treating $N(t)$ as a continuous function (the actual discrete sums can easily be deduced from the integrals below), we may write

$$N_0 = \frac{S\tau_\mu E_+ - \mathrm{RT}}{\tau_\mu^2 E_+ E_- - T^2}, \qquad B = \frac{R\tau_\mu E_- - ST}{\tau_\mu^2 E_+ E_- - T^2},$$

where

$$S \equiv \int_{t_i}^{t_f} N(t)dt, \qquad R \equiv \int_{t_i}^{t_f} N(t)\exp(t/\tau_\mu)dt,$$
$$E_\pm \equiv \pm[\exp(\pm t_f/\tau_\mu) - \exp(\pm t_f/\tau_\mu)],$$

and $T \equiv t_f - t_i$, the time interval over which the integrals are evaluated.

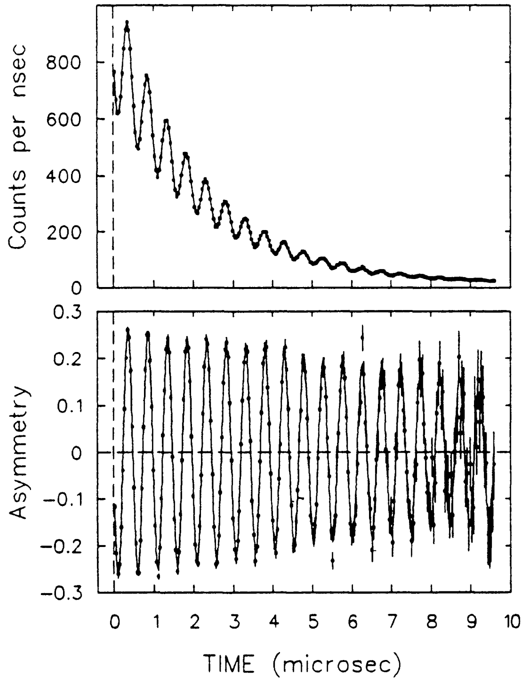

FIG. 2. Top: "Raw" time spectrum from a simple TF-μ^+SR experiment: The overall exponential decay reflects the muon lifetime, and the precession of the μ^+ spins is manifest in the superimposed oscillations as the muon polarization sweeps past the positron detector. Bottom: TF-μ^+SR asymmetry spectrum obtained from the raw time spectrum by subtracting any time-independent background and dividing out the exponential distribution of muon decay times.

Except for an empirical multiplicative constant, $A(t)$ represents the time evolution of the muon polarization, much like a free-induction decay (FID) signal in NMR.

For the next decade or so after 1957, μSR was developed primarily in the guise of a series of experiments using muons to test the predictions of quantum electrodynamics (QED) with unprecedented accuracy (Combley *et al.*, 1981, Hughes, 1988). Meanwhile, the muon's weak-interaction properties were being precisely determined in an ongoing (Stoker *et al.*, 1985) sequence of measurements of the *Michel parameters* (Michel, 1950) of normal muon decay. Many of today's areas of application of μSR began as peripheral problems in those fundamental physics experiments.

With the advent of the "meson factories" in the 1970s came a hundredfold to thousandfold increase in the intensity of muon beams, dramatically accelerating the development of new experimental techniques and the discovery of new applications for μSR. When the potential of μSR became apparent, most meson factories invested in the upgrading of muon beamlines and μSR facilities that came on line in the early 1980s. Since then, the techniques of μSR have been discovered by the chemistry and solid-state physics communities, and what was once an esoteric oddity has become one of the fastest-growing areas of intermediate-energy science.

Today's μSR is a standard magnetic-resonance tool, almost exclusively devoted to disciplines rarely associated with subatomic physics—primarily chemistry and condensed-matter physics. A list of some of these areas of application will appear later in this article, but first it is necessary to describe the fundamental physics that makes μSR possible and to explain the basic techniques of its use.

1. BASIC MUON PHYSICS

1.1 Muon Production

Although muons are produced in a variety of high-energy processes and elementary-particle decays such as $K \rightarrow \mu + \nu$ decay (Yamanaka *et al.*, 1984), μSR requires low-energy muons in order to stop the beam in samples of convenient thickness (≤ 1 cm), and these are available in the required intensities only from the ordinary two-body decay of charged pions:

$$\pi^+ \rightarrow \mu^+ + \nu_\mu \quad \text{or} \; \pi^- \rightarrow \mu^- + \bar{\nu}_\mu, \tag{3}$$

from which the muon emerges (in the rest frame of the pion) with a momentum of 29.79 MeV/c and a kinetic energy of 4.119 MeV. The lifetime of a free charged pion is $\tau_\pi = 26.03$ ns. Because the neutrino is only produced with negative helicity (spin antiparallel to momentum) and the antineutrino only with positive helicity—a succinct description of P nonconservation adequate for our purposes here—the simultaneous conservation of linear and angular momentum forces the μ^+ also to have negative helicity (and the μ^-, positive helicity) in the rest frame of the pion. Thus, muons emitted from pion decay at rest are also 100% spin polarized opposite to (for μ^+) or along (for μ^-) the direction of their momenta. This is the greatest advantage of μSR as a magnetic-resonance technique: whereas NMR and ESR rely upon a thermal-equilibrium spin polarization, usually achieved at low temperatures in strong magnetic fields, μSR begins with a *perfectly* polarized probe, regardless of conditions in the medium to be studied. It also implies that muon spin degrees of freedom usually start their evolution as far from thermal equilibrium as conceivable.

Most μ^+ beams today are literally emitted from π^+ decay at rest in the surface layer of the primary target, where the pions themselves are produced by collisions of high-energy protons with target nuclei—hence, the common mnemonic name, *surface muons* (Bowen, 1985). Unfortunately, this mode is not available for negative muons because a π^- stopping in the production target almost always undergoes nuclear capture from low-lying orbitals of pionic atoms before it has a chance to decay. This problem was actually solved long before the surface-muon beam was invented, when many fundamental physics experiments with muons were primarily concerned with the "heavy electron" behavior of the μ^-. In so-called "conventional" muon channels, pions are allowed to decay *in flight* down a relatively long straight section where the decay muons are collected by axial or alternating-gradient magnetic fields; the muons emitted "backward" in the pion rest frame have quite different momenta from those emitted forward (or from the pions themselves), and can thus be selectively extracted by a bending magnet. The disadvantages of such *backward muon beams* are their relatively higher momentum (usually ~ 50–100 MeV/c), their larger momentum spread (and, therefore, lower stopping density), and their much larger phase space (and, therefore, lower luminosity). Today, they are rarely used for μ^+SR, but for μ^-SR there is no alternative.

1.2 Muon Decay

Books have been written on the subject of this section (Primakoff, 1975); there follows only a synopsis of those aspects of normal muon decay ($\mu^+ \rightarrow e^\pm + \nu_{e,\mu} + \bar{\nu}_{\mu,e}$) that are essential to a qualitative understanding of μSR. The foremost of these is, of course, the propensity of the muon-decay positron (electron) to be emitted along (opposite to) the spin of the μ^+ (μ^-). This second example of P nonconservation in the weak interaction is what allows us to read out the information encoded in the evolution of an initially polarized muon spin ensemble. The information is delivered to the experimenter in the form of rather high-energy (up to 52 MeV) positrons or electrons, which readily penetrate sample holders, cryostats, or ovens and the detectors used to establish the time and direction of the muon decay.

The decay probability of the muon, illustrated in Fig. 3, depends upon the $e^\pm$ energy $x \equiv \epsilon_e/\epsilon_{\max}$ (where $\epsilon_{\max} = 52.83$ MeV is the maximum possible total relativistic energy of the $e^\pm$) and the angle θ between the muon spin direction and the direction and the direction of $e^\pm$ emission as

$$dP(x,\theta) = E(x)[1 + a(x)\cos\theta]dxd(\cos\theta). \tag{4}$$

The *asymmetry* factor a depends upon the $e^\pm$ energy as

$$a(x) = \pm(2x - 1)/(3 - 2x), \tag{5}$$

and the normalized $e^\pm$ energy spectrum has the form

$$E(x) = 2x^2(3 - 2x). \tag{6}$$

Figure 4 shows the functions $E(x)/2$, $a(x)$, and their product. Note that a changes sign at low energy; however, very few positrons are

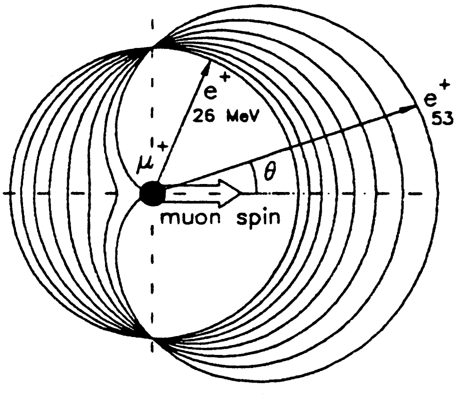

FIG. 3. Angular distribution of the e^+ from $\mu^+ \rightarrow e^+ + \nu_e + \bar{\nu}_\mu$ decay: The asymmetry (anisotropy) of the distribution is 100% for the highest $e^\pm$ energy $\epsilon_{\max} = 52.83$ MeV and zero (i.e., an isotropic distribution) for $\epsilon_e = \epsilon_{\max}/2$; for $\epsilon_e \rightarrow 0$ (not shown), the asymmetry is negative.

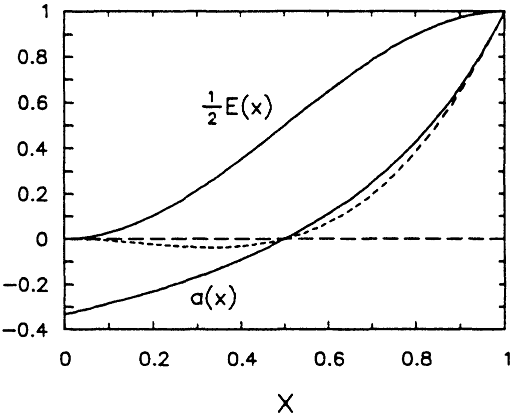

FIG. 4. Solid lines: energy spectrum $E(x)/2$ ($e^{\pm}$ energy $\epsilon_e = 52.83x$ MeV) of $e^{\pm}$ from normal $\mu^{\pm}$ decay and ϵ_e dependence of μ^+ decay asymmetry $a(x)$ (degree of correlation between e^+ momentum and μ^+ spin direction). (For $\mu^- \to e^-\bar{\nu}_e\nu_\mu$, a has the opposite sign) Dashed line: weighted $\mu^+ \to e^+\nu_e\bar{\nu}_\mu$ asymmetry spectrum, the product of $E(x)/2$ and $a(x)$.

emitted with such low energies (see above), and those which are will usually not be detected (see below).

1.2.1 Asymmetry Calibration The theoretical average asymmetry,

$$\langle a \rangle = \int_0^1 a(x)E(x)dx = \pm\tfrac{1}{3}, \qquad (7)$$

is never realized in practice (except by accident), for several reasons.

First, radiative corrections subtly distort the low-energy end of the spectrum (Sachs and Sirlin, 1975).

Second, the low-energy $e^{\pm}$ are easily stopped in even thin layers of material, such as the wrappings of scintillation counters: indeed, since this effect raises the average asymmetry, it is common practice to insert degraders between the muon stopping region and the $e^{\pm}$ detectors, when feasible, up to an optimum range (e.g., ≈ 3–4 cm of graphite) that maximizes the product A^2N, where A is the empirical asymmetry (defined below) and N is the $e^{\pm}$ rate reaching the detector. (One should generally use low-Z materials for degraders in order to minimize bremsstrahlung and pair production, which can greatly confuse the issue, unless one wishes to *use* such effects to bypass strong magnetic fields—see below.)

Third, a typical $e^{\pm}$ detector intercepts a rather large solid angle and, thus, averages $\cos\theta$ appreciably. The optimal detector geometry for general-purpose time-differential (TD)-μSR is a cube centered on the target where the muons stop, with each of the six faces a single detector. (Although, in principle, one could gain back some asymmetry by using position-sensitive detectors and weighting individual events according to the projections of the $e^{\pm}$ track along different axes, the complications of counting statistics—not to mention the expense of processing such information at very high rates!—are not usually justified by the marginal improvement.)

Finally, because the decay $e^{\pm}$ will follow a helical path in an applied magnetic field **H** (which is an essential part of many μSR experiments), the curling up of $e^{\pm}$ orbits causes the efficiency of their detection to be a function of both their energy and the applied field strength; the apparent direction of emission may also be affected. At p_e = 30 MeV/c (a typical $e^{\pm}$ momentum) the radius of curvature ρ_e of an orbit perpendicular to an applied magnetic field H = 1 T is exactly 10 cm, as given by the familiar formula

$$\rho\,(\mathrm{cm}) = \frac{p\,(\mathrm{MeV/c})}{0.3\,H\,(\mathrm{kOe})}. \qquad (8)$$

This effect becomes problematic at high fields ($H \gtrsim 5$ T $\Rightarrow \rho \lesssim 2$ cm), in which case μSR requires very small samples and detectors. Another possible approach to this problem is to generate bremsstrahlung photons intentionally from the decay e^+, using a lead converter near the target, and then reconvert the photons in front of the detector after they have traversed the high-field region.

Taken together, these systematic effects make the empirical asymmetry A in a given detector perplexingly dependent upon the thickness, geometry, and material of target and detectors, not to mention magnetic field. Attempts to correct analytically for $e^{\pm}$ energy, etc., are usually only reliable to within a few percent.

Fortunately, there are many materials (including most metals) in which the μ^+ suffers negligible depolarization; thus, the usual method for *calibrating the asymmetry* in μ^+SR is to measure first a dummy sample of dimensions identical to those of the real sample, but made from aluminum or silver. If great care is taken to arrange the two samples in exactly the same position, this will determine the *empirical maximum asymmetry* A_0 to within a few percent of itself; typical values are 0.2 to 0.3. However, materials of the same nominal thickness (in g/cm^2) will often exhibit different dE/dx and multiple scattering of the $e^{\pm}$, so that this calibration method cannot generally be trusted to better than ~1%.

In μ^-SR, as we shall see, such calibrations are much more difficult because of the loss of μ^- polarization in the initial cascade to the ground state of the muonic atom. To the author's knowledge, there is no satisfactory solution for the A_0 calibration problem in μ^-SR. The usual method is to calibrate on graphite, which gives a residual μ^- polarization of about (17 ± 2)% of the initial muon beam polarization (Kuno *et al.*, 1984).

1.3 The "Heavy Electron" versus the "Light Proton"

Asymmetry calibration is perhaps the least of the qualitative differences between μ^+SR and μ^-SR, all of which are consequences of the opposite charges of the μ^+ and μ^-. In elementary-particle physics, the muon is often described as a "heavy electron" in reference to its family resemblance to other leptons; it is the μ^- that is so designated. Indeed, the μ^- does just what one might expect of a heavy electron: It undergoes Coulomb capture into atomic orbitals analogous to those of electrons but 206.768 times closer to the nucleus and unrestricted by the Pauli exclusion principle; thus, the muon quickly (usually in ~10^{-14} s) cascades to the 1*s* ground state, emitting photons (known as muonic x rays) and/or Auger electrons (ejected from low-lying orbitals in the same atom by muonic x rays) on the way. Once there, the significant overlap between muon and nuclear wave functions (in heavy nuclei, the muon spends most of its time inside the mean nuclear radius) gives the weak interaction between the muon and the proton a chance to act, resulting in nuclear capture ($\mu^- + p \rightarrow n + \nu_\mu$), which appreciably reduces the μ^- lifetime (and the fraction of muons decaying to electrons for μ^-SR to detect).

By contrast, the positive muon avoids positively charged nuclei and will capture its own electron when it can, to form the hydrogen-like atom muonium (μ^+e^-, usually abbreviated Mu). In fact, the qualitative behavior of the μ^+ in matter resembles that of a light proton far more than that of its closer antilepton relative, the positron. The μ^+ mass is 0.1126 times that of the proton but 206.768 times that of the positron: its magnetic moment is 3.1832 times that of the proton but only 0.48363% of the positron's. The Mu atom is almost identical (except in mass) to the hydrogen atom, whereas the positronium atom (e^+e^-, or Ps) has no nucleus and half the binding energy of H. Positronium also annihilates very quickly, whereas weak electron capture in muonium ($\mu^+ + e^- \rightarrow \nu_\mu + \nu_e$) only shortens the μ^+ lifetime by ~1 part in 10^{10}.

As a result, μ^-SR differs qualitatively from μ^+SR.

1. *Lifetime(s):* The μ^+ lifetime is independent of its environment, while that of the μ^- depends strongly upon the Z of the nucleus to which it becomes attached: a sample containing several elements will have as many μ^- lifetimes, the ratios of whose probabilities cannot be reliably estimated from the Fermi-Teller "Z law" (Fermi and Teller, 1947) and, therefore, must be determined by fitting the time distribution of decay or capture products.
2. *Muon capture:* A μ^- that undergoes nuclear capture ($\mu^- + p \rightarrow n + \nu_\mu$) does not produce a decay electron for detection in μ^-SR; the μ^-SR event rate per incident muon is thus proportional to the μ^- lifetime, which in high-Z elements can be as little as 4% of the free muon lifetime. This rate loss can often be offset by increasing the μ^- beam intensity, except when the sample contains both heavy and light nuclei.
3. *Heterogeneous signal:* Whereas the μ^+ produces only positrons and undetected neutrinos, a μ^- entering some high-Z samples like ^{238}U is likely to cause emission of x rays and Auger electrons during the atomic cascade, fission fragments and neutrons (on average, several per muon) from either nuclear muon capture or nuclear internal conversion of muonic x rays, and the occasional decay electron. The μ^-SR signal is, therefore, noisy and heterogeneous. This does make it rather interesting from the points of view of atomic and nuclear physics, but such complica-

tions are so far unwelcome in historically typical μ^-SR applications.

4. *Lost polarization:* Spin-orbit coupling in the atomic cascade leaves the μ^- in its ground-state 1*s* orbital with a dramatically reduced spin polarization, typically $\lesssim$20%. Because statistical uncertainties shrink as $N^{-1/2}$, one must accumulate 25 times more events to achieve the same statistical accuracy in measuring a five times smaller signal. This seemingly irreducible disadvantage is the main limitation of μ^-SR applications and is the reason why μSR is often treated as synonymous with μ^+SR. However, since μ^-SR offers complementary and unique information, it will always be an essential part of the μSR repertoire. There is also a chance that new techniques, such as x-ray tagged μ^-SR, may help overcome this disadvantage.
5. *Hyperfine effects:* The hyperfine coupling between the muon and nuclear spins in nonzero-spin nuclei is enormous, often large enough to cause Auger emission of *K* and *L* electrons (Winston, 1963). The two hyperfine states F^+ and F^- of the 1*s* muonic atom (in which the muon spin is respectively parallel and antiparallel to the nuclear spin) therefore produce distinct μ^-SR signals, each with its own characteristic precession frequency, in such systems. This was long thought to limit useful μ^-SR to spinless elements, but quite a few nuclei with spin $I > 1/2$ have been found to exhibit potentially useful μ^- SR signals, albeit at reduced amplitudes (Brewer, 1984a, 1984b).

Taking into account the different lifetimes, cascade depolarization, and spin states of negative muons in different muonic atoms, we may generalize Eq. (1) to obtain the most general form for *electron-triggered* μ^-SR in a sample composed of different elements:

$$N(t) = B + \sum_i N_i^0 \exp(-t/\tau_i)[1 + A_i(t)], \qquad (9)$$

in which $N_i^0 = N_0 f_i(\tau_i/\tau_\mu)$, where N_0 is an overall normalization, f_i is the fraction of muons captured on the *i*th type of nucleus, and τ_i is the μ^- lifetime in that species of muonic atom and hyperfine state; $A_i(t) = A_0 P_i(t)$, where A_0 is a common maximal asymmetry factor and $P_i(t)$ is the polarization of muons in the *i*th type of muonic atoms. For TF-μ^-SR, this polarization will have the form $P_i(t) = P_i^0 G_i(t)\cos(w_i t + \phi_i)$, where P_i^0 is the polarization remaining in that atom's given hyperfine state after the cascade, $G_i(t)$ is the corresponding relaxation function (see below), ω_i is its characteristic precession frequency in the applied magnetic field, and ϕ_i is the initial phase.

In the interest of simplicity, further discussion of μ^-SR will be neglected in favor of μ^+SR, which occupies most of the attention of the μSR community.

2. μSR TECHNIQUES

Ideally (and often in reality), the μSR experimenter has control over the orientation of the detectors, the applied magnetic field, and (within some range) the muons' spin polarization. The beam momentum can be deflected as well, but this is rarely desirable. In order to designate different orientation choices consistently, the labeling conventions defined in Fig. 5 have been devised specifically for μ^+SR experiments using surface muons, which are originally polarized opposite to their momentum but whose spins can be rotated 90° by a Wien filter in the beamline (Brewer, 1981). (Such flexibility is not available for conventional or backward $\mu^\pm$ beams, which will be neglected here partly for that reason.) The standard detector array consists of six counters aligned with the positive and negative coordinate axes and labeled *F* (forward), *B* (backward), *U* (up), *D* (down), *L* (left), and *R* (right), according to a "beam's-eye view" naming convention. Note that the unrotated muon polarization points toward the *B* counter and, in the spin-rotated mode, toward the *U* counter; the latter depends, of course, on the orientation of the fields in the Wien filter.

2.1 Time-Differential μSR in Transverse Field

The simplest and most familiar time-differential (TD)-μSR technique is the transverse-field (TF) muon spin *rotation* experiment, in which an external magnetic field is applied perpendicular (transverse) to the muon polarization, causing the muon spins to precess (rotate) about the field.

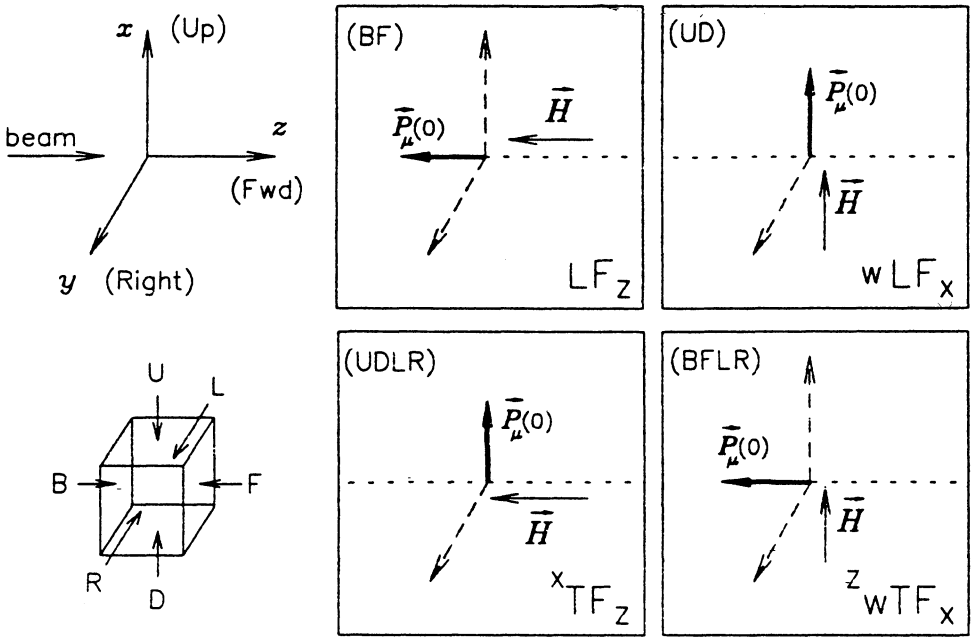

FIG. 5. Coordinate system and labeling conventions for surface-muon μSR experiments. Note that the superscript on the left indicates the direction of the incoming muon polarization, while the subscript on the right indicates the direction of the applied field (if any); for longitudinal field (LF) and by continuation for zero field (ZF), both will always be the same, but for TF, there are, in principle, two possible arrangements for each choice of super(sub)script (for instance, $^zwTF_y$ vs $^xwTF_y$). "Weak" (w) field means "not strong enough to deflect the muon beam appreciably."

2.1.1 The Basic Technique In transverse-field (TF)-μSR, a magnetic field **H** is applied perpendicular to the initial muon spin direction, causing Larmor precession of the muon polarization about **H**. This arrangement varies from the primitive version depicted in Fig. 1 to the 4-counter, spin-rotated ^{x}TF$_z$-μSR apparatus shown in Fig. 6 and the ^{z}TF$_x$-μSR arrangement in Fig. 7, each of which illustrates *quadrature* detection (two orthogonal pairs of opposing counters in the plane perpendicular to the applied field); but the time spectra from individual counters all have the qualitative appearance shown in Fig. 2.

2.1.1.1 Frequency. The most obvious parameter observable in a TF-μSR spectrum is the *muon precession frequency* $\omega_\mu = \gamma_\mu B$, which (thanks to our precise knowledge of γ_μ) is equivalent to the local magnetic field B at the muon. Since B may be affected by diamagnetism, paramagnetism, and contact hyperfine interactions with polarized electrons (Knight shifts) in the medium, B is generally different from the applied field H_0; this difference is often the main focus of the TF-μSR experiment. An example is shown in Fig. 8.

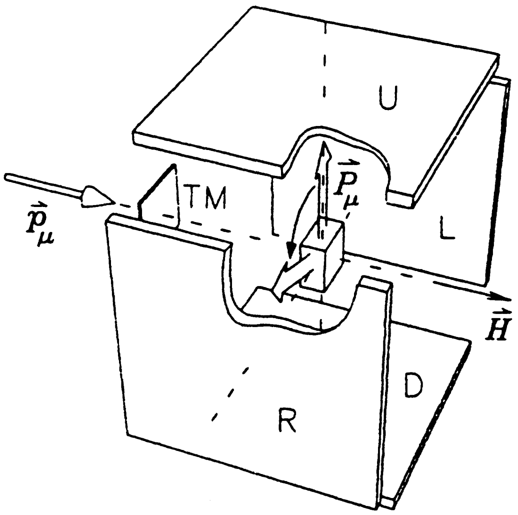

FIG. 6. Counter arrangements for ^{x}TF$_z$-μSR, in which the muon beam has been passed through a Wien filter to rotate the muon spins until they are perpendicular to their momenta. The momentum $\mathbf{p}_\mu$ is undeflected in the magnetic field $\mathbf{H} = H_z \| \mathbf{p}_\mu$, but the polarization $\mathbf{P}_\mu$ precesses in the x-y plane.

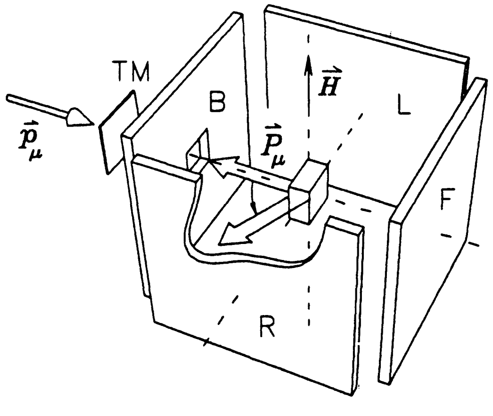

FIG. 7. Counter arrangements for ${}^{z}w\mathrm{TF}_x$-μ^+SR, in which the μ^+ beam still has its polarization antiparallel to its momentum. The field must therefore be perpendicular to both $\mathbf{p}_\mu$ and $\mathbf{P}_\mu$, which restricts use with surface muon beams to weak fields $H \lesssim 100$ Oe.

2.1.1.2 Asymmetry. The second obvious parameter characterizing muon precession in a transverse field is the amplitude, or *asymmetry*, of the precession signal. As mentioned earlier, the absolute calibration of this amplitude is tricky, but one can usually convert the initial amplitude into a *residual polarization* with an accuracy of a few percent. In metallic samples, the initial polarization is generally consistent with 100%, but in insulators or semiconductors, liquids, or gases, some fraction of the muons (ranging from 0 to 100%) either form muonium (Mu) or experience some other form of depolarization in the early ($\lesssim$1 ns) stages of thermalization in the sample.

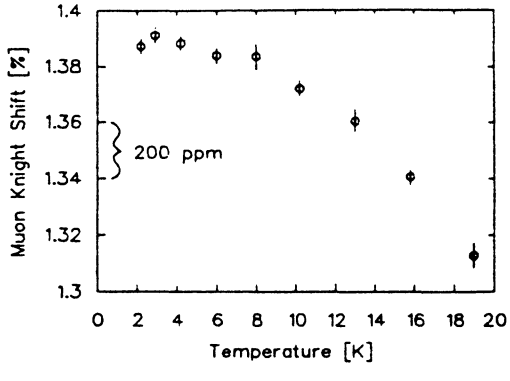

FIG. 8. Positive muon Knight shift $K_\mu \equiv (B_\mu - H_0)/H_0$ (where H_0 is the applied magnetic field and B_μ is the field at the muon) as a function of temperature in a single crystal of antimony with the c axis parallel to the applied field (1.5 T). The precision (approximately ±50 ppm) is typical of what can easily be achieved with routine methods in modest TF-μSR experiments; much higher precision is possible with refined techniques (see the later section on strobo-μSR).

The effect of Mu formation in wTF is to cause the muon polarization to precess in the opposite sense to that of muons in diamagnetic environments, roughly 103 times faster; this results in dramatic "dephasing" if the Mu atoms subsequently react at exponentially distributed times to enter diamagnetic states. If all this takes place within a few ns, all one can observe is the net effect on the subsequent diamagnetic μ^+ precession signal, namely, a reduction of asymmetry and a simultaneous shift of the apparent initial phase of precession. If the magnetic field is in the z direction and the initial μ^+ polarization is in the x direction, we may define the *complex* muon polarization $\tilde{P}(t) \equiv P_x(t) + iP_y(t)$, in terms of which the overall polarization of an ensemble of muons starting as Mu atoms and reacting at a rate Λ to form some diamagnetic species is given in the low-field limit (averaging over high-frequency hyperfine oscillations) by

$$\tilde{P}(t) \approx \frac{1}{2}\frac{\omega_{\mathrm{Mu}}+\omega_\mu}{\omega_{\mathrm{Mu}}+\omega_\mu-i\Lambda}\exp[-(\Lambda+i\omega_{\mathrm{Mu}}t)] + \frac{1}{2}\left[\frac{i\Lambda}{i\Lambda-(\omega_{\mathrm{Mu}}+\omega_\mu)}+\frac{\Lambda^2}{\Lambda^2+\omega_0^2}\right]e^{i\omega_\mu t}, \qquad (10)$$

where the diamagnetic precession frequency is ω_μ and the muonium precession frequency is $\omega_{\mathrm{Mu}} \approx 103\omega_\mu$ in the opposite sense. This sort of simple *residual polarization* picture also describes many other *delayed formation* scenarios.

2.1.1.3 Initial Phase of Precession. A more subtle aspect of the residual polarization picture is the shift of the apparent initial phase of the μ^+ precession due to (e.g.) the formation of short-lived Mu atoms precessing in the opposite sense. Measurement of such phase shifts has often proved valuable in sorting out "fast chemistry" effects (Brewer *et al.*, 1974).

2.1.1.4 "Relaxation." Following thermalization of translational degrees of freedom, which may take as little as 100 ps in solids, the muon precession signal may still decrease in amplitude because of dephasing

(T_2 effects) or true irreversible relaxation processes (T_1 effects). The latter are more often studied in longitudinal field (LF), where the distinction between T_1 and T_2 is not subject to so much semantic debate. (See Sec. 2.2.3.) Depolarization due to inhomogeneous magnetic fields often takes the form of a Gaussian relaxation function $\exp(-\frac{1}{2}\sigma^2 t^2)$, as shown in Fig. 9.

2.1.1.5 Muonium Precession. All the features just described for muon precession at the diamagnetic Larmor frequency $\omega_\mu = \gamma_\mu H$ are also often seen for *muonium precession* in ωTF-μSR experiments. The main differences are that $\omega_{\mathrm{Mu}} = \gamma_{\mathrm{Mu}} H$ is roughly 103 times larger than ω_μ in the same field H (because of the huge magnetic moment of the electron locked to the muon spin by the hyperfine interaction), that the sense of ωTF Mu precession is opposite to that of the free μ^+ (because of the opposite sign of the dominating e^- moment), that half the muon polarization appears lost in most experiments [because of the fast (4463 MHz for Mu in vacuum) hyperfine oscillations between the $|\Uparrow\downarrow\rangle$ and $|\Downarrow\uparrow\rangle$ states, where $\Updownarrow$ refers to the electron spin and $\updownarrow$ refers to that of the muon], and that ω_{Mu} splits into two frequencies for $H \gtrsim 20$ Oe because the hyperfine interaction is finite. This phenomenon will be discussed in more detail below.

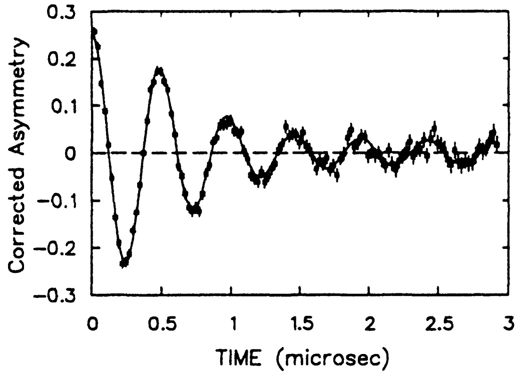

FIG. 9. Depolarization of muons in a sintered powder of superconducting $La_{1.85}Sr_{0.15}CuO_4$ at 10 K in an applied magnetic field of 150 Oe. The relaxation is due to local field inhomogeneities caused by flux exclusion and a vortex lattice in this type-II superconductor; it is adequately described by a Gaussian relaxation function in which $\sigma \propto \lambda^{-2}$, where λ is the London penetration depth. This has been much used for initial estimates of λ.

2.1.2 High Field Since the period of muon precession is 7.38 ns in a field of 1 T and is inversely proportional to H, it is not difficult to achieve a magnetic field strong enough to challenge the time resolution of most μSR spectrometers. Moreover, the radius of decay positron orbits shrinks with increasing field until experiments in $H \gtrsim 5$ T require small detectors within 1–2 cm of the sample. Nevertheless, the effort of miniaturization and state-of-the-art time resolution is often justified by the improved resolution for Knight shifts and/or access to intrinsically high-field phenomena. In high fields, a standard TD-μSR time spectrum consists of a large number of small time bins, each of which may have a rather low number of counts and a correspondingly large statistical uncertainty; this circumstance makes the usual laboratory-frame asymmetry plot rather uninformative to the eye and expensive to fit by χ^2 minimization, motivating new methods of data analysis.

2.1.2.1 Fourier Transforms. Often, especially when high frequencies are involved, one would rather see a frequency spectrum than a time spectrum. The art of converting the latter into the former is subject to continuing evolution and perpetual debate, but virtually all μSR facilities include the fast Fourier transform (FFT) in their standard data-analysis tools. A complete account of the merits and hazards of such treatments is beyond the scope of this article, but it is useful to point out several ubiquitous features that are ignored at one's peril.

First, as shown in Fig. 10, the fact that the time range of a μSR spectrum necessarily starts at $t = 0$ means that the imaginary (odd in t) part of the resultant frequency spectrum is dramatically broadened by the tacit presence of the Heaviside function; this has nothing to do with either the muon lifetime or the finite range of positive time. Thus, the real part of the amplitude (the proper objective of our FFT) may be quite narrow, but the square root of the power spectrum (the sum of the squares of the real and imaginary parts) will still be very broad. Various tricks are available for unmixing the real and imaginary parts.

Second, because of the finite muon lifetime, there are fewer statistics in the bins at late times, and consequently the "error bars" in the asymmetry spectrum grow exponen-

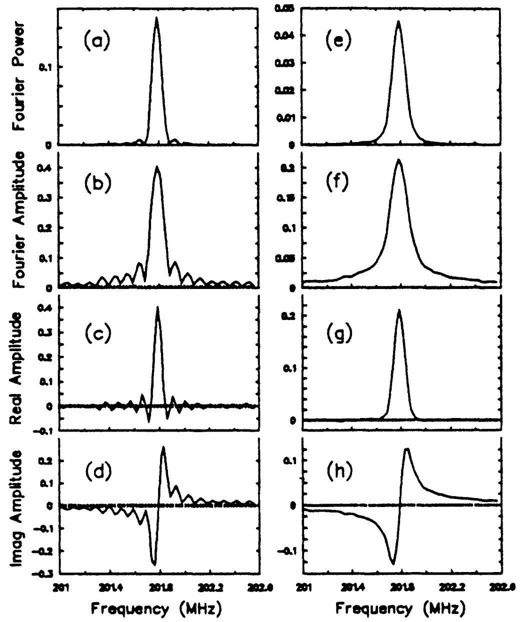

FIG. 10. Results of complex FFT of a quadrature μ^+SR time spectrum with a slow relaxation rate: (a)–(d) the power, the amplitude (square root of the power), the real and the imaginary parts of the frequency spectrum resulting from FFT of the raw complex time spectrum over 10.24 μs at a field of 1.489 T; (e)–(h) the same quantities for the same data *apodized* by a Gaussian envelope function with a 4-μs time constant.

tially with time as $\exp(+t/2\tau_\mu)$. (See Fig. 2.) This introduces noise in an annoyingly nonuniform way, which may, in principle, be corrected for using not-so-fast Fourier transform methods. These *noise effects* are always subject to suppression *without limit* (except for the finite patience of the experimenter) by simply taking more statistics: although the finite muon lifetime is responsible, it does not fix any absolute limits on precision—only practical ones.

Third, because experimenters' patience is indeed finite, most μSR time spectra only extend to 10–20 μs; thus, the FFT is cut off abruptly at some time limit, introducing the "ringing" effect evident on the left side of Fig. 10. This can be suppressed by *apodization*—multiplying the "raw" time spectrum by an *envelope function* that causes the product to go to zero gently before the abrupt end of the time range. As evident from Fig. 10, well-chosen apodization removes the ringing caused by FFT on a finite time interval but introduces a line-broadening effect that must be taken into account in the interpretation.

In general, there are no time-space ↔ frequency-space interconversions or interpretations that make everyone happy; hence, the earlier reference to such devices as "art." However, like other art forms, the FFT can produce some very pleasing results: for example, Fig. 11 shows the frequency (and, thus, the internal magnetic field) spectrum from μ^+SR in superconducting $YBa_2Cu_3O_{6.95}$.

2.1.2.2 Rotating Reference Frames. For TF-μSR in high magnetic field (HTF-μSR), another essential tool is the *rotating reference frame* (RRF) transformation, an example of which is shown in Fig. 12. The following description will omit the details of transforming discrete spectra and give only a simplified treatment in terms of an idealized continuous time spectrum. In fact, the RRF transformation works best on *orthogonal pairs* of spectra—as, for instance, when the magnetic field acts in the z direction and detectors are situated in the $\pm x$ and $\pm y$ directions defining the plane of precession of the muon polarization—where a *complex* asymmetry spectrum $\tilde{A}(t)$ can be defined as (e.g.) $\tilde{A}(t) \equiv A_x(t) + iA_y(t)$. In this case, the RRF transformation is simply

$$\tilde{A}_{\mathrm{RRF}}(t) = \tilde{A}(t)e^{-i\Omega t}, \qquad (11)$$

where Ω is the RRF frequency—chosen ar-

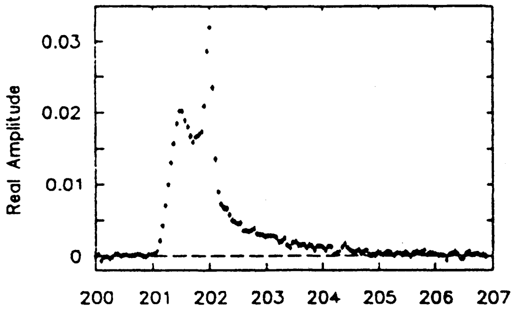

FIG. 11. Real part of frequency spectrum taken at a temperature of 6 K for positive muons in a single (3 × 3 × 0.2 mm^3) crystal of superconducting $YBa_2Cu_3O_{6.95}$ (a type-II superconductor with T_c = 92 K) cooled in a field of 1.489 T applied along the crystalline c axis. The sharp peak at 202 MHz is due to muons missing the superconductor and stopping in a normal region where the field distribution is not broadened by the vortex lattice.

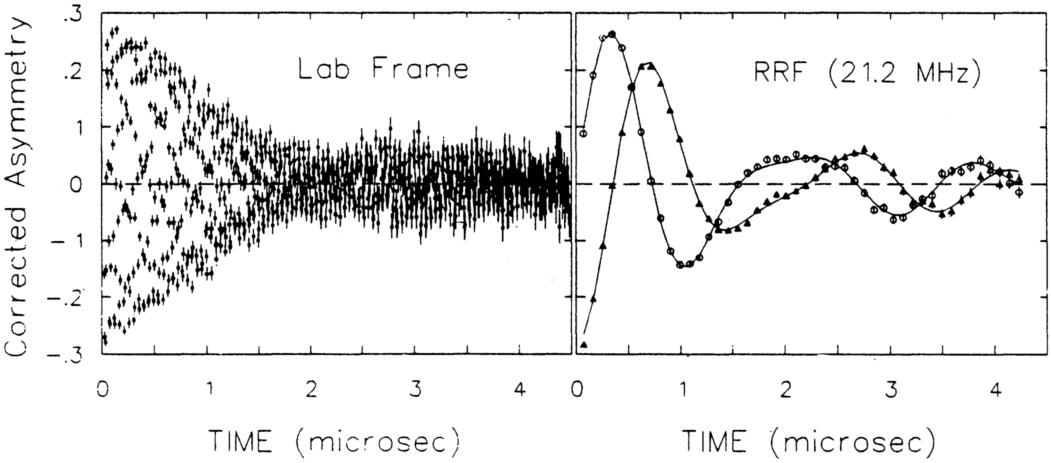

FIG. 12. The first 4.5 μs of a μ^+SR time spectrum taken at roughly 0.2 T in the laboratory frame where the data are recorded and in the rotating reference frame (RRF) to which it has been transformed numerically for display and fitting. The RRF frequency of 21.2 MHz is chosen to produce a slow precession signal for visual clarity. Because the original spectrum is real (no orthogonal detector arrays), the maximum amplitude in each of the complex RRF spectra (real part: circles; imaginary part: triangles) is actually a factor of 2 smaller than in the lab frame. The solid line in the RRF spectra is a fit by three frequencies; a single frequency (or even two) would result in a much poorer fit to the data.

bitrarily to produce a transformed $\tilde{A}_{\mathrm{RRF}}(t)$ with the desired characteristics. For instance, if the muons precess at a high frequency ω_μ, one may select Ω slightly smaller than ω_μ to produce an $\tilde{A}_{\mathrm{RRF}}(t)$ that varies relatively slowly with time; the resulting complex spectrum can be *packed* (n_p original time bins → one coarser time bin), in which case the statistical uncertainties of individual bins are reduced by a factor roughly equal to $\sqrt{n_p}$.

There are several advantages to such a transformation: first, and most obvious, is the reduction (by a factor of n_p) of the number of bins to be fitted in analysis programs, a straightforward improvement of efficiency; second, and probably more important, is the transformation of a large number of narrow time bins (with statistical uncertainties often larger than the signals under scrutiny) into a small number of well-defined bins representing the signal in a form that allows convenient display and visual inspection.

In the event that a single spectrum contains several signals at drastically different frequencies ω_i, separate RRF transformations at frequencies $\Omega_i \approx \omega_i$ can be used to isolate each signal for easy fitting.

It is also possible to perform a RRF transformation on a pure real spectrum (i.e., one without orthogonal detector axes) if care is taken to select Ω and n_p so as to "bin out" the spurious signal at the RRF frequency (Riseman and Brewer, 1990).

2.1.3 Paramagnetic States The use of Fourier transforms to reveal *line shapes* is actually a recent application in μSR; the oldest, and still most widespread, use of FFT in μSR is for extracting the frequency spectrum of muons in *paramagnetic* states such as muonium or radicals (molecules containing one or more unpaired electrons), in which the muon spin is strongly coupled to electron spins (or, in principle, orbital moments) by *hyperfine* interactions. The behavior of the simplest case, two spin-$\frac{1}{2}$ particles coupled by a scalar contact interaction in an applied magnetic field, is pictured in Fig. 13. Such couplings cause the μ^+e^- spin system (for example) to respond as a whole to applied magnetic fields, often producing rich and informative structure in the frequency spectrum (Patterson, 1988). A classic example is shown in Fig. 14 and a more recent case in Fig. 15.

2.2 Time-Differential μSR in Longitudinal and Zero Field

Referring again to Fig. 5, consider now the time evolution of the muon polarization in a magnetic field parallel to its initial direction

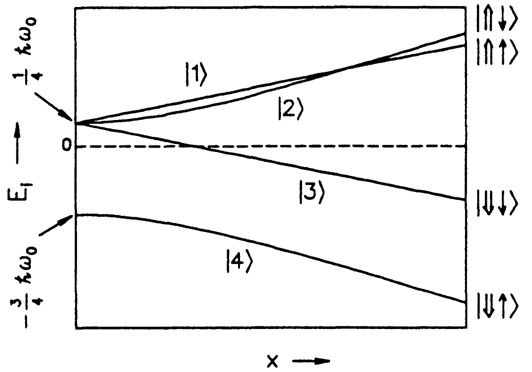

FIG. 13. Breit-Rabi diagram showing the energy levels of a system of two spin-$\frac{1}{2}$ particles of opposite sign and different magnetic moments—e.g. muonium—as functions of the reduced field $x \equiv H/H_0$, where H_0 (1585 Oe for Mu in vacuum) is a characteristic hyperfine field. For the purpose of illustration, unphysical values of moments and coupling constants have been used. The hyperfine frequency $\nu_0 \equiv \omega_0/2\pi$ has the value 4463 MHz for muonium in vacuum. In zero field, the three triplet ($J = 1$) eigenstates $|1\rangle$, $|2\rangle$, and $|3\rangle$ are degenerate, and the singlet ($J = 0$) ground state $|4\rangle$ is $\hbar\omega_0$ lower in energy. At high reduced field ($x \rightarrow \infty$), the eigenstates are $|1\rangle \rightarrow |\Uparrow\uparrow\rangle$, $|2\rangle \rightarrow |\Uparrow\downarrow\rangle$, $|3\rangle \rightarrow |\Downarrow\downarrow\rangle$, and $|4\rangle \rightarrow |\Downarrow\uparrow\rangle$ ($\Updownarrow$ refers to the electron spin, and $\updownarrow$ refers to that of the muon.) Note that state $|1\rangle$ is lower in energy than state $|2\rangle$ above $H_c = \omega_0(\gamma_e - \gamma_\mu)/2\gamma_e\gamma_\mu$ (16.386 T for Mu in vacuum).

(longitudinal field or LF). If the muon polarization initially has no components perpendicular to the local field, then none will develop, and only the polarization along the initial direction needs to be measured. Figure 16 illustrates the most common configuration, and Fig. 17 shows an alternative scheme sometimes used with spin-rotated beams.

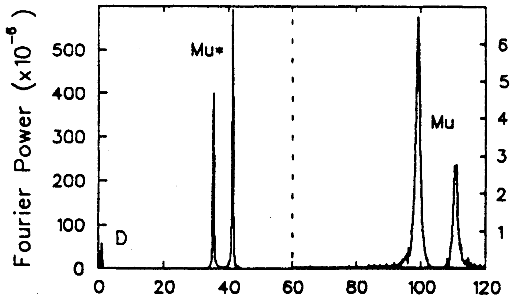

FIG. 14. Frequency power spectrum of positive muons in a pure silicon crystal at 10 K and 75 Oe transverse field showing the diamagnetic (D) signal, the rapidly diffusing muonium (Mu) signals, and the signals from so-called "anomalous" muonium (Mu*), which has been shown to be a muonium atom localized on a Si-Si bond center. Note vertical scale change at 60 MHz. After Brewer *et al.* (1973).

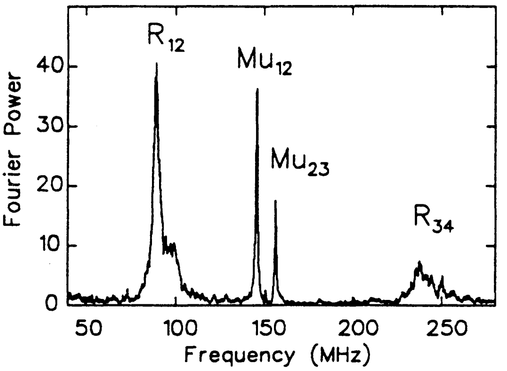

FIG. 15. Frequency power spectrum for positive muons in powdered buckminsterfullerite (crystalline C_{60}) at room temperature and 108 Oe (Kiefl *et al.*, 1992), showing simultaneously the signals from muons in the $C_{60}Mu\cdot$ radical (R_{12} and R_{34}) and the endohedral muonium (Mu@C_{60}) atom (Mu_{12} and Mu_{23}). In paramagnetic molecules with nuclear moments, such radical signals cannot be seen in low field because of nuclear hyperfine broadening.

2.2.1 Two-Counter Asymmetry

The perceptive reader will have noticed several figures with vertical axes labeled "Corrected Asymmetry." This is in reference to the fact that μSR time spectra taken in a longitudinal geometry cannot be converted to asymmetry spectra by use of Eq. (2), because neither N_0 nor B can be extracted numerically from the data without some model of the time dependence of the longitudinal polarization.

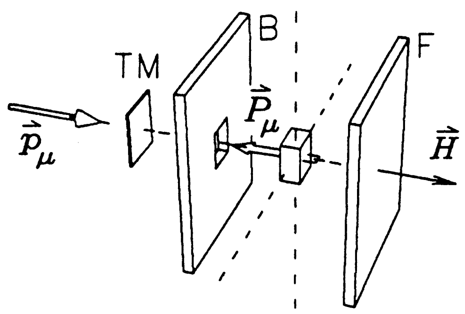

FIG. 16. Longitudinal-field μSR counter arrangement for LF_x-μ^+SR, in which the μ^+ beam is polarized antiparallel to its momentum. As $\mathbf{H} = H\hat{z} \parallel \mathbf{p}_\mu \parallel \mathbf{P}_\mu$, any magnetic field strength may be used.

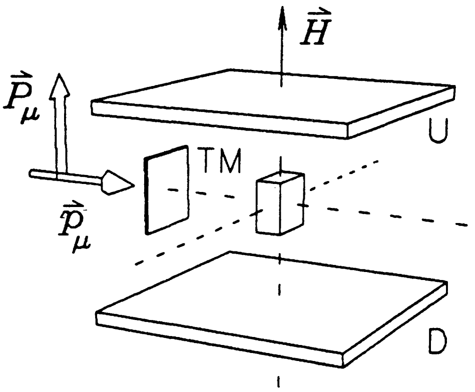

FIG. 17. Longitudinal-field μSR counter arrangement for LF$_x$-μSR, in which the muon beam arrives spin rotated by a Wien filter. Since the field is now perpendicular to $\mathbf{p}_\mu$, use with surface-muon beams is restricted to weak fields $H \lesssim 100$ Oe.

Instead, one combines the time spectra from two detectors on opposite sides of the sample, such as "U" and "D" in Fig. 17, in the following way. First, define the parameters $\epsilon_{U,D}$ = efficiency of U or D e detector, $B_{U,D}$ = background in said e detector (measured using "$t < 0$ bins"), $A_{U,D}$ = intrinsic asymmetry of e detector [count rate $\sim 1 \pm A_{U,D}$ for muons fully polarized along (opposite) detector symmetry axis], and $P_x(t)$ = muon polarization along axis. Thus,

$$N_{U,D}(t) = B_{U,D} + N_0\epsilon_{U,D}[1 \pm A_{U,D}P_x(t)],$$

where N_0 is a common normalization. The *experimental asymmetry* $a(t)$ is then obtained from either

$$a(t) \equiv \frac{[N_U(t) - B_U] - [N_D(t) - B_D]}{[N_U(t) - B_U] + [N_D(t) - B_D]} \quad \text{or}$$

$$a(t) = \frac{(1 - \alpha) + (1 + \alpha\beta)A_UP_x(t)}{(1 + \alpha) + (1 - \alpha\beta)A_UP_x(t)}, \tag{12}$$

where

$$\alpha \equiv \frac{\epsilon_D}{\epsilon_U} \text{ and } \beta \equiv \frac{A_D}{A_U}. \tag{13}$$

Thus, $(1 - \alpha)/(1 + \alpha)$ is the "baseline" asymmetry for totally unpolarized muons.

2.2.1.1 The "Corrected" Asymmetry. In addition to the obvious "baseline shift," there is also a more subtle distortion in $a(t)$ for $\alpha\beta \neq 1$. A plot of $a(t)$ looks as if it has a nonlinear scale for the abscissa. To remove these distortions completely, one must somehow independently determine α and β (usually by fitting data taken in wTF with the same geometry) and then apply the correction

$$A_UP_x(t) = \frac{(\alpha - 1) + (\alpha + 1)a(t)}{(\alpha\beta + 1) + (\alpha\beta - 1)a(t)}. \tag{14}$$

These corrections apply equally to TF-μSR asymmetry spectra formed from opposing pairs of detectors; in fact, these are usually used to fit for α. However, β must be determined from simultaneous fits to the opposing raw spectra $N_{U,D}(t)$ in TF. It is not unusual to assume $\beta = 1$, although, in principle, one should always determine this empirical parameter as accurately as possible.

2.2.2 Zero Field By extension, the field may be zero (ZF-μSR), in which case all the same arguments hold as for LF-μSR, and one measures the time evolution of the muon polarization along its original direction. In μSR, this is just a routine extension of LF, but it bears emphasis since ZF is not so simple (though not impossible) in other magnetic-resonance techniques.

2.2.3 "Relaxation" One consequence of the ease with which one can reduce the field to zero in a LF-μSR experiment is that conventional notions of *longitudinal* (T_1) vs *transverse* (T_2) relaxation processes often become confused and subject to bitter semantic arguments. In NMR, longitudinal relaxation generally occurs in a strong LF, so that a change of polarization requires unambiguous transitions between Zeeman energy eigenstates of the probe spin. It is then easy to define the longitudinal relaxation rate T_1^{-1} in terms of such spin-lattice relaxation processes. Moreover, in strong transverse fields where the Zeeman energies are much greater than any local couplings (such as dipole-dipole interactions between the probe spin and nearby magnetic moments), it is easy to define a transverse relaxation rate T_2^{-1} in terms of the *dephasing* of probe spins precessing at slightly different frequencies as a result of small differences in the local field strength at different sites.

As noted by Kubo and Toyabe (1966), these distinctions are blurred and the terminology becomes less useful as the applied field becomes comparable to the local fields and eventually goes to zero. Basically, if the components of local fields transverse to the applied field are non-negligible in the vector sum forming the total field at the probe, then, even in this classical picture, the relaxation phenomena become quite complicated and are still being worked out today (Dalmas de Réotier *et al.*, 1992). A broad review of this subject is impossible here, but a few examples can help to illustrate the potential of ZF- and LF-μSR.

2.2.3.1 Nuclear Dipolar Relaxation. In the presence of strong electric-field gradients, nuclei with electric quadrupole moments (such as Cu) exert an effective classical dipolar field on the muon:

$$\mathbf{B}_{\mathrm{dip}} = \hbar\gamma_n J_q[3(\hat{r}\cdot\hat{q})\hat{r} - \hat{q}]/r^3, \qquad (15)$$

where γ_n is the nuclear gyromagnetic ratio, J_q is the component of nuclear spin along the electric-field gradient direction $\hat{q}$, and $\hat{r} = \mathbf{r}/r$ where $\mathbf{r}$ is the vector of length r from the muon to the nucleus. The muon usually has several near-neighbor nuclei generating a broad and nearly isotropic distribution of nuclear dipolar fields. Assuming a Gaussian distribution of internal fields with random orientation leads to a muon polarization function (Hayano *et al.*, 1979)

$$g_{zz}^{\mathrm{KT}}(t) = \tfrac{1}{3} + \tfrac{2}{3}(1 - \Delta^2 t^2)\exp(-\tfrac{1}{2}\Delta^2 t^2), \qquad (16)$$

which at early times ($t \ll \Delta^{-1}$) approaches a simple Gaussian form $G_{zz}(t) \sim \exp[-\Delta^2 t^2]$. The interpretation of Δ is defined by

$$\Delta^2/\gamma_\mu^2 = \tfrac{1}{2}(\langle B_x^2\rangle + \langle B_y^2\rangle); \qquad (17)$$

thus, Δ is γ_μ times half the mean squared internal field in the plane perpendicular to the initial muon polarization (taken in this notation to be along the $\hat{z}$ direction).

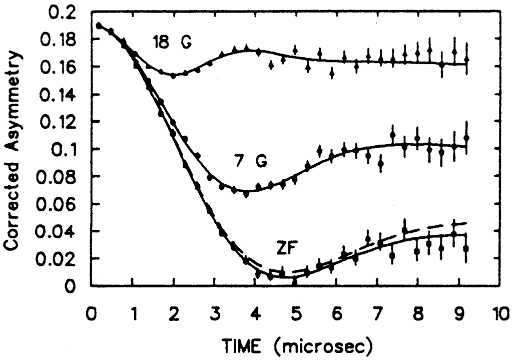

FIG. 18. Zero-field (ZF) and weak (7 and 18 Oe) longitudinal field (wLF)-μ^+SR in a single crystal of pure copper at 45 K with the muon polarization initially along the ⟨111⟩ axis of the crystal. At this temperature, the muons are almost static in the Cu lattice (hopping rate $\lesssim$ 0.1 μs^{-1}). The solid lines show fits by the exact spin Hamiltonian between the muon in an octahedral interstitial site and its six nearest-neighbor Cu spins; the dashed line shows a fit by a simple Gaussian Kubo–Toyabe function [Eq. (16)].

2.2.3.2 Motional Narrowing. Figure 18 shows a famous example of nuclear dipolar relaxation of muons in copper metal for ZF and *w*LF. At the temperature of 45 K, the muons are almost perfectly static in the Cu lattice; at higher temperatures, they diffuse by thermally activated "hopping" between adjacent octahedral interstitial sites, causing a reduction of the relaxation rate and a change of its shape toward a slow exponential decay (often referred to as "motional narrowing," in reference to the analogous phenomena in NMR, where one observes a resonance line shape whose width is proportional to the relaxation rate). Considerable literature is devoted to the mathematics of "dynamicizing" static relaxation functions (Kehr *et al.*, 1978, Hayano *et al.*, 1979, Celio, 1987). At *lower* temperatures, the muons again begin to hop as quantum mechanical tunneling becomes important; this phenomenon (and others like it) is an important illustration of quantum diffusion in the presence of dissipation (Luke *et al.*, 1991), a major area of application of μSR, which provides a light interstitial probe ideally suited to testing theories of quantum dissipation (Kagan and Klinger, 1974; Kondo, 1986; Yamada *et al.*, 1986; Kagan and Prokof'ev, 1990, 1991, 1992).

2.2.3.3 Nuclear Dipolar Oscillations. In Fig. 18, we also see a slight difference between the simple Kubo–Toyabe function [Eq. (16)] and the exact quantum mechanical solution of the coupled equations of motion of the muon and its six nearest-neighbor Cu spins (Celio, 1986). This reflects the fact that the nuclear moments are not simply static dipoles producing a field at the muon site, but also precess in the dipolar field due to the muon. Such a classical picture quickly becomes useless in picturing the actual evolution of such spin systems, particularly when the nuclei are few in number and have no electric quadrupole interactions, as in the case of the $F\mu F^-$ ion formed when positive muons are implanted into any ionic fluoride compound such as LiF (shown in Fig. 19).

2.2.4 True Relaxation In a strong longitudinal field, the muon's *spin up* and *spin down* states are good eigenstates of the Zeeman Hamiltonian, and so the muon polarization will remain static ("locked" by the applied field H) unless some magnetic perturbation drives them in resonance at their Larmor frequency $\omega_\mu = \gamma_\mu H$. Such perturbations with finite spectral density at the muon's Larmor frequency can cause its spin to flip as in magnetic resonance and lead to true relaxation (involving irreversible transitions between energy levels) at a rate T_1^{-1}. For a simple spin-lattice relaxation process with a characteristic correlation time τ_c for fluctuations of local fields of strength δ/γ_μ (whether caused by fluctuations of the fields themselves or by hopping of the muon between sites with different fields), the longitudinal relaxation rate is given by

$$T_1^{-1} = 2\delta^2\tau_c/(1 + \omega_\mu^2\tau_c^2), \tag{18}$$

which leads to a "T_1 minimum" $T_1(\min) = \omega_\mu/\delta^2$ at $\omega_\mu\tau_c = 1$.

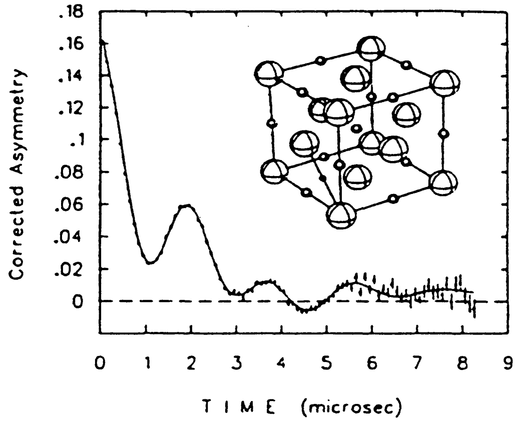

FIG. 19. Zero-field μ^+SR asymmetry spectrum in LiF at 89 K, showing the oscillatory behavior of the $F\mu F^-$ ion as a nearly isolated system of three spin-$\frac{1}{2}$ moments coupled by dipole-dipole interactions. Inset: The rock salt crystal structure showing the μ^+ site (black dot) between two F^- ions (large spheres), which are actually pulled in slightly toward the μ^+ (Brewer *et al.*, 1986).

2.2.4.1 Longitudinal-Field Muonium Relaxation. In very low field, the triplet component of the muonium atom spin system can be treated as a single spin-1 particle with approximately the magnetic moment of an electron. In this case, Eq. (18) applies as well for Mu as for the diamagnetic μ^+ (with $\omega_{\mathrm{Mu}} \approx 103\omega_\mu$ substituted for ω_μ), and we may see the behavior shown in Fig. 20 as the Mu hopping rate τ_c^{-1} changes with temperature.

In higher fields, the Mu spin system has four eigenstates, and transitions among levels become more complicated. Nevertheless, an analogous system of rate equations (Yen, 1988) can be used to extract the Mu hopping rate as a function of temperature, as shown in Fig. 21. This has allowed precise measurements of the quantum diffusion of muonium in insulators (Kadono, 1990), an important test ground for theories of quantum dissipation (see above).

2.3 Time-Integral μSR

In time-integral (I)-$\mu^\pm$SR, one simply scales the total $e^\pm$ count rate in some direction with no regard for the time of arrival of the muons. As a result, there is no rate limitation due to "pileup": Any number of muons may be in the sample at once. As for TD-μSR, there are both TF and LF versions of I-μSR.

2.3.1 Transverse-Field I-μSRotation At cw cyclotron accelerators, the beam arrives at the production target in few-ns pulses separated by the cyclotron's rf period ($1/\nu_{\mathrm{rf}} \sim$

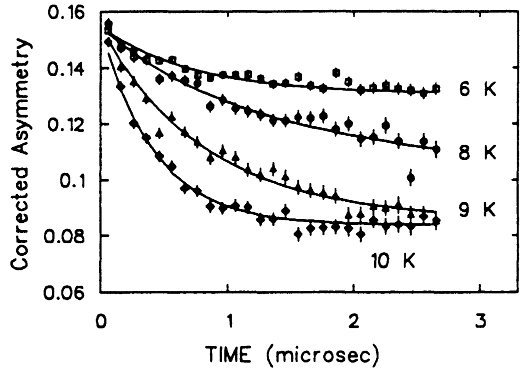

FIG. 20. Longitudinal relaxation of muonium atoms in solid nitrogen at 10 K (diamonds), 9 K (triangles), 8 K (circles), and 6 K (squares) for an applied LF of 8 Oe.

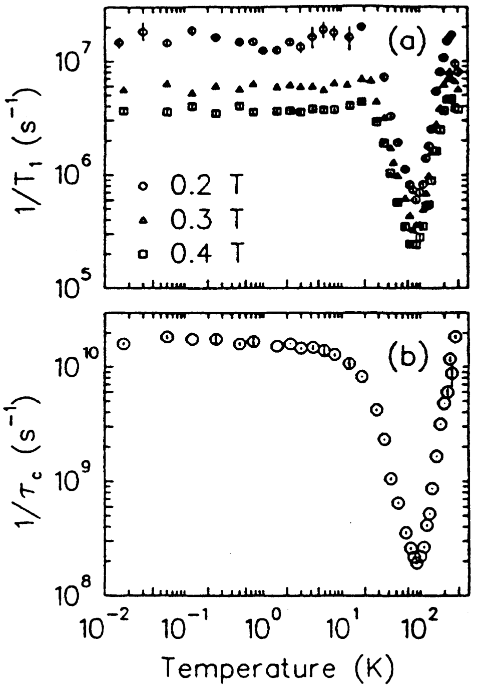

FIG. 21. (a) Longitudinal relaxation rate T_1^{-1} of muonium atoms in a pure GaAs crystal as a function of field and temperature. (b) The deduced Mu hopping rate τ_c^{-1} as a function of temperature. (Schneider *et al.*, 1992.)

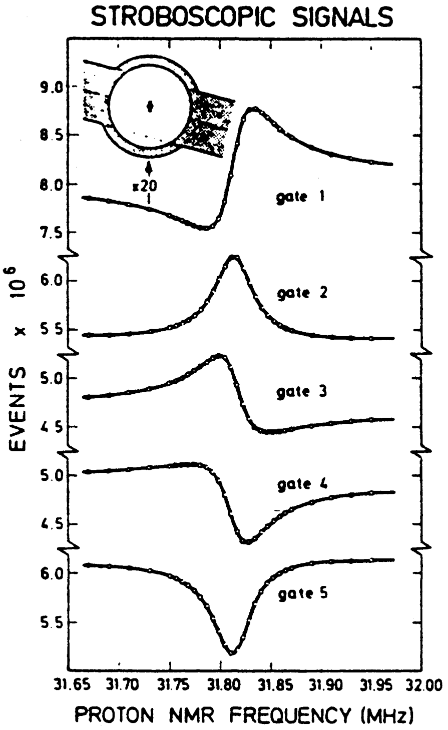

FIG. 22. Resonance signals in a strobo-μSR experiment.

20–50 ns), which must be stable to high precision. As a result, the muon beam also arrives with the same rf structure, though somewhat smeared out by the pion decay lifetime. If each muon stops in a magnetic field H of just the right strength to match an integral number n of muon-spin Larmor precession periods (n/ν_μ) to the rf period, then the polarization of muons arriving in subsequent rf "buckets" will be *in phase* with those arriving previously. Usually, n is limited to 1–3 because of the finite width of the muon buckets. The rf period is broken up into a small number of *timing gates*, during which muon decays in a given direction (determined by a counter) are accumulated in different scalers.

As illustrated in Fig. 22, the rate in a given gate as a function of magnetic field exhibits a resonance as the rf period matches an integral number of Larmor periods. In the absence of spin relaxation, the resonance line shape is Lorentzian with a width of $1/\tau_\mu$ due to the finite muon lifetime; as relaxation is included, the line shape becomes more complicated but may be fitted by a model function. All such resonance lines for the various gates are fitted simultaneously to determine the position and width of the line.

This stroboscopic μSR technique was developed to make a precise measurement of the muon's magnetic moment (Camani *et al.*, 1978) and has since been adapted to muon Knight-shift measurements (Gygax *et al.*, 1984), where its unlimited rate gives it an advantage over TD-μSR methods. "Strobo-μSR," as it is called, is best suited to situations where the muon precesses at a single well-defined frequency and where any relaxation is either uninteresting or fast enough ($\gtrsim 0.454\ \mu s^{-1}$) to be easily unfolded from the "natural linewidth." Splittings of less than this linewidth cannot be unambiguously resolved by strobo-μSR. However, single frequencies (and thus, e.g., muon Knight shifts) can be routinely measured with a precision of $\lesssim 5$ ppm by this method.

2.3.2 Longitudinal-Field IμSRelaxation

In ZF and LF, the I-μSR technique is much simpler, consisting of simply scaling decay counts in the forward (+) and backward (−) directions separately and measuring the resultant asymmetry as a function of independent variables like magnetic field or temperature. The relevant terms are defined as follows: R = rate of arrival of muons; $\epsilon_\pm$ = efficiency of e detector along (+) or opposite (−) μ polarization; $B_\pm$ = background rate in said e detector; $A_\pm$ = intrinsic asymmetry of e detector [count rate ~ $(1 \pm A_\pm)$ for muons fully polarized along (opposite) detector axis of symmetry]; $G_{zz}(t)$ = longitudinal relaxation function. The number of counts in time $T \gg T_\mu$ is

$$N_\pm = B_\pm T + RT\epsilon_\pm \pm RT\epsilon_\pm A_\pm \mathcal{L}(G_{zz}). \tag{19}$$

where

$$\mathcal{L}(G_{zz}) \equiv \int_0^\infty e^{(-t/\tau_\mu)} G_{zz}(t) \frac{dt}{\tau_\mu} \tag{20}$$

is the *Laplace Transform* of $G_{zz}(t)$.

The experimental asymmetry is

$$\mathcal{A} \equiv \frac{N_+ - N_-}{N_+ + N_-} \tag{21}$$

or

$$\mathcal{A} = \frac{b_- + (1 - \alpha) + (1 + \alpha\beta)A_+\mathcal{L}(G_{zz})}{b_+ + (1 + \alpha) + (1 - \alpha\beta)A_+\mathcal{L}(G_{zz})}, \tag{22}$$

where $b_\pm \equiv (B_+ \pm B_-)/R\epsilon_+$, $\alpha \equiv \epsilon_-/\epsilon_+$, $\beta \equiv A_-/A_+$. Useful approximations are $B_\pm \approx 0$, $A_\pm \approx A$, and $\epsilon_+ \approx \epsilon_-$, giving $\alpha \approx \beta \approx 1$, $b_\pm \approx 0$, and

$$\mathcal{A} \approx \frac{1 - \alpha}{1 + \alpha} + \frac{2}{1 + \alpha} A_+\mathcal{L}(G_{zz}), \tag{23}$$

where $(1 - \alpha)/(1 + \alpha)$ is the baseline asymmetry for totally unpolarized muons.

2.3.3 Avoided Level-Crossing Resonance

The LF I-μSR method discards all details of the time dependence of the muon polarization, preserving only the Laplace transform of $G_{zz}(t)$; thus, it may appear to be a retrogressive step. Indeed, at a *pulsed* μSR facility, where all the muons arrive at once (within some δt), there is little motive for ignoring the time dependence. However, most muon channels at meson factories are able to produce at least an order of magnitude more muons than can be accommodated in a conventional TD-μSR experiment because of the pileup ambiguity alluded to earlier. Thus, I-μSR trades off detailed information for higher sensitivity by accepting the full muon stop rate.

This is particularly useful when one is looking for *conditions of enhanced relaxation*, the most common being those values of the applied magnetic field for which the muon Zeeman splitting matches a transition between energy levels of the nuclear spins, leading to the possibility of a "flip-flop" of spins in which the muon and the nucleus simultaneously change levels with no change in total energy.

Such conditions are referred to as either "level-crossing resonance" (LCR) or "avoided level crossing" (ALC), depending mainly upon one's laboratory affiliation; the two acronyms refer to the same phenomenon, but one emphasizes the fact that levels never actually cross in quantum mechanics.

2.3.3.1 Nuclear Quadrupolar μALCR. The first such resonance studied in μSR was actually observed with TD-μSR methods, following a suggestion by Abragam (1984). In that experiment (Kreitzman *et al.*, 1986), the μ^+ Zeeman splitting at 81 Oe matches the (mostly) electric quadrupolar splitting between $m_I = \pm\frac{3}{2}$ and $m_I = \pm\frac{1}{2}$ levels of the spin-$\frac{3}{2}$ Cu nuclei, causing the resonant relaxation shown in Fig. 23.

This measurement and its later improvements (Luke *et al.*, 1991) allowed a precise determination of the electric-field gradient

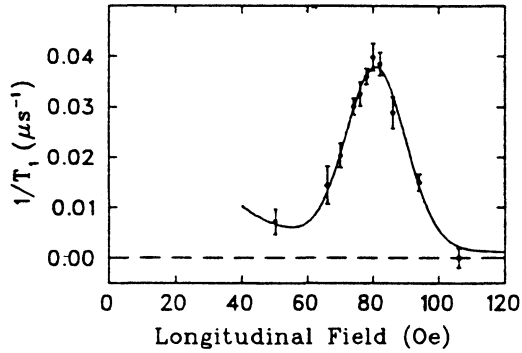

FIG. 23. Longitudinal relaxation rate of muons in a single crystal of copper with the applied field (and the muon polarization) along the ⟨111⟩ crystalline axis (Kreitzman *et al.*, 1986).

produced by the μ^+ at the Cu sites; similar experiments have since revealed nuclear quadrupolar μALCRs in solids containing ^{17}O, ^{14}N, and other nuclei. The main application is in determining the muon site in crystals.

2.3.3.2 "Muonated" Radicals. Shortly after the discovery of nuclear quadrupolar μALCR, it was realized that a much stronger resonance could occur through the hyperfine interactions of unpaired *electrons* with both the μ^+ and the nuclear spins. The quantum mechanics of this system is considerably more complicated, but the vague notion of level matching is still applicable: When the energy difference between two states in which both the muon and the nuclear spins have flipped approaches zero, resonant muon depolarization can occur.

One of the first applications of this principle to *paramagnetic* systems was in the case of the $C_6F_6Mu\cdot$ radical, a paramagnetic molecule formed by addition of a Mu atom to a double bond in the hexafluorobenzene ring (Kiefl *et al.*, 1986).

In these measurements, a time-integral method is always used, which is sometimes subject to "noise" as beam conditions shift without notice: to combat this problem and produce the smooth resonance patterns shown in Fig. 24, a *field-differential* technique is often employed: The experimental asymmetry $\mathscr{A}$ defined in Eq. (22) is measured for short time intervals (seconds) with a "toggle field" $\pm\delta H$ applied alternately along (+) or opposite (−) the main longitudinal field H_0, and the difference between these two asymmetries ($\mathscr{A}^+ - \mathscr{A}^-$) recorded as a function of H_0. This produces a signal analogous to a true differential resonance ($d\mathscr{A}/dH_0$), except that the finite difference ($\pm\delta H$) can actually be larger than the natural width of the resonance in some cases, producing a "bump" followed by a mirror-image "dip" $2\delta H$ later.

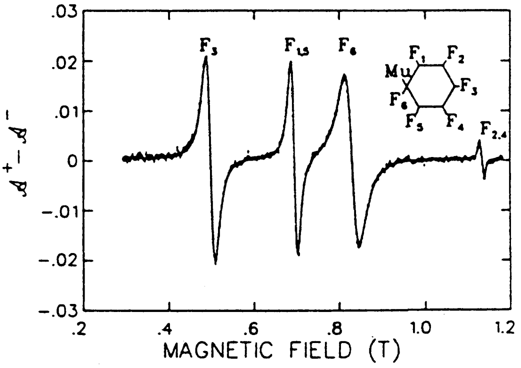

FIG. 24. Field-differential μALCR in the $C_6F_6Mu\cdot$ radical (Kiefl *et al.*, 1986)

Many such radicals have now been studied by this method, resulting in much detailed information about both the structure and the dynamics of these molecules, whose difference from the analogous radicals formed by H-atom addition is often insignificant. Several new species never observed in any other way have been discovered using μALCR spectroscopy.

2.3.3.3 Muonium in Semiconductors. Another paramagnetic system amenable to study by these methods is the muonium-like center in semiconductors, where the unpaired electron may have hyperfine interactions not only with the μ^+ but also with neighboring nuclei of the lattice. The resulting spectroscopy can be very rich, as illustrated in Fig. 25, and reveals the muonium site, as well as a great deal of information about its electronic wave function. This has been used to obtain most of what is known

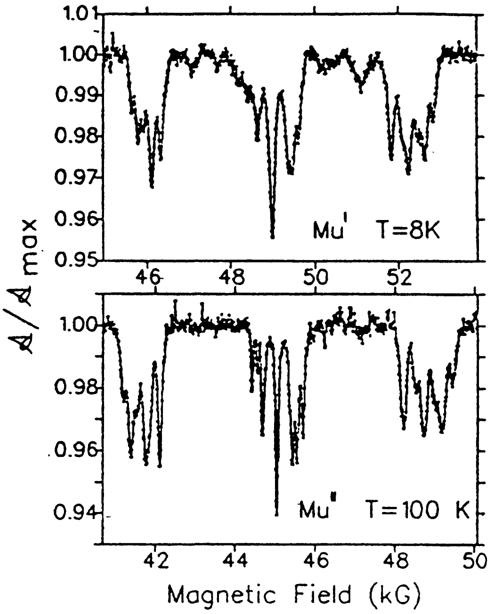

FIG. 25. Muonium-nuclear ALCR spectra in a single crystal of CuCl (Schneider *et al.*, 1990). The Mu^{I} (low temperature) and Mu^{II} (high temperature) muonium states have the same hyperfine interactions with neighboring nuclei but slightly different ω_0 values.

about isolated atomic hydrogen in Si, GaAs, GaP, and CuCl (Kiefl and Kreitzman, 1992).

2.4 Muon Spin Resonance

Traditional magnetic-resonance techniques involve intentional irradiation of the probe spin system with photons of energies equal to transitions between its eigenstates. None of the μSR techniques so far described have involved such irradiation, but true *resonance* techniques are also used in μSR, as outlined below. As in NMR, ESR, ENDOR, etc., the variety of irradiation schemes is huge and constantly growing; only a few simple generalizations are possible in limited space.

2.4.1 rf I-μSResonance The first and most familiar form of muon spin *resonance* uses an rf field in the frequency range of about 5–500 MHz to drive the polarization of muons in diamagnetic states (i.e., "bare" muons) in resonance at fields of 300 Oe to 4 T. The rf field may be generated in a simple coil at the lower frequencies or a cavity at the higher. In some cases, the rf frequency can be swept through resonance, but most often the main field H_0 is swept, as in NMR. A good recent survey of these techniques and applications may be found in Nishiyama (1992).

For the higher frequencies, the same apparatus may be used at lower fields to study paramagnetic species such as muonated radicals. An interesting variation of this is seen in Fig. 26.

The time scale for muon spin resonance is restricted by the muon lifetime; to be observed, a resonance effect must affect the muon polarization in $\lesssim 10$ μs, implying an rf field $H_1 \gtrsim 2$ Oe, which is not difficult to achieve but represents fairly high rf power. Because of power supply limitations and Ohmic heating, such irradiations are less efficient at cw facilities than at pulsed-beam facilities (Nishiyama, 1992), where all the muons can be irradiated simultaneously with timed bursts of rf power; nevertheless, rf-μSR is quite feasible at cw facilities (Kreitzman, 1990) and has enjoyed a rapid growth of applications virtually everywhere.

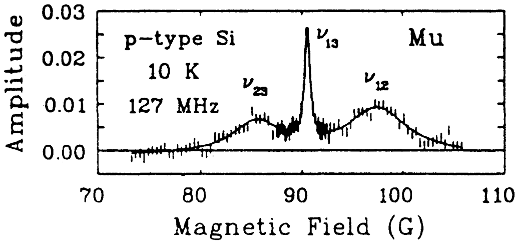

FIG. 26. Muonium resonance in mildly p-type Si at 10 K in a 127-MHz rf field, showing power-broadened resonances at the two muonium transition frequencies ν_{12} and ν_{23} (see Fig. 13), as well as a sharp two-photon resonance at the frequency ν_{13}.

One advantage of rf-μSR is that it can identify *final states* of muons in situations where stochastic processes render impossible the usual means of identification of these states by their distinctive time spectra. A typical example is the *delayed formation* of diamagnetic molecules following initial formation of Mu atoms: In TF-μSR, the early Mu precession quickly dephases the muon polarization so that no signal can be seen from the diamagnetic final state unless the reaction times are short compared to the Mu frequency; in strong LF, however, not only is the Mu polarization "held" by the applied field, but the population of the final state as a function of time can be determined by delaying the rf pulse (at the diamagnetic-resonance frequency) relative to the time of arrival of the muons.

Perhaps the most exciting feature of rf-μSR is its ability to reveal the time evolution of the muon polarization *during* the rf irradiation, a capability denied to conventional magnetic-resonance techniques, which detect the same sort of electromagnetic signal as they use to drive the probe spins. This is a consequence of the fact that μ^+SR detects high-energy positrons whose registration by the counters is unaffected by the rf fields.

2.4.2 Muon Spin Echoes Many modern NMR techniques involve *pulses* of rf power designed to rotate the probe's polarization through various angles in the rotating reference frame. This is more difficult for muons because of their short lifetime—to be useful in μSR, a $\pi/2$ pulse must be no more than $\lesssim 1$ μs long, which means an rf magnetic field of $H_1 \gtrsim 75$ Oe. This is barely possible with today's rf techniques. It is particularly difficult at cw facilities, where the rf pulse must be *asynchronously triggered* by the arrival of individual muons, in order to occur at the same time (relative to "$t = 0$" when the muon arrives in the sample) for each. Nevertheless, such *spin echo* techniques (nicknamed μSE) were first successfully demonstrated at TRIUMF, a cw facility (Kreitzman *et al.*, 1988).

As mentioned earlier, μSR detection techniques allow observation of the time evolution of the muon polarization *during* the rf irradiation. This can be especially informative in the case of μSE pulse sequences. An example is shown in Fig. 27.

A bonus of rf-μSR at pulsed facilities is the chance to overcome the *time resolution* limitations of pulsed beams (muons arrive in bunches $\gtrsim$50 ns long) by stopping the muons in a strong LF and then using a high-power phase-locked $\pi/2$ pulse to rotate them all *in phase* into the plane perpendicular to $\mathbf{H}_0$, after which they will exhibit HTF-μSR precession with all the same observables as enjoyed at cw facilities.

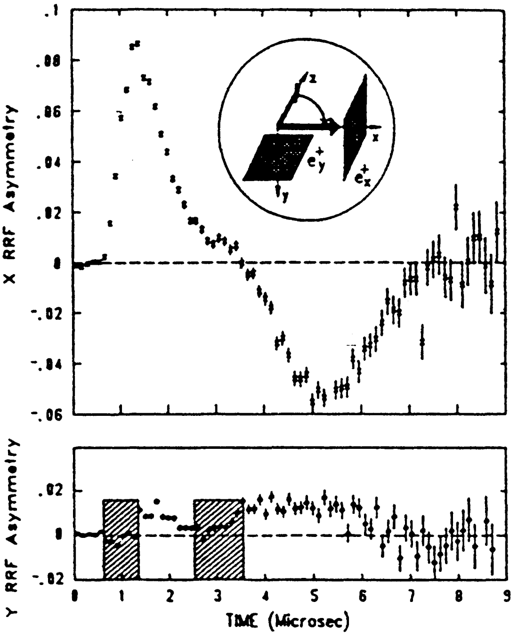

FIG. 27. Muon spin echo in the rotating reference frame during a $(\pi/2)_y$–τ–π_y pulse sequence (shaded regions) where the muons arrive initially polarized along $\mathbf{H}_0 = H_0\hat{z}$, and the detectors are in the x and y directions. The small signal observed along the y axis indicates that the rf was slightly off resonance. Inset: The geometry of the positron detectors and the polarization during the $\pi/2$ pulse (Kreitzman *et al.*, 1988).

3. μSR APPLICATIONS

Examples of selected experimental results having been shown above, this section will focus mainly on the "scientific demography" of μSR.

3.1 Muonium Chemistry

One of the main themes of μSR could be characterized as "The Chemistry of a Light Isotope of Hydrogen." The muonium atom and the H atom have the same size, and almost the same reduced mass and ionization potential, and both obey the Born-Oppenheimer approximation (the electron wave functions adiabatically adjust to the slow motions of the nuclei) that allows relatively simple calculations of chemical reaction processes and molecular structure.

3.1.1 Chemical Kinetics This strong similarity, combined with the unprecedented *isotopic* difference (a ratio of ~18 in mass between Mu and D) and the enhanced quantum behavior of the light Mu atom, makes muonium chemistry a very important testing ground for the few *ab initio* theories of chemical reaction kinetics (Fleming and Senba, 1992).

Once the *differences* between Mu and H chemistry are well understood (as is now the case for many types of reactions), measurements of Mu reactivity (which are often easy) can be used to predict reliably the reactivity of H under circumstances where H atoms cannot even be detected. This is now entirely feasible but has not yet attracted much attention from the "practical" side of the chemistry community.

3.1.2 Radicals Following chemical reaction, Mu is often incorporated into *radicals* (molecules with unpaired electrons) where a weakened hyperfine interaction persists between the electron and muon spins (and also between the electron and any other nuclei with magnetic moments). These molecules resemble their hydrogenic analogs (where the μ^+ is replaced by a proton) so closely that most of their reaction kinetics are nearly identical. This allows use of the "muonated" version to learn the chemical behavior of the "protonated" version even when information on the latter is unavailable by other methods. In some cases, μSR has confirmed the very *existence* of radicals that have never been observed by any other means.

3.1.3 Molecular Structure By using μALCR techniques, one can use μSR to measure not only the hyperfine coupling of the muon to the unpaired electron but also the couplings of that electron to the *other* nuclei; thus, the muon can be used to read out the molecular structure of regions of the molecule quite far from the muon's site, much as in a conventional ENDOR experiment (Percival *et al.*, 1987). The details of the *structure* of muonium- and hydrogen-substituted versions of radicals are sometimes influenced by the lighter mass of the muon: Bond lengths to Mu tend to be slightly longer, and bond angles are often changed by the larger zero-point motions of Mu. This, combined with the fact that the nuclear hyperfine couplings can often be measured more accurately by μSR than by ENDOR, makes μALCR a very important probe of molecular structure (Roduner, 1988).

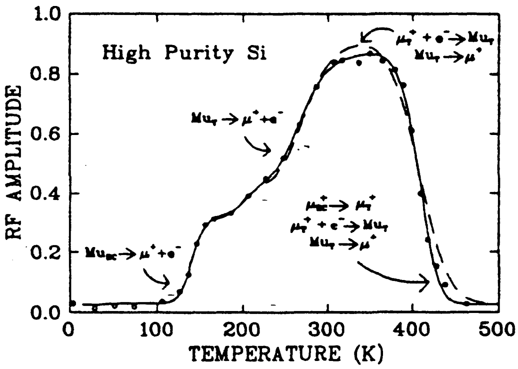

FIG. 28. The rf-μ^+SR diamagnetic-resonance amplitude in pure silicon, showing evidence for thermally activated transitions between various charge states and lattice sites of the μ^+: Mu_{BC} = anomalous muonium (also called Mu*) localized at a Si-Si bond-centered site: Mu_T = quasi-free Mu atom diffusing rapidly between tetrahedral sites; μ_{BC}^+ and μ_T^+ indicate freshly ionized muons at the corresponding sites.

3.2 Condensed-Matter Physics

The applications of μSR in condensed-matter physics also include many comparisons between the μ^+ and the proton, insofar as their behaviors in matter differ only by virtue of the muon's lighter mass; however, there are perhaps a greater variety of applications that might be titled, "The Muon as a Magnetic Probe." Both types will be surveyed briefly below.

3.2.1 Hydrogen in Semiconductors Closely related to the muonium chemistry applications discussed above are the uses of μ^+SR to learn the location, electronic structure, and dynamical behavior of Mu and/or H atoms in semiconductors, where H is known to be an ubiquitous impurity but where far more is known about isolated H atoms from μSR than from any method that observes H itself (Patterson, 1988, Kiefl *et al.*, 1988). Since H is known to affect the electronic properties of Si and GaAs and is now being used specifically for the purpose of *passivation* of electrically active impurities, this knowledge is vital to the semiconductor industry; consequently, μ^+SR spectroscopy in semiconductors is one of the most important applications of μSR in solids.

Figure 28 shows the temperature dependence of the probability of finding the muon in an ionized state (either μ^+ or Mu^-) at times comparable to the muon lifetime (long after any initial formation or reaction of paramagnetic states), obtained by rf-μSR at the diamagnetic frequency. Comparison of these results with those for highly doped *p*- or *n*-type Si reveals the interactions between H-like species and impurities (Hitti *et al.*, 1994).

3.2.2 Quantum Diffusion As mentioned earlier, the theory of *quantum dissipation* (quantum tunneling or bandlike propagation in the presence of stochastic or other interactions with the host medium) is currently a very important field of condensed-matter physics, having implications for the electronic and diffusive transport properties of all types of materials (Stamp and Zhang, 1991; Kagan and Prokof'ev, 1992). Because of its light mass, its affinity for electrons, and its repulsion from nuclei, the μ^+ (and its neutral version, Mu) is an apt probe of such behavior, the ideal "light interstitial" (Flynn and Stoneham, 1970; Kagan and Klinger, 1974; Petzinger, 1982). Study of the *quantum diffusion* of μ^+ and Mu in solids has therefore constituted a significant fraction of the μSR community's experimental effort over the last two decades (Kadono, 1992), and the results of such experiments have attracted considerable theoretical interest (Kagan and Prokof'ev, 1992).

***3.2.2.1 Motion of the μ^+* in Metals.** In impure or imperfect metallic crystals, the μ^+

tends to be trapped at defects or at least to be localized in their vicinity; the resultant temperature dependence of motional narrowing effects, used to measure μ^+ mobility, can be quite complicated (Möslang *et al.*, 1983, Petzinger, 1982). In pure, well-annealed crystals, however, one often observes the intrinsic interactions between the μ^+ and the lattice. The hopping of positive muons in various pure metals has been studied extensively by means of ZF- and wLF-μ^+SR measurements such as those shown in Fig. 18 (Luke *et al.*, 1991), as well as by TF-μ^+SR motional narrowing experiments. These studies have revealed a qualitatively consistent temperature dependence typified by the classic example shown in Fig. 29. At high temperatures (above about 80 K in the case of Cu), the muon exhibits semiclassical thermally activated hopping over the energy barriers between adjacent interstitial sites; at lower temperatures (below about 20 K for Cu), the μ^+ actually hops faster as T decreases, because of the increasing probability of quantum tunneling between sites, as the disorder due to lattice vibrations is reduced. The power law $\tau_e^{-1} \propto T^{-\alpha}$ is predicted by theory (Kondo, 1992) to have a much smaller exponent α for screened positive muons in metals than in insulators, as a result of the additional dissipation in electronic degrees of freedom (often referred to as "electron drag" effects). At still lower T, the intrinsic disorder of even pure high-quality crystals begins to inhibit tunneling once again. Similar results have been observed in aluminum using indirect methods involving trapping of the μ^+ at impurities.

3.2.2.2 Muonium Diffusion in Nonmetals. In ionic crystals, the μ^+ tends to form a hydrogen bond with the most negative species in the lattice, such as F^- or O^{--} ions. Muon diffusion is thus suppressed until quite high temperatures, typically ≃200 K (Kiefl *et al.*, 1990). If the muon forms muonium, however, the Mu atom is more or less decoupled from local electric fields and usually diffuses freely through the lattice; interestingly, Mu diffusion is quite similar in both covalent and ionic nonmetals. An example of Mu diffusion in the semiconductor GaAs was shown earlier in Fig. 21 (Schneider *et al.*, 1992); a startlingly similar result was obtained (Kiefl *et al.*, 1989) in an ionic insulator (KCl), as shown in Fig. 30.

Somewhat different Mu hopping behavior is observed in the van der Waals insulator s-N_2 formed by crystallization of solid nitrogen, as shown in Fig. 31. Such so-called *cryocrystals* have very low Debye temperatures and quite weak couplings between the neutral Mu atom and the host lattice; they are, therefore, interesting systems in which to test theoretical predictions of the temperature dependence of Mu quantum diffusion (Storchak *et al.*, 1994).

An interesting effect has been observed in *superconducting aluminum*, which can be driven normal by an applied magnetic field: At the same T, muon diffusion is considerably *faster* in the superconducting (SC) state than in the normal state, because the opening of a SC gap at the Fermi surface effectively quenches the dissipation due to screening electrons (Kondo, 1992). Thus, SC

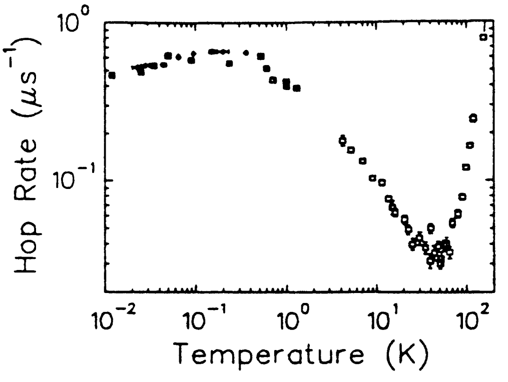

FIG. 29. Temperature dependence of the μ^+ hopping rate τ_c^{-1} in pure copper crystals (Luke *et al.*, 1991).

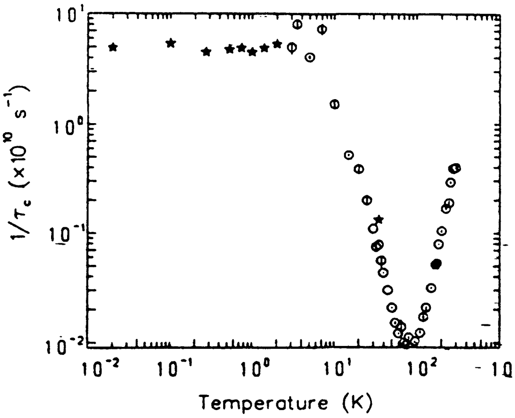

FIG. 30. Temperature dependence of Mu hopping rate in KCl (Kiefl *et al.*, 1989).

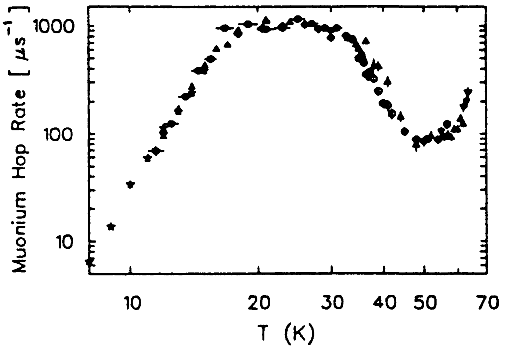

FIG. 31. Temperature dependence of the muonium hopping rate in solid nitrogen cryocrystals (Storchak *et al.*, 1994).

Al acts like an insulator with respect to muon quantum diffusion.

3.2.3 Magnetism

Perhaps because it so easily generates rich and precise data, *magnetism* is currently the largest area of application of μSR.

In magnetically ordered materials, the muon (μ^+ or μ^-) generally samples the local magnetic field at some preferred site or sites and reveals that field through its characteristic Larmor precession frequency (Seeger and Schimmele, 1992). The local field $\mathbf{B}_\mu$ is composed of several contributions:

$$\mathbf{B}_\mu = \mathbf{H}_0 + \mathbf{B}_\mathrm{L} + \mathbf{B}_\mathrm{dip} + \mathbf{B}_\mathrm{hf}, \tag{24}$$

where $\mathbf{H}_0$ is the externally applied field (if any), $\mathbf{B}_\mathrm{L}$ is the *Lorentz field* produced by magnetic "charges" induced on the interior of a hypothetical spherical cavity around the muon site due to the average bulk magnetization of the medium, $\mathbf{B}_\mathrm{dip}$ is the *dipolar field* due to all the microscopic moments within the Lorentz cavity, and $\mathbf{B}_\mathrm{hf}$ is a *hyperfine field* transmitted to the muon by its Fermi contact interactions with conduction electrons. The last two components are usually the focus of experiments on magnetically ordered systems, where $\mathbf{B}_\mu$ is usually measured with $\mathbf{H}_0 = 0$, as well as measurements of the muon Knight shift $K_\mu \equiv (B_\mu - H_0)/H_0$ in paramagnets.

Interpretation of these fields requires a knowledge of the muon site, including zero-point motion and any tunneling between nearby sites, and can be quite difficult; nevertheless, in many cases, it is easier than trying to sort out the "core polarization" effects of various interacting atomic electrons that transmit the hyperfine field to the nucleus in (e.g.) NMR.

One attractive feature of μSR in magnetism studies is that measurements of internal magnetic fields can be performed on unaligned polycrystalline or powdered samples almost as easily as on perfect single crystals. In zero applied field, if all the local fields have the same *magnitude*, the only effect of random *directions* is to reduce the muon precession amplitude to $\frac{2}{3}$ of the value that would be observed if the fields were all perpendicular to the muon polarization. Thus, ZF-μSR, which requires a minimum of apparatus, is often adequate for very detailed investigations of the temperature dependence of magnetic order and dynamics. Moreover, using ZF-μSR, one can measure B_μ in crystals exhibiting antiferromagnetism or even more exotic forms of magnetic order as easily as in simple ferromagnets. A classic example is shown in Fig. 32 for the intermetallic compound MnSi, where the magnetic field due to itinerant-electron moments forms a long-wavelength spiral with a small component along the axis of the spiral—so-called *helimagnetic* order.

Disordered magnetic systems (such as *spin glasses*) can also be studied rather easily by μSR, leading to a strong μSR presence in the currently popular area of frustrated and low-dimensional magnetism. A classic example is shown in Fig. 33 and a more recent case in Fig. 34.

3.2.4 Superconductivity

Because muon decay provides a nonelectromagnetic "sig-

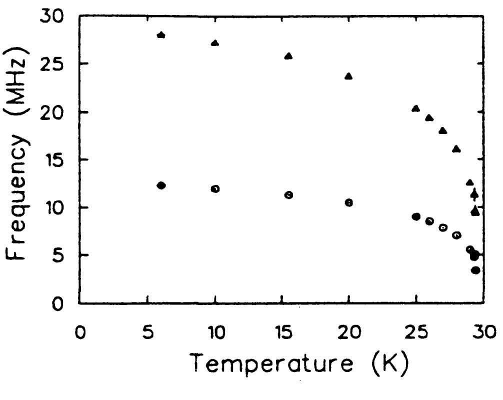

FIG. 32. Temperature dependence of the magnitudes of the internal fields at two different μ^+ sites in a single crystal of the itinerant-electron helimagnet MnSi.

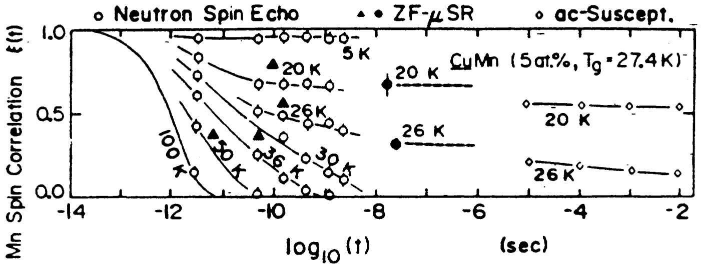

FIG. 33. Comparison of neutron spin-echo (NSE), ZF-μSR, and ac-susceptibility measurements on the spin glass CuMn (5 at. %), showing the time dependence of the static order parameter ξ (related to the fraction of Mn moments that have not yet moved appreciably) at several temperatures below the nominal glass transition at T_g = 27.4 K (Uemura *et al.*, 1984).

nal,'' μSR is able to probe magnetic fields inside superconductors without distortion. This gives μSR a unique window on the magnetic behavior of superconductors. Since 1987, the μSR community has exploited this advantage in the study of high-temperature superconductors (HTSC).

3.2.4.1 Phase Diagrams. All of the copper oxide superconductors show either antiferromagnetic (AFM) or spin-glass-like (SGL) order very close to (and sometimes overlapping) the superconducting (SC) phase, as first demonstrated for $YBa_2Cu_3O_x$ by μSR experiments in 1988. Since then, the temperature vs doping phase diagrams of most HTSC materials have been mapped out using μSR.

3.2.4.2 Magnetic Penetration Depth. The first application of μSR to HTSC was in measuring the magnetic penetration depth λ by fitting the high transverse field (HTF)-μSR time spectrum by a simple Gaussian relaxation function and thus estimating the second moment of the frequency distribution due to the vortex lattice of magnetic flux penetrating a type-II superconductor in the mixed state. Figure 35 shows the earliest results, obtained in January 1987 (Aeppli *et al.*, 1987). Since then, an enormous body of data has been accumulated by this method, which is still the most reliable way to estimating λ in sintered, random powders. However, as evident in Fig. 11, for well-characterized single crystals one finds a line shape much less symmetric than a Gaussian, as predicted by London and later Gor'kov models of the vortex lattice. The ability to make such measurements and interpret them correctly in terms of λ and ξ is a recent achievement of μSR, which is now poised to exploit these capabilities to legitimize (or, in some cases, debunk) the many published μSR results worldwide interpreting the Gaussian-fitted

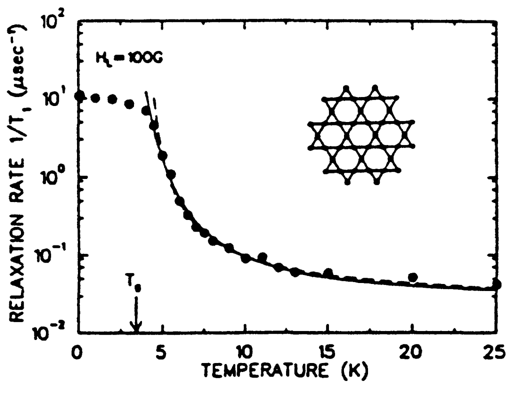

FIG. 34. Muon spin-relaxation rate T_1^{-1} due to fluctuating local fields in the antiferromagnetically frustrated two-dimensional kagomé lattice (inset) of $SrCr_8Ga_4O_{19}$ (Uemura *et al.*, 1994). The increase of T_1^{-1} with decreasing T is a manifestation of the gradual slowing down of Cr spin fluctuations; the saturation of this effect below T_g shows that the Cr moments never freeze even as $T \rightarrow 0$. This is in marked contrast with the situation in dilute spin glasses like CuMn, where all fluctuations cease as $T \rightarrow 0$.

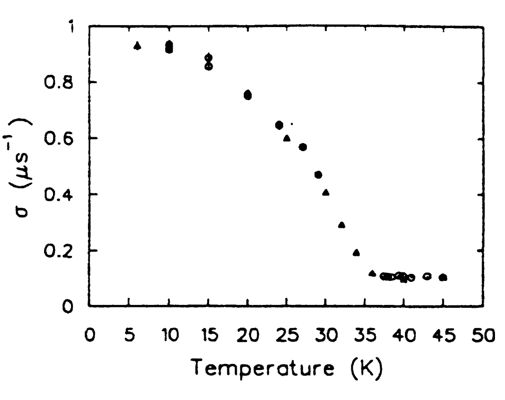

FIG. 35. First TF-μ^+SR measurement of the temperature dependence of the μ^+ dephasing rate σ (proportional to λ^{-2}, the inverse square of the magnetic penetration depth, which is, in turn, proportional to the superconducting carrier density n_s) in a high T_c cuprate superconductor—in this case, an unaligned sintered powder sample of $La_{1.85}Sr_{0.15}CuO_4$ in an applied field of 0.4 T (Aeppli *et al.*, 1987).

linewidths (for sintered powder or aligned HTSC samples) in terms of λ.

3.2.4.3 Universal Correlations between T_c and n_s. A collection of such results has been assembled by Uemura *et al.* (1989) to produce the plot shown in Fig. 36, commonly known as an "Uemura plot." The "universal" linear dependence of T_c on n_s in the *underdoped* (less than optimal n_s) regime gives way to a "turnover" and eventual linear *decrease* of T_c with increasing n_s in the *overdoped* regime (Niedermayer *et al.*, 1993).

***3.2.4.4* Non-*s*-Wave Pairing?** Whether it be crudely characterized in terms of a Gaussian relaxation rate σ or more accurately represented by the second moment of a line shape, the muon dephasing rate due to the vortex lattice in a type-II superconductor is proportional to the inverse square of the London penetration depth, λ^{-2}, which in turn is proportional to the superconducting carrier density n_s. Thus, a measurement of the temperature dependence of σ is (apart from a normalization) also a measurement of the excitation of normal carriers across the superconducting energy gap that characterizes the pairing mechanism. A conventional BCS "*s*-wave" superconductor has a prescribed T dependence of n_s, variations from which are taken as evidence for unconventional pairing mechanisms. While all this should, in the author's opinion, be taken with a grain of salt, still μSR has been an important contributor to the debate over the still-mysterious mechanism for HTSC. While the μSR data for the 18-K superconductor K_3C_{60} are consistent with *s*-wave BCS superconductivity, the μSR results on certain organic and heavy-fermion superconductors show a T dependence strongly suggestive of nodes in the gap (momentum-space directions in which the superconducting gap vanishes) (Uemura, 1992). A very similar T dependence has been observed for n_s in various samples of $YBa_2Cu_3O_{6.95}$—most notably in high-quality single crystals—as shown in Fig. 37.

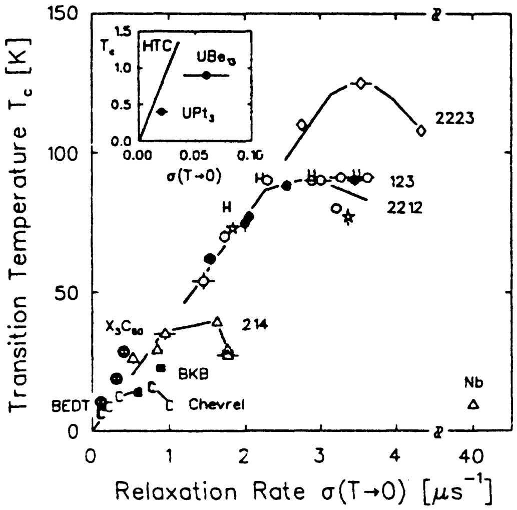

FIG. 36. Correlations between the superconducting carrier density n_s (as measured by the μ^+ dephasing rate σ) and the superconducting transition temperature T_c for a wide variety of "exotic" superconductors including many HTSC; conventional BCS superconductors have higher n_s with lower T_c.

4. μSR FACILITIES

No practical guide to μSR would be complete without some discussion of the *logistics* of μSR experiments and the *facilities* required for its pursuit.

4.1 Logistics of Accelerator Experiments

Because μSR requires muons, which must be produced by large, high-intensity accelerators, μSR beam time will probably always be a scarce commodity for which users will compete fiercely. Such competition has

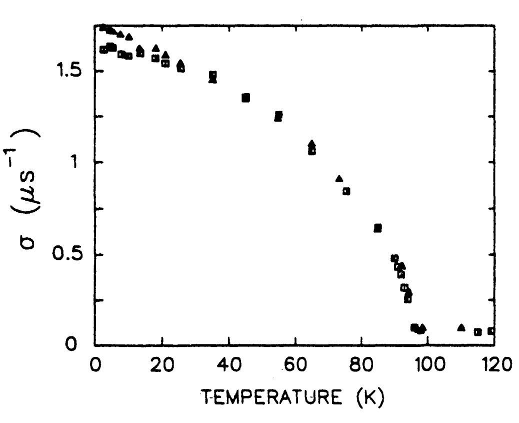

FIG. 37. Temperature dependence of the μ^+ dephasing rate σ (proportional to the superconducting carrier density n_s) in single crystals of $YBa_2Cu_3O_{6.95}$ in a TF-μ^+SR experiment at 0.5 T (triangles) and 1.5 T (squares). An "*s*-wave" BCS superconductor is expected to show very weak T dependence near T = 0; in contrast, these data indicate an almost linear decrease from T = 0.

a beneficial effect on the quality of the resulting research, but the scarcity of muons does regulate the growth of the field. It will probably always be the case that μSR experiments will be screened in advance by peer review committees and allocated beam time every few months for periods ranging from a few days to several weeks. While this sort of schedule may often hamper spontaneity, it is well matched to the cycle of experimental conception, design, execution, data analysis, and publication, as evidenced by the high output of most μSR facilities.

Furthermore, because a typical μSR experiment involves so many diverse technologies (from magnetic beam optics to charged-particle detectors to magnets and cryostats to fast electronics and extensive on-line computer software), it makes an excellent training ground for students, who, once having mastered the techniques of μSR, can apply them almost unaltered to a wide variety of scientific topics.

4.2 Standard μSR Hardware

Most accelerator laboratories with μSR programs also maintain some sort of μSR *user facility* providing standard μSR spectrometers, electronics, and data-acquisition computers; it is now rare for individual experimental groups to provide their own equipment, except for specialized cryostats, ovens, sample cells, or data-analysis software. Some of the standard components to be found at most μSR user facilities are listed below.

4.2.1 Detectors The workhorse of μSR is the plastic scintillation counter, in which a flash of light (generated as an ionizing particle passes through the scintillating plastic) is transmitted down clear plastic light guides by total internal reflection to a photomultiplier tube, which, in turn, emits an electrical pulse that is transmitted down a coaxial cable to a fast timing discriminator module in the counting room outside the experimental area (from which experimenters are excluded while the beam is on, because of the appreciable radiation levels in the muon beam itself). The result is a logical timing pulse whose arrival time corresponds within ~1 ns (and a fixed delay) to the time the particle passed through the detector. The M and E detectors in Fig. 1 would normally be plastic scintillation counters for the incoming muon and outgoing positron respectively, and would normally provide a resolution of 0.5–2 ns on the corresponding time interval. With careful design of the counters and special photomultiplier tubes, cables, and fast electronics, it is possible to achieve time resolutions of ~100 ps (Holzschuh, 1983; Keller *et al.*, 1987), but this is not yet common.

While scintillation counters are simple, reliable, and versatile and have excellent performance in the time domain, they have a number of frustrating features: First, even if the scintillator itself is small, the light guides are bulky and inflexible (flexible fiber optics can be used but so far only at the expense of timing degradation and/or loss of light-collection efficiency); this becomes increasingly inconvenient as increasing stopping luminosity reduces the scale of μSR samples and experiments. Second, most photomultiplier tubes are very sensitive to magnetic fields and must be magnetically shielded and/or removed to field-free regions by long light guides, which again reduce time resolution. Third, scintillation counters have only crude position sensitivity, which is generally of secondary importance to μSR experiments, but could be useful. For example, by distinguishing one incoming muon from another and matching the outgoing positrons to the muons from which they arise, one may overcome the pileup rate limitation normally imposed by allowing only one muon in the sample at a time; this technique has been demonstrated but never widely applied.

Potentially promising alternatives include modern proportional counters and solid-state barrier detectors, but neither of these have the required time resolution in their standard configurations; more development work is needed.

4.2.2 Electronics Raw counter pulses are fed to fast discriminators—usually constant-fraction discriminators (CFDs)—which generate uniform timing logic pulses if the raw pulse height is above a set threshold. The logic pulses are processed by a simple "fast electronics" circuit shown schematically in Fig. 38.

The incoming muon generates a pulse in the M counter, which ultimately sends a start pulse to a fast time digitizer ("clock"); later,

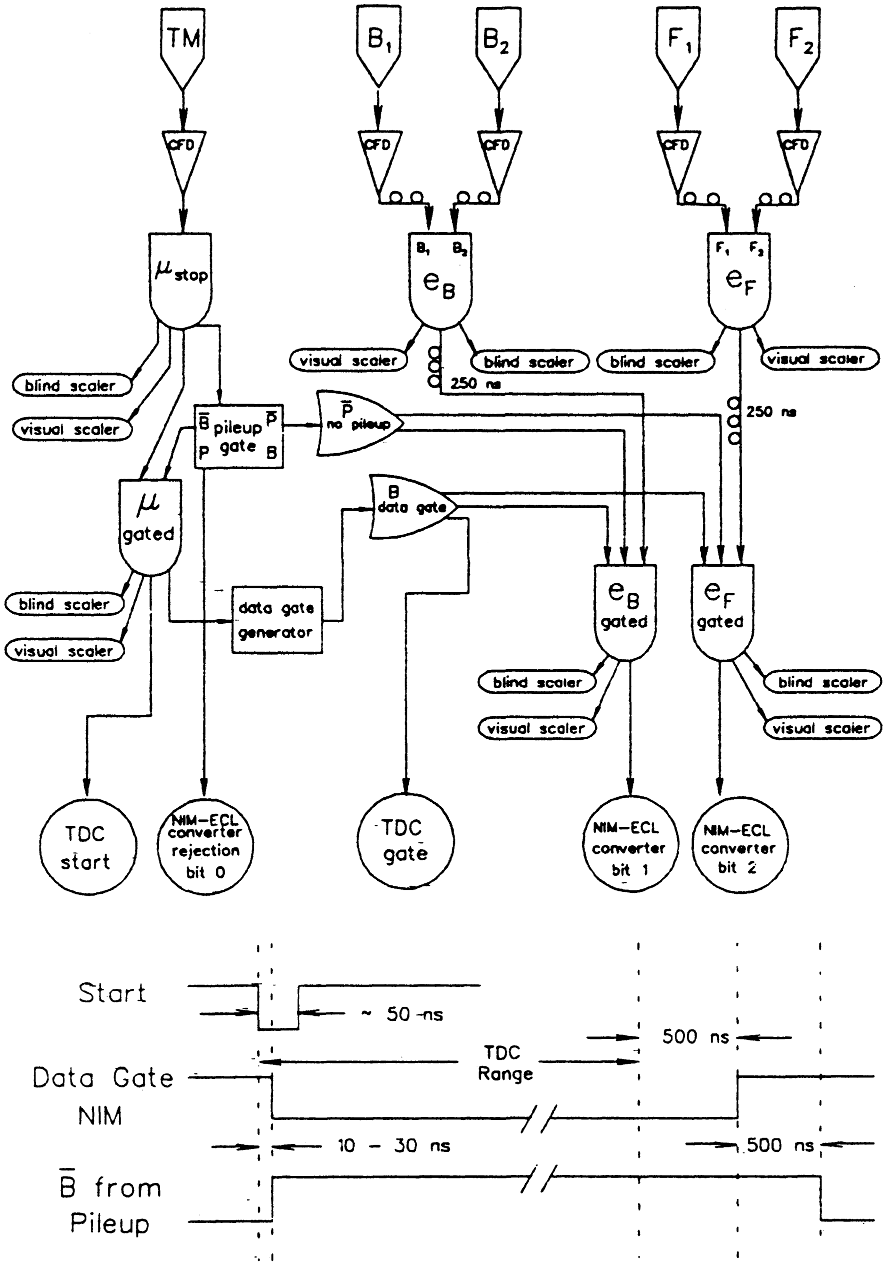
TM
B1
B2
F1
F2
CFD
μ stop
B1 B2
eB
F1 F2
eF
blind scaler
visual scaler
250 ns
B̄ pileup P̄
P gate B
P̄ no pileup
B data gate
μ gated
data gate generator
eB gated
eF gated
TDC start
NIM-ECL converter rejection bit 0
TDC gate
NIM-ECL converter bit 1
NIM-ECL converter bit 2
Start
~ 50 ns
TDC Range
500 ns
Data Gate NIM
10 – 30 ns
B̄ from Pileup

FIG. 38. Schematic electronics logic diagram for two "channels" of a typical time-differential (TD)-μSR experiment (see Fig. 1).

the muon decays, triggering a stop pulse in the *E* counter, which stops the clock; the time interval is digitized, and the corresponding bin in a time histogram is incremented. The additional logic modules ensure that events with "second muons" are rejected with uniform efficiency throughout the gate *T*, lest rate-dependent distortions spoil the resultant time spectra (histograms), which are the final output of the experiment.

Similar care must be taken to reject "second *E*" events in which a (possibly accidental) count in an *E* counter renders the identity of the stop pulse ambiguous. This function is handled internally in the LRS 4204 time-to-digital converter, which is now the standard μSR clock primarily because it provides this function along with excellent time resolution and the ability to directly increment time bins in a CAMAC Histogramming Memory, thus relieving the data-acquisition computer of all event-processing duties except periodic histogram and scaler readout.

There are numerous improvements and adaptations of this basic TD-μSR arrangement, of course, as well as several entirely different electronics setups for time-integral (I)-μSR techniques.

4.2.3 μSR Spectrometers The local facility almost always supplies a choice of several μSR spectrometers, whose function is to control the magnetic field at the sample (usually by means of three orthogonal sets of Helmholtz coils, one of which can generate a large main field) and to provide mounting brackets for counters, cryostats, or other paraphernalia. Often a standard set of counters is an integral part of the spectrometer, as well. There is such a huge variety of such spectrometers that it would be pointless to try to describe a "typical" version in any detail.

4.2.4 Target Vessels and Cryostats Since the muon's initial polarization is independent of the state of its environment, μSR samples may be gases, liquids, or solids at any temperature, pressure, or magnetic or electric field. This versatility is reflected in the variety of sample environments (target vessels) used in μSR experiments, ranging from large "gas cans" (for studying hydrogen gas at low pressure and high temperature) to miniaturized cryostats (for studying small crystals at low temperature and high magnetic field). The most common use of μSR is currently in low-temperature condensed-matter physics, so that most μSR spectrometers have general-purpose cryostats built in (but removable). However, in general, the user is responsible for the sample environment.

4.3 μSR Software

Because μSR data is always intimately connected with computers (indeed, one needs computer graphics just to see what one is measuring) and because there is such a premium on efficient use of beam time, which, in turn, demands that one know what one has learned from one run before setting the independent variables for the next (it is not unusual for a μSR experimenter to generate, while a spectrum is still accumulating, the figures that will be published later to describe it), one of the most important components of any μSR user facility is a suite of process-control, data-presentation, and data-analysis software to provide the user with all the mathematical tools described in this article, and more. Since implementation of such utilities is nontrivial, some effort is now being made to define international standards (data formats, process control, data analysis, *etc.*), so that the μSR community can cooperate on such software development and share the benefits thereof.

5. SPECULATIONS

The felicitous history of μSR has reached a point of crisis in the 1990s: Accelerator facilities rarely capture the fascination of the subatomic physicists who build them for more than about two decades, and the meson factories have reached that difficult age. While μSR facilities require only modest investments for qualitative improvements using existing accelerators, the big machines themselves need expensive upgrades to maintain world-class status as subatomic facilities. Thus, the future of μSR remains inextricably linked to that of subatomic physics—more specifically, to that of high-intensity, intermediate-energy, fixed-target hadron accelerators. This has always been the case and will remain so until someone invents a tabletop device capable of generating muon beams. In spite of the implied mutual liability, the author has always considered this unlikely symbiosis to be one of the greatest charms of μSR.

Works Cited

Abragam, A. (1984), *C.R. Acad. Sci. Ser. B* **299**, 95–99.

Aeppli, G., Ansaldo, E. J., Brewer, J. H., Cava, R. J., Kiefl, R. F., Kreitzman, S. R., Luke, G. M., Noakes, D. R. (1987), *Phys. Rev. B* **35**, 7129–7132.

Bowen, T. (1985), *Phys. Today* **38** (7), 22–34.

Brewer, J. H., Crowe, K. M., Gygax, F. N., Johnson, R. F., Patterson, B. D., Fleming, D. G., Schenck, A. (1973), *Phys. Rev. Lett.* **31**, 143–146.

Brewer, J. H., Crowe, K. M., Gygax, F. N., Johnson, R. F., Fleming, D. G., Schenck, A. (1974), *Phys. Rev. A* **9**, 495–507.

Brewer, J. H. (1981), *Hyperfine Int.* **8**, 831–834.

Brewer, J. H. (1984a), *Hyperfine Int.* **17–19**, 879–884.

Brewer, J. H. (1984b), *Hyperfine Int.* **17–19**, 873–878.

Brewer, J. H., Kreitzman, S. R., Noakes, D. R., Ansaldo, E. J., Harshman, D. R., Keitel, R. (1986), *Phys. Rev. B* **33**, 7813–7816.

Camani, M., Gygax, F. N., Klempt, E., Rüegg, W., Schenck, A., Schilling, H., Schulze, R., Wolfe, H. (1978), *Phys. Lett. B* **77**, 326–330.

Celio, M. (1986), *Phys. Rev. Lett.* **56**, 2720–2723.

Celio, M. (1987), *Helv. Phys. Acta* **60**, 600–610.

Combley, F. H., Farley, F. J. M., Picasso, E. (1981), *Phys. Rep.* **68**, 93–119.

Cronin, J. W. (1981), *Rev. Mod. Phys.* **53**, 373–383.

Dalmas de Réotier, P., Yaouanc, A., Meshkov, S. V. (1992), *Phys. Lett. A* **162**, 206–212.

Fermi, E., Teller, E. (1947), *Phys. Rev.* **72**, 399–408.

Fitch, V. L. (1981), *Rev. Mod. Phys.* **53**, 367–371.

Fleming, D. G., Senba, M. (1992), in: T. Yamazaki, K. Nakai, N. Nagamine (Eds.), *Perspectives of Meson Science*, Amsterdam: North-Holland, pp. 219–264.

Flynn, C. P., Stoneham, A. M. (1970), *Phys. Rev. B* **1**, 3966–3978.

Friedman, J. I., Telegdi, V. L. (1957), *Phys. Rev.* **105**, 1681–1682.

Garwin, R. L., Lederman, L. M., Weinrich, M. (1957), *Phys. Rev.* **105**, 1415–1417.

Gygax, F. N., Hintermann, A., Rüegg, W., Schenck, A., Studer, W., Van der Wal, A. J. (1984), *J. Less-Common Met.* **101**, 97–113.

Hayano, R. S., Uemura, Y. J., Imazato, J., Nishida, N., Yamazaki, T., Kubo, R. (1979), *Phys. Rev. B* **20**, 850–859.

Hitti, B., Kreitzman, S. R., Estle, T. L., Lichti, R. L., Chow, K. H., Schneider, J. W. (1994), *Hyperfine Int.*, **85–87**, in press.

Holzschuh, E. (1983), *Phys. Rev. B* **27**, 102–111.

Hughes, V. W. (1988), *Phys. Scr.* **T22**, 111–118.

Kadono, R. (1990), *Hyperfine Int.* **64**, 615–633.

Kadono, R. (1992), in: T. Yamazaki, K. Nakai, K. Nagamine (Eds.), *Perspectives of Meson Science*, Amsterdam: North-Holland, pp. 113–116.

Kagan, Yu., Klinger, M. I. (1974), *J. Phys. C* **7**, 2791–2807.

Kagan, Yu., Prokof'ev, N. V. (1990), *Phys. Lett. A* **150**, 320–326.

Kagan, Yu., Prokof'ev, N. V. (1991), *Phys. Lett. A* **159**, 289–294.

Kagan, Yu., Prokof'ev, N. V. (1992), in: A. J. Leggett, Yu. M. Kagan (Eds.), *Modern Problems in Condensed Matter Science*, Amsterdam: North-Holland.

Kehr, K. W., Honig, G., Richter, D. (1978), *Z. Phys. B* **32**, 49–58.

Keller, H., Kiefl, R. F., Baumeler, Hp., Kündig, W., Patterson, B. D., Imazato, K., Nishiyama, K., Nagamine, K., Yamazaki, T., Grynszpan, R. I. (1987), *Phys. Rev. B* **35**, 2008–2014.

Kiefl, R.F., Kreitzman, S. R. (1992), in: T. Yamazaki, K. Nakai, K. Nagamine (Eds.), *Perspectives of Meson Science*, Amsterdam: North-Holland, pp. 265–292.

Kiefl, R. F., Kreitzman, S. R., Celio, M., Keitel, R., Luke, G. M., Brewer, J. H., Noakes, D. R., Percival, P. W., Matsuzaki, T., Nishiyama, K. (1986), *Phys. Rev. B* **34**, 681–684.

Kiefl, R. F., Brewer, J. H., Kreitzman, S. R., Luke, G. M., Riseman, T. M., Estle, T. L., Celio, M., Ansaldo, E. J. (1988), in: G. Ferenczi (Ed.), *Proceedings of the 15th International Conference on Defects in Semiconductors, Budapest, Mater. Sci. Forum* **38–41**.

Kiefl, R. F., Kadono, R., Brewer, J. H., Luke, G. M., Yen, H. K., Celio, M., Ansaldo, E. J. (1989), *Phys. Rev. Lett.* **62**, 792–795.

Kiefl, R. F., *et al.* (1990), *Phys. Rev. Lett.* **64**, 2082–2085.

Kiefl, R. F., *et al.* (1992), *Phys. Rev. Lett.* **69**, 2005–2008.

Kondo, J. (1986), *Hyperfine Int.* **31**, 117–133.

Kondo, J. (1992), in: T. Yamazaki, K. Nakai, K. Nagamine (Eds.), *Perspectives of Meson Science*, Amsterdam: North-Holland, pp. 137–160.

Kreitzman, S. R. (1990), *Hyperfine Int.* **65**, 1055–1070.

Kreitzman, S. R., Brewer, J. H., Harshman, D. R., Keitel, R., Williams, D. Ll., Crowe, K. M., Ansaldo, E. J. (1986), *Phys. Rev. Lett.* **56**, 181–184.

Kreitzman, S. R., Williams, D. Ll., Kaplan, N., Kempton, J. R., Brewer, J. H. (1988), *Phys. Rev. Lett.* **61**, 2890–2893.

Kubo, R., Toyabe, T. (1966), in: R. Blinc (Ed.), *Magnetic Resonance and Relaxation*, Amsterdam: North-Holland, p. 810.

Kuno, Y., Imazato, K., Nishiyama, K., Nagamine, K., Yamazaki, T., Minamisono, T. (1984), *Phys. Lett. B* **148**, 270–284.

Lee, T. D., Yang, C. N. (1956), *Phys. Rev.* **104**, 254–258.

Lee, T. D., Yang, C. N. (1957), *Phys. Rev.* **105**, 1671–1675.

Luke, G. M., Brewer, J. H., Kreitzman, S. R., Noakes, D. R., Celio, M., Kadono, R., Ansaldo, M. (1991), *Phys. Rev. B.* **43**, 3284–3297.

Michel, L. (1950), *Proc. Phys. Soc. London, Sect. A* **63**, 514–531.

Möslang, A., Graf, H., Balzer, G., Recknagel, E., Weidinger, A., Wichert, Th., Grynszpan, R. I. (1983), *Phys. Rev. B* **27**, 2674–2681.

μSR Newsletter (1974), **1**, (15 January).

Niedermayer, Ch., Bernhard, C., Binninger, U., Glückler, H., Tallon, J. L., Ansaldo, E. J., Budnick, J. I. (1993), *Phys. Rev. Lett.* **71**, 1764–1767.

Nishiyama, K. (1992), in: T. Yamazaki, K. Nakai, K. Nagamine (Eds.), *Perspectives of Meson Science*, Amsterdam: North-Holland, pp. 199–218.

Patterson, B. D. (1988), *Rev. Mod. Phys.* **60**, 69–159.

Percival, P. W., Kiefl, R. F., Kreitzman, S. R., Garner, D. M., Cox, S. F. J., Luke, G. M., Brewer, J. H., Nishiyama, K., Venkateswaran, K. (1987), *Chem. Phys. Lett.* **133**, 465–470.

Petzinger, K. G. (1982), *Phys. Rev. B* **26**, 6530–6546.

Primakoff, H. (1975), in: V. W. Hughes, C. S. Wu (Eds.), *Muon Physics*, Vol. 2, *Weak Interactions*, New York: Academic, pp. 3–48.

Rasetti, F. (1944), *Phys. Rev.* **66**, 1–5.

Riseman, T. M., Brewer, J. H. (1990), *Hyperfine Int.* **63**, 249–252.

Roduner, E. (1988), *The Positive Muon as a Probe in Free Radical Chemistry*, Lecture Notes in Chemistry, Vol. 49, Berlin: Springer.

Sachs, A. M., Sirlin, A. (1975), in: V. W. Hughes, C. S. Wu (Eds.), *Muon Physics*, Vol. 2, *Weak Interactions*, New York: Academic, pp. 49–81.

Schneider, J. W., *et al.* (1990), *Phys. Rev. B* **41**, 7254–7257.

Schneider, J. W., *et al.* (1992), *Phys. Rev. Lett.* **68**, 3196–3199.

Seeger, A., Schimmele, L. (1992), in: T. Yamazaki, K. Nakai, K. Nagamine (Eds.), *Perspectives of Meson Science*, Amsterdam: North-Holland, pp. 293–382.

Stamp, P. C. E., Zhang, C. (1991), *Phys. Rev. Lett.* **66**, 1902–1905.

Stoker, D., Balke, B., Carr, J., Gidal, G., Jodido, A., Shinsky, K. A., Steiner, H. M., Strovink, M., Tripp, R. D., Gobbi, B., Oram, C. J. (1985), *Phys. Rev. Lett.* **54**, 1887–1890.

Storchak, V., Brewer, J. H., Hardy, W. N., Johnston, S., Kreitzman, S. R., Morris, G. D. (1994), *Phys. Rev. Lett.* **72**, 3056–3059.

Uemura, Y. J. (1992), in: T. Yamazaki, K. Nakai, K. Nagamine (Eds.), *Perspectives of Meson Science*, Amsterdam: North-Holland, pp. 87–112.

Uemura, Y. J., Harshman, D. R., Senba, M., Ansaldo, E. J., Murani, A. P. (1984), *Phys. Rev. B* **30**, 1606–1608.

Uemura, Y. J., *et al.* (1989), *Phys. Rev. Lett.* **62**, 2317–2320.

Uemura, Y. J., Keren, A., Le, L. P., Luke, G. M., Wu, W. D. (1994), *Hyperfine Int.*, **85–87**, in press.

Winston, R. (1963), *Phys. Rev.* **129**, 2766–2785.

Wu, C. S., Ambler, E., Hayward, R. W., Hoppes, D. D., Hudson, R. P. (1957), *Phys. Rev.* **105**, 1413–1415.

Yamada, K., Sakurai, A., Miyazima, S., Hwang, H. S. (1986), *Prog. Theor. Phys.* **75**, 1030–1043.

Yamanaka, T., Ishikawa, T., Ohtake, S., Iwasaki, M., Akiba, Y., Yamazaki, T., Hayano, R. S. Schnetzer, S. R., Taniguchi, T. (1984), *Hyperfine Int.* **17–19**, 901–906.

Yen, H. K. (1988). M.Sc. thesis, University of British Columbia.

Further Reading

Several excellent reviews of the *applications* of μSR have been written in recent years. The following are specially recommended:

Schenck, A. (1986), *Muon Spin Spectroscopy: Principles and Applications in Solid State Physics*, Bristol, UK: Adam Hilger.

Cox, S. F. J. (1987), *J. Phys, C* **20**, 3187–3319.

The Proceedings of all six International Conferences on μSR to date have been published in *Hyperfine Interactions* as special volumes **6** (1979), **8** (1981), **17–19** (1984), **31** (1986), **63–65** (1990), and **85–87** (1994). These (and any subsequent Proceedings) are the ideal sources for state-of-the-art developments in μSR.

MUONIC, MESONIC, AND BARYONIC ATOMS AND MOLECULES

F. J. HARTMANN, *Physik-Department, Technische Universität München, Garching, Germany*

	Introduction	55
1.	**Energy Levels in Muonic, Mesonic, and Baryonic Atoms**	56
2.	**Formation of Exotic Atoms**	58
3.	**Deexcitation of the Exotic Atom**	59
3.1	Radiative Transitions	59
3.2	Auger Transitions	60
3.3	Stark Mixing	60
3.4	Transfer Processes	61
3.5	Coulomb De-excitation	63
3.6	Chemical De-excitation	63
3.7	The Exotic-Atom Cascade	63
3.8	Exotic-Particle Polarization during the Cascade	63
4.	**Hadronic Atoms: the Strong Interaction**	64
5.	**Experiments on Exotic Atoms**	65
5.1	Intensity of Exotic X Rays	66
5.2	Capture Ratios of Muons in Compounds	68
5.3	Precision Measurements of Transition Energies in Muonic and Pionic Atoms	68
5.4	μ^- Spin Rotation in Semiconductors	69
6.	**Application of Exotic Atoms**	69
6.1	Application of Muonic X Rays for Elemental Analysis	69
6.2	Application of Muonic X Rays in Diagnostics	70
7.	**Muonic Molecules**	70
7.1	Definition	70
7.2	Formation of Muonic Molecules	71
7.3	Energy Levels in Muonic Molecules	72
7.4	Deexcitation of Muonic Molecules	72
7.5	Disappearance of Muonic Molecules: Muon-Induced Fusion	72
7.6	The Cycle of Muon-Catalyzed *dt* Fusion	73
7.7	The Prospects of Energy Production by Muon-Catalyzed *dt* Fusion	74
7.8	A High-Flux Neutron Source Based on Muon-Catalyzed *dt* Fusion	75
	Glossary	75
	Works Cited	76
	Further Reading	77

INTRODUCTION

Muonic, mesonic, and baryonic atoms, usually also called exotic atoms, are atoms in which, besides electrons, an elementary particle of *negative* charge and sufficiently long lifetime is bound in an orbit around the nucleus. Table 1 gives information on the elementary particles for which exotic atoms have been observed. In *muonic atoms*, the exotic particle is a muon; in *mesonic* atoms, it is a meson (pion, kaon); and in *baryonic* atoms, it is a baryon (antiproton, sigma hyperon).

The existence of the respective exotic atoms usually was proven shortly after the discovery of the particle itself. muonic atoms, e.g., were postulated by Fermi *et al.* (1947) to explain observations by Conversi *et al.* (1947) that the disappearance rate of muons strongly depended on the element in which the muon was stopped. Characteristic x rays from pionic atoms were first observed in 1952 by Camac *et al.* (1952). Kaonic atoms were first seen at Argonne National Laboratory in 1965 (Burleson, 1965), and antiprotonic atoms were discovered in 1970 at CERN (Bamberger *et al.*, 1970), but only after a special antiproton beam line had been set up. Nowadays, muonic and

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Table 1. Elementary particles that form exotic atoms (Particle Data Group, 1992). The symbols ν_e and ν_μ denote electron neutrino and muon neutrino, respectively. Antiparticles are marked with a bar.

Particle	Symbol	Mass (MeV/c^2)	Lifetime	Spin	Year of discovery	Main decay modes
Muon	μ^-	105.65839(3)	2.197 μs	1/2	1937	$\mu^- \rightarrow e^- + \bar{\nu}_e + \nu_\mu$
Pion	π^-	139.5679(7)	26.03 ns	0	1952	$\pi^- \rightarrow \mu^- + \bar{\nu}_\mu$
Kaon	K^-	493.646(9)	12.37 ns	0	1947	$K^- \rightarrow \mu^- + \bar{\nu}_\mu$ $\rightarrow \pi^- + \pi^0$
Antiproton	$\bar{p}$	938.2723(3)	Stable	1/2	1955	
Sigma hyperon	Σ^-	1197.43(6)	147.9 ps	1/2	1968	$\Sigma^- \rightarrow n + \pi^-$

pionic atoms may be easily obtained at so-called meson factories, proton accelerators with energies in the 600-MeV to 1-GeV region and a proton current of about 1 mA. Such factories are operated at Paul Scherrer Institut (PSI), Villigen, Switzerland, at the Los Alamos Meson Physics Facility, Los Alamos, New Mexico, U.S.A. (LAMPF), and at TRIUMF, Vancouver, British Columbia, Canada. Antiprotons to form antiprotonic atoms are produced abundantly in a dedicated antiproton facility [Low-Energy Antiproton Ring (LEAR)] at CERN, Geneva, Switzerland.

1. ENERGY LEVELS IN MUONIC, MESONIC, AND BARYONIC ATOMS

The mass of the elementary particle forming an exotic atom is at least 207 times as large as the electron mass; hence, the radii of the exotic-particle orbits in such an atom are much smaller and the binding energies much larger than for an electron in an ordinary atom. Already for rather large excitations, the exotic particle is inside the electron cloud, and the exotic-particle–nucleus system is hydrogenlike, with the electromagnetic interaction between the exotic particle and the nucleus screened only slightly by that part of the electron cloud that is between particle and nucleus. Therefore, all the effects known from hydrogenlike electronic atoms, like fine structure, hyperfine structure, etc., may show up in exotic atoms; they are, however, mostly of completely different importance. As in the hydrogen atom, the bound states in exotic atoms usually are characterized by the quantum numbers n (principal quantum number), l (orbital angular momentum quantum number), and j (total angular momentum quantum number). As a first guess, the binding energy of the exotic particle in an atom with atomic number Z may be calculated from the Bohr formula,

$$E_n = \tfrac{1}{2}\mu c^2 \, [\alpha Z/n]^2, \tag{1}$$

derived long before a quantum-mechanical treatment of the hydrogenlike atom was possible. The radius of the exotic particle's orbit comes out to be

$$r_n = a_0(m_e/\mu)(n^2/Z), \tag{2}$$

while the velocity is given by

$$v_n = v_0 \, Z/n. \tag{3}$$

Here μ, the reduced mass of the muon, meson, or baryon, reads

$$\mu = m_x M/(m_x + M), \tag{4}$$

with m_x and M the exotic-particle mass and the mass of the nucleus, respectively. The use of μ instead of m_x takes the movement of the exotic particle and nucleus around a common center of gravity into account. The electron mass is denoted by m_e, and α is the fine-structure constant ($\alpha = 1/137.036$); $a_0 = 52.92$ pm and $v_0 = \alpha c = 2.188 \times 10^6$ m/s are Bohr radius and velocity, respectively.

As can be easily seen from Eq. (3), the motion of the exotic particle in the lowest levels of heavy exotic atoms is relativistic. Hence, a relativistic wave equation has to be used to describe the interaction of the exotic particle and the nucleus. For particles with half-integer spin (like muons or antiprotons), this is the Dirac equation. The energy eigenvalues (total energy) $E(n,l,j)$ for a Coulomb ($1/r$) potential (valid for a pointlike nucleus) are given by

$$E(n,l,j) = \mu c^2 \left[1 + \left(\frac{Z\alpha}{n - |\kappa| + \sqrt{\kappa^2 - Z^2\alpha^2}} \right)^2 \right]^{-1/2}, \tag{5}$$

with $j = l \pm \frac{1}{2}$ and $|\kappa| = j + \frac{1}{2}$. Hence, for each l (except $l = 0$), there exist two states with different j and different binding energy; the l state has a fine structure.

For spin-0 particles (pions and kaons) the corresponding wave equation is the Klein–Gordon equation, sometimes also called the relativistic Schrödinger equation. Here the energy eigenvalues for a Coulomb potential are given by

$$E(n,l) = \mu c^2 \left[1 + \left(\frac{Z\alpha}{n - l - \frac{1}{2} + \sqrt{(l + \frac{1}{2})^2 - Z^2\alpha^2}} \right)^2 \right]^{-1/2}. \tag{6}$$

In the early 1960s, large discrepancies were found between the binding energies in heavy exotic atoms derived from the first experiments on the characteristic x rays from exotic atoms and the energy values calculated from Eqs. (5) and (6), respectively: In a more realistic picture, the finite size of the nucleus has to be taken into account, as the exotic particle is very near to the nucleus. The nuclear charge distribution is most widely assumed to be a Fermi distribution, with the radial charge density $\rho(r)$ given by

$$\rho(r) = \frac{N}{1 + \exp[4 \ln 3(r - a)/t]}.$$

In this equation, $N \approx (3Z/4\pi)[a^3 + (\pi t/4 \times \ln 3)^2 a]^{-1}$ is determined by the condition

$$4\pi \int_0^\infty \rho(r) r^2 \, dr = Z;$$

a is the half-density radius and t the skin thickness of the nucleus, i.e., the distance between the points where the charge distribution reaches 10 and 90%, respectively, of the maximum value.

There are other effects besides the nuclear size that influence the exotic-particle binding energy:

1. The vacuum polarization, i.e., the generation of virtual e^-e^+ (and $\mu^-\mu^+$) pairs in the Coulomb field of the nucleus, may be explained by quantum electrodynamics (QED) from the fact that even the vacuum has, under certain conditions, the properties of a dielectric medium. Vacuum polarization leads to an increase of the binding energy, which for elements in the lead region reaches about 70 keV.
2. The self-energy due to the interaction of the exotic particle with its own radiation field may also be explained by QED. This effect is less important than in electronic hydrogen.
3. The polarization of the nucleus is brought about by attraction of the protons in the nucleus by the negatively charged exotic particle.
4. The screening of the nuclear charge by that part of the electron cloud that is between exotic particle and nucleus diminishes the nuclear charge seen by the exotic particle and, hence, its binding energy.

As in electronic atoms, also in exotic atoms a nonspherical nuclear-charge distribution leads to an electric *hyperfine* splitting (hfs) of the levels. In most cases, only the interaction of the quadrupole moment of the nucleus with the exotic particle plays a role. The hfs reaches values in the keV region. For muonic Au, e.g., the quadrupole hfs splitting of the $2p_{3/2} \rightarrow 1s_{1/2}$ transition is nearly 20 keV.

Finally, again similar to the electronic atom, in an exotic atom with a spin-$\frac{1}{2}$ particle (like the muon) and a nucleus of nonzero spin, the interaction between nuclear and particle magnetic moment has to be taken into account. The resulting *magnetic* hfs of the levels reaches values up to a few keV, in muonic ^{209}Bi, e.g., roughly 15 keV for the magnetic hfs of the $2p_{1/2} \rightarrow 1s_{1/2}$ transition. Here the finite size of the nucleus (and of its magnetization distribution) leads to a reduction of the magnetic hfs compared to the values for a point nucleus (Bohr–Weisskopf effect).

For hadronic atoms, the strong interaction of the exotic particle with the nucleus results in a measurable shift of the lowest level populated before nuclear capture and also in a level broadening due to the very short lifetime of this level.

Typical level energies and corrections for some muonic atoms are given in Table 2; here the level shifts produced by all the electrodynamic effects just described are not hidden by strong-interaction shifts (cf. Sec. 4).

2. FORMATION OF EXOTIC ATOMS

When an energetic exotic particle enters a target (with a kinetic energy in the MeV region), it is slowed down by interaction with the target electrons. Down to velocities of the order of magnitude of the Bohr velocity (i.e., at energies of around 2.8 keV for muons or 25 keV for antiprotons, to give two examples), the slowing down is rather well described by the Bethe formula (Bethe, 1930). At lower energies, fewer and fewer atomic electrons contribute to stopping because they are too tightly bound to the nucleus to be knocked out of the atom by the exotic particle: So-called shell corrections have to be applied to the Bethe formula. Formation of exotic atoms takes place only after the negative elementary particle has been slowed down to energies around and below 100 eV. It is also commonly accepted that this *Coulomb capture* (called this to distinguish the process from nuclear capture mediated by strong interaction) takes place by the Auger effect: The exotic particle goes over from a free state to a bound orbit around the nucleus, and an electron is emitted that also takes the largest fraction of the energy released during the capture process. It is not so well known from which shell this Auger electron stems and which is the first bound orbit of the exotic particle in the atom. There are, however, indications that this first orbit has a rather large n. *Muonic* atoms with $n \geq 20$ have been detected via the emitted characteristic muonic x rays; furthermore, laser-stimulated de-excitation of a long-lived state with $n = 39$ and $l = 35$ in *antiprotonic* He, identified by the transition energy, established very recently that levels with very large n are populated by Coulomb capture in He.

The probability $p(l)$ for the population of states with angular momentum l by capture was for a long time assumed to follow the relation

$$p(l) \propto 2l + 1,$$

which means that each of the $2l + 1$ l-substates with different magnetic quantum number m ($m = -l, -l + 1, -l + 2, \ldots, l - 2, l - 1, l$) is populated with equal probability. A hand-waving argument for $p(l)$ to be roughly proportional to l is the following: If we assume the exotic particle to hit the atom

Table 2. Electromagnetic binding energies and energy corrections in muonic atoms (Barrett, 1977). The difference between binding energy and point-nucleus energy is mostly due to the finite-size effect. VP: vacuum-polarization correction; EN: corrections for electron screening and nuclear polarization; LS: correction for vacuum polarization by generation of virtual $\mu^-\mu^+$ pairs and for the self-energy of the muon.

Exotic atom	Level nl_j	Binding energy (keV)	Point nucleus energy (keV)	Corrections (eV)		
				VP	EN	LS
μ^7Li	$1s_{1/2}$	24.95	24.92	61	0	0
μ^{12}C	$1s_{1/2}$	100.37	100.38	398	6	−4
μ^{24}Mg	$1s_{1/2}$	397.59	403.97	2213	56	−46
μ^{56}Fe	$1s_{1/2}$	1731.57	1915.29	11 828	553	−420
μ^{56}Fe	$2p_{1/2}$	481.36	479.91	1882	19	3
μ^{56}Fe	$2p_{3/2}$	477.21	475.55	1823	19	−5
μ^{208}Pb	$1s_{1/2}$	10 596.51	20 992.36	67 154	5076	−2950
μ^{208}Pb	$2p_{1/2}$	4813.78	5385.41	32 436	1130	−331
μ^{208}Pb	$2p_{3/2}$	4628.41	4837.25	29 891	1116	−664
μ^{208}Pb	$4f_{5/2}$	1201.19	1197.37	3791	16	13
μ^{208}Pb	$4f_{7/2}$	1191.99	1188.30	3682	18	−9

with velocity v and impact parameter (i.e., distance from the center of the atom) b, it brings in an angular momentum $L = mbv$. As $L \approx \hbar l$ and the probability for the exotic particle to hit the atom at impact parameter b is proportional to the area of a thin ring of inner radius b, $p(l) \propto l$ follows. Deviations from the "statistical" distribution $p(l) \propto 2l + 1$ have been found for muonic atoms, with the distribution rather flat for the transition elements and also for elements in the rare-earth region.

Coulomb capture of exotic particles may be treated either by use of quantum mechanics or in a semiclassical approximation. The complex many-body problem of the exotic-particle interaction with the nucleus and all the atomic electrons has been too complicated up to now to be solved satisfactorily by quantum mechanics, except for the simplest atoms like hydrogen or helium. Semiclassical calculations, although certainly more superficial, have brought more understanding. They are the more adequate as the first bound orbits in exotic atoms have large n, and, hence, as a consequence of the correspondence principle, the results of quantum-mechanical and classical calculations should coincide.

In all semiclassical theories, the following assumptions are made:

1. The exotic particle moves along a classical trajectory in the electric field of the nucleus that is screened by the atomic electrons;
2. it loses energy by a frictional force proportional to its velocity $\mathbf{F} \propto -\mathbf{v}$, which is brought about by the Coulomb interaction of the particle with atomic (and valence) electrons;
3. the particle is assumed to be captured as soon as its total energy falls below zero during the passage through the atom.

The first model of this kind was developed by Fermi and Teller (1947). In their calculations, the probability for capture was simply assumed to be proportional to the rate of energy loss at very low energies in the electron gas of the respective atom. The per-atom probability $P(Z)$ for Coulomb capture came out in this model to be proportional to the atomic number Z, $P(Z) \propto Z$. In more recent calculations, a less pronounced Z dependence was found, with $P(Z)$ even oscillating with Z at periods correlated with those of the periodic table (see, e.g. Daniel, 1980, and Sec. 5.2).

3. DE-EXCITATION OF THE EXOTIC ATOM

The cascade of the exotic particle down the ladder of excited levels proceeds by a number of processes: emission of radiation, Auger effect, Stark mixing, and Coulomb and chemical de-excitation. For exotic hydrogen atoms, transfer to heavier atoms may interfere with the cascade. All processes are now described in more detail.

3.1 Radiative Transitions

In a radiative transition, the exotic atom de-excites, as is already indicated by the name, by emission of quantum radiation. The probability for an electric dipole (E1) transition—the normal transition in exotic as in normal atoms—from level (n_i, l_i) to level (n_f, l_f) is given by a formula derived from that valid for hydrogenlike electronic atoms (Bethe and Salpeter, 1977):

$$\Gamma_{\text{Rad}} = \frac{4}{3}\frac{\alpha c}{a_0}\left[\frac{m_e}{\mu}\right]^2\left[\frac{\Delta E}{\alpha m_e c^2}\right]^3 \frac{\max(l_i, l_f)}{2l_i + 1} I^2_{i\to f}. \tag{7}$$

Here $I_{i\to f}$ is the matrix element

$$I_{i\to f} = \int_0^\infty R^*(n_i, l_i) r^3 R(n_f, l_f)\, dr, \tag{8}$$

with $R(n, l)$ the normalized wave functions for the hydrogenlike exotic atom and ΔE the transition energy. Only transitions with $|\Delta l| = |l_i - l_f| = 1$ are allowed. Because of the ΔE^3 factor, transitions with maximum possible change $\Delta n = n_i - n_f$ are favored, although $I_{i\to f}$ becomes smaller with increasing Δn. Radiative transitions constitute the main de-excitation mechanism for the late stage of the exotic-atom cascade. Besides E1 transitions, also E2 (electric quadrupole) transitions have been observed, but this only in the Lyman series, i.e., for transitions to the 1s level, in heavy atoms.

3.2 Auger Transitions

In an Auger transition, the exotic atom de-excites by emission of an atomic electron, either from the host atom itself (*internal* Auger effect) or from an atom the exotic atom collides with (*external* Auger effect). The expressions for the rate Γ_{Auger} for the internal Auger effect are rather complicated (for details, see Akylas and Vogel, 1978). A formula given by Ferrell (1960) connects Γ_{Auger} with the rate Γ_{Rad} for a radiative dipole transition between the same levels:

$$\frac{\Gamma_{\text{Auger}}}{\Gamma_{\text{Rad}}} = \frac{\sigma^{Z-1}(\Delta E)}{(Z-1)^2\sigma_{\text{T}}},$$

and may serve for a fast calculation of Γ_{Auger}. Here $\sigma^{Z-1}(\Delta E)$ is the cross section for the photoeffect at energy ΔE in the element with atomic number $Z - 1$, and σ_{T} denotes the Thomson cross section, i.e., the classical cross section for scattering of an electromagnetic wave on a free electron, $\sigma_{\text{T}} = 8\pi r_e^2/3 = 0.665 \times 10^{-28}$ m^2 with $r_e = 2.818$ fm the classical electron radius.

Besides dipole transitions, monopole and quadrupole transitions play a role, however minor; changes in the angular momentum with $|\Delta l| \leq 2$ are possible. The Auger effect favors low-energy transitions, especially those with small Δn in the first stages of the exotic-atom cascade.

Figure 1 shows rates for radiative and Auger transitions between circular orbits for initial levels with $n \leq 6$. The rates are orders of magnitude larger than those for hydrogen-like electronic atoms.

The *external* Auger effect is most important for exotic hydrogen, as there the only electron has been ejected from the host atom during Coulomb capture; hence, the internal Auger effect can no longer take place.

3.3 Stark Mixing

Stark mixing is important for the cascade in exotic hydrogen and helium atoms because these systems are small and neutral or only singly charged and can closely approach other atoms in the target. Inside the electron cloud, they feel a strong electric field $\mathbf{E}$ from the nucleus that destroys the spherical symmetry of the exotic atom, as is well known from the Stark effect in electronic atoms. A dipole moment $\boldsymbol{\mu}$ is induced in the exotic atom by the electric field; by interaction of the field with this dipole moment, different angular-momentum states are mixed, as the angular momentum is no longer conserved; still conserved, however, is its projection in the direction of the electric field (characterized by the magnetic quantum number m).

The rate of Stark transitions from a state with quantum numbers n, l, and m to a state with quantum numbers n, l', and m' is proportional to the matrix element $M_{l,l'}$ for the interaction energy $\boldsymbol{\mu} \cdot \mathbf{E}$. A calculation of this

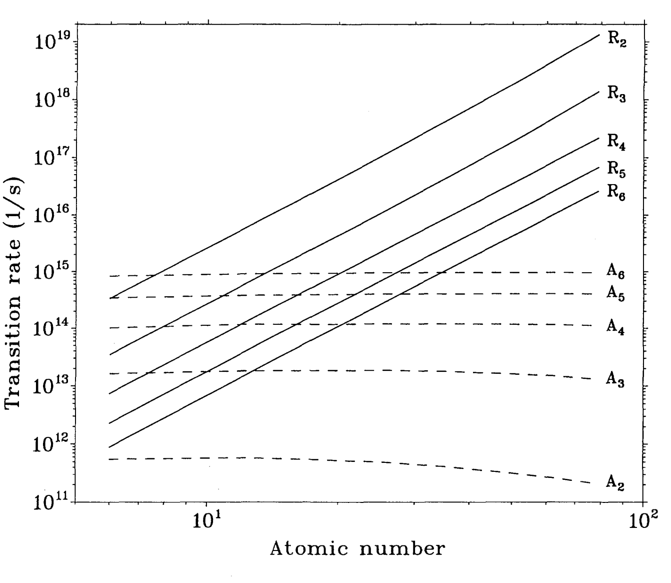

FIG. 1. Radiative (R_2 to R_6) and Auger (A_2 to A_6) rates in muonic atoms for the transitions (n, $l = n - 1$) → ($n - 1$, $l = n - 2$), $n = 2$ to 6, as a function of the atomic number.

matrix element with hydrogenlike wave functions leads to nonzero results only for $m = m'$ and $l = l' \pm 1$, with values given, e.g., by

$$M_{l,l+1} = \frac{3}{2} a_0 \frac{m_e}{\mu} e|\mathbf{E}|n \sqrt{\frac{(n^2 - l^2)(l^2 - m^2)}{4l^2 - 1}}.$$

To estimate the influence of the Stark mixing for exotic hydrogen atoms in hydrogen (certainly the most important case), one may approximate the electric field that the exotic atom feels during the collision with a hydrogen atom by a constant field that the atom experiences at the site of the Bohr orbit, $|\mathbf{E}| \approx e/4\pi\epsilon_0 a_0^2 = 5.14 \times 10^{11}$ V/m. For $\mu^- p$ and $\pi^- p$ atoms, e.g., the transition rate then becomes $\omega_{\text{Stark}} = M_{l,l'}/\hbar \approx 2 \times 10^{14} n\sqrt{n^2 - 1}\ \text{s}^{-1}$ if we take $l = 1$ and $m = 0$. The characteristic duration τ_{Stark} of a collision is roughly $\tau_{\text{Stark}} = a_0/v$, with a_0 standing for the dimension of the atom and v the exotic-atom velocity; thus, at a μp energy of 1 eV, $\tau_{\text{Stark}} \approx 5 \times 10^{-14}$ s and $\omega_{\text{Stark}}\tau_{\text{Stark}} \gg 1$ follows: There will be many Stark transitions between different l states with the same n during one collision (complete Stark mixing), and the angular-momentum states will be completely mixed already after one collision. The exotic-hydrogen atom will experience, however, such a large electric field only for central collisions with impact parameters smaller than a radius ρ_{lim} that is of the order of magnitude of the radius of the hydrogen atom.

Figure 2 compares radiative, Auger, and Stark transitions in pionic hydrogen for a gaseous target with 0.001 of the density of liquid hydrogen.

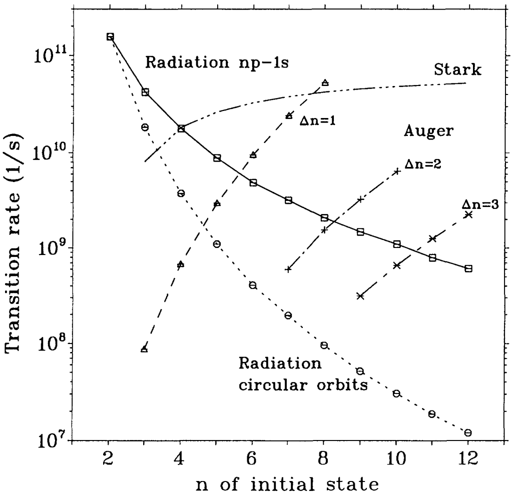

FIG. 2. Rates for radiative, Auger, and Stark transitions in pionic hydrogen for an atomic density that is one thousandth of the density of liquid hydrogen (Markushin, 1990).

3.4 Transfer Processes

Transfer is the reaction

$$X^- Z_0 + Z \rightarrow Z_0 + X^- Z,$$

with X^- an exotic particle. It plays a role mainly for exotic hydrogen atoms ($Z_0 = 1$), as here the *neutral* $X^- h$ system (where h stands for a proton, deuteron, or triton, respectively) can penetrate the electron shell of the atom Z (in the following, only this type of transfer will be discussed). We can simplify the problem if we assume the $X^- h$ atom to be in the ground ($1s$) state already and to have traversed the electron cloud of atom Z. The question then is the following: What is the probability for a transfer of X^- from the $1s$ orbit of the $X^- h$ atom to a bound orbit in $X^- Z$ while the nuclei h and Z come near and separate again? As has been shown, the probability for such a transfer becomes especially large if there are so-called crossings in a diagram which displays the energy of the systems ($X^- h(1s) + Z$) and ($h + X^- Z$), respectively, as a function of the distance R between h and the nucleus with charge Z. Such a diagram, also called correlation diagram and well known in molecular-orbital theory, is shown in Fig. 3 for the muonic proton-fluorine system. At the crossings, the energies of these two systems are equal.

The energy $\epsilon_{1s}^{(h)}$ for one limiting case, $X^- h$ in the ground state and the atom Z a distance R apart (point A in Fig. 3), is given approximately by

$$\epsilon_{1s}^{(h)} \approx -\frac{m_p}{1 + m_p}\left(\frac{1}{2} + \frac{9Z^2}{4R^4}\right),$$

with ϵ, R, and the proton mass m_p given in muonic Rydberg units, $R_{h_\mu} = 5.62$ keV, muonic Bohr units, $a_\mu = 256$ fm, and units of the muon mass, respectively. The second term in this equation is a (weak) van der Waals shift. In the second limiting case, X^- in an orbit with

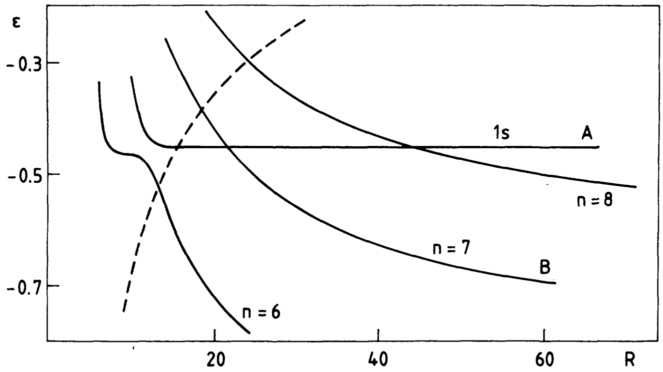

FIG. 3. Correlation diagram for the system $\mu + p +$ fluorine (Holzwarth and Pfeiffer, 1975). Radius R and total energy ϵ in muoatomic units (unit radius 256 fm, unit energy 5.6 keV). 1*s*: energy of the 1*s* level in the μp system; $n = 6$, $n = 7$; $n = 8$: energy of the levels with the respective n in the muonic fluorine atom.

principal quantum number n around nucleus Z and h some distance away (point B in Fig. 3, as an example), the energy ϵ_n^Z is given approximately by

$$\epsilon_n^Z \approx -\frac{m_Z}{1 + m_Z}\frac{Z^2}{2n^2} + \frac{Z - 1}{R},$$

with the same units and m_Z again given in muon-mass units. Here the second term corresponds to a Coulomb repulsion of the exotic atom (charge $Z - 1$) by the proton. Naturally, the two nuclei h and Z cannot come closer than a distance determined by the combined repulsion between h and Z and attraction between X^- and h and Z, respectively. It is shown in Fig. 3 as dashed line. Two crossings can be seen in the allowed R region, one between $\mu p(1s)$ and $\mu\mathrm{F}(n = 7)$ at $r_c \approx 21$ muoatomic units (or 5.4 pm) and one between $\mu p(1s)$ and $\mu\mathrm{F}(n = 8)$ at $r_c \approx 42$ muoatomic units (or 10.7 pm). A large rate is expected for the muon transfer from p to F, especially to levels with $n = 7$ and $n = 8$, and has been found by experiment.

The rate for transfer via crossing states, λ_{tr}, was recently recalculated (Sayasov, 1990) to be (for liquid-hydrogen density)

$$\lambda_{\mathrm{tr}} = 2.6 \times 10^{10}\,(Z/\sqrt{M_p})\,F(\xi)\,\mathrm{s}^{-1},$$

with $\xi = (3\sqrt{2M_p}Z/8r_c)^{2/3}$, r_c the separation of the nuclei at the crossing point, and M_p the mass of the exotic hydrogen atom normalized to the μp mass. The term $F(\xi)$ is shown in Fig. 4 as a function of ξ, together with experimental $F(\xi)$ values derived from measured transfer rates. The model succeeds in explaining the large rate difference between muon transfer from p to Ne ($\lambda_{\mathrm{tr}} = 5.6 \times 10^9$ s^{-1}) and that from d to Ne ($\lambda_{\mathrm{tr}} = 1.4 \times 10^{11}$ s^{-1}), both at liquid-hydrogen density.

Transfer plays an especially prominent role in hydrogen *mixtures* where it is a step on the way from the exotic atom to exotic molecules (see Sec. 7). For energy reasons, transfer is only possible from the lighter to the heavier hydrogen isotope: Muons in the heavier muonic atom are, because of their larger reduced mass, more strongly bound.

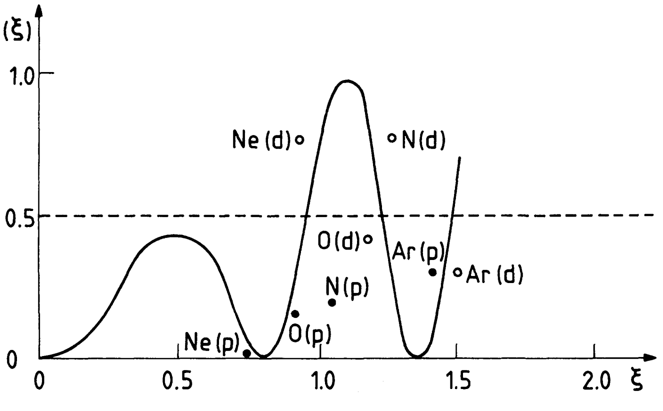

FIG. 4. Normalized transfer rate $F(\xi) = \lambda_{\mathrm{tr}}/[2.6 \times 10^{10} Z M_p^{-1/2}]$ as a function of ξ (Gershtein, 1993). Solid line: calculations. Circles and dots: experimental results. For further explanations, see text.

3.5 Coulomb De-excitation

Coulomb de-excitation is a transition of the kind

$$(X^-p)_n + p \rightarrow (X^-p)_{n-k} + p'.$$

It is important only for the exotic hydrogen atom cascade because only the *neutral* exotic hydrogen atom may approach other atoms (mostly those with higher Z) close enough to make fast de-excitation possible. The mechanism is very similar to that for transfer, as also here the transition proceeds especially fast if crossings exist in the correlation diagram between the initial and the final levels of the exotic hydrogen atom.

3.6 Chemical De-excitation

At the early stage of the exotic-*hydrogen* atom cascade, chemical de-excitation, i.e., the de-excitation of the exotic atom by excitation of rotations and vibrations or by dissociation of hydrogen molecules, is possible.

3.7 The Exotic-Atom Cascade

The exotic-atom cascade shall now, as an example, be described in more detail for a muonic atom of medium Z, e.g., μ^-Ar. After capture into a highly excited orbit, the muon particle cascades down the ladder of excited states in the exotic atom. The only de-excitation mechanism in this stage is the Auger effect. By a series of successive Auger transitions, the exotic atom becomes highly ionized. Replacement of the missing electrons proceeds by electron capture into the multiply charged $\mu^-\mathrm{Ar}^{m+}$ ion in collisions with other Ar atoms in the gas. The collision rate depends strongly on the gas pressure and is, at low pressure at least, too small to keep up with the ionization of the exotic atom by the Auger effect. Thus, the μ^-Ar atom finally may become completely stripped of electrons. From this moment on, the cascade proceeds only by radiative transitions down to the $\mu^-\mathrm{Ar}(1s)$ level. It should be noted in parentheses that, for metals and even for insulators, replacement of electrons is fast enough to let the muonic atom be only weakly deprived of electrons during the cascade: The ionized exotic atom distorts the solid-state potential and strongly attracts electrons from the vicinity. As a consequence, radiative transitions become dominant in this case only when the muon has reached levels with $n \leq 10$ (depending on the atomic number of the host atom; see Fig. 1).

The exotic-atom cascade usually is finished within picoseconds. After reaching the ground state of the muonic atom, the muon either decays or reacts with a proton of the nucleus by weak interaction,

$$p + \mu^- \rightarrow n + \nu_\mu. \tag{9}$$

The muon accordingly disappears from the $1s$ level with a rate that is the sum of the muon-decay rate and the nuclear-capture rate. The probability for nuclear absorption, proportional to first order to the sum of the squared wave function at the sites of the protons in the nucleus,

$$w_{\text{abs}} \propto \sum_{\text{all protons}} |\psi(r_{\text{proton}})|^2,$$

should follow a Z^4 law because there are Z protons in the nucleus, and $|\psi(r_{\text{proton}})|^2$ should be $\approx |\psi(0)|^2 \propto Z^3$ for a hydrogenlike atom with a point nucleus of charge Z. Saturation effects are expected for high Z because the large neutron excess in heavy elements limits the number of empty states in which the neutron formed by reaction (9) can be accommodated. The lifetime of the μ^- in the $\mu^-Z(1s)$ state decreases from 2.2 μs in μ^-H to 200 ns in μ^-Fe ($Z = 26$) and 75 ns in μ^-Pb ($Z = 82$).

Exotic He is a very special case. Here, as has been verified for pionic and kaonic He and especially for antiprotonic He (Yamazaki *et al.*, 1993), a small fraction ($\approx$3%) of the exotic He atoms are formed in long-lived states: The exotic atom can deexcite from circular orbits (i.e., those with $l = n - 1$) or nearly circular orbits only by the rather slow radiative transitions because Auger transitions with small angular-momentum transfer are forbidden for energy reasons, and Stark mixing is strongly suppressed for the $(\alpha X^- e^-)$ system, as a result of shielding by the remaining electron.

3.8 Exotic-Particle Polarization during the Cascade

One quantity with some relevance for applications [e.g., μ^- spin rotation (μ^-SR) experiments] has been left out of consideration up to now: the polarization of the exotic particle during slowing down, capture, and cascade to the 1*s* level in the exotic atom. Polarization is the correlation between the spin of a particle and a selected direction, e.g., the direction of flight; a particle is said to be completely polarized if its spin is parallel or antiparallel to this direction. Polarization clearly is most relevant for muons, as only they have a nonzero spin *and* survive the cascade down to the ground state of the exotic atom with very high probability.

Negative muons are usually produced in meson factories by the two-body decay of π^- (spin zero), $\pi^- \rightarrow \mu^- + \bar{\nu}_\mu$. In the pion rest frame, the μ^- spin is parallel to its momentum, as the $\bar{\nu}_\mu$ is completely polarized, the two decay particles fly in opposite directions, and the total spin after decay must be zero. In the laboratory frame, the polarization is no longer 100% because the center-of-mass velocity changes the direction of the muon but not its spin; polarization may be still rather high (around 70%), however. This means that the spin of the muons in the beam preferentially points into the direction of flight. During slowing down, the direction of the muon spin is not changed because the forces leading to the moderation do not act on the spin. The polarization after Coulomb capture has been shown (Mann and Rose, 1961) to be roughly $\pm\frac{1}{3}$ of the initial polarization for muon capture into the states with $j = l \pm \frac{1}{2}$ and for large l. Strong depolarization may occur during the muonic cascade, certainly not by dipole transitions (radiative or Auger), as they do not act on the muon spin, but by interaction between exotic-particle spin and angular momentum. As a result, the muons starting from the state $(n, l, j = l + \frac{1}{2})$ will preserve $\frac{1}{3}$ of their initial polarization if hyperfine interactions of the muon with the (possibly) nonzero spin of the nucleus do not lead to additional depolarization. Those muons starting from $(n, l, j = l - \frac{1}{2})$ will reach the 1*s* state completely depolarized, however.

Polarization of the muon becomes evident, through parity nonconservation in μ decay, by the asymmetry in the angular distribution $W(\theta)$ of the decay electrons with respect to the particle spin,

$$W(\theta) = 1 - a\cos(\theta). \tag{10}$$

In this formula, a is an energy-dependent factor, which reaches 1 for maximum decay-electron energy (roughly 50 MeV).

4. HADRONIC ATOMS: THE STRONG INTERACTION

In all exotic atoms except those with muons, the exotic particle disappears by strong interaction with the nucleus—in all but the lightest atoms, before it has reached the 1*s* level. In order to describe the interaction, a *complex* strong-interaction potential has to be added to that of the electromagnetic interaction, complex because it has to describe the absorption of the hadron. It may be generated by superimposing the various hadron-nucleon potentials, taking the influence of the nuclear medium into account. If the hadron-nucleon interaction is not too strong, the potential can be assumed to be proportional to the product of an effective scattering length $\bar{a}$, which characterizes the strength of the strong interaction of the hadron with a single nucleon, and the nucleon density $\rho(\mathbf{r})$, the sum of proton and neutron densities, $\rho(\mathbf{r}) = \rho_p(\mathbf{r}) + \rho_n(\mathbf{r})$:

$$V_{\text{Strong}} = 4\pi/\mu[\bar{a}\rho(\mathbf{r})].$$

This *Ansatz* is usually used in antiprotonic atoms.

It has turned out that such a simple description is inadequate for the pion-nucleus potential. Rather elaborate potentials resulted, including not only local terms, i.e., those depending on **r**, but also nonlocal terms, i.e., those containing the gradient of **r**, $\nabla\mathbf{r}$. A widely used parametrization of the pion-nucleus potential is the Ericson-Ericson potential (Ericson and Ericson, 1966) given by

$$\begin{aligned} -(\mu/2\pi)V_{\text{Strong}}(\mathbf{r}) = {} & (1 + \mu/M) \\ & \times [b_0\rho(\mathbf{r}) + b_1\delta\rho(\mathbf{r})] \\ & + (1 + \mu/2M)B_0\rho^2(\mathbf{r}) \\ & - \boldsymbol{\nabla}\frac{\alpha(\mathbf{r})}{[1-\frac{4}{3}\pi\zeta\alpha(\mathbf{r})]}\boldsymbol{\nabla}, \end{aligned}$$

with

$$\alpha(\mathbf{r}) = \frac{c_0\rho(\mathbf{r}) + c_1\delta\rho(\mathbf{r})}{1 + \mu/M} + \frac{C_0\rho^2(\mathbf{r})}{1 + \mu/2M}.$$

Here $\delta\rho(\mathbf{r}) = \rho_n(\mathbf{r}) - \rho_p(\mathbf{r})$, μ and M are the reduced hadron mass and the nucleon mass, respectively, b_0, b_1, c_0, and c_1 are adjustable complex parameters standing for the different possible interaction modes, and the terms with B_0 and C_0 represent pion absorption on two nucleons [for this reason, they are quadratic in $\rho(\mathbf{r})$]. The parameter ζ, finally, characterizes the strength of short-range correlations between the scattering nucleons in the term $1 - \frac{4}{3}\pi\zeta\alpha(\mathbf{r})$, which is the analog to the term found when calculating the susceptibility in dense polarizable media (Lorentz term). Even more elaborate potentials have been recently used (Konijn *et al.*, 1990).

As pointed out before, strong interaction results in an appreciable shift and width of the last level populated before absorption. But already the preceding x-ray transitions are attenuated by a nonzero probability for hadron absorption. Table 3 shows examples of shifts and widths of the lowest levels accessible in pionic and antiprotonic atoms.

Absorption of the pion by the nucleus (with an energy release ΔE approximately equal to the rest energy of the pion, $\Delta E \approx 139$ MeV) results in the evaporation of a number of light particles (neutrons, protons, deuterons, tritons). In antiproton annihilation, the energy release is more than an order of magnitude larger ($\Delta E \approx 1876$ MeV, twice the rest energy of the $\bar{p}$). Here part of the mean number of five pions (neutral and charged) generated during annihilation on *one* nucleon miss the nucleus, as the annihilation takes place at its outer periphery. Nevertheless, in the mean, three pions hit the nucleus, which, because of a strong resonance peak in the pion-nucleon cross section (the so-called Δ_{33} resonance), is black for these pions. The nucleus is heated up by the energy deposited by the pions, and cools down by the emission of fast light particles and later by neutron evaporation. For a medium-mass nucleus like Mo ($A \approx 100$), up to 30 nucleons are emitted. The energy spectra of protons, deuterons, tritons, and α particles emitted after $\bar{p}$ annihilation in ^{238}U are shown in Fig. 5.

Table 3. Typical strong-interaction shifts and widths in pionic and antiprotonic atoms (see Batty *et al.*, 1991, Heitlinger *et al.*, 1992, and Konijn *et al.*, 1990, for references). The shift is defined as the difference between measured binding energies and those calculated taking only the electromagnetic interaction into account. A negative energy shift corresponds to a repulsive strong-interaction potential.

Exotic atom	Level (n, l)	Strong interaction Shift ϵ (keV)	Width Γ (keV)
π-^{3}He	1*s*	0.032(3)	0.028(7)
π-^{4}He	1*s*	−0.076(2)	0.045(21)
π-^{10}B	1*s*	−3.25(8)	1.58(16)
π-^{11}B	1*s*	−4.17(9)	1.77(16)
π-^{16}O	1*s*	−18.7(4)	7.9(3)
π-^{18}O	1*s*	−23.3(3)	6.3(4)
π-^{24}Mg	1*s*	−80.5(12)	24.3(16)
π-^{24}Mg	2*p*	0.125(4)	0.0725(18)
π-^{40}Ca	2*p*	1.77(8)	1.64(11)
π-^{93}Nb	2*p*	−11(3)	64(8)
π-^{93}Nb	3*d*	0.74(2)	0.402(16)
π-^{208}Pb	3*d*	22.7(24)	47(4)
π-^{208}Pb	4*f*	1.68(2)	0.98(5)
$\bar{p}$-^{1}H	2*p*	−0.730(20)	1.12(6)
$\bar{p}$-^{16}O	3*d*	−0.112(20)	0.495(45)
$\bar{p}$-^{17}O	3*d*	−0.140(46)	0.54(15)
$\bar{p}$-^{18}O	3*d*	−0.195(20)	0.64(4)
$\bar{p}$-^{23}Na	3*d*	−2.08(30)	2.9(11)

5. EXPERIMENTS ON EXOTIC ATOMS

Since the early days of exotic-atom research, investigators have mainly observed the properties of radiation and particles

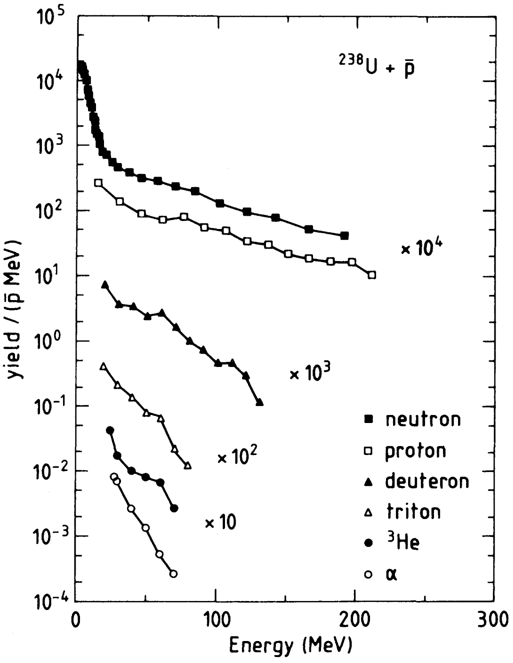

FIG. 5. Energy spectrum of annihilation products from antiproton annihilation in $\bar{p}$-^{238}U (Machner, 1993).

emitted during the atomic cascade and the decay or annihilation of the exotic particle:

1. energy, intensity, anisotropy, and polarization of x rays from exotic atoms (from now on, called exotic x rays),
2. energy and intensity of electrons emitted in the Auger effect,
3. emission-time delay, intensity, and anisotropy of electrons from μ decay in muonic atoms,
4. emission-time delay, energy, and multiplicity of annihilation products from hadronic atoms.

To collect such information, the following detectors have been employed:

1. scintillation counters with NaI(Tl), bismuth germanate (BGO), BaF, and CsI crystals for the detection of high-energy quantum radiation;
2. lithium-drifted Si and Ge detectors and hyperpure-Ge detectors mainly for the measurement of energy and intensity of medium- and low-energy quantum radiation;
3. crystal spectrometers, mainly for the precise determination of the energy of low-energy quantum radiation;
4. semiconductor spectrometers for the detection of electrons from exotic-particle Auger effect;
5. semiconductor and magnet spectrometers for the measurement of multiplicity and energy of charged particles generated by and following exotic-particle absorption by the nucleus;
6. scintillation counters with liquid organic scintillators for the detection of neutrons from exotic-particle annihilation and with plastic scintillators for the detection of electrons from muon decay.

Only in a very few cases has it been attempted to stimulate excitation or deexcitation of the exotic atom by laser irradiation because, on account of the restricted intensity of the exotic-particle beams, the experiments are tedious, and long measuring times are necessary. On the other hand, the unprecedented accuracy of the laser wavelength determination makes it possible to measure level-energy differences in exotic atoms with uncertainties orders of magnitude smaller than is possible with detectors; thus, such experiments are very valuable for the precise determination of the tiniest QED corrections in the level energies of exotic atoms. Examples of experiments aimed at the excitation or deexcitation of exotic atoms by laser irradiation are

1. the stimulation of the muonic ($2s \rightarrow 2p$) transition in muonic helium by laser irradiation,
2. laser-stimulated ($3d \rightarrow 3p$) transitions in muonic hydrogen,
3. the stimulated deexcitation of long-lived, high-lying states ($n \approx 38$) in antiprotonic helium, also by irradiation with a laser.

Representing the large number of experiments on muonic, mesonic, and baryonic atoms in the past decades, some examples will be described now.

5.1 Intensity of Exotic X Rays

A typical experimental arrangement for the detection of exotic x rays is shown in Fig. 6. A telescope with scintillation counters S_1, S_2, S_3, and S_4 is set up to detect the particles coming down the beam line. Coincidence of signals from S_1, S_2, and S_3 with no signal from S_4 indicates a stop of the particle in the target. With a moderator of adjustable thickness, the exotic particles are slowed down to stop in the target. The Ge detectors serve to measure the muonic x rays emitted within a very short ($\approx$50 ns) time window after the stop. A typical x-ray spectrum, in this case the Lyman series of muonic Fe, is shown in Fig. 7. Transitions from $l = 1$ levels with n up to $n = 20$ to the $1s$ level can be separated, but transitions from even higher levels can be seen.

To get an idea about the de-excitation of the exotic atom, the cascade of transitions is simulated with competition between radiative and Auger transitions, depletion of the electron shell by Auger transitions, and electron replacement taken into account. For hadronic atoms, nuclear absorption has to be taken into consideration too. One usually does *not* start the simulations at the highly excited states populated at atomic capture, because the transition rates are poorly known there. A good choice for the beginning of the simulation is a state with an orbit similar in size to the *K*-electron orbit because, from then

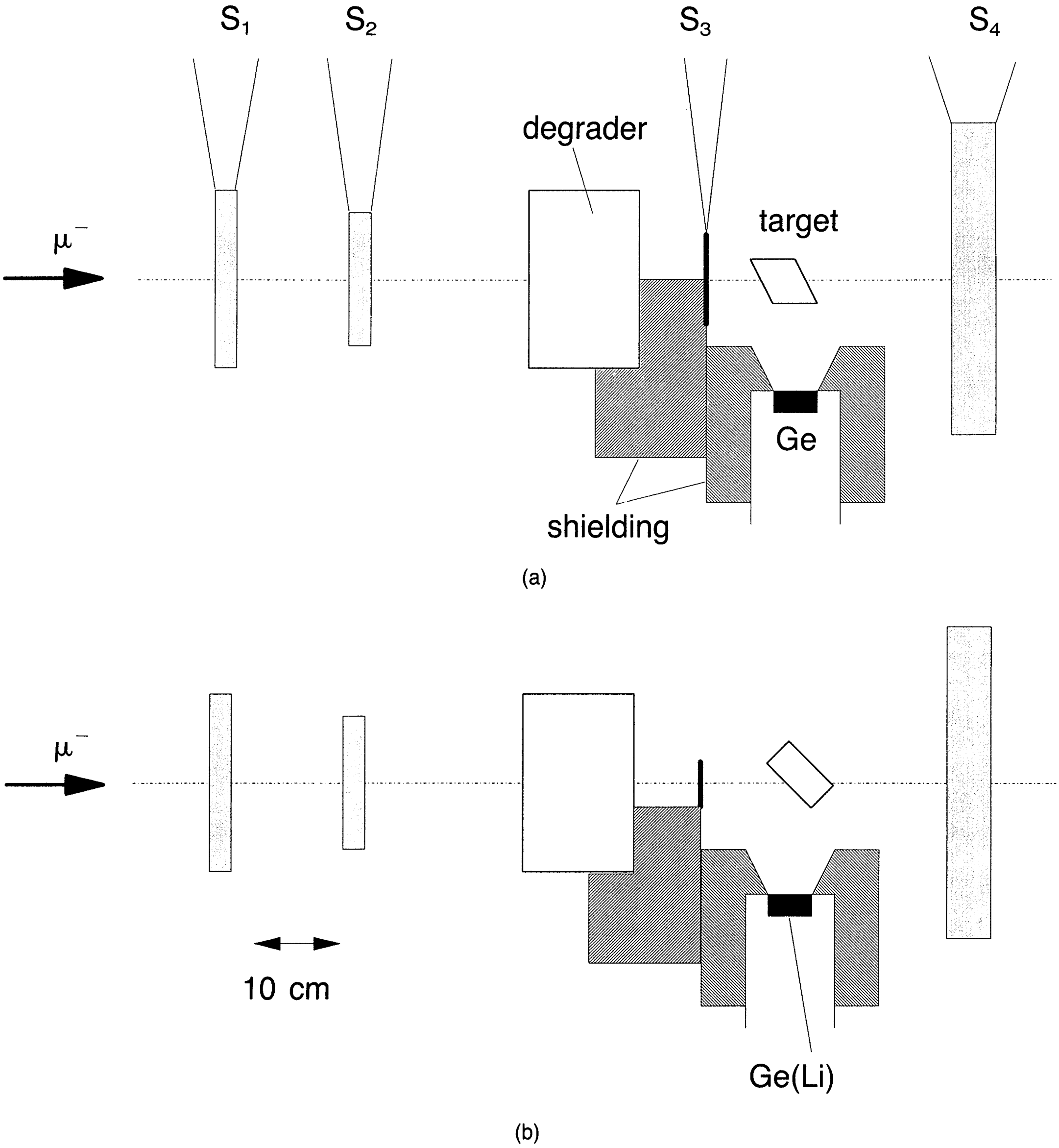

FIG. 6. Typical experimental arrangement for the measurement of muonic x-ray intensities. Upper half: side view: lower half: top view. S_1, S_2, S_3, S_4: scintillation counters; intrinsic Ge: hyperpure-Ge detector; Ge(Li): Li-drifted Ge detector.

on, the problem is hydrogenlike and the calculation of rates straightforward. In general, at the beginning of the simulation, angular-momentum states at only *one n* are assumed to be populated, with a probability distribution characterized by one or two adjustable parameters. The exotic x-ray intensities measured may then be compared with the results of the cascade calculations, and the parameters of the calculation, characterizing the initial distribution and depletion and replacement of the electrons, may be adjusted to get best agreement between simulated and measured x-ray intensities. More than 80 such intensities were recorded for muonic Fe, to give an example.

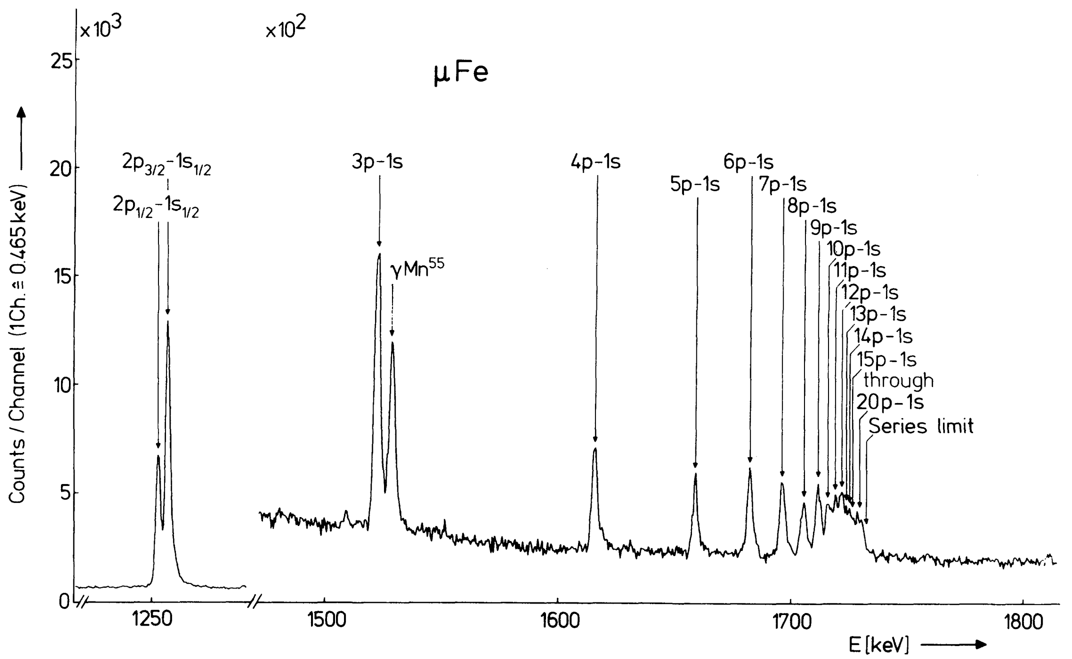

FIG. 7. Muonic Lyman series from muonic Fe (Hartmann *et al.*, 1976). Transitions with $n > 20$ are clearly visible.

5.2 Capture Ratios of Muons in Compounds

In muonic atoms, the sum of the Lyman-series x-ray intensities from a certain element is equal to the number of muons captured by the atoms of this element. This is because, except for atoms with $Z < 6$ or Z very large, each muon captured into the atom makes one and only one *radiative* transition to the 1*s* level of the muonic atom. In atoms with $Z < 6$, the muons reaching the 2*s* levels are trapped there because transitions to the 2*p* level are forbidden for energy reasons, and those to the 1*s* level are monopole transitions and, hence, strongly suppressed; for very heavy atoms, on the other hand, the transition energies in the final stage of the cascade become high enough to make de-excitation by neutron emission from the nucleus possible and probable.

Relative per-atom capture probabilities have been derived from Lyman-series x-ray intensities for 150 different compounds, alloys, and mixtures. From this large data set, it was possible to extract the per-atom capture probabilities for 65 elements (normalized to $Z = 8$). They are shown in Fig. 8. A pronounced periodicity of $P(Z)$ with Z is visible (cf. Sec. 2).

The chemical bond was shown to have an influence on the capture probability; this "chemical effect," however, changes $P(Z)$ by not more than 5%.

5.3 Precision Measurements of Transition Energies in Muonic and Pionic Atoms

Precision measurements of transition energies in exotic atoms serve to determine fundamental constants or quantities important for the understanding of properties of nuclei and of elementary particles.

One example is the measurement of the $3p \rightarrow 1s$ transition energy in pionic hydrogen with a crystal spectrometer. Such a measurement yields the strong-interaction shift ϵ_{1s} and width Γ_{1s} of the 1*s* level in pionic hydrogen and serves to deduce the very fundamental pion-nucleon strong-interaction parameters. A double-focusing crystal spectrometer with Si crystals cut in the (111) direction was used. Position-sensitive CCD (charged-coupled device) x-ray detectors served to detect the scattered x rays. The energy of the $3p \rightarrow 1s$ transition was measured

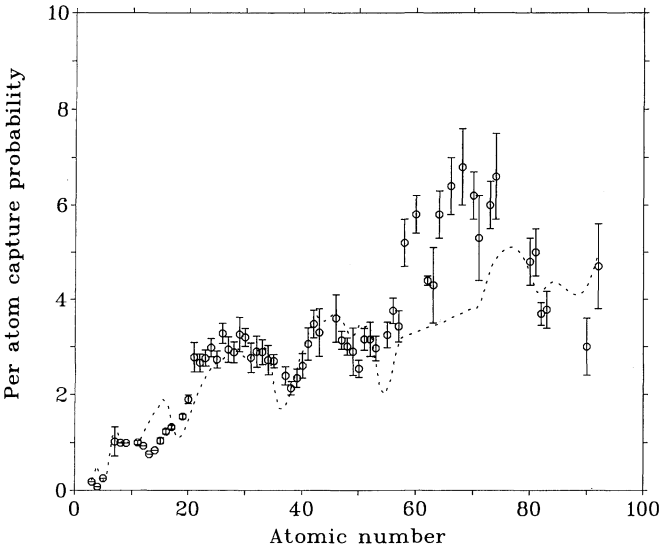

FIG. 8. Per-atom capture probability $P(Z)$ as a function of Z. Dashed line: result from a semiclassical calculation taking the atomic radii into account (Daniel, 1979).

with a relative error of only 100 ppm. Hence, the strong-interaction shift could be determined to better than 5%.

Similar measurements have been performed to determine the energy of the $4f \rightarrow 3d$ transition in pionic ^{24}Mg to derive the rest mass of the pion with high accuracy.

Numerous experiments with semiconductor detectors have been devised over the decades to measure muonic x-ray energies precisely, preferably from the Lyman series, in order to determine the finite-size energy corrections (cf. Sec. 1). From these corrections, nuclear-size parameters (radius, skin thickness, and deformation if applicable) could be deduced by comparing the energy eigenvalues of the Dirac equation for the potential of the extended nucleus with experiment. This method complements the method of high-energy electron scattering on nuclei.

5.4 μ^- Spin Rotation in Semiconductors

As outlined in Sec. 3.7, the spin polarization of the muon in the 1*s* level of a muonic atom is roughly $\frac{1}{6}$ of the initial polarization of the muon in the particle beam, but may strongly depend on hyperfine interactions between the muon and the electric and magnetic field at the site of the muonic atom. The muonic atom acts, so to say, as an extended nucleus with charge $Z - 1$ that feels the electronic structure of the surrounding medium. Thus μ spin rotation (μ^-SR) has become a tool in solid-state physics, although not nearly as extensively used as spin rotation with positive muons (μ^+SR) (see MUON SPIN ROTATION/RELAXATION/RESONANCE).

One such experiment performed only recently was μ^-SR in Si, pure and heavily doped with boron, respectively. An asymmetry of decay-electron emission as an indicator for the muon polarization (see Sec. 3.8) was visible at low temperatures and increased strongly with increasing boron-doping level. The result was interpreted as follows: The $2p \rightarrow 1s$ transition energy of 400 keV is large enough to let the μ-Si atom leave its place in the lattice through recoil, and, hence, a quasi-Al ion moves to an interstitial place in the lattice. Only at low temperatures does it stay at this place and is the muon polarization not immediately destroyed. Nevertheless, because of the interaction of the muon spin with an *unpaired* electron in a quasi-Al^{++} state, strong depolarization takes place. Strong boron doping apparently favors occupancy of a quasi-Al$^+$ state with *no unpaired* electrons where the muon does not suffer strong depolarization.

6. APPLICATION OF EXOTIC ATOMS

6.1 Application of Muonic X Rays for Elemental Analysis

As rather high-energy x rays characteristic of the element that captured the muon are emitted when muons are stopped in a mixture or compound of elements and as the Coulomb-capture probability is rather well known for a large number of elements (see Sec. 5.2), the negative muon can be used as a probe for elemental analysis. The method is nondestructive, as only a tiny activation of the sample is to be expected; it is sensitive to all elements except (perhaps) the lightest ones; it may be applied (and has been applied) in surface analysis because muons with kinetic energies in the keV region, which can be produced in modern meson factories with a sufficient rate, are stopped in the first few layers of the sample; it certainly is not cheap, but rather simple, as the energy of the muonic x rays is in a region easily accessible to measurements with semiconductor detectors. The sensitivity of the method is about 10^{-4} atomic abundance. As an example, Fig. 9 shows the muonic x ray spectra from body and glaze of a bowl manufactured in Cairo in the 14th century. It is easily noted that Pb and Sn lines appear in the spectrum from the glaze, and the Ca intensity is increased, as well.

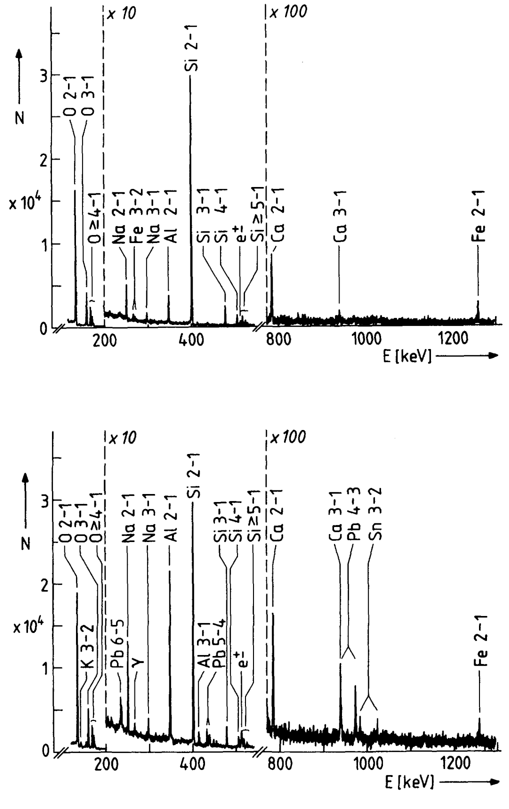

FIG. 9. Muonic x-ray spectra from body (top) and glaze (bottom), respectively, of an islamic bowl (Cairo, 14th century) (Köhler *et al.*, 1981).

6.2 Application of Muonic X Rays in Diagnostics

As pointed out more than 20 years ago, muons may also be used to determine the elemental composition of tissue and to replace neutron activation analysis because muon analysis is less destructive to the organism. Normal and pathological tissues are expected to differ, e.g., by their oxygen content. These expectations have been confirmed by experiments on the liver from a healthy dog and from a dog having suffered from liver cancer.

It seems to be feasible now to apply muons for the scanning of the human body. Both the path of the muon entering the body and the track of the decay electron may be determined with the help of multiwire chambers. The characteristic x rays from the atom that finally captures the muon may also be detected. Thus, even a complete scan of the elemental composition of the body might become feasible.

7. MUONIC MOLECULES

7.1 Definition

Muonic molecules are systems in which two nuclei are bound by the attracting force of a muon, rather similar to an electronic molecule where the binding force is exerted by electrons. In principle, exotic molecules could be formed with each kind of negatively charged exotic particle. The short lifetime and/or the strong interaction of the particle has made observation of exotic molecules other than those with muons impossible up to now. Muonic-molecule formation has been studied nearly exclusively for molecules formed by nuclei of the hydrogen isotopes (protium, deuterium, and tritium); besides these molecules, only those formed between hydrogen and He have found some interest. The large mass of the muon compared to the electron ($m_\mu \approx 207 m_e$) has two effects: The distances in these muonic molecules are about 207 times smaller, and the vibration frequency is 3000 times larger than in the corresponding electronic molecules. This in-

Table 4. Muonic molecules formed from hydrogen isotopes. The terms J and v denote the rotational and vibrational quantum numbers, respectively. The molecular formation rates quoted are normalized to liquid hydrogen density (4.25×10^{22} atoms/cm^3).

Exotic molecule	First bound level (J, v)	Binding energy (eV)	Molecular formation rate (μs^{-1})
$pp\mu$	(1,0)	107.312	2.2
$pd\mu$	(1,0)	97.587	5.9
$pt\mu$	(1,0)	99.190	6.5
$dd\mu$	(1,1)	1.967	...[a]
$dt\mu$	(1,1)	0.632	...[a]
$tt\mu$	(1,1)	45.236	3.0

[a] The molecule formation is resonant, and, hence, the formation rate is temperature dependent.

creases fusion probability by tens of orders of magnitude and leads to fast nuclear fusion. In the following, we will deal exclusively with muonic molecules with hydrogen nuclei.

7.2 Formation of Muonic Molecules

Muonic molecules are formed by collision reactions between muonic hydrogen atoms and electronic hydrogen molecules. Correspondingly, six different types of molecules may be formed between the three kinds of muonic hydrogen atoms and electronic hydrogen molecules. They are listed in Table 4. As transfer (cf. Sec. 3.4) from the lighter hydrogen isotope to the heavier one is much faster than molecule formation, the generation of muonic molecules is appreciable only for collisions of muonic hydrogen atoms with molecules formed by atoms of equal or smaller mass. Two formation mechanisms are possible:

1. Nonresonant formation. The formation energy of the muonic molecule is used to eject an electron from the partner molecule, e.g.,

   $$d\mu + H_2 \rightarrow [(pd\mu)pe]^+ + e^- .$$

   This process is comparatively slow and can take place for all kinds of muonic molecules between hydrogen isotopes (cf. Table 4).
2. Resonant formation. The energy released during the formation of the muonic molecule is used to excite rotations (quantum number K) and vibrations (quantum number v) in the muomolecular complex, e.g.,

   $$t\mu + D_2 \rightarrow [(dt\mu)dee]^*_{vK}.$$

   Here the $dt\mu$ molecular ion is one center of the complex, the second deuteron the other. Resonant formation becomes effective when the sum of the binding energy of the muonic molecule $|\epsilon_{11}|$ and of the kinetic energy ϵ_0 of the muonic atom matches the excitation energy ΔE_ν of the electronic molecule (cf. Fig. 10),

   $$\epsilon_0 + |\epsilon_{11}| = \Delta E_\nu. \qquad (11)$$

   This may happen in two ways:
   —During slowing down, the muonic hydrogen atom passes kinetic-energy regions where condition 11 is fulfilled.
   —When the muonic atoms have been

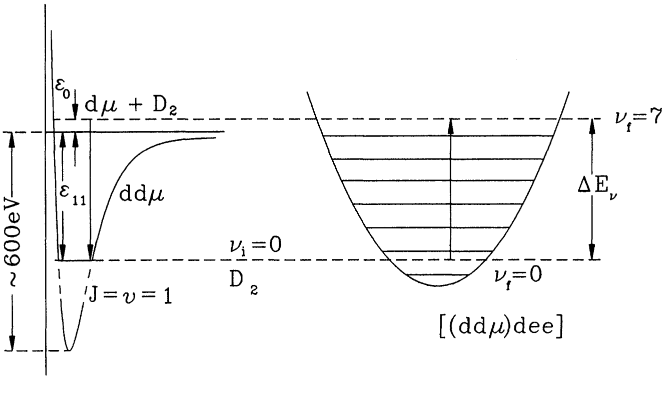

FIG. 10. The energy-level scheme for resonant $dd\mu$ molecule formation. The resonance condition is most easily met for the mesomolecular complex $[(dd\mu)dee]^*_{K,\nu}$ in the vibrational state with $\nu = 7$.

slowed down to thermal energies, which means that their kinetic-energy distribution is Maxwellian with a most probable energy $\epsilon_{\max} = \frac{3}{2}kT$ (with T the absolute temperature), there are always muonic hydrogen atoms with energies for which Eq. (11) is satisfied. Their fraction, however, depends strongly on T and should reach a maximum at a temperature for which $\epsilon_{\max} = \epsilon_0$.

For the most loosely bound levels in the $dd\mu$ and $dt\mu$ molecules, nonresonant formation is impossible, as the small energy gained in molecule formation (below 2 eV) is much smaller than the ionization energy of the host molecule (ionization potential $\approx$15 eV).

7.3 Energy Levels in Muonic Molecules

The calculation of binding energies in muonic molecules has been a successful application of three-body theory, and a high accuracy has been achieved during several decades of research. Typical results for the binding energy of the first state populated after muomolecule formation may be found in the third column of Table 4. Only small corrections, mainly for vacuum polarization and finite nuclear size (cf. Sec. 1), have to be added to the nonrelativistic binding energy. The respective values for the $dt\mu$ molecule, as an example, are nonrelativistic binding energy 660.2 meV, vacuum polarization and finite-size corrections 16.6 and 13.3 meV, respectively, total binding energy 631.8 meV.

The measurement of the muonic-molecule formation rate is a very sensitive tool to determine the binding energy of the (rotational quantum number $J = 1$, vibrational quantum number $v = 1$) level in the $dd\mu$ and $dt\mu$ molecules where resonant molecule formation takes place: As the energies of the rotational and vibrational states of the *electronic* host molecule are very well known, and the kinetic energy of the muonic atom *and* of the molecule it collides with are only some tens of meV in thermal equilibrium, the resonance condition poses a stringent condition on the binding energy of the muonic molecule. By changing the temperature of the mixture, one may realize small changes in $\epsilon_{\max}$ and, hence, may scan slowly through the resonances. Thus, the level energy in muonic molecules may be determined to some tenths of a meV, a tiny fraction of the total binding energy.

7.4 Deexcitation of Muonic Molecules

As stated earlier (cf. Table 4) the muonic molecules are preferentially formed in excited rotational-vibrational states with $J = 1$. For muonic molecules with identical nuclei ($pp\mu$, $dd\mu$, $tt\mu$), electromagnetic E1 transitions to lower states are suppressed by the Pauli principle. In the $pd\mu$ and $pt\mu$ molecules, mostly formed in the ($J = 1, v = 0$) state, transitions by the Auger effect of the kind

$$[(pd\mu)_{J=1,v=0}\, pee] \rightarrow [(pd\mu)_{J=0,v=0}\, pe]^+ + e^-$$

are fast (rates of the order of 10^{11} s^{-1}). In the $dt\mu$ molecule, finally, a competition takes place between de-excitation to the ground state ($J = 0$, $v = 0$) and the fusion reaction. Therefore, in 80% of the cases, fusion takes place from the ($J = 0$, $v = 1$) state, while in the other cases the molecule reaches the ground state.

7.5 Disappearance of Muonic Molecules: Muon-Induced Fusion

The muonic molecule may disappear by nuclear fusion or by muon decay. The rate λ_L^{Jv} for fusion from a state with orbital angular momentum L between the nucleons is given by

$$\lambda_L^{Jv} = K_L \rho_L^{Jv}.$$

In this expression, K_L is the nuclear-reaction constant, closely related to the cross section for nuclear fusion at zero kinetic energy between the nucleons, and ρ_L^{Jv} is the probability for the nuclei to be at zero distance. Only fusion reactions with $L = 0$ and $L = 1$ play a role. Fusion energies and rates for the different muon-induced fusion reactions are given in columns 3 and 4 of Table 5.

After muon-induced fusion, the muon is not necessarily lost for further fusions. Only in the reactions with gamma emission is the probability high (cf. column 5 of Table 5) that, immediately after fusion, the muon "sticks" to the fusion product ("initial sticking"), as the recoil kinetic energy of the generated nucleus is small. In all other cases (and especially for dt fusion), the probability of muon

Table 5. Fusion reaction(s) by strong interaction in muonic molecules.

Exotic molecule	Main fusion reactions	Energy gain (MeV)	Fusion rate (s^{-1})	Sticking probability
$pd\mu$	$pd\mu \rightarrow \mu^3He + \gamma$	5.5	2.6×10^5	0.99
$pt\mu$	$pt\mu \rightarrow \mu^4He + \gamma$	19.6	0.7×10^5	0.94
$dd\mu$	$dd\mu \rightarrow \mu + {}^3He + n$	3.3	4.3×10^8	0.12
	$\rightarrow t\mu + p$	4.0		
$dt\mu$	$dt\mu \rightarrow \mu + {}^4He + n$	17.6	1.2×10^{12}	0.0055
$tt\mu$	$tt\mu \rightarrow \mu + {}^4He + 2n$	11.3	1.5×10^7	0.14

"rejuvenation" is high. This is partly because the generated nucleus has too high a velocity to allow the muon to stick to it, but partly also because of the phenomenon that, in case the muon initially sticks to the nucleus, the muon is stripped with high probability from the muon-nucleus system during the deceleration ("reactivation"). The reactivation probability is the higher the larger the recoil energy of the μ-nucleus system generated by the fusion process.

Because the sticking probability is so low and muomolecule formation and muon-induced fusion are so fast, the cycle of muon-induced deuterium-tritium fusion is a very special case and is usually meant when one speaks about muon-catalyzed fusion, μCF.

7.6 The Cycle of Muon-Catalyzed *dt* Fusion

The cycle of muon-catalyzed *dt* fusion comprises muoatomic, muomolecular, and nuclear-fusion processes. It is shown in Fig. 11. Muons slowed down to an energy of around 10 eV in a mixture of deuterium and tritium form $D_2^+\mu^-$ and $T_2^+\mu^-$ complexes, which decay forming excited $d\mu$ or $t\mu$ atoms (kinetic energy ≈ 1 eV). After muon capture

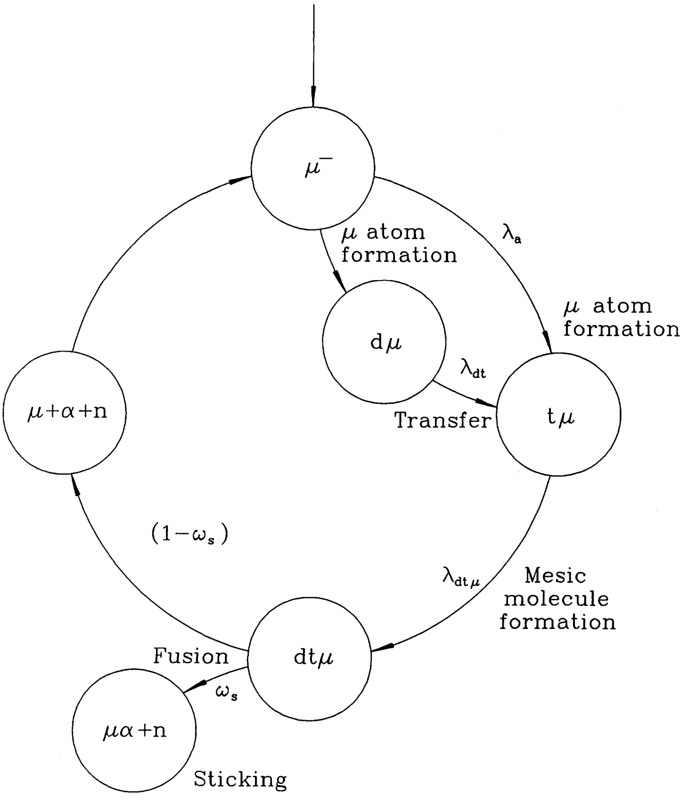

FIG. 11. The main muon catalysis cycle in a mixture of deuterium and tritium. λ_a, λ_{dt}, $\lambda_{dt\mu}$: rates for atom formation, transfer, and mesonic molecule formation, respectively; ω_s: probability for sticking.

to deuterium, fast transfer to tritium takes place (but, for energy reasons, not in the reverse direction; see Sec. 3.4). Calculations have revealed that transfer should be much faster if the muon is still in an excited state of the $d\mu$ atom. After transfer, the $t\mu$ system has appreciable kinetic energy as, e.g., the muon-binding energies in the $d\mu(1s)$ and $t\mu(1s)$ states differ by as much as 45 eV. The $t\mu$ system is slowed down by collisions with other atoms; the deceleration becomes rather slow at energies around 3 eV because there is a minimum in the scattering cross section (Ramsauer–Townsend effect). Finally, $dt\mu$ molecules are formed. Deexcitation and fusion compete in the muomolecule. After fusion, the muon sticks to the charged fusion product, an α particle, for only 0.7% of the cases. After this "initial" sticking, 30% of the trapped muons are reactivated during deceleration of the $\alpha\mu$ system. The rejuvenated muon can catalyze further fusions until it decays during the cycle or finally sticks to the fusion product where it is trapped until it decays. Up to 150 ± 20 fusions catalyzed by one muon were observed in liquid deuterium-tritium mixtures at tritium concentrations around 50%. Since, for each fusion event, 17.6 MeV of energy is produced, the energy release per muon is ≈ 2 GeV. Furthermore, for each fusion, a neutron of 14.1 MeV is produced and can be used to fission ^{238}U or to breed ^{239}Pu. This has posed the question of energy production by μCF.

7.7 The Prospects of Energy Production by Muon-Catalyzed *dt* Fusion

To judge today's perspective of energy production by muon-catalyzed fusion, one has to compare the energy necessary to produce one muon with the energy gain from about 150 catalyzed dt fusions. Muon production has to proceed via pion generation by nuclear reactions of GeV protons or deuterons with light nuclei (e.g., Be or C) and successive pion decay to convert pions into muons. Shooting deuterons of 4 GeV on carbon, for example, the mean energy necessary to create one π^- is about 6 GeV, and, at a conversion efficiency of 0.75, roughly 8 GeV have to be spent per produced muon. By increasing the accelerator efficiency, this amount can, perhaps, be decreased by another 25%, but still the energy necessary to produce one muon is three times larger than the energy gained by fusion processes catalyzed by this muon.

One way to make the energy balance in a μCF-based device positive is the muon-catalytic hybrid reactor (MCHR) (Petrov, 1985). The scheme is shown in Fig. 12. A beam of deuterons with ≈ 1 GeV/nucleon is directed to a target T, made from Be or Li. A fraction of 30% of the beam energy is used to create pions; the other 70% are used to fission ^{238}U or breed ^{239}Pu. The pions produced decay in a converter C into muons. The muons are guided to the so-called synthesizer S filled with a mixture of D_2 and T_2, where muon-catalyzed fusion takes place. Neutrons gen-

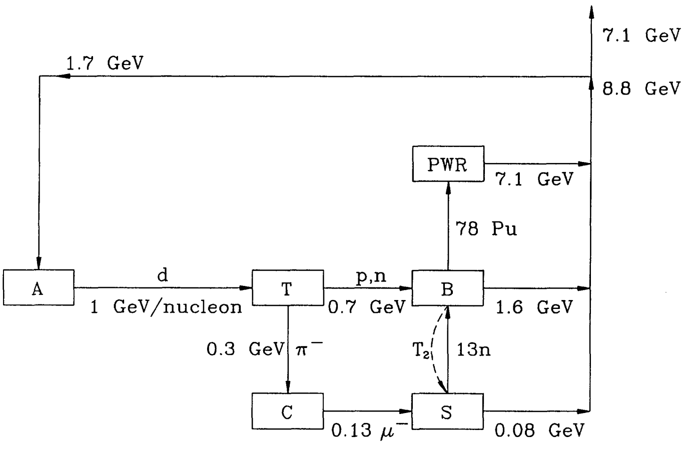

FIG. 12. Concept for a hybrid reactor based on muon-catalyzed *dt* fusion (Petrov, 1985). The energies given in the figure are energies needed or gained per nucleon. A: accelerator; T: target for pion production (Li or Be); C: converter $\pi^- \rightarrow \mu^-$, S: synthesizer (a high-density mixture of deuterium and tritium); B: blanket (a mixture of U and Li); PWR: nuclear power reactor.

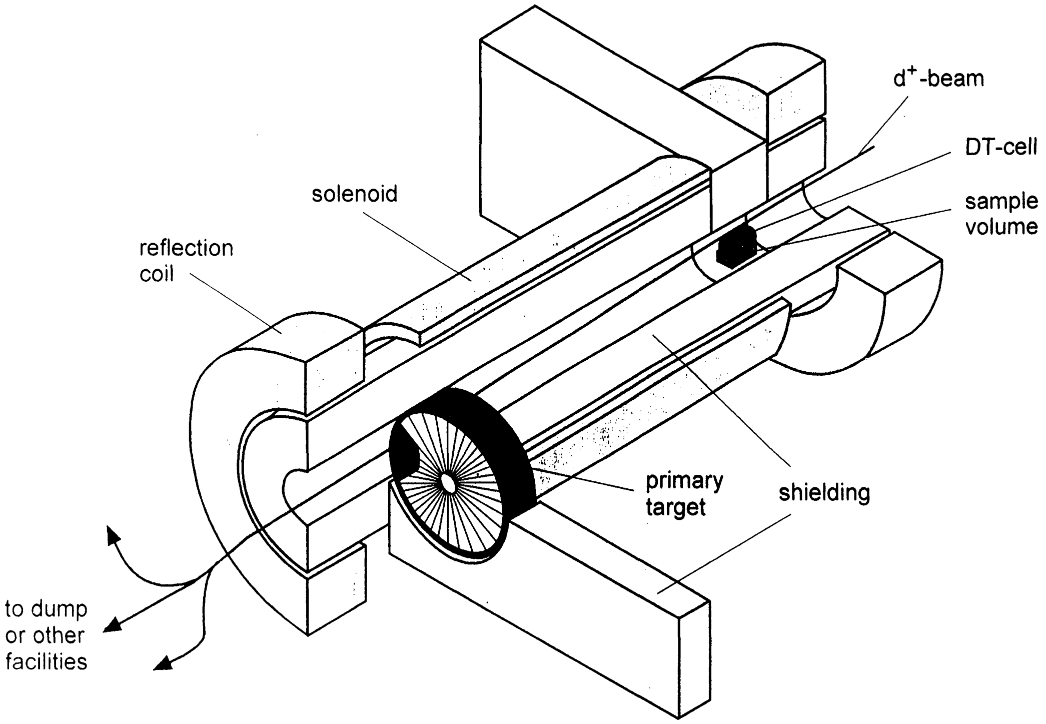

FIG. 13. Scheme for a high-flux source for 14.1 MeV neutrons based on muon-catalyzed fusion (Petitjean *et al.*, 1993).

erated in the synthesizer hit a blanket B (^{6}Li mixed with ^{238}U); tritium and plutonium are bred, which can be used to replace the tritium used in the fusion process and to feed thermal nuclear reactors that generate energy in a conventional way, respectively.

Doubts about the consequences of the MCHR for the environment and the very ambitious technology make it rather unlikely that this interesting scheme will become reality in the near future. Another system using muon-catalyzed fusion, which could be built using current technology, however, is a high-flux source for high-energy (14.1 MeV) neutrons necessary to test "first wall" and blanket materials for a future thermal-fusion reactor.

7.8 A High-Flux Neutron Source Based on Muon-Catalyzed *dt* Fusion

The conceptual design for an intense 14-MeV neutron source based on muon-catalyzed fusion (Petitjean *et al.*, 1993) is rather similar to the first stages of the MCHR (cf. Fig. 13). An intense deuteron beam interacts with a 30- to 50-cm-long carbon target. Muons from the backward decay of the generated pions are guided by a 5- to 10-T magnetic field to a target with a deuterium-tritium mixture of high density. With a deuteron beam of 12 mA at 1.5 GeV, about 10^{16} π^-/s can be generated, and up to 10^{15} μ^-/s may stop in the DT target. If we assume 100 fusion cycles to be catalyzed by one muon, a source strength of $\approx 10^{17}$ n/s results. The part of the beam not interacting with the carbon target (roughly 30–40%) can be used for other facilities (e.g., a spallation neutron source).

GLOSSARY

Antiproton: Antiparticle of the proton. Proton and antiproton differ in their charge. Annihilation of an antiproton with a proton sets free an energy of nearly 2 GeV, which goes mostly into pions.

Auger Effect: Transition of a particle (electron, exotic particle) from a free to a bound state or from one bound state to another one of lower energy with emission of an atomic electron.

Baryonic Atom: Exotic atom with a baryon (i.e., a strongly interacting elementary particle at least as heavy as the antiproton).

Cascade: The sequence of processes during the de-excitation of muonic, mesonic, and baryonic atoms. It comprises the de-excitation processes themselves (radiative, Auger, Stark transitions, etc.) and the refilling of the electron shell depleted by Auger transitions.

Coulomb Capture: Capture of a negatively charged elementary particle into an atom by virtue of the Coulomb interaction.

Coulomb Potential: Potential between two pointlike charges. The Coulomb potential decreases as $1/r$, where r is the distance between the charges.

Dirac Equation: Relativistic wave equation to describe particles with half-integer spin (like muons, antiprotons).

Fermi Distribution: Distribution function devised by Fermi. In exotic atoms used to describe the radial charge density in the nucleus.

Fine Structure: Splitting of energy levels in a muonic, mesonic, or baryonic atom with the same principal and angular-momentum quantum numbers. The fine-structure splitting is due to the interaction between spin and orbital angular momentum of the exotic particle and occurs only for particles with nonzero spin.

Hadronic Atom: Exotic atom with a strongly interacting elementary particle (pion, kaon, antiproton, Σ hyperon).

Hyperfine Structure: Splitting of energy levels in a muonic, mesonic, or baryonic atom with the same total angular momentum of the exotic particle; due to the interaction with the nuclear moments (e.g., the electric quadrupole moment or the magnetic dipole moment).

Kaon: Strongly interacting elementary particle with a mass of 966 electron masses and a life time *in vacuo* of $\approx$12 ns.

Klein–Gordon Equation: Relativistic wave equation to describe particles with integer spin (like pions or kaons).

Mesonic Atom: Exotic atom with a meson (pion, kaon).

Muon: Elementary particle. The muon is very similar to the electron, only its mass is $\approx$207 times as large as that of the electron, and its lifetime is only 2.2 μs.

Muon-Induced Fusion: Process of nuclear fusion of (hydrogen) nuclei bound in muonic molecules. Because of the small distance between the nuclei brought about by the (heavy) muon, the fusion process is very fast.

Muonic Molecules: Dinuclear molecules, usually with hydrogen isotopes, where a muon acts as the negative particle binding the hydrogen nuclei.

Nuclear Polarization: The polarization of the nucleus by the electric charge of the exotic particle in muonic, mesonic, and baryonic atoms. It plays a role only for the most tightly bound orbits of the exotic atom.

Pion: Elementary particle with a mass of $\approx$273 electron masses and a lifetime *in vacuo* of 26 ns. The particle interacts strongly with nucleons. Besides positive and negative pions, also neutral pions are known.

Ramsauer–Townsend Effect: Pronounced minimum in the cross section for the scattering of muonic hydrogen atoms by other hydrogen atoms.

Stark Mixing: Transitions between states with different orbital angular momentum (quantum number l)—but usually with the same principal quantum number—during encounters of an exotic atom with other atoms. In this case, the spherical symmetry of the exotic atom is destroyed, and l is no longer a good quantum number. Stark mixing plays a very important role in exotic hydrogen atoms.

Transfer: Transition of an exotic particle from an exotic hydrogen atom (μp, μd, or μt) to another (heavier) atom. Transfer plays a very important role in the muon-induced-fusion cycle.

Vacuum Polarization: The QED correction to the electromagnetic binding energy that is most important in muonic, mesonic, and baryonic atoms. It is due to the creation of virtual electron-positron (or $\mu^-\mu^+$) pairs in the electric field of the nucleus.

Works Cited

Akylas, V., Vogel, P. (1978), *Comp. Phys. Commun.* **15**, 291–302.

Bamberger, A., Lynen, U., Piekarz, H., Povh, B., Ritter, H. G., Backenstoss, G., Bunaciu, T., Egger, J., Hamilton, W. D., Koch, H. (1970), *Phys. Lett.* **33B**, 233–235.

Barrett, R. C. (1977), in: V. W. Hughes, C. S. Wu (Eds.), *Muon Physics*, Vol. 1, New York: Academic, pp. 309–322.

Batty, C. J., Friedman, E., Gal, A., Kalbermann, G. (1991), *Nucl. Phys. A*, **535**, 548–572.

Bethe, H. (1930), *Ann. Phys. (Leipzig)* **5**, 325–400.

Bethe, H. A., Salpeter, E. E. (1977), *Quantum Mechanics of One- and Two-Electron Atoms*, New York: Plenum.

Burleson, G. R., Cohen, D., Lamb, R. C., Michael, D. N., Schluter, R. A., White, T. O. (1965), *Phys. Rev. Lett.* **15**, 70–72.

Camac, M., McGuire, A. D., Platt, J. B., Schulte, H. J. (1952), *Phys. Rev.* **88**, 134; (1955), *ibid.* **99**, 897–905.

Conversi, M., Pancini, E., Piccioni, O. (1947), *Phys. Rev.* **71**, 209–210.

Daniel, H. (1979), *Z. Phys.* **A291**, 29–31.

Daniel, H. (1980), *Ann. Phys. (NY)* **129**, 303–319.

Ericson, M., Ericson, T. E. O. (1966), *Ann. Phys. (NY)* **36**, 323–362.

Fermi, E., Teller, E., Weisskopf, V. (1947), *Phys. Rev.* **71**, 314–315.

Fermi, E., Teller, E. (1947), *Phys. Rev.* **72**, 399–408.

Ferrell, R. A. (1960), *Phys. Rev. Lett.* **4**, 425–428.

Gershtein, S. S. (1993), in: L. Schaller, C. Petitjean (Eds.), *Muonic Atoms and Molecules*, Basel: Birkhäuser, pp. 169–186.

Hartmann, F. J., von Egidy, T., Bergmann, R., Kleber, M., Pfeiffer, H.-J., Springer, K., Daniel, H. (1976), *Phys. Rev. Lett.* **37**, 331–334.

Heitlinger, K., Bacher, K., Badertscher, A., Blüm, P., Eades, J., Egger, J., Elsener, K., Gotta, D., Morenzoni, E., Simons, L. M. (1992), *Z. Phys.* **A342**, 359–368.

Holzwarth, G., Pfeiffer, H.-J. (1975), *Z. Phys.* **A272**, 311–313.

Köhler, E., Bergmann, R., Daniel, H., Ehrhart, P., Hartmann, F. J. (1981), *Nucl. Instrum. Methods* **187**, 563–568.

Konijn, J., de Laat, C. T. A. M., Taal, A., Koch, J. H. (1990), *Nucl. Phys. A* **519**, 773–804.

Machner, H. (1993), Kernforschungsanlage Jülich Report No. KFA-IKP(1)-1993-4.

Mann, R. A., Rose, H. E. (1961), *Phys. Rev.* **121**, 293–301.

Markushin, V. M. (1990), in: L. M. Simons, D. Horvath, G. Torelli (Eds.), *Electromagnetic Cascade and Chemistry of Exotic Atoms*, New York/London: Plenum, pp. 73–96.

Particle Data Group (1992), *Phys. Rev. D* **45**, S1–S584.

Petitjean, C., Atchison, F., Heidenreich, G., Walter, H. K., Amelotti, F., Andreani, R., de Marco, F., Monti, S., Pillon, M., Vecchi, M., Markushin, V. E., Ponomarev, L. I., Niebuhr, C. (1993), Paul-Scherer-Institut Report No. PSI-PR-93-09.

Petrov, Yu. V. (1985), *Nature* **285**, 466–468.

Sayasov, Yu. S. (1990), *Helv. Phys. Acta* **63**, 547–548.

Yamazaki, T., Widmann, E., Hayano, R. S., Iwasaki, M., Nakamura, S. N., Shigaki, K., Hartmann, F. J., Daniel, H., von Egidy, T., Hofmann, P., Kim, Y.-S., Eades, J. (1993), *Nature* **361**, 238–240.

Further Reading

General:

C. J. Batty (1982), *Sov. J. Part. Nucl.*, **13**, 71–96.

Horvath, D., Lambrecht, R. M. (1984), *Exotic Atoms—A Bibliography 1939–1982*, Amsterdam: Elsevier.

Hughes, V. W., Wu, C. S. (Eds.) (1977), *Muon Physics*, Vols. I and III, New York: Academic.

Kim, Y. N. (1971), *Mesic Atoms and Nuclear Structure*, Amsterdam/London: North-Holland.

Schaller, L., Petitjean, C. (Eds.) (1993), *Muonic Atoms and Molecules*, Basel: Birkhäuser.

Simons, L. M., Horvath, D., Torelli, G. (Eds.) (1990), *Electromagnetic Cascade and Chemistry of Exotic Atoms*, New York/London: Plenum.

West, D. (1958), *Rep. Prog. Phys.* **21**, 271–311.

Yamazaki, T., Nakai, K., Nagamine, K. (Eds.) (1992), *Perspectives of Meson Science*, Amsterdam: North-Holland.

Section 1:

Engfer, R., Schneuwly, H., Vuilleumier, J. L., Walter, H. K., Zehnder, A. (1974), *At. Data Nucl. Data Tables* **14**, 509–597.

Section 2:

Leon, M., Seki, R. (1977), *Nucl. Phys.* **282**, 445–460.

Leon, M., Miller, J. H. (1977), *Nucl. Phys.* **282**, 461–473.

Vogel, P., Haff, P. K., Akylas, V., Winther, A. (1975), *Nucl. Phys.* **254**, 445–479.

Section 3:

Eisenberg, Y., Kessler, D. (1961), *Nuovo Cimento*, **19**, 1195–1210.

Evseev, V. S. (1977), in: V. W. Hughes, C. S. Wu (Eds.) (1977), *Muon Physics*, Vol. III, New York: Academic, pp. 235–298.

Section 5:

von Egidy, T., Jakubassa-Amundsen, D. H., Hartmann, F. J. (1984), *Phys. Rev. A* **29**, 455–461.

Section 6:

Sakamoto, K., Hosoi, Y., Nagamine, K. (1992), in: T. Yamazaki, K. Nakai, K. Nagamine (Eds.), *Perspectives of Meson Science*, Amsterdam: North-Holland, pp. 487–491.

Section 7:

Bogdanova, L. N. (1992), *Surv. High Energy Phys.* **6**, 177–221.

Breunlich, W., Kammel, P., Cohen, J. S., Leon, M. (1989), *Annu. Rev. Nucl. Part. Sci.* **39**, 311–356.

Ponomarev, L. J., Gershtein, S. S. (1977), in: V. W. Hughes, C. S. Wu (Eds.), *Muon Physics*, Vol. III, New York: Academic, pp. 141–233.

Ponomarev, L. I. (1990), *Contemp. Phys.* **31**, 219–245.

MUONIUM AND POSITRONIUM

TOSHIMITSU YAMAZAKI, *Institute for Nuclear Study, University of Tokyo, Tanashi, Tokyo 188, Japan*

YASUO ITO, *Research Center for Nuclear Science and Technology, University of Tokyo, Tokai, Ibaraki 319-11, Japan*

Introduction 79
1. **Basic Properties of Free Muonium** 81
1.1 Muonium Spin Dynamics 81
1.2 Muonium in Longitudinal Magnetic Field 81
1.3 Muonium in Low Transverse Field 82
1.4 Muonium in High Transverse Field 82
2. **Behavior of Muonium in Matter** 83
2.1 Muonium Chemistry 83
2.2 Muonium States in Solids 84
2.3 Quantum Diffusion of Muonium 86
3. **Basic Properties of Positronium** 86
3.1 Structure and Annihilation of Positronium 86
3.2 Formation of Positronium 87
3.3 Methods for Studying Positronium 88
4. **Behavior of Positronium in Matter** 88
4.1 Positronium in Gases 88
4.2 Positronium in Insulators 89
4.3 Positronium in a Metal Surface and in Voids 91
5. **Advanced Source of Muons and Positrons** 91
5.1 Ultralow-Energy Muon and Muonium Beams 91
5.2 Advanced Techniques of Positronium Studies 91
6. **Test of Quantum Electrodynamics** 92
6.1 Hyperfine Structure 92
6.2 1s–2s Energy Interval and Lamb Shift 92
6.3 Positronium Annihilation Rate 93
Works Cited 94
Further Reading 95

INTRODUCTION

Muonium (μ^+e^-, atomic symbol Mu) and positronium (e^+e^-, atomic symbol Ps) are the simplest atoms comprising only charged leptons: the electron (e^-,e^+) and the muon (μ^-,μ^+). They are hydrogenlike atoms with the proton in H replaced by either μ^+ or e^+. The proton has a finite size, which may affect the binding energies, while μ^+ and e^+ are believed to be point objects. Thus, Mu and Ps not only provide ideal means for testing quantum electrodynamics, but also serve as unique probes of matter. The basic properties of μ^+ and e^+ are compared to those of p^+ in Table 1, and those of Mu and Ps are given in Table 2 in comparison with the properties of H. The level diagrams of the three light isotopes of hydrogen, namely, Ps, Mu, and H, are presented in Fig. 1.

Both Mu and Ps are unstable atoms. Although studies of small hydrogen impurities in matter are very interesting and important, they are practically difficult. On the other hand, implanted muons form hydrogenlike impurities and probe their behavior and local electronic structures. The lifetime of Mu is limited by the free lifetime of μ^+ itself (2.2 μs). The characteristic features of muon polarization and decay asymmetry permit unique spectroscopic uses of muons and muonium for studying the local electronic structure, chemical bonding, chemical reactions, diffusion, trapping, magnetism, etc. (see MUON SPIN ROTATION/RELAXATION/RESONANCE).

The e^+ is a stable particle in vacuum, but

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Table 1. Basic properties of positron, muon, and proton.

Property	e^+	μ^+	p^+
Rest mass (MeV)	0.511	105.6596 = 206.7685 m_e = 0.1126 m_p	938.2592 = 1836.4 m_e
Spin	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
Magnetic moment (erg/G)	$\mu_e = 9.2848 \times 10^{-21}$	$\mu_\mu = 4.490\,48 \times 10^{-23}$ $= 0.004\,836\ \mu_e$ $= 3.183\,34\ \mu_p$	$\mu_p = 1.410\,62 \times 10^{-23}$ $= 0.001\,52\ \mu_e$
g factor	2.002 319	2.002 332	2.675 20
Gyromagnetic ratio	2.802 47 MHz/G	13.5544 kHz/G	4.257 707 kHz/G
Lifetime	∞[a]	2.1971 μs	∞
Decay	(Pair annihilation)	$e^+\bar{\nu}_\mu \nu_e$	

---

[a] Stable in vacuum, but annihilates with e^- in matter. The lifetime of e^+ by pair annihilation depends on the electron density (see text).

Table 2. Basic properties of positronium, muonium, and hydrogen.

Property	Positronium	Muonium	Hydrogen
Mass	2 m_e	$m_\mu + m_e = 207.8\ m_e$	$m_p + m_e = 1837.6\ m_e$
Reduced mass	$m_e/2$	0.9952 m_e	0.9994 m_e
Spin	0,1	0,1	0,1
Bohr radius (nm)	0.105	0.053 17	0.052 94
Ionization energy (eV)	6.77	13.549	13.595
de Broglie wavelength (300 K) (nm)	5.428	0.2979	0.1004
Thermal velocity (300 K) (m/s)	6.7×10^4	0.75×10^4	0.25×10^4
Hyperfine frequency (MHz)	203 400	4463.3	1420.4

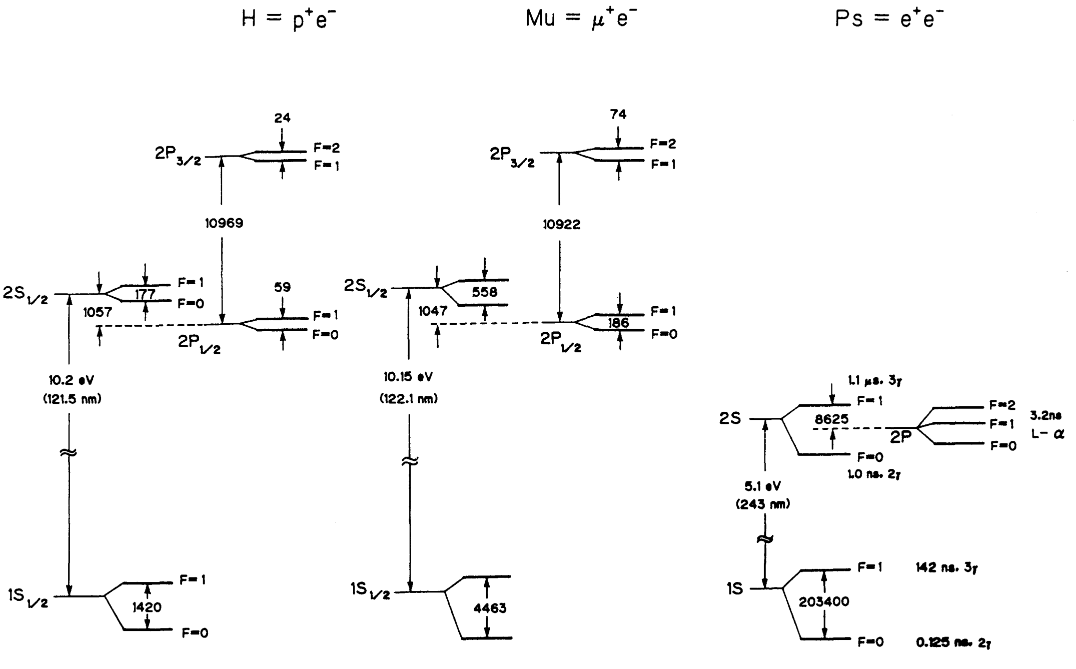
H = p⁺e⁻
Mu = μ⁺e⁻
Ps = e⁺e⁻
24
2P3/2
F=2
F=1
10969
2S1/2
F=1
177
F=0
1057
59
F=1
F=0
2P1/2
10.2 eV
(121.5 nm)
1S1/2
F=1
1420
F=0
74
2P3/2
F=2
F=1
10922
2S1/2
558
1047
F=1
186
F=0
2P1/2
10.15 eV
(122.1 nm)
1S1/2
4463
1.1 μs. 3γ
F=1
2S
8625
2P
F=0
1.0 ns. 2γ
F=2
F=1
F=0
3.2ns
L- α
5.1 eV
(243 nm)
1S
F=1
203400
F=0
142 ns. 3γ
0.125 ns. 2γ

FIG. 1. Energy levels of muonium and positronium compared to those of the hydrogen atom. The energies for the fine and hyperfine structures are in MHz. For positronium, the lifetimes and annihilation modes, or the lifetime of Lyman-α emission, are given in parentheses.

annihilates with one of the electrons in a material substance. In vacuum, Ps has a mean lifetime of 1.25×10^{-10} s (p-Ps) and 1.42×10^{-7} s (o-Ps). The lifetime of e^+ in matter depends on the local electron density and is the source of microscopic information concerning the electronic states probed by the e^+ and Ps. Positron annihilation spectroscopy has been developed and is used in various fields of materials science (see POSITRON ANNIHILATION SPECTROSCOPY).

Although the existence of Mu and Ps was naturally expected since the discovery of μ^+ and e^+, it took quite some time until Mu and Ps were really discovered. Muonium was first identified by Hughes *et al.* (1960) in Ar gas from the characteristic precession frequency of 1.39 MHz/G in a small transverse field. Experimental evidence for Ps was first obtained by Shearer and Deutsch (1949) as a long-lived component in the slowing down of e^+ in nitrogen and other gases.

1. BASIC PROPERTIES OF FREE MUONIUM

Although the muonium mass is about nine times smaller than the hydrogen mass, its electronic properties (e.g., the Bohr radius and ionization energy) are very close to those of hydrogen, since the reduced mass of Mu is only 0.4% smaller (see Table 2).

1.1 Muonium Spin Dynamics

Muonium is a system of two Dirac particles with the spin Hamiltonian

$$\mathcal{H} = g_e\mu_e\mathbf{B}\cdot\mathbf{S} - g_\mu\mu_\mu\mathbf{B}\cdot\mathbf{I} + A\mathbf{I}\cdot\mathbf{S}, \tag{1}$$

where g_e and g_μ are the electron and muon g factors, respectively, and μ_e and μ_μ are the corresponding Bohr magnetons. As in the hydrogen atom, the ground-state energy level is split into triplet (total angular momentum quantum number $F = 1$) and singlet ($F = 0$) states with a splitting equal to the hyperfine coupling constant (A), which in terms of the frequency corresponds to 4463 MHz (Fig. 1).

Muonium is formed when a μ^+ captures an electron. Under normal experimental conditions, the μ^+ is spin polarized. Since the electrons are unpolarized, muonium atoms with spin states $|\alpha\rangle_\mu|\alpha\rangle_e$ and $|\alpha\rangle_\mu|\beta\rangle_e$ are formed with a population of 50% in each state, where α and β denote up and down spin, respectively. The latter state (often mistakenly called "singlet Mu") oscillates between the two eigenstates $|F,m\rangle = |1,0\rangle$ and $|0,0\rangle$ (the true singlet state) at the hyperfine frequency of $\nu_0 =$ 4463 MHz. The former, accurately called "triplet Mu," is one of the eigenstates and is stable in a zero or longitudinal field. In general, Mu formed in matter may be different from free Mu because of possible coupling with neighboring atoms/molecules.

1.2 Muonium in Longitudinal Magnetic Field

When a magnetic field is applied in the direction of the μ^+ spin polarization, half of the initial muon polarization (i.e., in triplet Mu) does not evolve. At zero field, the observed μ^+ spin polarization is 50% of the initial μ^+ polarization, because μ^+ in singlet Mu evolves at the hyperfine frequency. The functions for the four spin states in an applied field B are

$$|1\rangle = |\alpha\alpha\rangle, \tag{2}$$

$$|2\rangle = s|\alpha\beta\rangle + c|\beta\alpha\rangle, \tag{3}$$

$$|3\rangle = |\beta\beta\rangle, \tag{4}$$

and

$$|4\rangle = c|\alpha\beta\rangle - s|\beta\alpha\rangle, \tag{5}$$

where

$$s = \frac{1}{\sqrt{2}}\left(1 - \frac{x}{\sqrt{1+x^2}}\right)^{1/2}, \qquad c = \frac{1}{\sqrt{2}}\left(1 + \frac{x}{\sqrt{1+x^2}}\right)^{1/2}. \tag{6}$$

Here, $x = B/B_0$ is the magnetic field intensity measured in units of the internal magnetic field B_0, which is defined by

$$B_0 = \frac{h\nu_0}{g_e\mu_e + g_\mu\mu_\mu} = \frac{2\pi\nu_0}{\gamma_e + \gamma_\mu}. \tag{7}$$

The value of B_0 for free Mu is 1585 G. The energy diagram, the so-called "Breit–Rabi" diagram, is shown in Fig. 2.

With increasing longitudinal field, the μ^+

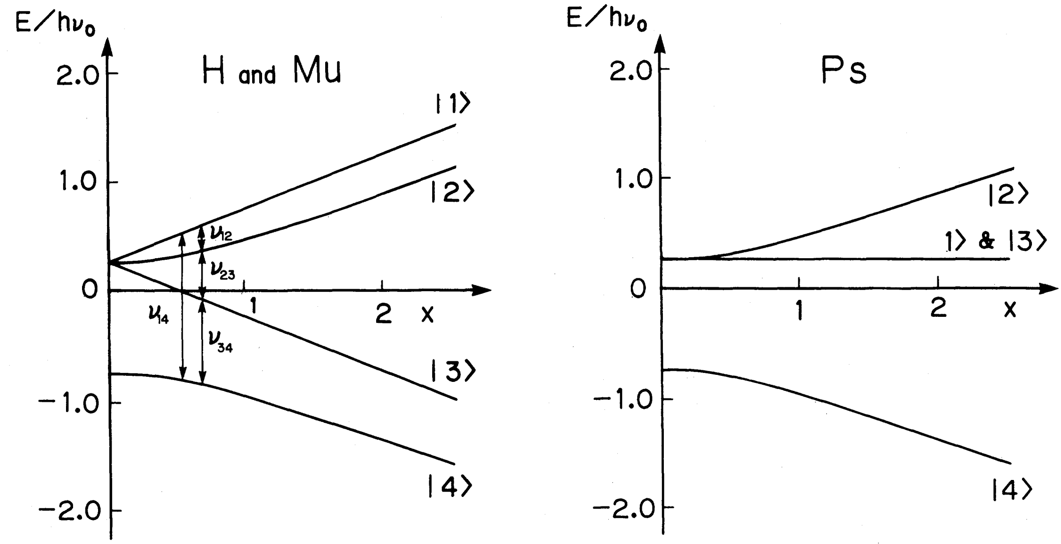

FIG. 2. Breit–Rabi diagram of H, Mu, and Ps. The energy levels of H, Mu, and Ps are described by common equations when the magnetic field is normalized to the internal magnetic field as $x = B/B_0$, and the energy normalized to the hyperfine coupling energy as $E/h\nu_0$. This results in very similar diagrams for H and Mu. For Ps, the levels $|1\rangle$ and $|3\rangle$ are degenerate at any magnetic field intensity as a result of a cancellation of the magnetic moments of positron and electron. B_0 = 505 G for H, 1585 G for Mu, and 36 245 G for Ps. ν_0 = 1420 MHz for H, 4463 MHz for Mu, and 203 400 MHz for Ps.

and e^- spins become decoupled as a result of the Paschen–Back effect, and the consequence is an increase in the observable μ^+ polarization,

$$P = \tfrac{1}{2}\left(1 + \frac{x^2}{1 + x^2}\right). \tag{8}$$

"Longitudinal field repolarization" refers to an experiment in which the μ^+ spin polarization is measured as a function of the longitudinal magnetic field. Here, not only the existence of Mu is tested, but also the *hf* coupling constant is roughly determined using Eq. (8).

1.3 Muonium in Low Transverse Field

In a low transverse magnetic field, the μ^+ spin in triplet Mu evolves (ν_{12} and ν_{23} in Fig. 2) according to the Larmor precession frequency,

$$\nu_M = (\gamma_M/2\pi)B, \tag{9}$$

where γ_M is the gyromagnetic ratio of Mu, namely,

$$\begin{aligned}\frac{\gamma_M}{2\pi} &= \frac{1}{2}\frac{g_e\mu_e - g_\mu\mu_\mu}{h} = \frac{1}{2}\frac{\gamma_e - \gamma_\mu}{2\pi} \\ &= -1.394\ \text{MHz/G}.\end{aligned} \tag{10}$$

Since the magnetic moment of Mu is nearly equal to that of the electron, the Larmor precession of Mu is about 100 times faster than that of the μ^+:

$$\nu_\mu = (\gamma_\mu/2\pi)B, \tag{11}$$

where

$$\begin{aligned}\gamma_\mu/2\pi &= g_\mu\mu_\mu/h \\ &= 0.013\,5544\ \text{MHz/G}.\end{aligned} \tag{12}$$

Thus, one can easily discriminate between μ^+ and Mu. At a moderate transverse field, ν_{12} and ν_{23} are not the same and the Mu precession contains a low-frequency beat with

$$\Delta\nu = \frac{\nu_{12} - \nu_{23}}{2} \sim -\frac{1}{4}\frac{\gamma_e^2 B^2}{2\pi\nu_0}. \tag{13}$$

From the observed beat, one can determine ν_0.

1.4 Muonium in High Transverse Field

At a sufficiently high field ($x^2 \gg 1$), the μ^+ spin evolution manifests two frequencies at

$$\begin{aligned}\nu_{12} &= \frac{\nu_0}{2} - \frac{\nu_0}{2}\left[(x^2 + 1)^{1/2} - \frac{\gamma_e - \gamma_\mu}{\gamma_e + \gamma_\mu}x\right] \\ &\simeq \frac{\nu_0}{2} - \nu_\mu\end{aligned} \tag{14}$$

and

$$\nu_{34} = \frac{\nu_0}{2} + \frac{\nu_0}{2}\left[(x^2+1)^{1/2} - \frac{\gamma_e - \gamma_\mu}{\gamma_e + \gamma_\mu}x\right] \simeq \frac{\nu_0}{2} + \nu_\mu. \tag{15}$$

From a pair of observed frequencies, ν_{12} and ν_{34}, one can precisely determine ν_0. A direct measurement of ν_0 is also possible by applying an rf field so as to cause transitions between the hyperfine doublet at zero field or between $|1\rangle$ and $|2\rangle$ ($|3\rangle$ and $|4\rangle$) at high field.

2. BEHAVIOR OF MUONIUM IN MATTER

Long-lived muonium has been identified by its characteristic precession frequency [Eq. (9)] in various substances: in gases [rare gases (except He and Ne), H_2, N_2, and vapors of molecular gases], in liquids [cryoliquids (including all liquified rare gases), water, alcohols, and some saturated organic liquids], in insulating solids (e.g., quartz, alkali halides, metal oxides, ice, and some organic solids), and in semiconductors (Si, Ge, GaAs, etc.). The muonium is formed when the ionization potential of host atoms/molecules is approximately less than the binding energy (13.5 eV) of Mu so that the electron transfer from a host atom to the μ^+ is energetically possible. The exceptional media are He and Ne, where the ionization energy (24.6 and 21.6 eV, respectively) is much larger than 13.5 eV. No muonium exists in metals, since the μ^+ charge is shielded by free electrons.

2.1 Muonium Chemistry

Like the hydrogen atom, Mu is a reactive species, and its reaction rate is measured from enhanced relaxation of the observed signal of Mu spin precession as a function of the amount of added reactants. This technique has been applied to studies of various reactions of Mu (substitution, abstraction, and spin-conversion reactions). The difference between the Mu chemical reactions, such as Mu + $F_2 \rightarrow$ MuF + F, and the analogous reactions of H atoms reveal remarkable isotope effects, which can be ascribed to quantum tunneling of Mu due to its light mass (Garner *et al.*, 1978).

When Mu reacts rapidly in a substance, it is chemically incorporated into a molecule. When it settles down in a diamagnetic species, such as MuOH, there is no hyperfine interaction for the μ^+, whose spin therefore precesses with approximately the frequency of a bare μ^+ [Eq. (11)]. Such Mu states are called "diamagnetic muons," and their observed polarization is denoted by "P_D." The fractions of diamagnetic muons and muonium formed in typical substances are given in Table 3.

Table 3. μ^+ and Mu formation in matter. P_D, polarization of diamagnetic muon; P_M, polarization of muonium; P_R, polarization of Mu-substituted free radicals; P_L, lost (or missing) polarization.

Substance		P_D	P_M	P_R	P_L
He	Gas	1.0	0	0	0
Ne	Gas	1.0	0	0	0
Ar	Gas	0.25	0.75	0	0
Kr	Gas	0	1.0	0	0
Xe	Gas	0	1.0	0	0
H_2	Gas, 3 atm	0.4	0.6	0	0
N_2	Gas, 1 atm	0.15	0.6	0	0
H_2O	Liq.	0.62	0.20	0	0.18
CH_3OH	Liq.	0.61	0.20	0	0.20
Hexane	Liq.	0.65	0.13	0	0.22
CCl_4	Liq.	1.0	0	0	0
C_6H_6	Liq.	0.15	0	0.65	0.20
Polyethylene		0.64	—	—	—
Quartz	Crystal	0.21	0.7	—	0.1
Graphite	Crystal	1.0	0	0	0
Diamond	Crystal, 4.2 K	0.1	0.2	0.12[a]	0.6

[a] The value of P_R for diamond refers to Mu*.

When Mu adds to unsaturated bonds, Mu-containing free radicals are produced. The μ^+ spin evolution in Mu-substituted free radicals is complicated because of the splitting of energy levels caused by superhyperfine coupling of the unpaired electron with the nuclear spins. At a sufficiently high field, however, the superhyperfine interactions are decoupled and the μ^+ spin evolves according only to its own hf coupling with the unpaired electron spin, showing the two frequencies given in Eqs. (14) and (15). The hyperfine frequency ν_0 is, however, considerably different from the free value. Such a state can be regarded as being a pseudo Mu with a larger μ^+–e^- distance. Different Mu-substituted free radicals can be identified from the hf constant ν_0 (Roduner *et al.*, 1978). Typical examples are shown in Fig. 3. Various Mu-containing free radicals are produced, and their structures have been studied. The fraction of Mu-substituted radicals, denoted by P_R, is presented also in Table 3.

The sum of P_D, P_M, and P_R is not necessarily equal to 1. The fraction of unobservable μ^+ spin polarization in a transverse field precession, $P_L = 1 - P_D - P_M - P_R$, is called "lost polarization" or "missing polarization" (see Table 3). Missing polarization may arise when the rate of the reaction that leads to diamagnetic muons or to Mu-substituted free radicals is not very high, i.e., on the order of the hyperfine frequency (4463 MHz) or less, and the μ^+ spin precession becomes out of phase. Missing polarization can be restored when a high longitudinal field is applied.

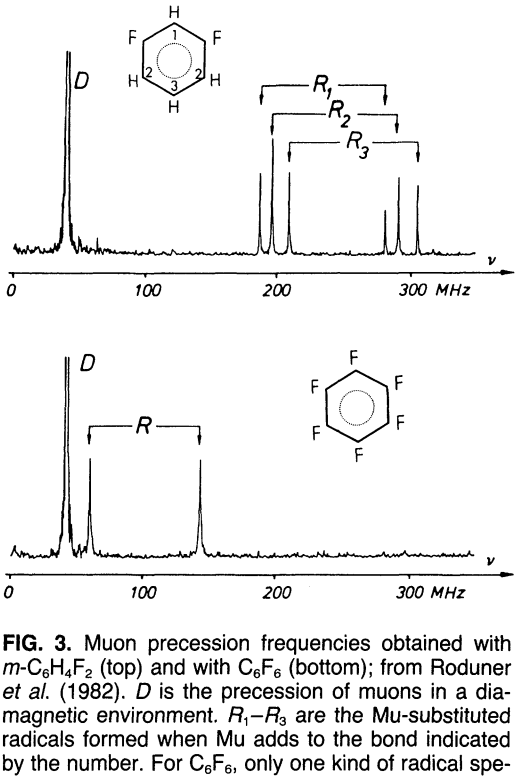

FIG. 3. Muon precession frequencies obtained with *m*-$C_6H_4F_2$ (top) and with C_6F_6 (bottom); from Roduner *et al.* (1982). *D* is the precession of muons in a diamagnetic environment. R_1–R_3 are the Mu-substituted radicals formed when Mu adds to the bond indicated by the number. For C_6F_6, only one kind of radical species is formed with Mu adding to any bonding.

Muonium states and Mu-substituted free radicals, when coupled with nearby nuclear moments, can be studied by the muon-level-crossing resonance (LCR) method, as first proposed by Abragam (1984), in which the μ^+ spin polarization is measured in a variable longitudinal magnetic field. At a particular magnetic field, where one of the Breit–Rabi levels crosses another one involving a nuclear spin flip, resonant transfer of polarization from μ^+ to the nuclear spin occurs. This results in a decrease in the μ^+ spin polarization. Each group of equivalent nuclear spins gives rise to a single resonance at a field B_R given by

$$B_R = \tfrac{1}{2}\left[\frac{A_\mu - A_n}{\gamma_\mu - \gamma_n} - \frac{A_\mu + A_n}{\gamma_e}\right], \tag{16}$$

where A_μ and A_n are the hyperfine coupling constants of the μ^+ and the nucleus, respectively, with the unpaired electron, and γ_μ, γ_e, and γ_n are the gyromagnetic ratios of the μ^+, e^-, and nucleus. The first successful application of this method was carried out by Kiefl *et al.*, (1986), who measured the modulated asymmetry $A^+ - A^-$ of integral counts of μ–e decay as a function of the longitudinal field. An example of the LCR spectrum is given in Fig. 4.

2.2 Muonium States in Solids

Muonium is formed in most insulators and semiconductors. Typical values of the hyperfine constant of Mu in solids are presented in Table 4. In most substances (quartz is exceptional), the hyperfine constant is smaller than the vacuum value; this is ascribed to the increased size of Mu in the crystal (Brewer *et al.*, 1973).

It is very important to study the location and local electronic structure of hydrogen impurities of an extremely small amount in

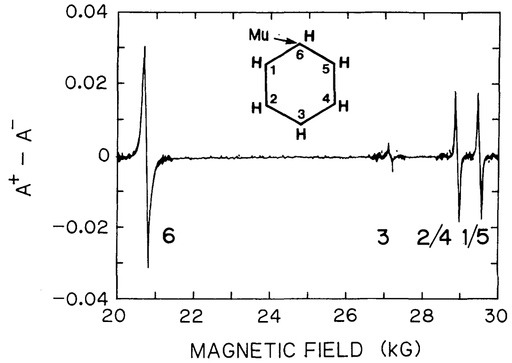

FIG. 4. Muon level-crossing spectrum of C_6H_6Mu in neat benzene at 300 K (Percival, 1987). The muon spin polarization is seen to be depolarized resonantly at four magnetic field intensities, from which the hyperfine frequencies are calculated by using Eq. (16) as 126.1 MHz for H(6), −25.1 MHz for H(1,5), 7.47 MHz for H(2,4), and 36.2 MHz for H(3), where H(*n*) indicates the hydrogen atom position.

semiconductors, and this can be done experimentally only by using the muonium probe. Comprehensive studies have been carried out, as reviewed by Patterson (1988). The μSR frequency spectrum at high transverse field for all the diamond-lattice semiconductors at low temperatures reveals the presence of not only a normal Mu but also an anomalous one, characterized by very low frequencies that depend on the crystal orientation with respect to the applied field. The hyperfine structure of the normal Mu is isotropic, while that of the anomalous one, which is denoted by Mu*, is anisotropic. The experimental data are well reproduced (Patterson *et al.*, 1978) by introducing an anisotropic hyperfine coupling term [generalization of Eq. (1)],

$$A_\perp(I_xS_x + I_yS_y) + A_\parallel I_zS_z. \tag{17}$$

The hyperfine parameters $A_\perp$ and $A_\parallel$ obtained experimentally for typical solids are presented in Table 4.

The level-crossing resonance method was applied to Mu* in GaAs, and the hyperfine coupling constants of the nearest-neighbor ^{75}As and ^{69}Ga were determined (Kiefl *et al.*, 1987). From this information, the site of Mu* was unambiguously determined; while the normal Mu is in a tetrahedral interstitial site, the Mu* is located at or near the bond center with the unpaired electron centered on the two nuclei. Similar information was obtained also for Si using the hyperfine coupling with ^{29}Si. The dynamical behavior with characteristic temperature dependences of Mu and Mu* in semiconductors is an interesting problem.

The relaxation of the Mu signal in a solid is related to Mu diffusion and trapping characteristics. In quartz (Brewer, 1981), the small Mu relaxation rate ($\lambda = 0.2\ \mu s^{-1}$) below 80 K is due to an interaction of Mu with the nuclear dipole moment of ^{29}Si, whose natural abundance is 4.7%. Above 80 K, the relaxation rate increases rapidly with temperature until the precession signal becomes unobservable around 140 K. This relaxation is caused by Mu diffusion among sites that are equivalent but have different orientations of the principal axes of the hyperfine tensor. Above 150 K, the precession signal becomes observable and its relaxation rate decreases

Table 4. Typical Mu and Mu* states in solids. The hyperfine constants given here are at the lowest temperatures.

		Hyperfine constant		
Substance	State	A	$A_\parallel$	$A_\perp$
Quartz	Mu	4463		
Diamond	Mu	3721(21)		
	Mu*		167.983(57)	−392.586(55)
Si	Mu	2006.3(2)		
	Mu*		16.819(11)	92.59(5)
Ge	Mu	2359.5(2)		
	Mu*		27.269(13)	131.037(34)
GaP	Mu	2914(5)		
	Mu*		219.0(2)	79.48(7)
GaAs	Mu	2883.6(3)		
	Mu*		217.8(2)	87.74(6)
ZnS	Mu	3547.8(3)		

with temperature because of a "motional narrowing" effect. The temperature dependence of the Mu and Mu* relaxation rates has also been extensively studied for Si and Ge. The dynamic change of a Mu state to another Mu or μ^+ state can be studied by the rf resonance method under a longitudinal field (Nishiyama *et al.*, 1985).

2.3 Quantum Diffusion of Muonium

The diffusion of muons and muonium is expected to show anomalous quantum behavior because of the muon's unique mass, which is only $\frac{1}{9}$ of the hydrogen mass. Kagan and Klinger (1974) predicted a divergent behavior ($T^{-\alpha}$ dependence, $\alpha = 7$–9) of the diffusion rate toward $T \rightarrow 0$. The diffusion of the μ^+ in metals shows such a unique behavior characteristic of quantum diffusion, namely, the diffusion rate (probed by nuclear dipolar relaxation) decreases with the decrease of temperature, but reaches a minimum at around 70 K and then increases according to a power law with $\alpha \sim 0.6$ (Kadono *et al.*, 1985). The small power was explained as the effect of conduction electrons by Kondo (1984) and by Yamada (1984). A longitudinal-field spin-relaxation experiment on Mu in alkali halides by Kiefl *et al.* (1989) revealed an even more dramatic effect with a sharper increase in the diffusion rate with $\alpha \sim 3$ toward lower temperature, as shown in Fig. 5.

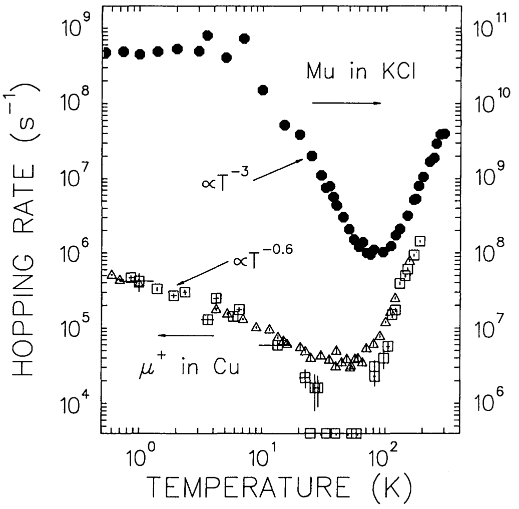

FIG. 5. Temperature dependence of the diffusion rate of Mu in KC*l* (Kiefl *et al.*, 1989) and μ^+ in Cu metal (Kadono *et al.*, 1985). Each shows a minimum at around 70 K, characteristic of quantum diffusion, though the power parameter in the low-temperature regime is different.

The Mu was also observed to be diffusing in ice at temperatures as low as 8 K; the low-temperature diffusion is considered to be due to quantum tunneling (Leung *et al.*, 1987). Muonium radicals can traverse a chain of $(CH)_n$ bonds. In *cis*-polyacetylene, Mu radicals are stable, while in *trans*-polyacetylene, they traverse as a soliton wave (Nagamine *et al.*, 1984).

3. BASIC PROPERTIES OF POSITRONIUM

3.1 Structure and Annihilation of Positronium

Positronium comprises a positron and an electron and is the lightest hydrogenlike atom. Since its reduced mass ($\mu = m_e/2$) is approximately half that of the H atom, the fundamental atomic parameters can be compared to those of the H atom through a scaling factor of 2. Hence, its ionization potential is 6.8 eV, the Bohr radius is 0.105 nm, and so on (see Table 2). Since its center of mass is situated just between the positron and the electron, its Bohr radius is, itself, the diameter, and there is no nucleus on which the center of mass can be located; i.e., the Born–Oppenheimer approximation cannot be applied. Because of the extremely small mass and absence of a nucleus, the chemical properties of the Ps are substantially different from those of the Mu and H atoms.

Ps is a system of two Dirac particles, and the spin Hamiltonian has four eigenstates as in the cases of H and Mu (Fig. 1). The singlet state ($|0,0\rangle$) is called parapositronium (*p*-Ps), and the triplet states ($|1,-1\rangle$, $|1,0\rangle$, and $|1,1\rangle$) are called orthopositronium (*o*-Ps). Following the selection rule of positron annihilation (see POSITRON ANNIHILATION SPECTROSCOPY), *p*-Ps annihilates with an intrinsic lifetime of

$$1/\lambda_S = 125 \text{ ps} \tag{18}$$

into two gamma rays, each having an energy of 511.0 keV. The *o*-Ps annihilates with an intrinsic lifetime of

$$1/\lambda_T = 142 \text{ ns} \tag{19}$$

into three gamma rays with a continuous energy spectrum ranging from 0 to 511.0 keV.

Various reactions can affect Ps annihilation, and the processes leading to a reduction in the *o*-Ps lifetime are collectively called "quenching." "Pickoff" annihilation is an important quenching process which takes place whenever an *o*-Ps is produced in matter, since the positron in the *o*-Ps can annihilate with one of the surrounding electrons having a singlet spin configuration to the e^+. The rate of pickoff annihilation is proportional to the overlap integral of the wave function of the outer electrons with the e^+. Although the pickoff annihilation mainly controls the lifetime of the *o*-Ps in condensed matter, it only slightly affects the lifetime of the *p*-Ps, since the *p*-Ps annihilation rate is already fast.

Other quenching processes are Ps compound formation, oxidation, and spin-conversion reactions. Examples are

$$o\text{-Ps} + \text{Cl} \rightarrow \text{PsCl} \quad \text{(compound formation)}, \tag{20}$$

$$o\text{-Ps} + \text{Fe}^{3+} \rightarrow e^+ + \text{Fe}^{2+} \quad \text{(oxidation)}, \tag{21}$$

and

$$o\text{-Ps} + \text{Fe}^{2+} \rightarrow p\text{-Ps} + \text{Fe}^{2+} \quad \text{(spin exchange)}. \tag{22}$$

The *o*-Ps lifetime can also be reduced in a magnetic field; this effect is called "magnetic quenching." Since e^+ and e^- have the same mass but opposite signs of the electric charge, the consequence of the Zeeman effect on the Ps involves a mixing of only the two $m = 0$ spin states, $|1,0\rangle$ and $|0,0\rangle$. At low fields, the field perturbed annihilation rates are approximated as

$$\lambda(|1,0\rangle) \sim \lambda_T + x^2\lambda_S/4 \tag{23}$$

and

$$\lambda(|0,0\rangle) \sim \lambda_S, \tag{24}$$

and at high fields, both reach the same rate,

$$\lambda(|1,0\rangle) = \lambda(|0,0\rangle) = (\lambda_S + \lambda_T)/2. \tag{25}$$

Here, $x = B/B_0$ is the field intensity measured in units of the internal field of the Ps, as in the case of Mu, Eq. (7),

$$B_0 = A/2g_e\mu_e. \tag{26}$$

3.2 Formation of Positronium

Positronium is formed when a positron picks up an electron from the substance. The simplest case is electron abstraction in a positron-molecule collision:

$$e^+ + M \rightarrow \text{Ps} + M^+. \tag{27}$$

The threshold (E_{th}) of this reaction is 6.8 eV below the ionization potential (I_M) of molecule M; $E_{\text{th}} = I_M - 6.8$ eV. Its cross section is of the same order as those of excitation and ionization. Ps formation through reaction (27) is considered to be efficient when the energy of the incident e^+ is in the Øre gap, i.e., between E_{th} and the first electronic excitation energy (E_{ex}), where Ps formation is the only possible electronic process. The idea that Ps formation takes place in the energy region of the Øre gap is called either the "Øre model" or the epithermal model (Goldanskii, 1968). This model is considered to be valid for rare gases, where the e^+ energy loss, being due solely to elastic collisions, is slow in the energy region of the Øre gap.

There is another model, called the spur model, where the Ps is considered to be formed as a result of the recombination of an e^+ with one of the excess electrons (e^-_{excess}) generated in the terminal positron spur (Mogensen, 1974),

$$e^+ + e^-_{\text{excess}} \rightarrow \text{Ps}. \tag{28}$$

Since excess electrons are involved in this mechanism, Ps formation should parallel the observations of radiation chemistry, such as electron and hole scavenging reactions as well as solvation effects. For the condensed phases of molecular substances, experimental results concerning Ps formation and radiation chemistry show strong parallelism in support of the spur model (Abbe *et al.*, 1981; Ito, 1988). In high-pressure gases, there is also some evidence for the spur reaction process (Jacobsen, 1986).

3.3 Methods for Studying Positronium

The common techniques of positron-annihilation spectroscopy, measurements of the lifetime, angular correlation, Doppler-broadened positron annihilation radiation (DBPA), and three-photon annihilation are also used in studying Ps (see POSITRON ANNIHILATION SPECTROSCOPY for the details of the techniques). Examples of the lifetime and angular correlation spectra in water are shown in Fig. 6. The lifetime spectrum is decomposed into the *p*-Ps, e^+, and *o*-Ps components in the order of increasing lifetime. In the angular correlation spectrum, e^+ annihilation and *o*-Ps pickoff annihilation produce broad components due to large momentum of the annihilating electron. Annihilation of *p*-Ps gives a narrow component, since its center-of-mass momentum is small. The energy of the DBPA carries momentum information that is equivalent to that obtained from an angular correlation measurement; however, because of insufficient energy resolution, DBPA is used only complementarily. The magnetic quenching effect provides an excellent means to confirm the existence of Ps. In special cases, the formation of an excited Ps is detected by measuring the Lyman-α photons (243 nm) of the Ps (Mills *et al.*, 1975).

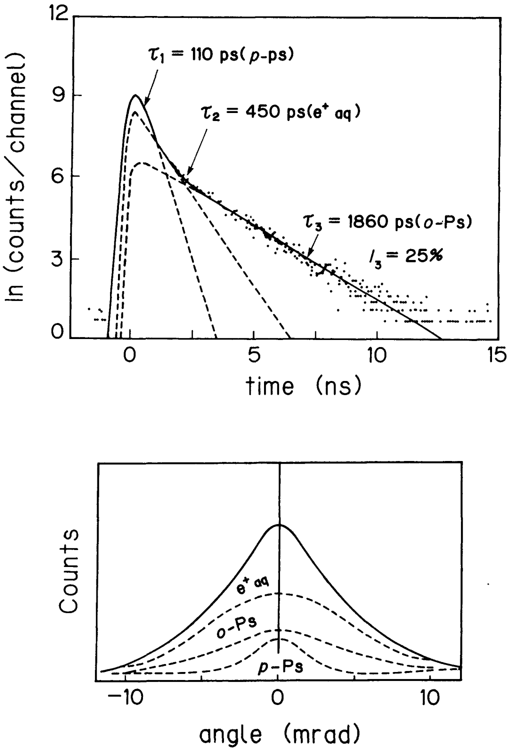

FIG. 6. Typical lifetime spectrum and angular correlation of positron-annihilation gamma rays in water. The observed spectra (solid lines) are decomposed into three components (broken lines) corresponding to different positron states: *p*-Ps, solvated positron e^+_{aa}, and *o*-Ps.

4. BEHAVIOR OF POSITRONIUM IN MATTER

4.1 Positronium in Gases

The positron lifetime spectrum in gases consists of three components: *p*-Ps, e^+, and *o*-Ps. The *p*-Ps annihilates with an intrinsic lifetime of 125 ps. The *o*-Ps annihilation rate is the sum of its intrinsic rate (λ_T = 1/142 ns^{-1}) and the pickoff annihilation rate, which is proportional to the gas density (n_M),

$$\lambda_t = \lambda_T + \pi r_e^2 \, c n_M {}^1Z_{\text{eff}}. \tag{29}$$

Here, ${}^1Z_{\text{eff}}$ is the effective number of electrons having a singlet spin configuration to e^+ per molecule, and $4{}^1Z_{\text{eff}}$ is to be compared with the total number of electrons (Z) in a molecule. In many cases, $4{}^1Z_{\text{eff}}$ is smaller than Z (see Table 5), since the e^+ in the Ps cannot penetrate into molecules because of a repulsion incurred by its partner electron. When $4{}^1Z_{\text{eff}} > Z$, as for O_2 and CCl_2F_2, it is an indication of a chemical reaction between the *o*-Ps and the molecules.

Positrons that have not formed a Ps further slow down and reach equilibrium annihilation. The shoulder in Fig. 7 corresponds to annihilation during this slowing-down process, and the width (τ_s) is the slowing-down time to the equilibrium state, which is inversely proportional to the gas density. This shoulder is observed mainly in monatomic gases; the slowing down is normally too fast to be observed in polyatomic gases. The equilibrium annihilation rate is proportional to the gas density as

$$\lambda = \pi r_e^2 \, c n_M Z_{\text{eff}}, \tag{30}$$

where Z_{eff} is the effective number of electrons per molecule. Normally, Z_{eff} is greater than the total number of electrons in a molecule, because the electron density at the e^+ is en-

Table 5. Annihilation parameters of e^+ and o-Ps in selected gases.

Gas	Molecular density (Amagat)	No. of electrons Z	$^1Z_{\text{eff}}$	Z_{eff}	$\tau_s n_M$ (Amagat · ns)
He	18–60	2	0.35	3.0	1600
Ne	7–39	10	0.23	6.0	1700
Ar	1–280	18	0.35	18–27	360
Kr	1–117	36	0.45	40–67	290
Xe	1–6	54	1.03	320	200
H_2	19–170	2	0.186	12–13	—
N_2	1–234	14	0.26	19–31	—
CO_2	1–48	22	0.48	50–120	—
O_2	7–215	16	14–86	19–26	—
NO_2	—	19	~10	—	—

hanced by e^+ interaction with the induced electric dipole moment. When Z_{eff} is far greater than Z, this is an indication of a strong interaction between the e^+ and the molecule.

The linearity in n_M of Eqs. (29) and (30) holds over a wide range of molecular densities. At high densities, both the o-Ps and e^+ lifetimes deviate from the linearity as a result of their localization. At high densities, a Ps can be localized in a hole that is made by the Ps's repulsive force. This is called a "Ps bubble" (Ferrell, 1957); the o-Ps lifetime becomes longer than that in the average gas density. The size of a Ps bubble in liquid He is about 1.5 nm, and the o-Ps lifetime is ≈100 ns. Positrons can be localized in a cluster or "droplet" of molecules formed by the positron's attractive force, in which case the e^+ lifetime becomes shorter.

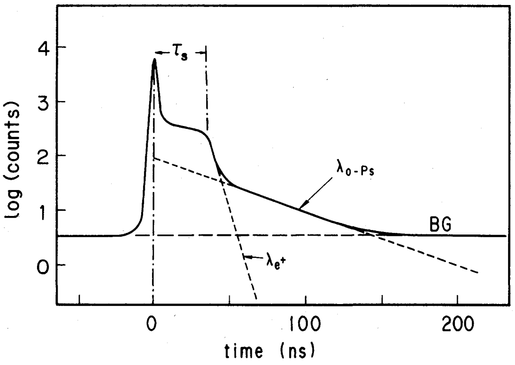

FIG. 7. Positron lifetime spectrum in He gas at 7.5 K (Canter *et al.*, 1975). The prompt component contains p-Ps annihilation and other fast annihilation in the source and wall of the chamber, and the longest-lifetime component corresponds to the annihilation of o-Ps. The intermediate component is due to annihilation of e^+, whose lifetime spectrum shows a nonexponential shoulder (time width τ_s) followed by an exponential decay corresponding to e^+ annihilation in equilibrium state.

A study of Ps-gas interactions (Griffith and Heyland, 1978) involves the difficulty of stopping high-energy positrons in a gaseous media; it has been customary to use high-pressure gases. By using silica aerogel, a low-density aggregate of ultrafine silica powder, as the media for producing Ps, it is possible to inject positrons in low-pressure gases at near to 1 atm (Kakimoto *et al.*, 1990). By combining this method of Ps formation with the magnetic quenching technique, the process of Ps thermalization at very low energies (from several tenths of eV down to thermal) was studied, and the momentum transfer cross section was obtained.

4.2 Positronium in Insulators

Ps is formed in most insulating materials. When a p-Ps annihilates from its free state in crystals of insulating materials, the p-Ps exhibits a sharp, narrow component corresponding to its nearly zero momentum. Satellite peaks are also observed in single crystals of ice (Mogensen *et al.*, 1971; West *et al.*, 1981), quartz (Greenberger *et al.*, 1970; Berko *et al.*, 1977), and ionic crystals (Kasai *et al.*, 1988) at the projections of the reciprocal-lattice vectors, reflecting the symmetry of the crystal lattice. When p-Ps annihilates from a trapped state, the width of the peak becomes broader as a result of the zero-point momentum. The o-Ps is mostly identified from its long lifetime. The long lifetime, however, is not direct evidence for the existence of o-Ps, and for a critical identification, it is necessary to resort to the magnetic quenching effect.

There are at least three components in most

Table 6. Lifetimes of e^+ and *o*-Ps in selected substances.

Substance	e^+ lifetime (ns)	*o*-Ps lifetime (ns)	*o*-Ps intensity (%)
Al	0.166	—	No Ps formation
Si crystal	0.220	—	No Ps formation
Si monovacancy	0.270	—	No Ps formation
Si divacancy	0.320	—	No Ps formation
H_2O	0.4	1.85	28
CH_3OH	0.4	3.9	21
Hexane	0.4	3.9	42
Benzene	0.45	3.15	43

lifetime spectra of liquids and solids of insulators. They are *p*-Ps, e^+, and *o*-Ps in the order of increasing lifetime (see Fig. 6 and Table 6). When different phases are contained, as in polymers, there can be more positronium components. The longest lifetime component is normally due to *o*-Ps pickoff annihilation in the holes of insulators. Because of an exchange interaction between e^- in Ps and the electrons in the surrounding molecules, Ps is repelled to free volume regions. The *o*-Ps lifetimes in the holes in solids and in bubbles in liquids are well correlated with the radius (R) as

$$\frac{1}{\tau_3} = 2\left[1 - \frac{R}{R + \Delta R} + \frac{1}{2\pi}\sin\frac{2\pi R}{R + \Delta R}\right] \text{ns}^{-1}, \tag{31}$$

which describes the pickoff annihilation of *o*-Ps based on the overlap probability of the e^+ in *o*-Ps with one of the electrons in the inner surface of the hole with an effective thickness of ΔR (Tao, 1972; Eldrup *et al.*, 1981). In Fig. 8, the experimental and the theoretical [Eq. (31)] results for the correlation between the hole size and the *o*-Ps lifetime are shown. The *p*-Ps also enters the same holes, but mostly annihilates by itself. In the self-annihilation of *p*-Ps, the momentum associated with the zero-point oscillation in the holes is reflected in the linewidth of the angular correlation or the Doppler-broadened annihilation radiation. For a spherical potential well of radius R with infinite height, the width of the angular correlation curve (Γ (FWHM)) is given by

$$\Gamma R = 1.66 \text{ mrad nm}. \tag{32}$$

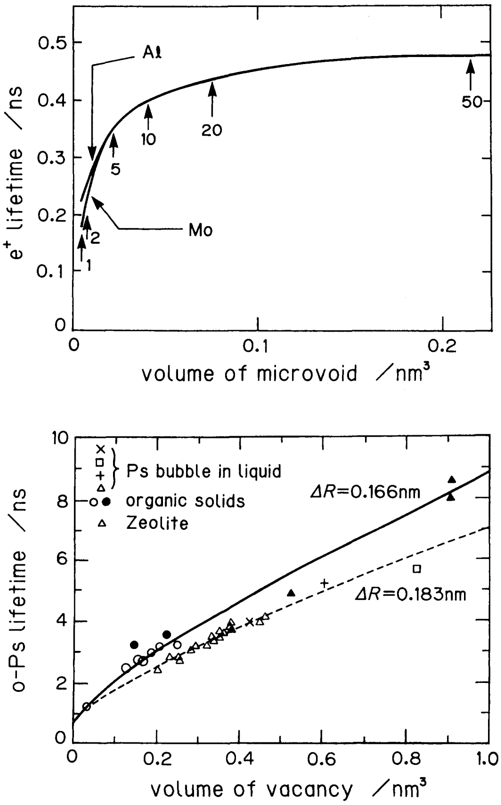

FIG. 8. Sensitivities of e^+ lifetimes toward microvoid size in metals (upper) and of *o*-Ps lifetimes toward vacancies or defects in insulators (lower). The upper figure is the theoretical expectation for Mo and Al (reproduced from Hautojarvi *et al.*, 1977), and the number in the figure describes the number of monovacancies contained in the microvoid. The lower figure contains experimental data points together with the results of fitting Eq. (31) (the solid and the broken lines) with the parameters indicated (Nakanishi *et al.*, 1988).

4.3 Positronium in a Metal Surface and in Voids

Positronium is not formed in the bulk of metals and alloys because of a shielding effect of the free electrons, and the positron exists as a Bloch state with the existence probability peaked at interstitial sites or, in the presence of vacancy-type defects, efficiently trapped in one of them. Because of the sensitivities of positron annihilation to the electron density and to the electron momentum, trapping results in an increased positron lifetime and a decreased momentum distribution (see Positron Annihilation Spectroscopy).

Positronium has been found in the voids of Al, Nb, and V (Jensen *et al.*, 1985; Hasegawa *et al.*, 1989), and the Ps states are known to be strongly affected by the internal and surface conditions of the voids where gas impurities are normally contained. Positronium is also formed when an e^+ is re-emitted from the surface of metals, by picking up a free electron (Mills, 1983). When an e^+ is injected at a low glancing angle to a metal surface, Ps is also produced by an electron pickup reaction (Gidley *et al.*, 1987). The Ps formation potential (E_{Ps}) at a surface is given by

$$E_{\mathrm{Ps}} = \phi_+ + \phi_- - 6.8\ \mathrm{eV}, \qquad (33)$$

where ϕ_+ and ϕ_- are the work functions of the positron and electron, respectively.

5. ADVANCED SOURCE OF MUONS AND POSITRONS

5.1 Ultralow-Energy Muons and Muonium Beams

When pions decay in flight, high-energy muons are produced. Muons with a smaller energy can be obtained when pions are stopped and made to decay near to the surface of the pion-producing target. The energy of these muons (so-called surface muons) is 4.1 MeV, which is normally sufficiently low for experiments in which a high μ^+ stopping rate in thin targets or in gases is required. For future applications, however, muons with a much lower energy are desired. An attempt to extract positive muons from the surface of materials, like extracting positrons of negative work function, has been made, and muons of about 10 eV were found to be emitted from a thin surface of rare gas solids (Harshman *et al.*, 1986, 1987). This phenomenon is interpreted as being due to the large band gap of the surface material; the ionization loss process in the final stage of muon stopping ceases, and still energetic (~10 eV) muons come out of the surface. The typical yield in solid Ar is around 10^{-5} per incident surface μ^+. Very recently, Morenzoni *et al.* (1994) observed that these μ^+'s retain the initial polarization.

Muonium can be produced as a very low-energy beam by an electron pickup reaction of low-energy μ^+,

$$\mu^+ + M \rightarrow \mathrm{Mu} + M^*. \qquad (34)$$

Such a scheme was used for studies of the muonium Lamb shift (Oram *et al.*, 1984; Woodle *et al.*, 1990), though the intensity is very small.

A striking phenomenon has been found: when μ^+ is stopped in a hot W target (Mills, *et al.*, 1986) or in $\mathrm{SiO_2}$ powders (Beer *et al.*, 1986), a fraction of the μ^+'s that have diffused to the surface is evaporated in the form of thermal-energy Mu. The yield of the thermal Mu is large, of the order of 10% of the incident surface muons. The thermal Mu emitted into a vacuum is important and useful, not only for fundamental studies (QED tests), but also as a possible source of a slow μ^+ beam, which may be formed after the ionization of thermal Mu by resonant multiphoton absorption (Chu *et al.*, 1988). A dedicated attempt to produce an ultralow-energy μ^+ beam is in progress (Nagamine and Ishida, 1992).

5.2 Advanced Techniques of Positronium Studies

Because of recent developments of slow positron beam techniques (see Positron Annihilation Spectroscopy), various new techniques for Ps studies have become possible. The cross sections of positron collision, including those of Ps formation, have already been measured for many molecules (Griffith and Heyland, 1978; Humberston, 1979). Attempts are being made to enhance the intensity and quality, i.e., monochromaticity and beam size, of a slow positron beam in order to extend the technique to measurements of the differential cross sections and positron reactions. It also becomes possible to obtain an energy-controllable Ps beam. When positron beams are passed through either dilute gases or very thin foils, a Ps beam can be generated via the forward-peaked nature of the cross sections of the charge-exchange reactions [Eq. (27)]. The energy of the Ps beam can be controlled by varying the energy of the incident positrons (Charlton and Laricchia, 1990). Such a Ps beam should not only augment our understanding of various fundamental interactions of Ps, but should also be used as a probe of new techniques, such as low-energy positronium diffraction (Roellig *et al.*, 1987).

The formation of cold positrons is also an important subject. A successful technique has been to use a kind of Penning trap, the upstream part of which comprises a concentration gradient of coolant N_2 gas; the downstream comprises a low potential and a high-vacuum trapping region. Positrons were cooled by exciting the vibrational energy levels of N_2 molecules down to room temperature and were stored for more than 100 s (Surko *et al.*, 1989). Cold positrons will become a powerful tool in Ps studies, which will be extended to the field of exotic-particle science.

6. TEST OF QUANTUM ELECTRODYNAMICS

Atomic hydrogen is the simplest nuclear atom, and its hyperfine splitting constant is one of the most accurately measured physical quantities. The Lamb shifts and the 2s → 1s energy have also been precisely measured. However, a theoretical treatment of the hydrogen atom cannot preclude uncertainties resulting from the finite size, mass, and polarizability of the proton; a comparison of the experimental data with theory to better than a few ppm has thus not been possible. As a result, special interest has been given to purely leptonic atoms, Mu and Ps, which are free of any nuclear effect. Compared to the hydrogen atom, however, Mu and Ps involve more experimental uncertainties because of their short lifetimes.

The QED test can be performed by precise measurements of any observable quantities—the hf splittings, the annihilation rates, the fine structure of the excited energy levels, and so on. See, for instance, Sapirstein and Yennie (1990) in the book of Kinoshita (1990). The presently known values of the hyperfine splitting, the $2S_{1/2} \rightarrow 2P_{1/2}$ classical Lamb shift, and the $n = 2 \rightarrow n = 1$ transition energy for H, Mu, and Ps, both experimental and theoretical, are summarized in Table 7. The comparison shows that in the case of the Mu and Ps, QED has been tested to higher orders. As the experimental precision improves, further meaningful information concerning QED corrections will be revealed.

6.1 Hyperfine Structure

The best experimental value of the ground-state hyperfine frequency [$\nu_{\rm hf}$ (1s)] for muonium was obtained by an rf resonance observation of the hyperfine frequencies (ν_{12} and ν_{34}) at a high longitudinal field (Mariam *et al.*, 1982). Muonium was produced in 1.5-atm Kr gas. The experimental result has a precision of 0.04 ppm and shows excellent agreement with the theoretical value.

The most precise experiment for determining $\nu_{\rm hf}$ for positronium was carried out by Ritter *et al.* (1984). The transition from *o*-Ps to *p*-Ps caused by an rf field can be signaled by the quenching of long-lived positronium. Although the precision is 3 ppm, it reveals some discrepancy with the quoted theoretical value. This difference may be ascribed to uncalculated higher order terms of QED.

6.2 1s–2s Energy Interval and Lamb Shift

The production of thermal Mu from SiO_2 powders into a vacuum was used to observe 3-photon ionization of Mu in a pulsed muon beam at KEK (Chu *et al.*, 1988) and later at ISIS of the Rutherford Appleton Laboratory (Maas *et al.*, 1994). A pulsed laser beam in the ultraviolet range (wavelength of 244 nm) was applied to the Mu emission region. The latest experimental value is shown in Table 7. It has a precision of 0.03 ppm. This value is in excellent agreement with the calculated value of 0.002 ppm precision, which includes various higher order corrections of QED.

The 1s–2s level splitting in positronium was determined by Chu *et al.*, (1984). A pulsed laser was used to induce 3-photon ionization of Ps. For this purpose, positrons were first

Table 7. Comparison of the hyperfine coupling constants (ν_{hf}), the Lamb shifts $S = E_{2S_{1/2}} - E_{2P_{1/2}}$, and the $2s(F = 1)$–$1s(F = 1)$ energies E_{2s-1s} of the three hydrogenic atoms, between experiment and theory. The theoretical values are taken from Sapirstein and Yennie (1990).

		H	Mu	Ps
$\nu_{\text{hf}}^{(1s)}$ (kHz)	Exp.	1420 405.751 7667(9)[a]	4463 302.88(16)[b]	203 389 100(740)[c]
	Theor.	1420 405(2)	4463 303.11(1.3)	203 404 500(930)
S (MHz)	Exp.	1057.845(9)[d]	1062(13)[e]	8621.6(28)[f]
	Theor.	1057.864(11)[g]	1047.49(1)	8625.14
E_{2s-1s} (MHz)	Exp.	2 466 061 413.8(1.5)[h]	2455 529 002(80)[i]	1233 607 218.9(11)[j]
	Theor.		2455 528 934(4)	1233 607 228.4(19)

---

[a] Hellwig *et al.* (1970); Essen *et al.* (1971)
[b] Mariam *et al.* (1982)
[c] Ritter *et al.* (1984)
[d] Lundeen and Pipkin (1981)
[e] Weighted average of Oram *et al.* (1984) and Woodle *et al.* (1990)
[f] Weighted average of Mills *et al.* (1975) and Hatamian *et al.* (1987)
[g] $\langle r_p^2 \rangle^{1/2} = 0.833$ fm is assumed. The uncertainty of ± 0.030 fm in $\langle r_p^2 \rangle^{1/2}$ causes a respective error of ± 0.009.
[h] Beausoleil *et al.* (1987)
[i] Maas *et al.* (1994)
[j] Chu *et al.* (1984)

stored and then injected pulsewise into a clean Al target surface, which desorbed free thermal Ps in a vacuum.

A low-energy muonium beam (described in Sec. 5.1) contains metastable 2s muonium. Such a Mu(2s) beam was used to obtain information concerning the "classical" Lamb shift, namely, $S = E_{2S_{1/2}} - E_{2P_{1/2}}$, by applying a quenching electric field along the flight region, where Lyman-α x rays were detected. A more accurate measurement was carried out by causing the $2s \rightarrow 2p$ transition by using resonating microwaves. The average experimental value from the two experiments (Oram *et al.*, 1984; Woodle *et al.*, 1990) is compared with the theoretical value in Table 7. The present experimental value of 0.13% accuracy does not allow a precision test of various corrections of QED.

The classical Lamb shift in positronium was measured by using a positronium beam containing Ps(2s) (Mills *et al.*, 1975; Hatamian *et al.*, 1987). The microwave resonance was probed by the emission of Lyman-α x rays. The experimental value with 300-ppm accuracy is in good agreement with theory.

6.3 Positronium Annihilation Rate

The theoretical annihilation rate of p-Ps (Harris and Brown, 1957) is

$$\Gamma(1^1S_0 \rightarrow 2\gamma) = \frac{\alpha^5 mc^2}{2\hbar}\left[1 - \alpha\left(\frac{5}{\pi} - \frac{\pi}{4} - \tfrac{2}{3}\alpha^2 \ln \alpha + \ldots\right)\right] = 7.986654(1) \times 10^9\ \text{s}^{-1}, \qquad (35)$$

where the $\alpha^2 \ln \alpha$ term was added by Caswell and Lepage (1979). The experimental p-Ps annihilation rate obtained by a technique involving the Zeeman mixing of the two $m = 0$ states (Gidley *et al.*, 1982) is 7.994(11) ns^{-1}, in agreement with the theory.

The theoretical annihilation rate of o-Ps (Adkins, 1983; Caswell and Lapage, 1979) is

$$\Gamma(1^3S_1 \rightarrow 3\gamma) = \alpha^6 \frac{mc^2}{\hbar}\frac{2(\pi^2 - 9)}{9\pi}\left[1 - (3.273 \pm 0.001)\alpha + \tfrac{1}{3}\alpha^2 \ln \alpha + \ldots\right] = 7.03830(7) \times 10^6\ \text{s}^{-1}. \qquad (36)$$

The o-Ps annihilation rate has been measured directly by producing o-Ps in low-pressure gases or at the surface of solid materials in a vacuum; the values have been interpolated to an ideal noninteracting condition in order to obtain the intrinsic rate. All of the results are larger than the theoretical value. The most accurate value to date is 7.0482(15) $\times 10^6$ s^{-1} (Nico *et al.*, 1992), which is 6 SD above the theoretical value. This discrepancy has not yet been explained.

Works Cited

Abbe, J. Ch., Duplatre, G., Maddock, A. G., Talamoni, J., Haesler, A. (1981), *J. Inorg. Nucl. Chem.* **43**, 2603–2610.

Abragam, A. (1984), *C.R. Acad. Sci. (Paris) Ser. II* **299**, 95.

Adkins, G. S. (1983), *Phys. Rev. A* **27**, 530–532.

Beer, G. A., Marshall, G. M., Manson, G. R., Olin, A., Gelbart, Z., Kendall, K. R., Bowen, T., Halverson, P. G., Pifer, A. E. (1986), *Phys. Rev. Lett.* **57**, 671–674.

Berko, S., Maghgooie, M., Mader, J. J. (1977), *Phys. Lett.* **63A**, 335–338.

Beausoleil, R. G., MacIntyre, D. H., Foot, C. J., Hidum, E. A., Couillaud, B., Hänsch, T. W. (1987), *Phys. Rev.* **A35**, 4878–4881.

Brewer, J. H., Crowe, K. M., Gygax, F. N., Johnson, R. F., Patterson, B. D., Fleming, D. G., Schenck, A. (1973), *Phys. Rev. Lett.* **31**, 143–149.

Brewer, J. H. (1981), *Hyperfine Int.* **8**, 375–380.

Canter, K. F., McNutt, J. D., Roellig, L. O. (1975), *Phys. Rev. A* **12**, 375–385.

Caswell, W. G., Lepage, G. P. (1979), *Phys. Rev. A* **20**, 36–43.

Charlton, M., Laricchia, G. (1990), *J. Phys. B* **23**, 1045–1078.

Chu, S., Mills, A. P., Jr., Hall, J. L. (1984), *Phys. Rev. Lett.* **52**, 1689–1692.

Chu, S., Mills, A. P., Jr., Yodh, A. G., Nagamine, K., Miyake, Y., Kuga, T. (1988), *Phys. Rev. Lett.* **60**, 101–104.

Eldrup, M., Lightbody, D., Sherwood, J. N. (1981), *Chem. Phys.* **63**, 51–58.

Essen, L., Donaldson, R. W., Bangham, M. J., Hope, E. G. (1971), *Nature* **229**, 110–111.

Ferrell, R. A. (1957), *Phys. Rev.* **108**, 167–168.

Garner, D. M., Fleming, D. G., Brewer, J. H. (1978), *Chem. Phys. Lett.* **55**, 163–167.

Gidley, D. W., Mayer, R., Frieze, W. E., Lynn, K. G. (1987), *Phys. Rev. Lett.* **58**, 595–601.

Gidley, D. W., Rich, A., Sweetman, E., West, D. (1982), *Phys. Rev. Lett.* **49**, 525–528.

Goldanskii, V. I. (1968), *At. En. Rev.* **6**, 3–148.

Greenberger, A., Mills, Jr., A. P., Thompson, A., Berko, S. (1970), *Phys. Lett.* **32A**, 72–73.

Griffith, T. C., Heyland, G. R. (1978), *Phys. Rep.* **C39**, 169–277.

Harris, I., Brown, L. M. (1957), *Phys. Rev.* **105**, 1656–1661.

Harshman, D. R., Warren, J. B., Beveridge, J. L., Kendall, K. R., Kiefl, R. F., Oram, C. J., Mills, A. P., Jr., Crane, W. S., Rupaal, A. S., Turner, J. H. (1986), *Phys. Rev. Lett.* **56**, 2850–2853.

Harshman, D. R., Mills, A. P., Jr., Beveridge, J. L., Kendall, K. R., Morris, G. D., Senba, M., Warren, J. B., Rupaal, A. S., Turner, J. H. (1987), *Phys. Rev. B* **36**, 8850–8853.

Hasegawa, M., Yoshinari, O., Matsui, H., Yamaguchi, S. (1989), *J. Phys. E: Cond. Matter* **1**, SA77–SA81.

Hatamian, S., Conti, R. S., Rich, A. (1987), *Phys. Rev. Lett.* **58**, 1833–1836.

Hautojarvi, P., Heinio, J., Manninen, M., Nieminen, R. (1977), *Philos. Mag.* **35**, 973–981.

Hellwig, H., Vessot, R. F. C., Levine, M. W., Zitzewitz, P. W., Allan, D. W., Glazer, D. J. (1970), *IEEE Trans. Instrum.* **IM-19**, 200.

Hughes, V. W., McColm, D. W., Ziock, K., Prepost, R. (1960), *Phys. Rev. Lett.* **5**, 63–65.

Humberston, J. W. (1979), *Adv. At. Mol. Phys.* **15**, 101–166.

Ito, Y. (1988), in: D. M. Schrader, Y. C. Jean (Eds.), *Positron and Positronium Chemistry*, Amsterdam, Oxford, New York, Tokyo: Elsevier, pp. 120–158.

Jacobsen, F. M. (1986), *Chem. Phys.* **109**, 455–464.

Jensen, K. O., Eldrup, M., Singh, B. N. (1985), *J. Phys. F: Met. Phys.* **15**, L287–L293.

Kadono, R., Imazato, J., Matsuzaki, T., Nishiyama, K., Nagamine, K., Yamazaki, T., Richter, D., Welter, J.-M. (1985), *Phys. Lett. A* **109**, 61–64.

Kagan, Yu., Klinger, M. I. (1974), *J. Phys.* **C7**, 2791–2807.

Kakimoto, M., Hyodo, T., Chang, T. B. (1990), *J. Phys.* **B23**, 589–597.

Kasai, J., Hyodo, T., Fujiwara, K. (1988), *J. Phys. Soc. Jpn.* **57**, 329–341.

Kiefl, R. F., Kreitzman, S. R., Celio, M., Keitel, R., Luke, G. M., Brewer, J. H., Noakes, D. R., Percival, P. W., Matsuzaki, T., Nishiyama, K. (1986), *Phys. Rev. A* **34**, 681–684.

Kiefl, R. F., Celio, M., Estle, T. L., Luke, G. M., Kreitzman, S. R., Brewer, J. H., Noakes, D. R., Ansaldo, E. J., Nishiyama, K. (1987), *Phys. Rev. Lett.* **58**, 1780–1783.

Kiefl, R., Kadono, R., Brewer, J. H., Luke, G. M., Yen, H. K., Celio, M., Ansaldo, E. J., (1989), *Phys. Rev. Lett.* **62**, 792–795.

Kinoshita, T. (Ed.) (1990), *Quantum Electrodynamics*, Singapore: World Scientific.

Kondo, J. (1984), *Physica B* **125**, 279–285; **B126**, 377–384.

Leung, S.-K., Brodovich, J.-C., Newman, K. E., Percival, P. W. (1987), *Chem. Phys.* **114**, 399–409.

Lundeen, S. R., Pipkin, F. M. (1981), *Phys. Rev. Lett.* **46**, 232–235.

Maas, F. E., Braun, B., Geerds, H., Jungmann, K., Matthias, B. E., zu Putlitz, G., Reinhard, I., Schwarz, W., Willmann, L., Zhang, L., Baird, P. E. G., Sandars, P. G. H., Woodman, G. S., Eaton, G. H., Matousek, P., Toner, W. T., Towrie, M., Barr, J. R. M., Ferguson, A. I., Persand, M. A., Riis, E., Berkeland, D., Boshier, M. G., Hughes, V. W., Woodle, K. A. (1994), *Phys. Lett.* **A 187**, 247–254.

Mariam, F. G., Beer, W., Bolton, P. R., Egan, P. O., Gardner, C. J., Hughes, V. W., Lu, D. C., Souder, P. A., Orth, H., Vetter, J., Moser, U., zu Putlitz, G. (1982), *Phys. Rev. Lett.* **49**, 993–996.

Mills, A. P., Jr., (1983), "Experimentation with Low-Energy Positron Beams," in: W. Brandt, A. Dupasquier (Eds.), *Positron Solid State Physics*, Amsterdam, New York, Oxford: North-Holland.

Mills, A. P., Jr., Berko, S., Canter, K. F. (1975), *Phys. Rev. Lett.* **34**, 1541–1549.

Mills, A. P., Jr., Imazato, J., Saitoh, S., Uedono, A., Kawashima, Y., Nagamine, K. (1986), *Phys. Rev. Lett.* **56**, 1463–1466.

Mogensen, O. E., Kvajic, G., Eldrup, M., Milosevic-Kvajic, M. (1971), *Phys. Rev. B* **4**, 71–73.

Mogensen, O. E. (1974), *J. Chem. Phys.* **60**, 998–1004.

Morenzoni, E., Kottmann, F., Maden, D., Matthias, B., Meyberg, M., Prokscha, Th., Wutzke, Th., Zimmermann, U. (1994), *Phys. Rev. Lett.* **72**, 2793–2796.

Nagamine, K., Ishida, K., Matsuzaki, T., Nishiyama, K., Kuno, Y., Yamazaki, T., Shirakawa, H. (1984), *Phys. Rev. Lett.* **53**, 1763–1766.

Nagamine, K., Ishida, K. (1992), in: T. Yamazaki, K., Nakai, K. Nagamine (Eds.), *Perspectives of Meson Science*, Amsterdam: North-Holland, Chap. 14.

Nakanishi, H., Jean, Y. C. (1988); in: D. M. Schrader, Y. C. Jean (Eds.), *Positron and Positronium Chemistry*, Amsterdam: Elsevier, Chap. 5.

Nico, J. S., Gidley, D. W., Skasley, M., Zitzewitz, P. W. (1992), *Mater. Sci. Forum* **105–110**, 401–409.

Nishiyama, K., Morozumi, Y., Suzuki, T., Nagamine, K. (1985), *Phys. Lett.* **A111**, 369–372.

Oram, C. J., Baily, J. M., Schmor, P. W., Fry, C. A., Kiefl, R. F., Warren, J. B., Marshall, G. M., Olin, A. (1984), *Phys. Rev. Lett.* **52**, 910–913.

Patterson, B. D. (1988), *Rev. Mod. Phys.* **60**, 69–159.

Patterson, B. D., Hintermann, A., Kündig, W., Meier, P. F., Walder, F., Graf, H., Recknagel, E., Weidinger, A., Wichert, T. (1978), *Phys. Rev. Lett.* **40**, 1347–1350.

Percival, P. W., Kiefl, R. F., Kreitzman, S. R., Garner, D. M., Cox, S. F. J., Luke, G. M., Brewer, J. H., Venkateswaran, K. (1987), *Chem. Phys. Lett.* **133**, 465–470.

Ritter, M., Egan, P. O., Hughes, V. W., Woodle, K. A. (1984), *Phys. Rev. A* **30**, 1331–1344.

Roduner, E., Brinksman, G. A., Louwrier, W. F. (1982), *Chem. Phys.* **73**, 117–130.

Roduner, E., Percival, P. W., Fleming, D. G., Hochmann, J., Fisher, H. (1978), *Chem. Phys. Lett.* **57**, 37–44.

Roellig, L. O., Weber, M., Berko, S., Brown, B. L., Canter, K. F., Lynn, K. G., Mills, A. P. Jr., Tang, S., Viescas, A. (1987), in: J. W. Humberston, E. A. Armour (Eds.), *Atomic Physics with Positrons*, NATO Advanced Study Institutes, Series B, No. 169, New York, London: Plenum Press, pp. 233–239.

Sapirstein, J., Yennie, D. R. (1990), in: T. Kinoshita (Ed.), *Quantum Electrodynamics*, Singapore: World Scientific, pp. 560–672.

Shearer, J. W., Deutsch, M. (1949), *Phys. Rev.* **76**, 462.

Surko, C. M., Leventhal, M., Passner, A. (1989), *Phys. Rev. Lett.* **62**, 901–904.

Tao, S. J. (1972), *J. Chem. Phys.* **56**, 5499–5510.

West, R. N., Mayers, J., Walters, P. A. (1981), *J. Phys. E: Cond. Matter* **14**, 477–488.

Woodle, K. A., Badertscher, A., Hughes, V. W., Lu, D. C., Ritter, M. W., Gladisch, M., Orth, H., zu Putlitz, G., Eckhause, M., Kane, J., Mariam, F. G. (1990), *Phys. Rev. A* **41**, 93–105.

Yamada, K. (1984), *Prog. Theor. Phys.* **72**, 195–201.

Further Reading

Ache, H. J. (Ed.) (1979), *Positronium and Muonium Chemistry*, Advances in Chemistry Series Vol. 175, Washington, D.C.: American Chemical Society.

Chappert, J., Grynszpan, R. I. (Eds.) (1984), *Muons and Pions in Materials Research*, Amsterdam, Oxford, New York, Tokyo: North-Holland.

Fleming, D. G., Senba, M. (1992), in: T. Yamazaki, K. Nakai, K. Nagamine (Eds.), *Perspectives of Meson Science*, Amsterdam: North-Holland, Chap. 8.

J. W. Humberston, E. A. G. Armour, (Eds.) (1987), *Atomic Physics with Positrons*, New York, London: Plenum Press.

Kadono, R. (1992), in: T. Yamazaki, K. Nakai, K. Nagamine (Eds.), *Perspectives of Meson Science*, Amsterdam: North-Holland, Chap. 4.

Kinoshita, T. (Ed.) (1990), *Quantum Electrodynamics*, Singapore: World Scientific.

Schultz, P. J., Lynn, K. G. (1988), *Rev. Mod. Phys.* **60**, 701–779.

Schenck, A. (1985), *Muon Spin Rotation Spectroscopy*, Bristol, Boston: Adam Hilger Ltd.

Schrader, D. M., Jean, Y. C. (Eds.) (1988), *Positron and Positronium Chemistry*, Amsterdam, Oxford, New York, Tokyo: Elsevier.

Walker, D. C. (1983), *Muon and Muonium Chemistry*, London, New York: Cambridge University Press.

MUSIC, ELECTRONIC

ROBERT MOOG, *Big Briar, Inc., Asheville, North Carolina, U.S.A.*

Introduction 97
1. **Technical Principles of Musical Tone Production** 98
1.1 Pitch, Loudness, Brightness, and Timbre 98
1.2 Envelopes 99
1.3 Modulation 100
1.4 Differences between Acoustic and Electronic Musical Instruments 101
2. **Early Electronic Musical Instruments** 101
2.1 Monophonic Instruments 101
2.2 Polyphonic Instruments 103
2.3 Early Encoded-Performance Instruments 105
3. **The Impact of the Tape Recorder** 106
3.1 The Tape Recorder as a Composing Tool 106
3.2 Audio Signal-Processing Techniques 107
3.3 The Classical Tape Studio ... 108
4. **The Advent of Voltage Control** 109
4.1 Technical Principles of Voltage Control 109
4.2 Electronic Music Synthesizers 110
5. **Computer Music** 115
5.1 Early Applications of Computers 115
5.2 Sampling Technology 116
5.3 Contemporary Sound Synthesis Techniques 118
5.4 Linear Sound Processing Techniques 119
5.5 Computer Systems as Composing Tools 121
5.6 Interactive Systems for Improvising 122
5.7 Performer Interfaces and Performance Techniques 122
6. **Future Directions for Electronic Music Technology Development** 123
6.1 Models for Sound Generation and Processing 123
6.2 Electroacoustic Transducer Design 125
6.3 Real-Time Control Interfaces 125
6.4 Digital Electronics, Computer Software, and Artificial Intelligence 126
Works Cited 126
Further Reading 127

INTRODUCTION

Electronic music is both a body of technology and a medium of musical expression. Seen from the instrument designer's point of view, electronic music is a collection of electronic devices and systems for use by musicians to generate, process, and shape musical tones, and to organize the tones in time to produce musical works. From the musician's point of view, electronic music is a medium of artistic expression—a set of resources that enable one to work with sound in ways that are different from those of traditional acoustic musical instruments. Thus, our discussion of electronic music includes musical and cultural as well as technical developments.

The technology of electronic music has been under development since the beginning of the twentieth century, during which time the state of electronics has advanced through several important stages. The seminal development was the triode vacuum tube. A wide variety of oscillator and amplifier circuits were designed, and, starting in the 1920s, many novel electronic musical instruments were introduced as commercial products. Then, after World War II, high-quality tape recorders became available. In addition to conventional record-playback applications,

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

tape recorders enabled musicians to modify individual sounds, to splice sounds together with unprecedented precision, and to combine lines of sound into complete pieces of music. A few years later, transistors and then integrated circuits enabled equipment designers to increase the complexity and decrease the size and price of electronic equipment by orders of magnitude. And finally, digital computers, first as research installations and then more and more as desktop tools, have enabled musicians to use the power of digital data processing for musical ends.

For musicians, the advent of electronics bore cultural as well as technological consequences. As cultural leaders spoke glowingly of the benefits of electricity and electronics (Lenin proclaimed that "Communism is socialism with electrification"), musicians and inventors explored the new capabilities of electronics. Radio and the electric phonograph brought music of high technical quality into people's homes and fostered a common cultural datum. After World War II, experimentation with electronic music as a compositional medium accelerated as studios and research centers around the world explored the capabilities of tape recorders. At the same time, vernacular music utilized more and more electronic technology until, by the 1960s, rock and roll was virtually completely dependent on electronics. A furious spate of musical experimentation ensued throughout the world during the late 1960s and early 1970s, during which time awareness of the electronic music medium entered the popular culture. Then, during the 1980s and 1990s, the popular music industry adopted electronic music techniques, while the proliferation of personal computers and music software placed potent resources at the fingertips of professional musicians and amateurs alike. Most recently, the exponentially growing power of personal computers has made practical the integration of electronic music techniques with image and video processing, thus making possible the development and proliferation of desktop multimedia production systems.

Throughout the development of electronic music, several strands of commonality have characterized the medium, and have served to differentiate it from traditional acoustic music production. Two features of the medium have existed from the very beginning: The sounds of electronic music are produced by electronic oscillators or from electronically recorded data, and musical control over individual tones is generally specified in terms of general physical parameters such as frequency and amplitude. Other capabilities have become important more recently: data-storage capabilities of tape recorders, the ability allowed musicians by the computer media to compose music off line in a studio environment, and artificial intelligence software to establish interactive composing environments based on musical rules. Each of these features has its own set of advantages and limitations for musicians.

This article will first give some basic technical concepts and define terms related to music production. Then, in the remainder of the article, specific developments will be discussed. We will start with the first electronic musical instruments and generally follow the evolution of the medium. The evolutionary tree has two distinct branches: analog and digital. The analog branch extends to the 1970s and includes the tape recorder, "classical studio" signal generating and processing techniques, and voltage-controlled synthesizers. The digital branch begins in the 1950s and includes sound-production techniques, music composing systems, algorithmic composition, and real-time performance systems.

1. TECHNICAL PRINCIPLES OF MUSICAL TONE PRODUCTION

1.1 Pitch, Loudness, Brightness, and Timbre

Let us imagine that we hear a violinist playing a single note on the open A string as steadily as she can. We observe that the acoustic wave form thus produced repeats about 440 times per second, and has a reasonably steady amplitude. We also observe that the wave shape is complex and difficult to describe. These observations—of frequency, amplitude, and wave shape—are objective in nature.

Next, we imagine that we ask another musician to listen to the tone and to describe what he hears. He may say, "That is a violin tone. The pitch is A-440 and it is being played mezzo-forte." His observations are subjective. The variables that he uses to describe the tone—pitch, loudness, brightness, and timbre—are related but not identical to our objective variables of frequency, amplitude, and wave shape. The objective variable most closely related to pitch is average frequency, while the objective variable most closely related to loudness (dynamic level) is average amplitude.

What are the objective correlates of timbre, that property of a sound that enables a listener to identify the source from which it comes? Not too long ago, musical acousticians generally associated timbre with wave shape. Early textbooks on musical acoustics often showed oscillographs of a few cycles of a musical instrument's wave form, accompanied by text that linked the instrument's perceived timbre with specific features of the illustrated wave form. Or the spectral distribution (the wave form's Fourier transform) would be shown as a series of precise integral harmonics of the wave form's fundamental frequency, suggesting that the energy of each harmonic relative to the tone's total energy remained constant as the tone evolved.

It is now understood that wave shape *per se* (or harmonic content *per se*) is only weakly correlated with timbre. Suppose we electronically sample just one complete cycle of the violin tone, and then cause that one cycle to repeat continuously. If we play that signal for our musician-listener, he may say, "The pitch is A-440 and it is medium-loud. But it sounds like a buzzer, not a violin."

Our musician-listener is showing us that perception of timbre (tone quality) is determined not only by the steady-state shape of a sound wave form, but by how the objective variables, especially the wave form itself, change as the tone evolves. Every traditional musical instrument imparts characteristic fluctuations that we as listeners use to identify it. Put another way, if we regard a musical tone as a message quantum, then the steady-state properties of the wave form (or spectrum) are information carriers, and the temporal variations of these properties bear the musical information. Fig. 1 shows how the wave form of one particular violin tone changes as the tone evolves.

Rough relationships between musical variables and their objective correlates are shown in Table 1.

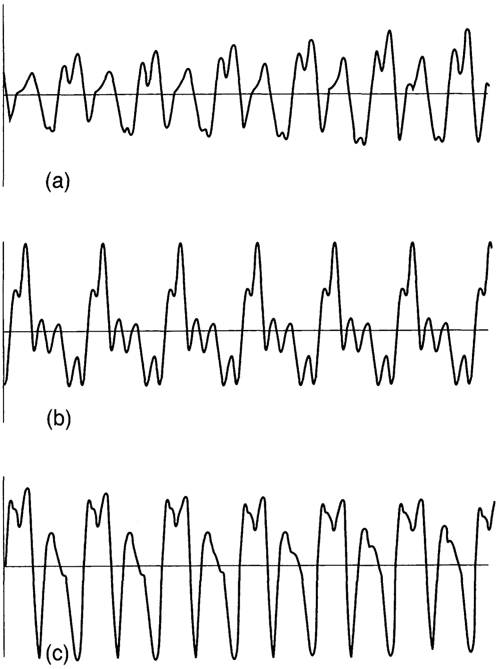

FIG. 1. Wave-form segments of a 660-Hz violin tone at three different times in its evolution: (a) 0.05 s after onset; (b) 0.50 s after onset; (c) 0.58 s after onset.

1.2 Envelopes

In general, musical tones are not steady, but evolve from onset to completion in characteristic ways. A violin tone, for instance, begins with a short burst of noise (complex,

Table 1.

Musical Variable	Objective Correlates
Pitch	Average frequency (strong correlation)
Loudness	Average amplitude (strong correlation) Average wave shape (some correlation)
Brightness	Average wave shape (strong correlation) Average amplitude (some correlation)
Timbre	Wave-shape fluctuations (strong correlation) Frequency fluctuations (some correlation) Amplitude fluctuations (some correlation) Average wave shape (weak correlation)

nonperiodic wave shape) as the bow scrapes the string, then evolves into a periodic wave form as the string begins to vibrate. As the tone evolves, the player imparts slowly varying changes to the tone's amplitude, frequency, and wave shape by changing the bow's position and velocity and the position of her finger on the string.

A graph of the temporal change of a tone's variable is called an envelope. Envelopes may describe how the overall amplitude of a tone evolves, how the amplitude of a component of the tone evolves, or how a single variable such as frequency fluctuates. Envelope features that are common to the tones of a given instrument largely define that instrument's timbre, while envelope features that vary from note to note carry the expressive nuance of the music. Four envelopes of a single cello tone are shown in Fig. 2.

Players of traditional acoustic instruments shape envelopes intuitively and continuously while they play. Electronic musicians, on the other hand, often specify envelope shapes explicitly and quantitatively, "off line," before the tones are actually produced.

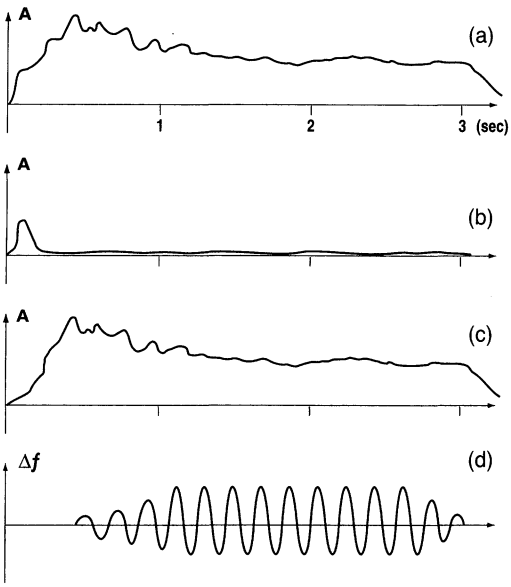

FIG. 2. Envelopes of a 3-s, 82-Hz cello tone: (a) overall average amplitude; (b) amplitude of noise (starting transient); (c) amplitude of periodic component; (d) frequency fluctuation (vibrato) of periodic component.

1.3 Modulation

In the parlance of electronic music, periodic or quasiperiodic fluctuation of a tone's variable is called modulation. Common types of modulation are vibrato, which is generated by varying frequency smoothly at a rate of about 6 Hz; trill, which is an abrupt frequency modulation; and tremolo, which is a smooth amplitude modulation. Modulations at subsonic rates create fluctuations that are slow enough for the ear to follow explicitly, whereas modulations at audio rates create sum and difference frequencies that are generally perceived as timbral changes. Frequency modulation will be discussed further below in Sec. 5.1.

It is often convenient to use envelopes to specify temporal changes of a modulation, just as it is useful to describe the evolution of a tone itself in terms of one or more envelopes. As an example, Fig. 3(a) shows the envelope of a synthetic vibrato (frequency modulation), the amplitude of which increases and

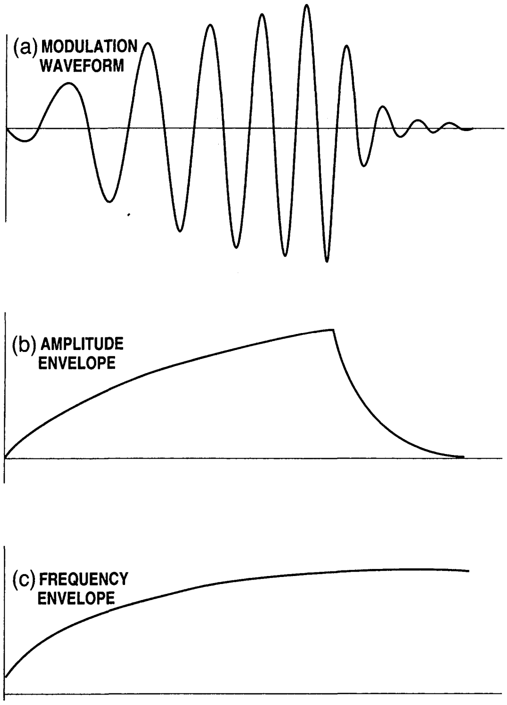

FIG. 3. Wave form of an electronically generated modulation function, with two envelopes that describe the temporal evolution of the modulation's amplitude and frequency: (a) modulation wave form; (b) amplitude envelope; (c) frequency fluctuation envelope.

then decreases, and the rate of which increases monotonically; Figs. 3(b) and 3(c) show the amplitude and frequency envelopes of the modulation. This example illustrates a hierarchy of envelopes, in which envelopes 3(b) and 3(c) specify temporal variations of function 3(a), which then defines changes in an audio signal.

1.4 Differences between Acoustic and Electronic Musical Instruments

Acoustic musical instruments differ inherently from electronic musical instruments, both in the tone-production mechanisms and in the means of musical control.

The tone-producing mechanisms of acoustic musical instruments are fundamentally different from those of electronic instruments. Acoustic musical instruments are distributed systems whose vibrations generally give rise to complex, constantly changing wave forms. Electronic oscillators, on the other hand, are generally nondistributed systems that produce simple, static wave forms.

With respect to musical control, acoustic instruments generally place the musician in intimate contact with the vibrating elements. This provides tactile feedback and enables the musician to exercise detailed control over the sound. On the other hand, electronic instruments are generally controlled through transducers that are physically separate from the tone producers, and provide little or no direct tactile feedback.

The history of the electronic music medium has been marked by ongoing technical developments to improve the musical richness of electronically produced tones, and to increase the amount of control that musicians can exert over the evolution of the tones. But it has also been marked by explorations by creative musicians who discovered new ways of producing music for which the capabilities of electronic instruments are admirably suited. The remainder of this article traces the remarkably multifaceted history of the electronic music medium.

2. EARLY ELECTRONIC MUSICAL INSTRUMENTS

Pitched musical instruments tend to fall into two classes: monophonic and polyphonic. Monophonic instruments produce only one tone at a time, but the pitch of that tone can be varied over a wide range. Polyphonic instruments can produce many tones at a time, but the pitch of each tone is fixed. The trumpet is an example of a monophonic instrument, while the piano is a polyphonic instrument. String instruments are midway between the two, in that they can produce a few tones at a time, and the pitch of each tone can be varied over a limited range.

2.1 Monophonic Instruments

In order to build a monophonic electronic instrument with a pitch range of several octaves (one octave = 2:1 frequency ratio), one needs an audio-frequency oscillator with a frequency range of at least 10:1, along with the means to change the frequency rapidly and precisely. The first instrument to meet these requirements was the Theremin, which was designed in 1920 by the Russian physicist Leon Theremin (Glinsky, 1992). Theremin used two nearly identical high-frequency (170 kHz) oscillators, plus a circuit that produced the difference (beat) frequency. By varying the frequency of one of the oscillators by only 1%, the beat frequency would change from zero to 1.7 kHz, about two octaves above A-440!

To enable a musician to control the beat frequency, Theremin connected an *LC* resonant circuit to one of the oscillators. The capacitive element of the circuit was a slim brass rod, called the pitch antenna. Bringing one's hand within a few inches of the pitch antenna would increase its capacitance by a fraction of a picofarad, lower the oscillator frequency by about 1%, and sweep the frequency of the beat tone through most of the musical pitch range!

To control the loudness of the tone, Theremin introduced a third high-frequency oscillator with its own "volume antenna" circuit, the output of which varied the gain of the instrument's audio amplifier.

Thus, Theremin not only provided an instrument with a wide pitch range but also introduced a completely novel means of control: motion of the player's hands in space, without touching the instrument at all! Professor Theremin is shown demonstrating his instrument in Fig. 4.

In 1929, Theremin licensed the Radio Corporation of America (RCA) to build and sell

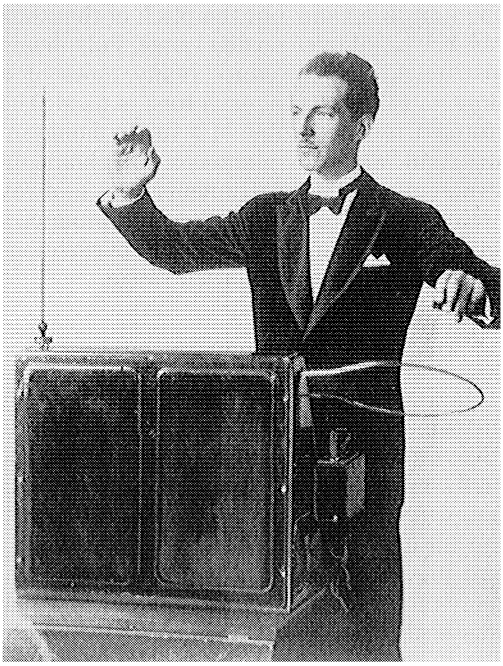

FIG. 4. Leon Theremin demonstrating the Theremin (ca. 1927).

Theremins. RCA built several hundred instruments, using the same chassis that it had designed for its superheterodyne radio receivers. Fig. 5 is a simplified block diagram of the RCA circuitry. Fig. 6 is a photo of the main chassis of the RCA Theremin.

The Ondes Martenot is another early monophonic electronic instrument that uses a beat-frequency oscillator configuration. Developed by the French engineer Maurice Martenot, the Ondes Martenot employs a six-octave keyboard in conjunction with a flexible band to enable the "ondiste" to play either discrete or continuous pitch changes. Loudness and timbre control is achieved through a series of buttons and levers, which are played with the left hand (Laurendeau, 1990).

Although the Theremin and the Ondes Martenot both use a beat-frequency oscillator for tone production, their characteristic timbres are different. This is because of the difference in the control means. The theremin responds immediately to every motion of the player's hands, and therefore has a fluid, ethereal quality. The sound of the Ondes Martenot is more steady and less fluid because mechanical elements have to be moved in order to effect pitch changes.

The Trautonium, developed by Friedrich Trautwein, is another interesting scheme to provide a wide frequency range with novel pitch-control means. The Trautonium uses a gas-filled triode as a relaxation oscillator. The frequency-determining device is a long resistive element whose resistance the player varies by pressing on a conductive ribbon that is suspended over it, then sliding her finger along the ribbon. The oscillator wave form

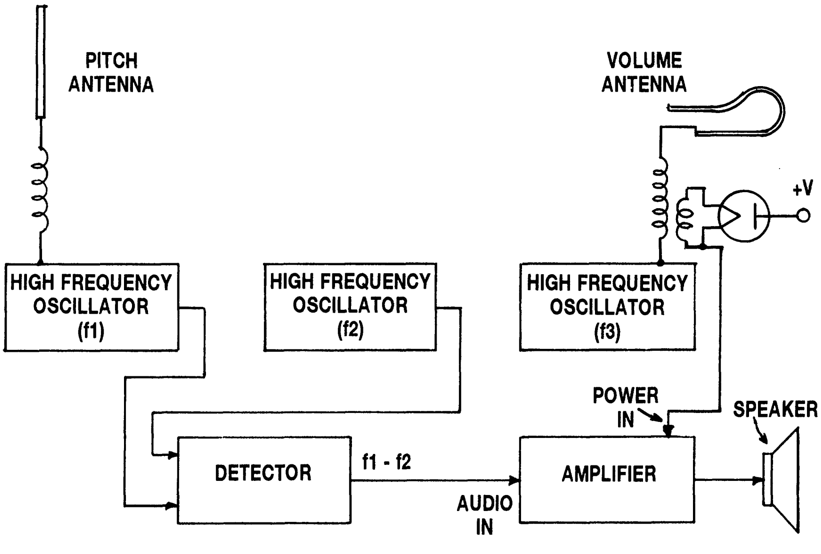
PITCH ANTENNA
VOLUME ANTENNA
+V
HIGH FREQUENCY OSCILLATOR (f1)
HIGH FREQUENCY OSCILLATOR (f2)
HIGH FREQUENCY OSCILLATOR (f3)
POWER IN
SPEAKER
DETECTOR
f1 - f2
AUDIO IN
AMPLIFIER

FIG. 5. Block diagram of the RCA Theremin (1929).

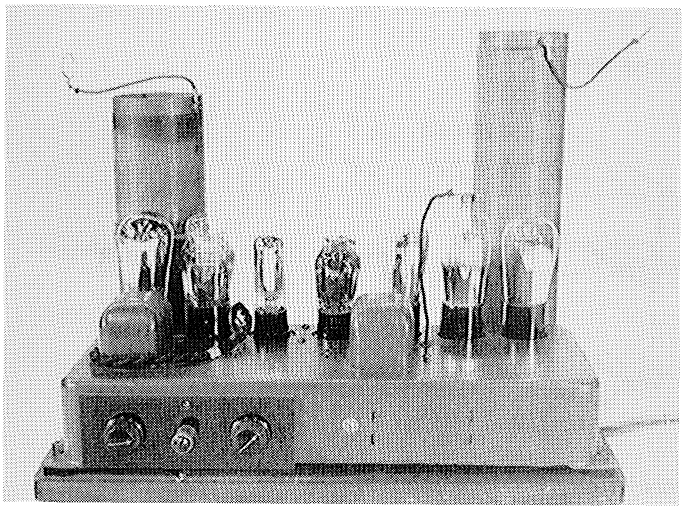

FIG. 6. Main chassis of the RCA Theremin. The volume antenna inductor is on the left, and the pitch antenna inductor is on the right.

has a sharp edge, which gives rise to the full range of harmonics. The harmonics may be filtered by one or more resonant circuits to simulate formants, the pronounced resonances in the spectra of vocal and certain acoustical musical instruments. The Trautonium was manufactured by Telefunken for a short period of time during the 1930s (Rhea, 1979b).

The Theremin, the Ondes Martenot, and the Trautonium were the three earliest electronic musical instruments to be manufactured commercially. With all three, the most distinguishing feature is the novel means with which the instrument's tone is controlled by the player. The originality and imagination displayed by the instruments' inventors appears to be a product of the optimism of the early days of electronic technology.

2.2 Polyphonic Instruments

Unlike the monophonic instruments described above, early polyphonic electronic instruments were close adaptations of their acoustic counterparts. In particular, the control interfaces of nearly all polyphonic electronic instruments were conventional organ-style or piano-style keyboards, and the instruments themselves were often called organs or pianos.

Many polyphonic instruments employed the motion of mechanical elements to generate periodic wave forms, and electronic amplification and loudspeakers to convert the wave forms into sound. The Hammond Organ, introduced commercially in 1935, used tiny rotating steel disks ("tonewheels") to generate a multiplicity of sine waves. Each key activated one switch for each of eight harmonics, while an array of drawbar switches above the keyboards determined the tones' brightness and timbre by setting the relative amplitudes of the harmonics. Vibrato was introduced with an *LC* delay line that was mechanically scanned at rates of up to 6 Hz. Overall volume was controlled with a foot-operated potentiometer. A simplified schematic is shown in Fig. 7, while an early model of the Hammond Organ is illustrated in Fig. 8.

Hammond tonewheel organs remained in production for nearly five decades, and have been widely used for popular and religious music performance (Rhea, 1977b). They have been notably ubiquitous on radio, perhaps because the instrument's audio bandwidth matches that of AM radio.

Continuous-tone polyphonic instruments using optical wave-form storage were developed during this period. In these instruments, transparent disks containing many concentric tracks were rotated at constant speed. Endless loops of tone wave forms were drawn in each of the tracks. Each disk was placed between a light source and an array of photodetectors. A mechanical shutter, opened by the depression of a key, activated a single track. Such instruments with rotating optical disks offer the advantages of ar-

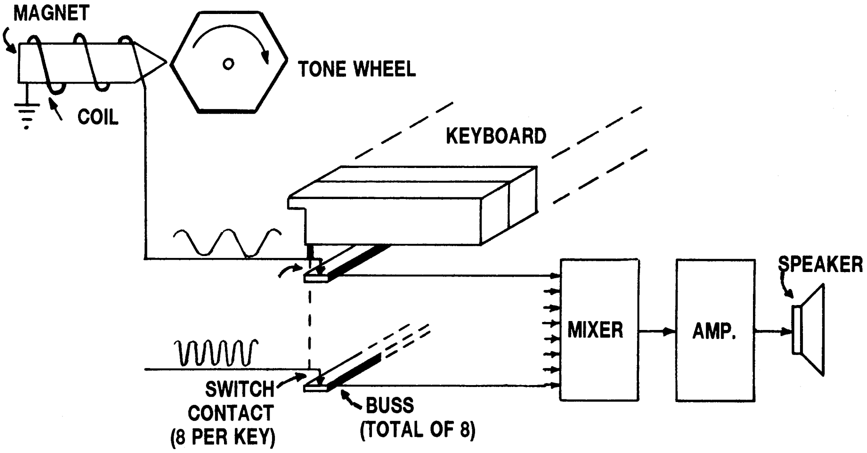
MAGNET
TONE WHEEL
COIL
KEYBOARD
SPEAKER
MIXER
AMP.
SWITCH CONTACT (8 PER KEY)
BUSS (TOTAL OF 8)

FIG. 7. Simplified diagram of the Hammond Organ.

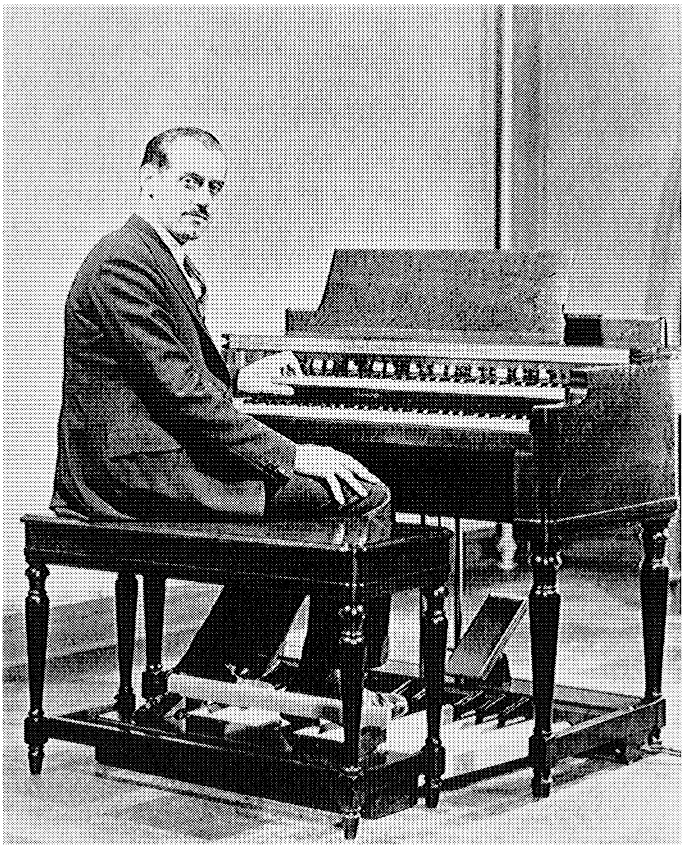

FIG. 8. The Hammond Organ (1935).

bitrarily complex wave forms. One of the earliest examples of a "photoelectric organ" is the Cellulophone, constructed during the early 1930s in France by P. Toulon (Rhea, 1977a).

Many workers developed electronically amplified pianos in the 1930s. Design considerations for these efforts included not only transducer configuration but also transducer placement and hammer design. One of the most successful developers of amplified pianos was Benjamin Miessner, who introduced the Miessner Electronic Piano as a commercial product in 1932, and then licensed several piano manufacturers to build similar instruments. Miessner also developed an amplified pianolike instrument in which brass reeds instead of strings were struck with hammers. This design formed the basis of the Wurlitzer Electronic Piano, which soon became a staple of the popular music arsenal (Rhea, 1978a).

Because of the expense and relative instability of early vacuum tubes, the development of polyphonic instruments using electronic oscillators began somewhat later in the 1930s. In France, Edouard Coupleux and Joseph Givelet built several large scale all-electronic organs that were installed in several churches and radio stations (Rhea, 1978b). In 1939, the Hammond Novachord was introduced (Fig. 9). It was the first commercial polyphonic instrument to use electronic master oscillators for the top octave of notes, to which all lower octaves were synchronized. The Novachord included means for varying amplitude envelopes and formants, and achieved considerable popularity despite its notable complexity, weight, and price (Rhea, 1979a).

2.3 Early Encoded-Performance Instruments

Encoded-performance instruments are systems with which a musician may record explicit coded "performance instructions" off line, on a medium such as punched paper, marked paper, or rigid templates, and then realize a piece of music by having the machine scan the recording medium at a specified rate. The recording medium is directly analogous to a traditional composer's score, and the machine itself takes the place of one or more traditional performers.

The earliest successful electronic encoded-

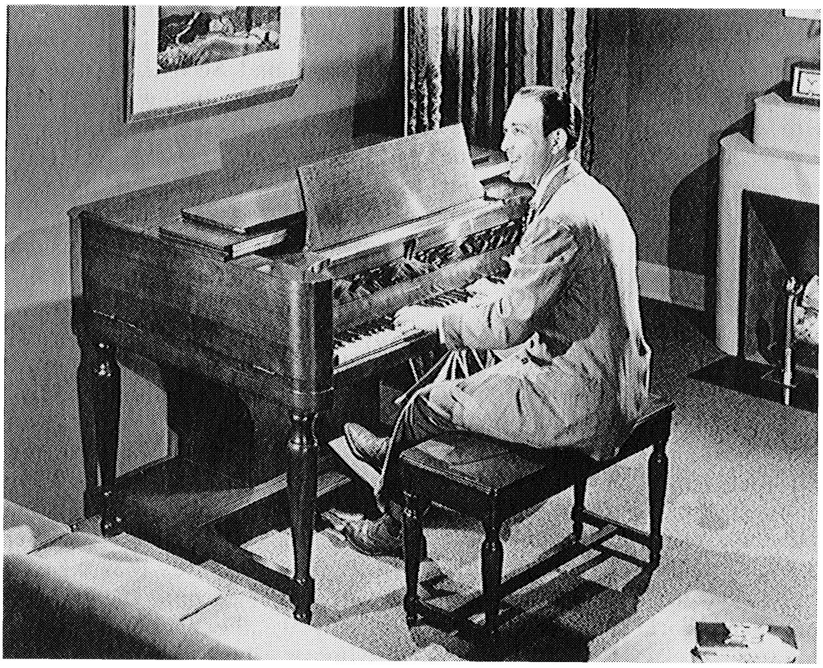

FIG. 9. The Hammond Novachord (ca. 1939).

performance instrument was built and demonstrated by Coupleux and Givilet at the Paris exposition of 1929. The device included four *LC* oscillators and related wave-shaping circuitry, which were controlled by a paper roll and a player-piano-like pneumatic mechanism. The device was capable of continuous control of frequency (by moving iron cores in and out of inductors) and amplitude (by moving potentiometers), and discrete selection of several options (Rhea, 1979c).

In 1945, John Hanert developed an experimental encoded-performance instrument for the Hammond Instrument Company. The device, which became known as the Hanert Electrical Orchestra, employed a score reader mechanism that rolled down a long table on which cards containing the score were placed. The score consisted of conductive marks that the score reader detected as it rolled along. The score reader then generated keying and control signals that activated several racks of tone-production circuitry (Rhea, 1979d).

Whereas devices like those of Coupleux–Givilet and Hanert were primarily technological showpieces, other encoded-performance instruments were designed with specific compositional goals in mind. One such device is the Free Music Machine of Percy Grainger and Burnett Cross. Grainger envisioned a type of music devoid of beats, fixed pitches, and fixed harmonies. He and Cross designed an encoded-performance instrument consisting of four oscillators, each equipped with potentiometers for continuously changing frequency and amplitude. The potentiometers were attached to levers that followed contours cut onto rolls of heavy paper. As the paper scrolled, the levers followed the contours, thereby changing the frequencies and amplitudes (Rhea, 1979e).

The above-described instruments constitute only a tiny fraction of musical instruments and automata using electronic technology that were designed and built before the end of World War II. Most were highly experimental; a few were put into commercial production. Technically speaking, most were innovative and original. However, musically speaking, most (but not all) were designed for traditional modes of composition and performance. As World War II drew to a close, the advance of electronic technology accelerated, the world of experimental music blossomed, and the electronic music medium developed into a significant and innovative component of the musical mainstream.

3. THE IMPACT OF THE TAPE RECORDER

3.1 The Tape Recorder as a Composing Tool

Tape recorders became available to musicians a few years after the end of World War II. For the first time, musicians had a high-quality sound-recording medium that was convenient to use and manipulate. With a tape recording machine, razor blade, and some splicing tape, one could connect one sound to another and perform simple temporal transformations on individual sounds. Put another way, the tape recorder was a highly versatile, easy-to-use composing tool, offering a composer access to any sound whatsoever that could be picked up by a microphone or generated in an electronic instrument.

Musician experimenters quickly found that tape manipulation offered several very useful sound transformations. First, a length of tape on which a single sound is recorded can be cut into pieces to shorten or rearrange the sound. Or, the same piece of tape can be spliced into a loop to form a continuously repeating sound. Second, the tape can be turned around and run backward, thus reversing the evolution of the sound. And third, the speed of tape travel can be changed simply by changing the surface velocity of the capstan.

Just these three simple tape manipulations placed a wealth of new sound material at the disposal of experimental composers. To illustrate, suppose we have recorded a single piano tone that lasts 2 s. The tone has an initial "thump" attack transient, after which both its amplitude and harmonic content slowly decrease. We first cut off the attack, then select a 1-s segment of the remaining sound. We rerecord that segment, giving us two identical 1-s segments. We now splice the two segments together to form a loop, in such as way that the end of one segment is spliced onto the end of the second. Finally, we play back that loop at half speed, giving us a low-pitched sonorous drone that rises and falls in amplitude and brightness every 4 s. The drone does not sound like a piano tone at all, and in fact is a completely new musical sound.

3.2 Audio Signal-Processing Techniques

As experimental musicians learned the compositional capabilities of the tape recorder, they also explored the sound modification capabilities of other electronic instruments. They found that several basic signal processors that were originally developed for laboratory, recording studio, or military use were also capable of musically useful sound transformations. The more important of these signal-processing techniques will now be described.

Mixing is the linear addition of two or more audio signals. A mixer is basically an analog adder circuit, preceded by potentiometers (often called faders) at each of the inputs. A mixer enables one to combine two or more signals; by turning one fader up and another fader down, one can "crossfade" between two sounds.

Filtering is a signal-processing operation with a device, called a filter or equalizer, whose gain is a settable function of frequency. Filters are used to emphasize certain portions of the frequency spectrum, and attenuate other portions. Filters "color" sounds that are passed through them, in direct analogy to optical filters, which reshape the spectra of light passing through them. Common types of electronic music filters include low-pass, high-pass, and parametric filters, and bandpass filter banks. A low-pass filter permits frequencies up to the "cutoff frequency" to pass relatively unattenuated, and attenuates frequencies above cutoff. Similarly, a high-pass filter attenuates frequencies below its cutoff. A parametric filter has one or more resonances, each of the center frequencies and bandwidths of which can be varied. Bandpass filter banks (commonly called graphic equalizers) consist of an array of filters that pass contiguous bands of the frequency spectrum. The gains of the filters are individually variable, often by an array of slider potentiometers that provide a graphic indication of the overall frequency response of the bank.

Filters are capable of changing the timbre of a sound dramatically, especially in cases where a filter's operating points are changed as the sound evolves.

Ring modulation is the arithmetic multiplication of two signals. The output of a ring modulator contains the sum and difference frequencies of all frequency components of the two input signals, but not the input signals themselves. In a typical application, a sound to be processed is applied to input *A* while a sine wave of frequency *f* is applied to input *B*. The output signal contains two sets of partials. (Note: The term *partial* is used here to denote *any* component frequency of a musical sound, whereas the term *harmonic* generally refers to component frequencies that are integral multiples of the sound's fundamental frequency.) One set consists of the partials of input *A* shifted up in frequency by *f*, while the other set is the partials of input *A* shifted down by *f*. The term *partial* is used here instead of *harmonic* because the frequencies of harmonics are generally assumed to be integral multiples of a tone's fundamental frequency, or nearly so, whereas a partial may be of any frequency. Musical sounds with significant "out-of-tune" partial energy (i.e., containing strong partials whose frequencies are nonintegral multiples of the fundamental) are called *clangorous* sounds. They generally have a less "pure" quality and a more vague sense of pitch than tones arising from nearly periodic wave forms. Tympani and gongs are two examples of instruments that produce clangorous sounds. Ring modulators are particularly useful for producing clangorous sounds in which the frequency relationships among its partials can be quickly and precisely varied.

Frequency shifting (Klangumwandlung) is similar to ring modulation, but the output of a frequency shifter contains only one set of shifted partials. In general, *Klangumwandler*-produced tones are less clangorous than ring modulator tones.

Ring modulation and frequency shifting provide new sounds that retain the temporal variations of an existing sound, but have a different set of partials.

Reverberation and *echo* are two closely related methods of processing a signal by repeating it at different times. Reverberation simulates the multiplicity of closely spaced reflections that one hears in an enclosed space, while echo consists of fewer, more widely spaced repetitions of the original signal. *Tape echo* takes advantage of the fact that many tape recorders have a playback head that is separate from but spaced near to the recording head. If a sound to be processed is mixed with a certain amount of the playback head signal, and the mixture is fed back to the record head, then the recorded signal includes an attenuated replica (echo) of what was recorded a short time before. The strength and duration of the echoes can be controlled by varying the gain of the feedback loop, the spacing of the echoes can be controlled by varying the tape speed or the distance between the heads, and musically interesting evolutions of the echoes can be achieved by placing signal processors such as a filter or ring modulator in the feedback loop.

Electronic switching is the processing technique by which an output signal alternates periodically between two input signals. Electronic switching can proceed so rapidly that the qualities of the original input signals are completely obscured, and a new sonic texture emerges.

3.3 The Classical Tape Studio

During the late 1940s and early 1950s, experimental composers in Europe and the United States rapidly developed techniques and gathered equipment to compose complete pieces of music directly on tape. Their experimentations were both technical and aesthetic. As these explorers worked with their newly found equipment and honed their newly learned techniques, they discovered sonic relationships that went far beyond those of traditional instrumental music. Sounds that earlier had been considered musically irrelevant noises, when treated electronically, often acquired musical substance. At the same time, composition by tape recorder allowed music production to be done with precision and versatility.

In Europe, the first institutions to support the development of the electronic music medium were the radio stations. Tape composition studios were established at the French Radio in Paris in 1948, at the Northwest German Radio in Cologne in 1951, and at the Italian Radio in Milan in 1953. At the Paris studio, Pierre Schaeffer and his colleagues focused their efforts on composition with "natural" sounds—sounds that started out as recordings of acoustical events. They called their compositional approach "Musique Concrete" (Schaeffer, 1952). Werner Meyer-Eppler, Karlheinz Stockhausen, and Herbert Eimert took a diametrically opposite approach at the Cologne studio, preferring to create their music from the simplest of electronically generated waves, a compositional approach that they called "Elektronisches Musik" (Eimert, 1956).

Tape composition developed in the United States first in the music departments of universities. Perhaps the best-known American tape music pioneer is Vladimir Ussachevsky. He began his explorations in 1951 at Columbia University, and, with his colleagues, established the Columbia Princeton Electronic Music Center in 1961 in New York City (Moog, 1981).

Professional audio tape recorders were the basic pieces of composing equipment in all of these early studios. Signal-generating and -processing devices, generally off-the-shelf test gear, standard audio broadcasting and recording devices, and war-surplus communications equipment, were mounted in racks. Inputs and outputs were wired to patch panels for convenience of interconnection. The European studios were usually built to allow several people to work simultaneously. The composer would give instructions to one or more technicians, who would then actually manipulate the equipment. In contrast, the American studios tended to be more compact, so that one person could perform all operations (Fig. 10).

The above-described studio configuration, along with the techniques that were used, came to be known as the Classical Tape Composition Studio (Ussachevsky, 1958). As consumer-grade tape recorders became available and the prices of electronic equipment (especially war-surplus equipment and equipment in kit form) dropped, small schools and private musicians as well as large institutions set up tape music studios. Most of the work done in these studios was experimental in nature; by the early 1960s, the world of commercial music had only a few brief encounters with the world of experimental electronic music.

The classical tape studio offered musicians a wealth of new sounds, plus techniques for modifying and composing sounds that excited and inspired many composers. But a growing number of composers felt a need for more control over the evolution of individual sounds. The signal-generating and -processing instruments of the classical tape

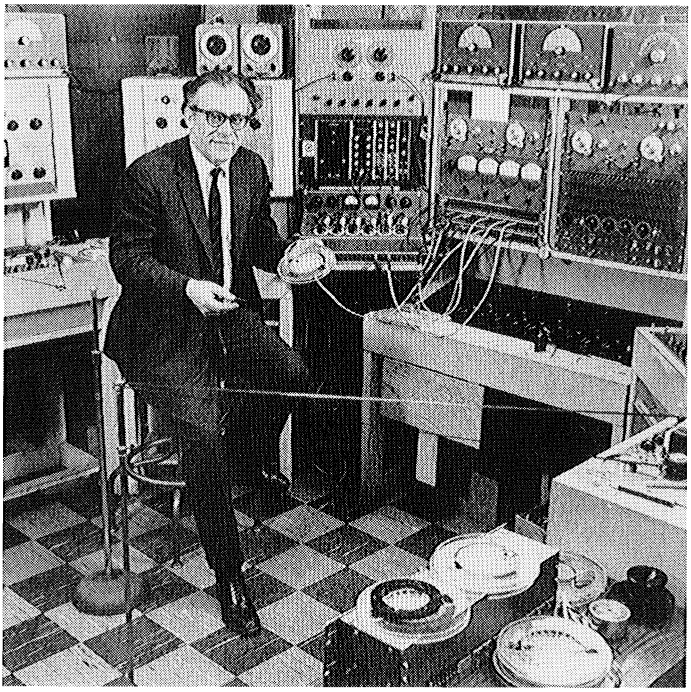

FIG. 10. Vladimir Ussachevsky in a studio of the Columbia–Princeton Electronic Music Center (1965).

studio were settable by knobs or sliders. Rapid, precise change of parameters such as frequency of filter cutoff and gain was difficult if not impossible. Furthermore, few of these instruments were optimally suited for music production. For instance, most waveform generators were standard test oscillators with frequency ranges that were switched by factors of 10—a ratio that is appropriate for technical measurements but has virtually no musical significance. Starting in the early 1960s, these problems were addressed by equipment designers who were working with the newly available silicon junction transistors.

4. THE ADVENT OF VOLTAGE CONTROL

4.1 Technical Principles of Voltage Control

A voltage-controlled device is one in which one or more of its operating points can be varied by changing the magnitude of an externally applied voltage, called the control voltage. For instance, in a voltage-controlled oscillator (VCO), a control voltage changes the frequency, while in a voltage-controlled amplifier (VCA), a control voltage changes the gain.

The introduction of inexpensive, high-quality silicon junction transistors in the early 1960s facilitated the development of voltage-controlled devices with wide range and high accuracy. Silicon junction transistors are three-terminal amplifying devices with the following remarkable characteristics:

1. The range of collector currents over which the device operates accurately is well in excess of three decades.
2. Transconductance is directly proportional to the collector current over the entire operating range.
3. The collector current increases exponentially with respect to the base-to-emitter voltage over the entire operating range.
4. The device operates with high stability and low noise.

This combination of attributes has enabled equipment designers to build multiple–wave-form VCOs that cover the entire audio-frequency range as well as musically useful subsonic frequencies, voltage-controlled filters (VCFs) whose characteristic frequencies may be swept over the entire audio-frequency range, and VCAs with dynamic ranges that approach the limits of human hearing.

The exponential relationship between input voltage change and output current change has proven to be particularly useful in voltage-controlled electronic music devices. This is because physical variables such as frequency and amplitude are generally proportional to the exponentials of the corresponding subjective musical values. In particular, musical pitch intervals are frequency ratios. Stated another way, frequency is exponentially related to pitch. Similarly, audio signal amplitude is exponentially related to subjective loudness.

A VCO in which the frequency is exponentially related to the magnitude of its control voltage turns out to be ideal for electronic music applications. In a typical VCO, a control voltage increase of 1 V doubles the frequency, or raises the pitch by one octave. A 12-tones-to-the-octave equally tempered chromatic scale is generated simply by increasing the control voltage in $\frac{1}{12}$-V steps; pitch transposition is accomplished by adding multiples of $\frac{1}{12}$ V to the control voltage that produces the pitch pattern. Furthermore, timing relationships are just as simply manipulated when a VCO is operated in the subsonic frequency region. The tempo of a subsonic wave form can be doubled by a 1-V control voltage increase. The musician thinks of the control voltage as a measure of a musical value, and the exponential conversion within a voltage-controlled device automatically produces the correct operating point.

Another desirable property of voltage-controlled devices is that their operating points can be changed essentially instantly, and can therefore be used to impart rapid, precise variations to individual sounds. The task of designing new timbres with voltage-controlled equipment thus reduces to the specification of voltage variations to change the sound's variables as individual sounds evolve. These voltage variations take the form of envelopes, modulations, random variations, and arbitrarily shaped nuances (Moog, 1965).

Wave-form–generating devices such as voltage-controlled oscillators, signal-processing devices such as voltage-controlled filters and amplifiers, and generators of envelope, modulation, and other control signals may be assembled into a system for producing sounds of considerable complexity and musical interest. Control signals may be generated automatically in real time, prerecorded as a score of an encoded-performance instrument, or shaped continuously by a musician manipulating an appropriate set of transducers. The timbres thus produced may span the range of musical tones from pitched tones like those of orchestral instruments to sounds whose partials and modes of evolution are unlike any "natural" sounds. Systems such as these became known as Electronic Music Synthesizers, and beginning in the late 1960s, became commercially important resources for experimental and commercial musicians alike.

4.2 Electronic Music Synthesizers

The term "synthesizer" was coined before the development of transistorized voltage-controlled instruments. Applied to an electronic musical instrument, the term suggests that the sounds are produced, or synthesized, from separate elements.

One of the best-known precursors of voltage-controlled synthesizers is the RCA Electronic Sound Synthesizer, a large-scale experimental encoded-performance instrument, two of which were designed and built immediately after World War II at RCA Laboratories (Fig. 11). The RCA Synthesizer uses hundreds of vacuum tubes to produce four musical sounds simultaneously. The score is a paper roll, on which rows of holes are punched to specify the frequencies, envelopes, formants, and other variables of the four tones. The time resolution of the score is $\frac{1}{30}$ s. The score's drive mechanism is mechanically synchronized with a disc-recording lathe. The lathe partitions the disc surface into six separate bands, which may be recorded on one at a time and then played back simultaneously. Thus, a completed recording may have as many as 24 perfectly synchronized tones, which would then be mixed into a final two-track tape recording (tape recorders of more than two tracks had not yet been developed) (Babbitt, 1964).

The first voltage-controlled music production systems were designed not as stand-alone encoded-performance instruments, but as components of a classical tape studio. Each of the generating and processing instruments was built as an individual module. The composer would interconnect the modules with

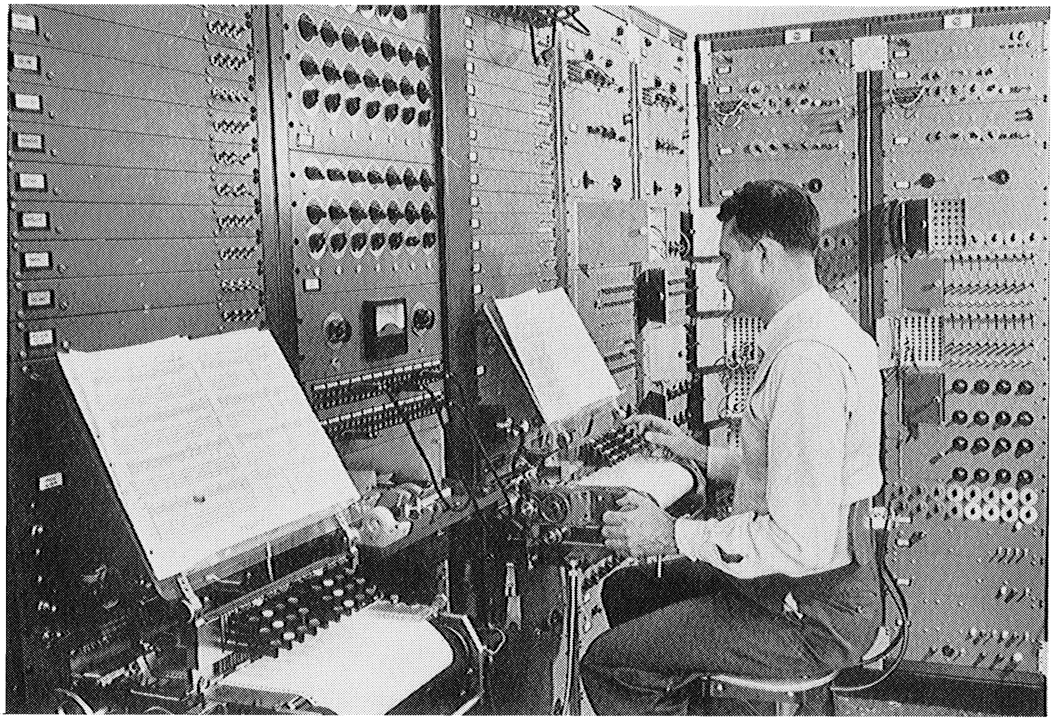

FIG. 11. The RCA Electronic Sound Synthesizer (1955).

patch cords to produce a single sound or sequence of sounds, which would then be recorded on tape (Moog, 1967). However, musicians and equipment designers alike soon realized that, by adding programming and sequencing devices to produce control voltage changes for large-scale musical gestures, one could use voltage-controlled systems to produce complete musical works with little or no tape splicing. As the 1960s progressed, commercial and popular music producers as well as experimental composers came to use modular voltage-controlled synthesizers, and several companies began manufacturing them. Fig. 12 shows a representative commercial modular voltage-controlled synthesizer.

The late 1960s was a time of widespread musical experimentation. "Synthesizer music" was frequently heard on records and radio, in live performance, and as film and television backgrounds. Among these, the record album "Switched-on Bach," by W. Carlos and B. Folkman, stands out as a watershed event. Carlos and Folkman used a modular synthesizer, a small mixing board, and an eight-track tape recorder to realize a complete album of the music of J. S. Bach (Fig. 13). Using a conventional keyboard mechanism to generate $\frac{1}{12}$-V steps for the notes of the chromatic scale, Carlos and Folkman performed and recorded one line of music at a time onto a single track of the eight-track tape. When all tracks were satisfactorily recorded, they were mixed to a two-track master tape.

A student of Vladimir Ussachevsky, Carlos had the unique talents of understanding how to produce musically appealing timbres with electronic equipment, and how to use these timbres, which were distinctly different from those of acoustic instruments, to realize traditional music. Released at the end of 1968, "Switched-on Bach" captured the public's ear and soon became the largest-selling album of Bach's music (Berger, 1969).

"Switched-on Bach," and some other synthesizer records of the 1960s, inspired many popular musicians to perform live on synthesizers. Because of their complexity, modular synthesizers were ill-suited to the demands of the stage, and so equipment manufacturers began to produce instruments

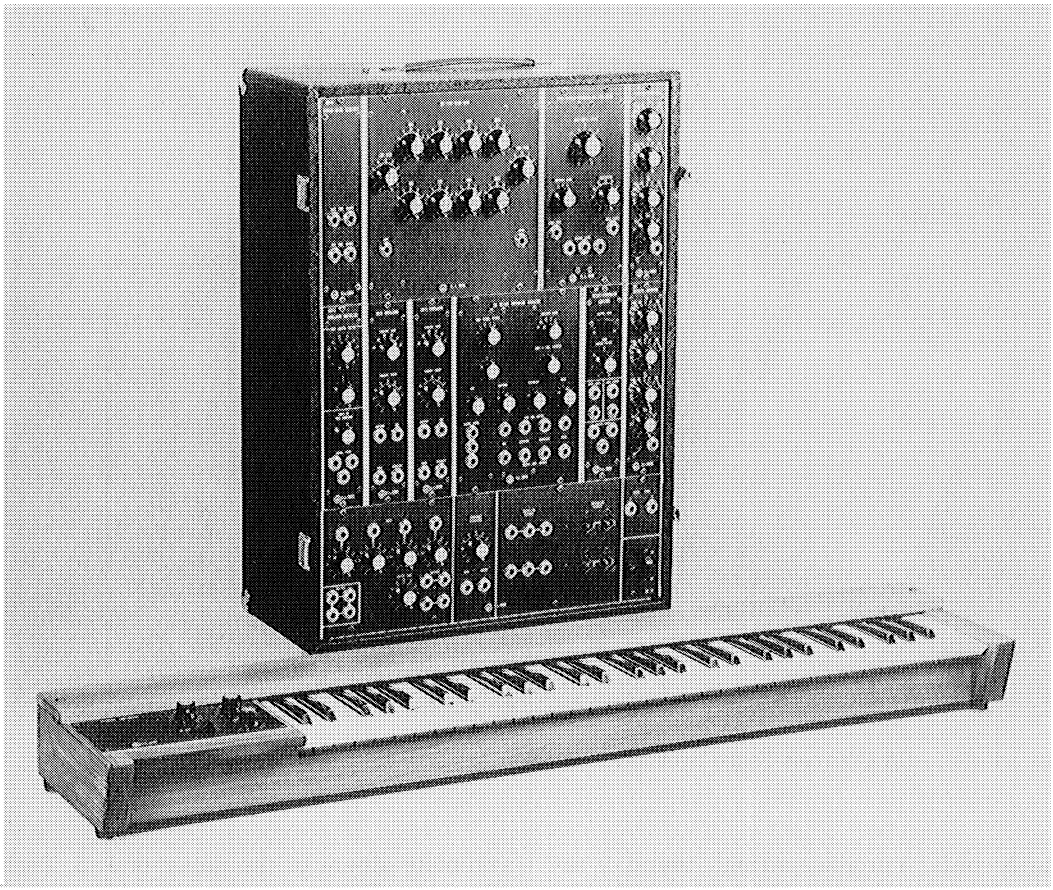

FIG. 12. A modular voltage-controlled synthesizer (ca. 1970).

with simpler capabilities, and with fixed interconnections and fast-acting switches instead of patch cords. By the early 1970s, the considerable sonic resources and flexibility of modular, patch-cord-connected systems had evolved downward to an easy-to-use but relatively standard, fixed set of functions. Two or three VCOs for pitched sounds, plus a noise generator for pitchless or weakly pitched sounds, produced the starting audio wave forms. The VCO generally produced sawtooth and rectangular wave forms for high harmonic contect, and triangular wave forms when low harmonic content was called for. One VCF, usually a four-pole low-pass filter, shaped the spectral distribution of the combined wave forms, and one or two VCAs were used to shape the overall amplitude evolution. A low-frequency oscillator (LFO) produced subsonic sine control wave forms (for textures such as vibrato and tremolo) and square control wave forms (for textures such as trill). Envelope-defining control signals were produced by envelope generators (EGs); a common envelope shape had a four-segment attack-decay-sustain-release (ADSR) configuration (Fig. 14). The instrument was played by means of a keyboard that produced a frequency control voltage according to which key was depressed, and an envelope-start signal whenever any key was depressed. Most instruments had input ports for extra control signals from devices like foot pedals and joysticks (Fig. 15).

Despite the skepticism of the musical instrument retailing community, these "compact performance synthesizers" soon became extremely popular with performing musicians. With the circuit operating points immediately accessible with front panel knobs,

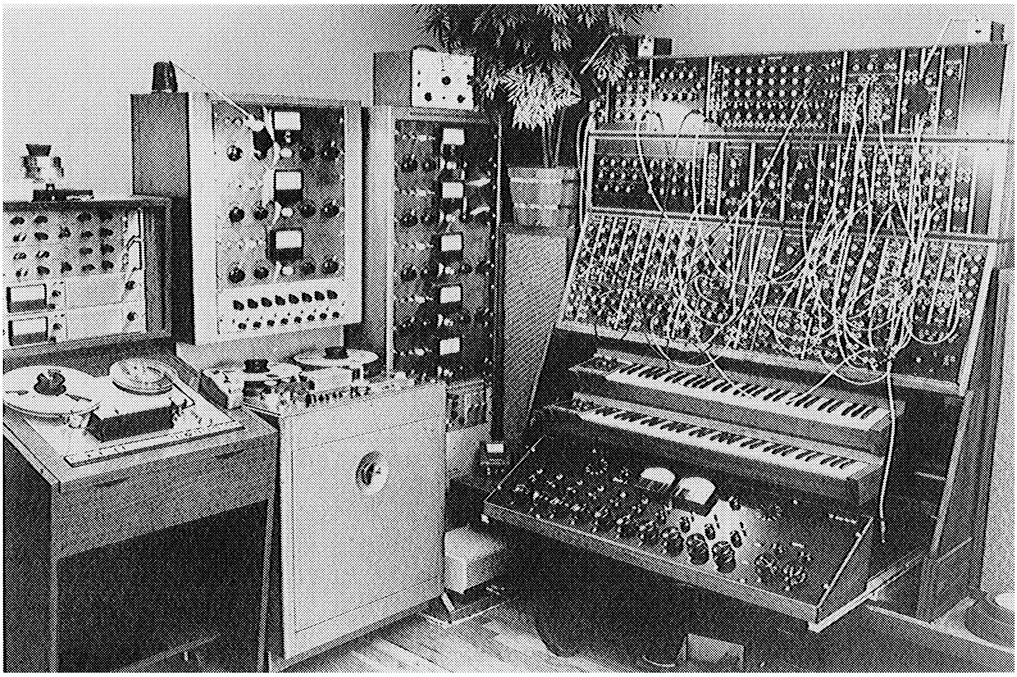

FIG. 13. The studio of W. Carlos, where "Switched-on Bach" was realized.

and most interconnections accessible with front panel switches, musicians quickly discovered the correlations between operating-point changes and timbral results. If, for instance, a rising envelope signal were used to raise the cutoff frequency of the low-pass filter over a time of a tenth of a second or so, then the resulting tone would have a horn-like attack. If, on the other hand, that same envelope signal were set up to descend at the beginning of the tone, thereby rapidly lowering the filter's cutoff frequency, then the resulting tone would have a plucked-string-like quality.

The first compact performance synthesizers produced one or two tones at a time and required the musician to set the front panel controls for a given timbre (Fig. 16). As the 1970s progressed, two technical advances contributed to the further evolution of commercial synthesizers. One of these is the microprocessor, while the other is the introduction of high-quality analog integrated circuits containing entire voltage-controlled signal-generating and -processing functions. Together, these advances enabled equipment designers to build "programmable polyphonic synthesizers." These instruments contained a small number of "voice circuits," each of which contained all the sound-generating and processing functions for one sound, plus a microprocessor-based scanning

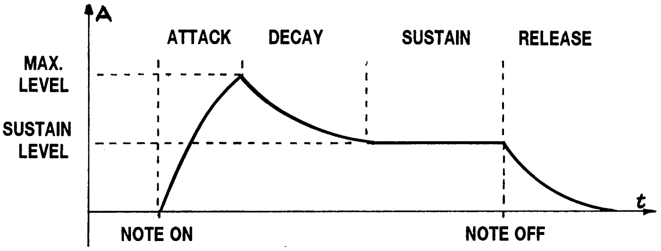

FIG. 14. An "ADSR" envelope.

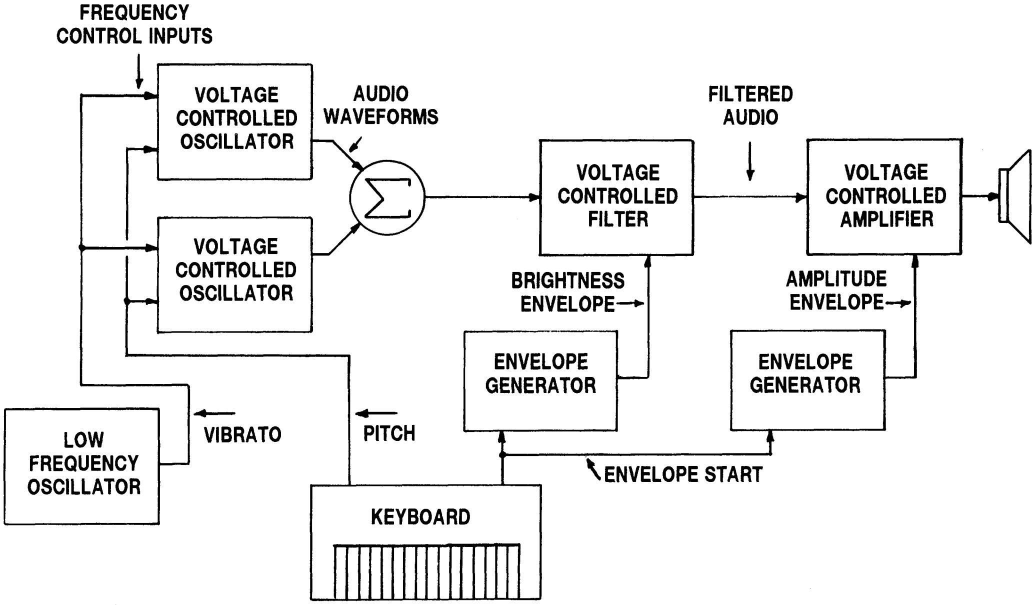
FREQUENCY CONTROL INPUTS
VOLTAGE CONTROLLED OSCILLATOR
AUDIO WAVEFORMS
VOLTAGE CONTROLLED OSCILLATOR
$\sum$
VOLTAGE CONTROLLED FILTER
FILTERED AUDIO
VOLTAGE CONTROLLED AMPLIFIER
BRIGHTNESS ENVELOPE
AMPLITUDE ENVELOPE
ENVELOPE GENERATOR
ENVELOPE GENERATOR
LOW FREQUENCY OSCILLATOR
VIBRATO
PITCH
ENVELOPE START
KEYBOARD

FIG. 15. Block diagram of a "standard synthesizer voice."

and data-storage system. The front panel controls were not connected directly to the sound-generating circuits. Instead, the microprocessor scanned the front panel controls, stored the settings for later editing or retrieval, and then routed the data to the voice circuits to set their operating points. The microprocessor also scanned the keyboard and, following nontrivial voice-assignment rules, associated a voice circuit with a newly depressed key.

Voltage-controlled devices were first de-

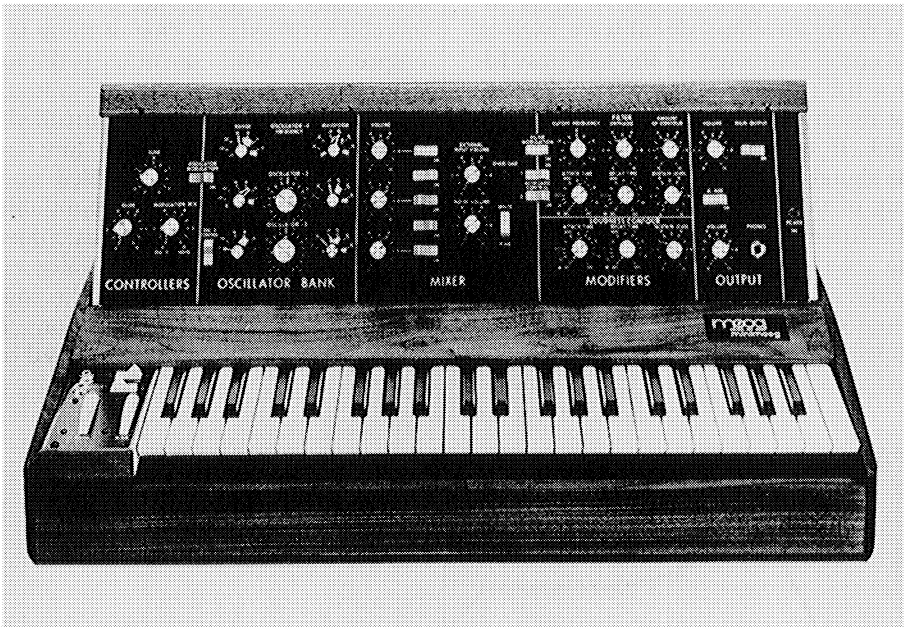
CONTROLLERS
OSCILLATOR BANK
MIXER
MODIFIERS
OUTPUT
moog

FIG. 16. A compact performance synthesizer (ca. 1972).

signed as instruments for the classical tape studio. Their users were experimental musicians who wanted maximum versatility, and were much less concerned with speed or ease of use. Over a period of no more than 15 years (1964–1979), they evolved into electronic-organ-like performance instruments. Instead of understanding the principles and practice of synthesis with voltage-controlled functions, players of latter-day voltage-controlled synthesizers are able to acquire complete banks of preprogrammed "synth sounds" on digital media. Beginning in the 1980s, many of the functions of voltage-controlled performance synthesizers have been ported over to all-digital instruments, which have become ubiquitous enough to be considered folk instruments.

Since the 1960s, experimental musicians seeking to develop their electronic resources further have gravitated to the digital computer. Advances in the realm of computer music have proceeded rapidly, with respect both to the performance/cost ratios of digital hardware and to the capabilities of computer software.

5. COMPUTER MUSIC

5.1 Early Application of Computers

Today's personal computers, together with a modest (by current standards) amount of sound-generation circuitry, are capable of producing complex, high-quality music in real time. Digital computers did not always offer this capability. Computers of the 1950s and early 1960s required hours just to compute the data from which a few seconds or minutes of music could be constructed. Then the data had to be processed by a digital-to-analog converter, an operation that taxed the then-available conversion hardware even to produce musical material of limited dynamic range and frequency response. For this reason, the earliest computer synthesis explorations centered on simple, sparse musical material.

The "Music *N*" (Music I–Music V) programs, developed during the 1960s at Bell Labs under the direction of Max Mathews, are representative of the music synthesis programs of this time. In writing the programs, Mathews addressed the problems of

1. how to minimize the vast amounts of data that are needed to generate musical sounds, and
2. what sort of a language could be employed to describe complex sequences of sounds.

Mathews's approach was to define the music in terms of *notes*, generate each note with one or more *unit generators*, and use stored wave forms as the starting material for the unit generators (Mathews, 1969).

Although Mathews's unit generators were subroutines of the Music *N* programs rather than hardware-based devices, they were similar in concept to voltage-controlled modules. Unit generators included such functions as oscillators (wave form generators with amplitude and frequency control inputs), multiple-input mixers (adders), and two-input multipliers. In the Music *N* programs, single-line instructions linked outputs to inputs, and numbers in decimal notation defined the wave forms, envelopes, and note durations. Most time-varying functions were defined in terms of straight line segments. Like a system of voltage-controlled-modules, a Music *N* "instrument" could produce tones whose complexity depended on how many unit generators were used. Unlike the analog counterpart, however, the complexity of a Music *N* instrument was limited not by hardware, but rather by constraints on available computing time. Figure 17 shows a very simple Music V instrument, along with the high-level instructions needed to set it up and have it produce two notes.

The program design principles articulated by Mathews became the basis for the development of many computer synthesis approaches. These range from simple (but computation intensive) additive synthesis schemes involving continuous control over the amplitudes of individual harmonics, to nonlinear schemes that produce complex musical tones with high efficiency. The latter is illustrated by John Chowning's explorations with frequency modulation (Chowning, 1973).

Electronic engineers are acquainted with frequency modulation (FM) as a communications technique. If the frequency of one sine wave (called the carrier) is varied in proportion to the amplitude of another sine wave

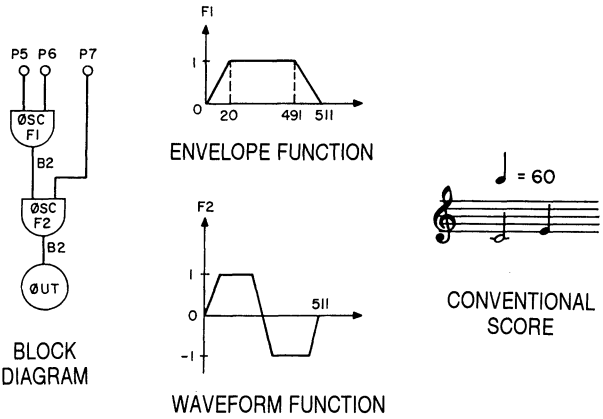

```
 1  INS 0 1 ;
 2  ØSC P5 P6 B2 F1 P30 ;
 3  ØSC B2 P7 B2 F2 P29 ;
 4  ØUT B2 B1 ;
 5  END ;
 6  GEN 0 1 1 0 0 .99 20 .99 491 0 511 ;
 7  GEN 0 1 2 0 0 .99 50 .99 205 −.99 306 −.99 461 0 511 ;
 8  NØT 0 1 2 1000 .0128 6.70 ;
 9  NØT 2 1 1 1000 .0256 8.44 ;
10  TER 3 ;
```

COMPUTER SCORE: Lines 1-4 specify the instrument; line 6 specifies the waveform; line 7 specifies the envelope; lines 8-9 specify the notes.

FIG. 17. Example of a simple Music V program: (a) block diagram; (b) envelope function; (c) wave-form function; (d) conventional "score"; and (e) the computer score. In the computer score, lines 1 through 5 specify the block diagram, line 6 specifies the envelope function, line 7 specifies the wave-form function, and lines 8 and 9 "play" the two tones.

(called the modulator), then the resulting frequency-modulated wave contains a rich set of sum and difference frequencies. If both carrier and modulator are audio-frequency waves, then the sum and difference frequencies are also audible partials. By varying just the frequency ratio between the two sine waves, and the amplitude of the modulator, one can produce an extremely wide variety of complex, musically rich wave forms.

The results that Chowning and his colleagues obtained from their work with FM were subsequently licensed to the Yamaha Corporation, and have formed the technical basis for some of the most successful digital keyboard instruments to have been introduced. A representative model is the Yamaha DX7, whose sounds are widely heard in current popular music (Fig. 18). Figure 19 shows an illustration from the DX7 user's manual that explains, in terms that musicians have come to understand, how FM produces musical tones.

5.2 Sampling Technology

In synthesizing musical tones, one generally starts with simple wave forms and time-varying control functions, and then builds up complexity by combining components and processing operations. All of the synthesis methods described above work this way.

Another approach to achieving a desired timbre is to start with a sound that is already complex, and perform simple operations on it to change its character. The existing sound is usually a recording of a "natural" sound. This approach works well when the desired sound is close to some already existing sound. The technique of recording individual sounds, then playing them back at specified times and with specified modifications, is called *sampling*.

The advantages of sampling are obvious, especially when one considers the needs of popular-music performers. The first commercially viable sampling instruments used

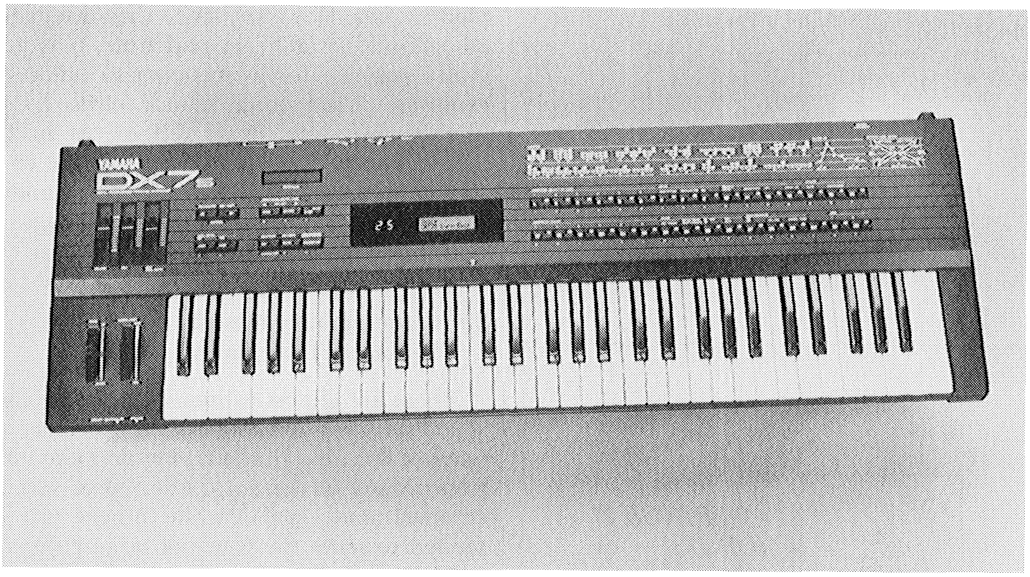

FIG. 18. The Yamaha DX7 synthesizer (ca. 1983).

lengths of magnetic tape on which were recorded choral, string ensemble, and horn ensemble tones, all of which were difficult to synthesize. The tape segments were set in motion by pressing keys on a conventional keyboard; a single recording typically lasted 8 s or so, usually long enough to play one note in a piece of music. When the keys were released, the tape segments were quickly returned to their starting positions. These instruments bore the trade names Mellotron and Chamberlain, and were widely used throughout the 1960s and 1970s, despite their mechanical complexity and tendency for the tapes to break during performance.

Equipment designers understood that individual sounds could be stored as digital recordings in semiconductor memory and then read out in response to a signal such as a key depression. However, the high price of semiconductor memory made this approach impractical until the late 1970s, when the Fairlight Computer Music Instrument was introduced (Fig. 20). The Fairlight was a small general-purpose business computer that was equipped with a piano-style keyboard and a multiplicity of peripheral "voice" cards, each of which had enough RAM (random-access memory) for one sound, and a digital-to-analog converter. In addition to scanning the keyboard and providing appropriate commands to the voice cards to play their sounds, the Fairlight provided the means for recording individual sounds as digital files, graphically editing the wave forms of those sounds with a light pen, storing individual sound files on 8-in. floppy disks, and loading sounds from disk to the voice card memories. The Fairlight and other disk-based sampling instruments enjoyed great popularity among producers of commercial and popular music.

The Kurzweil K250 was the first sampling instrument to have a library of sound files stored in semiconductor ROM (read-only memory) instead of on disk (Fig. 21). The ROM storage allowed the player to switch sound files instantly, instead of having to change disks and wait for the new sounds to load.

Since the introduction of the K250, the popularity of ROM-based sample-playing instruments has grown rapidly. Commercially successful instruments range from digital pianos and "band-in-a-box" peripherals for personal computers, to sample-playing modules with a full range of high-quality orchestral timbres and a broad palette of signal-processing resources to modify the timbres of the samples.

What are the relative advantages of sample-playing instruments over instruments that synthesize sounds from simple components? Sample players produce realistic simulations

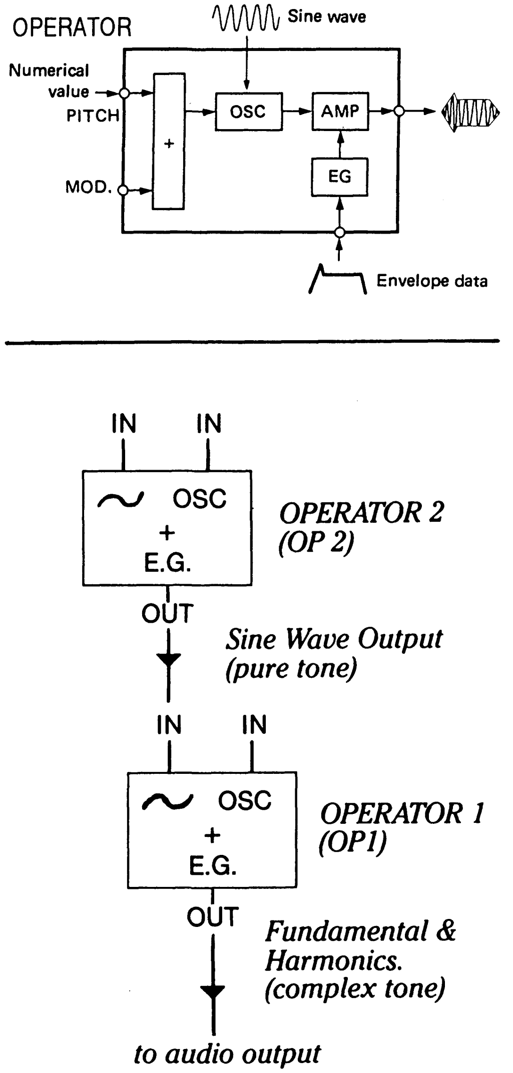

FIG. 19. An FM "operator," and a block diagram showing how two operators may be "patched" together to produce a complex tone.

of already familiar timbres. However, since their sounds are based on fixed recordings, the complex microstructure of each sound's evolution is difficult to change, and successive tones tend to sound alike. That is not a serious drawback in an instrument like a digital piano, because successive sounds of an acoustic piano also tend to sound alike. If an orchestral timbre such as a violin tone is played, however, successive tones should not sound exactly alike.

Synthesized timbres, on the other hand, are rarely as complex or "realistic" as sampled sounds. However, since they are generated "from scratch" in real time, it is generally possible for a musician to tailor the evolution of individual sounds, either by explicitly including small differences in the specifications of sounds as part of the composing process, or by using performance transducers to impart continuous changes in real time, as the tones are generated.

5.3 Contemporary Sound Synthesis Techniques

There are almost as many computer-based sound synthesis techniques in use as there are ways of defining time-varying periodic functions! Many of these are tailored to specific compositional goals, while others are intended to serve the needs of large groups of musicians. Four techniques, which are briefly described below, are representative examples of those that are currently in use.

Additive synthesis is synthesis of a periodic wave form by generating individual sinusoidal harmonics, imparting the appropriate slowly varying amplitude and frequency envelopes to them, and then combining them to achieve a desired sound (Snell, 1985). In theory, additive synthesis may be used to create any periodic function whatsoever. However, practical considerations tend to limit its usefulness. Since a complex low-pitched sound may contain hundreds of audible harmonics, additive synthesis tends to be extremely computation-intensive. Furthermore, additive synthesis cannot produce sounds with noise (random) components.

Nonetheless, several devices employing additive synthesis have been produced. Some enable the musicians to analyze a musical tone into its component harmonics, edit the envelopes of the harmonics, and then recombine the harmonics into a new sound that is closely related to the starting sound. This technique is called *resynthesis*, and is used by musicians who want to make small, carefully controlled timbral variations. The user interface of one such device, the Technos Axcel synthesizer, is a touch-sensitive surface on which a musician can draw envelope shapes directly with his finger.

Nonlinear distortion is a processing technique whereby a sound wave form is multiplied by a function that varies both with time and with the instantaneous amplitude of the

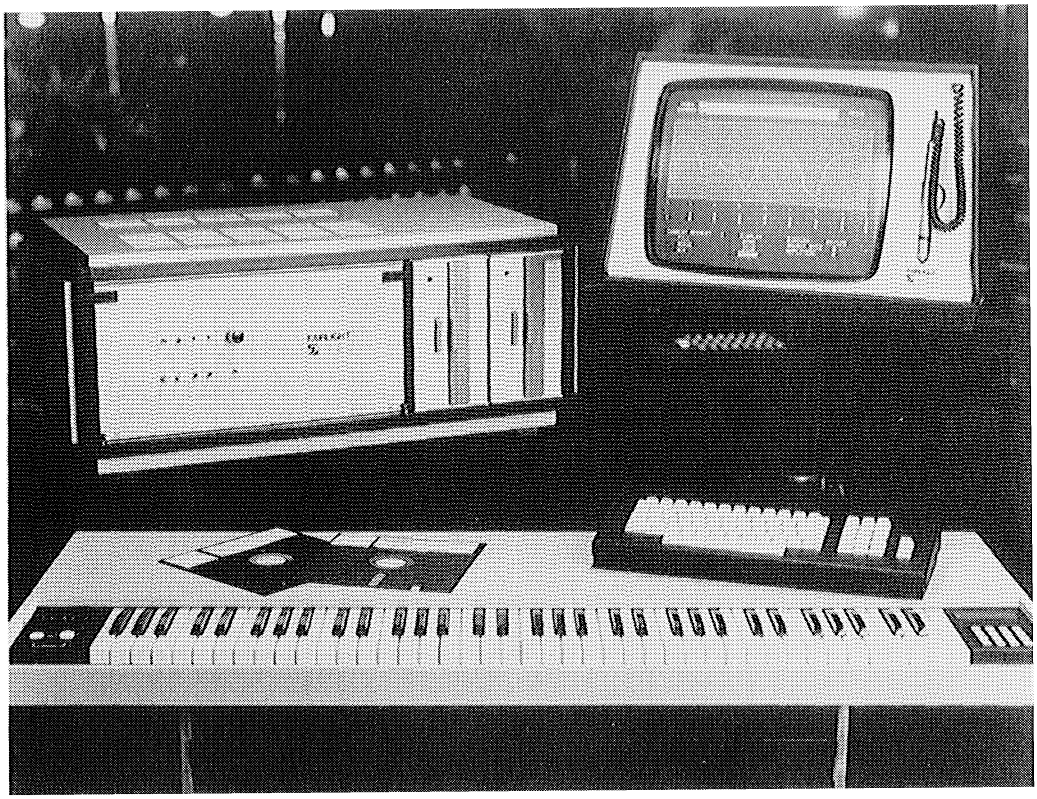

FIG. 20. The Fairlight "Computer Music Instrument" (ca. 1978).

starting wave form. This technique generally increases the harmonic content (Roads, 1985b). Useful distortion functions range in complexity from simple "clipping" of the starting waveform to complex, rapidly varying distortions that heighten the amplitudes of specific harmonics or groups of harmonics. Nonlinear distortion is especially useful for imparting timbral variation to sounds that are already defined, such as sounds produced by a sampling instrument.

Granular synthesis is the production of a wide variety of sonic textures by breaking continuous sounds up into very brief "granules." The composer specifies parameters such as the starting wave form, the average length and repetition time of the granules, the degree of randomness of length and repetition time, and the attack and decay times of the granules (Roads, 1985a). Music using granular synthesis tends to be pointillistic when the repetition time is long compared to the granule length, and shimmering and sonorous when the granules occur densely enough to overlap.

Models of distributed vibrating systems (either real or imagined) may be constructed with computer programs or dedicated digital circuitry. Properly excited and controlled, these models have the potential for producing complex wave forms of the type associated with acoustic musical instruments (Borin *et al.*, 1992). The Yamaha model VL1 Physical Modeling Synthesizer is an example of a commercial instrument employing such models (Fig. 22). *Waveguide modeling* is another modeling technique, in which waveguide boundaries, surface discontinuities, and exciting functions are specified, often as functions of time.

5.4 Linear Sound Processing Techniques

Just as many tape music composers preferred to compose by processing recorded "real-world" sounds, computer musicians

FIG. 21. The Kurzweil 250 (1985).

often prefer to start with recorded sounds, and then treat them with linear (nondistorting) processes before assembling them into musical structures. These processes are generally computation intensive, and rely heavily on the availability of fast multiplying devices and routines. Here are examples of linear processes that are currently in use.

Filtering is the process in which audio signals are passed through models of devices in

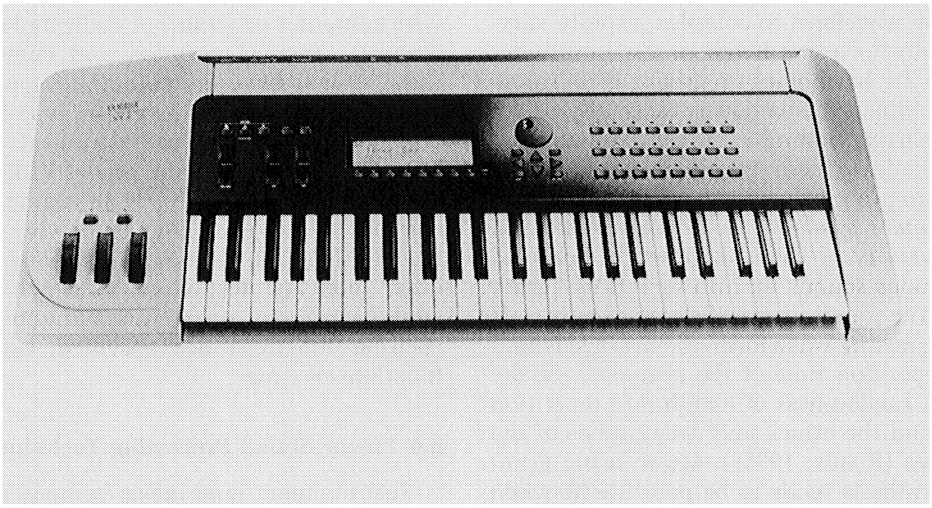

FIG. 22. The Yamaha VL1 synthesizer incorporates physical modeling of orchestral instruments (1994).

which the gain and phase shift are functions of frequency. Structures with multiple resonances (such as the bodies of string instruments) may be simulated, or filters with sharp cutoffs may be used to eliminate portions of the audio-frequency spectrum completely from the starting audio. All-pass filters are used as phase-shifting devices; when the phase-shifted output is added to the unshifted input signal, nulls occur in the composite frequency response, imparting the characteristic "comb filter" quality to the sound (Smith, 1989).

Linear predictive coding is a specialized filtering technique in which an input signal is analyzed to yield a time-based series of *n*-pole infinite impulse response (IIR) filters whose poles approximate the formants of the analyzed signal. The analysis is repeated as often as a hundred times a second, thereby providing a frequency response that follows the changing formant structure of the analyzed material. The frequency response is then used as a filter to process a second, relatively static signal (Makhoul, 1975). Linear predictive coding is especially valuable in analyzing vocal sounds and imparting vocal articulation and formant structure to simple wave forms with high harmonic content.

Reverberation is the creation and control of a multiplicity of closely spaced echos of the original audio material. Reverberation simulates the acoustic effect of the sound source that is enclosed in a reflective space. In digital reverberation systems (which consist of little more than a multiplicity of digital delay lines), the parameters of reverberation—spacing and decay rate of echoes, dependence of the decay rate on frequency, and elapsed time to the first echo—may be programmed to simulate listening spaces ranging in volume from that of a 55-gallon drum to Westminster Cathedral (Moorer, 1979). Used as a compositional resource, reverberation enables the composer to regard the perceived spatial environment as a manipulable compositional element. And, by making the delay times very long, say greater than 0.1 s or so, one can produce discrete echoes.

The perceived *spatial location and motion* of a sound can be controlled by a combination of simple processes. In addition to the standard bag of tricks employed by recording engineers to define left-to-right location, one can control the ratio of reverberated to unreverberated sound to influence the perception of distance between the listener and the sound source. One can also vary the sound's frequency and amplitude to simulate the Doppler effect of a sound source moving rapidly past the listener (Chowning, 1971).

5.5 Computer Systems as Composing Tools

At the present time, musicians are able to connect a wide variety of readily available electronic music devices into computer-controlled systems. This is made possible by MIDI (Musical Instrument Digital Interface), a communications protocol that has been adopted by the electronic musical instrument industry. MIDI employs a serial interface and simple two-conductor cables to interconnect system components. The MIDI codes enable information on when notes begin and end, how continuous variables are being changed, which "program" (set of parameter values that determine timbre) has been selected, and so forth, to be sent from one system device to the others (Rothstein, 1992).

The central component of a MIDI system is usually a personal computer that is equipped with MIDI-specific software. The most widely used class of MIDI programs are the so-called sequencer programs, which enable MIDI messages to be recorded, edited, and played back in loose analogy to a tape recorder. Sequencers have graphic interfaces that display the MIDI messages as musical scores, graphs, and alphanumeric lists. With a MIDI sequencer and a small number of MIDI-controlled sound-producing and -processing devices, one can realize a wide variety of music of moderate complexity.

MIDI was drawn up in response to the needs of musicians who are interested in popular, rather than experimental, music. For this reason, the orientation underlying the specification of MIDI codes tends to favor conventional pop music. Even so, many ingenious programs serving the needs of experimental musicians have been developed. Just a few examples are David Zicarelli's *MAX*, an unusually versatile screen-based patching and mapping system, Wayne Kirby's *Serious Composer*, an algorithmic composition program based on fractals, and Laurie Spiegel's *Music Mouse*, an improvisation environment in which one first specifies boundary conditions for a set of musical events, and then manipulates the events with a computer mouse.

5.6 Interactive Systems for Improvising

Improvising systems are algorithmic composition programs that are designed to be used in real time. They analyze inputs from acoustic or electronic tone sources, then apply composition rules to formulate a musical response. The composition rules are invariably complex, since the system must attempt to simulate the creative intuition of a human performer, anticipate the probabilities of a wide range of possible inputs, and continuously correct its output in a relatively graceful way. In a very real sense, the rules that guide an improvising system are the expression of a musical personality, and formulating the rules requires considerable musical insight.

A simple example of an improvising system is Roger Dannenberg's Blues Accompaniment program (Dannenberg and Mont-Reynaud, 1987). The program assumes that the live performer will play a blues line with the standard 12-measure blues chord progression in mind, and that he will play within a narrow tempo range. When the musician starts, the program seeks to determine the beat and location within the chord progression. Once it makes a decision, the system produces the chord pattern, while updating the likelihood of other chords being more or less "correct" with each successive beat.

Robert Rowe's program *Cypher* is another example of an interactive improvising system (Rowe, 1992). Rather than address the relatively rigid rules of a conventional blues improvisation, Cypher works on a looser set of composition rules. In Cypher, musical ideas are derived from the inputs provided by the improvising musicians, transformed algorithmically, and then played back to provide new musical ideas for the performers themselves to develop.

5.7 Performer Interfaces and Performance Techniques

The current state of the electronic music art is such that the means for creating sounds of high musical quality, and composing these sounds off line into musical works of moderate complexity, are widely available at relatively reasonable cost. However, real-time performance control over musical variables leaves something to be desired. The performance interfaces of most currently available electronic musical instruments are mechanically primitive organ-style or piano-style keyboards. Few instruments enable musicians to impart the sorts of complex, multivariable nuances that traditional acoustic musicians take for granted.

The development of alternative performance interfaces continues to proceed, however. In the commercial musical instrument world, several designs based on traditional performance interfaces are available. Among these are guitarlike, clarinet-like (Fig. 23), and drumlike devices whose outputs are MIDI messages rather than sound wave forms. Experimental musicians often favor more general, less conventional control interfaces such as touch-sensitive surfaces (Fig. 24).

Many experimental performers are working with "virtual reality" systems—multiple transducers, the outputs of which are used to create real-time computer models of a moving hand or body. The DataGlove is one such device. With a DataGlove and appropriate software, one can "play a virtual instrument" simply by moving one's fingers in space.

Work is also proceeding on adapting traditional keyboard design for use in electronic music systems. The complex tactile feedback provided by the keyboard of an acoustical piano has been electronically simulated by equipping each key of an experimental keyboard with a motion sensor and a linear motor, and providing the appropriate feedback from sensor to motor. In another development, an experimental keyboard has been equipped with sensors that continuously detect the vertical position of the key as well as two dimensions of the position of the performer's finger on the key surface, thereby providing continuous control over three variables of each sound (Fig. 25) (Moog and Rhea, 1990).

We conclude this section with a description of Max Mathews's Radio Baton. The Radio Baton is a conducting device. That is, it is intended to be used in a situation where a piece of music is already composed as a series of computer instructions, and is to be "performed" under the global control of a

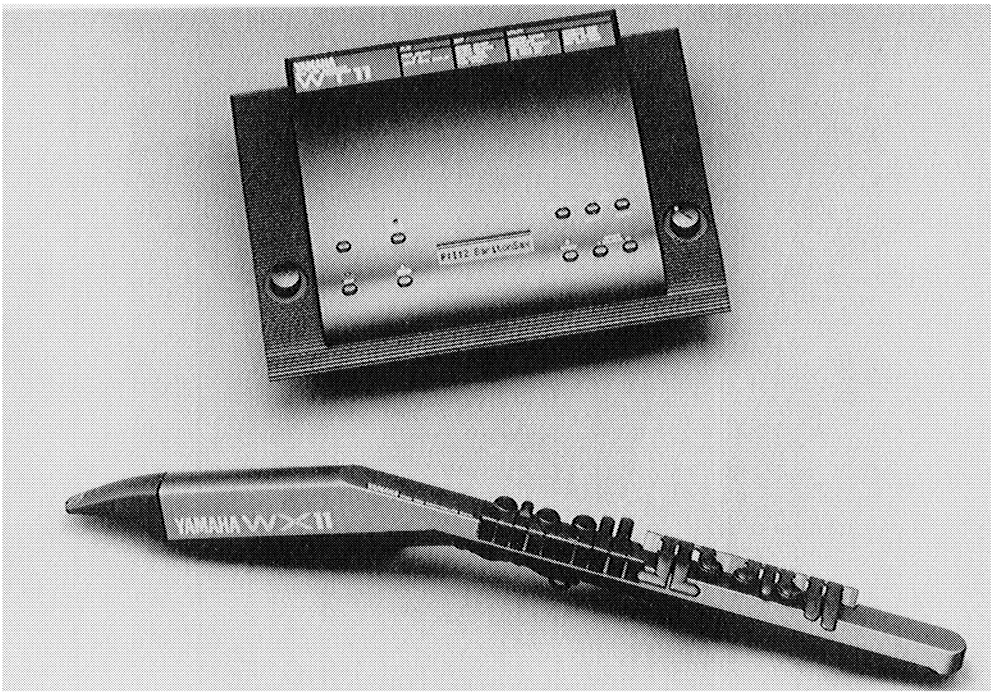

FIG. 23. The Yamaha WX11 wind controller. In addition to containing an array of switches for selecting pitches, the device contains a pressure transducer for continuously registering the player's breath pressure (1990).

"conductor." For instance, if the piece were of traditional tonal music, the score would specify the order of notes to be played, but not the overall tempo or dynamics. Mathews's Radio Baton consists of two drumsticklike batons, plus a surface over and on which the conductor moves the batons. As a transducer system, the Radio Baton uses radio-frequency fields to (continuously) determine the positions of the batons; from the conductor's point of view, the device enables him to use simple, intuitively appropriate gestures to control how the system "reads" the score (Fig. 26).

6. FUTURE DIRECTIONS FOR ELECTRONIC MUSIC TECHNOLOGY DEVELOPMENT

Despite its important role in contemporary experimental and art music, and its significant stake in the current musical instrument marketplace, electronic music technology has by no means reached its full maturity. The explosive development of digital electronics and software has opened up exciting new possibilities for musical sound generation, processing, performance, and composition. These possibilities have, in turn, underscored the need for better real-time human/machine interfaces, and encouraged researchers to explore further the psychoacoustic and cybernetic aspects of music production and listening. Some important opportunities for applying the principles of physics and engineering to the further development of electronic music technology follow.

6.1 Models for Sound Generation and Processing

The current trend in digital electronic music systems is toward increasing use of models of vibrating systems (as opposed to fixed wave forms) to generate and process musical tones. Models may be based on traditional acoustic musical instruments (for producing familiar timbres) or on imaginary vibrating devices (for exploring new timbres). The important requirements of such models

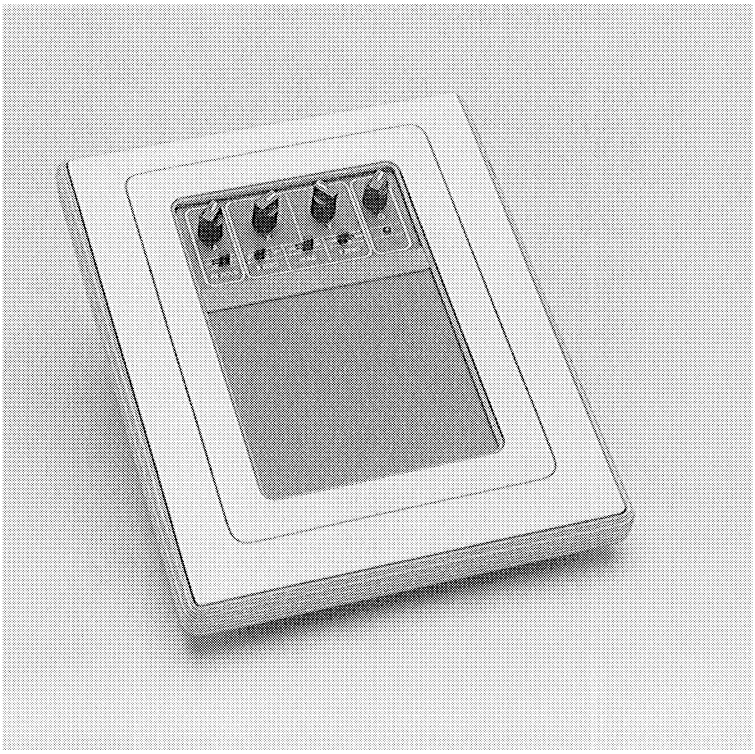

FIG. 24. A "touch-plate controller," which produces control signals proportional to two dimensions of the position of a player's finger on its surface, plus a third signal proportional to the area of the finger in contact with the surface (1984).

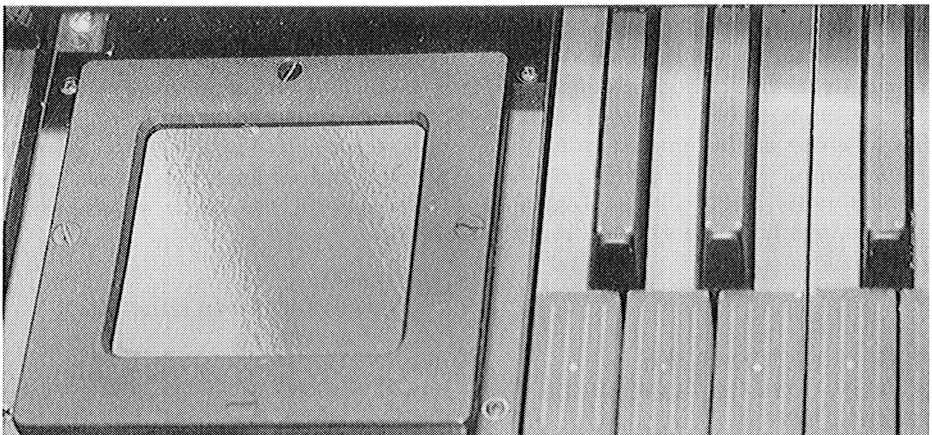

FIG. 25. A portion of a multiple touch sensitivity keyboard. The large touch plate on the left controls global parameters, while each of the keys on the right has its own small touch surface for shaping the parameters of a single sound event (1993).

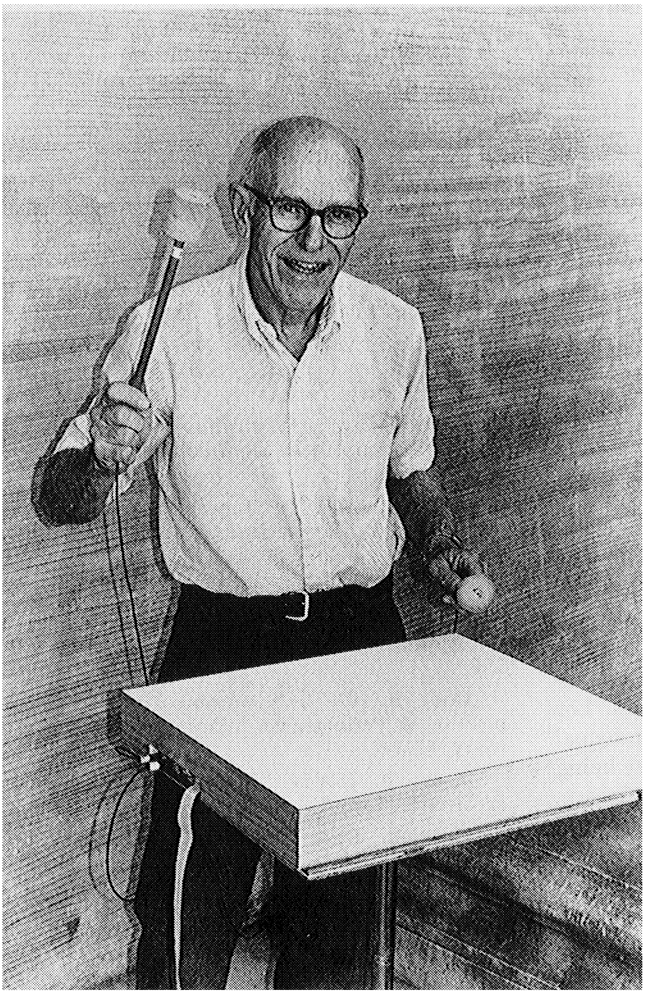

FIG. 26. Max Mathews and his Radio Baton.

are that their output signals be musically interesting, and that they contain a multiplicity of continuously variable parameters that the musician can manipulate in order to impart musical gestures. The disciplines of musical acoustics, psychoacoustics, and vibration analysis are appropriate starting points for the development of models for musical tone generation.

6.2 Electroacoustic Tranducer Design

Electronic music always needs one or more electroacoustic transducers in order to be heard. Current loudspeaker and headphone technology has evolved to meet the requirements of two-channel (stereo) wideband audio signal reproduction. Electronic musical instruments and performance systems often require more application-specific electroacoustic transducers. One common example is the speakers used in guitar amplifiers, which are deliberately designed to be somewhat nonlinear and of narrow bandwidth. Another example is a system of a large number of speakers that are distributed throughout a listening space in order to create a multichannel acoustic field. Opportunities exist for collaborations among electronic musicians, speaker designers, and architectural acousticians.

6.3 Real-Time Control Interfaces

A good real-time control interface between a musician and his instrument

1. senses hand and other body motions accurately and rapidly, and
2. provides equally accurate and rapid tactile feedback.

Resolution and accuracy requirements are typically on the order of 1 mm for position detection, and on the order of 1 g for force detection. Mechanical bandwidth (response speed) requirements are on the order of hundreds of hertz. Sensors and tactile feedback mechanisms must be rugged, reliable, smooth-acting, aesthetically pleasing—and inexpensive. The studies of materials science, sensor design, actuator (motor) design, and ergonomics can be applied to the further development of user interfaces for electronic music production.

6.4 Digital Electronics, Computer Software, and Artificial Intelligence

For all of human history, musical instrument builders have looked to the most sophisticated technologies of their times. For this reason, virtually all of the new musical instruments and systems of our time are being built with digital electronics and software. As the distinction between hardware and software continues to blur, and ever-increasing percentages of the value of electronic music devices become directly traceable to the encoded intelligence of computer programs, workers in electronic music technology are gravitating toward the resources of software-based artificial intelligence. Already, many music producers, commercial as well as experimental, do much of their work at computer terminals. The instrument builder, traditionally an expert at fabricating and assembling physical materials, will tend more and more to build software-based music production environments. To do this, he will draw on the principles of physical modeling, psychoacoustics, and ergonomics.

Works Cited

Babbitt, M. (1964), "An Introduction to the R.C.A. Synthesizer," *J. Mus. Theory* **VIII**, 251–265.

Berger, I. (1969), "The 'Switched-on Bach' Story," *Saturday Rev.* **January 25**, 45–47.

Borin, G., De Poli, G., Sarti, A. (1992), "Algorithms and Structures for Synthesis Using Physical Models," *Comp. Mus. J.* **16** (4) 30–42.

Chowning, J. (1971). "The Simulation of Moving Sound Sources," *J. Audio Eng. Soc.* **19**, 2–6.

Chowning, J. (1973), "The Synthesis of Complex Audio Spectra by Means of Frequency Modulation," *J. Audio Eng. Soc.* **21**, 526–534.

Dannenberg, R., Mont-Reynaud, B. (1987), "Following an Improvisation in Real Time," *Proceedings of the 1987-International Computer Music Conference*, San Francisco: Computer Music Association, pp. 241–248.

Eimert, H. (1956), *Electronic Music*, Ottawa: National Research Council of Canada, Technical Translation TT-612.

Glinsky, A. V. (1992), *The Theremin in the Emergence of Electronic Music*, Ph.D. Thesis, New York University, New York.

Laurendeau, J. (1990), *Maurice Martenot, Luthier de l'Electronique*, Montreal: Louise Courteau Editrice Inc.

Makhoul, J. (1975), "Linear Prediction: A Tutorial Review," *Proc. IEEE* **63**, 561–580.

Mathews, M. (1969), *The Technology of Computer Music*, Cambridge, MA: MIT Press.

Moog, R. (1965), "Voltage-Controlled Electronic Music Modules," *J. Audio Eng. Soc.* **13**, 200–206.

Moog, R. (1967), "Electronic Music, Its Composition & Performance," *Electronics World* **February**, 42–46.

Moog, R. (1981), "The Columbia/Princeton Electronic Music Center," *Keyboard* **May**, 22–32.

Moog, R., Rhea, T. (1990), "Evolution of the Keyboard Interface," *Comp. Mus. J.* **14** (2), 52–60.

Moorer, J. (1979), "About This Reverberation Business," *Comp. Mus. J.* **3** (2), 13–28.

Rhea, T. (1977a) "The Cellulophone," *Keyboard* **September**, 56.

Rhea, T. (1977b), "The Hammond Organ," *Keyboard* **May**, 47, and **June**, 47.

Rhea, T. (1978a), "B. F. Miessner's 'Stringless Piano'," *Keyboard* **April**, 62.

Rhea, T. (1978b), "Early One-Tune-Per-Note Organs," *Keyboard* **December**, 69.

Rhea, T. (1979a), "The Novachord," *Keyboard* **January**, 62.

Rhea, T. (1979b), "The Trautonium," *Keyboard* **April**, 76.

Rhea, T. (1979c), "The First Synthesizer," *Keyboard* **August**, 70.

Rhea, T. (1979d), "The Hanert Synthesizer," *Keyboard* **September**, 78.

Rhea, T. (1979e), "The Grainger-Cross Free Music Machine," *Keyboard* **November**, 90.

Roads, C. (1985a), "Granular Synthesis of Sound," in: C. Roads, J. Strawn (Eds.), *Foundations of Computer Music*, Cambridge, MA: MIT Press, pp. 145–159.

Roads, C. (1985b), "A Tutorial on Nonlinear Distortion or Waveshaping Synthesis," in: C. Roads,

J. Strawn (Eds.), *Foundations of Computer Music*, Cambridge, MA: MIT Press, pp. 83–94.

Rothstein, J. (1992), *MIDI—A Comprehensive Introduction*, Madison, WI: A-R Editions, Inc.

Rowe, R. (1992), *Interactive Music Systems*, Cambridge, MA: MIT Press.

Schaeffer, P. (1952), *A la Recherche d'une Musique Concrete*, Paris: Editions du Seuil.

Smith, J. (1989), "Fundamentals of Digital Filter Theory," in: C. Roads (Ed.), *The Music Machine*, Cambridge, MA: MIT Press, pp. 509–519.

Snell, J. (1985), "Design of a Digital Oscillator That Will Generate up to 256 Low-Distortion Sine Waves in Real Time," in: C. Roads, J. Strawn (Eds.), *Foundations of Computer Music*, Cambridge, MA: MIT Press, pp. 289–325.

Ussachevsky, V. (1958), "The Process of Experimental Music," *J. Audio Eng. Soc.* **6**, 202.

Further Reading

Appleton, J., Perera, R. (Eds.) (1975), *The Development and Practice of Electronic Music*, Englewood Cliffs, NJ: Prentice-Hall, Inc.

Chamberlin, H. (1985), *Musical Applications of Microprocessors*, 2nd ed., Hasbrouck Heights, NJ: Hayden Book Co.

Davis, D. (1988), *Computer Applications in Music—A Bibliography*, Madison, WI: A-R Editions.

Davis, D. (1982), *Computer Applications in Music—A Bibliography, Supplement 1*, Madison, WI: A-R Editions.

Deutsch, D. (1982), *The Psychology of Music*, New York: Academic Press.

Mathews, M., Moore, F., Risset, J. (1974), "Computers and Future Music," *Science* **183**, 263–268.

Mathews, M., Pierce, J. (Eds.) (1989), *Current Directions in Computer Music Research*, Cambridge, MA: MIT Press.

Moog, R. (1977), "Electronic Music," *J. Aud. Eng. Soc.* **25**, 855–861.

Moog, R. (1986), "Digital Music Synthesis," *Byte* **June**, 155–168.

Moore, F. (1990), *Elements of Computer Music*, Englewood Cliffs, NJ: Prentice Hall, Inc.

Pellman, S. (1994), *An Introduction to the Creation of Electroacoustic Music*, Belmont, CA: Wadsworth.

Pierce, J. R. (1992), *The Science of Musical Sound*, New York: W. H. Freeman.

Prieberg, F. (1960), *Musica ex Machina*, Berlin: Verlag Ullstein.

Rhea, T. (1972), *The Evolution of Electronic Musical Instruments in the United States*, Ph.D. Thesis, George Peabody College for Teachers, Nashville.

Roads, C. (Ed.) (1990), *The Music Machine*, Cambridge, MA: MIT Press.

Roads, C. (Ed.) (1985), *Composers and the Computer*, Madison, WI: A-R Editions.

Rothstein, J. (1992), *MIDI—A Comprehensive Introduction*, Madison, WI: A-R Editions, Inc.

Strawn, J. (Ed.) (1985), *Digital Audio Signal Processing: An Anthology*, Madison, WI: A-R Editions, Inc.

Schwartz, E. (1973), *Electronic Music: A Listener's Guide*, New York: Praeger.

Strange, A. (1983), *Electronic Music: Systems, Techniques, and Controls*, 2nd ed., Dubuque, IA: William C. Brown.

Vail, M. (1993), *Vintage Synthesizers*, San Francisco: GPI Books.

Winckel, F. (1963), "The Psycho-Acoustical Analysis of Structure as Applied to Electronic Music," *J. Mus. Theory* **7** (2), 194–246.

MUSICAL INSTRUMENTS

Thomas D. Rossing, *Department of Physics, Northern Illinois University, Dekalb, Illinois, U.S.A.*

	Introduction	129
1.	**String Instruments**	130
1.1	Vibration of Strings	130
1.1.1	Plucked Strings	130
1.1.2	Struck Strings	130
1.1.3	Bowed Strings	132
1.2	Plucked String Instruments	132
1.2.1	Guitars	132
1.2.2	Lutes	134
1.3	Bowed String Instruments	136
1.3.1	Violins	136
1.3.2	Other Instruments in the Violin Family	138
1.3.3	Viols	139
1.4	Piano	139
1.4.1	Piano Soundboards	140
1.4.2	Sound Decay	141
2.	**Wind Instruments**	143
2.1	Vibration of Air Columns	143
2.2	Reed and Lip Vibrations	144
2.3	Brass Instruments	145
2.3.1	The Trumpet	145
2.3.2	Valves and Slides	147
2.3.3	The French Horn	147
2.3.4	The Trombone	148
2.4	Woodwind Instruments	148
2.4.1	The Clarinet	148
2.4.2	The Saxophone	149
2.4.3	Double-Reed Instruments	149
2.4.4	The Flute	150
2.4.5	The Recorder	151
2.5	Pipe Organs	151
2.5.1	Sound Generation in Flue Pipes	154
3.	**Percussion Instruments**	154
3.1	Vibrations of Bars, Membranes, and Plates	154
3.1.1	Vibrations of Bars	155
3.1.2	Vibrations of Membranes	157
3.1.3	Vibrations of Plates	157
3.2	Drums	158
3.2.1	Kettledrums	158
3.2.2	Snare Drums	159
3.2.3	Indian Drums	160
3.3	Mallet Instruments	160
3.3.1	Marimba	160
3.3.2	Xylophone	162
3.3.3	Vibes	162
3.3.4	Chimes	162
3.4	Cymbals, Gongs, and Steel Pans	162
3.4.1	Cymbals	162
3.4.2	Gongs and Tam-Tams	163
3.4.3	Steel Pans	163
3.5	Bells	164
3.5.1	Carillon Bells	164
3.5.2	Vibrations of Carillon Bells and Church Bells	165
3.5.3	Handbells	165
3.5.4	Ancient Chinese Two-Tone Bells	165
4.	**Research Opportunities**	167
	Glossary	170
	Works Cited	171
	Further Reading	171

INTRODUCTION

Musical instruments have played an important role in nearly all cultures throughout history. First came the human voice (which we consider to be a musical instrument), then percussion instruments, and finally primitive string and wind instruments. All of these major families of instruments have developed along many different paths according to the culture that nurtured them. Together, they weave an interesting acoustical story.

The most common way to classify musical instruments is into the traditional families of string, wind, and percussion instruments, depending upon the nature of the primary vibrator. (Sometimes the alternative designations chordophones, aerophones, membranophones, and idiophones are used.) Of course, many instruments incorporate exci-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ters or resonators of a different nature from the primary vibrator: woodwind instruments make use of vibrating reeds, string instruments have rigid soundboards, xylophones and marimbas have air-filled resonators, etc.

An interesting acoustical characteristic of musical instruments is the positive or regenerative feedback that sustains oscillations in many instruments. In brass and woodwind instruments, the regenerative feedback is so strong that it creates regimes of oscillation. In bowed string instruments, it is also strong, but in plucked or hammered string instruments and in percussion instruments, it is hardly a factor at all. In instruments with little or no positive feedback, once the sound is initiated, the player has little means for changing the pitch or timbre. The human voice incorporates a modest amount of feedback.

The scientific study of musical instruments has attracted the attention of a number of famous scientists: Rayleigh, Helmholtz, Tyndall, Raman, Savart, and Saunders, to name a few. The study of musical instruments is an important part of the interdisciplinary field of musical acoustics. In addition to studying the production of sound by musical instruments, musical acousticians also study the acoustics of concert halls (see ACOUSTICS, ARCHITECTURAL), the recording and reproduction of sound (see SOUND REPRODUCTION), and the perception of musical sound (see ACOUSTICS, PSYCHOLOGICAL).

1. STRING INSTRUMENTS

1.1 Vibration of Strings

The equation for transverse motion of a completely flexible string can be written

$$\frac{\partial^2 y}{\partial t^2} = \frac{T}{\mu}\frac{\partial^2 y}{\partial x^2} = c^2\frac{\partial^2 y}{\partial x^2}, \tag{1}$$

where T is tension, μ is mass per unit length, and $c = \sqrt{T/\mu}$ is the speed of transverse waves. The general solution of Eq. (1) can be written in a form credited to d'Alembert (1717-1783):

$$y = f_1(ct - x) + f_2(ct + x). \tag{2}$$

The function $f_1(ct - x)$ represents a wave traveling to the right, and $f_2(ct + x)$ represents a wave traveling to the left. The nature of the functions f_1 and f_2 is arbitrary; they could be pulses, or they could be sinusoidal. In fact, these two independent functions can be chosen so that their sum represents any desired initial displacement $y(x,0)$ and velocity $\dot{y}(x,0)$.

A string of length L, fixed at its two ends, will have normal modes of vibration that can be written

$$y_n(x,t) = A_n \sin(2\pi f_n t + \phi_n)\sin(2\pi f_n x/c), \tag{3}$$

where $f_n = nc/2L$ is the frequency of the nth normal mode, A_n is its amplitude, and ϕ_n is its phase. The displacement y of the string can then be written as a sum of the normal modes:

$$y = \sum_n y_n. \tag{4}$$

The force of excitation determines the amplitude A_n and phase ϕ_n of each normal mode.

Plucking, bowing, or striking the same string would excite the same normal modes but with quite different values of the C_n's and ϕ_n's. Furthermore, the vibrational spectrum of a string depends not only on the manner of excitation but also on the point of excitation.

1.1.1 Plucked Strings When a string is plucked at its midpoint, the resulting vibration will be made up of the odd-numbered modes ($n = 1, 3, 5, 7, \ldots$) with amplitudes $C_n = 1, 1/9, 1/25, \ldots$ and alternating phases ($\phi_n = 0, 180°, 0, 180°, \ldots$). The mode amplitudes fall off as $1/n^2$ (12 dB/octave), and the string exerts a force on the end fixtures (e.g., the bridge of a guitar) that alternates in direction and whose waveform approximates a square wave (see Chap. 2 in Fletcher and Rossing, 1991).

Figure 1 shows the addition of modes to obtain the shape of a string plucked at one-fifth of its length. Note that the fifth mode is missing, also the tenth and fifteenth, as shown in the spectrum in Fig. 1. As a general rule, when a string is excited at a fraction $1/\beta$ of its length, the amplitude dips to zero for $n = \beta, 2\beta, 3\beta, \ldots$.

1.1.2 Struck Strings When a string is

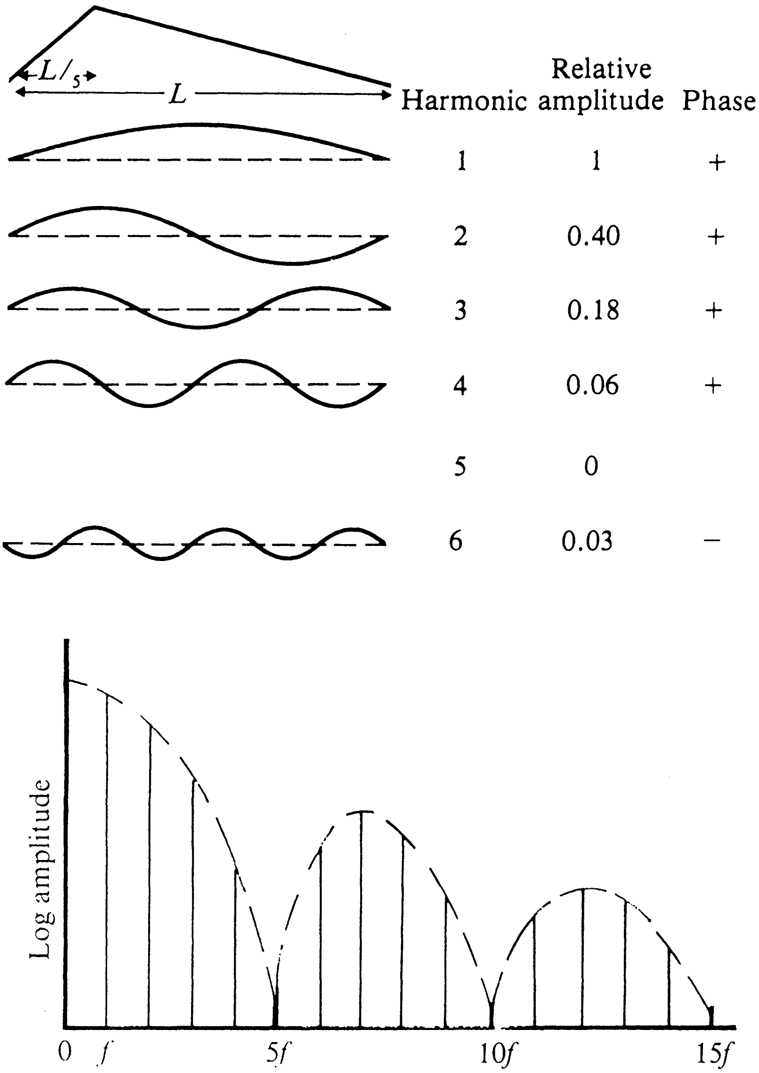

Harmonic	Relative amplitude	Phase
1	1	+
2	0.40	+
3	0.18	+
4	0.06	+
5	0	
6	0.03	−

FIG. 1. Addition of modes to obtain the shape of a string plucked at one-fifth of its length. The mode spectrum at the right shows the relative amplitudes of the modes.

plucked, it is given an initial displacement $y(x,0)$, but the initial velocity is assumed to be zero. When a piano string is struck by a hammer, however, the opposite is true: it is given an initial velocity but the initial displacement is zero. Figure 2 shows the displacement and velocity for a long string struck by a hard narrow hammer having a velocity V.

A short time t after being struck, a portion of the string with length $2ct$ and mass $2\mu ct$ is set into motion. As this mass increases and becomes comparable to the hammer mass M, the hammer is slowed down and would eventually be stopped. With a string of finite length, however, reflected impulses return while the hammer still has appreciable velocity, and these reflected impulses react with the hammer, causing it to be thrown back from the string.

The interaction of a felt piano hammer with a piano string is considerably more complicated than suggested by the simple hard, narrow hammer approximation. In general, mode amplitudes in a struck string fall off approximately as $1/n$ (6 dB/octave), which is

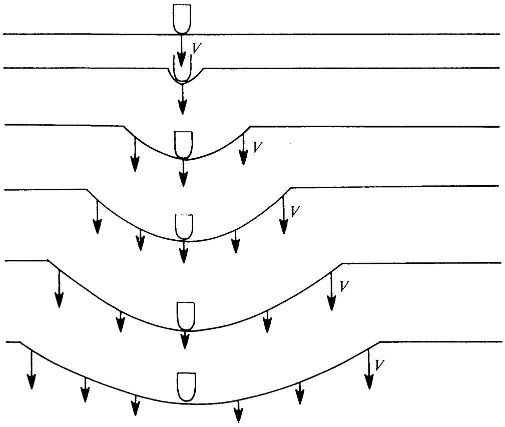

FIG. 2. Displacement and velocity of a long string at successive times after being struck by a hard narrow hammer having a velocity *V*. (From Fletcher and Rossing, 1991.)

considerably slower than those of a plucked string (which go nearly as $1/n^2$ or 12 dB/octave). If the string is struck at $1/\beta$ of its length, the amplitude drops to zero for $n = \beta, 2\beta, 3\beta, \ldots$, as in the plucked string case.

1.1.3 Bowed Strings When a bow is drawn across the string of a violin, the string appears to vibrate back and forth smoothly between two curved boundaries, much like a string vibrating in its fundamental mode. However, this appearance of simplicity is deceiving, as shown by Helmholtz (1877) over 100 years ago. The string more nearly forms two straight lines with a sharp bend at the point of intersection. This bend races around the curved path that we see, making one round trip for each period of the vibration.

The Helmholtz-type motion of a bowed string is shown in Fig. 3. As the bow moves ahead at constant speed, the bend races around a curved path. The graph on the right-hand side shows the position of the point of contact between bow and string at successive times; the letters correspond to the frames in the left half of the figure.

The action of the bow on the string is often described as a "stick and slip" action. The bow drags the string along until the bend arrives and triggers the slipping action [frame (a) in Fig. 3], which continues until the string is picked up once again by the bow [frame (c)]. From frames (c) to (i), the string moves at the speed of the bow. The bend moves along the string at the wave speed $c = \sqrt{T/\mu}$.

The envelope of the bend (dashed curve in Fig. 3) is composed of two parabolas, and it has an amplitude that is proportional, within limits, to the bow velocity. For a given bow velocity, it increases as the string is bowed closer to one end. Thus, a violinist has two ways to control loudness: bowing speed and bowing position. (Bow force has relatively little effect on amplitude, provided it stays within the range for which the stick and slip action works.)

The actual string motion may be a combination of several Helmholtz-type motions. The finer details of bowed string motion are elegantly treated in a book by Cremer (1981).

1.2 Plucked String Instruments

1.2.1 Guitars The modern guitar, which is a descendent of the 16th century Spanish vihuela, has 6 strings about 65 cm in length, tuned to E_2, A_2, D_3, G_3, B_3, and E_4 (f = 82, 110, 147, 196, 247, and 330 Hz). (Throughout this article, we will use the American Standard notation, in which A_4 has a frequency of 440 Hz. "Middle C" is denoted as C_4. The numbers change between B and C.) The top is usually cut from spruce or rosewood, planed to a thickness of about 2.5 mm, while the back is usually of hardwood with about the same thickness.

Acoustic guitars can be classified into four families of design: classical, flamenco, flat-top (or folk), and arch-top. Classical and flamenco guitars have nylon strings; flat-top and arch-top guitars have steel strings. Steel string guitars usually have a steel rod embedded inside their neck, and their soundboard has crossed bracing, whereas the braces on classical guitar soundboards are usually arranged in a fan-shaped pattern, as shown in Fig. 4.

A plucked guitar string exerts a series of impulses on the bridge, which is transmitted to the top plate. The top plate, in turn, transmits vibrational energy to the back plate and the enclosed air. At higher frequencies, sound is radiated mainly by the top plate, but at lower frequencies, sound radiation from the back plate and the sound hole are major contributors to the overall guitar sound. The mechanical frequency response of the guitar body to a large extent determines the quality of the radiated sound.

The mechanical frequency response and

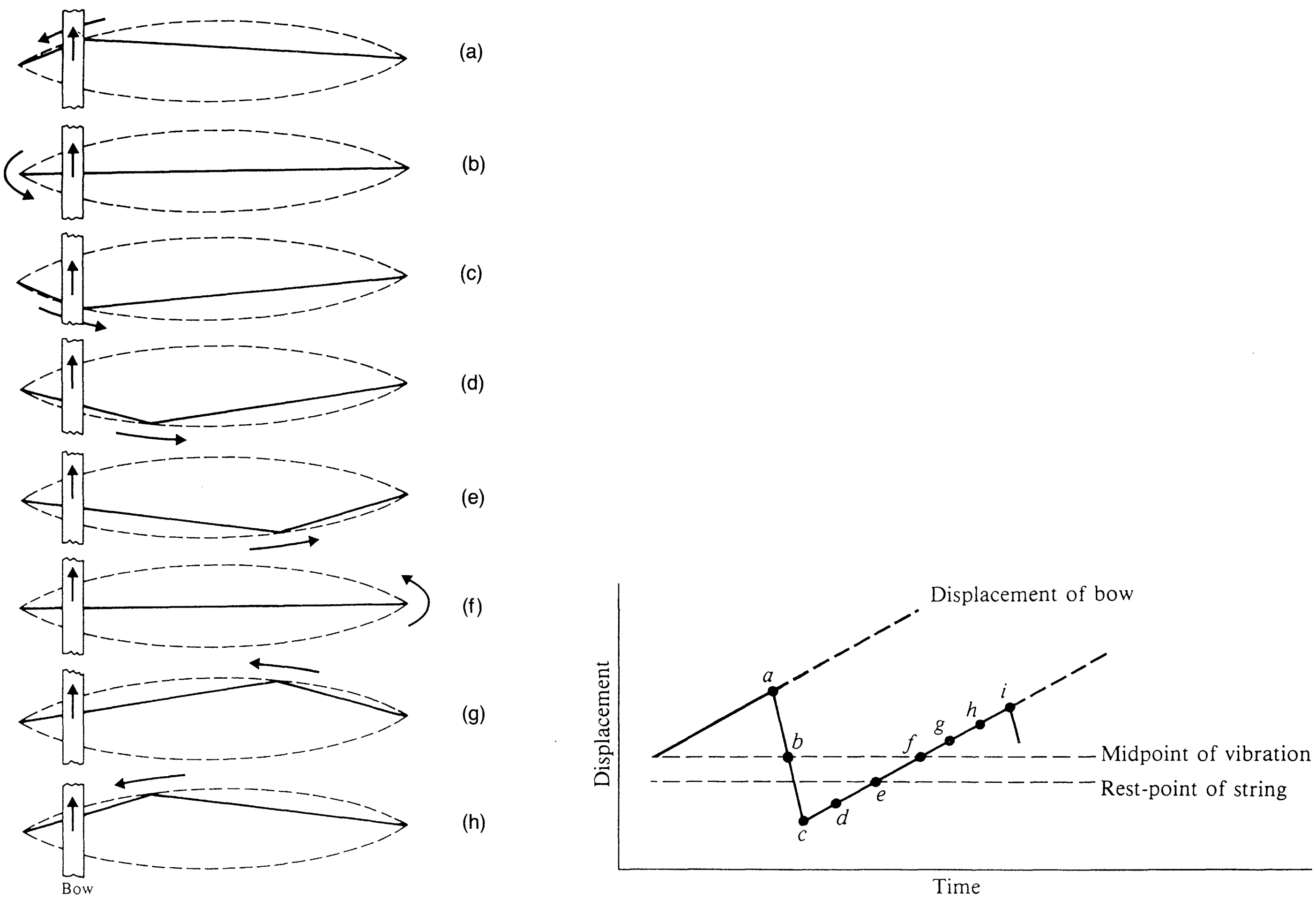
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
Bow
Displacement of bow
Displacement
a
b
c
d
e
f
g
h
i
Midpoint of vibration
Rest-point of string
Time

FIG. 3. Motion of a bowed string. Left: Time analysis of the motion, showing the shape of the string at eight successive times during the cycle. Right: Displacement of the bow (dashed line) and the string at the point of contact (solid line) at successive times. The midpoint of the string vibration is displaced slightly from the rest point of the unbowed string. The letters correspond to the letters in the time analysis at the left. (From Rossing, 1990.)

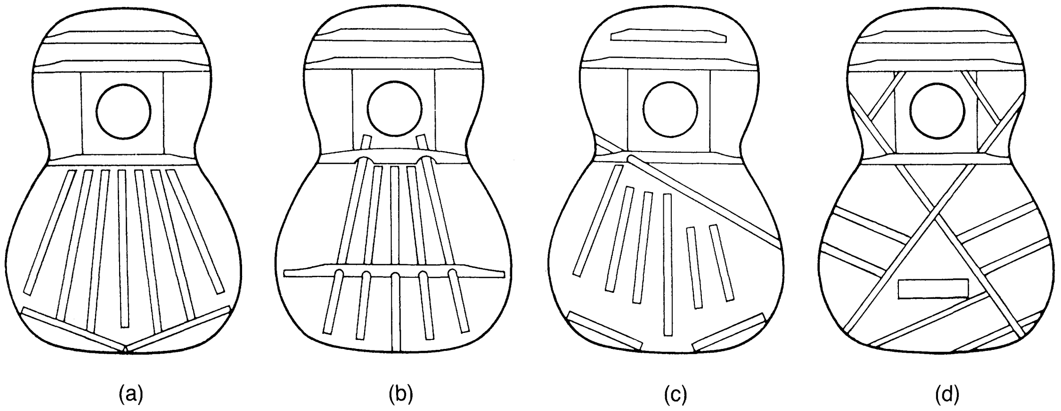

FIG. 4. Various designs for bracing a guitar soundboard: (a) traditional (Torres) fan bracing; (b) Bouchet (France); (c) Ramirez (Spain); (d) crossed bracing. (Rossing, 1990.)

radiated sound spectrum of a Martin D-28 folk guitar are shown in Fig. 5. The low-frequency sound spectrum is dominated by mechanical resonances at about 100, 190, and 205 Hz. These three resonances are due to coupled motions of the top plate, the back plate, and the enclosed air as shown in Fig. 6(a).

At 103 Hz, the top plate and back plate move in opposite directions, and so air is pumped in and out of the sound hole. At 204 Hz, on the other hand, the air flow through the sound hole is opposite in phase to the motion of the plates. This can be likened to the two modes of a mechanical system consisting of two masses connected by springs; in the mode of lower frequency, the masses move in the same direction, but in the mode of higher frequency, they move in opposite directions. At 193 Hz, the two plates move in the same direction.

The resonances at 376 and 436 Hz result from the "see saw" motion of the top and back plates around nodal lines running lengthwise. In the lower one, the top and back plates move in opposite directions so that air sloshes from one side to the other, while in the upper resonance, the two plates move together, resulting in less air motion.

The holographic interferograms in Fig. 7 show the motion of the top plate in a classical guitar at ten of its resonances. The lowest one at 268 Hz corresponds to the 376-Hz mode in the folk guitar, whose cross-bracing gives the top plate considerably more transverse stiffness.

1.2.2 Lutes Lutes of various types have

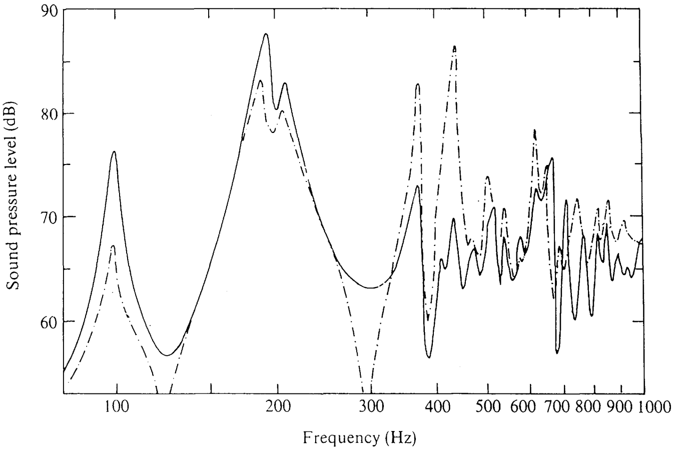

FIG. 5. Mechanical frequency response and sound spectrum 1 m in front of a Martin D-28 folk guitar driven by a sinusoidal force of 0.15 N applied to the treble side of the bridge. Solid curve, sound spectrum; dashed curves, acceleration level at the driving point. (Fletcher and Rossing, 1991.)

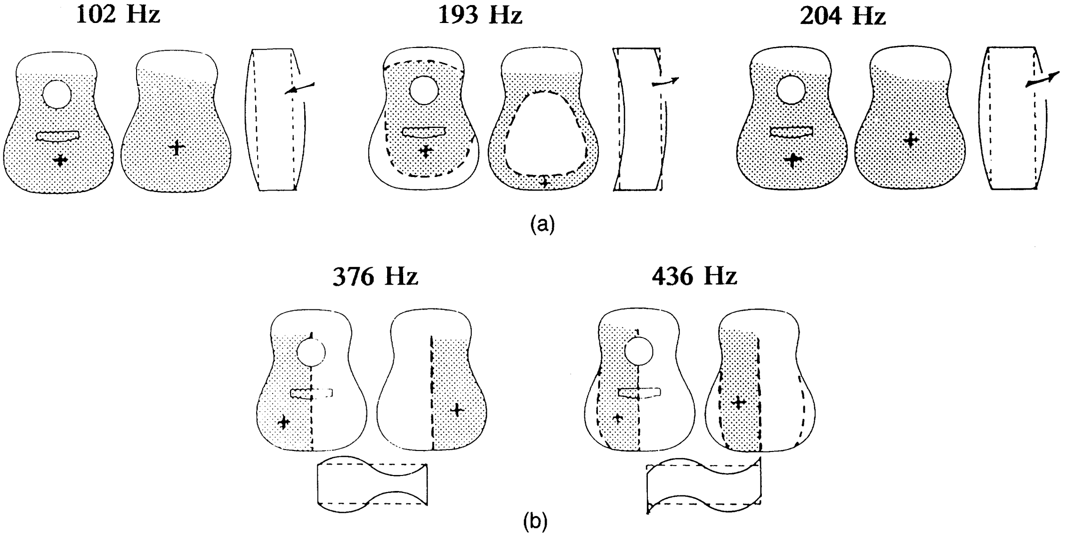
102 Hz
193 Hz
204 Hz
(a)
376 Hz
436 Hz
(b)

FIG. 6. Motion of the top plate, back plate, and enclosed air at five low-frequency resonances of a Martin D-28 folk guitar: (a) resonances in which the top and back plates each vibrate in a single phase; (b) resonances in which a nodal line divides the top and back plates into halves. (Fletcher and Rossing, 1991.)

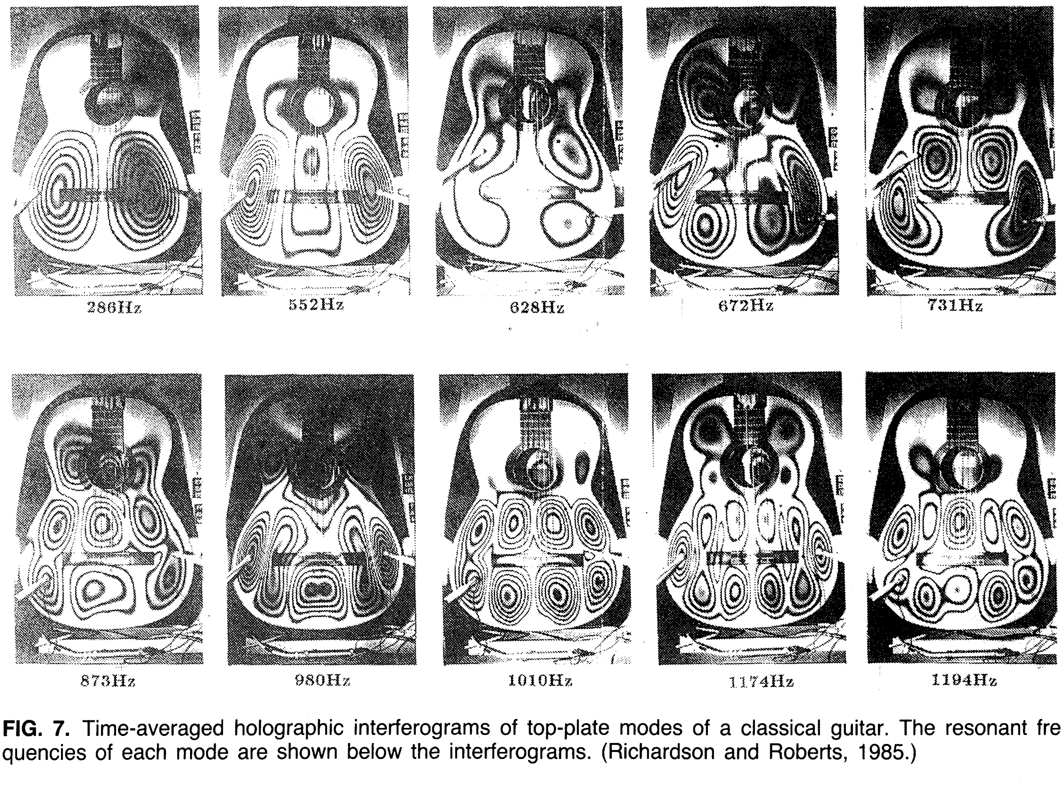
286Hz
552Hz
628Hz
672Hz
731Hz
873Hz
980Hz
1010Hz
1174Hz
1194Hz

FIG. 7. Time-averaged holographic interferograms of top-plate modes of a classical guitar. The resonant frequencies of each mode are shown below the interferograms. (Richardson and Roberts, 1985.)

existed throughout the Near East and Far East for many centuries. The instrument that developed great popularity in Europe during the 16th and 17th centuries typically had 11 strings arranged in six courses (all but the uppermost consisting of two unison strings) that might be tuned A_2, D_3, G_3, B_3, E_4, and A_4, although the tuning was often changed to fit the music being played. The pear-shaped body of the lute is fabricated by gluing together a number of thin wooden ribs. The table or soundboard (of spruce, cedar, or cypress) is braced by transverse bars above, below, and at the soundhole.

Although the acoustics of the lute appears to be an interesting subject, only a few studies on their acoustical behavior have been reported. Firth (1977) measured the input admittance and sound radiation of a modern lute and also investigated the modes of vibration. Nodal patterns for several modes are shown in Fig. 8.

1.3 Bowed String Instruments

1.3.1 Violins The violin is the most studied of all musical instruments in the symphony orchestra. Outstanding contributions to our understanding of violin acoustics have been made by such distinguished scientists as Felix Savart (France, 1791–1841), Hermann von Helmholtz (Germany, 1821–1894), Lord Rayleigh (England, 1842–1919), C. V. Raman (India, 1888–1970), and Frederick Saunders (United States, 1875–1963). More recently, the work of Professor Saunders has been continued by members of the Catgut Acoustical Society, an organization that publishes an excellent journal describing current research on the acoustics of string instruments.

A violin has four strings of steel, gut, or nylon (wrapped with silver, aluminum, or steel), tuned to G_3, D_4, A_4, and E_5. The top plate is generally carved from Norway spruce (*Picea abies* or *Picea excelsis*) and the back, ribs, and neck from curly maple (*Acer platanoides*). The fingerboard is made of ebony and the bow stick of pernambuco (*Caesalpinia echinata*), a very dense wood. Running longitudinally under the top is a bass bar, and a sound post is positioned near the treble foot of the bridge.

Why are some violins, especially those crafted by the Italian masters, so highly treasured, while other instruments that appear to be virtually identical to them are considered inferior to them in quality? This is a question that has intrigued violin researchers for many years. One important fac-

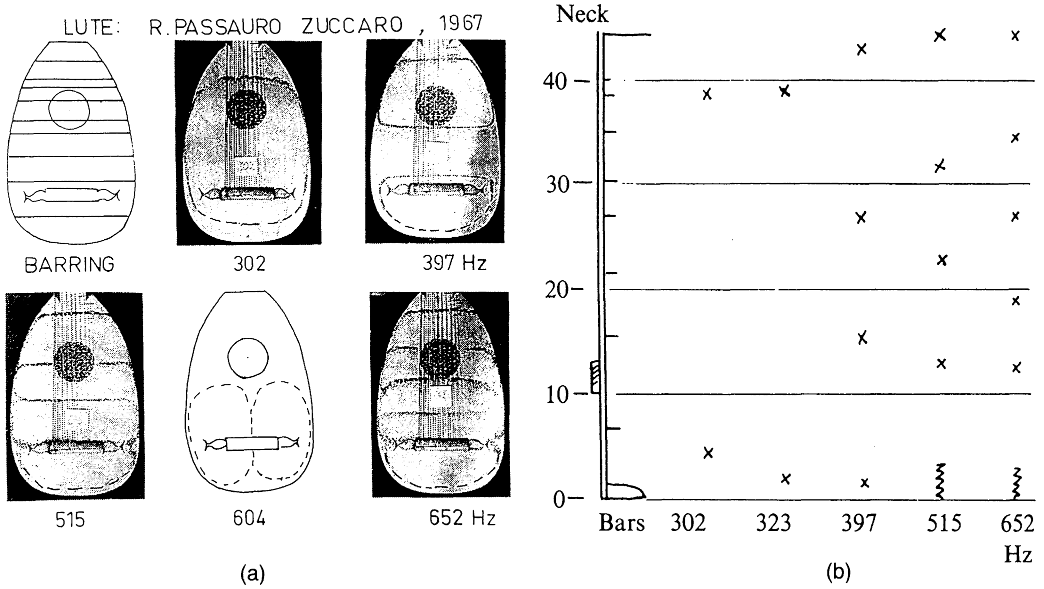

(a)

(b)

FIG. 8. (a) Barring pattern and nodal patterns in the top plate of the lute at five resonances and (b) locations of nodes compared to the bridge and the bars (Firth, 1977).

tor in the sound quality and playability of a violin is the vibrational behavior of its body. We often describe this in terms of normal modes of vibration, which consist of coupled motions of the top plate, the back plate, and the enclosed air. Smaller contributions are made by the ribs, neck, and fingerboard.

One way to determine the normal modes of vibration is to apply an oscillating force to the bridge and observe the motion of the various parts of the violin. Observing the motion may be done electrically, by means of accelerometers, or optically, by means of holographic interferometry.

Six low-frequency modes of vibration of a violin are shown in Fig. 9. The C_1 and C_2 body ("corpus") modes radiate little sound, although they contribute to the "feel" of the violin. The other four modes are indicated on the frequency response curve in Fig. 10, where the vertical coordinate is the velocity divided by the force (called mechanical admittance or mobility).

The A_0 ("air") mode and the T_1 ("top") mode both involve considerable motion of air in and out of the *f*-holes. Both modes radiate sound efficiently, and they dominate the low-frequency sound spectra of most violins. Above 1 kHz, the vibrational modes or resonances are bunched together, as can be seen in Fig. 10, and it is difficult to identify the individual modes.

One of the most critical steps in constructing a violin is carving the top and back plates. Since both plates have an arch, they are carved from blocks of wood thicker at the center. As the plates near their desired shapes, the violin-maker tests them by listening to "tap tones." To hear these, the plate is held in a particular way and tapped at certain spots. The trained ear of the violin maker can extract useful information by noting the pitch and the decay time of the tap tones.

The analysis of the vibrations of the plates has been refined considerably by the application of electronics and optics. One rewarding technique has been the observation of individual plate modes of vibration by Chladni patterns. The plate is usually mounted horizontally and excited by a loud-speaker. The two most important modes of vibration are those shown in Fig. 11. (The plates shown are for a viola, but violin and cello plates show similar patterns.) If the frequency of the lower mode (a) in the top plate matches that of the back plate, then tuning the upper modes (b)

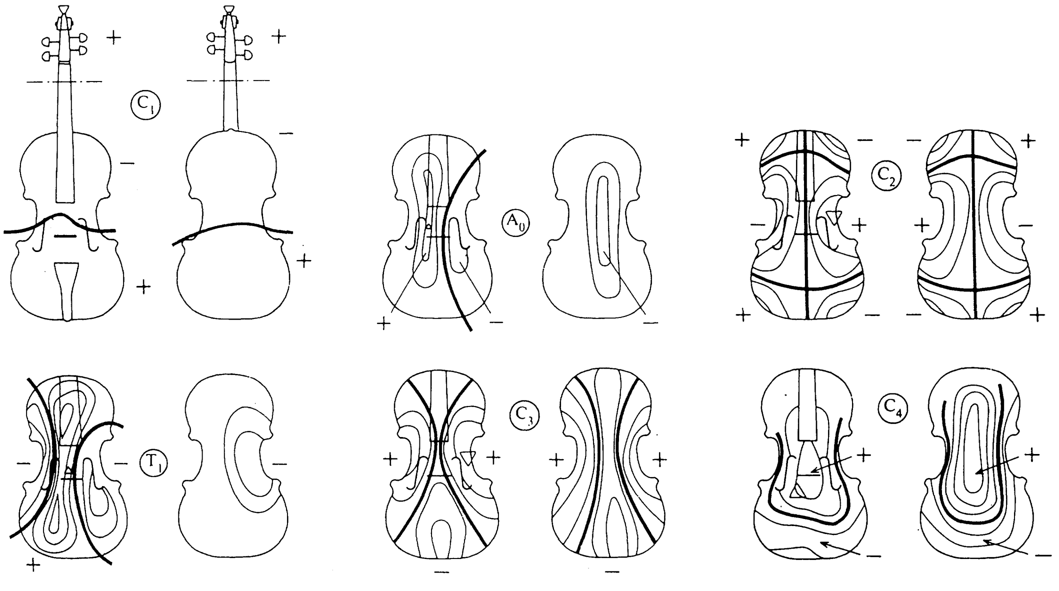

FIG. 9. Modal shapes for six modes in a violin. C_1 (185 Hz): one-dimensional bending; A_0 (275 Hz): air flows in and out of the *f*-holes; C_2 (405 Hz): two-dimensional flexure; T_1 (460 Hz): mainly motion of the top plate; C_3 (530 Hz), C_4 (700 Hz): two-dimensional flexure. Top plate and back plate are shown for each mode. The heavy lines are nodal lines; the direction of motion is indicated by + or −. The drive point is indicated by a small triangle. (After Alonso Moral and Jansson, 1982a.)

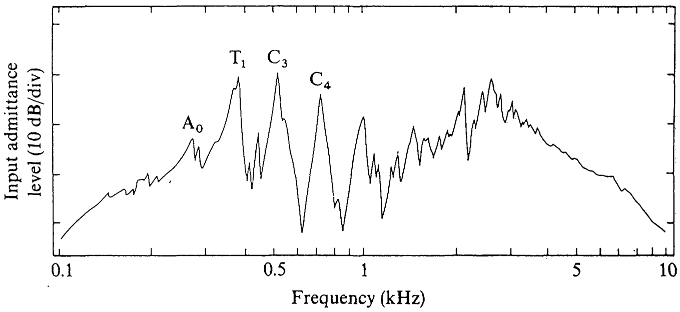

FIG. 10. Input admittance (driving-point mobility) of a Guarneri violin driven on the bass bar side (Alonso Moral and Jansson, 1982b).

to match each other in frequency or to be a semitone different will usually produce plates of high quality (Hutchins, 1977). The upper ("ring") mode (b) is often tuned an octave higher than the lower ("x") mode. Of course, the shapes of the nodal lines are equally as important as the frequencies of the modes.

The violin bridge transforms the motion of the vibrating strings into periodic driving forces applied by its two feet to the top plate of the instrument. Violin bridges typically have strong resonances around 3000 and 6000 Hz, as shown in Fig. 12. The lower resonance is due to a rocking-bending motion, whereas the upper one is due to a symmetrical up-and-down motion. These resonance frequencies can easily be changed by cutting away wood at the appropriate places, and thus, shaping the bridge is a convenient way to alter the frequency response of a violin.

In order to "darken" the sound of a violin, the player may attach a mute to the bridge. The additional mass of the mute shifts the bridge resonances to lower frequencies, changing the sound spectrum of the instrument. Typical violin mutes have masses of about 1.5 grams.

1.3.2 Other Instruments in the Violin Family

The other string instruments in the violin family are the viola, cello (violoncello), and double bass (also known as contrabass, string bass, or bass viol). The general shape and construction of the viola and cello are similar to those of the violin. The strings, like those of the violin, are tuned in intervals of

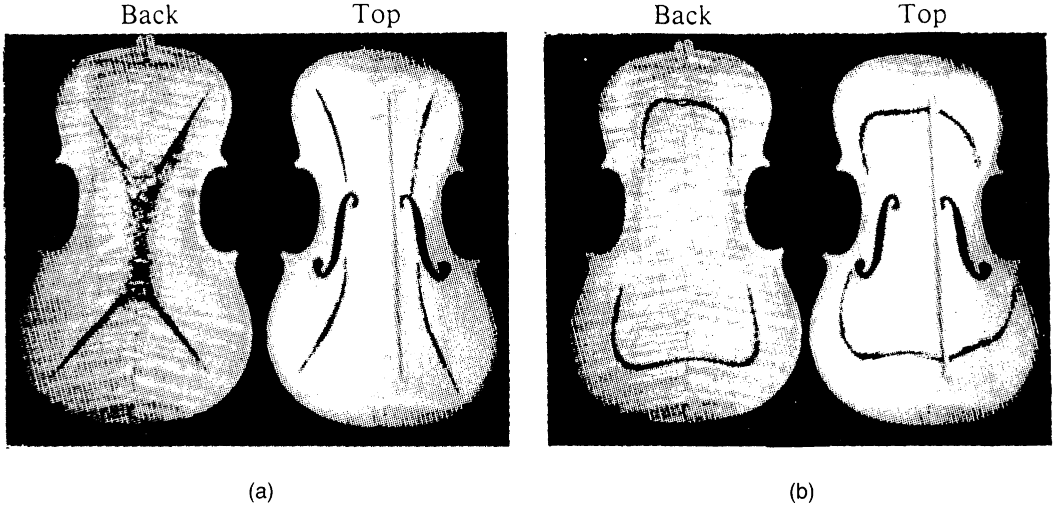

(a) (b)

FIG. 11. Chladni patterns showing modes of vibration in the viola top and back plates (from Hutchins, 1977).

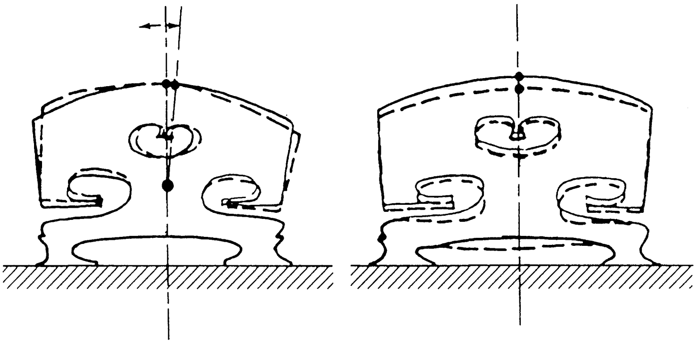

FIG. 12. First two vibrational modes (resonances) of a violin bridge.

fifths: C_3, G_3, D_4, A_4 in the viola and C_2, G_2, D_3, A_3 in the cello. Although the viola is tuned a fifth below the violin, its main air resonance is typically less than a third below that of the violin, and the main top plate resonance is a little over a third lower.

The lower resonances of a cello are related to its string frequencies in about the same way as those of the viola. However, the tall bridge of the cello results in a large driving force and a strong response near the main top plate (T_1) resonance. Also, the second air resonance is nearly an octave above the main air (A_0) resonance and thus reinforces this resonance by strengthening the second harmonic.

The double bass is tuned in fourths rather than fifths, which reduces the distances that the hand must travel along the long fingerboard. It has relatively narrow sloping shoulders and a flat back like members of the viol family.

In 1958, Carleen Hutchins and other members of the Catgut Acoustical Society set about to develop a new family of fiddles, with resonances scaled to those of a fine violin. Their research led to a family of eight new instruments. The family has two instruments pitched above the violin: the treble and the soprano. The alto, which is tuned like the viola, is either held under the chin (if one has long arms) or played vertically on a peg in the fashion of a cello. The tenor tunes between the viola and cello, and the baritone has cello tuning. Finally there are two basses, with their strings tuned in fourths. These new instruments have appeared in several concerts and have aroused considerable interest, both musically and scientifically.

1.3.3 Viols A viol differs from a violin primarily in that its back is flat and braced by three cross bars, the middle one supporting the sound post. The top plate has a shallower arch than that of the violin, and it has *c*-holes rather than *f*-holes. The ribs are considerably higher than those of the violin.

Viols were popular instruments during the 16th and 17th centuries, dropped from favor in the 18th century, and have enjoyed a considerable renaissance in the 20th century. The three most common members of the viol family are the treble, tenor, and bass. Each has six strings tuned in fourths except for a major third at the center.

1.4 Piano

The pianoforte, invented in 1709 by Bartolomeo Cristofori in Florence, was a direct descendent of the harpsichord. Many years of development led to the modern piano, which has become the most popular and versatile of all musical instruments. It has a range of more than seven octaves (A_0 to C_8) and a wide dynamic range. Pianos vary in size from small home uprights or spinets to large concert grands.

A typical concert grand piano has 243 strings, varying in length from about 2 m at the bass end to about 5 cm at the treble end. Included are eight single strings wrapped with one or two layers of wire, five pairs of strings (also wrapped), seven sets of three wrapped strings, and 68 sets of three unwrapped strings. Smaller pianos may have fewer strings, but they play the same number of notes: 88. The soundboard is nearly always made of spruce, about 1 cm thick, with its grain running the length of the piano. Ribs on the underside of the soundboard stiffen it in the cross-grain direction. The soundboard is the main source of radiated sound, as is the top plate of a violin or cello.

The upright piano developed about the middle of the 19th century. The nearly rectangular soundboard and the strings in upright pianos are vertical, and the hammers travel horizontally; thus, the action is different from that of the grand piano. In full-size upright pianos, which stand 130 to 150 cm ($4\frac{1}{2}$ to 5 feet) in height, the striking mechanism or action is located some distance above the keys and connected to them mechanically by thin rods called stickers. In studio uprights or console pianos, which stand about 100 to 130 cm ($3\frac{1}{4}$ to $4\frac{1}{4}$ feet) in height, the action is mounted directly over the keys without stickers. In small spinet pianos (less than 100 cm in height), the action is below the keys and drop stickers transmit key motion to the action.

Piano strings make use of high-strength steel wire. Efficiency of sound production calls for the highest string tension possible, while at the same time, minimizing inharmonicity calls for using the smallest string diameter (core diameter in a wrapped string) possible. This results in tensile stresses of around 1000 $\mathrm{N/mm^2}$, which is about half the yield strength of steel wire. For steel with an elastic modulus of 2×10^{11} $\mathrm{N/m^2}$, this results in an elongation of about $\frac{1}{2}\%$ when the string is under tension. Fortunately, when strings break, it is usually near the keyboard end so that the broken strings recoil away from the pianist.

A completely flexible string would vibrate in a series of modes that are exact harmonics of a fundamental. Actual strings have some stiffness, however, which provides a restoring force (in addition to the tension), slightly raising the frequency of all the modes. The additional restoring force is greater in the case of the higher modes because the string makes more bends. Thus the modes are spread apart in frequency and are no longer exact harmonics of a fundamental. In other words, a real string with stiffness is partly stringlike and partly barlike.

The inharmonicity of strings (that is, the amount by which the actual mode frequencies differ from a harmonic series) is found to vary with the square of the mode number. Thus, the second mode ("harmonic") is shifted four times as much as the fundamental.

Inharmonicity is especially noticeable in the case of the large bass strings, because the stiffness of a string increases sharply with its radius. Since wrapped strings are more flexible than are solid strings of the same diameter, the inharmonicity of the bass strings is reduced substantially by the use of wrapped rather than solid strings of the same weight. (The lower strings on a guitar and violin are wrapped rather than solid, for the same reason.)

An approximate expression for the inharmonicity $I_n = f_n/nf_1^0$ of the nth partial (spectral component) of a solid string can be written as

$$I_n = Bn^2, \tag{5}$$

where the inharmonicity constant B is given by

$$B = \frac{\pi^2\beta 2}{8} = \frac{\pi^3 a^4 E}{8L^2T}, \tag{6}$$

where a = radius, E = elastic modulus, L = length, and T = tension.

The inharmonicity in strings is the main reason why pianos are "stretch-tuned" (i.e., tuned so that octaves have frequency ratios slightly greater than 2). If they were not, the upper partials of a note would be slightly sharp with respect to the notes in the upper octaves to which they correspond, and undesirable beats would result when chords are played.

1.4.1 Piano Soundboards The soundboard is the main radiating member in the instrument, transforming some of the mechanical energy of the strings and bridges into acoustic energy. In addition, it opposes the vertical components of string tension that act on the bridges. This vertical force is in the range of 10–20 N per note, or a total of about 900–1800 N. The acoustical and structural functions are not totally independent of one another.

Nearly all piano soundboards are made by gluing strips of spruce together and then adding ribs that run at right angles to the grain of the spruce. The ribs are designed to add enough cross-grain stiffness to equal the natural stiffness of the grain along the wood, which is typically about 20 times greater than across the grain. In some pianos, the net cross-grain stiffness, including the contribution of the treble bridge, ends up being greater than the stiffness along the grain, a condition that is described as overcompensation. However, overcompensation is generally avoided in pianos of high quality. The unloaded soundboard is usually not a flat panel, but has a crown of 1–2 mm on the side that holds the bridges. When the strings are brought up to tension, the downward force of the bridge causes a dip in the crown, and in older pianos, the downward bridge force may have permanently distorted the soundboard.

Mode shapes of the soundboard of a 9-ft concert grand piano are shown in Fig. 13. At the lowest frequency, the entire soundboard moves in a single phase, but at the higher frequencies, the soundboard vibrates in several segments.

An idealized plate, clamped at its edges, has a radiation resistance that increases with increasing frequency and equals the characteristic impedance of air at high frequency. This assumes an optimum impedance match between the plate and its surroundings and no internal losses. Losses do occur, however, and these result in a decrease in sound radiation in the upper treble range. At the bass end, below the critical frequency (i.e., the frequency at which the speed of bending waves in the soundboard equals the speed of sound in air), the radiation efficiency drops dramatically. This results in a favored region for acoustic radiation between about 200 and 2000 Hz.

1.4.2 Sound Decay Since the strings are the principal reservoir for storing vibrational energy in a piano, the rate of sound decay is mainly determined by how rapidly energy is extracted from the strings. This depends upon the coupling between the strings, the bridge, and the soundboard, as well as the tuning of the unison strings.

Most piano notes show a compound decay curve: the initial decay rate is several times greater than the final decay rate. Initial decay rates may vary from about 4 dB/s at the bass end of the scale to 80 dB/s at the treble end. Decay times for a full 60-dB decay may vary from 0.2 to 50 s.

The compound decay curve gives rise to a very important feature of piano sound. The rapid transfer of energy to the soundboard initially results in a high rate of sound radiation and gives the listener the impression of loudness. At the same time, the slow final decay rate gives the impression of a sustained sound.

At least two different phenomena contribute to the compound decay: a change in the predominant direction of vibration of the string from perpendicular to parallel during decay and the coupling between the strings in a unison group. The mechanical impedance of the soundboard, which varies considerably with location and with frequency, is always much greater than the impedance of the strings. Typically, the impedance ratio is on the order of 200:1. Thus, vibrational energy is transferred rather slowly from the strings to the soundboard. The mismatch is even greater for vibrations parallel to the soundboard, since the soundboard stiffness is very large in this plane.

Figure 14 compares the decay rates for vibrations perpendicular to and parallel to the soundboard in the same string. Note the slower decay rate for parallel vibrations as a result of the greater stiffness of the soundboard in its own plane. The hammer excites mainly perpendicular vibrations in the string, but because of the spiral wrapping and the mechanical termination of the string at the bridge, the polarization changes in the course of time. This leads to a compound decay curve with a rapid initial decay (when perpendicular vibrations are dominant) and a slower final decay (due to poor coupling of the parallel vibrations).

When a hammer strikes a tricord (set of three unison strings), it sets all three strings into vibration with the same phase. Thus, they all exert vertical forces with the same phase on the bridge, and energy is transferred at a maximum rate. Because of very small differences in frequency, however, they soon get out of phase, and the resultant force on the bridge is diminished. Thus, the decay rate starts high and diminishes, a second reason for the compound decay curves. A skilled piano tuner can balance the fast initial sound decay and the slowly decaying aftersound in a piano by the tuning of the unisons. If the unison strings are closely tuned, the aftersound will develop more slowly since it will require a longer time for the unison strings to lose phase coherence.

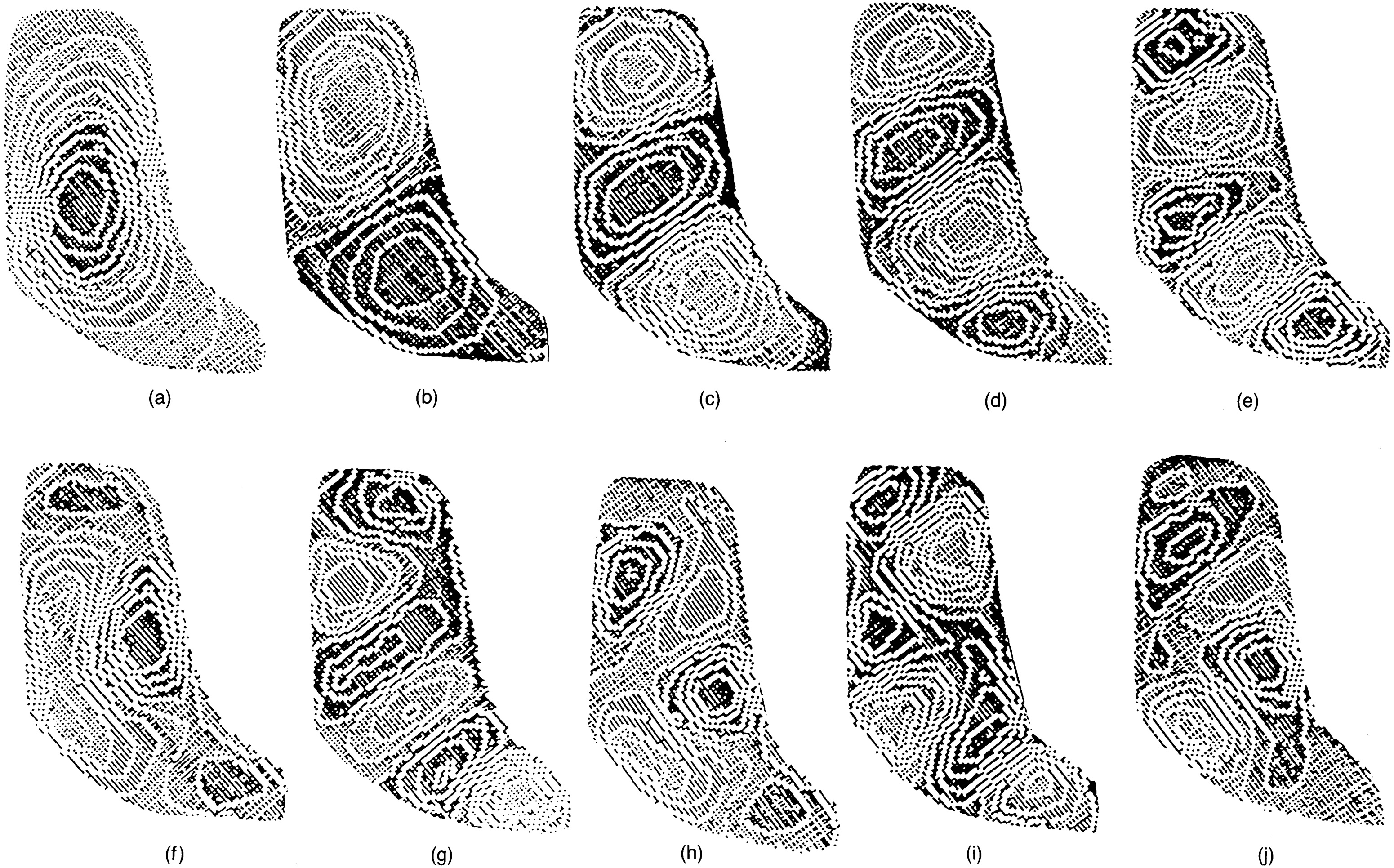

FIG. 13. Mode shapes on the soundboard of a 9-ft concert grand piano: (a) 52 Hz; (b) 63 Hz; (c) 91 Hz; (d) 106 Hz; (3) 141 Hz; (f) 152 Hz; (g) 165 Hz: (h) 179 Hz; (i) 184 Hz; (j) 188 Hz. (Kindel, 1989.)

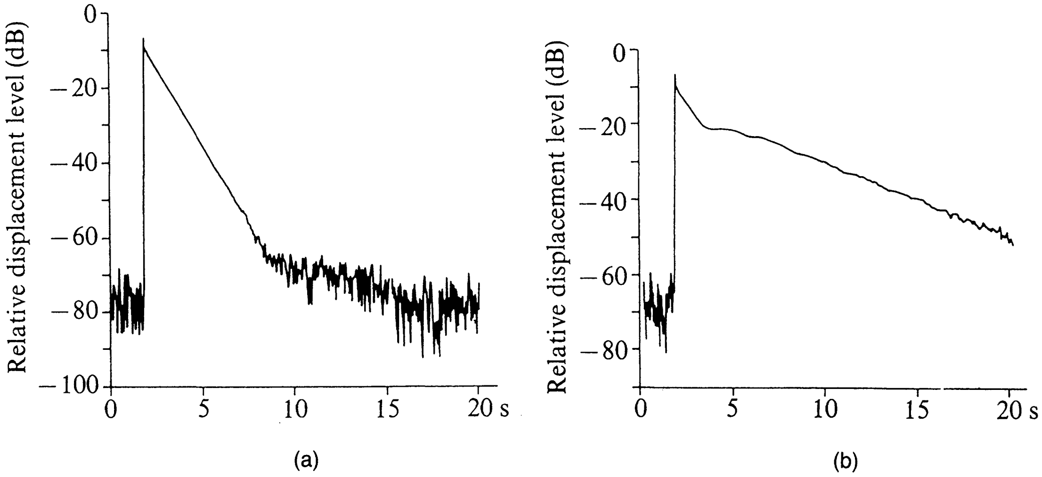

(a) (b)

FIG. 14. Decay curves for vibrations in a single D: string (f = 311 Hz) of a grand piano: (a) perpendicular polarization and (b) parallel polarization (Weinreich, 1977.)

2. WIND INSTRUMENTS

2.1 Vibration of Air Columns

Longitudinal waves in an air column behave much the same way as the transverse waves on a string discussed in Section 1.1. The acoustic pressure p (the fluctuation of pressure above and below atmospheric pressure) can be written as

$$\frac{\partial^2 p}{\partial t^2} = \frac{K}{\rho}\frac{\partial^2 p}{\partial x^2} = c^2 \frac{\partial^2 p}{\partial x^2}, \qquad (7)$$

where K is the bulk modulus, ρ is the density, and c is the wave speed. For propagation of sound waves in air, the adiabatic bulk modulus is the appropriate one to use, and the wave velocity c (in m/s) can be written as

$$c = \sqrt{\gamma RT/M} = 20.1\sqrt{T}, \qquad (8)$$

where T is absolute temperature. At 20 °C (T = 293 K), for example, c = 343 m/s.

A harmonic plane wave propagating in the x direction can be written as $p(x,t) = A \sin [k \times (ct - x)]$, where $k = 2\pi/\lambda$. In a pipe of length L with open ends (so that $p = 0$ at $x = 0$ and $x = L$), this leads to standing waves, and the pressure equation for the nth mode is similar in form to Eq. (3) for the transverse motion of a vibrating string:

$$p_n(x,t) = A_n \sin(2\pi f_n t + \phi_n) \sin(2\pi f_n x/c), \qquad (9)$$

where $f_n = nc/2L$ is the frequency of the nth normal mode, A_n is its amplitude, and ϕ_n its phase.

In actual practice, the sound pressure does not drop to zero exactly at the end of a pipe but rather a small distance ΔL beyond the open end. The end correction ΔL for a cylindrical pipe of radius r without a flange is approximately 0.61 r (at low frequency when $r \ll \lambda$). Twice this amount should be added to the physical length of a pipe with two open ends (an "open" pipe) to obtain the acoustic length.

In a pipe with one open end and one closed end (a "closed" pipe), the mode frequencies are given by $f_n = nc/4L$ (n = 1, 3, 5 . . .). The lowest mode has half the frequency of the open pipe, and only odd-numbered harmonics are present. Note that at a closed end, where the acoustic pressure is maximum, the acoustic velocity and the displacement are minimum, while the opposite is true at an open end.

The modes or resonances of a conical pipe have essentially the same frequencies as an open pipe of the same length. This statement is true even if the cone is truncated, which may appear paradoxical at first glance. As a sound wave travels toward the small end of a cone, its pressure amplitude increases. Pressure distributions for the first three modes of a cone are shown in Fig. 15, along with the corresponding modes of an open pipe.

The cylindrical pipe in Fig. 15 is an example of an acoustic horn, a horn being de-

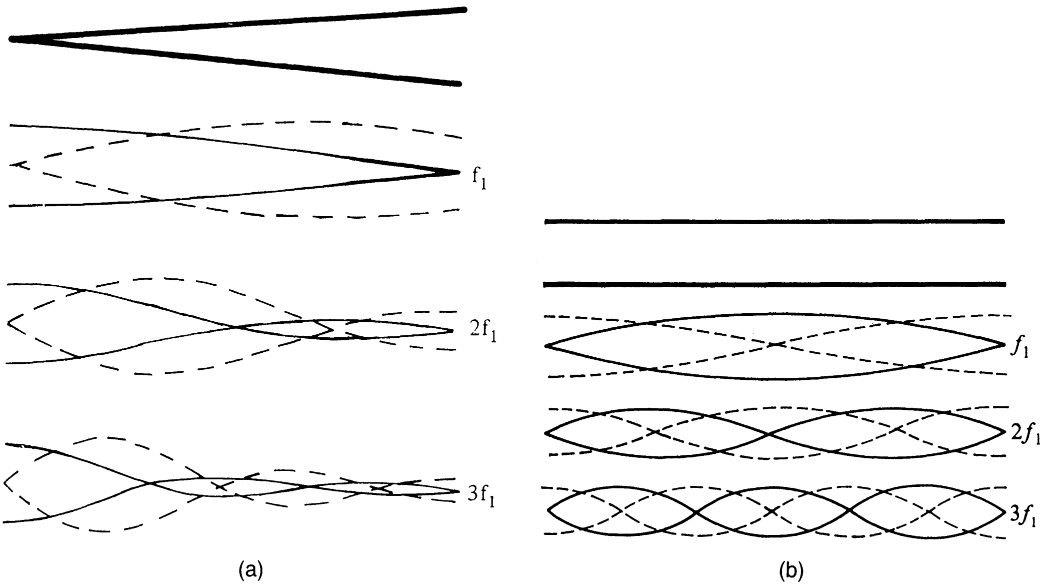

FIG. 15. Pressure (solid line) and velocity (dashed line) distributions for the first three modes of (a) a cone and (b) an open pipe. (From Rossing, 1990.)

fined as a conduit whose length is large compared with its lateral dimensions. Horns of various types are used in instrument bores, many of which can be approximated by Bessel horns whose radii are given by

$$r = bx^{-\epsilon} \tag{10}$$

If $\epsilon = 0$, the horn is cylindrical; if $\epsilon = -1$, it is conical. Most brass instruments have horn profiles that are nearly cylindrical for about the first half of their length, then expand with a flare that is well approximated by a Bessel horn (see Chap. 8 of Fletcher and Rossing, 1991).

2.2 Reed and Lip Vibrations

Wind instruments are made to sound either by blowing a jet of air across an opening (as in flutes and flue organ pipes) or by buzzing together the lips or a thin reed and its support (as in trumpets, clarinets, oboes, or reed organ pipes). In the latter case, the reed that controls the air flow (we can consider the brass player's lips as a type of reed) acts as an inward-striking valve that is blown closed [Fig. 16(a)] or as an outward-striking valve that is blown open [Fig. 16(b)]. A woodwind reed acts as a blown-closed valve, while the brass player's lips act as a blown-open valve.

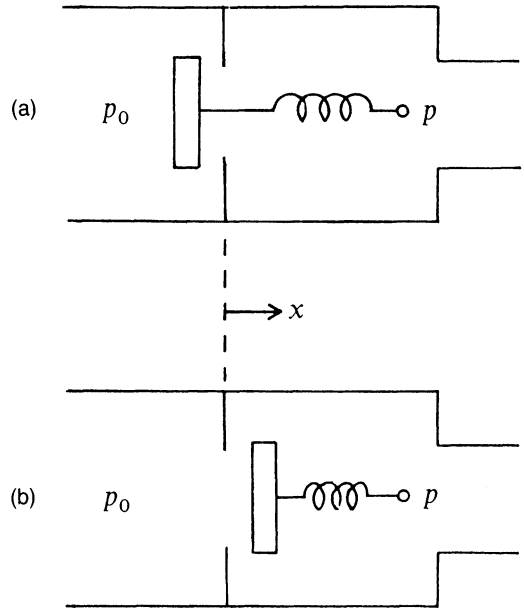

FIG. 16. The two basic types of a pressure-controlled valve. In each case, the valve cap is attached to a spring. In (a), the blowing pressure p_0 tends to blow the valve closed, while in (b), the blowing pressure tends to open the valve. The internal pressure p is the control variable and the coordinate x gives the position of the cap relative to its seat. (From Fletcher and Rossing, 1991.)

For the simple valves in Fig. 16, $x - x_0 = \gamma(p - p_0)$, where x is the position of the valve cap relative to its seat, and x_0 is its equilibrium position [$x_0 < 0$ in Fig. 16(a), $x_0 > 0$ in Fig. 16(b)]. The average volume flow U through the valve can be written as

$$U = |ax_0(p - p_0)^{1/2} + b(p - p_0)^{3/2}|, \qquad (11)$$

where a and b are positive constants.

For the blown-closed valve in Fig. 16(a), this leads to the solid curve in Fig. 17, which has a region of negative resistance extending from point A to point C. Over this range, regenerative feedback can occur over a wide frequency range up to the reed resonance, but a pipe resonator is generally required. For the blown-open valve in Fig. 16(b), x_0 is positive and Eq. (9) leads to the dashed curve in Fig. 17 which has no region of negative resistance. For such a valve, sound generation relies on phase shifts that occur near the resonance frequency of the valve (see Chap. 13 in Fletcher and Rossing, 1991).

Although in principle it is possible to generate a steady tone using only the mouthpiece of a wind instrument, in practice a stable tone requires positive feedback from the air column. Cooperation between the air column and the lips or reed is similar to the electrical feedback that sustains oscillations in an audio generator.

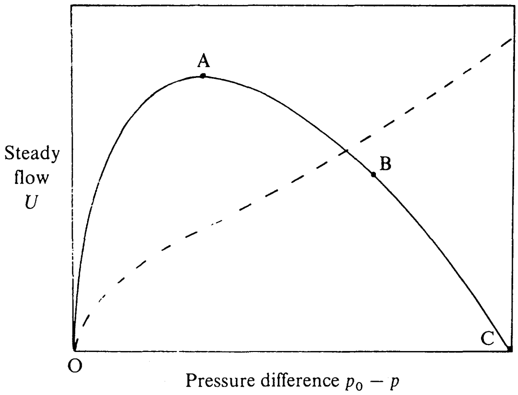

FIG. 17. The steady flow characteristic of a pressure-controlled valve of the type shown in Fig. 16(a) (solid line). The valve is blown completely closed at point C. The valve admittance is negative over the range AC. The broken line shows, to a different scale, the static flow behavior of a valve with a configuration as in Fig. 16(b). (From Fletcher and Rossing, 1991.)

2.3 Brass Instruments

The principal members of the brass family are the trumpet, French horn, trombone, and tuba. Their playing ranges are indicated in Fig. 18. Other important brass instruments are the cornet, fluegelhorn, bugle, and baritone horn.

Brass instruments typically have four sections: a mouthpiece, a tapered mouthpipe, a cylindrical section, and a bell, as shown in Fig. 19. The trumpet, French horn, and trombone have cylindrical sections of considerable length, while the fluegelhorn, baritone, and tuba, often characterized as instruments of conical bore, are tapered throughout much of their length.

2.3.1 The Trumpet The modern trumpet is a descendent of the valveless baroque or natural trumpet, which itself developed through many years of experimentation. The trumpet plays a very important role in symphony orchestras, military bands, jazz ensembles, and wind ensembles of all kinds. The most common trumpet is tuned in B♭, although trumpets tuned in other keys such as D, C, and F are also used. The B♭ trumpet bore is about 140 cm long and has a fundamental of B_2♭, although the lowest useful note is B_2♭.

One way to study the resonances of a pipe (or a wind instrument) is to measure its acoustic impedance as a function of frequency. (Acoustic impedance is defined as the sound pressure divided by the volume velocity.) The impedance of a B♭ trumpet is shown in Fig. 20. Note that the enhancement of impedance (pressure) peaks in the vicinity of the mouthpiece resonance (around 800 Hz). These large pressure peaks enhance the ability of the air column to control the rather massive lips of the player. Skillful design of the bore has brought the peak frequencies (except for the lowest one) very nearly into a harmonic relationship so that the air column resonances will respond to the partials of the desired tone.

Arrows at the top of Fig. 20 indicate the peaks (2, 4, 6, and 8) that correspond to the partials of B_3♭ (f = 233 Hz; this note is the written C_4 for an instrument tuned in B♭). When the note is blown very softly, the player's lips vibrate nearly sinusoidally, and only peak 2 is of importance. As the level of loudness increases, more harmonics are pro-

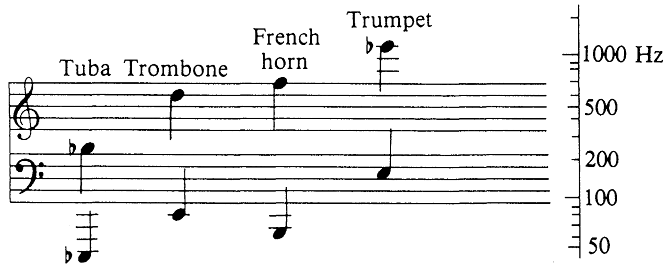

FIG. 18. Playing ranges of tuba, trombone, French horn, and trumpet.

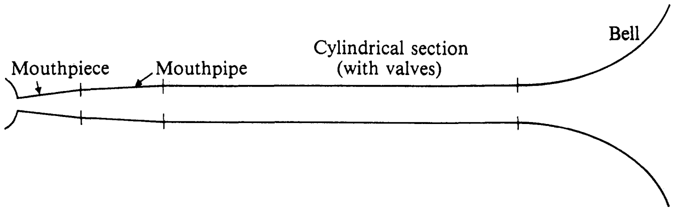

FIG. 19. A cross section of a brass instrument bore.

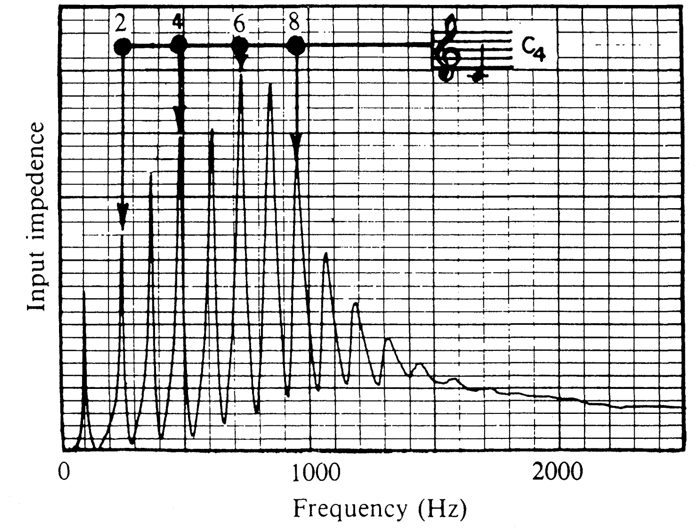

FIG. 20. Impedance curve of a B♭ trumpet, showing peaks that form a "regime of oscillation" for C_4 (from Benade, 1976).

duced, which excite peaks 4, 6, and 8. The tone takes on increasing stability as the lips become subject to control by pressure impulses corresponding to all four peaks.

The lowest peak in Fig. 20 lies below the fundamental frequency (about 117 Hz, corresponding to B_2♭). It is possible to play the "missing" fundamental note, sometimes called the pedal tone, by relying on feedback from other peaks having harmonically related frequencies. It is also possible to play at the frequency of the lowest peak, but that is not a useful note at all. Notes above F_5 (698 Hz) or so become increasingly difficult to play. The peaks that correspond to the harmonics are small, which indicates that the corresponding resonances are too weak to be of much help in controlling the lips and stabilizing the tone. Because peak 6 is very tall, F_5 is quite easy to play softly, indicating a strong resonance; little increase in stability is noted during crescendo, however, because peak 12, which lies near the second harmonic, is small. Above B_6♭ (written C_6), even the fundamental is weakly supported by the air column, and the lips must provide their own stability, which makes great demands on the player's ability to control his or her lips.

2.3.2 Valves and Slides On a brass instrument of fixed length (like the military bugle), it is possible to play eight or more notes, corresponding to the peaks in Fig. 20, and a few more may be added by skillful playing. To play the remaining notes of the scale, however, the acoustical length of an instrument must be varied. This is done by moving a slide (trombone) or by inserting lengths of additional tubing by means of valves (trumpet, French horn, baritone, tuba, etc.).

The playing positions of a trombone slide are shown in Fig. 21. Seven positions of the slide are needed to play all desired notes of the scale; note that the seven positions are not spaced equally. Going down a semitone decreases the frequency by about 6%, so the instrument should be lengthened by about this same amount. As the instrument gets longer, a 6% increase in length also becomes progressively longer.

In most brass instruments, however, a fundamental problem arises in designing the additional lengths of tubing for each valve. If the lengths of tubing are designed so that the first and second valves can lower the pitch by a whole tone and a semitone, respectively, then the combination will be inadequate to lower the pitch by three semitones, as illustrated by the following: To lower the pitch a whole tone, the first valve must add a length of tubing 12.2% of the total length L, or $0.122L$. To lower the pitch a semitone requires the addition of 5.9% of the total length, or $0.059L$. Thus the two valves together increase the length by 18.1%, or $0.181L$, but three semitones requires an increase of 18.9%.

Various compromises have been proposed to deal with this tuning problem. The valve slides for the first and second valves may be made slightly longer than the optimum length, thus accepting a slight flattening of the first two semitones in order to reduce the sharpening of the third one. Similarly, the third valve slide may be designed to give a shift slightly more than three semitones. Since the discrepancy is usually greatest when the third valve is used in combination with the others, most trumpets have a provision for moving the third valve slide with the left hand while the instrument is being played. Beyond that, the brass player must "lip" the various notes into tune, that is, make small corrections in frequency by changing the tension of the lips.

The valve problem becomes increasingly difficult in the lower brass instruments. A fourth valve is frequently added to the tuba; it lowers the pitch by a fourth, and thus substitutes for the troublesome combination of valves 1 and 3. Even a fifth or sixth valve is sometimes added, their function varying in the different tubas.

2.3.3 The French Horn The French horn in common use is a "double horn" composed of two horns tuned to F_2 and $B_2\flat$, which share the same mouthpiece and bell. The three main valves each have two sets of windways and valve slides, and a fourth valve is used to switch from one horn to the other. The B♭ horn is preferred for playing high notes, because it has stronger resonances at the higher frequencies, and they are spaced further apart.

Having a bore length of about 375 cm (30% longer than the trombone), the French horn has many resonances, and much of the time the horn is played at a pitch corresponding to one of the higher resonances. This means that many notes on a French horn lack the stability of notes normally played on a trumpet or trombone, where there are several prominent resonances corresponding to the overtones of the note played.

Placing the hand in the bell of the horn makes it easier to play the higher notes. By

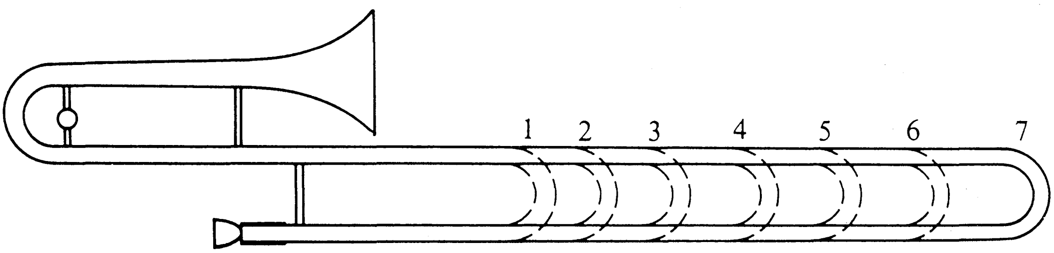

FIG. 21. Playing positions of a trombone slide.

inhibiting the radiation and increasing the reflection of the higher frequencies, it allows standing waves to build up and leads to more usable resonances at the higher frequencies. Placing the hand in the bell also lowers the pitch. If the hand is inserted as far as possible, a "stopped" tone is produced, which appears to be about a semitone higher than normal. Apparently all the resonances are lowered to the point where the next higher resonance is only a semitone above the desired note, which can be confirmed by holding a note as the hand is slowly inserted to the stopped position. If no attempt is made to hold the original pitch with the lips, the frequency will fall slowly. On the other hand, if the player tries to hold the pitch, the note jumps up to the next higher resonance at some point as the hand is inserted.

2.3.4 The Trombone The trombone, like the trumpet, has a bore that is predominantly cylindrical. Its tubing has about twice the length of a trumpet, however, so that its main resonances occur at frequencies that are approximately harmonics of a missing fundamental near $B_1\flat$ (58 Hz). The lowest note that is played in "first" position (i.e., with the slide fully retracted, see Fig. 21) is normally $B_2\flat$, although the $B_1\flat$ pedal note in a trombone is a more useful note than in a trumpet, and occasionally it is called for in musical scores.

The playing range of the ordinary tenor trombone extends up to D_5 (the tenth harmonic of the fundamental $B_1\flat$). The bass trombone has a slightly larger bore and a mouthpiece of larger volume than the tenor, making it easier to play the lower notes. Most bass trombones, and some tenor trombones as well, include a rotary valve and additional tubing to lower the fundamental to F_1. An excellent discussion of the acoustics of brass instruments appears in Chap. 9 of Campbell and Greated (1987).

2.4 Woodwind Instruments

The principal woodwinds in the symphony orchestra are the flute, clarinet, oboe, and bassoon; auxiliary instruments include the piccolo, English horn, bass clarinet, and contrabassoon. The saxophone is the principal woodwind in jazz bands and also plays an important role in marching and concert bands and, occasionally, in symphony orchestras. Old instruments, such as recorders and crumhorns, have recently experienced a revival in interest.

Most woodwinds were originally constructed from wood; hence their family name. Modern woodwinds may be either wood, metal, or plastic, although wood is still the preferred material for woodwinds other than flutes, piccolos, and saxophones. Like the brass instruments, woodwinds make use of feedback from an oscillating air column to control the flow of air input and maintain the oscillations. The flow-control valve may be either a vibrating reed or an oscillating stream of air. The feedback mechanisms in these two cases are quite different, as will be discussed in this section.

In woodwinds, the resonances of the air column are tuned by opening and closing tone holes with the fingers and with mechanical keys. Sound is radiated from the open tone holes, so that the radiation pattern becomes more complex than that of the brass instruments, which radiate virtually all their sound from the bell.

2.4.1 The Clarinet The clarinet is a single-reed instrument with a cylindrical bore about 15 mm in diameter. The cane reed is clamped by the ligature against a specially designed surface of the mouthpiece called the table. The opening between the tip of the mouthpiece and the reed is about 1 mm. When the clarinet is played, the lower lip pushes the reed in to about half this distance, and it vibrates around this position. For soft tones, the tip of the reed does not touch the mouthpiece; therefore, the flow of air is not interrupted. As the blowing pressure is increased, the amplitude of the reed increases, until for loud tones the tip of the reed touches the mouthpiece during about half of the cycle. Thus the flow of air, which is more or less sinusoidal for soft playing, takes on more harmonics as the blowing pressure increases.

A clarinet reed has a natural frequency of its own at around 2000–3000 Hz; oscillations at this frequency are prevented largely by the damping action of the lower lip pressed against the reed. If the clarinet is blown with the teeth against the reed rather than the lip, unpleasant squeaks and squeals appear.

The B♭ clarinet has a wide playing range of $3\frac{1}{2}$ octaves divided into three registers, as shown in Fig. 22. The low or *chalumeau* register extends from D_3 to E_4, and a second register, the *clarion*, plays a 12th above this, since the first overtone of a cylindrical closed tube is the third harmonic of the fundamental. Notes above $B_5\flat$ can be played in the *altissimo* register, which uses the third mode (fifth harmonic) of the air column. Often the "throat tones" (G_4 to $A_4\sharp$), which use open keys in the throat of the clarinet, are included in the chalumeau register.

The low tones of a clarinet are particularly rich in harmonics, although the second harmonic and other even-numbered harmonics are weak throughout the chalumeau register. The preponderance of odd-numbered harmonics gives the clarinet its "hollow" or "woody" quality. In the clarion register, the even-numbered harmonics are strong, resulting in quite a different tone quality from the lower register.

Orchestral clarinetists alternate between clarinets tuned in A and B♭, the choice usually depending on the key signature of the music. The bass clarinet, which plays an octave lower than the B♭ clarinet, is also used in orchestras and bands. Other clarinets are the E♭ soprano and E♭ alto. The C clarinet, used in the Classical period, is rarely seen today, but the basset horn in F (an alto clarinet with a narrow bore) is played in early-music ensembles.

When a pipe has more than one open hole, its acoustical behavior exhibits several interesting features. If the open holes are regularly spaced, they constitute a tone-hole lattice, not unlike a lattice of atoms in a crystal or a line of beads spaced equally on a string. The open tone-hole lattice acts as a filter that transmits waves of high frequency but reflects those of low frequency. The critical frequency above which sound waves can propagate through a lattice of tone holes is called the *cutoff frequency* of the lattice, which has been found to be an important factor in determining the timbre of a woodwind instrument.

The open-tone-hole cutoff frequency in most good clarinets is around 1500 Hz. A clarinet with a higher cutoff frequency tends to have a "bright" tone; an instrument with a lower cutoff frequency has a "dark" tone. In an experiment in which two matching clarinets were reworked to raise the cutoff frequency slightly in one and to lower it in the other, it was found that players of classical music preferred the one with the lower cutoff frequency, while jazz clarinetists chose the one with the higher cutoff and "bright" tone (Benade, 1976).

2.4.2 The Saxophone The saxophone is a conical-bore single-reed instrument invented in 1846 by Adolphe Sax. Although used occasionally in symphony orchestras, its main use has been in all types of bands and in ensembles that play jazz and popular music. The four main members of the saxophone family are the soprano in B♭, the alto in E♭, the tenor in B♭, and the baritone in E♭. Saxophones are characterized by a large bore diameter, a low input impedance, and a louder sound than the other woodwinds.

Since the saxophone has a conical bore, its spectrum has even-numbered harmonics as well as odd-numbered ones. Its cone angle (3° to 4°) is considerably greater than that of the oboe or bassoon, however, and so saxophone tone has fewer prominent harmonics than the double reed instruments. The larger bore of a saxophone allows correspondingly large tone holes that radiate efficiently, and thus it can be played loudly.

2.4.3 Double-Reed Instruments The family of orchestral double reeds includes the oboe, English horn, bassoon, and contrabassoon. Basically, they are conical tubes with

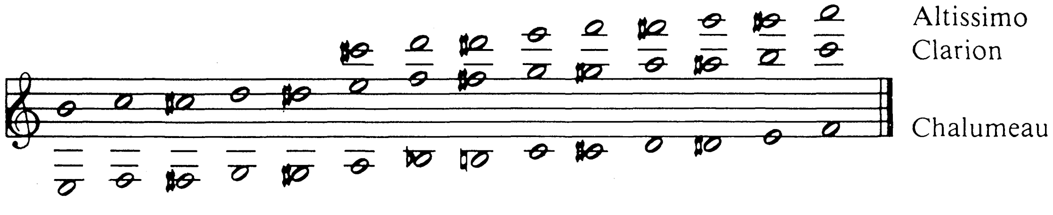

FIG. 22. Three registers of the clarinet. Shown are the notes fingered; notes sounded on a B♭ clarinet are one whole tone lower.

the tip of the cone cut off and a reed attached. All these instruments use a double reed consisting of two halves of cane beating against each other. Small mouthpieces with a single reed have been used in place of the double reed, but they have not become very popular because they change the quality of the instrument.

The oboe bore is a nearly straight cone about 60 cm in length with a cone angle of about 1.4°. Since the resonances of a cone are an octave apart, the registers of the oboe are also an octave apart, extending from D_4 to C_5 and D_5 to C_6, although additional keys and fingerings extend the playing range from $B_3\flat$ to G_6. Oboes are usually constructed from three pieces of wood: a top section, a bottom section, and a bell, with all three sections having tone holes.

Quality oboes have cutoff frequencies that are nearly constant throughout their playing range and, for different instruments, vary from about 1100 to 1500 Hz (Benade, 1976). A higher cutoff frequency results in a bright tone, and a lower cutoff frequency in a dark tone, just as in the case of clarinets. The spectra of oboe sounds show substantial amounts of the higher harmonics along with broad resonances or formants in the neighborhood of 1000 and 3000 Hz. These formants may be related to the mechanical properties of the reed.

The English horn or cor Anglais is an alto version of the oboe with a pear-shaped bell that gives a distinctive tone quality to certain notes near its resonance. Composers have taken note of this distinctive tone quality in writing solos for this instrument. The English horn is tuned in F, a fifth below the oboe.

The bassoon has a nearly conical bore with a cone angle slightly less than 1° and with a total length of about 254 cm. The tube makes several bends in order to be of a manageable size. The playing range of the bassoon extends from $B_1\flat$ to about C_5 (58–523 Hz), while the contrabassoon plays an octave lower. Open tone-hole cutoff frequencies for quality bassoons range from about 350 to 500 Hz (Benade, 1976). Like the oboe, the bassoon has many sharp resonances, and the tone is rich in harmonic overtones.

2.4.4 The Flute A flow-controlled air stream, such as that which occurs in a blown bottle or a flute, collaborates with an air column to oscillate at frequencies at which the air column has dips in its impedance curve (that is, where the pressure is at a minimum and, therefore, the flow-controlling jet deflection is at a maximum). This is opposite to the woodwind reed or brass player's lips, which are controlled by pressure peaks.

The flute is basically a cylindrical pipe, about 19 mm in diameter, except for the head joint, which tapers to about 17 mm at the mouth or embouchure hole. Its vibration modes thus form a series of frequencies that includes all the harmonics of the fundamental. The resonance frequencies of the pipe are influenced by the presence of the air stream, however. In particular, the frequency of the lowest resonance is raised when the blowing pressure increases, whereas the second resonance remains more or less unchanged. The range of the flute is normally B_3 to D_7.

The flute plays in three registers, and the player changes registers, without the aid of register keys, by adjusting the technique of blowing. The three parameters at the player's control are the blowing pressure, the length of the air jet, and the area of lip opening. The technique used by most flute players includes adjustments in all three of these (Fletcher, 1975).

The most efficient excitation of the fundamental takes place when the time for the air jet to travel across the embouchure hole is about half the period of an oscillation. If the jet travel time is shortened much below this, the fundamental will not sound. Thus, to sound the second register, the player moves the lips forward to decrease the jet length and/or increases the blowing pressure. At the same time the blowing pressure is increased, the size of the lip opening is decreased to maintain the loudness and tuning. Studies of a number of flute players by Fletcher (1975) show that to shift to a register an octave higher, a flute player will typically double the blowing pressure, reduce the jet length by 20%, and reduce the lip opening by about 30% (Doubling the blowing pressure increases the jet speed by 40%, since jet speed is approximately proportional to the square root of pressure.)

The angle of blowing depends on the shape of the player's lips, but typically ranges from 25° to 40° below the horizontal. Most players use a shallower angle for low notes and for loud playing.

The narrow end of the head joint is closed by a cork plug whose position can be adjusted by a screw. Changing the length of the small cavity between the cork and the embouchure hole can have a substantial effect on the tuning and "playability" of certain notes.

The piccolo is about one-half the length of the flute, and it sounds one octave higher. An alto flute, tuned a fourth lower, and a bass flute tuned an octave lower than the standard flute, are occasionally seen.

2.4.5 The Recorder The recorder, also known as the *Blockflöte*, flauto dolce, and English flute, is an early member of the flute family, now experiencing a revival. It has a reverse conical bore, tapering inward toward the foot, and a whistle-type mouthpiece with a fixed windway. The three most commonly played recorders are the descant (soprano) in C, the treble (alto) in F, and the tenor in C. A bass in F, a sopranino in F, and a great bass in C complete the family. Each instrument has a two-octave normal playing range.

The fixed windway makes the recorder easy to sound but somewhat difficult to play in tune at different dynamic levels. Without the flexibility to adjust the embouchure in the manner of the flautist, the recorder player changes to the upper register by half-opening the thumb hole and increasing the blowing pressure slightly. Wind pressure in an alto recorder varies from about 100 N/m^2 for the lowest note to about 500 N/m^2 for the highest.

The recorder has only eight tone holes. Therefore, in order to play a full chromatic scale, liberal use is made of cross-fingerings. Cross-fingerings leave one or more tone holes open and close other tone holes below these. In general, cross-fingered notes are not as stable as other notes, and the recorder player often changes the fingering slightly from instrument to instrument. Since the pitch produced by a given fingering may rise as much as a semitone (100 cents) from low to high blowing pressure, a skilled recorder player often changes fingerings with dynamic level.

2.5 Pipe Organs

The pipe organ has been called the "king of musical instruments." No other instrument can match it in size, range of tone, loudness, or complexity. The modern pipe organ includes a large variety of pipes, arranged into divisions or organs. The principal division of the organ is called the great organ, and it usually contains the most stops or voices. The pipes in the swell division are usually enclosed behind a set of shutters that can be opened or closed to change the loudness.

Organ pipes are organized into ranks of similar pipes. One rank of pipes will include one pipe for each note (61 in the divisions operated from a keyboard, 32 in the pedal division). Each stop on the organ usually corresponds to one rank of pipes, except in the case of mixture stops, which involve several ranks. On smaller organs, a rank of pipes may be included in more than one stop.

There are two basic types of organ pipes; flue (labial) pipes and reed (lingual) pipes. Flue pipes produce sound by means of a vibrating air jet, in a manner similar to the flute and the recorder. Reed pipes use a vibrating brass reed to modulate the air stream.

The essential parts of a metal flue pipe are shown in Fig. 23(a). The air jet passes through a flue or windway, a narrow opening between the languid and the lower lip. Sound is produced when the air jet encounters the upper lip and oscillates back and forth, sometimes blowing into the pipe, sometimes blowing out through the mouth of the pipe. The large flue pipes usually have ears on either side of the mouth to guide the air jet.

The stopped wooden flue pipe, shown in Fig. 23(b), uses a similar mechanism to produce sound. Since the the resonator is now a closed pipe, it need be only half as long as an open pipe in order to sound the same pitch. Wooden pipes usually have a square cross section and produce a sound with a flute-like quality.

There are three families of flue pipes: (1) diapasons, (2) flutes, and (3) strings. Pipes of the flute family usually have the least overtone content, while the bright-sounding strings have the most. String pipes are generally slender cylinders, whereas diapasons are open cylinders of somewhat greater diameter. Flute pipes come in several sizes and shapes, are constructed of either wood or metal, and may be open or closed. Closed pipes sound mainly the odd-numbered harmonics of the fundamental.

The reed pipe, shown in Fig. 23(c), has a

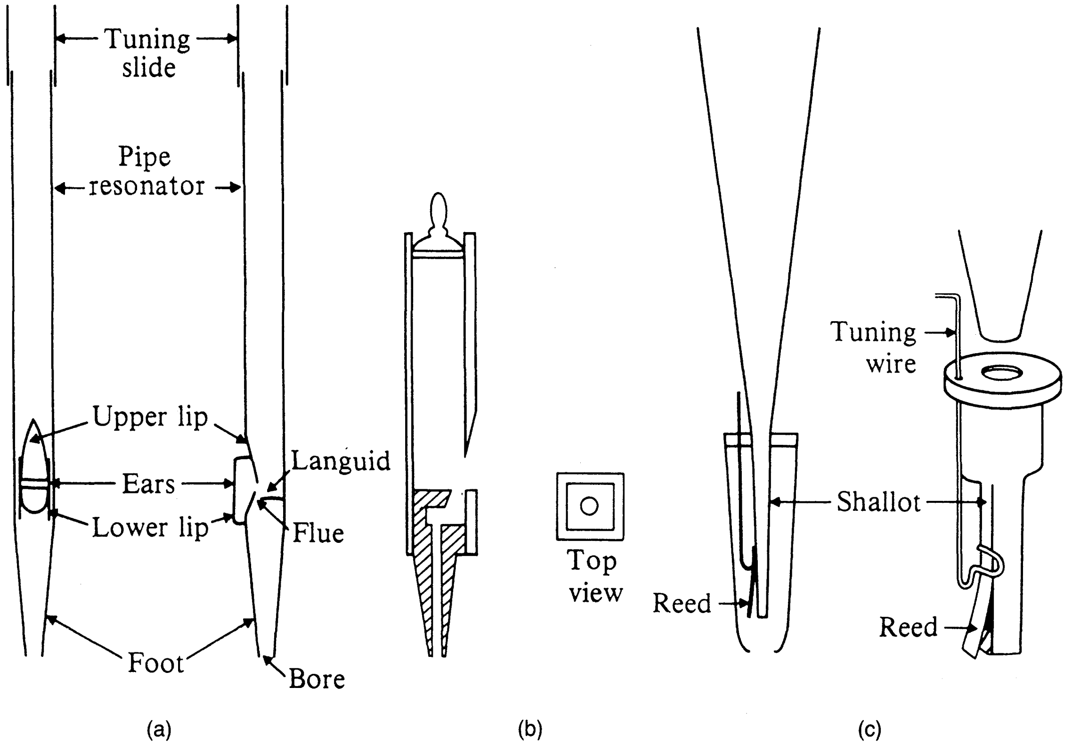

FIG. 23. Organ pipes: (a) open flue pipe of metal; (b) stopped wood flue pipe; (c) reed pipe, (From Rossing, 1990.)

vibrating reed or tongue, which modulates the flow of air passing through the shallot into the resonator. The reed is pressed against the open side of the shallot by a wire that can be adjusted up and down to tune the vibrating reed. Reed pipes are rich in harmonics. If the resonator is cylindrical, the odd-numbered harmonics will be favored, as in a clarinet. Conical resonators, however, will reinforce all harmonics, both odd and even.

Figure 24 shows several different pipes and the names given to the corresponding organ stops. Resonators of several different shapes are shown, including open and closed cylinders, cones, inverted cones, rectangular cylinders, and closed pipes with chimneys. (The chimney is tuned to boost one of the upper harmonics, around the fifth or sixth.)

Figure 25 compares the first four resonance frequencies of open and closed cylindrical, conical, and reverse conical pipes that can be used as organ pipe resonators. Note that the resonance frequencies of cone are essentially the same as those of an open cylinder, although the pressure amplitudes are quite different, as shown in Fig. 15 (Ayers *et al.*, 1985).

The *scale* of a rank of pipes refers to the ratio of diameter to length for the pipe of lowest pitch. Large-scale (large-diameter) pipes tend to have a dominant fundamental and fewer harmonics, whereas small-scale pipes have more harmonics. The reason for this is related to a frequency dependence of the end correction of a pipe. For an open cylindrical pipe, the end correction at the open end is approximately 0.6 times the radius, whereas at the mouth of a flue pipe, one adds approximately 2.7 times the radius to calculate the effective length (Strong and Plitnik, 1983). The end correction decreases with frequency, so the pipe effectively shortens for the higher partials; thus, the pipe resonances are slightly less than an octave apart. However, the spectrum of the sound source (the oscillating jet) has exact harmonics. Thus in a large-scale pipe, only the first few harmonics in the source are reinforced by the natural frequencies of the pipe. For pipes of small scale, such as the strings, the end correction

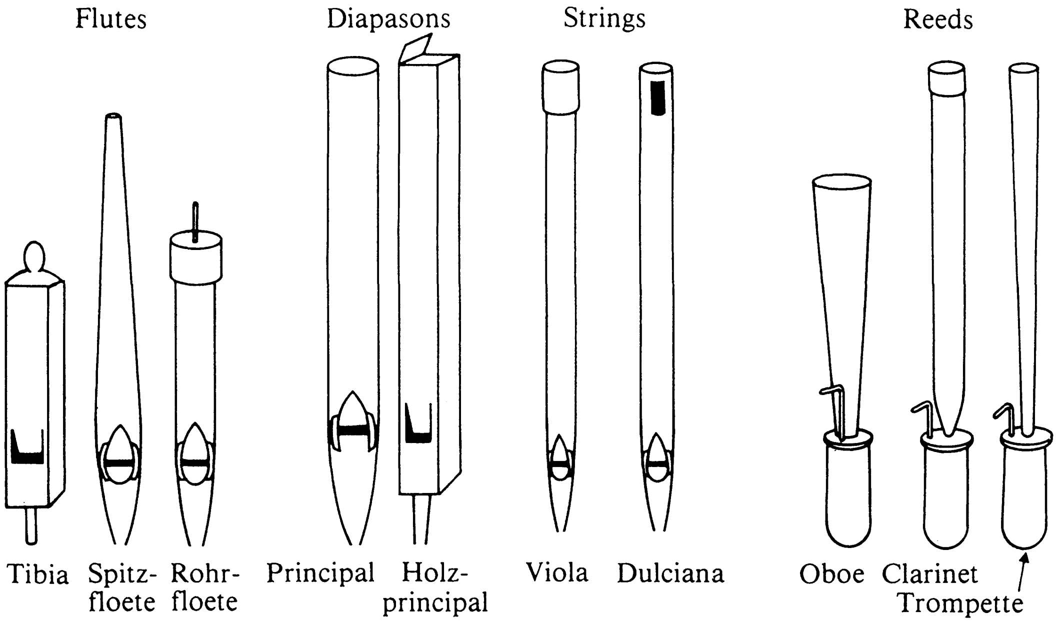
Flutes
Diapasons
Strings
Reeds
Tibia
Spitz-floete
Rohr-floete
Principal
Holz-principal
Viola
Dulciana
Oboe
Clarinet
Trompette

FIG. 24. Organ pipes of various families with the names of the corresponding stops (from Rossing, 1990).

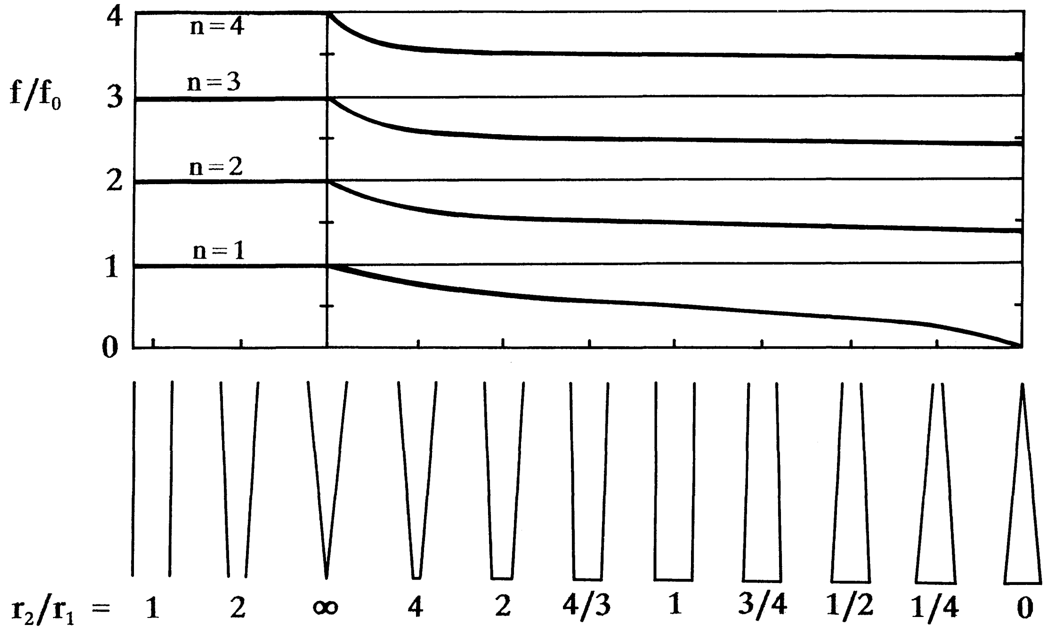

FIG. 25. First four resonance frequencies of open and closed cylindrical and conical pipes.

is much smaller to start with; thus, the pipe resonances more nearly match the harmonics of the oscillating jet.

Each rank of pipes is usually graduated in diameter according to a fixed relationship for that rank. For large-scale pipes, the diameter may reduce to one-half at the 17th note of the scale; for small-scale pipes, the diameter will reduce at a slower rate than this.

Normally, the mouth width, mouth height ("cut-up"), and width of the flue opening follow the same scale as the pipe diameter. The mouth width, for example, may vary from 0.6 times the pipe diameter (large-scale flute) to 0.8 times the diameter (string or diapason). Similarly, the cut-up may vary from 0.25 times the mouth width (diapason) to 0.4 times the width (flute); the flue width is typically about 0.03 times the mouth width.

2.5.1 Sound Generation in Flue Pipes

When air is admitted at the foot of a flue organ pipe, it flows upward and forms a sheetlike jet as it emerges from the flue (see Fig. 23). The jet flows across the mouth of the pipe and strikes the upper lip, where it interacts both with the lip itself and with the air in the pipe resonator. The physics of this interaction is discussed in Chapter 16 of Fletcher and Rossing (1991).

Jets can be described as either *laminar* (arranged in layers or streamlines) or *turbulent* (characterized by eddies or vortices). Jets in organ pipes are almost completely turbulent, which makes it quite difficult to describe them mathematically. Somewhat paradoxically, however, a fully turbulent jet in an organ pipe is more stable than a laminar one, and organ builders go to some lengths to ensure fully turbulent jets by cutting fine nicks along the edge of the languid, for example.

A fully turbulent jet mixes gradually with the surrounding air, thus slowing down and broadening out in a rather simple way. The jet velocity, which is maximum at the center, falls off in a bell-shaped curve whose width increases linearly with distance from the flue. Conservation of flow momentum requires that the velocity decreases with the square root of the distance from the flue slit.

The air flow during the attack transient in an organ pipe, although quite complicated, can be studied by means of Schlieren photographs. Figure 26 shows the jet at intervals of 0.1 s as the jet first strikes the edge, forming vortices. Eventually, this settles down to an acoustically driven jet oscillation.

Assuming that standing waves already exist in the pipe resonator, the air motion at the mouth associated with these standing waves will be essentially at right angles to the jet. This sets up a wavelike disturbance in the jet that propagates along the jet at a little less than half the speed of the air at the center of the jet. The amplitude of this wavelike disturbance grows nearly at an exponential rate along the jet, typically doubling during each millimeter of travel. The growth rate is found to be greatest when the wavelength along the jet is about six times the width of the jet at that location. On the other hand, when the length of the wave is less than the width of the jet, it does not grow at all, so that only long wavelengths can propagate with a large amplitude along long jets.

The large acoustic standing waves in the pipe drive the tip of the jet alternately inside and outside the upper lip. If the puffs of air enter the pipe in step with the standing waves, energy will be added. The important relationship between the acoustical air motion in the pipe mouth and the time the air pulse arrives inside the upper lip is determined by the time required for a wave on the jet to travel from the flue slit to the upper lip. This is, in turn, determined by the mouth height (cut-up) and the blowing pressure.

One of the most critical of all the steps in organ building is voicing the pipes, which means making coarse and fine adjustments in the various parts of the pipe so that it "speaks" properly. Much of the voicing can be done in the organ-builder's shop, but the final voicing or finishing is done after the organ is installed, and it takes into account the acoustics of the room, as well as the acoustics of the organ. One of the objectives of voicing is to achieve a uniformity of loudness and timbre within each rank of pipes. Another is to adjust the initial transient or attack of each pipe.

3. PERCUSSION INSTRUMENTS

3.1 Vibrations of Bars, Membranes, and Plates

Percussion instruments generally use one or more of the following basic types of vibrators: strings, bars, membranes, plates, air

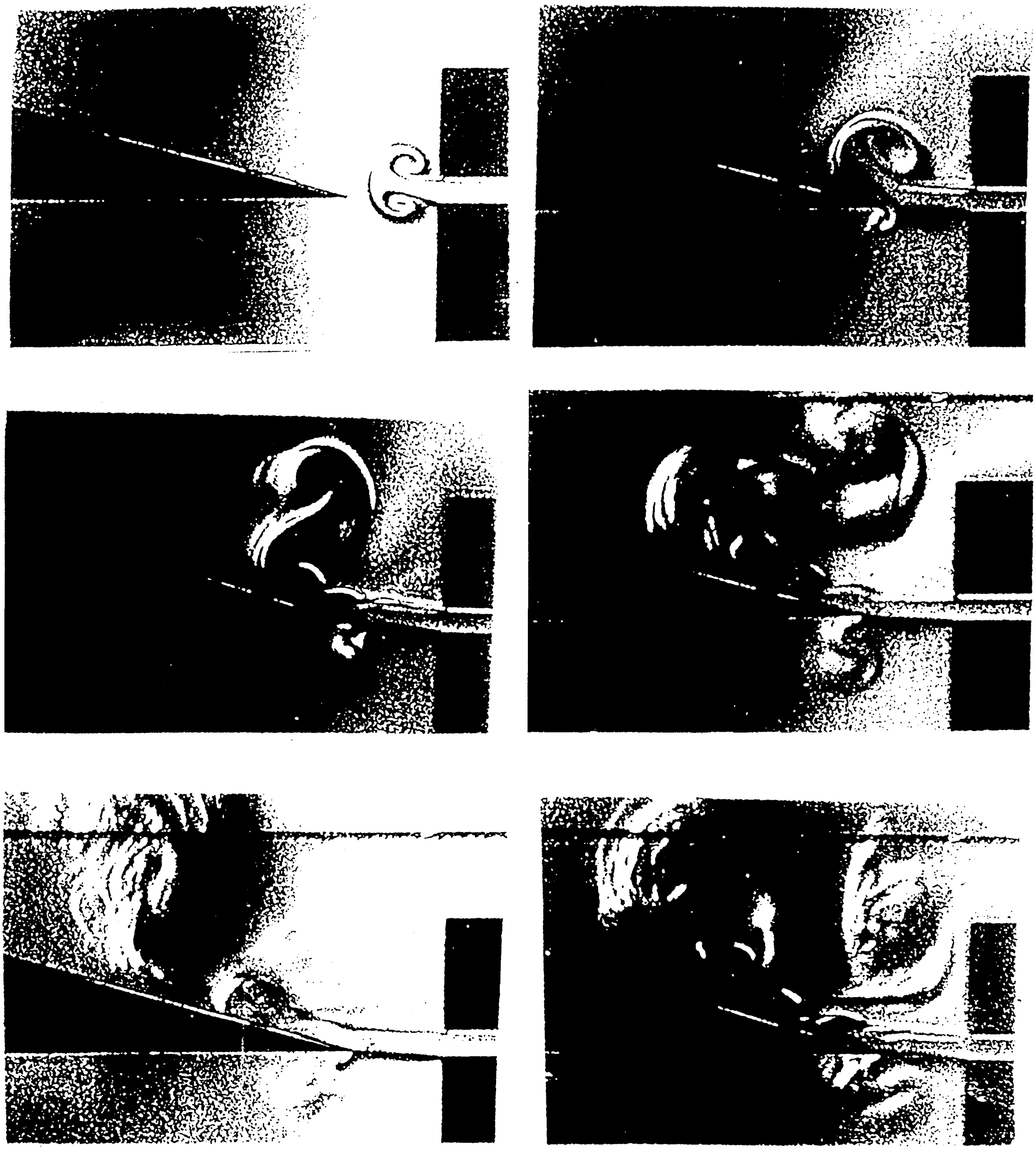

FIG. 26. Schlieren photographs showing air flow at intervals of 0.1 s during the attack when a jet strikes a sharp edge such as the labium of an organ pipe (photos furnished by A. Hirschberg).

columns, or air chambers. Strings and air columns tend to produce harmonic overtones; the others, in general, do not. Bars, membranes, and plates are three classes of vibrators whose modes of vibration are not related harmonically, and thus the overtones they sound will not be harmonics of the fundamental tone. The inharmonic overtones of these complex vibrators give percussion instruments their distinctive timbres.

3.1.1 Vibrations of Bars Although the compressional and torsional wave speeds in a bar show little or no dispersion (i.e., frequency dependence), this is definitely not the case with bending waves, whose speed is nearly proportional to $\sqrt{f}$. Bending waves involve both compressional and shear strains.

When a bar is bent, the outer side is stretched and the inner side is compressed; somewhere between is a neutral axis whose length remains unchanged. The simplest theory for such bending motion is the Euler-Bernoulli beam theory, which gives accurate results in thin bars and rods at low frequency. The Euler-Bernoulli equation of motion can be written

$$\frac{\partial^2 y}{\partial t^2} = \frac{-EK^2}{\rho}\frac{\partial^4 y}{\partial x^4}, \qquad (12)$$

where E is Young's modulus, K is the radius of gyration for the cross section (about the neutral axis), and ρ is density. This equation of motion differs from the wave equations for compressional and torsional waves in that it is a fourth-order equation; $f(vt \pm x)$ is not a solution. Harmonic solutions may be obtained by substituting $y(x,t) = Y(x)e^{j\omega t}$, giving a solution with four arbitrary constants to be determined from the end conditions,

$$\begin{aligned} Y(x) = {} & A\cosh(\omega x/v) + B\sinh(\omega x/v) \\ & + C\cos(\omega x/v) + D\sin(\omega x/v). \end{aligned} \qquad (13)$$

Harmonic waves propagate at a phase velocity that is proportional to the square root of frequency and also to the square root of the longitudinal wave speed c_L:

$$v = \sqrt{2\pi f c_L}. \qquad (14)$$

The dispersion relationship $\omega = c_L K k^2$ has the same form as that of de Broglie waves for a free particle. The group velocity is twice the phase velocity ($v_g = d\omega/dk = 2v$).

Although the relatively simple Euler-Bernoulli beam theory gives an accurate description of bending waves in a thin beam or rod at low frequency, it predicts too low a wave speed in a thick bar or a bar vibrating at high frequency. For one thing, it assumes that plane sections remain plane, which is equivalent to neglecting shear deformations. For another thing, it neglects rotary inertia. In Timoshenko beam theory, appropriate terms are added to the equation of motion to account for the effects of shear deformation and rotary inertia.

The modes of transverse vibration in a bar or rod depend upon the end conditions. Three different end conditions are commonly considered: free, simply supported (hinged), and clamped. For each of these, a pair of boundary conditions can be written. At a free end, there is no torque and no shearing force, so that the second and third derivatives are both zero; at a simply supported end, there is no displacement and no torque, so that y and its second derivative are zero; at a clamped end, y and its first derivative are zero.

There are six different combinations of these end conditions, each leading to a different set of vibrational modes. Three of the more common combinations are shown in Fig. 27.

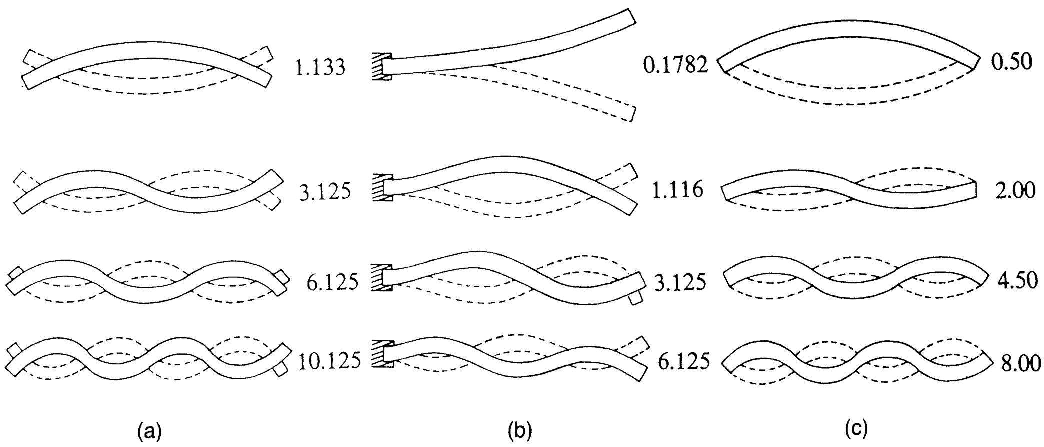

FIG. 27. Bending vibrations of (a) a bar with two free ends, (b) a bar with one clamped end and one free end, and (c) a bar with two supported (hinged) ends. The numbers are relative frequencies for a thin bar; to obtain actual frequencies, multiply by $(\pi K/L^2)/\sqrt{E/\rho}$.

In a bar with free ends, the vibrational frequencies are given by

$$f_n = \frac{\pi K}{8L^2}\sqrt{\frac{E}{\rho}}(2n+1)^2 = \frac{0.113h}{L^2}\sqrt{\frac{E}{\rho}}(2n+1)^2,\ n = 1, 2, 3\ \ldots, \tag{15}$$

where h is the thickness of the bar (greater precision results from using 3.011 in place of $2n + 1 = 3.0$ when $n = 1$).

3.1.2 Vibrations of Membranes The wave equation for a thin membrane can be written as

$$\frac{\partial^2 z}{\partial t^2} = \frac{T}{\sigma}\nabla^2 z = c^2\nabla^2 z, \tag{16}$$

where T is the tension, σ is the mass per unit area, and c is the wave velocity. For a circular membrane, the solutions are generally written as Bessel functions of order m, and the zeros of these functions give the frequencies of the various modes of vibration. The first 14 modes of an ideal membrane are shown in Fig. 28. The (m, n) mode corresponds to the nth zero of the Bessel function of order m.

The normal-mode frequencies of real membranes may be quite different from those of an ideal membrane given in Fig. 28. The principal effects in the membrane acting to change the mode frequencies are air loading, bending stiffness, and stiffness to shear. In general, air loading lowers the modal frequencies, while the other two effects tend to raise them. In thin membranes, air loading is usually the dominant effect.

The effect on frequency of the air loading depends upon the comparative velocities for waves in the membrane and in air, and also upon whether the air is confined in any way. A confined volume of air (as in a kettledrum, for example) will raise the frequency of the axisymmetric modes, especially the (0,1) mode. When a membrane vibrates in open air, however, all the modal frequencies are lowered, the modes of lowest frequency being lowered the most. The confining effect of the kettle enhances this mass loading and resultant frequency lowering in the nonaxisymmetric modes such as (1,1) and (2,1).

3.1.3 Vibrations of Plates The equation of motion for bending or flexural waves in a plate is

$$\frac{\partial^2 z}{\partial t^2} + \frac{Eh^2}{12\rho(1 - \nu^2)}\nabla^4 z = 0, \tag{17}$$

where ρ is density, ν is Poisson's ratio, E is Young's modulus, and h is the plate thickness. Bending waves in a plate are dispersive; that is, their velocity depends on frequency:

$$v(f) = \sqrt{1.8fhc_L}, \tag{18}$$

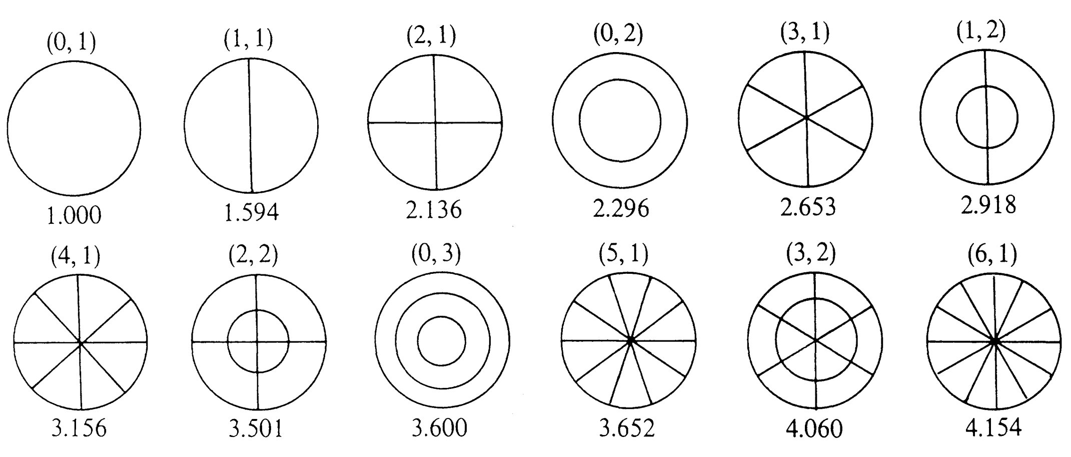

FIG. 28. First 14 modes of an ideal membrane. The mode designation (m,n) is given above each figure and the relative frequency below. To convert these to actual frequencies, multiply by $(2.405/2\pi a)\sqrt{T/\sigma}$, where a is the membrane radius.

where c_L is the longitudinal wave speed given by $c_L = \sqrt{E/\rho(1 - \nu^2)}$.

For a circular plate, ∇ is expressed in polar coordinates, and $Z(r,\phi)$ can be a solution of either $(\nabla^2 + k^2)Z = 0$ or $(\nabla^2 - k^2)Z = 0$. Solutions of the first equation contain the ordinary Bessel functions $J_m(kr)$; solutions to the second, the hyperbolic Bessel functions $I_m(kr) = j^{-m}J_m(jkr)$, where $j = \sqrt{-1}$. Thus the possible solutions are given by a linear combination of these Bessel functions times an angular function:

$$Z(r,\phi) = \cos(m\phi + \alpha)[AJ_m(kr) + BI_m(kr)]. \tag{19}$$

The hyperbolic Bessel functions are to the hyperbolic sines and cosines as the ordinary Bessel functions are to the ordinary sines and cosines.

If a plate is clamped at its edge $r = a$, then $Z = 0$ and $\partial Z/\partial r = 0$. The first of these conditions is satisfied if $AJ_m(ka) + BI_m(ka) = 0$, and the second if $AJ'_m(ka) + BI'_m(ka) = 0$.

A plate with a free edge is more difficult to handle mathematically. The boundary conditions used by Kirchhoff lead to a rather complicated expression for k_{mn}, which reduces to $(2n + m)\pi/2r$ for large ka. The modal frequencies f_{mn} are given in Table 1 for modes with m nodal diameters and n nodal circles.

3.2 Drums

Some drums, such as kettledrums, tabla, and boobams, convey a strong sense of pitch, and others do not; in the latter group are the bass drum, snare drum, tenor drum, tom-toms, bongos, congas, and many drums mainly of African and Oriental origin. As vibrating systems, drums can be divided into three categories: those consisting of a single membrane coupled to an enclosed air cavity (such as kettledrums), those consisting of a single membrane open to the air on both sides (tom-toms, congas) and those consisting of two membranes coupled by an enclosed air cavity (bass drums, snare drums). One can also categorize drums by their ethnic origin—for instance, Oriental, African, or Latin American—or according to the types of musical performance with which they are associated, such as symphonic, military, jazz, ethnic, dance, and marching bands.

Table 1. Vibration frequencies of a thin circular plate with free edge. Lowest mode frequency $f_{20} = 0.2413\ c_L h/a^2$.

m	n	f/f_{20}
2	0	1.00
0	1	1.73
3	0	2.33
1	1	3.91
4	0	4.11
5	0	6.30
2	1	6.71
0	2	7.34
3	1	10.07
1	2	11.40
4	1	13.92
2	2	15.97
5	1	18.24
3	2	21.19
4	2	27.18
5	2	33.31

3.2.1 Kettledrums

Kettledrums, or timpani, are usually considered to be the most important drums in modern orchestras. Most modern timpani have a pedal-operated tensioning mechanism in addition to six or eight tensioning screws around the rim of the kettle. The pedal typically allows the player to vary the tension over a range of at least 3:1, which corresponds to a tuning range in excess of a musical sixth. At one time, most timpani heads were made of calfskin, but this material has gradually been replaced by Mylar (polyethylene terephthalate). Because of its homogeneity, Mylar is insensitive to humidity and is easier to tune than are heads made of natural materials. A thickness of 0.19 mm is considered standard for Mylar timpani heads. Timpani kettles are roughly hemispherical; copper is the preferred material, but fiberglass and other materials are also used.

Although the modes of vibration of an ideal membrane are not harmonic, a carefully tuned kettledrum is known to sound a strong principal note with two or more harmonic overtones. Lord Rayleigh recognized the principal note (at frequency f_1) as coming from the (1,1) mode and identified overtones about a perfect fifth ($f/f_1 = 1.50$), a major seventh (1.88), and an octave (2.00) above the principal tone (Rayleigh, 1894). He identified these overtones as originating from the (2,1), (3,1), and (1,2) modes, respectively, which in an ideal membrane should have frequencies of 1.34, 1.6, and 1.83 times the frequency of the (1,1) mode. Rayleigh's results are quite remarkable, considering the equipment available to him.

More recent measurements have indicated that the (1,1), (2,1), and (3,1) modes in a timpani have frequencies nearly in the ratios 1:1.5:2. The frequencies of the (4,1) and (5,1) modes are typically 2.44 and 2.90 times that of the fundamental (1,1) mode, within about half a semitone of the ratios 2.5 and 3, respectively. Thus the family of modes having 1–5 nodal diameters radiates prominent partial tones having frequency ratios of nearly 2:3:4:5:6, giving the timpani a strong sense of pitch.

How are the inharmonic modes of an ideal circular membrane shifted in frequency so that a series of prominent harmonic partials appears in the sound of a carefully tuned kettledrum? Four effects seem to contribute:

1. The membrane vibrates in a sea of air, and the mass of this air sloshing back and forth lowers the frequencies of the principal modes of vibration.
2. The air enclosed by the kettle has resonances of its own that interact with those modes of the membrane that have the right symmetry.
3. The bending stiffness of the membrane, like the stiffness of piano strings, raises the frequencies of the higher overtones.
4. Drumheads have a rather large stiffness to shear, and so they resist the type of distortion needed to deflect a membrane (as if it were being wrapped around a bowling ball) without wrinkling it.

Studies have shown that air loading, which lowers the low-frequency modes, is mainly responsible for establishing the harmonic relationship of kettledrum modes (Rossing, 1982). Other effects act merely to fine-tune the frequencies but may have a considerable effect on the rate of decay of the sound. The stiffness (or pressure) of the air enclosed in the kettle raises the frequencies of the axially symmetric modes, especially the (0,1) mode.

3.2.2 Snare Drums Snare drums are used in many types of music ensembles, including jazz bands, marching bands, and symphony orchestras. The orchestral snare drum is a two-headed instrument, about 35 cm in diameter and 13 to 20 cm deep. Strands of wire or gut stretch across the lower, or snare, head. When the upper, or batter, head is struck, the snare head vibrates against the snares. Alternatively, the snares can be moved away from the head to give a totally different sound.

In a two-headed drum, there is appreciable coupling between the two heads, especially at low frequency. This coupling may take place acoustically through the enclosed air or mechanically through the shell and leads to the formation of mode pairs. In the first two modes of vibration of the drum, the batter and snare heads move in the manner of the (0,1) membrane mode of an ideal membrane, as shown in Fig. 29. In the lower-frequency member of the pair, both heads move in the same direction, and in the higher-frequency mode, they move in opposite directions.

A simple two-mass model describes these first two modes reasonably well. The batter head is represented by a mass m_b and a spring

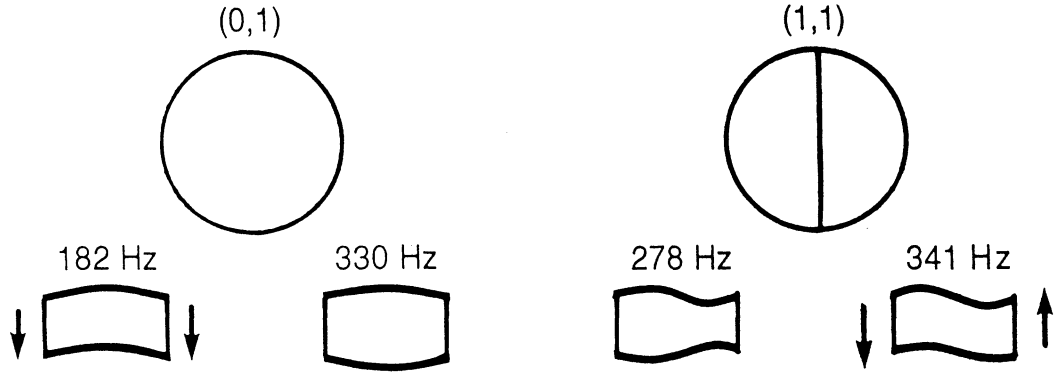

FIG. 29. Lowest four vibrational modes of a snare drum. The modal designations (*m*,*n*) refer to the membrane modes shown in Fig. 28. Mode frequencies shown are for a typical snare drum.

with stiffness K_b, the snare head by a mass m_s and a spring constant K_s; the enclosed air constitutes a third spring with constant K_c connecting the masses. This system is known to have two modes of vibration (Rossing, 1992), whose frequencies are given by

$$\omega^2 = \tfrac{1}{2}(\omega_b^2 + \omega_s^2 + \omega_{cb}^2 + \omega_{cs}^2) \pm \tfrac{1}{2}\sqrt{[(\omega_b^2 + \omega_{cb}^2) - (\omega_s^2 + \omega_{cs}^2)]^2 + 4\omega_{cb}^2\omega_{cs}^2} \tag{20}$$

where

$$\omega_b^2 = K_b/m_b, \quad \omega_s^2 = K_s/m_s, \quad \omega_{cb}^2 = K_c/m_b, \quad \omega_{cs}^2 = K_c/m_s. \tag{21}$$

3.2.3 Indian Drums Foremost among the drums of India are the tabla of northern India and the mridanga of southern India. The overtones of both of these drums are tuned harmonically by loading the drumhead with a paste of starch, gum, iron oxide, charcoal, or other materials.

The tabla has a rather thick head made from three layers of animal skin. (Calf, sheep, goat, and buffalo skins are used in different regions.) The shell is traditionally made of wood, but currently may be made of metal as well. The innermost and outermost layers of the head are annular, or ring shaped, and the layers are braided together at their outer edges and fastened to a leather hoop. Small straws or strings are placed around the edge between the outer and middle head. Tension is applied to the head by means of a long leather thong that weaves back and forth (normally 16 times) between the top and bottom of the drum. The tension in the thong can be changed by moving small wooden cylinders up or down. To tune the overtones of the tabla harmonically, one applies many thin layers of black paste to the center of the head, building up a circular patch. The paste consists of boiled rice and water with heavy particles such as iron oxide or manganese dust added to increase the density. After it is applied, each layer is allowed to dry and is then rubbed with a smooth stone until tiny cracks appear in the surface. The patch ends up with a slightly convex surface. Fine-tuning of the head is accomplished by upward or downward taps on the hoop with a small hammer.

The tabla we have described is usually played along with a larger drum, known by various names: banya, bayan, bhaya, dugga, or left-handed tabla. The head of this larger drum is also loaded (slightly off center), and the shell may be of clay, wood, or metal. To play the tabla, the drummer rests the edge of the palm on the widest unloaded portion of the membrane. This constraint causes the nodal patterns to be quite symmetrical. Releasing the palm pressure produces a sound of different quality.

The mridanga or mridangam is an ancient, two-headed drum that functions, in many respects, as a tabla and banya combined into one. The smaller head, like that of the tabla, is loaded with a patch of dried paste, while the larger head is normally loaded with a paste of wheat and water shortly before playing.

A succession of Indian scientists, beginning with C. V. Raman, have studied the acoustical properties of these drums. Raman and his colleagues recognized that the first four overtones of the tabla are harmonics of the fundamental mode. Later they identified these five harmonics as coming from nine normal modes of vibration, several of which have the same frequencies. The fundamental is from the (0,1) mode; the second harmonic is from the (1,1) mode; the (2,1) and (0,2) modes provide the third harmonic; the (3,1) and (1,2) modes similarly supply the fourth harmonic; and three modes, the (4,1), (0,3), and (2,2), contribute to the fifth harmonic.

Figure 30 shows the nodal patterns of the nine normal modes corresponding to the five harmonics. Also shown are some of the combination modes whose vibrational frequencies correspond to the five tuned harmonics.

3.3 Mallet Instruments

3.3.1 Marimba Marimbas and xylophones both make use of tuned bars of rosewood or fiberglass synthetic material. The *marimba* typically includes 3 to $4\frac{1}{2}$ octaves of tuned bars, graduated in width from about 4.5 to 6.4 cm ($1\frac{3}{4}$ to $2\frac{1}{2}$ in.). Beneath each bar is a tubular resonator tuned to the fundamental frequency of that bar. When the marimba is played with soft mallets, it produces a rich mellow tone. The playing range of a large concert marimba is A_2 to C_7 (f = 110 to 2093 Hz), although bass marimbas extend to C_2 (f = 65 Hz).

A deep arch is cut in the underside of marimba bars, particularly in the low register.

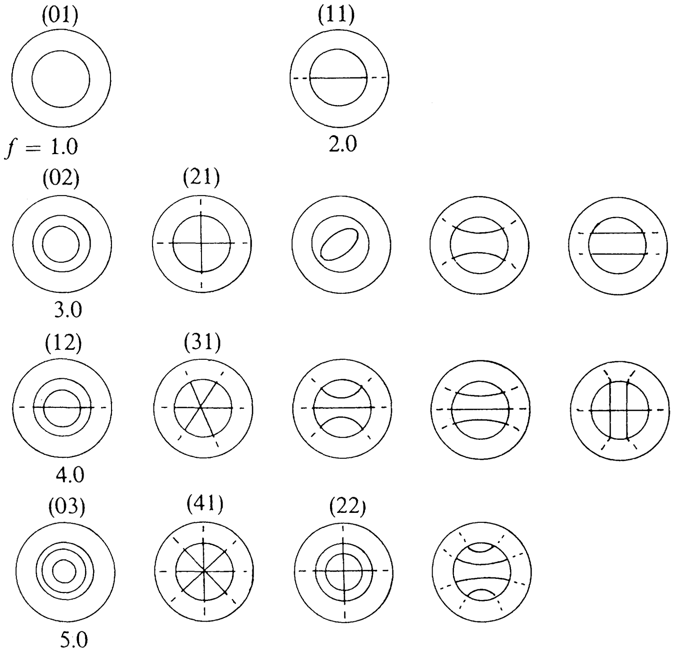

FIG. 30. Nodal patterns of the nine normal modes plus seven of the combination modes that correspond to the five harmonics of a tabla or mridanga head (from Fletcher and Rossing, 1991).

This arch serves two useful purposes: It reduces the length of bar required to reach the low pitches, and it allows tuning of the overtones (the first overtone is normally tuned two octaves above the fundamental). Figure 31 shows a scale drawing of a marimba bar, and it also indicates the positions of the nodes for each of the first seven modes of vibration.

Marimba resonators are cylindrical pipes tuned to the fundamental mode of the corresponding bars. A pipe with one closed end and one open end resonates when its acoustical length is one-fourth of a wavelength of the sound. The purpose of the tubular resonators is to emphasize the fundamental and also to increase the loudness, which is done at the expense of shortening the decay time of the sound.

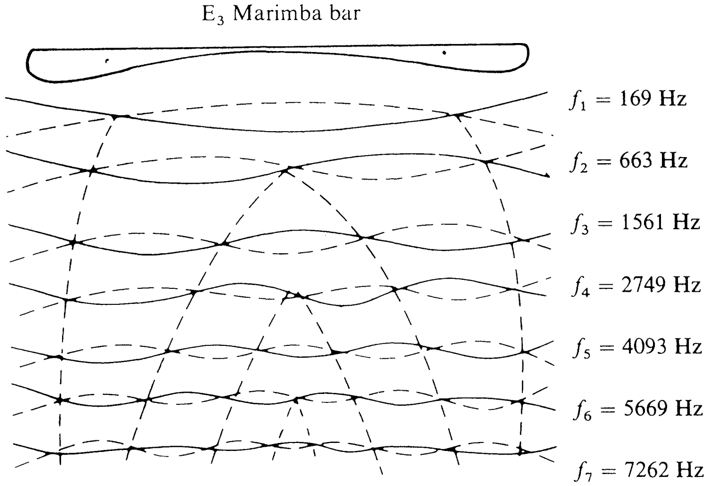

FIG. 31. Scale drawing of a marimba bar tuned to E_3 (f = 265 Hz). The dashed lines locate the nodes of the first seven modes.

3.3.2 Xylophone The term *xylophone* can refer to any percussion instrument made of wood, and most musical cultures feature such instruments. American xylophones typically cover a range of 3 to $3\frac{1}{2}$ octaves extending from F_4 or C_5 to C_8 (f = 349–4186 Hz) and, like the marimba, may have bars of synthetic material or rosewood. In some European countries, however, the term xylophone refers to marimbas as well as American xylophones. Modern xylophones are nearly always equipped with tubular resonators to increase the loudness of the tone.

Xylophone bars are also cut with an arch on the underside, but the arch is not as deep as that of the marimba, since the first overtone is tuned to a musical 12th above the fundamental (that is, three times the frequency of the fundamental). Since a pipe closed at one end can also resonate at three times its fundamental resonant frequency, a xylophone resonator reinforces the 12th as well as the fundamental. This overtone boost, plus the hard mallets used to play it, give the xylophone a much crisper, brighter sound than the marimba. We have found that careful overtone tuning is usually ignored in the upper register, as in the case of the marimba. (Rossing, 1976).

3.3.3 Vibes A very popular mallet percussion instrument is the *vibraphone* or *vibraharp*, as they are designated by different manufacturers. *Vibes*, as they are popularly called, usually consist of aluminum bars tuned over a three-octave range from F_3 to F_6 (f = 175–1397 Hz). The bars are deeply arched so that the first overtone has four times the frequency of the fundamental, as in the marimba. The aluminum bars tend to have a much longer decay time than the wood or synthetic bars of the marimba or xylophone, and so vibes are equipped with pedal-operated dampers.

The most distinctive feature of vibes, however, is the vibrato introduced by motor-driven discs at the top of the resonators, which alternately open and close the tubes. The vibrato produced by these rotating discs or pulsators consists of rather substantial fluctuation in amplitude ("intensity vibrato") and a barely detectable change in frequency ("pitch vibrato"). The speed of rotation of the discs may be adjusted to produce a slow vibe or a fast vibe. Often vibes are played without vibrato by switching off the motor. They are usually played with soft mallets or beaters, which produce a mellow tone, although some passages call for harder beaters.

Because vibraphone bars have a much longer decay time than do marimba and xylophone bars, the effect of the tubular resonators on decay time is more dramatic. At 200 Hz (A_3), for example, we have measured a decay time (60 dB) of 40 s without the resonator and 9 s with the tube full open. For A_5, we measured 24 s with the resonator closed and 8 s with it open.

3.3.4 Chimes Chimes, or tubular bells, are usually fabricated from lengths of brass tubing $1\frac{1}{4}$ to $1\frac{1}{2}$ in. in diameter. The upper end of each tube is partially or completely closed by a brass plug with a protruding rim. This rim forms a convenient and durable strike point.

An interesting acoustical property of chimes is that there is no mode of vibration with a frequency at, or even near, the pitch of the strike tone one hears. The frequencies excited when a chime is struck are very nearly those of a free bar described earlier. Modes 4, 5, and 6 appear to determine the strike tone. This can be understood by noting that these modes for a free bar have frequencies nearly in the ratio $9^2:11^2:13^2$, or 81:121:169, which are close enough to the ratio 2:3:4 for the ear to consider them nearly harmonic and to use them as a basis for establishing a pitch. The largest near-common factor in the numbers 81, 121, and 169 is 41, and so the subjective strike tone lies an octave below the frequency of mode 4.

3.4 Cymbals, Gongs, and Steel Pans

3.4.1 Cymbals Cymbals are among the oldest of musical instruments and have had both religious and military use in a number of cultures. The Turkish cymbals generally used in orchestras and bands are saucer shaped with a small dome in the center, in contrast to Chinese cymbals, which have a turned-up edge more like a tamtam. Orchestral cymbals are usually between 40 and 55 cm in diameter and are made of bronze. A leading manufacturer of cymbals, the Avedis Zildjian Company, claims that its secret process for treating cymbal alloys was discovered in 1623.

Many different types of cymbals are used in orchestras, marching bands, concert bands, and jazz bands. Orchestral cymbals are often designated as "French," "Viennese," and "Germanic" in order of increasing thickness. Jazz drummers use cymbals designated by such onomatopoetic terms as "crash," "ride," "swish," "splash," "ping," and "pang." Cymbals range from 20 to 75 cm (8 to 30 in.) in diameter.

Given the wide ranges in diameter and thickness, considerable variety of cymbal tone is available to the percussionist. Furthermore, a good cymbal can be made to produce many different tones by using a variety of sticks and striking it at several different places. A large cymbal struck gently near the rim produces a low sound, not unlike that of a small tamtam. The fullest sound is obtained by a glancing blow about one-third of the way in from the rim.

A cymbal may be excited in many different ways. It may be struck at various points with a wooden stick, a soft beater, or another cymbal. The onset of sound is quite dependent on the manner of excitation. The coupling between vibrational modes in a cymbal is strong, however, so that a large number of partials quickly appear in the spectrum, however it is excited.

By comparing sound spectra recorded immediately after striking and after varying time intervals later, we make the following observations about cymbal sound:

1. The sound level below about 700 Hz shows a rather rapid decrease during the first 200 ms, after which it decays slowly. This is apparently a result of conversion of energy into modes of higher frequency.
2. Several strong peaks in the 700–1000-Hz range build up between 10 and 20 ms, then decay.
3. Sound energy in the important 3- to 5-kHz range peaks about 50–100 ms after striking.
4. Sound in the range of 3–5 kHz, which gives the cymbal its "shimmer," is often the most prominent feature from about 1 to 4 s after striking.
5. The low frequencies again dominate the lingering sound, but at a much lower level, so that they are rather inconspicuous (Rossing and Shepherd, 1983).

3.4.2 Gongs and Tam-Tams Gongs play a very important role in Oriental music, but they enjoy considerable popularity in Western music as well. They are usually cast of bronze with a deep rim and a protruding dome. Gongs used in symphony orchestras usually range from 0.5 to 1 m (20 to 38 in.) in diameter, and are tuned to a definite pitch. When struck near the center with a massive soft mallet, the sound builds up relatively slowly and continues for a long time if the gong is not damped.

Massive gongs are central to every gamelan (ensemble) in Indonesia. Of considerable interest to the acoustician are the gongs used in Chinese opera orchestras, which glide downward or upward in pitch after being struck, because of their special shapes that lead to nonlinearities characterized by a hardening or softening spring constant (Rossing and Fletcher, 1983).

Tamtams are similar to gongs in appearance and are often confused with them. The main differences between the two are that tamtams do not have the dome of the gong, their rim is not as deep, and their metal is thinner. Tamtams sound a much less definite pitch than do gongs. In fact, the sound of a tamtam may be described as somewhere between the sounds of a gong and a cymbal. The sound of a large tamtam develops slowly, changing from a sound of low pitch at strike to a collection of high-frequency vibrations, which are described as "shimmer." These high-frequency modes fail to develop if the tamtam is not hit hard enough, indicating that the conversion of energy takes place through a nonlinear process. Figure 32 shows the shapes of a number of different gongs and tamtams.

3.4.3 Steel Pans An instrument of rather recent origin that has achieved great popularity is the Caribbean steel pan or steel drum. It has become the foremost musical instrument in its home country (Trinidad) and in other Caribbean countries; steel bands are becoming increasingly common in Europe and the United States as well. Steel pans are generally played in ensembles ranging in size from six or eight players to 100 or more. The instrument is occasionally used as a solo instrument accompanied by an orchestral ensemble.

Steel drums or pans are usually fabricated from 55-gal oil drums. The drums in a steel

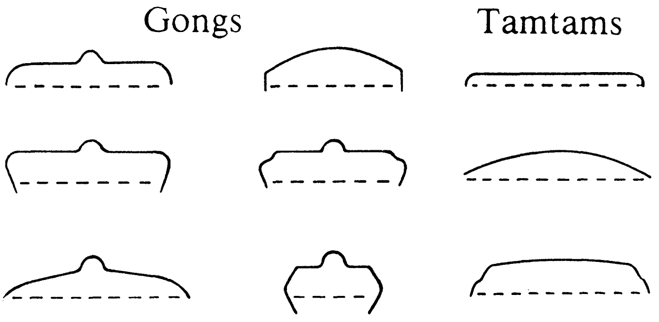

FIG. 32. Typical shapes of gongs and tamtams. Note the deeper rims and tuning domes of the gongs.

band are known by various names such as single tenor, soprano, double tenor, guitar, cello, and bass. The single tenor or lead pan has from 26 to 32 different notes, but each bass drum has only 3 or 4; hence, the bass drummer typically plays on six drums in the manner of a timpanist.

The first step in making a steel drum is to hammer the end of the oil barrel to the shape of a shallow basin. Then a pattern of grooves is cut with a nail punch in order to define sections of the various notes. Next, each section is "ponged up" with a hammer, and after heating and hardening the drum, each section is tuned by the skillful use of the hammer.

Different pan makers use different patterns for their pans, and these patterns change from time to time. A steel band may typically employ the following instruments:

1. single tenor (also called lead, soprano, or ping-pong): one pan of $2\frac{1}{2}$ octaves (28 notes) ranging from D_4 to F_6;
2. double tenor: two pans of 15 notes each from $A_3\flat$ to $C_6\sharp$;
3. double second: two pans of 14 notes each from $F_3\sharp$ to A_5;
4. cello: three drums of seven notes each from B_2 to G_4; and
5. bass: six drums of three notes each from C_2 to F_3.

Other designs may include a guitar consisting of two drums of 10 notes each, a tenor bass consisting of four drums with five notes each, and a bass set of six drums with four notes each.

A skilled steel drum maker tunes at least one overtone of each note to a harmonic of the fundamental (usually the octave). Sometimes it is possible to tune another mode to the third or fourth harmonic by adjusting internal "tension" along a boundary.

The sound spectra of steel drums are surprisingly rich in harmonic overtones. These harmonic overtones appear to have three different physical origins:

1. radiation from higher modes of vibration tuned harmonically by the tuner;
2. radiation from nearby notes whose frequencies are harmonically related to the struck note; and
3. nonlinear motion of the note area vibrating at its fundamental frequencies.

3.5 Bells

Bells have been a part of nearly every culture in history. Bells existed in the Near East before 1000 B.C., and a number of Chinese bells from the time of the Shang dynasty (1600–1100 B.C.) are found in museums throughout the world. In 1978, a set of 65 tuned bells from the 5th century B.C. was discovered in the Chinese province of Hubei.

3.5.1 Carillon Bells Bells developed as Western musical instruments in the 17th century when bell founders discovered how to tune their partials harmonically. The founders in the Low Countries, especially the Hemony brothers (François and Pieter) and Jacob van Eyck, took the lead in tuning bells, and many of their fine bells are found in carillons today.

The carillon also developed in the Low Countries. Chiming bells by pulling ropes attached to the clappers had been practiced for some time before the idea of attaching these ropes to a keyboard or handclavier occurred to bell ringers in the 16th century. Many mechanical improvements during the 17th and 18th centuries, including the so-called breached wire system and the addition of foot pedals for playing the larger bells, led to development of the modern carillon. Today, the term carillon is reserved for an instrument of 23 (two octaves) or more tuned bells played from a clavier (smaller sets are called "chimes"). The largest carillon in existence is the 74-bell (6 octave) carillon at Riverside Church in New York whose largest bell is more than 18 000 kg (20 tons).

3.5.2 Vibrations of Carillon Bells and Church Bells Perhaps it is stretching the imagination a bit to think of a bell as being a plate, but the general principles of its vibrational behavior are similar. Although the mathematical description of the vibrations of a bell are understandably complex, the principal modes, at least, can be described by specifying the number of circular nodes and meridian nodes. The lowest mode of vibration (called the hum tone), for example, has four meridian nodes, so that alternate quarters of the bell essentially move inward and outward.

Figure 33 shows the principal vibrational modes for a carillon bell. The mode called the third is tuned a minor third above the strike tone, whereas the upper third is usually a major third above the octave. The strike tone is determined by the octave, the 12th, and the upper octave, whose frequencies have the ratios 2:3:4, just as in chimes. Unlike chimes, however, carillon bells have a mode called the prime or fundamental with a frequency at or near the strike tone. Careful studies have shown, however, that the pitch of the strike tone is determined by the three modes mentioned previously rather than by the prime.

Vibrational frequencies of groups O-X in a church bell with a D_5 strike note are shown in Fig. 34. Each group includes modes having the same number of nodal circles positioned at about the same height on the bell. Horizontal lines on the axis at the right indicate relative strengths of several partials in the bell sound (these vary considerably with location, of course). Arrows denote the three partials [octave or nominal, 5th (12th), and upper octave] that determine the strike note.

A new type of carillon bell has been developed at the Royal Eijsbouts Bellfoundry in The Netherlands. The new bell replaces the strong minor-third partial with a major-third partial, thus changing the tonal character of the bell sound from minor to major. This requires an entirely new bell profile. The new bell design evolved partly from the use of a technique for structural optimization using finite-element methods on a digital computer. This technique allows a designer to make changes in the profile of an existing structure and then to compute the resulting changes in the vibrational modes (Lehr, 1987).

3.5.3 Handbells Handbells also date back to at least several centuries B.C., although tuned handbells of the present-day type were developed in England in the 18th century. One early use of handbells was to provide tower bellringers with a convenient means to practice change ringing. In more recent years, handbell choirs have become popular in schools and churches—some 20 000 choirs are reported in the United States.

Although they are cast from the same bronze material and cover roughly the same range of pitch, the sounds of church bells, carillon bells, and handbells have distinctly different timbres. In a handbell, only two modes of vibration are tuned (although there are three harmonic partials in the sound), whereas in a church bell or carillon bell, at least five modes are tuned harmonically. A church bell or carillon bell is struck by a heavy metal clapper in order to radiate a sound that can be heard at a great distance, where the gentle sound of a handbell requires a relatively soft clapper.

In the so-called English tuning of handbells, followed by most handbell makers in England and the United States, the (3,0) mode is tuned to three times the frequency of the (2,0) mode. The fundamental (2,0) mode radiates a rather strong second-harmonic partial, however, so that the sound spectrum has prominent partials at the first three harmonics. Some Dutch founders aim at tuning the (3,0) mode in handbells to 2.4 times the frequency of the fundamental, giving their handbell sound a minor-third character somewhat like a church bell. Such bells are usually thicker and heavier than bells with the English-type tuning.

3.5.4 Ancient Chinese Two-Tone Bells Much interest has developed in the acoustical properties of ancient Chinese two-tone bells whose almond-shaped cross section consists of two segments of a circle joined at

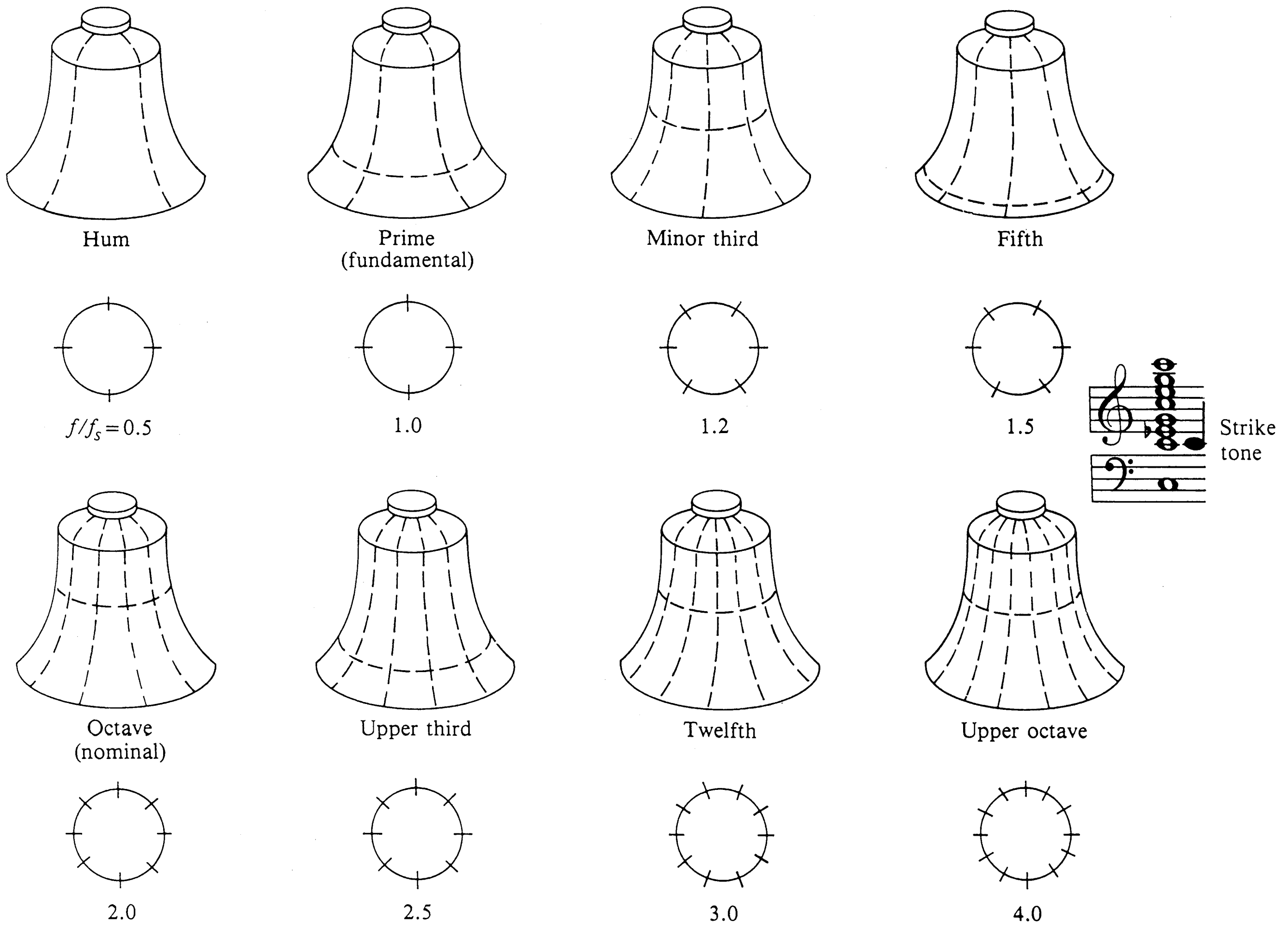
Hum
$f/f_s = 0.5$
Prime (fundamental)
1.0
Minor third
1.2
Fifth
1.5
Strike tone
Octave (nominal)
2.0
Upper third
2.5
Twelfth
3.0
Upper octave
4.0

FIG. 33. The first eight vibrational modes of a carillon bell or tuned church bell. Dotted lines indicate the approximate locations of nodes. Frequencies relative to the strike tone are given. The notes that correspond to these in a C_4 bell are shown on a musical staff. In some bells, only five modes are tuned harmonically. (From Rossing, 1990.)

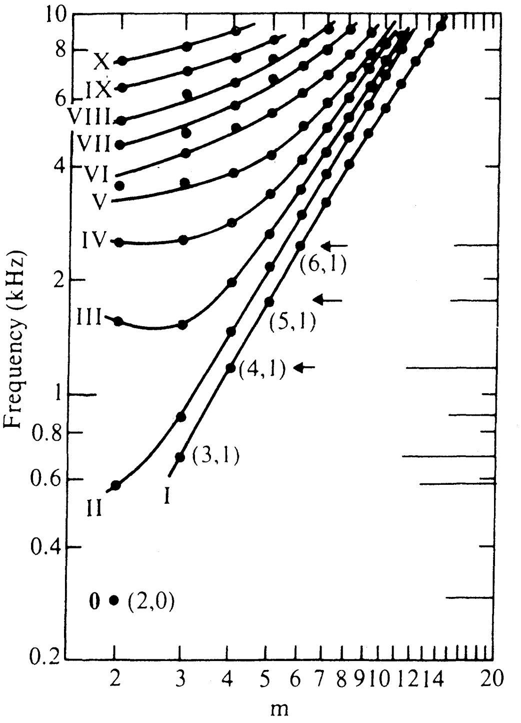

FIG. 34. Vibrational frequencies of groups O-IX in a D_5 church bell. Also shown on the right are the relative strengths at impact of several partials in the bell sound. Arrows denote the three partials in group I that determine the strike note. (From Rossing and Perrin, 1986).

the *xian* or spine. Because they have almond-shaped, rather than round, cross sections, the vibrational modes are split into frequency doublets, and thus, they sound tones with two different pitches when struck at two different points (sometimes referred to as the *sui-yin* and *gu-yin* strike points). This interest was stimulated by the discovery in 1977 of a rack of 65 bells in the tomb of the Marquis Yi of Zheng. These bells, carefully tuned and richly inscribed, dated from the 5th century B.C.

Holographic studies of two-tone bells have shown that the vibrational modes tend to occur in pairs, one with a node at the xian and one with an antinode at that location. The mode with a node at the spine generally has the higher frequency.

Some modes of vibration in a modern Chinese two-tone bell are illustrated in the holographic interferograms in Fig. 35. The interval between the two fundamental pitches, determined by the frequencies of the $(2,0)_a$ and $(2,0)_b$ modes, is a minor third ($f_{20b}/f_{20a} = 1.20$). This is the closest interval on the Western musical scale in about half the ancient bells that have been studied, the others varying from a minor second to a minor sixth. Modal frequencies are shown as a function of the number of complete nodal meridians *m* in Fig. 36.

4. RESEARCH OPPORTUNITIES

Support for research on musical instruments has always been difficult to obtain, and so many physicists who do research in this area earn their livelihood through teaching or research in other areas. Nevertheless, there is an active community of musical acousticians around the world, and international meetings offer opportunities to share research results. The Acoustical Society of America has a Technical Committee on Musical Acoustics that arranges technical sessions at most meetings of that society. The Catgut Acoustical Society stresses research on bowed string instruments, but its meetings include sessions dealing with other musical instruments as well.

Digital computers and digital instrumentation have led to great progress in musical acoustics research in recent years. Digital computers have themselves become musical instruments of great versatility (see MUSIC, ELECTRONIC). Not only can digital computers generate musical sounds, but they can be employed by composers to "write" music as well.

The analysis of musical instruments and the way in which they produce musical sound has been greatly aided by two techniques that rely on digital computers. Experimental modal testing or modal analysis makes use of a computer to determine the most plausible normal modes of vibration in a musical instrument based on its response to impact or steady-state excitation. Finite-element methods, using a digital computer, theoretically determine the modes of vibration and response function. Used together, these two methods constitute a very powerful technique for understanding sound production in existing musical instruments and the design of new instruments. One example of the latter is the development of carillon bells in which the dominant minor-third partial is replaced by a major third, thus producing an

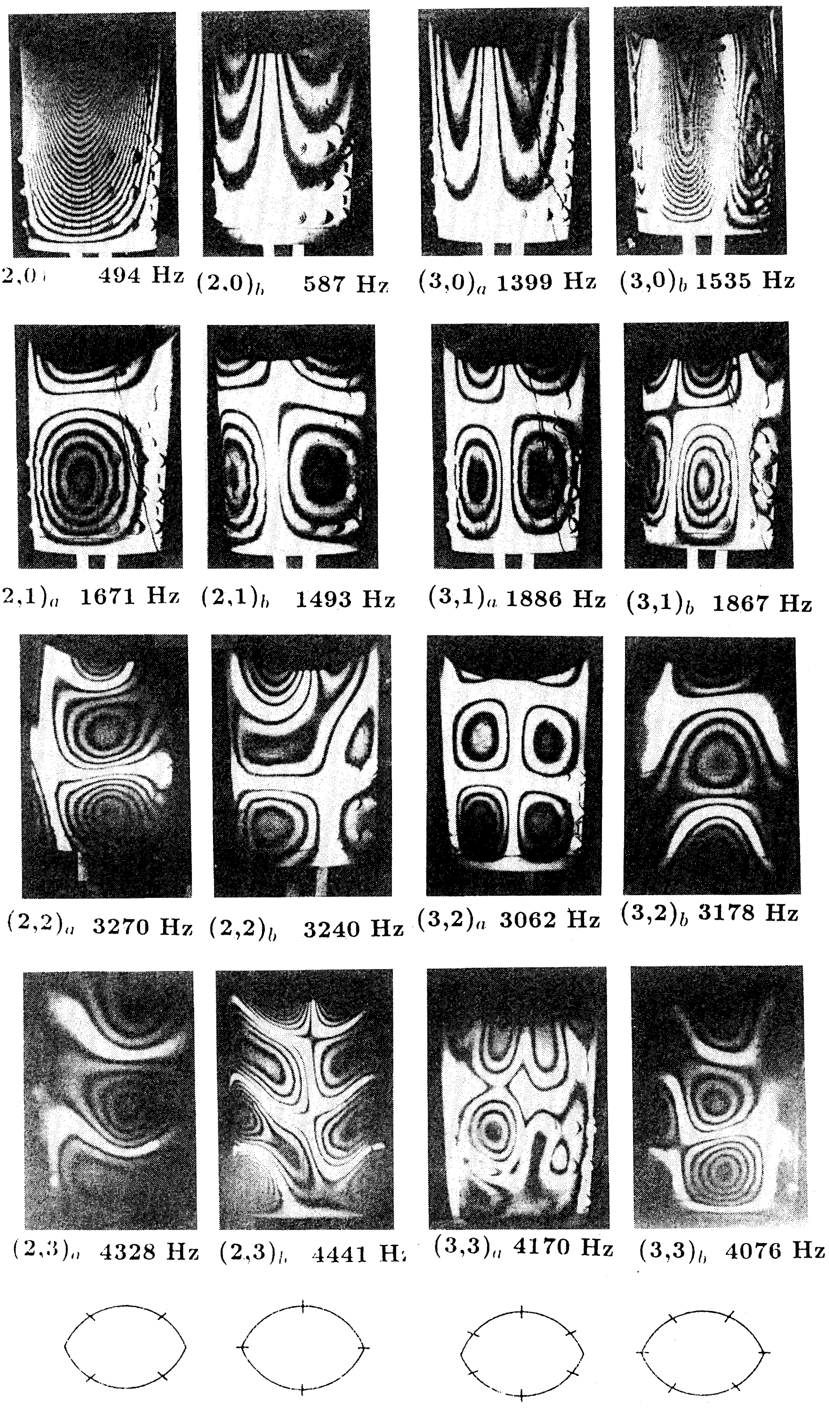

$(2,0)_a$ 494 Hz $(2,0)_b$ 587 Hz $(3,0)_a$ 1399 Hz $(3,0)_b$ 1535 Hz

$(2,1)_a$ 1671 Hz $(2,1)_b$ 1493 Hz $(3,1)_a$ 1886 Hz $(3,1)_b$ 1867 Hz

$(2,2)_a$ 3270 Hz $(2,2)_b$ 3240 Hz $(3,2)_a$ 3062 Hz $(3,2)_b$ 3178 Hz

$(2,3)_a$ 4328 Hz $(2,3)_b$ 4441 Hz $(3,3)_a$ 4170 Hz $(3,3)_b$ 4076 Hz

FIG. 35. Some modes of vibration in a modern Chinese two-tone bell (from Tsai, *et al.*, 1992).

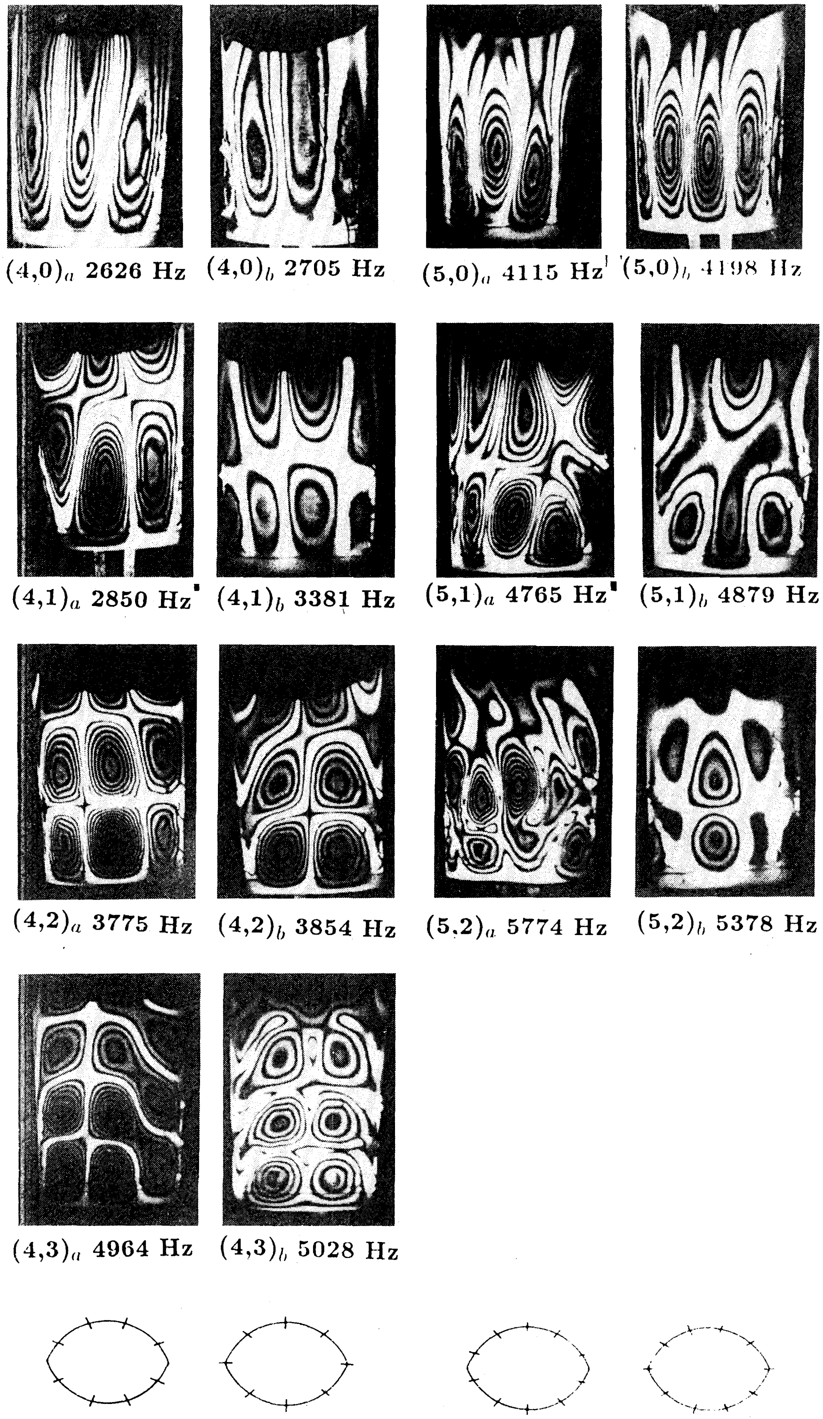
$(4,0)_a$ 2626 Hz
$(4,0)_b$ 2705 Hz
$(5,0)_a$ 4115 Hz
$(5,0)_b$ 4198 Hz
$(4,1)_a$ 2850 Hz
$(4,1)_b$ 3381 Hz
$(5,1)_a$ 4765 Hz
$(5,1)_b$ 4879 Hz
$(4,2)_a$ 3775 Hz
$(4,2)_b$ 3854 Hz
$(5,2)_a$ 5774 Hz
$(5,2)_b$ 5378 Hz
$(4,3)_a$ 4964 Hz
$(4,3)_b$ 5028 Hz

FIG. 35. Continued.

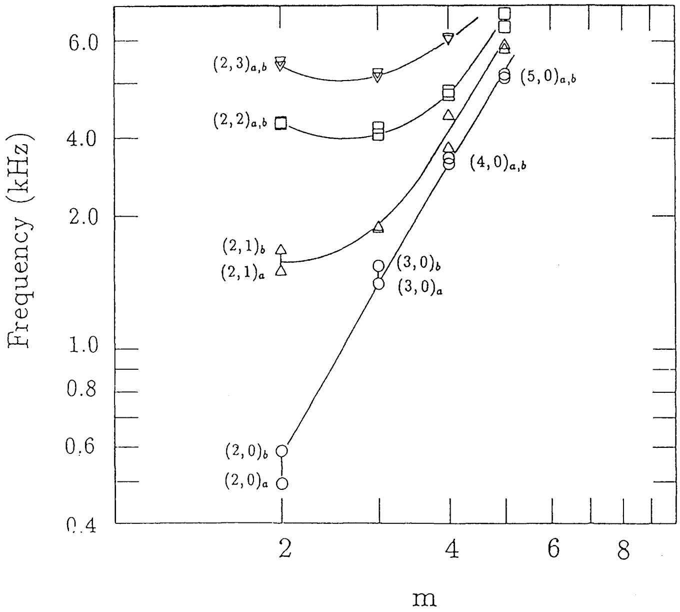

FIG. 36. Mode frequencies of a modern Chinese two-tone bell as functions of the number of nodal meridians *m* and nodal circles *n* (from Tsai, *et al.*, 1992).

entirely new musical character; this development was made possible by the use of finite-element methods on a digital computer.

Many musical instruments require a certain type of wood, and the growing scarcity of some wood is cause for alarm. The same is true of animal materials, such as gut for strings, hides for drumheads, and wool for felt hammers. This situation has led to extensive research on synthetic materials, but much more is needed.

Another concern of musicians is that high-quality handmade musical instruments are becoming unaffordable by many performers. More research is needed in order to raise the quality of affordable mass-produced instruments. Also, a better understanding of musical instruments will lead to the development of instructional aids that supplement or replace costly private instruction in learning to play musical instruments.

GLOSSARY

Admittance: Reciprocal of impedance. Mechanical admittance (mobility) is velocity divided by force. Acoustical admittance is flow velocity divided by pressure.

Bessel Horn: A family of horns whose radii r are given by $r = bx^{-\epsilon}$, which form the basis for the air columns in many wind instruments.

Carillon: A musical instrument consisting of 23 or more tuned bells whose clappers are mechanically coupled to a clavier or keyboard.

Cent: 1/100 of a semitone in the scale of equal temperament. An octave is equal to 1200 cents.

Chladni Patterns: A pattern created by sprinkling a fine powder on a vibrating plate. The powder tends to collect along the nodal lines, thus making them visible.

Cross Fingering or Fork Fingering: A woodwind fingering that leaves one or more tone holes open and closes other tone holes below this.

Harmonics: Tones or partials whose frequencies are integral multiples of a fundamental frequency.

Impedance: Reciprocal of admittance. Mechanical impedance is force divided by velocity. Acoustical impedance is pressure divided by flow velocity.

Mobility: Mechanical admittance. Velocity divided by force.

Partial Tones or Partials: Single-frequency components of a tone.

Radiation Resistance: Real part of the radiation impedance, which is a measure of the fluid loading on a vibrating surface.

Scale of Equal Temperament: A musical scale in which each semitone increments the frequency by a factor $2^{1/12} = 1.05946 \simeq 1.06$. In this scale, each semitone interval is exactly 100 cents.

Sound Pressure Level: Logarithmic expression of sound pressure expressed in decibels (dB). The sound pressure level is $L_p = 20 \log (p/p_0)$, where p is sound pressure, and $p_0 = 2 \times 10^{-5}$ N/m^2 = 20 μPa.

Sound Spectrum: A "recipe" for a sound that expresses, in graphical form, the sound-pressure levels and the frequencies of the various partials or spectral components. In the spectrum of a musical sound, these partials are often (but not always) harmonics of a fundamental tone.

Tricord: A set of three piano strings tuned to the same frequency and struck by the same hammer.

Works Cited

Alonso Moral, J., Jansson, E. (1982a), *Acustica* **50**, 329–377.

Alonso Moral, J., Jansson, E. (1982b), Report STL-QPSP 203, Stockholm: Speech Trans. Lab., Royal Inst. Tech. (KTH), pp. 60–75.

Ayers, R. D., Eliason, L. J., Mahgerefteh, D. (1985), *Am. J. Phys.* **53**, 528–537.

Benade, A. H. (1976), *Fundamentals of Musical Acoustics*, New York: Oxford University Press. Reprint (1990), New York: Dover.

Campbell, M., Greated, C. (1987), *The Musician's Guide to Acoustics*, New York: Schirmer.

Cremer, L. (1981), *Physik der Giege*, Stuttgart: Hirtzal Verlag. English translation by J. S. Allen (1984), Cambridge, MA: MIT Press.

Firth, I. (1977), *Catgut Acoust. Soc. Newsletter* No. 27, 12.

Fletcher, N. H. (1975), *J. Acoust. Soc. Am.* **57**, 233–235.

Fletcher, N. H., Rossing, T. D. (1991), *Physics of Musical Instruments*, New York: Springer-Verlag.

Helmholtz, H. L. F. (1877), *On Sensations of Tone*, 4th ed.; translation by A. J. Ellis (1954), New York: Dover.

Hutchins, C. M. (1977), *Catgut Acoust. Soc. Newsletter* No. 28, 22–24.

Kindel, J. (1989), *Modal Analysis and Finite Element Analysis of a Piano Soundboard*, M. S. Thesis, Univ. of Cincinnati (unpublished).

Lehr, A. (1987), *The Designing of Swinging Bells in the Past and Present*, Asten, The Netherlands: Athanasius Kircher Foundation.

Richardson, B. E., Roberts, G. W. (1985), *Proc. SMAC 83*, vol. 2, Stockholm: Royal Swedish Academy of Music, p. 285–302.

Rossing, T. D. (1976), *Phys. Teach.* **14**, 546–556.

Rossing, T. D. (1982), *Sci. Am.* **247** (5), 172–178.

Rossing, T. D., Fletcher, N. J. (1983), *J. Acoust. Soc. Am.* **73**, 345–351.

Rossing, T. D., Perrin, R. (1986), *Appl. Acoust.* **20**, 41–70.

Rossing, T. D. (1990), *The Science of Sound*, Reading, MA: Addison-Wesley.

Rossing, T. D. (1992), *Phys. Today* **45** (3), 40–47.

Rossing, T. D., Shepherd, R. B. (1983), *Proceedings of the 11th International Congress on Acoustics*, vol. 6, Paris, pp. 329–333.

Strong, W. J., Plitnik, G. R., (1983), *Music, Speech, and High Fidelity*, 2nd ed., Provo, UT: Soundprint.

Tsai, J., Jiang, Z., Rossing, T. D. (1992), *Acoust. Australia* **20**, 17–19.

Weinreich, G. (1977), *J. Acoust. Soc. Am.* **62**, 1474–1484.

Further Reading

Rossing, T. D. (1987), "Resource Letter MA-2: Musical Acoustics," *Am. J. Phys.* **55**, 589–601. An annotated list of 414 references on musical acoustics.

Fletcher, N. H., Rossing, T. D. (1991), *Physics of Musical Instruments*, New York: Springer-Verlag.

Hutchins, C. M. (1975, 1976), *Musical Acoustics*, Parts 1 and 2, Stroudsburg, PA: Dowden, Hutchinson, and Ross. Reprints of 60 articles on bowed string instruments.

Elementary textbooks on musical acoustics:

Backus, J. (1977), *The Acoustical Foundations of Music*, 2nd ed., New York: Norton.

Benade, A. H. (1976), *Fundamentals of Musical Acoustics*, New York: Oxford; reprint (1990), New York: Dover.

Campbell, M., Greated, C. (1987), *The Musician's Guide to Music*, New York: Schirmer.

Hall, D. E. (1991), *Musical Acoustics*, 2nd ed., Pacific Grove, CA: Brooks/Cole.

Rossing, T. D. (1990), *The Science of Sound*, 2nd ed., Reading, MA: Addison-Wesley.

Strong, W. J., Plitnik, G. R. (1992), *Music, Speech, Audio*, Provo, UT: Soundprint.

NANOPHASE MATERIALS

RICHARD W. SIEGEL, *Materials Science Division, Argonne National Laboratory, Argonne, Illinois, U.S.A.*

Introduction 173
1. **Background and General Principles** 174
2. **Synthesis and Processing** 175
2.1 Cluster Assembly 175
2.2 Mechanical Attrition 181
3. **Structure** 183
3.1 Grains and Pores 183
3.2 Dislocations 184
3.3 Grain Boundaries 186
3.4 Stability and Strains 188
4. **Properties** 189
4.1 Chemical Properties 189
4.2 Mechanical Properties 191
4.3 Other Physical Properties 195
5. **Future Directions** 196
6. **Acknowledgments** 197
Works Cited 197
Further Reading 200

INTRODUCTION

Nanophase materials, with their grain sizes or phase dimensions in the nanometer size regime, represent a subset of the broad class of nanostructured materials artificially synthesized with microstructures modulated in zero to three dimensions on length scales less than 100 nm that it has become possible to create over the past several years. The various types of nanostructured materials share three fundamental features: constituent atomic domains (grains, layers, or phases) spatially confined to less than 100 nm, significant atom fractions associated with interfacial environments, and interactions between the atomic domains. Nanostructured materials (see Fig. 1) thus include atom clusters and filaments—which can be considered as essentially zero-dimensionally modulated materials—, one-dimensionally modulated multilayers, two-dimensionally modulated ultrafine-grained overlayers, coatings, or buried layers, and their three-dimensionally modulated analogues, nanophase materials. Intermediate dimensionalities can also exist, and nanocomposite materials, containing multiple phases, can range from the most conventional case, in which a nanoscale phase (zero-dimensionally modulated) is embedded in a phase of conventional sizes, to the case in which all the constituent phases are of nanoscale dimensions (three-dimensionally modulated and nanophase or multilayered). While the man-made synthesis of these materials is rather recent, naturally formed nanostructured materials are found in the earliest meteorites and in a wide variety of biological systems.

Increasing interest in recent years has focused on a wide variety of nanostructured materials, since it is anticipated that their properties will be different from, and often superior to, those of conventional materials that have phase or grain structures on a coarser size scale (Kear *et al.*, 1989; Hadjipanayis and Siegel, 1994), and the initial investigations have indicated this to be the case. Moreover, it is possible to engineer these properties by controlling and varying the sizes of the constituent domains and the manner in which they are assembled (Siegel, 1993). This interest has been stimulated not only by the recent efforts and successes in synthesizing a variety of fascinating atom clusters, zero-dimensionality quantum-well structures, and one-dimensionally modulated multilayered materials with nanometer-scale modulations, but also by the rapidly developing potential for synthesizing three-dimensionally modulated, bulk nanophase materials via the assembly of clusters of atoms (Andres *et al.*, 1989) and a variety of nanocomposite materials (Roy *et al.*, 1988; Burggraaf *et al.*, 1989; Niihara, 1991). The possibilities to create new nanostructured materials with unique or improved properties are thus expected to have

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

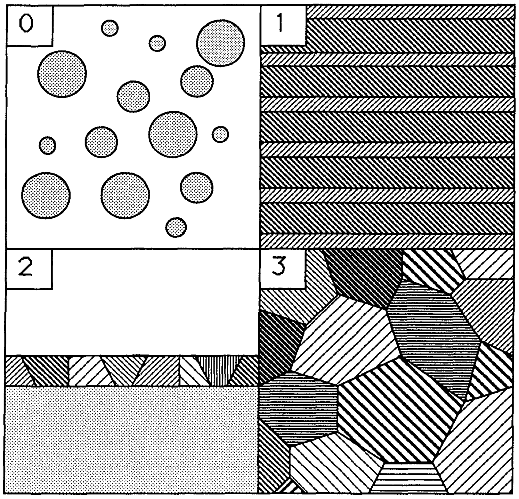

FIG. 1. Schematic of the four basic types of nanostructured materials, classified according to integral modulation dimensionality: zero, clusters of any aspect ratio from 1 to ∞; one, multilayers; two, ultrafine-grained overlayers (coatings) or buried layers; three, nanophase materials (Siegel, 1993).

a significant impact on our ability to engineer advanced materials with controlled chemical, electronic, magnetic, mechanical, and optical properties with attendant useful technological applications.

In the past few years, atom clusters with average diameters in the range of 5–50 nm of a variety of materials, including metals and ceramics, have been synthesized by evaporation and condensation in high-purity gases followed by consolidation *in situ* under ultrahigh-vacuum conditions to create nanophase materials. These new ultrafine-grained materials have properties that are often significantly different and considerably improved relative to those of their coarser-grained counterparts. The property changes result from their small grain sizes, the large percentage of their atoms in grain-boundary environments, and the interactions between their grains. Since their properties can be engineered during the synthesis and processing steps, cluster-assembled nanophase materials appear to have great technological potential beyond the current scientific interest in their grain-size–dependent properties. Recent research on nanophase materials is reviewed in this article and some future applications are considered.

1. BACKGROUND AND GENERAL PRINCIPLES

Evidence from the earliest meteorites that have been found and studied suggests that nanophase materials have been part of our universe from its beginnings after the Big Bang; it has been suggested that primordial materials with nanometer-scale phase structures condensed from our solar nebula (Blander and Katz, 1967; Blander and Abdel-Gawad, 1969; Grossman, 1972). Many biological materials, often constituents of a hierarchically structured system, are also found to be nanostructured (Aksay *et al.*, 1992) as they have evolved to optimize their performance. More recently, during the past two decades, the planned human synthesis of nanophase materials has begun. New methodologies for synthesizing ultrafine-grained nanophase materials are being developed to create new levels of property engineering through the sophisticated control of scale, morphology, and architecture.

Atom clusters in the nanometer size regime, containing hundreds to tens of thousands of atoms, can now be economically produced by either physical or chemical processes in sufficient numbers that they can be assembled into technologically useful nanophase materials. Such materials can take advantage of and incorporate a variety of size-related effects in condensed matter. These effects range from electronic quantum size effects caused by spatial confinement of delocalized valence electrons to altered cooperative (many-body) atom phenomena, such as lattice vibrations or melting, to the suppression of such lattice-defect mechanisms as dislocation generation and migration in confined grain sizes.

The desirable sizes of the constituent atomic ensembles are generally below 100 nm, since it is in this size range (and often below 10 nm) that various properties begin to change significantly as a result of a variety of confinement effects. A property will be altered when the entity or mechanism (or combination thereof) responsible for that property is confined within a space (defined by the size of the ensemble of atoms) smaller than some critical length associated with that entity or mechanism. Thus, for example, a metal that is conventionally ductile by virtue of the usual ease in creating and moving dislocations through its crystal lattice will become significantly harder when grain sizes are reduced to the point where dislocation sources are no longer able to operate at low levels of applied stress. Since the stress to operate a Frank–Read dislocation source, for example, is inversely proportional to the spacing between its pinning points, a critical length in this case is that for which the stress to operate this source becomes larger than the conventional yield stress for the given metal. Such confinement only appears to be different from that usually encountered in the technical literature, where, for example, the optical absorption properties of a quantum-well semiconductor device are blue shifted (to shorter wavelengths) because of the size of the well becoming comparable to and smaller than the effective size of the excitonic state responsible for this absorption. The specifics are quite different, but the underlying general principle of confinement is not.

Control of the size or sizes and morphologies of the phase domains or granular entities (e.g., clusters) being assembled is therefore of primary importance in any of the methods for the synthesis of nanostructured materials. Beyond this, chemical composition of the phases and cleanliness of the interfaces between phases or grains must be controlled as well. In bulk cluster-assembled nanophase materials, such control appears to be readily available. Before proceeding with a description of their synthesis and processing, it is useful to list some of the unique advantages of the assembly of nanophase materials under controlled atmospheres from gas-condensed clusters; they are as follows:

1. The <100-nm sizes and surface cleanliness of the atom clusters allow conventional restrictions of phase equilibria and kinetics to be overcome during material synthesis and processing; this results from the combination of short diffusion distances, high driving forces, and uncontaminated surfaces and interfaces available with cluster building blocks.
2. The large fraction of atoms residing in the grain boundaries and interfaces of these materials allows for the special interatomic arrangements in interfaces to constitute significant volume fractions of material; thus, novel material properties may result from these less than perfectly ordered atomic environments.
3. The reduced size scale and large surface-to-volume ratios of the individual nanophase grains can be predetermined and can significantly alter a variety of physical and chemical properties; the degree of porosity of the cluster assembly can also be controlled.
4. A wide range of materials can be produced in this manner, including metals and alloys, intermetallic compounds, ceramics, and semiconductors; and such materials can be synthesized to contain crystalline, quasi-crystalline, or amorphous structures.
5. The extensive possibilities for reacting, coating, and mixing *in situ* various types, sizes, and morphologies of clusters create significant future potential for the synthesis of a variety of new multicomponent composites with nanometer-sized microstructures and engineered properties that can be both multifunctional and hierarchical.

2. SYNTHESIS AND PROCESSING

2.1 Cluster Assembly

Many opportunities exist to create nanophase materials assembled from atom clusters of metals, ceramics, and semiconductors synthesized by means of a number of physical and chemical methods that appear to have very broad technological potential in the area of advanced materials. One of the traditional methods for synthesizing ultrafine powders or colloidal suspensions has been chemical precipitation. A wide variety of these methods, including sol-gel synthesis and the inverse micelle method, for example, have been successfully applied to the synthesis of nanometer-sized clusters with narrow size distributions. A number of high-temperature gas-reaction methods are also now becoming available for the synthesis of nanoscale clusters or larger powders with nanoscale substructures. Additional references to the variety of interesting methods available can be found elsewhere (Andres *et al.*, 1989; Aksay *et al.*, 1989; Rieke *et al.*, 1990; Averback *et al.*, 1991; Whitesides *et al.*, 1991; Heuer *et al.*, 1992; Komarneni *et al.*, 1993). Invariably, in each of the solution-chemical precipitation methods, the synthesized clusters are burdened by surface layers from the solutions in which they are formed. These surface layers are often intentionally introduced to cap the clusters at a given size or to keep them from agglomerating and forming large pores or to keep the clusters isolated from one another in order to control their interaction in the assembled material. If the clusters are to be assembled into monolithic bulk solids, where no such separation is desired, then the surface layers can become contaminants in the grain boundaries of the cluster-consolidated material, and processing and properties can be adversely affected. Such problems can be avoided in the gas-condensation method for synthesizing nanophase materials (Gleiter, 1989; Uyeda, 1991; Siegel, 1991), which will be highlighted in the present article.

The synthesis of ultrafine-grained materials by the *in situ* consolidation in vacuum of nanometer-size gas-condensed ultrafine particles or atom clusters first suggested by Gleiter (1981), was applied initially to metals (Birringer *et al.*, 1984, 1986). By consolidating clusters in this manner, materials with a large fraction of their atoms in grain boundaries could be formed. This method was subsequently applied to the synthesis of nanophase ceramics (Siegel and Hahn, 1987; Siegel *et al.*, 1988). The study of the gas condensation of ultrafine particles or atom clusters has, however, had a rather longer history stretching back to the formation and use of "smokes," such as carbon, or bismuth "blacks," for a variety of applications (Pfund, 1930). Scientific research into the controlled production of ultrafine particles by means of the gas-condensation method has been more recent (Kimoto *et al.*, 1963; Granqvist and Buhrman, 1976; Thölén, 1979), but still preceded cluster assembly and therefore provided an important basis for this work. In addition, of course, the previous knowledge of powder metallurgy and ceramics generated over an even longer period provided much needed background information for the work on nanophase materials to progress. The application of these ideas in recent years to the synthesis of a variety of nanophase metals and ceramics has built upon this broad scientific and technological base.

Early research on atom clusters formed via the gas-condensation method defined most of the important experimental parameters that control the sizes of the clusters formed in the conventional gas-condensation method that has been used to synthesize nanophase materials. The pioneering work of Uyeda and coworkers (Kimoto *et al.*, 1963) demonstrated that a wide range of metallic ultrafine particles (between 10 and 100 nm in diameter) could be condensed in a low-pressure Ar atmosphere and that the particle sizes could be controlled by varying the gas pressure in the range of about 1–30 Torr (0.13–4 kPa). Subsequent work by Thölén (1979) extended this method to a number of additional metals and to a study of the nucleation, growth, and coalescence of the clusters. However, the most detailed study to date of the conventional gas-condensation process for forming ultrafine metal particles or clusters via condensation in naturally convecting inert gases (He, Ar, or Xe) was carried out by Granqvist and Buhrman (1976). Some of their results will be discussed further below, but it was these early studies that elucidated the essential parameters (primarily type of gas, gas pressure, and evaporation rate) that control the formation of gas-condensed atom clusters that have made possible the synthesis of cluster-assembled nanophase materials in recent years.

The basic aspects of the generation of atom clusters via gas condensation can be described using the conceptual model shown in Fig. 2. A precursor material, either an element or compound, is evaporated in a gas maintained at a low pressure, usually well below one atmosphere. The evaporated atoms or molecules lose energy via collisions with the gas atoms or molecules and undergo a homogeneous condensation to form atom

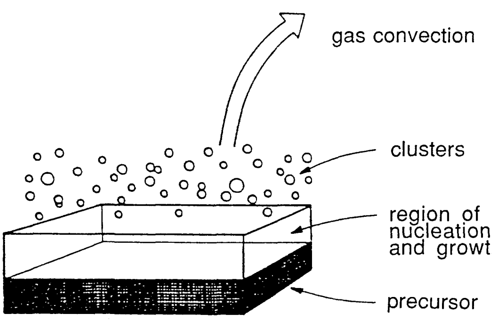

FIG. 2. Conceptual model for cluster formation via gas condensation (Granqvist and Buhrman, 1976).

clusters in the highly supersaturated vicinity of the precursor source. In order to maintain small cluster sizes, by minimizing further atom or molecule accretion and cluster-cluster coalescence, the clusters once nucleated must be removed rapidly from the region of high supersaturation. Since the clusters are already entrained in the condensing gas, this is readily accomplished by setting up conditions for moving this gas. Such gas motion has generally been driven by natural convection under the combined action of gravity and the temperature difference between the precursor source and a cooled thermophoretic cluster collection surface. However, a forced gas flow can also be used, with significant advantages in terms of both cluster-size control and process efficiency.

It should be emphasized that there are thus only three fundamental rates, which function relative to one another, that essentially control the formation of the atom clusters in the gas-condensation process. They are

1. the rate of supply of atoms to the region of supersaturation where condensation occurs,
2. the rate of energy removal from the hot atoms via the condensing medium, the gas, and
3. the rate of removal of the clusters once nucleated from the supersaturation region.

Other factors can also affect the clusters finally collected, particularly those that result in significant cluster-cluster coalescence, but these three rates represent the core of the process. Accordingly, the smallest cluster sizes for a given precursor are obtained for a low evaporation rate and condensation in a low pressure of a light inert gas, such as He. These conditions lead to a lower supersaturation of precursor atoms in the gas, slower removal of energy from the evaporated atoms (via the lighter gas atoms at lower pressure), and more rapid convective gas flow also as a result of the lower gas pressure. The rapid gas flow is significant, since it guarantees a shorter dwell time of the gas-entrained condensed clusters in the supersaturated region in which, if they remained, they would grow further. The effects of these controllable experimental parameters (gas type and pressure and evaporation rate) on the average cluster sizes produced under conditions of natural convective gas flow are shown in Figs. 3(a) and 3(b).

A typical apparatus (Siegel and Eastman, 1989) that uses such a process for the synthesis of nanophase materials via the *in situ* consolidation of gas-condensed clusters is shown schematically in Fig. 4. It consists of an ultrahigh-vacuum (UHV) system fitted with two resistively heated evaporation sources (which operate in the manner of Fig. 2), a cluster collection device (liquid-nitrogen–filled cold finger) and scraper assembly, and *in situ* compaction devices for consolidating the powders produced and collected in the chamber. Before making the powders, the UHV system is first evacuated by means of a turbomolecular pump to below 10^{-5} Pa and then back-filled with a controlled high-purity gas atmosphere at pressures of about a few hundred pascals. For producing metal powders this is usually an inert gas, such as He, but it can alternatively be a reactive gas or gas mixture if, for example, clusters of a ceramic compound are desired.

The clusters have rather narrow size distributions, as shown in Fig. 5, and are collected via thermophoresis on the surface of the cold finger in the form of very open, fractal structures. The clusters are weakly held on the collector surface by means of van der Waals type forces and are easily removed from this surface with a Teflon scraper. Upon removal, the clusters fall like snow from the surface and are funneled into a set of compaction devices (see Fig. 4) capable of consolidation pressures up to about 1–2 GPa, in which the nanophase samples are formed at room temperature, or at elevated temperatures if needed. The pellets formed in this conventional research apparatus are typically about 9 mm in diameter and 0.1 to 0.5 mm thick, depending upon the amount of material made (usually a few hundred milligrams) and the experiments to be performed. The sizes of these samples have been more a matter of laboratory convenience than any real limitation of the gas-condensation method itself. All of the fundamental rates involved in the cluster synthesis can be significantly increased above those in use in the laboratory, and scale-up leading to commercially viable production rates is already a reality. The scraping and consolidation are performed under UHV conditions after removal of the inert or reactive gases from the

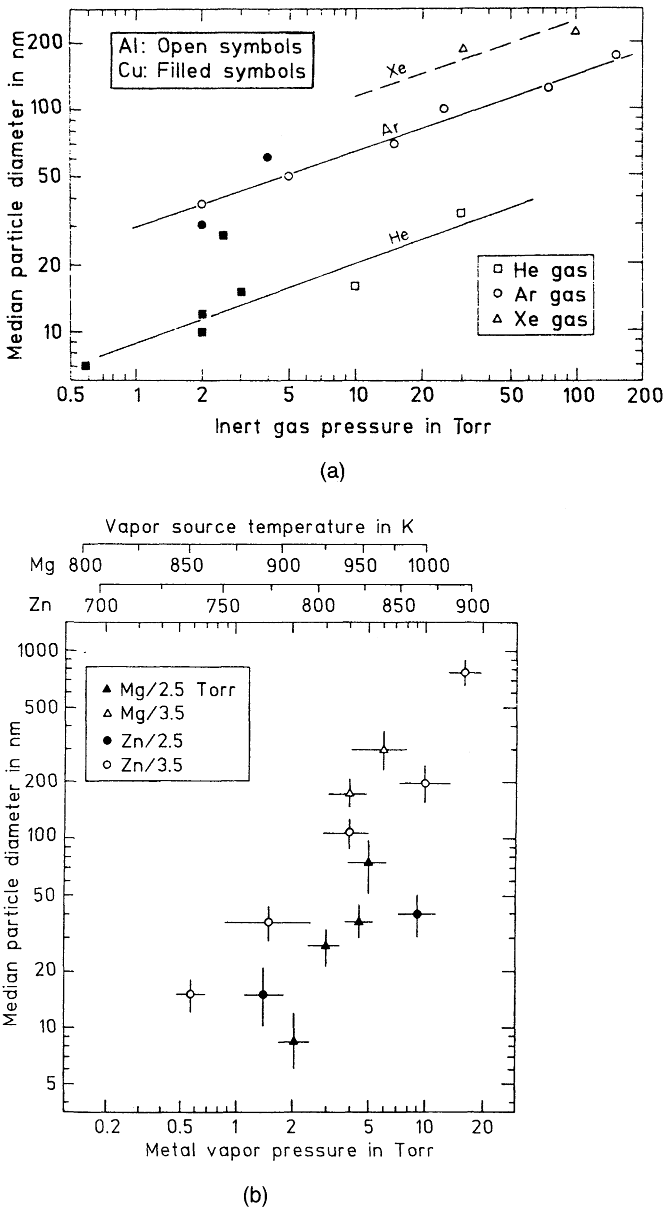

(a)

(b)

FIG. 3. (a) Median particle diameter versus pressure of He, Ar, or Xe gas for clusters of Al and Cu formed via gas condensation; the straight lines only serve to guide the eye. (b) Median particle diameter versus vapor pressure at the metal surface (or source temperature) for Mg and Zn evaporated and gas-condensed in two different Ar pressures (2.5 and 3.5 Torr) (Granqvist and Buhrman, 1976).

chamber, in order to maximize the cleanliness of the particle surfaces and the interfaces that are subsequently formed. Also, any possibility of trapping remnants of these gases in the nanophase compact is minimized by consolidation in vacuum. It should be noted in this regard that the total surface area of the nanophase powders produced in a given run is so great that, for a residual gas pressure of less than 10^{-5} Pa in a volume the size of the UHV chamber used, rather little gas contamination of the cluster surfaces would be expected.

Since the as-collected gas-condensed clusters are generally aggregated in rather open fractal arrays, their consolidation at pressures of 1–2 GPa is easily accomplished, even at room temperature. The difficulties in consolidating the hard equiaxed agglomerates of fine powders resulting from conventional wet-chemistry synthesis routes are mostly avoided. The sample densities resulting from cluster

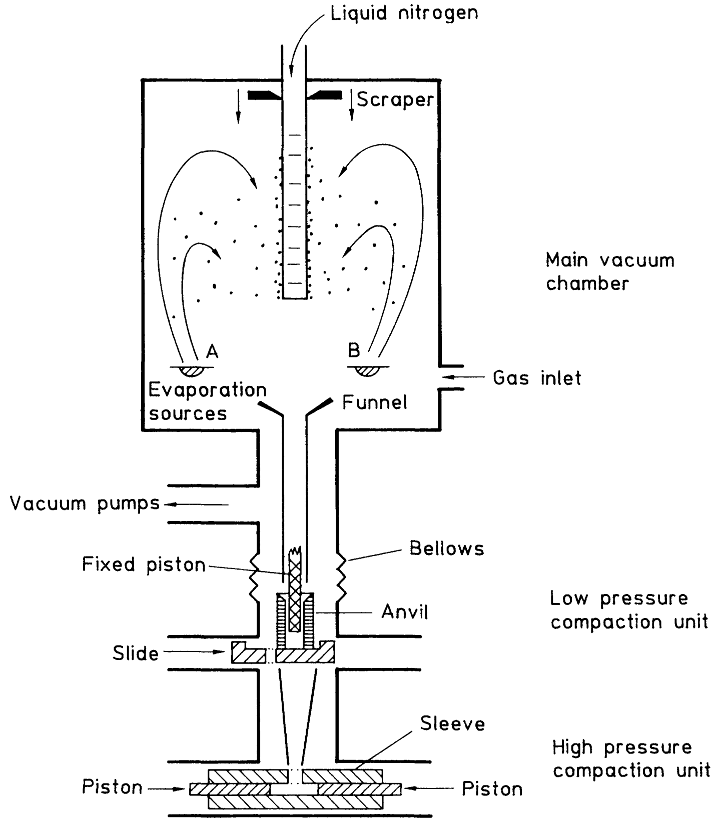

FIG. 4. Schematic drawing of a gas-condensation chamber for the synthesis of nanophase materials. Precursor material evaporated from sources A and/or B condenses in the gas and is transported via convection to the liquid-nitrogen–filled cold finger. The clusters are then scraped from the cold finger, collected via the funnel, and consolidated first in the low-pressure compaction unit and then in the high-pressure compaction unit, all in vacuum (Siegel and Eastman, 1989).

consolidation at room temperature have ranged up to about 97% of theoretical for nanophase metals and up to about 75–85% of theoretical for nanophase oxide ceramics. This as-consolidated, green-state porosity will be considered further in Sec. 3, but it probably represents (at least in part) a manifestation of powder agglomeration leading to voidlike flaws. Fortunately, these appear to be capable of being removed by means of cluster consolidation at elevated temperatures and pressures without significant attendant grain growth.

If an elemental precursor is evaporated in an inert gas atmosphere, then the atom clusters formed and collected are the same material, only in a reconstituted form. However, if clusters of a compound, such as a ceramic oxide, are desired, the process can become somewhat more complex. For example, in order to produce nanophase TiO_2 with a rutile structure and 12-nm average grain size, Ti metal clusters condensed in He were first collected on the cold finger and subsequently oxidized by the introduction of oxygen into the chamber (Siegel *et al.*, 1988). A similar method has been used to produce α-Al_2O_3 (Eastman *et al.*, 1989) with an 18-nm average grain size after oxidizing Al clusters in air at 1000 °C. If the vapor pressure of a compound is sufficiently large, as in the cases of MgO and ZnO, for example, it is possible to sublime the material directly from the oxide precursor in a He atmosphere containing, in addition, a partial pressure of O_2 to attempt to maintain oxygen stoichiometry during cluster synthesis. Such a method has been used (Eastman *et al.*, 1989) to produce nanophase MgO and ZnO with average grain sizes down to about 5 nm. Frequently, however, oxygen stoichiometry is not maintained.

In the case of nanophase TiO_2 cited above, the oxygen deficiency, while still present, is rather small and easily remedied as a result of the small grain sizes and short diffusion distances involved. Raman spectroscopy has been a useful tool in studying the oxidation state of nanophase TiO_2 because of the intense and well-studied Raman bands in both the anatase and rutile forms of this oxide and

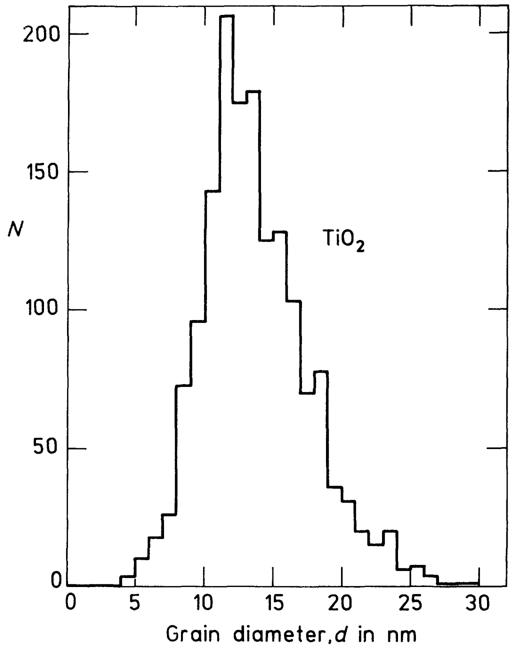

FIG. 5. Grain-size distribution for a nanophase TiO_2 (rutile) sample compacted to 1.4 GPa at room temperature, as determined by dark-field transmission electron microscopy (Siegel *et al.*, 1988).

the observation that these bands were affected in nanophase samples (Melendres *et al.*, 1989). A series of Raman spectra from two as-consolidated titanium dioxide nanophase samples and from one of these samples annealed in air until fully oxidized to TiO_2 is shown in Fig. 6. The band broadening observed in the nanophase samples (and also band shifting in both the anatase and rutile phases) was confirmed (Parker and Siegel, 1990a) to be the result of an oxygen deficiency, which could be subsequently removed in these samples by annealing in air. A subsequent calibration of this deviation from stoichiometry (Parker and Siegel, 1990b), shown in Fig. 7, indicated that $TiO_{1.89}$ was the actual material produced in the apparatus of Fig. 4, but that it could be easily oxidized to fully stoichiometric TiO_2, if desired, without sacrificing its small grain size (12 nm). Also, if intermediate deviations from stoichiometry were sought, in order to select particular material properties sensitive to the presence of oxygen deficient defects, they could be readily achieved as well.

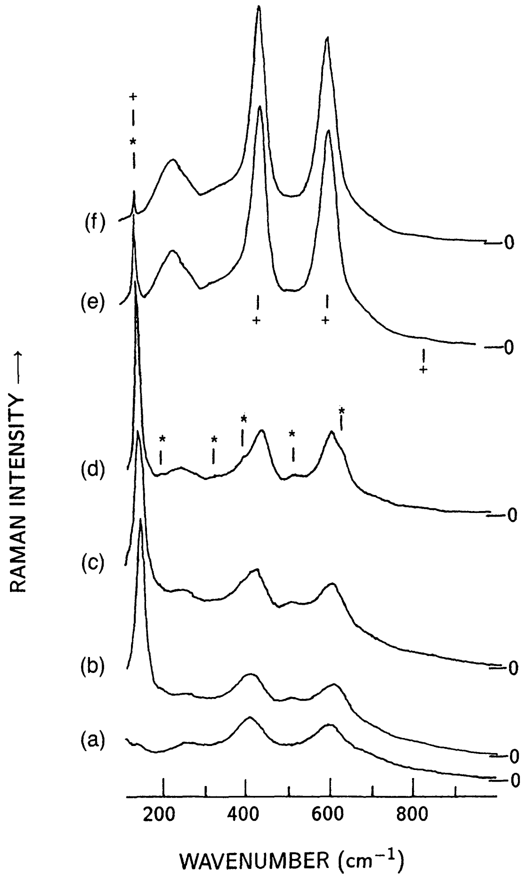

FIG. 6. (a), (b) Raman spectra of two areas of an as-compacted nanophase TiO_2 sample compared with spectra for the area used in (b) after successive annealing in air for 0.5 h at (c) 300 °C, (d) 500 °C, (e) 700 °C, and (f) 1000 °C (* indicates anatase $k = 0$ phonons; + indicates rutile $k = 0$ phonons) (Melendres *et al.*, 1989).

Most of the atom clusters assembled into nanophase materials to date have been generated from Joule-heated evaporation sources. However, such sources have limitations that need not be suffered, since a wide variety of other sources are also available. The primary limitations are source-precursor incompatibility, temperature range, uniformity and control, and dissimilar evaporation rates for different constituents in an alloy or compound precursor. Each of these limitations can be avoided by a host of alternative sources that have been developed over the years of ultrafine particle research, but that are only now beginning to enter the field of nanophase materials synthesis (Uyeda, 1991; Siegel 1991). Among other sources for bringing atom supersaturations into a condensing gas

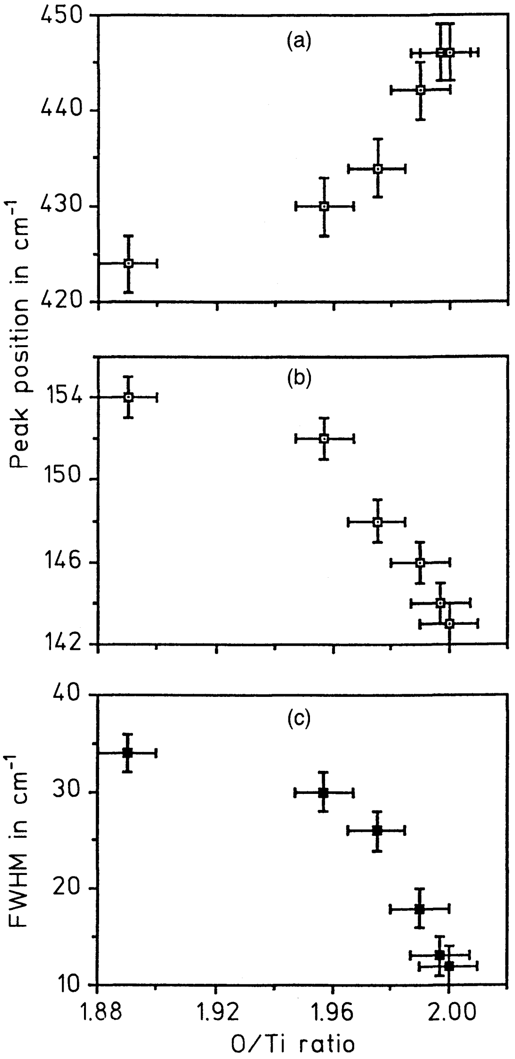

FIG. 7. Variation with O/Ti ratio of the peak position of (a) the rutile "447 cm^{-1}" vibrational mode and (b) the anatase "143 cm^{-1}" vibrational mode, as well as (c) this anatase mode's full width at half maximum (FWHM) (Parker and Siegel, 1990b).

medium that have been successfully used to produce clusters or ultrafine particles are Joule-heated ovens, sputtering, electron-beam heating, laser ablation, and plasma methods. This wide variety of evaporation methods will allow for greatly increased flexibility in the use of refractory or reactive precursors for clusters, and will be especially useful as one moves toward synthesizing technological quantities of more complex multicomponent or composite nanophase materials.

2.2 Mechanical Attrition

Before turning to the structure and properties of nanophase materials, it might be useful to mention the mechanical attrition method (Hellstern *et al.*, 1989; Koch *et al.*, 1989; Fecht *et al.*, 1989; Koch, 1993). Mechanical attrition produces its nanostructures, not by cluster assembly, but by means of the structural decomposition of coarser-grained structures induced by severe mechanical deformation. Nanometer-size grains nucleate within the localized shear bands of heavily deformed materials, converting a coarse-grained structure to nanophase. The heavy deformation is usually induced by means of high-energy ball milling, but can result as well from surface wear phenomena (Ganapathi and Rigney, 1990) and other methods for introducing high densities of deformation, such as the high-energy shear process (Valiev, 1993). A series of transmission electron micrographs and associated electron-diffraction patterns showing this grain refinement in AlRu as a function of ball-milling time is presented in Fig. 8. The individual grains resulting from mechanical attrition are never isolated clusters, and much synthesis and processing flexibility is lost. Nevertheless, ultrafine grain sizes can be readily achieved by this rather straightforward method, albeit with probable contamination from the sources of mechanical work. However, the method is capable of producing commercial quantities of material, and for those applications in which careful control of purity or grain morphology are not important, and for which the desired properties are indeed accessible in the mechanically attrited nanostructures, it can be a useful source of these materials.

The variation with grain size as a function of ball-milling time in Pd (Eckert *et al.*, 1992) is shown in Fig. 9(a). This type of saturation behavior is found to be typical for ball-milled metals and clearly shows that grain sizes down into the nanometer regime can be readily achieved, although it is more easily accomplished in relatively harder materials, as indicated by the relationship between minimum grain size and bulk modulus shown in Fig. 9(b). Mechanical attrition exercised at low temperatures (cryomilling) can effectively extend the range of applicability, as

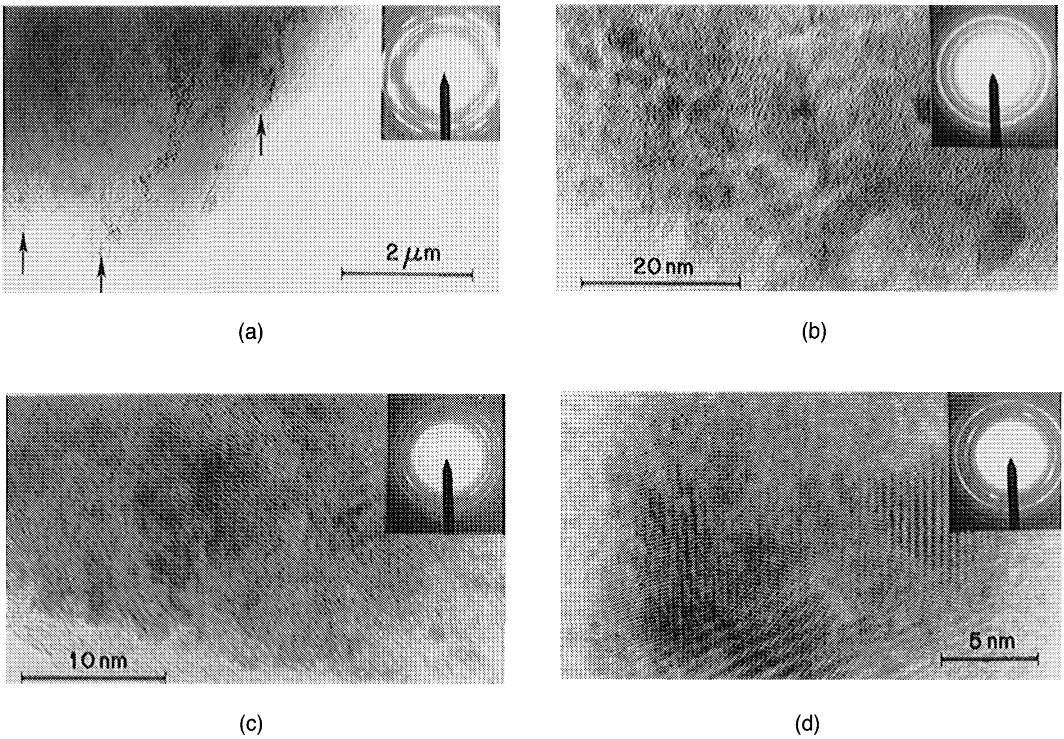

FIG. 8. Variation of the microstructure of AlRu with time of high-energy ball milling as seen by TEM and electron diffraction patterns shown in the insets (Fecht *et al.*, 1989): (a) After 10 min, arrows indicating highly deformed shear bands; high resolution images and diffraction patterns after (b) 10 min, (c) 2 h, and (d) 64 h.

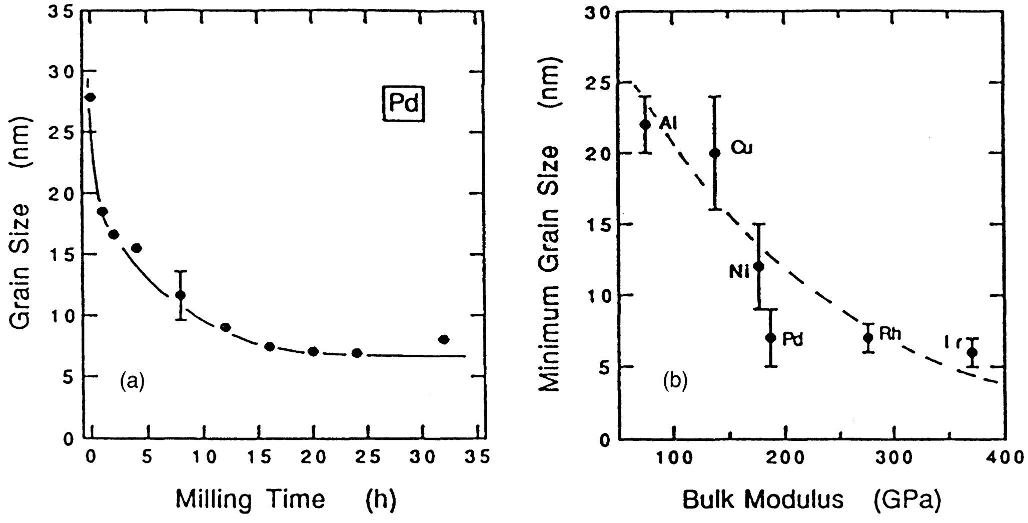

FIG. 9. (a) Average grain size, determined by x-ray line broadening, of Pd as a function of ball-milling time. (b) Minimum average grain size attained by ball milling of various fcc metals plotted against their bulk modulus values (Eckert *et al.*, 1992).

shown by Luton *et al.* (1989) in their work on dispersion-strengthened Al.

3. STRUCTURE

The structures of nanophase materials on a variety of length scales have an important bearing on their properties. They are dominated by ultrafine grain sizes and the large number of associated interfaces. In addition, however, other structural features, such as pores (and larger flaws), grain-boundary junctions, and other crystal-lattice defects that can depend upon the manner in which these materials are synthesized and processed, play a significant role. It has become increasingly clear that all of these structural aspects of nanophase materials must be considered in trying to understand fully the properties of these new materials.

3.1 Grains and Pores

Our present knowledge of the grain structures of nanophase materials, whether formed by cluster consolidation or by intense mechanical deformation, has resulted primarily from direct observations using transmission electron microscopy (TEM) (Siegel *et al.*, 1988; Thomas *et al.*, 1989, 1990; Wunderlich *et al.*, 1990). A typical high-resolution image of a nanophase Pd sample formed by the consolidation of gas-condensed clusters is shown in Fig. 10. TEM has shown that the grains in such nanophase compacts are essentially equiaxed, similar to the atom clusters from which they were formed, although departures from spherical structures are expected simply from the efficient packing of the clusters during consolidation. The grains also appear to retain the narrow (ca. ±25% FWHM) log-normal size distributions typical of the clusters formed in the gas-condensation method (see Fig. 5), since measurements of these distributions before or after cluster consolidation by dark-field electron microscopy yield similar results. Grain-size distributions in deformation-produced nanostructures can be somewhat broader.

The observations (Nieman *et al.*, 1991) that the densities of nanophase materials consol-

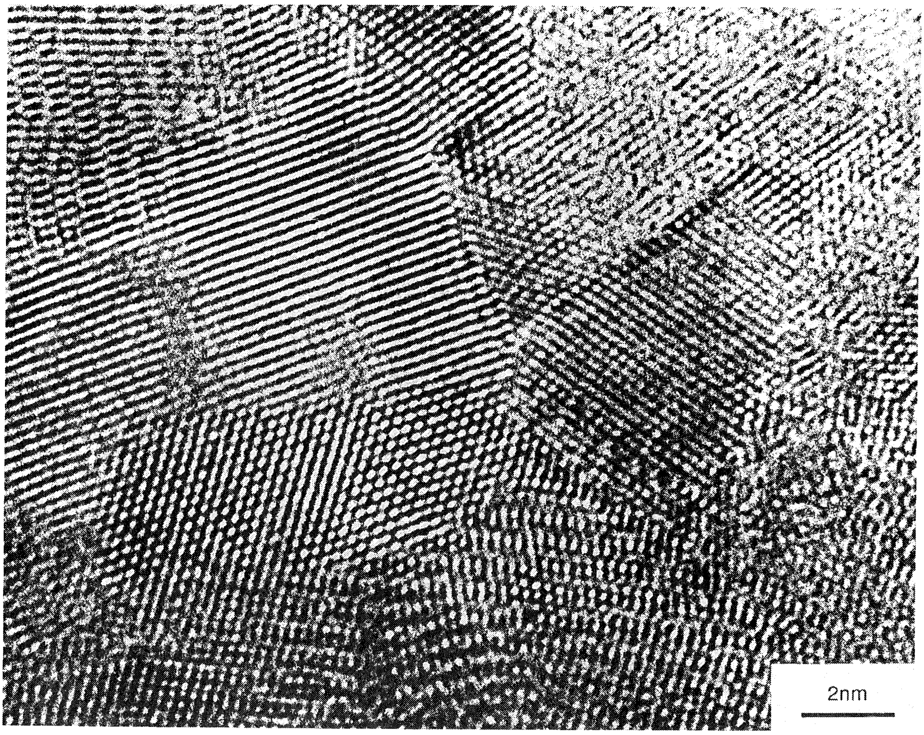

FIG. 10. High-resolution transmission electron micrograph of a typical area in nanophase palladium (Thomas *et al.*, 1990).

idated from equiaxed clusters extend well beyond the theoretical limit (78%) for the densest packing of undeformed identical spheres indicate that an extrusion-like deformation of the clusters must result during the consolidation process, filling (at least partially) the pores among the grains. Such observations indicate that cluster extrusion in forming nanophase grains may result from a combination of deformation and diffusion processes. These processes are also evident from scanning tunneling and atomic force microscopy observations (Sattler *et al.*, 1994) on nanophase Pd and Ag and from computer simulations of the consolidation of Pd clusters (Liu *et al.*, 1994).

Observations by both electron and x-ray scattering indicate, however, that no apparent preferred crystallographic orientations or texture of the grains results from their uniaxial consolidation and that the grains in the nanophase compact are essentially randomly oriented with respect to one another. This is interesting with respect to the fact that such deformation and annealing textures commonly observed in conventional grain-size materials are the result of dislocation glide processes, which seem to be suppressed in nanophase materials, as will be discussed in the following. Indeed, only very few dislocations are observed within the grains in these these ultrafine-grained materials, and these are normally in immobile or locked (i.e., sessile) configurations.

All of the nanophase materials consolidated to date from clusters at room temperature have invariably possessed some porosity. This porosity has ranged from about 15–25% for ceramics to less than 5% for metals. Significant porosity can also result from the deformation or crystallization synthesis of nanophase materials. Evidence for porosity has been obtained by positron annihilation spectroscopy (PAS) (Schaefer *et al.*, 1987, 1988; Siegel *et al.*, 1988), precise densitometry (Nieman *et al.*, 1991), and porosimetry (Hahn *et al.*, 1990a; Wagner *et al.*, 1991) measurements, and small-angle neutron scattering (SANS) (Wagner *et al.*, 1991; Sanders *et al.*, 1993).

PAS is primarily sensitive to small pores, ranging from single vacant lattice sites to larger voids, but PAS can probe these structures enclosed in the bulk of the material, even as a function of sintering as shown in Fig. 11 and discussed in Sec. 4.1. Porosimetry measurements using the BET (Brunauer–Emmett–Teller) N_2 adsorption method (Gregg and Sing, 1982), on the other hand, probe only pore structures open to the free surface of the sample, but BET can yield pore-size distributions (although with some questions regarding their validity at nanometer pore sizes), which are unavailable from PAS. Such a set of pore-size distributions in TiO_2 is shown in Fig. 12. Densitometry using an Archimedes method integrates over all structures in the sample, including grains, pores (open or closed), and density decrements at defect sites as well. SANS is quite sensitive to pores in the 1–100-nm size range, but deconvolution of the scattering data can be difficult when a broad spectrum of scattering centers is present, as is often the case in nanophase materials. However, even in such a case, it has been possible (Sanders *et al.*, 1993) to analyze SANS data from cluster-consolidated Pd in terms of a population of small (ca. 1-nm–diameter) voids presumed to be located at grain-boundary intersections (triple junctions) and grain-sized voids; however, the possible scattering from grain boundaries or larger flaws was not taken into account in the analysis.

A variety of such measurements to date have shown that the porosity in nanophase metals and ceramics is primarily smaller than or equal to the grain size of the material (although some larger porous flaws have been observed by optical and scanning electron microscopy). The porosity is frequently associated with the grain-boundary junctions (triple junctions) and especially, but not only, in ceramics, it is interconnected and intersects the specimen surfaces. Fortunately, consolidation at elevated temperatures should be able to remove this porosity uniformly without sacrificing the ultrafine grain sizes in these materials. Some evidence for this has already been obtained by means of the sinter forging of nanophase ceramics (Owen and Chokshi, 1993).

3.2 Dislocations

Dislocations are most often the type of lattice defect responsible for the permanent or plastic deformation in metals and nonmetal crystalline materials. As a dislocation moves through the crystal lattice under the influ-

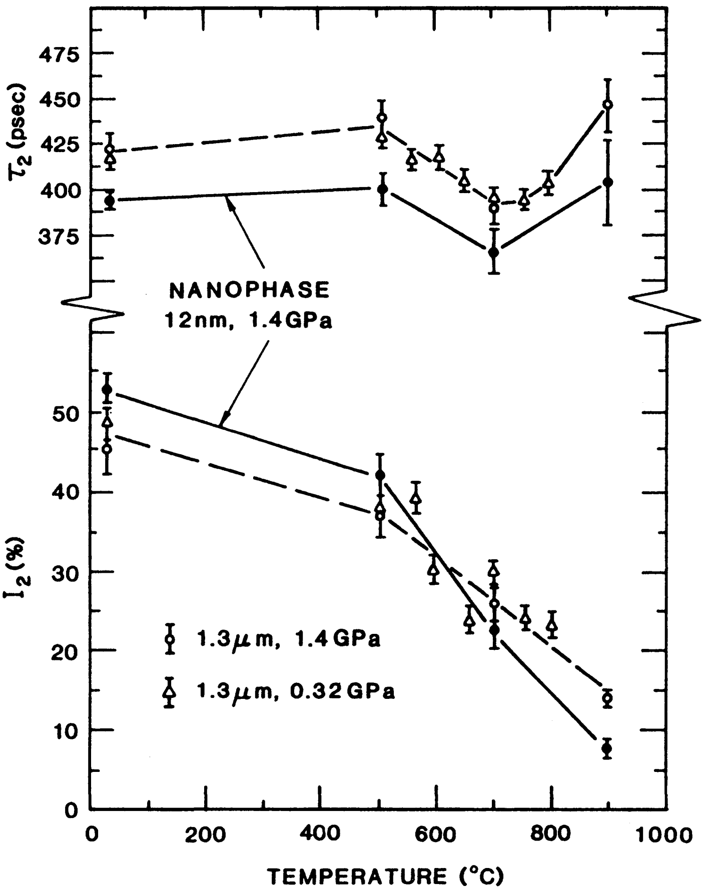

FIG. 11. Results of two-component lifetime fits to positron annihilation data from three TiO_2 samples as functions of sintering temperature; τ_2 corresponds to positrons trapped at voids (pores) and I_2 is the relative intensity of this component. A 12-nm grain-size nanophase sample (filled circles) compacted at 1.4 GPa is compared to 1.3-μm grain-size samples compacted at 1.4 GPa (open circles) and 0.32 GPa (triangles) from commercial powder. The PAS data were taken at room temperature; no sintering aids were used (Siegel *et al.*, 1988).

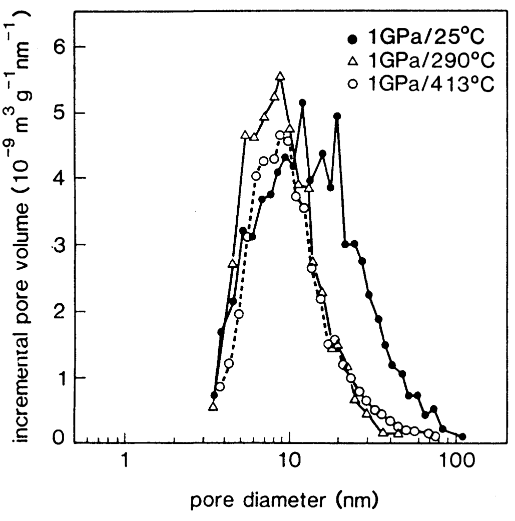

FIG. 12. Pore-size distributions in nanophase TiO_2 consolidated under 1 GPa pressure at room temperature, 290 °C, and 413 °C, as measured by the BET method (Wagner *et al.*, 1991).

ence of an externally applied stress, its motion imparts a permanent microscopic displacement that taken together with the motion of other dislocations can lead to a macroscopic shape change in the material. The presently available experimental evidence suggests that dislocations are seldom present in nanophase materials (Gao and Gleiter, 1987; Thomas *et al.*, 1990; Milligan *et al.*, 1993). When they are observed, it is primarily either in materials at the upper end of the grain size range (Morris and Morris, 1991) or in limited instances (Wunderlich *et al.*, 1990; Jain and Christman, 1994) in immobile or locked configurations. The reason for this substantial lack of dislocations is that image forces exist in finite atomic ensembles that tend to pull mobile dislocations out of the grains, especially when they are small, in analogy with the forces on a point electrical charge near a free surface of a conducting body. Since mobile dislocations are not initially available in sufficient numbers to effect plasticity in nanophase materials, new mobile dislocations must be created or other deformation mechanisms, such as grain-boundary sliding, must come into play.

Dislocations can be created or can multiply from a variety of sources. A simple but representative example is the Frank–Read dislocation source, shown in Fig. 13, in which a dislocation line, pinned between two pinning points that prevent its forward motion on a slip plane, can bow out between these pinning points to form a new dislocation, if the stress acting on the pinned dislocation is sufficient. The critical stress to operate such a Frank–Read source is inversely proportional to the distance between the pinning points and hence will also be limited by the grain size, which limits the maximum distance between such pinning points. This suggests that dislocation multiplication in nanophase metals will become increasingly difficult as the grain size decreases (see Sec. 4.2). The critical stress will eventually, at sufficiently small grain sizes, become larger than the yield stress (at which plastic deformation begins) in the conventional material and could even approach the theoretical yield strength of a perfect, dislocation-free single crystal.

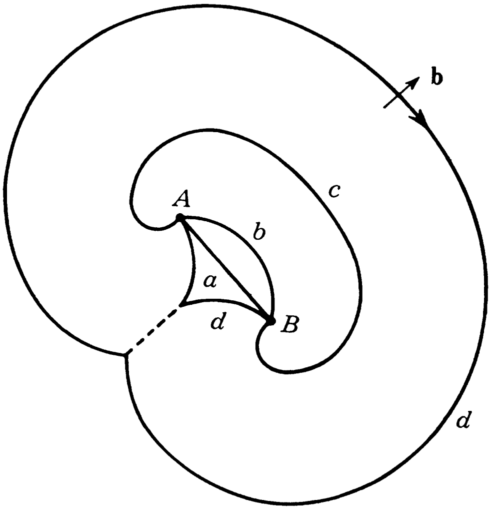

FIG. 13. Schematic representation of a Frank–Read dislocation source. A segment of dislocation loop **a** of length L and Burgers vector **b** is pinned between points A and B in a material of shear modulus G. When the stress applied to the dislocation line segment exceeds a critical stress, $\sigma_{\text{crit}} = kGb/L$, the dislocation bows out ($a \rightarrow b \rightarrow c \rightarrow d$) until it creates a new dislocation loop. The constant of proportionality, k, is 0.5 for an edge dislocation and 1.5 for a screw dislocation.

3.3 Grain Boundaries

As a result of their ultrafine grain sizes, nanophase materials have a significant fraction of their atoms in grain-boundary environments, where they occupy positions relaxed from their normal lattice sites. For high-angle grain boundaries (interfaces between grains with more than a few degrees of orientational mismatch) in conventional materials, these relaxations extend over about two atom planes on either side of the boundary, with the greatest relaxation existing in the first plane (Wolf and Lutsko, 1988; Wolf and Yip, 1992). Thus, for an average grain diameter range between 5 and 10 nm, where much of the research on nanophase materials has focused, grain-boundary atom percentages range between about 5 and 50% (see Fig. 14). Since such a large fraction of their atoms reside in grain boundaries, these interface structures can play a significant role in affecting the properties of nanophase materials.

Several early investigations of nanocrystalline metals, including x-ray diffraction (Zhu *et al.*, 1987), Mössbauer spectroscopy (Herr *et al.*, 1987), positron lifetime studies (Schaefer

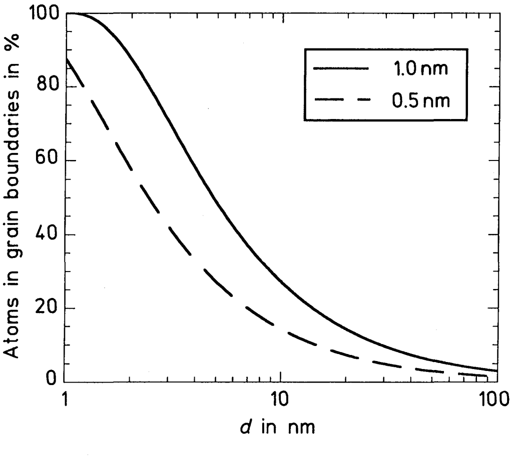

FIG. 14. Percentage of atoms in grain boundaries (including grain-boundary junctions) of a nanophase material as a function of grain diameter, assuming that the average grain-boundary thickness ranges from 0.5 to 1.0 nm (ca. 2 to 4 atomic planes wide) (Siegel, 1991).

et al., 1987, 1988), and extended x-ray absorption fine-structure (EXAFS) measurements (Haubold *et al.*, 1988, 1989), were interpreted in terms of grain-boundary atomic structures that may be random, rather than possessing either the short-range or long-range order normally found in the grain boundaries of conventional coarser-grained polycrystalline materials. This randomness was variously associated (Birringer and Gleiter, 1988) with either the local structure of individual boundaries (as seen by a local probe such as EXAFS or Mössbauer spectroscopy) or the structural coordination among boundaries (as might be seen by x-ray diffraction). However, direct observations by high-resolution electron microscopy (HREM) have indicated that their structures are rather similar to those of conventional high-angle grain boundaries. An extensive review of these results has appeared elsewhere (Siegel, 1992).

The direct imaging of grain boundaries with HREM can avoid the complications that may arise from porosity and other defects in the interpretation of data from less direct methods, such as x-ray scattering and Mössbauer spectroscopy. A HREM study (Thomas *et al.*, 1989, 1990) that included both experimental observations and complementary image simulations indicated no manifestations of grain-boundary structures with random displacements of the type or extent suggested by earlier x-ray studies on nanophase Fe, Pd, and Cu (Zhu *et al.*, 1987; Haubold *et al.*, 1988, 1989). Contrast features at the observed grain boundaries that might be associated with disorder did not appear wider than 0.4 nm, indicating that any significant structural disorder that may be present essentially extends no further than the planes immediately adjacent to the boundary plane. Such localized lattice relaxation features are typical of the conventional high-angle grain-boundary structures found in coarse-grained metals. HREM investigations of grain boundaries in nanophase Cu (Ganapathi and Rigney, 1990) and Fe alloys (Trudeau *et al.*, 1991) produced by surface wear and high-energy ball milling, respectively, appear to support this view. Recent nuclear magnetic resonance (NMR) studies of cluster-consolidated nanophase Ag (Suits *et al.*, 1993) have indicated that the grain boundaries in this material have an electronic structure consistent with that for conventional high-angle boundaries. However, there are indications (Tschöpe and Birringer, 1993; Valiev *et al.*, 1991) that metastable grain-boundary configurations do exist in some cases and that these configurations can be transformed by low-temperature annealing to more stable states. Whether this behavior is associated with the intrinsic grain-boundary structure itself or with extrinsic grain-boundary dislocation configurations and/or strains remaining from synthesis or processing remains to be clarified.

Indeed, as shown in Fig. 10 and in similar electron micrographs, the nanophase grain boundaries appear to be essentially low-energy configurations exhibiting flat facets interspersed with steps. Such structures could arise only if sufficient local atomic motion occurred during the cluster consolidation process to allow the system to reach at least a local energy minimum. These observations suggest at least two conclusions (Siegel and Thomas, 1992): first, that the atoms that constitute the grain-boundary volume in cluster-assembled nanophase materials have sufficient mobility during cluster consolidation to accommodate themselves into relatively low-energy grain-boundary configurations; and second, that the local driving forces for grain growth are relatively small, despite the large amount of energy stored in the many grain boundaries in these materials.

The foregoing discussion suggests that nanophase materials should be a valuable resource for studying the average properties of grain boundaries. The high number density of such defects in these materials enhances their influence on macroscopic properties, allowing these effects to be studied by a variety of experimental techniques. Indeed, a number of the effects observed in nanophase materials to date that have been deemed unusual may simply result from so many grain boundaries being available for study in a sample for the first time. For the careful study of grain-boundary properties to be successful in the future, however, specimen porosity will need to be removed, via consolidation at elevated temperature and/or pressure, so that its property contributions can be eliminated. By varying the grain size in a series of samples, the effects from interfaces and junctions in nanophase materials could be effectively separated in such future studies.

3.4 Stability and Strains

The experimental observations of narrow grain-size distributions, essentially equiaxed grain morphologies, and low-energy grain-boundary structures in nanophase materials suggest that the inherent resistance to grain growth observed for cluster-assembled nanophase materials (see Fig. 15) results primarily from a sort of frustration (Siegel, 1990). It appears that the narrow grain-size distributions normally observed in these nanophase materials coupled with their relatively flat grain-boundary configurations place these ultrafine-grained structures in a local minimum in energy from which they are not easily extricated. Such frustration would also likely be increased by the multiplicity of grain-boundary junctions in these materials. There are normally no really large grains to grow at the expense of small ones through an Otswald ripening process, and the grain boundaries, being essentially flat, have no local curvature to tell them in which direction to migrate. Their stability thus appears to be analogous to that of a variety of closed-cell foam structures with narrow cell-size distributions, which are deeply metastable despite their large stored surface energy. Such a picture now appears to have some theoretical support (Rivier, 1992).

It should be pointed out that exceptions to this frustrated grain growth behavior would be expected if considerably broader grain-size distributions were accidentally present in a sample, which would allow a few larger grains to grow at the expense of smaller ones, or if significant grain-boundary contamination were present, allowing enhanced stabilization of the small grain sizes to further elevated temperatures. Occasional observations of each of these types of behavior have been made. One could, of course, intentionally stabilize against grain growth by appropriate doping by insoluble elements or composite formation in the grain boundaries. For cluster-assembled materials, such stabilization should be especially easy, since the grain boundaries are available as cluster surfaces prior to consolidation. The ability to retain the ultrafine grain sizes of nanophase materials is important when one considers the fact that it is their grain size and large number of grain boundaries that determine to a large extent their special properties.

Strains are a matter of fact in nanophase materials. Simply as a result of the large number of grain boundaries, and the concomitant short distances between them, the intrinsic strains associated with such interfaces (Wolf and Lutsko, 1988; Cammarata and Sieradski, 1989) are always present in these materials. Beyond these intrinsic strains, there may also be present extrinsic strains associated with the particular synthesis method. For example, intense plastic deformation synthesis of nanophase materials may lead to additional residual strains (Valiev, 1993) that can be subsequently relieved by low-temper-

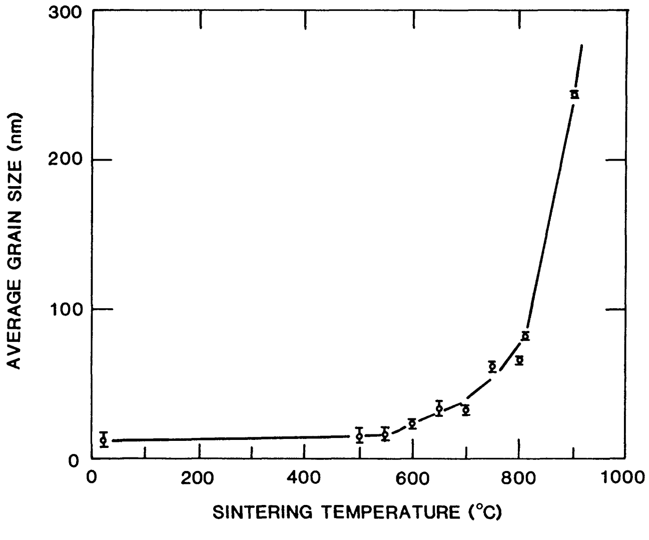

FIG. 15. Variation of average grain size with sintering temperature (0.5 h at each) for a nanophase TiO_2 (rutile) sample compacted to 1.4 GPa at room temperature, as determined by dark-field transmission electron microscopy (Siegel *et al.*, 1988).

ature annealing, leaving only those intrinsic strains due to the presence of the high-angle grain boundaries. Evidence for the strains in nanophase materials is now becoming available. X-ray line-broadening investigations of both cluster-consolidated (Nieman *et al.*, 1991) and ball-mill-attrited metallic nanophase samples (Eckert *et al.*, 1992) have indicated residual strains of about (0.1–1)%, which are consistent in magnitude with the strains expected from conventional high-angle grain boundaries (Wolf and Yip, 1992) in the grain size ranges investigated. Also, recent NMR measurements (Suits *et al.*, 1994) on cluster-consolidated nanophase Cu have yielded strain values in this range.

4. PROPERTIES

The unique properties of nanophase materials result from an interplay among their three fundamental features: constituent atomic domains (e.g., grains) spatially confined to less than 100 nm, significant atom fractions associated with interfacial environments (e.g., grain boundaries or free surfaces), and interactions between their domains. In some cases one of these features dominates, in other cases another. Research on a variety of chemical, mechanical, and other physical properties is beginning to yield a glimmer of understanding of just how this interplay manifests itself in the properties of these new materials.

4.1 Chemical Properties

Nanophase materials exhibit properties that are different and often considerably improved in comparison with those of conventional coarse-grained structures. Because of their small sizes and radii of curvature, coupled with their surface cleanliness, the constituent clusters of nanophase materials can react with one another rather aggressively, even at relatively low temperatures. For example, nanophase TiO_2 (rutile) exhibits significant improvements in both sinterability and resulting mechanical properties relative to conventionally synthesized coarser-grained rutile (Siegel *et al.*, 1988; Averback *et al.*, 1989; Hahn *et al.*, 1990a; Mayo *et al.*, 1990). Nanophase TiO_2 with a 12-nm initial mean grain diameter has been shown to sinter under ambient pressures at temperatures 400 °C to 600 °C lower than the conventional sintering temperature for coarse-grained rutile, without the need for any normally required compacting or sintering aid such as polyvinyl alcohol, as shown in Fig. 16. Furthermore, it has been demonstrated (Hahn *et al.*, 1990a) that sintering the same nanophase material under pressure (1 GPa) or with appropriate dopants such as Y can further reduce the sintering temperatures, while suppressing grain

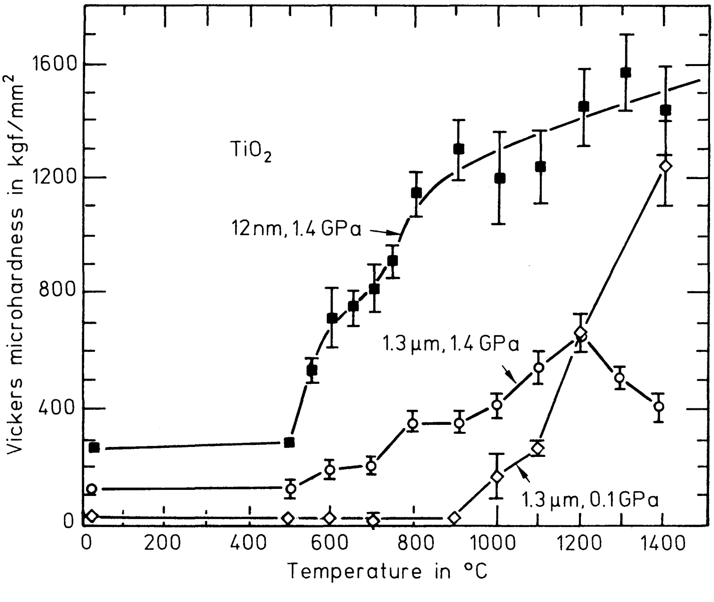

FIG. 16. Vickers microhardness of TiO_2 (rutile) measured at room temperature as a function of 0.5-h sintering at successively increased temperatures. Results for a nanophase sample (filled squares) with an initial average grain size of 12 nm consolidated at 1.4 GPa are compared with those for coarser-grained samples with 1.3-μm initial average grain size sintered with (diamonds) or without (circles) the aid of polyvinyl alcohol from commercial powder consolidated at 0.1 GPa and 1.4 GPa, respectively (Siegel *et al.*, 1988).

growth as well, thus allowing for the unique possibility of sintering nanophase ceramics to full density while retaining their ultrafine grain size. The resulting fracture characteristics (Li *et al.*, 1988b; Averback *et al.*, 1989; Mayo *et al.*, 1990; Höfler and Averback, 1990) developed for sintered nanophase TiO_2 are as good, or in some aspects improved, relative to those for conventional coarser-grained rutile.

Sintering behavior has also been followed by PAS and SANS measurements. It was already discussed in Sec. 3.1 how PAS can be a useful tool in the study of the ultrafine-scale porosity inherent in as-consolidated nanophase compacts; it can probe such porosity as a function of sintering temperature as well, to observe densification via the removal of voids. An example of PAS lifetime results (Siegel *et al.*, 1988) used to follow the sintering behavior of nanophase TiO_2 was already shown in Fig. 11. The intensity I_2 of the lifetime (τ_2) signal corresponding to positron annihilation from void-trapped states in the nanophase sample is seen to decrease rapidly during sintering above 500 °C as a result of the densification of this ultrafine-grained ceramic, even though rapid grain growth does not set in until above 800 °C (see Fig. 15). Furthermore, the variation of τ_2 with sintering indicates that there is a redistribution of void sizes accompanying this densification. Similar behavior is also observed for the coarser-grained samples investigated, but, as expected, the densification proceeds more slowly in these latter samples and the average pore sizes are larger according to the larger values of τ_2. The redistribution of pore sizes can also be monitored by means of BET measurements (if the pores are still open to the sample surfaces so that N_2 can enter) as demonstrated by Hahn *et al.* (1990a) and Wagner *et al.* (1991). Nanophase TiO_2 in its as-consolidated state and as a function of sintering in air was also followed by SANS (Epperson *et al.*, 1989). While SANS contrast can yield information regarding the nature of the intergrain nanophase boundaries, particularly their average density *vis-à-vis* the grain density (Jorra *et al.*, 1989; Epperson *et al.*, 1989, 1990), it can also yield information about the presence of voids (or porosity) and their removal during the sintering process. However, the clear separation of void and grain-boundary contributions to small-angle neutron scattering is still an open question, as discussed in Sec. 3.1, even though an attempt at using pressure-assisted sintering to remove the voids has recently been made (Wagner *et al.*, 1991).

Atomic diffusion in nanophase materials, which can have a significant bearing on their mechanical properties, such as creep and superplasticity, and other physical properties as well, has been found to be very rapid. Measurements of self-diffusion and impurity diffusion (Horváth *et al.*, 1987; Mütschele and Kirchheim, 1987; Horváth, 1989; Hahn *et al.*, 1989; Schumacher *et al.*, 1989; Averback *et al.*, 1989) in as-consolidated nanophase metals (Cu,Pd) and ceramics (TiO_2) indicate that atomic transport can be orders of magnitude faster in these materials than in coarser-grained polycrystalline samples, exhibiting surface-diffusion-like behavior. However, the very rapid diffusion in as-consolidated nanophase materials appears to be intrinsically coupled with the porous nature of the interfaces in these materials. It has been shown in at least one case (Hf in TiO_2) that the rapid surfacelike diffusivities can be suppressed back to conventional values by sintering samples to full density (Averback *et al.*, 1989). Nonetheless, there exist considerable possibilities for efficiently doping nanophase materials at relatively low temperatures by means of the rapid diffusion available along their ubiquitous grain-boundary networks and interconnected porosity, with only short diffusion paths remaining into their grain interiors, to synthesize materials with tailored chemical, mechanical, or other physical properties.

The chemical reactivity of nanophase materials, with their potentially high surface areas compared to conventional materials, can also be rather striking. Since clusters can be assembled by means of a variety of methods, for example, there can be an excellent degree of control over the total available surface area in the resulting self-supported ensembles. Thus, one can maximize porosity for obtaining very high surface areas, remove most of it via consolidation, but retain some to facilitate low-temperature doping or other processing, or fully densify the nanophase material. Also, composition control can be readily achieved, since rapid diffusion paths are plentiful and diffusion distances are short in the clusters. Measurements (Beck and Siegel,

1992) of the decomposition of H_2S over lightly consolidated nanophase TiO_2 at 500 °C demonstrate the enhanced chemical reactivity of nanophase materials rather well. Figure 17 shows the reaction rates for S removal from H_2S via decomposition for nanophase TiO_2 (rutile) compared with that for a number of other commercially available forms of TiO_2 having either the rutile or anatase structure. It can be seen that the nanophase sample was far more reactive than any of the other samples tested, both initially and after extended exposure to the H_2S. Indeed, only the nanophase titania demonstrated a nonzero reaction rate for S removal after 7 h. This greatly enhanced and sustainable activity was shown to result from a combination of unique features of the nanophase material, its high specific surface area combined with its rutile structure and its oxygen-deficient composition (see Sec. 2.1).

Another important aspect of materials with reduced spatial dimensions in the nanometer size regime is phase stability. As one reduces the sizes of constituent phases sufficiently, into the nanoscale (mesoscopic) regime where condensed matter behaves differently than either atomic or molecular species or bulk solids, equilibrium phase relations will change because of the expected changes in a variety of electronic and thermodynamic parameters of the confined atomic system. For example, the melting temperature of small clusters is suppressed well below that for the bulk solid (Buffat and Borel, 1976; Berry *et al.*, 1984; Han and Whetten, 1988; Goldstein *et al.*, 1992), the result of a higher effective pressure in the confined system, the Gibbs–Thomson effect. Related effects are also seen in cluster-consolidated materials in which high-pressure polymorphic structures have been stabilized (Li *et al.*, 1988a; Skandan *et al.*, 1992). Thus, the effects of reduced spatial dimensions on phase stability are important, since one cannot expect that the phase equilibria known for conventional materials systems will apply in general to those of significantly reduced dimensions.

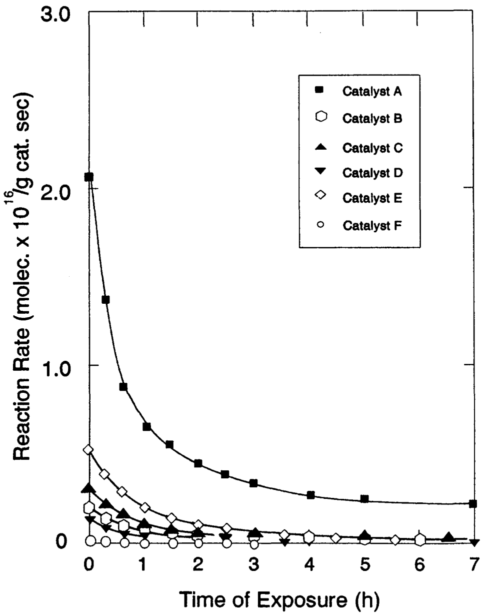

FIG. 17. Reaction rates for H_2S decomposition as a function of exposure time at 500 °C for nanophase TiO_2 compared with that from several commercially available TiO_2 materials and a reference (catalyst A, 76 m^2/g nanophase rutile; B, 61 m^2/g anatase; C, 2.4 m^2/g rutile; D, 30 m^2/g anatase; E, 20 m^2/g rutile; F, reference alumina) (Beck and Siegel, 1992).

4.2 Mechanical Properties

The mechanical properties of nanophase materials are generally rather different from those of their coarse-grained counterparts (Siegel and Fougere, 1994). This difference results primarily from spatial confinement and induced accommodation strains in nanophase metals and from small grain sizes and short diffusion distances in nanophase ceramics. Because of their ultrafine grain sizes, nanophase ceramics are easily formed, as has been clearly evident from the sample compaction process (Siegel and Hahn, 1987; Siegel *et al.*, 1988) and from early demonstrations via deformation (Karch *et al.*, 1987) as well. However, the degree to which nanophase ceramics are truly ductile is still only beginning to be understood.

Nanoindenter measurements on nanophase TiO_2 (Mayo *et al.*, 1990) and ZnO (Mayo *et al.*, 1992) have demonstrated that a dramatic increase of strain-rate sensitivity occurs with decreasing grain size, as shown in Fig. 18, that is remarkably similar for these two materials, even though they had as-consolidated densities of about 75% and 85%, respectively. Since this strong grain-size dependence was found for sets of samples in which the porosity was changing very little, it appears to be an intrinsic property of these

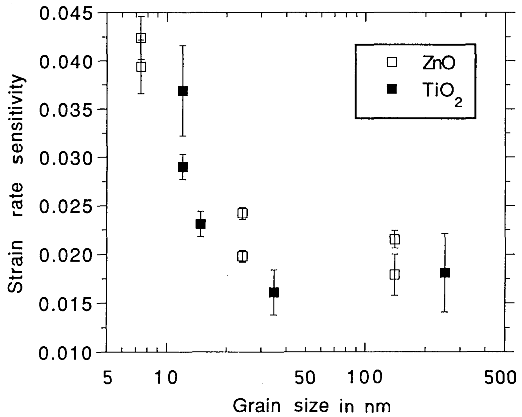

FIG. 18. Strain-rate sensitivity of nanophase TiO_2 (Mayo *et al.*, 1990) and ZnO (Mayo *et al.*, 1992) as a function of grain size. The strain-rate sensitivity was measured by a nanoindentation method, and the grain size was determined by dark-field TEM.

ultrafine-grained ceramics. The strain-rate sensitivity m is related to the stress σ applied to a sample and the resulting sample strain rate $d\epsilon/dt$ by the empirical expression $\sigma = K(d\epsilon/dt)^m$, where K is a constant. Values of m can vary from 0 to 1, the limits, respectively, for perfectly brittle and perfectly ductile behavior. The strain-rate sensitivity values at the smallest grain sizes yet investigated (12 nm in nanophase TiO_2 and 7 nm in ZnO) indicate not only ductile behavior of these nanophase ceramics at room temperature, but also a significant potential for increased ductility at even smaller grain sizes and at elevated temperatures. The maximum m values measured in these studies, about 0.04, are approximately one-quarter that for Pb at room temperature, for example. However, no superplasticity has yet been observed in nanophase materials at room temperature, which would be indicated by m values greater than about an order of magnitude higher than the maximum observed. Nevertheless, it already seems clear that in the future, at smaller grain sizes and/or at elevated temperatures, superplasticity of these materials will indeed be observed.

The possibilities for plastic forming of nanophase ceramics to near net shape appear to be well on their way to realization. Karch and Birringer (1990) have demonstrated that nanophase TiO_2 could be readily formed to a desired shape with excellent detail below 900 °C, and the fracture toughness was found to increase by a factor of 2 as well. The ability to deform fully dense nanophase TiO_2 extensively at elevated temperatures (ca. 800 °C) without cracking or fracture has also been demonstrated (Hahn *et al.*, 1990b; Guermazi *et al.*, 1991), as shown in Fig. 19. While these latter demonstrations have been accompanied by significant grain growth in the samples at the elevated temperatures employed, it can be expected that lower-temperature studies [below $(0.4\text{–}0.5)T_m$] in the future will also allow for near net-shape forming of nanophase ceramics with both their ultrafine grain sizes and their attendant properties retained.

The enhanced strain-rate sensitivity at room temperature found in the nanophase ceramics TiO_2 and ZnO (Mayo *et al.*, 1990, 1992) appears to result from increased grain-boundary sliding in this material, aided by the presence of porosity, ultrafine grain size, and probably rapid short-range diffusion as well. This behavior is therefore dominated by the presence of the numerous interfaces in these materials and the very short diffusion distances involved in effecting the necessary atomic healing of incipient cracks for grain-boundary sliding to progress at the strain rates utilized without fracturing the sample. Extrapolating from this apparently generic behavior, one can expect that grain-boundary sliding mechanisms, accompanied by short-range diffusion-assisted healing events, would be expected to dominate increasingly the deformation of a variety of nanophase materials. Enhanced forming and even superplasticity in a wide range of nanophase materials, including intermetallic compounds, ceramics, and semiconductors, might become a reality. Indeed, the ability to deform nanophase TiO_2 extensively at elevated temperatures near 800 °C without cracking or fracture, as indicated above, has already been clearly demonstrated, and similar results have been recently seen (Jain and Christman, 1994) in a mechanically attrited nanophase intermetallic compound (which is conventionally brittle), $Fe_{28}Al_2Cr$. Consequently, increased opportunities for the superplastic near-net-shape forming of many rather conventionally brittle and difficult-to-form materials could result.

The predominant mechanical property change resulting from reducing the grain sizes of nanophase metals, in contrast to the be-

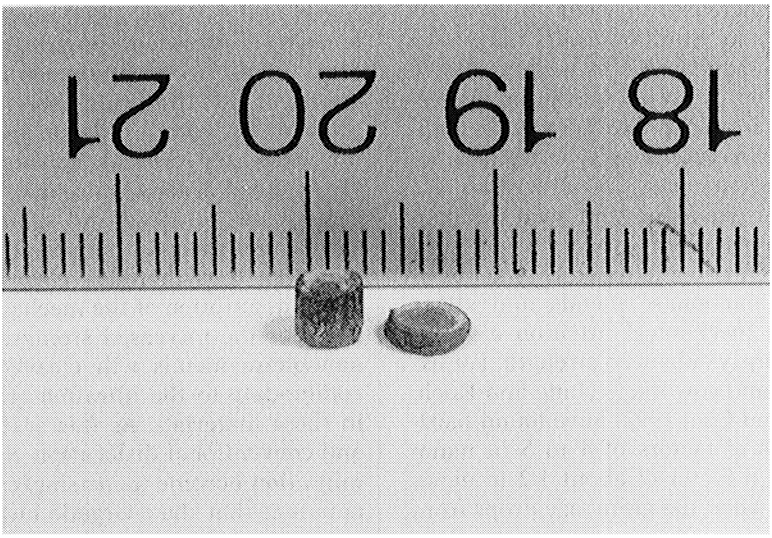

FIG. 19. Nanophase TiO_2 sample before and after compression at 810 °C for 15 h (Hahn *et al.*, 1990b). The total true strains were as high as 0.6, which represents deformation to a final thickness of less than 2 mm from an initial length of 3.5 mm at about 0.5 of its melting temperature (1830 °C). The grain size increased from 40–50 nm to about 1 mm. The small rule divisions are millimeters.

havior found for nanophase ceramics, is the significant increase in their strength. While the microhardness of as-consolidated nanophase oxides is reduced relative to their fully dense counterparts (Fig. 16), as a result of significant porosity in addition to their ultrafine grain sizes, the case for nanophase metals is quite different. Figures 20 and 21 show microhardness and stress-strain results for nanophase Pd and Cu compared with similar results for their coarser-grained counterparts (Nieman *et al.*, 1989, 1990, 1991). In their as-consolidated state, nanophase Pd samples with 5–10-nm grain sizes have been observed to exhibit up to about a 500% increase in hardness over coarser-grained (ca. 100 μm) samples (Nieman *et al.*, 1989), with

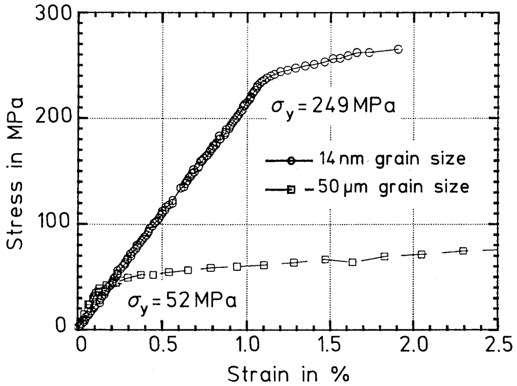

FIG. 20. Stress-strain curve for a nanophase (14 nm) Pd sample compared with that for a coarse-grained (50 μm) Pd sample (Nieman *et al.*, 1990). The strain rate $d\epsilon/dt = 2 \times 10^{-5}\ s^{-1}$.

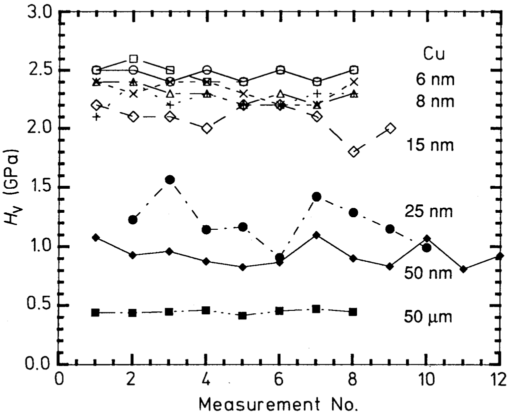

FIG. 21. Vickers microhardness (H_V) measurements at a number of positions across several nanophase Cu samples ranging in grain size from 6 to 50 nm, compared with similar measurements from an annealed conventional 50-μm grain-size Cu sample (Nieman *et al.*, 1991).

concomitant increases in yield stress σ_y, as shown in Fig. 20. Similar results have been observed in nanophase Cu as well, as shown in Fig. 21. The common strengthening behavior found in nanophase Pd and Cu indicates that this response is generic to nanophase metals, at least those with a fcc structure. The likelihood that such mechanical behavior is more broadly generic to nanophase metals is increased by the observations that nanophase metals and alloys produced via mechanical attrition also exhibit significantly enhanced strength. For example, Koch and coworkers (Jang and Koch, 1990; Koch and Cho, 1992) have found hardness increases of factors of 4 to 5 in nanophase Fe and a factor of about 1.2 in nanophase Nb_3Sn when the grain size drops from 100 to 6 nm.

In apparent contrast to these observations of enhanced strength with decreasing grain size, Chokshi *et al.* (1989) have reported a softening with decreasing grain size in the nanometer regime for cluster-assembled Cu and Pd samples, which was rationalized in terms of their expectation (Karch *et al.*, 1987; Birringer *et al.*, 1988) of room-temperature diffusional creep in these ultrafine-grained metals. A similar apparent softening was also reported (Chang *et al.*, 1991) for TiAl. The rapid atomic diffusion observed in nanophase materials (see Sec. 4.1), along with their nanometer grain sizes, has suggested that a large creep enhancement might result in nanophase materials, even at room-temperature. However, constant-stress creep measurements on nanophase Pd and Cu (Nieman *et al.*, 1990, 1991) show that the observed creep rates at room temperature are at least three orders of magnitude smaller than predicted on the basis of a Coble creep model, in which the creep rate varies as $D_b d^{-3}$, where D_b is the grain boundary diffusivity and d is the mean grain size. Such creep resistance will need to be explored further at elevated temperatures in these and other nanophase materials. However, it now appears that this apparent softening may only be a product of the manner in which the grain sizes were increased in these samples by annealing (Fougere *et al.*, 1992).

The increased strength observed in ultrafine-grained nanophase metals, although apparently analogous to conventional Hall–Petch strengthening (Hall, 1951; Petch, 1953; Siegel and Fougere, 1994) observed with decreasing grain size in coarser-grained metals, must result from fundamentally different mechanisms. The grain sizes in the nanophase metals considered here are smaller than the necessary critical bowing lengths for Frank–Read dislocation sources to operate at the stresses involved and smaller also than the normal spacings between dislocations in a pileup. It is therefore clear that an adequate description of the mechanisms responsible for the increased strength observed in nanophase metals will clearly need to accommodate to the ultrafine grain-size scale in these materials. As this scale is reduced, and conventional dislocation generation and migration become increasingly difficult, it is apparent that the energetic hierarchy of microscopic deformation mechanisms will become successively reached. Thus, easier mechanisms (such as dislocation generation from Frank–Read sources) will become frozen out at sufficiently small grain sizes and more energetically costly mechanisms will become necessary to effect deformation. Hence, grain confinement appears to be the dominant cause for the increased strength of nanophase metals. It can be generally expected that as nanophase grain sizes decrease and fall below the critical length scale for a given mechanism to operate, the associated property will be significantly changed. Clearly, much work in this area remains.

Nanocomposites consisting of metallic phases, ceramic and metallic phases (cermets), and ceramic phases in a variety of modulation dimensionalities also have considerably enhanced mechanical properties, including increased strength and fracture toughness. Prominent examples of these are ceramic nanocomposites in which nanoscale phases are combined with coarser-grained structures by means of mechanical milling (Niihara, 1991; Niihara *et al.*, 1993), and WC/Co nanophase cermets formed by a high-temperature thermochemical spray conversion method (Kear and McCandlish, 1993). This dramatically improved mechanical behavior, which should find use in a wide variety of wear and high-strength applications, results from a combination of the attributes of these nanocomposites including defect confinement, local and long-range strains, and increased complexity of fracture propagation.

4.3 Other Physical Properties

Rather limited research has been carried out so far on the other physical properties of cluster-assembled nanophase materials. However, there appear to be interesting prospects, based upon what little has been accomplished and on the expectations for confined systems of atoms in which the sizes of constituent domains fall below the critical length scales pertinent to a given property. A few examples of physical property changes that can occur in nanophase samples are presented here. Many more can be expected in the near future, as research begins to focus on these aspects of nanophase materials.

In magnetic multilayers, such as those formed by alternating layers of ferromagnetic Fe and Cr, the material can be nanostructured so that its electrical resistance is significantly decreased (by up to a factor of 2 depending upon the Cr-layer thickness) by the application of a magnetic field of 2 T (Baibach *et al.*, 1988). Such an effect, called giant magnetoresistance, occurs when the magnetic moments of the neighboring alternating layers (Fe/Cr) are arranged in an antiparallel fashion, so that application of the magnetic field overcomes the antiferromagnetic coupling and aligns the layers into a condition of parallel ferromagnetic ordering, strongly reducing the electron scattering in the system. It is important to control the nature of the interfaces between the layers, including both their structure and chemistry, in order to maximize this effect. Magnetoresistive materials are already being used in the magnetic recording industry as read heads, by virtue of their lower noise and improved signal-handling capabilities. It is now expected that nanophase materials will have a significant future impact in this area as new materials are developed with stable giant magnetoresistance at room temperature that can operate at low magnetic fields around 10^{-3} T. Recent discoveries that nanocomposites of magnetic cobalt clusters embedded in a nonmagnetic matrix of copper or silver (Berkowitz *et al.*, 1992; Xiao *et al.*, 1992) or magnetic platelike NiFe deposits embedded in silver (Hylton *et al.*, 1993) also exhibit giant magnetoresistance should hasten the advent of their useful application. The three-dimensional modulation and greater degrees of freedom in creating nanophase composite materials should allow for the engineering of these magnetic devices. Magnetic refrigeration applications of nanophase materials are also being explored (Shull *et al.*, 1993) with the anticipation that materials with usefully high efficiencies can be developed by nanostructuring.

The possibilities for other types of devices, such as chemical and physical sensors, resulting from nanophase materials are also apparent. Nanophase TiO_2 with a 12-nm grain size was doped at about the 1 at. % level with Pt diffused in from the surface. After annealing in air for 4 h at about 500 °C, the ac conductivity of the sample was measured as a function of temperature. The strongly nonlinear, and reversible, electrical response shown in Fig. 22, caused presumably by the Pt doping into the band gap of this wide–band-gap (3.2 eV) semiconductor, suggests that the rather easy impurity doping of nanophase electroceramics through rapid diffusion down their many grain boundaries and interconnected pores may lead to a wide range of interesting device applications. Also, the enhanced low-temperature sinterability of nanophase ceramics, without the need for any possibly contaminating additives, should help with device compatibility problems frequently encountered in such applications. However, much work remains to be done in this area.

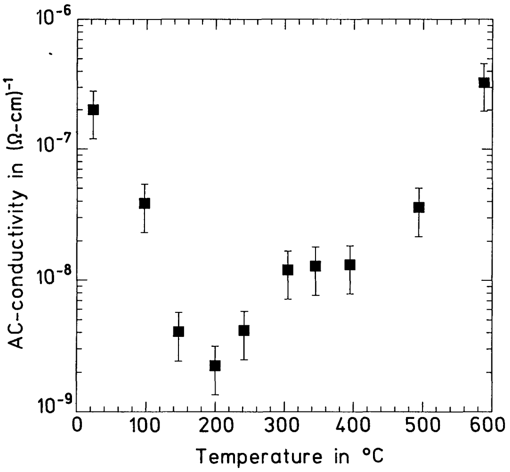

FIG. 22. The ac conductivity of Pt-doped nanophase TiO_2 as a function of temperature. The sample was preannealed in air at about 500 °C for 4 h prior to the conductivity measurements; the electrical response is reversible with temperature (Siegel, 1991).

The ability to control the porosity of nanophase ceramics will apparently lead to some rather useful optical applications. It has been demonstrated (Skandan *et al.*, 1991) that the oxidation step in the synthesis of cluster-consolidated nanophase Y_2O_3 can be controlled so that the as-consolidated material is effectively transparent. With such control, the resulting green-state porosity has a size distribution sufficiently small and narrow that light scattering is reduced and the material, although porous, is rather transparent as shown in Fig. 23. The ability thus to reduce the size of the porosity in such material to well below the wavelengths of visible or other radiation, coupled with the possibilities for efficient doping of these materials cited previously, should lead to a variety of interesting optical properties and related applications.

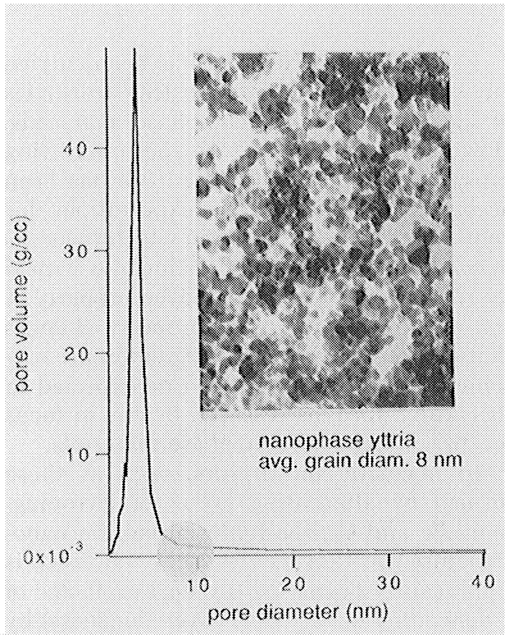

FIG. 23. Pore-size distribution in nanophase Y_2O_3 (the powder is shown in the inset TEM) consolidated at room temperature under 136 MPa pressure to a density of about 40% of the theoretical value. The transparent sample is shown on the axis at about its average grain diameter, 8 nm (courtesy of J.C. Parker, Nanophase Technologies Corporation).

5. FUTURE DIRECTIONS

It is very clear that in such a broad field as described in this article, we are just beginning to scratch the surface of the tremendous opportunities for synthesizing and applying nanophase materials. On the basis of the limited knowledge that has already been accumulated, the future appears to hold great promise for nanophase materials. The property changes that are found to occur when grain sizes are scaled down into the nanometer regime and interfaces take up significant volume fractions of the material can make significant differences in the range of use of a variety of materials. The cluster sizes utilized to date (down to about 5 nm) indicate that the high reactivities and short diffusion distances available in cluster-assembled materials can have profound effects upon the both the processing characteristics of these materials and their properties. These characteristics should be further enhanced as even smaller and more uniformly sized clusters become available in sufficiently large numbers to effect their assembly into usable and commercially viable materials.

The high diffusivities along their grain-boundary networks, with only a few atomic jumps separating grain interiors from grain boundaries, should make possible efficient impurity doping of these materials. Nanophase insulators and semiconductors, for example, could be easily doped with impurities at relatively low temperatures, allowing efficient introduction of impurity levels into their band gaps and control over their electrical and optical properties. Moreover, the ability to produce via cluster assembly fully dense ultrafine-grained nanophase ceramics with controlled or flaw-free microstructures that are readily formable and exhibit ductility should have a significant technological impact in a wide variety of applications. For example, near-net-shape forming of nanophase ceramic parts with complex and ultrafine detail would seem to be possible. Subsequent controlled grain growth could then be used to alter the grain-size–dependent properties of these ceramics.

Research on nanophase materials is now rapidly expanding, but much work still remains to be done. Further research on the synthesis of a broader range of nanophase materials, encompassing metals, alloys, ceramics, semiconductors, and composites, will have to be carried out in order to arrive at a fuller appreciation of just how broad an impact nanophase materials will have on materials technology. Investigations of their structure that will need to accompany such research will begin to elucidate the interplay between the effects of spatial confinement and large numbers of interfaces on the various chemical and physical properties of these new materials. A knowledge of the variation of such properties with the detailed structures and the synthesis and processing parameters of a variety of nanophase materials should eventually lead to an understanding of these new materials and consequently to the realization of their full impact on applied physics, materials science, and technology.

6. ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy, BES-Materials Sciences, under Contract W-31-109-Eng-38. The author thanks his many collaborators without whose efforts and contributions this work would not have been possible.

Works Cited

Aksay, I. A., McVay, G. L., Ulrich, D. R. (Eds.) (1989), *Processing Science of Advanced Ceramics*, Materials Research Society Symposium Proceedings No. 155, Pittsburgh: MRS.

Aksay, I. A., Baer, E., Sarikaya, M., Tirrell, D. A. (Eds.) (1992), *Hierarchically Structured Materials*, Materials Research Society Symposium Proceedings No. 255, Pittsburgh: MRS.

Andres, R. P., Averback, R. S., Brown, W. L., Brus, L. E., Goddard, W. A., III, Kaldor, A., Louie, S. G., Moskovits, M., Peercy, P. S., Riley, S. J., Siegel, R. W., Spaepen, F., Wang, Y. (1989), *J. Mater. Res.* **4**, 704–736.

Averback, R. S., Hahn, H., Höfler, H. J., Logas, J. L., Chen, T. C. (1989), in: B. M. DeKoven, A. J. Gellman, R. Rosenberg (Eds.), *Interfaces between Polymers, Metals, and Ceramics*, Materials Research Society Symposium Proceedings No. 153, Pittsburgh: MRS, pp. 3–12.

Averback, R. S., Nelson, D. L., Bernholc, J. (Eds.) (1991), *Clusters and Cluster-Assembled Materials*, Materials Research Society Symposium Proceedings No. 206, Pittsburgh: MRS.

Baibach, M. N., Broto, J. M., Fert, A., Nguyen, F. D. V., Petroff, F., Etienne, P., Creuzet, G., Friederich, A., Chazelas, J. (1988), *Phys. Rev. Lett.* **61**, 2472–2475.

Beck, D. D., Siegel, R. W. (1992), *J. Mater. Res.* **7**, 2840–2845.

Berkowitz, A. E., Mitchell, J. R., Carey, M. J., Young, A. P., Zhang, S., Spada, F. E., Parker. F. T., Hutten, A., Thomas, G. (1992), *Phys. Rev. Lett.* **68**, 3745–3748.

Berry, R. S., Jellinek, J., Natanson, G. (1984), *Phys. Rev. A.* **30**, 919–931.

Birringer, R., Gleiter, H., Klein, H.-P., Marquardt, P. (1984), *Phys. Lett. A* **102**, 365–369.

Birringer, R., Herr, U., Gleiter, H. (1986), *Suppl. Trans. Jpn. Inst. Met.* **27**, 43–52.

Birringer, R., Gleiter, H. (1988), in: R. W. Cahn (Ed.), *Encyclopedia of Materials Science and Engineering*, Suppl. Vol. 1, Oxford: Pergamon, pp. 339–349.

Birringer, R., Hahn, H., Höfler, H., Karch, J., Gleiter, H. (1988), *Defect Diffusion Forum* **59**, 17–31.

Blander, M., Katz, J. L. (1967), *Geochim. Cosmochim. Acta.* **31**, 1025–1034.

Blander, M., Abdel-Gawad, M. (1969), *Geochim. Cosmochim. Acta* **33**, 701–716.

Buffat, Ph., Borel, J.-P. (1976), *Phys. Rev. A* **13**, 2287–2298.

Burggraaf, A. J., Keizer, K., van Hassel, B. A. (1989), *Solid State Ionics* **32/33**, 771–782.

Cammarata, R., Sieradzki, K. (1989), *Phys. Rev. Lett.* **62**, 2005–2008.

Chang, H., Höfler, H. J., Altstetter, C. J., Averback, R. S. (1991), *Scripta Metall. Mater.* **25**, 1161–1166.

Chokshi, A. H., Rosen, A., Karch, J. Gleiter, H. (1989), *Scripta Metall.* **23**, 1679–1683.

Eastman, J. A., Liao, Y. X., Narayanasamy, A., Siegel, R. W. (1989), in: I. A. Aksay, G. L. McVay, D. R. Ulrich (Eds.), *Processing Science of Advanced Ceramics*, Materials Research Society Symposium Proceedings No. 155, Pittsburgh: MRS, pp. 255–266.

Eckert, J., Holzer, J. C., Krill, C. E., III, Johnson, W. L. (1992), *J. Mater. Res.* **7**, 1751–1761; and in: W. A. T. Clark, C. L. Briant, U. Dahmen (Eds.), *Structure and Properties of Interfaces in Materials*, Materials Research Society Symposium Proceedings No. 238, Pittsburgh: MRS, pp. 745–750.

Epperson, J. E., Siegel, R. W., White, J. W., Klippert, T. E., Narayanasamy, A., Eastman, J. A., Trouw, F. (1989), in: L. E. McCandlish, D. E. Polk, R. W. Siegel, B. H. Kear (Eds.), *Multicomponent Ultrafine Microstructures*, Materials Research Society Symposium Proceedings No. 132, Pittsburgh: MRS, pp. 15–20.

Epperson, J. E., Siegel, R. W., White, J. W., Eastman, J. A., Liao, Y. X., Narayanasamy, A. (1990), in: S. M. Shapiro, S. C. Moss, J. D. Jorgensen (Eds.), *Neutron Scattering for Materials Science*, Materials Research Society Symposium Proceedings No. 166, Pittsburgh: MRS, pp. 87–92.

Fecht, H. J., Hellstern, E., Fu, Z., Johnson, W. L. (1989), *Adv. Powder Metall.* **1–2**, 111–122.

Fougere, G. E., Weertman, J. R., Kim, S., Siegel, R. W. (1992), *Scripta Metall. Mater.* **26**, 1879–1883.

Ganapathi, S. K., Rigney, D. A. (1990), *Scripta Metall. Mater.* **24**, 1675–1678.

Gao, P., Gleiter, H. (1987), *Acta Metall.* **35**, 1571–1575.

Gleiter, H. (1981), in: N. Hansen, A. Horsewell, T. Leffers, H. Lilholt (Eds.), *Deformation of Polycrystals: Mechanisms and Microstructures*, Roskilde: Risø National Laboratory, pp. 15–21.

Gleiter, H. (1989), *Prog. Mater. Sci.* **33**, 223–315.

Goldstein, A. N., Echer, C. M., Alivisatos, A. P. (1992), *Science* **256**, 1425–1427.

Granqvist, C. G., Buhrman, R. A. (1976), *J. Appl. Phys.* **47**, 2200–2219.

Gregg, S. J., Sing, K. S. W. (1982), *Adsorption Surface Area and Porosity*, New York: Academic Press.

Grossman, L. (1972), *Geochim. Cosmochim. Acta* **36**, 597–619.

Guermazi, M. Höfler, H. J., Hahn, H., Averback, R. S. (1991), *J. Am. Ceram. Soc.* **74**, 2672–2674.

Hadjipanayis, G. C., Siegel, R. W. (Eds.) (1994), *Nanophase Materials: Synthesis–Properties–Applications*, Dordrecht: Kluwer.

Hahn, H., Höfler, H., Averback, R. S. (1989), *Defect Diffusion Forum* **66–69**, 549–554.

Hahn, H., Logas, J., Averback, R. S. (1990a), *J. Mater. Res.* **5**, 609–614.

Hahn, H. Logas, J., Höfler, H. J., Kurath, P., Averback, R. S. (1990), in: M. J. Mayo, M. Kobayashi, J. Wadsworth (Eds.), *Superplasticity in Metals, Ceramics, and Intermetallics*, Materials Research Society Symposium Proceedings No. 196, pp. 71–76.

Hall, E. O. (1951), *Proc. Phys. Soc. London B* **64**, 747–753.

Han, M. Y., Whetten, R. L. (1988), *Phys. Rev. Lett.* **61**, 1190–1193.

Haubold, T., Birringer, R., Lengeler, B., Gleiter, H. (1988), *J. Less-Common Metals* **145**, 557–563.

Haubold, T., Birringer, R., Lengeler, B., Gleiter, H. (1989), *Phys. Lett. A* **135**, 461–466.

Hellstern, E., Fecht, H. J., Fu, Z., Johnson, W. L. (1989). *J. Appl. Phys.* **65**, 305–310.

Herr, U., Jing, J., Birringer, R., Gonser, U., Gleiter, H. (1987), *Appl. Phys. Lett.* **50**, 472–474.

Heuer, A. H., Fink, D. J., Laraia, V. J., Arias, J.-L., Calvert, P. D., Kendall, K., Messing, G. L., Blackwell, J., Rieke, P. C., Thompson, D. H., Wheeler, A. P., Veis, A., Caplan, A. L. (1992), *Science* **255**, 1098–1105.

Höfler, H. J., Averback, R. S. (1990), *Scripta Metall. Mater.* **24**, 2401–2406.

Horváth, J., Birringer, R., Gleiter, H. (1987), *Solid State Commun.* **62**, 319–322.

Horváth, J. (1989), *Defect Diffusion Forum* **66–69**, 207–228.

Hylton, T. L., Coffey, K. R., Parker, M. A., Howard, J. K. (1993), *Science* **261**, 1021–1024.

Jain, M., Christman, T. (1994), *Acta Metall. Mater.* **42**, 1901–1911.

Jang, J. S. C., Koch, C. C. (1990), *Scripta Metall. Mater.* **24**, 1599–1604.

Jorra, E., Franz, H., Peisl, J., Wallner, G., Petry, W., Birringer, R., Gleiter, H., Haubold, T. (1989), *Philos. Mag. B* **60**, 159–168.

Karch, J., Birringer, R., Gleiter, H. (1987), *Nature* **330**, 556–558.

Karch, J., Birringer, R. (1990), *Ceram. Int.* **16**, 291–294.

Kear, B. H., Cross, L. E., Keem, J. E., Siegel, R. W., Spaepen, F., Taylor, K. C., Thomas, E. L., Tu, K.-N. (1989), *Research Opportunities for Materials with Ultrafine Microstructures*, Vol. NMAB-454, Washington, DC: National Academy.

Kear, B. H., McCandlish, L. E. (1993), *Nanostruct. Mater.* **3**, 19–30.

Kimoto, K., Kamiya, Y., Nonoyama, M., Uyeda, R. (1963), *Jpn. J. Appl. Phys.* **2**, 702–713.

Koch, C. C., Jang, J. S. C., Gross, S. S. (1989), *J. Mater. Res.* **4**, 557–564.

Koch, C. C., Cho, Y. S. (1992), *Nanostruct. Mater.* **1**, 207–212.

Koch, C. C. (1993), *Nanostruct. Mater.* **2**, 109–129.

Komarneni, S., Parker, J. C., Thomas, G. J. (Eds.) (1993), *Nanophase and Nanocomposite Materials*, Materials Research Society Symposium Proceedings No. 286, Pittsburgh: MRS.

Li, Z., Hahn, H., Siegel, R. W. (1988a), *Mater. Lett.* **6**, 342–346.

Li, Z., Ramasamy, S., Hahn, H., Siegel, R. W. (1988b), *Mater. Lett.* **6**, 195–201.

Liu, C.-L., Adams, J. B., Siegel, R. W. (1994), *Nanostruct. Mater.* **4**, 265–274.

Luton, M. J., Janath, C. S., Disko, M. M., Matras, S., Vallone, J. (1989), in: L. E. McCandlish, D. E. Polk, R. W. Siegel, B. H. Kear (Eds.), *Multicomponent Ultrafine Microstructures*, Materials Research Society Symposium Proceedings No. 132, Pittsburgh: MRS, pp. 79–86.

Mayo, M. J., Siegel, R. W., Narayanasamy, A., Nix, W. D. (1990), *J. Mater. Res.* **5**, 1073–1082.

Mayo, M. J., Siegel, R. W., Liao, Y. X., Nix, W. D. (1992), *J. Mater. Res.* **7**, 973–979.

Melendres, C. A., Narayanasamy, A., Maroni, V. A., Siegel, R. W. (1989), *J. Mater. Res.* **4**, 1246–1250.

Milligan, W. W., Hackney, S. A., Ke, M., Aifantis, E. C. (1993), *Nanostruct. Mater.* **2**, 267–276.

Morris, D. G., Morris, M. A. (1991), *Acta Metall. Mater.* **39**, 1763–1770.

Mütschele, T., Kirchheim, R. (1987), *Scripta Metall.* **21**, 135–140; *ibid.*, 1101–1104.

Nieman, G. W., Weertman, J. R., Siegel, R. W. (1989), *Scripta Metall.* **23**, 2013–2018.

Nieman, G. W., Weertman, J. R., Siegel, R. W. (1990), *Scripta Metall. Mater.* **24**, 145–150.

Nieman, G. W., Weertman, J. R., Siegel, R. W. (1991), *J. Mater. Res.* **6**, 1012–1027.

Niihara, K. (1991), *J. Ceram. Soc. Jpn.* **99**, 945–952.

Niihara, K., Nakahira, A., Sekino, T. (1993), in: S. Komarneni, J. C. Parker, G. J. Thomas (Eds.), *Nanophase and Nanocomposite Materials*, Materials Research Society Symposium Proceedings No. 286, Pittsburgh: MRS, pp. 405–412.

Owen, D. M., Chokshi, A. H. (1993), *Nanostruct. Mater.* **2**, 181–187.

Parker, J. C., Siegel, R. W. (1990a), *J. Mater. Res.* **5**, 1246–1252.

Parker, J. C., Siegel, R. W. (1990b), *Appl. Phys. Lett.* **57**, 943–945.

Petch, N. J. (1953), *J. Iron Steel Inst.* **174**, 25–28.

Pfund, A. H. (1930), *Rev. Sci. Instrum.* **1**, 397–399.

Rieke, P. C., Calvert, P. D., Alper, M. (Eds.) (1990), *Materials Synthesis Utilizing Biological Processes*, Materials Research Society Symposium Proceedings No. 174, Pittsburgh: MRS.

Rivier, N. (1992), in: P. Jena, S. M. Khanna, B. K. Rao (Eds.), *Physics and Chemistry of Finite Systems: from Clusters to Crystals*, Dordrecht: Kluwer, pp. 189–198.

Roy, R., Komarneni, S., Yarbrough, W. (1988), in: J. MacKenzie, D. Ulrich (Eds.), *Ultrastructure Processing of Advanced Ceramics*, New York: Wiley Interscience, pp. 571–588.

Sanders, P. G., Weertman, J. R., Barker, J. G., Siegel, R. W. (1993), *Scripta Metall. Mater.* **29**, 91–96.

Sattler, K., Raina, G., Ge, M., Venkateswaran, N., Xhie, J., Liao, Y. X., Siegel, R. W. (1994), *J. Appl. Phys.* **76**, 546–551.

Schaefer, H. E., Würschum, R., Scheytt, M., Birringer, R., Gleiter, H. (1987), *Mater. Sci. Forum* **15–18**, 955–960.

Schaefer, H. E., Würschum, R., Birringer, R., Gleiter, H. (1988), *Phys. Rev. B* **38**, 9545–9554.

Schumacher, S., Birringer, R., Straub, R., Gleiter, H. (1989), *Acta Metall.* **37**, 2485–2488.

Shull, R. D., McMichael, R. D., Ritter, J. J., Bennett, L. H. (1993), in: S. Komarneni, J. C. Parker, G. J. Thomas (Eds.), *Nanophase and Nanocomposite Materials*, Materials Research Society Symposium Proceedings No. 286, Pittsburgh: MRS, pp. 449–453.

Siegel, R. W., Hahn, H. (1987), in: M Yussouff (Ed.), *Current Trends in the Physics of Materials*, Singapore: World Scientific, pp. 403–419.

Siegel, R. W., Ramasamy, S., Hahn, H., Li, Z., Lu, T., Gronsky, R. (1988), *J. Mater. Res.* **3**, 1367–1372.

Siegel, R. W., Eastman, J. A. (1989), in: L. E. McCandlish, D. E. Polk, R. W. Siegel, B. H. Kear (Eds.), *Multicomponent Ultrafine Microstructures*, Materials Research Society Symposium Proceedings No. 132, Pittsburgh: MRS, pp. 3–14.

Siegel, R. W. (1990), in: M. J. Mayo, M. Kobayashi, J. Wadsworth (Eds.), *Superplasticity in Metals, Ceramics, and Intermetallics*, Materials Research Society Symposium Proceedings No. 196, pp. 59–70.

Siegel, R. W. (1991), in: R. W. Cahn (Ed.), *Materials Science and Technology*, Vol. 15: *Processing of Metals and Alloys*, Weinheim: VCH, pp. 583–614.

Siegel, R. W. (1992), in: D. Wolf, S. Yip (Eds.), *Materials Interfaces: Atomic-Level Structure and Properties*, London: Chapman and Hall, pp. 431–460.

Siegel, R. W., Thomas, G. J. (1992), *Ultramicroscopy* **40**, 376–384.

Siegel, R. W. (1993), in: M. Nastasi, D. M. Parkin, H. Gleiter (Eds.), *Mechanical Properties and Deformation Behavior of Materials Having Ultra-Fine Microstructures*, Dordrecht: Kluwer, pp. 509–538.

Siegel, R. W., Fougere, G. E. (1994), in: G. J. Hadjipanayis, R. W. Siegel (Eds.), *Nanophase Materials: Synthesis–Properties–Applications*, Dordrecht: Kluwer, pp. 233–261.

Skandan, G., Hahn, H., Parker, J. C. (1991), *Scripta Metall. Mater.* **25**, 2389–2393.

Skandan, G., Foster, C. M., Frase, H., Ali, M. N., Parker, J. C., Hahn, H. (1992), *Nanostruct. Mater.* **1**, 313–322.

Suits, B. H., Siegel, R. W., Liao, Y. X. (1993), *Nanostruct. Mater.* **2**, 597–602.

Suits, B. H., Meng, M., Siegel, R. W., Liao, Y. X. (1994), *J. Mater. Res.* **9**, 336–342.

Thölén, A. R. (1979), *Acta Metall.* **27**, 1765–1778.

Thomas, G. J., Siegel, R. W., Eastman, J. A. (1989), in: B. M. DeKoven, A. J. Gellman, R. Rosenberg (Eds.), *Interfaces between Polymers, Metals, and Ceramics*, Materials Research Society Symposium Proceedings No. 153, Pittsburgh: MRS, pp. 13–20.

Thomas, G. J., Siegel, R. W., Eastman, J. A. (1990), *Scripta Metall. Mater.* **24**, 201–206.

Trudeau, M. L., Van Neste, A., Schulz, R. (1991), in: R. S. Averback, D. L. Nelson, J. Bernholc (Eds.), *Clusters and Cluster-Assembled Materials*, Materials Research Society Symposium Proceedings No. 206, Pittsburgh: MRS, pp. 487–492.

Tschöpe, A., Birringer, R. (1993), *Acta Metall. Mater.* **41**, 2791–2796.

Uyeda, R. (1991), *Prog. Mater. Sci.* **35**, 1–96.

Valiev, R. Z., Krasilnikov, N. A., Tsenev, N. K. (1991), *Mater. Sci. Eng. A* **137**, 35–40.

Valiev, R. Z. (1993), in: M. Nastasi, D. M. Parkin, H. Gleiter (Eds.), *Mechanical Properties and Deformation Behavior of Materials Having Ultra-Fine Microstructures*, Dordrecht: Kluwer, pp. 303–308.

Wagner, W., Averback, R. S., Hahn, H., Petry, W., Wiedenmann, A. (1991), *J. Mater. Res.* **6**, 2193–2198.

Whitesides, G. M., Mathias, J. P., Seto, C. T. (1991), *Science* **254**, 1312–1319.

Wolf, D., Lutsko, J. F. (1988), *Phys. Rev. Lett.* **60**, 1170–1173.

Wolf, D., Yip, S. (Eds.) (1992), *Materials Inter-*

faces: Atomic-Level Structure and Properties, London: Chapman and Hall.

Wunderlich, W., Ishida, Y., Maurer, R. (1990), *Scripta Metall. Mater.* **24**, 403–408.

Xiao, J. Q., Jiang, J. S., Chien, C. L. (1992), *Phys. Rev. Lett.* **68**, 3749–3752.

Zhu, X., Birringer, R., Herr, U., Gleiter, H. (1987), *Phys. Rev. B* **35**, 9085–9090.

Further Reading

Among the numerous references cited above, there are several comprehensive treatments of nanophase materials that can be usefully consulted. They are as follows:

Andres, R. P., Averback, R. S., Brown, W. L., Brus, L. E., Goddard, W. A., III, Kaldor, A., Louie, S. G., Moskovits, M., Peercy, P. S., Riley, S. J., Siegel, R. W., Spaepen, F., Wang, Y. (1989), *J. Mater. Res.* **4**, 704–736.

Gleiter, H. (1989), *Prog. Mater. Sci.* **33**, 223–315.

Hadjipanayis, G. C., Siegel, R. W. (Eds.) (1994), *Nanophase Materials: Synthesis-Properties-Applications*, Dordrecht: Kluwer.

Jena, P., Khanna, S. N., Rao, B. K. (Eds.) (1992), *Physics and Chemistry of Finite Systems: from Clusters to Crystals*, Dordrecht: Kluwer.

Kear, B. H., Cross, L. E., Keem, J. E., Siegel, R. W., Spaepen, F., Taylor, K. C., Thomas, E. L., Tu, K.-N. (1989), *Research Opportunities for Materials with Ultrafine Microstructures*, Vol. NMAB-454, Washington, DC: National Academy.

Nastasi, M., Parkin, D. M., Gleiter, H. (Eds.) (1993), *Mechanical Properties and Deformation Behavior of Materials Having Ultra-Fine Microstructures*, Dordrecht: Kluwer.

In addition, a series of symposia on the subject of nanophase and related materials has been sponsored by the Materials Research Society since 1988. The proceedings of these symposia are published in volumes 132 (1989), 206 (1991), 286 (1993), and 351 (1994) of the *MRS Symposium Proceedings* series. Current developments in the field can be followed in *Nanostructured Materials*, an Acta Metallurgica, Inc. international journal published by Pergamon Press since 1992. A series of International Conferences on Nanostructured Materials is also being held and the proceedings of the first (Cancun, 1992) and second (Stuttgart, 1994) of these are published in separate volumes (3 and 6, respectively) of *Nanostructured Materials*.

NATURAL GAS

Robert J. Finley, *Bureau of Economic Geology, The University of Texas at Austin, Austin, Texas, U.S.A.*

Introduction 201
1. **Definition of Natural Gas** 202
1.1 Origin of Natural Gas 202
1.2 Migration and Trapping of Natural Gas 203
1.3 Access to Natural Gas Resources 203
1.4 Overview of World Gas Consumption, Production, and Reserves 203
2. **The Natural Gas Resource Base** 204
2.1 World Natural Gas Resources 204
2.2 Natural Gas Resources in Canada 205
2.3 Natural Gas Resources in Mexico 206
2.4 Natural Gas Resources in the United States 206
2.4.1 United States Resource Estimates Through Time 207
3. **History of Natural Gas Resource Development** 207
3.1 United States Natural Gas Development 207
3.2 Canadian Natural Gas Development 208
3.3 Mexican Natural Gas Development 209
4. **Technologies for Natural Gas Development** 209
4.1 Delineation of Reservoir Types 209
4.2 Technologies for Conventional Reservoir Development 210
4.2.1 Exploration Technologies 210
4.2.1.1 Three-Dimensional Seismic Imagery 210
4.2.1.2 Basin Analysis and Sequence Stratigraphy 210
4.2.2 Development Technologies 210
4.2.2.1 Reservoir Characterization Technology 210
4.2.2.2 Geophysical Detection Technologies 211
4.2.2.3 Horizontal Drilling 211
4.3 Technologies for Unconventional Reservoir Development 211
4.3.1 Advanced Extraction Technologies 212
4.3.1.1 Hydraulic Fracturing 212
4.3.1.2 Cavity Completions 212
4.4 Reserve Growth in Mature Fields 212
5. **Gas Gathering, Processing, and Distribution** 213
6. **Utilization and Applications** 214
6.1 Industrial Natural Gas Use 214
6.2 Residential Natural Gas Use 215
6.3 Commercial Natural Gas Use 215
6.4 Electric Utility Use of Natural Gas 216
6.5 Transportation Use of Natural Gas 216
7. **World Natural Gas Trade** 216
7.1 Liquefied Natural Gas Trade 216
8. **Summary and Future Outlook** 217
Glossary 218
Works Cited 218
Further Reading 219

INTRODUCTION

In 1992, natural gas was the third leading primary source of energy in the world, following crude oil and coal, accounting for 71.2 quadrillion Btu (quads) (75.2×10^{12} MJ) of production, or 22.8% of a total of 311.8 quads (329.2×10^{12} MJ) (British Petroleum Com-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

pany, 1993). One quad is approximately the energy equivalent of 28.3×10^9 m^3 of natural gas. The former Soviet Union was the major producer of natural gas in 1990, accounting for 39% of production, followed by the United States (24%) and Canada (5%). Although some natural gas is shipped internationally through pipelines and a small amount is shipped as liquefied natural gas (LNG), most natural gas is consumed in or near the country of origin. Thus, the three leading producers are also the leading consumers.

The prominence of natural gas will likely increase with increasing concern for the environmental consequences of fossil fuel use. Natural gas combustion results in minimal atmospheric emissions of sulfur oxides and particulates and far lower emissions of carbon monoxide, carbon dioxide, and nitrogen oxides than either coal or oil. In stationary (power plants, boilers, etc.) uses, natural gas emits roughly half as much carbon dioxide per unit of energy as coal and 30% less than oil.

Concerns about the natural gas resource base, especially in the United States, that existed in the late 1970s have largely dissipated. Such concerns led to restrictions on natural gas use in the United States that have now been lifted. Assessments of the resource base (proved reserves plus resources) in the United States (lower 48 states) have increased substantially from 1980 through 1992 to a level of 36.66×10^{12} m^3 (National Petroleum Council, 1992). Worldwide, proved reserves as of 1 January 1993 were 137.9×10^{12} m^3, of which 4.53×10^{12} m^3 were U.S. proved reserves (lower 48 states). With substantial reserves and even greater volumes of resources remaining to be developed, natural gas usage will increase as energy demand grows because of its low rate of combustion emissions, recognized availability, and improved technology for production, even from the geologically complex reservoirs that hold large natural gas resources.

1. DEFINITION OF NATURAL GAS

Natural gas is a naturally occurring mixture of hydrocarbon and nonhydrocarbon gases occurring in geologic formations beneath the earth's surface. The hydrocarbon gases are of very low molecular weight with the principal combustible component being methane, CH_4. A typical natural gas contains 70% to nearly 100% methane, 1% to 10% ethane, and smaller amounts of higher molecular weight hydrocarbons. Nonhydrocarbon components may include carbon dioxide, nitrogen, hydrogen sulfide, helium, water vapor, and hydrogen. Natural gas is referred to as sour if sufficient sulfur compounds are present to require removal for use of the gas. If referred to as wet, substantial and extractable quantities of C_2 to C_5 hydrocarbons, commonly referred to as natural gas liquids, are present. Most natural gas requires at least the removal of water vapor to bring it up to pipeline quality (American Gas Association, 1990; Kvenvolden, 1988).

1.1 Origin of Natural Gas

Natural gas forms thermogenically from the alteration and degradation of organic matter trapped within fine-grained sedimentary rocks. Rocks such as shales with less than 0.5% organic carbon content are poor source rocks, those with 1.0% to 2.0% are good, and those with over 4.0% organic carbon are very good. The type of original organic matter, temperature history of the source rocks, and geologic time during which source materials were buried determine the volume and chemical content of the resulting hydrocarbons forming natural gas. The same controls apply to the formation of oil. Natural gas may also form microbially through carbon dioxide reduction to produce methane; the resulting gas is quite dry with only very small amounts of hydrocarbons heavier than methane (Waples, 1985). A general progression of hydrocarbon generation with increasing time and temperature is from biogenic gas to oil, light oil, wet gas, and finally dry gas. Not all stages are present in the history of any one natural gas accumulation. The type of hydrocarbon generated is highly dependent on whether source materials are oil-prone or gas-prone. Most natural gas deposits in the earth contain thermogenic methane generated at the high-temperature end of this progression (Hunt, 1979). Another theory of the origin of natural gas holds that natural gas may be formed within the earth's crust exclusive of an organic source (Gold and Soter, 1980; Gold, 1987), but this theory is not widely held by petroleum geologists.

1.2 Migration and Trapping of Natural Gas

Once natural gas is generated, it may be expelled from the organic-rich source rock, such as a low-permeability shale, into a more permeable rock, such as a sandstone, and migrate in several stages until trapped by a low-permeability barrier. Such a barrier may be another shale or a chemical sediment such as salt or gypsum. The geometry of these barriers is quite varied, and trapping of migrated natural gas often depends on the structural configuration of the rock in addition to the rock type. The result is a natural gas reservoir from which natural gas can be technically and economically recovered. Migration is driven by the lower density and resultant buoyancy of natural gas relative to oil and formation waters. A reservoir rock, such as a sandstone, with permeability of >0.1 millidarcy from which gas can be readily extracted forms a *conventional* natural gas reservoir.

Some organic-rich source rocks such as coal and highly organic black shales retain recoverable gas that may be produced by tapping into the natural fracture systems within these source rocks. As gas (and, as in the case of coal, formation water) is withdrawn from the fracture system, reservoir pressure is reduced and additional gas desorbs from the rock matrix and can be produced. Such reservoirs require special treatments, such as the creation of a hydraulic fracture extending away from the wellbore or a cavity around the wellbore at the level of the coalbed, to produce gas economically. In some cases, sandstone reservoirs have such low permeability (<0.1 millidarcy) that they do not produce gas without special well treatments to dissolve minerals that have filled the reservoir pore space or to fracture the sandstone artificially near the well. Such low-permeability (tight) sandstones and potentially productive coals and shales are termed *unconventional* reservoirs. As conventional reservoirs are depleted, especially in maturely explored and developed regions such as onshore in the lower 48 states of the United States (compared to offshore production in the U.S. Gulf of Mexico), unconventional reservoirs are becoming increasingly important as natural gas sources.

1.3 Access to Natural Gas Resources

Natural gas resources will be exploited by continued exploration for undiscovered accumulations, by advanced recovery in known conventional reservoirs where the efficiency of recovery can be improved as more of a given field is drilled, and through development of unconventional reservoirs. The location of most unconventional resources is generally known for the United States, but only about 35% of the in-place resource may be exploited with current technology and well spacing. Economic extraction of this gas is highly dependent on continuing technological advancement. The same is true for advanced recovery of conventional resources where, in geologically complex reservoirs, only about 65% of the in-place gas may typically be reached. Additional gas supplies for U.S. and other major markets are available through pipeline imports and imported LNG.

Because transoceanic trade in LNG is limited by costs to liquefy, ship, and regasify, future gas supplies will be primarily dependent on exploration and production technologies disseminated on a worldwide basis to develop gas resources marketable through pipelines. By the year 2000, world LNG trade is expected to represent only about 4% of production, with supplies close to markets retaining a significant cost advantage (MacDougall and Linder, 1992).

1.4 Overview of World Gas Consumption, Production, and Reserves

World consumption of natural gas has grown steadily over the last two decades from about 1.08×10^{12} m^3 in 1970 to about 2.12×10^{12} m^3 in 1990. Total world consumption stood at 2.13×10^{12} m^3 in 1992. Between 1970 and 1990, North American consumption declined, while that of European countries not members of the Organization for Economic Cooperation and Development (OECD), notably the former Soviet Union, increased substantially. The former Soviet Union saw an increase in consumption of 92%, from 394×10^9 m^3 to 758×10^9 m^3 between 1980 and 1990 alone (British Petroleum Company, 1991, 1993).

Among the ten leading consuming nations, Germany, Japan, Italy, and France have substantial deficits between domestic production and consumption that must be made up by imports (Table 1). Among these ten nations, the former Soviet Union and the Netherlands exported natural gas to other parts of Europe, and Canada exported gas to the United States. Other exporting nations contributing to European supply are Algeria and Norway. Germany, the European nation with the largest consumption of imported natural gas, received 22.4 × 10^9 m^3 of imports from the former Soviet Union and 19.6 × 10^9 m^3 of imports from the Netherlands in 1990 (Energy Information Administration, 1993a). Gazprom, the Russian state-owned natural gas producer, is constructing new pipelines to increase its export capability to Western European markets. Gazprom aims to increase its total annual production to 739 × 10^9 m^3 in 1998 from 653 × 10^9 m^3 in 1992. In 1992, Gazprom delivered 113 × 10^9 m^3 to Western Europe and the same amount to the newly independent Baltic states (Wood, 1993).

Japan receives substantial volumes of LNG from Indonesia (24.4 × 10^9 m^3 in 1992) and from the Middle Eastern countries of United Arab Emirates, Iran, and Iraq (3.32 × 10^9 m^3 in 1992), with the balance of supply from other sources (Energy Information Administration, 1993a).

Consuming countries with significant reserves are led by the former Soviet Union (55.0 × 10^{12} m^3) and in particular Russia with 47.6 × 10^{12} m^3 of reserves as of the end of 1992. Saudi Arabia, Canada, and the Netherlands have significant reserve levels in relation to domestic consumption (Table 1). Iran has the most significant reserves of natural gas in the Middle East (19.8 × 10^{12} m^3, 1992) followed by Qatar (6.4 × 10^{12} m^3, 1992) and Abu Dhabi (5.3 × 10^{12} m^3, 1992). Other significant reserves are held by Nigeria (3.4 × 10^{12} m^3), Venezuela (3.6 × 10^{12} m^3), and Indonesia (1.8 × 10^{12} m^3), all as of the end of 1992 (British Petroleum Company, 1993). All of these countries already are, or plan to become, involved in international natural gas trade, either through pipeline exports or through LNG export facilities.

2. THE NATURAL GAS RESOURCE BASE

2.1 World Natural Gas Resources

Increased exploration and development of the resource base has increased world proved reserves of natural gas from 46 × 10^{12} m^3 in 1970 to 75 × 10^{12} m^3 in 1980 and 137.9 × 10^{12} m^3 in 1990 (British Petroleum Company, 1993). The former Soviet Union holds the largest proved reserves with 40% of the world total (Table 2). The former Soviet Union is also the world's leading natural gas producer, with Russia accounting for 79% of that production in 1991 (Anonymous, 1992a). Among Middle Eastern countries, Iran holds 19.8 × 10^{12} m^3 or about 14% of the world total in a region with 31.3% of the world total (Table 2). Among the North Sea producing countries, Norway and the Netherlands each

Table 1. Consumption (1991), production (1991), and reserves (1992) of natural gas for the 10 leading consuming countries (from British Petroleum Company, 1993; Energy Information Administration, 1993a).

Area	Consumption (10^9 m^3)	Production (10^9 m^3)	Reserves (10^{12} m^3)
Former Soviet Union	711.7	812.2	55.0
United States	541.4	504.3	4.7
Canada	70.0	112.5	2.7
Germany	105.2	25.8	0.3
United Kingdom	61.7	54.8	0.5
Japan	54.0	2.0	[a]
Italy	51.4	17.9	[b]
Netherlands	43.3	81.8	1.9
France	37.8	3.0	[b]
Saudi Arabia	33.0	33.2	5.2

[a] Less than 0.05.
[b] Less than 0.6 for OECD Europe other than Germany, Netherlands, Norway, and United Kingdom.

Table 2. World proved reserves of natural gas, 1 January 1993 (from Anonymous, 1992a; British Petroleum Company, 1993).

Area	10^{12} m^3	10^{12} ft^3	Share of total (%)	R/P ratio[a]
North America	7.4	262.8	5.4	12.0
USA	4.7	167.1	3.4	9.4
Canada	2.7	95.7	2.0	23.5
Latin America	7.3	259.5	5.4	75.8
OECD Europe	5.3	191.8	3.8	27.6
Former Soviet Union	55.0	1942.3	39.8	67.8
Non-OECD Europe	0.6	21.0	0.4	15.7
Middle East	43.1	1520.1	31.0	[b]
Africa	9.7	346.9	7.1	[b]
Asia and Australasia	9.5	341.0	6.9	52.5
World total	137.9	4885.4		64.8

[a] Reserve/production ratio at year end 1992.
[b] Over 100 years.

hold about 1.9 × 10^{12} m^3 of reserves, totaling 2.8% of world reserves. Norway has the potential to become a future exporter of LNG.

Assessments of world resources beyond proved reserves are not well defined outside of North America. The Canadian Energy Research Institute (CERI) calculated 221.9 × 10^{12} m^3 of potential reserves (proved undeveloped and undiscovered potential) outside of North America (MacDougall and Linder, 1992), yet this number seems conservative compared to the volume of proved reserves outside North America. Of the resources assessed by CERI, 37% are in the former Soviet Union and 33% are in the Middle East (Table 3). Major former Soviet Union fields are now in production in the Yamal Peninsula area of western Siberia. These fields contain about as large a volume in possible reserves as in already explored reserves (defined as proved plus probable), thereby substantially increasing reserves by additional development in these known fields alone, without an increment of undiscovered potential in the region. Yamal fields discovered to date contain approximately 16.6 × 10^{12} m^3 of gas and additional potential for undiscovered resources is foreseen (Anonymous, 1992b).

Table 3. World undeveloped and undiscovered natural gas resource potential outside North America (from MacDougall and Linder, 1992).

	Resources	
	10^{12} m^3	10^{12} ft^3
Latin America	14.9	528
Western Europe	9.6	338
Former Soviet Union	81.4	2877
Eastern Europe	0.6	21
Middle East	74.2	2619
Africa	15.1	534
Asia	11.7	415
Pacific Rim	14.4	508
Total	221.9	7840

2.2 Natural Gas Resources in Canada

Canada has significant natural gas reserves, and significant volumes of undiscovered resources lie almost entirely within the Western Canadian Sedimentary Basin (Alberta, northeast British Columbia, and southern Saskatchewan) and the northern Frontier areas (Mackenzie Delta/Beaufort Sea area and Arctic Islands). Of 2.7 × 10^{12} m^3 of proved reserves (1 January 1993), less than 2% are found in eastern Canada. Total resources assessed by the U.S. National Petroleum Council (NPC) amount to (17.6–21.0) × 10^{12} m^3 (Table 4), although the potential nonconventional resource may be more than double the volumes assessed (National Petroleum Council, 1992). Some opportunities for offshore resource development also exist in eastern Canada (offshore Newfoundland and Nova Scotia) and in western Canada (offshore British Columbia).

The U.S. represents a major market for Canadian natural gas production at the present time. In 1991, 47 × 10^9 m^3 was exported to the United States by pipeline, representing 45% of Canadian production. The re-

Table 4. Natural gas resources of Canada, 1 January 1991 (National Petroleum Council, 1992).

	Current technology		2010 technology[a]	
	10^{12} m^3	10^{12} ft^3	10^{12} m^3	10^{12} ft^3
Proved	2.0	72	2.0	72
Conventional				
Reserve appreciation	0.6	22	0.7	24
New fields	2.8	99	3.1	109
Frontier	8.3	293	9.0	317
Subtotal	11.7	414	12.8	450
Nonconventional				
Coalbed methane	2.3	80	3.7	129
Tight gas	1.6	55	2.5	89
Subtotal	3.9	135	6.2	218
Total resources	17.6	621	21.0	740

[a] Basis—Technically recoverable resources incorporating technology advancement through 2010.

maining Canadian production was split almost equally between Ontario and Québec and western Canadian markets.

2.3 Natural Gas Resources in Mexico

Mexico's natural gas production is predominantly associated with the nation's oil production. Current (1992) gas production is about 31.3 × 10^9 m^3 per year (British Petroleum Company, 1993), 87% of which comes from the onshore and offshore basins in the Southern Isthmus area. Proved reserves of Mexico are estimated at 2.0 × 10^{12} m^3 and total assessed resources at 7.1 × 10^{12} m^3 under assumptions of current technology (National Petroleum Council, 1992). The current reserves-to-production ratio is about 51:1.

Exports of natural gas from the United States to Mexico are presently growing as industrialization proceeds in the border area, thus increasing demand for electricity and for process heat. New cross-border interconnections are being added. United States exports to Mexico in 1991 were 1.71 × 10^9 m^3 and in the first half of 1992 were 1.25 × 10^9 m^3 (White, 1992a).

2.4 Natural Gas Resources in the United States

The most recent U.S. assessment of natural gas resources for the lower 48 states indicates 4.53 × 10^{12} m^3 of proved reserves (lower 48 states) and 32.13 × 10^{12} m^3 of resources (at levels of technology expected to be developed through 2010) as of 1 January 1991 (National Petroleum Council, 1992). The resources beyond proved reserves are about equally divided between conventional and unconventional reservoir types (Table 5). Larger volumes of in-place resources exist beyond those assessed within this recent evaluation but will only be recoverable with advanced technological development.

Resource estimates made in the period 1980–1992 have reflected progressively greater understanding of the resource base. In particular, this understanding has incorporated new knowledge of reserve growth, also termed reserve appreciation, of old fields and of new technology applicable to the production of all classes of natural gas resources. This is particularly true for the nonconventional re-

Table 5. U.S. lower 48 states natural gas resource base (National Petroleum Council, 1992).

	10^{12} m^3	10^{12} ft^3
Proved reserves	4.53	160
Conventional resources		
Reserve appreciation	5.75	203
New fields	11.69	413
Subtotal	17.44	616
Nonconventional resources		
Coalbed methane	2.77	98
Shales	1.61	57
Tight sandstones	9.88	349
Other	0.42	15
Subtotal	14.69	519
Total resources[a]	36.66	1295

[a] Technically recoverable resource base as of 1 January 1991 incorporating technology advancement through 2010.

sources such as those in low-permeability (tight) sandstones, coalbeds, and fractured shales. As the resource base has been more accurately assessed and as all categories of the resource base have been included in the assessment process, the total volume of assessed resources has grown. It is important to recognize that more than half, over 194 000 wells, of all the producing gas wells ever drilled in the United States [total 394 915 from 1918 to 1990 (DeGolyer and MacNaughton, 1991)] were completed between 1974 and 1990. As a result, substantially more is known in the early 1990s about the geology and engineering properties of natural gas reservoirs in the United States than just a decade ago. Much of this same knowledge will ultimately be applicable worldwide to advanced reservoir development.

2.4.1 United States Resource Estimates through Time Natural gas within any producing area, regional or national, consists of proved reserves in known fields, inferred reserves (the next most certain category of reserves) and undeveloped resources within these fields (the source of reserve growth), and undiscovered resources. Other than proved reserves, all volumes of future natural gas supply are estimates based on information derived from past and current experience in gas production and reservoir development. Even proved reserves are subject to periodic revision (Finley and Fisher, 1991). In 1988, the U.S. Department of Energy (DOE) evaluated the then-current resource assessments of the Potential Gas Committee, the U.S Geological Survey, and the Minerals Management Service to create a new, integrated assessment of U.S. natural gas resources (Finley *et al.*, 1988). An important element of the DOE-sponsored study was the first-time quantification of nonassociated gas reserve growth potential based on depositional complexity (reservoir heterogeneity) of many types of gas reservoirs. The 1988 DOE assessment of natural gas resources for the lower 48 states defined 29.98×10^{12} m^3 of reserves and recoverable resources with current technology. Proved reserves made up 4.50×10^{12} m^3 of the total. Of that total, 3.37×10^{12} m^3 represented nonassociated reserve growth resources related to depositional complexity within known reservoirs. The balance derives from such approaches as drilling deeper for new reservoirs or extending field boundaries. The total assessed reserve growth resource was 8.15×10^{12} m^3.

Many analysts regarded the DOE assessment as high in terms of the total volume of resources. Since that time, however, assessments by other organizations, public and private, have confirmed resource bases of over 28.3×10^{12} m^3. Enron Corp. estimated reserves plus recoverable resources of 33.69×10^{12} m^3 (with advanced technology) as of 1990, and the National Research Council (NRC) assessed 40.50×10^{12} m^3 of reserves plus resources as of year-end 1986 (National Research Council, 1990; Enron Corp., 1991). The Gas Research Institute (GRI), in their 1992 analysis of long-term trends in U.S. gas supply and prices, indicates 34.2×10^{12} m^3 of reserves plus resources as of 1 January 1991 (Holtberg *et al.*, 1992). Most recently, the NPC has assessed a supply base (reserves plus resources) of 36.66×10^{12} m^3 with advanced technology (Table 5). Including the NPC study, the average of the five assessments (DOE, Enron, NRC, GRI, and NPC) cited here is about 35×10^{12} m^3, a much more substantial resource than had been assumed prior to 1988. Note that somewhat different assumptions, such as those regarding the use of current or future technology, apply to each assessment. In comparison, the U.S. Geological Survey's assessment as of 1 January 1980 for the U.S. lower 48 states was only 23.33×10^{12} m^3 of reserves plus resources (Dolton *et al.*, 1981) and as of 1 January 1987 was 17.3×10^{12} m^3 (Mast *et al.*, 1989).

3. HISTORY OF NATURAL GAS RESOURCE DEVELOPMENT

3.1 United States Natural Gas Development

Total U.S. cumulative production of natural gas amounts to about 21.7×10^{12} m^3 as of 1 January 1991; only about 142×10^9 m^3 of that volume has been produced in Alaska and the balance in the lower 48 states (Potential Gas Committee, 1991). Annual dry gas production for the total United States reached 615×10^9 m^3 in 1973 and declined to 456×10^9 m^3 in 1986 before returning to about 503×10^9 m^3 for 1991 and 1992 (Energy Information Administration, 1993b). The history of natural gas production in the United States has been one of interplay between government regulation from the wellhead to the local distribution company vs free market forces. Only as of 1 January 1993 have the last wellhead price controls been removed; however, since most gas was already deregulated, little effect on market prices was seen on that date. Transportation by pipelines is now predominantly a service provided for gas not owned by the pipelines, a transition that began in the mid 1980s and continues into the early 1990s.

The gas industry in the United States began in the mid to late 1800s with gas originally used for lighting and with much of that gas being manufactured gas from coal or wood. Manufactured gas continued to be important through World War II because there was no inexpensive means to transport natural gas long distances beyond the producing regions. By 1938, interstate movement of natural gas amounted to only 11.3×10^9 m^3 annually. That same year the first major piece of federal regulatory legislation was passed as The Natural Gas Act of 1938, designed to protect consumers from perceived monopolistic pricing practices of gas utilities and pipelines (National Petroleum Council, 1992).

A surplus of natural gas existed during the 1940s and 1950s as a by-product of oil production. The Railroad Commission of Texas took action in 1947–1948 to limit flaring of gas associated with oil production, thus reducing the waste of natural gas and promoting expanded natural gas use. Following World War II, long-distance interstate pipeline transportation became feasible beyond the six principal producing states of that time, namely Texas, Louisiana, California, Oklahoma, West Virginia, and Kansas. In 1945, interstate transportation volumes reached about 85×10^9 m^3. By 1960, volumes had increased to over 340×10^9 m^3 and by 1969 to 538×10^9 m^3. Much of this supply was destined for operation of a greatly expanded industrial base and for millions of new homes constructed during the post-war period.

One effect of federal regulation was a long period (1950–1970) of low gas prices and major demand growth. Exploration for and development of natural gas resources became less economically attractive during this period as a result of low wellhead prices and rising costs, and the result was a decline in total domestic gas production beginning in late 1973. Service curtailments developed, particularly during the severe winter of 1976–1977. Because federal market regulation did not apply within the states, a dual market developed consisting of higher-priced (more market sensitive) intrastate gas compared to lower-priced, regulated interstate gas. In response to supply and market issues, Congress passed the Natural Gas Policy Act (NGPA) in 1978, which established pricing categories for all gas produced and sold in the United States and helped spur resource development, particularly in some of the more difficult-to-produce resource categories such as gas from tight sandstones. The Fuel Use Act of 1978 was passed to prohibit use of natural gas in power plants and large industrial boilers. The perception was that natural gas was a limited resource and should be preserved for "higher" uses such as home heating.

As the United States entered the 1980s, price-induced conservation, long-term fuel switching, industrial restructuring away from gas-consuming heavy industries, and efficiency improvements reduced total consumption from around 563×10^9 m^3 in 1980 to about 477×10^9 m^3 in 1983 (Energy Information Administration, 1993b). Incentives under the NGPA supporting resource development combined with declining consumption to create a "gas bubble" of excess deliverability beginning in 1982–1983 (National Petroleum Council, 1992). This excess, and the orientation of the Federal Energy Regulatory Commission (FERC) toward a restructuring of the industry toward a freer and less regulated environment, led to gas-on-gas price competition and low spot-market natural gas prices that prevailed into the early 1990s. By the latter half of 1992, it began to appear that supply and increasing demand had come into a balance as the FERC continued to foster deregulation of the industry.

3.2 Canadian Natural Gas Development

Natural gas was discovered in Ontario and Alberta, Canada, in 1880 and 1884, respectively. In Alberta, some of the first pipelines were built to supply natural gas to the cities of Calgary and Edmonton between 1912 and 1923. Large-scale use of natural gas and widespread exploration of the Western Canadian Sedimentary Basin, mostly for oil, both began in 1947. Production from fields in British Columbia was connected to markets in the Vancouver area in the late 1940s and early 1950s, and the TransCanada pipeline was completed in 1958, linking western source areas to markets in the eastern provinces of Ontario and Quebec. A pipeline to the California market in the United States was completed in 1961 (National Petroleum Council, 1992).

Canadian natural gas production has increased steadily during the 1980s; 1984 production was 79 × 10^9 m^3, whereas 1991 production was 105 × 10^9 m^3. Of 1991 production, 47 × 10^9 m^3 was exported to the United States, more than doubling the export level of 22 × 10^9 m^3 of 1984 (Arthur Anderson & Co., 1987; National Petroleum Council, 1992). Canada's share of the U.S. market was 8.7% in 1991, increasing from 4.6% in 1986, as a result of expansion of cross-border pipeline capacity.

3.3 Mexican Natural Gas Development

Natural gas produced in Mexico has been primarily associated gas related to that nation's oil production from the Gulf of Campeche offshore and the Reforma area onshore between the cities of Veracruz and Villahermosa. Annual production in 1992 was about 31.3 × 10^9 m^3 (British Petroleum Company, 1993). Development of the gas resource was concurrent with the development of Mexico's oil resources during the 1970s and 1980s. The main trunk line system for natural gas distribution is a 1245-km-long pipeline system that extends from the Reforma area north to Monterrey. The pipeline system was completed in 1980 with additional lines extending farther north and also northwest and west to Mexico City and the Pacific coast.

In the early 1970s and from 1980 to 1984, the United States imported natural gas from Mexico; U.S. imports peaked at 3 × 10^9 m^3 in 1984. The U.S. exports to Mexico were never more than 0.11 × 10^9 m^3 annually between 1977 and 1988 but in 1989 jumped to 0.48 × 10^9 m^3 and in 1991 reached 1.7 × 10^9 m^3 annually. These supplies have been flowing into northern Mexico to meet demand from increased industrial development and also to replace supplies from southern Mexico that are now being used to displace fuel oil use in and around Mexico City. This displacement is an effort to improve air quality in that region. Industrial expansion possible under the North American Free Trade Agreement will likely lead to additional demand for U.S. gas exports to Mexico through the 1990s with that demand growing at an annual rate of 7% to 10% through 2000 (White, 1992a).

4. TECHNOLOGIES FOR NATURAL GAS DEVELOPMENT

4.1 Delineation of Reservoir Types

Technologies for natural gas development can be related to major categories of the natural gas resource base from which the gas is produced: associated reservoirs, nonassociated conventional reservoirs, and nonassociated unconventional reservoirs. Associated reservoirs are oil reservoirs that also produce natural gas. Such gas may be present in solution with oil or as a free gas cap or both. Associated gas may be produced and marketed or may be reinjected (cycled) to maintain reservoir pressure. The technologies related to associated gas resources are those of oil exploration and development of oil resources, many of which are identical for natural gas with the major exception of unconventional gas reservoirs.

Nonassociated conventional reservoirs have sufficient permeability so that the trapped gas will readily flow to the wellbore. They also typically have sufficient storage capacity in pores within the rock matrix so that they are not dependent on gas flow in natural fractures to provide a production pathway to the wellbore. The amounts of oil produced are minimal, although "wet" gas reservoirs contain economically recoverable quantities of natural gas liquids (NGLs) such as propane and butane. Currently (1992–1993) in the United States, readily producible reservoirs with a high content of NGLs are economically attractive because of the demand for NGLs as chemical feedstocks.

Nonconventional nonassociated reservoirs include low-permeability sandstones (and carbonates), coalbeds capable of gas production, and fractured shales. In the United States, reservoirs are considered low permeability, or tight, if their *in situ* matrix permeability is 0.1 millidarcy or less. Hydraulic fracturing to create a man-made fracture propped open with clean sand injected into the formation is the typical method used to produce natural gas from tight sandstones. This process may be helped or hindered by the presence of natural fractures. Coalbeds, shales, and tight sandstones have limited matrix porosity (storage capacity) but natural fractures ("cleat" in coal beds) act as flow pathways to the wellbore. Coalbeds produce gas that is adsorbed onto the coal surface after reservoir pressure is reduced, a process requiring varying amounts of initial water production. Fractured shales, some of which are quite silty, also produce gas from fractures and from adsorbed gas within the matrix of the rock. The amount of gas present in shale reservoirs is related to the organic content of the rock, and the shales act as the source, the reservoir, and the trap for the natural gas. The location of unconventional reservoirs is well known in many gas-producing regions of the United States and Canada, which contrasts with the need to search for conventional natural gas traps. Unconventional reservoirs are more sensitive to the levels of technology available for extraction from a known accumulation and whether or not such technology can be economically deployed at a given market price.

4.2 Technologies for Conventional Reservoir Development

4.2.1 Exploration Technologies Exploration for conventional natural gas today benefits from advanced geophysical technologies, new concepts of basin analysis and the sequences in which basin-filling rocks were deposited, satellite remote sensing, and a host of powerful computer-based information-handling and interpretation capabilities. Geophysical technologies include standard two-dimensional (2-D) seismic surveys but with improved resolution and the ability to estimate reservoir attributes and fluid content. Two-dimensional seismic surveying yields individual cross-section lines but without the ability to construct such sections at orientations other than that in which the line was acquired. Nor can 2-D data be used to construct images of continuous surfaces through a potentially gas-bearing rock volume to examine reservoir structure and quality.

4.2.1.1 Three-Dimensional Seismic Imagery Three-dimensional (3-D) data, acquired on closely spaced grid lines at right angles to each other, can be used to reconstruct seismic images at any orientation through the data volume and to image surfaces at any level within the volume. Typically, 3-D data are used for field development. Recently, however, 3-D seismic surveys are being deployed over large areas to define complex structures, especially in the U.S. Gulf of Mexico where migration of bedded salt has occurred. This offshore province yielded 26% of total U.S. natural gas production in 1991 and has, on average, produced about two-thirds gas and one-third oil in terms of its total supply mix. The expense of drilling wells in deep waters (200–1000 m) in the offshore U.S. Gulf of Mexico mandates that 3-D seismic surveying be utilized to reduce the risk of offshore exploration and development.

4.2.1.2 Basin Analysis and Sequence Stratigraphy Explorationists today are integrating seismic, well, and biostratigraphic data to understand the depositional framework and structural style of a basin. Sediment input and relative changes in the sea level are key controls on depositional sequences that define parts of a basin potentially prospective for the trapping of natural gas and oil. Such controls influence the distribution of source beds, the possible timing of hydrocarbon migration, and the relative position of reservoir rocks as they exist today. The process of evaluating all such controls and the resulting reservoir geometry is termed *sequence stratigraphy*.

4.2.2 Development Technologies Advanced technologies for development of natural gas reservoirs include innovations in reservoir geological characterization, geophysical detection, and drilling and production techniques. In mature producing trends in the United States and Canada, improved extraction from known reservoirs through advanced technologies has significant potential to sustain the yield of natural gas from known producing regions, thus extending the economic life of surface production infrastructure and the natural gas pipeline network.

4.2.2.1 Reservoir Characterization Technology Geologists and reservoir engineers are cooperating to an ever greater degree to define the detailed internal character of reservoirs, even where wells are separated by several hundred to 1000 m. Geological aspects of reservoir characterization involve predictions about the interwell area that can only be indirectly assessed from information collected at the wellbore or obtained seismically between wellbores. Nonuniform deposition of the rocks forming the reservoir unit, cementation of pore spaces between grains that reduces gas storage capacity, and structural changes after burial, such as faulting, affect flow of natural gas to the wellbore. Such variabilities in the reservoir rock, termed *reservoir heterogeneities*, have resulted in near-original reservoir pressures being encountered during drilling in reservoirs thought to have been largely depleted. Incomplete recovery of gas from a field because of flow boundaries within the reservoir can be overcome by additional selective drilling or recompletion of other, untapped reservoirs once the primary reservoir target has been depleted (see Sec. 4.4). Incremental gas recovery in mature fields in the United States is a major topic of current research that is expected to help sustain gas supplies well beyond current reserve estimates.

4.2.2.2 Geophysical Detection Technologies Advanced geophysical detection technologies applied to gas resource development are used to define the physical framework of the reservoir within a producing field. This framework includes the composition of the reservoir rock and distribution of pore space within the reservoir, and, under certain conditions, yields the detection of potential indicators of natural gas vs formation water within the reservoir. The latter is termed *bright spot technology*. Virtually all of this type of geophysical work is now done with 3-D seismic data (see above) as a result of the much greater detail available. Automated routines now available on workstations that have 2–10 (or more) gigabytes of memory allow rapid interpretation of a 3-D data set. Because of the importance of careful targeting of new wells in reservoir development, the proportion of 3-D seismic surveying to all seismic lines shot in the Gulf of Mexico and the North Sea has gone from 10% in 1980 to 60% in 1989 (Hansen *et al.*, 1989).

Future geophysical technologies are now being designed to avoid problems with surface conditions that absorb seismic energy. Cross-borehole tomographic techniques are emerging that use an energy source in one well and a receiver array in another well to image the interwell area. This approach is akin to computerized axial tomography (CAT) scanning, a diagnostic medical technique consisting of multiple x-ray scans through the body. The distance between wells with 1992 technology is limited to 200–350 m, somewhat less than typical gas well spacing. Higher-energy sources should extend application of this technique to wellbores 600 m or more apart, which will result in more frequent application in gas field development.

4.2.2.3 Horizontal Drilling Among advanced drilling and production technologies, horizontal drilling offers the opportunity to drain naturally fractured reservoirs more efficiently. It is also potentially effective to overcome depositional variations that compartmentalize reservoirs (reservoir heterogeneity) and prevent producible gas from flowing to the wellbore. Horizontal drilling has been a technical and economic success in the Austin Chalk trend, a carbonate oil reservoir with large volumes of associated gas, in south and central Texas, U.S.A. Experiments are now underway to horizontally drill fractured shale reservoirs, a type of unconventional gas reservoir, to increase production. Horizontal drilling of low-permeability sandstone reservoirs has been successful in the North Sea, offshore of the Netherlands, to increase both oil and gas production. There, multiple hydraulic fracture treatments have been carried out from the horizontal segment of the well, a technique that has widespread potential application to low-permeability natural gas reservoirs. Although horizontal drilling requires experience in each region where gas is produced, successful horizontal wells may have productivity 2.5–5 times that of vertical wells at 1.5–2.5 times the cost once that experience is gained.

4.3 Technologies for Unconventional Reservoir Development

Unconventional reservoirs of natural gas are ones whose location is usually well known as a result of previous exploration and development activities but that often have not been as extensively exploited as the conventional accumulations. The primary unconventional reservoirs include low-permeability (tight) sandstones, coalbed methane, and fractured shales, all of which are contributing to current world gas supplies. Other unconventional natural gas sources are known, such as gas entrained in geopressured brines, and gas hydrates, which are physical combinations of gas and water in which the gas molecules fit into a crystal structure resembling that of ice. Gas hydrates are extensive along some submerged continental margins and along the North Slope of Alaska. Very little is known about the economics of gas hydrate development. Extracting gas from geopressured brines suffers from the low energy content of the brines, typically 0.6–1.1 m^3 of gas per barrel (159 l) of brine, and the disposal problems associated with production of large volumes of saline fluids.

4.3.1 Advanced Extraction Technologies

Near-term commercially viable unconventional gas resources will be extracted through improved technology applied to tight sandstones, coalbeds, and fractured shales. The technologies are those of improved extraction from largely known accumulations and, therefore, are characterized as extraction techniques rather than either exploration or development technologies.

4.3.1.1 Hydraulic Fracturing The process of hydraulic fracturing involves pumping mixtures of fluids and a solid, such as loose quartz sand, into a natural gas reservoir with the objective of fracturing the reservoir rock. The resulting fracture in the rock, especially if held open by sand or other propping material carried into the reservoir, forms a direct, high-permeability pathway to the wellbore. Hydraulic fracturing is designed to overcome the low permeability of a rock formation that otherwise would not yield gas to a wellbore at economic rates sufficient to justify drilling and operating the well. A key issue in hydraulic fracturing is control of the geometry of the hydraulic fracture to maximize its extension in the reservoir, to limit its height to the target productive interval, and to know its orientation.

The geometry of the fracture is controlled by the stress state in the reservoir, which relates both to the regional basin-scale characteristics, such as the tension direction generally normal to the shoreline of the U.S. Gulf of Mexico, and to the physical character of individual beds (toughness relative to fracture propagation). If a hydraulic fracture penetrates upward rather than outward from a well, it achieves minimal contact with gas-saturated rock and production is inefficient. Excessive formation water may be brought into the well instead of natural gas, and more wells will need to be drilled to extract the gas resource from a given area.

Advanced techniques are being applied to model fracture propagation and to test formations in advance with very small-scale fracture treatments for analysis of fracture orientation and of pumping pressures required to induce fracturing. The most sophisticated techniques involve modeling fracture propagation through analysis of downhole pressures, flow rates, and fluid characteristics as the fluids and sand proppant are being pumped into the formation. Pumping rates, sand concentrations, and fluid properties are then varied to ensure that the maximum volume of treatment fluids penetrate the targeted reservoir interval.

4.3.1.2 Cavity Completions A new technique applied to coalbeds capable of yielding methane to wellbores is cavity completion. This technique involves pulsing a well by alternately injecting fluid and then allowing the fluid to flow back to the surface under rapid release of pressure. The target coalbeds immediately below the casing of the wellbore break down along their natural fractures, termed cleat, into particles of coal that are transported to the surface during the pulsing process. The result is a cavity that connects the wellbore to the coal cleat system. Methane is then produced from the well through desorption from the coal, typically after a period of dewatering that may be limited to days or that may take months to achieve economic gas production rates. Cavity completions have been successful in the San Juan Basin of northeast New Mexico, U.S.A., where extensive research into their development was carried out by GRI in 1990–1992.

4.4 Reserve Growth in Mature Fields

Not all improvements in natural gas production technologies in the early 1990s relate to field techniques for reservoir stimulation or well completion. The concept of reserve growth in mature fields targets a resource that has been assessed at between 6.68×10^{12} m^3 (National Petroleum Council, 1992) and 7.48×10^{12} m^3 (Energy Information Administration, 1990) but will probably be closer to 8.5 $\times 10^{12}$ m^3 in the U.S. lower 48 states once fully developed. The availability of this gas resource is dependent on understanding reservoir compartmentalization inherited from the original depositional system of the reservoir rock and its subsequent modification during burial. Thus, reserve growth is dependent on the concept of reservoir heterogeneity and an understanding of how the reservoir framework was deposited.

Because gas is a much more mobile fluid in the reservoir than oil, gas reservoirs of conventional permeability have been developed in the United States on relatively wide well spacings (259–130 hectares per well) compared to oil (16–8 hectares per well, or less). It was largely assumed that the mobility of natural gas would overcome reservoir heterogeneity at the production time scale. We now know that this is not the case in all reservoirs. Those with complex depositional geometries or those that have an areally or vertically variable reservoir quality cannot be uniformly drained by widely spaced wells. It has long been known that reservoirs broken up by faulting must be more extensively drilled, but recognition of the extensive gas reserve growth potential related to the former factors only emerged in the United States during the period 1988–1992. Certainly, the increasing maturity of the U.S. resource base, while seen as a problem from the viewpoint of exploration, now offers the required data base for reserve growth realization. Resultant reservoir development and redevelopment approaches will extend to applicable reservoirs beyond the United States as current world gas reserves mature and must be replaced.

5. GAS GATHERING, PROCESSING, AND DISTRIBUTION

Beyond the wellhead, gas gathering, processing, and distribution begins in the field where gas is produced. Water and associated oil may be separated from the gas stream in equipment serving one or several wells. A series of small-diameter pipelines then delivers gas to a natural gas processing plant. Natural gas liquids and impurities, such as carbon dioxide, hydrogen sulfide, and nitrogen, may be separated at a processing plant that serves one or more fields within a producing trend. The natural gas leaving such a plant would typically be of a quality suitable for consumption [approximately 1056 MJ (10^6 btu) per 28 m^3 (10^3 ft^3)].

In the United States, there are about 1.9 million km of pipelines that gather and move natural gas from the producing regions to distributors and end users. The U.S. natural gas demand has a high degree of seasonality with 60% of consumption occurring during the winter heating season (October through March). As a result, underground storage is needed to balance the relatively constant flow from natural gas reservoirs with seasonality, peak-month, and even peak-day usage that can vary with weather patterns. Peak-month delivery may be 4–8 times low-month delivery, and peak winter day usage may be 1.5–3 times greater than average winter day requirements. There are now about 370 underground storage fields in the United States with a working gas capacity of about 113 $\times 10^9$ m^3. These fields can provide (between 1.5 and 1.8) $\times 10^9$ m^3 per day on a peak day. The estimated 1991 peak-day U.S. capability to transport natural gas was 3.39 $\times 10^9$ m^3, whereas the estimated 1991 peak-day U.S. demand was 2.89 $\times 10^9$ m^3. The estimated annual delivery capability in 1991 was 680 $\times 10^9$ m^3, compared to actual 1991 consumption of 544 $\times 10^9$ m^3. Of the latter volume, 47 $\times 10^9$ m^3 was supplied through net pipeline imports from Canada (National Petroleum Council, 1992; Gas Research Institute, 1993).

The business of natural gas distribution changed markedly in the United States in the decade of the 1980s. Pipelines previously played the role of middleman between producers and the local gas utilities or local distribution companies (LDCs). Pipelines purchased the gas from producers, transported it as system gas owned by the pipeline, and sold it to the LDCs. Contractual relationships tended to be long term and involved relatively few parties, and the industry's interstate activities were heavily regulated. As the industry entered the 1990s, the majority of interstate gas transported by the pipelines was not owned by those pipelines; the pipeline company only provided the transportation service. Pipelines are now much less in the business of aggregating supplies. Independent gas marketers, individual producers, and

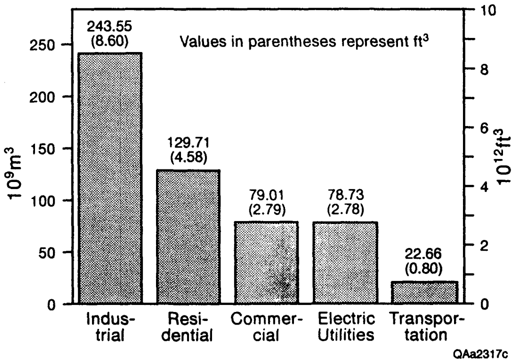

FIG. 1. Natural gas consumption in the United States by sector, 1991 (Energy Information Administration, 1992).

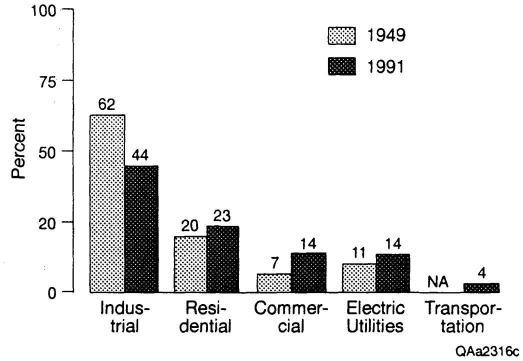

FIG. 2. Comparison of U.S. natural gas consumption by sector, 1949 and 1991 (Energy Information Administration, 1992).

groups of producers are selling gas, and the pipelines are under regulatory pressure to unbundle their services to distinguish, for example, costs of transportation and of storage. The LDCs and large end users are now in the business of building their own supply portfolios, arranging supply contracts, and balancing purchases with demand. Trends toward disaggregation of functions within the natural gas business are likely to continue particularly with respect to "open access" to space on pipelines for all gas shippers.

6. UTILIZATION AND APPLICATIONS

Natural gas can fill a variety of energy needs in fuel and power applications and as a raw material or feedstock for petrochemicals, fertilizers and other applications. Among the five natural gas–consuming sectors in the United States, industrial uses represented the greatest demand, followed by residential, commercial, electric utilities, and transportation (Fig. 1). Over the last 40 years, however, the volume of the industrial market has declined as a component of the total gas market, whereas the commercial sector has doubled its share of U.S. total gas use (Fig. 2). Total gas use peaked with top production in the early 1970s, then declined irregularly through 1986, but has seen a rebound up to 1991 to the consumption level of 1975 (Fig. 3) (Energy Information Administration, 1993a). These recent increases in U.S. gas use are in part driven by new end-use technologies for natural gas, removal of regulations on natural gas use in boiler and utility applications, and recognition of natural gas's benefits as an environmentally benign, domestically abundant fuel. It would be appropriate to consider current and future utilization of natural gas in each of the five major use sectors for the United States and other leading consuming nations (Tables 1 and 6).

6.1 Industrial Natural Gas Use

Industrial energy demand is the largest single category of natural gas use in the United States, Germany, and Canada (Fig. 1; Table

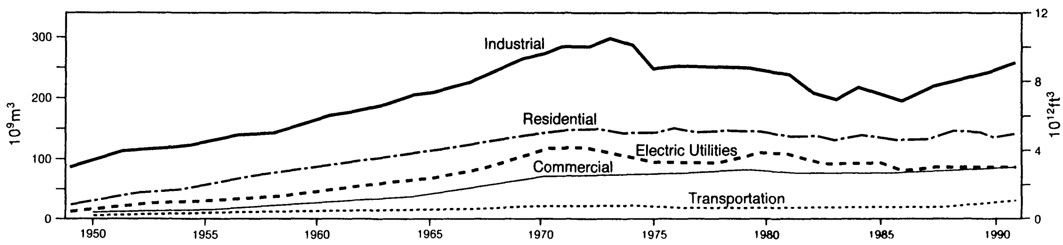

FIG. 3. History of U.S. natural gas consumption, 1949 through 1991 (Energy Information Administration, 1992).

6). To meet U.S. industrial demands, natural gas use for fuel and power grew between 5 and 10% annually between 1986 and 1991. This rate of growth is expected to slow from the mid-1990s to 2000 because natural gas prices are expected to increase at relatively greater rates than other fuels. Cogeneration of process steam and power represents 40% of the industrial power category of natural gas use and is the largest single category of use (Holtberg *et al.*, 1992).

Natural gas is used in the production of ammonia, hydrogen, and methanol, with hydrogen primarily being produced and used by oil refineries. Approximately 60% of U.S. natural gas use as feedstock is for ammonia production and about 25% is for hydrogen production. Methanol production accounts for about 17% of feedstock demand. Methanol is used in developing other chemicals as gasoline additives. Total industrial feedstock consumption of natural gas is projected to decline by about 15% by 2010, primarily as a result of reduced ammonia consumption for domestically produced fertilizers (Holtberg *et al.*, 1992).

6.2 Residential Natural Gas Use

Natural gas accounts for the largest portion, 45%, of total U.S. residential energy consumption, followed by electricity at 32%. Space heating accounted for 65% and water heating about 15% of 1990 residential gas consumption. Cooking, clothes drying, etc., accounted for the balance. The share of U.S. gas consumption used for space heating is expected to decline to 55% by 2010 as increased equipment efficiencies and improved thermal integrity of buildings more than offsets projected growth in the number of space-heating installations (Holtberg *et al.*, 1992).

The residential gas market is the major part of the "core" market for natural gas. Residential natural gas customers supplied by LDCs do not have alternative fuel capabilities and are thus captive with respect to their fuel choice. Such customers generally pay the highest prices for natural gas and benefit from the most highly developed parts of the LDCs distribution system. Because two-thirds of U.S. residential consumption is for space heating, year-to-year residential demand is highly dependent on winter weather severity, as it is worldwide for major countries using gas for residential consumption (Table 6).

6.3 Commercial Natural Gas Use

Natural gas consumed in the commercial sector (Table 6) is predominantly for space heating, accounting for 50% of gas use in this sector in the United States. The other half of U.S. commercial gas use is spread among water heating (5%), cogeneration (7%), and the balance for other applications such as cooking, process heat, etc. Gas consumption for space cooling in 1990 accounted for just over 1% of U.S. commercial gas consumption, but space cooling is a major potential growth area using technologies such as pack-

Table 6. Gas consumption by sector, 1990, in seven leading consuming nations, 10^9 m^3. Percentage indicates percent of total reported consumption accounted for by the five sectors indicated (from Energy Information Administration, 1993b; Japan Gas Association, 1993; United Nations, 1993).

	Former Soviet Union (71%)[a]	United States (93%)	Germany (88%)	Canada (94%)	United Kingdom (97%)	Netherlands (97%)	Japan (not available)
Industrial	49.88	198.75	22.89	37.02	18.28	9.23	4.02
Residential	9.29[b]	124.35	16.43	13.92	28.48	9.63	7.76
Commercial	14.80	74.28	7.77	10.82	8.33	7.78	2.56
Electric utility	247.56	78.93	12.01	2.21	1.10	9.96	43.30[c]
Transport[d]	0.37	18.69[e]	...[f]	0.07	...[f]	...[f]	...[f]

---

[a] Not included are petrochemical feedstocks, 14.8%, and losses in transport and distribution, 13.6%.

[b] 1991 data; 1990 data not available.

[c] Estimated.

[d] Vehicles, except in U.S.

[e] Mostly pipeline compressor fuel (see text).

[f] Negligible volume.

aged cogeneration systems, heat pumps, and engine-driven chillers (Holtberg *et al.*, 1992). Use of such technologies would tend to balance the seasonal demand for natural gas, thus making more efficient use of the entire gas production and distribution infrastructure where climatic conditions lead to demand for cooling in commercial (and residential) facilities.

6.4 Electric Utility Use of Natural Gas

Among the leading consuming nations, the states of the former Soviet Union consume the most natural gas for electric generation, more than three times U.S. consumption (Table 6). Current (1991) U.S. natural gas usage of 78.93 × 10^9 m^3 per year for electricity generation is slightly greater than commercial sector consumption. Gas consumption in the U.S. electric utility sector is expected to grow as the result of the need to meet the Clean Air Act Amendments of 1990 and the need for new peaking and intermediate load service, particularly in the second half of the 1990s. Between 1995 and 2000, projected U.S. gas usage for electricity generation is expected to grow from 88 × 10^9 m^3 to 119 × 10^9 m^3, or at almost 6% per year. Most of this growth (31 × 10^9 m^3) is expected to be the result of cofiring of coal and natural gas (Holtberg *et al.*, 1992). Federal restrictions on use of natural gas for boiler fuel has held down gas use in this sector in the United States until the early 1990s.

6.5 Transportation Use of Natural Gas

This use sector includes both pipeline transportation and use in vehicles; vehicular use is the smallest use sector worldwide (Table 6). Three-quarters of the current natural gas used for transportation in the United States is consumed by natural gas pipeline compressors. Such use is expected to grow as part of the overall transportation use growth of 3% to 4% per year from 1990 through 2010. Use of compressed natural gas in U.S. vehicles will grow much more substantially at a rate of 26% per year from about 0.11 × 10^9 m^3 in 1990 to about 10.8 × 10^9 m^3 in 2010. Fleet vehicle use is the most likely source of this growth in view of public policy requirements for use of alternative fuels and given the limited number of public refueling stations. A projection of U.S. natural gas vehicle use sees 3.8 million out of 185 million vehicles using natural gas by 2010 (Holtberg *et al.*, 1992).

7. WORLD NATURAL GAS TRADE

Given the extensive reserves and resources of natural gas available in the world, international natural gas trade is predicted to expand from about 283 × 10^9 m^3 annually in 1990 to about 793 × 10^9 m^3 annually in 2015 (MacDougall and Linder, 1992). Pipeline delivery of gas accounted for 74% of such trade in 1990, and the balance was LNG. It is expected that increases in North American natural gas trade via pipeline will require that additional lines and interconnects be built. Current pipeline facilities are very near capacity, according to preliminary data for 1992 throughput volumes to the United States from Canada. With respect to U.S.–Mexico trade, more interconnects are planned to allow U.S. gas to flow into northern Mexico. Such trade has the potential to expand following the 1993 ratification of the North American Free Trade Agreement.

7.1 Liquefied Natural Gas Trade

Trade in LNG is dependent on facilities for liquefaction, shipping by LNG tanker, and regasification. In the Atlantic market, Algeria is currently the prime potential supplier. Facilities planned or under discussion in Venezuela, Nigeria, and Norway would add to the current 34 × 10^9 m^3 of Algerian liquefaction (design) capacity. The United States has four facilities capable of receiving LNG cargoes: Everett, Massachusetts; Cove Point, Maryland; Elba Island, Georgia; and Lake Charles, Louisiana. The Everett and Lake Charles facilities were operated in 1991 with an estimated combined volume for the year of 1.74 × 10^9 m^3 (White, 1992b).

Of major impact in the Atlantic market would be Nigeria's Bonney LNG project and Venezuela's Cristóbal Colón LNG project. The Bonney project is scheduled to begin deliveries in 1997 with the bulk of the shipments going to Italy, Spain, and France. The Cristóbal Colón project is designed to begin deliveries of over 5.7 × 10^9 m^3 per year to the United States and possibly Europe in 1997 or 1998. Venezuela has significant natural gas reserves (3.12×10^{12} m^3) but limited domestic demand, leading to the opportunity to export natural gas. Venezuelan LNG facilities will have the advantage of being closer to the United States than other LNG projects, which should decrease transportation costs. One assessment suggests, however, that the Venezuelan project will not be cost competitive with Algerian supplies until after 2010 (National Petroleum Council, 1992).

In the Pacific, Indonesia dominates the LNG market with a 50% share of liquefaction capacity. Malaysia, Brunei, Australia, Abu Dhabi, and the United States (Alaska) also export within the region, and together all countries have a total annual export capacity of 57×10^9 m^3. The U.S. LNG exports to Japan amounted to 1.5×10^9 m^3 in 1991 from a liquefaction plant located at Cook Inlet, Alaska. The Trans Alaska Gas System (TAGS) project is proposed to export initially about 10.1×10^9 m^3 per year of North Slope gas to Pacific nations, particularly Japan, South Korea, and Taiwan, beginning in 2000. The project will involve a 1300 km pipeline from North Slope oil fields to a terminal in Valdez, Alaska for tanker loading. Gas from these fields will reportedly not be available until 2000 (or even later) because of the need to keep North Slope oil reservoirs under pressure to maintain oil recovery efficiency.

8. SUMMARY AND FUTURE OUTLOOK

Natural gas, principally methane, is an abundant resource in the United States and worldwide. The United States (lower 48 states) has an assessed resource base as of 1 January 1991 of 36.66×10^{12} m^3 (including 4.53×10^{12} m^3 of proved reserves) produceable using advancing technology through 2010. Worldwide proved gas reserves as of 1 January 1993 were 137.9×10^{12} m^3. In the United States, natural gas use has increased 22% between 1986 and 1992, primarily in recognition of natural gas's suitability for industrial uses (low emissions, shorter lead times for gas-fired equipment), including cogeneration. Much of the future supply to continue this growth and to expand gas use in the future, such as an alternative fuel in the transportation sector, will be dependent on advanced recovery technology applied to a better geologically characterized resource base. Unconventional domestic resources such as low-permeability sandstones, coalbeds, and shales will provide increasing volumes of reserves, and old conventional reservoirs in mature fields will be reexplored and redeveloped using new concepts of reservoir geology. Advanced geophysics, horizontal drilling, improved reservoir characterization, improved hydraulic fracturing, and much greater capability to model and develop reservoirs will add to natural gas availability at moderate costs.

Natural gas use in the United States will remain strong in the industrial and commercial sectors but could be affected by continuing low to moderate world oil prices. Dual-fuel–capable facilities in the industrial and electric utility sectors are highly sensitive to delivered price on a Btu basis unless overridden by emissions requirements. End-user concern with natural gas supply has shifted from the volume of the resource base to issues of deliverability. Major U.S. companies have been shifting operations overseas, leading to concerns that the smaller companies and independent producers will be required to provide an increasing share of the domestic supply. The outlook for that supply remains good in the face of these changes as technology leads to improved per-well recoveries and technology-transfer activities by government, private research groups, universities, and trade organizations continue to expand. Such technologies are likely to become an increasingly important U.S. export commodity.

North American and world gas trade will increase as demand increases as a result of industrialization and concurrent concerns for the environment. Most such trade will be by pipeline, but new LNG projects are expected to become economic over the next two decades in several countries as demand grows and resources are developed beyond domestic needs. The United States itself may become an exporter of LNG from Alaska to the Pacific Rim unless pipeline facilities to the lower 48 states become economic and are constructed.

Whereas natural gas was once regarded as limited in availability, it is now known to be an abundant resource whose deliverability is dependent on efficient conversion of resources to proven reserves. In the United States, a balance between supply and demand, encouraged by reduced government regulation and improved efficiency in production and distribution, is emerging in 1992-1993. The U.S. oversupply that has persisted as the so-called "gas bubble" since 1982 has ended. Continued end-use market penetration by natural gas will likely be fostered by environmental benefits of natural gas use but remains fully dependent on the confidence in supply available to the end user.

GLOSSARY

Associated Gas: Gas produced concurrently with oil that was either in solution with oil in the reservoir or in contact with the crude oil as a gas cap within the reservoir.

Conventional Reservoirs: Typically, sandstones and carbonates with greater than 0.1 millidarcy permeability capable of readily flowing gas to a wellbore.

Conventional Resources: Resources included in this category are natural gas and natural gas liquids that exist in conventional reservoirs or in a fluid state amenable to extraction techniques employed in traditional development practices. They occur as discrete accumulations.

Field: A single pool or multiple pools of hydrocarbons grouped on, or related to, a single structural or stratigraphic feature that has trapped natural gas and from which the gas is being produced.

Inferred Reserves: That part of the identified economic resources, over and above measured and indicated reserves, that will be added through extensions and revisions and with the addition of new pay zones in discovered fields. Equivalent to probable resource.

Natural Gas: A naturally occurring mixture of hydrocarbon and nonhydrocarbon gases, with methane being the principal combustible component.

Nonassociated Gas: Free natural gas not in contact with oil, nor dissolved in crude oil, in the reservoir.

Reserve Growth: The addition of proved reserves from a developed reservoir, or series of reservoirs comprising a gas field, as a result of more intensive development (new drilling, well recompletions) within the established boundaries of the field. Undrained compartments, deeper previously undrilled reservoirs, and shallow bypassed reservoirs may all contribute new reserves.

Reserves: Part of the identified economic resource that is estimated from geologic evidence supported directly by engineering data. They are demonstrated with reasonable certainty to be recoverable in future years from known reservoirs under existing economic and operating conditions and are generally equivalent to "proved reserves."

Unconventional Reservoirs: Typically, low-permeability sandstones, shales, and coals with matrix permeability of less than 0.1 millidarcy requiring special treatment to flow gas to the wellbore. Flow in natural or hydraulically induced fractures is essential to economic production rates.

Undiscovered Resources: Resources estimated to exist, on the basis of broad geologic knowledge and theory, outside of known fields or known accumulations. Also included are resources from undiscovered pools within known fields to the extent that they occur as unrelated accumulations controlled by distinctly separate structural features or stratigraphic conditions.

Works Cited

American Gas Association (1990), *Changes in Natural Gas Recovery Technology & Their Implications,* a report of A.G.A. Planning and Analysis Group, September 1990, Arlington, Virginia, 34 pp.

Anonymous (1992a), "Worldwide Production Report," *Oil Gas J.* **90** (52), 39–43.

Anonymous (1992b), "Russia Pins Energy Hopes on Western Siberia Gas," *Oil Gas J.* **90** (36), 17–20.

Arthur Andersen & Company (1987), *Natural Gas Trends, 1987–88 Edition,* Chicago, IL: Cambridge Energy Research Associates, 154 pp.

British Petroleum Company (1991), *BP Statistical Review of World Energy: June 1991,* London: Ashdown Press Ltd.

British Petroleum Company (1993), *BP Statistical Review of World Energy: June 1993,* London: Leycol Printers Ltd.

DeGolyer and MacNaughton (1991), *Twentieth Century Petroleum Statistics: 1991,* 47th ed., Dallas: DeGolyer and MacNaughton.

Dolton, G. L., Carlson, K. H., Charpentier, R. R., Coury, A. B., Crovelli, R. A., Frezon, S. E., Khan, A. S., Lister, J. H., McMullin, R. H., Pike, R. S., Powers, R. B., Scott, E. W., Varnes, K. L. (1981), *Estimates of Undiscovered Recoverable Conventional Resources of Oil and Gas in the United States,* United States Department of the Interior, Geological Survey Circular 860, Alexandria, VA: U.S. Geological Survey.

Energy Information Administration (1990), *The Domestic Oil and Gas Recoverable Resource Base: Supporting Analysis for the National Energy Strategy,* Washington, DC: Office of Oil and Gas, U.S. Department of Energy, Service Report SR/NES/90-05, 56 pp.

Energy Information Administration (1992), *Geologic Distributions of U.S. Oil and Gas,* Washington, DC: Office of Oil and Gas, U.S. Department of Energy, DOE/EIA 0557, 137 pp.

Energy Information Administration (1993a), *Annual Energy Review 1992,* Washington, DC: Office of Energy Markets and End Use, U.S. Department of Energy, DOE/EIA-0384(92), 350 pp.

Energy Information Administration (1993b), *Monthly Energy Review, February 1993,* Washington, DC: Office of Energy Markets and End Use, U.S. Department of Energy, DOE/EIA-0035(93/02), 164 pp.

Enron Corp. (1991), *Enron Corp's Outlook for Natural Gas: Fueling the Future into the 21st Century,* Houston, Texas: Enron Corp.

Finley, R. J., Fisher, W. L. (1991), *Natural Gas Reserve Additions: Recent Trends and Future Potential in the Lower 48 States,* Society of Petroleum Engineers Hydrocarbon Economics and Evaluation Symposium, Dallas, Texas, April 11–12, 1991, SPE Paper No. 22037, pp. 209–216.

Finley, R. J., Fisher, W. L., Seni, S. J., Ruppel, S. C., White, W. G., Ayers, W. B., Jr., Dutton, S. P., Jackson, M. L. W., Banta, N. (1988), *An Assessment of the Natural Gas Resource Base of the United States,* Bureau of Economic Geology, Report of Investigations No. 179, Austin, Texas: The University of Texas at Austin.

Gas Research Institute (1993), *Policy Implications of the GRI Baseline Projection of U.S. Energy Supply and Demand to 2010,* Washington, DC: Gas Research Institute.

Gold, T. (1987), *Power from the Earth: Deep Earth Gas—Energy for the Future,* J. M. Dent and Sons Ltd., 208 pp.

Gold, T., Soter, S. (1980), "The Deep Earth-Gas Hypothesis," *Sci. Am.,* **242** (6), 154–161.

Hansen, T., Kingston, J., Kjellesvik, S., Lane, G., l'Anson, K., Naylor, R., Walker, C. (1989), "3-D Seismic Surveys," *Oilfield Rev.,* **1** (3), 54–61.

Holtberg, P. D., Woods, T. J., Lihn, M. L., Koklauner, A. B. (1992), *1992 Edition of the GRI Baseline Projection of U.S. Energy Supply and Demand to 2010,* Washington, DC: Gas Research Institute.

Hunt, J. M. (1979), *Petroleum Geochemistry and Geology,* San Francisco, CA: W. H. Freeman and Company.

Japan Gas Association, (1993), "Gas Industry in Japan," briefing paper, Tokyo, Japan.

Kvenvolden, K. A. (1988), in: L. B. Magoon (Ed.), *Petroleum Systems of the United States,* USGS Bulletin 1870, Washington, DC: U.S. Geological Survey, Department of the Interior, pp. 44–45.

MacDougall, M. W., Linder, P. T. (1992), *Long-Term Outlook for World Gas Trade: 1990–2015,* Calgary, Alberta: Canadian Energy Research Institute, Study No. 46, ISBN 0-920522-78-5, 271 pp.

Mast, R. F., Dolton, G. L., Crovelli, R. A., Root, D. H., Attanasi, E. D., Martin, P. E., Cooke, L. W., Carpenter, G. B., Pecora, W. C., Rose, M. B. (1989), *Estimates of Undiscovered Conventional Oil and Gas Resources in the United States—A Part of the Nation's Energy Endowment,* United States Department of the Interior, U.S. Geological Survey, Minerals Management Service, Washington, DC: U.S. Government Printing Office, 44 pp.

National Petroleum Council (1992), *The Potential for Natural Gas in the United States, Executive Summary,* Washington, DC: U.S. Department of Energy.

National Research Council (1990), *Fuels to Drive our Future,* Washington, DC: National Academy Press.

Potential Gas Committee (1991), *Potential Supply of Natural Gas in the United States (December 31, 1990),* Colorado School of Mines, Golden, Colorado: Potential Gas Agency.

United Nations (1993), *Annual Bulletin of Gas Statistics for Europe,* Geneva: Economic Commission for Europe.

Waples, D. W. (1985), *Geochemistry in Petroleum Exploration,* Geological Sciences Series, Boston, MA: International Human Resources Development Corporation.

White, B. (1992a), *The U.S.–Mexican Natural Gas Relationship,* Gas Energy Review, Vol. 20, No. 12, Arlington, Virginia: American Gas Association, pp. 18–27.

White, B. (1992b), *LNG Import/Export Projects—Current Status Report,* Gas Energy Review, Vol. 20, No. 10, Arlington, Virginia: American Gas Association, pp. 10–20.

Wood, P. (1993), *Russia and Ukraine Face Off Over Prices, Payments for Gas,* Natural Gas Week International, a supplement to Natural Gas Week, Washington, DC: The Oil Daily Company.

Further Reading

American Gas Association (1988), *The Future of Natural Gas in the United States,* Arlington, Virginia: Planning and Analysis Group, American Gas Association, April 1988. A general review of U.S. gas use and supply through 2000.

American Gas Association (1990), *Glossary for the Gas Industry,* prepared by Statistics and Load Forecast Methods Committee, Arlington, Virginia: Planning and Analysis Group, The American Gas Association, Fifth Edition. A glossary of natural gas industry terms that includes conversion factors and definitions of abbreviations.

American Petroleum Institute (1991), *1991-92 Guide to Petroleum Statistical Information,* 7th ed., Washington, DC: Central Abstracting & Information Services. A guide to public and private sources of statistical information related to natural gas and oil worldwide.

Barker, C. (1979), *Organic Geochemistry in Petroleum Exploration,* Tulsa, OK: American Association of Petroleum Geologists Department of Education Activities. An introduction to the fundamental ideas of organic geochemistry.

Energy Information Administration (1992), *Geologic Distributions of U.S. Oil and Gas: July 1992.* Washington, DC: U.S. Government Printing Office. The second report in an ongoing series that provides results of domestic oil and gas field-level data constructions, and analyses thereof, founded on EIA's unique collection of annual domestic proved reserves estimates.

Energy Modeling Forum (1988), *North American Natural Gas Markets, Volumes I–III,* Stanford University, Stanford, California, EMF Report 9, variously paginated. A summary of the research by an Energy Modeling Forum working group on the evolution of the North American natural gas markets between 1988 and 2010.

Enron Corp (1993), "The Outlook for Natural Gas," *The Enron Outlook,* March 1993, Houston, TX: Enron Corp. Enron Corporation's outlook for supply and demand for natural gas in North America.

Finley, R. J. (1993), "A Positive Assessment of the U.S. Natural Gas Resource Base," in: S. L. McDonald and M. Mohammadioun (Eds.), *The Role of Natural Gas in Environmental Policy,* Austin, TX: Bureau of Business Research, University of Texas at Austin, pp. 1–7. Describes U.S. natural gas supplies in the context of demand based on environmental policies.

Garrett, C. M., Jr., Hocott, C. R., Finley, R. J., Galloway, W. E., Siqueira, C., Davis, M. D., Murray, R. C. (1985), *Analysis of Negative Revisions to Natural Gas Reserves in Texas,* Bureau of Economic Geology, The University of Texas at Austin, Austin, Texas, Final Report prepared for the Gas Research Institute under Contract No. 5083-800-0908, GRI-85/0111. An assessment of why U.S. natural gas reserves were substantially revised downward.

Julius, D., Mashayekhi, A. (1990), *The Economics of Natural Gas: Pricing, Planning and Policy.* Oxford: Oxford University Press. A comprehensive survey of the major issues involved in the pricing of gas and in the planning of its development.

National Research Council (1991), *Undiscovered Oil and Gas Resources: An Evaluation of the Department of the Interior's 1989 Assessment Procedures.* Washington, DC: National Academy Press. A review, conducted by the National Research Council's Committee on Undiscovered Oil and Gas Resources, of the procedures of the Department of the Interior's most recent resource assessment, which was published in 1989.

U.S. Congress, Office of Technology Assessment (1985), *U.S. Natural Gas Availability: Gas Supply Through the Year 2000,* Washington, DC, OTA-E-245, February 1985. A thorough review of all categories of U.S. natural gas supply with excellent synopses of extraction technologies applicable to each supply category.

NETWORK THEORY, ELECTRICAL

EDWIN C. JONES, JR., *Electrical Engineering and Computer Engineering Department, Iowa State University, Ames, Iowa, U.S.A.*

Introduction 222
1. **Basic Concepts** 223
1.1 Electric Current 223
1.2 Voltage 224
1.3 Power 224
1.4 Linear and Nonlinear Circuits 224
1.5 Time-Varying Circuits 224
1.6 Discrete-Time Circuits 224
1.7 Superposition 224
2. **Continuous-Time Circuits** 225
2.1 Node 225
2.2 Kirchhoff's Current Law 225
2.3 Loop 225
2.4 Kirchhoff's Voltage Law 225
2.5 Branch 225
2.6 Tellegen's Theorem 225
3. **Circuit Elements** 225
3.1 Resistors 225
3.2 Capacitors 226
3.3 Inductors 226
3.4 Mutual Inductance 227
3.5 Gyrator 227
3.6 Sources 227
3.6.1 Independent Voltage Sources 227
3.6.2 Independent Current Sources 228
3.6.3 Dependent or Controlled Sources 228
3.6.3.1 Voltage-Controlled Voltage Source (VCVS) 228
3.6.3.2 Current-Controlled Voltage Source (CCVS) 228
3.6.3.3 Current-Controlled Current Source (CCCS) 228
3.6.3.4 Voltage-Controlled Current Source (VCCS) 229
4. **Circuit-Analysis Techniques** 229
4.1 Series and Parallel Combinations 229
4.2 Tee-Pi or Wye-Delta Conversions 231
4.3 Thévenin Theorem 231
4.4 Norton Theorem 231
4.5 Thévenin–Norton Theorem Comparisons 232
4.6 Thévenin–Norton Theorem Example 232
4.7 Ladder Networks 232
5. **Formal Circuit-Analysis Methods** 234
5.1 Nodal and Loop Analysis 234
5.2 Nodal Analysis 234
5.3 Nodal Analysis with Controlled Sources 235
5.4 Loop Analysis 235
5.5 Loop Analysis with Controlled Sources 236
5.6 Solution of Node-Voltage and Loop-Current Equations 236
6. **Laplace Transforms in Circuit Analysis** 236
6.1 Introduction to Laplace Transforms 236
6.2 Definition of a Laplace Transform 236
6.3 Laplace Transform Theorems 237
6.4 Laplace Transform Pairs 237
7. **Circuit Analysis with Energy-Storage Elements** 238
7.1 Initial Conditions 238
7.2 An Example with Nodal Analysis 238
7.3 An Example with Loop Analysis and Initial Conditions 239
7.4 State-Variable Methods 239
7.5 Solution of State-Variable Equations 241
8. **Sinusoidal Steady-State Analysis** 242
8.1 Description of the Technique 242
8.2 Definition of a Phasor 242
8.2.1 Phasor Algebra 243
8.2.2 Differentiation and Integration 243

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

8.3 Susceptance and Reactance .. 243
8.4 Admittance and Impedance .. 243
8.5 Network Theorems and Solution Techniques 243
8.6 Power and Energy Considerations 244
8.6.1 Power Factor 244
8.6.2 Reactive Volt-Amperes 245
8.6.3 Power and Vars 245
8.7 Three-Phase Systems 245
8.7.1 Balanced Three-Phase Systems 245
8.7.2 Balanced Loads 246
8.7.3 Unbalanced Loads and Symmetrical Components ... 246
9. Two-Port Analysis 246
9.1 Definition of a Port 246
9.2 Open-Circuit Impedance Parameters 247
9.3 Short-Circuit Admittance Parameters 247
9.4 Hybrid Parameters 247
9.5 Transmission-Line Parameters 247
9.6 Two-Port Parameter Conversions 248
9.7 Equivalent Circuits for Two-Ports 248
9.8 Two-Port Analysis 248
10. Discrete Circuits 249
10.1 Introduction 249
10.2 Discrete Signal Representation 249
10.2.1 Analytic Representation 249
10.2.2 Sequence Representation 249
10.2.3 Numerical or Graphical Representation 249
10.3 Discrete-Signal Operations .. 249
10.4 Signal-Flow and Block Diagrams 250
10.4.1 Signal-Flow Graph Definitions 250
10.4.2 Signal-Flow Graph Reductions 250
10.4.3 Mason's Rule or Formula 250
10.5 Discrete Equations 252
10.6 A Compound-Interest Example 252
10.7 The α-β Radar Equations 253
10.8 Numerical Solutions of Difference Equations 254
10.9 The z Transform 254
10.9.1 Definition of the z Transform 254
10.9.2 Examples of z Transforms ... 254
10.9.3 z-Transform Theorems 254
10.9.4 z-Transform Tables 254
10.10 Formulation and Solution of z-Transform Equations 255
10.11 Shannon's Sampling Theorem 257
11. Suggestions for Further Study 257
11.1 Nonlinear and Time-Varying Networks 257
11.2 Computer-Based Circuit Analysis 257
11.3 State-Variable Methods for Continuous-Time Circuits 258
11.4 Topology 258
11.5 Stability 258
11.6 State-Variable Methods for Discrete-Time Circuits 258
11.7 The Modified z Transform ... 258
Glossary 258
Works Cited 260
Further Reading 260

INTRODUCTION

Network theory, also called circuit theory, is the study of the analysis of electric circuits or networks. In turn, these networks are interconnections of electrical devices such as resistors, capacitors, electronic devices, motors, sources of electrical energy, and many others. A few of these devices are described, and others are described elsewhere in this Encyclopedia. Such circuits are described mathematically by ordinary differential equations. The circuits may, however, be linear or nonlinear, time-invariant or time-varying.

The equations needed are given by the describing equations of the elements and two basic principles, known as Kirchhoff's laws. Kirchhoff's current law states that the sum of currents leaving a node is zero, while the voltage law states that the sum of the voltages around a closed path is zero.

Analysis methods proceed from the basic principles, and the choice of method depends on the complexity of the network, the needed information, and the preferences of the analyst. Some analysis methods depend on circuit reductions and equivalent circuits. Other methods require formal equation-writing methods, and include nodal analysis, loop analysis, and state-variable techniques.

Although classical differential equation theory can be used, most circuit analysts use operational methods, especially the Laplace transform. The Laplace transform is especially important for linear, time-invariant networks, and allows complete solutions that consider initial conditions as a part of the complete solution. Linear circuit equations are also solved frequently with computer techniques. An important program is SPICE, the acronym for *Simulation Program with Integrated Circuit Emphasis*. The Electronics Research Laboratory at the University of California at Berkeley developed SPICE in 1975, and made it available to the public. Many improvements and enhancements have been made, and the program is now the most widely used circuit-analysis program. The personal computer version of SPICE is known as PSPICE, available from the MicroSim Corporation. State-variable equations can be solved with Laplace transforms, but are typically solved by computer-based numerical methods when the circuits are nonlinear.

Virtually all of the electrical energy transmission and distribution in the world involves sinusoidal voltage sources. Because of the importance of the topic, special methods for analysis of networks operating in the sinusoidal steady state are needed. These require a special transform commonly called a phasor transform. This transform can be considered as a special case of the Laplace transform, or it can be developed independently. The latter approach is used here.

Electrical networks or circuits often are used to modify signals that carry information in electrical form. A very general type of network is the two-port, a circuit that typically has a pair of input terminals, a pair of output terminals, and the property that, at each pair, the current into one lead is equal to the current out of the other lead. It is often desirable to convert one two-port representation to another, and this process is described.

In recent years, circuits or networks described by difference equations have become quite important. They are the foundation for most digital signal-processing techniques, including image processing. Signals in such circuits are described by sequences in which the discrete quantity (independent variable) is an integral multiple of a time or space unit, and the signals are undefined between integers. Three operations are commonly used: addition, multiplication by a constant, and delay. Such circuits may be solved by techniques quite similar to the techniques used in ordinary differential equations. They may also be solved numerically. An operational method is the z transform. This transform is closely related to the Laplace transform, and it is defined and uses are developed.

1. BASIC CONCEPTS

Network theory, also known as circuit theory, refers to the body of knowledge that describes the behavior of interconnected electrical devices and components. The devices included may be resistors, capacitors, inductors, transformers, transistors, relays, motors, diodes—and there are many more, including sources of electrical energy. The terms network theory and circuit theory will be used interchangeably in this article. Circuits are used in the processing of energy and information in electrical form. Circuits of interest have the property that there is always at least one continuous path through the circuit; that is, it is possible to find a point at which one can start a traverse of the circuit, go through at least two elements, and return to the starting point.

Networks and circuits are described by ordinary differential equations. In general, the independent variable is time, and the circuits are sometimes known as continuous-time circuits. The use of ordinary differential equations results from a condition that is put on circuit theory, which is that the spatial effects of electrical quantities may be neglected. When spatial effects are significant, electromagnetic field theory must be used, a topic that requires the use of partial differential equations. The electrical properties of immediate interest are current and voltage.

1.1 Electric Current

In SI units, the fundamental electrical quantity is current. Current is considered to be the motion of charged particles. The smallest unit of charge is postulated as the equivalent of the charge on one electron, and there are two types of charge, called positive and negative charge. In SI units, the unit of charge is the coulomb (C). The charge on one electron is equal to 1.602×10^{-19} C. A charge flow rate of 1 C/s is defined as 1 ampere (A). By convention, current is considered to be the flow of positive charges. When it is necessary to consider the flow of negative charges, appropriate modifiers are used. A time-varying current will be written as $i(t)$.

1.2 Voltage

The motion of charged particles requires the expenditure of energy or results in the release of energy. The voltage between two points in space is defined as the work per unit charge (joules/coulomb) required to move a charge from one point to the second. Often one point is selected as a reference point, or a point at which the voltage is defined to be zero. When this is done, the voltage difference becomes the voltage at the second point. In general, the choice of reference point is arbitrary. In SI units, the unit of voltage is the volt (V). A time-varying voltage will be written as $v(t)$.

1.3 Power

The power delivered by an electrical source (sources are discussed in Sec. 3.6) to an electrical device is given by

$$p(t) = v(t)i(t), \tag{1}$$

where $p(t)$ is the power delivered to the device. See Fig. 1. The choice of sign is important—the choice made here is commonly called "the passive sign convention." When the power, as given by Eq. (1), is positive, then energy is being delivered to the device in question.

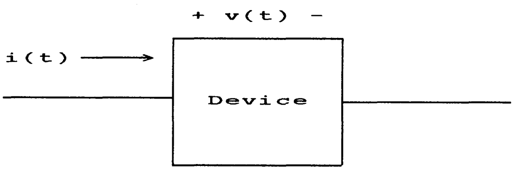

FIG. 1. Sign conventions for electrical devices.

1.4 Linear and Nonlinear Circuits

A linear circuit is a network or circuit that is completely described by a set or system of linear differential equations. In many circuits that are fundamentally nonlinear, the operating conditions are such that it is possible to use the mathematics of linear differential equations to analyze the circuit behavior. This technique is called "small-signal analysis." It depends on use of Taylor-series models, and the neglect of terms of order higher than one. It is necessary to explicitly state the limitations—an example might be that a certain current never goes negative.

A circuit that contains one or more nonlinear elements is fundamentally nonlinear. In most cases, such circuits must be analyzed numerically. While the principles of analysis of linear and nonlinear circuits are similar, the details differ substantially. In this article, the principal focus is on linear circuits.

1.5 Time-Varying Circuits

If one or more of the elements or devices in a circuit varies with time, then the circuit is said to be time-varying. The mathematical techniques used for analysis depend on the type of variation. In many cases the networks must be analyzed numerically. In this article, the principal focus is on circuits in which the component or element values are constant with respect to time, so that the networks are actually described by differential equations with constant coefficients.

1.6 Discrete-Time Circuits

The latter part of this article is devoted to networks or circuits that are described by a set or system of difference equations. The qualities of interest are known and defined at discrete points in time, but are not known or defined between those times. The times may or may not be uniformly spaced.

1.7 Superposition

If a circuit is linear, the principle of superposition applies. This principle states that, for a given system (network) having a general excitation $x(t)$ that yields a response $y(t)$, if the response to an excitation $x_1(t)$ is $y_1(t)$, and the response to an excitation $x_2(t)$ is $y_2(t)$, then any excitation

$$x(t) = a_1x_1(t) + a_2x_2(t)$$

yields the response

$$y(t) = a_1y_1(t) + a_2y_2(t) \tag{2}$$

for all values of the constants a_i.

2. CONTINUOUS-TIME CIRCUITS

In this section, some important terms in network analysis will be introduced. They include elements, nodes, loops, Kirchhoff's laws, and Tellegen's theorem.

2.1 Node

The junction of three or more wires from circuit elements in a circuit or network is called a *node*. (It should be noted that some writers used the term "node" to identify a junction of two or more elements, and reserve the term "essential node" for a junction of three or more elements.) One of the important laws of circuit analysis, Kirchhoff's current law, makes use of this definition of a node.

2.2 Kirchhoff's Current Law

Kirchhoff's current law, often abbreviated KCL, states that at any node in a circuit or network,

$$\Sigma i = 0. \tag{3}$$

The law states that the sum of currents leaving (or entering) a node is always zero. This principle results from Gauss's law or the principle of charge conservation. In circuit theory, charge conservation leads to the postulate that it is not possible for electrical charge to accumulate at a node.

2.3 Loop

In a circuit or network, there are one or more closed paths, or *loops*. Such paths start at a node, traverse two or more elements, and terminate at the starting point. The second law of circuit analysis, Kirchhoff's voltage law, makes use of this definition. If the circuit is planar, that is, if it may be drawn on a planar surface without any wires crossing, then the loop is frequently called a mesh.

2.4 Kirchhoff's Voltage Law

Kirchhoff's voltage law, often abbreviated KVL, states that around any loop in a circuit or network,

$$\Sigma v = 0. \tag{4}$$

The law states that the sum of voltages around a closed path or loop in a circuit is equal to zero, or that the voltage between any two points is independent of the path used between the two points.

2.5 Branch

A branch of a network or circuit is an element or combination of elements that provides a path from one node to another.

2.6 Tellegen's Theorem

This theorem (Tellegen, 1952) applies to any circuit, linear or nonlinear. It states: Let N be a network of b branches and n nodes. If KVL is satisfied by the branch voltages $v_1, v_2, \ldots, v_b$ in every loop, and if KCL is satisfied by the branch currents $i_1, i_2, \ldots, i_b$ at every node, then

$$\sum_{k=1}^{b} v_k i_k = 0. \tag{5}$$

When this theorem is applied, the sign convention of Fig. 1 must be observed.

3. CIRCUIT ELEMENTS

3.1 Resistors

Electrical components in which the voltage across the device is proportional to the current through the device are known as resistors. The proportionality constant is called the resistance, and the relation is known as Ohm's law,

$$v(t) = Ri(t). \tag{6}$$

In fundamental units, resistance has units of (kg m^2)/(ampere^2 second^3). The unit of resistance is called the "ohm," and the standard symbol is Ω. The reciprocal of resistance is called conductance, and the reciprocal unit is the siemens (S). Formerly, this unit was called "mho," ohm spelled backwards, but this term is now obsolete. The graphical symbol is shown in Fig. 2(a).

The power dissipated by a resistor is given by

$$p(t) = v(t)i(t) = i^2(t)R = v^2(t)/R. \tag{7}$$

3.2 Capacitors

A capacitor is a circuit element that stores energy in an electric field. This property is called capacitance, and the unit of capacitance is the farad (F). The two approved graphical symbols are shown in Fig. 2(b). The capacitor is described by the equations

$$i(t) = C\frac{dv}{dt}, \quad v(t) = \frac{1}{C}\int_0^t i(\tau)d\tau + v(0), \tag{8}$$

$$W = \tfrac{1}{2}Cv^2, \tag{9}$$

where W = energy stored in the capacitor in joules, τ = a dummy variable representing time in seconds, and C = capacitance in farads.

3.3 Inductors

An inductor is a circuit element that stores energy in a magnetic field. This property is called inductance, and the unit of inductance is the henry (H). The approved graphical symbol is shown in Fig. 2(c). The inductor is described by the equations

$$v(t) = L\frac{di}{dt}, \quad i(t) = \frac{1}{L}\int_0^t v(\tau)d\tau + i(0), \tag{10}$$

$$W = \tfrac{1}{2}Li^2, \tag{11}$$

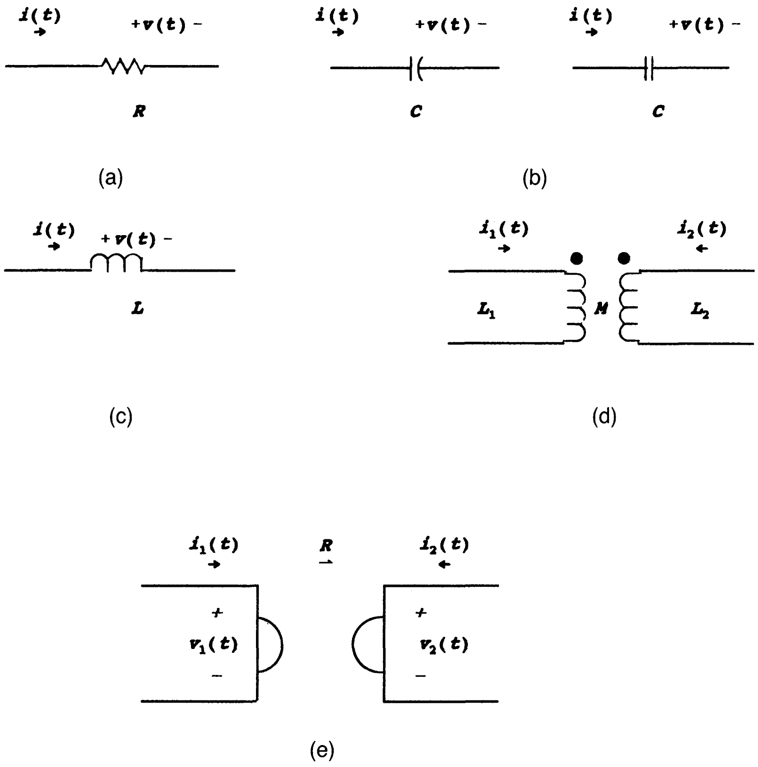

FIG. 2. Graphical symbols for circuit elements: (a) resistor, (b) capacitor, (c) inductor, (d) coupled coils, and (e) gyrator.

where W = energy stored in the inductor in joules, τ = a dummy variable representing time in seconds, and L = inductance in henrys.

3.4 Mutual Inductance

When two coils or inductors exist in close proximity, or are wound on the same coil form, a changing current in one coil will induce a voltage in the second. This effect forms the basis for transformers, a commonly employed electrical element. Figure 2(d) shows a symbolic representation of a pair of coupled coils. The dots represent the direction of the winding of the coils on the coil form. The convention is that currents entering (or leaving) dotted terminals develop additive fluxes. A pair of equations is needed to describe the coupled coil pair:

$$\begin{aligned} v_1(t) &= L_1 \frac{di_1}{dt} + M \frac{di_2}{dt}, \\ v_2(t) &= M \frac{di_1}{dt} + L_2 \frac{di_2}{dt}, \end{aligned} \tag{12}$$

where the L's represent the self inductance of the coils, and M represents the mutual inductance.

Application of the principle of conservation of energy shows that $L_1L_2 - M^2 \geq 0$. A "coefficient of coupling"

$$k^2 = \frac{M^2}{L_1L_2} \leq 1 \tag{13}$$

may be defined as a figure of merit for coupled coils.

3.5 Gyrator

Tellegen (Tellegen, 1948) postulated the fundamental nonreciprocal electrical element in 1948, and named it the "gyrator." With modern technology excellent gyrators are built. The defining equations are

$$\begin{aligned} v_1 &= Ri_2, \\ v_2 &= -Ri_1, \end{aligned} \tag{14}$$

where R is known as the "gyration resistance." A commonly used symbol (there is no standard symbol) is shown in Fig. 2(e). Gyrators are used to control power in electronic circuits, especially microwave circuits. The concept is used by circuit theorists to develop circuit and network properties.

3.6 Sources

Electrical circuits must have an energy source. In network theory, the goal is to have mathematical models that describe the sources adequately, regardless of the types of sources represented. One convenient classification is to consider independent and dependent or controlled sources. Another classification, which applies to independent sources, is to consider constant (dc) sources, sinusoidal (ac) sources, and general time-varying sources.

3.6.1 Independent Voltage Sources The independent voltage source model postulates a device or system that supplies electrical energy at a voltage unaffected by the load or device that is connected to the source. Many devices are well described by combinations of an ideal source and one or more components such as a resistor. An ideal voltage source will supply whatever current is needed by the components connected to it.

A battery is modeled as a constant (with respect to time) voltage. Often these are called "dc" sources. While dc strictly means "direct current," the term is often used as a synonym for a constant source. Figure 3(a) shows the common graphical symbol for a constant voltage source.

Much electrical energy is distributed with sinusoidal voltage sources. Often these are called "ac" (for "alternating current") sources. In Sec., 8.2, a powerful method of circuit analysis with ac sources will be developed. It is known as "phasor analysis," and is based on the fact that a sinusoidal function is described by three quantities: amplitude V_m, frequency f, and phase α. A common representation of an ac voltage source is shown in Fig. 3(b).

In many applications, a more general model is needed. In Fig. 3(c) a general time-varying voltage source is indicated. The function for $v(t)$ may be known analytically, numerically, or by its statistical properties. Methods for analysis of circuits with such sources will be developed in later paragraphs.

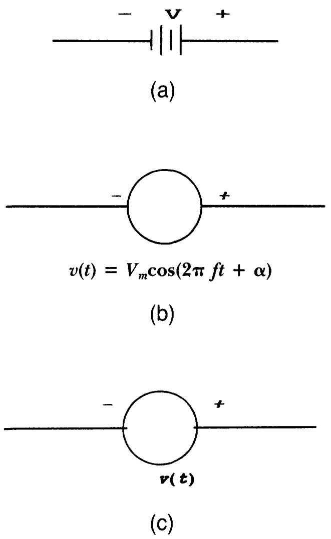

FIG. 3. Independent voltage sources: (a) constant (dc) voltage source, (b) sinusoidal (ac) voltage source, and (c) general time-varying voltage source.

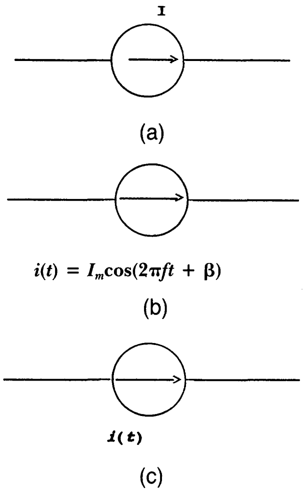

FIG. 4. Independent current sources: (a) constant (dc) current source, (b) sinusoidal (ac) current source, and (c) general time-varying current source.

3.6.2 Independent Current Sources The independent current source model postulates a device or system that supplies electrical energy at a current value unaffected by the load or device that is connected to the source. Many devices are well described by combinations of current sources and components such as resistors. An ideal current source will develop whatever voltage is needed by the components or load connected to its terminals.

A constant (dc) current source is indicated in Fig. 4(a). The current is constant with respect to time as well as connected load.

A sinusoidal (ac) current source is indicated in Fig. 4(b). As with ac voltage sources, the sinusoidal function is described by its amplitude, frequency, and phase. The phasor analysis techniques to be developed will be applicable to these functions.

A general time-varying current source is depicted in Fig. 4(c). As with voltage sources, the function $i(t)$ may be known analytically, numerically, or statistically.

3.6.3 Dependent or Controlled Sources Circuits that contain electronic devices such as amplifiers are frequently analyzed and designed using linear circuit models for the electronic devices. Such models contain a minimum of three terminals. In general there is an input terminal pair and an output terminal pair. One terminal may be common to both input and output pairs. Voltages can be measured between terminals, and there is current between terminals. If one such voltage or current is a known function of another current or voltage, then the second quantity is said to control the first, or the first is dependent on the second. Four such models can be identified. Examples of their use will be included later in examples.

3.6.3.1 Voltage-Controlled Voltage Source (VCVS). The model of Fig. 5(a) applies to the case in which the voltage across one terminal pair, v_{xz}, is proportional to the voltage across another pair, $v_{yz'}$. The proportionality factor is A, often but not always a constant. The device is, in general, not reciprocal.

3.6.3.2 Current-Controlled Voltage Source (CCVS). Figure 5(b) indicates that the voltage at one terminal pair, v_{xz}, is proportional to and controlled by the current $i_{yz'}$. The proportionality factor is shown as g, chosen because of a convention that will appear later in the section on two-ports. In general, the device being modeled is not reciprocal.

3.6.3.3 Current-Controlled Current Source (CCCS). When the current between one pair of terminals, i_{xz}, is proportional to the current between the terminals of another pair, $i_{yz'}$, the model is called a current-controlled current source, depicted in Fig. 5(c). The proportionality factor is β. In general, the device being modeled is not reciprocal.

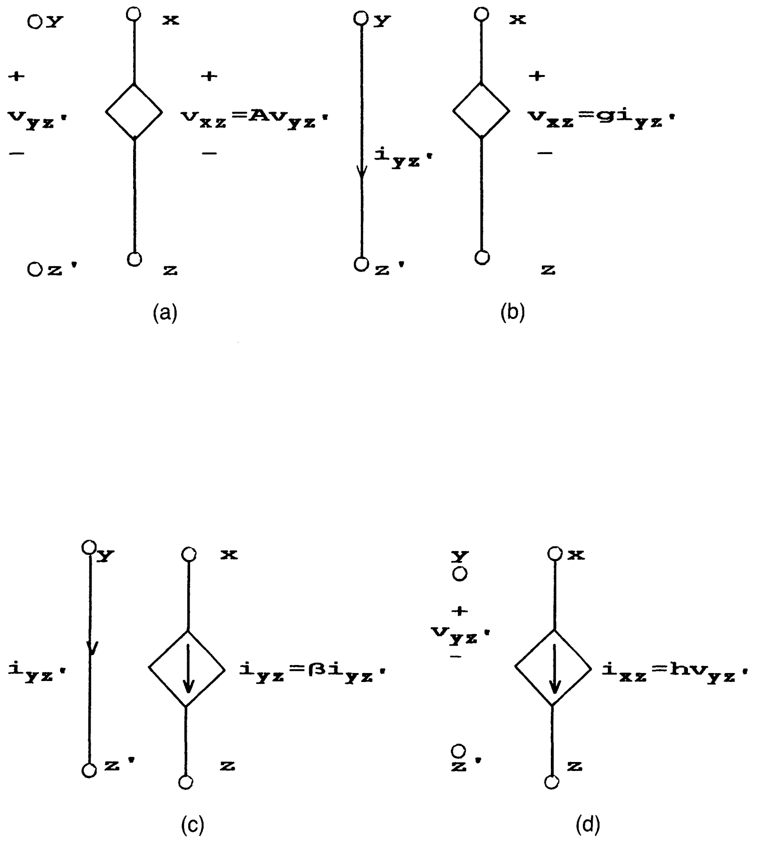

(a) (b) (c) (d)

FIG. 5. Controlled sources: (a) voltage-controlled voltage source, (b) current-controlled voltage source, (c) current-controlled current source, and (d) voltage-controlled current source.

3.6.3.4 Voltage-Controlled Current Source (VCCS). When the current between the terminals of one pair, i_{xz}, is proportional to the voltage at another pair, $v_{yz'}$, the model is called a voltage-controlled current source. The proportionality factor is called h. The model is shown in Fig. 5(d). In general, the device being modeled is not reciprocal.

4. CIRCUIT-ANALYSIS TECHNIQUES

To solve an electric circuit or network means, at a minimum, to find the voltage across and the current through each element of the network. Many methods for doing this exist, and several important methods will be discussed in this section and in following sections. The first to be discussed are methods for replacing certain combinations of elements with simpler combinations. The reason for this is to reduce the complexity of the analysis and to give the analyst greater insight into the most important aspects of the overall circuit behavior. Equivalents for combinations of elements will be presented, followed by equivalents for sources and elements.

4.1 Series and Parallel Combinations

If two or more electrical circuit elements carry the same current, the elements are said to be in *series*. If two or more elements are connected across the same voltage, the elements are said to be in *parallel*. In neither case is being connected across equal or equivalent voltages, or carrying equal or equivalent currents, sufficient. Figure 6 shows equivalent network elements for series combinations of resistors, capacitors, and inductors. In the latter case, two cases must be distinguished when there is mutual inductance between the coils. Figure 7 shows equivalent network elements for parallel combinations of resistors, capacitors, and inductors. Though derivations are not included in this article, the equivalents may be developed by using the

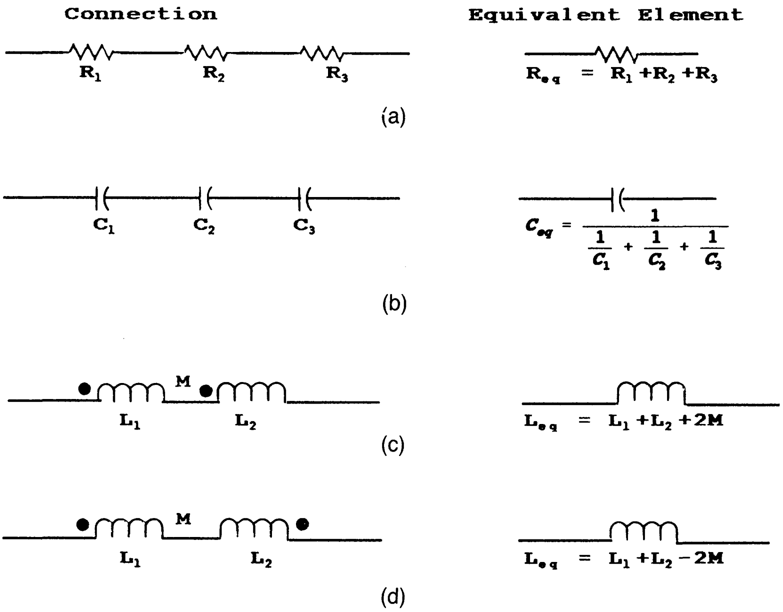
Connection
Equivalent Element
R_1
R_2
R_3
$R_{eq} = R_1 + R_2 + R_3$
(a)
C_1
C_2
C_3
$C_{eq} = \frac{1}{\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3}}$
(b)
M
L_1
L_2
$L_{eq} = L_1 + L_2 + 2M$
(c)
M
L_1
L_2
$L_{eq} = L_1 + L_2 - 2M$
(d)

FIG. 6. Equivalent elements, series circuits: (a) resistances in series, (b) capacitors in series, (c) inductors in series, additive fluxes, and (d) inductors in series, subtractive fluxes.

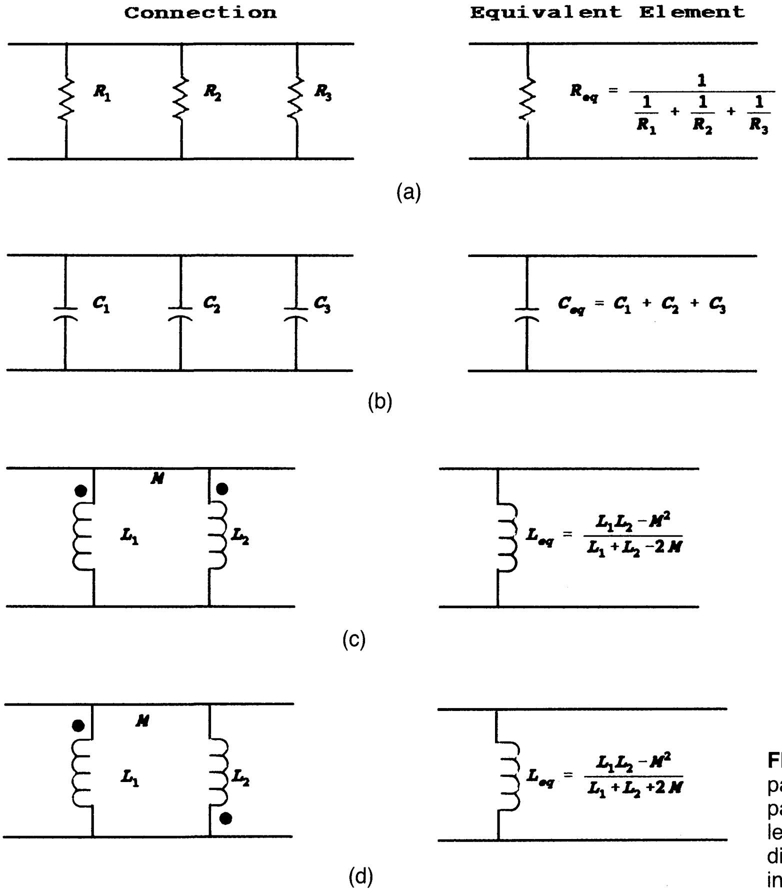
Connection
Equivalent Element
R_1
R_2
R_3
$R_{eq} = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}}$
(a)
C_1
C_2
C_3
$C_{eq} = C_1 + C_2 + C_3$
(b)
M
L_1
L_2
$L_{eq} = \frac{L_1 L_2 - M^2}{L_1 + L_2 - 2M}$
(c)
M
L_1
L_2
$L_{eq} = \frac{L_1 L_2 - M^2}{L_1 + L_2 + 2M}$
(d)

FIG. 7. Equivalent elements, parallel circuits: (a) resistors in parallel, (b) capacitors in parallel, (c) inductors in parallel, additive fluxes, and (d) inductors in parallel, subtractive fluxes.

defining equations for the elements in conjunction with Kirchhoff's laws.

4.2 Tee-Pi or Wye-Delta Conversions

Figure 8 shows two commonly used ways to connect three elements as parts of a circuit. The first, Fig. 8(a), is called a "tee" or "wye" network, the second, Fig. 8(b), a "pi" or "delta" network. When one or the other of these element combinations is apparent in a network, conversion to the other often enables the analyst to replace one combination of elements with an equivalent combination, which in turn leads to further simplifications, such as a possible series or parallel reduction. The equations are

$$\begin{aligned} R_a &= (R_1R_2 + R_1R_3 + R_2R_3)/R_3, \\ R_b &= (R_1R_2 + R_1R_3 + R_2R_3)/R_2, \\ R_c &= (R_1R_2 + R_1R_3 + R_2R_3)/R_1; \end{aligned} \tag{15}$$

$$\begin{aligned} R_1 &= R_aR_b/(R_a + R_b + R_c), \\ R_2 &= R_aR_c/(R_a + R_b + R_c), \\ R_3 &= R_bR_c/(R_a + R_b + R_c). \end{aligned} \tag{16}$$

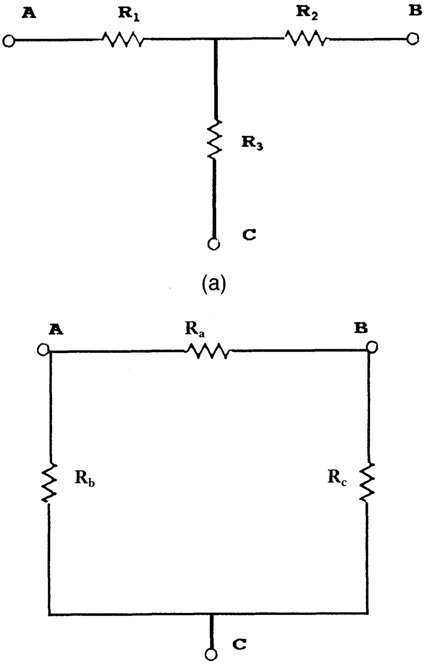

(a)

(b)

FIG. 8. Tee-pi or wye-delta conversion: (a) Tee- or wye-connected elements, and (b) pi- or delta-connected elements.

Though shown here for resistors, the equations may be generalized to impedances. More will be said about this point later in the article.

4.3 Thévenin Theorem

In many applications, an electrical network is used to control or modify an electrical signal delivered to some device. A common example is an amplifier-loudspeaker combination. An analyst might be interested in understanding how various loudspeakers perform with a given amplifier, but does not want to make repeated analyses of the amplifier, since the principal interest is in the behavior of the loudspeaker in its setting. This analyst needs a way to model the amplifier in a simplified but equivalent manner, and to use this model in conjunction with the electrical models of the various loudspeakers being considered. More specifically, since the loudspeaker normally has two terminals, then it would be convenient to represent the amplifier as a device with two terminals (a terminal pair) and other necessary components. A pair of theorems developed by Thévenin and Norton provide a means for doing this.

The Thévenin theorem states that, at a terminal pair, any linear network may be replaced by a series combination of a voltage source and a resistance (or impedance—to be discussed later); see Fig. 9(a). The voltage is equal to the voltage at the terminal pair when the external load is removed or open-circuited, and the resistance is equal to the resistance calculated or measured at the terminal pair with all of the independent sources de-energized.

De-energization means, in this context, that the source voltages and currents of *independent* sources are set to zero, but that the source resistances (impedances) are unchanged. *Controlled* or *dependent* sources are not de-energized or modified in any way.

4.4 Norton Theorem

This theorem, which is a companion to the Thévenin theorem, states that, at a terminal pair, any linear network may be replaced by the parallel combination of a current source and a resistance (or impedance); see Fig. 9(b). The current is equal to the current that flows through the short circuit when the external

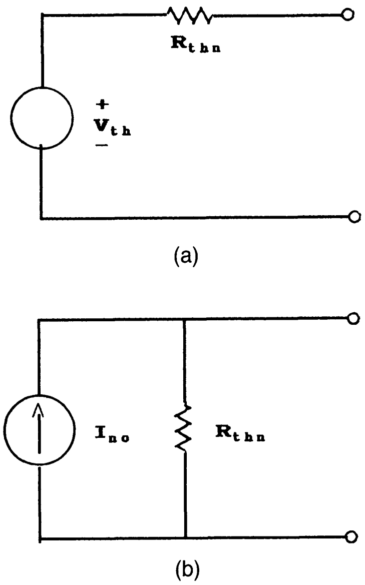

(a)

(b)

FIG. 9. (a) Thévenin and (b) Norton equivalent circuits.

load is short circuited. The resistance is equal to the resistance measured or calculated at the terminal pair with all independent sources de-energized. As before, de-energization means that independent source voltages and currents are set to zero but source resistances are unchanged. Dependent or controlled sources are not de-energized.

4.5 Thévenin–Norton Theorem Comparisons

The similarity of these two theorems suggests that they are closely related. Some authors combine them and speak of a Thévenin–Norton theorem. Since the definitions of the equivalent impedances are identical, these impedances themselves are identical. If the Thévenin equivalent for a circuit or network is known, then the Norton equivalent current is found from the relation

$$V_{\text{th}} = I_{\text{no}} R_{\text{thn}}. \tag{17}$$

4.6 Thévenin–Norton Theorem Example

Figure 10(a) shows a network with a controlled source. The Thévenin voltage is calculated, in this case, directly, without any modification of the circuit. The current through the 5-Ω resistor is zero because of the open circuit at terminals A and B. Thus, $V_{CB} = 15$ V. The resulting current through the dependent current source is $(2\ \Omega^{-1})$ (15 V) = 30 A. From this result, $V_{AC} = (2\ \Omega)(30\ \text{A}) = 60$ V. Thus, $V_{AB} = 15 + 60 = 75$ V, and since this is the open-circuit voltage at the terminals, it is the Thévenin voltage.

In Fig. 10(b), the equivalent resistance is calculated. To begin, the independent source of 15 V is temporarily set equal to zero. Since the resistance of an ideal voltage source is zero, the voltage source is replaced by a short circuit. The Thévenin resistance can be calculated by applying, conceptually, a test source at the output terminal pair. The ratio of the test voltage to the resulting current is the needed resistance. Since the circuit is linear, any convenient value may be chosen for the source. Call the voltage source applied V_T, and the resulting current I_T. The current through the 5-Ω resistor is, by Kirchhoff's current law, also I_T. Thus, $V_{CB} = (5\ \Omega)I_T$. The current through the 2-Ω resistor is $I_T + (2\Omega^{-1})V_{CB} = 11I_T$. Therefore, $V_{AC} = (2\ \Omega)(11I_T) = (22\ \Omega)I_T$, and $V_T = V_{AB} = V_{CB} + V_{AC} = (27\ \Omega)I_T$. Finally, the ratio between this current and voltage is 27 Ω, which is the Thévenin–Norton resistance.

Figure 10(c) shows the circuit prepared for Norton current determination. The output terminal pair is short circuited, conceptually, and the current through this short circuit is calculated from the modified circuit. (It can also be measured when the current can be short circuited without damage.) The resulting current is labeled I_{no}. In this circuit, $V_{CB} = 15\ \text{V} - (5\ \Omega)I_{\text{no}}$. Writing a Kirchhoff's voltage law equation around the closed path gives $(5\ \Omega)I_{\text{no}} + (2\ \Omega)[I_{\text{no}} - (2\ \Omega^{-1})V_{CB}] - 15\ \text{V} = 0$. A simultaneous solution of these two equations gives $I_{\text{no}} = 25/9$ A, which is the required value.

Figure 10(d) shows the combination of these results into both the Norton and Thévenin equivalent circuits. The work may be checked by computing the ratio of the Thévenin voltage to the Norton current, which is 27 Ω, the same value that was computed in Fig. 10(d).

4.7 Ladder Networks

A ladder network is a special but frequently occurring circuit configuration. An example, with resistors only, is shown in Fig.

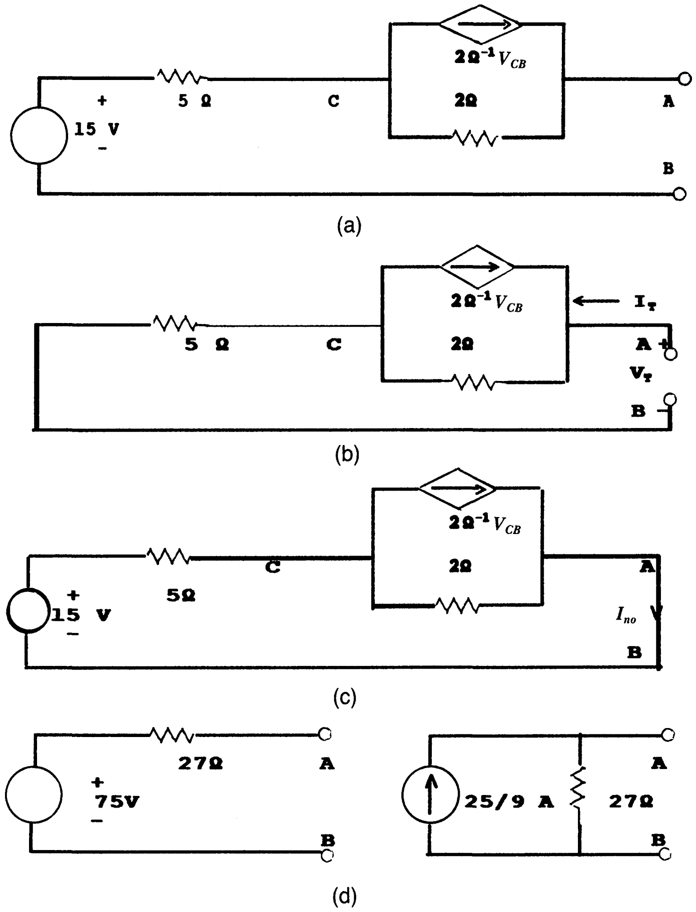

(a)

(b)

(c)

(d)

FIG. 10. Thévenin–Norton circuit determination. (a) Example circuit; (b) Thévenin–Norton resistance calculation; (c) illustrating Norton current calculation, and (d) Thévenin and Norton equivalent circuits.

11(a). The principal characteristics are that there is one terminal, labeled 2-2′, that is common to input and output, and that no node is "bridged." There would never be an element connecting, for example, terminal 1 to 4. Two special analysis methods are described for this configuration.

The first method starts with use of series and parallel combinations alternately to reduce the network to a single equivalent element and a source. To do this, the analyst starts at the element farthest from the source, and works toward the source. Currents through and voltages across each element are then found by reversing the process, starting at the source, and working toward the load.

The second method uses the principle of superposition. The analyst assumes a current (or voltage) at the output, computes the voltage (or current) at the output, and then uses KCL and KVL to work back toward the source. Eventually a source magnitude is determined, which is probably different from the actual source. However, because of superposition, the ratio of the actual current (or voltage) to the assumed current (or voltage) is equal to the ratio of the actual source to the source value resulting from the assumption.

Figure 11(b) shows an example of this solution method. In this example, voltages are stated with respect to the reference node. Assume a current of 1 A through the 9-Ω resistor, giving a voltage at terminal 2 of 9 V. The current through the 7-Ω resistor is also 1 A, so that the voltage at terminal 4 is 16 V. Then, the current through the 16-Ω resistor is 1 A, and the current through the 4-Ω resistor is 2 A. Next, the voltage at terminal 3 is found to be $16 + 2(4) = 24$ V. Thus, the current through the 6-Ω resistor is 4 A, leading to a current of 6 A through the 2-Ω resistor. Finally, the resulting voltage at terminal 1 would be 36 V. Since the actual voltage at terminal 1 is

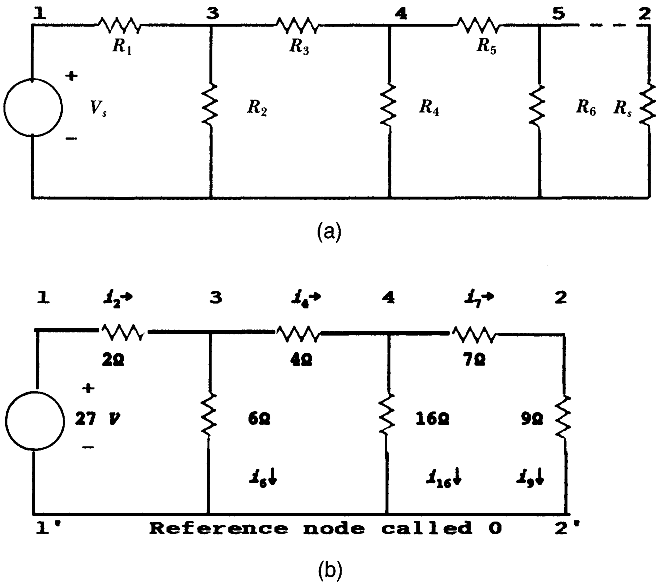

FIG. 11. Ladder networks. (a) General ladder network, (b) sample ladder network calculation.

27 V, the currents in all the resistors are scaled by 27/36 = $\frac{3}{4}$. The solution for the actual current in the 9-Ω resistor is 0.75 A.

5. FORMAL CIRCUIT-ANALYSIS METHODS

5.1 Nodal and Loop Analysis

A combination of *b* branches or elements interconnected to form a circuit or network requires, for a complete analysis, the determination of *b* voltages and *b* currents, or 2*b* quantities. Thus, 2*b* equations are needed. Half of these are provided by the defining equations of the branches, e.g., Ohm's law. The remaining equations are provided by systematic application of Kirchhoff's laws. There are two methods, one based on each law. Both use the definitions of nodes and loops given earlier.

Suppose that a circuit having *b* elements has *n* (essential) nodes. Kirchhoff's current-law equations can be written at each, but one will be redundant, because it is the linear sum of the remaining $n - 1$ equations. Thus, in a circuit with *n* (essential) nodes, $n - 1$ independent KCL equations can be written. To complete the analysis, $b - (n - 1)$ KVL equations must be written. In general, this is possible, though in some cases, especially when the networks are nonplanar, care must be taken to ensure that the equations are independent. In practice, it is usually sufficient to use one set or the other. The choice depends on which technique yields the smaller number of equations, which technique yields the most useful and usable information, and the preference of the analyst.

5.2 Nodal Analysis

Figure 12 shows a typical node in a circuit that is isolated for study. The voltage at this node is determined with respect to some reference node in the circuit. The voltage at the

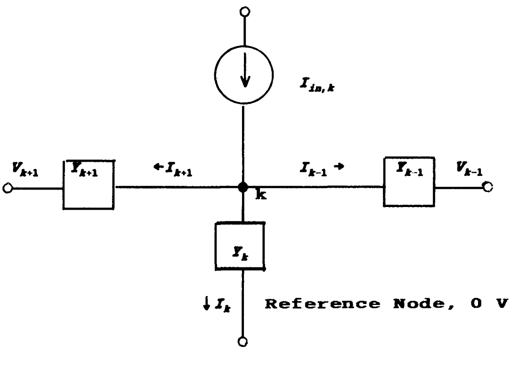

FIG. 12. Isolated node for KCL equations.

reference is arbitrarily called 0 V. One element is shown connected from the node in question to the reference node, though this is not necessary. Two other elements and a current source are shown.

Kirchhoff's current law written at this node yields

$$I_{k-1} + I_k + I_{k+1} = I_{in,k}. \tag{18}$$

A typical expression for the current through each element can be written

$$I_{k+i} = (V_k - V_{k+i})Y_{k+i}, \tag{19}$$

where the Y_k represent the branch relations, and are written in *admittance* form. Once all of the possible equations are written, simplified, and organized into matrix form, the result is

$$\begin{bmatrix} Y_{11} + Y_{12} + Y_{13} + \cdots & -Y_{12} & -Y_{13} & \cdot \\ -Y_{21} & Y_{21} + Y_{22} + Y_{23} + \cdots & -Y_{23} & \cdot \\ -Y_{31} & -Y_{32} & Y_{31} + Y_{32} + Y_{33} + \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ \cdot \\ \cdot \end{bmatrix}. \tag{20}$$

The square matrix is a complete description of the circuit. The column matrix (vector) of voltages represents the dependent variables, while the column matrix of currents represents the independent variables.

5.3 Nodal Analysis with Controlled Sources

When a branch contains a controlled source of any of the four types, it is usually convenient to use the Thévenin–Norton theorem to convert the controlled source to a voltage-controlled current source (unless it is already in this form). When this is done, the right-hand side of the previous equation will contain the dependent variables (voltages) as well as the independent current sources. The dependent-variable terms should then be transposed to the left-hand side of the matrix equation. A major result of this effort is that the square matrix is no longer symmetric.

5.4 Loop Analysis

Figure 13 shows a typical loop in a circuit that is isolated for study. A loop current is defined as indicated, and loop currents from neighboring loops are also indicated. As indicated, each element current is the difference between two loop currents. This formulation of the loop presumes that the network is planar, i.e., it can be drawn in a plane without wires crossing. If the network is not planar, then some element currents will be composed of three or more currents, and some signs in the final result may change. Kirchhoff's voltage law written around the loop states that

$$V_{k-1} + V_k + V_{k+1} + V_{k+2} = V_{in,k}. \tag{21}$$

An expression for each of the voltages is

$$V_{k+i} = (I_k - I_{k+i})Z_{k+i}, \tag{22}$$

where the Z_k describe the elements and are called *impedances*. When all of the possible equations are written and simplified, and the results are organized, a general result is

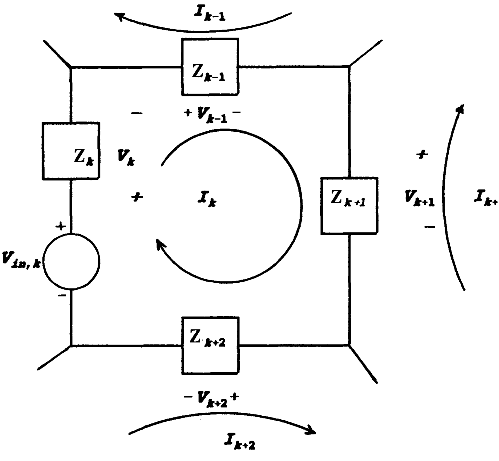

FIG. 13. Isolated loop for KVL equations.

$$\begin{bmatrix} Z_{11}+Z_{12}+Z_{13}+\cdots & -Z_{12} & -Z_{13} & \cdot \\ -Z_{21} & Z_{21}+Z_{22}+Z_{23}+\cdots & -Z_{23} & \cdot \\ -Z_{31} & -Z_{32} & Z_{31}+Z_{32}+Z_{33}+\cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ \cdot \\ \cdot \end{bmatrix}. \quad (23)$$

The square matrix is a complete description of the network, while the column matrix or vector of currents shows the dependent variables, the loop currents, and the column matrix of voltages shows the independent variables or forcing functions in the various loops.

5.5 Loop Analysis with Controlled Sources

When an element contains a controlled source, it is convenient to use the Thévenin–Norton theorem to convert all such sources to current-controlled voltage sources. When this is done and the equations are written, the right side of Eq. (23) will contain the dependent variables (loop currents) in addition to the independent voltage sources. These dependent-variable terms should then be transposed to the left side of the matrix equation. The resulting square matrix will no longer be symmetric, but it will describe the network completely.

5.6 Solution of Node-Voltage and Loop-Current Equations

If the sources are dc sources and the elements are resistive, the elements in the square matrices of Eqs. (20) and (23) are numerical terms, as are the vectors representing the dependent and independent variables. Calculators and computer programs that handle such equations are normally used to obtain the numerical solutions for the dependent variables, especially when there are three or more equations.

When the sources are sinusoidal (ac) and a steady-state solution is required, then all of the terms in the square matrices are complex numbers, and both the dependent and independent variables are phasors, also complex numbers. These concepts are discussed in Sec. 8.2. Computer programs and calculators for handling complex arithmetic facilitate the numerical solutions for the dependent variables.

If the terms in square matrix are generalized admittances or impedances, as discussed in immediately following paragraphs on the solution of circuit equations by Laplace transforms, then solutions will be quotients of polynomials in the Laplace variable s.

6. LAPLACE TRANSFORMS IN CIRCUIT ANALYSIS

6.1 Introduction to Laplace Transforms

Since linear circuits or networks are described by linear differential equations, most often with constant coefficients, analysts can use any of the techniques available for solving such equations. A powerful technique is the Laplace transform. This technique, an important part of operational calculus, provides the analyst a powerful method for finding both steady-state and transient components of the solution simultaneously. The technique requires transformation of circuits and sources through application of certain theorems. Solution for unknown dependent variables is carried out algebraically, and inverse transforms are computed. It is necessary, however, to know initial conditions, the energy stored in capacitors and inductors.

6.2 Definition of a Laplace Transform

If a function of time $f(t)$ is known, and defined for $t \geq 0$, then the (single-sided) Laplace transform is given by

$$\mathscr{L}f(t) = F(s) = \int_{0^-}^{\infty} f(t)\epsilon^{-st}dt, \quad (24)$$

where s is a complex variable,* $s = \sigma + j\omega$,

---

*In these equations, j is used for the imaginary operator, $j = \sqrt{-1}$. Mathematicians and physicists typically use i for the imaginary operator, while electrical engineers use j. This choice is made to avoid confusion with the use of i, which is the standard symbol for current.

Table 1. Laplace transform theorems.

Operation	Theorem
Derivative	$\mathcal{L}\frac{df(t)}{dt} = sF(s) - f(0^-)$
nth-order derivative	$\mathcal{L}\frac{d^nf}{dt^n} = s^nF(s) - s^{n-1}f(0^-) - \cdots - f^{(n-1)}(0^-)$
Integral	$\mathcal{L}\int_{0^-}^{t} f(x)dx = \frac{F(s)}{s}$
Time shift	$\mathcal{L}[f(t - t_0)u(t - t_0)] = \epsilon^{-t_0 s}F(s)$ where u is the unit step function
Frequency shift	$\mathcal{L}\epsilon^{-at}f(t) = F(s + a)$
Frequency scaling	$\mathcal{L}f(at) = \frac{1}{a}F\left(\frac{s}{a}\right), a > 0$
Initial value	$\lim_{t\to 0} f(t) = \lim_{s\to\infty} sF(s)$ provided the limit exists
Final value	$\lim_{t\to\infty} f(t) = \lim_{s\to 0} sF(s)$ provided the limit exists
Constant multiplier	$\mathcal{L}kf(t) = k\mathcal{L}F(s)$
Addition	$\mathcal{L}[a_1 f_1(t) + a_2 f_2(t)] = a_1F_1(s) + a_2F_2(s)$ a_1 and a_2 are constants
Convolution	$F(s) = F_1(s)\,F_2(s), f(t) = \int_0^t f_1(\tau)f_2(t - \tau)d\tau = \int_0^t f_2(\tau)f_1(t - \tau)d\tau$

chosen so that the integral will converge. In turn, σ is the real part of the variable, and ω is the imaginary part. The frequency of sinusoidal excitation functions, ω, is measured in rad/s.

6.3 Laplace Transform Theorems

If the function $f(t)$ has the Laplace transform $F(s)$, then the theorems of Table 1 apply. In these theorems, the term $f(0^-)$ represents the initial condition, or the value of f at $t = 0$.

6.4 Laplace Transform Pairs

Table 2 presents a listing of 12 common time functions and their Laplace transforms. These are sufficient for much of the analysis that is necessary. References include more extensive tables.

It should be noted that the time functions and their Laplace transforms form a biunique pair. In other words, a given time function has one and only one transform, and a given transform has one and only one inverse.

Table 2. Laplace transform pairs.

Name	Time function $f(t)$, $t > 0^-$	Laplace transform $F(s)$
Unit impulse	$\delta(t)$	1
Unit step	$u(t)$	$1/s$
Unit ramp	t	$1/s^2$
nth-order ramp	t^n	$n!/s^{n+1}$
Exponential	ϵ^{-at}	$1/(s + a)$
Damped ramp	$t\epsilon^{-at}$	$1/(s + a)^2$
Cosine	$\cos\omega t$	$s/(s^2 + \omega^2)$
Sine	$\sin\omega t$	$\omega/(s^2 + \omega^2)$
...	$(1/2\omega^3)(\sin\omega t - \omega t \cos\omega t)$	$1/(s^2 + \omega^2)^2$
...	$(1/2\omega)t \sin\omega t$	$s/(s^2 + \omega^2)^2$
Damped cosine	$\epsilon^{-at} \cos\omega t$	$(s + a)/[(s + a)^2 + \omega^2]$
Damped sine	$\epsilon^{-at} \sin\omega t$	$\omega/[(s + a)^2 + \omega^2]$

7. CIRCUIT ANALYSIS WITH ENERGY-STORAGE ELEMENTS

7.1 Initial Conditions

If a circuit has initial energy stored in it, that is, if any of the capacitors is charged when the circuit is energized or reconfigured, or if any of the inductors has a nonzero initial current when the circuit is energized or reconfigured, then these conditions must be determined before a complete analysis can be completed. In general, this will require knowledge of the circuit just before the circuit to be analyzed is connected. Normal circuit-analysis methods can be used to determine the initial conditions.

7.2 An Example with Nodal Analysis

Figure 14 shows a network with three (essential) nodes and three energy-storage elements. The earlier discussion shows that two independent KCL equations can be written, and are needed. The voltage source is a unit step function, and it is assumed that there is no initial energy stored in the network. When the defining equations for capacitors and inductors of Secs. 3.2 and 3.3 are combined with the theorems of Table 1, it is found that, for the inductor, $I(s) = V(s)Y(s) = V(s)(1/sL)$, and for the capacitors, $I(s) = V(s)sC$. For the resistors, $I(s) = V(s)\ (1/R)$. The terms from Table 1 that refer to initial conditions are zero.

The following treatment shows the writing of two Kirchhoff's current-law Laplace-domain equations for the network and a complete solution for $v_2(t)$. The two KCL equations are

$$V_3\left(1 + sC_1 + \frac{1}{sL}\right) - V_2\left(\frac{1}{sL}\right) = \frac{1}{s},$$
$$-V_3\left(\frac{1}{sL}\right) + V_2\left(1 + sC_2 + \frac{1}{sL}\right) = 0. \qquad (25)$$

The solution for $V_2(s)$ is

$$V_2(s) = \left(\frac{1}{LC_1C_2}\right) s^{-1}\left\{s^3 + s^2\left(\frac{1}{C_1} + \frac{1}{C_2}\right) + s\left[\frac{1}{C_1C_2} + \frac{1}{L}\left(\frac{1}{C_1} + \frac{1}{C_2}\right)\right] + \frac{1}{LC_1C_2}\right\}^{-1}. \qquad (26)$$

The specific results will depend on the numerical values chosen for the two capacitors and the inductor. If $L = 1.345$ H, $C_1 = 0.680$ F, and $C_2 = 4.375$ F, Eq. 26 becomes

$$V_2(s) = \frac{0.25}{s(s^3 + 1.70s^2 + 1.60s + 0.50)}. \qquad (27)$$

In general, the denominator polynomial must be factored, and computer techniques are employed in all but the simplest cases. When the denominator of Eq. (27) is factored, the result is

$$V_2(s) = \frac{0.25}{s(s + 0.50)(s^2 + 1.20s + 1.00)}. \qquad (28)$$

Expansion into partial fractions provides a convenient method for finishing the solution. In this case, the result is

$$V_2(s) = \frac{0.50}{s} + \frac{-0.769}{s + 0.50} + \frac{0.135 - j0.139}{s + 0.60 + j0.80} + \frac{0.135 + j0.139}{s + 0.60 - j0.80},$$
$$V_2(s) = \frac{0.50}{s} - \frac{0.769}{s + 0.50} + \frac{0.270\,(s + 0.60) - 0.278(0.80)}{(s + 0.60)^2 + (0.80)^2}. \qquad (29)$$

Finally, the inverse transform and solution is

$$v_2(t) = 0.50 - 0.77\,\epsilon^{-0.50t} + [0.27\cos(0.80t) - 0.28\sin(0.80t)]\,\epsilon^{-0.60t}. \qquad (30)$$

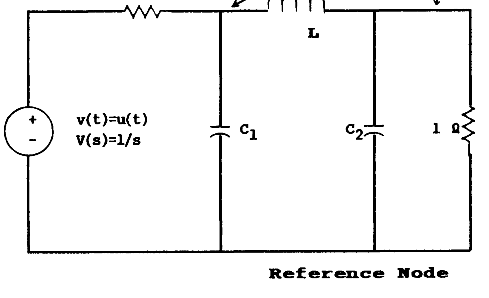

FIG. 14. Nodal analysis without initial conditions.

7.3 An Example with Loop Analysis and Initial Conditions

Figure 15 shows a circuit with a voltage source, a load resistor, and three energy-storage elements. The excitation function is a unit step function. Assume that the initial energy stored in the capacitor is such that the initial voltage is +4 V, and that the initial energy in the $\frac{1}{6}$-H inductor is such that the initial current is 3 A. Table 1, in conjunction with the defining relations for an inductor, shows that, for the $\frac{5}{6}$-H inductor, $V(s) = I(s)(\frac{5}{6}s)$, while for the second inductor, $V(s) = \frac{1}{6}[sI(s) - 3]$. For the capacitor,

$$V(s) = \frac{I(s)}{(12/25)s} + \frac{4}{s},$$

The $4/s$ term represents the initial condition, and in Table 2, it represents the integral from $t = 0-$ to $t = 0+$. (Some authors define the integral from $t = 0+$ and show the initial condition explicitly.)

The equations become:

$$\frac{1}{s} = \frac{5s}{6} I_1 + \frac{1}{(12/25)s}(I_1 - I_2) + \frac{4}{s}, \tag{31}$$

$$0 = \frac{1}{(12/25)s}(I_2 - I_1) - \frac{4}{s} + \frac{1}{6}sI_2 - \frac{1}{6}3 + I_2(1). \tag{32}$$

Solution for $I_2(S)$ gives, after some manipulation and simplification,

$$I_2(s) = \frac{3s^3 + 24s^2 + 7.5s + 15}{s^4 + 6s^3 + 15s^2 + 15s}. \tag{33}$$

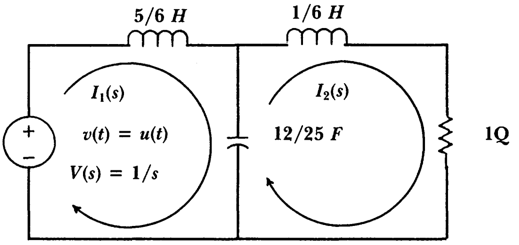

FIG. 15. Loop analysis with initial conditions.

Expansion in partial fractions gives

$$I_2 = \frac{1}{s} - \frac{11.64}{s + 2.32} + \frac{6.82 - j1.16}{s + 1.84 - j1.75} + \frac{6.82 + j1.16}{s + 1.84 + j1.75}, \tag{34}$$

$$I_2 = \frac{1}{s} - \frac{11.64}{s + 2.32} + \frac{13.64\,(s + 1.84) + 2.32(1.75)}{(s + 1.84)^2 + (1.75)^2}. \tag{35}$$

The inverse transform of Eq. (35) gives the time-domain solution that is required:

$$i_2(t) = 1 - 11.64\,\epsilon^{-2.32t} + \epsilon^{-1.84t}(13.64 \cos 1.75t + 2.32 \sin 1.75t). \tag{36}$$

It should be noted that this solution does satisfy the initial condition of 3 A in the $\frac{1}{6}$-H inductor and thus in the resistor.

7.4 State-Variable Methods

A technique for solving virtually all types of circuit problems is known as the state-variable method. It is valid for nonlinear and time-varying networks as well as for linear networks. It is readily adapted to digital computer solutions. The technique has become common both in control systems analysis and in network or circuit theory, and requires the analyst to put the circuit equations into a set of first-order differential equations. Invariably matrix analysis methods are used, as will be done here. The technique will be introduced in this article, and the references give much additional information.

Equations written in state-variable form are usually expressed as

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t), \end{aligned} \tag{37}$$

where $\mathbf{x}$ is a matrix (vector) of dependent or response variables, $\mathbf{u}$ is a matrix of forcing or excitation functions, $\mathbf{A}$ is a matrix that describes the network, $\mathbf{B}$ is a matrix showing input connections to circuit, $\mathbf{C}$ is a matrix relating output variables to dependent variables, $\mathbf{D}$ shows direct input-output connections, and $\mathbf{y}$ is the matrix of output variables.

The key term in Eq. (37) is the matrix **A**, as it contains the information that relates currents and voltages in all network elements. If the network is linear and time-invariant, then all terms of the matrix **A** are constants. Otherwise these terms are nonlinear, time-varying, or both. The matrix **A** is square.

In this section, the formulation and solution of the equations for a commonly occurring case will be considered, and the technique will be presented by example. There are situations in which this method may not apply without modification. Since derivatives of dependent variables are needed, and since the basic element equations for inductors and capacitors require derivatives of capacitor voltages and inductor currents, the formulation of state-variable equations focuses on these quantities as an appropriate choice of state variables.

Figure 16(a) shows a network with a voltage source, a current source, two resistors, and two energy-storage elements. For this problem, the output signal is taken as the voltage across the resistor R. A good technique (Rohrer, 1970) for writing the equations is to draw an auxiliary circuit, as Fig. 16(b), in which the inductors are replaced by independent current sources, and the capacitors are replaced by independent voltage sources. Then, Kirchhoff's laws are written for the network. Generally, a mixture of the current and voltage laws is needed. For the circuit under consideration,

$$\begin{aligned} i_c(t) &= -Gv_c(t) + i_L(t) + i_{in2}(t), \\ v_L(t) &= -v_c(t) - Ri_L(t) + v_{in1}(t). \end{aligned} \tag{38}$$

After the element relations for the capacitor and inductor are substituted for the inductor voltage and capacitor current terms, and the equations are rewritten and combined into matrix form, the circuit description is

$$\begin{aligned} \begin{bmatrix} \dot{v}_C(t) \\ \dot{i}_L(t) \end{bmatrix} &= \begin{bmatrix} -\dfrac{G}{C} & +\dfrac{1}{C} \\ -\dfrac{1}{L} & -\dfrac{R}{L} \end{bmatrix} \begin{bmatrix} v_C(t) \\ i_L(t) \end{bmatrix} \\ &\quad + \begin{bmatrix} \dfrac{1}{C} & 0 \\ 0 & \dfrac{1}{L} \end{bmatrix} \begin{bmatrix} i_{in2}(t) \\ v_{in1}(t) \end{bmatrix}, \\ v_0 &= [0 \quad R] \begin{bmatrix} v_C \\ i_L \end{bmatrix} + 0. \end{aligned} \tag{39}$$

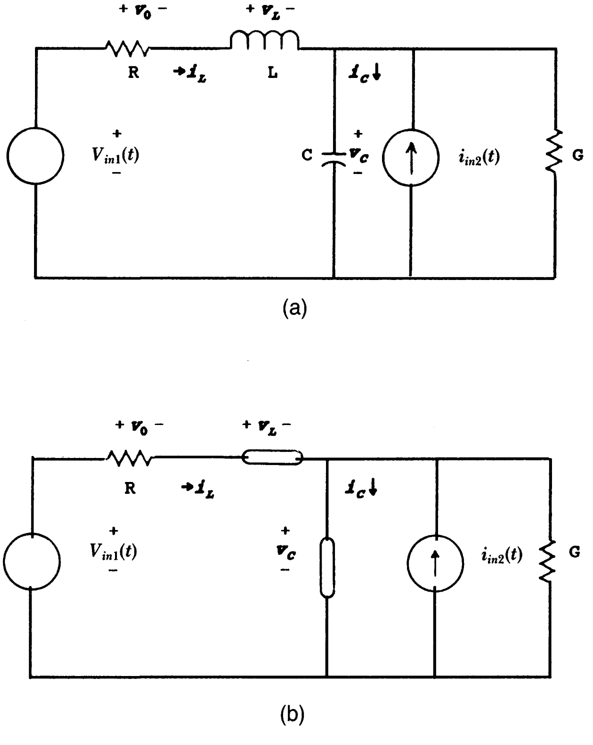

FIG. 16. Circuit to illustrate state-variable analysis techniques. (a) Original circuit, and (b) circuit with energy storage elements replaced by sources.

This example shows a technique that will work in most commonly encountered situations. If, however, a circuit or network contains a node at which the only branches connected are inductors and current sources, or if it contains a loop with only voltage sources and capacitors, then additional care is needed. Methods for dealing with this situation are discussed in many more advanced circuit theory textbooks, such as Rohrer (1970), Balabanian *et al.* (1969), or Desoer and Kuh (1969).

7.5 Solution of State-Variable Equations

Equations such as Eqs. (39) can be solved by a variety of methods. Laplace transform methods can be used. Extensions of classical differential equation solutions are possible, and will be illustrated here. Let $C = \frac{1}{3}$ F, $G = \frac{2}{3}$ S, $L = 1$ H, and $R = 6\ \Omega$. With these substitutions, Eqs. (39) become

$$\begin{bmatrix} \dot{v}_C(t) \\ \dot{i}_L(t) \end{bmatrix} = \begin{bmatrix} -2 & 3 \\ -1 & -6 \end{bmatrix} \begin{bmatrix} v_C(t) \\ i_L(t) \end{bmatrix} + \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} i_{in2}(t) \\ v_{in1}(t) \end{bmatrix},$$
$$v_0 = [0 \quad 6] \begin{bmatrix} v_C \\ i_L \end{bmatrix} + 0; \tag{40}$$

that is,

$$\mathbf{A} = \begin{bmatrix} -2 & 3 \\ -1 & -6 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix},$$
$$\mathbf{C} = [0 \quad 6], \mathbf{D} = 0.$$

The zero-input response, which corresponds to the complementary function in conventional differential-equation solution techniques, is obtained from Eqs. (40) when the input functions are zero, or $u = 0$. This solution is

$$\mathbf{x}(t) = \epsilon^{\mathbf{A}(t)}\mathbf{x}(0), \quad t \geq 0;$$
$$\mathbf{y}(t) = \mathbf{C}\epsilon^{\mathbf{A}(t)}\mathbf{x}(t), \quad t \geq 0. \tag{41}$$

The exponential matrix in Eq. (41) is found from the eigenvalues of the matrix **A**, which in turn are found from its characteristic polynomial

$$\Delta(\lambda) = \det|\mathbf{A} - \lambda\mathbf{I}|. \tag{42}$$

Here **I** is the identity matrix, and from the values given, the characteristic polynomial is found to be

$$\Delta(\lambda) = \begin{vmatrix} -2-\lambda & 3 \\ -1 & -6-\lambda \end{vmatrix}$$
$$= \lambda^2 + 8\lambda + 15 = (\lambda + 3)(\lambda + 5). \tag{43}$$

These eigenvalues are distinct, and thus the exponential matrix function can be written as

$$\epsilon^{\mathbf{A}t} = \mathbf{f}(\lambda) = \alpha_0(t)\mathbf{I} + \alpha_1(t)\mathbf{A}, \tag{44}$$

where **I** is again the identity matrix. According to the Cayley–Hamilton (Balabanian *et al.*, 1969) theorem, any function of a square matrix can be satisfied as a scalar function by its eigenvalues. This fact is used to determine

$$f(\lambda) = \epsilon^{\lambda t} = \alpha_0(t) + \lambda\alpha_1(t), \tag{45}$$

where λ is an eigenvalue of **A**. When the eigenvalues -3 and -5, respectively, are substituted into Eq. (45), the results are

$$\epsilon^{-3t} = \alpha_0(t) - 3\alpha_1(t),$$
$$\epsilon^{-5t} = \alpha_0(t) - 5\alpha_1(t). \tag{46}$$

Thus,

$$\alpha_0(t) = (5\epsilon^{-3t} - 3\epsilon^{-5t})/2,$$
$$\alpha_1(t) = (\epsilon^{-3t} - \epsilon^{-5t})/2. \tag{47}$$

The exponential matrix of Eq. (44) is then found, and from this, the zero-input response is found as

$$\begin{bmatrix} v_C(t) \\ i_L(t) \end{bmatrix} = \begin{bmatrix} \dfrac{3\epsilon^{-3t} - \epsilon^{-5t}}{2} & \dfrac{3\epsilon^{-5t} - 3\epsilon^{-3t}}{2} \\ \dfrac{\epsilon^{-3t} - \epsilon^{-5t}}{2} & \dfrac{3\epsilon^{-5t} - \epsilon^{-3t}}{2} \end{bmatrix} \begin{bmatrix} v_C(0) \\ i_L(0) \end{bmatrix}, \tag{48}$$

where the term $[v_c(0) \quad i_L(0)]^t$ is the initial state.

The preceding term corresponds to the familiar complementary function or homogeneous solution for differential equations. It is also necessary to compute the particular integral or particular solution. This is done with the generalized convolution integral, Table

1, which gives

$$\mathbf{x}(t) = \int_{t_0}^{t} \epsilon^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)d\tau. \tag{49}$$

The complete solution for $\mathbf{x}(t)$ is then the sum of Eqs. (41) and (49). The complete output expression $\mathbf{y}(t)$ is given by

$$\mathbf{y} = \mathbf{C}\int_{0}^{t} \epsilon^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)d\tau + \mathbf{D}\mathbf{u}(t). \tag{50}$$

If the current source is a step function of strength 2 A, and the voltage source is a step function of strength 4 V, then the excitation function becomes

$$\mathbf{u}(t) = \begin{bmatrix} 2 \\ 4 \end{bmatrix} u(t).$$

After some work, solution of Eq. (49) gives

$$\begin{bmatrix} v_C(t) \\ i_L(t) \end{bmatrix} = \begin{bmatrix} 3.2 - 5.0\,\epsilon^{-3t} + 1.8\,\epsilon^{-5t} \\ 2/15 - 1.8\,\epsilon^{-5t} + 5/3\,\epsilon^{-3t} \end{bmatrix} u(t) \tag{51}$$

as the zero-state response, and the complete solution for the network is the sum of Eqs. (48) and (51).

8. SINUSOIDAL STEADY-STATE ANALYSIS

8.1 Description of the Technique

An important forcing function to consider is a sinusoidal signal. One reason is the fact that, throughout the world, virtually all electric energy is distributed by voltages and currents that are sinusoidal. A second reason is that designers find that analysis of electrical systems with sinusoidal forcing functions yields good insight to the understanding of how the system will process other signals. The statement is true for high-power electrical distribution systems and for relatively low-power communications and related systems. The phrase *sinusoidal steady state* refers to the behavior of the circuit after the transient terms (terms arising from the complementary function) have become negligible.

When the circuit or network is linear, then the steady-state response to a sinusoidal forcing function is another sinusoidal signal at the same frequency. The relative amplitudes and phases may vary, but the frequency does not. This observation is crucial to development of the techniques. In essence, the process is one of finding an equivalent circuit in the frequency domain, analyzing it algebraically rather than through the use of differential equations, and interpreting the final results as needed. Since the frequency of the response signal is identical with that of the input signal, the analyst need only find the amplitudes and phases of the response signals, whether currents or voltages.

There are two approaches to the topic. One is to develop the topic from basic principles. The second is to consider sinusoidal steady state as a special case of Laplace analysis in which the complex variable s is replaced by its imaginary component, $s = j\omega$. The results are interpreted in sinusoidal terms. The first approach will be the principal approach used in this article. Further, $\omega = 2\pi f$, where f is the frequency in hertz (Hz).

The process is called *phasor analysis*. It makes use of the imaginary operator and Euler's relation:

$$\begin{aligned} &j = \sqrt{-1}, \\ &\epsilon^{jx} = \cos x + j \sin x. \end{aligned} \tag{52}$$

Euler's relation may be solved for the trigonometric terms, giving

$$\begin{aligned} \cos x &= (\epsilon^{jx} + \epsilon^{-jx})/2, \\ \sin x &= (\epsilon^{jx} - \epsilon^{-jx})/j2. \end{aligned} \tag{53}$$

A form of Euler's relation that often appears in phasor analysis is

$$\epsilon^{j(\omega t+\alpha)} = \cos(\omega t + \alpha) + j \sin(\omega t + \alpha). \tag{54}$$

8.2 Definition of a Phasor

The phasor representation or phasor transform of a sinusoidal function is, by definition, a complex number that contains both the amplitude and phase information of the original. Specifically, when

$$v(t) = V_m \cos(\omega t + \alpha) = V_m \sin(\omega t + \alpha + \pi/2), \tag{55}$$

the phasor representation of $v(t)$ is

$$V = V_m \epsilon^{j\alpha}. \tag{56}$$

It should be noted here that electric power engineers usually use a scaling factor for phasors. The alternative definition is given as

$$V_{\text{eff}} = (V_m/\sqrt{2})\, \epsilon^{j\alpha}, \tag{57}$$

and the justification will be presented in the section on power calculations. The definition may be used in reverse to convert a phasor to a time function, provided the frequency is known.

8.2.1 Phasor Algebra Phasors are complex numbers, and it is frequently necessary to perform extensive arithmetic computations with phasors. Phasor algebra follows all of the rules of complex-number arithmetic and algebra, or of two-dimensional vector algebra. Modern calculators and mathematical computer programs have greatly facilitated the arithmetic calculations that are required, as have circuit-analysis programs.

8.2.2 Differentiation and Integration Suppose a phasor is represented by

$$P_1 = P_m \epsilon^{j\omega t} \epsilon^{j\theta}. \tag{58}$$

where the frequency term is included in order to study differentiation and integration. Differentiation with respect to time becomes

$$\frac{dP_1}{dt} = j\omega P_m \epsilon^{j\omega t} \epsilon^{j\theta} = \omega P_m \epsilon^{j\omega t} \epsilon^{j(\theta + \pi/2)}, \tag{59}$$

while integration becomes

$$\int P_1 dt = \frac{1}{j\omega} P_m \epsilon^{j\omega t} \epsilon^{j\theta} = \frac{1}{\omega} P_m \epsilon^{j\omega t} \epsilon^{j(\theta - \pi/2)}. \tag{60}$$

8.3 Susceptance and Reactance

Table 3 shows the ratios of voltage phasors to current phasors, or their reciprocals. The former are called *reactances*, while the latter are called *susceptances*. When the element being considered is a capacitor, the adjective capacitive is used. When the element is an inductor, the adjective used is inductive.

Table 3. Capacitor and inductor equivalents in phasor domain.

Phasor ratio	Capacitor equivalents	Inductor equivalents
$\dfrac{V(j\omega)}{I(j\omega)}$ (Reactance)	$\dfrac{1}{j\omega C}$	$j\omega L$
$\dfrac{I(j\omega)}{V(j\omega)}$ (Susceptance)	$j\omega C$	$\dfrac{1}{j\omega L}$

8.4 Admittance and Impedance

Analysis of circuits operating in the sinusoidal steady state is typically begun by replacing each capacitor and each inductor by its susceptance or reactance. These terms may be single numbers or may be functions of frequency (in rad/s). Resistors are unaffected by this transformation. In turn, these terms may be combined following the series/parallel rules of Figs. 6 and 7, or the tee-pi conversions of Fig. 8. The result is a complex number that may or may not be a function of frequency. If the term is the ratio of a voltage phasor to a current phasor, it is called *impedance.* If the term is the ratio of a current phasor to a current phasor, it is called *admittance.*

8.5 Network Theorems and Solution Techniques

The circuits being studied are linear. The Thévenin and Norton theorems apply. Kirchhoff's laws still apply. The process of writing either loop-current or node-voltage equations still applies. These features make the method attractive and useful. The results of the analyses will again be phasor currents and voltages, of the same frequency as the excitation function. Sometimes these are reconverted to time-domain functions, but often they are left as phasor quantities and interpreted directly. This is possible because, in a linear circuit, the frequency components in the response functions are identical with those of the source. New signal frequencies can be generated only by nonlinear networks.

Figure 17 shows an analysis of a circuit that has three energy-storage elements and two nodes. The current source has a frequency of 1.25 rad/s. Specifically, Fig. 17(a) shows the basic circuit, while Fig. 17(b) shows the circuit prepared for phasor analysis at the spec-

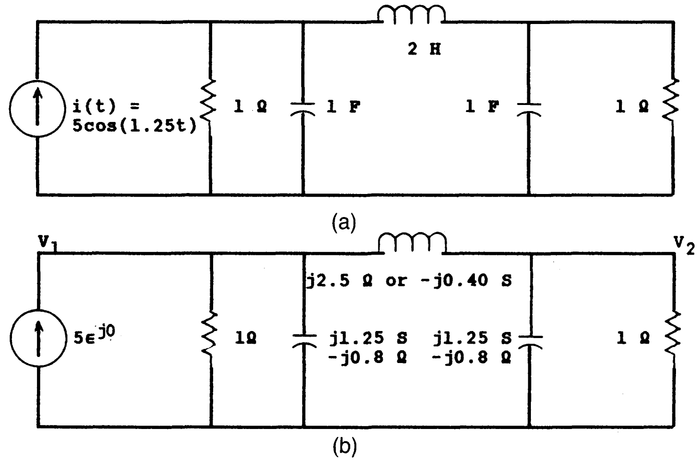

FIG. 17. (a) Two-node circuit for phasor analysis. (b) Phasor-domain equivalent circuit.

ified frequency. The following steps show the writing of Kirchhoff's current-law equations at the two nodes in matrix form, the solution for the two phasor voltages, and the reconversion to the time domain for the final answer:

$$\begin{bmatrix}5\\0\end{bmatrix} = \begin{bmatrix}(1.00 + j1.25 - j0.40) & -(-j0.40)\\ -(-j0.40) & (1.00 + j1.25 - j0.40)\end{bmatrix}\begin{bmatrix}V_1\\V_2\end{bmatrix}, \tag{61}$$

$$\begin{bmatrix}V_1\\V_2\end{bmatrix} = \begin{bmatrix}3.055 - j2.155\\ -1.103 - j0.284\end{bmatrix} = \begin{bmatrix}3.378\,\epsilon^{-j0.614}\\ 1.139\,\epsilon^{-j2.890}\end{bmatrix}, \tag{62}$$

$$\begin{bmatrix}v_1(t)\\v_2(t)\end{bmatrix} = \begin{bmatrix}3.738\cos(1.25t - 0.614)\\ 1.139\cos(1.25t - 2.890)\end{bmatrix}. \tag{63}$$

8.6 Power and Energy Considerations

When a sinusoidal voltage $v(t) = V_m \cos(\omega t + \alpha)$ is impressed across a resistor of resistance R, then there is a current $i(t) = I_m \cos(\omega t + \alpha)$ through the resistor, where $I_m = V_m/R$. The instantaneous power is given by Eq. (1), and, over an integral number of cycles, the average power is given by

$$P_{\text{avg}} = \frac{V_m I_m}{2} = \frac{V_m^2}{2R} = \frac{I_m^2}{2}R. \tag{64}$$

This amount of power is equal to the power that would be delivered by a constant voltage of magnitude $V_{\text{eff}} = V_m/\sqrt{2}$ or a constant current of magnitude $I_{\text{eff}} = \sqrt{2}I_m$. This leads to the definition of *effective values* of ac voltages and currents, which are equal to the peak values divided by $\sqrt{2}$.

The concept of effective values is in quite common use. For example, a typical light bulb available in North America might be rated at 60 W, 120 V. The 120-V specification is an effective-value voltage specification, while the average power consumed by the bulb is 60 W.

When Eqs. (56) and (57) were introduced, it was noted that the latter gives an alternative definition of a phasor. Equation (57) uses effective values in its definition, as it facilitates power and energy calculations without the necessity of retaining the scaling factor. There is not a basic difference, however, but it is necessary that the analyst be aware of which definition is being used, and to be consistent.

A more general definition for effective values is known as the *root mean square* (rms) value of the function, which applies to any signal, not simply sinusoidal signals. The term describes the defining equation, which is

$$F_{\text{rms}} = \left[\frac{1}{T}\int_{t_0}^{t_0+T} [f(t)]^2 dt\right]^{1/2}. \tag{65}$$

8.6.1 Power Factor When circuits or systems operate in the sinusoidal steady state, the current and voltage are generally not in phase; in other words, there is a phase difference between them. One effect of this fact is that the average power delivered to such systems is less than the product of effective values of current and voltage, though that term is significant. Specifically, if

$$v(t) = V_m \cos(\omega t + \alpha) \tag{66}$$

and

$$i(t) = I_m \cos(\omega t + \beta), \tag{67}$$

a computation of the average power delivered over an integral number of cycles gives

$$\begin{aligned} P_{\text{avg}} &= \frac{V_m I_m}{2} \cos(\alpha - \beta) \\ &= V_{\text{eff}} I_{\text{eff}} \cos(\alpha - \beta). \end{aligned} \tag{68}$$

The angle $\alpha - \beta$, which is the phase difference between the voltage and current, is called the *power factor angle*, and its cosine is called the *power factor*. It represents the ratio of the average power to the product of current and voltage for a device. For example, a motor rated at 120 V, 500 W, and 6.5 A would have a power factor of $500/(120 \times 6.5) = 0.641$.

8.6.2 Reactive Volt-Amperes When the voltage across and the current through a device are given by Eqs. (66) and (67), respectively, a computation of the power delivered to the device as a function of time shows

$$\begin{aligned} p(t) = \frac{V_m I_m}{2} &[\cos(\alpha - \beta) \\ &+ \cos(2\omega t + \alpha + \beta)]. \end{aligned} \tag{69}$$

There is a constant term and a double-frequency term that represents the energy that is interchanged between the electric and magnetic fields of the source and device. This new term is called *reactive volt-amperes* (vars). This quantity must be considered in the analysis and design of electrical systems, especially large power systems.

8.6.3 Power and Vars It may be shown that

$$\begin{aligned} \text{vars} &= \frac{V_m I_m}{2} \sin(\alpha - \beta) \\ &= V_{\text{eff}} I_{\text{eff}} \sin(\alpha - \beta). \end{aligned} \tag{70}$$

A combination of Eqs. (68) and (70) shows that, if the phasor voltage across and the phasor current through a device are given, respectively and in effective values, by

$$\begin{aligned} V_1 &= V_{\text{eff}} \epsilon^{(j\alpha)}, \\ I_2 &= I_{\text{eff}} \epsilon^{(j\beta)}, \end{aligned} \tag{71}$$

then the equation

$$\begin{aligned} VA &= V_{\text{eff}} I_{\text{eff}} [\cos(\alpha - \beta) + j \sin(\alpha - \beta)] \\ &= V_{\text{eff}} I^*_{\text{eff}}, \end{aligned} \tag{72}$$

where the asterisk denotes the complex conjugate, may be used to find both average power and vars. The real part of the expression is the average power, while the imaginary part is the vars.

8.7 Three-Phase Systems

Virtually all bulk electric energy distribution is accomplished through the use of three-phase systems. There are several important reasons for this. The first is that a three-phase system distributes energy with less losses than a single-phase system would. The second is that it uses materials more efficiently. The third is that it can be shown that a (balanced) three-phase system distributes electric power at a constant rate, i.e., without a double-frequency term such as appears in Eq. (69). As a consequence, three-phase generators develop and deliver constant power, and three-phase motors develop constant torque. The double-frequency torque component in a single-phase motor does not exist in a three-phase motor. Throughout this section, effective values of voltages and currents will be used.

8.7.1 Balanced Three-Phase Systems A balanced three-phase source consists of three voltage (or current) sources that are sinusoidal, equal in magnitude, and separated in phase by 120° or $2\pi/3$ rad. In time-domain and phasor format, a typical set of voltages is given by Table 4. (At a frequency of 60 Hz,

Table 4. Phasor representations of three-phase voltages.

Time format		Phasor format
$v_{ab} = V_{\text{eff}} \cos(377t)$	$= v_{ao}$	$V_{\text{eff}} \epsilon^{j0}$
$v_{bc} = V_{\text{eff}} \cos(377t - 2\pi/3)$	$= v_{bo}$	$V_{\text{eff}} \epsilon^{(-j2\pi/3)}$
$v_{ca} = V_{\text{eff}} \cos(377t + 2\pi/3)$	$= v_{co}$	$V_{\text{eff}} \epsilon^{(j2\pi/3)}$

the frequency in rad/s is $2\pi \times 60 \cong 377$ rad/s.) These three sources may be connected in either of two ways to form a balanced three-phase source. One is called a delta (or mesh or pi) connection, the second a wye (or star or tee). Both are shown in Fig. 18. Three wires or lines will be needed to connect the sources to a load. These are labeled *A*, *B*, and *C* in Fig. 18. The individual source voltages are known as *phase voltages*, and the voltages between any pair of lines are known as *line voltages*. For the wye connection, a fourth line, labeled "*O*," is possible. Sometimes it is utilized in practice, but not always. For the delta system, it can be shown that, in phasor notation only,

$$\begin{aligned} V_{AB} &= V_{ab} = V_{\text{eff}}\epsilon^{(j0)}, \\ V_{BC} &= V_{bc} = V_{\text{eff}}\epsilon^{(-j2\pi/3)}, \\ V_{CA} &= V_{ca} = V_{\text{eff}}\epsilon^{(j2\pi/3)}, \end{aligned} \qquad (73)$$

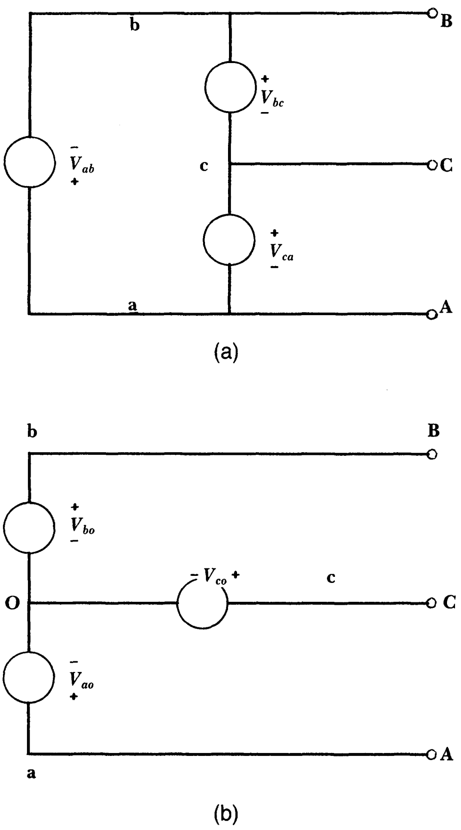

FIG. 18. (a) Delta-connected three-phase source. (b) Wye-connected three-phase source.

whereas in the wye system,

$$\begin{aligned} V_{AB} &= V_{ao} - V_{bo} = \sqrt{3}\, V_{\text{eff}}\epsilon^{(j\pi/6)}, \\ V_{BC} &= V_{bo} - V_{co} = \sqrt{3}\, V_{\text{eff}}\epsilon^{(-j\pi/2)}, \\ V_{CA} &= V_{co} - V_{ao} = \sqrt{3}\, V_{\text{eff}}\epsilon^{(j5\pi/6)}. \end{aligned} \qquad (74)$$

These results show that in a balanced wye system, the line voltages are equal to the phase voltages multiplied by $\sqrt{3}$, and are shifted in phase by 30°. In a balanced delta system, the line and phase voltages are equal. Similarly, it can be shown that, in a balanced delta system, the line currents are equal to the phase currents multiplied by $\sqrt{3}$, and they are shifted in phase by 30° with respect to the phase currents. In a balanced wye system, the line and phase currents are equal.

8.7.2 Balanced Loads A balanced three-phase load consists of three equal impedances connected either in delta or wye. Equations (15) and (16) can be used to convert from one to the other. Because of the symmetry of a balanced source and a balanced load, analysis reduces to analysis of a single phase, with appropriate shifts in phase for the other two phases.

8.7.3 Unbalanced Loads and Symmetrical Components A three-phase system that has nonsymmetrical loads, or sources that differ in magnitude, or sources that differ in phase by other than 120°, is said to be unbalanced. Circuits of this type may be analyzed by conventional nodal or loop analysis, or by any other conventional technique. There is a technique for solution of unbalanced networks that makes use of symmetrical components. In essence, this technique requires the analyst to formulate unbalanced systems as the sum of three balanced systems, and to use the techniques of balanced network analysis on the equivalent networks. The technique is described in several books (Anderson, 1973; Stevenson, 1982).

9. TWO-PORT ANALYSIS

9.1 Definition of a Port

A common use of electrical circuits is to connect a source of electrical energy or information to a load, often in such a way as to modify the signal in a prescribed way. In its basic form, such a circuit has one pair of input terminals and one pair of output terminals. Each of the four wires will have a current in or out of the circuit. Between any pair of terminals there will be a voltage.

If the structure of the circuit is such that the current into one terminal is equal to the current out of a second terminal, then that terminal pair is called a *port*. By Kirchhoff's current law, the second pair of terminals will also be a port. The circuit of Fig. 19 has two such ports, and is called a two-port. The input-output circuit variables are completely characterized, at the terminals, by the two currents and two voltages indicated in Fig. 19.

A particularly important type of two-port is called a filter. In the article ELECTRONIC CIRCUITS filters are discussed at length. In this section, a variety of relations among the terminal voltages and currents will be discussed. Six such sets are found to be useful. They are known as the open-circuit impedance parameters, short-circuit admittance parameters, hybrid parameters (two types), and transmission-line parameters (two types). For resistive circuits, all sets of parameters will be composed of constants. For circuits operating in the sinusoidal steady state, the parameters will be complex numbers for single-frequency operation, or functions of ω when the frequency is a variable. If Laplace analysis is being done, the parameters will be functions of the Laplace variable s. In the following, functional notation will be suppressed.

9.2 Open-Circuit Impedance Parameters

If the two currents are considered to be the independent variables and the two voltages are the dependent variables, then one can write

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \end{bmatrix}, \quad \text{or} \quad [V] = [z][I]. \tag{75}$$

The four numbers or functions in the square matrix characterize the network. They may be computed using any conventional method of circuit analysis, and often are directly computed by computer-based software. They may also be measured. For example, the term z_{21} can be computed or measured as the ratio V_2/I_1 when I_2 is set equal to zero (for computations), or made zero by an appropriate open circuit when measurements are being taken.

9.3 Short-Circuit Admittance Parameters

If the two voltages are in independent variables, and the two currents are the dependent variables, then one can write

$$\begin{bmatrix} I_1 \\ I_2 \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}, \quad \text{or} \quad [I] = [y][V]. \tag{76}$$

Measurements and computations follow principles similar to those of open-circuit impedance parameters.

9.4 Hybrid Parameters

Voltages and currents may be mixed in their roles as independent and dependent variables in two ways, as indicated:

$$\begin{bmatrix} V_1 \\ I_2 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} I_1 \\ V_2 \end{bmatrix} \quad (h \text{ parameters}), \tag{77}$$

$$\begin{bmatrix} I_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} \begin{bmatrix} V_1 \\ I_2 \end{bmatrix} \quad (g \text{ parameters}). \tag{78}$$

9.5 Transmission-Line Parameters

The voltage and current at the input port may be used as independent variables with the output quantities as dependent variables, or the roles may be reversed.

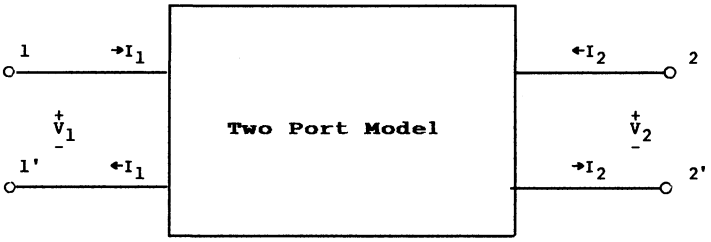

FIG. 19. Prototype to define ports of a network.

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix}$$

(chain or transmission-line parameters), (79)

$$\begin{bmatrix} V_2 \\ I_2 \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} V_1 \\ I_1 \end{bmatrix}$$

(inverse transmission-line parameters). (80)

9.6 Two-Port Parameter Conversions

Any set of two-port parameters may be converted to any other set. For example, Eq. (75) may be solved for the two currents, and the result compared with Eq. (76). This comparison shows that

$$\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = \begin{bmatrix} \dfrac{z_{22}}{|z|} & \dfrac{-z_{12}}{|z|} \\ \dfrac{-z_{21}}{|z|} & \dfrac{z_{11}}{|z|} \end{bmatrix}, \tag{81}$$

where $|z| = z_{11}z_{22} - z_{12}z_{21}$. Complete tables are given in Huelsman (1963) and in Fink and Beaty (1993), Article 2.

9.7 Equivalent Circuits for Two-Ports

Equivalent circuits may be drawn for any set of two-port parameters. The process is shown in Fig. 20 for the [g] parameters, Eq. (78), but the technique is quite similar for the other combinations. Figure 20(a) shows the equivalent circuit.

9.8 Two-Port Analysis

From the equivalent circuit for a two-port, circuit analysis is possible. Figure 20(b) shows a two-port, with hybrid-g parameters, a current source, and a load resistor. As shown here, the current source has a very high output resistance or impedance. If it were present and of such a size as to be significant in circuit operation, it would be added to g_{11}. The output voltage is $V_2 = -I_2R_L$. Analysis of the equivalent circuit using Kirchhoff's laws and some simplification shows that

$$\frac{V_{\text{load}}}{I_{\text{source}}} = \frac{V_2}{I_1} = \frac{R_L g_{21}}{(g_{11}g_{22} - g_{12}g_{21}) + g_{11}R_L}. \tag{82}$$

Similar analyses can be performed for any configuration that may arise.

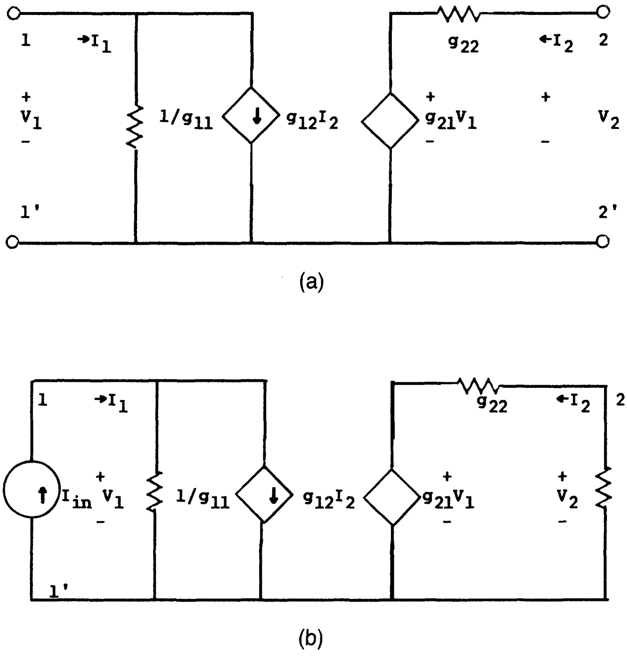

FIG. 20. Two-port analysis techniques. (a) Equivalent circuit for hybrid-g parameters, and (b) two-port with source and load.

10. DISCRETE CIRCUITS

10.1 Introduction

Discrete circuits are circuits are circuits or networks used to process discrete data. In turn, discrete data are data defined at specific instants, but undefined at other times. Discrete data may also be defined in spatial terms, but such topics are not considered in this article. As a nontechnical example, the number of people attending each of a sequence of concerts in a theater may be known, but no data are given—and it is possible that there is no interest—in the number of spectators in the theater at times when there is no concert. Another, often used, example is the set of equations that leads to the radar equations (Cadzow and Van Landingham, 1985). The data arise, in one scenario, when a radar transmitter on one airplane sends pulses that are reflected back from a second airplane. Specifically, the time for return reflections from each pulse is measured. No data are collected or available between pulses. The equations themselves will be developed in Sec. 10.7.

Discrete data are available in two ways. It is important to emphasize that the concept of discrete data applies to the independent variable—time or a spatial measurement. If the dependent variables are expressed in or converted to digital formats, the process is called *digitization* or *quantization*. Continuous-time data may be converted to discrete-time data formats through some sort of sampling process. An important principle that applies here is Shannon's sampling theorem, discussed in Sec. 10.10. This is a necessary step when the data are to be processed by a digital computer, often the mode of implementation of a discrete circuit. The second is from transducers that give discrete data directly.

Discrete circuits are implemented with electronic circuits or with digital computers. In this discussion, however, the interest is in the terminal behavior of the circuits, not in the detailed operation of the circuits. This section considers linear, time-invariant circuits only.

10.2 Discrete Signal Representation

In this discussion, it will be assumed that the discrete data are taken at integral multiples of some time unit that is usually called the *sampling interval*. The integer k will denote the time, kT, where T is the sampling interval. This regular spacing of the data is not necessary, but it is convenient.

10.2.1 Analytic Representation Discrete signals may be represented in analytic form. Several examples are shown in Table 5.

10.2.2 Sequence Representation Discrete signals may be represented as a sequence of numbers, as

$$f(k) = \{5, 4, \underset{\uparrow}{-3}, -2, 8, 9, \ldots\} \tag{83}$$

The vertical arrow below one element of the sequence denotes the value at $k = 0$. This is an example of a sequence with nonzero values for negative k.

10.2.3 Numerical or Graphical Representation Discrete data may be represented in numerical form, and stored numerically. The sample space may be multidimensional, for example time and space, and then several indices are needed. The data may also be represented or shown graphically, with the index k being the independent variable. Figure 21 shows a possible representation. It is essential that the form of the graph not suggest any sort of line or connection between ordinates as k varies.

10.3 Discrete-Signal Operations

There are three basic operations in linear, time-invariant discrete-circuit analysis and design. The first of these, illustrated in Fig.

Table 5. Analytic representation of discrete functions.

Name	Analytic representation
Exponential	$f(kT) = f(k) = A\epsilon^{-kT}$
Step	$u(kT) = u(k) = 1$, all $k \geq 0$
Ramp	$f(kT) = f(k) = kT$, $k \geq 0$
Impulse	$f(kT) = f(k) = \begin{Bmatrix} 1, k = 0 \\ 0, k \neq 0 \end{Bmatrix}$
Complex periodic exponential (N defines the period)	$f(kT) = f(k) = A\epsilon^{j(2\pi/N)\mathrm{kT}}$
Sinusoidal	$f(kT) = f(k) = A\cos\left(\frac{2\pi kT}{N} + \alpha\right)$

Note: k takes on integral values only.

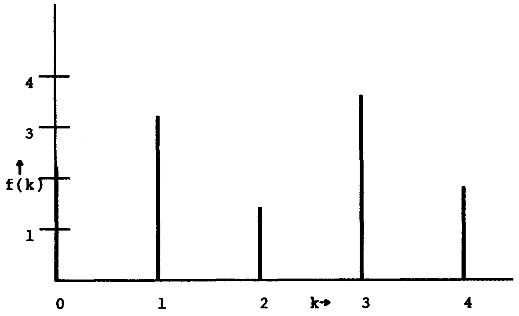

FIG. 21. Graphical representation of discrete data.

22(a), is addition. Addition includes subtraction if signs are appropriate. The second, illustrated in Fig. 22(b), is multiplication by a constant. Sometimes, this is called gain or amplification. (If the gain were a function of the index k, then the circuit would be time-varying. If two signals were multiplied together, the circuit would be nonlinear.) The third operation, illustrated in Fig. 22(c), is delay. A general term that includes both gain and delay is transmittance. The output of a delay block is identical with the input, but it is delayed in time, normally by one sampling interval, though this is not essential. Delay, like energy-storage elements in ordinary circuits, provides memory. The notation z^{-1} is a commonly used symbol for delay, because of its association with the z transform, to be introduced in Sec. 10.9.

10.4 Signal-Flow and Block Diagrams

Discrete circuits are commonly shown by block diagrams or by signal-flow diagrams. The two ideas are quite similar, and are used in similar ways. The block diagram uses geometric figures to describe the three basic operations, and directional lines to show how signals move through the system. A typical block diagram is shown in Fig. 23(a). Figure 23(b) shows the same system described by a signal-flow graph, which uses directional lines to show signal propagation through a system, and coefficients on or near the lines.

10.4.1 Signal-Flow Graph Definitions It is convenient to define a number of terms associated with signal-flow graphs.

Node: A point in a signal-flow graph at which two or more elements connect.

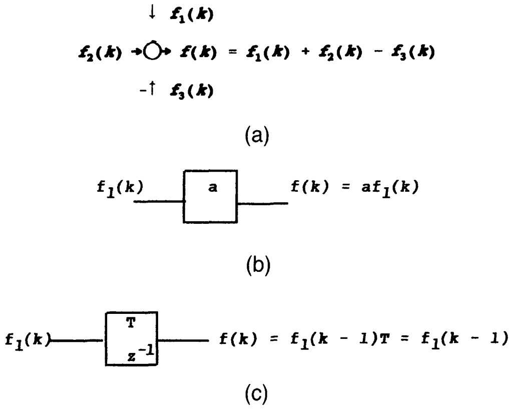

FIG. 22. Discrete-circuit operations: (a) summing, (b) multiplication or gain, and (c) delay.

Branch: An element that shows signal propagation from one node to another. A branch is directional.

Input node: A node that has only outgoing branches.

Output node: A node that has only incoming branches.

Path: A set of branches traversed in the same direction. In this definition, an element may be included more than once.

Forward path: A path that starts at an input node, ends on an output node, and does not include any branch or node more than one time.

Loop: A path that starts and terminates on the same node, and does not traverse any node or branch more than once.

Forward path gain: The product of branch gains or transmittances is called the path gain. If the path is a forward path, then the path gain is called forward path gain.

Loop gain: The product of branch gains or transmittances around a loop.

10.4.2 Signal-Flow Graph Reductions At times it is useful to simplify or to expand a signal-flow graph either to facilitate analysis or to enhance system understanding. Figure 24 illustrates three equivalents that are commonly used. They may be extended to a wide variety of situations.

10.4.3 Mason's Rule or Formula A powerful rule was developed by Mason (1953, 1956) for use in the analysis of signal-flow graphs. Although presented here in connection with discrete circuits, it is equally ap-

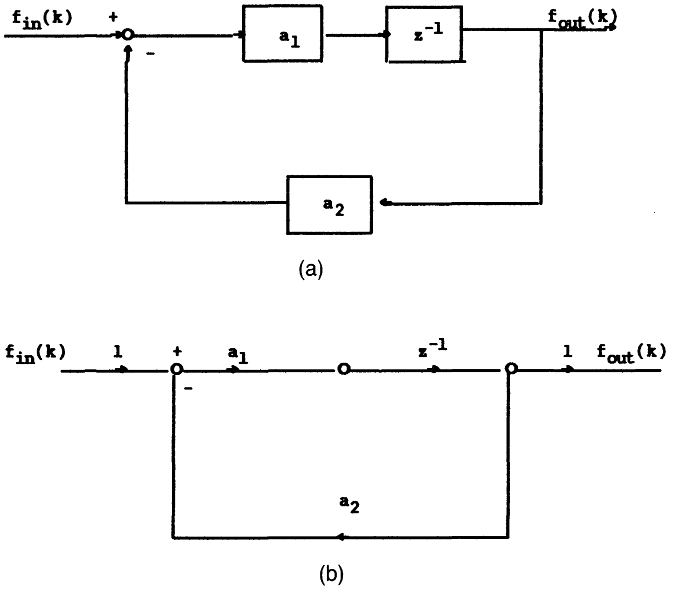

FIG. 23. Equivalent graphical representations of a single discrete circuit: (a) block diagram, and (b) signal-flow diagram.

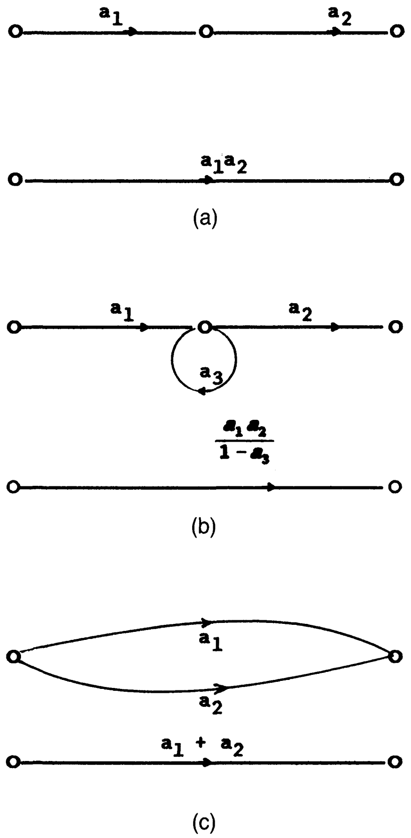

FIG. 24. Signal-flow graph reduction techniques: (a) paths in cascade, (b) removal of self loop, and (c) parallel paths.

plicable to continuous-time circuits. The rule is general, easily applied, and readily applicable to digital computer analysis. The rule states that the overall or general transmittance from an input node to an output node is given by

$$M = \frac{y_{\text{out}}}{x_{\text{in}}} = \frac{1}{\Delta} \sum_k M_k \Delta_k, \tag{84}$$

where $\Delta = 1 -$ (sum of all individual loop transmissions) + (sum of the products of the transmittances of loops that do not touch, taken two at a time) − (sum of the products of the transmittances of loops that do not touch, taken three at a time) + $\cdots$, M_k = path transmittance of the kth forward path, and Δ_k = value of that part of Δ that arises from terms that do not touch the kth forward path.

As an example of the use of Mason's rule is shown in Fig. 25. The diagram shown has seven nodes and ten transmittance paths. The input node is labeled x, the output y. There are three loops, one of which (this type of loop is called a "self-loop") does not "touch" the other two. There are two transmittance paths, one of which touches all three loops, the other of which does not touch two of the loops. Thus,

$$\begin{aligned}\Delta = 1 - a_8 - a_4a_9 - a_4a_5a_{10} \\ + a_8a_4a_9 + a_8a_4a_5a_{10},\end{aligned} \tag{85}$$

$$M_1\Delta_1 = (a_1a_3a_4a_5a_6a_7)(1), \tag{86}$$

$$M_2\Delta_2 = (a_1a_2)(1 - a_4a_9 - a_4a_5a_{10}), \tag{87}$$

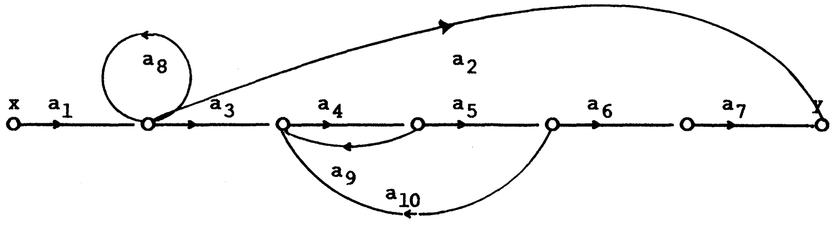

FIG. 25. Signal-flow graph for analysis.

so that

$$M = \frac{y_{\text{out}}}{x_{\text{in}}} = \frac{a_1a_3a_4a_5a_6a_7 + a_1a_2(1 - a_4a_9 - a_4a_5a_{10})}{1 - a_8 - a_4a_9 - a_4a_5a_{10} + a_8a_4a_9 + a_8a_4a_5a_{10}}. \tag{88}$$

10.5 Discrete Equations

A general type of expression that describes a finite discrete circuit is

$$\begin{aligned} y(k) = {} & \alpha_n x(k) + \alpha_{n-1}x(k-1) + \alpha_{n-2}x(k-2) \\ & + \cdots + \alpha_0 x(k-n) - \beta_{n-1}y(k-1) \\ & - \beta_{n-2}y(k-2) - \cdots - \beta_0 y(k-n). \end{aligned} \tag{89}$$

This is a generic nth-order *difference equation* with constant coefficients. It relates the output or dependent variable $y(k)$ to the input or independent variable $x(k)$. In general, $x(k)$ is known in some way—analytically, numerically, or stochastically. To solve the equation means then to find a $y(k)$ that satisfies Eq. (89) for all values of k.

Several ways for solving equations like Eq. (89) exist. Occasionally, solutions can be developed by algebraic methods. It is possible to build a theory of difference equations that is analogous in nearly every respect to the theories of differential equations. Concepts like complementary function and particular solution apply. Similarly, a powerful transform method has been developed. It is known as the z transform, and is discussed in Sec. 10.7. Numerical techniques are also commonly employed.

10.6 A Compound-Interest Example

The calculations for compound interest are a good illustration of a discrete equation and its solution. Suppose a person deposits an amount of money PR regularly at the beginning of an interest period, and that all money on deposit earns interest at a rate R over this period. If the annual interest rate is r, compounded N times annually, then $R = r/N$. Call the total amount on deposit y. The following set of equations illustrates the basic compounding process.

$$\begin{aligned} y(1) &= PR, \\ y(2) &= (1+R)y(1) + PR = (1+R)PR + PR, \\ y(3) &= (1+R)y(2) + PR \\ &= (1+R)^2PR + (1+R)PR + PR, \\ &\;\;\vdots \\ y(n) &= [(1+R)^{n-1} + (1+R)^{n-2} \\ &\quad + \cdots + (1+R) + 1]PR, \end{aligned} \tag{90}$$

where $y(n)$ is the amount on deposit at the end of the nth compounding period immediately after a deposit has been made. There is an algebraic identity that states

$$\sum_{k=0}^{n} \alpha^k = \frac{1 - \alpha^{n+1}}{1 - \alpha}.$$

If this identity is used with the preceding equation, with $\alpha = 1 + R$, then, after some manipulation, an expression for the amount on deposit is

$$\begin{aligned} y(n) &= \left[\frac{(1+R)^n - 1}{R}\right]PR \\ &= \left[\frac{(1+r/N)^n - 1}{r/N}\right]PR. \end{aligned} \tag{91}$$

Suppose a person deposits \$100 into a savings account at the beginning of a month, and makes 24 such deposits. Further, consider a 9% interest rate, with monthly compounding, so that $N = 12$. Equation (91) shows that, immediately after making the 24th deposit, the saver will have \$2618.85 in the account.

10.7 The α-β Radar Equations

A frequently quoted example of discrete equations is a set of equations obtained for the purpose of analyzing data from radar pulses sent from one airplane to another (Cadzow and Van Landingham, 1985). One purpose of the system might be to avoid collisions. In a basic representation, the transmitter on one airplane sends short pulses of electromagnetic energy. These are reflected by the second airplane, and received by the first. Figure 26(a) typifies the sent and received pulses. The received pulses are contaminated with electrical noise, and thus are not rectangular. The time of pulse travel to and from the "target" ΔT must be measured in the presence of noise, and these estimates are the only data obtained.

Define the following terms:

$u(k) = (c/2)\Delta T$, the measurement of distance to the target, where c is the velocity of light.

$y_v(k)$ = an estimate of the velocity of the target.

$y_p(k-1)$ = a prediction of the distance to the target, based on earlier data.

$y(k)$ = present value of the range of the target.

The basic ideas are to

1. use the preceding value of position, the latest value of velocity, and T to predict where the target will be at the kth pulse;
2. use the predicted range to the target and the new measurement of range to determine a new value for the range; the two pieces of data are weighted by factors $1 - \alpha$ and α respectively;
3. use the preceding velocity and an estimate of the velocity change based on the new target range and the predicted range, weighted by a factor β, to compute a new value of velocity.

The equations are

$$\begin{aligned} y_p(k) &= y(k-1) + Ty_v(k-1), \\ y(k) &= y_p(k) + \alpha[u(k) - y_p(k)], \\ y_v(k) &= y_v(k-1) + (\beta/T)[u(k) - y_p(k)]. \end{aligned} \tag{92}$$

The determination of the factors α and β has been studied a great deal as radar systems have developed and improved. Figure 26(b) shows a flow graph of these equations.

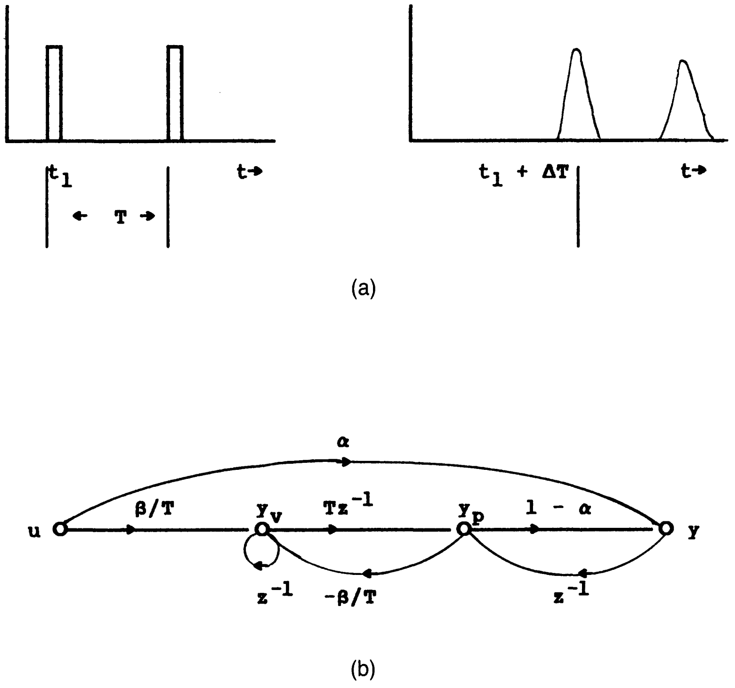

FIG. 26. (a) The transmitted and received pulses in radar. (b) The signal-flow diagram for deducing the velocity of the target.

10.8 Numerical Solutions of Difference Equations

Numerical solutions are carried out in a sequential fashion, and are readily adapted to digital computer solutions. The technique is illustrated by an example. Let a sample of Eq. (89) be given by

$$y(k) = 2x(k) - x(k-1) - 1.5x(k-2) - 0.5y(k-1), \qquad (93)$$

where

$$x(k) = \begin{cases} 2, k \geq 0 \text{ even}, \\ 1, k \geq 0 \text{ odd}, \end{cases}$$

and the initial condition is $y(k-1) = 0$. Further, $x(k-1) = x(k-2) = 0$. The solution becomes

$$y(0) = 2x(0) = 4$$

$$y(1) = 2x(1) - x(0) - 0.5y(0) = 2 - 2 - 2 = -2$$

$$y(2) = 2x(2) - x(1) - 1.5x(0) - 0.5y(1) = 4 - 1 - 3 + 1 = 1$$

$$y(3) = 2x(3) - x(2) - 1.5x(1) - 0.5y(2) = 2 - 2 - 1.5 - 0.5 = -2$$

$$y(4) = 2x(4) - x(3) - 1.5x(2) - 0.5y(3) = 4 - 1 - 3 + 1 = 1.$$

10.9 The *z* Transform

A powerful transform method has been developed for analysis and design of discrete data systems. It has much in common with the Laplace transform, and it allows the use of algebraic methods in many aspects of the circuit analysis and design. In this section, the topic will be defined and introduced, theorems will be stated, tables will be given, and sample problems will be worked.

10.9.1 Definition of the *z* Transform Suppose a discrete function sequence $\{e(k)\}$, $k \geq 0$, is being considered. The z transform of $\{e(k)\}$, denoted by $E(z)$, is defined as a power series in z^{-k} whose coefficients are equal to the values of the sequence. Thus,

$$E(z) = Z[\{e(k)\}] = \sum_{k=0}^{\infty} e(k)z^{-k}. \qquad (94)$$

In general, z is a complex number.

10.9.2 Examples of *z* Transforms Let the sequence be given as $e(k) = \{1, -\frac{1}{2}, \frac{1}{4}, -\frac{1}{8}, \frac{1}{16}, \ldots\}$. From Eq. (94),

$$E(z) = 1 - \tfrac{1}{2}z^{-1} + \tfrac{1}{4}z^{-2} - \tfrac{1}{8}z^{-3} + \cdots \qquad (95)$$

If the identity

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots, \ |x| \leq 1 \qquad (96)$$

is used, then

$$E(z) = \frac{1}{1 - (-\frac{1}{2})z^{-1}} = \frac{z}{z + \frac{1}{2}}. \qquad (97)$$

As a second example, suppose that the sequence is

$$e(k) = \epsilon^{-akT}. \qquad (98)$$

Then,

$$E(z) = 1 + \epsilon^{-aT}z^{-1} + \epsilon^{-2aT}z^{-2} + \cdots \qquad (99)$$

Application of Eq. (96) to this sequence gives

$$E(z) = \frac{z}{z - \epsilon^{-aT}}. \qquad (100)$$

10.9.3 *z*-Transform Theorems Table 6 presents, without proof, a variety of theorems found to be useful with z transforms. Proofs of these theorems follow from the definitions. In this table, it is assumed that $E(z)$ is the z transform of the sequence $e(k)$.

10.9.4 *z*-Transform Tables Table 7 presents, without proof, a limited set of z transforms for a commonly used set of sequence functions. Additional tables may be found in some of the references.

Table 6. z-transform theorems.

Operation	Theorem
Addition and subtraction	$Z[\{e_1(k)\} \pm \{e_2(k)\}] = E_1(z) \pm E_2(z)$
Multiplication by constant	$Z[a\{e(k)\}] = aE(z)$
Real translation, part 1	$Z[\{e(k - n)\}] = z^{-n}E(z)$
Real translation, part 2	$Z[\{e(k + n)\}] = z^n[E(z) - \sum_{k=0}^{n-1} e(k)z^{-k}]$
Complex translation	$Z[\{\epsilon^{ak}e(k)\}] = E(z\epsilon^{-a})$
Initial value	$e(0) = \lim_{z\to 0}E(z)$ (provided limit exists)
Final value	$\lim_{k\to\infty}e(k) = \lim_{z\to 1}(z - 1)E(z)$ (provided limit exists)

10.10 Formulation and Solution of *z*-Transform Equations

Example 1. Equation (93), which has been solved numerically, will be used to illustrate a z-transform solution. It is repeated here:

$$y(k) = 2x(k) - x(k-1) - 1.5x(k-2) - 0.5y(k-1) \qquad (101)$$

where

$$x(k) = \begin{cases} 2, & k \text{ even}, \\ 1, & k \text{ odd}. \end{cases}$$

Denote the z transform of $y(k)$ by $Y(z)$, and the z transform of $x(k)$ by $X(z)$. If the theorems from Table 6 are used, and a transform of the equation is taken, with initial conditions (values for $k < 0$) that are zero, then the transformed equation becomes

$$Y(z) = 2X(z) - z^{-1}X(z) - 1.5z^{-2}X(z) - 0.5z^{-1}Y(z). \qquad (102)$$

The input sequence $x(k)$ can be written as the sum of two sequences, $\{\frac{3}{2}, \frac{3}{2}, \frac{3}{2}, \ldots\}$ and $\{\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, \ldots\}$, or in another form, as $\frac{3}{2}\{1, 1, 1, \ldots\}$ and $\frac{1}{2}\{1, -1, 1, -1, 1, \ldots\}$. The latter two sequences have, from Table 7, z transforms of

Table 7. A short table of z transforms

Name	Function	z transform
Unit impulse	$\delta(k) = 1, k = 0$ $=0, k \neq 0$	1
Unit step	$u(k) = 1, k \geq 0$	$\dfrac{z}{z-1}$
Alternating step	$u(k) = (-1)^k, k \geq 0$	$\dfrac{z}{z+1}$
Unit ramp	$k, k \geq 0$	$\dfrac{Tz}{(z-1)^2}$
nth-order ramp	$k^n, k \geq 0$	$\lim_{a\to 0}(-1)^n \dfrac{\partial^n}{\partial a^n}\left[\dfrac{z}{z - \epsilon^{-aT}}\right]$
Exponential	ϵ^{-akT}	$\dfrac{z}{z - \epsilon^{-aT}}$
Damped ramp	$t\epsilon^{-akT}$	$\dfrac{Tz\epsilon^{-aT}}{(z - \epsilon^{-aT})^2}$
Cosine	$\cos(akT)$	$\dfrac{z(z - \cos(aT))}{z^2 - 2z\cos(aT) + 1}$
Sine	$\sin(akT)$	$\dfrac{z\sin(aT)}{z^2 - 2z\cos(aT) + 1}$
Damped cosine	$\epsilon^{-akT}\cos(bkT)$	$\dfrac{z(z - \epsilon^{-aT}\cos(bT))}{z^2 - 2\epsilon^{-aT}\cos(bT)z + \epsilon^{-2aT}}$
Damped sine	$\epsilon^{-akT}\sin(bkT)$	$\dfrac{z\epsilon^{-aT}\sin(bT)}{z^2 - 2\epsilon^{-aT}\cos(bT) + \epsilon^{-2aT}}$

$$\tfrac{3}{2}\frac{z}{z-1} \quad \text{and} \quad \tfrac{1}{2}\frac{z}{z+1}, \tag{103}$$

respectively. When the sum of these two terms is substituted into Eq. (102) and the resulting equation is solved for $Y(z)$, the result is

$$\begin{aligned} Y(z) &= \tfrac{3}{2}\frac{2z^2 - z - 1.5}{(z-1)(z+0.5)} \\ &\quad + \tfrac{1}{2}\frac{2z^2 - z - 1.5}{(z+1)(z+0.5)}. \end{aligned} \tag{104}$$

Because the degree of the numerator is equal to the degree of the denominator in both parts of Eq. (104), a direct partial-fraction representation is not possible. Examination of Table 7 shows that invariably a factor z appears in the numerator in the transform. Thus, a technique that virtually always works is to expand $Y(z)/z$ in partial fractions, and then to multiply through by z at the end. When this is done, the result is

$$\begin{aligned} Y(z) &= 4.5 - \tfrac{1}{2}\frac{z}{z-1} - \frac{z}{z+0.5} \\ &\quad - 1.5 + \frac{z}{z+0.5} + \frac{1.5\,z}{z+1}. \end{aligned} \tag{105}$$

As it happens, some cancellations are possible. When these are done and the information in Table 7 is used to take the inverse transform, the result is

$$y(k) = 3\delta(k) - \tfrac{1}{2}u(k) + 1.5(-1)^k. \tag{106}$$

Evaluation of Eq. (106) gives the same numerical results that were given earlier in the numerical solution.

Example 2. This example will show some computational details, especially some that arise in the computation of the inverse z transform. When roots are complex, then the residues in the inverse z transform expansion are complex numbers, which may be unwieldy to deal with. Euler's relation [Eqs. (52) and (53)] may be used to arrange the results in real-number terms only. Suppose that the transfer function for a discrete system, as a specific example of Eq. (88) and Fig. 25, is given by

$$\frac{Y(z)}{X(z)} = \frac{z^3 + 2z^2 + 5z - 5}{z^3 - 1.70z^2 + 1.59z - 0.623} \tag{107}$$

and that the forcing function $X(z)$ is a step function, so that

$$X(z) = \frac{z}{z-1}. \tag{108}$$

When $Y(z)$ is computed by first factoring the denominator, finding $Y(z)/z$, and then multiplying through by z, the result is

$$\begin{aligned} Y(z) &= \frac{11.24\,z}{z-1} + \frac{0.868\,z}{z-0.70} \\ &\quad + \frac{(-5.552 + j\,0.995)z}{z - 0.50 - j\,0.80} \\ &\quad + \frac{(-5.552 - j\,0.995)z}{z - 0.50 + j\,0.80}. \end{aligned} \tag{109}$$

While the first two terms need no special attention, the latter are difficult to interpret without modification. Consider the auxiliary equation

$$Y(z) = \frac{a + jb}{z - \alpha - j\beta} + \frac{a - jb}{z - \alpha + j\beta} \tag{110}$$

whose inverse transform can be written as

$$y(k) = (a + jb)\,\gamma^k \epsilon^{jk\theta} + (a - jb)\,\gamma^k \epsilon^{-jk\theta}, \tag{111}$$

where

$$\begin{aligned} \gamma &= \sqrt{\alpha^2 + \beta^2}, \\ \theta &= \arctan(\beta/\alpha). \end{aligned} \tag{112}$$

When the two real terms from Eq. (111) are collected and the two imaginary terms are collected, and in addition, the numerator and denominator of the real terms are multiplied by 2, while for the imaginary terms, $j2$ is used, then

$$y(k) = 2a\gamma^k \frac{\epsilon^{jk\theta} + \epsilon^{-jk\theta}}{2} - 2b\gamma^k \frac{\epsilon^{jk\theta} - \epsilon^{-jk\theta}}{j2}. \tag{113}$$

With this auxiliary result, the inverse transform from Eq. (109) can be written

$$\begin{aligned} y(k) &= 11.24u(k) + 0.87(0.70)^k \\ &\quad - 11.10\,(0.94)^k \cos(1.01k) \\ &\quad + 1.99\,(0.94)^k \sin(1.01k). \end{aligned} \tag{114}$$

Figure 27 shows a graph of this function for

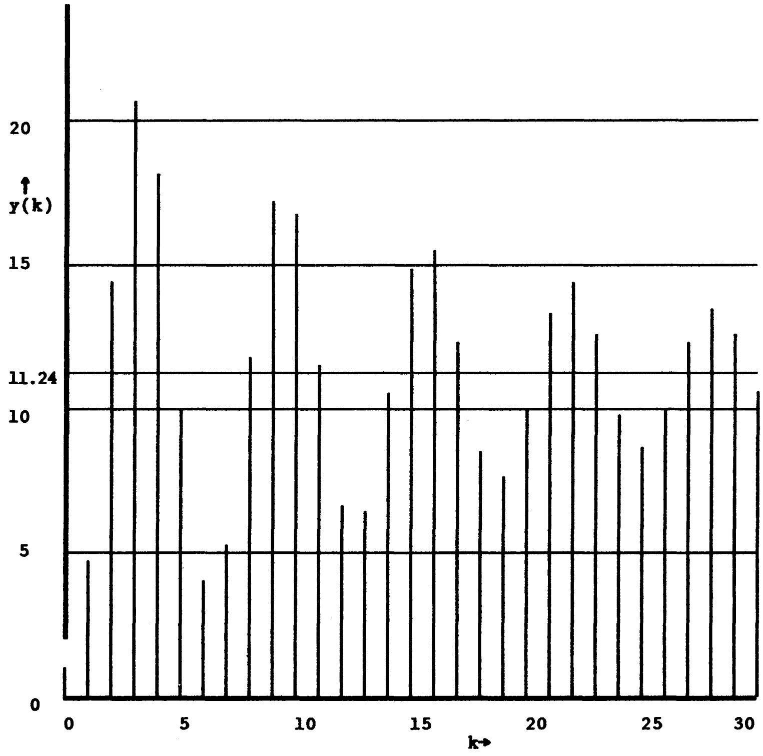

FIG. 27. Graph of Eq. (114).

values of k up to 30. It is seen that this function approaches its final value of 11.24 slowly, primarily because of the large coefficients and the fact that the magnitude of the complex roots is close to unity.

10.11 Shannon's Sampling Theorem

An important question that arises in the design of systems that require sampling of a continuous-time system signal in order to obtain a discrete representation of that signal is the question of how often the signal must be sampled in time in order to obtain a satisfactory representation of that signal. Shannon answered this question in a classic paper (Shannon, 1949) that shows the minimum sampling frequency must be at least twice the highest frequency component of the signal. For example, if a signal has frequency components from very low frequency to, say, 20 kHz, then the minimum sampling frequency must be 40 kHz, or the maximum sampling interval can be no longer than 25 μs. This rate is commonly called the Nyquist rate, because of earlier work done on the problem by Nyquist (1928).

11. SUGGESTIONS FOR FURTHER STUDY

11.1 Nonlinear and Time-Varying Networks

It was indicated at the beginning that this article would focus on linear, time-invariant networks. The study of nonlinear networks is interesting and challenging. Time-varying networks present a different set of challenges, and the combination is still rich with areas of research. Some of the items listed as suggestions for further reading include excellent discussions of these concerns.

11.2 Computer-Based Circuit Analysis

The advent of the digital computer has given circuit analysts and designers powerful tools to work with. These tools are refined regularly and rapidly, so that the specific details change rapidly. Software is available for desktop computers and for centralized computing facilities. The reader interested in this subject is advised to check books that have been published quite recently, and to also study software manuals carefully. Some suggestions are included.

11.3 State-Variable Methods for Continuous-Time Circuits

The techniques presented here for state-variable analysis work well in most cases, but, as indicated, loops with nothing but capacitors and voltage sources, or nodes with nothing but current sources and inductors, require special treatment.

11.4 Topology

Circuits or networks have been studied extensively using topological methods. These methods yield great insight into analysis and design methods, ensure that the equations written are independent, and make possible the sophisticated circuit analysis and design programs that are available. Some titles are suggested.

11.5 Stability

Active networks, in their basic form, are networks or circuits that contain controlled sources. Because of the energy such sources can provide, it is possible that such networks can become unstable. In general, this is not a desirable behavior. Methods exist for predicting such instability, and there are design methods for eliminating it.

11.6 State-Variable Methods for Discrete-Time Circuits

It is possible to develop a state-variable technique for discrete circuits. The study is particularly useful in the design and analysis of digital control systems and in signal-processing techniques. The technique has a lot in common with what was presented for continuous-time circuits, but significant differences in details must be considered.

11.7 The Modified *z* Transform

The *z*-transform techniques presented do not allow for study between sampling intervals. The modified *z* transform (Phillips and Nagle, 1990) allows such study when needed.

GLOSSARY

Admittance: The ratio of the transform (phasor or Laplace) of a current to the transform of a voltage.

Block Diagram: A graphical representation of the interconnection of the elements of a circuit or system. See **Signal-Flow Graph**.

Branch: A circuit element connecting two nodes in a circuit.

Capacitor: An electric circuit element having the property that the current through the device is proportional to the time derivative of the voltage across the device.

Circuit: An interconnected collection of electrical devices such as resistors, capacitors, electronic devices, and electrical energy sources. Also called **networks**.

Continuous-Time Circuit: A circuit or network that is described by a set of continuous-time integro-differential equations.

Current: The flow of electrically charged particles.

Dependent or Controlled Source: A source of electrical energy or signal whose current or voltage is controlled by or proportional to another current or voltage in the circuit.

Discrete-Time Circuit: A circuit or network that is described by a set of discrete-time difference equations.

Effective value: Also known as the root mean square (rms) value of the function:

$$F_{\text{rms}} = \left[\frac{1}{T}\int_{t_0}^{t_0+T} (f(t))^2\,dt\right]^{1/2}.$$

Gyrator: The fundamental nonreciprocal electrical element. The defining equations are

$$v_1 = R\,i_2,$$
$$v_2 = -R\,i_1,$$

where R is known as the "gyration resistance."

Impedance: The ratio of the transform (phasor or Laplace) of a voltage to the transform of a current.

Independent Source: A source of electrical energy or signal that is controlled by an agent outside the circuit.

Inductor: An electric circuit element having the property that the voltage across the device is proportional to the time derivative of the current through the device.

j: The imaginary operator, defined by $j^2 = -1$. Also denoted by i.

Kirchhoff's Current Law (KCL): The statement that the sum of currents leaving (or entering) a node is zero.

Kirchhoff's Voltage Law (KVL): The statement that the sum of voltages around a loop is zero.

Laplace Transform: If a function of time $f(t)$ is known, and defined for $t \geq 0$, then the (single-sided) Laplace transform is given by

$$\mathscr{L}f(t) = F(s) = \int_0^\infty f(t)\, \epsilon^{-st}\, dt,$$

where s is a complex variable, $s = \sigma + j\omega$, chosen so that the integral will converge. In turn, σ is the real part of the variable, and ω is the imaginary part.

Linear Circuit: A circuit or network described by a set of linear integro-differential equations.

Loop: A path through a circuit that begins and ends at the same node.

Mason's Rule: A technique for formulating input-output relations from flow graphs or block diagrams.

Mesh: A path in a planar circuit, i.e., one that can be drawn on a planar surface without wires crossing.

Node: In a circuit or network, the junction of three or more wires from circuit elements. Sometimes called "essential node."

Network: A term that is synonymous with circuit.

Norton Theorem: A theorem stating that, at a terminal pair, any linear network may be replaced by the parallel combination of a current source and a resistance or impedance. The current is equal to the current that flows through the short circuit when the external load is short circuited. The resistance is equal to the resistance measured or calculated at the terminal pair with all independent sources de-energized.

Path: A traverse through two or more branches in a circuit.

Phasor: A representation of a sinusoidal function that is a complex number that contains both the amplitude and phase information of the original. When

$$v(t) = V_m \cos(\omega t + \alpha) = V_m \sin(\omega t + \alpha + \pi/2),$$

the phasor representation of $v(t)$ is

$$V = V_m \epsilon^{j\alpha}.$$

The definition may be used in reverse to convert a phasor to a time function, provided the frequency is known. Sometimes called a phasor transform.

Port: A circuit or network with one or more terminal pairs having the property that the current into one wire is equal to the current out of a second wire. These two wires form a terminal pair that is called a port.

Resistor: An electric circuit element having the property that the voltage across the device is proportional to the current through the device.

Shannon's Sampling Theorem: A theorem that states that, in order to preserve the original content, the sampling frequency used must be at least twice the highest frequency component of the signal being sampled. Sometimes called the Nyquist rate.

Signal-Flow Graph: A graphical representation of the interconnection of the elements of a network or system. See block diagram.

Superposition: A principle stating that, for a given system (network) having a general excitation $x(t)$ that yields a response $y(t)$, if the response to an excitation $x_1(t)$ is $y_1(t)$, and the response to an excitation $x_2(t)$ is $y_2(t)$, then any excitation

$$x(t) = a_1x_1(t) + a_2x_2(t)$$

yields the response

$$y(t) = a_1y_1(t) + a_2y_2(t)$$

for all values of the constants a_i.

Tellegen's Theorem: This theorem states: Let N be a network of b branches and n nodes. If KVL is satisfied by the branch voltages $v_1, v_2, \ldots, v_b$ in every loop, and if KCL is satisfied by the branch currents $i_1, i_2, \ldots, i_b$ at every node, then

$$\sum_{k=1}^{b} v_k i_k = 0.$$

Thévenin Theorem: A theorem stating that, at a terminal pair, any linear network may be replaced by a series combination of a voltage source and a resistance or impedance. The voltage is equal to the voltage at the terminal pair when the external load is removed or open-circuited, and the resistance is equal to the resistance calculated or measured at the terminal pair with all of the independent sources de-energized.

Two-Port Parameters: A set of equations that describes a circuit having two ports by its terminal behavior.

Voltage: The energy, per unit charge, required to move a charge from one point in space to another.

***z* Transform:** A power series in z^{-k} whose coefficients are equal to the values of a sequence. Thus,

$$E(z) = Z[\{e(k)\}] = \sum_{k=0}^{\infty} e(k)\, z^{-k}.$$

In general, z is a complex number.

Works Cited

Anderson, P. M. (1973), *Analysis of Faulted Power Systems*, Ames: Iowa State University Press.

Balabanian, N., Bickart, T. A., Seshu, S. (1969), *Electrical Network Theory*, New York: John Wiley and Sons.

Cadzow, J. A., Van Landingham, H. F. (1985), *Signals, Systems, and Transforms*, Englewood Cliffs, N.J.: Prentice-Hall Book Company, Inc.

Desoer, C. A., Kuh, E. S. (1969), *Basic Circuit Theory*, New York: McGraw-Hill Book Company.

Fink, D., Beaty, W. (1993), *Standard Handbook for Electrical Engineers*, 13th ed., New York: McGraw-Hill Book Company.

Huelsman, L. P. (1963), *Circuits, Matrices, and Linear Vector Spaces*, New York: McGraw-Hill Book Company.

Mason, S. J. (1953), "Feedback Theory—Some Properties of Signal Flow Graphs," *Proc. Inst. Radio Eng.* 41:9, pp. 1144–1156.

Mason, S. J. (1956), "Feedback Theory—Further Properties of Signal Flow Graphs," *Proc. Inst. Radio Eng.* **44** (7), 920–926.

Nyquist, H. (1928), "Certain Topics in Telegraph Transmission Theory," *Trans. Am. Inst. Electrical Eng.* **47** (4), 617–644.

Phillips, C. L., Nagle, H. T. (1990), *Digital Control Systems: Analysis and Design*, 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall Book Company.

Rohrer, R. A. (1970), *Circuit Theory: An Introduction to the State Variable Approach*, New York: McGraw-Hill Book Company.

Shannon, C. E. (1949), "Communication in the Presence of Noise," *Proc. Inst. Radio Eng.* **37** (1), 10–21.

Stevenson, W. D., Jr. (1982), *Elements of Power System Analysis*, 4th ed., New York: McGraw-Hill Book Company.

Tellegen, B. D. H. (1948), "The Gyrator, a New Electric Network Element," *Philips Res. Rep.* **3**, 81–101.

Tellegen, B. D. H. (1952), "A General Network Theorem, with Applications," *Philips Res. Rep.* **7**, 259–269. [A related article is Tellegen, B. D. H. (1953), "A General Network Theorem, with Applications," *Proc. Inst. Radio Eng., Australia* **14**, 265–270.]

Further Reading

Chan, Shu-Park (1969), *Introductory Topological Analysis of Electrical Networks*, New York: Holt, Rinehart and Winston.

Chan, Shu-Park, Chan, Shu-Yun, Chan, Shu-Gar (1972), *Analysis of Linear Networks and Systems*, Reading, MA: Addison-Wesley Publishing Company.

Chua, L. O. (1969), *Introduction to Nonlinear Network Theory*, New York: McGraw-Hill Book Company.

Chua, L. O., Desoer, C. A., Kuh, E. (1987), *Linear and Nonlinear Circuits*, New York: McGraw-Hill Book Company.

Conant, R. (1993), *Engineering Circuit Analysis with PSPICE and PROBE*, New York: McGraw-Hill Book Company.

Franklin, G. F., Powell, J. D., Workman, M. L. (1990), *Digital Control of Dynamic Systems*, 2nd ed., Reading, MA: Addison-Wesley Publishing Co.

Gardner, M. F., Barnes, J. L. (1942), *Transients in Linear Systems*, New York: John Wiley and Sons, Inc.

Hayt, W. H., Kemmerly, J. E. (1993), *Engineering Circuit Analysis*, 5th ed., New York: McGraw-Hill Book Company.

Keown, J. L. (1991), *PSPICE and Circuit Analysis*, New York: Merrill Publishing Company.

Mayhan, R. J. (1984), *Discrete-Time and Continuous-Time Linear Systems*, Reading, MA: Addison-Wesley Publishing Co.

McGillem, C. D., Cooper, G. R. (1991), *Continuous and Discrete Signal and System Analysis*, 3rd ed., Philadelphia: Saunders College Publishing (a division of Holt, Rinehart, and Winston).

Monssen, F. (1993), *PSPICE with Circuit Analysis*, New York: Merrill, Macmillan Publishing Company.

Nilsson, J. W. (1993), *Electric Circuits*, 4th ed., Reading, MA: Addison-Wesley Publishing Co.

Rashid, M. H. (1990), *SPICE for Circuits and Electronics Using PSPICE*, Englewood Cliffs: Prentice-Hall Book Company.

Seshu, S., Reed, M. B. (1961), *Linear Graphs and Electrical Networks*, Reading, MA: Addison-Wesley Publishing Company.

Spence, R. (1970), *Linear Active Networks*, New York: Wiley-Interscience, Inc.

Strum, R. D., Kirk, D. E. (1988), *First Principles of Discrete Systems and Digital Signal Processing*, Reading, MA: Addison-Wesley Publishing Company.

Strum, R. D., Kirk, D. E. (1994), *Contemporary Linear System—Using MATLAB*, Boston: PWS Publishing Company.

Thorpe, T. W. (1992), *Computerized Circuit Analysis with SPICE*, New York: John Wiley and Sons.

Tuinenga, P. W. (1988), *A Guide to Circuit Simulation and Analysis Using PSPICE*, Englewood Cliffs: Prentice-Hall Book Company.

Van Valkenburg, M. E. (Ed.) (1974). *Circuit Theory: Foundations and Classical Contributions*, Stroudsburg, PA: Dowden, Hutchinson, and Ross Publishing.

Van Valkenburg, M. E. (1982), *Linear Circuits*, Englewood Cliffs: Prentice Hall Book Company.

NETWORKS, COMPUTER

EHUD GAVRON, *Aces Research, Inc., Tucson, Arizona, U.S.A.*

Introduction 263

1. **A Brief History of Computer Networks** 263

1.1 ARPANET 264

1.2 Internet 264

1.3 Ethernet 264

2. **Types of Computer Networks** 264

2.1 Local-Area Networks 264

2.2 Metropolitan-Area Networks 265

2.3 Wide-Area Networks 265

2.4 Internetworks 265

3. **Geometries of Computer Networks** 265

4. **Switching Techniques for Computer Networks** 266

4.1 Circuit Switching 266

4.2 Message Switching 267

4.3 Packet Switching 267

5. **ATM and SONET** 268

6. **Computer Network Media** ... 268

7. **Digital Signal-Encoding Techniques** 269

8. **Network Protocol Models** ... 270

9. **Network Security** 272

9.1 Physical Integrity 272

9.2 Data Confidentiality 272

9.3 Denial of Service 273

9.4 Recovery Plans 273

10. **Network Alternatives** 273

11. **Conclusions** 273

Glossary 273

Further Reading 274

INTRODUCTION

The explosion in both the popularity and the use of computers has created a situation where it is no longer simply enough to process data on one computer. The need frequently arises to share data and resources between many computers. This is the primary purpose of utilizing a computer network.

As network use becomes more widespread, however, its use is also becoming more difficult. The advantages of using computer networks are such that the added difficulty is viewed merely as a hurdle to overcome. With companies attempting to minimize costs it becomes apparent that sharing expensive peripherals such as printers and disk drives between many users is a goal that is highly desirable. Making a large number of devices available to a large number of users increases the utilization of resources while only marginally increasing costs. Additionally, a computer network can eliminate dangerous dependences on a single resource and allow for more robust infrastructure. No longer is the failure of a single computer likely to halt work in a department, because there are several other machines or servers that can perform the same task. Networks have allowed the computing world to step away from the peer-to-peer computing model, where all computers are equal, into client-server relationships, where servers offer their resources to systems that desire to use those services (clients). Client-server computing and its tremendous popularity are easily linked to the advances in available network technologies. Networks have advanced from their infancy in the 1970s into large and often complex constructs. Understanding these networks, their uses, and their limitations will become an increasingly critical skill in business and society.

1. A BRIEF HISTORY OF COMPUTER NETWORKS

The first real computer network was designed and built in the 1950s at the Massachusetts Institute of Technology's (M.I.T) Lincoln Laboratory. The Semi-Automatic

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Ground Environment (SAGE) system consisted of 23 computer networks, each of which connected roughly 100 radar sites to a single computer. Data were sent through voice-grade phone lines at a rate of approximately 1300 bits per second (bps).

In 1964 the first large commercial computer network was built by IBM for American Airlines. The SABRE system heavily built upon work done for the Department of Defense (DoD) on the SAGE system, and indeed involved many of the same individuals who had worked on SAGE.

1.1 ARPANET

In 1971 the Advanced Research Projects Agency (ARPA) (now known as DARPA) heavily promoted the creation of the ARPA Network (ARPANET) to connect military installations and universities with widely varying environments. The technique used to connect these institutions was called "packet switching" (see Sec. 4.3). Work on ARPANET was often groundbreaking. It was the test-bed for analytical and performance models for networks. Additionally, many techniques such as flow control, fault-tolerant performance, and layered protocols received much of their initial testing on the ARPANET. In 1983 ARPANET was divided into two separate networks: the Defense Data Network (DDN) and a new ARPANET. Today ARPANET is almost completely phased out as hosts have been moved to direct Internet connectivity.

1.2 Internet

The Internet is a worldwide internetwork of hosts communicating via the TCP/IP protocol suite. Internetworks are multiple networks connected such that data can pass between the various networks. The Internet is the only current global-scale internetwork. TCP/IP is a developing suite of communication protocols that was formed as a direct result of the DARPA research. Its strengths are in fast and simple scalability, although as local-area networks have expanded into the global Internet, a potential shortage of unique worldwide addresses is possible in the near future. This same global Internet is widely being touted as the Information Superhighway of the future. The Internet of today owes much of its structure and existence to the ARPANET of old, both in structure and in design.

1.3 Ethernet

Ethernet appeared from Digital Equipment Corporation, Intel, and Xerox (DIX) in 1974 and remains today the dominant Local-Area Network (LAN) technology. Its original goal was to devise a way to send information between many office machines. Its basic protocol was a refinement of earlier work with wireless data transmission from Hawaii to the mainland. Originally specifying a one megabit per second network (1Mb/s), the DIX consortium later succeeded in revising the standard by an order of magnitude, making Ethernet a 10-Mb/s LAN medium. The development of other local-area network protocols (token-ring, Apple's localtalk, Banyan's vines) occurred at about the same time, but these other protocols have never gained the popularity and use that Ethernet has.

2. TYPES OF COMPUTER NETWORKS

Computer networks come in many shapes and sizes. The different characteristics of networks results in several different categories of networks. The first characteristic often used to separate networks is the size and complexity of the network. This segmentation, in general, leads to three different classifications of networks: Local-Area Networks (LANs), Metropolitan-Area Networks (MANs), and Wide-Area Networks (WANs). Additionally, there are internetworks that bridge the gap between individual networks.

2.1 Local-Area Networks

LANs are by far the most common type of telecommunication network. As its name implies, a LAN is typically a network that spans a small geographic region. Devices, or nodes, in a LAN are typically never further apart than a few kilometers. LANs are generally owned by a single organization and contain no more than a few hundred nodes. The cabling between the nodes is typically installed and owned by the organization that manages the LAN. The data transmission speeds are typically in the 10-Mb/s range. However, as fiber-optic links become more prevalent, these speeds are constantly increasing.

2.2 Metropolitan-Area Networks

MANs are the next largest category of networks. A MAN typically covers a larger geographic region than a LAN. However, there is no set distance or size at which a LAN and a MAN are distinct; thus the line between them is often blurred. One characteristic that does often distinguish the two is that the cable for a MAN is often owned by local public data carriers (i.e., local phone companies) and leased by the organization. Because of the higher costs of leasing cable, data transmission rates are typically between 56 kb/s and 1.544 Mb/s (referred to by telephone companies as T-1 circuits). These are also available in Europe under the slightly different configurations of 64 kb/s and 2.048 Mb/s (known as E-1 circuits).

2.3 Wide-Area Networks

WANs are the largest of the discrete networks. WANs often cross state and even country boundaries to create a single network for large organizations. The costs associated with such networks are high as, in the United States, the cable must invariably be leased from interexchange carriers (IXCs), also known as "long-distance carriers." These carriers (e.g., AT&T) provide these high-bandwidth lines at a relatively high cost. The speeds of such connections generally range from 56 kb/s to 45 Mb/s (T-3 circuits). Several intracontinental fiber-optic connections are currently available, promising significantly more bandwidth for future use. These larger data pipes (150 Mb/s to 1.5 Gb/s) have necessitated the design of protocols capable of rapidly processing data at such rates. These are discussed in more detail under ATM and SONET.

2.4 Internetworks

The concept of connecting various networks together provides a framework for integrating the advantages of multiple networks with simple scalability. Such integrated networks are known as internetworks. Internetworks are designed for circumstances when applications or needs transcend a single network and must be able to span several networks. These networks connect several small networks to create larger networks. Examples of this might be the connection of several MANs to create a WAN. Internetworks are usually only used by large organizations that can afford not only the cabling costs associated with the connections, but also the cost in personnel needed to administer the smaller networks and the larger networks simultaneously.

3. GEOMETRIES OF COMPUTER NETWORKS

Computer networks may also be characterized by their geometry in addition to their size. There are many different possible topologies. Different geometries may often be used for different sized networks, and for networks with different functions. In general there are four different types of computer network geometries: bus, ring, star, and mesh. These geometries are illustrated in Fig. 1.

Based on the same technology found inside the modern computer, connecting all "cards" to a chassis or "motherboard," a bus geometry is characterized by a single com-

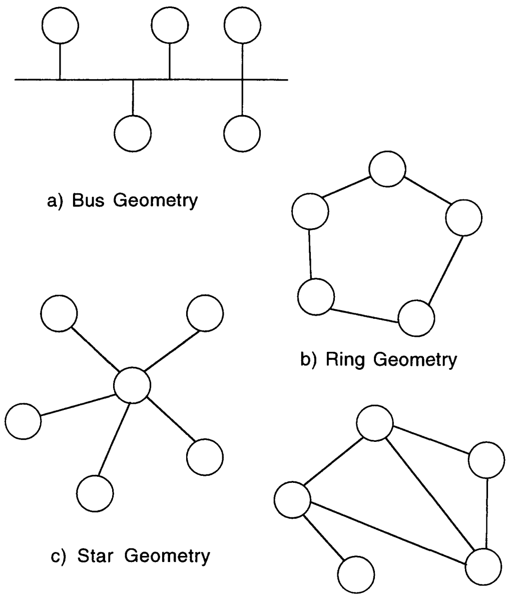
a) Bus Geometry

b) Ring Geometry

c) Star Geometry

d) Mesh Geometry

FIG. 1. Computer network geometries.

mon cable, or bus, in which nodes are connected via short cables. This is the most common configuration for LANs as it is an architecture of lesser complexity which performs its functions better under low network loads.

The ring (or loop) topology consists of all the nodes being connected in series along a closed path. Normally this path involves data transmission in a single direction, either clockwise or counterclockwise. However, there are technologies, primarily Fiber Distributed Data Interface (FDDI), that utilize a double ring where data flows on one ring clockwise and on the other ring counterclockwise. The advantage of using dual rings is that not only does it increase the amount of data that can be sent and received in a given time, but it is also much more fault tolerant. A FDDI ring by design includes some fault tolerance, in that segmentation of the ring will cause FDDI frames to reverse direction at the break point. Thus a single break is not sufficient to eliminate more than the general throughput of the ring. Multiple failures, however, will isolate the ring into broken segments. A dual-ring configuration provides redundancy via a completely separate fiber-optic path for data to follow. A failure on one ring is compensated by the existence of a second complete ring. The advantage to using rings at all is that they are normally implemented in such a manner as to allow them to do very well at high network loads.

A star topology, based on the model of a single computer with many outlying connections to terminals, has a single central node that is connected via point-to-point links to the other nodes. This single node acts as a central switching point for all of the other nodes. The disadvantage of this configuration is that the central node is a single point of failure; should it fail, the entire network will be useless. The star's advantage is that it does an excellent job of keeping very busy nodes from bogging down slower nodes with their network traffic.

Finally, the mesh (or general) geometry is a general configuration that allows for multiple routes for data. This is the topology that is most commonly seen for WANs. In particular, the ARPANET was developed using this configuration and the current NSFNET is still of this configuration type. Its main advantage is fault tolerance; the failure of a single node is normally acceptable because there are multiple routes to each node. However, in general, it is more expensive in terms of both cabling costs and network equipment costs because of this redundancy.

4. SWITCHING TECHNIQUES FOR COMPUTER NETWORKS

In computer networks, there are three primary switching techniques. These techniques allow sharing of communication facilities by multiple users. These techniques are circuit switching, message switching, and packet switching. A form of packet switching utilizing fixed-length packets of information is known as cell-switching, and is the basis for such protocols as ISDN (Integrated Services Digital Network) and ATM (Asynchronous Transfer Mode). The emphasis is therefore on covering the details of packet switching as it is currently the dominant technique. Figure 2 shows the different times and techniques involved with each switching technology.

4.1 Circuit Switching

Circuit switching is what is used by the average individual most because it is used for ordinary telephone calls. The basic construct is the simplest of the three techniques. Use is first initiated by a connections phase, during which a circuit is established between a source and destination. The second phase is the actual message phase, where data are transmitted back and forth. Finally, there is the disconnection phase, where the circuit is reset. An example of circuit switching can be shown in a telephone call. The connection phase occurs when the phone number is dialed and while waiting for the recipient to answer. The message phase is the actual conversation, and the disconnection phase is when both of the parties hang up. The disadvantage of circuit switching technology is that in computer time the connection phase is relatively long. This automatically burdens the end system with a high overhead before a message is even sent. However, as the connection phase times become shorter, circuit switching may become more prevalent.

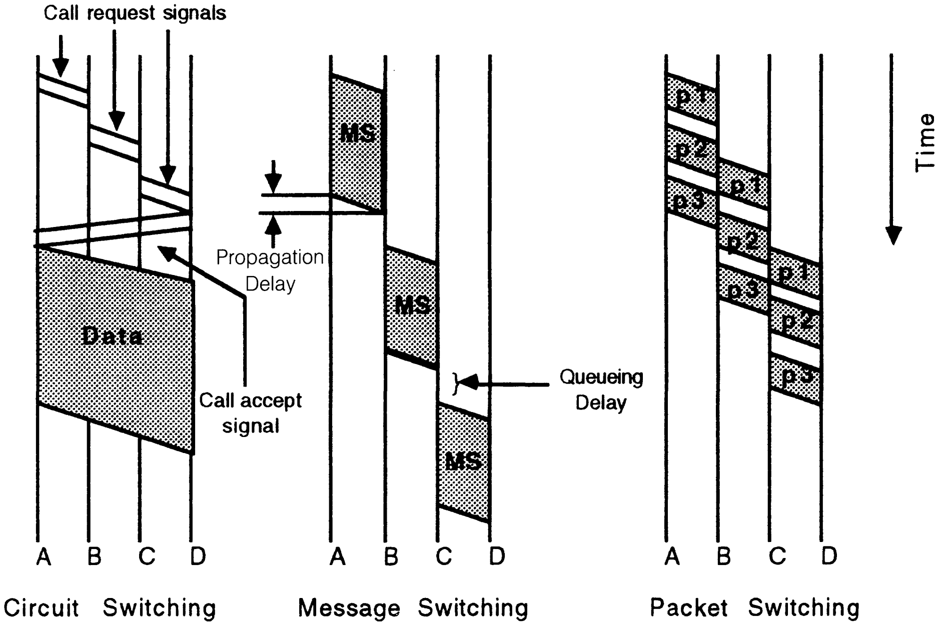

FIG. 2. Switching techniques.

4.2 Message Switching

Message switching shares a single connection between multiple users. In this scheme, there is no connection or termination phase because a message is simply sent from link to link until it reaches its destination. It is sometimes also referred to as store-and-forward switching, since messages are stored at intermediate nodes on their way to their destinations. The basic concept can be easily illustrated. If a message needs to get from point A to point C, and it passes through point B, it can be sent to point B first where it will then be forwarded to point C. The message would first go from A to B. It would then sit at B until a connection to C became available. In the interim someone else can also send a message from A to B. Finally, when the B to C link becomes available the message would be sent, and hopefully shortly thereafter any other messages would also continue on to point C. The advantage to this system is that it allows multiple users to use a single channel with little added delay over circuit switching. However, the time gained in not having a connection phase can often be lost in queuing delay as messages sit at intermediate nodes.

4.3 Packet Switching

Finally, there is packet switching. Packet switching is equivalent to message switching for short messages. Any message that is larger than some specified length is broken into smaller pieces, known as packets. These packets, each with a certain amount of extra data known as a header, are then transmitted through the network individually. The main advantage of packet switching is that it allows for an effect known as pipelining. Pipelining occurs when the gaps that are normally associated with message switching on a network get filled by parts of the packets, thus leaving little idle time on any of the segments. This effect is visible in Fig. 2.

There are two basic approaches to packet switching: virtual circuit and datagram. Virtual circuits are again analogous to setting up a telephone call. One host determines a route to another host and then all subsequent packets follow that same route. The advantage to virtual-circuit packet switching is that it allows for smaller headers on the packets, as they simply must know a circuit number. The disadvantage is that if some node in the circuit should go down, all packets will be lost. Datagram routing is more closely analogous to sending a letter. There is no time required to set up a route, and different packets may take different routes to get to their location. As with mailing a letter, datagram routing is next-hop routing. In next-hop routing, each host only needs to know the next host responsible for routing the datagram to its eventual destination. Just as the letter is given to the carrier who gives it to the post office, etc., next-hop routing makes possible large internetworks with minimal routing path information storage. The advantage of datagram packet switching is that it is more robust; if a node on the way to a destination should go down, the packets will simply use a different route if it is available. The disadvantages are that each packet must have more information about the destination, and the order of packet delivery could vary. This requires more robust network programs on the receiving end node.

A variant of packet switching is called "cell-switching" or "cell-relay," and it uses fixed-length data packets known as "cells." The chief advantage to the use of the fixed-length cells as compared to variable-length packets is that equipment can be optimized to buffer, copy, and route data by quickly scanning multiple headers in multiple cells.

5. ATM AND SONET

In the early 1990s, the introduction of globally available fiber-optic technology had prompted many of the IXCs to create nationwide (and in some cases international) fiber-optic networks. These networks typically used FDDI at its initial rate of 100 Mb/s. As technology advanced, SONET (Synchronous Optical NETwork) was introduced to provide the much-needed higher rates of speed. SONET rates are measured in scales from OC-1 to OC-12, with the number being a multiple of 50 Mb/s. Thus OC-3, which is typically discussed as appropriate for replacing FDDI, is roughly 150 Mb/s.

With the availability of faster networking, the technology to do high-speed cell routing became greatly needed. The telephone company ISDN offering was heavily modified and changed to become ATM. ATM is expected to be the backbone networking standard into the next century as it provides access to bandwidth-intensive video workstations, sound transmissions, and data networking.

6. COMPUTER NETWORK MEDIA

In the 1980s, one of the hot topics of discussion was the choice of media for use in LANs, MANs, and WANs. In the early 1980s, the only real choice in network media was coaxial cable. However, in recent years, the shift has been away from coaxial cable for the LAN applications and toward twisted-pair copper wire, because of its lower cost. On the MAN and WAN side the trend has definitely accelerated toward optical fiber.

Coaxial cable is so named because of its construction. It consists of a center conducting cable, surrounded by an insulator, which is in turn surrounded by another conductor, which is finally covered with an outer insulator. This construction gives coaxial cable very good immunity to electromagnetic interference (EMI) that often affects other cables. It is capable of being used for cable runs anywhere from a few meters to a few kilometers. However, its cost effectively prevents it from achieving widespread use in both MAN and WAN applications.

Twisted-pair cable has rapidly become the cable of choice for LAN installation. Twisted pair is constructed with two copper wires that are insulated, and then twisted around each other. This twisting helps to eliminate the effect of outside electromagnetic interference on the cable. As the number of twists per inch increases, so does the effectiveness of the noise reductions. Unfortunately, this increase in the number of twists increases the cost of the wire. The technology has evolved to the point where twisted pair can be used for cable runs up to 1 km and data rates of up to 100 Mb/s, with two current schemes competing for the opportunity to provide this standard. CDDI (Copper Distributed Data Interface) and 100-Mb/s Ethernet comprise these two camps.

Optical fiber is the medium of choice for MANs and WANs. Optical fiber is essentially a long transparent core of glass or plastic compounds, surrounded by another layer of the material with different optical characteristics than the core. The advantages of optical fiber are numerous. First, it is immune to electromagnetic interference. Second, there is little signal loss to the medium itself. As a result, cable runs can be longer (up to roughly 90 km) and tend to have fewer errors. Finally, the data rates for optical fiber can exceed 1 Gb/s! The main disadvantage of optical fiber used to be the cost. However, by the late 1980s fiber cost began to drop rapidly as better production methods and high volumes lowered manufacturing costs. A relevant disadvantage of optical fiber is that the devices that convert the optical signals to electrical signals and vice versa are still relatively expensive.

Connecting remote locations has become easier with the use of radio networking. Broadband radio networking utilizes large-bandwidth allocations of frequency space to transmit the full bandwidth of the network over radio waves. Spread spectrum radio networking utilizes multiple smaller-bandwidth channels for similar effect. A large contributor to the growth of the latter has been the Federal Communication Commission (FCC) deregulation of low-power spread-spectrum communication.

7. DIGITAL SIGNAL-ENCODING TECHNIQUES

Digital signal encoding is used to modify the binary data that computers and networks communicate with in order to improve communication performance. There are three main criteria used to determine which encoding techniques are to be used, with a goal of maximum signal with minimal noise. These are spectrum shaping, clock extraction, and error-detection capability. Spectrum shaping is designed to minimize the impact of low-frequency noise on a signal. The ability for the receiver to determine the clocking frequency of the sender with no additional data is defined as clock extraction. Finally, error detection is the characteristic that allows the receiver to determine whether or not the data on the line are actually valid. The four encoding schemes covered are polar nonreturn to zero (NRZ), unipolar return to zero (RZ), bipolar RZ, and Manchester. These four techniques are illustrated in Fig. 3.

Polar NRZ is the most common encoding technique between computers and remote terminals. A full-width positive pulse indicates a one, and a full-width negative pulse indicates a zero. The weakness of polar NRZ is that it does poorly on computer network media that do not use low frequencies (i.e., coaxial cable and twisted pair). Additionally, error detection for polar NRZ is poor and clock synchronization is not inherent to the technique.

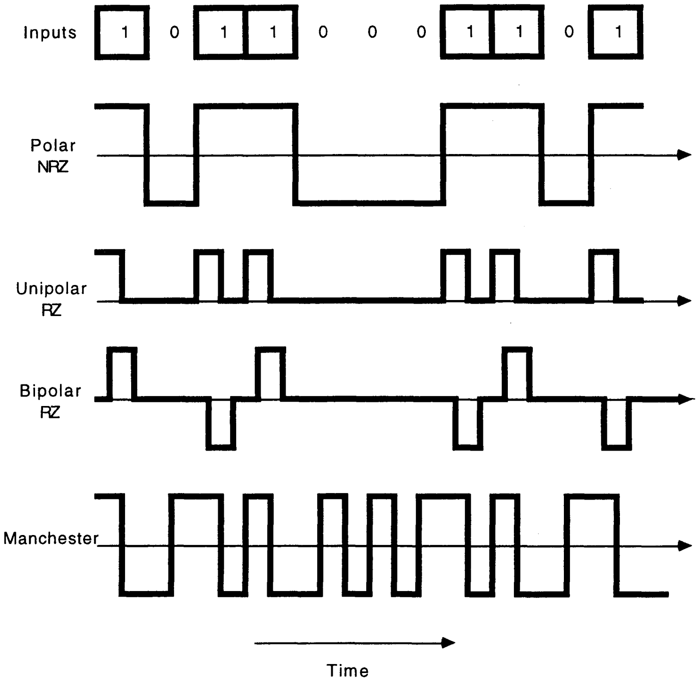

FIG. 3. Digital signal-encoding techniques.

Unipolar RZ encoding uses a half-width positive pulse to represent a one and no pulse to represent a zero. Its clock extraction is poor, although it does have moderately good error detection. However, it shares the disadvantage of polar NRZ and is not particularly useful for networks that are even as small as LANs.

Bipolar RZ uses alternating positive and negative half-width pulses to indicate ones, and no pulse to indicate zeros. A noteworthy feature of this technique is that it helps eliminate the effects of dc voltages on the lines. Clock extraction is not inherent to the encoding, but error detection is good, although not as good as for Manchester encoding. This technique is often used for WANs that utilize cable as opposed to satellites.

Manchester (or biphase) encoding uses a positive and negative half-width pulse to represent a one, and a negative and positive half-width pulse to represent a zero. This is in fact the dominant encoding technique for LANs. Clock extraction for Manchester is simple and error detection is excellent.

8. NETWORK PROTOCOL MODELS

Commonly agreed upon conventions for communications are typically referred to as protocols. A protocol is a set of rules stating how information is going to be exchanged. In most situations in life, fairly informal protocols may be used. However, for computer networks, a very formal protocol set must be used in order to ensure that all computers involved can communicate effectively. In computer networks, the protocol models that are typically used are referred to as layered protocols. These protocols consist of several layers, each with its own function and definition, that combine to make up the entire protocol model. By far the most common network protocol model is the International Standards Organization (ISO) Open Systems Interconnection (OSI) architecture. This is typically referred to as the ISO/OSI model, or more informally as the Seven-Layer Model. However, a second networking protocol developed for the United States Department of Defense (DoD) is the most widely implemented model. This model is a direct result of the ARPANET project from the 1970s and 1980s. Figure 4 shows both of these models.

The ISO/OSI model is broken into seven layers: physical, data link, network, transport, session, presentation, and application. Each of these layers has a separate function, and can communicate with the layers both above and below it bidirectionally. A brief description of the seven layers follows.

1. The physical layer provides media-level (e.g., electrical, light-oriented, radio waves, etc.) functional and procedural character-

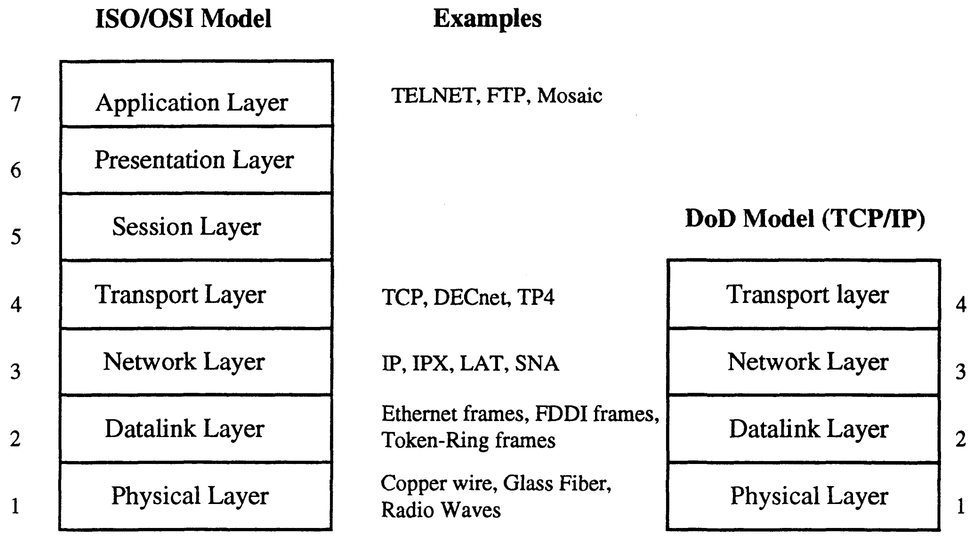
ISO/OSI Model
Examples
7
Application Layer
TELNET, FTP, Mosaic
6
Presentation Layer
5
Session Layer
DoD Model (TCP/IP)
4
Transport Layer
TCP, DECnet, TP4
Transport layer
4
3
Network Layer
IP, IPX, LAT, SNA
Network Layer
3
2
Datalink Layer
Ethernet frames, FDDI frames, Token-Ring frames
Datalink Layer
2
1
Physical Layer
Copper wire, Glass Fiber, Radio Waves
Physical Layer
1

FIG. 4. Network protocol models.

istics involved in transparently passing the bits of computer data across the network.

2. The data-link layer provides functional means of transferring data between network components and may do some error correction. It also regulates flow control, and media access control.
3. The network layer provides switching and routing functions to different network hosts and masks data transfer details from higher layers.
4. The transport layer provides data transfer to upper layers and ensures reliability and cost-effective data transfer. This layer does so without presenting lower-layer information to the higher layers. It is therefore said to provide a seemingly transparent connection between two communication endpoints.
5. The session layer provides a mechanism for structuring dialogs between different application processes.
6. The presentation layer serves to make differences in data representation between applications transparent to the application.
7. The application layer is concerned only with the semantics of the applications. Applications all reside in this layer.

The DoD model is much more flexible and allows applications to skip layers that would otherwise be required in the ISO/OSI model. Where the ISO model emphasizes the creation of a structured model, the DoD model demonstrates true functional internetworking. This may be partly because the ISO model was constructed by theoreticians with network ideals, while the DoD model was designed by practitioners documenting an existing model. This DoD advantage makes it especially suited for use on internetworked WANs and MANs. With the emphasis on internetworking, the DoD model is a connectionless model designed to handle short interactive data transfers. The DoD model consists of four layers, which are defined as follows:

Layer I (network access layer) contains protocols that control the flow of data on the network from host to host.

Layer II (Internet layer) provides the functions necessary to allow data to traverse between networks from host to host. Additionally, this layer may handle error correction, message segmentation, and assembly and message identification.

Layer III (host-host layer) contains functions needed to process data between processes on different network hosts. It may include setting up logical connections, error control, and flow control.

Layer IV (process/application layer) contains protocols specific to a given process or application. This may include connecting to remote hosts, transferring files, and exchanging messages.

The DoD model is used heavily by advocates of client-server computing because of its inherent efficiencies. In the early days of computing, all processing was done on one machine in a batch environment. The same machine read the input, processed the data, and generated output. Over time, multiple machines were used in such a fashion, with one doing input, one doing processing, and yet another doing output.

In the last few years, the norm has changed to utilize machines of differing power, where the large, powerful machines make their processing power available to smaller, less powerful machines. In such an event, the large machine is said to be "providing a service" to the small machine by making its (the large) CPU horsepower available.

This usage has expanded now; if the hardware offers cycles then it is a "compute server." If the software offers other software programs the ability to process data for them, it is an "application server." Programs and hardware that use these services are called "clients."

For example, if there are 30+ IBM PC clones, and 100+ Macintoshes, each of these have local disks, but it would be unreasonable to buy everybody large disks. Instead, one can have one machine dedicated to providing everyone disk space; it would be purchased with large disk storage facilities. These it would provide to anyone who wished to use them over the network. Thus it is a "file server," and any PC or Macintosh that uses it is called a "client."

In general, there are two philosophies of structuring information processing. Peer-processing, where all computers are created equal and all are equally capable of performing a job, is the first such philosophy or model. This is often found in large computer centers filled with minicomputers or mainframes. The other is client-server, where a minority of large systems provide services (disk, CPU, memory, data processing, results processing, application data, etc.) to smaller systems. This latter philosophy works well with cost-conscious computing establishments. As the global Internet has grown to include clients and servers of all kinds, the concept of a "global backplane" has emerged, where the physical location of the clients and servers is not at all important, but rather their connection to the global network, and being able to access available resources, are. In the "global backplane" topology, all computers are connected together, and a suitably authorized client may access any available server.

9. NETWORK SECURITY

As computer networks become more prevalent and higher volumes of data flow on them, the security of both the network and the data becomes a greater concern. In the infancy of computer networks, dependency on computers was minimal, and as a result a loss of service was not crippling. However, this has changed and today's organizations have yet another mission-critical component to keep a watchful eye over.

Threats to computer network security are numerous. The primary concerns are as follows: the physical integrity of the network, the confidentiality of the data, denial of services, and recovery plans. These issues are briefly discussed, and methods of minimizing the threat are provided. It should be noted that this discussion is in no way complete or all-encompassing, as this is the subject of numerous books.

9.1 Physical Integrity

The physical integrity of the network is probably the issue that should be first addressed when looking at network security issues. Even after eliminating all other threats, no matter how sophisticated the attack, if the equipment is left without adequate protection, the network can be shut down with nothing more complicated than a pair of wire cutters or the stroke of a key. The key to physical security is to locate network-critical items in restricted areas. If an organization has a single server, then placing it in the hallway where anyone could turn it off would not be as wise a decision as placing it in a locked room. The same holds true for the cables between computers. The normal practice is to run the cables underground where they are difficult to get to, or through the actual structure of a building. Unfortunately, with the increasing use of the global internetworks by anyone with the financial resources, control of all access to the cables or even interconnected machines is difficult to achieve. In this case, the best practice is to isolate the network in a manner that still allows the realization of most of the benefits without the liabilities. This is normally done by placing a "firewall" between the organizational network and the networks under external control. A firewall is a computer that is designed to stop external access and prevents outsiders from being able to glean information that they could use to attack the organizational network.

9.2 Data Confidentiality

Data confidentiality is perhaps the issue that people concern themselves with most. The thought of proprietary data getting into the hands of competitors or external agencies is normally a fear that will motivate decision makers to ensure that their network is secure. However, most attacks do not attempt to get proprietary data, but instead attempt to get access control mechanisms such as passwords and user names that will allow someone access to the network at a later time. In part, this vulnerability is a direct result of the networking protocols themselves. The protocols normally transmit the data with no sort of basic encryption used. This makes the "sniffing" or viewing of the data a relatively trivial task for someone who is knowledgeable. The extremely paranoid network users (primarily governments) actually place their cables in gas-pressurized lines so that they can detect if someone attempts to tap into the cable in order to view their data. However, this is impractical for most nongovernmental organizations. The best practical method of preventing these types of attacks is to make sure that the networks are physically secure, and to remove some level of trust from the users and replace it with wary education. Most security attacks are the direct result of an internal user either giving their password out to someone, or doing the actual damage themselves.

9.3 Denial of Service

Denial of service attacks are relatively rare, but need to be discussed, if only briefly. A denial of service attack does not attempt to access the data, but instead it attempts to prevent effective use of the equipment from occurring. An example might be to try to fill up the disk drives so that data storage can no longer be done. Another example might be to produce CPU-intensive applications in an attempt to prevent real processing from being done. Again, the best method of preventing these attacks is to make sure that the network is physically secure, and that users are educated against providing access control mechanisms to unauthorized parties. An additional method of gaining access is the physical one, where denial attacks may come from programs brought in by individuals using the network. These programs may contain hidden programs within them, called "Trojan horses," that can destroy important data as well as infiltrate other systems. Programs that spread in this manner are also referred to as "viruses."

9.4 Recovery Plans

The final portion of network security is a backup and recovery plan. No matter how good the security is, it is not possible to plan for all eventualities. In this case, a well thought-out method of backing up data and storing them so that they can be recovered later is invaluable. A very good practice is to back up the data (normally to digital tapes) and then to remove them from the physical proximity of the normal computing facilities or to put them in fireproof/waterproof safes. Off-site storage is by far the safest practice, but is sometimes expensive.

10. NETWORK ALTERNATIVES

While a majority of the users of computer networks seek to connect their networks to larger networks that can provide those services they seek, alternative providers have also proliferated. Such companies as Compuserve, Delphi, America On-Line, and others have packaged network-oriented services for those not directly networked. Using a standard telephone modem (MOdulator, DEModulator), a device that converts digital data into audio, customers of these providers can access the provider database, which is itself linked into the global information base. Some companies also offer connection via local telephone numbers that connect via X.25 protocol networks to the provider's base system.

11. CONCLUSIONS

Computer networks are becoming more prevalent and necessary elements of not only business, but society. They are available in homes, businesses, schools, and governmental organizations. Networks today are already very large, and look only to get larger. Fortunately, the basic concepts of LANs, MANs, and WANs are sufficiently broad, and the concepts are sufficiently advanced, to allow computer networking technology to continue to evolve without being limited by older, legacy technologies.

GLOSSARY

ARPA: Advanced Research Projects Agency.

ARPANET: ARPA-funded research network.

ATM: Asynchronous transfer mode.

CDDI: Copper distributed data interface.

DARPA: Defense Advanced Research Projects Agency.

DIX: Digital, Intel, and Xerox.

DDN: Defense Data Network.

FDDI: Fiber distributed data interface.

IP: Internet protocol.

ISDN: Integrated services digital network.

IXC: Interexchange carriers.

LAN: Local-area network.

MAN: Metropolitan-area network.

NSFNET: National Science Foundation-funded research network.

OSI: Open Systems Interconnect.

SONET: Synchronous optical network.

TCP: Transport control protocol.

WAN: Wide-area network.

Further Reading

Abrams, M. D., Podell, H. J. (1986), *Computers and Network Security*, Washington, DC: IEEE Computer Society Press.

Comer, D. E. (1991), *Internetworking with TCP/IP*, Englewood Cliffs, NJ: Prentice-Hall.

Halsall, Fred (1992), *Data Communications, Computer Networks and Open Systems*, Reading, MA: Addison-Wesley.

Kessler, G. C., Train, D. A. (1991), *Metropolitan Area Networks; Concepts, Standards and Service*, Reading, MA: Addison-Wesley.

Nemzow, Martin, A. W. (1992), *The Ethernet Management Guide: Keeping the Link*, Reading, MA: Addison-Wesley.

Spragins, J. D., Hammond, J. L., Pawlikowski, K. (1991), *Telecommunications Protocols and Design*, Reading, MA: Addison-Wesley.

Stallings, W. (1993), *Networking Standards: A Guide to OSI, ISDN, LAN, and MAN Standards*, Reading, MA: Addison-Wesley.

Stevens, R. W. (1990), *Unix Network Programming*, Englewood Cliffs, NJ: Prentice-Hall.

NEURAL NETWORKS

Tommaso Toffoli, *Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*

Introduction 275
1. **Neurons as Circuit Elements** 276
1.1 Stylized Neurons 276
1.2 Why Neurons? 277
2. **Historical Notes** 277
3. **Tutorial** 278
3.1 Associative Memories 278
3.2 The Hopfield Model 278
3.3 Discussion 279
4. **Structures** 280
4.1 Generalities 280
4.2 Feedforward Networks 280
4.3 Convolution Networks 281
4.4 Autonomous Networks 281
4.5 Boltzmann Machines 281
4.6 Physical Models 282
4.7 Cellular Automata 283
5. **Programming** 283
5.1 Back-Propagation 284
5.2 Energy Function 284
5.3 Simulated Annealing 285
5.4 Learning 286
5.5 Realization Capabilities 286
6. **Tasks** 287
6.1 Recognition, Classification, Generalization 287
6.2 Discovery 287
6.3 Optimization 287
6.4 Inference 288
6.5 Control 288
6.6 Routine Applications 289
7. **Neural "Alchemy"** 289
7.1 Analog versus Digital 290
7.2 Random Interconnection 290
7.3 Programming by Example versus Hard Coding 291
8. **Complex-System Dynamics and Emergent Behavior** 291
8.1 Collective Behavior 292
8.2 Neural-Network Programming and Statistical Mechanics 293
9. **Conclusions and Perspectives** 293
Glossary 294
Works Cited 295
Further Reading 296

INTRODUCTION

Neural networks are digital or analog circuits consisting of a large number of simple elements, and designed in such a way as significantly to exploit aspects of collective behavior—rather than rely on the precise behavior of each individual element. Historically, interest in neural networks arose from the realization that conventional digital computers, in spite of their enormous raw processing power, compare poorly in many tasks with the nervous system of animals. By mimicking structural aspects of the latter, would one more easily reproduce the strong points of its behavior? Thus, neural networks are ostensibly an alternative type of computing hardware, loosely patterned (both in the nature of the circuit elements and in the way they are interconnected) after the animal nervous system.

Today, however, it is becoming clear that—rather than just another type of computing medium—neural networks represent a different *conceptual approach* to computation, depending in an essential way on the use of statistical concepts. In this sense, the theory of neural networks plays in information processing a role analogous to that of statistical mechanics in physics.

From a descriptive viewpoint, we shall be concerned with what neural networks *are* ("structure") and what they *do* ("behavior"). However, we shall never leave far behind questions of a more critical nature: What makes it possible for a certain kind of network to perform certain kinds of tasks? Intuitively, *how* does it work, and *why* should

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

it work at all? This theme is well developed in Hertz *et al.* (1991), to whom we are greatly indebted here.

A typical application for neural networks is to help in making decisions based on a large number of input data having comparable *a priori* importance—for instance, reconstructing the characters on a license plate (a few bits of information) from the millions of pixels of a noisy and somewhat blurred and distorted camera image.

In general, the neural-network approach seems best suited to computational problems of large width and moderate depth—"democratic" rather than hierarchical algorithms. Keep in mind that word recognition out of connected speech—which is a hard task for conventional computers—is performed by our brain with a latency of just a fraction of a second, and thus cannot involve more than a few levels of neurons.

Neural-network design and analysis typically assume a regime of high hardware redundancy. It then becomes both possible and desirable to program a network for a given task by indirect methods (training by example, successive approximations, simulated annealing, etc.). Indeed, the metaphor of a network "learning" its task instead of being "programmed" for it is one of the most appealing—and elusive—aspects of this discipline. By empirical means, it is not hard to come up with a neural-network design that works for a certain toy problem. It is much harder to prove the correctness of the design, and rationally determine its potential and limitations. The importance of theoretical work in this context cannot be overstated.

In spite of the nominal emphasis traditionally being placed on neural networks as a distinguished kind of *hardware*, it must be clear that ultimately one is speaking of a distinguished class of *algorithms*; as a matter of fact, many neural-network applications are routinely and satisfactorily run on ordinary digital computers.

1. NEURONS AS CIRCUIT ELEMENTS

1.1 Stylized Neurons

The human brain consists of about 10^{11} neurons of various types; each neuron typically connects, via an axon that eventually branches out into strands and substrands, to many thousand neurons (Fig. 1). The firing of a neuron is mostly an all-or-nothing business; this discrete character is retained as the pulse travels down an axon. However, upon arrival at a destination neuron the pulse is handled by a synaptic interface characterized by an analog parameter (typically, an excitation or inhibition weight) whose value may be to some extent history-dependent. The complete physiological picture is rather complex.

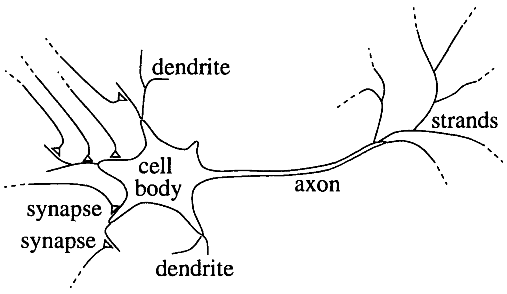

FIG. 1. Schematic structure of a typical brain neuron.

Though neuronlike circuit elements are not strictly essential to neural networks in the modern sense of the term (see Secs. 6.4 and 1.7), historically they have played a dominant role. A drastically simplified model of a neuron, proposed by McCulloch and Pitts (1943), is shown in Fig. 2. The neuron can be in one of two states, +1 and −1, which may be thought of as "on" and "off," or "true" and "false," this state appears at the neuron's output. The inputs may come from other neurons or from external stimuli. State updating may be synchronous (all neurons are

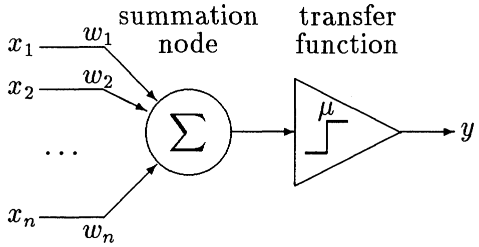

FIG. 2. McCulloch–Pitts neuron. The summation node constructs the weighted sum (with coefficients w_1, w_2, . . .) of the inputs. Depending on whether or not this sum exceeds a threshold μ, an output of +1 or −1 is returned by the transfer function.

updated simultaneously at times $t = 0, 1, 2, \ldots$) or asynchronous (each neuron is updated at random times with a given probability per unit time). The new state of the neuron is determined by the inputs as follows. Input x_j is multiplied by a weight w_j, representing the strength of the corresponding synaptic connection (positive weights correspond to excitatory synapses, negative weights to inhibitory ones). The contributions from all inputs are added and compared with a threshold μ; the neuron turns on if the threshold is exceeded, and off otherwise.

The overall response of the neuron is thus of the form

$$y = g\left(\sum_{j=1}^{n} w_j x_j\right), \tag{1}$$

where g—the *transfer function*—is given by

$$g(x) = \mathrm{sgn}(x - \mu). \tag{2}$$

[By $\mathrm{sgn}(x)$ one denotes the signum function, which returns $+1$ or -1 depending on whether $x \geq 0$ or $x < 0$.]

It is often convenient to replace the step function (2) by a more general transfer function. A continuous function such as the *sigmoid* (Fig. 3), defined by

$$g(x) = \tanh(\beta x) \equiv \frac{e^{\beta x} - e^{-\beta x}}{e^{\beta x} + e^{-\beta x}} \tag{3}$$

(where β is an adjustable parameter), allows one to write formally the partial derivatives of the neuron's output with respect to its inputs (cf. Sec. 5.3): the signum function is recovered in the limit as $\beta \to \infty$.

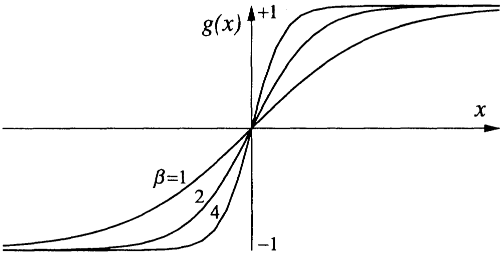

FIG. 3. Sigmoid transfer functions are often used in analog networks. The parameter β gives the *small-signal gain*. The *saturation levels*, ±1, are independent of β.

1.2 Why Neurons?

McCulloch–Pitts neurons are computation universal circuit elements (Minsky, 1967); in fact, they can simulate conventional logic gates such as AND, OR, and NOT. (For instance, with a two-input neuron, weights of 1 for each input, and a threshold of $\frac{1}{2}$ the neuron will fire only when both inputs are turned on, yielding an AND gate.) But why, then, not use gates to begin with?

The neuron is optimized for a different circuit architecture—one that demands elements capable of responding in a nontrivial way to a *large* number of inputs. A generic Boolean function of n inputs can be thought of as a lookup table with one independent entry for each combination of input values; its complexity (intuitively, its material cost) grows exponentially with n. To have a complexity that grows only linearly with n—and thus a general design that is materially realizable even when n is large—one must drastically restrict the class of functions computable by the circuit element. This is achieved by the two-stage design of Fig. 2: a summation node followed by a transfer function. The first stage deals with all n inputs, but only in an additive way; while the second stage, which has only one argument, contributes the required nonlinearity.

2. HISTORICAL NOTES

Neural networks started out, with McCulloch and Pitts' neurons, as an exercise in mathematical biology. The first networks to use many-input neurons systematically were the perceptrons (Rosenblatt, 1962), in which neurons are arranged in regular layers with no feedback from a layer to previous ones (Sec. 4.2). The very book that celebrated perceptrons (Minsky and Papert, 1969) also brought about their untimely death by pointing at a serious limitation of one-layer perceptrons while casting doubts about the existence of systematic learning algorithms for multilayer ones.

After that, interest in neural networks remained sparse for 20 years, with occasional contributions from physiologists (Marr, 1969) and physicists (Caianiello, 1961). The 1980s saw a sweeping revival with new ideas from statistical mechanics and dynamical systems, such as energy function and stable attractors (Hopfield, 1982); new programming techniques, such as the back-propagation learning algorithm (Rumelhart *et al.*, 1986); and, of course, the availability of computing machinery of ever increasing performance.

Today, neural networks are used routinely in many specialized applications, chiefly in low-level image and speech processing, and sensor–actuator integration in motor control; they are also widely used for a variety of noncritical tasks where adequate training by example can be imparted rapidly and economically by nonspecialists: data presorting, screening of applications, poll analysis, quality control, etc. (see Sec. 6.6). However, it still is not clear whether they will ever represent a major component of everyday computation. On the theoretical side, much of the initiative and of the conceptual machinery for fresh developments have been coming from the statistical mechanics community. On the architectural side, computer designs based on elements that are simpler, more numerous, and more heavily interconnected than in traditional architectures are preached by many with almost religious fervor; neural networks fit in very well with the "connectionist gospel" (Hillis, 1985). However, the tenets of this gospel are constantly being revised under the pressure of technological expediency and, ultimately, of fundamental physical constraints: massively parallel architectures tend to favor uniform and local interconnections and limited fanout of signals. Perhaps a final assessment of the potential of neural networks will only come when we know how the brain actually manages to do what it does.

In the meantime, neural networks have matured enough to provide substantial conceptual and practical contributions to the study of the brain itself. This is the domain of *computational neuroscience*, with which we shall not deal here (but see Further Reading).

3. TUTORIAL

As an introductory tutorial, we shall present the Hopfield model (Hopfield, 1982) in its use as an associative memory.

3.1 Associative Memories

In a conventional random-access memory, p patterns $x^1, x^2, \ldots, x^p$ (entries) are stored in such a way that, when presented with the address μ (*key*), the memory returns the entry x^μ. The association between entries and keys is arbitrary; i.e., there are no semantic connections between a key and the corresponding entry. In *associative* (or "contents-addressable") memories retrieval is keyed by a pattern x rather than by an explicit address μ. The memory's task is to return the entry $\bar{x}$ that, according to given criteria, is the "closest match" to key x. It may be reasonable to have the associative memory return "no match" in certain ambiguous situations—for example, if the key is equidistant from the two closest entries ("is \$100 closer to \$101 or to \$99?"), or if it is so far from all of the entries that a match would be meaningless ("is a turnip closer to a book or to a bicycle?").

For sake of illustration, we will take patterns to be bitmap images, each consisting of N pixels (values $+1$ and -1) indexed by $j = 1, 2, \ldots, N$. The "closeness" of the match between two patterns x and x^μ will be measured in terms of their *Hamming distance*, i.e., the number of bit positions on which the two patterns disagree.

3.2 The Hopfield Model

The desired Hopfield model will be constructed with one neuron per pixel position; thus, the collection x_j of neuron states is a pattern. Every neuron is formally connected to all other neurons with a synaptic weight w_{ij} from neuron j to neuron i (of course, lack of physical interconnection can be represented by $w_{ij} = 0$). There are no external inputs. Since information may flow from one neuron back to itself, directly or through other neurons, this is a network with feedback; indeed, since there are no external inputs, it is an *autonomous* network (Sec. 4.4). A threshold at 0 will be used as a transfer function. Thus, the dynamics of the network is given by

$$x_i := \operatorname{sgn}\left(\sum_j w_{ij} x_j\right). \tag{4}$$

Updating, represented by the assignment symbol ":=", is done asynchronously in this case.

Started with an arbitrary initial pattern x and driven by the above dynamics, the network will describe a trajectory through the space of all possible patterns. Our goal is to program the network's weights so that, for any reasonable pattern x used as a key (cf. Sec. 3.1), the trajectory starting from it will eventually settle on the matching entry $\bar{x}$. Intuitively we want to construct around each of the entries a *basin of attraction* (Fig. 4) such that any trajectory starting within a basin will come to rest on the entry itself (the basin's *attractor*).

Analogy with plausible neurological mechanisms (Hebb, 1949) suggests the following weight assignment ("Hebb rule"):

$$w_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} x_i^{\mu} x_j^{\mu}. \tag{5}$$

It turns out that this assignment substantially achieves the goal, *provided* that the entries are sufficiently distant from one another.

Note that increasing the number N of pixels makes more room for distance between the entries, and thus tends to increase the storage capacity of the associative memory. It has been shown that, for a given number N of pixels, recall errors are rare for $p \ll N$ but grow catastrophically as soon as p approaches $0.138N$. If no errors are tolerated, the storage capacity is proportional to $N/\log N$. The cost of the network (measured in terms of the number of synapses, N^2) grows as the square of its storage capacity.

For unreasonable keys (see above), instead of "no match" this network may return a spurious entry (i.e., a pattern not in the set $x^1, x^2, \ldots x^p$). This issue will be further examined in Sec. 5.3.

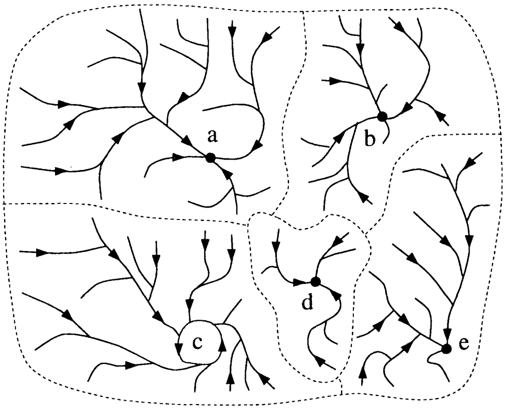

FIG. 4. Basins of attraction. Attractor *c* is a cycle rather than a point.

3.3 Discussion

Let us review the types of activities that were involved in this application, as they are representative of the whole neural network approach.

We started with a well defined *task*—to realize an associative memory for patterns consisting of arrays of N bits.

We then constructed a *structure* hopefully suitable for the task. This was an autonomous, fully interconnected network of McCulloch–Pitts neurons, having the same geometric layout as the pattern themselves.

Next, we *programmed* the network with the desired collection of patterns, using the Hebb rule as a programming algorithm. One could compute and enter the weights [Eq. (5)] by hand. What is remarkable about this algorithm, however, is that it could be easily carried out in parallel by the network itself in a special *learning* mode, as follows. On entering this mode, all weights are cleared. The first entry pattern is then "shown" to the network, i.e., the neuron states are set to the values of the corresponding pixels. Within each neuron, each synapse will compare its input signal with the neuron state; if they are equal, the synapse weight increments by $1/N$; otherwise, it decrements by the same amount. This teaching procedure is repeated for every entry pattern. By the time we are done, the network will have "learned" the correct weights.

At this point, the network can be *used* in its normal operating mode. A "key" pattern is written in the initial state of the network; the network is allowed to evolve according to its internal dynamics until it attains equilibrium; the matching "entry" pattern is read off the final state.

Finally, we *analyzed* the performance of the network in the given task (error probability, storage capacity, cost). We discovered that, as we attempt to cram in more and more patterns, at a certain point the network undergoes a phase transition and suddenly loses its functionality as an associative memory.

4. STRUCTURES

Below, we give a brief classification of neural networks by structure.

4.1 Generalities

In terms of interconnection, a computing network is usually drawn as a directed graph whose nodes represent the computing elements and whose arcs represent the signals flowing in and out of these elements—including external input and output signals (Fig. 5). The names of signal variables, such as *a* and *b* in the figure, are placed next to the corresponding arcs. A fanout node is one that produces at its outputs multiple copies of its single input. A directed path from a node to itself is a *cycle*; in computational terms, a cycle represents feedback.

In the neural-network literature, somewhat different conventions are used in drawing these graphs. In general, a computing element may have many distinct outputs. In a neural network, each neuron has a single output—which reflects the neuron's state (Sec. 1.1). However, copies of this output are typically sent to many destinations. This *fanout* of signals is usually represented as taking place within a neuron; thus, in graphs such as Fig. 6, it is understood that all the arcs leaving a node carry the same signal. It then becomes more practical to place signal labels not on the arcs that carry them but on the nodes that originate them. Since also external input signals usually fan out to different destinations, the source of each external input is indicated by an auxiliary node (a black dot in Fig. 6)—which, however, is only a graphical device and is not counted among the network nodes. Moving signal names from arcs to nodes leaves room for placing on each arc, as a label, the weight of the synapse to which the arc is attached.

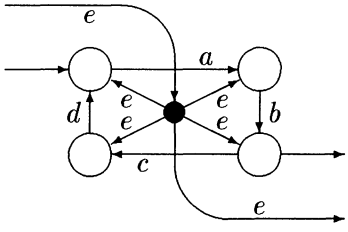

FIG. 5. Nodes and arcs in a directed graph. Here, arcs *a*, *b*, *c*, *d* make up a feedback cycle; the black dot represents a fanout node.

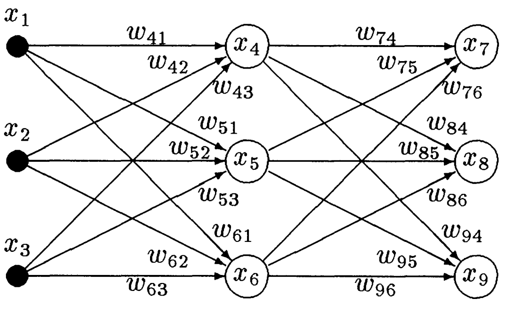

FIG. 6. Two-layer perceptron.

4.2 Feedforward Networks

A network is called *feedforward* if it has no cycles. Such a network acts as a memoryless transducer: information enters the network through its inputs, flows in one direction from node to node, and leaves the network through its outputs. A typical example is the *perceptron* (Fig. 6), where the nodes are arranged in regular layers, each corresponding to a different processing stage.

In principle, a two-layer perception can be programmed to compute an arbitrary Boolean function of its inputs, with the first layer being dedicated to ANDing appropriate subsets of the network inputs or their complements, and the second to ORing the intermediate results. However, this approach corresponds to brute-force lookup. Depending on the nature of the computing task, extra levels may make it possible to achieve the same results in a more structured way and with fewer nodes and wires.

For example, in the *cognitron* (Fukushima, 1975) successive layers are meant to deal with a hierarchy of features in image processing. Within each level, the cognitron architecture supports the idea of *locality*—which nodes are "near" to which other nodes: on each level, a neuron connects only to a small number of neurons of the previous level that are near to one another. At the bottom level, locality is interpreted geometrically; on higher levels, it may refer to distance in some parameter space (slope, contrast, etc.).

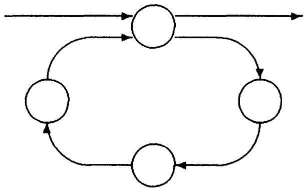

FIG. 7. In a convolution network, some information is recirculated through feedback loops.

4.3 Convolution Networks

Much as feedforward networks, convolution networks have external inputs and outputs, with information generally traveling in the direction from the former to the latter. However, some information is recirculated through feedback loops (Fig. 7), allowing more recently acquired information to interact with earlier information. Convolution networks may be viewed as transducers with memory.

For instance, if the input stream is connected speech and the task is speech segmentation, a convolution network may be used to implement a "sliding window" a few seconds wide. Through taps on this window, the segmentation algorithm can evaluate each input sample in the context of a few earlier and later samples.

4.4 Autonomous Networks

An autonomous network has no external inputs or outputs (Fig. 8); it can be thought of as an isolated dynamical system governed by time-independent dynamics.

For computational purposes, this isolation is broken at time t_0, when the current state is flushed out and replaced by an arbitrary initial state, and then at time t_1, when the system is halted and its final state examined. In this context, an autonomous network is equivalent to a feedforward network consisting of an indefinite number of identical layers (Fig. 9). Each iteration of the original network corresponds, in the new network, to going through one more processing layer.

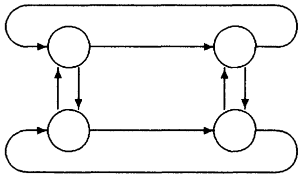

FIG. 8. Autonomous network.

Depending on a delicate balance between reversibility and irreversibility, linearity and nonlinearity, excitability and refractoriness, short- and long-range interconnections, etc., certain autonomous networks may display remarkable invariant features of a statistical character (equilibrium states, attractors, limit cycles, phase transitions, etc.). These features, which are implicitly contained in the dynamical laws, are explicitly "developed" by letting the system evolve; the result may depend in a highly nontrivial way on the initial conditions. When understood and controlled, this correspondence between initial state and developed features provides a rich computational repertoire (cf. Sec. 3).

4.5 Boltzmann Machines

A "canonical" type of autonomous network is represented by Boltzmann machines (Hinton and Sejnowski, 1986). These are networks with symmetric connections and stochastic dynamics, with probability P for a neuron x_i to fire given by an expression of the form

$$P[x_i := 1] = f\left(\beta \sum_j w_{ij} x_j\right), \qquad (6)$$

where β is an adjustable parameter (see discussion in Sec. 5.3). Boltzmann machines (Fig. 10) are more general than Hopfield networks in that they may include among the x_i, beside the variables in whose terms a problem is stated, additional "slack" variables that rep-

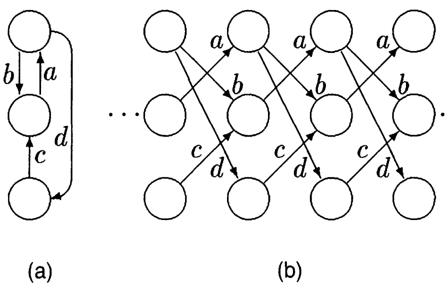

(a) (b)

FIG. 9. By "unrolling in time" the autonomous network in (a) one obtains the equivalent feedforward network in (b).

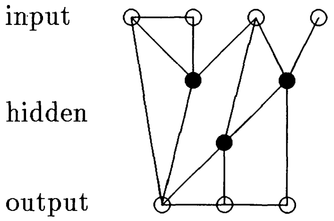

FIG. 10. Boltzmann machine. The hidden units represent slack variables—additional neurons whose state in not initialized or read out by the user. No direction arrows are drawn on the links, since the connections must be symmetrical. The distinction between hidden and visible nodes, and, for the latter, between input and output nodes, is not a structural one, but just reflects the intended use of the network.

resent extra degrees of freedom. The nodes carrying these variables are called *hidden units*—as contrasted to the other nodes, or *visible units*.

Because of the symmetry of connections, Boltzmann machines have a well-defined energy function E (see Sec. 5.2). The corresponding equilibrium distribution is the Boltzmann–Gibbs distribution

$$P(x) = Z^{-1}e^{-\beta E(x)}. \tag{7}$$

The function that expresses the normalization coefficient Z, namely

$$Z = \sum_{\text{all}\,x} e^{-\beta E(x)}, \tag{8}$$

is called the *partition function*, and plays a fundamental role in statistical mechanics (Baierlein, 1971). From distribution (7) one can in principle derive all the statistical properties of the equilibrium state.

The presence of hidden units, which during training are not forced to particular input values and are not requested to converge to particular output values, gives the network more freedom to adjust visible-unit values so as to obey all the given constraints. Without such flexibility, certain problems have no solutions (they are overspecified), and the network cannot be expected to learn to solve them.

4.6 Physical Models

Often, the activity of a neural network can be motivated or illustrated by a physical model—mechanical, electrical, optical, chemical. Sometimes, the physical process itself can be used as a practical way to carry out the desired computation.

Let us consider an array of raw data, $\bar{x}_i$, which are to be smoothed to values $\bar{x}_i$ (Sec. 6.3) by some form of low-pass filtering. In the mechanical model of Fig. 11, the x's are represented by the vertical positions of a number of unit masses, and $\bar{x}$'s by the positions of an equal number of anchor points. The mass at x_i is connected to the corresponding anchor point at $\bar{x}_i$ by means of a spring of stiffness α, and to its neighbors at $x_{i\pm1}$ by springs of stiffness β. Higher-order filtering can be achieved by adding connections to more distant neighbors such as $x_{i\pm2}$.

The cost function to be minimized is the energy stored in the springs, given by

$$E = \tfrac{1}{2}\alpha \sum_i (x_i - \bar{x}_i)^2 + \tfrac{1}{2}\beta \sum_i (x_i - x_{i-1})^2 \tag{9}$$

(this is the usual sum of the squares of the errors). A larger value for β gives greater smoothness; a larger value for α, greater fidelity to the original data. The configuration of least energy, corresponding to the solution of the problem, is easily found by physical means. In fact, as soon as one introduces some damping in the system [viz., a dissipative term in Eq. (9)], this configuration becomes an attractor of the dynamics (cf. Sec. 3.2): from any initial configuration for the x_i, the network will converge toward it. The fastest convergence rate is achieved with critical damping.

The same solution can be arrived at by means of the electrical model of Fig. 12. The $\bar{x}$ and x data are represented by voltages. The current through each resistor is proportional to the voltage difference between the two ends of the resistor, while the dissipation through the same resistor is proportional to the square of the current. At equilibrium, the voltages

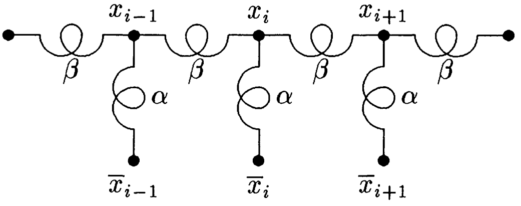

FIG. 11. A mechanical network for "smoothing" a one-dimensional array of raw data. The $\bar{x}_i$ are the raw data, the x_i, the conditioned ones.

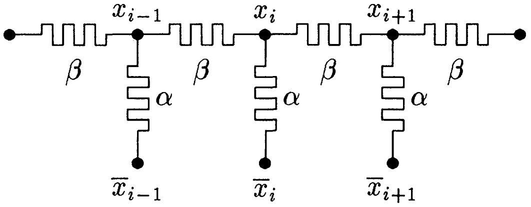

FIG. 12. An electrical model for data smoothing.

are such as to minimize the total dissipation in the network, which is given by a formula identical to Eq. (9).

A typical visual scene contains smooth areas bounded by sharp edges; the latter would be degraded by a smoothing-only algorithm. In order to enhance the edges, a cost function such as Eq. (9) could be augmented with appropriate "reward" and "penalty" terms. A more sophisticated approach is to let overstressed springs "snap" (or, in the electrical model, overheated wires "burn"); on a second pass through the optimization algorithm, the two severed ends would be free to converge to different values. See Koch *et al.* (1986) for an extensive discussion of methods of this kind in early vision processing.

4.7 Cellular Automata

In contrast to the "engineering" flavor (springs, resistors, etc.) of the above models, cellular automata (Burks, 1970) provide a computing-network metaphor for more universal aspects of physics—such as uniformity, locality, spacetime structure, inertia, microscopic reversibility, etc. (Toffoli and Margolus, 1987).

For our purposes, cellular automata are autonomous networks (Sec. 4.4) obeying the following constraints (Fig. 13). Nodes are arranged in a uniform array. To each node, or *cell,* there is associated a state variable ranging over a finite set—typically just a few bits' worth of data. Time advances in discrete steps, and the dynamics are given by an explicit rule—say, a lookup table—through which at every step each cell determines its current state from the current state of its neighbors. Gone are two more typically "neural" features of neural networks, i.e., arbitrary interconnection, and provisions for an arbitrary number of inputs at each node (cf. Sec. 1.2). Communication between distant nodes is still possible, but through a sequence of action-by-contact interactions rather than by action-at-a-distance.

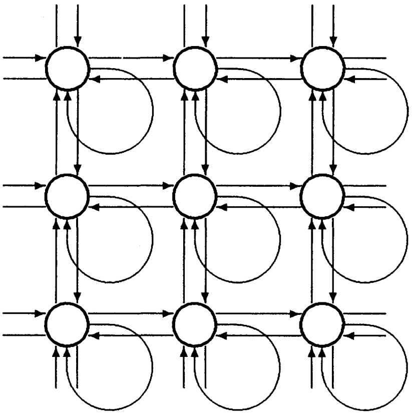

FIG. 13. A two-dimensional cellular automaton with first neighbor interconnection. Note the feedback loop from each node to itself.

In spite of important differences in motivation and emphasis, the cellular automata approach has much to share with that of neural networks; they are both concerned with the synthesis of complex macroscopic behavior out of the collective action of a large number of simple primitives (Wolfram, 1986).

Within the connectionist tradition, the cellular-automation scheme of computation is unique in the way it lends itself to extremely efficient concrete implementations. This feature provides a strong incentive to recast many neural-network techniques into a cellular-automaton format. Of course, cellular automata are the ideal substrate for many computing activities having a strong spatial component (e.g., image manipulation).

5. PROGRAMMING

An ordinary computer is programmed by writing an explicit sequence of instructions: the programmer should be able to describe and justify what each line of code "does." The typical neural network is programmed for its task by assigning the synaptic weights. Note, however, that a synaptic weight reflects the projection on a one-dimensional scale of parameters that inhabit a multidimensional space, and in turn will affect the values of a large number of computation variables. The programming of such weights is typically done by indirect means, usually by iterating a training procedure that is applied in parallel to the entire network. For example, we have seen in Sec. 3.3 that an associative memory *à la* Hopfield may be programmed by "showing" to it, in learning mode, the desired target patterns; each pattern leaves an "imprint" of itself in the form of an additive contribution to all the network weights.

A simple analogy for this situation is found in the Fourier transform $\mathcal{F}:\{f \rightarrow g\}$. Information about each point of the original function f is thinly and uniformly spread over the entire range of the new functions g. Intuitively, every point of g contains an infinitesimal amount of information about every point of f, and vice versa.

The most general indirect method of programming is, of course, trial and error: the weights are adjusted until the desired input/output behavior (or a minimum in an appropriate error function) is attained. Given the exponential size of the network parameter space, such an approach is hopeless unless one can rely on some sense of direction. Usable convergence criteria vary enormously depending on the nature of the network, the nature of the application, and the sophistication of the available theoretical tools. Except in very special cases (e.g., linear programming), in order to achieve a *global* minimum of the error function one needs strategies to avoid entrapment in *local* minima. These two concerns, namely local optimization by continuous means and global exploration by judiciously timed jumps, are illustrated below by, respectively, the back-propagation algorithms and the simulated annealing algorithm.

5.1 Back-Propagation

In the two-layer perceptron of Fig. 6, let the indices i, j, and k run respectively over the output nodes (second layer of neurons), the "hidden" nodes (first layer), and the network inputs; thus, for instance, w_{ij} will denote the synaptic weight of the connection from hidden neuron j to output neuron i.

Let x^μ denote an input pattern ($\mu = 1, \ldots, p$), y^μ the corresponding output pattern for a given set of weights, and $\bar{y}^\mu$ the desired output pattern for that input. The overall error between actual and desired response will be measured by the following error function (sum of squares):

$$E = \tfrac{1}{2} \sum_{ij} (y_i^\mu - \bar{y}_i^\mu)^2. \tag{10}$$

With a neuron of the form in Eq. (1), the variable term in Eq. (10) is

$$y_i^\mu = g\left(\sum_j w_{ij} g\left(\sum_k w_{jk} x_k^\mu\right)\right). \tag{11}$$

Under the assumption that g is a differentiable transfer function such as the sigmoid (Fig. 3), E is a continuous, differentiable function of every weight, and can be minimized by letting the w's slide in the direction of the gradient ("gradient descent"). Similar considerations apply to the case of more than two layers.

Because of the nesting of weights from different layers in Eq. (11), it is convenient to compute the errors and consequently adjust the weights layer by layer, starting with the output layer and proceeding toward the input. This *back-propagation algorithm* is local (weight adjustments for each neuron can be performed on the basis of the information available at the neuron's input and output), and is not computationally very demanding (if n is the number of synapses, it requires calculating order n derivatives, rather than order n^2 as in the minimization of E by simultaneously adjusting all the weights).

Many variants of the back-propagation algorithm have been studied, the main objectives being faster convergence and avoidance of local minima. Overall, theoretical and empirical results confirm the viability of this general strategy.

5.2 Energy Function

The following discussion is complemented by the entries on annealing, dissipation, energy, energy function, entropy, and temperature in the Glossary.

The Hamming distance between two patterns (Sec. 3.1) is defined independently of what patterns we are interested in retrieving. Can we define a real-valued function on individual patterns such that the resulting ordering of patterns would provide a "sense of direction" in solving a specific problem? In physics, energy plays such a role. For instance, in a conservative system, state b can be reached from state a only if $E(b) = E(a)$; in a dissipative system, only if $E(b) \leq E(a)$. Moreover, in a dissipative system, the states for which E has a local minimum are the attractors of the dynamics.

The idea of energy function in neural networks was pioneered by Hopfield (1982). In the example of Sec. 3, with weights w_{ij} given by the Hebb rule as in Eq. (5), let us define the *energy* of a pattern x by

$$E(x) = -\tfrac{1}{2} \sum_{ij} w_{ij} x_i x_j. \tag{12}$$

Then

1. the stored patterns $x^1, \ldots, x^p$ correspond to local minima for E, and
2. the dynamics in Eq. (4) are suitable for finding these local minima, since on each trajectory the energy E decreases or remains constant.

A sufficient condition for a quantity of the form of Eq. (12) to be an energy function is that the weights be symmetrical, i.e., $w_{ij} = w_{ji}$ (which is the case in the Hopfield model). See Sec. 4.5 for further details.

The importance of the energy function is that it gives a descriptive rather than a prescriptive characterization of the solutions (cf. variational principles vs. dynamical laws in physics). The problem is stated in terms of an energy landscape, and one is free to develop different strategies to explore this landscape in a search for solutions. Even one's idea of what should constitute an "acceptable solution" may be influenced by the nature of the landscape.

5.3 Simulated Annealing

Simulated annealing (Kirkpatrick *et al.*, 1983) is a broad strategy for searching for deep minima in an abstract energy-function landscape. It mimics the controlled, gradual cooling of a sample of material aimed at relieving internal stresses. If the sample is cooled too rapidly ("quenching"), it may remain frozen in a high-energy metastable state very close to the initial state; if it is cooled too slowly, it may tumble all the way down to the ground state (the absolute minimum of energy), thus erasing all memory of its initial state. Different annealing schedules allow one to "dial in," as it were, terminal states having different degrees of stability.

In physics, temperature measures the intensity of the randomizing influences that affect a system's nominal dynamics. As remarked above, in the deterministic neuron dynamics of Eq. (4) the value of a pixel is complemented if and only if the new configuration has less energy—no matter how small the energy difference. To represent temperature in the network, we introduce a stochastic dynamics derived from Eq. (4) as follows.

The output from the summation node in neuron i,

$$h_i = \sum_j w_{ij} x_j, \tag{13}$$

represents the "tendency" (on a scale from $-\infty$ to $+\infty$) for the neuron to fire. We rescale this tendency to the interval $(-1,+1)$ by sending h_i through a sigmoid of the form of Eq. (3), and we interpret the result as the expectation value of the neuron state. The corresponding probability P for neuron i to fire is given by

$$P[x_i := 1] = 1/(1 + e^{-2\beta h_i}). \tag{14}$$

For $\beta = \infty$, Eq. (14) yields back Eq. (4): there is no thermal agitation. For $\beta = 0$, the neuron will go with equal probabilities to states $+1$ or -1: thermal agitation completely swamps any internal tendencies. In fact, β can be identified with the temperature parameter—and its inverse, $T = 1/\beta$, with the absolute temperature—of a statistical mechanical system (Katz, 1967).

In simulated annealing, one starts the search for a minimum of the energy function with the system at a high temperature, and gradually reduces it. The details of this thermal schedule influence not only the speed with which a solution is reached, but also the nature of the solution.

Suppose there is a path of monotonically decreasing energy from point P to point Q; even if P and Q are near to one another, this path may be long and circuitous because of energy barriers that block more direct paths. (Given the high dimensionality of the typical problem space, the number of paths from P to Q may be astronomical.)

If the temperature is high, the system can easily go across high (but thin) barriers, thus availing itself of much shorter paths between P and Q. Intuitively, at temperature T the system sees an energy landscape smoothed out (cf. Sec. 4.6) on a scale $\Delta E \approx T$, and may be made to follow an energy-function gradient averaged out on this scale; finer details are virtually ignored. Once the approximate bottom of a valley is thus rapidly reached, one may reduce the temperature, thus exposing ripples on a finer scale and allowing further descent.

On the other hand, if the initial temperature is too high, the system may escape from the original basin of attraction and land in a different basin. Though desirable in certain circumstances, such wholesale erasure of initial-state information would clearly defeat the purpose in the associative memory task discussed above.

5.4 Learning

We have seen that neural networks are typically programmed by indirect means (i.e., without direct external access to the synaptic weights), for which "teaching" is an appropriate metaphor. Actually, the term usually employed for this process is "learning"—which emphasizes the network's rather than the programmer's viewpoint. Intuitively, the programmer influences the network's "genotype" by applying pressure on its "phenotype." Unlike biological evolution, besides mechanisms of a Darwinian kind—or *genetic algorithms* (Holland, 1975)—like simulated annealing (Sec. 5.3), it is sometimes possible to set up more efficient, Lamarckian mechanisms, like back-propagation (Sec. 5.1).

In supervised learning, the network's role is to internalize appropriately information presented by the teacher during a training session: explicit data to be memorized (cf. Sec. 3), examples of the desired behavior, or scores evaluating the current behavior of the network and aimed at steering it toward the desired one (cf. Sec. 6.4). Once the training phase is over, the values of the network's parameters are "set"; the network is then ready for the operational phase, where the network's state variables come into play.

Some neural networks may be designed to operate in an unsupervised learning mode. Typically, the network will be expected to identify and extract significant features of the input stream without the complicity of a teacher, and possibly to apply this knowledge to subsequent segments of the input stream (cf. Sec. 6.2). Here, of course, the distinction between network parameters and state variables depends on the point of view and on the time scale.

5.5 Realization Capabilities

Without loss of generality, the target task for a neural network can be thought of as a function f to be computed, and the programmed network as an approximate realization of this function, obtained by composition of given primitives. Thus, the study of what can be computed by a neural network is basically a chapter of functional analysis.

It is true that the approximation error ϵ_{approx}—i.e., the distance between the target function and the function computed by the network when optimally programmed—can be made as small as desired by increasing the number n of nodes; however, as n increases, the number N of trials required to train the network must increase as well.

Typically, one does not explicitly know the target function. Moreover, though the actual network function is in principle knowable, a detailed knowledge of it becomes more and more expensive as the number of nodes increases. What one can reasonably demand is an estimate of this function, subject to a certain estimation error ϵ_{estim}. Thus, a reliable programming methodology must have access to a way to assess both the approximation error and the estimation error.

Such assessments have been made by Barron (1991) for networks with a sigmoidal transfer function (Fig. 3), under specific smoothness assumptions for f, yielding

$$\begin{aligned}\epsilon_{\text{approx}} &\approx 1/n,\\ \epsilon_{\text{estim}} &\approx (nd/N)\log N\end{aligned} \qquad (15)$$

(where d is the number of arguments of f). The approximation error is inversely proportional to the number n of nodes; however, for the two errors to remain comparable, one needs

$$N/\log N \approx n^2,$$

i.e., the number N of trials must increase somewhat faster than the square of the number of nodes. This sets a practical limit on the number of nodes.

The above bounds provide qualitative indications of the trade-offs that one may expect in more general situations.

6. TASKS

Below, we briefly review typical tasks for neural networks. The same kind of abstract task (e.g., optimization) can be found in different applications (e.g., vision, control, scheduling). Vice versa, many applications entail, at different stages, different kinds of tasks; for example, automated vehicle driving may require first recognition of the landscape and then appropriate control responses.

6.1 Recognition, Classification, Generalization

Recognition and, more generally, classification are among the tasks that are most commonly pressed upon neural networks. Usually, the classes one is interested in are not predefined by formal criteria, but happen to be distinguished from one another by the values of a large number of variables—even though some overlap may occur in some of the variables. (For example, an overwhelming number of factors may lead one to classify a kiwi as a "bird," even though the lack of wings, by itself, would tend to disqualify it.)

In this context, writing an explicit computer program to do the classification would be an immense and at the same time ill-defined task. Too much cleverness in devising compact decision tables might lead to untenable criteria—classification dead ends—as new data accumulate. In a sense, specifying a neural network (such as a perceptron or a Hopfield model) for this task is a way of giving an implicit definition of the classes themselves, based on the training data's statistics and "internal evidence" rather then externally imposed criteria.

When the basins of attraction (Fig. 4) are large, so that they encompass patterns that must be very different from the "type" attractor, classification turns into generalization: is an airplane more like a car or like a bird? Again, a specific neural network in this context must be viewed as a creative way of *proposing* generalization criteria where none are extant, rather than *implementing* given ones. In special situations, this generalization mechanism may lead to such consistent, objective results that it qualifies as an inference tool (cf. Sec. 6.4).

6.2 Discovery

Most unsupervised learning is aimed at identifying information-rich descriptors for a collection of objects or a distribution of samples. For example, in police files, "hair color" conveys much information in France but little in Iceland or China. The prototypical task in this area is *principal-component analysis*: in a distribution over a vector space of dimension N, is there a subspace of dimension $M < N$ that accounts for a large fraction of the data variance?

A simple approach to this problem, involving an M-node, one-layer, linear feedforward network, was proposed by Sanger (1989). The node's response is

$$y_i = \sum_j w_{ij} x_j, \tag{16}$$

and the incremental learning rule is

$$\Delta w_{ij} = \eta y_i \left(x_j - \sum_{k=1}^{i} y_k w_{kj} \right) \tag{17}$$

(where η is a learning rate parameter). By iterating this rule, the M vectors that make up the $M \times M$ matrix w_{ij} converge to a set of orthonormal vectors that represent the first M principal components of the distribution.

The principal application of this kind of analysis is adaptive data compression.

6.3 Optimization

In classification, one seeks a pattern (from a collections of "types") from which the given pattern has minimal distance. In optimization, one seeks the (unknown) pattern that minimizes the value of an assigned *cost function*. Optimization has long been the chief concern of operations research (*q.v.*), and has spawned a prodigious amount of literature.

In this field, one can expect novel contributions (whether conceptual or practical) from neural networks only in special circumstances: a cost function that depends on a large number of variables, modest coupling between variables, willingness to accept as a solution a "reasonably low" value of the cost function (rather than an absolute minimum), or stress on practical implementability and real-time performance.

The benchmark for combinatorial optimization methods is the traveling salesman problem, which provides nontrivial constraints. Neural-network approaches to this problem have been proposed by Hopfield and Tank (1986) and many others. The emphasis is, for the moment, conceptual, since the results are not generally competitive with more conventional algorithms. Useful results based on optimization algorithms are routinely achieved in more down-to-earth tasks, such as are found in early visual processing (cf. Sec. 4.6).

6.4 Inference

What comes after 1, 2, 3, 4? After 2, 4, 6, 8? After 1, 4, 9, 16? For each of these sequences, though many continuations are defensible, one stands out so compellingly that we view it as the "right" answer. Can one make neural networks reason in this fashion, and infer from a limited number of examples the entire set or sequence from which the examples "must" have been drawn?

A good paradigm for this kind of task is provided by an experiment discussed by Carnevali and Patarnello (1987), where the objective is to teach a network to perform binary addition by showing to it a number of examples. For definiteness, consider a box with two eight-bit inputs, x_1 and x_2, and one eight-bit output, y. The box contains a few hundred NAND gates (here is an example of a "neural network" consisting of Boolean gates rather than McCulloch–Pitts neurons), which are to be connected as a feedforward network (Sec. 4.2) from the inputs x_1, x_2 to the output y. For any given interconnection pattern, the box will compute a certain function f of the arguments x_1, x_2. The domain of f consists of $K = 2^8 \times 2^8$ points, and f itself can be identified with a lookup table of K eight-bit entries.

Our goal is to arrive at an interconnection pattern for which the box will compute the function

$$y = \bar{f}(x_1,x_2) \equiv x_1 + x_2 \pmod{2^8}. \tag{18}$$

[For simplicity, we ask that addition in Eq. (18) be performed modulo 2^8—i.e., ignoring the carry from the most significant digit—so as to have an eight-bit result.]

In the experiment, the network interconnection pattern, initially random, is gradually steered toward the goal by a simulated annealing procedure (Sec. 5.3). However, our only guide in this steering will be provided by the value of f on a small number k of "test points" (to be specific, let $k = \sqrt{K}$), where we strive to make f coincide with $\bar{f}$. That is, as an energy function we shall use the Hamming distance between f and $\bar{f}$, but computed only on the k test entries rather than on all the K entries of the table. The details of the thermal schedule are not essential here. The training will terminate when the network yields the correct results (i.e., $y = x_1 + x_2$) on the k test points.

Once the training is over, almost invariably the network will be found to compute $\bar{f}$ correctly not only on the k test points, but on all the K input points: the network has successfully inferred what function the samples came from!

Is such degree of success an exception or the rule? We defer discussion of this remarkable result to Sec. 7.3.

6.5 Control

Consider a robotic arm consisting of n consecutive segments, as in Fig. 14. The near end of the first segment is hinged on a base plate, while the far end of the last segment carries a tool. The tool's position in space x

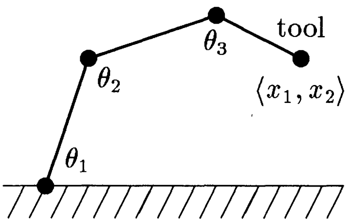

FIG. 14. Robotic arm with three joints, in two dimensions.

$= \langle x_1, \ldots, x_m \rangle$ is a function $x = f(\theta)$ of the set of joint angles $\theta = \langle \theta_1, \ldots, \theta_n \rangle$ (where θ_1 is the angle between the base plate and the first segment, θ_2 that between the first segment and the second, and so forth). The number of joints is usually somewhat larger than the number of degrees of freedom of the tool end, so that the same tool position may be achieved by different settings of the joint angles.

While the driving software has direct control of the joint angles θ, the desideratum is the tool position x. A typical control problem, which we shall briefly discuss here for sake of illustration, is the *inverse kinematics* problem (Craig, 1989); namely, for any desired x, to find a θ for which $x = f(\theta)$.

The function f itself is explicitly given in closed form and is easy to compute (the θ_i are handled one at a time in sequence). However, inverting f (a nonlinear, many-to-one function of many variables) is an awkward task, for which semiempirical methods based on neural networks are well suited.

Learning feedback may be provided by a camera that compares the actual position x' achieved by the tool with the target position x. In the specialized learning approach discussed by Psaltis *et al.* (1988), one makes do without a camera by taking advantage of the fact that f is a known, continuous function of x. In Fig. 15, g represents a feedforward neural network which, given a target position x, strives to output an appropriate value for θ. Learning feedback for g is provided by another feedforward network, f, which encodes the known function $x = f(\theta)$ and does not need any learning itself (its parameters are fixed). If the value of θ produced by g is such that $x = f(\theta)$, then $x' = x$; if not, the error $\epsilon = x' - x$ is back-propagated first through the stages of f, so as to suggest a new value for the intermediate target θ, and then through g itself, so as to suggest new settings for g's synapses.

Kinematics, is, of course, but a first step. More demanding control tasks to which neural networks are being applied include dynamics, trajectory planning, and integration of vision with mechanical control.

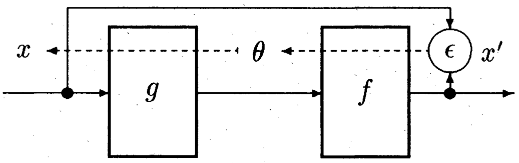

FIG. 15. Feedback loop arrangement in specialized learning (after Kröse *et al.*, 1993).

6.6 Routine Applications

In numerical terms, perhaps the most common application of neural networks is coarse, routine, noncritical classification of objects or events, where training (typically, programming "by example" using some variant of the back-propagation method) can be carried out in a casual and inexpensive way by nonspecialists. For example, which tax returns should be earmarked for a second look by the auditor? The selection criteria are easily updated simply by adding to the training set any returns that seem for some reason remarkable. If the network errs by selecting an unsuspicious return, a few minutes of auditor time will be wasted; if it errs by overlooking a "fishy" return, no liabilities will be incurred.

Inexpensive hardware and software are available that can handle a variety of nondemanding recognition, classification, generalization, and inference tasks. The usefulness of these tools is, of course, highly dependent on the quality of the documentation and the skill of the user.

Besides professionally approached applications, there is room for an enormous variety of applications where some "roll your own" preprocessing by neural networks, no matter how rudimentary and how naively programmed, may have (at least for some of the parties involved) an economical advantage over no preprocessing at all: real-estate searches, mailing lists, insurance application screening, preventive health screening, quality control, inventory control.

7. NEURAL "ALCHEMY"

Many a hardware or architecture feature was at one time or another touted as "the magic element" that somehow would give neural networks properties that transcended those of conventional computers. Here we shall try to separate the real relevance of these features from their mythological aspects.

7.1 Analog versus Digital

Analog circuitry vied seriously with digital circuitry in the early days of scientific computation. Today, analog circuitry is still competitive in certain specialized real-time tasks, such as TV signal processing.

Analog elements are capable of impressive nonlinear feats: for example, a matched transistor pair can be used to compute accurately logarithms or exponentials with a dynamic range of 10^8. However, neural networks are not geared to exploiting such sophisticated "calculus" building blocks; all the nonlinearity they need can be provided by low-accuracy "squashing" functions (Figs. 2 and 3). In neural networks, most of the action is in a linear analog operation, i.e., the weighted sum of n arguments (cf. Sec. 1.2); this can be implemented by a simple passive device, namely, a network of n resistors tied together into a summing junction (Fig. 16). This device only need be followed by an inverting amplifier/limiter to yield a complete neuron (Fig. 16) having computing capabilities equivalent to those of a McCulloch–Pitt neuron. (The adjustable resistors of Fig. 16 only provide positive weights; the effect of a negative weight is achieved by driving an input with an inverted signal. Note that the inverting amplifier already provides an inverted output signal; for a noninverted signal, one would use one more inverting stage.)

Note that, with fixed input resistors of equal value and appropriate threshold bias, this circuit becomes an n-input NOR gate. Indeed, RTL (Resistor Transistor Logic) digital gates are realized in this fashion.

The threshold provided by the single transistor stage in Fig. 16 is not very sharp; indeed, one can think of it as a sigmoid with a maximum slope of approximately ten (cf. Fig. 3). While two or three cascaded limiting stages would give an excellent approximation to a digital output, the single-stage circuit can be used as an analog-input, analog-output neuron.

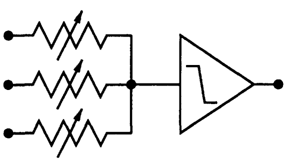

FIG. 16. A neuron consisting of a voltage summing junction (the adjustable resistors set the synapse weights), and an inverting, limiting amplifier.

If, with the same amount of machinery as a digital circuit, a neural network can handle real-valued variables, does it not have, in a sense, "infinitely more" computing power? Can something like this account for the brain's performance?

There is no doubt that, given comparable resources, a verbatim physical implementation of an analog computation may in certain cases outperform by a few orders of magnitude a digital simulation (e.g., by floating-point circuitry) of the same computation. This point has been well argued by Mead (1989). What should be clear, however, is that the advantage—both in terms of capacity and speed—is in the form of a *constant* factor. In fact, because of thermal noise and fabrication tolerances, the nominally continuous range of an analog variable is actually equivalent to a modest number of distinguishable states. Moreover, when one changes one of the inputs in Fig. 16, the new voltage at the summing node is approached exponentially with a time constant τ_{analog}: to have k significant digits, one must wait for a time $\approx k\tau_{\text{analog}}$. If the same input data were encoded as binary strings and processed by a serial digital adder with a clock period τ_{digital}, one would get k digits in a time $\approx k\tau_{\text{digital}}$.

Thus, the analog aspects of neural networks may be good heuristics and may be retained to advantage in a physical implementation, but they do not hold the key computational capabilities that transcend those of digital networks.

7.2 Random Interconnection

Formally, a neuron can directly connect with as many other neurons—and as distant—as desired. Random interconnection is often regarded as a distinguishing trait of neural networks, and one that could give them special computational capabilities. However, even though random interconnection may be a useful metaphor at the conceptual design stage, eventually the material cost of interconnection must be reckoned with.

With no constraints on interconnection, the number of distinct nodes that can be reached from a given node in n steps grows exponentially with n. In a three-dimensional world, if we assume that a "step" represents a fixed unit of cost (such as transmission time or dissipated power), this number can only grow polynomially, i.e., as n^3. The cost of embedding an exponential network in a polynomial space soon becomes a dominant aspect of design: our brain devotes perhaps nine-tenths of its resources to interconnection.

In light of the above considerations, the strict discipline of the cellular-automaton network format (Sec. 4.7) no longer appears a capriciously imposed handicap: the contrast between "random interconnection" and "local and uniform interconnection" is one of heuristic approach, not of computational capabilities.

7.3 Programming by Example versus Hard Coding

A special appeal of neural networks is the new programming style they bring in (cf. Sec. 5.4). By giving a number of concrete examples, one might be able to tell a network *what* to do without having to explain in detail *how* to do it.

To evaluate the reasonableness of these claims in a particular context, one must ask a number of questions. How precisely is the desired task identified by a given set of examples? Does the network have, to begin with, adequate resources to carry out the task it is supposed to learn? Is the training procedure capable of reaching its goal, and of doing so in an acceptable time?

These questions are hard to answer without making unduly restrictive hypotheses. Typical of the state of the art is a thorough analysis by Baum and Haussler (1989). This is concerned with the task of generalizing, from training examples, a pass/fail criterion to be applied to items occurring with a certain probability distribution; its conclusions may be used as an indication of what one can reasonably expect in a wider range of tasks.

Here, we shall paraphrase a study by Carnevali and Patarnello (1987), chosen for its direct intuitive appeal, which throws light on the potential and the intrinsic limitations of programming by example.

With reference to the model of Sec. 6.4, consider the space S of all possible interconnection patterns (which here we shall call *structures*) achievable with n gates; the size N of this space increases exponentially with n. For a given n, certain functions will be realized by many different structures; others by few; while others may have no realization at all. Let S_f be the set of structures that realize a function f, and N_f its size. Intuitively, if we are wandering at random through S, the chances of finding a structure that implements f are proportional to N_f/N. Indeed, we can use $\Delta S = \log(N/N_f)$—the amount of information that must be supplied by the training procedure to home in on the subset S_f—as a measure of the complexity of the problem f. Remarkably, numerical evidence suggests that ΔS is to a good approximation independent of n.

Given an appropriate energy function E for f, simulated annealing will lead one toward structures of lower and lower energy, i.e., structures that compute f to greater and greater accuracy. Indeed, if S_f is not empty, one will eventually reach a structure for which $E = 0$: the network will have "learned" f, essentially by rote memorization.

To attempt inference, we provide an energy function E' that only reflects the network's error on a fraction k/K of f's domain. Annealing will then home in on a superset S'_f of S_f, of size N'_f. The probability that a structure for which $E' = 0$ will also be one for which $E = 0$ is N_f/N'_f. Thus, inference will be successful only on those problems for which the number of solutions increases slowly as we relax the specifications.

In conclusion, both the cost of learning and the probability of successful inference depend on the underlying combinatorics—how many different structures realize a given function—and thus widely vary from problem to problem. Learning a problem by rote has a cost ΔS that reflects the relative difficulty (with respect to all possible problems) of putting together a solution with the given resources; once that price is paid, learning is all but guaranteed. On the other hand, the chances of correct inference depend on the rate of change of a problem's difficulty as its specifications are relaxed; this rate is presumably very hard to estimate *a priori*.

8. COMPLEX-SYSTEM DYNAMICS AND EMERGENT BEHAVIOR

Historically, much neural network research has been conducted with ostensibly utilitarian goals; a neural network was supposed to be "for something." Complex-system dynamics may be viewed as the study of neural networks for their own sake. What are the modes by which complex macroscopic behavior may emerge from the collective action of microscopic components? What kinds of behavior are more likely to arise in this way?

8.1 Collective Behavior

Here we shall use the term *assembly* for any system consisting of a large number of simple, identical nodes interconnected in a simple, regular way. In other words, an assembly is a network whose structural "texture" is the same—or at least has the same statistics—everywhere. Macroscopic properties of an assembly are, by definition, those obtained by coarse-grain averaging—i.e., by counting the occurrences of individual node states over volume elements that encompass a large number of nodes but are still much smaller than the entire assembly. Finally, by an *architecture* we shall mean the set of all assemblies that have a given basic structure but differ in the values of certain adjustable parameters. For instance, the set of all two-dimensional cellular automata having two states per cell and first-neighbor interconnection is an architecture whose variable "parameter" is the cell rule itself (see Sec. 4.7).

In Sec. 7.3, we concluded that a predisposition for a network to learn certain tasks must be balanced—in what is in effect a constant-sum game—by an aversion to learning other tasks. Analogous trade-offs must be expected in emergent phenomena. Only if the behavior we desire belongs to the "natural" macroscopic repertoire of a given architecture do we have a good chance to elicit it by simply adjusting a few parameters. (Otherwise, to obtain that behavior we will have to program the network explicitly at a microscopic level—which, of course, is what is done in traditional digital-circuit design.)

In this regard, cellular automata provide instructive and well-understood examples (cf. Fig. 17). It must be noted that, for a given cellular automata architecture, almost all rules (see Glossary) lead to trivial macroscopic behavior: i.e., from most initial states the system immediately attains equilibrium in the form of a maximally random state. For nontrivial behavior to emerge, the rule must obey constraints that commute with coarse-grain averaging. Typical constraints of this kind are conservation laws and symmetries. For example, if a microscopic rule is translation-invariant, so must be all macroscopic properties.

For two-state cellular automata, let us consider the small subset of rules that conserve the number of 1's (thought of as "particles" in a "vacuum" of 0's). Particle conservation clearly withstands the ravages of coarse-grained averaging. It turns out that from almost all these rules there emerges a characteristic macroscopic behavior; namely, the evolution of particle density obeys the diffusion equation [Fig. 17(a)]. Further, within this subset of particle-conserving rules, let us

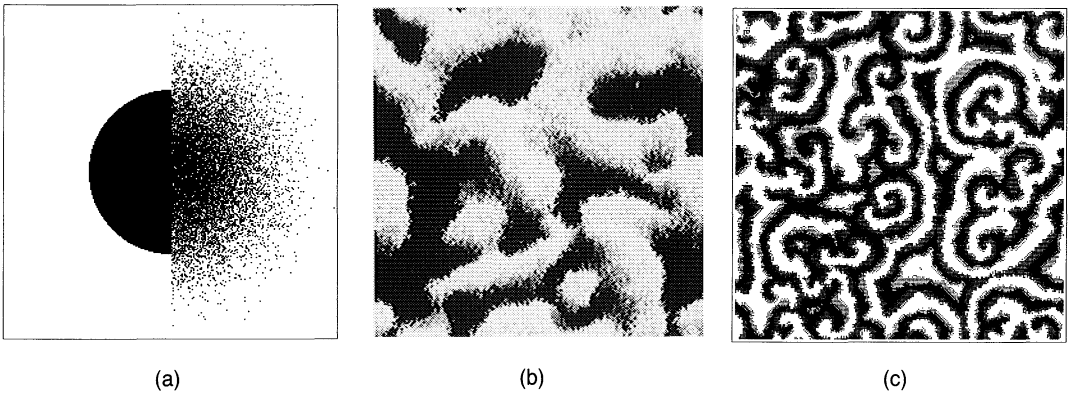
(a) (b) (c)

FIG. 17. Samples of emergent behavior in cellular automata. (a) Diffusion; this is a composite figure, the left half from the initial state (a solid disc), the right half from a later state. (b) Phase separation, in three dimensions, as a consequence of a "voting majority" rule. (c) Traveling waves in an excitable medium.

look at the small fraction that depart from diffusive behavior. These are rules that obey additional constraints. For example, those that conserve momentum (the x-component of momentum is the number of particles traveling in the positive x direction minus that traveling in the opposite direction) macroscopically lead to the Navier–Stokes equation, i.e., hydrodynamic behavior (Frisch *et al.*, 1986)—except, of course, for a small subset that obey further constraints; and so forth.

In Sec. 7.3, the structures that computed binary addition correctly 10% of the time were very few (with respect to the set of all possible structures), but, of these, almost all did it correctly a full 100% of the time: intuitively, on this task it takes some rare perversity to manage to go part of the way without going all the way. Analogously, in cellular automata one has to look hard for rules that conserve particles; but when one is found, it is a rare conspiracy that will make the rule not yield the diffusion equation.

In brief, while certain traits of the rule are essential to obtaining a desired type of macroscopic behavior, others are irrelevant: a whole class of microscopically distinct rules are, for macroscopic purposes, equivalent. The chances of finding a rule of this class are proportional to the size of the class relative to the larger volume of rule space through which one is conducting the search. Fixing at least some of the essential traits drastically decreases this volume and thus increases the chances of success.

8.2 Neural-Network Programming and Statistical Mechanics

In physics, thermodynamical states are equivalence classes (established according to certain natural criteria) of microscopic states. The dynamical aspects of thermodynamics are but the expression of counting arguments: basically, a system will progress from state A to state B if

1. the size of class B is much larger than that of class A, and
2. there are no constraints to prevent the microscopic state from reaching class B from class A.

Similar considerations apply to the programming of neural networks. The role of microscopic states is played by microscopic (i.e., completely specified) rules (such as the lookup table for a cellular automaton, the list of synaptic weight for a network of neurons, or the circuit diagram for a network of Boolean gates); the role of thermodynamical state is played by the network's macroscopic behavior. During training, a network will progress from behavior A to behavior B under conditions analogous to **1** and **2** above.

Thus, much of the enormous wealth of concepts and techniques of statistical mechanics is directly applicable to neural networks, even though the latter may have nothing to do with physics. Though ordinarily found only in physics textbooks, entropy, energy, and temperature are, at the bottom, combinatorial concepts that belong to everybody. If it was physicists that happened to discover, name, and use them, it is because their discipline confronted them with massive combinatorics problems before anybody else.

9. CONCLUSIONS AND PERSPECTIVES

Much of the expertise developed in dealing with neural networks as tools actually has to do with a better understanding of the tasks to which they were addressed. Much as statistical mechanics rapidly outgrew its thermodynamical origins, this understanding transcends the original breeding ground of "network of neurons" and is transferable to a wider range of problems and tools. Specifically, neural networks retain their relevance as a high-level programming environment even when the target hardware is a conventional digital computer.

Neural networks are not an improved "brand" of hardware or programming style, to be used—on the same tasks—as a drop-in replacement for more traditional brands. They address different tasks, and thus are meant to play a complementary rather than a competitive role. However, the range of tasks for which neural networks are best suited is still poorly known.

The "market share" of neural networks is likely to be very responsive to the nature and the availability of different kinds of "raw computational materials." Just to illustrate a possible scenario, nanoscale device physics may soon offer "quantum dot" arrays (Reed, 1993)—near–atomic-scale devices fabricatable in astronomical numbers but having very primitive computing capabilities and connected only on a local scale. Initially, quantum dots are expected to operate in a quasianalog regime, and to be affected by substantial noise; extracting useful computation from such devices would be a challenge well suited for neural networks. On the other hand, as soon as quantum dots attained a digital, low-noise regime, the door would be open for competition from more traditional, gate-level design approaches.

GLOSSARY

Additive: In the present context, synonymous with **Linear**.

Almost All: Strictly, this means "all, except a subset of measure zero." More loosely, when the set s under consideration is seen as a generic element of a sequence $\{s_i\}$ of sets (e.g., increasingly large neural networks of a given kind), it is convenient to say that a property is true for "almost all" elements of s if the fraction of s_i on which the property is true goes to 1 as $i \to \infty$.

Analog: An analog signal is one designed to convey a choice out of a continuous range (e.g., any voltage between 0 and 5 V). The actual number of distinguishable choices is of course limited by noise and other tolerances. See **Digital**.

Annealing: A procedure by which a system is subjected to a randomizing procedure (e.g., shaking, heating) of controlled intensity in order to remove internal stresses and leave the system in a more stable state.

Attractor: In a dynamical system, a point or a cycle that is asymptotically stable: if the system is on an attractor, after a small perturbation it will tend to return toward it. See **Basin of Attraction**.

Autonomous: A system is autonomous if its evolution is not dependent on the state of variables external to the system itself. Intuitively, "isolated" or "self-contained."

Basin of Attraction: The set of all initial states whose trajectories converge toward a given attractor. See **Attractor**.

Computation Universal: A set of circuit elements is computation universal if, for any computable function, out of these elements one can synthesize a computing network capable of computing that function. The concept of "computable function" is independently defined and reasonably well agreed upon.

Deterministic: A system is deterministic if its state at any future time is uniquely determined by its current state. See **Stochastic**.

Digital: A digital signal is one designed to convey one of two choices, or, more generally, one of a finite, fixed (and typically small) number of choices. See **Analog**.

Dissipation, Dissipative: A mechanism consisting of macroscopic parts that interact according to specified laws is, on a finer scale, made up of an astronomical number of microscopic parts. Perfect decoupling between the nominal, macroscopic "modes of operation" and the microscopic modes is in practice impossible to achieve. Energy injected into the macroscopic modes eventually redistributes itself as thoroughly as possible through the system, without regard for nominal boundaries between macroscopic and macroscopic modes; thus, there is a tendency for energy to flow from macroscopic to microscopic modes until equilibrium is attained. This tendency is called dissipation. An analogous situation arises in any system consisting of a large number of weakly coupled parts. See **Energy Function**.

Emergent: A quantity, property, or behavior that, though not defined in terms of microscopic dynamics, is well characterized at the level of macroscopic statistics. For example, "pressure"—the average collision rate—is meaningless for an individual molecule but well defined for a gas near equilibrium.

Energy: A physical quantity whose main property is that it is conserved in a wide range of circumstances even in systems that are not perfectly decoupled from the rest of the world. This property greatly constrains—and in certain contexts completely determines—the range of admissible dynamical or statistical behavior. See **Entropy, Temperature, Energy Function**.

Energy Function: The amount of energy stored in the mechanical modes of operation, as a function of the system's mechanical state (see **Dissipation**). Knowledge of the energy function allows one to predict the path that the mechanical state will follow on its way to equilibrium (cf. **Energy**).

Entropy: Entropy quantifies by how much a partial description of a system's state fails to be a complete description; basically, it counts (on a logarithmic scale) how many possibilities are consistent with the given description. In a physical system, this number of possibilities depends, among other things, on the ways one can apportion energy, momentum, material particles, etc. between different parts of the system (see **Temperature, Energy**), and leads to important statistical predictions derivable from purely combinatorial arguments. Similar arguments can be used to study the statistics of more abstract systems.

Exponential: A function $f(x)$ is exponential if it grows approximately as a^x, for some fixed a; it is polynomial if it grows approximately as x^b, for some fixed b. For large enough x, an exponential function grows faster than any polynomial function.

Feedback: Feedback occurs whenever the current state of a system variable depends to some extent on the state of the same variable at a previous time. A feedback loop is a path, in the system's structure, through which this dependence make itself felt.

Hamming Distance: Given two sets of binary digits in a one-to-one correspondence (e.g., two strings of the same length, two arrays of the same format), their Hamming distance is simply the number of positions in which they do not agree.

Hidden Unit: A neural-network node not used as an input or an output by the algorithm under consideration. See **Visible Unit, Slack Variable**.

Inversion: Inverting a function f means finding all values x such that $f(x) = y$.

Linear: A system is linear if

1. for any two states a, b the "sum" (in some specified sense) $a + b$ of these states is also a state, and
2. the evolution law f is such that $f(a + b) = f(a) + f(b)$.

Monotonic: A function $y = f(x)$ is monotonic if the increment Δy corresponding to a given increment Δx is always of the same sign (either ≥ 0 or ≤ 0) throughout the range of x.

Network: In the present context, the set of interconnections or pathways through which system variables influence one another.

Polynomial: See **Exponential**.

Reversibility: A system is reversible if its state at any previous time is uniquely determined by its current state. See **Deterministic**.

Slack Variable: Strictly, a variable such that the given problem has solutions for a whole range of values of the variable itself. More loosely, a variable added in the hope that, with this extra degree of freedom, solutions will be easier to find. Adding **hidden units** to a network increases the number of ways a solution compatible with the given inputs can be constructed.

Stochastic: If the evolution in time is not uniquely determined by the current state (see **Deterministic**), but nonetheless definite probabilities can be assigned to the different possibilities, the system is stochastic. Intuitively, "nondeterministic but statistically well-characterized."

Symmetry: A system is symmetric with respect to a given group of transformations if its behavior is unchanged under those transformations. For example, addition is symmetric with respect to permutations of its input terms, while subtraction is not.

Temperature: Measures by how much the amount of disorder in a system can be increased as a new amount of energy becomes available to be "reshuffled" between its parts (see **Energy, Entropy**). Analogs of temperature arise whenever a fixed amount of something has to be distributed between many parts.

Transfer Function: Intuitively, the analog equivalent of a conversion table such as a tax table. The level, range, or scale of a variable x constructed within part of a system may not be suitable for input to another part; a transfer function $y = f(x)$ performs the appropriate conditioning (e.g., level-shifting, range-compression, shaping).

Visible Unit: A neural-network node used as an input or an output by the algorithm under consideration. See **Hidden Unit**.

Works Cited

Baierlein, R. (1971), *Atoms and Information Theory—An Introduction to Statistical Mechanics*, San Francisco: W. H. Freeman.

Barron, A. R. (1991), "Approximation and Estimation Bounds for Artificial Neural Networks," in: L. C. Valiant, M. K. Warmuth (Eds.), *Proceedings of 4th Annual Workshop on Computation Learning Theory (COLT '91)*, San Mateo, CA: Morgan-Kaufman.

Baum, Eric B., Haussler, David (1989), "What Size Net Gives a Valid Generalization?" *Neural Comput.* **1**, 151–160.

Burks, A. (Ed.) (1970), *Essays on Cellular Automata*, Urbana, IL: University of Illinois Press.

Cajaniello, E. R. (1961), "Outline of a Theory of Thought and Thinking Machines," *J. Theor. Biol.* **1**, 204–235.

Carnevali, P., Patarnello, S. (1987), "Exhaustive Thermodynamic Analysis of Boolean Learning Networks," *Europhys. Lett.* **4**, 1199–1204.

Craig, J. J. (1989), *Introduction to Robotics*, Reading, MA: Addison-Wesley.

Frisch, U., Hasslacher, B., Pomeau, Y. (1986), "Lattice-Gas Automata for the Navier-Stokes Equation," *Phys. Rev. Lett.* **56**, 1505–1508.

Fukushima, K. (1975), "Cognitron: A Self-Organizing Multilayered Neural Network," *Biol. Cybernetics* **20**, 121–136.

Hebb, D. O. (1949), *The Organization of Behavior*, New York: Wiley.

Hertz, J., Krogh, A., Palmer, R. G. (1991), *Introduction to the Theory of Neural Computation*, Redwood City, CA: Addison-Wesley.

Hillis, D. (1985), *The Connection Machine*, Cambridge, MA: MIT Press.

Hinton, G. E., Sejnowski, T. J. (1986), "Learning and Relearning in Boltzmann Machines," in: D. E. Rumelhart, J. L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, Cambridge, MA: MIT Press, Chap. 7.

Holland, J. (1975), *Adaptation in Natural and Artificial Systems*, Ann Arbor, MI: University of Michigan Press (republished by MIT Press, 1992).

Hopfield, J. J. (1982), "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," *Proc. Nat. Acad. Sci. USA* **79**, 2554–2558.

Hopfield, J. J., Tank, D. W. (1986), "Computing with Neural Circuits: a Model," *Science* **233**, 625–633.

Katz, A. (1967), *Principles of Statistical Mechanics*, San Francisco: W. H. Freeman.

Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P. (1983), "Optimization by Simulated Annealing," *Science* **220**, 671–680.

Koch, C., Marroquin, J., Yuille, A. (1986), "Analog 'Neuronal' Networks in Early Vision," *Proc. Nat. Acad. Sci. USA* **83**, 4263–4267.

Kröse, B. J. A., van der Smagt P. P. (1993), *An Introduction to Neural Networks*, Amsterdam: The University of Amsterdam.

Marr, D. (1969), "A Theory of Cerebellar Cortex," *J. Physiol.* **202**, 437–470.

McCulloch, W. S., Pitts, W. (1943), "A Logical Calculus of Ideas Immanent in Nervous Activity," *Bull. Math. Biophys.* **5**, 115–133.

Mead, C. (1989), *Analog VLSI and Neural Systems*, Reading, MA: Addison-Wesley.

Minsky, M. (1967), *Computation: Finite and Infinite Machines*, Englewood Cliffs, NJ: Prentice-Hall.

Minsky, M. L., Papert, S. A. (1969), *Perceptrons—An Introduction to Computational Geometry*, Cambridge, MA: MIT Press (republished with additions in 1988).

Psaltis, D., Sideris, A., Yamamura, A. A. (1988), "A Multilayer Neural Network Controller," *IEEE Control Systems* **8** (4), 17–21.

Reed, M. (1993), "Quantum Dots," *Scientific American* **268** (1), 118–123.

Rosenblatt, F. (1962), *Principles of Neurodynamics*, Portage, MI: Spartan.

Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986), "Learning Representations by Back-Propagating Errors," *Nature* **323**, 533–536.

Sanger, T. D. (1989), "Optimal Unsupervised Learning in a Single-Layer Linear Feedforward Neural Network," *Neural Networks* **2**, 459–473.

Toffoli, T., Margolus, N. (1987), *Cellular Automata Machines—A New Environment for Modeling*, Cambridge, MA: MIT Press.

Weisbuch, G. (1991), *Complex Systems Dynamics*, Redwood City, CA: Addison-Wesley.

Wolfram, S. (ed.) (1986), *Theory and Applications of Cellular Automata*, Singapore: World Scientific.

Further Reading

The book by Hertz *et al.* (1991) is one of the best all-around introductions to the theory of neural networks. More descriptive works are Kröse and van der Smagt (1993), for neural networks, and Weisbuch (1991), for complex systems. Articles on neural networks are regularly to be found in a number of journals, including *Complex Systems, Neural Computation, Physical Review Letters, Physical Review A, Neural Networks, Europhysics Letters, Journal of Physics, Applied Optics, Cognitive Science, Biological Cybernetics*, and *Neural Networks*, and occasionally in journals such as *Science* and *Nature*. The bimonthly *Neural Computation* deals with the physiological aspects of neural computation. Another rich source of articles is the proceedings of several IEEE conferences on neural networks—specifically, *Advances in Neural Information Processing Systems*, published annually (since 1988) by Morgan-Kaufman.

Katz (1967) is a sophisticated introduction to statistical mechanics from the information-theory viewpoint; Baierlein (1971), a more leisurely one. A direct road from information theory to inference theory—with thermodynamics as a corollary rather than as a starting point—has been traced by E. T. Jaynes in a series of papers. We recommend "Where Do We Stand on Maximum Entropy?" reprinted in his collection, Jaynes, E. T. (1983), *Papers on Probability, Statistics, and Statistical Mechanics*, Dordrecht: Reidel, pp. 210–315, as well as "How Does the Brain Do Plausible Reasoning?" in: G. J. Erickson and C. R. Smith (Eds.) (1988), *Maximum-Entropy and Bayesian Methods in Science and Engineering*, Vol. 1, Boston: Kluwer Academic, pp. 1–24.

16
Neurobiophysics

Oleg V. Kolomytkin and Andrew A. Marino

16.1 **Introduction 525**
16.2 **Ion Channels 526**
16.2.1 Overview 526
16.2.2 Na^+ Channels 528
16.2.3 K^+ Channels 529
16.2.4 Ca^{2+} Channels 529
16.2.5 Cl^- Channels 530
16.2.6 Ligand-Gated Channels 530
16.3 **Biophysics of Neurons 531**
16.3.1 Resting Membrane Potential 531
16.3.2 Chemical-to-Electrical Transduction 533
16.3.3 Signal Summation 534
16.3.4 Action Potential 537
16.3.5 Electrical-to-Chemical Transduction 539
16.4 **Neuronal Systems 541**
16.4.1 Overview 541
16.4.2 Sensory Systems 541
16.4.3 Neural Networks 544
16.4.4 Volume Transmission 545
16.4.5 Brain Electrical Activity 547
16.4.6 Learning and Memory 548
16.4.7 A Nonlinear Future 551
Glossary 552
References 554
Further Reading 555

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

16.1
Introduction

Neurobiophysics is the study of the structure and function of the nervous system from the perspective of physics. In 1902, Julius Bernstein hypothesized that cells were ionic solutions surrounded by thin membranes having permeability properties that resulted in the establishment of an electrical potential across the membrane, and that during nervous system activity, the potential decreased as a result of changes in the permeability. These ideas led to the work of Alan Hodgkin and Andrew Huxley, which is presently recognized as the prototypical physical explanation of electrical conduction in nerve cells. Attempts to understand the nervous system were also made from other scientific perspectives including neuroanatomy, neurophysiology, neurochemistry, and molecular biology. The distinctions between the various areas of neuroscience are often difficult to discern, but for our purposes studies that emphasize the role of physical laws, elucidation of physical mechanisms, and the use of mathematical analysis are within neurobiophysics and will be considered here.

The basic structural and functional unit of the nervous system is the *neuron*, which is a cell specialized to receive information and influence other neurons or effector cells (Figure 16.1). The complexity of the nervous system varies with the degree of evolutionary development of the organism; in mammals, the system is composed of more than 10^{11} neurons and perhaps 10 times that number of support cells (the neuroglia). There are many different kinds of neurons, but they can be classified into a small number of groups on the basis of common properties.

Information enters a neuron through an elaborate network of processes called *dendrites* and exits through a single process, the *axon*. Neurons communicate with one another by means of specialized contacts called *synapses*. A typical neuron simultaneously receives numerous signals at synapses on the dendrites or cell body. Each signal produces a decrease (excitatory) or increase (inhibitory) in the local *resting membrane potential* that propagates passively to the cell's trigger zone, usually the axon *hillock*. The contribution of a particular input decreases exponentially with increasing distance between its location and the hillock, and the cell responds to the instantaneous sum of the individual contributions. If the net change in the *membrane potential* at the

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

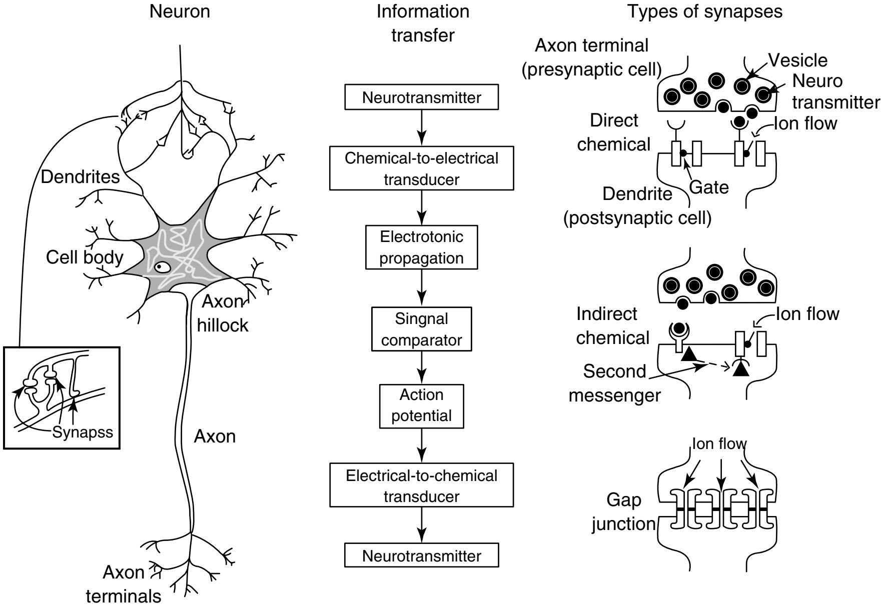
Neuron
Dendrites
Cell body
Axon hillock
Synapss
Axon
Axon terminals
Information transfer
Neurotransmitter
Chemical-to-electrical transducer
Electrotonic propagation
Singnal comparator
Action potential
Electrical-to-chemical transducer
Neurotransmitter
Types of synapses
Axon terminal (presynaptic cell)
Vesicle
Neuro transmitter
Direct chemical
Ion flow
Gate
Dendrite (postsynaptic cell)
Indirect chemical
Ion flow
Second messenger
Ion flow
Gap junction

Fig. 16.1 Information transfer in the neuron. The presynaptic and postsynaptic neurons are defined by the direction of information transfer between them. In a chemical synapse, the cells are separated by a narrow gap across which a neurotransmitter diffuses and binds to receptors on the postsynaptic cell. The binding triggers the opening of membrane channels, allowing ions to pass from the interstitial fluid into the postsynaptic cell. In the direct chemical synapse, the receptor, membrane pore, and gate consist of a unitary protein complex. Alternatively, the receptor and channel may be physically separated but linked by intermediary second messengers (indirect chemical synapse). In gap junction synapses, ions pass between the cytoplasm of adjacent cells; gap junction channel conductance is relatively large (100–200 pS) compared with that of the channels at chemical synapses. Typically, directly gated channels mediate neuronal activity and indirectly gated channels modulate the excitability of neurons.

hillock induced by the summed inputs is a depolarization that exceeds a threshold value, an *action potential* (spike pulse) is generated and propagates along the axon. When the action potential arrives at the axon terminal, voltage-sensitive Ca^{2+} *ion channels* are activated thereby triggering release of *neurotransmitters* into the synaptic cleft.

Prior to discussing the action potential, we will describe the structure and function of ion channels. Thereafter, we will consider the dynamical properties of systems of neurons.

16.2
Ion Channels

16.2.1
Overview

The neuronal membrane is an electrically insulating phospholipid bilayer about 5 nm thick; the electrical activity of the neuron arises from the flow of ions through ion channels embedded in the membrane. A gated ion channel is a functional unit containing a pore through which ions may pass, a gate, and a sensor capable of opening or closing the gate in response to a signal. There are two classes of gated ion channels, depending on the nature of the signal to which they are responsive. *Voltage-gated* channels have an ion conductance that depends on the cell's membrane potential, which provides the driving force for the ions and also affects the probability that the channel is open. The conductance of *ligand-gated channels* is directly or indirectly dependent on the binding of a neurotransmitter to the sensor portion of an ion channel or to a membrane receptor protein, respectively (Figure 16.1). Most gated channels exhibit only one of the two forms of gating behavior. Nongated channels are essentially membrane pores lacking gates and sensors.

The *selectivity* of a channel refers to the ion species that will pass through its pore. Voltage-gated channels are denoted by the ion that passes through most readily; the main types are Na^+, K^+, Ca^{2+}, Cl^-, and an ion channel that allows all small cations to pass (called a nonselective cation channel). Ligand-gated channels are labeled by the ligand that opens the channel. As examples, a Na^+ channel has Na^+ as the main permeant ion and a nicotinic acetylcholine channel is a transmembrane protein having a receptor capable of binding the neurotransmitter acetylcholine, which results in the passage of ions through the channel pore. The response of a neuron to its environment is mainly determined by the gating and selection characteristics of its ion channels, and by the density and distribution of each channel type in the neuronal membrane. Neurons contain: (i) nongated channels that serve to establish the membrane potential; (ii) ligand-gated channels that subserve reception of the input signal; (iii) voltage-gated Na^+ and K^+ channels that function in a synchronized fashion to permit propagation of an action potential; (iv) voltage-gated Ca^{2+} channels that participate in the transduction of the action potential into the chemical signal that constitutes the neuron's output; (v) *gap junction* channels; and (vi) channels that are sensitive to various intracellular chemical signals.

Studies involving *site-directed mutagenesis* and X-ray crystallography have revealed some of the basic structural features of ion channels, for example the location of the ion-selective filter for voltage-gated Na^+, Ca^{2+}, and K^+ channels (Gouaux and Mackinnon, 2005). The filter consists of negatively charged amino-acid residues and lies in the channel's pore-lining region (P segments). By changing specific amino acids in the P segments, the permeant-ion profile of Na^+ and Ca^{2+} channels can be interchanged. The primary structures of K^+ channels are more heterogeneous, but location of the selectivity filter has been narrowed down to a three-amino-acid motif in the P segment.

All known voltage-gated channels sense voltage changes by means of a protein subunit called the S4 segment, which was identified on the basis of measurements of the *gating current*. A span of at least seven S4 amino acids is exposed extracellularly during activation of K^+ channels, suggesting that either the S4 amino acids moved extracellularly or the rest of the channel protein moved intracellularly. A typical gating current consists of the movement of 10 positive charges through a distance up to 0.3 nm.

Gap junctions are ion channels that connect the cytoplasm of adjacent cells, thereby allowing passage of ions and signaling molecules smaller than ≈1 kDa. Transmembrane protein subunits called *connexins* form a pore that constitutes one half of the gap junction; the other half is formed by a similar structure on an adjacent cell. Heterology among the connexin subunits allows differences in permeability, conductance, and gating of gap junctions by voltage or various ligands (Hervé *et al.*, 2007).

16.2.2
Na^+ Channels

Na^+ channels respond rapidly to depolarization; their basic role in the nervous system is to generate the initial portion of the action potential. Nine members of the voltage-gated sodium-channel family are known (Benarroch, 2007); their amino-acid identity is greater than 50%.

A Na^+ channel consists of one α and $1-4\beta$ protein subunits (Figure 16.2); the α subunit determines the channel's selective and gating properties that are modulated by the β subunits. An α subunit has four homologous domains (I–IV) that each contains six hydrophobic amino-acid sequences (S1–S6) arranged to form a pore selective for the passage of Na^+. The channel's voltage sensitivity is located in S4, a highly conserved amino-acid sequence composed of positively charged amino acids. During depolarization, the S4 region moves toward the outer side of the membrane thereby allowing the channel to conduct ions. The selectivity filter, the narrowest part of the pore (0.3×0.5 nm), is formed by the extracellular region between S5 and S6 (P-loops) of all four domains; the filter permits single Na^+ and associated water molecules to pass, but excludes the larger K^+, negatively charged glutamic acid residues, and Cl^-.

Three states of voltage-gated Na^+ channels have been identified: deactivated, activated, and inactivated; the channel is open only in the activated state. Transitions between states can be described as a Markovian process or by the Hodgkin–Huxley formalism (see below).

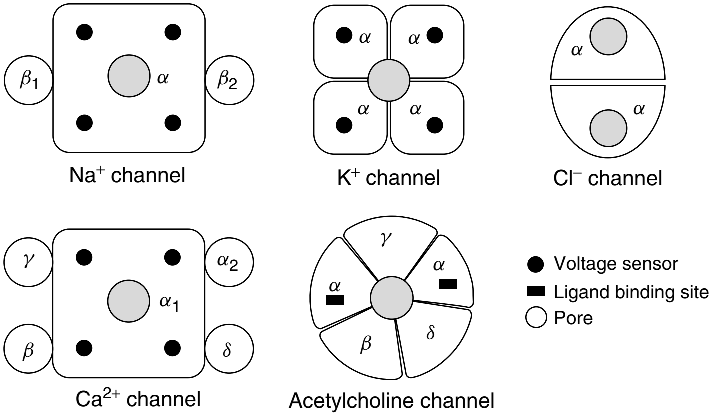

Fig. 16.2 Subunit composition of ion channels (depicted from above the plasma membrane).

16.2.3
K^+ Channels

The K^+ family of voltage-gated ion channels is the largest and most diverse group of voltage-gated ion channels (Doyle *et al.*, 1998). K^+ channels subserve many functions, including return of the membrane potential to its preexisting level, formation of trains of action potentials, and the occurrence of rhythmic activity. There are at least four major classes of voltage-gated K^+ channels in the nervous system, and perhaps 10 times that number of K^+ channel subtypes. The main class is the *delayed rectifier*, which is the axonal K^+ channel that opens with depolarization and is largely responsible for repolarizing the axon membrane following an action potential. Other important types of K^+ channels include (i) an axonal channel that opens rapidly upon depolarization and then quickly closes; (ii) an inward-rectifying channel that opens only with hyperpolarization; and (iii) a channel that is sensitive to intracellular Ca^{2+}. In the case of the Ca^{2+}-sensitive K^+ channel, the Ca^{2+} concentration shifts the channel voltage dependence (Hille, 1992). There is only about a 10% variation in the total amino-acid sequence among the various K^+ channels.

The molecular mass of the K^+ channel protein is about 25% of that of the Na^+ channel protein, and consists of six transmembrane segments. By analogy with the known structure of the Na^+ channel, a K^+ channel is thought to be a tetramer with a central pore (Figure 16.2). Five residues from each subunit form the selectivity filter; their negatively charged carbonyl oxygen atoms are directed toward the pore and interact with the K^+, effectively replacing its hydration shell; the Na^+ does not penetrate the filter because the ion's interaction with its hydration shell is greater than with the carbonyl oxygen atoms in the pore.

The conductance of some K^+ channels appears to exceed the value associated with movement of K^+ in aqueous solutions. One possibility is that the channel can accommodate more than one ion at a time. If so, the presence of the first ion might reduce the electrostatic forces on the second ion, thereby facilitating its passage through the pore.

Axons from mollusks, arthropods, annelids, and vertebrates contain essentially identical Na^+ channels and K^+ channels, indicating that the evolution of these channels was essentially complete at the time of the common ancestor of these phyla, approximately 500 million years ago.

16.2.4
Ca^{2+} Channels

There are many types of voltage-gated Ca^{2+} channels in the nervous system, including types N, L, P, Q, R, and T. They all open with depolarization and appear to have a common subunit composition (Figure 16.2), but differ in voltage dependence, ionic selectivity, and pharmacology. The α subunit forms the Ca^{2+}-selective pore and contains the voltage-sensing machinery or ligand-binding sites. The subunit consists of four homologous domains (I–IV), each containing six transmembrane α-helices; the α_1 subunit (190 kDa) determines most of the channel's properties; there are at least four variations of α_1. The α_2, β, γ, and δ subunits modulate the properties of the Ca^{2+} channel.

Neurotransmitter secretion at nerve terminals is a well-studied Ca^{2+}-dependent process. N-type voltage-gated Ca^{2+} channels in the presynaptic terminal open in response to the depolarization produced by the arrival of the action potential, thereby triggering membrane-bound vesicles containing neurotransmitter to fuse with the axon membrane, resulting in release of the neurotransmitter into the synapse. The Ca^{2+} channels in the axon terminal provide the only mechanism whereby the action potential can be transduced into a form capable of carrying information across a chemical synapse.

16.2.5
Cl^- Channels

The superfamily of Cl^- channels consists of approximately 13 members; factors that can regulate Cl^- channel activity include voltage, Ca^{2+} concentration, extracellular ligands, and pH (Suzuki, Morita and Iwamoto, 2006). Chloride channels are important for maintaining resting membrane potential and normal cell volume; they conduct Cl^- as well as other small anions including HCO_3^-, I^-, SCN^-, and NO_3^-. Chloride channel subunits are between 1 and 12 transmembrane segments.

16.2.6
Ligand-Gated Channels

Tab. 16.1 Neurotransmitters organized by chemical properties.

Monoamines and Acetylcholine
Acetylcholine
Dopamine
Epinephrine
Histamine
Norepinephrine
Serotonin
Amino Acids
Aspartate
Gamma aminobutyric acid
Glutamate
Glycine
Serine
Purines
Adenosine
Adenosine triphosphate
Gases
Carbon monoxide
Nitric oxide
Lipids
2-arachidonoylglycerol
Anandamide
Peptides
Beta-endorphin
Bombesin
Bradykinin
Corticotrophin releasing hormone
Galanin
Neuropeptide Y
Neurotensin
Orexin
Oxytocin
Peptide YY
Somatostatin
Substance P
Vasoactive intestinal polypeptide
Vasopressin

Ligand-gated ion channels are specialized for converting neurotransmitters (Table 16.1) into graded electrical signals; the similarity in structure of ligand-gated channels suggests that they are a protein superfamily. The channels open transiently following binding of a neurotransmitter, thereby producing a postsynaptic potential as a consequence of the ion flux; ligand-gated channels are usually voltage insensitive. In one type of ligand-gated channels, the receptor, membrane pore, and gate consist of a unitary protein structure (*direct chemical synapse*); in another type (*indirect chemical synapse*) the receptor and pore portions of the channel are physically separated but linked by intermediary substances known as second (or higher-order) messengers (Figure 16.1). Ligand-gated channels have one or more binding sites for a particular neurotransmitter or second messenger, and a characteristic ion selectivity.

The most studied ligand-gated channel is the nicotinic acetylcholine receptor (nACr) channel (Figure 16.2), which is found at the neuromuscular junction and at other locations in the nervous system; the nACr serves as a prototype for the less studied ligand-gated channels. The receptor molecule is ~9 nm in diameter, and protrudes from the membrane surfaces ~6 nm into the extracellular space and ~2 nm into the cytosol. When two acetylcholine molecules bind to the receptor, a conformational change occurs that opens an aqueous pore 2–3 mm in diameter for about a millisecond; thereafter the molecules disassociate from the receptor and are hydrolyzed by acetylcholinesterase. The nACr excludes anions, possibly because of the negatively charged amino acids at its mouth. The channel is formed from five subunits (stoichiometry shown in Figure 16.2); the α subunits bind acetylcholine with high affinity, one molecule of which must bind to each α subunit for the channel to open efficiently. Site-directed mutagenesis of the cDNAs of the α subunits has shown that the binding sites for acetylcholine are located near two cysteine residues on the extracellular portion of the subunit. The four subunits are encoded by different but homologous genes; each subunit appears to consist of four membrane-spanning regions arranged in such a way that specific regions of each of the subunits face each other to create the membrane-spanning pore.

16.3
Biophysics of Neurons

16.3.1
Resting Membrane Potential

The membrane potential of the neuron E_m is established by the same mechanism as in other cells (Figure 16.3). The Na^+/K^+ pump, a transmembrane energy-consuming enzyme, moves three Na^+ out of the cell and two K^+ in for each molecule of adenosine triphosphate converted to adenosine diphosphate. Nongated ion channels permit passive transmembrane flow of ions down their electrochemical potential gradients, principally Na^+ and K^+.

E_m is given by the Goldman–Hodgkin–Katz (GHK) equation (Hille, 1992):

$$E_m = \frac{RT}{F} \ln \frac{P_{\mathrm{Na}}[\mathrm{Na}^+]_0 + P_{\mathrm{K}}[\mathrm{K}^+]_0 + P_{\mathrm{Cl}}[\mathrm{Cl}^-]_i}{P_{\mathrm{Na}}[\mathrm{Na}^+]_i + P_{\mathrm{K}}[\mathrm{K}^+]_i + P_{\mathrm{Cl}}[\mathrm{Cl}^-]_0} \tag{1}$$

where R is the universal gas constant, T is the absolute temperature, F is Faraday's constant, the internal and external concentrations of Na^+ are, respectively, represented by $[Na^+]_i$ and $[Na^+]_0$, P_{Na} is the membrane permeability for Na^+ ions, and the concentrations and permeabilities of K^+ and Cl^- are defined similarly. If the permeability of any two of the ions is zero, the GHK equation reduces to the Nernst equation for the third ion and the resulting potential is the *reversal potential* for that ion. In the resting state, the cell membrane is more permeable to K^+ than to Na^+ and Cl^-; consequently the resting potential is close to the K^+ reversal

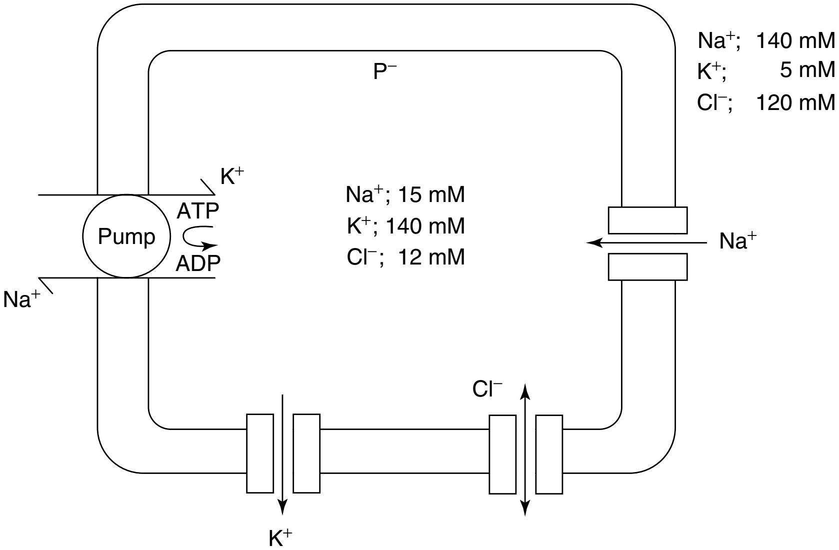
P⁻
Na⁺; 140 mM
K⁺; 5 mM
Cl⁻; 120 mM
K⁺
ATP
Pump
ADP
Na⁺
Na⁺; 15 mM
K⁺; 140 mM
Cl⁻; 12 mM
Na⁺
Cl⁻
K⁺

Fig. 16.3 The pump-leak model for the origin of the cell resting membrane potential. The interaction among the pump and leak channels results in a negative membrane potential (cell interior with respect to exterior). P^-, negatively charged impermeant proteins. Typical intracellular and extracellular concentrations of several important ions are shown.

potential, typically −65 mV (cell interior negative).

An equivalent circuit model of the neuronal membrane (Figure 16.4) permits consideration of other conditions, such as the response of a cell to an applied voltage or current. In these cases, the membrane potential departs from the resting potential and the equivalent circuit can be used to analyze the resulting time-dependent and steady-state current changes.

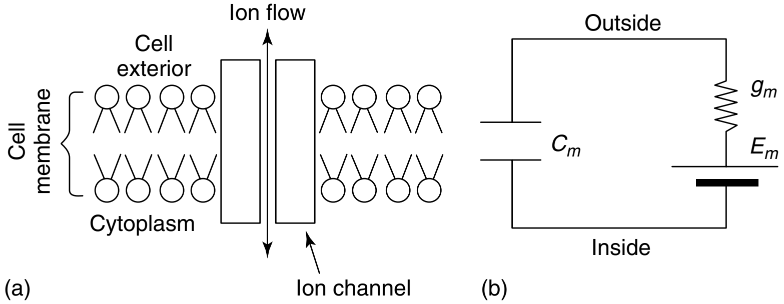
Ion flow
Cell exterior
Cell membrane
Cytoplasm
Ion channel
(a)
Outside
g_m
C_m
E_m
Inside
(b)

Fig. 16.4 A cell membrane and its equivalent circuit. (a) Each ion-conducting pathway in the membrane contributes to the membrane potential. (b) Lumped presentation of the permeability pathways. C_m, g_m, membrane capacitance and conductance, respectively. E_m, membrane potential.

16.3.2
Chemical-to-Electrical Transduction

Synapses are structures that facilitate transmission of signals between neurons. Two general types of synapses are recognized, depending on the source of the ions that enter the postsynaptic cell (Figure 16.1). In a chemical synapse (the characteristic linkage between neurons in the mammalian nervous system), there is a 20–30 nm gap between membranes of two cells; the diameter of the gap is 1–2 μ m. A neurotransmitter (Table 16.1) synthesized and secreted by the presynaptic cell diffuses across the gap and reversibly binds to *receptors* on the postsynaptic cell. As a result, membrane channels in the postsynaptic cell open or close, thereby altering ion flow between the interstitial fluid and the neuron interior. Typical binding constants are in the range of 10^5–10^7 M^{-1}. In some cases, the neurotransmitter receptor and the pore through which the ions pass are part of a unitary transmembrane protein complex; in other cases they are distinct proteins and the events at each site are coupled by intracellular second-messengers (Figure 16.1).

Direct gating of ion channels involves a change in the conformation of only a single macromolecule, and can therefore occur on the order of milliseconds. Channels activated by second messengers are slower because they involve a series of reactions. In both cases, the ion flow produces a change in the membrane potential of the postsynaptic cell in the vicinity of the entry point of the ions into the cell. If Cl^- enters, the resting membrane potential (typically about −65 mV) becomes more negative, resulting in a hyperpolarization; entry of cations produces a depolarization. Hyperpolarization of the neuronal membrane inhibits neuronal activity (Eqs (1) and (3)), and depolarization produces the opposite effect. The neurotransmitter is removed from the synaptic gap by a pump in the presynaptic membrane or by means of enzymatic degradation.

In a gap junction synapse (Figure 16.1), the presynaptic and postsynaptic cells are linked by conducting channels that permit ionic flow between them. Gap junctions (also called *electrical synapses*) are relatively rare in mammalian nervous systems compared with those of lower vertebrates and invertebrates.

Postsynaptic membranes provide chemical-to-electrical transduction. The nACr in the postsynaptic membrane of a neuromuscular synapse is a representative transducing element (neurotransmitter-gated ion channel). The receptor shows little selectivity among cations, and consequently the relative contributions to the channel current are determined by the cationic driving forces. For a typical neuron, the Na^+ concentration is far from Nernst equilibrium (unlike K^+ concentration). Consequently both the Na^+ gradient and the membrane voltage act to drive Na^+ into the cell, rendering it the principal nACr current. The net result of the acetylcholine-induced cationic conductance is the production of an electrical depolarization in the vicinity of the membrane containing the nACr (positive shift from the resting value). The total transmembrane current at the synapse is the sum of the ion flow through several hundred thousand such transmitter-gated channels, each of which has an identical conductance but an open-time governed by stochastic processes.

The biophysical principles governing synapses at the neuromuscular junction also apply to synapses in the central nervous system (CNS). However, signal transduction in the CNS is complicated by several factors.

1. A typical CNS neuron receives many simultaneous excitatory and inhibitory inputs (synapses involving vertebrate skeletal muscle are always excitatory).
2. Many different neurotransmitters may be involved in signal transduction by one neuron, and a given neurotransmitter may have more than one kind of receptor at the cell membrane.
3. The actual role of the neuron in the signaling pathway is determined by the sum of its excitatory and inhibitory inputs, and not merely by the occurrence of the inputs (at the neuromuscular junction each synaptic potential tends to produce an action potential).
4. The indirect mechanism (Figure 16.1) for the effect of neurotransmitters on membrane potential can result in (i) channels that open or close at the resting potential; (ii) transient changes in membrane voltage that last much longer than those caused by directly gated channels; and (iii) second messengers that cause effects in addition to those on channel conductance (alterations in receptors for other neurotransmitters and in gene expression, for example).

16.3.3
Signal Summation

A transient neurotransmitter-induced voltage change at a synapse has no

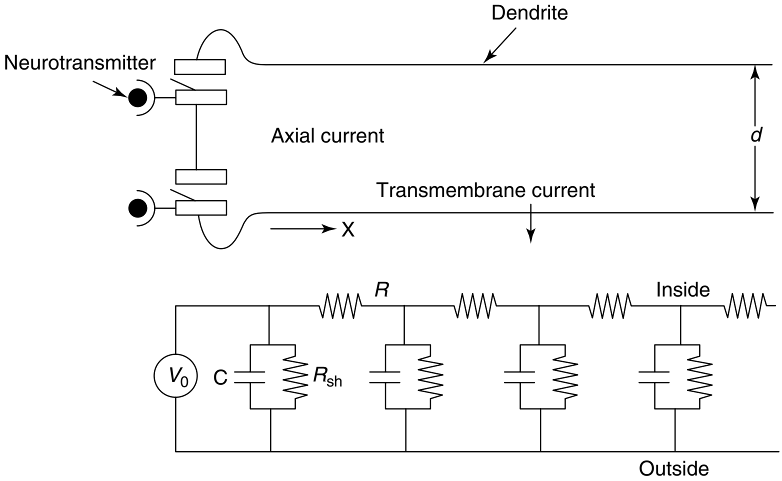

Fig. 16.5 Electrotonic spread of voltage in a dendrite and the corresponding equivalent circuit. (a) Binding of neurotransmitter molecules initiates a change in the membrane potential (V_0) induced by ion flow through neurotransmitter-gated channels in the dendrite. (b) V_0 propagates through the dendrite, which is represented as a series of segments with capacitance C and resistance R_{sh}, connected by resistance R.

individual significance with regard to the information transmitted by the postsynaptic neuron; physiological significance resides in the sum of the transients that propagate to the axon hillock. Signal propagation can be modeled in terms of the response of a lossy insulated wire embedded in a conducting medium (Figure 16.5) (Deutsch and Micheli-Tzanakou, 1987). The dendrite is conceptually divided into a series of isopotential segments of length ∂x and diameter d represented by a membrane capacitance C in parallel with a transmembrane resistance R_{sh}; each adjacent pair of segments is connected by the axonal resistance R, where $R = 4\rho\partial x/\pi d^2$ and ρ is the resistivity of the cytoplasm. A voltage transient occurs at a synapse on the distal portion of the dendrite as a result of an ion flux, and a portion of the resulting current charges the local membrane capacitance thereby increasing or decreasing the membrane potential (depending on the charge of the permeant ion). The remaining current splits and either charges the capacitance of the second segment, passes through the membrane and completes the circuit back to its source, or continues on to the next segment. This process continues until all the current has leaked out and returned to its source via the extracellular fluid, which is assumed to have negligible resistance. From Ohm's and Kirchoff's laws, the equation governing the spread of the potential $V(x, t)$, measured from the resting potential, is

$$\lambda^2 \frac{\partial^2 V}{\partial x^2} = \tau_m \frac{\partial V}{\partial t} + V \tag{2}$$

where $\lambda = \sqrt{R_{\mathrm{sh}}/R}$ is the length constant and $\tau_m = R_{\mathrm{sh}}C$ is the time constant. In terms of C_m and R_m, which are the capacitance and resistance per unit area, respectively, we have $\lambda = \sqrt{R_m d/4\rho}$ and $\tau_m = R_m C_m$; typically, C_m is assumed to be about $1\,\mu\mathrm{F\ cm}^{-2}$. Both τ_m and λ depend on the type of neuron; values for the hippocampal pyramidal cell are $\tau_m = 15\text{–}70\ \mathrm{ms}$ and $\lambda = (0.5 - 1.5)l_0$, where l_0 is the length from the tip of the dendrite to the cell body.

The general solution for $V(x, t)$ is known (Rall, 1960), but the behavior of the potential can be inferred from a consideration of the simplest special cases. The steady-state solution for an infinitely long dendrite (applicable to slow synaptic potentials and to background depolarizations, as in cells in the retina) is $V = V_0 e^{-x/\lambda}$. Since $\rho \approx 1\ \Omega m$ in all neurons, the spatial characteristics of the potential are determined by the membrane resistance and the dendrite diameter. The spread of the potential in finite dendritic systems depends on the nature and extent of the branching that occurs. If the diameters of the dendrites at a branching point are such that $d_0^{3/2} = \sum_i d_i^{3/2}$, where d_0 is the diameter of the parent dendrite and d_i are the diameters of each of the daughter branches, then the branches are electrically equivalent to the stem, and the potential spreads through the entire system as it would in an infinite dendrite (Rall, 1960). This model is useful for calculating the passive spread of potential in dendritic systems (*electrotonus*), and can be used to estimate λ for the dendritic tree.

The membrane time constant is an important determinant of the time course of the postsynaptic potential, but a precise description depends upon many factors including the model chosen. The essential feature of all realistic models is the prediction of an amplitude diminution and pulse-width increase as the postsynaptic potential propagates from its origin. Thus synapses near the cell body facilitate relatively large and rapid responses, whereas distant inputs lead to weaker and slower changes in the membrane potential at the cell hillock (site of generation of the action potential).

Neurons have evolved mechanisms by which they can receive synapses at their distal dendrites and still transmit large, rapid postsynaptic potentials to the axon hillock. These mechanisms include

1. the presence of a high specific membrane resistance;
2. production of a particularly large postsynaptic potential; and
3. the occurrence of active membrane processes in dendrites, for example voltage-gated ion channels responsible for the action potential (see below).

The net effect of simultaneously activated synapses depends on their location in the dendritic tree. When neighboring synapses are activated simultaneously, the conductance changes interact nonlinearly, thereby precluding a general analysis of the responses. At widely separated synapses, the overall response may be described in terms of a superposition of the individual postsynaptic potentials.

Dendritic spines (Figure 16.6) are narrow projections from the dendrites; they are found on many types of neurons and can be the locus of synaptic inputs. The structural and electrophysiological characteristics of spines are not well understood; they may exhibit properties and functions not occurring elsewhere in the dendritic system. For example, the dendritic spine might provide a high-resistance path for a postsynaptic potential into the dendritic tree; this would have the effect of electrically isolating its postsynaptic potential from those induced at synapses on the dendrite itself. Consequently, the dendrite–synapse postsynaptic potential and the spine–synapse postsynaptic potential would add in a more nearly linear fashion than if both potentials occurred side by side directly on the dendrite.

Although the basic response of each portion of the dendritic membrane to a voltage change is a propagating, diminishing voltage transient, the overall response of the neuron is difficult to characterize because of its highly complex morphology. A neuron (Figure 16.6) may contain 30 000 synapses, any combination of which may simultaneously transmit either inhibitory or excitatory postsynaptic potentials, all of which are processed in parallel. Specialized computer programs have been developed to accommodate many degrees of freedom needed to apply the cable equation to realistic models of actual neurons (McKenna, Davis and Zornetzer, 1992).

The role of dendrites was traditionally thought to be limited to that of supporting the passive unidirectional propagation of current, but new evidence suggests that dendrites can also integrate and process bidirectional information. Back-propagation of axon-derived action potentials (*antidromic action potentials*) has been demonstrated in the hippocampus (Magee and Johnston, 1997) and neocortex (Markram *et al.*, 1997), and shown to be necessary for long-term potentiation. Subthreshold synaptic inputs that were followed rapidly by the arrival of back-propagating action potentials facilitated amplification of subsequent subthreshold synaptic inputs. Variations in

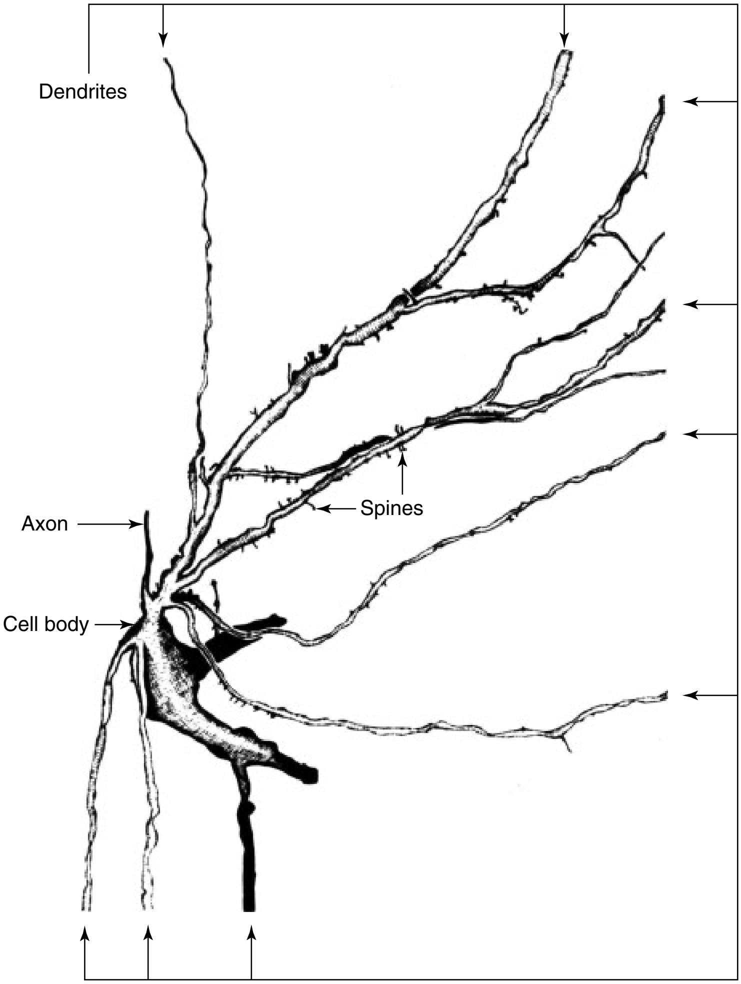

Fig. 16.6 Drawing of a sensory neuron from the dorsal horn of the spinal cord of a Macaque monkey. Only a portion of the dendritic tree and axon is shown. A typical sensory neuron receives about 20 000 synapses. The shading indicates relative depth. Courtesy of John A. Beal.

the timing of the signals altered the *plasticity* of the synapses (Sejnowski, 1997), which has implications for the process of memory formation (see below). Thus, it appears that dendrites may have an important and previously unrecognized role in synaptic plasticity.

If the voltage sum from all synaptic inputs exceeds a threshold (typically $\approx$30 mV) at the axon hillock, then an action potential is generated and propagates along the axon.

16.3.4
Action Potential

An action potential is a local, transient change in membrane potential that propagates at 0.5–110 m s^{-1} (depending on the type of neuron). The biophysical process underlying development of the action potential was elucidated by Hodgkin and Huxley in a classic series of studies that employed the *voltage-clamp* technique (Figure 16.7) (Hodgkin and Huxley, 1952). The voltage-clamp (applied voltage, V_m) was particularly useful because it permitted direct control over the fundamental variable (voltage dependence of membrane conductance).

The current across the membrane was described in terms of a capacitive and three ionic components (Figure 16.8):

$$I_m = C\frac{dV_m}{dt} + g_{\mathrm{Na}}(V_m - E_{\mathrm{Na}}) + g_{\mathrm{K}} \times (V_m - E_K) + g_{\mathrm{L}}(V_m - E_{\mathrm{L}}) \quad (3)$$

where C is membrane capacitance and E_{Na}, E_{K}, and E_{L} are the reversal potentials for the Na^+, K^+, and leak ions, respectively. The electrical excitability of the membrane is contained in the voltage- and time-dependent conductances, g_{K} and g_{Na}; g_{L} is a leakage conductance (undetermined ionic composition). By varying V_m and the concentrations of the ions and by using radioactive K^+,

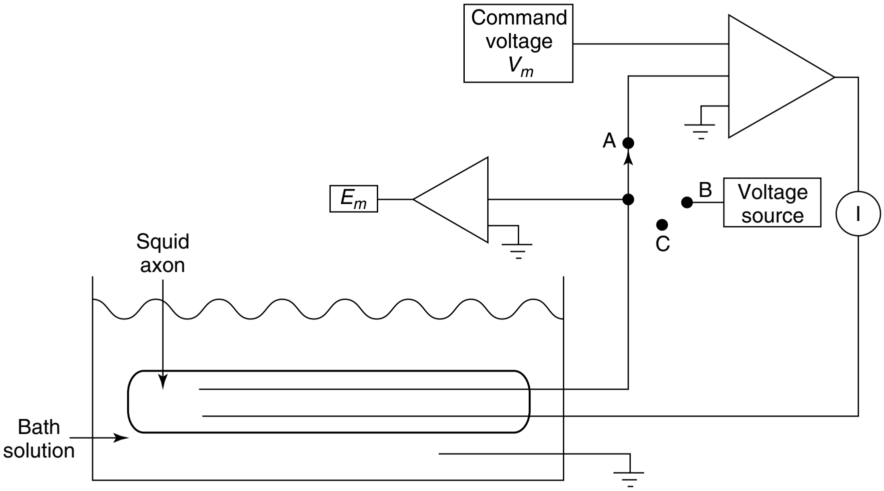

Fig. 16.7 The voltage-clamp technique and measurement of the action potential in the squid axon. For voltage-clamp measurements (switch position A), a voltage-recording electrode and a current-delivering electrode are placed intracellularly. The output of the feedback amplifier is determined by the difference between the command voltage (V_m) and the membrane potential (E_m). For nonzero differences, the amplifier drives a current through the membrane in such a direction as to reduce the difference to zero. When a voltage step is applied, the membrane capacitance becomes charged in a time in the order of microseconds; thereafter, the capacitive current does not contribute to the membrane current. The circuit permits an abrupt displacement of the membrane potential that can be maintained indefinitely at the new value as the membrane current is measured. The design of the electrodes prevents the flow of longitudinal currents. In switch position B, the voltage-clamp is removed, and a brief (100 μ s) voltage pulse is applied between the voltage measuring and bath-solution electrodes; if the switch is then moved to the open-circuit position C, the axon develops an action potential or returns to baseline, depending on whether the depolarizing pulse reached the threshold level.

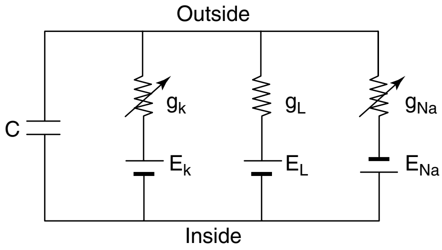

Fig. 16.8 Pathways for current flow in the Hodgkin–Huxley model of the axon. In voltage-clamp studies, capacitive current flows only during the time the potential is changing; thereafter, the membrane current consists solely of an ion flow through the three ion-conductive branches represented by Na^+, K^+, and leak conductances. The batteries depict the Nernst equilibrium potentials; time- and voltage-dependent conductances are employed to explain the response of the membrane to changes in voltage.

Hodgkin and Huxley showed that the early inward current in the voltage-clamped squid axon was due to Na^+ entering the axon, and that the later-appearing outward current was due to K^+ leaving the axon (Figure 16.9). To provide a basis for reconstructing the action potential, they measured the time variations of g_{Na} and g_K for various values of the membrane potential using an ion-substitution method (pharmacological methods are now used for dissecting the Na^+ and K^+ currents). The mathematical description that they formulated to explain the transmembrane current responsible for the action potential consisted of three empirical equations

$$\begin{aligned} g_n(V_m, t) &= \bar{g}_n \Phi^\alpha X^\beta \\ \dot{\Phi}(V_m, t) &= \frac{1}{\tau_\Phi}(\Phi_\infty - \Phi) \\ \dot{X}(V_m, t) &= \frac{1}{\tau_\chi}(X_\infty - X) \end{aligned} \tag{4}$$

where n is Na^+ or K^+; Φ and X are gating variables for activation and inactivation, respectively, representing the fraction of the maximum conductance at any given time and voltage; $\overline{g_n}$ is the maximal value of the conductance, and α and β are constants. The time constants (τ_Φ, τ_χ) and the steady-state values for activation and inactivation (Φ_∞, X_∞) are functions of ion type (Na^+, K^+) and V_m, and are adjusted empirically. For Na^+ channels $\alpha = 3$, $\beta = 1$, for K^+ channels $\alpha = 4, \beta = 0$.

The Hodgkin–Huxley equations (Eq. (4)) and the cable equation (Eq. (2)) are sufficient to explain propagation of the action potential, the existence of the voltage threshold, and the velocity of propagation. Despite some limitations, the model is the generally accepted explanation of the origin and characteristics of the action potential in the peripheral and central neurons of invertebrates and vertebrates. The complex pattern of electrical activity found in the nervous system arises from the interplay of the Hodgkin–Huxley mechanism, neuronal structural complexity, and ion-channel diversity.

When the action potential arrives at the axonal terminal, it is transduced into a chemical signal.

16.3.5
Electrical-to-Chemical Transduction

Neurotransmitter molecules are stored in the axon terminal in specialized organelles called synaptic vesicles. When the action potential arrives at the

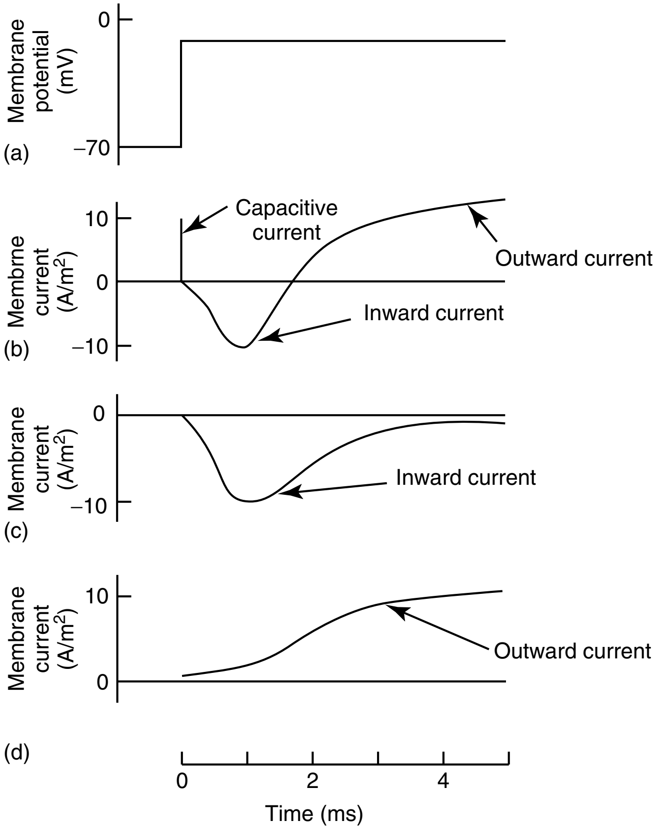

Fig. 16.9 Current flows across the squid axon under voltage-clamp conditions. During a depolarization of the membrane potential (a), an early inward and later outward current is observed following the capacitive transient. The ionic current flow (b) can be resolved into the sum of an inward Na$^+$ current (c) and an outward K$^+$ current (d).

terminal, the depolarization opens voltage-gated Ca^{2+} channels that are concentrated there, and the resulting Ca^{2+} influx triggers a series of protein interactions that cause the vesicle membranes to fuse with the presynaptic membrane resulting in neurotransmitter release into the synaptic cleft (Kandel, Schwartz and Jessell, 2000). In some cases, the vesicles fuse only slightly with the membrane rather than integrating fully into it, thereby resulting in faster recycling of the vesicle membrane. Thus, presynaptic membranes provide electrical-to-chemical transduction.

A small baseline level of neurotransmitter release into the synaptic cleft occurs spontaneously under resting conditions. The effect of the arrival of each action potential at the axon terminal can be viewed as a Bernoulli trial with regard to the fate of individual vesicles. A higher frequency of action potentials results in a longer time during which the membrane is depolarized, thereby permitting a greater Ca^{2+} influx and a greater number of fused vesicles; thus, neurotransmitter release is a graded response.

16.4
Neuronal Systems

16.4.1
Overview

Specialized neurons or neuroepithelial cells called *sensory receptors* detect external stimuli and transduce them into electrical signals, the language of the nervous system. The signals are processed by groups of neurons in local neuronal networks whose activity may be synchronized with that in other regions of the nervous system. Synchronization is accomplished by neurons that are hardwired between local networks and by *volume transmission* mediated by electrical and chemical messengers. The ultimate result of the sensory inputs and subsequent regional interactions is the behavior manifested by the organism. Not surprisingly, mechanistic knowledge of neuronal systems is inversely related to the complexity of the level of the activity under consideration. There is convincing evidence that sensory receptor function is mediated by changes in conductance of ion channels, but the mechanisms for higher brain functions, such as behavior, *memory*, and consciousness, are poorly understood.

16.4.2
Sensory Systems

Mechanical, energetic, and chemical stimuli are transduced by receptor cells (Table 16.2), resulting in electrical activity that serves as the basis of conscious or unconscious perception of the internal and external environment, and that controls the autonomic regulatory systems in the body. Some life forms possess sensory capabilities not presently known to occur in human beings. In the case of electroreception, for example, the catfish *Kryptopterus bicirrhis* can detect the presence of 2 μV m^{-1} at 10 Hz (Kolomytkin *et al.*, 2007). Magnetoreception, which was known to occur in bacteria and birds, was recently discovered in human beings (Carrubba *et al.*, 2007).

In all known sensory systems, stimulus transduction is mediated by a change in conductance of ion channels. The stimulus may interact with a membrane receptor protein coupled to a channel protein via a second-messenger system, as in the cases of chemical and light stimuli. Alternatively, the stimulus may act directly on the channel to produce a deformation that alters conductance, as in the sense of touch. In all cases, transduction ultimately produces a change in channel conductance that gives rise to a change in membrane potential called the *receptor potential.* The receptor potential is smoothly graded in proportion to the strength of the stimulus, and propagates electrotonically to the presynaptic membrane (neuroepithelial cell) or to the site of generation of an action potential (neuron).

Signal transduction can be illustrated by considering the retinal photoreceptor cells (Figure 16.10). The cells have a membrane potential of about −40 mV in the absence of light, and become hyperpolarized in proportion to the intensity of the absorbed light (Figure 16.10a), thereby reducing the amount of neurotransmitter (glutamate) normally released onto neurons synapsed with

Tab. 16.2 Main types of sensory modalities.

Sense	Stimulus	Receptor Cell	Receptor Location
Mechanical			
Balance	Mechanical	Hair cells	Vestibular organ
Hearing	Mechanical	Hair cells	Cochlea
Pressure	Mechanical	Neurons	Tissue
Vascular pressure	Mechanical	Neurons	Blood vessels
Muscle stretch	Mechanical	Neurons	Muscle spindle
Muscle tension	Mechanical	Neurons	Tendons
Joint position	Mechanical	Neurons	Ligaments
Osmotic pressure	Osmotic pressure	Osmoreceptors	Hypothalamus
Pain	Various	Neurons	Many organs
Energetic			
Vision	Photons	Photoreceptors	Retina
Temperature	Temperature	Neurons	Tissue
Electroreception	Electric fields	Electroreceptors	Skin (brain?)
Magnetoreception	Magnetic fields	Magnetoreceptors	Unknown
Chemical			
Arterial oxygen	O_2 tension	Neurons	Carotid body
Glucose	Glucose	Glucoreceptors	Hypothalamus
pH (cerebrospinal fluid)	Ions	Ventricle cells	Medulla
Taste	Chemicals	Taste-bud cells	Tongue
Smell	Molecules	Olfactory receptors	Nose
Pain	Various	Neurons	Many organs

the photoreceptor cell. The molecular details (Figure 16.10b) involve opsin, a G-protein-coupled receptor, and 11-*cis*-retinal, a light-absorbing chromophore bound to the opsin. In the absence of light, high levels of cyclic guanosine 3′-5′ monophosphatase (cGMP) keep cGMP-gated Na^+ channels open, thereby allowing an inward Na^+ current. Photon absorption by the chromophore activates transducin (the G protein) that induces phosphodiesterase to hydrolyze cGMP into GMP; this leads to the closure of Na^+ channels, which causes the cell to become hyperpolarized (Fesenko, Kolesnikov and Lyubarsky, 1985).

Evolution employed the basic physical principles governing the interaction of force and matter to produce many different mechanisms for transmembrane ion flow. Receptors that transduce mechanical force or pressure (mechanoreceptors) are examples. When the stimulus is applied to the gate of the ion channels, changes occur in transmembrane ion flow that result in the generation of receptor potentials according to the GHK equation. Some mechanoreceptors operate by means of cytoskeletal proteins that transmit force to the channel gate; in other cases force is applied by means of lateral tension in the cell membrane. Some mechanoreceptors trigger an action potential only when they produce a threshold receptor potential, while others continuously trigger action potentials with a certain frequency. In the latter cases, the force changes the probability of the channels

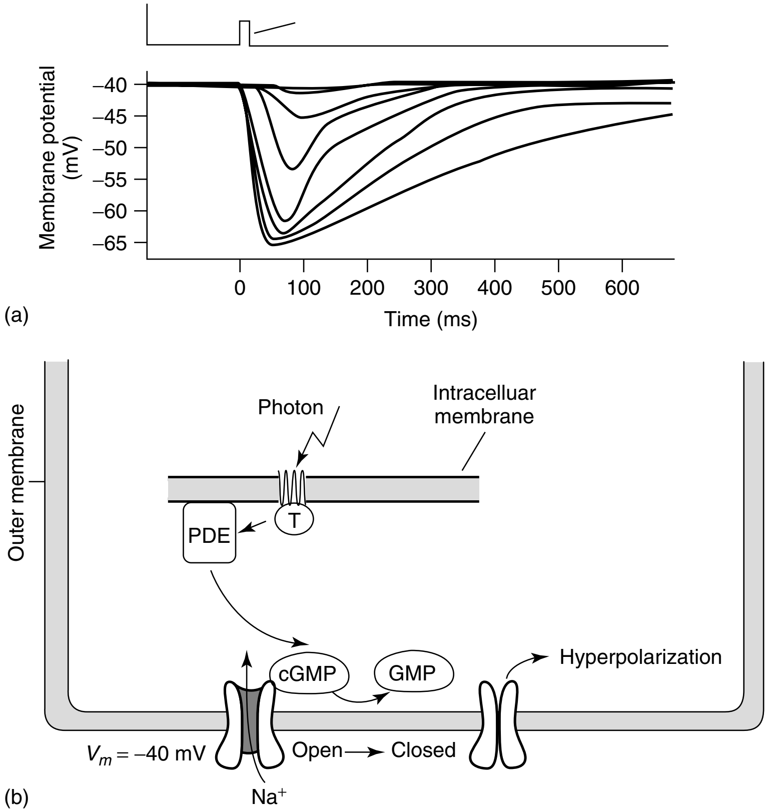

Fig. 16.10 Phototransduction in the retina. (a) Photoreceptor cell stimulated with progressively more intense light flashes becomes correspondingly more hyperpolarized. (b) One photon is sufficient to cause a change in the membrane conductance because of the amplification that occurs in the transduction cascade (a single activated chromophore can cause the hydrolysis of $> 10^5$ cGMP molecules). T, transducin; PDE, phosphodiesterase; and cGMP, cyclic guanosine 3′-5′ monophosphatase.

to be in the open state, thereby allowing continuous modulation of the frequency of the action potentials.

The hair cells in the inner ear have a well-understood force-transduction mechanism (Figure 16.11). Projections from the cells (stereocilia) contain ion channels whose gates are linked by specialized proteins. Deflection of the stereocilia by sound pressure results in a bending moment that changes the tension in the links, thereby changing the probability of the channel to be in the open state (Kachar *et al.*, 2000). The stiffness of the stereocilia in hair cells of the bullfrog is about 1 mN m^{-1}, so that a force of 100 pN is sufficient to produce a deflection of ≈100 nm, which stretches the tip links by about 12 nm. The tip link works as a spring, shifting the channel gate by ∼2 nm, which is sufficient to open most channels. Myosin I molecules maintain resting tension on the channels to bias them to the most sensitive part

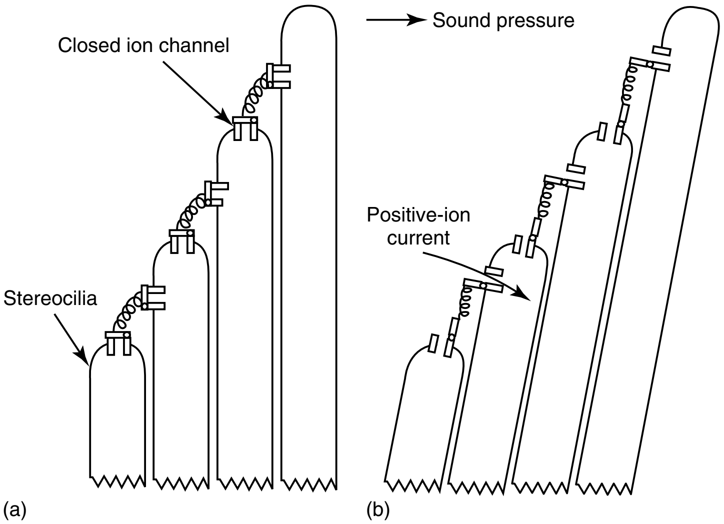

Fig. 16.11 Sound transduction by hair cells. (a) Stereocilia on hair cells contains ion channels whose gates are mechanically connected. (b) Sound pressure results in bending moments that open the channel gates.

of their activation curve (Sukharev and Corey, 2004).

The mechanisms responsible for the operation of exotic sensory systems such as electroreceptors are far less understood. It was suggested that Ca^{2+} channels in the apical and basal membrane provided the sensitive element (Bullock *et al.*, 2005). However, the electric-field sensitivity of all known Ca^{2+} channels is too low to account for the observed behavior of electroreceptors. One possibility is that the sensitive element is a gel particle connected to an ion-channel gate in the cell's apical membrane. In theory, mechanical displacement of the gel in small electric fields could open an ion-channel gate (Kolomytkin *et al.*, 2007).

16.4.3
Neural Networks

Receptor cells have no functional significance within the nervous system unless they are part of a network. Network complexity varies across the phylogenetic scale from a simple two-cell reflex response to the highly complex neural networks that mediate human consciousness and behavior.

Mathematical neural networks have been developed (Section 16.4.3) in the hope that they could model biological learning. In such a network a neuron is modeled as a summing node that weights each input and transforms the sum in a predetermined manner to yield an output (Figure 16.12). Usually three or more layers of model neurons interconnected in such a way that the outputs of the model neurons in one layer are the inputs to the next layer. The output, f, of each neuron depends on other functions $f = f(g_1, g_2, \ldots, g_n)$, where g_i are functions of other functions $g_i = g_i(h_1, h_2, \ldots, h_m)$, and so on up to the first layer where the arguments are the input signals. The *back-propagation*

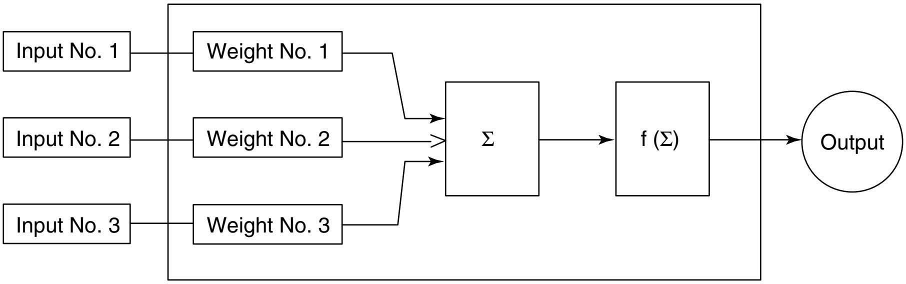

Fig. 16.12 Typical model neuron in an artificial neural network. The output is the transformed sum of the weighted inputs. A threshold is a common transfer function; in such a case, the weighted sum becomes the output only if it exceeds a specified level.

algorithm is an example. By following a systematic iterative procedure, the weights and transfer function of each model neuron can be adjusted so that the network produces correct outputs. In a sense, the artificial networks mimic the learning behavior of biological neurons, although they do not simulate the actual learning procedures used by the brain. The time required for an artificial network to learn increases rapidly as the size of the network increases, which is not true for brain networks. It is now clear that processing of information within the CNS is more complicated than that embodied in mathematical neural networks.

Biological neural networks are spatially localized functional units containing input and output neurons and local neurons that facilitate information-processing within the network. Operations within local networks are mediated by a combination of synaptic patterns (Figure 16.13). Synaptic divergence is a morphological pattern that amplifies neuronal activity by distributing it to many cells. Neurotransmitter release does not necessarily occur at each synapse within a particular morphological unit; in the neuromuscular junction, for example, only about 10% of the synapses are activated by entry of the action potential into the presynaptic axon terminal (Shepherd, 1990). Synaptic convergence facilitates the spatial and temporal integrative function of the dendritic tree (Figure. 16.13b and c). Presynaptic inhibition is a neuronal connectivity pattern in which a specific cell may be simultaneously presynaptic to one cell and postsynaptic to a second cell (Figure 16.13d); this arrangement permits a neuron to modify the activity of the cell without actually synapsing with it. The dendritic architecture together with the variety of possible synapses provide a wide range of computational possibilities, even in highly simplified networks (Figure 16.14). In general, the performance of biological neuronal networks is degraded if some of the component neurons fail, but catastrophic failure of the network does not occur.

16.4.4
Volume Transmission

In the hardwired model of the CNS, information transfer is mediated by synapses between adjacent cells (Figure 16.1), however, it can also occur via a nonsynaptic mechanism known

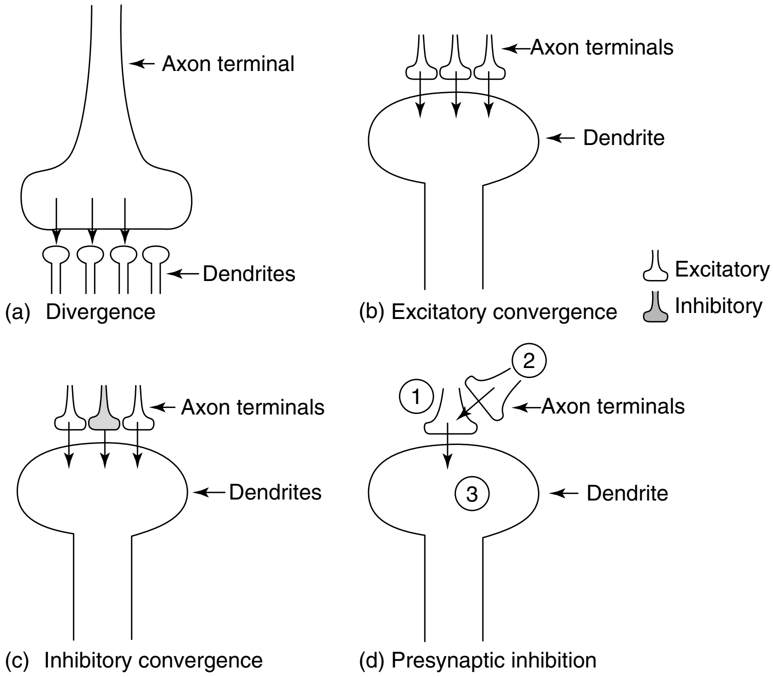
Axon terminal
Dendrites
(a) Divergence
Axon terminals
Dendrite
(b) Excitatory convergence
Excitatory
Inhibitory
Axon terminals
Dendrites
(c) Inhibitory convergence
1
2
Axon terminals
3
Dendrite
(d) Presynaptic inhibition

Fig. 16.13 Basic form of neuronal connectivity. (a) A divergent pattern produces a gain in activity (unless all but one of the synapses are silent). (b) and (c) Convergent patterns, which mediate neuronal integration. (d) Presynaptic inhibition. The effect of cell 1 on cell 3 is modified by cell 2, which does not directly affect cell 3 (adapted from Shepherd, 1990).

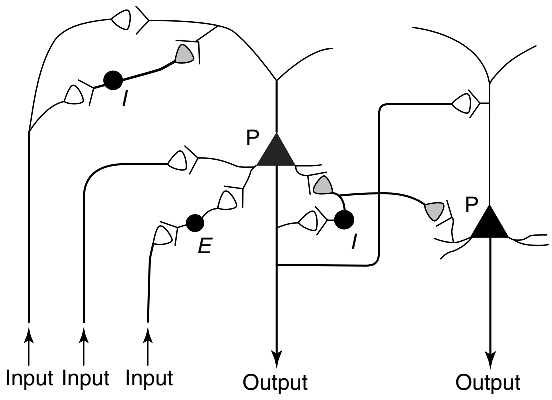
I
P
E
I
P
Input
Input
Input
Output
Output

Fig. 16.14 Basic circuit organization of the cerebral cortex. Pyramidal neurons P receive inputs, generate outputs, and interact with one another. Local neurons may be inhibitory (*I*) or excitatory (*E*). Because of the cell density, dendritic architecture, and synaptic organization, there exist a vast number of computational possibilities (adapted from Shepherd, 1990).

as *volume transmission*. For example, potentials produced by groups of cells propagate in the electrically conductive extracellular fluid and induce effects in distant cells such as synchronizing or altering their firing probability (Bullock, 1997). For the special case of *n* point current sources in a homogeneous conductor, the potential is

$$V(\mathbf{r}, t) = \frac{1}{4\pi\sigma} \sum_{i=1}^{n} \frac{I_i(t)}{R_i} \tag{5}$$

where $I_i(t)$ is the total current flowing from the ith point source into a medium of conductivity σ and R_i is the distance of the ith source from the field point $\mathbf{r}$. Thus, in principle, ionic flow in each neuron makes a contribution to the electrical environment of all other neurons in the CNS.

Diffusion of neuroactive substances through the brain (*volume diffusion*) is another important form of volume transmission. Neurotransmitters or other molecules may be released in a region not containing synapses (or may escape from the region of a synapse) and diffuse to other parts of the CNS through the cerebrospinal fluid in the extracellular space of the brain, which occupies about 20% of the brain's volume, thereby reaching distant targets.

Neuropeptides (a subset of neurotransmitters) are good candidates for volume transmitters because they are present in the extracellular fluid for relatively long periods and have relatively high affinity for their receptors. Neuropeptide Y, which is released in certain regions of the thalamus and hypothalamus, is an example. The receptors for neuropeptide Y are located several millimeters from the points of release; since it diffuses through the brain at about 1 mm h^{-1}, the spatial separation of secretion and binding of neuropeptide Y suggests that it may mediate slow information transfer within the brain.

Nitric oxide and carbon monoxide are also possible volume signaling agents. Nitric oxide is produced in both neurons and glia, and can quickly diffuse through the tissue.

In addition to isotropic diffusion, preferential diffusion along extracellular fiber bundles within the brain may occur (Bjelke *et al.*, 1995).

The traditional view was that neurotransmitter release always involved calcium-dependent fusion of neurotransmitter-laden vesicles at neuronal presynaptic membranes. However, *uptake carriers* can also secrete neurotransmitter (Atwell, Barbour and Szatkowski, 1993) via a nonvesicular process that does not require Ca^{2+} and that usually occurs away from the synaptic cleft. Thus, uptake carriers provide a possible source of diffusable neuroactive agents.

16.4.5
Brain Electrical Activity

Postsynaptic potentials and action potentials propagate electronically by volume conduction, yielding the *electroencephalogram* (EEG), a nonstationary, time-dependent voltage measured on the scalp. Spontaneous changes in the EEG are pathognomonic for some diseases, including brain tumors, epilepsy, and infection. Changes in the EEG induced by sensory or cognitive stimuli are called *evoked potentials*; they are used to study brain function and to detect abnormalities in the nervous system.

Even though it is universally agreed that brain electrical activity is nonlinear, the mathematical methods normally used to extract information from the EEG have been linear; chief among them are Fourier analysis and time averaging. Phase-space-based methods and other nonlinear techniques hold great promise for revealing previously unsuspected aspects of organized neuronal activity.

For example, consider the problem of determining whether environmentally strong electromagnetic fields (EMFs) are transduced by the mammalian nervous system. In principle, a solution could be found by comparing the EEG in the presence and absence of an EMF; a statistically reliable difference would indicate that transduction had occurred. This problem defied solution by traditional linear methods of analysis but was solved (Carrubba *et al.*, 2007) using a nonlinear method known as *recurrence analysis* (Webber and Zbilut, 1994), which is capable of detecting deterministic dynamical activity in a nonstationary signal such as the EEG (Figure 16.15).

16.4.6
Learning and Memory

Memory, the accessible storage of information, is encoded in multiple regions in the brain by different neuronal circuits depending on whether the memories are explicit (facts) or implicit (skills, habits, and behaviors) (Kandel, Schwartz and Jessell, 2000). The encoding is believed to consist of changes in the patterns of

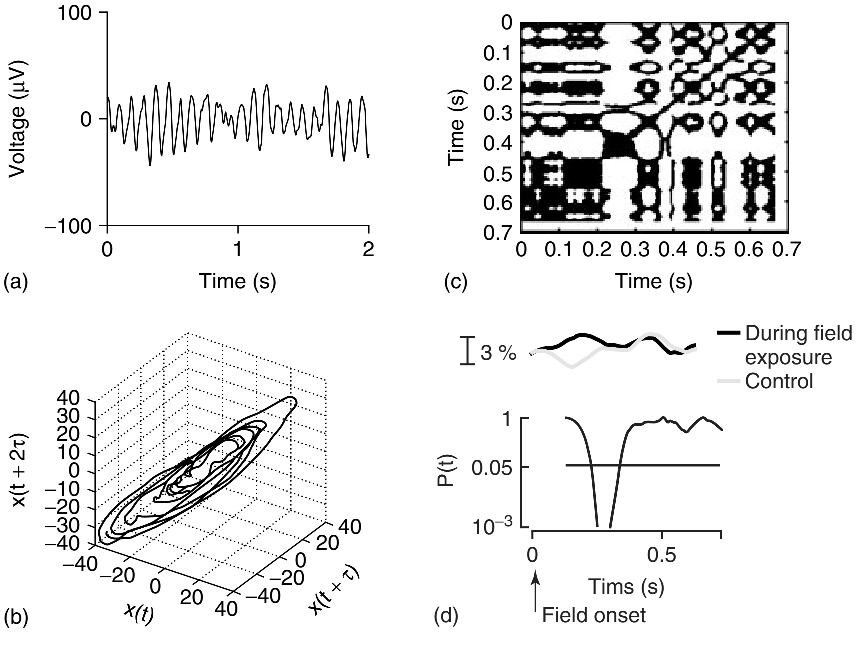

Fig. 16.15 Use of recurrence analysis to detect an onset evoked potential triggered by a 2-gauss, 60-Hz magnetic field. (a) Typical electroencephalogram recorded from the scalp of a 20-year-old male. (b) Appearance of the electroencephalogram in a three-dimensional phase-space ($\tau = 3.3$ ms). (c) Recurrence plot corresponding to the phase-space plot. (d) Upper curves, numerical characterization of a series of recurrence plots, using the recurrence analysis parameter percent recurrence (Webber and Zbilut, 1994). The gray curve is the average value of a series of control epochs. The solid curve is the average of the experimental epochs (field turned on at $t = 0$). The subject exhibited a nonlinear magnetosensory evoked potential about 300 ms after field onset. P, probability of a difference between exposed and control epochs (Carrubba *et al.*, 2007).

synaptic connections among neurons. If the pattern plasticity is labile, the result is *short-term memory*; if the change is permanent, the information is essentially imprinted on the organism resulting in *long-term memory*.

The NMDA glutamate postsynaptic receptor is believed to play a crucial role in formation of long-term memory (Kandel, Schwartz and Jessell, 2000). The receptor's ion channel opens when glutamate and glycine are bound to the receptor while the postsynaptic membrane is depolarized. The necessary conditions are provided by the occurrence of repeated nerve pulses and the presence of other glutamate receptors. The low-selective cation channel of the NMDA receptor transfers Ca^{2+} from outside to inside the cell.

It is assumed that long-lasting improvements in synaptic transmission occur when the intracellular calcium concentration exceeds a critical threshold in the postsynaptic compartment. One possible explanation involves the irreversible activation of a calcium–calmodulin-dependent kinase (CaMK) by Ca^{2+} ions. After activation, CaMK can phosphorylate and regenerate itself independently of Ca^{2+} (Figure 16.16). Mathematical models have predicted that spontaneous reversion to the inactive state would be unlikely, and the active Ca^{2+}-independent state could potentially last a lifetime. The kinase

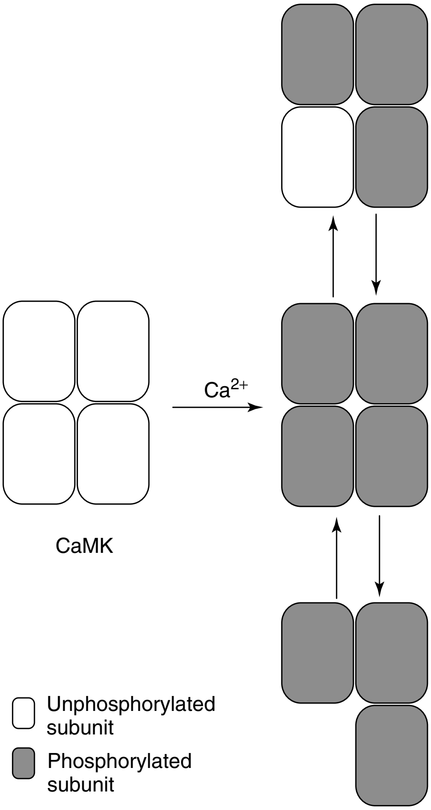

Fig. 16.16 Irreversible activation of calcium–calmodulin-dependent kinase (CaMK). In the inactivated state, CaMK's subunits are unphosophorylated. Stimuli that induce memory activate CaMK by Ca^{2+}-dependent phosphorylation of the subunits; thereafter, autophosphorylation occurs independently of Ca^{2+}. Thus, if a subunit becomes dephosphorylated (top), CaMK can return to full activity by means of a Ca^{2+}-independent process. If a subunit becomes unbound or degraded (bottom), the subunit is replaced independently of Ca^{2+} (Hanson and Schulman, 1992).

could, therefore, mediate long-term potentiation in synaptic function by phosphorylating another protein, for example AMPA receptor (Nayak, Moore and Browning, 1996).

Other mechanisms involved in the production of late long-term memory involve induction of specific changes in gene expression and protein synthesis by persistent activation of protein kinases (Pastalkova *et al.*, 2006; Serrano, Yao and Sacktor, 2005).

It has been argued that memory is distinct from events that occur at the cellular and molecular levels, and therefore there is no necessary conceptual link between behavior and the underlying physical processes (Skinner, 1966). If the issue of memory formation is approached from this perspective, the goal becomes one of understanding the principles that govern formation of the neural networks that mediate memory (as opposed to the goal of discovering molecular mechanisms). One of the most advanced efforts in this regard is that of Tsien and his colleagues. A broad outline of their work is given below; readers interested in the details should consult the original reports (Lin *et al.*, 2007; Lin *et al.*, 2005; Lin, Osan and Tsien, 2006).

When a memory-producing event (MPE) occurs, the baseline spiking activity of the neurons involved in encoding the memory exhibit four different kinds of changes (Figure 16.17a). Tsien and colleagues developed a device that allowed simultaneous monitoring of hundreds of randomly selected neurons in the hippocampus of an awake freely moving mouse, and then observed the responses to various MPEs (Figure 16.17b). When the results were analyzed using a pattern-recognition algorithm and then projected to a three-dimensional space (D-space) whose axes were abstract concepts that bore no relation to ordinary physiological parameters, clustering of the patterns from a particular mouse subjected to different MPEs was observed (Figure 16.17 c), indicating that different events were encoded by different changes in spike activity. Once the memories were generated and characterized by D-space analysis, they could be observed to occur spontaneously, suggesting that the mouse was remembering a particular event.

Further analysis of the structure of each of the D-space clusters led to the discovery of the existence of groups of neurons (cliques) that were activated by some MPEs but not others. For example, a particular clique might be activated by a fall or a shake but not a drop, whereas another clique might be activated by a fall or a drop, but not by a shake. Analysis of the differential response of the cliques in the context of different MPEs provided good evidence for the existence of a hierarchical basis for memory formation that consisted of a kind of binary code represented by the activation and nonactivation of particular neural cliques (Figure 16.17d).

In summary, neural cliques appear to serve as functional coding units that give rise to memories. Different cliques extract distinct features of an event; the activated cliques are linked in the brain to form an accessible, categorical, hierarchical representation of the event. This structure permits new memories to be formed by means of modifications or substitutions in the code for previous memories. For example, this time the man standing on the corner is wearing a white shirt. Thus the brain is not like a camera that records every detail, but

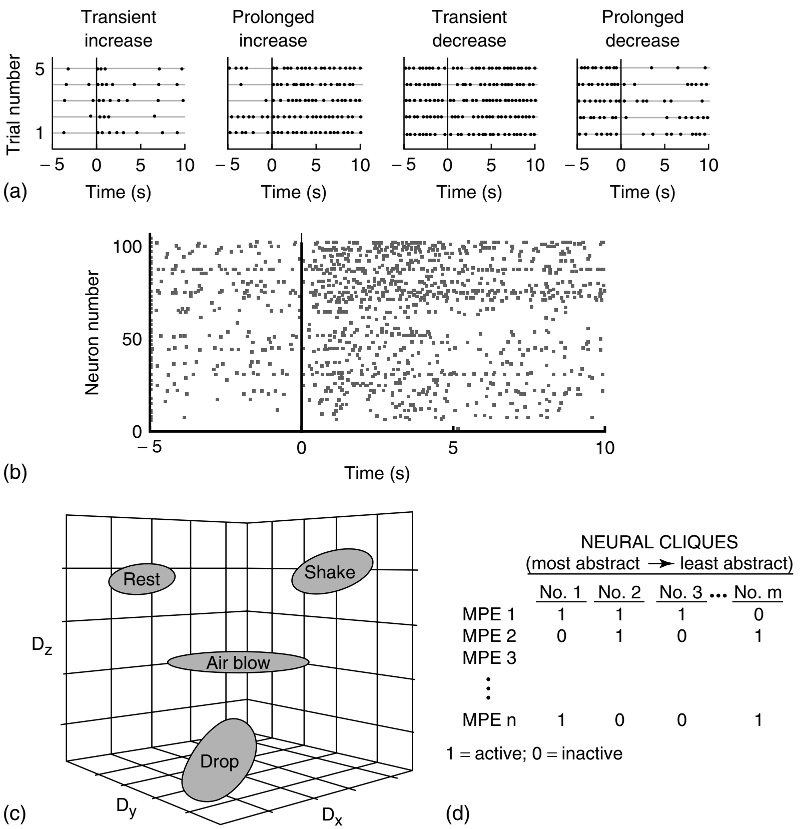

Fig. 16.17 Experimental detection of memory formation in a mouse. (a) Spike raster plots for five repetitions of an event, showing the four major types of changes that occur in the spike patterns (stimulus applied at 0 seconds). (b) Spike raster plots of 100 simultaneously recorded neurons in a mouse (stimulus applied at 0 seconds). (c) Locations in D-space associated with four different memory-producing events (MPEs). (d) Binary codes for specific MPEs.

rather functions by abstracting features that are common to quite different MPEs. A hierarchical organization of neural cliques also provides a general means for representing other types of information such as sensory perception.

16.4.7
A Nonlinear Future

As we said in the beginning, neurobiophysics is the study of the structure and function of the nervous system as seen from the perspective of physics, which is the view that a proper description of nature's behavior should be formulated in mathematical terms involving abstract concepts. Neurobiophysics began with the development of the Hodgkin–Huxley (HH) equations. Although they are still regarded as the prototypical explanation for the development of the action potential, there is no reasonable hope of being able to explain the behavior of neuronal networks based on the reductionistic approach embodied in the HH equations.

Virtually all the great advances in neuroscience have been based on a reductionistic approach to the nervous system, which is more or less equivalent to assuming that it is essentially linear and seeking the knowledge that could be gained under this assumption. Perhaps most of what can be understood within this perspective has already been elucidated. If so, future advances will require that neuroscience become more physics-like with regard to reliance on abstract concepts and mathematical description. It is interesting to note that at least two recent developments regarding the operation of the nervous system, the discovery of the human magnetic sense (Carrubba *et al.*, 2007) and an explanation of the basis for memory formation (Lin *et al.*, 2007; Lin *et al.*, 2005; Lin, Osan and Tsien, 2006), were based explicitly on the notion that the phenomena studied were nonlinear in nature. The hallmark of both experimental approaches was the reliance on abstract concepts and mathematical analysis and the avoidance of analytical techniques that amount to averaging away the phenomenon of interest. We expect that further advances will be made when this approach is followed in the context of other problems associated with the nervous system.

Glossary

Action potential: a self-propagating electrical signal initiated in a neuron when the membrane potential in a localized area of the cell membrane becomes depolarized beyond a threshold value.

Antidromic action potential: movement of an action potential in a direction opposite to that of normal propagation. Back-propagation.

Axon: the neuronal process that transmits the action potential from the cell body to the synapse.

Back-propagation algorithm: an iterative mathematical procedure for training an artificial neural network.

Connexin: protein subunit of a gap junction.

Delayed rectifier: an axonal K^+ channel that opens with depolarization and is largely responsible for repolarizing the axon membrane following an action potential.

Dendrites: branches that extend from the neuronal cell body and function to receive messages from other neurons.

Dendritic spines: narrow projections from dendrites present on many types of neurons; they can be the locus of synaptic inputs.

Direct chemical synapse: a synapse in which neurotransmitter molecules bind to ionotropic receptors (ion channels) in the postsynaptic membrane. The binding produces a conformational change in the receptor, thereby allowing ions to flow in or out of the cell, resulting in a change of membrane potential.

Electroencephalogram (EEG): a spontaneous nonstationary electrical signal consisting of the sum of the action potentials and synaptic potentials occurring throughout the brain.

Electrotonus: passive spread of electrical potential; occurs in the neuronal membrane.

Evoked potential: an electrical potential recorded on the scalp following presentation of a sensory stimulus to the subject (to be distinguished from the electroencephalogram, which is a spontaneous electrical potential).

Gap junctions: ion channels that connect the cytoplasm of adjacent cells, thereby allowing ions and small signaling molecules to pass between the cells.

Gating current: the movement of electrical charge in voltage-gated channels, which occurs during formation of a transmembrane ion pore.

Hillock: location in a neuron where the action potential is initiated.

Indirect chemical synapse: a synapse in which neurotransmitter molecules bind to receptors in the postsynaptic membrane, leading to activation of a second-messenger signaling system (usually G proteins). The messenger molecules open ion channels thereby allowing ions to flow in or out of the cell, resulting in a change in membrane potential.

Ion channels: transmembrane proteins that allow ions to pass through the cell membrane.

Ligand-gated channels: ion channels whose conductance depends directly or indirectly on the binding of a neurotransmitter to the channel.

Long-term memory: information storage in the nervous system that causes a permanent change in synaptic patterns.

Membrane potential: electrical potential of the cell interior relative to the cell bath.

Memory: accessible storage of information in the brain.

Neuron: the electrically active cell of the nervous system.

Neurotransmitter: a chemical agent synthesized and secreted by a presynaptic cell that effects a communication with a postsynaptic cell.

Plasticity: the tendency of synapses and neural circuits to change as a result of activity.

Receptor potential: initial response of a receptor cell to a stimulus, consisting of a change in voltage across the receptor membrane.

Receptors: cells that directly transduce a chemical, mechanical, or energetic stimulus to produce the receptor potential.

Recurrence analysis: a mathematical technique for detecting nonlinear deterministic activity in a time series such as a biological signal.

Resting membrane potential: the membrane potential under steady-state conditions (no net current or applied voltage).

Reversal potential: for a channel with one permeant ion, the equilibrium potential determined from the Nernst equation.

Selectivity: the permeant ion species of a channel pore.

Sensory receptor: a structure that recognizes a stimulus in the external or internal environment of an organism.

Short-term memory: information storage resulting in a temporary change in synaptic patterns.

Site-directed mutagenesis: a form of genetic engineering that permits changes in specific nucleic acids in a given gene.

Synapses: specialized junctions between neighboring neurons that facilitate transmission of signals between the cells.

Uptake carriers: membrane proteins that facilitate uptake or release of neurotransmitters, ions, and other substances.

Voltage-clamp: a technique for studying transmembrane ion kinetics in which the membrane potential is held at a predetermined value.

Voltage-gated: ion channels whose conductance depends on the membrane potential.

Volume diffusion: a form of volume transmission in which neurotransmitters diffuse through the cerebrospinal fluid, thereby reaching distant receptors.

Volume transmission: information transfer within the nervous system by means other than a synapse.

References

Atwell, D., Barbour, B., and Szatkowski, M. (1993) Nonvesicular release of neurotransmitters. *Neuron*, **11**, 401–7.

Benarroch, E.E. (2007) Sodium channels and pain. *Neurology*, **68**, 233–36.

Bjelke, B., England, R., Nicholson, C., Rice, M., Lindberg, J., Zoli, M., Agnati, L., and Fuxe, K. (1995) Long distance pathways of diffusion for dextran along fibre bundles in the brain. Relevance for volume transmission. *Neuroreport*, **6**, 1005–9.

Bullock, T.H. (1997) Signals and signs in the nervous system: the dynamic anatomy of electrical activity is probably information rich. *Proc. Natl. Acad. Sci., U.S.A.*, **94**(1), 1–6.

Bullock, T.H., Bennett, M.V., Johnston, D., Josephson, R., Marder, E., and Fields, R.D. (2005) Neuroscience. The neuron doctrine, redux. *Science*, **310**, 791–93.

Carrubba, S., Frilot II, C., Chesson Jr, A.L., and Marino, A.A. (2007). Evidence of a nonlinear human magnetic sense. *Neuroscience*, **144**, 356–67.

Deutsch, S. and Micheli-Tzanakou, E. (1987) *Neuroelectric Systems*, New York University Press, New York.

Doyle, D.A., Morais, C.J., Pfuetzner, R.A., Kuo, A., Gulbis, J.M., Cohen, S.L., Chait, B.T., and MacKinnon, R. (1998) The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science*, **280**, 69–77.

Fesenko, E.E., Kolesnikov, S.S., and Lyubarsky, A.L. (1985) Induction by cyclic GMP of cationic conductance in plasma membrane of retinal rod outer segment. *Nature*, **313**, 310–13.

Gouaux, E. and Mackinnon, R. (2005) Principles of selective ion transport in channels and pumps. *Science*, **310**, 1461–65.

Hanson, P.I. and Schulman, H. (1992) Inhibitory autophosphorylation of multifunctional Ca^{2+}/calmodulin-dependent protein kinase analyzed by site-directed mutagenesis. *J. Biol. Chem.*, **267**, 17216–17224.

Hervé, J.C., Bourmeyster, N., Sarrouilhe, D., and Duffy, H.S. (2007) Gap junctional complexes: from partners to functions. *Prog. Biophys. Mol. Biol.*, **94**, 29–65.

Hille, B. (1992) *Ionic channels of excitable membranes*, Sinauer, Sunderland MA.

Hodgkin, A. and Huxley, A. (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, **117**, 500–44.

Kachar, B., Parakkal, M., Kurc, M., Zhao, Y., and Gillespie, P.G. (2000) High-resolution structure of hair-cell tip links. *Proc. Natl. Acad. Sci., U.S.A.*, **97**, 13336–41.

Kandel E.R., Schwartz J.H., and Jessell T.M. (eds) (2000). *Principles of Neural Science*, McGraw-Hill, New York.

Kolomytkin, O.V., Dunn, S., Hart, F.X., Frilot, C., Kolomytkin, D., and Marino, A.A. (2007) Glycoproteins bound to ion channels mediate detection of electric fields: a proposed mechanism and supporting evidence. *Bioelectromagnetics*, **28**, 379–85.

Lin, L., Chen, G., Kuang, H., Wang, D., and Tsien, J. (2007) Neural encoding of the concept of nest in the mouse brain. *Proc. Natl. Acad. Sci., U.S.A.*, **104**, 6066–71.

Lin, L., Osan, R., Shoham, S., Jin, W., Zuo, W., and Tsien, J. (2005) Identification of network-level coding units for real-time representation of episodic experiences in the hippocampus. *Proc. Natl. Acad. Sci., U.S.A.*, **102**, 6125–30.

Lin, L., Osan, R., and Tsien, J. (2006) Organizing principles of real-time memory encoding: neural clique assemblies in universal neural codes. *Trends Neurosci.*, **29**, 48–57.

Magee, J.C. and Johnston, D. (1997) A synaptically controlled, associative signal for hebbian plasticity in hippocampal neurons. *Science*, **275**, 209–13.

Markram, H., Lubke, J., Frotscher, M., and Sakmann, B. (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, **275**, 213–15.

McKenna T., Davis J., and Zornetzer S.F. (eds) (1992). *Single Neuron Computation*, Academic Press, Boston.

Nayak, A.S., Moore, C.I., and Browning, M.D. (1996) Ca^{2+}/calmodulin-dependent protein kinase II phosphorylate of the presynaptic protein synapsini I is persistently increased during long term potentiation. *Proc. Natl. Acad. Sci., U.S.A.*, **93**, 15451–56.

Pastalkova, E., Serrano, P., Pinkhasova, D., Wallace, E., Fenton, A., and Sacktor, T. (2006) Storage of spatial information by the maintenance mechanism of LTP. *Science*, **313**, 1141–44.

Rall, W. (1960) Membrane potential transients and membrane time constant of motoneurons. *Exp. Neurol.*, **2**, 503–32.

Sejnowski, T.J. (1997) The year of the dendrite. *Science*, **275**, 178.

Serrano, P., Yao, Y., and Sacktor, T. (2005) Persistent phosphorylation by protein kinase Mzeta maintains late-phase long-term potentiation. *J. Neurosci.*, **25**, 1979–84.

Shepherd G.M. (ed) (1990). *The Synaptic Organization of the Brain*, Oxford Press, New York.

Skinner, B.F. (1966) The phylogeny and ontogeny of behavior. Contingencies of reinforcement throw light on contingencies of survival in the evolution of behavior. *Science*, **153**, 1205–13.

Sukharev, S.and Corey, D.P. (2004). Mechanosensitive channels: multiplicity of families and gating paradigms. *Sci. STKE*, Feb. **3**(219), re4.

Suzuki, M., Morita, T., and Iwamoto, T. (2006) Diversity of Cl(-) channels. *Cell. Mol. Life Sci.*, **63**, 12–24.

Webber, C.L. and Zbilut Jr, J.P. (1994) Dynamical assessment of physiological systems and states using recurrence plot strategies. *J. Appl. Physiol.*, **76**, 965–73.

Further Reading

Byrne J.H. and Roberts J.L. (eds) (2004). *From Molecules to Networks*, Elsevier, New York.

Dayan, P. and Abbott, L.F. (2005). *Theoretical Neuroscience*, MIT Press, Cambridge.

Izhikevich, E.M. (2007). *Dynamical Systems in Neuroscience*, MIT Press, Cambridge.

Koch, C. (2004). *Biophysics of Computation*, Oxford University Press, New York.

Niedermeyer, E. and Silva, F.L.d. (1999). *Electroencephalography: Basic Principles, Clinical Applications and Related Fields*, 4th edn, Urban and Schwarzenberg, Baltimore.

Nunez P.L. and Srinivasan R. (eds) (2006). *Electric Fields in the Brain: The Neurophysics of EEG*, Oxford University Press, New York.

Squire, L.R. and Kandel, E.R. (2000). *Memory: From Mind to Molecules*, Scientific American, New York.

Yuste, R. and Tank, D.W. (1996). Dendritic integration in mammalian neurons, a century after Cajal. *Neuron*, **16**, 701–16.

Zoli, M. and Agnati, L.F. (1996). Wiring and volume transmission in the central nervous system: The concept of closed and open synapses. *Prog. Neurobiol.*, **49**, 363–80.

NEUTRAL ATOMIC AND MOLECULAR COLLISION PROCESSES

M. KIMURA, *Argonne National Laboratory, Argonne, Illinois, U.S.A., and Department of Physics, Rice University, Houston, Texas, U.S.A.*

L. K. JOHNSON, *The Aerospace Corporation, Los Angeles, California, U.S.A.*

Introduction 323
1. **Overview** 324
2. **Theoretical Framework** 326
2.1 Inelastic Processes 326
2.1.1 High-Energy Collisions (>10 keV) 326
2.1.2 Intermediate-Energy Collisions (10 eV–10 keV) 327
2.1.3 Low-Energy Collisions (≤10 eV) 328
2.2 Elastic Processes 329
2.2.1 Glory Scattering 329
2.2.2 Rainbow Scattering 330
2.2.3 Orbiting Resonance 330
3. **Experimental Framework** 330
3.1 High to Superthermal Energies (10 eV–1 MeV) 331
3.1.1 Charge-Transfer Beams 331
3.1.2 Sputtered Beams 331
3.1.3 Seeded Supersonic Beams 331
3.1.4 Other Methods 332
3.2 Low to Thermal Energies (<10 eV) 332
3.3 Beams of Excited Species 332
3.3.1 Thermal Dissociation 332
3.3.2 Electron Impact 333
3.3.3 Optical Excitation 333
3.3.4 Chemical Enhancement 333
4. **Observables Relevant to Applications** 333
4.1 Differential and Total Cross Sections and Rate Constants 333
4.2 Stopping Power 333
4.3 Diffusion Constant and Mobility 333
5. **Application and Relevance** 334
5.1 Scientific Applications 334
5.1.1 Dynamics: The Determination of Interaction Potential 334
5.1.2 Spectroscopy: The Determination of Physical Constants and Identification 335
5.2 Technological Applications 335
5.2.1 Quantum Electronics: The Gas Laser 335
5.2.2 Plasma and Fusion Research 335
5.2.3 Surface Sciences 335
5.2.4 Energy 336
6. **Conclusion** 337
7. **Acknowledgment** 337
Glossary 337
Works Cited 338
Further Reading 338

INTRODUCTION

Historically, the interactions between neutral atoms and molecules have been of the greatest interest to chemists and engineers seeking to determine reaction rates as functions of various parameters such as temperature and pressure. An applied chemical process is typically investigated by modeling the complex of possible reactions, using experimentally determined rates to parametrize the system of equations constituting the model. Our understanding of a process of applied interest (combustion of gasoline in air, for example) thus relies directly on the measured rate constants. Other complex processes that are not principally chemical also rely on a database of rate constants or collision cross sections; good examples would be the behavior of fusion-confinement plasmas, molecular-beam-induced epitaxy of semiconductors, or the characteristics of the earth's atmosphere from ground level to space. Atomic and molecular experimental studies, along with

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

quantum solid-state and surface physics (which are all similar in technique), frequently are the principal connections between our best physical understanding of the microscopic world and our most sophisticated and profound technological challenges. Coupled with recent experimental advancements such as laser technologies, the understanding of these processes is increasingly of fundamental importance for manipulating materials at atomic and molecular levels. We are now in an era of nanometer technology.

1. OVERVIEW

Processes that may be possible upon collisions of neutral particles can be classified into one group involving species with quantum states at the ground (lowest) levels and another group involving species with quantum states at excited levels. Species in the two groups are formed as follows:

1. Ground state:

$A + BC(v,J) \rightarrow A + BC^+ + 1e^-$, ionization; (p1a)
$\rightarrow A^+ + BC + 1e^-$, stripping; (p1b)
$\rightarrow A + BC^*, A^* + BC$, electronic excitation; (p2a)
$\rightarrow A^- + BC^+, A^+ + BC^-$, electron transfer; (p2b)
$\rightarrow A + BC^* (v',J')$, rovibrational excitation; (p3)
$\rightarrow AB + C, AC + B, ABC$, chemical reaction. (p4)

2. Excited state: all processes [(p1) to (p4)], plus

$A^* + BC(v,J) \rightarrow A + BC^*$, excitation transfer; (p5)
$\rightarrow A + BC(v',J')$, quenching; (p6)
$\rightarrow A + BC^+ + 1e^-$, Penning ionization; (p7)
$\rightarrow ABC^+ + 1e^-$, associative ionization. (p8)

The species indicated by *A*, *B*, and *C* may each represent a simple atom or a complex molecule. The asterisk represents a species in an excited state [an electronic state or, for a molecule, a rotationally and vibrationally (rovibrationally) excited state that is designated by a set of vibrational and rotational quantum numbers (v,J)]. Processes (p1a) and (p1b) correspond to ionization events, with an electron being released from a target (ionization) or from a projectile (stripping). Similarly, electronic excitation (p2a) is possible for a target or a projectile or both. Upon collision, electron(s) may be captured either by a projectile to form $A^- + BC^+$ or by a target to form $A^+ + BC^-$. In chemical reaction (p4), reaction mechanisms are classified as either complex mechanisms, where reactants pass through a transition-state complex as an intermediate state, or direct mechanisms, where no complex is formed. The criterion defining these two types of mechanisms is whether the lifetime of a reaction complex is longer (in a complex mechanism) or shorter (in a direct mechanism) than its rotational time ($\sim 10^{-12}$ s). Chemical reaction (p4) (reactive scattering) frequently accompanies rovibrational excitation in products. The quenching process (p6) includes both electronic deexcitation to a lower level and rovibrational deexcitation. Penning ionization (p7) occurs when the excitation energy of A^* exceeds the ionization potential of the *BC* target. When the final products form a molecular ion, the process is an associative ionization (p8). An electronically excited molecule normally decays by dissociation or, for highly excited cases, by ionization.

Figure 1 is a schematic diagram showing

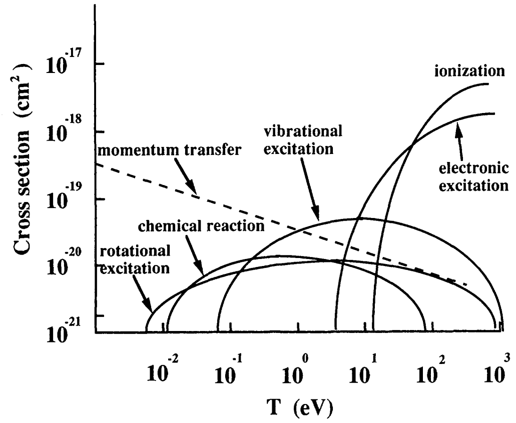

FIG. 1. Schematic diagram of probable processes and the magnitudes of their cross sections for a hypothetical collision species as a function of collision energy (*T*).

the probable processes and the approximate magnitudes of the corresponding cross sections for a hypothetical collision species as a function of collision energy. The actual magnitudes of the cross sections vary from system to system. When a polar molecule or a highly polarizable atom is involved in the collision, the inelastic cross sections are normally large, with magnitudes as large as about 10^{-15} cm^2 for chemical reaction and 10^{-17} cm^2 for vibrational excitation. When a nonpolar molecule or a slightly polarizable atom is a collision partner, the magnitudes of these cross sections are reduced significantly. Upon collision, electronic excitation and ionization (stripping) may be the dominant processes above a few tens of electron volts and up to the high-energy (>10 keV) region. If one member of the pair is a molecular species and an additional internal degree of freedom (e.g., a heavy nucleus) is involved, then vibrational and rotational (rovibrational) excitation processes become important in the low-energy (≲10 eV) region. In the same energy region (≲10 eV), chemical reaction (reactive scattering) may compete, if a colliding pair approaches close enough at sufficiently low energy to exchange a particle. Elastic (momentum transfer) processes are always associated with any collision and have their own interesting features, which are closely connected with various types of resonances. For neutral pairs, the interaction is weaker than for ionic pairs, with a long-range attractive van der Waals interaction with a dominant asymptotic term of R^{-6} when the two particles are separated by a distance R. When one of the two particles is an ion, the dominant asymptotic interaction is due to attractive dipole polarization and is of the form R^{-4}. At small R, strong overlap of electron charge clouds causes the interaction to become repulsive. At intermediate R, a balance between attractive and repulsive interactions sometimes causes a small minimum (bonding) at the equilibrium distance, forming a stable molecule.

As an actual example, Fig. 2 shows cross sections of processes resulting from collisions of H atoms with H_2 molecules. Because the H_2 molecule is nonpolar, the magnitudes of the cross sections are relatively small. Processes for ionization, dissociative ionization, collisional dissociation, and electronic exci-

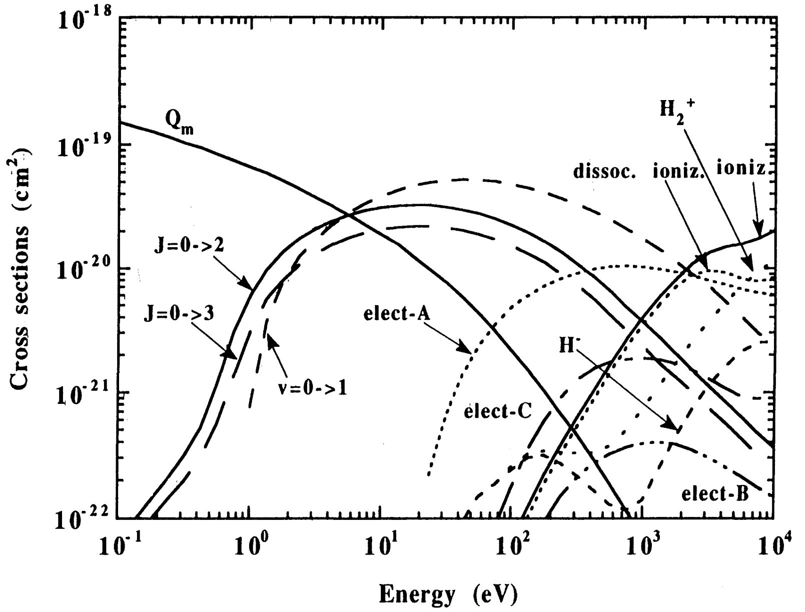

FIG. 2. Representative examples of various inelastic processes resulting from collisions of H atoms with H_2 molecules (adapted from Phelps, 1990). Notations in the figure are as follows: Q_m, momentum transfer; $J = m \rightarrow n$, rotational transition from m level to n level; $v = m \rightarrow n$, vibrational transition from m level to n level; elect-A, electronic excitation of the H_2 to the $A\Sigma$ state; elect-B, electronic excitation to $B\Sigma$; elect-C, electronic excitation to $C\Sigma$; H^-, H^- formation.

tation (elect-A, elect-B, and elect-C) are dominant above a few tens of electron volts. The H^- formation resulting from charge transfer is also a large contributor in this energy domain. Various vibrational excitation processes are dominant in the energy region between a few tens of electron volts and a few tenths of an electron volt; below this energy region, various rotational excitations dominate.

Theoretically, in the high-energy domain (>1 keV), the collision takes place impulsively, and the projectile leaves the interaction region very quickly. Hence, the interaction time is short (i.e., weak interaction), and the couplings among the various final states are considered to be weak. Thus, perturbation theory (such as the Born series), which normally accounts for only the initial and final states, can be applied to describe the collision dynamics reasonably well. As the energy decreases to a few kiloelectronvolts and further to electron volts, the interaction times become longer, causing a greater disturbance to the target. Therefore, various collision processes may occur simultaneously in a single collision. Because these processes couple strongly with each other in a complex manner as a function of collision energy and time, the processes all must be treated equally in any theoretical model. The model based on this physical picture is termed a close-coupling method. In this energy domain, the relative velocity of heavy particles (nuclei) is so much smaller than that of orbital electrons that the colliding particles are considered to form a collision complex or a quasi-molecule during the collision. Therefore, a complete understanding of collision dynamics in this energy domain requires an accurate knowledge of the adiabatic potential curves (surfaces in three dimensions) of the colliding system, obtained by a large-scale quantum-chemical procedure.

Experimentally, the study of neutral atomic and molecular collisions is a relatively mature field, although recent advances in experimental technique have brought renewed attention to previously impossible measurements. The emphasis in recent experimental work has been on well-prepared and well-analyzed quantum states of collision partners and products, as well as on examination of collisions under single-collision conditions. Early collision work examined the characteristics of atomic and molecular mixtures over a range of temperatures. These techniques usually are not sufficiently specific about which among processes (p1) to (p4) dominates. The evolution of the charged-particle beam, available over a wide range of energies, and the fact that an ion beam can be neutralized via process (p2b) provide experimentalists with good control over the initial dynamics of the neutral collision partners. Advances in beams produced by supersonic jets, pulsed or continuous, provide good control over the initial dynamics and states at low interaction energies, where cooling in the expansion of the jet flow conveniently restricts the number of populated rovibrational molecular states. In addition, techniques to detect relatively small (1–1000) numbers of atoms and molecules allow investigators to isolate collisions sufficiently to distinguish among the host of possible collision paths. Finally, the desire to prepare a specific quantum state or to use spectroscopic sensitivity to isolate product states (or both) has led to the widespread adoption of tunable lasers in experimental neutral-collision physics.

2. THEORETICAL FRAMEWORK

2.1 Inelastic Processes

2.1.1 High-Energy Collisions (>10 keV) When the velocity (v_{rel}) of the incident atom is larger than the orbital velocities (v_{el}) of the electrons in the target atom or molecule, electron transfer (p2b) rapidly becomes unimportant compared with ionization (p1a) and electronic excitation (p2a). This situation simplifies the theory substantially. Cross sections for this case can be evaluated on the basis of a Born-type expansion or its distorted-wave generalization. When the particle is sufficiently fast (but still nonrelativistic) and it interacts through the potential V, the general solution $\Psi(\xi)$ (where ξ represents all coordinates collectively) of the Schrödinger equation can be written formally by using the Green's function $G(k,\xi,\xi')$, as (atomic units are used throughout)

$$\Psi(\xi) = \Phi(\xi) + \int G(k,\xi,\xi')V(\xi',\xi')\Psi(\xi')d\xi', \quad (1)$$

where $\Phi(\xi)$ is a solution of the homogeneous equation of total angular momentum J,

$$\left[\frac{d^2}{d\xi^2} + k^2 - \frac{J(J+1)}{\xi}\right]\Phi(\xi) = 0. \tag{2}$$

Starting from the incident plane wave $\Phi_{k_i}(\xi)$ with wavelength k_i as a zero-order approximation, we may write $\Psi(\xi)$ in the Born series (which is a perturbation expansion) in powers of the interaction potential,

$$\begin{aligned}\Psi_k(\xi) = \Phi_{k_i}(\xi) &+ \int G(k,\xi,\xi')V(\xi')\Phi_{k_i}(\xi')d\xi' \\ &+ \int G(k,\xi,\xi')V(\xi')G(k,\xi',\xi'')V(\xi'') \\ &\times \Phi_{k_i}(\xi'')d\xi' d\xi'' + \dots .\end{aligned} \tag{3}$$

Correspondingly, with Eq. (3), the scattering amplitude can be written in the Born series as

$$f = -\frac{1}{4\pi}\langle\Psi_{k_f}|V + VGV + VGVGV + \dots|\Psi_{k_i}\rangle. \tag{4}$$

By substituting the zero-order plane-wave approximation for Ψ_{k_i} and Ψ_{k_f} and retaining only the first term of V, we obtain the first Born scattering amplitude.

The cross section for a transition from state i to state f, calculated in the lowest order in the interaction potential V between the projectile and target (the first Born approximation), is specifically

$$\sigma_{fi} = \frac{\mu^2}{2\pi k_i^2}\int_{K_{\min}}^{K_{\max}} K dK \int_0^{2\pi} d\phi |T_{fi}^B(K)|^2, \tag{5}$$

with

$$T_{fi}^B(\mathbf{K}) = \int d^3R\, e^{i\mathbf{K}\cdot\mathbf{R}} \int d^3r \varphi_f(r)V(r)\varphi_i(r). \tag{6}$$

Here, μ is the reduced mass of the colliding system, r represents the collective position of the electrons centered at the target, R is the internuclear separation, ϕ is the azimuthal angle, $\hbar\mathbf{k}_i$ and $\hbar\mathbf{k}_f$ are the momentum of the projectile before and after the collision, $\hbar\mathbf{K} = \hbar(\mathbf{k}_i - \mathbf{k}_f)$ is the momentum transfer, $\varphi_i(r)$ and $\varphi_f(r)$ are electronic wave functions of the target for the initial and final atomic states, and $V(r)$ is the interaction potential between two particles.

As the energy decreases, the first Born approximation [Eqs. (5) and (6)] is considered to be less accurate. For such a case, agreement can be achieved by retaining higher-order terms in the perturbation series in Eq. (4); however, rigorous evaluation of higher-order terms is extremely difficult, and only some approximate attempts have been made to date.

2.1.2 Intermediate-Energy Collisions (10 eV–10 keV) When $v_{\rm rel}$ approaches or becomes less than $v_{\rm el}$, the interaction time becomes longer, causing stronger disturbance to the target. The condition $v_{\rm el} \simeq v_{\rm rel}$ is also favorable for a jump by a target electron to the projectile (electron capture). Therefore, transition probabilities for many inelastic events become comparable in magnitude, as exemplified in Fig. 2. Furthermore, these processes couple strongly with one another. This situation requires the theory to treat of all the involved processes equally. The close-coupling method has been developed specifically for this purpose. In this energy region, a semiclassical treatment, in which electronic states are solved quantum mechanically while nuclear motions are treated classically, is still considered to be valid. The assumption is that the nuclei move along a classical path specified by the internuclear separation vector $\mathbf{R}(t) = \mathbf{v}_{\rm rel}t + \mathbf{b}$, where $\mathbf{v}_{\rm rel}$ is the relative velocity and $\mathbf{b}$ is the impact parameter, and that the electronic wave function $\Psi(\mathbf{r},t)$ satisfies a time-dependent Schrödinger equation of electronic Hamiltonian $H_{\rm el}$,

$$i\frac{\partial\Psi(\mathbf{r},t)}{\partial t} = H_{\rm el}\Psi(\mathbf{r},t). \tag{7}$$

The solution of Eq. (7) can be obtained by expanding the total scattering wave function $\Psi(\mathbf{r},t)$ in terms of a known basis set of electronic wave functions $\varphi_i(\mathbf{R}(t),\mathbf{r})$ in the form

$$\begin{aligned}\Psi(\mathbf{r},t) = &\sum_i a_i(t)\varphi_i(\mathbf{R}(t),\mathbf{r}) \\ &\times \exp\left(-i\int^t E_i(\mathbf{R}(t'))dt'\right).\end{aligned} \tag{8}$$

By substituting Eq. (8) into Eq. (7) and projecting onto any basis function $\varphi_j(\mathbf{R},\mathbf{r})$, one obtains the set of first-order coupled equations for time-dependent coefficients $a_i(t)$,

$$i\sum_j S_{ji}\frac{da_j(t)}{dt} = \sum_{i\neq j} V_{ij}a_i, \tag{9}$$

where S is the overlap matrix $S_{ji} = \langle\varphi_j|\varphi_i\rangle$, and V_{ij} describes the coupling matrix

$$V_{ji} = \left\langle \varphi_j \middle| H_{\text{el}} - \frac{\partial}{\partial t} \middle| \varphi_i \right\rangle.$$

The coupled Eqs. (9) are normally solved numerically, subject to the initial condition $a_i(t \rightarrow -\infty) = \delta_{ij}$, to extract the scattering amplitude $a_k(t \rightarrow +\infty)$ of a transition to state k. The square of the scattering amplitude gives a transition probability, and the cross section is defined as the integration of the probability over impact parameter b,

$$\sigma_k(E) = 2\pi \int bdb|a_k(E,b)|^2. \tag{10}$$

To describe the collision dynamics correctly and hence obtain accurate transition probabilities, selecting a best-suited basis set $\varphi_i(\mathbf{R}(t),\mathbf{r})$ for the expansion in Eq. (8) is important. In the kiloelectronvolt regime, the collision time is sufficiently short that the colliding particles are considered to retain their atomic character during the collision (a distant collision). In this circumstance, using an atomic orbital (AO) for $\varphi_i(\mathbf{R}(t),\mathbf{r})$ as an expansion basis set is adequate and correct; however, as energy decreases into the electronvolt region, the electronic wave function has time to adjust to the changing potential field of two nuclei, and the colliding particles are regarded correctly as forming a quasi-molecule during the collision (a close collision). In this case, a Born–Oppenheimer molecular orbital (MO) is a suitable choice for the expansion basis set to describe the collision system correctly. The theoretical treatment based on this MO is summarized in Sec. 2.1.3.

When the AO expansion is used and only initial and final states are retained (two-state approximation), the coupled Eq. (9) is known to recover the first Born approximation of Eqs. (5) and (6) in the impact-parameter representation.

2.1.3 Low-Energy Collisions ($\leq$10 eV)

When $v_{\text{rel}} \ll v_{\text{el}}$ the electronic wave functions have time to adjust partially to the changing nuclear potential field (i.e., to near-adiabatic collision conditions). The incompleteness of this adjustment causes mixing of the molecular states and leads to electronic transitions. For an extremely slow collision, the electrons completely adjust to the slowly changing potential field; the electronic state behaves adiabatically, with its energy staying always on a single Born–Oppenheimer energy surface, which corresponds to an eigenstate of the molecular electronic Hamiltonian. In this case, no electronic transition takes place; however, even if no electronic transition occurs in the collision, rovibrational excitation and chemical reactions are possible because of their smaller energy defects (electronic excitation, 1–10 eV; vibrational excitation, 0.01–1 eV; and rotational excitation, <0.01 eV). In this energy region, two difficult tasks must be accomplished to calculate the transition probabilities accurately.

First, the semiclassical approximation discussed in Sec. 2.1.2 is no longer valid because the de Broglie wavelength of the relative motion of the heavy particles (nuclei) is no longer small compared with atomic dimensions; in principle, a fully quantum mechanical treatment for both electronic and nuclear motions is required. Many electronic and rovibrational motions couple strongly, making it extremely difficult to solve rigorously the multistate quantum-mechanical close-coupled equations.

Second, the details of collision dynamics are very sensitive to every aspect of computation and are particularly sensitive to the quality of the adiabatic potential (the energy surface) of the collision system. Thus, highly precise calculation of molecular electronic states as functions of the distance and geometry of the colliding particles is essential for accurate determination of collision dynamics and cross sections. To illustrate the complexity of solving the coupled equations, a formal form of the coupled equations (a simplified version) is given as

$$\left[\frac{1}{2\mu}\frac{d^2}{dR^2} - E - \frac{J(J+1)}{2\mu R^2}\right]\Psi_\alpha(R)$$
$$= \sum_{\alpha'} \langle\alpha|V|\alpha'\rangle\Psi_{\alpha'}(R). \tag{11}$$

In Eq. (11), α designates a set of all quantum numbers at a given state specifying the electronic and rovibrational states of the colliding system. The summation of the coupling matrix element on the right-hand side includes, in principle, all combinations of the other quantum states coupled. The number of states needing to be considered easily exceeds a few hundred when one tries to calculate rigorously, for example, rovibrational excitation cross sections at collision energies of a few electron volts. Therefore, to reduce the number of coupled equations to a numerically tractable level, various types of approximations are used at various levels of computation.

Very powerful quantum-chemistry structure codes developed by various groups enable us to determine potential curves (surfaces) for simpler systems like atom–diatomic-molecular collision systems; however, even with these powerful computer programs, obtaining adiabatic potentials with reasonable accuracy is still exceedingly difficult and time consuming, particularly when excited states must be determined.

Additional difficulty arises when electronic continuum (ionized) states couple strongly with other degrees of motion. A good example of this is Penning ionization (p7), in which the initial excited state is embedded in the final continuum state, and ionization is always possible. High-quality computation of excited molecular states is still far more difficult than that for the ground state because of the strong mixing of many states lying nearby. Rigorous treatment of continuum states is also extremely difficult, if not impossible. A procedure normally proposed is a discretized sampling of continuum energies, followed by treatment of continuum states as if they were discrete; however, this procedure has not been applied extensively to studies of the continuum problem arising from low-energy collision systems.

2.2 Elastic Processes

Low-energy elastic-scattering cross sections (total or differential cross sections) resulting from neutral-particle collisions often show various types of structures, due to interferences or resonances, as functions of energy or scattering angle. Careful analysis of these structures provides detailed information on the interaction potential between colliding particles. In addition, the momentum-transfer cross section, the cross section that accounts for large-angle elastic scattering, is an essential physical quantity for describing diffusion and mobility of a particle through a gas or condensed matter and hence is practically important.

In the classical two-body collision, the classical deflection function X is defined as

$$X(E,b) = \pi - 2b \int_{R_0}^{\infty} \frac{dR}{[1 - V(R)/E - b^2/R^2]R^2}, \tag{12}$$

where b is the impact parameter, E is the collision energy at the center of mass of the system, and $V(R)$ represents the interaction potential. The lower limit of the integrand, R_0, is the classical turning point. $X(E,b)$ can easily be evaluated if the interaction potential $V(R)$ is known. When $V(R)$ is repulsive or attractive, $X > 0$ or $X < 0$, respectively.

The classical differential scattering cross section can be expressed as a function of scattering angle θ,

$$\sigma(\theta) = \frac{2\pi b db}{2\pi \sin\theta d\theta}. \tag{13}$$

By combining Eqs. (12) and (13), this cross section is re-expressed as

$$\sigma(\theta) = \frac{b}{\sin\theta \, |dX/db|}. \tag{14}$$

Several important features in the differential cross section are summarized subsequently.

2.2.1 Glory Scattering

Because the particle feels a balance between the attractive nature and the repulsive nature of the potential $V(R)$, at a certain impact parameter (or scattering angle θ_g), $\sin\theta_g = 0$. For such a circumstance, from Eq. (14), the classical differential cross section diverges and gives an infinite value, called the glory scattering; however, this argument is purely classical, and if the quantum-mechanical treatment is used, the value of the differential cross section remains finite with a maximum, because many partial waves contribute and cause interference.

2.2.2 Rainbow Scattering Because the potential has attractive and repulsive natures, the classical deflection function possesses either a maximum or a minimum. Therefore, at a certain impact parameter, the condition $dX/db = 0$ is met, leading to the divergence of the differential cross section Eq. (14). This peak is called the rainbow scattering. Again, this argument is purely classical; and if the quantum-mechanical treatment is used, the differential cross section remains finite, with merely a shallow dip at the rainbow angle because of contributions of many partial waves.

2.2.3 Orbiting Resonance In the classical picture, when the projectile approaches the target with a certain impact parameter, the projectile may circle the target several times before departing from the interaction region. This phenomenon, the orbiting resonance, is particularly seen when the incident energy equals the magnitude of the effective potential (the sum of the interaction potential and the centrifugal potential). Quantum mechanically, the phenomenon corresponds to resonances because the projectile stays for a longer period of time near the target.

3. EXPERIMENTAL FRAMEWORK

Collision processes are typically examined by performing scattering experiments, which conventionally consist of several common experimental elements, regardless of interaction energy. The collision partners must be prepared in well-defined initial states before the collision occurs. One would like to know the initial momentum of each collision partner as well as the complete quantum state information. After the partners are allowed to interact, some product of the collision process must be detected. A complete measurement of a scattering event includes the final states and the momentum of each exiting particle. In addition, for an absolute measurement, the quantities of initial and final particles must be measured independently during the experiment. In practice, such definition is almost never achieved. For example, absolute results can be obtained by scaling relative data to absolute measurements available in the literature. However, the states and momenta of the collision system are rarely fully measured; consequently, assumptions must be made for comparison with theory or to use experimental results in a predictive model for applied work. Nearly every experiment has substantial shortcomings, and conscientious authors candidly note such limitations.

The desire to restrict the momenta of particles in the collision system under study has led to several common experimental techniques. Atoms and molecules participating in the collision under study are initially widely separated (centimeters to meters); they come together to interact (10^{-9} m) and subsequently separate again. Thus, these experiments take place, in some sense, in the gas phase. Because no other collisions are allowed except the interaction under study, experiments generally require the long mean free paths available in high-vacuum apparatus. Initial particle trajectories are typically restricted to one dominant direction in the apparatus, forming them into a *molecular beam*; such a beam is amenable to restriction of particle speeds in the beam. Thus, a collision partner can be prepared with a well-defined momentum before the collision. Techniques for selecting formation and velocity vary with energy and can influence the initial quantum state of the colliding particles, as discussed subsequently.

A collision event can take place via the interaction of two beams or by placing a cell containing the other collision partner in the path of the beam. In the latter case, the nomenclature *projectile* and *target* is appropriate. In nearly every case, the laboratory reference frame is not the center-of-mass (c.m.) frame, and a transformation to the c.m. frame is required to compare experimental results with theory.

The interaction between particles takes place under carefully controlled conditions. In particular, the process under study must be distinguishable from other competing processes. The initial particle densities or fluxes are kept low so that multiple collisions do not dominate. More will be said about this subsequently.

After the particles separate, some product of the collision must be collected. This collection can be somewhat artificially separated into analysis and detection. The products typically pass through the analyzer, which restricts their properties, and then arrive at the products detector. Analyzers can restrict the quantum state or the momentum of products (or both). A large variety of analyzers exists. Some of the more common analyzers are discussed subsequently. The detector converts the arrival of the analyzed product into an electrical signal, which is conventionally stored as a count in a computer data file. Sometimes the product of interest is a photon, and so optical filters or spectrometers constitute the analyzer, and detectors like photomultipliers or charge-coupled devices are used. More frequently, the product of interest is an atom, molecule, ion, or electron. In some cases, an ion or electron current can conveniently be collected directly from an electrode inside the apparatus. In other cases, the flux is so small that amplifying detectors based on secondary electron ejection and cascade (such as a Channeltron or microchannel plate) must be used. Products that do not carry an electrical charge can be analyzed and detected in several ways. High-velocity neutral particles may transfer enough energy to a detector to generate a secondary electron, which is amplified and detected in the usual way. Neutral products can be ionized for analysis and detection by using electron impact or multiphoton techniques, but the effect of the ionization process (such as dissociative ionization) must be taken into account. Finally, the neutral product can be excited with a probe photon into an upper state that emits upon transition to a lower state, where the emission is spectroscopically detected. This technique generally requires the high photon flux of a laser and is consequently called laser-induced fluorescence. The technique must be used carefully, because competing mechanisms also depopulate the upper state and because the technique relies on published oscillator strengths.

The experimentalist often struggles to produce enough signal to overcome instrumental noise. This problem can be overcome by improving the efficiency of detection or the efficiency of the analyzer (or both) or by increasing the number of collisions occurring. However, the number of collisions cannot be raised indiscriminately, because unwanted multiple collisions will affect the results.

3.1 High to Superthermal Energies (10 eV–1 MeV)

Five orders of magnitude are considered as "high energy" because the experimental techniques for beam generation and detection in this range are similar. Atomic and molecular physics and nuclear physics overlap at the high end of this energy region. Our discussion is limited to electronic interactions between neutral atoms and molecules.

3.1.1 Charge-Transfer Beams Neutral beams with energies above 10 eV are typically generated by passing an ion beam of the desired energy through a cell containing gas. Charge-transfer events at high impact parameters produce fast neutral particles whose trajectories are virtually unchanged. The neutral beam coming from the target cell passes through a downstream aperture to select these coaxial neutral particles. Charge transfer occurs most readily with an identical species; producing neutral rare-gas beams is therefore relatively straightforward. Charge transfer is also favored when the ionization potentials of the target and projectile are similar. For this reason, krypton (ionization potential = 14.0 eV) is used as a target for protons to produce H atom beams.

3.1.2 Sputtered Beams A fast neutral beam can be produced by bombarding solid targets with an intense ion beam of kiloelectronvolt energy. More than 90% of the sputtered particles that leave the target are neutral, falling into a broad energy spectrum ranging from 0.1 to 100 eV. Therefore, a velocity selector is essential for obtaining a monoenergetic beam.

3.1.3 Seeded Supersonic Beams A high-energy beam of neutral atoms and molecules can be obtained by supersonic expansion of a gas. In this method, a small fraction of the desired heavy gas is mixed with light inert carrier gas and expanded through a nozzle of 0.1-mm diameter that is heated to about 1000 °C. The final beam, which has an energy range up to a few tens of electron volts, is produced by skimming and collimation. This technique is particularly useful when it is applied to fragile molecules that would easily be broken up in a high-temperature or violent environment.

3.1.4 Other Methods Another technique for the production of neutral beams from ions involves photodetachment from a negative-ion state. The negative-ion beam intersects an intense pulsed or continuous laser; barring multiphoton effects, the resulting neutral beam is entirely in the ground electronic state. The cross section for photodetachment is low enough, however, that intense laser fields (intracavity Ar^+, for example) are required.

Methods of ion-beam production will not be discussed here, except to note that the characteristics of ion sources (such as electron energy, pressure, size of the ion extraction region, and impurities or mixtures of source gases) do affect the qualities (particularly the energy spread and quantum state) and flux of the extracted beam (see PARTICLE SOURCES, ION). For example, ionized oxygen extracted from a conventional plasma source at kiloelectronvolt energies will be in a mixture of ion states, including the $O^+(^4S)$ ground and the $O^+(^2D)$ metastable excited states. Various measures have been developed for preparing ion beams in a known quantum state or with a low energy spread; the applicability of these techniques to neutral studies is limited to cases where the preparation technique does not adversely affect the neutral beam qualities (such as intensity or divergence).

3.2 Low to Thermal Energies (<10 eV)

Most thermal-energy-beam sources employ a flow system in which the beam material flows as a vapor from an oven into a chamber of lower pressure. In the chamber, collimation and differential pumping are used to produce a final well-defined beam for experimentation (in the collision chamber). Two different regimes characterize the beam flow, depending on the size of the mean free path of atoms or molecules in the source relative to the slit width. For flow in which collisions among atoms or molecules are rare, or molecular effusion, molecular motions are mutually independent; for flow in which atoms or molecules undergo many collisions, or hydrodynamic flow, energy transfer among various internal motions and translational motions occurs frequently.

A molecular-effusion source has an advantage for predicting the beam properties easily and accurately because it operates at thermal equilibrium. The beam leaves a chamber, which contains the beam material as a vapor, via a narrow slit. The gas pressure in the source can be controlled to produce molecular effusion by making the width of the slit smaller than the mean free path of the molecules. The beam thus produced has the Maxwellian velocity distribution, which can control the energy resolution of the beam. For some purposes, some type of velocity selection may be needed.

The hydrodynamic source has an advantage of higher beam intensities than those in molecular-effusion sources. This beam, produced by replacing the thermal equilibrium gas in the molecular-effusion source with a supersonic jet of gas, leaves a chamber of high pressure through a nozzle. Therefore, a large jet of hydrodynamic gas flow can be obtained.

3.3 Beams of Excited Species

The production of beams of excited (electronic or rovibrational) species is also important for applications. The lifetimes of these species vary significantly, from 10^{-1} s for metastable electronic states from which transitions are optically forbidden to less than 10^{-5} s for electronic states that can make optically allowed transitions. Usually, the lifetimes for rovibrational excited states lie between these extremes; however, in general, the techniques are simply a combination of the beam production method already described with a method to produce excited states.

Depending on whether the decay of the desired excited state is optically allowed or forbidden, the specific technique and device to produce the species vary. Thermal dissociation, electron impact, optical excitation, and chemical enhancement are the most commonly used techniques.

3.3.1 Thermal Dissociation Thermal dissociation is one of the simplest techniques for producing atomic beams of excited species with translational kinetic energies of kT in an oven or arc source. The device simply consists of a tungsten tube into which the desired gas is introduced, then heated for dissociation. With this technique, reasonably high-intensity beams of atomic H gas can be obtained.

3.3.2 Electron Impact Electron impact is particularly useful for producing, through electron exchange, optically forbidden metastable states that have a longer lifetime of 10^{-3} s. Negative ions formed temporarily are also known to decay to specific rovibrational states of a molecule; hence, negative-ion resonance can be used selectively for this purpose. As alternative techniques with similar electron scattering properties, radio frequency and microwave discharge are commonly used.

3.3.3 Optical Excitation Optical excitation is essentially the same as electron impact, but a photon is used as the exciting source. This method is suited for production of optically allowed excited states. However, its usefulness depends on absorption intensity.

3.3.4 Chemical Enhancement By controlling, with the help of a laser, a chemical process in which the specific reaction cross section is known, chemical enhancement is useful for production of beams with state-selected quantum states (vibrational and rotational excited states) of molecules.

All methods for beam production discussed previously are also useful as cluster beam sources.

Detection systems for beams of excited species are essentially a combination of those described previously.

4. OBSERVABLES RELEVANT TO APPLICATIONS

Various physical observables that can be determined rigorously, through experimental and theoretical work, are useful for applications. Some of the most fundamental quantities are summarized here.

4.1 Differential and Total Cross Sections and Rate Constants

The differential cross section, usually determined in an experiment, is defined as the number of particles scattered at the scattering center in direction (θ,ϕ), per unit solid angle, per unit incident flux. The total intensity of scattering, the total cross section, defined by integrating the differential cross section over all solid angles, has a dimension of area or $(\text{length})^2$. Since the incident particles normally have a distribution of velocity, $f(v)$, the rate constant $k(v)$ may be defined as

$$k(v) = \int_0^\infty dv\, vf(v)\, \sigma(v). \tag{15}$$

4.2 Stopping Power

When energetic particles traverse a material, they lose their kinetic energies in a series of collisions with constituent atoms and molecules in the material. The average rate of energy loss in a collision that causes a transition from i to j is an important measure characterizing energetic particles and is defined as

$$-\left.\frac{dE}{dt}\right|_{i\to j} = v_A \Delta E \sigma_{i\to j} n_B \equiv v_A n_B S_{ij}. \tag{16}$$

Here, v_A is the velocity of the incident particle A, n_B is the density of B atoms or molecules in the material, and ΔE is the average energy loss of the incident particle A for the process $i \to j$ with cross section $\sigma_{i\to j}$. The quantity S_{ij}, sometimes referred to as the stopping cross section, has units of (area × energy). By summing over all possible processes, one can obtain the energy loss per unit of path length (dx) or the stopping power:

$$-\frac{dE}{dx} = n_B \sum_j \Delta E_j \sigma_{i\to j}. \tag{17}$$

The stopping power is directly measurable experimentally.

4.3 Diffusion Constant and Mobility

As a slow neutral particle or ion moves through a gas or condensed matter under the influence of a static, uniform electric field or as a result of density or temperature gradients that act as thermodynamic driving forces, the particle or ion gains energy from the field and gains or loses energy through collisions, finally attaining a steady-state condition. In a steady state, the particle flux is proportional to the applied field $\mathscr{E}$ (external field or density/temperature gradient) with a multiplication constant K. The constant K is called the mobility (or diffusion constant) of the atoms and is usually expressed in units of cm^2/$\mathscr{E}$-s.

5. APPLICATION AND RELEVANCE

Knowledge of the spectroscopy, dynamics, and production of specific species in collision processes involving neutral atoms and molecules is important because of its intrinsic scientific interest; its significance in other fields of science; and its application in technological and engineering areas. Here we summarize representative examples of scientific and technological applications of this knowledge.

5.1 Scientific Applications

5.1.1 Dynamics: The Determination of Interaction Potential Measurement of the differential cross section as a function of scattering angle at a given collision energy provides information on the nature of the interaction between two particles. For simpler, but difficult, systems like H + H and H + He, a rigorous quantum-mechanical calculation for elastic and inelastic processes can easily be performed. The result thus obtained can be compared readily with the measurement as a test of the precision of the interaction potential used. Repeating the calculation with refined potential parameters allows one to achieve better agreement with the measurement and, hence, better understanding of the electronic structure of the system and its collision dynamics. For example, an experimental differential cross section for the small-angle elastic collision of He atoms with He atoms at 0.5 keV is shown in Fig. 3(a), along with the results of quantum-mechanical calculations using different types of interaction potentials. Accurate determination of the interaction potential of the He-He system is crucial in calculations of bulk properties of liquid and solid He and transport properties of He vapor in conjunction with superfluidity and Bose–Einstein condensation. The potentials used in the calculation are shown in Fig. 3(c). The experimental data agree well with calculations using three types of potentials in the scattering angle range below 0.1°, suggesting that all three potentials at large R (the long-range attractive part) are reasonable; however, at angles larger than 0.1°, the experi-

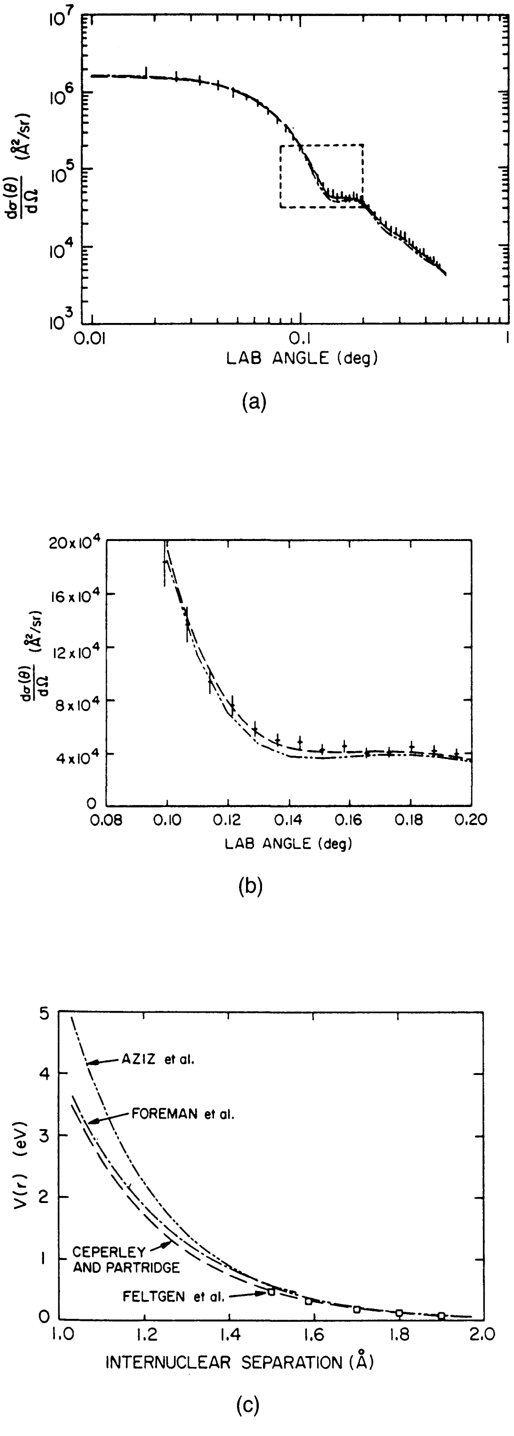

FIG. 3. (a) Measurements and theoretical calculations of differential elastic cross sections in collisions of He atoms with He atoms (from Nitz *et al.*, 1987). (b) Enlarged portion at $0.08 < \theta < 0.2$ of the differential elastic cross section shown in (a). (c) Proposed potential curves of the He-He ground state (from Nitz *et al.*, 1987).

mental data compare more favorably to the results of the calculation using the least steep portion of the potential at $R < 1.8$ Å.

5.1.2 Spectroscopy: The Determination of Physical Constants and Identification Probing molecular beams, coupled with photons that considerably increase spectral resolution over a wide energy range, are a central feature of modern spectroscopy. The interactions between atoms or molecules and a radiation field can be analyzed to reveal a variety of physical properties that characterize atomic or molecular structures and electronic states with higher precision, permitting higher efficiency and optimization in the design of experimental and engineering apparatus. Magnetic resonance, electron resonance, infrared laser, microwave, infrared, ultraviolet, and photofragment spectroscopies have particularly benefitted.

5.2 Technological Applications

5.2.1 Quantum Electronics: The Gas Laser Taking the He-Cd$^+$ ion laser as an example, we describe here the basic principle of the laser and its relationship with collisions of neutral atoms. The Cd$^+$ ions excited during collisions with He atoms decay to lower states, emitting photons that are known to be an efficient laser source. The fundamental goal for this laser is to maintain a large fraction of excited Cd^{+*} states so that collision-induced photon emission occurs continuously (inversion of population). To this end, two processes are known to be efficient. One is based on charge transfer by the ion (p2b),

$$\mathrm{He}^+ + \mathrm{Cd}(5s^2\ {}^1S_0) \rightarrow \mathrm{He} + \mathrm{Cd}^{+*}(6f\,{}^2F, 6g\,{}^2G). \tag{18}$$

The excited Cd^{+*} ion decays to a lower level by emitting a photon ranging in wavelength from 6360 to 7237 Å. The second process is based on Penning ionization (p7):

$$\begin{aligned}&\mathrm{He}^*(2^3S) + \mathrm{Cd}(5s^2\ {}^1S_0)\\ &\quad\rightarrow \mathrm{He}(1^1S) + \mathrm{Cd}^{+*}(4d^9 5s^2\ {}^2D) + 1e^-. \end{aligned}\tag{19}$$

Here, the excited Cd$^{+*}(4d^9 5s^2)$ ions emit laser photons of 3250 and 4416 Å. As this example shows, knowledge of the various inelastic processes in the atomic collision is essential for developing an efficient laser.

5.2.2 Plasma and Fusion Research Neutral-particle beams are frequently used for plasma diagnosis. Injection of neutral particles into a plasma or fusion reactor and study of the density distribution, charge distribution, and emitted photons or the scattering angles of the neutral-particle flux exiting from the plasma can provide information on the plasma's temperature and density and various parameters that characterize the spatial and time distribution of the plasma. In addition, careful analysis of this information may lead to an understanding of the distribution of an impurity and its mean charge distribution in the plasma environment.

5.2.3 Surface Sciences Neutral-atom scattering from surfaces has become a vitally important technique for the study of surfaces. In particular, He scattering has several advantages because it is a highly surface-sensitive, nondestructive probe. Unlike what happens in electron or heavy-ion scattering, low-energy (millielectronvolt) He atoms colliding with a surface do not penetrate deep into the bulk. Thus, any information taken away by the scattered He atom can be related to the nature of the outermost surface layer of the bulk solid, whether the information concerns structure or dynamics. Furthermore, analysis of the spectrum is easier because the He beam is inert and nonreactive and has no internal degrees of freedom except large electronic excitation energies. Hence, the surface structure or dynamics of virtually any material, in the form of a static state, under the influence of an external field, or during surface processing, can be studied without altering the material.

Structural and dynamic phenomena are directly accessible to He scattering. By measuring elastic scattering, we can reconstruct the map of the electronic charge density of the atoms on the surface. In inelastic He scattering, phonons, surface vibrations, and even adsorbate vibrations leave their signatures in the scattered-He spectrum. As a representative example, a He scattering intensity spectrum from a Pt(111) surface with and without a low density of randomly adsorbed CO molecules is illustrated in Fig. 4(a). This experiment provides information on the density and location of adsorbed CO molecules. This type of experiment also gives a useful

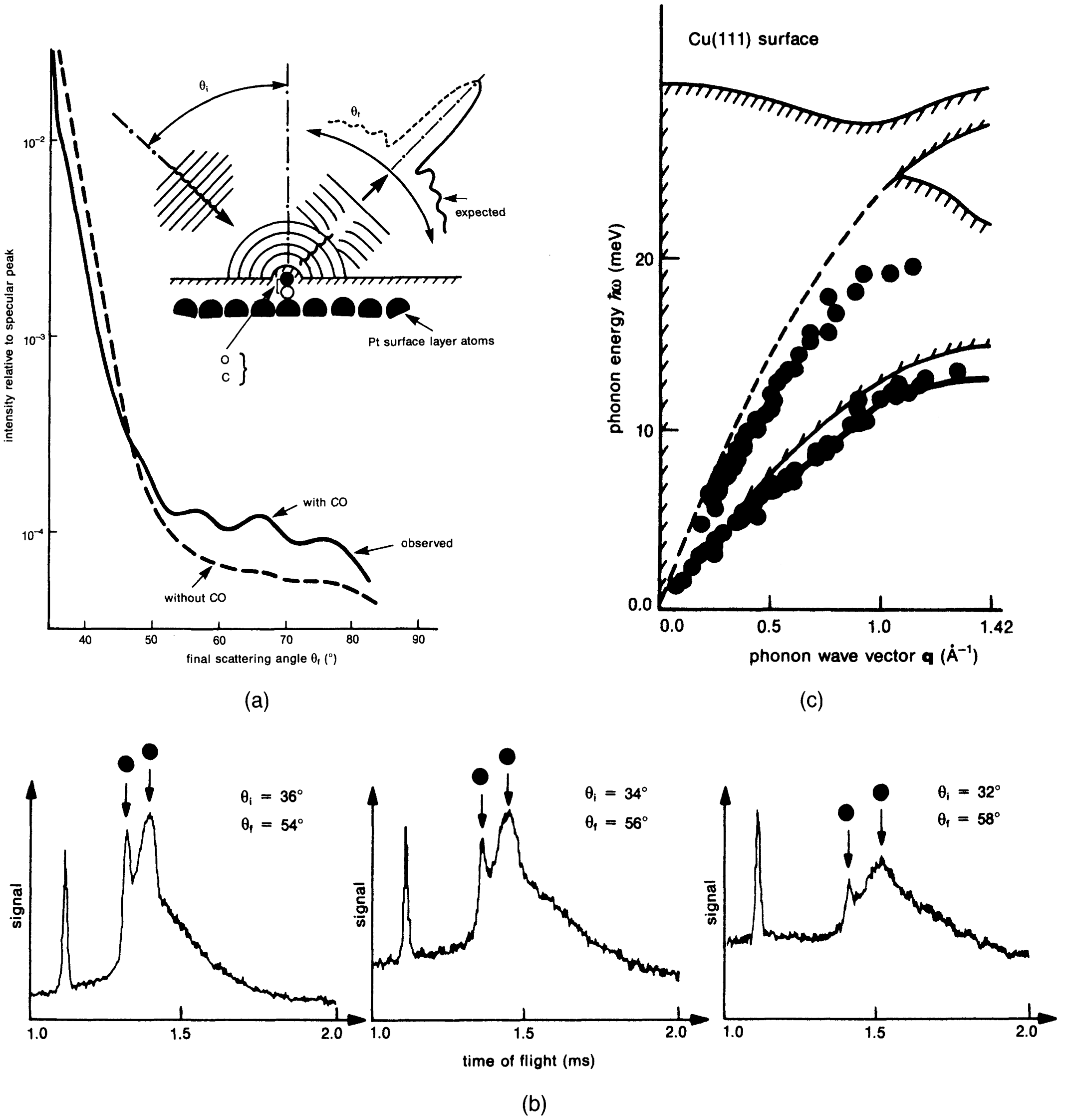

FIG. 4. (a) Helium scattering intensity spectrum from Pt(111) surface (from Lahee and Toennies, 1993). (b) Time-of-flight spectra for He atoms scattered from Cu(111) surface (from Lahee and Toennies). (c) Surface phonon dispersion relation curves (from Lahee and Toennies).

method for analysis of the structures and densities of defects. Figure 4(b) displays a series of time-of-flight energy-loss spectra for He atoms scattered from a Cu(111) surface. Each peak shows the signal of energy loss of the He atom. The relative shape of the peaks represents the signature of single-phonon creation. From these energy-loss spectra, the dispersion curves can be constructed, as shown in Fig. 4(c). The dispersion curves contain much useful information on the forces binding atoms at the surface.

5.2.4 Energy *Combustion.* Combustion is essentially a gas-phase reaction. Detailed kinetic knowledge is required to understand the reaction process, and this knowledge is an important ingredient in models to optimize fuel efficiency, for example. Knowledge needed includes the identities and concentrations of combustion species and unwanted products (pollutants) in both ground and excited states, the reaction dynamics, and the reaction cross section or rate coefficients.

Magnetohydrodynamic electrical power. The essential idea in magnetohydrodynamics (MHD) is to extract an electrical current from a hot gas without a mechanical device. A hot gas, produced, for example, by burning a fossil fuel, is exhausted, after ionization, through a magnetic field. The interactions of the magnetic field and the flow velocity create an electric field transverse to the flow direction, producing an electric current across the field that can be extracted for commercial use. One of the major advantages of MHD electrical power is its increased efficiency over conventional mechanical generators. Penning ionization (p7) is known to play an essential role in the physics of this process.

6. CONCLUSION

We have provided the information necessary for understanding a variety of phenomena arising from the application of neutral atomic and molecular processes in experimental and theoretical instrumentation. Recent advances in technologies, such as the production of high-power, high-precision lasers and high-intensity beams, have significantly improved our basic understanding of chemical processes. We are now in an era of femtochemistry, in which an experiment can directly monitor bond breaking and bond formation during a chemical reaction in units of femtoseconds (10^{-15} to 10^{-12} s). We can witness every movement of the atoms and molecules in the chemical process at any instant. These refinements in experimental information on chemical processes demand more refined and precise theoretical treatments to match the experimental advances. With the recent development of the scanning tunneling microscope, nanometer technology or angstrom technology has emerged and is becoming the standard in high-tech industries such as the semiconductor industries. Combining femtochemistry with nanometer technology is the ultimate goal in the application of chemical processes at atomic and molecular levels within the foreseeable future.

ACKNOWLEDGMENT

This work was supported by the U.S. Department of Energy, Office of Energy Research, Office of Health and Environmental Research, under Contract No. W-31-109-Eng-38.

GLOSSARY

Born Approximation: One of the perturbation theories that is frequently used for solving quantum-mechanical scattering problems. Interaction potentials between a projectile and target are often chosen as perturbations to cause a transition in the system. When the interaction is of the lowest order, it is termed the first Born approximation.

Born–Oppenheimer Approximation: An approximate method to solve the Schrödinger equation for a molecular system on the basis that large differences in masses of electrons and nuclei leads to the separation of the electronic wave function and the nuclear wave function.

Charge-Transfer Beam: A neutral-particle beam produced by effectively using the charge transfer of an ion beam through a neutral gas.

Close-Coupling Method: A theoretical approach for treating many processes simultaneously and equally. Usually, the coupled equations are solved numerically to extract scattering amplitudes or the scattering *S*-matrix.

Momentum-Transfer Cross Section: Cross section for a collision with momentum transfer but no change in the internal state of a target. Momentum transfer occurs because momentum is conserved in the system. The momentum-transfer cross section σ_{MT} for scattered angle θ is defined as

$$\sigma_{\mathrm{MT}} = 2\pi \int d\theta \sin\theta(1 - \cos\theta)\frac{d\sigma_{\mathrm{el}}}{d\Omega}, \qquad (20)$$

where $d\sigma_{\mathrm{el}}/d\Omega$ is the differential elastic cross section per unit of solid angle.

Penning Ionization: In collisions between an excited species and a ground-state species, ionization by interaction may be possible if the excitation energy of the excited species exceeds the ionization energy of the target. This process is called Penning ionization. A good example is $\mathrm{He}^*(2^2P) + \mathrm{Ar} \rightarrow \mathrm{He} + \mathrm{Ar}^+ + 1e$, where the excitation energy of $\mathrm{He}(2^2P)$ is 21 eV, and the ionization potential of Ar is 16 eV.

Perturbation Theory: In quantum mechanics, the total Hamiltonian is divided into two parts. The Schrödinger equation corresponding to one part can be solved exactly for the energy and the wave function (the unperturbed wave function). The other part corresponds to a correction that is approximately calculated by using the unperturbed wave function. The results from the unperturbed and perturbed portions are added to give an approximate total energy of the system.

Seeded Supersonic Beam: A neutral-particle beam obtained by using supersonic expansion of a gas.

Sputtered Beam: A fast neutral-particle beam produced by bombarding solid targets with energetic ions.

Stopping Power: An experimentally measurable quantity that describes the rate of average energy loss of a particle passing through a material.

Works Cited

Lahee, A., Toennies, P. (1993), *Phys. World* **6** (4), 61–66.

Nitz, D. E., Gao, R. S., Johnson, L. K., Smith, K. A., Stebbings, R. F. (1987), *Phys. Rev. A* **35**, 4541–4547.

Phelps, A. V. (1990), *J. Phys. Chem. Ref. Data* **19**, 653–675.

Further Reading

Fluendy, M. A. D., Lawley, K. P. (1973), *Chemical Applications of Molecular Beam Scattering*, London: Chapman and Hall.

Kimura, M., Lane, N. F. (1989), in: Bates, D., Bederson, B. (Eds.), *Advances in Atomic, Molecular and Optical Physics*, Vol. 26, New York: Academic Press, p. 79.

Lahee, A., Toennies, P. (1993), *Physics World* **6** (4), 61–66.

Massey, H. S. W., McDaniel, E. W., Bederson, B. (Eds.), (1982), *Applied Atomic Collision Physics*, Vols. 1–5, New York: Academic Press.

Nitz, D. E., Gao, R. S., Johnson, L. K., Smith, K. A., Stebbings, R. F. (1987), *Phys. Rev. A* **35**, 4541–4547.

Phelps, A. V. (1990), *J. Phys. Chem. Ref. Data* **19**, 653–675.

Prigogine, I., Rice, S. A. (Eds.), (1975, 1992), *Advances in Chemical Physics*, Vols. 30 and 83, New York: John Wiley & Sons.

Scoles, G. (Ed.), (1988), *Atomic and Molecular Beam Methods*, Vols. 1 and 2, Oxford, UK: Oxford University Press.

NEUTRON DIFFRACTION

THOMAS VOGT, *Physics Department, Brookhaven National Laboratory, Upton, New York, U.S.A.*

	Introduction	339
1.	**Scattering Theory**	341
2.	**Neutron Diffraction**	342
2.1	Medium-Range Order in Glasses	343
2.2	Neutron Powder Diffraction	344
2.2.1	*Ab Initio* Structure Determination	344
2.2.2	The Rietveld Method	346
2.2.3	The Analysis of Mesoscopic Structures using High-Resolution Powder Diffraction	346
2.3	Single-Crystal Neutron Diffraction	348
2.3.1	Two-Dimensional Position-Sensitive Detectors	348
2.4	Neutron Reflectometry	349
	Works Cited	352
	Further Reading	352

INTRODUCTION

Neutrons constitute a very important probe for condensed matter research. This is a consequence of their unique properties. The neutron is an uncharged particle. To detect neutrons one must therefore rely on nuclear reactions producing charged particles. The nuclear reaction on which the most common detector for thermal neutrons, the ^{3}He gas detector, is based is

$$n + {}^3\mathrm{He} \rightarrow {}^3\mathrm{H} + {}^1\mathrm{H} + 0.764\ \mathrm{MeV}. \qquad (1)$$

The proton (^{1}H) and triton (^{3}H) move away in opposite directions from the neutron absorption site and collide with other gas molecules, ionizing them (see DETECTORS, PARTICLE, TRACKING). The neutron has a mass of 1.0087 u. As for any elementary particle, the wave-particle formalism using de Broglie's equation $\lambda = h/mv$ allows the energy to be expressed as

$$\begin{aligned} E(\mathrm{meV}) &= 0.08617\,T\,(\mathrm{K}) = 5.227\,v^2\,(\mathrm{km\ s}^{-1}) \\ &= 81.81/\lambda^2\,(\mathrm{\AA}). \end{aligned} \qquad (2)$$

Its mass allows the neutron to be slowed down after its creation in a fission or spallation process. This so-called moderation process occurs mostly via inelastic neutron-proton collisions. The neutrons adapt to the temperature of the hydrogen-containing moderator. In a research reactor, this moderation process leads to a distribution of neutron wavelengths between 1 and 3 Å (for 300 K, λ = 1.78 Å) and energies between 80 and 10 meV. These values correspond to typical interatomic distances and energy values of thermal excitations (phonons, magnons) in condensed matter. Inelastic neutron scattering first made the phonon "visible" and provided various crucial experiments exploring the dynamical nature of condensed matter (see NEUTRON SCATTERING).

Here we shall focus on structural investigations using neutron diffraction. Neutrons do not primarily interact with the electrons of matter but instead with the nucleus. The nuclear forces have a very short range, on the order of a few femtometers (1 fm = 10^{-15} m).

Thermal neutrons "see" very little when passing through matter, since the size of the scattering centers is about 10^5 times smaller than the distances between them. An important consequence of this is that the nucleus can be effectively considered as a point scatterer and the neutron is therefore scattered isotropically. The differential cross section describing the interaction of neutrons with the nucleus is a constant, $\delta\sigma/\delta\Omega = b^2$. In contrast to this, x rays are not scattered isotropically since the size of the "electron cloud" is comparable to the size of the x-ray wavelengths typically used in a diffraction experiment. This is why one observes a decrease

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

of the interaction between x rays and electrons at high scattering angles, and it is difficult to measure intensities at very high scattering angles (see X-RAY DIFFRACTION).

The constant *b* is the so-called scattering length and is a property of the nucleus with a given atomic number *Z* and weight *A* as well as the spin state relative to that of the neutron. The interaction of neutrons with the nucleus allows negative scattering lengths corresponding to a phase change of π during the scattering process.

Two main fields of applications in scattering experiments arise from the above:

1. The low-*Z* ("lighter") elements may scatter more strongly than the "heavier" ones. Neighboring elements may vary quite a bit in their scattering lengths. To distinguish, for instance, between Mn and Fe, which have one electron difference, using x rays is virtually impossible without measuring at an absorption edge. With neutrons, however, there is a big difference in their coherent scattering length ($b_{\mathrm{Mn}} = -3.73$ fm, $b_{\mathrm{Fe}} = 9.45$ fm) as shown in Fig. 1. Neutron diffraction is an indispensable tool to determine cation distributions in ceramics and minerals containing elements with little difference in *Z*.
2. Isotopes of the same atom (same *Z* but different atomic weight *A*) may also vary dramatically in their scattering power and might even have scattering lengths with opposite signs, as in the case of hydrogen (−3.74 fm) and deuterium (6.674 fm) or the isotopes ^{62}Ni (−8.7 fm) and ^{61}Ni (7.60 fm) or ^{48}Ti (−6.08 fm) and ^{50}Ti (6.18 fm). Experiments exploiting this by changing the scattering length of a particular element via isotopic substitution are called contrast variation experiments and are extensively used in structural investigations of a wide variety of matter, including liquids, metals, polymers, and glasses. In biological applications, the variation of the H_2O/D_2O ratio changes the "visibility" of water in a diffraction experiment and might give important clues to its location, for instance in proteins.

When low-energy ("cold") neutrons scatter at small angles from biological samples in solution, the measured intensity depends on the "contrast" between the molecules and the solvent; that is, the difference between their scattering-length densities. If one is dealing with a multicomponent solution, as one often does in biology, one can match the scattering-length density of the solvent by adjusting the ratio $H_2O:D_2O$ and thus eliminate the scattering contribution of that par-

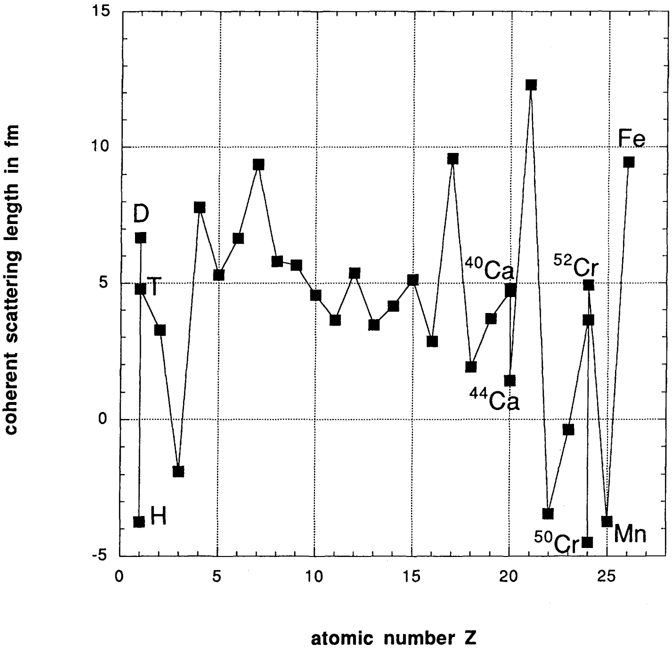

FIG. 1. The variation of the coherent scattering length in femtometers for the elements $Z = 1$ (H, D, T) to $Z = 26$ (Fe). For the cases of $Z = 1$, 20, 24 the coherent scattering lengths of the major isotopes are shown. For further details see text.

ticular component. This effect can also be used when studying metal or polymer alloys, liquids, and glasses by varying the isotopic ratio of a particular element that has isotopes with opposite scattering lengths (e.g., Ni, Ti). See Jacrot (1976) for more details.

Since the interaction of neutrons with the nuclei of matter occurs within a much smaller area than the one between x rays and electrons, and since there is no Coulomb barrier to overcome because the neutron is uncharged, the penetration depth of neutrons is much bigger than that of x rays. Neutrons are a true bulk probe. This has the advantage that sophisticated sample environments such as pressure cells, furnaces, cryostats, and magnets can be more easily penetrated and, provided good collimation, their scattering will not mask the actual measurement. *In situ* experiments are generally more cumbersome to do with x rays.

On the other hand, the typical flux of a monochromated neutron beam is in the order of 10^6 to 10^8 neutrons $\mathrm{cm}^{-2}\ \mathrm{s}^{-1}$. This is comparable to the flux of molecules in a vacuum and explains why in neutron scattering experiments—especially when measuring weak signals as in inelastic scattering—typical sample volumes of about 1 cm^3 have to be used to achieve reasonable counting rates.

One property that really put neutrons in the spotlight of attention about 50 years ago was the fact that the neutron has a magnetic moment of $1.9132\mu_N$, which interacts via a dipole-dipole interaction with the magnetic moments of unpaired electrons. The first solid-state magnetic structure investigation using neutrons was performed by Shull and Smart (1949). They showed that at low temperature the MnO diffraction pattern reveals extra Bragg peaks ("magnetic" Bragg peaks) because the magnetic moments of the manganese are aligned along the edges of the unit cell pointing in opposite directions. This leads to a magnetic cell twice as big as the chemical cell. This was the experimental proof of antiferromagnetism as predicted by Néel. It is safe to say that the scientific contributions that neutron scattering provided in the field of magnetism are of paramount importance and will continue to be. Neutron diffraction is for the moment the only method to determine magnetic structures of bulk samples with atomic resolution.

The magnitudes of magnetic and nuclear scattering are comparable. However, because of the spatial dimension of the unpaired electron density, the scattering process is no longer isotropic, and scattering-angle–dependent magnetic form factors describing the strength of the scattering as a function of the scattering angle exist similar to the ones observed in x-ray scattering. This form factor is the Fourier transform of the unpaired electron density. For more details, see MAGNETIC ORDERING IN SOLIDS.

The neutron has a spin angular momentum of $\frac{1}{2}$. One is able to polarize, that is, separate, neutrons into beams with spins of $\pm\frac{1}{2} \times h/2\pi$. One can polarize neutrons using Bragg reflections if the nuclear and magnetic contributions to the scattering amplitude compensate exactly for one of the two spin states. This is achieved, for instance, when using the (111) reflection of Heusler alloys such as $\mathrm{Cu_2MnAl}$ and CoFe.

Since only the perpendicular component of the magnetic moment with respect to the scattering vector contributes to magnetic scattering, nuclear and magnetic scattering contributions can be separated (see NEUTRON SCATTERING).

1. SCATTERING THEORY

Independent of the specific type of interaction a probe has with matter (electrostatic for electrons, electromagnetic for x rays, and nuclear for neutrons), a similar formalism is used for scattering phenomena, upon which we will rely in the following (see Fig. 2).

There are two principal types of scattering: elastic scattering, where the probing particle is deflected without energy loss or gain during the scattering process, and inelastic scattering, where the probing particle loses or gains energy. The energy gain or loss is called the energy transfer (for more details see NEUTRON SCATTERING). In both cases, the particle is scattered by the scattering angle 2θ. Scattering processes are often depicted using scattering triangles where a scattering vector $\mathbf{Q}$ is defined by the vector relationship $\mathbf{Q} = \mathbf{k} - \mathbf{k}'$ as shown in Fig. 2. Here $\mathbf{k}$ and $\mathbf{k}'$ are the vectors representing the incident and diffracted spherical wave fronts of the neutron. As can easily be demonstrated by simple trigonometric arguments, in the elastic case $\mathbf{Q}$ equals $(4\pi \sin\theta)/\lambda$. The scattering vec-

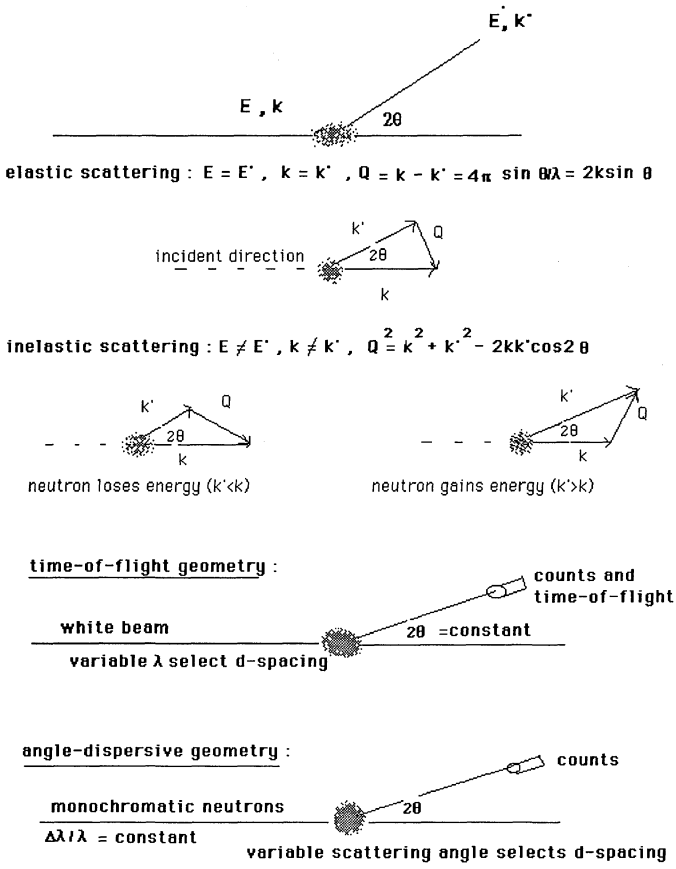
E', k'
E , k
2θ
elastic scattering : E = E' , k = k' , Q = k - k' = 4π sin θ/λ = 2ksin θ
k'
Q
incident direction
2θ
k
inelastic scattering : E ≠ E' , k ≠ k' , Q² = k² + k'² - 2kk'cos2θ
k'
Q
2θ
k
neutron loses energy (k'<k)
k'
2θ
Q
k
neutron gains energy (k'>k)
time-of-flight geometry :
counts and time-of-flight
white beam
2θ =constant
variable λ select d-spacing
angle-dispersive geometry :
counts
monochromatic neutrons
2θ
Δλ/λ = constant
variable scattering angle selects d-spacing

FIG. 2. Basic scattering terms used in the text. For more details see text.

tor multiplied by $h/2\pi$ is referred to as the momentum transfer.

If the probing particles interact with the scattering centers in such a way that the scattered waves have a phase relationship with each other, one speaks of coherent scattering and interference is possible. Diffraction or Bragg scattering is the simplest form of coherent scattering. If the probing particles interact in an independent, not phase-related manner with the different scattering centers, one refers to this as incoherent scattering. Here no interference is possible. The intensities from the different scattering centers simply add up. In the case of neutrons this is very important, since one can obtain information about a specific scattering species at different positions and times (see NEUTRON SCATTERING). In diffraction experiments the background can be dominated by incoherent scattering, often due to the presence of hydrogen in the material, and when performing, e.g., powder diffraction experiments, where the Bragg reflections are much weaker than those of single crystals, the samples containing hydrogen will often have to be deuterated to improve the signal-to-noise ratio.

2. NEUTRON DIFFRACTION

The basis of all diffraction is Bragg's law (see CRYSTALLOGRAPHY):

$$\lambda = 2d \sin\theta. \tag{3}$$

There are two principal ways of doing diffraction experiments:

1. If one measures with a monochromatic beam obtained by the use of an appropriate single-crystal monochromator (see MONOCHROMATORS), thus leaving the selected narrow wavelength band $\Delta\lambda/\lambda$ constant and varying the scattering angle θ,

one is using a so called angle-dispersive or single-wavelength setup.

2. If, on the other hand, one leaves the scattering angle θ constant using a beam with a broad distribution between $\lambda_{\min}$ and $\lambda_{\max}$, one is measuring in an energy-dispersive setup. When using neutrons this is also referred to as the time-of-flight technique.

The angle-dispersive setup is mostly used at steady-state research reactors and x-ray synchrotrons.

Differentiating Bragg's law one obtains

$$\Delta\lambda/\lambda = \Delta d/d + \Delta\theta \cot\theta. \tag{4}$$

For both energy and time-of-flight setup, the resolution is given by

$$\Delta d/d = \Delta\theta \cot\theta. \tag{5}$$

To obtain a good resolution (small $\Delta d/d$) a high scattering angle has to be chosen. Neglecting collimation, the intensity depends on the wavelength band reflected off the monochromator:

$$\Delta\lambda/\lambda = \Delta\theta_M \cot\theta_M. \tag{6}$$

The quantity $\Delta\theta_M$ is the mosaic spread of the monochromator crystal reflecting at a monochromator scattering angle $2\theta_M$, referred to as the takeoff angle (for more see MONOCHROMATORS).

The advantage of the time-of-flight setup is that one can choose the scattering angle so that little or no parasitic scattering from the sample environment masks the pattern of the sample. This is very important, for instance, in high-pressure work: in the angle-dispersive technique one has to use a very small beam to reduce the background scattering from the pressure cell. In the time-of-flight technique one can position the detector at a scattering angle of 90°, where one does not detect any Bragg reflections originating from the pressure cell. This does, however, lower the resolution and limit the **Q** range. Most of the high-pressure work using neutrons these days is done using neutrons created in spallation sources.

In the time-of-flight technique, the time it takes the neutrons to travel from the source to the detector ("time of flight") allows the separation of the Bragg peaks corresponding to their different d spacings. If L is the flight path, ΔL its uncertainty (since the neutrons are generated in a finite moderator slice and detected in a finite detector), t the time of flight, Δt its uncertainty, and θ the scattering angle with its uncertainty $\Delta\theta$, then the resolution of a time-of-flight powder diffractometer is given by

$$\Delta\mathbf{Q}/\mathbf{Q} = [(\Delta t/t)^2 + (\Delta L/L)^2 + (\cot\theta\Delta\theta)^2]^{1/2}, \tag{7}$$

and the flight time t can be related to the flight path L, the d spacing d_{hkl}, and the wavelength λ as follows:

$$t = 505.555\, L\,[\mathrm{m}]\, d_{hkl}\lambda\,[\text{Å}]\, \sin\theta_0. \tag{8}$$

This explains why a long flight path L and a high scattering angle (near 180°, the so-called backscattering geometry) are advantageous. The flight path on the high-resolution powder diffractometer at the ISIS neutron spallation source in the United Kingdom, for instance, is 96 m. If one were to use only an array of detectors covering a small angular range close to 180°, the largest d spacing that could be observed using a neutron pulse with a wavelength distribution $\lambda_1 < \lambda < \lambda_2$ is $\lambda_2/2$, whereas at $2\theta = 90°$ a larger d spacing of $\lambda_2/2^{1/2}$ can be detected. Since the observation of large d spacings is crucial for the determination and refinement of magnetic structures as well as for the indexing of unknown phases during the course of *ab initio* structure determinations, constant-wavelength diffractometers offer an advantage for this kind of investigations. On the other hand, constant-wavelength instruments are not able to provide intensities in such a wide range of low d spacings. This is important for the accurate determination of atomic displacement parameters, strain tensors, and particle size distributions, provided all wavelength-dependent corrections are taken into account appropriately.

2.1 Medium-Range Order in Glasses

Silicate glasses were assumed to be built of silicate frameworks consisting of SiO_4 tetrahedra, in between which the cations were randomly distributed as network modifiers. Neutrons are of tremendous use when investigating structural features of glasses, liquids, and dense gases since one can obtain intensities from high scattering angles and separate the various partial structure factors (see CRYSTALLOGRAPHY) of a multicomponent system by measurements using different isotopic compositions.

Neutron scattering experiments using three specimens of a calcium silicate glass with different isotopic concentrations of calcium revealed that nonrandom correlations beyond the nearest-neighbor distances exist among the cations. (See Fig. 1 for the different scattering lengths of the Ca isotopes.) Gaskell *et al.* (1991) provided the first experimental evidence of medium-range order in the cation distribution of this calcium silicate glass and sparked a series of other investigations. By measuring the differences between the scattered intensities of the glass containing the natural isotopic mixture of calcium and one with equal amounts of natural Ca and ^{44}Ca and the difference between a glass containing a mixture enriched with ^{44}Ca and pure ^{44}Ca one removes all terms contributing to the total structure factor other than the Ca-centered correlations such as Ca-Si, Ca-O, and Ca-Ca. Then, by taking the difference of these differences, all terms except the Ca-Ca correlations cancel. Even though this double-difference signal is very weak and subject to systematic and random errors, the authors clearly observed a second neighbor distance near 6.4 Å. This is consistent with a calcium ordering scheme in which each calcium is located at the center of an octahedron of oxygens. They share edges and create sheetlike regions within the silicate matrix. This arrangement resembles the connectivity in the mineral wollastonite. If all the sheets were connected smoothly beyond the medium-range order, this crystalline structure would be realized. This work has led to a substantial change of our view on glasses: the cations, which have been previously regarded as network discontinuities, are actually ordering elements, and the long-range disorder in glasses coexists with short- and medium-range order. Glasses and their corresponding crystalline structures may have more in common than previously thought.

2.2 Neutron Powder Diffraction

The recent discovery of high-temperature superconducting ceramics and the unraveling of their structural details demonstrated the indisputable utility of neutron powder diffraction. Structure solution and refinement are no longer the exclusive domains of single-crystal diffraction. Powder diffraction is the only tool to obtain structural information of many samples like catalytically active zeolites, and low-temperature and high-temperature phases where single crystals are not available. At the same time, information about the mesoscopic properties (e.g., texture, particle size, stacking faults) of the sample can be obtained. Modern powder methods have developed from a qualitative method previously used for phase identification to a quantitative method to detect phases and determine their volume fraction and atomic and mesoscopic structure.

2.2.1 *Ab Initio* Structure Determination

This area has created a lot of excitement in recent years, since historically it was the exclusive domain of single-crystal diffraction. A powder diffraction experiment provides two types of information: the position of the Bragg peaks, which is determined by the unit cell size and symmetry, and the intensity, which depends on the position of the atoms within the unit cell. Once a structural model is established (for further details see CRYSTALLOGRAPHY) this model will have to be adjusted by refining its crystallographic parameters (e.g., positions of the atoms within the unit cell) with respect to the measured quantities.

The availability of highly collimated monochromatic neutron and synchrotron powder diffractometers made it possible to obtain good enough data to tackle the task of structure determination from powder diffraction data. The first goal in that process is always the determination of the unit cell. Here, many times electron diffraction can provide help. After determining the unit cell and narrowing down the possible space groups, the deconvolution of the individual integrated intensities is the next step. This is referred to as pattern matching or whole pattern fitting and does not require a structural model. Given a space group, the diffraction pattern is refined using the individual intensities of the Bragg reflections, the cell constants, which determine their position, and the so-called machine parameters of the diffractometer describing the resolution (*UVW*, peak-shape parameters, etc.) as variables. The obtained intensities can then be used to obtain a structural model via direct methods or Patterson procedures (see CRYSTALLOGRAPHY for more details). The actual refinement is then usually done with the Rietveld technique once a structural model has been established. If one has a reasonable model, a Rietveld refinement can be used to explore the fit of various models to the data and reject or approve them. Zachariasen (1949) solved the monoclinic structure of α- and β-UF_5 this way already in 1949. This trial-and-error approach is still used quite often. An example is given by Cockcoft and Fitch (1988). When solving the low-temperature structure of the molecule SF_6, they used the intramolecular distances and angles as chemical restraints in the least-squares refinements to ensure that the SF_6 octahedra did not distort beyond unphysical limits while exploring various intermolecular configurations.

The use of these subsidiary conditions by introducing prior chemical knowledge (e.g., distance and/or angle restraints) forces the minimized quantity of the least-squares refinement to satisfy the diffraction profile *and* the imposed model. This can be used, as in the above-mentioned case, to reduce the number of parameters or to obtain accurate structural parameters when refining very complex structures such as the one mentioned in Fig. 3.

One of the first structure determinations

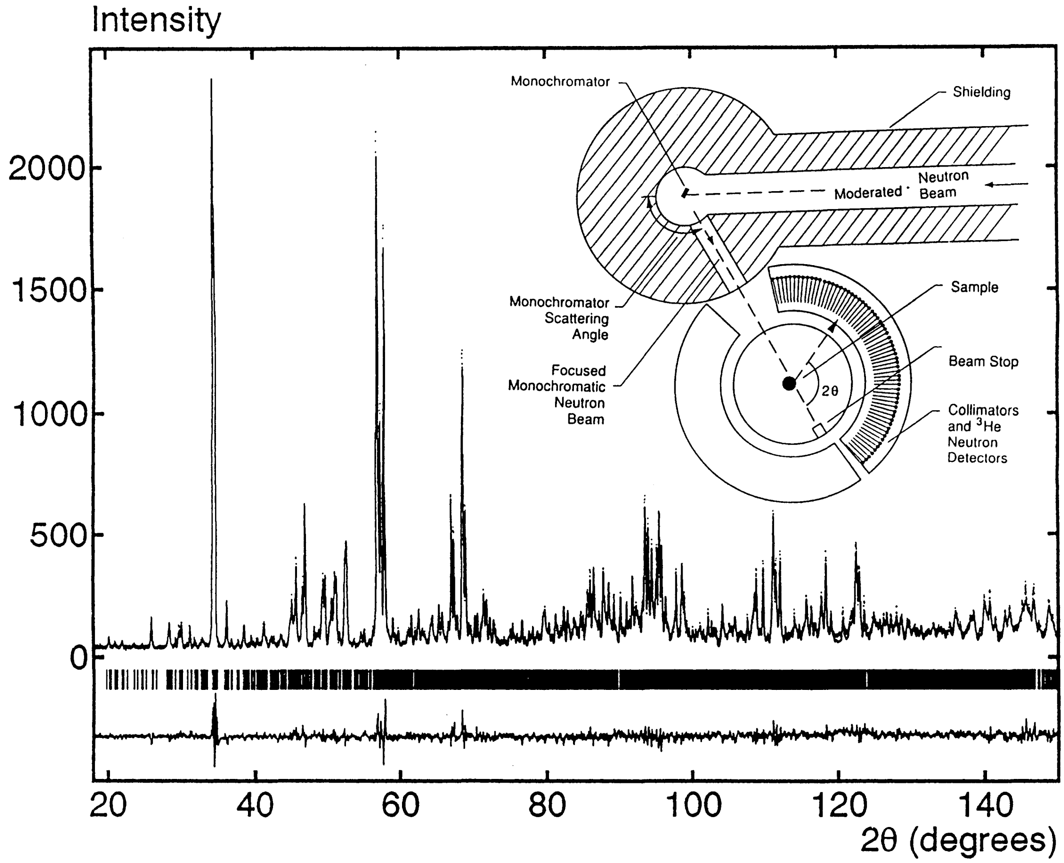

FIG. 3. A typical high-resolution neutron powder diffraction pattern recorded with the High-Resolution Neutron Powder Diffractometer at the High-Flux Beam Reactor at Brookhaven National Laboratory using a wavelength of 1.88 Å. A sketch of the basic scattering geometry of the instrument is shown. The material examined is the high-temperature form of Bi_2MoO_6, a prototypical catalyst. It has a monoclinic unit cell with a = 17.2623(1) Å, b = 22.4297(2) Å, c = 5.584 86(5) Å, and β = 90.4975(6)°. The structure was refined using the Rietveld technique (see text for details) using 2559 reflections whose positions in 2θ are indicated by the tick marks below the pattern and refining 146 parameters. The differences between the refined model and the data are given in the lower part of the plot.

in modern days that focused the general attention to solving structures from neutron powder diffraction data was done by Attfield in 1986. One of the more complicated structures solved was $Ga_2(HPO_3)_3 \cdot 4H_2O$, a novel framework structure with 29 atoms in the asymmetric unit cell and 117 structural parameters (Morris *et al.*, 1993). When solving structures using only powder diffraction the complementarity of neutron and x ray synchroton radiation is often very helpful. For a recent review see Cheetham *et al.* (1992).

One of the main reasons why "classical" direct methods to solve structures from powder data alone are not always successful is their need for atomic resolution and an extensive sampling of reciprocal space. The most crucial step in extracting information out of a complex powder pattern is the way heavily overlapped reflections are dealt with. Instead of ignoring them or assigning equal intensities to each particular reflection, maximum-entropy methods attempt to deal with them so that they are as noncommittal as possible with respect to the other data. For an example see Tremayne *et al.* (1992). The basic theory of the use of maximum entropy in crystallography and in powder diffraction in particular has been laid out by Bricogne (1991).

Besides solving structures, maximum-entropy methods are also used to reconstruct Patterson and Fourier densities. Papoular *et al.* (1991) established the superior quality of reconstructed densities when using maximum-entropy methods compared to standard Fourier methods. The low noise and smaller truncation effects allow the display of the disordered density.

2.2.2 The Rietveld Method The Rietveld technique makes use of the fact that the peak shape of Bragg reflections can be described analytically and the variation of their widths (FWHM, full width at half-maximum) can be expressed as a convolution of the optical characteristics of the diffractometer (takeoff angle θ_M, collimation, etc.) with sample-related effects (strain, particle size broadening, etc.). This allows the least-squares refinement of crystal structure parameters by fitting the diffraction profile without explicitly extracting structure factors or integrated intensities. Assuming a Gaussian shape for each Bragg reflection and referring to the original notation of Rietveld (1969), the contribution of a reflection k to a measured profile point y_{ik} at the position $2\theta_i$ is

$$y_{ik} = aj_kL_kS_k^2 \times 2[\ln 2/H_k\pi^{1/2}]^{1/2} \times \exp[-4\ln 2\{(2\theta_i - 2\theta_k)/H_k\}^2], \tag{9}$$

with a being the scale factor relating the integrated intensities to the modulus of the structure factor, j_k the multiplicity of the reflection, L_k the Lorentz factor, S_k the sum of nuclear and magnetic contributions, $2\theta_k$ the calculated position of the Bragg reflection, and H_k the FWHM, which in the simplest case can be expressed as a function of 2θ according to Cagliotti *et al.* (1958) in the following way:

$$H_k = (U\tan^2\theta_k + V\tan\theta_k + W)^{1/2}. \tag{10}$$

Rietveld refinement nowadays is the method with which neutron and x-ray powder diffraction data are analyzed. David *et al.* (1993) demonstrated the structural insight and details that modern high-resolution neutron powder diffraction can reveal when determining the structure of buckminsterfullerene (C_{60}) as a function of temperature. A very careful analysis of the low-temperature phase showed that the C_{60} molecules reorient between two energetically very close orientations. In one of the orientations, the pentagons face carbon-carbon double bonds in a neighboring molecule; in the other, energetically less favorable orientation the hexagons face the carbon-carbon double bonds. These results underline the fact that modern neutron powder diffraction can yield very accurate results, comparable in some cases to those obtained by single-crystal diffraction.

2.2.3 The Analysis of Mesoscopic Structures Using High-Resolution Powder Diffraction With the development of high-resolution neutron powder diffractometers in the range of $\Delta d/d \sim 5 \times 10^{-4}$, it was quickly realized that the resolution was so good that the mesoscopic structure of the sample became accessible and sometimes gave problems that had to be dealt with to be able to determine the atomic structure correctly.

Under mesoscopic structures we summarize structural inhomogeneities in materials beyond interatomic distances on length scales of tens to hundreds of angstroms, e.g., twin and domain structures, antiphase boundaries, or stacking faults. Since these effects manifest themselves on length scales beyond atomic resolution they reveal their presence in powder pattern mainly as deviations from the "ideal" symmetrical peak shape and resolution. The modeling of the peak profile parameters becomes more complicated than indicated in the simple expression (9) by Rietveld, since the isotropic peak shape, which is independent of the lattice direction, has to be replaced by an anisotropic one. The physical origin of these profile shape variations with the lattice direction can be very diverse and include variations of interatomic distances due to internal stresses, variations in stoichiometry, stacking faults, dislocations, and many more. These variations are then convoluted with the optical properties of the diffractometer (e.g., collimation, takeoff angle). This is why the knowledge of the peak shape and resolution function (variation of the full width at half-maximum with the scattering angle 2θ) of a given high-resolution powder diffractometer is of crucial importance for these studies.

The two most commonly encountered mesoscopic structures in high-resolution powder diffraction are strain and particle size effects, which lead to a broadening of reflections. If one compresses a polycristalline material in a particular direction, then the *d* spacings are varied depending on the direction and strength of the stress with respect to the individual crystallites. If one measures the *d* spacings in different orientations of the unit cell to the stress, one observes variations. These strains can be converted into stresses using the elastic constants of the probed material. They are typically in the order of 10^{-4} in $\Delta d/d$, the resolution now accessible with high-resolution neutron powder diffractometers. The knowledge of the stress distribution in matter is not only important for applied science (engineering parts such as railway tracks, welds, etc.) and earth sciences (e.g., stress distribution in rocks), but may also play a major role in investigations of phase transitions, lattice defects, and compositional variation in materials.

Differentiating Bragg's law one obtains

$$\Delta d/d = -\Delta 2\theta(2\tan\theta) \qquad (11)$$

or

$$\Delta 2\theta = -2\epsilon \tan\theta \quad \text{with} \quad \epsilon = \Delta d/d.$$

This shows that strain will manifest itself in the high-angle reflections.

One of the basic assumptions of the kinematical scattering theory applied in general to analyze diffraction data is that of an infinite lattice that allows the reciprocal lattice points to be treated as δ-functions. If the particle size diminishes, the coherently diffracting domains have a finite size, smearing out all the reciprocal lattice points to the same spread. The size broadening is described by the Scherrer relation:

$$\Delta 2\theta = \text{const}\ \lambda/\cos\theta\, T, \qquad (12)$$

T being the thickness of the coherently diffracting domain. Both effects, particle size and strain broadening, can be present in the same material, and both will manifest themselves at high angles, where they can be separated as a result of their different variation with 2θ (size broadening increases with $1/\cos\theta$, strain broadening increases with $\tan\theta$). Equation (12) shows that particle size broadening is wavelength dependent. Time-of-flight neutron data have to be corrected for this, if the quantities that are to be determined like accurate atomic displacement parameters are not to be corrupted by systematic errors. One of the basic assumptions of the Rietveld method was the simple angular variation of the FWHM of a reflection profile with 2θ usually expressed using Cagliotti's formula (Eq. 10). With the continuing efforts of improving reactor- and spallation-source–based high-resolution neutron powder diffractometers, this simple angular variation will have to be replaced by more complex ones.

The investigation of mesoscopic features in materials will play a major role in the future. Their importance, especially in superconducting ceramics, has been recognized. The following examples illustrate this further.

1. Medarde *et al.* (1992) were able to model the anisotropic broadening of all reflections *hkl* with $h > 0$ in $La_{1.5}Sr_{0.5}NiO_{3.6}$ and understand its origin in an oxygen vacancy ordering introducing microstrain in the NiO_2 basal plane.
2. Roessli *et al.* (1993) found linewidth and shape anomalies *only* in the magnetic reflections of the low-temperature magnetic phase of $HoBa_2Cu_3O_7$ at 7 mK. These asymmetric line shapes could be ex-

plained as magnetic stacking faults along the *c* axis of the structure resulting from finite magnetic correlations that are in the order of 29 Å. The same type of magnetic mesoscopic structure was found at 10 K in samples of $PrBa_2Cu_3O_6$ by Guillaume *et al.* (1993).

The last two examples again underline the unique information one can obtain from neutron diffraction data with regard to magnetic structures.

2.3 Single-Crystal Neutron Diffraction

Classical areas of single-crystal neutron diffraction are

1. the location of "light" or neighboring elements;
2. the determination of magnetic structures;
3. the high-resolution structural investigations to detect subtle changes of crystal structures near structural phase transitions by analyzing atomic displacement parameters or to obtain accurate atomic positions for electron density studies.

In materials science and structural chemistry, a lot of problems concerned with the first two points can nowadays be tackled using powder diffraction. However, when one attempts to solve magnetic structures or refine biological structures, single-crystal diffraction will always be the method of choice. Protein crystallography using neutrons has not shown its full power yet, since it is difficult to grow single crystals of sufficient size and the fluxes of monochromatic beams available at the moment are often not high enough to tackle interesting problems. Even if oversimplified, the physicist's description of a protein as a molecule made up to roughly 50% by hydrogen serves to make the point concerning the inherent potential of neutron diffraction in the field of biological structures. The availability of new sources providing high-flux neutron beams together with advances in instrumentation—and here especially in two-dimensional position-sensitive detectors—will provide a way out of the current dilemma.

The high-resolution single-crystal experiments profit from the accuracy of high-angle intensities. The Fourier transform of the atomic displacement parameters (see CRYSTALLOGRAPHY) represents the probability of locating an atom in real space. This so-called probability density function reveals the dynamic and static disorder of atoms. At high temperatures or close to phase transitions the underlying harmonic approximation of the potential function may no longer be valid. By using higher expansion terms of the atomic displacement parameters these disorder phenomena can be quantified. The fact that the neutron scattering length is constant in **Q** space results in very accurate high-angle structure factors needed to determine reliable atomic displacement parameters. For more details see Kuhs (1992).

2.3.1 Two-Dimensional Position-Sensitive Detectors Single-crystal neutron diffraction is changing rapidly as a result of improvements in instrumentation and particularly of the availability of two-dimensional position-sensitive detectors. The classical method of measuring a diffraction pattern is to measure the intensity step by step using a single counter. This is very time-consuming and not very efficient since in general the number of reflections "available" in a given solid angle is in the dozens to hundreds in the case of proteins. The gain by measuring simultaneous reflection is enormous. Multidetectors can be one-dimensional when measuring simultaneously in one direction. They are used in powder diffraction where a whole powder pattern can be recorded "at once" without stepping the counter. For single-crystal diffraction one needs two directions along which one would like to measure simultaneously.

The use of two-dimensional position-sensitive detectors (2D-PSD) in single-crystal neutron diffraction has been shown to be very suitable in single-crystal time-of-flight measurements at spallation sources (Schultz *et al.*, 1984) and in reactor-based macromolecular crystallography (Schoenborn *et al.*, 1984), since many Bragg reflections are excited simultaneously and thus the time of an experiment can be reduced considerably. In protein crystallography this can determine if an experiment can be done at all.

Recently Lehmann *et al.* (1989) and McIntyre (1992) made the case for using a two-dimensional detector when analyzing single crystals with smaller unit cells. Besides the obvious advantages of a much faster measurement time, better separation from undesired parasitic scattering due to sample environments such as furnaces and pressure cells, and a more accurate integration of weak Bragg intensities because of the much better determination of the background, they demonstrated that when studying crystals with poor quality or twinning, the 2D-PSD offers enormous advantages. McIntyre and Renault (1989) were able to separate overlapping reflections of a twinned sample of $YBa_2Cu_3O_{7-\delta}$ and integrate the four domain orientations individually (see Fig. 4). They revealed that the difference in intensity between certain pairs of individual overlapping reflections was sensitive to the difference in the occupancy of oxygen sites in the plane between pairs of closest barium atoms.

Another neutron scattering technique that has benefited from the use of 2D-PSDs is small-angle neutron scattering (SANS). With SANS one probes mesoscopic features like aggregates of atoms or structural inhomogeneities in matter (voids in solids, precipitates, etc.). One uses long-wavelength neutrons, typically beyond 10 Å, and detects the elastic coherent scattering to obtain information about size and shape of the studied aggregates. The installation of cold sources (e.g., liquid-hydrogen moderator) to obtain higher fluxes of cold neutrons and the development of 2D-PSDs has allowed exciting new experiments in this field. Recently the detection of flux lines in high-temperature superconducting single crystals using SANS was a major step forward in our understanding of these materials (see NEUTRON SCATTERING). Neutron scattering is the only method by which one can measure *directly* the spatial variation of a magnetic field in the bulk of a superconductor.

2.4 Neutron Reflectometry

Scattering from a surface or interface without changing the energy of the probing particle and at the same glancing angle of incident and reflected beam is referred to as elastic, specular scattering (see Fig. 5). If the wave vector $\mathbf{Q}$ is such that $2\pi/Q$ is much bigger than the interatomic distances in the medium, it can be described as a continuum.

In vacuum, the neutron has no potential energy and the kinetic energy is $(h/2\pi)k_0^2/2m$. The potential energy of a neutron in the continuum limit is $(h/2\pi)bN/m$, N being the number of scattering nuclei and b the scattering amplitude. In a medium, the total energy is the sum of potential and kinetic energy. In analogy to classical optics, the medium is characterized by a refractive index:

$$n = [(1/4\pi)bN/K_0^2]^{1/2}. \qquad (13)$$

One can see that the variation of the refractive index is a measure of the density variation $\rho(z) = bN$ perpendicular to the surface. By measuring the reflectivity as a function of $\mathbf{Q}$ one can determine the thickness of a thin film and its scattering density ρ since in the Born approximation the reflectivity is given as

$$|R|^2 = [(4\pi)^2/\mathbf{Q}^4]2\rho^2(1 - \cos Qd). \qquad (14)$$

Even though it is not possible to resolve the atomic structure of the film if $\mathbf{Q}/2\pi$ is much bigger than the interatomic distances, changes in the scattering density $\Delta\rho$ with angstrom resolution can still be observed. One can measure reflectivities at $\mathbf{Q}/2\pi$ values in the region of interatomic distances and then can determine the atomic structure. This is then a normal diffraction experiment.

Neutron reflectometry is a rapidly expanding field and has allowed, for instance, a better understanding of block copolymers—polymers made out of two different types of monomers linked to each other. They are used as, e.g., surfactants, adhesives, and compatibilizing agents in polymer blends. The different polymer blocks tend to separate, leading to either a lamellar structure of alternating layers or a rod structure with one block being embedded in a matrix of the other block. The lamellar structures are ideal study objects for neutron reflectometry, since varying the contrast through selective deuteration allows one or the other block to be highlighted.

If the material is saturated magnetically, the scattering length b incorporates the mean magnetic and nuclear scattering lengths of the material. So the product $(b_n \pm b_m)N$ is a measure of the sum of the coherent nuclear and magnetic scattering-length densities. The sign depends on the spin component of the neutron and is determined by whether the magnetic moment is aligned parallel (spin up) or antiparallel (spin down) to an applied ex-

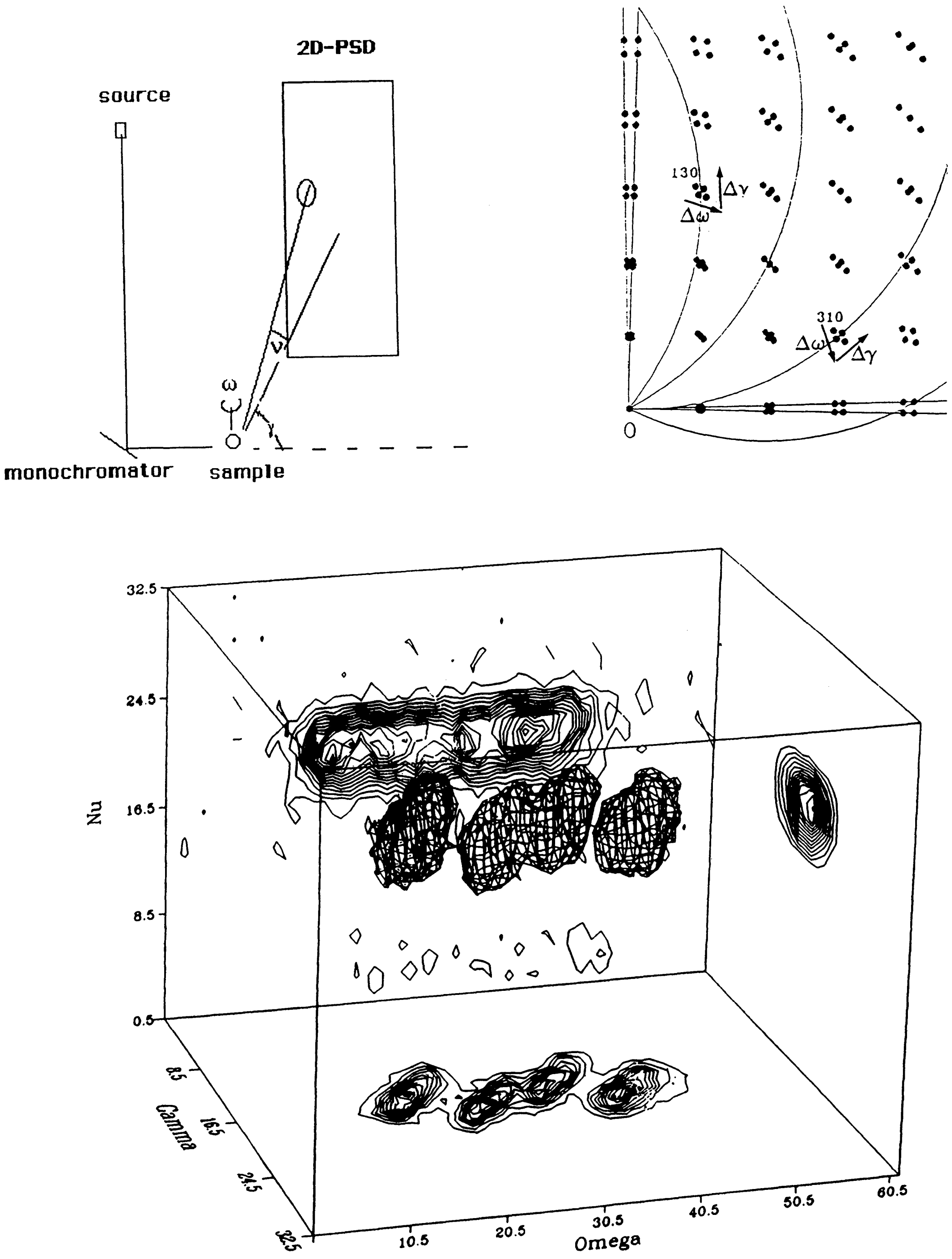

FIG. 4. The basic scattering geometry of a single-crystal diffractometer using a two-dimensional position-sensitive detector. The 420 and 240 Bragg reflections of a twinned $YBa_2Cu_3O_{7-\delta}$ crystal obtained when turning around ω are shown. MacIntyre and Renault (1989) were able to separate nearly a complete set of untwinned single-crystal intensities using 2D-PSD data. The ω projections show that a single detector would record reflections that are not resolved when performing the same scan.

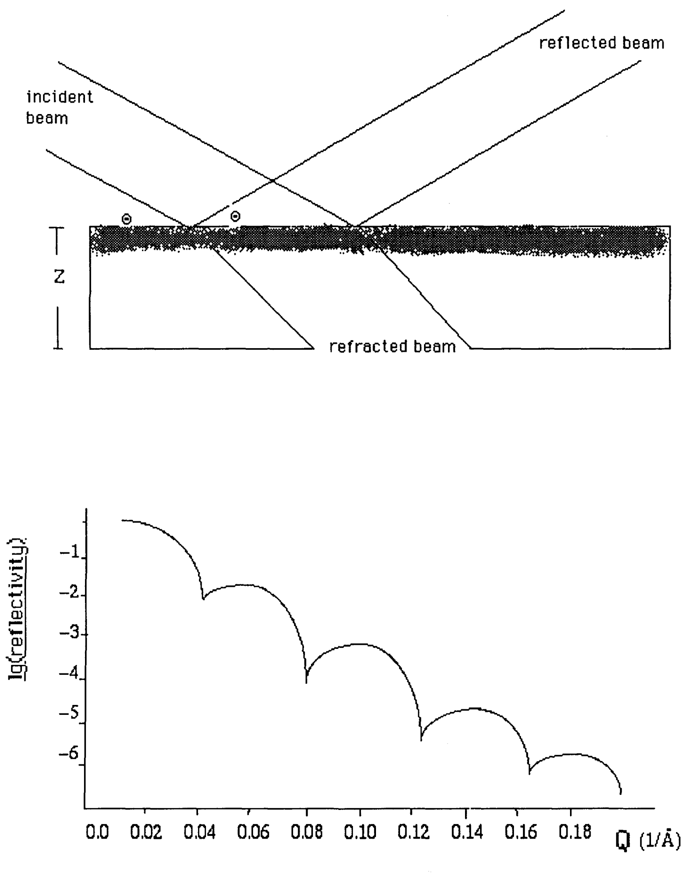

FIG. 5. The principal arrangement of specular reflection experiments is shown. The glancing angles θ of incident and reflected beams are the same. The density profile along z is probed in such an experiment. Below, the variation of the neutron reflectivity of a thin Ni film on a silicon substrate is sketched. The thickness of the film can be obtained directly from the period of oscillation of the reflectivity, since the reflectivity $|R|^2 = (16\pi^2/\mathbf{Q}^4)2\rho^2(1 - \cos Qd)$, where d is the film thickness and ρ the scattering density along z.

ternal field. Thus the critical angle for the two spin states is different and allows their separation.

It is straightforward to calculate a specular reflectivity for a given model, but to solve the inverse problem of determining a chemical and/or magnetic density profile from measured reflectivity curves is much more demanding and not always successful. For details see Felcher and Russell (1991).

If one is interested in the structure of an interface, and particularly in its roughness, it is sometimes not possible to distinguish between different models by measuring only the specular reflectivity, since one averages over the surface perpendicular to **Q**. By measuring in a nonspecular setting one might be able to distinguish between different interface models by the shape and amplitude of the diffuse scattering, resulting from a different in-plane roughness.

By applying Snell's law and setting the refracted angle in the medium equal to zero, one obtains a critical value of the wave-vector transfer $Q_c = (16\pi\rho)^{1/2}$ below which total reflection will occur. Below Q_c the kinetic energy of the neutron is smaller than the potential barrier of the medium. This is the principle of neutron guides now widely used to transport neutrons by internal reflections to experiments situated up to 100 m away from the core or target. By curving them one can reduce the background consisting of γ rays and fast neutrons. They are highly polished, nickel-plated glass tubes with rectangular cross sections. With natural nickel $\theta_c \sim 0.1°$ at $\lambda = 1$ Å. The value of θ_c can be increased by 20% by the use of ^{58}Ni. A further increase can be obtained by using multilayer materials (e.g., NiTi with $\theta_c = 3\theta_{c\mathrm{Ni}}$). Multilayers are composites made of repetitive blocks of discrete material characterized by a lattice spacing d separated by another block of crystalline or amorphous material. By uniformly varying the layer distance, the Bragg reflection can be separated in such a way that its tail runs into the edge of the total reflection at θ_C. This leads to a substantial gain of reflectivity in comparison to normal guides and can be used to make magnetic as well as nonmagnetic mirrors. The challenge for the future of neutron scattering will be whether the tremendous potential of polarized neutron scattering using high-flux spin-polarized beams can be realized. This requires substantial instrumental developments as well as a new generation of high-flux reactors and spallation sources.

Works Cited

Attfield, J. P., Sleight, A. W., Cheetham, A. K. (1986), *Nature* **322**, 620–622.

Bricogne, G. (1991), *Acta Crystallogr.* **A47**, 803–829.

Cagliotii, G., Paoletti, A., Ricci, F. P. (1958), *Nucl. Instrum. Methods* **3**, 223.

Cheetham, A. K. (1992), *Angew. Chem., Int. Ed. Engl.* **32**, 1557–1570.

Cockcroft, J. K., Fitch, A. F. (1988), *Z. Kristallogr.* **184**, 123–145.

David, W. I. F., Ibberson, R. M., Matsuo, T. (1993), *Proc. R. Soc. London A* **442**, 129–146.

Felcher, G. P., Russell, T. P. (1991), *Physica B* **173**, 1–210.

Gaskell, P. H., Eckersley, M. C., Barnes, A. C., Chieux, P. (1991), *Nature* **350**, 675–677.

Guillaume, M., Fischer, P., Roessli, B., Podlesnyak, A., Schefer, J., Furrer, A. (1993), *Solid State Commun.* **88** (1), 57–61.

Jacrot, B. (1976), *Rep. Prog. Phys.* **39**, 911.

Kuhs, W. F. (1992), *Acta Crystallogr.* **A48**, 80–98.

Lehmann, M. S., Kuhs, W. F., McIntyre, G. J., Wilkinson, C., Allibon, J. R. (1989), *J. Appl. Crystallogr.* **22**, 652–568.

McIntyre, G. (1992), *Neutron News,* **3** (2), 15–19.

McIntyre, G., Renault, A. (1989), *Physica* **B156/157**, 880–883.

Medarde, M., Rodriguez-Carvajal, J., Obradors, X., Sayagues, M. J., Vallet, M., Gonzalez-Calbet, J. (1992), *Physica B* **180/181**, 402.

Morris, R. E., Harrison, W. T. A., Nicol, J. M., Wilkinson, A. P., Cheetham, A. K. (1993), *Nature* **359**, 519–522.

Papoular, R. J., Prandl, W., Schiebel, P. (1991), in: C. R. Smith *et al.* (Eds.), *Maximum Entropy and Bayesian Methods*, Norwell, MA: Kluwer Academic Publishers, pp. 359–376.

Rietveld, H. (1969), *J. Appl. Crystallogr.* **2**, 65.

Roessli, B., Fischer, P., Staub, U., Zolliker, M., Furrer, A. (1993), *Europhys. Lett.* **23** (7), 511–515.

Schoenborn, B. P. (1984), in: B. P. Schoenborn (Ed.), *Neutrons in Biology*, New York: Plenum, pp. 261–279.

Schultz, A. J., Srinivasan, K., Teller, R. G., Williams, J. M., Lukehart, C. M. (1984), *J. Am. Chem. Soc.* **106**, 999–1003.

Shull, C. G., Smart, J. S. (1949), *Phys. Rev* **76**, 1256.

Tremayne, M., Lightfott, P., Metha, M. A., Bruce, P. G., Harris, K. D. M., Shankland, K., Gilmore, C. J., Bricogne, G. (1992), *J. Solid State Chem.* **100**, 191–196.

Zachariasen, W. H. (1949), *Acta Crystallogr.* **2**, 296.

Further Reading

Bacon, G. E. (1975), Neutron Diffraction, Oxford: Clarendon Press.

Lovesey, S. W. (1984), *Theory of Neutron Scattering from Condensed Matter*, Vols. 1 and 2, Oxford: Clarendon Press.

Price, D. L., Skold, K. (1986), *Introduction to Neutron Scattering in Methods of Experimental Physics*, Vol. 23, Part A, Orlando, FL: Academic, p. 1.

Rossat-Mignod, J. M. (1987), "Magnetic Structures," in: *Methods of Experimental Physics*, Vol. 23, Part C, Orlando, FL: Academic, p. 69.

White, J. W., Windsor, C. G. (1984), *Rep. Prog. Phys.* **47**, 707.

Windsor, C. G. (1981), *Pulsed Neutron Scattering*, London; Taylor and Francis.

7
Neutron Spectroscopy

Anton Heidemann

7.1 Introduction 185
7.2 Theoretical Foundations of Neutron Scattering 187
7.2.1 Typical Response Functions 190
7.2.2 Magnetic Scattering 190
7.3 Techniques in Neutron Scattering 191
7.3.1 Neutron Sources 191
7.3.1.1 Nuclear Fission Reactors 191
7.3.1.2 Spallation Sources 191
7.3.2 Moderators 192
7.3.3 Neutron Optics 192
7.3.4 Neutron Scattering Instruments 194
7.3.4.1 Instruments without Energy Analysis 194
7.3.4.2 Instruments with Energy Analysis 194
7.4 Neutron Spectroscopy: Techniques and Applications 194
7.4.1 Time-of-Flight (TOF) Spectroscopy 194
7.4.1.1 Direct Geometry Spectrometers 195
7.4.1.2 Indirect (Inverted) Geometry Spectrometers 197
7.4.2 Three-Axis Spectroscopy (TAS) 198
7.4.2.1 Principle 198
7.4.2.2 Layout 199
7.4.2.3 The TAS Resolution Function 200
7.4.2.4 Applications 200
7.4.2.5 Comparison between TAS and TOF Spectroscopy of Coherent Excitations in Single Crystals 200
7.4.3 Back-Scattering (BS) Spectroscopy 201
7.4.3.1 Principle 201
7.4.3.2 Existing Instruments 202
7.4.3.3 Applications 202
7.4.4 Neutron Spin Echo (NSE) Spectroscopy 203
7.4.4.1 Principle 203
7.4.4.2 Implementation 204

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

7.4.4.3 NSE Variants 205
7.4.4.4 NSE versus Standard Inelastic Spectrometers 206
7.5 Conclusion 206
Acknowledgments 207
Glossary 207
References 207
Web sites 208

7.1
Introduction

The neutron is composed of one up and two down quarks with charges of 2/3 and −1/3, respectively. The total charge of the neutron is zero. Its internal structure leads to an electrical charge distribution. This results in a nonzero magnetic moment and an electric polarizability. Hot, thermal, cold, and very cold neutrons have wavelengths ranging from 0.1 to 100 Å and energies from a few electronvolts to microelectronvolts or from terahertz to megahertz. They are therefore an ideal probe for the study of the atomic order and the atomic motions in condensed matter. Neutrons behave like waves. Therefore, a neutron beam can be monochromatized by Bragg reflection from crystals. Neutrons also behave like particles flying with velocities in the range from kilometers per second to meters per second. Their energy can therefore be determined by time-of-flight (TOF) measurements. The zero net charge of the neutron means that it interacts very weakly with matter and penetrates deeply into the sample and the sample container. Neutrons interact directly with the nuclei of the atoms. This short-range interaction can be considered point like characterized by a single parameter, with the scattering length a of the order of 10^{-12} cm. Figure 7.1 shows the variation of the scattering length for the different elements. There is no systematic variation across the periodic table as there is for X-rays and electrons.

The magnetic moment of the neutron interacts with the unpaired electron spins of magnetic atoms with strength comparable to that of the nuclear interaction. Therefore, the neutron is also a powerful probe of magnetic properties of solids. The effective magnetic scattering length a_m for an atom with a magnetic moment of μ Bohr magnetons is given by

$$a_m = 0.27 \times 10^{-12}\ \mu(\text{cm}) \qquad (7.1)$$

Neutrons are spin 1/2 particles. A neutron beam can therefore be polarized. A polarized neutron beam entering a magnetic field perpendicular to its magnetic moment will undergo Larmor precession. This property is used in the neutron spin echo (NSE) technique.

The neutron beams used in scattering experiments can be produced either by a nuclear reactor or by a neutron spallation source. Both types of sources can be either continuous or pulsed. However, the majority of research reactors are

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

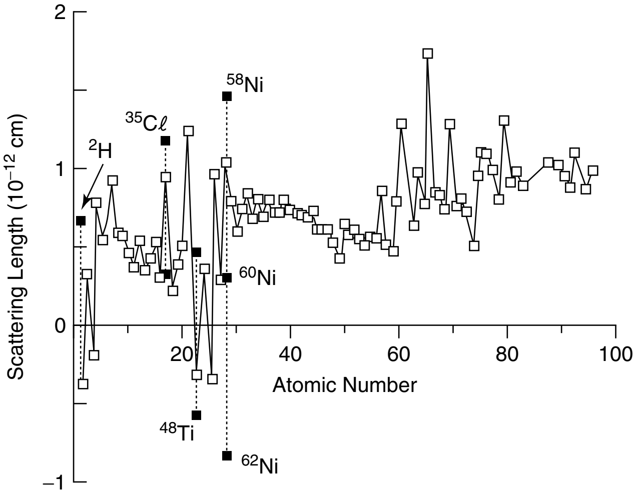

Fig. 7.1 Variation of the coherent scattering length *a* with atomic number (Price and Sköld, 1986).

continuous sources, and most of the existing spallation sources are pulsed. The latter are well suited to TOF techniques. In a neutron scattering experiment, one measures the wave vector transfer **Q** or the momentum transfer $\hbar\mathbf{Q}$ and the transfer of energy $\hbar\omega$ of the neutron to the sample

$$\mathbf{Q} = \mathbf{k}_i - \mathbf{k}_f \tag{7.2}$$

$$\hbar\omega = E_i - E_f = \left(\frac{\hbar^2}{2m}\right)\left(k_i^2 - k_f^2\right) \tag{7.3}$$

Equation (7.2) describes the momentum conservation and Eq. (7.3) the energy conservation. $k = 2\pi/\lambda$ is the magnitude of the wave vector and λ is the wavelength of the neutron. The subscripts i and f refer to the initial and final neutron beam.

A few additional remarks concerning the scattering triangle may be useful (see Figure 7.2). The scattering event is characterized by 6 variables: 3 from each vector $\mathbf{k}_i$ and $\mathbf{k}_f$. However, we are only interested in **Q** and $\hbar\omega$. Therefore, we have 2 redundant variables. They should be chosen to optimize the experiment. For noncrystalline samples, only $|\mathbf{Q}| = Q$ and $\hbar\omega$ are meaningful. Here, we have four redundant variables.

The following relations are valid for the energy E and the velocity v of neutrons:

$$E[\text{meV}] = 2.072\ k^2[\text{Å}^{-2}] \tag{7.4}$$

$$v[\text{m/s}] = \frac{3956}{\lambda[\text{Å}]} \tag{7.5}$$

Energy conversion:

$$1\ \text{meV} = 0.2418\ \text{THz} = 11.6045\ \text{K}$$
$$= 8.006554\ \text{cm}^{-1} \tag{7.6}$$

We distinguish between two types of experiments:

1. Neutron scattering without energy analysis: this is the domain of diffraction, reflectometry, small-angle and diffuse scattering.
2. Neutron scattering with energy analysis. This is the domain of *neutron spectroscopy*.

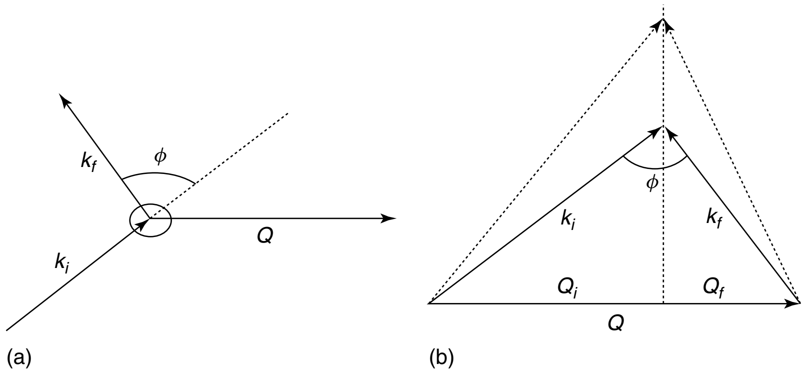

Fig. 7.2 Scattering in (a) real and (b) reciprocal space. The dotted line in (b) shows an alternate configuration leading to the same momentum and energy transfers (**Q**, ω) (Hippert *et al.*, 2005).

In this monograph, we only briefly mention neutron scattering without energy analysis. The main part of the chapter is devoted to neutron spectroscopy covering both *inelastic neutron scattering* (*INS*) and *quasi-elastic neutron scattering* (*QENS*).

Figure 7.3 shows the areas in the (Q, ω) and (r, t) planes, which are covered by neutron spectroscopy. It also shows the areas investigated by other scattering techniques such as electromagnetic and electron scattering. It is obvious that neutron spectroscopy is complementary to these other techniques.

A very comprehensive review of neutron spectroscopy has been published recently (Hippert *et al.*, 2005). The interested reader is referred to this book for further reading.

7.2 Theoretical Foundations of Neutron Scattering

The quantity that essentially results from neutron scattering experiments is the double differential cross section $d^2\sigma/d\Omega dE_f$. It describes the probability with which an incident neutron of energy E_i is scattered by an angle ϕ into a solid angle volume element $d\Omega$ and an energy element between E_f and $E_f + dE_f$. $d^2\sigma/d\Omega dE_f$ can be derived in first-order perturbation theory (Born approximation). van Hove (1954) has shown that $d^2\,\sigma/d\Omega dE_f$ is connected with correlation functions, which describe the motion of the scattering particles in space and time. Therefore, they allow the calculation of $d^2\,\sigma/d\Omega dE_f$ in terms of simple physical models.

In the following, we briefly describe the main ingredients of this theory. Details can be found in (Lovesey, 1984; Scherm, 1972).

The Hamiltonian of the total system, consisting of the sample and the neutron, is given by Eq. (7.7)

$$H = H_0 + \frac{p_n^2}{2m} + V \qquad (7.7)$$

Here, H_0 is the Hamiltonian of the unperturbed system, V is the interaction potential between the neutron and the scattering particles, and ${p^2}_n/2m$ is the kinetic energy of the neutron, $\mathbf{p}_n = \hbar\mathbf{k}$.

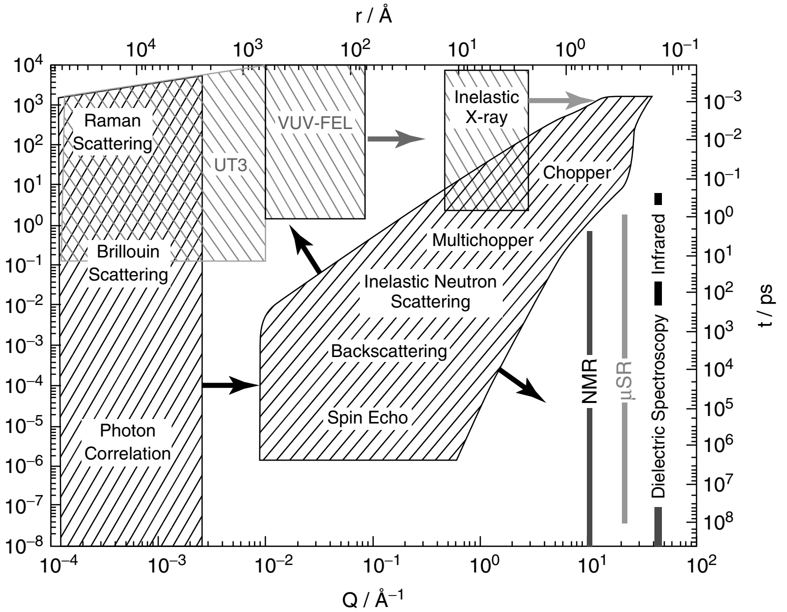

Fig. 7.3 Distance–time (r, t) and ($\mathbf{Q}$, ω) scales probed by neutron spectroscopy. (Pynn, unpublished.)

The incident and scattered neutrons are characterized by plane waves with spin σ: $|\mathbf{k}\sigma>$. In the following, we consider, for simplicity, only the case of unpolarized neutrons.

We denote the eigenfunctions and eigenvalues of the unperturbed system by $|n>$ and E_n with

$$H_0|n>= E_n|n> \tag{7.8}$$

Using Fermi's golden rule, the transition probability per unit time of the system from an initial state $|\mathbf{k}_i n_0>$ to a final state $|\mathbf{k}_f n_1>$ is given by

$$w(\mathbf{k}_i,\, n_0 \longrightarrow \mathbf{k}_f,\, n_1) = \frac{2\pi}{\hbar}|<\mathbf{k}_f\; n_1| \times V|\mathbf{k}_i\; n_0>|^2\delta\left(E_{n_1} - E_{n_0} - \hbar\omega\right) \tag{7.9}$$

We have to sum up over all transitions, where the wave vector changes from $\mathbf{k}_i$ to $\mathbf{k}_f$:

$$w(\mathbf{k}_i \longrightarrow \mathbf{k}_f) = \sum_{n_0}\sum_{n_1} p(n_0) \times w(\mathbf{k}_i, n_0 \longrightarrow \mathbf{k}_f, n_1) \tag{7.10}$$

$p(n_0)$ is the statistical weight of the state $|n_0>$:

$$p(n_0) = \frac{\exp\left(\dfrac{-E_{n_0}}{k_BT}\right)}{\sum\limits_{n_0}\exp\left(\dfrac{-E_{n_0}}{k_BT}\right)} \tag{7.11}$$

With normalized plane waves

$$\mathbf{k}> = L^{-3/2}\ \exp(i\mathbf{k}\cdot\mathbf{r}) \tag{7.12}$$

and the Fermi pseudopotential for V (Fermi, 1936)

$$V(\mathbf{r}-\mathbf{r}_i) = \left(\frac{2\pi\hbar^2}{m}\right) a_i\delta(\mathbf{r}-\mathbf{r}_i) \tag{7.13}$$

we obtain finally

$$\begin{aligned}\frac{d^2\,\sigma}{d\Omega\; dE_f} &= \left(\frac{k_f}{k_i}\right)\cdot N^{-1}\sum_{n_0}\sum_{n_1} p(n_0)\\ &\times\sum_i\sum_j a_i a_j < n_0|\exp(-i\mathbf{Q}\cdot\mathbf{r}_i)|\\ &\times n_1 >< n_1|\exp(i\mathbf{Q}\cdot\mathbf{r}_j)|\\ &\times n_0 > \delta(E_{n_1}-E_{n_0}-\hbar\omega)\end{aligned} \qquad (7.14)$$

where a_i is the scattering length of the ith nucleus. N is the number of scattering centers.

Introducing the coherent and incoherent scattering functions (the dynamic structure factors) $S_{\text{coh}}(\mathbf{Q}, \omega)$ and $S_{\text{inc}}(\mathbf{Q}, \omega)$, we get

$$\begin{aligned}&\frac{d^2\sigma}{d\Omega\; dE_f}\\ &= \hbar^{-1}\left(\frac{k_f}{k_i}\right)\cdot[< a >^2\; S_{\text{coh}}(\mathbf{Q}, \omega)\\ &\quad + (< a^2 > - < a >^2)S_{\text{inc}}(\mathbf{Q}, \omega)]\end{aligned} \qquad (7.15)$$

$$\begin{aligned}&S_{\text{coh}}(\mathbf{Q}, \omega)\\ &= N^{-1}\sum_{n_0}\sum_{n_1} p(n_0)\sum_i\sum_j < n_0\\ &\quad\times|\exp(-i\mathbf{Q}\cdot\mathbf{r}_i)|n_1 >< n_1|\\ &\quad\times\exp(i\mathbf{Q}\cdot\mathbf{r}_j)|\\ &\quad\times n_0 > \delta(E_{n_1}-E_{n_0}-\hbar\omega)\end{aligned} \qquad (7.16)$$

$$\begin{aligned}&S_{\text{inc}}(\mathbf{Q}, \omega)\\ &= N^{-1}\sum_{n_0}\sum_{n_1} p(n_0)\sum_{j=1}^{N}| < n_1|\\ &\quad\times\exp(i\mathbf{Q}\cdot\mathbf{r}_j)|n_0 > |^2\\ &\quad\times\delta(E_{n_1}-E_{n_0}-\hbar\omega)\end{aligned} \qquad (7.17)$$

We assume that the sample contains a mixture of isotopes with different scattering lengths $a_{j.}$ that are randomly distributed over the positions r_j. By this, we introduce the following averages:

$< a >= a_{\text{coh}}$ is the coherent scattering length.

$< a^2 > - < a >^2 = a^2{}_{\text{inc}}$ is the square of the incoherent scattering length. Its origin can be either isotopic or spin incoherence: for nuclei with a spin I not equal to zero, the interaction depends on the orientation between the spin of the neutron and the nucleus, with scattering lengths a_+ and a_-. In the case where there is no correlation between the nuclear spins and their sites, we get

$$a_{\text{coh}} = \frac{(I+1)a_+ + Ia_-}{(2I+1)} \qquad (7.18)$$

$$a^2_{\text{inc}} = \frac{I(I+1)(a_+ - a_-)^2}{(2I+1)} \qquad (7.19)$$

The scattering functions depend only on the properties of the unperturbed sample and not on the properties of the scattered particle. They can be calculated if the eigenfunctions of the sample are known.

van Hove (1954) established the following relations:

$$\begin{aligned}&S_{\text{coh}}(\mathbf{Q}, \omega)\\ &= (2\pi)^{-1}\int\exp[i(\mathbf{Q}\cdot\mathbf{r}-\omega t)]\\ &\quad\times G(\mathbf{r}, t)d\mathbf{r}dt\end{aligned} \qquad (7.20)$$

$$\begin{aligned}&S_{\text{inc}}(\mathbf{Q}, \omega)\\ &= (2\pi)^{-1}\int\exp[i(\mathbf{Q}\cdot\mathbf{r}-\omega t)]\\ &\quad\times G_s(\mathbf{r}, t)d\mathbf{r}dt\end{aligned} \qquad (7.21)$$

G and G_s are correlation functions in space and time. S_{coh} and S_{inc} are their four-dimensional Fourier transforms in $(\mathbf{Q}, \omega)$ space. The so-called intermediate scattering functions $I(\mathbf{Q}, t)$ and $I_s(\mathbf{Q}, t)$

are defined as

$$I(\mathbf{Q}, t) = \int \exp(i\mathbf{Q} \cdot \mathbf{r}) G(\mathbf{r}, t) d\mathbf{r} \quad (7.22)$$

$$I_s(\mathbf{Q}, t) = \int \exp(i\mathbf{Q} \cdot \mathbf{r}) G_s(\mathbf{r}, t) d\mathbf{r} \quad (7.23)$$

These functions are measured in NSE spectroscopy (see Section 7.4.4).

So far, all the calculations were done in the framework of quantum mechanics. However, using the correlation function approach of van Hove, things can be simplified in the classical limit. Examples are the calculation of S_{inc} for simple cases like translational diffusion in liquids and solids, reorientations of molecular groups in crystals at not too low temperatures, phonon spectra, and so on. The quantum mechanical approach is needed for the case of tunneling of molecular groups in condensed matter at low temperature.

Collective excitations (phonons, magnons) with dispersion, where ω depends on $\mathbf{q} = \mathbf{Q} - \boldsymbol{\tau}$ (see Section 7.4.2) and the intensity depends on $\mathbf{Q}$, are determined via the measurement of S_{coh}. Information about local single-particle motions (diffusion, reorientation of molecules) is obtained from the measurement of S_{inc}. Here, only the intensity depends on $\mathbf{Q}$ and the frequency is independent of $\mathbf{q}$.

7.2.1
Typical Response Functions

An excitation at a given frequency ω_q with an infinite lifetime can be represented by

$$\mathbf{S}(\mathbf{Q}, \omega) = \frac{\mathbf{A}(\mathbf{Q})}{\left(1 - \exp\left(\frac{-\hbar\omega}{k_B T}\right)\right)} \times [\delta(\omega - \omega_q) - \delta(\omega + \omega_q)] \quad (7.24)$$

where $A(Q)$ is a dimensionless structure factor.

If the excitation has a finite lifetime, the delta functions can be replaced by Lorentzians. Overdamped excitations can be described by quasi-elastic Lorentzians:

$$\mathbf{S}(\mathbf{Q}, \omega) = \frac{\mathbf{A}(\mathbf{Q})}{\left(1 - \exp\left(\frac{-\hbar\omega}{k_B T}\right)\right)} \times \frac{1}{\pi} \frac{\Gamma_Q}{\omega^2 + \Gamma^2{}_Q} \quad (7.25)$$

where Γ_Q is the half width at half maximum (HWHM) of the Lorentzian.

7.2.2
Magnetic Scattering

Neutrons are scattered by the unpaired electrons of atoms. The double differential magnetic cross section can be written as (Lovesey, 1984)

$$\frac{d^2\sigma}{d\Omega dE_f} = \hbar^{-1} \left(\frac{k_f}{k_i}\right) \cdot a_m{}^2 S_{\text{mag}}(\mathbf{Q}, \omega) \quad (7.26)$$

The dynamic structure factor is given by

$$S_{\text{mag}}(\mathbf{Q}, \omega) = \frac{1}{2\pi} \int dt \, \exp(-\mathbf{i}\omega t) \times < \mathbf{M}^*{}_\perp(-\mathbf{Q}, 0) \cdot \mathbf{M}_\perp(\mathbf{Q}, t) > \quad (7.27)$$

The magnetic interaction operator $\mathbf{M}(\mathbf{Q}, t)$ is the Fourier transform of the total magnetization density $\mathbf{M}(\mathbf{r}, t)$. Therefore, neutrons observe the sum of spin and orbit contributions. Only magnetic moments perpendicular to $\mathbf{Q}$ are measured. This is expressed by

$$\mathbf{M}_\perp = \mathbf{Q} \times (\mathbf{M} \times \mathbf{Q}) \quad (7.28)$$

rather than by **M** occurring in Eq. (7.27). Magnons are described by $S_{\text{mag}}(\mathbf{Q}, \omega)$

7.3
Techniques in Neutron Scattering

7.3.1
Neutron Sources

Neutron scattering suffers from low data rates, particularly in comparison with X-rays or infrared spectroscopy.

Much effort has therefore been concentrated on increasing the fluxes of neutron sources. Presently, there exist two types of high-flux neutron sources.

7.3.1.1 Nuclear Fission Reactors

The principle is based on the following nuclear reaction:

$$\text{n}_{\text{thermal}} + \text{U}^{235} \longrightarrow 2 \text{ fission fragments} + 2.5\,\text{n}_{\text{fast}} + 180\text{ MeV} \tag{7.29}$$

Here, the HFR of the ILL in France (ILL) is the world's best steady state reactor specifically designed for neutron scattering experiments. Its thermal flux near the reactor core is 1.5×10^{15} neutrons $\text{s}^{-1}\,\text{cm}^{-2}$. The reactor became critical in 1972. Since then no research reactor with a higher flux has become operational. There have been attempts to develop nuclear research reactors producing a significantly higher flux (ANS at Oak Ridge National Laboratory (ORNL), USA in the 1980s), but they were not built. The reasons are manyfold: very expensive to build and to operate, problems with radiation damage and heat removal (very short lifetimes of reactor components). The reactor FRMII similar to the ILL reactor with a flux about 50% of the latter has become critical in Munich (FRMII) recently.

7.3.1.2 Spallation Sources

Spallation is a common nuclear reaction occurring for high-energy protons bombarding heavy nuclei. The principle is based on the following nuclear reaction:

$$p + \text{heavy nucleus} \longrightarrow \text{spallation fragments} + n + 24\text{ MeV} \tag{7.30}$$

An important limiting factor in determining the maximum neutron output of a particular type of source is the rate of removal of the heat deposited in the target by the nuclear reaction. The energy released per useful neutron is about 3 times smaller in spallation compared to fission. If, in addition, the spallation source is pulsed, the heat generated in the target will again be reduced. Therefore, a modern pulsed high-flux spallation source like the future project of the European Spallation Source (ESS) can, in principle, achieve a peak flux 40 times higher than the average steady flux of the ILL reactor.

Until recently, the spallation source ISIS at the Rutherford laboratory in United Kingdom can be considered as the world leading pulsed neutron source. It has a cold *peak* flux similar to the continuous flux of the ILL reactor.

As TOF is used on the ISIS spectrometers, this means that the data rate on instruments on both sources will be similar. ISIS however gains in the high-energy epi-thermal range. With the event of the spallation source SNS becoming operational at ORNL, the situation will change: in the final state this new source should produce a significantly higher flux than ISIS. However, one has to wait until all the technical problems are solved to be able to achieve the nominal design flux of the SNS.

For the sake of completeness, we mention the Swiss Neutron Spallation Source

(SINQ), which is a *continuous* spallation source at the Paul Scherrer Institute (PSI).

7.3.2
Moderators

Both types of neutron sources produce high-energy neutrons in the megaelectronvolt range. To become useful in neutron scattering, these neutrons have to be slowed down by so-called moderators. As neutrons have a mass very similar to the mass of the proton, they can be slowed down very efficiently by interaction with the latter in hydrogen or deuterium containing materials like H_2O or D_2O. Therefore, light or heavy water at room temperature are used as moderators to produce "thermal" neutrons with energies around 20 meV. Liquid H_2 or D_2 at 20 K is used to produce cold neutrons with an energy around 2 meV. Hot neutrons with 150 meV are moderated with graphite at very high temperatures of about 2000 K.

7.3.3
Neutron Optics

Neutron scattering is an intensity-limited technique: even on a high-flux reactor like the HFR at the ILL the incident flux of a high-resolution back-scattering (BS) spectrometer like the IN10 is down to 10^4 s^{-1} cm^{-2}. All possible tricks in neutron optics must therefore be played to optimize the flux of the instruments. There exist quite a number of neutron optical beam definition devices (Anderson *et al.*, 1999). In the following, we describe the more important ones.

Neutron guides: they permit to transport neutron beams over long distances without significant loss of the order of 1% m^{-1}. This makes it possible to gain space around a nuclear reactor or to perform high-resolution TOF experiments using very long (100 m or more) distances. The implementation of a group of thermal and cold neutron guides at the early stage of construction of the ILL facility is part of the ILL's success story. The principle of neutron guides is a mirror reflection of neutrons. The refractive index n for neutrons in matter is close to but smaller than one:

$$n = 1 - \frac{\lambda^2 N_v a_{\mathrm{coh}}}{2\pi} \tag{7.31}$$

N_v is the number density of nuclei in the sample. The glancing angle for cold neutrons is of the order of 1°. It increases with increasing wavelength.

Good neutrons guides are made of materials with high N_v and a_{coh} values. Normal neutron guides use mirrors made of channels of glass plates with or without Nickel coating. In order to increase the transmitted flux, currently super-mirror coatings are used. Super mirrors are made out of a sequence of alternating coatings of varying thickness. In this way, the glancing angle can be increased substantially. Flux gain factors of up to one order of magnitude become reality.

Soller collimators: These are mechanical devices that permit a divergent neutron beam to be collimated. They are composed of a number of equidistant neutron-absorbing thin blades separated by spacers.

Choppers: These enable a pulsed neutron beam to be produced. This is either achieved by a rotating mechanical device like a disc with windows or for polarized neutrons with an electromagnetic device.

A Fermi chopper is basically made out of a curved Soller collimator that rotates around an axis perpendicular to the neutron beam.

Crystal monochromators: Using Bragg reflection of a collimated white neutron beam on a crystal, we can obtain a monochromatic beam with wavelength λ:

$$\lambda = 2d_{hkl}\ \sin\theta_B \tag{7.32}$$

d_{hkl} and θ_B are the d spacing of the lattice plane *(hkl)* and the Bragg angle.

Two types of crystals are used: mosaic crystals and perfect crystals. The former (Cu, Be, pyrolythic graphite) are the main components of three-axis spectrometers, the latter (Si, CaF_2) are used on BS instruments.

Focusing devices: These are typically curved systems like crystal arrays on a cylindrical or spherical surface. They allow spatial and monochromatic focusing of a neutron beam. Again, large area (20 m^2) crystal arrays are needed in some cases (see Section 7.4.3).

Polycrystalline filters: They are used to remove unwanted radiation from the beam while maintaining as high a transmission as possible for neutrons of the required energy.

Velocity selectors: These are fast rotating mechanical devices with a helical neutron path. They allow the preparation of a monochromatic beam of cold and thermal neutrons with a typical band width of 15% in wavelength.

Polarizers: A neutron beam can be polarized by reflection from a super-mirror made of alternating magnetic and nonmagnetic coatings or by Bragg reflection from a magnetic crystal like a Heusler alloy.

Spin analyzers: The same technique is used as for the polarizers. However, spin analyzers covering a large solid angle are realized by devices filled with polarized He^3 gas.

Spin orientation devices: π and $\pi/2$ Flippers: A flipper action is based on a nonadiabatic transition of a neutron spin through a magnetic field. By using a flat coil (a few millimeters thick) perpendicular to the neutron beam with the appropriate field inside, the polarized neutron can by flipped by π or $\pi/2$.

Precession coils: In Section 7.4.4, about NSE, we discuss that the Fourier time is proportional to the field integral.

Large diameter long solenoids with strong field integrals and homogeneous windings are needed.

Fresnel coils: These are special devices used in NSE to correct for field integral inhomogeneities. Current loops with a spiral structure where the radius changes as $r \sim \phi$ are the appropriate devices.

Doppler drives: This mechanical or electromechanical device makes it possible to modulate the velocity of Bragg reflected neutrons via the longitudinal neutron Doppler effect. They are used on BS spectrometers. Velocity variations up to a few meters per second are possible at present.

Neutron detectors: The principle of neutron detection is based on neutron absorption by nuclei. Gas detectors use He^3 or $B^{10}F_3$ gas in an electric field. Scintillator detectors use Li^6 in some solid Li containing compound. Here, the photon output triggered by neutron absorption is detected by charge-coupled devices (CCDs) or photomultipliers. The challenge is to develop multidetectors that cover very large areas up to tenths of square meters.

7.3.4
Neutron Scattering Instruments

In the early days of neutron scattering, instruments were invented by the pioneers like Shull, Wollan and Koehler (1951) for diffractometers, Brockhouse (1961) for three-axis spectrometers, Maier-Leibnitz (1966) for neutron optics and Mezei (1972) for the NSE technique. Today, powerful software tools are available to optimize neutron scattering instruments consisting of the components mentioned above. We distinguish between two types of instruments.

7.3.4.1 Instruments without Energy Analysis

With these instruments, one measures the angular distribution of neutrons scattered by the sample. The incident neutron energy is defined by a crystal monochromator, a velocity selector, or by TOF. Three different types of instruments can be distinguished.

Diffractometers measure the scattered intensity over a wide range of scattering angles either simultaneously or in specific scans. The main domain of application is crystallography, the study of atomic and magnetic structures of condensed matter including liquids (Bacon, 1962; Kisi and Howard, 2008).

Small-angle neutron scattering (SANS) instruments (Williams, May and Guinier, 1994) measure the scattering of neutrons at small angles. This makes it possible to study objects and structures in the range of about 10–1000 Å at low resolution. Typical examples are microstructures in alloys and global structures of very large molecules (polymers, biological cells). The fact that the two hydrogen isotopes H and D have scattering lengths of opposite sign allows one to use isotope labeling and contrast matching methods to make visible or invisible parts of the objects under study.

In reflectometers (Daillant *et al.*, 1999) a cold neutron beam is reflected from the sample as from a mirror. The reflected intensity (specular or nonspecular) is detected at small angles. A reflectometer is therefore similar to a SANS instrument.

Neutron reflectometry is the technique for the study of planar structures with a wide range of materials from magnetic multilayers to biological systems at the solid–liquid interface.

7.3.4.2 Instruments with Energy Analysis

Instruments with energy analysis cover the domain that we call *neutron spectroscopy*. We distinguish two types of experiments:

INS with $\hbar\omega$# 0: All kinds of periodic motions (collective and (or) single particle) give rise to INS.

QENS: Here, one measures the broadening of the "elastic" line around $\hbar\omega \sim 0$. All kinds of diffusive motions give rise to QENS (Springer and Lechner, 2005).

7.4
Neutron Spectroscopy: Techniques and Applications

7.4.1
Time-of-Flight (TOF) Spectroscopy

In the TOF technique, a burst of polychromatic neutrons is produced and the times taken by the neutrons to travel from the source of the burst to the detector are measured (Windsor, 1981). TOF spectrometers can be divided into two classes: direct and indirect geometry instruments. The accessible regions in $(\mathbf{Q}, \omega)$ space are different for both geometries. They can be calculated using the equations governed by the conservation laws of momentum and energy transfer (Eqs. (7.2) and (7.3)). They are shown in Figure 7.4.

7.4.1.1 Direct Geometry Spectrometers

In direct geometry spectrometers (Figure 7.4(a)), the incident pulsed beam is monochromatized by a device like a crystal or a chopper system. The final energy E_f is determined by TOF. The scattered neutrons are detected by a detector bank covering (if possible) the whole range of scattering angles with a large vertical aperture. QENS and neutron energy gain spectroscopy are the main applications.

Examples are the TOF spectrometers IN4 and BRISP for thermal neutrons and IN5, IN6 for cold neutrons at the ILL. IN5 is a high-precision instrument used to study low-energy transfer processes as a function of momentum transfer Q.

As shown in Figure 7.5, IN5 has an array of four disk choppers spinning at up to 20 000 revolutions per minute to pulse and to monochromatize the incident beam. Wavelengths from 1.8 to 20 Å are available yielding elastic energy resolutions between 1 meV and 1 μeV. IN5 is a very flexible instrument.

Applications include

- diffusion in disordered systems such as liquids, molecular crystals, amorphous solids (glasses), polymers, hydrogen-metal systems, super-ionic conductors;
- dynamics of soft matter including gels, polymers, proteins, and biological membranes;
- rotational tunneling in molecular crystals;
- crystal field splitting;
- spin dynamics in High-T_c superconductors;
- critical scattering phenomena in dense gases and solids;
- spin dynamics of molecular spin clusters.

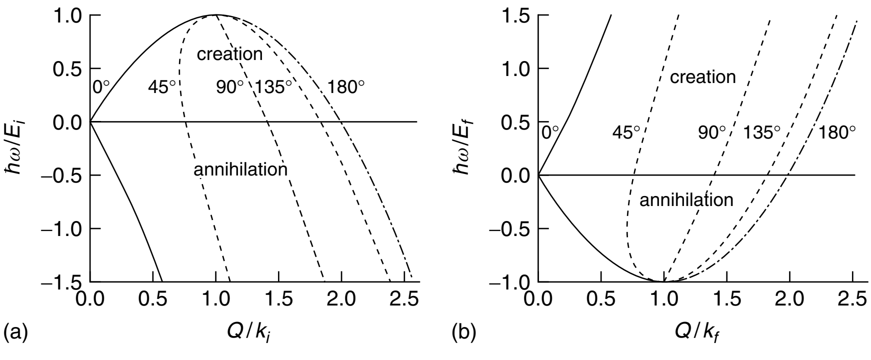

Fig. 7.4 Scattering parabolas for different scattering angles. (a) k_i = constant (b) k_f = constant (Hippert *et al.*, 2005).

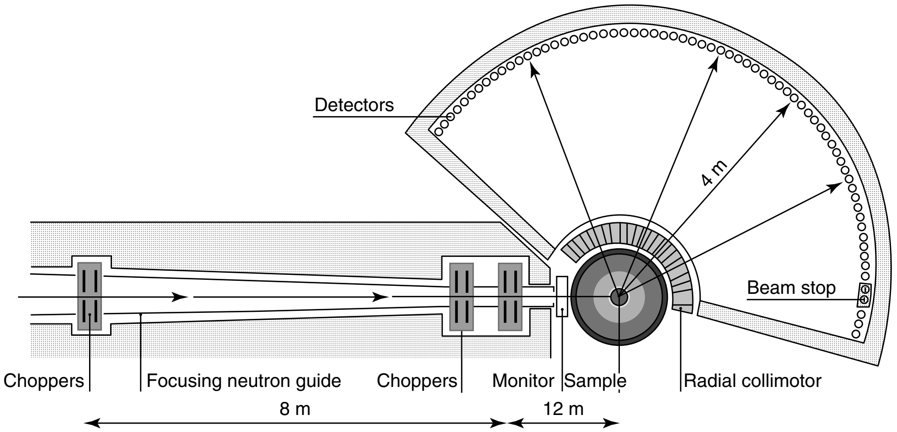

Fig. 7.5 Layout of the direct geometry TOF spectrometer IN5 at the ILL.

A beautiful example of a result obtained on IN5 from magnetic excitations in the molecular spin cluster in Mn_{12}-acetate is shown in Figure 7.6. Mn_{12}-acetate contains ferrimagnetic spin clusters with a total spin $S = 10$. The magnetic ground-state splitting is determined by magnetic INS (Mirebeau *et al.*, 1999). Figure 7.6 clearly shows the effect of temperature on the population of the magnetic sublevels.

The TOF spectrometer IN6 at the ILL employs a vertically curved triplet pyrolythic graphite crystal together with a Fermi chopper. Time focusing is used here to improve the energy resolution. Incident wavelengths between 4 and 6 Å are available with elastic resolutions from 170 to 50 μeV. The strength of IN6 is the very high sensitivity and high data rate. BRISP is a thermal neutron *Brillouin* scattering spectrometer optimized to operate at *small* scattering angles. Applications include coherent magnetic excitations and density fluctuations of disordered systems at low Q.

IN4 is a TOF spectrometer for thermal neutrons used to study mainly single-particle excitations in condensed matter. Applications include the study of energy level spacing of magnetic ions, crystal field splittings, valence fluctuations, heavy fermions, vibrational states in amorphous solids and polycrystals, molecular excitations, and matrix isolated systems.

Similar spectrometers exist at the spallation source ISIS at the Rutherford laboratory such as HET (High-Energy Transfer), MARI, and MAPS. Investigations on HET have broadened from studies of high-energy magnetic excitations and the dynamics of hydrogen-metal systems into the fields of quantum magnetism, non-Fermi liquids, and Brillouin scattering.

MAPS has changed the way the neutron community thinks about inelastic neutron scattering. Its huge array of position sensitive detectors has created a survey technique that is able to map vast areas of the Brillouin zone, making it possible to see the unexpected. It is able to reveal broad features that could be dismissed as background on a three-axis spectrometer.

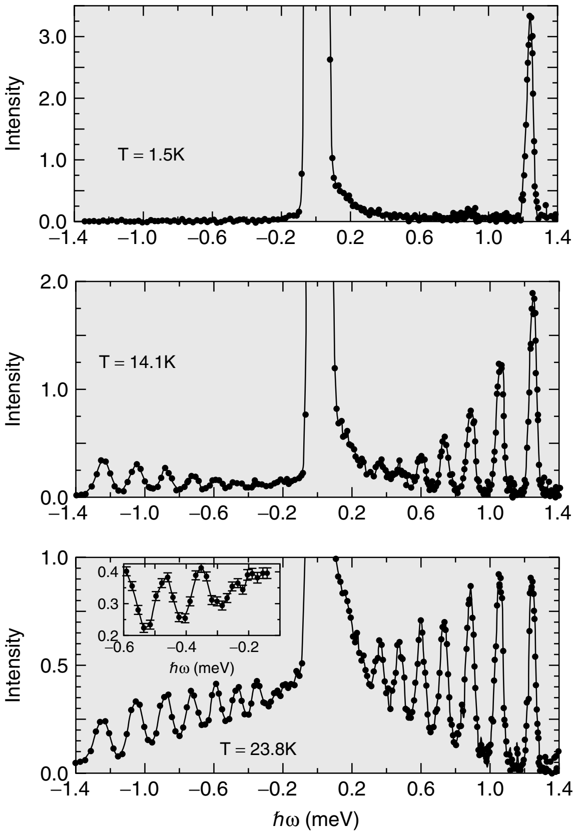

Fig. 7.6 Energy spectra from the Mn_{12}-acetate molecular spin cluster obtained on IN5 (Mirebeau *et al.*, 1999).

The position-sensitive detectors give near-continuous coverage over a large solid angle detector array in the forward direction. The pixel size in reciprocal space is significantly smaller than the resolution volume defined by the other instrumental contributions. In contrast, conventional detectors on HET and MARI integrate along one direction in reciprocal space, which overwhelms the intrinsic resolution in that direction. With MAPS, there is complete freedom to construct scans along any direction in reciprocal space and project data onto any plane in reciprocal space.

MAPS is optimized to measure collective high-energy magnetic excitations in single crystals with varying energy resolution depending on choice of the monochromating chopper.

7.4.1.2 **Indirect (Inverted) Geometry Spectrometers**

In indirect (inverted) geometry spectrometers, an incident white pulsed beam hits the sample (Figure 7.4b). The final energy is defined by a crystal or a filter analyzer. The initial energy is determined by TOF. Again one tries to cover a large solid angle either by using

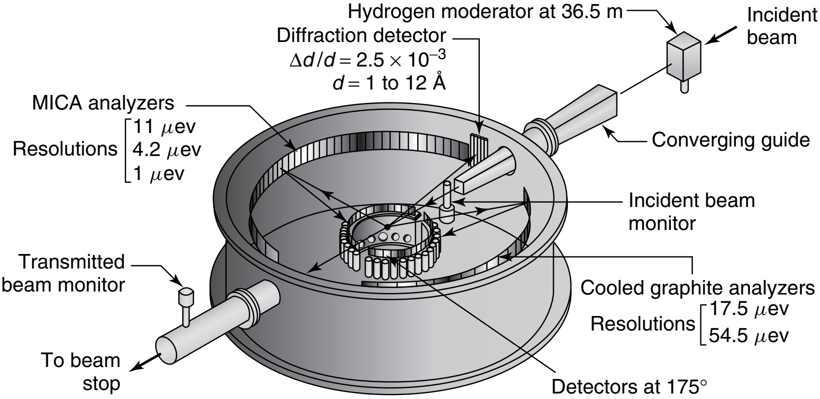
Hydrogen moderator at 36.5 m
Incident beam
Diffraction detector
Δd/d = 2.5 × 10⁻³
d = 1 to 12 Å
MICA analyzers
Resolutions
11 μev
4.2 μev
1 μev
Converging guide
Incident beam monitor
Transmitted beam monitor
Cooled graphite analyzers
Resolutions
17.5 μev
54.5 μev
To beam stop
Detectors at 175°

Fig. 7.7 The IRIS spectrometer at ISIS.

many analyzer crystals simultaneously or by a crystal filter. QENS and neutron scattering with energy loss are the main applications. This kind of geometry is rarely used on a continuous neutron source.

Examples on the pulsed source ISIS are the spectrometers TOSCA, PRISMA, and IRIS (Figure 7.7). PRISMA can perform scans along a defined direction in **Q** space with its multiarm analyzer system.

IRIS is a high-resolution TOF BS spectrometer: the pulsed white beam travels through a 30-m long neutron guide before hitting the sample. The scattered beam is analyzed in energy and momentum transfer Q by a large solid angle, spherically curved crystal analyzer in near-BS geometry. The long flight path allows the incident energy to be defined with the same precision as the analyzer defines the final energy, that is, $\Delta E/E =$ 1% or 15 μeV at $\hbar\omega = 0$. Neutron energy loss spectroscopy down to 7 meV is possible. A similar instrument is in the commissioning phase at the SNS. It will have an elastic energy resolution of about 2 μeV with a final energy of 2 meV.

Neutron BS is described in more detail in Section 7.4.3.

7.4.2
Three-Axis Spectroscopy (TAS)

7.4.2.1 Principle

In order to obtain information on the excitation spectrum at a given point **Q** in reciprocal space, a more sophisticated procedure has to be adopted than just scanning the energy of the scattered neutrons. At each point of the scan the scattering triangle is modified, so that $\mathbf{k}_\mathrm{i}$ and $\mathbf{k}_\mathrm{f}$ will close at the same wave vector transfer **Q** but the length of one of the vectors is varied to provide for the required energy transfer. We define

$$\mathbf{Q} = \boldsymbol{\tau}_{\mathbf{hkl}} + \mathbf{q} \tag{7.33}$$

where $\boldsymbol{\tau}_{\mathbf{hkl}}$ is the reciprocal lattice vector and **q** the reduced wave vector within the Brillouin zone. Therefore, in addition to the $\mathbf{k}_i$ and $\mathbf{k}_f$ modification, all the angles of the scattering triangle will also change.

There are two ways to achieve this. In the TOF technique, all energy transfers are measured simultaneously in all scattering directions, and later, offline,

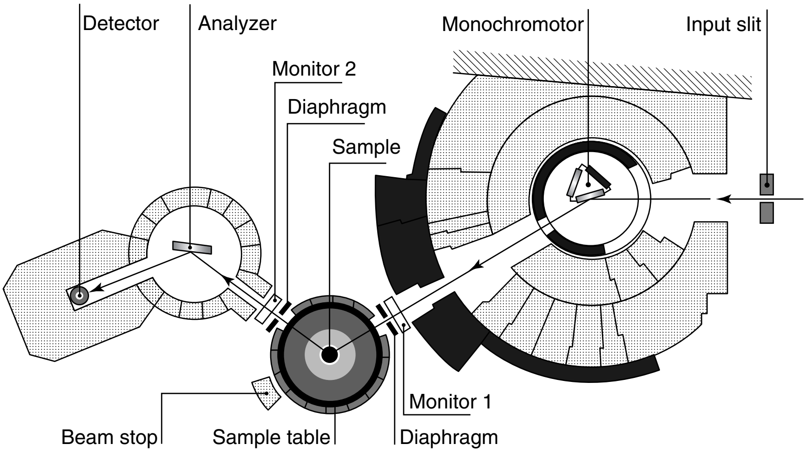

Fig. 7.8 Layout of the TAS spectrometer IN8 at the ILL.

one selects from the huge amount of data only those points that contain data in a specific **Q** direction. This procedure became feasible only with the rapid progress in computing technology. In the three-axis spectroscopy (TAS) technique, one takes data directly and only in a specific **Q** direction. There is a long ongoing discussion on which is the "better" technique. Currently, it is clear that both techniques work with their intrinsic advantages and drawbacks. At the end of this chapter, we try to compare the two.

7.4.2.2 **Layout**

A typical setup of a TAS spectrometer is schematically displayed in Figure 7.8.

The incident and scattered neutron wave vectors, $\mathbf{k}_i$ and $\mathbf{k}_f$, are selected by Bragg reflection from the monochromator and analyzer crystals, respectively. Large crystals of Cu, Zn, or pyrolytic graphite (PG) with a mosaic width of 20–30 minutes have been used in combination with Soller collimators defining the beam divergence of the incident and diffracted beam. On modern instruments, the mosaic crystals are segmented into plates of a few centimeters in size and mounted on mechanical devices (benders), which permit to control their individual orientation in order to act as curved mirrors and to focus the neutron beam horizontally and/or vertically on the sample and detector. More recently, elastically bent perfect Si or Ge crystals, offering improved focusing properties, thanks to the absence of random mosaic block misorientations, have been employed for work requiring higher resolution. The monochromatic incident flux is monitored by a low-efficiency detector called *monitor 1*. Another similar monitor 2 is placed in the scattered beam to detect the possible presence of strong Bragg peaks, which might give rise to spurious signals in the inelastic spectra. The detector is usually a single He^3 proportional gas tube. The monochromator and analyzer crystals diffract, together with the nominal wavelength λ, also its harmonics, $\lambda/2$ $\alpha\nu\delta$ $\lambda/3$, which may be at the origin of spurious effects. They can be eliminated by appropriate filters. Typical examples of state-of-the-art spectrometers are IN1,

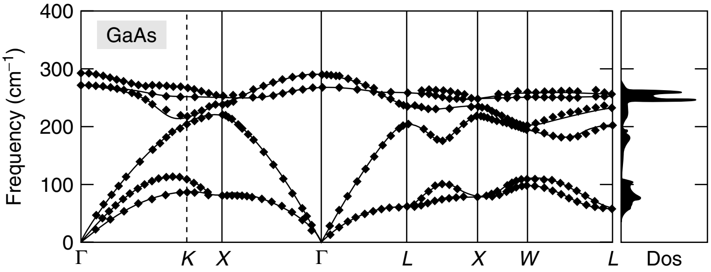

Fig. 7.9 Phonon dispersion curves for GaAs at 12 K. The calculated density of states is shown on the right.

IN8, IN14, and IN20 at the ILL. IN1 is a TAS spectrometer for hot neutrons, IN8 and IN20 for thermal neutrons, and IN14 for cold neutrons. All of them can also use polarized neutrons with or without polarization analysis.

7.4.2.3 The TAS Resolution Function

The resolution volume, that is the region of the ($\mathbf{Q}, \omega$) space, which corresponds to the momentum end energy transfers of neutrons arriving in the detector in a given spectrometer configuration, can be approximated by a four-dimensional ellipsoid. This resolution volume can be calculated analytically in the Gauss approximation or, more recently, numerically by powerful Monte-Carlo methods with programs like RESTRAX (Saroun and Kulda, 1997) or McStas (Lefmann *et al.*, 2000).

7.4.2.4 Applications

The main application of TAS spectroscopy is the determination of collective excitations like phonons and magnons by coherent scattering on single crystalline samples. The goal is to measure the dispersion relations $\hbar\omega(\mathbf{q})$ in single crystals, where $\hbar\omega$ is the energy and $\mathbf{q}$ the wave vector of the excitation. Dispersion-less excitations, where $\hbar\omega$ does not depend on $\mathbf{q}$, originating from single-particle excitations are better determined by TOF spectroscopy.

Figure 7.9 shows a comparison between measured (Strauch and Dorner, 1990) and calculated (Giannozzi *et al.*, 1991) phonon dispersion curves in GaAs. The dispersion curves consist of three acoustic and three optical branches.

7.4.2.5 Comparison between TAS and TOF Spectroscopy of Coherent Excitations in Single Crystals

- **Advantage of TAS:** The energy of the excitation is determined at a selected point in reciprocal space and the scan is carried out in a single well-defined direction chosen by the experimentalist.

- **Disadvantages of TAS:** Spurious peaks may be produced by higher harmonic Bragg peaks from the monochromator and analyzer crystals.

  Unexpected excitations may be missed.

- **Advantage of TOF:** An overview over the excitations of a more or less substantial fraction of reciprocal space is obtained in one run. So unexpected features may be detected.

- **Disadvantage of TOF:** The dispersion relation in a specific $\mathbf{q}$-direction has to be extracted by the computer out of a huge amount of data.

Globally, one could say that the two techniques are complementary to each other. Further information about TAS spectroscopy can be found in (Shirane, Shapiro and Tranquada, 2002).

7.4.3
Back-Scattering (BS) Spectroscopy

7.4.3.1 Principle

Like three-axis instruments, BS spectrometers use single crystals as monochromators and analyzers, but with *fixed* Bragg angles equal to or close to 90° (Birr, Heidemann and Alefeld, 1971). In this way, a very high energy resolution of the order of microelectronvolts is achieved. This can be shown by differentiating the Bragg equation:

$$\frac{\Delta E}{E} = 2\frac{\Delta \lambda}{\lambda} = 2\left(\frac{\Delta d}{d} + ctg\theta\ \Delta\theta\right) \tag{7.34}$$

For a Bragg angle of 90°, the geometry term becomes zero in first order. $\Delta d/d$ describes a lattice parameter variation. In perfect crystals, it is determined by primary extinction and has typical values around 10^{-5}:

$$\frac{\Delta d}{d} \equiv \frac{\Delta \tau}{\tau} = \frac{16\pi\ F_\tau N_0}{\tau^2} \tag{7.35}$$

Here, F_τ is the structure factor of the reciprocal lattice vector $\boldsymbol{\tau}$, N_0 is the number density of the unit cells.

For a perfect silicon crystal with (111) orientation, one calculates an energy width $\Delta E = 2E\Delta\tau/\tau = 0.077\ \mu\text{eV}$. On a real BS spectrometer, the total energy resolution is calculated by the convolution of the monochromator and the analyzer contribution including the geometry term in second order.

For $\theta_B = 90°$ the latter is given by

$$\left(\frac{\Delta k}{k}\right)_{\text{geom}} = \frac{(\Delta\theta)^2}{8} \tag{7.36}$$

This yields a typical value for the energy resolution $\Delta E = 0.2\ \mu\text{eV}$ (full width at half maximum (FWHM)) at an incident energy $E_i = 2$ meV for the case of a BS spectrometer using silicon crystals in (111) orientation in good agreement with experimental data.

As the Bragg angle is fixed, the energy scan can only be performed by varying the length of $\mathbf{k}_i$ either by moving the monochromator periodically parallel to the lattice vector $\boldsymbol{\tau}$ using the longitudinal Doppler effect or by scanning the lattice parameter d of the monochromator by temperature variation. In the former case, Doppler velocities v_{doppler} of a few meters per second yield energy transfers up to 50 μeV for 2-meV neutrons:

$$\delta E_{\text{doppler}} = 2E_i \frac{v_{\text{doppler}}}{v_i} \tag{7.37}$$

Equation (7.37) is valid in first-order $v_{\text{doppler}} \ll v_i$.

In the latter case, temperature variations up to 300 °K on a CaF_2 crystal with (422) orientation yield energy transfers up to 200 μeV for thermal neutrons.

Because of the very high energy resolution, the incident flux of a BS spectrometer is low. To overcome this problem, one uses a large spherically curved multicrystal analyzer covering the whole scattering angular range with a huge solid angle. Analyzer surface areas up to 20 m^2 are not unusual. The analyzed neutrons are detected by a multidetector.

The Q resolution of a BS spectrometer is dominated by the incident and scattered beam divergence. The former depends on the geometry of the primary spectrometer, the latter on the geometry of the crystal analyzer/multidetector. The compromise between intensity and Q resolution leads to values of the latter between 5 and 10%. The problem of separating the back-scattered from the incident beam in exact BS can be solved by the use of a deflector crystal and (or) a chopper and TOF discrimination. Higher order contamination can be eliminated by the latter technique at the same time.

7.4.3.2 Existing Instruments

Typical BS instruments are IN10 and IN16 (see Figure 7.10) for cold neutrons and IN13 for thermal neutrons at the ILL.

Similar BS spectrometers exist at NIST and the FRMII reactor. Both use a so-called phase-space transformer (PST), which allows an intensity gain of a factor 2 to 4 (Schelten and Alefeld, 1984). This device transforms a collimated white beam into a divergent monochromatic beam by Bragg reflection from a mosaic crystal moving *parallel* to the scattering plane at high speed. Again the Doppler effect is used. The price to pay is the increased background produced by the proximity of the strong white beam.

The TOF BS spectrometers IRIS at ISIS and BASIS at SNS have been mentioned in section 7.4.1. Here, BS is only used in the secondary spectrometer and TOF in the primary spectrometer; this is the way to go on a pulsed source.

7.4.3.3 Applications

The main application of BS spectroscopy is high-energy resolution incoherent neutron scattering of low-energy single-particle excitations in the

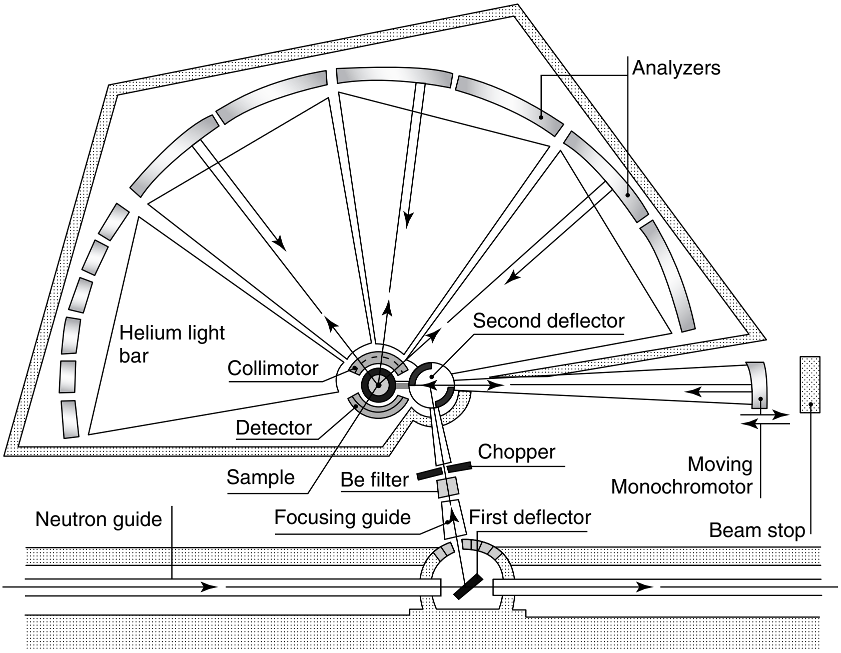
Analyzers
Helium light bar
Collimotor
Second deflector
Detector
Chopper
Sample
Be filter
Moving Monochromotor
Neutron guide
Focusing guide
First deflector
Beam stop

Fig. 7.10 Layout of the BS spectrometer IN16 at the ILL.

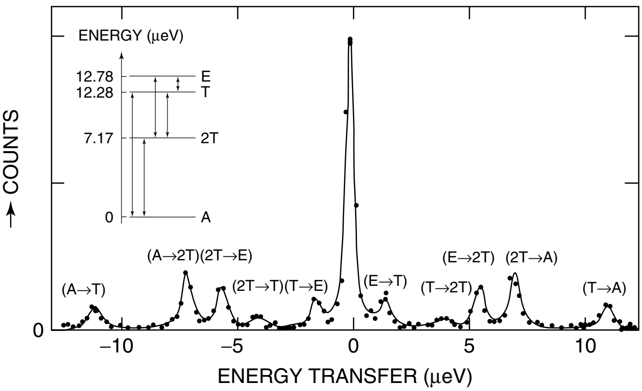

Fig. 7.11 Tunneling spectrum of NH_4^+ ions in NH_4ClO_4 at 4 K (Prager and Alefeld, 1976; Press, 1981).

microelectronvolt or gigahertz range with moderate momentum transfer resolution: tunneling spectroscopy (Prager and Alefeld, 1976; Prager and Heidemann, 1997; Press, 1981), hyperfine interactions (Heidemann, 1972; Word, Heidemann and Richter, 1977; Chatterji and Frick, 2004), slow motions in condensed matter: viscous liquids, glass dynamics (Mezei, Pappas and Gutberlet, 2003), liquid and plastic crystals, soft matter, and self diffusion. For further information, see (Hippert *et al.*, 2005).

Figure 7.11 shows the tunneling spectrum of NH_4^+ ions in NH_4ClO_4 at 4 K measured on IN10. The NH_4^+ ions behave like quantum rotators at low temperatures. Analyzing the energies and intensities of the observed transitions yields the overlap matrix elements (Prager and Alefeld, 1976; Press, 1981).

7.4.4
Neutron Spin Echo (NSE) Spectroscopy

7.4.4.1 Principle

Neutron scattering suffers from the strong correlation between energy resolution and intensity, that is, high-energy resolution is only obtained at the expense of low intensity. In BS spectroscopy, this problem is bypassed by the use of *spatial focusing* of the scattered (and incident) neutron beam resulting in only moderately good resolution in wave vector transfer **Q**.

The NSE technique (Mezei, 1972) allows the energy resolution to be decoupled from the intensity in cases whenever Q resolution is not so important or is determined by the angular resolution rather than the wavelength distribution. This is achieved by a very special "time-focusing" technique using the precession of neutron spins in a magnetic field.

A polarized neutron beam entering a region with a magnetic field B perpendicular to the magnetic moment of the neutrons will undergo precession with the Larmor frequency ω_L:

$$\omega_L = \gamma B \text{ where } \gamma = \frac{2.916\ \text{kHz}}{\text{Oe}} \tag{7.38}$$

The total precession angle after traversing a distance l with velocity v will be $\phi_1 = \omega_L\ l_1/v_1$. If v has a finite distribution, the beam will appear to be depolarized after a short distance. Assume now that this beam hits a sample.

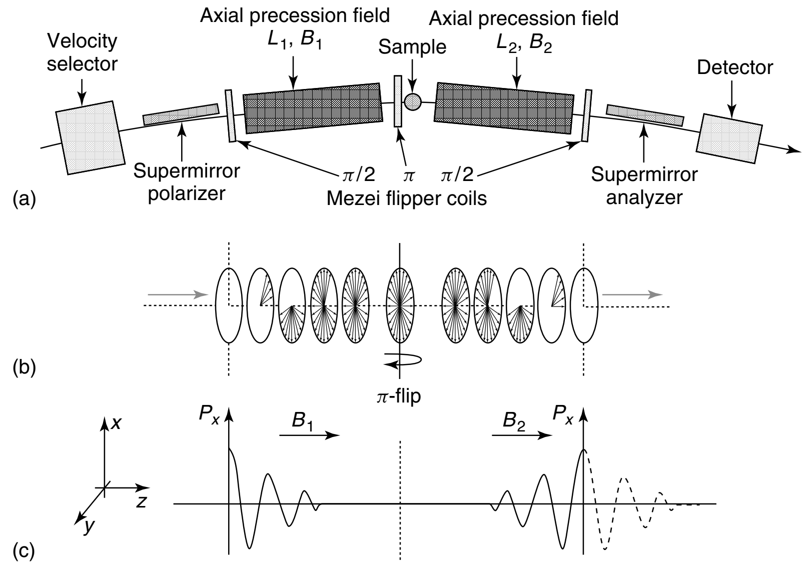
Velocity selector
Axial precession field L_1, B_1
Sample
Axial precession field L_2, B_2
Detector
Supermirror polarizer
$\pi/2$ π $\pi/2$
Mezei flipper coils
Supermirror analyzer
(a)
(b)
π-flip
x
z
y
P_x
B_1
B_2
P_x
(c)

Fig. 7.12 Layout of an NSE spectrometer: (a) principal components, (b) dephasing and rephasing of the neutron spins, (c) x-component of the net neutron polarization (Hippert *et al.*, 2005).

The scattered beam would now travel through another region with an opposite magnetic field $-B$. Then the total precession angle will be

$$\phi_{\text{tot}} = \omega_L \left(\frac{l_1}{v_1} - \frac{l_2}{v_2} \right) \tag{7.39}$$

in the case of elastic scattering and if $l_1 = l_2$, ϕ_{tot} is independent of v and we recover the incident beam polarization.

In the case of a small energy change $\hbar\omega \ll mv^2$ we obtain in first-order approximation,

$$\phi_{\text{tot}} = \left(\frac{\hbar\, \gamma B l}{mv^3} \right) \omega \tag{7.40}$$

Analyzing the polarization P_{final} after the second flight path yields

$$P_{\text{final}}(\mathbf{Q}, t_F) =< \cos\Phi >= \int \cos(\omega\, t_F) \times S(\mathbf{Q}, \omega) d\omega \equiv S(\mathbf{Q}, t_F) \tag{7.41}$$

where $t_F = \hbar\ \gamma B l / m v^3$ is the so-called Fourier time. NSE therefore directly measures the intermediate scattering function. It is important to note that the Fourier time t_F is proportional to $1/v^3$ or to λ^3. So with very cold long-wavelength neutrons the time domain can be extended to the near microseconds range or, in terms of energy transfer, nanoelectronvolt values can be measured.

7.4.4.2 **Implementation**

A velocity selector with a coarse wavelength resolution of the order of 15% selects a poorly monochromatic beam, which is polarized (Figure 7.12). After passing a $\pi/2$ flipper, the beam starts precessing in a solenoid magnetic field, traverses a π flipper and hits the sample. The scattered beam precesses in the second solenoid field in the opposite direction, traverses a second $\pi/2$ flipper, enters an analyzer, and is finally detected with a counter.

This layout is used by the NSE spectrometer IN11 at the ILL. A typical field of applications of NSE spectroscopy is coherent quasi-elastic small-angle scattering. An example is the slow motion polymer dynamics in the megahertz to gigahertz range.

Figure 7.13 shows experimental data of $S(Q, t_f)$ from IN15 for polyethylene together with theoretical predictions. Solid lines represent the model of reptation by de Gennes (1981). Note that the Fourier time extends to 200 ns.

7.4.4.3 NSE Variants

A number of variants of this technique exist. The monochromatization by a velocity selector can be replaced by a TOF analysis of a white neutron beam as implemented on IN15 at the ILL. The NSE tool can also be used as an option on a TAS spectrometer with polarization analysis (TASSE) (Mezei, 1980). Here the application is the measurement of the line width of elementary excitations without and with dispersion. For the latter case, special tricks have to be played in order to turn the ($\mathbf{q}$, ω) ellipsoid by tilting the angles of the precession fields. The IN20 spectrometer at the ILL allows use of this technique. Another variant is the so-called resonance or zero field neutron spin echo (ZFNSE). It was invented by Gähler and Golub (1987).

Principle The neutron precession in the solenoids is replaced by the passage of the neutrons through a strong high-frequency electromagnetic field. This technique is a promising option for high-resolution inelastic scattering on TAS spectrometers. A dedicated instrument of this kind, the TRISP spectrometer, became operational recently at the FR-MII reactor.

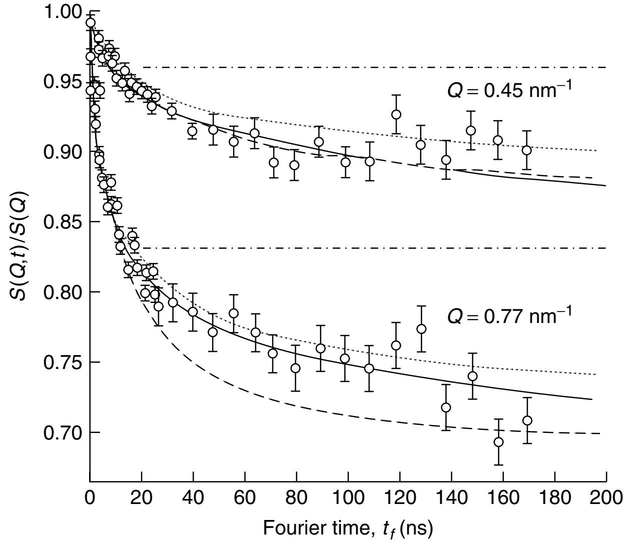

Fig. 7.13 Plot of $S(Q, t_f)$ data for polyethylene measured on IN15 (Hippert *et al.*, 2005).

7.4.4.4 NSE versus Standard Inelastic Spectrometers

On all instruments working in $(\mathbf{Q}, \omega)$ space, the instrumental resolution function has to be deconvoluted from the measured spectrum to obtain the pure sample response $S(\mathbf{Q}, \omega)$. In $(\mathbf{Q}, t)$ space, the deconvolution becomes a simple division. This is an obvious advantage. Samples containing nuclei that scatter spin incoherently such as H, V, Co are not good candidates for NSE because they decrease the scattered polarization by a factor 1/3. Incoherent scattering, in general, is better investigated on spectrometers working in $(\mathbf{Q}, \omega)$ space. A sample with a strong elastic peak and weak inelastic excitations close to the elastic line is also not a good candidate for NSE: the resulting $S(t)$ curve would be a cosine oscillation with a small amplitude sitting on a high constant background. Good samples for NSE are those giving rise to a strong quasi-elastic signal at small momentum transfer, as from deuterated polymer solutions or from critical scattering near a phase transition. For further information, see (Mezei, Pappas and Gutberlat, 2003).

7.5
Conclusion

We have seen that neutron spectroscopy covers a large range of applications. This is again summarized in Figure 7.14, which shows the applications covered by neutron spectroscopy in the (Q, ω) plane. The region inside the trapezoidal box is

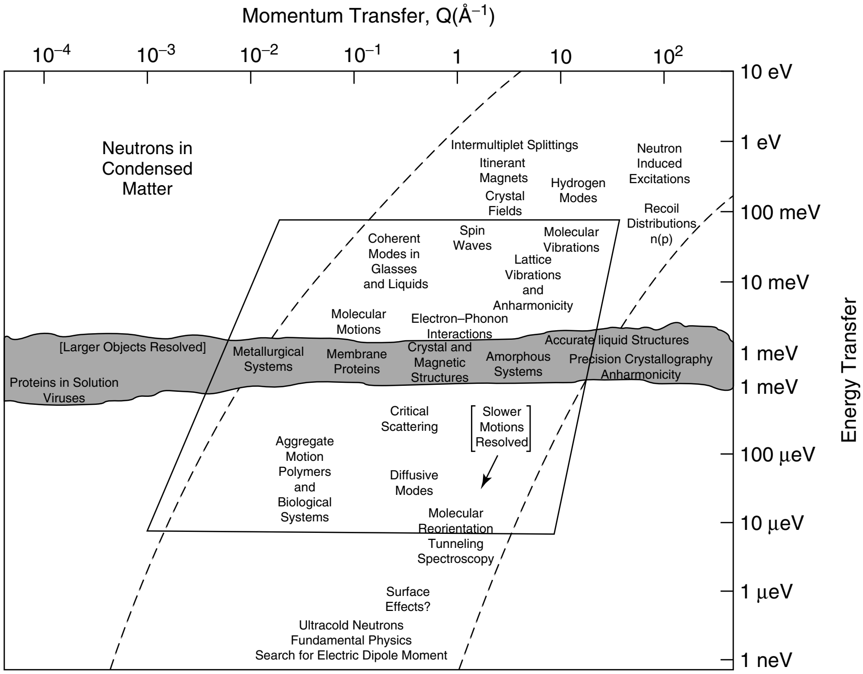
Momentum Transfer, Q(Å^{-1})
10^{-4}
10^{-3}
10^{-2}
10^{-1}
1
10
10^{2}
10 eV
1 eV
100 meV
10 meV
1 meV
1 meV
100 μeV
10 μeV
1 μeV
1 neV
Energy Transfer
Neutrons in Condensed Matter
Intermultiplet Splittings
Itinerant Magnets
Crystal Fields
Hydrogen Modes
Neutron Induced Excitations
Recoil Distributions n(p)
Coherent Modes in Glasses and Liquids
Spin Waves
Molecular Vibrations
Lattice Vibrations and Anharmonicity
Molecular Motions
Electron–Phonon Interactions
[Larger Objects Resolved]
Metallurgical Systems
Membrane Proteins
Crystal and Magnetic Structures
Amorphous Systems
Accurate liquid Structures
Precision Crystallography
Anharmonicity
Proteins in Solution
Viruses
Critical Scattering
Slower Motions Resolved
Aggregate Motion Polymers and Biological Systems
Diffusive Modes
Molecular Reorientation
Tunneling Spectroscopy
Surface Effects?
Ultracold Neutrons
Fundamental Physics
Search for Electric Dipole Moment

Fig. 7.14 Applications of neutron scattering in condensed matter research (Lander and Emery, 1985).

the range probed by TAS spectroscopy. The shaded region corresponds to an expanded cut in the energy scale (Lander and Emery, 1985).

Acknowledgments

The author is indebted to D. Gray for checking the English.

Glossary

Doppler Drive: Device to change the neutron velocity using Doppler effect.

Fermi Chopper: Mechanical device to pulse and monochromatize a neutron beam.

Nuclear Spallation: Process in which a heavy nucleus fragments into lighter nuclei and neutrons, when hit by highly energetic particles like protons from an accelerator.

Neutron Back-Scattering: Method to achieve very high energy resolution in neutron scattering using Bragg reflection with a Bragg angle of 90°.

Neutron Guide: Device to transport a neutron beam over long distances.

Neutron Three-Axis Spectrometer: Instrument consisting of three axes: monochromator, sample, and analyzer.

Neutron Spin Echo: Method to achieve very high energy resolution in neutron scattering using the Larmor precession of neutrons.

Phase-Space Transformer: Device to transform a white collimated neutron beam into a monochromatic divergent beam using the Doppler effect.

Super-Mirror: Mirror for neutrons with an increased glancing angle using multilayer coating.

Scattering Length: Parameter defining the strength of the neutron sample interaction.

Time-of-Flight Method: Determination of the neutron velocity by measuring its time of flight over a given distance.

Velocity Selector: Mechanical device to monochromatize a neutron beam.

References

Anderson, I.S., Brown, P.J., Carpenter, J.M., Lander, G., Pynn, R., Rowe, J.M., Schärpf, O., Sears, V.F. and Willis, B.T.M. (1999) *International Tables for Crystallography C*, 2nd edn, Kluwer Academic Publishers, p. 426.

Bacon, G.E. (1962), *Neutron Diffraction*, Clarendon Press, Oxford.

Birr, M., Heidemann, A. and Alefeld, B. (1971) *Nucl. Instrum. Methods*, **95**, 435.

Brockhouse, B.N. (1961), *Inelastic Neutron Scattering in Solids and Liquids*, IAEA, p. 113.

Chatterji, T. and Frick, B. (2004) *Solid State Commun.*, **131**, 453.

Daillant, J. and Gibaud, A. (eds) (1999) *X-Ray and Neutron Reflectometry: Principles and Applications*, Springer Verlag.

Fermi, E. (1936) *Ric. Sci. Vicotruiz.*, **7**, part 2, 13.

Gähler, R. and Golub, R. (1987) *Z. Phys.*, **B65**, 269.

de Gennes, P.G. (1981) *J. Physique*, **42**, 73.

Giannozzi, P., Gironcoli, S., Pavone, P. and Baroni, S. (1991) *Phys. Rev.*, **B43**, 7231.

Heidemann, A. (1972) *Z. Phys.*, **238**, 208.

Hippert, F., Geissler, E., Hodeau, J.L., Lelievre-Berna, E. and Regnard, J.R. (eds) (2005) *Neutron and X-Ray Spectroscopy*, Springer Verlag.

van Hove, L. (1954) *Phys. Rev.*, **95**, 249.

Kisi, E.H. and Howard, C.J. (2008) *Applications of Neutron Powder Diffraction*, Oxford Series on Neutron Scattering in

Condensed Matter, vol. 15, Oxford University Press.

Lander, G.H. and Emery, V.J. (1985) *Nucl. Instrum. Methods*, **B 12**, 525.

Lefmann, K., Nielson, K., Tennant, A. and Lake, B. (2000) *Physica B*, **276–278**, 152–153.

Lovesey, S.W. (1984) *Theory of Neutron Scattering from Condensed Matter*, Clarendon Press, Oxford.

Mezei, F. (1972) *Z. Phys.*, **255**, 146.

Mezei, F. (ed.) (1980) *Neutron Spin Echo*, Lecture Notes in Physics, vol. 128, Springer Verlag.

Mezei, F., Pappas, C. and Gutberlet, T. (eds) (2003) *Neutron Spin Echo Spectroscopy: Basics, Trends and Applications*, Lecture Notes in Physics, vol. 601, Springer Verlag.

Mirebeau, I., Hennion, M., Casalta, H., Andres, H., Güdel, H.U., Irodova, A.V. and Caneschi, A. (1999) *Phys. Rev. Lett.*, **83**, 628.

Prager, M. and Alefeld, B. (1976) *J. Chem. Phys.*, **65**, 4927.

Prager, M. and Heidemann, A. (1997) *Chem. Rev.*, **97**, 2933.

Press, W. (1981) *Single-Particle Rotations in Molecular Crystals*, Springer Tracts in Modern Physics, vol. 92.

Price, D.L. and Sköld, K. (1986) *Neutron Scattering*, Methods of Experimental Physics, vol. 23, Academic Press, London, Part **A**, p. 1.

Saroun, J. and Kulda, J. (1997) *Physica B*, **234–236**, 1102.

Schelten, J., Alefeld, B. (1984) Jülich Report.

Scherm, R. (1972) *Ann. Phys.*, **7**, 349.

Shirane, G., Shapiro, S.M. and Tranquada, J.M. (2002) *Neutron Scattering with a Triple-Axis Spectrometer, Basic Techniques*, Cambridge University Press.

Shull, C.G., Wollan, E.O. and Koehler, W.C. (1951) *Phys. Rev.*, **84**, 904.

Springer, T. and Lechner, R.E. (2005) Diffusion studies of solids by quasi-elastic neutron scattering, in *Diffusion in Condensed Matter-Methods, Materials, Models*, (eds P. Heitjans and J. Kärger), Springer Verlag.

Strauch, D. and Dorner, B. (1990) *J. Phys.: Condens. Matter*, **2**, 1457.

Williams, C., May, R.P. and Guinier, A. (1994) Small angle scattering of X-rays and neutrons, in *Characterisation of Materials*, Materials Science and Technology, vol. 2B (ed. E. Lifshin), VCH Verlagsgesellschaft, Weinheim, p. 611.

Windsor, C.G. (1981) *Pulsed Neutron Scattering*, Taylor & Francis, London.

Maier-Leibnitz, H. (1966) *Nucleonik*, **8**, 61.

Word, R., Heidemann, A. and Richter, D. (1977) *Z. Phys.*, **B28**, 23–30.

Web sites

BS review: http://wwwold.ill.fr/YellowBook/IN16/BS-review/index.htm.

ESS: http://neutron.neutron-eu.net/n_ess.

FRMII: http://www.frm2.tu-muenchen.de.

ILL: http://www.ill.eu/.

Cicognani, G., *The Yellow Book: Guide to Neutron Scattering Facilities, ILL.*

Copies can be obtained from: ILL, SCO, BP156, F-38042 Grenoble Cedex 9 (France).

ISIS: http://www.isis.rl.ac.uk/.

NIST: http://rrdjazz.nist.com.

SINQ: http://sinq.web.psi.ch/.

SNS: http://neutrons.ornl.gov/.

NONHOMOGENEOUS FLOWS

H. G. E. Hentschel, *Emory University, Atlanta, Georgia, U.S.A.*

Introduction 379
1. **Surface Flows** 381
1.1 Dynamics of Liquid–Liquid Interfaces 382
1.1.1 The Saffman–Taylor and Taylor Instabilities 382
1.1.2 Flow in Hele–Shaw Cells 383
1.1.3 Laplace's Equation and Diffusion-Limited Fingering 385
1.2 Dynamics of Liquid–Gas Interfaces 386
1.2.1 Surface Gravity Waves 386
1.2.2 Capillary Waves 387
1.3 Flow on Surfaces 388
1.3.1 Wetting, and Capillary Flow 388
1.3.2 Droplet Flow 389
1.4 Surface Evolution and Mixing 390
1.4.1 Deterministic Mixing 390
1.4.2 Mixing by Diffusion 392
2. **Multiphase and Anisotropic Flows** 392
2.1 Liquid–Solid Phase Flows 393
2.1.1 Slurry Flows 393
2.1.2 Flows in Polymer Solutions and Melts 394
2.2 Solid–Gas Phase Flows 395
2.2.1 Flows in Granular Material 395
2.2.2 Granular Topography and Self-Organized Criticality 396
2.3 Liquid–Gas and Liquid–Liquid Phase Flows 397
2.3.1 Bubble and Emulsion Rheology 397
2.3.2 Cavitation 398
2.3.3 Foam Structure and Stability 398
2.3.4 Foam Rheology and Texture Coarsening 399
2.4 Anisotropic Flows 400
2.4.1 Nematic Liquid-Crystal Rheology 400
2.4.2 Dynamics of Electrorheological Fluids 401
3. **Flow in Porous Media** 402
3.1 Microscopic Equations of Motion 402
3.2 Macroscopic Transport in Porous Media 403
3.3 Drainage and Imbibition 404
3.4 Models for Fluid Flow in Porous Media 406
3.4.1 Invasion Percolation 406
3.4.2 Gradient Percolation 407
3.4.3 Macroscopic Langevin Equations 409
3.5 Hydrodynamic Dispersion in Porous Media 410
3.5.1 Dispersion of Conserved Solute Particles 411
3.5.2 Grain Trapping in Porous Media 412
Glossary 412
Works Cited 414
Further Reading 416

INTRODUCTION

Any flow involving more than a single phase—liquid, solid, or gas—may be classified as nonhomogeneous. If the interface between the phases is macroscopic, we are interested in the dynamics of this evolving interface or, more generally, in surface flows—flows in porous media, flow on two-dimensional surfaces, or flow on one-dimensional fibers. Such nonhomogeneous flows are dominated by surface tension and wetting effects. The phenomenon of deterministic mixing by chaotic advection can also be considered a surface evolution phenomenon.

On the other hand, if the interfacial structure between the phases occurs on a microscopic scale, we have nonhomogeneous materials with rheological behavior very different from that encountered in the case of a simple Newtonian fluid. The material can then be considered as a dispersion of several differ-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ent pure phases, such as granular materials (solid in gas), foams (gas bubbles in liquids), slurries (solid in liquid), or emulsions (droplets of one immiscible fluid in another). Orientational ordering at the macroscopic scale can also result in non-Newtonian anisotropic flows, as is the case for nematic liquid crystals or electrorheological fluids.

A special, but technologically very important, case is that of fluid flow in porous media. The explicitly nonuniform boundary conditions in porous media, due to the tortuous connectivity of its pores and throats, result in a wealth of nonhomogeneous flows and interfacial evolution.

In Sec. 1, we consider flows involving macroscopic surfaces. These flows may involve either two-fluid hydrodynamic instabilities, such as the Taylor and Saffman–Taylor instabilities, or interfacial flow phenomena on solid substrates, such as droplet motion on surfaces and fibers. At the interface between liquids and gases, both surface gravity and capillary waves can occur. Surface evolution due to deterministic mixing and diffusion will also be described.

In Sec. 2, we consider the rheological behavior and flow exhibited by a variety of multicomponent systems, including foams, slurries, and granular materials. The flows exhibited by complex molecular materials, such as polymer solutions, and orientationally ordered fluids, such as nematic liquid crystals and electrorheological fluids, are also briefly discussed.

Finally, in Sec. 3, we consider a variety of flows observed in porous media. Topics such as the interfaces observed during drainage of wetting fluids by nonwetting fluids, imbibition of wetting fluids by porous media, and the anomalous transport of passive scalars in porous media will be examined.

Historically, the field of nonhomogeneous flows has been an active area of research since at least the middle of the 19th century. Lamb in his *Hydrodynamics*, originally published in 1879, considered such topics as cavitation, the motion of solids through liquids, and tidal and surface waves. G. I. Taylor contributed greatly to our understanding of surface instabilities and dispersion in porous media in the first half of this century. More recently, de Gennes has emphasized the importance of scaling arguments in the analysis of nonhomogeneous flows. In particular, scaling arguments have been used to extract the relevant time and length scales, as well as the dimensionless parameters describing various nonhomogeneous flow regimes, and, in so doing, have helped to emphasize the relative importance of various physical mechanisms in creating the particular flow. In addition, topics such as flow in porous media and multiphase flows have been greatly enriched by new concepts in statistical mechanics, such as fractals, self-affine surfaces, percolation, and self-organized criticality. Technologically, nonhomogeneous flows have been central to an astounding variety of industrial processes.

In the oil industry, secondary enhanced oil recovery is a major task. When oil wells are first dug, perhaps only 10% of the oil in an oil-rich sediment will flow back under its own pressure. To release the remaining oil deposits, water is often pumped into the porous sediments to displace the more viscous residual oil. This immediately leads to a classic instability in two-phase fluid flow, the Saffman–Taylor instability, where displacing a more viscous fluid (oil) through a porous medium using a less viscous one (water) leads to a fingering interface, which, from a technological viewpoint, reduces the areal sweep efficiency for oil compared to that for a flat interface, resulting in oil being left in the sediment. Typically, even after secondary oil recovery, only one-third of the oil has been recovered.

All processes in which mixing is used involve nonhomogeneous flows. These include materials blending in the polymer, chemical, and food industries. In addition, mixing flows affect boundary-layer drag in pipes, nozzles, and jets and in naval technology, such as in flow near ships' hulls. The technological importance extends well beyond these areas, however, as any process involving combustion will also be concerned with turbulent mixing, as will such fields as bioengineering and physiology, where knowledge of the flow of blood in fluid vessels is required. Indeed, when geological, oceanographic, and astronomic mixing processes are included, such as mixing in the earth's mantle, atmospheric turbulence, and mixing in the cores of stars, 40 orders of magnitude in Reynolds number may be involved.

Wetting of surfaces and fibers constitutes another class of flows—this time on dimensionally reduced domains—that have important technological consequences. Thus, technologies such as lubrication, processing of fibers, droplet creation, and imbibition by porous media are all controlled by the surface tensions and long-range forces at phase interfaces.

Multiphase flows, such as foams, slurries, and granular flows, form the basis of many industries.

Foams, which are liquid–gas mixtures, are used in fire fighting, as well as to transport pulverized coal in pipelines. In the oil industry, as foam has a low mobility in porous media, this property can be used for blocking and diverting oil flows in desired directions, as well as to transport sand in the technology of enhanced oil recovery. Formations are fractured to enhance oil flowback, and these fractures are kept open using proppants such as sand that are carried to the site of action without settling by foams that then degrade, allowing easy cleanout. Foams are also used in the paper industry to disperse pigments and other paper coatings, while, in the polymer industry, liquid foam flow is an important intermediary in the formation of synthetic cellular solids.

Liquid–solid mixtures in the form of slurries and suspensions also have many industrial applications. They appear in the food and paint industries in the form of pulps, inks, paints, coatings, and emulsions, while, in biotechnology, such problems as coagulation of blood, water purification, and the flows of enzymes all depend on the rheology of suspensions.

Granular flows are two-phase solid–gas mixtures, and such flows are intrinsic to the mining and building industries, among others. Tasks such as funneling flows in hoppers, material crushing for buildings and its resulting flows under stresses, and segregation of granular materials of many different sizes all depend on the very unusual rheologies displayed by granular flows.

Flows in porous media varying in scale from geological deposits to membranes are vital to technologies as varied as the water industry and biotechnology. In the field of hydrogeology, the demand for pure water supplies means that it is vital to have knowledge of the manner in which effluents and other environmental pollutants are carried and dispersed by water in porous media. The related field of filtration is required for water purification. For biotechnology, chromatography and electrophoresis are of ever-increasing importance for the separation of large polymeric molecules, and the manner in which such polymers move in porous media in the presence of electric fields will determine sample purity, while semipermeable flows through membranes are especially important for dialysis.

1. SURFACE FLOWS

By surface flows, we mean all flows that involve in one way or another an evolving macroscopic interface between two phases. We shall consider three types of nonhomogeneous flows under this heading: the evolution of interfaces in liquid–liquid immiscible flows; waves at liquid–gas interfaces; and capillary flow of liquids on solid substrates.

In immiscible two-fluid flows, the behavior of this interface is all important, and in those cases where the interface is unstable to small perturbations in curvature, such as is the case when a less viscous fluid is used to drive a more viscous one, an initially flat surface can develop into a complex asymptotic pattern. The final structure is strongly dependent on boundary conditions, and both asymptotic fingering and dendritic branching patterns in the same universality class as two-dimensional diffusion-limited aggregation have been observed.

At liquid–gas interfaces, wave motion is often observed. Wave motion is due to the interaction between an inertial term tending to keep a system moving at a constant velocity when disturbed, and a restoring force tending to bring the system back to equilibrium. The inertial term, however, causes the system to overshoot the equilibrium point, setting up an oscillatory motion. At liquid–gas interfaces, the restoring forces include gravity and surface tension.

Another class of flows involving surfaces comprises those where the flows actually occur on a solid substrate, such as on two-dimensional planes or one-dimensional fibers . In this case, the flow is strongly influenced by surface tension and the wetting properties of the liquid.

1.1 Dynamics of Liquid–Liquid Interfaces

Here we examine the time evolution of interfaces between two immiscible liquids, each liquid segregated into a single simply connected domain of macroscopic scale.

1.1.1 The Saffman–Taylor and Taylor Instabilities How does such an interface between two immiscible fluids develop with time under an applied pressure? The important criterion for surface evolution is stability. If an initially flat interface is stable, then it will remain flat with time, and any small perturbations due to noise will decay. Under certain flow conditions, however, the flat interface can become unstable, and the interface develops with time into a complex curved surface. Two important causes of such instabilities are the Taylor and Saffman–Taylor instabilities.

Consider the experimental situation when one immiscible fluid displaces another in a porous medium. Taylor (1950) showed that the interface can become unstable when a fluid of higher density ρ_2 is forced, usually by gravity, into another fluid of lower density ρ_1 (for example, the interface between a heavy oil lying on top of less dense water is unstable). Saffman and Taylor (1958) pointed out that a related instability occurs when a fluid of lower viscosity η_2 is driven with a velocity U into a fluid of greater viscosity η_1. The flat surface is then replaced by a highly irregular contour with fingerlike projections [see Fig. 1(a)]. Such convoluted interfaces also occur in enhanced oil recovery when water is pumped in at a point [see Fig. 1(b)] into sediments to release the residual oil. The importance of this fingering in enhanced oil recovery is apparent: The fraction of oil recovered, or areal sweep efficiency, is reduced compared to that expected for a flat displacing interface.

The full set of equations describing the Taylor and Saffman–Taylor instabilities are the following:

1. Darcy's law in a gravitational field, which relates the flow rate $\mathbf{v}$ of a liquid of viscosity η to the pressure gradient ∇p in a porous medium of permeability κ (equivalent to Ohm's law for the electric current in a resistor),

$$\nabla p_\alpha = -(\eta_\alpha/\kappa_\alpha)\mathbf{v}_\alpha - g\rho_\alpha \cos\theta\,\mathbf{i}. \tag{1}$$

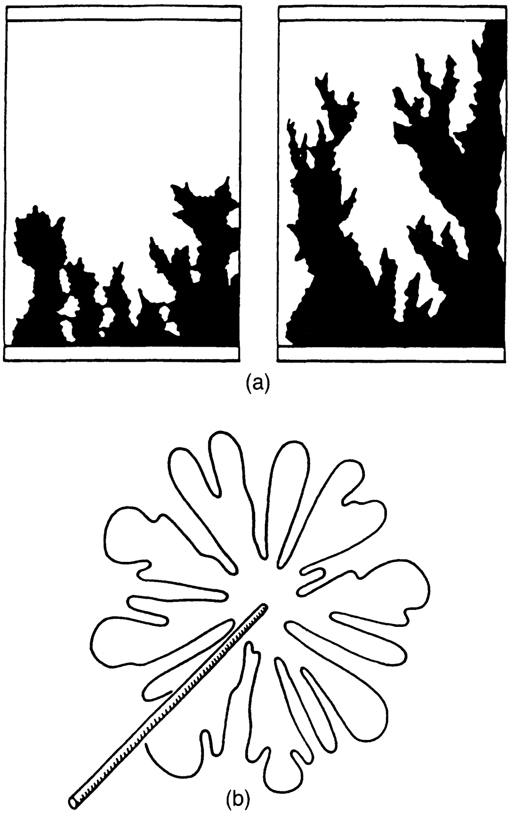

FIG. 1. The developing interface of immiscible two-fluid flow: (a) dyed water displacing glycerine in a porous medium; (b) air injected in at a point in a Hele–Shaw cell filled with glycerine.

Here the subscripts $\alpha = 1,2$ represent respectively the displaced and displacing fluids; also, the sediment is assumed to be so aligned that an angle θ exists between the Earth's gravitational field g and the sediment, which is taken to lie in the x-y plane (see Fig. 2).

2. The incompressibility conditions

$$\nabla \cdot \mathbf{v}_\alpha = 0. \tag{2}$$

3. The pressure difference across any point $\mathbf{s}$ of the two-dimensional curved interface due to the effective surface tension σ in the porous medium,

$$p_2(\mathbf{s}) - p_1(\mathbf{s}) = \sigma[R_1^{-1}(\mathbf{s}) + R_2^{-1}(\mathbf{s})], \tag{3}$$

where $R_1(\mathbf{s})$ and $R_2(\mathbf{s})$ are the two principal radii of curvature of the interface at $\mathbf{s}$.

The origin of surface tension lies at the

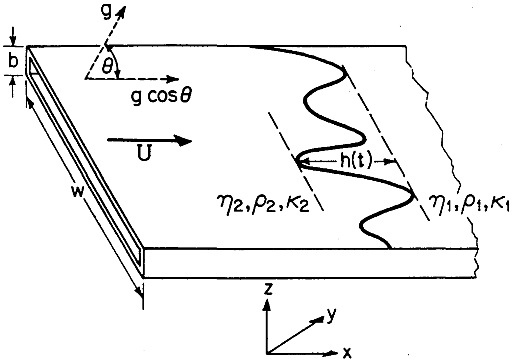

FIG. 2. Schematic diagram of the Saffman–Taylor and Taylor instabilities in a Hele–Shaw cell.

molecular level in the observation that the interaction energy per molecule deviates from its bulk value near an interface, and, in general, this deviation is in the positive direction because of the broken attractive bonds that exist near an interface. Consequently, the energy per unit area of a surface $\sigma = (\partial G/\partial A)_{T,P}$ is positive, or, equivalently, a normal force of size $df = \sigma dl$ exists on any line dl drawn on the surface, tending to it tear apart. This second definition gives rise to the term "surface tension." For curved surfaces with principal radii of curvature R_1 and R_2, the pressure drop Δp required across the surface to keep such a conformation in being is then given by Eq. (3). That this statement is likely to be true can be most easily appreciated by considering the case of a surface defined by the function $z(x, y)$. Consider for the moment the case where the surface is only curved in the x direction $z(x)$. Draw two lines on the surface of length δy parallel to the y axis a distance δx apart, and consider the total force δf_z in the z direction on this rectangle of area $\delta y \delta x$ due to surface tension. By definition,

$$\delta f_z = \sigma \delta y \left[\left(\frac{\partial z}{\partial x} \right)_x - \left(\frac{\partial z}{\partial x} \right)_{x+\delta x} \right] \approx -\sigma \delta y \delta x \frac{\partial^2 z}{\partial x^2}.$$

As the pressure is force/area, we see that the surface tension has induced a pressure $p_{\text{tension}} = df_z/\delta y \delta x = -\sigma\, \partial^2 z/\partial x^2$ across the interface for stability. But on a slowly varying surface, the curvature in the x direction is given by $R_x^{-1}(\mathbf{s}) \approx -\partial^2 z/\partial x^2$. With a more rigorous argument, Eq. (3) can be derived exactly.

The initial instability of a flat surface if surface tension is neglected can be studied using linear-stability analysis. When such a procedure is carried out, the condition for unstable growth is

$$[(\eta_1/\kappa_1) - (\eta_2/\kappa_2)]U + g \cos \theta(\rho_2 - \rho_1) > 0, \tag{4}$$

which includes the Taylor instability and Saffman–Taylor instability conditions as special cases. Experimentally, the gravity-dominated Taylor instability has been observed in oils of similar viscosity but different densities, while the Saffman–Taylor instability has been studied in detail in Hele–Shaw cells.

1.1.2 Flow in Hele–Shaw Cells The Taylor and Saffman–Taylor instabilities are very relevant to the oil industry, concerned as it is with the extraction of oil from geological sediments using water pumped in at high pressure. To study such flows in the laboratory, scaled-down versions of the resulting flow must be created. As such oil-bearing sediments are often effectively two dimensional in character—they have a depth b much less than their width W, or length L (typically, $b \sim 10$ m, $W \sim L \sim 1000$ m)—motion in oil sediments is often analyzed by considering two-fluid flow between glass plates, a device known as a Hele–Shaw cell (typically, $b \sim 0.1$ cm, $W \sim 10$ cm, $L \sim 100$ cm). This use of Hele–Shaw cells is possible because Darcy's law $\mathbf{v} = -(\kappa/\eta)\nabla p$ also applies to these cells with the permeability κ replaced by $b^2/12$ (typically, for porous media, 10^{-7} cm^2 $< \kappa < 10^{-4}$ cm^2, depending on whether the sand is fine or coarse packed, while $b^2/12 \sim 10^{-3}$ cm^2 for a Hele–Shaw cell).

In Saffman–Taylor experiments in Hele–Shaw cells, the less viscous fluid is forced in with a velocity U, and, as expected, an initial instability appears (see Fig. 3). Asymptotically, one finger becomes dominant (see Fig. 4) and grows at the expense of the rest. This finger has a well-defined shape with width λW. Experimentally, the parameter λ has been found to vary from $\lambda \to 1$ as $U \to 0$, to, apparently, $\lambda \to \frac{1}{2}$ as $U \to \infty$, though at high velocities spatiotemporal fluctuations appear.

A great deal of theoretical effort (see, for example, Bensimon *et al.*, 1986, and references therein) has gone into trying to understand the shape and stability of this asymptotic finger for two reasons. First, flow in

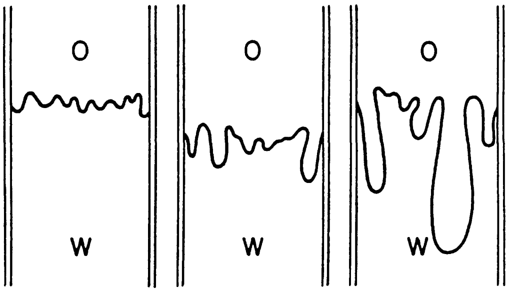

FIG. 3. The growth of an unstable flat interface in the Saffman–Taylor instability (W, water; O, oil); redrawn from Maher (1985).

Hele–Shaw cells represents perhaps the simplest pattern-forming flow that can be seen experimentally and that may yet be theoretically tractable because of its quasi-two-dimensional nature. Second, this finger is closely related to a large class of important problems involving pattern formation in interfaces, including dendrite formation and directional solidification.

Originally, Saffman and Taylor tried to find a relationship between λ and U for the asymptotic finger, while neglecting surface tension, by formulating the flow as a free-boundary problem in the plane, which they solved using conformal mapping. By convention, the glass plates in a Hele–Shaw cell are taken to be separated in the z direction, and the flow is taken to be in the x direction; then the shape of the finger found by Saffman and Taylor can be written

$$\frac{x - Ut}{W} = \frac{1-\lambda}{\pi}\ln\cos(\pi y/\lambda W). \tag{5}$$

Unfortunately, there are two separate problems with the solution given by Eq. (5): It is not unique, as, for each U, a continuous family of solutions parametrized by λ exist; and all these solutions are linearly unstable to small perturbations.

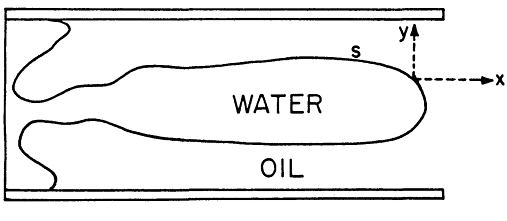

FIG. 4. The asymptotic finger seen in two-fluid flow confined to a Hele–Shaw cell.

Both problems can be remedied by keeping surface tension (Chuoke *et al.*, 1959). Obviously, in the experimental situation, surface tension does exist, and a unique finger is selected for each U. The effect of surface tension is to stabilize the smallest-wavelength fluctuations that have the largest curvatures. A linear-stability analysis of the flat interface confirms physical intuition. Adding a small perturbation to the flat interface $h(y, t) = A_k(t)\cos ky$ results in growth of the k wave-vector mode as

$$\frac{dA_k}{dt} = kA_k\left[U - \frac{\sigma b^2}{12(\eta_1-\eta_2)}k^2\right]. \tag{6}$$

As the minimum value for a wave vector is $k = 2\pi/W$, for very small forcing velocities $U < (\pi^2/3)(b/W)^2\sigma/(\eta_1 - \eta_2)$, the interface is stable. Notice that the stability appears to depend on only one dimensionless parameter $d = (\pi^2/3)(b/W)^2[\sigma/(\eta_1 - \eta_2)U]$, rather than two separate dimensionless parameters—the capillary number $Ca = (\eta_1 - \eta_2)U/\sigma$ [which measures the ratio of viscous $(\eta_1 - \eta_2)U/r$ to capillary forces σ/r for an immiscible two-fluid interface of curvature $1/r$ flowing with velocity U] and the aspect ratio W/b, which is dependent solely on the Hele–Shaw cell geometry. Thus, the condition for stability of a flat interface can be rewritten $d > d_c = 1$. Also, by keeping the surface tension σ, a unique scaling relationship between λ and U can be found (Pitts, 1980) if one equates the change in pressure due to the flow with the change in pressure across the interface at the tip and sides of the finger due to variations of transverse curvature. Again, this relationship depends solely on the parameter d,

$$\lambda = \lambda(d), \tag{7}$$

and obeys the asymptotic conditions $\lambda \to 1$ as $d \to \infty$ and $\lambda \to \frac{1}{2}$ as $d \to 0$.

But what about the stability of this asymptotic finger? Using the flat-interface results as a first estimate, one might expect that, for $d \to \infty$, the finger would be stable, but, for $d < d_{\text{finger},c} \approx 1$, the interface would become unstable. Certainly, experimentally, for high flow rates U, the asymptotic finger appears to become unstable, and time-dependent chaotic fluctuations as well as finger bifurcations set in that distort the shape (Maher, 1985; Tabeling and Libchaber, 1986). Recent theoretical work (DeGregoria and Schwartz, 1986; Bensimon, 1986), however, would suggest that as $d \rightarrow 0$, the asymptotic finger remains stable to infinitesimal perturbations, though unstable to finite perturbations. The size of the finite perturbations that can destroy the finger becomes, however, asymptotically small: $\delta h_{\text{noise}} \sim \exp(-c\, d^{-1/2})$ as $d \rightarrow 0$. Thus, as any experimental setup will always have some noise, there will always appear to exist a critical value for d dependent on the noise such that, for $d < d_c(\text{noise}) \sim 1/(\ln \delta h_{\text{noise}})^2$, the finger will show spatiotemporal fluctuations.

1.1.3 Laplace's Equation and Diffusion-Limited Fingering Asymptotic fingers are generated in a Hele–Shaw cell in the presence of surface tension. But what happens when the viscosity ratio is much greater than one, $\eta_1/\eta_2 > 1$, and there is no surface tension, so that $d = 0$? It is observed experimentally that bifurcation in fingering occurs on all scales, and the flow takes on a dendritic structure reminiscent of diffusion-limited aggregation.

Diffusion-limited aggregation was originally introduced by Witten and Sander (1981) as a model for aggregation of soot particles. Imagine that carbon atoms performing a Brownian walk travel until they touch the soot aggregate and that, at this point, they stick irreversibly. What will be the shape of the resulting aggregate? Probably, before the first simulations were performed, one would have guessed that a compact cluster with a rough surface would be seen. In fact, a fractal dendritic object occurs [see Fig. 5(b)]. The reason for such growth is screening. The particle performing a random walk sticks on first contact and consequently cannot enter deep into the resulting fjordlike voids to create a compact object. Diffusion-limited aggregation might have remained a model for a certain class of aggregation processes, except that it was shown by Niemeyer *et al.* (1984) to be isomorphic to a solution of Laplace's equation (because of the random walk taken by the aggregating particles) with moving boundary conditions and, thus, a basic universal paradigm controlling the structure of such apparently diverse phenomena as electrochemical deposition (Brady and Ball, 1984; Matshushita *et al.*, 1984), dielectric breakdown (Niemeyer *et al.*, 1984), and even aspects of biological growth, such as retinal vasculature (Family *et al.*, 1989). Only the nature of the field obeying Laplace's equation changes: for example, the concentration field in the case of aggregation or the electrical potential in the case of dielectric breakdown.

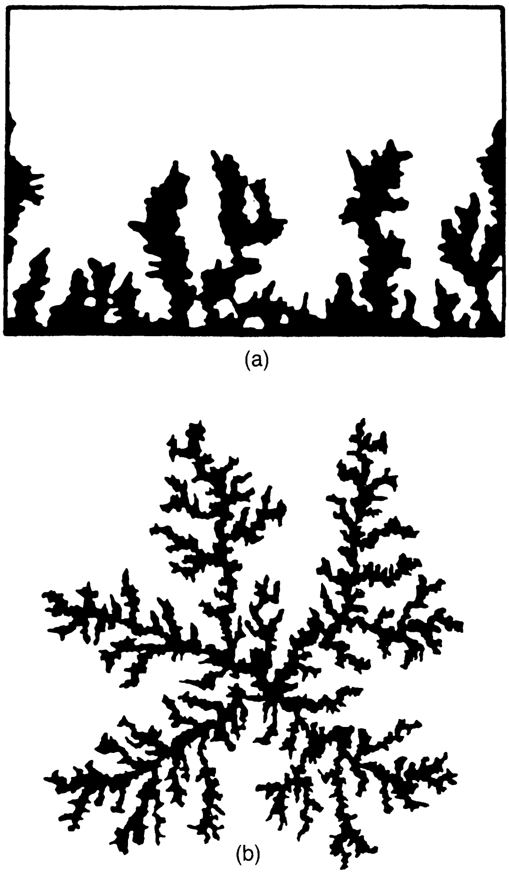

FIG. 5. Diffusion-limited aggregate grown (a) on a flat interface, (b) from a point seed. Compare with the two-fluid flow experiments in Fig. 1.

Two-fluid flow in porous media was also shown by Paterson (1984) to obey Laplace's equation in a certain limit, with the field this time being the pressure field (compare Fig. 5, which shows computer simulations of diffusion-limited aggregation with various boundary conditions, with Fig. 1). The evolving interface in immiscible two-fluid flow can therefore be shown to be in the same universality class as diffusion-limited aggregation. To see this, consider Eqs. (1–3).

First, neglect g and η_2, in which case Eq. (1) reduces to $\nabla p_2 = 0$ for the displacing phase. This has the solution $p_2 = 0$, and specifically, at the interface, $p_2(\mathbf{s}) = 0$. Now neglect surface tension σ; then $p_1(\mathbf{s}) = 0$ for the displaced fluid interface also. Using Eq. (2), it can now be seen that in the more viscous phase, the pressure field $p_1(\mathbf{x})$ and interfacial velocity $\mathbf{v}_1(\mathbf{s})$ obey identical equations to diffusion-limited aggregation (DLA):

$$\nabla^2 p_1 = 0, \tag{8}$$

with $p_1(\mathbf{s}) = 0$ and p_1(other boundaries) constant, while the boundary moves as $\mathbf{v}_1(\mathbf{s}) = -(\kappa_1/\eta_1)\nabla p_1(\mathbf{s})$. And, indeed, the argument appears to be correct. Experiments on two-fluid flow (Nittmann *et al.*, 1985) with $\eta_1 \gg \eta_2$ and negligible surface tension σ, by using water ($\eta_2 \approx 10^{-2}$ P) to displace polysaccharide solutions ($\eta_1 \approx 10^2$ P) in a Hele–Shaw cell, yield a fractal interface with DLA exponents.

For finite values of the capillary number $Ca = (\eta_1 - \eta_2)U/\sigma$ and the viscosity ratio $M = \eta_2/\eta_1$, a more complex crossover behavior is seen. At short times $t \ll t_c(Ca, M)$, the flow appears to be a dendritic fractal in the same universality class as DLA with an inner length scale $l(Ca, M)$ controlled by capillarity and viscosity ratio. At long times, however, $t \gg t_c(Ca, M)$, crossover to compact flow is observed, though the interface may have a rough self-affine form (see Sec. 3).

1.2 Dynamics of Liquid–Gas Interfaces

Interfaces between gas and liquids also exist, and the most commonly observed dynamics for this case are waves.

1.2.1 Surface Gravity Waves The large-scale surface evolution to be observed on seas and lakes is in the form of surface gravity waves. This wave motion is caused by the interaction of an inertial term, the liquid density ρ_0, tending to keep the liquid moving at a constant velocity, while gravity $\mathbf{g}$, taken to act in the z direction, is the anisotropic resorting force leading to oscillations. Assume that the equilibrium liquid–gas interface (typically water–air) lies in the $z = 0$ plane and consider a small perturbation $h(\mathbf{r}, t)$. This perturbation is assumed small enough that linear theory is applicable, and, specifically, the interface is single valued at each point (a condition violated by the large-scale wave motions seen in stormy seas).

In the water phase $z < h(\mathbf{r}, t)$, the velocity field $\mathbf{v}(\mathbf{x}, t)$ may be taken to obey

$$\rho_0 \frac{\partial \mathbf{v}}{\partial t} = -\nabla p + \rho_0 \mathbf{g}, \tag{9}$$

where $p(\mathbf{x}, t)$ is the pressure field in the water and dissipation has been neglected. The additional constraint of incompressibility gives $\nabla \cdot \mathbf{v} = 0$. Looking for irrotational potential flow solutions $\mathbf{v} = \nabla\phi$ implies that the velocity potential obeys Laplace's equation

$$\nabla^2 \phi = 0, \tag{10}$$

while the pressure field is given by

$$p = p_a - \rho_0 \frac{\partial \phi}{\partial t} - \rho_0 g z. \tag{11}$$

Here, the constant of integration p_a is the air pressure at the interface.

Eq. (11) can be used to yield the boundary condition for the velocity potential at the interface,

$$\left.\frac{\partial \phi}{\partial t}\right|_{z=h} = -gh, \tag{12}$$

while, if the water is of depth d, then the normal component of the velocity is zero at this depth $v_z(z = -d) = 0$, or, in terms of the velocity potential,

$$\left.\frac{\partial \phi}{\partial z}\right|_{z=-d} = 0. \tag{13}$$

As the boundary condition at the interface Eq. (12) is expressed in terms of the surface conformation h, the equation of motion for the interface itself, given by $dh/dt = v_z(z = h)$, is needed to complete the free-boundary problem involved. In terms of the velocity potential, the interface therefore obeys

$$\frac{\partial h}{\partial t} + \nabla\phi \cdot \nabla h = \left.\frac{\partial \phi}{\partial z}\right|_{z=h}, \tag{14}$$

and, if we combine Eqs. (12) and (14), the explicit surface profile can be deleted in favor of the velocity potential ϕ alone. In the linear approximation, this yields the interface boundary condition

$$\left.\frac{\partial^2\phi}{\partial t^2}\right|_{z=0} = -\left.\frac{g\partial\phi}{\partial z}\right|_{z=0}. \tag{15}$$

Laplace's equation, together with the boundary conditions at the interface [Eq. (15)] and on the sea bottom [Eq. (13)], yields traveling-wave solutions $\phi(\mathbf{r}, t) = \Phi(z)\exp(ikx - i\omega t)$ provided

$$\omega^2 = gk \tanh kd. \tag{16}$$

Thus, in contrast to sound waves in bulk fluids, which are nondispersive and longitudinal in nature, surface gravity waves can be strongly dispersive, in addition to which the resulting velocity field during propagation has both longitudinal and transverse components.

Two limiting cases can be derived from Eq. (16). For not too long-wavelength surface waves on deep water, $kd \to \infty$, and the dispersion relation reduces to $\omega^2 = gk$. Thus, the phase velocity of of deep-water waves increases with wavelength, $c_k = \omega/k = (g/k)^{1/2}$, as does the group velocity, which is exactly half the phase velocity: $u_g(k) = d\omega/dk = c_k/2$. This increase of velocity with wavelength can be seen when a stone is thrown into a pool: After a time t, the localized initial disturbance spreads into a sinusoidal pattern in which the long-wavelength modes have traveled further, $u_g(k)t$, than the shorter wavelengths, which remain near the origin of the disturbance.

The other limit consists of long-wavelength waves in shallow channels. In this case, $kd \to 0$, and the dispersion relation reduces to $\omega = (gd)^{1/2}k$. Such channel waves are nondispersive with a speed of propagation $c = (gd)^{1/2}$, which, for typical channel depths, is a much slower mode of propagation than sound. In addition, for such long-wavelength modes, and in contrast to sound, the wave motion is essentially transverse. The dependence of speed on depth also gives an indication of the way waves break on a shoreline. The closer to the shoreline, the slower the speed of wave propagation, and this will result in a shocklike phenomenon of waves breaking on the ocean front as the faster, deep-water waves catch up with slower waves near the shoreline.

1.2.2 Capillary Waves Including surface tension results in a pressure drop $\Delta p = \sigma(1/R_1 + 1/R_2)$ across the liquid–gas interface. Clearly, the influence of surface tension becomes ever more significant at larger curvatures, and, therefore, small-wavelength oscillations must finally be dominated by surface tension. Such waves, known as capillary waves or surface ripples, can be seen when the surface of a glass of water is gently disturbed.

In the linear approximation for waves traveling in the x direction, the only nonzero surface curvature lies in the x direction and is given by $\Delta p = \sigma(1/R_1 + 1/R_2) \approx -\sigma\, \partial^2 h/\partial x^2$. Therefore, the excess pressure in the water above the air pressure is given by

$$p_e(z) = \rho_0 g h - \sigma\frac{\partial^2 h}{\partial x^2}. \tag{17}$$

In consequence, for an oscillation of wave number k, the dispersion behavior for surface gravity waves may be generalized to include the effects of surface tension by replacing g by $(g + \sigma k^2/\rho_0)$:

$$\omega^2 = (g + \sigma k^2/\rho_0)\, k \tanh kd. \tag{18}$$

It is apparent from Eq. (18) that crossover to surface-tension–dominated ripples occurs at $k_c \approx (\rho_0 g/\sigma)^{1/2}$. For water, this corresponds to a wavelength of about 1.7 cm.

For deep-water waves, the dispersion relation [Eq. (18)] reduces to $\omega^2 = (g + \sigma k^2/\rho_0)k$, and, therefore, the phase velocity can be written

$$c_k = \left[\frac{g+\sigma k^2/\rho_0}{k}\right]^{1/2}. \tag{19}$$

Note that Eq. (19) implies that the speed of the wave motion has a minimum value at the crossover to capillary-dominated behavior, $k_c \approx (\rho_0\, g/\sigma)^{1/2}$. For water, this corresponds to a speed of 23 cm/s. At even smaller wavelengths, surface tension dominates surface fluctuations entirely, in which case the dispersion curve for pure capillary waves finally appears as

$$\omega^2 = \sigma k^3/\rho_0. \tag{20}$$

Despite the rich variations that surface waves have been shown to exhibit, the whole discussion here has been within the linear approximation, and both dissipation and nonlinear wave phenomena have been neglected. The complexity of real sea-wave motion attests to the rich store of new phenomena to be expected if such nonlinear behavior is included [see, for example, Lighthill (1978)].

1.3 Flow on Surfaces

In flows on solid substrates, once again surface tension plays a dominant role in controlling both form and flow. Surface tension will sometimes tend to minimize surface areas, accounting for such observations as the fact that droplets are spherical, or that wetted fibers tend to stick together, while in other cases it encourages wetting of the complete substrate.

1.3.1 Wetting, and Capillary Flow Consider a column of liquid in a capillary tube of radius r, and specifically, its curvature at the top of this column where the liquid and vapor phases meet. It could either be concave, in which the case the liquid is wetting the solid substrate forming the capillary tube (water in glass, for example), or convex, in which case the liquid is nonwetting (mercury in glass, for example). For a nonwetting fluid to enter a capillary tube and displace the air, an external pressure must be applied to overcome capillary forces resisting entry of the liquid, while for a wetting fluid, the capillary forces will encourage entry even against an external gravitational field. The same phenomena of drainage and imbibition occur during flow in porous media (see Sec. 3).

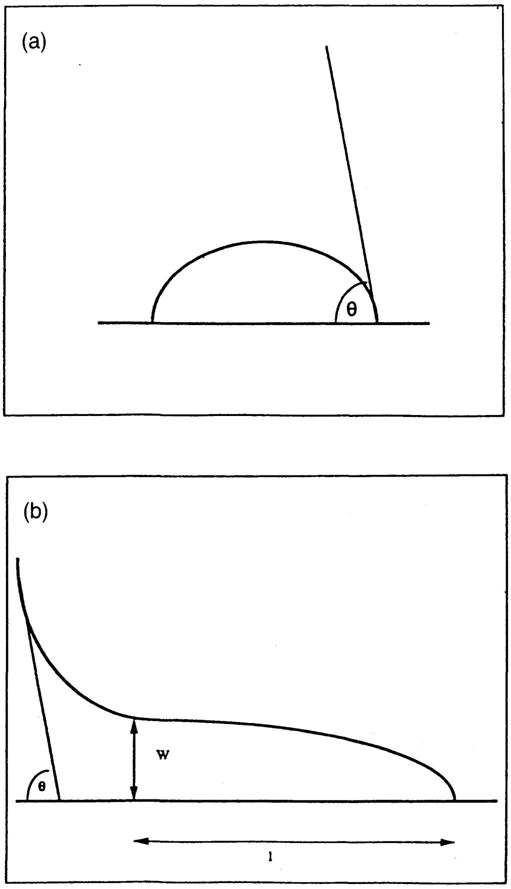

FIG. 6. Schematic diagram showing (a) a droplet in mechanical equilibrium, (b) a spreading droplet on a flat substrate.

To answer the question as to whether a liquid will or will not wet a solid surface, the angle between the tangent to the liquid–vapor interface where it meets the substrate and the substrate itself is required. This angle is known as the static wetting angle θ [see Fig. 6(a)]. For mechanical equilibrium to exist for a droplet on a solid substrate, the sum of all the forces due to the surface tension for each interface, $\sigma_{LV} = \sigma$, σ_{LS}, and σ_{SV} (L = liquid, V = vapor, S = substrate), must vanish at the contact point or

$$\sigma_{SV} - \sigma_{LS} - \sigma \cos \theta = 0. \tag{21}$$

This is Young's equation for the wetting angle θ of a stable nonwetting liquid–vapor interface lying on a substrate. If we use a similar argument for the dynamic wetting angle θ_d during the motion of a liquid up a capillary tube, then there exists a pressure $p_{\text{tension}} = 2(\sigma_{SV} - \sigma_{LS} - \sigma \cos \theta_d)/r$ due to capillarity. Thus, the pressure drop across the column of height h is $\Delta p = [2(\sigma_{SV} - \sigma_{LS} - \sigma \cos \theta_d)/r - \rho g h]$, and, if Poiseuille flow $v = r^2 \, \Delta p / 8\eta h$ is occurring in the capillary tube, then

$$\frac{dh}{dt} = \frac{r^2}{8\eta h}\left[\frac{2(\sigma_{SV} - \sigma_{LS} - \sigma \cos \theta_d)}{r} - \rho g h\right]. \tag{22}$$

Wetting fluids may therefore be expected to rise a distance $h_{\max} = 2(\sigma_{SV} - \sigma_{LS} - \sigma \cos \theta_d)/r\rho g$ in capillary tubes, though this will occur with an ever-decreasing velocity as $h_{\max}$ is approached, and to reach this height requires theoretically an infinite amount of time.

1.3.2 Droplet Flow If we define a spreading coefficient $S = \sigma_{SV} - \sigma_{LS} - \sigma$, then Eq. (21) implies that, for $S > 0$, a stable droplet is impossible, as Young's equation cannot be obeyed for any static wetting angle, and the droplet will flow over the solid surface, forming a thin film of microscopic dimensions. The dynamics of this wetting process are rather complex, for even the question of the existence of a dynamic wetting angle θ_d has to be treated with some care. A cursory definition of a macroscopic dynamic contact angle $\theta_{d,\text{macro}}$ as the angle between the nominal contact line and the substrate surface would suggest that this contact angle has a singularity as the spreading velocity U tends to zero, for the static contact angles defined in the limits of receding ($U \to 0^-$) and advancing ($U \to 0^+$) contact lines are often observed to be different. A microscopic examination of the region near the spreading droplet in the vicinity of the apparent contact point at which the liquid, vapor, and substrate phases meet, however, would reveal that the nominal macroscopic contact line is actually preceded by a mesoscopically thin precursor film [see Fig. 6(b)] due to a disjoining pressure Π that reflects the existence of attractive long-range van der Waals forces between the liquid and substrate phases (Joanny and de Gennes, 1984; de Gennes, 1985). Thus, the apparent singularity is a macroscopic artifact, and a separate microscopic dynamic contact angle $\theta_{d,\text{micro}}$ can be defined by the angle between the spreading surface and the tangent between the macroscopic droplet and the tip of this precursor film. The thickness W of this precursor film, as opposed to its spreading length l, is controlled by a balancing of the van der Waals forces, which will tend to make this film microscopically thin, indeed of atomic dimensions, with the short-range capillary forces that will tend to discourage the existence of such a film at all.

Consider a droplet spreading on a solid surface whose normal lies in the z direction, and suppose the droplet is spreading in the x direction. The dynamic droplet shape $h(x, t)$ during spreading can be related to the spreading velocity $U(t)$ [see, for example, Joanny (1990)]. If the velocity field in the droplet has a parabolic profile in the z direction as in Poiseuille flow, then the average velocity $U(x, t)$ is given by

$$U(x, t) = \left(\frac{h^2}{3\eta}\right)\frac{dp}{dx}, \qquad (23)$$

where the pressure $p = \sigma d^2h/dx^2 + \rho gh + \Pi(h)$. Solving the equations of motion for sessile drops, that is, small drops where the hydrostatic pressure can be neglected, yields

$$h(x, t) = [9\eta U(t)/\sigma]^{1/3}\, x \log^{1/3}(kx), \qquad (24)$$

where k is a constant of integration. Thus, apart from logarithmic correction, the drop profile near the interfacial contact point is wedgelike, with a dynamic contact angle

$$\theta(t)_{d,\text{macro}} = [9\eta U(t)/\sigma]^{1/3}. \qquad (25)$$

Equation (25) is Tanner's law, which states that a universal relation exists between the contact-line velocity and the dynamic contact angle.

Equation (25) can also be used to find the scaling behavior of droplet radius with time, $R(t)$. For a nonvolatile liquid, the total volume V is conserved during spreading, but for a wedge-shaped drop, $V \sim R^3(t)\theta_{d,\text{macro}}(t)$, and, as $U(t) \sim dR/dt$, the droplet radius increases as

$$R(t) \sim V^{3/10}[\sigma t/\eta]^{1/10}. \qquad (26)$$

Similar calculations can carried out for pendant-shaped drops dominated by gravity, yielding $R(t) \sim V^{3/8}[\rho g t/\eta]^{1/8}$.

It should be emphasized that the results above imply a very idealized form of wetting dynamics. First, to study mathematically the dynamics of wetting, it is not enough to assume Poiseuille flow for the droplet. Clearly, stick boundary conditions on the substrate surface are incompatible with a moving contact line, but what are the correct boundary conditions? Possibilities would include a frictional force proportional to the surface velocity v introduced at the substrate surface, resulting in a stress tensor $\sigma_{x,z} = \gamma v$; another possibility would be that the main body of fluid in the droplet before the nominal contact line does indeed obey stick boundary conditions, but slip boundary conditions are applied to the precursor film, whose length diverges as $l = vt$, and, consequently, the molecules in the precursor film are free to roll at the substrate surface.

Second, volatility and contamination have been neglected; these can lead to temperature and surface-tension gradients, resulting in additional forces (the Marangoni effect). In fact, the alteration in dynamics can be dramatic: For example, if surfactant solutions are spread on moist surfaces, large surface-tension gradients will appear, leading to a fingering of the contact line (Marmur and Lelah, 1981); this spreading is an order of magnitude faster than surface-tension- or gravity-dominated wetting. Also, if the substrate surface is not smooth, then the wetting dynamics may resemble more closely flow in porous media than the creeping flow described here. Finally, liquids have been assumed to be Newtonian; for non-Newtonian rheologies, the shape of a spreading droplet may vary considerably from the caplike droplets seen in simple wetting. For example, shear-thinning liquids typically develop a protrusion that extends beyond the cap, forming a macroscopic foot (Brochard and de Gennes, 1984).

So far, the wetting of two-dimensional surfaces has been examined. But, technologically, the wetting of fibers is just as important. One of the most interesting instabilities concerns the wetting–nonwetting transition on fibers. Will a liquid prefer to form an annulus of thickness l enveloping the fiber, which can be taken to have the form of a cylinder of radius r and length L, or to form a series of droplets instead, as is often seen on spider's webs after rain? The total capillary energy per unit length of fiber E/L above that for the dry fiber may be written

$$E/L = 2\pi r(\sigma_{LS} - \sigma_{SV}) + 2\pi(r + l)\sigma, \tag{27}$$

and, therefore, when $E/L < 0$, or the spreading coefficient $S > \sigma l/r$, the fiber will be wetted. Thus, it is more difficult to wet a fiber than a surface made from the same material.

1.4 Surface Evolution and Mixing

Surfaces also evolve in a complex manner when mixtures are stirred. The process occurs in two stages. First, deterministic stirring can lead to a stretching and folding of material surfaces reminiscent of the manner in which a baker creates millefeuille pastry. Such stretching and folding lie at the heart of chaotic mixing. Then, when the folded layers of the liquids give rise to large enough gradients in their respective concentrations, diffusion takes over from deterministic mixing as the most powerful mechanism equilibrating the separate liquid phases, resulting ultimately in a homogeneous mixture.

1.4.1 Deterministic Mixing Consider the evolution of immiscible interfaces by external deterministic mixing. It would certainly not surprise anyone that with random stirring, efficient mixing of liquids can be made to occur, but what may surprise one is that periodic stirring can be very efficient in mixing. Under periodic stirring, a stretching and folding of the material surfaces can occur in which the interfacial region grows with time, forming a highly interwoven convoluted structure having fractal properties, though often, ordered flow regions also coexist with the chaotic mixing regions and appear to be separated from them by impenetrable barriers. Clearly, from a technological viewpoint, such barriers are undesirable if enhanced mixing is required.

Such mixing behavior has not only been observed in the laboratory but also on a large scale in atmospheric and oceanographic flows. For example, vortices and jets have been observed to act as trapping barriers to mixing and may be the cause of the persistence of the Great Red Spot on Jupiter. In addition, vortexlike flows in the Atlantic have been observed to trap for years water and animal life from the Mediterranean where these eddies originated.

Typically [see Ottino (1990) for a comprehensive review], during a mixing process, the material can be considered to have no back reaction on the flow field, so that the material behaves as a passive scalar, and the resulting concentration of one component $c(\mathbf{r}, t)$ is given by

$$\frac{\partial c}{\partial t} + \mathbf{v} \cdot \nabla c = D\nabla^2 c, \tag{28}$$

where the velocity $\mathbf{v}(\mathbf{r}, t)$ is a known deterministic function of space and time. In addition, the flow may be taken to be incompressible, and, therefore, $\nabla \cdot \mathbf{v} = 0$. The term $D\nabla^2 c$ represents diffusion due to thermal fluctuations, which are the dominant contribution to mixing at small scales (see Sec. 1.4.2).

How can this completely deterministic stirring flow field $\mathbf{v}(\mathbf{r}, t)$ chaotically advect the material surfaces and induce the observed complexity and inhomogeneities that characterize such stirred fluids (see Fig. 7)? The answer appears if, for the moment, we neglect diffusion and consider the motion of any marked particle in this flow field. It will obey the deterministic dynamical system

$$\frac{d\mathbf{x}}{dt} = \mathbf{v}(\mathbf{x}, t), \tag{29}$$

which, as $\nabla \cdot \mathbf{v} = 0$, is the equation of motion for a conserved flow and therefore reminiscent of N-dimensional nonintegrable Hamiltonian systems obeying

$$\begin{aligned} \frac{\partial p_\alpha}{\partial t} &= -\frac{\partial H}{\partial q_\alpha}, \\ \frac{\partial q_\alpha}{\partial t} &= \frac{\partial H}{\partial p_\alpha}, \end{aligned} \tag{30}$$

where p_α, q_α $(1 \le \alpha \le N)$ are the canonical momentum and position variables describing the dynamics. Such mechanical systems also conserve volume in phase space by Liouville's theorem. Conserved flows obeying Eq. (29) can have chaotic trajectories for the autonomous case $\mathbf{v}(\mathbf{x}, t) = \mathbf{v}(\mathbf{x})$ in three or more dimensions, and, even in two dimensions, chaotic trajectories are possible for time-periodic flows $\mathbf{v}(x, y, t + \tau) = \mathbf{v}(x, y, t)$. Indeed, two-dimensional periodic flows can be made isomorphic to periodically forced one-dimensional Hamiltonian systems with the real space of the fluid motion being replaced by the phase space $(x \rightarrow p, y \rightarrow q)$. Thus, by studying the dynamics of Hamiltonian systems, we can understand aspects of deterministic mixing.

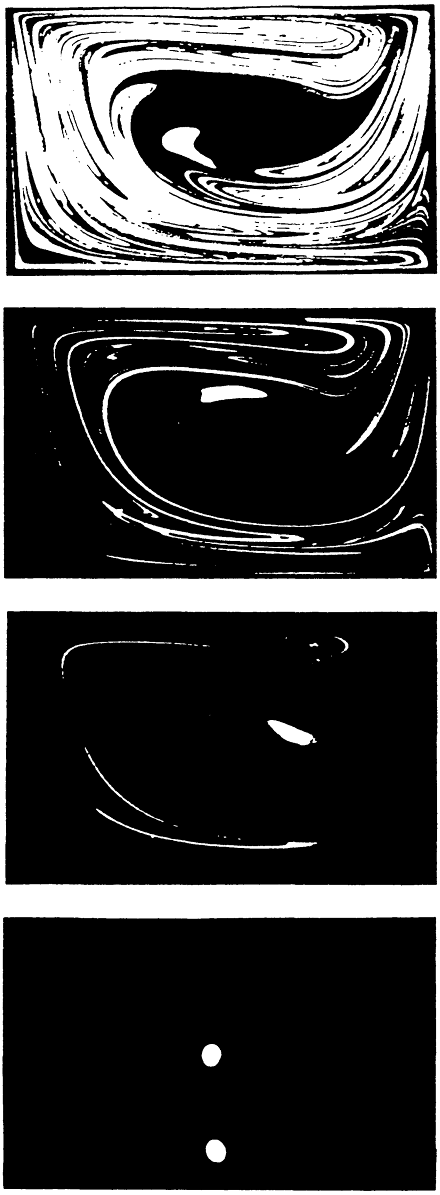

FIG. 7. Stretching and folding of two passive blobs under periodic mixing; note the appearance of regular regions within the chaotic flow; redrawn from Ottino (1990). Reproduced with permission from the *Annual Review of Fluid Mechanics*, Volume 22, ©1990, by Annual Reviews Inc.

Flows in integrable N-dimensional Hamiltonian systems have N constants of motion and are therefore confined to N-dimensional submanifolds of the complete $2N$-dimensional phase space with the topology of a torus. Thus, using the analogy introduced above, for integrable stirring, one would expect material trajectories to be confined to submanifolds of the stirred domain. This is not what is observed experimentally. Real stirring is equivalent to a nonintegrable Hamiltonian system, and the question arises of what dynamical systems theory can tell us about this case. Perhaps the most important property of nonintegrable Hamiltonian flows is given by the KAM (Kolmogorov–Arnol'd–Moser) theorem, which describes what happens to integrable quasiperiodic flows on tori as nonintegrable terms are added to the Hamiltonian. The motion on an N-torus consists of N in general incommensurate periodic motions and is therefore quasiperiodic. Nonintegrable terms destroy all tori eventually, and commensurate-frequency tori immediately (though they are of measure zero compared with the incommensurate case); but what is important is that most tori with incommensurate frequencies are at first only distorted, the most stable tori, which last longest under nonintegrable perturbations, being those with the most incommensurate frequencies. Once a torus is destroyed, a trajectory is free to wander in an apparently random manner, resulting in great sensitivity of the trajectories to their initial conditions. Thus, typical regions of a deterministic chaotic flow are separated by surviving tori through which no trajectory can penetrate, accounting for the observed spatial structure in chaotic advection.

As smaller and smaller length-scale structure is created during this stretching and folding process, however, there must come a point where this analogy with Hamiltonian flows breaks down, for diffusion will destroy all tori at small enough scales (the dissipative term dominates the inertial transport term, $D\nabla^2 c \gg \mathbf{v}\cdot\nabla c$, at small enough scales), and the final stage of any mixing process will be dominated by diffusion.

1.4.2 Mixing by Diffusion Diffusion is the transport mechanism by which thermal fluctuations equilibrate density inhomogeneities in a multiphase medium. Consider a drop of ink placed in a quiescent glass of water; with time, the fluids will mix totally, and, on a macroscopic level, the densities of both water and ink will appear constant. To find the equations of motion for such diffusion, we need to consider the equations of continuity for multicomponent miscible fluids.

Miscible fluids can be described by their local mass densities $\rho_i(\mathbf{r}, t)$ (n components in the general case). The current of species i is given by the sum of convective and diffusive contributions:

$$\mathbf{J}_i = \rho_i \mathbf{v} + \mathbf{J}_{i,\text{diff}}, \tag{31}$$

where the assumption is being made in Eq. (31) that the local velocity of each species is identical, $\mathbf{v}_i = \mathbf{v}$.

The diffusive flux can be related to the thermodynamic force by

$$\mathbf{J}_{i,\text{diff}} = -\lambda\nabla\mu_i = -\lambda\frac{\partial\mu_i}{\partial\rho_i}\nabla\rho_i = -D_i\nabla\rho_i. \tag{32}$$

In Eq. (32), μ_i is the chemical potential of species i, and the mass diffusion constant for species i is therefore given by $D_i = \lambda\partial\mu_i/\partial\rho_i$. With these constitutive relations, the equations of continuity for a multicomponent miscible fluid become

$$\frac{\partial\rho_i}{\partial t} + \nabla\cdot(\rho_i\mathbf{v}) = D_i\nabla^2\rho_i. \tag{33}$$

The total density of the fluid is $\Sigma_i\rho_i = \rho$, and this total fluid density obeys $\partial\rho/\partial t + \nabla\cdot(\rho\mathbf{v}) = 0$, as $\Sigma_i\, \mathbf{J}_{i,\text{diff}} = 0$. This last result is a consequence of the fact that the densities of only $n - 1$ components in an n-component mixture can be independent. It is also possible to define the specific density of each species as $\rho_i^* = \rho_i/\rho$, in which case $\Sigma_i\, \rho_i^* = 1$. The specific density obeys the related equation of continuity

$$\frac{\partial\rho_i^*}{\partial t} + \mathbf{v}\cdot\nabla\rho_i^* = D_i\nabla^2\rho_i^*, \tag{34}$$

accounting for Eq. (28) describing mixing.

To estimate the power of diffusion in mixing surfaces over a distance L, note that the time T_{diff} taken for an inhomogeneity to spread a distance L by diffusion scales as $T_{\text{diff}} \sim L^2/D$. As the deterministic mixing time is $T_{\text{det}} \sim L/U$, where $U \sim \max|v(\mathbf{x}, t)|$, diffusion must win at small scales $L \ll D/U$.

2. MULTIPHASE AND ANISOTROPIC FLOWS

Fluids having Newtonian rheologies (see RHEOLOGY) are isotropic and have a stress proportional to rate of strain, and, indeed, such assumptions describe remarkably accurately the flow properties of simple molecular fluids. Here, we consider the flows of that large class of technologically important nonhomogeneous materials that break one or both of these assumptions.

An important class of flows comprises those in materials that contain dispersions of one pure phase in another, separated by interfaces of microscopic or mesoscopic size. Such multiphase flows can on a macroscopic scale be treated as continuous but usually have non-Newtonian rheologies. This would include foams, which consist of a gaseous phase separated by liquid interfaces; emulsions, which consist of droplets of two immiscible liquid phases; slurries, which are high concentrations of solid particles dispersed in a liquid phase; and granular flows, which on a mesoscopic scale are solid particles separated by a gas phase. Coarse-graining of the mesoscopic dynamics results in a new effective rheology whose properties are often far from Newtonian, showing strong shear-thinning or -thickening behavior; or, if Newtonian, then having transport coefficients that are strongly dependent on the structure at this mesoscopic scale.

Macroscopic anisotropic flows can also occur in molecular fluids and mixtures. Such fluids are orientationally ordered on a macroscopic scale and include liquid crystals and electrorheological fluids.

2.1 Liquid–Solid Phase Flows

Numerous technologically important materials can be thought of as consisting of solid particles immersed in a liquid phase. This applies not only to suspensions and slurries, where the suspended particles are of macroscopic or mesoscopic size, but also to polymer solutions (see POLYMER DYNAMICS), which have a somewhat analogous structure in the semidilute regime, with their radius of gyration replacing the hard-sphere radius.

2.1.1 Slurry Flows A concentrated mixture of solid particles immersed in a liquid forming a suspension can be argued to lead to even more complex flow fields than are to be expected for flows in porous media (see Sec. 3). The reason for this is that, while porous media present a quenched random topography (that is, a fixed-in-time though random environment) in which the flow must occur, the suspended-particle geometry in which liquid flow occurs in a slurry is controlled by a complex feedback process between flow and geometry. In principle, for a fixed multiparticle geometry, the complete Stokes flow must be solved; then this flow in turn will result in hydrodynamic forces on the suspended particles, leading to a complex suspension dynamics [see, for example, Batchelor (1967)].

The resulting macroscopic rheology of concentrated suspensions can be studied as a function of the volume fraction ϕ of solid particles suspended in the liquid phase, provided one is interested in shear flows that are varying over much larger length scales than the interparticle distance. In the semidilute suspension limit, the effective viscosity can be expanded in a power series in this volume fraction:

$$\eta = \eta_0[1 + 5\phi/2 + O(\phi^2)]. \tag{35}$$

The first term in this expansion, derived originally by Einstein (1906), can be found by calculating the effect on the flow field of a single spherical particle and then summing these contributions from all particles, as though independent, into a new effective stress tensor. To first order the effect of the suspension on rheology is to increase the fluid viscosity in the universal manner given by Eq. (35); that it should increase is clear, for the effect of the suspended particles can only increase dissipation during flow. The term quadratic in the volume fraction is the first contribution that requires for its calculation that the suspended particles are no longer assumed to be independent; and such calculations lead to additional direct interaction and Brownian forces on the particles, as well as a changed hydrodynamic stress.

The coefficients of the higher-order terms in the volume fraction are not universal constants but depend on the Peclet number $Pe = \gamma a^2/D$, where γ is the shear rate, a is the particle radius, and D is the particle diffusion constant. The Peclet number here measures the ratio of two time scales—a diffusive time scale $t_D \sim a^2/D$ to a shear time scale $t_{\text{shear}} \sim 1/\gamma$—for particle motion. Because of this dependence of the effective viscosity on Peclet number, the flow rheology cannot be Newtonian and, in general, can be expected to show shear-thinning or shear-thickening behavior. Indeed, measurements of η/η_0 vs the Peclet number typically show a dramatic shear-thinning behavior for suspensions at Pe

~ 10, with separate plateaus in the effective viscosity for much lower and much higher Peclet numbers (see Fig. 8).

For concentrated solutions, we move into the slurry flow regime [see, for example, de Kruif (1990)]. Such flows are highly shear thinning, but the viscosity begins to diverge,

$$(\eta/\eta_0) \sim [1 - \phi/\phi_m]^{-2}, \tag{36}$$

as the volume fraction ϕ approaches the value $\phi_m \approx 0.63$, which is the limiting value for the volume fraction of random close-packed spheres. Indeed, even before this limit has been reached, phase transitions to ordered solid phases for $\phi > 0.5$ have been observed, and at high shear rates, the suspension can become anisotropic, as light- and neutron-scattering studies confirm.

2.1.2 Flows in Polymer Solutions and Melts Consider a semidilute polymer solution formed from the polymerization of a number concentration c of monomers into polymer molecules consisting of N contiguous monomers. It is known (Flory, 1969) that in a good solvent, such linear polymer chains take up the conformation of a self-avoiding random walk of radius of gyration $R_g \sim N^{3/5}$. This observation allows one to estimate the effect of the polymer on the solvent viscosity. By treating the polymers as spheres of radius $\sim R_g$ and using the results for solid suspensions [Eq. (35)] to calculate their rheology in the semidilute limit (in the concentrated regime, this scaling analogy will break down as individual polymer molecules begin to interpenetrate), one would expect

$$\delta\eta/\eta \sim \phi \sim cR_g^3/N \sim cN^{4/5}. \tag{37}$$

Such scaling arguments, which were first introduced by Flory (1971) and de Gennes (1984) into polymer physics, are very powerful and have been verified for numerous static and dynamic properties of polymers.

As a further example of dynamical scaling behavior, consider the self-diffusion D. For a solid particle of radius R, Stokes calculated that $D = kT/6\pi\eta R$. Therefore, for a polymer molecule, one may estimate $D \sim kT/6\pi\eta R_g$ (de Gennes, 1977), implying a scaling of the diffusion coefficient with polymer size of $D \sim N^{-3/5}$, which again has been confirmed by light-scattering experiments (de Gennes, 1984). Using this diffusion coefficient, one may introduce a time scale $\tau_{\text{Zimm}} = R_g^2/D \sim 6\pi\eta R_g^3/kT$ for polymer molecule self-diffusion across a distance of the polymer size. Thus, only shear flows greater than the inverse time scale $\gamma \gg 1/\tau_{\text{Zimm}}$ can be expected to influence polymer dynamics significantly.

Semidilute polymer solutions are also shear thinning and are often represented by a viscosity which has a constitutive form of the Bird–Carreau type,

$$\frac{\eta}{\eta_0} \approx \left[1 + \frac{dv/dz}{\gamma_c}\right]^n, \tag{38}$$

with $n < 0$ for shear-thinning behavior. At high shear rates ($\gamma \gg 1/\tau_{\text{Zimm}} \sim kT/6\pi\eta R_g^3$) and in turbulent flows, however, a strongly shear-thickening behavior may be expected, as dynamic coil–stretch transitions of the polymer configuration can occur, resulting in an enhanced viscosity.

In contrast with dilute or semidilute polymer solutions is the flow behavior that occurs in polymer melts. Here the individual polymer chains interpenetrate much like a bowl of spaghetti. From the standpoint of rheology, a strong viscoelastic behavior is observed—polymer melts behave like highly viscous fluids when perturbed slowly, but any high-frequency perturbation will result in rubberlike behavior. The crossover time scale $\tau \sim \tau_0 N^a$ for such viscoelastic behavior can be

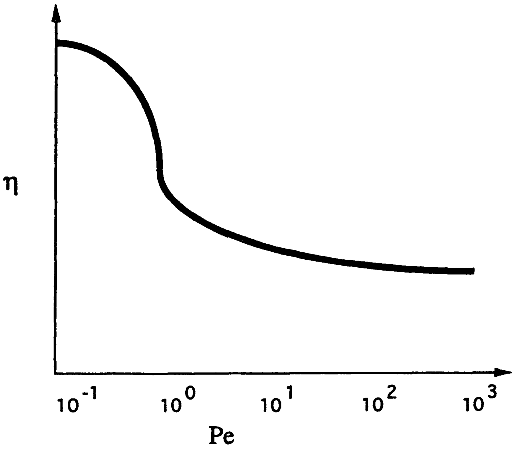

FIG. 8. Schematic diagram showing the dependence of the effective suspension viscosity on Peclet number.

of the order of seconds, and the exponent is approximately $a \approx 3$ or slightly larger. A simple argument leading to a value of $a = 3$ is a consequence of de Gennes's (1984) reptation model for polymer dynamics in melts. The idea is that each polymer molecule moves by Brownian motion in a tube created by the local configuration of the other polymers. The viscoelastic time scale can be estimated as the time for a typical molecule to travel its own length in such a tube,

$$\tau \sim L^2/D_{\text{tube}} \sim N^2/D_{\text{tube}} \sim N^3. \tag{39}$$

In deriving Eq. (39), the scaling behavior of $D_{\text{tube}} \sim N^{-1}$ was used, and this scaling behavior can be estimated from the frictional force f_{friction} that the tube exerts on the polymer. By the fluctuation-dissipation theorem, $D_{\text{tube}} \sim kT/f_{\text{friction}}$. As the frictional force can be expected to be proportional to the length of the polymer chain, we indeed find that $D_{\text{tube}} \sim N^{-1}$, and, therefore, $\tau \sim N^3$, as advertised.

2.2 Solid–Gas Phase Flows

2.2.1 Flows in Granular Material Granular materials consist of a high volume fraction of solid particles separated by a gaseous phase. In contrast to slurry flow, nearly all of the momentum is carried by these solid particles. The viscoelastic properties of granular materials can be expected to depend strongly on the local applied stresses. As stress is increased, local bonds between grains are broken, and this failure leads to quasistatic flow of moving particle clusters. At very high stresses, a "melting" transition occurs, leading to a rapid flow regime. Indeed, in many respects, granular flows have analogies to molecular dynamics. Apart from the "phase transitions" mentioned above, in the free-flow regime, the velocity of individual grains can be considered as a superposition of a local mean velocity and a random "thermal" motion

$$\mathbf{v}_i = \langle \mathbf{v} \rangle(\mathbf{r}) + \delta \mathbf{v}_i, \tag{40}$$

where a "granular temperature" can be defined by $kT = \rho a^3 \langle (\delta \mathbf{v}_i)^2 \rangle$, in analogy with normal thermal fluctuations (Ogawa, 1978). Here, ρ is the grain material density, and a is the grain radius. A fundamental difference from molecular systems, however, is that all granular collisions are inelastic, and, therefore, energy is dissipated during intergrain collisions and transformed into heat. Thus, the granular temperature $T(\mathbf{r})$ is, in contrast to simple fluids, a local property of the flow considered.

The experimentally observed behavior of the stress tensor for granular flows and slurries is

$$\sigma_{ij} = \rho a^2 f_{ij}(\phi) \gamma^2, \tag{41}$$

where γ is the shear rate (Campbell, 1990).

The dependence on shear of both the stress tensor and the granular temperature can be derived from scaling arguments based on the observation that, whereas for molecular gases, $|\langle \mathbf{v} \rangle| \ll |\delta \mathbf{v}_i|$, in granular flows, $|\langle \mathbf{v} \rangle| \sim |\delta \mathbf{v}_i|$. This observation allows us to estimate the stress tensor as $\sigma \sim \rho \langle \mathbf{v} \rangle^2 \sim \rho a^2 \gamma^2$, in agreement with Eq. (41), and the dependence on shear of the granular temperature as

$$kT = \rho a^3 \langle (\delta \mathbf{v}_i)^2 \rangle \sim \rho a^3 \langle \mathbf{v} \rangle^2 \sim \rho a^5 \gamma^2.$$

This last scaling relation can be used to estimate the effective granular viscosity for simple shear flows as

$$\eta \sim \rho a^2 \gamma \sim [\rho kT/a]^{1/2}, \tag{42}$$

and, thus, granular flows are shear thickening, with a viscosity proportional to shear or to the square root of the granular temperature. Similar arguments for the diffusion of "heat" in the presence of temperature gradients in granular flows yield, for the thermal conductivity, $\kappa \sim T^{1/2}$. These estimates are similar to those derived in the kinetic theory of gases for the transport coefficients, with the substitution of a granular temperature for the real temperature.

The dependence of the stress tensor and, consequently, the effective viscosity, on the solid volume fraction $f_{ij}(\phi)$ is also complex. As most of the momentum is carried by the solid, one might estimate $f_{ij}(\phi) \sim \phi$, and, indeed, there is a range of ϕ around $\phi \approx 0.5$ where this estimate appears to hold (see Fig. 9). But, for both limits, $\phi \rightarrow 0$ and $\phi \rightarrow 1$, $f_{ij}(\phi)$ diverges. As for the case of slurries [see Eq. (36)], at high volume fraction, no shear flow is possible, and this accounts for the divergence at the high end. At low volume frac-

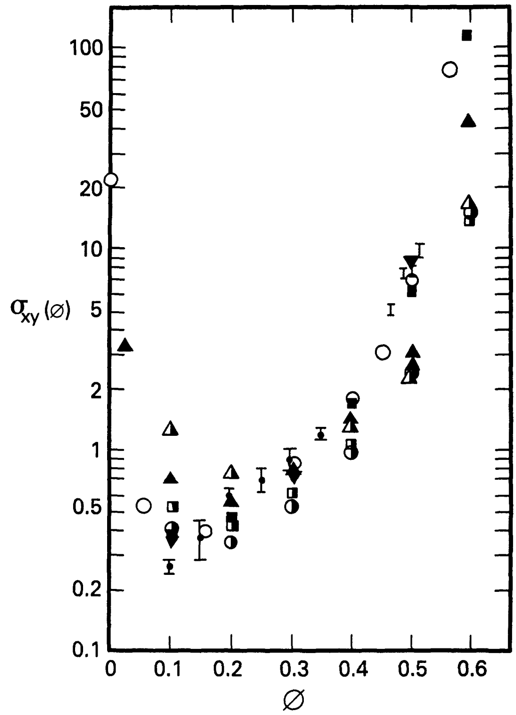

FIG. 9. Dependence of the stress tensor on the solid fraction ϕ in granular flows from various experimental studies and computer simulations; note the divergence for $\phi \rightarrow 0$ and $\phi \rightarrow 1$; redrawn from Campbell (1990). Reproduced with permission from the *Annual Review of Fluid Mechanics*, Volume 22, ©1990, by Annual Reviews Inc.

tion, the effective granular temperature diverges, as dissipation tends to zero at a much faster rate $\sim\phi^2$ (because of the lack of inelastic collisions) than energy input through gravity and the granular boundaries $\sim\phi$, resulting in an effective shear-thickening flow.

When granular materials are shaken, a fluidization transition occurs, followed by convective motions by the granular material reminiscent of the Rayleigh–Bénard convection. Consider shaking a container containing a granular material. Typically, a type of viscoelastic behavior is observed. At low frequencies, the granular material acts as a solid, at high frequencies, a fluidization transition occurs, and, at higher frequencies still, convective cells appear. If the motions of individual grains are followed in computer simulations, then at low frequencies, smooth trajectories are seen, while at higher frequencies, the granular temperature increases, and the individual grains perform Brownian motion.

An important property of polydisperse granular flows is segregation. In granular flows consisting of varying sized particles, the small grains will tend to migrate to the bottom of a container when shaken, while the large grains will be found at the top. This phenomenon, which is the opposite of diffusive mixing, can be understood in terms of the effect of gravity in the presence of hard-core repulsive interactions. The only way a large grain could possibly sink in the container would be for the random shaking to create a vacancy large enough to accommodate the grain volume. Clearly, the larger the volume, the more improbable that this event occurs—as small grains will immediately tend to fill any small vacancy due to gravity—resulting in the observed segregation.

2.2.2 Granular Topography and Self-Organized Criticality So far, we have been considering the rheology of granular flows. Granular systems and powders do not, however, only support fluidized motion; they also can behave like solids and support finite shear stresses without flow. Specifically, they can support surface structures with hill-like topographies without volume flow. If the angles of these hills are denoted by θ, then it is possible to define two angles: θ_r, the angle of repose, and θ_s, the maximum angle of stability, so that for $\theta < \theta_r$, any hill is stationary, and for $\theta > \theta_s$, the angle is unstable, leading to avalanches resulting in a new angle of tilt $\theta < \theta_s$; for $\theta_r < \theta < \theta_s$, the hills are marginally stable. The dynamics of avalanche-like behavior in this marginally stable region have recently become an active area of research with the introduction of the concept of self-organized criticality.

Self-organized criticality is the process whereby a nonequilibrium system self-organizes itself into a marginally stable macroscopic state having no intrinsic length scale and showing self-similar critical fluctuations on all length scales. Such self-organization is argued to be robust, similar to the attractor of a dynamical system with a large basin of attraction, so that for a large set of initial conditions, the system will always end up in the self-organized critical state. Typical examples of phenomena said to exhibit such self-organized criticality include earthquakes, the structure of macroscopic terrains such as mountains and sand piles, and diffusion-limited aggregation.

In the case of sand piles, the idea is that on continually adding grains to the sand surface, columns of grains will form until a gravity-induced instability results in their collapse. The resulting occurrence of an avalanche will self-organize the surface into a new marginally stable form with a self-similar distribution of hill sizes. A variety of models for avalanche flow have been simulated. The original discrete sandpile model of Bak *et al.* (1987) for self-organized criticality consisted of adding grains randomly to a lattice to form columns of height h_i at each site i. To simulate the instability due to gravity of tall columns on a two-dimensional lattice, the rule was introduced that if $h_i > h_{\text{crit}}$, then the column would lose four grains, $h_i \rightarrow h_i - 4$, while, in parallel, the nearest-neighbor columns would each gain one grain, $h_i \rightarrow h_i + 1$, conserving the granular number in the absence of external addition. This rule does indeed lead to avalanche-like behavior when a series of neighboring columns all reach marginal stability together. In addition, the fluctuations of loading times, that is, the times between avalanches, do occur with a $1/f$-like power spectrum associated with processes involving no intrinsic time scale, while the resulting topography, as well as the size distribution of avalanches, has no intrinsic length scale associated with it. Other models followed, such as replacing a critical column height h_{crit} by a critical slope $|h_i - h_j| = S_{\text{crit}}$ for avalanches to occur (Kadanoff *et al.*, 1989); they all exhibited similar types of scaling behavior, showing that self-organized criticality is a robust and ubiquitous phenomenon.

In concert with simulations of models, theoretical analyses of the current fluctuations and avalanches in a flowing sandpile can be made using driven Langevin equations incorporating the symmetries and conservation laws of the discrete models (see Sec. 3.4.3). For example, the sand forming a hill has a macroscopic flow direction, and we consider small undulations $h(\mathbf{r}, t)$ transverse to the macroscopic slope. Hwa and Kardar (1989) showed that the resulting equation for these spatiotemporal fluctuations is

$$\frac{\partial h}{\partial t} = \nu_{\|}\partial_{\|}^2 h + \nu_{\perp}\nabla_{\perp}^2 h - \frac{\lambda}{2}\partial_{\|}h^2 + \eta(\mathbf{r}, t), \tag{43}$$

which is anisotropic because of the existence of a macroscopic flow direction; and the noise due to the added sand grains is taken to be white, $\langle\eta(\mathbf{r}, t)\eta(\mathbf{r}', t')\rangle = 2D\delta(\mathbf{r} - \mathbf{r}')\delta(t - t')$.

Experiments on natural sandpiles present important differences from the predictions of self-organized criticality because dissipation effects are not correctly incorporated in the simple sandpile models described above, but qualitative similarities remain, including the fact that characteristic quantities such as angles of tilt and the loading times have no characteristic angle or time scales, but rather are given by distributions.

2.3 Liquid–Gas and Liquid–Liquid Phase Flows

2.3.1 Bubble and Emulsion Rheology

Closely related in structure and rheology to suspensions of solid particles in liquids are emulsions consisting of drops of one immiscible liquid in another, and suspensions of gas bubbles in a liquid medium.

If Stokes's equations are solved for the case of droplets of viscosity η_{drop} dispersed in a continuous phase of viscosity η_0, and the droplets are assumed to be perfectly spherical with no deformation, then Taylor (1932) first showed that, to first order in the volume fraction, Einstein's expression for the effective viscosity given by Eq. (35) is generalized to

$$\eta = \eta_0\left[1 + \frac{5\lambda+2}{2(\lambda+1)}\phi + O(\phi^2)\right], \tag{44}$$

where $\lambda = \eta_{\text{drop}}/\eta_0$. An examination of Eq. (44) shows that, as $\lambda \rightarrow \infty$, Einstein's result for solid particles is recovered, while for bubbles, the limit $\lambda \rightarrow 0$ needs to be taken with the result

$$\eta_{\text{bubbles}} = \eta_0[1 + \phi]. \tag{45}$$

Equations (44) and (45) are only valid to lowest order in the volume fraction, and for spherical droplets. For small departures from sphericity, the effective viscosity can also be expected to depend on the capillary number $Ca = \eta a\lambda/\sigma$; and, as for solid suspensions, the higher-order terms will depend on the Peclet number $Pe = \lambda a^2/D$ in the case of simple shear flows. For more complex flow patterns, droplet interaction with and deformation in the flow can be expected to complicate the rheology.

2.3.2 Cavitation Cavitation is the tendency of liquids to form bubbles spontaneously when subjected locally to negative pressures, as they cannot withstand the tension involved. Indeed, a liquid, when subject to pressures $p(\mathbf{r})$ below the vapor pressure p_v of the gaseous phase, will respond by creating bubbles to relieve this tension. Typically, such bubbles will nucleate around microbubbles of the vapor phase already dissolved in the liquid, or other impurities, and the conditions for cavitation to occur are typically created by the motion of solid bodies through the liquid phase.

To see that this is the case, consider a constant flow field with velocity $\mathbf{v}$ in a gravitational field $\mathbf{g}$. The energy density e at any point $\mathbf{r}$ in the flow is related to the pressure field $p_0(\mathbf{r})$ by Bernoulli's equation $e = \rho v^2/2 + p_0(\mathbf{r}) + \rho\mathbf{g}\cdot\mathbf{r}$. Now place a streamlined obstacle in the flow field. In the presence of this obstacle, the velocity field changes to $a(\mathbf{r})\mathbf{v}$. Consequently, the energy density of the flow may now be written $e' = \rho\alpha(\mathbf{r})^2 v^2/2 + p(\mathbf{r}) + \rho\mathbf{g}\cdot\mathbf{r}$, where $p(\mathbf{r})$ is the pressure field in the liquid in the presence of the obstacle. For steady potential flow, however, the energy density $e' = e$ remains unchanged, and therefore

$$\frac{p(\mathbf{r}) - p_v}{\rho v^2/2} = \frac{p_0(\mathbf{r}) - p_v}{\rho v^2/2} - \frac{\alpha(\mathbf{r})^2 - 1}{2}. \tag{46}$$

As cavitation is likely to occur whenever the local pressure is below the vapor pressure in the gas phase, a local dimensionless cavitation number $K(\mathbf{r}) = [p_0(\mathbf{r}) - p_v]/(\rho v^2/2)$ can be defined, which, whenever it falls below some critical value K_c that depends on the shape and size of the obstacle involved, will make it likely that cavitation occurs at that point.

Typical conditions under which cavitation would be observed would be near the blades of a rotating propeller. As the rotation speed of the propeller increases, the cavitation number decreases, and greater regions of the flow lie below K_c.

2.3.3 Foam Structure and Stability As the volume fraction of gas droplets in a liquid phase, or *foam quality*, tends to one, a transition in structure takes place from a dispersed phase of bubbles into a cell-like structure consisting of regions of the gaseous phase separated by domain walls of the liquid phase. Usually, the liquid phase will contain surfactants, that is, surface-active agents such as soap, consisting of molecules that tend to gather at the liquid–gas interface and reduce surface tension. Such surface-active additives are usually necessary in order to create the large interfacial areas that are characteristic of foams.

Two-dimensional foam patterns are best described as a distribution of cell-like domains, most of which are pentagons or hexagons, though small numbers of other polygons are also seen (see Fig. 10). The number of foam cells N_{cell}, number of edges N_{edge}, and number of vertices N_{vert} are not independent but related topologically by Euler's formula

$$N_{\text{vert}} + N_{\text{cell}} = N_{\text{edge}} + 1. \tag{47}$$

For three-dimensional foams, the gas bubbles consist of polyhedra with cell-like faces, and, again, a topological relationship can be derived, relating the number of polyhedra N_{poly} to the number of vertices, edges, and cells:

$$N_{\text{vert}} + N_{\text{cell}} = N_{\text{edge}} + N_{\text{poly}} + 1. \tag{48}$$

For two-dimensional foams, three cells meet at each vertex, forming angles of 120° (this structure can be understood on the basis of surface tension, which will tend to reduce surface area—thus, if four cells meet at a vertex, then surface energy will be reduced if this configuration is replaced by two vertices, each connected to three neighboring cells). This topology can be used to calculate the average number of faces n in a typical cell. For a large foam pattern where boundary effects can be neglected, $N_{\text{vert}} \approx nN_{\text{cell}}/3$, while $N_{\text{edge}} \approx nN_{\text{cell}}/2$. Substituting these estimates into Eq. (47) yields $n = 6$, and, therefore, the typical cells in foam are hexagons.

This extraction of structure from topology can be extended to the stability of individual cells. As was first pointed out by von Neumann (1952), to lowest-order approximation, whether an individual cell grows or decays depends only on the number of edges n in that cell. In fact, the cell area a_n of an n-sided cell grows or shrinks as a result of gaseous dif-

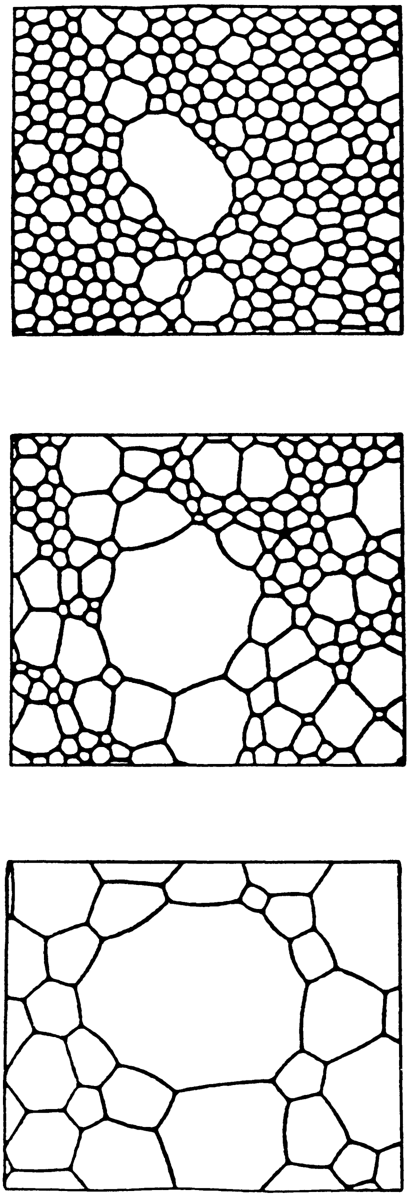

FIG. 10. Coarsening of a two-dimensional foam pattern; note also the topology of the foam cells, edges, and vertices; redrawn from Glazier (1989).

fusion through the cell membrane as

$$\frac{da_n}{dt} = \kappa(n - 6). \tag{49}$$

The derivation of Eq. (49) can be understood by treating the cell as a regular polygon consisting of n edges, each of length l and radius of curvature R. Then one would expect the rate of change of area due to gaseous diffusion to be proportional to both the total surface length of the polygon nl and the pressure drop $\Delta p \sim \sigma/R$. Thus, $da_n/dt \sim nl\Delta p \sim nl\sigma/R$, but for an n-sided polygon, from geometric arguments, $R = 3nl/(6 - n)$, yielding Eq. (49).

2.3.4 Foam Rheology and Texture Coarsening Because foam is not in a state of equilibrium, foam rheology is time and history dependent. Nevertheless, certain general properties about foam rheology can be made [see, for example, Kraynik (1988)].

Foams have strong viscoelastic properties and can behave either as elastic media or as liquids, depending on the stress applied. For small applied stresses below a material-dependent yield stress, $\sigma_{\alpha\beta} < \sigma_{\text{yield}}$, foam will deform elastically and therefore have an infinite viscosity. Above this yield stress, $\sigma_{\alpha\beta} > \sigma_{\text{yield}}$, however, foam behaves as a shear-thinning liquid with an effective viscosity that can approximately be parametrized as

$$\eta = \sigma_{\text{yield}}/\gamma + \eta_\infty. \tag{50}$$

The transport parameters in Eq. (50) are strongly dependent on foam quality and, therefore, time dependent as coarsening takes place in the foam in question.

The boundary conditions for foam flows also tend to differ from those applied to typical Newtonian fluids. Instead of stick boundary conditions as in Poiseuille flow, foams tend to slip at walls with a slip velocity $v_{\text{slip}} = \Psi_w\sigma_w$, where σ_w is the wall shear stress and Ψ_w is a slip coefficient. In addition, for small applied stresses below the yield stress, $\sigma_{\text{slipyield}} < \sigma < \sigma_{\text{yield}}$, though no relative motion of the cell or polyhedral domains will occur, plug flow is still possible. For even smaller stresses $\sigma < \sigma_{\text{slipyield}}$, even plug flow is impossible.

Foams cannot be treated as systems with a static structure and a consequent static rheology, as they are constantly coarsening with time. Coarsening is the property of foams whereby large cells tend to grow at the expense of small ones, which can also burst, leading to fewer but larger cells with time, a well-known property of foam, as anyone who has studied the head of a glass of beer knows.

In real foams, this has many causes: Free-energy minimization will tend to reduce surface area, and this will encourage large bubbles to grow at the expense of small ones; many-faceted cells will be favored at the expense of few-faceted ones, as Eq. (49) suggests; liquid films between cells will rupture and bubbles coalesce, again encouraging coarsening. One main cause of such rupturing is the thinning of bubble walls due to drainage in the presence of gravity. Gravity will also tend to cause an overall spatial variation in foam texture and, consequently, transport properties.

von Neumann's equation [Eq. (49)] can be used to estimate this coarsening process in a foam of total area A. Assume that polygons with greater than six sides grow, while three-, four-, and five-sided polygons are continually bursting, so that their numbers are decreasing in time. The manner in which the average number of cells $N(t)$ in this piece of foam diminishes in time as a result of this coarsening process and the concomitant increase in their average polygonal area $a(t)$ can be estimated from scaling arguments. First, conservation of area implies that $N(t)a(t) = A$. Second, assuming that the loss of cells comes solely from the disappearance of three-, four-, and five-sided polygons at a rate set by von Neumann's law, one would expect

$$\begin{aligned}\frac{dN}{dt} &\sim \sum_{n=3,4,5} a_n^{-1}(t)N_n(t)\frac{da_n(t)}{dt}\\ &\sim -\kappa \sum_{n=3,4,5} a_n^{-1}(t)N_n(t)(n-6),\end{aligned}$$

where $N_n(t)$ is the number of n-sided polygons at time t. To estimate the size of this sum, assume that in the scaling state, $a_n(t) \sim a(t)$ and $N_n(t) \sim N(t)$, in which case $dN/dt \sim -N(t)^2$. Thus, the number of cells diminishes with time as $N(t) \sim t^{-1}$, while their average area grows linearly with time, $a(t) \sim t$.

The effect of surfactant on texture coarsening can also be important. The relationship between the concentration of surfactants c in the bulk of the liquid, the concentration $\Gamma(c)$ at the liquid–gas interface, and the reduction in surface tension is $d\sigma/d(\ln c) = -kT\Gamma(c)$. For small concentrations of surfactant, $\Gamma(c) \sim c$, and, therefore, the reduction in surface tension is linear with volume surfactant concentration; but this relationship must break down, as only a finite concentration of surfactant molecules can occupy the interface, and, therefore, there exists an asymptotic limit to surface-tension reduction. Thus, during drainage of fluid from films in the presence of gravity, surface-tension gradients will appear, which will result in forces (the Marangoni effect) that can change the drainage process significantly.

The same considerations apply to bubble collapse due to gaseous diffusion through the cell wall under the pressure $\Delta p(t) = 2\sigma/R(t)$. Solving for the dependence of radius on time, one finds $R_0^2 - R(t)^2 \sim \sigma t$, where R_0 is the initial bubble radius, and, therefore, the bubble disappears in a finite time. In the presence of surfactant, however, as the bubble decreases in radius and therefore in surface area, the surfactant concentration at the surface increases. Therefore, the surface tension decreases, resulting in a bubble that lasts a much longer time than the simple estimate $\tau \sim R_0/\sigma$ would suggest. A good review of foam structure and dynamics can be found in Glazier (1989).

2.4 Anisotropic Flows

Though clearly all the flows we have considered thus far involve interfaces between phases at either a macroscopic or a mesoscopic scale, in the main they remain macroscopically isotropic. That is, their flow properties do not have any macroscopic directionality with flow. There are classes of liquids, however, that do have a specific macroscopic orientation associated with their structure, for example, nematic liquid crystals (see LIQUID CRYSTALS, STRUCTURE OF) and electrorheological fluids (see ELECTRORHEOLOGY), and, in consequence, their flow properties are profoundly changed from simple Newtonian flow.

2.4.1 Nematic Liquid-Crystal Rheology Liquid crystals are liquids consisting of long rigid organic molecules that, because of short-range repulsion, tend to align, like logs in a river, forming a phase intermediate between a simple liquid and a crystal. If the ordering is in both the position and orientation of the molecules, then the phase is called a smectic, while if the positions of the molecules are random but orientational ordering survives, then a nematic phase exists. Such a phase can be described macroscopically by an order parameter $\mathbf{n}(\mathbf{r})$, called the director, which gives the local average orientation of the molecules at point **r**. The existence of such a director will affect flow in several ways, as the liquid possesses macroscopic orientational order.

First, the stress tensor can no longer be expected to be symmetric as is the case for simple fluids, because macroscopic torques

$$\Gamma_\alpha(\mathbf{n}, \nabla\mathbf{n}) = \epsilon_{\alpha\beta\chi}\sigma_{\beta\chi} \tag{51}$$

can be applied to to nematic liquid crystals whose magnitude can be calculated from the distortion energy density of the nematics [see, for example, de Gennes (1974)]. These torques will distort the constant orientation of the nematic phase; and if in the undistorted state the director is taken to be a constant vector in the z direction, $\mathbf{n}(\mathbf{r}) = n_z\mathbf{k}$, then the basic distortions in orientation displayed by the nematic phase are splays $\partial n_x/\partial x$ or $\partial n_y/\partial y$, twists $\partial n_y/\partial x$ or $\partial n_x/\partial y$, and bends $\partial n_x/\partial z$ or $\partial n_y/\partial z$ in the director.

Second, the dissipative contributions to the stress tensor $\sigma'_{\alpha\beta}$ in the equation of motion for the momentum density

$$\rho dv_\alpha/dt = -\partial_\alpha p + \partial_\beta\sigma'_{\alpha\beta} \tag{52}$$

no longer depend only on the symmetric rate-of-strain tensor $S_{\alpha\beta} = \partial_\alpha v_\beta + \partial_\beta v_\alpha$ (assuming the nematic to be incompressible), but will also depend on the rate of rotation of the director

$$\mathbf{N} = d\mathbf{n}/dt - \mathbf{\Omega} \times \mathbf{n} \tag{53}$$

additional to that ($\mathbf{\Omega} = \frac{1}{2}\nabla \times \mathbf{v}$) induced by vorticity alone.

To write down the constitutive relation for $\sigma_{\alpha\beta}$, one symmetry can, however, be invoked. Because the flow for nematics will depend on the orientation of the rodlike molecules, but will remain unaffected by a flip through 180° (consider again the analogy to logs in a river; the tops and bottoms of the log are identical in contrast to dipole moments), it follows that the stress tensor must be invariant under the flip $\mathbf{n}(\mathbf{r}, t) \rightarrow -\mathbf{n}(\mathbf{r}, t)$ and the associated flips in the rate of rotation $\mathbf{N}(\mathbf{r}, t) \rightarrow -\mathbf{N}(\mathbf{r}, t)$. Therefore, the Leslie stress tensor for nematics can be written

$$\begin{aligned}\sigma'_{\alpha\beta} = {} & \alpha_4 S_{\alpha\beta} + \alpha_1 n_\gamma S_{\gamma\delta} n_\delta n_\alpha n_\beta + \alpha_5 n_\alpha n_\gamma S_{\gamma\beta} \\ & + \alpha_6 n_\gamma S_{\gamma\alpha} n_\beta + \alpha_2 n_\alpha N_\beta + \alpha_3 N_\alpha n_\beta,\end{aligned} \tag{54}$$

where the α's are transport coefficients of which only $\alpha_4 = \eta$, the shear viscosity, survives for a simple fluid. Notice that the Leslie stress tensor obeys the flip symmetry and is linear in the rate of strain $S_{\alpha\beta} = (\partial_\alpha v_\beta + \partial_\beta v_\alpha$. For a nematic, therefore, effective shear viscosities can be defined depending on the direction of the rate of shear relative to the director, and, in consequence, despite the linearity in rate of shear of the stress tensor, nematics can be expected to display a non-Newtonian orientational dependence in their flow characteristics.

2.4.2 Dynamics of Electrorheological Fluids In contrast with nematic liquid crystals, electrorheological fluids consist of polarizable particles dispersed in nonpolar solvents, such as silica gel in oil. They also exhibit orientational characteristics, forming in the presence of strong electric fields columns of particles that aggregate together at long times (see ELECTRORHEOLOGY).

The dynamics of column formation and the observed coarsening in electric fields have been studied extensively recently (Halsey, 1992). The orientational ordering appears to involve two time scales: the time $\tau_1(E)$ after switching on the field for columns of the polarizable particles to form; then a longer time scale $\tau_2(E)$, on which the columns appear to drift together as a result of thermal fluctuations, ending finally in one giant aggregate. Scaling arguments can be made to estimate these time scales. If the polarizable particles are taken to have a radius r_p and polarizability α, then the drift velocity $\mathbf{v}$ of the particles in an electric field can be found by balancing the attractive dipole-dipole force $f \sim d^2/r_p^4$, causing chain formation (the induced dipole moment per particle $\mathbf{d} = \alpha\mathbf{E} \sim \beta r_p^3\mathbf{E}$), to the Stokes friction $f \sim \eta v r_p$ leading to $v \sim d^2/\eta r_p^5 \sim E^2/\eta$. Thus, the time for column formation in a container of length scale L can be expected to scale as $\tau_1(E) \sim L/v \sim L\eta/E^2$. The sharp decrease in column formation time with increasing applied field has been observed experimentally (Ginder and Elie, 1992). From scaling, we may also expect that $\tau_2(E)/\tau_1(E) = f(\lambda)$, where the dimensionless constant $\lambda = (\beta E)^2 r_p{}^3/kT$ measures the ratio of dipole-dipole interactions $d^2/r_p{}^3$ to thermal fluctuations kT.

In consequence of this chain formation, in large electric fields the viscosity has been observed to increase by a factor of $\eta(E \text{ large})/\eta \sim 10^5$. The actual response of electrorheological fluids in applied electric fields E to stress is viscoelastic, with a yield stress that scales as $\sigma_{\text{yield}} \sim CE^2$. Under small stresses $\sigma \ll \sigma_{\text{yield}}$, the fluid behaves as an elastic solid, while if large stresses $\sigma \gg \sigma_{\text{yield}}$ are applied, the electrorheological fluid will flow, though with a markedly increased effective viscosity $\eta(E)$.

3. FLOW IN POROUS MEDIA

Studies of both miscible and immiscible flows in porous media represent a rich field of present-day research dependent on many of the concepts introduced previously.

In the case of immiscible flows, the structures of the rough interfaces occurring during drainage and imbibition experiments often turn out to be fractal or self-affine (see FRACTAL GEOMETRY). This interfacial structure can be expected to be dependent on many variables, including whether the flow is wetting or nonwetting, because of the large liquid–solid interface involved; whether the strength of capillary forces is significant, because in porous media of low permeability κ and consequent small pore sizes $a \sim \kappa^{1/2}$, the pressure drops due to surface tension σ in pores, $\Delta p_{\text{tension}} \sim \sigma/a \sim \sigma/\kappa^{1/2}$, can be large compared to the viscous pressure drop, $\Delta p_{\text{viscous}} \sim U\eta/a \sim U\eta/\kappa^{1/2}$; the magnitude of the viscosity ratios $M = \eta_1/\eta_2$, because of the existence of the Saffman–Taylor instability; and whether gravity interacts on the evolving interface, because of the Taylor instability—gravity may be expected to have a significant effect when the hydrostatic pressure drops $\Delta p_{\text{gravity}} \sim \Delta\rho g\kappa^{1/2}$ that occur across pores become large in comparison with capillary and viscous pressure drops.

In the case of miscible flows, such questions as the nature of the anomalous diffusion, which controls dispersion of passive scalars, and how these transport coefficients scale with flow velocity in porous media have been studied.

3.1 Microscopic Equations of Motion

A porous medium can be thought of as consisting of a tortuous, multiply connected domain of pores connected by throats. Such porous media are created naturally by grains of sand or other material, typically random in shape and coming in a distribution of sizes (though in laboratory simulations, often monodisperse glass beads are used instead), that are packed into a random configuration. Typically, these configurations do not vary over time, and, hence, the porous medium forms a quenched (that is, time independent) random topography through which the fluid must flow. The terms pores and throats are not well-defined concepts, but represent the idea that the topology of a porous medium can usually be considered to consist of fairly large interstitial holes connected by long thin channels. The Reynolds numbers for typical flows in porous media are usually very small, $Ua/\nu \ll 1$, and, for such creeping flows, neglect of inertial terms is valid. Therefore, in principle, for the case of multiphase flow, the full set of equations that need to be solved are Stokes flow in a gravitational field $\mathbf{g}$,

$$\frac{\partial \mathbf{v}_\alpha}{\partial t} = -\frac{1}{\rho_\alpha}\nabla p + \nu_\alpha \nabla^2 \mathbf{v}_\alpha + \mathbf{g}, \tag{55}$$

supplemented by the incompressibility conditions $\nabla \cdot \mathbf{v}_\alpha = 0$. Here, the subscripts $\alpha = 1,2$ represent respectively the displaced and displacing fluids [oil (1) being displaced by water (2), for example] in the case of two-phase flow.

The boundary conditions are all important for flows in these contorted domains with large liquid–liquid and liquid–solid interfaces. At any point $\mathbf{s}$ of the curved liquid–liquid interface between the immiscible fluids, the normal component of the velocity across the interface must be continuous, while the pressure difference due to surface tension σ is discontinuous, leading to the boundary conditions

$$\begin{gathered} \mathbf{v}_{1,n}(\mathbf{s}) = \mathbf{v}_{2,n}(\mathbf{s}) = 0, \\ p_2(\mathbf{s}) - p_1(\mathbf{s}) = \sigma[R_1^{-1}(\mathbf{s}) + R_2^{-1}(\mathbf{s})], \end{gathered} \tag{56}$$

where $R_1(\mathbf{s})$ and $R_2(\mathbf{s})$ are the two principal radii of curvature of the interface at $\mathbf{s}$. In contrast, at the liquid–solid interface, stick boundary conditions can be assumed:

$$\mathbf{v}_1\,(\text{substrate}) = \mathbf{v}_2\,(\text{substrate}) = 0. \tag{57}$$

The complexity of this flow lies in the stick boundary conditions at the liquid–substrate interface and the free boundary conditions at the liquid–liquid interface. The shape of the domain is typically so tortuous that analytical approaches to studying such flows starting from the microscopic equations Eqs. (55)–(57) are hopeless, and the major theoretical tools developed to deal with this class of flows have been the derivation of macroscopic phenomenological equations to account for gross flow behavior. More recently, this approach has been supplemented by computer simulations on simplified models, such as invasion percolation and gradient percolation, coupled with scaling arguments to account for both interfacial roughness and transport.

3.2 Macroscopic Transport in Porous Media

Structure and transport in porous media have been described macroscopically in terms of a few parameters, such as porosity ϕ, permeability κ, effective surface tension σ_{eff}, and saturation S, which represent simple macroscopic statistical averages of the very complicated nonhomogeneous flow taking place in the porous medium [see, for example, Adler and Brenner (1988)]. Phenomenologically, Darcy's law,

$$\mathbf{v} = -(\kappa/\eta)\nabla p, \tag{58}$$

has been used to describe the flow dynamics. It represents a equation similar to Ohm's law, describing the linear relationship between flux and applied pressure for single-phase flow in porous media. Such equations of motion and the macroscopic transport coefficients involved are averages on length scales L much larger than the pore size, $L \gg a$, and, consequently, can only be expected to be valid on macroscopic flow scales. The porous medium is also usually supposed to be homogeneous and isotropic on macroscopic scales. If any of these conditions are not met, then the phenomenology must be altered to take the new conditions into account. For example, if the porous medium is heterogeneous on all scales considered, then all averages will depend on sample size, while if the medium is macroscopically anisotropic, then tensor equations can be expected to describe the flow, and Darcy's law will be generalized to $v_i = -(\kappa_{ij}/\eta)\,\partial p/\partial x_j$.

The porosity ϕ is the volume fraction not occupied by the solid matrix, that is, the volume fraction of the throats and pores making up the flow domain in a fully connected porous medium. Typical values for the porosity might vary from 5% to 20% in sandstones or limestone.

The permeability to single-phase flow κ can be found from Darcy's law by measuring the flux of liquid as a function of the pressure drop. Its dimensions are length squared, $[\kappa] = L^2$, and its magnitude can be estimated for many porous media as $\kappa \sim a^2$, where a is a typical pore size. It must be remembered, however, that permeability depends not only on porosity but also on connectivity, and this estimate may break down for poorly connected media—a porous medium can have a high porosity, yet if no connected path exists from one end of the porous medium to the other, its permeability is zero. For typical porous media forming geological reservoirs, the pore sizes can vary from 10^{-6} to 10^{-2} cm, and, therefore, the permeability will vary to a much greater degree than porosity, covering up to eight orders of magnitude in the range $10^{-12}\ \text{cm}^2 < \kappa < 10^{-4}\ \text{cm}^2$, depending on pore size and connectivity [see, for example, Pittman (1983)].

The pressure differences across interfaces when measured on a macroscopic scale represent a weighted average of the pressure differences locally at each throat and pore, which will depend on the instantaneous position of the interface and the wettability of the underlying substrate. Consequently, the effective surface tension σ_{eff} measured macroscopically in a porous medium can be very different from the microscopic value σ between two pure phases. For example, in the case of oil and water, the ratios $\sigma_{\text{eff}}/\sigma \approx 5.5$ for oil-wetted media and $\sigma_{\text{eff}}/\sigma \approx 306$ for water-wetted media have been found.

In the case of immiscible two-phase flow, a permeability for both the displaced and displacing species can be defined from the pair of Darcy's-law equations $\mathbf{v}_1 = -(\kappa_1/\eta_1)\nabla p$ and $\mathbf{v}_2 = -(\kappa_2/\eta_2)\nabla p$; but, in this case, the permeability of each species will be a function of the saturation S, the volume fraction of each species in the interstitial flow domain. Calling the saturation of the displaced fluid S_1 and that of displacing fluid S_2, then, naturally, $S_1 + S_2 = 1$. The two-phase permeabilities κ_1 and κ_2 are much more complex than those for single-phase flow and depend not only on saturation, but on the surface tension between the phases σ, on the viscosity ratio M, and also on the history of the flow, as strong evidence for hysteresis is often observed during drainage and imbibition experiments.

3.3 Drainage and Imbibition

Two of the most important classes of flow in porous media are drainage and imbibition. In a drainage experiment, a porous medium is first saturated with a wetting liquid ($S_1 = 1, S_2 = 0$), and then a nonwetting fluid is used to displace the wetting fluid. For this to occur, an external pressue p must be applied. For small pressures $p < p_{c2}$, small amounts of the wetting fluid are displaced, to be replaced by small nodules of nonwetting fluid, but there is no macroscopic flow of the displacing liquid, and the permeability of the displacing phase is zero, $\kappa_2 = 0$ (see Fig. 11). When a pressure is applied sufficient to overcome the capillary forces, $p > p_{c2}$, then a macroscopic flow of the nonwetting phase is suddenly seen, and $\kappa_2 > 0$. Thus, a critical pressure p_{c2} exists for drainage to occur, and one can estimate this critical pressure, which is due to surface tension, as $p_{c2} \sim \sigma/a$. One can also define X_1 and X_2, the volume fractions of interstitial region occupied by the displaced and displacing phases, which are parts of domains spanning the entire sample and through which macroscopic flow can therefore occur; in contrast, the saturations S_1 and S_2 will not distinguish between spanning clusters and small isolated nodules of fluid. Clearly, $X_2(p) = 0$ for $p < p_{c2}$, and $X_2(p) > 0$ for $p > p_{c2}$. A percolation transition has occurred at p_{c2}.

Percolation transitions are geometric phase transitions that involve the clusters formed from sites (or bonds) on a lattice that are occupied randomly with a probability P [see, for example, Stauffer (1985)]. For very small P, only isolated clusters of occupied sites exist. But, as $P \rightarrow P_c$, the size of the typical clusters diverges as $\xi \sim (P_c - P)^{-\nu}$, and an incipient infinite cluster that spans the lattice appears (see Fig. 12). That such a cluster must appear is clear because when $P = 0$, no such cluster exists, $X(0) = 0$, and when $P = 1$, every site belongs to this cluster, $X(1) = 1$. A geometric phase transition is involved because the behavior of $X(P)$ is not analytic; rather $X(P) = 0$ for $P < P_c$, and $X(P) \sim (P - P_c)^{\beta}$ for $P > P_c$. One of the most important developments in modern physics has been the realization that certain properties of percolation, as well as other phase transitions, are universal (see UNIVERSALITY) and, therefore, independent of the microstructure of the lattice—for example, critical exponents, such as β and ν—while

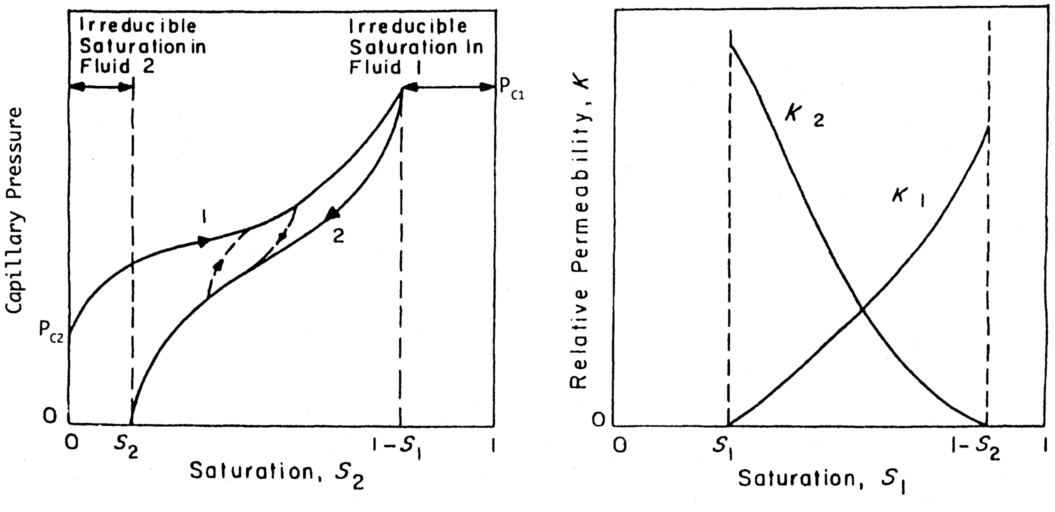

FIG. 11. Plots of the capillary pressure and relative permeabilities as function of saturation during drainage and imbibition; redrawn from Adler and Brenner (1988). Reproduced, with permission, from the *Annual Review of Fluid Mechanics*, Volume 20, ©1988, Annual Reviews Inc.

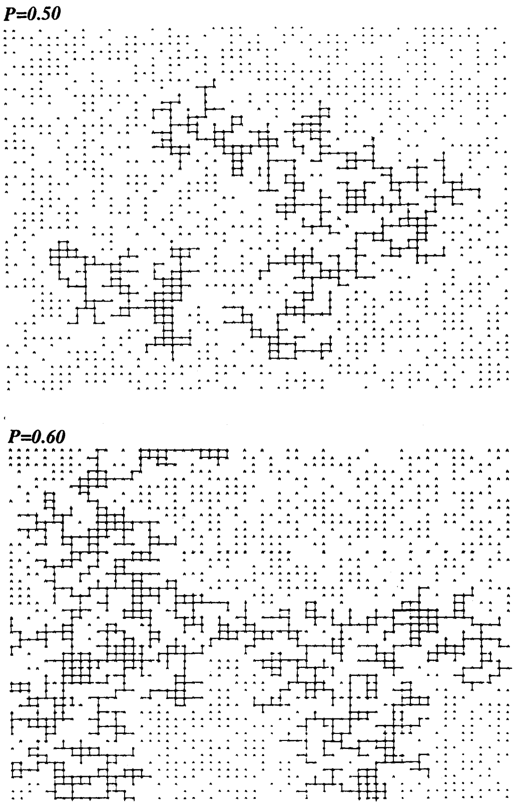

FIG. 12. The percolation transition on a square two-dimensional lattice. Here $P_c = 0.5928\ldots$, and the largest clusters for (a) $P = 0.5$ and (b) $P = 0.6$ are shown; redrawn with permission from Stauffer (1985).

others are strongly lattice dependent, such as P_c.

Returning now to flow in porous media, the universality does not reside in the pressure p_{c2}—indeed, the estimate $p_{c2} \sim \sigma/a$ is dependent on both pore size and surface tension—but rather in the exponents t and β, which control respectively the manner in which the permeability $\kappa_2(p) \sim (p - p_{c2})^t$ and the probability $X_2(p) \sim (p - p_{c2})^\beta$ that a piece of the displacing fluid belongs to a spanning cluster increase for pressures above the critical pressure $p > p_{c2}$. The exponents t and β, among many others, are the same as for percolation and depend only on dimension. For example, the exact value for the percolating fraction exponent is $\beta = 5/36$ in two dimensions, while in three dimensions, it is $\beta \approx 0.417$; similarly, the permeability exponent t, which is identical to the conductivity exponent in random resistor networks, is $t \approx 1.3$ in two dimensions, while in three dimensions, it is $t \approx 2.0$. One's intuitions might suggest that $\kappa(p) \sim X(p)$ and, therefore, $t = \beta$, but only a fraction of the pores and throats constituting the spanning cluster actually have pressure drops across them and, therefore, carry flowing fluid. Thus, t depends not only on the spanning cluster geometry but also on the fluid flow on that cluster, though, as these scaling relations are valid as $p \to p_{c2}$, we can expect $t > \beta$.

As the pressure of the nonwetting fluid is increased well beyond p_{c2}, more and more of the wetting fluid is displaced from the porous medium, and both $\kappa_1 > 0$ and $\kappa_2 > 0$. At some pressure $p_{c1} > p_{c2}$, a second percolation transition occurs when the displaced wetting fluid itself no longer spans the porous medium [$X_1(p) = 0$ for $p > p_{c1}$, and $X_1(p) > 0$ for $p < p_{c1}$], but rather forms a set of disconnected clusters. At this point, $\kappa_1 = 0$, and an irreducible fraction of the wetting fluid $S_{1,\mathrm{irred}}$ is left behind in the porous medium.

If the external pressure is now reduced again, the wetting fluid will begin to replace the nonwetting fluid. This reverse process is called imbibition. More generally, an imbibition experiment would begin with a porous medium filled with a nonwetting fluid, which is replaced by a wetting fluid—for example, water soaking into a piece of paper. No external pressure needs to be applied here because of the existence of a positive spreading coefficient $S = \sigma_{SV} - \sigma_{LS} - \sigma$.

It might appear that a cyclic reversible process of drainage followed by imbibition could be set up. This is not the case; the process is hysteretic because, though an experiment with a porous medium soaked with a wetting fluid could be set up, after one drainage and imbibition cycle an irreducible amount of nonwetting fluid $S_{2,\mathrm{irred}}$ would remain in the medium at the point when the capillary pressure is zero. After one cycle, at least an irreducible minimum of each phase must exist in the porous medium, $S_{1,\mathrm{irred}} < S_1 < 1 - S_{2,\mathrm{irred}}$ and $S_{2,\mathrm{irred}} < S_2 < 1 - S_{1,\mathrm{irred}}$, and in this region, each phase spans the porous medium. The irreducible minimum of each phase is the saturation at the percolation transition; thus, at $S_1 \approx S_{1,\mathrm{irred}}$ where the wetting permeability tends to zero as $\kappa_1 \sim (p_{c1} - p)^t$, and $S_2 \approx S_{2,\mathrm{irred}}$, where the nonwetting permeability tends to zero as $\kappa_2 \sim (p - p_{c2})^t$.

3.4 Models for Fluid Flow in Porous Media

The discussion on drainage and imbibition above combines empirical observations with percolation theory. But the theoretical isomorphism drawn between conductivity on a percolating lattice and two-fluid flow in a porous medium is tentative: For example, in contrast to exact percolation, there is no reason in two-fluid flow below the percolation transition why locally large pressure gradients cannot exist capable of mobilizing individual blobs of fluid that can then flow or fuse with other clusters. In addition, real two-fluid flow has a large number of variables that can be expected to affect the interfacial structure and flow but that do not appear in normal percolation theory, including surface tension σ, gravity g, viscosity ratios M, flow rate U, the nature of the porous medium, and wetting properties. All these variables, which will give rise to a large phase diagram, need to be investigated. What is more, the saturation of each species in the porous medium and the permeability only give partial information on structure. More recently, fractal and self-affine properties of interfaces in two-phase flow have been the subject of detailed investigations.

To simulate aspects of this phase diagram and to study interfacial structure, new models and approaches have been introduced. Invasion percolation (de Gennes and Guyon, 1978; Wilkinson and Willemsen, 1983) has been used to study low–capillary-number [$Ca = (\eta_1 - \eta_2)\ U/\sigma \ll 1$] creeping flows dominated by surface tension. In the opposite limit of high capillary numbers, surface tension can be neglected, and diffusion-limited fingering is observed. In the presence of gravitational fields, the relevant dimensionless control variable is the Bond number $B = \Delta\rho g\kappa/\sigma$, and the model of gradient percolation (Sapoval *et al.*, 1985; Gouyet *et al.*, 1990) can be used to study its influence on flow.

Hand in hand with new models has gone the development of new theoretical approaches to understanding flow in porous media. For example, the influence of the porous medium on two-fluid flow has been analyzed theoretically in terms of effective Hamiltonians that were shown to lie in the same universality class as random-bond and random-field Ising models (Brochard and de Gennes, 1983), while the realization that surfaces in imbibition experiments are often self-affine can be accounted for in terms of the self-affine solutions of a class of forced partial differential equations (Edwards and Wilkinson, 1982; Kardar *et al.*, 1986).

3.4.1 Invasion Percolation How do drainage interfaces depend on capillary number Ca and the viscosity ratio M of the fluids involved? To answer this and related questions, models for porous media and for the flows in such quenched random topographies are introduced and simulated.

To simulate fluid flow in porous media, lattice models are often used to represent the pores and throats. The pores can be represented by sites on this lattice, while the throats are represented by bonds. The quenched random nature of the porous medium is represented by assigning a set of random variables r_{ij} to the bonds $\langle ij \rangle$ between sites i and j, representing the variable radius of the throat considered.

To model drainage of a wetting fluid by a nonwetting fluid in the limit of low capillary numbers involves creeping flows, and the contribution of viscosity to the pressure drop across a throat can be neglected. Therefore, the process of drainage in this limit is entirely due to the capillary forces $p_{ij} = 2\sigma/r_{ij}$ across each throat. Assume that an external pressure p is applied and that flow can occur in any throat provided $p > p_{ij}$. If all throats obeying $p > p_{ij}$ were filled with the nonwetting fluid, the model would be isomorphic to the pure percolation problem discussed in Sec. 3.3. But, in drainage, the external pressure is applied along one edge of the porous medium, and the displacement of wetting fluid by the nonwetting invader begins at this edge. To model this flow, the invasion percolation model was introduced (de Gennes and Guyon, 1978; Wilkinson and Willemsen, 1983) describing the process of filling the active ($p > p_{ij}$) sites. An invasion edge is chosen (hence the name invasion percolation) where the external pressure is presumed to be applied, and, sequentially, the most active bond available at that point in time, that is, the bond with largest radius r_{ij}, is filled with nonwetting fluid. In consequence, several differences exist from normal percolation.

First, invasion percolation is a dynamic process. Second, throats can only be occupied by the nonwetting fluid if there exists a contiguous path of active sites from the initial invasion edge of the nonwetting fluid. Therefore, for $p < p_c$, there exist no disconnected clusters of occupied sites as in true percolation, but, rather, a finite invasion of the porous medium by the nonwetting fluid occurs. Third, trapping of the wetting fluid by the nonwetting phase can occur because of the incompressibility of the fluids involved. Consider an active site that is a candidate for the replacement of the wetting phase by the nonwetting phase. If a contiguous path exists to the opposite face of the porous medium from which the drainage of the wetting phase is occurring, invasion is possible; but if the candidate active throat filled with wetting fluid is surrounded by the invading nonwetting phase, then invasion of the throat is impossible, as the ejected wetting fluid would have nowhere to go. Simulations (Chandler *et al.*, 1982) appear to yield exponents close to those one would predict on the basis of pure percolation arguments, and so do real experiments involving displacement of glycerol by air (Lenormand and Zarcone, 1985). For example, the saturation $S_2(W)$ on a finite-size lattice of width W scales at breakthrough on a two-dimensional lattice as $S_2(W) \sim W^{-0.17}$. Finite-size scaling arguments would predict $S_2(W) \sim W^{-\beta/\nu}$, where $\beta/\nu = 15/144 \approx 0.11$ for normal percolation.

Invasion percolation should be valid at $Ca = 0$. For finite flow rates, the pressure drops due to viscosity must be included in simulations (Chen and Wilkinson, 1985; Lenormand *et al.*, 1988). If Poiseuille flow is assumed in each throat, then the flux J_{ij} of fluid across each bond $\langle ij \rangle$ if pressures p_i and p_j exist at sites i and j is

$$J_{ij} = \pi r_{ij}^4 (p_i - p_j - p_{ij})/8l\eta \tag{59}$$

when the pressure drop is greater than the capillary pressure, $p_i - p_j > p_{ij}$, and the flux is zero, $J_{ij} = 0$, otherwise. In Eq. (58), l is the length of the throat, and η is the average viscosity. Incompressibility demands that $\Sigma_j J_{ij} = 0$, where the sum is over nearest-neighbor sites j of each site i. This set of equations has been simulated by Lenormand *et al.* (1988) as a function of the capillarity Ca and viscosity ratio $M = \eta_2/\eta_1$. Three qualitative "phases" of the flow can be seen in both numerical simulations and real experiments (see Fig. 13): a stable flat interface between the immiscible fluids, viscous fingering, and capillary fingering.

In the limit of very small capillary numbers, capillary fingering is observed, described by invasion percolation. As the capillary number increases and, therefore, surface-tension effects are less dominant, a stable flat interface appears, provided a more viscous fluid is used to displace a less viscous one, $M > 1$, as the Saffman–Taylor stability analysis would suggest. What is perhaps most surprising about these simulations is that crossover does not occur when $Ca \approx 1$, but rather when $Ca \leqslant 10^{-2}$. In the opposite limit, $M \ll 1$ and $Ca \gg 10^{-2}$, viscous fingers appear, as diffusion-limited arguments would suggest.

3.4.2 Gradient Percolation The influence of gravity on the stability of two-fluid flow has been shown by Taylor (1950) to be substantial. What are the consequences for drainage in a gravitational field? Without a gravitational field, breakthrough of the nonwetting phase always occurs whatever the system size under the application of a finite pressure. Against-gravity breakthrough will never occur in an infinite system. Rather, a macroscopic invasion front of length z_m can be achieved, depending on the applied pressure, and the interface of the front is rough—indeed, fractal. Experimentally (Clement *et al.*, 1985; Gouyet *et al.*, 1990), the structure of the invasion front can be studied by, for example, forcing nonwetting Wood's metal, a liquid alloy, through packs of crushed glass, then freezing the metal and examining slices at various heights z (assume that the invasion front begins at $z = 0$, and that the slices are taken for $z > 0$, in a gravitational field pointing downward, $\mathbf{g} = -g\mathbf{k}$).

The saturations of both the displaced $[S_1(z)]$ and displacing $[S_2(z)]$ phases are functions of the height z, and the same applies to the probability of belonging to a spanning cluster for both the displaced $[X_1(z)]$ and displacing $[X_2(z)]$ phases. For small z, the crushed glass is almost completely saturated by the Woods metal, and, therefore, $S_2(z) \approx 1$, with only a few unconnected blobs of the displaced medium, $S_1(z) \approx 0$, in this case, air [similarly, $X_1(z) = 0$, while $X_2(z) > 0$]. At some critical height z_{c1}, a percolation transition occurs in the displaced phase, and, for $z_{c1} < z < z_{c2}$, both the wetting and nonwetting

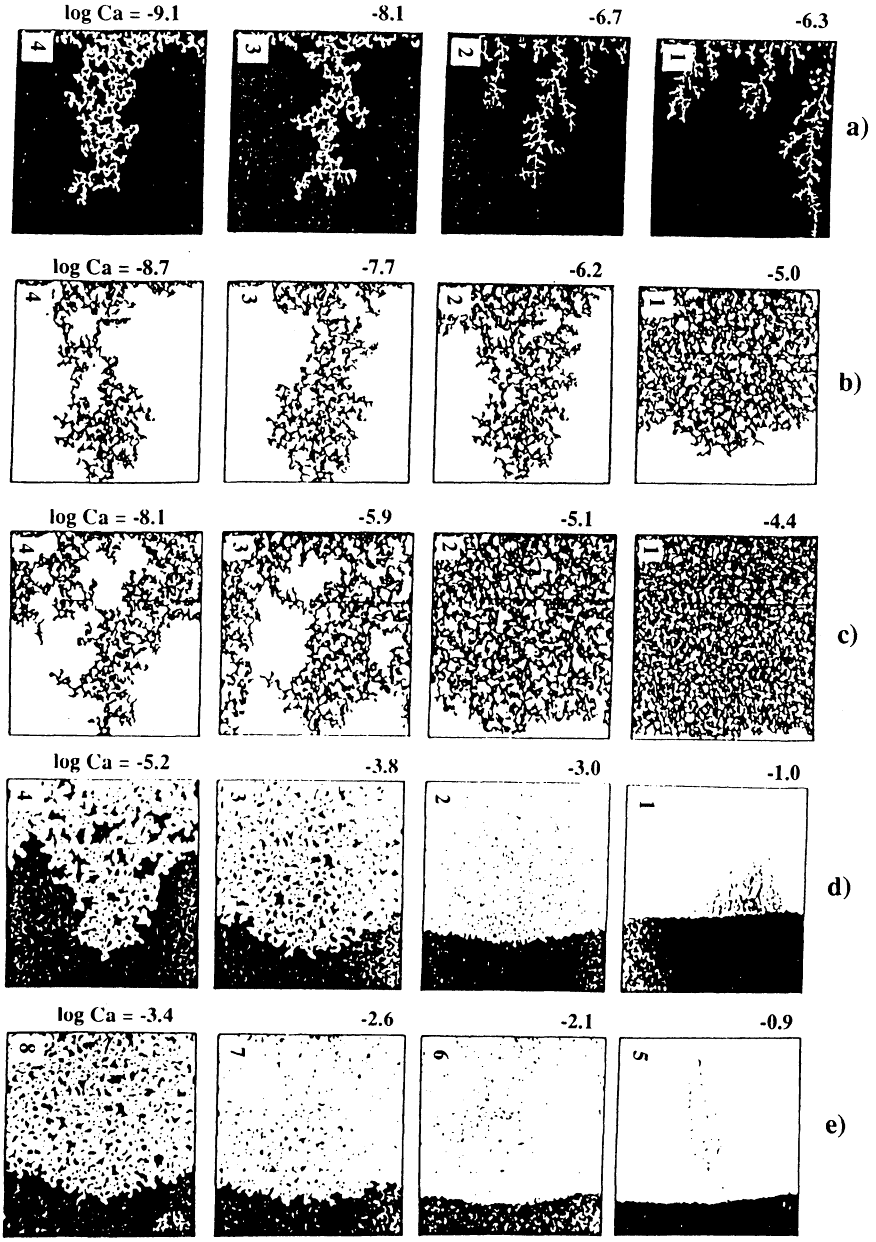
log Ca = -9.1
-8.1
-6.7
-6.3
4
3
2
1
a)
log Ca = -8.7
-7.7
-6.2
-5.0
4
3
2
1
b)
log Ca = -8.1
-5.9
-5.1
-4.4
4
3
2
1
c)
log Ca = -5.2
-3.8
-3.0
-1.0
4
3
2
1
d)
log Ca = -3.4
-2.6
-2.1
-0.9
8
7
6
5
e)

FIG. 13. The two-phase interface at various capillary numbers and viscosity ratios M. (a) The transition from viscous fingering (1) to capillary fingering (4) for an unstable Saffman–Taylor interface ($\log M = -4.7$); (b),(c) the interfaces are stable according to the Saffman–Taylor criterion ($\log M = 0.7$ and $\log M = 1.9$), but the low capillary numbers yield capillary fingers modeled by invasion percolation; (d),(e) the interfaces are even more stable according to the Saffman–Taylor criterion ($\log M = 2.0$ and $\log M = 2.9$), and the interfaces vary between apparently flat on the large scale (though perhaps having a self-affine structure at smaller scales) and incipient capillary fingering; redrawn with permission from A. Coniglio in *Hydrodynamics of Dispersel Media*, Eds. J. P. Hulin, A. M. Cazabat, E. Guyon, F. Carmona Amsterdam: North-Holland (1990).

phases are connected [and, therefore, both $X_1(z) > 0, X_2(z) > 0$]. But the displacing phase becomes ever more sparse as the displacing phase percolation transition at $z = z_{c2}$ approaches; and for $z > z_{c2}$, only the displaced phase is connected [in this region, $X_1(z) > 0$, while $X_2(z) = 0$], with the displacing phase forming a critical tail in the form of a rough interface with a fractal dimension $D \approx 2.5$ and a width $W(B)$ that depends on the Bond number $B = \Delta\rho\gamma\kappa/\sigma$, a measure of the ratio of hydrostatic to surface-tension forces on the fluid in a pore.

The creation and properties of this nonhomogeneous flow structure can be understood in terms of invasion percolation in a gravitational field or gradient percolation (Sapoval *et al.*, 1985; Gouyet *et al.*, 1990). The terminology gradient percolation comes from the influence of the gravitational field on the probability $P(z)$ for an active bond to exist at a height z. In normal invasion percolation, we have $P(z) = P$, a constant dependent only on the constant applied pressure. In a gravitational field, however, the effective hydrostatic pressure at a height z is $\Pi(z) = \rho g(z_m - z)$, where the constant z_m depends only on the applied pressure. Therefore, if $P^*(\Pi)d\Pi$ is the probability for the existence of capillary pressures in the porous medium between Π and $\Pi + d\Pi$, then the way in which the active bond probability $P(z)$ depends on z is

$$P(z) = \int_0^{\rho g(z_m - z)} P^*(\Pi)d\Pi. \tag{60}$$

If we assume that the throat sizes are equally distributed, then $P(z)$ decreases with a constant gradient $|dP/dz|$ as $P(z) = 1 - |dP/dz|z$.

Using this gradient dependence for $P(z)$ and assuming the validity of the usual percolation exponents, several estimates can now be made for the properties of the invasion front during drainage in a gravitational field. The two percolation transitions can be seen to occur at heights $z_{c1} = (1 - P_{c1})/|dP/dz|$ and $z_{c2} = (1 - P_{c2})/|dP/dz|$, while the width of the critical tail can be estimated as $W(B) \sim \xi \sim [P(z) - P_{c2}]^{-\nu}$. To find the Bond number dependence of the width, one estimates from scaling that $P(z) - P_{c2} \approx |dP/dz|(z - z_{c2}) \sim |dP/dz|W(B)$, and, therefore, one finds self-consistently that $W(B) \sim |dP/dz|^{-\nu/(1+\nu)}$. Finally, as the gradient is proportional to the ratio of the hydrostatic pressure $\Delta\rho g\kappa^{1/2}$ to the capillary pressure $\sigma/\kappa^{1/2}$ and, therefore, to the Bond number, $|dP/dz| \sim B = \Delta\rho g\kappa/\sigma$, one finds a singular form for the width of the interface (Wilkinson, 1984), which should diverge as the gravitational field tends to zero as

$$W(B) \sim \kappa^{1/2} B^{-\nu/(1+\nu)}. \tag{61}$$

The observed fractal dimension of the rough interface can also be predicted from percolation theory as the fractal dimension of percolation clusters $D = d - \beta/\nu$. In three dimensions, $D \approx 2.55$, in close agreement with experimental observation.

3.4.3 Macroscopic Langevin Equations

When the process of imbibition rather than drainage is considered, then the interface on a very large scale appears to be flat. Microscopically, this makes good sense. If the interface of a nonwetting fluid is considered, it will have great difficulty in flattening neighboring fingers, as this may involve flow through small pores, which is discouraged; for a wetting fluid, however, capillary forces will encourage such flow and, in so doing, relax incipient fingers (Lenormand and Zarcone, 1985). But, on closer examination, at smaller scales, the rough surfaces are not really flat at all, but to a good approximation can be described as single-valued fluctuations in height $h(\mathbf{r}, t)$ on a flat substrate; that is, overlaps are unimportant. Even more importantly, they appear to be self-affine. That is, they have a type of self-similarity under changes of scale, as do fractals, but, in contrast with fractals, they are not isotropic: If surfaces in windows of length L and length λL are compared, then the height fluctuations do not become statistically self-similar if an equal change of scale in the direction of the fluctuations $h \rightarrow \lambda h$ is made, as would be the case for fractals, but, rather, the change of scale $h \rightarrow \lambda^\alpha h$ must be made, where α, known as the Hurst exponent, is a property of the self-affine interface.

The growth and form of such interfaces have been the subject of extensive investigations in recent years and have resulted in the recognition (Family and Vicsek, 1985) that the width fluctuations $W(L, t) = \langle|h(\mathbf{r} + \mathbf{L}, t) - h(\mathbf{r}, t)|\rangle$ of interfaces grown on substrates of dimension L^{d-1} all obey a special form of self-affine dynamic scaling: Initially, the fluctuations increase with time,

$$W(L, t) \sim t^{\beta} \quad \text{if } t \ll t_L, \tag{62a}$$

and then saturate to a length-scale–dependent final value,

$$W(L, t) \sim L^{\alpha} \quad \text{if } t \gg t_L. \tag{62b}$$

The crossover time scale $t_L \sim L^z$ is also length-scale dependent, with an exponent $z = \alpha/\beta$. This last identity can be derived from the existence of a scaling form $W(L, t) = t^{\beta} f(t/L^z)$.

In addition to scaling, a great degree of universality was discovered in the actual values of the critical exponents α, β, and z seen in such apparently unrelated phenomena as the interface of two-phase flow during imbibition, surfaces grown by molecular beam epitaxy, and the structure of surfaces in growing bacterial colonies. This universality was confirmed by computer simulations of various microscopic models for surface roughening, which concluded that universality classes exist for apparently unrelated models creating self-affine interfaces, with the values within a given class dependent only on certain conservation laws and symmetries and on the spatial dimensionality d.

To account for this universality, the underlying kinematics of rough surface growth were incorporated into Langevin equations. These equations incorporate the notion that, under driving with a random force $\eta(\mathbf{r}, t)$, the height fluctuations $h(\mathbf{r}, t)$ in interfacial growth grow with a velocity λ normal to the interface and relax at a rate D because of surface diffusion or surface tension (surface tension would be the relaxation mechanism in imbibition, while an isomorphic equation with the surface tension replaced by surface diffusion describes molecular beam epitaxial growth of surfaces). Typical of such Langevin equations is the Kardar–Parisi–Zhang (KPZ) equation (1986) for the height fluctuations $h(\mathbf{r}, t)$ in an interface growing with a velocity λ normal to the interface:

$$\frac{\partial h}{\partial t} = D\nabla^2 h + \frac{\lambda}{2}(\nabla h)^2 + \eta(\mathbf{r}, t), \tag{63}$$

where $\langle \eta(\mathbf{r}, t)\eta(\mathbf{r}', t')\rangle = \Gamma\delta(\mathbf{r} - \mathbf{r}')\delta(t - t')$. The KPZ equation is one of a zoo of such equations, which differ depending on the symmetries and conservation laws that need to be incorporated. The Hwa–Kardar equation (43) describing height fluctuations in flowing sandpiles is another example. The exponents α and β are therefore not independent, and for the KPZ equation, the scaling relation $\alpha + \alpha/\beta = 2$ can be derived.

The actual values of the exponents require further analytical or numerical work, but, in $d = 2$, the dynamic renormalization group (Medina *et al.*, 1989) gives the exact result $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{3}$. There are no exact results in $d > 2$, but, on the basis of numerical evidence alone, Kim and Kosterlitz (1989) have suggested $\alpha = 2/(d + 2)$ and $\beta = 1/(d + 1)$.

How well do such equations agree with experimental data on imbibition? The results are mixed. Typical values for the exponents in two-dimensional experiments on imbibition (Horvath *et al.*, 1991) involving water displacing air in paper gives $\alpha \approx 0.75$, $\beta \approx 0.6$. Therefore, scaling laws such as $\alpha + \alpha/\beta = 2$ are obeyed to great accuracy, but the individual exponents are very far from theoretical predictions.

There could be many causes for this variance. The noise may not be δ correlated in time, or the porous medium itself can be considered to contribute a quenched random noise $\zeta(\mathbf{r}, h)$ obeying $\langle \zeta(\mathbf{r},h)\zeta(\mathbf{r}',h')\rangle = G\delta(\mathbf{r} - \mathbf{r}',)\,\delta(h - h')$ to the growth process. This noise could cause local pinning of the interface, leading when depinning occurs to sudden bursts of growth, or even the existence of a global pinning-depinning transition. This is an active area of present research, and a definite reason for the variance between theory and experiment remains, at present, unknown.

3.5 Hydrodynamic Dispersion in Porous Media

The discussion of flow of wetting fluids in porous media has focused so far on the structure of the self-affine interface during imbibition. Such fluids can, however, also carry dispersions of solute particles. Examples would include sediments or pollutants in groundwater. The manner in which such suspensions are dispersed by flow is also a matter of technological importance. Here, we consider the dispersion of such solutes dissolved in a liquid flowing in a porous medium and the magnitude of the diffusion coefficients involved.

3.5.1 Dispersion of Conserved Solute Particles Consider a pulse of passive scalar particles, that is, particles whose motion is controlled by the velocity field in the porous medium but which do not react back on the flow, released inside the fluid flowing in a porous medium. They will disperse under the combined action of convection and molecular diffusion with time. This dispersion will be anisotropic because of the existence of a macroscopic flow U, say in the x direction. In consequence, the convection-diffusion equation for the particle density $c(\mathbf{r}, t)$ can be written

$$\frac{\partial c}{\partial t} + U\frac{\partial c}{\partial x} = D_\parallel \frac{\partial^2 c}{\partial x^2} + D_\perp \nabla_\perp^2 c, \tag{64}$$

where, from scaling, one would expect the longitudinal and transverse effective diffusion constants to be of the same order of magnitude, $D_\perp \sim D_\parallel$. How do these effective diffusion constants $D_\perp = D_\perp(D, U, \kappa)$ and $D_\parallel = D_\parallel((D, U, \kappa)$ depend on the molecular diffusion constant D, the macroscopic flow velocity U, and the permeability κ? A great deal of information can be derived from scaling arguments.

First, as the flow velocity tends to zero, $U \to 0$, the dispersion must be isotropic, and, as there is no convection, molecular diffusion is the only source of dispersion:

$$D_\parallel \approx D_\perp \approx D. \tag{65}$$

In the opposite limit $U \to \infty$, the dispersion is convection dominated, and, from scaling, one expects

$$D_\perp \sim D_\parallel \sim U\kappa^{1/2}. \tag{66}$$

This result can be understood as follows: The streamlines in porous media are extremely complex—streamlines in different parts of the same pore or throat will separate into different pores and throats; velocity in large pores and throats is slow, while in small pores and throats, it is fast; the complex connectivity of the porous medium allows streamlines to separate and then rejoin later. The consequence is that mechanical mixing will dominate, and the contribution of molecular diffusion to dispersion can be neglected. To estimate the effective diffusion constant in the high-flow limit, one can argue that the relaxation time scale $\tau \sim a/U$ is the time necessary for a particle to cross a pore, and, therefore, $D_\perp \sim D_\parallel \sim \langle x^2 \rangle/\tau \sim a^2/(a/U) \sim U\kappa^{1/2}$.

This scaling argument depends on the existence of a single flow rate U in the porous medium for its validity. Experimentally, logarithmic corrections are often observed to scaling:

$$D_\perp \sim D_\parallel \sim U\kappa^{1/2} \log(U\kappa^{1/2}/D), \tag{67}$$

and these logarithmic deviations from scaling can be understood on the basis of the existence of a distribution of flow velocities in the pores. There does not exist a single transit time $\tau \sim a/U$, but, rather, a distribution with moments given by $\langle t^n \rangle \sim \int (a/v)^n P(v) dv$, where $P(v)dv$ is the probability of a velocity in the pore between v and $v + dv$. If $P(v) \sim v$ as $v \to 0$, then logarithmic corrections to the high–flow-rate effective dispersion may be expected.

At intermediate flow rates, an alternative scaling form is also possible. Taylor (1953) considered the problem of dispersion in a pipe and pointed out that because Poiseuille flow in a pipe has a parabolic profile in velocity (maximum at the center of the pipe, zero at the edges because of stick boundary conditions), molecular diffusion across streamlines will enhance dispersion above that for pure molecular diffusion alone. A similar argument should be valid for porous media: In a time $\tau \sim a^2/D$ that the particle diffuses across the pore as a result of molecular diffusion, the actual variance in position between particles near the edge of the pore and the pore center is $\langle x^2 \rangle \sim (U\tau)^2 \sim U^2a^4/D^2$. Therefore, dispersion in this regime is controlled by an effective Taylor diffusivity of size

$$D_\perp \sim D_\parallel \sim \langle x^2 \rangle/\tau \sim a^2U^2/D \sim U^2\kappa/D. \tag{68}$$

In conclusion, a scaling form for the effective longitudinal and transverse diffusivities may be written:

$$\begin{aligned} D_\parallel(D, U, \kappa) &= U\kappa^{1/2} f_\parallel(Pe), \\ D_\perp(D, U, \kappa) &= U\kappa^{1/2} f_\perp(Pe), \end{aligned} \tag{69}$$

where $Pe = U\kappa^{1/2}/D$ is the dimensionless Peclet number, which is the ratio of the convective time scale $\tau_{\text{convective}} \sim a/U$ to the diffusive time scale $\tau_{\text{molecular}} \sim a^2/D$ in the porous medium.

Thus far, the dispersion in porous media has been assumed to be Gaussian, albeit with an effective anisotropic diffusion constant depending on the flow. Actual experiments, such as injecting glycerol with colored dyes into two-dimensional porous media, suggest that while the average concentration of tracer particle does indeed obey a convective-diffusive equation, fluctuations exist about this average. Feder *et al.* (1990) studied the equiconcentration surfaces of the dispersed dye and found that the surfaces are not smooth lines, but self-affine, with $\alpha \approx 0.58$ in two dimensions.

3.5.2 Grain Trapping in Porous Media

The material forming the solute in Sec. 3.5.1 was implicitly assumed to be of microscopic dimensions, certainly much smaller than any pore size. But what happens to grains of material that start to approach the pore size or throat radius as they flow through the porous medium? Clearly, the possibility exists that such material will be trapped or pinned at a site. Perhaps later this same grain will be released as a result of Brownian noise, only to be retrapped somewhere else at a later time. The total effect of such trapping is to slow down greatly the average velocity of such grains from the average flow velocity U.

de Gennes (1979) gave an estimate for this slowing down. Suppose the grains have a radius R, while the pores have a distribution of sizes with a probability $P^*(a)da$ that their radius lies between a and $a + da$. Then the probability of being trapped in a pore is $P_{\mathrm{trap}} = \int_0^R P^*(a)da$. Therefore, the average distance traveled between traps is $\langle x \rangle \sim \kappa^{1/2}/P_{\mathrm{trap}}$, while the typical time between such trappings is $\tau = \langle x \rangle / U \sim \kappa^{1/2}/UP_{\mathrm{trap}}$. Using the high–flow-rate form for the effective diffusion constants $D_\perp \sim D_\parallel \sim U\kappa^{1/2}$, one can estimate the dispersion in these distances as $\langle y^2 \rangle \sim \langle x^2 \rangle - \langle x \rangle^2 \sim \kappa/P_{\mathrm{trap}}$.

Once it is trapped, the next question is how long such a grain is likely to remain trapped before it is released as a result of thermal fluctuations. This can be estimated from the Fokker–Planck equation $\partial P/\partial t + U\partial P/\partial x = D\partial^2 P/\partial x^2$ for the probability $P(x,t)$ that a grain travels a distance x in a time t against the flow U, where the diffusion constant $D \sim kT/6\pi\eta R$. The equation has an equilibrium solution $P(x) \sim \exp(-Ux/D)$, and, therefore, the probability of moving a pore distance $a \sim \kappa^{1/2}$ is $P(a) \sim \exp(-U\kappa^{1/2}/D)$. If the escape attempt rate is D/κ due to Brownian motion, then the release time can be estimated as $\tau_{\mathrm{release}} \sim \kappa/D \exp(U\kappa^{1/2}/D)$, and, therefore, the average grain velocity is reduced from its nominal value U due to trapping to

$$\begin{aligned} U_{\mathrm{grain}} &\sim U\tau/\tau_{\mathrm{release}} \\ &\sim (D/\kappa^{1/2}P_{\mathrm{trap}}) \exp(-U\kappa^{1/2}/D). \end{aligned} \tag{70}$$

The swift decrease in grain mobility with increasing Peclet number is very evident in expression (70).

GLOSSARY

Attractor: A term from dynamical systems theory to describe the region of phase space that a trajectory approaches asymptotically. To be an attractor, this region must be stable under small perturbations of the trajectory. The attractor itself could be a point or a limit cycle, or have some more complex topology, such as a fractal set.

Basin of Attraction: The basin of attraction for a given attractor is that region of phase space for which all trajectories beginning in this basin will asymptotically reach the attractor.

Brownian Motion: The noise due to thermal fluctuations. At the microscopic level, this noise manifests itself as random forces acting on the individual molecules with a variance proportional to the temperature, causing the molecules to perform a random walk. At the macroscopic level, the noise is manifested as diffusion.

Chaos: The state of a deterministic dynamical system whose time evolution shows extremely sensitive dependence on initial conditions—a consequence of which is that detailed prediction of the space and time development of the flow is impossible, and statistical methods have to be used to extract average properties of both spatial and temporal fluctuations.

Conformal Mapping: A mapping using an analytic function in the complex plane. Its importance lies in the fact that conformal mapping techniques can be used to solve Laplace's equation in the presence of complex boundary conditions by transforming the boundary into a simpler shape.

Constitutive Relation: The equations of motion derived for hydrodynamic densities in the form of the equations of continuity are all expressed in terms of stress tensors $\sigma_{\alpha\beta}$ and fluxes and, therefore, do not form a closed set of equations. To achieve this eminently desirable result, explicit expressions for these stresses and fluxes, called constitutive relations, are required. Typically, fluxes are assumed to be proportional to thermodynamic forces (Darcy's law gives velocity fields proportional to pressure gradients, Ohm's law gives electric currents proportional to electric-potential gradients, for example). For Newtonian fluids, the phenomenological assumption made is that the stress $\sigma_{\alpha\beta}$ is proportional to the rate-of-strain tensor $\partial_\alpha v_\beta + \partial_\beta v_\alpha$; the constant of proportionality is the viscosity.

Dynamical Systems: Systems described by deterministic simple differential equations, or mappings. Recent developments in the study of the time evolution of nonlinear dynamical systems have led to the realization that even very simple deterministic equations can lead to very complex trajectories that show extreme sensitivity to initial conditions (see **Chaos**).

Fokker–Planck Equation: Stochastic processes, that is, dynamical systems in the presence of noise, will lead to an ensemble of dynamical behaviors, one set of trajectories for each realization of the noise, and, consequently, unique trajectories have to be replaced by a probability distribution for dynamical variables with time. The equations of motion for these probability distributions are known as Fokker–Planck equations.

Fractals: A word coined by Mandelbrot to describe objects with a box-counting dimension D larger than their topological dimension. If a fractal of size L is covered with boxes of size l, and the number of such boxes required to cover the object scales as $N(l) \sim (L/l)^D$, then its fractal dimension is D. In general, such objects are exactly or stochastically scale invariant and often created by some recursive process. This means that there is no length scale associated with the object in question, and it appears self-similar under magnification.

Hamiltonian Dynamics: Dynamical laws describing a special subset of dynamical systems that conserve energy and phase-space volume. They come originally from a generalized form of Newton's law of motion and describe the motion of mechanical systems without friction. Because of phase-space conservation, there cannot be any attractors in a Hamiltonian flow. In general, a system with N degrees of freedom will be described by N coordinates q_α and N conjugate momenta p_α. A function known as a Hamiltonian $H(p_\alpha,q_\alpha)$ can be built from these coordinates and momenta, such that the dynamical system obeys $\partial p_\alpha/\partial t = -\partial H/\partial q_\alpha, \partial q_\alpha/\partial t = \partial H/\partial p_\alpha$.

Hysteresis: When a physical system does not show reversible behavior coming back to its original state as a control parameter is varied in a cyclic manner, the system is said to show hysteresis. Originally, the term was applied to the magnetic behavior of ferromagnets in external magnetic fields.

Langevin Equations: Equations of motion for dynamical systems to which random fluctuations and also dissipative frictional forces have been added. The fluctuating and dissipative terms are not usually independent, but intimately related via the fluctuation-dissipation theorem. Langevin equations are a different language in which to describe the same stochastic phenomena that Fokker–Planck equations do.

Linear Stability Analysis: The analysis of the manner in which infinitesimal perturbations applied to a system grow or decay. As small perturbations always exist, even a single growing mode implies instability. Linear stability analysis can only show that a system is stable to infinitesimal fluctuations; it says nothing about finite perturbations.

Phase Diagram: The behavior of physical systems changes as control variables such as temperature, magnetic field, temperature gradients, flow rates, etc. are altered. Typically, however, a given phase or spatiotemporal pattern changes smoothly and is stable for a region of values of these control parameters, followed by abrupt bifurcations to new phases or flow patterns. Phase diagrams are schematic representations of these various stable regimes and the lines of marginal stability between them.

Phase Space: An abstract space whose coordinates are all the variables needed to describe the temporal behavior of a dynamical system. For example, for a Hamiltonian system with N degrees of freedom, its dimensionality is $2N$. The instantaneous state of the system is then given by a single point in this phase space, and the motion of this point gives the trajectory of the system with time.

Phase Transition: The macroscopic phase of a system can be described by its symmetry and by the existence of a set of macroscopic variables that describe the degree of order (translational, rotational, magnetic, etc.) present in the system. As control parameters change, bifurcations between phases can occur, called phase transitions. These can either be discontinuous, resulting in the appearance of a nonzero new order parameter (a first-order phase transition), or continuous, in which case the new order parameter grows from zero as the system is brought through the transition (a second-order phase transition).

Pinning: In porous media, quenched random impurities can lock the interface at a given position and only release it when surface-tension forces become very large. In certain cases, this local pinning can become so strong and occur at so many points that the interface becomes globally pinned.

Polydisperse: Often used in the context of polymers and granular materials to describe polymer chains or grain sizes that are not all of the same magnitude but follow some distribution.

Power Spectrum: Fluctuating temporal signals can be thought of as consisting of a linear combination of modes fluctuating at different frequencies. The power spectrum $S(f)$ gives the power of fluctuations at each frequency f. Therefore, a single periodic mode would have a power spectrum consisting of a single δ function at that frequency, while the opposite extreme would be represented by white noise with a constant power spectrum.

Quenched Randomness: Any randomness that does not change with time, typically spatial randomness. The opposite is annealed randomness, which, over the time scale of the experiment, samples the complete distribution of random ensembles, as is the case for Brownian noise.

Renormalization Group: The group formed by the set of operators used to integrate out all features below a given length scale in, and therefore to coarse-grain, a physical system. From their influence with length scale on the parameters describing the new coarse-grained system, critical exponents can be found.

Scaling Arguments: A theoretical approach to understanding the spatial or temporal properties of a physical process by combining dimensional analysis with the relevant length or times scales involved. Often, exponents can be derived uniquely by such arguments, which depend for their validity on deep intuition about the physical processes involved.

Self-Affine Surfaces: A self-affine surface is self-similar under magnification, but, in contrast with fractals, this magnification must be anisotropic. Typically, a surface of height $h(\mathbf{x})$ is self-affine if $|h(\mathbf{X} + \mathbf{L}) - h(\mathbf{X})| \sim L^{\alpha}$, and, therefore, a change in magnification by a factor λ in the plane must be supplemented by a factor λ^{α} in the height direction for the surface to remain self-similar. Surfaces which are stochastically self-affine often characterize interfacial growth.

Stress Tensor: Normally written $\sigma_{\alpha\beta}$, it is the force per unit area in direction α applied by the rest of the fluid on a material particle surface whose normal is in direction β (thus, the normal stress is the negative of the pressure exerted by the fluid particle $\sigma_{\alpha\alpha} = -p$, while $\sigma_{\alpha\beta}$ with $\alpha \neq \beta$ is the tangential stress applied to the material particle surface). For a simple Newtonian fluid, the stress tensor must be symmetric, $\sigma_{\beta\alpha} = \sigma_{\alpha\beta}$.

Tensors: Mathematical objects that transform under groups of symmetry operations in the same manner as products of vectors.

Universality: In phase transitions and critical phenomena, the idea that certain quantities—critical exponents, amplitude ratios—are independent of the detailed microscopic interactions present, but depend only on very general properties, such as symmetry of the interactions, and dimensionality.

Works Cited

Adler, P. M., Brenner, H. (1988), *Annu. Rev. Fluid Mech.* **20**, 35–59.

Bak, P., Tang, C., Weisenfeld, K. (1987), *Phys. Rev. Lett.* **59**, 381–384.

Batchelor, G. K. (1967), *An Introduction to Fluid Dynamics*, Cambridge, UK: Cambridge University Press.

Bensimon, D. (1986), *Phys. Rev. A* **33**, 1302–1308.

Bensimon, D., Kadanoff, L. P., Liang, S., Schraiman, B. I., Tang, C. (1986), *Rev. Mod. Phys.* **58**, 977–999.

Brady, R. M., Ball, R. C. (1984), *Nature* **309**, 225–229.

Brochard, F., de Gennes, P. G. (1983), *J. Phys. Lett. (Paris)* **44**, L785–L791.

Brochard, F., de Gennes, P. G. (1984), *J. Phys. Lett. (Paris)* **45**, L597–L602.

Campbell, C. S. (1990), *Annu. Rev. Fluid Mech.* **22**, 57–92.

Chandler, R., Koplik, J., Lenormand, R., Willemsen, J. F. (1982), *J. Fluid Mech.* **119**, 249.

Chen, J. D., Wilkinson, D. (1985), *Phys. Rev. Lett.* **55**, 1892–1895.

Chuoke, R. L., van Meurs, P., van der Poel, C. (1959), *Pet. Trans. AIME* **216**, 188.

Clement, E., Baudet, C., Hulin, J. P. (1985), *J. Phys. Lett. (Paris)* **46**, L1163–L1171.

DeGregoria, A. J., Schwartz, L. W. (1986), *J. Fluid Mech.* **164**, 383.

Edwards, S. F., Wilkinson, D. R. (1982), *Proc. Roy. Soc. London, Ser. A* **381**, 17–31.

Einstein, A. (1906), *Ann. Phys. (Leipzig)* **19**, 289.

Family, F., Vicsek, T. (1985), *J. Phys. A* **18**, L75–L81.

Family, F., Masters, B. R., Platt, D. E. (1989), *Physica D* **38**, 98–103.

Feder, J., Maloy, K. J., Jossang, T. (1990), "Experiments on Diphasic Flows: Front Dynamics in Model Porous Media," in: J. P. Hulin, A. M. Cazabat, E. Guyon, F. Carmona (Eds.), *Hydrodynamics of Dispersed Media*, Amsterdam: North-Holland, pp. 231–248.

Flory, P. J. (1969), *Statistics of Chain Molecules*, New York: Interscience Publishers.

Flory, P. J. (1971), *Principles of Polymer Chemistry*, Ithaca, NY: Cornell University Press.

de Gennes, P. G. (1974), *The Physics of Liquid Crystals*, Oxford, UK: Oxford University Press.

de Gennes, P. G. (1977). *Makromolecules* **9**, 587.

de Gennes, P. G., Guyon, E. (1978), *J. Mech.* **17**, 403.

de Gennes, P. G. (1979), *C.R. Acad. Sci. (Paris)* **289**, 329–331.

de Gennes, P. G. (1984), *Scaling Concepts in Polymer Physics*, Ithaca, NY: Cornell University Press.

de Gennes, P. G. (1985), *Rev. Mod. Phys.* **57**, 827–863.

Ginder, J. M., Elie, L. D. (1992), in R. Tao (Ed.), *Proc. Conf. Electrorheological Fluids*, Singapore: World Scientific, pp. 23–36.

Glazier, J. A. (1989). *Dynamics of Cellular Patterns*, Ph.D. Thesis, The University of Chicago.

Gouyet, J. F., Rosso, M., Clement, E., Baudet, C., Hulin, J. P. (1990), "Invasion Percolation in Model Porous Media under Gravity," in: J. P. Hulin, A. M. Cazabat, E. Guyon, F. Carmona (Eds.), *Hydrodynamics of Dispersed Media*, Amsterdam: North-Holland, pp. 179–192.

Halsey, T. C. (1992), *Science* **258**, 761–766.

Horvath, V. K., Family, F., Vicsek, T. (1991), *J. Phys. A* **24**, L25–L29.

Hwa, T., Kardar, M. (1989), *Phys. Rev. Lett.* **62**, 1813–1816.

Joanny, J. F., de Gennes, P. G. (1984), *C. R. Acad. Sci. (Paris)* **299**, 279.

Joanny, J. F. (1990), "Spreading Kinetics," in: J. P. Hulin, A. M. Cazabat, E. Guyon, F. Carmona (Eds.), *Hydrodynamics of Dispersed Media*, Amsterdam: North-Holland, pp. 17–27.

Kadanoff, L. P., Nagel, S. R., Wu, L., Zhou, S. (1989), *Phys. Rev. A* **39**, 6524–6537.

Kardar, M., Parisi, G., Zhang, Y. (1986), *Phys. Rev. Lett.* **56**, 889–892.

Kim, J. M., Kosterlitz, J. M. (1989), *Phys. Rev. Lett.* **62**, 2289–2292.

Kolmogorov, A. N. (1954), *Dok. Akad. Nauk SSSR* [*Sov. Phys. Dokl.*] **98**, 527–532.

Kraynik, A. M. (1988), *Annu. Rev. Fluid. Mech.* **20**, 325–357.

de Kruif, C. G. (1990), "The Rheology of Colloidal Dispersions in Relation to Relation to their Microstructure," in: J. P. Hulin, A. M. Hulin, A. M. Cazabat, E. Guyon, F. Carmona (Eds.), *Hydrodynamics, of Dispersed Media*, Amsterdam: North-Holland, pp. 79–101.

Lamb, H. (1879), *Hydrodynamics*; republished (1945), New York: Dover.

Lenormand, R., Touboul, E., Zarcone, C. (1988), *J. Fluid Mech.* **189**, 165.

Lenormand, R., Zarcone, C. (1985), *Phys. Rev. Lett.* **54**, 2226–2229.

Lighthill, J. (1978), *Waves in Fluids*, Cambridge, UK: Cambridge University Press.

Maher, J. (1985), *Phys. Rev. Lett.* **54**, 1498–1501.

Marmur, A., Lelah, M. D. (1981), *Chem. Eng. Commun.* 13, 133.

Matshushita, M., Sano, M., Hayakawa, Y., Honjo, H., Sawada, Y. (1984), *Phys. Rev. Lett.* **53**, 286–289.

Medina, E., Hwa, T., Kardar, M., Zhang, Y. (1989), *Phys. Rev. A* **39**, 3053–3075.

Moser, J. (1973), *Stable and Random Motions in Dynamical Systems*, Princeton, NJ: Princeton University Press.

Niemeyer, L., Pietronero, L., Wiesmann, H. J. (1984), *Phys. Rev. Lett.* **52**, 1033–1036.

Nittmann, J., Daccord, G., Stanley, H. E. (1985), *Nature* **314**, 141–144.

Ogawa, S. (1978), in: *Proc. U.S.-Japan Seminar Contin.-Mech. and Stat. App. Mech. Granular Mater.*, Tokyo: Gukujutsu Bunken Fukyukai, pp. 208–217.

Ottino, J. M. (1990), *Annu. Rev. Fluid. Mech.* **22**, 207–253.

Paterson, L. (1984), *Phys. Rev. Lett.* **52**, 1621–1624.

Pittman, E. D. (1983), in: D. L. Johnson, P. N. Sen (Eds.), *Physics and Chemistry of Porous Media*, New York: AIP, pp. 1–19.

Pitts, E. (1980), *J. Fluid Mech.* **97**, 53.

Saffman, P. G., Taylor, G. I. (1958), *Proc. Roy. Soc. London, Ser. A* **245**, 312–329.

Sapoval, B., Rosso, M., Gouyet, J. F. (1985), *J. Phys. Lett. (Paris)* **46**, L149–L156.

Stauffer, D. (1985), *Introduction to Percolation Theory*, London: Taylor & Francis.

Tabeling, P., Libchaber, A. (1986), *Phys. Rev. A* **33**, 794–796.

Taylor, G. I. (1932), *Proc. R. Soc. London, Ser. A* **138**, 41–48.

Taylor, G. I. (1950), *Proc. R. Soc. London, Ser. A* **201**, 192–196.

Taylor, G. I. (1953), *Proc. R. Soc. London, Ser. A* **219**, 186–203.

von Neumann, J. (1952), "Discussion," in: *Metal Interfaces*, Cleveland: American Soc. Metals, pp. 108–110.

Wilkinson, D., Willemsen, J. F. (1983), *J. Phys. A* **16**, 3365–3376.

Wilkinson, D. (1984), *Phys. Rev. A* **30**, 520–531.

Witten, T. A., Jr., Sander, L. M. (1981), *Phys. Rev. Lett.* **47**, 1400–1403.

Further Reading

The Works Cited represent but a minute fraction of the available literature on nonhomogeneous flows and were chosen with three purposes in mind: as a historical paper representing the beginning of a new approach or area in nonhomogeneous flows; because the paper contains a specific piece of information deemed important for the particular areas mentioned in this article; or because the work, usually a book, represents a significant source of information and can be used either as a starting point for further reading or as a database of references.

In this last instance, however, it should be kept mind that research in nonhomogeneous flows is a very rapidly developing subject, and the latest references will only be found in the latest volumes of such journals as *Physical Review A* and *E*, *Physical Review Letters*, and *Journal of Fluid Mechanics*. Review articles on various aspects of nonhomogeneous flows may be found in *Reviews of Modern Physics*, *Annual Reviews of Fluid Mechanics*, and Conference proceedings.

Several books and articles are particularly appropriate to special areas of this article. Batchelor (1967) is a good general reference on fluid mechanics. Pelce, P. (Ed.) (1990), *Dynamics of Curved Fronts*, Perspectives in Physics, Boston: Academic Press, is a good collection of classic papers in interfacial evolution. Lighthill (1978) is probably the best all-round modern text on waves in fluids. Ottino (1990) is a good reference on mixing, and its foundation in Hamiltonian mechanics can be found in Arnol'd, V. I. (1978), *Mathematical Methods of Classical Mechanics*, Berlin: Springer-Verlag, and Moser (1973). de Gennes is a useful reference for many areas mentioned in this article, including liquid crystals (1974), and the dynamics of wetting (1985). Finally, as an example of the application of scaling methods to modern statistical mechanics with which fluid mechanics is becoming ever more tightly bound, it is impossible to find a superior text to de Gennes (1984), where polymer configurations and flows are considered. The classic text on flow through porous media is Bear, J. (1972), *Dynamics of Fluids in Porous Media*, New York: American Elsevier.

NONLINEAR ACOUSTICS

See ACOUSTICS, NONLINEAR

NONLINEAR DYNAMICS

See CHAOTIC PHENOMENA

NONLINEAR OPTICS

See OPTICS, NONLINEAR

NUCLEAR MAGNETIC RESONANCE

See MAGNETIC RESONANCE, NUCLEAR

NUCLEAR MAGNETIC RESONANCE, DETERMINATION OF STRUCTURES BY

See STRUCTURE DETERMINATION BY NUCLEAR MAGNETIC RESONANCE

NONLINEAR SYSTEMS

HENRY D. I. ABARBANEL,* *Department of Physics and Marine Physical Laboratory, Scripps Institution of Oceanography, University of California–San Diego, La Jolla, California, U.S.A.*

Introduction 417
1. **Nonlinear Systems in Physical Problems** 418
1.1 Thermal Convection 418
1.2 Nonlinear Circuits 419
1.3 Planetary Motions 420
1.4 Other Chaotic Systems 422
2. **Hamiltonian Systems and Dissipative Systems** 422
2.1 Phase-Space Volume 422
2.2 Attractors 423
2.3 Regular and Periodic Behavior 423
2.4 Strange Attractors and Nonperiodic Behavior 424
2.5 Unstable Periodic Orbits 424
3. **Maps and Flows** 424
3.1 Stability of Maps or Flows .. 425
3.2 One-Dimensional Maps: Universality 426
4. **Dimensions of Strange Attractors** 427
4.1 Box Counting 427
4.2 Generalized Dimensions; D_q 427
4.3 Statistics on Strange Attractors 429
5. **Lyapunov Exponents** 429
5.1 Oseledec's Multiplicative Ergodic Theorem 429
5.2 Positive Exponents and Strange Attractors 430
5.3 Global and Local Exponents 430
6. **Information Theory and Strange Attractors** 430
7. **Analysis of Experimental Data** 431
7.1 Time-Delay Embedding 431
7.2 Average Mutual Information 432
7.3 Global Embedding Spaces ... 432
7.3.1 True Vector Fields; False Nearest Neighbors 433
7.3.2 Local Dimensions 434
7.4 "Noise" versus Chaos 434
7.5 Dimensions from Data 435
7.6 Lyapunov Exponents from Data 435
7.7 Using Invariants of the Dynamics 436
7.7.1 Classification and Identification 436
7.7.2 Models and Observations 436
7.8 "Control" of Chaos 437
Glossary 437
Works Cited 438
Further Reading 439

INTRODUCTION

The study of the nonlinear behavior of physical systems dates back in some sense to Newton (1687), where the planetary equations of motion under $1/r^2$ gravitational forces were the first example of nonlinear interaction among the degrees of freedom

$$\mathbf{x}_j(t), \frac{d\mathbf{x}_j(t)}{dt}, \tag{1}$$

for the $j = 1, 2, \ldots, N$ planetary bodies. The remarkable fact that the case of two bodies ($N = 2$) with $1/r^2$ forces is exactly integrable gave rise to a worldview that for over two centuries limited inquiry into the fascinating aspects of the nonlinear properties of these equations when $N \geq 3$. The difficulty of the problem was known to many of the great nineteenth-century mechanicists and was chronicled by Whittaker (1937). The first recognition of the insolubility of the problem on Newtonian terms was perhaps that of Poincaré (1892). His work stands clearly as the beginning of the study of the effects of nonlinearity in a qualitative fashion, and the slow recognition that we are unlikely to repeat the

---

*Institute for Nonlinear Science.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

wonderful accident of Newton's discovery of integrable dynamics and must instead turn to statistical and qualitative versions of our description of physical systems.

Since the work of Lorenz (1963), it has become clear that dissipative systems are equally as rich as the pure, friction-free equations of Newton and Hamilton. Since dissipation appears in all applications in physics, the properties of these systems are worthy of study on their own.

The physical world is replete with nonlinear systems, and, indeed, the view that only nonlinear systems are of interest to physicists is easy to defend. This article cannot cover all of physics, of course, and so the scope of the article is limited to examples from a few areas of physics and to the simplest examples of nonlinear behavior.

Our plan is to describe in physical terms the origin of the critical unexpected phenomenon appearing in nonlinear physical systems: *nonperiodic evolution of the orbits* in any system with three or more degrees of freedom. This is absolutely the most striking distinction between nonlinear behavior and the familiar linear behavior, which is composed of superpositions of sinusoidal modes or resonances of the source of observations. The origin of this nonperiodic behavior is *instability* in the physical processes arising because there are alternative, more efficient ways to transfer energy through the system than linear conduction or diffusion. The way in which these instabilities are tamed and used by physics is the most amazing part of the tale, and we shall attempt to convey that while presenting some of the tools that have been developed to analyze signals composed of bounded and folded instabilities.

This motion has been called *chaos*, which would seem to imply it is no different from "noise" and unpredictable. These motions are completely deterministic, however, and are slightly predictable. Indeed, chaos is evolution of a physical system that is intermediate between the completely irregular signals we characterize as noise and the completely regular signals we see in superpositions of sine waves. Chaos has a broad Fourier power spectrum and is nonperiodic, just as noise would be, but it is structured in the space of its own state variables, and noise is not (Abarbanel *et al.*, 1993). This structure in state space and the limited predictability of chaos are the keys to utilizing chaotic physical systems in applications.

1. NONLINEAR SYSTEMS IN PHYSICAL PROBLEMS

1.1 Thermal Convection

A paradigm for nonlinear systems has emerged in the Lorenz model, which is based on the partial differential equations for convection in the lower atmosphere when it is considered as a fluid heated from below. One starts with the Navier–Stokes equations for fluid velocity, coupled with the heat equation for the temperature, and makes a three-mode approximation to the motion. The rescaled mode amplitudes satisfy

$$\begin{aligned}\frac{dx(t)}{dt} &= \sigma[y(t) - x(t)],\\ \frac{dy(t)}{dt} &= rx(t) - x(t)y(t) - y(t),\\ \frac{dz(t)}{dt} &= x(t)z(t) - bz(t),\end{aligned} \tag{2}$$

where

$$\sigma = \nu/\kappa$$

is the Prandtl number,

$$r = q\alpha\Delta Th^3/\kappa\nu \tag{3}$$

is a reduced Rayleigh number, and b is a horizontal length scale.

The solutions to this equation are stable and decay to the point $(x, y, z) = (0, 0, 0)$ when $r < 1$. They are stable and decay to one of the points $(\pm\sqrt{b(r-1)}, \pm\sqrt{b(r-1)}, r-1)$ when $1 < r < \sigma(\sigma + b + 3)/(\sigma - b - 1)$; and when r exceeds this value, the orbits of the system become *nonperiodic*. The Fourier power spectrum of an orbit becomes broad and continuous; there are no peaks indicating strong periodic behavior, and the spectrum falls off rapidly with Fourier frequency. For very large r, the orbit again becomes periodic.

The solution to the Lorenz equations with $\sigma = 10$, $r = 45.92$, and $b = 6$ has been widely studied as an example of chaotic motions. In

Fig. 1, we display the time trace of $x(t)$ resulting from the Lorenz equations integrated with sampling time $\tau_s = 0.01$ in dimensionless units. The data displayed are $x(n) = x(t_0 + n\tau_s)$. In Fig. 2, we have the Fourier spectrum of this time series showing its broadband, continuous nature. In Fig. 3 is the phase-space portrait for the Lorenz orbits in $(x(t), y(t), z(t))$ space. The irregularity of the orbits for $x(t)$ results from the projection of the geometric figure here onto the $x(t)$ axis.

The Lorenz equations also arise in several other settings. In the case of a liquid confined to a torus, with the lower part of the torus heated to $T_0 + \Delta T$ and the upper half to T_0, the Lorenz equations with $b = 1$ are a nearly exact representation of the motion. For a laser near the lasing threshold, one again finds the Lorenz equations with the variables representing light intensity, population inversion, and medium polarization. The widespread appearance of the equations and their simple form and rich solution structure have given impetus to their detailed study over the past three decades.

1.2 Nonlinear Circuits

The appearance of chaotic motions that are nonperiodic has been known in nonlinear circuits for over 50 years (VanderPol, 1922). The dynamical origin of the ensuing "noise" was not recognized for some time. Indeed, the main effort for decades in work with electronics and circuits has been to banish this "noise" and emphasize the linearity of circuit response. Since all circuit elements are nonlinear at some level, the opportunity to take advantage of the chaotic behavior of circuits in applications has only emerged with the knowledge we now have of how to characterize the chaos and how to model it without detailed circuit equations, especially for the critical nonlinear elements. An example of the observed signal from a circuit with a hysteretic nonlinearity is shown in Fig. 4. It is broadband spectrally. Using a technique to be introduced later, its portrait in phase space is shown in Fig. 5. Here we see the state-space structure, even though the time trace is complicated and the Fourier spectrum is broad.

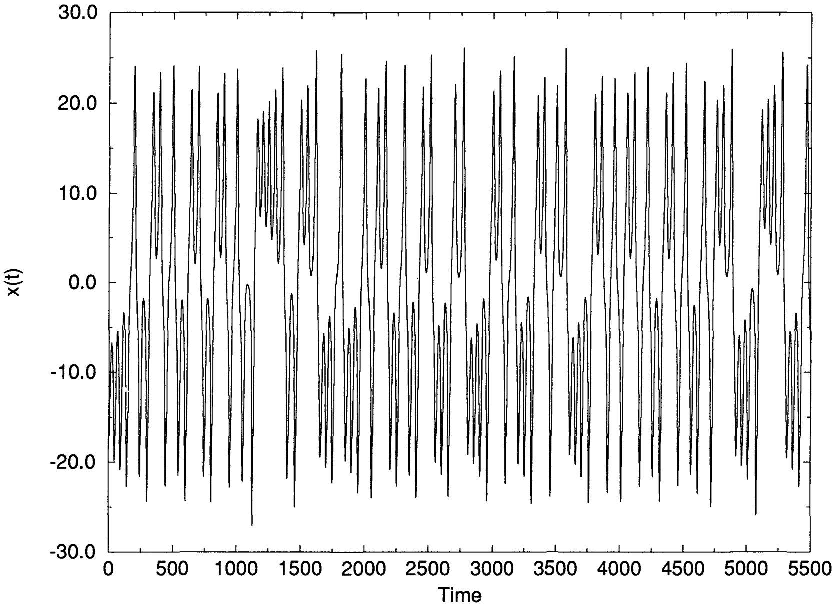

FIG. 1. The time series for one of the components of the Lorenz equations (2). The equations were solved with a Runge–Kutta integrator using a time step of 0.01. $\sigma = 16$, $b = 4$, $r = 45.92$. The irregularity in time is one of the characteristics of chaos.

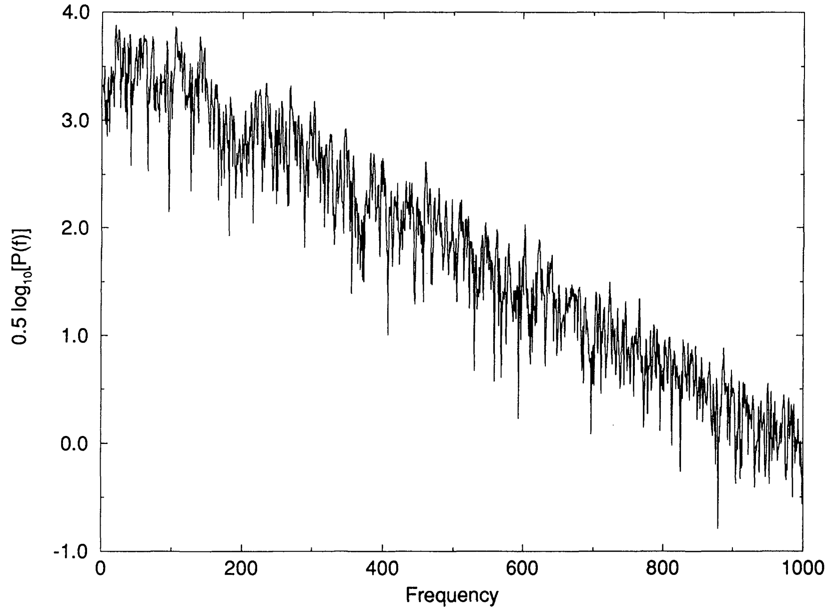

FIG. 2. The Fourier power spectrum of the time series in Fig. 1. The power spectrum is broad and continuous. No sharp peaks indicating periodicity are seen.

The possibility of applications flowing from chaotic operation of such circuits is based on the geometry seen in this figure. "Noise" would be isotropic and unstructured in phase space—here we see structure that can be exploited.

1.3 Planetary Motions

The motions of the bodies in our solar system are, in detail, far from the simple periodic motions characteristic of Keplerian orbits. The planetary motions are well described without dissipation, so that the basic equations are those of Hamiltonian mechanics. The motions become chaotic through sequences of instabilities associated with the resonances in mechanical systems. In the 1950s and 1960s, through the work of Kolmogorov, Arnol'd, and Moser (Arnol'd and Avez, 1968), it became clear that associated with these resonances are orbits that find their way throughout phase space rather than being

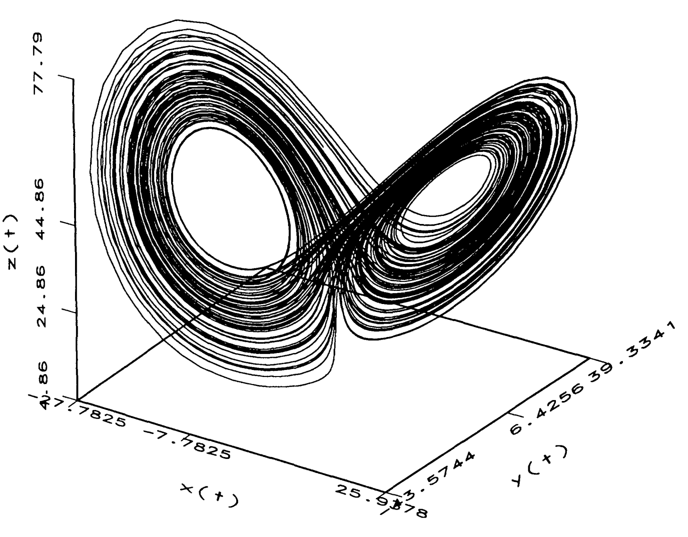

FIG. 3. The Lorenz attractor as seen in three dimensions using as coordinates the solutions ($x(t)$, $y(t)$, $z(t)$) of the Lorenz equations.

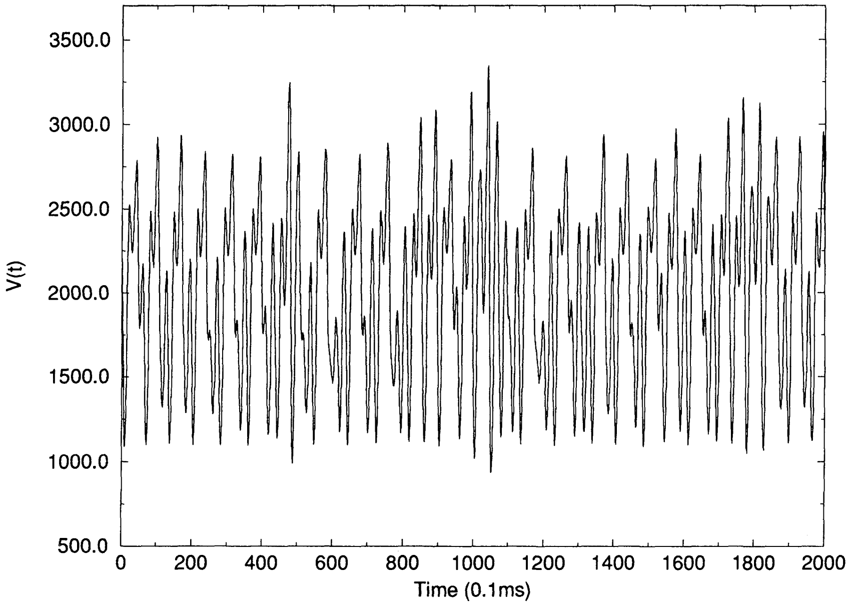

FIG. 4. The time series of a voltage from a nonlinear hysteretic circuit. The sampling time $\tau_s = 0.1$ ms in this experiment. The irregularity of the time series is characteristic of chaotic behavior.

confined to small subportions of phase space, as they would if the motion were regular and integrable. Indeed, in almost every problem of planetary motions in our solar system (Wisdom, 1987), there should be some regions of phase space where chaotic motions are seen. This leads to substantial deviations from regular behavior, and it points out the underlying reason why efforts to solve the three-body problem or to apply classical perturbation theory to many-body dynamics often led to disappointing results or outright failure. Basically, the motions that were described by that perturbation theory were so far from the assumed orbit, to which small corrections were to be found, that power-series expansions had no chance of succeeding. Pioncaré showed that the standard power-series approximations would at best be asymptotic and thus eventually fail, regardless of

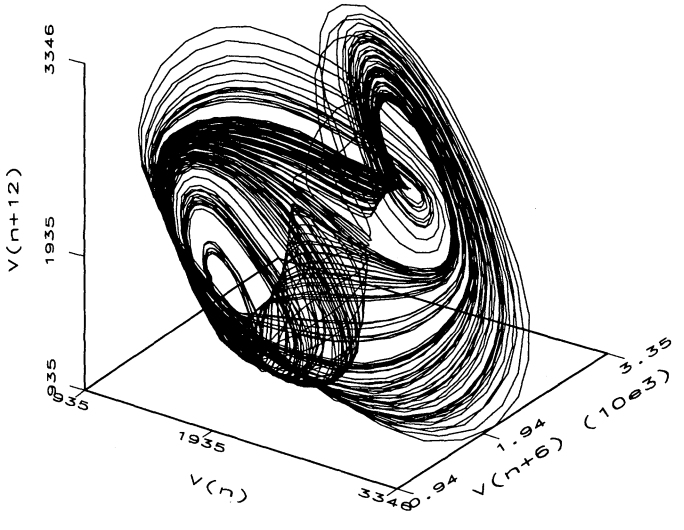

FIG. 5. Using time-delayed values of the voltage in Fig. 4, ($V(t)$, $V(t + T)$, $V(t + 2T)$), as coordinates we display the attractor associated with the nonlinear hysteretic circuit. $T = 6\tau_s$ and is chosen by the use of average mutual information. See Fig. 8.

how accurate they might be in the short run. Wisdom (1987) also points out that the time scale for chaotic phenomena to exhibit themselves can often be much longer than tens of millennia and might easily be missed if one were not aware either in simulations or other calculations of their possibility.

1.4 Other Chaotic Systems

Chaos appears in many other physical systems: in plasma physics, in the motion of particles in accelerators, in the description of earthquake motions, and so forth. By restricting ourselves to the examples above, we are trying to avoid a simple listing of almost all of physics in order to allow us to illustrate how one would think about specific issues. Chaos also appears in biological systems and chemical systems, many of which are of substantial interest. We shall have little to say about these systems, as the details of their structure would take us too far afield. It is important to note that the hallmark of the presence of chaotic behavior is that, when a physical system is out of equilibrium and is transporting energy or momentum through it from some source, such as heat in the case of convection, to some sink, then it now appears that as the ratio of energy transported to energy dissipated increases, chaos is the mode preferred to transport these quantities most effectively. Transport through many harmonic modes is unstable to chaos, which is nonperiodic motion.

2. HAMILTONIAN SYSTEMS AND DISSIPATIVE SYSTEMS

Physical systems divide rather naturally into those that are closed and are described by the conservative dynamics of Hamilton and open systems where energy flows in at some scales and flows out at others. The Lorenz system, which has its origin in viscous fluid flows, is of the second kind, while planetary motions and motions associated with large-scale atmospheric and oceanic dynamics are often nearly Hamiltonian. The key distinction from the point of view of analyzing the nonlinear dynamics is that, as they evolve, conservative systems preserve volume in phase or state space, while dissipative systems do not. This apparently small distinction gives rise to a world of interesting phenomena unanticipated by Newtonian mechanics.

2.1 Phase-Space Volume

Conservative systems can be described by Hamilton's equations. In most physical problems, we can find canonical coordinates $\mathbf{q}(t) = (q_1(t), \ldots, q_N(t))$ and $\mathbf{p}(t)$ that, given a scalar function $H(\mathbf{q}, \mathbf{p})$, the Hamiltonian, evolve as

$$\frac{d\mathbf{q}(t)}{dt} = \frac{\partial H(\mathbf{q}, \mathbf{p})}{\partial \mathbf{p}}, \qquad \frac{d\mathbf{p}(t)}{dt} = -\frac{\partial H(\mathbf{q}, \mathbf{p})}{\partial \mathbf{q}}. \tag{4}$$

In these canonical coordinates, the volume of phase space is preserved. This follows from the general result that, if a dynamical system with variables $\mathbf{x}(t) = (x_1(t), \ldots, x_M(t))$ satisfies

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{F}(\mathbf{x}(t)), \tag{5}$$

then a volume $V(t)$ evolves as

$$\frac{dV(t)}{dt} = \int_{V(0)} d^M x \nabla_{\mathbf{x}} \cdot \mathbf{F}(\mathbf{x}(t)). \tag{6}$$

For Hamiltonian mechanics,

$$\mathbf{F}(\mathbf{q}, \mathbf{p}) = \left[\frac{\partial H}{\partial \mathbf{p}}, -\frac{\partial H}{\partial \mathbf{q}}\right] \tag{7}$$

and

$$\nabla_{\mathbf{x}} \cdot \mathbf{F}(\mathbf{x}) = 0. \tag{8}$$

For the Lorenz system given before, $\nabla_{\mathbf{x}} \cdot \mathbf{F}(x, y, z) = -(\sigma + b + 1) < 0$, and so volumes in phase space for that dynamics obey $V(t) = V(t = 0) \exp[-(\sigma + b + 1)t]$, which goes to zero rapidly as time unfolds. Volumes in dissipative system shrink to zero with time, so that any collection of initial conditions will evolve to a set with dimension less than that of the state space. In a Hamiltonian system, the volume is preserved in canonical coordinates, so that any set of initial conditions in dimension $2N$ will remain of dimension $2N$, however twisted the individual orbits become.

This would be a fact of no special importance, except that Hamiltonian systems are rare in physics. Indeed, they are not generic, since even the slightest dissipation of energy from the spatial and temporal scales describing some phenomenon leads to the decrease in phase-space volume associated with those scales. In viscous fluid flow, macroscopic motions dissipate their energy into microscopic motions, resulting in no motion on macroscopic scales unless energy from outside the system is available. In the Lorenz system, energy is input from the temperature difference and lost to viscosity and thermal dissipation. The balance between these effects has remarkable consequences.

When phase-space volumes shrink, the points on any orbit evolve to a set with dimension less than that of the overall phase space in which the equations lie. For example, suppose the phase space is one dimensional, with a dynamical variable $x(t)$ satisfying

$$\frac{dx(t)}{dt} = -\nu x(t), \quad \nu > 0. \tag{9}$$

Then the solution $x(t) = x(0)e_{-\nu t}$ rapidly goes to the zero-dimensional limit $x = 0$, regardless of $x(0)$. Here $\nabla_{\mathbf{x}} \cdot \mathbf{F}(\mathbf{x}) = -\nu$. The dimension of the solution after a long time is less than the dimension of the space in which the solution wanders.

2.2 Attractors

The shrinking of phase-space volume leads to the notion of an *attractor,* which is the set of points in the phase space to which solutions go after transients have died out. The part of phase space from which all points go to the attractor is called the *basin of attraction*. In the simple example $\dot{x}(t) = -\nu x(t)$, the attractor is the origin, and the basin of attraction is the whole real line. A system can have many attractors with complex boundaries between their basins of attraction. The attractors can be as simple as a point, which has dimension zero, or as interesting as others we now discuss. A Hamiltonian system has no attractor, but phase-space points wander all over the surface of constant $H(\mathbf{q}, \mathbf{p})$, since $H(\mathbf{q}(t), \mathbf{p}(t)) = 0$. In fact, Gibbsian statistical mechanics starts with the hypothesis that this surface is uniformly populated and with this assumption introduces the microcanonical ensemble. Dissipative systems do not follow this notion.

Physically, a dissipative system, when driven weakly, will move toward a state of no motion. So the attractor would just be a point, which has dimension zero, regardless of the dimension of the overall space. When the driving or energy input is increased, instabilities arise, and the geometry of the motion changes and becomes richer. Among the first things to happen is that a periodic orbit may emerge from the instability.

2.3 Regular and Periodic Behavior

Evolution to a point attractor can become unstable when the driving forces are increased. This is seen in the simple example of motion in a plane with polar coordinates (r, θ):

$$\begin{aligned} \frac{dr(t)}{dt} &= ar(t) - r^3(t), \\ \frac{d\theta(t)}{dt} &= 1. \end{aligned} \tag{10}$$

For $a < 0$, all points in the plane move to the origin, which is the attractor of dimension zero. For $a > 0$, all points in the plane move to and then around the circle of radius $\sqrt{a}$. The stable motion for long times, namely the attractor, is periodic with frequency unity, and the attractor has dimension one. In each case, the basin of attraction is the whole plane. This change in the structure and topology of time asymptotic orbits is called a *bifurcation*.

Seen in terms of the cartesian coordinates $x = r \cos \theta$, $y = r \sin \theta$, the fixed point (0, 0) for the $\dot{x}$, $\dot{y}$ dynamics is stable for $a < 0$. Near (0, 0), small deviations satisfy

$$\begin{aligned} \frac{dx(t)}{dt} &= ax(t) - y(t), \\ \frac{dy(t)}{dt} &= ay(t) + x(t), \end{aligned} \tag{11}$$

which has solutions proportional to $\exp[\lambda t]$ with $\lambda = a + i$. So two complex conjugate eigenvalues of the linearized problem cross the imaginary axis in the λ plane when the fixed point becomes unstable. This Hopf–Andronov bifurcation or instability is generic for sets of differential equations describing physical phenomena. When we place the convection problem described in Sec. 1.1 in a rotating frame of reference, the first instability encountered as r increases can be a periodic orbit, which is said to bifurcate from the fixed point describing pure conduction.

The possibility of transition from a fixed long-term behavior to periodic orbits has been understood for a long time. In fact, a popular theory about turbulence held that it was composed of a sequence of Hopf–Andronov instabilities resulting in more and more complex time traces associated with many, many new frequencies in a fluid. The attractor would progress from a circle to a torus to tori in higher dimensions. It would always be regular, and its dimension would be integer. Experimentally, this is now known not to be correct (Gollub and Benson, 1980).

2.4 Strange Attractors and Nonperiodic Behavior

It was realized somewhat over 20 years ago (Ruelle and Takens, 1971) that when one has a dynamical evolution which shrinks phase-space volumes, the attractor need not be a regular geometrical figure with an integer dimension. These arguments, connected with observations of Smale (1967), suggested that, in fact, regular motions were unlikely, namely were not generic, for realistic dynamical evolution and in physical systems should rarely be observed. The catchy name of *"strange" attractors* was given to the geometric figures that would appear in these circumstances.

Several features of strange attractor motion soon became apparent. First, the dimension of the attractor was typically not an integer. Second, the motion on the attractor was nonperiodic, as observed by Lorenz in 1963. Third, the attractors could be compact in phase space, even while all motions were unstable. Fourth, the orbits on the attractor were "sensitive to initial conditions," meaning that two orbits very nearby would separate rapidly, exponentially fast actually, while moving about on this compact object.

Amazingly, this set of attributes, so divorced from Newtonian expectations, has been verified over and over again in experimental observations. Some of the evidence will be touched on as we proceed.

2.5 Unstable Periodic Orbits

One may wonder what happened to the periodic orbits that were emerging from the instabilities as the forcing parameters of a system were increased. These orbits remain in the phase space, but they are unstable. This means that any orbit from an arbitrary initial condition will wander near an unstable periodic orbit and then wander away to another and then to another forever. Since the Fourier spectrum of orbits on strange attractors is broad and may have no sharp peaks, unstable periodic orbits of all frequencies can be contained within the motion on a strange attractor. The collection of all of these unstable periodic orbits may be used to describe the attractor and may be viewed as some kind of nonlinear basis set for all motions. In any case, it suggests that many applications that are based on *stable* periodic motions could well be realized even in the presence of chaos using the periodic but unstable motions that abound in them. Below, we shall see that the presence of these unstable periodic orbits allows one to "control" chaos while exploiting the benefits of the broad regions of phase space visited by orbits on a strange attractor.

In Plate 1, we display the unstable periodic orbits up to period nine that lie within the strange attractor generated by the map of the plane to itself $(x(n), y(n)) \rightarrow (x(n+1), y(n+1))$,

$$\begin{aligned} x(n+1) &= 1.4 + 0.3y(n) - x(n)^2, \\ y(n+1) &= x(n), \end{aligned} \tag{12}$$

which is due to Hénon (1976). The plethora of unstable orbits is rather clear from this Plate. The unstable periodic orbits are detected within the strange attractor by placing small circles about each of the points generated by iterating the map and asking at how many steps, if at all, the orbit returns to the small circle. If it does return as the circle is made small, this locates the unstable periodic orbit.

3. MAPS AND FLOWS

All nonlinear systems depend on time in a continuous fashion. This corresponds to our daily experience, and we describe the movement of a system in time with a multivariate state space vector $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_d(t))$. When we actually make observations, we see the state of the system only at discrete times that are multiples of some sampling time τ_s characteristic of our measurement instruments. This means we see only $\mathbf{x}(n) = \mathbf{x}(t_0 + n\tau_s)$, that is, discrete samples of the state of the system. The dynamics associated with this changes from differential equations to *maps* or nonlinear systems in discrete time. So the continuous time dynamics, known as a *flow*,

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{F}(\mathbf{x}(t)), \tag{13}$$

becomes

$$\mathbf{x}(n + 1) = \mathbf{f}(\mathbf{x}(n)). \tag{14}$$

If τ_s is short, then $\mathbf{f}(\mathbf{x}) \approx \mathbf{x} + \tau_s\mathbf{F}(\mathbf{x})$.

Maps also arise when we view the system as it passes through some designated subspace of the whole phase space. For example, if we have a nonlinear system in three dimensions $\mathbf{x}(t) = (x(t), y(t), z(t))$ and view the state every time $z(t_n) = 0$, then we produce a two-dimensional map $(x(t_n), y(t_n)) \rightarrow (x(t_{n+l}), y(t_{n+l}))$ that replaces the dynamics. The qualitative features of the discrete map are often similar enough to that of the original flow that they may be used as diagnostics for the flow itself.

3.1 Stability of Maps or Flows

Maps and flows typically depend on parameters describing physical properties of the system, such as viscosities, conductivities, length scales, etc., and aspects of external forcings, such as temperature, external magnetic or electric field strengths, and so forth. As we vary these parameters, the stability of solutions to the continuous or discrete dynamics will change. For example, if we have a time-independent solution $\mathbf{x}_0$ of some discrete dynamics $\mathbf{x}_0 = \mathbf{f}(x_0, \mu)$, where μ is a set of parameters, then we can examine the evolution of this system in the neighborhood of $\mathbf{x}_0$ by setting $\mathbf{x}(n) = \mathbf{x}_0 + \delta\mathbf{x}(n)$, where, to linear order in $\delta\mathbf{x}(n)$, we have in vector notation

$$\delta\mathbf{x}(n + 1) = \mathbf{Df}(\mathbf{x}_0) \cdot \delta\mathbf{x}(n), \tag{15}$$

and

$$Df_{ab}(\mathbf{x}) = \frac{\partial f_a(\mathbf{x})}{\partial x_b}, \tag{16}$$

is the Jacobian matrix for this dynamics. The linear stability of the solution $\mathbf{x}(n) = \mathbf{x}_0$ rests on the eigenvalues of $\mathbf{Df}(\mathbf{x}_0)$. If they all lie within the unit circle, then

$$\delta\mathbf{x}(n + L + 1) = \underbrace{\mathbf{Df}(\mathbf{x}_0) \cdots \mathbf{Df}(\mathbf{x}_0)}_{L \text{ times}} \cdot \delta\mathbf{x}(n) \tag{17}$$

decays rapidly to zero, since $\delta\mathbf{x}(n + L + 1)$ behaves as $(\text{eigenvalue})^L\delta\mathbf{x}(n)$. If any eigenvalues lie outside the unit circle, the solution $\mathbf{x}_0$ is unstable. These eigenvalues move with the parameter settings μ, and, as they move out of the unit circle, the fixed point $\mathbf{x}_0$ loses stability, and a new solution dictated by the nonlinear properties of $\mathbf{f}(x)$ emerges. This kind of change in the solution to a dynamical evolution, discrete or continuous, is called a *bifurcation* and usually results in a change in the geometric structure of the solution.

The physical reason associated with this change in solution type is connected with the passage of energy through the nonlinear system. As the system is forced more and more vigorously, the optimum method for passage of this energy through the system changes. One can see this quite clearly in the free-energy arguments given by Chandrasekhar (1961) connected with the transition from heat conduction to fluid convection in a fluid heated from below. The dimensionless ratios of energy input to energy loss or dissipation characterize the allowed states of the nonlinear system, and as they are varied, the system will undergo bifurcations from one form of solution to another.

The various kinds of allowed bifurcations are well studied (Drazin, 1993). Many have been seen in physical systems. The simplest bifurcations are associated with changes in a single parameter, such as a Reynolds number or a Rayleigh number. The patterns of changes in solutions are well classified for this case. When more than one parameter is altered at a time, the class of allowed solutions is much richer (Guckenheimer and Holmes, 1986).

3.2 One-Dimensional Maps: Universality

A class of bifurcations that has been studied in detail is that of smooth one-dimensional maps $x(n + 1) = f(x(n))$, where $f(x)$ is differentiable. The paradigm for these maps is the logistic map $f(x) = rx(1 - x)$, which undergoes a sequence of bifurcations that results in periodic solutions with twice the period of the previous solution. This period-doubling bifurcation sequence (Collet and Eckmann, 1980) has *universality* properties that often appear in physical experiments. The fixed point at $x_0 = 1 - 1/r$ remains stable until $r = 3$; then two solutions (period doubling) emerge. At $r \approx 3.449$, this becomes unstable to four solutions, then eight, etc. If the lth solution appears at r_l, then

$$r_l \to K - \text{const.} \times \delta^{-1}, \tag{18}$$

where $K \approx 3.5700$, and $\delta = 4.6692\cdots$ is a new constant appearing for all $f(x)$ with smooth maximum. The other constants vary with the map. Another universal constant is associated with the rescaling properties of the solutions for the logistic map. The universal behavior translates into expected strengths for the Fourier modes of the period-doubled solutions, and these are confirmed by experiments where the dynamics appears to lie nearly on a one-dimensional subspace of the whole state space.

The universality properties of one-dimensional maps stand out as a special feature of the analysis of nonlinear systems. In most realistic systems, a sequence of period doubling may occur in some part of parameter space, but many other sequences of bifurcation or instability are also present. If one sees a period-doubling sequence in observations, then we expect to see both universality properties of period doubling:

1. the exponential decrease in the value of the bifurcation parameter between bifurcations; and
2. a specific order in which various periodic solutions occur (Devaney, 1989).

These are then useful for further analysis.

The sequence of bifurcations for the logistic map $x \to rx(1 - x)$ is shown in Fig. 6. The final 200 points after 450 iterations of the map

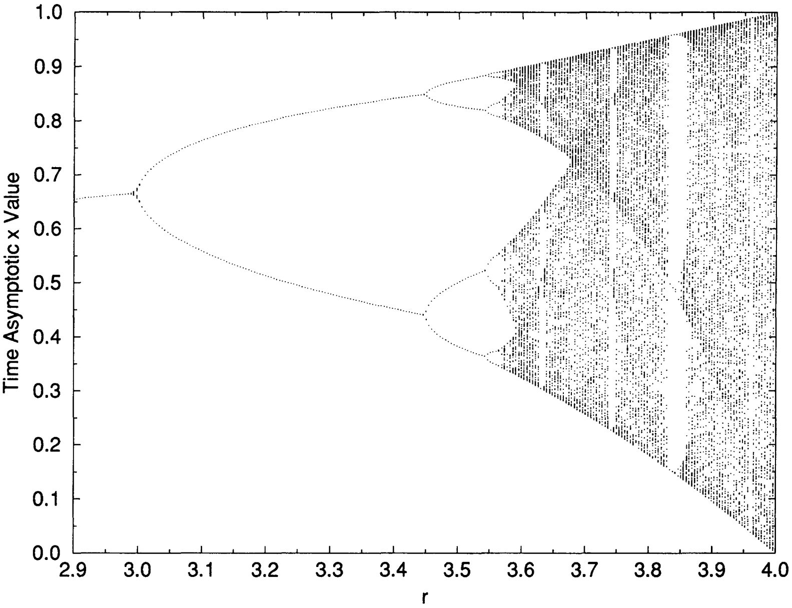

FIG. 6. Bifurcation diagram for the logistic map $x \to rx(1 - x)$ for $2.9 \leq r \leq 4$. The final 200 points from iterating the map for 450 steps are displayed at each value of r. $x(0) = 0.54$ in each case. The apparent "gap" near $r \approx 3.85$ is a region of period-three orbits, each part of which then undergoes a sequence of period-doubling bifurcations. Indeed, periodic orbits and chaotic regions are thoroughly mixed in this diagram. With only 450 points at each value of r, the full structure is not revealed.

starting from $x_0 = 0.54$ are displayed. The structure in the region $r \geq 3$ is clear.

4. DIMENSIONS OF STRANGE ATTRACTORS

A solution of a nonlinear system describes a sequence of points $\mathbf{x}(n)$ in a d-dimensional state space. One of the clearest properties of a set of points is its dimension. A surprise about chaotic dynamics of nonlinear systems is that the set of points it visits may have a dimension d_A that is not integer. This sounds peculiar or strange at first, thus the name strange attractor. Recall that the orbits are nonperiodic and cannot close, yet they must lie on a subspace of the whole state space because the dynamics dissipates energy, and so the orbit must visit an infinite number of distinct points in the state space. Under such circumstances, the dimension of that infinite set of points may be unexpected.

The idea of sets of points with noninteger dimension is really quite old (Cantor, 1883), but until experiments in physics revealed fractional dimensional sets, called *fractals*, in observations (Brandstater and Swinney, 1987), the idea remained a mathematical curiosity. Its appearance in experiments makes it a compelling notion for physics, and the idea of dimension of the long-time orbit of a dynamical system turns out to be quite useful for classifying the source of signals.

4.1 Box Counting

The crudest notion of a dimension of a set is associated with counting the number of boxes of linear side r required to capture the whole set when r becomes small. This number for any regular set (a point with $d = 0$, a line segment with $d = 1$, an area with $d = 2$, etc.) behaves as

$$N(r) \approx \text{const.}/r^d, \tag{19}$$

and, taking logarithms to eliminate the constant, we are led to expect the dimension of a set of points to be

$$d_A = \lim_{r\text{small}} \frac{\log N(r)}{\log(1/r)}. \tag{20}$$

It is not hard to find sets of an infinite number of points having dimension d_A that is not an integer. Take the set formed by successively removing from a unit line segment the middle $\frac{1}{5}$ of what remains. At the first stage, we remove the piece from $\frac{2}{5}$ to $\frac{3}{5}$. Then we remove the pieces from $(\frac{2}{5})^2$ to $\frac{6}{25}$ and the middle fifth of the piece from $\frac{3}{5}$ to 1. Continuing in this manner, we see that, at the mth stage, we have $N(r) = 2^m$ and $r = (\frac{2}{5})^m$; this leads to

$$d_A = \frac{\log 2}{\log(\frac{5}{2})} \approx 0.7564\cdots. \tag{21}$$

This *Cantor set* (Cantor, 1883) is quite typical of the structures one encounters in the results of numerical calculations of nonlinear evolution rules and from experiment.

For example, if one takes data from the Lorenz attractor for $(x(t), y(t), z(t))$ and evaluates the number of boxes required to cover the data points as the size of the boxes shrinks, then $d_A \approx 2.06$ emerges. This tells us that the geometric structure in Fig. 3 is nearly two dimensional, but the excursions of the orbit out of the plane sparsely fill the third dimension, and a cross section of the orbit would reveal a Cantor set of dimension about 0.06. The importance of the fractal dimension d_A lies in its being one of the characteristic invariants of the strange attractor. The orbit of the system as it moves around the strange attractor is enormously sensitive to small perturbations. No two orbits are identical. A quantity such as d_A is a statistical property of the density of points on the attractor $\rho(x)$ that characterizes its geometry. To explore this further, we turn to more general concepts of dimension.

4.2 Generalized Dimensions; D_q

The density of points $\rho(\mathbf{x})$ on an attractor is inhomogeneous in a typical case arising in physics. We can characterize this inhomogeneity in the distribution with appropriately chosen moments. If the orbit visits the $\mathbf{x}(n) = \mathbf{x}(t_0 + n\tau_s)$, then, absent any error in measurement or computation, the density is a sequence of delta functions at the location of the orbit points. This means that a natural density on the attractor for M data points will be

$$\rho(\mathbf{x}) = \frac{1}{M}\sum_{k=1}^{M} \delta^d(\mathbf{x} - \mathbf{x}(k)). \tag{22}$$

This is too singular to deal with, but we expect that noise or any errors will coarse-grain our ability to measure these points, and, thus, we can associate with the small measurement volume V_j near a point $\mathbf{x}(j)$ the normalized frequency of visit to that volume

$$\begin{aligned} p(\mathbf{x}(j), r) &= \int_{V_j} d^d x \rho(\mathbf{x}) \\ &= \frac{1}{M}\sum_{k=1}^{M} \theta(r - |\mathbf{x}(k) - \mathbf{x}(j)|), \end{aligned} \tag{23}$$

where the step function $\theta(w) = 0$, $w < 0$, and $\theta(w) = 1$, $w > 0$. The length r is a typical size of the volume V_j. This probability or frequency is no longer singular. As M becomes large and r becomes small, it captures the essential idea we have of number of points in a small volume on the attractor. The moments of this quantity are useful for defining generalized dimensions. We consider (Grassberger, 1983)

$$C_q(r) = \sum_{j=1}^{M} p(\mathbf{x}(j), r)^q. \tag{24}$$

For $q = 0$, this gives just the number of boxes needed to cover the set of orbit points and scales with r as $r^{-D_0} = r^{-d_A}$. This defines the fractal dimension d_A, which we have identified with the "box-counting dimension" D_0. For $q \neq 0$, we expect this to scale as

$$C_q(r) \approx r^{(q-1)D_q}, \quad r \text{ small}, \tag{25}$$

defining the generalized dimension D_q.

In this discussion, we encounter quite a few "dimensions." The integer dimension of the phase space of a system we have been calling d. The fractal dimension of an attractor that we observe in the phase space we have called d_A and identified with the power D_0, also fractional, entering in the small-r behavior of $C_0(r)$. The generalized dimension D_q is also fractional, as a rule.

Since $p(\mathbf{x}(j), r) \leq 1$, $p(\mathbf{x}(j), r)^{q_1} \geq p(\mathbf{x}(j), r)^{q_2}$ if $q_2 \geq q_1$, implying $D_{q_1} \geq D_{q_2}$. This means we have a whole spectrum of dimensions, which are all equal in the rare case when the attractor is uniformly populated. The various D_q can be calculated using the "correlation integrals" $C_q(r)$, and one finds that the D_q vary smoothly in the neighborhood of $q = 0$. Here $q = 1$ is a special case worth noting. Since

$$D_q = \lim_{r \text{small}} \frac{1}{q-1} \frac{\log C_q(r)}{\log r}, \tag{26}$$

we need to take the limit to find

$$C_1(r) = \sum_{j=1}^{M} p(\mathbf{x}(j), r) \log[p(\mathbf{x}(j), r)],$$

and

$$D_1 = \lim_{r \text{small}} \frac{\log C_1(r)}{\log r}. \tag{27}$$

This defines the information dimension.

In a typical experiment involving fluid flow between rotating cylinders, it is possible to evaluate the $C_q(r)$ using an "embedding" method to be described below. Values for D_1 and D_2 have been extracted from the observations. They vary with the relative rotation rates of the cylinders and increase with this rate, as one might expect from their interpretation as a sense of the number of degrees of freedom excited in the fluid. The observed values (Brandstater and Swinney, 1987) range smoothly between nearly 2 and about 4.5 as the speed of the rotating cylinders is increased by about a factor 2. Because of the accuracy of real data from these experiments, the difference between D_1 and D_2 is not distinguishable.

All of these dimensions are of importance because they serve to classify the attractor and thus are useful in experiment for identifying the source of a chaotic signal, so that one can recognize it in any future measurement. The point here is that any given orbit of the system is unstable to small perturbations of the orbit. Such perturbations, as we shall see in the next section, grow exponentially rapidly, so that the original orbit and any orbit near it separate in phase space while both are moving about on the attractor. They move quite differently on the attractor, so that they rapidly become uncorrelated. This means that every measurement of an orbit will yield a different answer in detail, and orbits from experiment and from models cannot be compared. One must compare items not sensitive to initial conditions or perturbations. The D_q are among such quantities.

4.3 Statistics on Strange Attractors

Since any given orbit $\mathbf{x}(t)$ or $\mathbf{x}(n)$ is extremely sensitive to perturbations or changes in initial conditions, we cannot compare the computed orbits from a model to any that are observed. We must have some basis on which to test models and to compare one experiment with another, and this basis is given by the statistical properties of strange attractors (Abarbanel *et al.*, 1993; Eckmann and Ruelle, 1985).

The statistics are based on the distribution of points $\rho(\mathbf{x})$ we saw above. Any function on the state space $g(\mathbf{x})$ allows us to define a quantity, its mean over the attractor weighted with the density of points, that is unchanged under changes in the initial condition. The mean of $g(\mathbf{x})$ is

$$\begin{aligned}\bar{g} &= \int d^dx g(\mathbf{x})\rho(\mathbf{x}) \\ &= \frac{1}{M}\sum_{k=1}^{M} g(\mathbf{x}(k)). \end{aligned} \qquad (28)$$

If we now estimate this using $g(\mathbf{f}(\mathbf{x}))$, where we assume $\mathbf{x}(k+1) = \mathbf{f}(\mathbf{x}(k))$, that is, we start one step further along on the orbit, we find that $\bar{g}$ is unaltered to order $1/M$. So, as M becomes large, the mean taken with $\rho(\mathbf{x})$ as the probability density is an invariant of the dynamics. For the correlation functions $C_q(r)$ that define the generalized fractal dimensions D_q, we must take

$$\begin{aligned} g(\mathbf{x}) &= \frac{1}{M}\sum_{n=1}^{M} \theta(r - |\mathbf{x} - \mathbf{x}(n)|), \\ C_q(r) &= \int d^dx\rho(\mathbf{x}) \frac{g(\mathbf{x})^{(q-1)}}{q-1}. \end{aligned} \qquad (29)$$

This means that the D_q are invariants of the motion and can be used to classify the source of an observation. This is very useful, as we can now compare the D_q, or, actually, the entire function $C_q(r)$, as seen in observations with other observations to establish if we are seeing a known system, or with computations on a model as part of the verification process for the model itself.

5. LYAPUNOV EXPONENTS

The D_q and associated moments of other physically interesting quantities are properties of the static, time-evolved distribution of points on the attractor, as represented by $\rho(\mathbf{x})$. There are other invariants associated with the dynamical evolution on the attractor, and they give a sense of the instability associated with the formation of the attractor, as well as of the ability to make predictive models of the nonlinear system.

The quantities that govern the stability of an observed orbit $\mathbf{x}(n)$ are called *Lyapunov exponents*. They are defined by looking at the linearized evolution of perturbations to a solution to $\mathbf{x}(n+1) = \mathbf{f}(\mathbf{x}(n))$, which we have seen obeys

$$\delta\mathbf{x}(n+1) = \mathbf{Df}(\mathbf{x}(n)) \cdot \delta\mathbf{x}(n), \qquad (30)$$

so that

$$\begin{aligned} \delta\mathbf{x}(N+L+1) &= \mathbf{Df}(\mathbf{x}(L+n)) \\ &\quad \cdot \mathbf{Df}(\mathbf{x}(N+L-1)) \\ &\quad \cdots \mathbf{Df}(\mathbf{x}(n)) \cdot \delta\mathbf{x}(n) \\ &= \mathbf{Df}^L(\mathbf{x}(n)) \cdot \delta\mathbf{x}(n). \end{aligned} \qquad (31)$$

The eigenvalues of the L times composed Jacobian matrices $\mathbf{Df}(\mathbf{x})$, which we have called $\mathbf{Df}^L(\mathbf{x})$, determine the stability of the observed orbit under small perturbations.

5.1 Oseledec's Multiplicative Ergodic Theorem

In 1968, Oseledec (Oseledec, 1968) proved the remarkable result that, if we consider the orthogonal matrix (T denotes transpose)

$$\mathbf{OSL}(\mathbf{x}, L) = \{\mathbf{Df}^L(\mathbf{x})^T \cdot \mathbf{Df}^L(\mathbf{x})\}^{1/2L}, \qquad (32)$$

its eigenvalues as $L \to \infty$ are independent of $\mathbf{x}$ for $\mathbf{x}$ within the basin of attraction of any attractor. These eigenvalues, written as $\exp[\lambda_1]$, $\exp[\lambda_2]$, ..., $\exp[\lambda_d]$, define the Lyapunov exponents λ_a, which we order as $\lambda_1 \geq \lambda_2 \geq \cdots$. The independence of $\mathbf{x}$ on the λ_a is known as the multiplicative ergodic theorem. The remainder of the theorem proves that the eigendirections associated with the $\exp[\lambda_a]$ exist. It is very important that associated with the linearization of the nonlinear dynamics is a set of quantities that are independent of starting location within a basin of attraction. This means that each of the λ_a serves as a characteristic of the attractor itself and is not associated with any particular orbit. Further, it is easy to show that, under any smooth change of coordinates $\mathbf{x} \to H(\mathbf{x})$ where derivatives of $H(\mathbf{x})$ exist in phase space near the attractor, the values of the λ_a are unchanged. To be more precise, when one makes a smooth change of variables, the limit $L \to \infty$ acquires a piece that is of order $1/L$ associated with the change of coordinates. This vanishes as $L \to \infty$, of course, but changes the interpretation of local properties on the attractor.

If the orbit satisfies a differential equation, that is, $\mathbf{x}(t)$ comes from a flow, then at least one of the λ_a is identically zero. This is seen by considering a perturbation $\delta\mathbf{x}(n)$ that is precisely along the direction of $\mathbf{f}(\mathbf{x}(t)) = \dot{\mathbf{x}}(t)$. If the system producing the signal is Hamiltonian, then the symmetry of Hamiltonian dynamics guarantees that the sum of the exponents is zero (nonshrinking volumes), and the exponents come in pairs (λ, $-\lambda$). This tells us that, when we have observed data from which we evaluate Lyapunov exponents, we can determine without knowing the equations of motion if the system is a flow or a map and whether it is Hamiltonian or not.

5.2 Positive Exponents and Strange Attractors

If one or more of the λ_a is positive, the perturbation is unstable. It does not grow without bound, however, but is folded back into the region of the attractor by the dissipation associated with the dynamics or, in the case of a Hamiltonian system, moves out of a region of resonance and wanders through phase space. What is critical is that, since the error $\delta\mathbf{x}(n)$ grows exponentially rapidly and soon reaches the size of the attractor itself, predictability is lost on a time scale of order $1/\lambda_1$.

Line segments in the d-dimensional phase space grow as $\exp[\lambda_1]$; areas grow as $\exp[\lambda_1 + \lambda_2]$; three-dimensional volumes as $\exp[\lambda_1 + \lambda_2 + \lambda_3]$; d-dimensional volumes, as $\exp[\lambda_1 + \lambda_2 + \cdots + \lambda_d]$. Since we know that volumes in the whole space must shrink, we must have $\lambda_1 + \lambda_2 + \cdots \lambda_d < 0$. This means that, somewhere between line segments and d-dimensional volumes, there is a sense of a dimension associated with a subvolume of the whole space where the subvolume neither grows nor shrinks. This allows one to define a *Lyapunov dimension* D_L (Kaplan and Yorke, 1979), which should be essentially the dimension of the attractor. The Lyapunov dimension is defined by considering that integer K where $\Sigma_{a=1}^{K} \lambda_a > 0$ but $\Sigma_{a=1}^{K+1} \lambda_a < 0$; then

$$D_L = K + \sum_{a=1}^{K} \lambda_a/|\lambda_{K+1}|. \tag{33}$$

The term D_L is conjectured to be equal to D_1, the information dimension.

5.3 Global and Local Exponents

The Lyapunov exponents λ_a tell us what happens to a perturbation a long time after it is made. More to the point in considering prediction, we need to know what happens a finite time after a perturbation is made, since we can only predict for a short time in any case. The answer to this question is given by the local Lyapunov exponents $\lambda_a(\mathbf{x}, L)$, which are the logarithms of the eigenvalues of the matrix $\mathbf{OSL}(\mathbf{x}, L)$ given above. The $\lambda_a(\mathbf{x}, L)$ satisfy $\lambda_a(\mathbf{x}, L) \to \lambda_a$ for large L, but the variations of the $\lambda_a(\mathbf{x}, L)$ as we visit parts of the attractor are quite large for small L. The variance of the values of $\lambda_a(\mathbf{x}, L)$ about the mean falls to zero as $1/L^p$ where $\frac{1}{2} \le p \le 1$. Local exponents can be evaluated from observed data, as we will discuss below.

6. INFORMATION THEORY AND STRANGE ATTRACTORS

The presence of positive Lyapunov exponents but finite-size attractors distinguishes nonlinear systems from linear systems in a striking way. Linear systems have all $\lambda_a \le 0$, or they make no sense. Associated with this is the notion of a chaotic nonlinear system as a generator of information in the direct sense of Shannon (1948). The idea is that if we have a resolution in any observation given by a size R in phase space, then two state points $\mathbf{x}_1$ and $\mathbf{x}_2$ within the ball of size R cannot be distinguished. A time T later, the distance between these two points has grown to $|\mathbf{x}_1 - \mathbf{x}_2| \exp[\lambda_1 T]$, and when this is larger than $2R$, we can distinguish the two points that were hidden by our instrument resolution at the initial moment. The instability associated with $\lambda_1 > 0$ has revealed information not measurable at the start.

This suggests that the theory of information may well play a role in the description of nonlinear systems that is not realized in linear analyses. This led Fraser and Swinney (1986) to suggest that when determining when two events are independent in a nonlinear system, one should use the notion of *mutual information* (Gallager, 1968). We associate with the state-space sequence $\mathbf{x}(j)$ a probability distribution $P_x(\mathbf{x})$. Similarly, with another sequence $\mathbf{y}(n)$, we associate $P_Y(\mathbf{y})$. If we measure $\mathbf{x}(j)$, how much do we learn about $\mathbf{y}(n)$? The answer (in bits) is

$$I_{XY}(\mathbf{x}(j), \mathbf{y}(n)) = \log_2 \left[\frac{P_{XY}(\mathbf{x}(j), \mathbf{y}(n))}{P_X(\mathbf{x}(j)) P_Y(\mathbf{y}(n))} \right], \tag{34}$$

where $P_{XY}(\mathbf{x}, \mathbf{y})$ is the joint distribution of both kinds of events. If the measurements $\mathbf{x}$ and $\mathbf{y}$ are independent, then $P_{XY}(\mathbf{x}, \mathbf{y}) = P_X(\mathbf{x}) P_Y(\mathbf{y})$, and $I_{XY}(\mathbf{x}, \mathbf{y}) = 0$. This is to be seen as a statistic on the attractor where $\mathbf{x}$ and $\mathbf{y}$ are drawn from points on the attractor. As pointed out earlier, the average of this over the attractor with the natural density leads to an invariant. This yields the *average mutual information*

$$I_{XY} = \sum_{j,n} I_{XY}(\mathbf{x}(j), \mathbf{y}(n)) P_{XY}(\mathbf{x}(j), \mathbf{y}(n)), \tag{35}$$

which is characteristic of the systems X and Y and not of any particular orbits $\mathbf{x}(j)$ or $\mathbf{y}(n)$.

In analyzing chaotic systems, we may ask what "correlation" two events $\mathbf{x}(j)$ and $\mathbf{y}(n)$ have with each other. The average mutual information statistic allows us to probe the nonlinear connection between these two events, and there is some evidence that this probe can be much more sensitive than the familiar linear cross correlation $\Sigma_n \mathbf{x}(n + m)\mathbf{y}^T(n)$.

If the measurements $\mathbf{x}(j)$ and $\mathbf{y}(n)$ are the same, then the average mutual information is

$$h_X = -\sum_n P_X(\mathbf{x}(n)) \log P_X(\mathbf{x}(n)), \tag{36}$$

which looks like the Boltzmann entropy for this system. It is called the Kolmogorov–Sinai entropy, and a result due to Pesin (Eckmann and Ruelle, 1985) connects it with the positive Lyapunov exponents of the source:

$$h_X = \sum_{\lambda_a > 0} \lambda_a. \tag{37}$$

This corresponds to our heuristic explanation of the role played by information in describing nonlinear systems.

7. ANALYSIS OF EXPERIMENTAL DATA

The developments discussed above are useful as insights into how one thinks "nonlinearly." When confronting experiments on physical systems and when choosing how to model or control or predict in physical systems, we must understand how to extract from these observations characteristics such as fractal dimensions and Lyapunov exponents.

The challenge is more severe since in most observations, we measure one, or sometimes a few, variables of the physical dynamics. In a fluid dynamical experiment, we may measure the fluid velocity or temperature or density at some points, perhaps over a long period of time. Unfortunately, we do not measure the fluid velocity everywhere in the fluid or even usually measure all components of the velocity. Rarely does one also measure pressure and density and velocity and temperature, except at a few locations within the fluid. In a nonlinear circuit, we typically observe one or a few voltages in the circuit and rarely measure all independent voltages and currents describing the circuit.

7.1 Time-Delay Embedding

Nonetheless, there is a striking result due to Ruelle and elaborated on by Takens (1981) and Mañé (1981) that accurate measurement of a single variable, call it $x(n)$, will suffice for reconstructing all of the statistical information we seek. The idea is that knowledge of all the time derivatives of the observed variable would capture the full operating phase space of the system. We could try to create these derivatives from the $x(n)$ via

$$\dot{x}(n) = \frac{x(t_0+(n+1)\tau_s)-x(t_0+n\tau_s)}{\tau_s}; \tag{38}$$

for finite τ_s, this would be a linear filter of our data and one that requires very accurate knowledge of the data $x(n)$ because the filter involves differences. Accurate evaluations of small differences between large numbers require very accurate measurements. Ruelle observed that for purposes of providing coordinates for the phase space of the system, we do not need the derivatives, but the time-delayed variables $x(t_0 + (n + 1)\tau_s)$ would do just fine; indeed, that is the only new information in the filter we call $\dot{\mathbf{x}}(n)$. This suggests building d-dimensional spaces from vectors composed of time delays of the observations:

$$\begin{aligned}\mathbf{y}(n) &= \{x(t_0 + n\tau_s), x(t_0 + (n + T)\tau_s), \\ &\quad \ldots, x(t_0 + [n + (d - 1)]T\tau_s)\} \\ &= [x(n), x(n + T), \ldots, x(n + (d - 1)T)].\end{aligned} \tag{39}$$

The Takens–Mañé results tell us on the basis only of geometry that if we choose $d > 2d_A$, where d is integral while d_A need not be, then we have sufficient coordinates to represent faithfully the space of the source of the $x(n)$. The essential idea behind the result is that we are observing the strange attractor of the nonlinear system as projected on the observation axis $x(n)$. To "unfold" this projection, we must provide enough coordinates to allow the orbit to evolve without intersecting itself. In a space of dimension d, subspaces of dimension d_1 and d_2 intersect generically in a space of dimension $d_1 + d_2 - d$. Self intersections of a space with itself, $d_1 = d_2 = d_A$, are avoided if the dimension of the intersections is less than zero; so $d > 2d_A$. This is a sufficient condition and is independent of the value of the time lag T, itself a multiple of τ_s.

To use this result, we must find rules that allow us to choose a good time delay T and a dimension d, called the embedding dimension, using properties of the data.

7.2 Average Mutual Information

Since the unfolding result holds for arbitrary T, we need another principle to choose it. The choice of T too small would be uninteresting, since then the coordinates $x(n)$ and $x(n + T)$ of our data vector $\mathbf{y}(n)$ are essentially the same; the dynamics would not have had enough time to act to distinguish the two states at those times. If the lag T is too large, however, then the instabilities in the system should have destroyed the connection between $x(n)$ and $x(n + T)$, and so T too long is not helpful. We require a way to choose an intermediate time lag T.

For this, we turn to average mutual information, as that tells us the nonlinear "correlation" between $x(n)$ and $x(n + T)$. We form the distribution of $x(n)$, $P(x(n))$, which is also $P(x(n + T))$, if there are enough data. Then we form the joint distribution $P(x(n), x(n + T))$, and the average mutual information

$$\begin{aligned}I(T) = &\sum_{x(n+T)} P(x(n), x(n + T)) \\ &\times \log_2 \left[\frac{P(x(n), x(n+T))}{P(x(n))P(x(n+T))}\right].\end{aligned} \tag{40}$$

When this quantity, which satisfies $I(T) \geq 0$, has its first minimum, we have a compromise time delay that is quite useful for forming data vectors $\mathbf{y}(n)$ that have rather independent components $x(n + kT)$, $k = 1, 2, \ldots, d - 1$.

For data from the hysteretic circuit discussed earlier, we have the average mutual information curve shown in Fig. 7. Look again at Fig. 5 to see the strange attractor in phase space made out of vectors $(x(n), x(n + T_0), x(n + 2T_0))$ where $T_0 = 6$ is chosen from the average mutual information curve.

7.3 Global Embedding Spaces

The choice of the embedding dimension d required by the data is not as prescriptive as the choice of time lag T. In fact, the route to take is guided by the Takens result about unfolding the attractor by adding dimensions to the data vector. As one adds dimensions, we undo apparent intersections of the orbit with itself due to projection of a higher dimensional object into a lower dimension. If we keep track of these intersections and establish when they disappear, we should be

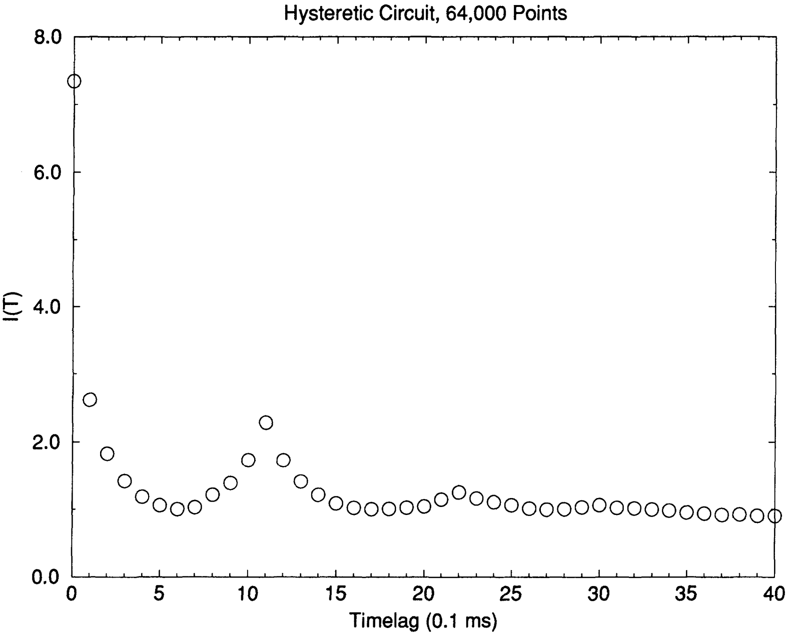

FIG. 7. Average mutual information $I(T)$ for data from the nonlinear hysteretic circuit as shown in Fig. 4. The first minimum of $I(T)$ is at $T = 6$, and this value is used in creating time-delayed coordinates for the display of the attractor in Fig. 5.

able to establish an embedding dimension for the data vectors that is necessary according to the data. The criterion $d > 2d_A$ is sufficient and generic.

It is important from a practical point of view to establish the necessary dimension, since all additional components of the data vector will be occupied by contamination or noise, which respects no finite dimension. To alleviate the effects of this, we must work in as low a dimension as is permitted by the data. It is also true that the amount of calculation one needs to do to evaluate fractal dimensions or Lyapunov exponents rises rapidly with the embedding dimension, so that having the smallest d one can utilize is quite important.

7.3.1 True Vector Fields; False Nearest Neighbors The key to establishing the necessary dimension for a set of data is the idea that orbit points should be near each other only by virtue of the dynamics, not because of a projection from another part of the attractor because we work in too small a space. Dynamically nearby orbit points will not move away from each other as we increase the dimension, but points nearby because of projection will move apart and stay apart. This suggests the method of *false nearest neighbors*, which systematically adds components to the time-delay data vector $\mathbf{y}(n)$, asking at each dimension what percentage of neighbors at the previous dimension are no longer neighbors. When this number drops to zero, the necessary dimension for unfolding has been reached. The key to this procedure is a systematic search for the neighbors of any data point, and this can be done quite efficiently (Kennel *et al.*, 1992). In Fig. 8, we have the percentage of false nearest neighbors for the data from the hysteretic circuit discussed before. Clearly, the number of false neighbors vanishes at $d = 3$; the use of $T = 6$ comes from mutual information.

The other geometric test for embedding dimension rests on the observation that the direction of the vector field $\mathbf{F}(\mathbf{x}(t))$ is unique in small regions of phase space. Thus, if one can carve up the phase space surrounding an attractor and determine when the local vector field is unique in each small sector, then the minimum embedding dimension will have been uncovered (Kaplan and Glass, 1992). This procedure works well in low dimensions and in the absence of any contamination of the

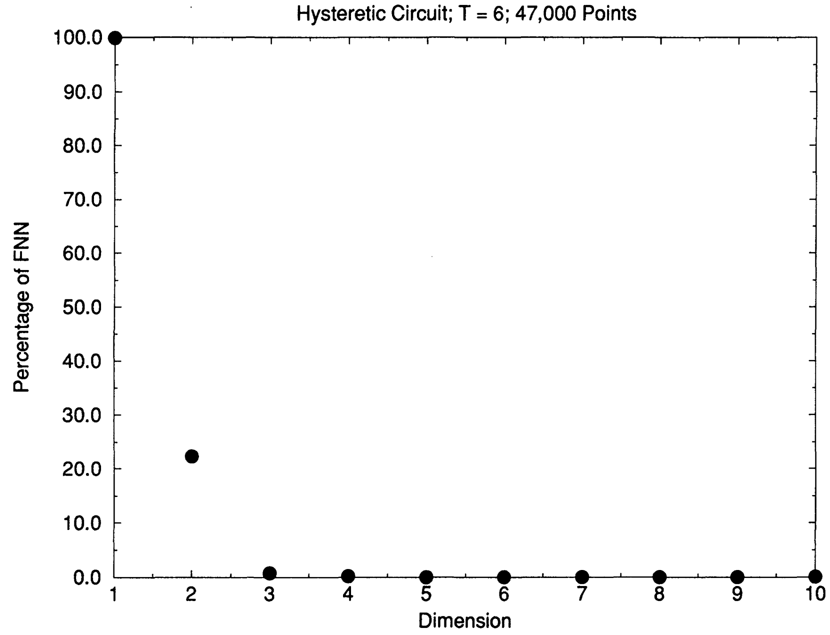

FIG. 8. The percentage of false nearest neighbors for the data from the nonlinear hysteretic circuit seen in Fig. 4. Using the time delay determined by average mutual information shown in Fig. 8, we see that false neighbors go to zero at $d = 3$. This tells us that the attractor can be unfolded in three dimensions, and the display of Fig. 5 is based on this information.

data. In high dimensions, the number of regions one must explore becomes quite large. Further, the accuracy with which one can determine by finite differencing the direction of $\dot{\mathbf{x}}(t)$ becomes quite dependent on the number of data, as distances on an attractor are approximately proportional to

$$(1/N)^{1/D_1} \tag{41}$$

for N data points.

7.3.2 Local Dimensions Even when one has determined the global embedding dimension required to unfold the attractor, the number of degrees of freedom required to describe the dynamics has been given an upper bound. As a contrived example, one can image a flow in two dimensions which evolves on a twisted surface or a Möbius strip. The global embedding dimension is $d = 3$, but the dynamics is two dimensional. If one wishes to evaluate Lyapunov exponents for these data, we had better know that only two of them are dynamical and the third purely geometric.

To discover the local dimension, one needs to work in the global space that separates all points properly and then move locally onto the attractor to establish what dimension captures the evolution of points from neighborhood to neighborhood. Adding this dynamical aspect to the determination of a dimension works rather well and is rather robust to external noise (Abarbanel and Kennel, 1993).

7.4 "Noise" versus Chaos

The whole discussion of dimension raises the very interesting issue of what constitutes "noise" and how we are to distinguish it from chaos. It should be clear that noise in a classical setting, compared to quantum mechanical spontaneous emission, for example, may well be seen as very high-dimensional dynamics. High is rather much in the eye of the beholder or in the ability of the computer. If the dimension of a signal is determined to be larger than $d = 25$, let us say, then, for practical purposes, it may as well be considered infinite dimensional and treated as a stochastic source of contamination of other signals; that is, we treat it just as a traditional view of noise would dictate. Somewhere between a low dimension, $d = 5$, say, and this higher dimension, we lose our ability in a practical sense to analyze and use properties of the structure in state space that are an intrinsic part of the solutions to nonlinear systems. This is an artificial barrier, which we may overcome dimension by dimension as computational capability increases. One should recognize that there is no fundamental difference between noise and chaos, as the former is simply the latter in high dimensions, and, as time goes by, the boundary between our treatment of the two is likely to become less and less clear. Basically, one can say that, when it is possible in a practical sense to utilize the state-space structure of a signal, one should do it. When it appears not to be possible, statistical methods of a traditional sort are all we can turn to.

7.5 Dimensions from Data

To determine fractal dimensions from observations of data, the route is now clear. From the observed scalar data $x(n)$, we need to find an appropriate time delay $T\tau_s$ and an appropriate embedding dimension d with which to make data vectors $\mathbf{y}(n) = [x(n), x(n + T), \ldots, x(n + (d - 1)T)]$. With these data, we need to form the number density in any sphere of radius r about each phase space point $\mathbf{x}$. This is

$$n(\mathbf{x}, r) = \frac{1}{M} \sum_{n=1}^{M} \theta(r - |\mathbf{y}(n) - \mathbf{x}|), \tag{42}$$

as we noted before. The moments of this with the natural density $\rho(\mathbf{x})$ give the correlation functions

$$C_q(r) = \frac{1}{q-1} \int d^d x n(\mathbf{x}, r)^{(q-1)} \rho(\mathbf{x}), \tag{43}$$

and this scales for small r as $r^{(q-1)D_q}$. Computing the $C_q(r)$ has become a kind of cottage industry. The most popular is the evaluation of D_2, for which an enormous body of literature exists (Abarbanel *et al.*, 1993).

The accurate computation of $C_q(r)$ and the extraction of the D_q by looking at plots of log $C_q(r)$ vs log r involve many numerical issues, not the least of which has to do with the fact that this plot is often not even linear. However, when one is quite fortunate, one can extract from these plots acceptable values of the fractal dimensions and use these values for classifying the attractor. Many have focused on these numbers as a sign of the low dimensionality of the source producing the observations, and while that is true, the utility of the D_q as classifiers is much more significant. If it is only low dimensionality one wishes to identify, the geometrical methods above (Sec. 7.3.1) are less computationally demanding and more robust against signal contamination.

The collection of D_q's serves to identify the attractor with which one is dealing. What constitutes a "complete" collection of these invariants is an unanswered question at this time. The utility of a collection of different quantities with which to classify an attractor is clear, and it is quite likely that, absent a clear answer on the completeness of some set of descriptors, one will proceed in a practical sense to use all that can be evaluated reliably.

7.6 Lyapunov Exponents from Data

The route to the determination of Lyapunov exponents from data is also reasonably clear. One must construct data vectors $\mathbf{y}(n)$ as described and then use these to form numerical estimates of the Jacobians $\mathbf{Df}(\mathbf{x})$ at various points in the phase space. The general formulation of this requires creating local maps in the state space that evolve whole phase-spatial neighborhoods into other neighborhoods. In this way, temporal data can be converted into data in phase space for the numerical evaluation of the space derivatives required in the Jacobians. If the local map $\mathbf{f}(x)$ is represented in terms of some basis set of known functions $\phi_k(\mathbf{x})$,

$$\mathbf{f}(x) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(\mathbf{x}), \tag{44}$$

then the coefficients $\mathbf{a}_k$ can be determined by looking at each data point $\mathbf{y}(n)$ and its neighbors and the temporally subsequent data point $\mathbf{y}(n + 1)$ and its neighbors. Then a least-squares fit to the evolution of the set of neighbors near $\mathbf{y}(n)$ into the set of neighbors near $\mathbf{y}(n + 1)$ gives the $\mathbf{a}_k$. The local Jacobian follows as

$$Df_{jm}(\mathbf{y}(n)) = \sum_{k=1}^{K} (a_k)_j \left. \frac{\partial \phi_k(\mathbf{x})}{\partial x_m} \right|_{\mathbf{x}=\mathbf{y}(n)}. \qquad (45)$$

The Oseledec matrix can then be evaluated as a product of local Jacobians and with some care (Abarbanel *et al.*, 1993) can be diagonalized to determine both the local and global Lyapunov exponents.

7.7 Using Invariants of the Dynamics

Of what use can these invariants be?

7.7.1 Classification and Identification First, as we have suggested, these quantities such as fractal dimensions and Lyapunov exponents are averages over the attractor of physically motivated statistics. They are independent of initial conditions, as the particular orbits of the system are not. Since one may evaluate these quantities from observed data and since they are independent of the coordinate system in which one evaluates them, they naturally have a role in classifying the system that produced the signal. If one is concerned with knowing whether the signal that she sees this week is from the same source as the signal seen last week, then these are the identifiers which she must use for those purposes. From the point of view of system design for this or that specific application, these are the parameters toward which one can direct the design procedures. Suppose one wants a source of signals that has less or more predictability; then attaching the capability for evaluating Lyapunov exponents to the design tools will enable this aspect of the design to be achieved.

At the very least, whether one wishes to manipulate these quantities for design purposes or catalog them for identification uses, they provide the correct taxonomy for recognition of the source and the basis for listing its physical properties. In this fashion, they replace the familiar resonant frequencies of a system, which is a concept tuned to linear dynamics, while the material we have been discussing is focused on nonlinear systems.

7.7.2 Models and Observations The critical role played by the invariants arises when one is trying to make models of the physics of the source of observations. The models which one can make basically fall into two categories:

1. Black-box-like models, which provide local or global "fits" to the data in the multidimensional reconstructed phase space of the data vectors $\mathbf{y}(n)$. Such models consist of parametric rules for reproducing the observed $\mathbf{y}(n) \rightarrow \mathbf{y}(n+1)$, and the parameters are chosen by some kind of minimization procedure,most efficiently a least squares estimation. For local modeling, we create rules that fit the way the neighborhood of points in phase space around the point $\mathbf{y}(n)$ evolves into the neighborhood of points around $\mathbf{y}(n+1)$. Some basis set of functions is chosen for this fit, but if the neighborhoods are small enough, polynomials will do fine. If not, more sophistication is called for (Abarbanel *et al.*, 1993).

   Global models perform the same tasks of a least-squares fit to parameters but use all the data on the attractor at once. The benefit of the global models is their compact form, which is used to "explain" all of the data, but one must have a sensible choice of model to begin with. The number of models is clearly infinite, and incorrect choices may be hard to distinguish from one another. Local models are less elegant and use more computations, but can basically be used to fit the evolution of nearly anything in phase space. Local models put together can certainly reproduce any dynamics, but at the cost of a large lookup table, which essentially translates the neighborhood-to-neighborhood behavior into a numerical rule. Enormously accurate predictive models can be achieved with high enough order local models.

   Since each form of model works by evaluating its parameters by a least-squares fit, there is no guarantee that the properties such as fractal dimensions or Lyapunov exponents will be properly reproduced, since they represent other aspects of the dynamics. This allows one to use the invariants as constraints on model making and may usefully limit the variation in parameters that one need explore. Indeed, it happens that unconstrained modeling may result in very bad values for Lyapunov exponents, while constrained models, differing only slightly from the very best least-squares fit, do substantially better.

2. Physically based models. From the point of view of doing physics with chaotic nonlinear systems, the most important aspect of the invariants is to allow one to compare observations with models. Suppose one has observations of a plasma under some circumstances where the signals are chaotic and has a proposed model for the dynamics based on approximations to the electrodynamics and possibly thermonuclear reactions of the constituents of the system. Then using the invariants extracted from the observations and comparing these to the invariants evaluated directly from the model may allow one to exclude the model or change parameters in the model to correspond to the experiment or verify one's physical sense of the dynamics. In some sense, the usual activities of model making and verification on the part of physicists are moved to this arena and away from the Fourier-based tradition that is appropriate for linear dynamics.

7.8 "Control" of Chaos

It is not always a welcome discovery to learn that the physical system one has created is operating in a chaotic mode. If this is not the desired operating condition, one can utilize methods to "control" that chaotic motion. The basic idea is to locate some of the infinite number of unstable periodic orbits that comprise a strange attractor, and, in a locally linear analysis about these periodic orbits, identify the stable and unstable directions associated with perturbations near the orbits. Then one can devise small changes in the parameters of the system that drive observed evolution onto the stable direction of one of these periodic orbits (Ott *et al.*, 1990). Repeated application of the required small parameter changes drives the system onto the stable direction of the unstable periodic orbit that was chosen. Basically, one is turning an autonomous problem where a parameter is held constant into a nonautonomous problem where an external force is altering the parameter according to the system location in phase space. Thus, we create a new system in which the previously unstable periodic orbit has become stable.

This method has proven successful in stabilizing the unstable periodic orbits of chemical reactions, cardiac cells, infrared lasers, magnetoelastic ribbons, and nonlinear circuits, among other systems. It suggests that one might actually wish to drive a system of interest into a chaotic regime where it visits an enormously larger piece of state space than in regular motion and then stabilize the evolution to one or more chosen periodic orbits at will. This kind of flexibility will no doubt be quite useful in applications of these methods.

GLOSSARY

Attractor: The set of points in phase space visited by the solution to an evolution equation long after transients have died out. An attractor can have an integer dimension—a regular attractor—or a fractional dimension—a *strange* attractor.

Average Mutual Information: The amount (in bits if base-two logarithms are used) one learns about one observation from another observation on the average over all measurements. This is a statistic introduced in communications theory by Shannon (1948), which describes chaotic nonlinear systems as they are information sources.

Basin of Attraction: The set of points in phase space that, taken as initial conditions for an evolution rule, lead to the same attractor. The boundaries between basins of attraction can be fractional dimensional.

Bifurcation: A change in the state or topology of the solution to an evolution equation when a parameter is varied. A bifurcation is associated with changes in stability of solutions to dynamical nonlinear equations.

Cantor Set: An uncountably infinite set of points in R^d (a) which contains no subinterval of R^d, (b) which is closed, and (c) each point of which is an accumulation point. The dimension of such a set is typically fractional. We have called this dimension d_A, since, in physical applications, the Cantor set is an attractor of a dynamical system.

Chaos: Deterministic evolution of a nonlinear system that is between regular behavior and stochastic behavior or " noise." This motion of nonlinear systems is slightly predictable and nonperiodic, and specific orbits change exponentially rapidly in response to changes in initial conditions or orbit perturbations.

Embedding Dimension: The dimension of phase space required to unfold the attractor of a nonlinear system from the observation of scalar signals from the source. This dimension is integer.

Flows and Maps: Flows are evolution rules in continuous time. Maps are evolution rules in discrete time. Maps can come from flows by finite sampling associated with real experiments or from Poincaré sections, which view a continuous orbit stroboscopically.

Fractal Dimensions: The dimensions of an infinite set of points associated with the way the density of points scales with small volumes surrounding the points. Fractional values for the dimensions are typical for strange attractors of nonlinear systems.

Hamiltonian Systems: Laws of evolution corresponding to closed systems or systems without dissipation. The equations of motion come from a scalar "Hamiltonian" and, when expressed in canonical coordinates, the Hamiltonian flow preserves volumes in phase space.

Lyapunov Exponents: The rate at which nearby orbits diverge from each other after small perturbations when the evolution of a nonlinear system is chaotic. There are M exponents for evolution of an M-dimensional flow or map. For a dissipative system, the sum of these exponents is negative. For a flow, one exponent must be zero.

Phase Space or State Space: Nonlinear systems are described by multidimensional vectors labeled by continuous or discrete time. The space in which these vectors lie is called state space or phase space. The dimension of phase space or state space is integer.

Works Cited

Abarbanel, H. D. I., Brown, R., Sidorowich, J. J. ("Sid"), Tsimring, Lev Sh. (1993), "The Analysis of Observed Chaotic Data in Physical Systems," *Rev. Mod. Phys.* **65**, 1331–1392.

Abarbanel, H. D. I., Kennel, M. B. (1993), "Local False Nearest Neighbors and Dynamical Dimensions from Observed Chaotic Data," *Phys. Rev. E* **47**, 3057–3068.

Arnol'd, V. I., Avez, A. (1968), *Ergodic Problems of Classical Mechanics*, New York: Benjamin.

Brandstater, A., Swinney, H. L. (1987), "Strange Attractors in Weakly Turbulent Couette-Taylor Flow," *Phys. Rev. A* **35**, 2207.

Cantor, G. (1883), "Über Unendliche, Lineare Punktmannichfaltigkeiten," *Math. Ann.* **21**, 545–591; see also Hausdorff, F. (1919), "Dimension und Äusseres Mass," *Math. Ann.* **79**, 157–179.

Chandrasekhar, S. (1961), *Hydrodynamic and Hydromagnetic Stability*, Cambridge, UK: Cambridge University Press.

Collet, P., Eckmann, J.-P. (1980), *Iterated Maps on the Interval as Dynamical Systems*, Boston: Birkhauser.

Devaney, R. L. (1989), *An Introduction to Chaotic Dynamical Systems*, 2nd ed., Redwood City, CA: Addison-Wesley.

Drazin, P. G. (1993), *Nonlinear Systems*, Cambridge, U.K.: Cambridge University Press.

Eckmann, J.-P., Ruelle, D. (1985), "Ergodic Theory of Chaos and Strange Attractors," *Rev. Mod. Phys.* **57**, 617.

Fraser, A. M. (1989a), "Information and Entropy in Strange Attractors," *IEEE Trans. Info. Theory* **35**, 245.

Fraser, A. M. (1989b), *Physica D* **34**, 391.

Fraser, A. M., Swinney, H. L. (1986), *Phys. Rev. A* **33A**, 1134.

Gallager, R. G. (1968), *Information Theory and Reliable Communication*, New York: John Wiley and Sons.

Gollub, J. P., Benson, S. V. (1980), "Many Routes to Turbulent Convection," *J. Fluid Mech.* **100**, 49–470.

Grassberger, P. (1983), "Generalized Dimensions of Strange Attractors," *Phys. Lett. A* **97**, 227; see also Hentschel, H. G. E., Proccacia, I. (1983), "The Infinite Number of Generalized Dimensions of Fractals and Strange Attractors," *Physica D* **8**, 435; see also Renyi, A. (1970), *Probability Theory*, Amsterdam: North Holland.

Guckenheimer, J., Holmes, P. J. (1986), *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, 2nd ed., New York: Springer-Verlag.

Hénon, M. (1976), "A Two Dimensional Mapping with a Strange Attractor," *Commun. Math. Phys.* **50**, 69.

Kaplan, D. T., Glass, L. (1992), "Direct Test for Determinism in a Time Series," *Phys. Rev. Lett.* **68**, 427.

Kaplan, J. L., Yorke, J. A. (1979), "Chaotic Behavior in Multidimensional Difference Equations," *Lecture Notes in Mathematics* **730**, 228.

Kennel, M. B., Brown, R., Abarbanel, H. D. I. (1992), "Determining Minimum Embedding Dimension using a Geometrical Construction," *Phys. Rev. A* **45**, 3403–3411.

Lorenz, E. N. (1963), "Deterministic, Nonperiodic Flow," *J. Atmos. Sci.* **20**, 130–141.

Mañé, R. (1981), in: D. Rand, L. S. Young (Eds.), *Dynamical Systems and Turbulence, Warwick, 1980*, Lecture Notes in Mathematics Vol. 898, Berlin: Springer, p. 230.

Newton, I. (1687), *Philosophia Naturalis Prin-*

cipia Mathematica, Imprimatur S. Pepys, Reg. soc. prases. Julii 5. 1686. Londini, Jussu Sociaetatis Regia ac Typis Josephi Streater. Prosat apud plures Bibliopolas. Anno, 1687.

Oseledec, V. I. (1968), "A Multiplicative Ergodic Theorem. Lyapunov Characteristic Numbers for Dynamical Systems," *Trudy Mosk. Mat. Obsc.* **19**, 197; *Moscow Math. Soc.* **19**, 197.

Ott, E., Grebogi, C. Yorke, J. A. (1990), "Controlling Chaos," *Phys. Rev. Lett.* **64**, 1196–1199.

Poincaré, H. (1892), *Les Méthodes Nouvelles de Méchanique Celeste*, Paris: Gauthier-Villars.

Ruelle, D., Takens, F. (1971), "On the Nature of Turbulence," *Commun. Math. Phys.* **20**, 167; see also Newhouse, S., Ruelle, D., Takens, F. (1978), "Occurrence of Strange Axiom A Attractors Near Quasiperiodic Flows on T^m ($m = 3$ or more)," *Commun. Math. Phys.* **64**, 35.

Shannon, C. E. (1948), "A Mathematical Theory of Communication," *Bell Syst. Tech. J.* **27**, 379–423 [Part I], 623–656 [Part II].

Smale, S. (1967), "Differentiable Dynamical Systems," *Bull. Am. Math. Soc.* **73**, 747.

Takens, F. (1981), in: D. Rand, L. S. Young (Eds.), *Dynamical Systems and Turbulence, Warwick, 1980*, Lecture Notes in Mathematics Vol. 898, Berlin: Springer, p. 366.

VanderPol, B. (1992), "On Oscillation Hysteresis in a Simple Triode Generator," *Philos. Mag.* **43**, 700–719.

Whittaker, E. T. (1937), *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, 4th ed., London, New York: Cambridge University Press.

Wisdom, J. (1987), "Chaotic Dynamics in the Solar System," *Icarus* **72**, 241–275.

Further Reading

Abarbanel, H. D. I., Rabinovich, M. I., Sushchik, M. M. (1993), *Introduction to Nonlinear Dynamics for Physicists*, Singapore: World Scientific.

Arrowsmith, D. K., Place, C. M. (1990), *An Introduction to Dynamical Systems*, Cambridge, U.K.: Cambridge University Press.

Baker, G. L., Gollub, J. P. (1990), *Chaotic Dynamics*, Cambridge, U.K.: Cambridge University Press.

Lichtenberg, A. J., Lieberman, M. A. (1983), *Regular and Stochastic Dynamics*, Berlin: Springer-Verlag.

Moon, F. C. (1992), *Chaotic and Fractal Dynamics*, New York: John Wiley and Sons.

Ott, E. (1993), *Chaos in Dynamical Systems*, Cambridge, U.K.: Cambridge University Press.

Ottino, J. M. (1989), *The Kinematics of Mixing: Stretching, Chaos, and Transport*, Cambridge, U.K.: Cambridge University Press.

Shinbrot, T., Grebogi, C., Ott, E., Yorke, J. A. (1993), *Nature* **363**, 411–417.

Thompson, J. M. T., Stewart, H. B. (1986), *Nonlinear Dynamics and Chaos*, New York: John Wiley and Sons.

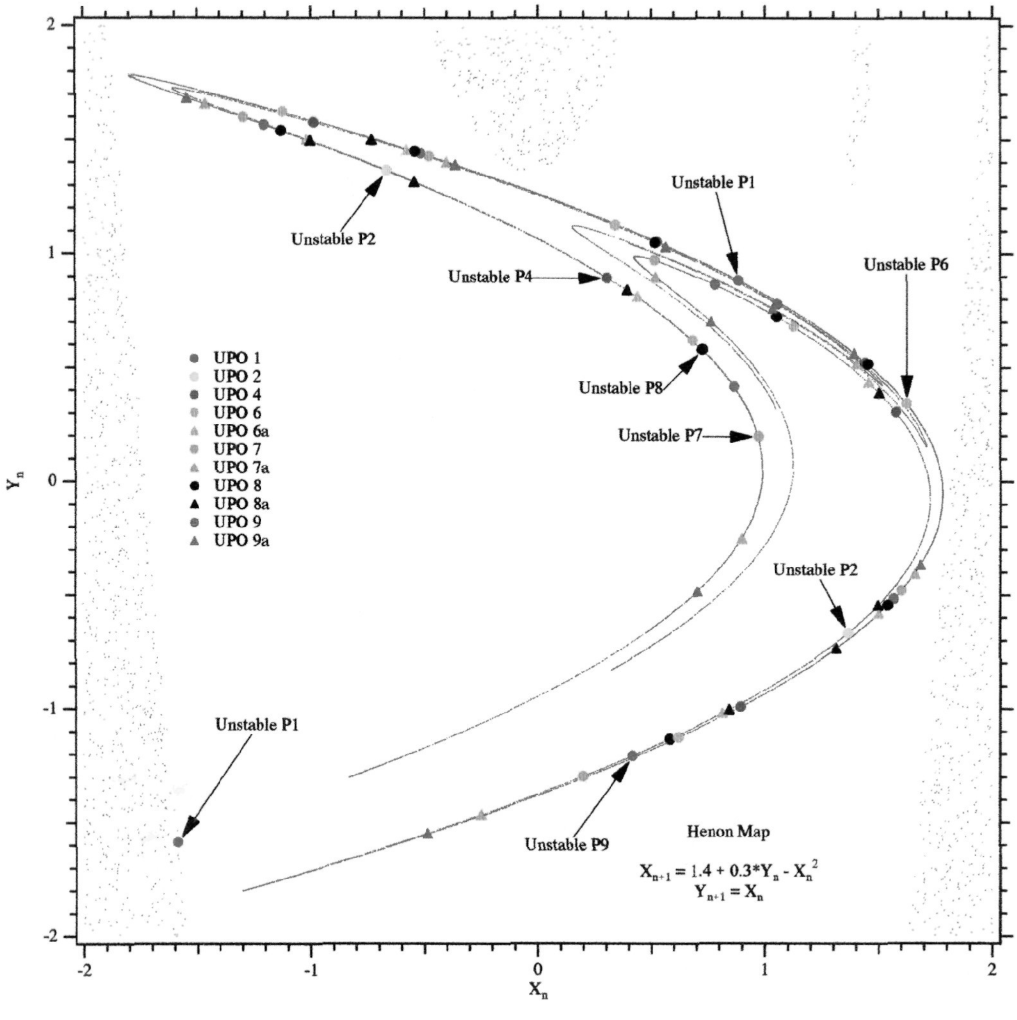

Vol. 11 – PLATE 1. The unstable periodic orbits in the Hénon map of the plane to itself for periods up to period nine. The orbits are labeled as P*N* for period *N*. The designation UPO stands for Unstable Periodic Orbit.

NUCLEAR ENERGY, FISSION

Walter Y. Kato, *Brookhaven National Laboratory, Upton, New York, U.S.A.*

Introduction 441
1. **Reactor Types** 444
2. **Nuclear Physics** 445
3. **Reactor Design** 449
4. **Power Reactors** 455
5. **Reactor Safety** 459
6. **Advanced Water Reactors** ... 467
7. **Economic Considerations** 470
8. **Radioactive Waste Management** 471
9. **Nuclear Regulation** 472
9.1 USNRC 472
9.2 IAEA 474
10. **Conclusion** 475
Works Cited 475
Further Reading 476

INTRODUCTION

2 December 1992 marked the 50th anniversary of the initiation of the first self-sustaining nuclear chain reaction of a manmade nuclear reactor in the world. That reactor, called Chicago Pile No. 1 (CP-1), was constructed in secrecy by the Manhattan Project at the beginning of World War II underneath the west stands of the University of Chicago squash court by a team of physicists under the leadership of Enrico Fermi (Smyth, 1945; Rhodes, 1986). CP-1 was composed of graphite blocks of varying lengths and pressed natural uranium oxide and uranium metal cylinders. It was designed and constructed to prove that a self-sustaining fission chain reaction could exist and to confirm theoretical physics calculations. The success of the self-sustaining chain reaction reactor in 1942 led to the design and construction of a large, graphite-moderated, natural-uranium–fueled, water-cooled plutonium production reactor at Hanford, Washington.

The plutonium production reactor used the nuclear reaction

$$^{238}\text{U} + \text{neutron} \rightarrow {}^{239}\text{U} \xrightarrow{\beta^-} {}^{239}\text{Pu}$$

for the production of plutonium for nuclear weapons. The chemical separation of plutonium from irradiated natural uranium appeared to be a much easier process at that time than the isotopic separation of the ^{235}U isotope from natural uranium. The Manhattan Project also initiated in 1942 the design and construction of a large ^{235}U isotope separation plant based upon a diffusion process at Oak Ridge, Tennessee. This led in 1945 to the production of sufficient quantitites of highly enriched uranium containing ^{235}U for the construction of the first atomic weapon.

The Soviet Union also started on the construction of a graphite-moderated natural-uranium reactor under the leadership of I.V. Kurchatov in Moscow in 1943. This resulted in the Φ-1 reactor achieving a self-sustaining chain reaction on 25 December 1946 at what is now known as the I.V. Kurchatov Institute of Atomic Energy in Moscow. Following attainment of the self-sustaining chain reaction in the Φ-1 reactor, the Soviet scientists and engineers designed and constructed graphite-moderated, water-cooled, natural-uranium–fueled reactors for the production of plutonium in the Urals region of the Soviet Union by 1948.

Although the first nuclear reactors were constructed for plutonium production for nuclear weapons, it was recognized early that the energy released from ^{235}U fission could be used for the production of electrical power, motive power for ships, and district heating. With the end of World War II, large quantities of separated ^{235}U isotopes, which could be used to fuel nuclear power reactors, became available from the separation plants at Oak Ridge.

Nuclear reactors to produce self-sustaining chain reactions utilize the following

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

nuclear reactions:

$$^{235}\text{U} + n \rightarrow \text{fission products} + \text{kinetic energy} + 2.42n$$

or

$$^{239}\text{Pu} + n \rightarrow \text{fission products} + \text{kinetic energy} + 2.87n.$$

The first nuclear reactor that produced electricity was Argonne National Laboratory's Experimental Breeder Reactor No. 1 (EBR-1) (Lichtenberger *et al.*, 1955), which was a 1.2 MW(t) [200 kW(e)] sodium-potassium–cooled fast reactor located at the National Reactor Testing Station (NRTS), which is now known as the Idaho National Engineering Laboratory (INEL), located about 60 miles west of Idaho Falls, Idaho. EBR-1 first produced electricity on 20 December 1951.

In the middle 1940s, the U.S. Navy under the leadership of Adm. Hyman Rickover became interested in developing nuclear power for powering submarines. This led to the establishment of the Knolls Atomic Power Laboratory (KAPL) near Schnectady, New York, and the Bettis Atomic Power Laboratory (BAPL) near Pittsburgh, Pennsylvania, for the development of a nuclear power system for submarines. The General Electric Company (GE) operated KAPL and the Westinghouse Electric Corporation operated BAPL for the U.S. Atomic Energy Commission. BAPL developed the concept of a pressurized, water-cooled and -moderated nuclear reactor, while KAPL developed the concept of a liquid-metal–cooled reactor for motive power for the submarines. The BAPL water-moderated and -cooled reactor became the forerunner of the commercial pressurized-water reactors (PWR) that were later manufactured by Westinghouse, Combustion Engineering, and Babcock & Wilcox corporations.

Argonne National Laboratory (ANL) originated the concept of the boiling-water reactors (BWR). The concept of the BWR was demonstrated by the BORAX [15 MW(t)] Reactor series of experiments (Dietrich *et al.*, 1955) at the NRTS begining around 1952, which led to the construction of the Experimental Boiling Water Reactor (EBWR) (Harrer *et al.*, 1955), a 20 MW(t) [5 MW(e)] prototype reactor at the ANL site near Chicago in 1961. The EBWR and the Vallecitos BWR (VBWR), later designed and constructed by GE in Vallecitos, California, were the forerunners of the large commercial BWRs.

Although fission reactors essentially produce heat energy, it is important to note the significant differences between fission reactors and chemical-combustion, fossil-fueled plants for electricity generation. Fission reactors produce heat through the fission process as a result of a nuclear interaction between a neutron and the fissile isotope nucleus. The fissioning of ^{235}U results in the energy release of 1 MW-d (24 000 kW h) per gram of ^{235}U. On the other hand, a chemical combustion (oxidation) process of methane gas, for example, is a result of the oxygen molecule combining chemically with carbon molecules, resulting in the heat energy release of about 1.6 kW h per gram of methane (CH_4) gas. This has the following consequences:

1. Fission reactors, in addition to producing heat energy, produce neutron and gamma radiation that must be contained by shielding.
2. Nuclear reactors produce gaseous and particulate fission products, which are radioactive and decay with the emission of beta, gamma, or alpha rays on varying time scales. The decay rate of a radioactive material represents the activity of the material. The time it takes for the activity to decay by a factor of 2 is known as the half-life of the radioactive material. Radioactive materials must also be contained by shielding. They must be handled carefully and prevented from being released to the environment for long periods of time.
3. Nuclear reactors have the ability to produce greater than design-basis heat energy; therefore, they require well-designed and reliable control mechanisms.
4. Finally, nuclear reactors, because of their varying half-life fission products, produce significant amounts of heat energy called decay heat even though the fission process has been terminated. This is in contrast to chemical-combustion plants, which produce no heat once the combustion process is stopped. The amount of decay heat produced after termination of the fission process is illustrated in Table 1. The existence of decay heat results in the requirement that nuclear fuel elements must be cooled for months or years after

Table 1. Decay heat from fission reactors.

Time after reactor shutdown	Approximate percentage of full power production
1 sec	6
1 h	1
28 h	0.5
12 d	0.3
4 m	0.13

the termination of the fission process in order to avoid melting of the fuel and releasing of the fission products.

The ability of the ^{235}U isotope nucleus to capture a neutron and either be fissioned or become the ^{236}U isotope and emit a gamma ray depends on the probability of interaction between the neutron and target nuclei, which is measured by the *interaction cross section*. Various types of interactions are possible depending on the kinetic energy of the neutron. Three principal reactions for the ^{235}U isotope and neutrons are scattering, capture, and fission at low neutron energies. Neutrons have a kinetic velocity or energy measured in m/s or eV. As an example, neutrons at equilibrium at room temperature have a velocity of 2200 m/s or an energy of 0.025 eV. Cross sections, which are conventionally designated by the symbol σ, are measured in units of cm^2 as a function of the kinetic energy of the neutron. A cross section of 10^{-24} cm^2 is called a cross section of 1 barn. The capture (σ_c) and fission (σ_f) cross sections for ^{235}U at 0.025 eV are 98 and 582 barns, respectively. Isotopes such as ^{233}U and ^{239}Pu, also, readily fission at low neutron energies since they have fission (σ_f) cross sections of 527 and 742 barns, respectively, at 0.025 eV, whereas ^{238}U fissions with neutrons having energies greater than about 0.68 MeV.

The fissioning of the nucleus of a fissionable isotope such as ^{235}U produces on the average two fragments called fission products plus on the average 2.44 neutrons. Fission products are isotopes having atomic mass about half of the uranium atomic mass and are usually unstable, which means that they are radioactive and decay by emitting a beta particles, gamma rays, or possibly alpha particles. The neutrons emitted at fission have an average energy of 1.98 MeV, and these neutrons could cause additional fissions if another ^{235}U isotope is in proximity. This could create a self-sustaining chain reaction. Unfortunately, natural uranium is a mixture of the two isotopes ^{235}U (0.72%) and ^{238}U (99.28%), which have fission cross sections of about 4.2 and ~0.6 barns at 0.025 eV and 2 MeV, respectively, and capture cross sections of about 2.7 and 0.4 barns at 0.025 eV and 2 MeV, respectively. The significance of these values is that it is not possible to have a self-sustaining chain reaction with natural uranium alone because the capture cross section at 2 MeV is not significantly less than the fission cross section, which means that most of the neutrons would be captured or escape out of the system without causing additional fissions.

In order to produce a self sustaining chain reaction, it is necessary either to separate the fissionable ^{235}U isotope from natural uranium and use uranium enriched in ^{235}U or to decrease the energy of the fission neutrons by neutron moderation. The original nuclear reactor CP-1 used natural uranium as fuel and pure graphite as moderator. The graphite, having a low capture cross section and low atomic mass, would slow the neutrons from 2 MeV to 0.025 eV without being captured by the graphite or escaping out of the reactor system. Neutrons, when slowed down, would be captured by the ^{235}U nucleus and cause fission. The trick in making a self-sustaining chain reaction is to minimize leakage and captures by the moderator or structural materials while maximizing fissions by the ^{235}U nucleus. It is necessary to use high-purity graphite because industrial graphite contains small amounts of such materials as boron and vanadium, which have high neutron absorption cross sections. Other materials, in addition to high-purity graphite, that could act as moderators are light water, heavy water (D_2O), and beryllium because of their low atomic mass numbers. Unfortunately, hydrogen has a fairly high capture cross section, and therefore it is not suitable as a moderator with natural uranium. Heavy water can be used as a moderator with natural uranium, but it requires considerable effort to separate heavy water from ordinary water. If natural uranium could have its ^{235}U content increased to between 2.5 and 3.2%, then you could have a self-sustaining chain reaction using light water as moderator.

Today, natural uranium consists principally of the two isotopes ^{238}U (99.2%) and ^{235}U (0.72%) with half-lives of 4.47×10^9 yr and

7.04×10^8 yr, respectively. Natural uranium had a higher content of ^{235}U (i.e., 1.6%) a billon (10^9) years ago because of the difference in the ^{238}U and ^{235}U half-lives. Two billion years ago the ^{235}U content in natural uranium was closer to 3%. Under proper conditions, the natural uranium of two billion years ago could have produced a self-sustaining chain reaction with light water as moderator.

In May 1972, H. Bouzigues, a staff member of the French nuclear fuel processing plant at Pierrelatte, discovered that some of the natural uranium coming from the Oklo region of the Republic of Gabon in West Africa had a ^{235}U content of 0.7171% instead of 0.7202% found in normal natural uranium (Cowan, 1976). Other samples contained only 0.44% of U^{235}. A serious investigation into this discrepancy by the French Commissariat à l'Energie Atomique (CEA) resulted in the discovery that nature had constructed the world's first nuclear fission reactor possibly 1.7 to 1.9 billion years ago in a rich deposit of uranium ore in the Oklo region. That a self-sustaining fission reactor had existed was deduced from the other stable isotopes that were located near the region where the uranium depleted in ^{235}U was found. A study of the absolute amount of such rare earth elements as neodymium, lanthanum, and gadolinium and metals such as zirconium, ruthenium, and rhodium, and of their isotopic composition, led to the firm conclusion that a natural uranum fission reactor had existed there. The concentration of the uranium ore, the availability of water possibly in the form of water of crystalization in sedimentary ore or saturated water in the ground, and the lack of neutron absorbers, such as boron or lithium compounds, resulted in a self-sustaining chain reaction, which released an estimated 15 000 MW y of energy. Estimates of the duration of the fission reactors range from 150 000 to 1.5×10^6 years, $(1.7 \text{ to } 1.9) \times 10^9$ years ago. In the Oklo region, the French have found six reactor zones that were located in regions with exceptionally rich ores. In all six areas, the ^{235}U was found to be significantly depleted, leading to the conclusion that reactors had operated at more than one site. It is, thus, concluded that CP-1, the graphite natural uranium reactor, was the first manmade fission reactor but not the world's first fission reactor.

1. REACTOR TYPES

One approach to catagorizing reactors is by their function or use. Three principal categories are research, power generation, and material production. In the research category can be included zero-power or critical experiments, training reactors, and reactors that are operated for the production of neutrons. Examples of research reactors are the High Flux Beam Reactor (HFBR), as shown in Fig. 1 (Hendrie, 1965), at Brookhaven National Laboratory (BNL), the High Flux Isotope Reactor (HIFR) (Swartout *et al.*, 1965) at Oak Ridge National Laboratory (ORNL); and the Advanced Test Reactor (ATR) at the Idaho National Engineering Laboratory (INEL) (IAEA, 1989). These reactors have beam ports through which beams of neutrons can be removed for neutron scattering and activation experiments. They can also be used for radioisotope production. Research reactors are operated at power levels from a few watts or kilowatts to tens of megawatts. The heat that is produced by the fission process in these reactors is essentially waste heat and is disposed to the environment. Included in the research category are zero-power reactors or critical experiments, which are flexible-geometry, low-power reactors and operate at a few milliwatts to about 100 W for short periods of time to verify reactor physics calculations. CP-1, the first manmade reactor, was essentially a critical experiment, since it was operated at very low power levels and was used to verify reactor physics calculations. The objective of critical experiments is to measure such parameters as critical mass, neutron flux and power distribution, neutron energy spectrum, etc., while varying the fuel composition or geometry. Since handling of fuel is usually carried out by hand, it is necessary to minimize the total amount of fission energy produced by the reactor for any given run to minimize radiation levels from the fuel. An example of an operating critical experiment is the Zero Power Plutonium Reactor (ZPPR) (Davey, 1972), as shown in Fig. 2, operated by ANL at INEL since 1967.

Power generation reactors are those reactors that have been designed for the generation of electricity or process heat. These reactors operate at power levels of a few hundred to thousands of megawatts. Since the thermal efficiency of power reactors is about $\frac{1}{3}$, it requires about a 3000-MW power plant

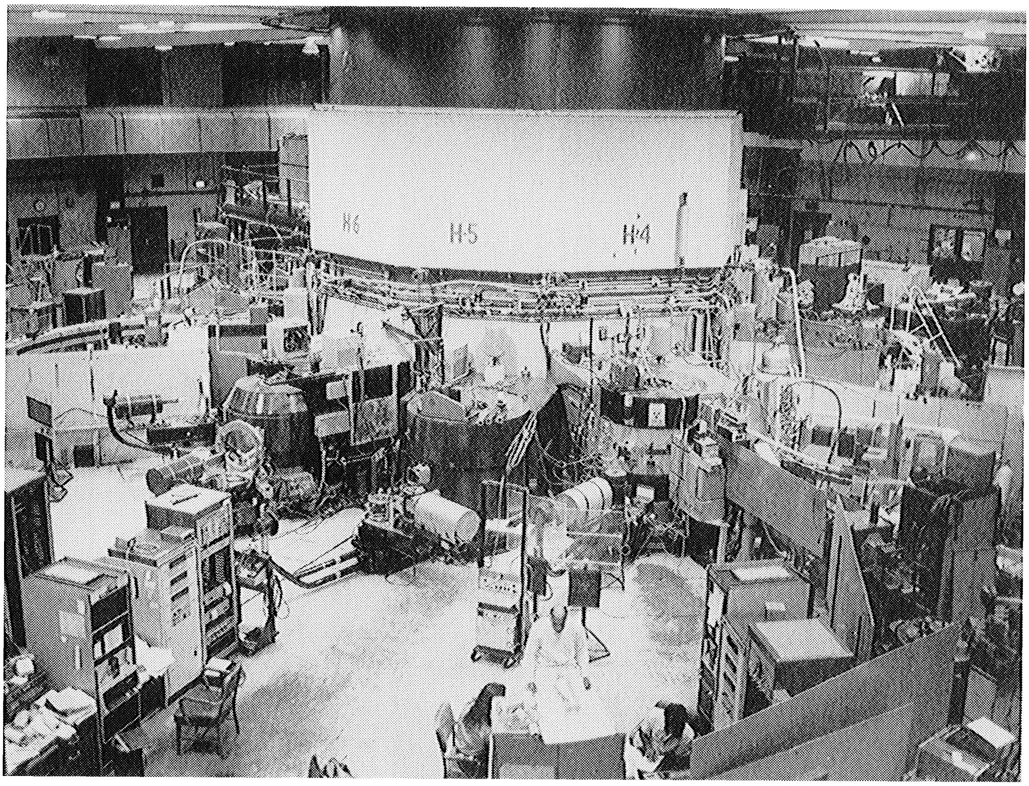

FIG. 1. Experimental area of BNL High Flux Beam Reactor (HFBR)

for the generation of 1000 MW of electrical power. There are a number of different types of power reactors, namely, light-water moderated and cooled reactors (LWRs), heavy-water moderated and cooled reactors (HWRs), heavy-water moderated light-water cooled reactors, graphite-moderated light-water cooled reactors, graphite-moderated gas-cooled reactors (HTGRs), and unmoderated sodium-cooled fast reactors (LMRs). Examples of these commercial power generation reactors will be given later.

A third category comprises the material production reactors. These reactors are designed for the production of large quantities of plutonium, tritium, or possibly ^{233}U isotopes for weapons use. These reactors are either graphite-moderated light-water cooled reactors, such as the production reactors at Hanford, Washington, or heavy-water moderated and cooled reactors, such as at the Savannah River Plant in South Carolina, and are operated at a few hundred to a few thousand megawatts.

2. NUCLEAR PHYSICS

An understanding of some nuclear physics is essential to the understanding of this brief exposition of nuclear reactor theory and operation. The reader is referred to such texts as Glasstone and Sesonske (1987), Lamarsh (1977), or Stamm'ler and Abbate (1983) for more detailed discussions on nuclear physics and nuclear reactor theory.

Neutrons and protons make up the the principal particles in atomic nuclei. Fission or breakup of heavy atomic nuclei, such as lead, tungsten, uranium, and heavier nuclei can occur by bombardment of these elements by particles such as neutrons, protons, alphas, or high-energy gamma rays. The higher the energy of the incident particle, the easier it is for fission to occur. Charged particles because of the Coulomb barrier must be accelerated to high energies to cause fission. Only neutrons, however, can cause fission at energies below 2 MeV and only in uranium or higher atomic mass nuclei. Neutrons can be

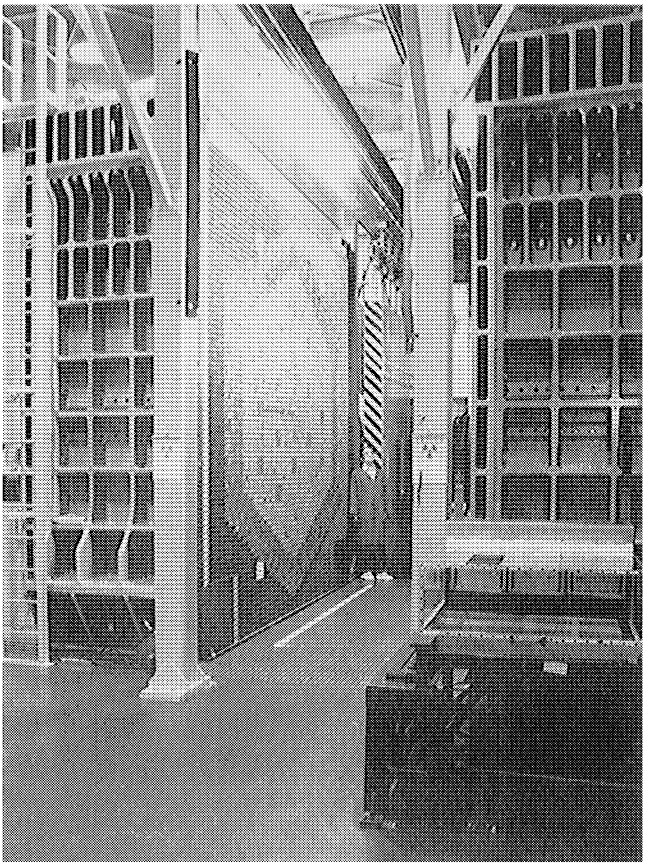

FIG. 2. Zero Power Plutonium Reactor (ZPPR) at ANL.

produced through such nuclear interactions as the (α, n) reaction using naturally occurring radioactive elements such as radium or polonium, which emit alpha particles during their decay, and elements such as beryllium. The incidence of alpha particles on beryllium nuclei results in the production of neutrons. Radium-beryllium neutron sources produce neutrons whose kinetic energy is about 4.8 MeV. With the availability of various isotopes of plutonium, plutonium-beryllium neutron sources have become available.

Although ^{235}U can fission at most neutron energies, it has a higher probability of causing fission when the incident neutrons have energies of about 0.025 eV. ^{238}U, on the other hand, will only fission with neutrons greater than 0.68 MeV. The probability of fission occurring in ^{238}U at energies greater than 0.68 MeV is considerably less than the fission of ^{235}U at 0.025 eV. The reason that the compound nucleus ^{236}U, which is formed by the absorption of a neutron by the ^{235}U nucleus, fissions is that the binding energy for the ^{236}U nucleus is less than the energy of the combined ^{235}U nucleus and neutron. The same is true for such isotopes as ^{233}U, ^{239}Pu, and ^{241}Pu, which are also fissile isotopes.

The breakup of the nuclei in the ^{235}U fission process produces on the average 2.44 neutrons and two nuclei, each having about half of the atomic mass of the original nuclei. The neutrons that are emitted at fission have a distribution of kinetic energy, as shown in Fig. 3, with an average value of 1.98 MeV. The technical problem in designing a self-sustaining reactor system is how to slow the neutrons to about 0.025 eV without the neutrons leaving the system or being absorbed by some other non-fissioning nuclei.

In addition to neutrons causing fission in ^{235}U or other isotopes, there are other reactions such as capture, elastic and inelastic scattering, and alpha, proton, or two-neutron production that can occur. Illustrations of absorption cross section values for various

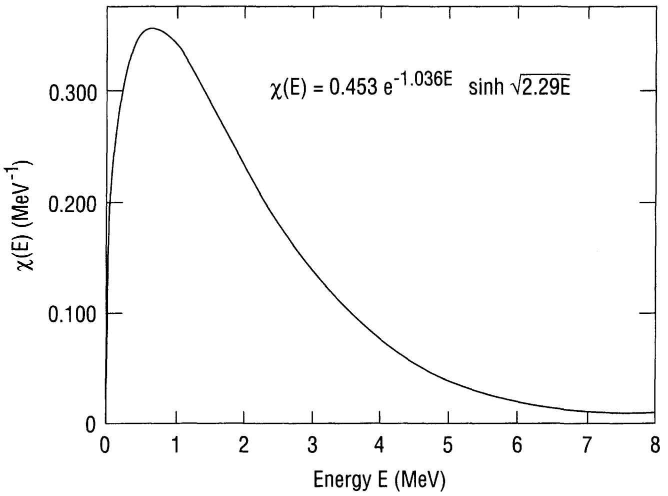

FIG. 3. Prompt neutron spectrum from fission.

isotopes at incident neutron energies of 0.025 eV are given in Table 2. The interaction of neutrons with different isotopic nuclei varies with the nuclei as well as with the energy of the incident neutrons.

Illustrations of the total neutron cross sections for ^{238}U and fission cross sections for ^{235}U nuclei are shown in Figs. 4 and 5, respectively. Typical neutron parameters for ^{235}U and ^{239}Pu are given in Table 3. Although ^{235}U fission has been used as an illustration in this discussion, isotopes such as ^{233}U and ^{239}Pu will also fission with neutrons at thermal energies. These isotopes do not exist in nature in any significant quantitites since they have shorter half-lives than ^{238}U and ^{235}U and have decayed since the beginning of the universe. Both ^{233}U and ^{239}Pu have been made in large quantities through the use of reactors.

Table 2. Thermal absorption cross sections of typical isotopes at $E = 0.025$ eV.

Isotope	Cross section (b)
^{1}H	332×10^{-3}
^{2}H	0.53×10^{-3}
^{6}Li	940
^{7}Li	37×10^{-3}
^{10}B	3837
^{11}B	5.5×10^{-3}
^{12}C	3.4×10^{-3}
^{54}Fe	2.25
^{197}Au	98.8
^{232}Th	7.4
^{235}U	680
^{238}U	2.7
^{239}Pu	1011

It can be seen from these illustrations that there are two components to the cross sections: a smoothly varying 1/v component and a resonance structure component. Single-level and multilevel resonance formulations based on the Breit–Wigner resonance formula have been developed to fit the resonance structure. The single-level Breit–Wigner resonance formula for neutron capture is

$$\sigma_c = \frac{\lambda_r^2 g}{4\pi} \frac{\Gamma_n \Gamma_\gamma}{(E - E_r)^2 + \Gamma^2/4},$$

where σ_c is the capture cross section of the resonance at energy E_r, λ_r is the wavelength of the neutron with energy E_r, g is a statistical factor, Γ_n and Γ_γ are the neutron width and radiation width, respectively, and Γ is the total width of the resonance at one-half its height.

Measurements, evaluations, and compilations of neutron cross sections for different reactions as functions of incident neutron energy for most isotopes have been carried out. In the United States, the National Nuclear Data Center (NNDC) at BNL is responsible for the compliation of neutron cross sections and resonance parameters for all isotopes. The NNDC has developed a CSISRS (Cross Section Information Storage and Retrieval System), which is an automated compilation of unevaluated experimental data. Originally, NNDC developed the Evaluated Nuclear Data Files (ENDF) A and B. ENDF/A contains both complete and incomplete data sets for many

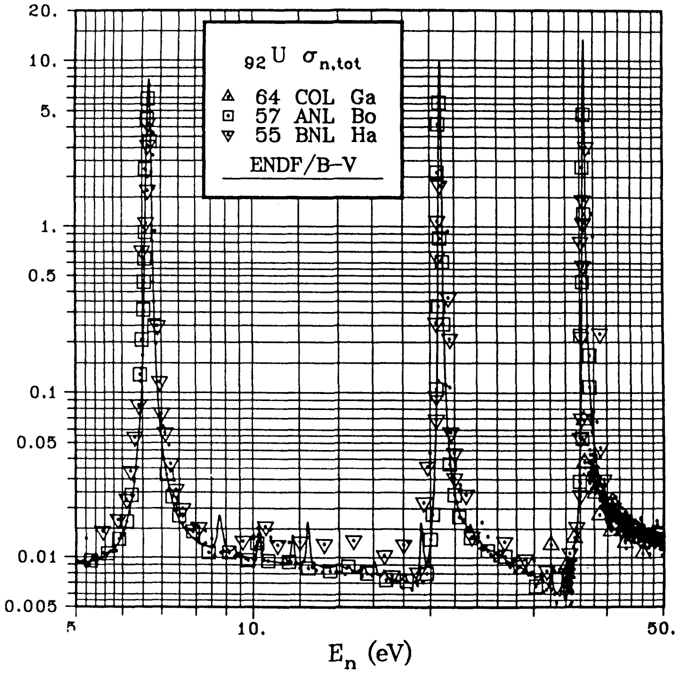

FIG. 4. Total cross section for ^{238}U

isotopes. ENDF/B, on the other hand, contains a unique complete cross section set needed for reactor design. Today, ENDF/B is the principal repository for evaluated neutron-induced–reaction data in the energy range from 10^{-5} eV to 20 MeV. The evaluation of the data for ENDF/B is carried out by the Cross Section Evaluation Working Group (CSEWG), a cooperative effort by technical experts from national laboratories,

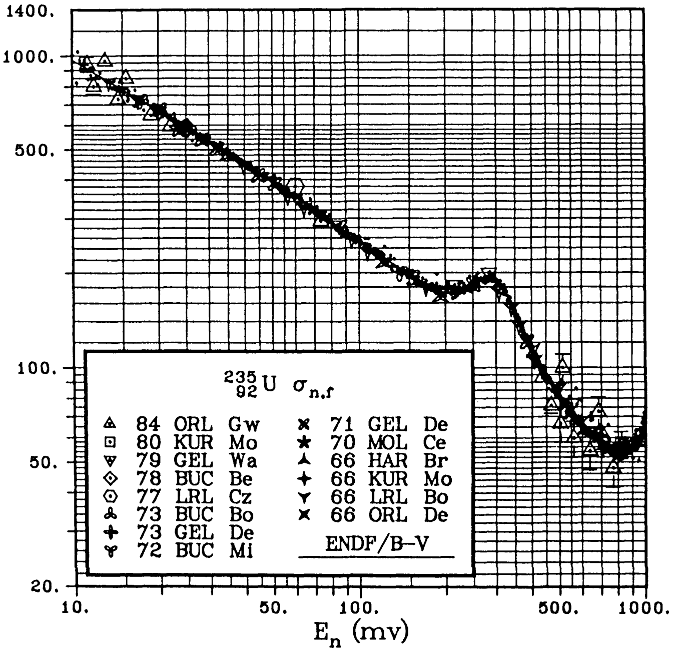

FIG. 5. Fission cross section for ^{235}U.

Table 3. Neutron parameters for ^{235}U and ^{239}Pu.

	Thermal $E = 0.025$ eV		Fast $E = 2$ MeV	
	^{235}U	^{239}Pu	^{235}U	^{239}Pu
Fission cross section (b)	582	742	2.7	2.2
Capture cross section (b)	98	269	1	0.71
No. of neutrons/fission	2.418	2.871	2.44	2.88
No. of neutrons/capture	2.068	2.108	2.21	2.60
Fraction of delayed *n*'s	0.0065	0.0021	0.0064	0.00206

industry, and academia. The ENDF/B file and other data sets, in addition to being accessible on-line via computer terminals have been published by the NNDC (Mughabghab *et al.*, 1981; McLane *et al.*, 1988). The ENDF/B neutron cross section files are used extensively in reactor physics calculations to calculate fundamental parameters, such as critical mass and power distributions, for reactors. Measured cross sections and evaluations are routinely exchanged with nuclear data compilation centers in Europe, the Former Soviet Union, Japan, and China. Periodically, conferences are held among technical experts in nuclear data to exchange data, methods for evaluation, and low-energy nuclear theory.

When a low-energy neutron interacts with a ^{235}U nucleus and causes fission, on the average, 2.44 neutrons are emitted instantaneously and a small number of neutrons are emitted a fraction of a second or seconds later from the decay of unstable fission-product nuclei. These delayed neutrons enable us to control the chain reactions in fission reactors more easily. The total energy released during fission is on the average about 200 MeV, which includes the contributions from gamma rays, kinetic energy of the fission products, and neutron kinetic energy. It is this energy that is utilized for the generation of heat, which is subsequently converted to electrical power.

3. REACTOR DESIGN

Since 2.44 neutrons are produced on the average whenever the ^{235}U nuclei fissions, it is possible to have a continuing chain of fission events if at least one of the neutrons produced at fission causes fission in another ^{235}U nucleus. The job of a reactor designer is to maximize the probability of neutrons born in fission to survive and cause fission in a ^{235}U or other fissile nucleus by

a. minimizing the probability of a neutron leaving the system;
b. minimizing a neutron from being captured by nonfissile nuclei;
c. maximizing a fission event occurring by increasing the number of ^{235}U or other fissile nuclei per unit volume; and
d. slowing down the average velocity of the neutrons from that at 1.98 MeV to 0.025 eV.

The designer does a trade-off study of these four approaches in designing a reactor. To minimize the probability of a neutron escaping the system, the designer can surround the reactor core, which contains the fissile and moderator materials, with a reflector consisting of material that has a large scattering cross section and a low absorption cross section. Materials such as water, graphite, beryllium, iron, or stainless steel have been used as reflectors. To minimize the probability of a neutron being captured by nonfissile nuclei, it is necessary for the designer to use low-absorption materials that have high strength at operating temperatures, such as aluminum or zirconium for structural and fuel cladding materials, to minimize the amount of nonfissile material utilized. To increase the number of fissile materials per unit volume, the designer would have to use as high a density as possible of fissile materials, generally metals or oxides, with uranium enriched in the ^{235}U isotope. Increasing uranium enrichment increases the cost of the material as well as fabrication costs because of production and materials accountability issues. Finally, the designer selects the operating neutron energy spectrum of the reactor, which determines the type of coolant and moderator used. For a thermal energy spectrum reactor, whose average neutron energy is around 0.025 eV, the designer has the choice of moderating materials such as water, heavy water, and graphite. If water or heavy water is chosen, these fluids can also act as coolants or heat removal fluids at the same time. If beryllium or graphite is chosen as moderator, gases such as helium or liquids such as water could be selected as the coolant.

Some of the essential parameters that a reactor designer must determine by analysis so that the reactor can be started up, maintained at steady-state operation over its lifetime, and shut down as required are listed here:

1. Criticality: A reactor is said to attain criticality when the chain reaction in the system is self-sustaining.
2. Critical mass: The amount of fuel required to maintain a steady-state or self-sustaining chain reaction.
3. Multiplication: The ratio of the number of neutrons in one generation over the number in the previous generation is defined as the multiplication of the system and is usually designated by k. When $k = 1$, the reactor is said to be critical. The ratio of the change in multiplication, due to effects such as fuel or control-rod movement, material density changes resulting from temperature change, or fissile material depletion or buildup, to the multiplication itself ($\Delta k/k$) is called the reactivity of the system and is a measure of the supercriticality or subcriticality of the system.
4. Power distribution: The spatial distribution of heat energy generation throughout the reactor core and reflector as a function of fuel burnup.
5. Control worth: The amount of neutron absorber materials, such as boron, cadmium, etc., in solid or in liquid solution required in the core to control the reactor. The absorber materials are often called poisons since they tend to terminate the chain reactions. Absorber material, such as soluble burnable poisons, is introduced into the coolant at the initial stages of a new core to balance the later buildup of fission product poisons.
6. Temperature effects on operation: The effect on reactivity due to changes in the system due to temperature effects is called temperature coefficient of reactivity. Temperature effects on the system can be due to thermal expansion of fuel and structural materials, density changes in materials such as fuel and coolant due to temperature, and changes in the width of neutron resonance cross sections for the fertile and fissile material resulting from the Doppler effect.
7. Fuel burnup: The destruction of the fissile isotopes due to fission is called burnup.
8. Fission product production, which introduces absorption into the reactor core.
9. Kinetic parameters, such as neutron lifetime and effective delayed neutron fraction, for prediction of the transient behavior of the reactor under normal and abnormal conditions.
10. Effects of radiation damage on reactor core materials.
11. Radiation levels from direct core radiation, and activation of coolants and coolant impurities.

To give the reader an idea of how a reactor is designed, a brief description of the thinking process that a designer undertakes in reactor design is presented. Assuming that only purified natural uranium metal is available to the designer, he would have to select a neutron-moderating material that would have the smallest capture cross section and at the same time would have the largest slowing-down potential per collision. The average logarithmic fractional loss in energy due to an elastic collision of the neutron and a nuclei is given by

$$\xi \equiv \ln \frac{E_1}{E_2} = 1 + \frac{\alpha}{1-\alpha} \ln \alpha,$$

where

$$\alpha \equiv \left(\frac{A-1}{A+1}\right)^2 .$$

Here, A is the atomic mass of the scattering nuclei, E_1 is the energy of the incident neutron, and E_2 is the energy of the neutron after scattering. Some of the scattering properties for the lighter elements are given in Table 4. It can be seen that the designer has choices among heavy water (D_2O), graphite, and beryllium for a nuclear reactor using natural uranium. Of these, purified graphite is the most plentiful, least expensive, and easiest to manufacture. D_2O is very plentiful; however,

Table 4. Scattering properties of various materials.

Element	Mass no.	Average change in lethargy	Average number of collisions to 1 eV from 2 MeV
H	1	1.0	15
D	2	0.726	20
Be	9	0.207	70
C	12	0.158	92
O	16	0.120	121
U	238	0.0083	1700

it would require an isotope-separation plant to separate the D_2O from H_2O since D_2O constitutes only 1.5% of ordinary water in nature.

Now that the designer has selected graphite as the moderator and natural uranium metal as the fuel, he must choose the shape and dimensions of the fuel. Depending upon the purpose of the reactor (i.e., research, electrical power generation, or material production), the designer must decide upon the total amount of power to be generated by the reactor and its power density. He must also decide the type of cladding that the uranium metal fuel will require. The cladding is to prevent the oxidation of the uranium as well as to contain the fission products and avoid possible chemical interactions of the uranium with the coolant. The choice of a cladding material is determined by the desired operating temperature and the neutronic and structural properties of the proposed cladding material. The cladding material must be able to maintain its strength at operating temperatures as well as be able to withstand radiation damage due to neutron interactions with the cladding material during the lifetime of the fuel in the reactor. The fuel cladding material and the structural material must be chemically compatible with the proposed coolant. In order to minimize parasitic neutron losses to the cladding material, it is desirable to use materials such as aluminum or zircaloy, a zirconium–aluminum alloy. High-strength materials such as stainless steel would not be acceptible for natural uranium reactors because of high neutron capture cross section unless the designer decided to use higher enriched uranium as fuel. Even zirconium would have to be highly purified in order to remove the small amount of hafnium that normally occurs with zirconium, since hafnium has very large absorption resonances.

Assume that the designer has selected air as the coolant and aluminum as the cladding material. He must decide upon the shape and dimensions of the fuel and the cladding thickness. Uranium metal must be cast and machined in an inert atmosphere, since uranium metal is pyrophoric and will oxidize very rapidly in air. The designer must also conduct many reactor physics analyses to determine the shape and amount of fuel that must be embedded in the graphite in order to make a self-sustaining chain reaction of fission events.

Earlier, a chain reaction was described as one initiated by a neutron from a neutron source or cosmic-ray event causing a chain reaction by producing a fission event in a ^{235}U nucleus, which in turn produced neutrons that again caused fission in another ^{235}U nucleus. The occurrence of many parallel chain reactions without dying out gives rise to a self-sustaining nuclear reactor.

The self-sustaining chain reactions may be described quantitatively by the multiplication factor, k. If k is greater than 1, the number of fissions is increasing from one generation to another, which would mean that the total number of fissions in the system or reactor is increasing, therefore, the total power (energy being generated) of the reactor is increasing. If k is less than 1, then the power of the reactor is decreasing and the chain reactions will be terminated and the reactor is said to be shut down. If k is exactly equal to 1, the reactor is said to be stable and at a constant power level. The operator of a reactor must, therefore, bring together enough fuel in the form of ^{235}U or other fissile material and moderator material and remove any absorbing materials that might be used as control rods in order to achieve a $k = 1$ critical system. In order to be able to control the reactor, the designer will have designed into the reactor control rods, which consist of removable neutron absorber material such as boron, or cadmium. The operator attains criticality ($k = 1$) by the removal of the control rods after the reactor has been loaded with fuel. After attaining criticality, the operator can reach an operating power level by removing additional control rods, making the reactor supercritical, or $k > 1$. As the total power generation of the reactor is increased, the operator must now introduce coolant into the reactor to remove the heat generated. The operator balances the heat-removal capability of the system to the heat generation of the reactor. When the heat removal from the reactor is balanced with the heat being generated, the operator then inserts control rods so that $k = 1$ and the reactor operates at constant power. To shut down the reactor, the operator inserts the control rods so that $k < 1$ and the chain reaction is terminated.

The multiplication factor and other reactor parameters, such as power distribution, neutron flux distribution, etc., of a system can be analytically calculated by using the Boltzmann neutron transport theory to describe the geometry and material composition of the system (Lamarsh, 1977; Bell and Glasstone, 1970). The simplest approach would be to use a one-dimensional, one–energy-group diffusion-theory approximation to the Boltzmann transport equations to calculate the k of a system. If fission neutrons were born at thermal energy instead of with a distribution with an average energy of 1.98 MeV, it would be possible to describe a reactor analytically with a one–energy-group diffusion theory, where the rate of change with time of number of neutrons in a system is equal to the production of neutrons minus the leakage and the absorption of neutrons in the system.

Unfortunately, since fission neutrons are born with a continuous energy distribution up to about 10 MeV with an average energy of 1.98 MeV, it is necessary to solve multiple–energy-group transport or multiple–energy-group diffusion-theory equations to determine the reactor-physics parameters. In addition to being multiple-group, these equations may have to be multidimensional in order to describe a complex geometry system. Such systems of multidimensional, multiple-group partial differential equations were initially solved using finite-difference methods using fast, large-memory, mainframe computers. Today, diffusion-theory equations are usually solved using various nodal methods, while transport-theory equations still use finite-difference methods. Examples of reactor physics computer codes that numerically solve the diffusion- or transport-theory equations are given in Table 5 (National Energy Software Center, 1991). As work stations have become more powerful with larger memories and with higher computational speeds, many reactor physics computer codes have been adapted for use on work stations.

Table 5. Examples of reactor analysis computer codes.[a]

Title	Description
	Cross section generation
NJOY	Nuclear data processing code system for producing pointwise and multigroup neutron and photon cross sections from ENDF/B data.
	Static design studies
TWODANT	One and two-dimensional, multigroup, discrete-ordinates transport theory code.
1DB	One-dimensional, multigroup diffusion theory code.
3DB	Three-dimensional, multigroup diffusion theory code.
	Space-time–dependent kinetics
DIF3D	Three-dimensional, multigroup diffusion-theory, time-dependent code.
	Monte Carlo
MCNP	Monte Carlo neutron and photon transport code.

[a] For additional reactor codes, see Radiation Shielding Information Center (1993) or Energy Science and Technology Software Center (1993).

In the diffusion-theory approximation, the neutron is considered as a function of space and energy only. Because there is an angular dependence of elastically and inelastically scattered neutrons, it is necessary to take into consideration this angular dependence for accurate analysis. Simplifying assumptions in the usual diffusion equations introduce inaccuracies near interfaces of the reactor core with vacuum or at core-reflector, core-control rod, or core-blanket interfaces. Interfaces of the core and strongly neutron-absorbing media and fuel and coolant cells are also difficult to model accurately. In such cases, differential transport (Boltzmann equation) theory models, which include higher order angular elastic scattering effects, are used.

In order to determine the k of a system, it is first necessary to develop averaged group cross sections for each energy group that is planned to be used for the diffusion- or transport-theory equations. The averaged group cross sections are developed by averaging the microscopic cross sections in the ENDF/B file over an energy spectrum in each energy group. Computer codes such as NJOY (RISC Software Directory, 1993) may be used to generate the group cross sections. Of course, increasing the number of energy groups increases the complexity of the computation and the computer time required for the solution of the problem. Following the generation of the group cross sections, reactor-physics computer codes, as shown in Table 5, are used to calculate the reactor-physics parameters needed in the design, such as the multiplication, power distribution, etc.

Neutron movement from production, through migration, to absorption or leakage from the system can be considered a stochastic process or a random-walk process with a probability distribution for each type of interaction, such as angular scattering, energy loss or gain, or fission neutron energy, in the system. A process of following the transport of neutrons stochastically in the reactor system has been developed and is called the Monte Carlo method. This method essentially follows the life of a single neutron until it is absorbed or leaks from the system. Neutrons that are produced as a result of fission from this neutron are also followed as a chain of events or reactions. It is necessary to follow the history of many hundreds of thousands of initiating neutrons in order to obtain a statistically meaningful value for the multiplication constant or other physics parameter. Computer codes such as MCNP (Briesmeister, 1986) have been developed using the Monte Carlo method to model complex-geometry reactor systems. In order to model reactor systems accurately, it is necessary that a sufficient number of neutron interaction histories be calculated. The Monte Carlo method is a large consumer of computing time on the fastest mainframe computer, so that obtaining a million histories to achieve a convergence of k to $\pm 0.1\%$ is very time-consuming and expensive. With the advent of parallel computers and fast and relatively inexpensive work stations, using the Monte Carlo method to solve the neutron-transport problem has become routinely possible. It is anticipated that, in the near future, a fast parallel computer having 1000 or more parallel computing units could produce 10^7 to 10^9 histories in a short time. The reactor designer will then be able to carry out more accurate design analysis of complex reactor systems less expensively.

The use of diffusion or transport equations, discussed above to describe a reactor, results in a description of the steady-state condition of the reactor. In the steady-state condition, the neutron flux is independent of time and only a function of position in the reactor. The total power generated in the reactor is determined by integrating over space and energy the neutron flux, the number of fissile atoms, and their fission cross section. The reactor can be critical at any power level from a few microwatts or milliwatts to many thousands of megawatts provided that the heat generated is removed from the reactor to avoid melting of the fuel. The power level of the reactor is controlled by inserting or removing neutron absorbers from the reactor. Decreasing the neutron absorbers in the reactor will increase the k_{eff} so that, when the multiplication becomes greater than 1, the neutron flux will start to increase exponentially. The fact that some of the neutrons produced at fission are delayed and are not emitted instantaneously allows the reactor power to be increased slowly under control. It would be difficult without delayed neutrons to control the reactor, because the reactor power would increase very quickly exponentially with a short reactor period when the multiplication became greater than one. The reactor period is defined as the time required for the power to increase by a factor e. The rate of increase of the neutron density is a function of the prompt-neutron lifetime of the system and the degree to which k_{eff} is greater than 1.0. The prompt-neutron lifetime, l_p, is the average time elapsed between the time when the neutron is born at fission and the time it is absorbed or lost out of the reactor. It can be considered as having two components: the time it takes for the neutron to be slowed down to thermal energy from fission energy and the time it takes for the neutron to diffuse and become absorbed or lost out of the reactor. The slowing-down time is much smaller than the diffusion time, so that the prompt-neutron lifetime may be equated to the diffusion time. The prompt-neutron lifetime for thermal reactors can vary from a few tenths of a millisecond to over a hundred milliseconds, depending on the moderator. The neutron lifetime for fast-spectrum reactors is in the range of a few tenths of microseconds to a few microseconds.

In a thermal reactor, where most of the fissions occurring in the reactor are from ^{235}U, the fraction of delayed neutrons is approximately 0.0065. When the k_{eff} is >1.0 but <1.0065, the reactor is said to be delayed critical and the reactor power will increase with a period of seconds to minutes. An illustration of how rapidly the reactor power would increase is shown in Table 6, where the reactor period is given as a function of neutron lifetime and the reactivity of the system. When k_{eff} is >1.0065, the reactor is said to be prompt critical and the reactor power will increase exponentially with a very short period, as can be seen in the last line of Table 6. The rate of increase of the reactor power above prompt critical is dependent on the degree of prompt criticality and the neutron lifetime of the system. The reactor that becomes prompt critical will increase its power level by many orders of magnitude in a very short time until other mechanisms, such as temperature effects or destruction of the core configuration, terminate the reactor power excursion.

The k_{eff} of the reactor at any given time is determined by the geometry and the materials in the reactor, as well as the temperatures of the fuel, moderator, and coolant of the reactor. The temperature changes the density of the materials as well as their position in the reactor and, hence, the k_{eff} of the system. An increase in temperature also changes the effective width and height of the resonance cross sections in such materials as ^{238}U and ^{235}U through the Doppler effect. This effect increases the number of resonance level neutrons absorbed in these materials. An increase in ^{238}U absorptions would tend to decrease the reactivity, while increasing the number of fissions in ^{235}U would tend to increase reactivity. This change in reactivity due to the broadening of the resonance levels by a change in fuel temperature is called the Doppler temperature coefficient. The reactivity effect due to the creation of voids in the coolant due to vaporization of the coolant is called the void coefficient of reactivity. In calculating changes in power level due to changes in reactivity, it is necessary to take into consideration the reactivity coefficients due to such effects as Doppler, void formation, expansion effects, and density changes.

The reactor designer must choose the appropriate geometries and materials for the reactor core so that it will have an overall negative power reactivity coefficient, so that when the power increases the reactivity of the reactor decreases. Usually, for thermal reactors, where the ^{235}U enrichment is only a few percent, the ^{238}U Doppler effect will provide a negative Doppler fuel-temperature coefficient. The coolant or moderator void coefficient could be positive or negative depending on the degree of moderation. In the western world, LWRs are designed with a negative void coefficient so that the reactor's reactivity will decrease with an increase in power due to decreasing moderator density and increasing fuel temperature. On the other hand, the Soviet-designed RBMK, a graphite-moderated, light-water–cooled reactor, had a positive void coefficient of reactivity, which led to the catastrophic Chernobyl-4 accident.

The study of the behavior of the time variation of the neutron flux when the multiplication is greater or less than unity is called reactor kinetics. To calculate accurately the neutron flux when the reactor becomes superdelayed critical or subdelayed critical, it is necessary to solve a time-dependent diffusion or transport set of equations. For simple geometry systems, it can be assumed that the spatial and energy distribution of the neutron flux does not change with time. This significantly eases the effort to calculate the neutron flux change with time. It is, however, necessary to take into consideration the fact that the rate of emission of delayed neu-

Table 6. Reactor period (s) for various values of reactivity for ^{235}U system.

Reactivity	Lifetime = 10^{-3} s	Lifetime = 10^{-4} s	Lifetime = 10^{-5} s
0.001	60	60	60
0.003	10	10	10
0.004	3.5	2.5	2.5
0.007	0.8	0.2	0.14
0.010	0.3	0.04	0.0003

trons falls into six different time-decay groups, each of which is characterized by a definite exponential decay rate. The average energy of the delayed neutrons emitted is also substantially lower than the average prompt neutron energy at fission. The variation of the neutron flux with time is calculated by solving the point reactor kinetics equation (Glasstone and Sesonske, 1987), which incorporates the effects of the six groups of delayed neutrons.

In complex-geometry reactor systems where there is asymmetry in reactivity changes due to nonuniform control rod movements or temperature effects, it may be necessary to use one-dimensional or multidimensional kinetics equations in order to analyze power transients accurately. Large neutron flux asymmetries due to nonuniformities in neutron absorber distribution or spectral changes due to temperature effects require the use of multidimensional and multiple–energy-group time-dependent solutions of the transport equations for transient analysis. The transient excursion of the Chernobyl-4 reactor is a good example of a transient that requires the solution of such multidimensional, time-dependent equations.

4. POWER REACTORS

The development of ^{235}U enrichment production plants during World War II made possible the development of large nuclear power plants for electricity generation in the 1950s and 1960s because of the availability of large quantities of enriched uranium. Although highly enriched uranium or 93% ^{235}U was used primarily for weapons and naval reactors, enriched uranium also became available for the development of research, test, and demonstration power plants. Eventually in the United States, the utility and reactor vendor industry settled down on the choice of either boiling-water or pressurized light-water reactors using 2–5% ^{235}U-enriched uranium oxide fuel. The light-water reactors have become the predominant type of reactors used for electric power generation in the world. In the late 1960s through the 1980s, GE was the BWR manufacturer in the United States, while Westinghouse, Combustion Engineering Co., and Babcock and Wilcox were the PWR manufacturers. Since the early 1980s, only Westinghouse, GE, and ABB (ASEA Brown Boveri) Combustion Engineering have remained in the reactor vendor business. Combustion Engineering was taken over by ABB, which formed ABB-CE.

In 1991, there were approximately 111 nuclear power plants licensed to operate in the United States with a total operating capacity of 101 GW(e), which resulted in about 22% of the total electricity being generated from nuclear plants. There are nuclear power plants operating in 32 countries worldwide, which supply more than 17% of the world's electricity. In 1992, there were at least 411 reactors with a net electrical power capacity of over 320 GW(e). In 1991, France generated 72.7% of its electricity from nuclear plants, as compared with Belgium with 59.3%, and Sweden with 51.6%. A list of nuclear capacity, in countries worldwide, is shown in Table 7.

As of 1992, there were 237 PWRs [207 GW(e)] and 88 BWRs [72 GW(e)] out of a total of 411 nuclear power plants in worldwide operation. In both categories, ordinary water is used as moderator and coolant. The fuel is uranium dioxide enriched in ^{235}U content to between 2.5 and 5%. In the BWR case, the coolant pressure and temperature are such that water is in a boiling state as it exits from the core region, and steam is directly passed into a turbine for electric power generation. A schematic diagram of a BWR is shown in Fig. 6. A detailed schematic of the reactor portion is shown in Fig. 7, followed by a schematic arrangement of the steam supply system in Fig. 8.

In the case of the PWR, high coolant pressures are maintained so that water does not reach the boiling temperature and the heated water is fed directly into a steam generator. The secondary side of the steam generator produces steam, which is then fed into a turbine for electricity generation. A schematic diagram of a PWR is shown in Fig. 9. A schematic diagram of the reactor coolant system for a four-loop PWR is shown in Fig. 10, while the details of the core of a PWR are shown in Fig. 11. A typical PWR fuel assembly is shown in Fig. 12, while a steam-generator cross section is shown in Fig 13.

The Soviet Union developed light water reactors designated VVER, which are similar to the western pressurized-water reactors.

Table 7. World commercial reactors in 1991.

Country	Number of operating reactors	Capacity [GW(e)]	Electricity generation % nucl., 1990
S. Africa	2	1.8	5.6
Argentina	2	0.69	19.8
Belgium	7	5.5	60.1
Brazil	1	0.62	1.0
Bulgaria	6	3.7	35.7
Canada	20	14.0	14.8
China	1	0.3	—
Czech and Slovak Republics	8	3.2	28.4
Finland	4	2.3	35.0
France	55	56.5	74.5
Germany	21	22.5	33.1
Hungary	4	1.7	51.4
India	9	1.8	2.2
Japan	43	33.2	27.1
S. Korea	9	7.22	49.1
Lithuania	2	2.7	55
Mexico	1	0.65	2.6
Netherlands	2	0.5	4.9
Pakistan	1	0.13	1.1
Russia	28	17.8	11
Slovenia	1	0.62	25
Spain	9	7.1	35.9
Sweden	12	10	45.9
Switzerland	6	4.88	42.6
Taiwan	6	5.14	34.6
Ukraine	14	12.1	25
United Kingdom	35	11.9	19.7
United States	108	98.2	20.6

Significant differences between the early VVERs and the western PWRs include the lack of a containment, the lack of multiple-loss-of coolant emergency cooling systems, and horizontal steam generators. Containment buildings are constructed around the newer VVERs.

Three other types of reactors have been built in western countries for electric power generation. These include the CANDU, a heavy-water–moderated, heavy-water–cooled tubular reactor designed and built by Atomic Energy of Canada, Ltd. (AECL), and a high-temperature gas-cooled reactor (HTGR) built by General Atomics Corporation of San Diego, California, which is shown schematically in Fig 14. Liquid-metal–cooled reactors (LMRs) (Waltar, 1981) have been built by ANL (EBR-I and EBR-II), GE (SEFOR), and Westinghouse (FFTF) (Peckinpaugh, 1982, 1986). A schematic view of a liquid-metal–cooled reactor (LMR) is shown in Fig 15.

The liquid-metal–cooled reactors designed and constructed by ANL, under the sponsorship of the Atomic Energy Commission, are based upon a metallic fuel element with a sodium-potassium (NaK) alloy as coolant for the EBR-I and sodium as coolant for the EBR-II. Both EBR-I (Lichtenberger *et al.*, 1955) and EBR-II (Lentz *et al.*, 1986) were built at the National Reactor Testing Station (NRTS) in Idaho. The EBR-I had a small compact core and produced 1.2 MW of heat energy. The NaK alloy conducted heat away from the reactor core to a steam generator, where the heat energy produced steam. A small turbine produced about 200 kW of electricity in EBR-I. EBR-I operated with neutrons whose spectrum had an average energy of about 100 keV as opposed to thermal reactors whose average neutron energy is 0.025 eV. In addition to demonstrating the possibility of electricity generation, EBR-I showed that it was possible to produce more plutonium than the amount of ^{235}U consumed. It was a converter of ^{238}U atoms into plutonium, ^{239}Pu, with a conversion ratio of a little greater than 1.0. The conversion ratio, or the breeding ratio, is defined as the ratio of the average of number of fissile atoms pro-

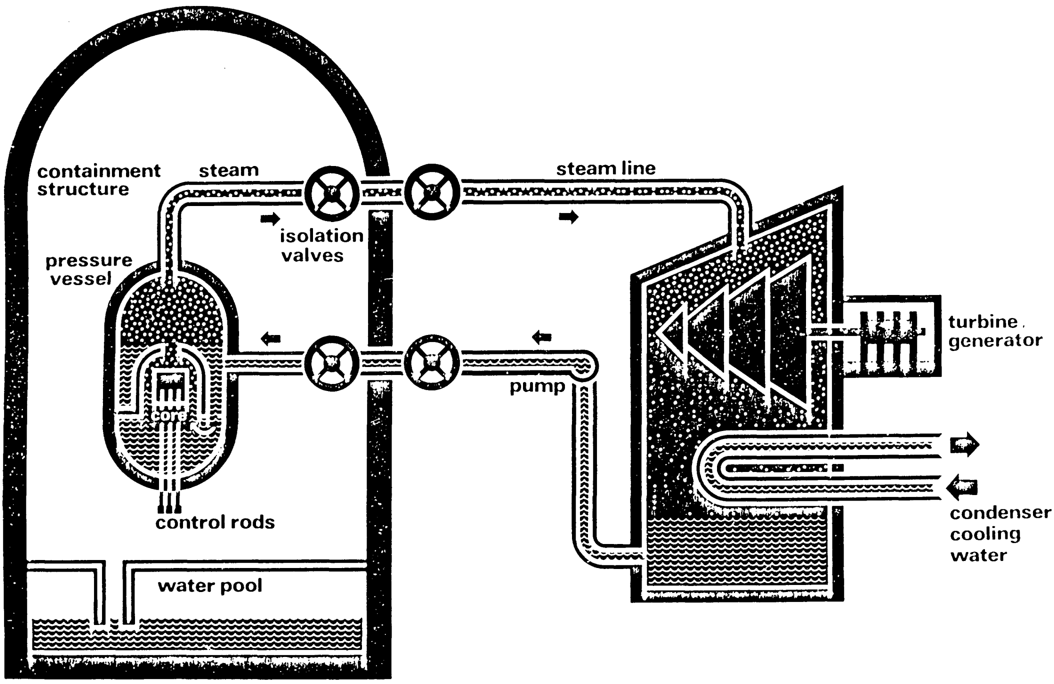

FIG. 6. Schematic diagram of BWR.

duced per fissile atom consumed in the reactor. When the new fissile atoms produced are different from the fissile atoms consumed, the reactor is said to be a converter, while if the new fissile atoms are the same as the fissile atoms consumed, the reactor is said to be a breeder reactor. Since ^{235}U constitutes only 0.72% of natural uranium, a breeder reactor could utilize the remaining ^{238}U to produce fissile atoms. During the early development of the nuclear industry, it had been forecast that there would be many 1000 MW(e) nuclear power reactors in operation. It was anticipated that these thermal reactors would consume the world's available ^{235}U in about half a century, indicating the necessity for near-term plutonium breeder reactors in order to maintain the nuclear option. Following the accident at the Three Mile Island–2 (TMI-2) nuclear power plant in April 1979, there was a considerable slowdown in the construction of power plants, because it was necessary for the existing plants and plants under construction to meet additional regulatory requirements. This gave rise to significant construction cost increases as well as operation and maintenance costs. As of 1993, there had been no new orders for nuclear power plants in the United States since 1978 because of a number of factors including no urgent need for additional generation capacity, the uncertainy in financing large capital projects, the uncertainty in regulations, and increased public opposition.

The FSU, in addition to the VVER series, which is similar to the western PWRs, built the RBMK type of power generation plant, which is a variation of their plutonium production reactor. It is a graphite-moderated, light-water–cooled, boiling-water, channel-type reactor. The Chernobyl-4 reactor, which exploded catastrophically in April 1986, is a typical RBMK-type reactor, which generated about 1000 MW(e). Instead of the reactor being contained in a thick pressure vessel, fuel elements were located in 1661 vertical high-pressure pipes or tubes. The reactor is designed so that water reaches saturation temperature about $\frac{1}{3}$ of the way up the channel with nucleate boiling in the remaining $\frac{2}{3}$ of the channel. The fuel was 2% enriched uranium oxide clad with zirconium containing 1% niobium. It was designed so that it could have on-line refueling, thus increasing its op-

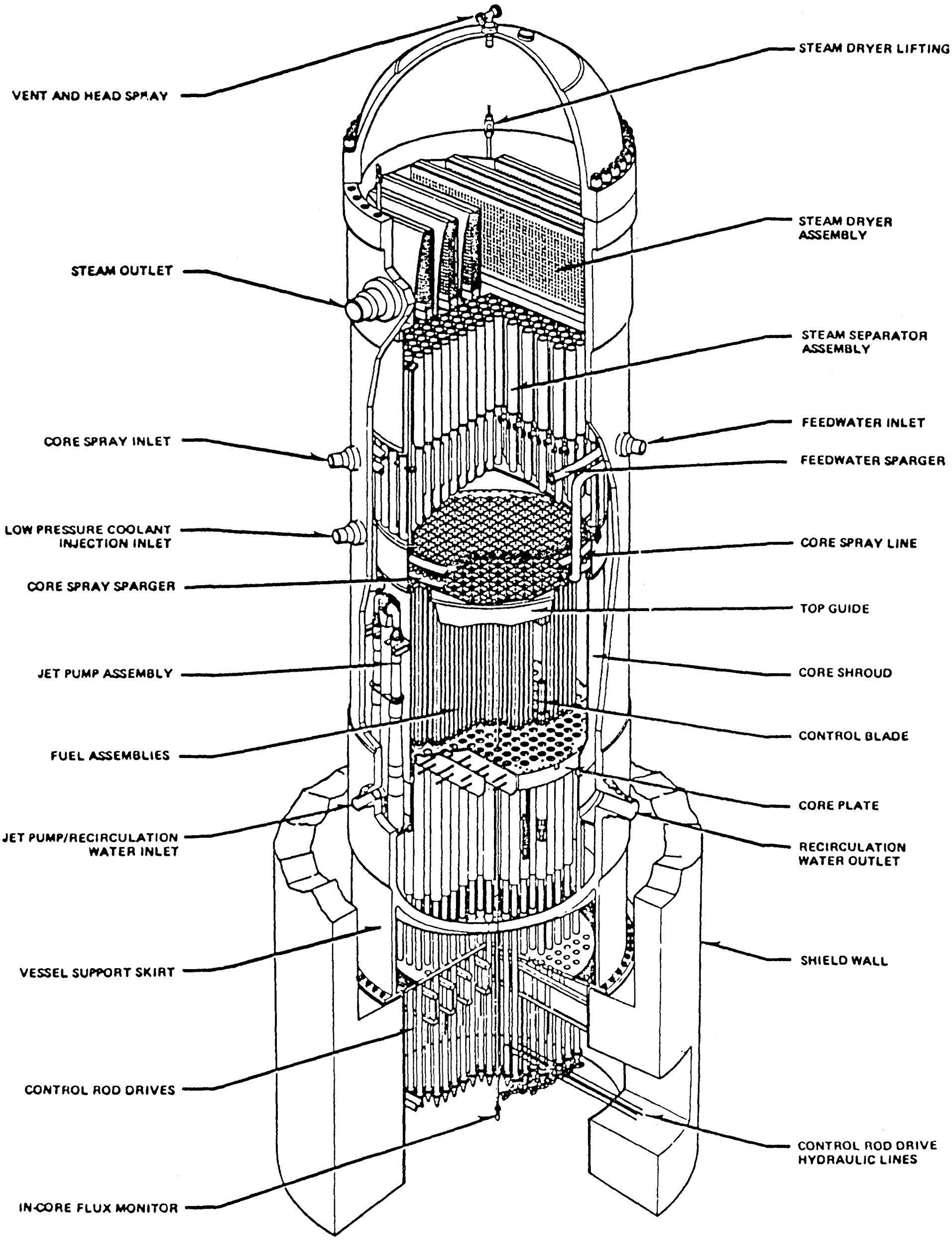
STEAM DRYER LIFTING LUG
VENT AND HEAD SPRAY
STEAM DRYER ASSEMBLY
STEAM OUTLET
STEAM SEPARATOR ASSEMBLY
FEEDWATER INLET
CORE SPRAY INLET
FEEDWATER SPARGER
LOW PRESSURE COOLANT INJECTION INLET
CORE SPRAY LINE
CORE SPRAY SPARGER
TOP GUIDE
JET PUMP ASSEMBLY
CORE SHROUD
CONTROL BLADE
FUEL ASSEMBLIES
CORE PLATE
JET PUMP/RECIRCULATION WATER INLET
RECIRCULATION WATER OUTLET
SHIELD WALL
VESSEL SUPPORT SKIRT
CONTROL ROD DRIVES
CONTROL ROD DRIVE HYDRAULIC LINES
IN-CORE FLUX MONITOR

FIG. 7. BWR reactor assembly.

erating capacity. Its inlet temperature was 270° C and average outlet temperature was 284° C, with an operating pressure of 986 psig. Although the reactor itself was contained in a steel compartment, the entire reactor did not have a containment building in the Western sense. The RBMK was also designed with a positive void coefficient of reactivity, as well as having a control-rod design such that, when the rod was being inserted into the core from its full out position, it would add reactivity instead of decreasing reactivity. Following the accident at Chernobyl, the RBMK reactors were modified so that the insertion of the

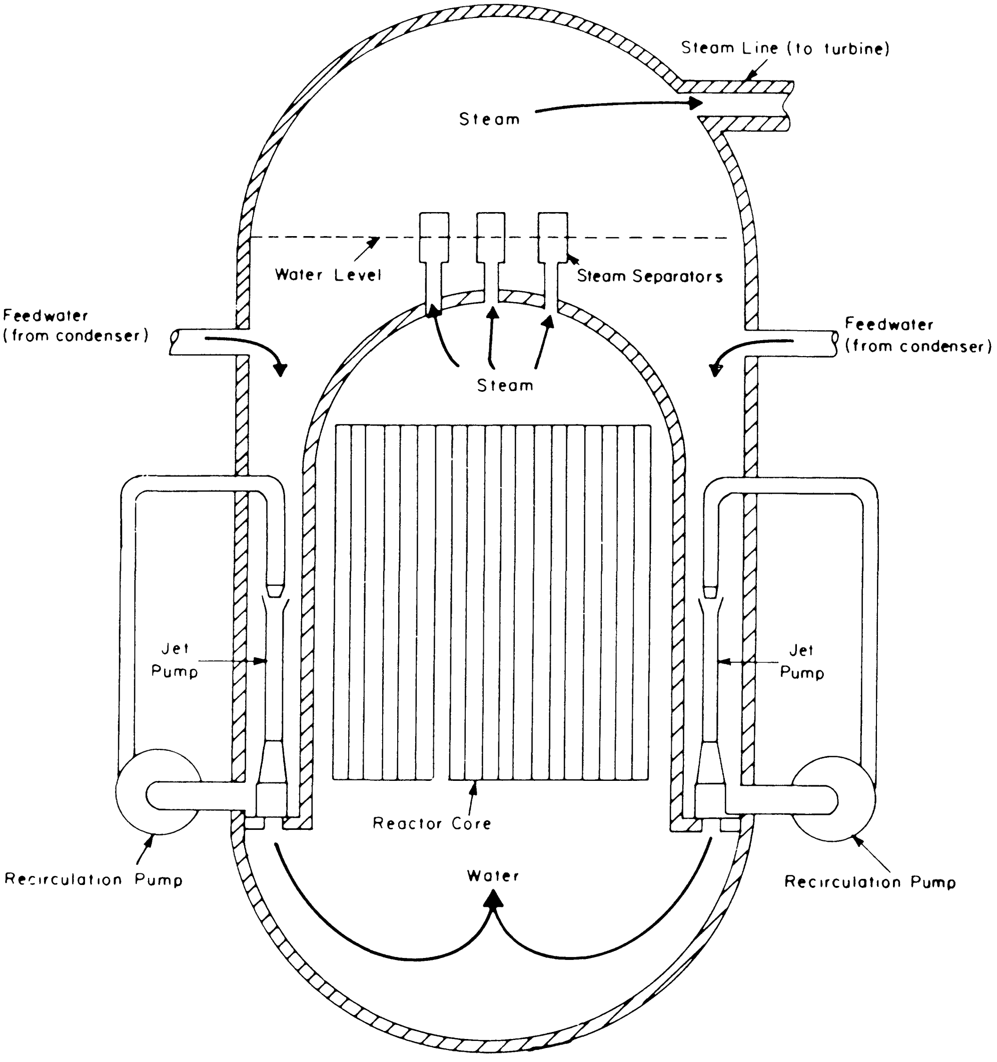
Steam Line (to turbine)
Steam
Water Level
Steam Separators
Feedwater (from condenser)
Feedwater (from condenser)
Steam
Jet Pump
Jet Pump
Reactor Core
Recirculation Pump
Water
Recirculation Pump

FIG. 8. Schematic diagram of BWR nuclear steam supply system.

control rods from the full out position does not initially add reactivity.

5. REACTOR SAFETY

Fissile isotope fuel in commercial nuclear power plants is consumed generating electricity in the range of 30 000 to 50 000 MW-d per metric ton of uranium. The energy produced by the fuel in MW-d per metric ton is a measure of the fuel consumed or burned. Since about a third of the fuel load is replaced per year after a year's operation, there is considerable buildup of fission products in the fuel at any given time. It is essential that these radioactive fission products be prevented from being released to the environment. This is accomplished by maintaining a defense-in-depth concept by, first, using an oxide fuel which has a high melting point. Second, the fuel has zircalloy cladding, which has high heat-transfer characteristics and high mechanical strength. Third, various engineered safeguards are built into the design of the plant to maintain core cooling under most accident conditions. Fourth, the reactor core is placed inside a high-strength pressure ves-

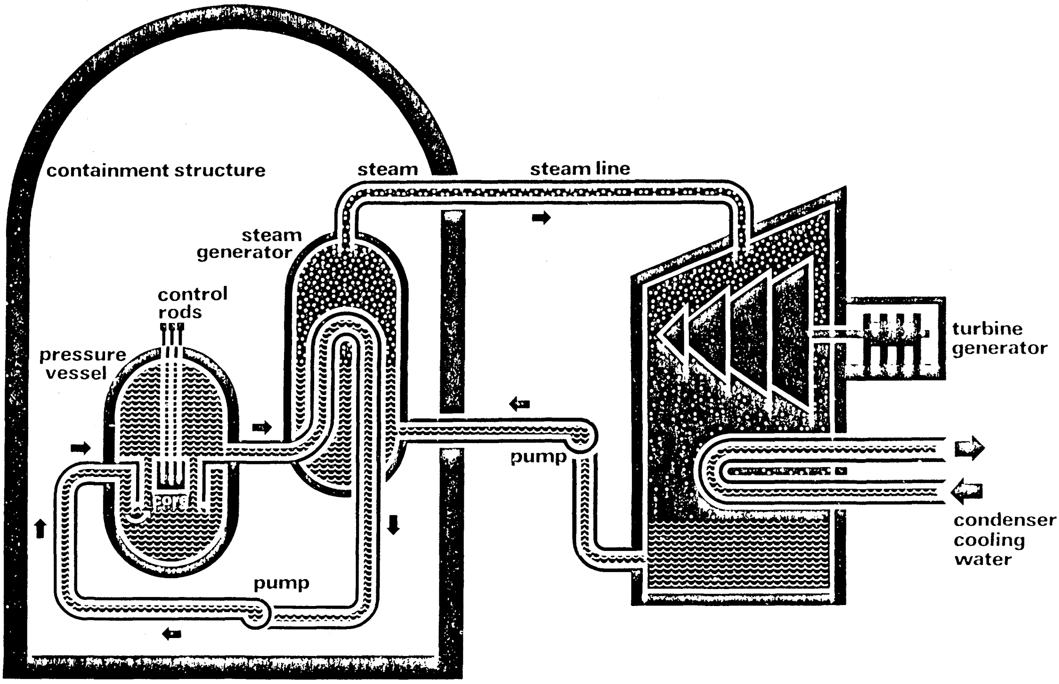

FIG. 9. Schematic diagram of PWR System.

sel and primary circuit. Fifth, the entire reactor and primary circuit are placed inside a steel and/or concrete containment building. Sixth, instrumentation and controls are built into the system so that automatic shutdown of the reactor and automatic operation of the engineered safeguards will take place without operator intervention. Highly trained operators, maintenance personnel, and supervisors complete the safeguards for the safe operation of a nuclear power plant. In addition, the Nuclear Regulatory Commission (NRC) has resident site inspectors at each nuclear power plant to audit the operations of the plant continuously.

It is of interest to consider the engineered safeguards that have been built into the power plant to maintain adequate cooling of the reactor core at all times. Maintaining core cooling of the reactor for a significant period of time after shutdown is important because of the need to remove the decay heat associated with fission product decay.

The engineered safeguards consist of instrumentation and control circuits, redundant heat-removal systems, emergency core-cooling systems, redundant emergency power sources, containment cooling systems, and strong steel and concrete containments. The instrumentation and control system is probably the most important engineered safeguard, since it is through this system that the reactor can be automatically shut down and the operators can detect abnormal situations and take corrective action. Since a reactor can generate more power than its design cooling capability, it is necessary to have instrumentation that continuously monitors the rate of power increase as well as controlling its operating power level. Usually, neutron and fission detectors are used to monitor the rate of change of power as well as the power levels. Neutron detectors are usually ion chambers filled with boron trifluoride gas and operated in an ion-current or pulse mode. Neutrons interacting with ^{10}B produce alpha particles that cause ionization in the ion chamber. Fission detectors are ion chambers filled with argon gas and internal plates or cylinders coated with fissile material. Neutrons causing fission in the ion chamber result in ionization of the gas in the chamber by the fission products. Again, the ionization current in the chamber, which is proportional to the fission

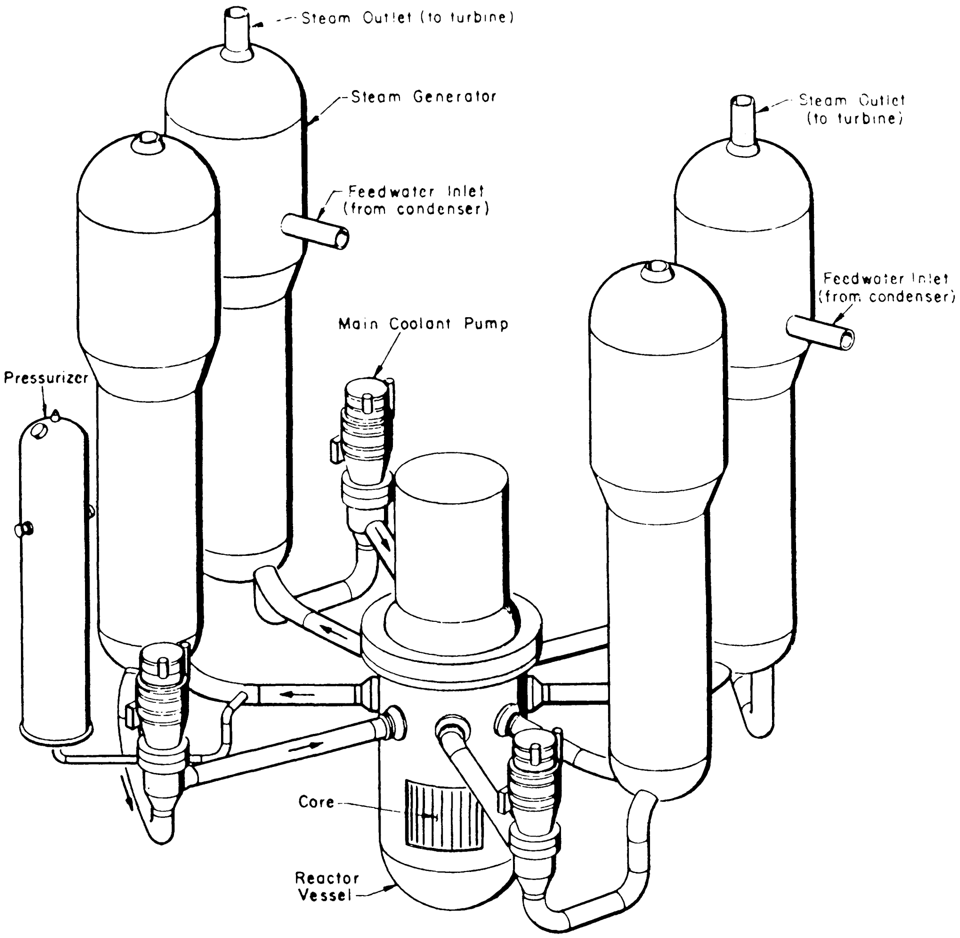
Steam Outlet (to turbine)
Steam Generator
Feedwater Inlet (from condenser)
Steam Outlet (to turbine)
Feedwater Inlet (from condenser)
Main Coolant Pump
Pressurizer
Core
Reactor Vessel

FIG. 10. Schematic diagram of PWR reactor-coolant system.

rate, is measured. In addition, instrumentation is available to monitor the distribution of power being generated in the core. Reactors have automatic power trip or "scram" circuits that will automatically cause insertion of the control and safety rods when the power level exceeds preset limits, or when the rate of increase of power exceeds preset limits. There are also temperature gauges that monitor coolant temperatures in various parts of the primary circuit and can also cause reactor scram when the values exceed preset limits. Radiation detectors throughout the containment building and turbine building monitor these areas for possible inadvertent release of radioactivity through leaks in piping coupled with defective fuel cladding or cladding rupture, which would allow fission products to escape from the fuel.

Since it is essential to maintain coolability of the reactor core by avoiding loss of circulation capability or loss of coolant, a number of redundant systems to maintain coolability have been installed in LWRs. The operation of valves and pumps requires electrical power; therefore, off-site power is brought into the plant by two redundant approaches. In addition, a battery system and emergency diesel generators are provided at the power station for operation in the event of loss of off-site power.

In the case of PWRs, which have primary and secondary coolant systems for heat removal, two or more auxiliary feed-water pumps and corresponding feed-water circuits and supply are provided to maintain adequate secondary water supply to the steam generators. During certain modes of normal

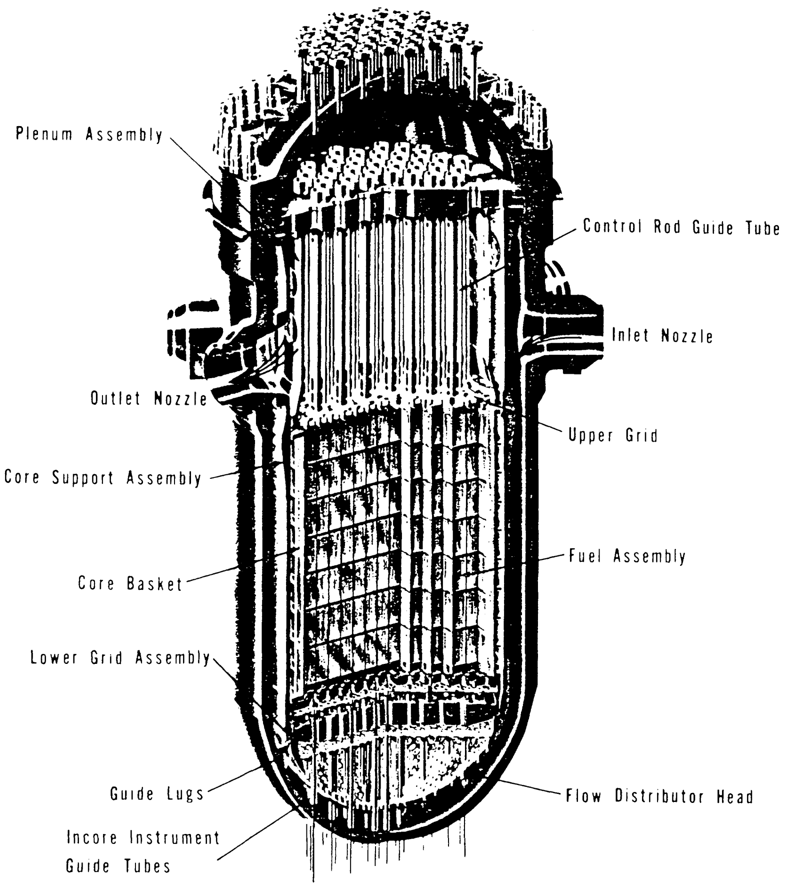
Plenum Assembly
Control Rod Guide Tube
Inlet Nozzle
Outlet Nozzle
Upper Grid
Core Support Assembly
Fuel Assembly
Core Basket
Lower Grid Assembly
Guide Lugs
Flow Distributor Head
Incore Instrument
Guide Tubes

FIG. 11. PWR reactor core assembly.

operation and in emergency situations, steam can be released to the atmosphere as a heat sink. Two separate emergency core-cooling systems (ECCS) are also provided that can introduce water into the primary circuit. One system is composed of several large-volume accumulators containing borated water, which will automatically supply water to the primary system when its pressure drops below preselected values. A second ECCS system is available, where a pump supplies borated water from a large-volume supply tank. The ECCS was designed to handle large pipe-break loss-of-coolant accidents (LOCA). Since the Three Mile Island Unit 2 (TMI-2) accident in April 1979, where the core was uncovered because of a stuck-open pressurizer valve resulting in a loss of primary coolant water, consideration is also given to the prevention of small-break LOCAs (SBLOCAs). In addition to these mechanisms for heat removal, there is a smaller capacity residual-heat–removal system that is available to the primary system for removing heat from the core after it has been shut down. Since large pumps require their own cooling water for operation, there are redundant service water systems for maintaining the operation of pumps. These varied and redundant safety systems have been introduced to be able to maintain core cooling under all circumstances.

Before a nuclear power station can be constructed and operated, the NRC requires for review and approval technical information about a proposed nuclear power plant, such as detailed descriptions of the site, design of the power plant, operation, maintenance, and management procedures; probabilistic safety assessment; accident analysis, including deterministic normal and abnormal transients;

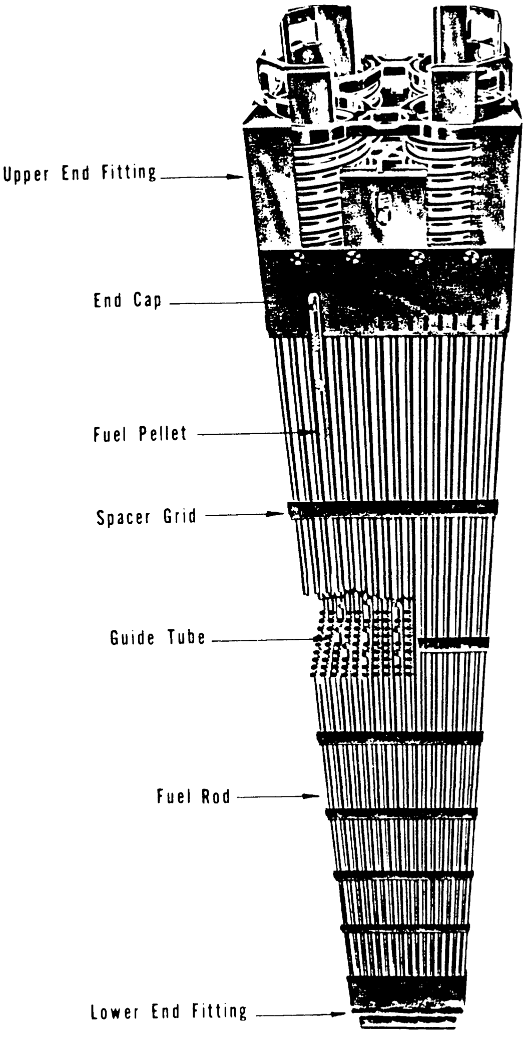

FIG. 12. PWR fuel assembly.

and beyond-design-basis accident analysis. To carry out reactor accident analysis, it is necessary to have an understanding of reactor physics (i.e., neutronics) of the reactor as well as the thermal hydraulics or heat-transfer mechanisms of the plant. In some instances, coupled neutronics and thermal hydraulics analysis is essential for the analysis of transients during operation.

A great deal of LWR safety research under the sponsorship of the U.S. Atomic Energy Commission (USAEC), the U.S. Nuclear Regulatory Commission (USNRC), industry, and the Electric Power Research Institute (EPRI) has been carried out since the late 1960s in the development and verification of complex thermal hydraulics accident analysis computer codes. The RELAP (Ransom *et al.*, 1985, 1987) and TRAC (Schnurr *et al.*, 1992; Borkowski *et al.*, 1992) family of computer codes have been developed by the Idaho National Engineering Laboratory (INEL) and Los Alamos National Laboratory (LANL), respectively, under the sponsorship of the USNRC for the thermal hydraulics analysis of LOCAs and SBLOCAs. RELAP-5, for example, is a complex computer code that solves six thermal hydraulic equations and considers energy, mass, and momentum for water in the liquid and vapor phases for the modeling of various components of the primary and secondary circuits of a PWR. A version of RELAP-5 also exists that can model BWRs. The reactor core is usually modeled on a three-dimensional basis, whereas the remainder of the components are modeled on a one-dimensional basis. Such complex computer codes as RELAP-5 to simulate SBLOCAs have required, until recently, the use of large mainframe computers such as the IBM 3090 or Cray high-speed computers. With the introduction of high-speed and large-memory work stations, such codes as RELAP-5 have been adapted to these smaller computers. A significant amount of experimental work, such as the LOFT (Loss of Flow Test) series of experiments at INEL as well as separate effects experiments, has been carried out to verify RELAP-5 analyses. In addition, international cooperative programs under the leadership of the USNRC for RELAP-5 verification have been carried out by research groups in the European Community, Japan, Korea, and Taiwan.

The TMI-2 accident showed the necessity of improving our understanding of severe accidents, since a large fraction of the core became molten and flowed to the bottom of the pressure vessel. Thermal hydraulic computer codes such as RELAP-5 are adequate for analyzing a transient to the onset of core uncovery but breaks down when there is massive degradation of the fuel. In order to simulate the degradation of the fuel mathematically, computer codes such as MELCOR (Summers *et al.*, 1991) have been developed by Sandia National Laboratory (SNL) and SCDAP/RELAP5 MOD2 (Allison *et al.*, 1989) by the INEL under the sponsorship of the USNRC. MELCOR is an attempt to follow the severe-

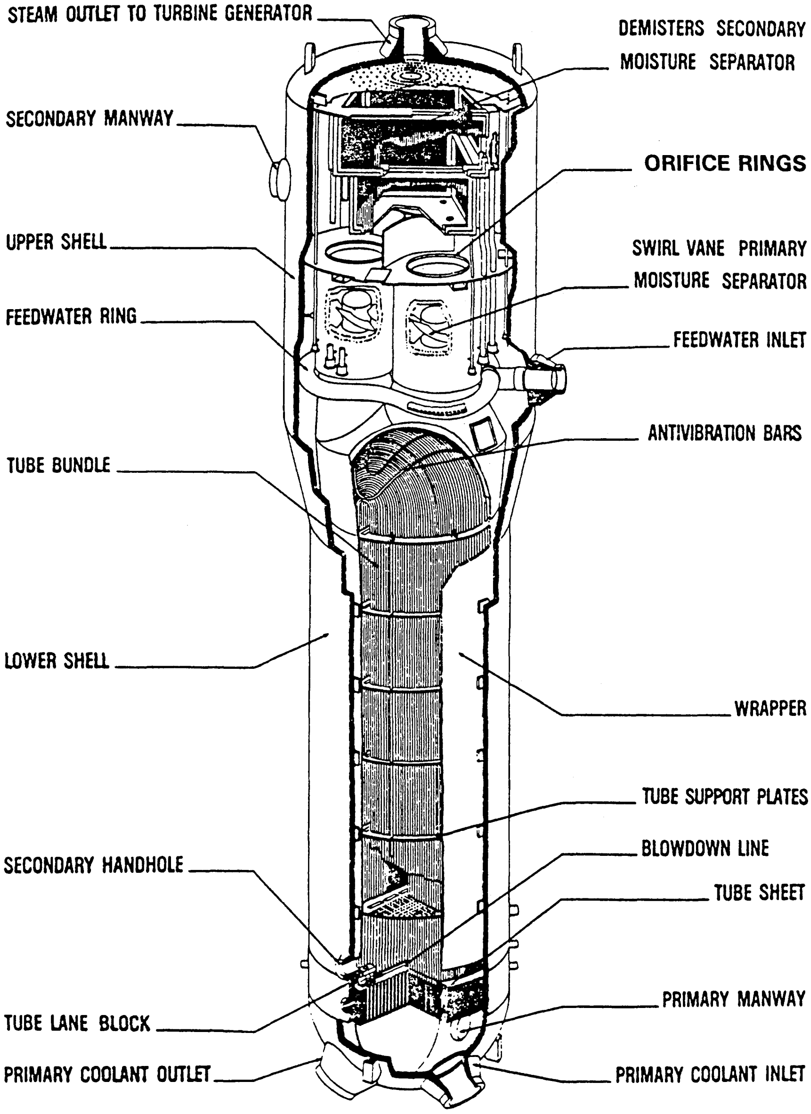
STEAM OUTLET TO TURBINE GENERATOR
DEMISTERS SECONDARY MOISTURE SEPARATOR
SECONDARY MANWAY
ORIFICE RINGS
UPPER SHELL
SWIRL VANE PRIMARY MOISTURE SEPARATOR
FEEDWATER RING
FEEDWATER INLET
ANTIVIBRATION BARS
TUBE BUNDLE
LOWER SHELL
WRAPPER
TUBE SUPPORT PLATES
SECONDARY HANDHOLE
BLOWDOWN LINE
TUBE SHEET
PRIMARY MANWAY
TUBE LANE BLOCK
PRIMARY COOLANT OUTLET
PRIMARY COOLANT INLET

FIG. 13. PWR steam generator.

accident phenomenon from the overheating of the fuel rods to the molten core slumping to the bottom of the pressure vessel, interacting with the steel pressure vessel, and finally flowing out of the pressure vessel onto the reactor cavity concrete pad beneath the reactor. Core–concrete interaction is then followed by a buildup of pressure and rupture of the containment due to combustion of hydrogen produced from the oxidation of the zirconium of the cladding combined with other phenomena such as direct containment-atmosphere heating by molten-core particles. The rupture of the containment with the subsequent release of fission products to the environment and its consequences are simulated by other computer codes such as the Melcor Accident Consequence Code System (MACCS) (Chanin *et al.*, 1990). Severe-accident analysis is an attempt to provide a quantitative basis to define emergency preparedness and evacuation zones surrounding a nuclear power plant, as well as to determine the design safety margins in the con-

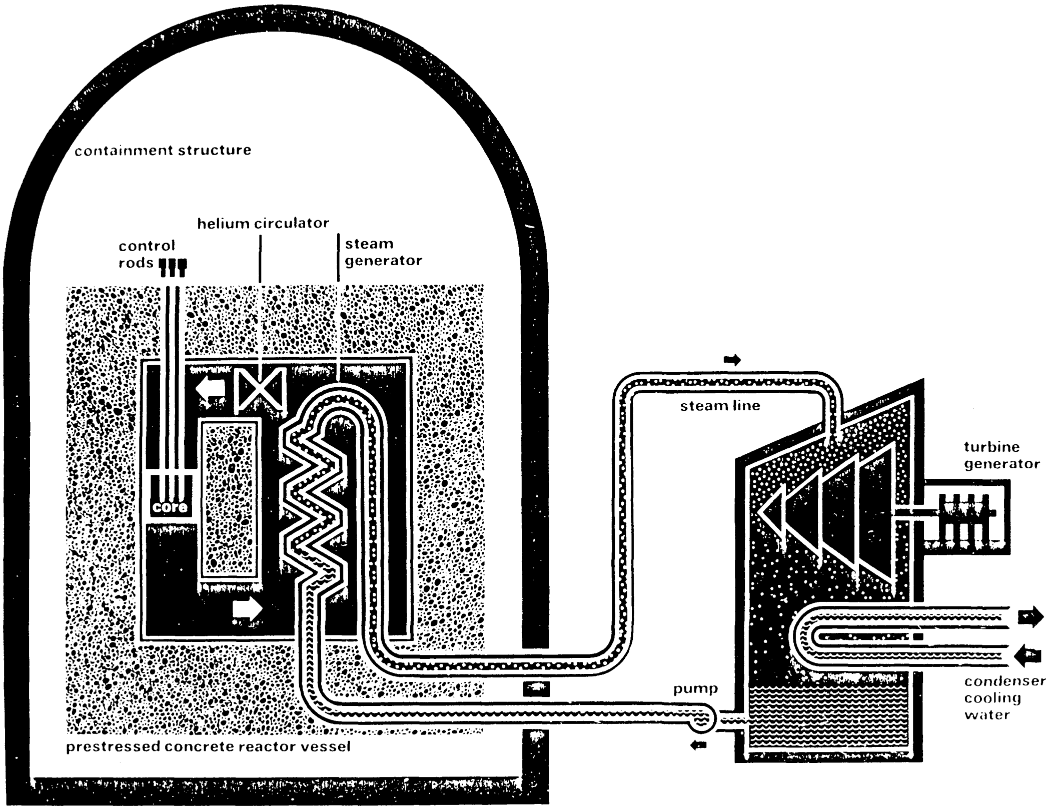

FIG. 14. Schematic diagram of HTGR.

tainment structures and other barriers to fission product release.

Since the mid-1970s probabilistic risk analysis (PRA) (Fullwood and Hall, 1987), more recently called probabilistic safety assessment (PSA), has become a popular way to analyze complete power plants to identify design deficiencies. The approach is to develop fault trees for all of the systems in a plant and to try to quantify overall failure probabilities of systems by providing failure probabilities of components based upon experience. The probability of a core becoming molten, or core melt probability, is calculated. The rupture of containment and resulting release of fission products to the environment can also be calculated on a probabilistic basis. Finally, with use of wind pattern frequencies and other weather statistics and population densities surrounding the plant site, consequences such as deaths and early cancers can be calculated on a probabilistic basis. The first complete analysis of power plants using PRAs was carried out in the Reactor Safety Study, WASH-1400 (NUREG 75/014) in 1973–1975. A more recent study on the assessment of severe accident risks for a set of commercial LWRs in the United States by the USNRC resulted in the report "Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants," published by the USNRC as NUREG 1150 in October 1990. Although these studies indicated that the probability of a core melt was in the range of 10^{-5} to 10^{-6} per reactor year with accompanying release of fission products, there has been a significant decrease of public confidence in commercial nuclear power as a result of the accidents at TMI-2 in April 1979 (Rogovin *et al.*, 1980) in the United States and Chernobyl Unit 4 (U.S. Nuclear Regulatory Commission, 1987; IAEA, 1992) in the Soviet Union in April 1987.

In the case of TMI-2 core, uncovery took

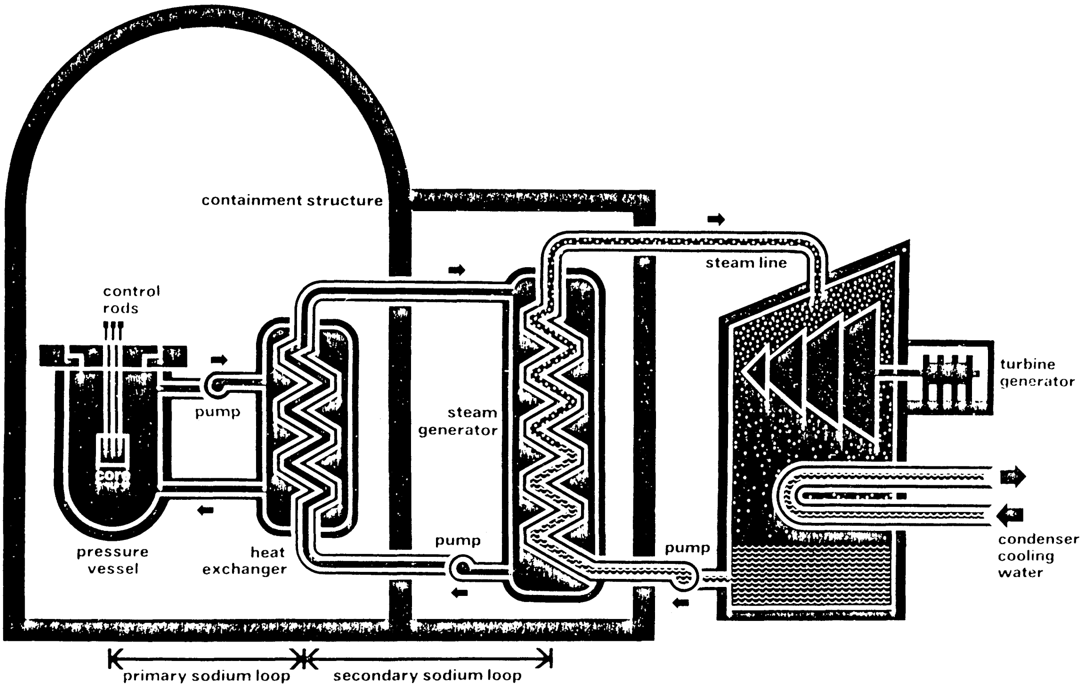

FIG. 15. Schematic diagram of LMFBR.

place as a result of a stuck-open motor-operated pressurizer relief valve coupled with operator errors. As a consequence, a large fraction of the core became molten and settled in the bottom of the pressure vessel. Fortunately, a sufficient amount of core cooling water was poured into the pressure vessel so that the vessel remained intact. Although the core was completely destroyed, most of the fission products were contained in the primary circuit. Some gaseous fission products such as the noble gases and radioiodine were released to the containment building, with only a small quantity of radioactivity being released to the atmosphere and surrounding environment. The report of Rogovin *et al.* (1980) estimates that about 2.4×10^6 Ci of short-lived noble gases and 15 Ci of ^{131}I in gaseous form were released to the environment. Because of the uncertainty in the status of the reactor in the early stages of the accident and media confusion, there was consideration given to large-scale evacuation of the population surrounding the TMI site. However, this was not carried out because the primary circuit and the containment remained intact and most of the fission products were contained. Although the reactor plant was a financial loss and the cleanup required large expenditure of funds, no lives were lost. The probability of any latent cancers has been estimated to be very small, such that it is not possible to distinguish cancers resulting from the accident from cancers due to other sources. A consequence of the TMI-2 accident has been the requirement of an emergency preparedness zone, as well as emergency procedures for evacuation and public notification, around nuclear power plant sites.

The Chernobyl-4 accident, on the other hand, released large quantities of radioactivity to the environment, required significant evacuation of the population around the site, and has taken considerable land out of use because of radioactive contamination of the soil. The Chernobyl nuclear power station is located about 60 miles north of the city of Kiev in Ukraine, one of the former Soviet Union republics. There were four RBMK-type nuclear power plants in operation at the station with an additional two under construction at the time of the accident in April 1986. The Chernobyl-4 accident was a result of design deficiencies and gross operator error. The lack of an adequate containment compounded the consequences because of the release of such long-lived radioactive isotopes as cesium-137 and strontium-90 and short-lived iodine-131 on a worldwide basis in the northern hemisphere. The Chernobyl-4, as described earlier, was a graphite-moderated, light-water–cooled, pressure-tube reactor. Design deficiencies, such as the reactor having a positive coolant void coefficient, control rods adding reactivity when being inserted from their full out position, and reactor instability at low power, coupled with operator error, such as bypassing scram circuits, operating with too many control rods in their full out position, and conducting an experiment without fully understanding the physics and thermal hydraulics of the system, caused the catastrophic accident. Decreasing the flow in the pressure tubes by deactivating the coolant pumps during a special experiment caused water boiling in the core, which resulted in increased reactivity. Power, pressure, and temperature increased. Coupled with the addition of reactivity from control rod insertion, this caused the rupture of pressure tubes, adding further voiding which in turn added more reactivity. The addition of large amounts of reactivity made the reactor superprompt critical, which essentially blew the reactor apart, resulting in molten fuel, fission products, and hot graphite being expelled out of the building and destroying the building structure itself. It has been estimated that about 50×10^6 Ci of noble gases and another 50×10^6 Ci of other radionuclides, which constituted about (3–4)% of the total fission-product inventory, were released to the environment (U.S. Nuclear Regulatory Commission, 1987). The initial wind pattern at the time distributed the fission products and fine vaporized fuel toward the west and northwest. Significant quantities of radioactivity were detected a few days after the accident in Finland and Sweden. Later, winds dropped the fission products in Eastern and Western Europe.

The magnitude of the potential property loss and health effects from the release of large quantitites of fission products and other radioactive reactor components make it mandatory that designers, constructors, operators, and regulators of nuclear power plants take utmost care in carrying out their respective responsibilities, with emphasis on reactor safety, to avoid future release of significant quantities of radioactivity.

6. ADVANCED WATER REACTORS

Since 1985, EPRI has led a utility, government, and industry advanced light-water reactor program. This program includes the simplification of overall design, including the design of passive emergency core-cooling, decay-heat–removal, and containment-cooling systems. The objective of the program is to seek greater safety margins, greater ease of construction, improved reliability and availability, improved maintainability, lower costs, and greater ease of operation over existing large LWRs (Stahlkopf *et al.*, 1988). The U.S. Department of Energy (DOE) since 1986 has had contracts with Westinghouse and GE for the detailed design, development, and certification for the APWR, AP600, and ABWR and the SBWR, respectively. The Westinghouse APWR, GE ABWR, and ABB Combustion Engineering's System 80 PWR are evolutionary LWRs improving upon the current version of LWRs. The GE SBWR and Westinghouse AP600 are mid-sized LWRs with passive safety features.

In addition to the LWRs being developed in the United States, other reactors such as the Modular High Temperature Gas-Cooled Reactor (MHTGR) by General Atomic; the advanced liquid-metal–cooled fast-spectrum reactor (PRISM) by GE; the light-water reactor (PIUS) by ABB Atom; and the heavy-water–moderated and –cooled reactor (CANDU-3) by Atomic Energy of Canada, Ltd. (AECL) are some of the advanced reactors that are being designed. A compilation of the characteristics of some of these reactors is provided in Table 8. Institutes of the Commonwealth of Independent States (CIS), states of the former USSR, which are forming private and independent organizations, are also working on advanced pressurized light-water reactors such as the VVER 91 and 92 (Gagarinski *et al.*, 1992). A French and German consortium, NPI, is also working on advanced PWRs.

An ABWR (Anon., 1992) of 1350 MW(e) has been developed jointly by GE, Hitachi, and Toshiba. The Tokyo Electric Power Company initiated the construction of two ABWRs in

Table 8. Design characteristics of new reactors (Nuclear News, 1992).

Design characteristics	System 80+	ABWR	SBWR	APWR	AP600	MHTGR	CANDU-3	PIUS	ALMR
Reactor type	Pressure l. water	Boiling l. water	Boiling l. water	Pressure l. water	Pressure l. water	Gas-cooled	Pressurized D_2O	Pressurized H_2O	Liq. metal fast breeder
Lead designer	ABB-CE	GE	GE	Westhse.	Westhse.	General Atomics	AECL	ABB ATOM	GE/ANL
MW(e) (net)	1300	1300	640	1050	600	538(4)	450	640	1440(3)
MW(t)	3817	3926	2000	3150	1940	1400(4)	1440	2000	4245(3)
Coolant	L. water	L. water	L. water	L. water	L. water	Helium	D_2O	H_2O	Liq. sodium
Moderator	L. water	L. water	L. water	L. water	L. water	Graphite	D_2O	H_2O	None
Fuel material	UO_2/PuO_2	UO_2	UO_2	UO_2	UO_2	UCO	Nat. UO_2	UO_2	U25%/Pu10%/Zr metal alloy
Cladding material	Zircal-4	Zircal-2	Zircal-2	Zircal-4	Zircaloy	Refractory coated particles	Zircaloy-4	Zircaloy-4	HT-9 ferritic alloy
Fuel geometry	16 × 16	8 × 8 or 9 × 9	8 × 8 or 9 × 9	17 × 17	17 × 17	Hexagonal graphite blocks	37-element fuel bundle	18 × 18	217 pin in hexagonal bundle
No. fuel assemblies	241	872	732	193	145	660	232 fuel channels	213	66 driver fuel
Pin diameter (mm)	9.7	12.3	12.3	9.5	9.5	13	13.1	9.5	7.2
Equivalent core diameter (mm)	3650	3708	4880	3370	2922	1650 3500	4912	3760	1570
Fuel length (mm)	3810	3708	2743	3658	3658	7925	5944	2500	1350
Average power density (kW/L)	95.5	50.6	41.0	96.2	78.8	5.9	12.78	72.3	180
Coolant inlet temperature (°C)	292	216	216	287	276	258	268	260	338
Coolant outlet temperature (°C)	324	288	288	325	312	687	310	290	485
System pressure (MPa)	15.41	7.07	7.07	15.51	15.41	6.38	9.90	8.99	Unpressurized
No. of loops	2	N.A.	N.A.	3	2				

1989 at its Kashiwazaki–Kariwa Nuclear Power Station. Some features that improve construction costs and maintainability and provide additional margins of safety include internal recirculation pumps, which eliminate external recirculation pumps and piping; no large coolant nozzles below the top of the core; new pressure suppression containment, which permits rapid vessel depressurization; improved core and fuel design that maximizes performance and fuel economy; and a triple-redundant ECCS core protection system.

The APWR (Anon., 1992) that has been designed by Westinghouse, Mitsubishi Heavy Industries of Japan, and five Japanese PWR owner utilities is a four-loop, 1350-MW(e) power plant. Its most significant difference from the exisiting PWRs is that it has a 15–20% lower power density, which provides a greater margin of safety. In addition, a larger pressure vessel provides a larger core inventory of water above the core, which enhances safety and decreases the required capacity of the ECC system. A larger diameter containment building provides greater ease in maintenance. Economic goals are to reduce the capital costs by 15% and fuel-cycle costs by 20% over existing PWRs. The reduction in the fuel-cycle cost is achieved by the reduction of the power density, which leads to a reduction of the initial enrichment required to achieve the same energy output. A further reduction of the initial enrichment is accomplished by shifting the spectrum of the core during early operation of the core by displacing the moderator with a mechanical water displacer. This results in increased production of plutonium by slightly hardening the spectrum for increased resonance neutron absorption. This plutonium is then burned later in the cycle when the displacer is removed, producing a softer spectrum and increased core reactivity.

The AP600 (Anon., 1992; Vijuk and Bruschi, 1988; Bruschi and Anderson, 1991) is a two-loop, 600-MW(e) PWR system that includes a simplified reactor coolant system and a balance of plant, passive safety, and plant arrangement systems with a modular construction approach. Some of the key design features include a reactor coolant system that uses sealed canned motors mounted directly on the steam generators, a core design resulting in a low core power density (80 kW/L), a reactor core surrounded by a stainless steel and water reflector, and passive safety features that operate following plant transients or accidents. The passive safety system provides for residual heat removal, reactor coolant makeup for inventory control, safety injection, and containment spray cooling. A passive residual-heat–removal heat exchanger (PRHR-HX) has been provided to remove the decay heat by natural circulation from the core in the event that the normal or startup feedwater systems (i.e., steam-generator heat-removal systems) are not functioning. In the AP600, if there is steam released to the atmosphere inside the steel containment, natural circulation is used to transfer the heat to the steel shell. The AP600 has been designed to have significantly fewer valves and large pumps, less piping, and less control cabling and ducting in order to minimize construction costs.

General Electric has been working on the design of the Simplified Boiling Water Reactor (SBWR) (McCandless and Redding, 1989) since 1982 with the objectives of providing power generation costs cheaper than coal, simplified safety systems, shorter construction schedules, and a 600-MW electrical rating. In the SBWR design, the power density was decreased to 42 kW/L, and natural circulation was chosen for providing coolant flow through the core. A gravity-driven core-cooling system provides a simple approach to emergency core cooling, which eliminates the need for pumps and emergency diesels. It has a gravity driven cooling pool that can keep the core covered. The SBWR pressure vessel is designed so that there are no large pipes attached to the pressure vessel near or below the core elevation, which ensures core coverage for all design basis events. In addition to the passive cooling pool, the SBWR uses a pressure suppression pool, isolation condenser, and passive containment cooling system for its passive emergency response.

The PIUS (Process Inherent Ultimate Safety) reactor concept by ABB Atom of Sweden (Hannerz, 1988) is a 640-MW(e) PWR-type reactor that is located near the bottom of a deep reactor pool. The pool is compartmentalized so that the reactor core is near the bottom of a chimneylike structure submerged in a large pool. The pool contains water with a high boron concentration in a prestressed concrete vessel. The reactivity is controlled by the boron concentration in the coolant water and by temperature so that there are no control rods in the reactor. The moderator temperature coefficient is strongly negative throughout the operating cycle. During normal operation, the power is controlled by using the negative moderator temperature coefficient without adjusting the coolant boron concentration. The PIUS is an innovative concept since fuel integrity is maintained by an inherent passive coolant system, and there should be no core uncovery or fuel overheating in any credible accident because of the large volume of available coolant water.

7. ECONOMIC CONSIDERATIONS

The generation of electricity through the use of nuclear energy must be competititve with the cost of power generation from the burning of fossil fuels such as coal and oil. The cost of electricity generation may be considered to be composed of capital carrying charges, operation and maintenance costs, and fuel costs. The capital carrying charges include the depreciation and amortization of the costs of building and financing the plant. The capital costs of a plant may also change depending on the amount of additional investments required during its operating life. In the case of nuclear plants, the capital carrying charges are the predominant costs for the generation of electricity. The large capital carrying costs include the high cost of construction, as well as the interest charges accumulated during the length of time required to construct the plant, and decommissioning costs.

The Ahearne Committee Report on Nuclear Power (National Research Council, 1992; Ahearne, 1993) presents the Energy Information Administration's 1988 figures for the comparison of nuclear- and coal-generated electricity. Table 9 lists the highest, lowest, and average total generation costs for nuclear and coal in 1988.

Table 9. Average electricity generation costs for 1988 (cents/kWh) (National Research Council, 1992).

			Highest/Lowest	
	Nuclear	Coal	Nuclear	Coal
Capital carrying	3.4	1.1	9.4/0.4	5.4/0.3
Operation & maintenance	1.5	0.4	1.2/0.8	0.7/0.5
Fuel	0.8	1.7	0.7/0.5	2.5/1.4
Total	5.6	3.1	11.3/1.6	8.5/2.2

The reasons that existing nuclear power plants have higher capital carrying charges than coal plants are that the nuclear power plant equipment and building have been more expensive to construct, they have taken longer to construct and hence accumulate more interest charges, and finally they have decommissioning costs that must be taken into consideration. In the late 1960s and early 1970s, at the beginning of the commercial nuclear power industry, it was believed that operation and maintenance (O&M) costs would be fairly low. After the TMI-2 accident in April 1979, there was an increase in regulatory requirements, which significantly increased the O&M costs.

The Ahearne committee reports that plants that went into commercial operation during the period 1978–1989 required an average construction time of 11.1 years in the United States as compared with 5.9 years in France and 4.7 years in Japan. Since a lengthy construction period with the attendant interest charges has a significant impact on the capital costs, there is a strong incentive to minimize the construction time in the United States.

As there have not been any orders for new nuclear power plants since 1979, it is difficult to estimate the cost of electricity generation from new nuclear power plants because of the uncertainty of the capital costs for the advanced light-water reactors (ALWR) that the reactor vendors would probably supply to the utilities. The ALWRs are being designed with simplification as well as ease in construction, operation, and maintenance in mind. In addition, with the NRC tending toward a single construction and operating license approval process, which would decrease the construction time, the capital cost should be more predictable for the new plants.

In 1992, there were a couple of nuclear power plants that the utility owners decided to shut down because of the costs to replace steam generators, such as in the case of Portland General Electric's Trojan power plant whose alternate source of electricity from hydropower is cheaper. The Yankee Rowe plant in Massachusetts was shut down because of the uncertainties in radiation embrittlement of the pressure vessel, which caused concern in the NRC. As the nuclear power plants begin to approach the end of their 40-year operating license period, the owners must consider the cost effectiveness of pursuing license extension. Such an extension might require additional capital investment, which, together with higher O&M costs, may make nuclear electricity less competitive compared with electricity generated from other sources. Nuclear electricity is highly competitive in countries such as Japan and France, which have no indigenous fossil fuels. Limitations on the production of greenhouse gases, such as CO_2, and acid rain restrictions on fossil fuel combustion in both electricity generation and transportation may result in the increased use of nuclear power plants for electricity generation in the future in the United States.

8. RADIOACTIVE WASTE MANAGEMENT

Although the production of electricity using fission energy has a number of advantages, one of the disadvantages is that the resultant radioactive substances created must be prevented from being dispersed in the environment for many years. In addition to the high-level waste (i.e., fission products and transuranium isotopes) that is produced during reactor operation, there are other low-level wastes generated during the fuel cycle that must be managed (see NUCLEAR WASTE MANAGEMENT).

During the fuel production process from the mining and milling operation through the conversion to uranium hexafloride, the gaseous-diffusion, centrifuge, or atomic-vapor–laser enrichment processes, and the fuel fabrication process, there are mill tailings, sludge from chemical processes, scrap, and fabrication residue containing uranium and its daughter products that must be considered as low-level waste. During reactor operation, resins from the cleanup of the primary cooling water, filters from air cleaning, sludge and waste from evaporator concentrates, and airborne releases of radioactive noble gases and tritium are additional sources of low-level radioactive waste. If the spent fuel from commercial nuclear power plants is reprocessed, as it is in France, the United Kingdom, Japan, and the Former Soviet Union, and as is the fuel from military reactors in the United States, there is high-level waste consisting of the fission products and transuranic elements that must be disposed of. In addition, there will be cladding hulls with residual fuel, transuranic solids from the chemical processing, and gaseous radioactive wastes such as the noble gases, tritium, and iodine. Finally, in the decontamination and decommissioning of the nuclear plant, there will be large components, such as the pressure vessel and the reactor internals including the control rods, internal shield, etc., which will have become radioactive from interactions with neutrons. In all of the components of the primary circuit, there will be activated residue from small impurities in the water, which is difficult to remove. These components in total must be considered for radioactive material disposal. The concrete shielding and other support structures surrounding the pressure vessel will also have become activated to some small degree so that they may have to be considered as low-level waste.

The United States national policy in 1994 is to not reprocess spent fuel from commercial nuclear power plants but to develop a final storage facility for the spent fuel to discourage the use of plutonium so as to encourage nonproliferation. Although it had been planned by the Department of Energy (DOE) to characterize three possible sites for permanent storage of spent fuel, the U.S. Congress in 1987 enacted legislation restricting site characterization to the Yucca Mountain site in the State of Nevada. Because of public opposition from the State of Nevada and its citizens, initiation of the characterization studies was delayed until late 1991 by the refusal of the State to issue various permits required for the study. The spent fuel contains long-lived fission products and transuranic elements, which must be contained for tens to hundreds of thousands of years. The proposed concept is to encapsulate the spent fuel and then store the fuel in underground storage facilities at Yucca Mountain. The controversy involves the question of whether the Yucca Mountain area will remain stable against earthquakes or volcanic action for a long period so that the disposal facility will be free from ground water.

In countries such as France and Japan, where fuel reprocessing is carried out, the fission products are separated from the transuranic elements. The fission products are then vitrified (i.e., embedded in glass) and then encapsulated for final disposal in underground salt mines. The transuranic elements such as plutonium are separated for use as fuel in liquid-metal–cooled breeder reactors or are mixed with enriched uranium oxide for light-water reactor fuel [mixed oxide (MOX) fuel]. A high degree of separation of the transuranic elements from the spent fuel in fuel reprocessing would ease the requirements for the high-level waste respository, since most of the fission products have significantly shorter half-lives than the alpha-active transuranic elements.

The DOE, in 1993, carried out alternative studies for the disposal of the transuranic elements. Concepts such as developing reactors specifically for burning or transmuting the transuranic elements, or using high-current, high-energy proton accelerators for transmuting the transuranic elements to fission products with shorter half-lives, are being considered. The principle in transmutation is to design a system with as hard a neutron energy spectrum as possible so that the capture-to-fission ratio for the transuranic elements is as low as possible. This minimizes the number of higher isotopes being produced and maximizes the fission of the fissionable transuranic elements. In addition, it is desirable, from a safety point of view, to have a system which has as high a power density as possible to minimize the total amount of transuranic inventory in the system. The concept of using accelerators for transmutation is based on the ability to construct a linear proton accelerator that produces high-energy protons in the range of 1.6 GeV with currents of 25 mA or greater. The high-energy, high-current protons are targeted on a heavy element such as lead or tungsten, which produces copious amounts of high energy neutrons. These neutrons then cause the fission of the transuranic elements in the system. The long–half-life transuranic elements are, thus, transmuted by fission into shorter–half-life fission products. Since the accelerator system can be operated with a multiplication of $k < 1$, or less than critical, the concern for supercriticality accidents would be diminished.

Very hard-spectrum reactors using liquid-metal cooling could be designed for the transmuting of transuranic elements. Such reactors could also be designed to produce electrical power.

9. NUCLEAR REGULATION

9.1 USNRC

The federal government in the United States has the responsibility for regulating the production and utilization of atomic energy in accordance with the Atomic Energy Act of 1946 and the Atomic Energy Act of 1954. The Energy Reorganization Act of October 1974 (Public Law 93–438) established the Nuclear Regulatory Commission (NRC) and separated the regulatory functions from the promotional functions of the Atomic Energy Commission (AEC). The Energy Reorganization Act, at the same time, created the Energy Research and Development Administration (ERDA), which took over the other functions of the AEC including its promotional developmental functions. The ERDA eventually became the Department of Energy by the Act of 1977. The NRC is composed of five members appointed by the President of the United States. Its principal functions are the licensing and regulation, including safeguards, of the construction and operation of nuclear reactors and the facilities for processing, handling, and transporting nuclear materials, including special nuclear materials, and the transport and storage of radioactive waste. It also has the responsibility for conducting research necessary for the performance of its licensing and regulatory functions. In carrying out its regulatory functions, the Commission reviews designs and construction quality assurance procedures, as well as conducting inspections of nuclear facilities and enforcing its regulations. The NRC maintains site inspectors at each of the commercial nuclear power plants in the United States.

The Commission carries out its responsibilities by promulgating rules and regulations, which become a part of the Code of Federal Regulations (CFR), a manual for the executive departments and the agencies of the federal government. Title 10, Chapter 1 of the CFR has been reserved for the NRC regulations. This chapter is divided into various parts, as shown in Table 10. The NRC also issues Regulatory Guides which

1. describe methods which the regulatory staff finds acceptable in implementing regulations;
2. discuss techniques used by staff to evaluate specific problems; and
3. provide guidance to license applicants.

The Commission has the three statutory offices of Nuclear Reactor Regulation, Nuclear Regulatory Research, and Nuclear Material Safety and Safeguards to carry out its principal duties. It has an Executive Director of Operations who coordinates the activities of the three principal offices, as well as five regional offices, as of 1993, which provide oversight of the nuclear facilities in their respective geographical areas. It also has the Advisory Committee on Reactor Safeguards (ACRS), the Advisory Committee on Nuclear Waste, and Atomic Safety and Licensing Boards to provide the Commission with advice.

The Commission had, up to 1990, required the submission of a preliminary safety analysis report (PSAR) for review and approval prior to the initiation of construction of a nuclear facility and a final safety analysis report (FSAR) before initiation of operation of the facility. The content requirements for SARs are given in NRC Regulatory Guide 1.70 and include such items as detailed description of the site, plant, reactor, engineered safety features, radioactive waste management, radiation protection, conduct of operations, safety assessments including accident analyses, and technical specifications.

Table 10. Code of Federal Regulations, Title 10, Chapter 1.

Part 1:	Statement of Organization and General Information
Part 2:	Rules of Practice
Part 20:	Radiation Protection Standards
Part 50:	Licensing of Production and Utilization Facilities
Part 51:	Licensing and Regulatory Policy and Procedures for Environmental Protection
Part 52:	Early Site Permits; Standard Design Certifications; and Combined Licenses for Nuclear Power Plants
Part 55:	Operators' Licenses
Part 70:	Special Nuclear Materials
Part 100:	Reactor Siting Criteria

The objective of the safety assessment of nuclear power plants is to provide an analysis and evaluation of the design and performance of the structures, systems, and components of the facility to assess the risk to public health resulting from operation of the plant. The analysis must include margins of safety during normal and transient conditions anticipated during the life of the facility and adequacy of the facility to prevent accidents and mitigate the consequences of accidents.

The range of accidents which should be considered are

1. events of moderate frequency with no abnormal radioactive releases from the facility,
2. events of small probability with potential for small radioactive release from the facility,
3. potentially severe accidents of extremely low probability to establish performance requirements of engineered safety features and to evaluate the acceptability of the site.

For new nuclear power plants, the Commission now also requires the submission of a Probabilistic Safety Assessment (PSA) [formerly called Probabilistic Risk Assessment (PRA)] based upon the detailed design of the plant. It should also be recognized that the Environmental Protection Agency (EPA) requires an Environmental Impact Statement for new large projects.

Following a detailed review of the SAR by the Commission staff, a review by the ACRS is conducted. Before a license is granted for the start of construction, the Commission usually conducts a public hearing at which any opposition views are heard.

In 1990, the NRC promulgated 10CFR52, "Early Site Permits; Standard Design Certifications; and Combined Licenses for Nuclear Power Plants," which significantly changed the rules and regulations governing the licensing of nuclear power plants. The NRC now issues a Combined Operating License (COL), which essentially combines the former construction permit and the operating license. Reactor vendors now submit an application for design certification of a specific nuclear reactor design. In parallel, a utility may submit an application for an early site permit. Following completion of the design certification process and an early site permit, an application for the COL may be considered. The technical information that is required of applicants for a COL is essentially the same as for the previous construction permits and operating licenses except for information technically relevant to a specific site. In addition, such analysis as a design-specific probabilistic risk assessment and proposed tests, inspections, and acceptance criteria are also required. The NRC staff and ACRS conduct a thorough review of the COL application, which is followed by a mandatory public hearing before a COL is issued. Construction of the nuclear power plant may be started after the issuance of the COL. Following completion of the construction, the NRC must issue a finding that the plant as constructed meets all acceptance criteria before operation is permited.

9.2 IAEA

The International Atomic Energy Agency (IAEA) (Scheinman, 1987) was originally established to prevent the proliferation of nuclear weapons and to encourage the use of nuclear energy for peaceful uses. Since the Chernobyl accident in 1986, there has been increased emphasis on helping member states with reactor safety. Before Chernobyl, the IAEA organized international conferences and workshops on nuclear safety, as well as providing operational safety review teams (OSART) for the evaluation of the safety of nuclear power plants at the request of member states. In the late 1970s, the IAEA developed an Incident Reporting System (IRS) for the compilation and distribution of unusual events occuring at nuclear power plants. The IAEA IRS contains unusual-event reports compiled by the Agency as well as those compiled by the Nuclear Energy Agency (NEA) of the Organization of Economic Cooperation and Development (OECD). Also included in the complilation of the unusual events is the analysis of unusual events.

Following the Chernobyl-4 accident in 1986, an International Nuclear Safety Advisory Group (INSAG) was established by the Director General of the IAEA with the intention of strengthening the IAEA's contribution to ensuring the safety of nuclear power plants. The INSAG's responsibilities include "serving as a forum for the exchange of information on nuclear safety issues of international significance and formulating where possible common safety concepts." One of its first products has been the issuance of the "Basic Safety Principles for Nuclear Power Plants" (IAEA, 1988), which outlines specific safety principles that should be followed for the design, construction, and operation of nuclear power plants.

The IAEA at the request of member states will conduct peer review of safety activities in those states. Peer review of three categories of activities has been carried out. These are regulatory activities, engineering matters, and operational safety. The IAEA has organized independent regulatory review teams and developed nuclear safety standards under the heading of regulatory activity peer reviews. They have conducted peer reviews of PSAs, safety analysis, siting, and design concepts under the category of engineering safety review service. In the operational safety category, they have conducted reviews of the operating process, event assessment, and feedback of experience.

Another area where the IAEA has been very active is in the organization of Assessment of Safety Significant Events Teams, which are sent to plants that have had unusual events. They have had over 50 separate missions to over 21 countries.

In the early 1990s, experts from the IAEA and NEA [the Nuclear Energy Agency of the Organization for Economic Cooperation and Development (OECD)] developed the International Nuclear Event Scale (INES), shown in Table 11, which would categorize unusual events so that it would be easier for the public to understand the severity of unusual events or incidents. The INES has categories

Table 11. International nuclear event scale (INES).

0	Below scale (deviation, no safety significance)
1	Anomaly
2	Incident
3	Serious incident
4	Accident without significant off-site risk
5	Accident with off-site risk
6	Serious accident (significant off-site impact)
7	Major accident (widespread health and environmental effects)

from 0 to 7, where 7 represents a major accident that has widespread health and environmental effects. The TMI-2 accident would be a category 5, whereas the Chernobyl-4 accident would be classified as a 7. Although some member states such as Japan, France, and the United States have developed their own nuclear event scale, the IAEA INES is becoming accepted on a worldwide basis.

10. CONCLUSION

This has been a brief review of the development of nuclear fission reactors over the past 50 years. Although nuclear energy was born of military necessity, commercial nuclear electricity with over 411 large nuclear power plants in operation worldwide in 1993 has become a mature, major new source of energy for the world. In 1953, 10 years after the construction of the first nuclear reactor, President Eisenhower of the United States initiated the Atoms for Peace Program, which has since spread the knowledge of nuclear fission technology worldwide. It is up to mankind to determine whether nuclear fission will be used as a source of clean energy for the improvement of man's living standards throughout the world or for destructive purposes through its use for nuclear weapons. If used carefully and properly, nuclear fission, as described in this article, is an environmentally benign technology which can provide clean heat and electrical energy to help fulfill the increasing energy demands by the world's current population of about 5×10^9 people and for a projected $(8–9) \times 10^9$ people in the years 2020–2030 and beyond. With the depletion of fossil fuels such as wood, oil, and coal and the concern over the increasing greenhouse gases in the atmosphere from combustion of fossil fuels, it is urgent that people universally accept the use of nuclear fission for peaceful purposes. Issues such as prevention of nuclear weapons proliferation, radioactive waste management, safety, and economics must be addressed to the satisfaction of the concerned communities before there will be universal acceptance of this very important and useful technology.

Works Cited

Ahearne, J. F. (1993), "The Future of Nuclear Power," *Amer. Sci.* **81**, 24–35.

Allison, C. M., Johnson, E. C., Berna, G. A., Gheng, T. C., Hagrman, D. L. (1989), *SCDAP/RELAP5/MOD2 Code Manual*, Vols. 1 and 2, NTIS No. NUREG/CR-5273-V1/HDM, Springfield, VA: National Technical Information Service.

Anon. (1992), "The New Reactors," *Nucl. News* **35** (12), 66–90.

Bell, G. I., Glasstone, S. (1970), *Nuclear Reactor Theory*, New York: Van Nostrand Reinhold.

Borkowski, J. A., Wade, N. L., Giles, M. M., Rouhani, S. Z., Shumway, R. W. (1992), *TRAC-BF1/MOD1: An Advanced Best Estimate Computer Program for BWR Accident Analysis*, Vols. 1 and 2, NTIS No. NUREG/CR-4356-V1/HDM, Springfield, VA: National Technical Information Service.

Briesmeister, J. F. (Ed.) (1986), *MCNP—A General Monte Carlo Code for Neutron and Photo Transport. Version 3A, Rev. 2*, NTIS No. DE87000708/HDM, Springfield, VA: National Technical Information Service.

Bruschi, H., Anderson, T. (1991), "Turning the Key," *Nucl. Eng. Int.* **36** (448), 15–22.

Chanin, D. I., Spring, J. L., Ritchie, L. T., Jow, H. N. (1990), *Melcor Accident Consequence Code System (MACCS), Users Guide*, Vol. 1, NTIS No. NUREG/CR-4691-V1/HDM, Springfield, VA: National Technical Information Service.

Cowan, G. A. (1976), "A Natural Fission Reactor," *Sci. Amer.* **235** (1), 36–47.

Davey, W. G. (1972), "The Demonstration Reactor Benchmark Program," in: *Proceedings of the Topical Meeting on New Developments in Reactor Physics and Shielding*, Vol. 2, Rept. No. CONF-720901, Washington, DC: U.S. Atomic Energy Commission, p. 789.

Dietrich, J. R., Lichtenberger, H. V., Zinn, W. H. (1955), "Design and Operating Experience of a Prototype Boiling Water Power Reactor," in: R. A. Charpie *et al.* (Eds.), *Proceedings of the International Conference on the Peaceful Uses of Atomic Energy*, Vol. 3, New York: United Nations, p. 56.

Energy Science and Technology Center (1993), *Software Listing*, Report No. ESTSC-2/Rev. 1, Oak Ridge, TN: Oak Ridge National Laboratory.

Fullwood, R. R., Hall, R. E. (1987), *Probabilistic Risk Assessment in the Nuclear Power Industry*, Oxford, United Kingdom: Pergamon.

Gagarinski, A. Y., Ignatiev, V. V., Novikov, V. M., Subbotin, S. A. (1992), "Advanced Light Water Reactors: Russian Approaches," *IAEA Bull.* **2** (2), 37–40.

Hannerz, K. (1988), "Making Progress on PIUS Design and Verification," *Nucl. Eng. Int.* **33** (412), 29–31.

Harrer, J. M., Jameson, A. S., West, J. M. (1955), "The Engineering Design of a Prototype Boiling Water Power Plant," in: R. A. Charpie *et al.* (Eds.), *Proceedings of International Conference on the Peaceful Uses of Atomic Energy*, Vol. 3, New York: United Nations, p. 250.

Hendrie, J. M. (1965), "The Brookhaven High Flux Beam Reactor (HFBR)," in: A. de Calmès *et al.* (Eds.), *Proceedings of the International Confer-*

ence on the Peaceful Uses of Atomic Energy, Vol. 7, New York: United Nations, p. 373.

IAEA (1988), *Basic Safety Principles for Nuclear Power Plants*, Safety Series No. 75-INSAG-3, Vienna: International Atomic Energy Agency.

IAEA (1989), *Directory of Nuclear Research Reactors*, Vienna: International Atomic Energy Agency.

IAEA (1992), *The Chernobyl Accident: Updating of INSAG-1*, Report No. INSAG-7, Vienna: International Atomic Energy Agency.

Lamarsh, J. R. (1977), *Introduction to Nuclear Engineering*, Reading, MA: Addison-Wesley.

Lentz, G. L., Buschman, H. W., Smith, R. N. (1986), "EBR-II: Twenty Years of Operating Experience," in: *Proceedings of the Symposium on Fast Breeder Reactors: Experience and Trends*, Vol. 1, Vienna: International Atomic Energy Agency, p. 171.

Lichtenberger, H. V., Thalgott, F. W., Kato, W. Y., Novick, M. (1955), "NaK Cooled Fast Reactor (EBR-I)," in: R. A. Charpie *et al.* (Eds.), *Proceedings of the International Conference on the Peaceful Uses of Atomic Energy*, Vol. 3, New York: United Nations, p. 345.

McCandless, R. J., Redding, J. R. (1989), "Simplicity: the Key to Improved Safety, Performance and Economics," *Nucl. Eng. Int.* **34** (424), 20–24.

McLane, V., Dunford, C. L., Rose, P. F. (1988), *Neutron Cross Sections*, Vol. 2, San Diego: Academic.

Mughabghab, S. F., Divadeenam, M., Holden, N. E. (1981), *Neutron Cross Sections*, Vol. 1, New York: Academic.

National Energy Software Center (1991), "Compilation of Program Abstracts," Report No. ANL-7411, Argonne, IL: Argonne National Laboratory.

National Research Council (1992), *Nuclear Power, Technical and Institutional Options for the Future*, Washington, DC: National Academy Press.

Peckinpaugh, C. L. (1986), "Fast Flux Test Facility: First Three Years of Operation," in: *Proceedings of a Symposium on Fast Breeder Reactors: Experience and Trends*, Vol. 1, Vienna: International Atomic Energy Agency, p. 203.

Peckinpaugh, C. L., Bennett, R. A., Wykoff, W. R. (1982), "Fast Flux Test Facility, Operational Results," in: *Proceedings of the International Conference on Nuclear Power Experience*, Vol. 5, Vienna: International Atomic Energy Agency, p. 155.

Radiation Shielding Information Center (1993), *RSIC Software Directory*, Oak Ridge, TN: Oak Ridge National Laboratory.

Ransom, V. H., Wagner, R. J., Trapp, J. A., Feinauer, L. R., Johnsen, G. W. (1985), *RELAP5/MOD2 Code Manual*, Vol. 1, "Code Structure, Systems Models, and Solution Methods," NTIS No. NUREG/CR-4312/HDM, Springfield, VA: National Technical Information Service.

Ransom, V. H., Wagner, R. J., Trapp, J. A., Johnsen, G. W., Miller, C. S. (1987), *RELAP5/MOD2 Code Manual*, Vol. 2, "Users Guide and Input Requirements," NTIS No. NUREG/CR-4312-REV-1/HDM, Springfield, VA: National Technical Information Service.

Rhodes, R. (1986), *The Making of the Atomic Bomb*, New York: Simon and Shuster.

Rogovin, M., Frampton, G. T., Jr., Cornell, E. K., DeYoung, R. C., Budnitz, R. (1980), *Three Mile Island: A Report to the Commissioners and to the Public*, Vol. 1, NTIS No. NUREG/CR-1250-V-1/HDM, Springfield, VA: National Technical Information Service.

Scheinman, L. (1987), *The International Atomic Energy Agency and World Nuclear Order*, Washington, DC: Resources for the Future.

Schnurr, N.M., Steinke, R. G., Spore, J. W., Martinez, V. (1992), *TRAC-PF1/MOD2 Code Manual*, NTIS No. NUREG/CR-5673-V2/HDM, Springfield, VA: National Technical Information Service.

Smyth, H. D. (1945), *Atomic Energy for Military Purposes*, Princeton, NJ: Princeton University Press.

Stahlkopf, K. E., DeVine, J. C., Sugnet, W. R. (1988), "US ALWR Programme sets out Utility Requirements for the Future," *Nucl. Eng. Int.* **33** (412) 16–19.

Stamm'ler, R. J. J., Abbate, M. J. (1983), *Methods of Steady-State Reactor Physics in Nuclear Design*, New York: Academic.

Summers R. M., Cole, R. K., Boucheron, E. A., Carmel, M. K., Dingman, S. E. (1991), *MELCOR 1.8.0: A Computer Code for Nuclear Reactor Severe Accident Source Term and Risk Assessment Analyses*, NTIS No. NUREG/CR-5531HDM, Springfield, VA: National Technical Information Service.

Swartout, J. A., Boch, A. L., Cole, T. E., Cheverton, R. D., Adamson, G. M., Winters, C. E. (1965), "High Flux Isotope Reactor (HFIR)," in: A. de Calmès *et al.* (Eds.), *Proceedings of the International Conference on the Peaceful Uses of Atomic Energy*, Vol. 7, New York: United Nations, p. 360.

U.S. Nuclear Regulatory Commission (1987), *Report on the Accident at the Chernobyl Nuclear Power Station*, NTIS No. NUREG-1250/HDM, Springfield, VA: National Technical Information Service.

Vijuk, R., Bruschi, H. (1988), "AP600 Offers a Simple Way to Greater Safety, Operability, and Maintainability," *Nucl. Eng. Int.* **33** (412), 22–28.

Waltar, A. E. Reynolds, A. B. (1981), *Fast Breeder Reactors*, New York: Pergamon, p. 772.

Further Reading

Duderstadt, J. J., Hamilton, L. J. (1976), *Nuclear Reactor Analysis*, New York: John Wiley & Sons.

Glasstone, S., Sesonske, A. (1987), *Nuclear Reactor Engineering*, New York: Van Nostrand Reinhold.

Greenspan, H., Kelber, C. N., Okrent, D. (1968), *Computing Methods in Reactor Physics*, New York: Gordon and Breach, 1968.

Henry, A. F. (1975), *Nuclear Reactor Analysis*, Cambridge, MA: MIT Press.

Hsu, Y. Y., Graham, R. E. (1976), *Transport Processes in Boiling and Two-Phase Systems*, Washington, DC: Hemisphere Publishing.

Ott, K. O., Bezella, W. A. (1989), *Introductory Nuclear Reactor Statics*, LaGrange Park, IL: American Nuclear Society.

Ott, K. O., Neuhold, R. J. (1985), *Introductory Nuclear Reactor Dynamics*, LaGrange Park, IL: American Nuclear Society.

Tong, L.S., Weisman, J. (1979), *Thermal Analysis of Pressurized Water Reactors*, 2nd ed., LaGrange Park, IL: American Nuclear Society.

NUCLEAR FUELS AND ISOTOPES

Allen G. Croff, *Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A.*

Introduction 480
1. **Overview** 480
2. **Uranium Mining and Milling** 482
2.1 Uranium Sources and Mining Techniques 482
2.2 Uranium Milling and Refining 482
2.3 Purification and Conversion of UOC to UF_6 483
3. **Isotope Enrichment** 483
3.1 Uranium 484
3.1.1 Electromagnetic Isotope Separation (EMIS) 484
3.1.2 Gaseous Diffusion 484
3.1.3 Gas Centrifuge 485
3.1.4 Laser Isotope Separation (LIS) 486
3.1.5 Aerodynamic Processes 488
3.2 Stable Isotopes 488
3.2.1 Deuterium 488
3.2.2 Lithium 489
3.2.3 Boron 489
3.2.4 Other Light Isotopes 490
3.2.5 Heavier Stable Isotopes 490
4. **Conversion of UF_6 to Other Species** 490
4.1 Conversion of UF_6 to UO_2 490
4.2 Conversion of UF_6 to Uranium Metal 490
5. **Fuel Fabrication and Refabrication** 491
5.1 Oxide Pellet Fuels 491
5.2 Metal Fuels 491
6. **Effects of Fuel Irradiation** 492
6.1 Irradiation Processes and Products 492
6.2 Distribution and Properties of Fuel Irradiation Products 494
6.2.1 Distribution of Irradiation Products 494
6.2.2 Properties of Irradiation Products 494
7. **Spent-Fuel Reprocessing** 495
7.1 Purex Reprocessing 496
7.1.1 Head End 496
7.1.1.1 Remove Nonfuel Components 496
7.1.1.2 Segmentation of Fuel 496
7.1.1.3 Dissolution 496
7.1.2 Product Recovery 497
7.1.2.1 Primary Decontamination 497
7.1.2.2 Partitioning 497
7.1.2.3 Solvent Extraction Equipment 497
7.1.3 Product Purification 498
7.1.3.1 Uranium Purification 498
7.1.3.2 Plutonium Purification 498
7.1.4 Product Conversion 498
7.1.4.1 Uranium Conversion 498
7.1.4.2 Plutonium Conversion 498
7.1.4.3 Coconversion 498
7.2 Reprocessing Variations 498
7.2.1 Purex Variations 499
7.2.2 Aqueous Reprocessing Variations 499
7.2.3 Nonaqueous Reprocessing Variations 499
7.2.3.1 Pyroprocesses 499
7.2.3.2 Fluoride Volatility Processes 500
7.2.4 Continuous Reprocessing 500
8. **Storage and Transportation of Nuclear Materials** 501
8.1 Storage 501
8.1.1 Water Pools 501
8.1.2 Dry Storage 502
8.2 Transportation 502
8.2.1 Regulations 502
8.2.2 Shipping Casks 502
9. **Beneficial Use of Radioisotopes** 504
9.1 Production and Recovery of Radioisotopes 504
9.1.1 Production by Irradiation 504
9.1.2 Production by Decay 504
9.1.3 Recovery of Radioisotopes 505
9.2 Useful Radioisotopes and Their Applications 505
9.2.1 Medical Diagnostics and Treatment 505

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

9.2.1.1 Diagnosis 505
9.2.1.2 Treatment 505
9.2.2 Industrial Applications 505
9.2.2.1 Sterilization 505
9.2.2.2 Radiation-Induced Reactions 506
9.2.2.3 Measurement and Testing ... 506
9.2.3 Basic Research in the Physical Sciences 506
9.2.4 Radiotracer Applications 506
9.2.5 Power Production 506
9.2.6 Military Applications 507
Glossary 507
Works Cited 510
Further Reading 510

INTRODUCTION

This article addresses the necessary or desirable operations that are performed to prepare the nuclear fuel required to sustain nuclear reactors and to manage the fuel constituents after removal from the reactor. Taken collectively, these operations are called a *nuclear fuel cycle*.

These operations include the following:

1. obtaining raw materials to make fuel;
2. processing the fuel materials to achieve the desired chemical and isotopic composition;
3. fuel manufacture;
4. processing of used fuel to recover valuable constituents, including optional recovery of radioactive isotopes for beneficial uses; and
5. transportation and storage of fuel materials at various facilities.

The design and operation of the nuclear reactors and the management of nuclear wastes are described in separate articles in this series entitled NUCLEAR ENERGY, FISSION, and NUCLEAR WASTE MANAGEMENT.

1. OVERVIEW

Since the practical implementation of nuclear power during World War II, many different types of nuclear reactors have been conceived, and many of these have been built. For most of these, multiple fuel cycles are possible, yielding a large number of reactor–fuel cycle combinations. However, for descriptive purposes, these can be grouped into three categories: once through, reprocessing and recycle, and continuous recycle. These three fuel cycle categories are depicted in Fig. 1.

A *once-through fuel cycle* involves the production of fissile material having a concentration and total mass adequate to sustain a nuclear chain reaction. This material is then fabricated into fuel assemblies that have optimal nuclear physics properties, provide for removal of the heat generated during fission, and keep the radioactive products of fission separated from the coolant. This fuel (called fresh fuel) is transported to the reactor, where it is inserted into a core containing many assemblies and used to generate heat, which then generates high-pressure steam that is used to generate electricity. At some point, the fissile material is sufficiently depleted so that fission will no longer occur (i.e., the reactor is subcritical). The oldest fuel in the reactor core is then discharged and stored in a manner that provides protection from its intense radiation and heat removal. At some point, the spent fuel will be packaged and sent to a deep-mined cavern (i.e., a repository) for permanent disposal.

The spent fuel from a once-through fuel cycle contains significant amounts of residual fissile material. In a fuel cycle involving *reprocessing and recycle*, the fuel is reprocessed to recover the fissile material, and this material is refabricated into recycle fuel for use in another reactor. The most promising of these cycles involves recovering the slightly enriched uranium and plutonium from the spent fuels. The uranium would be re-enriched to concentrate the fissile isotope ^{235}U. The plutonium (produced by irradiating ^{238}U with neutrons) is typically composed of about $\frac{2}{3}$ fissile isotopes and would be diluted with natural uranium to achieve the proper fissile enrichment, which can range from 3 to 25%. Systems have also been proposed in which ^{232}Th is irradiated to generate fissile ^{233}U, which is then diluted with ^{232}Th to form recycle fuel. The fuel assemblies made from these materials are then inserted into the reactor and used to generate electricity. Upon

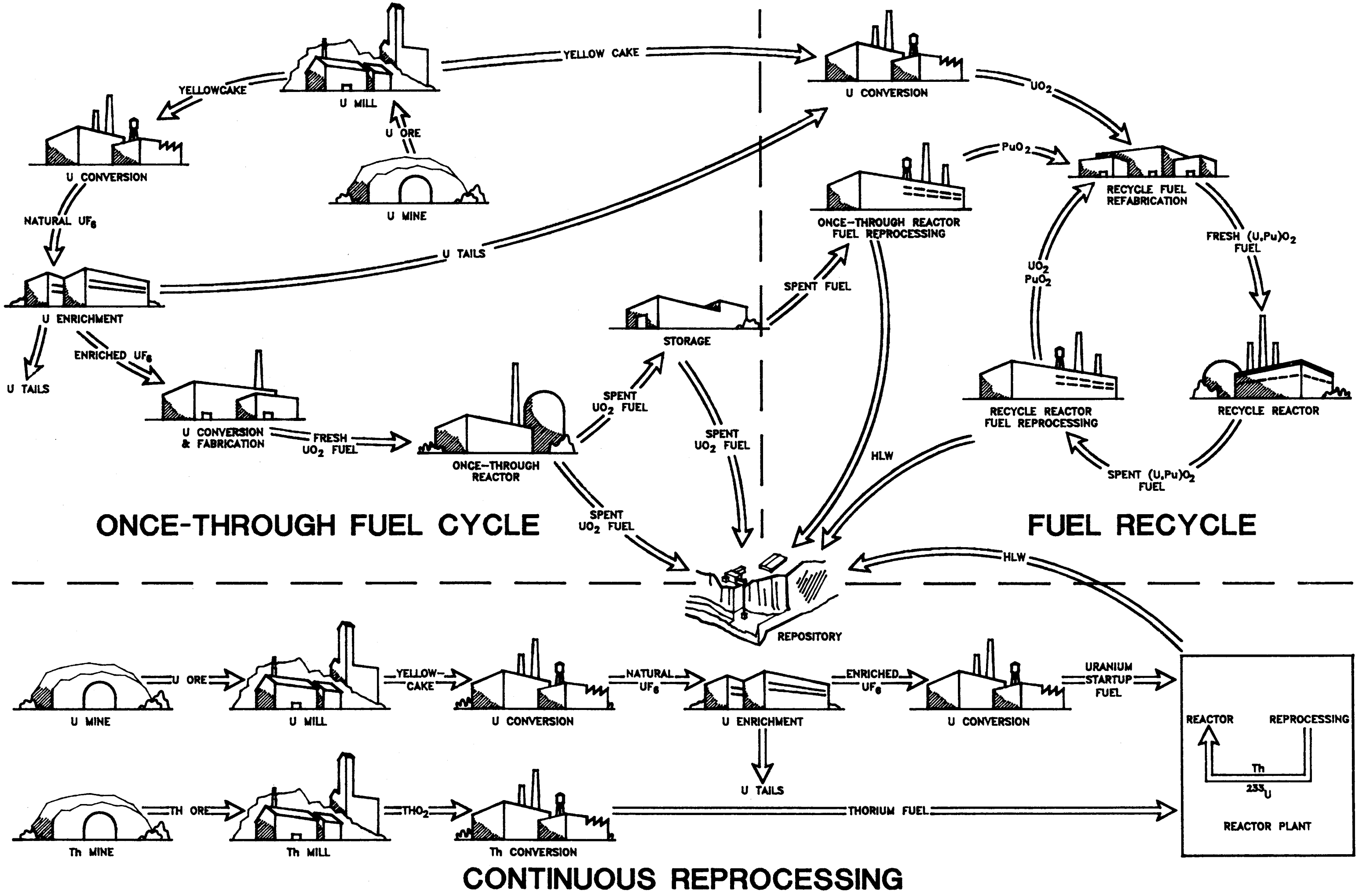
YELLOWCAKE
U MILL
YELLOW CAKE
U CONVERSION
U ORE
U MINE
U CONVERSION
NATURAL UF$_6$
U TAILS
U ENRICHMENT
ENRICHED UF$_6$
U TAILS
U CONVERSION & FABRICATION
FRESH UO$_2$ FUEL
ONCE-THROUGH REACTOR
SPENT UO$_2$ FUEL
STORAGE
SPENT FUEL
SPENT UO$_2$ FUEL
SPENT UO$_2$ FUEL
ONCE-THROUGH REACTOR FUEL REPROCESSING
HLW
UO$_2$
PuO$_2$
RECYCLE FUEL REFABRICATION
FRESH (U,Pu)O$_2$ FUEL
UO$_2$
PuO$_2$
RECYCLE REACTOR FUEL REPROCESSING
RECYCLE REACTOR
SPENT (U,Pu)O$_2$ FUEL
ONCE-THROUGH FUEL CYCLE
FUEL RECYCLE
HLW
REPOSITORY
U MINE
U ORE
U MILL
YELLOW-CAKE
U CONVERSION
NATURAL UF$_6$
U ENRICHMENT
ENRICHED UF$_6$
U CONVERSION
URANIUM STARTUP FUEL
U TAILS
REACTOR
REPROCESSING
Th
^{233}U
REACTOR PLANT
Th MINE
TH ORE
Th MILL
THO$_2$
Th CONVERSION
THORIUM FUEL
CONTINUOUS REPROCESSING

Fig. 1. Relationship of operations in once-through, recycle, and continuous reprocessing nuclear fuel cycles.

discharge, they are again reprocessed, and the cycle is repeated. Depending on the specific features of the reactor design, the overall recycle loop may be a net producer (i.e., a breeder) or net consumer (i.e., a converter) of fissile material.

In a fuel cycle involving *continuous recycle*, the fuel is dissolved or suspended in a liquid such as water, molten salt, or molten metal. The liquid is continuously circulated through a loop. At some point in the loop, its configuration is such that it is critical, generating the desired heat. At other points, the heat is removed to generate steam, and, at still other locations, the fuel is continuously reprocessed to remove unwanted by-products. Many such systems have been proposed since the 1940s, but they have not been well developed as compared to the batch systems described above.

The remainder of this article will more specifically address most of the steps depicted in Fig. 1. Exceptions to this are descriptions of specific types of nuclear reactors and management of radioactive wastes, both of which are the subject of separate articles (see NUCLEAR ENERGY, FISSION; NUCLEAR WASTE MANAGEMENT).

2. URANIUM MINING AND MILLING

2.1 Uranium Sources and Mining Techniques

The only naturally occurring fissile isotope is ^{235}U, which constitutes 0.71% of natural uranium. It is this material that is the wellspring of all nuclear fuel cycles. Uranium is a ubiquitous element, being widely distributed in the earth's crust at an average concentration of about 0.0003%. At this concentration, recovery is not economic. However, as a result of geologic and environmental processes, uranium has been concentrated in veins at many locations. A few high-grade veins containing the mineral pitchblende (U_3O_8) have been found to contain uranium at concentrations of 30–70% U_3O_8, mostly in Africa. Significant ore bodies having uranium concentrations up to 15% are now being developed in Canada. However, these are relatively uncommon, and most of the high-concentration veins were mined decades ago. Much more common are veins of carnotite ($K_2O \cdot 2UO_3 \cdot V_2O_5 \cdot 3H_2O$) ore that contain 0.1 to 0.2% uranium, typically in a sandstone matrix. These are widely distributed in the western United States and Canada in a band covering the eastern to middle portion of the Rocky Mountains, and in similar geologies in other countries. It is economically feasible to mine carnotite ores at uranium concentrations of 0.1% or above. Total known world resources of uranium outside the centrally planned economies are estimated to be 3.1 million metric tons (OECD, 1992).

The mining of carnotite ores is accomplished using a variety of methods. Standard tunnel mining has often been used, although less commonly so now because of the relatively high cost of such operations in combination with low uranium prices. Open-pit mining is often used for deposits relatively near the surface. A newer technique, called *in situ* leaching, involves injecting leaching solutions into a subsurface ore zone through a well and then drawing the uranium-bearing solution back to the surface through another well. The leaching chemicals are very similar to those used in uranium ore milling, described below.

Other dilute sources of uranium are also economic to mine when the uranium is a by-product of the production of other economic values. Significant amounts of uranium are generated as a by-product of phosphate mining. Most of this production occurs in the Gulf Coast states in the United States, and similar locations elsewhere. Uranium is also produced as a by-product of copper and gold mining at various locations throughout the world.

Recovery of uranium from seawater has been and continues to be studied, typically using ion-exchange technology. While the "ore" is easily accessible, the low concentration of the uranium and fouling of the ion exchange material continue to make this an uneconomic source.

2.2 Uranium Milling and Refining

Uranium milling is the concentration and purification of the uranium in ores. Ores are typically ground to 20 to 200 mesh to provide adequate contact with reagents. The ground ore is then contacted with an aqueous acid or alkaline leach, depending on the specific properties of the ore. Sulfuric acid is most commonly used, but all of the mineral acids are used to some extent. Alkaline leaching is performed with solutions of the carbonates and bicarbonates of sodium or ammonia. The acid-based process is generally more economic unless the ore contains large amounts of alkaline materials, which would require excessive acid consumption. In either case, the result is a selective dissolution of uranium from the ore.

The acid or alkaline slurry of uranium-bearing solution and solids is then filtered to remove the solids. These solids are somewhat radioactive, containing all of the decay products of the uranium that have achieved equilibrium levels during millions of years of decay. Most noteworthy of these are ^{226}Ra (with a half-life of about 1600 years) and its chemically inert gaseous decay product ^{222}Rn. These radionuclides are primarily responsible for the radiological effects of the tailings. The tailings are typically managed by simply creating small hills of the solids near the mill, along with engineered features to manage the runoff resulting from rain and to control the release of ^{222}Rn.

The uranium is recovered from the aqueous leach solution by precipitation using a variety of chemicals such as ammonia, magnesia, or caustic soda. The product of the precipitation is the diuranate of ammonia, magnesium, or sodium, with a typical chemical formula being $(NH_4)_3U_2O_7$. Depending on the specifics of the process, this product is then washed and dried to yield a 75-to 100%-pure U_3O_8 product. Although all of the diuranate intermediates and the impure product are commonly called *yellowcake*, only ammonium diuranate has the characteristic yellow color and these products are more properly called *uranium ore concentrates* (UOC).

2.3 Purification and Conversion of UOC to UF_6

The impure UOC product must be purified to be acceptable in subsequent processes. This is generally performed at a centralized facility separate from the mine and mill, and this same facility often includes the processes for conversion to UF_6. Purification is typically accomplished by dissolving the yellowcake in concentrated nitric acid and extracting the uranium from the acid by *solvent extraction*: contacting it with immiscible tributylphosphate (TBP) diluted with kerosene or *n*-dodecane and then back-extracting the uranium from the TBP solution using dilute nitric acid. Purification may also be accomplished using other solvent extraction reagents or ion exchange techniques. The uranium solution is then typically evaporated and calcined (to yield high-purity UO_3) and then reduced by hydrogen gas to yield uranium dioxide (UO_2) and nitric acid for internal recycle.

Uranium dioxide can be taken directly to fresh fuel fabrication for reactors that do not require isotopically enriched uranium. However, many reactors require uranium enriched so that it has a ^{235}U concentration of 1–5%, with the required concentration depending on the reactor design and operating parameters. All currently operating enrichment processes require the feed material to be in the form of uranium hexafluoride (UF_6), thus requiring *conversion* of the oxide to the hexafluoride. This conversion is accomplished in two steps. The first is to react the UO_2 with anhydrous hydrogen fluoride (HF) in a multiple-stage reactor, which yields uranium tetrafluoride, UF_4. The UF_4 is then reacted with elemental fluorine gas (F_2) to yield UF_6.

Purification and conversion can be carried out simultaneously by fluorinating the UOC directly by first using the HF-F_2 sequence described above. The result is that the uranium and a few other elements form fluorides that are easily volatilized, leaving most of the impurities in a solid residue. The gaseous fluoride stream is then further purified by fractional distillation to yield a pure UF_6 product.

3. ISOTOPE ENRICHMENT

The function of *isotope enrichment* processes is to increase the relative abundance of selected isotopes in an element. This also necessarily produces another output stream (typically called *tails*) that is depleted in the particular isotope. In some cases, the degree of enrichment is almost complete. For example, to produce "heavy water," the relative abundance of deuterium ($^2H_2$) is increased from the very low 0.015% natural abundance in water to about 99.75%. In other cases, the concentration of a particular isotope may be increased only a few percentage points. For example, uranium is enriched from its natural 0.71% ^{235}U concentration to less than 5% for most reactor applications.

3.1 Uranium

Uranium is the most extensively enriched material in the world. The first successful large-scale enrichment of uranium isotopes was achieved during World War II by the United States in its effort to obtain enough highly enriched (>90% ^{235}U) uranium to produce a nuclear weapon. A large number of uranium isotope separation methods were studied during this period, and most were able to enrich uranium to varying degrees. However, the cost of enrichment was very high for many of these processes, and less costly alternatives were subsequently developed. The following discussion describes uranium enrichment processes in their approximate order of large-scale development or use.

3.1.1 Electromagnetic Isotope Separation (EMIS) EMIS capitalizes on the fact that charged particles move in circular paths when subjected to a magnetic field perpendicular to the particle path. Different paths are followed by charged particles of different masses, with the path of a heavier particle having a slightly larger diameter. Thus, when ionized (charged) atoms such as those of uranium are injected into a strong magnetic field, the ^{235}U ions follow a different path than the ^{238}U ions, and the former may be collected in a different location from where the latter impinge (Fig. 2). Recovery of the separated material is then simply a matter of chemical processing.

The separation factor (a measure of the amount of enrichment attained with a single enrichment stage) is very large for the EMIS process. In theory, complete isotopic separation can be achieved in a single stage. In practice, the extent of enrichment attained is limited by ion beam dispersion, interactions with solid materials, and other second-order effects. However, highly enriched uranium can be produced with only two consecutive stages of EMIS.

The EMIS machines used during World War II in the United States to enrich uranium were called *calutrons*. Although the enrichment achieved in a single stage was high, the throughput of any one device was relatively small. As a result, a large number of calutrons were operated in parallel to produce the required amount of product (several tens of kilograms).

Although the calutrons were successful, they were expensive to operate because they used large amounts of electricity for the amount of enriched uranium produced and because they were very labor intensive. As a result, use of this technology for uranium enrichment ceased soon after the War, and most of the calutrons were dismantled.

3.1.2 Gaseous Diffusion Another method of uranium enrichment studied during World War II—*gaseous diffusion*—proved to be much less costly than EMIS, requiring far fewer people, operations, and units of electricity per unit of product. In the gaseous diffusion process, UF_6, which is gaseous at relatively low temperatures and pressures, is allowed to diffuse through very small openings in porous tubes called *barriers* contained within large cylindrical housings. The $^{235}UF_6$ molecules go through the pores a little faster than the heavier $^{238}UF_6$ molecules, and so the gas that passes through the barrier is slightly enriched in the lighter isotope. The slightly enriched product from this stage is then pumped to the next-higher stage, where the process is repeated. The gas that does not pass through the tubes is slightly depleted in $^{235}UF_6$ and is pumped to the next lower stage, where the process is repeated. In this way, the enriched uranium is moved continuously upward and repeatedly enriched in ^{235}U, while the depleted uranium is moved continuously downward and repeatedly depleted in ^{235}U. The assemblage of interconnected stages is called a *cascade*. Along with the housing and tubes (called a *convertor*), each stage also requires a compressor to recompress the gas received from other stages and a heat exchanger to remove the heat of compression. Such an arrangement is shown in Fig. 3.

As implied above, the amount of enrichment achieved by a single stage is small. In fact, the maximum theoretical amount by which the ^{235}U can be enriched in a single stage is given by the square root of the ratio of the masses of the heavier and lighter uranium molecules, which is $(352/349)^{0.5} = 1.004$. As a consequence, to enrich natural uranium containing 0.71% ^{235}U to 3% (an enrichment typically required by reactors) while produc-

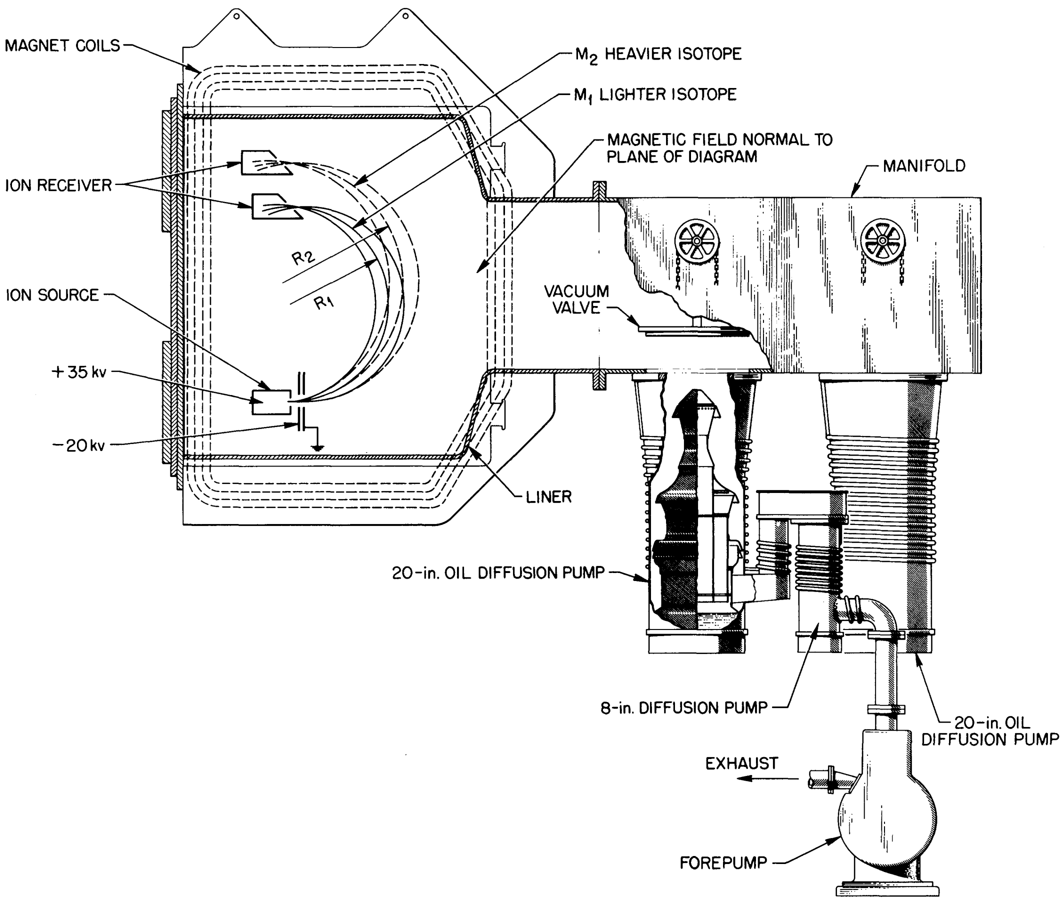
MAGNET COILS
M_2 HEAVIER ISOTOPE
M_1 LIGHTER ISOTOPE
MAGNETIC FIELD NORMAL TO PLANE OF DIAGRAM
MANIFOLD
ION RECEIVER
R2
R1
ION SOURCE
VACUUM VALVE
+35 kv
−20 kv
LINER
20-in. OIL DIFFUSION PUMP
8-in. DIFFUSION PUMP
20-in. OIL DIFFUSION PUMP
EXHAUST
FOREPUMP

Fig. 2. Schematic diagram of electromagnetic isotope separation (a Calutron).

ing a tails stream containing 0.2% ^{235}U would require 1272 theoretical stages. This number is increased by the inevitable inefficiencies in any real process.

Despite the large number of stages required, the large physical size of the plants (which are multi-story buildings about 1 km long and 100 m wide), and the relatively large amounts of electricity required (a single plant requires on the order of 1000 MW), the large throughput of the plant and the low operation and maintenance cost resulted in the product cost being sufficiently low to make nuclear fuels economic. Gaseous diffusion plants for enriching uranium for both power reactors and military purposes have been built in the United States, Russia, France, Great Britain, and China. Construction of additional plants is not expected because the high electricity consumption makes the product relatively expensive compared to other technologies that have become available.

3.1.3 Gas Centrifuge A *gas centrifuge* separates uranium isotopes by placing UF_6 in an artificially created high-gravity environment, thus magnifying the importance of the small mass difference between ^{235}U and ^{238}U. The high-gravity environment is created by spinning a right-circular cylinder called a *rotor* around its axis at very high speeds, conceptually similar to the spin cycle on a washing machine. For a given rotor diameter, the separation factor is a function of the rate at which the centrifuge rotates. This rate is limited by the mechanical properties of the ro-

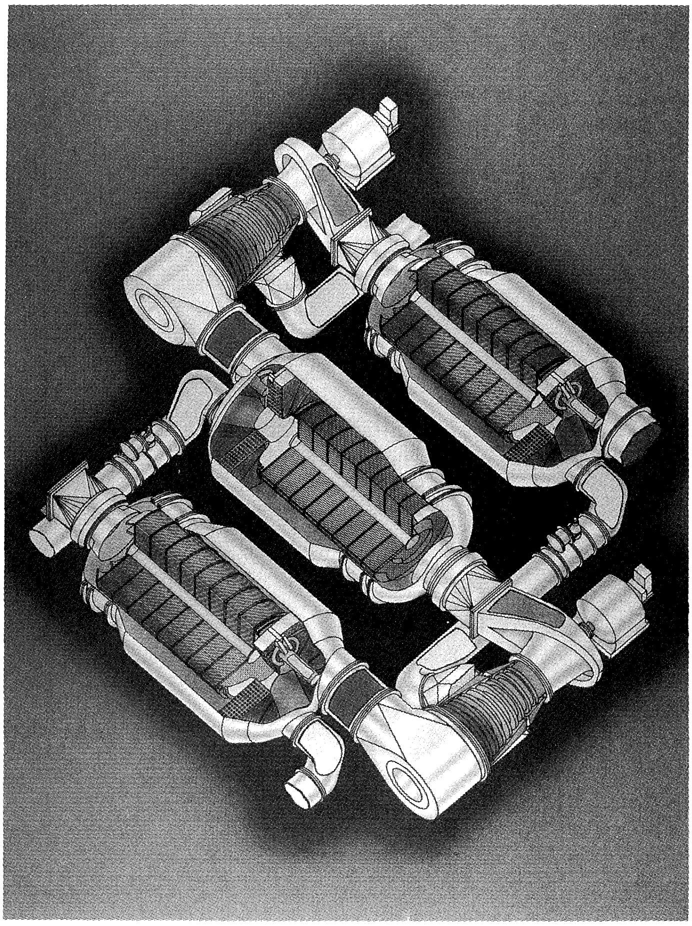

Fig. 3. Diagram of interconnected stages in a gaseous diffusion plant cascade.

tor materials, which may be advanced metal alloys or composites.

Because of the circulating flow within the centrifuge, each functions as a "miniature cascade," with multiple effective stages within each physical device. A schematic of one of the early centrifuge designs depicting this circulation is shown in Fig. 4. The net result is that the number of physical stages (i.e., centrifuge machines) required is far fewer than that required for gaseous diffusion, although a number of centrifuges may have to be operated in parallel to achieve the desired throughput. As a result of the processes within the centrifuge being reversible (the pressure drop in gaseous diffusion is irreversible), the electricity requirement for gas centrifuges is estimated to be about 4% of that for gaseous diffusion. However, the centrifuges are very sophisticated devices, requiring advanced materials to achieve economic operating speeds without destroying themselves. On balance, the cost of product from the gas centrifuge process appears to be less expensive than that from gaseous diffusion, but not greatly less.

Gas centrifuge plants for uranium enrichment are currently operating in the Netherlands, Great Britain, Germany, and Russia. Plants are under construction in other countries. A small pilot plant was built and operated in the United States, and construction started on a demonstration plant, but the project was cancelled in favor of a laser isotope separation process.

3.1.4 Laser Isotope Separation (LIS) A *laser isotope separation* (LIS) process separates uranium isotopes by capitalizing on slight differences in the wavelength of light

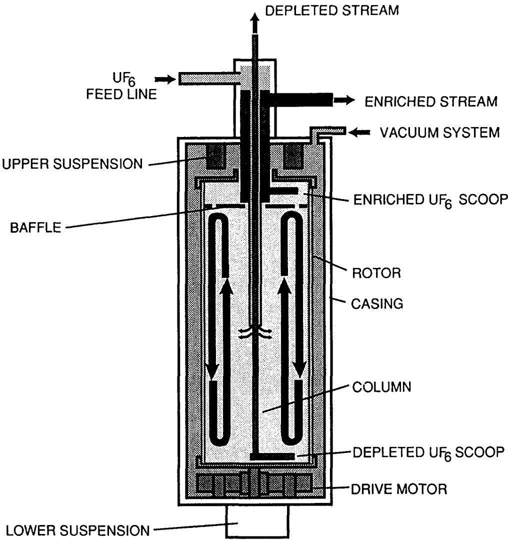
DEPLETED STREAM
UF_6 FEED LINE
ENRICHED STREAM
VACUUM SYSTEM
UPPER SUSPENSION
ENRICHED UF_6 SCOOP
BAFFLE
ROTOR
CASING
COLUMN
DEPLETED UF_6 SCOOP
DRIVE MOTOR
LOWER SUSPENSION

Fig. 4. Schematic diagram of a gas centrifuge for uranium enrichment.

that will excite uranium metal or uranium-bearing molecules to a species that can be readily separated. The variation in wavelengths is a result of the differing atomic properties of ^{235}U and ^{238}U that are a consequence of the difference in the masses of the two nuclei. The laser power requirements are significant (multikilowatt to megawatt). It was only with advanced developments in high-power lasers that the wavelengths could be adequately controlled to achieve the desired enrichment while producing the power required for the needed throughput.

As with EMIS, the selectivity of LIS processes is very high. In theory, only one stage would be required to separate the isotopes completely, while in practice, only two stages may be needed. The process requires much less electricity than EMIS. However, the cost of the equipment (especially the high-powered lasers) required for LIS and the precision with which it must be controlled is high. While this process is still under development, it is thought that the initial cost of product will be about the same as that from gas centrifuges.

The *atomic vapor laser isotope separation* (AVLIS) process (Fig. 5) is carried out by introducing uranium metal (or uranium metal alloy) vapor into a sealed vessel under low-pressure conditions. The uranium is exposed to laser light at one or more suitable wave-

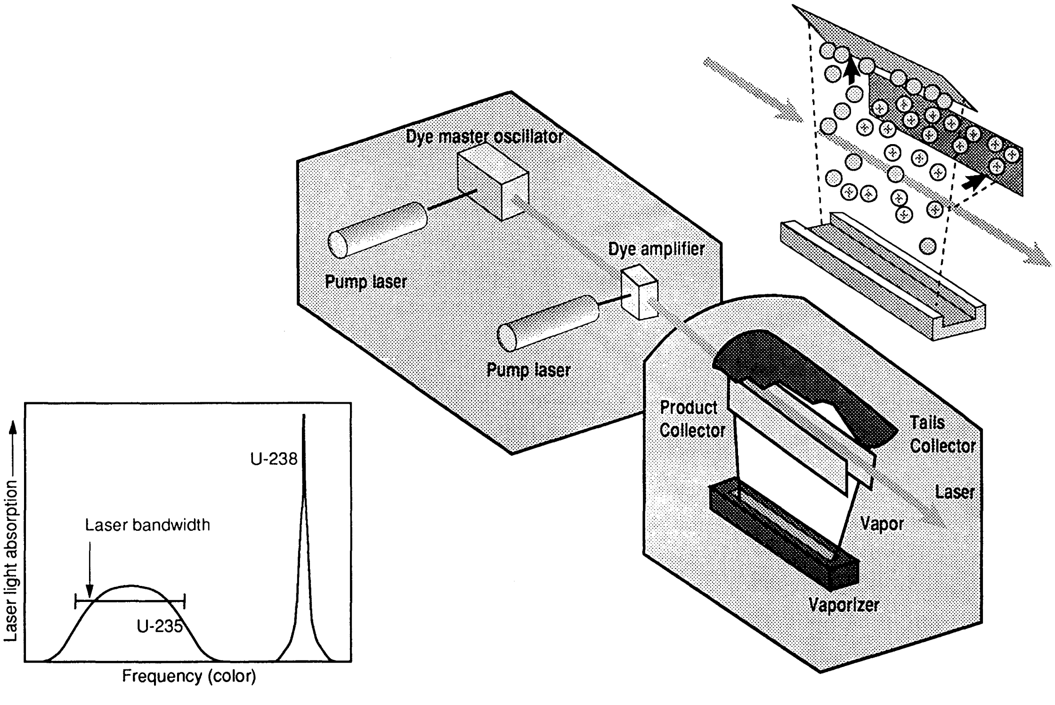
Dye master oscillator
Pump laser
Dye amplifier
Pump laser
Product Collector
Tails Collector
Laser
Vapor
Vaporizer
Laser light absorption
U-238
Laser bandwidth
U-235
Frequency (color)

Fig. 5. Diagram of a laser isotope separation module.

lengths, resulting in the selective ionization of the ^{235}U fraction. The ionized material can then be diverted from the un-ionized (^{238}U) material with an electrostatic field and collected.

The *molecular laser isotope separation* (MLIS) process is conceptually similar to AVLIS. However, in this case, the feed material is the molecule UF_6. Upon exposure to laser light of suitable wavelengths, the UF_6 is dissociated by photolysis to form UF_5 and atomic fluorine (which may quickly combine to the more stable F_2 molecule). The UF_5 is a solid entrained within the flowing UF_6 gas, and can be separated from the gas stream by standard techniques such as by filtration with gaseous-diffusion–plant barrier material. Rapid separation of the UF_5 is required to prevent recombination with the atomic and molecular fluorine present in the gas stream.

3.1.5 Aerodynamic Processes Aerodynamic processes capitalize on the centrifugal forces resulting from forcing a gas (in this case, a mixture of UF_6 and hydrogen) through a small, curved nozzle shown in Fig. 6 (Becker *et al.* 1975) at high speed. The heavier fraction tends to be concentrated near the periphery of the nozzle, allowing a knife edge to separate it from the lighter fraction. The enriched and depleted streams from each stage are moved to other stages in a manner very similar to that used in gaseous diffusion.

As with gaseous diffusion, aerodynamic processes are inherently irreversible. As a consequence, the cost of enriched product appears to be at least as large as that of gaseous diffusion and higher than that of gas centrifuges. Small production plants based on these processes have been built and operated in Brazil and South Africa.

3.2 Stable Isotopes

A wide variety of separated *stable isotopes* (i.e., nonradioactive) are used for various aspects of nuclear power, in scientific experiments, and as precursors for production of radioactive isotopes (via irradiation in particle accelerators or small reactors) used in an expanding variety of medical procedures.

3.2.1 Deuterium Deuterium (^2H or D) is a heavy isotope of hydrogen containing one proton and one neutron in the nucleus as compared to the single proton in normal hydrogen (^1H). Deuterium occurs naturally (DHO at low concentrations, D_2O at high concentrations) in concentrations of about 0.015 at.% in all water. While it has some minor scientific uses, by far the predominant use of heavy water is as a neutron moderator in some types of nuclear reactors such as the Canadian CANDU. Deuterium's nuclear properties result in its having an exceedingly small chance of undergoing undesirable neutron absorp-

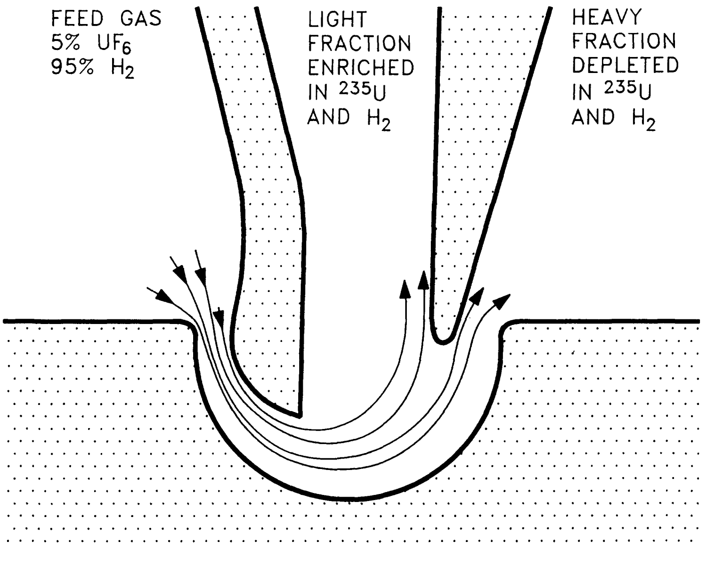

Fig. 6. Cross section of a basic separation nozzle used for uranium enrichment.

tion reactions, thus allowing a practical reactor to be operated using only natural (i.e., not enriched) uranium fuel. However, the savings in uranium enrichment costs are approximately balanced by the cost of having to enrich the deuterium to nearly 100% and the fact that a 600 MW_e CANDU reactor requires an inventory of about 450 metric tons of heavy water.

Many processes have been proposed and used since the first known heavy-water production plant was operated in 1934. Early plants were typically based on electrolysis of water, which, like the EMIS process for other stable isotopes, was effective but costly. Over the years, a wide variety of alternative processes have been developed and deployed to produce heavy water. The two general categories of processes that have been developed involve exchange of D and H between various chemicals such as ammonia, water, or hydrogen sulfide, and distillation of various chemicals such as water, ammonia, or methane.

After decades of experience, the preferred approach to producing essentially pure heavy water is the Girdler-sulfide, dual-temperature water–hydrogen-sulfide exchange process (GS) followed by water distillation. A flowsheet of the GS process is exceedingly complex because it involves extensive measures to conserve heat within the process. However, in conceptual terms, a tall vertical vessel is fed at the top with natural water at about 0° C. This is contacted with a gaseous stream of D_2S injected at the bottom of the vessel. The net result is that the deuterium preferentially moves to the aqueous phase (which becomes heavy water) and hydrogen to the gaseous phase. The D_2S is generated in a separate vessel by reacting part of the heavy water product with H_2S at a much higher temperature, where the formation of D_2S is favored, producing water depleted in deuterium. The sulfide compound is continuously recycled within the process. The product enrichment of this process is typically about 15 at.% deuterium.

The second step of this approach involves taking the partially enriched product from the GS process and fully enriching it in a simple water distillation column. The lighter H_2O favors the gaseous (steam) phase. Thus, injecting the partially enriched liquid product from GS into the top of a vessel which has H_2O steam rising from the bottom has the effect of stripping the rest of the H_2O from the liquid, yielding the desired high-purity product.

3.2.2 Lithium Lithium has two stable isotopes. Lithium-7 (abundance of 92.5 at.%) has a relatively low probability of undergoing undesirable interactions with neutrons, and its compounds have found limited use in controlling the water chemistry in nuclear reactors. ^{6}Li has a relatively large probability of undergoing interactions with neutrons, with the products being stable helium and radioactive ^{3}H (called tritium). Tritium is used in industry in a variety of applications related to radioluminescence. Tritium production does not require that the precursor material be enriched in ^{6}Li, and so the demand for enriched lithium products has been limited to the small demands for ^{7}Li and enrichment activities related to the production of ^{6}Li for nuclear weapons.

As with heavy water, a wide variety of processes have been investigated for the purpose of enriching lithium. In fact, the general categories of processes are the same as those for heavy water production (i.e., electrolysis, distillation, and chemical exchange), along with ion exchange processes. As a result of the limited demand for enriched lithium, the preferred process is not evident. However, there appears to be a tendency to favor a chemical exchange process involving contact of an aqueous solution of LiOH with a lithium amalgam (a solution of lithium in mercury).

3.2.3 Boron Boron has two stable isotopes. ^{10}B (abundance of 19.6 at.%) has an extremely high probability of undergoing reactions with neutrons, and it is used as a neutron absorber to control the chain reaction in nuclear reactors. Worldwide production is on the order of a few tons per year. ^{11}B has a relatively low probability of undergoing nuclear reactions. Although only a few small plants have been built, the preferred process involves a chemical exchange reaction between boron trifluoride (BF_3), which is gaseous at modest temperatures, and a complex of boron trifluoride and dimethyl ether [$(CH_3)_2O$], which acts as the liquid phase.

3.2.4 Other Light Isotopes There is also limited enrichment of the low-abundance, stable isotopes of carbon (^{13}C, abundance 1.11 at.%), nitrogen (^{15}N, abundance 0.365 at.%), and oxygen (^{17}O, abundance 0.037 at.%, and ^{18}O, abundance 0.204 at.%). All of these are used in scientific research to make biological compounds that can be traced through organisms and chemical reactions because of their enrichment. As with the other isotopes mentioned above, the preferred production methods involve distillation and chemical exchange. Worldwide production is estimated to be several kg/yr or less for these isotopes.

3.2.5 Heavier Stable Isotopes Beyond the stable isotopes listed above, demand for enriched products is typically small and depends on the identification of a specific use for the isotope. Most elements have been isotopically enriched in small amounts to provide relatively pure materials upon which to perform experiments concerning the unique nuclear properties of each. Other selected isotopes continue to be separated in larger quantities to provide precursor material to produce useful radioactive isotopes, especially for medical applications.

Although a few specialized enrichment processes have been used, most of the heavier elements are enriched using the EMIS process described earlier. While it is expensive to operate for the purpose of producing large quantities of enriched uranium, its flexibility and high enrichment per stage make it cost-effective for small scale separations.

4. CONVERSION OF UF_6 TO OTHER SPECIES

The fact that UF_6 is gaseous at modest temperatures makes it very useful for enrichment processes. However, this compound is not useful for the end applications of the uranium. As a result, the UF_6 must be converted to a chemical species that has properties appropriate to its end use.

4.1 Conversion of UF_6 to UO_2

Uranium dioxide (UO_2) is the species used to manufacture most nuclear reactor fuels, and it has few uses beyond this application. Three different processes have been developed to perform this function. All are used commercially.

The first method of production involves two steps. The first is gas-phase reduction of UF_6 by elemental hydrogen, yielding UF_4 and HF. The UF_4 is then reacted with steam, yielding UO_2 and a solution of HF.

The second method involves direct reaction of UF_6 with water to form UO_2F_2 and aqueous HF. This solution is then reacted with ammonium hydroxide (dissolved ammonia) to precipitate ammonium diuranate [$(NH_4)_2U_2O_7$) and ammonium fluoride. The ammonium diuranate is then filtered from the solution, dried, and reduced with hydrogen gas to yield ammonia and UO_2.

The third method involves reaction of the UF_6 with carbon dioxide and ammonia in aqueous solution to produce a precipitate of ammonium uranyl carbonate, $(NH_4)_4UO_2(CO_3)_3$. The ammonium uranyl carbonate is converted to UO_2 by reaction with steam and hydrogen gas at elevated temperatures.

In all cases, the characteristics of the product must be carefully controlled. Not only is the chemical composition of the product important, but its physical attributes such as grain size and sinterability are also very important in subsequent manufacture of fuel.

4.2 Conversion of UF_6 to Uranium Metal

Uranium metal has a variety of uses. It is used in nuclear fuels, typically alloyed with other metals (e.g., aluminum, zirconium) to provide acceptable mechanical properties. It is used to manufacture armor-piercing bullets for the military. The high density (19.1 g/cm^3) and pyrophoric nature of uranium metal when finely ground or molten make it preferable to lead (11.7 g/cm^3) or other materials in some applications. Uranium is also used to form radiation shields, especially for gamma rays. It has occasionally been used in other applications, such as counterweights in airplanes, where space is at a premium and its high density makes it cost-effective.

Only one process is used for the large-scale production of uranium metal. The first step is to reduce the UF_6 to UF_4 with hydrogen gas, as described in the previous section. The UF_4 is then mixed with a metal more reactive than uranium, typically magnesium or calcium, and an iodine catalyst. This mixture is placed in a sealed chemical bomb (a sturdy closed vessel), and the mixture is ignited electrically. The result is the reduction of the uranium to molten metal and the production of the molten fluoride of the reducing agent (e.g., CaF_2). The metal and fluoride salt are immiscible, and so, after cooling, the result is a solid layer of the fluoride salt atop a mass of uranium metal, often called a *button*.

5. FUEL FABRICATION AND REFABRICATION

The purpose of fuel fabrication and refabrication is to process raw nuclear materials into a form and configuration that are appropriate for the fuel of a particular reactor design. *Fabrication* involves processing unirradiated nuclear materials (i.e., uranium or thorium). *Refabrication* involves processing of materials recovered from spent fuels for recycle, with plutonium and ^{233}U being the most common examples.

Typically, the nuclear material begins in powder or metal form. This material is formed into larger shapes, generically called *fuel*, to withstand irradiation in the reactor better and to facilitate handling. The fuel is then enclosed in a metal container, generically designated as *cladding*, which serves the purpose of separating the radioactive fuel from the relatively clean fluid (e.g., water, liquid sodium, CO_2, gaseous helium) required to cool the fuel during irradiation. The units of clad fuel are called *fuel rods* or *fuel elements*. The fuel elements are then arranged into a fixed geometric array, called a *fuel assembly*. Dozens, if not hundreds, of fuel-cladding-assembly combinations have been developed over the last five decades of nuclear power. The following sections describe two of the more preferred combinations that have received some degree of large-scale use.

5.1 Oxide Pellet Fuels

By far the most popular fuel type is formed from uranium dioxide produced as described earlier. Most large reactors that produce electric power use this type of fuel, primarily because it is stable at relatively high temperatures. The following steps are typically followed to make oxide fuels:

1. If the fuel is to be composed of oxides of several elements (e.g., Pu-enriched uranium dioxide or $[Pu,U]O_2$), the oxides are thoroughly blended to achieve the desired enrichment, which would typically range from a few percent Pu (for thermal reactors) to 25% (for fast reactors).
2. The oxide powder is blended with binding agents and die lubricants.
3. The mixture is pressed into pellets that are the shape of a right circular cylinder, typically about the diameter of a pencil or somewhat larger.
4. The pellets are sintered at high temperatures to consolidate the powder into a solid pellet.
5. The pellets are inspected, ground to size, and placed in a vacuum to remove atmospheric gasses.
6. The pellets are then loaded into empty cladding tubes, which are typically 1–5 m long and made of Zircaloy (an alloy of zirconium with minor amounts of tin and iron) or stainless steel, and an end cap is attached by welding, yielding fuel rods.
7. The rods are inserted into *grid spacers* (metal squares subdivided into smaller squares sized to hold the rods in place with springs), and the entire assembly is bound together axially with end pieces that are connected with hollow tubes similar to those used for cladding.
8. The assembly is cleaned and readied for transport to the reactor.

A cutaway drawing of a typical fuel assembly is shown in Fig. 7.

5.2 Metal Fuels

A second common fuel type is made of metal and is commonly employed in test and research reactors where achieving high temperatures (which is required to produce electricity efficiently) is irrelevant. Following are the major steps typically required to make such fuel:

1. Uranium dioxide with an appropriate isotopic enrichment (which could range from natural to about 93% ^{235}U) is reduced from the oxide to the metal as described earlier.
2. The uranium metal is often alloyed with other metals (e.g., Al, Zr, Mo, Nb) to improve its physical characteristics.

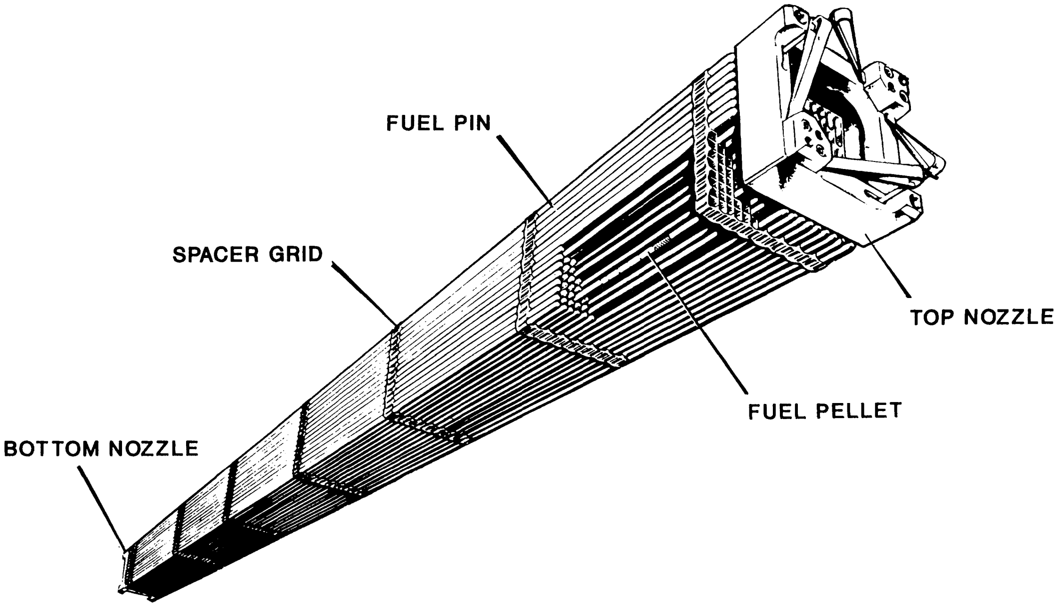

Fig. 7. Major components of a typical fuel assembly from a pressurized water reactor.

3. The alloy is then hot rolled and sized into plates in the range of 0.5–2 mm thick.
4. The plates are placed between two larger plates of cladding metal (typically Al or an Al alloy), which are then hot pressed to seal the edges into a fuel element.
5. The elements are then put into slots in two of four larger flat metal plates that form a box surrounding the plates but open at the ends to allow coolant to pass through. This assembly is typically about 1 m long and 10 cm on each side, although annular designs are also employed and size can vary widely.
6. The assembly is cleaned and prepared for transportation to the reactor.

There is substantial variability in these fuels, with fuel material made of uranium hydrides being relatively common and more advanced materials such as uranium silicides being developed.

6. EFFECTS OF FUEL IRRADIATION

6.1 Irradiation Processes and Products

As noted previously, there is a wide variety of fuel material, cladding material, coolants, and moderators in reactors throughout the world. Despite this diversity, the effects of irradiation are similar in all cases.

Conceptually, a nuclear reactor is a vessel in which a sufficient number of fuel assemblies are placed in close proximity in the *core* and controlled such that a stable fission rate is established in the fuel. Fission generates a significant amount of heat, and one purpose of the vessel is to contain the coolant required to remove the heat from the fuel. Each fission generates about 200 MeV of energy that is recovered as heat, and about 2.5 neutrons. As a consequence, a typical large nuclear power plant generating 1000 MW of electricity has a fission rate of about 9.4×10^{19} fissions/s and produces 2.3×10^{20} neutrons/s.

The 2.5 neutrons produced by each fission reaction are consumed by the nuclei of all elements in the reactor in the following ways:

1. Exactly 1.0 of the neutrons interacts with a nucleus and results in another fission, thus sustaining the chain reaction. This nucleus must be able to be fissioned by the neutrons produced by previous fissions. These neutrons have a spectrum of kinetic energies (i.e., speeds) ranging from fractions of an eV to about 3×10^{6} eV. The only elements that can be fissioned by neutrons having this range of energies are

the actinides. By definition, this includes all elements with atomic numbers greater than 89 (that of actinium). In practice, only a few actinides occur naturally or can be produced in sufficient quantities to constitute a fuel material: Th, U, and Pu. Isotopes such as ^{235}U and Pu^{239} can be fissioned by neutrons of all energies and are called *fissile*. Isotopes such as ^{238}U and ^{240}Pu can be fissioned by only high-energy neutrons and are called *fissionable*. The sum total of the interaction of the 2.5 neutrons with the various actinide isotopes present in the reactor at a given instant results in 1.0 producing a fission that sustains the chain reaction. The other 1.5 neutrons interact as described below, with the proportions of these reactions varying significantly depending on the reactor design.

2. A significant portion of the neutrons interact with the actinides by *neutron capture*. In this case, the actinide is not fissioned. Instead, the neutron combines with the actinide nucleus to form the next heavier isotope of the same actinide element. For example, ^{238}U captures a neutron to form ^{239}U. Uranium-239 has a relatively short half-life and decays to ^{239}Np, which also has a relatively short half-life and decays to ^{239}Pu. During the residence of the fuel, dozens of actinide isotopes are produced by a complex chain of neutron captures and radioactive decays. Many of these are sufficiently long-lived so that they remain in the fuel when it reaches the end of its useful life. All are radioactive, and the dominant decay mechanism is alpha decay, although the decay of some radionuclides results in the emission of neutrons.
3. In the fission process, the actinide nucleus is split into two separate nuclei having much lower atomic numbers. These nuclei are called *fission products*. The split is usually not symmetrical (i.e., the two fission-product nuclei have very different masses). Further, an entire range of fission product elements and isotopes are produced. The distribution of these is shown in Fig. 8 (Katcoff, 1958). As produced, they are all radioactive, and the dominant decay mechanism is beta decay. Many have short half-lives, and these decay to long-lived or stable (i.e., nonradioactive) fission products that are present when the fuel reaches the end of its useful life.

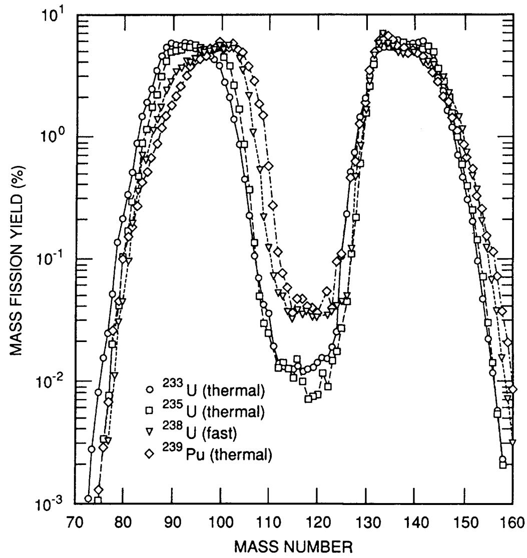

Fig. 8. Fission-product yield as a function of atomic mass number for 0.025 eV-neutron-induced fission of ^{233}U, ^{235}U, ^{239}Pu, and 14 MeV-neutron-induced fission of ^{238}U.

4. Another fate of the 1.5 neutrons not involved in sustaining the chain reaction is to interact with nonactinide nuclei by neutron capture, resulting in *activation products*. The nonactinide nuclei are most often the other materials present in the reactor (e.g., cladding metal, the reactor vessel, or neutron absorbers used to control the chain reaction). The neutron capture process is the same as that described previously for the actinides. Most of the activation products are relatively short-lived and decay to innocuous levels during irradiation. However, a few have longer half-lives and are present in the fuel and reactor structural materials when it reaches the end of its useful life.

At the end of its useful life, the uranium dioxide fuel typically used to generate electricity contains about 1% actinides heavier than uranium (primarily Pu), 3% fission products, a variable but very small amount of activation products, and about 96% uranium isotopes (^{238}U with about 0.8% ^{235}U and 0.4% ^{236}U). The proportions of constituents other than uranium isotopes are increased as the amount of power extracted from a unit of fuel (called *burnup*) increases. The useful life of the fuel is limited by the consumption of fissile isotopes, production of fission product neutron absorbers, and physical degradation of the fuel assembly structural materials (e.g., cladding).

6.2 Distribution and Properties of Fuel Irradiation Products

This section discusses the characteristics of the fuel after it reaches the end of its useful life. At this point, it is removed from the reactor vessel (i.e., *discharged*) and becomes known as *spent fuel*. The characteristics to be discussed are the location of the various products within the fuel assembly and the salient properties of the various products.

6.2.1 Distribution of Irradiation Products

The actinides in the spent fuel are contained almost exclusively within the fuel matrix. "Tramp" fuel on the cladding may result in a small amount being associated with it, but this is usually negligible.

The fission products are mostly contained within the fuel matrix. However, some of the fission product elements are gaseous (e.g., Kr, Xe) or are somewhat volatile at reactor operating conditions (e.g., Cs, I). Some of these elements will be present in nonfuel areas where it is relatively cool, such as in the top end of the fuel rod, which is designed as a gas plenum, or on the inside of cladding.

The activation products are distributed more widely. Impurities and the oxygen within the fuel are activated by irradiation and tend to remain there. The cladding and other metals used to make the fuel assembly and impurities within them also become activated. Most remain contained within the metals, although a small amount is dissolved by the coolant and finds its way to the reactor plant wastes. Although activation rates decrease as the distance from the reactor core increases, all of the other fixed internal parts of the reactor and the structure (typically steel and concrete) immediately surrounding the reactor become activated to varying degrees during their useful lives (typically 2–3 decades) and must be treated as radioactive waste when the plant is decommissioned. Finally, the coolant circulating through the core and the chemicals within it (if any) become activated to varying degrees, and the resulting radioactive isotopes find their way into reactor plant wastes during operation. Most of the activation products in the coolant are short lived, although a few are produced in significant quantities and are long lived (e.g., ^{3}H, ^{14}C).

6.2.2 Properties of Irradiation Products

The irradiation products in spent fuel could be characterized in a variety of ways, including element concentrations, isotope concentrations, chemical form, etc. However, the most relevant property of these products is that they are radioactive. As such, they have emissions of varying energy that include photons (gamma rays, x rays, bremsstrahlung), electrons (beta particles, Auger electrons), alpha particles, neutrons, neutrinos, and (very seldom) positrons. These emissions are generally well characterized and documented in large volumes (Lederer and Shirley, 1978).

These emissions have impact in two important ways. The first is that some of them (photons and neutrons) are considered to be *penetrating radiation* because they are capable of passing through a significant thickness of shielding material before being absorbed. As a consequence, any significant source of these must be shielded to protect human health. Calculation of this impact, called the *radiation dose*, is very complex because of its dependence upon the specific mix of isotopes, the age of the source, the geometric relationship of the source to the target, and the composition and thickness of materials between the source and target. Results of a calculation giving the dose rates from an unshielded, 10-year-old spent-fuel assembly are shown in Fig. 9 (Croff *et al.*, 1979). As a basis for comparison, a dose of about 500 rem has been considered to be lethal to 50% of human recipients, although medical advances evident in treating the casualties of the accident at the Chernobyl reactor in the Ukraine now indicate that the dose required to cause a 50% fatality rate may be significantly higher.

The second important impact of the radioactive emissions is that all (except neutrinos) are eventually absorbed in the material surrounding the spent fuel. The energy of these emissions is manifested as energy deposited in the material and the resultant heating of the material. A typical fuel assembly from a modern power reactor generates

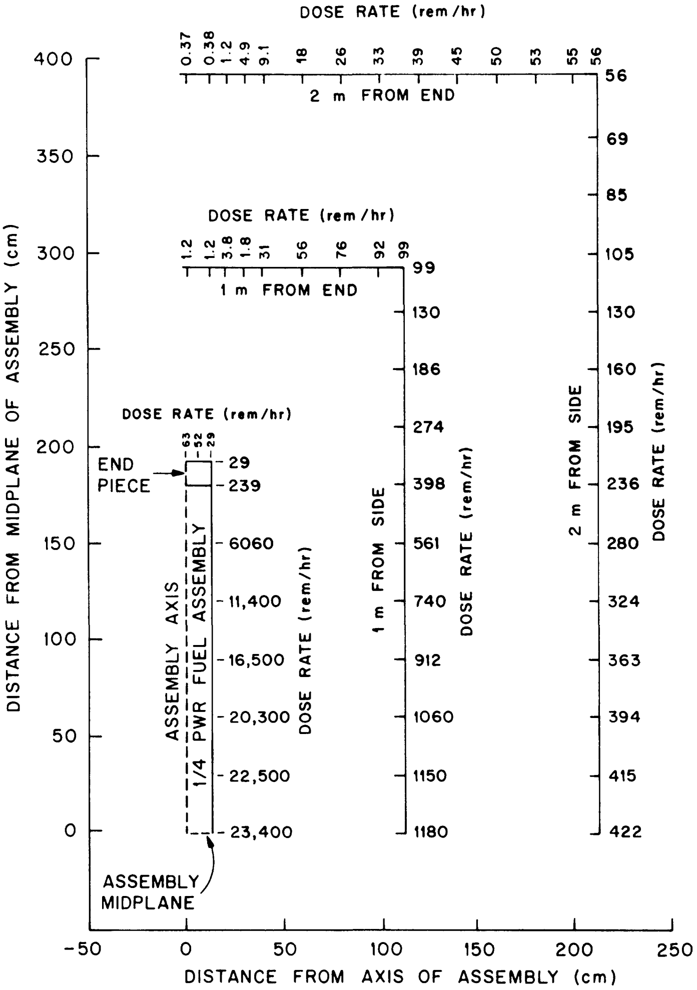

Fig. 9. Radiation dose from a 10-year-old spent pressurized water reactor fuel assembly that has been irradiated to 33 GWd per Mg of initial uranium.

heating of the material. A typical fuel assembly from a modern power reactor generates 17 300 kW of thermal power during irradiation. When the chain reaction ceases, the thermal power immediately declines to about 1200 kW. As shown in Fig. 10 (Croff and Alexander, 1980), the thermal power continues to decline at a decreasing rate thereafter until it finally attains negligible levels. However, for many years, the thermal power is sufficient to threaten the integrity of the fuel assembly if provisions are not made to remove the heat.

Radioactive isotopes can also impact living organisms if ingested or inhaled. For more discussion of these subjects, see ENVIRONMENTAL HEALTH AND SAFETY and NUCLEAR WASTE MANAGEMENT.

7. SPENT-FUEL REPROCESSING

After irradiation in a reactor and discharge from the reactor at the end of its useful life, spent fuel is stored for a significant amount of time (at least a couple of years, often many years) to allow shorter-lived irradiation products to decay. At this point, it becomes available for one of two possible dispositions. The first is that a determination is made that further chemical processing of the spent fuel is not worth while, and the spent

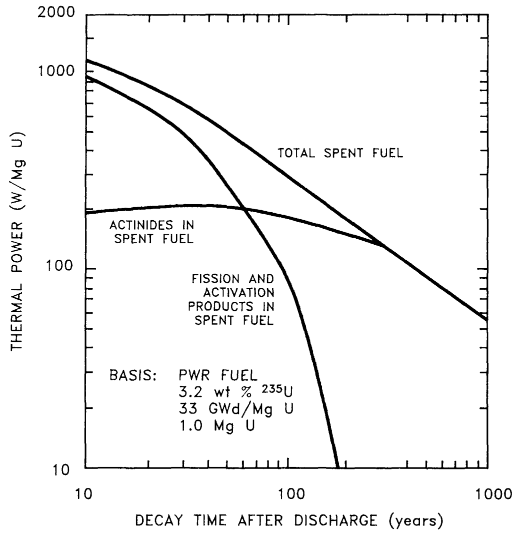

Fig. 10. Thermal power of major constituents in spent pressurized water reactor fuel that initially contained 1.0 Mg of uranium.

fuel should be treated as a waste (i.e., to have a once-through fuel cycle as described at the outset; see Fig. 1). In this case, the spent fuel would typically be stored for an extended time, sealed in a multiple-layer package, and sent to a permanent disposal site, where it would be buried deep underground (see article on NUCLEAR WASTE MANAGEMENT).

The second possible disposition is that further chemical processing of the spent fuel is worth while. This could be the case to recover various valuable constituents of the spent fuel, to change the form of the spent fuel to one more acceptable for waste disposal, or a combination of the two. The processing of the spent fuel is called *reprocessing*. The operations required for reuse of fissile and fertile constituents of the spent fuel are called *recycle*. Because of the intense radiation from the irradiation products, reprocessing must be carried out in remotely operated facilities that have thick, dense walls called *shielding* interposed between people and the nuclear material.

By far the most widely used reprocessing technology is the Plutonium-Uranium Extraction process, which is shortened to *Purex*. It has been used for decades to reprocess defense fuels to recover Pu for military purposes. More recently, large plants have been built in France and Great Britain to implement this process on civilian fuels, and more plants are under construction. The Purex process is described in some detail below as it is applied to oxide fuels clad in metals such as Zircaloy and stainless steel. This is followed by brief descriptions of variants of the Purex process and major alternative reprocessing technologies.

7.1 Purex Reprocessing

Reprocessing involves the application of a number of chemical processing steps to separate and purify the uranium and plutonium constituents of the spent fuel and to manage the residuum, which is considered to be waste. These are described in the following sections in the order in which they are performed.

7.1.1 Head End The *head end* of a reprocessing plant involves operations (generally physical) to expose the fuel material.

7.1.1.1 Remove Nonfuel Components. The first operation is to remove structural components that do not contain the fuel, such as end fittings and possibly the grid spacers. This may be accomplished by unscrewing threaded components or sawing and cutting. The removed metal pieces are treated as waste.

7.1.1.2 Segmentation of Fuel. The second operation is to expose the fuel material within the cladding in a process step called *segmentation*. This is typically accomplished with a mechanical shear, which is simply a mechanism to clamp a group of rods together with the ends projecting beyond a flat surface and chop off 2–4-cm segments with a hydraulically driven metal blade (much like chopping a bundle of celery stalks with a knife). It can also be accomplished with an abrasive saw or with a high-power laser.

7.1.1.3 Dissolution. The third operation is to dissolve the exposed fuel material in hot, concentrated nitric acid in an operation called *dissolution*. After being sheared off, the rod segments with the fuel exposed drop into a metal basket with small perforations. When the basket is sufficiently full, it is lowered into a vessel containing the acid. The acid is heated and sparged with air, and, over a period of time, most of the fuel matrix is dissolved. Fuel constituents that are not dissolved fall into three categories: cladding, insoluble metals, and volatile elements. The cladding does not dissolve in nitric acid, and remains in the basket in the form of *cladding hulls*. When fuel dissolution is complete, the cladding hulls are removed from the dissolver and treated as waste.

The insoluble metals typically comprise a small fraction of the fission products, which are themselves a small fraction of the fuel. It is the more noble fission products (i.e., Pd, Rh, Ru, Tc, Mo) that tend to be insoluble under dissolver conditions. These are typically removed from the nitric acid solution by centrifugation and may then be leached with more aggressive reagents (e.g., HNO_3-HF) to remove residual actinides before being treated as a waste or dissolved in aqua regia for further processing.

The volatile elements that evolve from the dissolver solution are called *off gas*. They may subsequently be captured to prevent their release to the environment and treated as waste. The inert gases Kr (which contains radioactive ^{85}Kr) and Xe are released quantitatively from the dissolver solution. Also released quantitatively under the highly acidic conditions are iodine (containing radioactive ^{129}I) and carbon as CO_2 (containing the radioactive activation product ^{14}C). Other species that are volatilized to some significant extent are activation- and fission-product tritium and Ru as RuO_4 (which contains radioactive ^{106}Ru). Nonradioactive oxides of nitrogen are also evolved from the dissolver solution.

7.1.2 Product Recovery The nitric acid solution containing most of the fuel constituents is then sent to two successive solvent extraction operations to recover impure uranium and plutonium in separate streams.

7.1.2.1 Primary Decontamination. In the first operation, the nitric acid solution undergoes multistage contact with a 30-vol% solution of tri-*n*-butyl phosphate (TBP) dissolved in an organic diluent such as *n*-dodecane. The aqueous and organic streams are immiscible and can be easily separated by gravity or centrifugation. The organic phase extracts virtually all (99+%) of the U and Pu (which are in the +6 and +4 valence states, respectively), a significant amount of the Np (which is in multiple valence states), and a significant amount of a few fission products such as Zr, Nb, and Ru. These leave the solvent extraction equipment in the organic stream. The nitric acid solution contains virtually all of the fission products, Am and Cm, and some of the Np. This solution is evaporated to concentrate it and recover the nitric acid. The remaining solution is called *high-level waste* and has little further value, except possibly for the recovery of small amounts of isotopes for specific beneficial uses.

The organic chemicals are recycled within the plant. However, because of the fission products, the organic extractant is exposed to intense radiation (primarily beta particles and gamma rays) while it is in contact with the aqueous solution. The result of this exposure is that some of the TBP is degraded to form undesirable impurities, with the levels increasing with increased exposure. It is necessary to remove the impurities from the solvent to make it acceptable for reuse. This can be accomplished by various means such as contacting it with an aqueous solution of sodium carbonate or passing it over a bed of silica gel.

7.1.2.2 Partitioning. The second product-recovery operation separates the U from the Pu. This is accomplished by contacting the organic stream laden with U and Pu with a dilute nitric acid stream containing a reductant such as ferrous sulfamate, hydrazine, or electrolytically generated U^{4+}. The reductant changes Pu to the +3 valence state, which is readily extractable by the nitric acid stream while the U remains in the organic stream.

7.1.2.3 Solvent Extraction Equipment. The key pieces of equipment in the above operations are the solvent extraction contactors. This is especially the case in the primary decontamination operation, where fast, efficient contact between immiscible phases followed by rapid, complete separation is important. The first equipment used was called a *mixer-settler*, which amounted to little more than mixing the organic and aqueous streams in a vessel using a powered agitator and then allowing the phases to separate by gravity in a baffled portion of the device. The two phases were then routed to neighboring mixer-settlers in counter-current fashion, much like the enriched and depleted streams in uranium enrichment described in Sec. 3.1.2. This equipment had relatively long contact times and low efficiencies, resulting in the development of more advanced equipment.

The successor to the mixer-settler in nuclear fuel reprocessing is the *pulse column*, which is presently in widespread use in most reprocessing plants. In this device, an aqueous (more dense) stream is introduced into the top of a vertical cylindrical column, and an organic (less dense) phase is introduced into the bottom. Spaced throughout the column are *sieve plates*, which are perforated horizontal metal plates. As the liquids pass through the plates, they are dispersed to promote intimate contact of the two phases. A *pulser* (much like a hydraulic cylinder) is periodically activated to force a small amount of liquid upward through the plates into the next stage. While this device was a distinct improvement over the mixer-settler, it still suffers from long contact times and requires large (and expensive) shielded process cells because the contactors are relatively tall.

The most recent advance in solvent extraction equipment is the *centrifugal contactor*. This is a piece of rotating equipment much like the gas centrifuge described in Sec. 3.1.3. The two streams are introduced, and, in one portion of the equipment, an agitator ensures good dispersion and contact between the phases, which quickly exit to another portion where they are quickly separated by centrifugal force. The equipment is relatively small and appears to meet the demands of fuel reprocessing quite well.

7.1.3 Product Purification At this point in the process, the separate U and Pu streams remain cross-contaminated (e.g., Pu in the U stream) as well as being contaminated with significant amounts of Np and fission products. The purpose of product purification is to remove these contaminants to acceptable levels.

7.1.3.1 Uranium Purification. The uranium is in the organic phase as it enters purification. It is then typically back extracted into dilute nitric acid and then subjected to one or two more cycles of TBP extraction and nitric acid back extraction to achieve the necessary purity. An aqueous solution of U may also be passed over a bed of silica gel for final purification. The purity requirements are higher if the uranium is being recycled for isotopic re-enrichment, as opposed to being simply used as feed material to be enriched with Pu.

7.1.3.2 Plutonium Purification. The plutonium is in the aqueous phase as it enters purification. It is typically subjected to one more cycle of TBP extraction and nitric acid back extraction. The final purification step consists of passing a nitric acid solution of the Pu through a column of anion-exchange resin to remove the residual contaminants.

7.1.4 Product Conversion The two product streams resulting from purification of U and Pu are aqueous acid solutions of the respective nitrates. The final forms of these products are generally UO_2 or UF_6 and PuO_2, respectively. Such conversion is usually done relatively promptly because the liquids are more vulnerable to leakage, criticality, and radiation effects and are expensive to store.

7.1.4.1 Uranium Conversion. The conversion of uranium nitrate solution to UO_2 generally proceeds as described in Sec. 2.2. Briefly, the uranium is precipitated as ammonium diuranate, which is then calcined to yield UO_2. This can be mixed with other fissile elements such as Pu to make mixed-oxide fuel, which can then be fabricated as described earlier and recycled to a reactor. Alternatively, the UO_2 can be further converted to UF_6 for re-enrichment, as described in Sec. 2.3.

7.1.4.2 Plutonium Conversion. The approaches used to convert plutonium nitrate to PuO_2, which is the only useful form in power production systems, are somewhat more diverse than for uranium. The plutonium can be precipitated with oxalic acid as plutonium oxalate, which is then filtered, dried, and calcined to PuO_2. The addition of hydrogen peroxide to the plutonium nitrate solution directly precipitates PuO_2, which can then be recovered through filtration and dried. Finally, there is some experience with direct evaporation and calcination of the nitrate solution to yield PuO_2, although this has not been used commercially.

7.1.4.3 Coconversion. Techniques have been developed and tested that form mixed $(U,Pu)O_2$ directly from a mixed-nitrate solution. These are called *coconversion processes*. They can involve direct calcination of the nitrate mixture to yield UO_3 and PuO_2, followed by hydrogen reduction to the mixed dioxide. Alternatively, the U and Pu can be coprecipitated and then calcined to the dioxide.

7.2 Reprocessing Variations

Numerous variations of reprocessing technology have been developed, and many have been deployed. Some of these variations represent the normal evolutionary improvement of any technology. Others include features necessary to produce somewhat different products or handle various types of spent fuel. Finally, fundamentally different approaches can be used for a specific fuel type.

7.2.1 Purex Variations A number of specific variations in the implementation of the Purex process are embodied in Sec. 7.1, and many more (primarily regarding chemical conditions in the processes and equipment differences) are possible, depending on cost estimates and the objectives of the Purex plant designer.

The only major variation of Purex that has been seriously considered (but not deployed) is to operate the process without the partitioning and purification operations (American Nuclear Society 1978). This yields a mixed U-Pu product that is contaminated with significant amounts of radioactive fission products for the purpose of making the material less desirable for diversion to undesirable uses. Such a material could only be recycled by designing facilities and equipment to carry out all steps of the fuel cycle with fully remote operations.

7.2.2 Aqueous Reprocessing Variations A number of aqueous reprocessing technologies other than Purex have been developed since the 1940s, when large-scale nuclear technology was born. Most of the early processes (e.g., bismuth phosphate precipitation, REDOX, Trigly, Butex) were displaced by the superior Purex process and others developed after Purex did not appear to offer enough advantages to justify the cost and risk of development. As a consequence, the Purex process continues to dominate the processing of uranium-based fuels to the exclusion of other aqueous processes.

Other aqueous processes are used in the processing of fuels that are not uranium based. The best known of these processes is *Thorex*, which is used for thorium-based fuels in which ^{233}U is produced by neutron capture. Thorex is similar to Purex in many respects, and the reader is referred to Sec. 7.1 for the operational sequence of the process. The primary difference in Thorex is that Th does not dissolve easily in nitric acid, and it is necessary to add HF to the dissolver solution to obtain an acceptable dissolution rate. However, the HF causes additional problems in the process, principally unacceptably high corrosion of metals. As a consequence, aluminum nitrate is added to the dissolver to complex the fluoride ion and keep corrosion to an acceptable level. The fluorides and aluminum report to the high-level waste and may result in problems or additional cost in its management. So far, Thorex has not been applied commercially.

7.2.3 Nonaqueous Reprocessing Variations Two major classes of nonaqueous processes have been developed: pyroprocesses and fluoride volatility. The potential advantages of these processes are that first, unlike organic solvents, radiation damage to the reagents does not create deleterious effects in the process, and second, the fundamentally different nature of the reagents offers unique possibilities concerning what is readily separated and what is not. The waste streams from these processes differ markedly from aqueous process wastes.

7.2.3.1 Pyroprocesses. *Pyroprocesses* are defined by the use of high-temperature (>500 °C) molten metals and salts as the reprocessing reagents. The feed material must typically be in the form of metal, and oxide fuels must be reduced to the metal (typically with Ca) before processing. Numerous pyroprocesses have been investigated over the last few decades, including volatilization, fractional crystallization, extraction with liquid metals, liquid extraction from fused salts, electrolysis, partial oxidation, skull reclamation, and salt transport. Most of these have not been developed to any appreciable extent because they manifested serious deficiencies compared to other processes, but those described below have been investigated to a significant extent.

A partial oxidation process (called *melt refining*) that involves melting uranium-based metallic spent fuel in a ZrO_2 crucible at 1400 °C has been applied to the recycle of fuel at the EBR-II test reactor in Idaho for almost 30 years. Many fission products are volatilized at this temperature, and other nonvolatile fission products that are more reactive than uranium and 5–10% of the feed uranium are oxidized by the ZrO_2 and form a separate layer (called a *skull*) in the crucible. The remaining metal is a mixture of fission products less reactive than uranium (e.g., Zr, Nb, Mo, Ru, Rh, Pd), which are collectively known as *fissium*, and uranium. This mixture is combined with fresh, enriched uranium metal, and the product is poured into glass molds to make recycle fuel. The fissium is highly radioactive, but its physical properties stabilize the uranium during irradiation in the reactor. The *skull reclamation* process uses a Mg-Zn alloy to reduce the valuable uranium in the skull selectively to the metal, which is soluble in the alloy. The Mg-Zn is then removed by distillation to recover the uranium.

An electrolysis process called *electrodeposition* is being investigated for processing metallic U-Pu fuels for advanced reactors cooled with liquid sodium. In this process, stainless-steel–clad fuel made of metallic U-Pu-Zr and a portion of the fission products are chopped and placed in a basket. The electrorefiner contains a bottom layer of molten Cd and a top layer of LiCl-KCl eutectic salt. The basket is lowered into the salt layer, and $CdCl_2$ is added to oxidize the alkali, alkaline-earth, and rare-earth metals in the spent fuel to their chlorides, which dissolve in the salt layer. Electrolysis occurs by making the basket and bottom Cd layer the anode. Two cathodes are immersed in the molten salt layer, with one part being steel and one part being molten Cd in a ceramic crucible. During electrolysis, U-Cd is deposited on the steel cathode, and a mix of U-Pu-Cd-lanthanides is deposited on the Cd cathode. The U-Pu product contains the other actinides heavier than U that were present in the spent fuel. Both products contain significant amounts of fission products and the Zr that was present in the initial fuel. When electrodeposition is complete, the two cathodes are removed to a retort where the Cd is volatilized. The remaining constituents are then removed to an injection melter, which melts the actinide products, adjusts the composition to the correct U-Pu-Zr proportions, and injects this mixture into fuel molds. The fuel is then recycled to the reactor.

7.2.3.2 Fluoride Volatility Processes. As might be evident from the section on uranium enrichment, fluorides of some elements are volatile at modest temperatures (a few hundred degrees Celsius or less). This property has been used as a basis for reprocessing spent fuel using a *fluoride volatility process*.

In this process, the fuel material is exposed by decladding operations much as in the Purex process. However, at this point, the fuel material is converted from the oxide to fluorides using a sequential combination of HF and F_2. Many of the fluorides are volatile but at varying temperatures. This permits a first separation of the U from other radionuclides by using simple distillation. The impure uranium stream from this operation is then further purified using techniques such as passing the gases through solid beds of NaF to remove undesirable elements by selective sorption of uranium. When this is used as a primary reprocessing technology, only the U is recovered, and the Pu follows the fission-product waste. As a consequence, this has been considered for only highly enriched uranium fuels where the Pu is mostly composed of ^{238}Pu, which is not worth recovering.

Fluoride volatility processes are capable of excellent decontamination from undesirable constituents and have been used to clean uranium feed to enrichment operations. A full-size reprocessing plant (Midwest Fuel Reprocessing Plant) in which the final cleanup operations were based on fluoride volatility processes was designed and built at Morris, Illinois, but never operated because of design inadequacies (especially inadequate surge capacity within the facility).

7.2.4 Continuous Reprocessing The reprocessing of spent fuel from an individual reactor, as described in Sec. 7.1, is performed intermittently on a group of discharged spent fuel assemblies sent to a separate facility. However, reactor systems have been developed and tested in which the fuel is circulated through a core region configured so that a nuclear chain reaction occurs and the fuel is heated. The fuel then flows to a heat exchanger, where the heat is removed by generating steam to make electricity in a turbine. The fuel (or a fraction thereof) is then circulated to a reprocessing operation that is an integral part of the fuel flow loop, where it is *continuously reprocessed* to remove deleterious fission products.

This concept has been tested in designs involving aqueous fuel [e.g., the Aqueous Homogeneous Reactor Experiment (Chapman *et al.*, 1963) and the Molten Salt Breeder Reactor (Weinberg *et al.*, 1970)]. In the aqueous case, reprocessing operations involve volatilization of the noble-gas fission products and use of the fact that actinide sulfates exhibit *retrograde solubility* (less soluble at higher temperatures) to precipitate them with some of the fission products and recover them with a cyclone. For the molten salt reactor, the reprocessing technology would be very similar to pyroprocessing as described in Sec. 7.2.3.1. These concepts have not been developed further than the test reactor phase, and there are no apparent plans to do so.

8. STORAGE AND TRANSPORTATION OF NUCLEAR MATERIALS

Because of economics of scale and other business factors, most facilities constituting the nuclear fuel cycle shown in Fig. 1 are generally not colocated. As a consequence, various radioactive product materials must be transported between facilities. This, in turn, dictates that the various facilities have significant capability for storing the feed and product materials in order to achieve economic transportation loads, to ensure uninterrupted facility operations, or to condition materials prior to processing.

Although there are radioactive materials transported between and stored at each of the facilities shown in Fig. 1, most of these operations are straightforward because the materials involved are only mildly radioactive. Considerations for transportation and storage of materials that have not been irradiated involve protecting them from dispersion and the natural elements (e.g., rain) and, in the case of precision-manufactured components such as fresh fuel assemblies, protecting them from physical damage during transport. These considerations, and the transportation and storage methods employed, are very similar to those used for mildly toxic industrial chemicals and other high-precision equipment and will not be detailed here. One exception is storage and transportation of highly enriched fissile material, which requires special considerations to provide adequate protection from illicit diversion and criticality control.

On the other hand, spent fuel and the high-level waste from the primary decontamination operation of fuel reprocessing are highly radioactive, emitting significant amounts of heat and penetrating radiation. The considerations in their transport and storage are unique and are described in some detail below.

8.1 Storage

8.1.1 Water Pools Spent fuel is most commonly stored in deep-water pools, which are usually rectangular and placed in a large open room with a high-capacity overhead crane. They are present at reactor plants, which typically have the discharged spent fuel from many years of operations stored in the pool since disposal facilities have not yet been built in any country. They are also present at fuel reprocessing plants and constitute the centerpiece of stand-alone facilities designed for long-term spent-fuel storage. Experience has shown that fuels clad in Zircaloy and stainless steel can be stored under water for decades with no significant degradation.

The pools have varying surface areas depending on the capacity of the facility. They are normally quite deep: 3–6 m plus the height of a fuel assembly. The water layer above the top of the fuel assemblies provides radiation shielding to the workers (e.g., crane operators) located around and above the pool, and one fuel-assembly height is required for storage. A portion of the pool is normally deeper yet by about an extra fuel-assembly length plus 1–2 m to accommodate a large, thick-walled vessel used to transport the fuel between sites (see Sec. 8.2 for details).

The pool is built with a grid structure of metal racks to hold the spent fuel in place in a configuration that eliminates the possibility of a critical reaction. The pools have a high mass and typically a high center of gravity, and extensive provisions are made to ensure integrity during potential seismic events. The water in the pool is circulated to maintain it at extremely high purity levels and to remove the decay heat from the spent fuel. Fuel assemblies in which the cladding of one or more elements has been penetrated are normally sealed in a metal can before being stored.

Other materials having significant radioactivity are often stored in the pools. Examples are contaminated or irradiated reactor components and metal hardware from dismantled fuel assemblies. In a reprocessing plant pool, the high-level waste packages and cladding waste packages are typically stored in the pool.

8.1.2 Dry Storage Storage of some materials under water may not be possible or desirable on the basis of technical or economic grounds. This may be the case because water will adversely affect the integrity of the fuel (e.g., fuels clad in Al and Mg or made of graphite), an existing pool is full and expansion is not possible, or the fuel characteristics are such that it is more economical to store it under dry conditions.

There are two common technologies for dry storage. One is to store the fuel in a container called a *storage cask* that is so massive that it provides adequate radiation protection. This cask is simply loaded with spent fuel or waste, sealed, and then placed outside on a concrete pad where heat is removed via conduction through the package and natural convection. This container is very similar to those used for transporting the spent fuel except that it may not be certified for off-site transportation. It is described in detail in Sec. 8.2.

The second common dry-storage technology is *dry wells*. These are typically an array of individual cells made of reinforced concrete and sized so that a fuel assembly can be lowered into them and a plug providing radiation shielding inserted into the top of the well. The wells are not normally ventilated, and heat removal occurs via conduction through the walls and then convection in a moving air stream. If the wells are ventilated and actively cooled, the gas stream must be purified before release. This type of design is also used in reprocessing plants for storage of high-level and other highly radioactive wastes.

8.2 Transportation

Spent fuel is transported between facilities via trucks, railroads, and ships in containers called *shipping casks* or *shipping packages*. This operation is challenging because the cask must simultaneously provide for radiation protection, heat removal, criticality protection, and barriers to the release of radionuclides from the spent fuel in the case of accidents.

8.2.1 Regulations Standards that casks must meet have been established by international organizations (IAEA, 1992) and national regulatory authorities and are remarkably uniform throughout the world. This is necessary because of the international nature of the nuclear fuel cycle and the need to move nuclear material between nations. The manner in which a particular designer meets these standards is flexible, depending on the specific characteristics of the fuel being transported. When a new shipping cask is to be built, or an existing cask significantly modified, it must be shown to meet the standards by actual testing, typically with a scale model, or by rigorous engineering analysis. A cask is subjected to the following tests in sequence:

1. a 9-m drop onto a flat, horizontal, unyielding surface so that it strikes in its most damaging orientation (typically an edge);
2. a 1-m drop onto a 0.15-m-diameter steel bar at least 0.2 m long, at its most vulnerable part;
3. a fire test in which the cask is totally engulfed in a 800 °C thermal environment for 30 min;
4. water immersion in which the cask is submerged under at least 0.9 m of water for 8 h, and a separate cask is completely immersed under 15 m of water for 8 h. In addition, for casks carrying over 1 MCi of radioactivity, the International Atomic Energy Agency requires that the immersion depth be increased to 200 m (a pressure chamber is used to simulate this depth).

To be acceptable, the cask must remain leaktight after this test sequence.

Myriad additional regulations are often imposed on the transportation operation *per se*. These include how the cask is fastened to the transport vehicle, radiation level limits, shipping documentation, labeling, driver qualifications, emergency response, and routing.

8.2.2 Shipping Casks The above requirements, when coupled with the need for a significant payload and removal of significant amounts of decay heat from the spent fuel, result in the cask having to be made almost exclusively of metal components. A typical schematic of a spent-fuel shipping cask is shown in Fig. 11. The cask has an inner and outer shell usually made of stainless steel. Between these shells is a very thick layer of denser metal to provide shielding from penetrating photon radiation. Iron, lead, and uranium are often used to provide this

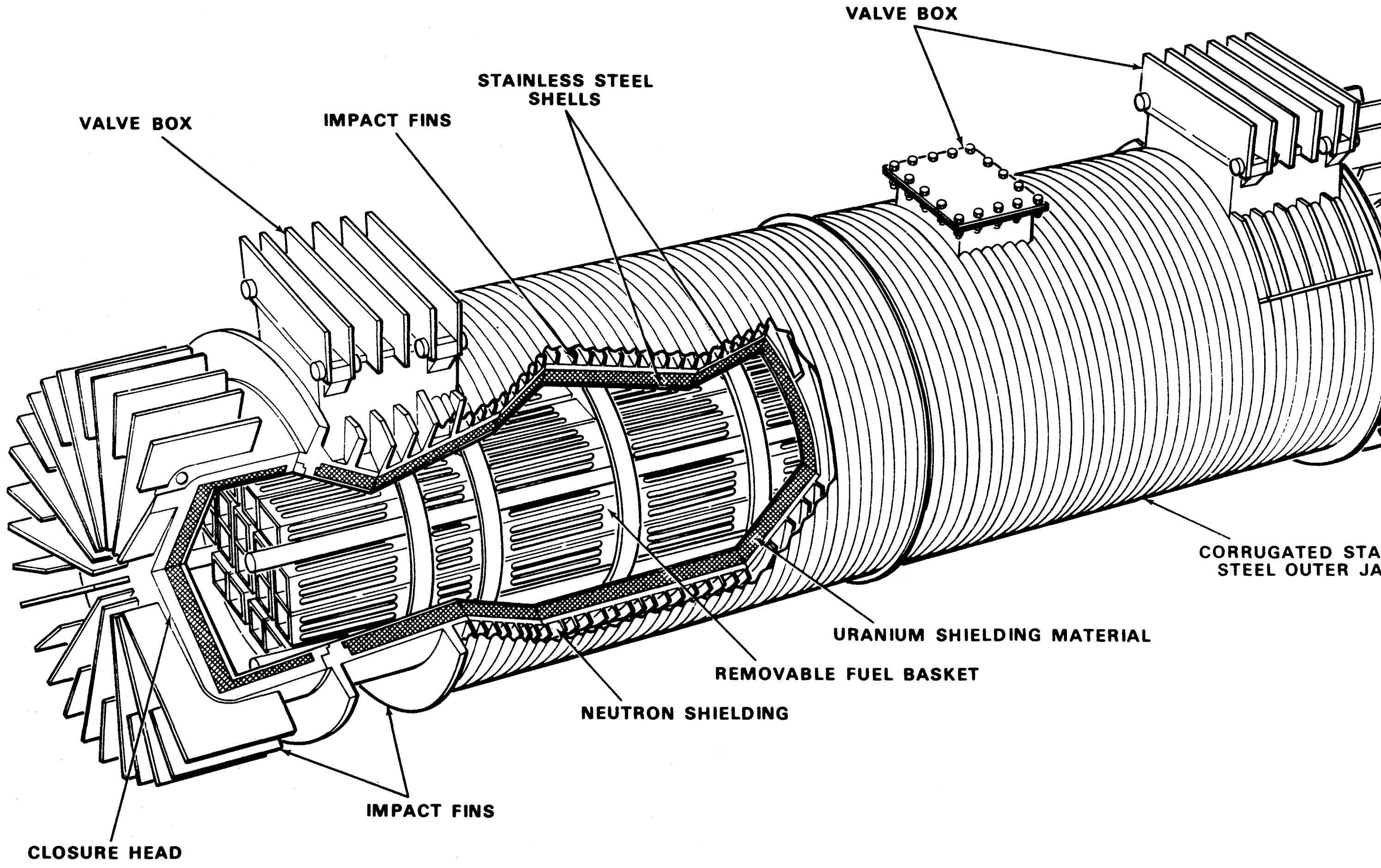
VALVE BOX
STAINLESS STEEL SHELLS
VALVE BOX
IMPACT FINS
CORRUGATED STA
STEEL OUTER JA
URANIUM SHIELDING MATERIAL
REMOVABLE FUEL BASKET
NEUTRON SHIELDING
IMPACT FINS
CLOSURE HEAD

Fig. 11. Schematic diagram of a typical spent-fuel shipping cask.

shielding. In addition, if the spent fuel emits a significant amount of neutron radiation, there is often a layer of hydrogenous material (e.g, polyethylene) that contains a good neutron absorber (e.g., boron). Casks typically weigh between 25 and 100 Mg.

A grid called a *basket* that is appropriately sized for the spent fuel being transported is inserted into the hollow cavity in the middle of the cask. The basket holds the fuel in place to prevent damage during normal transport and minimize damage during accident conditions. It is also important in assuring that the array of fuel assemblies will remain subcritical during transport.

The cask is typically loaded by removing its lid and immersing it in the pool with the open end upward. The fuel is then loaded into the cask with a crane, the lid placed on the cask (to provide radiation shielding), and the cask removed from the pool. Water is drained from the cask interior, the lid bolted down, the inside dried with a vacuum pump, the cask leak tested, and the interior filled with an inert gas (typically nitrogen). The outside of the cask is decontaminated to remove the trace radioactivity present from the pool water, loaded onto the transport vehicle (trailer or rail car), tied down, and then transported to its destination, where the process is reversed to unload the fuel. The cask is cooled by conducting the heat through the walls; no active cooling measures are employed, although fins are often used to enhance convective heat transfer on the outer surface.

9. BENEFICIAL USE OF RADIOISOTOPES

The preceding sections of this article have been concerned primarily with the beneficial use of a very limited number of radioisotopes: ^{232}Th/^{233}U and ^{235,238}U/$^{239-242}$Pu, with the latter halves of the pairs being created during fuel irradiation. Thorium and ^{235,238}U occur naturally, although the latter often requires enrichment to achieve usable concentrations of ^{235}U. However, a large number of other radioisotopes have found beneficial uses in smaller, but nevertheless extremely important, applications. The purpose of this section is to describe the production and uses of radioisotopes other than those used in the manufacture of power reactor fuels.

9.1 Production and Recovery of Radioisotopes

Radioisotopes other than those that occur naturally are produced by two methods: irradiation of a target by nuclear particles and decay of other radioisotopes.

9.1.1 Production by Irradiation Radioisotope production by irradiation can be accomplished using nuclear reactors or particle accelerators (see NUCLEAR ENERGY: FISSION; ACCELERATORS, LINEAR). In the case of reactors, neutrons irradiate the target material and are captured or induce fission as a result. In the case of accelerators, the target is irradiated by high-energy charged particles such as protons, deuterons, or alpha particles. This results in the ejection of photons and/or other nuclear particles. In both cases, the target material, irradiating particle intensity, irradiation duration, and particle energy spectrum are selected to maximize the yield of the desired radioisotope or minimize contamination by undesirable isotopes.

Very often, the target material is a stable material enriched in a specific isotope (see Sec. 3.2.5) to maximize the yield and purity of the desired radioisotope. The targets are always small (micrograms to grams) in the case of accelerator irradiation, and in cases where expensive, enriched stable isotopes are used as targets in reactors. Targets are also small (ranging from 100 g for initial isotopes to micrograms for targets of the heaviest elements) when long-lived radioisotopes (the transuranic actinides in particular) are irradiated in reactors to produce heavier actinides. When the desired radioisotope is a fission product, the target may be nuclear fuel, which can be a relatively large assembly as described previously, or a target consisting of a few grams of enriched uranium.

9.1.2 Production by Decay In a few cases, very pure radioisotopes can be recovered by decay. Specifically, a pure, relatively long-lived isotope is isolated and allowed to decay. Periodically, its decay product is separated (or *milked*) from the parent for a particular use. This production method is significantly dependent upon serendipity, requiring the ability to isolate a suitable long-lived parent that does not produce competing decay products and that produces the desired product at an adequate rate. This method is used, for example, in the preparation of ^{99m}Tc from decay of fission product ^{99}Mo.

9.1.3 Recovery of Radioisotopes In general terms, radioisotopes are recovered by dissolving the target and applying appropriate chemical or physical processes to isolate a relatively pure product. When production is by decay, the parent is often maintained in dissolved form to facilitate separation. In the case of fission products, the feed material is most often the high-level waste produced after the primary decontamination operation in reprocessing. In other cases, the target may be dissolved in any of a number of reagents and the separation effected using a variety of techniques (e.g., solvent extraction, precipitation, ion exchange, vaporization). These techniques are specific to the nature of the target and product, as well as the product purity specifications. In most cases, the processing also involves recovery of the residual target material, which is often very valuable. The recovery operations take place in facilities ranging from standard ventilated laboratory hoods to elaborate cells with thick radiation shielding, depending on the characteristics and amount of the radioisotopes being processed. In very special cases, a few atoms of exotic radioisotopes such as some of the very heavy elements are separated by recoil from thin films that are bombarded with particles from accelerators.

9.2 Useful Radioisotopes and Their Applications

A wide variety of radioisotopes are regularly used in the modern world. Some of these applications are very specialized and esoteric, relating only to the frontiers of knowledge (e.g., the very heavy synthetic elements). Other applications occur every day in familiar settings and affect the lives of many people (e.g., medical diagnostics/treatment and the use of radioisotopes in smoke detectors). The following provides a framework of the applications of radioisotopes that is used here for the purposes of more detailed discussions: medical diagnostics and treatment, industrial applications, basic research in the physical sciences, radiotracer applications, power production, and military applications.

9.2.1 Medical Diagnostics and Treatment By far the most rapidly growing use of radioisotopes is in the medical community (see NUCLEAR MEDICINE). Within this community, radioisotopes are used for diagnosis (determining the cause of symptoms) and treatment (curing the problem that causes the symptoms). Approximately 300 radiopharmaceuticals are presently available, and this number is growing rapidly.

9.2.1.1 Diagnosis. The chemical properties of many isotopes result in their concentrating in particular organs when introduced into the body. The radioactive emanations from the decay of these isotopes can then be detected, providing an indication of the activity of the organ or even an image of the organ that can achieve remarkable clarity through computer enhancement. Although a wide variety of radioisotopes are used for this purpose (e.g., ^{131}I, ^{201}Tl, ^{47}Sc), the most commonly used is ^{99m}Tc, which is produced by decay from its longer-lived ^{99}Mo parent. By incorporating the ^{99m}Tc into various chemical compounds, it can be caused to concentrate in almost any organ of the human body. It is estimated that ^{99m}Tc is used in about 30 000 medical procedures per day.

9.2.1.2 Treatment. After a problem is diagnosed, radioisotopes are often used to effect a cure. In general, this involves use of the radioactive emanations of a radioisotope to destroy specific tissues such as tumors. For example, gamma rays from ^{131}I are often used in the treatment of the thyroid gland by introducing it into the body, and ^{60}Co is used as an external gamma-ray source for treating a variety of tumors. Californium-252 decays in part by the emission of neutrons and has been found to be effective in treating cancer of the cervix and prostate.

9.2.2 Industrial Applications

9.2.2.1 Sterilization. The quest to improve the quality of health and foodstuffs has resulted in an increasing need for sterilization. Regarding the former, sterilization of a wide variety of medical equipment is required. Because much of this material is now made of plastics and polymers, the previous practice of using high temperatures is unacceptable. Sterilization of foodstuffs is increasingly desirable to prevent spoilage while retaining the full nutritional content and taste, and, again, this is not possible by using high-temperature or drying technologies.

It has been found that irradiating the equipment or food with massive amounts of gamma radiation is very effective in sterilization without impairing other characteristics or making the target material radioactive. The medical equipment can be sterilized after being sealed in impermeable packages, and sterilized food can last for months without refrigeration. Although any radioisotope that emits high-energy gamma rays can be used in this application, by far the most common are ^{60}Co (produced by neutron irradiation of natural ^{59}Co), ^{192}Ir (produced by neutron irradiation of natural ^{191}Ir), and ^{137}Cs (a fission product). Gamma ray sterilization has also been tested on sewage sludges, although application has been very limited.

9.2.2.2 Radiation-Induced Reactions. Gamma rays (typically from the same sources as described immediately above) have also found use in creating desirable chemical reactions. One of the most common is the use of gamma rays to induce the polymerization of a monomer (e.g., linking ethylene monomer to form polyethylene). Gamma rays from radioisotopes can also be used to break chemical bonds when required, and this has been used to control the molecular weight of Teflon®.

9.2.2.3 Measurement and Testing. Radioactive emanations from radioisotopes can be used in a number of ways related to industrial measurement and testing.

1. A gamma-ray source and a number of detectors can determine the level or density of liquid in a tank as a means of process control.
2. Beta particles (e.g., from ^{90}Sr) can be used to measure the thickness of thin layers on a substrate such as metal plating or paint as a means of process or quality control.
3. Neutron and gamma radiation are used in petroleum exploration to determine the characteristics of the surrounding rock down-hole, which is called *well logging*.
4. Neutron and gamma radiation are used to radiograph industrial equipment to detect imbedded flaws (similar to an x ray of the human body).
5. Radioisotopes are used to make self-luminous materials (e.g., watch dials, egress lighting).
6. Americium-241 is used in many home smoke detectors, where the difference between radiation-induced ionization of the smoke and clean air results in activation of the alarm.

9.2.3 Basic Research in the Physical Sciences The properties of radioisotopes have made them very useful in research concerning the physical sciences. In some instances, the radioisotope is the subject of study. For example, the heavy actinide isotopes (e.g., einsteinium, fermium) are studied to learn more about the structure of the nucleus and the laws that govern it, as well as the chemical properties of these exotic elements. In most other cases, the properties of the radioisotope emanations are used as a tool to learn more about a specific phenomenon or object. One of the most common examples is *activation analysis*, in which an object is briefly exposed to neutrons and the induced radioactivity measured as a means to determine nondestructively the atomic composition of the object. Applications of activation analysis range from industry to criminology. The use of radioisotopes to determine the age of an object (e.g., ^{14}C dating) is also quite common in the archaeological and geological sciences.

9.2.4 Radiotracer Applications A *radiotracer* is a radioisotope that is used to follow the movement or reaction path of a material. For example, a biological compound into which radioactive tritium has been introduced can be used to study the behavior of the compound in the human body. Radiotracers are widely used in many applications described above, including medicine (e.g., examining blood circulation patterns), industry (e.g., tracking the movement of process fluids and leak detection), and the sciences (e.g., monitoring the movement of groundwater around waste-disposal sites). Radiotracers may be natural (e.g., ^{14}C, ^{226}Ra) or artificial (i.e., manufactured and introduced into the system being studied).

9.2.5 Power Production Some radioisotopes are used indirectly in the production of nuclear power. For example, a neutron source is needed to initiate the chain reaction in a nuclear reactor, especially a new reactor containing only unirradiated fuel. These sources are often made of an alpha-emitting radioisotope (e.g., ^{210}Po) and an element that releases neutrons when an energetic alpha particle is absorbed (e.g., beryllium) or of a radioisotope such as ^{252}Cf that decays in part by *spontaneous fission* and consequent neutron emission.

A much more common use of radioisotopes is as a source of power for specialized applications. Medium-lived radioisotopes such as ^{90}Sr, ^{238}Pu, and ^{244}Cm have been and continue to be used for this purpose. The heat from the radioisotope may be used without conversion (i.e., as a source of space or process heating). However, the heat is typically converted to electricity directly by using the heat to create a flow of electrons in a thermoelectric generator or indirectly by vaporizing a fluid which drives a turbine that turns an electric generator.

The major attribute of such power sources is a high power output per unit weight combined with a very long useful lifetime. The cost of radioisotopes makes their use as a power source impractical for normal applications. However, they have found extensive use in specialized applications such as powering equipment (e.g., military outposts, buoys, communications) where refueling would be more expensive, powering the instrumentation aboard spacecraft (especially interplanetary probes), and in early artificial hearts.

9.2.6 Military Applications There are specialized military applications for specific radioisotopes. One of the most common is the use of tritium in nuclear weapons, which can greatly increase the explosive power of the device. This occurs when the fissioning of the uranium or plutonium primary system provides sufficient energy to initiate a reaction in which two tritium nuclei are fused, with the resulting release of considerable additional energy. Other radioisotopes such as ^{252}Cf and ^{244}Pu have found use in diagnosing the sequence of events that occurs in underground nuclear weapons tests. Californium-252 is also used to radiograph intact military aircraft for flight-induced structural damage, an approach that has proven to be less costly than disassembling the aircraft and using more conventional radiographic techniques.

GLOSSARY

Activation Analysis: Analysis of the elemental content of a material by briefly irradiating it with neutrons and measuring the radioactive emissions from the resultant short-lived radioactive products.

Activation Product: The radioactive product of neutron capture by a material originally composed of nonradioactive isotopes.

Atomic Vapor Laser Isotope Separation: A physical method of uranium enrichment based on the selective adsorption of laser light by vaporized ^{235}U atoms to produce charged atoms (ions).

Burnup: A measure of fuel consumption (and power production) in a nuclear reactor. In the strict sense, burnup is the depletion of fissile material through fission or other processes and is expressed as a percentage. In this sense, fuel could have 75% burnup (in which case 75% of the initial ^{235}U nuclei have fissioned or been converted to another isotope). Burnup is more commonly used to express the amount of energy produced per metric ton of uranium metal (MWd/t). This is more a measure of the total number of fissions that occur, because heat is produced by fissioning of initial ^{235}U nuclei, fissioning of Pu produced in the reactor, and other nuclear processes. While it would be more correct to use the term "exposure" in this context, "burnup" continues in popular usage. Commercial power reactors attempt to maximize the amount of energy produced (i.e., fuel consumed) and thus typically have high burnups (greater than 10 000 MWd/t). "Production reactors" are operated to maximize production of weapons-grade ^{239}Pu. This is typically limited by the generation of ^{240}Pu that occurs at higher burnups. Production reactors thus typically have burnups of much less than 10 000 MWd/t.

Calutron: An electromagnetic isotope separation device used to enrich uranium during World War II and subsequently to produce a variety of enriched stable isotopes for research and medical applications.

Cascade: Isotope separation devices (gaseous diffusion or centrifuge) connected so as to multiply the amount of enrichment and/or depletion achievable by a single device.

Centrifugal Contactor: A solvent extraction device that mixes two immiscible liquids of differing density and then forces their rapid separation with the use of centrifugal force.

Cladding: The protective surrounding of reactor fuel materials to prevent their corrosion or erosion by the reactor coolant, to increase the strength of the fuel material, and to contain fission products that would contaminate the coolant.

Cladding Hulls: Pieces of nuclear fuel cladding remaining after fuel dissolution.

Coconversion: A process wherein uranium and plutonium concurrently and together undergo conversion from nitrate salts to oxide.

Continuous Recycle: A nuclear fuel cycle in which the fuel is dissolved or suspended in a liquid such as water, molten salts, or molten metals and continuously circulated through a loop involving fission, heat removal, waste product removal, and fuel renewal.

Continuous Reprocessing: Reprocessing technology used in continuous (as opposed to a discrete fuel batch) recycle, which involves on-line reprocessing of nuclear fuel.

Conversion: Changing the chemical form of uranium or plutonium to one desirable as feed to a subsequent step of the nuclear fuel cycle [e.g., yellowcake to UF_6, U/Pu nitrates to $(U,Pu)O_2$, UF_6 to UO_2].

Core: The power-producing heart of a nuclear reactor, containing the nuclear fuel that undergoes fission and releases energy. The core contains the fuel, the moderator (if required), and the control rods. The core is often surrounded by a reflecting material (e.g., beryllium) that bounces stray neutrons back into the fuel.

Dissolution: Immersion of fuel rod pieces resulting from segmentation in solvents, typically hot nitric acid, to dissolve the fuel.

Dry Well: A storage area containing air surrounded by thick radiation shielding in which spent fuel or nuclear waste is stored while being cooled by conduction of heat to an air stream outside the cell walls.

Electromagnetic Isotope Separation: An isotope separation or enrichment process that capitalizes on the different radii of the circular paths in a magnetic field of ions having different atomic masses.

Enrich: Increase the relative abundance of an isotope of an element, for example, to increase the ^{235}U concentration from its natural value of 0.72% to 3–5% for use in civilian power reactors.

Fertile Material: A material that is not fissionable by thermal (low-energy) neutrons, but that can be converted into a fissile material through neutron capture by irradiation in a reactor. There are two basic types of fertile materials: ^{238}U and ^{232}Th. When these fertile materials capture neutrons, they are converted into fissile ^{239}Pu and ^{233}U, respectively.

Fissile Isotope: An isotope that can fission by absorbing a thermal neutron and can support a thermal neutron chain reaction as well. Examples are ^{239}Pu and ^{235}U.

Fission Products: The complex mixture of elements produced as a result of nuclear fission. Fission products decay and form additional products, which may be radioactive or stable. The mixture of fission products so formed contains about 200 different isotopes of over 30 elements.

Fissionable Isotope: An isotope that can be fissioned by only high-energy neutrons.

Fluoride Volatility Process: A reprocessing technology based on the low-temperature volatility of fluorides of key elements (such as uranium).

Fuel Assembly: A number of fuel rods held in a fixed geometric array.

Fuel Fabrication: Processing of unirradiated nuclear and structural materials into a chemical and physical form and configuration that are appropriate for the fuel of a specific reactor design.

Fuel Refabrication: Processing of nuclear materials recovered from spent (irradiated) fuels and structural materials into a form and configuration that are appropriate for the fuel of a specific reactor design.

Fuel Rod (or Fuel Element): A container of nuclear fuel, usually a metal tube, suitable for insertion in the core of a reactor.

Gas Centrifuge: A device for enriching uranium (and other isotopes) in which gaseous UF_6 is contained in a hollow cylindrical rotor that is spun at high speed. The heavier $^{238}UF_6$ molecules tend to move closer to the rotor wall than the lighter $^{235}UF_6$ molecules, thus allowing enrichment to occur.

Gaseous Diffusion: A process for enriching uranium that uses gaseous UF_6 as the working fluid. The lighter $^{235}UF_6$ molecules flow through a porous barrier slightly faster than the heavier $^{238}UF_6$ molecules, thus allowing enrichment to occur.

Grid Spacers: Metal component of a fuel assembly used to hold fuel rods in place laterally.

Head End: The initial step of reprocessing, which includes spent fuel segmentation and dissolution.

High-Level Waste: The highly radioactive waste remaining after reprocessing.

Ion Exchange: A solid or liquid material with replaceable ions that can exchange with other ions in the surrounding solution.

Isotope: A nuclear species that is unique because of the number of neutrons and protons it contains. Most elements are composed of a mixture of isotopes, each with the same number of protons but differing in the number of neutrons. Isotopes may be nonradioactive or radioactive.

Isotope Enrichment: A process to increase the relative abundance of selected isotopes of an element.

Laser Isotope Separation: An enrichment process in which a selected isotope is excited by a laser beam tuned to a specific frequency (color). The excited isotope can then be removed by, for example, electromagnetic or chemical means. Important examples include Atomic Vapor Laser Isotope Separation (AVLIS) and Molecular Laser Isotope Separation (MLIS).

Metric Ton (t): 1000 kilograms or 1 Mg. Also written as tonne. Equal to 2200 pounds.

Milk: Periodically separate a desirable decay product being continuously produced by a long-lived radioactive parent.

Mixer-Settler: A solvent extraction device that mixes two immiscible liquids of differing densities and allows them to settle under normal gravitational force to effect the separation of components.

Molecular Laser Isotope Separation (MLIS): Laser isotope separation in which the feed is a uranium compound, typically UF_6.

Neutron Capture: A process in which an isotope absorbs a neutron but does not fission, thereby creating a heavier isotope of the same element.

Nuclear Fuel: The nuclear material that sustains a chain reaction in a nuclear reactor.

Nuclear Fuel Cycle: The sequence of operations involved in supplying fuel for nuclear power generation: irradiating the fuel in a nuclear reactor, handling and treating the fuel elements following discharge from the reactor, and disposing of the radioactive wastes. If spent fuel is to be reprocessed, the fuel cycle is called a "closed" cycle. If the fuel is not to be reprocessed, it is a "once-through" cycle.

Off Gas: Gases evolved during reprocessing of spent fuel.

Once-Through Nuclear Fuel Cycle: A nuclear fuel cycle in which the fuel constituents are not recovered for reuse, and the spent fuel is managed intact.

Penetrating Radiation: Particle or electromagnetic emissions, resulting from the decay of radioactive isotopes or devices such as an x-ray machine, that can penetrate significant thicknesses of solid materials.

Pulse Column: A vertical pipe in which a dissolved material (e.g., plutonium) can be extracted or exchanged by the contracting of two immiscible solutions of differing densities. The lighter solution is usually fed into the bottom of the column, while the heavier solution is fed near the top. The process is enhanced by filling the column with suitable materials to produce droplets of one solution (e.g., perforated horizontal plates) and periodically pulsing the flow.

Purex: A reprocessing technology involving aqueous nitric acid dissolution of spent fuel and recovery and separation of uranium and plutonium using solvent extraction from nitric acid solutions using tributyl phosphate.

Pyroprocessing: A reprocessing technology that employs high temperatures and typically uses molten metals and salts as the reprocessing reagents.

Radiation Dose: A measure of the radiation energy absorbed in materials, especially living organisms.

Radionuclide: Same as a radioactive isotope.

Radiotracer: A radioactive isotope used to follow the movement or reaction path of a material.

Recycle: Reuse of the fissile and fertile products from reprocessing. Recycle involves reprocessing and refabrication.

Reprocessing: Chemical processing of spent nuclear fuel to recover desired constituents, typically uranium and plutonium.

Reprocessing and Recycle: A nuclear fuel cycle in which fuel constituents are recovered for reuse.

Segmentation: A process used to reduce fuel rods to short pieces that are open at the ends, thus exposing the spent fuel.

Separation Factor: The ratio of isotopic concentration in the product (enriched) stream to the concentration in the waste (tails) stream of a single stage of an enrichment process.

Shipping Cask (or Shipping Package): A container with thick walls of radiation-absorbing material used to transport spent fuel while providing protection from radiation and preventing the release of nuclear material during an accident.

Solvent Extraction: Contacting two immiscible liquids to selectively move a dissolved constituent of one liquid to the other.

Spent Fuel: Fuel that has been removed from a nuclear reactor because it has reached the end of its useful life.

Spontaneous Fission: A mode of radioactive decay in which a nucleus fissions without being struck by a neutron. Observed only in heavy actinide isotopes.

Stable Isotopes: Nonradioactive isotopes of an element.

Stage: A single step in an isotope separation or chemical separation process. In an isotope enrichment cascade, a single stage consists of all those separators that receive feed material of the same enrichment. The number of stages in a cascade primarily depends on the desired cascade product enrichment and waste stream assays and on the separation factor of the process. The definition is the same for chemical separation, but stages are not necessarily physically distinct units as is common in isotope enrichment.

Tails: Material from isotope enrichment that is depleted in the isotopes of interest.

Thorex: A variant of the Purex process that is used to recover thorium and uranium from thorium-based fuels.

Tonne (abbreviation t): A metric ton, 1000 kg.

Tributyl Phosphate (TBP): A water-immiscible liquid used in reprocessing. It is used to extract uranium and plutonium from an acidic aqueous solution. TBP is usually diluted with an organic solvent such as kerosene.

Uranium Concentrates: The product of a uranium ore concentration plant. Uranium concentrates typically contain 65 to 85% uranium, expressed as U_3O_8. Concentrates are usually in the form of UO_3, U_3O_8, or yellowcake.

Uranium Milling: Concentration and purification of uranium in ores.

Yellowcake: Ammonium or sodium diuranate, chemical forms of uranium concentrates.

Works Cited

American Nuclear Society (1978), "CIVEX: Solution to Breeder/Diversion Dilemma?" *Nucl. News* **21** (5), 32–37.

Becker, E. W., Bier, W., Ehrfeld, W., Schubert, K., Schutte, R., Seidel, D. (1975), *Uranium Enrichment by the Separation Nozzle Process*, Gesellschaft für Kernforschung mbH, Karlsruhe, Germany, Report KfK-2235.

Chapman, R. H., Spiewak, I., Tobias, M. L., Vondy, D. R. (1963), "Design of Small Aqueous Homogeneous Breeder Reactor," *Nucl. Sci. Eng.* **15** 347–353.

Croff, A. G., Hermann, O. W., Alexander, C. W. (1979), *Calculated, Two-Dimensional Dose Rates from a PWR Fuel Assembly*, Oak Ridge National Laboratory Report No. ORNL/TM-6754 [available from the U.S. Department of Energy Office of Scientific and Technical Information, P.O. Box 62, Oak Ridge, TN 37831].

Croff, A. G., Alexander, C. W. (1980), *Decay Characteristics of Once-Through LWR and LMFBR Spent Fuels, High-Level Wastes, and Fuel-Assembly Structural Material Wastes*, Oak Ridge National Laboratory Report No. ORNL/TM-7431 [available from the U.S. Department of Energy Office of Scientific and Technical Information, P.O. Box 62, Oak Ridge, TN 37831].

International Atomic Energy Agency (1992), *Transport of Radioactive Materials*, IAEA/PI/A.2E 92–04899 [available from the International Atomic Energy Agency, Division of Public Information, P.O. Box 100, A-1400 Vienna, Austria].

Katcoff, S. (1958), "Fission Product Yields from U, Th, and Pu," *Nucleonics* **16** (4), 78–93.

Lederer, C. M. Shirley, V. S. (1978) *Table of Isotopes*, 7th ed., New York: Wiley.

Organization for Economic Cooperation and Development (OECD) and the International Atomic Energy Agency (1992), *Uranium 1991: Resources, Production, and Demand* [available from OECD Publications and Information Centre, Suite 700, 2001 L Street NW, Washington, D.C. 20036–4910, and numerous other locations worldwide].

Weinberg, A. M., Rosenthal, M. W., Briggs, R. B., Kasten, P. R., Haubenreich, P. N., Engel, J. R., Grimes, W. R., McCoy, H. E., Beatty, R. L., Cook, W. H., Gehlback, R. E., Kennedy, C. R., Koger, J. W., Litman, A. P., Sessions, C. E., Weir, J. R., Whatley, M. E., McNeese, L. E., Carter, W. L., Ferris, L. M., Nicholson, E. L., Scott, D., Eatherly, W. P., Bettis, E. S., Robertson, R. C., Perry, A. M., Bauman, H. F. (1970), series of papers on the molten salt breeder reactor, *Nucl. Appl. Tech.* **8** (2), 105–215.

Further Reading

Benedict, M., Pigford, T. H., Levi, H. W., (1981), *Nuclear Chemical Engineering*, 2nd ed., New York: McGraw-Hill. A comprehensive technical book on most of the processes involved in the nuclear fuel cycle.

Gregory, J. N., (1966), *The World of Radioisotopes*, Sydney: Halstead Press. (1966). Although a bit dated, still provides a reasonably comprehensive overview of the uses of radioisotopes.

Lederer, C. M., Shirley, V. S., (1978), *Table of Isotopes*, 7th ed., New York: Wiley. An extremely detailed compendium of all isotopes known at the time of publication and their nuclear characteristics, including half-lives and radioactive emanations.

Long, J. T., (1967), *Engineering for Nuclear Fuel Reprocessing*, New York: Gordon and Breach. A comprehensive technical book on the equipment used for nuclear fuel reprocessing.

National Research Council, (1982), *Separated Isotopes: Vital Tools for Science and Medicine*, Washington, DC: National Academy Press. Provides a good overview of the production and uses of enriched stable isotopes, which are often the precursors of radioisotopes.

Organization for Economic Cooperation and Development and the International Atomic Energy Agency, (1983), *Uranium Extraction Technology*, Vienna: OECD and IAEA. An overview of uranium mineralogy, extraction, and processing.

Wymer, R. G., Vondra, B. L., Jr. (Eds.), (1981), *Light-Water Reactor Nuclear Fuel Cycle*, Boca Raton, FL: CRC Press. A more descriptive book on the nuclear fuel cycle with emphasis on reprocessing and waste management technology.

NUCLEAR MEDICINE

THOMAS J. RUTH, *UBC/TRIUMF PET Program, University of British Columbia and TRIUMF, Vancouver, British Columbia, Canada*

	Introduction	513
1.	**Historical Background**	513
2.	**Radioisotopes; Radionuclide Production**	514
3.	**Radioactive Tracers**	517
4.	**Radiopharmaceuticals**	518
5.	**Radioimmunoassays**	519
6.	**Imaging Techniques**	520
6.1	Planar Imaging	520
6.2	Positron-Emission Tomography	521
6.3	Single-Photon–Emission Computed Tomography	522
7.	**Functional Imaging**	523
7.1	Central Nervous System	523
7.2	Cardiac Imaging	527
7.3	Bone Scans	529
7.4	Endocrine Systems	530
7.5	Kidney	531
7.6	Pulmonary System	532
7.7	The Liver, Spleen, and Bile	532
7.8	Tumor and Inflammatory Processes	533
8.	**Radiation Dosimetry**	533
9.	**Radiotherapy**	536
10.	**Future Aspects of Nuclear Medicine**	538
	Glossary	540
	Works Cited	540
	Further Reading	541

INTRODUCTION

Nuclear medicine makes use of the fact that certain radioisotopes emit γ rays with sufficient energy that the γ rays can be detected outside of the body. These radioisotopes are attached to biologically active compounds, called radiopharmaceuticals, which either localize in certain bodily tissues or follow a particular biochemical pathway. Modern nuclear medicine encompasses many disciplines, including the use of *in vitro* and *in vivo* techniques. However, the following discussion will concentrate on the uses of radioactive substances for the diagnosis or therapeutic treatment of human disease.

1. HISTORICAL BACKGROUND

Nuclear medicine (NM) has its origins in the pioneering work of the Hungarian doctor G. Hevesy, who, in 1924, used radioactive isotopes of lead as tracers in bone studies. Shortly thereafter, R. H. Stevens made intravenous injections of radium chloride to study malignant lymphomas (1926).

However, it was not until the discovery of artificially produced radioactive isotopes that the number of available species suitable for use as tracers began to increase. The 1931 invention, by Ernest Lawrence, of the cyclotron made it possible to produce radioactive isotopes of a number of biologically important elements.

The use of these artificially produced radiotracers continued with J. G. Hamilton and R. Stone using radioactive sodium clinically in 1937. In 1938, S. Hertz, A. Roberts, and R. D. Evans used radioactive iodine in the study of thyroid physiology, followed, in 1939, by J. H. Lawrence, K. G. Scott, and L. W. Tuttle studying leukemia with radioactive phosphorus. By 1940, J. G. Hamilton and M. H. Soley were performing studies in iodine metabolism by the thyroid gland *in situ* by use of radioiodine in normal subjects and in patients with various types of goiters.

In 1941, the first medical cyclotron was installed at Washington University, St. Louis, Missouri, where radioactive isotopes of phosphorus, iron, arsenic, and sulfur were produced. With the development of the fission process, during the second World War, most radioisotopes of medical interest began to be

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

produced in nuclear reactors. After World War II, the wide use of radioactive materials in medicine established the new field of what was then called atomic medicine, and only later became known as nuclear medicine. Radioactive carbon, tritium, iodine, iron, and chromium found increasing use in the study of disease processes.

In the 1950s, B. Cassen developed the concept of the rectilinear scanner, which opened the way to obtaining in a short amount of time the distribution of radioactivity in a subject. This development was followed by production of the first γ camera by H. Anger in 1953 (see BIOMEDICAL ENGINEERING). The original design was modified later in the decade to what is now known as the Anger scintillation camera, thus heralding the modern era of γ cameras whose principles are still in use today.

D. Kuhl and colleagues introduced the concept of transaxial tomogram using a transverse imaging device. Extensions of this work led to almost exact tomographic sections. Hounsfield introduced the x-ray–transmission computed tomography by application of the reconstruction techniques used in radio astronomy. Its extension to emission computed tomography (ECT) is the basis of single-photon–emission tomography (SPECT) and positron-emission tomography (PET).

P. Richards was instrumental in the development of the ^{99}Mo/^{99m}Tc generator system following its invention at Brookhaven National Laboratory in 1966. Technetium-99*m* produced via this generator system has become the most widely used radionuclide in nuclear medicine today, accounting for as much as 85% of all *in vivo* diagnostic procedures.

The modern era of nuclear medicine may become known as molecular medicine as the field translates advances in molecular biology and biochemistry into treatment of humans and the diagnosis of disease and anatomical abnormalities.

2. RADIOISOTOPES; RADIONUCLIDE PRODUCTION

Radionuclide production is indeed true alchemy, that is, converting the atoms of one element into those of another. This conversion involves the altering of the number of protons and/or neutrons in the nucleus (target). If a neutron is added without the emission of particles, then the resulting nuclide will have the same chemical properties as the target nuclide. If, however, the target nucleus is bombarded by a charged particle, for example, a proton, the resulting nucleus will usually be that of a different element. The exact types of nuclear reactions a target undergoes depend on a number of parameters, including the type of bombarding particle and the energy of these projectiles.

The binding energy of nucleons in the nucleus is of the order of 8 MeV, on average. Therefore, if the incoming projectile has more than this amount of energy, the resulting reaction may cause other particles to be ejected from the target nucleus. By carefully selecting the target nucleus, the bombarding particle, and its energy, it is possible to produce a specific radionuclide.

Specific activity is a measure of the number of radioactive species (atoms or molecules) as compared with the total number of species present in the sample. The specific activity is usually expressed in terms of radiation units per mass unit. The traditional units have been Ci/mole (Ci/g) or fraction thereof. If the only atoms of the labeled element present in the sample are those of the radionuclide, then the sample is referred to as *carrier free*. For example, a compound labeled with ^{99m}Tc will be carrier free, since there are no stable isotopes of technetium. However, in most cases, there are small quantities of unlabeled compounds that have a similar chemical behavior and can act as *pseudocarrier*. The specific activity of an isotope or radiopharmaceutical is important in determining the chemical or biological effect the substance may have on the system under investigation.

As an example of specific activity, assume that glucose has been labeled with 10 mCi of ^{11}C with a half-life of 20.3 min in such a way that one of the glucose carbon atoms is the radioactive label and the remaining five carbons are natural nonradioactive carbon. Its carrier free specific activity (SA) would be as follows: atoms of $^{11}\mathrm{C} = N = dN/\lambda dT = 10 \times 3.7 \times 10^7/0.693(20.3 \times 60) = 6.5 \times 10^{11}$. For every radioactive carbon atom, there is one glucose molecule, and, using Avogadro's number, the number of moles is then 1.08×10^{-12}, and the SA $= 9.3 \times 10^9$ Ci/mol.

If the radionuclide had been ^{14}C with its 5730-yr half-life, following the same process but using the λ for ^{14}C, the SA would be 62 Ci/mole. If the radiolabeled glucose had been prepared in a plant, the naturally occurring glucose would have lowered the SA because of the nonradioactive glucose molecules. Therefore, it is easy to see that the short lived radioisotopes potentially have a much higher specific activity.

The use of nuclear reactors for the production of radioisotopes relies on the fact that during the fission process in a reactor, there are large numbers of neutrons produced and thermalized by the surrounding media, usually water or graphite. These neutrons are ideal for initiating (n,γ) reactions. In some reactors, the higher-energy neutrons are used to produce radioisotopes via other reactions, for example, (n,p) or (n,α) reactions. Table 1 illustrates some neutron-induced reactions.

The fission process is a source of several widely used radioisotopes. For example, ^{90}Sr, ^{99}Mo, ^{131}I, and ^{133}Xe are all produced in reactors by fission and can be separated from the uranium fuel cells or from targets of enriched ^{235}U placed in the reactor for radioisotope production directly. The major drawback of using fission-produced materials is the production not only of isotopes of the desired species, but of large quantities of both radioactive waste material and other radionuclides as well. In producing ^{131}I from fission, the isotopes ^{127}I and ^{129}I also are formed, thus reducing the specific activity.

It is ironic that the first radionuclides artificially produced were produced on Lawrence's cyclotrons, but it was another 30 years before accelerator-produced radionuclides began to play a major role in the production of medically important radiopharmaceuticals. The principal advantage of accelerator-produced radioisotopes is the high specific activities that can be obtained. Another not insignificant advantage is that a smaller amount of radioactive waste is generated from charged-particle reactions, especially at low (30 MeV) bombarding energy.

Table 1. Reactor produced radionuclides.

Reactions
$^{14}N(n,p)^{14}C$
$^{31}P(n,\gamma)^{32}P$
$^{32}S(n,p)^{32}P$
$^{50}Cr(n,\gamma)^{51}Cr$
$^{124}Xe(n,\gamma)^{125}Xe \rightarrow {}^{125}I$

Cyclotrons designed for producing medical radioisotopes were initially capable of accelerating protons, deuterons, $^{3}He^{+2}$, and α particles. However, the primary radioisotopes used in medical applications can all be produced by protons, and the simplicity of design for proton-only cyclotrons has resulted in cyclotrons that accelerate H^{-} ions capable of two or more simultaneous beams of varying energies and intensities. The modern cyclotron is completely computer controlled and is capable of running for many days with minimal attention.

The production mode of the radionuclides used for diagnostic and therapeutic purposes uses both cyclotrons and reactors, as shown in Table 2. Table 3 shows the quantities produced for the common commercially available radionuclides and the change in demand over the last decade.

As can be seen from Table 4, the PET radionuclides are produced from either proton or deuteron reactions. Since most early PET research was performed at major research laboratories having accelerators capable of proton or deuteron production, these reactions were the standard. However, in the early 1980s, small, compact, proton-only cyclotrons became available, and cyclotrons specifically designed for producing PET radionuclides were installed in many hospitals. The major drawback of these proton-only cyclotrons is that, in some cases, an enriched target material must be used for sufficient product to be generated.

The success of the small low-energy cyclotron encouraged research into the design of even lower-energy accelerators, i.e., linear accelerators and cyclotrons of a few MeV extracted energy. To date, these accelerators are in the developmental stage, but there could be a proliferation of these devices in most medical centers if they prove inexpensive to purchase and operate.

The equation for the rate of production is

$$R = I\sigma t, \tag{1}$$

where R is the number of nuclei formed per second, I is the flux of bombarding particles per second, σ is the cross section (probability of the reaction occurring) in cm^2, and t is the target thickness expressed as the number of

Table 2. The production modes of some radionuclides widely used for diagnostic or therapeutic purposes. The principal mode of decay is given, along with the major γ-ray energies and their intensities (I) in percent of total decay rate.

Nuclide	Principal decay mode	E(I) keV (%)	Reaction
^{67}Ga	EC	93 (37) 185 (20)	^{68}Zn(p,2n)
^{99}Mo/^{99m}Tc	IT	140 (98)	^{98}Mo(n,γ), fission product
^{111}In	EC	171 (90)	^{112}Cd(p,2n)
^{123}I	EC	159 (83)	^{124}Xe(p,2n)^{123}Cs → ^{123}Xe →
^{131}I	β^-	364 (81)	^{130}Te(n,γ)^{131}Te →
^{127}Xe	EC	203 (68)	^{133}Cs(p,spall)
^{133}Xe	β^-	81 (36)	Fission product
^{201}Tl	EC	167 (11)	^{203}Tl(p,3n)^{201}Pb →

nuclei per cm^2 (n/cm^2). It is of historical interest to note that the unit for cross section is the barn, which is defined as 10^{-24} cm^2. The expression barn comes from the fact that the probability of a neutron interacting with a target is proportional to the area of the nucleus, which, compared to the size of the neutron, is as big as a barn.

The rate of production is, of course, affected by the fact that the resulting nuclide is radioactive and, thus, undergoes radioactive decay. For short-lived nuclides, the competing reaction rates, production and decay, will come to equilibrium at sufficiently long bombardment times, since the rate of decay is proportional to the number of radionuclei present. The point where equilibrium has been reached is called saturation, meaning that, no matter how much longer the irradiation occurs, the production rate is equal to the rate of decay, and no additional product will be formed. At shorter irradiation times, the fraction of product produced is related to the saturation factor given by $1 - e^{-\lambda t}$, where λ is the decay constant of the decaying nuclide, and to the bombardment time t. It is evident that an irradiation equivalent to one half-life would result in 50% of saturation. For practical reasons, an irradiation rarely exceeds three half-lives (90% of saturation), except for the shortest-lived radioisotopes.

For long-lived species, the quantity produced is usually expressed in terms of integrated dose or total beam flux (μA-h). For example, with a long-lived radionuclide such as ^{82}Sr ($t_{1/2}$ = 25 d), the amount produced will be essentially the same whether it is produced from 100 μA in 1 h or 50 μA in 2 h (both represent 100 μA-h of beam).

Finally, the other source of radionuclides used in medicine is the generator. A radioactive generator takes advantage of the cases where one longer-lived (parent) radionuclide decays, usually by β^- emission, to a shorter-lived (daughter) radionuclide. The chemical differences in the two elements are exploited to separate the daughter product from the parent. The parent radionuclide is produced by one of the methods described above and then attached to an inert substance from which the desired product can be eluted. The product can be used directly, as in the case of ^{82}Rb$^+$ from the Sr/Rb generator, or after undergoing a chemical reaction, as in the case

Table 3. Some of the principal commercially available radionuclides and the quantities, in curies, used over the last decade. Note the rate of growth for ^{123}I and ^{201}Tl.

Nuclide	$t_{1/2}$	Retail consumption (1982)	Retail consumption (1987)	Retail consumption (1990)
^{67}Ga	78.3 d	800	800	820
^{99}Mo/^{99m}Tc	66 h/6 h	100 000 (^{99}Mo)	120 000	150 000
^{111}In	68 h	150	160	185
^{123}I	13.2 h	75	1250	3100
^{131}I	8.0 d	10 000	10 000	14 000
^{127}Xe	36.4 d	100	100	100
^{133}Xe	5.2 d	25 000	25 000	45 000
^{201}Tl	73 h	500	2500	6000

Table 4. The most commonly used positron emitters and typical reactions for their production; chemical forms of precursors for labeling more complex chemical species and typical radiopharmaceuticals.

Nuclide	$t_{1/2}$ (min)	Reaction	Precursor	Radiopharmaceutical
^{11}C	20.3	^{14}N(p,α)	CO$_2$, CH$_3$I	Many
^{13}N	9.97	^{16}O(p,α)	NH$_3$, NO$_3^-$	NH$_3$
^{15}O	2.03	^{14}N(d,n) ^{15}N(p,n)	O$_2$, H$_2$O, CO	same as precursor
^{18}F	109.7	^{18}O(p,n) ^{20}Ne(d,α)	F$^-$, F$_2$	Fluorodeoxyglucose Fluorodopa

of ^{99m}Tc from the Mo/Tc generator (see below).

The equilibrium equations (Sorenson and Phelps, 1987) that reflect the relative radioactivity of parent and daughter are given by the general equation

$$A_d(t) = A_p(0)[\lambda_d/(\lambda_d - \lambda_p)] \times (e^{-\lambda_p t} - \mathrm{e}^{\lambda_d t}) + A_d(0)e^{-\lambda_d t}, \qquad (2)$$

where A is the radioactivity for daughter (d) and parent (p), respectively. The initial radioactivity present is expressed as $A(0)$ for both parent and daughter. The first term in Eq. (2) accounts for the growth of the daughter as a function of the decay of the parent and the disappearance of the daughter due to its decay, and the last term accounts for the presence of any daughter nuclei at zero time.

All generator systems used routinely in nuclear medicine form equilibrium between parent and daughter. For the situation where the parent has a half-life much longer than the daughter, for example, with ^{68}Ge/^{68}Ga and ^{82}Sr/^{82}Rb, the change in the amount of the parent during the time for steady state to be reached will be negligible; that is, $\lambda_p \ll \lambda_d$ and $e^{-\lambda_p t} \approx 1$. Thus, by using Eq. (2), the quantity of daughter radioactivity at any time can be expressed by

$$A_d(t) = A_p(0)\,(1 - e^{-\lambda_d t}). \qquad (3)$$

Thus, after a very long time when $e^{-\lambda_d t} \approx 0$, the daughter and parent radioactivities are approximately equal. The steady-state condition is referred to as secular equilibrium.

In the case of the ^{99}Mo/^{99m}Tc generator, the parent (^{99}Mo) decays at a rate ($t_{1/2}$ = 66 h, where $t_{1/2} = \ln 2/\lambda$) long in comparison to its daughter (^{99m}Tc $t_{1/2}$ = 6 h), such that $\lambda_p < \lambda_d$, so that $e^{-\lambda_d t}$ is negligible in comparison to $e^{-\lambda_p t}$ when t is sufficiently long. Under these conditions, Eq. (2) can be solved so that the ratio of the daughter radioactivity to that of the parent is given by

$$A_d/A_p = T_p/(T_p - T_d), \qquad (4)$$

where T is the half-life for each species, respectively. This condition is referred to as transient equilibrium, even though, in the strict sense, it is a steady state and not a true equilibrium.

3. RADIOACTIVE TRACERS

The term radiotracers refers to a radioactive species that is used to follow (trace) the uptake into or function of an organ system in a living plant or animal. Initially, the radiotracers used were radioactive isotopes of naturally occurring elements or chemical congeners of these elements, as in the case where de Hevesy used Pb to trace bone activity (see above). Other early uses of simple elements as radiotracers were the use of radioiodine to monitor iodine uptake in the thyroid and the use of radioactive phosphorus to monitor metabolism.

The characteristics of an ideal tracer include the following:

1. The behavior of the tracer should be identical or related in a known manner to the substance to be traced; and
2. the quantity of tracer used should not have an effect on the process to be studied.

Inherent in the first of these characteristics is that there should be negligible effects due to the different isotopes present, as in the case of ^{131}I substituted for the natural isotope of iodine, ^{127}I. The rates of movement of atoms or molecules vary inversely as the square roots of their masses. Thus, for very light nuclei, e.g., isotopes of hydrogen, there can be a 73% difference between the diffusion rates of ^{1}H and ^{3}H. The difference in the diffusion rates for more massive isotopes is much less, as in the case of the iodine isotopes, where the difference is about 1.5%. Fortunately, this effect is generally of no consequence in nuclear medicine.

Radioactive isotopes of the elements of sodium and potassium (^{24}Na, ^{42}K) have been used to measure the sodium and potassium content of the body, employing the *isotope dilution* technique. Analysis by isotope dilution involves adding a known mass and specific activity of a substance to a mixture containing an unknown quantity of the substance and taking an aliquot of the mixture and determining the new specific activity of the substance under investigation. It can be shown that the mass of the substance of interest in the unknown mixture can be expressed as

$$M_u = M_1(S_1/S_2 - 1), \tag{5}$$

where M_u is the mass of the substance in the system under investigation, S_1 and S_2 are the specific activities of the tracer before and after its addition to the system, respectively, and M_1 is the mass of the labeled substance added.

4. RADIOPHARMACEUTICALS

The term radiopharmaceutical is derived from the fact that a radionuclide has been attached to a biologically active compound and is used for the diagnosis or the therapeutic treatment of human diseases. Radiopharmaceuticals differ in one major aspect from regular pharmaceuticals in that they are given in such small concentrations that they do not elicit any pharmacologic response. Because of this difference, there have been several attempts to change the name used to describe these substances. Present-day radiopharmaceuticals are used for diagnostic purposes in about 95% of the cases, and the remainder are used in therapy. The therapeutic effect is in response to the radioactivity associated with the radiopharmaceutical and not to the chemical nature of the compound, which is administered in very small concentrations.

In order for a radiotracer (radiopharmaceutical) to be used in humans safely, it must meet quality standards that include chemical and radiochemical purity, that it be sterile and free from pyrogenic material. The ideal radiopharmaceutical should

1. be readily available at a low cost;
2. be a pure γ emitter, that is, have no emission of particles such as α's and β's that contribute radiation dose to the patient but do not provide any diagnostic information (see Sec. 8);
3. have a short biological and physical half-life so that it is eliminated from the body as quickly as possible;
4. have a high target to nontarget ratio so that the resulting image has a high contrast; that is, the background does not blur the image; and
5. possess the proper metabolic activity in that it follows or is trapped in the metabolic process of interest.

The ability to measure regional biochemical function requires a careful design process with these principles in mind. However, it is not possible in reality to meet all of these criteria. For example, all decay processes involve the emission of particles, as in the case of the pure γ emitters, which have Auger electrons emitted during some fraction of the decays. Thus, it is necessary to address the following steps (Eckelman, 1991) in the development of a biochemical probe:

1. develop a radiotracer that binds preferentially to a specific site;
2. determine the sensitivity of the radiotracer to a change in biochemistry; and
3. find a biochemical change as a function of a specific disease that matches that sensitivity.

Figure 1 illustrates the ability to image different functional systems using the same radionuclide by varying the compound to which it is attached.

A large number of radiotracers have been synthesized to probe metabolic turnover such as oxygen consumption, glucose utilization, and amino-acid synthesis. Enzymatic activity, neurotransmission, receptor density, and occupancy have all been measured via appropriately designed radiotracers. It should be pointed out that the development of radiotracers for PET fundamentally violates rule number 2 for the ideal tracer, since, by na-

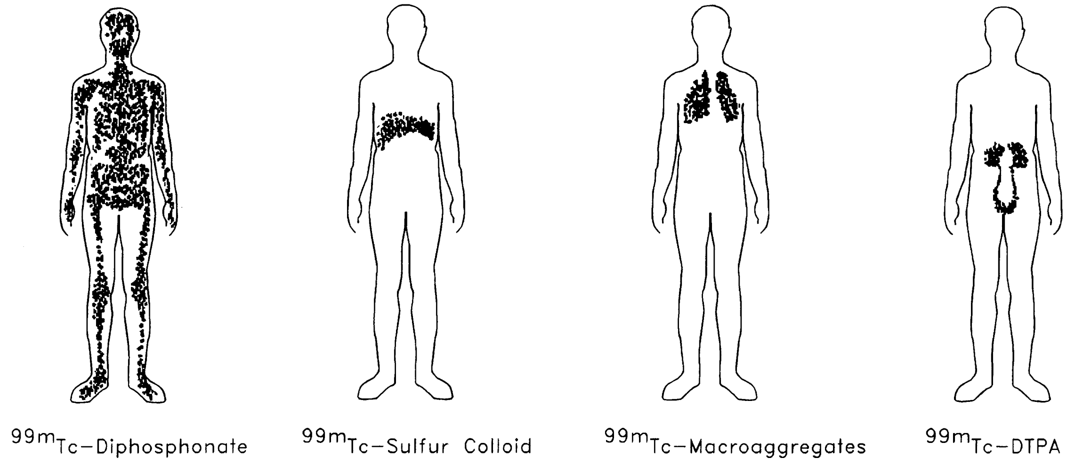
99mTc–Diphosphonate
99mTc–Sulfur Colloid
99mTc–Macroaggregates
99mTc–DTPA

FIG. 1. Dependence of distribution of radioactivity on the radiopharmaceutical used. Left to right: the skeleton imaged by ^{99m}Tc-methylene diphosphonate, the liver and spleen with ^{99m}Tc sulfur colloid, the lungs with ^{99m}Tc albumin microspheres, and the kidney-ureter-bladder system with ^{99m}Tc DTPA (diethylenetriaminepenta-acetic acid). (Used with permission of J. S. Robertson, Mayo Clinic.)

ture, the PET radioisotopes emit β^+ particles. However, the resulting coincident γ rays from the β^+ annihilation form the basis for the technique, which will be described below.

In addition to these principles, the synthetic chemist must plan on how to insert the radioisotope into the molecule at a point in the synthesis so that there is minimal handling, and late enough in the synthesis process to minimize loss due to chemical yield and radioactive decay. For these reasons, the preparation of radiopharmaceuticals requires planning and techniques not encountered by traditional synthetic chemistry.

5. RADIOIMMUNOASSAYS

On the basis of the radioisotope techniques developed earlier in the 1950s, S. Berson and R. Yalow published in 1959 the first use of radioimmunoassay (RIA) to measure the concentration of insulin in unextracted human plasma. The RIA principle is simple and is illustrated by the schematic in Fig. 2.

The concentration of an unknown amount of unlabeled antigen can be determined by comparing its inhibitory effect on the binding of radioactively labeled antigen to specific antibody with the inhibitory effect of

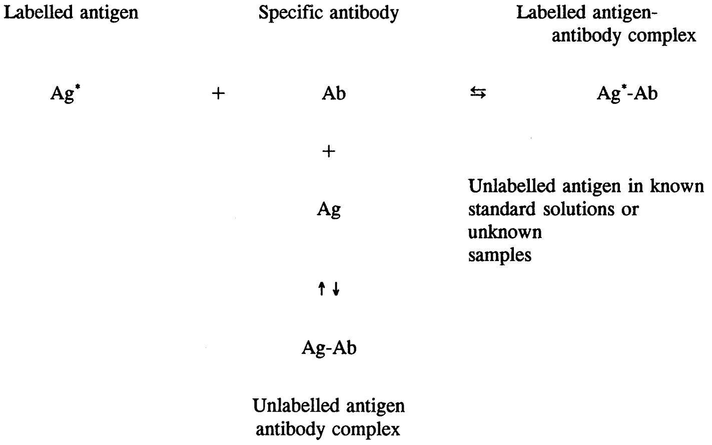
Labelled antigen
Specific antibody
Labelled antigen-antibody complex
Ag*
+
Ab
⇆
Ag*-Ab
+
Ag
Unlabelled antigen in known standard solutions or unknown samples
↑↓
Ag-Ab
Unlabelled antigen antibody complex

FIG. 2. Schematic representation of the reactions used in radioimmunoassay.

known standards (Yalow, 1978). Because of the use of radioactive tracers, the method is highly sensitive for detecting very small quantities. Concentrations of less than 1 p*M* can be determined for some species. The technique involves the separation of the labeled antigen of interest into bound and unbound fractions after its interaction with an antibody in the presence of an unknown amount of unlabeled antigen. The ratio of the bound to free fractions of the labeled antigen is compared to the binding of known standards.

In order to use this method to measure the concentration of a substance, the antigen in the samples to be tested and the antigen in standard samples must have identical immunologic behavior. Therefore, it is not necessary for the unknown and the standards to be chemically identical or to have identical biological behavior. The original technique has been modified by making use of a variety of radiotracers (Harbert and da Roche, 1984) to measure concentrations of vitamins, enzymes, peptides, serum proteins, hormones, viruses, drugs, and tumor antigens, as shown in Table 5.

Commercially available *kits* provide the materials for performing RIA for most of these substances. A typical kit would consist of a series of standard samples, each containing a specific amount of an unlabeled antigen, a vial of labeled antigen, a vial of antibody solution, and a substance used to precipitate the antigen-antibody complex. In most cases, ^{125}I is the radioactive isotope employed.

6. IMAGING TECHNIQUES

The following sections describe the techniques used to obtain *in vivo* functional images following the injection of radiotracers. The underlying basis for the techniques is that the radioisotopes employed must emit photons of sufficient energy that they penetrate the body so that they may be detected. All modern nuclear-medicine imaging devices make use of a scintillator coupled to a device that converts the emitted light into an electrical pulse. The following sections describe the more common methods presently in use.

6.1 Planar Imaging

By far the most common imaging device in nuclear medicine is the planar camera or Anger camera, named after its inventor, H. Anger. The basic components of the camera include a thin crystal of NaI scintillator coupled to a cluster of photomultiplier tubes and an (x, y) positioning circuit, and a readout device that may be an oscilloscope or film. In between the radioactive source and the detector is a collimator constructed of lead metal. The collimator has one or more holes drilled through it to allow the γ rays through. Since γ rays cannot be bent or focused, the collimator's function is to absorb those γ rays that do not pass through the openings.

In its simplest form, the collimator has a single pinhole and acts like a camera lens. Other collimators are constructed with the holes converging, diverging, or parallel to the imaginary line connecting the object to be imaged and the camera face. The converging collimator has the effect of magnifying the imaged object, while the diverging collimator reduces the image. The parallel collimator is used for high resolution. Regardless of which collimator is used, they all absorb a large fraction of the photons emitted by the radiotracer in the patient.

Table 5. A few examples of the application of radioimmunoassay.

Test	Substrate	Radioisotope	Applications
Schilling test	Vitamin B_{12}	^{57}Co	Indication of absorption of vitamin B_{12} as in pernicious anemia.
Blood volume	Serum albumin	^{131}I	Determination of blood volume in pre- and postoperative patients
Red blood cell survival	Red cells	^{51}Cr	Hemolytic anemia patients have a very short red cell survival time.
Plasma iron turnover	Ferrous citrate	^{59}Fe	Hematologic disorders have abnormal iron turnover rates

The NaI scintillator has been designed to minimize multiple interactions with the incident γ rays so that the position of interaction can be accurately determined. Typical scintillation cameras have detectors 25–45 cm in diameter with thicknesses in the range of 0.64–1.27 cm.

The (x, y) positioning circuit relies on the light output from the many (19–91) photomultiplier tubes (PM tubes) mounted on the back of the scintillator. The photomultiplier tube closest to the position of the γ ray's interaction will receive the maximum amount of light, while the other PM tubes will receive light in proportion to the solid angle subtended by the tube at the point of interaction. The positioning circuit sums the output of the PM tubes and produces x and y pulses proportional to the (x, y) coordinates of the γ-ray interaction.

A more detailed discussion of the Anger camera can be found in BIOMEDICAL ENGINEERING.

6.2 Positron-Emission Tomography

Positron-emission tomography (PET) makes use of the unique characteristics of the radionuclides that decay by emitting positrons. When these positrons lose energy through their interactions with the electrons of the surrounding matter, each positron will annihilate with an electron, giving rise to two photons, or γ rays, which radiate at 180° from each other. This conversion of matter into energy is an example of the energy–mass relationship described by Einstein. With an appropriate array of detectors, these coincident annihilation γ rays can be detected. Thus, the annihilation event is self-collimating; that is, the back-to-back photons define the line of response between two detectors. By mathematically back-projecting the lines of response, it is possible to reconstruct a density map for the location of the radioactivity. If the positron radionuclide is attached to a biologically active compound (radiopharmaceutical), then the spatial and temporal distribution in an organ or a functional system can be determined.

The spatial resolution of a PET system is determined by the ability to measure the origin of the two annihilation photons. The theoretical limits arise from the range of the positrons in tissue and the lack of true collinearity of the emitted photons. The first limit is determined by the kinetic energy of the positron, which for the typical PET radionuclides have a range of 2–20 mm in tissue. The second limiting factor arises from the fact that the annihilation event occurs when the positron-electron pair still have a small amount of kinetic energy, and, in order to conserve momentum, the two photons are emitted at an angle slightly different from the 180° for the at-rest situation. The effect of this lack of collinearity depends on the detector separation, which means the larger the diameter of the tomograph, the greater the effect. Thus, the best resolution that can be expected for a PET scanner is of the order of 2 mm. Because positrons are emitted from the decaying nucleus over an energy spectrum and only a very small fraction have the maximum range, the broadening associated with positron range amounts to only a fraction of a millimeter for ^{18}F and ^{11}C, the two most widely used positron-emitting nuclides.

Initially, PET scanners were constructed with a single ring of detectors. However, the limited axial field of view made it difficult to perform meaningful studies. Subsequent designs employed multiple rings of detectors, separated by lead or tungsten septa to collimate the γ rays into planes or slices. Thus, the tomographs could acquire data over an extended field of view, and reconstructed images could be displayed in three dimensions.

The design of the detector system has the constraints of high efficiency to minimize the radioactive dose given to the patient and a small detector for finer resolution with the added complication and expense of more detectors and the associated electronics. The modern tomograph designers have solved this problem by using a block detector, which has many small detectors sharing a smaller subset of photomultiplier tubes (PMT); for example, a block of detector material is cut into an 8×8 array sharing four PMTs. Using Anger logic, the location (which crystal segment) of the incident photon can be accurately determined.

The strength of PET is the ability to determine the amount of attenuation caused by the surrounding material, e.g., the torso in a cardiac scan, by using an external positron source, usually ^{68}Ga in equilibrium with its parent ^{68}Ge. Since the count rate is determined by the coincidence rate, only those γ

rays that are not absorbed by the intervening material are registered. Thus, by comparing the count rate with and without the attenuating material (the patient), it is possible to correct for the attenuation that the γ rays originating internally will experience.

The sequence of measurements in a PET scan include a blank scan with a positron source without the patient, followed by the same measurement with the patient in place. This determines the attenuation factors. This is followed by removal of the external source and the injection of the PET radiotracer, and a third scan. The reconstructed image of this last measurement is corrected for attenuation.

The other factors affecting the quality of a PET scan are the amount of scatter of the γ rays and the number of random coincidences. Scatter is caused by one, or both, of the coincident γ rays scattering from the subject or from other material in the scanner, giving rise to an incorrectly encoded line of response. The γ rays are detected by different pairs of detectors than if they had not been scattered, which results in the blurring of the image. Scatter is corrected for by determining the scatter function at the edge of the field of view and extrapolating a profile across the field of view. The correction is applied before attenuation corrections and reconstruction are performed.

Random coincidences arise from the limitations of the detector design, which requires a finite timing window to allow the two γ rays to be registered in the opposing detectors and their electronics. The coincidence window for modern tomographs is usually set at between a couple of nanoseconds and 20 ns. The random rate is determined by

$$r = \tau N_1 N_2, \qquad (6)$$

where τ is the coincidence resolving time and N_1 and N_2 are the count rates in the opposing detectors, respectively. Obviously, the shorter the resolving time, the lower the random coincidence rate; and conversely, the higher the individual detector counting rates, the higher the random coincidence rate. Because the random events have no directional character, the effect is to add noise uniformly throughout the image. Corrections for random events are made either by determining the single rates directly and applying Eq. (6) or, alternatively, by the use of a separate circuit with time delay to measure the random rate.

Another important factor in PET is that the radionuclides that decay by positron emission are the light elements associated with biological systems—carbon, oxygen, and nitrogen. Fluorine's atomic size makes it a suitable substitute for hydrogen. Thus, it is possible, in principle, to construct nearly any compound found in a biological system with a positron emitter. When using ^{18}F as the label, the resulting analog often has similar properties to the natural compound. The difficulty is, of course, the short half-lives of these isotopes, which makes the job of synthesizing them extremely difficult. The basic strategy is to insert the label as near the end of the synthesis as possible. There have been hundreds of compounds labeled with positron-emitting isotopes. However, simply having a labeled compound does not make it useful. What is required is a complete understanding of the metabolic processes the compound undergoes and how the label follows that process, because the PET image detects the location of the radionuclide, not its chemical form. Thus, modeling of a PET radiopharmaceutical is an important component in PET. Nevertheless, even with the complexity of synthesizing PET radiopharmaceuticals and the modeling required to understand the images, PET has proved to be an important tool in understanding basic *in vivo* biochemistry. There are a small number of compounds that are beginning to see use in clinical medicine and are mentioned in the relevant sections below. In particular, 2-[^{18}]F-fluoro-2-deoxyglucose (FDG) has been used to assess myocardium viability, as well as to detect hypermetabolic tissue associated with tumor growth.

6.3 Single-Photon–Emission Computed Tomography

As with PET, single-photon–emission computed tomography (SPECT) acquires views of the emitted photons from many different angles and reprojects these views to reconstruct the image or distribution of radioactivity in the object or patient. In SPECT, the radiopharmaceuticals used contain radioisotopes such as ^{99m}Tc that emit single photons (ones that are not in timed coincidence with one another). Directional information is achieved by collimating the photons incident on the detector. The collimator thus reduces the sensitivity of the camera, because all of the photons not parallel to the holes in the collimator are prevented from reaching the detector surface.

Since the reconstructed image contains the three-dimensional information on the distribution of radioactivity, SPECT also has the potential for quantification. The factors affecting this capability are similar to those in PET, for example, system sensitivity, dead time, spatial resolution, sampling interval, reconstruction filters, and the size of the object being imaged. Also, as in PET, the photons emerging from the subject are attenuated by the amount of matter between their origin and the detector and, of course, have a definite probability of being scattered along their path. These last two nonlinear effects are the most significant and difficult to correct for.

Because of the inherently lower energies (100–150 keV) of the photons emitted by radionuclides used in SPECT, the effect of attenuation can be quite dramatic, with reductions as great as a factor of 5 or more. Thus, with single-photon emitters, it is nearly impossible to determine whether data reflect a weak source near the surface or a stronger source located at some greater depth. In addition, the amount of scattering is strongly dependent on the energy of the photons, with the photons from ^{201}Tl having a scattering fraction of as much as 40–50% depending on the depth of the source.

Whereas scintillation camera images show the distribution of the radiopharmaceutical in defined regions in planar view, they suffer from the superimposition of organs and background contributions to the areas of interest. It is because of these shortcomings in planar imaging that SPECT has a major role to play in diagnostic imaging, regardless of whether SPECT can achieve the difficult task of providing quantitative information. The ability to view the distribution in three dimensions greatly affects the interpretation of the images. Figures 4 and 6 illustrate this point.

7. FUNCTIONAL IMAGING

Nuclear-medicine imaging differs from other radiological imaging in that the radiotracers used in nuclear medicine relate to the function of an organ system or metabolic pathway, and, thus, the tracing of these agents reveals the integrity of these systems or pathways. This is the basis for the unique information that a nuclear-medicine scan provides, as indicated in the variety of agents used to monitor the systems indicated in Table 6 and described in the following sections for various scanning procedures for the various organ or functional systems of the body. The structures of some of the ^{99m}Tc compounds are shown in Fig. 3.

7.1 Central Nervous System

Unlike other body tissues, the nervous system is highly discriminatory as to which solutes are transported into its tissues. This selective permeability is called the blood–brain barrier. The ability to measure changes in the blood–brain barrier (BBB) with radiophar-

Table 6. Standard nuclear-medicine procedures along with the radiopharmaceuticals commonly used for such procedures.

Standard radionuclide procedures	Radiopharmaceutical[a]
Brain perfusion	^{99m}Tc-HMPAO, ^{99m}Tc-ECD
Brain permeability	^{99m}Tc-DTPA
Blood-pool imaging	^{99m}Tc-RBC ^{99m}Tc-albumin
Bone imaging	^{99m}Tc-diphosphonates
Hepatobiliary	^{99m}Tc-DISIDA
Infection imaging	^{67}Ga-citrate Radiolabeled WBC
Liver–spleen	^{99m}Tc-colloid ^{99m}Tc-IDA
Lung	^{133}Xe (^{127}Xe) ^{99m}Tc-MAA
Myocardial perfusion imaging	^{201}Tl, ^{82}Rb, ^{18}FDG ^{99m}Tc-isonitriles ^{99m}Tc-teboroxines
Renal imaging	^{99m}Tc-DTPA ^{99m}Tc-MAG3
Thyroid imaging	^{123}I$^-$ ^{99m}Tc-pertechnetate
Tumor imaging	^{67}Ga-citrate Radiolabeled antibodies

---

[a] Abbreviations: HM-PAO, hexamethylpropylene amine oxime; ECD, ethyl cysteinate dimer; DTPA, diethylenediaminetetra-acetic acid; RBC, red blood cells; DISIDA, diisopropyliminodiacetic acid; WBC, white blood cells; IDA, iminodiacetic acid; MAA, macroaggregated albumin; FDG, fluorodeoxyglucose; and MAG3, mercaptoacetylglycylglycylglycine.

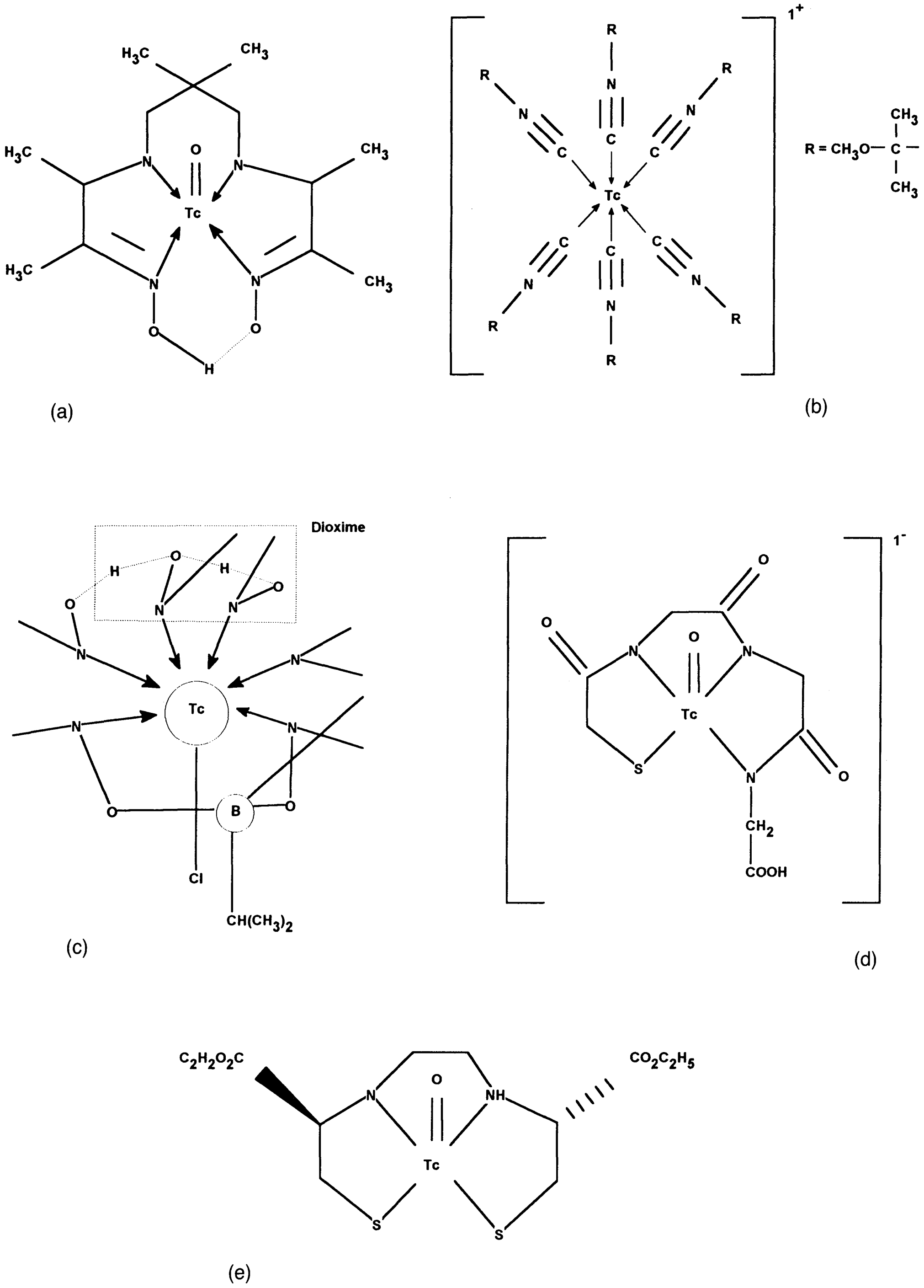
H3C
CH3
H3C
O
CH3
N
N
Tc
H3C
CH3
N
N
O
O
H
(a)
R
N
C
Tc
R = CH3O— C —
CH3
CH3
1+
(b)
Dioxime
H
O
H
O
N
N
O
N
N
Tc
N
N
O
B
O
Cl
CH(CH3)2
(c)
O
O
N
O
N
Tc
S
N
O
CH2
COOH
1-
(d)
C2H2O2C
CO2C2H5
N
O
NH
Tc
S
S
(e)

FIG. 3. Schematic of the structures of the ^{99m}Tc radiopharmaceuticals defined in the text. (a) Tc-HM-PAO, hexamethylpropylene amine oxime; (b) Tc-sestamibi, complexing with isonitriles; (c) Tc-teboroxime; (d) Tc-MAG3, mercaptoacetylglycylglycylglycine; and (e) Tc-ECD, ethyl cysteinate dimer.

maceuticals provided a major impetus for the growth of nuclear medicine over the last 25 years (Swanson *et al.*, 1990). In more recent times, the other radiological imaging techniques, such as computed tomography (CT) and, more recently, magnetic resonance imaging (MRI), have all but eliminated the use of radioactive agents to measure brain lesions.

However, recent developments in ligands that cross the BBB intact have opened the possibility of measuring functional changes in the brain associated with regional perfusion, metabolism, and the distribution and concentration of receptors. While very few of these new radiopharmaceuticals have achieved widespread clinical usage, the potential applications clearly indicate that brain imaging will again become an important aspect of nuclear medicine.

The brain is divided into two hemispheres along a fissure that runs from the anterior to the posterior. Each hemisphere has four lobes called the frontal, temporal, parietal, and occipital. The brain consumes up to 25% of the blood's oxygen supply and receives about 20% of all the blood that comes from the heart. The nervous system is bathed in a colorless fluid known as the cerebrospinal fluid.

The outer layer of the brain is composed of gray matter, which forms the cortex. Below this layer is the interconnecting tissue, made up of white matter interspersed with gray-matter areas. The gray matter is made up of neuronal endings, while the color of white matter is due to the myelin sheathes covering the axons.

Normal brain capillaries differ anatomically from capillaries found elsewhere in the body by their closed endothelial clefts and close approximation and continuity of the endothelial cells with the covering membrane. Molecules of glucose, water, and electrolytes apparently pass freely into the vascular space by passive diffusion and then into the brain cells themselves by active transport across its membrane, utilizing high-energy processes produced from the metabolism of glucose (the primary energy source of the brain).

Larger molecules are normally stopped at the endothelial lining, but, in disease, they may pass by pore filtration through the endothelial clefts widened by the disease process. These lesions that disturb the brain's integrity unselectively allow radiopharmaceuticals to permeate the nervous tissue.

Cerebral imaging normally consists of two phases (Harbert and da Roche, 1984), as follows:

1. a dynamic or angiographic study comprised of the arrival of the radioactive bolus in the cerebral hemisphere, which essentially constitutes a qualitative measure of brain perfusion, and
2. static images obtained approximately 1 h after the dynamic study, which provide a record of the distribution of the radiopharmaceutical in cerebral regions.

Scans at even later times can help delineate suspected lesions. The most commonly used agent for permeability studies is $^{99m}TcO_4^-$, pertechnetate, though the more expensive ^{99m}Tc-DTPA, diethylenetriamine-penta-acetic acid, is preferred because it clears more rapidly from the blood and does not accumulate in the blood vessels of the lateral ventricles (fluid-filled spaces in the brain), as does ^{99m}Tc-pertechnetate, resulting in a *hot spot* in the brain scan.

Single-photon–emission computed tomography (SPECT, see above) is gaining favor for three-dimensional perfusion studies of the brain. HM-PAO and ECD labeled with ^{99m}Tc are the agents of choice, hexamethylpropylene amine oxime and ethyl cysteinate dimer, respectively. These two compounds take advantage of the *p*H shift that occurs between the blood and the brain tissue (Saha, 1992). At the *p*H of blood (*p*H = 7.4), the agent is neutral and lipid soluble and can rapidly penetrate the intact blood-brain barrier. When it encounters the more acidic brain tissue (*p*H = 7.0), the agent acquires a charge, and, as an ion, it is incapable of reverse diffusion. Figure 4 depicts a blood flow image using SPECT and HM-PAO. A number of radioiodinated lipophilic agents, including IMP (isopropyl-*p*-iodoamphetamine) and HIPDM (*N*,*N*,*N*′-trimethyl-*N*-[2-hydroxy-3-methyl-5-iodobenzyl]-1,3-propanediamine), have shown promise but have been replaced by the technetium agents.

Although the concept of using positron emitters for biomedical investigations had been around since the 1951, positron-emission tomography (PET) did not become practical until the mid 1970s, based on the confluence of the PET scanner by Ter-Pogossian,

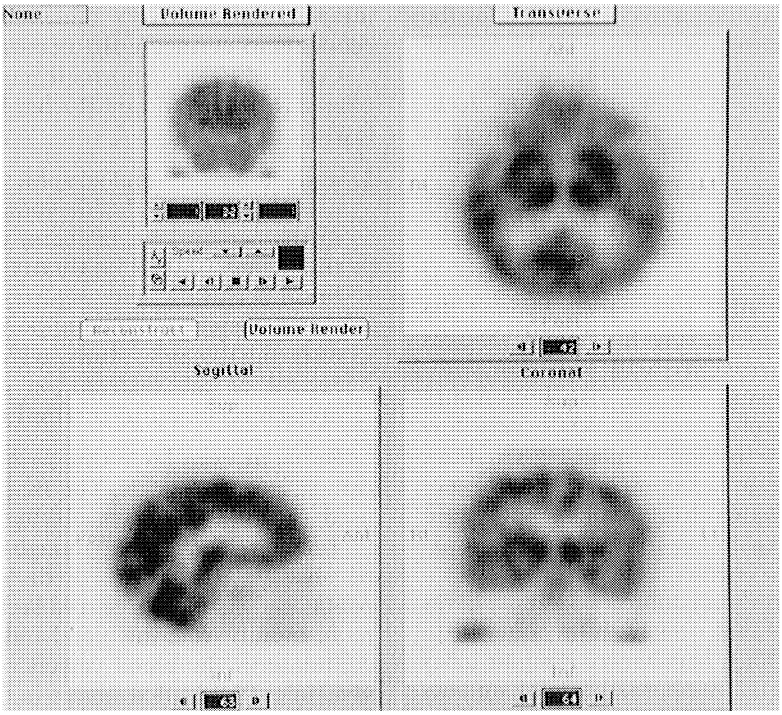

FIG. 4. SPECT image of the brain using ^{99m}Tc-HM-PAO, hexamethylpropylene amine oxime. The image is shown in a volume-rendered form, in the upper left, and three slice configurations: transverse, sagittal, and coronal. The abreviations are as follows: Ant—anterior (front); Post—posterior (back); Rt—right; Lt—left; Sup—superior (top); and Inf—inferior (bottom). (Courtesy I. Szaz, University Hospital, UBC Site.)

Phelps, and Hoffman (Ter-Pogossian *et al.*, 1975) in St. Louis and the development of the glucose analog ^{18}F-fluoro-deoxy-D-glucose (FDG) by Wolf and colleagues at Brookhaven National Laboratory (Reivich *et al.*, 1979). FDG is transported by plasma and crosses the BBB, where it is phosphorylated in a similar manner to glucose. The resulting FDG-6-phosphate is not a substrate for the other enzymes that are involved with glucose metabolism; thus, the labeled species is metabolically trapped. Once the relationship between FDG and glucose transport and uptake is taken into account, it is possible to obtain a quantitative measure of glucose metabolism on a region-specific basis. Mathematically, this relationship is achieved by determining the steady-state ratio of the net extraction of FDG to that of glucose at constant plasma levels of FDG and glucose and is referred to as the *lumped constant*. The FDG technique with PET has been used extensively in fundamental research of the functioning human brain to understand a number of normal brain functions, as well as disease processes. The clinical utility of this method has been employed to monitor tumor treatment and to locate the foci of epileptic seizures.

Oxygen consumption, blood flow, and blood volume have been measured using the ^{15}O-labeled molecules of O_2, H_2O, and CO, respectively. Neurotransmission and receptor density have also been measured for a number of neurological disorders, such as Parkinson's disease and schizophrenia. Figure 5 compares the uptake of ^{18}F-6-fluorodopa, FDOPA, in a normal subject and one affected by Parkinson's disease. The images in FDOPA scans reflect the uptake and conversion of fluorodopa to fluorodopamine and thus monitor the efficacy of the enzymatic systems involved in those processes. Research using PET is aimed not only at diagnosing disease but at an understanding of the disease process, as well as monitoring possible treatments.

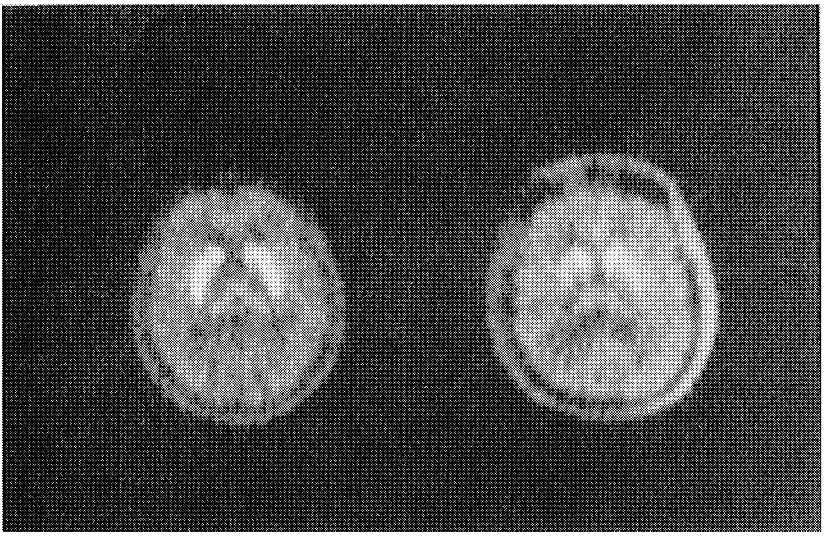

FIG. 5. PET image of ^{18}F-fluorodopa (FDOPA) uptake in (left) a normal patient and (right) a patient with Parkinson's Disease. Note the diminished uptake of FDOPA in the lower portion of the striatal region in the right image. (Courtesy UBC/TRIUMF PET Program.)

7.2 Cardiac Imaging

In general, nuclear medicine studies of the heart can be divided according to the clinical information sought (Swanson *et al.*, 1990). These studies include those that monitor the myocardium directly, such as perfusion assessment of metabolic activity, and determining whether the myocardium is healthy or infarcted. The imaging agents for perfusion use either potassium analogs or agents that are trapped in the myocardium in proportion to flow. The new technetium agents are included in the latter.

Among the perfusion agents, ^{201}Tl is the most widely used. Monovalent thallium has an ionic radius that is equivalent to that of K^+; thus, its uptake follows the active transport system for sodium and potassium. As much as 4% of the injected dose is taken up by the myocardial tissue, reaching its maximum in about 10–30 min after injection. The blood is cleared rapidly, with about 5% of the injected dose remaining in the blood pool after about 5 min.

^{201}Tl continuously redistributes between intracellular and interstitial compartments with differing rates for healthy and infarcted tissue. This behavior is typical for alkaline-like, singly charged cations. Thus, the equilibrium that is established upon redistribution can be used to distinguish between infarcted and ischemic tissue. In the case of ischemia, the diminished uptake disappears with redistribution, while persistent defects remain.

While the physiological behavior of ^{201}Tl is highly suited for perfusion imaging, its imaging characteristics are less than ideal. Principally, the energy of the emitted photons is extremely low (69–80 keV), which results in their being scattered, resulting in a loss in spatial resolution. Also, the half-life of ^{201}Tl (73 h) adds to the radiation burden for the patient. This radionuclide is cyclotron produced, making it expensive. Because of these negative attributes, there have been active attempts to find a ^{99m}Tc-based replacement.

Within the last decade, a series of ^{99m}Tc-labeled compounds have been proposed for myocardial perfusion studies (Saha, 1992). The

^{99m}Tc isonitriles are chemical complexes consisting of six monodentate CN*R* ligands surrounding a positively charged Tc^+ ion in the center, resulting in a stable cationic complex $Tc(CNR)_6{}^+$, where *R* represents one of the three functional groups tested thus far, tertiary butyl, carboxyisopropyl, and *hexakis*-2-methoxy-2-isobutyl groups. Of these agents, ^{99m}Tc-sestamibi (MIBI, where the complex is carboxyisopropyl isonitrile) has received the widest clinical usage. The myocardial uptake of this agent is proportional to regional myocardial blood flow. It does not redistribute with time, and, thus, a second injection must be performed the following day if a stress/rest protocol is desired. Figure 6 shows cardiac images acquired using ^{99m}Tc-MIBI.

Another class of technetium agents has been developed, consisting of boronic acid adducts of technetium dioxime (BATO). Unlike the isonitriles, the BATO complexes are neutral with respect to charge and are highly lipophilic. Lipid solubility is the uptake mechanism. ^{99m}Tc-teboroxime has undergone clinical trials. It consists of seven coordinate bonds coming from three dioxime molecules and one chlorine atom. ^{99m}Tc-teboroxime also does not redistribute and, thus, requires a second injection. BATO has a higher uptake in the myocardium than the isonitriles but does not clear as rapidly from the blood as does sestamibi and therefore does not have as high an image contrast.

$^{82}Rb^+$, as supplied from the ^{82}Sr generator, is the PET analog of K^+. Between 65 and 75% of the $^{82}Rb^+$ is extracted from plasma by the myocardium in the first pass at normal blood flow. Ammonia and water labeled with ^{13}N and ^{15}O, respectively, have also been used clinically to monitor cardiac blood flow. The use of these compounds is restricted by the fact that their short half-lives (^{13}N, $t_{1/2}$ = 10 min; ^{15}O, $t_{1/2}$ = 2 min) dictates production to be at or very near the site of use by an accelerator.

The metabolic function of the heart muscle can be monitored by a number of iodinated fatty acids (see below) or with the glu-

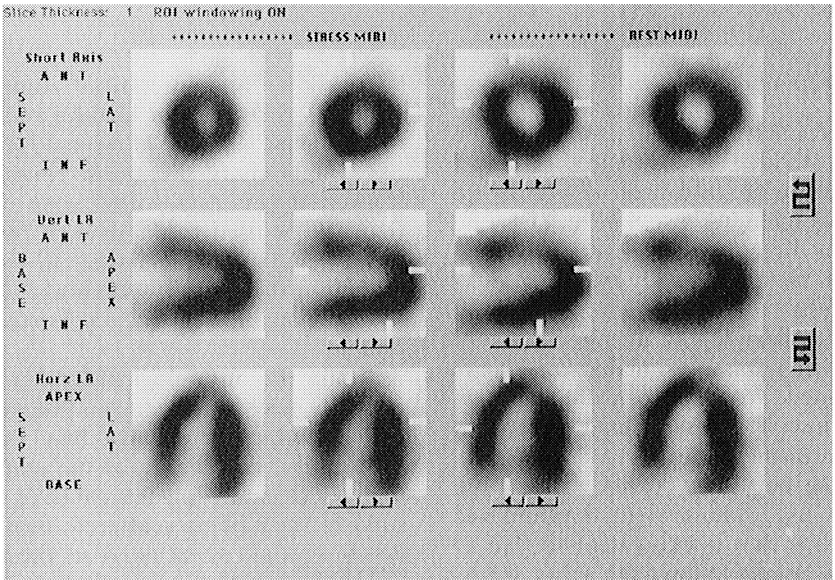

FIG. 6. SPECT image of the heart using ^{99m}Tc-MIBI, methoxy isobutyl isonitrile. The two columns on the left show the distribution in the heart at rest, and the two columns on the right show the distribution in the heart under stress. The top row has the images oriented in an oblique manner. This involves the reorientation of the heart images in such a way that slices at right angles to the long axis of the left ventricle become transverse or oblique images. The second row shows slices in a plane parallel to the long axis, from septum (the muscle that separates the the two chambers) to lateral wall (the thick outside wall of the heart), which are referred to as oblique/sagittal (Obl/Sag). The bottom row has the plane at right angles to oblique/sagittal and is the oblique/coronal (Obl/Cor). The abbreviations refer to ANT, anterior; INF, inferior; SEPT, septum; LAT, lateral; Base, base of the heart; APEX, the apex of the heart. (Courtesy I. Szaz, University Hospital, UBC Site.)

cose analog 2-fluoro(^{18}F)-2-deoxyglucose (FDG). FDG in conjunction with PET has the potential to measure the metabolic capacity of the myocardium quantitatively. Just as in the case for the brain, the metabolism of FDG essentially stops after phosphorylation, trapping the labeled analog in the cells.

The combination of ^{18}F-FDG and ^{13}N-ammonia with PET has received considerable interest in assessing the viability of the myocardium. Figure 7 compares a normal heart image with those that indicate ischemia. The ammonia scan measures blood flow and can detect obstructions in the blood supply to the myocardium, as shown in the figure. The FDG scan, however, indicates that the muscle is still capable of utilizing glucose and thus may return to normal function if reperfused via heart bypass surgery.

While FDG is an excellent compound for monitoring cardiac metabolism, its widespread use is hindered by the fact that these studies require a PET scanner, and, even though PET scanning is becoming more available, it is still only found in about 100 centers as of 1993. The primary energy source of the heart muscle under fasting conditions and rest is free fatty acids. A number of fatty acids have been labeled with ^{123}I. The most promising of these is ^{123}I-labeled *p*-iodophenylpentadecanoic acid (IPPA). The iodine atom is more strongly bound to aromatic carbons than to aliphatic carbons, thus minimizing the loss of the ^{123}I *in vivo*. With the favorable characteristics of ^{123}I and the rapid blood clearance of the fatty acids, IPPA holds promise for use in monitoring cardiac metabolism in the clinical setting, although the usefulness of such measurements has not been established.

Ventricular function is measured with blood-pool agents. These studies are performed by gating the data collection to the cardiac cycle. In this method, the radiopharmaceutical must reside in the vascular compartment for a sufficiently long time and at constant levels for the data acquisition.

7.3 Bone Scans

Although bone mineral is mainly composed of calcium, phosphate, and hydroxyl ions, living bone is a dynamic and physiologically active organ. The large adsorptive surface of bone provides chemically reactive sites that can be labeled with a wide variety of radionuclidic substances (e.g., Sr^{+2}, Pb^{+2}, Ra^{+2}, and Mg^{+2} for Ca^{+2}; F^- for OH^-; CO_3^{-2}, citrate, phosphate esters, diphosphonates, and pyrophosphate for phosphate).

In the formation of bone, amorphous calcium phosphate (ACP) precipitates on the surfaces of collagen fibers, where it grows over days and weeks to form hydroxyapatite [HA; $Ca_{10}(OH)_2(PO_4)_6$] (Swanson *et al.*, 1990). It has been shown that ^{99m}Tc-diphosphonate binds

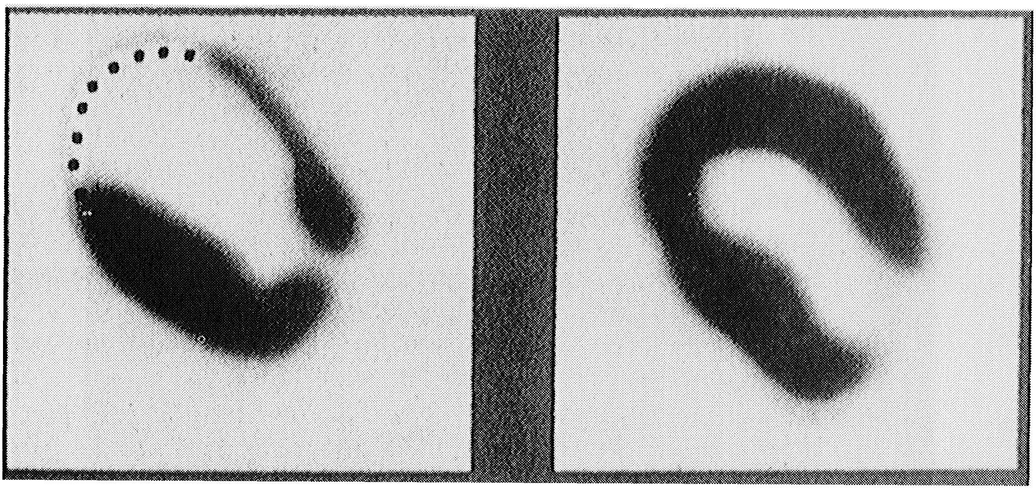

FIG. 7. PET image of the heart comparing blood flow with ^{13}N-ammonia on the left and metabolism with ^{18}F-FDG on the right. Note that there is a mismatch between flow and metabolism, indicating that the region of low flow is still capable of metabolizing glucose. (Courtesy M. Phelps, UCLA.)

to ACP to a higher degree than to crystalline HA. This selectivity forms the basis for detecting bone lesions, since areas of increased osteogenic activity contain higher concentrations of ACP than normal bone.

Shortly after its discovery in 1937, ^{18}F was discovered to have a high affinity for bone and teeth. This characteristic was exploited in the 1950s, when ^{18}F-fluoride was used extensively as a bone-scanning agent in nuclear medicine. With the advent of the ^{99m}Tc agents in the 1970s, the use of fluoride was replaced because of the better physical characteristics of half-life and photon energy, as well as the ease of delivery of ^{99m}Tc from a generator.

With the growing number of PET centers with their own cyclotrons, the interest in ^{18}F-fluoride as a bone-scanning agent has returned. The tomographic resolution and sensitivity for ^{18}F are superior to those of the single-photon systems. Thus, the resulting images can provide for accurate anatomical localization and improved lesion contrast. The first-pass extraction of fluoride from plasma is quantitative and in proportion to bone perfusion. Figure 8 illustrates the use of PET and whole-body ^{18}F-fluoride imaging.

7.4 Endocrine Systems

The endocrine systems regulate body functions through the secretion of various chemicals called hormones, from the Greek word meaning *arouse to activity*. The endocrine system includes the thyroid and parathyroid, pancreas, pituitary, gonads, and the adrenal glands. The thyroid gland helps control energy production, and the parathyroid provides for the distribution of calcium between the blood and bone. The adrenal glands regulate blood pressure, while the pancreas controls blood sugar levels. The pituitary controls growth, as well as regulating the activity of the gonads, the adrenal, and thyroid glands through the release of hormones. The gonads produce the germ cells of reproduction and a number of hormones related to the reproductive process.

The hormones generated by the thyroid that regulate the basal metabolic rate contain iodine. The iodine is accumulated in the thyroid from various food sources; thus, radioactive iodine becomes the simplest and nearly ideal tracer of metabolic function of the thyroid. With the production of ^{131}I from

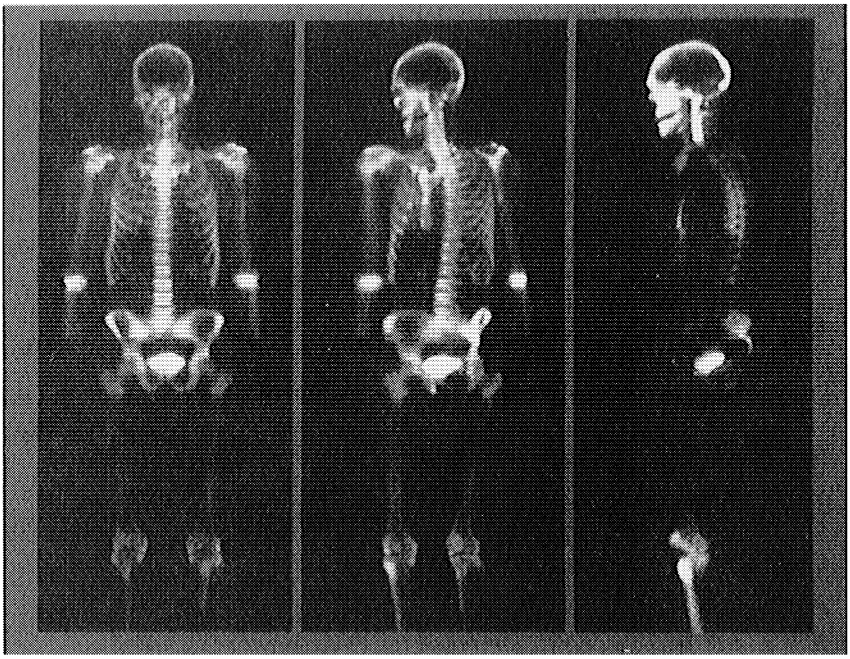

FIG. 8. Whole body PET scan of the skeleton using ^{18}F-fluoride in a normal male. The images from the left are the anterior projection, sagittal, and coronal images. (Courtesy C. Hoh, UCLA.)

reactors during and after the second World War, studies and treatment of thyroid diseases became widely available. There are a number of radioisotopes of iodine, but ^{123}I has the best physical properties for nuclear imaging, namely short half-life for low radiation exposure to patient, the energy of the principal γ ray matched to the present collimator systems on γ cameras, and no β-particle emission, which also reduces the dose to the patient in comparison to ^{131}I. Until recently, the production of ^{123}I required a high-energy accelerator. However, in the early 1980s, the use of a low-energy reaction on enriched ^{124}Xe to produce this radionuclide in high yield and extremely high purity increased the availability of this important radionuclide. The major drawback remains that its cost of production still makes the cheap and readily available ^{131}I competitive. As the reliability of ^{123}I becomes established, the universal acceptance of this radionuclide will be assured.

Radioiodine uptake studies utilizing sodium iodide are used to study hypothyroidism and hyperthyroidism, and to plan radioiodine treatment, the assessment of functional reserve, and the evaluation of thyroid gland autonomy. Radionuclide imaging of the thyroid is used to assess the size, position, and morphology of the thyroid gland, the detection of metastatic thyroid cancer and response to therapy, the evaluation of acute thyroid pain, the detection of accessory thyroid tissue, and the assessment of the functional status of nodules.

Because $^{99m}TcO_4^-$, pertechnetate, has a similar size and ionic characteristics to iodine (Swanson *et al.*, 1990), it is expected to be trapped in the thyroid. However, with the trapping, it is not incorporated into the functional hormonal system. Pertechnetate is used primarily for studying the morphologic structure of the thyroid.

Certain cells of the pancreas secrete enzymes that are used in the digestive system to react with proteins, carbohydrates, and lipids, while other cells secrete the hormones insulin and glucagon. Insulin controls the glucose level in blood. The amino acid methionine is incorporated in the digestive enzymes. Thus, an analog of methionine, selenomethionine, where the sulfur in methionine is replaced with radioactive ^{75}Se, has been used to image the pancreas to detect tumors, pancreatitis, and pseudocysts. However, because of high uptake in the liver and gastrointestinal areas, visualization of the pancreas is difficult.

The two adrenal glands positioned on top of the kidneys have two structures, an outer cortex and a central medulla. The adrenal cortical cells synthesize a number of steroid hormones from the circulating plasma cholesterol. Radioiodinated analogs of cholesterol, 19-iodocholesterol and 6β-iodomethyl-19-norcholesterol, have been used to image the cortical rim of the adrenal glands. Tumors and hyperfunction, Cushing's syndrome, can be diagnosed with increased uptake of these compounds.

7.5 Kidney

The urinary system is composed of the two kidneys, which are connected to the bladder via the ureters, and the bladder contents are discharged through the urethra. Just as in the case of the adrenal glands, morphologically the kidneys have an outer rim called the cortex and an inner area called the medulla. The nephron is the functional unit of the kidney and consists of tubules and ducts intermixed with the capillary bed of the blood. The nephrons perform three functions: the filtration of blood plasma, selective absorption of the materials the body requires, and the secretion of waste materials to the urine. The kidneys receive about 25% of the blood supply in each cycle, so that the entire blood volume passes through the kidneys every few minutes (Saha, 1992).

As part of the secretion process, there are two independent pathways for organic ions, and one for anions and one for cations. The pathway for anions has been exploited for imaging in nuclear medicine by the use of *o*-iodohippuric acid (OIHA), initially labeled with ^{131}I and more recently with ^{123}I. A new agent that is replacing OIHA is ^{99m}Tc-mertiatide or MAG3, a chelating agent based upon amide nitrogen and thiolate sulfur groups. Since the excretion rate of a drug by the kidneys depends on the molecular size, lipid solubility, and pK_a of the drug, and the volume of distribution of the drug in the body as well as the rate of transport through the renal tubules, it is possible to quantify renal function.

After injection of ^{99m}Tc-MAG3, dynamic images are acquired, followed by a delay to allow for plasma clearance, and a static image is then obtained. Renograms, or a display of the uptake and flow in the kidney, are obtained by placing a region of interest (ROI) on the temporal images of the kidneys and displaying the radioactivity present in the ROIs as a function of time.

^{99m}Tc-DTPA is often used to monitor the morphologic structure of the kidneys by acquiring a static image 30–60 minutes after injection (see Fig. 1).

7.6 Pulmonary System

The pulmonary system is imaged by the demonstration of pulmonary perfusion employing limited capillary blockade, as well as ventilation assessment with inspired inert gases, such as the noble gas xenon.

The airway of the respiratory system consists of the trachea, which divides into right and left main-stem bronchi, and these in turn divide to form lobar bronchi. The lobar divisions on the right are upper-, middle-, and lower-lobe bronchi, while, on the left, only upper- and lower-lobe bronchi exist. These lobes further divide into segments terminating in alveoli. In the adult, there are approximately 250 to 300 million alveoli, with an average diameter of 150 μm per alveolus. In addition to the direct pathway to the alveoli, there are pores and canals that interconnect them to allow collateral ventilation in the peripheral lung.

In the xenon ventilation studies, the patient breathes in the radioactive gas and holds his or her breath while the lungs are ventilated. Then the patient breathes while the air and xenon reach equilibrium, and, finally, the patient breathes in fresh air to wash out the xenon. Images are acquired in all three conditions. The ventilation studies can be used to determine if there is any airway obstruction, bronchitis, or emphysema.

For ventilation studies, ^{133}Xe is the most widely used noble gas, even though ^{127}Xe has better physical properties (higher photon yield, higher photon energy). The difference in availability has dictated the relative usage of these two isotopes of xenon (see Tables 2 and 3). Xenon-127 requires a high-energy accelerator to be produced, and the only accelerators producing ^{127}Xe at the present time are those associated with basic physics programs, thus making production and delivery unstable. Efforts to use the ^{126}Xe(n,γ) reaction may alter this in the future.

Perfusion studies employ the limited capillary blockade through the use of particles that become lodged in the capillary beds of the lung. The capillaries that surround the alveoli are between 7 and 10 μm in diameter. Thus, to achieve highest extraction efficiency, the ideal size for the particles is in the range of 10–20 μm. Macroaggregates (MAA) and human albumin microspheres (HAM) labeled with ^{99m}Tc have been used for 20 years to measure perfusion. More than 90% of the particles from either agent are extracted from the blood pool in the first passage through the lungs. Distribution in the lung is expected to be uniform, and, thus, the image appears "hot." Where blood flow is occluded or restricted because of emboli or vessel constriction, the particles will be prevented from passing these points of obstruction. The image thus appears with "cold spots" at these locations. Perfusion studies are used to diagnose obstructions caused by pulmonary embolism, tumor, infection, and tuberculosis.

7.7 The Liver, Spleen, and Bile

The lobules of the liver consist of two basic type of cells: hepatocytes and endothelial cells (Kupffer's cells). The hepatocytes make up about 85% of the cell population in the liver and are responsible for the major metabolic functions, while the endothelial cells make up the remaining 15% of all liver cells and serve to cleanse the blood. Each lobule consists of a central vein surrounded by a number of sinusoids, like spokes around the hub of a wheel. Sinusoids are vessels that transport blood from the portal vein to the central veins that eventually deliver blood to the hepatic vein (Saha, 1992).

The liver serves to metabolize the sugars and simple amino acids from digestion and to store the resulting peptides and proteins, as well as for the detoxification of harmful substances through methylation and conjugation. The liver excretes bile into the intestines through the hepatic duct.

The spleen consists of two main compartments. They are the white pulp, consisting of small lymphocytes and plasma cells, and red pulp, a swampy mass of vascular spaces whose supporting structure contains blood-cleansing cells. The two main functions of the spleen are antibody production in the white pulp and particle filtration in the red pulp. Blood entering the spleen through the splenic arterioles empties first into the white pulp, where plasma skimming occurs to remove soluble antigens. Passing through the white pulp, the blood enters the red pulp through an open meshwork of large oval cells at the junction of these two cell types. This cellular maze serves to remove abnormal cells and allows normal cells to proceed unhindered.

Labeled colloids are used to evaluate the blood-cleansing function of both the liver and the spleen. The most widely accepted agent today is ^{99m}Tc-sulfur colloid. After administering the ^{99m}Tc-sulfur colloid, the majority of the particles (0.3–1.0 μm) are accumulated by the reticuloendothelial cells of the liver (80–90%) and spleen (5–10%), and the remainder go to the bone marrow. In principle, the correlation between particle size and organ avidity for the colloid would have the spleen removing the larger particles, the smaller particles going to the liver and the smallest particles to the bone marrow. In the normal liver, a homogeneous distribution of radioactivity throughout the organ is expected. Any localized, space-occupying process in the liver may present as a focal area of decreased uptake referred to as a defect. Lesions as small as 8 mm in size may be identified, while those that are 20–25 mm in diameter are routinely imaged. Colloid scans are used to diagnose cirrhosis of the liver, abscesses, tumors, metastatic lesions, and hepatitis.

To evaluate the functional integrity of the hepatocytes, ^{99m}Tc-IDA has been developed. The ^{99m}Tc-IDA is given to fasting patients, where it rapidly appears in the liver and then progressively clears from the liver via the biliary tree, where major hepatic ducts and the gall bladder are able to be visualized.

7.8 Tumor and Inflammatory Processes

Although the precise mechanism by which tumor cells proliferate is not known, there are some common factors that can be exploited for imaging. With the increased metabolic activity and blood flow, along with high vascular permeability, agents that mark these events can be used to identify this change in function as compared to normal tissue. Tumors also have antigens associated with them that can be used to target a radiopharmaceutical specifically to all of the locations of the particular tumor.

The radiopharmaceuticals in routine clinical use are, in general, nonspecific with respect to tumor type and primarily reflect the altered functional state. In fact, ionic ^{67}Ga as gallium citrate was accidentally found to accumulate in soft-tissue tumors. It was quickly learned that ^{67}Ga localized in many nonmalignant areas, such as inflammatory lesions. It has thus become a common mode for evaluating patients with suspected infectious processes. Although the mechanism by which ^{67}Ga accumulates in these tissues is not understood, it is known that gallium quickly binds to transferrin in the blood and that this protein binding may be part of the process.

More recently, ^{111}In-labeled leukocytes have been used to detect abscesses in various parts of the body. Whereas ^{67}Ga is eliminated slowly through the gut, the labeled leukocytes are not excreted through the bowel and kidney. Thus, the abdominal background in this region is lower. The difficulty in labeling the leukocytes, however, is a major shortcoming.

As referred to in other sections, a number of other agents are used to monitor tumors associated with particular organs or functional systems. As the understanding of the use of monoclonal antibodies increases, the variety and specificity of these agents will increase their use.

8. RADIATION DOSIMETRY

The γ rays and particle emissions associated with the radionuclides used in nuclear medicine cause ionization to occur as they pass through matter. If the material these photons and particles are passing through is human tissue, these ionizations can break chemical bonds or kill cells if sufficient energy is deposited in these structures. Fortunately, in the case of nuclear-medicine imaging, the repair mechanisms of the human body usually prevent long-lasting damage from occurring. The restorative function in certain organs is more sensitive than others; thus, it becomes important to monitor the amount of energy deposited per unit mass in the various parts of the body by the decay of radioactive species. The energy absorbed from the ionizing radiation per unit mass of material is referred to as the absorbed dose. The *Système Internationale* (SI) unit for absorbed dose is the *gray* (Gy) and is equivalent to 1 J/kg. In spite of the efforts to make SI units the standard, the traditional units for radiation exposure and absorbed dose are commonly used. The term rad is used to express absorbed dose, and its relationship to the Gy is such that 100 rad = 1 Gy. The dose-equivalent unit (rem) was developed to account for the differences in the effects which various types of radiation have in causing biological damage. In radiation protection, the rem is defined as

$$\text{No. rems} = \text{No. rads} \times \text{QF} \times N$$

where QF is the quality factor and N the modifying factor of the radiation in question. The term QF is related to the amount of energy transfer to the surrounding medium the radiation traverses. The quality factor for photons is 1, while protons and neutrons have the value of 10 and heavy ions (α and other larger ions) have the value of 20, reflecting the greater loss of energy per unit length in matter for the heavier species. For practical purposes, the term N is assumed to be unity. The dose equivalent in SI units is the sievert (Sv), where 1 Sv is equal to 100 rem. Since it is difficult to determine the actual absorbed dose that radiation workers receive, it is common to refer to their radiation exposure in sieverts or rems and fractions thereof. However, for nuclear-medicine procedures, an effort is made to determine the patients' actual radiation exposure expressed in grays.

In order to estimate the absorbed dose due to a nuclear-medicine procedure, a method has been developed whereby dose estimates are calculated. These calculations require the knowledge of the biological distribution data and the physical properties of the radionuclide. The absorbed-dose method allows one to calculate the dose delivered to the tissue of interest, referred to as the target organ, from the radioactivity accumulated in a source organ, which may or may not be a number of sites within the body. The source organ can also be a target organ, which is generally the case when dealing with low-energy photons or particle emissions.

The actual process of calculating the dose (Loevinger *et al.*, 1988) includes the following steps:

1. The time integral of the amount of radioactivity (accumulated activity) that is in a source organ is determined by following the time course of radioactivity in the organ of interest.
2. The total amount of energy associated with the emissions of the radioactive species in the source organ is calculated. This is usually tabulated according to the type of radiation (penetrating or nonpenetrating), energy, and intensity (the frequency of the particular emission).
3. The fraction of the energy that is absorbed by the target organ is determined. The amount of energy absorbed depends on the type and energy of the emissions, as well as the physical relationship between the source and target organs, that is, the relative size, shape, and distance apart.

There are obvious difficulties in determining these quantities, especially in each human subject. Therefore, estimates and certain assumptions are made to simplify the calculations.

To calculate the cumulative activity A, the sum of all of the nuclear transitions in an organ is determined for the time interval of interest. Cumulative activity defines the time integral of the activity, which is proportional to the sum of all the nuclear transitions during a given time interval.

The rate of disappearance of the radioactive species from the body or organ within the body is due to the radioactive decay, which is dictated by the physical half-life of the radionuclide in question, and the biological processes that accumulate or eliminate the species. This latter process is measured as the biological half-life. Thus, the clearance from a source organ is measured by the effective half-life, a combination of the radioactive decay and biological elimination. Mathematically, this is represented by the sum of the decay constants for the physical and biological half-lives:

$$\lambda_e = \lambda_p + \lambda_b, \tag{7}$$

where the respective decay constants are equal to $(\ln 2)/t_{1/2}$ for each process, physical and biological. Residence time defines the average time that the administered radioactivity spends in the source organ. The residence time of a radiopharmaceutical in any particular organ is related to the fraction of the injected dose reaching the organ and to the chemical and biological processes that the radiopharmaceutical experiences.

Figure 1 illustrates how the distribution of different radiopharmaceuticals can affect the absorbed dose, not only to specific organs but also to the whole body. The compounds used in the figure are labeled with the same radionuclide, but the chemical form dramatically alters where they deposit in the body. The ^{99m}Tc-diphosphonate binds to the hydroxyapatite of bones and is distributed throughout the skeleton, while the spleen and liver have a high uptake of ^{99m}Tc-sulfur colloid. The lungs are clearly visible with ^{99m}Tc-albumin microspheres, and the kidney-ureter-bladder system concentrates the ^{99m}Tc-DTPA complex.

The distribution of the radiopharmaceutical will, of course, be dependent upon the functional integrity of the organ or functional system, which could be radically different in health and disease. Thus, to have as accurate an estimate of dose as possible, the actual distribution in the subject must be determined. Otherwise, the values used will only be population averages.

The data on the physical characteristics for the decay of the radionuclides used in nuclear medicine have, for the most part, been determined accurately. For a given radionuclide, each emission is characterized as to its energy and relative frequency. The product of these two values is the energy imparted per decay of the nuclide and referred to as the equilibrium absorbed-dose constant, symbolized by Δ. The value for this term in traditional units is g-rad/μCi-h, and Gy-kg/Bq-s in SI units.

The emissions associated with radioactive decay can be classified into two types of radiation, nonpenetrating and penetrating radiation. Nonpenetrating radiations are those particles emitted by the source that contribute a negligible dose to organs other than the source organ. These emissions include all of the particulate emissions, as well as very low-energy photons. Penetrating radiations are those particles emitted by the source that contribute a significant dose to organs other than the source organ. All of the emissions that can be detected outside of the body, as in the case for imaging, are obviously in this category.

The absorbed fraction is the fraction of energy emitted by the source organ that is absorbed by the target organ, and is given the symbol ϕ. The value of the absorbed fraction depends on the size, shape, and composition of the source organ, the attenuation factors between the two organs (distance and composition) of the intervening matter, along with the energy of the emission and its type. In the special case of nonpenetrating radiation (α and β particles, conversion and Auger electrons, and low-energy x rays or γ rays), where the source and target organs are one and the same, the value of ϕ is 1, and, for all other organs, ϕ equals 0. That is, all of the energy is absorbed in the source organ, while no energy is imparted to surrounding tissue.

Mathematical models have been derived to calculate the values of ϕ as a function of photon energy for pairs of organs in a standard adult of 70 kg. As an example, the fraction of absorbed energy in the heart (target organ) from a uniformly distributed source of 100-keV photons in the liver (source organ) equals 0.067. In establishing the values for ϕ, the reciprocity theorem has been used, which states that it does not matter which organ is the source and which the target, the relationship is the same. The energy absorbed per gram is the same for radiation traveling from organ A to organ B as it is in traveling from B to A.

To aid in the calculation of dose, a single parameter has been tabulated that is the product of the equilibrium absorbed-dose constant and the absorbed fraction per unit mass, and has been given the symbol S:

$$S = \Sigma \Delta \phi, \tag{8}$$

where $\Delta\phi$ is summed (Σ) over all emissions from the radionuclide.

The Medical Internal Radiation Dose Committee of the U.S. Society of Nuclear Medicine (MIRD Committee) has compiled tables of S values for organ pairs and all commonly used radionuclides in nuclear medicine. Thus, once the cumulated activity for a particular radiopharmaceutical is known for each organ, then the dose is determined from the product as in

$$D = AS, \tag{9}$$

for each organ in the body.

The *S* values are tabulated for about 20 organs acting both as source and target organs. The compiled values of the *S* are for a standard 70-kg man; therefore, in applying these values to other persons of differing size, the *S* values may not be correct.

Recently, the International Commission on Radiological Protection (ICRP) has adapted the concept of *effective dose equivalent* H_E for use in calculating dose to patients receiving radiopharmaceuticals (ICRP, 1988). The advantage of this approach is that a single figure can be used to express the radiation risk to patients undergoing different diagnostic procedures. The calculation is performed by summing the doses to the different organs in the body, weighted by a factor determined to reflect the relative sensitivity of that organ to radiation. For example, the weighting factors for gonads, red bone marrow, and skin are 0.20, 0.12, and 0.01, respectively. Tables of the effective dose per unit administered activity have been compiled by the ICRP for the standard radiopharmaceuticals. Table 7 gives some of these values. With this approach, it can be seen how the various procedures can be compared and, in the case of the same radiopharmaceutical, how the different radioisotopes affect the dose.

The calculation of absorbed dose is obviously a complex undertaking, but, as described here, a number of simplifications and assumptions have been made, so that useful values of estimated dose can be determined.

9. RADIOTHERAPY

The use of radionuclides for therapeutic purposes makes use of the modes of radioactive decay involving particulate emission. These modes include α and β decay (both positron and negatron), as well as any mode resulting in Auger electron and/or Coster–Kronig electron emission. The emission of γ rays may or may not accompany these particle emissions.

The use of particles emanating from the nucleus relies on the fact that these particles are stopped within a short distance. Alpha-emitting radionuclides are especially attractive in that α particles with 5–8 MeV energy have a range in tissue of about 40–80 μm. Figure 9 schematically illustrates the range of particles in tissue, along with a relative scale of tissue sizes to reflect the range along which the energy of the particles would be deposited. In choosing a therapeutic radionuclide, the precise site of localization of the radiopharmaceutical carrier in the tumor governs this choice. Thus, the Auger emitters must be transported to the nucleus if not incorporated into the DNA itself if *cell kill* is to be achieved. Table 8 gives some radionuclides that are considered to be candidates for use in radiotherapy.

Alpha particles with their high linear energy transfer (LET, the energy deposited per unit length by radiation as it passes through matter) are potentially lethal for cells if deposited on the cell surface or into the cytoplasm. There are very few acceptable α-emitting radionuclides because of the complex decay schemes associated with these isotopes. However, there are two candidates that have received special attention, ^{211}At and ^{212}Bi, both of which are cyclotron produced and have relatively short half-lives, which severely restrict their availability.

In addition to the types of particles emitted in their decay, there are a number of other considerations in assessing the suitability of radioisotopes for use in therapy. The physical half-life of a radionuclide used in therapy

Table 7. Radiation exposure due to selected radiopharmaceuticals. Note the factor of 100 difference between ^{123}I and ^{131}I in the thyroid uptake studies. This difference is due to the longer half-life of ^{131}I.

Radionuclide	Substance	*E* (mSv/MBq)
^{15}O	Oxygen gas, single breath, 20 s hold	4.2×10^{-4}
^{18}F	FDG	2.0×10^{-2}
^{99m}Tc	Tc-DPTA	5.2×10^{-3}.
^{99m}Tc	Pertechnetate	1.2×10^{-2}
^{123}I	Iodide (thyroid uptake 35%)	2.2×10^{-1}
^{131}I	Iodide (thyroid uptake 35%)	2.4×10^{1}
^{201}Tl	Thallium chloride	2.3×10^{-1}

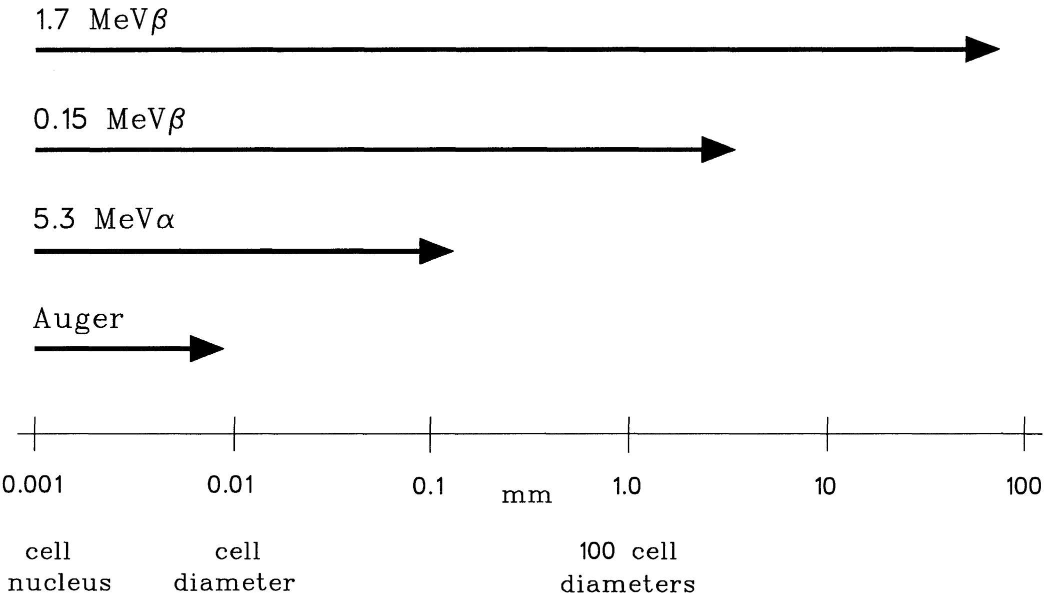

FIG. 9. Schematic representation of the ranges of particles in tissue in comparison to relative size of the cell structure.

must be long enough to allow the labeled species to clear the blood pool and concentrate in the region of interest but short enough to prevent the radionuclide from becoming a radiation burden to the rest of the body through metabolism of the primary agent. The physical half-life should be closely associated with the biological half-life of the radiopharmaceutical in the tumor. In this way, the bulk of the decays occur in the region where their damaging effects will be directed to the tumorous tissue and not to normal tissue.

One of the major difficulties in using pure β-emitting nuclides for therapy is the inability to detect their location accurately from outside of the body. However, a low-abundance γ ray with an energy of 100–200 keV can be used in imaging to assess localization and estimate delivered dose. Thus, the accompanying γ ray(s) should be in low abundance to remove radiation burden from surrounding healthy tissue.

The availability of a radionuclide and its associated chemistry determine its practical utility as far as receiving widespread clinical use is concerned. Just as in the case of the α emitters, even with their nearly ideal particulate energies, the difficulties in producing them have not only restricted the availability for clinical assessment but also basic research. The chemical form of the radionuclide has ramifications in determining how the radionuclide will be delivered to the site of interest. In this regard, ^{131}I has been con-

Table 8. Radionuclides with potential use as radiotoxic agents. The half-life of the radionuclide is included along with the particle emitted and its energy and the principal γ ray with its intensity (I) in percent.

Radionuclide	$t_{1/2}$	Particle	$E_{\max}$ (MeV)	E_γ (MeV) (I)
^{90}Y	64 h	β^-	2.3	None
^{111}In	2.83 d	Auger e^-	···	0.17 (90)
^{125}I	60.1 d	Auger e^-	···	0.035 (6)
^{131}I	8.1 d	β^-	0.6	0.36 (81)
^{153}Sm	46.7 h	β^-	0.81	0.103 (28)
^{186}Re	90.6 h	β^-	1.07	0.14 (9)
^{211}At	7.2 h	α	5.87 (7.45 from ^{211}Po)	0.69 (0.2)

sidered the nearly ideal radiotherapeutic agent for thyroid carcinoma because of the high natural uptake of iodine by the thyroid.

Radionuclides that are used in therapy are made in three different manners (see Sec. 2). The most common is from neutron capture, since neutron-rich radionuclides decay by β^- emission. The high fluxes of neutrons associated with nuclear reactors make it possible to produce large quantities of neutron-rich radionuclides by this method. However, the specific activity of reactor-produced radionuclides is generally low. An alternative source that is available indirectly from a reactor is from generators. Accelerator-produced radionuclides, as mentioned above, can, in principle, produce radionuclides with very high specific activity.

The interest in therapy with radionuclides has increased during the last decade, not only with the prospect of labeling site-specific compounds with these radiotoxic substances but also as palliative agents for treating metastatic tumors. Post-mortem studies indicate that a large majority of carcinomas develop bone metastases. While the exact mechanisms for causing pain in skeletal metastases are unknown, their management has been directed toward palliation. The use of targeted radionuclide therapy for this treatment is gaining favor. High specific uptake in bone lesions with rapid clearance from other tissue is an important consideration here, as in all tumor treatments using radiotoxic nuclides. Of the bone agents in use, ^{32}P orthophosphate and ionic ^{89}Sr have been used most extensively. More recently, ^{153}Sm-EDTMP (ethylenediaminetetramethylenephosphonic acid) and ^{186}Re-HEDP (1-hydroxymethylidene diphosphonate) show promise to be even more effective at rapid symptom relief.

The use of antibodies for direct treatment of tumors without damaging normal tissues has been of interest for at least a century. However, it was not until recently that the ability to isolate and label antibodies has made this approach possible on a large scale. Although a number of antibodies had been labeled with ^{131}I in the 1950s and 1960s, it was not until the mid 1970s, when the techniques for producing monoclonal antibodies were developed, that advances in radioimmunotherapy (RIT) became a reality.

There are five factors that determine the potential utility of radiolabeled antibodies for radioimmunotherapy:

1. the absolute quantity of radiopharmaceutical delivered to the tumor (target);
2. the amount of radioactivity in other organs (nontarget);
3. the radiosensitivity of the tumor;
4. the radiation dose delivery rate; and
5. the types and energies of radiations emitted.

There is still some dispute as to whether monoclonal antibodies are more efficacious than polyclonals for achieving an acceptable target to nontarget ratio. The isotopes ^{111}In, ^{90}Y, ^{125}I, and ^{186}Re have all been incorporated into antibodies and are the subject of clinical trials. While the results are promising, it is still too early to know in which direction the results will lead.

^{131}I is still widely accepted, even though its physical characteristics are less favorable than those of other radiotoxic nuclides. For example, its half-life of 8 d is, in general, longer than the biological half-life of most carriers. Also, its β energy ($E_{\max} = 0.6$ MeV) is considered less than optimum for large tumors, as is the large number of γ rays that accompany the β decay. What remains attractive about this isotope is the wide variety of compounds into which iodine can be chemically inserted.

What is still a big question mark in the use of targeted radionuclide therapy is a complete understanding of the dose delivered, not only to the tumor but to surrounding healthy tissue. Dosimetric treatments that encompass situations where the microdosimetry on the cellular level is combined with biologic data could lead to more realistic dose estimates. This is especially true for the Auger emitters. When the decay of an Auger electron emitter occurs in the vicinity of DNA, the dose to microscopic volumes within the DNA may prove to be even more meaningful than the dose to the whole cell. The time-dependent distribution of the radiotoxic substance is required to gain a fuller understanding of the relationship between dose and effect.

10. FUTURE ASPECTS OF NUCLEAR MEDICINE

Probably the biggest advances in nuclear medicine in the future will be in the area of therapy. As the ability to deliver radionuclides to specific sites becomes a reality through research in monoclonal antibodies and their fragments, the use of therapy with radionuclides will become increasingly possible. As such, the rhenium isotopes ^{188}Re, from the ^{188}W/Re generator system, and ^{186}Re will be attractive choices because of rhenium's similar chemistry to ^{99m}Tc. Also, most experts feel that the growth area for PET will be in oncology for both diagnosis and treatment assessment. The ability to acquire quantitative metabolic and receptor density images throughout the whole body will provide a needed tool in the fight against cancer.

As the present-day PET tomographs approach theoretical spatial resolution, the remaining technological advances will be made in volume imaging, where the sensitivity of the tomograph is maximized by the removal of interplane septa. Work in this area has already begun with extremely promising results, not only in the increased sensitivity but in spatial sampling.

There is increasing interest in monitoring the effect on the neuronal system of the heart as a function of ischemia and infarction. One candidate is *m*-iodobenzylguanidine (MIBG), a norepinephrine analog. MIBG could monitor the depletion of the catecholamine stores of the heart in disease. Imaging of the muscarinic acetylcholine receptors in the heart using radioiodinated 3-quinuclidinyl-4-iodobenzylate (QNB) has been proposed. A ^{11}C-labeled QNB has also been proposed for use with PET. The availability of SPECT imaging has been a driving force in the development of ^{123}I-labeled compounds that are analogs of receptor antagonist used in PET. A number of these are undergoing clinical trials.

As the chemistry for attaching technetium becomes even more sophisticated, there will be more attempts at incorporating ^{99m}Tc into agents that have receptor sensitivity, so that receptor imaging will no longer be associated solely with PET radiopharmaceuticals or their iodinated analogs. However, in the meantime, more radiopharmaceuticals labeled with ^{123}I will find their way into widespread clinical use as they are validated in the assessment of disease and the availability of ^{123}I at a reasonable cost becomes assured.

There will be continuing efforts toward improving SPECT cameras as well as the software for correcting for scatter and attenuation in SPECT. To take advantage of the new radiopharmaceuticals for brain and heart, improved collimators will be designed.

Since radiopharmaceuticals generally are used to monitor function, there will be increasing use of these techniques to monitor treatment, for example, in determining the effectiveness of organ transplants. As an example of PET volume imaging, Fig. 10 illustrates the functional integrity following fetal

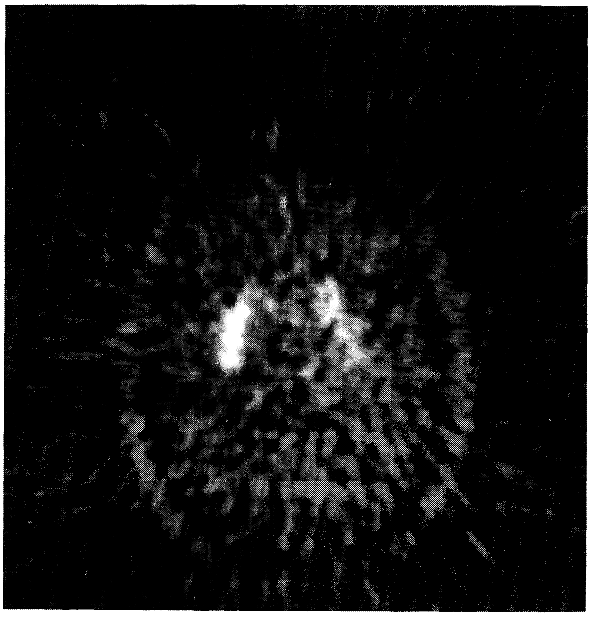

FIG. 10. PET image of ^{18}F-fluorodopa uptake in a patient who has received a fetal cell transplant. The arrows point to the four "hot spots" corresponding to the four surgical implants, indicating that the implanted cells are metabolically active with respect to the dopaminergic pathway. (Courtesy UBC/TRIUMF PET Program.)

cell transplantation. It can be expected that neuropsychiatric treatment will also be monitored by nuclear imaging processes as we gain a better understanding of the biochemistry of mental function.

As one of the pioneers of nuclear medicine, H. Wagner, recently said at the 1992 Annual Society of Nuclear Medicine meeting, nuclear medicine has begun the transformation of the art of diagnostic and therapeutic decision making into a quantitative science.

GLOSSARY

Alpha Decay: The mode of radioactive decay when the nucleus emits an α particle (the nucleus of ^{4}He), the product being two protons and two neutrons less.

Annihilation Radiation: The two 511-keV γ rays resulting from the annihilation of the β^+ with an electron in matter.

Antibody: A protein produced in the body in response to and as a counteragent to antigens.

Antigen: A substance that can induce the production of an antibody and bind with it specifically.

Auger Electron: The emission of an electron as an alternative to x-ray emission in an excited atom; thus, no x ray is emitted.

Becquerel: The SI unit of radioactivity, equivalent to one disintegration per second.

Beta Decay: The mode of radioactive decay when the nucleus emits an electron. In β^- decay, the atomic number increases by one, and, in β^+, decay the atomic number decreases by one.

Biological Half-Life: The time during which one-half of the amount of a substance is eliminated by a biological process such as urinary excretion.

Carrier Free: A term used in radiopharmaceutical sciences to indicate the absence of any stable atoms in a radionuclide sample. The term *no carrier added* is used to indicate the absence of purposely added nonradioactive atoms of the same element.

Curie: A measure of radioactivity equivalent to 3.7×10^{10} disintegrations per second.

ECD: The abbreviation for 1,1-ethylcysteinate dimer, which is used to monitor cerebral blood flow when labeled with ^{99m}Tc.

Electron Capture: The mode of radioactive decay when the nucleus captures an orbiting electron to convert a proton into a neutron; equivalent product to positron decay.

Electron Volt (eV): The kinetic energy required to move an electron through a potential difference of 1 V.

Half-Life ($t_{1/2}$): The time during which one-half the number of radioactive species decays. Same as physical half-life.

HIPDM: The abbreviation for *N*,*N*,*N*′-trimethyl-*N*-[2-hydroxy-3-methyl-5-iodobenzyl]-1,3-propanediamine, used as a blood-flow agent.

HM-PAO: The abbreviation for hexamethylpropylene amine oxime, which when labeled with ^{99m}Tc is used in cerebral blood-flow measurements.

IMP: The abbreviation for *N*-isopropyl-*p*-iodoamphetamine, which is used in assessing cerebral blood flow.

Infarct: An area of dead tissue resulting from a complete lack of blood circulation.

IPPA: The abbreviation for iodophenylpentadecanoic acid, a fatty acid used to monitor cardiac metabolism.

Ischemia: A condition where a tissue region has a deficiency in blood supply.

Kit: A package of the chemicals required to prepare a radiotracer.

Lumped Constant: The collection of mathematical terms used to express the relationship between the biological behavior of glucose and fluoro-deoxy-glucose.

MAG3: The abbreviation for triamidemonothiol-(N3M) complex ^{99m}Tc-mercaptoacetylglycylglycylglycine, which is used to monitor renal function.

MeV: Mega electron volts, a measure of energy equivalent to 1.6×10^{-19} joules.

RBC: The abbreviation for red blood cells.

Rectilinear Scanner: An imaging device whose detector is scanned mechanically in a rasterlike pattern over the area of interest.

Specific Activity: The measure of the amount of radioactivity per mass of sample.

Tritium: The only radioactive isotope of hydrogen, with 2 neutrons; it has a half-life of 12.3 years.

WBC: The abbreviation for white blood cells.

Works Cited

Eckelman, W. C. (1991), *Nucl. Med. Biol.* **18**, iii–vi.

Harbert, J., da Roche, A.F.G. (1984), *Textbook of*

Nuclear Medicine, 2nd ed., Vol. I, "Basic Science," Philadelphia: Lea & Febiger.

ICRP Publication 53 (1988), *Radiation Dose to Patients from Radiopharmaceuticals*, Oxford: Pergamon.

Loevinger, R., Budinger, T. F., Watson, E. E. (1988), *MIRD Primer for Absorbed Dose Calculations*, New York: Soc. Nuclear Medicine.

Reivich, M., Kuhl, D., Wolf, A., Greenberg, J., Phelps, M., Ido, T., Casella, V., Fowler, J., Hoffman, E., Alavi, A., Som, P., Sokoloff, L. (1979), *Circ. Res.* **44**, 127–137.

Saha, G. B. (1992), *Fundamentals of Nuclear Pharmacy*, 3rd ed., New York: Springer-Verlag.

Sorenson, J. A., Phelps, M. E. (1987), *Physics in Nuclear Medicine*, 2nd ed., New York: Grune & Stratton.

Swanson, D. P., Chilton, H. M., Thrall, J. H. (1990), *Pharmaceuticals in Medical Imaging*, New York: Macmillan.

Ter-Pogossian, M. M., Phelps, M. E., Hoffman, E. J., Mullani, N. A. (1975), *Radiology* **114**, 89–98.

Yalow, R. S. (1978), "Radioimmunoassay: A Probe for the Fine Structure of Biologic System," *Science* **200**, 1236–1245.

Further Reading

Burns, H. D., Gibson, R. E., Dannals, R. F., Siegl, P. K. S. (1993), *Nuclear Imaging in Drug Discovery, Development and Approval.* Boston: Birkhauser.

Diksic, M., Reba, R. C. (Eds.) (1990), *Radiopharmaceuticals and Brain Pathology Studied with PET and SPECT*, Boca Raton, FL: CRC Press.

Frost, J. J., Wagner, H. N. (Eds.) (1990), *Quantitative Imaging, Neuroreceptors, Neurotransmitters and Enzymes*, New York: Raven.

Harbert, J., da Roche, A.F.G. (1984), *Textbook of Nuclear Medicine*, 2nd ed., Vol. I, "Basic Science," Philadelphia: Lea & Febiger.

Loevinger, R., Budinger, T. F., Watson, E. E. (1988), *MIRD Primer for Absorbed Dose Calculations*, New York: Soc. Nuclear Medicine.

Sorenson, J. A., Phelps, M. E. (1987), *Physics in Nuclear Medicine*, 2nd ed., New York: Grune & Stratton.

Swanson, D. P., Chilton, H. M., Thrall, J. H. (1990), *Pharmaceuticals in Medical Imaging*, New York: Macmillan.

2
Nuclear Reactions

Carlos A. Bertulani

2.1 Introduction 47
2.2 Basic Principles 48
2.2.1 Conservation Laws 48
2.2.2 Kinematics 49
2.2.3 Cross Sections 50
2.2.4 Elastic Scatteringa 52
2.2.5 Reaction Cross Sections 53
2.2.6 Excitation Functions 53
2.3 Statistical Reactions 55
2.3.1 Compound Nucleus 55
2.3.2 Energy Spectrum of Neutrons 57
2.3.3 Resonances 59
2.3.4 Statistical Theory of Nuclear Reactions 62
2.3.5 The Optical Model 63
2.4 Direct Reactions 65
2.5 Heavy Ion Reactions 68
2.5.1 Types and Properties 68
2.5.2 Superheavy Elements 71
2.6 Electromagnetic Probes 72
2.6.1 Coulomb Excitation 72
2.6.2 Photonuclear Reactions and Giant Resonances 73
2.6.3 Electron Scattering 75
2.7 Relativistic Nuclear Collisions 76
2.7.1 Transport Theories and Equation of State 76
2.7.2 Kinematics 77
2.7.3 The Quark–Gluon Plasma (QGP) 79
2.7.4 Ultraperipheral Relativistic Heavy Ion Collisions 81
2.8 Nuclear Reactions in Stars 82

Encyclopedia of Nuclear Physics and its Applications, First Edition. Edited by Reinhard Stock.
© 2013 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2013 by Wiley-VCH Verlag GmbH & Co. KGaA.

2.8.1 Hydrogen and CNO Cycles 82
2.8.2 Synthesis of Heavier Elements 84
2.8.3 Thermonuclear Cross Sections 85
2.9 Reactions with Radioactive Nuclear Beams 86
Acknowledgments 90
References 90
Further Readings 91

2.1
Introduction

With the collision of two nuclei, a nuclear reaction can take place and, as in the case of a chemical reaction, the final products can be different from the initial ones. This process occurs when a target is bombarded by particles coming from an accelerator or a radioactive substance. It was in the latter way that Rutherford observed, in 1919, the first nuclear reaction produced in the laboratory,

$$\alpha + {}^{14}_{7}\mathrm{N} \rightarrow {}^{17}_{8}\mathrm{O} + \mathrm{p} \tag{2.1}$$

using α-particles from a ^{214}Bi sample.

As in Eq. (2.1), other reactions were induced using α-particles, the only projectile available initially. With the development of accelerators around 1930, the possibilities multiplied by changing the energy and mass of the projectile. At present, it is possible to bombard a target with protons of energy greater than 1 TeV ($1\ \mathrm{TeV} = 10^{12}\ \mathrm{eV} = 1.602 \times 10^{-7}$ J), and beams of particles as heavy as uranium are available for study of reactions with heavy ions.

Sometimes we can have more than two final products in a reaction, as in

$$\begin{aligned} \mathrm{p} + {}^{14}\mathrm{N} &\rightarrow {}^{7}\mathrm{Be} + 2\alpha, \\ \mathrm{p} + {}^{23}\mathrm{Na} &\rightarrow {}^{22}\mathrm{Ne} + \mathrm{p} + \mathrm{n} \end{aligned} \tag{2.2}$$

or just one, as in the capture reaction $\mathrm{p} + {}^{27}\mathrm{Al} \rightarrow {}^{28}\mathrm{Si}^*$, where the asterisk indicates an excited state that usually decays, emitting γ-radiation. Under special circumstances, more than two reactants are possible. Thus, for example, the reaction $\alpha + \alpha + \alpha \rightarrow {}^{12}\mathrm{C}$ can take place in the overheated plasma of stellar interiors.

The initial and final products can also be identical. This case characterizes a process that can be elastic, as in $\mathrm{p} + {}^{16}\mathrm{O} \rightarrow \mathrm{p} + {}^{16}\mathrm{O}$ where there is only transfer of kinetic energy between the projectile and the target, or inelastic, as in $\mathrm{n} + {}^{16}\mathrm{O} \rightarrow \mathrm{n} + {}^{16}\mathrm{O}^*$, where part of the kinetic energy of the system is used in the excitation of ^{16}O.

Naturally, nuclear reactions are not limited to nuclei. They could involve any type of particle and also radiation. Thus, the reactions

$$\begin{aligned} \gamma + {}^{63}\mathrm{Cu} &\rightarrow {}^{62}\mathrm{Ni} + \mathrm{p}, \\ \gamma + {}^{233}\mathrm{U} &\rightarrow {}^{90}\mathrm{Rb} + {}^{141}\mathrm{Cs} + 2\mathrm{n} \end{aligned} \tag{2.3}$$

are examples of nuclear processes induced by γ-radiation. In the first case a γ-ray knocks a proton off ^{63}Cu and in the second it induces nuclear fission of ^{233}U, with the production of two fragments and emission of two neutrons.

Unlike a chemical reaction, the resulting products of a nuclear reaction are not

Encyclopedia of Nuclear Physics and its Applications, First Edition. Edited by Reinhard Stock.
© 2013 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2013 by Wiley-VCH Verlag GmbH & Co. KGaA.

determined univocally: starting from two or more reactants there can exist dozens of final products with an unlimited number of available quantum states. As an example, the collision of a deuteron with ^{238}U can result in, among others, the following reactions:

$$\begin{aligned} d + {}^{238}\mathrm{U} &\rightarrow {}^{240}\mathrm{Np} + \gamma, \\ &\rightarrow {}^{239}\mathrm{Np} + \mathrm{n}, \\ &\rightarrow {}^{239}\mathrm{U} + \mathrm{p}, \\ &\rightarrow {}^{237}\mathrm{U} + \mathrm{t} \end{aligned} \tag{2.4}$$

In the first one, the deuteron is absorbed by uranium, forming an excited nucleus of ^{240}Np that de-excites by emitting a γ-ray. The second and third are examples of *stripping reactions*, in which a nucleon is transferred from the projectile to the target. The last one exemplifies the inverse process: the deuteron captures a neutron from the target and emerges as ^{3}H (tritium). This is denoted as a *pickup reaction*. Another possibility would be, in the first reaction, that ^{240}Np fissions instead of emitting a γ-ray, contributing with dozens of possible final products for the reaction.

Each reaction branch, with well-defined quantum states of the participants, is known as a *channel*. In Eq. (2.4), for the *entrance channel* d + ^{238}U, there are four possible *exit channels*. Note that a different exit channel would be reached if some of the final products were in the excited state. The probability that a nuclear reaction takes place through a certain exit channel depends on the energy of the incident particle and is measured by the *cross section* of that channel.

Nuclear reactions proceed through many possible distinct mechanisms. For instance, in *direct reactions*, the projectile and the target have an interaction of short duration, with possible exchange of energy or particles between them. Another mechanism involves the *fusion* of the projectile with the target, the available energy being distributed among all nucleons, forming a highly excited *compound nucleus*. The decay of the compound nucleus leads to the final products of the reaction.

In high-energy collisions, the nuclei fragment and particles that were not initially present are produced (for instance, pions, kaons, etc.). The reactions proceed through an intermediate phase in which the nuclear matter is compressed. At very high energies, the quarks and gluons inside nucleons become ‘‘free’’ for a short time, forming the *quark–gluon plasma*. The study of high-energy reactions with nuclei is very important for a better understanding of what happens during spectacular stellar explosions (supernovae) and in the interior of compact stars, as for instance, neutron stars. The study of nuclear reactions at high energies allows to obtain information on the *equation of state* of nuclear matter.

2.2
Basic Principles

2.2.1
Conservation Laws

Several conservation laws contribute to restrict the possible processes in a nuclear reaction.

1. *Baryonic number*. There is no experimental evidence of processes in which nucleons are created or destroyed without the creation or destruction of corresponding antinucleons. The application of this principle to low-energy reactions is still more restrictive. Below the threshold

for the production of mesons (∼140 MeV), no process related to the nuclear forces is capable of transforming a proton into a neutron and vice versa, and processes governed by the weak force (responsible for the β-emission of nuclei) are very slow in relation to the times involved in nuclear reactions ($\sim 10^{-22}$–10^{-16} s). In this way, we can speak separately of proton and neutron conservation; this should show up with same amounts in both sides of a nuclear reaction.

2. *Charge*. This is a general conservation principle in physics, valid in any circumstance. In purely nuclear reactions, it is computed making the sum of the atomic numbers, which should be identical, on both sides of the reaction.
3. *Energy and linear momentum*. These are two of the most applied principles in the study of the kinematics of reactions. Using them, angles and velocities are related to the initial parameters of the problem.
4. *Total angular momentum*. It is always a constant of motion. In the reaction

$$^{10}\mathrm{B} + {}^{4}\mathrm{He} \rightarrow {}^{1}\mathrm{H} + {}^{13}\mathrm{C} \qquad (2.5)$$

^{10}B has $I = 3$ in the ground state, whereas the α-particle has zero angular momentum. If it is captured in an s-wave ($l_i = 0$), the intermediate compound nucleus is in a state with $I_c = 3$. Both final products have intrinsic angular momenta equal to 1/2. Hence, their sum is 0 or 1. Therefore, the relative angular momentum of the final products will be $l_f = 2$, 3, or 4.

5. *Parity*. It is always conserved in reactions governed by nuclear interaction. In the previous example, ^{10}B, ^{4}He, and the proton have equal parities, while ^{13}C has odd parity. Therefore, if $l_i = 0$, we necessarily have $l_f = 3$. Thus, the orbital momentum of the final products of Eq. (2.5) is determined by the combined conservation of the total angular momentum and of the parity.
6. *Isospin*. This is an approximate conservation law that is applied to light nuclei, where the effect of the Coulomb force is small. A nuclear reaction involving these nuclei not only conserves the z-component of the isospin (a consequence of charge and baryonic number conservation) but also conserves the total isospin **T**. Reactions that populate excited states not conserving the value of **T** are strongly inhibited. An example is the reaction $\mathrm{d} + {}^{16}\mathrm{O} \rightarrow \alpha + {}^{14}\mathrm{N}$, where the excited state 0^+, with 2.31 MeV, of ^{14}N is about a 100 times more populated than the 1^+ ground state. Conservation of energy, angular momentum, and parity does not impose any prohibition for that channel, whose low occurrence can only be justified by isospin conservation: the ground states of the four participant nuclei in the reaction have all $\mathbf{T} = 0$ and the state 1^+ of ^{14}N has $\mathbf{T} = 1$.

2.2.2
Kinematics

We consider a typical reaction, in which the projectile a and the target A result in two products, b and B. This can also be expressed in the notation that we used so far, $\mathrm{a} + \mathrm{A} \rightarrow \mathrm{b} + \mathrm{B}$, or even in a more compact notation, A(a,b)B.

Often, a and b are light nuclei and A and B the heavy ones; the nucleus b is emitted

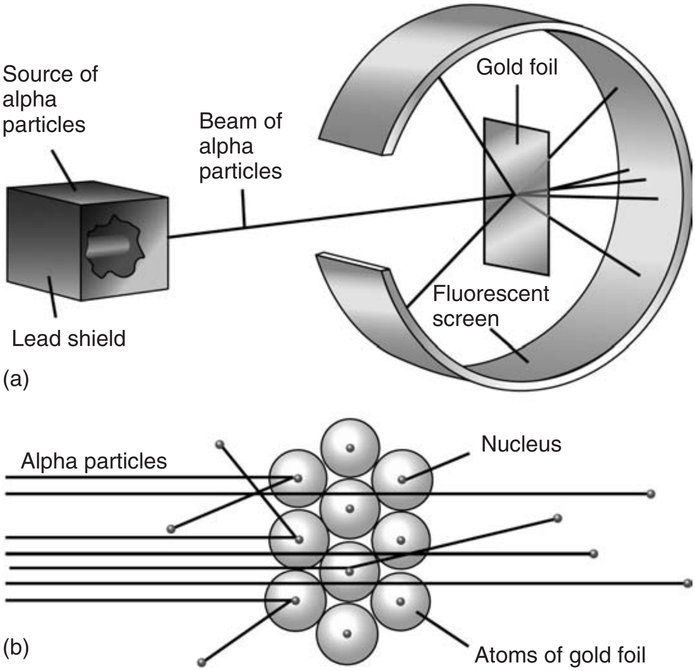
Source of alpha particles
Beam of alpha particles
Gold foil
Lead shield
Fluorescent screen
(a)
Alpha particles
Nucleus
(b)
Atoms of gold foil

Figure 2.1 (a) Rutherford's scattering experiment with α-particles scattering on a gold foil. (b) Microscopic interpretation of the experimental results.

at an angle θ and its energy registered in the laboratory system. The recoiling nucleus B has usually a short range and cannot leave the target. It is convenient to introduce the *Q-value of the reaction*, which measures the energy gained (or lost) because of the difference between the initial and final masses:

$$Q = (m_a + m_A - m_b - m_B)c^2 \quad (2.6)$$

Using energy and momentum conservation in the reaction, one gets

$$Q = E_b\left(1 + \frac{m_b}{m_B}\right) - E_a\left(1 - \frac{m_a}{m_B}\right) - \frac{2}{m_B}\sqrt{m_a m_b E_a E_b}\cos\theta \quad (2.7)$$

From this relation, one concludes that, when Q is negative, the incident particle has an energy threshold, E_t, which is a function of the angle θ, below which nuclei b are not observed:

$$E_t = \frac{-Q m_B (m_B + m_b)}{m_a m_b \cos^2\theta + (m_B + m_b)(m_B - m_a)} \quad (2.8)$$

This example shows the power of conservation laws in the analysis of nuclear reactions. Similar to this example, very useful relations can be derived in the relativistic regime, using other conserved quantities.

2.2.3
Cross Sections

Figure 2.1 shows schematically a typical scattering experiment. In fact, this is the sketch of Rutherford's experiment

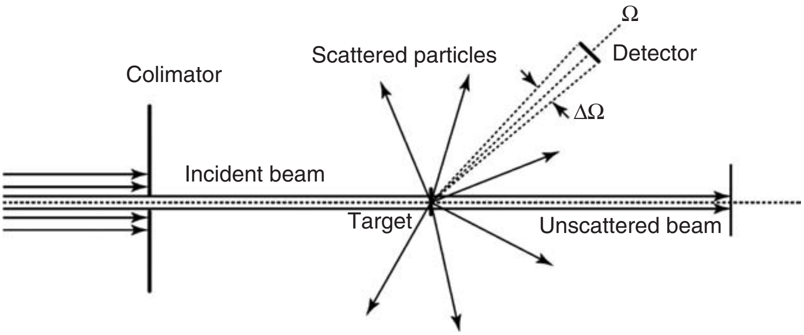

Figure 2.2 Schematic representation of a scattering experiment.

in 1910 [1]. Projectiles (here, α-particles) from a source pass through a collimator and collide with a target (gold foil). Some projectiles are scattered by the target and reach the detector (here, a fluorescent screen). Rutherford expected them to go straight through the target, with perhaps minor deflections. Most did go straight through, but, to his surprise, some particles bounced straight back! Rutherford hypothesized that the positive α- particles had hit a heavy mass of positive particles, which he called *the nucleus*.

Let us consider an experiment such as the one represented in Figure 2.2 that measures the count rate of events leading to the population of channel-α, $N_\alpha(\Omega,\ \Delta\Omega)$. Assuming that interactions between beam particles can be neglected, the count rate $N_\alpha(\Omega,\ \Delta\Omega)$ should be proportional to the incident flux J, to the fraction of solid angle $\Delta\Omega$, where the particles are scattered to, and to the number of target particles per unit volume. Hence, $N_\alpha(\Omega,\ \Delta\Omega)$ can then be written as $N_\alpha(\Omega,\ \Delta\Omega) = (\Delta\Omega \cdot n \cdot J)\ \mathrm{d}\sigma_\alpha(\Omega)/\mathrm{d}\Omega$. The constant of proportionality,

$$\frac{\mathrm{d}\sigma_\alpha(\Omega)}{\mathrm{d}\Omega} = \frac{N_\alpha(\Omega, \Delta\Omega)}{\Delta\Omega\ \cdot n\ \cdot J} \tag{2.9}$$

is called the *differential cross section* for channel-α. This quantity is very useful, as it does not depend on experimental details (detector size, incident flux, and target thickness). It depends exclusively on the physics of the projectile and the target particles.

One is often interested in the angle-integrated *total cross section* given by

$$\sigma_\alpha = \int \mathrm{d}\Omega \left[\frac{\mathrm{d}\sigma_\alpha(\Omega)}{\mathrm{d}\Omega}\right] \tag{2.10}$$

In nuclear physics, cross sections are usually measured in units of *barns*. One barn corresponds to an area of a circle of diameter approximately equal to 8 fm = 8×10^{-15} m. For comparison between theory and experiment, it is necessary to have cross sections in the same reference frame. Of course, the measured cross section is obtained in the laboratory frame, where the target is at rest. From the theoretical point of view, however, it is important to take advantage of the translational invariance of the projectile–target Hamiltonian, introducing the center of mass (CM) frame. The transformation is done using the laws of conservation of energy and momentum.

The Rutherford experiment has a simple classical description in which one assumes that the α-particles follow hyperbolic trajectories in the Coulomb field of the heavy target. Scattering at an angle θ can be easily calculated and with that one is able to calculate the Coulomb, or Rutherford cross section. The result is

$$\frac{d\sigma_R}{d\Omega} = \left(\frac{Z_P Z_T e^2}{4E}\right)^2 \operatorname{cosec}^4\left(\frac{\theta}{2}\right) \tag{2.11}$$

where Z_P (Z_T) is the projectile (target) charge number and E is the CM energy.

2.2.4
Elastic Scatteringa

When a beam of particles – represented in quantum mechanics by a plane wave – hits a nucleus, the wavefunction is modified by the presence of a scattering potential $V(r)$, responsible for the appearance of a phase in the outgoing part of the wave. Elastic scattering is just one of the channels along which the reaction can proceed and is known as *elastic channel*. Inelastic scattering and all the other channels are grouped in the *reaction channel*.

The occurrence of a nuclear reaction through a given reaction channel leads to a modification of the outgoing part of the wavefunction not only by a phase factor but also by a change in its magnitude, indicating that there is a loss of particles in the elastic channel. For a projectile with momentum $p = \hbar k$, this can be expressed by

$$\Psi \sim \frac{1}{2i}\sum_{l=0}^{\infty}(2l+1)i^l P_l(\cos\theta) \times \frac{\eta_l e^{i\left(\frac{kr-l\pi}{2}\right)} - e^{-i\left(\frac{kr-l\pi}{2}\right)}}{kr} \tag{2.12}$$

where the complex coefficient η_l is the factor mentioned above. If $\eta = 1$, the sum in Eq. (2.12) can be done analytically, leading to $\Psi \sim \exp(i\mathbf{k}.r)$, that is, a plane wave. But if $\eta_l = \exp[i\delta l]$, with δ_l real, the incoming and outgoing waves have the same magnitude, that is, the scattering is elastic.

The sum in Eq. (2.12) is known as the *partial wave expansion* of the scattering wave. The label $l = 0, 1, 2, \ldots =$ (s, p, d, … waves) denotes the contribution of a particular angular momentum (in units of $\hbar$) to the total wavefunction. Classically, the angular momentum of an incident particle is given by $l = kR$, where R is known as the *impact parameter*, which is the perpendicular distance to the target if the projectile were undeflected. In quantum mechanics l is not continuous, varying in steps of one.

We can rewrite Eq. (2.12) as a sum of a plane wave and a scattering outgoing wave in the form $\Psi \sim \exp(i\mathbf{k}.r) + f(\theta)e^{ikr}/r$, where $f(\theta)$ accounts for the distortion of the outgoing wave at the scattering angle θ. It is known as the *scattering amplitude*:

$$f(\theta) = \frac{1}{2ik}\sum_{l=0}^{\infty}(2l+1)(\eta_l - 1)P_l(\cos\theta) \tag{2.13}$$

One can now calculate the cross section by counting the number of particles that are scattered through an angle θ. This can be done by calculating the particle current associated with the wavefunction Ψ. One obtains the differential scattering cross section

$$\frac{d\sigma_e}{d\Omega} = |f(\theta)|^2 = \frac{1}{4k^2}\left|\sum_{l=0}^{\infty}(2l+1) \times (1-\eta_l)P_l(\cos\theta)\right|^2 \tag{2.14}$$

The total scattering cross section, Eq. (2.10), becomes

$$\sigma_e = \pi\lambda\!\!\!^{-}{}^2 \sum_{l=0}^{\infty}(2l+1)|1-\eta_l|^2 \tag{2.15}$$

with $\lambda\!\!\!^{-} = \lambda/2\pi = 1/k$.

2.2.5
Reaction Cross Sections

To calculate the reaction cross section, it is necessary to compute the number of particles that disappear from the elastic channel, which is measured by the flux of the probability vector current through a spherical surface of large radius centered at the target, calculated with the total wavefunction of Eq. (2.12). One finds

$$\sigma_r = \pi \lambda\!\!\!^{-2} \sum_{l=0}^{\infty}(2l+1)(1-|\eta_l|^2) \tag{2.16}$$

From Eqs. (2.15) and (2.16), when $|\eta_l| = 1$, the reaction cross section is zero and we have pure scattering. The contrary, however, cannot happen, as the vanishing of σ_e also implies the vanishing of σ_r. In general, there is a region of allowed values of η_l for which the two cross sections can coexist.

The maximum of σ_r happens for $\eta_l = 0$, which corresponds to total absorption. Let us suppose that the absorption potential is limited to the surface of a nucleus with radius $R \gg \lambda\!\!\!^{-}$, that is, that all the particles with impact parameter smaller than the radius R are absorbed. This is equivalent to saying that all particles are absorbed for $l \le R/\lambda\!\!\!^{-}$. In this case,

$$\sigma_r = \pi \lambda\!\!\!^{-2} \sum_{l=0}^{R/\lambda\!\!\!^{-}}(2l+1) = \pi(R+\lambda\!\!\!^{-})^2 \tag{2.17}$$

This is the value that would be intuitively adequate for the total cross section, that is, equal to the geometric cross section (the part $\lambda\!\!\!^{-}$ can be understood as an uncertainty in the position of the incident particle). However, we saw above that the presence of scattering is always obligatory. For $\eta_l = 0$, the scattering and reaction cross sections are identical, yielding the total cross section

$$\begin{aligned}\sigma &= \sigma_r + \sigma_e \\ &= \pi(R+\lambda\!\!\!^{-})^2 + \pi(R+\lambda\!\!\!^{-})^2 \\ &= 2\pi(R+\lambda\!\!\!^{-})^2\end{aligned} \tag{2.18}$$

that is twice the geometric cross section!

The presence of the scattering part, which gives an apparently unexpected result in Eq. (2.18), can be interpreted as the effect of diffraction of the plane waves at the nuclear surface. This effect leads to a ''shadow'' behind the nucleus, decreasing its apparent diameter so that, at a certain distance, the perturbation caused by the presence of the nucleus disappears and the plane wave is reconstructed. In this situation, we can say that the part of the beam that is diffracted has to be the same as the part that is absorbed, justifying the equality of σ_r and σ_e. The diffraction phenomenon appears clearly in the elastic scattering or inelastic angular distribution (differential cross section as a function of the scattering angle): Figure 2.3 exhibits angular distributions for the elastic scattering of 30 MeV protons on ^{40}Ca,^{120}Sn, and ^{208}Pb. The oscillations in the cross sections are characteristic of a *Fraunhofer diffraction*, similar to light scattering by an opaque disk. The angular distance $\Delta\theta$ between the diffraction minima closely follows the expression $\Delta\theta = \hbar/pR$, characteristic of diffraction phenomena.

2.2.6
Excitation Functions

When the projectile has a very low energy, $k \to 0$, and in particular, $l = kR \ll 1$. As an example consider the scattering of neutrons with $l = 0$ and ignore the spins of the neutron and of the target. In this case, the *Schrödinger equation* for the radial

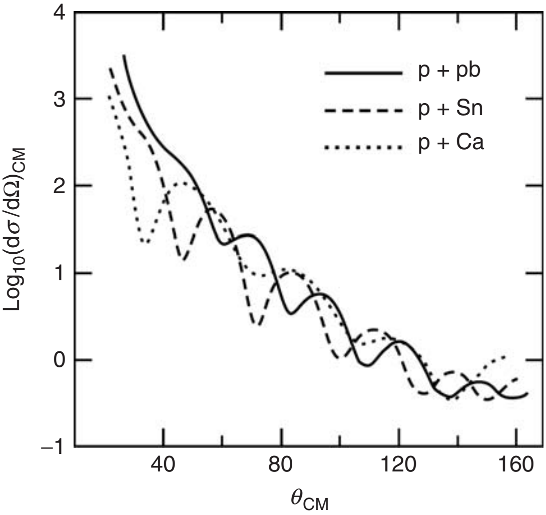

Figure 2.3 Elastic scattering angular distribution of protons of 30 MeV on Ca, Sn, and Pb. The curves are adjusted to the experimental data of Ref. 2

motion of the neutron with respect to the target is

$$\frac{d^2u_0}{dr^2} + k^2u_0 = 0 \quad (r \geq R) \qquad (2.19)$$

This is valid for the radial wavefunction u_0 at distances r larger than the *channel radius* $R = R_a + R_A$, with R_a and R_A being the radii of the projectile and the target, respectively. The solution of Eq. (2.19) is

$$u_0 = \eta_0 e^{ikr} - e^{-ikr} \quad (r \geq R) \qquad (2.20)$$

A radial wavefunction inside the nucleus should connect to the external function (2.20) with a continuous function and its derivative at $r = R$. That is, the function

$$f_l \equiv R\left[\frac{du_l/dr}{u_l}\right]_{r=R} \qquad (2.21)$$

must have identical values if calculated with the internal or the external function and this condition creates a relationship between f_l and η_l. Hence, the knowledge of f_l leads to the knowledge of the cross sections. For neutrons with $l = 0$, the application of Eq. (2.20) results in

$$f_0 = ikR\,\frac{\eta_0 + e^{-2ikR}}{\eta_0 - e^{-2ikR}} \qquad (2.22)$$

from which we extract

$$\eta_0 = \frac{f_0 + ikR}{f_0 - ikR}e^{-2ikR} \qquad (2.23)$$

If f_0 is a real number, then $|\eta_0|^2 = 1$. The reaction cross section, Eq. (2.16), will be zero and we have pure scattering.

Using Eq. (2.23), the scattering cross section, Eq. (2.15), can be written as

$$\sigma_{e,0} = \pi\lambda\!\!\!^{-2}|A_{\text{res}} + A_{\text{pot}}|^2 \qquad (2.24)$$

with $A_{\text{res}} = -2ikR/(f_0 - ikR)$ and $A_{\text{pot}} = \exp\,(2ikR) - 1$. The separation of the cross section in two parts has physical justification: A_{pot} does not contain f_0 and therefore does not depend on conditions inside the nucleus. It represents the situation where the projectile does not penetrate the nucleus as in the idealized situation where the nucleus is considered an impenetrable hard sphere. The wavefunction is zero inside the nucleus and u_0 vanishes at $r = R$, implying $f_0 \to \infty$ and $A_{\text{res}} \to 0$. Hence, A_{pot} is only responsible for scattering.

Substituting Eq. (2.23) in Eq. (2.16) and using

$$f_0 = f_R + \mathrm{i}f_I \tag{2.25}$$

we have

$$\sigma_{r,0} = \pi \bar{\lambda}^2 \frac{-4kRf_I}{f_R^2 + (f_I - kR)^2} \tag{2.26}$$

an equation that is useful when we study the presence of resonances in the *excitation function* (cross section as a function of the energy).

2.3
Statistical Reactions

2.3.1
Compound Nucleus

When a low-energy neutron (<50 MeV) enters the range of nuclear forces, it can be scattered or begin a series of collisions with the nucleons. The products of these collisions, including the incident particle, will continue in their course, leading to new collisions and new changes of energy. During this process, one or more particles can be emitted and they form, with the residual nucleus, the products of what is known as a *pre-equilibrium* reaction. At low energies, the largest probability is the continuation of the process so that the initial energy is distributed among all nucleons, with no emitted particle. The final nucleus with $A + 1$ nucleons has an excitation energy equal to the kinetic energy of the incident neutron plus the binding energy the neutron has in the new, highly unstable, nucleus [3]. It can, among other processes, emit a neutron with the same or smaller energy as that of the one absorbed. The de-excitation process is not necessarily immediate and the excited nucleus can live for a relatively long time. We say that there is, in this situation, the formation of a *compound nucleus* as intermediary stage of the reaction. In the final stage, the compound nucleus can evaporate one or more particles, fission, and so on. In our notation, for the most common situation in which two final products are formed (the evaporated particle plus the residual nucleus or two fission fragments, etc.) we write:

$$\mathrm{a} + \mathrm{A} \to \mathrm{C}^* \to \mathrm{B} + \mathrm{b}$$

the asterisk indicating that the compound nucleus C is in an excited state.

The compound nucleus lives long enough to "forget" how it was formed and the de-excitation to the final products b and B only depends on the energy, angular momentum, and parity of the quantum state of the compound nucleus. An interesting experimental verification was accomplished by Ghoshal in 1950 [4]. He studied two reactions that take to the same compound nucleus, ^{64}Zn*, and measured the cross sections of three different forms of decay, as shown below:

$$\begin{array}{ccccc} \mathrm{p} + {}^{63}\mathrm{Cu} \searrow & & \nearrow & {}^{63}\mathrm{Zn} + \mathrm{n} \\ & {}^{64}\mathrm{Zn}^* \to & & {}^{62}\mathrm{Cu} + \mathrm{n} + \mathrm{p} \\ \mathrm{a} + {}^{60}\mathrm{Ni} \nearrow & & \searrow & {}^{62}\mathrm{Zn} + 2\mathrm{n} \end{array} \tag{2.27}$$

If the idea of the compound nucleus is valid and if one chooses the energy of the proton and of the incident α-particle to produce the same excitation energy, then the cross section for each one of the three exit channels should be independent of

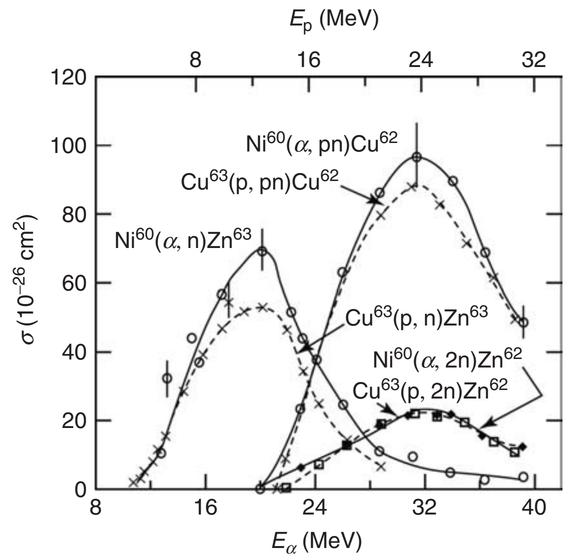

Figure 2.4 Cross sections for the reactions shown in Eq. (2.24). The scales of the upper axis (energy of the protons) and lower axis (energy of the α-particle) were adjusted to correspond to the same excitation energy of the compound nucleus [4].

the way the compound nucleus is formed. That is, the properties of the compound nucleus do not have any relationship with the nuclei that formed it. This is confirmed in Figure 2.4, where one sees clearly that the cross sections depend practically only on the exit channels.

The angular distribution of fragments, or evaporated particles, of a compound nucleus should be isotropic in the CM, and this is verified experimentally. However, the total angular momentum is conserved and cannot be ‘‘forgotten.’’ Reactions with large transfer of angular momentum, as when heavy ions are used as projectiles, can show a nonisotropic angular distribution in the CM system.

The occurrence of a nuclear reaction in two stages allows the cross section for a reaction A(a,b)B to be written as the product, $\sigma(\mathrm{a,b}) = \sigma_c(\mathrm{a,A})P(\mathrm{b})$, where $\sigma_c(\mathrm{a,A})$ is the cross section of formation of the compound nucleus starting from the projectile a and the target A and $P(\mathrm{b})$ is the probability that the compound nucleus emits a particle b leaving a residual nucleus B. If the quantum numbers of entrance and exit channels are well specified, that is, if the reaction begins at an entrance channel α and ends at an exit channel β, one can write

$$\sigma(\alpha, \beta) = \sigma_c(\alpha)P(\beta) \tag{2.28}$$

We can associate the probability $P(\beta)$ to the width Γ_β of the channel β and write

$$P(\beta) = \frac{\Gamma(\beta)}{\Gamma} \tag{2.29}$$

where Γ is the total width, that is, $\tau = \hbar/\Gamma$ is the half-life of disintegration of the compound nucleus. Equation (2.29) just expresses the fact that the decay probability through channel β is the decay rate through that channel divided by the total decay rate. In the competition between the several channels β, the nucleons have clear preference over the γ-radiation whenever there is available energy for their emission and, among the nucleons, the neutrons have preference as they do not have the

Coulomb barrier as an obstacle. Thus, in a reaction where there is no restriction for neutron emission, we can say that

$$\Gamma \cong \Gamma_n \tag{2.30}$$

where Γ_n includes the width for the emission of one or more neutrons.

The study of the function $P(\beta)$ is done in an evaporation model that leads to results in many aspects similar to the evaporation of molecules of a liquid, with the energy of the emitted neutrons having the form of a Maxwell–Boltzmann distribution (MB)

$$I(E) \propto E \exp\left(-\frac{E}{\theta}\right) \mathrm{d}E \tag{2.31}$$

with I measuring the amount of neutrons emitted with energy between E and $E + \mathrm{d}E$. The quantity θ, with dimension of energy, has the role of a *nuclear temperature*. It is related to the density of levels ω of the daughter nucleus B by

$$\frac{1}{\theta} = \frac{\mathrm{d}S}{\mathrm{d}E} \tag{2.32}$$

with

$$S = \ln \ \omega(E) \tag{2.33}$$

where $\mathrm{d}S/\mathrm{d}E$ is calculated for the daughter nucleus B at the maximum excitation energy that it can have after the emission of a neutron, that is, in the limit of emission of a neutron with zero kinetic energy.

The level density $\omega(E)$ is a measure of the number of available energy states for the decay of the compound nucleus in the interval $\mathrm{d}E$ around the energy E. In this sense, the relationship (2.33) is, neglecting the absence of the Boltzmann constant, identical to the thermodynamic relationship between the entropy S and the number of states available for the transformation of a system. Equation (2.32) is the well-known relation between the entropy and the temperature.

A simple expression for the energy dependence of the state density is provided by the equidistant spacing model, which assumes that the one-particle states are equally spaced with spacing d and that the total energy of the nucleus is simply obtained by adding the energies of the constituent nucleons. The solution of this problem can be obtained from statistical mechanics [6]:

$$\rho(E) \sim \exp\ (2\sqrt{aE})$$

Extensive analyzes of experimental data show that for nuclei far from the region of the magic nuclei, a varies linearly with A (or with N and Z), as shown in Figure 2.5:

$$a \cong \frac{A}{k}\ \mathrm{MeV}^{-1} \tag{2.34}$$

It is found that $k \cong 7.5 - 8$.

2.3.2
Energy Spectrum of Neutrons

The energy distribution of neutrons emitted by a compound nucleus has the aspect of the curve shown in Figure 2.6. Only the low-energy part obeys Eq. (2.31) and the reason is simple: the emission of a low-energy neutron leaves the residual nucleus with a large excitation energy, and the level density is very high. The large density of final states makes the problem tractable with the statistical model that leads to Eq. (2.31); in the opposite situation, it is the low-energy states of the residual nucleus. These isolated states appear as peaks in the tail of the distribution. When the emission is of a proton or of another charged particle, the form of Figure 2.6 is distorted, the part of low energy of the spectrum

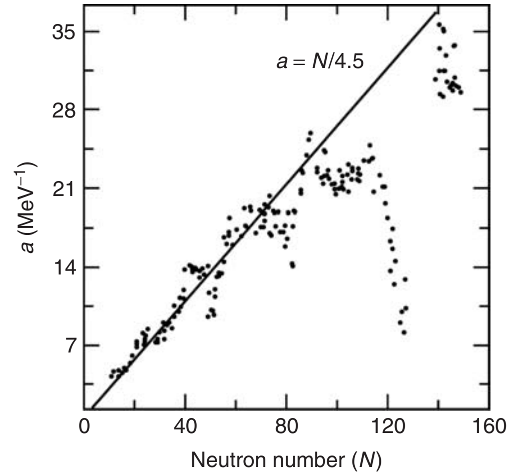

Figure 2.5 Values of the level density parameter a as a function of the neutron number. Far from magic regions, $a \approx N/4.5\ \text{MeV}^{-1}$, which approximately corresponds to $a = A/7.5\ \text{MeV}^{-1}$ [5].

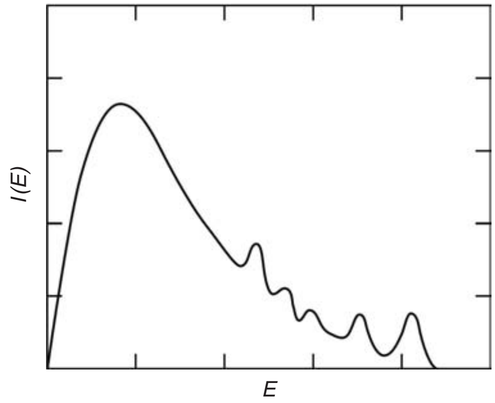

Figure 2.6 Energy spectrum of neutrons evaporated by a compound nucleus.

being suppressed partially by the Coulomb barrier.

Let us now assume that there is only elastic scattering or formation of a compound nucleus. We further assume that the elastic scattering is purely potential, without resonant elastic scattering. That is, there is no reemission of neutrons with energy equal to the incident projectile. This is equivalent to saying that the probability that the exit channel is the same as the entrance channel is very low. According to these assumptions, the cross section $\sigma_c(\alpha)$ of Eq. (2.28) is the reaction cross section of Eq. (2.16). We can still write the wavefunction inside the nucleus as just an incoming wave

$$u_0 \cong \exp(-\mathrm{i}Kr) \quad (r < R) \tag{2.35}$$

where $K = \sqrt{2m(E - V_0)}/\hbar$ is the wavenumber inside the nucleus, and it is assumed that the neutron with total energy E is subject to a negative potential V_0. Equation (2.35) is clearly a crude simplification when the incident neutron interacts in a complicated way with the other nucleons in the nucleus. It allows,

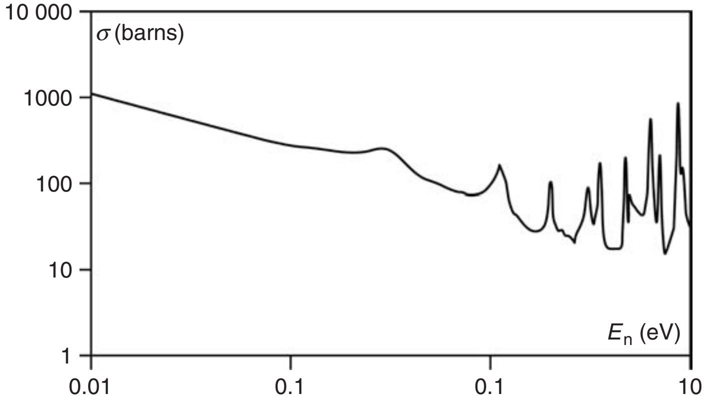

Figure 2.7 Total cross section for low-energy neutrons hitting ^{235}U.

however, to explain the average behavior of the cross sections for low energies. Starting with Eqs. (2.21) and (2.35), one determines the value of f_0:

$$f_0 = -\mathrm{i}KR \tag{2.36}$$

and from Eq. (2.26), we get

$$\sigma_c = \pi \lambdabar^2 \frac{4kK}{(k+K)^2} \tag{2.37}$$

for the cross section of compound nucleus formation for neutrons with $l = 0$. At low energies, $E|V_0|$, thus kK. Under these conditions, $\sigma_c = 4\pi/kK$, since $\lambdabar = k^{-1} \gg R$. Thus, σ_c varies with $1/k$. That is,

$$\sigma_c \propto \frac{1}{v} \tag{2.38}$$

where v is the velocity of the incident neutron. This is the well-known $1/v$ *law* that governs the behavior of the capture cross section of low-energy neutrons. Figure 2.7 exhibits the excitation function (cross section as function of the energy) for the reaction n+^{235}U. The cross section decays with $1/v$ up to 0.3 eV, where a series of resonances appear.

2.3.3
Resonances

To understand why there are resonances, we shall use again the simple model of a single particle subject to a square-well potential. We know that inside the well, the Schrödinger equation only admits solution for a discrete group of values of energy, $E_1, E_2, \ldots, E_n$. A particle is confined to the interior of the well by reflections that it has at the surface of the well. In these reflections, the wave that represents the particle should be in phase before and after the reflection and this only happens for a finite group of energies. Outside the well, the Schrödinger equation does not impose restrictions and the energy can have any value. But we know, from the study of the passage of a beam of particles through a potential step, that the discontinuity of the potential at the step provokes reflection even when the total energy of the particles is larger than the step, a situation where classically there would not be any difficulty for the passage of the particles. This reflection is partial and it becomes larger the closer the energy is to the height of the potential step. We can say that a particle with energy slightly positive is almost as confined as a particle inside the well. From this fact results the existence of almost bound states of positive energy known as *quasi-stationary states* or *resonances*. These resonances appear as peaks in the excitation function, a peak at a given energy meaning that the energy coincides with a given resonance of the nucleus.

The existence of resonances can also be inferred from the properties of the wavefunction. We consider only elastic scattering, with the other channels closed. The external and internal wavefunctions are both sine functions, the first with wavenumber $k = \sqrt{2mE}/\hbar$ and the second with $K = \sqrt{2m(E - V_0)}/\hbar$. If E is small and V_0 is about -35MeV, we have $K \gg k$. The internal and external parts should join at $r = R$ with continuous function and derivatives. As the internal frequency is much larger than the external one, the internal amplitude is quite reduced. There is a perfect matching between both and the internal amplitude is identical to the external one only at the proximity of the situation in which the derivative is zero. The energy for which this happens is exactly the energy of resonance.

Resonances appear in the excitation function at relatively low energies, where the number of open channels is not very large and it is possible to return to the entrance channel. To arrive at an expression of the cross section that describes a resonance, we rewrite Eq. (2.35) as,

$$u_l \sim \exp(-\mathrm{i}Kr) + b\exp(\mathrm{i}Kr)\ , \quad (r < R) \tag{2.39}$$

this time containing a second part, which takes into consideration the part of the wave that returns. This second part allows the existence of resonant scattering, where the incident particle is re-emitted with the same energy that it entered, after forming the compound nucleus. The complex amplitude b is always smaller than one, because there are no creation of particles in the region $< R$ in Eq. (2.26).

The second parenthesis in the denominator of Eq. (2.26) is never zero, since the numerator forces f_I to be always negative. If for f_R vanishes a certain energy, $\sigma_{r,0}$ passes by a maximum in that energy. We can tentatively identify these energies as being the energy of the resonances. Let us take the extreme case of a single resonance at the energy E_R, that is, $f_R = 0$ for $E = E_R$. We can expand f_R in a Taylor series in the neighborhood of a resonance, $f_R(E) = (E - E_R)(\mathrm{d}f_R/\mathrm{d}E)_{E=E_R} + \cdots$. Keeping just the first term of the expansion and using Eqs. (2.24) and (2.26), we get

$$\sigma_{e,0} = \pi \lambda\!\!\!^{-2} \left| \exp(2\mathrm{i}kR) - 1 + \frac{\mathrm{i}\Gamma_\alpha}{(E - E_R) + \mathrm{i}\frac{\Gamma}{2}} \right|^2 \tag{2.40}$$

$$\sigma_{r,0} = \pi \lambda\!\!\!^{-2} \frac{\Gamma_\alpha(\Gamma - \Gamma_\alpha)}{(E - E_R)^2 + \left(\frac{\Gamma}{2}\right)^2} \tag{2.41}$$

where we define

$$\Gamma_\alpha = -\frac{2kR}{(\mathrm{d}f_R/\mathrm{d}E)_{E=E_R}} \quad \text{and} \quad \Gamma = \frac{2kR - 2f_I}{(\mathrm{d}f_R/\mathrm{d}E)_{E=E_R}} \tag{2.42}$$

The energy Γ that appears in Eq. (2.40) is the total width of the resonance, $\Gamma = \Gamma_\alpha + \Gamma_\beta + \cdots$, that is, the sum of the widths for all the possible processes of decay of the nucleus, starting from the resonant state. Γ_α is the entrance channel width, and $\Gamma - \Gamma_\alpha$ is the sum of the widths of all the exit channels except α. If we

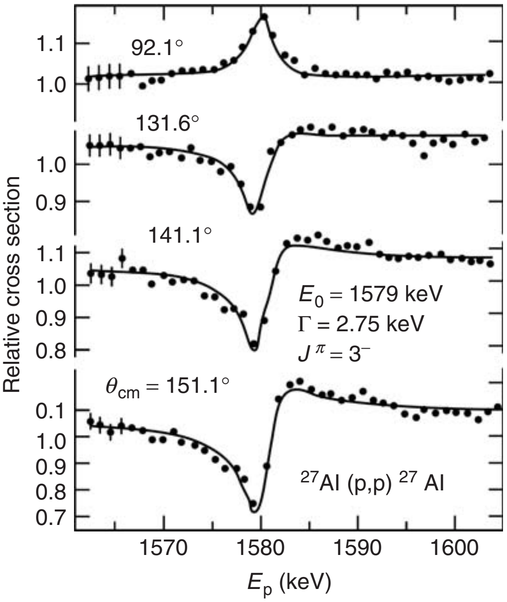

Figure 2.8 Differential cross section at four angles for the elastic scattering of protons off ^{27}Al, in units of the Rutherford cross section, in the neighborhood of the 1579 keV resonance. For larger angles, we have typical interference between the resonant scattering and the potential scattering [7].

restrict the exit channels to a single channel β, or we denote β as the group of exit channels except α, Eq. (2.41) is rewritten as

$$\sigma_{\alpha,\beta} = \pi \lambda\!\!\!^{-2} \frac{\Gamma_\alpha \Gamma_\beta}{(E - E_R)^2 + \left(\frac{\Gamma}{2}\right)^2} \tag{2.43}$$

which is the usual way of presenting the *Breit–Wigner formula*, which describes the form of the cross section close to a resonance. Let us recall that Eq. (2.43) refers to an incident particle of $l = 0$, without charge and without spin. If the spins of the incident and target particles are s_a and s_A, respectively, and the incident beam is described by a single partial wave $l \neq 0$, one can show that the cross section of Eq. (2.43) should be multiplied by the statistical factor $g = (2I + 1)/(2s_a + 1)(2s_A + 1)$, where I is the quantum number of the total angular momentum $I = \mathbf{s}_a + \mathbf{s}_A + \mathbf{l}$ of the compound nucleus. g reduces, naturally, to the unit in the case of zero intrinsic and orbital angular momenta.

If the exit channel is the same as the entrance channel α, the cross section should be obtained from Eq. (2.40) and its dependence in energy is more complicated because in addition to the resonant scattering there is the potential scattering, and the cross section from Eq. (2.40) will contain an interference term between both. The presence of interference results in a peculiar aspect of the scattering cross section, which differs from the simple form in Eq. (2.43) for the reaction cross section. This is seen in Figure 2.8 that shows the forms that a resonance can take in the scattering cross section.

The region of energy where resonances show up can extend to 10 MeV in light nuclei but it ends well before this in heavy nuclei. Starting from this limit, the increase in level density with energy implies that the average distance between levels is smaller than the width of the levels and individual resonances cannot be resolved experimentally. They form a continuum and this region is known as the *continuum region*. The cross section in this region fluctuates.

2.3.4
Statistical Theory of Nuclear Reactions

The fluctuating behavior of low-energy nuclear reactions is due to the interference of the reaction amplitudes corresponding to the excitation of each of the overlapping states, which vanish in the energy average of the cross section since these amplitudes are complex functions with random modulus and phase. Calling $\sigma_{\mathrm{CN}}(c)$ the cross section for the formation of a compound nucleus in the entrance channel c, and using the *reciprocity theorem*, which relates the cross section $\sigma_{cc'}$ to the cross section for the time-reversed process $c' \to c$, one gets

$$\sigma_{cc'}(E_{c'})\mathrm{d}E_{c'} = \sigma_{\mathrm{CN}}(c) \times \frac{(2I_{c'}+1)\mu_{c'}E_{c'}\;\sigma_{\mathrm{CN}}(c')\omega(U_{c'})\mathrm{d}U_{c'}}{\sum\limits_{c}\int_0^{E_c^{\max}}(2I_c+1)\mu_c E_c\;\sigma_{\mathrm{CN}}(c)\omega(U_c)\mathrm{d}U_c} \tag{2.44}$$

where I_c is the angular momentum and μ_c is the reduced mass in channel c. Ejectiles with energy in the range $E_{c'}$ to $E_{c'} + \mathrm{d}E_{c'}$ leave the residual nucleus with energy in the range $U_{c'}$ to $U_{c'} + \mathrm{d}U_{c'}$, where $U_{c'} = E_{\mathrm{CN}} - B_{c'} - E_{c'}$ and E_{CN} and $B_{c'}$ are, respectively, the compound nucleus energy and the binding energy of the ejectile in the compound nucleus. Equation (2.44) is the *Weisskopf–Ewing formula* for the angle-integrated cross sections [9]. To a good approximation, the level density $\omega(U) \propto \exp\,(U/T)$, so the ejectile spectrum given by the Weisskopf–Ewing theory is Maxwellian. It rises rapidly above the threshold energy, attains a maximum, and then falls exponentially.

The Weisskopf–Ewing theory is simple to use, but it has the disadvantage that it does not explicitly consider the conservation of angular momentum and does not give the angular distribution of the emitted particles. This is provided by the *Hauser–Feshbach theory* [10]. This theory takes into account the formation of the compound nucleus in states of different J and parity π. Let us consider the case of a reaction leading from the initial channel c to a final channel c'. If there is no *pre-equilibrium emission*, one may identify the compound nucleus formation cross section $\sigma_{\mathrm{CN}} = \sum_{J,\pi}\sigma_{\mathrm{CN}}^{J\pi}$ with the optical model reaction cross section $\sigma_R = \pi/k^2\sum_l(2l+1)T_l$, which, if the transmission coefficients $T_l = 1 - |S_l|^2$, do not depend on J. S_l is the average value of the scattering amplitude over several overlapping resonances.

Using the above assumptions, the reciprocity theorem, and following a similar derivation as with the Ewing–Weisskopf theory, the cross section for transition $c \to c'$ is given by the *Hauser–Feshbach* expression

$$\sigma_{cc'} = \frac{\pi}{k^2}\sum_J \frac{(2J+1)}{(2i_c+1)(2I_c+1)} \times \frac{\sum\limits_{s,l} T_l(c)\sum\limits_{s',l'} T_{l'}(c')}{\sum\limits_{c}\sum\limits_{s,l} T_l(c)} \tag{2.45}$$

The compound nucleus states may be both of positive and negative parity. Since parity is conserved, in evaluating the expression (2.45), one must take into account that the parity of compound nucleus states and the parity of the residual nucleus states may impose restrictions on the values of the emitted particle angular momentum. Thus, positive parity compound nucleus states decay to positive parity states of the residual only by even angular momenta and to negative parity residual nucleus states by odd angular momenta.

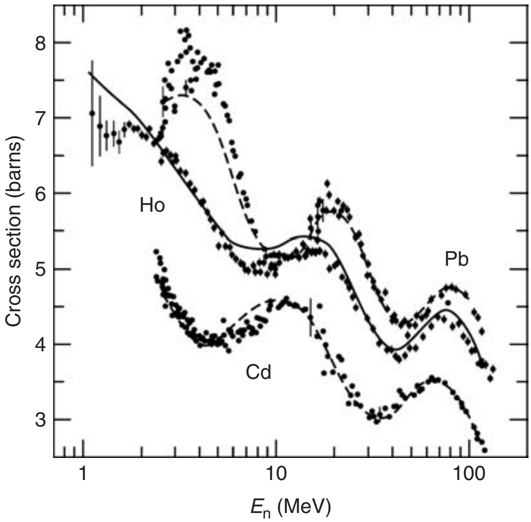

Figure 2.9 Total cross section for neutrons incident on cadmium, holmium, and lead, showing an oscillatory behavior. The curves for cadmium and lead are results of calculations with the optical model [8].

2.3.5
The Optical Model

Expression (2.18) shows that if a particle is always absorbed when it reaches the nucleus, the total cross section falls monotonically with the energy and grows linearly with $A^{\frac{1}{3}}$. In Figure 2.9, we see curves of total cross section of neutron scattering showing an oscillatory behavior for its energy dependence as well as for the mass dependence of the target. Their presence is mainly due to interference phenomena between the part of the incident beam that passes through the nucleus and the part that passes around it [11].

The basis of the *optical model* was established by Feshbach and collaborators in 1953 [12]. In this model, the interaction between the nuclei in a reaction is described by a potential $U(r)$, with r being the distance between the CM of the two nuclei. One replaces the complicated interaction that a nucleon has with the rest of the nucleus with a potential that acts on the nucleon. The potential $U(r)$ includes a complex part that takes into account the absorption effects, that is, the inelastic scattering. The nuclear scattering is treated in similar form as the scattering of light by a glass sphere and the name of the model is derived from this analogy.

In its most commonly used form, the optical potential is written as the sum

$$U(r) = U_R(r) + U_I(r) + U_D(r) + U_S(r) + U_C(r) \tag{2.46}$$

which contains parameters that can vary with energy and masses of the nuclei and that should be chosen by an adjustment to the experimental data. Obviously, the optical potential $U(r)$ will only make sense if these variations are small for neighboring masses or energies.

The first part of Eq. (2.46), $U_R(r) = -Vf(r, R, a)$, is real and represents a nuclear well with depth $-V$, being multiplied by a Woods–Saxon form factor $f(r, R, a) = [1 + \exp\{(r - R)/a\}]^{-1}$, where R is the radius of the nucleus and a measures the diffuseness of the potential, that is, the width of the region where the function f is sensibly different from 0 or

1. V, R, and a are treated as adjustable parameters.

The absorption effect or, alternatively, the disappearance of particles from the elastic channel is taken into account, including the two following imaginary parts $U_I(r) = -\mathrm{i}Wf(r, R_I, a_I)$ and $U_D(r) = 4\mathrm{i}a_I W_D \mathrm{d}f(r, R_I, a_I)/\mathrm{d}r$. An imaginary part produces absorption. It is easy to see this for the scattering problem of the square well: if an imaginary part is added to the well, $U(r) = -V_0 - \mathrm{i}W_0(r < r_0)$ and $U(r) = 0(r > r_0)$, it appears in the value of $K = [2m(E + V_0 + \mathrm{i}W_0)]^{1/2}/\hbar$. This will produce an exponentially decreasing internal wavefunction. Thus, it corresponds to an absorption of particles from the incident beam.

U_I is responsible for the absorption in the whole volume of the nucleus, but U_D, built from the derivative of the function f, acts specifically in the region close to the nuclear surface. These two parts have complementary goals: at low energies, there are no available unoccupied states for nucleons inside the nucleus and the interactions are essentially at the surface. In this situation, $U_D(r)$ is important and $U_I(r)$ can be ignored. On the other hand, at high energies, the incident particle has larger penetration and, in this case, the function $U_I(r)$ is important.

As with the shell model potential, used in nuclear structure, a spin–orbit interaction term is added to the optical potential. This term, which is the fourth part of Eq. (2.46), is usually written in the form

$$U_S(r) = \mathbf{s{\cdot}l}\ V_s\ \frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}f(r, R_S, a_S) \quad (2.47)$$

s is the spin operator and **l** the angular orbital momentum operator. As with $U_D(r)$, the part $U_S(r)$ is only important at the surface of the nucleus since it contains the derivative of the form factor f. The values of V_S, R_S, and a_S must be adjusted by experiment.

The presence of the term U_S is necessary to describe the effect of *polarization*. Through experiences of double scattering, it can be verified that proton or neutron beams suffer strong polarization at certain angles. This means that the quantity

$$P = \frac{N_\mathrm{c} - N_\mathrm{b}}{N_\mathrm{c} + N_\mathrm{b}} \quad (2.48)$$

where N_c is the number of nucleons in the beam with spin up and N_b with spin down, has a value significantly different from zero at these angles. With the inclusion of U_S, the optical model is able to reproduce in many cases the experimental values for the polarization given in Eq. (2.48).

Finally, a term corresponding to the Coulomb potential is added to Eq. (2.46) whenever the scattering involves charged particles. It has the form

$$U_C(r) = \frac{Z_1Z_2e^2}{2R_\mathrm{c}}\left(3 - \frac{r^2}{R_\mathrm{c}^2}\right)\ (r \le R_\mathrm{c}) \quad (2.49)$$

$$= \frac{Z_1Z_2e^2}{r}\ (r > R_\mathrm{c}) \quad (2.50)$$

where it is assumed that the nucleus is a homogeneously charged sphere of radius equal to the *Coulomb barrier radius* R_c, which separates the regions of nuclear and Coulomb forces.

Figure 2.10 exhibits the result of the application of Eq. (2.46) to the elastic scattering of 17 MeV protons on several light nuclei. The angular distribution is very well reproduced by the model, which also reproduces correctly the polarization given by Eq. (2.46) for copper as a function of the scattering angle.

The optical model has a limited set of adjustable parameters and is not capable to describe abrupt variations in the

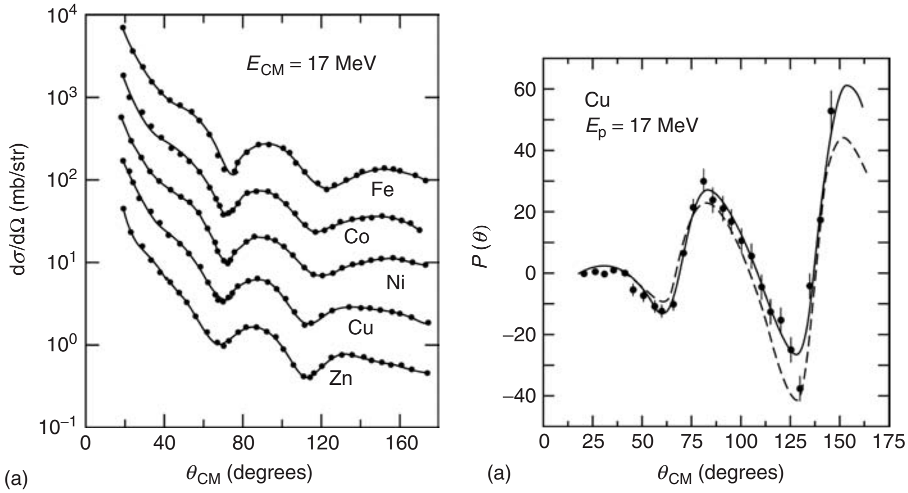

Figure 2.10 (a) Angular distribution of the elastic scattering of 17 MeV protons on nuclei in the region $Z = 26 - 30$. (b) The polarization value Eq. (2.48) in the scattering of 9.4 MeV protons on copper. The curves are, in both cases, obtained with fits from the optical model [13].

cross sections, as it happens for isolated resonances. However, it can provide a good description of the cross sections in the presence of the oscillations of large width in the continuous region, as it treats these as wave phenomena.

2.4
Direct Reactions

Direct reactions become more probable as one increases the energy of the incident particle: the wavelength associated with the particle decreases and localized areas of the nucleus can be "probed" by the projectile. In this context, peripheral reactions, where only a few nucleons of the surface participate, become important. Direct reactions happen during a time of the order of $10^{-22}\$$. Reactions with formation of compound nuclei can be up to six orders of magnitude slower. A reaction type at a given energy is not necessarily exclusive; the same final products can be obtained, part of the events in a direct manner, other parts through the formation and decay of a compound nucleus.

There are two characteristic types of direct reactions. In the first, the incident particle scatters inelastically and the transferred energy is used to excite a collective mode of the nucleus. Rotational and vibrational bands can be studied in this way. The second type involves a modification in the nuclear composition. Examples are transfer of nucleons, as *pickup* and *stripping* reactions. An important reaction of the latter kind is a *knockout* reaction where the incident particle knocks out a particle of the target nucleus and continues in its path, resulting in three reaction products. Reactions with nucleon exchange can also be used to excite collective states. An example is a pickup reaction where a projectile captures a neutron from a deformed target and the product nucleus is in an excited state belonging to a rotational band.

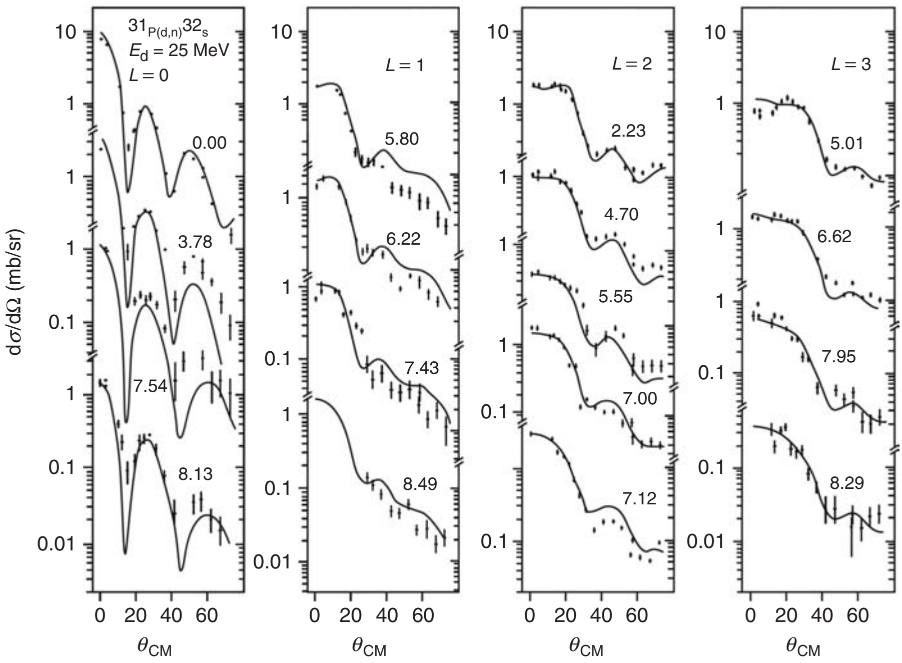

Figure 2.11 Angular distribution of the reaction $^{31}\mathrm{P}(d,n)^{32}\mathrm{S}$, with the transfer of a proton to several states of $^{32}\mathrm{S}$. The curves are results of DWBA calculations for the indicated l values [14].

Direct reactions exhibit a peculiar form of angular distribution, which allows us to extract information on the reaction mechanism with the employment of simple models. Typical examples are the stripping reactions (d,n) and (d,p), where the angular distribution of the remaining nucleon presents a forward prominent peak and smaller peaks at larger angles, with the characteristic aspect of a diffraction figure.

Figure 2.11 shows experimental results for the reaction $^{31}\mathrm{P}(d,n)^{32}\mathrm{S}$ [14]: angular distributions of the detected neutrons corresponding to each energy level of $^{32}\mathrm{S}$ are exhibited for several angular momenta. We see that the behavior of the cross sections is in agreement with qualitative predictions: the curves exhibit a first peak at a value of θ that grows with l. Other smaller peaks occur as θ increases. The increase of θ_l with l is an important characteristic that can be used to identify the value of the transferred momentum in a given angular distribution.

We consider an initial quantum state with particles of mass m_a hitting a target A of mass m_A. The final quantum state are particles of mass m_b moving away from nucleus B of mass m_B. The differential cross section for the process is given by

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{m_\mathrm{a} m_\mathrm{b} k_\mathrm{b}}{(2\pi\hbar^2)^2 k_\mathrm{a}} |V_\mathrm{fi}|^2 \tag{2.51}$$

involving the matrix element

$$V_\mathrm{fi} = \int \Psi_\mathrm{b}^* \Psi_\mathrm{B}^* \Psi_\beta(\mathbf{r}_\beta) V \Psi_\mathrm{a} \Psi_\mathrm{A} \Psi_\alpha(\mathbf{r}_\alpha) \mathrm{d}\tau \tag{2.52}$$

$\Psi_a, \Psi_b, \Psi_A,$ and Ψ_B are the internal wavefunctions of the nuclei a, b, A, and B. Ψ_α, Ψ_β are the wavefunctions of the relative motion in the entrance channel α and in the exit channel β. The integration volume $d\tau$ spans the coordinates of all particles. V is the perturbation potential that causes the ''transition'' from the entrance to the exit channel.

The use of plane waves for Ψ_α and Ψ_β in Eq. (2.52) is known as the *first Born approximation*. With it, we can arrive at an approximate expression for the behavior of the differential cross section. As the nuclear forces are of short range, one can restrict the integral Eq. (2.52) to regions where $\mathbf{r}_\alpha \cong \mathbf{r}_\beta = \mathbf{r}$. This leads to

$$V_{\mathrm{fi}} \cong \int \mathrm{d}\mathbf{r} \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r}) \left\{ \int \Psi_b^* \Psi_B^* V \Psi_a \Psi_A \, \mathrm{d}\tau' \right\} \tag{2.53}$$

where $\mathbf{k} = \mathbf{k}_\alpha - \mathbf{k}_\beta$. The global variables in $d\tau$ were separated into variables $d\mathbf{r}$ and $d\tau'$.

Expanding the plane wave in a Legendre polynomial series, we obtain

$$V_{\mathrm{fi}} \cong \sum_{l=0}^{\infty} \mathrm{i}^l (2l+1) \int j_l(kr) P_l(\cos\theta) F(\mathbf{r}) \mathrm{d}\mathbf{r} \tag{2.54}$$

where $F(\mathbf{r}) = \int \Psi_b^* \Psi_B^* V \Psi_a \Psi_A \mathrm{d}\tau'$ contains all the internal properties and is known as the *form factor* of the reaction. The action of V is restricted to the surface of the nucleus: outside the nucleus the action of V is limited by the short range of the nuclear forces and inside the nucleus there is a strong deviation to the absorption channel. Expression (2.54) becomes $V_{\mathrm{fi}} \cong \sum_{l=0}^{\infty} c_l j_l(kR)$, where the coefficients c_l contain information on the form factor $F(\mathbf{r})$. The index l can be identified as the angular momentum transferred and, for a reaction that involves a single value of l, we can write for the differential cross section:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} \propto |j_l(kR)|^2 \tag{2.55}$$

where the dependence in θ is contained in

$$k^2 = k_\alpha^2 + k_\beta^2 - 2k_\alpha k_\beta \cos\theta.$$

We have an oscillatory behavior for the angular distribution, the maxima separated by π from each other in the axis kR.

The Born approximation with plane waves predicts for certain cases the correct place of the first peaks in the angular distribution but without reproducing correctly the intensities. A considerable progress can be made in the perturbative calculations if, instead of plane waves in Eq. (2.52), we use *distorted* waves that contain, besides the plane wave, the part dispersed elastically by the optical potential of the target. The Born approximation with distorted waves, or distorted wave Born approximation (DWBA), became a largely employed tool in the analysis of experimental results of direct reactions. With it, one can try to extract with a certain reliability the value of the angular momentum l transferred to the nucleus in a stripping or pickup reaction. An example of this is the already mentioned stripping reaction, $^{31}\mathrm{P}(d,n)^{32}\mathrm{S}$, for deuterons of 25 MeV. For the energy levels shown in Figure 2.11, the assignments of the value of l for the level is, in most cases, univocal.

The angular momentum l transferred in a direct reaction generally modifies the value of the total angular momentum of the nucleus. If J_i is the spin of the target nucleus, the spin J_f of the product nucleus is limited to the values

$$\left| |J_\mathrm{i} - l| - \frac{1}{2} \right| \le J_\mathrm{f} \le J_\mathrm{i} + l + \frac{1}{2} \tag{2.56}$$

and the initial and final parities obey the relationship $\pi_\mathrm{i}\pi_\mathrm{f} = (-1)^l$. Equation (2.56) allows, with the knowledge of the target nucleus and of the transferred angular momentum, the determination of the parity of the product state formed and is a tool for the determination of its spin.

The knowledge of the transferred angular momentum value in a direct reaction opens the possibility to test the predictions of the shell model for the structure of nuclei. In a direct reaction, one assumes that the nucleon is located in an orbit of the nucleus with the same angular momentum as the transferred momentum in the reaction. In almost all cases studied with direct reactions, the value of the assigned l corresponds exactly to that predicted by the shell model. We know, however, that the real situation is more complicated, owing to the presence of the residual interactions that give place to configuration mixing. As a result, the cross section for the formation of a state i of the product nucleus is related to that calculated with DWBA for the formation from a single-particle state by

$$\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}\right)_{\mathrm{exp}} = \frac{2J_\mathrm{B}+1}{2J_\mathrm{A}+1}\mathcal{S}_{ij}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}\right)_{\mathrm{DWBA}} \tag{2.57}$$

where the *spectroscopic factor* $\mathcal{S}_{ij}$ measures the weight of the configuration j used in the DWBA calculation, in the final state i, with the sum-rule

$$\sum_j \mathcal{S}_{ij} = n_i \tag{2.58}$$

The sum (2.57) embraces all the nucleons in the configurations j of the product nucleus. The statistical weight $(2J_\mathrm{B}+1)/(2J_\mathrm{A}+1)$ that appears in the DWBA calculation involving the angular momentum of the target nucleus J_A and final nucleus J_B is explicitly given in Eq. (2.58).

2.5
Heavy Ion Reactions

2.5.1
Types and Properties

Heavy ion reactions (with $A > 4$) can be separated into three major categories.

1. Owing to their large charge, two heavy nuclei feel a strong mutual Coulomb repulsion. To produce a nuclear reaction, the projectile needs enough energy to overcome the Coulomb barrier. For a very heavy target, such as ^{238}U, about 5 MeV per nucleon is necessary. Then the wavelength of the projectile is small compared with the dimensions of the nuclei and classical and semiclassical methods become useful in the description of the reaction.
2. The projectile carries a large amount of angular momentum and a good part of it can be transferred to the target in the reaction. Rotational bands with several dozens of units of angular momentum can be created. In fact, heavy ion reactions are the best suited to feed high spin levels.
3. Direct reactions and formation of compound nucleus are also common processes in reactions with heavy ions. But some peculiarities of these are not found in reactions with projectile nucleons. One of these processes can be understood as intermediate between a direct reaction and the formation of a compound nucleus. Fusion does not occur but the projectile and target pass a relatively long time under the mutual action of the nuclear forces. Nuclear matter is exchanged between both and there is a strong heating of the two nuclei, with a large transfer of kinetic

energy to the internal degrees of freedom. These are the *deep inelastic collisions*.

The kind of process that prevails depends on the distance of closest approach d between the projectile and the target. If this distance is sufficiently large only the long-range Coulomb interaction acts and, for a classical hyperbolic trajectory, d is related to the impact parameter b and to the energy E of the projectile by $d = a/2 + [(a/2)^2 + b^2]^{1/2}$, where a is the distance of closest approach in a head-on collision. This is related to E by $a = Z_1 Z_2 e^2/4\pi_0 E$.

Experimentally, the variable under control is the energy E of the projectile and, for E sufficiently large, d can be small enough to enter the range of nuclear forces. Collisions near this limit are called *grazing* collisions and are characterized by values of b_{graz} and d_{graz}. Assuming that there is always reaction when $b < b_{\text{graz}}$, the reaction cross section σ_r can be determined geometrically by $\sigma_r = \pi b^2_{\text{graz}}$. The experimental determination of σ_r allows to establish the value $d_{\text{graz}} = 0.5 + 1.36\ (A_1^{1/3} + A_2^{1/3})$ showing that the distance of grazing collision is somewhat larger than that deduced from two touching spheres ($1.36 fm > r_0 = 1.2$ fm).

When the impact parameter is close to b_{graz}, one expects nuclear reactions of short duration, without the contribution of the compound nucleus formation. Such reactions are elastic and inelastic scattering and involve a transfer of few nucleons. When the incident energy is sufficiently high, small values of b can lead to the projectile penetrating the target. Depending on the energy and on the involved masses, the reaction can end in one of the processes below:

1. *Fusion* – this is the preferred process when one has light nuclei and low energy. There is formation of a highly excited compound nucleus that decays by evaporation of particles and γ-radiation emission, leading to a cold residual nucleus. If the energy in the CM is close to the Coulomb barrier energy, the cross section of compound nucleus formation starting from two nuclei is practically equal to the reaction cross section.
2. *Fission* – when the compound nucleus is heavy the fission process competes strongly with the evaporation of particles in each stage of the evaporation process. A very heavy compound nucleus with a large excitation energy has a very small probability of arriving to a cold residual nucleus without fission at some stage of the de-excitation. The role of the angular momentum l transmitted to the target nucleus is also essential. The fission barrier decreases with the increase of l and for a critical value l_{crit} the barrier ceases to exist. A nucleus with angular momentum greater than l_{crit} suffers immediate fission, and this is also a limiting factor in the production of superheavy elements.
3. *Deep inelastic collision (DIC)* – this is a phenomenon characteristic of reactions involving very heavy nuclei ($A > 40$) and with an incident energy of 1–3 MeV above the Coulomb barrier. In DIC, the projectile and the target spend some time under mutual action, exchanging masses and energy but without leading to the formation of a compound nucleus. The projectile escapes after transferring part of its energy and angular momentum to its internal degrees of freedom and to the

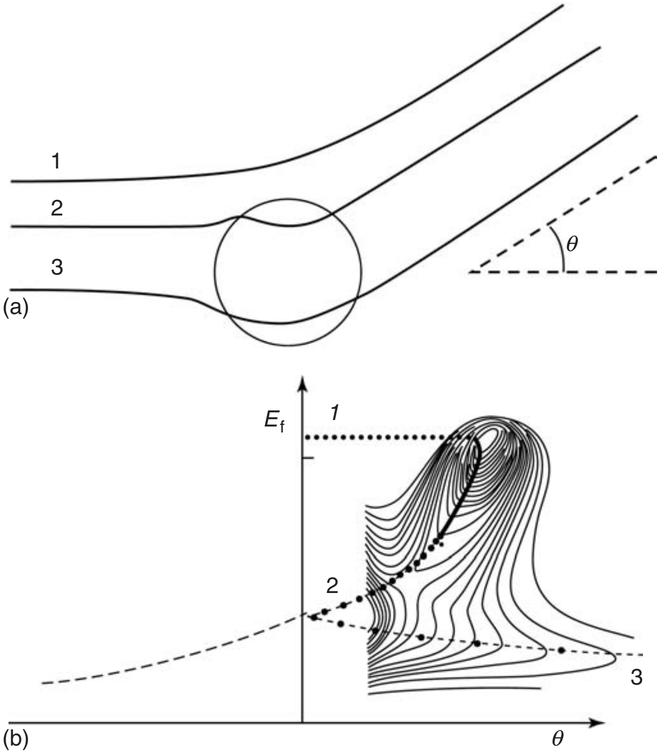

Figure 2.12 (a) Reactions with different impact parameters leading to the same scattering angle θ. (b) Lines of maxima in topographical diagrams of the final energy against the scattering angle.

target, with values reaching 100 MeV and 50ℏ.

One of the most interesting aspects of DIC is the correlation between the energy dissipated in the collision and the scattering angle in the CM. Let us look at Figure 2.12a. Trajectory 1 shows the projectile with an impact parameter that leads it out of the range of nuclear forces. The Coulomb scattering angle will become larger as the impact parameter decreases. In a graph as in Figure 2.12b, where one plots the final energy against the scattering angle, trajectories of type 1 are located in the upper branch, where there is no dissipation and the initial kinetic energy stays unaffected. This upper branch has a maximum value for the scattering angle. At a smaller impact parameter, the nuclear attractive force and, with that, the effects of dissipation of DIC begin to act. A given Coulomb scattering angle θ can also be reached by the combination of the nuclear and Coulomb forces. Only, now there is loss of energy and the events are located in branch 2 of Figure 2.12b. There still is no one-to-one correspondence between angle and energy because an infinite number of trajectories can lead to the same angle θ. Branch 2 should be understood as a line of maxima in a three-dimensional representation (called the *Wilczynski diagram*) where the axis perpendicular to the paper is proportional to the cross section $d^2\sigma/dEd\theta$. The same angle θ can also be obtained by trajectory 3, with a longer interaction time between the nuclei and a larger dissipation. As now the projectile is deflected toward the nucleus, the scattering angle would be formally $-\theta$ but, as there is no experimental distinction between θ and $-\theta$, these events appear as an independent ramification in the lower part of the diagram.

2.5.2
Superheavy Elements

The heaviest element found in Nature is $^{238}_{92}$U. It is radioactive, but it has survived since its formation in supernovae explosions because it has a decay half-life of the order of the age of the Earth. Elements with larger atomic number (transuranic) have shorter half-lives and have disappeared. They are created artificially through nuclear reactions using heavy elements as target. Initially, the projectiles used were light particles: protons, deuterons, α-particles and neutrons. The use of neutrons is justified because the β^--emission of the compound nucleus increases the value of Z and it was in this way [15] that the first transuranic element, neptunium, was obtained: $\text{n}+{}^{238}\text{U}\rightarrow{}^{239}\text{U}\rightarrow{}^{239}\text{Np}+\beta^-$.

Reactions with light particles can produce isotopes up to mendelevium ($Z = 101$), but it is not possible to go beyond that; the half-lives for α-emission or spontaneous fission become extremely short, making the preparation of a target impracticable. The alternative is to place a heavy element under the flux of very intense neutrons. This can be done using special reactors or using the remaining material of nuclear explosions. The elements einsteinium ($Z = 99$) and fermium ($Z = 100$) were discovered in this way in 1955 but the increasing competition of beta decay with alpha decay and with spontaneous fission prevents this method from being used for larger Z.Starting in 1955, heavy ion accelerators began to deliver beams with high enough intensity and energy to compete in the production of transuranic isotopes. The first positive result was the production of two californium isotopes ($Z = 98$) with the fusion of carbon and uranium nuclei:

$$\begin{aligned} {}^{12}_{6}\text{C} + {}^{238}_{92}\text{U} &\rightarrow {}^{244}_{98}\text{Cf} + 6\text{n}, \\ {}^{12}_{6}\text{C} + {}^{238}_{92}\text{U} &\rightarrow {}^{246}_{98}\text{Cf} + 4\text{n} \end{aligned} \tag{2.59}$$

This opened up the possibility of directly reaching the nucleus one wants to create from the fusion of two smaller nuclei. The difficulty of such task is that the cross sections for the production of heavy isotopes are extremely low. For example, the reaction ${}^{50}_{22}\text{Ti} + {}^{208}_{82}\text{Pb} \rightarrow {}^{257}_{104}\text{Rf} + \text{n}$, which produces the element rutherfordium, has a cross section of only 5 nb. A small increase in the charges drastically reduces this value. The cross section for the fusion reaction ${}^{58}_{26}\text{Fe} + {}^{208}_{82}\text{Pb} \rightarrow {}^{265}_{108}\text{Hs} + \text{n}$ is 4 pb. In comparison, the typical cross sections of DIC for heavy nuclei are in the range 1–2 b.

In spite of the experimental refinement that these low cross sections demand, one is able to produce isotopes with charge as heavy as $Z = 118$. The understanding of the mechanisms that lead to fusion is, however, not fully understood. According to the traditionally accepted model, the fusion of two nuclei proceeds in two stages: the formation of a compound nucleus and the de-excitation of the compound nucleus by evaporation of particles, preferentially neutrons. The difficulties for the materialization of the process in very heavy nuclei reside in both stages [16].

2.6
Electromagnetic Probes

2.6.1
Coulomb Excitation

Coulomb excitation is an inelastic scattering process in which a nucleus excites another nucleus with its electromagnetic field V. This field can be decomposed in terms of a series of multipoles, for example, $E1, E2, M1, \cdots$, which carry well-defined angular momenta and parities. At low bombarding energies, $E2$ (quadrupole) excitations are more common, while at higher energies, $E1$ (dipole) excitations dominate. The Coulomb excitation cross section in first-order perturbation theory is given by

$$\frac{d\sigma_{i\to f}}{d\Omega} = \left(\frac{d\sigma}{d\Omega}\right)_{\text{el}} \frac{16\pi^2 Z_2^2 e^2}{\hbar^2} \times \sum_{\pi\lambda\mu} \frac{B(\pi\lambda, I_i \to I_f)}{(2\lambda+1)^3} |S(\pi\lambda, \mu)|^2 \tag{2.60}$$

where $(d\sigma/d\Omega)_{\text{el}}$ is the Rutherford elastic scattering cross section, Z_2 is the charge of the projectile, $B(\pi\lambda, I_i \to I_f)$ is the *reduced matrix element* of the target, $\pi\lambda = E1, E2, M1, \ldots$ is the multipolarity of the excitation, and $\mu = -\lambda, -\lambda+1, \ldots, \lambda$. The orbital integrals $S(\pi\lambda, \mu)$ contain information on the dynamics of the reaction [17].

Because Coulomb excitation is an external process, that is, it occurs when the nucleons from one nucleus are outside the nuclear matter distribution from the other nucleus, the matrix elements for Coulomb excitation are the same as those for excitation by real photons (except for $E0$ Coulomb excitations, which are very small). Therefore, Coulomb excitation cross sections can be written as

$$\frac{d\sigma_C(E_x)}{dE_x} = \sum_{E\lambda} \frac{n_{E\lambda}(E_x)}{E_x} \sigma^{\gamma}_{E\lambda}(E_x) + \sum_{M\lambda} \frac{n_{M\lambda}(E_x)}{E_x} \sigma^{\gamma}_{M\lambda}(E_x) \tag{2.61}$$

where $\sigma^{\gamma}_{\pi\lambda}(E_x)$ are the *photonuclear cross sections* for the multipolarity $\pi\lambda$, E_x is the excitation energy, and $n_{M/E\lambda}(E_x)$ are known as *virtual photon numbers* [17].

The photonuclear cross sections are related to the reduced matrix elements, for the excitation energy E_x, through the relation [17]

$$\sigma_{\gamma}^{\pi\lambda}(E_x) = \frac{(2\pi)^3(\lambda+1)}{\lambda[(2\lambda+1)!!]^2} \left(\frac{E_x}{\hbar c}\right)^{2\lambda-1} \times \frac{dB(\pi\lambda, E_x)}{dE_x} \tag{2.62}$$

where dB/dE_x are the *electromagnetic response functions*, such that

$$B(\pi\lambda, I_i \to I_f) = \int dE_x \frac{dB(\pi\lambda, E_x)}{dE_x}$$

For differential cross sections, an analog expression exists, namely,

$$\frac{d\sigma_C(E_x)}{d\Omega} = \frac{1}{E_x} \sum_{\pi\lambda} \frac{dn_{\pi\lambda}}{d\Omega}(E_x, \theta) \sigma_{\gamma}^{\pi\lambda}(E_x) \tag{2.63}$$

where Ω denotes the solid scattering angle. This is not different from Eq. (2.60), only rewritten in a more practical way. It helps us to see how we can access information on phenomena induced by real photons by using the field of a charged particle.

In Figure 2.13, we show a calculation (with $E_\gamma \equiv E_x$) of the virtual photons for the $E1$ multipolarity, "as seen" by a projectile passing by a lead target at impact parameters equal to and exceeding $b = 12.3$ fm, for three typical bombarding energies. As the projectile energy increases,

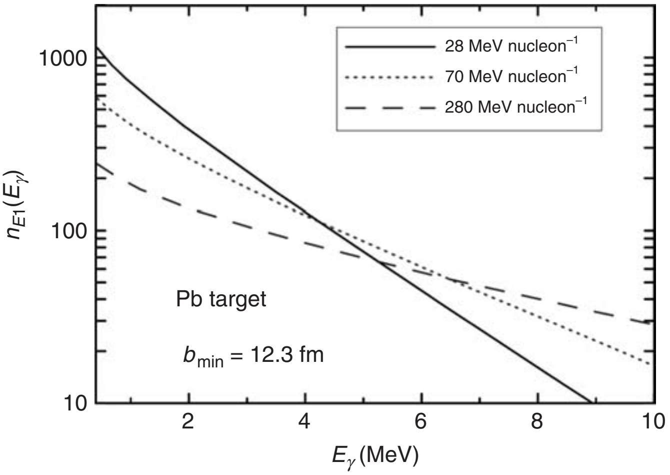

Figure 2.13 Number of virtual photons for the $E1$ multipolarity, "as seen" by a projectile passing by a lead target at impact parameters $b_{\min} = 12.3$ fm and larger, for three typical bombarding energies.

more virtual photons of large energy are available for the reaction. This increases the energy of states probed in the excitation process, making it possible to study giant resonances, lepton and meson production, and so on [17].

2.6.2
Photonuclear Reactions and Giant Resonances

A photonuclear reaction is a reaction resulting from the interaction of the electromagnetic radiation with a nucleus. Therefore, one can access information that are complementary to Coulomb excitation (and vice versa).

When the energy of the photon is located above the separation energy of a nucleon, the cross section of photoabsorption reveals the presence of characteristic sharp resonances. However, when the incident energy reaches the range of 15–25 MeV, a new change appears in the cross section, with the presence of a wide and large peak, called *giant electric dipole resonance* (GDR) ($E1$ excitation). Figure 2.14 exhibits the excitation function of photoabsorption of ^{120}Sn at photon energies around the GDR at 15 MeV.

The giant resonance occurs in nuclei along the whole periodic table, with the resonance energy decreasing as $E_{\text{GDR}}80/A^{1/3}$ for $A > 20$. Their widths are almost all in the range between 3.5 and 5 MeV. In few cases, they can reach 7 MeV. They are a collective excitation, that is, an excitation simultaneously involving many nucleons . The time-varying electric field of the photon is very effective in inducing collective oscillations of protons against neutrons.

The GDR resonance arises from an excitation that transfers by one unit of angular momentum to the nucleus ($\Delta l = 1$). If the nucleus is even–even it is taken to a 1^- state. The transition also changes the isospin of one unit ($\Delta T = 1$) and, so it is also called *isovector resonance*. The photon can excite giant isoscalar

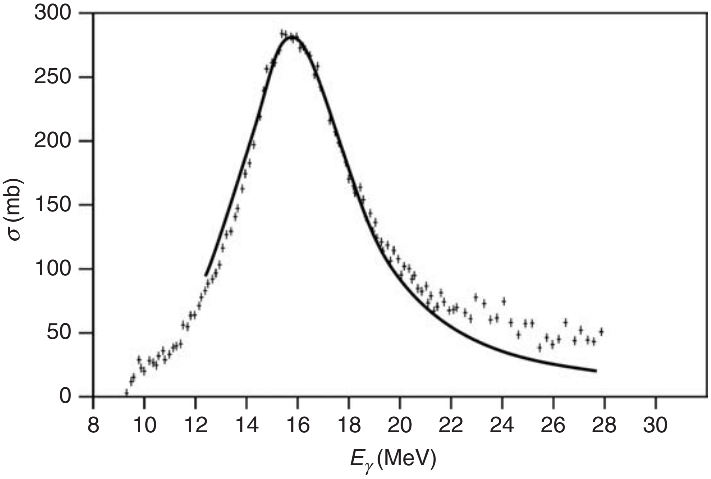

Figure 2.14 Giant resonance in the absorption of photons by ^{120}Sn [18].

resonances less effectively, but still with appreciable cross sections, with $\Delta T = 0$. But electric quadrupole ($\Delta l = 2$) and electric monopole ($\Delta l = 0$) are observed mostly in reactions with charged particles. In a giant electric quadrupole resonance, the nucleus oscillates between a spherical (assuming that this is the form of the ground state) and an ellipsoidal form. If protons and neutrons vibrate in phase, we have an isoscalar resonance ($\Delta T = 0$) and if they are oscillate in opposite phase we have an isovector resonance ($\Delta T = 1$). The giant monopole resonance is a very special case of nuclear excitation where the nucleus contracts and expands radially, maintaining its original form but changing its volume. It is also called *breathing mode*. It can also happen in isoscalar and isovector forms. It is an important way to study the *compressibility of nuclear matter*.

Besides electric giant resonances, associated with a variation in the form of the nucleus, magnetic giant resonances exist, which involve *spin vibrations*. In these, nucleons with spin up move out of phase with nucleons with spin down. The number of nucleons involved in the process cannot be very large because it is limited by the Pauli principle. The magnetic resonances can also separate into isoscalar resonances, where protons and neutrons of the same spin vibrate against protons and neutrons of opposite spin, and isovector, where protons with spin up and neutrons with spin down vibrate against their corresponding partners with opposite spins. The last cases are effectively probed in charge-exchange reactions, that is, when the projectile charge changes to $Z - 1$ or $Z + 1$, as in (d,p) and (d,n). They are known as *giant Gamow–Teller resonances*.

Giant resonances can also be produced in an excited nucleus, a case known as the *Brink–Axel hypothesis* [19]. A special case is when two giant resonances are excited simultaneously in the same nucleus. The double giant dipole resonance (or *multiphonon giant resonance* was observed for the first time in reactions with double charge exchange induced by pions in ^{32}S [20]. But Coulomb excitation proved to be the best probe, as shown theoretically in Refs [17, 21].

2.6.3
Electron Scattering

Electron scattering experiments have provided a rich database on most stable or long-lived nuclei. Electrons are structureless point-like objects that only interact electromagnetically. Therefore, electron scattering avoids the complexity of the strong interaction between the projectile and the target and provides clean information about the charge distribution in the nucleus.

Considering relativity and the spin of the electrons, the differential cross section for elastic scattering by spinless point-like nuclei can be expressed by the *Mott scattering formula* (for simplicity, we use $\hbar = c = 1$ units):

$$\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}\right)_{\text{Mott}} = \frac{Z_T^2 e^4 \cos^2(\theta/2)}{4p_0^2 \sin^4(\theta/2)[1+(2p_0/M)\sin^2(\theta/2)]} \tag{2.64}$$

where Z_{T_e} and M are the charge and mass of the target nucleus, respectively, and p_0 is the momentum of the incoming electron. As the nucleus is not a point-like particle, the formula is modified by adding the nuclear electric and magnetic form factors, which contain information of the charge and magnetization distributions inside the nucleus.

The *Rosenbluth formula* accounts for the electron spin and explicitly expresses the cross section for arbitrary nuclei as

$$\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}\right) = \left(\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}\right)_{\text{Mott}} \Big\{A_0(q^2) + B_0(q^2) \times \left[\frac{1}{2} + \left(1 + \frac{q^2}{4M^2}\right)\right]\tan^2(\theta/2)\Big\}$$

where $q^2 = (\mathbf{p}_\mathrm{f} - \mathbf{p}_\mathrm{i})^2 - (E_\mathrm{f} - E_\mathrm{i})^2$ is the 4-momentum transfer squared, where $\mathbf{p}_\mathrm{i}(\mathbf{p}_\mathrm{f})$ and $E_\mathrm{i}(E_\mathrm{f})$ are the electron incoming (outgoing) momentum and energy. $A_0(q^2)$ and $B_0(q^2)$, functions of q^2, are the form factors associated with the charge and magnetization distribution of the nucleus, respectively. For a spin-0 nucleus,

$$A_0(q^2) = \frac{G_E^2(q^2)}{1 + q^2/4M^2} \quad \text{and} \quad B_0(q^2) = 0 \tag{2.65}$$

where $G_E(q^2)$ is the *Sachs charge form factor*, and the factor $1 + q^2/4M^2$ is the kinematical recoil correction. For $q \to 0, q^2 \approx \mathbf{q}^2$, where $\mathbf{q}$ is the three-momentum transfer, and to leading order in powers of $\mathbf{q}$:

$$G_E(q^2) \simeq G_E(\mathbf{q}^2) = \int \rho_T(\mathbf{r}) \mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{r}} \mathrm{d}^3 r \tag{2.66}$$

One thus sees that electron scattering is related to the Fourier transform of the nuclear charge density, $\rho_T(\mathbf{r})$. If the function $G_E(\mathbf{q}^2)$ is mapped for a sufficient number of momentum transfers $\mathbf{q}^2$, the Fourier transform can be inverted and $\rho_T(\mathbf{r})$ can be mapped with precision. For heavy elements ($A > 10$), this technique shows that the charge radius is closely proportional to 1.12 $A^{1/3}$ fm, where A is the nuclear mass number. This indicates the density saturation of the nuclear matter.

When $\mathbf{q}^2 = 0$, inelastic electron scattering probes the same multipolarity transitions ($E1, E2, \ldots$) as in Coulomb excitation or with real photons. However, in contrast to the later probes, electron scattering also provides information on nuclear excitations for cases in which $\mathbf{q}^2 \neq 0$. This is useful in many aspects. For example, at high momentum transfers, the momentum transfer $\mathbf{q}$ can be absorbed by a nucleon in the nucleus, which acquires an energy of the order of $\mathbf{q}^2/2m_N^*$, thus probing the effective mass m_N^* of a nucleon in the nucleus. This is of relevance to understand the effects of nucleon–nucleon interactions in the nuclear environment. The q-region where this occurs is termed *quasi-free scattering*. As electrons can penetrate the nuclei, they can also more effectively probe monopole (or breathing mode) excitations of the nuclei, a case in which the nuclear mass distribution vibrates radially. This is of importance for determining the *compressibility modulus* of the nuclear matter. Finally, in high-energy electron scattering, the electron can penetrate deep inside the nucleons and probe the spin and charge distributions of quarks and gluons inside the nucleon [22].

2.7
Relativistic Nuclear Collisions

2.7.1
Transport Theories and Equation of State

As the bombarding energy in nucleus–nucleus collisions increases, the structure aspects of the nuclei become less relevant. Except for the bulk properties of the nuclei (size and number of nucleons), the physics involved is primarily due to the individual, and sometimes collective, hadronic collisions. Several theoretical tools are used to describe these reactions, and, in particular, we quote (i) *time-dependent Hartree–Fock* (TDHF), (ii) *anti-symmetrized molecular dynamics* (AMD), (iii) *Boltzmann–Uehling–Uhlenbeck* (BUU), and so on.

At intermediate energies of $E_{\text{lab}} \sim 100 - 1000$ MeV nucleon^{-1}, the nucleons and the products of their collisions can be described individually and their propagation can be described by semiclassical equations. One of such equations, and perhaps the most popular in such studies, is the so-called *BUU equation*:

$$\begin{aligned}\frac{\partial f}{\partial t} + \left(\frac{\mathbf{p}}{m} + \nabla_{\mathbf{p}} U\right)\cdot\nabla_{\mathbf{r}} f - \nabla_{\mathbf{r}} U\cdot\nabla_{\mathbf{r}} f = \\ \int \mathrm{d}^3 p_2 \int \mathrm{d}\Omega\, \sigma_{\mathrm{NN}}(\Omega)|\mathbf{v}_1 - \mathbf{v}_2| \\ \times\left\{f_1' f_2' \left[1 - f_1\right]\left[1 - f_2\right] - f_1 f_2 \left[1 - f_1'\right] \right. \\ \left. \times \left[1 - f_2'\right]\right\}. \qquad (2.67)\end{aligned}$$

If dN is the number of particles in the volume element $\mathrm{d}^3 r$ and whose momenta fall in the momentum element $\mathrm{d}^3 p$ at time t, then the distribution function $f(\mathbf{r}, p, t)$ is given by $\mathrm{d}N = f(\mathbf{r}, p, t)\mathrm{d}^3 r \mathrm{d}^3 p$. Thus, the BUU equation is an equation for the distribution function $f(\mathbf{r}, p, t)$. To account for the effect of each particle interacting with all others, one introduces the concept of *mean-field*, $U(\mathbf{r}, p, t)$. This mean-field exerts a force on each particle, given by $-\nabla_{\mathbf{r}} U(\mathbf{r}, p, t)$. In addition, the momentum dependence of the potential introduces a dependence through the derivative $-\nabla_{\mathbf{p}} U(\mathbf{r}, p, t)$.

Owing to the nucleon–nucleon collisions, the distribution function within $\mathrm{d}^3 r \mathrm{d}^3 p$ can also be modified by nucleons leaving (or entering) this volume. This is taken care by of the *collision term*, that is, the right-hand side of the BUU equation. σ_{NN} is the nucleon–nucleon differential scattering cross section, $\mathbf{v}_1$ and $\mathbf{v}_2$ are the velocities of two colliding nucleons. The first factor inside braces are for collisions populating the volume element and the second term for those depleting it. The factors $(1 - f)$ account for Pauli blocking of final occupied states. The integrals average over scattering angle and over all collisions within $\mathrm{d}^3 r \mathrm{d}^3 p$. The BUU equation falls in the category of what one calls *quantum transport theories*. Hadronic transport theories have been quite successful in applications, describing a multitude of measured particle spectra.

Equation (2.67) needs as basic ingredients the mean-field U and the cross section σ_{NN}. Because these two quantities are related to each other, one should in principle derive them in a self-consistent microscopic approach, as in the *Brueckner theory*.

An important ingredient in the transport theory calculations is the compressibility K of nuclear matter, which refers to the second derivative of the compressional energy E with respect to the density [1]

$$K = 9\rho^2 \frac{\partial^2}{\partial \rho^2}\left(\frac{E}{A}\right) \tag{2.68}$$

This is an important quantity, for example, for nuclear astrophysics. Supernova models might or not lead to explosions depending on the value of K. The central collisions of heavy nuclei are one of the few probes of this quantity in the laboratory. The dependence of the calculations on K follow from the dependence of the mean-field potential U ($U \sim E/A +$ kinetic energy terms) on the particle density ρ. A typical parametrization for U is the Skyrme parametrization $U = a\rho/\rho_0 + b(\rho/\rho_0)^\sigma$.

The output of Eq. (2.67) is the distribution function% $f(\mathbf{r}, p, t)$, which allows one to calculate many properties of heavy ion collisions. Let us quote *collective flows*, proton and neutron production rates (subthreshold and above threshold), pion and kaon yields, and so on. Combining Eq. (2.67) with a *phase–space coalescence model*, one can also calculate such quantities as exclusive flows and intermediate fragment formations.

The dynamics of the central high-energy reactions can be broken down into several stages. Baryon-density contour plots are shown in Figure 2.15 for 400 MeV nucleon^{-1}Au + Au collisions at $b = 0$, which will serve to illustrate our points.

Following an initial interpenetration of projectile and target densities, the NN collisions begin to thermalize matter in the overlap region making the momentum distribution there centered at zero momentum in the CM. The density in the overlap region rises above normal and a disk of excited and compressed matter forms at the center of the system. More and more matter dives into the region with compressed matter that begins to expand in transverse directions. At late stages, when the whole matter is excited, transverse expansion predominates.

2.7.2
Kinematics

In relativistic nucleus–nucleus collisions, other definitions are best suited for discussing relations between energies, momenta, and angles. The *rapidity* is a variable frequently used to describe the behavior of particles in inclusively measured reactions. It is defined by

$$y = \frac{1}{2}\ln\left(\frac{E + p_{\parallel}}{E - p_{\parallel}}\right) \tag{2.69}$$

which corresponds to $\tanh y = p_{\parallel}/E$, where y is the rapidity, $p_{\parallel}$ is the longitudinal momentum along the direction of the incident particle, E is the energy, both defined for a given particle. The accessible range of rapidities for a given reaction is determined by the available CM energy and all participating particles' rest masses. One usually gives the limit for the incident particle, elastically scattered at zero angle:

$$|y_{\max}| = \ln\left[\frac{E + p}{m}\right] = \ln(\gamma + \gamma\beta) \tag{2.70}$$

---

[1] The factor of 9 is a historical artifact from the convention originally defining K.

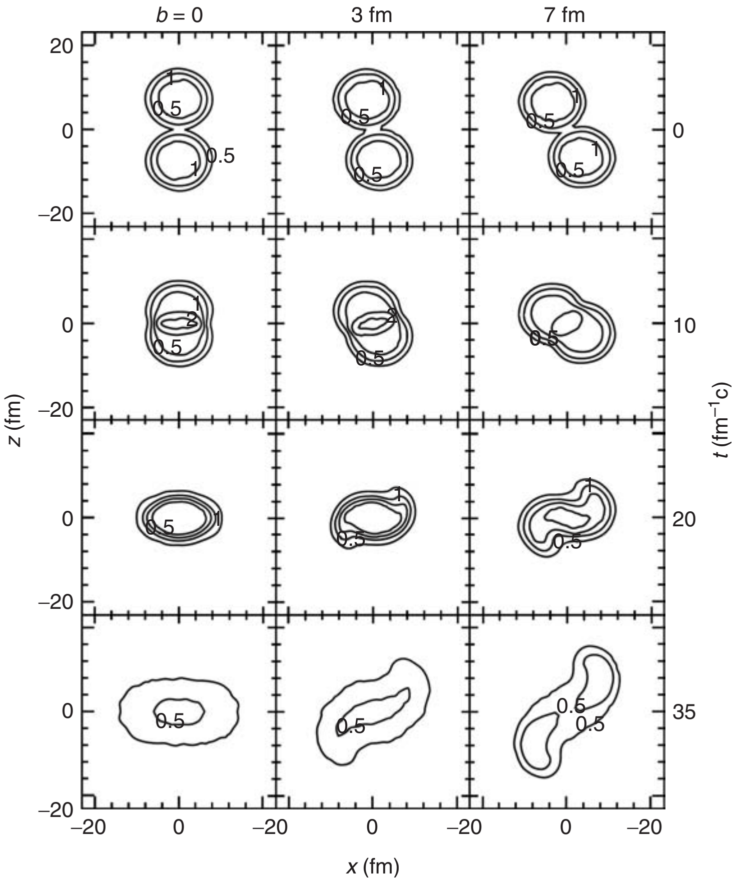

Figure 2.15 Contour plots of baryon density in the reaction plane in Au + Au collisions at 400 MeV nucleon^{-1}. The displayed contour lines are for the densities $\rho/\rho_0 = 0.1$, 0.5, 1, 1.5, and 2 [23].

where $\beta \equiv v$ is the velocity and all variables referring to the through-going particle given in the desired frame of reference (i.e., in the CM).

Note that $\partial y/\partial p_{\parallel} = 1/E$. A Lorentz boost β along the direction of the incident particle adds a constant, ln $(\gamma + \gamma\beta)$, to the rapidity. Rapidity differences, therefore, are invariant to a Lorentz boost. Statistical particle distributions are flat in y for many physics production models. Frequently, the simpler variable *pseudorapidity* η is used instead of rapidity.

The pseudorapidity is a handy variable to approximate the rapidity if the mass and momentum of a particle are not known. It is an angular variable defined by

$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right] \tag{2.71}$$

whose inverse function is $\theta = 2\arctan(e^{-\eta})$, where θ is the angle between the particle being considered and the undeflected beam. η is the same as the rapidity y if one sets $\beta = 1$ (or $m = 0$).

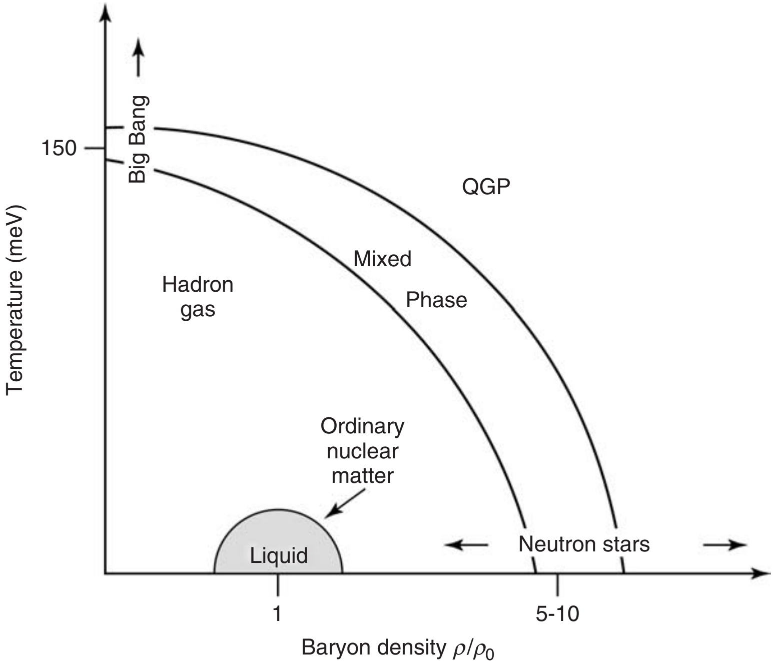

Figure 2.16 As water comes in different phases (solid, liquid, and gas), so nuclear matter can come in its normal hadronic form and at sufficiently high temperature and density, in the form of a deconfined state of quarks and gluons. The diagram shows how nuclear matter should behave as a function of density and temperature.

Statistical distributions plotted in pseudorapidity rather than rapidity undergo transformations that have to be estimated by using a kinematic model for the interaction.

2.7.3
The Quark–Gluon Plasma (QGP)

The primary motivation for studying ultra-relativistic heavy ion collisions is to gain an understanding of the equation of state of nuclear, hadronic, and partonic matter, commonly referred to as *nuclear matter*. Displayed in Figure 2.16 is a schematic phase diagram of nuclear matter. The behavior of nuclear matter as a function of temperature and density (or pressure), shown in Figure 2.16, is governed by its equation of state.

Conventional nuclear physics is concerned primarily with the lower left portion of the diagram at low temperatures and near normal nuclear matter density. Here normal nuclei exist and at low excitation, a liquid–gas phase transition is expected to occur. This is the focus of experimental studies using low-energy heavy ions. At somewhat higher excitation, nucleons are excited into baryonic resonance states, along with accompanying particle production and hadronic resonance formation. In relativistic heavy ion collisions, such excitation is expected to create hadronic resonance matter.

We now briefly discuss the QGP signatures in nucleus–nucleus collisions. For more details see, for example, Refs [24, 25]. One group of such signatures can be classified as *thermodynamic variables*. This class involves determination of the energy density ϵ, pressure P, and entropy density s of the interacting system as a function of the temperature T and the baryochemical potential μ_B. Experimental observables can be identified with these variables and thus their relative behavior can be determined. If a phase transition to QGP occurs, a rapid rise in the effective number of degrees of freedom, expressed by $/T^4$ or s/T^3, should be observed over a small range of T. The variables T, s, and can be identified with the average transverse momentum p_T, the hadron rapidity density $\mathrm{d}N/\mathrm{d}y$, and the transverse energy density $\mathrm{d}E_T/\mathrm{d}y$, respectively. The transverse energy produced in the interaction is $E_T = \sum_i E_i \sin\theta_i$, where E_i and θ_i are the kinetic energies of the ejectiles and the emission angles.

Electromagnetic (EM) probes, such as photons and leptons, provide information on the various stages of the interaction without modification by final state interactions. These probes may provide a measure of the thermal radiation from a QGP, if a region of photon energy, or equivalently lepton pair invariant mass, can be isolated for emission from a QGP relative to other processes. However, the yields for EM probes are small relative to background processes, which are primarily EM decays of hadrons and resonances. Lepton pairs from the QGP are expected to be identifiable in the 1–10 GeV invariant mass range. The widths and positions of the ρ, ω, and ϕ peaks in the lepton pair invariant mass spectrum are expected to be sensitive to medium-induced changes of the hadronic mass spectrum.

The production of J/Ψ particles in a quark–gluon plasma is predicted to be suppressed (Figure 2.17). This is a result of the *Debye screening* of a $c\bar{c}$ pair, initially formed in the QGP by fusion of two incident gluons. Less tightly bound excited states of the $c\bar{c}$ system, such as Ψ' and χ_c, are more easily dissociated and will be suppressed even more than the J/Ψ.

A long-standing prediction for a signature of QGP formation is the enhancement of strange hadrons. The production of strange hadrons relative to nonstrange hadrons is suppressed in hadronic reactions. This suppression increases with increasing strangeness content of the hadron. In a QGP, the strange quark content is rapidly saturated by $s\bar{s}$ pair production in gluon–gluon reactions, resulting in an enhancement in the production of strange hadrons. Thus, multistrange baryons and strange antibaryons are predicted to be strongly enhanced when a QGP is formed.

The connection between energy loss of a quark and the color-dielectric polarizability of the medium can be established in analogy with the theory of electromagnetic energy loss. Although radiation is a very efficient energy loss mechanism for relativistic particles, it is strongly suppressed in a dense medium by the *Landau–Pomeranchuk effect* [24]. Adding the two contributions, the stopping power of a quark–gluon plasma is predicted to be higher than that of hadronic matter. A quark or *gluon jet* propagating through a dense medium will not only lose energy but will also be deflected. This effect destroys the coplanarity of the two jets from a hard parton–parton scattering with the incident beam axis. The angular deflection of the jets also results in an azimuthal asymmetry. The presence of a quark–gluon plasma is

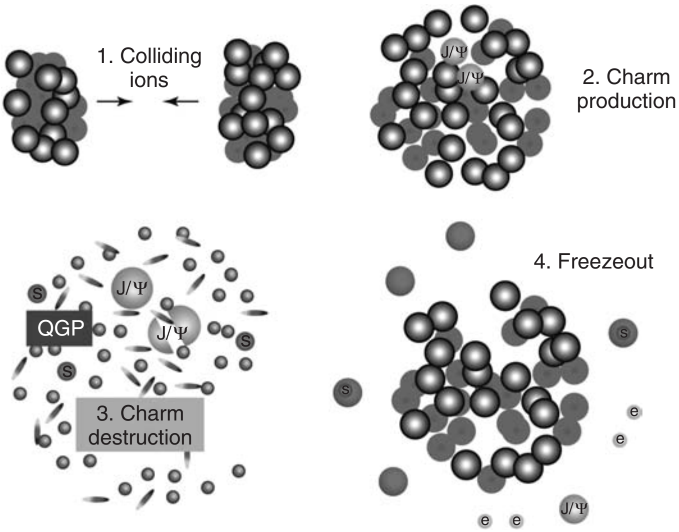

Figure 2.17 Formation and evolution of J/Ψ particles in relativistic heavy ion collisions. After the formation, the J/Ψs are dissociated in the plasma owing to color screening. The end effect is a smaller number J/Ψs than expected from pure hadron–hadron multiple collisions.

also predicted to attenuate the emission of jet pairs opposite to the trigger jet.

All the above quark–gluon plasma signatures have been studied to maturity at the CERN (European Organization for Nuclear Research) – Large Hadron Collider (LHC) and Super Proton Synchrotron (SPS) and the Brookhaven National Laboratory (BNL) – Relativistic Heavy Ion Collider (RHIC) heavy ion accelerator facilities. For more details, the following references discuss: (i) direct photons [26, 27], (ii) dileptons [28], (iii) J/Ψ production [29], and (iv) quark in-medium attenuation, also known as *highp$_T$ quenching* or *jet attenuation* [30].

2.7.4
Ultraperipheral Relativistic Heavy Ion Collisions

Relativistic heavy ion collisions have also been used to study nuclear photoexcitation and photoproduction of hadrons. The strong electromagnetic fields from a heavy ion allow for the study of multiphoton excitation of nuclear targets, for example, the double-giant resonance [17, 21]. Reactions such as multiple vector meson production can be used for studies involving polarized photons [17]. The high photon energies can be used to study the gluon density in heavy nuclei [31] at low Feynman x.

Photon–photon (or two-photon) processes have long been studied at e^+e^- colliders. They are an excellent tool for many aspects of meson spectroscopy and tests of QED [32]. At hadron colliders, they are also used to study atomic physics processes, often involving electrodynamics in strong fields. One striking success was the production of antihydrogen atoms at CERN's LEAR2 [33] and at the Fermilab Tevatron [34]. At the highest energy colliders, reactions such as $\gamma\gamma \to X$ may be used to probe the quark content and spin structure of meson resonances [17]. Production of meson or baryon pairs can also probe the internal structure of hadrons. At the LHC, electroweak processes such as $\gamma\gamma \to W^+W^-$ may be probed. The cross section for two-photon processes is, in analogy to Eq. (2.61), as follows [17]:

$$\sigma = \int \mathrm{d}E_1 \mathrm{d}E_2 \frac{n_1(E_1)}{E_2} \frac{n_1(E_2)}{E_2} \sigma_{\gamma\gamma\to X}(E_1, E_2) \tag{2.72}$$

where $n_i(E_i)$ is the number of virtual photons of energy E_i generated by the ion i, and $\sigma_{\gamma\gamma\to X}(E_1, E_2)$ is the cross sections for production of particle X by two real photons with energies E_1 and E_2.

2.8
Nuclear Reactions in Stars

2.8.1
Hydrogen and CNO Cycles

Energy production in stars is a well-known process. The initial energy that ignites the process arises from the gravitational contraction of a mass of gas. The contraction increases the pressure, temperature, and density, at the center of the star until the values are reached at which the thermonuclear reactions start, initiating the star's lifetime. The energy liberated in these reactions yield a pressure in the plasma, which opposes compression due to gravitation. Thus, an equilibrium is reached for the energy which is produced, the energy which is liberated by radiation, the temperature, and the pressure.

The Sun is a star in its initial phase of evolution. The temperature at its surface is 6000 °C, while in its interior the temperature reaches 1.5×10^7K, with a pressure given by 6×10^{11} atm and density 150g cm^{-3}. The present mass of the Sun is $M_\odot = 2 \times 10^{33}$g and its main composition is hydrogen (70%), helium (29%), and less than 1% of heavier elements, such as carbon, oxygen, and so on.

What are the nuclear processes that originate the huge thermonuclear energy of the Sun, and that have lasted 4.6×10^9 years (the assumed age of the Sun)? It cannot be the simple fusion of two protons, or of α-particles, or even the fusion of protons with α-particles, since neither $^2_2$He,$^8_4$Be, or $^5_3$Li, are stable. The only possibility is the proton–proton fusion in the form

$$\mathrm{p} + \mathrm{p} \to \mathrm{d} + \mathrm{e}^+ + \nu_e \tag{2.73}$$

which occurs via β-decay, that is, due to the weak interaction. The cross section for this reaction for protons of energy around 1 MeV is very small, of the order of 10^{-23} b. The average lifetime of protons in the Sun due to the transformation to deuterons by means of Eq. (2.73) is about 10^{10}y. This explains why the energy radiated from the Sun is approximately constant in time, and not an explosive process.

The deuteron produced in the above reaction is consumed almost immediately in the process

$$\mathrm{d} + \mathrm{p} \to {}^3_2\mathrm{He} + \gamma \tag{2.74}$$

The resulting $^3_2$He reacts by means of

$${}^3_2\mathrm{He} + {}^3_2\mathrm{He} \to {}^4_2\mathrm{He} + 2\mathrm{p} \tag{2.75}$$

which produces the stable nucleus $^4_2$He with a great energy gain, or by means of the reaction

$${}^3_2\mathrm{He} + {}^4_2\mathrm{He} \to {}^7_4\mathrm{Be} + \gamma \tag{2.76}$$

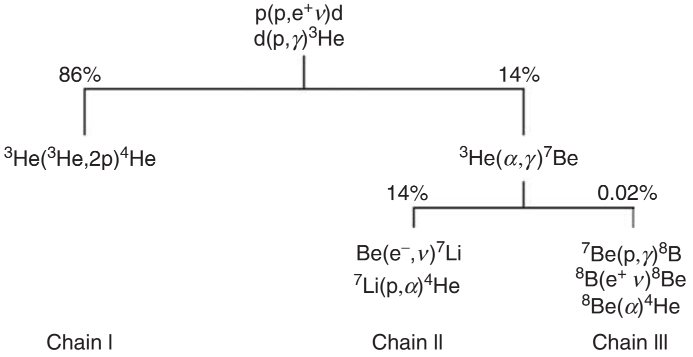

Figure 2.18 The p–p chain reaction (p–p cycle). The percentage for the several branches are calculated in the center of the Sun [35].

In the second case, a chain reaction follows as

$$ {}^7_4\mathrm{Be} + \mathrm{e}^- \to {}^7_3\mathrm{Li} + \nu_e, \quad {}^7_3\mathrm{Li} + \mathrm{p} \to 2({}^4_2\mathrm{He}) \tag{2.77}$$

or

$$\begin{gathered} {}^7_4\mathrm{Be} + \mathrm{p} \to {}^8_5\mathrm{B} + \gamma, \\ {}^8_5\mathrm{B} \to 2({}^4_2\mathrm{He}) + \mathrm{e}^+ + \nu_e \end{gathered} \tag{2.78}$$

The chain reaction (2.73)–(2.78) is called the *hydrogen cycle*. The result of this cycle is the transformation of four protons into an α-particle, with an energy gain of 26.7 MeV, about 20% of which are carried away by the neutrinos (Figure 2.18).

If the star contains heavier elements, another cycle can occur – the *carbon cycle* or *CNO cycle* [36]. In this cycle, the carbon, oxygen, and nitrogen nuclei are catalyzers of nuclear processes, with the end product also in the form $4\mathrm{p} \to {}^4_2$ He. Figure 2.19 describes the CNO cycle. Owing to the larger Coulomb repulsion between the carbon nuclei, it occurs at higher temperatures (larger relative energy between the participant nuclei), up to 1.4×10^7K. In the Sun, the hydrogen cycle prevails. However, in stars with larger temperatures, the CNO cycle is more effective.

After the protons are transformed into helium at the center of a star like our Sun, the fusion reactions start to consume protons at the surface of the star. At this stage, the star starts to become a *red giant*. The energy generated by fusion increases the temperature and expands the surface of the star. The star's luminosity increases. The red giant contracts again after the hydrogen fuel is burned.

Other thermonuclear processes start. The first is the helium burning when the temperature reaches 10^8K and the density becomes 10^6g cm^{-3}. Helium burning starts with the triple-capture reaction

$$3({}^4_2\mathrm{He}) \to {}^{12}_6\mathrm{C} + 7.65\mathrm{MeV} \tag{2.79}$$

followed by the formation of oxygen via the reaction

$${}^{12}_6\mathrm{C} + {}^4_2\mathrm{He} \to {}^{16}_8\mathrm{O} + \gamma \tag{2.80}$$

For a star with the mass of the Sun, helium burning occurs in about 10^7y. For a much heavier star, the temperature can reach 10^9K. The compression process followed by the burning of heavier elements

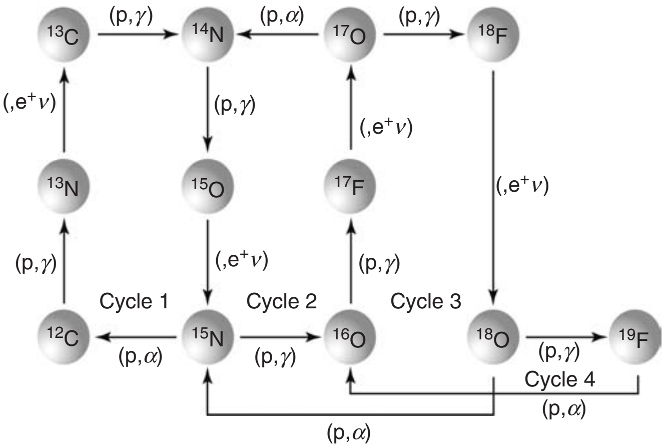

Figure 2.19 The CNO cycle. (Courtesy of Frank Timmes.)

can lead to the formation of iron and nickel. After that, the thermonuclear reactions are no more energetic and the star stops producing nuclear energy.

2.8.2
Synthesis of Heavier Elements

In Figure 2.20, we show the relative distribution of elements in our galaxy. It has two distinct regions: in the region $A < 100$ it decreases with A approximately like an exponential, whereas for $A > 100$ it is approximately constant, except for the peaks in the region of the magic numbers $Z = 50$ and $N = 50, 82, 126$.

The thermonuclear processes (2.73)–(2.80) can explain the relative abundance of $^4_2$He, $^{12}_6$C, and $^{16}_8$O. The processes occurring after $^4_2$He burning mainly form isotopes of $^{20}_{10}$Ne, $^{24}_{12}$Mg, and $^{28}_{14}$Si. We can understand the small abundance of the elements Li, Be, and B as due to the small velocity with which they are formed via the reaction (2.76) and the first equation of (2.77), while they are rapidly consumed by the second reaction in (2.77) and the first reaction in (2.78).

The synthesis of elements heavier than oxygen occur when, after helium burning, a new compression and heating of the star raises the temperature to values higher than 6×10^8K. This situation ignites an intense carbon burning:

$$\begin{aligned} {}^{12}_{6}\mathrm{C} + {}^{12}_{6}\mathrm{C} &\rightarrow {}^{20}_{10}\mathrm{Ne} + {}^{4}_{2}\mathrm{He} \rightarrow {}^{23}_{11}\mathrm{Na} + \mathrm{p} \\ &\rightarrow {}^{23}_{12}\mathrm{Mg} + \mathrm{n} \rightarrow {}^{24}_{12}\mathrm{Mg} + \gamma \end{aligned} \tag{2.81}$$

Carbon and oxygen can also burn simultaneously:

$$\begin{aligned} {}^{12}_{6}\mathrm{C} + {}^{16}_{8}\mathrm{O} &\rightarrow {}^{24}_{12}\mathrm{Mg} + {}^{4}_{2}\mathrm{He}, \text{etc.}, \\ {}^{16}_{8}\mathrm{O} + {}^{16}_{8}\mathrm{O} &\rightarrow {}^{28}_{14}\mathrm{Si} + {}^{4}_{2}\mathrm{He}, \text{etc.} \end{aligned} \tag{2.82}$$

For temperatures above 3×10^9K, more photonuclear processes appear. These yield more nuclei to be burned and heavier nuclei are produced:

$$\begin{aligned} \gamma + {}^{28}_{14}\mathrm{Si} &\rightarrow {}^{24}_{12}\mathrm{Mg} + {}^{4}_{2}\mathrm{He}, \\ {}^{4}_{2}\mathrm{He} + {}^{28}_{14}\mathrm{O} &\rightarrow {}^{32}_{16}\mathrm{S} + \gamma, \text{etc.} \end{aligned} \tag{2.83}$$

Owing to the large number of free neutrons, many (n, γ)-reactions (radiative neutron capture) elements in the mass range $A = 28, \ldots, 57$ are formed. This

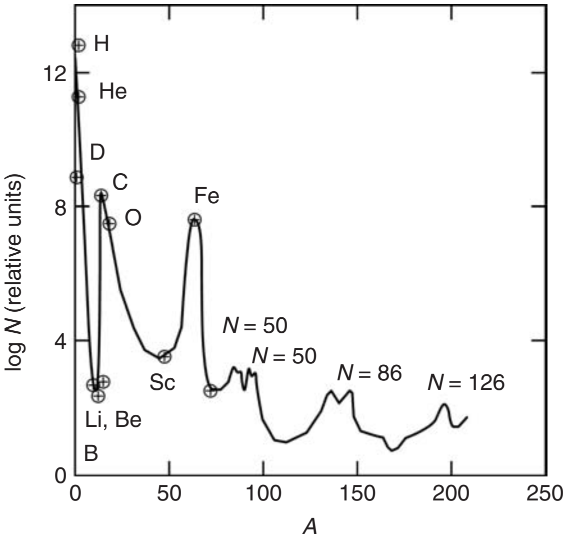

Figure 2.20 Relative distribution of elements in our galaxy.

leads to a large abundance of elements in the iron mass region, which have the largest binding energy per nucleon. For elements heavier than iron, the nuclear fusion processes do not generate energy.

For $A > 100$, the distribution of nuclei cannot be explained in terms of fusion reactions with charged particles. They are formed by the successive capture of slow neutrons and of β^--decay. The maxima of the element distribution in $N = 50, 82, 126$ are due to the small capture cross sections corresponding to the magic numbers. This yields a slew of isotopes at the observed element distribution.

2.8.3
Thermonuclear Cross Sections

The nuclear cross section for a reaction between target j and projectile k is defined by

$$\sigma = \frac{\text{number of reactions (target}^{-1}s^{-1})}{\text{flux of incoming projectiles}} = \frac{r/n_j}{n_k v} \tag{2.84}$$

where the target number density is given by n_j, the projectile number density is given by n_k , and v is the relative velocity between target and projectile nuclei. Then r, the number of reactions per cubic centimeter per second, can be expressed as $r = \sigma v n_j n_k$, or, more generally,

$$r_{j,k} = \int \sigma |v_j - v_k| \mathrm{d}^3 n_j \mathrm{d}^3 n_k \tag{2.85}$$

The evaluation of this integral depends on the type of particles and distributions that are involved. For nuclei j and k in an astrophysical plasma, obeying an MB distribution,

$$\mathrm{d}^3 n_j = n_j \left(\frac{m_j}{2\pi kT}\right)^{\frac{3}{2}} \exp\left(-\frac{m_j v_j^2}{2kT}\right) \mathrm{d}^3 v_j \tag{2.86}$$

Eq. (2.85) simplifies to $r_{j,k} =< \sigma v > n_j n_k$, where $< \sigma v >$ is the average of σv over the temperature distribution in Eq. (2.86).

In astrophysical plasmas with high densities and/or low temperatures, effects of electron screening become highly important. This means that the reacting nuclei, due to the background of electrons and nuclei, feel a different Coulomb repulsion than in the case of bare nuclei. Under most conditions (with nonvanishing temperatures), the generalized reaction rate integral can be separated into the traditional expression without screening and a screening factor

$$< \sigma v >^*_{j,k} = f_{\text{scr}}(Z_j, Z_k, \rho, T, Y_i) < \sigma v >_{j,k} \tag{2.87}$$

This screening factor is dependent on the charge of the involved particles, the density, temperature, and the composition of the plasma. Here Y_i denotes the abundance of nucleus i defined by $Y_i = n_i/(\rho N_A)$, where n_i is the number density of nuclei per unit volume and N_A Avogadro's number. At high densities and low temperatures, screening factors can enhance reactions by many orders of magnitude and lead to pycnonuclear ignition.

When in Eq. (2.85), particle k is a photon, the relative velocity is always c and quantities in the integral are not dependent on $d^3 n_j$. Thus, it simplifies to $r_j = \lambda_{j,\gamma} n_j$ and $\lambda_{j,\gamma}$ results from an integration of the photodisintegration cross section over a Planck distribution for photons of temperature T

$$r_j = \frac{1}{\pi^2 (c\hbar)^3} \int d^3 n_j \times \int_0^\infty \frac{c\sigma(E_\gamma) E_\gamma^2}{\exp(E_\gamma/kT) - 1} dE_\gamma \tag{2.88}$$

There is, however, no direct need to evaluate photodisintegration cross sections, because, due to detailed balance, they can be expressed by the capture cross sections for the inverse reaction $l + m \to j + \gamma$.

A procedure similar to Eq. (2.88) is used for electron captures by nuclei. Because the electron is about 2000 times less massive than a nucleon, the velocity of the nucleus j is negligible in the CM system in comparison to the electron velocity ($|v_j - v_e| \approx |v_e|$). The electron capture cross section has to be integrated over a Boltzmann, partially degenerate, or Fermi distribution of electrons, depending on the astrophysical conditions. The electron capture rates are a function of T and $n_e = Y_e \rho N_A$, the electron number density [37]. In a neutral, completely ionized plasma, the electron abundance is equal to the total proton abundance in nuclei $Y_e = \sum_i Z_i Y_i$ and $r_j = \lambda_{j,e}(T, \rho Y_e) n_j$.

This treatment can be generalized for the capture of positrons, which are in a thermal equilibrium with photons, electrons, and nuclei. At high densities ($\rho > 10^{12}$ g cm^{-3}), the size of the neutrino scattering cross section on nuclei and electrons ensures that enough scattering events occur to thermalize a neutrino distribution. The inverse process to electron capture (neutrino capture) can occur even then and the neutrino capture rate can be expressed similarly to Eq. (2.88), integrating over the neutrino distribution. Also inelastic neutrino scattering on nuclei can be expressed in this form. Finally, for normal decays, like β or α decays with half-life $\tau_{1/2}$, we obtain an equation similar to Eq. (2.88) or r_j of the previous paragraph with a decay constant $\lambda_j = \ln 2/\tau_{1/2}$ and $r_j = \lambda_j n_j$.

2.9
Reactions with Radioactive Nuclear Beams

The basic research activity in nuclear physics, driven by the desire to understand the forces that dictate the properties of nuclei, has spawned a large number of beneficial applications. Among its many progeny, we can count reactor- and spallation-based neutron sources, synchrotron radiation sources, particle physics, materials modification by implantation, carbon dating, and much more. It is an excellent example of the return to society of investment in basic research.

All these achievements have been realized by accelerating the 283 stable or long-lived nuclear species we find here on Earth. In recent years, however, it has become evident that it is now technically possible to create and accelerate unstable nuclei and that there are some 6000–7000 distinct nuclear species which live long enough to be candidates for acceleration. They are the nuclei within the so-called *drip-lines*, the point where the nucleus can no longer hold another particle. This has led to many new opportunities in industry, medicine, materials studies, and the environment.

Assume that a highly energetic uranium projectile ($N/Z \sim 1.6$) hits a target nucleus in an almost central collision, as shown in Figure 2.21. A part of the projectile (*participant*) is scrapped off and forms a highly excited mixture of nucleons with a part of the target. A piece of the projectile (*spectator*) flies away with nearly the same velocity as that of the beam. The neutron-to-proton ratio of the spectator is nearly equal to that of the projectile. Since the N/Z ratio of light nuclei (stable) is close to one, the fragment is far from the stability line. Statistically, a large number of fragments with different N/Z ratios are created and several new exotic nuclei have been discovered in this way. These nuclei can be collected in a secondary beam and further accelerated; they can induce reactions with a target nucleus. This method has become an important tool to study the properties of short-lived isotopes.

Experiments with secondary beams are limited by reaction cross section and *luminosity*. The luminosity L is defined as the product of beam intensity i and target thickness $t, L = i \times t$. The reaction rate N is the product of luminosity and reaction cross section $\sigma_r, N = \sigma_r \times L$. In most of the reactions, the usable target thickness is limited by the width of the excitation function (i.e., the cross section as a function of the excitation energy). Production reactions with a wide excitation function covering a broad energy range can profit in luminosity by the use of thick targets.

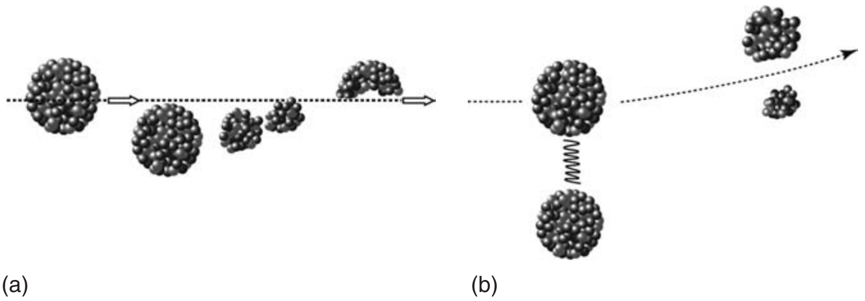
(a) (b)

Figure 2.21 (a) Schematic description of a nuclear fragmentation reaction producing rare isotopes. The lower fragments are called *participants*, while the upper one is called the *spectator*. Using uranium projectiles ($N/Z \sim 1.6$), one expects to produce (light) spectator nuclei of about the same N/Z ratio. (b) Coulomb fission of relativistic projectiles leading to the production of rare isotopes. For a heavy unstable projectile, an exchanged photon with the target can give it enough energy to fragment into several types of isotopes.

The condition for fragmentation of heavy ions is that the projectile should move faster than nucleons move inside the nucleus. The projectile energy should be sufficiently above the Fermi domain, for example, above 100 A MeV. The usable target thickness for these high energies is of the order of several grams per square centimeter, corresponding to 10^{23} atoms cm^{-2}. The excitation function for complete fusion of heavy ions, however, has a width of only 10 MeV. This corresponds to a usable target thickness of the order of 1 mg cm^{-2} or 10^{18}atoms cm^{-2}. Consequently, beam intensities for the investigation of complete fusion reactions must be by 4–5 orders higher to achieve the same luminosity as for fragmentation.

Figure 2.22 shows an example of the production cross sections for tin isotopes from complete fusion (dotted line), nuclear fragmentation (solid line), and Coulomb fission of ^{238}U (dashed line). The symbols represent experimental data. The fragmentation cross sections (solid line) have been calculated with a semiempirical code [Sue00].

The first experiments with unstable nuclear beams were designed to measure the nuclear sizes, namely, the matter distribution of protons and neutrons. For stable nuclei, such experiments are best accomplished with electron beams, which probe the nuclear charge (proton) distribution. Electron scattering experiments with unstable beams can only be performed in an electron–nucleus collider. However, the easiest solution is to measure the *reaction cross section* in collisions of unstable beams with a fixed target nucleus [38].

The reaction cross section in high-energy collisions is given by

$$\sigma_R = 2\pi \int [1 - T(b)] b db \qquad (2.89)$$

where

$$T(b) = \exp\left\{-\sigma_{\mathrm{NN}} \int_{-\infty}^{\infty} \mathrm{d}z \times \int \rho_P(\mathbf{r})\rho_T(\mathbf{R}+\mathbf{r})\,\mathrm{d}^3r\right\} \qquad (2.90)$$

with $\mathbf{R} = (\mathbf{b}, z)$. σ_{NN} is the nucleon–nucleon cross section at the corresponding bombarding energy,

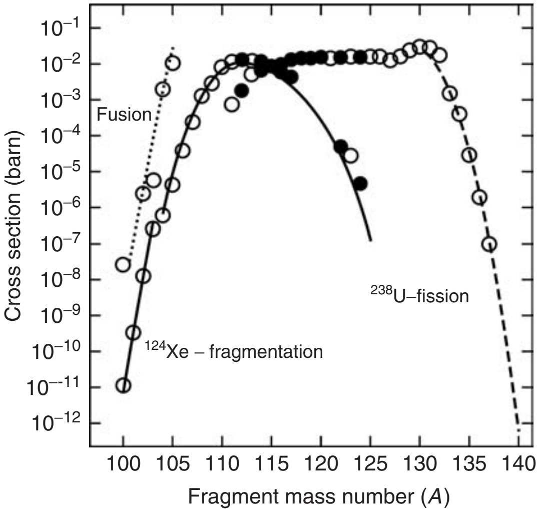

Figure 2.22 Production cross sections for the tin isotopes from complete fusion (dotted line), fragmentation (solid line), and projectile fission of ^{238}U (dashed line). The symbols represent experimental data. The fragmentation cross sections (solid line) have been calculated with a semiempirical code.

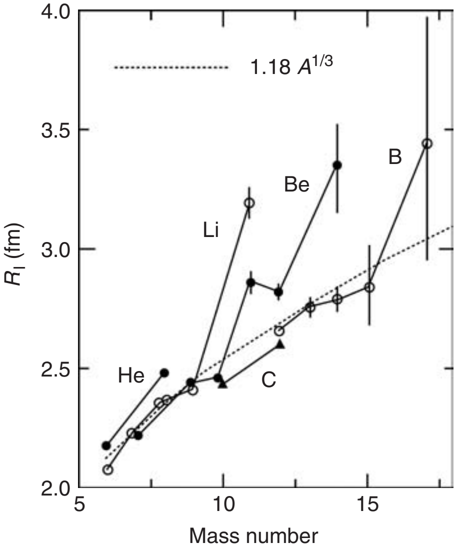

Figure 2.23 The matter density radii of several light nuclei compared to the trend $R \sim 1.18\ A^{1/3}$ fm (dashed line) for normal nuclei. The solid lines are guides to the eyes.

and $\rho_{P(T)}$ is the projectile (target) matter density distribution. $T(b)$ is known as the *transparency function* or *survival probability*. The factor $1 - T(b)$ is the probability that a reaction occurs for a collision with impact parameter b. The factor $\left[\sigma_{\mathrm{NN}} \int \rho_P \rho_T \mathrm{d}^3 r\right]^{-1}$ is interpreted as the mean free path for a nucleon–nucleon collision. The reaction cross sections can be calculated using the matter distributions of the target in Eq. (2.16).

Experimentally, a significant increase in the root mean square radii for the neutron-rich isotopes ^{6}He,^{8}He, and ^{11}Li (Figure 2.23) was observed. Thus, the addition of the neutrons to ^{4}He and ^{9}Li nuclei increase their radii considerably. This can be understood in terms of the binding energy of the outer nucleons. The large matter radii of these nuclei have lead the experimentalists to call them "*halo nuclei*." The binding energy of the last two neutrons in ^{11}Li is about 300 keV. In ^{6}He, it is 0.97 MeV. These are very small values and should be compared with $S_n =$ 6–8 MeV, which is the average binding of nucleons in stable nuclei. Abnormally large radii were also found for other light neutron-rich nuclei.

The matter density radii of these nuclei do not follow the observed trend $R \sim 1.18\ A^{1/3}$ fm of normal nuclei. Thus, the halo seems to be a common feature of loosely bound neutron-rich nuclei.

Several other methods have been devised to probe the structure of nuclei far from the valley of stability. Among these are (i) Coulomb dissociation [39], (ii) Trojan horse method [40], (iii) asymptotic normalization coefficients [41], (iv) heavy ion charge exchange [42], (v) knockout [43], and (vi) fusion reactions [44]. These methods yield different insights into the structure of exotic nuclei and comprise most of the experiments in radioactive beam facilities [45].

Reactions producing rare nuclear isotopes have opened a new research front in nuclear physics with applications in many areas of science: the rapid production of short-lived nuclei, many of which are poorly known, is important for astrophysics (r-process) and cosmology; nuclear medicine benefits from studies of new nuclear isotopes; and the list goes on.

However, the basic question still remains: what combinations of neutrons and protons can make up a nucleus? The experimental detection of new nuclear isotopes is an ongoing research that will ultimately lead to new insight and development of nuclear science, with enormous profit for mankind [46].

Acknowledgments

The author is grateful to Prof. Reinhard Stock and Prof. Konrad Gelbke for their support and encouragement. This work was supported in part by funds provided by the US Department of Energy (DOE) under contract No. DE-FG02-08ER41533.

References

1. Rutherford, E. (1911) *Philos. Mag.*, **21**, 669
2. Ridley, B.W. and Turner, J.F. (1964) *Nucl. Phys.*, **58**, 497.
3. Bohr, N. (1936) *Nature*, **137**, 344
4. Ghoshal, S.N. (1950) *Phys. Rev.*, **80**, 939
5. Facchini, U. and Saetta-Menichella, E. (1968) *Energ. Nucl.*, **15**, 54
6. Feshbach, H. (1992) *Theoretical Nuclear Physics - Nuclear Reactions*, John Wiley & Sons, Inc., New York.
7. Tveter, A. (1972) *Nucl. Phys.*, **A185**, 433
8. Marshak, H., Langsford, A., Wong, C.Y., and Tamura, T. (1968) *Phys. Rev. Lett.*, **20**, 554
9. Weisskopf, V.F. and Ewing, D.H. (1940) *Phys. Rev.*, **54**, 472
10. Hauser, W. and Feshbach, H. (1952) *Phys. Rev.*, **87**, 366
11. McVoy, K.W. (1967) *Ann. Phys.*, **43**, 91
12. Feshbach, H., Porter, C.E., and Weisskopf, V.F. (1953) *Phys. Rev.*, **90**, 166
13. Perey, F.G. (1963) *Phys. Rev.*, **131**, 745
14. Miura, K. *et al.* (1987) *Nucl Phys.*, **A467**, 79
15. McMillan, E. and Abelson, P.H. (1940) *Phys. Rev.*, **57**, 1185
16. Oganessian, Y.T. *et al.* (2006) *Phys. Rev.*, **C 74**, 044602
17. Bertulani, C.A. and Baur, G. (1986) *Phys. Lett.*, **B174**, 23; (1988) *Phys. Rep.*, **163**, 299
18. Leprêtre, A. *et al.* (1974) *Nucl. Phys.*, **A219**, 39
19. Axel, P. (1962) *Phys. Rev.*, **126**, 671
20. Mordechai, S. *et al.* (1988) *Phys. Rev. Lett.*, **61**, 531.
21. Schmidt, R. *et al.* (1993) *Phys. Rev. Lett.*, **70**, 1767.
22. Frankfurt, L., Strikman, M., and Weiss, C. (2005) *Ann. Rev. Nucl. Part. Sci.*, **55**, 403.
23. Danielewicz, P. (1995) *Phys. Rev.*, **C51**, 716.
24. Wong, C.Y. (1994) *Introduction to High Energy Heavy Ion Collisions*, World Scientific, Singapore.
25. Hwa, R.C. and Wang, X.N. (eds) (2004) *Quark-Gluon Plasma 3*, World Scientific, Singapore.
26. Aggarwal, M.M. *et al.* (2000) *Phys. Rev. Lett.*, **85**, 3595.
27. Bathe, S. *et al.* (2006) *Nucl. Phys.*, **774**, 103.
28. Damjanovic, S. *et al.* (2006) *Nucl. Phys.*, **A774**, 715.
29. Adare, A. *et al.* (2007) *Phys. Rev. Lett.*, **98**, 232302.
30. Buesching, H. *et al.* (2006) *Nucl. Phys.*, **774**, 103.
31. (a) Gonçalves, V.P. and Bertulani, C.A. (2002) *Phys. Rev. C*, **65**, 054905; (b) Adeluyi, A. and Bertulani, C.A. (2012) *Phys. Rev.*, **C 85**, 044904.
32. Bertulani, C.A., Klein, S., and Nystrand, J. (2005) *Ann. Rev. Nucl. Part. Sci.*, **55**, 271.
33. Baur, G. *et al.* (1996) *Phys. Lett.*, **B 368**, 251.
34. Blanford, G. *et al.* (1998) *Phys. Rev. Lett.*, **80**, 3040.
35. Bahcall, J.N. (1989) *Neutrino Astrophysics*, Cambridge University Press, Cambridge.
36. Bethe, H.A. (1939) *Phys. Rev.*, **55**, 434.
37. Fuller, G.M., Fowler, W.A., and Newman, M. (1985) *Ap. J.*, **293**, 1.
38. Tanihata, I. *et al.* (1985) *Phys. Rev. Lett.*, **55**, 2676.
39. Baur, G., Bertulani, C.A., and Rebel, H. (1986) *Nucl. Phys.*, **A458**, 188.
40. Baur, G. (1986) *Phys. Lett.*, **B178**, 135.
41. Xu, H.M., Gagliardi, C.A., Tribble, R.E., Mukhamedzhanov, A.M., and Timofeyuk, N.K. (1994) *Phys. Rev. Lett.*, **73**, 2027.
42. Steiner, M. *et al.* (1996) *Phys. Rev. Lett.*, **76**, 26.
43. Hansen, P.G. and Tostevin, J.A. (2003) *Annu. Rev. Nucl. Part. Sci.*, **53**, 219.
44. Canto, L.F., Gomes, P.R.S., Donangelo, R., and Hussein, M.S. (2006) *Phys. Rep.*, **424**, 1.

45. Geesaman, D.F., Gelbke, C.K., Sherrill, B.M., and Janssens, R.V.F. (2006) *Annu. Rev. Nucl. Part. Sci.*, **56**, 53.
46. Baumann, T. *et al.* (2007) *Nature*, **449**, 1022.

Further Readings

Bertulani, C.A. and Danielewicz, P. (2004) *Introduction to Nuclear Reactions*, IOP Publishing, London.
Bertulani, C.A. (2007) *Nuclear Physics in a Nutshell*, Princeton University Press, Princeton, NJ.
Danielewicz, P., Lacey, R., and Lynch, W.G. (2002) Determination of the equation of state of dense matter. *Science*, **298**, 1592.
Feshbach, H. (1993) *Theoretical Nuclear Physics, Nuclear Reactions*, Wiley-Interscience, New York.
Fröbrich, P. and Lipperheide, R. (1996) *Theory of Nuclear Reactions*, Oxford University Press, Oxford.
Iliadis, C. (2007) *Nuclear Physics of Stars*, John Wiley & Sons, Inc., New York.
Lilley, J.S. (2001) *Nuclear Physics: Principles and Applications*, John Wiley & Sons, Inc., New York.
Glendenning, N.K. (2004) *Direct Nuclear Reactions*, World Scientific, Singapore.
Wong, S.S.M. (2004) *Nuclear Physics*, John Wiley & Sons, Inc., New York.
Wong, C.Y. (1994) *Introduction to High Energy Heavy Ion Collisions*, World Scientific, Singapore.

1

Nuclear Structure

Jan Jolie

1.1 Introduction 5
1.2 General Nuclear Properties 6
1.2.1 Properties of Stable Nuclei 6
1.2.2 Properties of Radioactive Nuclei 7
1.3 Nuclear Binding Energies and the Semiempirical Mass Formula 8
1.3.1 Nuclear Binding Energies 8
1.3.2 The Semiempirical Mass Formula 10
1.4 Nuclear Charge and Mass Distributions 12
1.4.1 General Comments 12
1.4.2 Nuclear Charge Distributions from Electron Scattering 13
1.4.3 Nuclear Charge Distributions from Atomic Transitions 14
1.4.4 Nuclear Mass Distributions 15
1.5 Electromagnetic Transitions and Static Moments 16
1.5.1 General Comments 16
1.5.2 Electromagnetic Transitions and Selection Rules 17
1.5.3 Static Moments 19
1.5.3.1 Magnetic Dipole Moments 19
1.5.3.2 Electric Quadrupole Moments 21
1.6 Excited States and Level Structures 22
1.6.1 The First Excited State in Even–Even Nuclei 22
1.6.2 Regions of Different Level Structures 23
1.6.3 Shell Structures 23
1.6.4 Collective Structures 25
1.6.4.1 Vibrational Levels 25
1.6.4.2 Rotational Levels 26
1.6.5 Odd-A Nuclei 28
1.6.5.1 Single-Particle Levels 28
1.6.5.2 Vibrational Levels 28

Encyclopedia of Nuclear Physics and its Applications, First Edition. Edited by Reinhard Stock.
© 2013 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2013 by Wiley-VCH Verlag GmbH & Co. KGaA.

1.6.5.3 Rotational Levels 28
1.6.6 Odd–Odd Nuclei 28
1.7 Nuclear Models 29
1.7.1 Introduction 29
1.7.2 The Spherical-Shell Model 30
1.7.3 The Deformed Shell Model 32
1.7.4 Collective Models of Even–Even Nuclei 33
1.7.5 Boson Models 35
Glossary 40
References 41
Further Readings 42

1.1
Introduction

The study of nuclear structure today encompasses a vast territory from the study of simple, few-particle systems to systems with close to 300 particles, from stable nuclei to the short-lived exotic nuclei, from ground-state properties to excitations of such energy that the nucleus disintegrates into substructures and individual constituents, from the strong force that hold the atomic nucleus together to the effective interactions that describe the collective behavior observed in many heavy nuclei.

After the discovery of different kinds of radioactive decays, the discovery of the structure of the atomic nucleus begins with the fundamental paper by Ernest Rutherford [1], in which he explained the large-angle alpha (α)-particle scattering from gold that had been discovered earlier by Hans W. Geiger and Ernest Marsden. Indeed, Rutherford shows that the atom holds in its center a very tiny, positively charged nucleus that contains 99.98% of the atomic mass. In 1914, Henry Moseley [2, 3] showed that the nuclear charge number Z equaled the atomic number. Using the first mass separators, Soddy [4] was able to show that one chemical element could contain atomic nuclei with different masses, forming different isotopes. With the availability of α-sources, due to the works of the Curies in Paris, Rutherford [5] was able to perform the first nuclear reactions on nitrogen. The first attempt at understanding the relative stability of nuclear systems was made by Harkins and Majorsky [6]. This model, like many others of the time, consisted of protons and electrons. In 1924, Wolfgang Pauli [7] suggested that the optical hyperfine structure might be explained if the nucleus had a magnetic dipole moment, while later Giulio Racah [8] investigated the effect on the hyperfine structure if the nuclear charge were not spherically symmetric – that is, if it had an electric quadrupole moment.

All of these structure suggestions occurred before James Chadwick [9] discovered the neutron, which not only explained certain difficulties of previous models (e.g., the problems of the confinement of the electron or the spins of light nuclei), but opened the way to a very rapid expansion of our knowledge of the structure of the nucleus. Shortly after the discovery of the neutron, Heisenberg [10] proposed that the proton and neutron are two states of the nucleon classified by a new spin quantum number, the isospin. It may be difficult to believe today, 60 years after Chadwick's discovery, just how rapidly our knowledge of the nucleus increased in the mid-1930s.

Encyclopedia of Nuclear Physics and its Applications, First Edition. Edited by Reinhard Stock.
© 2013 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2013 by Wiley-VCH Verlag GmbH & Co. KGaA.

Hans A. Bethe's review articles [11, 12], one of the earliest and certainly the best known, discuss many of the areas that not only form the basis of our current knowledge but that are still being investigated, albeit with much more sophisticated methods.

The organization here will begin with general nuclear properties, such as size, charge, and mass for the stable nuclei, as well as half-lives and decay modes (α, β, γ, and fission) for unstable systems. Binding energies and the mass defect lead to a discussion of the stability of systems and the possibility of nuclear fusion and fission. Then follow details of the charge and current distributions, which, in turn, lead to an understanding of static electromagnetic moments (magnetic dipole and octupole, electric quadrupole, etc.) and transitions. Next follows the discussion of single-particle and collective levels for the three classes of nuclei: even–even, odd-A, and odd–odd (i.e., odd Z and odd N). With these mainly experimental details in hand, a discussion of various major nuclear models follows. These discussions attempt, in their own way, to categorize and explain the mass of experimental data.

1.2
General Nuclear Properties

1.2.1
Properties of Stable Nuclei

The discovery of the neutron allowed each nucleus to be assigned a number, A, the mass number, which is the sum of the number of protons (Z) and neutrons (N) in the particular nucleus. The atomic number of chemistry is identical to the proton number Z. The mass number A is the integer closest to the ratio between the mass of a nucleus and the fundamental mass unit. This mass unit, the unified atomic mass unit, has the value $1\ \mathrm{u} = 1.660538921(73) \times 10^{-27}\ \mathrm{kg} = 931.494061(21)\ \mathrm{MeV\,c^{-2}}$. It has been picked so that the atomic mass of a $^{12}\mathrm{C}_6$ atom is exactly equal to 12 u. The notation here is $^{A}\mathrm{X}_{N}$, where X is the chemical symbol for the given element, which fixes the number of electrons and hence the number of protons Z. This commonly used notation contains some redundancy because $A = Z + N$ but avoids the need for one to look up the Z-value for each chemical element. From this last expression, one can see that there may be several combinations of Z and N to yield the same A. These nuclides are called *isobars*. An example might be the pair $^{196}\mathrm{Pt}_{118}$ and $^{196}\mathrm{Au}_{117}$. Furthermore, an examination of a table of nuclides shows many examples of nuclei with the same Z-value but different A- and N-values. Such nuclei are said to be *isotopes* of the element. For example, oxygen (O) has three stable isotopes: $^{16}\mathrm{O}_8$, $^{17}\mathrm{O}_9$, and $^{18}\mathrm{O}_{10}$. A group of nuclei that have the same number of neutrons, N, but different numbers of protons, Z (and, of course, A), are called *isotones*. An example might be $^{38}\mathrm{Ar}_{20}$, $^{39}\mathrm{K}_{20}$, and $^{40}\mathrm{Ca}_{20}$. Some elements have but one stable isotope (e.g., $^{9}\mathrm{Be}_5$, $^{19}\mathrm{F}_{10}$, and $^{197}\mathrm{Au}_{118}$), others, two, three, or more. Tin ($Z = 50$) has the most at 10. Finally, the element technetium has no stable isotope at all. A final definition of use for light nuclei is a *mirror pair*, which is a pair of nuclei with N and Z interchanged. An example of such a pair would be $^{23}\mathrm{Na}_{12}$ and $^{23}\mathrm{Mg}_{11}$.

The nuclear masses of stable isotopes are determined with a mass spectrometer, and we shall return to this fundamental property when we discuss the nuclear binding energy and the mass defect in Section 1.3. After mass, the next property of interest is the size of a nucleus. The simplest assumption here is that the mass and charge form a uniform sphere whose size is determined by the radius. While not all nuclei are spherical or of uniform density, the assumption of a uniform mass/charge density and spherical shape is an adequate starting assumption (more complicated charge distributions are discussed in Section 1.4 and beyond). The nuclear radius and, therefore, the nuclear volume or size is usually determined by electron-scattering experiments; the radius is given by the relation

$$R = r_0 A^{1/3} \tag{1.1}$$

which, with $r_0 = 1.25$ fm, gives an adequate fit over the entire range of nuclei near stability. An expression such as Eq. (1.1) implies that nuclei have a density independent of A, that is, they are incompressible. A somewhat better fit to the nuclear sizes can be obtained from the Coulomb energy difference of mirror nuclei, which covers but a fifth of the total range of A. This yields $r_0 = 1.22$ fm. Even if the charge and/or mass distribution is neither spherical nor uniform, one can still define an equivalent radius as a size parameter.

Two important properties of a nuclide are the spin J and the parity π, often expressed jointly as J^π, of its ground state. These are usually listed in a table of isotopes and give important information about the structure of the nuclide of interest. An examination of such a table will show that the ground state and parity of all even–even nuclei is 0^+. The spin and parity assignments of the odd-A and odd–odd nuclei tell a great deal about the nature of the principal parts of their ground-state wave functions. A final property of a given element is the relative abundance of its stable isotopes. These are determined again with a mass spectrograph and listed in various tables of the nuclides.

1.2.2
Properties of Radioactive Nuclei

A nucleus that is unstable, that is, it can decay to a different or *daughter* nucleus, is characterized not only by its mass, size, spin, and parity but also by its lifetime τ and decay mode or modes. (In fact, each level of a nucleus is characterized by its spin, parity, lifetime, and decay modes.) The law of radioactive decay is simply

$$N(t) = N(0)\mathrm{e}^{-\lambda t} = N(0)\mathrm{e}^{-t/\tau} \tag{1.2}$$

where $N(0)$ is the number of nuclei initially present, λ is the decay constant, and its reciprocal τ is the lifetime. Instead of the lifetime, often the half-life $T_{1/2}$ is used. It is the time in which half of the nuclei decay. By setting $N(T_{1/2}) = N(0)/2$ in Eq. (1.2), one obtains the relation

$$T_{1/2} = \ln(2)\tau = 0.693\tau \tag{1.3}$$

The decay mode of ground states can be α, β, or spontaneous fission. Excited states mostly decay by γ-emission. More exotic decays are observed in unstable nuclei far from stability where nuclei decay takes place by emission of a proton or neutron.

In α-decay, the parent nucleus emits an α-particle (a nucleus of ${}^4\mathrm{He}_2$), leaving the daughter with two fewer neutrons and protons:

$$ {}^{A}\mathrm{X}_N \rightarrow {}^{A-4}\mathrm{Y}_{N-2} + {}^4\mathrm{He}_2 \tag{1.4}$$

The α-particle has zero spin, but it can carry off angular momentum. In β-decay the weak interaction converts neutrons into protons (β^--decay) or protons into neutrons (β^+-decay). Which of the two decays takes place depends strongly on the masses of the initial and final nuclei. Because a neutron is heavier than a proton, the free neutron is unstable against β^--decay and has a lifetime of 878.5(10) s. The mass excess in β^--decay is released as kinetic energy of the final particles. In the case of the free neutron, the final particles are a proton, an electron, and an antineutrino, denoted by $\overline{\nu}$. All of these particles have spin 1/2 and can also carry off angular momentum. In the case of β^+-decay, the final particles are a bound proton, an antielectron or positron, and a neutrino. Finally, as an alternative to β^+-decay the initial nucleus can capture an inner electron. In this so-called electron capture decay, only a neutrino, ν, is emitted by the final nucleus. In general, the decays can be written as

$$\beta^-\text{-decay}: \quad {}^{A}\mathrm{X}_N \to {}^{A}\mathrm{Y}_{N-1} + \mathrm{e}^- + \overline{\nu} \tag{1.5a}$$

$$\beta^+\text{-decay}: \quad {}^{A}\mathrm{X}_N \to {}^{A}\mathrm{Y}_{N+1} + \mathrm{e}^+ + \nu \tag{1.5b}$$

$$\beta^+\text{-decay (ec)}: {}^{A}\mathrm{X}_N + \mathrm{e}^- \to {}^{A}\mathrm{Y}_{N-1} + \nu \tag{1.5c}$$

One very rare mode of decay is double β-decay, in which a nucleus is unable to β-decay to a $Z+1$ daughter for energy reasons but can emit two electrons and make a transition to a $Z+2$ daughter. An example is ${}^{82}\mathrm{Se}_{48} \to {}^{82}\mathrm{Kr}_{46}$ with a half-life of $(1.7 \pm 0.3) \times 10^{20}$ years. Double β-decay is observed under the emission of two neutrinos. Neutrinoless double β-decay is intensively searched for in ${}^{76}\mathrm{Ge}$ because it is forbidden for massless neutrinos with definite helicities. Enriched Ge is hence used as it allows the use of a large single crystal as source and detector (for a review see [13]).

In spontaneous fission, a very heavy nucleus simply breaks into two heavy pieces. For a given nuclide, the decay mode is not necessarily unique. If more than one mode occurs, then the branching ratio is also a characteristic of the radioactive nucleus in question.

An interesting example of a multimode radioactive nucleus is ${}^{242}\mathrm{Am}_{147}$. Its ground state ($J^\pi = 1^-$, $T_{1/2} = 16.01\,\mathrm{h}$) can decay either by electron capture (17.3% of the time) to ${}^{242}\mathrm{Pu}_{148}$ or by β^- decay (82.7% of the time) to ${}^{242}\mathrm{Cm}_{146}$. On the other hand, a low-lying excited state at 0.04863 MeV ($J^\pi = 5^-$, $T_{1/2} = 152$ years) can decay either by emitting a γ-ray (99.52% of the time) and going to the ground state or by emitting an α-particle (0.48% of the time) and going to ${}^{238}\mathrm{Np}_{145}$. There is an excited state at 2.3 MeV with a half-life of 14.0 ms that undergoes spontaneous fission [14]. The overall measured half-life of ${}^{242}\mathrm{Am}_{147}$ is then determined by that of the 0.04863 MeV state. Such long-lived excited states are known as *isomeric states*. From this information on branching ratios, one easily finds the several partial decay constants for ${}^{242}\mathrm{Am}_{147}$. For the ground state, $\lambda_{\mathrm{ec}} = 2.080 \times 10^{-6}\,\mathrm{s}^{-1}$ and $\lambda_{\beta-} = 9.944 \times 10^{-6}\,\mathrm{s}^{-1}$, while for the excited state at 0.04863 MeV, $\lambda_\gamma = 1.439 \times 10^{-10}\,\mathrm{s}^{-1}$ and $\lambda_\alpha = 6.639 \times 10^{-13}\,\mathrm{s}^{-1}$ and for the excited state at 2.3 MeV, $\lambda_{\mathrm{SF}} = 49.5\,\mathrm{s}^{-1}$.

1.3
Nuclear Binding Energies and the Semiempirical Mass Formula

1.3.1
Nuclear Binding Energies

One of the more important properties of any compound system, whether molecular, atomic, or nuclear, is the amount of energy needed to pull it apart, or, alternatively, the energy released in assembling it from it constituent parts. In the case of nuclei, these are protons and neutrons. The binding energy of a nucleus ${}^{A}X_{N}$ can be defined as

$$B(A, Z) = ZM_{\mathrm{H}} + NM_{\mathrm{n}} - M_{\mathrm{X}}(Z, A) \quad (1.6)$$

where M_{H} is the mass of a hydrogen atom, M_{n} the mass of a neutron, and $M_{\mathrm{X}}(Z,A)$ the mass of a neutral atom of isotope A. Because the binding energy of atomic electrons is very much less than nuclear binding energies, they have been neglected in Eq. (1.6). The usual units are atomic mass units, u. Another quantity that contains essentially the same information as the binding energy is the *mass excess* or the *mass defect*, $\Delta = M(A) - A$. (Another useful quantity is the *packing fraction* $P = [M(A) - A]/A = \Delta/A$.) The most interesting experimental quantity $B(A,Z)/A$ is the binding energy per nucleon, which varies from somewhat more than 1 MeV nucleon^{-1} (1.112 MeV nucleon^{-1}) for deuterium (${}^{2}\mathrm{H}_{1}$) to a peak near ${}^{56}\mathrm{Fe}_{30}$ of 8.790 MeV nucleon^{-1} and then falls slowly until, at ${}^{235}\mathrm{U}_{143}$, it is 7.591 MeV nucleon^{-1}. Except for the very light nuclei, this quantity is roughly (within about 10%) 8 MeV nucleon^{-1}. A strongly bound light nucleus is the α-particle, as for ${}^{4}\mathrm{He}_{2}$ the binding energy is 7.074 MeV nucleon^{-1}. It is instructive to plot, for a given mass number, the packing fraction as a function of Z. These plots are quite accurately parabolas with the most β-stable nuclide at the bottom. The β^{-} emitters will occur on one side of the parabola (the left or lower two side) and the β^{+} emitters on the other side. For odd-A nuclei, there is but one parabola, the β-unstable nuclei proceeding down each side of the parabola until the bottom or most stable nucleus is reached. For the even-A nuclei, there are two parabolas, with the odd–odd one lying above the even–even parabola. The fact that the odd–odd parabola is above the even–even one indicates that a pairing force exists that tends to increase the binding energy of the even–even nuclei. See Figure 1.1 for the $A = 100$ mass chain. Other indications of the importance of this pairing force are the before-mentioned 0^{+} ground states of all even–even nuclei and the fact that only four stable odd–odd nuclei exist: ${}^{2}\mathrm{H}_{1}$, ${}^{6}\mathrm{Li}_{3}$, ${}^{10}\mathrm{B}_{5}$, and ${}^{14}\mathrm{N}_{7}$. For even-A nuclei, the β-unstable nuclei zig-zag between the odd–odd parabola and the even–even parabola until arriving at the most β-stable nuclide, usually an even–even one. If the masses for each A are assembled into a three-dimensional plot (with N running along one long axis, Z along a perpendicular axis, and $M(A,Z)$ mutually perpendicular to these two), one finds a "landscape" with a deep valley running from one end to the other. This valley is known as the *valley of stability*.

The immediate consequence of the behavior of $B(A,Z)/A$ is that a very large amount of energy per nucleon is to be gained from combining two neutrons and two protons to form a helium nucleus. This process is called *fusion*. The release of energy in the fission process follows from the fact that $B(A,Z)/A$ for uranium is less than for nuclei with more or less half the number of protons. Finally, the fact that the binding energy per nucleon peaks near iron is important to the understanding of those stellar explosions known as *supernovae*. In Figure 1.2, the packing fraction, $P = \Delta/A$, is plotted against A for the most stable nuclei for a given mass number. Note that P has a broad minimum near iron ($A = 56$) and rises slowly until lawrencium ($A = 260$). This shows most clearly the

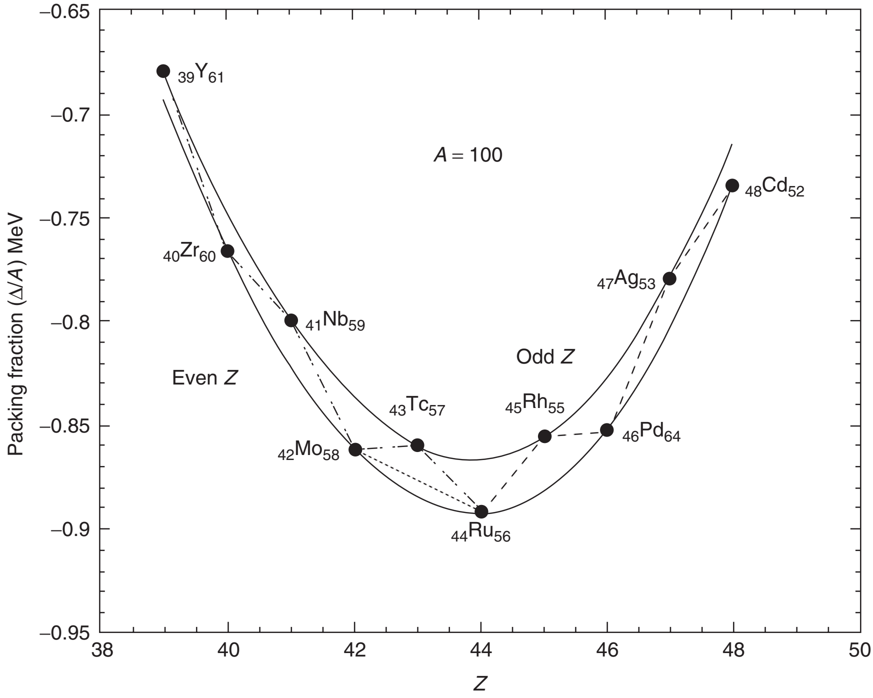

Figure 1.1 The packing fraction Δ/A plotted against the nuclear charge Z for nuclei with mass number $A = 100$. Note that the odd–odd nuclei lie above the even–even ones. The β^- transitions are indicated by - • -, the β^+ transitions by ---, while the double β-decay $^{100}_{42}\mathrm{Mo}_{58} \to ^{100}_{44}\mathrm{Ru}_{56}$ is denoted by• • •. Data from [14]. The double β-decay from [15].

energy gain from the fission of very heavy elements.

1.3.2
The Semiempirical Mass Formula

The semiempirical mass formula may be looked upon as simply the expansion of $B(A,Z)$ in terms of the mass number. Because $B(A,Z)/A$ is nearly constant, the most important term in this expansion must be the term in A. From Eq. (1.1) relating the nuclear radius to $A^{1/3}$, we see that a term proportional to A is a volume term. However, this term overbinds the system because it assumes that each nucleon is surrounded by the same number of neighbors. Clearly, this is not true for surface nucleons, and so a surface term proportional to $A^{2/3}$ must be subtracted from the volume term. (One might identify this with the surface tension found in a liquid drop.) Next, the repulsive Coulomb forces between protons must be included. As this force is between pairs of protons, this term will be of the form $Z\{Z-1)/2$, the number of pairs of Z protons, divided by a characteristic nuclear length or $A^{1/3}$. Two other terms are necessary in this simple model. One term takes into account that, in general, $Z \sim A/2$, clearly true for stable light nuclei, and less so for heavier stable

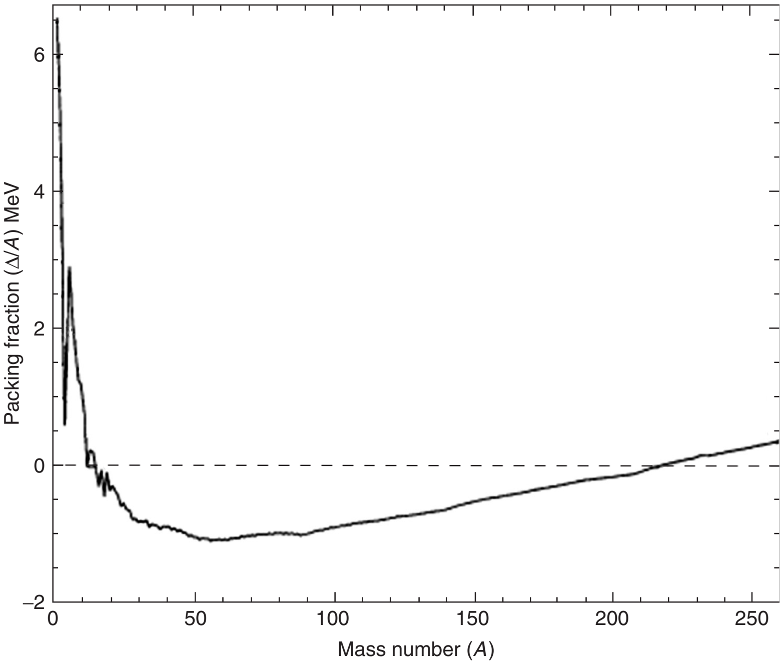

Figure 1.2 The packing fraction Δ/A plotted against the mass number A for all nuclei from ${}^{2}_{1}D_{1}$ to ${}^{260}_{103}Lr_{157}$. Data from [14].

nuclei where more neutrons are needed to overcome the mutual repulsion of the protons. This term is generally taken to be of the form $a_{\text{sym}}\,(N-Z)^2/A$. The other term takes into account the fact, noted in Section 1.3.1, that even–even nuclei are more tightly bound than odd–odd nuclei because all of the nucleons of the former are paired off. This is done by adding a term $\delta/2$ that is positive for even–even systems and negative for odd–odd systems and zero for odd-A nuclei. Thus, the two parabolas for even A are separated by δ. From Eq. (1.6), the semiempirical Bethe–Weisäcker mass formula then becomes $M(A,Z) = ZM_{\text{H}} + NM_{\text{n}} - B(A,Z)$ with

$$B(A,Z) = a_{\text{v}}A - a_{\text{s}}A^{2/3} - a_{\text{c}}Z(Z-1)A^{-1/3} - a_{\text{sym}}(N-Z)^2A^{-1} + \frac{\delta}{2} \qquad (1.7)$$

Originally, the constants were fixed by the measured binding energies and adjusted to give appropriate behavior with the mass number [16]. Myers and Swiatecki [17] (see also [18]) have included other terms to account for regions of nuclear deformation, as well as an exponential term of the form $-a_{\text{a}}A\exp(-\gamma A^{1/3})$, for which they provide no physical explanation beyond the fact that it reduces the deviation from experiment. Their model evolved into the macroscopic-microscopic global mass formula, called the *finite-range droplet model* (see [19]) and the *DZ-model* proposed by Duflo and Zuker [20], and more microscopic models, called *HFB* [21]. The many adjustable parameters of the available mass formulas are then fitted to masses of 1760 atomic nuclei [22]. The formulas fit binding energies quite well with errors below 1%, but still have problems to predict masses far from stability. As those are important for nuclear astrophysics, the measurement of masses of exotic nuclei is an important field today.

A number of consequences flow from even a superficial examination of Eq. (1.7). The fact that the binding energy per nucleon, B/A, is essentially constant with A implies that the nuclear density is constant and, thus, the nuclear force saturates. That is, nucleons interact only with a small number of their neighbors. This is a consequence of the very short range of the strong force. If this were not so, then each nucleon would interact with all others in the given nucleus (just as the protons interact with all other protons), and the leading term in $B(A,Z)$ would be proportional to the number of pairs of nucleons, which is $A(A-1)/2$ or roughly A^2. This would imply that B/A would go as A. Thus, not only does the nuclear force saturate (the Coulomb force does not) but it is also of very short range (that of the Coulomb force is infinite) as the sizes of nuclei are of the order of 3.0 fm (recall Eq. (1.1)).

1.4
Nuclear Charge and Mass Distributions

1.4.1
General Comments

In his 1911 paper, Rutherford was able to conclude that the positive charge of the atom was concentrated within a sphere of radius $<10^{-14}$ m (10 fm). This result came from α-particle scattering. However, for energetic enough α-particles, the scattering result will contain a component due to nuclear interactions of the α-particle, as well as the Coulomb interaction. For probing the structure of nuclei, electrons have the advantage that their scattering is purely Coulombic; however, to determine details of the internal nuclear structure, electron energies must be well over 100 MeV for their de Broglie wavelengths to be less than nuclear dimensions. Well before the existence of such high-energy electron beams, nuclear structure effects were extracted from information provided by optical hyperfine spectra. In particular, nuclear charge distributions (electric quadrupole moments) and current distributions (magnetic dipole moments) were deduced from very accurate optical measurements (see the following section). A result involving the innermost electrons of heavy atoms is the *isotope shift*, which can be observed in atomic X-rays. This arises because the nuclear radii for two different isotopes of the same atom will produce slightly different binding energies of their K-shell electrons. Thus, the K X-rays of these isotopes will be very slightly different in energy. As an example, the isotopic pair $^{203}\mathrm{Tl}_{122}$ and $^{205}\mathrm{Tl}_{124}$ have an isotope shift of about 0.05 eV. Another early method to determine the charge radius is to take the difference between the binding energies of two mirror nuclei (cf. Section 1.2.1). This leads to an expression that only involves a_c and, thus, the nuclear radius. This is useful for light nuclei for which mirror pairs occur.

With the advent of copious beams of negative muons, much more accurate optical-type hyperfine spectrum studies could be made. The process is quite simple, and the advantages obvious. By stopping negative muons in a target, an exotic atom is formed in which the muon replaces an orbital electron and transitions to the muonic K-shell follow. These transitions of the muon to the $1s_{1/2}$ state emit photons of the appropriate (but high) energies. (As the muon is more than 200 times as massive as the electron, the radii of the muon orbits are reduced by that amount, so that electron-shielding problems are much reduced.) The energies of the photons are such that the $2p_{1/2}-2_{3/2}$ splitting is easily measured (in $^{116}Sn_{66}$, it is 45.666 keV). Thus, both nuclear charge radii and isotope shifts are quite accurately determined. In some recent experiments, root mean square (RMS) charge radii have been measured with a precision of 2×10^{-18} m. As electron scattering and muonic atoms are the two methods of measuring characteristics of the nuclear charge radius most susceptible of the greatest accuracy, they will be discussed in turn.

1.4.2
Nuclear Charge Distributions from Electron Scattering

In any scattering experiment, what is measured is the differential cross section $(d\sigma/d\Omega)$. Rutherford developed an expression for α-particle scattering that can be used for low-energy, spinless particles incident on a spinless target. Both incident and target particles are assumed to be point particles. The differential cross section for scattering relativistic electrons off point-charged particles leads to the expression for Mott scattering, while, if the target particle has nonzero spin (there is then a magnetic contribution), one obtains the Dirac scattering formula for $(d\sigma/d\Omega)$. However, real nuclei are not point particles, so one needs to make use of the charge form factor $F(\vec{q})$ with $\vec{q}$ the transferred momentum. The form factor $F(\vec{q})$ is then the Fourier transform of the charge density $\rho_{ch}(\vec{r})$:

$$F(\vec{q}) = \int \rho_{ch}(\vec{r}) e^{i\vec{q}\cdot\vec{r}}\, d\tau \qquad (1.8)$$

If one restricts the problem to spherically symmetric distributions, the angular integration of the Fourier integral follows at once, so that

$$F(q) = \frac{4\pi}{q}\int \rho_{ch}(r)\sin(qr) r dr \qquad (1.9)$$

If the target nucleus has zero spin (applicable to all even–even nuclei), then the differential cross sections for a point target and a finite-sized target are related by

$$\frac{d\sigma}{d\Omega} = \left(\frac{d\sigma}{d\Omega}\right)_{\text{Ruth.}} F(q)^2 \qquad (1.10)$$

With the charge form factor determined experimentally, the inverse transform yields the radial charge density

$$\rho_{ch}(r) = \frac{1}{2\pi^2 r}\int F(q)\sin(qr) q dq \qquad (1.11)$$

If the target nucleus is not of spin zero, then an additional term containing the so-called transverse form factor, $F_T(q)$, is needed. (The form factor defined in Eq. (1.9) is sometimes called the *longitudinal form factor*.) In any event, the charge distribution must be normalized to the number of protons (Z) in the target nucleus.

At this point, there are two ways to proceed. The first is a model-independent analysis of the form factor, or, second, one can assume a model with several parameters and fit these to the data. Limiting oneself to small momentum transfers, one can obtain the form factor as a power series in q^2 by expanding $\sin(qr)$ in Eq. (1.9) in a power series of its argument. Keeping only the lowest term of order q^2, one obtains

$$F(q) = Z\left(1 - \frac{1}{6}q^2\langle r^2\rangle\right) \qquad (1.12)$$

with

$$\langle r^2 \rangle = \int r^2 \rho_{\mathrm{ch}}(r) 4\pi r^2 \mathrm{d}r \qquad (1.13)$$

the RMS radius of the charge distribution. It should be noted that this is not the nuclear radius R, which is usually taken as the radius of the constant-density sphere. This yields

$$\langle r^2 \rangle = \frac{3}{5} R^2 \qquad (1.14)$$

Data compilations [23] show that for most stable nuclei,

$$R \approx 1.25 A^{1/3} \qquad (1.15)$$

As was stated in Section 1.2.1, a better way to describe the charge distribution is to use a Fermi distribution which takes account of the constant charge density, ρ_0, at the center of the nucleus and the gradual decrease near the surface. This is achieved by

$$\rho_{\mathrm{ch}}(r) = \rho_0 \frac{1}{1 + \mathrm{e}^{\frac{r - R_{1/2}}{a}}} \qquad (1.16)$$

with $R_{1/2}$ the radius at half density and a the diffuseness parameter indicating the distance at which the density falls from 90 to 10% of the constant density ρ_0. For heavy nuclei, the following parameterization holds:

$$\begin{aligned} \rho_0 &= 0.17 \frac{Ze}{A}\,\mathrm{fm}^{-3} \\ a &= 2.4\,\mathrm{fm} \\ R_{1/2} &= 1.128 A^{1/3} - 0.89 A^{-1/3}\,\mathrm{fm} \end{aligned} \qquad (1.17)$$

There are enough experimental electron-scattering data available throughout all regions of the stable nuclei that quite accurate charge parameters exist for almost all of the systems. The compendium by de Vries *et al.* [23] lists these parameters fitted to the data for several distribution functions in addition to the two-parameter Fermi functions.

1.4.3
Nuclear Charge Distributions from Atomic Transitions

During the last decades, tremendous progress was obtained in the study of atomic transitions using high-precision laser spectroscopy. This allows the measurement of nuclear charge radii and also of nuclear moments for stable and even unstable isotopes. This is because the difference between a point nucleus and a finite-size nucleus causes a very small change in the Coulomb potential the atomic electrons feel. A small energy difference on the atomic levels results when we assume that the nucleus is a sphere with constant charge density. For 1s electrons one obtains

$$\Delta E_{1\mathrm{s}} = \frac{2}{5} \frac{Z^4 e^2}{4\pi\varepsilon_0} \frac{R^2}{a_0^3} \qquad (1.18)$$

with a_0 the Bohr radius. Because no point nucleus exists and the theoretical calculations are not accurate enough to calculate the small shift exactly, one generally measures isotope shifts as the frequency difference of atomic transitions measured in two isotopes of a given element. This then yield the differences in the nuclear radius. Starting from known radii of stable isotopes, it is then possible to determine the radii of unstable nuclei on which one cannot perform electron scattering. It is also possible to measure isotope shifts using optical transitions. Because these are caused by the outermost electrons, the shifts are very small in the order of parts per million. As indicated above, they are still within reach of modern laser techniques.

The small shift for 1s binding energies is related to the large difference between the Bohr radius and the nuclear radius. On replacing one electron by a muon, the muonic orbits shrink by a factor of 207, the mass difference between the heavy muon and the light electron. At the same time, the muon binding energy is increased by a similar factor, making the transition energies in the mega electron volt region. The energies are so high (in $^{238}\mathrm{U}_{146}$, the measured 2p–1s transition is about 6.1 MeV) that one must generate Dirac solutions for the muon moving in a Coulomb potential generated by a non-point-charge distribution. To these initial Dirac solutions, one must add corrections, which, in order of size, are vacuum polarization, nuclear polarization, the Lamb shift, and relativistic recoil. Electronscreening corrections are often included, but they are very tiny (for the ls muonic state in $^{238}\mathrm{U}_{146}$, this correction has the value of 11 eV).

As many nuclei are not spherical, several studies have used as the appropriate charge distribution a slightly modified form of Eq. (1.16), which includes the deformations

$$\rho_{\mathrm{ch}}(r) = \rho_0 \left\{ 1 + \exp\left(\frac{r - R_{1/2}\left(1 + \sum_{n=1} \beta_n Y_{n0}(\theta, \phi)\right)}{a} \right) \right\}^{-1} \tag{1.19}$$

Here the β_{n} are deformation parameters that determine the nuclear shape. As an example, the nuclear mean square radius may be expressed as

$$\begin{aligned} &\langle r^2 \rangle_{\mathrm{deformed}} \\ &\approx \langle r^2 \rangle_{\mathrm{sph}} \left(1 + \frac{5}{4\pi} \left(\beta_2^2 + \beta_4^2 + \cdots \right) \right) \end{aligned} \tag{1.20}$$

Experiments to fit a, c, and, in deformed regions, β_{n} have been made throughout the periodic table with results consistent with the electron data. However, to combine the results of electron-scattering experiments with those from muonic atoms, it is necessary to use the so-called Barrett moment

$$\left\langle r^k \mathrm{e}^{-\alpha r} \right\rangle = \frac{4\pi}{Z} \int_0^\infty \rho(r) \mathrm{e}^{-\alpha r} r^{k+2} \mathrm{d}r \tag{1.21}$$

where k and α are fitted to the experimental data. The muonic data are equivalent to data from electron-scattering experiments at low momentum transfer. The inclusion of the muonic Barrett moment improves the overall fit by reducing normalization errors. This then reduces the uncertainties over what would be obtained by fitting either the electron-scattering or the muonic atom data alone. Extensive tables of data fitted by various charge distribution models as well as model independent analyzes can be found in de Vries *et al.* [23].

1.4.4
Nuclear Mass Distributions

While the measurement of the charge distribution can be made using electromagnetic probes, this is not possible for the mass distribution because of the uncharged neutron. Instead, the nuclear strong force has to be used. This is more complicated as mostly both Coulomb force and strong force are present. Nevertheless, from α-scattering experiments, information of the mass distribution is obtained. There are also indirect ways in which one can get information on nuclear mass radii. One example is the dependence of α-decay rate on the nuclear radius that defines the Coulomb barrier. In deformed nuclei, this causes an anisotropy because the Coulomb barrier is lower in the direction of the longest axis, making the tunnel probability enhanced. A second way is to use pions instead of muons. These interact with the nucleus through both the Coulomb force and strong force, which, in comparison to muonic atoms, causes an extra shift that allows the determination of the mass radius. The result of these experiments on stable nuclei finds that the charge and mass radii are equal to within about 0.1 fm. This somewhat surprising result can be understood as a balance between the proton Coulomb repulsion that tends to push the protons to the outside and a strongly attractive neutron–proton strong force that tends to pull the extra neutrons to the inward.

Recently, the common opinion that the radii scale with $A^{1/3}$ was found to be heavily violated in more exotic nuclei.

Especially in light nuclei with a large neutron number, so-called halo-nuclei, strong deviations were observed (an early review is given in [24]). Using the radioactive beam techniques, very neutron-rich He, Li, and Be isotopes can be created and studied in the laboratory. It turned out that these loosely bound nuclei show very extended neutron radii whereby two neutrons are moving at radii similar to the radii of Pb isotopes. Moreover, as is the case of $^{11}\mathrm{Li}_8$, the bound system consists of three entities: two neutrons and a $^{9}\mathrm{Li}_6$ that cannot exist two by two, as the dineutron and $^{10}\mathrm{Li}_7$ are unbound. The research on exotic nuclei is still in its infancy and more exotic features such as proton halos or neutron skins are expected. They are of importance as they may influence the creation of the elements under astrophysical conditions.

1.5
Electromagnetic Transitions and Static Moments

1.5.1
General Comments

Static electromagnetic nuclear moments played an important role in the unscrambling of the detailed measurements of atomic optical hyperfine structure well before the gross components of atomic nuclei were in hand. Almost a decade before the discovery of the neutron, Pauli [7] suggested that the optical hyperfine splitting might be due, in part, to the interaction with a nuclear magnetic moment (μ). This suggestion lay fallow until 1930, when Goudsmit and Young, using the spectroscopic data of Schiller and of Granath, deduced the nuclear magnetic moment of ^{7}Li to be $\mu = 3.29\mu_{\mathrm{N}}$, where the nuclear magneton equals

$$\mu_{\mathrm{N}} = \frac{e\hbar}{2M_{\mathrm{p}}} = 5.050789 \times 10^{-27}\,\frac{\mathrm{J}}{\mathrm{T}} \tag{1.22}$$

This value is quite close to the currently accepted value ($\mu = 2.327\mu_{\mathrm{N}}$). Because of the existence, by then, of extensive hyperfine optical spectroscopic data, Goudsmit, in 1933, was able to publish a table of some 20 nuclear magnetic moments ranging from ^{7}Li to ^{209}Bi. In 1937, Schmidt published a simple, single-particle model of nuclear magnetic moments and supported it with the experimental moments of 32 odd-proton nuclei and 15 odd-neutron nuclei. This simple model yields what is now known as the *Schmidt limits*, within which almost all nuclear magnetic moments lie (see the following).

The suggestion that the nuclear electric quadrupole moments (Q) might also play an important role in optical hyperfine structure was again made before the discovery of the neutron. Racah [8] was the first to work out the theory associated with "nuclear charge asymmetry" and the interaction with the atomic electrons. Casimir [25], sometime later, developed the theory of nuclear electric quadrupole hyperfine interaction and applied it to ^{151}Eu and ^{153}Eu. In this paper, Casimir mentions work by Schiller and Schmidt, who determined Q for ^{175}Lu. A short time later, Gollnow [26] obtained $Q = 5.9$ b for this nucleus, quite close to the currently accepted value of 5.68 b. This very large quadrupole moment (very much larger than can be accounted for by the single-particle shell model) was to provide, 20 years later, strong impetus for the development of the collective model of the nucleus. In 1954, Schwartz [27] extended the theory of nuclear hyperfine structure to examine the magnetic octupole hyperfine interaction and calculated the first four nuclear magnetic octupole moments (O) from data of the hyperfine structure of the nuclear ground states. The next nuclear moment is the hexadecapole (H); however, no direct measurements of such static moments exist. What is known about these moments comes mainly from electromagnetic transitions of electrons and negative muons. For an in-depth theoretical study of all of these moments and how they can be used to test various nuclear models see, in particular, the text by Castel and Towner [28].

Nowadays, the measurement of moments is still very important to assess the single-particle structure of exotic nuclei and several powerful techniques have been developed in this domain [29]. Most information is, however, gathered via the determination of electromagnetic transitions by γ-ray spectroscopy. This is to a large extent due to the availability of large-volume semiconductor detectors for γ-ray detection and the high computing power that allows one to analyze more and more complex measured spectra using coincidence conditions. The recent development of γ-ray tracking detectors out of segmented Ge-detectors offers very high perspectives in the field of exotic nuclei [30].

1.5.2
Electromagnetic Transitions and Selection Rules

Without going into a detailed discussion on how matrix elements are calculated, we review here the calculation of electromagnetic transitions. The interested reader can find more details in Heyde [31]. The calculation of transitions and also moments involves the wavefunctions of nuclear states and forms a very sensitive probe for nuclear structure research. On the other hand these transitions and moments are electromagnetic in nature making the interaction very well understood. Using the long-wavelength approximation $\lambda \gg R$ and a multipole expansion of the electromagnetic operators, the transition rates per unit of time can be expressed as

$$T(L) = \frac{2}{\varepsilon_0 \hbar} \frac{L+1}{(L[(2L+1)!!]^2)} \left(\frac{\omega}{c}\right)^{2L+1} B(L) \tag{1.23}$$

with L the multipolarity, ω the angular frequency of the radiation such that $\hbar\omega \cong E_i - E_f$ up to a small nuclear recoil correction, and $B(L)$ the reduced transition probability

$$B(J_i \to J_f; L) = \sum_{M_i, M_f} \left| \left\langle \alpha_f; J_f M_f \right| O(LM) \left| \alpha_i; J_i M_i \right\rangle \right|^2 \tag{1.24}$$

in units of $e^2 b^L$ and $\mu_N b^{L-2}$ for electric and magnetic $B(LM)$ values. The labels α identify the initial and final states and $O(LM)$ is the electric or magnetic multipole operator of rank LM. As all states have good angular momentum, one can now use for the transition rates the Wigner–Eckart theorem to remove all reference to the M projections. This yields

$$B(J_i \to J_f; L) = \frac{1}{2J_i + 1} \left| \left\langle \alpha_f; J_f \left\| O(L) \right\| \alpha_i; J_i \right\rangle \right|^2 \tag{1.25}$$

The electric multipole operator for a number of point charges becomes

$$O(E; LM) = \sum_i e_{\text{eff}}(i) r_i{}^L Y_{L,M}(\theta_i, \phi_i) \tag{1.26}$$

with $e_{\text{eff}}(i)$ the effective charge of the ith nucleon. Here it is anticipated that owing to core polarization effects and truncations of the model space, other values than the free charges $+e(0)$ for proton (neutron) need to be used. Instead of the operator (Eq. (1.25)), one can also use a similar operator but using the nuclear electric charge density ρ_{ch} and an integration over the nuclear volume. These are used as several models describe the nucleus as a droplet (Section 1.7). For the magnetic multipole operator, we have

$$\begin{aligned} O(M; LM) = {} & \sqrt{L(2L+1)}\mu_N \sum_i r_i{}^{L-1} \\ & \times \left\{ \frac{e_{\text{eff}}(i)}{e} \frac{2}{L+1} [Y_{L-1} \otimes j_i]^{(L)} \right. \\ & + \left(g_s(i) - \frac{e_{\text{eff}}(i)}{e} \frac{1}{L+1} \right) \\ & \left. \times \left[Y_{L-1} \otimes s_i \right]^{(L)} \right\} \end{aligned} \tag{1.27}$$

with the effective gyromagnetic ratio g_s which also may differ from the free ones.

The multipole expansion and the fact that states in atomic nuclei have good angular momentum and parity leads to several selection rules. The first one is related to the vector coupling of the angular momentum and states $|J_i - J_f| \le L \le J_i + J_f$. The second is due to the parity of the operators, which clearly is $(-1)^L$ for the electric and $(-1)^{L-1}$ for the magnetic operator. Owing to this, electric and magnetic transitions of order L cannot take place at the same time between states, and moments such as the electric dipole moment are forbidden. While the selection rules allow the determination of the spin and parities of nuclear excited states, they are also (at a higher level) invaluable to test nuclear models (Section 1.7).

Weisskopf has estimated the so-called single-particle values or Weisskopf units (W.u.) by assuming that a single-particle makes a transition with multipole L from a state with spin $L + 1/2$ toward a state with spin $1/2$, that the radial part of the wavefunction can be approximated by a constant value up to the radius R, and that certain values for the effective charges hold. This leads to the following estimates for the half-lives corresponding to the single-particle values:

$$T_{1/2}(E1) = \frac{6.764 \times 10^{-6}}{E_\gamma^3 A^{2/3}} \quad \text{(s)}$$

$$T_{1/2}(M1) = \frac{2.202 \times 10^{-5}}{E_\gamma^3} \quad \text{(s)}$$

$$T_{1/2}(E2) = \frac{9.527 \times 10^{6}}{E_\gamma^5 A^{4/3}} \quad \text{(s)}$$

$$T_{1/2}(M2) = \frac{3.102 \times 10^{7}}{E_\gamma^5 A^{2/3}} \quad \text{(s)}$$

$$T_{1/2}(E3) = \frac{2.045 \times 10^{19}}{E_\gamma^7 A^{2}} \quad \text{(s)}$$

$$T_{1/2}(M3) = \frac{6.659 \times 10^{19}}{E_\gamma^7 A^{4/3}} \text{ (s)} \quad (1.28)$$

One notices that transition rates of the lower multipolarities are faster than the higher by orders of magnitude and that for a given multipolarity, the electric ones are about 100–1000 faster as the magnetic ones. Owing to the selection rules and enhanced quadrupole collectivity, only $E2$ and $M1$ transitions happen on similar timescales. In this case, one has a transition of mixed multipolarity.

The Weisskopf estimates are very crucial to determine whether a transition is caused by a single nucleon changing orbits or by several nucleons acting in a collective way. The measurement of transition rates of excited states delivers very important information on nuclear structure, but is also quite involved. One needs to measure the lifetime of a state, the (mixed) multipolarity, and the energy and intensities of the transitions deexiting a given state. To this end, γ-ray arrays consisting of several Ge-detectors are appropriate. Besides this, electromagnetic decay can also take place with the emission of conversion electrons. These electrons allow one to determine the multipolarity as well as to observe the by γ-emission forbidden $E0$ transitions (due to the fact that the photon has spin 1 with projection +1 and −1).

1.5.3
Static Moments

In contrast to the transition rates, the *multipole moments* are generally defined as the matrix element of the $M=0$ component of the moment operator for a single state with magnetic projection $M=+J$. Of the moments, the two lowest are the most important.

1.5.3.1 Magnetic Dipole Moments

If one assumes that the magnetic properties are associated with the individual nucleons, then the magnetic moment is defined as

$$\mu = \langle \alpha, JJ| \sum_i g_l(i) l_{z,i} + g_s(i) s_{z,i} |\alpha; JJ\rangle \quad (1.29)$$

where the sum extends over all of the A nucleons. Generally, this will not be needed; for instance, the magnetic moment of an odd-A nucleus will be generated by the last neutron and proton as the adjacent even–even ground state has no magnetic moment. In Eq. (1.29), g_l and g_s are the orbital and spin gyromagnetic ratios. They are often chosen as $0.7g_{\text{free}}$ where the free-particle values are $g_l = 1$, $g_s = 5.587$ for protons and $g_l = 0$, $g_s = -3.826$ for neutrons (all in μ_N). It is instructive to calculate the magnetic moment for a single nucleon (which, as explained earlier, is a good approximation for the ground state of odd-A nuclei). Using the total angular momentum and the Wigner–Eckart theorem, one deduces the single-particle magnetic moments for aligned and antialigned orbital momentum and spin:

$$\mu(j = l + 1/2) = j\left(g_l + \frac{(g_s - g_l)}{2}\right) \quad (1.30a)$$

$$\mu(j = l - 1/2) = \left(jg_l - \frac{j}{j+1}\frac{(g_s - g_l)}{2}\right) \quad (1.30b)$$

The magnetic moments plotted as a function of spin form the so-called Schmidt lines. It is interesting that, when plotted, almost all of these moments lie between the Schmidt lines. The moments that lie outside these limits occur mainly for some very light nuclei. One may conclude that the single-particle model does possess some validity. Another set of limits, the Margenau–Wigner (M-W) limits, is obtained by replacing the free-particle values for the orbital gyromagnetic ratios, g_l, by the uniform value Z/A. The justification for this is that one is in effect averaging over all states that lead to the correct nuclear spin. This calculation represents an early attempt to account for core contributions to the dipole-moment operator. Figure 1.3 and Figure 1.4 show plots of a number of the ground-state magnetic moments for odd-Z (Figure 1.3) and odd-N (Figure 1.4) nuclei.

Besides the ground state, excited states can have magnetic moment and their measurement is often used to extract information on the underlying single-particle structure. Common in nuclear structure physics is the use of g-factors that are analogous to the single-particle gyromagnetic ratios, except that they are dimensionless. They are defined as

$$\mu = gI\mu_\mathrm{N} \tag{1.31}$$

One way to determine the g-factor is to measure the Lamor frequency when excited nuclei are placed in an external magnetic field B. Then,

$$\omega_\mathrm{L} = \frac{gB\mu_\mathrm{N}}{\hbar} \tag{1.32}$$

The main problem hereby is to align the spins of an ensemble of atomic nuclei.

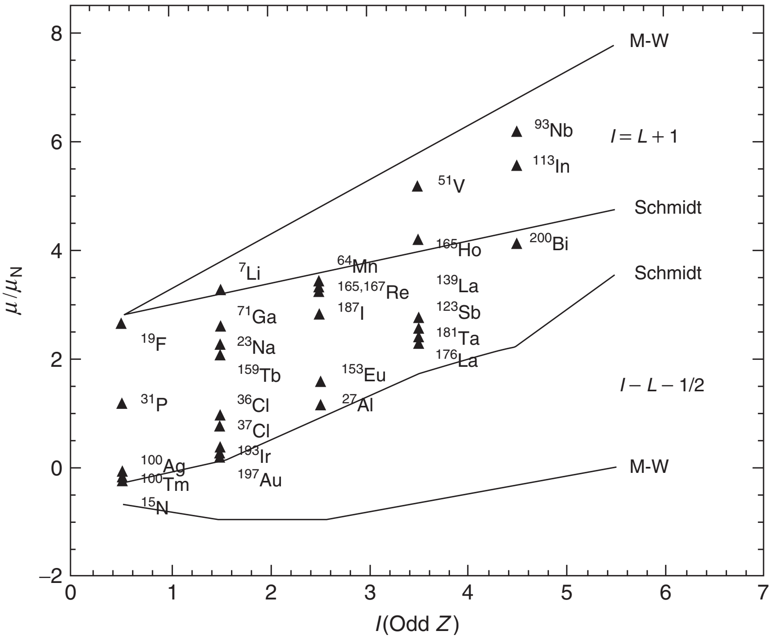

Figure 1.3 Nuclear magnetic moments in units of the nuclear magneton (μ/μ_N) plotted against the nuclear spin (I) for a number of odd-Z nuclei. The Schmidt limits, as well as the Margenau–Wigner (M-W) limits, are shown as solid lines. Data from [14].

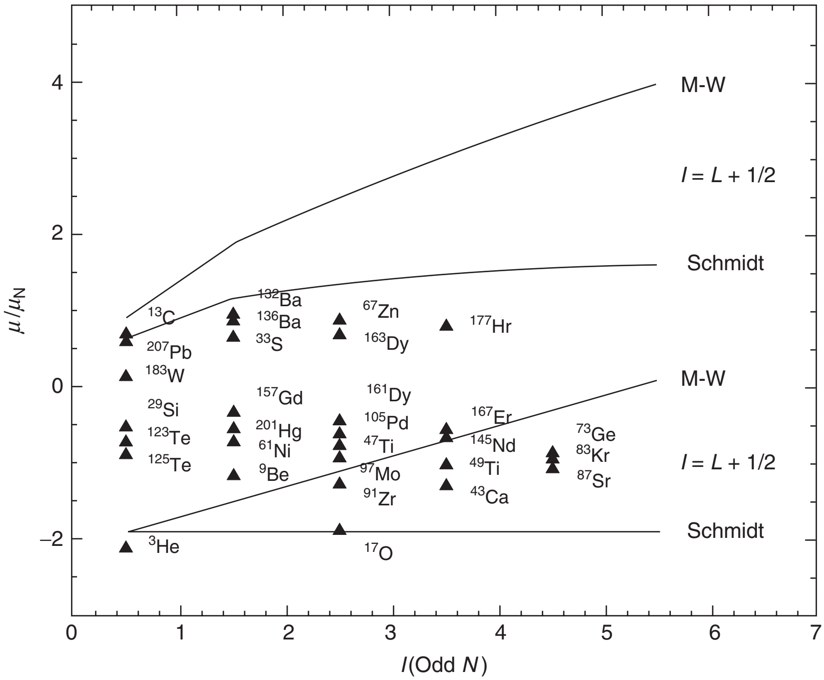

Figure 1.4 Nuclear magnetic moments in units of the nuclear magneton (μ/μ_N) plotted against the nuclear spin (I) for a number of odd-N nuclei. The Schmidt limits, as well as the Margenau–Wigner (M-W) limits, are shown as solid lines. Data from [14].

This has to be done by the nuclear reaction used, cooling in an external magnetic field, or via the observation of changes in angular correlations.

1.5.3.2 Electric Quadrupole Moments

As a general definition of the quadrupole moment, we have the expectation value of $(3z^2 - r^2)$. Using the proportionality of the quantity with $Y_{20}(\theta,\phi)$ one gets

$$Q(J) = \sqrt{\frac{16\pi}{5}} \times \left\langle \alpha; JJ \left| \sum_i \frac{e_{\text{eff}}\,(i)}{e} r_i^{\,2} Y_{2,0}(\theta_i, \phi_i) \right| \alpha; JJ \right\rangle \tag{1.33}$$

or, using the Wigner–Eckart theorem,

$$Q(J) = \sqrt{\frac{16\pi}{5}} \sqrt{\frac{J(2J-1)}{(2J+1)(2J+3)(J+1)}} \times \left\langle \alpha; J \left\| \sum_i \frac{e_{\text{eff}}\,(i)}{e} r_i^{\,2} Y_2(\theta_i, \phi_i) \right\| \alpha; J \right\rangle \tag{1.34}$$

The quadrupole moment can also be calculated for a single nucleon in an orbit $j = J$. This yields

$$Q(j) = -\frac{(2j-1)}{(2j+2)} \frac{e_{\text{eff}}(i)}{e} \langle r^2 \rangle \tag{1.35}$$

One thus obtains a negative quadrupole moment for a single nucleon. If the orbit is filled up to a single hole, a quadrupole moment as in Eq. (1.35) but with a positive sign is expected. We would like to illustrate the application of Eq. (1.35) with examples near the doubly magic nucleus $^{16}O_8$ (see also Section 1.7). The orbit that the ninth nucleon can occupy has $j = 5/2$ and we can use Eqs. (1.13) and (1.14) to estimate $<r^2>$. This yields for $^{17}F_8$ using the free-proton charge $Q = -5.9\,\text{fm}^2$, which is in excellent agreement with the experimental absolute value given in Firestone *et al.* (1996) of $|Q| = 5.8(4)\,\text{fm}^2$. For the odd neutron nucleus $^{17}O_9$, one finds experimentally $Q = -2.578\,\text{fm}^2$, which shows the need for effective charges and can be reproduced using $0.44e$ as neutron effective charge. Finally, if we place five neutrons in the $j = 5/2$ orbit, we have the $N = 13$ isotones and expect moments of $Q = +2.6\,\text{fm}^2$. Experimentally, one finds $Q = 20.1(3)\,\text{fm}^2$ in $^{25}Mg_{13}$ and $Q = 10.1(2)\,\text{fm}^2$ in $^{23}Ne_{13}$. This observation of much larger quadrupole moments occurs for most atomic nuclei having several nucleons in an orbit or in several orbits, indicating that the model is too simple and that all of the electric charges must be considered. This holds especially if the core is not spherically symmetric and the motion becomes collective.

1.6
Excited States and Level Structures

1.6.1
The First Excited State in Even–Even Nuclei

The most obvious characteristic of the various nuclei is that all of the even–even nuclei have ground-state spins and parities of 0^+. This not only categorizes one large group of nuclei but also indicates that the nuclear force is such that it couples, preferentially, pairs of like nucleons to angular momentum zero. The second observation is that the first excited state in even–even nuclei is almost always a 2^+ excitation. This can be understood by the combination of good total angular momentum and the Pauli principle. If one couples two nucleons in orbit j and implies the antisymmetrization, one obtains

$$\varphi(j,j;JM) = \frac{1}{2}\sum_{mm'}(jmjm'|JM)(\psi_1(jm)\,\psi_2(jm') - \psi_2(jm)\psi_1(jm')) \tag{1.36}$$

with the Clebsch–Gordan coupling coefficient for the angular momentum coupling. This can, however, be rewritten as the m-values are in the summation. Using the symmetry properties, one arrives at

$$\varphi\,(j,j;JM) = \frac{1}{2}\left(1-(-1)^{2j-J}\right) \times\sum_{mm'}(jmjm'|JM)\,\psi_1(jm)\psi_2(jm') \tag{1.37}$$

The phase factor, and the fact that $2j$ is an odd number, imply that states with odd values of J do not exist. One might wonder whether this is an essential property of fermions, but surprisingly it is not. Consider bosons with integer angular momentum ℓ. Then, owing to symmetrization, Eq. (1.36) becomes

$$\varphi(\ell,\ell;JM) = \frac{1}{2}\sum_{mm'}(\ell m\ell m'|JM) \times(\psi_1(\ell m)\psi_2(\ell m') + \psi_2(\ell m)\psi_1(\ell m')) \tag{1.38}$$

which yields after reordering,

$$\varphi(\ell,\ell;JM) = \frac{1}{2}\left(1+(-1)^{2\ell-J}\right)\sum_{mm'} \times(\ell m\ell m'|JM)\psi_1(\ell m)\psi_2(\ell m') \tag{1.39}$$

but now 2ℓ is even and again all states with odd J-values disappear. The fact that both bosons and fermions yield similar results is very important for nuclear models. It means that one can often describe even–even atomic nuclei using either fermionic nucleons or collective bosons such as phonons.

1.6.2
Regions of Different Level Structures

The existence of a low-lying 2^+ excitation can be explained in different ways, that is, short-range interaction, a collective quadrupole vibration, or as a first excited state of a rotational band. Which interpretation is right or better and which mixture of interpretations is right depends a lot on where on the Segre chart the nucleus is located and how the higher excitations behave. This is illustrated in Figure 1.5, which shows the known energies of the first excited states in the even–even Cd and Hf isotopes. One notices huge differences in the absolute excitation energies but also a quite smooth behavior. If one now also considers the next excited state, which is a 4^+ excitation and plots the ratio $R4/2$,

$$R4/2 = \frac{E_{\mathrm{exc}}(4^+)}{E_{\mathrm{exc}}(2^+)} \tag{1.40}$$

(see also Figure 1.5) more information can be obtained. In the Hf isotopes, a clear saturation near $R4/2 = 3.33$ occurs indicating that a rotational band, with its typical $I(I+1)$ energy dependence, is built on the ground state. For Cd, we observe very high 2^+ excitation energies and low $R4/2$ ratios at the beginning and end of the isotope chain. This is a clear evidence that at the extremes, a shell closure occurs, when $N = 50$ and 82. In the middle, one finds, compared to Hf, much higher 2^+ excitation energies and $R4/2$ ratios slightly above 2, indicating an anharmonic quadrupole vibrational nature at mid shell.

While the strong pairing of like nucleons can explain the ground state of even–even nuclei, it does not yield predictions for the ground-state spins and parities of odd-A and odd–odd nuclei. Nevertheless, it simplifies this task a lot, as in most cases one can conclude that the ground-state spin and parity of an odd-A nucleus is equal to the one of the last odd nucleon. This forms an important testing ground for the shell model, which predicts the single-particle energy, spin, and parity of the subsequent orbits.

1.6.3
Shell Structures

Various sets of nuclear data indicate that certain numbers of nucleons, either neutrons or protons, correspond to the filling of angular momentum “shells.” These are similar to the atomic shells often denoted as the K-shell, L-shell, and so on. In nuclei, the shell filling is similar but not identical, and the principal shell closings occur at experimentally observed numbers. These are for either *N or Z* equal to 2, 8, 20, 28, 50, 82, and 126. These are the *magic numbers*. They occur where there are drastic changes in neutron cross sections, nucleon separation energies, and so on. An interesting, but small, group of nuclei comprises the doubly magic ones – that is, nuclei with both neutron and proton numbers magic. The five stable ones are $^{4}\mathrm{He}_{2}$, $^{16}\mathrm{O}_{8}$, $^{40}\mathrm{Ca}_{20}$, $^{48}\mathrm{Ca}_{28}$, and $^{208}\mathrm{Pb}_{126}$. Note that no stable doubly magic nuclei with 50 and 82 exist. What is notable about the doubly magic nuclei is that their first excited states are at a very

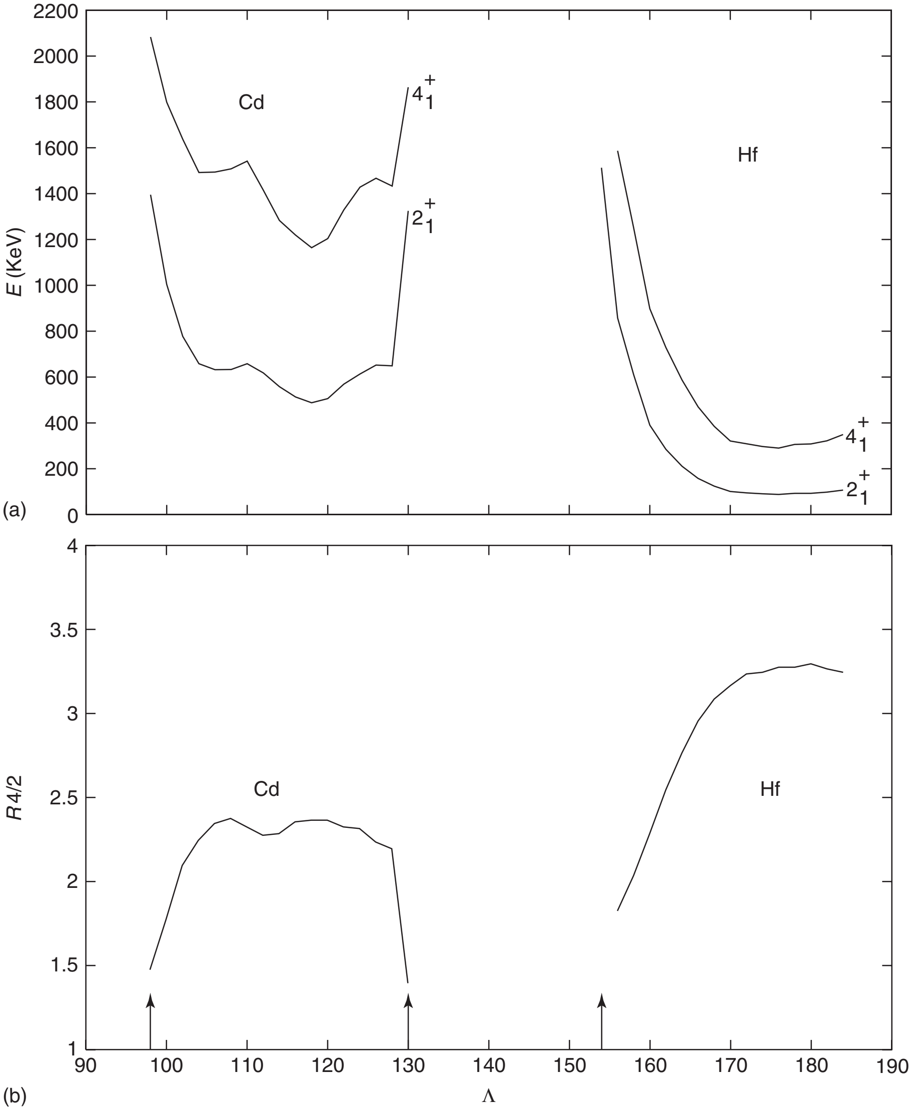

Figure 1.5 The known energies of the first excited states in the even–even Cd and Hf isotopes (a). The $R4/2$ ratios are also given (b). The left side represents the Cd isotopes and the right side the Hf isotopes. Arrows mark the closed shells at $A = 98$, 130, and 154.

high energy compared with their non-doubly magic neighbors. The energy in megaelectronvolts, spin, and parity of these first excited states is 20.1, 0^+; 6.05, 0^+; 3.35, 0^+; 3.83, 2^+; and 2.61, 3^-, respectively. Also the observed spins (and parities) of the first excited states is very different from that of all other even–even nuclei.

Because of these large gaps or changes in nuclear properties, they divide the low-lying excited states of nuclei into roughly three regions. These are, broadly, the nuclei in the neighborhood of the magic numbers that are dominated by shell or single-particle structures and those further away, which show collective behavior. Of particular interest are the ones found between these two, as they may give clues to the onset of collective motion in atomic nuclei and other systems. We mentioned earlier that 126 is a magic number, but strictly speaking, this holds only for neutrons as the unstable element with $Z = 126$ has not yet been observed. Nowadays, there is a strong interest to study whether magic numbers still stay valid in exotic nuclei far from stability and which is the next magic number for superheavy isotopes.

1.6.4
Collective Structures

Within major divisions (even–even, odd-A, and odd–odd), the excited levels are further divided into groups that are single-particle or collective (vibrational or rotational) in nature. However, because of the strong pairing force, even–even nuclei do not show single-particle excited levels but rather one- or two-particle-hole excitations or two-quasiparticle excitations. Here, we introduce the most general features of the two major classes of collective nuclei. We restrict ourselves to even–even nuclei for the sake of simplicity.

1.6.4.1 Vibrational Levels

The intermediate systems between the tightly bound magic nuclei and the deformed nuclei are the so-called vibrational nuclei. Here the number of nucleons outside of a deformable core is small, and the zero-point energy of the lowest oscillations is greater than the energy of deformation, so that the shape of the core is not stabilized. The motion then is of quantized surface oscillations with angular momentum two. One can introduce creation and destruction phonon operators Q^{+}_{LM} and Q_{LM} which fulfill the boson commutation rules:

$$[Q_{LM}, Q^{+}_{L'M'}] = \delta_{LL'}\delta_{MM'} \tag{1.41}$$

and for which all other commutators are zero. The angular momentum of the phonon is L, its magnetic projection M and its parity $\pi = (-1)^L$. The magnetic projection is mostly not of great importance. In this representation, the number operator counting the number of L phonons is given by

$$\widehat{N}_L = \sum_M Q^{+}_{LM} Q_{LM} \tag{1.42}$$

Using the number operator, the simplest Hamiltonian becomes

$$\widehat{H} = \sum_L \hbar\omega_L \left(\widehat{N}_L + L + \frac{1}{2}\right) \tag{1.43}$$

where the additional factors are chosen such that the Hamiltonian corresponds to the quantized harmonic oscillators for L phonons. Each of the phonons has energy $\hbar\omega_L$. Because most even–even nuclei have as first excited state a 2^+ state, the $L = 2$ quadrupole phonons are dominant. At higher energy, there is evidence for the presence of $L = 3$ octupole phonon making a 3^- state. Using the energy solution of the Hamiltonian equation (Eq. (1.43)),

$$E(N_L, L) = \sum_L \hbar\omega_L \left(N_L + L + \frac{1}{2}\right) \tag{1.44}$$

and the proper angular momentum coupling and symmetrization given in Eq. (1.39), one arrives at a very simple prediction. The one-quadrupole phonon state has $L=2$ and $E_{\text{exc}} = \hbar\omega_2$ and the two-phonon states have $L=0$, 2, 4 and $E_{\text{exc}} = 2\hbar\omega_2$. This exactly yields the $R4/2$ ratio of 2 observed in Figure 1.5.

A nice example of a ''good'' vibrational even–even nucleus is $^{110}\text{Cd}_{62}$. The first excited 2^+ state is at 0.6578 MeV followed by a 0^+, 2^+, 4^+ states at 1.473, 1.476, and 1.542 MeV. The ratio $R4/2=2.35$ is somewhat larger than the one the simple model gives.

1.6.4.2 **Rotational Levels**

If the number of nucleons outside the deformable core is such that the zero-point oscillations are much less than the energy of deformation, then the system will have a stable but deformed shape and one must quantize a rigid rotator. From classical considerations, we know that, if the system has a permanent, nonspherical shape, there exists a body-fixed system in which the inertial tensor $\mathfrak{I}$ is diagonal and is related to the laboratory-fixed system by an Euler transformation. We denote these two coordinate systems as (1,2,3) and (x,y,z), respectively. The inertial tensor then has components $\mathfrak{I}_1$, $\mathfrak{I}_2$, and $\mathfrak{I}_3$, so that the Hamiltonian is just

$$\widehat{H} = \frac{1}{2}\left(\frac{\widehat{L}_1^2}{\mathfrak{I}_1} + \frac{\widehat{L}_2^2}{\mathfrak{I}_2} + \frac{\widehat{L}_3^2}{\mathfrak{I}_3}\right) \tag{1.45}$$

with $\widehat{L}_{1,2,3}$ the body-fixed angular momentum. The simplest system is the one for which all moments of inertia are equal: $\mathfrak{I}_1 = \mathfrak{I}_2 = \mathfrak{I}_3 = \mathfrak{I}$. Then the energies are directly found:

$$\widehat{H} = \frac{\hbar^2}{2\mathfrak{I}}(L(L+1)) \tag{1.46}$$

The system specified by Eq. (1.45) possesses the symmetry properties belonging to the point group D_2 for which four representations exist. The Hamiltonian operator (Eq. (1.45)) does not mix different representations. The basis functions are those of a symmetric top, $|LMK>$ being diagonal in $\boldsymbol{L}^2$, $\boldsymbol{L}_z$, and $\boldsymbol{L}_3$ with the usual eigenvalues $L(L+1)$, M, and K, respectively. Here K is the projection of the angular momentum on the body-fixed frame. Applying this formula for the ground-state rotational band with $K=0$, we find now $R4/2=3.33$. However, states with different K-values are degenerated. This is not found in real deformed nuclei.

If the inertia tensor is such that $\mathfrak{I}_1 \neq \mathfrak{I}_2 \neq \mathfrak{I}_3$, an asymmetric top problem results. In this case, the wave functions are to be expanded in terms of the symmetric-top functions:

$$|LM\rangle = \sum_{K=-L}^{+L} A_K \, |LMK\rangle \tag{1.47}$$

The 1, 2, 3 labels of the momental ellipsoid's semiaxes are quite arbitrary and can be chosen in 24 different ways for right- (left-) handed systems. These 24 relabelings can be produced by three different relabeling transformations T_1, T_2, and T_3 [32], which correspond to label interchanges and operate on the $|LM>$ wave functions. They have the properties

$$T_1^2 = T_2^4 = T_3^3 = 1 \tag{1.48}$$

From these relations, one can generate Table 1.1, relating the four representations and the values of angular momentum and parity allowed in each. Note in particular that the A and B_2 representations are associated with positive-parity $(\pi=+)$ states, while B_1 and B_3 representations are associated with negative parity $(\pi=-)$

Table 1.1 The representation of the point group D_2 to which the states belong, and the allowed values of the angular momentum and parity associated with them.

Representation	K	Parity	Allowed L-values
A	Even	$+$	$L=0,2,2,3,4,4,4,5,5$
B_1	Even	$-$	$L=1,2,3,3,4,4,5,5,5$
B_2	Odd	$+$	$L=1,2,3,3,4,4,5,5,5$
B_3	Odd	$-$	$L=1,2,3,3,4,4,5,5,5$

states. From this table, we note that only the rigid rotator systems belonging to the A representation have a zero angular momentum state. They must be associated with the lowest levels of even–even nuclei and form the $K=0$ ground-state band with $K=0$ and $L=0, 2, 4, 6, \ldots$ followed by the γ-vibrational band with $K=2$ and $L=2, 3, 4, 5, \ldots$.

Negative-parity states in even–even nuclei could be associated with either the B_1 or B_3 representations. If one argues that, given two (or more) possibilities, the lowest set of states will be the most symmetric, then because the B_1 representation arises from sums over even K quantum numbers, one should associate this representation with the lowest-lying negative-parity states in even–even systems. Furthermore, in the deformed regions, the negative-parity bands in these nuclei not only lie above the ground-state rotational band but they also have as their band head a 1^- level. This is another indication that these negative-parity rotational bands belong to the B_1 representation.

The problem of solving the asymmetric, rigid rotor is far more involved than the symmetric rotor as all of the latter's eigenvalues can be given in closed form. To proceed along this line, one sets $\mathfrak{I}_0=\mathfrak{I}_1=\mathfrak{I}_2\neq\mathfrak{I}_3$. The Hamiltonian of the system then becomes

$$\widehat{H}_{\text{sym}}=\frac{1}{2}\left(\frac{\widehat{L}_1^2+\widehat{L}_2^2}{\mathfrak{I}_0}+\frac{\widehat{L}_3^2}{\mathfrak{I}_3}\right)=\frac{1}{2}\left(\frac{\widehat{L}^2}{\mathfrak{I}_0}+\left(\frac{1}{\mathfrak{I}_3}-\frac{1}{\mathfrak{I}_0}\right)\widehat{L}_3^2\right) \quad (1.49)$$

and both of the operators appearing are diagonal. Thus, we obtain immediately the eigenvalues

$$E_{\text{sym}}=\frac{h^2}{2}\left(\frac{L(L+1)}{\mathfrak{I}_0}+\left(\frac{1}{\mathfrak{I}_3}-\frac{1}{\mathfrak{I}_0}\right)K^2\right) \quad (1.50)$$

This expression is model independent and only depends upon the ratio of the moments of inertia $\mathfrak{I}_0/\mathfrak{I}_3$. There have been recent observations of "superdeformation" for rare earth nuclei with spheroidal-axes ratios of 2 : 1. For symmetric rotational systems with $K=0$, the ratio $R4/2$ is 10/3 rather than 2 for the vibrational systems. Thus, this ratio differentiates well between rotational and vibrational systems. In recent years, there was quite some interest in how these shape changes occur [33, 34], as they can be treated as quantum-phase transitions. Atomic nuclei that have an $R4/2$ ratio of around 2.9 lie at the critical point of the spherical-deformed phase transition.

1.6.5
Odd-A Nuclei

1.6.5.1 Single-Particle Levels

Near the magic numbers, the ground and lower excited states appear to be mainly single-particle states, that is, states that do not seem to possess collective properties. For instance, near the doubly magic nucleus $^{16}O_8$, one finds that both $^{15}N_8$ and $^{15}O_7$ have $1/2^-$ ground states and nearly identical lower excited states. This shows that the $N=Z=8$ shell is filled by a $p_{1/2}$ orbit. Above the closed shell are $^{17}F_8$ and $^{17}O_9$ both of which have the expected $5/2^+$ ground state from the $d_{5/2}$ orbit. These data clearly support the notion that angular momentum shells are filled in a manner similar to the electronic shells of the elements.

1.6.5.2 Vibrational Levels

The nucleons all possess intrinsic spin 1/2 and angular momentum l, so that the simplest model for vibrational odd-A nuclei that one might construct is to add a single nucleon to an even–even core in which the nucleons of each kind pair to zero angular momentum. If the core is a good vibrational nucleus ($R4/2 \sim 2$), then the ground state of the odd-A system will have the angular momentum properties of the odd nucleon. Furthermore, by coupling this nucleon to the excited states of the core, one should expect to be able to determine the angular momenta of the odd-A nucleus's excited states. An example is the coupling of a neutron to the vibrational nucleus $^{110}Cd_{62}$ (Section 1.6.2) to yield the odd-A nucleus $^{111}Cd_{63}$, which has a ground-state spin of $1/2^+$. One might expect that there would be two excited states built on the first excited 2^+ state with spins $3/2^+$ and $5/2^+$. This nucleus does indeed have a $3/2^+$, $5/2^+$ pair as its first excited states, which are at 245.4 and 342.1 keV. Coupling of the $1/2^+$ neutron with the two-phonon states would lead to excited state spins from $1/2^+$ through $9/2^+$. Such a set of states does appear at higher energies. However, it is difficult to pick out which states might belong to the two-phonon core states. Below them there is a first negative parity state with spin $11/2^-$ at 396.22 keV, which indicates that one has the presence of a low-lying $h_{11/2}$ neutron orbit.

1.6.5.3 Rotational Levels

One might expect that a simple model for the odd-A rotational levels would follow that of the vibrational levels. However, the experimental data do not show such a similar structure. As an example we consider $^{157}Gd_{93}$ which has a $3/2^-$ ground state. Following the reasoning for vibrational nuclei, one expects from the coupling with the first excited 2+ state a multiplet with spins $1/2^-$, $3/2^-$, $5/2^-$, and $7/2^-$. Instead, one observes a rotational-like band with the spin sequence: $5/2^-$, $7/2^-$, $9/2^-$, $11/2^-$, ... This is due to the fact that the ground-state band has also a given value of K and we conclude that $K=3/2$.

1.6.6
Odd–Odd Nuclei

As noted in Section 1.6.1, there are only four stable odd–odd nuclei; however, a great deal of data exist for all of this class of nuclei. Again, one might feel that these nuclei would be well represented by an even–even core plus a neutron–proton pair. Even if these are coupled to a spin-zero core, there is still considerable ambiguity as to how these couple to form the ground-state spin of the system. For instance, if the proton and neutron angular momenta are j_p, j_n, respectively, then their vector sum can lead to the range of J values

$$\left|j_{\rm p} - j_{\rm n}\right| \leq J \leq j_{\rm p} + j_{\rm n} \qquad (1.51)$$

One can obtain the values of $j_{\rm p}$ and $j_{\rm n}$ from the ground-state spins of the neighboring odd-A nuclei with one less neutron or one less proton, respectively. In order to determine more exactly the range of the possible ground-state spins of these nuclei, one must make use of Nordheim's rules, which are the following:Strong rule, for $\eta = 0$:

$$J = \left|j_{\rm p} - j_{\rm n}\right| \qquad (1.52)$$

Weak rule, for $\eta = \pm 1$:

$$J = \left|j_{\rm p} - j_{\rm n}\right| \ \text{ or } \ j_{\rm p} + j_{\rm n} \qquad (1.53)$$

with

$$\eta = (j_{\rm p} - l_{\rm p}) + (j_{\rm n} - l_{\rm n}) \qquad (1.54)$$

In general, these do not work all of the time. To predict the excited states spins is even more difficult, making odd–odd nuclei the most difficult to describe.

1.7
Nuclear Models

1.7.1
Introduction

The nucleus, being a many-body collection of A particles interacting through a short range but strong force, is not easily dealt with. From a theoretical point of view, making use of an accurate many-body calculation is impossible because these are beyond the scope of current theoretical methods. Thus, calculations based upon various simple models are able to provide a good deal of physical insight as well as making predictions of measurable properties, providing a way to categorize the vast amounts of data available from nuclear physics laboratories. While these approaches may seem to be organized in an historical order, all are to greater or less extent being continuously refined.

The nuclear many-body problem is not solved definitely, owing to the large number of strongly interacting particles. To illustrate the difficulties, we take the example of ^{208}Pb. If we consider up to two-body interactions between the nucleons, the general Hamiltonian is given by

$$H = \sum_{i=1}^{208} t_i + \frac{1}{2}\sum_{i \neq j}^{208} V_{ij} \qquad (1.55)$$

The radius of this nucleus is 7 fm, while the strong force between two nucleons varies significantly on the scale of 0.2 fm. If one would calculate only the two-body interaction on a mesh, one should foresee for each nucleon at least 100 mesh points in each direction and for all 208 nucleons the number of possible combinations would be $100^{3.208}$. For a fast computer today 10^9 operations per second are feasible, which gives 10^{16} operations per year. Thus to perform the calculation one would need 10^{1232} years, while the age of our universe is only 2×10^{10} years.

As an exact treatment is impossible, one of the major characteristics of nuclear structure calculations is the choice of the right approximations and truncations needed to make a calculation feasible while keeping the essential physics. Of great help thereby is the extensive use of symmetry concepts, which allow one to simplify the many-body problem and make the calculations tractable. Many quite different approaches have turned out to be successful for certain classes of atomic nuclei. Most theoretical models rely on spherical symmetry, which allow the use of the very powerful ‘‘Racah machinery,’’ originally developed for atomic physics by Guilio Racah. The introduction of irreducible tensors and coupling coefficients then provides the necessary technical skills to perform nuclear structure calculations.

The first approach is based on the experimental observation of magic numbers and leads to the nuclear shell model as described in Section 1.7.2. In Section 1.7.3 we introduce the deformed shell model, which allows the description of medium-heavy nuclei that are deformed. Section 1.7.4 is devoted to the collective model in which medium heavy and heavy nuclei are described by a quantum liquid droplet.

In Section 1.7.5, we will deal with a new theoretical approach to many-body problems originating from nuclear physics: the interacting boson approximation (IBA). Although the IBA is a method that was developed to describe the atomic nucleus, it has since been applied to molecules, quarks, and fullerenes. The major assumption of the interacting boson model (IBM) is twofold. First, because of the shell structure evinced by many experiments, only the valence nucleons are considered to be important for the low-energy excited states of the atomic nucleus. Those move to a first approximation in orbits formed by their commonly created average field, which is spherical symmetric. Secondly, between like nucleons, either protons or neutrons, there is a dominant pairing force when they occupy the same orbit. This allows to replace coupled fermion pairs by real bosons that are simpler to deal with.

1.7.2
The Spherical-Shell Model

One of the first and best understood nuclear models is the nuclear shell model. The major assumption of this model is that nucleons up to a good approximation move independently in a spherically symmetric averaged field $U(r_i)$ created by the other nucleons. This means that the many-body Hamiltonian is rewritten as

$$H = H_0 + V_{\text{res}} \tag{1.56}$$

with

$$H_0 = \sum_{i=1}^{N} t_i + \sum_{i=1}^{N} U(r_i) \tag{1.57}$$

and the residual interaction V_{res} is given by

$$V_{\text{res}} = \frac{1}{2}\sum_{i\neq j}^{N} V_{ij} - \sum_{i=1}^{N} U(r_i) \tag{1.58}$$

The shell model uses the ansatz (Eq. (1.57)) in which the single particles move in a potential well provided by the other particles. While it might seem improbable that nuclei made up of a large number of strongly interacting particles could in any way be represented by such a model, one must recall that the Pauli principle prevents most interactions of a nucleon in the nucleus. In general, scattering cannot take place, as most of the states into which a low-energy nucleon can scatter are filled. This implies that the mean free path for nucleon motion is quite long, and a single-particle model is worth exploring.

One obvious and important characteristic of odd-A nuclei is their ground-state spins and parities. Starting with the lightest nuclei, the general order of odd-A ground-state spins and parities is $1/2^{+}$ for neutron and proton and after the strongly bound ^{4}He $3/2^{-}$, $1/2^{-}$ then, after $^{16}O_8$, it becomes $5/2^{+}$, $3/2^{+}$, $1/2^{+}$ and then, after $^{40}Ca_{20}$, $7/2^{-}$, $5/2^{-}$, $3/2^{-}$, $1/2^{-}$ an order reminiscent of the atomic shells in which the alternating parities are related to the orbital angular momentum suggesting a sequence: s-p-d and s-f and p to which a spin one-half is coupled. To obtain a model with such an ordering of angular momentum l-values requires solving a single-particle Schrodinger equation with a reasonable potential function. The form of this potential is not too important because only minor changes in l-value order occur using one or another r dependence. The two simplest to calculate are the isotropic harmonic oscillator and the infinite square well. The latter is not realistic because the nucleon separation energy is quite finite and well-known for most nuclei. (One might also consider a potential function related to the charge distributions of Eqs. (1.16) and (1.19); however, these would require extensive numerical computing to extract the eigenvalues.) The angular momentum order of the levels and the number of particles in each for the isotropic harmonic oscillator is 1s (2); 1p (6), 1d, 2s (12); If, 2p (20); lg, 2d, 3s (30); and so on. (The semicolons separate the levels with different quanta of energy $\hbar\omega$, and most are degenerate.) The total number of levels thus becomes 2, 8, 20, 40, 70, 112, 168, which only agrees with the lowest observed "magic numbers" 2, 8, 20, 28, 50, 82, and 126 (cf. Section 1.6.1).

It took the genius of Maria Goeppert Mayer [35] and others [36] to realize that by adding a spin–orbit term of the form

$$V_{\mathrm{ls}} = -V(r)\,\vec{l}\cdot\vec{s} \tag{1.59}$$

the theoretical order could be brought into agreement with that observed. The minus sign guarantees that the $j = l + 1/2$ level will lie below the $j = l - 1/2$ level, as observed and the dependence on l makes that the effect increases with l leaving the shell closures at 2, 8, 20 unaffected while introducing a new shell closure at 28 due to the lowering of the $1f_{7/2}$ orbit. Further, the observed closures at 50, 82, 126, are obtained in a natural way. Also the spins of the odd-A ground states is mostly correctly predicted. The shell model explains via large gaps between the single-particle orbital energies the large binding energies of nuclei having Z or N a magic number making the number of stable isotopes/isotones particularly large, for example, there are 10 stable Sn isotopes with $Z = 50$. For the excited states, magic nuclei show them at very high excitation energies as one needs to break the closed shell and lift one particle over the large shell gap. As an example in the doubly magic nucleus ^{208}Pb, an energy of 2.6 MeV is needed to create the first excited 3^{-} state and even 4.085 MeV for the 2^{+} state.

The robustness of double-magic nuclei allows a major truncation for shell model calculations. It turns out that they can in first instance be treated as inert cores, whereby the spins of all of the particles in the core pair off to zero. Thereby, only the so-called valence nucleons need to be considered. This is of importance as with increasing number of nucleons and possible orbitals, the shell model allows for many different configurations. As angular momentum and parity are conserved, the residual interaction used can be diagonalized separately for a given spin and parity. Nevertheless, the dimensions become quickly huge.

Although the shell model is one of the older nuclear models, it has become of increasing use in the last decades. This is due to two key aspects. The first is the tremendous increase in computing power allowing large-scale shell model codes to be developed, including quantum Monte Carlo diagonalization. The second is the construction of effective *in medium* interactions starting from the nucleon–nucleon scattering data. Hereby the elimination of the hard core repulsion of the nucleon–nucleon interaction was crucial and produced a universal, so-called $V^{\text{low-}k}$ interaction. The recent progress opens the direct way from quantum chromodynamics (QCD) to the effective interactions, which is in particular very much needed for the study of exotic nuclei. The most involved calculations are done for light nuclei up to ^{11}B. Here no-core shell model calculations are able to describe accurately binding energies and excited states for very light nuclei [37].

1.7.3
The Deformed Shell Model

As the very large quadrupole moments of the nuclei in the middle of the higher shells cannot be explained by the simple spherical model, their explanation requires large contributions from the even–even core, which carries all (or almost all) of the charge. It can be shown that a single-particle moving in a nonrigid potential well will have a lower energy if the well is not spherical, rather than if it is spherical. As the lowering of the particle energy is proportional to the eccentricity, such a system will assume a deformed equilibrium shape. The nuclear Hamiltonian will now contain a term that produces deviations from spherical:

$$H = H_0 + H_\text{d} - V_\text{ls}(r) + \text{c}l^2 \tag{1.60}$$

where H_0 is the isotropic harmonic oscillator Hamiltonian and $V_\text{ls}(r)$ is the spin–orbit term, both of which have proved so successful for the spherical-shell model. (The l^2 term helps give the proper level order in the isotropic limit.) Expanding the nuclear surface in spherical harmonics gives

$$S(\theta,\phi) = R_0 \left[a_0 + \sum_{\lambda>1,\mu} a_{\lambda\mu} Y_{\lambda\mu}(\theta,\phi) \right] \tag{1.61}$$

where R_0 is the radius of the spherical nucleus and a_0 is unity to quantities of second order and assures that the volume remains constant for small deformations. Taking the lowest order, one has that

$$S(\theta,\phi) = R_0 \left[1 + \sum_{-2}^{+2} a_{2\mu} Y_{2\mu}(\theta,\phi) \right] \tag{1.62}$$

[This expression should be compared with the nuclear charge distribution of Eq. (1.19).] It is customary to take only the symmetric term a_{20} in this equation to be nonzero. This then leads to the deformation term in the Hamiltonian

$$H_\text{d} = -Ka_{20}\omega_0^2 r^2 Y_{20}(\theta,\phi) \tag{1.63}$$

The full Hamiltonian can now be diagonalized using a set of spherical shell model basis states, and the resulting deformed, single-particle states are called *Nilsson* states after Sven Gosta Nilsson [38], who first carried out this diagonalization. The operators in Eq. (1.60) do not connect oscillator states differing by one principal oscillator quantum number, N, but do connect states differing by two. Nilsson neglected matrix elements not diagonal in N. This is a reasonably good assumption as oscillator states differing by $N = 2$ are rather far apart, except for states belonging to large N and large deformation (i.e., nuclei with large A and large quadrupole moments). The Nilsson levels are labeled $\Omega|N,n_3,\Lambda]$. The symbol Ω is the projection of the particle angular momentum j on the three axes of the body-fixed axis system and is a good quantum number. In the limit of very large deformations, the single-particle energy is given by

$$E = \left(n_3 + \frac{1}{2}\right)\hbar\omega_3 + (n_\perp + 1)\hbar\omega_\perp \tag{1.64}$$

where the symbol ω_3 is the oscillator frequency along the symmetry axis and the symbol $\omega_\perp$ is the oscillator frequency in the perpendicular directions. Thus, in the limit of very large deformations, one has $N = n_3 + n_\perp$. In this same limit, the projection of the orbital angular momentum on the symmetry axis becomes the last label,

$$N = \pm n_\perp, \pm n_\perp - 2, \ldots, \pm 1 \text{ or } 0 \tag{1.65}$$

This scheme is used to label not only the ground states of deformed nuclei but also the excited state rotational band heads. The level order will now be L, $L+1$, $L+2$, …, as a result of core rotation, so that, when a level does not fit into this sequence, it must belong to a different single-particle band. One now proceeds to assign quantum numbers to the states by determining the odd-nucleon number; then, because the Nilsson energy levels are plotted against the core deformation, one assigns the deformation by finding that place where a Nilsson level appears with the measured ground-state spin. The next Nilsson level up should give the excited-band head, and so on. As an example, consider the nucleus $^{173}\mathrm{Lu}_{102}$, which has a $7/2^+$ ground state. Following the procedure above, the appropriate level is the 7/2[404], with excited members $9/2^+$, $11/2^+$, $13/2^+$. Just above the $9/2^+$ level is a close doublet with spins $5/2^-$, $1/2^-$, which has been assigned to the 1/2[541] Nilsson level. The $3/2^-$ level belonging to this band lies above the $9/2^+$ state. This nonuniform order of a 1/2 band is an example of Coriolis coupling, which mixes the level order of $\Omega = 1/2$ bands.

1.7.4
Collective Models of Even–Even Nuclei

In Section 1.6.4.2, where the rotational-level structures of even–even nuclei were discussed, it was pointed out that the moments of inertia play a crucial role to describe the atomic nucleus as a rigid-body momental ellipsoid. Bohr [32] proposed a model of the momental ellipsoid and thus of deformed even–even nuclei and the cores of odd-A and odd–odd nuclei. Several important assumptions were made in order to extract these moments of inertia. The first is that the nuclear core is incompressible, so that the nuclear density is constant. Second, the flow is assumed to be irrotational, so that a velocity potential exists and satisfies Laplace's equation. Finally, it must be assumed that the motion is such as to preserve the principal axis system. The moments of inertia for quadrupole deformations then are

$$\Im_k^2 = 4B_2\beta^2\sin^2\left(\gamma - \frac{2\pi}{3}k\right) \tag{1.66}$$

Here B_2 and β are the quadrupole mass and deformation parameters, while γ is an asymmetry parameter whose range is $0 \le \gamma \le \pi/6$. There are similar relations for other deformations such as octupole ($Y3$), hexadecapole ($Y4$), and so on. The Hamiltonian for this system is now

$$H_{\text{quad}} = \frac{1}{2}B_2(\dot{\beta}^2 + \beta^2\dot{\gamma}) + \frac{1}{B_2\beta^2}\sum_{k=1}^{3}\frac{L_k^2}{\sin^2\left(\gamma - \frac{2\pi}{k}\right)} + \frac{1}{2}C_2\beta^2 \tag{1.67}$$

If the system is not rigid, then the deformation and asymmetry parameters are no longer fixed, and the system can execute either deformation (β) vibrations or asymmetry (γ) vibrations, or both. The β-vibrational bands have angular momentum 0 – they are akin to molecular "breathing" modes – and the γ-vibrational bands have angular momentum 2. Thus, nuclei will exhibit ground-state bands where $n_\beta = 0$ and $n_\gamma = 0$, β-bands $n_\beta = 1$, $n_\gamma = 0$), γ-bands ($n_\beta = 0$, $n_\gamma = 1$), and higher order bands. Although the β-vibrations are the simplest, they are not really well established [39], much better established are γ-bands of which double excitations with $n_\gamma = 2$ are found in nature [40].

In his original paper, Bohr [32] argued that the nucleus would stabilize around $\gamma = 0$, so that the momental ellipsoid was symmetric (long and thin, cigarlike), while the surface itself was spheroidal. Then $\mathfrak{I}_3 = 0$. This leads to difficulties as Eq. (1.50) shows that the energy eigenvalues become infinite unless $K = 0$. Thus, the ground-state bands of even–even nuclei have $K = 0$, and the angular momentum sequence becomes 0, 2, 4, 6, . . . , which is what is observed. These models automatically conform to the A representation. Again, if the system is not rigid, against β-vibrations, there will be one or more β-bands above the ground-state rotational band that mimic the ground-state band in spin sequence, but because $<\beta^2>$ will be greater than for the ground state, the levels will be closer together.

Finally, in this model if γ is not constant, γ-vibrations can arise with angular momentum 2. (One might think of these as equatorial bulges moving around the nuclear equator. They are akin to the well-known Jacobi ellipsoids of a self-gravitating, rotating fluid.) These γ-vibrations then produce a γ-band that is different in spin sequence from the β- and ground-state bands. This occurs because $K_\gamma = 2$, not 0. The γ-band sequence is 2, 3, 4, 5, ... There can be several γ-bands with different K values and different spin sequences. The lowest one has $K_\gamma = 2$, while the next will have $K_\gamma = 0$ and 4, and so on.

It must be made clear that the Bohr model is only applicable for the low-lying positive-parity levels. The negative-parity levels belonging to the B_1 representation have moments of inertia that are considerably more complicated as they represent an octupole or pearlike shape.

An example of a well-developed rotational system is $^{154}\text{Sm}_{92}$ where $R4/2 = 3.25$. This nucleus has a well-developed ground-state rotational band with the spin sequence 0^+, 2^+, 4^+, . . . , 12^+ with the first excited state rather low in energy at 81.99 keV. It has a β-band with a 0^+ band head 1099.28 keV, followed by the first excited 2^+ state at 1177.78 keV ($\Delta E = 78.52$ keV). This is a striking example of the increase in the moment of inertia due to the quantum of β-vibrational energy. Furthermore, the $K = 2$ γ-band, with sequence 2^+, 3^+, 4^+, . . . , starts at 1440.05 keV.

Also to be seen is an octupole band starting with the 1^- level at 921.39 keV, followed by the 2^-, 3^-, 5^-, . . . in the expected order. Thus, one sees four well-developed rotational bands fitting quite nicely into the model scheme.

Further, the collective behavior manifests itself in the quadrupole moments and electric quadrupole transitions. The axially symmetric core collective model yields for a state with quantum numbers K and I a laboratory quadrupole moment Q:

$$Q = \frac{3K^2 - I(I+1)}{(I+1)(2I+3)} Q_0 \tag{1.68}$$

with an intrinsic quadrupole moment Q_0 given by

$$Q_0 = \frac{3}{\sqrt{5\pi}} Z R_0^2 \beta (1 + 0.16\beta) \tag{1.69}$$

The fact that by definition K has to be lower than I (Eq. (1.74)) produces almost always a change of sign between intrinsic and laboratory quadrupole moments. Related to the intrinsic quadrupole moments are the $B(E2)$ values. In between states belonging to a given K band, they are given by

$$B(E2;\, I_i \to I_f) = \frac{5}{16\pi} e^2 Q_0^2 \left\langle I_i K 20 | I_f K \right\rangle^2 \tag{1.70}$$

When comparing Eq. (1.70) directly with the data, one often extracts the transition quadrupole moment Q_t as being the Q_0 value obtained for a given transition. A very useful relation is obtained for $B(E2)$ ratios between states belonging to bands having different K quantum numbers. These relations are called the *Alaga rules* and yield

$$\frac{B(E2;\, I_i \to I_f)}{B(E2;\, I_i \to I'_f)} = \left(\frac{\left\langle I_i K 2 \Delta K | I_f K_f \right\rangle}{\left\langle I_i K 2 \Delta K | I'_f K_f \right\rangle} \right)^2 \tag{1.71}$$

whereby $\Delta K = K_f - K$ as is expected for an angular momentum projection quantum number. Using the Alaga rules it is possible to determine the K values of rotational bands.

If one does not follow the Bohr assumption that the nucleus will stabilize around $\gamma = 0$ but around some other value within its range, then one has to deal with an asymmetric rotator, which will have a considerably more complicated eigenvalue structure. Davydov and Filippov [41] made an extensive investigation and showed that one obtained better results with such a model than with a symmetric one. However, one loses the computational simplicity of the symmetric model.

1.7.5
Boson Models

The models discussed up until now and that have been used most extensively would seem to be almost mutually exclusive. The shell model and its modifications, such as Nilsson's extension to a deformed system, focus on the single-particle aspects of nuclear structure. On the other hand, the collective model of Bohr and Mottelson, and extensions, place the emphasis on the cooperative, or fluid, aspects of these systems. Thus, they seem to stand apart and be almost unrelated. This has led to the next step, the investigation of models that are called *microscopic models* or *calculations*. These include pairing models (similar to the Bardeen–Cooper–Schriefer theory of superconductivity) and the random-phase approximation, to name but two. A more recent and still developing set of microscopic models is called the *interacting boson models*, which we present here briefly.

The solutions of the nuclear shell model lead to single-particle orbits that can be occupied by the neutrons and protons. Because of the Pauli principle, each nucleon in an atomic nucleus can occupy only different orbits. It turns out that once more the short range of the nuclear force leads to a dominant component when one wants to take account of the residual interaction. The dominant component is called *pairing interaction* and energetically favors the formation of nucleon pairs made out of two nucleons moving in orbits |njm> and |nj-m>, coupling to a two-nucleon state with total spin $J = 0$. To a lesser extent, the formation of nucleon pairs with spin $J = 2$, 4, 6 in that sequence is also favored. As the number of protons and neutrons is generally different in not too light nuclei, the single-particle orbits occupied by both will generally be very different and thus the pairing will be effective only between like nucleons (either protons or neutrons). When one is mainly interested in the low-energy excitations of the atomic nucleus, one can treat this structure in terms of interacting, nucleon pairs via the residual interaction. However, a many-body problem composed of paired fermions, although more correct, is still very difficult to solve. Instead by replacing these $n = n_\nu + n_\pi$ fermions by

$$N = \frac{n_\nu + n_\pi}{2} \tag{1.72}$$

real bosons with $L = 0$, 2 (and eventually 4, 6, ...), one has enormously simplified the original n-body problem. For instance, one does not need to bother about which orbits are occupied by the protons and neutrons, while keeping the essential physics. This is the essence of what is now called the *interacting boson approximation*, which was proposed and elaborated by Iachello and Arima in the second half of the 1970s [42].

The simplest version of the model is called *IBM-1* and deals with s ($L = 0$) and d ($L = 2$) bosons. Considering the constraint that the number of bosons is fixed $N = n_\mathrm{s} + n_\mathrm{d}$ and limiting the Hamiltonian to one- and two-body interactions, one obtains the multipole expansion form

$$\begin{aligned}\widehat{H} = \varepsilon \widehat{n}_d + c_1 \widehat{L} \cdot \widehat{L} + c_2 \widehat{Q} \cdot \widehat{Q} \\ + c_3 \widehat{T}_3 \cdot \widehat{T}_3 + c_4 \widehat{T}_4 \cdot \widehat{T}_4\end{aligned} \tag{1.73}$$

with the operators

$$\widehat{L}_m = \sqrt{10} (d^\dagger \times \widetilde{d})^{(1)}_m \tag{1.74a}$$

$$\begin{aligned}\widehat{Q}_m = \quad & (s^\dagger \times \widetilde{d} + d^\dagger \times s)^{(2)}_m \\ & + \chi (d^\dagger \times \widetilde{d})^{(2)}_m\end{aligned} \tag{1.74b}$$

$$\widehat{T}_{3m} = (d^\dagger \times \widetilde{d})^{(3)}_m \tag{1.74c}$$

$$\widehat{T}_{4m} = (d^\dagger \times \widetilde{d})^{(4)}_m \tag{1.74d}$$

The factor $\sqrt{10}$ is used in order to make the dipole operator an angular momentum operator and the first term in Eq. (1.73) contains the d-boson number operator. The multipole expansion has six free parameters of which two are needed for the quadrupole–quadrupole interaction. The use of the multipole expansion is especially advantageous because numerical studies have shown that the values of the octupole and hexadecapole interaction are generally very small. Thus, up to a good approximation, one can describe a variety of atomic nuclei at low excitation energies by the simple four-parameter Hamiltonian:

$$\widehat{H}^\chi = \varepsilon \widehat{n}_d + c_1 \widehat{L} \times \widehat{L} + c_2 \widehat{Q}^\chi \times \widehat{Q}^\chi \tag{1.75}$$

The index in Eq. (1.75) makes explicit that the quadrupole operator contains an additional parameter.

Of importance in the development of the IBM is the use of dynamical symmetries providing analytic solutions of the many-body problem. Because there are in total six distinct bosons when considering the magnetic projections, the symmetry of the many-body problem is $U(6)$ while the physical symmetry is given by SO(3) the Lie algebra of the angular momentum. Three dynamical symmetries are obtained. They can be schematically represented by the chains

$$U(6)\left\{\begin{array}{l}\left.\begin{array}{l}\supset U(5)\\ \supset O(6)\end{array}\right\} \quad \supset O(5)\\ \supset SU(3)\end{array}\right\} \supset O(3) \tag{1.76}$$

Each of the three chains can be used to provide a complete basis to analytically solve the N s, d boson model. One can also rewrite the general Hamiltonian as a combination of all previously determined Casimir operators:

$$\begin{aligned}\widehat{H} = {} & \varepsilon C_1[U(5)] + \alpha C_2[U(5)] \\ & + \delta C_2[SU(3)] + \eta C_2[O(6)] \\ & + \beta C_2[O(5)] + \gamma C_2[O(3)]\end{aligned} \tag{1.77}$$

where we have omitted all constant terms that contribute only to the binding energies. The Casimir form Eq. (1.77), like Eq. (1.73), contains six parameters and, once more, all three forms can be transformed into each other. The advantage of the form Eq. (1.77) lies in the fact that the three different limits correspond to three classes of atomic nuclei: vibrational, rotational, and γ-unstable nuclei correspond to the Hamiltonian with $\delta = \eta = 0$ ($U(5)$ limit), $\varepsilon = \alpha = \eta = \beta = 0$ (SU(3) limit) and $\varepsilon = \alpha = \delta = 0$ ($O(6)$ limit) . In these cases, closed analytic expressions for excitation energies and wavefunctions can be obtained [42]. Two of the limits already had known counterparts in the collective model. It was the discovery of the third, the $O(6)$ limit, in the Pt nuclei at the end of the 1970s [43], that boosted the use of the model. Soon it became evident that it formed a good approximation for the nuclear many-body problem in many nuclei. Outside the limits the model needs to be solved numerically and the transitional nuclei can be described.

The IBM-1 model describes only positive-parity levels of even–even nuclei. To describe negative-parity states, one must use negative parity p and f bosons constructed in a similar way to the s and d bosons. Furthermore, the IBM-1 has no explicit neutron–proton degree of freedom. This is introduced in the IBM-2 where a neutron–proton label is introduced for the s and d bosons. As the neutron and proton bosons are distinguishable, the IBM-2 leads to new sets of states, called *mixed symmetry states* (in the neutron–proton degree of freedom). Furthermore, analytical solutions can be derived for the IBM-2 [44]. Mixed symmetry states have been observed in many atomic nuclei, especially in deformed nuclei where the lowest state is a 1^+ state called the *scissor state* [45], and also recently in vibrational nuclei [46].

Finally, if one wants to describe odd-A nuclei, a model denoted interacting boson–fermion model (IBFM) must be used. Odd-A nuclei always have an unpaired nucleon that cannot form an s or d boson because a partner is missing. Therefore, an odd–even nucleus is described in the model with N bosons and one fermion, the unpaired nucleon, which can occupy several orbitals j. In addition, for such systems dynamical symmetries, called *Bose–Fermi symmetries*, can be used, although their elaboration is much more complex [47]. The symmetry starts then from the product $U^{\mathrm{B}}(6) \times U^{\mathrm{F}}(M)$ with

$$M = \sum_j 2j + 1 \tag{1.78}$$

and ends with SU(2). In 1980, a daring extension of the model was proposed [48]. Using supersymmetry, the structure of some odd–even nuclei could be related

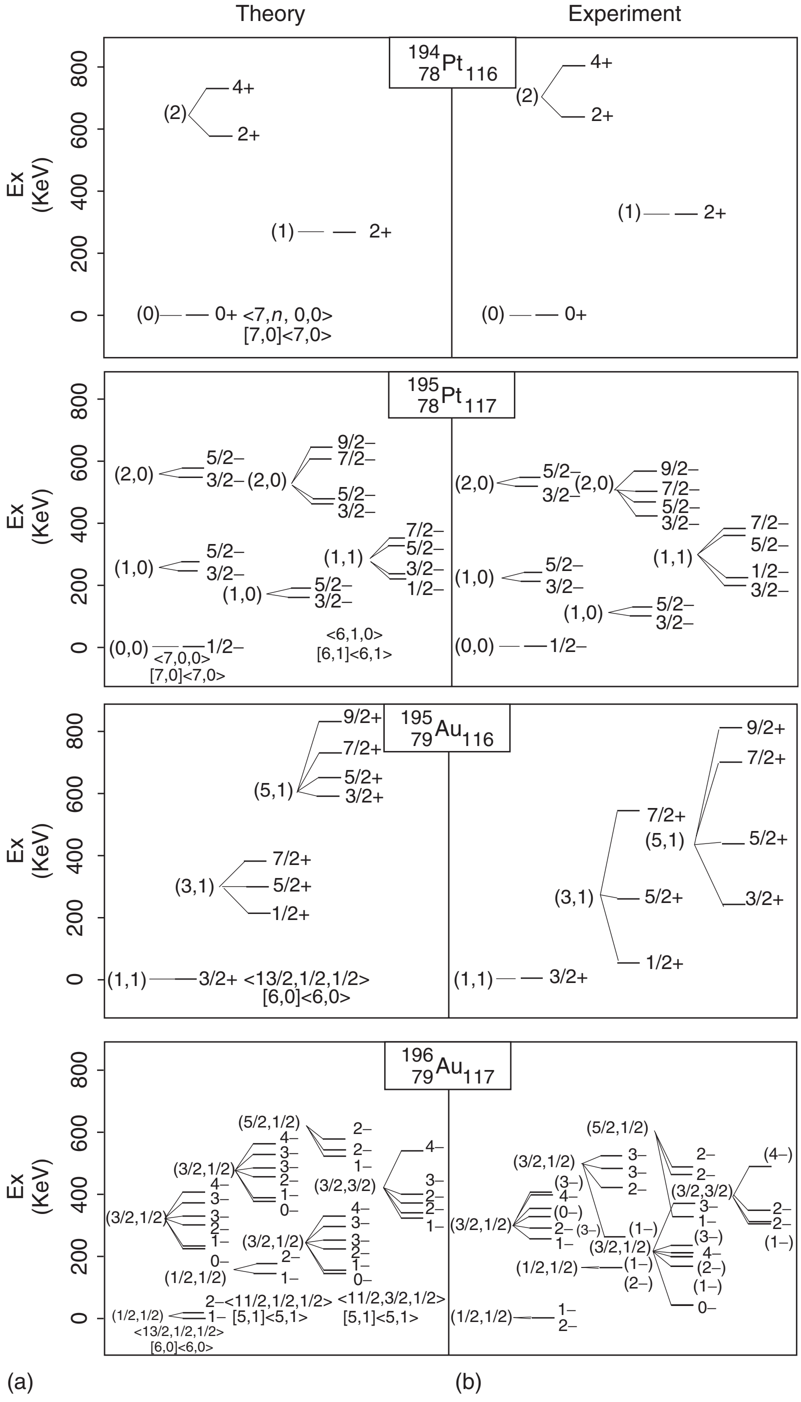
Theory
Experiment
Ex (KeV)
0
200
400
600
800
$^{194}_{78}Pt_{116}$
(2)
4+
2+
(1)
2+
(0)
0+ <7,n, 0,0>
[7,0]<7,0>
(2)
4+
2+
(1)
2+
(0)
0+
$^{195}_{78}Pt_{117}$
(2,0)
5/2–
3/2–
(2,0)
9/2–
7/2–
5/2–
3/2–
(1,0)
5/2–
3/2–
(1,1)
7/2–
5/2–
3/2–
1/2–
(1,0)
5/2–
3/2–
(0,0)
1/2–
<7,0,0>
[7,0]<7,0>
<6,1,0>
[6,1]<6,1>
(2,0)
5/2–
3/2–
(2,0)
9/2–
7/2–
5/2–
3/2–
(1,1)
7/2–
5/2–
1/2–
3/2–
(1,0)
5/2–
3/2–
(1,0)
5/2–
3/2–
(0,0)
1/2–
$^{195}_{79}Au_{116}$
(5,1)
9/2+
7/2+
5/2+
3/2+
(3,1)
7/2+
5/2+
1/2+
(1,1)
3/2+ <13/2,1/2,1/2>
[6,0]<6,0>
(5,1)
9/2+
7/2+
5/2+
3/2+
(3,1)
7/2+
5/2+
1/2+
(1,1)
3/2+
$^{196}_{79}Au_{117}$
(5/2,1/2)
2–
2–
1–
(3/2,1/2)
4–
3–
3–
2–
1–
0–
(3/2,3/2)
4–
3–
2–
2–
1–
(3/2,1/2)
4–
3–
3–
2–
1–
0–
(3/2,1/2)
4–
3–
3–
2–
1–
0–
(1/2,1/2)
2–
1–
2–<11/2,1/2,1/2>
[5,1]<5,1>
<11/2,3/2,1/2>
[5,1]<5,1>
(1/2,1/2)
1–
<13/2,1/2,1/2>
[6,0]<6,0>
(5/2,1/2)
(3/2,1/2)
3–
3–
2–
2–
2–
(4–)
(3/2,3/2)
2–
2–
(1–)
(3–)
4–
(0–)
2–
(3–)
1–
(1–)
(3/2,1/2)
3–
1–
(3–)
4–
(2–)
(1–)
0–
(3/2,1/2)
(1/2,1/2)
(1–)
(2–)
(1/2,1/2)
1–
2–
(a)
(b)

to that of the much simpler even–even nuclei via the embedding of the dynamical symmetries in a supersymmetry:

$$\begin{array}{ccccc} U(6/M) & \supset & U^{\mathrm{B}}(6) & \times & U^{\mathrm{F}}(M) \\ | & & | & & | \\ [N] & & [\mathrm{N}] & & [1^{m}] \end{array} \qquad (1.79)$$

connecting the N s, d bosons + 1 fermion problem to the N + 1 boson problem by the supersymmetric reduction rule $N = \mathrm{N} + m$. This embedding when realized in nature would reveal itself through the possibility to describe excited states in an odd–even nucleus and in the adjacent even–even nuclei using a common set of quantum numbers and parameters related via the supergroup. If dynamical symmetries are required, only a limited class of atomic nuclei can be described in this way. During the 1980s, experimental evidence was accumulated, but unfortunately it was not possible to determine completely the structure of the odd–even nucleus starting from the even–even nucleus alone.

Later an extension of supersymmetry was proposed, which incorporates a distinction between neutrons and protons. This extended supersymmetry allows one to describe a quartet of nuclei in a common framework. This quartet, called a *magic square*, consists of nuclei having the same total number of bosons (paired nucleons) and fermions (unpaired nucleons). It consists of an even–even nucleus, two odd–even nuclei and an odd–odd nucleus. The latter is interesting in two respects: its energy spectrum can be predicted from those of the other three members and heavy odd–odd nuclei are the most complex objects found in low-energy nuclear structure. If dynamical supersymmetry is able to describe these nuclei, which could not be described by other theoretical models, strong evidence for its existence can be obtained. This development is of importance not only for nuclear physics but for all other applications of supersymmetry in physics.

The odd–odd nucleus ^{196}Au is considered to be the ultimate test of supersymmetry in nuclear physics for three reasons. It is situated in a region of nuclei that was known to exhibit dynamical symmetries and supersymmetries. At the same time, it is the most difficult mass region in which to describe odd–odd nuclei. Finally, when its excitations with negative parity were predicted in 1989 on the basis of the three other nuclei, none of them were experimentally known. In a detailed spectroscopic study the excited states of ^{196}Au were determined and a good agreement with the supersymmetric model was obtained [49]. The results of this combined effort are illustrated in Figure 1.6. They reveal good agreement with the analytical prediction:

Figure 1.6 The lowest observed states (b), are compared to the theoretical result (a) for the four members of the quartet (note that one is dealing with holes in the doubly closed shell nucleus ^{208}Pb). The even–even nucleus ^{194}Pt, the negative parity states in the odd-neutron nucleus ^{195}Pt, the positive parity states in the odd-proton nucleus ^{195}Au and the negative-parity states in the odd–odd nucleus ^{196}Au up to 500 keV can be described using the formula (Eq. (1.88)) based on dynamical supersymmetry. Not shown are higher-lying states of the even–even and odd–even nuclei, which are also described by supersymmetry. Together with the spin and parities, the quantum numbers $[N_1,N_2]$, $<\Sigma_1,\Sigma_2>$, $<\sigma_1,\sigma_2,\sigma_3>$, and (τ_1,τ_2) are also indicated. (Based on [49].)

$$\begin{aligned}E = {} & A(N_1(N_1+5)+N_2(N_2+3)) \\ & +B(\Sigma_1(\Sigma_1+4)+\Sigma_2(\Sigma_2+2)) \\ & +B'(\sigma_1(\sigma_1+4)+\sigma_2(\sigma_2+2)+\sigma_3^2) \\ & +C(\tau_1(\tau_1+3)+\tau_2(\tau_2+1)) \\ & +DL(L+1)+EJ(J+1) \qquad (1.80)\end{aligned}$$

obtained using dynamical supersymmetry.

Glossary

Bosons: Quantum particles obeying Bose–Einstein statistics, so having integral spin.

Daughter: In a radioactive decay in which the radioactive nucleus A decays into nucleus B the nucleus B is said to be the daughter of nucleus A.

de Broglie wavelength: The wavelength λ of a particle given by the relation $\lambda = hp$, where h is Planck's constant and p the momentum of the particle.

Decay constant λ: If N nuclei are present at time t, then the decay constant is given by the number of decaying nuclei $-\mathrm{d}N/\mathrm{d}t$ divided by N.

Electron capture: A radioactive decay process, sometimes called *K capture*, where the decay energy Q is greater than the binding energy of one of the atom's cloud of electrons ($Q > BEe$). The nucleus absorbs one of the atomic electrons to undergo β^+ decay and emits thereby a neutrino. If $Q > 2m_\mathrm{e}c^2$, then electron capture competes with normal β^+ decay, whereby a positron and an electron neutrino are created.

Fermions: Half integer spin particles obeying Fermi–Dirac statistics.

Half-life $T_{1/2}$: In radioactive decay, the time in which half of the nuclei initially present decay.

Isobars: Nuclei with the same mass number A but with different numbers of protons (Z) and neutrons (N).

Isomeric state: An excited nuclear state whose half-life for γ-emission is quite long; similar to a metastable state of an atomic system.

Isotones: Nuclei with the same number of neutrons (N) but different numbers of protons (Z) and thus, different values for A.

Isotope shift: Small changes in the wavelengths of X-ray optical, and especially muonic transitions in going from one isotope to the next, which give a measure of the change in the nuclear radius as A changes by one unit.

Isotopes: Nuclei with the same number of protons (Z) but different numbers of neutrons (N) and thus, different values of A.

Lamb shift: A small quantum electrodynamic effect that is principally due to the radiative coupling of the orbiting electron (or muon) with the vacuum field and the finite nucleus.

Lifetime τ: The reciprocal of the decay constant. The lifetime is related to the half-life by $T_{1/2} = \ln(2)\tau$.

Mass defect: The difference between the mass of a nucleus of mass number A in atomic mass units less A, $\Delta = M(A) - A$.

Mass excess: Negative of *mass defect*.

Mirror pair: Two light nuclei with the same mass number A but with numbers of protons (Z) and neutrons (N) interchanged.

Muon: A member of the lepton family of elementary particles with spin one-half. The muon is 207 times as heavy as an electron and has a lifetime of

2.197×10^{-6} s. The muon, like the electron, has a negative charge.

Nuclear polarization: A small reduction in the Coulomb potential caused by the penetration of the nuclear volume by the bound orbiting particle (an electron or muon).

Packing fraction: The mass defect per unit mass number $P = \Delta/A$.

Parity π**:** The behavior of a state function upon reflection of the coordinates through the origin, $r \rightarrow -r$, then either $\phi(-r) = +\phi(r)$ and the state is said to have *positive parity* or $\phi(-r) = -\phi(r)$ and the state is said to have *negative parity*.

Pauli principle: The requirement that in a system of like particles obeying Fermi–Dirac statistics (half integer spin particles), no two particles can have the same set of quantum numbers. More generally, the Pauli principle requires wavefunctions for identical particles to be antisymmetric for fermions and symmetric for bosons upon interchange of two particles.

Phonon: In nuclear physics a phonon is a quantized surface vibration of a quantum fluid.

Vacuum polarization: A small radiative correction to the Coulomb potential of an atom arising from the emission and reabsorption of virtual positron–electron pairs.

Valley of stability: In a three-dimensional plot of the nuclear masses $M(Z,N)$ with the proton number Z as abscissa neutron number N as ordinate and $M(Z,N)$ normal to the Z–N-plane, the stable nuclei will be found along a region where $N \sim Z$, which appears to form a deep valley – *the valley of stability*.

Woods–Saxon potential: For a real potential function, this is written with an adjustable surface diffuseness parameter and a radius parameter R. For a complex potential, one adds a similar term, V_{Im}.

References

1. Rutherford, E. (1911) *Philos. Mag.*, **21**, 669–688.
2. Moseley, H.G.J. (1913) *Philos. Mag.*, **26**, 1024–1034.
3. Moseley, H.G.J. (1914) *Philos. Mag.*, **27**, 703–713.
4. Soddy, F. (1913) *Nature*, **92**, 399–400.
5. Rutherford, E. (1919) *Philos. Mag.*, **37**, 581–586.
6. Harkins, W.D. and Majorsky, S.L. (1922) *Phys. Rev.*, **19**, 135–156.
7. Pauli, W. (1924) *Naturwissenschaften*, **12**, 741–743. reprinted in Kronig, R., Weisskopf, V. F. (eds) (1964) *Collected Scientific Papers*, Vol. 2, New York: John Wiley & Sons, Inc., pp. 198–200.
8. Racah, G. (1931) *Z. Phys.*, **71**, 431–434.
9. Chadwick, J. (1932) *Nature*, **129**, 312.
10. Heisenberg, W. (1932) *Z. Phys.*, **77**, 1.
11. Bethe, H.A. and Bacher, R.F. (1936) *Rev. Mod. Phys.*, **8**, 82–229.
12. Bethe, H.A. (1937) *Rev. Mod. Phys.*, **9**, 69–244.
13. Avignone, F.T. III, Elliott, S.R., and Engel, J. (2008) *Rev. Mod. Phys.*, **80**, 481–516.
14. Lederer, C.M. and Shirley, V.S. (eds) (1978) *Table of Isotopes*, 7th edn, John Wiley & Sons, Inc., New York.
15. Kudomi, N., Ejiri, H., Nagata, K., Okada, K., Shibata, T., Shima, T., and Tanaka, J. (1992) *Phys. Rev.*, **46**, R2132–R2135.
16. Bohr, A. and Mottelson, B. (1969) *Nuclear Structure*, Vol. I, Benjamin, New York.
17. Myers, W.D. and Swiatecki, W.J. (1966) *Nucl. Phys.*, **81**, 1–60.
18. Myers, W.D. (1977) *Droplet Model of Atomic Nuclei*, IFI/Plenum, New York, and references therein.
19. Möller, P., Nix, J.R., Myers, W.D., and Swiatecki, W.J. (1995) *Atomic Data Nucl. Data Tables*, **39**, 185–225.
20. Duflo, J. and Zuker, A.P. (1995) *Phys. Rev. C*, **52**, R23–R27.
21. Goriely, S., Samyn, M., and Pearson, J.M. (2007) *Phys. Rev. C*, **75**, 064312.

22. Lunney, D., Pearson, J.M., and Thibault, C. (2003) *Rev. Mod. Phys.*, **75**, 1021–1082.
23. de Vries, H., de Jager, C.W., and de Vries, C. (1987) *Atomic Data Nucl. Data Tables*, **36**, 495–536.
24. Tanihata, I. (1995) *Prog. Part. Nucl. Phys.*, **35**, 505–574.
25. Casimir, H.B.G. (1935) *Physica*, **2**, 719–723.
26. Gollnow, G. (1936) *Z. Phys.*, **103**, 443–453.
27. Schwartz, C. (1955) *Phys. Rev.*, **97**, 380–395.
28. Castel, B. and Towner, I.S. (1990) *Modern Theories of Nuclear Moments*, Clarendon Press: Oxford, and references therein.
29. Neyens, G. (2003) *Rep. Prog. Phys*, **66**, 633–689.
30. Eberth, J. and Simpson, J. (2008) *Prog. Part. Nucl. Phys.*, **60**, 283–337.
31. Heyde, K.L.G. (1990) *The Nuclear Shell Model*, Springer Series in Nuclear and Particle Physics, Springer(A).
32. Bohr, A. (1952) *Kgl. Dan. Vid. Selsk. Mat. Fys. Medd.*, **26** (14).
33. Jolie, J., Cejnar, P., Casten, R.F., Heinze, S., Linnemann, A., and Werner, V. (2002) *Phys. Rev. Lett.*, **89**, 182502.
34. Cejnar, P., Jolie, J., and Casten, R.F. (2010) *Rev. Mod. Phys.*, **82**, 2155–2212.
35. See also: Mayer, M. G. (1948) *Phys. Rev*. **74**, 235–239; *Phys. Rev*, (1949), **75**, 1969–1970.
36. Haxel, O., Jensen, J.H.D., and Suess, H.E. (1949) *Phys. Rev.*, **75**, 1766.
37. Pieper, S.C., Varga, K., and Wiringa, R.B. (2002) *Phys. Rev.*, **C66**, 044310.
38. Nilsson, S.G. (1955) *Kgl. Dan. Vid. Selsk. Mat. Fys. Medd.*, **29**, 16.
39. Garrett, P.E. (2001) *J. Phys.*, **G27**, R1.
40. Börner, H.G. *et al.* (1991) *Phys. Rev. Lett.*, **66**, 691 and 2837.
41. Davydov, A.S. and Filippov, G.F. (1958) *Nucl. Phys.*, **8**, 237–249.
42. Iachello, F. and Arima, A. (1987) *The Interacting Boson Model*, Cambridge University Press, Cambridge.
43. Casten, R.F. and Cizewski, J.A. (1978) *Nucl. Phys.*, **A309**, 477.
44. Van Isacker, P., Heyde, K., Jolie, J., and Sevrin, A. (1986) *Ann. Phys.*, **171**, 253–296.
45. Richter, A. (1995) *Prog. Particle Nucl. Phys.*, **34**, 261.
46. Pietralla, N., von Brentano, P., and Lisetskij, A.G. (2008) *Prog. Particle Nucl. Phys.*, **60**, 225–282.
47. Iachello, F. and Van Isacker, P. (1991) *The Interacting Boson-Fermion Model*, Cambridge University Press, Cambridge.
48. Iachello, F. (1980) *Phys. Rev. Lett.*, **44**, 772–775.
49. Groeger, J. *et al.* (2000) *Phys. Rev. C*, **62**, 064304.

Further Readings

In this list, the symbols E, I, and A following the references indicate the level of difficulty: elementary, intermediate, and advanced, respectively.

Bethe, H.A. (1936) *Rev. Mod. Phys.*, **8**, 82–229 (I).

This and the preceding form the best review of the early state of the subject: Bethe, H.A. and Bacher, R.F. (1937) *Rev. Mod. Phys.*, **9**, 69–244(I).

Bohr, A. and Mottelson, B. (1969) *Nuclear Structure*, Vol. 1, Benjamin, New York(A).

This and the preceding, while old, are comprehensive and still useful: Bohr, A. and Mottelson, B. (1975) *Nuclear Structure*, Vol. 2, Benjamin, New York(A).

Bonatsos, D. (1988) *Interacting Boson Models of Nuclear Structure*, Clarendon, Oxford(A).

Castel, B. and Towner, I.S. (1990) *Modern Theories of Nuclear Moments*, Clarendon, Oxford(A).

An excellent text, more advanced than Pal (1983): Casten, R.F. (1990) *Nuclear Structure from a Simple Perspective*, Oxford University Press, New York(I).

An overview of the use of symmetry concepts in nuclear structure with many experimental examples: Frank, A., Jolie, J., and Van Isacker, P. (2008) *Symmetries in Atomic Nuclei: From Isospin to Supersymmetry*, Springer Tracts in Modern Physics, Springer (A).

A nice introduction into the practical aspects of the nuclear shell model: Heyde, K.L.G. (1990) *The Nuclear Shell Model*, Springer Series in Nuclear and Particle Physics, Springer (A).

An excellent introduction for students of which each new edition is substantially updated: Heyde, K.L.G. (1994) *Basic Ideas and Concepts in Nuclear Physics*, Fundamental and Applied Nuclear Physics, IOP (I).

Discusses many aspects, both theoretical and experimental, of this fascinating area of nuclear structure:
Janssens, R.V.F. and Khoo, T.L. (1991) Superdeformed nuclei. *Annu. Rev. Nucl. Part. Sci.*, **41**, 321–355 (A).

An ideal textbook for students:
Krane, K.S. (1987) *Introductory Nuclear Physics*, John Wiley & Sons, Inc., New York (I).

A compendium of nuclear data, of great value:
Lederer, C.M. and Shirley, V.S. (eds) (1978) *Table of Isotopes*, 7th edn, John Wiley & Sons, Inc., New York (A).

Nice treatments of several of the older microscopic models:
Pal, M.K. (1983) *Theory of Nuclear Structure*, Van Nostrand Reinhold, New York (E).

Very deep and detailed introduction in the algebraic methods used in the nuclear shell model and the interacting boson model:
Talmi, I. (1993) *Simple Models of Complex Nuclei, Contemporary Concepts in Physics*, Harwood Academic Publishers, Chur (A).

Another excellent, more advanced text:
Wong, S.S.M. (1990) *Introductory Nuclear Physics*, Prentice-Hall, Englewood Cliffs, NJ (I).

NUCLEAR WASTE MANAGEMENT

J. A. Klein, *Chemical Technology, Oak Ridge National Laboratory,* Oak Ridge, Tennessee, U.S.A*

Introduction 1
1. **Classification of Radioactive Waste** 2
1.1 Spent Fuel 2
1.2 High-Level Waste 2
1.3 Transuranic Waste 2
1.4 Intermediate- or Medium-Level Waste 2
1.5 Low-Level Waste 3
1.6 Mixed Waste 3
1.7 Uranium Mill Tailings 3
2. **Spent-Fuel Management** 3
2.1 Spent-Fuel Storage 3
2.2 Spent-Fuel Packaging and Transportation 3
2.3 Spent-Fuel Disposal 4
2.4 Representative Spent-Fuel Repository Designs 5
2.4.1 United States 5
2.4.2 Sweden 6
2.4.3 Germany 6
2.4.4 Canada 6
3. **High-Level Waste Management** 6
3.1 High-Level Waste Treatment 7
3.2 High-Level Waste Disposal .. 7
4. **Transuranic Waste Management** 7
4.1 Processing and Handling of Transuranic Waste 7
4.2 Waste Isolation Pilot Plant (WIPP) 8
5. **Low-Level Waste Management** 9
5.1 Low-Level Waste Generation 9
5.2 Low-Level Waste Treatment 9
5.2.1 Dewatering Low-Level Waste 10
5.2.2 Thermal Transformation 10
5.2.3 Sorting and Segregation 10
5.2.4 Decontamination 10
5.2.5 Mechanical Volume Reduction 10
5.2.6 Stabilization and Solidification 10
5.3 Low-Level Waste Disposal Methods 10
5.3.1 Above-Ground Disposal 11
5.3.2 Near-Surface Underground Disposal 11
5.3.3 Intermediate-Depth Disposal 11
5.3.4 Deep Disposal in Geologic Repositories 11
5.3.5 Ocean Disposal 12
5.4 Low-Level Waste Disposal in the United States 13
6. **Mixed-Waste Management** .. 14
6.1 Availability of Mixed-Waste Treatment Facilities 14
7. **Uranium Mill Tailings Management** 15
8. Quantities of Radioactive Waste 16
8.1 International Radioactive Waste Generation 16
8.2 U.S Radioactive Waste Generation 16
Glossary 17
Works Cited 17
Further Reading 18

INTRODUCTION

Because nuclear (radioactive) waste decays to background levels of radioactivity with time, the ultimate goal of radioactive waste management is quite obvious—store the material until it is no longer harmful. The difficulty with this approach is in the implementation of the theory. What does "no longer harmful" mean in both technical and public perception terms? What are the "best" ways to store material safely and in a cost-effective

---

*Managed by Martin Marietta Energy Systems, Inc. for the U.S. Department of Energy under Contract DE-AC05-84OR21400.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

manner? Only recently has public insistence been directed toward efforts to find environmentally acceptable methods of managing radioactive waste. This insistence has led to a more nearly complete understanding, by both scientists and policymakers, of the importance of developing cost-effective technologies to dispose of radioactive waste.

1. CLASSIFICATION OF RADIOACTIVE WASTE

A number of different schemes exist for the classification of radioactive waste. Some of these are based on the origin of the waste (defense/nondefense wastes, medical wastes, utility wastes, nuclear fuel cycle wastes, etc.); the type of material present in the waste (transuranic, mixed wastes, etc.); the level of radioactivity (high, low, intermediate, etc.); the form of the waste (dry solid, wet solid, liquid, etc.); and, in some cases, legally mandated definitions that may or may not be identical to a straightforward technical or intuitive definition. An example of the latter is the presence in the United States of waste that is defined as low-level but which may have a relatively high level of radioactivity. A general classification of the various radioactive waste categories currently used in various radioactive-waste–producing countries follows.

1.1 Spent Fuel

Spent fuel consists of irradiated fuel assemblies discharged from a nuclear reactor. Spent fuel may be stored at the reactor for reuse in the future, scheduled for reprocessing at some later date, or considered to be eligible for either permanent or temporary storage at a geological repository. In some definitions, especially if long-term disposal is being considered, the spent fuel may also be considered high-level waste. Spent fuel may derive from either commercial (utility) or noncommercial (research or defense) reactors. Spent fuel is usually highly irradiated (high activity), but it may be irradiated to a lesser degree.

1.2 High-Level Waste

High-level waste is usually defined as highly radioactive material resulting from the reprocessing of spent nuclear fuel. This material includes primarily the liquid waste remaining from the recovery of uranium and plutonium in a fuel reprocessing plant. High-level waste may also be in the form of sludge, calcine, vitrified glass, or other products into which such liquid waste is converted to facilitate its handling, storage, or disposal. High-level waste contains fission products that result in the release of considerable decay energy. For this reason, heavy shielding is required to control penetrating radiation, and provisions (e.g., cooling systems) are often needed to dissipate decay heat from the high-level waste.

1.3 Transuranic Waste

Only the U.S. Department of Energy (DOE) manages material specifically classified as "transuranic waste." In the United States, this term refers to radioactive waste that contains more than 100 nCi/g of alpha-emitting isotopes with atomic numbers greater than 92 and half-lives greater than 20 years. Such waste results primarily from fuel reprocessing and from the fabrication of plutonium weapons and plutonium-bearing reactor fuel. Generally, little or no shielding is required ("contact-handled" transuranic waste), but energetic gamma and neutron emissions from certain transuranic nuclides and fission-product contaminants may require shielding or remote handling ("remote-handled" transuranic waste). Although the United States is one of the few countries that recognize transuranic waste as a separate waste classification, any country with a plutonium nuclear weapons program or extensive spent-fuel reprocessing activities will generate similar material. Depending on the country and the specific radioactivity characteristics, this waste may be managed as either low- or high-level waste. Outside the United States, waste of this type is usually categorized as intermediate- or medium-level waste, but it also may be called transuranic-bearing waste or any of a number of similar terms.

1.4 Intermediate- or Medium-Level Waste

Intermediate- or medium-level waste is generally a non-U.S. category for waste that contains higher levels of radioactivity than low-level wastes and so requires some shielding. The boundary between this category and low-level waste varies depending on the country. Treatment and disposal of intermediate-level waste vary depending on the waste form and whether it contains short- or long-lived radioactive isotopes. In general, short-lived intermediate-level waste can be disposed of as low-level waste, but long-lived intermediate-level waste will need to be disposed of similarly to high-level waste.

1.5 Low-Level Waste

Low-level waste is generally considered radioactive waste not classified as spent fuel, high-level waste, transuranic waste, or any other specific category of radioactive waste. The radiation levels from this type of waste are generally quite low, although occasionally the radiation levels may be high enough to require shielding for handling and transport.

1.6 Mixed Waste

Mixed waste contains concentrations of both radioactive materials and nonradioactive hazardous chemicals. All radioactive waste forms may have the potential for being classified as mixed waste, but usually only mixed low-level radioactive waste requires the consideration of management practices separate from low-level radioactive waste. Management practices for spent fuel, high-level waste, and transuranic waste are generally stringent enough to minimize any problems arising from the hazardous components. The United States considers both the radioactive and the hazardous components as separate materials requiring independent regulation. However, most European communities treat low-level mixed waste as if it were strictly low-level waste. For them, the disposal risks for the hazardous components are less than the risks associated with the radioactive components.

1.7 Uranium Mill Tailings

Commercial uranium mill tailings are the earthen residues that remain after the extraction of uranium from ores. Tailings are generated in very large volumes and contain low concentrations of naturally occurring radioactive materials. Because they provide a potential health hazard, the isotopes that concern us most are ^{226}Ra and its daughter, ^{222}Rn.

2. SPENT-FUEL MANAGEMENT

Historically, storage facilities for both commercial and noncommercial spent fuel were limited in capacity. It was generally understood that the spent fuel would be stored only on-site for a limited time and then shipped to a reprocessing facility. However, only a few countries have reprocessed any spent fuel; and although a limited number of countries have decided to reprocess all their spent fuel, many countries have not committed to either the reprocessing or the direct disposal of spent fuel. Currently, direct disposal appears to be gaining favor. Because the fabrication, irradiation, handling, storage, possible fuel reprocessing, and long-term management of spent fuel are all integral parts of the nuclear fuel cycle, these topics are discussed in more detail in NUCLEAR FUELS AND ISOTOPES.

2.1 Spent-Fuel Storage

Storage sites for spent fuel usually consist of deep-water pools; however, many of these pools are rapidly reaching capacity. Utilities in the United States are currently dealing with this problem by reracking spent-fuel storage modules in existing pools, by shipping the spent fuel to other reactor sites that have surplus storage capacity; and by studying alternative storage options, including the use of dry storage casks for the on-site storage of spent fuel. It may also prove feasible to reuse these casks for the future shipment of spent fuel to a centralized storage facility or repository if one becomes available. The pros and cons of on-site storage vs a centralized storage facility or repository are actively being debated in the United States and in many other international communities.

2.2 Spent-Fuel Packaging and Transportation

Because spent fuel might not be reprocessed and because, even if it is to be reprocessed, it is unlikely to be reprocessed at the generation site, the transportation of spent fuel is inevitable. Because of the great deal of attention directed to the shipping of spent fuel, the shipping casks or packages used to transport it must satisfy a variety of societal, political, and technical requirements. Foremost among the latter are requirements that will prevent the release of any radioactive material into the environment and that will provide for radiation and criticality protection. Because the construction of nuclear facilities, fuel fabrication, and shipment of spent fuel are international businesses, the rules and regulations affecting the shipment of it throughout the world are quite similar.

As spent-fuel packages must withstand serious accidents, the criteria of no loss of shielding and no loss of containment have been taken by the international community to mean that certain test conditions must be met by any package designed to transport spent fuel. These mandated requirements are remarkably similar throughout the world. For the United States, the U.S. Nuclear Regulatory Commission (NRC) has determined that spent-fuel packages must experience no loss of integrity upon

1. a 30-ft free-fall drop onto an unyielding surface;
2. a puncture test of at least a 40-in. free-fall drop onto a 6-in.-diameter steel pin;
3. exposure to a thermal environment of 1475 °F for 30 min; and
4. immersion in water for 8 h.

Because of the various regulations and design requirements, spent-fuel shipping casks are invariably made from steel or steel composite (with lead or uranium) thick enough to meet structural and gamma-shielding requirements. Additional elastomer or resin material may also be included for neutron shielding. An appropriately designed liner, specific to the spent fuel being transported, holds the assemblies in place and prevents damage during transport. These casks typically weigh between 25 and 100 metric tons with rail, barge, or ship transportation being required for the larger casks.

In deference to the current uncertainties of whether reprocessing will occur and to the uncertainties of on- or off-site storage of the waste, a number of spent-fuel casks are being designed to meet both storage and transport requirements. These casks are basically similar to the transportation casks described previously, but they have been modified as necessary to provide longer storage life.

2.3 Spent-Fuel Disposal

If the direct disposal of spent fuel is eventually determined to be the disposal method of choice, the fuel will be sent either immediately or after some period of interim storage to a geological repository. For those countries that must dispose of both spent fuel and high-level waste, these materials will be handled in a common repository with possible differences only in specific handling methods and the final repository cells. The differences are usually trivial. The following discussion pertains to both.

Inherently, all proposals for the safe disposal of spent fuel require the safe storage of radioactive isotopes until they decay to nonradioactive elements or to nonhazardous levels of radioactivity (as defined by the host country). Depending on the specific materials and radionuclides present, this requirement mandates storage times of thousands to millions of years. Because these times are much greater than the design lifetimes of conventional manufactured structures, alternative methods of disposal are often required. The obvious method of choice is a deep geological repository, especially in underground rock formations that have not changed for millions of years and thus are considered inherently stable.

Although candidate geological formations have isolated material for sufficiently long periods of time, representative repository designs generally include additional engineered barriers that add greater long-term assurances that material will not migrate out of the formation. These barriers may include long-lived container compositions, incorporation of the spent fuel into inert waste forms, and special backfill material in completed waste cells.

Similar geological repositories have already been used for low-level and hazardous waste. However, the disposal of spent fuel differs considerably in that the heat content of the waste packages (due to radioactive decay) may damage both the package and the structural integrity of the geological formation. Thus, spent-fuel repository designs call for the spreading out of the packages over large areas of the repository. In contrast, low-level or transuranic repository designs call for relatively large rooms filled with waste packages.

Once a suitable geological formation has been selected, underground space needs to be excavated along with suitable access shafts or tunnels. The excavation of the final waste storage areas may occur before storage or may occur simultaneously with storage, with new excavation occurring in areas remote from ongoing waste emplacement. Throughout the world, a number of different rock types are being considered for individual geological repositories; examples include salt, clay, crystalline granites, and tuff. None of these rock types can be called a perfect choice, because each type has its own inherent advantages and disadvantages. There are difficulties in mining, in isolating the waste from groundwater, and in the long-term stability of the formation factor in the choice of geologic formation. Salt is generally considered the easiest and fastest material to excavate, and granite, the slowest. Granite, however, is more stable and allows also for deeper repositories. None of the rock types is overwhelmingly superior for waste isolation. The total effectiveness of a repository is determined by the combination of a favorable geology and an engineering design that mitigates any disadvantages.

Before the emplacement of spent fuel within the repository, the waste will have to be packaged in final canisters or packages. Various options are being considered for the required packages. One option consists of having the spent fuel shipped to the repository in typical spent-fuel shipping casks and transferred to the final disposal package at the repository. Other designs incorporate the use of the actual shipping container as a self-shielded waste package for direct emplacement within the underground repository. The packaging of the spent fuel can be done at the repository, at the reactor before shipment, or at a centralized spent-fuel storage facility where the economics of scale may make packaging more cost effective.

The time interval from spent-fuel generation to actual emplacement within a geological repository may be short or may encompass several decades. No specific requirements determine the time frame involved, although interim storage costs may dictate shorter time intervals, while the desirability of taking advantage of the drop-off in decay heat with time favors longer times.

2.4 Representative Spent-Fuel Repository Designs

Although, currently, no country has an operating geological repository for spent fuel, a number of countries have repositories either under construction or far enough along in their design to permit a discussion of the proposed repositories. Several, though not all, of the various international designs are discussed to present some indication of the similarities and differences among the various designs. The information presented here is as reported in the open literature and is therefore not definitive, but rather gives a working understanding of the proposed designs (Forsberg, 1991). Several countries with major nuclear programs (e.g., France, the United Kingdom, and Japan) have not yet published any detailed plans on their respective repository programs. In a majority of countries, the final decision whether to reprocess spent fuel has not been made. Thus, most of the repository designs are flexible enough to allow for the disposal of spent fuel or high-level waste.

2.4.1 United States The United States is considering building its first repository for both spent fuel and high-level waste in a tuff (fused volcanic deposits) formation at the Yucca Mountain Repository Site in Nevada. The current design includes provisions for disposing of commercial and defense-related spent fuel, high-level waste from defense activities, and a small amount of high-level waste from the historical commercial reprocessing of spent fuel. This repository will have a capacity of some 70 000 metric tons initial heavy-metal equivalent (MTIHM). Typical depths for placement of packages into this repository will be 300 m. The spent-fuel or high-level waste packages (identical in outer dimensions) are to be placed in widely separated bore holes placed in the floor of the disposal tunnels. These tunnels are to be arranged in a series of panels (compartments) with one panel receiving waste, a second ready to receive waste, and a third under construction. Because of heat dissipation requirements, the bore holes are to be widely spaced. The U.S. design calls for the potential retrieval of the waste if the repository does not prove suitable. This requirement has resulted in engineered designs for bore-hole liners and other handling and emplacement components to allow for future retrieval.

2.4.2 Sweden The Swedish design for the final disposal of spent fuel is unique in that it relies heavily on the waste package to provide the primary barrier to environmental releases. The reference waste package design consists of a solid copper package with the spent-fuel assemblies embedded within it. These packages are to be placed 500 m underground into crystalline rock (granite, gneiss, or gabbro). This unique design is based on the knowledge that native copper in granitic formations is quite stable for long geological periods of time. Copper can deform over time; therefore, the package fabrication procedure involves embedding the spent-fuel assemblies within a solid copper cylinder, filling the remaining void spaces with copper powder, and hot-pressing the entire package to produce a single solid copper cylinder. Because copper in nature has been shown to be stable at relatively low temperatures, current plans call for a minimum of 30 to 40 years of interim storage before final packaging to allow for the drop-off in decay heat. Operation is expected around the year 2020.

2.4.3 Germany The German design calls for the joint disposal of high-level waste and spent fuel from reactors of western design at a depth of some 870 m in the Gorleben salt dome. The high-level waste will result from the reprocessing of German spent fuel in France and Great Britain. Direct emplacement of spent fuel will result from unique situations that would make reprocessing difficult or very expensive. Although current German law calls for the reprocessing of spent fuel, direct spent-fuel emplacement will also occur if the reprocessing of reactor fuel is terminated. For spent fuel generated at Soviet-designed reactors within Germany, the preferred option is to return the spent fuel to Russia for reprocessing. The German design calls for the use of the Pollux cask concept, which includes the option for use as a transport cask, for long-term interim storage, or as a final disposal package. The Pollux design provides excellent radiation shielding, thus permitting the direct emplacement of packages at the end of tunnels, backfilling around the package with salt, and repeating the process until the tunnel is full. Salt slowly creeps (deforms) with time and, thus, over tens of years, the tunnels will be self-sealing. Current plans call for initial emplacement of material in the year 2008.

2.4.4 Canada One Canadian design for a spent-fuel repository is based on the direct disposal of spent fuel 500 to 1000 m underground in plutonic formations sometime after the year 2025. Because the Canadian commercial reactor program is based on CANDU reactors that use natural uranium fuel, several unique factors are involved in the Canadian repository design. The low burnup of CANDU fuel implies very large volumes of spent fuel with low decay-heat releases from the spent-fuel packages. The combination of these factors is a reference Canadian repository design of over 190 000 MTIHM capacity, but with larger spent-fuel waste packages and closer spacing of waste packages in the repository than occur with other countries' repository designs.

3. HIGH-LEVEL WASTE MANAGEMENT

High-level waste, which is generated by the reprocessing of spent nuclear fuel and irradiated targets, generally contains more than 99% of the nonvolatile fission products produced in the fuel or targets during reactor operation. The high-level waste from a facility that recovers uranium and plutonium contains about 0.5% of these elements, while the high-level waste from a facility that recovers only uranium contains about 0.5% of the uranium and essentially all of the plutonium. Most of the current U.S. inventory of high-level waste is the result of defense-related programs and is stored at DOE sites. A small amount of commercial high-level waste was generated during the period 1966–1972. Other countries that have or plan to have a significant commercial fuel reprocessing capability (i.e., France and Great Britain with the world's only commercial reprocessing facilities) might well have a different mix of commercial and defense-program-related high-level waste. The production of high-level waste as a waste product is an inherent part of the nuclear fuel cycle, and, as such, the various spent-fuel reprocessing technologies that generate high-level waste are discussed in detail in NUCLEAR FUELS AND ISOTOPES.

3.1 High-Level Waste Treatment

A number of highly publicized recent events in the United States concerning the existence of leaking liquid high-level waste storage tanks highlight the potential limitations and inherent difficulties in the long-term storage of liquid high-level waste. Although the tanks in question were of the old single-shell design, a number of countries (e.g., France, Australia, and the United States) have embarked on vigorous research and development efforts to stabilize high-level waste safely (American Nuclear Society, Inc., and American Society of Civil Engineers, 1992). These efforts are directed toward the conversion of liquid high-level waste to a form that simplifies the long-range waste management problems and often includes a major reduction in waste volumes. In the United States, high-level wastes are currently stored as neutralized sludge, salt, acid liquid, alkaline liquid, and calcine. Only the calcine is a form that can be stored safely in stainless steel bins in concrete vaults. Current plans call for all of the U.S. high-level waste eventually to be incorporated into a stable matrix using borosilicate glass in several different melters and processes. The French use a closely coupled calcination and borosilicate glass vitrification process that appears to be relatively flexible and inexpensive. Several other processes have also gathered attention on an international basis, including the Australian SYNROC process. SYNROC is basically a synthetic rock that consists of a number of titanate mineral phases. The main benefit of this process is that it offers exceptionally high resistance to leaching by groundwater. Although a number of countries have shown interest in alternative waste forms, only the borosilicate matrix has been, and is being, developed commercially.

3.2 High-Level Waste Disposal

To date, it appears that none of the major producers of high-level waste have directly disposed of any of this material. In numerous cases, this liquid waste has been stabilized and solidified by a number of processes, including evaporation, calcination, and vitrification of the waste into a glass matrix. Ultimately, high-level waste probably will be disposed of similarly to spent fuel in a geologic repository. Although a number of countries, including France and the United States, have some of their high-level waste in a form ready to be disposed of, none has apparently done so. The final repository designs generally call for the emplacement of high-level waste and spent fuel depending on each country's decision whether to reprocess spent fuel. These designs are discussed in Sec. 2.4.

4. TRANSURANIC WASTE MANAGEMENT

4.1 Processing and Handling of Transuranic Waste

Transuranic waste is generated primarily by activities associated with research and development, spent-fuel reprocessing, plutonium recovery, weapons manufacturing, environmental restoration, and decontamination and decommissioning projects. About half of the countries having nuclear reactors produce transuranic waste from at least one of these sources. Most transuranic waste exists in solid form (e.g., protective clothing, paper trash, rags, glass, miscellaneous tools, and equipment). However, some transuranic waste is in liquid form (sludges), the result of the chemical processing for recovery of plutonium or other transuranic elements. Before 1970, all U.S. DOE–generated transuranic waste was disposed of on-site in shallow landfill-type configurations. In 1970, DOE concluded that waste containing long-lived alpha-emitting radionuclides should have greater confinement from the environment. Thus, all transuranic waste generated in the United States since the early 1970s has been segregated from other waste types and placed in retrievable storage pending shipment and final disposal in a permanent geologic repository. This retrievable stored transuranic waste is contained in a variety of packages (metal drums and wooden and metal boxes) in earth-mounded berms, concrete culverts, or other types of facilities.

The majority (>90%) of transuranic waste contains primarily plutonium, which emits alpha particles and low-energy photons. Therefore, the packaging is designed to provide sufficient containment and shielding to minimize personnel exposure problems. This waste form is referred to as "contact handled." Less than 3% of transuranic waste contains activation materials and fission products that decay by beta emission and produce penetrating gamma and x-ray radiation. This waste is referred to as "remote handled" if the radiation level at the surface of the packaging exceeds 200 mrem/h.

It is estimated that as much as 50 to 60% of the transuranic waste is also mixed waste in that it contains hazardous constituents defined and regulated by the U.S. Environmental Protection Agency (EPA). Examples of transuranic mixed waste are radionuclide-contaminated spent solvents, discarded materials contaminated with both solvents and radioactive materials, scintillation fluids, and discarded contaminated lead shielding. However, the proposed disposal plans for transuranic waste are such that any concerns about the management of the hazard constituents should be minimal.

Current plans call for the retrievable stored transuranic waste and newly generated transuranic waste from U.S. defense-related activities to be shipped to the Waste Isolation Pilot Plant (WIPP), located near Carlsbad, New Mexico. The disposition of buried transuranic wastes and transuranic waste generated from nondefense activities is uncertain at this time.

4.2 Waste Isolation Pilot Plant (WIPP)

Except for those countries that manage transuranic or transuranic-like waste as low-level waste, most transuranic-waste–generating countries have adopted an interim storage concept with the final disposal as yet to be decided upon. In the past, Germany disposed of transuranic waste in the Asse salt mine, but it has discontinued that operation. Other countries are considering repository projects; for example, the United States is proceeding with the WIPP. The WIPP has been designed as a facility for the permanent isolation of U.S. defense-generated transuranic wastes, but it will also provide an underground test facility in which the safe disposal of high-level waste in salt can be demonstrated. Initial construction of the underground facility was begun in 1981. Currently, the WIPP facility is fully prepared to receive some transuranic waste to conduct various experiments related to regulatory compliance of the repository. The final decisions permitting the initial emplacement of transuranic waste have not yet been made.

Because of the public's and regulators' strong interest in the initial disposal of transuranic waste, a detailed set of waste acceptance criteria has been drafted. These rules are concerned with a wide variety of issues, including gas generation, combustibility, toxic and corrosive compounds, surface dose rate, surface contamination, and criticality. Generator, interim storage, and shipping sites must maintain detailed documentation to meet these criteria and be allowed to ship material to the WIPP.

Contact-handled transuranic waste will be transferred to the WIPP in specially designed and constructed packaging containers referred to as TRUPACT-II containers. Although designed for easily shielded contact-handled transuranic waste, the container consists of several sealed containment vessels and a sandwichlike foam and stainless steel shell that increases the package strength and safety to withstand postulated accidents associated with transport. TRUPACT-II has a capacity of 14 standard 55-gal drums and is about 3 m in diameter and 3 m tall. Three TRUPACT-IIs can be carried on a flatbed truck trailer. Shipment of the remote-handled transuranic waste will take place in a special shipping container that is somewhat similar to the various designs for a spent-fuel cask.

The WIPP site, which is located in the Permian Basin deposits of the southwestern United States, contains salt formations that are about 900 m thick, some of the thickest formations in the United States. Present mining activities take place 650 m below the surface near the middle of the salt bed. This formation was chosen because the area appears to be geologically stable with little earthquake activity, salt deposits demonstrate the absence of any circulating groundwater, salt is relatively easy to mine, and the salt has the ability to heal fractures and creep to fill voids and seal a waste repository. Aboveground facilities house the site personnel and equipment needed for WIPP operations and research activities. Included are separate areas for the contact- and remote-handled transuranic wastes. A series of four shafts connect the surface facilities to the underground area, making it possible to lower transuranic waste to its destination and to transport workers, equipment, mined salt, and fresh air.

Underground WIPP facilities consist of transuranic waste storage rooms 4 m high, 10 m wide, and 90 m long; these rooms are separated from each other by 30-m pillars. As one room is filled with waste, the next room is being excavated, and some of the salt rubble is used to backfill the voids in the room that is being filled. Contact-handled transuranic waste containers will be removed from the TRUPACT-II transporter and transferred to the underground storage area. There they will be moved to and stacked in the appropriate storage room. Remote-handled transuranic waste, however, will need to be handled by specially designed and shielded handling equipment. Actual waste containers will be remotely emplaced within predrilled holes in the storage room wall and immediately plugged. Other nonstorage rooms have been excavated for experimental high-level waste heat-dissipation tests and as shop and maintenance areas.

Near the end of the five-year demonstration and evaluation period, a decision will be made regarding permanent disposal of transuranic wastes at the WIPP. If a decision for permanent disposal is made, the facility will operate as a repository for an additional 20 years. The total anticipated volume of transuranic waste to be stored at the WIPP is about 200 000 m^3. Remote-handled transuranic waste is expected to require only 5000 m^3 of the total available volume.

Plans are to use both a traditional cement-based grout and natural salt plugs to cap the waste-filled underground rooms and to seal the vertical shafts. The cement grout provides a well-characterized short-term seal, and the salt plugs provide a deformable seal that will fuse with the surrounding formation. Plans are in place for the final decommissioning of the WIPP, including decontaminating the surface areas and dismantling of the surface facilities. Permanent surface markers will be placed on the site, and a long-term monitoring program will be conducted.

5. LOW-LEVEL WASTE MANAGEMENT

5.1 Low-Level Waste Generation

As indicated in the classification section, the usual way of determining what is included in low-level waste is to describe what it is not. In most countries, low-level waste is that radioactively contaminated material that is not covered by the definitions for spent fuel, high-level waste, transuranic waste, uranium mill tailings, or any other form of waste. Generally, low-level waste contains relatively small amounts of radioactivity, but occasionally it does have a level of radioactivity high enough to warrant special treatment and disposal. Principal generators of low-level waste include academic and medical institutions, nuclear power plants, and other industries. Of obvious importance is the generation of low-level waste in defense-related programs for those countries having such programs. A majority of low-level waste contains very low levels of radioactivity and generally requires no shielding to protect workers or the general public and decays in fewer than 100 years to levels that are generally considered acceptable. The remainder (3–5%) requires some degree of shielding and takes longer to decay to acceptable levels of radioactivity. This material may not be acceptable for near-surface disposal without including additional engineering barriers. The highest-activity low-level waste is often slated for disposal in a geologic repository along with spent fuel and high-level waste.

The types of low-level waste being generated include the widest range of materials, including contaminated trash, paper, glass, clothing, plastics, equipment, tools, sludges and slurries, and organic and aqueous liquids. Solids, liquids, or gases can all be included as low-level waste. Normally liquid wastes are stabilized or solidified before disposal, and gaseous wastes need to be filtered or scrubbed to remove contaminants before discharge to the environment.

5.2 Low-Level Waste Treatment

The best way to manage low-level waste is not to generate it. The use of substitute materials or good housekeeping practices can go a long way toward eliminating, or at least minimizing, the generation of low-level waste. Sometimes, however, it is not possible to eliminate a particular low-level waste; then various treatment methods have to be relied upon to reduce the volume or simplify the handling of the low-level waste. Numerous such technologies are available, and it would be impossible to describe all of them in detail. Generally, they can be grouped according to their principal objectives: dewatering, thermal transformation, sorting and segregation, decontamination, mechanical volume reduction, and stabilization and solidification.

5.2.1 Dewatering Low-Level Waste Liquid wastes can be concentrated and, thus, their volumes reduced by evaporation, incineration, ion exchange, filtration, reverse osmosis, and precipitation. Other technologies (e.g., distillation, ultrafiltration, centrifugation, and crystallization) can be and probably have been used to treat low-level waste, but the former are more commonly called upon.

5.2.2 Thermal Transformation Incineration of low-level waste is one of the two most widely used technologies for the volume reduction of solid low-level waste; the other technology is compaction. Incineration is also heavily used for a number of liquid low-level waste streams, especially those involving combustible organic materials. Some incineration of noncombustible liquid streams can be handled by mixing these streams with combustible low-level waste or by using a supplemental fuel. Although highly efficient (with reduction ratios of 100 to 1 being achieved for liquid low-level wastes), the capital costs are significant, and secondary waste streams are produced that have to be disposed of safely. These wastes include waste gases that need to be monitored and, if necessary, scrubbed. Fly ash and bottom ash are produced. The radioactivity present in the liquid low-level waste tends to concentrate in the bottom ash, and this material will need to be stabilized and disposed of safely.

5.2.3 Sorting and Segregation Sorting and segregation processes can be highly cost effective when the disposal costs for low-level waste are relatively high. The sorting process can take place either before the discarded material is labeled "low-level" waste or before actual disposal of the material. The effective use of sorting of radioactively contaminated items (and parts of items) from the bulk of the discarded material has been shown to cut down drastically on waste volumes. Alternatively, known contaminated material can be sorted into combustible material, waste suitable for decontamination, compactable waste, and noncompactable waste.

5.2.4 Decontamination Decontamination is the process whereby the surface contamination of solid low-level waste is either removed or reduced. The ultimate goal is to reduce the contamination to a level whereby reuse or nonradioactive waste disposal is feasible. Decontamination methods include mechanical (steam, water, or sand blasting), chemical (acids, alkalies), and ultrasonic (vibrating liquid baths). In some cases, the reuse or recycle of the decontaminated material is possible. Surface-contaminated scrap metal is often reclaimed by smelting. In some cases, minimally contaminated metal ingots that can be used in the production of shielding for spent fuel and high-level waste storage, shipping, and disposal casks are produced.

5.2.5 Mechanical Volume Reduction Numerous mechanical volume-reduction techniques [baling, shredding, and compaction (both conventional and supercompaction)] are available for use on dry low-level waste. Volume reduction values of 3 to 10 are not unusual. Mechanical volume reduction and other forms of volume reduction are heavily used when the disposal costs are high. Often, the final products are high-density disks that can be repacked in standard barrels to facilitate handling and disposal.

5.2.6 Stabilization and Solidification Solidification is generally used to convert low-level waste to a form that, upon its disposal, resists degradation and any release of contaminants to the environment. The waste to be treated can be liquid, slurry, sludge, wet solid, or dry solid. In most cases, the material to be stabilized is simply mixed with an appropriate solidification agent and allowed to solidify in the final disposal container. Some of the most widely used solidification agents are cements, asphalt, polymers, and molten glass. The form of the waste (liquid or solid) and the chemical and thermal stability of the waste determine the optimal solidification agent.

5.3 Low-Level Waste Disposal Methods

The various methods for the final disposal of low-level waste can be grouped into a number of generic categories: above-ground disposal, near-surface underground disposal, intermediate-depth disposal, deep disposal in geologic repositories, and ocean disposal. All specific methods for the disposal of low-level waste have their own advantages and disadvantages. All the low-level waste-producing countries have assumed the responsibility of disposing of their own low-level waste (i.e., very little, if any, transferring of low-level waste from one country to another exists) and, thus, need to optimize the cost-benefit tradeoffs (including safety) of the various options under their particular set of circumstances.

5.3.1 Above-Ground Disposal For above-ground disposal, reinforced concrete vaults or buildings are constructed directly on the surface. Any waste stored in the vault will be isolated from the environment as long as the integrity of the building is maintained. Above-ground disposal vaults are relatively inexpensive and can be easily located and monitored. However, they do lack the environmental protection and the barrier to radionuclide migration that any surrounding soil would offer. This type of disposal is often used for both short- and long-term interim storage of low-level waste when permanent disposal is to occur at a later time.

5.3.2 Near-Surface Underground Disposal Technologies fitting this category have historically been used for the disposal of the vast majority of low-level waste throughout the world. Various modifications of this category are generally the current method of choice for most countries that generate low-level waste because of the low cost and adequate isolation of the low-level waste. For most of these technologies, the waste is simply emplaced within about the first 10 m of the surface and then capped with a layer of soil. The most common near-surface disposal methods include trenches and below-grade or soil-mounded vaults or bunkers. In the past, water infiltration into near-surface disposal units has been linked to the loss of integrity in the overlying cap caused by the compaction and settling of the waste. Regulations requiring the compaction of the low-level waste before its emplacement, along with the use of high-integrity containers and concrete overpacks, help prevent many of these problems. In addition, a number of designs call for the stabilization of the emplaced waste with either a loose soil, sand, or gravel fill, or an injected concrete grout around, and in some cases into, the waste packages. After a disposal unit is filled, a cap composed of single or multiple layers of soil or impermeable synthetic membranes is used to seal the unit from precipitation and ground-water runoff. Although most low-level waste disposal units have been the below-grade type, where the elevation of the surface drainage is above the buried waste, a number of countries have used above-grade facilities, where the surface drainage elevation is below the lowest waste level. Earth-covered bunkers have been used successfully in France for over 20 years, and above-grade tumuli are currently being evaluated for some of the new generation of low-level waste disposal units being considered for use in the United States (Fig. 1).

5.3.3 Intermediate-Depth Disposal Augered holes in the earth's surface have been successfully used for the emplacement of low-level waste at a number of DOE sites (Fig. 2). This method involves the construction of concrete- or metal-lined holes, up to 3 m in diameter and up to 20 m or more in depth, in suitable soil. After the emplacement of the low-level waste, the holes are typically filled with a cement or concrete grout, and a concrete cap is added. Although this is a possibly acceptable method for the disposal of low-level waste, the difficulty in locating and monitoring suitable sites will generally favor other disposal methods. Also included in this category are the use of deep trenches and old pit mines. Unless readily available, these methods are costlier than shallow trenches and yet do not add any protection against releases to the environment.

5.3.4 Deep Disposal in Geologic Repositories Geologic repositories located up to 1500 m below the surface are generally being considered as sites for the disposal of most of the world's high-level waste. These deep repositories, although expensive, generally provide excellent barriers to the migration of radionuclides into the environment and to human intrusion into the disposal unit. A number of European countries are also planning to use such repositories for the disposal of low-level waste. Germany disposed of low-level waste in the Asse salt mine from 1967 to 1978. Sweden has built, and the United Kingdom plans to develop, geologic repositories for low-level waste. The United States is considering

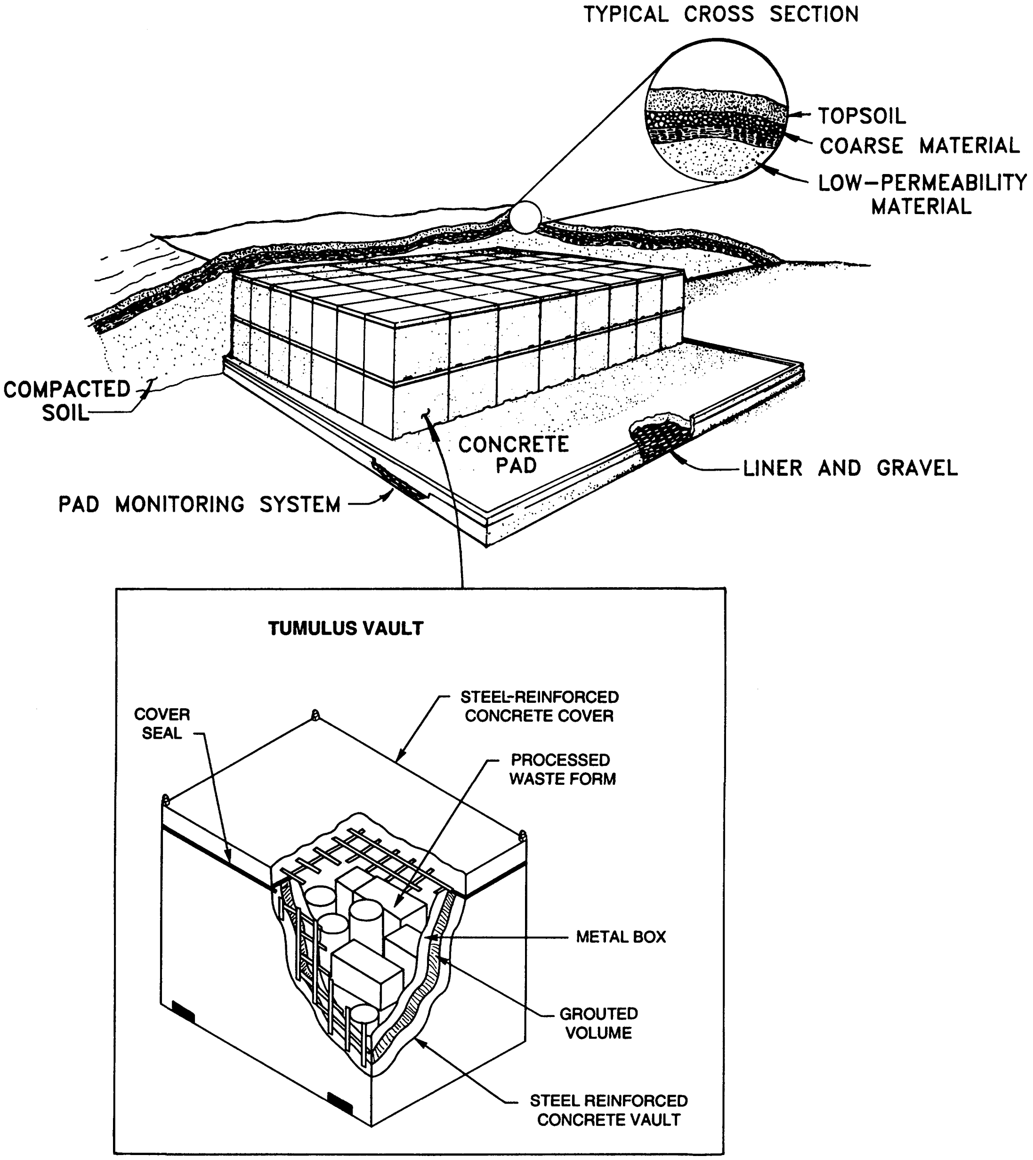
TYPICAL CROSS SECTION
TOPSOIL
COARSE MATERIAL
LOW-PERMEABILITY MATERIAL
COMPACTED SOIL
CONCRETE PAD
LINER AND GRAVEL
PAD MONITORING SYSTEM
TUMULUS VAULT
COVER SEAL
STEEL-REINFORCED CONCRETE COVER
PROCESSED WASTE FORM
METAL BOX
GROUTED VOLUME
STEEL REINFORCED CONCRETE VAULT

FIG. 1. Above-grade tumuli for the disposal of low-level radioactive waste (as being considered in the United States).

such repositories only for spent fuel, high-level waste, and transuranic waste. In most cases, deep geologic repositories are not considered cost-effective for disposing of low-level waste. The additional protection gained by disposing of material at such depths is often not necessary, and locating such repositories can be difficult and time-consuming.

5.3.5 Ocean Disposal In the 1950s and 1960s, the United States disposed of about 90 000 containers of low-level waste at various depths and distances from shore at several Atlantic and Pacific Ocean sites. Since 1967, a number of European countries have cooperated in a joint disposal of some 36 000 containers off the English coast. Although this is not currently practiced in the United States, several other countries do practice limited ocean disposal of radioactive waste, including low-level waste, without a great deal of publicity.

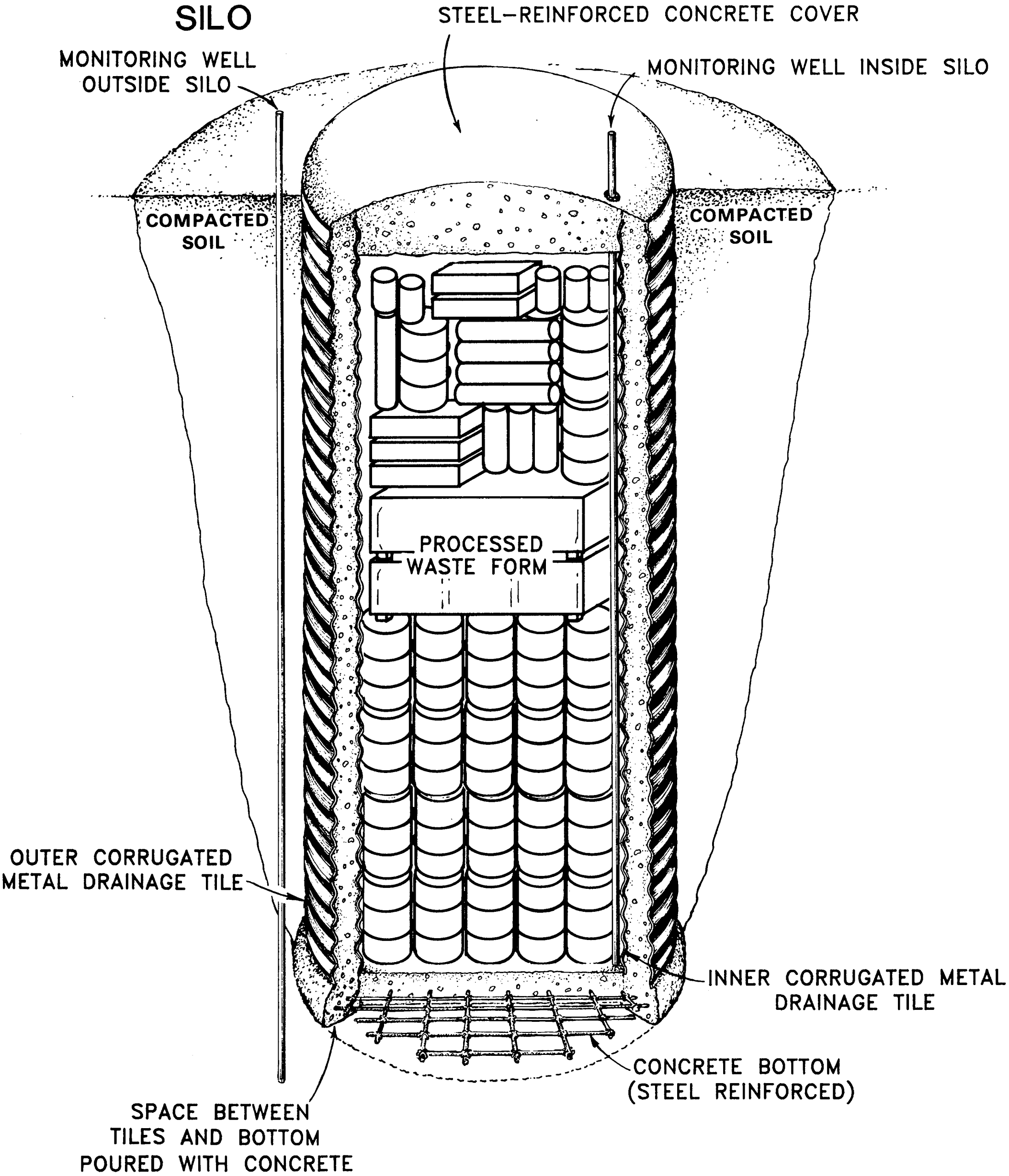
SILO
STEEL-REINFORCED CONCRETE COVER
MONITORING WELL OUTSIDE SILO
MONITORING WELL INSIDE SILO
COMPACTED SOIL
COMPACTED SOIL
PROCESSED WASTE FORM
OUTER CORRUGATED METAL DRAINAGE TILE
INNER CORRUGATED METAL DRAINAGE TILE
CONCRETE BOTTOM (STEEL REINFORCED)
SPACE BETWEEN TILES AND BOTTOM POURED WITH CONCRETE

FIG. 2. Augured holes (silos) for the disposal of low-level radioactive waste.

5.4 Low-Level Waste Disposal in the United States

The disposal of commercial low-level waste in the United States dates back to 1962 with the opening of the Beatty, Nevada, disposal site. Since then, a number of other sites have been opened and subsequently closed; only two (Hanford, Washington, and Barnwell, South Carolina) are still accepting low-level waste. The shipping of the country's waste to the two remaining sites developed into a general concern that the responsibility for accepting low-level waste for disposal should not rest with only a few states and that a coordinated national plan was needed. In response, in 1982, the U.S. Congress passed an act (the Low-Level Radioactive Waste Policy

Act) making each state responsible for its own low-level waste and encouraging the formation of regional interstate compacts to deal with the disposal problems. Although the law laid out the milestones leading to the operation of new regional facilities by 1 January 1993, none have been built, and it remains to be seen what the full impact of the law will be on those industries producing low-level waste.

6. MIXED-WASTE MANAGEMENT

Because mixed waste contains both radioactive materials and nonradioactive hazardous chemicals, it raises a number of legal, regulatory, institutional, and management issues that are unique. All radioactive waste types may, in part, contain some mixed wastes. However, the dominating radioactive characteristics of spent fuel, high-level waste, and transuranic waste generally dictate the methods by which these materials need to be treated, handled, stored, and disposed of. Thus, only mixed low-level waste is generally considered for mixed-waste management.

While low-level waste is somewhat ill-defined throughout the world, the definition of what constitutes a hazardous material is even less exact. Thus, the combination of the two definitions can be either extremely specific or quite confusing, or there may not be a definition at all. A majority of countries consider the low-level components to be of dominant concern and treat their low-level mixed waste as if it were only a radioactive waste. Other countries highlight the hazardous chemicals and manage the waste primarily on the basis of its hazardous components.

In the United States, the radioactive components of mixed waste are subject to the Atomic Energy Act, which, for government sources, is administered by DOE. Commercial sources are administered by the Nuclear Regulatory Commission (NRC) (unless a state has obtained agreement-state status). The hazardous components of all mixed wastes in the United States are subject to two federal statutes that are administered by the Environmental Protection Agency (EPA) (unless a state has obtained an authorization status): the Resource Conservation and Recovery Act [RCRA (1976, 1984)] and the Toxic Substances Control Act [TSCA (1976)]. Thus, the treatment, handling, and disposal of mixed wastes are subject to the regulations of the EPA and NRC (or the authorized and agreement states) or DOE.

Typically, mixed waste includes a variety of contaminated materials—for example, liquid scintillation fluids, air filters, cleaning materials, engine oils and grease, chlorinated or fluorinated organics, pesticides, lead wastes, mercury wastes, other toxic metals, paint residues, photographic materials, soils, building materials, and decommissioned plant equipment. It has been estimated that about 5 to 10% of U.S. low-level waste may, in fact, be properly classified as mixed waste. The characterization as mixed wastes is often quite difficult because such wastes may have different blends of hazardous (chemical or physical) and radioactive components that preclude accurate characterization. In addition, several processes may be involved in generating the wastes, including accidents that may not be amenable to characterization. Further complicating the issue is the fact that regulatory procedures for determining the hazardous components and their toxicities are rapidly changing.

6.1 Availability of Mixed-Waste Treatment Facilities

If, in a particular country, mixed waste is to be treated as either low-level waste or hazardous waste, then the treatment facilities or methods used for disposal are somewhat well defined. If, however, the mixed waste is jointly regulated—both as a radioactive material and as a hazardous waste (as in the United States)—then the nature and availability of mixed waste treatment and disposal facilities become much more difficult to assess.

In most cases, the usual procedure is to treat or remove the hazardous component by the most expeditious and effective treatment technology. The resulting treated waste, residue, or immobilized material can then be managed as if it were a low-level waste. In some cases, such as waste being held for the decay of radioactive isotopes or the presence of low concentrations of certain isotopes such as 0.05 μCi/g or less of ^{3}H or ^{14}C in the United States, the waste may be managed as if it were a hazardous material. The meager number of waste management facilities that meet all the technical requirements to handle both hazardous and radioactive materials and that have all the necessary permits and licenses to do so limits the availability of mixed-waste management facilities to the generators of this waste. In the United States, the nonavailability of licensed or permitted facilities to treat certain categories of mixed waste and the general prohibition on the long-term storage of hazardous wastes have placed numerous mixed-waste generators in a difficult position.

7. URANIUM MILL TAILINGS MANAGEMENT

Uranium mill tailings are the residual wastes of milled ore that remain after the uranium values have been recovered. Depending on the ores' chemical characteristics, uranium mills use either an acid or alkaline leach process to recover uranium. Currently, mills having more than 96% of the U.S. milling capacity use the acid leach process; the rest use the alkaline process or a number of nonconventional methods. Acid leaching is preferred for ores with low lime content. Those with high lime content require excessive quantities of acid for neutralization and, for economic reasons, are best treated by alkaline leaching. In either leach process, most of the uranium is dissolved together with other materials present in the ore (e.g., iron, aluminum, and other impurities). After the ore is leached, the uranium-laden leach liquor is removed from the tailings solid by decantation. Mill tailings from both processes consist of slurries of sands and claylike particles called slimes.

Because the amount of uranium (by weight) extracted from the ore during milling is relatively small, the dry weight of the tailings produced is nearly equal to the dry weight of ore processed. Typically, dry tailings are composed of 70 to 80 wt % sand-sized particles and 20 to 30 wt % finer-sized particles. The tailings contain other naturally radioactive substances (e.g., radium) that are responsible for more than three-fourths of the radioactivity that was originally present in the ore. After thorough washing, the tailings are pumped as a slurry to a tailings pond where the emplaced waste is drained and the water evaporated.

Because huge amounts of tailings are produced from even a modest uranium mill industry, large tailings have historically been produced. In the United States, over 100×10^6 m^3 of commercial uranium mill tailings are distributed over about 24 tailings sites. An equal number of DOE sites possess an estimated additional 30×10^6 m^3 of mill tailings. Several of the largest commercial sites contain over 10×10^6 m^3 of tailings, in piles covering over 100 ha each. Only two of these are currently active, adding some 360 000 m^3 of tailings in 1991. Of the inactive sites, their respective status ranges from unstabilized, to covered with wood chips, to partially stabilized, to fully stabilized with soil and surface plantings. In the future, most of the U.S. production of uranium is expected to come from nonconventional extraction operations (*in situ*, by-product, etc.) with imports and U.S. inventory drawdown projected to make up 80% of U.S. requirements through the year 2005. On a worldwide basis, at least 200×10^6 m^3 of tailings exist, with an additional $(10\text{–}20) \times 10^6$ m^3 of tailings being added each year.

In the past, failures of the holding dams at mill tailings impoundments have occurred. In addition, there was the historical use of the easy-to-handle, readily available, and expensive mill tailings as a source of landfill. Individuals and institutions used the mill tailings in the construction of homes, commercial buildings, and even schools. Because of such events and situations, the creation of large tailings piles is now widely regarded by the public and various regulators as unacceptable. Alternative disposal practices are being actively pursued. The obvious solution is to require the burial of tailings below grade in specially excavated pits or as backfill in deep pit and underground mines so as to reduce radioactive exposure levels to natural background levels. Use of cost-benefit, as low as reasonably achievable (ALARA), and sound engineering analyses leads to the conclusion that the various uncertainties in the predicted values of collective doses to the surrounding population and the costs associated with various waste management options often make sound choices for the future management of mill tailings quite difficult.

In the United States in 1986, the EPA issued rules on ^{222}Rn emissions from tailings piles. Mill owners will have to phase out the use of large existing tailings piles. New tailing piles must be contained in small (i.e., less than 16 ha) impoundments or disposed of by continuous dewatering and burial with no more than 4 ha uncovered at any one time.

8. QUANTITIES OF RADIOACTIVE WASTE

8.1 International Radioactive Waste Generation

Information on the quantities of radioactive waste being managed by the various international waste producers is often difficult or impossible to obtain. Accurate numbers for only spent fuel are available for most countries. Table 1 presents data on the amount of spent fuel generated through 1991 (The Uranium Institute, 1992). A comment needs to be made that some of this quantity has been reprocessed in France, Great Britain, and the United States and that additional material may be reprocessed in the future. Thus, these numbers represent total generation, not material in storage awaiting disposal. Additional quantities of spent fuel have been produced in Soviet-supplied reactors, but these numbers are not included.

Table 1. Spent-fuel production through 1991.

Country	MTIHM
Argentina	1 800
Belgium	1 400
Brazil	150
Canada	15 000
Finland	450
France	16 000
Germany	4 000
India	1 800
Italy	2 000
Japan	9 000
Netherlands	250
Slovenia	140
South Africa	400
South Korea	1 200
Spain	3 300
Sweden	3 000
Switzerland	1 200
Taiwan	1 200
United Kingdom	34 000
United States	24 000

8.2 U.S. Radioactive Waste Generation

The DOE annually publishes a reference document that quantifies the inventories of all radioactive material being generated, stored, or disposed of in the United States (Martin Marietta Energy Systems, Inc., 1994). Table 2 provides a summary of the quantities of radioactive waste being stored in the United States as of 31 December 1993. It must be noted that some differences in definitions for the various waste categories exist between the commercial and the DOE waste-generating industries. In the United States, low-level waste accounts for over 85% of the stored radioactive waste but less than 1% of the radioactivity.

Table 2. Radioactive waste inventories as of 31 December 1993.

Waste category	Total amount in storage	Comment
Spent fuel	28 000 MTIHM[a]	Does not include approximately 2500 MTHM of DOE-owned spent fuel
High-level waste	403 500 m^3	Material predominantly unsolidified waste
Transuranic waste	232 600 m^3	Includes both buried material and waste to be sent to WIPP
Low-level waste	4 561 400 m^3	Includes both commercial and defense material
Uranium mill tailings	118 600 000 m^3	Includes only commercial tailings.
Mixed waste	190 700 m^3	Does not include 2,100 m^3 of commercial mixed waste as of 31 December 1990

---

[a] Metric tons initial heavy-metal equivalent.

GLOSSARY

Activation Product: A radioactive material produced by bombardment with neutrons, protons, or other nuclear particles.

Burnup (Spent Fuel): The total energy released per initial unit mass of reactor fuel as a result of fission.

Calcine: A form of high-level waste produced from spent-fuel reprocessing waste by heating to a temperature below the melting point to bring about a loss of moisture and oxidation to a chemically stable form.

Canister: A metal container used for the storage or disposal of heat-producing solid radioactive waste.

Contact-Handled Waste: Radioactive waste with a low surface dose rate and minimal heat generation that permits handling by contact methods.

Decay, Radioactive: The transition of a nucleus from one energy state to a lower one, usually involving the emission of a photon, an electron, or a neutron.

Enrichment (Reactor Fuel): A nuclear fuel cycle process in which the concentration of fissionable uranium (i.e., ^{235}U) is increased above its natural level of 0.71%.

Fertile Nuclide: A nuclide capable of being transformed into a fissile nuclide by neutron capture.

Fissile Nuclide: A nuclide capable of undergoing nuclear fission with neutrons.

Fission Products: Nuclides produced either by fission of heavier nuclides or by the subsequent decay of the nuclides thus formed.

Fuel Assembly: A grouping of nuclear fuel rods that remains integral during the charging and discharging of a reactor core.

Geologic Repository: A facility that has an excavated subsurface system for the permanent disposal of spent fuel, high-level waste, and possibly other waste forms.

Half-Life, Radioactive: For a single radioactive decay process, the time required for the activity to decrease to half its initial value.

Heavy Metal: Mass of the fissile and fertile (heavy metal) content of nuclear fuel.

High-Level Waste: Generally, the highly radioactive material resulting from the reprocessing of spent nuclear fuel, including the liquid waste produced directly in reprocessing and the solid material derived from such liquid waste.

Intermediate-Level Waste: Radioactive waste that contains higher levels of radioactive than most low-level waste.

Low-Level Waste: For most countries, radioactive waste not classified as spent fuel, high-level waste, transuranic waste, or any other specific category of radioactive waste. Radiation levels are usually quite low, but occasionally the radiation levels may be high enough to require shielding.

Mixed Waste: Waste that includes concentrations of both radionuclides and hazardous chemicals.

Nuclear Fuel Cycle: The complete series of steps involved in supplying fuel for nuclear reactors. It includes mining, refining, enrichment, fabrication of fuel elements, use in a reactor, chemical processing to recover the fissionable material in the spent fuel, reenrichment of the fuel material, refabrication of new fuel elements, and management of radioactive waste.

Remote-Handled Waste: Radioactive waste with a high surface dose rate and/or heat generation that requires remote handling and/or shielding.

Reprocessing, Fuel: The chemical/mechanical processing of irradiated nuclear reactor fuel to remove fission products and recover fissile and fertile material.

Spent Fuel: Nuclear fuel that has been permanently discharged from a reactor after it has been irradiated.

Transuranic Waste: Radioactive waste that contains elevated concentrations of alpha-emitting isotopes with atomic numbers greater than 92. Some countries (e.g., the United States) may have a more specific definition.

Uranium Mill Tailings: Earthen residues that remain after the extraction of uranium from its ores.

Vitrification: The conversion of radioactive waste materials into a glassy or noncrystalline solid for subsequent disposal.

Works Cited

American Nuclear Society, Inc., and American Society of Civil Engineers (1992), *High-Level Radioactive Waste Management, Proceedings of the Third International Conference, Las Vegas, Nevada, April 12–16, 1992,* New York: ASCE. Third in a series of international conferences devoted solely to the subject of high-level waste management.

Forsberg, C. W. (1991), *Representative Geological*

Repository Designs for Spent Fuel and High-Level Waste, K/ITP-393, Oak Ridge, TN: Martin Marietta Energy Systems, Inc.

Martin Marietta Energy Systems, Inc. (1994), *Integrated Data Base Report—1994: U.S. Spent Fuel and Radioactive Waste Inventories, Projections, and Characteristics,* DOE/RW-0006, Rev. 10, Oak Ridge, TN: Oak Ridge National Laboratory.

The Uranium Institute (1992), *Radioactive Waste and the Nuclear Fuel Cycle,* London: The International Industrial Association for Energy from Nuclear Fuel.

Further Reading

League of Women Voters Education Fund (1985), *The Nuclear Waste Primer—A Handbook for Citizens,* New York: Nick Lyons Books. Although somewhat out of date, a comprehensive view of the political implications in managing radioactive waste. A newer version of this publication may be available.

Moghissi, A. A., Godbee, H. W., Hobart, S. A. (Eds.) (1986), *Radioactive Waste Technology,* New York: American Society of Mechanical Engineers. Contains a great deal of information on radioactive waste management for spent fuel, high-level waste, and low-level waste.

Post, R. G., Wacks, M. E. (Eds.) (1992), *Waste Management '92, Proceedings of the Symposium on Waste Management at Tucson, Arizona, March 1–5, 1992,* Tucson: Arizona Board of Regents. The latest in a series of annual waste management symposia that cover all aspects of radioactive waste management.

Yang, Y. S., Saling, J. H. (1990), *Radioactive Waste Management,* New York: Hemisphere Publishing Corporation. A detailed look at the entire subject of radioactive waste management.

NUCLEAR WEAPONS

See WEAPONS, NUCLEAR

NUCLEIC ACIDS

MAXIM D. FRANK-KAMENETSKII, *Department of Biomedical Engineering and Center for Advanced Biotechnology, Boston University, Boston, Massachusetts, U.S.A.*

Introduction 20
1. **Nucleic Acid Structures** 22
1.1 Major Helical Structures of DNA 22
1.1.1 B-DNA 22
1.1.2 B′-DNA 23
1.1.3 A-DNA 23
1.1.4 Z-DNA 23
1.1.5 psDNA 24
1.2 RNA Structure 24
1.3 DNA Topology 24
1.3.1 Knots of DNA 24
1.3.2 DNA Supercoiling 25
1.4 Unusual Structures and Structural Transitions 25
1.4.1 Melting of DNA 25
1.4.2 B–Z Transition 26
1.4.3 Cruciforms 26
1.4.4 H-DNA 27
1.4.5 Quadruplexes 27
1.4.6 RNA Unusual Structures 27
2. **Methods to Study Nucleic Acid Structures** 28
2.1 General Methods 28
2.1.1 X-ray Analysis 29
2.1.1.1 Fiber Diffraction 29
2.1.1.2 X-ray Crystallography 29
2.1.2 Nuclear Magnetic Resonance (NMR) 29
2.1.3 Microscopy 30
2.1.3.1 Regular Electron Microscopy 30
2.1.3.2 Cryo-Electron Microscopy ... 30
2.1.3.3 Scanning Tunneling and Force Microscopy 31
2.1.4 Optical Methods 31
2.1.4.1 UV Spectroscopy 31
2.1.4.2 Circular Dichroism (CD) 31
2.1.4.3 Infrared and Raman Spectroscopy 31
2.1.4.4 Fluorescence 31
2.1.5 Theoretical Methods 31
2.1.5.1 Conformational Analysis 32
2.1.5.2 Theoretical Models 32
2.2 Special Methods 33
2.2.1 Gel Electrophoresis 33
2.2.1.1 Background 33
2.2.1.2 Regular Gel Electrophoresis and DNA Sequencing 33
2.2.1.3 Pulse Gel Electrophoresis 34
2.2.1.4 Separation of DNA Topological Conformers 35
2.2.1.5 Two-Dimensional Gel Electrophoresis 35
2.2.2 Chemical, Photochemical, and Enzymatic Probing 36
3. **Biological Functions of Nucleic Acids** 37
3.1 Replication 37
3.1.1 DNA Replication in Cells 37
3.1.2 Replication of Viruses 37
3.2 Transcription 38
3.2.1 Transcription in Prokaryotes . 38
3.2.2 Transcription in Eukaryotes 38
3.3 Translation Process and Genetic Code 39
4. **Biotechnology** 40
4.1 Major Tools of Genetic Engineering 40
4.1.1 Chemical Synthesis of DNA .. 40
4.1.2 Restriction Enzymes 40
4.1.3 Reverse Transcriptase 40
4.1.4 Plasmids 40
4.1.5 Polymerase Chain Reaction (PCR) 40
4.2 The Coming of the Biotechnology Age 41
4.2.1 Protein Superproduction 41
4.2.2 Protein Engineering 41
4.2.3 New Pharmacology 41
4.2.4 Forensic Applications 41
4.2.5 Transgenic Animals 41
4.2.6 Human Genome Project 41
4.2.7 Diagnostics and Correction of Genetic Defects 41
Glossary 42
Works Cited 44
Further Reading 45

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

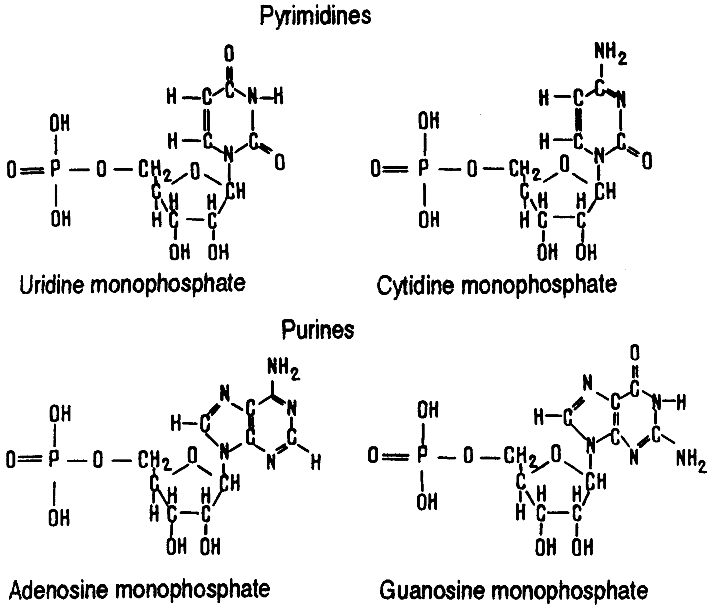

FIG. 1. Chemical formulas of residues of RNA, i.e., of nucleotides. At the top are the pyrimidine nucleotides (U and C), and below, the purine nucleotides (A and G). Nucleotides within DNA differ in that, instead of the right-hand lower OH group, they simply have H. In addition, DNA, instead of the uridine nucleotide, includes the thymidine nucleotide whose top C-H group in the ring is replaced by the C-CH_3 group.

INTRODUCTION

Nucleic acids play a crucial role in all living organisms because they are the key molecules responsible for storage, duplication, and realization of genetic information. There are two types of nucleic acids; ribonucleic acid, or RNA, and deoxyribonucleic acid, or DNA. They are both heteropolymeric molecules consisting of residues (nucleotides) of four types. Four RNA nucleotides are shown in Fig. 1. Each consists of a phosphate group, sugar (these two elements are identical for all four nucleotides), and a nitrous base, which determines the type of the nucleotide. DNA nucleotides are very similar to RNA nucleotides. The main difference is that the right-hand OH group in the sugar is replaced with just H (hence, "deoxy"). Three bases (A,G, and C) are the same for DNA and RNA. Only the fourth base is different: Instead of uracil (U), DNA carries thymine, T (see Fig. 2). The DNA chain is shown in Fig. 3.

In living nature, as we observe it today, DNA plays the most important role because genetic information is stored and duplicated in the form of DNA in all living cells and organisms, without any exceptions. Only in some viruses, which are not living creatures because they cannot reproduce themselves outside cells, is genetic information carried in the form of RNA molecules. The most notable class of such viruses are retroviruses, which include many cancer-inducing viruses as well as HIV, the AIDS virus.

The genetic message is "written down" in the form of continuous text consisting of four letters (DNA nucleotides A, G, T, and C). This continuous text, however, is subdivided, in its biological meaning, into sections. The most significant sections are genes, parts of DNA, which carry information about the amino-acid sequences of proteins. Whereas DNA stores

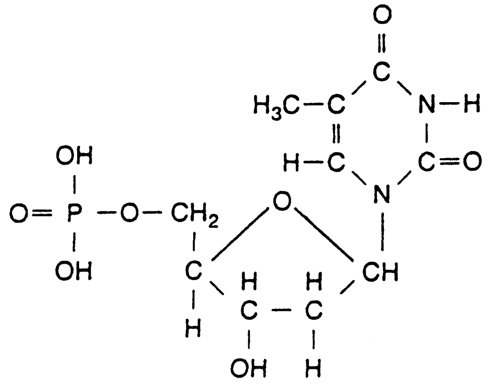

FIG. 2. Thymidine monophosphate is a thymine nucleotide that is part of DNA. The remaining three DNA nucleotides have similar structure, but each has a nitrous base of its own (the top group). These three bases (adenine, guanine, and cytosine) are similar in DNA and RNA (see Fig. 1).

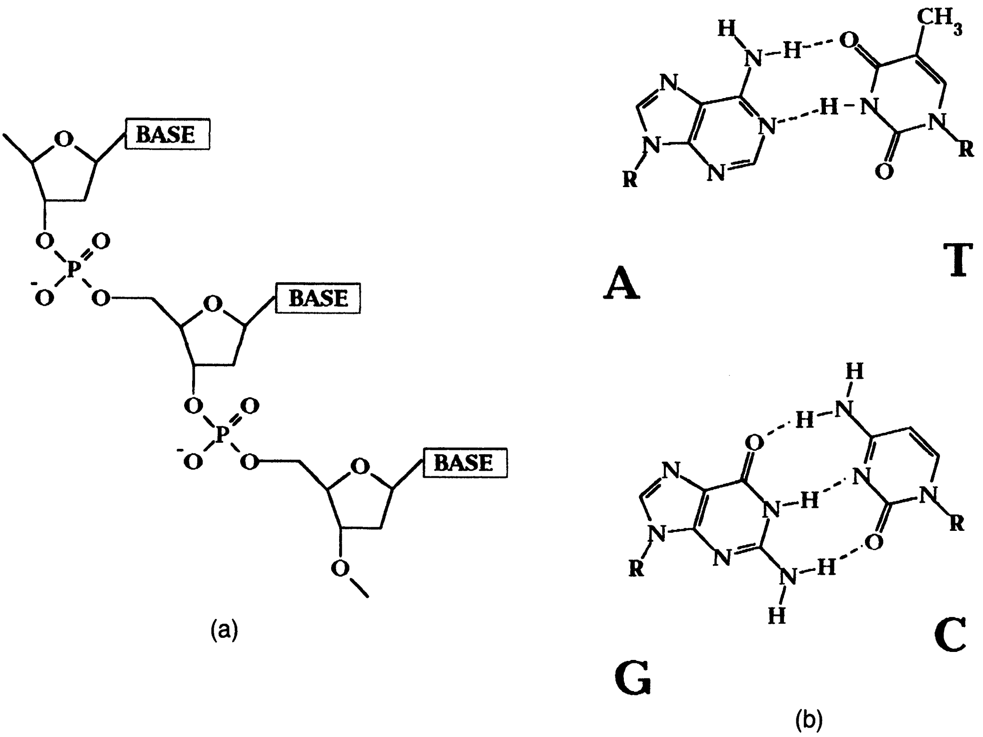

(a)

(b)

FIG. 3. (a) The DNA single strand and (b) the Watson–Crick complementary base pairs.

and duplicates genetic information, RNA plays a key role in its realization, i.e., in the process of synthesis of a protein molecule in accordance with the nucleotide sequence of the corresponding gene.

This process consists of two major steps. First, a RNA copy of the gene is made with the help of a special enzyme, RNA polymerase. The process is called transcription, and the RNA copy of the gene is called messenger RNA (mRNA). Messenger RNA is an exact copy of the gene, in which all thymines (T) are replaced with uracils (U). In the second step, called translation, mRNA is used as a template for protein synthesis in a special cell machinery, the ribosome. The ribosome is a complex aggregate of special ribosomal RNA (rRNA) and a number of proteins. It translates the nucleic-acid text into amino-acid language. In so doing, it uses a special dictionary, the genetic code. In the process of translation, an additional very important class of small RNA molecules plays a crucial role. These RNA molecules consist of about a hundred nucleotides each and are called transfer RNA (tRNA).

The above scheme, known as the central dogma of molecular biology, gives a very general description of how the nucleotide sequence of a gene is realized in the form of the amino-acid sequence of a protein molecule. The scheme is really very general because, for instance, it does not show any difference between prokaryotes, unicellular organisms carrying no cell nucleus, and eukaryotes, complex organisms that have cell nuclei. In practice, however, these two major classes of living organisms significantly differ in the organization of their genomes and its functioning. This difference manifested itself most vividly during the advent of the applied branch of molecular biology, genetic engineering. When pieces of DNA from eukaryotes were incorporated into prokaryotic DNA, they failed to produce any proteins. Only after the fundamental difference was realized in the genome organization of simple microorganisms on the one hand and of complex higher organisms on the other were methods developed that permitted genes of eukaryotes to be expressed in prokaryotes. This opened the way for innumerable applications of molecular biology and marked the beginning of a new era of technological revolution—the era of biotechnology.

The following are the objectives of this article:

Provide an overview of our present knowledge of physical structures of nucleic acids. We describe the most common structures, as well as some new structures discovered recently. The most important structural transitions, such as DNA melting and the B–Z transition, are explained. Special attention is paid to topological properties of DNA, which play a significant role in its functioning within the cell.

Describe some of the major experimental methods used to study nucleic acids, including x-ray diffraction, nuclear magnetic resonance (NMR), electron microscopy, spectroscopic methods, gel electrophoresis, and chemical, photochemical, and enzymatic probing. The potencies of the methods are compared and their advantages and disadvantages discussed.

Describe major biological phenomena where nucleic acids play a crucial role, DNA replication, transcription, and translation. We describe these processes not only in bacteria, where they obey the central dogma of molecular biology, but also in higher organisms and their viruses, for which significant deviations from the central dogma are common, such as RNA splicing, synthesis of DNA on the RNA template, etc.

Provide the background of biotechnology and its prospective. We give the basic ideas for breakthrough discoveries that led to the advent of biotechnology, including restriction enzymes and DNA cloning, DNA sequencing methods, and the polymerase chain reaction (PCR).

1. NUCLEIC ACID STRUCTURES

Traditionally, nucleic acids' structural versatility was considered much narrower than the structural versatility of proteins. This viewpoint well fitted "the division of power" between nucleic acids and proteins in the present-day living creatures. The major function of nucleic acids is to store and reproduce genetic information, while proteins perform an innumerable variety of reactions within the cell and also are its main construction blocks. Major discoveries of the last two decades, however, have significantly changed this attitude. The structural and functional versatility of nucleic acids have proved to be much greater than anybody anticipated before. This is especially true for RNA, which not only adopts extremely complex spatial structures, rivaling the complexity of proteins' spatial structures, but can perform catalytic functions. Such catalytically active RNA molecules have received the special name of ribozymes. The discovery of ribozymes and of the remarkable versatility of RNA structure completely changed the attitude toward the pivotal question of which type of major biopolymers, proteins or nucleic acids, had been the prebiotic molecule, the forefather of the living cell. Before, everybody believed that only protein could fulfill this function. Now, the common view is that this should be the RNA molecule.

However, in the present-day cell, RNA plays just an auxiliary role. The major nucleic acid now is DNA. As a result, DNA structures attract the most attention.

1.1. Major Helical Structures of DNA

In spite of the enormous versatility of living creatures and, accordingly, variability of genetic texts that DNA molecules in different organisms carry, they all have virtually identical physical, spatial structure: the double-helical B form discovered by Watson and Crick (1953). Sequences of the two strands of the double helix obey the complementary principle. This principle is the most important law in the field of nucleic acids, and, probably, is the most important law of the living nature. It declares that, in the double helix, A always opposes T and vice versa, whereas G always opposes C and vice versa.

1.1.1 B-DNA B-DNA (see Fig. 4) consists of two helically twisted sugar-phosphate backbones stuffed with base pairs of two types, AT and GC. The helix is right-handed with ten base pairs per turn. The base pairs are isomorphous: The distances between glycosidic bonds, which attach bases to sugar, are virtually identical for AT and GC pairs. Because of this isomorphism, the regular double helix is formed for an arbitrary sequence of nucleotides, and the fact that DNA should form a double helix imposes no lim-

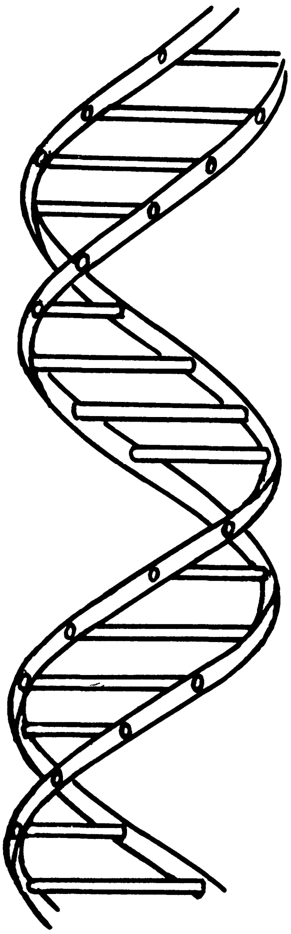

FIG. 4. The double helix.

itations on DNA texts. The surface of the double helix is by no means cylindrical. It has two very distinct grooves: the major groove and the minor groove. These grooves are extremely important for DNA functioning because, in the cell, numerous proteins recognize specific sites on DNA via binding with the grooves.

Each nucleotide has direction: If we treat the sugar as a ring lying in the plane, then the CH_2 group will be above the plane. This carbon is designated as 5′, whereas the carbon atom connecting the sugar to the next nucleotide in the polynucleotide chain is designated as 3′. Thus, the chemical direction is inherent in a single-stranded nucleic acid. In the DNA double helix, the two strands have opposite directions. In B-DNA, base pairs are planar and perpendicular to the axis of the double helix.

Under normal conditions in solution, often referred to as "physiological" (neutral pH, room temperature, about 200 mM NaCl), DNA adopts the B form. All available data indicate that the same is true for the totality of DNA within the cell. It does not exclude, however, the possibility that separate stretches of DNA carrying special nucleotide sequences would adopt other conformations.

1.1.2 B'-DNA Up to now, only one such conformation is demonstrated, beyond any doubts, to exist under physiological conditions. When several A residues in one strand (and, accordingly, several T in the other DNA strand) occur, they adopt the B′ form. In many respects, the B′ form is similar to the classical B form, but there are also significant differences. The main difference consists in the fact that base pairs in B'-DNA are not planar: They form a kind of propeller with the propeller twist of 20°.

Stretches of A residues produce bends in the double helix. Such bends play a very important role in DNA functioning. Although the structural basis of these bends is not fully understood the involvement of the B′ form in the DNA bending is very probable. In spite of its importance, the B′ form does not differ dramatically from the B conformation. Other helical conformations have been found in solution that are significantly different from B-DNA.

1.1.3 A-DNA Similarly to B-DNA, the A form can be adopted by an arbitrary sequence of nucleotides. As in B-DNA, in A-DNA the two complementary strands are antiparallel and form right-handed helices. Normal DNA undergoes a transition from the B to A form under drying. In A-DNA, the base pairs are planar, but their planes form a considerable angle with the axis of the double helix. In so doing, the base pairs shift from the center of the duplex, forming a channel in the center.

If any, A-DNA plays a rather modest role in DNA functioning. There are data indicating that some proteins induce the transition from B to A form. The reason for this remains to be elucidated.

1.1.4 Z-DNA Z-DNA presents the most striking example of how different from the B form the DNA double helix can be. Although in Z-DNA the complementary strands are antiparallel as in B-DNA, unlike in B-DNA, they form left-handed, rather than right-handed, helices. There are many other dramatic differences between Z- and B-DNA.

Not any sequence can adopt the Z form. To adopt the Z form, the regular alternation of purines (A or G) and pyrimidines (T or C) along one strand is strongly preferred. However, even this is not enough for Z-DNA to be formed under physiological conditions. Nevertheless, Z-DNA can be adopted by DNA stretches in a cell as a result of DNA supercoiling (see below). The biological significance of Z-DNA, however, remains to be elucidated.

1.1.5 psDNA The complementary strands in a DNA duplex can be parallel. Such parallel-stranded (ps) DNA is formed most readily if both strands carry only adenines and thymines, and their sequence excludes formation of the ordinary antiparallel duplex. If these requirements are met, the parallel duplex is formed under quite normal conditions. It is right-handed, but the AT pairs are not regular, Watson–Crick ones, but rather so-called reverse Watson–Crick.

1.2 RNA Structure

Quite naturally, similarity between DNA and RNA in their chemical nature entails significant similarity in structures they adopt. RNA can also form the double helix with complementary base pairs AU and GC. However, the major RNA duplex conformation resembles A-DNA rather than B-DNA. In general, because of the bulkier OH group in the sugar ring, the versatility of RNA duplexes is rather narrower than those of DNA.

In the cell, different types of RNA molecules (mRNA, rRNA, tRNA, etc.) exist as single strands. Of course, these strands are not unfolded but fold into complex spatial structures. Mutually complementary stretches of the same molecules form numerous short duplexes, which are the major structural motif of RNA molecules. Single-stranded regions form loops of different types. Sometimes, very complex folding patterns occur, called pseudoknots.

Spatial structure plays an important role in RNA functioning. For instance, rRNA organized in a complex spatial structure forms a scaffolding to which numerous ribosomal proteins attach to form the functional ribosome. The spatial structure of tRNA molecules is specially designed to permit them to fulfill the central role in realization of the genetic code. One part of tRNA carries an anticodon, a trinucleotide complementary to a codon (a trinucleotide corresponding to one amino acid in the genetic code), whereas the amino acid that corresponds to the codon in the genetic code is attached to one terminus of the molecule.

1.3 DNA Topology

DNA very often operates within the cell in the form of a closed circular molecule in which both strands form closed circles (Fig. 5). The closed circular state of DNA is extremely important because, even if DNA is linear as a whole, like the chromosomal DNA of eukaryotes, its parts are organized in closed domains, the closure being fulfilled by proteins. The physical properties and physiological behavior of closed circular (cc) DNA are different in many respects from those of linear DNA molecules. These differences stem from the fact that ccDNA acquires, as compared to linear DNA, a new fundamental feature, topology. There are two levels of DNA topology.

1.3.1 Knots of DNA The double helix as a whole may form a simple circle (i.e., a trivial knot) or be knotted (see Fig. 6). The knotting of DNA may be accomplished by ran-

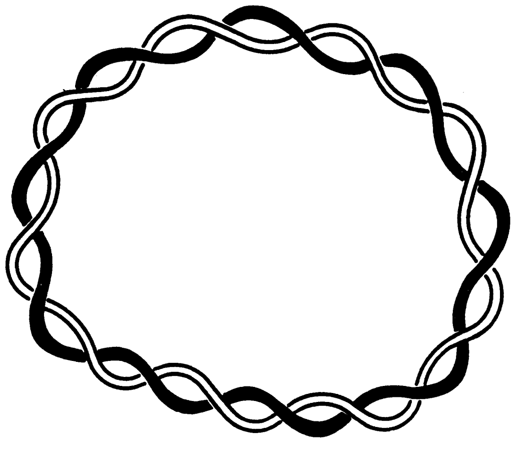

FIG. 5. In a circular closed DNA, two complementary strands form linkage of a high order.

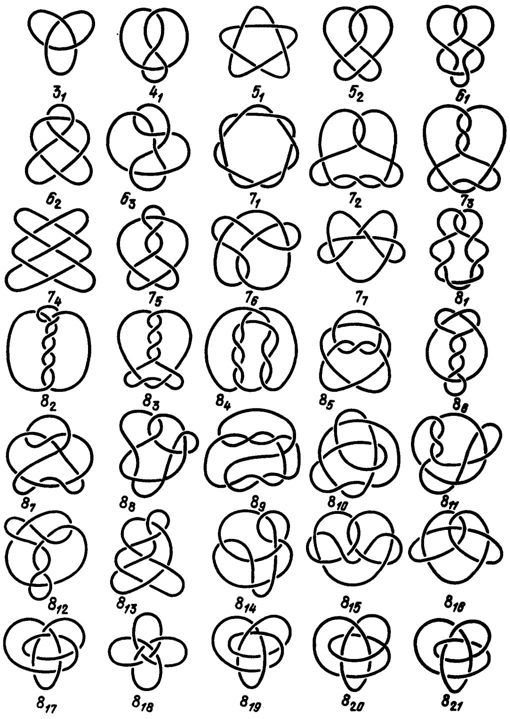

FIG. 6. Knots.

dom closure of linear molecules with "cohesive" ends (mutually complementary single-stranded overhangs). There are also various enzymes that do the job. The most important class of such enzymes is known as DNA topoisomerases II. The type of knot is a topological invariant; i.e., it cannot be changed (converted to another type of knot) without breakage of both DNA strands. As a result, identical ccDNA molecules belonging to different knot types may be separated in gel (Shaw and Wang, 1993).

1.3.2 DNA Supercoiling The two complementary strands in ccDNA are topologically linked with one another (see Fig. 5). The degree of linkage, which is designated as *Lk*, the DNA linking number, is a topological invariant, which cannot be changed by any deformation of the DNA strands without strand breaks. The *Lk* value is easily calculated as the number of times one strand pierces the surface bounded by the other strand. If N is the number of base pairs in DNA and Γ_0 is the number of base pairs per turn of the double helix under given ambient conditions, then the difference

$$\tau = Lk - N/\Gamma_0 \tag{1}$$

is the number of superhelical turns in ccDNA under given conditions. If $\tau \neq 0$, then ccDNA cannot form a planar ring with the helical twist corresponding to the Γ_0 value. Either the helical twist should change, or the DNA molecule as a whole should writhe in space. Actually, with $\tau \neq 0$, both things happen: DNA adopts writhe, and its twist changes. The distribution of energy associated with DNA supercoiling (i.e., with $\tau \neq 0$) between twist and writhe is determined by the torsional and bending rigidities, which have been measured with good accuracy (Frank-Kamenetskii, 1990a). As a result, theoretical knowledge of DNA supercoiling is fairly accurate (Vologodskii *et al.*, 1992).

Closed circular DNA molecules extracted from cells of eukaryotes are always negatively supercoiled ($\tau < 0$; the double helix is "underwound"). Usually, superhelicity is characterized by superhelical density:

$$\sigma = \tau\Gamma_0/N. \tag{2}$$

Special enzymes, topoisomerases, regulate DNA superhelicity in the cell. Supercoiling is also significantly affected by replication and transcription. DNA topology plays an extremely important role in DNA functioning (Cozzarelli and Wang, 1990).

1.4 Unusual Structures and Structural Transitions

If negative supercoiling in ccDNA becomes too high, the bending and torsional degrees of freedom may become insufficient to accommodate the superhelical energy. Under these circumstances, the integrity of the DNA double helix breaks down in its most "weak" sites. These sites, which have the potential to adopt structures significantly different from B-DNA, can undergo structural transitions into unusual (non-B-DNA) structures.

1.4.1 Melting of DNA On heating or under superhelical stress, the DNA complementary strands separate, forming single-stranded loops. In the case of linear duplex molecules, the process ends up, at sufficiently high temperature, with complete separation of the two strands. The process has been comprehensively studied both experimentally and theoretically (Wada and Suyama, 1986) and is known as DNA melting (the terms denaturation or helix-coil transition can be also found in the literature). AT pairs have significantly lower stability than GC pairs.

With increasing temperature, long DNA sections melt out. The melting of each such section is a cooperative, all-or-none phenomenon. However, the melting of a long enough molecule becomes a series of such cooperative processes, each occurring at its own temperature (Fig. 7). Roughly speaking, sections more enriched with CG pairs melt out at higher temperatures. Thus, DNA melting is not a classical phase transition because two different "phases," helical and melted, coexist with each other in one and the same molecule.

When two mutually complementary single-stranded DNA molecules are mixed, the process reverse to melting occurs. It is called renaturation or annealing.

DNA melting is an exceptionally important phenomenon. It occurs in the cell when DNA replicates and when DNA polymerase "reads" the mRNA copy of the gene. Cycles of melting and annealing are utilized in PCR machines, the most remarkable tools of the current biotechnological revolution (see Sec. 4.1.5).

Melting of AT-rich regions in ccDNA may occur with increasing negative supercoiling. However, this does not happen often because other unusual structures defeat melting in competition for the energy of supercoiling.

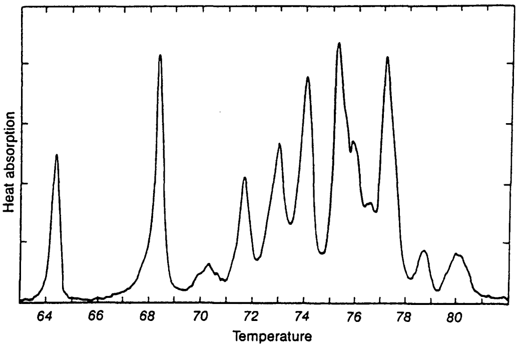

FIG. 7. Melting of DNA. This curve is also often called the differential melting curve. The curve was obtained for DNA that has the code name of ColEI and contains about 6500 nucleotide pairs.

1.4.2 B–Z Transition Negative supercoiling most favors formation of Z-DNA because, in this case, the maximal release of superhelical stress per base pair adopting a non-B-DNA structure is achieved. As a result, although under physiological conditions the Z form is energetically very unfavorable as compared with B-DNA, it is easily adopted in negatively supercoiled DNA by appropriate DNA sequences (with alternating purines and pyrimidines).

Linear DNA with the appropriate sequence adopts the Z conformation at a very high salt concentration (about 3 M NaCl).

1.4.3 Cruciforms Another structure readily formed under negative supercoiling is cruciform, which requires palindromic regions (see Fig. 8). To form a cruciform, a palindromic region must be longer than a certain minimum. For example, six–base-pair–long palindromes recognized by restriction en-

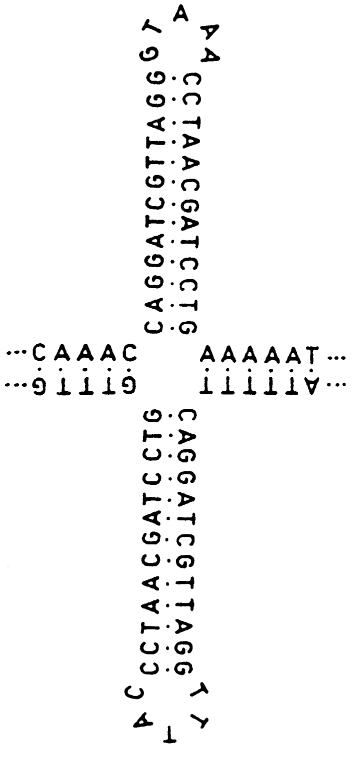

FIG. 8. A cruciform formed in ColEI DNA when the molecule is in a superhelical state.

zymes do not form cruciforms under any conditions.

1.4.4 H-DNA H-DNA forms a special class of unusual structures, which are adopted under superhelical stress by sequences carrying purines (A and G) in one strand and pyrimidines (T and C) in the other, i.e., homopurine-homopyrimidine sequences (Frank-Kamenetskii, 1990b). The major element of H-DNA is a triplex formed by one half of the insert adopting the H form and by one of two strands of the second half of the insert (Fig. 9). Two major classes of triplexes are known—pyrimidine-purine-pyrimidine (PyPuPy) and pyrimidine-purine-purine (PyPuPu). Figure 10 shows the canonical base triads entering these triplexes.

Always two isomeric forms of H-DNA are possible, which are designated as H-y3, H-y5, H-r3, and H-r5, depending on which kind of triplex is formed and which half of the insert forms the triplex (see Fig. 9). H-DNA may be considered as an intramolecular triplex (it is often referred to in this way). Its formation under physiological conditions occurs only under superhelical stress.

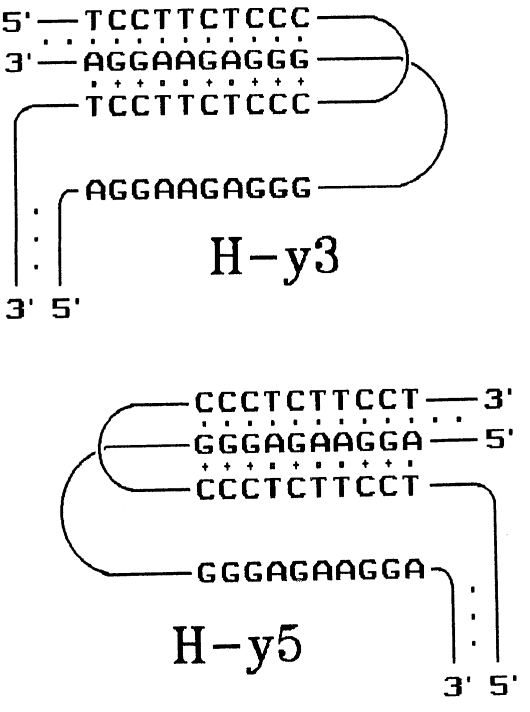

FIG. 9. DNA's H-form structure. Two possible "isomeric" variants of the structure are shown. The Watson–Crick pairing is designated by filled circles, while the GC Hoogsteen pairing, involving the presence of an extra proton, is designated by plus symbols.

The discovery of H-DNA stimulated studies of intermolecular triplexes, which may be formed between homopurine-homopyrimidine regions of duplex DNA and corresponding pyrimidine or purine oligonucleotides. These studies have led to the development of very promising approaches to sequence-specific targeting of duplex DNA (Cherny *et al.*, 1993).

1.4.5 Quadruplexes Of all nucleotides, guanines are the most versatile in forming different structures. They may form GG pairs, but the most stable structure, which is formed in the presence of monovalent cations (especially potassium), is the G4 quadruplex (see Fig. 11). G quadruplexes may exist in a variety of modifications: all-parallel, all-antiparallel, and others. As a result, G quadruplexes are easily formed both intermolecularly and intramolecularly, again with a variety of modifications.

A totally unusual quadruplex was discovered by Gehring *et al.* (1993). It contains two hemiprotonated parallel-stranded duplexes consisting of unusual CC^+ pairs. The two parallel-stranded duplexes are associated in a mutually antiparallel manner so that CC^+ base pairs from one duplex are "layered" by CC^+ pairs from the other duplex, thus alternating along the structure.

1.4.6 RNA Unusual Structures Most DNA unusual structures have their analogs in the RNA world. Specifically, the formation of triplexes was first demonstrated by Felsenfeld *et al.* (1957) for artificial RNA chains. Solitary RNA base triads are often met in tRNA structures. Triplexes are formed by mixed DNA-RNA hybrids, and some of them are rather stable (Roberts and Crothers, 1992).

RNA molecules form G quadruplexes, and there are exciting data indicating that RNA G quadruplexes could be an important element in the structure of HIV RNA. However, in general, studies of RNA unusual structures lag behind the corresponding DNA structures because methods to study unusual structures are much better developed in the case of DNA than in the case of RNA.

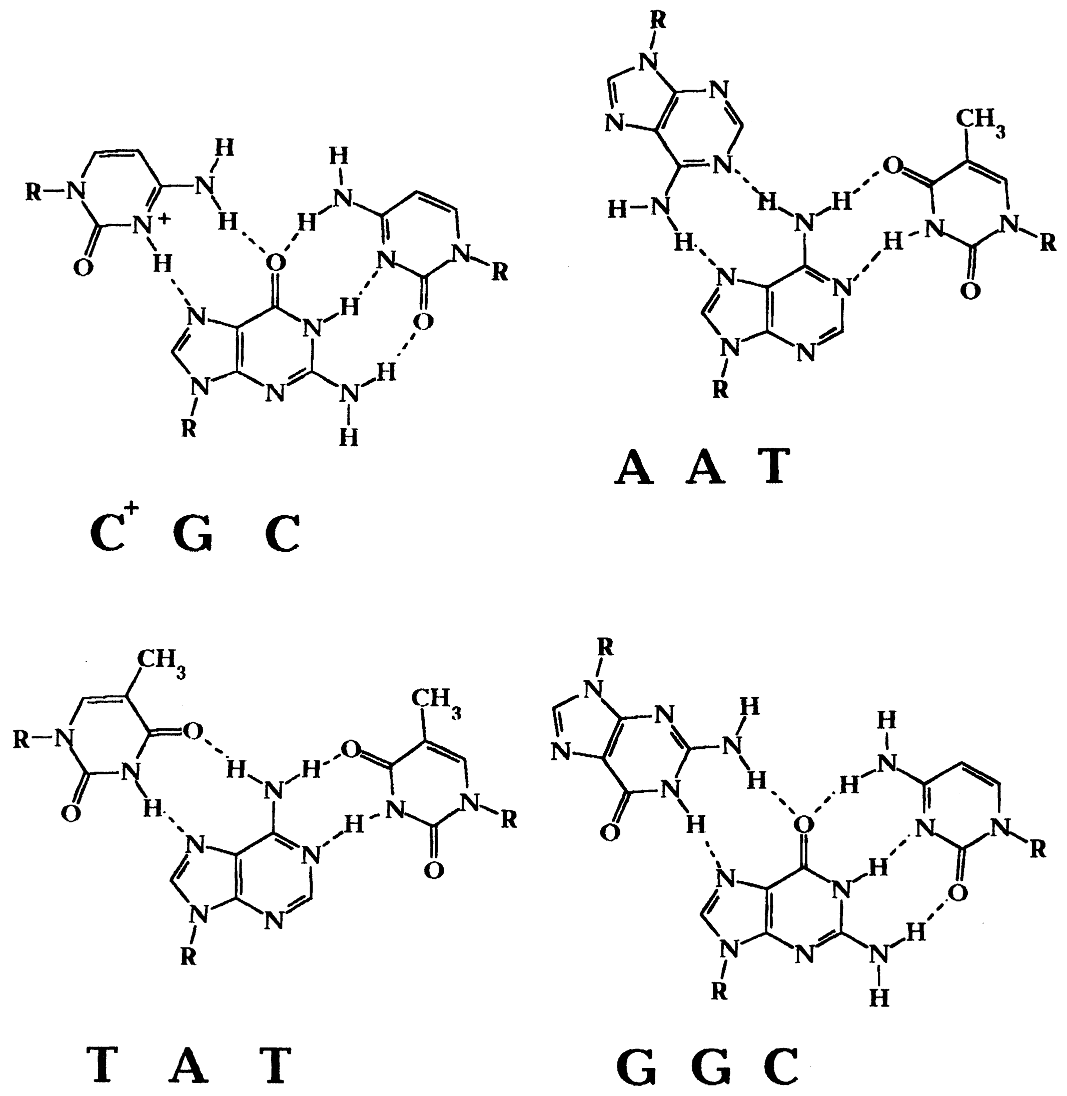

(a) (b)

FIG. 10. The structure of (a) pyrimidine (TAT and C$^+$GC) and (b) purine (AAT and GGC) base triads of which the DNA triple helix is made.

2. METHODS TO STUDY NUCLEIC ACID STRUCTURES

The number of methods that are used to study the structure of nucleic acids is truly enormous. Methods to study DNA alone occupy two volumes of the Methods in Enzymology series (Lilley and Dahlberg, 1992a,b). We survey only those of them that are most informative and are widely used. Along with general methods to study molecular structures, a number of special methods have been elaborated to study nucleic acids. But even the regular methods, such as x-ray analysis or electron microscopy, are often significantly modified to fit the needs of studying such complex molecules as nucleic acids.

2.1 General Methods

In this section, we discuss methods that are generally used to study molecular structures. We do not describe the background of

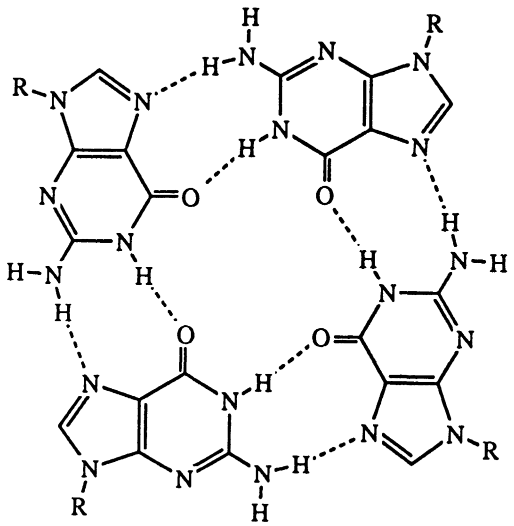

FIG. 11. G quadruplex.

the methods, because they are widely known, but rather concentrate on specific questions of their application to studing nucleic acids.

2.1.1 X-Ray Analysis As in other fields where molecular structure is essential, x-ray analysis occupies a unique position among methods of studying nucleic acid structure as the only direct method that permits elucidation of the structure in all details. X-ray crystallography is absolutely indispensable in the study of the detailed structure of complexes of nucleic acids with proteins, which is most essential for understanding how nucleic acids function in the cell.

However, for a long period of time, the whole edifice of molecular biology relied on one of the indirect versions of x-ray analysis, fiber diffraction, rather than on classical x-ray crystallography.

2.1.1.1 Fiber Diffraction. Long molecules of nucleic acids cannot be crystallized. As a result, from the early 1950s till the mid-1970s, only x-ray diffraction from nucleic acid fibers was used to elucidate the structure of nucleic acids. Such data were used by Watson and Crick (1953) to propose structures for B- and A-DNA. Fiber diffraction is an essentially indirect method, and, to elucidate structure from the data on fibers, one must heavily rely on theoretical approaches, such as conformational analysis.

2.1.1.2 X-Ray Crystallography. The first good crystals of nucleic acids were obtained for the case of relatively short nucleic acids, the tRNA molecules, and those were the first nucleic acids for which structure was solved by classical x-ray crystallography.

The first DNA crystals became available only in the late 1970s, after remarkable progress in chemical synthesis of short DNA pieces had been achieved. This led to many discoveries, first of all of left-handed Z-DNA by Rich and co-workers (Wang *et al.*, 1979).

In recent years, a lot of very detailed data on the structure of DNA have been obtained by x-ray crystallography, including detailed study of the B, B′, A, and Z forms (Dickerson, 1992). Among the most recent achievements of the method is solution of the structure of the G quadruplex (Kang *et al.*, 1992).

In spite of remarkable accomplishments, serious limitations are inherent in the method. It is extremely hard to obtain good crystals of a nucleic acid even if it adopts only a unique structure. For instance, for DNA or RNA triplexes, sufficiently good crystals are still unavailable. Even if this difficulty is overcome, it sometimes appears that the structure in a crystal is significantly perturbed by interaction with neighboring molecules. It is especially true of such subtle, but very important from a biological viewpoint, deformations as bending of the double helix. Thus, the data obtained by the methods of x-ray crystallography should always be correlated with the results of indirect methods, which permit the study of nucleic acid in solution. X-ray crystallography is totally useless for studying such biologically significant problems as DNA supercoiling or the structure of large RNA molecules.

2.1.2 Nuclear Magnetic Resonance (NMR) The role of NMR (see MAGNETIC RESONANCE AND QUADRUPOLE RESONANCE, NUCLEAR) constantly increases, and, in recent years, it has even started to compete successfully with x-ray crystallography. This has become possible as a result of the development of two-dimensional proton NMR techniques, especially nuclear Overhauser-effect spectroscopy (NOESY). The great advantage of NMR, as compared with x-ray crystallography, consists in the fact that it does not require crystals. As a result, structures of some nucleic acids that resist crystallization, like triplexes, have become subjects of very fruitful study by NMR (Rajagopal and Feigon, 1989).

A great advantage of NMR consists in the possibility of studying structural fluctuations, or "breathing," of nucleic acids by following proton exchange in nucleic acid bases. Such studies made it possible to find out important characteristics of base-pair fluctuational openings in DNA (Gueron *et al.*, 1987).

Turning to the limitations of the method, it should be emphasized that resolution of even the most powerful NMR spectrometers permits the study of only short nucleic acids containing about a dozen distinguishable nucleotides. Although crystals are not needed, only very concentrated solutions can be studied. Therefore, like x-ray crystallography, NMR is useless for studying many biologically relevant problems of nucleic acid structure. In spite of its limitations, proton NMR has firmly occupied the second position, after x-ray crystallography, among methods to study nucleic acid structure.

2.1.3 Microscopy Although microscopy looks like the most direct way to visualize structure, in application to nucleic acids it has too numerous limitations to occupy a position ahead of x-ray crystallography or NMR. Nevertheless, in recent years, the role of microscopy in the field of nucleic acids has significantly increased as a result of progress in regular electron microscopy of nucleic acids and their complexes and the development of new techniques, cryo-electron microscopy and scanning tunneling microscopy.

2.1.3.1 Regular Electron Microscopy. In regular electron microscopy, nucleic acids are placed on the grid, dried, and contrasted by one or another method. In recent years, the most popular technique of contrasting consists in staining with uranyl acetate. As a result, for instance, duplex DNA molecules are clearly seen (see Fig. 12).

Regular electron microscopy permits one to see complexes of nucleic acids with proteins. Poor resolution usually does not permit one to observe the internal structure of the complex but permits mapping of the location of the protein on DNA (see Fig. 12).

A limitation of regular electron microscopy stems from poor resolution and the possibility of significant perturbation of nucleic acid structure in the process of sample preparation. And still, in many cases, the method provides the most convincing evidence.

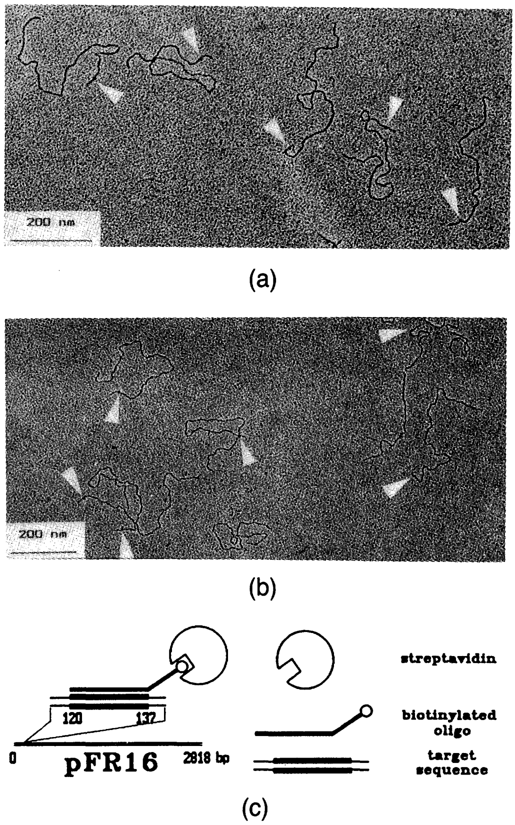

FIG. 12. Electron-microscopy visualization of triplexes formed between duplex DNA and corresponding oligonucleotides [(a) pyrimidine or (b) purine]. The oligonucleotides were chemically tagged by biotin molecules, and then the protein streptavidin was added, which strongly binds to biotin [see the chart, (c)]. The sites of triplex formation are seen as "beads" due to streptavidin molecules and are indicated by the white arrowheads. The data are from Cherny *et al.* (1993).

2.1.3.2 Cryo-Electron Microscopy. The method is based on obtaining vitrified water solutions via very quick cooling of an extremely thin (in the micron range) sample. As a result, the molecule is "frozen" in the state it adopted in solution before cooling. In recent years, the method has received numerous applications in molecular biology in general and in the field of nucleic acids and their complexes with proteins in particular (Dubochet *et al.*, 1992).

The great advantage over regular electron microscopy consists in avoiding the harsh procedures of sample preparation, which strongly limit the value of the data obtained by regular electron microscopy.

The major problem of cryo-electron microscopy stems from the low contrast of biological molecules. Without staining or other contrasting, they are barely visible in an electron microscope. Nevertheless, DNA molecules and their complexes with proteins are extensively studied by the method (Dubochet *et al.*, 1992).

2.1.3.3 Scanning Tunneling and Force Microscopy. These are very new methods in general and in the field of nucleic acids in particular. Spectacular images of DNA helices and even bases obtained by scanning tunneling microscopy (STM) have been widely publicized and created great excitement. However, the nature of these images remains obscure because DNA molecules are not conductors and are not expected to be seen in STM. The recent data of Dunlap *et al.* (1993) strongly indicate that the STM images of DNA could be artifacts.

By contrast, scanning force microscopy (SFM) provides reliable images of DNA molecules and their complexes with proteins (Bustamante *et al.*, 1993). At present, the resolution of this method is not much higher than that of regular electron microscopy. However, there are reasonable hopes that the resolution will be significantly improved in the course of further development of the method.

2.1.4 Optical Methods All optical methods that are traditionally applied to study molecular structures are widely used to study nucleic acids. They are indispensable in routine investigations because they are cheap and quick. The great advantage of those that use light in the visible and UV region consists in the possibility of study of very dilute solutions of nucleic acids, for which intermolecular interaction can be completely neglected. On the other hand, all these methods are essentially indirect and provide any structural information only after careful assignment of particular spectra or special changes with the help of more direct methods described above.

2.1.4.1 UV Spectroscopy. Nucleic acid bases absorb UV radiation around 260 nm. The intensity of this absorption, which is easy to measure with regular spectrophotometers, changes when, for instance, the DNA double helix melts. Hence, UV spectroscopy has been extensively applied to study DNA melting. In application to RNA, the method is not very informative because, in single-stranded RNA molecules, the variety of structures is too large to be elucidated by this method. Changes of UV absorption are too small for B-to-A and B-to-Z transitions.

2.1.4.2 Circular Dichroism (CD). CD spectra in the vicinity of 260 nm are much more sensitive to DNA helical structure than UV absorption. B-DNA, A-DNA, and Z-DNA have characteristic and very different CD spectra (Johnson, 1990), and this fact is extensively used in the study of structural transitions in DNA between different helical structures (Ivanov and Krylov, 1992). As with UV spectra, CD spectra are much less informative in the case of RNA than in the case of DNA.

2.1.4.3 Infrared and Raman Spectroscopy. Infrared and Raman spectra are sensitive to nucleic acid structure. Correspondingly, IR and Raman spectroscopies are used to study nucleic acids. However, the main limitation of these methods stems from the fact that they require high concentrations of nucleic acids. As a result, these methods are much less popular than UV and CD spectroscopies.

2.1.4.4 Fluorescence. Nucleic acids practically do not emit absorbed radiation. However, some RNA molecules, along with the four canonical bases, contain unusual bases, and some of them are fluorescent. This fact is extensively used to study the RNA molecules that carry such bases. Regularly, fluorescence methods are used via binding to nucleic acids of strongly fluorescent dye molecules, such as ethydium.

Fluorescence sensitization and quenching due to excitation energy transfer between the donor molecule of electronic excitation and the acceptor molecule of the excitation are also used. For instance, to check the polarity of strands the donor molecule is tagged to, say, 3'-end of one strand, whereas the acceptor molecule is tagged to the 3' end of the complementary strand. Then the two strands anneal, forming a duplex. Strong excitation energy transfer and acceptor indicates that a parallel-stranded duplex is formed.

2.1.5 Theoretical Methods The paper that signified the beginning of extensive studies of nucleic acids and their biological role was purely theoretical (Watson and Crick, 1953). Since then, theory has been playing a very important role in study of nucleic acids.

2.1.5.1 Conformational Analysis. Conformational analysis was especially important during the era of fiber diffraction. In fact, what Watson and Crick did in their classical paper (Watson and Crick, 1953) was a very simple, but exceptionally efficient, variety of conformational analysis. Since then, the method has been extensively used to refine structures solved by x-ray crystallography (Dickerson, 1992). The method is also often used to predict new structures. For instance, parallel-stranded DNA duplexes were first predicted theoretically and then found experimentally.

2.1.5.2 Theoretical Models. As in the study of any important physical object, a number of simplified theoretical models of nucleic acids exist, different models being used to analyze different properties. Figure 13 presents schematics of some of these models.

The DNA double helix may be treated as an isotropic elastic rod [Fig. 13(a)]. In the framework of this model, the DNA molecule is described by only three parameters: bending and torsional rigidities and the diameter. The model has proved to be extremely useful to analyze hydrodynamics and other properties of linear DNA, when it behaves as a usual polymeric molecule. It also permitted provision of the comprehensive theoretical treatment of DNA topology of both levels, knotting (Klenin *et al.*, 1988; Shaw and Wang, 1993) and supercoiling (Vologodskii *et al.*, 1992).

A quite different, but also very successful, model treats the DNA double helix as consisting of base pairs of two types: closed and open [Fig. 13(b)]. This is the helix-coil model, which has permitted quantitative explanation of all major features of DNA melting (Wada and Suyama, 1986).

The polyelectrolyte model [Fig. 13(c)] treats DNA itself just as a charged cylinder but allows for mobile counterions surrounding the double helix (Frank-Kamenetskii *et al.*, 1987).

To predict RNA structure, a simplified model of RNA folding is widely used [Fig. 13(d)].

There are many other theoretical models of DNA and RNA. Usually, the theoretical models are treated within the framework of statistical mechanics. However, in many cases, such consideration has proved to be insufficient because very slow relaxation is rather typical for nucleic acids. Such slow relaxational processes were very thoroughly studied both theoretically and experimentally in the case of DNA melting (Anshelevich *et al.*, 1984). One cannot correctly describe RNA folding without allowance for slow relaxational processes. Such processes are very essential in many cases of formation of DNA

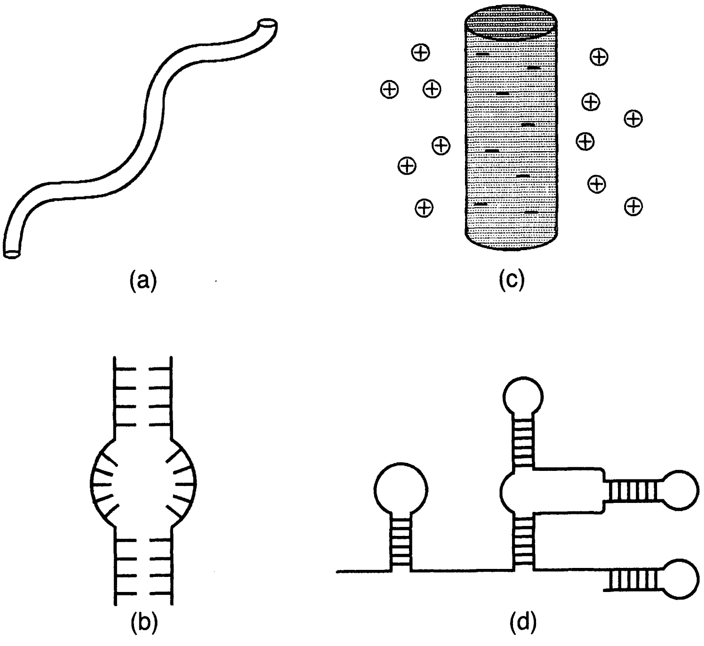

FIG. 13. Theoretical models of nucleic acids: (a) elastic-rod model, (b) helix-coil model, (c) polyelectrolyte model, (d) model of RNA folding.

unusual structures under negative superhelical stress.

2.2 Special Methods

In this section, we consider the most important methods specially developed to study nucleic acids. These methods have been introduced relatively recently (in the past 20 years), but, in many cases, they have pushed aside the traditional methods.

2.2.1 Gel Electrophoresis Gel electrophoresis is a simple technique, introduced in the early 1970s, which truly revolutionized the study of nucleic acids and, subsequently, the whole field of molecular biology. All developments in this field in the past 20 years are connected, directly or indirectly, with the gel electrophoresis method. Gel electrophoresis has pushed aside ultracentrifugation as the method used to separate nucleic acids.

2.2.1.1 Background. Gel electrophoresis differs from electrophoresis in solution only in the nature of the medium in which molecules are separated in the electric field. In the case of gel electrophoresis, the medium is a gel, a polymer network. The most popular in the field of nucleic acids are gels made of polyacrylamide or agarose. Originally, the great advantage of gels in the separation of nucleic acids was discovered purely empirically. The understanding came later after some ideas of P.-G. De Gennes were borrowed from polymer physics, namely the notion of reptation of polymer molecules.

As in regular electrophoresis, the electrophoretic mobility μ is defined as the proportionality coefficient between velocity of movement, v, and the electric field E:

$$v = \mu E. \tag{3}$$

The electric force applied to a nucleic acid of length L is proportional to L (because each residue carries a negative charge). Hence

$$\mu \sim LD, \tag{4}$$

where D is the diffusion coefficient:

$$D = \langle x^2 \rangle / \tau, \tag{5}$$

where τ is the characteristic time for a nucleic acid molecule to go out of its original "tube" and $\langle x^2 \rangle$ is the mean-square shift of the molecule after it leaves the tube.

All movements other than within the "tube" are forbidden in the gel. As a result, the molecule experiences Brownian motion only along its own axis (i.e., within the "tube"). Because the friction in the course of such movement is proportional to L,

$$\tau \sim LL^2 = L^3. \tag{6}$$

For the ideal polymer coil,

$$\langle x^2 \rangle \sim L,$$

and we finally obtain

$$\mu \sim 1/L, \tag{7}$$

whereas, without the gel, similar considerations would lead to the lack of dependence of μ on L. Equation (7) explains why the gel is so efficient a medium to separate molecules of nucleic acids according to their lengths during electrophoresis.

The consequences of Eq. (7) are really far-reaching. The entire idea of genetic engineering, i.e., reshuffling of DNA pieces extracted from different organisms, became feasible only after two major breakthroughs: the discovery of restriction enzymes, which cut long DNA molecules into shorter pieces recognizing special short nucleotide sequences, and the implementation of gel electrophoresis to separate the pieces obtained after cutting. Each piece forms its own band in the gel after the electric field is switched off. Then the gel is cut by an ordinary razor to obtain one unique piece of DNA.

2.2.1.2 Regular Gel Electrophoresis and DNA Sequencing. Gel electrophoresis is indispensable in all methods of DNA sequencing, which have revolutionized the entire field of biology. We explain the basic idea of application of regular gel electrophoresis to DNA sequencing using the chemical method of Maxam and Gilbert (1977) as an example. We explain the essence of the techniques using the example of single-stranded DNA fragments. The technique also happens to be applicable to double-stranded DNA fragments.

So, suppose we have a sample consisting of identical single-stranded DNA fragments of an unknown sequence. First, using a specific enzyme, radioactive phosphorus ^{32}P is attached to one (particular) end of the fragment. We shall refer to the molecule's labeled end as the beginning, the other end being unlabeled. The sample is then divided into four parts.

Then a substance which breaks the DNA strand after an adenine nucleotide (A) is added to the first fraction. In so doing, the reaction conditions are chosen in such a way that, on the average, about one A per fragment will be attacked during the reaction. The reagent will transform the original mixture of uniform-length fragments into one of fragments of varying length. In the process, we are concerned solely with the labeled fragments. If, say, adenines occupy positions 1, 3, 7, 13, 21, 25, and 26 on the original fragment, we may assume that the reagent is added to result in the appearance of labeled fragments with lengths of 1, 3, 7, 13, 21, 25, and 26 nucleotides. Not a single fragment of a different length is expected to appear.

Similarly, the three other parts of the original solution are treated with substances breaking the chain after T, G, and C. After this, all four samples are separated in a parallel manner in the same apparatus for gel electrophoresis. After the electric field is cut off, a photographic plate is placed on the gel's surface to imprint on it the labeled bands in the gel.

The experiment's result is diagrammed in Fig. 14. The sequence that can be read directly on the electrophoregram is shown to the left. (Actually, three, rather than four, different reagents are sufficient.)

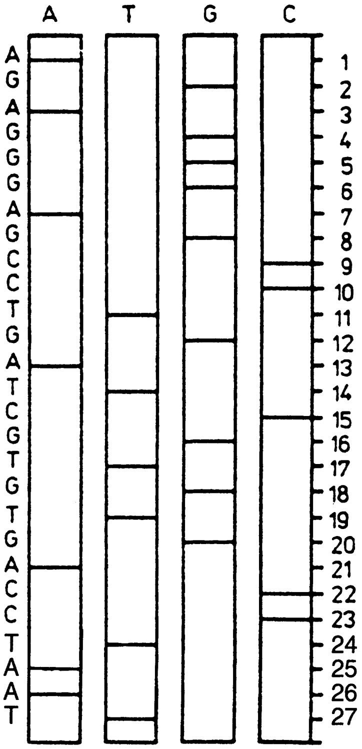

FIG. 14. The Maxam–Gilbert method (chart).

The DNA sequencing techniques have been developed to perfection, based on different conceptual approaches. The most fertile approach proved to be that of Sanger *et al.* (1977), who developed a method of DNA sequencing that uses enzymes working on DNA. Instead of radioactive labels, the approach uses nucleotide analogs linked to different dyes, which makes sequencing an ecologically much cleaner technique.

2.2.1.3 Pulse Electrophoresis. Even gel electrophoresis has its limitations. According to Eq. (7), with increasing length of molecules, their electrophoretic mobility decreases. Therefore, to separate very long molecules in a practically acceptable time scale, one needs to increase the electric field. However, the treatment in Sec. 2.2.1.1 is valid for the case of very small electric fields, which do not deform the molecules [otherwise Eq. (4) would not be valid]. At high fields, the DNA molecule (we talk about DNA because only DNA molecules could be so long that the problems arise) straightens along the electric field. As a result, it moves not like a polymer coil but like a rodlike, straight object. The electrophoretic mobility of such a molecule does not depend on its length independently of whether it moves in pure solvent or in gel, because, in this case, the friction and the driving force are both proportional to L. Thus, in a strong field, separation with respect to length occurs only for a short time before the DNA molecules are straightened. It is totally senseless to conduct electrophoresis longer than this time because all molecules, independently of their length, will just shift by the same distance. Does this mean that long DNA molecules cannot be separated in gel?

Schwartz and Cantor (1984) found a simple way out of the deadlock. If, soon after the molecules are straightened, the direction of the electric field is significantly changed, say by 90°, then, before the molecules will be straightened in the new direction, they assume again the shape of a polymer coil and will be separated for the same short time as while moving in the first direction. After straightening in the second direction, the field again is switched to the first direction, etc. As a result of such cycles or pulses, the molecules effectively move in the diagonal direction, and the separation takes place throughout the duration of the experiment.

Pulse electrophoresis dramatically increased the range of lengths of DNA molecules that can be separated in gels. The method permits separation of entire chromosomal DNA molecules. Implementation of this method opened the way to such ambitious projects as the Human Genome Project, which is designed to sequence the entire human genome (see Sec. 4.2.6).

2.2.1.4. Separation of DNA Topological Conformers. In the field of DNA topology and supercoiling, gel electrophoresis created a kind of revolution similar to those in other branches of nucleic acid research and applications. Although for a different reason than linear DNA, closed circular DNA molecules belonging to different topological classes move in gel with different velocities. As a result, knots of different types can be separated in gel (Shaw and Wang, 1993). The same is true for different topoisomers, ccDNA molecules differing in the linking number *Lk* (see Fig. 15). There it is not the *Lk* value according to which DNA molecules are actually separated, but the writhing, i.e., the spatial shape of the double helix (see Sec. 1.3.2). More precisely, the mobility depends on the absolute value of writhing.

2.2.1.5. Two-Dimensional Gel Electrophoresis. The regular, one-dimensional gel electrophoresis does not separate supercoiled molecules that have the same absolute value of the number of superhelical turns τ but different sign because those molecules have the same absolute value of writhing. When a structural transition into an unusual structure occurs, although the *Lk* value does not change, both twisting and writhing change (their sum, which is the linking number, remaining unchanged). As a result, a topoisomer carrying an unusual structure may move in gel with the same speed as another topoisomer without an unusual structure. To avoid such confusion, two-dimensional gels are used.

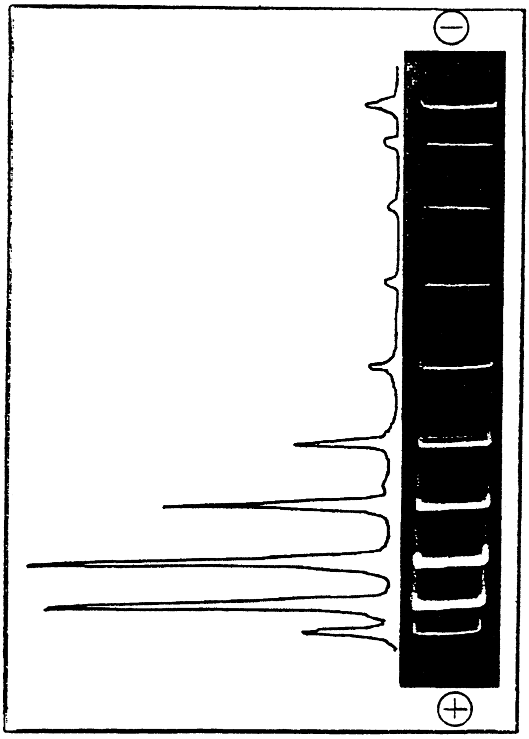

FIG. 15. Separation of DNA molecules differing by the number of superhelical turns, done by the gel electrophoresis technique. The experiment was conducted with DNA of a small pA03 plasmide, containing 1683 nucleotide pairs. Originally, the molecules were inserted at the top, near the negative electrode (the place is not shown in the figure).

A specially prepared mixture of different topoisomers of one and the same DNA, carrying an insert capable of changing into an alternative structure, is placed in the left top angle of a quadrangular gel plate (see Fig. 16). Then an electric field is applied to force DNA molecules to move from top to bottom along the left edge of the plate. Following the separation of topoisomers in the first direction, the gel is saturated with chloroquine molecules, which lessen superhelical stress. The chloroquine concentration is chosen in such a way as to make the superhelical stress insufficient for the formation of an unusual structure. After that, the direction of the electric field is changed to force molecules to move from left to right. As a result, the sequence of spots in the second direction corresponds to the topoisomers' sequence.

The uppermost spot in Fig. 16 corresponds to zero topoisomer, i.e., to a relaxed and nonsuperhelical DNA. The spots coming from that spot clockwise correspond to positive topo-

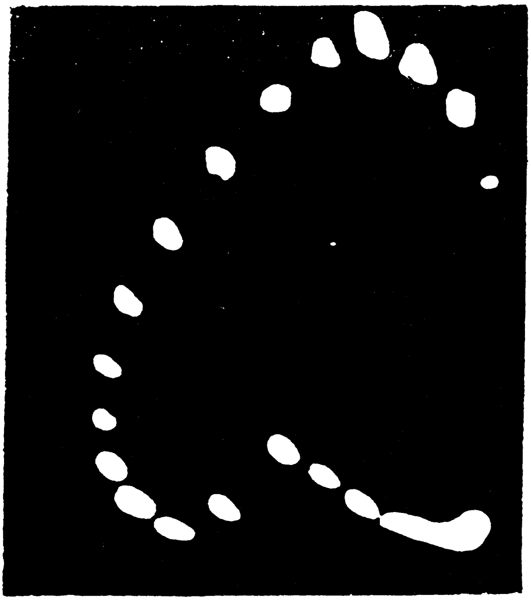

FIG. 16. A typical pattern of two-dimensional gel electrophoresis, observed during formation in DNA of an unusual structure.

isomers; those going anticlockwise, to the negative. One can clearly see the mobility drop, observed in this case between −10 and −12 topoisomers. This means that in topoisomers −12, −13, . . ., an unusual structure is present, while in topoisomers . . ., −9, −10, it is absent. Topoisomer −11 occupies an intermediate position: It carries the unusual structure during, roughly, half the time of its movement in the gel.

2.2.2 Chemical, Photochemical, and Enzymatic Probing A great variety of special approaches have been elaborated to study nucleic acid structures (Lilley and Dahlberg, 1992a,b). They are grounded on the different reactivities of nucleic acids adopting different structures with respect to chemical, photochemical, and enzymatic reactions. In many cases, these methods make it possible to arrive at very specific conclusions about the structure of a particular region of a nucleic acid under conditions that totally exclude application of not only x-ray crystallography or NMR but even spectroscopy and other indirect physical methods. Sometimes, the methods under consideration are applicable even *in vivo*.

To explain the general ideology underlying these methods, let us consider a specific example. In Sec. 1.4.4, we mentioned intermolecular triplexes, which are formed between homopurine-homopyrimidine regions of DNA and corresponding homopurine or homopyrimidine oligonucleotides. If such a complex is actually formed, the reactivity of the N7 position of guanine should dramatically decrease because this nitrogen is sheltered in the triplex by the Hoogsteen pairing (see Fig. 10).

The chemical agent used is dimethyl sulfate (DMS), which reacts with the N7 position of guanine, alkylating it. This alkylation occurs in single-stranded as well as in duplex nucleic acids. However, it cannot take place in triplexes. As a result, in the complex of duplex DNA with an oligonucleotide, which forms a triplex, all guanines in the duplex outside the triplex zone will be alkylated by DMS, whereas guanines within the triplex zone will remain unmodified.

Then the DNA piece under study is end-labeled and subjected to hot piperidine treatment, just as in the Maxam–Gilbert sequencing technique (actually, it is DMS that is used in the Maxam–Gilbert method as a reagent on guanines). Piperidine will convert the sites of alkylated guanines into chain breaks. Such breaks will never occur in the triplex zone. Figure 17 shows the pattern that is obtained after separation of the fragments in gel and radioautography.

The above example is a specific case of the footprinting assay. Such assays can be applied, for instance, to complexes of DNA with proteins to find out which sequences are recognized by the proteins. Instead of DMS, DNAase I, which cuts the uncovered DNA duplex, is often used. The yield of some photoproducts that can also be converted to strand breaks dramatically decreases when a duplex region is covered by a protein or an oligonucleotide. Hence the photo-footprinting assay is useful.

Some chemical reagents, like diethyl pyrocarbonate, potassium permanganate, and osmium tetroxide, do not react with bases in the double helix but react with open bases. The products can be converted into chain breaks. These reagents are widely used to detect open regions. Single-strand–specific nucleases, which digest single strands but do not digest duplex, are used in a similar way.

Chemical, photochemical, and enzymatic

probing are extremely powerful methods of detecting unusual structures, such as Z-DNA, cruciforms, H-DNA, and G quadruplexes (Lilley and Dahlberg, 1992a,b).

3. BIOLOGICAL FUNCTIONS OF NUCLEIC ACIDS

Nucleic acids occupy the most basic position in all living organisms. In this section, we briefly consider how nucleic acids fulfill the most significant of their numerous physiological functions.

3.1 Replication

The most basic function of nucleic acids consists in storage and reproduction of genetic information, which is encoded in the sequence of nucleotides. In principle, one strand of a nucleic acid carries the complete information, and many viruses actually carry only single-stranded DNA or RNA molecules. However, in all living cells, information is stored and reproduced in the form of the DNA double helix.

3.1.1 DNA Replication The ability of nucleic acids to replicate is inherent in the complementary principle. In autonomous living organisms, from unicellular up to the human, genetic information is stored in the form of duplex DNA. Schematically, replication consists in separation of the complementary DNA strands and synthesis on each of the strands of complementary ones (see Fig. 18). In practice, the process is conducted by a special replicating machinery consisting of a number of proteins. The central protein of the machinery is the DNA polymerase enzyme. Other important proteins are single-strand–binding (SSB) and DNA-helicase, which unwind the double helix. Topological problems, which are inevitable because of the necessity of unwinding of the two parent strands, are solved by topoisomerases.

Replication is initiated by binding to the specific sites on DNA of special short RNA oligonucleotides, the primers. On one of the two DNA strands, the synthesis of the complementary strand is carried out continuously. On the opposite strand, the synthesis is often interrupted because the polarity of the second strand is opposite to the direction of movement of the replication machinery along DNA. The gaps between fragments formed on the second strand are filled by DNA ligase.

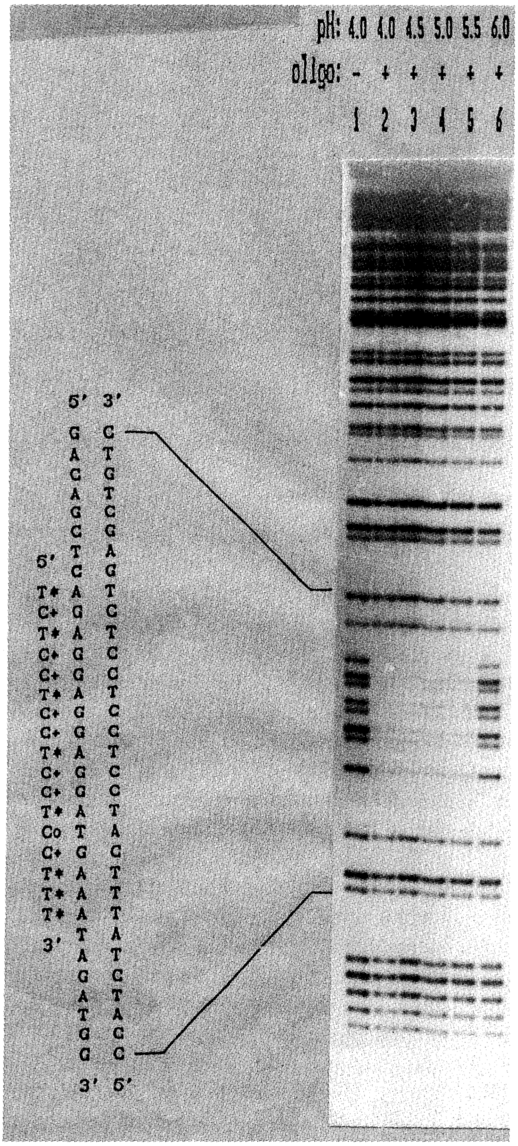

FIG. 17. The result of a footprinting experiment with dimethyl sulfate of a complex of duplex DNA carrying a homopurine-homopyrimidine insert with corresponding pyrimidine oligonucleotide. The data are from Cherny *et al.* (1993).

3.1.2 Replication of Viruses The above scheme is universal for cells. Viruses, the cellular parasites, cannot reproduce themselves autonomously. They penetrate the cell and

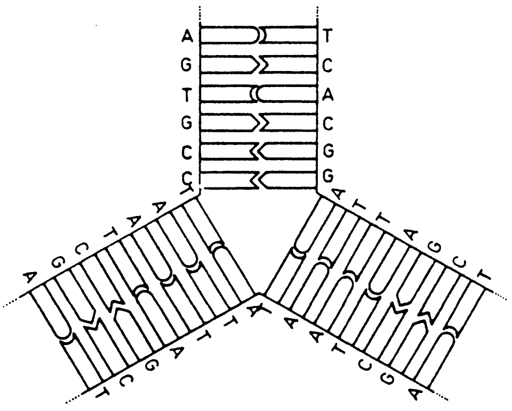

FIG. 18. Replication of duplex DNA.

replicate using the cellular resources. Viruses exhibit extreme versatility in the forms of storage of genetic information and, accordingly, in the fashion of replication. The genetic material of viruses can be in the form of linear duplex DNA, closed circular duplex DNA, linear single-stranded DNA, closed single-stranded DNA, or, last, linear single-stranded RNA.

Some RNA-containing viruses replicate with the help of RNA-dependent RNA polymerase. In many cases, however, first DNA is synthesized on the RNA template by the special enzyme reverse transcriptase, or RNA-dependent DNA polymerase. Then the closed circular double-stranded, so-called replicative, form of DNA is obtained. Then DNA replicates, and, in the final stage, virus RNA is synthesized on the DNA template. This complicated process is typical for retroviruses, the most important class of viruses from the medical viewpoint, which comprises many oncogenic (cancer inducing) viruses and HIV, the AIDS virus.

3.2 Transcription

Genetic information is expressed via the process of protein synthesis. The first step of this process is synthesis of the RNA copy of a gene, i.e., of messenger RNA (mRNA). This process is significantly different for prokaryotes and eukaryotes.

3.2.1 Transcription in Prokaryotes Prokaryotes generally obey the central dogma of molecular biology mentioned in the beginning of this chapter. The special enzyme RNA polymerase recognizes its binding sites on duplex DNA (promoters), binds to them, and synthesizes the full-length copy of the gene. The process is regulated by various mechanisms. Positive regulation occurs either via synthesis of entirely new RNA polymerase, which recognizes other promoters, or via protein factors, which bind with RNA polymerase and change its specificity. Negative regulation occurs via special sites on DNA (operators) to which proteins, called repressors, can bind. This binding prevents RNA polymerase from synthesizing RNA.

Messenger RNA serves as a template for protein synthesis on ribosomes. In prokaryotes, these two processes are coupled (see Fig. 19). The transcription process is often associated with formation of positive supercoils ahead of RNA polymerase and negative supercoils behind it. These topological problems are resolved by topoisomerases (Cozzarelli and Wang, 1990).

3.2.2 Transcription in Eukaryotes In eukaryotes, the process of synthesis of messenger RNA is much more complicated than in prokaryotes. First of all, in eukaryotes, DNA sequences corresponding to protein molecules are not continuous. They are interrupted by other sequences. The pieces corresponding to proteins have been named exons. The interrupting sequences were named introns. The role of introns still remains obscure, although the intron-exon or-

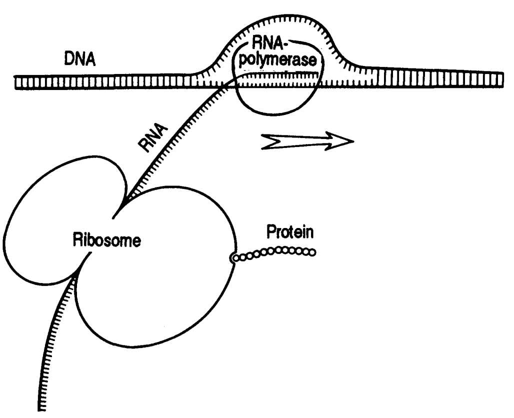

FIG. 19. RNA polymerase moves along DNA, synthesizing RNA. The ribosome copies information from RNA by synthesizing protein, in accordance with the genetic code.

ganization of eukaryotic genes has been known since the late 1970s.

The eukaryotic cell has several different RNA polymerases and numerous proteins (transcription factors) that modulate their activity and specificity. As a result, the mRNA synthesis is carried out by a complicated machinery consisting of many protein molecules.

The full-length copy of the entire sequence of the gene, including all its exons and introns, is synthesized by the transcription machinery. This full-length copy of the gene, which is called pre-mRNA, is converted into mRNA through a special maturation process, or splicing. In the course of splicing, all introns are deleted, and exons are "spliced" to form the maturated molecule of mRNA with the sequence corresponding to the amino-acid sequence of the protein molecule.

This ready-made mRNA is transported, in the form of a complex with special proteins, from the nucleus to the cytoplasm, where it strips the transporting proteins and serves as a template in the process of protein synthesis on ribosomes, i.e., translation. Thus, in contrast to prokaryotes, in eukaryotes transcription and translation are completely separated from each other.

Protein synthesis is extensively regulated on the mRNA level. Special protein factors, which bind with mRNA recognizing its various structural elements, modulate the rate of degradation of mRNA and its affinity to ribosomes. Thus, the yield of a particular protein is regulated the entire way from DNA to the spot where the protein molecule is synthesized, and even beyond this (see the next section).

3.3 Translation Process and Genetic Code

The translation process is the final stage of the protein synthesis. In this process, the genetic code plays the key role. The message brought by mRNA in the form of a nucleotide text is read by a ribosome, and the corresponding protein molecule is synthesized. Because a nucleotide text consists of four different letters, whereas sequences of proteins consist of 20 different amino-acid residues, the genetic code consists of triplets; i.e., one amino acid is encoded by a sequence of three adjacent nucleotides, which are called codons. The entire number of codons is $4^3 = 64$. Therefore, the genetic code is degenerate, several codons corresponding to one amino acid. There are also termination codons. There are no initiation codons, their role under appropriate conditions being played by the codons AUG and GUG, which normally correspond to methionine and valine. Figure 20 shows the genetic code.

In the realization of the translation process, the tRNA molecules play the key role. Each tRNA molecule carries an anticodon, the triplet complementary to the codon. A special enzyme recognizes the particular tRNA molecule and "charges" it with the corresponding amino acid. The ribosome creates conditions for binding (via codon-anticodon complementary recognition) of mRNA codons with the charged tRNA molecules and then adds to the growing protein chain the amino-acid residue brought by the tRNA molecule.

After the protein chain is synthesized, special large protein molecules, called chaperones, create conditions for folding of the chain into a biologically active spatial structure. After this, the protein molecule is ready to fulfill its function. However, even after the protein molecule is ready, its rate of degradation continues to be controlled by the cell.

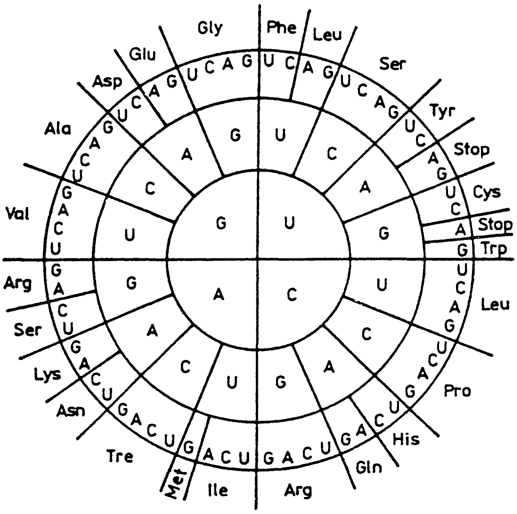

FIG. 20. The genetic code. The first letter of the codon is located in the central ring, the second in the first ring, and the third in the second. Written in the outer ring are the abbreviated designations of amino acids.

4. BIOTECHNOLOGY

The advent of the biotechnology era is the result of the entire accumulated knowledge about nucleic acids and the enzymes working with them. However, among countless achievements that have led to biotechnology and genetic engineering, there are a very few that have most contributed to the elaboration of techniques permitting interference into the most basic processes of life. In this section, we briefly survey these landmarks and then outline the major achievements and prospective of biotechnology.

4.1 Major Tools of Genetic Engineering

The biotechnology revolution is grounded on genetic engineering, a set of relatively simple chemical, biochemical, and biophysical techniques that have been developed during the past two decades. Some of them, like gel electrophoresis, are explained in other sections of this article. In this section, we briefly survey other very significant methods of genetic engineering.

4.1.1 Chemical Synthesis of DNA Achievements in the chemical synthesis of DNA had led, by the early 1980s, to the creation of robots that quickly and efficiently synthesize DNA oligonucleotides containing up to several dozen residues. This opened the way for obtaining synthetic genes. However, the most widespread application of DNA synthesis consists in the creation of so-called DNA probes, relatively short pieces that serve as primers for polymerases in numerous projects, most notably in PCR machines (see below).

4.1.2 Restriction Enzymes These enzymes recognize short sequences of duplex DNA (usually containing six base pairs) and specifically cut the duplex at these sites. Restriction enzymes are, along with gel electrophoresis (see Sec. 2.2.1), the major tools of genetic engineering.

4.1.3 Reverse Transcriptase The intron-exon organization of the genes of eukaryotes prevents their direct utilization in genetic engineering. When transferred to prokaryotes, these genes just do not express because, in prokaryotes, the mRNA maturation process cannot occur. Therefore, eukaryotic genes suitable for transfer to prokaryotes should not contain introns. This is often achieved by extracting maturated mRNA and synthesizing on it the corresponding DNA with the help of reverse transcriptase (RNA-dependent DNA-polymerase).

4.1.4 Plasmids Usually, alien genes are introduced not into the chromosome of a prokaryotic cell but into special small circular molecules, plasmids, which multiply together with the cell. Genetic engineers have constructed many artificial plasmids, which carry multiple recognition sites for restriction enzymes. When DNA fragments obtained as the result of cutting by a restriction enzyme are mixed with a plasmid cut with the same enzyme, the result is new plasmids carrying the fragments inserted into them. This is most common way of obtaining recombinant DNA molecules.

4.1.5 Polymerase Chain Reaction (PCR) This technique has revolutionized genetic engineering in recent years and had an impact comparable to that of the discovery of restriction enzymes and the implementation DNA sequencing methods. It makes it possible to multiply pieces of DNA starting from literally a single DNA copy.

The amplification is performed in a special PCR machine. The machine is loaded with the DNA molecule containing the piece to be amplified (it could be really only *one* DNA molecule to begin with), thermally stable *Taq* DNA-polymerase, and two DNA primers specially synthesized for the particular piece. One primer is complementary to the left-hand flanking sequence of the piece, the second one to the right-hand flanking sequence, but to the opposite strand. The PCR machine works as a cycler, each cycle consisting in heating the sample up to subboiling temperature (94 °C) with subsequent cooling down to 72 °C (the optimal temperature for *Taq* polymerase isolated from thermophilic bacteria). As a result of the first heating, the original DNA molecule melts, and the two strands separate (see Sec. 1.4.1). The subsequent cooling entails the binding of the primers, which are present in macroscopic quantity, to the complementary sites on the two single-stranded DNA molecules. The polymerase immediately starts to extend the primers from their 3′ ends using the complementary strands as the templates. Thus, the first cycle ends up with two partially double-stranded DNA molecules. The common duplex part of these two molecules encompasses only the piece to be amplified plus the two flanking primers. In the subsequent heating-cooling cycles, only this common part is amplified. After *n* cycles, one obtains 2^n duplex molecules corresponding to the piece flanked by the two primers.

4.2 The Coming of the Biotechnology Age

Biotechnology applications of the methods of genetic engineering are truly breathtaking and hardly embraceable. We provide a very brief survey of only some of them.

4.2.1 Protein Superproduction By incorporating genes of virtually any organism into bacteria, genetic engineers have made it possible to produce any protein in considerable quantities. On the one hand, this has enormously stimulated basic research on proteins and their functions, and, on the other hand, it has opened fantastic opportunities for producing proteins for medicine.

4.2.2 Protein Engineering A special technique has been elaborated to substitute nucleotides in cloned genes. The method consists in synthesis of a DNA probe that embraces the site to be changed and carries the desired substitution. This oligonucleotide is annealed with the single-stranded copy of the gene cloned in a special vector DNA. Then DNA polymerase is used to synthesize the double-stranded DNA, which will carry a mismatch in the site of substitution. After replication of the vector within the cell, two versions of the gene are obtained: the original, wild-type copy and the changed, mutant copy. The mutant copy is selected by special technique, again using the DNA probe.

As a result of the advent of protein engineering, unprecedented opportunities have appeared to change purposefully the amino-acid sequences of proteins. This has completely changed the way to study protein structure and function.

4.2.3 New Pharmacology Genetic engineering has made it possible to obtain in large quantities any proteins, even those that are present in the entire human body in very few copies. This opportunity is widely used by pharmaceutical companies to produce a rapidly increasing variety of biologically active proteins unavailable before, such as human growth hormone, human insulin, human interferon, etc.

4.2.4 Forensic Applications The implementation of PCR and other routine techniques of manipulation with DNA promises to revolutionize forensic science. Indeed, all people, with the exception of identical twins, differ from each other in a significant part of their DNA texts. So if a criminal left a tiny portion of his cells in any form (blood, sperm, etc), his DNA can be multiplied by PCR and then analyzed and compared with the DNA analysis of suspects. Although the new DNA forensic methods still need development in order to be widely applied, there is no doubt that they will become very useful in the nearest future.

4.2.5 Transgenic Animals Using microinjections of alien DNA into the cell nucleus and other methods, researchers routinely modify cells' genetic material. If cells are modified at the early stage of their differentiation, entire animals with directionally changed genomes can be produced. This has opened breathtaking opportunities for studying the role of separate genes on the level of the entire organism. Special laboratory animals with significant modifications of heredity, which are not encountered in natural animals, have been engineered. This can lead to a revolution in methods of creation of new species of domestic animals.

4.2.6 Human Genome Project The most ambitious idea that has emerged from achievements in nucleic acid research was sequencing the entire human genome, i.e., the Human Genome Project. The Human Genome Project is expected to be completed in the very early years of the next century. This landmark achievement will undoubtedly lead to a new revolution in medicine.

4.2.7 Diagnostics and Correction of Genetic Defects As a result of the Human Genome Project and related studies on human genetics, remarkable progress is underway in the tracing and sequencing of genes responsible for various hereditary diseases. In parallel, more data have been accumulating about a substantial genetic component in such very common diseases as cancer and atherosclerosis. As a result, the coming revolution in medicine is foreseen as the development and implementation in humans of methods of diagnosis and correction of wrong genes. Very efficient diagnostic methods have been developed and implemented for some genetic defects—for instance, in the case of sickle-cell anemia.

However, efficient methods to correct the established genetic defects are still lacking. Their development is the most challenging unsolved problem in the field of nucleic acids.

GLOSSARY

Adenine: A chemical group that is part of DNA and RNA. It is one of the four bases of nucleic acids. The abbreviated designation is A.

Amino Acid: A chemical compound with the structure $H_2N–CHR–COOH$, where R is any radical. Amino acid is the starting product for protein synthesis.

Amino-Acid Residue: A chemical group with the structure $–HN–CHR–CO–$, which is a residue of protein chain. It is what remains of an amino acid when it is built into the protein chain.

Base (Nucleic or Nitrous): A class of chemical compounds that includes adenine, guanine, thymine, cytosine, and uracil.

Chaperones: Specialized proteins that carry out inside the cell the folding of a polyamino-acid chain, newly synthesized on a ribosome, into a native protein molecule.

Chimerical DNA: An artificial molecule, put together by genetic engineering techniques with sections of different natural DNA. The terms "hybrid" and "recombinant" DNA have the same meaning.

Codon: A term connected with the genetic code. Denotes a trinucleotide corresponding to one amino-acid residue. There are several meaningless (nonsense) codons not corresponding to any amino acid. They play the role of stop signals during protein synthesis by mRNA on ribosomes. They are called terminating codons. Initiating codons serve as signals for starting protein synthesis.

Coil (Polymer): A notion of polymer physics. Serves to denote the form a polymer molecule assumes in space. In consequence of thermal motion, the form of a polymer coil constantly changes.

Cruciform: A DNA structure that can be formed in palindrome sequences.

Cytosine: A chemical group that is part of DNA and RNA. One of the four bases of nucleic acids. The abbreviated designation is C.

Degeneracy of the Code: One of the properties of the genetic code, which is that several codons may correspond to the same amino acid.

Deoxyribonucleic Acid: The full name of the DNA molecule.

DNA-Polymerase: An enzyme responsible for DNA synthesis on a DNA template. The process is called DNA replication.

Eukaryotes: Organisms provided with a cell nucleus.

Exon: A DNA segment that stores information about a part of the amino-acid sequence of protein.

Gel: A polymer network saturated with a solvent.

Gene: The principal notion of classical genetics which for a long time understood the term to mean an indivisible particle of heredity. During the 1950s and 1960s, the word "gene" was taken to mean a continuous DNA section, recorded on which, in the form of a nucleotide sequence, is information about the amino-acid sequence of one protein. The word is still used as the name for a DNA segment. But in some cases, it denotes a continuous segment that corresponds to only a part of the protein chain, while in others, it means a set of segments corresponding to a complete protein molecule. And it may well be that the same DNA segment simultaneously belongs to two and even three genes.

Genome: The entirety of genetic information of the organism.

Guanine: A chemical group that is part of DNA and RNA. It is one of the four nucleic acid bases. The abbreviated designation is G.

Hybrid DNA: An artificial molecule put together through genetic engineering techniques with segments of different natural DNA. The terms "recombinant" and "chimerical" DNA have the same meaning.

Intron: A DNA section dividing exons.

Ligase: An enzyme curing ruptures in DNA.

Linking Number: A quantitative characteristic of the degree to which two contours are linked. The linking number equals the number of times that one contour pierces the surface spread tightly across the other contour. It is denoted by *Lk*.

Nuclease: An enzyme that splits DNA or RNA.

Nucleotide: A residue of DNA and RNA.

Oligonucleotide: A short single-stranded piece of nucleic acid.

Oncogene: A gene causing cancer.

Palindrome: A phrase that reads the same from left to right or from right to left. In the context of DNA texts, such palindromes are called "mirror repeats." A real DNA palindrome is a segment of double helix that has the same sequence when reading either strand in the same direction, dictated by the chemical structure of DNA strands—for example,

→
ATGCGCAT
....
TACGCGTA
←

Plasmid: A circular DNA molecule multiplying together with bacteria and capable of passing from cell to cell.

Primer: A single-stranded oligonucleotide that binds, via complementary pairing, to DNA or RNA and serves for the priming of polymerases working on both DNA and RNA.

Prokaryotes: Single-cell organisms with no cell nuclei.

Promoter: A DNA segment with which RNA polymerase binds to set in motion mRNA synthesis.

Purine: A class of chemical compounds that includes adenine and guanine.

Pyrimidine: A class of chemical compounds that includes thymine, uracil, and cytosine.

Recombinant DNA: An artificial molecule put together by genetic engineering techniques from segments of different natural DNA. The terms "hybrid" and "chimeric" DNA have the same meaning.

Replication: Doubling of genetic material. DNA synthesis on DNA.

Repressor: A protein binding very strongly with an appropriate DNA segment between the promoter and the gene itself. By associating with DNA, repressor prevents the progress of RNA polymerase from promoter to gene and thus blocks mRNA synthesis. Serves to regulate transcription.

Reptation: Snakelike movement of a polymer molecule through the polymer network. The notion is crucial for the theoretical understanding of DNA separation in gel.

Restriction Endonuclease: An enzyme cutting the double helix in places with a definite nucleotide sequence. It is the chief instrument of genetic engineering, and often is called just restriction enzyme.

Restriction Fragment: A piece of DNA cut out of the molecule with the help of restriction endonucleases.

Reverse Transcriptase: An enzyme responsible for DNA synthesis on an RNA template. The process is called reverse transcription.

Ribosome: An involved complex of RNA and proteins, which is responsible for the translation process in the cell.

RNA Polymerase: An enzyme synthesizing mRNA on a DNA template. Carries out the transcription process.

Splicing: The process of mRNA maturing in eukaryotes, which results in the discharge of introns, while exons combine into one RNA chain.

Superhelicity: A characteristic of a circular closed DNA. Superhelicity arises when the linking number Lk in the DNA differs from the value N/Γ_0, where N is the number of base pairs in the DNA and Γ_0 is the number of base pairs per turn of double helix in linear DNA placed in the same conditions.

Template: A polymer molecule whose sequence is used for charting a sequence for another polymer molecule. DNA serves as a template for DNA synthesis during replication and for RNA during transcription. RNA serves as a template for protein synthesis during translation and for DNA during reverse transcription.

Topoisomerases: A class of enzymes that change the topology of a circular closed DNA.

Topoisomers: DNA molecules identical in chemical terms but different in topology (by the type of knot, or by the linking number).

Transcription: RNA synthesis on a DNA template. Reverse transcription is DNA synthesis on a RNA template.

Translation: Protein synthesis by an mRNA template on ribosome.

Thymine: A chemical group that is part of DNA. Is one of four DNA bases. The abbreviated name is T.

Uracil: A chemical group that is part of RNA. It is one of four RNA bases. The abbreviated designation is U.

Vector: A genetic engineering term. This is the name of a carrier DNA molecule (plasmid, virus, and so on), within which the desired gene is cloned.

Virus: A cellular parasite, one of the simplest objects of living nature. Outside the cell, a virus is a molecular complex consisting of nucleic acid (DNA, sometimes RNA) and several proteins forming the virus membrane. The cell's penetration by a virus (or its nucleic acid) switches the resources of the cell over to synthesizing the virus nucleic acid and proteins. When the cell's resources are exhausted, its membrane bursts, spilling out "ready-made" virus particles. Animals' viruses have a much simpler structure than bacteria viruses (bacteriophages). Animal viruses are incapable of injecting their nucleic acid into the cell, and they get inside the cell together with food. Viruses cause many infectious diseases, such as the grippe, smallpox, poliomyelitis, hepatitis, and AIDS. In some cases, finding itself inside the cell, the virus does not undo it but builds its DNA into that of the cell, after which the virus DNA begins to multiply together with the cell DNA. The behavior of the cell itself may, however, undergo a sharp change in the process.

Works Cited

Anshelevich, V. V., Vologodskii, A. V., Lukashin, A. V., Frank-Kamenetskii, M. D. (1984), *Biopolymers* **23**, 39–58.

Bustamante, C., Keller, D., Yang, G. (1993), *Curr. Opinion Struct. Biol.* **3**, 363–372.

Cherny, D. I., Malkov, V. A., Volodin, A. A., Frank-Kamenetskii, M. D. (1993), *J. Mol. Biol.* **230**, 379–383.

Cozzarelli, N. R., Wang, J. C. (Eds.) (1990), *DNA Topology and Its Biological Effects*, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Dickerson, R. E. (1992), in: D. M. J. Lilley, J. E. Dahlberg (Eds.), *DNA Structures. Part A. Synthesis and Physical Analysis of DNA*, Methods in Enzymology Vol. 211, San Diego: Academic, pp. 67–111.

Dubochet J., Adrian, M., Dustin, I., Furrer, P., Stasiak, A. (1992), in: D. M. J. Lilley, J. E. Dahlberg (Eds.), *DNA Structures. Part A. Synthesis and Physical Analysis of DNA*, Methods in Enzymology Vol. 211, San Diego: Academic, pp. 507–518.

Dunlap, D. D., Garcia, R., Schabtach, E., Bustamante, C. (1993), *Proc. Natl. Acad. Sci. USA* **90**, 7652–7655.

Felsenfeld, G., Davies, D. R., Rich, A. (1957), *J. Am. Chem. Soc.* **79**, 2023–2025.

Frank-Kamenetskii, M. D., Anshelevich, V. V., Lukashin, A. V. (1987), *Usp. Fiz. Nauk* **151**, 595–618. [*Sov. Phys. Usp.* **30**, 317–330].

Frank-Kamenetskii, M. D. (1990a), in: W. Saenger (Ed.), *Landolt-Börnstein Series, Group VII: Biophysics, Nucleic Acids, Subvol. 1C*, Heidelberg: Springer, pp. 228–240.

Frank-Kamenetskii, M. D. (1990b), in: N. R. Cozzarelli, J. C. Wang (Eds.), *DNA Topology and Its Biological Effects*, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory, pp. 185–216.

Gueron, M., Kochoyan, M., Leroy, J.-L. (1987), *Nature* **328**, 89–92.

Gehring, K., Leroy, Y.-L., Gueron, M. (1993), *Nature* **363**, 561–565.

Ivanov, V. I., Krylov, D. Yu. (1992), in: D. M. J. Lilley, J. E. Dahlberg (Eds.), *DNA Structures. Part A. Synthesis and Physical Analysis of DNA*, Methods in Enzymology Vol. 211, San Diego: Academic, pp. 111–127.

Johnson, W. C. (1990), in: W. Saenger (Ed.), *Landolt-Börnstein Series, Group VII: Biophysics, Nucleic Acids, Subvol. 1C*, Heidelberg: Springer, pp. 1–59.

Kang, C. H., Zhang, X., Ratliff, R., Moyzis, R., Rich, A. (1992), *Nature* **356**, 126–131.

Klenin, K. V., Vologodskii, A. V., Anshelevich, V. V., Dykhne, A. M., Frank-Kamenetskii, M. D. (1988), *J. Biomol. Struct. Dyn.* **5**, 1173–1185.

Lilley, D. M. J., Dahlberg, J. E. (Eds.) (1992a), *DNA Structures. Part A. Synthesis and Physical Analysis of DNA*, Methods in Enzymology Vol. 211, San Diego: Academic.

Lilley, D. M. J., Dahlberg, J. E. (Eds.) (1992b), *DNA Structures. Part B. Chemical and Electrophoretic Analysis of DNA*, Methods in Enzymology Vol. 212, San Diego: Academic.

Maxam, A. M., Gilbert, W. (1977), *Proc. Natl. Acad. Sci. USA* **74**, 560–564.

Rajagopal, P., Feigon, J. (1989), *Nature* **339**, 637–640.

Roberts, R. W., Crothers, D. M. (1992), *Science* **258**, 1463–1466.

Sanger, F., Nicklen, S., Coulson, A. R. (1977), *Proc. Natl. Acad. Sci. USA* **74**, 5463–5466.

Schwartz, D. C., Cantor, C. R. (1984), *Cell* **37**, 67–75.

Shaw, S. Y., Wang, J. C. (1993), *Science* **260**, 533–536.

Vologodskii, A. V., Levene, S. D., Klenin, K. V., Frank-Kamenetskii, M. D., Cozzarelli, N. R. (1992), *J. Mol. Biol.* **227**, 1224–1243.

Wada, A., Suyama, A. (1986), *Prog. Biophys. Mol. Biol.* **47**, 113–157.

Wang, A. H.-J., Quigley, G. J., Kolpak, F. K, Crawford, J. L., van Boom, J. H., van der Marel, G., Rich, A. (1979), *Nature* **282**, 680–685.

Watson, J. D., Crick, F. H. C. (1953), *Nature* **171**, 737–738.

Further Reading

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J. D. (1989), *Molecular Biology of the Cell*, 2nd ed., New York: Garland.

Cantor, C. R., Schimmel, P. R. (1980), *Biophysical Chemistry*, Parts 1–3, San Francisco: Freeman.

Cozzarelli, N. R., Wang, J. C. (Eds.) (1990), *DNA Topology and Its Biological Effects*, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Frank-Kamenetskii, M. D. (1993), *Unraveling DNA*, New York: VCH.

Lewin, B. (1990), *Genes IV*, Oxford: Oxford University.

Lilley, D. M. J., Dahlberg, J. E. (Eds.) (1992), *DNA Structures. Part A. Synthesis and Physical Analysis of DNA*, Methods in Enzymology Vol. 211, San Diego: Academic.

Lilley, D. M. J., Dahlberg, J. E. (Eds.) (1992), *DNA Structures. Part B. Chemical and Electrophoretic Analysis of DNA*, Methods in Enzymology Vol. 212, San Diego: Academic.

Ptashne, M. (1992), *A Genetic Switch*, 2nd ed., Palo Alto, CA: Cell Press & Blackwell.

Watson, J. D., Gilman, M., Witkowski, J., Zoller, M. (1992), *Recombinant DNA*, 2nd ed., Scientific American Books, New York: Freeman.

Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A., Weiner, A. M. (1987), *Molecular Biology of the Gene*, 4th ed., Menlo Park, CA: Benjamin/Cummings.

Weaver, R. F., Hedrick, P. W. (1992), *Genetics*, 2nd ed., Dubuque, IA: Wm. C. Brown.

NUMERICAL METHODS

See SCIENTIFIC COMPUTING BY NUMERICAL METHODS

OCCUPATIONAL HEALTH AND SAFETY

See SAFETY AND HEALTH IN THE WORKPLACE

OCEANOGRAPHY

Curtis A. Collins and Peter C. Chu, *Naval Postgraduate School, Monterey, California, U.S.A.*

	Introduction	47
1.	**Physical Characteristics of Seawater**	48
1.1	Equation of State	48
1.2	Hydrostatic Equation	49
1.3	Sound in the Ocean	50
1.4	Light in the Ocean	50
2.	**Ocean Structure**	51
2.1	Vertical Structure	51
2.2	*T-S* Diagram	53
2.3	Ocean Currents	53
3.	**Ocean Measurements**	53
3.1	Shipboard Profiling	54
3.2	Moored Instruments	55
3.3	Lagrangian Measurements	55
3.4	Satellite Measurements	56
3.4.1	Satellite Altimetry	56
3.4.2	Synthetic-Aperture Radar	56
3.4.3	Satellite Radiometer	57
4.	**Marine Hydrodynamics**	57
4.1	Coriolis Force	58
4.2	Pressure-Gradient Force	58
4.3	Tidal Forces	58
4.4	Gravity	59
4.5	Viscous Forces	59
4.6	Geostrophic Balance	59
4.7	β Effect	59
4.8	Conservation Principles	60
4.9	Topographic Forcing	60
5.	**Ocean Circulation Theory**	61
5.1	Wind-Driven Circulation	61
5.1.1	Ekman Transport	61
5.1.2	Sverdrup Transport	62
5.1.3	Westward Intensification	62
5.1.3.1	Stommel's Model	63
5.1.3.2	Munk's Model	63
5.2	Thermohaline Circulation	63
5.2.1	Thermocline Theory	63
5.2.2	Thermohaline Forcing	64
5.2.3	Ventilation	64
5.2.4	Convection	65
5.2.5	The Conveyor Belt	65
6.	**Ocean Waves**	65
6.1	Acoustic Waves	66
6.2	Gravity and Capillary Waves	66
6.3	Rossby Waves	67
6.4	Solitary Rossby Waves	68
6.5	Kelvin Waves	69
7.	**Instabilities, Eddies, and Fronts**	69
7.1	Rossby Radius of Deformation	69
7.2	Baroclinic Instability	69
7.3	Barotropic Instability	70
7.4	Mesoscale Eddies	70
7.5	Quasigeostrophic Turbulence	70
7.6	Submesoscale Coherent Vortices	71
7.7	Fronts	71
8.	**Small-Scale Turbulence and Mixing**	71
8.1	Instability and Turbulence	72
8.2	Generation and Maintenance of Turbulence	72
8.3	Statistical Characteristics of Turbulence	72
8.4	Double Diffusion	72
8.5	Microstructure	73
9.	**Concluding Remarks**	73
	Acknowledgments	74
	Glossary	74
	Works Cited	75
	Further Reading	75

INTRODUCTION

Oceanography deals with the application of fundamental scientific principles to the ocean environment. This results in a variety of subfields in oceanography (physical, chemical, and biological oceanography and marine geology), which must often be integrated to understand ocean processes. This report will focus on physical oceanography. Physical oceanography is the study of the physical structure, motion, and mixing of ocean waters.

The two traditional approaches used in

digital Encyclopedia of Applied Physics

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

physical oceanography are the descriptive approach and the theoretical approach. The descriptive approach uses observations to determine the structure of the ocean while the theoretical approach applies Newton's laws of motion and thermodynamical laws to understand how forces affect the structure of the ocean. These two approaches complement one another.

As a recognizable scientific discipline, oceanography is young, beginning with the voyage of *HMS Challenger* from 1872 to 1876. The first systematic large-scale expedition to study primarily the physical aspects of the ocean circulation was the German *Meteor* expedition to study the South Atlantic from 1925 to 1927. Worldwide surveys of the ocean were conducted during the International Geophysical Year in 1957–1958, and the first satellite mission that was dedicated to studies of the ocean was *Seasat*, which observed the oceans for 100 days in 1978.

This report is organized into nine sections. First, we describe the physical characteristics of seawater, then the character of the ocean's waters and currents, and then the instruments used in the practice of physical oceanography. Next, we introduce marine hydrodynamics and then develop the physics of the mean ocean circulation. Ocean waves are then described. The final sections deal with mesoscale motions, small-scale processes, and concluding remarks. A list of references and additional reading is provided for those who need more detailed information.

1. PHYSICAL CHARACTERISTICS OF SEAWATER

The density of water is determined by the temperature T, salinity S, and pressure p. Temperature is measured in degrees Celsius. Salinity represents the weight of dissolved solids (in grams) in a kilogram of seawater. In 1978, oceanographers adopted the Practical Salinity Scale, which defined a practical salinity unit (psu) in terms of the conductivity of seawater at 15 °C referenced to the conductivity of a potassium chloride solution at 15 °C (see UNESCO, 1981). This provided an accurate, reproducible primary standard and a better measure of the effect of the dissolved solids on density, and supported *in situ* measurement methods. The unit of pressure used by oceanographers is the decibar, db, where 1 db = 10 000 Pa.

In addition to modifying the conductivity of water, salt modifies the physical characteristics of water in another important way. The temperature of maximum density of seawater decreases with increasing salinity at a rate greater than the freezing point, so that for salinity greater than 24.7 psu, seawater freezes before the maximum density is reached. Other physical properties of water, e.g., viscosity, surface tension, and thermal conductivity, are also modified.

Typical ranges of salinity and temperature found at the sea surface in the open ocean are 32–37 psu and −1 to 29 °C, respectively. Salinities lower than these values are associated with melting ice, rivers, or intense rainstorms. Larger salinities are encountered in semienclosed evaporative basins such as the Red Sea or Mediterranean. Since water is slightly compressible, *in situ* temperature increases below about 3500 db because of adiabatic effects. These adiabatic changes can be taken into account by using a potential temperature, θ, which nominally uses the sea surface as a reference pressure. The mean temperature and salinity for the world ocean are 3.5 °C and 34.7 psu.

1.1 Equation of State

The equation relating density to temperature, salinity, and pressure is called the *equation of state*. The equation of state in current use, EOS80, was established by the Joint Panel on Oceanographic Tables and Standards (UNESCO, 1981) and uses the functional form

$$\rho(S,T,p) = \frac{\rho(S,T,0)}{1 - p/K(S,T,p)}, \qquad (1)$$

where $\rho(S,T,0)$ is the density of seawater at one standard atmosphere and $K(S,T,p)$ is the bulk modulus, and both are nonlinear functions of S, T, and p. Typical density values for the ocean range from 1020 to 1060 kg/m^3. The density is usually represented by a density anomaly γ:

$$\gamma(S,T,p) = \rho(S,T,p) - 1000\ \mathrm{kg/m^3}. \qquad (2)$$

The nonlinearity of the equation of state has several important properties. These can be deduced from the curved shape of constant-density lines on the *T-S* diagram (Fig. 1). At temperatures close to zero, the constant-density lines become nearly vertical, indicating that salinity changes have a much greater effect on density than they do at warmer temperatures. At warmer temperatures, the slope of these constant-density lines decreases, indicating that temperature changes contribute more to density variability. When two different water masses of the same density but differing temperature and salinity are mixed, temperature and salinity are conserved so that the resulting water mass will lie on a line connecting these two points on the *T-S* diagram; but, as indicated by the curve of the constant-density lines, the resulting mixture will be heavier (more dense) than the original waters. This provides a mechanism of generating internal buoyancy changes that is called either cabelling or thermobaric instability.

1.2 Hydrostatic Equation

In most oceanographic applications, vertical accelerations can be neglected so that pressure and depth z are related by the hydrostatic equation

$$\frac{\partial p}{\partial z} = -\rho g, \tag{3}$$

where g represents gravitational acceleration. Zonal variations of g are usually considered, and g is estimated as follows:

$$g(\phi) = 9.790\,318(1.0 + 5.2788 \times 10^{-3} \sin^2\phi + 2.36 \times 10^{-5} \sin^4\phi)\ \mathrm{m^2s}, \tag{4}$$

where ϕ is latitude (in degrees). Note that 1 db corresponds to about 1 m and that 1010 db is about 1000 m.

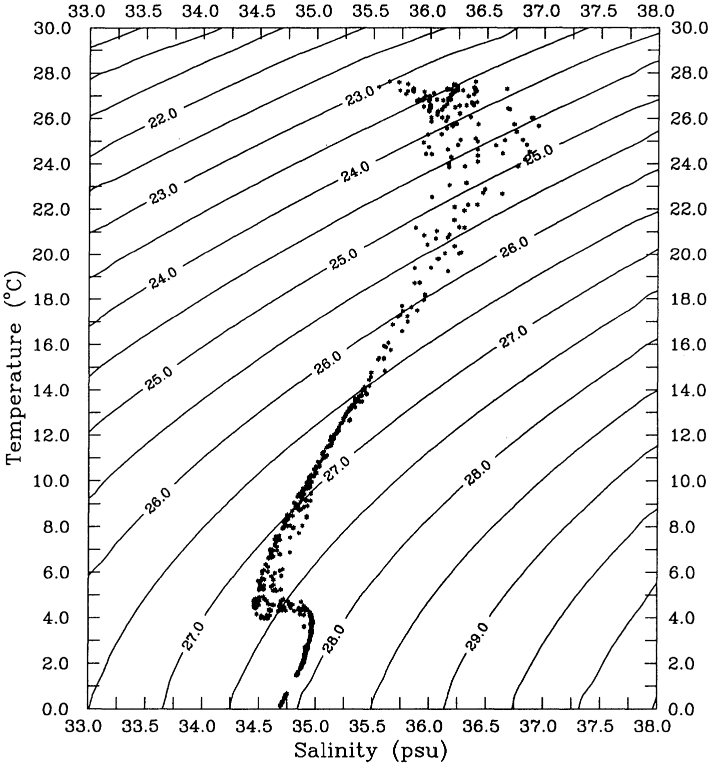

FIG. 1. *T-S* diagram for waters of the Tropical Atlantic along 23° W from 6° N to 9° S. Readily identifiable water masses include subtropical subsurface water at 26 °C, 36.5 psu, Antarctic intermediate water at 4 °C, 34.5 psu, North Atlantic deep water at 4 °C, 35 psu, and Antarctic bottom water at 2 °C, 35 psu. The curved lines are contours of a density anomaly, γ_t which is referenced to atmospheric pressure and has units of kg/m^3.

1.3 Sound in the Ocean

The speed of sound in seawater, U_a, in units of m/s, is given by

$$U_a^2 = \left(\frac{\partial p}{\partial \rho}\right)_a, \tag{5}$$

where $(\partial p/\partial \rho)_a$ is the reverse adiabatic density gradient. In practice, the speed of sound is usually not computed in this manner; instead, empirical formulas are used that express sound velocity as a function of S, T, and p and that are derived from laboratory measurements. Typical sound-speed values range from 1400 to 1700 m/s. To a first approximation, sound is nondispersive in the ocean. Refraction of sound is important, creating both convergence and shadow zones. A simple model for the attenuation of sound (which does not take into account geometric spreading, refractive loss, and volume and boundary scattering loss) is

$$I(x) = I_0 e^{-\kappa x}, \tag{6}$$

where $I(x)$ is the intensity of sound at distance x from the source and I_0 is the intensity of sound at the same frequency at the location of the source. The attenuation coefficient κ, in units of dB/m, is given as

$$\kappa = 4.343\left[\frac{1.7 \times 10^8(\frac{4}{3}\mu_F + \mu_F')f^2}{\rho_F c_F^3} + \left(\frac{SA'f_{rm}f^2}{f^2 + f_{rm}^2}\right)(1 - 1.25 \times 10^{-2}p) + \frac{A''f_{ra}f^2}{f^2 + f_{ra}^2}\right],$$

where μ_F is the dynamic coefficient of shear viscosity for fresh water, μ_F' is the dynamic coefficient of bulk viscosity for fresh water, f_{rm} is the relaxation frequency (kHz) for magnesium sulfate, f_{ra} is the relaxation frequency (kHz) for boric acid, A' is 2.03×10^{-5} dB/(kHz psu m), A'' is 1.2×10^{-4} dB/(kHz m), S is salinity (psu), and f is the frequency in kHz. At intermediate frequencies (100–1000 kHz), absorption is mainly due to boric acid and magnesium sulfate chemical relaxation processes. At low frequencies (below 100 Hz), attenuation is almost independent of frequency, and the attenuation coefficient is about 0.003 dB/km. This means that low frequencies can easily insonify the ocean, i.e., 15 W of transmitted power can be detected at a distance of 2000 km. In part, this is because sound speed increases with increasing pressure, creating a duct that focuses acoustic energy (called the SOund Fixing And Ranging or SOFAR channel) at about 1-km depth.

Current applications include not only finding submarines and fish, and mapping the seafloor, but inverse problems including acoustic tomography and ocean acoustic thermometry. The latter application may lead to the first quantitative measure of "greenhouse" warming of the ocean.

1.4 Light in the Ocean

The behavior of light in the ocean affects not only the heating of the upper layer but also the growth of plants (and hence the oceanic food chain). Light also has marine applications for charting, and finding fish, submarines, and mines, or measuring bioluminescence.

Seawater has both inherent and apparent optical properties (see OPTICS, UNDERWATER). Inherent properties can be measured on a seawater sample in the laboratory (if, of course, suspended particles, etc., were not affected by removal from the sea). Apparent properties must be measured *in situ*. The inherent properties include the beam attenuation coefficient c, the volume absorption coefficient a, and the volume scattering coefficient b, where $c = a + b$. A beam transmissometer is used to measure c (m^{-1}) over a fixed distance, so that for 1 m

$$T = I_1/I_0 = e^{-c}; \tag{7}$$

and c may be written

$$c = c_\lambda = a_{\lambda w} + a_{\lambda d} + a_{\lambda p} + b_{\lambda w} + b_{\lambda d} + b_{\lambda p}, \tag{8}$$

where the subscript λ indicates dependence on wavelength, w indicates pure water, d is dissolved matter, and p is suspended particles.

Apparent properties include the radiance (W m^{-2}) and the diffuse attenuation coefficient for downwelling (indicated by a superscript consisting of an arrow pointing downward) irradiance, $k_\lambda^\downarrow$ (W m^{-2}), which are both strongly λ dependent. Downwelling irradiance attenuation $k_\lambda^\downarrow(z)$, where z is depth, better represents the depth to which a lidar can detect an object than does $c_\lambda(z)$. It is always true that $a_\lambda < k_\lambda < c_\lambda = a_\lambda + b_\lambda$. A good estimate of the upwelling solar irradiance $k_\lambda^\uparrow$ can be obtained from satellite observations of the sea surface (see Sec. 3.4.3). The apparent properties are connected to the inherent ones via the radiative transfer equation [see Eq. (13)].

Within the ocean, light is mostly absorbed but is also scattered, by both seawater and particles within the water including plankton. The absorption results in heating of the water. The euphotic zone refers to the upper layer of the ocean in which the intensity of light is greater than 1% of its value at the sea surface. The attenuation of light with depth depends not only on the wavelength of light but also the clarity of the water. The clarity is mostly controlled by biological activity although at the coast, sediments transported by rivers or resuspended by wave action can strongly attenuate light.

In terms of clarity, waters may be classified according to Jerlov (1976) as coastal (types 1–9) or oceanic (types I–III). For oceanic waters, blue light (400–500 nm) is transmitted to the greatest depth, while coastal waters favor transmission of green-yellow light (650–580 nm) because plankton and organic compounds absorb blue light. Diffuse attenuation coefficients may vary from 0.01 to 1 m^{-1} for the oceanic and coastal regimes, respectively, in the latter case limiting light to the upper few meters.

2. OCEAN STRUCTURE

The physical structure of the ocean is largely governed by the shape of the ocean basins and atmospheric forcing and to a lesser extent by mixing and internal dynamic processes. Continents prevent ocean waters from freely circulating around the Earth except in the Southern Ocean between 56° S and 63° S. A system of midocean ridges splits the major oceans (Pacific, Atlantic, Southern, and Indian) into a series of smaller basins. In addition to these ridges, major topographic features include abyssal plains, trenches, submarine canyons, seamounts, and islands. The mean depth of the ocean is 3.8 km; the greatest depth, 12 km, is found in Marianas Trench. Note that these dimensions, e.g., 4 km in the vertical and 4000 km or more in the horizontal, imply aspect ratios of 1/1000 for vertical sections presented in an atlas or report. Semienclosed seas such as the Mediterranean, Red, Black, Baltic, Arctic, Japan, and Bering Seas or gulfs such as the Persian, Mexico, or California exchange water with adjacent oceans to varying degrees, often introducing waters of distinct characteristics.

Next to the coast, shallower waters overlie the continental shelf. The continental shelf is found inshore of a sharp break in topography that occurs at an average depth of 130 m. Continental shelves are regions of high fishery productivity, oil production, and recreational boating. Because of the shallow water, circulation on the continental shelf can have a different character from that found offshore. Oceanographers refer to the region that is found further inshore, where ocean swell begins to feel the bottom in about 10 m of water, as the "nearshore." Here surf forms, and waves run up onto the beach. Other distinct geographical features along the coast include bays, estuaries, fjords, and river mouths. Note that continental shelves are separated from deep ocean basins by a region where depth increases rapidly, which is called the continental slope.

The atmosphere drives the ocean both through the stress that the winds exert upon the water and through the buoyancy effects of heat, evaporation, and rainfall. Over much of the globe, the atmosphere has a distinct meridional character consisting of trade winds in the tropics and westerlies in the midlatitudes. In the Indian Ocean, these are modified by the Southwest Monsoon, which begins in late May, and the Northeast Monsoon, which begins in October. Continents serve to modify air masses further in such a way that cold, dry air may be found over the ocean off east coasts during the winter.

2.1 Vertical Structure

A major feature of the ocean is its layered structure. Radiation impinging on the ocean surface heats both the ocean surface and the lower atmosphere. As noted above, seawater is a good absorber of radiation, so that typically 60% of the penetrated radiation is absorbed within the first meter of the ocean, around 80% is absorbed within 10 m, and only about 1% of the entering radiation remains at 140-m depth in the clearest tropical ocean waters. The turbulent fluxes of heat, water, and momentum across the ocean surface result in vigorous mixing in the uppermost layer of the ocean. The turbulent kinetic energy generation by the wind stress and the upward buoyancy flux will mix the seawater vertically, eventually forming a vertically uniform layer called the upper mixed layer.

Below the upper mixed layer, the water can be divided into two zones in terms of the rate at which temperature decreases with depth: a region of rapid change of temperature with depth, called the thermocline, and a region where temperature changes slowly with depth, called the deep zone (Fig. 2). In mid- and high latitudes, the thermocline can be subdivided into an upper or seasonal thermocline and a lower or permanent thermocline.

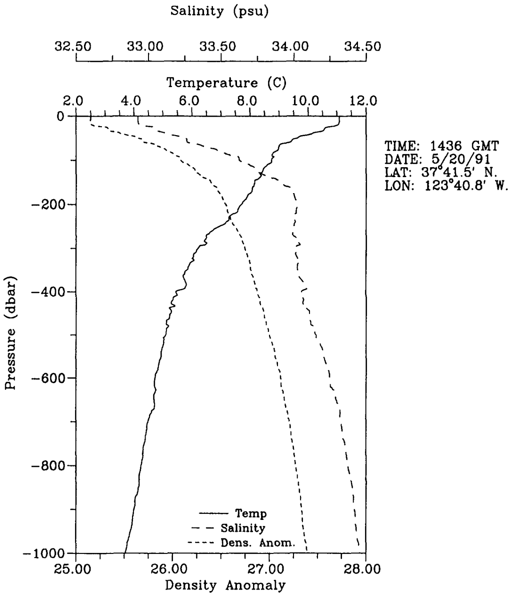

FIG. 2. Vertical profiles of temperature, salinity, and density for an oceanographic station off San Francisco, California. An oceanographic station consists of observations of temperature and salinity at a fixed location. Data were obtained by lowering an instrument that measures conductivity, temperature, and pressure (or depth), called a CTD, from the *R/V Pt. Sur* on May 20, 1991. The mixed layer is confined to the upper 15 m, and the seasonal thermocline/halocline/pycnocline extends to a depth of 80–100 m at this location.

At low and mid latitudes, despite a steady input of heat and downward mixing by eddy diffusion, the depth of the thermocline does not appear to change at any given location. It has been postulated that this is due to upward advection of water of the order of cm/day due to sinking of water at high latitudes.

Density increases with depth and has the same kind of vertical zonation as temperature. The zone of rapid density increase is called the pycnocline. The stability E is defined as

$$E = -\frac{1}{\rho}\frac{\partial \rho}{\partial z} - \frac{g}{U_a^2}. \tag{9}$$

The first term is the *in situ* vertical density gradient and the second term the adiabatic gradient ($\simeq 400 \times 10^{-8}$ m^{-1}). The requirement of static stability means that E must be positive. Situations where E is negative and the water column is not stable can exist, e.g., convection due to surface cooling or subsurface hot springs; when using a surface reference pressure, deep densities (below 4 or 5 km) can appear to be less than those at shallower depths, but this is due to inaccuracy of EOS80 and the situation is corrected by using a reference pressure that is near the observed density inversion. Equation (9) can be expanded to include vertical gradients of temperature and salinity:

$$E = \alpha\left(\frac{\partial T}{\partial z} + \Gamma\right) - \tilde{\beta}\frac{\partial S}{\partial z}, \tag{10}$$

where $\alpha = -\rho^{-1}\partial\rho/\partial T$ is the thermal expansion coefficient of seawater at the surface, $\tilde{\beta} = \rho^{-1}\partial\rho/\partial S$ is the haline contraction coefficient, and Γ is the adiabatic temperature gradient. This indicates that subsurface warming (called a temperature inversion) does not violate the requirement for stability if it is less than the adiabatic gradient or the salinity increases sufficiently.

The vertical distribution of salinity cannot be summarized as easily as that for density and temperature. Upper-layer salinity tends to be high in low to mid latitudes, decreasing to subsurface minimum at 600–1000 m. In the tropics, a shallow salinity maximum is usually found at about 100-m depth.

At high latitudes, upper ocean waters are fresher than those at depth. The region of rapid salinity change with depth is called the halocline. Since salinity usually does not dominate density changes, it can be used to trace the horizontal flow pattern.

2.2 T-S Diagram

The temperature-salinity (*T-S* or *θ-S*) diagram is used as a characteristic diagram by oceanographers. This is because temperature and salinity are conservative properties when a water mass is not in the mixed layer. This means that water masses must acquire temperature and salinity properties at the ocean surface and that these properties are determined by the local climate; when the water later sinks along density surfaces, it carries these properties with it. As a consequence, only a limited number of temperature and salinity combinations occur. These can be recognized by plotting observations for a series of oceanographic stations as a function of temperature and salinity.

A water type appears as a point on a *T-S* diagram, and a water mass is represented by a line and is formed from the mixing of water types. Using this representation, the global ocean has a distinct geography with specific regions readily identifiable by their *T-S* characteristics (Fig. 1).

2.3 Ocean Currents

In the mean, a steady ocean circulation exists. This circulation is the result of both thermohaline forcing and forcing by the wind. Currents are classified by their depth (surface, intermediate, deep, and bottom) or by their forcing mechanism. This is not unambiguous because some currents extend from the surface to the bottom of the ocean. The pattern of surface currents was first studied because of their effect on navigation. Today, the principal focus for such studies is to understand the poleward transport of heat by the ocean.

The steady-state circulation is characterized by subpolar and subtropical gyres with zonal tropical currents and an eastward flow around Antarctica called the Circumpolar Current. The strongest currents occur as poleward flow at the western boundaries of subtropical gyres and include the Gulf Stream in the North Atlantic, the Kuroshio in the North Pacific, and the Agulhas in the South Indian Ocean. The eastern-boundary currents of subtropical gyres are weaker, flow equatorward, and are characterized by upwelling at the coast. These include the California Current in the North Pacific, the Canary Current in the North Atlantic, the Benguela Current in the South Atlantic, and the Humboldt or Peru Current in the South Pacific. In the South Indian Ocean, the eastern boundary current, called the Leeuwin Current, flows poleward.

The system of equatorial currents includes the westward-flowing North and South Equatorial Currents. These currents are separated by the eastward-flowing North Equatorial Counter Current, which, in the Pacific, is found between 4° N and 10° N. At the equator, a strong subsurface eastward flow is usually found between 75 and 200 m depth and is called the Equatorial Undercurrent. It is usually about 400 km wide, and speeds in excess of 1 m/s are not uncommon.

The Indian Ocean circulation is dominated by the monsoons. A strong western-boundary current, the Somali Current, develops in response to the Southwest Monsoon. The monsoon currents in the North Indian Ocean reverse and flow in the direction of the monsoon. When eastward winds occur at the equator, a strong eastward current develops at the surface as well.

Intermediate, deep, and bottom currents are the result of thermohaline processes. These currents are also intensified at western boundaries.

3. OCEAN MEASUREMENTS

The equations of motion and state specify the variables that must be measured experimentally. These include temperature, salinity, and pressure for the equation of state and the velocity of the water. Related variables include the height of sea level and wind stress. Although the ocean environment is hostile, e.g., cold, great pressures, and often large accelerations at the sea surface, recent technical developments have greatly improved our ability to obtain data of the requisite accuracy and precision for study of most physical processes. Our major challenge is to obtain sufficient synoptic data to be able to identify unambiguously the processes that control ocean circulation.

3.1 Shipboard Profiling

Instruments that are routinely used from shipboard include conductivity-temperature-depth (CTD) profilers, expendable bathythermographs (XBT), and acoustic-Doppler current profilers (ADCP). The ship itself is a key element of these systems. Research vessels range in length from 30 to 100 m and have special capability for maneuvering at low speed, maintaining station, and handling equipment over the side. Protection from wind and seas is provided for scientists working on deck, and special electronic, biological, and chemical laboratories are needed. Larger ships must be capable of extended cruises of more than 40 days at sea and 20 000-km ranges.

CTDs are lowered by a winch and conducting cable when the ship is hove to (CTDs can also be towed, cycling up and down through the water column, but depth is limited to about 200 m by cable drag) (Fig. 3). Temperature measurement is by means of a platinum resistance thermometer (PRT) and a faster-response thermistor. Conductivity is measured by means of paired electrodes mounted inside a 4-mm-diam channel that is 30 mm long. Pressure is measured by a strain gauge. Data are sampled at 30 Hz and transmitted up the cable as a FM signal. The fast-response thermistor compensates for the mismatch in time constants of the PRT and the conductivity electrodes, allowing the computation of salinity without severe spikes. The accuracy of the system is 0.005 mS, 0.005 °C, and 0.1 of full scale for pressure. Corresponding precisions are 0.001 mS, 0.005 °C, and 0.1. The CTD is usually attached below a rosette that allows water samples to be collected at discrete depths by Niskin bottles for shipboard analysis. The Niskin bottles are open cylinders with endcaps, which are held together with an elastic band. During descent, the rosette holds the endcaps back, allowing water to flow freely through the bottles. During ascent, the rosette is electrically triggered, releasing the endcaps and closing the bottles. The Niskin bottles shown in Fig. 3 hold 5 L. Salts from these samples are necessary to maintain the accuracy of the CTD conductivity measurements. Final data processing produces a table of temperature and salinity values averaged over successive 2-db intervals. A typical CTD cast takes about 45 minutes per 1000 meters.

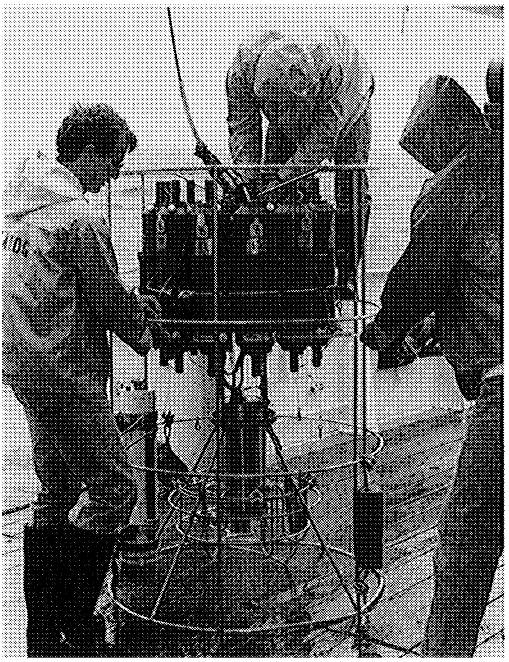

FIG. 3. Ocean profiling system. Vertical profiles of conductivity and temperature are obtained by a conductivity-temperature-depth (CTD) instrument and water sampling bottles. The CTD is the stainless steel cylinder that is mounted in the lower part of the cage. Water sampling (Niskin) bottles are held by a rosette at the top of the cage. The oceanographers standing on deck are filling seawater sample bottles from the Niskin bottles while the oceanographer standing on the cage is resetting the rosette.

XBTs are designed to be launched while a ship is under way. They were originally designed for use in antisubmarine warfare by naval vessels, but the simplicity and reliability have led to widespread use by merchant vessels in climate-related ocean studies. Temperature measurement is by means of a thermistor mounted on the nose of a 1-kg lead weight. Depth is determined by using an empirically determined rate of fall. Data are transmitted back to the ship by means of two-strand copper wire. XBTs are available for 450-m, 750-m, and 1500-m depths. Systematic errors of 0.2–0.4 °C occur in these measurements, and random errors are about half this amount; depth accuracy ranges from 2% to 10%, depending on the type (depth) of probe used (the increase with depth is due to errors in fall rate and microscopic leaks in wire insulation). Expendable probes that measure sound velocity and velocity shear are also available.

ADCPs are mounted in the hull of a ship, run continuously, look downward, and in good weather allow sampling to 400-m depth. Paired transducers look fore–aft and port–starboard, and the Doppler shift in gated returns (caused by plankton drifting in the water) is transformed to horizontal velocity. Typically, a pulse repetition rate of 1 Hz is used, and data are resolved in 8-m depth bins. Averaging over 3-min intervals reduces random errors to 1 cm/s. Current velocities are obtained by removing the ship's motion from ADCP output. ADCPs can also be lowered on a wire with a CTD, yielding velocity profiles that extend throughout the entire water column. The Global Positioning System (GPS) provides accurate navigational data for ADCP operation, including ship's heading.

3.2 Moored Instruments

Current measurements that are made at a fixed point in space are called "Eulerian" measurements. Eulerian measurements are made by mooring current meters. Currents are measured either

1. mechanically by means of orthogonal rotors or a rotor and vane,
2. acoustically using changes in travel time or an ADCP, or
3. electromagnetically in higher flow regimes associated with waves and surf.

In deep-sea applications, the mooring may be 4–6 km long, and is held to an anchor by means of a release that is interrogated acoustically from shipboard. A deep-sea mooring is typically left in place for two years.

Temperature and pressure measurements are easily made from moorings, but conductivity measurements of the precision and accuracy required for most deep-sea applications are not, because of problems of biological growth and electrode oxidation.

3.3 Lagrangian Measurements

Measurements that are of quantities that are traveling with the fluid are called "Lagrangian" measurements. Lagrangian measurement systems include surface (called drifters) and subsurface (called floats) drifting buoys. Surface drifting buoys must be properly coupled to the surface layer of the ocean, requiring a subsurface drogue that has a drag that is at least 40 times that exerted on the surface float. Drifter positions are currently determined by satellite at least six times per day with position accuracy (2σ) of 600 m. Future applications will use GPS with continuous position keeping of 100-m accuracy. Observations of near-surface temperature and vertical temperature gradients can also be made from drifters.

Subsurface floats are positioned acoustically. Floats that signal are called SOFAR floats while those that listen are called RAFOS floats. These floats may be isobaric or isopycnal or profile a short distance above and below these levels to measure potential vorticity. Moored autonomous listening stations (sound sources) or fixed Navy listening stations provide the requisite positioning network for SOFAR (RAFOS) floats. RAFOS floats surface when their mission is completed, relaying data via satellite.

A new system of Autonomous Lagrangian Circulation Explorer (ALACE) floats is currently being developed. These floats drift at depth but surface at regular intervals so that their position can be determined by satellite and subsurface ocean observations relayed to shore. Current floats are capable of 50–100 depth cycles. Further developments of this type of float, adding CTD, electromagnetic measurements, and perhaps "wings" to maneuver the float, promise to make ALACE a key component of a global ocean-observing system.

Anthropogenic chemical tracers are also an effective Lagrangian technique for determining ocean circulation. Tritium, produced by atmospheric atomic bomb testing, and cesium-137 and strontium-90, associated with nuclear power plant operations, have been used to trace horizontal and vertical circulation. Since tritium decays to helium-3 at a known rate, the ratio of tritium to helium-3 can be used to determine the length of time since the water parcel was exposed to the ocean surface; the utility of this technique is limited by the introduction of helium-3 into the ocean by submarine vents. Freon compounds (10, 11, 12) can also be used to trace and date ocean waters.

3.4 Satellite Measurements

Satellite techniques promise continuous global coverage. The principal problem with these techniques is that the atmosphere gets in the way and the ocean is nearly opaque to electromagnetic radiation. Satellite applications that we will discuss include satellite altimetry, synthetic-aperture radar, and ocean-surface temperature measurements. Other satellite techniques effectively map ice cover, and new satellite-positioning systems [the Global Positioning System (GPS)] and data relay capabilities are essential for modern experimental programs. Two kinds of sensors (active and passive) are available for measurements. The active sensors transmit radio signals, comparing the transmitted with the reflected or scattered signals that are received by the sensors from an area on the earth's surface. Passive sensors do not transmit a signal; instead, the emitted radiance from the earth's surface is compared to that received from space or a blackbody. Satellite altimetry and the synthetic-aperture radar are active sensors. Satellite radiometers are passive sensors.

3.4.1 Satellite Altimetry If the ocean were motionless, the shape of its surface would be determined entirely by the gravitational acceleration of the Earth. The spatial variation of the gravity field generates a sea-surface topography with a global range in height of nearly 160 m relative to the center of the earth. But forces associated with ocean currents produce departures from this static topography called *dynamic ocean topography*. The dynamic ocean topography has a global range in height of about 2 m, about 1% of the height range of the static or gravitational topography (Plate 1).

Satellite altimetry maps the dynamic ocean topography in three steps. First, the radar altimeter sends short pulses of microwave energy toward the ocean; the round-trip travel times of the reflected pulses yield the distance between the spacecraft and the sea surface (additionally, the shape of the reflected pulse can be used to determine wave height and sea-surface wind speed). Second, precise determination of satellite orbit permits these distances to be translated into a global map of sea level relative to the center of the Earth. Third, a map of the gravitational sea-surface topography (which must be obtained independently) must be subtracted from the satellite-derived topography, yielding the dynamic ocean topography. Methods for determining ocean current speed and direction from the dynamic ocean topography are described below.

Errors are associated with absorption of microwave energy by atmospheric water vapor and pulse-arrival delays due to electron release by ionization of gases in the ionosphere caused by sunlight. The largest source of error is associated with the determination of satellite orbit.

Satellite altimeter measurements have been made on Skylab (1973), Goes-3 (1975–1978), Seasat, and Geosat (1985–1989), and are currently being made by *Topex/Poseidon*. Topex has an orbital altitude of 1136 km and observes the Earth from 66° N to 66° S. Measurement tracks across the Earth are reproduced within 1 km over every 10-day repeat cycle. Widest spacing between tracks, at the equator, is 315 km. The altimeter operates at dual frequencies, 5.3 and 13.6 GHz, permitting correction for ionospheric electron delays. Topex also carries a microwave radiometer that operates at frequencies of 18, 21, and 37 Ghz to estimate total atmospheric water vapor content. Topex is tracked by three systems: ground-based lasers, measurement of Doppler shifts of microwave frequencies transmitted by a global network of ground-based beacons, and GPS. Sea-level topography is being measured with 2-cm accuracy.

3.4.2 Synthetic-Aperture Radar A synthetic-aperture radar (SAR) is a pulse-Doppler radar that transmits short radio pulses and maps the radar reflectivity of the sea over areas 50–100 km on a side, with a resolution of 10–40 m. From SAR images of ocean surface, we can "see" and study meanders of currents, eddies, storm waves, ice floes, leads, ice ridges, and polynyas. The perturbed sea surface ζ modulates the reflectivity of the radio pulse. The radar cross section becomes

$$\sigma = \langle\sigma\rangle + \delta\sigma, \tag{11}$$

where $\langle\sigma\rangle$ is the spatially averaged cross section per unit area of the sea, and $\delta\sigma$ is the deviation induced by surface long waves. A complex modulation transfer function R is used to link $\delta\sigma$ to ζ:

$$\delta\sigma/\langle\sigma\rangle = R\zeta. \tag{12}$$

When R is determined, the surface fluctuations can be measured by $\delta\sigma$.

Three processes cause variability in radar cross section: surface tilting, hydrodynamic interaction, and velocity bunching. The modulation caused by surface tilting and hydrodynamic interactions depends on the wave direction. Consider a SAR viewing the sea at an incidence angle of 23°: With an amplitude of 3 m and a wavelength of 200 m, the modulation caused by tilting is roughly ±3 dB for waves traveling toward the range direction (detectable by SAR), and nearly ±0.3 dB for waves traveling in the azimuthal direction (not detectable by SAR).

3.4.3 Satellite Radiometer Sea-surface temperature (SST) has a 150-year history of global measurement and is intimately linked with climate and global change. Although historical measurements were made by naval and merchant vessels, space-based measurements of SST have improved remarkably in the past decade and now represent one of our most important data sets for monitoring the global environment. Satellite measurement of SST is made possible by satellite radiometers (Plate 2).

The physical principle for the satellite radiometer is radiative transfer through an emitting and absorbing atmosphere with the absorption coefficient κ_λ, which is depicted by the radiative transfer equation

$$L^\lambda = L_S^\lambda \exp[-\tau(0,H)] + \int_0^H L_B^\lambda(z)\, \kappa_\lambda(z) \exp[-\tau(z,H)]\, dz, \qquad (13)$$

where

$$\tau(z,H) = \int_z^H \kappa_\lambda(z')\, dz'$$

is the optical thickness (or opacity) of the atmospheric layer with thickness H. The superscript λ indicates the wavelength of the radio waves. The terms L_S^λ, L^λ, and L_B^λ are the source radiance, the radiance seen by an observer viewing the source, and the radiance of the blackbody at the same temperature as the absorbing air, respectively. The first term on the right-hand side of Eq. (13) gives the total attenuation due to absorption; the second term gives the emission $L_B^\lambda \kappa_\lambda$ from any point z along the path attenuated by the air along the remainder of the path (Plate 2). The radiative transfer equation is complicated, and one needs to know L_B^λ (or temperature) and κ_λ (or atmospheric constituents) to solve it. At the atmospheric windows, however, the absorption coefficient κ_λ is small such that the second term of the right-hand side of Eq. (13) is negligible. Therefore, the source radiance (or the surface temperature) can be easily determined by the satellite radiometer.

The primary instrument for production of operational satellite-derived SST is the Advanced Very High Resolution Radiometer (AVHRR) carried on two polar-orbiting meteorological satellites. These satellites are sun-synchronous with local equator crossing times of 0730 and 1330 for descending orbits. The AVHRR scans the Earth perpendicular to the satellite direction of travel, providing twice-daily global coverage at 4-km ground resolution. The AVHRR has four or five channels, which use the 3–4-μm or 10–12-μm atmospheric windows, and carries an on-board blackbody viewed by the instrument, and this, along with a view of space, is used to perform a linear calibration. Multichannel observations are used to correct for atmospheric effects to produce a global SST data set. This is done after navigation of the images, land masking, eliminating observations of clouds, and averaging into 8-km areas. The bias and scatter of these data, when compared to drifting buoy measurements, are less than 0.3° and 0.7° C, respectively. Some of this difference is due to the fact that the radiometers measure the skin temperature of the ocean while buoy measurements are made in the upper meter of the ocean.

AVHRR images with 1-km resolution are broadcast directly to ground stations. These high-resolution images are useful for delineating fronts and eddies in the ocean that are associated with biological productivity and that also have tactical significance for naval warfare.

4. MARINE HYDRODYNAMICS

The equations of motion for the ocean are the momentum equations (application of Newton's law to a unit mass of seawater) and the continuity equation indicating conservation of mass per unit volume:

$$\frac{d\mathbf{V}}{dt} = \mathbf{F}_C + \mathbf{F}_p + \mathbf{F}_t + \mathbf{F}_g + \mathbf{F}_v, \tag{14}$$

$$\frac{d\rho}{dt} + \rho\nabla \cdot \mathbf{V} = 0, \tag{15}$$

where **V** is the velocity vector relative to the Earth, ρ is water density, and $\mathbf{F}_C$, $\mathbf{F}_p$, $\mathbf{F}_g$, $\mathbf{F}_t$, and $\mathbf{F}_v$ are various forces (per unit mass) exerted on the water parcels: Coriolis force, pressure-gradient force, gravity, tidal forces, and viscous forces.

4.1 Coriolis Force

The Earth is a rotating planet with the angular velocity of 2π/day. Any water parcel at rest relative to the Earth in fact moves with a velocity of as much as 460 m/s relative to a "fixed" star. Rather than include these velocities and accelerations, a coordinate system is generally chosen that is fixed with respect to the Earth. In this coordinate system, any moving water parcel "feels" an apparent force called the Coriolis force due to the rotation of the Earth:

$$\mathbf{F}_C = -2\mathbf{\Omega} \times \mathbf{V}, \tag{16}$$

where Ω is the angular velocity of the Earth.

4.2 Pressure-Gradient Force

If the pressure (p) on a level surface is inhomogeneous, the pressure gradient leads to a force exerted on the water parcel:

$$\mathbf{F}_p = -(1/\rho)\,\nabla p, \tag{17}$$

which is called pressure-gradient force.

4.3 Tidal Forces

In principle, the ocean responds to all gravitational fields exerted on it, but in practice, except for the Earth's own gravity, only the gravitational forces exerted by sun and moon are strong enough to induce detectable tidal excursions. If the ocean is attracted by the sun or moon, water parcels at different locations experience different forces that lead to a tidal force, which is decomposed into components normal and parallel to the ocean surface

$$\mathbf{F}_t = \frac{GMR}{D^3}\left(3\cos^2\chi - 1,\ -\tfrac{3}{2}\sin 2\chi\right). \tag{18}$$

Here G is the Newtonian gravitational constant, χ is the angle between the local vertical axis and the line between the center of the Earth and the external terrestrial body, M is the mass of the external terrestrial body, D is the distance between the Earth and the external terrestrial body, and R is the radius of the Earth. The 11 most important tides with commonly accepted nomenclature are listed in Table 1.

Table 1. Major tidal modes.

	Tidal mode	Period, T (h)
Semidiurnal Tides		
M_2	principal lunar	12.421
S_2	principal solar	12.000
N_2	elliptical lunar	12.658
K_2	declination luni-solar	11.967
Diurnal Tides		
K_1	declination luni-solar	23.935
O_1	principal lunar	25.819
P_1	principal solar	24.066
Q_1	elliptical lunar	26.868
Long-Period Tides		
Mf	fortnightly lunar	327.86
Mm	monthly lunar	661.31
Ssa	semiannual solar	4383.04

If the Earth were uniformly covered with a fluid envelope, these forces would give rise to an equilibrium tide having a maximum elevation of 0.35 m for the lunar (M2) tide and 0.16 m for the solar (K1) tide. In the ocean, this behavior is considerably modified. The phase speed of the tidal wave is approximated by $(gH)^{0.5}$ (where H is the water depth), ocean basins have natural periods of oscillation, tides are affected by the Coriolis force, and land masses can reflect the tides. A given tidal constituent may have amphidromic points where no vertical displacements occur.

Tides can behave as either standing or progressive waves. Tides are generally amplified in shallow water, and here tides are classified as either diurnal, semidiurnal, or a combination of diurnal and semidiurnal called "mixed." Tidal currents associated with rising sea level are called *flood*, while those associated with falling sea level are called *ebb*. Tides cause considerable mixing in shallow water, can generate internal tidal waves when waters are stratified, and, with the rate of river discharge, serve to control the character of estuaries and bays. Bores occur on some rivers, and, at a few locations, tides are used for power generation.

4.4 Gravity

Gravity is the strongest force acting on the sea. It is a body force exerted on the water column, and is the sum of the Earth's gravitational force and the centrifugal force. It is taken in the vertical direction downward:

$$\mathbf{F}_g = -g\mathbf{k} = \frac{d\Phi}{dz}, \tag{19}$$

where $\mathbf{k}$ is the vertical unit vector (directed positive upward from a level surface), g is the gravitational acceleration given by Eq. (4), and Φ is called the *geopotential*, which has units of $m^2\ s^{-2}$. The geopotential defines a level surface to which a motionless ocean would conform (see discussion in Sec. 3.4.1). Gravity provides the restoring force for sea and swell and the energy for thermohaline circulation. It also acts to compress water at depth.

4.5 Viscous Forces

The molecular viscosity of seawater is the mechanism by which mechanical energy is removed. The viscous force is always a resistance force and is expressed by

$$\mathbf{F}_\nu = \nu\nabla^2\mathbf{V}; \tag{20}$$

here ν is the kinetic molecular viscosity, which is a function of temperature, salinity, and pressure and has a value of about $10^{-6}\ \mathrm{m}^2\ \mathrm{s}^{-1}$ for seawater.

Although damping due to molecular viscosity is small, it is important for small-scale processes, e.g., acoustic and capillary waves. It is neglected for larger-scale processes for which the damping due to eddy mixing and diffusivity is represented by means of eddy viscosity.

4.6 Geostrophic Balance

For the large-scale motions in the ocean, the Coriolis force and the pressure-gradient force are the two dominating forces, and the momentum equation becomes geostrophically balanced. The water moves along isobars.

The pressure-gradient force is estimated through the use of the geopotential. From Eqs. (19) and (3),

$$d\Phi = gdz = -(1/\rho)\,dp = -\alpha dp, \tag{21}$$

and here α is specific volume. Integrated over a given pressure interval, this yields *dynamic height*. In practice, an anomaly δD is used; results are divided by ten, and the resulting unit is called *dynamic meters*. Since currents at depth in the ocean are typically slow, an assumption of a *level of no motion* is made, i.e., the lower-pressure surface is assumed to be parallel with a geopotential surface, and a geostrophic current velocity v_g, normal to two CTD stations, a and b, can be calculated:

$$v_g = (\delta D_a - \delta D_b)/fL, \tag{22}$$

where L is the distance between the stations.

4.7 β Effect

The Coriolis parameter ($f = 2\Omega\ \sin\phi$) depends on latitude. If the latitudinal range of motion exceeds 20°, the Coriolis parameter cannot be treated as a constant. To keep this dependence as simple as possible, a linear function is obtained by expanding f into a Taylor series about $y_0 = a\phi_0$ ($a \sim$ Earth's radius),

$$\begin{aligned} f &= f_0 + \beta(y - y_0) + \cdots \\ &= f_0\left[1 + \frac{\beta}{f_0}(y - y_0) + \cdots\right], \end{aligned} \tag{23}$$

where

$$\beta = \left.\frac{df}{dy}\right|_{\phi_0} = \frac{2\Omega\cos\phi_0}{a}. \tag{24}$$

This linear relationship between f and y is called the β effect. In the equatorial region, $f_0 \simeq 0$, the second term on the left-hand side of Eq. (23) becomes dominant and the β effect is evident. In the polar regions, $\beta \simeq 0$, the β effect is negligible. If the zonal variation is much smaller than the Earth's radius, $|y - y_0 \ll a$, the latitudinal variation of the Coriolis parameter is negligible:

$$|(\beta/f_0)(y - y_0)| \ll 1.$$

Therefore, it is reasonable to replace f by the constant f_0 except where it is differentiated by y and to change df/dy to the constant β. This is called the β-plane approximation.

4.8 Conservation Principles

With neither a heat source nor sink, the second law of thermodynamics requires entropy to be conserved, and therefore the potential temperature θ should also be conserved:

$$\frac{d\theta}{dt} = 0. \tag{25}$$

Similarly, with no source or sink of salt, salt should be conserved, i.e.,

$$\frac{dS}{dt} = 0. \tag{26}$$

Further, if water is incompressible, mass conservation leads to

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0, \tag{27}$$

which indicates that horizontal divergence will induce vertical convergence, and vice versa.

If there are neither heat nor salt sources or sinks, nor friction nor diffusion, a mathematical manipulation of momentum, heat, and continuity equations leads to conservation of potential vorticity (Π),

$$\frac{d\Pi}{dt} = 0. \tag{28}$$

4.9 Topographic Forcing

The ocean has sides (e.g., coasts) as well as a bottom. These boundaries are generally assumed to be rigid,

$$\mathbf{V} \cdot \mathbf{n} = 0, \tag{29}$$

indicating that no water particle can penetrate the boundaries. Here **n** is the unit vector in the outward normal direction to the rigid boundary.

In the open ocean, bathymetry is expressed by a two-dimensional function $z = -H(x,y)$. Therefore, the rigid-bottom boundary condition becomes

$$w = -u\frac{\partial H}{\partial x} - v\frac{\partial H}{\partial y}, \quad \text{at } z = -H(x,y). \tag{30}$$

Vertical motion is then generated by horizontal currents flowing over the bottom. This forced vertical motion is important in studies of deep-water processes.

For coastal regions, the water is generally considered as barotropic. Its potential vorticity is $\Pi = (f + \zeta)/H$, where $\zeta = \partial v/\partial x - \partial u/\partial x$ is the vertical component of the relative vorticity. Since the magnitude of ζ is much smaller than f, $|\zeta| \ll |f|$, potential vorticity conservation becomes

$$\frac{d\Pi}{dt} = u\frac{\partial}{\partial x}\left(\frac{f}{H}\right) + v\frac{\partial}{\partial y}\left(\frac{f}{H}\right) = 0, \tag{31}$$

which means that the current vectors are parallel to the isolines of the f/H field. This forcing is very important for topographic waves in coastal regions.

5. OCEAN CIRCULATION THEORY

Motions are generated both by surface wind stress (the wind-driven circulation) and by buoyancy flux (the thermohaline circulation). While it is difficult to say that one type of forcing is more important than the other, aspects of wind forcing have been easier to treat analytically.

5.1 Wind-Driven Circulation

The atmospheric forcing on the ocean takes two forms. One, called the *inverted barometer effect*, is the force exerted on the irregular ocean surface that is associated with atmospheric-pressure differences. The second is the wind stress. Except for severe weather systems (e.g., typhoons) in which the horizontal pressure gradient is large, wind stress is generally considered the major mechanical forcing of the ocean.

Winds blowing over the ocean surface drive different kinds of motion (gravity and capillary waves, turbulence, circulation, etc.) through frictional effects near the ocean surface. The surface wind stress, defined as the turbulent momentum flux at the ocean surface, is on the order of 1 dyn/cm^2 and is related to the surface wind speed $\mathbf{V}_a$ by the drag law

$$\boldsymbol{\tau} = C_D \rho_a |\mathbf{V}_a| \mathbf{V}_a, \tag{32}$$

where C_D ($\sim 10^{-3}$) is the drag coefficient, whose magnitude is affected by the roughness, the buoyancy flux, and the waves and turbulence near the ocean surface.

5.1.1 Ekman Transport When the wind blows over the ocean surface, the friction between the air and the ocean will slow the wind and cause ripples, waves, and currents in the ocean. The stress of the wind on the water involves a transfer of turbulent fluxes of momentum and kinetic energy from the atmosphere to the ocean. In the ocean, the turbulent kinetic energy transferred from the atmosphere decreases from the surface downward and will vanish at some depths. A depth of 100 m is usually considered the practical lower limit for wind-driven currents. The layer that contains the turbulent flux is called the surface layer, or the Ekman layer.

Imagine that the ocean surface layer is divided into thin horizontal layers. The frictional force (per unit area) exerted between the two neighboring layers is represented by a horizontal stress (X,Y), which is the retarding force for the upper layer and the driving force for the lower layer. At the ocean surface, (X,Y) is the wind stress. If these stresses are balanced by the Coriolis force, the Ekman velocity (u_E,v_E) satisfies

$$\frac{\partial u_E}{\partial t} - f v_E = \frac{1}{\rho}\frac{\partial X}{\partial z}, \tag{33a}$$

$$\frac{\partial v_E}{\partial t} + f u_E = \frac{1}{\rho}\frac{\partial Y}{\partial z} \tag{33b}$$

[this solution was derived by Ekman (1905) to explain why Arctic icebergs drifted 40° to the right of the wind]. Notice that currents driven by the pressure-gradient force are not included in the Ekman flow. For steady state ($\partial/\partial t = 0$), the vertical variation of the Ekman velocity, called the Ekman spiral, shows that the surface current is directed 45° to the right of the wind stress in the Northern Hemisphere and 45° to the left in the Southern Hemisphere. As depth increases, the direction of water transport deviates to the right in the Northern Hemisphere relative to the overlying water in the same fashion as the surface water is transported by the wind. Friction causes each succeeding layer to decrease in speed.

The total water transport in the Ekman layer is calculated by the vertical integration of the Ekman velocity:

$$(U_E,V_E) = \int (u_E,v_E)\,dz,$$

which is also called the Ekman transport. Vertical integration of Eq. (33) leads to

$$\rho\left(\frac{\partial U_E}{\partial t} - fV_E\right) = \tau_x, \quad \rho\left(\frac{\partial V_E}{\partial t} + fU_E\right) = \tau_y, \tag{34}$$

where (τ_x,τ_y) is the surface wind stress. In steady conditions, the Ekman transport is directed 90° to the right (or left) of the surface wind stress in the Northern (or Southern) Hemisphere. The discontinuity of the Ekman transport at the equator results in equatorial upwelling (downwelling) for westward (eastward) winds.

The wind stress varies from place to place, and hence so does the Ekman transport. This leads to convergence or divergence of water mass in the Ekman layer, i.e., causes the water mass to flow horizontally into some regions and out of others. This will result in vertical motion. For example, if the Ekman transport is horizontally convergent in a particular region, vertical downward motion away from the Ekman layer is induced in order to conserve mass. If we assume no significant distortion of the ocean free surface, a vertical integration of the continuity equation across the Ekman layer yields the vertical velocity (called Ekman pumping velocity) at the base of the Ekman layer:

$$w_E = \nabla \cdot \mathbf{U}_E = \mathbf{k} \cdot \mathrm{curl}(\boldsymbol{\tau}/\rho f). \tag{35}$$

This vertical velocity, driven by the surface wind stress, exchanges water between the surface layer (under the direct effect by the atmosphere) and the deeper water. In subtropical gyres (equatorward of 45 deg) w_E is usually negative: The surface water is pumped downward into the deeper layer. In subpolar gyres, w_E is generally positive: The water mass is pumped upward from the interior into the surface layer. Therefore, Ekman pumping is a crucial link between the surface wind forcing and deep-ocean dynamics.

5.1.2 Sverdrup Transport Ekman flow excludes the currents forced by the pressure gradient. The Sverdrup transport includes both the Ekman transport and the geostrophic transport that results from the pressure gradients set up by the convergent or divergent Ekman-layer flow. Let the mass transport function with respect to the ocean bottom (H) and the integrated pressure (P) be defined as

$$\frac{\partial M}{\partial x} = \int_{-H}^{0} v\,dz, \quad \frac{\partial M}{\partial y} = -\int_{-H}^{0} u\,dz,$$
$$P = \int_{-H}^{0} \frac{p}{\rho}\,dz. \tag{36}$$

Vertical integration (from $-H$ to 0) of the momentum equation (14) after neglecting the advection term and the tidal forcing leads to

$$-\frac{\partial^2 M}{\partial t \partial y} - f\frac{\partial M}{\partial x} = -\frac{\partial P}{\partial x} + \tau_x, \tag{37a}$$

$$\frac{\partial^2 M}{\partial t \partial x} + f\frac{\partial M}{\partial y} = -\frac{\partial P}{\partial y} + \tau_y. \tag{37b}$$

Sverdrup (1947) first obtained the steady-state solution of Eq. (37) for the mass transport M under the β-plane approximation:

$$\frac{\partial M}{\partial x} = \frac{1}{\beta}\left(\frac{\partial \tau_y}{\partial x} - \frac{\partial \tau_x}{\partial y}\right), \tag{38}$$

which shows the congruity between the north–south ocean flow and the wind-curl patterns. Note that the Sverdrup transport was obtained only when the zonal dependency of the Coriolis parameter is considered. If the β effect is negligible (f = const), there is no Sverdrup transport. In the low to middle latitudes of the Northern Hemisphere (15°–50° N), the surface wind-stress curl is usually anticyclonic, and drives a southward Sverdrup transport. Sverdrup further showed that if Eq. (38) is integrated for M from an east coast, where the mass transport diminishes, a current counter to the prevailing winds would occur in the Eastern Pacific Ocean under the Inter-Tropical Convergence Zone (ITCZ). Such a current is observed, the North Equatorial Counter Current. (Oceanographers have subsequently defined a unit of transport as a Sverdrup, i.e., 10^6 m^3/s = 1 Sv.)

5.1.3 Westward Intensification If you look at a pilot chart of ocean currents, you immediately notice that the strongest currents (e.g., the Gulf Stream and Kuroshio) appear at the western boundaries of the ocean and that currents elsewhere are noticeably weaker. This east–west asymmetry is called *westward intensification.*

As noted in Sec. 4.7, the zonal change of the Coriolis parameter (β effect) must be taken into account. In low to middle latitudes, the anticyclonic surface wind-stress curl drives equatorward Sverdrup transport. This equatorward flow must return poleward somewhere in order to satisfy continuity. Will this occur at the eastern or western boundary? Because the Coriolis effect causes equatorward flow to turn toward the west, the return flow must take place at the western boundary. Stommel (1948) and Munk (1950) offered theoretical explanations for the westward intensification. Both models are based on potential vorticity conservation,

$$\begin{aligned}\frac{d\Pi}{dt} &= \frac{\partial \Pi}{\partial t} + (\mathbf{V}\cdot\nabla)\Pi \\ &= \frac{1}{\rho h^2}\,\mathbf{k}\cdot(\text{curl } \boldsymbol{\tau}|_0 - \text{curl } \boldsymbol{\tau}|_{-h}) + A\nabla^2\Pi,\end{aligned} \tag{39}$$

which indicates that the time-rate change of the potential vorticity is caused by the surface wind stress ($\boldsymbol{\tau}|_0$), bottom friction ($\boldsymbol{\tau}|_{-h}$), and the horizontal diffusion ($\nabla^2\Pi$). Here, h is the ocean depth and is assumed to be a constant. The magnitude of the relative vorticity (ζ) for large-scale motions in the ocean is much smaller than the magnitude of the planetary vorticity f, so that the advection of the potential vorticity can be linearized:

$$\mathbf{V}\cdot\Pi = (\beta/h)\,v. \tag{40}$$

For the steady state, potential vorticity advection is balanced by friction and diffusion. The difference between the two models is that there is no horizontal diffusion in Stommel's model, and no bottom friction in Munk's theory.

5.1.3.1 Stommel's Model. By neglecting horizontal diffusion and parametrizing the bottom friction by the Rayleigh friction, $\tau_{-h}/\rho = R(u,v)$, the barotropic potential vorticity equation (39) becomes

$$\begin{gathered}\nabla^2\psi + \frac{h}{R}\beta\frac{\partial\psi}{\partial x} = \frac{\mathbf{k}}{\rho R}\,\text{curl } \boldsymbol{\tau}_0, \\ u = -\frac{\partial\psi}{\partial y}, \quad v = \frac{\partial\psi}{\partial x}.\end{gathered} \tag{41}$$

Stommel (1948) applied this model to a rectangular ocean. The shores of the ocean were at $x = 0, L_x$, $y = 0, L_y$. The surface wind stress was chosen as a sinusoidal function $\mathbf{k}\cdot\boldsymbol{\tau}_0/\rho = F\sin(\pi y/L_y)$. The model was integrated for three cases: $f = 0$, $f =$ const, and $f = f_0 + \beta y$. The nonrotating, wind-driven system features a broad circulation exhibiting no tendency toward western boundary intensification. The uniformly rotating, wind-driven system has a similar pattern (i.e., no westward intensification). The final case, which used a β plane, has intense crowding of streamlines toward the western boundary of the ocean.

5.1.3.2 Munk's Model. By neglecting bottom friction, Munk (1950) simplified the barotropic potential vorticity equation (39):

$$\left[-A\nabla^4 + \beta\frac{\partial}{\partial x}\right] = \mathbf{k}\cdot\text{curl } \boldsymbol{\tau}_0. \tag{42}$$

His solutions show westward intensification similar to Stommel's model.

5.2 Thermohaline Circulation

Thermohaline circulation problems have proven more difficult, both theoretically and experimentally. On the experimental side, this has been due to the relatively weak flows and the depth at which they occur, as well as the fact that the polar regions, where much of the convective forcing occurs, are not easily accessible. The two major deep-water masses are the North Atlantic Deep Water and Antarctic Bottom Water. North Atlantic Deep Water is formed at high latitudes in the North Atlantic from overflow waters from the Nordic seas (a production rate of about 6 Sv) and in the Labrador Sea (about 8 Sv). Most Antarctic Bottom Water is formed in the Weddell Sea at a rate of 2–4 Sv.

Stommel (1958) proposed a circulation model for the deep circulation given these two sources of deep water. Upwelling in the interior is fed by slow eastward flow from equatorward-flowing deep western-boundary currents, which carry water from the deep-water formation regions. Measurements subsequently confirmed the deep western-boundary currents.

5.2.1 Thermocline Theory Broecker (1991), Munk (1966), Schmitz and McCartney (1993), and Luyten *et al.* (1983) argue that the difficulty in modeling the thermocline is due to the tendency of the ocean to preserve both density and potential vorticity over great distances from source areas of both fields. The problem of the vertical structure of thermocline must be coupled to the horizontal structure of the ocean circulation.

A thermocline equation can be derived from the geostrophic and hydrostatic balance of the large-scale ocean circulation, and from the advective–diffusive balance of the density field:

$$J\left(\frac{\partial Q}{\partial z}, \frac{\partial^2 Q}{\partial z^2}\right) + \cot\phi \frac{\partial Q}{\partial \lambda}\frac{\partial^3 Q}{\partial z^3} = \mu \frac{\partial^4 Q}{\partial z^4} \cos\phi \sin\phi, \tag{43}$$

where ϕ and λ are latitude and longitude, respectively, $J(F,G)$ is the Jacobian for any two functions F and G; $\mu \equiv \delta_D/\delta_A$ is the ratio of the diffusive to advective length scales; Q is defined as

$$Q(\phi,\lambda,z) = -\int_z^0 p dz' + \sin^2\phi \int_0^\lambda w_E d\lambda'. \tag{44}$$

The thermocline equation (43) is very difficult to solve, although solutions have been obtained for special cases.

5.2.2 Thermohaline Forcing The combined effect of fluxes of heat and salt on the ocean causes buoyancy contrasts, and induces motion called the thermohaline circulation. The most important thermodynamic forcing is the buoyancy flux at the air–ocean interface:

$$B = \alpha(LE + Q)/c_{pw} + \beta(E - Pr)\, S_0 + \beta(F - M)(S_0 - S_f), \tag{45}$$

where Q is the sum of the radiative and turbulent heat fluxes at the air–ocean interface (upward positive), E the evaporation rate, Pr the precipitation rate, F the ice freezing rate, L the latent heat of vaporization, M the ice melting rate, C_{pw} the specific heat of seawater, S_0 the upper-ocean salinity, and S_f the ice salinity ($S_f \sim 0.7$ g/kg). Since $\alpha L/c_{pw} \simeq 0.12$ and $S_0 \simeq 0.03$, the contribution of evaporation rate to the buoyancy flux from the first term on the right-hand side of Eq. (45) is 4 times greater than that from the second term. The values of B are on the order of 10 g/cm^2 yr, which corresponds to a turbulent kinetic energy generation rate of the order of 3×10^{-4} cm^2 s^{-3}. Upward surface buoyancy flux B (heat loss from the ocean surface, or excess evaporation over precipitation, or salt injection due to ice formation) increases the surface density and makes the water parcel descend. The circulation that is driven by the buoyancy flux is called the thermohaline circulation.

5.2.3 Ventilation Ventilation is any process that brings subsurface water to the surface so that heat, momentum, and gases can be exchanged with the atmosphere. Ventilation of deeper water occurs primarily through two processes. The first is the Ekman pumping of water onto density surfaces. The region on an isopycnal surface affected by this process is called the ventilation region. The second process is deep convection. Although a number of estimates of the mean ventilation time for the deep ocean have been proposed, that of about 1400 years, based on radiocarbon, is the most widely quoted. The ventilation time T of the deep ocean is usually measured by the carbon isotope ratios:

$$T = \frac{V_D}{R} = \left[\frac{(^{14}C/C)_S}{(^{14}C/C)_D} - 1\right] T_{^{14}C}, \tag{46}$$

where V_D and R are the volume and the ventilation rate of the deep sea, respectively. Here $(^{14}C/C)_S$ and $(^{14}C/C)_D$ are the carbon isotope ratios for the mean surface and mean deep ocean. The term T_{14_C} is the mean half-life of ^{14}C.

Ventilation is fundamental to understanding the large-scale motion in the deep sea. To understand deep-sea circulation, one needs to answer the dual questions of how a fluid below the surface layer (which is directly driven by the wind) is set into motion, and what determines the depth to which that motion penetrates. Since the ocean has no significant frictional stresses between layers, only the surface water that is immediately acted upon by the wind stress can brought into motion. The existence of subsurface motion therefore requires either ventilation or convective mixing. Ventilation requires the water parcel trajectories to thread back and intersect near-surface regions directly affected by Ekman pumping. Convective mixing leads to momentum transfer between different layers.

This scheme creates at least two regions in the thermocline of the North Atlantic. One region is connected directly to northern latitudes where water is ventilated to the atmosphere and forced downward by Ekman pumping, subsequently being subducted and conserving potential vorticity. Another region is not ventilated to the atmosphere. The unventilated region sweeps across the entire low-latitude North Atlantic from the eastern boundary at the depth of 27.4 kg/m^3 isopycnal. Observed distributions of dissolved oxygen and tritium on this surface are indicative of poor ventilation in this region.

Subduction is a companion concept to ventilation. In the subtropical gyre, surface waters are pumped down into the main thermocline (ventilation) by the action of the prevailing winds. Under mechanical and buoyancy forcing, the mixed layer cools and burrows into the thermocline, entraining thermocline water. Stratification existing in the thermocline determines how the mixed layer responds to the forcing. Alternatively, the mixed layer may warm and retreat, leaving behind newly stratified fluid, which then subducts into the deeper layers.

5.2.4 Convection Generally the ocean is stable, i.e., the lighter water lies over denser water. In polar regions and the Mediterranean Sea, the stratification sometimes becomes quite weak. Slight negative buoyancy in the surface water caused by atmospheric cooling or salt injection will make the water column statically unstable and trigger deep convection. Water columns with vertically homogeneous hydrographic structure (called "chimneys") have been observed in the Arctic, Antarctic, and Mediterranean; they have been observed near ocean boundaries as well as in the open ocean.

Near-boundary convection requires three ingredients:

1. a reservoir,
2. a source of dense water within the reservoir,
3. a reason for the dense water to leave the reservoir and to descend the slope to the deep layer (Killworth, 1983).

Open-ocean convection usually requires a cyclonic gyre and "preconditioning." Upward "doming" of isopycnals in the center of the cyclonic gyre reduces static stability of water column within the gyre. The preconditioning creates a region of very weak static stability within the dome of the gyre where a sufficient surface forcing can drive deep convection. Many processes can be thought of as preconditioning, e.g., topographic forcing, baroclinic instability, etc.

An example of open-ocean convection is a column of water (14-km radius) that was observed within the central region of the Weddell Gyre west of Maud Rise (Gordon, 1978). The normal Antarctic stratification was absent, and the column appeared as a cold, low-salinity, high-oxygen, cyclonic flow (surface velocity about 50 cm/s), which extended to at least 4000 m.

5.2.5 The Conveyor Belt Recent observations and numerical modeling suggest the existence of a "conveyor belt" in the ocean that may serve as a mechanism for long-term climate variability. Deep waters in the northern North Atlantic are convected, sinking to depth and flowing southward along the western boundary of the North and South Atlantic Oceans as the North Atlantic Deep Water (NADW) with high salinity content, hence forming the conveyor belt's lower limb. After passing 30° S, this lower limb of the conveyor joins the Circumpolar Current along the perimeter of the Antarctic continent, along with older deep waters recirculated into the Antarctic from the deep Pacific and Indian Oceans. The lower limb water returns to the surface in the northern Indian and Pacific Oceans, feeding the upper limb of the conveyor. The imported water that feeds the upper limb of the conveyor belt in the Atlantic can have three components: Antarctic surface waters passing through the Drake Passage, Indian surface waters passing around the tip of Africa via the Agulhas Current, and intermediate waters formed at the northern perimeter of the Atlantic segment of the Antarctic. Different conveyors (or climatic regimes) can be envisioned in which only one or two of these three components are active.

6. OCEAN WAVES

An ocean basin filled with slightly compressible and electrically conducting seawater on a weakly magnetized rotating Earth will support a variety of waves. An ocean wave is generally depicted by a periodic function. A one-dimensional wave propagating in the x direction is usually described by a sinusoidal function

$$\zeta = A \cos(kx - \omega t), \tag{47}$$

where ζ is any wave variable, A the wave amplitude, k the wave number, and ω the wave frequency. The quantity $L = 2\pi/k$ is called the wavelength; $T = 2\pi/\omega$ is called the period. The nondimensional quantity $kx - \theta t$ is called the phase angle. The wave propagates with a phase speed

$$C = \omega/k. \tag{48}$$

When the phase speed (C) varies with the wave number (k), the waves with different wavelengths travel with different phase speeds, which is called the wave dispersion. The wave packet, a group of waves with different wavelengths, travels with the group velocity

$$C_g = \frac{\partial \omega}{\partial k}. \tag{49}$$

Hydroelectromagnetic waves (Alfvén waves) are possible in the ocean, but the Earth's magnetic field is so weak that the electromagnetic restoring forces are insignificant compared to other restoring forces; so Alfvén waves are not important in the ocean. Acoustic waves exist as a result of compressibility. Gravity waves arise through the restoring force of gravity. Short, high-frequency capillary waves occur at the interface between air and water through the restoring force of surface tension. Inertial waves appear as a result of the Coriolis force, acting at right angles to the velocity vector. Finally, very slow, planetary waves develop from the variation of quasigeostrophic potential vorticity.

These five basic types of oceanic waves (acoustic, capillary, gravity, inertial, and planetary waves) generally occur together, with five basic restoring forces all acting simultaneously to produce more complicated and mixed waves, such as inertial–gravity waves, mixed Rossby–gravity waves, etc. The effect of the lateral boundary on the inertial–gravity waves leads to a special type of wave named after Lord Kelvin, who first found a solution describing the wave.

6.1 Acoustic Waves

The propagation of sound in the ocean was discussed in Sec. 1.3. The propagation of acoustic waves in the ocean is governed by

$$\nabla^2 p - \frac{1}{U_a^2}\frac{\partial^2 p}{\partial t^2} + \frac{\mu}{\rho U_a^2}\nabla^2\frac{\partial p}{\partial t} = 0, \tag{50}$$

where μ is molecular viscosity of seawater. The ocean surface is a very good reflector/scatterer of sound energy. The ocean bottom, composed of hard rock to soft muds, has large variability in reflection, scattering, and absorption. We usually solve the wave-propagation equation subject to various surface and bottom boundary conditions and obtain information about the propagation and attenuation.

6.2 Gravity and Capillary Waves

The wind generates surface waves. At very low wind speeds, small capillary waves develop on the surface. Surface waves have frequencies ω much higher than the inertial frequency (Coriolis parameter) f and buoyancy frequency N. Therefore, the Coriolis, buoyancy, and viscous terms are usually neglected in the momentum equations. Under these conditions, the fluid motion will be *irrotational.* A scalar velocity potential ϕ can be defined to represent the vector velocity:

$$\mathbf{V} = \nabla\phi. \tag{51}$$

The incompressibility condition, $\nabla \cdot \mathbf{V} = 0$, implies that the velocity potential (ϕ) satisfies Laplace's equation,

$$\nabla^2\phi = 0. \tag{52}$$

To solve this equation, we need surface and bottom boundary conditions. Consider a uniform, frictionless ocean of depth H, whose surface is subject to a small wavelike perturbation

$$\xi(x,y,t) = \xi_0 \cos(kx + ly - \omega t). \tag{53}$$

The pressure across the air–ocean interface should be continuous, i.e.,

$$p = p_a - \tau_s\left(\frac{\partial^2 \xi}{\partial x^2} + \frac{\partial^2 \xi}{\partial y^2}\right). \tag{54}$$

Here p and p_a are oceanic and atmospheric pressures at the ocean surface, and τ_s is surface tension. If $\tau_s = 0$, we have the pure gravity waves; if $\tau_s \neq 0$, we have the gravity-capillary waves. Combining these two conditions, the surface boundary ($z = 0$) condition for ϕ can be derived:

$$\frac{\partial^2\phi}{\partial t^2} + \frac{1}{\rho}\left[-\tau_s\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) + \rho g\right]\frac{\partial\phi}{\partial z} = 0, \quad z = 0. \tag{55}$$

The ocean bottom is generally considered a rigid boundary. Since this is a linear system, ϕ should have the same harmonics as ξ, i.e.,

$$\phi = \phi_0(z)\sin(kx + ly - \omega t). \tag{56}$$

Substitution of Eq. (56) into Eq. (55) leads to an ordinary differential equation for ϕ_0 with z. By using the surface and bottom boundary conditions, we will have the dispersion relation

$$\omega^2 = (gk_h + k_h^3\tau_s/\rho)\tanh k_h H, \tag{57}$$

where $k_h = \sqrt{k^2 + l^2}$ is the horizontal wave number. The dispersion relation shows that the gravity-capillary waves are generally dispersive ($c_g = \partial\omega/\partial k \neq 0$).

For the pure gravity waves, $\tau_s = 0$, and therefore, the dispersion relation becomes

$$\omega^2 = gk_h \tanh k_h H. \tag{58}$$

If the horizontal wavelength $L_h = 2\pi/k_h$ is much shorter than the ocean depth (deep water), $k_h H = 2\pi H/L_h \gg 1$ and $\tanh k_h H \simeq 1$, and so the deep-water gravity-wave phase speed becomes

$$c = (g/k_h)^{1/2}, \quad k_h H \gg 1, \tag{59}$$

while the group velocity is

$$c_g = c/2. \tag{60}$$

The deep-water gravity waves are dispersive with the phase speeds depending inversely on the square root of the wave number; therefore, longer waves travel faster than short ones. In addition, the group velocity is $\frac{1}{2}$ of the phase speed, so that within a traveling wave packet, wave crests first appear toward the rear of the group. An initially localized packet consisting of many wave components spreads out in space and time as it propagates.

If the horizontal wavelength is much longer than the ocean depth (shallow water), $k_h H = 2\pi H/L_h \ll 1$ and $\tanh k_h H \simeq k_h H$, and so the shallow-water gravity-wave phase speed becomes

$$c = c_g = (gH)^{1/2}, \quad k_h H \ll 1. \tag{61}$$

Thus, the shallow-water gravity waves are nondispersive. Earthquakes generate a special kind of gravity wave called a tsunami, which behaves as a shallow-water wave even in deep water.

6.3 Rossby Waves

For large-scale motion, departure from geostrophic balance (sometimes called geostrophy) is small. Such flow is called quasigeostrophic flow. The quasigeostrophic potential vorticity (QGPV) Π_{QG} is

$$\Pi_{QG} = f_0 + \beta y + \zeta_g + \Lambda\frac{\partial}{\partial z}\left(\frac{f_0^2}{N^2}\frac{\partial\Phi}{\partial z}\right), \tag{62}$$

where $\zeta_g = (\partial^2/\partial x^2 + \partial^2/\partial y^2)\Phi$ is the geostrophic vorticity, Φ is the geopotential computed by Eq. (19), $N^2 = E$ is the square of the buoyancy frequency, $\Lambda = 0$ refers to the barotropic mode, and $\Lambda = 1$ to the baroclinic mode.

Rossby waves are associated with the β effect. The conservation of quasigeostrophic potential vorticity,

$$\left(\frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y}\right)\Pi_{QG} = 0, \tag{63}$$

leads to a special form of oscillation on the planetary scale. An individual water parcel experiences no change in its QGPV. If a water parcel moves poleward, then as f increases, ζ_g has to decrease for the barotropic mode ($\Lambda = 0$) in order to keep Π_{QG} constant. If, furthermore, the geostrophic vorticity is mostly curvature, this means that a poleward-moving water parcel loses cyclonic curvature and ultimately must curve anticyclonically and move equatorward. The converse is true for a water parcel moving equatorward. This mechanism leads to the generation of Rossby waves (Rossby, 1936).

If the disturbances are sufficiently small, the advection terms in the QGPV conservation equation (63) can be neglected,

$$\frac{\partial}{\partial t}\left[f_0 + \beta y + \nabla_H^2 \Phi' + \Lambda \frac{\partial}{\partial z}\left(\frac{f_0^2}{N^2}\frac{\partial \Phi'}{\partial z}\right)\right] = 0, \tag{64}$$

where Φ' is the geopotential perturbation and $\nabla_H^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the horizontal Laplacian. If we assume no vertical velocity at either the surface or bottom, the sinusoidal wave solution has the form (when N = const)

$$\Phi' = \mathrm{Re}\left[\cos\frac{n\pi z}{H}\, e^{i(kx+ly-\omega t)}\right], \tag{65}$$

where H is the ocean depth. Substitution of Eq. (65) into Eq. (64) leads to the Rossby-wave dispersion relation

$$\omega = -\frac{\beta k}{k^2 + l^2 + \Lambda m_n}, \qquad m_n = \frac{n\pi f_0}{NH}, \quad n = 1, 2, \ldots. \tag{66}$$

For baroclinic Rossby waves, $\Lambda = 1$, formula (66) shows the dispersion relation of the nth baroclinic mode. For barotropic Rossby waves, $\Lambda = 0$, the dispersion relation becomes

$$\omega = -\beta k/(k^2 + l^2). \tag{67}$$

Rossby waves are highly dispersive and propagate westward. Rossby-wave phase speed and group velocity can be easily computed from the dispersion relation:

$$\mathbf{C} = \left(\frac{\omega}{k}, \frac{\omega}{l}\right), \quad \mathbf{C}_g = \left(\frac{\partial \omega}{\partial k}, \frac{\partial \omega}{\partial l}\right). \tag{68}$$

If the barotropic Rossby waves propagate only in the x direction ($l = 0$), the phase speed is

$$C_x = \omega/k - \beta/k^2 \tag{69}$$

and group velocity is

$$C_{gx} = \frac{\partial \omega}{\partial k} = \frac{\beta}{k^2}. \tag{70}$$

The highly dispersive nature of these waves is indicated by the fact that the phase speed and group velocity have the same magnitude but opposite directions.

6.4 Solitary Rossby Waves

Rossby waves are usually nonlinear and dispersive. The degree of nonlinearity depends on the amplitude and length of the waves, while the phase dispersion is determined by their length relative to the vertical length scales related to stratification and water depth. In general, these two dynamic effects tend to balance in such a way that nonperiodic, solitary waves emerge from initial disturbances.

Consider barotropic Rossby waves ($\Lambda = 0$) being excited from a zonal basic flow ($\bar{U}$) and propagating in a zonal channel bounded by southern and northern boundaries $y = y_S$, $y = y_N$. Introducing a scaled coordinate $\xi = \epsilon^{1/2}(x - Ct)$ moving with the wave at speed C, we seek a steady solution

$$\Phi = -\frac{1}{f_0}\int_{y_S}^{y} \bar{U}\,dy + \epsilon\Phi_1(\xi,y) + \epsilon^2\Phi_2(\xi,y) + \cdots, \tag{71a}$$

$$C = C_0 + \epsilon C_1 + \epsilon^2 C_2 + \cdots, \tag{71b}$$

where ϵ is a measure of the wave amplitude. Notice that the disturbances are no longer sinusoidal functions. Substitution of Eq. (71) into Eq. (63) leads to nonlinear equations for the components Φ_n, $n = 1, 2, \ldots$. If the first component is separated into

$$\Phi_1 = A(\xi)\psi(y), \tag{72}$$

the function $A(\xi)$ satisfies the steady Korteweg–de Vries (KdV) equation

$$\frac{dA}{d\xi} + rA\frac{dA}{d\xi} + s\frac{d^3A}{d\xi^3} = 0, \tag{73}$$

and solvability to $O(\epsilon^2)$ is insured if r and s satisfy certain conditions.

The soliton solutions are a linear sum of a barotropic soliton and a baroclinic soliton (first baroclinic mode). If a rigid lid is imposed, the barotropic solitary Rossby wave has an eastward phase speed ($c > 0$), and the baroclinic solitary Rossby wave has a phase speed that is either westward with a faster speed than βa_1^2 ($c < -\beta a_1^2$), or eastward ($c > 0$). Here a_1 is the Rossby radius of deformation for the first baroclinic mode. When the barotropic and baroclinic solitons are combined, only the baroclinic part moves west.

6.5 Kelvin Waves

When planetary waves approach the coastal zone, their behavior changes. The lateral boundary supports a different type of wave called a Kelvin wave. Consider a straight coastline, placing the x axis along the coast and the y axis perpendicular to the coast so that offshore is positive. Assume that the offshore variation of wave height and alongshore velocity exponentially decays. Since the coast is a rigid boundary, v vanishes at $y = 0$; therefore, v should vanish for the other values of y as well. Kelvin-wave solutions can be obtained by letting $v = 0$ in the momentum and continuity equations,

$$\frac{\partial u}{\partial t} = -g\frac{\partial \eta}{\partial x}, \tag{74a}$$

$$\frac{\partial \eta}{\partial t} + H\frac{\partial u}{\partial x} = 0, \tag{74b}$$

$$fu = -g\frac{\partial \eta}{\partial y}, \tag{74c}$$

where η is the surface elevation. Solutions of Eq. (74) are

$$\begin{aligned} &\eta = A_\eta e^{-y/a}\cos(kx - \omega t), \\ &u = \eta(g/H)^{1/2}, \quad v = 0, \end{aligned} \tag{75}$$

where A_η is the amplitude for wave variable η, and a is the barotropic Rossby radius of deformation [see Eq. (76)]. The solution shows two important features of coastal trapped Kelvin waves:

1. The amplitude is greatest at the boundary and decreases exponentially away from it, and
2. propagation is with the coast at the right (left) in the Northern (Southern) Hemisphere.

Kelvin waves also occur along the equator where f changes sign, propagating from west to east. Abrupt changes in topography can also support Kelvin waves; in this case, they are called double Kelvin waves because there is motion on both sides of the depth change.

7. INSTABILITIES, EDDIES, AND FRONTS

The kinetic energy of the ocean is dominated by synoptic or mesoscale eddies, rotating about nearly vertical axes and extending throughout the water column. The scale of these eddies is controlled by the baroclinic radius of deformation. Typical scales are 100 km and 100 d with a rotational velocity of about 10 cm/s. Mesoscale eddies are generated by baroclinic and barotropic instabilities.

7.1 Rossby Radius of Deformation

Two fundamentally important frequencies in ocean dynamics are f and N, representing buoyancy and rotational effects. For small-scale motion, buoyancy dominates, while for large-scale motion, the rotational effect is dominant. There must exist a horizontal length scale at which rotation becomes important. This length scale, a, is called the Rossby radius of deformation, and the barotropic mode is given by

$$a = \sqrt{gH}/|f| \tag{76}$$

while the nth baroclinic mode is

$$a_n = \frac{NH/n\pi}{|f|}. \tag{77}$$

If $f = 10^{-4}\ \mathrm{s}^{-1}$ (about 45° N), and $H = 4$ km, then the barotropic Rossby radius $a \simeq 2000$ km. For shallower water on the continental shelf, $H \simeq 40$ m, then $a \simeq 200$ km. The baroclinic Rossby radius is much shorter. The typical value for the first baroclinic mode is 10–30 km, which is a natural scale for the ocean and is often associated with eddies, fronts, and coastal processes.

7.2 Baroclinic Instability

Baroclinic instability originates from the horizontal inhomogeneity of the density field and allows eddies to grow with energy coming from the mean potential energy field. The horizontal density gradient is related to the vertical shear of the geostrophic flow (U_g,V_g) by the *thermal wind* equation:

$$\frac{\partial U_g}{\partial z} = \frac{g}{f\rho}\frac{\partial \rho}{\partial y}, \quad \frac{\partial V_g}{\partial z} = -\frac{g}{f\rho}\frac{\partial \rho}{\partial x}. \tag{78}$$

The existence of vertical shear leads to baroclinic instability and eddy generation. In this instability process, energy is taken from the large-scale potential energy distribution and put into disturbance-scale kinetic and potential energies. Stommel (1965) suggested that the enormous amount of potential energy present in the stratification of the ocean and associated with the mean circulation is a possible source for baroclinic instability.

7.3 Barotropic Instability

Barotropic instability originates from the lateral shear of the flow and is an important mechanism to generate eddies in the tropics. For simplicity, consider a zonal current U that changes with latitude, i.e., $U = U(y)$. If the lateral shear satisfies the condition (Kuo, 1949)

$$\beta - \frac{d^2U}{dy^2} = 0 \tag{79}$$

somewhere in the flow, then the flow U is barotropically unstable. Eddies generated by the shear in the mean flow gain energy from the mean flow kinetic energy.

Within the ocean, eddy generation is the result of a mixture of baroclinic and barotropic instability.

7.4 Mesoscale Eddies

Mesoscale eddies are energetic disturbances of currents and the associated fields of salinity, temperature, and density (Robinson, 1983). More kinetic energy is contained in these features than in any other form of oceanic motion. These eddies are partly organized and yet highly irregular, with spatial scales of 10–100 km and temporal scales of weeks to months. The energy is distributed unevenly in space, with the greatest energy associated with western-boundary currents. Mesoscale eddies provide an important mechanism for the horizontal stirring of the ocean, and for transfers of energy, momentum, and water properties. Numerical model results (Schmitz and Holland, 1986) suggest that these eddies may also drive the recirculation region associated with subtropical gyres. The mesoscale features can also be found from satellite altimetry data. Geostrophic currents are maintained by horizontal pressure gradients and are expressed as sea-surface slopes relative to the level of constant geopotential (the geoid). All mesoscale current systems exhibit high levels of sea-height variability due to meandering, changes in speed or current reversal, vortex shedding, and migration of eddies. For example, western-boundary currents such as Gulf Stream and Kuroshio are characterized by high rms sea-height variability (Plate 3).

An important category of mesoscale eddy is called a *ring*. Rings associated with the Gulf Stream have been well documented (see Plate 2) (Richardson, 1983). These rings appear to form from cut-off Gulf Stream meanders. Two types of rings have been observed, warm-core rings to the north of the Stream and cold-core rings to the south. Cold-core rings are cyclonic, are carried to the west by the Gulf Stream recirculation, and are reabsorbed by the Gulf Stream to the south of Cape Hatteras. A young cold-core ring has rotational velocities of 1.5 m/s and is remarkably similar to a Rankin vortex: Speed increases with distance from the center (a nearly solid rotating core), reaching a maximum near a radius of 30–60 km, and then drops off toward the outer limits of the ring. Ten cold-core rings have been observed at a given time to the south of the Stream.

Warm-core rings are anticyclonic and are formed at a rate of about five per year. They may be reabsorbed by the Gulf Stream, but a few impact the continental shelf/slope where they are the most energetic source of variability on the Slope Water (Joyce, 1991). Rings are also observed in other strong western-boundary currents, most notably the Kuroshio and Agulhas.

7.5 Quasigeostrophic Turbulence

It is possible to extend the concepts of eddies and planetary waves to two-dimensional quasigeostrophic turbulence. Beginning first with an f plane in the absence of friction or forcing, conservation equations for energy E,

$$\frac{\partial E}{\partial t} = 0, \quad E = \int E(k)dk, \tag{80}$$

and for enstrophy Z,

$$\frac{\partial Z}{\partial t} = 0, \quad Z = \int k^2E(k)dk, \tag{81}$$

require an energy transfer to smaller wave numbers and an enstrophy transfer to larger wave numbers. This means that vorticity and eddy fields change with time, with eddies increasing in size and vorticity contours increasing in length. The rate of cascade controls the spectral slope so that for the energy cascading range,

$$E(k < k_0) \sim k^{-3/5}, \qquad (82)$$

and for the enstrophy cascading range,

$$E(k > k_0) \sim k^{-3}. \qquad (83)$$

There is some suggestion that the ocean behaves in this manner, but in numerical models, spectral slope is related to the friction parameterization.

With β, linear Rossby waves will propagate to the west and disperse energy in space, but can also concentrate energy at a specific λ. The transition between turbulence and waves occurs for wave steepness

$$k_\beta = (\beta/2U)^{1/2}, \qquad (84)$$

and energy and enstrophy cascades may be impeded. For a typical midlatitude synoptic motion, $k_\beta \sim 70$ km. This means that eddies in the energy cascade region increase scale to k_β, then propagate as Rossby waves. The waves undergo transition to zonal anisotropy, eventually creating a field of alternating zonal jets with a width scale of $2U_{RMS}/\beta$.

Stratification and topography further modify the behavior of quasigeostrophic turbulence. With stratification, vertical transfers of energy occur through interface vortex stretching and nonlinear modal coupling. Transfer of energy to both smaller and larger scales is possible. The latter occurs through interaction of barotropic and baroclinic wave triads. Barotropic triads involve energy transfer to larger scales only, while for baroclinic triads layers act independently for $k > k_{Rd}$, and energy is transferred to smaller scales for $k < k_{Rd}$, where k_{Rd} is the wave number for the Rossby radius of deformation for the first baroclinic mode.

In general, the effect of topography on two-dimensional turbulence is dependent on the ratio

$$r = \left(\frac{V}{fL}\right)\left(\frac{f}{H}\frac{\partial H}{\partial x}\right)^{-1} \qquad (85)$$

and the amount of energy in each layer. With strong lower-layer flow, cascade occurs until they reach the wavelength of the topography and become locked to f/H.

7.6 Submesoscale Coherent Vortices

Submesoscale coherent vortices are also observed in the ocean (McWilliams, 1985). Lenses of Mediterranean water are observed throughout the Eastern North Atlantic and are readily identified by their water properties and are called *Meddies*. Meddies have diameters of 30–70 km and thicknesses of 0.5–1 km; they are anticyclonic, 15–20 exist at any one time, and they cover about 5% of the area of the Canary Basin. Their core appears to be in solid-body rotation with a period of five days. They appear to erode from their sides and have 2–4-year life-time (Armi *et al.*, 1989). Stratification is minimum in their center, and their motion is primarily through advection by the mean flow. Similar features have been observed in other locations where strong water mass contrasts exist, e.g., the Gulf of California (Bray, 1988).

7.7 Fronts

Fronts are regions with sharp horizontal gradients in temperature, salinity, or density. They occur on a variety of scales: The Antarctic Polar Front circumscribes the Earth at 60° S while in coastal waters fronts associated with internal waves or river plumes may be less than 1 km in length. Fronts may occur in the temperature or salinity field as well as the density field. More generally, fronts are formed by horizontal convergence, by a water-mass boundary, or by coastal upwelling. The Antarctic Polar Front and the Kuroshio/Oyashio Front are regions where intermediate water masses are formed. Surface jets are associated with density fronts.

8. SMALL-SCALE TURBULENCE AND MIXING

Although eddies and quasigeostrophic turbulence serve to stir the ocean horizontally, smaller-scale motions are necessary to connect these scales to scales on which molecular diffusion can act. At the boundaries of the ocean, wind stress and shear provide mechanisms for mixing, while at fronts intrusive features serve to enhance mixing, and within the interior of the ocean, processes associated with internal waves provide energy for mixing. Since turbulence is discussed elsewhere in this Encyclopedia (see TURBULENT FLOWS), we will concentrate on those aspects that are unique to the ocean.

8.1 Instability and Turbulence

There are two major types of instabilities in the ocean relating to the onset of turbulence: static (gravitational) and dynamic. The static stability is given by Eq. (9), so that instability occurs when $E = N^2 < 0$. Dynamic instability usually appears at the interface of two layers moving initially at different velocities. Such an instability is called Kelvin–Helmholtz instability. Two nondimensional numbers, Reynolds number (*Re*) and Richardson number (*Ri*), are used to identify the onset of turbulence in the ocean:

$$Re = LU/\nu, \quad Ri = N^2/U_z, \tag{86}$$

where L, U, and U_z are scales for length, velocity, and vertical gradient of velocity, ν is the kinematic coefficient of molecular viscosity, and N is the buoyancy frequency. If the Reynolds number is greater than some criterion

$$Re > Re_{cr}\,(\sim 10^6), \tag{87a}$$

or if the Richardson number is smaller than some criterion

$$Ri < Ri_{cr}\,(\sim 0.25), \tag{87b}$$

turbulence will occur. These criteria provide useful guidelines for the ocean.

8.2 Generation and Maintenance of Turbulence

The onset of turbulence may occur suddenly through a breakdown of originally nonturbulent flow in certain localized regions. The cause of the breakdown is an instability caused by processes naturally occurring within the ocean, e.g., breaking internal waves, convective overturning, etc. Once turbulence is produced and becomes fully developed in the sense that its statistical properties achieve a steady state, the instability mechanism is no longer required to sustain turbulence. The turbulent kinetic energy (TKE) is converted from the mean flow kinetic energy by the shear production (*S*), and from the potential energy by the buoyancy production (*B*), damped by viscous dissipation (*D*), and transported from or to other regions of flow (*Tr*). This leads to the TKE equation

$$\frac{d(\text{TKE})}{dt} = S + B - D + Tr. \tag{88}$$

Thus, in order to maintain turbulence, one or more generating mechanisms must be active continuously.

8.3 Statistical Characteristics of Turbulence

To describe a turbulent field, it is necessary to know the probability density functions (pdf), which require measurements with sensitive and stable instruments for a sufficiently long period of time and at a fine spatial grid. The problem should become much simpler if the turbulent field could be assumed homogeneous and stationary. In such a case, the turbulent field can be depicted by its spectral density function, $E(k)$, which is generally in power-law form:

$$E(k) \simeq k^{-n}, \tag{89}$$

where k is the wave number. Laboratory results indicate that the exponent n equals 5/3 for the inertial–convective range, and equals 1 in the convective–diffusive range. Oceanic spectra of velocity and temperature appear to behave in this manner.

8.4 Double Diffusion

The ocean generally has a multilayer structure, caused by the stratification of two different components: heat and salt. The rate of molecular diffusion of heat is about 100 times that for salt. Consider a layer of warmer, saltier water above cooler, fresher water, such that the upper layer has the same density as or less density than the lower layer. Then, the saltier water at the interface will lose heat much faster than it will lose salt. If the density difference between the layers is small, the saltier water above may become heavier than the cooler, fresher water below, sink into the lower layer, and form tall thin convective cells, called "fingers." In these cells, descending branches of water diffuse heat horizontally more rapidly than salinity and, therefore, become heavier than surrounding water and continue to sink. A similar mechanism occurs in the rising branch. This process is called double-diffusive convection. The density ratio,

$$R_p = \alpha \Delta T / \tilde{\beta} \Delta S, \qquad (90)$$

is a controlling parameter in the growth rates of salt fingers and the resulting fluxes of heat and salt. The terms ΔT and ΔS are the temperature and salinity jumps between the two layers. There exists a criterion R_{pcr} such that if

$$R_p \leq R_{pcr},$$

double-diffusion takes place. Here R_{pcr} ranges from 1.2 to 2.2. In the laboratory, salt fingers have lengths of 20–30 cm with a spacing between fingers of about 1 cm and a temperature difference of 0.1 °C. In the ocean, this produces a "staircase" structure in temperature, salinity, and density where the steps are meters thick.

Although the occurrence of double diffusion is related to R_{pcr}, the rate of double diffusion has not been parametrized in terms of large-scale oceanic fields. The mass flux associated with double diffusion is downward so that it does not help to maintain the thermocline.

8.5 Microstructure

Bounds for mixing processes can be inferred from large-scale distributions. For example, a vertical advective–diffusive model of the ocean interior,

$$K_v \frac{d^2T}{dz^2} - w \frac{dT}{dz} = 0, \qquad (91)$$

governs the vertical distribution of temperature so that for the vertical velocity $w \simeq 4$ m/yr, the vertical mixing coefficient $K_v = 10^{-4}$ m^2/s provides a bound for other estimates of vertical mixing. The term K_v can be estimated from sufficient measurements of temperature microstructure from free-fall probes. In regions with no lateral gradients, we find

$$K_v = \mu k C, \qquad (92)$$

where the Cox number $C = \overline{(\partial T'/\partial z)}^2 (\partial \bar{T}/\partial z)^2$, and the multiplier μ takes a value between 1 and 3 depending upon layering or isotropy at small scales (Garrett, 1979). Measurements yield $K_v \simeq 10^{-6}$ m^2/s, implying that mixing takes place either at boundaries or water-mass formation regions but at rates that do not seem sustainable. Internal waves or internal tides or biological activity may provide sufficient energy for the missing mixing energy. In general, vertical mixing appears to be inversely proportional to N.

9. CONCLUDING REMARKS

In this review, there were a number of important areas that are normally considered within the purview of physical oceanography that we were unable to describe. These include coastal (Huyer, 1990) and estuarine (Officer, 1976) circulation, sea ice (*q.v.*) and related processes, internal waves, and electromagnetics and associated measurements (Sanford, 1971). The latter may be increasingly important as abandoned submarine cables are provided for research use.

Oceanographers are currently engaged in a number of important programs to understand ocean circulation better. Global surveys using carefully calibrated CTDs and water samplers for collection of nutrient, tritium, radiocarbon, and Freon are underway. These data, with *Topex/Poseidon* altimetric measurements, will provide a better description of thermohaline circulation and poleward transport of heat. They will also serve to constrain numerical ocean circulation models, which in turn can be used to study physical processes and design future experiments. In the Tropical Pacific, moored buoys are providing real-time oceanographic data via satellite, better to understand and predict *El Niño*. Modern supercomputers allow the coupling of global ocean and atmosphere models and hence a better understanding of climate. These computers also allow ocean models to be used for forecasts, which can be substantially improved through better ocean data assimilation techniques.

The ocean dominates the surface of the Earth. It has tremendous capacity for heat, carbon dioxide, and organic waste as well as biological productivity. We need to be able to use this resource in a balanced, ecologically sound manner without inducing anthropogenic change. Toward this end, we need a global ocean-observing system that will allow a continuous quantitative assessment of the state of the ocean.

ACKNOWLEDGMENTS

Ching-Sang Chiu provided valuable comments on Sec. 1.3, and Stevens Tucker assisted with Sec. 1.4. Albert Semtner and Carl Wunsch graciously provided Plates 1 and 3, respectively, and Paul Jessen constructed Figures 1 and 2.

GLOSSARY

Baroclinic: The state of the ocean in which surfaces of constant pressure (isobars) intersect surfaces of constant density (isopycnals).

Baroclinic Instability: A hydrodynamic instability arising from a baroclinic flow and vertical shear, allowing eddies to grow with energy coming from the mean potential energy field.

Barotropic: The state of the ocean in which surfaces of constant pressure (isobars) coincide with surfaces of constant density (isopycnals).

Barotropic Instability: A hydrodynamic instability arising from lateral shear in a current, which allows eddies to grow with energy coming from the mean flow kinetic energy.

Ekman Layer, Ekman Transport, and Ekman Pumping: The Ekman layer is the oceanic planetary boundary layer (typically 100 m thick) through which the stresses exerted on the water surface are transmitted into the ocean. For steady wind stress in the open ocean, the surface wind stress is balanced by the Coriolis force. The Ekman transport is the vertically integrated Ekman-layer currents, and is to the right (left) of the wind in the Northern (Southern) Hemisphere and depends only on the magnitude of the wind stress and the latitude. Ekman pumping is the vertical motion in the Ekman layer forced by convergence and divergence of the surface flow, associated with spatial variability of wind stress.

Convection: Vertical motion caused by dense water lying over lighter water.

CTD: An instrument to measure vertical profiles of conductivity and temperature vs depth (pressure).

Double Diffusion: Convection caused by different diffusion coefficients for heat and salt.

Geostrophic Flow: Flow due to the balance between Coriolis and pressure-gradient forces.

Mesoscale Eddies: Horizontal vortices whose size is controlled by the Rossby radius of deformation, e.g., 10–100 km.

Nutrients: Nonconservative chemicals (phosphate, silicate, nitrate) that identify water properties and are used to trace flow patterns.

Ocean Acoustic Tomography: A technique for determining ocean density and velocity structure by measuring the arrivals of specific acoustic rays.

Potential Density, and Potential Temperature: The density and temperature a water parcel would attain if raised (or lowered) adiabatically to the surface or other specified reference pressure.

Quasigeostrophic Flow: For large-scale motion, departure from geostrophic balance is small. Such flow is called quasigeostrophic flow.

Stratification: The vertical structure of the ocean.

Subduction: Newly stratified water leaving the ocean-surface mixed layer to enter deeper layers.

Subtropical Gyre: A basin-scale subtropical anticyclone.

Sverdrup Transport: The wind-driven meridional transport in the ocean interior. It includes both the Ekman transport and the geostrophic transport caused by Ekman pumping.

Thermohaline Circulation: The flow field associated with the sinking of dense water at high latitudes, its subsequent equatorward flow in deep western-boundary currents, and finally deep flow into the ocean interior and upwelling.

T-S Diagram: A characteristic diagram used to study water properties and to identify water masses and water types.

Ventilation: Any process that brings deep water to the surface so that it exchanges heat, momentum, and gases with the atmosphere.

Western-Boundary Intensification: The appearance of strong currents (e.g., the Gulf Stream and the Kuroshio) along the western boundaries of the oceans, as contrasted to weak currents in the interior and at the eastern boundary. This large-scale east–west asymmetry of the ocean circulation is called western-boundary intensification.

Works Cited

Armi, L., Herbert, D., Oakey, N., Price, J. F., Richardson, P. L., Rossby, H. T., Ruddick, B. (1989), "Two Years in the Life of a Mediterrean Salt Lens," *J. Phys. Oceanogr.* **19**, 354–370.

Bray, N. A. (1988), "Water Mass Formation in the Gulf of California," *J. Geophys. Res.* **93**, 9223–9240.

Broecker, W. S. (1991), "The Great Ocean Conveyor," *Oceanography* **4**, 79–89.

Ekman, V. M. (1905), "On the Influence of the Earth's Rotation on Ocean Currents," *Arch. Math. Astron. Phys.* **2**, (11).

Garrett, C. (1979), "Mixing in the Ocean Interior," *Dyn. Atmos. Oceans,* **3**, 239–265.

Gordon, A. L. (1978), "Deep Antarctic Convection West of Maud Rise," *J. Phys. Oceanogr.* **8**, 600–612.

Huyer, A. (1990), "Shelf Circulation," in: B. Le Mehaute and D. M. Hanes (Eds.), *The Sea,* Vol. 9A, New York: John Wiley and Sons, pp. 423–466.

Jerlov, N. G. (1976), *Marine Optics,* Amsterdam: Elsevier Scientific Publishing Co.

Joyce, T. M. (1991), "Review of U.S. Contributions to Warm-Core Rings," *Rev. Geophys.* Suppl. 610–616.

Killworth, P. D. (1983), "Deep Convection in the World Ocean," *Rev. Geophys. Space Phys.* **21**, 1–26.

Kuo, H. L. (1949), "Dynamical Instability of Two-Dimensional Nondivergent Flow in a Barotropic Atmosphere," *J. Meteorol.* **6**, 105–122.

Luyten, J. R., Pedlosky, J., Stommel, H. (1983), "The Ventilated Thermocline," *J. Phys. Oceanogr.* **13**, 292–309.

McWilliams, J. C. (1985), "Submesoscale Coherent Vortices in the Ocean," *Rev. Geophys.* **23**, 165–182.

Munk, W. H. (1950), "On the Wind-Driven Ocean Circulation," *J. Meteorl.* **7**, 79–93.

Munk, W. H. (1966), "Abyssal Receipes," *Deep-Sea Res.* **13**, 707–730.

Officer, C. B., (1976), *Physical Oceanography of Estuaries and Associated Coastal Waters,* New York: Wiley.

Richardson, P. L. (1983), "Gulf Stream Rings," in: A. R. Robinson (Ed.), *Eddies in Marine Science,* New York: Springer-Verlag, pp. 16–65.

Robinson, A. R. (1983), *Eddies in Marine Science,* New York: Springer-Verlag.

Sanford, T. B. (1971), "Motionally Induced Electric and Magnetic Fields in the Sea," *J. Geophys. Res.* **76**, 3476–3492.

Schmitz, W. J., Jr., Holland, W. R. (1986), "Observed and Modeled Mesoscale Variability Near the Gulf Stream and Kuroshio Extension," *J. Geophys. Res.* **91**, 9624–9638.

Schmitz, W. J., Jr., McCartney, M. S. (1993), "On the North Atlantic Circulation," *Rev. Geophys.* **31**, 29–49.

Stommel, H. (1948), "The Westward Intensification of Wind-Driven Ocean Currents," *Trans. Am. Geophys. Union* **99**, 202–206.

Stommel, H. (1958), "The Abyssal Circulation," *Deep-Sea Res.* **5**, 80–82.

Stommel, H. (1965), *The Gulf Stream: A Physical and Dynamical Description,* 2nd ed., Berkeley: Univ. of California Press.

Sverdrup, H. U. (1947), "Wind-Driven Currents in a Baroclinic Ocean, with Application to the Equatorial Currents of the Eastern Pacific," *Proc. Natl. Acad. Sci. U.S.A.* **33**, 318–326.

UNESCO (1981), *Tenth Report of the Joint Panel on Oceanographic Tables and Standards,* UNESCO Technical Papers in Marine Sci., Vol. 36, Paris: UNESCO.

Further Reading

Apel, J. R. (1987), *Principles of Ocean Physics,* New York: Academic Press.

Chu, P. C., Gascard, J. C. (1991), *Deep Convection and Deep Water Formation in the Oceans,* Elsevier Oceanography Series, Vol. 57, Amsterdam: Elsevier.

Clay, C. S., Medwin, H. (1977), *Acoustical Oceanography,* New York: John Wiley and Sons.

Gill, A. E. (1982), *Atmosphere-Ocean Dynamics,* New York: Academic Press.

Godin, G. (1972), *The Analysis of Tides,* Toronto: Univ. of Toronto Press.

Kinsman, B. (1965), *Wind Waves,* Englewood Cliffs, NJ: Prentice Hall, Inc.

Pedlosky, J. (1987), *Geophysical Fluid Dynamics,* 2nd ed., New York: Springer-Verlag.

Pickard, G. L., Emery, W. J. (1990), *Descriptive Oceanography: An Introduction,* 5th ed., Oxford, England: Pergamon Press.

Pond, S., Pickard, G. L. (1983), *Introductory Dynamical Oceanography,* 2nd ed., Oxford, England: Pergamon Press.

Preisendorfer, R. W. (1976), *Hydrologic Optics,* Vols. I–IV, Honolulu, HI: National Oceanic and Atmospheric Administration.

Robinson, A. R. (1983), *Eddies in Marine Science*, New York: Springer-Verlag.

Semtner, A. J., Chervin, R. M. (1992), "Ocean General Circulation from a Global Eddy-Resolving Model," *J. Geophys. Res.* **97**, 5493–5550.

Stommel, H. (1965), *The Gulf Stream: A Physical and Dynamical Description*, 2nd ed., Berkeley: Univ. of California Press.

Stommel, H. (1987), *A View of the Sea*, Princeton: Princeton University Press.

Stommel, H., Moore, D. W. (1989), *An Introduction to the Coriolis Force*, New York: Columbia University Press.

Warren, B. A., Wunsch, C. (1981), *Evolution of Physical Oceanography*, Boston: MIT Press.

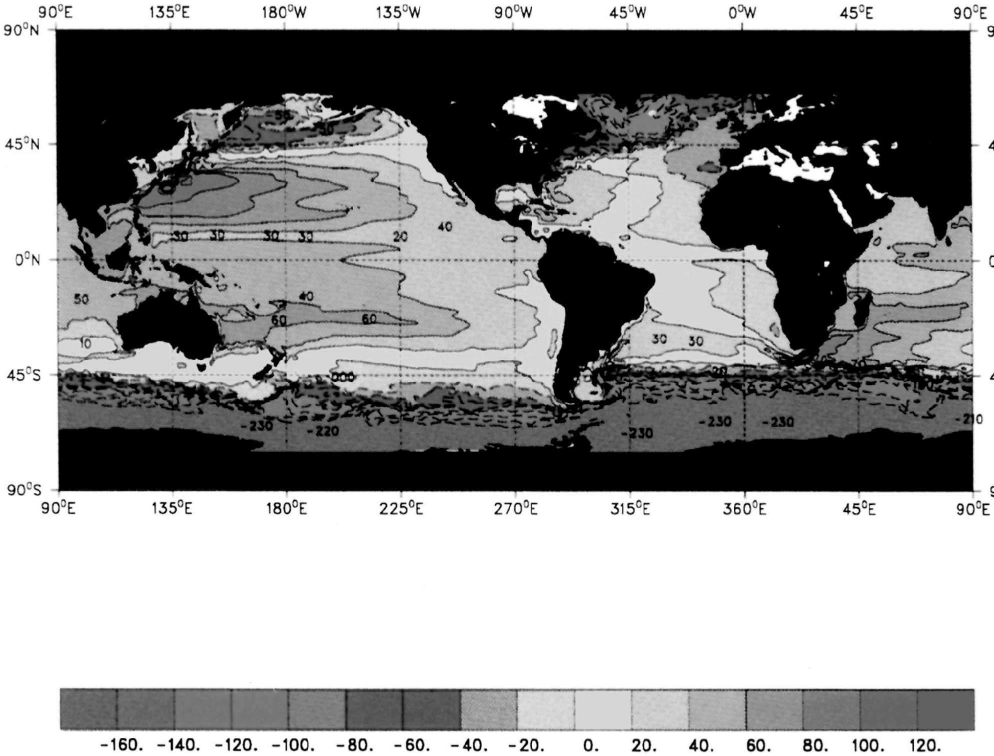
90°E
135°E
180°W
135°W
90°W
45°W
0°W
45°E
90°E
90°N
45°N
0°N
45°S
90°S
90°E
135°E
180°E
225°E
270°E
315°E
360°E
45°E
90°E
50
30
30
30
30
30
20
40
50
40
60
60
10
30
30
-230
-220
-230
-230
-230
-210
-160.
-140.
-120.
-100.
-80.
-60.
-40.
-20.
0.
20.
40.
60.
80.
100.
120.

Vol. 12 – PLATE 1. Dynamic topography of an ocean model sea surface. Results were obtained from a global eddy-resolving ocean model. Units are centimeters. (Courtesy A. Semtner.)

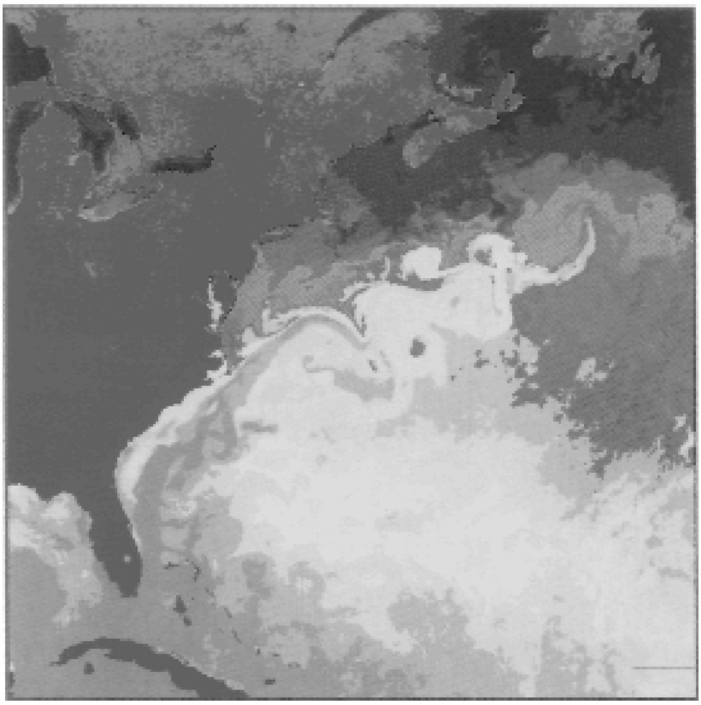

Vol. 12 – PLATE 2. Composite Advanced Very High Resolution Radiometer (AVHRR) image of the Western North Atlantic during April 1984. The image was produced by compositing data from 35 satellite passes obtained during the first week in April; this was necessary to remove clouds. The image is printed with an artificial color scale; Reds and oranges are warm, 24°–28° C; yellows and greens 17°–23° C; blues 10°–16° C; and purples are cold, 2°–9° C. The Gulf Stream appears as a band of warm water, originating in the Florida Straits, separating from North America at Cape Hatteras, and thence flowing west in the North Atlantic, developing meanders. Two cold-core rings appear to be almost pinched off from the Stream to the south of the southern tip of Nova Scotia. (Courtesy NASA.)

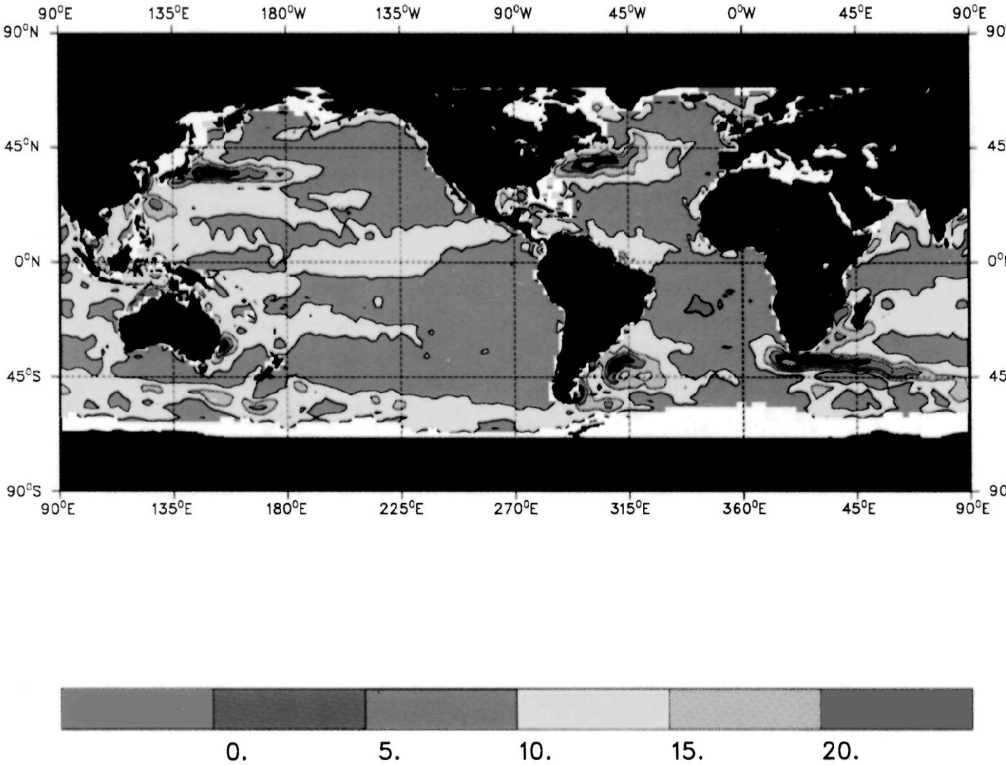
90°E
135°E
180°W
135°W
90°W
45°W
0°W
45°E
90°E
90°N
45°N
0°N
45°S
90°S
90°E
135°E
180°E
225°E
270°E
315°E
360°E
45°E
90°E
0.
5.
10.
15.
20.

Vol. 12 – PLATE 3. Root mean square variability of sea surface height determined by *Geosat*. Contours are centimeters. (Courtesy C. Wunsch.)

OIL SHALES AND TAR SANDS

Paul A. Petzrick, *1116 Gumbottom Road, Crownsville, Maryland, U.S.A.*

	Introduction	77
1.	**Origin, Properties, and Characteristics**	78
1.1	Origin and General Characteristics	78
1.2	Relevant Theories Regarding the Origin of Hydrocarbons	80
1.3	Specific Characteristics and Properties	81
2.	**Resources**	81
2.1	Major Deposits	81
2.1.1	Western Canada Basin	81
2.1.2	Eastern Venezuela Basin	82
2.1.3	Aldan Arch (Russia)	83
2.1.4	Green River Formation (U.S.A.)	83
2.1.5	Great Appalachian Basin (U.S.A.)	84
2.1.6	Baltic Basin (Europe)	84
2.2	Historically Significant Resources	84
2.3	Summary of World Resources	85
3.	**Production**	86
3.1	Historical Production	86
3.2	Current Production	87
3.2.1	Brazil	87
3.2.2	Canada	88
3.2.3	China	88
3.2.4	Estonia	88
3.2.5	Russia	88
3.2.6	Trinidad	88
3.2.7	United States	88
3.3	Prospects for Future Production	89
3.3.1	Australia	89
3.3.2	United States	89
3.3.3	Venezuela	89
4.	**Processing and Applications**	89
4.1	Mining	89
4.1.1	Surface Mining	89
4.1.2	Underground Mining	89
4.2	Crushing	90
4.3	Direct Use	90
4.4	Methods of Oil and Gas Production	90
4.4.1	Surface Pyrolysis	91
4.4.2	*In Situ* Extraction	92
4.5	Upgrading and Refining	92
5.	**Economic Aspects**	92
5.1	Historical Profitability	94
5.2	Current Profitability	94
5.3	Prospects for Future Financial Feasibility	94
5.4	Role of Byproducts and Specialty Products	95
5.5	Relationship to Petroleum Recovery Technology	96
6.	**Occupational, Environmental, and Legal Issues**	96
6.1	Historical Health Concerns	97
6.2	Environmental Issues	97
6.3	Legal Issues	97
	Glossary	98
	Works Cited	98
	Further Reading	99

INTRODUCTION

Oil shale and tar sand are relatively recent terms for rocks containing hydrocarbon materials that are of marginal economic interest to present society but have been used for centuries. Interest in oil shales and tar sands (OS&TS) has declined over the last decade, repeating historic cycles related to oil prices and the prognosis for petroleum supplies. The profile of OS&TS use since prehistoric time reflects widespread availability and relative ease of extraction of useful products from these resources by primitive technologies. Large-scale production of liquid fuel from OS&TS using modern technologies, however, has not generally been competitive with modern petroleum operations. The history of

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

small-scale operations exploiting these ubiquitous solid hydrocarbons, including asphaltites and lake asphalt, to produce a variety of products has been more stable. The components of the hydrocarbon spectrum, whether progressing from oil shale to petroleum to tar sands or from peat to coals, are formed and transformed under a great variety of conditions from diverse materials. As a result, the hydrocarbon resources found today have an enormous variety of physical properties and compositions. The terminology that has evolved for these resources is inconsistent. Labels, from ancient use of the word "asphalt" to the modern misnomers, oil shale and tar sand, are usually related to first use, discoverer, or location of first discovery rather than to composition. Historically, the efficiency of OS&TS extraction technologies has continually improved, whereas use of these substances has reflected cycles in markets and world economics. Just as OS&TS became important resources when wood and whale oil became scarce, they may become crucial resources as petroleum becomes harder to find and recover. OS&TS are comparable to other natural resources in that the occupational health, environmental, and legal issues associated with their extraction and use are receiving more attention as the world's growing population becomes more sensitive to global health and environmental issues and organizes to deal with them.

1. ORIGIN, PROPERTIES, AND CHARACTERISTICS

OS&TS are important components of the spectrum of hydrocarbon resources found at or near the Earth's surface. Modern geology has established with certainty that these ubiquitous resources are part of a continuum of organic materials and the conversion products thereof present in the Earth's crust. This spectrum of sometimes valuable materials graduates from just-fallen plant, animal, and marine life, to rocks containing the petroleum precursor kerogen, to petroleum (natural bitumens, oil, and natural gas), and on to petroleum degradation products resulting from the varying degrees of geologic processing to which the original buried organic materials were subjected. The resultant resources grade from one to another with great variety depending upon how the important parameters of depositional condition, pressure, temperature, and geochemistry changed over time. This continuum, as currently understood by geologists, is depicted in Fig. 1 with annotations of theoretical estimates of the order of magnitude of kerogen or equivalent energy value present in each of the important hydrocarbon forms (Durand, 1980, p. 29).

1.1 Origin and General Characteristics

Most deposited organic material is initially converted to the solid hydrocarbon called kerogen by diagenesis as the combination of descending mineral and organic materials is transformed into sedimentary rock. A minor portion is converted directly to natural bitumen. Rocks containing significant quantities of kerogen are called oil shales. In addition to being insoluble in most solvents, kerogen does not melt upon heating but is converted to vapors, some of which can be condensed to oil.

Kerogen is widely recognized as the precursor of petroleum. If, however, the host rock is not subsequently subjected to the geologic pressures and temperatures required to convert the kerogen to oil and gas, the oil shale (and coal) remains for man to find and exploit for its hydrocarbon content. All sedimentary rocks laid down since life began on Earth contain some kerogen. In most of the great depositional basins, the processes of deposition, burial, and geologic change continued to the point that some kerogen underwent catagenesis, and great quantities of petroleum were produced. In many cases, the petroleum migrated to the traps that are today's oil and natural gas reservoirs. As petroleum migrated, some of the liquids saturated porous or unconsolidated formations, or filled joints or voids; there the petroleum underwent thermal or biological degradation. The lighter fractions evaporated, leaving the heavy fractions to cool and with their new host material to form a reservoir with zero mobility. These reservoirs that do not produce petroleum by conventional primary methods are now called tar sands.

Most hydrocarbon materials that fit the definition of tar sands are natural asphalts,

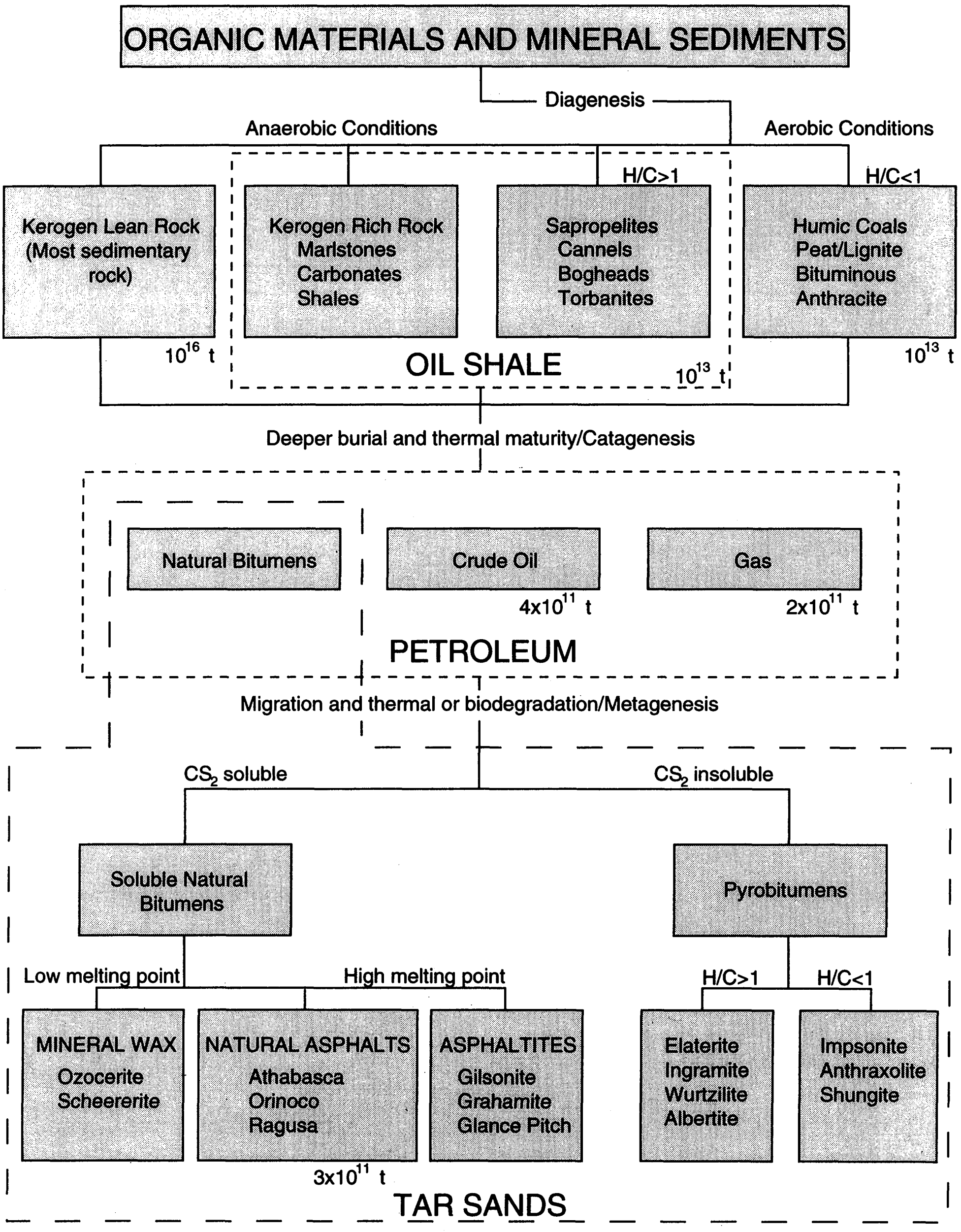
ORGANIC MATERIALS AND MINERAL SEDIMENTS
Diagenesis
Anaerobic Conditions
Aerobic Conditions
H/C>1
H/C<1
Kerogen Lean Rock (Most sedimentary rock)
Kerogen Rich Rock Marlstones Carbonates Shales
Sapropelites Cannels Bogheads Torbanites
Humic Coals Peat/Lignite Bituminous Anthracite
10^{16} t
OIL SHALE
10^{13} t
10^{13} t
Deeper burial and thermal maturity/Catagenesis
Natural Bitumens
Crude Oil
Gas
4×10^{11} t
2×10^{11} t
PETROLEUM
Migration and thermal or biodegradation/Metagenesis
CS_2 soluble
CS_2 insoluble
Soluble Natural Bitumens
Pyrobitumens
Low melting point
High melting point
H/C>1
H/C<1
MINERAL WAX Ozocerite Scheererite
NATURAL ASPHALTS Athabasca Orinoco Ragusa
ASPHALTITES Gilsonite Grahamite Glance Pitch
Elaterite Ingramite Wurtzilite Albertite
Impsonite Anthraxolite Shungite
3×10^{11} t
TAR SANDS

FIG. 1. The spectrum of hydrocarbons in the Earth's crust.

which are soluble, melt on heating, vaporize at higher temperatures, and can be cooked out of their host materials in boiling water. In a few cases, nearly pure petroleum moved into joints in adjacent rock and degraded into highly paraffinic material called mineral wax. In other isolated cases, petroleum carrying fine sediment escaped into joints in adjacent rock beds and was converted into solid hydrocarbons called asphaltites. At a few famous locations, such as Pitch Lake on Trinidad Island, a substantial amount of petroleum carrying fine sediment found its way to the surface, where it ponded and cooled to form asphalt lakes. At a few locations, such as New Brunswick, Canada, geologic conditions were such that *insoluble* natural bitumens formed, usually in veins or in voids in rock. These solid hydrocarbons have characteristics much like the kerogen in oil shale or coal and are called pyrobitumens.

As geologists, geophysicists, and geochemists better recognized the differences between kerogen as a precursor of petroleum and the products of petroleum degradation, the latter became known as "asphalts and allied substances." As the term asphalt came to be used for products manufactured from petroleum, the natural products became known as "natural asphalts and related substances." The modern term "tar sands" initially referred to deposits such as those in Canada and Venezuela that are mixtures of natural asphalt and sand, but the definition of tar sands evolved to recognize all host mineral matrices and to rely on viscosity to differentiate tar sands from heavy oil. Whether intended or not, the resultant definition includes all of the solid products of natural petroleum degradation, as well as the natural bitumens left behind in petroleum source rock, petroleum migration paths, and petroleum reservoirs. Tar sands are so delineated in Fig. 1.

In most cases, petroleum lost its mobility by degradation or metamorphosis after moving up along faults and layers of porous materials from where it was formed. Both oil shales and tar sand, therefore, are generally found at shallower depths than petroleum.

Resources cannot be classified as oil shale solely on the basis of composition because oil shale is an economic term. Although frequently misused as a lithologic term, oil shale means only that the rock is of interest for the production of oil. No minimum yield is specified, nor should one be. For instance, in China lean shale of 20 L/t has been processed for oil for 70 years because it had to be removed to reach the coal below it. At the rich end of the spectrum, there has been a 150-year controversy about what is coal and what is shale or pyrobitumen. To avoid this controversy, Fig. 1 includes only kerogenous rocks that are good sources of oil as oil shale. Humic coals have hydrogen-to-carbon atomic ratios of less than unity and were not considered good sources of oil using the simple pyrolysis practiced when the term oil shale came into use. The sapropelites, which can be lithologically shale or coal depending on the amount of mineral matter they contain, have hydrogen-to-carbon atomic ratios greater than unity and are excellent sources of oil using simple pyrolysis. The Australian and European oil industries were initiated on sapropelites and considered them the feedstock of choice. The early North American oil industry preferred the famous pyrobitumens from New Brunswick, albertite and sterralite, as feedstock but quickly turned to using sapropelites as pyrobitumen resources were exhausted (Ramsay, 1910).

1.2 Relevant Theories Regarding the Origin of Hydrocarbons

Theories about the origin of hydrocarbons are important for predicting their future availability and finding additional resources. Historically, petroleum and asphalt were believed to be primordial materials of the Earth that rose from its bowels through fissures in the crust for man to find and use. However poorly founded on knowledge, this was an abiogenic theory for the origin of hydrocarbons. As it became necessary to find and develop oil reservoirs systematically, petroleum geologists associated petroleum with organic materials deposited in sedimentary basins, and the biogenic theory of the origin of petroleum became dominant. In recent decades, as space exploration has provided new insights into the existence of hydrocarbons in the universe, the possibility of at least an abiogenic contribution to the availability of hydrocarbons in the Earth's crust is gaining credibility. A deep biogenic gas theory suggests that organic material falling to the ocean floor is pushed under the continental plates at subduction zones and is ultimately converted to methane or hydrogen-rich thermogenic gas.

The combination of these theories suggests that descending sedimentary materials *may be* enriched with hydrogen by ambient gas rising in the crust of the Earth. The nature of the sediment falling with the organic material and the geochemistry of the sedimentary basin will determine the type of rock formed with the kerogen. The role of hydrogen enrichment in the conversion of organic material to kerogen and the subsequent conversion of kerogen to petroleum is uncertain. Many petroleum geochemists now believe that water plays an important role in petroleum formation. When kerogen is converted to petroleum at greater depths, escaping gases may enrich younger layers of sediment; therefore hydrogen enrichment does not depend on deep gas or an abiogenic source of hydrogen or hydrocarbons.

1.3 Specific Characteristics and Properties

The specific characteristics and properties of OS&TS are important for planning extraction techniques and predicting economics. Table 1 is a summary of mineral content of the three geologic types of oil shale present in the world's largest known deposits. Some physical characteristics and properties of these shales are presented in Table 2. The properties of tar sands vary with those of the host rock and the impregnating hydrocarbon. Table 3 presents a summary of the known important properties of the tar sands in three of the world's largest deposits. The properties presented in these tables are only representative of the world's OS&TS; each deposit is unique, and the characteristics and properties of each must be determined to plan extraction. Characteristics and properties also vary by location and horizon within the deposit.

2. RESOURCES

The kerogen in sedimentary rocks is converted to petroleum as depositional basins mature, and petroleum degrades into tar sands as basins become overmature. OS&TS deposits are therefore best understood when the depositional basin in which they were formed is viewed as a whole.

2.1 Major Deposits

The geography and geology of six of the great depositional basins of the world are discussed below with a brief description of the OS&TS deposits found in them. These descriptions should be viewed as important data points in the Earth's crust and its history that illustrate the great variety and broad distribution of OS&TS. The quantity of kerogen in oil shales and asphalt in tar sands in just these six basins is sufficient to serve man for many centuries.

2.1.1 Western Canada Basin The Western Canada Basin (Meyer, 1987, p. 6) covers 770 000 km^2 in Alberta; tar sands deposits cover 31 000 km^2. Hydrocarbon deposits are found in this basin in the entire range of sedimentary rocks from Devonian to late Cretaceous. The depth and thermal maturity of

Table 1. Typical composition of oil shales.

Component	Location		
	Green River Formation	Appalachian Basin	Baltic Basin
	Composition (% by weight)		
Kerogen	13.4	13.9	25–30
Quartz	17	30	6.5
Analcine	12		
Feldspars			4.9
Orthoclase	3	9	
Albite	7		
Buddingtonite	3.5		
Clays			10.1
Illite	4	22	
Kaolinite	0.4	8	
Chlorite		2	
Mixed Clays		4.8	
Carbonates			
Dolomite	26	2	
Calcite	13		40.7
Siderite		0.7	
Pyrites	0.7	7.6	3
Al, Ca, Mg, Fe, and Ti oxides (not accounted for above)			5.3

Table 2. Typical characteristics and properties of oil shales.

	Location		
Characteristic or Property	Green River Formation	Appalachian Basin	Baltic Basin
Depositional condition	Lake Basin	Epicontinental Sea	Marine Platform
Geological era	Early and Middle Tertiary	Silurian and Devonian	Ordovician
Typical oil yield (L/t)	100	50	233
Porosity (% of bulk vol.)	0.14	0.57	0.1
Gross heating value (GJ/t)	5.38	4.85	12.45
H/C ratio of kerogen	0.13	0.09	0.12
Water yield (L/t)	58.4	23–44	94
Compressive strength (kg/cm^2)			
Perpendicular to bedding planes	700	1900	914
Parallel to bedding planes	750	1650	907

the basin are greatest along its southwest edge, next to the Rocky Mountain uplift. Numerous oil and gas reservoirs have been delineated throughout the basin. The world's greatest deposits of tar sands are found in the shallow northeast portion of the basin. One of the geologic mysteries of this basin is that the combined volume of oil (1.35×10^{11} barrels) in the four largest tar sands deposits (Athabasca, Cold Lake, Wabasca, and Peace River) and the numerous oil reservoirs in the basin cannot be explained by the volume of source rock present. The asphalt of these tar sands deposits is biodegraded and is most logically explained as the residue of once-mobile petroleum that saturated the extensive deposits of unconsolidated sand on the northeast boundary of the basin. Current reservoir temperatures in the tar sands deposits are 13–21 °C. At these temperatures the hydrocarbon residue is not mobile. Considerable tar sands in this area are exposed or have limited overburden; therefore, they are suitable for large scale surface mining.

2.1.2 Eastern Venezuela Basin The Eastern Venezuela Basin (Meyer, 1987, p. 6) covers 155 000 km^2 and contains the largest combined accumulation of oil, heavy oil, and tar sands in the world. The tar sands deposits cover 41 000 km^2. The basin extends north from the Orinoco River to off-shore beyond the north coast of South America and off-shore to the east to include Trinidad. As in the Western Canadian Basin, most of the tar sands have accumulated in the shallow part of the

Table 3. Typical characteristics and properties of tar sands. Source: principally Meyer (1987).

	Location		
Property	Western Canada Basin	Eastern Venezuela Basin	Aldan Arch
Host material	Unconsolidated sands	Sandstone and surface lakes	Voids and joints
Bitumen condition	Biodegraded	Biodegraded	Thermal-degraded
Reservoir age	Cretaceous	Tertiary, Cretaceous	Riphean, Vendian, Cambrian
API gravity	3°–8°	<6°–10°	<10°
Composition		% by weight	
Sand	83	65–70	[a]
Bitumen	13	24–27	0.8–50
Water	4	6–8	[a]
Sulfur content of bitumen	4–5	3.2–3.8	[a]

---

[a] The Aldan Arch Region in Russia has only recently been recognized as surpassing in magnitude the tar sands resources of the Melekess Basin, thereby becoming the third largest source of tar sands in the world. The solid hydrocarbons of the Aldan Arch Region are distributed in nine zones and have highly variable properties.

basin, in this case the southern boundary area, formerly called the Orinoco Tar Belt. Some of the material, normally defined as tar sands with API gravities less than 10° at normal temperatures, can be exploited by primary methods because reservoir temperatures in the basin range from 38 to 85 °C. The name of the area has been officially changed to Orinoco Heavy Oil Belt to reflect the presence of some oil that can be produced by primary methods.

Two famous lake asphalt deposits, Trinidad's Pitch Lake and Bermudez Pitch Lake near Guanoco, Venezuela, punctuate the northeast corner of the Eastern Venezuela Basin. These unusual occurrences result from very sediment-laden, heavy, warm oil flowing up along faults in this area to the surface.

2.1.3 Aldan Arch (Russia) The Aldan Arch (Meyer, 1987, p. 77) is a large uplift in Russia's Siberian Platform. This single structure with 1 300 000 km^2 underlain by asphalt rocks and asphaltite deposits will probably prove to be at least the third largest tar sands deposit in the world. The origin of the collection of tar sands in this relatively undeveloped area is more easily explained than is the origin of the Alberta tar sands. The Siberian Platform accumulated 9000 m of sedimentary rock from Riphean time (925×10^6 years ago) to Ordovician time. Considerable oil had formed in this great mass of sediment when magma moved under the area, creating the Aldan uplift. Heat from that intrusion drove the volatiles from large oil deposits that had moved into porous formations or the voids and joints created as the area was disturbed, leaving behind extensive asphaltites and other tar sands.

2.1.4 Green River Formation (U.S.A.) The Green River Formation (Russell, 1990, p. 85) of the Rocky Mountain Region of the United States is the largest and richest resource known on Earth. The term resource is used to describe this complex geologic formation instead of OS&TS deposit because of the unique collection of minerals that are commingled with the kerogen. The formation contains enough kerogen to produce 2×10^{12} barrels of oil, but the formation is also a remarkable collection of massive quantities of unique saline minerals that exist only in minute quantities elsewhere in the world. The formation covers 41 000 km^2 and is up to 2200 m deep.

The genesis of this anomaly was a large, inland stratified body of water that remained remarkably stable for 10^7 years from 60 to 50 Ma ago during the Tertiary Period. With no outlet, the water became very salty at depth but supported verdant algae blooms. These blooms produced the large volume of organic matter that became kerogen as 10 Ma of sediment piled up. The water existed for most of the period in two lakes: Lake Uinta in what is now northwest Colorado and northeast Utah, and Lake Gosuite in what is now southwest Wyoming. These lakes gradually diminished in size and became localized into six basins; clockwise around the Uinta Mountains the basins are the Green River, Great Divide, Washakie, Sand Wash, Piceance Creek, and Uinta. Each basin received slightly different runoff and developed different geochemistry.

The Piceance Creek Basin supported the most consistent growth of algae and was warmer and deeper than the other basins; consequently, surface effects did little to disturb normal stratification of the water. The deep center of the lake became a giant chemical system producing massive quantities of the saline minerals nahcolite, dawsonite, and sodium chloride in what is known as the salt cap area. The production and deposition of algae produced an amazingly uniform 640 m of oil shale. Natural gas has been found in and under the oil shale horizons of the Piceance Basin.

The Uinta Basin was not as deep, and algae growth was not as great as in the Piceance Basin. As a result, oil shale in the Uinta Basin is not as rich, except for the Mahogany Zone of rich shale that pervades the entire Green River Formation. An earlier sediment basin near the Uinta Mountains was buried and heated sufficiently to produce the Cedar Rim and Blue Bell oil fields of the Altamont Trend. This earlier basin was also the likely source of the petroleum that migrated upward and formed the tar sands of Asphalt Ridge; the famous vein asphaltite, gilsonite; the pyrobitumen, wurtzilite; and the mineral wax, ozocerite, in and near the Uinta Basin.

Algae growth in the more northern basins was not as great nor as persistent; thus, these basins have only layers of rich oil shale. In the Green River Basin, varying periods of the unique geochemistry laid down layers of trona (a hydrous acid sodium carbonate) of sufficient thickness to be mined adjacent to the richest oil shale horizons. The entire Green River Formation was subsequently lifted in the mountain-building process, and the great quantities of oil shale remain as originally formed.

2.1.5 Great Appalachian Basin (U.S.A.) When the North American continent was still connected to a great land mass to the east during middle Devonian time, great rivers fed the giant inland Chattanooga Sea, which stretched from the present Appalachian Mountains to the center of the Great Plains. Organic growth in the inland sea was sufficient to produce some rich oil shale and enormous quantities of lean oil shale over a wide area. Variations in river flows produced clastic intrusions from the east, and subsequent glaciers and erosion buried or carved away great parts of the oil shale. Earlier depositions produced oil and gas under parts of the shale, but the Devonian/Mississippian/Pennsylvanian shales were never buried deep enough or otherwise matured to convert the kerogen in them to petroleum. Collectively, what remains of these shales has been estimated to contain sufficient kerogen to produce 2.6×10^{12} barrels of oil. Scattered sapropelic and tar sands deposits, including some asphaltite veins, dot the large basin. The shales range over 650 000 km^2 and vary in depth from a few meters to 250 m. Overburden varies from zero at outcroppings to hundreds of meters of glacial till. Much of the area has a limestone cap over the oil shale. Even the richest shales in this complex basin have a serious deficiency of hydrogen and leave considerable carbon when pyrolyzed to produce oil (Russell, 1990, p. 85).

2.1.6 Baltic Basin (Europe) The Baltic Basin (Russell, 1990, p. 487) contains the most-used oil shale deposit of this century. The basin contains two oil shale seams. The rich middle Ordovician carbonate or kukersite provides the feedstock for the Estonian oil shale industry. The lower Ordovician Dictyonema shale contains phosphorite but has not been exploited. The kukersite extends over just 5200 km^2 in the area between the Gulf of Finland and Lake Peipus in Estonia and extends east in Rûssia. Overburden varies from zero at the outcrop near the Baltic Sea to 250 m at the southern limits of the shale of commercial interest. Mining is beset by problems with groundwater and weakened areas that must be left as pillars during underground mining. Numerous conflicting estimates of the amount of kerogen in the basin suggest the potential to produce 0.6×10^{12} barrels of oil.

2.2 Historically Significant Resources

The above discussion of six of the world's giant OS&TS deposits confirms the enormous magnitude of these resources in the Earth's crust. Several smaller deposits have had unique roles in the history of use of OS&TS and are noted to illustrate the broad distribution and historical importance of even small resources. The known history of native asphalt begins with hydrocarbon-saturated soil that became known as the slime pits of Hetite in ancient Babylon, and nearly pure asphalt fished out of the Dead Sea (Abraham, 1960, p. 9). Native asphalt oozed into the Dead Sea, solidified, and broke off to rise to the surface of the water in chunks the size of horses. This source was economically valuable in its time; authorities are reported to have issued permits to fish for asphalt in the tepid water of the Dead Sea. These sources and numerous other seeps in the Middle East provided the material for the applications identified in ancient Egypt and Mesopotamia dating from before 5000 B.C. (Abraham, 1960, p. 23).

Oil shale has an even longer history of use. Bracelets of French Autun oil shale found on skeletons of Neolithic man of the late stone age may have been used to repel insects (Russell, 1990, p. 545). The first confirmed production of shale oil has been traced by its name, "Giant's Blood," to a legend that dates to 800 A.D. and the Seefeld oil shale of Austria (Russell, 1990, p. 532). Oil from this shale high in the Austrian Alps near Tyrol gained fame for the treatment of skin diseases and ultimately became the only authentic source of Ichthyol (Blade, 1938). A modern sulfonated version of this medication is still produced at outcroppings of the Seefeld oil shale in Austria and Italy.

The first recorded European use of hydrocarbons in the New World was the caulking of ships at Lake Trinidad during Columbus' third voyage (Abraham, 1960, p. 41). Although discovered first by Europeans, this sticky lake asphalt proved too difficult to haul back to Europe, and so a harder hydrocarbon deposit near Havana, Cuba, became the primary source of exports to Europe for the next century. Europe had numerous small tar sand deposits, and many, such as Seyssel, were worked to exhaustion as the demand for asphalt grew. The Seyssel tar sand deposit in France's Rhone Valley was discovered in 1735. Production there led to the discovery of modern road paving, as pieces of asphalt that fell off the wagons were compacted by subsequent traffic. Seyssel became the most important source of asphalt in Europe until it was exhausted. The asphalt rocks of Ragusa, Sicily, are reported to have been mined sporadically since Roman time. The Wietzer tar sand deposit in Germany was the scene of J. Taube's 1769 development of the hot-water process of extraction and distillation of asphalt that is the basis of modern commercial production of oil from tar sands (Abraham, 1960, p. 45).

The Scottish oil shale deposits near the Firth of Forth are the most famous as sources of oil. Although devoid of any advantage for mass production, these resources supported the Scottish industry for its entire 103 years of competition with the American petroleum industry. The deposits are actually nine thin seams of shale buried in 900 m of calciferous sandstone at the base of a carboniferous system of rocks (Sneddon, 1938, p. 54). The entire system is highly folded and steeply dipping. The shales are lenticular deposits cut by severe faults and subject to flooding by groundwater. The yield varied from 450 L/t initially to as low as 60 L/t—not likely candidates for a role in history.

The best-known oil shale of the 19th century was the rich torbanite of New South Wales, Australia. At the peak of the gas-lighting era, it was discovered that the addition of just 5% of this ultrarich (up to 900 L/t) shale to coal increased the luminosity of the resulting gas by a factor of 6 (Carne, 1903, p. 97). Demand for this unique shale pushed the industry for its production into the wilderness of New South Wales. Pieces of shale left in river valleys by floods were traced upstream to outcroppings in the Sydney Permian Basin. Individual deposits were thin seams of ox-bow shape exposed by the intense erosion of the area. Exposures usually occurred at the top of long talus slopes topped by sheer sandstone cliffs. The shale seams were under intense pressure from extensive overburden but could be mined easily by hand methods because the pressure caused the shale to splinter when struck with a pick. The industry moved to ever more remote areas, solving onerous materials handling problems, as the small deposits were exhausted and new ones discovered. Ultimately, the industry spawned its own railroad to the port at Sydney, where daily boat loads of raw shale were shipped to the gas light systems of the world. The seconds, or leaner shale (less than 300 L/t), were discarded at the mine mouth or processed in local plants for oil, wax, and other products.

The fame of the Australian torbanites led to testing of North American sapropelic rocks of similar appearance as a substitute for the imported product. The spread of agriculture led to discovery of numerous small deposits of cannel and boghead materials in the eastern United States and Canada that were initially identified as coal. Even albertite and stellarite were mistakenly identified by the courts as coal (Gesner, 1865, p. 49). North American sapropelic rocks, whether subsequently proven to be coal or shale, and pyrobitumens became the feedstock of choice for the American pyrolysis industry that blossomed in 1855 and that ended quickly with large-scale production of petroleum in 1859.

2.3 Summary of World Resources

Any tabulation of world OS&TS must recognize that these resources have been neglected for the last hundred or so years during which petroleum has dominated the supply of hydrocarbons worldwide. Except in the most recent two decades, when various oil crises generated interest in OS&TS, these solid hydrocarbons received little attention by the world's major hydrocarbon producers, who were in the position to define and explore them. Even recent interest has been limited to the giant deposits that could produce enough oil to replace petroleum. Most discoveries of deposits and extension of data about known deposits have been anecdotal to the search for petroleum. Generally, the more thoroughly a depositional basin is explored the more OS&TS are found. Thus in the 20 years between 1964, when Ball (1965) did a comprehensive survey of tar sand resources in the United States, and 1983, when Lewin and Associates, Inc. (1983) completed a similar survey, U.S. tar sand resources grew almost tenfold from 5.5×10^9 barrels to 54×10^9 barrels. This suggests that largely unexplored developing countries and undeveloped frontier areas such as arctic Alaska, Canada, and Russia, where some oil has been found, may have enormous undiscovered OS&TS resources. Table 4 presents estimates of the order of magnitude of oil in total resources of OS&TS.

3. PRODUCTION

As depicted in Fig. 1, oil shales, petroleum, and tar sands form a spectrum of hydrocarbon resources. Man has used various forms of these natural resources for centuries as fuels, lubricants, paving, building and manufacturing materials, cosmetics, and medicine. Archaeologists and historians have traced the use of petroleum, natural asphalt, and asphalt rocks to the earliest civilizations in the Middle East prior to 5000 B.C. Archaeologists have also found evidence of centuries of use of natural hydrocarbons by Chinese, Japanese, and aboriginal Americans of both continents. As the demand for hydrocarbons reached massive quantities during the industrial age, petroleum, which is the easiest form to extract and transport, became the dominant resource used and the use of OS&TS declined.

Table 4. Ultimate recoverable OS&TS resources.

Region	10^9 barrels of oil Oil shale	Tar sands
Africa	4,000	1
Asia	5,500	492
Europe	1,400	21
North America		
Canada	10	2,516
United States	2,990	56
Oceania	1,000	
South America		
Brazil	2,000	
Venezuela		1,182
Total	17,000	4,000

Sources: Preparation of this summary of OS&TS resources revealed numerous conflicts in data regarding these resources. The oil shale amounts are from Duncan (1965) for just the richest (over 104 L/t) shale, and the tar sands amounts are from Meyer (1987). The totals exceed the amounts derived by Durand (1980) of 3.7×10^{12} barrels for oil shale and 2.2×10^{12} barrels for tar sands. Duncan's total for oil shales yielding over 21 L/t is 2.092×10^{15} barrels.

3.1 Historical Production

Cycles of use of OS&TS predate their competition with petroleum. The ancient uses of asphalt in the Middle East faded as Roman building technology using cement became dominant. Subsequently, the growth of cities fostered the broad use of charcoal as fuel, and societies remote from petroleum sources learned about liquid hydrocarbons as a byproduct of charcoal production. A demand for liquid products, initially easily met with byproducts of charcoal, had been created, and when the forests were used up in populated areas, the availability of fuel of any type became a daily problem. Artificial tar sands, prepared by saturating earth with petroleum or other liquid hydrocarbons, were burned as a substitute for charcoal. Various patented or manufactured solid fuels using hydrocarbon binders became items of commerce. In 1838, builders, especially of railroads, rediscovered the advantages of pickling wood with liquid hydrocarbons. Coupled with the growing demand for liquid fuel for lighting, the resultant demand for liquid hydrocarbons became overwhelming. Prices for whale oil in the United States rose to \$16/L by 1846.

The oil shortage of 1838 triggered the first modern effort to perfect mass production of oil from solid hydrocarbons. Much was already known from the gas-lighting industry, which had started to produce gas from coal, other solid hydrocarbons including oil shales, and various organic materials by pyrolysis during the late 18th century. Petroleum was not unknown at the time, but its transportation problems had not been solved. As early man used more hydrocarbons, he found a more widespread and certain source in mining and cooking oil from rocks than in traveling great distances to skim oil from water wells, ponds, springs, and streams or primitive oil wells. Considering the primitive nature of transportation at the time, this was probably the correct economic choice. Even so, where liquid hydrocarbons could be found close to the Earth's surface, this relatively convenient resource was exploited. For instance, in 1838 there were at least 82 crude wells near Baku and 3500 pits for the production of oil on an island in the Caspian Sea. Marco Polo had described exporting oil from this area by animal train in 1298. Muspratt (1860, Vol. 1) also reported 520 oil wells operating in Burma exporting 400 000 hogsheads [$(0.6–1.4)\times10^6$ barrels) of oil per year prior to 1859.

It is a staggering paradox, therefore, that the petroleum industry waited until 1859 to explode onto the world from North America. This paradox did give rise to a short but flourishing and relatively modern oil shale industry. After a jump start from this short market pull and subsequent government protection from the predatory marketing practices of the U.S. petroleum industry, oil shale plants operated in Scotland, France, Spain, Estonia, Sweden, Germany, Brazil, and Australia. Some plants survived for many years in competition with the petroleum industry. In North America, there would be no such longevity. The initial oil shale plants of 1855 in Boston and New York, which operated on imported torbanites and pyrobitumens, and the Breckenridge, Kentucky, plant, which used cannel coal as feedstock, quickly grew to over 55 plants in the eastern United States by 1859. The largest was a 25 000-L/d plant near Pittsburgh. All were shut down or converted to oil refineries as large-scale petroleum production started in Pennsylvania in 1859.

Tar sands were well known in the 1830s, but this sticky material that caked so easily was hard to work with for the production of oil or fuel. Concerning this period, Muspratt (1860) reported what he considered many impractical attempts to use native asphalt in applications for which it was not suited. Natural asphalts, which appeared to have found their broadest use in early civilizations, made a dramatic comeback as road-paving material during the late 19th century. As the demand for paved roads spread across North America, dozens of local tar sands deposits were mined to supply paving material. The Kyrock Mine in Kentucky was a typical operation that shipped asphalt rock by river barge and rail to pave hundreds of miles of roads in midwest America. Two such mines still operate in Uvalde, Texas, supplying paving material to a distance of 175 km from the mines. Paving was the dominant use of tar sands for 50 years from 1870 to 1920. Production of paving material reached 0.25 Mt per year in both Europe and North America. The production of asphalt from petroleum was patented in 1894 and was gradually perfected. After 1920, the use of synthetic asphalt prepared from petroleum and local aggregates became more economical for most paving. In Russia, where distribution of petroleum products is still a problem, the local production and use of natural asphalts are reaching the levels reported during the early part of this century in Europe and the United States.

Among the tar sands, asphaltites in particular have been the basis of stable industries producing highly varied products for years. Grahamite, initially mined in West Virginia during the 1850s and burned as coal, has since been used to produce tar, lubricants, illuminating gas, and varnish. The gilsonite industry of Utah, in operation since 1888, has produced ink, insulation, paving material, roofing, varnish, asphalt additives, sealers, gasoline, and carbon control rods for nuclear reactors.

3.2 Current Production

OS&TS were worked in at least eight countries in 1992.

3.2.1 Brazil Brazil has nine significant oil shale formations. The Brazilian oil shale industry started in 1881 with 20 Scottish retorts operating on Tertiary shale at Taubate in San Paulo State. This plant operated until 1920, sat idle until 1941, and was then renovated and expanded for wartime production of oil, wax, and gas. A second company operating on Cretaceous shale in Bahia erected a plant of 52 Scottish retorts in 1884 and produced illuminating gas, oils, paraffin, candles, soap, and sulfuric acid. Production was equivalent to about a thousand barrels of oil per day. Several other plants produced shale oil products at intermittent times throughout World War II. Oil shale was also used as railroad locomotive fuel (Russell, 1990, p. 172). The high water content of the Cretaceous oil shale was a problem. In an attempt to improve the economics of Brazil's oil shale operations, the government investigated the dryer Irati shale. The Taubate plant was close to the Irati Formation; it was purchased by the national government in 1951 and became the start of the Petrobras research and development program that is currently demonstrating large-scale production of oil from the Irati Formation using the Petrosix technology (see Sec. 4.4.1).

3.2.2 Canada In 1967, after a half century of diverse trials, Great Canadian Oil Sands Ltd. (now Suncor) put the first modern commercial plant for the production of oil from tar sands into operation in Alberta. The Suncor plant is producing at design capacity of 65 000 barrels per day. A more recent effort, the joint venture Syncrude Canada Ltd., attained sustained production of 109 000 barrels per day during 1983 and produced at design capacity of 129 000 barrels per day during 1985. These operations use open-pit mining and hot-water separation of the tar and sand in surface plants. The development of horizontal drilling technology may result in larger-scale oil production from these massive resources by *in situ* methods.

3.2.3 China Oil shale is widely distributed in China; deposits are known to exist in more than a dozen provinces. Two deposits are currently worked. Fushun, in Manchuria, is the most famous and is the world's oldest continuous oil shale operation currently in production. The shale is lean, 20 L/t, but generally has to be removed to get to coal beds. Started as a local activity shortly after discovery of the shale in 1909, the operation has been expanded and renewed several times. In 1924, the Japanese added facilities at Fushun based on Estonian technology. The Japanese plant produced 5500 barrels per day during World War II. The Chinese took over the plant after the war, rebuilt it during 1954, and expanded it during the 1970s. Peak production had reached 7.5×10^6 barrels per year but has dropped to less than 0.5×10^6 barrels per year. A second oil shale operation was started in Maoming, southern China, in the 1950s using richer shale of 75 L/t.

3.2.4 Estonia Kukersite of the Baltic Basin has long been recognized for its fuel value and has been used extensively because of the unusual richness of so large a deposit. Farmers also burned it on their fields to use the ash for fertilizer and as a soil conditioner. Estonia became an independent country after World War I; three private companies and the Estonian government built plants for oil production. By 1938, the Estonian portion of the basin produced a million barrels of oil per year and employed 6000 people, meeting the fuel needs of two-thirds of Estonia's heavy industry. The industry was devastated during World War II, but produced at the rate of 1.75×10^6 barrels of oil per year for a period during German occupation. In 1945, Estonia became part of the Soviet Union. As part of implementing a succession of Five Year Plans, the Soviet central government built major gas and electric generation plants fueled by kukersite in the Baltic Basin. These plants produced gas and over 3000 MW of electricity for Leningrad and the northwest Soviet Union. The industry is now in decline because natural gas pipelines have been extended to the area, and the power plants have major environmental problems.

3.2.5 Russia Since the 17th century, many projects in Russia have produced natural bitumen and oil shale products (Russell, 1990, p. 470). None can be classified as large-scale oil producers, but the diversity of resources, lack of transportation, and great local demand for a wide range of products have prompted the greatest diversity in use of oil shale and tar sand products in the world. There are numerous local tar sands operations in the former republics of the Soviet Union. For example, a surface mining operation started in the 1950s in Kazakhstan is now producing more than 4 Mt per year. A large portion of the material is used directly for road paving. A 1991 report (Vakhitov, 1991) cited uses for fuels (in liquid, gas, and solid forms), petrochemical feedstock, varnishes, paints, insulating material, metalliferous products, biostimulants, building materials, and paving material.

3.2.6 Trinidad Asphalt from Trinidad's Pitch Lake continues to be produced by the government-owned Trinidad Lake Asphalt Company, Ltd. at the rate of 820 barrels per day, down from peak production of 3100 barrels per day during 1964.

3.2.7 United States Local paving still supports continuous operation of the two tar sand mines near Uvalde, Texas. Limited local operations of this type, once widespread, also continue intermittently, as needed for local paving, at Asphalt Ridge near Vernal, Utah. In the eastern Uinta Basin, commercial production of gilsonite began during 1888 and continues. Also, in the Green River Formation, the important saline minerals associated with oil shale are being exploited; trona has been mined in Wyoming since 1947, and solution mining of nahcolite in Colorado was initiated during 1990.

3.3 Prospects for Future Production

In addition to continuation of the operations listed in Sec. 3.2 at about their present levels, the following projects probably will go into production during the next decade.

3.3.1 Australia A large multinational oil company and an Australian oil shale resource company have persisted for several years in preparations to produce oil from shale in Queensland. Assuming reasonable cooperation from the Australian government, this group is expected to operate an open-pit mine and to retort shale using a modern surface pyrolysis technology within the next decade.

3.3.2 United States With enormous, rich oil shale deposits and widespread shallow tar sands and sapropelic deposits, the United States is a logical location for increased use of these resources. Small projects producing premium hydrocarbon products, mineral coproducts, and chemical by-products are more likely to be feasible earlier than large-scale projects competing with the petroleum industry.

3.3.3 Venezuela The great Orinoco tar sands have not yet been put into production. The high viscosity (100 000 mPa s at ambient temperature) of the bitumen is a problem, but Venezuela has learned a great deal about handling heavy oils from its experience throughout the Eastern Venezuelan Basin since the 1950s. Neither heating nor dilution with lighter crudes is considered to be a cost-effective method for extracting the viscous tar. The decision has been made to try to produce, transport, and burn the bitumen directly as an emulsion using a brine present in the deposit. Venezuela is testing world markets for the new power plant fuel called Orimulsion.

4. PROCESSING AND APPLICATIONS

The method of extracting useful products from OS&TS depends on the product desired, the formation geometry, and the physical properties and characteristics of the pay zone. Attempts to classify extraction methods tend to obscure the range of choices and combinations available to operators. The general categories of mining and direct combustion, mining and surface processing, *in situ* processing, and combinations of these methods have all been used or demonstrated at large scale.

4.1 Mining

Mining methods for OS&TS are very sensitive to the properties of the particular deposit under consideration and the overburden covering the pay zone. The two general categories of mining are surface and underground.

4.1.1 Surface Mining Open-pit mining is the method of choice for mining both OS&TS when depths of overburden permit. Most OS&TS in consolidated formations require blasting and mucking to remove the ore. For unconsolidated tar sand formations and weak shales, direct loading with large shovels or ripping and bulldozing is possible. Suncor's bucket-wheel loader and conveyor system in its Alberta mine is the leading example of large-shovel mining of tar sands. The rich shale of the Baltic Basin can be ripped directly from the mine face and loaded to rail cars or trucks with large shovels. The tar-impregnated Anacacho Limestone of Uvalde, Texas, is blasted and mucked with large shovels and load-haul-dumps, but the weak sandstone of Asphalt Ridge near Vernal, Utah, is ripped with bulldozers and pushed directly to the mixing table for further processing. The unique properties of gilsonite permit very efficient mining of this asphaltite even from very narrow vertical veins using high-pressure water jets. In each case, the physical properties of the ore body, volume of operation, and economics determine the exact choice of method and equipment.

4.1.2 Underground Mining Where depth of overburden requires underground mining, the operator is faced with a far broader set of problems. Access, whether by vertical, horizontal, or sloping adit, must be planned and mined out. A strong competent horizon of rock must be selected for the mine roof. Ventilation, dewatering (if needed), and emergency egress for every stage of mine development must be carefully planned. Room-and-pillar or chamber-and-wall mining have been the methods of choice for most large underground oil shale mines. The heights of rooms or chambers in the Green River Formation have dictated special drilling and blast hole loading equipment to reach the entire face of the mine. Similarly, specialized roof bolting equipment and techniques have evolved for large oil shale mines. The U.S. Bureau of Mines did considerable development work to perfect equipment and techniques for large production mines in the Green River Formation. Cleveland-Cliffs, Inc., Mobil Oil, and Unocal have all operated room-and-pillar mines safely in the Green River Formation. Numerous mechanical miners to replace blasting and mucking have been proposed for OS&TS operations, but although several have been carefully tested, none have been used in large-scale production. The Scottish and Australian oil shale mines worked only narrow seams of a meter or two compared to the 27-m cuts used in the hundreds of meters deep ore body of the Green River Formation. During 100 years of operation, the Scottish mines evolved from crude hand-pick operations to very sophisticated electrical operations (Sneddon, 1938, p. 53).

4.2 Crushing

An extensive choice of crushing equipment designed primarily for crushing limestone is available and very satisfactory for most OS&TS. Crushing tar sands from unconsolidated formations can be as simple as driving over the run-of-mine material with a bulldozer, as is done to produce road paving material at Asphalt Ridge. For harder ores, gyratory, jaw, cone, and roll crushers have been used successfully. Particle size is usually dictated by the choice of subsequent oil extraction process or firing method.

4.3 Direct Use

Historical evidence suggests that oil shales have been used directly as fuel for hundreds of years. The use of kukersite as power plant fuel in Estonia is the leading contemporary example of this practice.

The outstanding direct use of tar sands has been for road paving. Processing for this application may range from simple compaction of run-of-mine material to some crushing and softening with hydrocarbon solvent, as is done at Uvalde, Texas, or blending with water, as is done at Asphalt Ridge in Utah.

4.4 Methods of Oil and Gas Production

Converting of kerogen in oil shale to oil and separating bitumen from the host tar sand material have many similarities, but the properties of bitumen offer more choices for extracting useful products. The bitumen in tar sands melts, can be decomposed or cracked at high temperatures, can be washed off the host material with hot water in some cases, and is soluble. Kerogen in oil shale does not melt and is insoluble; therefore, it is generally separated from the host rock by application of heat to decompose the solid kerogen into gases and residual carbon. This process is called pyrolysis and is completed in a reaction chamber called a retort for both surface and *in situ* operations. The atmosphere in which pyrolysis occurs is critical because the newly formed products of pyrolysis are very reactive with any other elements present in the pyrolysis zone, particularly oxygen. Retorts are therefore designed to have inert gases sweep the product gases out of the pyrolysis zone to a condenser or condensing zone in the retort where the oil and uncondensable gas fractions are separated. The residual carbon and uncondensable gases may be burned to provide heat for the process. Maximum oil production requires pyrolysis at the lowest possible temperature (about 480 °C) to avoid unnecessary cracking of the hydrocarbon molecules, which reduces oil yield. Higher temperatures can be used to produce lighter oil and a greater volume of gas at the expense of oil yield. Supercritical extraction techniques have been tested on oil shale to improve oil yield. This involves washing intermediate products of kerogen conversion out of the host rock at lower temperatures than used for pyrolysis.

Where permeability of the host rock is sufficient or can be created, several methods of *in situ* heating are technically feasible for both oil shales and tar sands. The basic objective of both surface and *in situ* operations is to complete the minimum decomposition of the hydrocarbon component necessary to get it free of the mineral matrix and do as little heating of the mineral matter as possible. Minimum heating of the mineral matter is particularly important for carbonate shales because the carbonate decomposition reaction if triggered is highly endothermic and generates unwanted carbon dioxide but no useful products.

4.4.1 Surface Pyrolysis The most recently developed surface pyrolysis technology is the Alberta Oil Sands Technology and Research Authority (AOSTRA) Taciuk Processor (ATP). Developed for tar sands, the ATP has also been tested on oil shale and has been used to remove hydrocarbons from contaminated soil. This rotary retort, developed for thermal treatment of Alberta tar sands, is the method that has been chosen for the Australian oil shale project previously mentioned (Kuss, 1994). The ATP is illustrated in Fig. 2.

Another surface pyrolysis technology suitable for both oil shales and tar sands is the Lurgi–Ruhrgas technology developed in Germany to process coal fines. This technique, which features a lift pipe to recycle hot solids, has been successfully tested for processing the diatomaceous earth tar sands of the McKittrick Formation in California and for Green River oil shale. It has been found useful on a wide range of feedstocks. Figure 3 shows a configuration of the Lurgi–Ruhrgas technology suitable for processing tar sands.

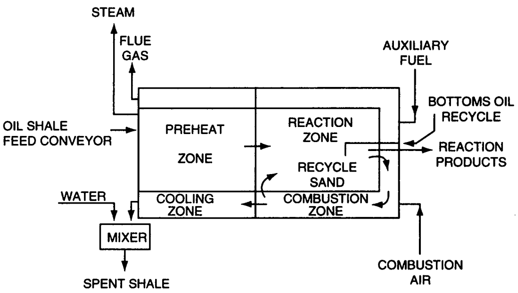

FIG. 2. AOSTRA Taciuk Processor (ATP). The ATP consists of a single, horizontal, rotating vessel in which separate compartments or zones are maintained by the geometry of the equipment and flow of gases. The reaction zone is heated by a combination of conduction from the combustion zone, where residual carbon is burned off the spent feedstock, and a portion of the hot spent feedstock, which is recycled to the reaction zone.

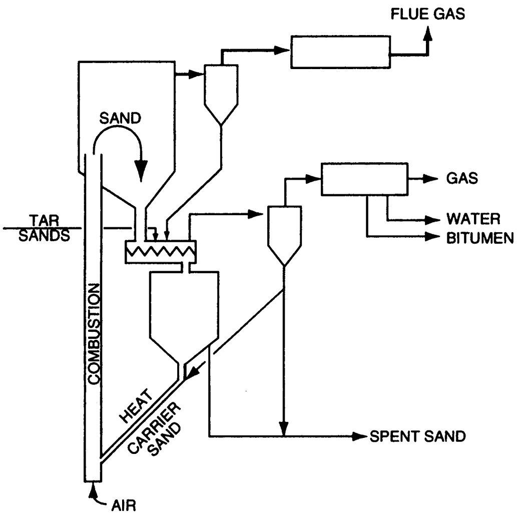

FIG. 3. Lurgi–Ruhrgas process. The unique feature of this system is the lift pipe in which residual carbon is burned off spent feedstock to provide heat. Burned feedstock is carried to the retort for solid-to-solid heat transfer to raw feedstock.

The largest surface oil shale pyrolysis reactor currently operating is the Petrosix 11-m vertical-shaft retort used in Brazil's development program. The modern vertical-shaft retort is the product of years of evolution that can be traced to the original change from horizontal- to vertical-shaft retorts in the Scottish oil shale industry. James Young and the Scottish oil shale companies shifted from the fixed horizontal retorts that had been adopted from the coal gasification industry to higher-throughput, vertical-shaft retorts to produce oil. A succession of outstanding engineers following Young improved the Scottish retorts and set the standard of efficiency for the world. The Scottish designs were exported to France, Australia, Spain, Brazil, and the United States. These retorts were indirectly heated by coal. When U.S. Bureau of Mines engineers set out to develop a high-efficiency, high-throughput oil shale retort specifically for the Green River Formation shale, they studied the Scottish design and elected to develop a vertical-shaft retort, called the Gas Combustion Retort (GCR), that would burn the uncondensable gases of the retorting process as fuel. Engineers who worked on that program went on to develop and build the Petrosix GCR for Brazil and the Paraho version of the GCR, which has been used extensively at pilot scale for various research and testing programs in the United States. The Paraho version is illustrated in Fig. 4 as an example of a modern, high-throughput, vertical-shaft retort.

Upon completion of the Bureau of Mines GCR demonstration program, a consortium of major oil companies leased the pilot and demonstration plants at Anvil Points, Colorado, and conducted proprietary testing. Subsequently, three of the major oil companies developed alternative concepts. Union Oil Company (later Unocal) perfected its rock pump retort to the demonstration stage. During 1949, Standard Oil Company (later Exxon) tested and patented a fluid-bed retort. Chevron demonstrated an ebullient or boiling-bed retort during the 1970s. The latter two designs are extensions of oil company experience in operating sand crackers and fast catalytic crackers in modern refineries.

During the most recent period of high interest in oil shale, the Kiviter retort replaced older equipment for the production of liquids from kukersite in Estonia. China continued to use the Pintsch retorts. The Institute for Gas Technology, with support from the U.S. government, developed hydroretorting in a vertical-shaft retort for the hydrogen-deficient Devonian shales.

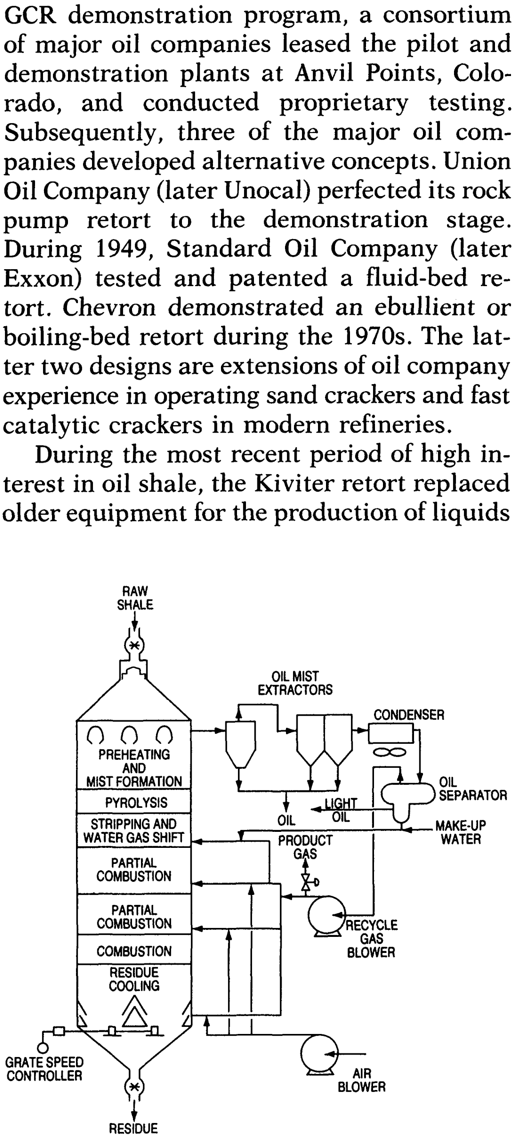

FIG. 4. Paraho retorting process. The Paraho process is typical of vertical-shaft retorts in which crushed shale with the fines removed descends through the retort under the influence of gravity. Zones for each step in processing the shale are maintained by managing gas flow in the retort. The direct combustion mode is shown here. When operated in the indirect mode, process gas is burned in a separate furnace and hot gases carry heat to the retort.

4.4.2 *In situ* Extraction *In situ* pyrolysis or heating of both oil shales and tar sands in place to produce oil has been tested extensively. One particularly successful operation involved using cheap hydroelectric power to heat shallow beds of shale in Sweden. The U.S. government provided extensive support of *in situ* OS&TS research and development during the 1970s. The cooperative program with industry resulted in the Geokinetics horizontal *in situ* process for shallow beds of oil shale and the Occidental vertical modified *in situ* process for deep, thick oil shale beds being tested to demonstration scale. These concepts are illustrated in Figs. 5 and 6.

4.5 Upgrading and Refining

Extensive testing and practice in the former oil shale industries of the world and Canada's tar sand industry have adequately proven that shale oil and bitumen from tar sands can be refined into the full range of petroleum products. Some shale oils contain high concentrations of nitrogen or arsenic and must be hydrotreated to be suitable as feedstock for conventional refineries. Trace metals, sulfur, and chlorides can also be a problem, just as they are with some crude oils. Historically, most oil shale plants were so isolated that small refineries of custom design were collocated with them to process their output. This was also the case for the Utah gilsonite industry. When it was decided to make gasoline from this unique resource, a small refinery of special design was constructed at Fruita, Colorado. The larger oil shale plants contemplated during the 1970s would have included upgrading facilities, as did Unocal's demonstration plant. Minimal treatment in custom-designed upgrading facilities is required to make feedstock suitable for modern refineries.

5. ECONOMIC ASPECTS

The monumental works of Abraham (1960) and Russell (1990) chronicled the rise and fall of OS&TS industries around the world. The

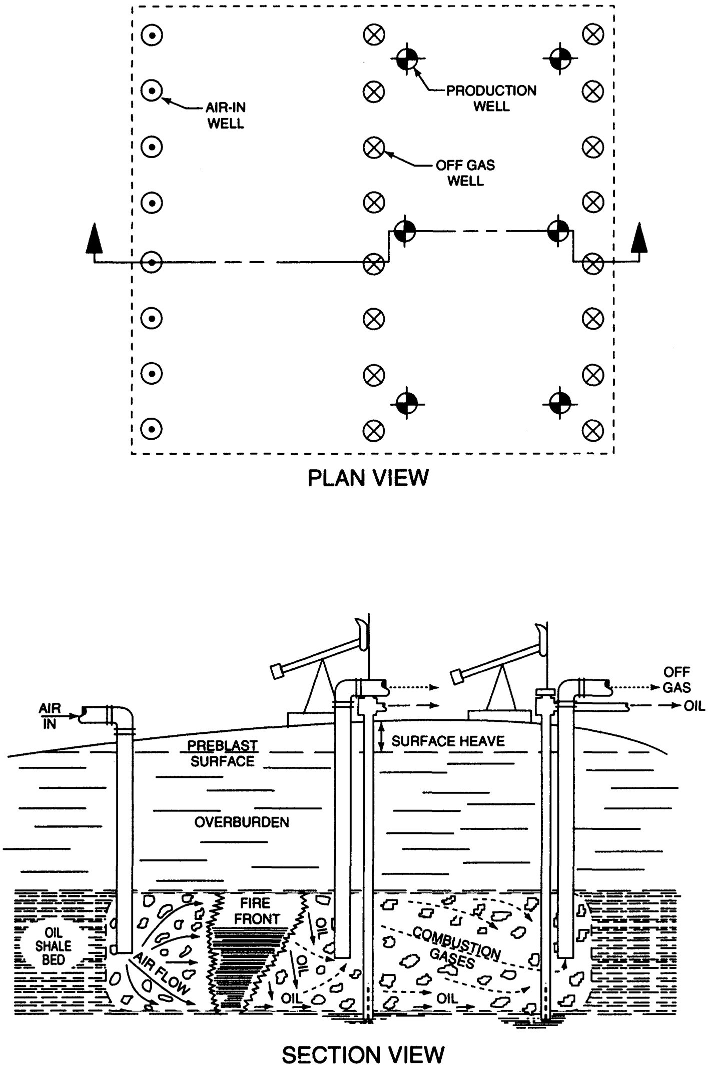

FIG. 5. Plan view and cross section of typical Geokinetics horizontal *in situ* process. This scheme for retorting shallow beds of shale in place depends on uniform blasting of the shale and surface heave to create void space for gas to flow through the shale. The shale is set afire at the air inlet wells, and air flow is managed to move hot inert combustion products through the balance of the shale ruble.

existence of an industry cannot be assumed to prove independent economic viability because special tax structure and tariff policy are well known to have sustained industries that otherwise would not have been profitable. The simplicity of just heating rock to get oil accounts for the long history of use of OS&TS. The economic competition between petroleum and even this simple method of extraction accounts for the enormous variety of methods tried, as inventors attempted to improve the economics of operating with solid feedstock. No "cotton gin" of pyrolysis technology, however, was developed to fix the cost of producing shale or tar sands oil at or near that of petroleum extraction. At the time of the birth of the modern petroleum industry (1859), Muspratt (1860, Vol. 2, p. 99) reported a dozen methods for heating both hydrocarbons that melted (resins and asphalt) and others that vaporized or decomposed (kerogen and organic waste). By the time the U.S. Bureau of Mines surveyed oil shale patents a century later (Klosky, 1949, 1958), more than 1500 patents had been registered, and there are now about 2000. No commercial-scale production of oil from shale or tar sands since the emergency fuel production required during World War II, other than projects that received government subsidies, has survived more than a few years in the era of cheap Middle East oil. Neither have the opportu-

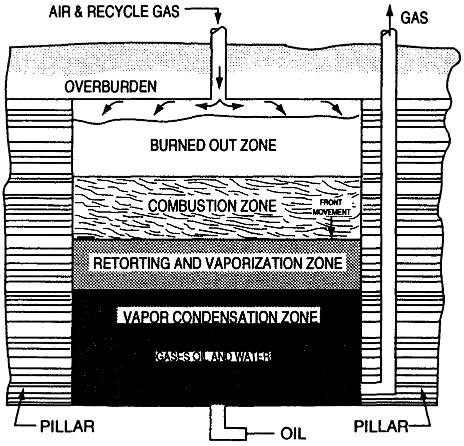

FIG. 6. Occidental vertical modified *in situ* process. This concept was developed specifically for the deep, thick shale beds of the Green River Formation. About 20% of the shale in the retort area is mined; the balance of the shale is then carefully blasted using the mined-out volume to permit expansion and uniform distribution of void space throughout the retort. A combustion zone is started at the top of the retort and moved down through the shale rubble by management of combustion air and recycled gases. Full-scale retorts would contain 350 000 m^3 of shale rubble.

nities offered by the fact that the bitumen in tar sands is soluble resulted in success. Most solvent projects failed because of the loss of expensive solvent adhering to the host rock or sand. The economics of all OS&TS projects were seriously impaired by the extra cost of mining and handling solid feedstock and disposal of spent feedstock. The materials handling problem has forced the development of ever more efficient surface plants and provided the impetus to experiment with *in situ* methods.

5.1 Historical Profitability

The Australian oil shale industry of New South Wales is a concise example of an industry's response to competition, outside influences, government policy, and technological change. As shown in Fig. 7, the small industry started experimentally in 1855, reached continuous production status in 1865, and flourished in relative isolation until the financial panic of 1873. After that, the struggling industry needed tariff protection from the predatory marketing practices of Standard Oil Company. The Provincial Government of New South Wales provided that protection, and the fledgling industry recovered and grew with the local economy. During 1902, central government came to Australia and canceled the protective tariffs in response to perceived greater national interests, and the oil shale industry went into decline. The gas mantle, invented in 1911, eliminated the export market for Australia's raw shale, and the tenacious little industry died shortly thereafter. It sprang to life again in response to the emergency needs in World War II and disappeared when the emergency was over (Sell, 1951).

Other OS&TS industries cannot be so neatly capsulized, but all that had reasonably sustained operations began and grew under special circumstances and declined when those conditions ended.

5.2 Current Profitability

The current large-scale operations in Brazil and Canada are not economical in competition with the petroleum industry. These projects barely meet operating costs and are currently paying little return on capital investment. Newell (1993) most recently reported operating costs of \$15/bbl (Canadian) in a facility that would now cost (\$8–\$10)$\times 10^9$ to build. World oil prices at the time of this announcement were in the range of \$17/bbl. The industries in Russia, Estonia, and China are not yet operating in true free-market economies. The proposed project for Queensland, Australia, has requested special tax treatment to compete with the petroleum industry. No current large-scale OS&TS operation can compete with the petroleum industry as a source of refinery feedstock as long as petroleum prices remain low. However, small tar sands operations, producing local paving material, have a long history of profitability. The gilsonite industry of Utah has a checkered history but overall has been profitable.

5.3 Prospects for Future Financial Feasibility

Both upstream and downstream costs are increasing for the petroleum industry and can be expected to increase at even greater rates as convenient reserves are used up and more production must come from sensitive and remote areas. Nevertheless, world petroleum

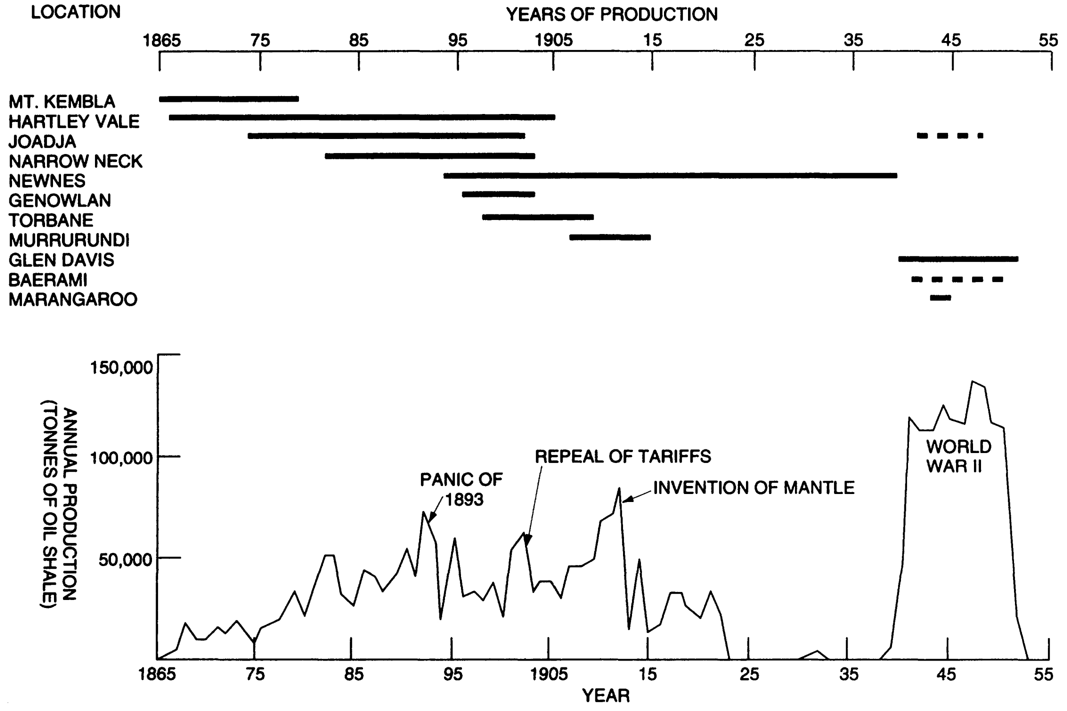

FIG. 7. Australian oil shale industry production.

producers can be expected to continue to operate as an imperfect Von Stackelburg monopoly (Greene and Leiby, 1993) and manage petroleum prices to a degree that makes large-scale production of refinery feedstock from OS&TS a financially unacceptable venture. Petroleum producers will keep prices between the long-term and short-term cost of producing petroleum, and both of these boundaries will increase with time. Unfortunately, justification of large OS&TS operations to produce oil in competition with petroleum requires a sharp increase in world petroleum prices at a predictable time. Sharp increases in petroleum prices, such as occurred in the 1970s, are certain to occur again, but their timing is uncertain. The most likely scenario will be more projects that expand the envelope of primary, secondary, and tertiary recovery of crude oil into the most convenient and most easily worked tar sands, including depleted shallow petroleum reservoirs. This movement forward along the geologic spectrum from conventional petroleum to heavy oil, extra heavy oil, and then tar sands is more likely to happen than any giant step to oil shale operations in competition with petroleum. Canadian *in situ* tar sands operations using horizontal drilling are the most likely candidates for success in this scenario. The production cost for this extraction technology is expected to be $6.00/bbl (Canadian). In any country where modern transportation is available, oil shale can compete with petroleum as a source of liquid fuel only if given special tax treatment, import protection, royalty forgiveness, or even direct government support. In some countries, these incentives may be provided to develop domestic resources, improve balance of trade, and create domestic employment.

5.4 Role of Byproducts and Specialty Products

The wide range of products and characteristics of shale oils and bitumen from tar sands makes them excellent sources of petrochemical feedstocks. The current liquid production in Estonia is used almost entirely for petrochemical production. The historically successful oil shale industries usually produced a range of products in addition to oil and gas. Oil shale development is most likely to occur as small projects justified by the high dollar value of specialty products, special petrochemical feedstocks, coproducts, and byproducts. The oil shale resources of the United States are most amenable to this type of development. The Green River Formation, as mentioned in Sec. 2.1.4, is a veritable warehouse of saline minerals and hydrocarbons. The New Paraho Corporation has found the high nitrogen content of Green River Formation shale oil, which is usually a problem for refining, to be an advantage in producing modern asphalt paving modifiers that greatly improve the performance of conventional asphalt produced from petroleum derivatives. Several test strips of highway paving using their Shale Oil Modified Asphalt Technology (SOMAT) under a variety of climatic condition show that the shale oil–derived modifier produces asphalt pavement of such superior characteristics compared to conventional asphalt that its production at commercial scale is economically justified (Lukens, 1993).

In addition to the current production of trona and nahcolite from the Green River Formation, dawsonite in the Piceance Basin is an enormous resource for potential domestic production of alumina. The potential for multimineral operations improves the prospects for overcoming the barriers to successful oil shale operations. The University of Kentucky is pursuing a similar approach in investigating integrated production of several products from Devonian shale.

5.5 Relationship to Petroleum Recovery Technology

Eventual large-scale production of shale oil in competition with petroleum will hinge on a better understanding of how nature converts kerogen to petroleum, including the role of water and ambient gases that may be present in the natural process. This requires fundamental research in aqueous organic chemistry and how the atmosphere in which pyrolysis occurs influences reaction rates and pyrolysis products. Similarly, large-scale production of refinery feedstock from tar sands in competition with petroleum will depend on better understanding of the thermal and biological degradation processes that generate tar sands. Some large oil companies and governments are sponsoring research in these areas.

The terms "oil shale" and "tar sands" may be misnomers, but they identify the largest and most widely distributed hydrocarbon resources in the world, with a close technological and geological relationship to petroleum. These resources represent an assured future source of petrochemical feedstock, even if hydrocarbons are no longer used as fuels for environmental or economic reasons. The technology for extraction of useful products from these ubiquitous resources, whether by mining and surface processing or by *in situ* methods, is particularly crucial when related to present tertiary petroleum extraction technology. Reservoir engineers concur that current primary petroleum extraction methods leave behind as much as two-thirds of the oil in the ground. Although the world has extracted between $(435 \text{ and } 440)\times 10^9$ barrels of oil from petroleum reservoirs and will probably extract that amount again, twice as much oil will remain in petroleum reservoirs and have to be extracted by methods akin to the extraction technologies for OS&TS. This fact alone establishes an enormous economic reward for developing the most efficient OS&TS extraction technologies. The real value of such technologies will only become apparent in the future, as they are used to clean out petroleum reservoirs. Economics will most likely dictate that exposed deposits and shallow reservoirs will be first worked by these methods, followed by ever deeper operations.

Ultimately, mankind may have to resort to use of the less economic portions of the spectrum of hydrocarbon materials available to us, such as fuels manufactured from renewable biomass, unconventional gas resources, and dirtier coals. Large-scale OS&TS operations have favorable economics in such scenarios.

6. OCCUPATIONAL, ENVIRONMENTAL, AND LEGAL ISSUES

The coincidence of renewed interest in OS&TS during the 1970s, and implementation of the environmental ethic of the 1960s, created a clash of interests that delayed many OS&TS projects. The long history of OS&TS operations, frequently at isolated locations, provides many opportunities for case studies of the effects of long-term exposure to raw OS&TS and their products on flora, fauna, and human populations. The common source and close relationship between raw OS&TS and their products and petroleum and petroleum products provide little reason to expect surprise environmental impacts from future OS&TS operations.

6.1 Historical Health Concerns

Health problems in the early oil shale industries depended on the nature of the feedstock and had much to do with the primitive state of technology and industrial hygiene. Antisell (1859) reported on a Rhenish feedstock that "on removing the cap of the retort, a strong smell of arsenic is perceived, and the workmen suffer from colic, ulcers at the root of the nose, of the joints, and an irritable condition of the skin." As early as 1920, Scottish shale oils, particularly the high-temperature fractions, were shown to have higher levels of carcinogenicity than other mineral oils tested (Key, 1974). As a consequence, the health histories of Scottish oil shale workers have been the subject of extensive studies. These studies confirmed that there was a higher incidence of skin cancer, especially scrotal cancer, among workers with prolonged exposure to Scottish shale oil and shale oil products. Classic high rates of scrotal cancer reported among British textile workers who oiled spinning jennies with shale-derived lubricants decreased as alternative lubricants were used. In general, the incidence of disease decreased as special precautions were instituted and improved hygiene became standard practice. Studies of the health of oil shale workers and residents in the area of the modern oil shale industry in Estonia reveal that an excess of skin cancer occurred only in women who worked in the industry for 10 or 20 years (Coffin, 1981). Investigations of health problems at isolated oil shale operations and villages in South Africa and Australia show that the workers and their families were generally very healthy but suffered from lack of immunity to communicable diseases brought in by visitors (Knapman, 1988). No unusual incidence of disease has been reported from current oil shale operations in Brazil or tar sand operations in Canada.

6.2 Environmental Issues

OS&TS share a common environmental problem with coal and petroleum: their use as fuel releases carbon dioxide and sulfur compounds to the environment. The carrying capacity of the environment, therefore, may limit future use of these fuels. In addition, OS&TS have the unique problem that the mining and extraction of their hydrocarbon content at commercial scale leaves enormous quantities of solid residue. Some shale oils, like some crude oils, contain enough heavy metals to be of concern. Many OS&TS contain connate water or water of hydration that results in an overall excess of water from their processing, which must be treated as wastewater. Air emissions are a potential problem for processing or burning OS&TS just as they are for any other hydrocarbon. Pyrolysis products such as shale oil are generally viewed as containing more carcinogens and toxic substances than other hydrocarbons. Special precautions must be taken regarding pyrolysis products from any source until their exact nature is known. The emissions from direct combustion of kukersite in Estonia are reported to be a severe environmental problem. The Estonian power plants do not have modern emission controls. Regardless of location, large-scale OS&TS operations will produce major socioeconomic impacts, population displacements, and transportation problems.

All of the above problems have been studied extensively (Bates, 1981). For the most part, there are technical solutions to the environmental issues of OS&TS development. Ultimately, each nation and the world collectively will have to make decisions regarding the tradeoffs involved in development of natural resources as population growth and world development demand greater quantities and varieties of products. The risks, dangers, and environmental damage of producing petroleum in sensitive frontier areas and transporting petroleum and petroleum products over great distances must be weighed against the environmental costs of meeting society's demands from more distributed alternatives.

6.3 Legal Issues

The marginal economics of OS&TS can make ordinary legal issues into insurmountable barriers to development. Among the legal issues are ownership of resources, mining and environmental laws, use of public lands and resources, and tax structure. The same laws and tax structure that are a nuisance to petroleum development can be an absolute barrier to OS&TS development because of the longer lead time and higher capital cost of the latter. Governments that seek to promote development of these alternatives to petroleum must be sensitive to the expectations of entrepreneurs evaluating higher-risk projects. Government ownership of the oil shale plant in Brazil, near-partnership operation of the tar sands plants in Canada, and support of clean fuels production in the United States are examples of government actions needed to promote large-scale production of alternatives to petroleum.

GLOSSARY

API Gravity: An arbitrary scale expressing the gravity or density of petroleum and petroleum products. The measuring scale is calibrated in terms of degrees API and is calculated by the following formula:

$$\text{Degrees API} = 141.5 \times [\text{specific gravity at } 15.56\ ^\circ\text{C}]^{-1} - 131.5.$$

Catagenesis: Regressive stage in organic metamorphism in which chemical bonds in large molecules of kerogen are broken by thermodynamic factors to form smaller petroleum (oil and gas) molecules.

Coal: Sedimentary rock containing sufficient carbonaceous material, usually of vegetable origin, to be of interest as fuel.

Dawsonite: A mineral consisting of basic sodium carbonate occurring in white, bladed crystals.

Diagenesis: Reconstructive process by which changes are produced in sedimentary rock during or immediately after their deposition and which is caused by such forces as weight of overburden or hot water. Stage in organic metamorphism in which disseminated organic material is converted to large molecules of kerogen.

Kerogen: The insoluble hydrocarbon material in sedimentary rock that upon destructive distillation forms oil and gas.

Lithologic: Relating to the character of rock in terms of its structure, mineral composition, color, and texture.

Metagenesis: Stage of organic metamorphism in which petroleum is altered by thermal or biological degradation under conditions that change the carbon structure and produce solid or nearly solid hydrocarbons.

Nahcolite: A mineral consisting of natural sodium bicarbonate.

Natural Asphalt or Bitumen: Soluble natural hydrocarbon material having less than 10° API gravity at ordinary temperatures.

Oil shale: Rock containing sufficient kerogen to be of economic interest as a source of oil.

Pyrobitumens: Natural bitumens that as a result of metamorphosis are insoluble in carbon disulfide and have characteristics similar to kerogen but are degradation products of petroleum.

Retort: A closed vessel or chamber with an outlet, used for distillation, sublimation, or decomposition by heat.

Sapropelites: Coal or oil shale formed from organic-rich sediment of marine, estuarine, or lacustrine deposition consisting of largely organic debris from aquatic plants and animals.

Tar Sands: Hydrocarbon-bearing materials formed by petroleum impregnation of mineral matter or voids along its migratory path in which the petroleum is subsequently degraded or metamorphosed leaving hydrocarbon material having an *in situ* viscosity greater than 10 000 cP at reservoir conditions. (*In situ* viscosity data are frequently not available, and so a gravity cutoff value of less than 10° API gravity or zero mobility at ordinary temperatures is used as a default value. This definition has evolved to rely on *in situ* viscosity as the key physical property and omits reference to mineral matter present with the hydrocarbon. Thus, the definition includes all of the solid products of petroleum degradation as well as natural bitumen left behind as petroleum is expelled or removed from reservoirs or migratory paths.)

Trona: Hydrous acid sodium carbonate in crystals or fibrous or columnar masses as a deposit from various soda-brine springs and lakes.

Works Cited

Abraham, H. (1960), *Asphalts and Allied Substances, Their Occurrence, Modes of Production, Uses in the Arts, and Methods of Testing*, 6th ed., vol. I, New York: Van Nostrand.

Antisell, T. (1859), *The Manufacture of Photogenic or Hydro-Carbon Oils from Coal and Other Bituminous Substances, Capable of Supplying Burning Fluids*, New York: D. Appleton and Co., p. 128.

Ball Associates, Ltd. (1965), *Surface and Shallow Oil-Impregnated Rocks and Shallow Oil Fields in the United States*, Oklahoma City: Interstate Oil Compact Commission.

Bates, E. R., Thoem, T. L. (Eds.) (1981), *Environmental Perspective on the Emerging Oil Shale Industry*, Washington, DC: U.S. Government Printing Office, EPA-600/2-80-205a.

Blade, O. C. (1938), *Ichthyol—Its Source and Properties*, Washington, DC: U.S. Bureau of Mines (USBM Information Circular 7042).

Carne, J. E. (1903), *The Kerosene Shale Deposits of New South Wales*, Sydney, Australia: Geological Survey of New South Wales, p. 97.

Coffin, D. (1981), in: Bates E. R., Thoem, T. L. (Eds.), *Environmental Perspective on the Emerging Oil Shale Industry*, Washington, DC: U.S. Government Printing Office, EPA-600/2-80-205a, p. 100.

Duncan, D. C., Swanson, V. E. (1965), *Organic-Rich Shale of the United States and World Land Areas*, Washington, DC: U.S. Geological Survey (USGS Circular 523).

Durand, B. (1980), *Kerogen Insoluble Organic Matter from Sedimentary Rocks*, Bayeux: Institut Francais du Petrole.

Gesner, A. (1865), *A Practical Treatise on Coal, Petroleum, and Other Distilled Oils*, New York: Bailliere Brothers, p. 49.

Greene, D. L., Leiby, P. N. (1993), *The Social Costs to the U.S. of Monopolization of World Oil Market, 1972–1991*, Oak Ridge, TN: Oak Ridge National Laboratory (ORNL-6744).

Key, M. M. (1974), in: Committee on Science and Technology, *Oil Shale Technology*, Washington, DC: U.S. Government Printing Office, p. 458.

Klosky, S. (1949), *An Index of Oil-Shale Patents*, Washington, DC: U.S. Bureau of Mines (USBM Bulletins 467 and 468).

Klosky, S. (1958), *Index of Oil-Shale and Shale-Oil Patents: 1946–56, Parts I, II, and III*, Washington, DC: U.S. Bureau of Mines (USBM Supplements to USBM Bulletins 467 and 468).

Knapman, L. (1988), *Joadja Creek, The Shale Oil Town & Its People 1870–1911*, Sydney, Australia: Hale & Iremonger Pty Ltd., p. 60.

Kuss, V. H. (1994), *1993 Annual Report—Central Pacific Minerals, N.L.*, Canberra City, Australia: Central Pacific Minerals, N.L., p. 4.

Lewin and Associates, Inc. (1983), *Major Tar Sand and Heavy Oil Deposits of the United States*, Oaklahoma City: The Interstate Oil Compact Commission.

Lukens, L. A. (1993), *SOMAT Field Evaluation Program, Fourth Year Annual Report, U.S. Highway 191, Moab, Utah*, Lakewood, CO: The New Paraho Corporation.

Meyer, R. F. (Ed.) (1987), *Exploration for Heavy Crude Oil and Natural Bitumen, AAPG Studies in Geology #25*, Tulsa: The American Association of Petroleum Geologists.

Muspratt, S. (1860), *Chemistry, Theoretical, Practical, and Analytical, as Applied and Relating to the Arts and Manufactures*, London: William Mackenzie.

Newell, E. (1993), "Oilsands Play Larger Role in Canadian Oil Production," *Oil Gas J.* **91** (31), 25–30.

Ramsay, W. (Ed.) (1910), *Seventh International Congress of Applied Chemistry, Section IVa 1. Organic Chemistry and Allied Industries*, London: Partridge and Cooper, Ltd., p. 23.

Russell, P. L. (1990), *Oil Shales of the World, Their Origin, Occurrence & Exploitation*, New York: Pergamon Press.

Sell G. (Ed.) (1951), *Oil Shale and Cannel Coal*, Vol. 2 London: The Institute of Petroleum, p. 186.

Sneddon, J. B. *et al.* (1938), in: A. E. Dunstan (Ed.), *Oil Shale and Cannel Coal*, London: The Institute of Petroleum.

Vakhitov, G. G. (1991), "Soviet Production of Natural Bitumen," in: *Pace Synthetic Fuels Report*, December 1991, Houston: The Pace Consultants, Inc., pp. 3–25.

Further Reading

Descriptions of continuing developments regarding tar sands can be found in the publications of The United Nations Institute for Training and Research/United Nations Development Programme Information Centre for Heavy Crude and Tar Sands and the publications of the Alberta Oil Sands Technology Authority. The *Pace Synthetic Fuels Report* and *Oil and Gas Journal* provide continuous coverage of developments in both oil shales and tar sands. The Oil Shale Association supports the continued collection of oil shale literature in the Tell Ertl Oil Shale Archives at the Colorado School Of Mines. A list of relatively recent textbooks covering kerogen, conversion of kerogen to petroleum, and degradation of petroleum to tar sands follows.

Brooks, J., Welte, D. (1984), *Advances in Petroleum Geochemistry*, Vol. 1, Orlando, FL: Harcourt Brace Jovanovich.

Lee, S. (1991), *Oil Shale Technology*, Boca Raton, FL: CRC Press, Inc.

Schumacher, M. M. (1982), *Heavy Oil and Tar Sands Recovery and Upgrading International Technology*, Park Ridge, CA: Noyes Data Corporation.

Shell International Petroleum Company Ltd. (1983), *The Petroleum Handbook*, Amsterdam: Elsevier.

OPERATIONS RESEARCH

Hugh J. Miser, *Professor Emeritus of Industrial Engineering and Operations Research, University of Massachusetts at Amherst, Amherst, Massachusetts, U.S.A.*

	Introduction	101
1.	**Development of the Profession**	102
1.1	Professional Societies	102
1.2	Journals	103
1.3	Books	103
1.4	Education	103
2.	**Operations Research as a Science**	103
3.	**The Context**	104
4.	**Theories and Techniques**	105
4.1	Linear Programming	106
4.2	Descendant Programming Models	107
4.2.1	Convex Programming	107
4.2.2	Quadratic Programming	108
4.2.3	Flow through Networks	108
4.2.4	Integer Programming	108
4.2.5	Other Developments	108
4.3	Dynamic Programming	109
4.4	Queuing Theory	109
4.5	Simulation	110
4.6	Other Theories and Techniques	110
4.7	A Perspective on Standard OR Models and Techniques	111
5.	**Practice**	111
5.1	The Practice Processes	111
5.1.1	Formulation	112
5.1.2	Research	112
5.1.3	Evaluation and Presentation	112
5.1.4	Implementation and Evaluation	112
5.1.5	Variations	113
5.2	Systems Analysis	113
5.3	The Relation between Analyst and Client	113
5.4	Examples of Good Practice	114
5.4.1	Improving the Distribution of Blood to Hospitals	114
5.4.2	Protecting an Estuary from Flooding	115
6.	**Present State and Future Prospects**	116
	Acknowledgments	116
	Glossary	117
	Works Cited	117
	Further Reading	118

INTRODUCTION

Operations research is unusual among the sciences in that its beginnings—and even its name—can be traced unambiguously to an event in history: military operational trials being conducted on England's east coast in the summer of 1938.

In 1935, as part of Great Britain's effort to respond to the growing threat of German air power, scientists began an urgent series of developments and experiments aimed at detecting flying aircraft at a distance; by 1937, devices that came to be known as radar were in place and experiments were being conducted. Their technical success led in 1938 to large-scale operational trials involving both the radar sets and the fighter-plane defenses; they were aimed at testing the combined effectiveness in actual operations against an attacking force. In connection with this work, A. P. Rowe referred to its analysis as "operational research"—a term he is thought to have originated.

When the term crossed the Atlantic, it became "operations research," and the two synonymous usages remain today: operational research in the United Kingdom and parts of Europe, and operations research elsewhere; both are commonly abbreviated as OR. In the early 1950s, the term *management science* (MS) was introduced in the U.S. with a virtually synonymous meaning; thus, by the middle 1980s, the common subject was often referred to as OR/MS.

The 1938 partnership between the Royal

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Air Force operational people and the scientists, who not only worked with the radar but also analyzed the results of the operational trials, was successful both in characterizing the results of the trials and in showing how the operations could be improved. The next section sketches the growth of this seminal experience into the activities of an important science and an influential profession.

As operations research developed, its field of inquiry came to be recognized as the operations of systems consisting of nature, men, and machines, this last term including not only man's technical artifacts but also his laws, common practices, human behavior, and social structures and customs.

Attempts to generalize this description based on the concept of a system have not proved to be fruitful. Rather, the concept of an *action program* (Boothroyd, 1978; summarized in Miser and Quade, 1988, pp. 489–507) is more useful: a function, operation, activity, or response that is related to and given coherence by a human objective, need, or problem, together with the system of people, equipment, portion of nature, organizational elements, and management or social structure involved. OR can then be defined as the science of action programs, its work including both developing new knowledge and applying it to the problems of real action programs.

Two U.S. pioneer operations analysts (as OR workers engaged in practice are sometimes called), the physicist Philip L. Morse and the chemist George E. Kimball, observed early that "large bodies of men and equipment carrying out complex operations behave in an astonishingly regular manner, so that one can predict the outcome of such operations to a degree not foreseen by most natural scientists." Much of the success of the wartime work arose from the ability of the analysts to identify aspects of complex operations that exhibited such regularities.

1. DEVELOPMENT OF THE PROFESSION

The success of the small British team that was conducting operations research in 1938 set a pattern that was soon widely imitated; well before the end of World War II, most of the military services and their major commands in Britain, Canada, and the U.S. had established multidisciplinary teams to do such work (McCloskey, 1987). G. Majone has observed that the success of these teams did not arise from deep new theories of operational phenomena but rather the craft skills of the scientists applied to analyzing complex operations and assembling theories from readily available scientific tools.

However, one new theoretical development did emerge: search theory (Stone, 1989). It deals with an activity common in warfare: searching from a moving platform (such as a ship or an aircraft) for a fixed or moving target (such as a hostile submarine, ship, or aircraft). The goal is to detect, locate, and identify the target so that appropriate action can be taken. In a peacetime setting, this theory has been applied successfully to finding a sunken treasure ship off the U.S. Atlantic coast (Stone, 1992).

The wartime analysts emerged with the strong feeling that, if the man/machine/nature systems of warfare could be subjected to successful analysis, there should be many operations in peacetime society that could also benefit from such work. This feeling was soon converted to reality, as the community of analysts grew and the context of their inquiries spread widely, not only into varied operating situations, but also around the globe. Soon valuable new theories and applications were emerging from the work.

1.1 Professional Societies

Within the dozen years after the war, professional OR societies had been established in the United Kingdom (1948), the U.S. (1952), France (1956), India (1957), and Japan (1957). An International Federation of Operational Research Societies (IFORS) came into being in 1959; by 1973, 26 national OR societies were members, and by 1993 there were 41, with five other professional groups also adhering to the federation (including The Institute of Management Sciences, founded in the U.S. in 1954). The world-wide OR/MS community in 1993 was estimated to consist of over 35 000 persons.

The principal IFORS activities are these:

1. Conducting triennial international meetings for the exchange of information and discussion of international OR concerns.

The first meeting, at which IFORS was designed and agreed upon, occurred in 1957; the 13th meeting took place in 1993. For the first 12 of these meetings there has been a published volume of proceedings.

2. Publishing, since 1961, *International Abstracts in Operations Research (IAOR)*, a journal devoted to abstracts of the world OR literature. Three independently published volumes (Batchelor, 1959, 1962, 1963) contain abstracts of the literature for the years before 1961.
3. Publishing, beginning in 1994, *International Transactions in Operational Research*, a journal that will publish refereed papers from the 13th and later triennial meetings, as well as other essays of international interest.

1.2 Journals

The first two journals in the field were the *Operational Research Quarterly* (now the *Journal of the Operational Research Society*), founded in the United Kingdom in 1950, and *Operations Research*, founded in the U.S. in 1952. By 1973, there were 15 primary OR journals and 22 others of significant interest to OR workers. In 1993, *IAOR* covered the literature in 35 primary journals, 76 supplemental journals, and 67 specialist journals.

This literature and the professional activities that have prompted its growth since 1938 represent a tremendous development of interest in OR, both the validity of its theories and the effectiveness of their applications to real-world problems.

1.3 Books

Three books were written at the end of the war about the wartime work. Only one, however, was published promptly (Morse and Kimball, 1951); the other two (Johnson and Katcher, 1973; Waddington, 1973) were held up for many years by military classification concerns. Thus, the Morse and Kimball book was the first to offer a coherent account of a major body of OR work.

By the early 1950s, several anthologies of articles had appeared, but the first textbook emerging from an educational program did not appear until 1957 (Churchman *et al.*, 1957). However, the decade of the 1950s saw some 20 more specialized treatises appear. In the ensuing decades, this stream became a flood, with both textbooks and specialized treatises emerging in numbers almost too great to count.

Of the textbooks, Harvey M. Wagner's volume (Wagner, 1969) was widely regarded as setting a new standard and was awarded the Operations Research Society of America's Lanchester Prize. Recently, Hillier's (1990) has been widely used.

1.4 Education

In the early postwar years, a person interested in working in OR could learn about it only from the scattered literature and/or by joining a group engaged in such work, where the knowledge and craft skills of experienced analysts could be passed along. By the early 1950s, however, short courses of varying durations appeared in several countries, and by the end of the decade a few more formal university curricula had been established. By 1973, there were 53 such curricula in the U.S., and by 1993 the number in the U.S. and Canada had grown to over 100. Some included undergraduate as well as graduate training. Growth in other countries had also occurred on roughly the same scale.

Further, well before the 1990s, graduate training in these OR curricula had become recognized as an important qualification for entry into professional OR work, whether teaching, research, or practice.

2. OPERATIONS RESEARCH AS A SCIENCE

Following Ravetz (1971; for a summary focused on OR, see Miser and Quade, 1988, pp. 469–489), science may be described in general as "craft work operating on intellectually constructed objects," each object defining a class of phenomena. Scientific work is thus aimed at establishing new properties of the objects and verifying that they reflect the reality of the classes of phenomena that they represent.

This description has four implications:

1. The intellectual objects—which most scientists call theories but which OR workers usually call models—are created by the imagination, informed by earlier knowl-

edge of the phenomena and objects that have described aspects of them successfully, as well as innovative ideas or new evidence from reality. For example, there is a queuing theory that describes a wide variety of situations in which there are arrivals that demand service, which may be obtained immediately, or after a wait until earlier arrivals are served.

2. There is a continuing reference to the phenomena of reality, in OR's case those of action programs, or nature/man/machine systems. For queuing theory to be representative of some reality, such as the operations of a bank's customer services, the behavior of both the arrivals and the servers would have to be observed and characterized so that the theory could be based on these characterizations.
3. Scientific inquiry then becomes the search for new properties of the classes of phenomena both by manipulating the objects (that is, the theories or models), and by seeking new evidence from reality as a basis for revising them. In the banking example, after the queuing theory has been adapted to the observed behavior of the arrivals and servers, it can be manipulated to explore what might occur if the arrangements in the bank were changed, such as adding more servers, or improving their skills in various ways.
4. The new properties deduced from the objects—theories or models—must then be compared with the appropriate aspects of the phenomena of reality. In the banking example, the theory's predictions for changed conditions would be compared with an actual realization of these conditions in order to confirm the theory as a representation of reality, or to suggest how it might have to be revised.

Since all of science shares this generic description, it is important to note how the different sciences—such as physics, chemistry, biology, or operations research—are distinguished: not by their theories, models, or methods, many of which are widely shared, but by the portions of reality that they are undertaking to understand, explain, and solve problems in (Kemeny, 1959).

Within the framework established by this conception, it is possible to distinguish three classes of problems, depending on their goals: to paraphrase Ravetz, scientific problems, where the goal of the work is to establish new properties of the objects of inquiry, and the ultimate function is to achieve new knowledge; technical problems, where the function to be performed specifies the problem; and practical problems, where the goal of the task is to serve or achieve some human purpose and the problem is brought into being by recognizing a problem situation in which some aspect of human welfare could be improved.

In the queuing situation being used as an example, the scientific problem is to develop a model describing a queuing process, the technical problem is to achieve practical calculations from the model (usually in this case a rather complicated and difficult process), and the practical problem is to use the results to solve practical problems, such as those of improving a bank's service, or—to mention the purpose for which the theory was originally developed—to design an efficient and effective telephone exchange.

In some sciences, the professional groups that solve scientific problems are distinguished from those that work on technical and practical problems (although there is, of necessity, considerable overlap in these activities); for example, chemists solve scientific problems, while chemical engineers work primarily on technical and practical problems.

It is important to note that, in the case of operations research, there has not yet been such a split: operations research comprehends the work in all three classes of problems, with an attendant richness of interaction among its workers.

Thus, we can now describe operations research as the science of action programs, encompassing work on scientific, technical, and practical problems. It should be noted that, when the basic phenomena underlying a problem situation are not adequately understood, to solve a practical problem may involve solving scientific problems to obtain the needed models and technical problems to fashion the technical means needed to support the work toward ameliorating the practical problem.

3. THE CONTEXT

Within the broad contextual framework suggested by the term action program, it is possible to distinguish two major categories in which OR work takes place: processes common to a variety of action programs, and action programs distinguished by their practical context. Table 1 lists common processes to which OR work has contributed, and Table 2 shows areas of application in which significant OR work has been done. Since these tables list the standard terms of classification used by *International Abstracts in Operations Research*, work in any of these areas can be traced through this journal.

4. THEORIES AND TECHNIQUES

In contrast with the wartime years of operations research, when only one essentially new theory was devised, the postwar years have been a time of extraordinary fecundity for the development of models and solution techniques (often called algorithms in OR). While many have been borrowed from other fields and adapted for use in OR contexts, others have grown up largely in OR. Also, the uses of many have spread to other scientific fields.

Table 1. Common processes studied by operations research.

- Accidents
- Allocation: resources
- Behavior
- Bidding
- Communication
- Computers
  - Calculation
  - Data structure
  - Information
- Control
- Cybernetics
- Decision
  - Applications
  - Rules
  - Studies
- Demand
- Design
- Distribution
- Equipment
- Experiment
- Facilities
- Financial
- Forecasting: applications
- Gaming
- Growth
- Information
- Innovation
- Inspection
- Inventory
  - Order policies
  - Storage
- Investment
- Learning
- Location
- Maintenance, repair, and replacement
- Management
- Marketing
- Materials
- Measurement
- Organization
- Performance
- Personnel and manpower planning
- Planning
- Production
  - Flexible mfg. systems
  - Just-in-time
  - Material requirements planning
- Project mgmt.
- Quality and reliability
- Queues: applications
- Research
- Risk
- Scheduling
- Search
- Service
- Storage
- Supply
- Systems
- Testing
- Timetabling
- Values
- Vehicle routing and scheduling
- Work

Source: *International Abstracts in Operations Research* **44** (1993), p. 13.

Table 2. Areas of application studied by operations research.

- Advertising
- Agriculture and food
- Biology
- Commerce
- Communications
- Community OR
- Construction and architecture
- Cutting stock
- Ecology
- Economics
- Education
- Energy
- Engineering
- Finance and banking
- Forestry
- Geography and environment
- Government
- Health services
- Law and law enforcement
- Libraries
- Manufacturing industries
- Marketing
- Medicine
- Meteorology
- Military and defense
- Mineral industries
- Petroleum
- Politics
- Public service
- Recreation and tourism
- Science
- Social
- Space
- Sports
- Transportation
  - Air
  - General
  - Rail
  - Road
  - Water
- Urban affairs
- Water
- World affairs

Source: *International Abstracts in Operations Research* **44** (1993), p. 14.

Table 3 offers a list of the model and technique areas being actively pursued in OR in 1993. The terms shown are the standard classifications used in *International Abstracts in Operations Research*, so that, again, work in any of these areas can be traced through this journal.

As this table suggests, the most extensive borrowing from other fields is from probability theory (*q.v.*) [including especially stochastic processes (*q.v.*)] and statistics; indeed, surveys of OR practitioners show that, of the various techniques they employ, most are from probability and statistics. This is a natural consequence of the activities of exploring the behavior of action programs that not only yield numerical data—often in great volume—but also significant variations in the observations.

The reader will also note, however, bor-

Table 3. Model and technique areas OR is actively exploring.

- Adaptive processes
- Artificial intelligence
  - Decision support
  - Expert systems
- Calculus of variations
- Combinatorial analysis
- Computational analysis
  - Parallel computers
  - Personal computers
  - Supercomputers
- Control processes
- Cost-benefit analysis
- Decision theory
  - Multiple criteria
- Differential equations
- Fuzzy sets
- Game theory
- Gradient methods
- Graphs
- Heuristics
- Information theory
- Lagrange multipliers
- Markov processes
- Matrices
- Networks
  - Flow
  - Path
  - Scheduling
- Neural networks
- Numerical analysis
- Optimization
  - Simulated annealing
- Probability
- Programming
  - Assignment
  - Branch and bound
  - Convex
  - Critical path
  - Dynamic
  - Fractional
  - Geometric
  - Goal
  - Integer
  - Markov decision
  - Mathematical
  - Multiple criteria
  - Network
  - Nonlinear
  - Parametric
  - Probabilistic
  - Quadratic
  - Transportation
  - Traveling salesman
- Queues: theory
- Sets
- Simulation
  - Analysis
  - Applications
  - Languages and programs
- Statistics
  - Data envelopment analysis
  - Decision
  - Distributions
  - Empirical
  - Experiment
  - General
  - Inference
  - Multivariate regression
  - Sampling
- Stochastic processes
- Time series and forecasting methods

Source: *International Abstracts in Operations Research* **44** (1993), p. 14.

rowings from classical mathematics, such as the calculus of variations, combinatorial analysis (which has been spurred significantly by OR applications), differential equations, Lagrange multipliers, matrices (an essential ingredient in programming models), numerical analysis, optimization (a field whose development has been greatly spurred by OR work), probability (much used in view of the variabilities often encountered in the phenomena of action programs), and set theory.

In what follows, there is not space for even a cursory survey of all of the models and techniques commonly used in operations research. Thus, the ensuing subsections take up only a few of the most important that are regarded as especially associated with OR.

4.1 Linear Programming

The structure of the linear programming model, together with a procedure for obtaining solutions from it, was devised by George B. Dantzig in 1947 (Dantzig, 1963). While the original setting and motivation for it were largely economic, it soon became clear that it had many applications elsewhere, including business and industry, as well as in other sciences.

The simplest form of the linear programming model is this:

$$\text{Maximize} \quad c_1x_1 + \cdots + c_nx_n,$$

where the variables $x_1, \cdots, x_n$ satisfy the conditions

$$x_j \geq 0, \quad j = 1, \cdots, n,$$

$$a_{11}x_1 + \cdots + a_{1n}x_n \leq b_1,$$

$$\cdot \quad \cdot \quad \cdot$$

$$a_{m1}x_1 + \cdots + a_{mn}x_n \leq b_m,$$

and the c_j, b_i, and a_{ij} are constants.

One interpretation of this model that encompasses a wide variety of applications is this: There are n activities X_j, each producing x_j units of product, where each unit yields a value c_j. Each unit of the jth product consumes an amount a_{ij} of the ith resource, of which there is a maximum amount b_i that may not be exceeded by the sum of all of the activities. The problem then is to program a mix of activities that maximizes the total value (spoken of as the objective function) while remaining within the constraints on the resources.

The linearity implied by the statement of the model implies two basic assumptions underlying it: proportionality (that is, the resources used by an activity and their contribution to the objective function are proportional to the level of the activity) and additivity (that is, the combined use of resources by all of the activities is the sum of the individual uses). The third assumption is non-negativity (that is, negative quantities of activities are not possible).

The linear inequalities the variables x_j must satisfy correspond to the geometrical fact that a point $x = (x_1, \ldots, x_n)$ satisfying the model's conditions lies within or on the surface of a convex polyhedron (that is, one in which a line connecting any two points in or on the polyhedron also lies in or on it). If the objective function does not become arbitrarily large for points x on the surface of the polyhedron, then the maximum occurs at a vertex. Thus, the solution to a linear programming problem is one of finding the vertex, and therefore the point x, at which the objective function attains its maximum.

In addition to formulating the problem in this way, Dantzig devised a systematic and relatively economical procedure, called the *simplex method,* for discovering the point at which the objective function attains its maximum (Dantzig, 1963).

The model stated above has a dual:

$$\text{Minimize} \quad b_1y_1 + \cdots + b_my_m,$$

where the variables $y_1, \ldots, y_m$ satisfy the conditions

$$y_i \geq 0, \quad i = 1, \ldots, m,$$

$$a_{1j}y_1 + \cdots + a_{mj}y_m \geq c_j, \quad j = 1, \ldots, n.$$

If there is a maximum for the first problem there is a minimum for its dual and these two numbers are equal. This duality theorem is essentially equivalent to the minimax theorem of game theory (*q.v.*).

With the programming problem now expressed as minimizing an objective function, other applications come to mind. For example, there is a classical diet problem for mixing livestock and poultry feed that can be formulated in this way: If a_{ij} denotes the number of units of the jth nutrient in one unit of the ith food, c_j is the minimum amount of the jth nutrient needed for satisfactory health, b_i is the unit price of the ith food, and y_i is the number of units of this food to be purchased, then find the point $(y_1, \ldots, y_m)$ that minimizes the objective function and produces a satisfactorily mixed diet at minimum cost.

Since manual calculations by the simplex method become very laborious even for problems of modest size, computer programs for this method were introduced as soon as computers became powerful enough to handle them. By now, both the methods of solution and the computer programs for them have evolved to the point where problems of fairly large size can be handled on personal computers and problems with thousands of variables and constraints can be dealt with successfully at modest cost on large machines.

The resulting tools have greatly expanded the possible applications in many fields, including agriculture, business, government, manufacturing and service industries, transportation, and economics, and OR workers have achieved thousands of such applications. The theoretical basis for such work has also expanded similarly. By 1993, hundreds of papers were appearing yearly on both theory extensions and their important applications.

4.2 Descendant Programming Models

The value of a theory or model and its associated solution technique can be judged by the vigor with which they are extended either by theoretical developments or a widened coverage of the phenomena of reality. Thus, the ensuing value can be judged in part, first, by the scope of the phenomena that the model helps to understand and, second, by the importance of the problems the technique can deal with; but this judgment must also include, third, their fecundity in prompting the development of descendant models and techniques, that is, ones suggested by the basic model and its central technique but demanding important new approaches. The previous subsection suggests that linear programming has high value on the basis of the first two of these criteria; this subsection describes some descendant theories that support its value based on the third. In each case, there are also descendant solution techniques.

4.2.1 Convex Programming Let x be a set $(x_1, \ldots, x_n)$ of points in n-dimensional space. Then, a function $f(x)$ is a convex function if it is defined over a set of points x in a convex set and if the set of points $z \geq f(x)$ is convex in the $(n + 1)$-dimensional space (x,z).

Early investigators were able to amend the simplex method to apply to programs where the constraints and the objective function are convex.

4.2.2 Quadratic Programming A quadratic form in n variables is the sum of terms $a_{ij}x_ix_j$, where both i and j take on values from 1 to n. Such a form is positive semidefinite if its value is greater than or equal to zero for all values of the x's. By suitable transformations of the variables, it is possible to set such a form equal to the sum of squares of the new transform variables.

The transformed quadratic form can now be recognized as a convex function, which allows programs with positive semidefinite objective functions to be solved for the minimum value of this function by adjusting the methods for convex functions. A similar statement can be made for programs seeking the maximum of the negative semidefinite quadratic objective function. These special cases allow some improvements in the efficiency of the solution procedure.

4.2.3 Flow through Networks Consider a network connecting two nodes, a source and an destination, by means of a series of intermediate arcs that meet at intermediate nodes. Flows can take place in the arcs and through the intermediate nodes. Each arc has two flow capacities associated with it, one for each direction. The problem is to find the maximal flow that the network permits from the source to the destination. Efficient procedures for solving this problem are available.

It has also been generalized—for example, for a network with many sources and destinations and stipulated amounts at both, to minimize the costs of the flows that will be required for the flows from the sources to achieve the stipulated amounts at the destinations at minimum transportation cost—and suitable algorithms have been developed. This model and its associated solution algorithm have been applied, for example, to the problem of assigning children to schools.

4.2.4 Integer Programming It was recognized early that there were many attractive problems for which the variables in a program are constrained to be integers. Unfortunately, simply approaching these problems with continuous variables and their solution techniques does not work satisfactorily, so that special procedures must be devised for them.

The pioneer work in this regard was done by R. E. Gomory in 1958 (Gomory, 1991). Since then, tremendous advances have been made in developing solution algorithms, and they have been applied to numerous important problems, some quite large.

The well-known traveling salesman problem can also be formulated as an integer program. This problem arises when a salesman desires to keep his total travel distance or time to a minimum while making one visit to each of a number of cities and returning to his starting point at the end of his trip. Similar problems can occur in a wide variety of situations.

Further, this line of research has generated its own descendant models and their applications, including those for so-called mixed-integer problems (where some of the variables are continuous but others are restricted to integral values).

In cases where it is possible *a priori* to state upper and lower bounds for the maximum value, there is a branch-and-bound procedure that has had many applications. It has also been much developed since it was first set forth.

4.2.5 Other Developments As Table 3 indicates, there are a number of other descendant programming models and solution techniques. While there is not space here to describe them all, a few may be mentioned.

It can happen that a programming problem is stated such that there is no vector x that satisfies all of the constraints. If one can establish penalty costs for violating one or more of the constraints, then goal programming finds a solution that violates the constraints in such a way as to minimize the total penalty cost.

The programming structure has also been used to formulate models in which the functions are nonlinear (as, for example, in convex and quadratic programming, as mentioned earlier), and various special cases of this situation have yielded reasonably economical solution techniques.

So far, the programming formulations have treated the coefficients in the constraints as completely known. In many practical situations, however, these coefficients may not be known with precision, or may in fact be subject to random variations. Chance-constrained programming considers such cases under a few simplifying assumptions. Stochastic programming deals with more general cases.

For a modern overview of linear programming and its descendant theories, see Nemhauser *et al.* (1989).

4.3 Dynamic Programming

The word programming in the name of this subject might suggest that it is another theory that descended from linear programming, but this is not the case. Rather, dynamic programming is concerned with a sequence of decisions that must be made over time.

Unlike linear programming, which has specific models that can be solved by various techniques, dynamic programming deals with a structure and an analytic approach that can be applied to a variety of models.

The dynamic programming approach deals with problems that can be put into this structure:

1. The decision variables and their associated constraints are grouped according to stages to be considered sequentially.
2. For a given stage, the only information about previous stages relevant to selecting optimal values for the current decision variables is summarized by a quantity called the state variable.
3. For a given stage in its current state, the current decision has a forecastable influence on the state at the next stage.
4. The optimality of the current decision is judged in terms of its forecasted impact on both the present stage and all later stages.

There is a canonical form for the central recursion relation that can be developed in this way: Consider an N-stage process that is n stages from completion. Let s stand for a state of the system, S_n for the set of all possible states at the nth stage, d_n for the decision made at the nth stage, and $D_n(s)$ for the set of all the feasible values of d_n for the system in the nth state. Also for the nth state, let $R_n(s,d_n)$ be the immediate value of decision d_n and $T_n(s,d_n)$ the transformed state of the system at stage $n - 1$. If $f_n(s)$ is the value of the optimal decision at the nth stage, then we can say that it depends not only on the state the process is in at the nth stage, but also on the number of stages left to be completed. Then the dynamic programming recursion states that $f_n(s)$ is the optimum (that is, either the maximum or minimum) over all $d_n(s)$ in $D_n(s)$ of $R_n(s,d_n) + f_{n-1}(T_n(s,d_n))$ for every s in S_n.

Perhaps the most striking aspect of this procedure is that it begins with the last stage and works toward the beginning. Its advantage over the complete enumeration of all possibilities at all of the stages is that it optimizes at each stage over a limited number of decision variables.

Here are some problems that can be formulated with a dynamic programming structure: inventory rules indicating when to replenish an item and by what amount; capital budgeting for allocating scarce resources to new ventures; scheduling for routine and major overhauls of complex machinery; and formulating a long-range strategy for replacing depreciating assets. For further examples, see the classic book by Bellman and Dreyfus (1962).

4.4 Queuing Theory

Queuing is one of the commonest processes in developed societies: telephone calls arrive at an exchange, customers arrive at lunch counter, vehicles arrive at a car wash, aircraft arrive at an airport, and so on, each demanding service; when others are ahead of them, they wait in a queue until served. The theory of this process began early in this century in connection with telephony, but did not move significantly out of this context until it was adopted in the early postwar years by operations research, after which it experienced an explosive growth.

The simple examples just mentioned suggest the three aspects of a queue that are important for developing a theory of it: the arrival process, the service mechanism, and the queue discipline.

Consider the simplest case, in which the arrivals are random over time, and the service times are exponentially distributed. If λ and μ are the average arrival and service rates, respectively, then the average traffic density ρ is λ/μ. For the number N of customers in the system, the queue length Q, the waiting time W, and the total time D the customer is in the system, if we denote the probability that $Q = j$ ($j = 0, 1, \ldots$) by $Pr[Q = j]$, the average or expected number of customers in the system by $E[N]$, the expected length of the queue by $E[Q]$, the expected time a customer waits until service begins by $E[W]$, the expected total delay for a customer (waiting time plus service time) by $E[D]$, and the variance of the queue length by $Var[Q]$, then the following relations can be derived from the theory:

$$Pr[Q = j] = (1 - \rho)\,\rho^j, \quad E[N] = \rho/(1 - \rho),$$

$$E[Q] = \rho^2/(1 - \rho), \quad E[W] = \rho/(1 - \rho)\mu,$$

$$E[D] = 1/(1 - \rho)\,\mu, \quad Var[Q] = \rho/(1 - \rho)^2.$$

There are a tremendous number of other queuing processes with other assumptions about the arrival processes, the service-time distributions, and the queue disciplines, but, unfortunately, only a few yield simple formulas like those just set forth. Rather, more complicated expressions must be dealt with, approximations employed, or the queuing process simulated on a computer (Gross and Harris, 1985).

4.5 Simulation

Any theory or model is, of course, a representation—that is, a simulation of a sort—of some aspect of the phenomenon that it purports to describe. However, the term simulation here means something different: instead of representing a process by a model, it is a somewhat simplified ideal process that imitates selected aspects of the one from reality.

Consider an operational process where it is possible to subdivide it into a series of stages or steps, such as the queuing process just discussed: An arriving customer either takes a place at the end of the queue if there is one, or, if there is no queue, proceeds to obtain the desired service; whenever a service is completed, the queue, if there is one, is reduced by one waiting customer. This example can easily be simulated by hand, provided the analyst has been able first to establish a series of random arrivals and exponential service times to draw from as the simulation proceeds. However, it can also be programmed on a computer.

Similarly, many processes, such as for highway toll-collection operations, air-traffic arrivals at airports, banking service operations, or telephone ordering operations, can be simulated, either to solve problems or to see how revised procedures might work.

It is usual to represent the chosen aspects of the operations by a computer program that imitates them; OR workers usually call such a representation a simulation, but in other fields it is also called a Monte Carlo process. This procedure has been used quite often by OR workers when no simplifying theories have existed to make the analysis more compact and efficient. They do this in spite of the procedure's shortcomings: that specific assumptions about the situation have to be made instead of more general ones that a theory could incorporate, that a large number of trials (or "runs") may be needed to achieve a stable set of results, and that complicated situations may call for very large computer programs whose results may be difficult to interpret and apply.

Indeed, there is a standing—if informal—rule of action among OR workers that simulation is to be shunned if any analytic models can be made to represent the situation adequately, and that therefore simulation should be a last resort. Nevertheless, OR workers have found many situations in which simulation has been helpful to decision makers, and there has been substantial development in this field. Whereas in the early days of OR it was necessary to construct computer simulations from the basic elements and principles, by 1993 it was far more common for an analyst to adapt a commercially available simulation package. The journal *OR/MS Today* (December 1993, pp. 62–78) describes the essential properties of 55 such packages (see also Law and Kelton, 1991).

Simulation has an advantage in that it can portray to a user or client the steps in the process, thus conveying an important sense of reality. And it is a technique that can be used when analytic models cannot cover the problem situation adequately. Historically, successful simulations in OR have run all the way from relatively simple ones conducted by hand on a large sheet of paper to very complicated ones managed on large computers.

4.6. Other Theories and Techniques

As Table 3 indicates, there are many other theories and techniques that occupy a place in modern OR work; STOCHASTIC PROCESSES (for a modern overview, see also Heyman and Sobel, 1990), GAME THEORY, ARTIFICIAL INTEL-

LIGENCE, COMPUTER DATA BASES, and COMPUTER GRAPHICS are discussed elsewhere in this Encyclopedia. Space in this introductory overview does not permit others to be described, but they can be traced easily in the literature through the references and *International Abstracts in Operations Research*. For a survey of computing from an OR perspective, see Coffman *et al.* (1992).

4.7 A Perspective on Standard OR Models and Techniques

Over the last four decades, the community of OR workers primarily interested in scientific problems has not only developed many new models, but also algorithms (that is, solution techniques) for them where they were appropriate. This growing community of OR scientists has exhibited great skill and inventiveness in this process. Thus, such tools exist in many of the categories listed in Table 3 (as hinted at in the brief introductory descriptions above).

Many of these models and algorithms can be thought of as standard in the sense that they are reference points for OR workers looking at the phenomena of action programs with a view to constructing models to represent them and algorithms to obtain specific results. Thus, they can be considered to be an OR analyst's tool kit. And, of course, they form the substance of advanced treatises and courses in OR theory.

On the other hand, OR workers facing technical or practical problems can relatively seldom simply apply one of these standard tools; rather, the problem structure and its elements usually demand that the tools be adapted to the new situation. Indeed, how to be effective in adapting the tools in the kit of standard models and algorithms can be considered one of the essential craft skills of the OR practitioner.

It has been pointed out that OR comprehends work on all three classes of problems: scientific, technical, and practical. The close association of work on all three classes of problems has an important benefit for operations research: Not only do the model and algorithmic tools help practitioners quite directly, but also the process of looking at new situations and their phenomena has a feedback benefit to OR's theorists, as it often brings new modeling and algorithmic possibilities to them that might otherwise be overlooked.

5. PRACTICE

An OR analyst working to solve a practical problem is said to be involved in practice. Such work may entail, as has frequently been the case, work on technical problems when instrumentalities need to be designed to achieve desired goals, or on scientific problems when the underlying phenomena are not yet understood well enough. In any case, regardless of what subproblems the work may require solving, the overall goal of practicality qualifies the whole as being practice.

An OR analyst is often consulted when someone with a suitable responsibility or interest in an action program discerns a problem situation that needs improvement. While this person—and perhaps others—may have diagnosed the problem and may even have a belief about the course to take toward its possible solution, it is commonly the case that the forces actually at the core of the problem situation lie buried deeply enough to make such an early diagnosis questionable, and the preconceived fix inappropriate. Thus, typically it is best for the analyst—or the team of analysts if the problem situation is large and complex—to approach it with an open mind and explore it thoroughly before deducing its essential properties and using them to devise a scheme for ameliorating its undesirable aspects.

5.1 The Practice Processes

While each situation in practice may properly be regarded as unique, it is nevertheless possible to describe a central analysis process that contains elements central to most—if not all—of practice, as follows.

Figure 1 offers a synoptic view of these elements. It proceeds from the general recognition that a problem situation exists to the implementation of some policy or course of action and the evaluation of its effects. Since each situation has its own unique properties, few OR practice engagements follow such a procedure exactly—and sometimes not even approximately. However, it is common for at least some of these elements to occur during the work. Thus, Fig. 1 is not intended to be

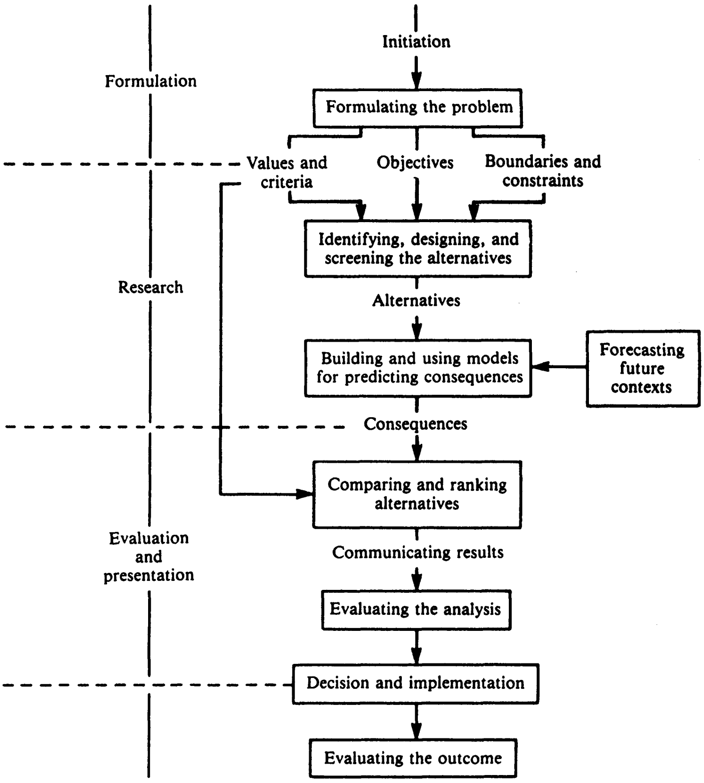
Formulation
Research
Evaluation and presentation
Initiation
Formulating the problem
Values and criteria
Objectives
Boundaries and constraints
Identifying, designing, and screening the alternatives
Alternatives
Building and using models for predicting consequences
Forecasting future contexts
Consequences
Comparing and ranking alternatives
Communicating results
Evaluating the analysis
Decision and implementation
Evaluating the outcome

FIG. 1. Important elements in an operations research practice engagement that runs from problem formulation through research and implementation to evaluating the outcome. [From Miser and Quade (1988), p. 23; reproduced by permission.]

prescriptive; rather, it offers a suggestive structure as the basis for discussion.

5.1.1 Formulation The work begins with a thorough exploration of the problem situation. It is vital for the client and people in the action program to cooperate in this work. The purpose is to formulate the problem to address. Once this is done and the client has agreed with the analysis team on the diagnosis, it is possible to plan the work to be done.

This early work also identifies the values and criteria that should inform the choice of what eventually will be done to ameliorate the client's concerns, sets up the objectives to be sought by the solution, and agrees with the client on the boundaries and constraints that must be observed in devising it.

5.1.2 Research This stage extends the information- and data-gathering that began in the formulation stage. The findings that emerge from processing these results allow the analysis team to identify, design, and screen possible alternative courses of action that may help with the problem. Against this background, the analysis team can build models capable of deducing the consequences of adopting each of the alternatives chosen for further investigation within the framework of possible future conditions.

5.1.3 Evaluation and Presentation With estimates of the consequences in hand, the analysts may compare—and possibly rank—the alternatives against the criteria chosen earlier in the analysis, plus any new ones that may have emerged during the work. These findings must then be presented to the client and other parties at interest in a way that enables them not only to appreciate the results but also have at least a broad overview of the logic and its factual basis that produced them. These understandings, combined with the client's wider and more general knowledge of the situation, may then enable him or her to adopt a suitable policy or course of action—and, when necessary, persuade the various parties at interest to accept it.

5.1.4 Implementation and Evaluation The client, and not the analysts, must decide on what to do and how to carry it out effectively, and the client must see to the implementation and take full responsibility for it. Nevertheless, experience shows that it is very important for the analysis team, or at least some analysts who understand and appreciate what was done, to work cooperatively throughout the implementation stage.

Later, if an evaluation of the efficacy of the results is undertaken, it is also desirable for someone familiar with the input analysis to be involved.

5.1.5 Variations While it is possible to specify a core diagram of the principal elements of OR practice, it must be admitted immediately that few, if any, such engagements follow this outline exactly. Rather, since each problem situation is different, the analysis must be adapted to it. Thus, in studying a series of cases, one finds variations such as these:

1. Instead of proceeding linearly from the top to the bottom of the diagram, the work cycles from intermediate stages back to earlier ones as the research program brings new insights and fresh intermediate results that may prompt reconsidering the beginning foundations of the work.
2. Some work may be aimed more at fleshing out the client's understanding of the situation than prompting a significant change, so it may stop at one of the intermediate stages.
3. The relative effort expended in the various stages may vary tremendously from case to case: one case may have to expend its major effort in just the information- and data-gathering stage, after which what needs to be done may be fairly apparent without much further analysis. Another case may proceed fairly expeditiously through the outline of Fig. 1 and then have a very long and complicated period of work to achieve what may appear to an outsider to be the implementation of a fairly simple set of proposals.
4. In some cases, an intermediate stage may dominate the work, as a result of such factors as technical difficulty in devising proper models, major uncertainties in forecasting future conditions, complexities of the underlying situation, and so on.

In any case, the procedure outlined here as the basis for discussion must be regarded as one that has stitched together the key elements that may enter OR practice to varying extents depending on the peculiarities of the situation being analyzed.

5.2 Systems Analysis

There is a branch of operations research devoted to problems of such large scale that they almost always involve major efforts by teams of analysts drawn from a variety of disciplines in addition to operations research. In the mid-1950s, it came to be called systems analysis (in this meaning not to be confused with the same term used in the computer community). While there is a separate literature about such work (Miser and Quade, 1985, 1988; Miser, 1995), its essential properties are sufficiently close to those of OR as it is usually understood that inferences back and forth between the two kinds of work are quite appropriate.

5.3 The Relation between Analyst and Client

Based on close study in a variety of fields of the ideal relation that should exist between professional practitioner and client, Schön (1983; for a relevant summary, see Miser and Quade, 1988, pp. 508–520) advocates a "reflective contract" that works in this way: "... in a reflective contract between a practitioner and client, the client does not agree to accept the practitioner's authority but to suspend disbelief in it. He agrees to join the practitioner in inquiring into the situation for which the client seeks help; to try to understand what he is experiencing and to make that understanding accessible to the practitioner; to confront the practitioner when he does not understand or agree; to test the practitioner's competence by observing his effectiveness and to make public his questions over what should be counted as effectiveness; to pay for services rendered and to appreciate competence demonstrated. The practitioner agrees to deliver competent performance to the limits of his capacity; to help the client understand the meaning of the professional's advice and the rationale for his actions, while at the same time he tries to learn the meanings his actions have for the client; and to reflect on his own tacit understanding when he needs to do so in order to play his part in fulfilling the contract."

Not all OR practice engagements realize this ideal, of course, but successful practitioners achieve a suitable form of it in their work. In any case, the practical arrangements aimed at achieving the ideal of the reflective contract must, of necessity, be evolved in the light of the circumstances and people peculiar to each engagement. It is common, as one way of supporting this concept in practice, for a member of the client's staff to work with the analysis team throughout the OR engagement. And, of course, it is wise for the client and the analysis team to have frequent interactions as the work proceeds.

5.4 Examples of Good Practice

Fortunately for someone interested in seeking out cases of good practice, there is a convenient central source: Since 1973 the journal *Interfaces* has specialized in publishing papers on such work. Assad *et al.* (1992) offer a selection of 19 cases from this source, accompanied by discriminating discussions. Miser (1995) centers his attention on cases, and Miser and Quade (1985, 1988) accompany their general discussions of analysis principles with illustrative cases.

Two areas often associated with operations research are discussed elsewhere in this Encyclopedia: MANUFACTURING ENGINEERING, and COMPUTER-AIDED MANUFACTURING. Three areas of application had by 1994 been given overviews in handbook volumes: logistics of inventory and production (Graves *et al.*, 1993), marketing (Eliashberg and Lilien, 1993), and problems in the public sector (Pollock *et al.*, 1994).

Each year there is a muscular and carefully administered competition in the U.S. for the best papers on practice. Since 1975, *Interfaces* has published the finalist papers in this competition; since there are five or more finalists in each competition, this competition alone offers a large number of papers and a wide variety of work. For example, in recent years there have been papers on computer-aided train dispatching, planning the management of the freshwater resources of the Netherlands, how OR is used in a large petroleum company, the size and deployment of a sales force, scheduling police patrols in a major city, assigning operators in a large urban transportation system, analyzing the basis for U.S. Medicare payments to hospitals, improving passenger traffic management at a major international airline, and considering the effectiveness of a needle-exchange system as a device for reducing the incidence of AIDS among drug addicts in American cities. Two contrasting cases, the first of which won this prize, are described briefly below.

5.4.1 Improving the Distribution of Blood to Hospitals Human blood, a living tissue of unique medical value, is a perishable product. In the U.S. when this study was being done, it had a fixed lifetime during which it could legally be used for transfusion to a patient of the proper type and after which it had to be discarded; this lifetime, since extended, was 21 days.

The blood is collected in units of one pint (approximately a half liter) from volunteer donors at various collection sites such as a regional blood center (RBC). After a series of typing and screening tests, it is shipped to a hospital blood bank (HBB) in the region of the RBC. The units are then stored at the HBB, where they are available to satisfy the random daily demands for transfusions to patients. Since in general not all units demanded and assigned to a patient are used, a unit can be assigned several times during its lifetime before it is transfused or outdated and discarded.

To manage the blood resources of a region is a difficult task. The blood distribution problem is complex, because of the blood's perishability, the uncertainties of its availability to the RBC, and the random nature of the demands and usages at the HBB's. Superimposed on these complexities are the large variations in the sizes of the hospitals—and therefore their HBB's—to be supplied, in the relative occurrences of the eight different blood groups, and in the mixes of whole blood and red blood cells called for when this study was done (since then the development of other blood products has complicated matters further). Finally, on the one hand, there is the imperative of having blood available when and where needed (although elective surgery can be postponed because of a blood shortage, this act incurs additional costs for all concerned), while on the other hand there is the desire to operate efficiently and economically, and thus to avoid wasting the precious individual gifts of blood.

The two most common performance measures for an HBB are the shortage rate (that is, the proportion of days when supplementary unscheduled deliveries from the RBC have to be made to satisfy the hospital's demand), and the outdate rate (that is, the proportion of the hospital's blood supply that is discarded because of its becoming outdated). Suitable calculations convert these measures for individual hospitals to similar ones for a region.

In 1973, two analysts in the U.S., one a blood expert and the other an OR analyst, reported a study that went most of the way toward solving this problem of managing blood supplies efficiently for the Greater New York Blood Program, which supplies over two hundred hospitals. After studying the patterns of demand and supply, they were able to characterize the situation with relatively simple models that were variants of standard ones adapted to suit this new purpose. On the basis of these models they were able to devise a decision support system for an RBC that addressed these questions: What are the minimum achievable outdate and shortage rates that can be set for the region? What distribution policy will achieve these targets? What levels of supply are needed to achieve alternative targets?

The blood management system they devised was characterized by centralized management at the RBC, prescheduled deliveries to the HBB's supplemented by emergency deliveries when needed, and a distribution policy according to which some blood was rotated among the hospitals. The operation of this system is based on a programming model whose objective was to optimize the allocation of the regional blood resources while observing policy constraints.

The final step was to implement this programmed blood distribution system (PBDS) in a trial region on Long Island near New York City. The success of this trial led to the system being spread throughout the region served by the Greater New York Blood Program.

Before PBDS came into effect on Long Island, the RBC made an average of 7.8 blood deliveries per week to each hospital in the region, all unscheduled, and the outdate rate was 0.20. After PBDS was implemented, the average number of deliveries dropped to 4.2, of which only 1.4 were unscheduled, and the average outdate rate fell to 0.04, which appears to be about the lowest possible for this situation (some wastage being an inevitable consequence of the random demand and limited supply). These management improvements represented substantial cost savings.

The analysis designed the PBDS so that it could easily be adapted to other regions, and some adaptations were made.

For the prize-winning account, see Brodheim and Prastacos (1979); for an extended summary accompanied by discussions of its lessons for practice, see Miser and Quade (1985, pp. 68–79 and *passim*).

While this analysis was conducted by a team of two analysts, one with an expertise in blood properties and detailed management and the other with a background in OR, the next example involved a large interdisciplinary team led by an OR expert and including representatives of many scientific specialities.

5.4.2 Protecting an Estuary from Flooding In 1953 a severe storm in the North Sea flooded much of the delta region of the Netherlands, killing several thousand people. Determined not to allow this to happen again, the government started a program to increase the protection from floods by constructing a new system of dams and dikes. By the mid 1970s this system was complete except for the protection of the largest estuary, the Oosterschelde. Three alternatives for this last segment were under consideration: building an impermeable dam to close off the estuary from the sea, building a flow-through dam with gates that could be closed during a storm, and building large new dikes around the estuary.

In 1975, the Netherlands Rijkswaterstaat (the government agency responsible for water control and public works) and The Rand Corporation of Santa Monica, California (a research institution with substantial experience in systems analysis), began a joint project with a view to helping decide what should be done. It set out to determine the major consequences that would follow from implementing each of the three alternatives for protecting the estuary.

These consequences were grouped into eight categories: financial costs; security from flooding; effects on jobs and profits in the fishing industry; changes in recreational opportunities; savings to carriers and customers of the inland shipping industry; changes in production, jobs, and imports for the 35 industrial sectors of the national economy; changes of the populations of the species that make up the ecology of the region; and, finally, the social impacts—displacement of households and activities and significant effects on the regional economies. A major uncertainty was the severity and frequency of the super-storms that make the provisions for protection necessary.

By intention, the study did not conclude by recommending a particular alternative. Rather, it clarified the issues by comparing, in a common framework, the many different impacts of the three alternatives. The choice among them was left to the political process, where the responsibility properly resides. (To go further by trying to rank the alternatives was inappropriate—as it usually is in a complex study of policy.)

There was no dominant alternative; rather, each was found to have a major disadvantage that might be considered serious enough to render it politically unacceptable. The storm-surge barrier alternative (that is, the flow-through dam) was by far the most costly; the impermeable dam was the worst for the ecology; and the open case, with new dikes around the estuary, lacked security.

The Rijkswaterstaat supplemented this work with several special studies of its own and submitted a report to the Cabinet recommending the more costly storm-surge barrier plan to the Parliament. It was adopted in 1976, and construction was completed about a decade later.

This study was not only a major effort, involving dozens of analysts and specialties, it also developed some novel ideas about how to present multifaceted results to policy makers with divergent backgrounds and concerns. Thus, in addition to helping the Dutch government reach a decision about an overwhelmingly important problem, it broke new ground on how to carry through and present the results of such a complicated analysis. A summary volume (Goeller *et al.*, 1977) is supported by a variety of volumes on specialized contributing analyses; for an overview and comments on the professional implications of the work, see Miser and Quade (1985, pp. 89–109 and *passim*).

6. PRESENT STATE AND FUTURE PROSPECTS

There is a tendency in mature knowledge-based professions for the academic theorists to have acquired dominance over the professional view of quality, relevance, indices of prestige, and rewards (Abbott, 1988). While this force is visible in operations research, it has yet to achieve the dominance shown elsewhere. Thus, OR workers devoted to successful practice continue to be honored along with their colleagues who center their attention on solving scientific problems. The resulting mix of interests and the communication among them still can be seen as one of the OR profession's great strengths, on which much of its current success has been built.

On the other hand, the creation of a profession with an identified body of knowledge and clearly understood foci of practice, while it has produced the strengths leading to the profession's success, has also tended to inhibit its willingness to look outward to entirely new contexts in which many of society's new and pressing problems are found. Major problems seldom occur in unidisciplinary form; rather, their successful analysis demands multidisciplinary inquiry. While OR has traditionally either led or played an important part in such teams, this form of effort has not grown commensurate with the needs that exist for this form of work.

In spite of the tremendous success of the profession, both in developing theoretical tools and applying them to important problems, it has yet to develop a widely shared underlying philosophy for its work, or a well-based epistemology of practice.

Thus, the vitality of the profession in the future depends to an important extent on maintaining the balance between theory and practice, reaching out to more important and diverse problems as part of the team efforts that will be needed to address them, and developing widely shared underlying concepts as the basis for the profession's work. Some efforts to meet these challenges are being made, but it is not yet clear what their outcomes will be.

ACKNOWLEDGMENTS

The discussion of practice processes in Sec. 5.1 is drawn from an article written for Gass and Harris (1995), and is reproduced by permission of Kluwer Academic Publishers. Figure 1 is copyright by Elsevier Science Publishing Co. of New York and is reproduced by permission. Tables 1, 2, and 3 are drawn from *International Abstracts in Operations Research* **44** (1993), pp. 13–14; they are reproduced by permission of the International Federation of Operational Research Societies. The case summaries in Secs. 5.4.1 and 5.4.2 are drawn by permission from Miser and Quade (1985).

GLOSSARY

Action Program: A function, operation, activity, or response that is related to and given coherence by a human objective, need, or problem, together with the system of people, equipment, portion of nature, organizational elements, and management or social structure involved.

Algorithm: A set of rules or a procedure for solving a problem; in operations research the term usually refers to a method for obtaining numerical solutions from a model or theory.

Convex Polyhedron or Set: A set (or polyhedron) is convex if any two points in it or on its boundary can be connected by a straight line that lies in it or on its boundary.

Descendant Model or Theory: A variant or extension of an earlier model (or theory) that extends it to a new class of phenomena.

IFORS: The acronym for the International Federation of Operational Research Societies.

Linear Programming: A model, together with its solution algorithm, that seeks the maximum (or minimum) of a linear function of n variables where these variables are positive and subject to linear constraints.

Model: The term used by operations research workers for a theory.

Operations Research: The science devoted to understanding the behavior of operating systems, including nature, men, and machines, this last term including not only the artifacts of modern technology but also laws, common practices, human behavior, and social structures and customs. More briefly, operations research is the science of action programs. Its activities include work on scientific, technical, and practical problems.

OR: The acronym for operations research.

Practical Problem: A problem brought into being by recognizing a problem situation in which some aspect of human welfare could be improved and where the goal is to serve or achieve some human purpose.

Scientific Problem: A problem whose solution has the goal of establishing the properties of the object of inquiry, the ultimate function being to achieve new knowledge.

Systems Analysis: The portion of operations research devoted to addressing large-scale problems, usually with interdisciplinary teams.

Technical Problem: A problem specified by the function to be performed.

Theory: An intellectual construct representing a class of phenomena.

Works Cited

Abbott, A. (1988), *The System of Professions: An Essay on the Division of Labor*, Chicago: University of Chicago Press.

Assad, A. A., Wasil, E. A., Lilien, G. L. (Eds.) (1992), *Excellence in Management Science Practice: A Readings Book*, Englewood Cliffs, NJ: Prentice-Hall.

Batchelor, J. H. (1959), *Operations Research: An Annotated Bibliography*, St. Louis: St. Louis University Press. Covers the literature through 1957.

Batchelor, J. H. (1962), *Operations Research: An Annotated Bibliography*, Vol. 2, St. Louis: St. Louis University Press. Covers the literature for 1958 and 1959.

Batchelor, J. H. (1963), *Operations Research: An Annotated Bibliography*, Vol. 3, St. Louis: St. Louis University Press. Covers the 1960 literature.

Bellman, R. E., Dreyfus, S. E. (1962), *Applied Dynamic Programming*, Princeton: Princeton University Press.

Brodheim, E., Prastacos, G. (1979), *Interfaces* **9** (1), 3–20.

Boothroyd, H. (1978), *Articulate Intervention*, London: Taylor and Francis.

Churchman, C. W., Ackoff, R. L., Arnoff, E. L. (1957), *Introduction to Operations Research*, New York: Wiley.

Coffman, E. G., Jr., Lenstra, J. K., Rinooy Kan, A. H. G. (Eds.) (1992), *Computing*, Amsterdam: North-Holland.

Dantzig, G. B. (1963), *Linear Programming and Extensions*, Princeton: Princeton University Press.

Eliashberg, J., Lilien, G. L. (Eds.) (1993), *Marketing*, Amsterdam: North-Holland.

Goeller, B. F., Abrahamse, A. F., Bigelow, J. H., Bolten, J. G., de Ferranti, D. M., DeHaven, J. C., Kirkwood, T. F., Petruschell, R. L. (1977), *Protecting an Estuary from Floods—A Policy Analysis of* *the Oosterschelde: Vol. I*, Summary Report, R-2121/1-NETH, Santa Monica, CA: The Rand Corporation.

Gomory, R. E. (1991), in: J. K. Lenstra, A. H. G. Rinooy Kan, A. Schrijver (Eds.), *History of Mathematical Programming*, Amsterdam: North-Holland, pp. 55–61.

Graves, S., Rinooy Kan, A. H. G., Zipkin, P. (Eds.) (1993), *Logistics of Production and Inventory*, Amsterdam: North-Holland.

Gross, D., Harris, C. M. (1985), *Fundamentals of Queuing Theory*, 2nd ed., New York: Wiley.

Heyman, D. F., Sobel, M. J. (Eds.) (1990), *Stochastic Models*, Amsterdam: North-Holland.

Hillier, F. S. (1990), *Introduction to Operations Research*, New York: McGraw-Hill.

Johnson, E. A., Katcher, D. A. (1973), *Mines against Japan*, Silver Spring, MD: Naval Ordnance Laboratory.

Kemeny, J. G. (1959), *A Philosopher Looks at Science*, New York: Van Nostrand Reinhold.

Law, A. M., Kelton, W. D. (1991), *Simulation Modeling and Analysis*, 2nd ed., New York: McGraw-Hill.

McCloskey, J. F. (1987), *Oper. Res.* **35**, 143–152, 453–470, 910–925.

Miser, H. J. (Ed.) (1995), *Handbook of Systems Analysis: Cases*, Chichester, England: Wiley.

Miser, H. J., Quade, E. S. (Eds.) (1985), *Handbook of Systems Analysis: Overview of Uses, Procedures, Applications, and Practice*, Chichester, England: Wiley.

Miser, H. J., Quade, E. S. (Eds.) (1988), *Handbook of Systems Analysis: Craft Issues and Procedural Choices*, Chichester, England: Wiley.

Morse, P. M., Kimball, G. E. (1951), *Methods of Operations Research*, New York: Wiley.

Nemhauser, G. L., Rinooy Kan, A. H. G., Todd, M. J. (Eds.) (1989), *Optimization*, Amsterdam: North-Holland.

Pollock, S. M., Barnett, A., Rothkopf, M. H. (Eds.) (1994), *Operations Research and the Public Sector*, Amsterdam: North-Holland.

Ravetz, J. R. (1971), *Scientific Knowledge and Its Social Problems*, Oxford: Oxford University Press.

Schön, D. H. (1983), *The Reflective Practitioner: How Professionals Think in Action*, New York: Basic Books.

Stone, L. D. (1989), *Oper. Res.* **37**, 501–506.

Stone, L. D. (1992), *Interfaces* **22**(1), 32–54.

Waddington, C. H. (1973), *OR in World War 2: Operations Research against the U-Boat*, London: Elek Science.

Wagner, H. M. (1969), *Principles of Operations Research with Applications to Management Decisions*, Englewood Cliffs, NJ: Prentice-Hall.

Further Reading

Gass, S. I., Harris, C. M. (Eds.) (1995), *Encyclopedia of Operations Research and Management Science*, Boston: Kluwer Academic Publishers.

Miser, H. J. (1980), *Science* **209**, 139–146. Also in: P. H. Abelson, R. Kulstad (Eds.) (1980), *Science Centennial Review*, Washington, DC: American Association for the Advancement of Science, pp. 121–128.

Mitchell, G. (1993), *The Practice of Operational Research*, New York: Wiley.

Moder, J. J., Elmaghraby, S. E. (Eds.) (1978a), *Handbook of Operations Research: Foundations and Fundamentals*, New York: Van Nostrand Reinhold.

Moder, J. J., Elmaghraby, S. E. (Eds.) (1978b), *Handbook of Operations Research: Models and Applications*, New York: Van Nostrand Reinhold.

Quade, E. S. (1989), *Analysis for Public Decisions*, 2nd ed., New York: North-Holland.

Rivett, P. (1994), *The Craft of Model Building for Decision Analysis*, New York: Wiley.

OPTICAL AMPLIFIERS

See AMPLIFIERS, OPTICAL

ORDER-DISORDER TRANSITIONS

J. M. SANCHEZ, *Center for Materials Science and Engineering, University of Texas, Austin, Texas, U.S.A.*

Introduction: General Aspects of Order and Disorder 1
1. **Characterization of Ordering** 4
1.1 The Manifestations of Ordering: Examples 4
1.1.1 Ordering in Alloys: Intermetallic Compounds 4
1.1.2 Long-Period Structures 5
1.1.3 Ordering in Low-Dimensional Systems 6
1.2 Formal Description of Order 8
1.2.1 Occupation Numbers and Correlations in Crystals 8
1.2.2 Long- and Short-Range Order 11
1.2.3 Ordering at Finite Temperatures 12
2. **Summary** 15
Works Cited 16
Further Reading 16

INTRODUCTION: GENERAL ASPECTS OF ORDER AND DISORDER

In nature, order is a recurrent theme. As such, we are conditioned to expect and recognize order in most physical systems. This experience has left us with an intuitive concept of order that, although it has served us well, requires some elaboration to be applicable in the realm of science. In most instances, order can be defined precisely in terms of mathematical concepts and be uniquely characterized using experimental probes, such as diffraction of x rays or electrons. Thus, the definition of order can be made void of subjective and/or semantic overtones. It is perhaps more difficult to rely on our intuition to define the concept of *disorder*, i.e., the *lack of order*, which appears, at first sight, to be a relative concept. However, we can find equally precise mathematical definitions of disorder. We will come back to these in the following sections so that we can expand upon the intriguing phenomenon of the transition from a disordered to an ordered state.

Undoubtedly the most apparent transition from a state of order to one of disorder is that of melting. In this case, an "ordered" crystalline state, which is stable at low temperatures, gives way to a "disordered" liquid state at high temperatures. Among several important features of this transition is the fact that the discrete translational symmetry characteristic of the crystalline state (see CRYSTALLINE STATE) is replaced by the continuous translational symmetry of the liquid. In the latter, all points in space are equivalent. The consequences of melting, in terms of physical properties, are generally quite apparent.

Other order-disorder transitions are less evident, although they also have profound effects on physical properties. In alloys, for example, and without ever abandoning the ubiquitous crystalline state, atoms of different kinds may be distributed more or less randomly (disordered state) or following a well-defined pattern in a unit cell (ordered state). Examples of these two states are schematically shown in Fig. 1 on the unit cell of a body-centered-cubic (bcc) crystalline lattice. Figure 1(a) shows a possible ordered state in which atoms of one kind (open circles) occupy all the cube corners, whereas the other type (full circles) occupy the cube centers. The crystal itself is constructed by repeating this unit which, in crystallographic terminology, is referred to as a simple cubic unit cell with two atoms per cell. We note in passing that this is the structure adopted by many ionic crystals, a typical case being cesium chloride (CsCl), and by a number of metallic systems,

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

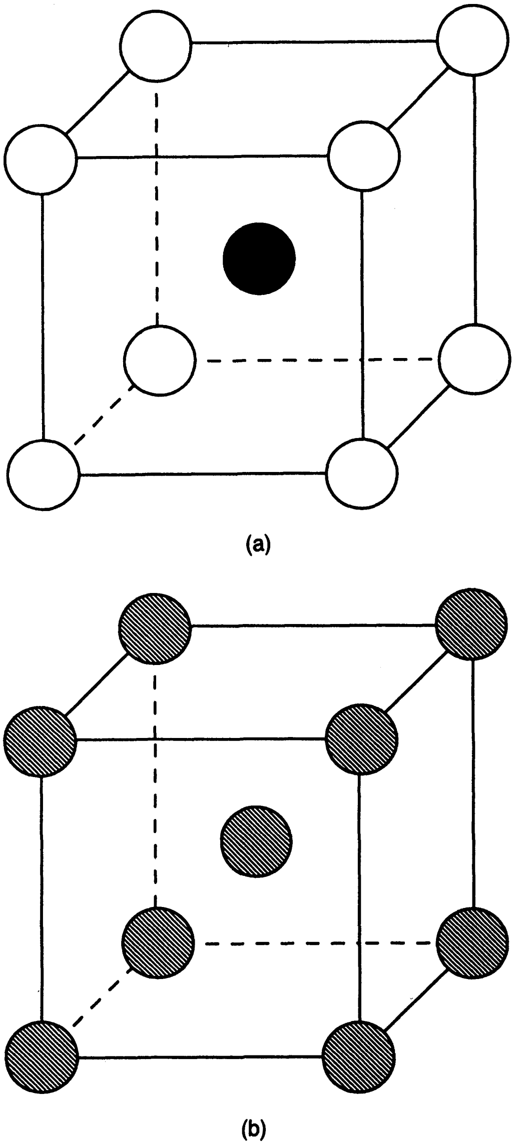

FIG. 1. Unit cells of (a) the primitive cubic CsCl and (b) the disordered bcc structures.

such as the low-temperature phase of copper-zinc (CuZn). Furthermore, since there are as many cube-corner as cube-center sites in the crystal, the numbers of Cs and Cl ions in CsCl, or of Cu and Zn atoms in CuZn, must be the same for perfectly stoichiometric compounds.

The representation of the disordered state is considerably more difficult since, strictly speaking, the unit cell is as large as the crystal itself and, thus, contains of the order of 10^{23} atoms. The disordered atomic arrangement is schematically shown in Fig. 1(b), in which an "average" atom (gray circle) is placed at each lattice site. As in the case of melting, the transition between the ordered and disordered states is also effected by temperature, and results in significant changes in the strength, brittleness, and electrical properties of the material.

Another easily recognizable order-disorder phenomenon is the spontaneous alignment of microscopic magnetic moments. Consider atoms having a well-defined magnetic moment and arranged on a crystalline lattice. At high temperatures, these atomic magnetic moments will be oriented in all possible directions such that the macroscopic, total magnetization of the crystal vanishes. Such a crystal is schematically shown in Fig. 2(a), which could be thought as a crude model of paramagnetic bcc iron. In many instances, including the ferrite phase of iron, the interactions between these atomic magnets are such that the energy of the crystal is lowered when the moments are aligned parallel to each other. Thus, the ground state, corresponding to the lowest energy state that can be achieved at low temperatures, is a ferromagnetic arrangement as shown in Fig. 2(b). In the ferromagnetic state, the macroscopic magnetization is nonzero and the material behaves as a permanent magnet (ferromagnet). Once again we find that macroscopic physical properties change significantly as the system undergoes an order-disorder transition.

There are certain common aspects to the three examples of order-disorder transitions discussed above. First and foremost is the fact that some physical property changes drastically in all cases. In melting the change is, primarily, in the ability of the material to withstand shear stresses; in the case of ordering in alloys, as in CuZn for example, it is in the strength or hardness of the material; and in the ferromagnetic transition is in the appearance of a permanent magnetic moment. These changes may reflect into either useful or detrimental properties. A material that is strong and ductile may become brittle upon ordering or, alternatively, some other desirable property, such as its electrical conductivity or high magnetic permeability, may be altered. These important practical aspects

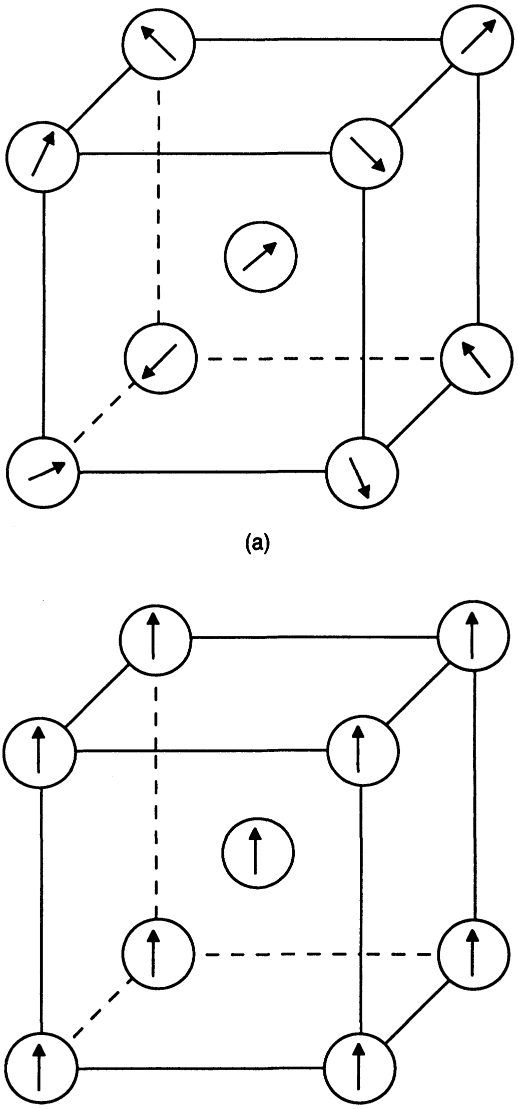

FIG. 2. Schematic representation of (a) paramagnetic and (b) ferromagnetic structures for the body-centered-cubic lattice.

have added to the interest in the study of order-disorder transitions in general terms as well as in specific systems.

A second aspect of commonality is that the transitions take place upon changes in some thermodynamic variable. In the cases discussed above, the thermodynamic variable is temperature although the transitions may also be effected by changes in pressure, in magnetic and/or electric fields, and, in the case of alloys, mixtures, or solutions, in the concentration of the different atomic or molecular components. Thus, the reasons for order-disorder transitions are to be found in the fundamental laws of thermodynamics and, more precisely, in the behavior of proper thermodynamic potentials or free energies.

A third aspect common to all order-disorder transitions is the existence of an *order parameter*. As a matter of convention, the order parameter is defined such that it vanishes in the disordered state whereas it takes finite values in the ordered state. The definition of the order parameter is not always intuitive, and it requires some characterization of the transition at hand before it can be identified. In the case of the paramegnetic-to-ferromagnetic transition, the order parameter may be defined as the macroscopic magnetization which, in turn, is the expectation value, or the average over the whole crystal, of the atomic magnetic moments. This average clearly vanishes in the paramagnetic state, where atomic magnetic moments point in all possible directions, and it is nonzero in the ferromagnetic state, where there is preferential orientation of these moments. In the case of ordering in alloys, as, for example, in the CuZn system, the order parameter is somewhat more difficult to visualize. However, a very precise definition can be given in terms of the probability of occupancy of cube-corner and cube-center sites by one type of atom, let us say Cu. Thus, the order parameter is given by the probability of a Cu atom occupying a cube-corner site *minus* the probability of a Cu atom occupying a cube-center site. Once again, the order parameter vanishes in the disordered state where Cu (and Zn) atoms occupy these two type of sites with equal probability and is nonzero in the order state. Unlike the case of the ferromagnet, where a nonzero order parameter can be put into evidence quite easily, the measurement of the order parameter in the case of the CuZn requires more sophisticated equipment such as an x-ray diffractometer.

A point of distinction between order-disorder transitions and general phase transitions is to be noted. Order-disorder transitions always involve a change in symmetry, which is not necessarily the case for phase transitions in general. For example, the translational symmetry is lost in melting, rotational symmetry is lost in the ferromagnetic state, and the *space group* of the crystal changes upon ordering in CuZn. In the last case, the crystal goes from a simple cubic structure in the ordered state to a bcc structure in the disordered state. This behavior is in contrast, for example, to the vapor-liquid phase transition at which the symmetry remains unchanged. The symmetry change characteristic of order-disorder transitions, generically known as *symmetry breaking*, usually provides powerful clues in the identification of the order parameter.

1. CHARACTERIZATION OF ORDERING

This section aims at providing a brief overview of the tools used in the description and characterization of ordering. The section begins with a few examples of ordered structures found in nature. These examples are chosen, primarily, because of their simplicity and as a way to introduce and highlight some of the relevant concepts. The elements needed for a more formal description of the ordered and disordered states are also given. The section concludes with a brief description of the thermodynamic principles involved in order-disorder transition, followed by examples of equilibrium phase diagrams.

1.1 The Manifestations of Ordering: Examples

1.1.1 Ordering in Alloys: Intermetallic Compounds Like most solids at low temperatures, alloys, which are mixtures of different atomic elements, adopt a crystalline structure. The phenomenon of ordering in these systems refers to the way in which the atomic elements are distributed in the crystal sites. An example of this type of ordering was discussed above and shown schematically in Fig. 1(b) for the case of the Cu_xZn_{1-x} alloys, where x stands for the concentration of Zn in the alloy. Clearly, the structure described by the unit cell of Fig 1(b) can be achieved only if the concentration x equals 0.5, i.e., at 1:1 stoichiometry. The ordered structure in this case is an example of an *intermetallic compound*. Compounds of other stoichiometries are always possible. In the case of Cu_xZn_{1-x}, two other low-temperature intermetallic compounds are observed in addition to CuZn, occurring at stoichiometries Cu_5Zn_6 ($x = \frac{5}{11}$) and $CuZn_4$ ($x = 0.2$). Because of their crystalline nature, intermetallic compounds necessarily belong to one of the 230 possible crystallographic space groups listed in the Henry and Lonsdale (1965) (see CRYSTALLINE STATE). Although the space-group notation provides an unambiguous way of describing the symmetry of the compounds, it is not particularly helpful in describing the atomic arrangements, i.e., the ordering of atoms in the unit cell. Thus, this space-group notation is not universally used to describe ordered structures, and a variety of other conventions abound. Among them is the use of a prototype structure to refer to all structures of a certain type. In this mode, one would refer to ordered CuZn as the "CsCl structure" [see Fig. 1(a)]. It is also common practice to use a symbol in order to identify a given ordered structure. For example, in this notation, the "CsCl" structure is known as the B2 structure, and the disordered body-centered-cubic CuZn is identified by the symbol A2. The most extensive compilation of compounds to date is to be found in Villars and Calvert (1985). Pearson introduced his own notation which is also commonly used in the literature. For example, the cubic Primitive structure of CuZn becomes the cP2 structure in Pearson's notation, where the 2 shows that two atoms are present in the unit cell.

Two closely related intermetallic compounds, which are manifestations of ordering in the face-centered-cubic lattice, are those known as the $L1_0$ and $L1_2$ structures (see Fig. 3). The $L1_2$ structure is simple (primitive) cubic (space group $Pm\bar{3}m$), whereas $L1_0$-type ordering, usually accompanied by a slight tetragonal distortion, results in a simple tetragonal structure (space group $P4/mmm$). The $L1_0$ and $L1_2$ structures are representative of the simplest type of ordering that can be realized in fcc lattices. As such, they are observed to occur in a large number of alloys. For example, in the Cu-Au system, the two possible $L1_2$ structures, namely Cu_3Au and $CuAu_3$, as well as the $L1_0$ CuAu are present. Furthermore, all three intermetallic compounds go through an order-disorder transition in the solid state, adopting an fcc structure at high temperatures before, eventually, melting. In view of this, the intermetallic compounds CuAu and Cu_3Au are commonly

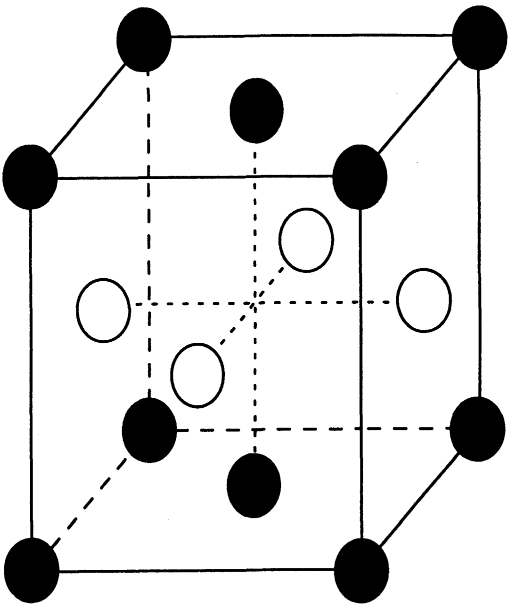

(a)

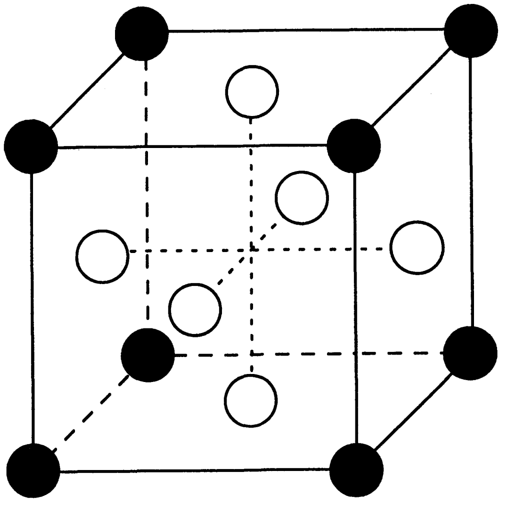

(b)

FIG. 3. Unit cells of (a) the body-centered tetragonal $L1_0$ and (b) the primitive cubic $L1_2$ structures.

used as the prototypes for the $L1_0$ and $L1_2$ structures, respectively.

1.1.2 Long-Period Structures

The simplicity of the low-temperature ordered compounds present in the Cu-Au systems is in sharp contrast to the appearance, around 1:1 stoichiometry and in a narrow temperature range of approximately 25 K, of an orthorhombic ordered structure known as CuAuII. Although CuAuII is closely related to the tetragonal $L1_0$ structure, the long orthorhombic lattice constant is approximately ten times that of $L1_0$. Structures such as that of CuAuII are known as *long-period* structures. At present, a seemingly endless list of long-period structures has been reported, with those related to the $L1_2$ structure of Cu_3Au being particularly common. Examples include Ag_3Mg, Cu_3Pd, Au_3Zn, Al_3Ti, etc. A characteristic feature of these long-period structures is that they can be constructed by introducing a periodic array of so-called *antiphase boundaries* (APB). Antiphase boundaries are planar defects that occur when two ordered domains meet. For example, an APB results if two otherwise identical ordered domains are displaced relative to each other by a distance that is less than a period of the lattice. Figure 4 shows schematically such an APB in the $L1_2$ structure.

A rather common structure, realized, for example, in the intermetallic compounds Al_3Ti and Ni_3Mo, can be obtained by arranging APBs every other $L1_2$ unit cell. The unit cell of the resulting tetragonal structure, known as DO_{22}, is shown in Fig. 5. The structure itself is made of a periodic distribution of APBs that separate domains two unit cells thick. Another example is the structure with periodicity 4, sometimes referred as the DO_{23}, shown in Fig. 6. DO_{22} and DO_{23} belong to a relatively simple sequence of structures for which the periodicity of the APBs is $2n$, where n is an integer number. It is interesting to note that the $L1_2$ structure corresponds to an infinite periodicity n and, thus, the terminology long-period structures is rather subjective.

More complex sequences of APBs are also

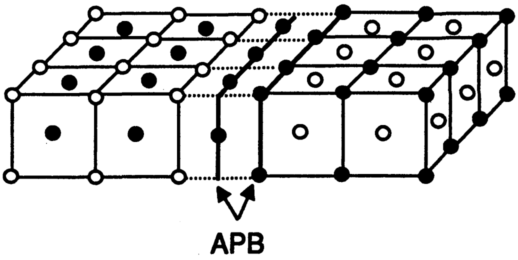

FIG. 4. Schematic representation of an antiphase boundary between two $L1_2$ ordered domains.

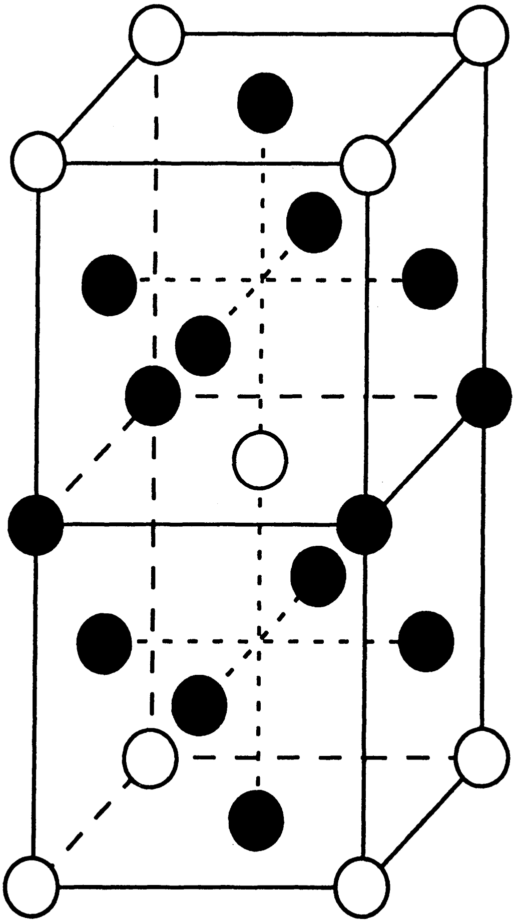

FIG. 5. Unit cell of the DO_{22} structure.

found in nature. To describe these structures a notation that indicates the sequence and the domain thickness in terms of the unit cell of the parent structure is commonly used. For example, the structure $\langle n_1 n_2 \ldots n_q \rangle$ corresponds to a sequence in which a domain of thickness n_1 is followed by an antiphase domain of thickness n_2, etc. As an example, the unit cell of the $\langle 211 \rangle$ ordered structure reported, for example, in the Pt_3V system is shown in Fig. 7.

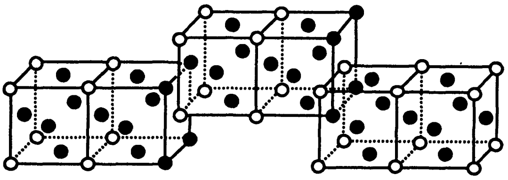

FIG. 6. Unit cell of a long-period structure generated by antiphase boundaries with periodicity 4.

1.1.3 Ordering in Low-Dimensional Systems When a physical phenomenon is strictly constrained to a surface, an interface, or an atomic plane in a crystal, the system may be construed as two dimensional. This reduced dimensionality has important consequences on general physical behavior and, in particular, on order-disorder transitions. The interest in the study of low-dimensionality systems is twofold. First, the physical understanding of phase transitions in two-dimensional systems is on very solid ground. For some simple models, such as the Ising model of two-dimensional ferromagnets, analytical results are known. Thus, these low-dimensional systems offer an excellent opportunity to test our most fundamental physical laws, models, and/or theories.

On the practical side, surfaces and interfaces are dominant in engineering systems and structures. For example, a modern microelectronics device is typically composed of several layers of different materials such as semiconducting silicon, noble metal gold, and transition metals—e.g., tungsten, titanium, and nickel—in close contact with each other. Surfaces also play a key role in friction, wear, and fracture of materials as well as in the performance of catalysts, adhesives, and composites and in biomaterials used for implants and prosthetics.

The order-disorder transition at a surface, for example, is generally found to differ considerably from that in the bulk. In many instances it is seen that, while the bulk undergoes a discontinuous or abrupt order-disorder transition, the surface orders continuously in the characteristic fashion of second-order transitions. This is the phenomenon of surface disorder or wetting, whereby a layer of the "disordered" or high-temperature phase develops at the surface as the transition temperature is reached from below. In the context of the solid-liquid transition, surface disorder is closely related to the phenomenon of premelting in which the surface region is seen to lose crystalline order at temperatures several degrees below the melting temperature of the bulk. The phenomenon of surface ordering has also been reported in ferromagnetic materials. In this case, the surface magnetization remains fi-

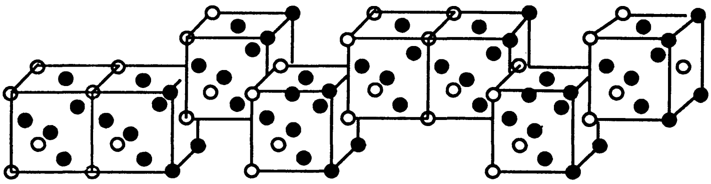

FIG. 7. Schematic representation of the ⟨211⟩ long-period structure.

nite at temperatures above the Curie temperature of the bulk.

The surface behavior in intermetallic compounds is governed by the interplay between at least one order parameter and the surface concentration. Furthermore, the symmetry breaking observed in the order-disorder transition is now constrained by the surface geometry and the crystallographic relation between the surface and the bulk. Consequently, surface-ordering phenomena and behavior in intermetallics are more complex than in the bulk. A case in point is the behavior of the (110) surface of Cu_3Au, for which two different surface-ordering transitions are reported to occur. This is in contrast to the single disorder-$L1_2$ transition taking place in the bulk at 663 K. Figure 8 depicts the atomic arrangements for each of the (110) surface phases, labeled 2x1 and 4x1, found in Cu_3Au. Note that perpendicular to the surface, the Cu_3Au crystal is made of alternating pure Cu and mixed (equiatomic) Cu-Au (100) planes. Thus, the 2x1 structure corresponds to a surface termination at a mixed (110) plane, whereas the formation of the 4x1 structure requires segregation of Cu to the surface.

Occasionally, a system is found where the order-disorder transitions can be described as a relatively simple two dimensional system. One such case is provided by the ordering of oxygen in the superconducting oxides, as exemplified by the celebrated $YBa_2Cu_3O_{6+x}$ system. The ordering of oxygen takes place, primarily, on one type of crystallographic plane. This ordering drives an important structural transition in the oxide from a high-temperature tetragonal structure to a low-temperature orthorhombic structure. The unit cells of both the tetragonal and orthorhombic structures are shown in Fig. 9 for the stoichiometric compound $YBa_2Cu_3O_7$. In the high-temperature tetragonal phase, shown in Fig. 9(a), the O1 and O5 oxygen sites are occupied with an equal probability of $\frac{1}{2}$, whereas the sites O2 and O3 are equivalent and occupied with probability 1. In the orthorhombic phase, Fig. 9(b), the O1 sites are now occupied with probability 1, leaving the O5 sites empty. In reference to the discussion of ordering in alloys, note that the $YBa_2Cu_3O_7$ system is

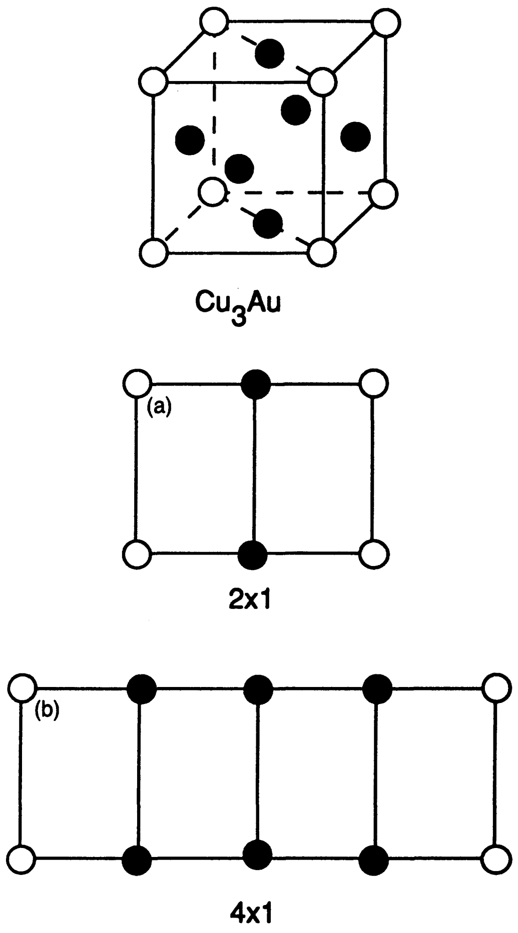

FIG. 8. Two-dimensional unit cells of ordered structures observed in the (110) surface of Cu_3Au. (a) The ordered structure with periodicity 2x1 appears at a temperature of 654 K, slightly lower than the bulk transition of 660 K due to surface segregation effects. (b) The transition to a 4x1 structure occurs at approximately 425 K.

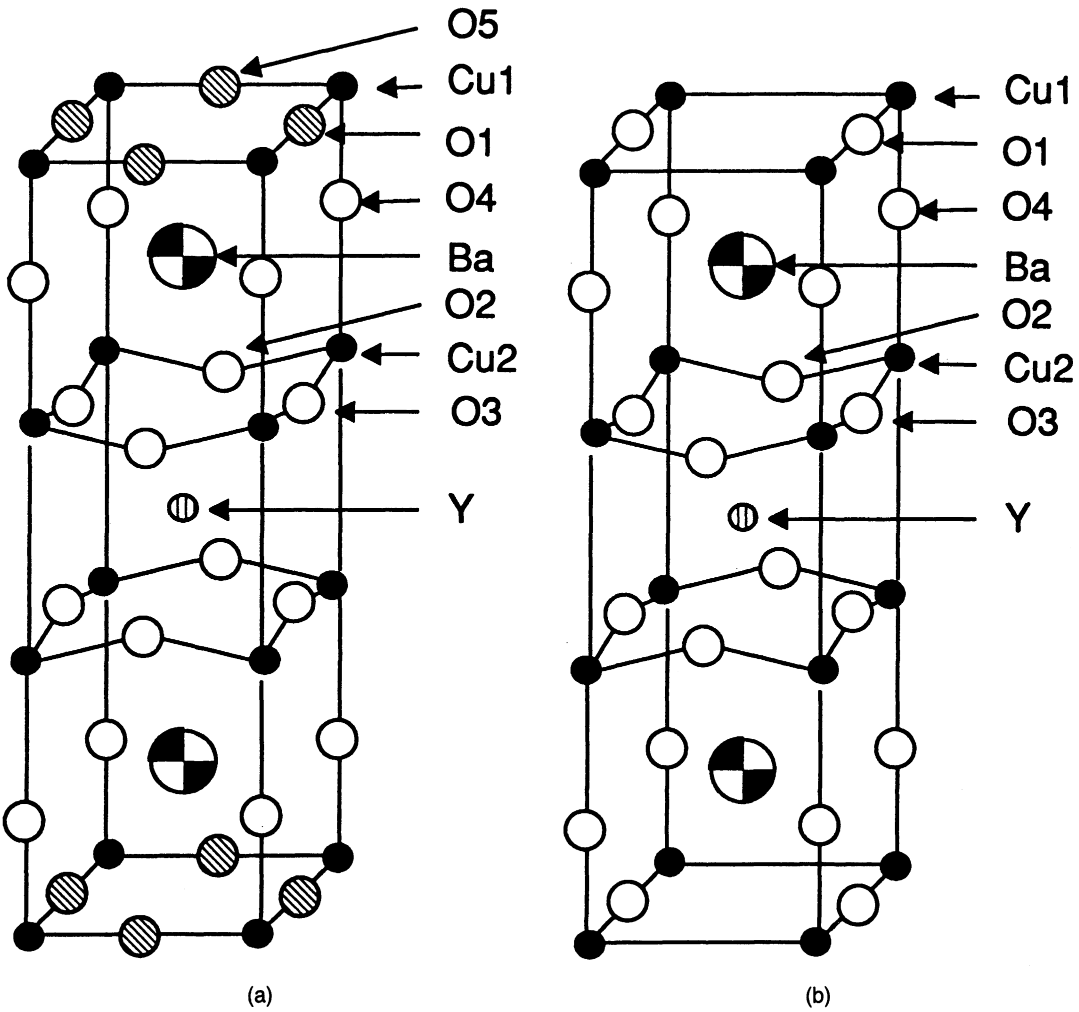

FIG. 9. The unit cells of the (a) high temperature tetragonal and (b) orthorhombic phases of $YBa_2Cu_3O_7$.

equivalent to that of a binary alloy in which the ordering species are vacant sites and oxygen ions. A number of possible ordered structures have been reported in $YBa_2Cu_3O_{6+x}$, a fact that reflects the relative complexity of the oxygen-oxygen interactions. However, two of the most prominent structures can be easily understood in terms of a simple model based on the repulsion between neighboring oxygen ions, coupled with an effective attraction between oxygen ions mediated by a Cu atom. The two-dimensional unit cells for these structures are shown in Fig. 10, where the repulsive interactions between nearest- (V_1) and next-nearest-neighbor (V_2) oxygen atoms and the attractive Cu-mediated interaction (V_3) are also shown.

The use of simple models base on short-ranged atomic interactions has traditionally played an important role in the classification and prediction of ordered structures in both two- and three-dimensional systems. For these models, the enumeration of the ordered states with the lowest energy, generically known as ground states, can be cast in the form of the classical linear programming problem of enumerating vertices of a convex polyhedron (see below).

1.2 Formal Description of Order

1.2.1 Occupation Numbers and Correlations in Crystals A useful concept in the description of order in crystalline systems is

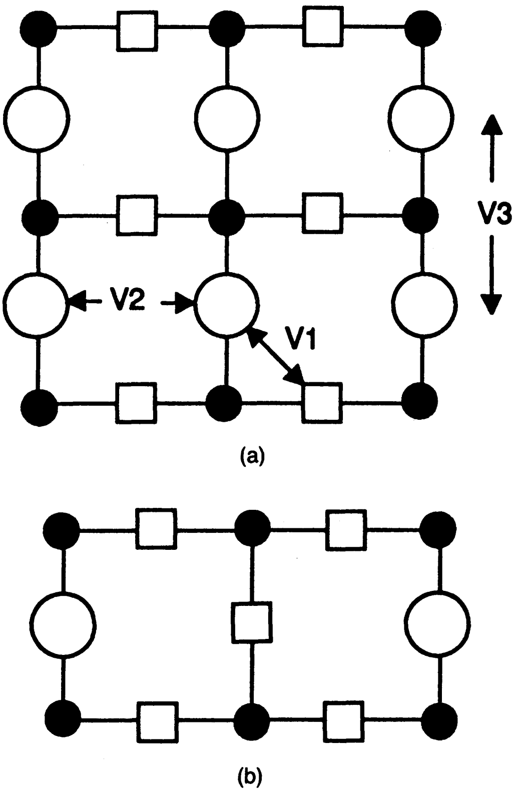

FIG. 10. Two-dimensional representation of oxygen ordering in the basal plane of the $YBa_2Cu_3O_{6+x}$ system corresponding to orthorhombic structures observed around (a) $x = 1$ and (b) $x = 0.5$.

that of an occupation number. In a binary alloy, for example, the occupation number σ_i takes two values, let us say $+1$ and -1. Each value of the occupation number indicates the atomic species located at the particular lattice site, which is labeled by the index i. The concept can be extended so that the occupation number refers, for example, to any number of atomic species, to the direction and magnitude of atomic magnetic moments, or, as in the case of the superconducting oxides discussed above, to vacant or occupied sites, A microscopic state of the system in which the identity of each atom at each crystal site is known can be fully specified by an N-dimensional vector $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$, where N stands for the total number of sites in the crystal. The total number of possible arrangements of A and B atoms, i.e., microstates, in a crystal with N lattice sites is 2^N.

A physical description of the system settles for a much less detailed knowledge in which only expectation values, or averages based on the equilibrium probability distribution, of certain key quantities are known. The calculation of such averages falls within the realm of statistical mechanics and thermodynamics and is essential for the study of the effect of temperature, pressure, and other thermodynamic variables on the order-disorder transition.

For a binary system, the average of the occupation number, denoted $\langle \sigma_i \rangle$, depends only on the concentration of each atomic species. If σ_i equals 1 for A atoms and -1 for B atoms, then $\langle \sigma_i \rangle = x_A - x_B$, where x_A and x_B are the concentrations of A and B atoms, respectively. The expectation value $\langle \sigma_i \rangle$ determines the composition of the binary alloy, but conveys no information on the state of order of the system. Some information regarding the atomic arrangement is provided by the pair correlation function $\langle \sigma_i \sigma_j \rangle$. These pair correlations depend strongly on the distance between sites i and j in the crystal. Note that for the perfectly ordered CuZn structure [Fig. 1(a)], the nearest-neighbor pair correlation equals -1, i.e., nearest-neighbor pairs are made of different atomic species, whereas it equals $+1$ for next-nearest neighbors in which case the two atoms are of the same type. On the other hand, for a random atomic distribution (totally disordered), the pair correlations vanish for 1:1 stoichiometry ($x_A = x_B = 0.5$). This stems from the fact that, for a random distribution, averages of products of the occupation numbers equal products of the averages; i.e., $\langle \sigma_i \sigma_j \rangle_{\text{rand}} = \langle \sigma_i \rangle \langle \sigma_j \rangle = (x_A - x_B)^2$.

The pair correlation functions contain enough information to distinguish between ordered and disordered states, although the characterization is not complete. For a complete description of the ordered state, all possible correlation functions are needed. The complete set of correlation functions includes averages of products of occupation numbers located on all possible sets of crystal sites. These sites can be visualized as forming well-defined geometrical figures, or clusters. In general, for a cluster α composed of crystal sites $\{i,j,\ldots,k\}$, i.e., $\alpha = \{i,j,\ldots,k\}$, the correlation function ξ_α is given by

$$\xi_\alpha = \langle \sigma_i \sigma_j \ldots \sigma_k \rangle.$$

As implied above, a full description of partially ordered and partially disordered states requires the characterization of an infinite number of correlation functions. Since, in practice, this cannot be achieved, approximations to the problem are called for. One approach consists of describing correlation functions up to a maximum cluster size. In this cluster approximation, the state of order is described by a finite set of correlation functions $\{\xi_\alpha\}$ corresponding to each of the subclusters of the maximum cluster. For example, a regular tetrahedron cluster made of nearest-neighbor sites is commonly used to describe ordering in the fcc lattice. At this level of approximation, four correlation functions are involved corresponding to the point, pair, triangle, and tetrahedron clusters. These clusters are shown schematically in Fig. 11. In general, the M-dimensional space spanned by the finite set of correlation functions ($M = 4$ for the tetrahedron approximation in the fcc lattice) is known as the configurational space of the system; i.e., any state of partial order corresponds to a point in this space. These M correlation functions are also subject to linear geometric constraints. For example, if the point correlation function $\xi_1 = \langle\sigma_1\rangle$ in a binary system is close to 1, implying that the system is made up almost entirely of A atoms, then the nearest-neighbor pair correlation function ξ_2 is also constrained to being close to 1. It can be shown that these constraints are linear, and that the space spanned by the correlation functions is a convex configurational polyhedron.

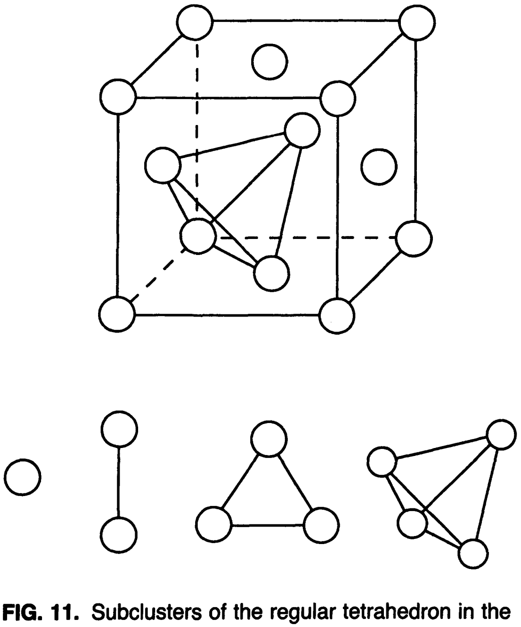

FIG. 11. Subclusters of the regular tetrahedron in the fcc lattice.

The cluster description of order is particularly useful for studying model systems with relatively short-range interactions. For these models, the expectation value of the energy is exactly given by a linear combination of correlation functions:

$$\langle E\rangle = \sum_{\beta\subseteq\alpha} V_\beta \xi_\beta,$$

where the sum is carried over all the subclusters β of the maximum cluster α, and where the coefficients V_β are known as *effective cluster interactions*. Since the energy is linear in the correlation functions and the configurational polyhedron is convex, it follows that the ground states are given by the vertices of the configurational polyhedron. Thus, a ground-state search is equivalent to finding all such vertices, itself a classical problem in linear programming. This procedure, although trivial in low dimensions (two or three), becomes rather cumbersome when the number of correlation functions is large, as would be the case for physical systems with long-range interactions.

The approximate description of order using finite clusters has been useful to describe and classify order structures in real systems. For example, a ground-state search for the fcc lattice with nearest- and next-nearest-neighbor effective interactions results in the structures shown in Fig. 12. In this simple model, the relevant parameter is the ratio R of first- to second-neighbor effective interactions. For example, for negative values of R, the model predicts the commonly observed A_3B ($Pm\bar{3}m$, Cu_3Au, $L1_2$) and AB ($P4/mmm$, CuAu, $L1_0$) structures. For $0 < R < 0.5$ the ground states are those labeled A_2B_2 ($I4_1/amd$), A_2B ($Immm$, Pt_2Mo), A_3B ($I4/mmm$, $TiAl_3$, DO_{22}) and A_5B. Finally, for R larger than 0.5, the ground states are indicated as AB ($R\bar{3}m$, CuPt, $L1_1$) and A_2B. The labeling scheme in Fig. 12 gives first the compound formula, which is followed, if available, by the space group, the prototype structure, and the corresponding structural symbol. This simple one-parameter model predicts a relatively large number of ordered structures, some of which are observed in nature. It should be noted that the number of possible ordered structures grows exponen-

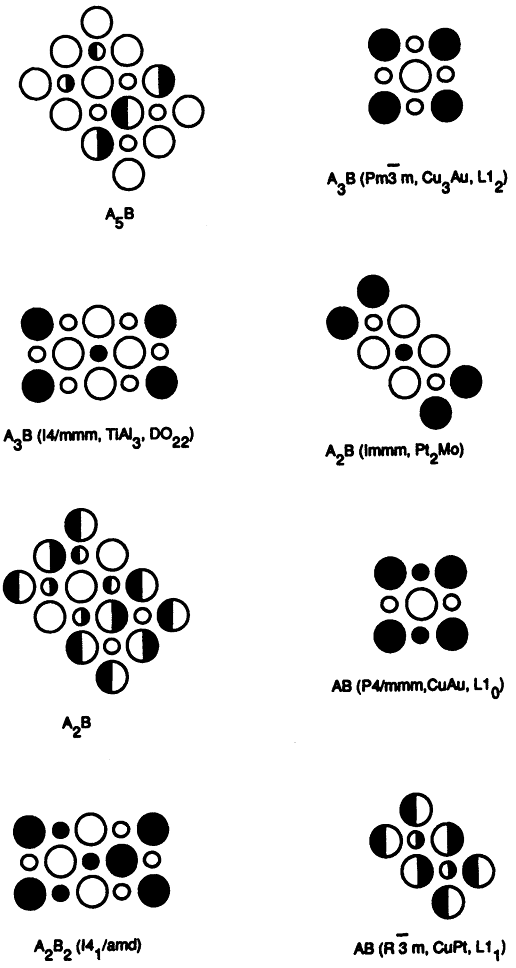

A_5B

A_3B ($Pm\bar{3}m$, Cu_3Au, $L1_2$)

A_3B ($I4/mmm$, $TiAl_3$, DO_{22})

A_2B ($Immm$, Pt_2Mo)

A_2B

AB ($P4/mmm$, $CuAu$, $L1_0$)

A_2B_2 ($I4_1/amd$)

AB ($R\bar{3}m$, $CuPt$, $L1_1$)

FIG. 12. Schematic representation of the unit cells for the ground states of the fcc lattice with nearest- and next-nearest-neighbor interactions. The two-dimensional drawings are projections of the unit cells onto the (001) crystallographic planes. Circles of different sizes corresponds to atoms located on consecutive (001) planes, with open and full circles representing *A* and *B* atoms, respectively. Half-open circles represent *A* and *B* atoms alternating along the direction perpendicular to the paper.

tially with the range of interactions, underscoring the richness of ordering phenomena in crystalline systems.

1.2.2 Long- and Short-Range Order The state of perfect order and that of total disorder are idealizations seldom realized in nature. The correlation functions introduced above provide a hierarchical framework to measure states of partial order (or disorder) that are relevant in the physical world.

In the disordered state, the correlation functions measure deviation from randomness and, as such, they provide an infinite set of what are generically known as *short-range order* (SRO) parameters. In practice, slight modifications of the correlation functions, the so-called Warren–Cowley SRO parameters, are used. These parameters have been adopted almost universally since they can be measured rather precisely using x-ray or neutron diffraction techniques. For a binary system, the Warren–Cowley SRO parameters are defined by

$$\alpha(r_{ij}) = \frac{\langle\sigma_i\sigma_j\rangle - \langle\sigma_i\rangle\langle\sigma_j\rangle}{(1 - \langle\sigma_i\rangle\langle\sigma_j\rangle)},$$

where r_{ij} stands for the distance between crystal sites i and j. These SRO parameters vanish for the random state, and are normalized in such a manner that they equal 1 for $r_{ij} = 0$, i.e., when the crystal sites are equal and, thus, $\sigma_i^2 = 1$.

If in the disordered state atoms of the same type have the tendency to cluster together, then the pair correlations for short distances will be positive and larger than the value corresponding to a random atomic distribution. Thus, in this case, the SRO parameters are positive. If the system has an ordering tendency, however, the SRO parameters will oscillate with distance, being negative (positive) for distances at which atoms of different (same) types are preferentially found. In the disordered state, the amplitude of these oscillations decreases with distance, vanishing for sufficiently large separation between the two crystal sites. That is to say, the atomic occupations of distant sites are uncorrelated.

The SRO parameters are functions of the thermodynamic state of the system and, therefore, change with temperature, pressure, and composition. At sufficiently high temperatures the system becomes increasingly disordered, and the Warren–Cowley parameters tend to zero. On the other hand, as the temperature is lowered, and, let us say, an order-disorder transition is approached, the oscillations of the SRO parameters in the disordered state persist over longer distances, decaying much more slowly with distance. At and below the order-disorder transition, the characteristic decay distance becomes infinite. Thus, *long-range order* (LRO) sets in and the atomic occupancy of distant sites is now strongly correlated.

The onset of LRO at the transition is described, appropriately enough, by one or more LRO parameters. As is the case with short-range order, there is significant latitude in the choice of LRO parameters. However, all practical definitions are related to the symmetry breaking that takes place at the transition. In particular, the change of symmetry taking place at the transition establishes uniquely the number of order parameters needed to describe the transition. It is common practice to group the different order parameters into the components of a vector, and to refer to the number of components as the dimension of the LRO parameter.

Returning to the example of CuZn, consider the transition from a disordered bcc structure [Fig. 1(b)] to the ordered $B2$ structure [Fig. 1(a)]. In addition to being endowed with the translational symmetry of the simple cubic lattice, the disordered system is invariant to a translation by a vector equal to half the cube diagonal, i.e., $(a/2)[1,1,1]$, where a is the cubic lattice constant. This symmetry element is lost in the ordered $B2$ structure, for which the cube-corner and cube-center sites are not equivalent. Thus, the expectation values of the occupation number at the cube corners, $\langle\sigma_0\rangle$, and at cube centers, $\langle\sigma_1\rangle$, are different. In this case, the transition is described by a unique LRO parameter, which can be defined as $\eta_1 = (\langle\sigma_1\rangle - \langle\sigma_0\rangle)/2$. This one-dimensional order parameter is nonzero in the ordered state, vanishing identically in the disordered state where $\langle\sigma_0\rangle = \langle\sigma_1\rangle$.

The concept of sublattice is often used to emphasize the nonequivalence of different sites in the unit cell of the ordered structure. A sublattice, which is not necessarily a Bravais lattice, connects all equivalent sites in a crystal. In the $B2$ structure, two interpenetrating simple cubic sublattices, displaced from each other by half the cube diagonal, are easily identified [see Fig. 1(a)]. The independent averaging of the occupation numbers on each sublattice leads naturally to a nonzero order parameter in the ordered state. This procedure is referred to as sublattice averaging and is commonly used in the thermodynamic description of the ordered state.

Sublattices and the associated LRO parameters may be defined for any ordered structure. Knowledge of the space group of the low-temperature structure uniquely determines the number of nonequivalent sites within the unit cell. These nonequivalent sites are known as Wyckoff positions and are listed in Henry and Lonsdale (1965), together with all symmetry elements, for all the 230 space groups. If the number of Wyckoff positions for a given structure is $m + 1$, the dimensionality of the LRO parameter is m, and the different components are given by

$$\eta_k = (\langle\sigma_k\rangle - \langle\sigma_0\rangle)/2 \quad (k = 1,\ldots,m).$$

For example, the number of nonequivalent sites in the primitive cubic $L1_2$ structure is 2, so that its LRO parameter is of dimension 1. On the other hand, the unit cell of the body-centered-tetragonal DO_{22} structure, with space group $I4/mmm$ (Fig. 5), contains three different Wyckoff positions, and the dimension of the LRO parameter is 2.

1.2.3 Ordering at Finite Temperatures

The transition between the ordered and disordered states is governed by the behavior of a thermodynamic potential or free energy. The relevant free energy is dictated by external constraints. For example, at a fixed temperature, the relevant thermodynamic potential is either the Helmholtz or the Gibbs free energy, depending on whether the volume or the external pressure is kept constant.

The fundamental thermodynamic principle involved in the study of phase stability and order-disorder transitions is that, under the proper set of conditions, the free energy is minimum for the equilibrium phase. Although the actual determination of the free energy, either experimentally or theoretically, may in itself be a rather involved task, much can be learned from general principles. An elegant and relatively simple theory of order-disorder transitions was advanced by Landau and Lifshitz (1969) on the basis of the expansion of the free energy in powers of the LRO parameter. Subsequent developments have established that the free energy is singular at the transition and, thus, such expansions are not generally valid. Nevertheless, the theory of Landau and Lifshitz is frequently used since it captures the most important aspects of order-disorder transitions.

For the case of a one-dimensional order parameter η, a fourth-order Landau and Lifshitz expansion is of the form

$$F(\eta,T,P) = F_0(T,P) + A(T,P)\eta^2 + B(T,P)\eta^3 + C(T,P)\eta^4,$$

where $F_0(T,P)$ is the free energy of the disordered state, i.e., for $\eta = 0$, and the coefficients $A(T,P)$, $B(T,P)$, and $C(T,P)$ are, in general, functions of temperature (T) and pressure (P). A fourth-order expansion of the free energy is usually sufficient to describe the thermodynamic behavior of the system provided that the coefficient $C(T,P)$ is positive. In the event that the fourth-order coefficient is negative, the free-energy expansion must be carried out to, at least, the next even power with a positive coefficient. In general, the calculation of the expansion coefficients is a difficult task that requires the adoption of a specific model for the atomic interactions. However, some general thermodynamic properties can be established using symmetry and group-theoretical arguments. In particular, Landau and Lifshitz studied the conditions under which the coefficient B in the free-energy expansion vanishes. Their analysis led to a set of celebrated rules for the occurrence of continuous phase transitions.

A vanishing third-order coefficient B, which is solely determined by the symmetry of the ordered phase, implies that the free energy is an even function of the LRO parameter, i.e., $F(\eta,T,P) = F(-\eta,T,P)$. Thus, assuming $C > 0$, a plot of the free energy vs the LRO parameter looks as shown schematically in Fig. 13(a). The several curves in the figure correspond to free energies at temperatures above, at, and below the order-disorder transition temperature T_c. At high temperatures, the minimum in the free energy occurs at $\eta = 0$ and the equilibrium state is disordered. At the transition temperature, the second-order coefficient $A(T,P)$ vanishes, becoming negative below T_c. Thus, below T_c, two minima in the free energy develop at finite values of η, indicating that the ordered state is the equilibrium thermodynamic phase. These two states, with equal free energies and ordered parameters of opposite signs, correspond to the same thermodynamic phase. This can be seen, for example, for the $B2$ structure shown in Fig. 1(b). In this case, a change in the sign of the order parameter corresponds to a simple exchange of cube-corner with cube-center sites. This operation leaves the $B2$ structure invariant and, therefore, symmetry imposes the requirement that the free energy be an even function of the LRO parameter; i.e., $F(\eta,T,P) = F(-\eta,T,P)$. In turn, this implies the vanishing of the coefficient B in the free-energy expansion.

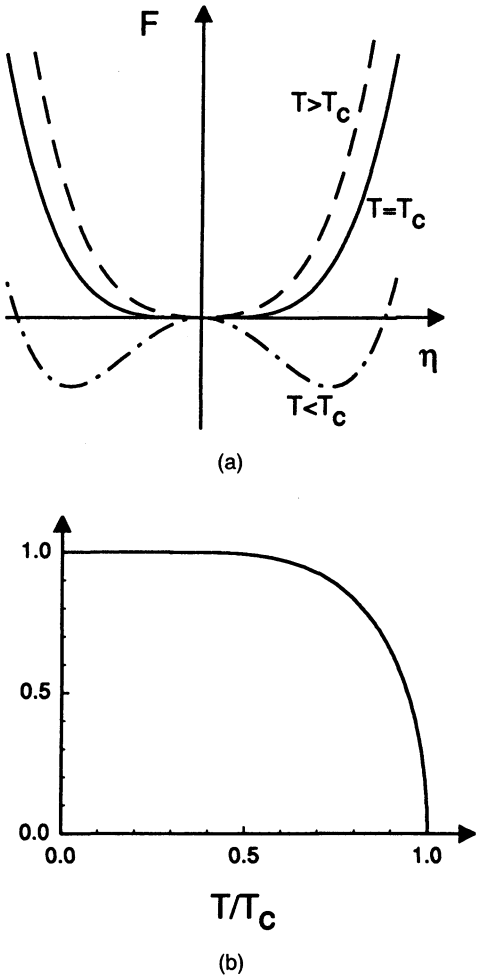

FIG. 13. Schematic (a) free-energy curves and (b) equilibrium long-range order parameter for systems displaying a second-order phase transition.

A plot of the value of η at which the minimum in the free energy takes place vs temperature is shown in Fig. 13(b). Note that the LRO parameter vanishes continuously as T_c is approached from below. This behavior of the LRO parameter is characteristic of *second-order* phase transitions.

The situation is significantly different when the third-order coefficient B in the free-energy expansion is nonzero, as shown in

Fig. 14. The free-energy curve may develop a local minimum at temperatures above the order-disorder transition. At the transition temperature, the LRO parameter jumps discontinuously from zero to a finite value corresponding to the ordered state. In this case, the order-disorder transition is said to be of *first order*.

For alloys, intermetallic compounds, and mixtures, the free energy is also a function of the concentrations of the different atomic or molecular components. For second-order transitions this concentration dependence of the free energy does not significantly alter the picture of order-disorder transitions given above. This is because at a second-order transition the compositions of the ordered and disordered phases are equal. For first-order transitions, however, an additional complication arises since the transition may now take place between phases of different average compositions. The situation is illustrated in Fig. 15, which shows schematic free-energy curves as a function of concentration for ordered and disordered phases in a binary alloy. The equilibrium condition of minimum free energy translates into a common-tangent rule that is also shown schematically in Fig. 15. The two points of tangency in Fig. 15 define three distinct regions: (I) a region where the disordered phase is stable; (II) a two-phase region where ordered and disordered phases coexist; and (III) a region where the ordered phase is stable. The information on the stability and coexistence between phases is commonly collected into graphs called *phase diagrams*. These phase diagrams are graphic representations of the temperature and composition ranges of stability and coexistence of all the thermodynamically stable phases in a given system. For example, the phase diagram of the Cu-Zn system is shown in Fig. 16, where the single-phase fields of the solid phases are indicated by Greek letters. The unlabeled regions in the figure correspond to two-phase fields, within which the adjacent phases coexist. The assignation of Greek letters to the stable phases is arbitrary and bears

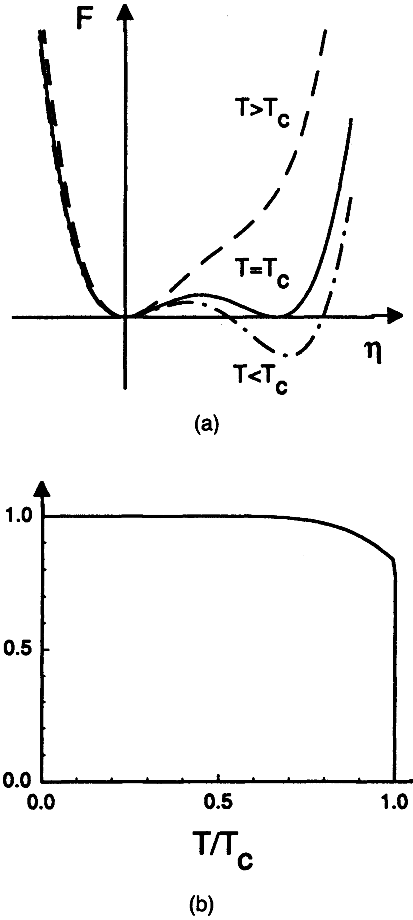

FIG. 14. Schematic (a) free-energy curves and (b) equilibrium long-range order parameter for systems displaying a first-order phase transition.

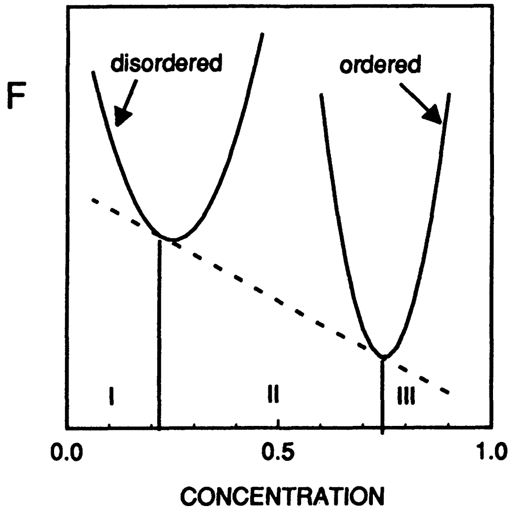

FIG. 15. Schematic free-energy curves vs composition for the ordered and disordered phases of a binary alloy that undergoes a first-order transition. The ordered and disordered phases are stable in regions I and III, respectively. For concentrations in region II both phases coexist.

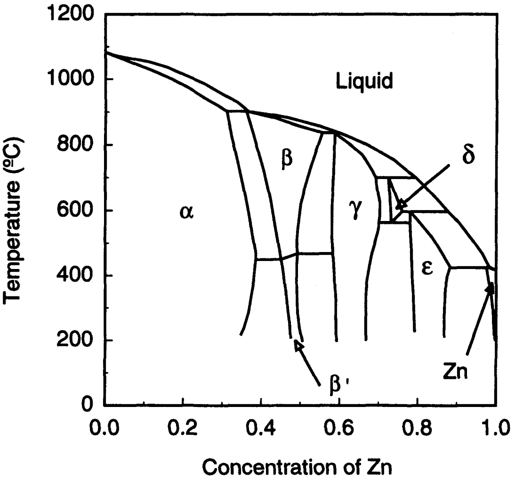

FIG. 16. Equilibrium phase diagram for the Cu-Zn system.

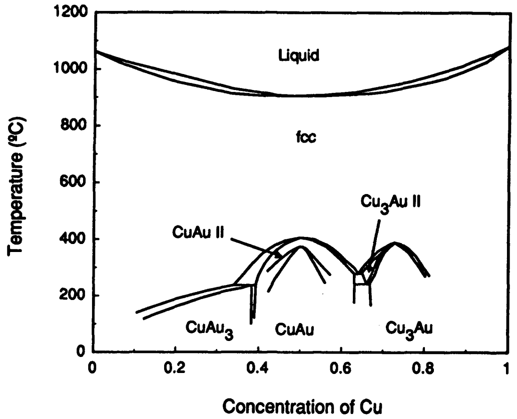

FIG. 17. Equilibrium phase diagram for the Cu-Au system.

no relation to the structure or to the symmetry of the phases. The phase indicated as α is a disordered fcc alloy. As the concentration of Zn increases beyond approximately 0.3, the α phase transforms, via first-order transitions, either to a bcc alloy (β) at high temperatures, or to an ordered *B*2 structure (β') at lower temperatures (Pearson's symbol *cP*2). The two intermetallic compounds labeled γ and ε have Pearson's symbols *tP*4 and *hP*2, respectively. The δ phase, occurring in a small concentration region at high temperatures, also has the *B*2 structure. At the Zn end of the phase diagram, the equilibrium phase is a hexagonal-closed-packed solid solution, with relatively low solubility of Cu.

Phase diagrams convey, in a concise and efficient manner, a great deal of information on the stability of different phases, on the occurrence of order-disorder transitions, and on the nature of these transitions. A well-known example is the CuAu phase diagram shown in Fig. 17. In this case, the system crystallizes from the liquid into a disordered fcc phase. As seen in the phase diagram, this disordered phase is stable over the entire composition range of the alloy. As the temperature is lowered, however, the phase diagram indicates that the disordered Cu-Au alloys undergo several ordering transitions. Around a Cu concentration of 0.25, the transition is from fcc solid solution to the $CuAu_3$ intermetallic compound that has the $L1_2$ structure. For larger concentrations of Cu the complexity and richness of the phase diagram increases: near a 1:1 stoichiometry the disordered solid solution first transforms into the CuAuII long-period structure, changing, at slightly lower temperatures, into the short-period CuAu ($L1_0$) phase. A similar sequence of transitions is found for higher Cu concentrations: the disordered phase also transforms to a long-period structure, labeled Cu_3AuII in Fig. 17, which, in turn, transforms to the Cu_3Au ($L1_0$) structure at lower temperatures. Finally, since all phases in Fig. 17 are separated by two-phase regions, it also follows from the phase diagram that all transitions are of first order.

2. SUMMARY

A few examples of well-known physical systems have been selected with the idea of exposing the fundamental and common aspects involved in the phenomenon of order-disorder transitions. Some well-known examples of ordering systems are followed through, starting with a description of their structure and symmetries and ending with a description of their equilibrium phase diagrams. The expectation is that readers will be able to transfer the knowledge gained here to other cases of interest since, quite clearly, a full treatment of the subject matter is beyond the scope of this article. There are excellent monographs, articles, and compilations that bring together the collective effort of over 50 years of intense research on order-disorder transitions. Some of these are listed below and recommended for further reading.

On a general note, it is perhaps worth while to reiterate two points that hopefully follow from the previous sections. First is the fact that the transition from order to disorder is accompanied by drastic changes in physical properties. Second, the diversity of ordered structures, either seen in nature or yet to be discovered, appear, if not endless, extremely rich. Consequently, understanding and controlling order-disorder transitions will continue to the be a key element in our efforts to tailor physical systems with desired properties.

Works Cited

Henry, N. F. M., Lonsdale, K. (Eds.) (1965), *International Tables for X-Ray Crystallography*, Birmingham, England: Kynoch Press.

Landau, L. D., Lifshitz, E. M. (1969), *Statistical Physics*, 2nd ed., Oxford: Pergamon Press.

Villars, P., Calvert, L. D. (1985), *Pearson's Handbook of Crystallographic Data for Intermetallic Phases*, Warrendale, OH: ASM International.

Further Reading

Krivoglaz, M. A., Smirnov, A. A. (1964), *The Theory of Order-Disorder in Alloys*, London: McDonald.

Ducastelle, F. (1991), "Order and Phase Stability in Alloys," in: F. R. de Boer, D. G. Pettifor (Eds.), *Cohesion and Structure*, Vol. 3, Amsterdam: North-Holland.

Kachaturyan, A. G. (1983), *The Theory of Structural Transformations in Solids*, New York: Wiley.

OSCILLATORS

PAUL D. COLEMAN, *Department of Electrical and Computer Engineering, University of Illinois, Urbana, Illinois, U.S.A.*

Introduction 17
1. **Oscillator Principles** 17
2. **Low-Frequency Triode Transistor Oscillators** 19
2.1 Two-Port Devices 19
2.2 One-Port Devices 21
3. **Microwave Conduction-Current Oscillators** 24
3.1 Reflex Klystrons 24
3.2 Magnetron 25
3.3 Backward-Wave Oscillators 27
3.4 Gyrotron 28
3.4.1 Gyrotron Principles 28
3.4.2 Mathematical Models of the Gyrotron 30
4. **Novel Conduction-Current Oscillators** 30
4.1 High-Energy Free-Electron Oscillator 30
4.2 Coherent Cherenkov Oscillators 32
4.2.1 The Boundary-Value Problem 32
4.2.2 Self-Excited Cherenkov Oscillator 33
5. **Relaxation Oscillators** 34
6. **Polarization-Current (Laser) Oscillators** 34
Works Cited 36

INTRODUCTION

The first generation of electromagnetic radiation was by H. Hertz in 1886, but the electronic age was born with the invention of the triode vacuum tube by L. de Forest in 1906. Vacuum-tube oscillators, in their many forms (triodes, klystrons, magnetrons, gyrotrons, backward-wave tubes, etc.), were the only sources of coherent electromagnetic radiation until the transistor was invented in 1947. In the last 40 years, the transistor and other solid-state devices have replaced the vacuum tube in all applications requiring power levels of less than a few watts at frequencies up to the order of 100 GHz. The vacuum tube cannot compete with the small size, cost, and heat dissipation of the transistor. However, for high-power applications, the vacuum tube is difficult to challenge, with relativistic free-electron oscillators, for example, being able to generate very large amounts of power from microwave to optical frequencies.

The development of vacuum-tube oscillators was driven by the challenge of extending their frequency of operation to higher and higher frequencies. The development of transistor and solid-state oscillators has been driven by circuit integration, cost, size, and heat dissipation. Both the vacuum tube and transistor are conduction-current devices and, hence, share basic physics problems and limitations.

In this article, the point of view taken is that only two types of amplifier oscillators exist: conduction current and polarization current. This point of view permits one to concentrate on the basic oscillator physics of each class rather than to regard various oscillators as fundamentally different.

1. OSCILLATOR PRINCIPLES

Sinusoidal oscillators in general consist of an amplifier with positive feedback and a resonator. Two classes of amplifiers exist: those based upon conduction currents and those based upon polarization currents. From a circuit's point of view, conduction-current devices can be two-port (three-terminal) or one-port (two-terminal), while polarization-current amplifiers are two-terminal devices.

The basic elements of a two-port oscillator are illustrated in Fig. 1. Power conservation for the steady state requires that

$$G\mathcal{P}_i = \mathcal{P}_i + \mathcal{P}_R + \mathcal{P}_F + \mathcal{P}_o, \tag{1}$$

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

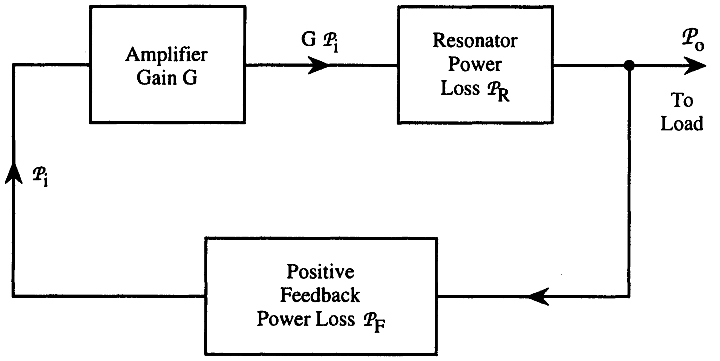

FIG. 1. Generic oscillator circuit.

where $\mathcal{P}_R$ and $\mathcal{P}_F$ are the power losses in the resonator and feedback system, $\mathcal{P}_o$ is the power output to the load, $\mathcal{P}_i$ the input power to the amplifier, and G the power gain. Gain nonlinearity must exist in the amplifier to limit $G\,\mathcal{P}_i$ to satisfy Eq. (1), although it could also be built into the feedback network. The oscillations build up from noise that always exists in a circuit at finite temperature, or from spontaneous radiation in the case of lasers. The two Barkhausen conditions for oscillation are as follows:

1. G must be greater than one; and
2. the phase shifts around the closed loop must be $2\pi n$ for the voltage and/or field to be single valued.

Maxwell's equations

$$\nabla \times \mathbf{E} + \mu_0 \frac{\partial \mathbf{H}}{\partial t} = -\mu_0 \frac{\partial \mathbf{M}}{\partial t}$$

and (2)

$$\nabla \times \mathbf{H} - \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \mathbf{J} + \frac{\partial \mathbf{P}}{\partial t}$$

along with Newton's and Schrödinger's equations are the starting points in determining the sources for electromagnetic (EM) fields and conditions for gain. It is readily seen that only two sources for EM fields exist: conduction currents **J** and polarization terms $\partial\mathbf{P}/\partial t$ and $-\mu_0\partial\mathbf{M}/\partial t$. From the invention of the triode vacuum tube by de Forest (1907) to the invention of the laser in the 1950s, the only amplifiers of EM fields were conduction-current devices. Now that polarization-current ($\partial\mathbf{P}/\partial t$) laser devices have been invented, one may wonder if any new breakthroughs on oscillators in a fundamental sense will be made in the future.

Powers ($\mathcal{P}$) absorbed or generated by a **J** or $\partial\mathbf{P}/\partial t$ device are given by the expressions

$$\mathcal{P} = \int \mathbf{E} \cdot \mathbf{J}\, dv$$

or (3)

$$\mathcal{P} = \int \mathbf{E} \cdot \frac{\partial \mathbf{P}}{\partial t}\, dv,$$

where v is the volume.

Since the coherent field **E** of interest will vary as $e^{j\omega t}$, **J** and/or $\partial\mathbf{P}/\partial t$ must have a Fourier component at ω for the average of $\mathcal{P}$ over a cycle to be different from zero. In particular, **J** and/or $\partial\mathbf{P}/\partial t$ must be out of phase with **E** for power to be generated. This means that in **J** devices the electron beam cannot be a dc beam but must be "bunched" or modulated in some manner. For a $\partial\mathbf{P}/\partial t$ or laser device, inverted populations or nonequilibrium distribution of the laser system must be obtained by means of the appropriate excitation or pumping scheme.

Given a two-level system with an electric dipole transition, it can be shown (Pantell and Puthoff, 1969) that **P** is given by the equation

$$\frac{\partial^2 \mathbf{P}}{\partial t^2} + \frac{2}{T_2}\frac{\partial \mathbf{P}}{\partial t} + \Omega^2 \mathbf{P} = \frac{2\Omega}{3\hbar}|\mu_{12}|^2 \times (N_1 - N_2)\mathbf{E}, \quad (4)$$

where $\hbar\Omega$ is the energy difference of the two levels, μ_{12} is the dipole matrix element (a measure of the coupling of the levels), N is the population, T_2 is a relaxation time, and

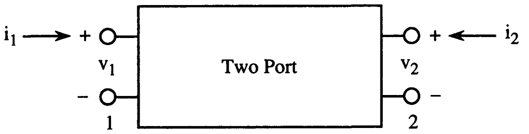

FIG. 2. Two-port circuit.

E is the field. Notice that for **E** sinusoidal, $\partial\mathbf{P}/\partial t$ will be sinusoidal, and that $\partial\mathbf{P}/\partial t$ will change signs when $N_2 > N_1$, the conditions for $\mathcal{P}_{AV}$ to be negative. From this point of view, there was no need to mention the words stimulated emission of radiation.

A principal practical motivation for the study of oscillators has been the quest for high oscillator frequencies, to provide more bandwidth and information capability. Conduction-current devices, with the exception of the free-electron laser, appear to have reached their practical frequency limit around 100 GHz. Lasers ($\partial\mathbf{P}/\partial t$ devices) can generate EM radiation almost across the spectrum from UHF to the ultraviolet.

In the next sections, various types of **J** and $\partial\mathbf{P}/\partial t$ devices will be examined from the point of view of the basic oscillator principles just described.

2. LOW-FREQUENCY TRIODE OR TRANSISTOR OSCILLATORS

2.1 Two-Port Devices

Low-frequency three-terminal, vacuum-tube and transistor oscillators employ lumped-circuit *LCR* circuit elements. These devices can be represented by the two-port network (Millman and Halhias, 1972) displayed in Fig. 2 with the hybrid equations

$$v_1 = h_{11}i_1 + h_{12}v_2, \tag{5}$$

$$i_2 = h_{21}i_1 + h_{22}v_2, \tag{6}$$

where the voltage and current variables represent small variations about the quiescent operating point. The *h*-parameter equivalent circuit, where terminals 1 and 2 are common, can be readily seen to be that given in Fig. 3. While the active elements on the circuits described here are semiconductor devices, earlier circuits used vacuum tubes in similar circuit configurations.

Given the *h* parameter of the amplifying two-port device, one needs to add an appropriate positive-feedback circuit to make an oscillator. The transfer gain A_f with feedback is given by the relationship

$$A_f = A/(1 + \beta A), \tag{7}$$

where β is a complex function of the frequency and *A* is zero-feedback gain. Oscillations can occur when $\beta A = -1$, the Nyquist criterion for instability.

The basic physics is concerned with processes within the "black box" two-port that yields the desirable *h* parameters of Eqs. (5) and (6) and how they might be changed. In both triode vacuum tubes and transistors, gain (or energy addition) is obtained by converting charge kinetic and/or potential energy into EM energy. In triodes and transistors, both dc ($\mathbf{E}_0$) and ac ($\mathbf{E}_1 \sin\omega t$) fields are present in the same region. Hence, the energy relations are from Newton's equation

$$m\mathbf{u}\cdot d\mathbf{u} = q\mathbf{E}_0\cdot d\mathbf{r} - q\mathbf{E}_1 \sin\omega t\cdot d\mathbf{r}. \tag{8}$$

The dc field is performing work on the charge *q*, which in turn is performing work on the ac field.

A key feature of a triode tube or bipolar junction transistor (BJT) transistor is the ability of the grid to control the flow of currents to the plate of the tube or the collector current to control the base current, respectively. By this means, charges are allowed into the interaction region of the device mainly

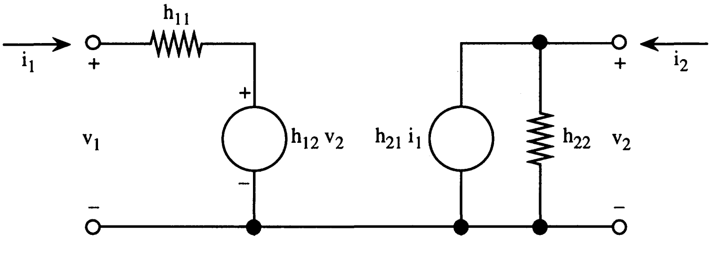

FIG. 3. Equivalent *h*-parameter circuit of a transistor.

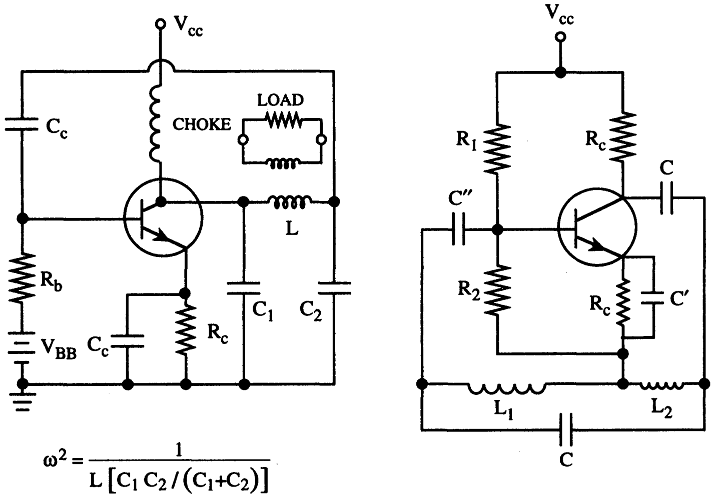

$$\omega^2 = \frac{1}{L\left[C_1 C_2/(C_1+C_2)\right]}$$

(a) Colpitts

(b) Hartley

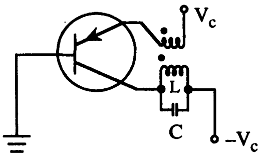

(c) Basic Transistor Oscillator Using L-C Circuit as the Resonator

FIG. 4. Basic Colpitts and Hartley oscillator configurations.

when they can deliver kinetic energy to the ac field.

A number of 180° feedback circuits have been developed to go with vacuum tubes and transistors to make oscillators, two of which have been given distinguishing names, Hartley and Colpitts. These low-frequency *L-C* oscillators (Hunter, 1956) are shown in Fig. 4, along with a simplified circuit to illustrate better the 180° feedback feature.

A phase shift of 180° in the positive-feedback circuit of an amplifier-oscillator can also be provided by a *RC* circuit as illustrated in Fig. 5. The loop gain condition yields the results

$$\omega_0 = [RC\sqrt{6 + 4R_C/R}]^{-1}$$

and (9)

$$\frac{R}{R_C} = \frac{h_{21} - 23}{58} > 0.$$

Piezoelectric quartz crystals with characteristic resonant frequencies can also be used in positive-feedback amplifier circuits to make an oscillator as shown in Fig. 6. Crystal oscillators can be made to be very frequency stable, having been used as frequency standards for many years.

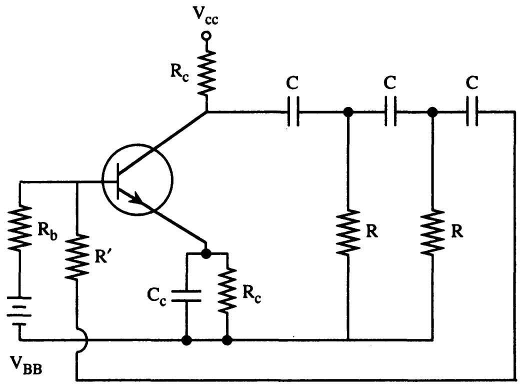

FIG. 5. Phase-shift oscillator circuit.

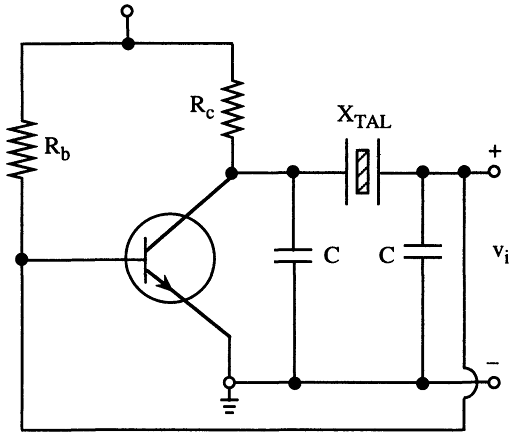

FIG. 6. Crystal oscillator circuit.

2.2 One-Port Devices

A number of two-terminal, one-port conduction-current devices have been discovered that display a negative ac resistance based upon a variety of physical processes. Their *I-V* characteristics are of the general form displayed in Fig. 7. Over the voltage range $V_1 < V < V_2$, assuming a linear relationship, it is seen that

$$I = a - V/r$$

and (10)

$$\frac{dV}{dI} = -r,$$

i.e., the device displays a negative differential resistance *r*.

To see, from a circuit point of view, how this diode amplifies, consider the simplified circuit of Fig. 8. The ac circuit equations are

$$e_s = i_D R + e_D = i_D(R - r)$$

with (11)

$$e = i_D R = e_s \frac{R}{R - r},$$

where e_s is the ac voltage signal and e the voltage output. Clearly, the output voltage e can exceed the input voltage e_s.

If the ac current and voltage amplitudes exceed the values $I_1 - I_0 = \Delta I$ and $V_2 - V_0 = \Delta V$, losses will occur to limit their values. The currents I_1 and I_2 and voltages V_1 and V_2 define the limits of the negative differential resistance region. Hence, the maximum ac power available from the diode is

$$\mathcal{P}_{AV} \le \tfrac{1}{2}(\Delta I)(\Delta V). \qquad (12)$$

The two-terminal devices of major interest are (Coleman, 1982; Liao, 1985) the Esaki tunnel diode, the Gunn diode, the IMPATT diode, and the resonant tunneling diode.

The Esaki diode consists of a heavily doped *p-n* junction where the Fermi levels are located within the bands themselves. As seen from Fig. 9, when a positive bias is applied electrons can tunnel from the filled electron states on the *n* side to the empty states on the *p* side at constant energy. With further

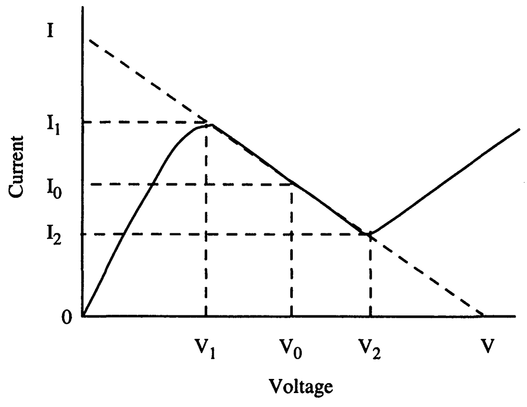

FIG. 7. Schematic *I–V* curves of a negative differential resistance device.

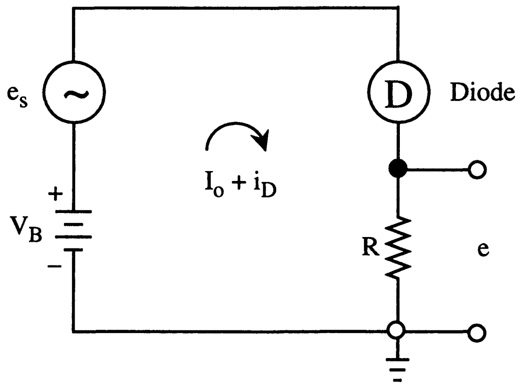

FIG. 8. Negative differential resistance diode amplifier circuit.

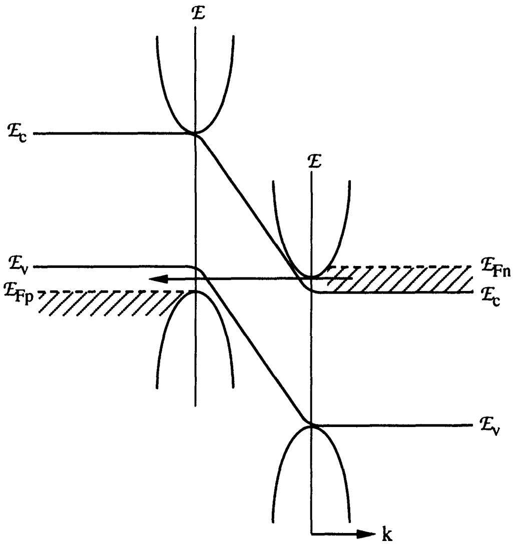

(a) Esaki Diode - Forward Bias

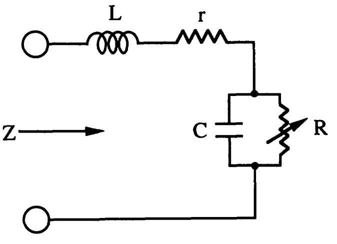

(b) Equivalent Circuit

FIG. 9. Esaki diode characteristics.

positive bias, the current decreases since there are fewer available states opposite the filled states. While the tunneling process is very fast, the negative resistance of the diode goes to zero at the frequency f_c given by the relationship

$$f_c = \frac{\sqrt{|R|/r - 1}}{2\pi C|R|}, \tag{13}$$

as can be determined from the equivalent circuit. With typical values for the circuit parameters, $f_c \sim$ 10–40 GHz.

The resonant tunnel diode (RTD) is a heterostructure with a double-barrier single-quantum well as illustrated in Fig. 10. Maximum current (resonant tunneling) occurs when the energy of the tunneling electrons is approximately in line with the quantized energy level for electrons in the quantum well. The equivalent circuit for the RTD is the same as that for the Esaki diode, but the cutoff frequency f_c can be made much greater, exceeding 100 GHz. Unfortunately, the product $(\Delta I) \times (\Delta V)$ of Eq. (12) is small so that only small amounts of power can be produced with this diode.

The Gunn diode is a bulk semiconductor electron-transfer device wherein electrons are transferred from the central valley of the material (i.e., GaAs) to an upper satellite valley where their velocity $(\hbar^{-1}d\mathscr{E}/dk)$ decreases when an electric field is applied to the semiconductor. This results in a negative differential (du/dE) velocity–field curve for the electric field greater than E_T as illustrated in Fig. 11. The usual semiconductor equations for the current J, electric field E, and electron density n (for one dimension),

$$J = qnu(E) - q\frac{\partial D(E)n}{\partial x} + \epsilon\frac{\partial E}{\partial t}, \tag{14}$$

$$\frac{\partial n}{\partial t} + \frac{1}{q}\frac{\partial J}{\partial x} = 0, \tag{15}$$

$$\frac{\partial E}{\partial x} = \frac{q(n - n_0)}{\epsilon}, \tag{16}$$

apply with the electron velocity u now a function of E. In general, these equations can only be solved by numerical methods. It is found that a number of modes of behavior exist (Copeland, 1967), viz., a transit-time dipole-layer mode, a limited space-charge accumulation mode, and a quenched dipole-layer mode.

Consider a dipole layer, illustrated in Fig. 12, moving from the negative to positive electrode of the semiconductor. The electric field inside the dipole is greater than the fields on either side. Thanks to the velocity property of the material, the electrons to the right of the dipole are moving more rapidly (running away from the dipole) leaving a depletion region while the electrons to the left of the dipole are catching up to form an accumulation region. The dc electron beam is being "bunched" to form the necessary ac current to drive the oscillator, with the frequency $f = u/L$. The dipole layer moves

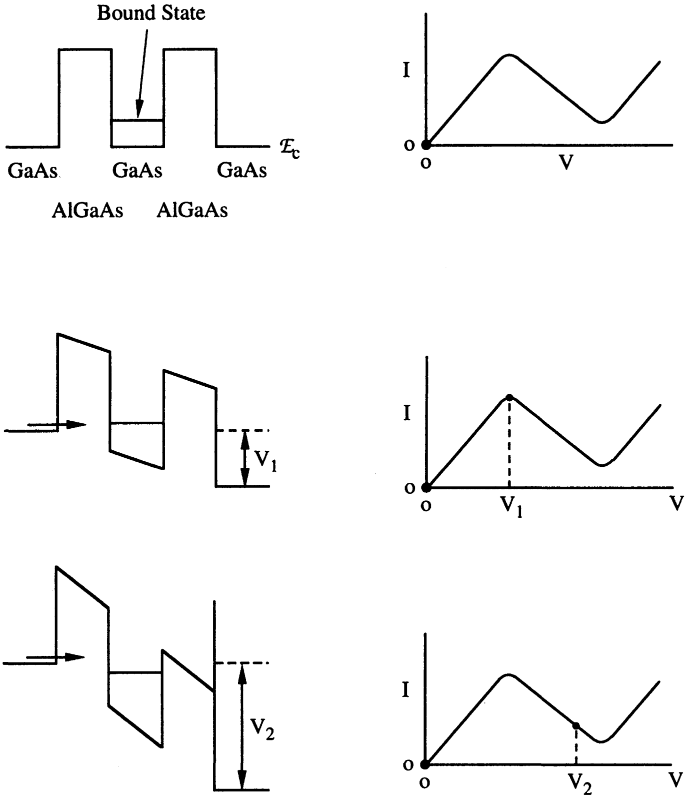
Bound State
GaAs
GaAs
GaAs
AlGaAs
AlGaAs
ℰc
I
o
o
V
V1
V2

FIG. 10. Resonant tunneling diode characteristics.

through the crystal and disappears at the anode. The electric field then rises to form a new dipole layer to repeat the process. Gunn-diode oscillators can yield several watts of pulsed power at frequencies as great as 100 GHz.

An IMPATT diode is an n^+pip^+ semiconductor structure shown schematically in Fig. 13. It consists of a narrow p avalanche region

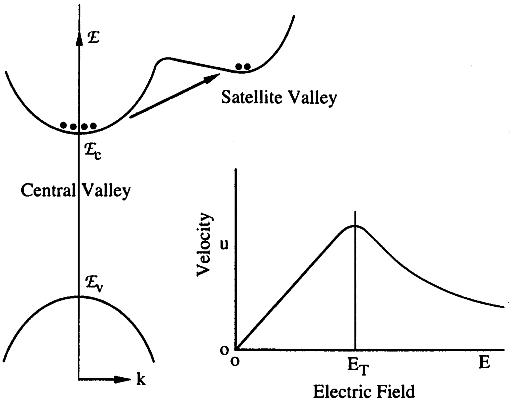
ℰ
Satellite Valley
ℰc
Central Valley
ℰv
k
Velocity
u
o
o
E_T
E
Electric Field

FIG. 11. Gunn diode characteristics.

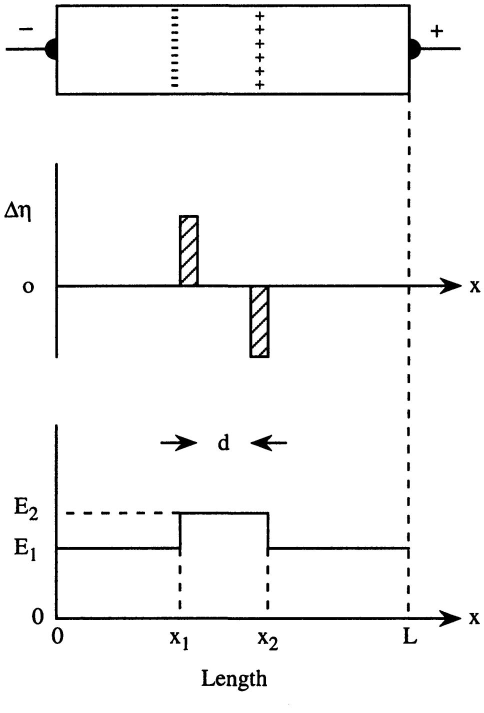
−
+
Δη
o
x
d
Field
E2
E1
0
0
x1
x2
L
x
Length

FIG. 12. Schematic of dipole drift in a Gunn diode.

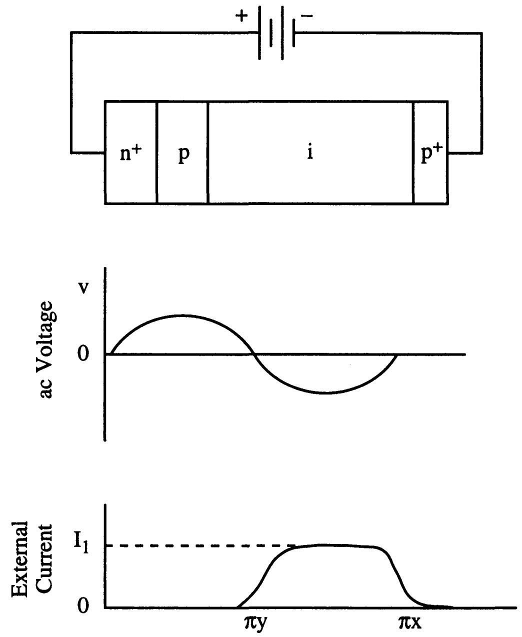

FIG. 13. Voltage and external current in an IMPATT diode.

where carrier multiplication by impact ionization occurs and a drift region where the carriers drift at saturated velocity. A phase-shift delay is caused by the avalanche process and the finite transit time of the carriers drifting through the drift region.

The average power for the wave forms shown is given by the expression

$$\begin{aligned}\mathcal{P}_{AV} &= \frac{1}{2\pi}\int_{y\pi}^{x\pi} I_1V_1 \sin\omega t(d\omega t)\\ &= \frac{V_1I_0}{\pi}\left(\frac{\cos y\pi - \cos x\pi}{x-y}\right),\end{aligned} \qquad (17)$$

where $I_1 = 2I_0/(x - y)$ and I_0 is the dc current. Maximum power produced by the diode is seen to be

$$(\mathcal{P}_{AV})_{\max} = -\frac{V_1I_0}{\pi}\left(\frac{1 + \cos(\pi x)}{x - 1}\right). \qquad (18)$$

The avalanche process contributes a 90° inductive phase lag to the current generated in the avalanche region. This current is then injected into the drift region where an additional 90° phase in the external current occurs, thus resulting in the ac voltage and current being 180° out of phase. The diode will oscillate at a frequency and level where the diode plus external circuit admittance is zero.

3. MICROWAVE CONDUCTION-CURRENT OSCILLATORS

The oscillators discussed thus far were essentially lumped-circuit devices with *LCR* positive-feedback circuits, although Esaki and resonant tunneling diodes plus Gunn and IMPATT diodes could be embedded into transmission lines and microwave resonators. At microwave and millimeter-wave frequencies, propagation effects are present that require wave and field analysis to describe the amplification process. For tube or free-electron devices, methods need to be employed to cope with circuit reactance, circuit-tube loss, and electron transit-time limitations to obtain these higher frequencies.

3.1 Reflex Klystrons

The reflex klystron (see Fig. 14) consists of an electron-emitting cathode, a current-control grid, an anode-reentrant microwave resonator, and a repeller electrode. Electrons injected from the cathode are first velocity

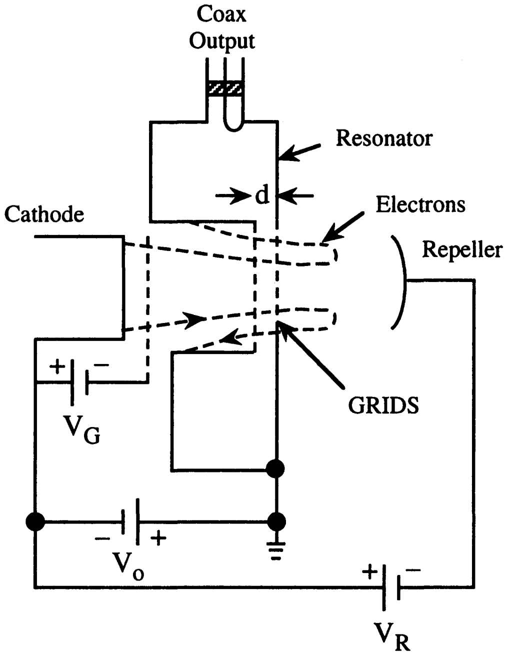

FIG. 14. Schematic diagram of a reflex klystron.

modulated by the cavity–gap field where the first electrons are decelerated and the later electrons accelerated to play a "handicap race" so as to yield a bunched ac electron beam. The electrons are then turned around by the repeller field and pass through the cavity–gap in the reverse direction 180° out of phase with the resonator field to give up a portion of their kinetic energy.

The determination of the ac currents in the bunched electron beam is obtained from Newton's equation as follows:

$$\int_{u_0}^{u} du = \frac{qE_1}{m} \int_{t_a - T/2}^{t_a + T/2} e^{j\omega t} dt, \tag{19}$$

where the transit time T across the resonator gap is given by $u_0 T \simeq d$ where d is the gap spacing, u_0 the dc beam velocity, and t_a the time the electrons reach the midpoint of the gap. Integration of Eq. (19) yields

$$\begin{aligned} u &= u_0 \left[1 + \frac{2V_1}{V_0} \left(\frac{\sin\theta/2}{\theta/2} \right) e^{i\omega t_a} \right] \\ &= u_0 [1 + 2\alpha A e^{i\omega t_a}], \end{aligned} \tag{20}$$

where $V_1 = E_1 d$, $m u_0^2 = 2qV_0$, $\theta = \omega T = \omega d/u_0$, $\alpha = V_1/V_0$, and $A = \sin\frac{1}{2}\theta/\frac{1}{2}\theta$. The electrons are returned to the resonator gap at time t_b given by

$$t_b = t_a + \frac{\ell}{u} \simeq t_a + \frac{\ell}{u_0} \left(1 - \frac{A\alpha}{2} \sin\omega t_a \right). \tag{21}$$

It is seen from Eqs. (20) and (21) that the transit time contained in the factor A is a key limitation in extending the frequency of a conduction-current device.

Finally, from conservation of electrons, the return current is

$$I_b = I_o \left| \frac{dt_a}{dt_o} \right| = I_0 \left| \frac{1}{1 - k \cos\omega t_a} \right|, \tag{22}$$

where $k = (A\alpha/2)(\omega d/u_0) = (A\alpha/2)(\omega T_0)$.

The current I_b is a periodic function, which can be analyzed into its Fourier components to yield for the fundamental

$$I_{b_1} = j2 I_0 J_1(k)\, e^{-j\omega t_o}, \tag{23}$$

where $J_1(k)$ is a first-order Bessel function, and the power generated,

$$\mathcal{P}_{AV} = V_1 I_0 J_1(k) \sin\omega t_o. \tag{24}$$

A reflex klystron can be frequency tuned mechanically by varying the gap spacing d, and electronically of the order of 1% by varying the repeller voltage. Reflex klystrons can be made to operate up to the order of 100 GHz. Power levels of the order of 10 MW can be obtained in the microwave range. For many years, they have been very useful oscillators for laboratory measurements, local oscillators in radars, linear accelerators, and missiles. Harmonic signals obtained from reflex klystrons driving Schottky diodes provided the first coherent millimeter–submillimeter radiation for the microwave spectroscopy of molecules. Microwave spectroscopy made key contributions to the invention of the NH_3 laser.

3.2 Magnetron

The magnetron (Coleman, 1982; Liao, 1985) is a crossed-field ($\mathbf{E} \cdot \mathbf{B} = 0$), termed M-type electron tube that was invented by A. W. Hull in 1921, although it was not developed into a high-power generator until World War II when the need for radar existed. Its high efficiency of 50%–80% makes it useful for many commercial uses such as in a microwave oven. It can produce pulsed powers as large as 10 MW in the microwave range, and it can be made to operate at frequencies approaching 100 GHz.

There are a number of different magnetron structures: a split-anode, a cylindrical traveling-wave slot, a cylindrical traveling-wave hole and slot, a rising sun, and a linear slotted magnetron, three of which are shown in Fig. 15. The electrons are tightly bunched when they are emitted from the large, high-current cathode, which leads to the high efficiency of the tube.

The fields in the interaction space are not simple and the charge motion, including space charge, is very complicated so that an analytical solution for a magnetron oscillator is not possible. Some of the features of a magnetron can be seen by considering a linear magnetron configuration.

Consider first the motion of electrons in crossed dc ($\mathbf{E} \cdot \mathbf{B} = 0$) fields where $\mathbf{E} = \hat{\mathbf{y}}\mathbf{E}$ and $\mathbf{B} = \hat{\mathbf{z}}\mathbf{B}$. Then Newton's equations are

$$\ddot{x} = -\omega_c \dot{y}$$

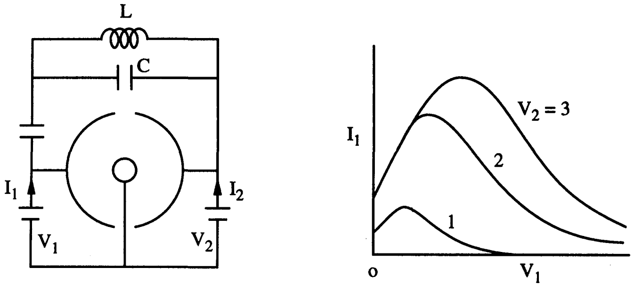

(a) Circuit Configuration (b) Current-Voltage Characteristics

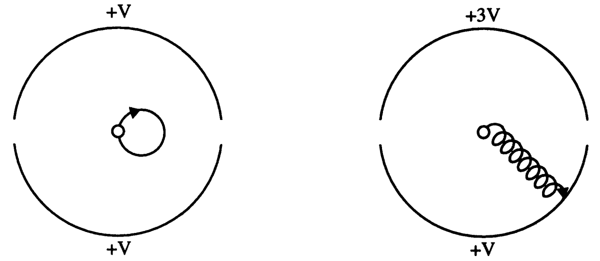

(c) Electron Paths in Split and DE Magnetron

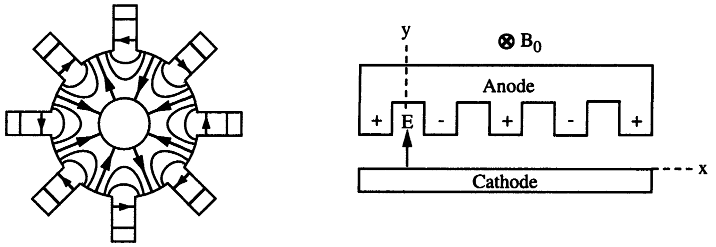

(d) Lines of Force in a Cylindrical Magnetron (e) Linear Magnetron Configuration

FIG. 15. Various magnetron configurations.

and

$$\ddot{y} = -eE/m + \omega_c\dot{x}, \tag{25}$$

where $\omega_c = eB/m$, with the solution for $\dot{x}$ being

$$\dot{x} = -\omega_c R \sin(\omega_c t - \alpha) + E/B \tag{26}$$

and

$$y = y_0 + R \sin(\omega_c t - \alpha). \tag{27}$$

Equation (26) suggests that to realize an oscillator a travelling-wave field moving with the drift velocity E/B should be used. Since E/B will be much less than the velocity of light c, a heavily loaded slotted circuit, shown in Fig. 15(e), will be needed to reduce the phase velocity to E/B.

Adding rf fields to Eq. (20) results in a problem too difficult to solve analytically. Hence, a word description of the results will be given. Electrons with certain phases drift toward the cathode while other electrons in other phases drift toward the anode and do work in the rf field. Here the kinetic energy stays approximately constant ($E/B \sim$ constant), but the full potential energy is delivered to the rf field. Only a small amount of energy is lost to the cathode.

For a cylindrical magnetron, Newton's equations become in polar coordinates (r,θ)

$$\ddot{r} - r\dot{\theta}^2 = -\omega_c r\dot{\theta} + \frac{e}{m}\frac{dV}{dr}, \tag{28}$$

where $E_r = -dV/dr$, and

$$\frac{1}{r}\frac{d}{dt}(r^2\dot{\theta}) = \omega_c\dot{r}, \tag{29}$$

where $\omega_c = eB/m$. Given that the radial and angular velocity are zero at the cathode where $r = r_c$, Eq. (29) yields

$$\dot{\theta} = \tfrac{1}{2}\omega_c\,[1 - (r_c/r)^2]. \tag{30}$$

Now the energy equation is

$$\tfrac{1}{2}m\,(\dot{r}^2 + r^2\dot{\theta}^2) = eV. \tag{31}$$

Eliminating $\dot{\theta}$ between Eqs. (30) and (31) and using the condition $\dot{r} = 0$ for $r = r_p$ yields the cutoff voltage

$$V_c = \frac{er_p^2}{8m}B\left[1 - \left(\frac{r_c}{r_p}\right)^2\right]^2, \tag{32}$$

the voltage below which no electrons will reach the plate. Oscillations are usually observed in the vicinity of cutoff.

The slot magnetron structure of Fig. 15(d) is a resonator with certain allowed frequencies, which can be found by requiring the phase shift around the circumference to be $2\pi m$ to bring the field back to its initial value. Hence, given the phase factor β_0 of the rotating electromagnetic wave, it is seen that

$$\beta_0 NL = 2\pi m, \tag{33}$$

where N is the number of slots, L the periodic distance, and m an integer.

The mode for $m = N/2$ where N is even is termed the π mode of the resonator. Other modes are double in that they allow traveling waves to propagate in either of two possible directions around the magnetron. The large currents in a magnetron require that space charge be included in the analysis of charge bunching, again a problem that cannot be solved analytically. Numerical calculations yield the results that the electrons bunch into a configuration like the spokes of a wheel (see Fig. 16) that rotate in synchronism with the tangential electric field.

Magnetrons are used in a pulsed mode with pulse lengths typically 1–10 μs at pulse repetition frequencies of the order of 100–400 Hz. For microwave oven applications, they can be operated for many minutes at power levels from 400 to 800 W.

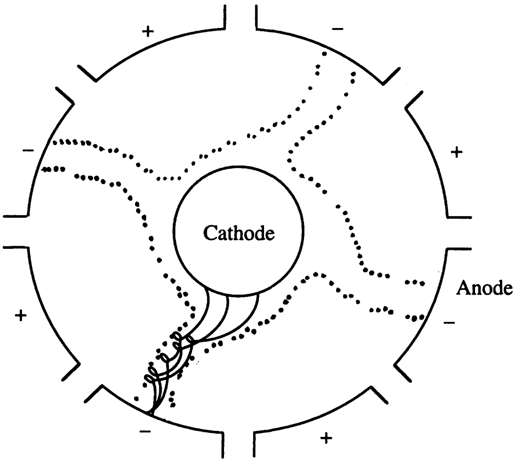

FIG. 16. Schematic of electron trajectories in a cylindrical magnetron.

3.3 Backward-Wave Oscillators

In order for a linear electron beam to interact with an electromagnetic traveling wave, be bunched, and deliver energy to the field, the electron velocity u_0 must be slightly greater than the phase velocity v_p of the wave. Since the normal transmission line or waveguide has a phase velocity $v_p \geq c$ where c is the velocity of light, some means must be found to reduce the phase velocity by a substantial amount if relativistic beam voltages are not used (Karp, 1955; Kompfner, 1947). Phase-velocity reduction of a waveguide can be made by making the structure periodic as illustrated in Fig. 17. What is also needed is a transverse-magnetic (TM) mode of propagation, where an axial field E_z exists on the axis and the transverse field is zero on the axis.

Periodic loading of the waveguide changes the ω vs β characteristics as shown in Figs. 17(c) and 17(d) so that the loaded guide now takes on filter stop-band properties. The calculation of β as a function of ω is a difficult task, but some general results can readily be obtained. Since the field is spatially periodic, one can expand the axial electric field in a

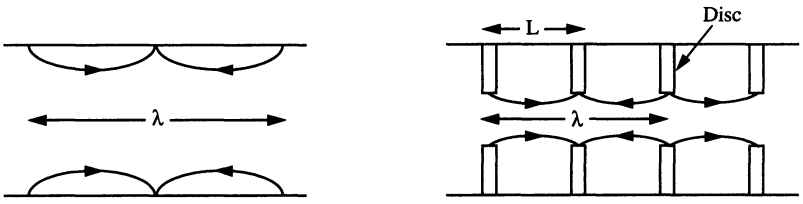

(a) Unloaded Guide

(b) Periodically Loaded Guide

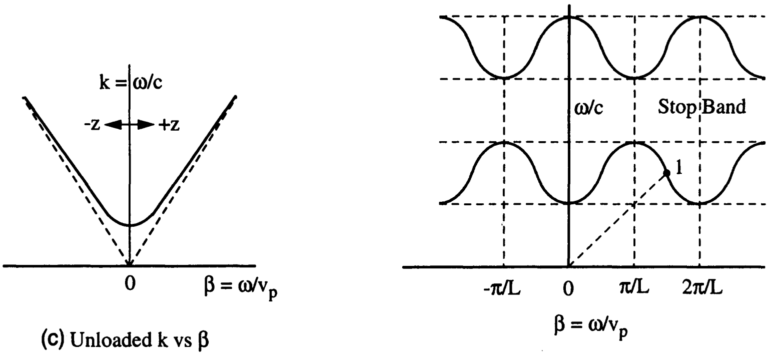

(c) Unloaded k vs β

(d) Loaded k vs β

FIG. 17. Characteristics of unloaded and periodically loaded waveguides.

Fourier series:

$$E_z = \sum_n E_n \exp\{j[\omega t - (\beta_0 + 2\pi n/L)z]\}$$
$$= \sum_n E_n \exp[j(\omega t - \beta_n z)], \qquad (34)$$

where

$$\beta_n = \beta_0 + 2\pi p/L = \omega/v_{pn}, \qquad (35)$$

i.e., the phase velocity will get smaller as n increases.

It will also be observed from Fig. 17(d) that for $n = 1$ (the first spatial harmonic), the group velocity v_{gr} is negative:

$$\frac{1}{v_{gr}} = \frac{d\omega}{d\beta_n} = \frac{d\omega}{d\beta_0} < 0$$

for operation at point 1 on the ω vs β curve.

The gain analysis of a backward-wave oscillator (BWO) is the same as a traveling-wave-tube amplifier (TWT). However, since the group velocity of the spatial harmonic field being used is negative, the power output is taken at the cathode end of the tube as illustrated in Fig. 18. The feedback is internal in the structure. If the beam velocity ($u_0 \sim v_p$) is changed, the frequency of the BWO is changed as point 1 moves along the ω vs β curve.

Backwave oscillators (also termed O-Carcinotrons), as developed by French scientists (Convert *et al.*, 1959), have reached 1 THz in frequency. These nonrelativistic beam tubes are strong competitors as continuous sources of millimeter and submillimeter waves, having achieved 1 W cw at 300 GHz with a beam voltage of the order of 10 kV.

3.4 Gyrotron

3.4.1 Gyrotron Principles The requirement of a synchronization of the electron velocity u_o and phase velocity v_p in a TWT or

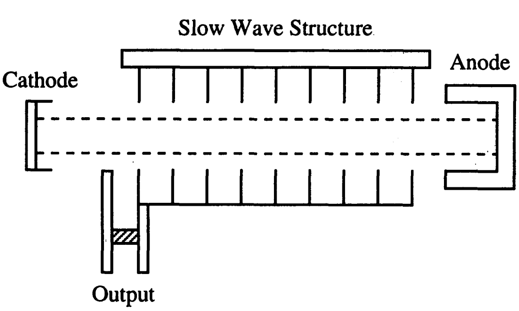

FIG. 18. Backward-wave oscillator schematic.

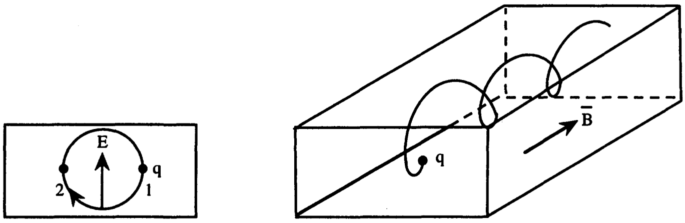

(a) End View of Charge Motion (b) Charge Motion in Waveguide

FIG. 19. Quasi-velocity matching.

BWO meant that the waveguiding structure of the tube had to be periodically loaded to achieve the condition $u_o \simeq v_p$. Loading the waveguide reduced its cross-sectional area. As the frequency of the tube was extended, the periodically loaded guide became uncomfortably small and the electron beam had to approach the loading elements more closely to maintain efficiency. What was necessary to solve this problem was to remove the periodicity from the waveguide and introduce the periodicity on the electron beam (Sprangle and Drobot, 1977). A quasisynchronism between the wave and beam is achieved by allowing the wave to skip past the beam, with the interaction taking place with different "parts" of the wave in consecutive cycles.

For example, consider a spiral electron beam in a field $\mathbf{B} = \hat{\mathbf{z}}B$ in a TE-mode waveguide as illustrated in Fig. 19(b). At point 1 of Fig. 19(a), the charge is doing work on the field. When the charge rotates with angular velocity $\omega_c = qB/m$ to position 2, the field $\mathbf{E}$ needs to reverse directions to continue the interchange. Hence, it can be seen that

$$t_{12} = \frac{\pi}{\omega_c} = \frac{d}{u_z} = \frac{d + \lambda q/2}{v_p} = \frac{d + \pi v_p/\omega}{v_p}; \tag{36}$$

solving for ω, one obtains a Doppler frequency relation

$$\omega = \frac{\omega_c}{1 - u_z/v_p} \tag{37}$$

for the quasisynchronous condition, where d is the distance traveled in the z direction with velocity u_z.

The electrons follow a helical trajectory around the lines of the axial dc magnetic field inside a TE-mode waveguide ($v_p > c$) as illustrated in the gyrotron schematic of Fig. 20. In the case of a linear, nonperiodic beam, the electron bunching is longitudinal. For the gyrotron with its periodic electron beam, the electron bunching is transverse or azimuthal phase bunching as indicated in Fig. 21. This bunching occurs because the cyclotron frequency qB/m depends on the relativistic mass $m = \gamma m_o$. It is not possible to solve the relativistic trajectory equations in closed form. However, the "handicap race" can readily be seen. The electron at position 1 is losing kinetic energy to the field; hence, it rotates faster with a smaller radius trying to catch up with electron at position 2. The electron at position 3 is gaining kinetic energy from the field and rotates more slowly with a larger radius, allowing electron 2 to catch up in position. At a later period of time, electrons 1, 2, and 3 will be lined up in azimuth.

The gyrotron shown in Fig. 20 uses a hollow electron beam and a TE_{01}-mode waveguide to obtain an E_ϑ field so that there are many orbits of the electrons as displayed in Fig. 21(b). Also, the electron beam must have most of its energy transverse to the axis of the tube (typically by a factor of 1.5–2).

In 1982, Temkin *et al.* (1982) built a gyrotron at 140 GHz with a pulsed power of 100

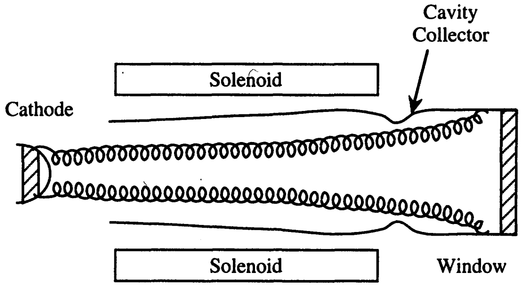

FIG. 20. Gyrotron circuit configuration.

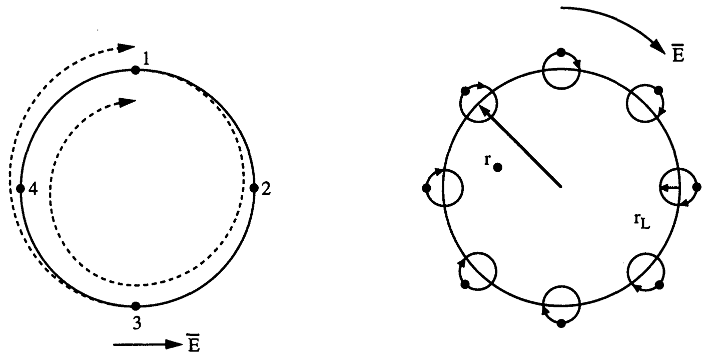

(a) Single Orbit

(b) Multiple Orbits

FIG. 21. Transverse electron bunching in a gyrotron.

kW and efficiency of 17% using a TE_{031}-mode resonator, 80-kV beam, and 5.4-T magnetic field. At lower frequencies (~30 GHz), scientists of the former Soviet Union achieved efficiencies of over 40%. For high-power pulsed operation, the gyrotron is in a class by itself.

3.4.2 Mathematical Models of the Gyrotron In the treatment of oscillators in this article, it was implied that conduction-current (**J**) devices could be analyzed by solving a Newton problem for the current and Maxwell's equations for the gain and fields. Of course, any classical problem can be quantized, even an *LC* circuit, but quantum effects would be indeed difficult to observe at 100 MHz. On the other hand, a polarization-current ($\partial\mathbf{P}/\partial t$) device requires a quantum treatment since the polarization **P** cannot be calculated classically. In this article, a $\partial\mathbf{P}/\partial t$ device has been called a laser and a **J** device a free-electron oscillator irrespective of how it is analyzed. The mathematical model cannot change the physics of the problem.

From this perspective, the gyrotron has had an interesting history. In the initial work to determine the operating principles of the gyrotron, a quantum-mechanical approach was used and the phase "electron cyclotron maser" was coined. It was implied that a gyrotron was a $\partial\mathbf{P}/\partial t$ device with observable quantum effects. While this is an interesting point of view, a quantum-mechanical model contributes little to the detailed analysis of the gyrotron interaction process.

4. NOVEL CONDUCTION-CURRENT OSCILLATORS

4.1 High-Energy Free-Electron Oscillator

In this review, oscillators have been classified as (1) conduction- or (2) polarization-current devices. The high-energy or relativistic free-electron oscillator has been termed a laser in the literature in spite of its being a conduction-current oscillator. Basically, the oscillator consists of an accelerator to produce a pulsed electron beam, a "wiggler" magnet to force the electrons to oscillate, and a Fabry–Pérot mirror system, as shown in Fig. 22. Because of relativistic effects, the radiation is strongly forward directed and Doppler shifted.

The wavelength λ_d of the radiation can be approximately calculated as seen in Fig. 22(b), by observing that while the radiation emitted by an electron at position 1 reaches position 3, the electron will have moved from position 1 to 2. Both the electron and wave will have undergone one cycle.

The "wiggler" magnet concept first appeared in the dissertation of P.D. Coleman (1951). Phillips (1960) developed the periodic magnetic field idea with his invention of the Ubitron, and in 1970 Madey (1971) extended the concepts to the optical range.

If the Fabry–Pérot mirror were not present, the electron emission would be strictly incoherent. With an optical field present to provide feedback, the electrons can be accelerated and decelerated slightly so that the faster electrons (more mass) can catch up with the slower electrons (less mass) to form

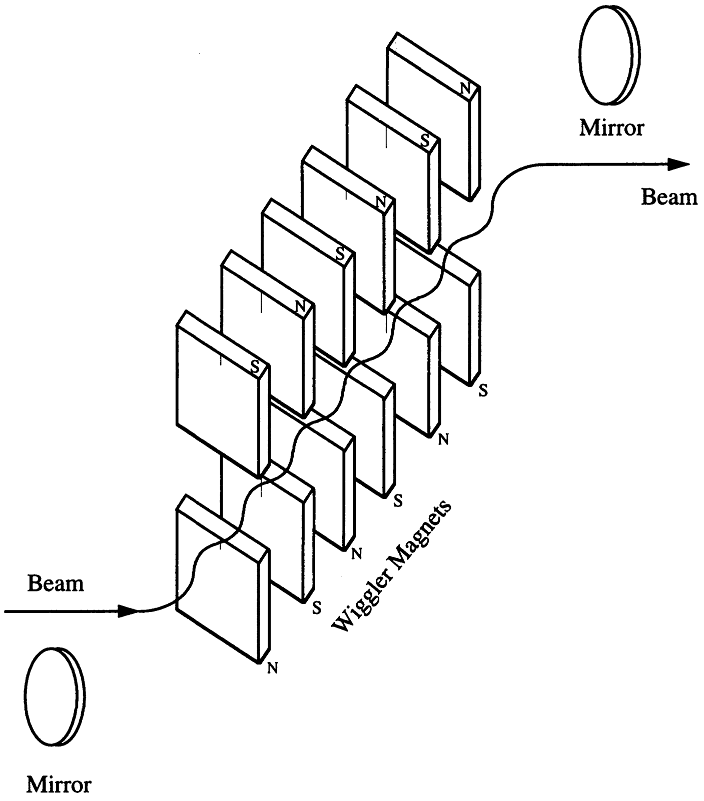
Mirror
Beam
N
S
N
S
N
S
S
N
S
N
S
N
Wiggler Magnets
Beam
Mirror

(a) Free Electron Oscillator Schematic

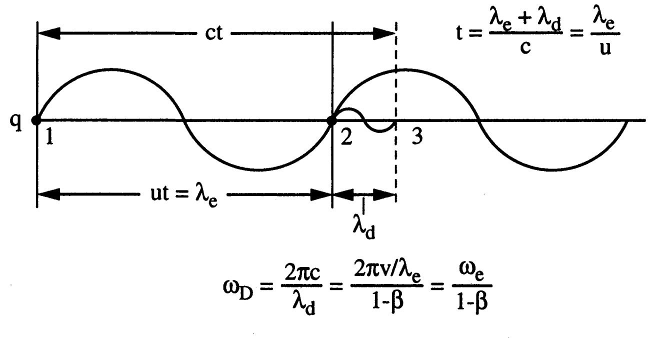
ct
$t = \frac{\lambda_e + \lambda_d}{c} = \frac{\lambda_e}{u}$
q
1
2
3
$ut = \lambda_e$
λ_d
$\omega_D = \frac{2\pi c}{\lambda_d} = \frac{2\pi v/\lambda_e}{1-\beta} = \frac{\omega_e}{1-\beta}$

(b) Doppler Frequency Shift

FIG. 22. Schematic drawing of a free electron oscillator.

bunches spaced at optical wavelengths (Parker *et al.*, 1982).

The relativistic equation of motion is

$$\frac{d}{dt}(m_o\gamma\mathbf{u}) = -e[\mathbf{E} + \mathbf{u} \times \mathbf{B}] + \mathbf{F}, \qquad (38)$$

where $\gamma = (1 - \beta^2)^{-1/2}$; $\beta = u/c$, $\mathbf{F}$ is the self-force, and $\mathbf{B} \simeq \mathbf{B}_0 \sin(2\pi x/\lambda_e)$. If the self-force is neglected, then

$$m_o c^2 \frac{d\gamma}{dt} = -e\mathbf{E} \cdot \mathbf{u} = \mathbf{u} \cdot \frac{d}{dt}(m_o\gamma\mathbf{u}), \qquad (39)$$

where $\mathbf{E}$ is the transverse electromagnetic (TEM) optical field in the resonator.

While the electron velocity is modulated, the greater change is in the modulation of the electron mass $m_o\gamma$. The electrons and wave are in resonance on the forward pass and out of resonance on the backward pass. It is a complex problem to calculate the velocity $\mathbf{u}$ and acceleration $\dot{\mathbf{u}} = d\mathbf{u}/dt$ of the charge e because of the relativistic equations.

The power radiated can be estimated from Larmor's formula (Stratton, 1941)

$$\begin{aligned}\mathcal{P} &= \frac{Q^2\dot{u}^2}{6\pi\epsilon c^3} \frac{1 - \beta^2 |(\mathbf{u}/u) \times (\dot{\mathbf{u}}/\dot{u})|^2}{(1 - \beta^2)^3} \\ &\to \frac{Q^2\dot{u}^2}{6\pi\epsilon c^3(1 - \beta^2)^2},\end{aligned} \qquad (40)$$

where if the electron bunch is small compared to the radiation wavelength $Q^2 = (ne)^2$, when n is the number of electrons in the bunch.

Pulse power levels (Szöke, 1981; Granatstein and Roberson, 1985; Brau and Newnam, 1987; Süsskind, 1988) in a free-electron oscillator can be enormous compared to other types of conduction-current devices. Robenson *et al.* (1987) have reported a peak laser power output of 6 × 10^8 W at a wavelength of 514 nm with a free-electron oscillator driven with a 1.3-GHz linear accelerator having the pulsed current output shown in Fig. 23. The efficiency of this oscillator is seen to be

$$\begin{aligned}\frac{\mathcal{P}_{\text{LASER}}}{\mathcal{P}_{\text{BEAM}}} &= \frac{6 \times 10^8}{100(120 \times 10^6)} \\ &= \frac{0.6\ \text{GW}}{12\ \text{GW}} = 0.05.\end{aligned} \qquad (41)$$

Power levels in the optical resonator can readily reach levels that damage the metal mirror of the Fabry–Pérot system.

It is tempting to use Larmor's formula [Eq. (40)] to estimate the peak power. The peak acceleration $\dot{u}$ can be approximated from the relation

$$\dot{u} \simeq c(qB/m). \qquad (42)$$

Then

$$\mathcal{P} = \frac{(nq)^2(qB/m_o)^2}{6\pi\epsilon c(1 - \beta^2)^2} \simeq 10^{-12}n^2\ \text{W}. \qquad (43)$$

The number of electrons in the 100-A, 20-ps pulse is 1.25 × 10^{10}; hence, $\mathcal{P} \simeq 0.16$ GW at the peak magnetic field.

4.2 Coherent Cherenkov Oscillators

4.2.1 The Boundary-Value Problem The generation of electromagnetic radiation (oscillators) can be based upon a number of physical phenomena. In the case of classical EM theory described by Maxwell's equations, it is the solutions to the equations, not the equations themselves, that provide a

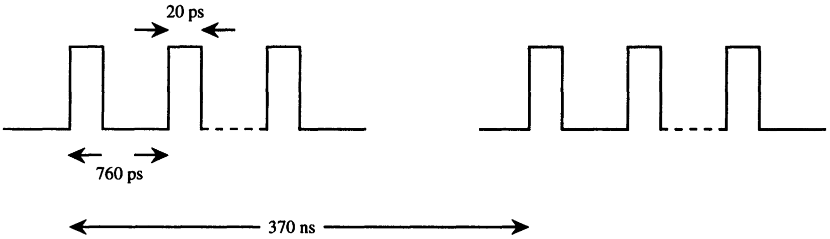

FIG. 23. Current pulses from accelerator driving the free electron oscillator.

mathematical description of the physical phenomena. From this point of view, the last fundamental phenomenon derived from Maxwell's equations was the Cherenkov effect, viz., a charge moving through a polarizable medium with a *constant* velocity faster than the medium's phase velocity will radiate. The Cherenkov effect can be studied by considering the excitation of the elementary dipoles by the passing charge or as a straightforward Maxwell boundary-value problem. It is not necessary for the charge to pass through the medium, just close to it.

While the Cherenkov effect was studied extensively by P. A. Cherenkov from 1924 to 1938 and a theory was constructed by Frank and Tamm (1937), coherent Cherenkov radiation in the microwave range was not demonstrated until 1960 by Coleman and Enderby (1960).

Consider the boundary-value problem of a bunched (ac) electron beam passing through a cylindrical hole in a dielectric as illustrated in Fig. 24. The equation for the vector potential $\mathbf{A} = \hat{\mathbf{z}}A$ is

$$\nabla^2\mathbf{A} - \mu\epsilon\frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu\mathbf{J}_n e^{jn\omega_0(t-z/u)} \tag{44}$$

for the nth harmonic current density $\mathbf{J}_n$.

In region II outside of the beam, Eq. (44) becomes

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{dR}{dr}\right) + \left(\frac{n\omega_0}{v}\right)^2[\mu\epsilon u^2 - 1]R = 0, \tag{45}$$

where $A_z = R(r)e^{jn\omega_0(t-z/u)}$. If

$$\mu\epsilon u^2 = (u/v_p)^2 > 1 \tag{46}$$

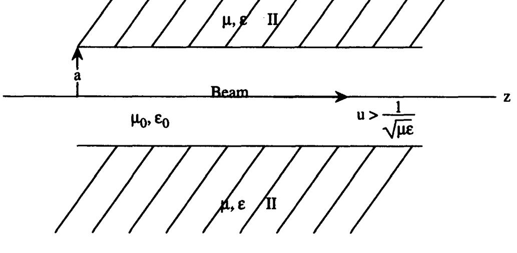

FIG. 24. Bunched cylindrical electron beam passing through hole in a dielectric.

(Cherenkov condition), then

$$R(r) = H_0^{(2)}(br)$$

(Hankel function—outgoing waves), where

$$b = (n\omega_0/v)[\mu\epsilon u^2 - 1]^{1/2}$$

For the driven oscillator system considered, coherent radiation will be produced at all harmonic frequencies $n\omega_0$ for which the Cherenkov condition is satisfied.

4.2.2 Self-Excited Cherenkov Oscillator A self-excited Cherenkov oscillator can be made by passing an electron beam through a dielectrically lined double-slab waveguide or a dielectrically lined cylindrical waveguide as illustrated in Fig. 25. Here the Cherenkov effect takes the form of a traveling-wave amplifier with internal feedback wherein the beam velocity u_0 is greater than the wave phase velocity.

It is of interest to briefly consider the physics of electronic motion in a traveling-wave field (Hutter, 1960). Here Newton's equation for one dimension takes the form

$$m_o\frac{du}{dt} = m_o\left(\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial z}\right) = qE_1e^{j\omega t}e^{-\Gamma z}, \tag{47}$$

with

$$u = u_0 + u_1e^{j\omega t}e^{-\Gamma z},$$

where u_0 is a constant velocity and u_1 is a small amplitude. Solving Eq. (47) one obtains

$$u_1 = \frac{qE_1/m}{j\omega - \Gamma u_0}. \tag{48}$$

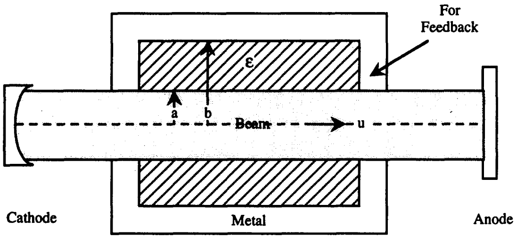

FIG. 25. Configuration of a Cherenkov oscillator.

Next let the charge density ρ and current J be

$$\rho = \rho_0 + \rho_1 e^{j\omega t} e^{-\Gamma z}, \tag{49}$$

$$J = \rho u = J_0 + J_1 e^{j\omega t} e^{-\Gamma z}, \tag{50}$$

with

$$\frac{\partial J}{\partial z} + \frac{\partial \rho}{\partial t} = 0. \tag{51}$$

Solving for J_1 yields the result

$$J_1 = \frac{j\omega}{u_0} \frac{J_0 E_1 e/m}{(j\omega - \Gamma u_0)^2}. \tag{52}$$

If the propagation constant Γ is purely imaginary, J_1 and E_1 are out of phase, no electron bunches are formed, and no work is done. However, if Γ is complex, work can be done, leading to attenuated or amplified waves. The value of Γ can be found by using J_1 in Maxwell's equations, applying boundary conditions for the problem, and obtaining a self-consistent solution.

Garate (1987) has analyzed a dielectrically lined double-slab waveguide driven by a relativistic electron-beam wave, using the procedure outlined, and obtained the data shown in Fig. 26. Von Laven *et al.* (1982) using a 1–40-A, 100–25-keV beam with a pulse duration of 4 μs, obtained 30–100 kW of pulsed power in the 30–50-GHz range using the structure shown in Fig. 25.

The use of microwave-type transmission lines and resonators in a Cherenkov oscillator will limit its frequency most likely below 300 GHz. The beauty of a Cherenkov oscillator is its simple beam-coupling structure. No high-technology, submicron fabrication techniques are needed to make the oscillator.

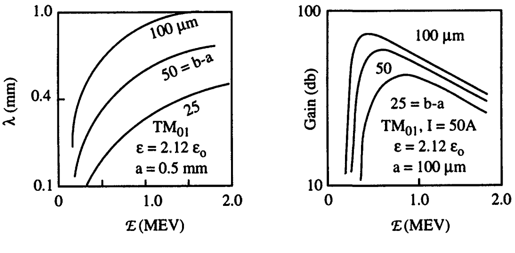

FIG. 26. Wavelength and gain vs electron energy of a Cherenkov oscillator.

5. RELAXATION OSCILLATORS

While the generation of sine waves is important, especially at high frequencies, there are many uses for other wave forms (pulses, rectangles, triangles, trapezoids, etc.)—generators, for example, in timing applications (Chance *et al.*, 1949). The relaxation oscillator that will be considered here is the astable plate-coupled multivibrator, which can provide nearly rectangular pulses of a specific duration. This multivibrator circuit has two triode tubes with *RC* coupling from the plate of each tube to the grid of the other tube as shown in Fig. 27.

The cycle of operation consists of alternate periods when one tube and then the other is cut off with rapid switching from one condition to the other. Suppose the plate current i_{b_2} increases, thereby lowering the voltage e_{b_2}. This in turn causes a sequence of actions with e_{c_1} decreasing, decreasing i_{b_1}, increasing e_{b_1}, increasing e_{c_2}, further increasing i_{b_2}. As the charge on capacitor C_1 decreases, e_{c_1} rises toward zero and, as soon as the cutoff voltage e_{c_0} is reached, turns tube 1 from off to on. Thus, the current is suddenly transferred from the plate circuit of one tube to that of the other. The various voltage wave forms are also displayed in Fig. 27.

The periods τ_1 and τ_2 depend on the *RC* time constant and the ratio of the voltages involved. For example,

$$\tau_2 = R_{g_2} C_2 \ln(e_2/e_{c_0}). \tag{53}$$

Switching times are much smaller than 1 μs, and the periods τ_1 and τ_2 can be changed by varying R_{g_1} and R_{g_2}.

6. POLARIZATION-CURRENT (LASER) OSCILLATORS

From a Maxwell-equation, mathematical point of view, conduction-current (**J**) and polarization-current ($\partial\mathbf{P}/\partial t$) oscillators represent different driving terms for the EM fields. Given **J** or $\partial\mathbf{P}/\partial t$, the mathematics of solving for the fields is the same. In this case, **J** or $\partial\mathbf{P}/\partial t$ must be out of phase more than 90° with

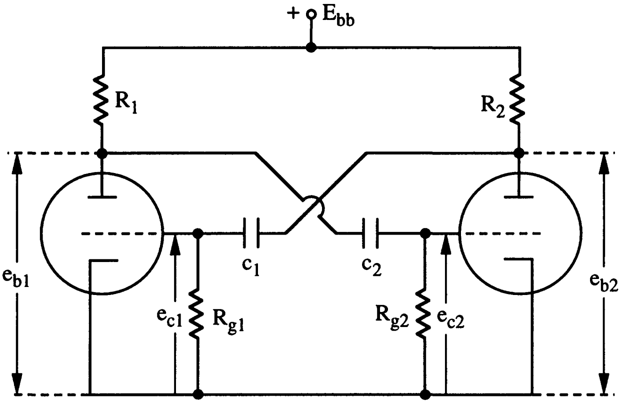

(a) Circuit Diagram

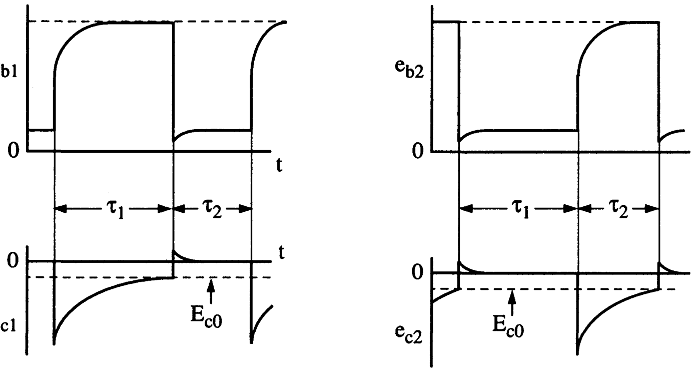

(b) Waveforms

FIG. 27. Astable multivibrator oscillator.

the field **E** to deliver energy. For a **J** device, this means generating an ac or bunched electron beam, and for a $\partial \mathbf{P}/\partial t$ device it means creating a non-thermodynamic-equilibrium distribution in a quantum system. In some cases, both can use the same feedback or resonator configuration, as, for example, the free-electron oscillator, which uses a Fabry–Pérot resonator. One key difference in the two devices is that the laser oscillator does not have a transit-time limitation on its frequency range.

However, a basic physics change had to be made in the analysis of **J** and $\partial \mathbf{P}/\partial t$ oscillators in that quantum mechanics has to replace Newtonian mechanics to calculate $\partial \mathbf{P}/\partial t$. This is not surprising since the laser is a quantum system. The equation for **P** for a two-level system, assuming an electric dipole transition, has been given in Eq. (4). It is to be observed that this second-order linear equation is of the same mathematical form as an *LRC* circuit driven by a voltage source. Since sinusoidal oscillations are being considered, this too is not unexpected.

The sign of $\partial \mathbf{P}/\partial t$ can be changed by making $N_2 > N_1$, i.e., a population inversion or change in the distribution function away from its thermodynamic equilibrium value. Equation (4) allows one to calculate the electric susceptibility χ_e since

$$\mathbf{P} = \epsilon_0 \chi_e \mathbf{E} = \epsilon_0(\chi_e' - j\chi_e'')\mathbf{E}, \tag{54}$$

and the loss α (or gain) of the EM wave traversing the medium,

$$\alpha = \omega\chi_e''/v_g, \tag{55}$$

where v_g is the group velocity.

Various types of lasers are described in other articles (see LASER PHYSICS; LASERS, FREE-ELECTRON; LASERS, GAS; LASERS, SEMICONDUCTOR; LASERS, SOLID STATE; ORGANIC DYE LASERS; TUNABLE DYE LASERS) and hence are discussed only to make comparisons with **J** devices. Conduction-current oscillators, as has been seen, have the challenges of transit time, circuit sizes, and losses, as they are extended to higher frequencies. Lasers have the challenges of devising specific excitation schemes to create highly non-thermodynamic-equilibrium distributions in the face of fast relaxation and damping processes.

In this author's opinion, while technological advances will be made on **J** devices in the future, finding new physics phenomena to exploit is becoming more difficult. No basic new **J** device has appeared for many years. A possible new frontier exists for $\partial \mathbf{P}/\partial t$ laser devices for nuclear energy-level systems.

Works Cited

Brau, C. A., Newnam, B. E. (Eds.), (1987), Special Issue on Free Electron Lasers, *IEEE J. Quantum Electron.* **QE-23**, 1468–1665.

Chance, B., Hughes, V. W., MacNichol, E. F., Jr., Sayer, D., Williams, F. C. (1949), *Waveforms*, MIT Radiation Laboratory Series, Vol. 19, New York: McGraw-Hill.

Coleman, J. T. (1982), *Microwave Devices*, Reston, VA: Reston Publishing Company.

Coleman, P. D. (1951), Ph.D. Dissertation, Massachusettes Institute of Technology.

Coleman, P. D., Enderby, C. E. (1960), "Megavolt Electronics Cerenkov Coupler for the Production of Millimeter and Submillimeter Waves," *J. Appl. Phys.* **31**, 1695–1696.

Convert, G., Yeou, T., Pasty, A. (1959), "Millimeter Wave O-Carcinotron," *Microwave Res. Inst. Symp. Ser. Polytech. Inst. Brooklyn* **9**, 313–339.

Copeland, J. A. (1967), "Characterization of Bulk Negative Resistance Diode Behavior," *IEEE Trans. Electron Devices* **ED-14**, 461–463.

de Forest, L. (1907), U.S. Patent 841387.

Frank, I. M., Tamm, I. (1937), "Coherent Visible Radiation of Fast Electrons Passing Through Media," *C. R. Acad. Sci. USSR* **14**, 109–114.

Garate, E. P. (1987), "High-Gain Cerenkov Free-Electron Laser at Far Infrared Wavelengths, *J. Quantum Electron.* **23**, 1627–1632.

Graratstein, V. L., Roberson, C. W. (Eds.) (1985), Special Issue on Free Electron Lasers, *J. Quantum Electron.* **QE-21**, 804–1119.

Hunter, L. P. (1956), *Handbook of Semiconductor Electronics*, New York: McGraw-Hill.

Hutter, R. G. E. (1960), *Beam and Wave Electronics in Microwave Tubes*, New York: D. Van Nostrand Co.

Karp, A. (1955), "TWT Experiments at Millimeter Wavelengths," *Proc. IRE* **43**, 41–46.

Kompfner, R. (1947), "The TWT as an Amplifier at Microwaves," *Proc. IRE* **35**, 124–127.

Liao, S. Y. (1985), *Microwave Devices and Circuits*, Englewood Cliffs, NJ: Prentice-Hall.

Madey, J. M. J. (1971), "Stimulated Emission of Bremsstrahlung in a Periodic Magnetic Field," *J. Appl. Phys.* **42**, 1906–1913.

Millman, J., Halhias, C. C. (1972), *Integrated Electronics*, New York: McGraw-Hill.

Pantell, R. H., Puthoff, H. E. (1969), *Fundamentals of Quantum Electronics*, New York: Wiley.

Parker, R. K., Jackson, R. H., Gold, S. H., Freund, H. P., Granatstein, V. L., Efthimion, P. E., Herdon, M., Kinkead, A. K. (1982), "Axial Magnetic Field Effects in a Collective Interaction Free-Electron Laser and Millimeter Wavelengths," *Phys. Rev. Lett.* **48**, 238–242.

Phillips, R. M. (1960), "The Ubitron," *IRE Trans. Electron Devices* **ED-7**, 231–241.

Robenson, K. E., Quimby, D. C., Slater, J. M., (1987), "The Tapered Hybrid Undulator of the Free-Electron Laser Oscillator Experiment," *J. Quantum Electron.* **9**, 1497–1513.

Sprangle, P., Drobot, A. T. (1977), "The Linear and Self-Consistent Nonlinear Theory of the Electron Cyclotron Maser," *IEEE Trans. Microwave Theor. Techniques* **MTT-25**, 528–544.

Süsskind, C. (Ed.) (1988), Special Issue Commemorating the Centennial of Heinrich Hertz, *IEEE Trans. Microwave Theor. Techniques* **MTT-36**, 802–858.

Szöke, A. (Ed.) (1981), Special Issue on Free Electron Lasers, *J. Quantum Electron.* **QE-17**, 1326–1514.

Stratton, J. (1941), *Electromagnetic Theory*, New York: McGraw-Hill.

Temkin, R. J., Kreischer, K. E., Mulligan, N. J., MacCabe, S., Fetterman, H. R. (1982), "A 100 kW, 140 GHz Pulsed Gyrotron," *Int. J. Infrared Millimeter Waves* **3**, 427–438.

Von Laven, S., Branscum, J., Golub, J., Layman, R., Walsh, J. (1982), "High Power Cerenkov Maser Oscillators," *Appl. Phys. Lett.* **41**, 408–410.

OSCILLOSCOPES, ANALOG AND DIGITAL

NELLIE BROCK, *Tektronix, Inc., Beaverton, Oregon, U.S.A.*

Introduction 37
1. **Basic Oscilloscope Concepts** 38
1.1 Analog Oscilloscopes 39
1.2 Digital Storage Oscilloscopes 41
1.3 General Wave-Form Capture Requirements 41
1.3.1 Bandwidth and Rise Time 41
1.3.2 Probing Effects 42
1.3.3 Transient vs Repetitive 43
2. **Resolution and Accuracy** 43
2.1 Analog Display Resolution 44
2.2 Digital Storage Resolution 44
2.3 Accuracy 44
2.4 Digital Time-Base Accuracy 44
3. **Memory and Record Length** 45
4. **Analog-to-Digital Conversion** 46
4.1 Digitizing Rate 46
4.2 Real Time and Equivalent Time 46
4.3 Converter Types 47
4.3.1 Successive Approximation 47
4.3.2 Flash Conversion 48
4.3.3 Charge-Coupled Device (CCD) 48
4.3.4 Scan Conversion 48
4.4 Vertical Quantizing 48
4.5 Effective Bits 49
5. **Display Techniques** 49
5.1 Dot Displays 49
5.2 Interpolated Displays 50
5.3 Peak Detection and Enveloping 50
5.4 Pre- and Post-Trigger Displays 51
6. **Useful Storage Bandwidth and Rise Time** 52
6.1 Useful Storage Bandwidth 53
6.2 Useful Storage Rise Time 53
7. **Signal Processing** 53
7.1 Save-on-Delta 53
7.2 Averaging 53
7.3 Integration and Differentiation 54
7.4 Fast Fourier Transform 54
7.5 Convolution and Correlation 54
8. **Measurement Automation** 54
8.1 Interfacing Techniques 54
8.2 Software Considerations 55
9. **Conclusion** 55
Glossary 56
Works Cited 57
Further Reading 57

INTRODUCTION

Oscillographs are a means of graphically studying oscillatory or cyclical phenomena—hence the term "oscillo" + "graph." With this fundamental definition, probably the earliest use of oscillography was the sundial. This simple device portrays the movement of the sun across the sky as a shadow cast across a scale to indicate the time of day.

A more sophisticated and widely familiar example is the seismograph used to record earth tremors and quakes. Here, a graphical record of seismic activity is penned onto a paper chart for detailed study. The process uses a slowly rotating drum onto which the blank graph paper is attached. The drum's rotation is controlled to correspond to a time scale. A seismic sensor detects earth tremors and converts them to a corresponding electrical signal which is used to drive a pen mechanically back and forth across the moving chart paper. The result is a graph of seismic amplitude vs time.

Historically, numerous general-purpose and highly specialized oscillographic devices have been developed to further the recording and study of physical phenomena. Prior to the advent of electronic devices, however, oscillography was generally confined to chart recorder and photographic methods.

It was not until the age of radio that an unavoidable need emerged for observing electronic wave forms. Engineers needed to

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

see into the radio circuits they were developing in order to measure wave-form distortion and other amplifier effects. Using cathode-ray tubes (CRT), which had previously been developed in the study of x rays, engineers built their own rudimentary oscilloscopes based on designs passed along through symposia and journals. These first instruments were referred to as oscillographs, but gained the name oscilloscope by the 1940s to distinguish them from the broader category of oscillographs. The term uses the root word scope, from the Greek "skopos," meaning an instrument for viewing, as in telescope.

The basic function of an oscilloscope is to provide an *x-y* graphical view of physical phenomena. In the most common applications today, the display is voltage amplitude vertically on the *y* axis vs time horizontally on the *x* axis. This is because an oscilloscope is an electronic instrument designed for acquisition and display of electrical signals. However, it is of the utmost importance to realize that the displayed electrical signal, or wave form, can represent virtually any physical phenomenon.

What the oscilloscope display represents is determined by the application and by the type of probe or sensor used at the oscilloscope input. For example, optical events can be captured and displayed via an optical-to-electrical converter. Similarly, a microphone can provide an audio signal input, an accelerometer can provide a motion signal, and a thermocouple can provide a temperature signal input.

Thus, by judicious choice of sensors or probes, an oscilloscope is potentially able to provide monitoring, measurement, and analysis of almost any physical phenomenon or event. The phrase "potentially able to" is used here because actual display, measurement, or analysis depends on the type of oscilloscope selected and its various capabilities.

The most basic selection issue, of course, is adequate frequency bandwidth of the oscilloscope for the event signals being investigated. For example, a 20-MHz oscilloscope is inadequate for acquiring and displaying 100-MHz signals. Even with adequate bandwidth, however, there can be certain classes of signals that still cannot be displayed adequately on certain types of oscilloscopes. For example, a fast transient signal often will not create a visible trace on the display of a conventional analog oscilloscope, but will be clearly visible on the display of a real-time digital oscilloscope or transient digitizer. On the other hand, extremely fast transients may only be brought to light by the use of fast analog techniques and special CRT photography.

The point here is that careful and informed oscilloscope selection is paramount in successful display, measurement, and analysis of physical phenomena. This requires an understanding of basic oscilloscope concepts, the different oscilloscope technologies available, and the key advantages of each. This article provides the basic information necessary for informed oscilloscope selection.

1. BASIC OSCILLOSCOPE CONCEPTS

The most basic elements of any oscilloscope are acquisition, processing, and display. These elements may be implemented in a wide variety of manners with either analog or digital technologies, or a combination of both.

Acquisition begins, fundamentally, as an analog process. It includes, first of all, probing or sensing the phenomenon to be displayed or measured. It also includes triggering on a particular event or aspect of the phenomenon and generally includes some preconditioning (amplification or attenuation) of the input signal.

Processing, in general, is anything that is done to or in conjunction with the input signal to provide a useful display of the signal. This almost always includes use of a time base to set a calibrated display sweep or to control a sampling interval. It may also include other functions, such as subtracting a ground reference or even performing a fast Fourier transform on the signal.

Display is the method of formatting and presenting the acquired signal for observation. This may include scale factors, measurement cursors, and even measurement results. Most often, the primary display is the oscilloscope screen. However, displays may include separate numeric readouts as well as results sent to a printer.

There are two classes of oscilloscopes available for signal acquisition, processing, and display. These are generally referred to as the *analog oscilloscope* and the *digital storage oscilloscope.* Figure 1 shows a basic block diagram of each oscilloscope class for purposes of comparison in the following discussion.

1.1 Analog Oscilloscopes

Beginning with the input signal from a probe or sensor, the first operation of an analog oscilloscope is input signal conditioning. This consists of either attenuation or amplification to provide an amplitude-calibrated signal within the vertical display range of the oscilloscope. This input signal drives the display trace vertically, similar to the y-axis pen drive on a plotter or chart recorder.

The input signal amplitude or y-axis drive provides one dimension of the display. The second dimension, the x-axis drive, is provided by a second signal. When the oscilloscope is operated in what is called an x-y mode, the x-axis drive signal is a second input signal. This produces an x-y display of one signal against another, such as in creating Lissajous patterns. However, typical oscilloscope usage, which is the focus here, is to plot or display signal amplitude (y) against time on the x axis. This is done by using a ramp signal generated within the oscilloscope to provide the x-axis or horizontal drive. The ramp signal used is often referred to as the horizontal sweep signal. The duration of this ramp and how it is triggered are key factors in providing the desired time-calibrated display of the signal being input to the oscilloscope.

As shown in Fig. 1(a), triggering information is obtained at the signal input stage. The setup of the triggering and horizontal systems also determines how the signal is displayed. This is illustrated in Fig. 2.

The trigger system settings determine the point on the input signal that causes triggering to occur. Triggering starts sweep generation by the horizontal system. The sweep signal is a linear ramp that drives the display trace left to right across the display at a constant sweep speed. The duration of this ramp, and thus the horizontal sweep speed, are determined by the oscilloscope's time-base setting. This provides horizontal time calibration of the display.

In single-shot triggering, the trigger is manually armed by pressing a trigger arm button. One single-sweep display occurs when the input signal crosses a preset trigger level. There are no further displays after that until the trigger is manually rearmed and another triggering event occurs. The single-shot mode is intended for use in acquiring single events such as a laser pulse, a sonar ping, or a mechanical impact signal.

A repetitive triggering mode is used for acquiring repetitive signals such as sinusoids. This is the triggering mode shown in

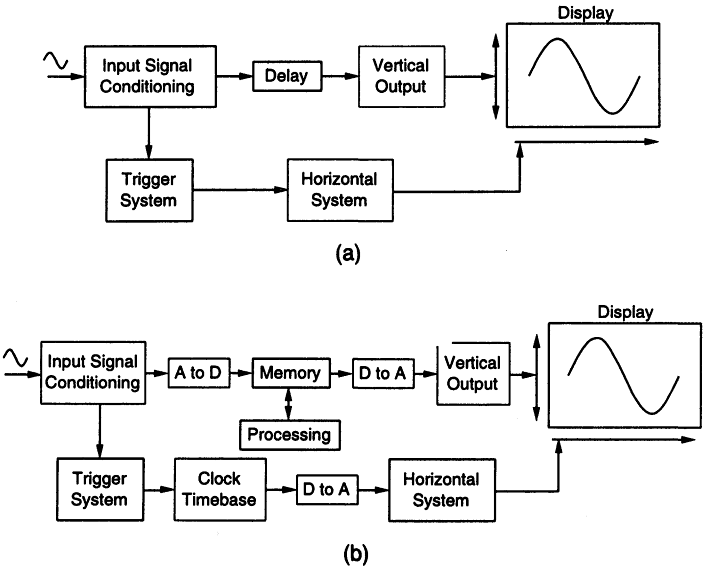

FIG. 1. Block diagram of (a) an analog oscilloscope and (b) a digital storage oscilloscope (DSO). The DSO has the same analog oscilloscope front-end requirements for signal conditioning and triggering.

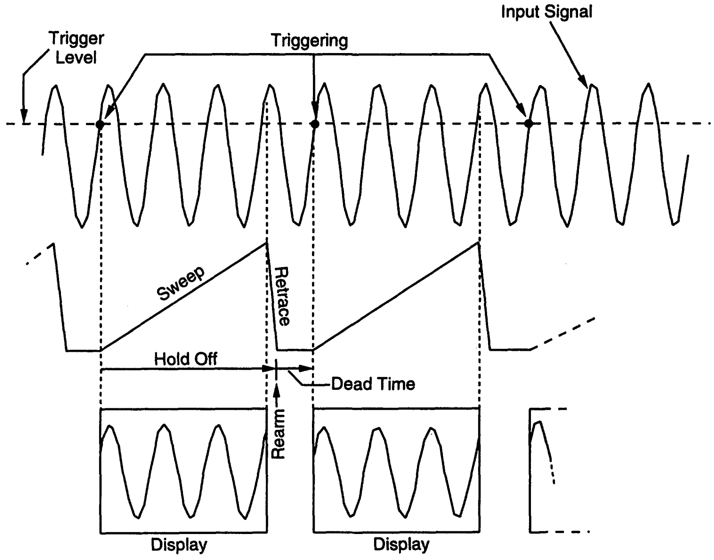

FIG. 2. A trigger level and triggering conditions are set up to determine the beginning point of waveform acquisition and display. Triggering also starts the sweep ramp for driving the display trace horizontally.

Fig. 2. Note that a sweep, and thus a display, are generated when the input signal reaches the trigger level setting. Once the sweep starts, further triggering is held off until sweep retrace is completed. At this point, the triggering circuitry is automatically rearmed. The next sweep (and display) then occurs when the input signal again crosses the trigger level.

The display on an analog oscilloscope is created by driving an electron beam across the phosphor-coated screen of a cathode ray tube (CRT). The input signal drives the CRT's vertical deflection plates to move the beam vertically, and the triggered sweep ramp (or other signal) drives the CRT's horizontal deflection plates to move the beam horizontally. The beam electrons strike and excite the CRT phosphor, causing it to emit light along the path traced by the electron beam.

In the case of single-shot triggering, the phosphor is excited by a single pass of the electron beam. However, the faster the beam is traveling (higher sweep speeds), the less the excitation and the dimmer the trace. The limit where portions of the displayed trace are no longer visible defines the CRT's writing speed. Thus, for viewing single-shot events on an analog oscilloscope, it is important to have a high–writing-speed CRT.

For repetitively triggered events, such as illustrated in Fig. 2, CRT writing speed is less of a concern. This is because the repetitively triggered sweeping of the beam regularly reexcites the phosphor in the trace path. As a result, repetitively triggered wave forms provide bright traces even when acquired at extremely fast sweep speeds.

It is important to note that the wave-form display process is analog and occurs in real time. That is, the actual analog signal being acquired is used directly to create the display trace. Thus, changes in the input signal are immediately reflected on the oscilloscope display in real time, as the signal changes actually occur. This is a key advantage when using the oscilloscope as a monitoring aid in making amplitude or timing adjustments on signals.

Another advantage is that momentary wave form disturbances, such as amplitude noise or time jitter, are also easily distinguishable. Such momentary or random signal disturbances cause corresponding deviations from the predominant trace path. Over multiple triggered traces, the result is less intense traces surrounding the more intense trace of the statistically dominant trace path. This intensity variation provides a subjective indication of signal noise statistics.

The real-time aspect of analog oscilloscopes can be a disadvantage, too. Since the display is a real-time display, transitory signal changes may fade from view before they can be studied. In the case of single pulses or other transient signals, the trace is a single brief flash across the screen. If the CRT's writing speed is not sufficient, there may be no trace at all. At the other extreme, a very low-frequency wave form may appear only as a bright dot moving slowly across the screen.

Many of these real-time display problems have been addressed through variations in CRT technologies. High-persistence phosphors have been used, for example, to slow trace fading in single-shot or low-frequency applications. Trace viewing time has been extended even more by use of storage meshes behind the CRT's phosphor. Another alternative is the use of CRT photography to capture transitory real-time displays for more detailed study.

1.2 Digital Storage Oscilloscopes

The digital storage oscilloscope (DSO) was developed to answer some of the shortcomings of analog oscilloscopes. Primarily, the DSO digitizes the input signal and stores it in digital memory. The key advantage is that the stored wave form can be easily reconstructed on the display at any time, even long after the actual analog signal is gone.

For transient events, the DSO essentially freezes the event on screen for detailed study. By contrast, single-shot capture of transient events with an analog oscilloscope provides only a brief trace flash on the display. Any detailed study of such analog wave-form displays must usually be done via a photograph of the display.

For any signal—single-shot or repetitive—digital storage offers the advantages of detailed analysis. For example, the mean value, root mean square value, or any other signal parameter can easily be computed from the stored values of the digitized signal. With appropriate computational support, advanced analysis techniques can also be applied. This can include performing fast Fourier transforms, convolution, correlation, or any other analytic process that can be programmed.

While the digital storage and processing of a DSO offer numerous advantages, there can also be some disadvantages. Understanding both the advantages and disadvantages relative to analog oscilloscopes is an important part of oscilloscope selection. However, before making a digital vs analog selection, it is important to consider the general wave-form capture requirements that apply to both.

1.3 General Wave-Form Capture Requirements

All oscilloscopes, regardless of their type, have an analog signal input system. This is necessary since the real world is analog. Even the digital signals used in digital systems are, in actuality, analog signals; they are time-continuous voltages that are simply level switched to represent digital data.

Consequently, whether considering an analog oscilloscope or a DSO, the first issue is the analog signal input requirements. This includes not only the oscilloscope's input signal conditioning but also the effects of probing or sensing.

1.3.1 Bandwidth and Rise Time Oscilloscopes are most commonly specified in terms of frequency bandwidth—e.g., a 200-MHz oscilloscope, a 200-MHz DSO. The bandwidth specification refers to the maximum-frequency sinusoid that can be acquired and displayed without significant signal attenuation. In other words, a 200-MHz oscilloscope can be expected to display any sinusoid up to 200 MHz without significant signal attenuation.

In the case of nonsinusoidal signals, the bandwidth specification must be interpreted in terms of the highest frequency component of significance in the signal being acquired. For example, repetitive nonsinusoidal wave forms, such as a square-wave signal, are made up of the fundamental frequency and some combination of harmonics. In general, such signals must be preserved out to the fifth harmonic for reasonably accurate observations and measurements. Thus, acquisition of a 40-MHz square wave would typically require a 200-MHz (5 × 40 MHz) oscilloscope to preserve signal amplitude and shape characteristics.

In many applications, the key consideration is measurement of a signal's rise time or some other timing aspect. For such applications, the oscilloscope's input bandwidth specification needs to be converted to a rise-time specification. According to Tektronix (1993a), the expression for bandwidth/rise-time conversion is

$$T_r = 0.35/f_{BW}, \qquad (1)$$

where T_r is rise time and f_{BW} is frequency bandwidth. Thus, the rise-time specification for a 200-MHz oscilloscope is 0.35/(200 MHz) = 1.75 ns. This expression can be derived from the Fourier transform relation between a Gaussian pulse's rise time and bandwidth. It provides a good approximation for most systems having a constant delay vs frequency.

An oscilloscope's rise time should be faster than the signal rise time being measured. How much faster depends on the measurement accuracy desired. The graph in Figure 3 provides a means of determining the necessary rise time for a desired accuracy. For example, if the oscilloscope's rise time is 3 times the rise time being measured (3:1 rise-time ratio), a 5% measurement error can be expected.

1.3.2 Probing Effects The type of probe or sensor used with the oscilloscope must also be considered as part of the overall measurement system. This is because the probe's impedance causes circuit-loading effects.

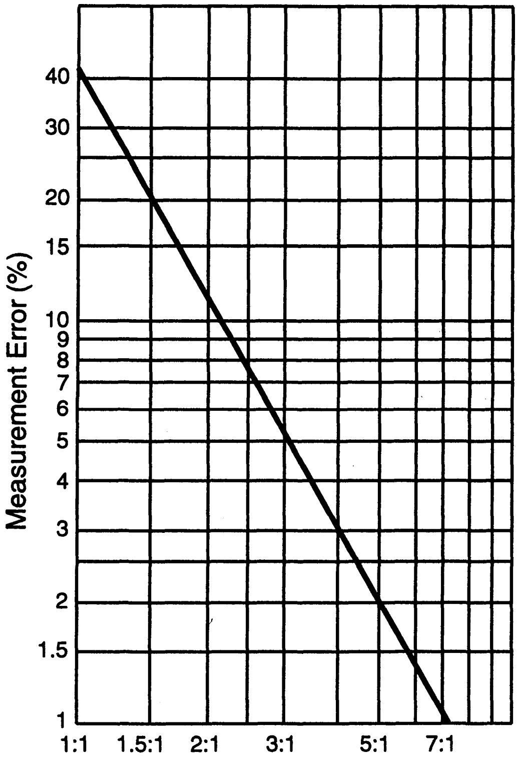

FIG. 3. Chart for determining rise-time measurement error. The amount of error depends on the ratio of the oscilloscope's rise time to the rise time being measured. A 3:1 ratio corresponds to 5% error.

For signal frequencies below 5 kHz, resistance is the dominant circuit-loading element. To minimize resistive loading, the probe used should have a resistance that is at least two orders of magnitude greater than the impedance of the circuit or sensor output being measured.

As frequencies progress above 5 kHz, probe inductance and capacitance have increasing loading effects. There are several steps that can be taken to reduce this reactive loading. To minimize inductive effects, the probe should be securely grounded with as short a ground lead as possible. To reduce capacitive effects from the probe, an attenuator probe (e.g., 10× attenuation) can be used. (Essentially, added resistive attenuation at or near the probe tip reduces the shunting effects of stray capacitance in the probe body and probe cable to the scope input.)

Additionally, the probe used should be adjusted to compensate for variations in the oscilloscope's input capacitance. Probe compensation is a simple procedure that is described in most oscilloscope manuals. Yet it is also one of the most often overlooked procedures, the result of which can be serious measurement errors, as reported by Tektronix (1993b).

Probes, like oscilloscopes, have rise-time and bandwidth limitations. As a result, probe selection must also be made to ensure adequate bandwidth or rise time for the signals to be measured. However, the probe's bandwidth or rise time cannot be considered alone. This is because the probe is an integral part of the oscilloscope's or DSO's input system.

Thus, the overall input system rise time must be considered in terms of the following relationship:

$$T_{r(\text{system})} = \sqrt{T^2_{r(\text{scope})} + T^2_{r(\text{probe})}}. \qquad (2)$$

The system rise time computed by Eq. (2) can be re-expressed as a system bandwidth using Eq. (1).

The key point to be realized from Eq. (2) is that inappropriate probe selection can significantly degrade the input bandwidth of an oscilloscope or DSO. To avoid this, manufacturers of high-quality instruments generally optimize their instruments for specific probe models. As reported by Tektronix (1993b), oscilloscope manufacturers often specify oscilloscope bandwidth at the probe tip when specific probe models are used.

1.3.3 Transient vs Repetitive Even with sufficient bandwidth or rise time, it is still possible that the instrument will not be able to provide an adequate display of the acquired signal. The classic example of this is in acquiring transient wave forms such as static discharges, mechanical impacts, and other one-time events.

When acquired with an analog oscilloscope, the transient signal causes a single-sweep trace across the display screen. For slow transients, the trace speed is slow enough that ample energy is imparted to the screen phosphor for a brief, but visible, trace in normal room light. On faster transients, the faster trace speed imparts less energy to the screen phosphor and the trace is dimmer.

As transient speed (and thus single-sweep trace speed) increases, a point is reached where the visual writing speed of the analog oscilloscope's display is exceeded. When this occurs, the trace is no longer visible in normal room light. However, the trace may still provide enough phosphor excitation to be visible through use of a viewing hood or via single-shot display photography. When the transient and its trace speed exceed the display's photographic writing speed, the trace is too dim for capture even with specialized photographic techniques.

In general, the photographic writing speed limit of an analog oscilloscope is significantly less than the bandwidth and horizontal sweep-speed limits. This means that the transient wave-form viewing capabilities will be substantially less than the repetitive wave-form viewing capabilities. The exception to this is microchannel plate (MCP) display technology. MCP oscilloscopes provide trace intensity multiplication of about 1000 times, as reported by Gustafson *et al.* (1985). This MCP intensity amplification is usually sufficient for single-sweep viewing up to the oscilloscope's rise-time specification at bandwidths of as much as 1 GHz.

With DSOs, display writing speed is not an issue. This is because the display trace does not have to be a real-time representation of the acquired signal. Instead, the signal is digitized and stored in memory, and then a display is generated from the stored wave-form data. As a result, both transient and repetitive signals can be displayed with equally clear and bright displays. Moreover, because the display is reconstructed from memory, transients are no longer a brief flash on the display, but are frozen on-screen for easy viewing.

The DSO's real-time sample rate is the limitation for capturing and displaying transient signals. Many DSOs use equivalent-time sampling for capture of repetitive wave forms up to the DSO's bandwidth limit. However, equivalent-time sampling, as will be discussed later, is a method of using lower real-time sample rates to create, effectively, high-speed sampling on repetitive wave forms. For single-sweep transient capture, the limit is always the actual real-time sampling rate.

Certain types of DSOs use real-time sampling exclusively to allow capture of transients up to the DSO's bandwidth and rise-time limit. Such DSOs are generally designed specifically for transient signal capture and are often referred to as transient digitizers.

The limitation on transient capture imposed by real-time sampling is essentially the time resolution provided by the sample rate. Similarly, the maximum equivalent-time sample rate imposes the upper limit on repetitive wave-form capture with a DSO. By comparison, analog oscilloscope limits for transient wave form capture are the display's writing speed and the bandwidth/rise-time specification for repetitive wave forms.

There is one other special limitation that needs to be considered regarding repetitive wave forms. This is the limitation of viewing very slow wave forms on analog scopes. Subhertz wave forms must be acquired on the lowest sweep-speed settings. As a result, the display on an analog oscilloscope will be just a bright dot moving across the screen. For such cases, there are two solutions for viewing a full trace of the slowly changing signal. One solution is to use an analog display storage oscilloscope or other high-persistence display technology. The other solution is to use a DSO to digitize the signal and re-create a full trace from memory.

2. RESOLUTION AND ACCURACY

Resolution and accuracy are sometimes confused in discussing oscilloscopes. In many cases they are wrongly taken to be synonymous.

Accuracy is the degree of conformity to an accepted standard. Resolution is the ability to distinguish differences. Thus, measurements between different things cannot be more accurate than the ability to resolve the differences. Consequently, to take full advantage of accuracy, there must be at least enough resolution to match accuracy and preferably more.

2.1 Analog Display Resolution

Analog oscilloscopes display wave forms against a calibrated display graticule. Generally, this graticule consists of eight major vertical divisions by ten major horizontal divisions. Each division is one centimeter square.

Typical spot sizes for the beam trace representing an acquired signal will range from 1/25 to 1/50 of a division, as reported by Tektronix (1992). The spot size can resolve half of a trace width. Therefore, the vertical display resolution is one part in 400 to one part in 800, while horizontal resolution is one part in 500 to one part in 1000. This is adequate for most on-screen measurement needs. Also, many analog oscilloscopes now have measurement cursors, or crosshairs, that can be positioned on the display to simplify waveform time and voltage measurements. It should be remembered, however, that such analog cursors are actually measuring screen position. Therefore, the cursors must be carefully positioned on the wave-form trace for accurate wave-form measurements.

2.2 Digital Storage Resolution

Digital storage oscilloscopes operate by sampling the input signal and digitizing the sample values. Thus, the time resolution of the stored signal data corresponds to the sampling interval (Δt), which is the reciprocal of the acquisition sampling rate.

DSO resolution may also be expressed in terms of record length. When a signal is captured with a DSO, a block of samples is taken for digitizing and storage. The number of stored samples for any given signal is referred to as the record length. Typical DSO record lengths are 512 and 1024 samples, although longer records are often available. A record length of 1024 samples implies a horizontal resolution of one part in 1024.

Vertical resolution corresponds to the digitizing resolution, which is one part in 2^B, where B is the number of digitizing bits. An 8-bit digitizer, for example, provides an amplitude resolution of one part in $2^8 = 256$; a 10-bit digitizer provides one part in $2^{10} = 1024$.

2.3 Accuracy

Oscilloscopes and DSOs have a variety of accuracy specifications. The most important of these include the amplitude accuracy of the vertical system and the time-base accuracy of the horizontal system.

Amplitude accuracy is simply the degree of accuracy to which a signal amplitude can be measured using the oscilloscope or DSO. In other words, how accurately can the amplitude of a 1-V, peak-to-peak sinusoid be measured. A fairly typical accuracy is 5%, which corresponds to one part in 20. This is well within the basic vertical resolution requirement cited above in 2.1 and 2.2 (resolution is greater than accuracy).

Time-base accuracy refers to ramp or clocking accuracy in reflecting a calibrated time horizontally across the display. In analog oscilloscopes, this is most directly related to the linearity of the ramp signal that drives the horizontal sweep. This is critical since the ramp horizontally drives the display, and any ramp nonlinearity translates to an effective speeding or slowing of the represented time. High-performance oscilloscopes address this by keeping sweep nonlinearities to 3% or less.

In DSOs, a ramp is often used to drive the horizontal sweep of the display system. The accuracy of this ramp is also important for measurements made directly from the display. However, since the display is reconstructed from digital memory, rather than real-time acquisition, it is not necessary to use as wide a range of ramp speeds as in analog oscilloscopes. Instead, DSO timing accuracy is based on the clocking accuracy of the digital time base used to drive sampling.

2.4 Digital Time-Base Accuracy

Digital time bases are generally driven by crystal-controlled clocks. This provides greater long-term accuracy and stability than the real-time methods of analog oscilloscopes.

In a DSO, the digital time base clocks the occurrence of signal sampling. The stability and accuracy of the clock determine the accuracy of where samples fall relative to each other on the signal. This is particularly important for equivalent-time sampling techniques since a complete sample record is built up over many signal repetitions. The longer it takes to build such a record, the more dominant clock instability and drift become in placing samples relative to each other and to the signal.

It is important to note that the sample interval (Δt) is assigned to the stored signal record as a single number. This sample interval is the reciprocal of the sample (or clock) rate. Thus, instabilities in actual clock rate result in a deviation of sample placement relative to sample interval. As indicated in Fig. 4, not only does this create a timing error, but it also results in a sample amplitude error for changing signal levels.

3. MEMORY AND RECORD LENGTH

DSOs, by definition, require memory for storage of the digitized wave form. However, it is common today for analog oscilloscopes to have digital memory, too. In analog oscilloscopes, this memory is used for storing front-panel setups for easy recall. Most DSOs offer similar setup storage and recall features. The advantage is that, once a complex measurement setup has been made, it can be stored in memory. Then, at the press of a few buttons, the stored settings can be recalled from memory for automatic setup of the oscilloscope or DSO.

A typical memory specification is the number of front-panel setups that can be stored in the DSO or oscilloscope. Another memory specification is the number of wave forms that can be stored in memory. The latter specification is exclusive to DSOs since analog oscilloscopes do not allow wave-form storage.

The number of wave forms that can be stored by a DSO is a function of available memory and the wave-form record length. As pointed out in previous discussion, record length is the number of stored samples used to represent the digitized wave form. The longer the record length (e.g., 1024 samples vs 512 samples), the greater the memory requirement for each stored wave form.

In some DSOs, record length is selectable. Thus, selecting a shorter record length allows more wave forms to be stored. On the other hand, a longer record length provides greater time resolution, but uses more wave-form memory, allowing fewer wave forms to be stored in a given memory size.

Also, memory can be partitioned differently in various DSOs. In some cases, there is fixed partitioning, allowing a fixed number of wave forms and front-panel setups to be stored. In other cases flexible partitioning allows trade-offs between the number of wave forms and setups stored.

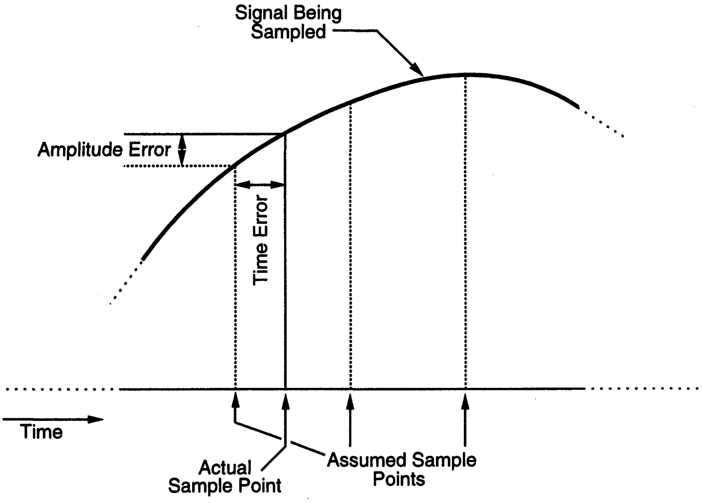

FIG. 4. Sample clocking inaccuracy results in two types of error. One is error in time placement of the sample. This results in a sample amplitude error that varies on the signal slope being sampled.

The key point is that, while memory and wave-form record length are important, the way memory is used is equally important. Applications requiring storage of numerous wave forms and few setups need memory partitioning in favor of wave forms. Applications requiring numerous front-panel setups generally benefit from memory partitioning in favor of setup storage.

4. ANALOG-TO-DIGITAL CONVERSION

Sampling and digitizing are the basic steps in analog-to-digital conversion (ADC) of wave forms. Up to this point, ADC has been discussed in a general sense. In reality, there are numerous ADC methods available. Each method has different advantages and limitations for wave-form acquisition and storage.

4.1 Digitizing Rate

The term "digitizing rate" is sometimes used in place of "sampling rate." To be correct, sampling rate is the rate at which amplitude samples are taken, and digitizing rate is the speed of converting an amplitude sample to a digital value. Thus digitizing and storage of an amplitude sample value must be completed before the next sample arrives for digitizing.

In terms of wave-form representation, the digitizing or sampling rate must be fast enough to define the wave form adequately. In the strictest sense, ADC must, at minimum, meet the Nyquist sampling rule. The Nyquist rule states that there must be two samples per cycle for the highest frequency of significance in the wave form being sampled.

For example, consider the case where the highest frequency of significance in a 20-MHz wave form is the fifth harmonic. Then, according to the Nyquist rule, the wave form must be sampled at a rate that is twice the frequency of the fifth harmonic. This would be a sampling rate of 200 MHz for the fifth harmonic of a 20-MHz wave form.

For most purposes, it is easier to think of sampling rate in terms of the DSO's bandwidth. At a minimum, the Nyquist sampling limit should be twice the DSO's bandwidth. For a 200-MHz DSO, this corresponds to a 400-MHz sampling rate at minimum. However, that also means that a 400-MHz sinusoid would be acquired with only two samples per cycle. While two samples per cycle may be adequate for mathematical description of the sinusoid, it is inadequate for graphical display of a recognizable sinusoid. Because of this, many DSOs use oversampling to provide more samples per cycle at the bandwidth limit. The effects of various numbers of samples per cycle on graphical representation are shown in Fig. 5.

4.2 Real Time and Equivalent Time

Real-time ADC means that all wave-form amplitude samples are taken, digitized, and stored in real time, as the wave form occurs. This is the ADC method necessary for capturing single-shot events. Samples are taken one after the other, in sequence, as the event transpires.

The difficulty with real-time ADC is the rates needed for adequate oversampling of the DSO's bandwidth. For example, with a 200-MHz DSO, oversampling at just 5 times the bandwidth means an ADC rate of 1 GHz. While 1-GHz and higher real-time ADC rates are certainly achievable, the technology is not inexpensive. Moreover, such high rates are not necessary when dealing with repetitive wave forms.

Repetitive wave forms can be repetitively acquired in a triggered mode (see Fig. 2). Since the wave-form repetitions are acquired over and over, a few samples can be taken at different points on each triggered acquisition. Over many triggered acquisitions, enough samples can be accumulated for a complete record of the wave form. This concept is illustrated further in Fig. 6. Essentially, the resulting composite of wave form samples is an equivalent-time representation of what would be achieved with a very fast real-time sampling rate.

Equivalent-time DSOs can capture repetitive wave forms with significant oversampling, even at their bandwidth limit. This is done economically by using low real-time sample rates and the correspondingly less-expensive ADC technology. For example, a 1-MHz real-time ADC can be used to achieve equivalent-time rates of hundreds of megahertz and even gigahertz for capturing repetitive wave forms. This assumes, of course,

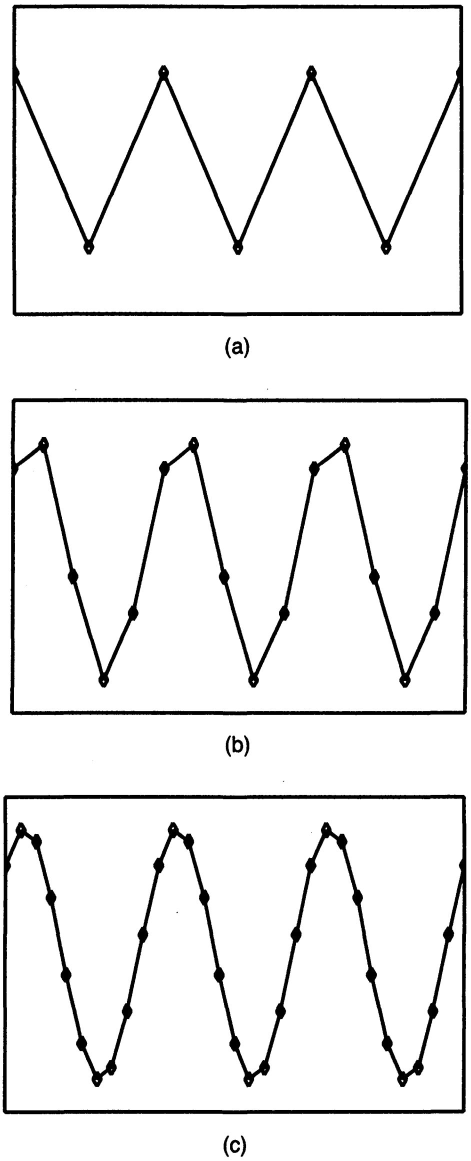

FIG. 5. The effects of various numbers of samples per cycle in accurate portrayal of a sinusoid—(a) two samples per cycle, (b) six samples per cycle, (c) 12 samples per cycle.

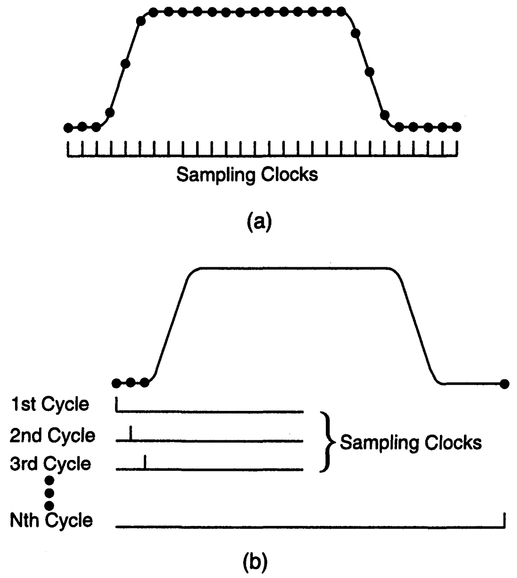

FIG. 6. Comparison of sampling methods. (a) In real-time sampling, samples are taken in sequence as the wave form actually occurs. (b) In equivalent-time sampling, samples are accumulated from numerous triggered acquisition cycles. For example, a 512-point record may take 512 acquisitions for complete sampling.

adequate analog bandwidth in the oscilloscope circuitry preceding the ADC.

There are some important cautions, however, when considering equivalent-time DSOs. Some DSOs with high equivalent-time capabilities gain their economic edge by using the lesser technology of very low real-time digitizing rates. This results in a very low single-shot capture capability for transients since single-shot capability is always limited to the real-time sample rate. Also, with a low real-time rate, full equivalent-time sampling can take longer because more repetitive acquisitions are necessary to build up a full wave-form sample record. Finally, any wave-form jitter or amplitude variations will have a degrading effect on the accuracy and appearance since the final record is a composite of many acquisitions over time. However, degradations from random effects can be reduced through the process of signal averaging.

4.3 Converter Types

A variety of converter types have been developed to meet various speed, resolution, and economic requirements. The major categories of converter types used in DSOs are successive approximation, flash conversion, charge coupled devices (CCDs), and scan conversion.

4.3.1 Successive Approximation These converters compare the input signal sample level to the output of a digital-to-analog converter (D to A). A bit-by-bit conversion is made starting with the most significant bit (MSB) and ending with the least significant bit (LSB).

Since each analog sample is digitized bit by bit, there is a trade-off between conversion speed and digital resolution (e.g., 8-bit vs 10-bit). For higher resolutions, longer conversion times must be accepted. To overcome the conversion-time trade-off, equivalent-time sampling methods are almost always used with successive approximation. However, the real-time conversion rate is still an important consideration for overall wave-form acquisition time and for any single-shot capture needs.

4.3.2 Flash Conversion These converters use a binary string of resistive dividers and comparators. The number of divider resistors and comparators is equal to the binary resolution of the converter. For example, an 8-bit converter would use an eight-level string of dividers and comparators.

The analog sample is applied to the top of the divider string, and the comparators output the corresponding binary levels in parallel. Since the conversion to all bits takes place in parallel, the method is quite fast, hence the name "flash" conversion.

However, there are several trade-offs to consider in flash conversion. One trade-off is that for higher resolution, more resistive elements and comparators must be strung together. This translates to greater converter complexity and expense. Also, the resistive divider string must be carefully designed to compensate for changing values at higher frequencies. Designing a high-performance flash converter is not a trivial task. Still, flash conversion offers an economical method for fast equivalent-time DSOs with real-time rates that are substantially higher than those using successive approximation.

4.3.3 Charge Coupled Device (CCD) A CCD is actually a very fast analog sampler and not a digitizer. Essentially, the CCD consists of a set of input cells, a memory section, and a set of output cells.

The key is that the CCD takes in analog samples at a high rate and transfers them to the memory section. The analog samples are held in memory and can then be read out and digitized at a slower rate, by successive approximation for example.

The advantage of the CCD technique is that high real-time sample rates can be achieved at modest cost. As reported by Tektronix (1992), a 100-MHz sample rate, 1-K record length, and 8-bit vertical resolution can be achieved at a modest cost compared to other methods used to achieve the same performance level.

4.3.4 Scan Conversion Scan converters typically provide the highest available real-time performance. As a result, they are only used in the class of DSOs referred to as transient digitizers.

A scan converter is basically a double-ended cathode-ray tube (CRT). One end provides a writing beam similar to that in a conventional analog oscilloscope CRT. However, instead of writing the wave-form trace onto the phosphor of a viewing screen, the trace is written onto a semiconductor target in real time. The target holds the trace, usually in the form of charge in a diode array. The reading beam at the other end of the CRT then scans the target and reads out the stored trace. This reading process can be done at a slow rate, allowing digitizing to be done by one of the more economical digitizing processes.

During scan conversion (reading of the target) no further wave-form information can be captured. However, this usually does not present a problem in transient wave-form capture since transients are infrequent occurrences.

4.4 Vertical Quantizing

Vertical quantizing is often referred to as quantizing noise. It is a function of digitizing, and thus occurs regardless of the method of digitizing.

Quantizing occurs when the analog value of a sample does not correspond exactly to one of the available digital levels. Since the digitizer cannot represent the analog level exactly, it must default to the closest digital level above or below.

This quantizing process has the effect of adding noise to the captured signal, hence the term "quantizing noise." The amplitude of this quantizing noise (or error) depends on the vertical resolution of the digitizer. An 8-bit digitizer, because of its lower resolution, causes higher quantizing noise amplitude than a 10-bit digitizer.

Quantizing noise is unavoidable. However, under certain circumstances, it can be significantly reduced, as reported by Ramirez (1985). This reduction occurs when repetitive wave forms are signal averaged by a processor having greater resolution than the digitizer. The result of such averaging can be that 8-bit digitizing can be cleaned up to be equivalent to 10-bit or even 12-bit digitizing.

4.5 Effective Bits

Digitizer resolution is generally stated as static resolution. That is, an 8-bit digitizer provides 8 bits of resolution in digitizing a dc voltage level. It also will generally provide 8-bit resolution for slowly changing wave forms. However, jitter, sampling uncertainty, and other effects begin to have a greater influence as wave-form frequency increases. These effects manifest themselves as increased noise on the digitized signal.

Since increased noise from sampling uncertainty is digitizing noise, it can, in effect, be categorized as quantizing noise. Thus, a higher quantizing noise translates to a lower digitizing resolution. This means that an 8-bit digitizer can actually appear to be of much lesser resolution when digitizing higher frequency signals. For example, at 75% of its sample rate limit, an 8-bit digitizer may appear to effectively be a 5-bit digitizer.

As mentioned previously, random noise effects can be overcome by signal averaging when repetitive wave forms are being acquired. Signal averaging cannot be done on single-shot or transient waveform acquisition. As a result, most transient digitizers have effective-bits specifications for different points across their bandwidth. In general, the higher the effective-bits specification, the higher the transient digitizer's quality.

5. DISPLAY TECHNIQUES

Digitized wave forms are stored in memory as an array of discrete amplitude values. These discrete values are then processed to present a visual display representing the original analog wave form. How this display is created is a matter of preference as well as a matter of how various types of wave forms should be represented for accurate interpretation.

5.1 Dot Displays

Dot displays are the simplest of all displays. These consist of simply displaying each sample value from memory as a dot. The vertical position of the dot represents amplitude and the horizontal position represents time.

In truth, the dot display is the most accurate representation of the acquired wave form. This is because the dots are direct plots of the acquired samples, and the acquired samples are the only recorded wave-form data. What happens between samples is speculation because no wave-form information was acquired between samples.

In practice, dot displays can be horribly confusing. This is shown in Fig. 7, where several sinusoids of different frequencies are represented by dot displays. The problem is that, when the dots are moved closer together horizontally, the eye is tricked into picking out different wave-form patterns. Generally, the eye picks out lower-frequency

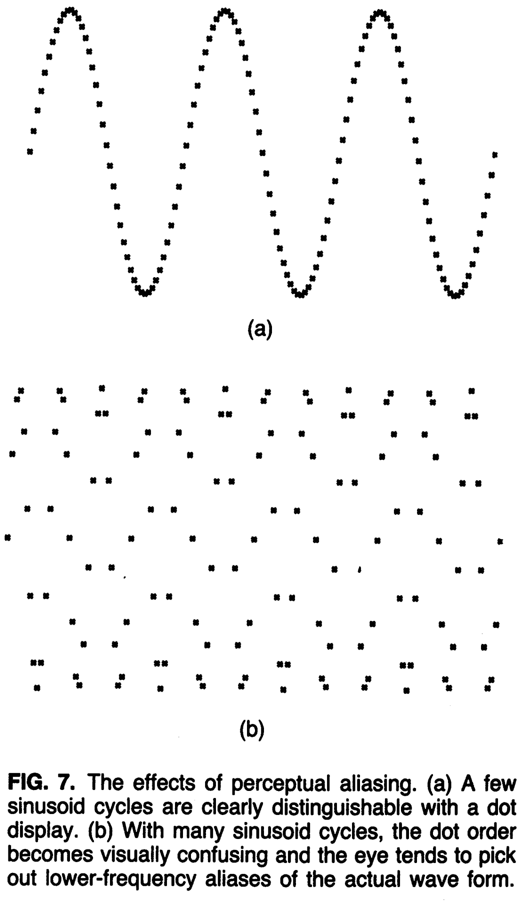

FIG. 7. The effects of perceptual aliasing. (a) A few sinusoid cycles are clearly distinguishable with a dot display. (b) With many sinusoid cycles, the dot order becomes visually confusing and the eye tends to pick out lower-frequency aliases of the actual wave form.

wave forms than actually represented by the dots. This effect is referred to as perceptual aliasing.

5.2 Interpolated Displays

To avoid perceptual aliasing, most DSOs provide interpolated displays. Most often the interpolation is linear. That is, a dot display is transformed by connecting the dots in time order with straight lines.

The effect of straight line interpolation is shown in Fig. 8. Note that, in comparison to the dot displays of Fig. 7, interpolated displays eliminate any confusion about the actual wave form path.

Linear interpolation provides a very close approximation of the actual wave shape for wave forms composed primarily of linear segments—e.g., square waves, ramps, triangular waves. For wave forms that follow nonlinear forms, such as sinusoids, other forms of interpolation may be preferred. Not surprisingly, sine interpolation is the most common form used in DSOs after linear interpolation. This is because sine interpolation provides a closer approximation for smooth wave forms.

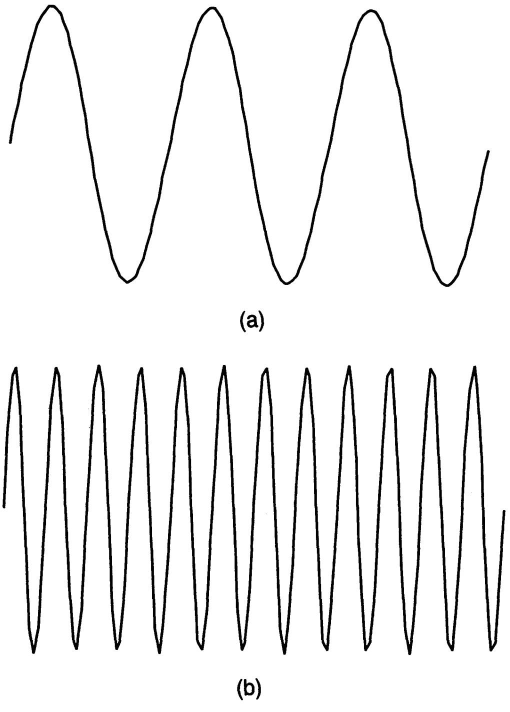
(a)

(b)

FIG. 8. Linear interpolation or vector display. When a dot display is straight-line connected in dot order, wave forms are easier to distinguish. Compare (b) here to (b) in Fig. 7.

5.3 Peak Detection and Enveloping

Peak detection and enveloping are both special methods of handling samples to reveal special wave-form situations. Peak detection, in general, ensures capture of wave-form peaks or spikes that many times are not captured with ordinary sample processing. Enveloping is somewhat different and allows an envelope of maximum wave-form excursions to be captured.

With ordinary sample processing, the wave-form record is filled with samples taken at equally spaced points in time (Δt). The time spacing, or Δt, corresponds to the display-time window (total horizontal sweep time) divided by the displayed record length. For example, say that the DSO's sweep-speed setting is 0.01 s/division. For a 10-division horizontal display, this corresponds to a display-time window of 0.1 s. Thus, with a displayed record of 1024 sample points, the sample spacing (Δt) is 0.1 s/1023 = 0.97752 μs.

The problem with ordinary sampling is that short-duration details, such as a noise spike, can easily fall between samples and be missed. The likelihood of this happening increases with the greater sample spacing that accompanies acquisition of slower wave forms.

Peak detection minimizes the loss of small details, such as spikes, that could fall between samples. The method involves always sampling at full speed regardless of the sweep-speed (time base) setting. With a 100-MHz real-time sample rate, this corresponds to a fixed sample interval (Δt) of 1 ns. However, on longer time windows this results in far more samples than there is record length storage. A 0.1-s window with a Δt of 1 ns corresponds to 100,000,000 samples. For a stored record length of 1024 samples, only selected samples are stored and the excess samples are discarded.

For peak detection, the maximum-amplitude sample in any stored sample interval is retained for the stored wave-form record. This is illustrated in Fig. 9. The result is that for any sweep speed, no matter how slow, there is a very high probability of capturing wave-

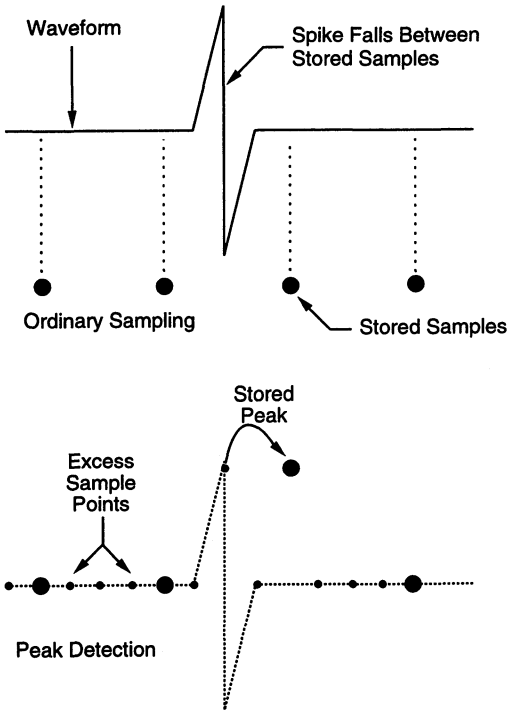

FIG. 9. Peak detection. With ordinary sampling, narrow wave-form spikes may occur between samples and may not be captured. Peak detection uses oversampling and picks the maximum valued sample for storage. This guarantees at least one sample on narrow spikes.

form peaks (or spikes) as narrow as the smallest sample interval. In the case of 100-MHz maximum sampling, this is referred to as 1-ns peak detection since peaks to 1 ns wide can be detected.

The key benefit of peak detection is that there is always an indication that signal spikes exist, even when operating at the slowest sweep speed. With this knowledge, a higher sweep speed can be used for more detailed capture and examination of the spikes. Without the initial indication provided by peak detection, signal spike problems could easily be missed.

Enveloping, another very useful process, can be implemented with or without peak-detection sampling. When implemented in conjunction with peak detection, enveloping is far more sensitive to minor wave-form changes.

Basically, enveloping involves the use of two stored wave-form records. One is used for maximum sample values and the other for minimum sample values. Over multiple wave-form acquisitions, an envelope display is created of maximum and minimum excursions of the wave form. Such an envelope display is very useful for detecting and studying jitter, drift, or other wave-form variations over time.

5.4 Pre- and Post-Trigger Displays

The majority of the displayed trace on an analog oscilloscope is a post-trigger display. It is the trace of the wave form occurring in time immediately after the trigger point (refer to Fig. 2).

In some oscilloscopes, a trigger hold-off feature allows the display to be delayed to a selected time after the trigger point. This is useful for viewing wave-form activity that occurs some time after a selected triggering point or triggering event. For example, the firing pulse for a laser could be used as a trigger, and trigger hold-off could be used to study laser activity some time after the firing pulse.

Often, it is important to see wave-form activity prior to the trigger point. This is referred to as pretrigger acquisition. Pretrigger acquisition is provided in some analog oscilloscopes by placing a fixed delay line in the signal path after the trigger-system pick-off point [refer to Fig. 1(a)]. This allows the displayed signal trace to include a portion of the signal occurring prior to the trigger point. The duration of this pretrigger portion of the display is equal to the time delay of the inserted delay line.

The pretrigger portion of the display shows wave-form activity prior to the trigger point. This is quite useful, for example, in capturing the starting point of pulses or transient wave forms. However, in analog oscilloscopes, delay-line pretriggering may not be sufficient for many needs. This will be especially true when the trigger level must be set high in order to avoid false triggers on noise or other aberrations (see Fig. 10).

DSOs get around the analog limits to pretrigger length in a fairly simple manner. This involves continuously sampling the input signal and shifting the samples through the record in memory. The segment of samples that is locked into memory as a stored wave-form record is determined by triggering on

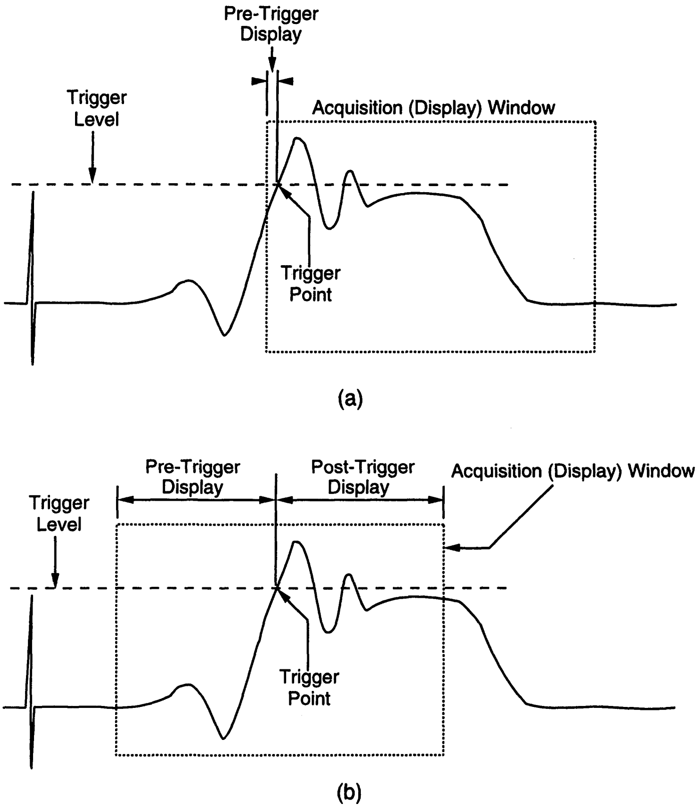

FIG. 10. Pre- and post-triggering. (a) Analog oscilloscopes use a fixed delay line to provide a small amount of pretrigger display. (b) Digital storage oscilloscopes can provide varying amounts of pre- and post-triggered wave-form information. This allows triggering above noise while still displaying the entire leading edge of a pulse.

the wave form and the amount of time that has been selected for pre- or post-triggering.

For example, let us say the DSO has been set up to provide half the display as pretrigger information and the other half as post-trigger information. The DSO is continuously acquiring samples and shifting them through memory while waiting for the triggering event. At the moment triggering occurs, samples existing in memory represent a record of 100% pretriggering information. Since the example setup calls for a record of 50% pretrigger and 50% post-trigger samples, the DSO continues acquiring samples after triggering until half a record length of post-trigger samples is in memory. At this point, the samples are frozen in memory as a wave-form record of half pretrigger and half post-trigger information, as shown in Fig. 10.

There are several key benefits to the adjustable pre- and post-triggering offered by DSOs. The first benefit is that trigger levels can be set high on the wave form to avoid false triggers, and the full wave form can still be viewed by using pretriggering to include the wave-form starting point. Another key benefit or use is the ability to look well ahead of the triggering point or event to study activity leading up to the triggering event. An example of the latter use would be study of charge buildup, corona, or other effects preceding high-voltage arcing.

6. USEFUL STORAGE BANDWIDTH AND RISE TIME

Previous discussion mentioned bandwidth and rise time as key specifications in selection and analog oscilloscope. The same analog specifications also apply to the signal input circuitry of a DSO. This is necessary for accurate presentation of the analog signal to the sampling and digitizing circuitry. Additionally, there are various other factors that affect the usefulness of the digitized signal display. Dot displays and perceptual aliasing are two of the factors already discussed.

Rather than trying to interpret the interactions of analog bandwidth and rise time and other various digitizing effects, combined figures of merit have been established for overall DSO evaluation. These figures of merit are referred to as useful storage bandwidth and useful storage rise time.

6.1 Useful Storage Bandwidth

In general, useful storage bandwidth (USB) is defined as follows for dot displays:

$$f_{\mathrm{USB}} = (\text{maximum digitizing rate})/25. \tag{3}$$

The value of 25 in the above relationship is the number of samples per cycle generally required for visual recognition of wave shapes on a dot display. For linearly interpolated display (vector connection of dots), the divisor value can be changed to 10.

6.2 Useful Storage Rise Time

With a DSO, the displayed rise time depends upon where samples fall on the actual rising wave-form transition. With a vector-interpolated display, the maximum rise-time errors on the DSO display closely follow the form of an analog oscilloscope's errors when the analog system has a rise time of 1.6 sample intervals. As a result, useful storage rise time (T_r) can be defined as follows:

$$T_r = (\text{minimum sample interval}) \times 1.6. \tag{4}$$

The above figure of merit applies to the worst-case error for rise-time measurement on a DSO vector display.

7. SIGNAL PROCESSING

One of the more powerful possibilities of a DSO is the ability to provide a wide variety of wave-form processing and analysis capabilities. Since the stored wave forms are essentially a series of numbers, virtually any kind of numerical analysis can be performed. Whether or not such analysis is available as a built-in feature depends on the particular DSO.

Most DSOs do provide measurement cursors for reading out the amplitude and time values of stored wave forms. Additionally, most DSOs provide the simpler built-in analysis functions such as wave-form maximum, minimum, average, and rms (root mean square) values.

High-end DSOs generally provide all of the simpler built-in functions as well as one or more advanced built-in functions. Some of the advanced functions that might be provided are briefly discussed below.

7.1 Save-on-Delta

Save-on-Delta is a function that relies on first storing an envelope of expected wave-form limits. Depending on the DSO implementation, the upper and lower limit traces can be generated manually or by downloading data to the DSO.

Once the upper and lower limit data, or envelope, is in place, Save-on-Delta acquisition of an actual wave form can begin. During acquisition, each sample of the wave form is compared to the upper and lower limit envelope. If the acquired samples do not violate the limits envelope, the wave form is not saved. If any sample does violate the limits envelope, the wave form is saved.

This acquisition mode is very useful for catching intermittent or infrequent wave-form conditions. A DSO can be left unattended in Save-on-Delta mode for as long as needed to capture limit-violation wave forms.

7.2 Averaging

Averaging can be applied to repetitively acquired wave forms to provide improvements in signal-to-noise ratio (SNR). Whenever possible, averaging should be applied to improve measurement accuracy.

The usual form of averaging is to select a number of acquisitions to be averaged—for example, 64 averages. Averaging then takes place by acquiring the first wave form. A second wave form is then acquired and automatically added to the previous wave form. This acquisition and summing of amplitude samples continues until the specified number of wave forms has been acquired and summed. The summed wave forms are then divided by the specified averaging number.

The result of averaging is that randomly occurring events (noise) on repetitively acquired wave forms become reduced, as reported by Ramirez (1985). When averaging is done at computation resolutions exceeding the digitizer's amplitude resolution, the effects of quantizing noise can also be reduced. This can result in a higher effective measurement resolution on the averaged wave form.

7.3 Integration and Differentiation

Stored wave forms can be integrated or differentiated using various numerical analysis routines. Typically integration and differentiation are done in a point-by-point (two point) process.

Integration is simply the process of calculating the rectangular area represented by a pair of points and summing those areas. The result of such integration provides a point-by-point plot of energy contained under the wave-form curve.

Differentiation is simply the point-by-point slope of a wave form. This is useful for defining break points on wave forms. A particularly useful application, as reported by Ramirez (1985), is in acquiring the impulse response of electronic circuits. Since a step voltage is easier to generate and control as a stimulus signal, a circuit step response is easier to acquire in most cases. When acquired with a DSO, the step response can be differentiated to provide an impulse response display.

7.4 Fast Fourier Transform

The classical Fourier transform is an analog method of transforming a time-domain signal to the frequency domain. This allows detailed analysis of a signal's frequency content. The discrete Fourier transform (DFT) is the classical numerical analysis approach to obtaining Fourier transform data. The fast Fourier transform (FFT) is simply a computationally efficient method of computing the DFT.

Some DSOs contain a built-in FFT algorithm. This allows acquired wave forms to be transformed from time-series displays to frequency-spectrum displays.

7.5 Convolution and Correlation

Convolution and correlation functions are very easy to compute in the frequency domain. For those DSOs having an FFT algorithm, the additional steps of providing convolution and correlations are often taken.

Convolution is a key computation used in transfer-function analysis. Correlation has many uses in signal identification and matching.

8. MEASUREMENT AUTOMATION

Since most oscilloscopes today are digitally controlled, they offer the opportunity for measurement automation. One form of automation has already been mentioned. That form is push-button storage and recall of instrument front-panel setups. Another form of automation just discussed is the signal processing capabilities that can be provided as built-in functions of DSOs.

Both forms of automation can be taken another step further by interfacing the instrument to a computer. When this is done, the automation resources are vastly expanded through software programming, disk storage, and setup and wave-form information transfers between the computer and instrument.

8.1 Interfacing Techniques

The basis for most stand-alone instrument interfacing today is the General Purpose Interface Bus (GPIB) as defined by the IEEE-488 specification. A single GPIB loop can contain up to 16 devices. One of the devices must be the GPIB controller (a personal computer for example), which leaves a potential of 15 devices for GPIB control. These devices can be DSOs, programmable analog oscilloscopes, arbitrary wave-form generators, power supplies, or any other type of GPIB instrumentation.

Another popular form of instrument interfacing is through an RS-232 port. This has become popular because most personal computers come standard with one or more RS-232 ports. The actual interfacing is still usually via the GPIB standard, but it is accomplished through RS-232/GPIB translation in the instrument's interfacing hardware and firmware.

RS-232 interfacing has another important advantage in that it is compatible with telephone modems. This allows an instrument to be remotely located and accessed through telephone lines and modem connection. If an RS-232/GPIB controller is used at the remote site, an entire GPIB instrument system can be remotely accessed and controlled over a single phone line.

8.2 Software Considerations

There are two key considerations in looking at software for automation applications. One consideration is the instrument interfacing and control software. The other is the wave-form handling and analysis software.

Most GPIB instrument manufacturers provide software drivers for their specific instrument models. These drivers are usually created so that they can be linked as modules into standard software platforms, such as BASIC or UNIX. A more important consideration, however, is command and syntax consistency across various instruments. If the manufacturer does not follow some form of codes and formats standard in creating drivers for various instruments, control of multiple instruments can become extremely complex.

After instrument control and wave-form transfer to the computer, there is the consideration of wave-form processing. This can be done by writing programs from the ground up using BASIC, C, or any other popular software platform. However, a tremendous amount of work can be saved by using any of several available wave-form processing packages. These packages provide specialized wave-form processing capabilities and often include many advanced analysis features such as FFT, convolution, correlation, modal analysis, and so forth. Essentially, once the acquired wave forms have been transferred to a computer, the analysis possibilities are endless.

9. CONCLUSION

Virtually any wave-form capture and analysis need can be met by one of the analog oscilloscopes or DSOs on the market today. The primary issue is matching application needs to a specific instrument type. This includes the following major steps:

1. Defining analog bandwidth and rise-time requirements.
2. Determining whether analog display or stored wave-form display is best for the application. Generally, digital storage is preferred for very slow wave forms, transient wave forms, or any wave form where substantial measurement or analysis is necessary. Analog oscillscopes are preferred for real-time wave-form viewing needs, such as in making circuit adjustments, and for multi-GHz transient recording.
3. Determining digital resolution in terms of both amplitude (digitizing resolution) and record length (generally longer records are needed for fast-rise, slow-decay wave forms).
4. Specifying sampling technique and rates—real time for single-shot transients and equivalent time for repetitive waveforms.
5. Wave-form storage requirements. Built-in instrument storage may be adequate for applications focused on a few wave forms, but large quantities of wave-form data usually require external storage (e.g., interfacing to a computer).
6. Considering the benefits that might be derived from the automated instrument control and advanced analysis techniques available through computer interfacing.

Careful consideration of the above issues in terms of specific application requirements will lead to a set of specific wave-form acquisition needs. This set of specific needs provides a clear definition of the specific instrument requirements for the application. This matching the tool to the job—rather than bending the job to fit an existing tool—leads to greater success in quickly obtaining useful measurement and analysis results.

It is important, however, to realize that both analog and digitizing oscilloscope technologies are marching to a quick pace and often leapfrog each other. This makes it risky to formulate any set of rules that dictates exclusive use of one technology over another for any given application need. Thus, evaluation of measurement needs against available oscilloscope technologies is an ongoing process.

As far as future trends are concerned, it is almost always safe to say that acquisition bandwidths and speeds will continue to increase. Also, with the digital age in full bloom, both analog and digitizing oscilloscopes are offering an expanding array of automation and measurement processing features. As a sidelight to this, industry observers speculate from time to time that oscilloscopes will eventually be compressed to a card that plugs into a personal computer. Current industry trends, however, suggest that the opposite is actually happening. More and more specialized computing power is being compressed into oscilloscopes. Whatever course future implementations take, the key oscillographic concepts will remain the same—acquisition, processing, and display of wave forms.

GLOSSARY

Aliasing: A condition where one frequency takes on the appearance of another frequency. In sampling, a minimum of two samples per cycle is required to avoid aliasing. In dot displays, perceptual aliasing can be prevented by connecting the dots in their correct time order.

Bandwidth: The continuous band of frequencies passed or transmitted by a device without undue attenuation. The bandwidth or pass-band edges are often defined as the points where amplitude becomes attenuated by 3 dB (half power).

Cathode-Ray Tube (CRT): A vacuum-tube display device consisting of an electron beam, an electronic drive and deflection system, and a phosphor-coated display screen. The deflection system traces the electron beam across the phosphor coating on the back of the screen causing luminescence of the phosphor along the beam path for a visible display of the trace on the front of the screen.

Equivalent-Time Bandwidth: The highest attainable bandwidth using equivalent-time sampling (one-half the equivalent-time sample rate).

Equivalent-Time Sampling: Any wave-form sampling method where a full wave-form record is built up over multiple, repetitive acquisitions of a wave form. Equivalent-time sampling allows capture of high-speed repetitive wave forms while using lower speed sampling and analog-to-digital conversion. Equivalent-time sampling cannot be applied to the capture of single-shot, transient wave forms.

Fast Fourier Transform (FFT): An efficient computer algorithm for computing the discrete Fourier transform of sampled waveform data. Time-domain samples are converted to frequency-domain data or vice versa.

Persistence: In reference to CRT phosphors, the property of persisting in light emission after excitation has ceased. Oscilloscope traces fade more slowly with high-persistence phosphors than with low-persistence phosphors.

Probe: Device used to sense or pick up signals and convey them to an instrument for observation or measurement. Voltage probes are commonly used with oscilloscopes, but current and temperature probes are generally available, too. In many cases, the probe is used to pick up the output of another sensor such as a strain gauge or accelerometer.

Probe Compensation: Adjustment of probe capacitance so that the probe attenuates all input signals equally across a range of frequencies.

Real-Time Bandwidth: The highest attainable bandwidth using real-time sampling (one-half the real-time sample rate).

Real-Time Sampling: Any wave-form sampling method where samples are taken sequentially in real time as the wave form actually occurs. Real-time sampling is required for the capture of single-shot, transient events and is equally applicable to the capture of any other class of wave form (repetitive, random, etc.).

Record Length: The number of samples acquired and stored as a digitized representation of a wave form.

Rise Time: The time required for the leading edge of a pulse to rise from 10 to 90% of its final value.

Sweep Speed: The speed at which the CRT display's electron beam is driven horizontally across the screen. Sweep speeds are usually in terms of seconds per display division, providing a time-calibrated display.

Time Base: In oscilloscopes, an electronic sweep or clocking system used to provide time-calibrated acquisition and display of wave-form information.

Triggering: The process where selected wave-form conditions cause acquisition of the wave form to begin.

Trigger Hold Off: A period of time when triggering is prevented. This is usually done to prevent a second trigger during the horizonal sweep period. Trigger hold off may also be used to provide a delayed sweep, allowing acquisition to occur at a selected period of time after the triggering event.

Trigger Point: The point on a wave form that causes triggering of wave-form acquisition.

Works Cited

Gustafson, M., Johnson, L., Caryl, M., Gifford, T., White, D. (1985), "State-of-the-Art Trends in Oscilloscopes," *Test & Measurement World*, April, 105–113.

Ramirez, R. W. (1985), *The FFT Fundamentals and Concepts*, Englewood Cliffs, NJ: Prentice-Hall, pp. 101, 151.

Tektronix (1992), *An Introduction to Digital Storage*, Report No. 41W-6051–4, Beaverton, OR: Tektronix, Inc.

Tektronix (1993a), *The Effects of Bandwidth on Measurement Accuracy*, Report No. 85W-8907–0, Beaverton, OR: Tektronix, Inc.

Tektronix (1993b), *ABC's of Probes*, Report No. 60W-6053–4, Beaverton, OR: Tektronix, Inc.

Further Reading

IEEE (1994), *IEEE Standard for Digitizing Waveform Instruments*, IEEE Standard 1057-1994, New York: IEEE.

Baumgart, C., Dunham, M., Moses, J. (1991), "Curve Fitting and Error Modeling for the Digitization Process Near the Nyquist Rate," *IEEE Trans. Instrum. Meas.* **IM-40**, 553–557.

Peetz, B. (1983), "Dynamic Testing of Waveform Recorders," *IEEE Trans. Instrum. Meas.* **IM-32**, 12–17.

Ramirez, R. W. (1985), *The FFT: Fundamentals and Concepts*, Englewood Cliffs, NJ: Prentice-Hall, Inc.

Tektronix (1992), *An Introduction to Digital Storage*, Report No. 41W-6051-4, Beaverton, OR: Tektronix, Inc.

Textronix (1992), *Real-Time Digital Signal Processing Techniques Address High-Resolution Measurement Requirements*, Technical Brief No. 3GW-8222-1, Beaverton, OR: Tektronix, Inc.

Tektronix (1992), *The XYZs of Using a Scope*, Report No. 03W-8605-0, Beaverton, OR: Tektronix, Inc.

Tektronix (1992), *Real-Time Versus Equivalent-Time Sampling*, Technical Brief No. 55W-8628-0, Beaverton, OR: Tektronix, Inc.

Tektronix (1993), *ABC's of Probes*, Report No. 60W-6053-4, Beaverton, OR: Tektronix, Inc.

Tektronix (1993), *Sampling Oscilloscope Techniques*, Technique Primer No 47W-7209, Beaverton, OR: Tektronix, Inc.

Tektronix (1993), *The Effect of Probe Input Capacitance on Measurement Accuracy*, Technical Brief No. 60W-8910-0, Beaverton, OR: Tektronix, Inc.

Tektronix (1993), *The Effects of Bandwidth on Measurement Accuracy*, Technical Brief No. 85W-8907-0, Beaverton, OR: Tektronix, Inc.

OSMOSIS

BENNY FREEMAN, *Department of Chemical Engineering, North Carolina State University, Raleigh, North Carolina, U.S.A.*

	Introduction	59
1.	**Molecular and Thermodynamic Description of Osmosis**	60
1.1	Molecular Basis of Osmosis	60
1.2	Thermodynamic Analysis of Osmosis	61
1.2.1	Osmotic Pressure of an Ideal, Nonionic Solution	61
1.2.2	The Influence of Solute Molecular Weight on Colligative Properties	62
1.2.3	Osmotic Pressure of a Solution Containing Several Solutes	63
1.2.4	Osmotic Pressure of a Dilute Ideal Electrolyte Solution	63
1.2.5	Osmotic Pressure of Dilute Ideal Protein Solutions	64
1.2.5.1	Donnan Equilibrium	65
1.2.6	Effect of Solution Nonideality on Osmotic Pressure	66
2.	**Experimental Methodology for the Determination of Osmotic Pressure**	68
3.	**Applications**	69
3.1	Reverse Osmosis	69
3.2	Dialysis	69
3.3	Electrodialysis	70
	Glossary	70
	Works Cited	71
	Further Reading	71

INTRODUCTION

Osmosis [from the Greek *ōsmos*, the action of pushing, derived from *ōthein*, to push (Morris, 1981)] is the phenomenon resulting from the tendency of a pure solvent to diffuse into a solution separated from the solvent by a membrane that is permeable only to the solvent (Atkins, 1990). Many important examples of osmosis are found in nature. Osmosis drives the transport of oxygen and carbon dioxide across capillary walls separating blood from air in the lungs. Fresh fruit will shrivel when placed in contact with strong sugar solutions, and dried fruit can swell in contact with water as a result of osmosis (Marick, 1969). The delicate balance of chemical substances inside living cells is maintained, in part, by osmosis. Furthermore, osmosis has been exploited scientifically as a sensitive probe of molecular weight of natural and synthetic macromolecules (Tanford, 1961). Reverse osmosis is an important industrial process for making drinking water from saltwater. A feed stream such as brackish water or seawater, is contacted with a membrane that is permeable to water but impermeable to microorganisms, colloids, dissolved salts, and organic species. The feed stream is pressurized to force permeation of pure, drinking-quality water across the membrane (Bhattacharyya and Williams, 1992). In the 1940s, the concept of artificial kidney dialysis was demonstrated and has become widely used to prolong the lives of people suffering from renal failure (Kessler and Klein, 1992). This technology is based on the osmotic transport of metabolic waste products from blood into a saline solution. The blood and saline solution are separated by a membrane that is permeable to the waste products.

Osmosis can be observed by separating a solution of sugar water from pure water by a semipermeable membrane made of a material such as cellulose, which permits small water molecules to permeate through the membrane but does not allow permeation of the larger solute molecules (in this case sugar) (Atkins, 1990). At the start of the experiment, the osmometer shown in Fig. 1(a), containing a sugar-water solution, is dipped into a bath of pure water to a depth such that the sugar solution in the tube is at the same level as

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

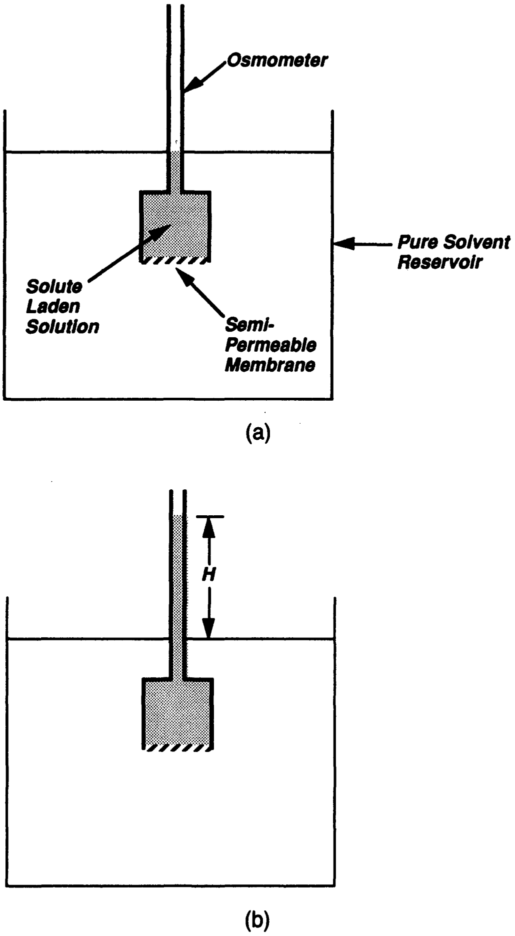

FIG. 1. Schematic of experiment to detect osmotic pressure of sugar water solution.

the pure water. As time proceeds, pure water diffuses across the semipermeable membrane into the solute-containing solution. As the pure water enters the sugar-water solution, it raises the level of the solution in the tube, and the pressure of the solution at the membrane face increases. Figure 1(b) illustrates the system at equilibrium, when there is no net transfer of solvent across the membrane, and the pressure of the sugar-water solution is higher than the pressure of the pure water in contact with the membrane. The value of the pressure difference across the membrane is $(\rho - \rho_{\text{air}})gH$, where ρ is the solution density, ρ_{air} is the density of air in the tubes above the solvent and solution, g is the average gravitational acceleration, and H is the difference between the level of sugar water in the capillary above the sugar-water reservoir and the level of pure water. This pressure difference is the osmotic pressure π.

Osmotic pressure is a thermodynamic property of a solution related to the solvent chemical potential in the solution (Denbigh, 1981). The experiment described in Fig. 1 is, therefore, an operational method for determining osmotic pressure, but the definition of osmotic pressure does not rely on the presence of a membrane or a bath of pure solvent. Like freezing point or boiling point temperatures, it is an intrinsic property of the solution. In this regard, the osmotic pressure does depend on solute concentration and nature since it is the presence of the solute that depresses the solvent chemical potential below its value in pure solvent. This chemical potential decrease, in turn, leads to experimentally observable manifestations of osmotic pressure (such as the increase in the height of solution in the capillary column described above). To provide some perspective of the magnitude of the osmotic effect, Table 1 presents values of osmotic pressure for several systems of interest. In this table, the increase in solution osmotic pressure with increasing solute concentration is observed.

1. MOLECULAR AND THERMODYNAMIC DESCRIPTION OF OSMOSIS

1.1. Molecular Basis of Osmosis

The presence of a solute in a solution can depress the solvent chemical potential by diluting the solvent (i.e., the solvent mole fraction is less in the solution than in pure solvent) (Denbigh, 1981). When a reservoir of solution is separated from pure solvent by a semipermeable membrane, the solvent chemical potential difference between the two solutions is a driving force for molecular diffusion of solvent from the solvent reservoir across the membrane into the solution. As solvent diffuses across the membrane, a pressure difference between the two reservoirs is established. The magnitude of this pressure difference increases until the solvent chemical potential in the solution is exactly equal to the chemical potential of the pure solvent. The equality of solvent chemical potential on both sides of the membrane is the criterion

Table 1. Typical values of osmotic pressure at 25 °C (Riley, 1990).

Solute	Concentration (mg/L)	Concentration (mol/L)	Osmotic pressure (bar)
NaCl	2 000	0.034 2	1.57
NaCl (isotonic solution)[a]	9 050	0.155	7.85
NaCl	35 000	0.60	27.4
Brackish water	2 000–5 000	0.034–0.086	1.0–2.7
Seawater	32 000		23.4
Sucrose	1 000	0.002 92	0.072
Dextrose	1 000	0.005 5	0.14

[a] Isotonic with mammalian plasma (Woodbury 1965).

for thermodynamic equilibrium (Denbigh, 1981). At equilibrium, there is solvent exchange between the two reservoirs due to molecular diffusion, but there is no net transport of solvent from one side of the membrane to the other. The pressure difference across the membrane at equilibrium is defined as the osmotic pressure.

1.2. Thermodynamic Analysis of Osmosis

The chemical potential μ_0 of a solvent in a solution at temperature T and pressure P relative to the chemical potential of pure solvent at the same temperature and pressure may be expressed as (Denbigh, 1981)

$$\mu_0 = \mu_0^0 + RT \ln a_0, \tag{1}$$

where μ_0^0 is the chemical potential of the pure solvent at temperature T and pressure P, R is the gas constant, and a_0 is the solvent activity. If the solution pressure is increased from P to $P + \pi$, the solvent chemical potential becomes

$$\mu_0 = \mu_0^0 + RT \ln a_0 + \int_P^{P+\pi} \bar{v}_0 dP, \tag{2}$$

where $\bar{v}_0$ is the partial molar volume of solvent in the solution (Prausnitz, 1969). Solvent in a solution will be in equilibrium with pure solvent when

$$\ln a_0 = -\frac{1}{RT} \int_P^{P+\pi} \bar{v}_0 dP. \tag{3}$$

The osmotic pressure π is, therefore, the pressure that must be applied to the solution to raise the solvent chemical potential in the solution to its pure-solvent value. The solvent partial molar volume is, to a very good approximation, independent of pressure (Denbigh, 1981), and therefore the osmotic pressure may be written as

$$\pi = -(RT/\bar{v}_0) \ln a_0. \tag{4}$$

1.2.1 Osmotic Pressure of an Ideal, Nonionic Solution The thermodynamic activity of solvent in an ideal solution is related to the solvent concentration as

$$a_0 = x_0, \tag{5}$$

where x_0 is the mole fraction of solvent (Denbigh, 1981). The reference state for chemical potential chosen for Eq. (1) is pure solvent and, accordingly, the solvent activity is 1.0 when the mole fraction of solvent is 1.0. As the solvent is diluted by solute, the solvent mole fraction and, in turn, activity decrease, which leads to a decrease in the solvent chemical potential according to Eq. (1). The solvent partial molar volume is equal to the pure-solvent molar volume in an ideal solution (Prausnitz, 1969). For an ideal solution, therefore, Eq. (4) becomes

$$\ln x_0 = -\pi \hat{v}_0/RT, \tag{6}$$

where $\hat{v}_0$ is the pure solvent molar volume. For dilute solutions containing a single solute,

$$\ln x_0 = -\pi \hat{v}_0/RT = \ln(1 - x_1) \approx -x_1, \tag{7}$$

where x_1 is the solute mole fraction. The solute mole fraction is the number of moles of solute, m_1, divided by the total number of moles of solvent and solute in the solution,

m_{tot}. In dilute solution, the total volume of solution, V, is not appreciably different from the product of m_{tot} and $\hat{v}_0$. In this case, Eq. (7) reduces to

$$\pi V = m_1 RT. \qquad (8)$$

Equation (8) is the van't Hoff equation for the osmotic pressure of a dilute ideal solution (Atkins, 1990). The similarity of this result with the ideal gas law illustrates the analogy between an ideal solution and an ideal gas (Atkins, 1990). Experimental determinations of the osmotic pressure of dilute solutions, which led to the empirical formulation of Eq. (8) well before its derivation from thermodynamic principles, inspired van't Hoff to develop the theory of ideal solutions based on the analogy with ideal gases (Mason, 1991). From Eq. (8), the osmotic pressure in an ideal solution is directly proportional to the number of solute molecules in the solution but does not depend upon the chemical nature of the solute.

Osmometry, the technique of using osmosis to determine the molecular weight of macromolecular species, has been widely practiced. To illustrate the sensitivity of osmotic pressure to solute molecular weight, Eq. (8) may be rewritten as follows:

$$\pi / c_1 = RT/M_1, \qquad (9)$$

where c_1 is the solute concentration (mass of solute/volume of solution) and M_1 is the solute molecular weight (g/mol). Experimentally, osmotic pressure is determined as a function of solute concentration at fixed temperature because, in real solutions, π / c_1 may not be independent of c. By plotting the ratio π / c_1 vs c_1 and extrapolating to infinite dilution (i.e., $c_1 \rightarrow 0$), RT/M_1 and, in turn, solute molecular weight may be determined. An example of using this technique to determine the molecular weight of a polymeric solute is presented later in this article.

1.2.2 The Influence of Solute Molecular Weight on Colligative Properties Solution properties such as the osmotic pressure of an ideal solution, which depend on the number, but not the chemical nature, of solute molecules in solution, are called colligative properties (Atkins, 1990). Other colligative properties, such as boiling point elevation, freezing point depression, and vapor pressure lowering, may be measured to determine solute molecular weight (Billmeyer, 1984). However, solution osmotic pressure is often the most practical colligative property to measure for high–molecular-weight solutes such as polymers. For example, Table 2 presents the coefficients characterizing the sensitivity of colligative properties to solute concentration when a high–molecular-weight solute is dissolved in benzene. For a solute molecular weight of 10 000 g/mol and a solute concentration of 0.1 g/cm^3, for example, the osmotic pressure would be 2.8 mm of benzene,

Table 2. Effect of solute molecular weight on colligative properties of a benzene solution. The pure-solvent normal freezing point is T_f. The freezing point depression is ΔT_f. The pure-solvent density and molar enthalpy of fusion are given by ρ and ΔH_f, respectively. The pure-solvent normal boiling point and enthalpy of vaporization are T_b and ΔH_v, respectively. The osmotic pressure coefficient and vapor pressure depression coefficient were estimated at 20 °C. The solute is assumed to be involatile. Moreover, it is presumed to be insoluble in solid benzene since these ideal freezing-point depression calculations are based on the assumption that the liquid containing the solute is in equilibrium with pure frozen solvent.

M_1 (g/mol)	Osmotic pressure coefficient (mm benzene cc/g)	Cryoscopic coefficient (K cc/g)	Ebullioscopic coefficient (K cc/g)	Vapor pressure depression coefficient (mm Hg cc/g)
	$\frac{\pi}{c_1} = RT\frac{1}{M_1}$	$\frac{\Delta T_f}{c_1} = \frac{RT_f^2}{\rho \Delta H_f}\frac{1}{M_1}$	$\frac{\Delta T_b}{c_1} = \frac{RT_b^2}{\rho \Delta H_v}\frac{1}{M_1}$	$\frac{\Delta p_d}{c_1} = p_0^* \hat{v}_0 \frac{1}{M_1}$
1 000	280	5.8	3.2	6.7
10 000	28	0.58	0.32	0.67
100 000	2.8	0.058	0.032	0.07
500 000	0.56	0.012	0.006 5	0.013
1 000 000	0.28	0.005 8	0.003 2	0.006 7

which is easily measurable using an apparatus similar to the one presented in Fig. 1. In contrast, the freezing point depression of benzene would be only 0.058 °C, the boiling point elevation of benzene would be even less, 0.032 °C, and the vapor pressure depression would be 0.067 mm Hg, a difference which would be difficult to detect experimentally.

1.2.3 Osmotic Pressure of a Solution Containing Several Solutes If there are N chemically distinct solutes in an ideal solution, the solvent activity in the solution is still given by Eq. (5). However, Eq. (7) must be replaced by

$$\ln x_0 = -\frac{\pi \hat{v}_0}{RT} = \ln\left(1 - \sum_{i=1}^{N} x_i\right) \approx -\sum_{i=1}^{N} x_i, \tag{10}$$

where x_i is the mole fraction of component i. If the solution is dilute, the total number of moles in the solution is approximately equal to the number of moles of solvent and, consequently, Eq. (10) becomes

$$\pi = RT \sum_{i=1}^{N} \frac{c_i}{M_i}, \tag{11}$$

where c_i and M_i are the concentration and molecular weight of solute species i. Using for the total solute concentration the definition

$$c = \sum_{i=1}^{N} c_i, \tag{12}$$

Eq. (11) becomes

$$\frac{\pi}{c} = \frac{RT}{c} \sum_{i=1}^{N} \frac{c_i}{M_i} = \frac{RT}{\bar{M}_n} \tag{13}$$

where $\bar{M}_n$ is the so-called number average molecular weight, which is defined as the mass of solute divided by the number of moles of solute molecules (Billmeyer, 1984). Osmotic pressure and, in fact, all colligative properties provide a measure of the number average molecular weight of solutes (such as polymers). Other relationships describing ideal-solution colligative properties may be generalized to solutions containing species of different molecular weight by replacing the solute concentration c_1 and solute molecular weight M_1 in the equations in Table 2 by the total solute concentration c and number average solute molecular weight $\bar{M}_n$, respectively.

1.2.4 Osmotic Pressure of a Dilute Ideal Electrolyte Solution The solute in many solutions of practical interest is an electrolyte (e.g., salt in saltwater or proteins in cells). The dissociation of the electrolyte must be accounted for in determining the osmotic pressure of the solution, since the number of individual solute molecules will change depending on the degree of dissociation of the electrolyte. This section presents the analytical fundamentals of osmotic pressure of electrolyte solutions.

If an electrolyte undergoes dissociation according to

$$M_{\nu_+}A_{\nu_-} \leftrightarrow \nu_+ M_{z_+} + \nu_- A_{\nu_-}, \tag{14}$$

where n_+ and n_- are the number of cations and anions, respectively, formed by dissociating one molecule of the parent electrolyte $M_{\nu_+}A_{\nu_-}$ into n_+ cations M_{ν_+} of charge z_+ and n_- anions A_{ν_-} of charge z_-, then the mole fraction of solvent in the solution is given by

$$x_0 = \frac{1}{1 + (M_0/1000)(n_+ + n_- + n_u)}. \tag{15}$$

The molality of dissolved cations (i.e., the number of moles of dissolved cations per kilogram of solvent) is n_+, n_- is the dissolved anion molality, and n_u is the molality of dissolved electrolyte that is undissociated. M_0 is the molecular weight of solvent in units of g/mole. Many electrolytes are dissolved in water, and Eq. (15) ignores the influence of hydronium and hydroxyl ions, which are always present in such solutions. If the electrolyte concentration is low (below approximately 0.01 molal) and the pH is markedly acidic or basic (i.e., $p\text{H} < 4$ or $p\text{H} > 10$), then Eq. (15) must be rewritten as follows to include the number of hydroxyl and hydronium ions in the solution (Tanford, 1961):

$$x_0 = \frac{1}{1 + (M_0/1000)(n_+ + n_- + n_u + n_{\text{OH}^-} + n_{\text{H}^+})}. \tag{16}$$

The remainder of this discussion is limited to cases where the concentrations of hydroxyl and hydronium ions are low relative to the concentration of electrolyte components. The following discussion must be modified accordingly if this assumption is not valid.

If n moles of parent electrolyte are added to 1 kg of solvent, the electrolyte concentrations are related by the following equation:

$$n = n_u + n_+/\nu_+ = n_u + n_-/\nu_-. \tag{17}$$

This relationship insures that the solution is electrically neutral (i.e., that the number of positive charges in the solution is exactly equal to the number of negative charges in the solution) (Atkins, 1990). The dissociation of the electrolyte is characterized by an apparent equilibrium dissociation constant, K, which is defined as (Denbigh, 1981)

$$K \equiv \frac{n_+^{\nu_+} n_-^{\nu_-}}{n_u} = \nu_+^{\nu_+} \nu_-^{\nu_-} \frac{\alpha^{\nu_+ + \nu_-}}{1 - \alpha} n^{\nu_+ + \nu_- - 1}, \tag{18}$$

where α is the fraction of parent electrolyte which dissociates ($\alpha \equiv 1 - n_u/n$). Values of K are tabulated for many electrolytes in standard references (e.g., Weast, 1979). As the tendency for the solute to undergo dissociation increases, the K value increases and vice versa. In aqueous solutions, the dissociation constant varies over an enormous range. For example, the dissociation constant of $(HPO_2)^{2-}$ to H^+ and $(PO_4)^{3-}$ is 2.2×10^{-13}, while K for hydriodic acid (HI) is approximately 10^{11} (Atkins, 1990).

The solvent activity of an ideal solution containing an electrolyte is equal to the solvent mole fraction, as before, and the osmotic pressure of an electrolyte solution is obtained by combining Eqs. (10) and (15) to give

$$\begin{aligned}\pi &= \frac{RT}{\hat{v}_0} \ln\left(1 + \frac{M_0}{1000}(n_+ + n_- + n_u)\right) \\ &\approx \frac{RT}{\hat{v}_0}\frac{M_0}{1000}(n_+ + n_- + n_u) \\ &= RT\frac{\rho_0}{1000} n[1 + \alpha(\nu_+ + \nu_- - 1)],\end{aligned} \tag{19}$$

where ρ_0 is the pure solvent density.

1.2.5 Osmotic Pressure of Dilute Ideal Protein Solutions Osmometry is a useful tool for characterizing the molecular weight of proteins such as hemoglobin, bovine serum albunim, and ovalbumin (Tanford, 1961). Proteins are macromolecules containing both weakly acidic and weakly basic structural elements. At low pH, they may carry a positive charge as high as +10 to +20 (measured in units of the charge of a proton) per 10 000 Da [1 Da (dalton) = 1/16 of the mass of an oxygen atom] (Tanford, 1961). As pH increases, weakly acidic groups on the protein chains acquire a negative charge, and the net charge of the protein molecule decreases. At very high pH, a protein molecule can carry a net negative charge of −10 to −20 per 10 000 Da (Tanford, 1961). The pH at which the protein has a net charge of zero is called the isoelectric point.

If protein molecules carrying a net charge of Z per molecule (measured in units of the charge of a proton) are dissolved at a concentration of n_p moles of protein per kilogram of solvent in a solution containing n moles of a fully dissociated electrolyte MA ($n_+ = n_- = 1$) per kilogram of solvent, the concentration of protein molecules will be related to the cation and anion concentration in the solution by (Tanford, 1961)

$$Zn_p = n_- - n_+ + (n_{OH^-} - n_{H^+}). \tag{20}$$

Equation (20) represents the condition for electrical neutrality. In many practical situations, the terms in parenthesis, which represent the effect of water ionization on electrical neutrality, may be neglected since hydroxyl and hydronium ion concentrations are often orders of magnitude lower than cation and anion concentrations. Ionization of water is therefore neglected in the following discussion. The protein molality of the solution would be n_p, the molality of cations would be n, and the molality of anions would be $n + Zn_p$.

The chemical potential of a fully dissociated electrolyte in solution is customarily written as the algebraic sum of the anion and cation chemical potentials. This convention is used since the individual anion and cation chemical potentials cannot be measured separately as a consequence of the requirement that the bulk solution remain electrically neutral (Denbigh, 1981). In a solution containing a polymer electrolyte (such as the protein solution described above), the chemical potential of a polymer molecule would represent, therefore, the chemical potential of the macromolecular ion plus the chemical potential of $|Z|$ small ions of opposite charge (Tanford, 1961). In order to define a chemical potential for macromolecular ions that more nearly reflects the properties of only the macromolecular component, Scatchard has suggested that the polymer electrolyte solution be viewed as comprised of the following three neutral pseudocomponents (Tanford, 1961):

- Component 0, solvent;
- Component 1, n_p moles of protein ions + $Zn_p/2$ moles of anions (A^-) − $Zn_p/2$ moles of cations (M^+);
- Component 2, neutral salt (MA).

Using these definitions, the chemical potential of component 1 will be equal to the algebraic sum of the chemical potential of the protein ions plus $Z/2$ times the difference between the chemical potentials of the anions and cations. For small ions, the difference in chemical potentials between cations and anions is believed to be very small (Tanford, 1961). Therefore, the chemical potential of component 1 in the mixture would be representative of the chemical potential of only the macromolecular ions.

Using Scatchard's convention, the concentrations of macromolecular ions, cations, and anions are given by (Tanford, 1961)

$$\begin{aligned} n_p &= n_1 \\ n_+ &= n_2 - Zn_1/2 \\ n_- &= n_2 + Zn_1/2. \end{aligned} \tag{21}$$

In these expressions, n_2 is the molality of neutral salt MA, and n_p is the polymer molality. The concentrations of anions and cations, expressed as molalities, are given by n_- and n_+, respectively. The polymer is added to the solution as $M_{|Z|}P$ if the polymer is negatively charged or as PA_Z if the polymer is positively charged. The total number of ions per kilogram of solvent is $n_p + 2n_2$. The solution can be formed experimentally by adding n_p moles of polymer and $n_2 - Zn_p/2$ moles of neutral salt to 1 kg of solvent. Note that to use this convention of component definition, the concentrations must be such that the number of moles of neutral salt added, $n_2 - Zn_p/2$, is positive.

The osmotic pressure of such a solution is given by an extension of Eq. (19) (Tanford, 1961):

$$\begin{aligned} \pi &= \frac{RT}{\hat{v}_0} \ln\left(1 + \frac{M_0}{1000}(n_+ + n_- + n_p)\right) \\ &\approx \frac{RT}{\hat{v}_0}\frac{M_0}{1000}(n_+ + n_- + n_p). \end{aligned} \tag{22}$$

1.2.5.1 Donnan Equilibrium. Living cells contain charged protein molecules and salt ions in an aqueous solution called plasma. The cells, which are bathed in a salt-containing, aqueous interstitial fluid, are impermeable to protein but permeable to the small salt ions. The presence of the protein inside the cell results in an unequal partitioning of the small ions between the inside and outside of the cell. This effect, called Donnan equilibrium, plays a central role in cell metabolic functions related to, for example, the propagation of nerve impulses (Woodbury, 1965). It is also of great importance in maintaining an osmotic pressure balance critical to the operation of living cells such as red blood cells, which contain proteins (hemoglobin) and are permeable to water and small ions such as Na^+, K^+, Cl^-, etc. (Tanner, 1959).

To develop a quantitative understanding of the Donnan effect, consider two solutions in contact via a membrane which is permeable to small ions and solvent but not to macromolecules. The solution on side I contains n_p^{I} moles of macromolecular ions, such as protein molecules, and $n^{\mathrm{I}} - Zn_p^{\mathrm{I}}/2$ moles of neutral salt per kilogram of solvent. The solution on side II of the membrane contains n^{II} moles of neutral salt per kilogram of solvent. The small cations and anions (always in equal amounts) can migrate across the membrane. At equilibrium, the chemical potential or, equivalently, activity of the neutral salt must be the same on both sides of the membrane. The activity may be expressed as the product of the constituent ion activities (Denbigh, 1981):

$$a = a_+a_-. \tag{23}$$

The condition of thermodynamic equilibrium requires that

$$a_+^{\mathrm{I}}a_-^{\mathrm{I}} = a_+^{\mathrm{II}}a_-^{\mathrm{II}}. \tag{24}$$

In a dilute solution, ion activity is taken to be equal to ion molality (e.g., $a_+^{\mathrm{I}} = n_+^{\mathrm{I}}$). Thus, the concentrations of small-molecule cations and anions on each side of the membrane are related by (Tanford, 1961)

$$n_+^{\mathrm{I}} n_-^{\mathrm{I}} = n_+^{\mathrm{II}} n_-^{\mathrm{II}}. \tag{25}$$

When combined with the electroneutrality condition expressed by Eq. (20), the concentration of ions may be written as

$$\begin{aligned}(n_+^{\mathrm{II}})^2 &= n_+^{\mathrm{I}}(n_+^{\mathrm{I}} + Zn_p)\\ (n_-^{\mathrm{II}})^2 &= n_-^{\mathrm{I}}(n_-^{\mathrm{I}} - Zn_p).\end{aligned} \tag{26}$$

From Eq. (26), the cation concentration will be higher on side II of the membrane than on side I and the anion concentration will be lower on side II if the macromolecular ions are positively charged (i.e., $Z > 0$) and vice versa if $Z < 0$. The difference in osmotic pressure between the solution containing the macromolecular ions and the solution without macroions is (Tanford, 1961)

$$\Delta\pi \equiv \pi^{\mathrm{II}} - \pi^{\mathrm{I}} = -\frac{RT}{\hat{v}_0}\ln\left[\frac{1 + (M_0/1000)(n_+^{\mathrm{I}} + n_-^{\mathrm{I}} + n_p)}{1 + (M_0/1000)(n_+^{\mathrm{II}} + n_-^{\mathrm{II}})}\right]. \tag{27}$$

1.2.6 Effect of Solution Nonideality on Osmotic Pressure Solvent activity in a nonideal solution is not equal to the mole fraction of solvent as shown in Eq. (5). Rather, the activity is written as the product of a convenient concentration scale (i.e., mole fraction, molality, volume fraction, etc.) and an activity coefficient γ_0, which characterizes the departure of the system from ideality (Prausnitz, 1969).

Many theoretical and empirical models are available for activity coefficients (Reid *et al.*, 1977). For uncharged polymer solutions, a widely used model is due to Flory and Huggins. In this model, the volume fraction of solvent in the solution is the concentration scale. The Flory–Huggins model yields the following expression for the solvent activity (Prausnitz, 1969):

$$a_0 = \phi_0 \exp\left[\left(1 - \frac{1}{r}\right)(1 - \phi_0) + \chi(1 - \phi_0)^2\right], \tag{28}$$

where ϕ_0 is the solvent volume fraction; χ and r are model parameters. The parameter χ characterizes the total intermolecular interaction energy between the molecules in the solution relative to the total intermolecular interaction energy for the pure components (Freeman, 1992). Good solvents have χ values less than 0.5; poor solvents have χ values greater than 0.5. A solvent for which χ is equal to 0.5 is called a theta solvent for the polymer (Billmeyer, 1984). The ratio of the polymer molecular volume to the solvent molecular volume is characterized by r, which is often set equal to the ratio of molar volumes of polymer and solvent. For high-molecular-weight polymers, $1/r$ is often negligibly small (Prausnitz, 1969).

Using the Flory–Huggins model, the following expression may be obtained for the osmotic pressure of a dilute, high-molecular-weight, polydisperse polymer solution (Young and Lovell, 1991):

$$\frac{\pi}{c} = \frac{RT}{\bar{M}_n}\left[1 + \frac{\bar{M}_n}{\hat{V}_0\rho_p^2}(\tfrac{1}{2} - \chi)c\right], \tag{29}$$

where ρ_p is the polymer density and, as in Eq. (12), c is the total polymer concentration. As indicated by Eq. (29), the ideal solution expression for the osmotic pressure [i.e., Eq. (13)] is recovered in the limit as polymer concentration goes to zero or as the Flory–Huggins χ parameter goes to $\frac{1}{2}$.

To characterize deviations from ideality that are not accurately described by the Flory–Huggins theory, Eq. (29) is often written as a power series (or virial series) in concentration (Tanford, 1961):

$$\frac{\pi}{c} = \frac{RT}{\bar{M}_n}[1 + Bc + Cc^2 + \ldots], \tag{30}$$

where the coefficients B and C are the second and third virial coefficients, respectively. The second virial coefficient can be written in term of the Flory–Huggins interaction parameter as (Young and Lovell, 1991)

$$B = \bar{M}_n(\tfrac{1}{2} - \chi)/(\hat{V}_0\rho_p^2). \tag{31}$$

Experimental osmotic pressure data for the polymer nitrocellulose in several solvents are presented in Fig. 2. The data presented in Fig. 2 are well represented by truncating Eq. (30) after the term containing the second virial coefficient. In thermodynamically good solvents, such as acetone, the second virial coef-

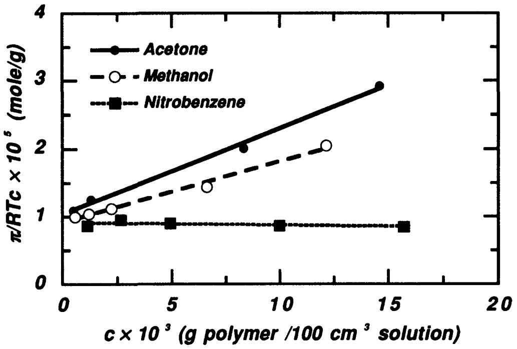

FIG. 2. Influence of solvent on results of osmotic pressure experiment to determine number average molecular weight of nitrocellulose.

ficient is large and positive, and the ratio of osmotic pressure to polymer concentration increases with increasing polymer concentration. The osmotic pressure in this nonideal solution depends, therefore, on both the number of solute molecules and the chemical nature of the solute (or, more precisely, on the interactions between the solute and solvent). In a thermodynamically poorer solvent, such as methanol, the sensitivity of π/c to polymer concentration is reduced. In a very poor solvent, such as nitrobenzene, the ratio π/c is almost insensitive to polymer concentration, indicating a χ value of approximately 0.5 or, equivalently, a vanishing second virial coefficient. When π/c is independent of polymer concentration, the solution is ideal, and the osmotic pressure depends only on the number of solute molecules, not on their chemical nature.

In systems containing electrolytes, activity coefficient models are available in the literature to account for nonideal solution behavior (Atkins, 1990). As described in Eq. (23), the activity for an electrolyte solution is typically written as the product of the constituent ion activities as (Tanford, 1961)

$$a = a_+ a_- = n_+ n_- \gamma_\pm^2, \tag{32}$$

where $\gamma_\pm$ is the mean activity coefficient for the cations and anions. A widely used model for the activity coefficient is given by the extended Debye–Hückel law (Atkins, 1990):

$$\log(\gamma_\pm) = -\frac{A|z_+ z_-|}{1 + BI^{1/2}} I^{1/2}, \tag{33}$$

where A and B are constants, and I is the ionic strength of the solution. The value of A is 0.509 $(\text{kg/mol})^{1/2}$ for an aqueous solution at 25 °C, and, in general, A is a function of the temperature and dielectric constant of the solution (Tanford, 1961). B is typically treated as an adjustable constant. The ionic strength of the solution is given by

$$I \equiv \tfrac{1}{2} \sum_i z_i^2 n_i, \tag{34}$$

where the summation extends over all different ions in the solution, and z_i and n_i are the charge and molality, respectively, of ion i (Tanford, 1961). This model can account for activity coefficients at concentrations as high as 0.1 mol/kg. There is no available theory for solutions which are substantially more concentrated (Atkins, 1990).

An example of osmosis in biological systems is presented in Fig. 3, which illustrates the dependence of hemoglobin osmotic pressure on hemoglobin concentration in water. As shown by the dashed line in Fig. 3, the osmotic pressure is not a linear function of hemoglobin concentration above approximately 7 g/(100 cm³), indicating that hemoglobin solutions are nonideal above this concentration.

As a practical application of the material presented in Sec. 1.2.5, it is interesting to note that at the concentration of hemoglobin in the blood [33 g/(100 cm³)], the osmotic pressure inside red blood cells due to hemoglobin is approximately 0.5 bar (380 mm Hg) (Lightfoot, 1974). The osmotic pressure of the serum proteins in contact with red blood cells is, on

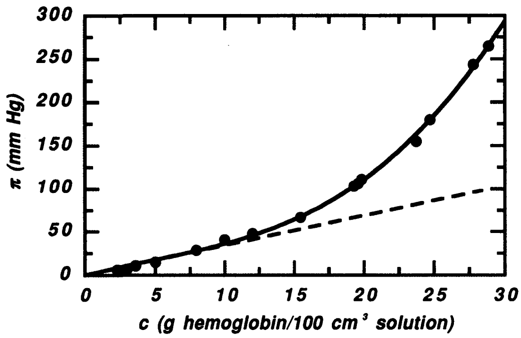

FIG. 3. Osmotic pressure of a hemoglobin solution at 0 °C. The dashed line indicates ideal solution behavior. (Lightfoot, 1974.)

average, 25 mm Hg (Lightfoot, 1974). The cell wall cannot support this large difference in osmotic pressure arising from the substantial difference in intracellular and intercellular protein concentration. Therefore, an opposing imbalance in electrolyte concentration is maintained which compensates for the unequal protein concentrations inside and outside the cell wall (Lightfoot, 1974).

2. EXPERIMENTAL METHODOLOGY FOR THE DETERMINATION OF OSMOTIC PRESSURE

Osmotic pressure may be determined directly using two broad classes of devices: static osmometers and dynamic osmometers (Young and Lovell, 1991). In both types of osmometers, the osmotic pressure is determined as a function of concentration [solute concentration is commonly less than 20 g/L (Young and Lovell, 1991)] at fixed temperature. The extrapolation of the ratio of π/c to zero concentration gives $RT/\bar{M}_n$, and the solute number average molecular weight may be computed directly.

In a static osmometer, the osmotic pressure is determined by measuring the equilibrium pressure difference across a membrane separating two solutions containing different concentrations of solute. A schematic of a static osmometer is presented in Fig. 4. Older static osmometers (e.g., Pinner–Stabin osmometers) have large cell volumes (3–20 cm^3) on either side of the membrane (Young and Lovell, 1991). Small-bore capillaries are used for the tubes above the cell. The use of small capillaries decreases the amount of solvent which must permeate through the membrane before the attainment of equilibrium. Normally, the amount of solvent added to the solution during the equilibration time does not significantly change the solute concentration (Young and Lovell, 1991). Modern static osmometers (e.g., Knauer and Wescan osmometers) have small cell volumes (on the order of 0.3–2 cm^3) and large membrane areas to minimize equilibration times (typically 30–60 minutes) (Young and Lovell, 1991). Pressure differences between the two sides of the membrane are typically monitored indirectly using pressure transducers (Billmeyer, 1984). Osmotic pressure is sensitive to temperature, and so temperature is typically controlled to ±0.01 °C (Young and Lovell, 1991).

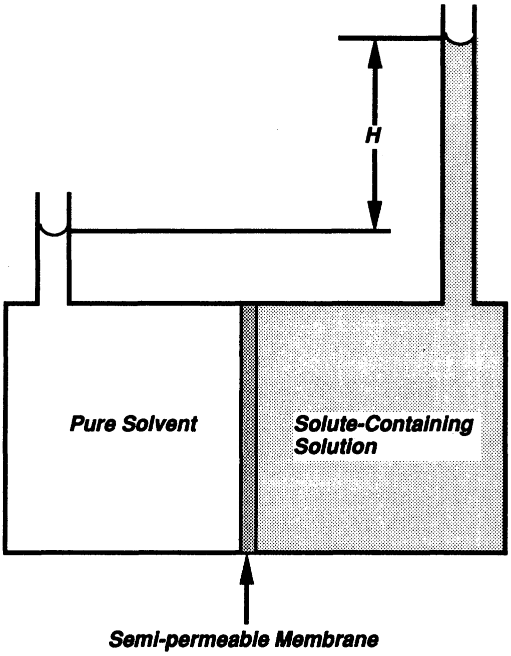

FIG. 4. Schematic of osmometer.

In a dynamic osmometer, the flow of solvent across the membrane is monitored, and a hydrostatic pressure just sufficient to prevent solvent flow through the membrane is applied to the solution (Young and Lovell, 1991). This hydrostatic pressure is the osmotic pressure of the solution.

Number average molecular weights between approximately 5×10^3 and 5×10^5 Da can often be determined to an accuracy of approximately 1% (Rodriguez, 1989). Real membrane materials are not completely impermeable to solutes. Most osmometer membrane materials are based on cellulose, and at molecular weights below 5×10^3 Da, many solutes can permeate through the membrane and will not, therefore, contribute to the observed osmotic pressure of the solution (Rodriguez, 1989; Young and Lovell, 1991). For this reason, membrane osmometry can seriously overestimate the number average molecular weight of samples containing a large number fraction of low–molecular-weight species. Above a molecular weight of approximately 5×10^5 Da, the osmotic pressure of the solution is very small, and the precision of the technique is limited, therefore, by the precision with which small pressure differences (between the two chambers connected by the membrane) can be determined (Billmeyer, 1984).

3. APPLICATIONS

3.1. Reverse Osmosis

The role of osmosis in industrial separation technologies such as reverse osmosis, dialysis, and electrodialysis are described in several review articles (Bhattacharyya and Williams, 1992; Eykamp and Steen, 1987; Kessler and Klein, 1992; Klein *et al.*, 1987; Kulkarni *et al.*; 1992, Strathmann, 1992). In reverse osmosis (also called hyperfiltration), two solutions of different solute concentrations are separated by a semipermeable membrane. The objective of the separation is to drive solvent (e.g., water) from a region of high solute concentration (e.g., seawater) across the membrane to a region of low solute concentration (e.g., fresh water). In order to transport solvent from the high–solute-concentration side to the low–solute-concentration side of the membrane, the solvent chemical potential on the high-concentration side must be raised to a value that is above the solvent chemical potential on the low–solute-concentration side. A hydrostatic pressure applied to the high–solute-concentration side of the membrane is typically used to increase the solvent chemical potential on the high–solute-concentration side of the membrane and drive solvent transport.

At steady state, the flux of solvent across the membrane is given by (Bhattacharyya and Williams, 1992; Kessler and Klein, 1992)

$$J_0 = A(\Delta p - \sigma\Delta\pi), \tag{35}$$

where σ is the Staverman reflection coefficient (defined below), J_0 is the solvent flux (volume of solvent permeating membrane per unit membrane area per unit time), A is the solvent permeability constant, Δp is the impressed hydrostatic pressure difference across the membrane, and $\Delta\pi$ is the osmotic pressure difference across the membrane (Bhattacharyya and Williams, 1992). Theoretical models are available to predict the dependence of the solvent permeability constant on material parameters (such as solvent diffusion coefficient and solubility in the membrane) and membrane thickness (Bhattacharyya and Williams, 1992).

At a solvent flux equal to zero, the Staverman reflection coefficient σ is the ratio of the applied pressure difference Δp to the osmotic pressure difference $\Delta\pi$ across the membrane. If σ is equal to 1, then the hydrostatic pressure difference that must be maintained across the membrane to prevent solvent flow due to osmosis is equal to the osmotic pressure difference between the two solutions. In this limit, the membrane is completely impermeable to solute. If the membrane is permeable to solute, the pressure difference needed to maintain zero solvent flux is less than the osmotic pressure difference predicted based on bulk solute concentrations and σ is, accordingly, less than 1 (Bhattacharyya and Williams, 1992).

The importance of osmosis in membrane separations such as reverse osmosis can be understood from Eq. (35). The osmotic pressure difference $\Delta\pi$ must be overcome by an applied hydrostatic pressure difference across a membrane to transport solvent through the membrane and, for example, desalinate water. If the osmotic pressure difference across the membrane is large, the applied hydrostatic pressure difference will be large which, in turn, increases the cost of the technique. From Table 1, the osmotic pressure of seawater is approximately 20 times that of brackish water. Consequently, reverse osmosis, which is very competitive economically with other technologies (such as multistage flash distillation, multiple effect evaporation, ion exchange, and electrodialysis) for brackish water desalination, is much less competitive for seawater desalination (Riley, 1990).

The desalination of water is the most widely known application of reverse osmosis. However, in addition to desalination of water, reverse osmosis is applied in the food industry to concentrate products such as milk, in the paper industry for the concentration of lignosulfonates, and in the chemical process industries to concentrate and recover a variety of industrial solutes (Eykamp and Steen, 1987).

3.2. Dialysis

Dialysis (*q.v.*) is a separation process in which a feed solution is passed across a semipermeable membrane that is permeable to solvent and small solutes in the feed but impermeable to large macromolecules in the feed (Klein *et al.*, 1987). Unlike reverse osmosis, dialysis membranes are designed to encourage the passage of small solute molecules while retaining only the species of highest molecular weight in the mixture. An important application of dialysis is hemodialysis. In this process, blood containing small–molecular-weight metabolic waste products is passed across a membrane that is permeable to the low–molecular-weight wastes but impermeable to higher–molecular-weight blood proteins (Klein *et al.*, 1987). This application has sparked the development of dialysis for use in industrial separations such as the desalting of cheese whey solids and the reduction of alcohol content in beer (Klein *et al.*, 1987). Dialysis based upon the Donnan effect described earlier is being tested in applications such as the softening of feedwater to reverse-osmosis membrane units and the removal of toxic or otherwise harmful cations from industrial wastewater streams (Klein *et al.*, 1987).

3.3. Electrodialysis

Electrodialysis involves the use of electrically charged membranes. An electric field is applied across the membrane to separate ionic solutes from an aqueous feed stream (Strathmann, 1992). The membranes used in such devices are ion selective (that is, they allow positively charged ions to permeate while excluding negatively charged ions or vice versa). The principal commercial applications of electrodialysis are desalination of water, the production of table salt (Strathmann, 1992), desalination of food such as cheese whey and sugar, and the production of chlorine in so-called "chlor-alkali cells" (Lonsdale, 1982).

GLOSSARY

Chemical Potential: An intensive thermodynamic function that is defined, mathematically, as the derivative of the extensive thermodynamic functions of internal energy, enthalpy, Gibbs free energy, and Hemholtz free energy with respect to the amount of one component in a mixture. More importantly for osmosis, the condition of chemical equilibrium between two or more phases is established when each substance has an equal value of its chemical potential in all phases into which this substance may partition. Difference in the chemical potential of a substance that may partition between two phases is the origin or driving force for the diffusion of that substance from a region of high chemical potential to a region of lower chemical potential.

Dalton (Da): A standard unit of molecular weight. One dalton is 1/16 the mass of an oxygen atom.

Dialysis: A separation process in which a feed solution is passed across a semipermeable membrane that is permeable to solvent and small solutes in the feed but impermeable to large macromolecules in the feed. Small solute molecules are separated from the feed solution in this process.

Electrodialysis: A separation process involving the use of electrically charged membranes and the application of an electric field across the membrane to separate ionic solutes from an aqueous feed stream.

Isotonic Solution: A solution that has the same osmotic pressure as a reference solution. The reference solution in biomedical applications is typically mammalian plasma.

Molality: The molal concentration of a solute. A one molal solution contains one mole of solute molecules per kilogram of solvent.

Osmotic Pressure: Operationally, the excess pressure that must be placed on a solution separated by a semipermeable membrane from a reservoir of pure solvent to prevent any net mass transport of solvent through the membrane. It is a thermodynamic property of the solution and is a measure of the depression in the solvent chemical potential due to the presence of solute molecules.

Osmometry: The analytical technique for measuring the osmotic pressure of a solution.

Reverse Osmosis: (Also called hyperfiltration) a separation process in which a solution at high pressure is passed across a semipermeable membrane that is permeable to solvent but impermeable to solute. The result is a concentration of the solute in the feed stream to the unit and the production of a stream of pure solvent that has permeated through the membrane.

Semipermeable Membrane: A membrane that is permeable to solvent molecules but impermeable to solute molecules.

Works Cited

Atkins, P. W. (1990), *Physical Chemistry*, San Francisco: W. H. Freeman.

Bhattacharyya, D., Williams, M. E. (1992), in: W. S. W. Ho, K. K. Sirkar (Eds.), *Membrane Handbook*, New York: Van Nostrand Reinhold, p. 263.

Billmeyer, F. W., Jr. (1984), *Textbook of Polymer Science*, New York: Wiley.

Denbigh, K. (1981), *The Principles of Chemical Equilibrium, 4th ed.*, Cambridge, U.K.: Cambridge Univ. Press.

Eykamp, W., Steen, J. (1987), in: R. W. Rousseau (Ed.), *Handbook of Separation Process Technology*, New York: Wiley, p. 826.

Freeman, B. D. (1992), in: S. L. Aggarwal, S. Russo (Eds.), *Comprehensive Polymer Science*, Oxford: Pergamon, p. 167.

Kessler, S. B., Klein, E. (1992), in: W. S. W. Ho, K. K. Sirkar (Eds.), *Membrane Handbook*, New York: Van Nostrand Reinhold, p. 161.

Klein, E., Ward, R. A., Lacey, R. E. (1987), in: R. W. Rousseau (Ed.), *Handbook of Separation Process Technology*, New York: Wiley, p. 954.

Kulkarni, S. S., Funk, E. W., Li, N. N. (1992), in W. S. W. Ho, K. K. Sirkar (Eds.), *Membrane Handbook*, New York: Van Nostrand Reinhold, p. 392.

Lightfoot, E. N. (1974), *Transport Phenomena with Living Systems*, New York: Wiley.

Lonsdale, H. K. (1982), "The Growth of Membrane Technology," *J. Membrane Sci.* **10**, 81–181.

Marick, L. (1969), in: W. H. Nault (Ed.), *The World Book Encyclopedia*, Chicago: Field Enterprises Education, p. 657.

Mason, E. A. (1991), "From pig bladders and cracked jars to polysulfones: an historical perspective on membrane transport," *J. Membrane Sci.* **60**, 125–145.

Morris, E. (1981), *The American Heritage Dictionary of the English Language*, Boston: Houghton Mifflin.

Prausnitz, J. M. (1969), *Molecular Thermodynamics of Fluid-Phase Equilibria*, Englewood Cliffs, NJ: Prentice-Hall.

Reid, R. C., Prausnitz, J. M., Sherwood, T. K. (1977), *The Properties of Gases and Liquids*, New York: McGraw-Hill.

Riley, R. L. (1990), in: R. W. Baker, W. Eykamp, R. L. Riley (Ed.), *Membrane Separations Systems: A Research Needs Assessment*, Washington, DC: Department of Energy, Chap. 5.

Rodriguez, F. (1989), *Principles of Polymer Systems*, New York: Hemisphere Publishing.

Strathmann, H. (1992), in: W. S. W. Ho, K. K. Sirkar (Eds.), *Membrane Handbook*, New York: Van Nostrand Reinhold, p. 217.

Tanford, C. (1961), *Physical Chemistry of Macromolecules*, New York: Wiley.

Tanner, R. (1959), in: J. L. Oncley (Ed.), *Biophysical Science—a Study Program*, New York: Wiley, p. 342.

Weast, R. C. (Ed.) (1979), *CRC Handbook of Chemistry and Physics*, Boca Raton, FL: CRC Press.

Woodbury, J. W. (1965), in: T. C. Ruch, H. D. Patton (Eds.), *Physiology and Biophysics*, Philadelphia: W. B. Saunders, p. 26.

Young, R. J., Lovell, P. A. (1991), *Introduction to Polymers*, London: Chapman & Hall.

Further Reading

Bhattacharyya, D., Williams, M. E. (1992), in: W. S. W. Ho, K. K. Sirkar (Eds.), *Membrane Handbook*, New York: Van Nostrand Reinhold, p. 263.

Billmeyer, F. W., Jr. (1984), *Textbook of Polymer Science*, New York: Wiley.

Denbigh, K. (1981), *The Principles of Chemical Equilibrium, 4th ed.*, Cambridge, U.K.: Cambridge Univ. Press.

Eykamp, W., Steen, J. (1987), in: R. W. Rousseau (Ed.), *Handbook of Separation Process Technology*, New York: Wiley, p. 826.

Kessler, S. B., Klein, E. (1992), in: W. S. W. Ho, K. K. Sirkar (Eds.), *Membrane Handbook*, New York: Van Nostrand Reinhold, p. 161.

Klein, E., Ward, R. A., Lacey, R. E. (1987), in: R. W. Rousseau (Ed.), *Handbook of Separation Process Technology*, New York: Wiley, p. 954.

Kulkarni, S. S., Funk, E. W., Li, N. N. (1992), in: W. S. W. Ho, K. K. Sirkar (Eds.), *Membrane Handbook*, New York: Van Nostrand Reinhold, p. 392.

Tanford, C. (1961), *Physical Chemistry of Macromolecules*, New York: Wiley.

Young, R. J., Lovell, P. A. (1991), *Introduction to Polymers*, London: Chapman & Hall.

PACKAGING TECHNOLOGY

John J. Ritsko, *IBM T.J. Watson Research Center, Yorktown Heights, New York, U.S.A.*

	Introduction	73
1.	**General Principles**	74
1.1	Interconnection Hierarchy	74
1.2	Electrical Design	74
1.3	Materials Characteristics	75
2.	**Chip-to-Package Interconnections**	77
2.1	Wirebonding	77
2.2	Tape Automated Bonding (TAB)	77
2.3	Solder-Bump Interconnections	77
3.	**Molded Plastic Packaging**	78
3.1	Lead-Frame Materials	78
3.2	Encapsulation Materials	79
4.	**Printed-Circuit Cards and Boards**	79
4.1	Structures	79
4.2	Materials and Processes	79
5.	**Ceramic Packaging**	80
5.1	Metal-on-Ceramic Packages	82
5.2	Multilayer Ceramic Substrates	82
5.3	Ceramic Materials	82
6.	**Thin-Film Wiring**	83
6.1	Materials	83
6.2	Sequential Patterning	83
7.	**Package-to-Package Interconnections**	84
7.1	Pins	84
7.2	Surface-Mount Technology	84
7.3	Connectors	85
7.4	Cables	85
8.	**Heat Transfer**	85
8.1	Thermal Engineering	86
8.2	Cooling Structures	86
9.	**Future Challenges, Emerging Technologies**	86
9.1	Optoelectronics	86
9.2	Electromagnetic Management	88
9.3	Low-Power, Low-Cost, Light-Weight Systems	88
	Works Cited	88

INTRODUCTION

The term "packaging" is used to describe the methods and structures by which electronic components are interconnected to make a functional system. The interconnection package must perform several key roles. It must provide mechanical support for the components as well as protection from damage due to handling. Sensitive active components such as transistors and integrated circuits must be protected from the external environment. The package must distribute power to active devices and remove the heat generated by them as well as providing for all required signal interconnections between components. Packaging technology is the set of designs, materials, and processes of which the interconnection packages are made.

In the earliest electronic systems, individual components, such as resistors, capacitors, tubes, and laser semiconductor devices, were interconnected by soldering their terminals or plugging them into sockets with connections made by discrete wires. As the complexity of electronic systems increased, denser and denser interconnection packages were required to reduce cost and save space as well as to increase the overall system performance. Printed circuit cards were introduced to interconnect discrete components efficiently, and integrated circuits allowed many active devices to be placed and interconnected on a single silicon chip. Nevertheless, the increasing complexity of electronic systems required ever more sophisticated packaging technology. Today, powerful computers utilize hundreds of thousands or even millions of circuits, and many systems are designed in such a way that their performance depends as much on packaging technology as on chip technology.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Modern packaging technology is a complex field concerning the fabrication of many different mechanical, electrical, and thermal structures using many different metals, polymers, glasses, and ceramics, which are formed in a wide variety of processes and which are optimized for both cost and performance. This article describes the principles and materials considerations of importance to packaging. It explains how integrated circuit chips are connected to the interconnection package and the various types of simple and complex packages in use today. Since only a general description can be provided here and the details are very important in the engineering of different packaging technologies, the reader is encouraged to refer to several recent books on packaging cited as references.

1. GENERAL PRINCIPLES

1.1 Interconnection Hierarchy

Complex systems consist of many levels of interconnections, as shown for a typical system in Fig. 1. The first-level package encapsulates the chip for protection from handling damage and the environment and provides metallic leads for power input as well as signal connections between chips. The simplest and lowest-cost package is a plastic or ceramic single-chip module (SCM). The chip-to-chip interconnections are then provided by the second level of packaging or multichip module (MCM), which may be a laminated plastic-based card or board (MCM-L) or a ceramic-based substrate (MCM-C), or a hybrid containing sequentially built thin-film wiring (MCM-D). For the highest-performance electrical interconnections between chips, bare chips are mounted on a multichip module that is then encapsulated as a whole, thus combining the first and second packaging levels.

1.2 Electrical Design

Signal interconnection between chip and package and between different packaging levels requires both making the number of required interconnections and providing the optimal electrical performance. The number of terminals required by a component is re-

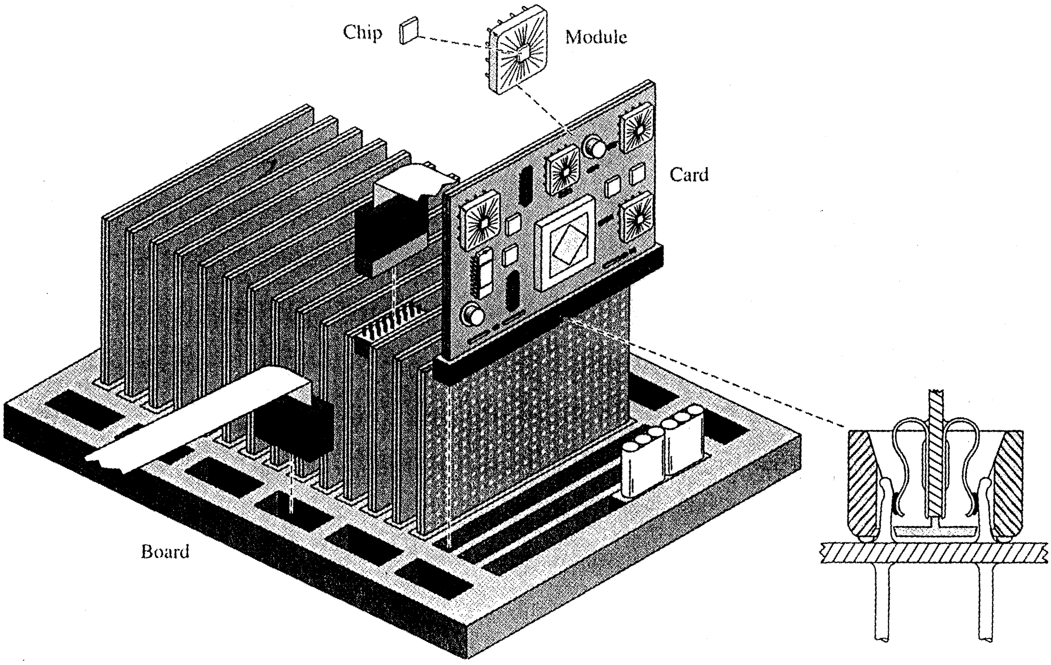

FIG. 1. Chips are mounted in single- or multiple-chip modules. The modules are soldered or pin attached to cards, which are plugged into connectors in a board. Cables connect signals between boards. (From Tummala and Rymaszewski, 1989, reproduced by permission.)

lated to the communication needed between this component and the rest of the system. A useful design guideline was given in an empirical rule devised by E. Rent in 1960 (Radke, 1969):

$$I = bC^p,$$

where I is the number of input and output terminals (I/O), C is the number of circuits on a package, b is a constant equal to approximately 2, and p is approximately $\frac{2}{3}$. In recent years, the degree of functionality contained in a package component has increased, requiring fewer interconnections to the next level than predicted by the original Rent's rule, but the above formula with lower values of p is still useful (Seraphim *et al.*, 1989). Bipolar chips containing 5000–20 000 circuits require 200–400 signal I/O, and large multichip modules containing millions of circuits require thousands of interconnections. The larger the number of I/O terminals on a given package the greater will be the amount of interconnection wiring required on the next level of packaging.

The delay caused by signal propagation between chips often limits overall system performance. The resistance R, capacitance C, and inductance L of the package wiring are key to improved performance. Proper electrical design minimizes both signal delay and noise. For short nets the RC time constant determines the package delay. Over longer distances signals travel in transmission lines in cards, boards, ceramic modules, or thin-film wiring with typical dimensions shown in Fig. 2. These signals travel with the velocity $V = c/\sqrt{\epsilon}$, where c is the speed of light and ϵ is the insulator dielectric constant. The time delay T in traveling a distance l in the package is then

$$T = (l/c)\sqrt{\epsilon}.$$

This delay is minimized by shortening the distance signals must travel by using finer, denser wiring and by reducing the dielectric constant of the insulator. The impedance of the transmission line, Z_0, is $\sqrt{L/C}$ where L is the inductance and C the capacitance per unit length. Z_0 is typically 40–80 Ω. In addition, the transmission-line designs shown in Fig. 2 contain parallel planes of wires, and the dimensions are chosen to reduce cross-coupling noise between lines to typically <10% of the signal amplitude by minimizing mutual inductance and interline capacitance.

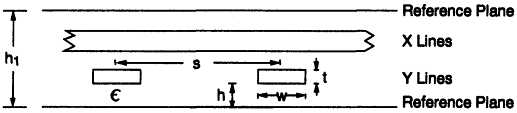

	Card/board	Ceramic module	Thin-film wiring
W	150–100	80	10–25
t	30–50	25	3–15
h	270–430	100	5–15
h_1	500–1200	300	20–36
s	500–400	400	25–75
ϵ	4.4–4.0	5–10	3–3.5
Z_0 (Ω)	80	50	40–50

FIG. 2. Cross section of transmission lines with typical dimensions (in micrometers). The term ϵ is the insulator dielectric constant and Z_0 the characteristic impedance.

Power must be distributed to the chips on an MCM by low-resistance power planes to minimize voltage drops across the package. When many drivers, N, are switched simultaneously on the chip, a noise voltage ΔV is induced on the chip reference voltage, related to the rate of driver current switching, di/dt, by the effective package inductance L_{eff}:

$$\Delta V = NL_{\text{eff}}\frac{di}{dt}.$$

The package must provide low enough L_{eff} to keep ΔV below typically 10–20% of the supply voltage. The low inductance is achieved by making the connection between the chip and power distribution planes as short as possible.

1.3 Materials Characteristics

Package structures are made from a wide variety of materials (Harper, 1991; Minges, 1989). The characteristics of some typical conductors and insulators are given in Tables 1 and 2. The most important properties are the electrical conductivity of metals, the dielectric constant of insulators, the thermal conductivity of the heat dissipation structures, and the relative thermal expansion coefficient of dissimilar materials that are bonded or joined together.

Table 1. Properties of insulator materials.

Insulator	Dielectric constant	Thermal expansion coefficient (10^{-7}/°C)	Thermal conductivity W/m°K
SiC	42	37	270
Si	11.7	27	170
Alumina	9.4	66	20
AlN	8.8	33	230
BeO	6.8	68	240
BN	6.5	37	600
Mullite	6.2	45	10
Diamond (CVD)	5.7	23	1500
Glass-ceramics	4–8	30–50	5
Epoxy-glass (FR-4)	4.7	110–150	0.2
Borosilicate glass	4.0	30	5
Epoxy-Kevlar (60%)	3.6	60	0.2
Polyimides	3–3.5	70–500	0.2
Benzocyclobutene	2.6	350–600	0.2
Polytetrafluoroethylene	2.2	200	0.1

Since the resistance of a conducting element increases with length but is inversely proportional to cross-sectional area, different types of interconnection make trade-offs between dimensions and resistivity to optimize performance. Copper is the conductor of choice for all interchip wiring because of its low resistivity and its superior chemical stability to silver; however, because of relative ease of fabrication higher-resistivity metals are often used, particularly in ceramic packages. Joining metals such as solders or brazes have high resistivity but are used for short, wide-area connections. Kovar is used for pins because of its mechanical properties, while the relatively large size of pins allows the use of a high-resistivity metal.

Since the performance of the package improves with lower dielectric constant and shorter wiring lengths, new materials or processes have been introduced to improve performance. Low-ϵ glass-ceramics have been developed to replace the more common alumina ceramic. Thin-film copper wiring (which produces the highest wiring density) is made with polymer insulators.

The thermal conductivity of plastics is relatively poor and heat is conducted away from high-power chips by ceramic or metal heat spreaders to air- or water-cooled heat sinks.

Electronic packages are comprised of dissimilar materials bonded or joined together, which expand at different rates on heating. The thermal cycles that occur during fabri-

Table 2. Properties of conductor materials.

Conductor	Electrical resistivity (10^{-6} Ωcm)	Thermal expansion coefficient (10^{-7}/°C)	Thermal conductivity W/m K
Alloy 42 (42% Ni, 58% Fe)	70	43	16
Kovar	50	53	17
Invar	46	15	11
Silver–palladium	20	140	150
Chromium	20	63	66
Pb 95%, Sn 5%	19	290	63
Pb 40%, Sn 60%	17	245	51
Au 80%, Sn 20%	16	159	57
Palladium	10.8	110	70
Platinum	10.6	90	71
Nickel	6.8	133	92
Molybdenum	5.2	50	146
Tungsten	5.5	45	200
Aluminum	4.3	230	240
Gold	2.2	142	297
Copper	1.7	170	393
Silver	1.6	197	418

cation or when systems are turned on and off result in stresses and/or strains which are proportional to the difference in thermal expansion coefficients of the different materials and the change in temperature. These stresses can result in fatigue failure such as delamination at interfaces of a multilayer structure, solder cracking, and metal fatigue. The design of a package for maximum reliability requires the choice of materials with as closely matched thermal expansion coefficients as possible given the fabrication and electrical design requirements.

2. CHIP-TO-PACKAGE INTERCONNECTIONS

2.1 Wirebonding

Wirebonding is the most common chip connection technology. Wires are directly and individually bonded to the input/output pads on the perimeter of the chip and then bonded to the package. The bonding process is ultrasonic, thermocompression, or thermosonic bonding applied by mechanical tools. The wires are aluminum (with 1% Si) or aluminum alloys or gold and are 25–38 μm in diameter for signal and as much as 125–150 μm in diameter for power connections if required. Chip pads can be placed on a 100–150-μm pitch, thus allowing several hundred I/O from large chips. Wirebonding is most commonly used with molded plastic SCMs (described below).

2.2 Tape Automated Bonding (TAB)

Tape automated bonding or TAB was developed as a highly automated technique for high-volume packaging of individual chips with relatively few connections, although it has evolved to accommodate over 300 I/O terminals. The TAB process begins with a continuous tape up to 70 mm or more wide as shown in Fig. 3(a). The tape is processed in a reel-to-reel method and contains sprocket holes for movement and alignment as shown in the figure. The tape is made of stable plastics such as polyimides, polyesters, etc., and contains copper conductors patterned using photolithographic techniques and metal etching processes. In order to isolate electrically the metal patterns associated with a given chip, mechanical punching is used. The plastic tape is either etched or prepunched to form a window in which the metal leads extend beyond the tape so that they can be bonded to chips containing metal bumps on their I/O pads. The bonding of all leads is done simultaneously in a thermocompression bonding process in an inner-lead bonding tool shown in Fig. 3(b). In this tool the chip is placed with its aluminum or gold bumps facing up. The TAB tape is placed over the chip with the inner lead bonds aligned to the metal bumps. A heated metal tip or thermode then uses compression to form the bonds between leads and bumps. The chips are retained on the tape for subsequent testing, encapsulation (if desired), and burn-in followed by excising of the individual chip-lead packages from the tape. TAB is a chip connection technology that can be used as part of a molded plastic- or ceramic-based single-chip module, but it can also function as a first-level package if the chip is encapsulated in plastic on the tape and the outer TAB leads are bonded directly to the second-level package.

2.3 Solder-Bump Interconnections

Solder-bump chip connections, also known as controlled-collapse chip connection (C4) or flip-chip bonding, involves solder bumps deposited by metal evaporation in a dense area array on wettable metal terminals on the chip and a matching footprint of solder-wettable terminals on the interconnection package. The upside down chip is aligned to the substrate and all joints are made simultaneously by reflowing the solder. The high surface tension of the solder causes self-alignment. For a given chip size this technology allows the greatest number of interconnections. The small bump size results in very low inductance. In a typical application, 100-μm-diam bumps are placed on a 225-μm grid allowing space for 784 connections on a 6.5-mm square chip. An example of this technology is shown in Fig. 4. Smaller solder bumps and tighter grids resulting in thousands of connections have been demonstrated (Tummala and Rymaszewski, 1989).

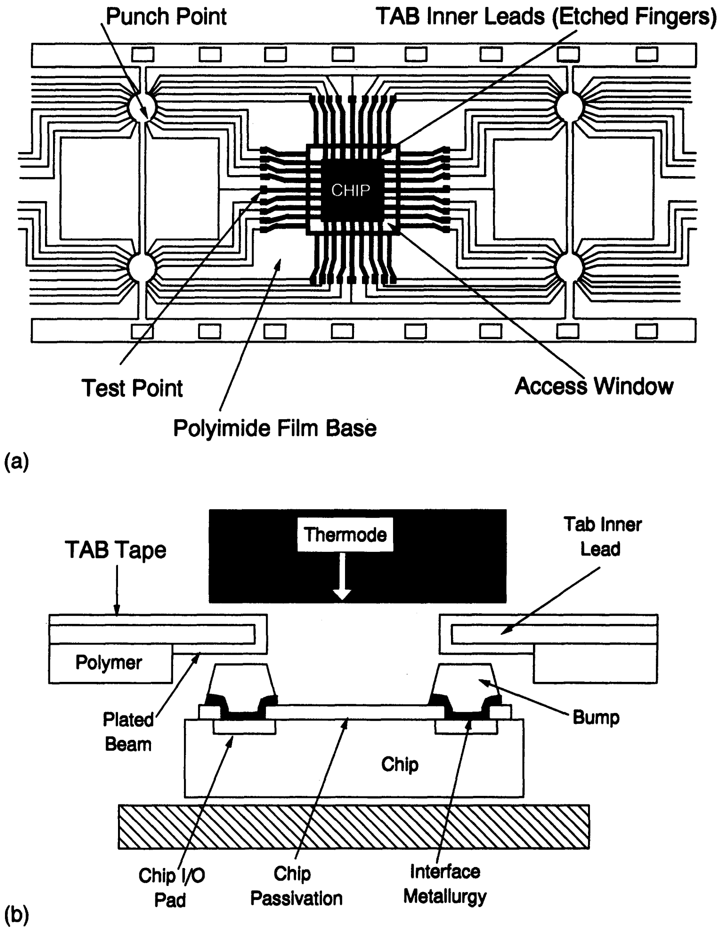
Punch Point
TAB Inner Leads (Etched Fingers)
CHIP
Test Point
Access Window
Polyimide Film Base
(a)
TAB Tape
Thermode
Tab Inner Lead
Polymer
Plated Beam
Bump
Chip
Chip I/O Pad
Chip Passivation
Interface Metallurgy
(b)

FIG. 3. (a) Tape automated bonding (TAB) assembly tape (adapted from Sinnadurai, 1985). (b) Inner lead bonding process (adapted from Tummala and Rymaszewski, 1989).

3. MOLDED PLASTIC PACKAGING

Transistors encapsulated in plastic appeared in the late 1950s, and because of the low cost and effectiveness of the technology, molded plastic packages are the most widely used single-chip modules today. They are called by many names such as dual in-line package (DIP), and variations like SIP and ZIP, the small outline package (SOP and SOJ), the plastic leaded chip carrier (PLCC), and plastic quad flat pack (PQFP), as well as others. Some of these are illustrated in Figure 5 where a cutaway view shows their basic construction. The lead frame is the supporting structure to which the chip is physically attached. Electrical connections are made by wirebonding and then the whole assembly is covered with a thermosetting molding compound (Goosey, 1985) in a transfer molding process.

3.1 Lead-Frame Materials

The most widely used metal for lead frames is Alloy 42 (42% Ni, 58% Fe), which has a thermal expansion coefficient closely matched to the Si chip, although its low thermal conductivity presents a problem for high-power chips since the lead frame is the main heat conduit away from the chip. To overcome the thermal limitation a variety of other lead-frame alloys are used (Tummala and Rymaszewski, 1989). The lead frames are 0.15–0.25-mm thick and are chemically etched or mechanically stamped from rolled strip stock. Chips are bonded to the lead frames by silicon/gold eutectic bonding or by use of silver-filled epoxy or polyimide adhesive paste. After encapsulation the leads are formed to special shapes for soldering to cards. In Fig. 5 the DIP leads are wider at the top than at the bottom, creating an integral standoff when

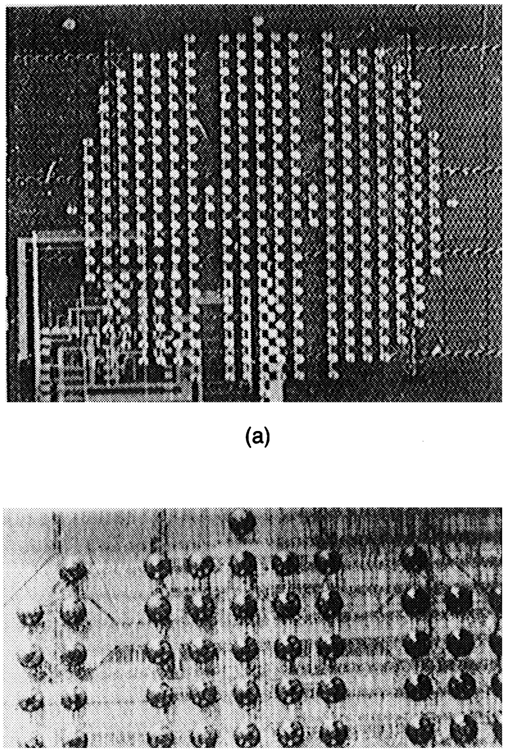

(a)

(b)

FIG. 4. (a) A 5000-circuit chip with 354 solder bumps in an area array; (b) a close-up view of solder bumps (courtesy of IBM Corp).

the leads are inserted into holes in cards. The PLCC leads shown in Fig. 5 have a "J bend" and the SOIC leads are "gull wing" shaped for surface-mount soldering.

3.2 Encapsulation Materials

Molding compounds are mixtures of materials optimized for a specific application. The most commonly used resin is cresol novolac epoxy, which has relatively low shrinkage on curing, low moisture absorption, chemical inertness, good adhesive properties, and a glass transition temperature of 150 °C. Other resins have glass transition temperatures between 125 and 200 °C. For applications requiring high heat resistance, low dielectric properties, and low-temperature properties silicone materials can be used. In addition to the resin, curing accelerators and curing agents, flame retardants, colorants, and inert fillers make up the molding compound. The filler comprises about 70% by weight and is needed to control shrinkage, improve thermal conductivity, and enhance thermal shock resistance. Powdered fused silica (SiO_2) is the primary filler used today. Nevertheless, the thermal expansion coefficients of molding compounds are typically (120–200) × 10^{-7}/°C. Care must be taken in the transfer molding process and the physical design of the chip, lead frame, and wire bonds to avoid stress-induced failures (Manzione, 1990).

4. PRINTED-CIRCUIT CARDS AND BOARDS

4.1 Structures

Printed-circuit technology is the lowest-cost method for interconnecting a large number of diverse components. The basic structure is shown in Fig. 6. The simplest cards are single sided and have wiring on only one surface. More complex cards are double sided and have wiring and components on both outside surfaces. The most complex cards have multilayer wiring and can support components on one or both outside surfaces. In cards and boards, plated copper wiring is connected between layers by plated through holes in a metal/polymer laminate. The plated through holes (PTH) are also used to mount chip carriers, passive devices, and connectors. The voltage reference planes distribute current, provide shielding, and control the impedance of the signal transmission lines. This basic structure is used to make a range of second-level packages, from inexpensive small cards containing one or two planes with signal linewidths of 1.0 mm and PTH diameter of 0.84 mm to high-cost, high-performance boards 600 × 700 mm in size and containing as many as 50 planes with signal linewidths of 0.05 mm and PTH diameter of 0.35 mm. Fig. 7 is a picture of a large card used in workstations.

4.2 Materials and Processes

The most common insulator is FR-4 (fire-retardant, epoxy-impregnated glass cloth), although polyimide-glass laminates are also used, and a variety of polymers and fillers are used for specific applications requiring low thermal expansion coefficient, low dielectric

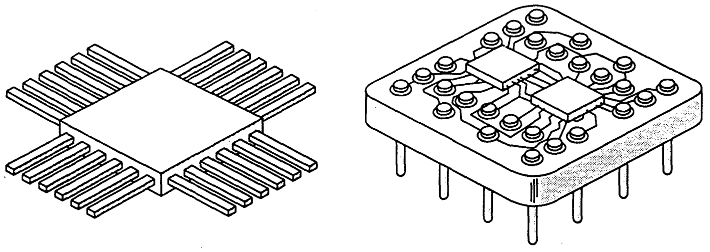

Flatpack

Pin grid array (PGA)

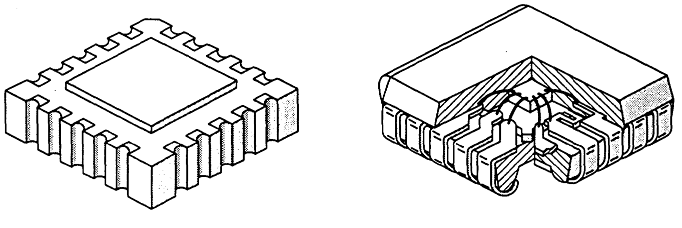

Leadless chip carrier (LCC)

Plastic leaded chip carrier (PLCC)

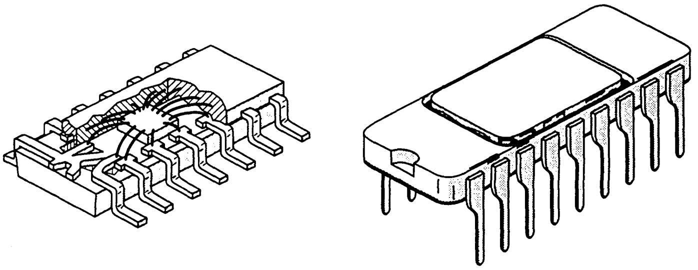

Small outline integrated circuit (SOIC)®

Dual in-line package (DIP)

FIG. 5. First-level packages (from Seraphim *et al.*, 1989, reproduced by permission) are commonly attached by solder to second-level wiring structures. In the LCC, the solder bridges into the groves shown.

constant, or high thermal stability (Seraphim *et al.*, 1989). High thermal stability is achieved with a number of polyimide materials. Low–dielectric-constant boards with very low dissipation at high frequencies are made of polytetrafluoroethylene (PTFE) (Goosey, 1985).

The fabrication process involves laminating copper on both sides of the insulating material (FR-4 or equivalent), drilling holes, and joining the two sides by electroless copper plating of the hole walls and surfaces. The copper wiring is then created by either a subtractive or an additive process. Both use photolithographic techniques to define a wiring pattern. In the subtractive process, the unwanted copper is etched away after through-hole plating. In the additive process, as the through holes are plated, copper is added in a pattern to the surface and finally coated with tin before the photoresist is removed and the base copper etched away. To construct a multilayer structure, partially cured "prepreg" (epoxy-glass laminate alone) is placed between each of the metallized double-sided laminates or cores to form a stack. Copper surface planes may be added and the entire stack laminated and treated as a core. Holes are drilled where desired to intersect internal copper pads joined to lines or voltage planes and the whole unit is plated again (Seraphim *et al.*, 1989).

5. CERAMIC PACKAGING

Ceramics are used to fabricate many of the chip carriers shown in Fig. 5. There are ceramic dual in-line packages (CERDIPS) and leadless chip carriers. While the shapes are similar to molded plastic packages, ceramic chip carriers are more rugged and have the

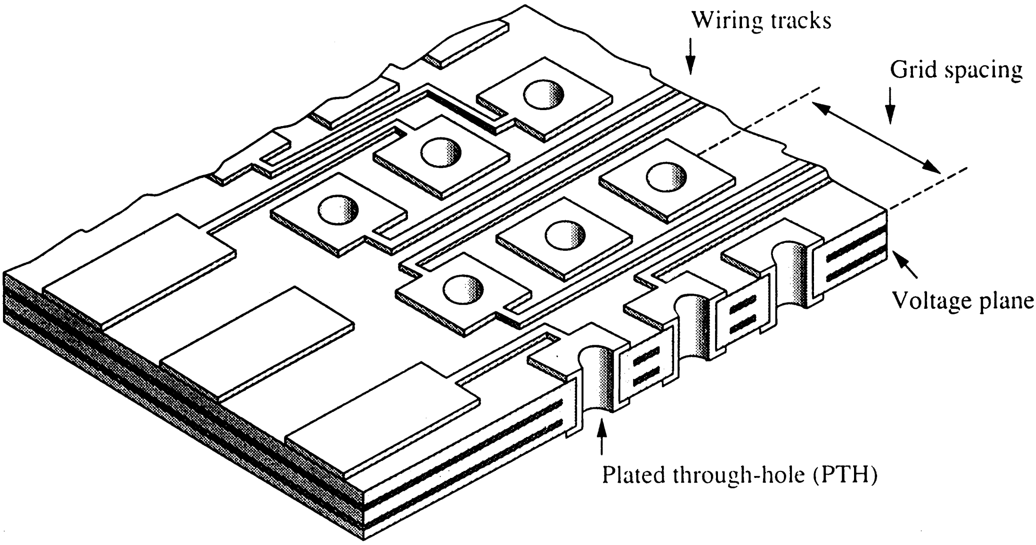

FIG. 6. Printed circuit card with two signal layers. There are two wiring tracks within the grid spacing (two lines per channel). (From Seraphim *et al.*, 1989, reproduced by permission.)

advantage that they can be made hermetic, i.e., the interior is totally isolated from the ambient atmosphere. Hermeticity is not achievable with plastic packages because of the gas permeability of plastics. Ceramic packages are sealed with metal or ceramic lids. Moreover, ceramics have thermal expansion coefficients well matched to semiconductors, thus reducing bonding stress, and they have relatively high thermal conductiv-

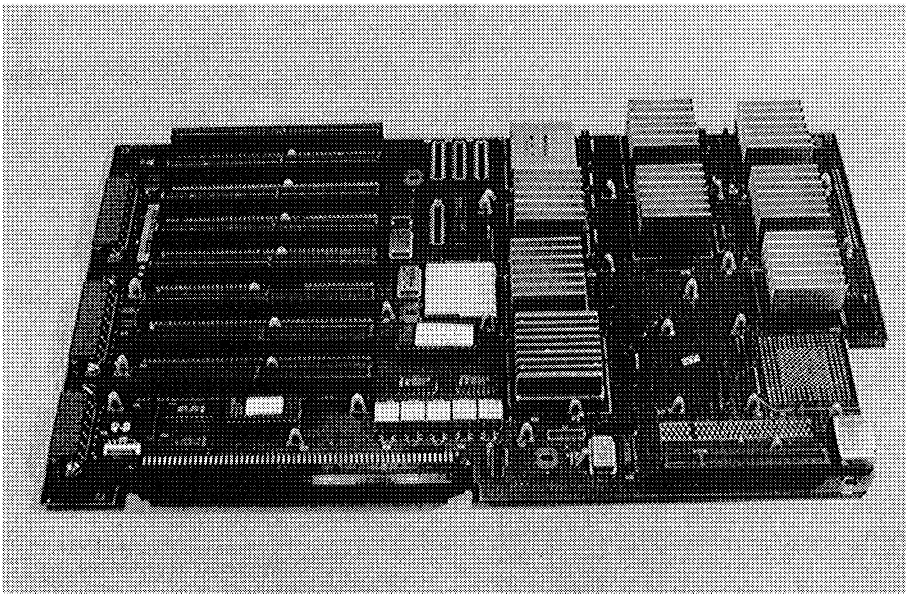

FIG. 7. This printed circuit card or "planar" is 250 × 420 mm^2 and interconnects the components of a workstation computer, including surface-mounted single-chip modules with heat sinks and connector blocks for additional small cards (courtesy of IBM Corp).

ity. In addition, the strength and rigidity of ceramics easily accommodate a large-area array of pins for dense interconnection to higher package levels.

5.1 Metal-on-Ceramic Packages

With the exception of the pin-grid array (PGA), the chip carriers in Fig. 5 have wirebond connections between chip and package I/O leads. In the ceramic PGA there is a single layer of metal on the ceramic surface to provide wiring between the components mounted on the PGA and make connections to the pins. This wiring is made by either thick- or thin-film processing. In the thick-film process, a screen or metal stencil is prepared with openings defining the circuit pattern. The screen is placed over the ceramic substrate. An ink paste containing metal particles is made to fill the openings in the stencil by use of a compliant blade or "squeegee" as in common silk-screen printing processes. The stencil is removed and the substrate is fired at high temperature, sintering the metal particles in the circuit pattern. In the thin-film process, metal is sputter deposited on the ceramic and patterned by photolithography and metal etching.

5.2 Multilayer Ceramic Substrates

Multichip ceramic substrates with multiple wiring layers are made by cosintering a ceramic and metal system. A powdered ceramic is mixed with solvents and organic binder polymer and tape cast and dried into a "green sheet," which is cut to size. Holes are punched for via interconnections and a thick-film metallization process used to fill the holes and define the circuit lines. After inspection the good sheets are stacked and laminated and then sintered at high temperature to form a solid body with metal conductors and ceramic insulation. Surface metals are overplated as required and pins brazed to bottom pads. Chips are mounted directly on the module, and it is hermetically sealed. An example is pictured in Fig. 8. The most complex ceramic MCMs today contain more than 60 layers of low-dielectric glass-ceramic with several hundred meters of cosintered copper wiring and are used to interconnect over a hundred chips.

5.3 Ceramic Materials

Alumina (Al_2O_3) is the most common ceramic in use. Thick-film processes typically employ 96% Al_2O_3, 3–3.5% SiO_2, and a small

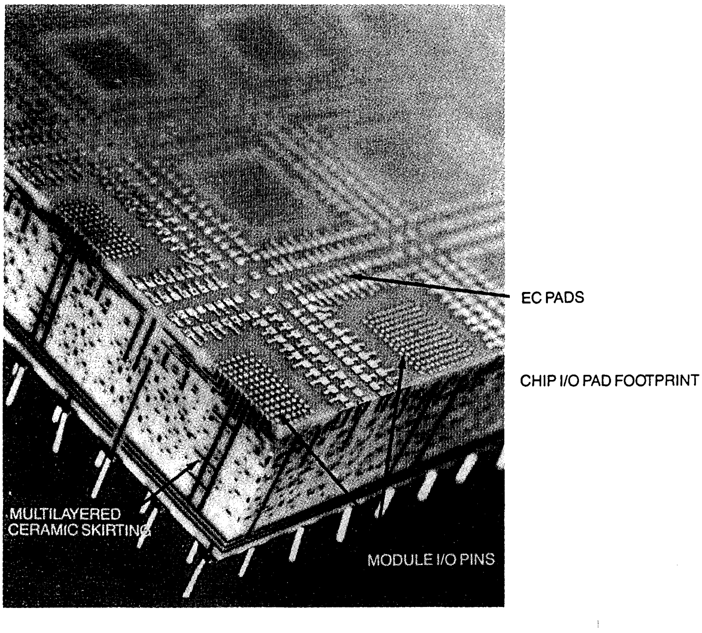

FIG. 8. Cross section of multilayer ceramic showing dense wiring (courtesy of IBM Corp., from Tummala and Rymaszewski, 1989, reproduced by permission).

amount of MgO and CaO. Other ceramics such as BeO, SiC, and AlN are used where higher thermal conductivity is desired, and glass-ceramics or mixtures of glass and ceramic that are compatible with copper metallization are used when a low dielectric constant is required for electrical performance. Since metals must be cosintered with the ceramic, the choice of metallization for a given ceramic is limited since the sintering temperature of the ceramic must be somewhat lower than the melting point of the metal. (Tummala and Rymaszewski, 1989).

6. THIN-FILM WIRING

Multilayer thin-film wiring is applied to the surface of a printed circuit card, ceramic package, or passive mechanical support where high performance, high density, and extreme compactness of interconnection wiring are required. The dimensions of transmission lines given in Fig. 2 indicate that roughly ten times as many wiring tracks can be made in a wiring layer using thin-film wiring as in thick-film card or ceramic wiring. Thus, roughly ten thick-film layers are equivalent to one thin-film layer in wirability, and thin-film layers are each only one tenth as thick.

6.1 Materials

The high electrical performance of thin-film wiring is due to the use of low–dielectric-constant insulators such as spray- or sputter-deposited borosilicate glass or quartz or, more commonly, polymers such as polyimides. Polyimides are chosen for their ability to form high-quality, defect-free thin films by spin or spray application and for high thermal stability, low moisture uptake, and good solvent resistance. Copper is the metal of choice, although aluminum and others are used. Since copper does not adhere well to polymers and can react with uncured material, very thin layers (≤1000 Å) of chromium, titanium, or other materials are placed between the copper conductor and the polymer to act as adhesion promoters and passivants.

6.2 Sequential Patterning

The very fine dimensions of thin-film wiring require sequential fabrication techniques; an example is given in Fig. 9. There are a variety of methods for metal and polymer deposition and patterning, and many different processes are in use. Metals may be deposited by evaporation, sputtering, electroless plating, or electroplating. They may be patterned using photolithographic techniques by wet etching as used in printed circuit cards, by plating in a resist pattern, or by lift-off processes used in making on-chip wiring. Polymer insulators may be patterned by wet chemical etching, reactive ion etching, or laser ablation, or they may be intrinsically photosensitive and directly patterned optically like a photoresist.

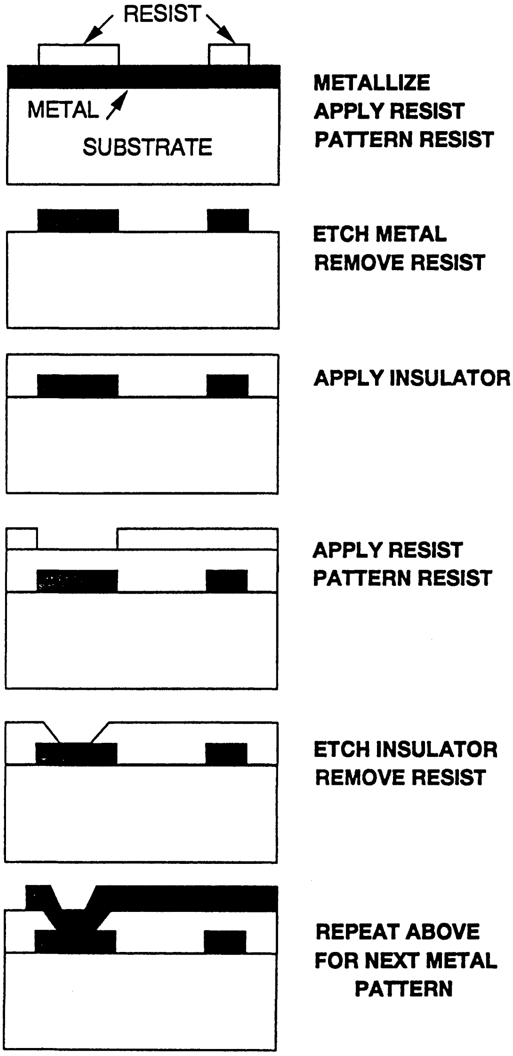

FIG. 9. Typical thin-film wiring fabrication process. In this figure, the sequential process flow is from top to bottom.

The choice of a particular patterning process for metal or polymer is dependent on cost and is often dictated by the aspect ratio (the ratio of height to width) of the most demanding feature required by the electrical design. Processes that are capable of forming high–aspect-ratio structures such as electroplating of a metal in a patterned photoresist or reactive ion etching of a polymer are generally more expensive than processes that can only form low–aspect-ratio structures such as wet etching of metals or polymers.

7. PACKAGE-TO-PACKAGE INTERCONNECTIONS

7.1 Pins

The most common interconnection method between first- and second-level packages is to solder the leads or pins of one component into the plated through holes of a card or board. Typically tin-lead eutectic solder (63% Sn, 37% Pb), which is liquid above 183 °C, is used in the wave soldering process shown schematically in Fig. 10. After flux is applied, the populated card is preheated and passed over a fountain of flowing solder where capillary forces cause the solder to fill the plated through holes.

In pin-grid array (PGA) and multilayer ceramic packages, pins are attached to the package before mounting on the card or board. In the simplest packages, pins are swaged into the package while more complex structures require pins that are brazed to pads on the bottom of the package. For multilayer ceramic packages, Ni/Au-plated Kovar pins are attached with eutectic Au-Sn (20% Au, 80% Sn) braze, which melts at 280 °C. Higher melting point (779 °C) but less expensive eutectic copper-silver braze or other materials are also used.

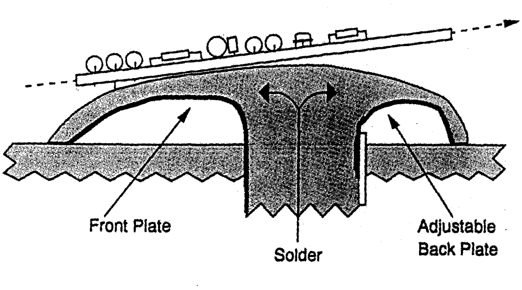

FIG. 10. Wave soldering (adapted from Seraphim *et al.*, 1989).

7.2 Surface-Mount Technology (SMT)

Relative to pin-in-hole mounting, surface-mount technology reduces package size and cost and improves board utilization. It does this by joining leads from components to the board surface on a finer pitch than the 2.54-mm hole spacing currently in use, thus allowing the whole board cross-sectional area to be used for wiring unobstructed by holes. Currently SMT connections are made with a pitch as fine as 0.3 mm, allowing several hundred perimeter bonding leads for a single chip module. Surface-mounted packages are either leadless or leaded chip carriers, some of which are shown in Fig. 5. For these packages a very short surface solder joint provides both electrical and mechanical connection between package and board. In leadless chip carriers, the solder must accommodate the shear strain caused by the relative displacement of package and board during thermal excursions. In leaded chip carriers, the strain is accommodated by elastic deformation of the leads. Since solder fatigue failure is a concern for SMT components, the contact metallurgy and intermetallic compound formation as well as the solder time and temperature cycle must be carefully controlled. Typically, eutectic Sn-Pb solder is applied by dipping or wave soldering, paste screening, or stenciling, plating, or application of preforms. The major joining techniques are wave soldering, vapor-phase reflow, infrared or hot-gas reflow, laser soldering, or hot-knife soldering. In vapor-phase reflow, the populated board with applied solder is placed in the saturated vapor of a boiling fluorocarbon. The vapor condenses on the board raising the temperature of all components to the fluorocarbon boiling point (typically 215 °C) thus reflowing the solder and joining all components simultaneously at uniform temperature.

While the vast majority of surface-mounted components are attached by solder, the use of conductive polymer adhesives is growing. Metal-particle–filled polymers such as epoxies form acceptable conductors in some applications. The principal metals used are silver, copper, or gold although silver/palladium alloys have been utilized (Goosey, 1985).

Highly loaded polymer films are isotropic conductors. Less highly loaded polymer films can be insulating until compressed between opposing metal pads. These materials form anisotropic "z-axis" conductive films and are finding increasing applications.

7.3 Connectors

A connector is a device to separate and reconnect electronic circuits repeatedly, and they come in many sizes and shapes. All involve the application of force by springs or other elements and a remakable low-resistance electrical contact surface. An example of a spring-loaded connector for a package with pins is shown in Fig. 11. The cross section of a card-edge connector is shown in Fig. 1, where spring-loaded contacts join to pads along both edges of a printed circuit card. Metal spring materials are most frequently beryllium copper or phosphor bronze, and interface contact metals are gold, palladium, or tin. The insulation that supports the metal contacts is most often an injection-molded plastic housing. Very high-density connectors are made from an area array of metal contact elements separated by a plastic insulator. The individual contact elements may be solid metal structures or shaped metal springs in a rigid plastic housing. The contact elements may also be metal structures in a compressible plastic medium that provides the contact force.

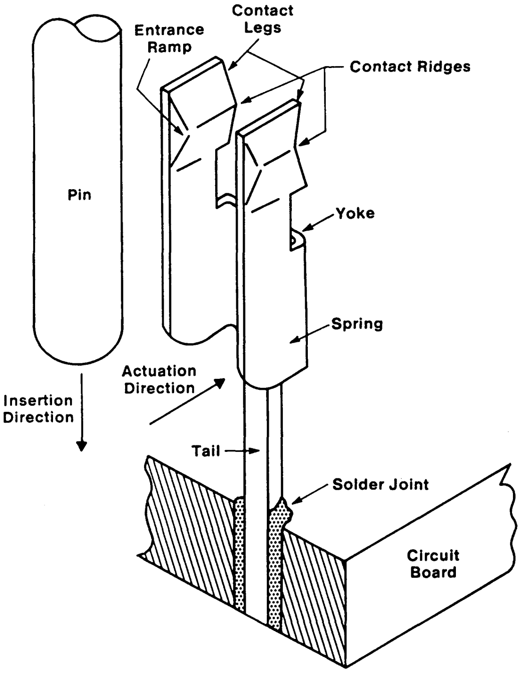

FIG. 11. Mechanical pin connector (from Tummala and Rymaszewski, 1989, reproduced by permission).

7.4 Cables

Cables are used to carry power, sense very low voltages or current, or transmit data over relatively long distances. Signal cables are flexible transmission lines containing a signal wire and a ground return conductor (a wire or shield, braid or continuous) separated by a dielectric and covered with a protective polymer jacket. Typical transmission lines are twin-lead, tri-lead, or coaxial. Since data are transmitted over long distances, the speed of transmission and the attenuation in the cable are important. For high speed, low–dielectric-constant polymers are used, and air is introduced to lower ϵ further. For example, expanded polytetrafluoroethylene has a dielectric constant of 1.3 so that signals travel at 88% of the speed of light. At high frequencies current flows in the skin depth, δ, on the outer surface of the inner conductor and the inner surface of the outer conductor. The skin depth δ is proportional to $1/\sqrt{f}$, where f is the frequency. Hence, the signal is attenuated by the skin-effect resistance and insulator dielectric loss. For low attenuation at high frequencies conductors are coated with silver or copper.

Electrical cables are used to interconnect packaged components within a system or for communication between systems. For short distances electrical cables can be used for high-speed data transmission, but for the highest data transmission rates or high-speed transmission over long distances, optical fiber cables are used. These cables consist of a graded-index SiO_2-based glass fiber in a protective plastic sheath. Optical fiber cables can carry data with minimal losses over several kilometers at rates in excess of Gbits/s.

8. HEAT TRANSFER

In order to maintain active device characteristics and ensure long lifetimes, semiconductor junctions must be maintained typically at 85 °C or less. The function of the package is to cool the chips by transferring the heat generated to an air or water reservoir or heat sink.

8.1 Thermal Engineering

The cooling structure of the package is characterized by the thermal resistance R, in °C/W, that it presents to a device generating heat Q (in watts, W), resulting in a temperature difference ΔT (in °C) between the device and the cooling fluid (a gas or liquid):

$$R = \Delta T/Q.$$

In low-power applications, R may be ~100 °C/W, but for densely packaged high-power chips, R must be engineered to be ≤ 10 °C/W. Heat is usually transferred to the cooling fluid by convection, which may be natural or forced (by a fan or pump) or due to boiling. The ability of the fluid to accept heat from area A (in m^2) is characterized by the heat transfer coefficient h (in W/m^2 °K):

$$Q = hA\Delta T.$$

Ranges of values for h are given in Table 3, but since h encompasses all the complex effects of convection, including fluid motion, fluid properties, and system geometry, the exact value for a particular cooling structure must be modeled and ultimately determined empirically.

8.2 Cooling Structures

Low-power electronic systems use natural-convection air cooling, but most computers use forced-air cooling, relying on circulating fans and finned metal heat spreaders as required to cool components. The multichip module packages require the most sophisticated cooling structures because they achieve high system performance by closely interconnecting many powerful chips. Figure 12 shows a few of the wide variety of structures in use. Depending on the thermal resistance required between the device and the cooling fluid, the heat may be removed from the chips through the interconnection package or from the back of the chip through various thermal structures such as solder, adhesives, paste, ceramic structures, metal plates, or metal pistons.

9. FUTURE CHALLENGES, EMERGING TECHNOLOGIES

Progress in package technology is characterized by steady evolutionary improvement of existing technology as well as revolutionary changes associated with new materials. For example, the outer lead bond pitch of surface-mount packages has continued to shrink over a 20-year period, and wiring density of cards and ceramic modules is increasing. There is ongoing work to reduce the dielectric constant of processable ceramics and polymers by introducing new materials. Other new materials such as synthetic diamond for thermal management could open new dimensions in packaging. But the key challenges to package technology will be, as always, to package new and ever more powerful systems at the lowest possible cost.

9.1 Optoelectronics

Optoelectronics devices and fiber-optic links are increasingly being used for high-bandwidth communications between electronic systems. While semiconductor devices used as transmitters (such as lasers and light-emitting diodes) and receivers can be interconnected and powered using the technologies described above, special package designs are required to couple the optical fiber to the semiconductor transmitter or receiver. For example, to avoid signal losses the alignment tolerance for single-mode fibers is about 0.5 μm. The coupling of fiber to device can

Table 3. Heat transfer coefficients (W/m^2 °K).

	Air (or N_2)	Fluorocarbon liquid	Water
Free convection	3–12	100–300	—
Forced convection	10–100	200–2000	3000–7000
Boiling	10 000	2000–6000	3000

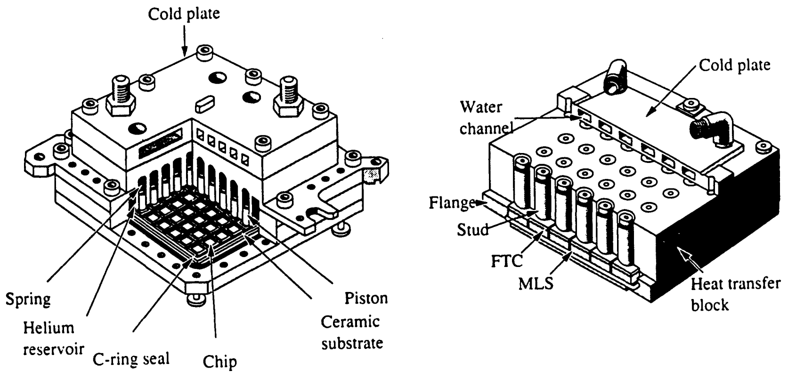
The IBM 3081 thermal conduction module

The NEC SX liquid cooling module

Cold plate
Spring
Helium reservoir
C-ring seal
Chip
Piston
Ceramic substrate
Water channel
Cold plate
Flange
Stud
FTC
MLS
Heat transfer block

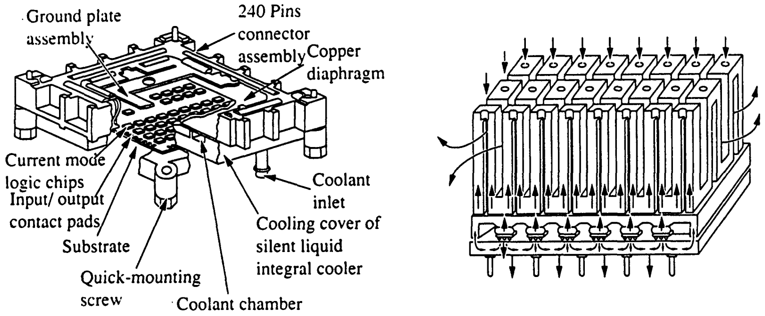
The Honeywell silent liquid integral cooler module

The IBM 4381 air-cooled module
Module cross-sectional view

Ground plate assembly
240 Pins connector assembly
Copper diaphragm
Current mode logic chips
Input/ output contact pads
Substrate
Quick-mounting screw
Coolant chamber
Coolant inlet
Cooling cover of silent liquid integral cooler

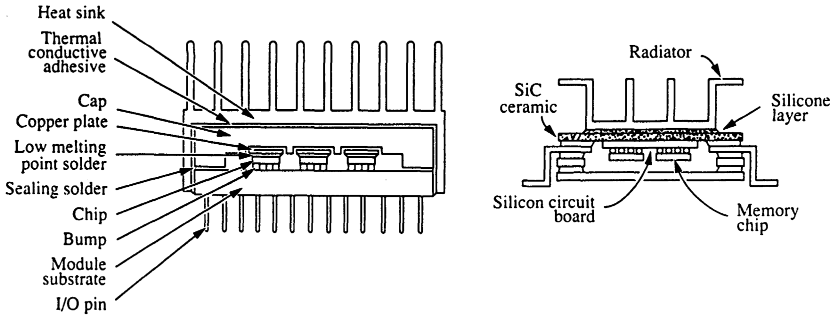
The Mitsubishi high thermal conduction module

The Hitachi SiC RAM module

Heat sink
Thermal conductive adhesive
Cap
Copper plate
Low melting point solder
Sealing solder
Chip
Bump
Module substrate
I/O pin
Radiator
SiC ceramic
Silicone layer
Silicon circuit board
Memory chip

FIG. 12. Schematic diagrams of multichip-module cooling structures (from Seraphim *et al.*, 1989, reproduced by permission).

be accomplished by precision mechanical mounting (with or without lenses), an adherent butt joint, or thin-film optical waveguides (Tummala and Rymaszewski, 1989). Future systems may require dense arrays of fibers coupled to small high-speed semiconductor switches that include optical transmitters and receivers.

9.2 Electromagnetic Management

Future packaging technology must successfully handle processor chips running at frequencies of hundreds of megahertz and the electromagnetic emissions and interference associated with these systems. In addition, as wireless communication becomes dominant for portable systems, the fabrication of high-frequency transmitters and receivers of electromagnetic waves will become important. Electromagnetic field modeling, component placement, and shielding will become key aspects of package technology. The need for low-cost, light-weight, moldable shielding materials could open up a range of applications for conducting polymer materials.

9.3 Low-Power, Low-Cost, Light-Weight Systems

As semiconductor technology reduces device size and power and careful design puts more and more function on a single chip, potentially very low-cost, reasonably powerful systems will be possible. Package engineers will be increasingly challenged to build thin, light-weight, but very robust structures that provide signal interconnection, power distribution, and cooling for these future portable systems.

Works Cited

Goosey, M. T. (1985), *Plastics for Electronics*, London: Elsevier Applied Science Publishers.

Harper, C. A. (Ed.) (1991), *Electronic Packaging and Interconnection Handbook*, New York: McGraw-Hill, Inc.

Manzione, L. T. (1990), *Plastic Packaging of Microelectronics Devices*, New York: Van Nostrand Reinhold.

Minges, M. L. (Tech. Chairman) (1989), *Electronic Materials Handbook*, Vol. 1, *Packaging*, Materials Park, OH: ASM International.

Radke, C. (1969), in: N. Garaffa, Jr. (Chmn.), *Proceedings of the 6th Annual SHARE–ACM–IEEE Design Automation Workshop*, New York: IEEE, pp. 257–267.

Seraphim, D. P., Lasky, R. C., Li, C. Y. (1989), *Principles of Electronic Packaging*, New York: McGraw-Hill Inc.

Sinnadurai, F. N. (1985), *Handbook of Microelectronics Packaging and Interconnection Technologies*, Ayr, Scotland: Electrochemical Publications Ltd.

Tummala, R. R., Rymaszewski, E. J. (1989), *Microelectronics Packaging Handbook*, New York: Van Nostrand Reinhold.

PALEOMAGNETISM

J. E. T. CHANNELL, *Department of Geology, University of Florida, Gainesville, Florida, U.S.A.*

Introduction 89
1. **Relevant Characteristics of the Geomagnetic Field** 89
2. **Origin of Magnetization in Rocks and Sediments** 90
2.1 Magnetic Minerals 90
2.2 Natural Magnetization Processes in Rocks and Sediments 91
3. **Laboratory and Analytical Techniques** 91
3.1 Demagnetization Techniques 91
3.2 Analysis of Demagnetization Data 92
3.3 Statistics 93
3.4 Field Tests for the Age of Magnetization 93
3.5 Identification of Carriers of Natural Remanence 94
4. **Applications of Paleomagnetic Studies** 95
4.1 Paleomagnetic Poles and Continental Drift 95
4.2 Tectonic Rotations in Mountain Belts 95
4.3 Magnetic Polarity Stratigraphy 95
4.4 Secular Variation Magnetic Stratigraphy in Lake Sediments 96
4.5 Rock Magnetic Stratigraphy 96
4.6 Insights into Geomagnetic Field Behavior from Secular Variation, Excursions, Polarity Transitions, Paleointensities, and Polarity Bias 96
Glossary 97
Works Cited 98
Further Reading 98

INTRODUCTION

Paleomagnetism is the study of the magnetization of rocks and sediments, and provides a record of the direction and intensity of the ancient geomagnetic field.

Paleomagnetism was a cornerstone of the "revolution" in the earth sciences that began in the 1950s and led to the development of the theory of plate tectonics. A series of landmark papers published in the 1950s on paleomagnetic data from Britain and North America provided the first land-based quantification of continental drift (e.g., Runcorn, 1956). Forty years on, paleomagnetism remains the primary means of reconstructing the relative positions of rock bodies in the geologic past (on scales from individual thrust sheets to entire continents). In addition, paleomagnetism has led to an improved understanding of the mechanisms for generation of the geomagnetic field, for secular variations, and for polarity reversals. The record of polarity reversals in rocks and sediments provides a unique means of global correlation, and has become the thread on which we hang our geologic time scales.

1. RELEVANT CHARACTERISTICS OF THE GEOMAGNETIC FIELD

About 90% of the geomagnetic field at the Earth's surface can be modeled by a dipole close to the center of the Earth. The nondipole components of the geomagnetic field contribute to the short-term (secular) variations of the field, and are responsible for the deviation of the north geomagnetic pole from the north pole (situated at the rotation axis of the Earth). It is generally accepted, on the basis of records for the last 10^5 years, that secular variation averages out over $\sim 10^4$ years. The time-averaged geomagnetic field is thought to approximate a geocentric axial

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

dipole field. This is of paramount importance to paleomagnetic studies because it allows us to predict the location of the ancient north pole relative to an observation point where the time-averaged field direction is recorded by ancient rocks or sediments.

The dipole component of the geomagnetic field has changed polarity numerous times in the geologic past. The present state of "normal" polarity has existed for about the last 780 000 years. Our template for the sequence of geomagnetic polarity reversals since about 150 Ma B.P. (before present) is based on magnetic anomalies from the oceans, which are produced by normally and reversely magnetized ocean crust formed at the mid-oceanic ridges. The so-called "seafloor spreading" process often occurred at rather constant rates over long periods of geologic time and spreading was usually symmetric either side of the ridge. For these reasons, the record of geomagnetic reversals in the ocean floor is often a faithful record of the sequence of polarity reversals through time. Statistical analysis of the reversal record since about 150 Ma B.P. indicates that there are too few short polarity intervals for a Poisson process; however, this may be attributed to the inadequacy of the record. It is generally considered that the lack of memory characteristic of the Poisson process is also a characteristic of polarity reversal history. Because of the dipole nature of the main field and the "geologically instantaneous" nature of the reversal process (the process takes about 3000 years), the record of individual reversals in rocks and sediments provides a unique global time marker, a unique means of global correlation.

2. ORIGIN OF MAGNETIZATION IN ROCKS AND SEDIMENTS

2.1 Magnetic Minerals

Iron oxide and iron sulfide minerals are the most common carriers of remanent magnetization in rocks and sediments. Titanomagnetites [$x\mathrm{Fe_2TiO_4} \cdot (1-x)\mathrm{Fe_3O_4}$] are present in many types of igneous and metamorphic rocks, and are therefore a detrital constituent in many sedimentary rocks. The titanomagnetites are ferrimagnetic at room temperature for values of x less than about 0.8. Many sediments, particularly deep-water pelagic limestones, contain magnetite ($\mathrm{Fe_3O_4}$) or low-titanium titanomagnetite that may be biogenic in origin, being derived from marine organisms, such as bacteria and some molluscs, which incorporate magnetite in their metabolism (see BIOMAGNETISM AND GEOMAGNETIC FIELD DETECTION BY ORGANISMS). Biogenic magnetite usually lies in the single-domain (SD) grain-size range, which is the optimal grain size for retention of magnetic remanence over geologically significant time intervals.

Titanohematites [$x\mathrm{FeTiO_3} \cdot (1-x)\mathrm{Fe_2O_3}$] are antiferromagnetic for $0 < x < 0.5$. The weak net magnetic moment in hematite (α-$\mathrm{Fe_2O_3}$) and titanohematites is due to a slight canting of the spin moments. Titanohematites are constituents in some igneous and metamorphic rocks. Hematite is a common constituent of sedimentary rocks and is an important carrier of magnetic remanence in clastic and chemical sediments. Hematite may occur in sediments as detrital "specularite" or as an authigenic pigment that grows during diagenesis.

Maghemite (γ-$\mathrm{Fe_2O_3}$) has the spinel structure of magnetite and the chemical composition of hematite. Maghemite is ferrimagnetic and has similar magnetic properties to magnetite. Maghemite is common as a low-temperature oxidation product of magnetite in both igneous and sedimentary rocks. It reverts to hematite at about 350 °C.

Goethite (α-FeOOH) is the most common iron oxyhydroxide mineral and is antiferromagnetic, the net moment being due to canting of spin moments. Goethite is an important product in low-temperature oxidation, particularly of iron sulfides, and is a constituent of manganese nodules and the Pacific red clays where it may precipitate directly from seawater.

Pyrrhotite (FeS_{1+x} where $x = 0$–0.14) is the most commonly occurring iron sulfide mineral that is ferrimagnetic. The ferrimagnetism is restricted to a limited compositional range ($0.1 < x < 0.14$). Pyrrhotite occurs in basic igneous rocks and can grow during sediment diagenesis in certain reducing environments. The most common iron sulfide mineral (pyrite, $\mathrm{FeS_2}$) is paramagnetic, and does not therefore carry magnetic remanence.

2.2 Natural Magnetization Processes in Rocks and Sediments

Thermal remanent magnetization (TRM) is the magnetization acquired by a magnetic mineral grain during cooling through its blocking temperature, above which the magnetic ordering of a grain is randomized by thermal energy. The blocking temperature is a function of grain size and composition and lies below the Curie (Néel) temperature. Igneous rocks cooling from magma generally acquire a TRM in the direction of the ambient geomagnetic field.

Chemical remanent magnetization (CRM) is produced by growth of a magnetic mineral grain through its blocking volume, below the blocking temperature of the grain. The blocking volume is the critical volume for a particular grain shape and composition, above which the grain can carry magnetic remanence. Magnetic mineral grains acquire a CRM as they grow through their blocking volume in the ambient geomagnetic field, and the direction of the CRM is usually maintained during further grain growth. CRM is an important mode of remanence acquisition during sediment diagenesis.

Detrital or depositional remanent magnetization (DRM) is an important magnetization process in sediments. Detrital or biogenic magnetic mineral grains with preexisting TRM or CRM can align themselves with the ambient field during deposition. The alignment occurs in the upper few centimeters of the soft sediment, and is therefore often referred to as postdepositional remanent magnetization (pDRM). The lock-in depth of DRM has been estimated experimentally and theoretically to be of the order of 10 cm, but will depend on grain size, porosity, sedimentation rate, and bioturbation depth.

Viscous remanent magnetization (VRM) is the remanence produced by a weak external field (such as the geomagnetic field) over a long period of time. VRM is an important source of secondary magnetization in rocks and sediments. The intensity of VRM is proportional to the logarithm of time that the sample experiences the weak field. The proportionality constant (viscosity coefficient) increases with temperature. Many rocks and sediments carry a VRM acquired during the present Brunhes interval of "normal" polarity of the geomagnetic field.

3. LABORATORY AND ANALYTICAL TECHNIQUES

The introduction of cryogenic magnetometers in the mid 1970s resulted in a significant reduction in magnetometer noise level, and opened up paleomagnetic research to a wider range of rocks and sediments (see Collinson, 1983). Modern cryogenic magnetometers can measure sample magnetization intensities as low as $\sim 10^{-5}$ A/m with sufficient precision for precise determination of paleomagnetic direction.

Rocks and sediments can become magnetized at any time during their history, and hence the natural remanent magnetizations (NRM) of rock and sediment samples are often the resultant of more than one magnetization component. TRM in igneous rocks and DRM in sedimentary rocks can usually be associated with the original cooling of the igneous rock or original deposition of the sedimentary rock and, in such cases, are referred to as "primary" magnetizations. Subsequent chemical alteration or heating of the rock can yield a CRM or TRM that would then be a "secondary" magnetization. VRM often has recent origin, having formed since the last geomagnetic polarity reversal. Laboratory techniques in paleomagnetism are designed to resolve the various magnetization components that constitute the total magnetization of the rock or sediment. The objective is usually to resolve the "primary" magnetization component, that associated with the formation of the rock or sediment. The age of the "primary" magnetization component can then be derived from the paleontological/radiometric age of the rock or sediment.

3.1 Demagnetization Techniques

The external field strength required to alter the magnetic remanence of a grain will depend on the coercivity of the individual grain, which is controlled by composition and grain size/shape. The dependence of coercivity on composition and grain size is utilized in *alternating-field* (AF) *demagnetization* to resolve the magnetization components carried by different grain populations. If the NRM of the sample is a composite of individual magnetization components that have distinct coercivity spectra, the AF method is suitable for resolving the individual components. The normal procedure is to demagnetize the sample progressively by a stepwise increase in peak alternating field, measuring the remaining magnetic remanence after each step. The range of peak alternating fields is typically 5–100 mT. For each demagnetization step, the sample is placed in a magnetically shielded coil and the alternating field is smoothly ramped down to zero in the absence of background field. Grain remanences with coercivity lower than the peak alternating field will be reoriented as the peak field is ramped down, and their contributions to the remaining remanence will cancel out. The net effect is to demagnetize these grains. As the peak alternating field is progressively raised, a larger proportion of the total magnetic grain population becomes demagnetized and ceases to contribute to the remanence of the sample. For a full description of the design of alternating-field demagnetizers see Collinson (1983).

The time required for magnetic remanence of a grain to decay in zero external field (the relaxation time) is a function of ambient temperature, grain size, and composition. The blocking temperature is the temperature at which the relaxation time falls within the time range of laboratory experiments ($\sim 10^3$–10^4 s). As the relationship between relaxation time and temperature is exponential, the exact length of the experiment does not, within reason, affect the observed blocking temperature. Two populations of magnetic grains with different origins, and therefore different compositions and/or grain sizes, might be expected to have distinct blocking-temperature spectra. This principle is utilized in *thermal demagnetization* to separate magnetization components carried by different magnetic grain populations. Normal practice is to heat the sample in progressive temperature increments (until the magnetization falls below the magnetometer noise level), allowing the sample to cool in field-free space after each heating increment. The remanence of all grains with blocking temperature less than the maximum temperature will become randomized as the sample cools in field-free space. The magnetization measured after each heating step can be used to determine the directions of magnetization components with distinct blocking-temperature spectra. For a description of the design of thermal demagnetizers, see Collinson (1983).

Leaching with hydrochloric acid can provide a means of selective destruction of magnetic grains. Such *chemical demagnetization* is applicable to siliceous rocks, particularly red sandstones, which contain a secondary magnetization due to authigenic growth of fine-grained pigmentary hematite. The dissolution rate of coarse-grained hematite of detrital origin is lower than that of the finer-grained pigmentary hematite of diagenetic origin, because of the difference in the ratio of surface area to grain volume. Leaching in 3–10 N hydrochloric acid, and remanence measurement at certain time intervals during the leaching, provides a means of resolving magnetization components with distinct "leaching spectra."

3.2 Analysis of Demagnetization Data

The most widely used method for analyzing changes in magnetization direction and intensity during demagnetization is the orthogonal projection, also referred to as the Zijderveld diagram (see Zijderveld, 1967; Dunlop, 1979; Collinson, 1983). This is a simple, but extremely useful, means of displaying the change in magnitude and direction of the magnetization vector during demagnetization. Straight-line segments on these plots indicate that the magnetization vector being removed has constant direction. This direction can be computed, and defined as a component of the magnetization. It is important to note that straight-line segments on orthogonal projections may not always represent single components. They may also represent composites of two or more components with identical demagnetization (coercivity, blocking, or leaching) spectra. The magnetization components cannot be resolved if the overlap in demagnetization spectra is too great. However, if the unblocking temperature spectra overlap, this does not necessarily mean that the coercivity spectra will overlap. The choice of demagnetization technique will be critical to whether the components of magnetization can be resolved.

The points on the orthogonal projection that define straight-line segments are commonly picked by eye. The component direction associated with a particular straight-line segment can then be computed by a least-squares technique (Kirschvink, 1980). This method gives a maximum angular deviation

(MAD) value, which provides an estimate of the goodness of fit of the straight line to the points on the projection.

In the special case when three magnetization components with overlapping demagnetization spectra are present, the Hoffman–Day method (Hoffman and Day, 1978) may help to resolve the component directions. The method involves plotting subtracted vectors on an equal-area stereographic projection, at each stage in the progressive demagnetization process. In a demagnetization interval characterized by two overlapping demagnetization spectra, the subtracted vectors will plot on a great circle between the two component directions. If three overlapping demagnetization spectra are present, the intersection of the two great circles can give an estimate of the direction associated with the component with intermediate blocking temperature, even though this component direction is not directly resolved by the demagnetization process. Orthogonal projection in this case would not resolve the intermediate component.

If two magnetization components with partially overlapping demagnetization spectra are present in the sample, the resultant vector will trace a great circle in stereographic projection as the overlapping interval is demagnetized (Halls, 1978). If one of the two components has dispersed directions, the great circles from individual samples will converge to give the direction of the undispersed component. The method is only applicable in the special case where one of the components is dispersed, as there will be no appreciable convergence of the great circles if both components have single directions. The dispersion of one of the components can occur in a number of situations. For example, the primary component in conglomerate clasts will be dispersed relative to the secondary component that postdates conglomerate deposition. In this case, the secondary component could be resolved by remagnetization circle analysis.

3.3 Statistics

When the the magnetization components for individual samples have been determined, the sample magnetization components are combined to produce a site mean direction. The site mean directions may then be combined to compute an overall mean direction.

The probability density function generally used in paleomagnetism to determine the mean and dispersion estimates is referred to as the Fisher distribution (Fisher, 1953). This distribution assumes circular symmetry about the mean direction. Each observation is considered as an independent unit vector, and the mean direction is determined by summing the Cartesian coordinates of each of the vectors. A resultant vector **R** of the distribution can be calculated as the vector sum of the vectors. The concentration parameter k for N vectors is calculated by

$$k = (N - 1)/(N - R).$$

For $k > 10$, the radius of the 95% circle of confidence, in which the true mean should lie at the 95% confidence level, can be calculated by

$$\alpha_{95} \sim 140^\circ/(kN)^{1/2}.$$

The use of Fisher statistics implies that the population of vectors has a Fisherian distribution, a condition that is often not satisfied. Although the Fisherian distribution is almost universally used as the basis for statistical treatment in paleomagnetism, the Bingham distribution is preferred by some workers as it is better suited to the bimodal (normal and reversed polarity) and azimuthal asymmetry which is inherent in paleomagnetic data (Onstott, 1980). When using Fisher statistics, each vector should be equally independent and it is therefore important to maintain the sample/site hierarchy.

3.4 Field Tests for the Age of Magnetization

As rocks and sediments can become magnetized at any time in their history, field tests that constrain the age of the magnetization are critical to any paleomagnetic study. The *fold test* involves the determination of dispersion of paleomagnetic vectors during unfolding of the structure, and therefore allows determination of the age of magnetization relative to the age of folding. The age of folding can usually be determined by stratigraphic constraints. The fold test is best suited to individual folds, although it can be adapted for regional use on tilted beds that are not part of a single structure. The test involves analysis of the dispersion of paleomagnetic vectors during progressive unfolding of the structure, by rotating the limbs of the structure to the horizontal about the strikes of the individual limbs. In cases where the structure has plunge, the plunge should be rotated to the horizontal prior to the rotation of the limbs about the strike. McElhinny (1964) used the ratio of the concentration parameter (k) before and after tilt correction to determine whether the fold test was positive (i.e., whether the magnetization was prefolding). Critical values of k_2/k_1 (k_1 and k_2 are values of k before and after tilt correction of the structure, respectively) are listed by McElhinny (1964) for vector populations of 2–20. In the last ten years, it has become increasingly apparent that magnetization components are sometimes coeval with deformation/folding. Statistical criteria for analysis of the dispersion of paleomagnetic vectors that are more rigorous and better suited to the progressive unfolding of structures have recently been introduced (e.g., McFadden, 1990).

The second category of field test is the *conglomerate test*, which can be applied to conglomerates, breccias, or agglomerates. If the magnetization component directions from the clasts are randomized, the magnetization component predates the formation of the clasts. The same magnetization component recognized in an adjacent *in situ* outcrop of the same lithology can then be considered to have age very similar to the stratigraphic age of outcrop. The test for randomness (Watson, 1956) can be applied to the vectors derived from the clasts.

The third category of field test is the *baked contact test* which can be applied to igneous intrusions and their thermally metamorphosed contact with the country rock (Wilson, 1962). If the igneous intrusion and the baked country rock carry the same magnetization component direction, and the country rock outside the aureole carries a different (preceding) magnetization, then this is considered to constitute a test for a "primary" TRM in the igneous body.

3.5 Identification of Carriers of Natural Remanence

Magnetic mineral phases tend to have characteristic modes of occurrence, and therefore, the identification of magnetic remanence carriers provides clues to the age of magnetization components. Recent advances in instrumentation have allowed complete magnetic hysteresis loops to be measured even on weakly magnetized sediments. For example, the alternating-gradient force magnetometer (μMAG) can be used to determine values of saturation magnetization (M_s), saturation remanence (M_{rs}), and coercive force (H_c) even on weakly magnetized limestones, giving good constraints on the magnetic mineral content of the sample. High-temperature experiments designed to determine Curie (Néel) temperatures, and low-temperature experiments designed to monitor transitions in magnetic behavior (Verwey and Morin transitions), are important means of determining the magnetic mineral content of rocks and sediments. However, in order to determine the carriers of natural remanent magnetization (NRM), the characteristics of the NRM itself must be determined as fully as possible.

As described above, progressive demagnetization is the standard procedure for determining the component content of the NRM. Thermal and alternating-field demagnetization characteristics indicate the blocking temperature and coercivity spectra, respectively, of NRM components.

Blocking temperature is a function of composition and grain size of the magnetic carrier. For single-domain (SD) grains of constant composition, the blocking temperature will be a function of grain size. Because of the large grain size of hematite at the SD/MD (single-domain to multidomain) threshold (~2 mm), the SD range for hematite covers most of the remanence-carrying grain sizes encountered in sediments. Therefore, in the case of hematite, blocking temperature is a useful guide to grain size. For titanomagnetites [$x\mathrm{Fe_2Ti0_4} \cdot (1 - x)\mathrm{Fe_30_4}$], the blocking temperature is a function of composition and grain size and shape. As x decreases, the Curie point increases almost linearly, and as blocking temperature cannot exceed the Curie point, the available blocking temperature range will increase as x decreases. For SD titanomagnetite grains of uniform shape and composition, the blocking temperature will be a function of grain volume/shape. The blocking temperature range in itself is not generally diagnostic of magnetic mineralogy, other than the fact that the maximum blocking temperature cannot exceed the Curie (Néel) point of the mineral composition. In the case of minerals with relatively low Curie (Néel) points, such as geothite (~120 °C), the blocking-temperature range is rather diagnostic because it is so restricted. Mineralogical changes as a result of heating during thermal demagnetization can obscure the blocking-temperature spectrum. For example, maghemite alters to hematite at about 350 °C, and iron sulfides will oxidize in air to form magnetite at about 300 °C.

The coercivities of remanence of hematite and geothite are typically about two and three orders of magnitude greater, respectively, than that of SD magnetite. For this reason, hematite and goethite magnetizations are very resistant to alternating-field (AF) demagnetization. For titanomagnetites, coercive force increases with x and with decreasing grain size. The grain-size effect will dominate in natural samples with fairly uniform composition.

4. APPLICATIONS OF PALEOMAGNETIC STUDIES

4.1 Paleomagnetic Poles and Continental Drift

The focus of the first modern paleomagnetic studies, carried out in the 1950s, was the documentation of continental drift. Deviating paleomagnetic data from European and North American rocks of the same age were shown to reflect the development of the North Atlantic Ocean since the formation of the rocks. As a result of the dipolar nature of the main time-averaged geomagnetic field, the paleolatitude of the rock body can be determined from the paleomagnetic inclination and the coordinates of the sampling site. It is customary to express a paleomagnetic direction recorded at a particular location as a virtual geomagnetic pole (VGP), determined assuming an axial geocentric dipole. Such paleomagnetic poles are a standard means of expressing paleomagnetic data, and the time-space distribution of paleomagnetic poles for a particular crustal block or continent is known as an apparent polar wander path (APWP). The APWP for a particular continent is an expression of the movement of the crustal block relative to the rotation axis of the Earth. True polar wander, the movement of the rotation axis relative to the entire lithosphere, is generally considered to be insignificant compared to apparent polar wander. Apparent polar wander paths are a principal means of reconstructing the relative position of crustal blocks in the geologic past and are continually updated by addition of new data. For a recent compilation of the APWPs for the major continents and the constraints that these data place on paleogeographic reconstructions, see Van der Voo (1993).

4.2 Tectonic Rotations in Mountain Belts

Paleomagnetic data have proved useful in documenting rotations of small-scale crustal blocks during deformation in mountain belts. One of the aims of geological studies in fold and thrust belts is the "palinspastic" reconstruction of the deformed belt by field mapping and seismic reflection studies of the subsurface structure. Such reconstructions require knowledge of the relative rotation of thrust sheets in the deformed belt. Paleomagnetic component directions that predate the deformation can be used to monitor these rotations. In zones of oblique convergence, large-scale vertical-axis rotations of thrust sheets relative to the adjacent stable continent have been recorded (e.g., Channell *et al.*, 1990). Clockwise and anticlockwise rotations are usually associated with dextral and sinistral shear, respectively, although the fault patterns that accommodate the observed rotations have not been well documented.

4.3 Magnetic Polarity Stratigraphy

Because of the dipole nature of the geomagnetic field, the recording of ancient polarity reversals in rocks and sediments provides a unique means for global correlation. No other type of event (isotopic, paleontologic, lithologic, or geochemical) recorded in rocks or sediments provides the same potential for precise global correlation. As our knowledge of the history of geomagnetic polarity reversal has improved, the geomagnetic polarity time scale (GPTS) has become central to our determination of geologic time. The template for the GPTS for the last 150 Ma is based on oceanic magnetic anomalies, which are produced by alternating zones of normal and reversed magnetized ocean crust formed at the mid-oceanic ridges. One of the roles of magnetic polarity stratigraphy is to correlate the GPTS to paleontologic events, radiometric ages, and isotopic events. Almost any stratigraphic study benefits from magnetostratigraphic data on account of the potential for precise global correlation, and for correlation to the GPTS and current geologic timescales (see King and Channell, 1991).

4.4 Secular Variation Magnetic Stratigraphy in Lake Sediments

"Magnetic stratigraphy" usually refers to magnetic polarity stratigraphy. Short-term "secular" variations of the geomagnetic field are generally considered to average out over intervals of about 10^4 years. Directional changes associated with secular variation provide a potential means of correlation in recent sediments. Whereas conventional polarity stratigraphy depends on the recognition of directional changes of remanent magnetization of about 180°, secular variations rarely exceed a few tens of degrees. Hence, sampling and measurement techniques in secular variation studies are particularly critical. The technique has been largely restricted to lake sediments because of their high sedimentation rates and relative lack of bioturbation. Most marine sediments are too bioturbated and have sedimentation rates that are too low for a high-fidelity record of secular variation. As secular variation is largely due to the nondipole components of the geomagnetic field, secular variations are not globally synchronous but can be used for local/regional correlation.

4.5 Rock Magnetic Stratigraphy

Rock magnetic stratigraphy involves the utilization of nondirectional magnetic properties of sediments as a means of correlation. The first rock magnetic parameter to be widely used for correlation was magnetic susceptibility. Magnetic susceptibility essentially monitors the concentration of ferrimagnetic minerals. The parameter provides a rapid means of sediment correlation, and is now a standard technique for correlation of lake and deep-sea sediment cores. The field of rock magnetic stratigraphy has broadened considerably in the last ten years with the use of magnetic parameters that are sensitive not only to magnetic mineral concentration but also to grain size and mineralogy. These techniques have been applied to the study of environmental changes in lake catchments by monitoring changes in magnetic mineralogy of the sediments (Thompson and Oldfield, 1986). Similar methods have been applied to marine sediments to study changes in sediment provenance and clastic/biogenic fluxes, which can often be related to paleoclimatic change (see King and Channell, 1991).

4.6 Insights into Geomagnetic Field Behavior from Secular Variation, Excursions, Polarity Transitions, Paleointensities, and Polarity Bias

Globally distributed secular-variation records provide a means of determining the history of drift of centers of secular variation, and hence establish the longevity of the westward drift observed in historical observatory records. The time-space configuration of the nondipole field revealed by detailed secular-variation-records will provide insights into geomagnetic field-generation mechanisms.

Brief departures of the geomagnetic field from the full polarity state (geomagnetic excursions) appear to have been recorded in the present normal-polarity interval (the Brunhes chron) at numerous locations, but many recorded excursions have not been documented at multiple sites and therefore cannot be firmly established as geomagnetic phenomena. The short duration ($\sim 10^3$–10^4 years) of geomagnetic excursions leads to difficulty in recording these phenomena in rock or sediment sections, as the rock/sediment record is usually episodic on these time scales. In addition, age control is often insufficient to establish unequivocal correlation for Brunhes excursions recorded at different locations. The best-documented excursion is from sediments at Mono Lake (California) where the excursion (dated at 27–29 ka B.P.) is followed by four successive recurrences of the excursion wave form (Lund *et al.*, 1988). An additional excursion at about 120 ka B.P. (Blake event) has been recorded at several locations in the Mediterranean region and in Atlantic cores. These and other records imply that the geomagnetic field exhibits a complete spectrum of behavior from high-amplitude secular variation to full polarity reversals. Brief polarity reversals that never stabilize, where the dipole field immediately reverts back to its initial polarity, have also been recorded. Detailed high-fidelity records of the process of polarity reversal, recorded in sediments with high accumulation rates, have tended to indicate that the dipole configuration of the field is maintained during the reversal process (Clement, 1992), and that the paleomagnetic pole follows preferred longitudinal paths during the reversal process.

Absolute geomagnetic paleointensity determinations in volcanic rocks generally utilize a method based on the so-called Thellier technique (Thellier and Thellier, 1959). The majority of the determinations are for the last few thousand years, and display considerable scatter due to short-term (secular) variation in geomagnetic intensity, nonrigorous experimental techniques, and use of unsuitable archaeological and volcanic materials. As volcanic eruption is discontinuous, the volcanic record of paleointensity is discontinuous. Each determination requires independent age control, and the uncertainties in these ages inhibit correlation to other paleointensity determinations.

More continuous paleointensity records for the last 100–140 ka have been obtained from deep-sea sediment cores (see Tauxe, 1993). These relative paleointensity records are obtained by normalizing the remanence (NRM) intensity using anhysteretic remanence (ARM), isothermal remanence (IRM), or susceptibility to compensate for variations in concentration of magnetic minerals. Good correlation of paleointensity records from widely spaced locations indicates that these data record the time-averaged dipole moment of the geomagnetic field, effectively averaging out secular variation. The scope of sediment paleointensity studies has been expanded back to 3.9 Ma B.P. by recent records from the equatorial Pacific (Valet and Meynadier, 1993). Several intriguing features appear in these records. Paleointensity lows in the Brunhes chron appear to correlate to short "polarity events" or excursions. Reversals tend to coincide with lows in the virtual axial dipole moments (VADMs), with abrupt increases immediately after reversal. The longer the duration of the polarity chron, the greater the abrupt increase in VADM at its base; and each polarity chron appears to be characterized by sawtooth decay of VADM leading up to the next reversal. This has led to the suggestion that the reversal frequency, and the occurrence of "polarity events" or excursions, are ultimately controlled by the intensity of the geomagnetic dipole moment, which also controls the stability of a particular polarity state.

As our record of polarity reversal history improves and is pushed back in time, we will be able to ascertain whether there is any long-term pattern or polarity bias in the record that may signify changes in convection in the Earth's core. Our present record is fairly complete for the last 150 Ma; however, this is a small fraction of the entire history of polarity reversal, which goes back at least 3000 Ma.

GLOSSARY

Anhysteretic Remanent Magnetization (ARM): The magnetic remanence produced in the laboratory by smooth decay of a strong alternating field in the presence of a weak direct field.

Apparent Polar Wander Path: The time-space distribution of paleomagnetic poles for a particular crustal block.

Blocking Temperature: The critical temperature above which the magnetic moment of a mineral grain becomes randomized by thermal energy. The relaxation time of the grain magnetization may increase from a few seconds to millions of years as the grain passes through its blocking temperature.

Blocking Volume: The threshold volume above which a grain can carry remanence at room temperature. Below this volume, the grain is superparamagnetic and the blocking temperature is below room temperature.

Coercivity of Remanence: The back field required to reduce a remanent magnetization to zero, after saturation.

Curie (Néel) Temperature: Temperature above which saturation magnetization is zero. It is a characteristic of a particular magnetic mineral.

Demagnetization: The process of progressive demagnetization of a magnetic remanence (using alternating fields, thermal, or chemical techniques). Progressive demagnetization is used as a means of resolving the component content of the remanence. The process is successful where the components have nonuniform responses to one of the demagnetization techniques.

Isothermal Remanent Magnetization (IRM): The magnetic remanence produced in the laboratory by temporary application of a strong direct (dc) field.

Magnetic Domains: Uniformily magnetized regions in a magnetic grain.

Magnetic Susceptibility: Measure of degree of magnetization of a substance in a weak direct field, equal to M/H (ratio of induced magnetization to applied field).

Multidomain Grain: A grain which contains more than one magnetic domain.

Natural Remanent Magnetization (NRM): The naturally occurring remanent magnetization of a sediment, rock, or mineral.

Normal Polarity: Polarity of the geomagnetic dipole field as of today.

Paleomagnetic Pole or Virtual Geomagnetic Pole (VGP): The pole calculated from a magnetization direction recorded at a particular location assuming that the geomagnetic field responsible for the magnetization direction was a geocentric dipole.

Reversed Polarity: Polarity of the geomagnetic dipole field opposite to that of today.

Secular Variations: Geologically short-term geomagnetic field variations, which are generally believed to average out in $\sim 10^5$ years.

Virtual Axial Dipole Moment (VADM): The magnetic moment of an axial geocentric dipole calculated from paleointensities observed at a particular location.

Works Cited

Channell, J. E. T., Oldow, J. S., Catalano, R., D'Argenio, B. (1990), "Paleomagnetically Determined Rotations in the Western Sicilian Fold and Thrust Belt," *Tectonics* **9**, 641–660.

Clement, B. M. (1992), "Evidence for Dipolar Fields during the Cobb Mountain Geomagnetic Polarity Reversals," *Nature* **358**, 405–407.

Collinson, D. W. (1983), *Methods in Rock Magnetism and Palaeomagnetism*, London: Chapman and Hall, 503 pp.

Dunlop, D. J. (1979), "On the Use of Zijderveld Vector Diagrams in Multicomponent Paleomagnetic Samples," *Phys. Earth Planet. Sci. Lett.* **20**, 12–24.

Fisher, R. A. (1953), "Dispersion on a Sphere," *Proc. R. Soc. London A* **217**, 295–305.

Halls, H. C. (1978), "The Use of Convergent Remagnetization Circles in Palaeomagnetism," *Phys. Earth Planet. Interiors* **16**, 1–11.

Hoffman, K. A., Day, R. (1978), "Separation of Multicomponent NRM: A General Method," *Earth Planet. Sci. Lett.* **40**, 433–438.

King, J. W., Channell, J. E. T. (1991), "Sedimentary Magnetism, Environmental Magnetism, and Magnetostratigraphy," *Rev. Geophys.* Suppl. (April), 358–370.

Kirschvink, J. L. (1980), "The Least-Squares Line and Plane and the Analysis of Paleomagnetic Data," *Geophys. J. R. Astr. Soc.* **62**, 699–718.

Lund, S., Liddicoat, J. C., Lajoie, K. R., Henyey, T. L., Robinson, S. W. (1988), "Paleomagnetic Evidence for Long Term (10^4 years) Memory and Periodic Behavior in the Earth's Core Dynamo Process," *Geophys. Res. Lett.* **15**, 1101–1104.

McElhinny, M. W. (1964), "Statistical Significance of the Fold Test in Paleomagnetism," *Geophys. J. R. Astron. Soc.* **8**, 338–340.

McFadden, P. L. (1990), "A New Fold Test for Paleomagnetic Studies," *Geophys. J. Int.* **103**, 163–169.

Onstott, T. C. (1980), "Application of the Bingham Distribution Function in Paleomagnetic Studies," *J. Geophys. Res.* **85**, 1500–1510.

Runcorn, S. K. (1956), "Paleomagnetic Comparisons Between Europe and North America," *Geol. Assoc. Can. Proc.* **9**, 77–85.

Tauxe, L. (1993), "Sedimentary Records of Relative Paleointensity of the Geomagnetic Field: Theory and Practice," *Rev. Geophys.* **31**, 319–354.

Thellier, E., Thellier, O. (1959), "Sur l'Intensité du Champ Magnétique Terrestre dans le Passé Historique et Géologique," *Ann. Geophys.* **15**, 285–376.

Thompson, R., Oldfield, F. (1986), *Environmental Magnetism*, London: Allen and Unwin, 227 pp.

Valet, J.-P., Meynadier, L. (1993), "Geomagnetic Field Intensity and Reversals during the Past Four Million Years," *Nature* **366**, 234–238.

Van der Voo, R. (1993), *Paleomagnetism of the Atlantic, Tethys and Iapetus Oceans*, Cambridge, U.K.: Cambridge University Press, 411 pp.

Watson, G. S. (1956), "A Test for Randomness of Directions," *Mon. Not. R. Astron. Soc.* **7**, 160–161.

Wilson, R. L. (1962), "The Paleomagnetism of Baked Contact Rocks and Reversals of the Earth's Magnetic Field," *Geophys. J. R. Astron. Soc.* **7**, 194–202.

Zijderveld, J. D. A. (1967), "A.C. Demagnetization of Rocks, Analysis of Results," in: D. W. Collinson, K. M. Creer, S. K. Runcorn (Eds.), *Methods in Paleomagnetism*, Amsterdam: Elsevier, 254–286.

Further Reading

Butler, R. F. (1992), *Paleomagnetism*, Boston: Blackwell Scientific Publications, 319 pp. A recent text dealing with basic principles of paleomagnetism.

Collinson, D. W. (1983), *Methods in Rock Magnetism and Palaeomagnetism*, London: Chapman and

Hall, 503 pp. A useful text, particularly for laboratory methods and instrumentation.

O'Reilly, W. (1984), *Rock and Mineral Magnetism*, Glasgow, Scotland: Blackie, 220 pp. Basic principles of rock magnetism.

Tarling, D. H. (1983), *Palaeomagnetism, Principles and Applications in Geology, Geophysics and Archaeology*, London: Chapman and Hall. A general text on paleomagnetism.

Thompson, R., Oldfield, F. (1986). *Environmental Magnetism*, London: Allen and Unwin, 227 pp. A text dealing with environmental magnetism, the use of magnetic properties of rocks and sediments as a monitor of environmental change, also some discussion of rock magnetic stratigraphy and secular variation stratigraphy.

Van der Voo, R. (1993), *Paleomagnetism of the Atlantic, Tethys and Iapetus Oceans*, Cambridge, U.K.: Cambridge University Press, 411 pp. A recent text dealing with the application of paleomagnetism to paleogeography and tectonics, including a useful database of paleomagnetic poles.

PARAMAGNETISM

YAKO YAFET, *AT&T Bell Laboratories (retired), Murray Hill, New Jersey, U.S.A.*

	Introduction	101
1.	**The Magnetic Moment**	102
1.1	Orbital Moment	102
1.2	Some Definitions and Units	103
1.3	Spin Moment	103
1.4	The Moment in a Magnetic Field	104
2.	**Atomic Paramagnetism**	104
2.1	Atomic States	104
2.2	Hund's Rules	105
2.3	Spin–Orbit Interaction	105
2.4	Landé *g* Factor	106
2.5	Magnetic Susceptibility	107
2.6	Saturation	108
3.	**Paramagnetism in Solids**	109
4.	**Ionic Solids**	110
4.1	Quenching of the Orbital Moment	111
4.2	Kramers's Theorem	112
4.3	Jahn–Teller Effect	112
4.4	g and g'	115
5.	**Conduction-Electron Paramagnetism**	116
5.1	Pauli Susceptibility	116
5.2	Measurement of χ_P	117
5.3	Localized Moments and Kondo Effect	117
5.4	Nonlocal Susceptibility and RKKY Interaction	118
5.5	Spin-Density Waves	118
5.6	Large g Factors in Semiconductors	118
6.	**Applications**	118
6.1	Magnetic Cooling	118
6.2	The Maser	119
6.3	Tunable Spin-Flip Raman Laser	119
7.	**Summary**	119
8.	**Acknowledgments**	120
	Glossary	120
	Works Cited	121
	Further Reading	121

INTRODUCTION

Paramagnetism may be defined as the study of the properties of matter that are related to the existence of free magnetic moments. A primary manifestation of paramagnetism is a positive magnetic susceptibility, but there are many phenomena that are connected with the existence of free moments and we shall examine some of them in this article.

Michael Faraday was the first person to study systematically the magnetic properties of matter. This work was published in three volumes in London from 1839 to 1855 under the title *Experimental Researches in Electricity*, and it provides a fascinating glimpse into Faraday's great mind. Having subjected all kinds of substances (including rarefied gases and such common items as dry mutton and fresh beef) to the action of a powerful electromagnet (Faraday, 1846a), he concluded, "All matter appears to be subject to the magnetic force as universally as it is to the gravitating, the electric and [other] forces." (Faraday, 1846b). He observed that *all* substances respond to a magnetic field in one of two ways: They orient their long dimension either across the "lines of force," or alongside them. He called the first kind "diamagnetic" (see DIAMAGNETISM) and kept the term "magnetic" for the second. "Magnetic" was then a general term for substances that are attracted to the lodestone and to magnetic metals such as iron. Diamagnetic materials are repelled by a magnet. The word "paramagnetic" was coined in 1850 from the Greek prefix "para," which means alongside (as in parallel), to describe paramagnetic behavior. "Ferromagnetic" was used by W. Thomson (Lord Kelvin), also in 1850, in the expression "feebly ferromagnetic," which might have meant "paramagnetic." A clear distinction between "ferromagnetic" and "paramag-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

netic" as they are used today was made by P. Curie in 1895.

The magnetic properties of matter are due primarily to the electrons. Nuclear magnetism is interesting in its own right, but it is of order 10^{-6} smaller than electronic magnetism and will not be considered here. Diamagnetism is a universal property of matter that follows from Lenz's law (see ELECTRODYNAMICS, CLASSICAL): All electrons contribute a diamagnetic moment that, omitting considerations of anisotropy, points in the direction opposite to that of the magnetic field. Paramagnetism on the other hand occurs only in materials that have spontaneous, randomly oriented moments in their ground state (with the exception of Van Vleck paramagnetism, which is discussed in Sec. 2.5). These moments tend to align themselves parallel to an applied field, but thermal agitation, or in the case of metals the exclusion principle, opposes that tendency and the net result is partial alignment along the field. This is paramagnetic behavior, which again may be subject to anisotropies.

Static paramagnetism, as all equilibrium magnetic phenomena, is quantum mechanical in origin. This follows from the Bohr–Van Leeuwen theorem (Van Vleck, 1932), which shows, in essence, that in classical systems all dependence of the Hamiltonian function on an applied magnetic field can be eliminated without affecting the values of thermodynamic averages. Our treatment will therefore be quantum mechanical, but nevertheless thinking along classical lines is still useful in providing physical insight. It is interesting that 30 years before quantum mechanics, Langevin (1905), by postulating the existence of magnetic moments, was able to derive the temperature dependence of paramagnetic susceptibilities that had been observed ten years earlier by P. Curie (Curie, 1905).

In this article, the paramagnetic properties of free atoms, of ionic salts containing metallic elements having partially filled shells, and of metals are considered. Some applications of paramagnetic substances, such as magnetic cooling, the maser, and the tunable spin-flip Raman laser, are also mentioned.

The basic concept in the theory of magnetism is that of the magnetic moment. In the next section, we define this and show that it is given by the derivative of the Hamiltonian with respect to the applied magnetic field. This property is used in Sec. 2.5 to calculate the thermal average of the moment per unit volume, and the susceptibility.

1. THE MAGNETIC MOMENT

The magnetic moment $\boldsymbol{\mu}$ of an electron is comprised of two parts, the orbital moment $\boldsymbol{\mu}_l$ and the spin moment $\boldsymbol{\mu}_s$. These two quantities will be defined, and the total moment $\boldsymbol{\mu} = \boldsymbol{\mu}_l + \boldsymbol{\mu}_s$ will be shown to satisfy the relation

$$\mu_\alpha = -\frac{\partial \mathcal{H}}{\partial H_\alpha}, \tag{1}$$

where $\mathcal{H} = T + V$ is the Hamiltonian function, the sum of the kinetic and potential energies, and α denotes a Cartesian component of the applied field $\mathbf{H}$ and of $\boldsymbol{\mu}$. Unless otherwise specified, the magnetic field will be taken in the z direction. The moments of individual electrons i are additive so that in many-electron systems, Eq. (1) is valid for the total moment, $\boldsymbol{\mu} = \Sigma_i \boldsymbol{\mu}_i$. To prove Eq. (1) in a one-electron system, the dependences of $\mathcal{H}$ on the orbital and spin moments will be considered separately.

1.1 Orbital Moment

The definition of the magnetic moment of a moving electron of charge $-e$ (note that e is positive) with respect to a point of origin is

$$\boldsymbol{\mu}_l = -(e/2c)\, \mathbf{r} \times \dot{\mathbf{r}}, \tag{2}$$

where c is the velocity of light $\mathbf{r}$ the position, and $\dot{\mathbf{r}}$ the velocity. The effect of an applied field $\mathbf{H}$ on the electron's motion is obtained by adding a vector-potential term (Van Vleck, 1932, Ch. I) to the momentum operator $\mathbf{p}$ in the Hamiltonian, thus:

$$\mathcal{H} \rightarrow (1/2m)[\mathbf{p} + e\mathbf{A}/c]^2 + V, \tag{3}$$

where $\mathbf{H}$ is derived from the vector potential $\mathbf{A}$,

$$\mathbf{H} = \nabla \times \mathbf{A}. \tag{4}$$

It is easily shown that the velocity is given by

$$\dot{\mathbf{r}} = (1/m)[\mathbf{p} + e\mathbf{A}/c], \tag{5}$$

and substituting for $\dot{\mathbf{r}}$ in Eq. (2) gives

$$\boldsymbol{\mu}_l = -(e/2mc)\,\mathbf{r} \times [\mathbf{p} + e\mathbf{A}/c]. \tag{6}$$

In atomic systems, it is convenient to use the symmetric gauge

$$\mathbf{A} = \tfrac{1}{2}\mathbf{H} \times \mathbf{r}, \tag{7}$$

and substitution of Eq. (7) into Eq. (3) followed by differentiation with respect to H gives, for a field in the z direction,

$$-\frac{\partial \mathcal{H}}{\partial H} = -\frac{e}{2mc}\left[(xp_y - yp_x) + \frac{eH}{2c}(x^2 + y^2)\right]. \tag{8}$$

The right-hand (r.h.) side of Eq. (8) is identical with the z component of $\boldsymbol{\mu}_l$ given by Eq. (6), and this proves Eq. (1) as far as the dependence on the orbital moment is concerned. The second term on the r.h. side of Eq. (8), proportional to H, is the induced diamagnetic moment, which always contributes a positive term to the energy. This term does not contribute to the paramagnetism of localized electrons and it will be neglected here. The first term, independent of H, is proportional to the z component of the orbital angular momentum. For non-s states, it represents a spontaneous, paramagnetic moment that, if the potential V is spherically symmetric, is free to take different orientations in space and hence to lower the magnetic energy.*

1.2 Some Definitions and Units

Before introducing the spin moment, it is convenient to give some definitions and establish a notation. We define the dimensionless angular momentum $\mathbf{l} = \mathbf{r} \times \mathbf{p}/\hbar$, where $\hbar$ is Planck's constant divided by 2π. The eigenvalues of the components l_α are integers ranging from $-l$ to $+l$ where l is the orbital quantum number. In the absence of the diamagnetic term, the orbital moment reduces to

$$\boldsymbol{\mu}_l = -(e\hbar/2mc)\,\mathbf{l}. \tag{9}$$

As $\mathbf{l}$ is dimensionless, it is natural to take as unit of magnetic moment the Bohr magneton $\beta = e\hbar/2mc$, which has the value 0.927×10^{-20} in centimeter-gram-second (cgs) units. The magnetomechanical ratio γ is the algebraic ratio of the magnetic moment to the angular momentum. For orbital motion, Eq. (9) gives the value

$$\gamma_{\text{orb}} = -e/2mc. \tag{10}$$

The magnetomechanical factor g' is defined as the absolute value of the ratio of magnetic moment in units of the Bohr magneton to the angular momentum in units of $\hbar$. For orbital motion, $g'_{\text{orb}} = 1$.

1.3 Spin Moment

The existence of an intrinsic electron spin angular momentum was postulated by Uhlenbeck and Goudsmit (1925) in order to explain the doubling of the optical spectral lines in the alkali atoms. The components of the spin vector $\hbar\mathbf{s}$ obey the angular momentum commutation rules, and in order to give a multiplicity of 2 to the optical spectra, the components of $\mathbf{s}$ must have magnitude $\frac{1}{2}$. To be consistent with the anomalous Zeeman effect (see below), the magnetomechanical ratio of the spin moment must have the value $2\gamma_{\text{orb}}$ with γ_{orb} given by Eq. (10), and hence the magnetomechanical factor of the spin moment must be twice that of the orbital moment, or $g'_s = 2$ (the more precise value, including radiative corrections is $g'_s = 2.0023$). With $g_s \equiv g'_s \cong 2$, the spin magnetic moment becomes $\boldsymbol{\mu}_s = -2\beta\mathbf{s}$ and the interaction of the total magnetic moment with an external field $\mathbf{H}$, called the Zeeman interaction, becomes finally

---

*This clear separation into a paramagnetic, H-independent moment and a diamagnetic moment linear in H is meaningful only when the electron's wave function is localized in space, as it is in atoms and molecules. The wave functions of electrons in energy bands (Bloch functions) extend over the whole volume of the sample, and the coordinates x and y can take on arbitrarily large values. The contributions of the two terms to the energy are then comparable, giving rise to the orbital diamagnetism of Bloch electrons, and it is simpler to deal directly with the complete Hamiltonian [Eq. (3)].

$$\mathcal{H}_Z = -(\boldsymbol{\mu}_l + \boldsymbol{\mu}_s)\cdot \mathbf{H} = \beta(\mathbf{l} + 2\mathbf{s})\cdot \mathbf{H}. \qquad (11)$$

Differentiation with respect to H gives Eq. (1).

1.4 The Moment in a Magnetic Field

In his investigations of the influence of a magnetic field on light propagating through glass, Faraday (1846c) had discovered that the field produces a rotation of the plane of polarization of plane-polarized light (the Faraday effect). Convinced that light and magnetism were related, he set out to find the effect of a magnetic field on the spectrum of a light source but was unable to reach a conclusion (as were other workers who tried in the following 50 years), primarily because the effect is very small and requires sensitive means of detection. Zeeman (1896) succeeded in showing that a magnetic field produces a definite broadening of the spectrum but he was not able to resolve the line splitting into what is now the familiar "Zeeman effect." From the different polarization properties of the center and edges of the broadened spectrum, he concluded that the line was split into three components. This observation was the first indication of the space quantization of angular momentum, even though it was still explainable by Lorentz's classical theory of a harmonically bound electron. Had Zeeman been able to resolve the broadened line, he would have found that it was not split into three (the normal Zeeman effect), but into a larger number of components, and this was no longer explicable by Lorentz's theory. The anomalous Zeeman effect, as this is called, depends in detail on the initial and final states of the emitting atom, and requires for its explanation that the spin moment have a magnetomechanical factor $g_s' \equiv g_s = 2$, and be coupled (by the spin–orbit interaction) to the orbital moment.

The Zeeman splittings determine the magnitude of the induced paramagnetic moment and of the susceptibility. Experimentally, the latter is determined by measuring the force exerted on the induced moment by an inhomogeneous field. In the Gouy method, this is done as follows. The specimen, in the form of a long thin rod of cross section A, is suspended between the vertical pole faces of an electromagnet with the field H at the center of the pole faces pointing horizontally. The specimen is midway between the poles but vertically off the magnet axis so that it is partly in the fringing field region where H has a gradient in the vertical direction x. If H is not too large ($H \lesssim 1$ T), the induced moment is proportional to H and the magnetization, which is the moment per unit volume M, can be written $M = \chi H$ where χ is independent of H. The force df_x exerted on the volume element $dv = Adx$ is

$$df_x = dm\,\frac{dH}{dx}. \qquad (12)$$

where $dm = \chi HAdx$ is the moment induced in the volume element. Substituting for dm and integrating Eq. (12) gives

$$f_x = \tfrac{1}{2}\,\chi A(H_{\max}^2 - H_{\min}^2), \qquad (13)$$

where $H_{\max}$ and $H_{\min}$ are the maximum and minimum values, respectively, of the field inside the specimen. Knowledge of A, $H_{\max}$, and $H_{\min}$ combined with a measurement of f_x determines the susceptibility.

The calculation of the paramagnetic susceptibility of free atoms is described in the next section. Knowledge of atomic paramagnetism provides the basis for analyzing the paramagnetic properties of solids.

2. ATOMIC PARAMAGNETISM

At a temperature T, the states that contribute to the paramagnetism are the ground state and those states that are within a thermal energy kT above it, k being Boltzmann's constant. A brief discussion of the structure of the atomic states is thus in order.

2.1 Atomic States

The electronic wave functions of many-electron atoms cannot be known exactly, and in order to classify and calculate them one makes use of a scheme based on perturbation theory and known as the central-field approximation (see ELECTRONIC STRUCTURE OF ATOMS AND MOLECULES). This is as follows.

The total atomic Hamiltonian is split into three parts in order of decreasing magnitude,

$$\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2 + \mathcal{H}_{so}. \qquad (14)$$

The term $\mathcal{H}_1$, the major part, is the sum over all electrons of the kinetic energy and of a suitably chosen spherical potential. $\mathcal{H}_1$ can be solved exactly in terms of one-electron orbitals (nl) of energy ϵ_{nl} where n and l are the principal and orbital quantum numbers (q.n.) respectively. The electrons are assigned the lowest energy levels in accordance with the Pauli principle, and they successively fill the (nl) shells, which can accommodate $2(2l + 1)$ electrons each. Filled shells have zero orbital and spin momenta and hence zero magnetic moment. Only atoms with partly filled shells are paramagnetic. The set of q.n.s (nl) that specify an eigenstate of $\mathcal{H}_1$ is called a configuration, and for paramagnetism it is sufficient to specify only the number of electrons in open shells. It is customary to use spectroscopic notation for the l values; thus $l = 0, 1, 2, 3, \ldots$ is written $s, p, d, f, \ldots$, etc. For example singly ionized iron, which has six electrons in the $3d$ and one in the $4s$ shells, is written $3d^6 4s^1$. Being the sum of one-electron energies ϵ_{nl} that are independent of "orientation" (i.e., azimuthal q.n.), the energy of a configuration that represents a partially filled shell is highly degenerate.

The effect of $\mathcal{H}_2$, which includes the electron–electron interaction minus an average one-electron potential that is included in $\mathcal{H}_1$, is to remove partially the orientational degeneracy of a configuration, thus splitting the configuration into "terms." Because $\mathcal{H}_2 < \mathcal{H}_1$, the mixing of configurations usually can be neglected. The total orbital ($\mathbf{L} = \Sigma_i \mathbf{l}_i$) and spin ($\mathbf{S} = \Sigma_i \mathbf{s}_i$) angular momenta, summed over all the electrons, commute with $\mathcal{H}_1$ and $\mathcal{H}_2$ and hence can be used to label the term states, each of which is associated with particular L and S values and has degeneracy $(2L + 1)(2S + 1)$. This classification scheme, which is based on the smallness of the spin–orbit interaction $\mathcal{H}_{so}$ [the last member of Eq. (14)], is known as Russell-Saunders coupling. Lastly, when $\mathcal{H}_{so}$ is taken into account, the degeneracy within each term is partially removed: Since $\mathcal{H}_{so}$, as well as $\mathcal{H}_1$ and $\mathcal{H}_2$, commutes with the total angular momentum $\mathbf{J} = \mathbf{L} + \mathbf{S}$, $\mathbf{J}$ is a good q.n. and the main effect of $\mathcal{H}_{so}$ is to split each term into multiplets (LSJ) with J running from $|L - S|$ to $L + S$. Each multiplet still has degeneracy $2J + 1$, and this is ultimately removed by an applied magnetic field that splits the multiplet into $2J + 1$ equidistant Zeeman levels. It is this splitting that causes the anomalous Zeeman effect. Because $\mathcal{H}_{so}$ does not commute with $\mathbf{L}$ or $\mathbf{S}$, the q.n.s L and S are now only approximate, and a multiplet (LSJ) may contain small admixtures of other multiplets ($L'S'J$) with $L' = L \pm 1$ and correspondingly $S' = S \mp 1$.

Multiplets having $J \neq 0$ possess a spontaneous moment while multiplets having $J = 0$ give rise only to an induced paramagnetic moment. The next two subsections describe how the lowest-energy term and the lowest multiplet belonging to that term are found.

2.2 Hund's Rules

Hund noted the empirical fact that the lowest-energy term of a configuration satisfies the following two conditions:

1. It has the largest S; and
2. if several terms have that same S value, it has the largest L among them.

Physically, these rules follow from the repulsive character of the electron–electron interaction: Concerning **1**, the Pauli principle keeps electrons of parallel spins apart from one another and hence lowers their mutual interaction energy. The larger the number of electrons with parallel spins, the larger the energy reduction. And concerning **2**, a large value of L allows the electrons to distribute themselves among a large number of orientations and thus to lower their energy even further.

Once the lowest term of a configuration is identified through Hund's rules, the ground state, which is the lowest multiplet level of that term, is determined by the spin–orbit interaction.

2.3 Spin–Orbit Interaction

An electron moving in an electric field $\mathbf{E}$ sees a magnetic field $\mathbf{H}_{orb} = \mathbf{v} \times \mathbf{E}/c$ proportional to the velocity $\mathbf{v}$. In the spherical potential of the atom, $\mathbf{E}$ is radial; hence $\mathbf{H}_{orb}$ is parallel to the orbital moment and there results a coupling between the spin and orbital moment of each electron i, given by

$$h_{so}(i) = \zeta_{nl} \mathbf{l}_i \cdot \mathbf{s}_i, \qquad (15)$$

where ζ_{nl} is the spin–orbit parameter of the shell (nl) obtained by integration over the square of the radial part of the corresponding wave function. The total spin–orbit interaction is $\mathcal{H}_{so} = \Sigma_i h_{so}(i)$.

By making use of the angular momentum commutation relations, it is shown that $\mathcal{H}_{so}$ commutes with **J** and that as a result of the Wigner–Eckart theorem it is equivalent, within the manifold of states belonging to the term (LS), to the operator

$$\mathcal{H}_{so} \rightarrow \Lambda \mathbf{L} \cdot \mathbf{S}. \tag{16}$$

The constant Λ is given by

$$\Lambda = \pm \zeta_{nl}/2S, \tag{17}$$

where the + and − signs apply respectively to configurations with less than half-filled and more than half-filled shells. In addition to the matrix elements of $\Lambda \mathbf{L} \cdot \mathbf{S}$, $\mathcal{H}_{so}$ has matrix elements between terms that differ in their (L,S) values by $\Delta L = \pm 1$, $\Delta S = \mp 1$, but this will not concern us here as their effect is usually small.

From Eq. (16), the multiplet energies with respect to the unsplit term energy are found to be

$$E_J = \tfrac{1}{2}\Lambda[J(J+1) - L(L+1) - S(S+1)]. \tag{18}$$

The energy difference between neighboring multiplets is thus proportional to the J value of the multiplet with the largest J,

$$\Delta E_J = E_{J+1} - E_J = \Lambda(J + l), \tag{19}$$

a relation that is known as the Landé interval rule.

It follows from Eqs. (17) and (18) that the ground multiplet state of a partly filled shell is $J = |L - S|$ or $J = L + S$ according to whether the shell is respectively less than or more than half-filled. Figure 1 shows the Hund's-rule ground term and the lowest multiplet for the different configurations d^n of a d shell. A similar figure, with maximum values $L = 6$ and $S = \frac{7}{2}$, would represent the $4f^n$ configurations of the rare earths (no confusion should arise between the superscript n and the principal quantum number n).

In the particular case of a shell that is half-filled, the ground state has $L = 0$, $S = J = (2l + 1)/2$, and the magnetic moment is entirely due to the spin with $g = 2$. A shell that is one electron short of being half-filled has a ground state $L = l$, $S = l$, and $J = 0$.

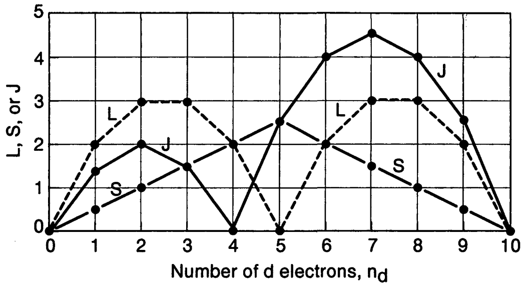

FIG. 1. Values of L, S, and J for the ground state of iron-group ions as the d shell is filled. Because of the change in sign of the spin-orbit constant Λ at $n_d = 5$, the values of J at a given L are higher for $n_d > 5$ than they are for $n_d < 5$. Only the values at integral n_d have physical meaning; the straight lines are a guide to the eye.

2.4 Landé *g* Factor

A magnetic field lifts the $(2J + 1)$-fold degeneracy of the multiplet states $|JJ_z\rangle$. Unlike the operator J_z, the magnetic moment $\beta(L_z + 2S_z)$ is not a constant of the motion, and in free ions it has matrix elements between multiplet states differing by $\Delta J = \pm 1$. However, the Zeeman interaction $\mathcal{H}_Z = \beta(\mathbf{L}_z + 2\mathbf{S}_z) \cdot \mathbf{H}$ is in practice much smaller than the spin–orbit interaction, and these off-diagonal matrix elements are usually unimportant. The Zeeman energies are then given by the expectation values of the magnetic moment over the $|JJ_z\rangle$ states,

$$\mathcal{H}_Z(JJ_z) = \langle JJ_z|\beta \mathbf{H} \cdot (\mathbf{L}_z + 2\mathbf{S}_z)|JJ_z\rangle \equiv g_J \beta H J_z, \tag{20}$$

where g_J is defined as the Landé g factor. The energy separation between consecutive levels is called the Zeeman splitting, equal in free atoms to $g_J \beta H$. The fact that these expectation values are proportional to J_z follows from the Wigner–Eckart theorem, which in systems having spherical symmetry requires the matrix elements of any vector operator (such as μ) over the states $|JJ_z\rangle$ to be proportional to those of the angular momentum itself. The coefficient of proportionality here is g_J.

The physical meaning of g_J can be seen in a semiclassical picture: The orbital and spin momenta, being strongly coupled by $\mathcal{H}_{so}$, precess around one another and around the direction of their resultant **J**. The magnetic moment precesses rapidly around **J** in the same manner, and since its components normal to **J** average to zero, it is only the component of $\boldsymbol{\mu}$ along **J** that couples to an external field. The term g_J is then obtained from the expectation value of this component in the direction of **J**,

$$g_J = \frac{\langle \boldsymbol{\mu} \cdot \mathbf{J} \rangle}{J(J+1)} \frac{1}{\beta}. \tag{21}$$

According to Eq. (21), g_J is the magnetomechanical ratio g' (as defined in Sec. 2.1) for the states of the (LSJ) multiplet. Explicit calculation of $\langle \boldsymbol{\mu} \cdot \mathbf{J} \rangle$ gives

$$g_J = \tfrac{3}{2} + \frac{S(S+1) - L(L+1)}{2J(J+1)}. \tag{22}$$

The proportionality factor in the Zeeman energies, Eq. (20), is traditionally called the spectroscopic splitting factor and denoted by g. From Eq. (20), g too is equal to g_J. The equality $g = g'$ is particular to spherical systems. For ions in solids, it will be seen in Sec. 4.4 that the two are in general different.

2.5 Magnetic Susceptibility

The magnetic susceptibility is defined as the derivative of the magnetization with respect to the applied magnetic field. At sufficiently low fields, the thermal average of the induced moment is linear in the field and the atomic susceptibility is a constant that can be calculated by expanding the energies of the low-lying states, n, to second order in the field. From Eq. (11),

$$E_n(H) = E_n^{(0)} - (\mu_z)_{nn}H + \sum_{n' \neq n} \frac{|(\mu_z)_{n'n}|^2}{E_n^{(0)} - E_{n'}^{(0)}} H^2, \tag{23}$$

where $E_n^{(0)}$ is the unperturbed energy and $\mu_z = -\beta(l_z + 2s_z)$. From Eq. (1) and the definition of the thermodynamic free energy, it follows that the paramagnetic moment is given by the negative of the derivative of the free energy with respect to H. The moments of atoms obey Boltzmann statistics, and hence for an assembly of N atoms per unit volume the magnetization $M(H,T)$ is given by

$$M(H,T) = -N\frac{\partial F(H,T)}{\partial H}, \tag{24}$$

where $F(H,T)$, the free energy per atom, is defined through

$$\exp\left[-\frac{F}{kT}\right] = \sum_n \exp\left[-\frac{E_n H}{kT}\right]. \tag{25}$$

The low-field ($\beta H \ll kT$) susceptibility follows from these relations and is given by

$$\begin{aligned}\chi(T) &= -N\left[\frac{\partial^2 F}{\partial H^2}\right]_{H\to 0} \\ &= N\left[\sum_n \left[\frac{(\mu_z)_{nn}^2}{kT} + 2\sum_{n'} \frac{|(\mu_z)_{n'n}|^2}{E_{n'}^{(0)} - E_n^{(0)}}\right] \cdot \exp\left(-\frac{E_n^{(0)}}{kT}\right)\right]\left[\sum_n \exp\left(-\frac{E_n^{(0)}}{kT}\right)\right]^{-1},\end{aligned} \tag{26}$$

where the states n and n' belong to the low-lying multiplets of the lowest term (LS).

The two terms in the numerator of Eq. (26) have different physical meanings: The first one is the contribution of a spontaneous moment that can orient itself in the external field. It exists only when n belongs to a multiplet having $J \neq 0$. The second one is a temperature-independent contribution of an induced moment known as Van Vleck paramagnetism (Van Vleck, 1932, Ch. 7). At low temperature, this term is important only when the ground multiplet has $J = 0$, but at temperatures comparable to the multiplet splitting* it makes a significant contribution even if $J \neq 0$. This is the reason for the discrepancies between the experimental and theoretical values of the effective magneton number in the cases of samarium and europium noted below (Table 1). In the induced moment, selection rules restrict the difference in the J values of the states n and n' to $\Delta J = \pm 1$.

A detailed account of the temperature variation of $\chi(T)$ is given by Van Vleck (1932, Ch. 9), and may be summarized as follows:

---

*Various units of energy are used in atomic physics, among them the kelvin (K), the wave number (cm^{-1}), and the electron volt (eV), which is equal to 1.6×10^{-19} J. In terms of the other units, 1 eV = 8065 cm^{-1} = 11 600 K.

Table 1. Effective magneton numbers p for trivalent rare-earth ions (near room temperature) [from Kittel (1986), p. 405].

Ion	Configuration	Basic level	p(calc.) $= g[J(J+1)]^{1/2}$	p(exp.), approximate
Ce^{3+}	$4f^1 5s^2 p^6$	$^2F_{5/2}$	2.54	2.4
Pr^{3+}	$4f^2 5s^2 p^6$	$^3H_4$	3.58	3.5
Nd^{3+}	$4f^3 5s^2 p^6$	$^4I_{9/2}$	3.62	3.5
Pm^{3+}	$4f^4 5s^2 p^6$	$^5I_4$	2.68	. . .
Sm^{3+}	$4f^5 5s^2 p^6$	$^6H_{5/2}$	0.84	1.5
Eu^{3+}	$4f^6 5s^2 p^6$	$^7F_0$	0	3.4
Gd^{3+}	$4f^7 5s^2 p^6$	$^8S_{7/2}$	7.94	8.0
Tb^{3+}	$4f^8 5s^2 p^6$	$^7F_6$	9.72	9.5
Dy^{3+}	$4f^9 5s^2 p^6$	$^6H_{15/2}$	10.63	10.6
Ho^{3+}	$4f^{10} 5s^2 p^6$	$^5I_8$	10.60	10.4
Er^{3+}	$4f^{11} 5s^2 p^6$	$^4I_{15/2}$	9.59	9.5
Tm^{3+}	$4f^{12} 5s^2 p^6$	$^3H_6$	7.57	7.3
Yb^{3+}	$4f^{13} 5s^2 p^6$	$^2F_{7/2}$	4.54	4.5

1. At low temperatures, $kT \ll \Delta E_J$ where ΔE_J is the energy separation of the ground multiplet J from the next higher one, only the states of the ground multiplet are populated. Usually $J \neq 0$, in which case the second term of Eq. (26), which reduces to a constant $N\alpha$, is negligible. The matrix elements $(\mu_z)_{nn}$ are equal to $-g_J J_z \beta H$, and summation over J_z gives the well-known Curie susceptibility,

$$\chi = N g_J^2 J(J+1)\beta^2/3kT. \tag{27}$$

The factor $p = g_J\sqrt{J(J+1)}$ is called the effective magneton number and provides a convenient way to tabulate the experimental data. Table I shows the values of p for trivalent rare-earth ($4f$) ions, deduced from room-temperature measurements, and for comparison the Hund's-rule values p(calc.). With the exception of Sm and Eu, the agreement between the two is very good even though the measurements were not made on free ions but on solid hydrated sulfates. As will be seen in Sec. 4.1, the reason for the good agreement is that rare-earth ions are little affected by their environment.

2. The discrepancy for Eu^{3+} arises because the Hund's-rule multiplet has $J = 0$ (the f shell of Eu^{3+} is one electron short of being half-filled). However the next level, $J = 1$, is separated in energy only by $\Delta E_J = 330$ K, which is comparable to room temperature. It turns out that the overall multiplet width is $21\Delta E_J$ and to obtain the susceptibility accurately, the Van Vleck part of Eq. (26) must be summed over all the multiplets. The result, which obviously depends on T, agrees with the measured value of p. The discrepancy in the case of Sm^{3+} has a similar origin: Even though the ground multiplet has $J = 5/2$, higher multiplets must be taken into account. These two cases illustrate the behavior of $\chi(T)$ when T is comparable with ΔE_J.

3. Finally when T is large compared to the overall width of the ground term multiplets, the expression of $\chi(T)$ become simple again. The Van Vleck terms cancel out in Eq. (26) since the Boltzmann factors effectively can be taken equal to unity, and L and S are decoupled and free to assume random orientations with respect to one other. The susceptibility is then the sum of independent orbital and spin contributions,

$$\chi = \frac{N\beta^2}{3kT}[L(L+1) + 4S(S+1)]. \tag{28}$$

2.6 Saturation

When the magnetic field is so large and the temperature so small that the ground-state Zeeman splitting is comparable to the temperature ($g_J\beta H \gtrsim kT$), the low-field expansion is not useful and the free energy must be calculated exactly. Replacing g_J by g in order to lighten the notation, Eq. (25) becomes

$$\exp\left(-\frac{F}{kT}\right) = \sum_{J_z=-J}^{+J} \exp\left(\frac{gJ_z\beta H}{kT}\right) \tag{29}$$

and the magnetization follows as

$$M(H,T) = NgJ\beta B_J(gJ\beta H/kT), \tag{30}$$

where in terms of $x = gJ\beta H/kT$, the function

$$B_J(x) = \frac{2J+1}{2J}\coth\frac{2J+1}{2J}x - \frac{1}{2J}\coth\frac{x}{2J} \tag{31}$$

is the Brillouin function, which approaches unity as $x \to \infty$. Thus, the magnetization saturates when the Zeeman energy is large compared to kT.

The classical limit of the Brillouin function is the Langevin function $L(x) = \coth x - 1/x$, which is obtained by letting the value of J approach infinity while $gJ\beta$ is kept constant and equal to the value of the classical moment μ. For large J, $B_J(x)$ approaches $L(x)$.

Figure 2 shows measurements of the magnetization in the saturation region, $x \gg 1$, for three ions having J values of $\frac{7}{2}$, $\frac{5}{2}$, and $\frac{3}{2}$ respectively. Gd^{3+} and Fe^{3+} have L = 0 in the ground state, and the moment is that of the spin alone, with $g = 2$. The ion Cr^{3+} has the configuration d^3 and as seen in Fig. 1, the free ion has $L = 3$, $S = \frac{3}{2}$, and $J = \frac{3}{2}$, resulting in a g value of $\frac{2}{5}$. However, the orbital moment is quenched in the solid (Sec. 4.1), and the magnetization is due entirely to the spin. The theoretical curve for Cr^{3+} in Fig. 2 is calculated with $g = 2$. Had $g = \frac{2}{5}$ been used instead, the curve would have strongly disagreed with the experimental points, thus providing confirmation of the quenching of the orbital moment.

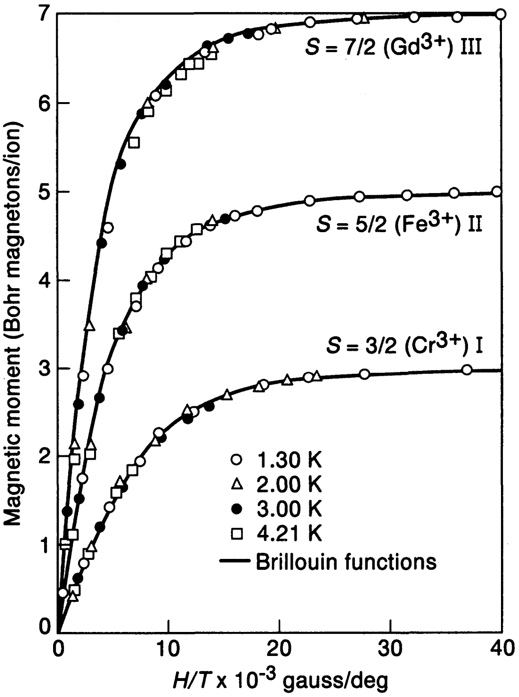

FIG. 2. Plot of magnetic moment vs H/T for (I) potassium chromium alum, (II) iron ammonium alum, and (III) gadolinium sulfate hexahydrate, taken at temperatures T = 1.3, 2.0, 3.0, and 4.21 K. The abscissa is proportional to the argument x of the Brillouin function in Eq. (31). Near-complete saturation is achieved at T = 1.3 K and H = 5 T (after Henry, 1952).

3. PARAMAGNETISM IN SOLIDS

The type of solid that forms when atoms come together depends on the valence electrons. Elements with open s or p shells form either metals, semiconductors (covalent solids), or molecular solids. In the latter two cases, strong bonds form between neighboring atoms, the electron spins pair off, the orbital moments are quenched as described below, and the solid is diamagnetic. Among paramagnetic solids are metals and salts of metallic ions having partially filled shells. The conduction electrons in metals and in doped semiconductors show a weak paramagnetism (Pauli paramagnetism) that is due predominantly to the spin. In addition, they have an orbital diamagnetism with susceptibility χ_L named after Landau (see DIAMAGNETISM) that in the free-electron model only is equal to minus one-third of the Pauli susceptibility. In real metals, χ_L depends on the band structure and the total susceptibility, including the filled-core diamagnetism and the Van Vleck paramagnetism of any partially filled bands, can be either positive or negative. Pauli paramagnetism is discussed in Sec. 5.

Strong, atomiclike paramagnetism occurs in the salts of the elements that have partially filled d shells (of the $3d$, $4d$, and $5d$ series) or f shells (the $4f$ rare earths and the $5f$ actinides). Most intensively studied have been the iron group ($3d$ series) and the $4f$ group, and these are the subject of this section. At high temperatures, the susceptibility obeys a Curie–Weiss law

$$\chi = N\,C/(T - \theta), \tag{32}$$

where the Curie constant C [see Eq. (27)] depends on the value of the effective Bohr magneton of the ion in the solid environment. In the rare-earth ions, according to Table 1, C is given by the full atomic moment, but Table 2 shows that in the ions of the iron group, C is given essentially by the spin-only moment, $2\beta S$: The orbital momentum is quenched (Sec. 4.1). The constant θ accounts for the exchange interactions between moments. In the mean-field approximation, these are represented by a molecular field proportional to the magnetization, $H_{\text{mol}} = \lambda M$, and the value that results for θ is $\theta = \lambda C$.

4. IONIC SOLIDS

The study of paramagnetic ions in solids is of interest because it reveals the changes that they undergo in a crystalline environment. These changes in turn provide information about the electronic state of the whole solid. Aside from its own interest, such a study provides opportunities for designing materials with properties suitable for applications, such as maser or laser materials. The susceptibility is a macroscopic, global quantity, which cannot by itself give much information about these effects. In the decades following World War II, the paramagnetism of solids was extensively studied by electron spin resonance (see ELECTRON PARAMAGNETIC RESONANCE), which measures directly the Zeeman splitting of the paramagnetic ground state. If the ions were free, this would of course be given by the Landé g factor, but in a solid, even the smallest interaction of the ion with the environment is often sufficient to change its Zeeman splittings radically.

The theoretical treatment of these interactions is known as "crystal-field" or "ligand-field" theory, and it is the subject of several excellent books. Discussions of the "spin Hamiltonian" and of the effective spin concept may be found in these. Only a brief account that emphasizes the physics is given here.

The ionic species under investigation is usually at the center of a cage of "ligands," which are negative ions such as F^- or O^{2-} ions belonging to a molecule such as H_2O. The ligands are often at the corners of an octahedron or a cube, and hence the symmetry of the central ion is cubic. Usually a small trigonal or tetragonal distortion is also present. Symmetry is very important in determining the g value in a spin-resonance [electron paramagnetic resonance (EPR)] experiment. The central ion together with the ligands forms a "complex" or complex ion that is a molecular entity that maintains its identity in a water solution. The effect of the charged ligands on the paramagnetic ion is represented by a local electric field (the crystal field) that has the site symmetry of the ion.

The notion of an electric field should not be taken too literally for it serves mostly as a convenient way to parametrize the interaction of the ion with the environment, including for instance the effect of the covalency with the ligands. The crystal-field interaction is written $\mathcal{H}_{\text{cf}}$, and it is added to the Hamiltonian [Eq. (14)]. The manner in and degree to which the states of the free ion are modified depend on the magnitude of $\mathcal{H}_{\text{cf}}$ relative to $\mathcal{H}_2$ and $\mathcal{H}_{\text{so}}$, and three cases can be distinguished: $\mathcal{H}_{\text{so}} > \mathcal{H}_{\text{cf}}$; $\mathcal{H}_2 > \mathcal{H}_{\text{cf}} > \mathcal{H}_{\text{so}}$; and $\mathcal{H}_{\text{cf}} > \mathcal{H}_2$.

Table 2. Effective magneton numbers for iron-group ions [from Kittel (1986), p. 406].

Ion	Configuration	Basic level	$p(\text{calc.}) = g[J(J+1)]^{1/2}$	$p(\text{calc.}) = 2[S(S+1)]^{1/2}$	$p(\text{exp})$[a]
Ti^{3+}, V^{4+}	$3d^1$	$^2D_{3/2}$	1.55	1.73	1.8
V^{3+}	$3d^2$	$^3F_2$	1.63	2.83	2.8
Cr^{3+}, V^{2+}	$3d^3$	$^4F_{3/2}$	0.77	3.87	3.8
Mn^{3+}, Cr^{2+}	$3d^4$	$^5D_0$	0	4.90	4.9
Fe^{3+}, Mn^{2+}	$3d^5$	$^6S_{5/2}$	5.92	5.92	5.9
Fe^{2+}	$3d^6$	$^5D_4$	6.70	4.90	5.4
Co^{2+}	$3d^7$	$^4F_{9/2}$	6.63	3.87	4.8
Ni^{2+}	$3d^8$	$^3F_4$	5.59	2.83	3.2
Cu^{2+}	$3d^9$	$^2D_{5/2}$	3.55	1.73	1.9

---

[a] Representative values.

1. $\mathcal{H}_{so} > \mathcal{H}_{cf}$ is the weak-field case, which occurs when the ligands have a minimal effect. In the rare-earth ions, the 4*f* shells lie inside the filled 5*s* and 5*p* shells, and as a result the inner 4*f* electrons are kept away from the ligands. As seen in Table 3, $\mathcal{H}_{cf}$ is small, of the order of 100 cm^{-1}, and its effect is limited to splitting the degenerate states of each multiplet and mixing their wave functions according to the local symmetry.
2. $\mathcal{H}_2 > \mathcal{H}_{cf} > \mathcal{H}_{so}$ is the intermediate crystal-field case, which occurs in many salts of the 3*d* group. The spatial extent of the 3*d* orbitals is comparable to that of the 4*f* in the rare earths, but there are no outer 5*s* or 5*p* shells to screen them from the ligands. The ground state is still given by Hund's rule, but the orbital levels are split by the crystal field and the orbital moment is partly or totally quenched. The term *J* is not a quantum number.
3. $\mathcal{H}_{cf} > \mathcal{H}_2$ is the strong crystal-field case, which occurs in some salts of the transition groups. The perturbation by the environment is so strong that the atomic one-electron orbitals (*nl*) have their $(2l + 1)$-fold degeneracy lifted, and the many-electron states must be expressed in terms of these split states. Hund's rule does not necessarily hold, and in fact, in octahedral symmetry the configurations d^4–d^7 do not obey it. Such ions are said to be in the low-spin state, $S = 1$ for d^4, $S = \frac{1}{2}$ for d^5, etc. When the ligands are so close that their wave functions overlap substantially those of the paramagnetic ion, covalent bonding takes place and the magnetic electrons become delocalized into molecular orbitals that are shared with the ligands. The orbital moments and the spin–orbit coupling are then affected through "orbital reduction" factors as the *d* orbitals now constitute only part of the wave function. Covalency strongly enhances the crystal-field splittings, which are as high as 30 000 cm^{-1} in some cyanides.

4.1 Quenching of the Orbital Moment

The $(2L + 1)$-fold degeneracy of the orbital states $|LL_z\rangle$ of a free ion reflects the freedom of the orbital momentum to assume any orientation (classically speaking) in its spherically symmetric environment. In a crystal the internal electric field has lower symmetry and is not radial. The orbit of an electron moving classically in such a field will deviate from being planar and precess in space in some complicated way. The angular momentum will fluctuate in time and either will average to zero or, if the symmetry is high, will have a nonzero average value only for certain orbits. This picture corresponds quantum mechanically to a mixing of the states having different values of L_z. The mixing of states strongly affects the expectation values of the Zeeman energies, and hence of the spectroscopic splitting factor. If the orbital degeneracy is completely lifted, the wave function must be real and the expectation value of the orbital momentum, which is a purely imaginary operator, must vanish. The orbital momentum is then completely quenched. The foregoing discussion has completely ignored the spin–orbit coupling, which must now be taken into account.

The spin–orbit coupling can be viewed as an internal magnetic field generated by the spin moment and acting on the orbital motion. Through the interaction $\zeta \mathbf{s} \cdot \mathbf{l}$, this field is coupled to the orbit of the electron *around* the nucleus, and in this way it opposes the quenching of $\mathbf{l}$. Whether quenching occurs or not depends therefore on the relative magnitudes of $\mathcal{H}_{cf}$ and $\mathcal{H}_{so}$, which are measured by the ratio of the crystal-field splitting Δ of the ground multiplet to the spin–orbit parameter ζ_{4f} or ζ_{3d}.

In the rare-earth ions, Table 3 shows that $\Delta \ll \zeta_{4f}$. The orbital moment is not quenched, but the crystal field mixes the $|JJ_z\rangle$ states of the ground multiplet and splits the multiplet energy. The angular momentum *J* remains as an approximate quantum number for clas-

Table 3. Ranges of the energies that determine the nature of the ground electronic state of a paramagnetic rare-earth (4*f*) ion, or an iron-group (3*d*) ion in a solid environment. E_c is the ground-term splitting, ζ_{nl} is the spin–orbit parameter, and Δ is the ground-state crystal-field splitting.

	E_c (cm^{-1})	ζ_{nl} (cm^{-1})	Δ (cm^{-1})
4*f* ions	$(5–30) \times 10^3$	600–3000	~100
3*d* ions	$(10–40) \times 10^3$	60–800	$(10–30) \times 10^3$

sifying the levels, and the ratio of magnetic to mechanical moment, g', is still given by the Landé g_J factor. However, the spectroscopic splitting factor is no longer equal to g_J but depends on the admixture coefficients of the $|JJ_z\rangle$ states in the crystal-field states. At temperatures T smaller than Δ, the susceptibility deviates from the Curie value both because of the splitting of the states and because of the difference of g from g_J. For T large compared to the overall Δ, the susceptibility should return to the free-ion value with $p = g_J\sqrt{J(J+1)}$ as the effect of the splittings averages out. This is borne out by the experimental results in Table 1, which show that the room-temperature susceptibility of rare-earth ions is not affected by their environment.

The ions of the iron group are usually at the opposite extreme, $\Delta \gg \zeta_{3d}$. The orbital momentum is then quenched and neither J nor g_J retains any validity. There is only a slight deviation of the g factor from $g_s = 2$ that is due to an admixture of orbital moment by the spin–orbit interaction. Table 2 shows that with the exception of Co^{2+}, the measured values of p are close to the spin-only value.

Although affecting the Curie susceptibility only slightly through the deviation of g from 2, the orbital moment contributes now a Van Vleck susceptibility χ_{VV}, which arises from the splitting of the d states by the crystal field and the fact that the orbital moment has matrix elements between the occupied and unoccupied crystal-field states. Being proportional to the product of the number of occupied and number of unoccupied states, χ_{VV} tends to be largest for half-occupied shells. In ionic salts, χ_{VV} is expected to be only a very small fraction of the Curie susceptibility, but in the transition metals the d electrons become delocalized into moderately broad energy bands and the spin susceptibility χ_s is reduced by a factor of order 10^2 compared to the Curie susceptibility (see Sec. 5). The term χ_{VV} becomes then comparable with χ_s. Indirect experimental evidence and an estimate of χ_{VV} came first in Knight-shift and susceptibility measurements in the intermetallics V_3Ga and V_3Si. More direct evidence was obtained later by a measurement of the g' factor in the transition metals themselves (see Sec. 4.4).

4.2 Kramers's Theorem

There is a limit to the removal of degeneracies by a crystalline field. A free ion has degeneracy $2J + 1$, which is even if the number of electrons is odd and, vice-versa, odd if the number of electrons is even. The ion may be a complex ion. Kramers's theorem states that in the presence of an arbitrary electric field and in zero magnetic field, all states of a system having an *odd* number of electrons must still have even degeneracy. The theorem is a consequence of the time-reversal invariance of the Hamiltonian including the spin–orbit interaction. The theorem is proved by showing that when the number of electrons is odd, the operation of time reversal applied to an eigenstate produces another eigenstate, which is orthogonal to the first and hence must be independent of it. The number of states that are degenerate with each other must therefore be even.

Ions with an *even* number of electrons, by contrast, may have their *spin* degeneracy lifted by an electric field of low symmetry. This occurs for instance in the eight-electron Ni^{2+} ion, which has an orbitally nondegenerate ground state with a spin of 1. The threefold spin degeneracy is completely lifted by a rhombic crystalline field acting in conjunction with the spin-orbit coupling (Griffiths and Owen, 1952).

In solids, the ground states of $3d$ and $4f$ ions with an odd number of electrons are usually Kramers doublets.

4.3 Jahn–Teller Effect

In the previous parts of this section, it has been tacitly assumed that paramagnetic ions situated at sites of high symmetry may have a ground state that is orbitally degenerate. By contrast, if the site is without any symmetry, then barring accidental degeneracy, all the states must be orbitally nondegenerate. Jahn and Teller (1937) asked the question, is the stability of a high-symmetry environment compatible with an orbitally degenerate electronic ground state? The Jahn–Teller (J–T) theorem states that excepting the case when the paramagnetic ion is part of a linear molecule (which does not concern us here), stability and orbital degeneracy are not compatible. When the spin is included, the theorem holds for the degeneracy including spin, except for the Kramers degeneracy, which cannot be removed by purely electrical forces. We shall not consider the spin any further here except to say that sometimes the spin–orbit coupling may stabilize the symmetric configuration. A simple system where the J–T theorem applies is illustrated in the following example, taken from their paper.

A planar molecule consists of a *diamagnetic* ion A located at point P and surrounded by four identical ligands B_1, B_2, B_3, B_4 situated symmetrically at a distance a [configuration C_0 in Fig. 3(a)]. This configuration is stable. Ion A is now replaced by ion M, which in the symmetric configuration has an orbitally degenerate ground state: For instance, M has one p electron outside of closed shells, and the orbitals p_x and p_y with wave functions proportional to x and y, respectively, are degenerate as a result of the fourfold local symmetry. The J–T theorem asserts that the configuration C_0 is now unstable with respect to the distortions C_1 and C_2 of Fig. 3(b). The physical reason for the instability is simply that the two degenerate states point in mutually orthogonal directions and a distortion of the symmetric configuration such as C_1 or C_2 will lower the energy of one of the states while raising that of the other, as shown in Fig. 3(c). An elementary calculation that assigns one electron charge, $-e$, to the ligands shows that if ions B_1 and B_3 are moved outward by a distance d while B_2 and B_4 are moved inward by the same distance (configuration C_1), the additional potential seen by an electron situated at the point (x,y,z) with respect to P is, in first order in d.

$$\Delta V(x,y) = -\tfrac{9}{2}(e^2/a^4)(x^2 - y^2)d. \tag{33}$$

Averaging Eq. (33) over the wave functions of p_x and p_y shows that the energy of p_x is lowered by

$$\Delta E(p_x) = -Vd \tag{34}$$

where the quantity V is positive, while the energy of p_y is raised by the same amount. The important fact is that Eq. (34) is linear in the nuclear displacement d. A nuclear configuration cannot be stable if the electronic energy of a neighboring configuration depends linearly on the nuclear displacement since the elastic energy $U = \frac{1}{2}Kd^2$, which provides the restoring force for displacements away from a position of stable equilibrium, is quadratic in the displacement. Minimizing the sum of the energy shift [Eq. (34)] and U, and setting $K = M\omega^2$ where M is the nuclear mass and ω the vibration frequency, the equilibrium value of d is found to be $d_{\text{eq}} = V/M\omega^2$, and the net lowering of the energy, or stabilization energy, is obtained as the negative of

$$E_{\text{J-T}} = V^2/2M\omega^2. \tag{35}$$

Note that the J–T theorem does not say how the system (the molecule) will respond to the instability, but only that the symmetric configuration is unstable. The simplest answer that comes to mind is that the nuclear configuration will distort in the manner suggested above until d assumes the value d_{eq} and the energy is lowered to its minimum, $-E_{\text{J-T}}$. Such a response of the system is called the "static J–T effect" as it assumes implicitly that the nuclei have infinite mass: Once at their position of minimum energy, they stay there forever. The consequence of such a response in an ionic solid would be the removal of all orbital degeneracies, and the

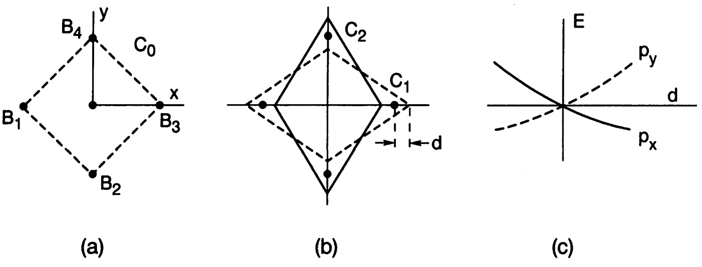

FIG. 3. The Jahn–Teller effect of a *p*-state ion in a symmetric planar molecule: (a) In the undistorted configuration C_0, the ion has fourfold symmetry and the states p_x and p_y are degenerate. The configuration C_0 is unstable against distortion. (b) Two equivalent distortions, C_1 and C_2. The symmetry of the central ion is lowered and the degeneracy is removed. (c) For small distortions, the splitting of the energies is linear in the amplitude *d* of the distortion.

lowering of the site symmetry should be observable as an anisotropic line (or g value) in an EPR experiment. In fact, the first observation of the J–T effect in a solid, made by EPR at room temperature (Bleaney and Ingram, 1950), showed a g value that was isotropic and different from $g_s = 2$.

The experiment was done on the d^9, Cu^{2+} ion at the center of an octahedral complex where, in the absence of any distortion, it would have had the twofold orbitally degenerate Γ_3 level as its ground state. A distorted octahedron, elongated along one of the [100] cubic axes, should have shown a resonance characteristic of tetragonal symmetry with $g_{\parallel} \neq g_{\perp}$. The observed isotropic g value was inconsistent with this, but it was also inconsistent with an undistorted configuration that, because of the twofold degeneracy, should have shown two resonance lines, split by the spin–orbit interaction, and having cubic but not spherical symmetry with $g_{111} \neq g_{100}$. The observed spectrum was explained by Abragam and Pryce (1950) as a manifestation of the "dynamic J–T effect," this terminology having been introduced later. The labeling of the effect as "dynamic" here means that because of their finite mass, the nuclei are able to move over an energy barrier from one distorted configuration to another under the action of the thermal vibrations of the surrounding lattice. At room temperature, these thermal excitations are sufficient to cause the activated transitions to occur at a rate that is large compared to the resonance frequency, and the three resonance lines that would have resulted from tetragonal distortions along any one of the three equivalent [100] cubic axes are motionally narrowed to a single isotropic line. Two years later (Bleaney and Bowers, 1952), the same system at 20 K showed in fact the three anisotropic lines of the "static J–T effect," indicating that at 20 K the transition rate between configurations was smaller than the EPR frequency. The same kind of distortion may thus appear as dynamic or static depending on the time resolution of the measuring apparatus as compared with the transition rate of the distortion between the different configurations.

The condition for the static J–T effect to occur is that the stabilization energy be large compared to the zero-point energy of the nuclei, $E_{J-T} \gg \hbar\omega$. This is referred to as the "strong J–T" condition. When this is not the case and $E_{J-T} \leq \hbar\omega$ ("weak J–T effect"), it is necessary to include the kinetic energy of the nuclear vibrations from the start in solving the problem, i.e., to consider the coupled electronic-vibrational system as a single dynamical system. The Hamiltonian of this system is called the vibronic Hamiltonian, and its states the vibronic states. The full dynamical nature of the J–T effect becomes apparent in the properties of these vibronic states.

It is clear that the vibronic Hamiltonian, inclusive of the J–T coupling, is invariant under the symmetry operations of the undistorted complex since these operations merely replace one distorted configuration by another. This can be seen in Fig. 3(b), where a rotation of a distorted configuration, accompanied by a rotation of the two electronic states, leaves things unchanged. It follows from this invariance that the J–T coupling *cannot lift* a degeneracy that is due to high symmetry, but only *replace* a purely electronic degeneracy by a vibronic degeneracy of the same character, i.e., the vibronic and the original electronic states must transform in the same way under the symmetry operations of the undistorted system. Physically, the electronic states and the nuclear vibrations have become strongly correlated.

Having the same twofold Γ_3 degeneracy as the unperturbed electronic states, the vibronic states should show an EPR spectrum having the same features as those of the electronic doublet described above. Such a spectrum was observed at very low temperatures, independently by Coffman (1965) and by Höchli and Estle (1967), on systems similar to those investigated by Bleaney and Ingram but having a weaker J–T coupling, $E_{J-T} \leq \hbar\omega$. Two anisotropic lines with $g_{111} \neq g_{100}$, characteristic of a Γ_3 level, were observed, but the anisotropy was smaller than that predicted by the known value of the spin–orbit constant. This difference was experimental proof of the presence of a vibronic doublet because such orbital reduction factors are predicted as a consequence of the smaller than unity overlap between the vibrational states associated with different electronic states. From the degree of the reduction, an estimate of the Jahn–Teller coupling could be obtained.

In principle, the static susceptibility should be affected too since the orbital reduction factors cause a change in the g values.

The experimental manifestations of the

dynamic J–T effect are diverse and fascinating. The interested reader (see Further Reading) may consult an excellent account by Sturge that includes physical discussions of the theory and the experimental results. More formal theoretical treatments can be found in the treatise by Abragam and Bleaney and in the review article of a key contributor to the theory, F. Ham. This section owes much to these sources.

4.4 *g* and *g′*

The magnetomechanical ratio in the z direction of a paramagnetic ion in a solid is defined as the ratio of the expectation values:

$$g' = -\frac{\langle \mu_z \rangle/\beta}{\langle J_z \rangle/\hbar}, \tag{36}$$

where μ_z and J_z are the z components of the magnetic moment and angular momentum, respectively. The negative sign reflects the sign of the electron charge so that g' as defined by Eq. (36) is positive. The spectroscopic splitting factor is a measure of the Zeeman splitting in a spin-resonance experiment, and for an effective spin of $\frac{1}{2}$ is defined by

$$g\beta H = \hbar\omega, \tag{37}$$

where ω is positive. For a free spin, $g = g_s = 2.0023$. The sign of g, which can be positive or negative, reflects the sense of precession of the moment around **H** in the solid (Blume *et al.*, 1969).

Early measurements of g and g' in the ferromagnetic 3*d* metals showed that they are not the same. Although both were found to have values close to $g_s = 2$, thus reflecting strong orbital quenching, they differed from 2 in opposite directions with $g > 2$ and $g' < 2$ (Kittel, 1949). The measured g values were close to those deduced from magnetic susceptibility measurements on the salts of the metals.

The first measurement of g' in a paramagnetic salt was made by Sucksmith (1930) on Dy_2O_3. Within experimental error, the result was $g' = g_j = \frac{4}{3}$, implying negligible orbital quenching in agreement with our remarks in Sec. 4.1.

The fact that $g \neq g'$ shows that the torque equation

$$\frac{d\langle \mathbf{J} \rangle}{dt} = \langle \boldsymbol{\mu} \rangle \times \boldsymbol{H} \tag{38}$$

does not hold in solids because it would imply $g = g'$. Thus from Eq. (38), $d\langle \mathbf{J} \rangle/dt$ is proportional to $\omega\langle \mathbf{J} \rangle$ and hence to $g\langle \mathbf{J} \rangle$, while $\langle \boldsymbol{\mu} \rangle$, from Eq. (36) is proportional to $g'\langle \mathbf{J} \rangle$.

The reason Eq. (38) does not hold is that through the interaction with the crystalline field, an exchange of angular momentum between the lattice and the moment is taking place as the latter precesses around the magnetic field (in the case of strong orbital quenching, this exchange is mediated by the spin–orbit interaction). Hence, part of the torque $\langle \boldsymbol{\mu} \rangle \times \mathbf{H}$ that is transferred by the microwave field to the angular momentum of the electron escapes into the lattice (or alternatively an additional torque is exerted by the lattice), and Eq. (38) does not hold. The situation is very similar (Kittel, 1949) to that of an electron in a lattice that is acted on by an external electric field: In free space, an electron of mass m acted on by a force **F** has an acceleration $\langle \dot{\mathbf{v}} \rangle = \mathbf{F}/m$, while in a crystal lattice, as a result of momentum transfer between the electron and the lattice, the acceleration is given by $\langle \dot{\mathbf{v}} \rangle = \mathbf{F}/m^*$, where m^* is the effective mass, here assumed isotropic. In this analogy, the g factor corresponds to m^* and g' corresponds to m.

From Eqs. (36) and (37), g' and g can be expressed as

$$g' = \frac{\langle L_z \rangle + 2\langle S_z \rangle}{\langle L_z \rangle + \langle S_z \rangle}, \tag{39}$$

$$g = \frac{\langle L_z \rangle + 2\langle S_z \rangle}{1/2}, \tag{39'}$$

where $\langle L_z \rangle$ and $\langle S_z \rangle$ are in units of $\hbar$.

In the case of 3*d* ions, strong orbital quenching gives $\langle S_z \rangle = \frac{1}{2}$ and $\langle L_z \rangle \ll \langle S_z \rangle$. Both g and g' are then close to 2, and Eqs. (39) and (39′) show that they satisfy the relation

$$g' = g/(g-1). \tag{40}$$

This is the Kittel–Van Vleck relation, and it is in fair agreement with experiment. Equation (40) has been generalized by Blume *et al.*

(1969) to cases where the g factor can be large and of either sign as for instance in the cases of some rare-earth ions and of the current carriers in semiconductors with very small band gaps.

As mentioned in Sec. 4.1, the 3*d* metals are expected to have a Van Vleck orbital susceptibility χ_{VV} (for which $g' = 1$) that is of the same order of magnitude as the spin susceptibility χ_s for which $g' = 2$. The actual value of g' can be measured in a gyromagnetic experiment and the relative magnitudes of χ_{VV} and χ_s can thereby be inferred. Such experiments, which are very delicate in paramagnetic metals as a result of the smallness of the χ's, were performed by Hughenin *et al.* (1971) on several transition metals. In V and Cr, the measured g' is close to 1.2, indicating the dominance of the Van Vleck term in these metals, while in Pd and Pt, which are at the end of the 4*d* and 5*d* series, the values of g' are about 1.7. The spin susceptibility is strongly exchange-enhanced in Pd and Pt and χ_{VV} is smaller than in V or Cr because the *d* shells are nearly full. As a result, the spin susceptibility is dominant.

5. CONDUCTION-ELECTRON PARAMAGNETISM

In metallic crystals, the valence states of neighboring atoms overlap strongly and the wave functions become delocalized. Ignoring at first the difficult problem of exchange and correlation, the delocalized, or conduction, electrons are assumed to move independently in an average potential that has the periodicity of the lattice. The valence states are broadened into energy bands of "Bloch states" $\Psi_{\mathbf{k}}$, which are modulated plane waves that depend on the wave vector $\mathbf{k}$, and the conduction electrons occupy these states subject to the Pauli principle that allows a maximum of two electrons of opposite spins to occupy each Bloch state. At $T = 0$, these are filled up to a maximum energy E_{F}, the Fermi energy, which ranges in different metals from about 3 to 15 eV, a value considerably larger than the room-temperature energy of $kT = 0.025$ eV. (No confusion between Boltzmann's constant k and $\mathbf{k}$ should arise.)

5.1 Pauli Susceptibility

One of the early successes of the theory of metals was the explanation by Pauli (1927) of their temperature-independent paramagnetism. The Langevin theory predicts a susceptibility $\mu^2/3kT$ per atom, which at the typical atomic densities of metals results in a volume susceptibility at room temperature of order 10^{-3}–10^{-4}. Instead, the observed susceptibilities were temperature independent and of order 10^{-6}. The explanation, well known now, is that the Pauli principle prevents the electrons that are well below the Fermi surface from reorienting their spin. Only those electrons having energies within a range kT of the Fermi surface are free to do so, contributing each a susceptibility $\mu^2/3kT$. For N conduction electrons per unit volume, the paramagnetic susceptibility becomes

$$\chi_{\mathrm{P}}^{(0)} \sim N\frac{kT}{E_{\mathrm{F}}}\frac{\mu^2}{3kT} \sim N\frac{\mu^2}{3E_{\mathrm{F}}}. \tag{41}$$

This is independent of T and smaller than the Curie value by a factor 10^2, as experimentally observed. For a free-electron gas, the actual calculated value is $\chi_{\mathrm{P}}^{(0)} = 3N\beta^2/2E_{\mathrm{F}}$, and for a general band structure having a density of states of one spin direction $n(E)$ per unit volume and unit energy range, the calculated value at zero temperature is

$$\chi_{\mathrm{P}}^{(0)} = \tfrac{1}{2}g^2\beta^2 n(E_{\mathrm{F}}), \tag{42}$$

where $g = 2$ for a free spin and $n(E)$ is evaluated at the Fermi energy. The orbital moment contributes to the Pauli susceptibility only through the spin–orbit coupling $\mathcal{H}_{\mathrm{so}}$. With the exception of points in $\mathbf{k}$ space at which Bloch states may be degenerate, the expectation value of $\mathcal{H}_{\mathrm{so}}$ over a Bloch state vanishes, and only admixtures of states from other bands by $\mathcal{H}_{\mathrm{so}}$ may give an orbital contribution to the g factor. These are usually small and shift the g value by a quantity Δg that ranges from 10^{-3} in Na to 0.2 in the heavy metals (Yafet, 1963). The reason for the smallness is that the interband energies at constant $\mathbf{k}$ are usually large compared to $\mathcal{H}_{\mathrm{so}}$. The g values at the Fermi surface, $g_{\mathbf{k}}$, depend on $\mathbf{k}$, but in a spin-resonance experiment the EPR line is narrowed by motional and exchange averaging to a width that is below the inhomogeneous width associated with the distribution of g values.

When one goes beyond the independent-electron approximation and includes the Coulomb interactions, the susceptibility can be calculated only approximately. Just as exchange interactions in free atoms are responsible for Hund's rule of maximum spin, here exchange interactions (somewhat tempered by correlations) between the conduction electrons tend to maximize the total number of electron pairs with parallel spins when a magnetic field is present. As a result, the susceptibility is enhanced over the noninteracting value $\chi_P^{(0)}$, and in the simplest approximation, the interacting susceptibility χ_P is given by

$$\chi_P = \chi_P^{(0)}/(1 - \nu). \tag{43}$$

The constant ν is the product of an effective exchange energy per electron and the density of states at the Fermi level normalized to one electron. If ν reaches the value 1, the system becomes ferromagnetically unstable. Paramagnetic metals have $\nu < 1$.

Since $\chi_P^{(0)}$ can be obtained from the calculated band structure, it is of interest to measure χ_P directly in order to compare the actual enhancement factor with the theoretical one.

5.2 Measurement of χ_P

The contribution of χ_P to the total χ was isolated by Schumacher and Slichter (1956) through the use of magnetic resonance. The spin paramagnetism χ_P is related to the area under the resonance absorption signal plotted as a function of the frequency. Precise measurements of the absolute value of such areas are very difficult to make. The problem was circumvented by observing the resonances of both electrons and nuclei on the same sample, changing only the magnetic field and leaving the frequency and all other parameters unchanged. The ratio of the areas under the two absorption curves, which involves only a relative measurement that can be done accurately, gives the ratio between the electron and the nuclear susceptibilities, and since the latter is simply the nuclear Curie susceptibility, χ_P may be deduced.

5.3 Localized Moments and Kondo Effect

When dissolved in nonmagnetic metals (Cu, Zn, Au), many elements of the 3*d* transition series show a Curie susceptibility, which indicates that they have a free localized moment. This is not totally unexpected as Hund's rule indicates that if the *d* electrons remain on the atom (as they must in order to keep it neutral) then they might well possess a moment. Friedel (1958) and Anderson (1961) developed the idea of the virtual bound state according to which the *d* electrons can transfer back and forth between the localized state and the conduction states of the host metal. This transfer broadens the localized level, and the higher the rate at which it occurs (which depends on the density of states), the larger the broadening. If the level becomes too broad, the two spin directions become equally populated in order to minimize the total band energy and the impurity atom becomes nonmagnetic. If the level remains fairly narrow, then Hund's rule wins, the energy of the unoccupied spin state rises above the Fermi level, and the impurity atom acquires a local moment.

The moment is coupled by exchange to the conduction electrons, and the exchange integral can be shown to be negative, which favors an antiparallel arrangement of the local spin with respect to the spins of the conduction electrons. Anomalous low-temperature behavior in the resistivity of normal metals with small amounts of magnetic impurities, and its interpretation by Kondo, led to the discovery that this exchange coupling causes the formation of a collective (with the conduction electrons) bound state that is a spin singlet. This Kondo state, as it is known, is at an energy below the Fermi level so that at zero temperature it is fully occupied. Being a singlet, it possesses no magnetic moment, and thus the local moment is compensated collectively by the spin moments of the conduction electrons. The binding energy depends exponentially on the reciprocal of the exchange integral and may thus vary over a broad range. The magnetic susceptibility of the impurity ion, which is Curie-type, $\chi = C/T$ at high temperature, saturates to a constant value $\chi = C/T_K$ as T approaches zero. The term T_K is a measure of the singlet binding energy.

5.4 Nonlocal Susceptibility and RKKY Interaction

It might be expected that if it were possible to apply a local magnetic field extending only within the space occupied by one atom in a metal, the polarization of the conduction electrons would extend up to some distance away because of the delocalized nature of their orbitals. Such a magnetic field is effectively realized by the exchange coupling between a local moment and the conduction electrons. The response of the electron gas to such a field is conveniently analyzed in terms of the wave-vector–dependent susceptibility: The local field is decomposed into Fourier components $H_{\mathbf{q}} \exp(i\mathbf{q}\cdot\mathbf{r})$. Each $H_{\mathbf{q}}$ induces a magnetic-moment density of the same wave vector, proportional to $M_{\mathbf{q}}^{(0)} = \chi_q^{(0)}$ where for convenience the noninteracting susceptibility is used. The dependence of $\chi_q^{(0)}$ on $\mathbf{q}$ is primarily determined by the shape of the Fermi surface. At wave vectors $\mathbf{q} = \mathbf{q}_c$ that span *locally parallel* areas of the Fermi surface (FS), $\chi_{q_c}^{(0)}$ has singularities that depend in detail on the local shape of the FS (Roth, 1966).

These singularities determine the dependence of the magnetization $\mathbf{M}(\mathbf{r})$ at large distances r. It is found that

1. $\mathbf{M}(\mathbf{r})$ is oscillatory in space, with a wave vector related to the spanning vector appropriate to the direction of $\mathbf{r}$; and
2. the amplitude of the oscillations decreases with distance as $1/r^3$ provided the FS has no unusual features at the spanning points.

For a gas of free electrons, the FS is a sphere and the oscillations are isotropic with a wave vector that is twice the Fermi wave vector, $2k_F$ (Ruderman and Kittel, 1954).

The long-range polarization $\mathbf{M}(\mathbf{r})$ causes an exchange interaction to occur between any two localized spin moments, such as those of the nuclear spins of the constituent atoms, or those of any localized electronic moments. This interaction is named the Ruderman–Kittel–Kasuya–Yosida (RKKY) interaction. It is responsible for the magnetic helical orderings of the core $4f$ moments in the heavy rare-earth metals (see Further Reading).

5.5 Spin-Density Waves

The interacting wave-vector–dependent susceptibilities have an enhancement factor similar to that of the Pauli susceptibility in Eq. (43), i.e.,

$$\chi_q = \chi_{\mathbf{q}}^{(0)}/(1 - \Omega v_c \chi_{\mathbf{q}}^{(0)}). \tag{44}$$

If it should happen that the denominator vanishes at some $\mathbf{q}_n$, then χ_{q_n} diverges, which means that the electron gas acquires a spontaneous long-range magnetization having spatial dependence $\exp(i\mathbf{q}_n\cdot\mathbf{r})$ as the spins order themselves in an antiferromagnetic helix. Such a configuration occurs as the ground state in metallic Cr below the Néel temperature $T_N = 311$ K (see Further Reading).

5.6 Large *g* Factors in Semiconductors

In the semiconductor InSb, the energy gap between conduction and valence band occurs at the center of the zone, $\mathbf{k} = 0$, and the magnitude of the gap is small compared to the spin–orbit splitting of the valence band. The magnitude of the g factor of the electrons at the bottom of the conduction band was predicted (Roth *et al.*, 1959) to be very large, $g \approx -50$, even though the wave function at this value of $\mathbf{k}$ is practically of pure s atomic character. Such large g factors were shown by Herring (1963) to arise from interatomic orbital currents circulating over large numbers of atomic sites by virtue of the smallness of the effective mass. Even larger g factors have been observed in the semimetal Bi.

6. APPLICATIONS

6.1 Magnetic Cooling

An early application of paramagnetism was the attainment of very low temperatures by the method of adiabatic demagnetization, proposed independently by Debye (1926) and Giauque (1927). This method makes use of the fact that a paramagnetic salt at a low temperature T has only two kinds of degrees of freedom, namely those of the vibrational motion of the nuclei that constitute the lattice, and those of the orientational dependence of the electronic magnetic moments of the ions (the nuclear moments are left out here). In the absence of a magnetic field, the entropy of the disordered, noninteracting moments is very large as each moment contributes its full orientational entropy $k \ln(2J + 1)$. A magnetic field H such that $g\beta H/kT \gg 1$ lowers this entropy dramatically as the moments become oriented. By contrast, the entropy of the lattice is very small at small T, only of order $k(T/\Theta)^3$ per degree of freedom, where the Debye temperature Θ is of order 400 K. The principle of the method is as follows. Starting from an initial $T_i \sim 1$ K that is reached by pumping on liquid He, the salt is magnetized isothermally so that its temperature remains at T_i. The initial field H_i is of order 1–5 T. The thermal contact with the outside is then broken, and the magnetic field is adiabatically and reversibly reduced to a very low value, $H_f \sim 10^{-3}H_i$.

From Eq. (29), the free energy and hence the entropy of the magnetic moment (or "spin") system depends only on the ratio $g\beta H/k_BT$. Hence, at constant entropy, the final temperature of the spin system is $T_f^{(s)} = (H_f/H_i)T_i^{(s)}$. The final temperature of the salt is attained after the lattice comes into thermal equilibrium with the spins by transfer of entropy between them. But since the entropy of the lattice is so much smaller than that of the spins, this transfer is negligible and the final temperature is simply $T_f \approx (H_f/H_i)T_i$. The limit of attainable values of T_f is determined by the temperature T_c at which the weak interactions (dipolar and exchange) between the spins lead to the formation of a cooperative magnetic state. Hence the choice of a suitable salt is guided by a compromise between minimizing these interactions and maintaining a sufficient density of magnetic ions. This is done by choosing suitable hydrated salts, and temperatures of order 10^{-3} K are thus reached. Lower temperatures, of order 10^{-6} K, are obtained by demagnetizing the nuclei of Cu.

6.2 The Maser

The maser (acronym for Microwave Amplification by Stimulated Emission of Radiation) is basically a microwave amplifier with a very low noise figure. It can be implemented by different physical systems, among them a paramagnetic salt having three low-lying spin levels A, B, C in order of increasing energy (Bloembergen, 1956). With the frequency differences denoted by ω_{AB}, ω_{BC}, and ω_{AC}, a microwave field of frequency ω_{AC} (the "pump" frequency) is used to induce transitions A → C, thereby equalizing the populations of these two levels. The signal that one wishes to amplify has frequency ω_{BC}. Application of this signal to the device induces a net rate of transitions C → B as long as the population of level C is kept higher than that of level B by the pump. The emitted radiation is in phase with and proportional to the amplitude of the applied signal, and hence the device acts as an amplifier. If the material is chosen such that the relaxation rates between the levels are favorable, the device can also operate as an oscillator. The noise figure is low because the maser must of necessity operate at a low temperature. The maser is thus particularly useful for amplifying weak signals that come from outer space and have themselves a low noise figure.

6.3 Tunable Spin-Flip Raman Laser

An infrared laser in the frequency range just below the energy gap ($E_G = 0.23$ eV) that separates the conduction and valence bands in the semiconductor InSb has been realized by inducing electronic Raman transitions in the conduction band (Patel and Shaw, 1971). A large magnetic field is used to quantize the valence- and conduction-band states into sharp Landau levels, and transitions between the two spin states of the lowest Landau level in the conduction band are induced by scattering incident light at a frequency ω_0 slightly below the gap energy.

The scattering cross section for spin reversal is large for two reasons:

1. The spin–orbit coupling is large compared to the band gap, which results in the Raman matrix element for spin reversal being of the same order as or larger than the matrix element for orbital transitions, which is proportional to the magnetic field (Yafet, 1966); and
2. the photon energy $h\omega_0$ is chosen to be nearly in resonance with the gap energy, which enhances the cross section further.

The laser operates at the frequency of the scattered light, $\omega_1 = \omega_0 - |g|\beta H/\hbar$, and since $|g|$ is of the order of 50, the frequency ω_1 can be tuned in a limited range by varying H. This flexibility allows the laser to be used as an analytical tool.

7. SUMMARY

A material is paramagnetic if it contains magnetic moments that are free to orient in an external magnetic field. Electron paramagnetism has been an invaluable probe, mainly through the use of paramagnetic resonance, for exploring the magnetic and electronic structure of solids. By this method, detailed studies of the low-lying states of paramagnetic ions in solids and of the relaxation processes of the magnetic energy have been made; subtle phenomena such as the Jahn–Teller effect have been elucidated.

The thermodynamic properties of spin systems are special in that the total number of states, and hence the entropy, are finite. In addition, the entropy can easily be manipulated by an external magnetic field. This flexibility has been used to reach very low temperatures, and also to achieve very high temperatures (or population inversion) by selective absorption of radiation. Spin systems are even used as model systems in the teaching of thermodynamics.

8. ACKNOWLEDGMENTS

The author is indebted to Prof. Sir Roger Elliott for making available his unpublished lecture notes on magnetism, and to Prof. Michael Sturge for a critical reading of the manuscript.

GLOSSARY

Boltzmann Factor: The population factor of a state in an ensemble that represents thermal equilibrium.

Complex: A compound ion that consists of a central metal ion surrounded by a cage of ligands and is sufficiently stable to retain its identity in a water solution.

Crystal Field: Electric field set up by the ligands that reflects the symmetry of the complex.

Debye Temperature: The temperature above which all acoustic modes of the lattice vibrations are appreciably excited and below which modes begin to freeze out.

Effective Mass: The apparent inertial mass, m^*, of conduction electrons in metals and semiconductors. Under the effect of an external force F, the acceleration of a wave packet in a spherical band is $\dot{v} = F/m^*$.

Effective Spin: If a paramagnetic ion in a solid has a level with degeneracy $2S' + 1$, the effective spin of that level is S'. The effective spin is defined in a more general way in the spin-Hamiltonian formalism.

Energy Band: A set of itinerant one-electron states in metals or semiconductors whose energies vary continuously within an energy range as the wave vector of the states ranges over the reduced Brillouin zone.

EPR: Electron paramagnetic resonance.

Fermi Surface: The surface of constant energy in $\mathbf{k}$ space that corresponds to the Fermi energy.

***g* Factor:** The proportionality coefficient that measures the splitting of a degenerate level by a magnetic field. In free atoms, g is equal to the Landé g factor.

Knight Shift: The shift of the nuclear resonance frequency in a metal, produced by the polarization of the conduction electrons in the applied magnetic field.

Ligand: A molecule or ion attached to the central ion of a complex (or complex ion).

Line of Force: A line formed by connecting neighboring points in space that are situated along the direction of the local magnetic field with respect to one another.

Magnetic Susceptibility: A measure of the response of a substance to an applied magnetic field. More precisely, the derivative of the magnetization with respect to the applied field.

Magnetization: The magnetic moment per unit volume of a piece of matter.

Paramagnet: A substance that contains elementary magnetic moments that tend to orient themselves along an applied field but that do not exhibit a macroscopic moment or any kind of magnetic ordering in the absence of a field.

Polarization of Light: The specification of the direction of the electric field of a light wave. In plane-polarized light, this direction stays the same at all points of space, at all times, while in circularly polarized light the direction of the field at each point in space rotates uniformly with time.

Van Vleck Susceptibility: The paramagnetic susceptibility that arises from off-diagonal matrix elements of the orbital magnetic moment.

Vibronic State: In a system that undergoes the Jahn–Teller effect, a state in which the electronic state of the paramagnetic ion and the state of the surrounding nuclei are strongly coupled. The total wave function is no longer a simple product of an electronic function and a function of the nuclear displacements, but a sum of such products.

Zeeman Energy: The energy of a spontaneous magnetic moment in a magnetic field.

Zeeman Splitting: The energy splitting of a multiplet level or of any degeneracy connected with the existence of a spontaneous moment by an external magnetic field.

Works Cited

Abragam, A., Pryce, M. H. L. (1950), *Proc. Phys. Soc. A* **63**, 409–411.

Anderson, P. W. (1961), *Phys. Rev.* **124**, 41–53.

Bleaney, B., Bowers, K. D. (1952), *Proc. Phys. Soc. A* **65**, 667–668.

Bleaney, B., Ingram, D. J. E. (1950), *Proc. Phys. Soc. A* **63**, 408–409.

Bloembergen, N. (1956), *Phys. Rev.* **104**, 324–327.

Blume, M., Geschwind, S., Yafet, Y. (1969), *Phys. Rev.* **181**, 478–487.

Coffman, R. E. (1965), *Phys. Lett.* **19**, 475–476.

Curie, P. (1905), *Ann. Chim. Phys. 7ème Ser.* **5**, 289.

Debye, P. (1926), *Ann. Phys.* **81**, 1154.

Faraday, M. (1846a), *Philos. Trans. R. Soc.* **136**, 21–40, articles 2279, 2280. Also in (1855) *Experimental Researches in Electricity*, Vol. 3, Series 20, London: Taylor & Francis.

Faraday, M. (1846b), *Philos. Trans. R. Soc.* **136**, 41–62, article 2420. Also in (1855) *Experimental Researches in Electricity*, Vol. 3, Series 21, London: Taylor & Francis.

Faraday, M. (1846c), *Philos. Trans. R. Soc.* **136**, 1–20, article 2152. Also in (1855) *Experimental Researches in Electricity*, Vol. 3, Series 19, London: Taylor & Francis.

Friedel, J. (1958), *Suppl. Nuovo Cimento* **7**, 287–311.

Giauque, W. F. (1927), *J. Am. Chem. Soc.* **49**, 1864–1870.

Griffiths, J. H. E., Owen, J. (1952), *Proc. R. Soc. A* **213**, 459–473.

Henry, W. E. (1952), *Phys. Rev.* **88**, 559–562.

Herring, C. (1963), in Yafet (1963), pp. 93–96, Appendix B.

Höchli, V. T., Estle, T. L. (1967), *Phys. Rev. Lett.* **18**, 128–130.

Hughenin, R., Pells, G. P., Baldock, D. N. (1971), *J. Phys. F: Met. Phys.* **1**, 281–289.

Jahn, H. A., Teller, E. (1937), *Proc. R. Soc. A* **161**, 220–235.

Kittel, C. (1949), *Phys. Rev.* **76**, 743–748.

Langevin, P. (1905), *J. Phys.* **24**, 678; *Ann. Chim. Phys. 8ème Ser.* **5**, 70.

Patel, C. K. N., Shaw, E. (1971), *Phys. Rev. B* 1279–1295.

Pauli, W. (1927), *Z. Phys.* **41**, 81.

Roth, L. M., Lax, B., Zwerdling, S. (1959), *Phys. Rev.* **114**, 90–104.

Roth, L. M. (1966), *Phys. Rev.* **149**, 519–525.

Ruderman, M. A., Kittel, C. (1954), *Phys. Rev.* **96**, 99–102.

Schumacher, R. T., Slichter, C.P. (1956), *Phys. Rev.* **101**, 58–65.

Sucksmith, W. (1930), *Proc. R. Soc. A* **128**, 276–293.

Thomson, W. (1850), *Philos. Mag.* **37**, 250.

Uhlenbeck, G. E., Goudsmit, S., (1925), *Naturwissenschaft 13*, 953–954; (1926) *Nature* **117**, 264–265.

Van Vleck, J. H. (1932), *The Theory of Electric and Magnetic Susceptibilities*, London: Oxford University.

Yafet, Y. (1963), in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, Vol. 14, New York: Academic, pp. 1–98.

Yafet, Y. (1966), *Phys. Rev.* **152**, 858–862.

Zeeman, P. (1896), *Proc. R. Acad. Amsterdam* **5**, 181; also (1897) *Philos. Mag.* **43**, 226–239.

Further Reading

Historical

Vershuur, G. L. (1993), *Hidden Attraction (The History and Mystery of Magnetism)*, New York: Oxford University Press.

General

Ashcroft, N. W., Mermin, N. D. (1976), *Solid State Physics*, New York: Saunders College, Chaps. 31 and 32.

Kittel, C. (1986), *Introduction to Solid State Physics*, 6th ed., New York: Wiley, Chaps. 14 and 16.

White, R. M. (1983), *Quantum Theory of Magnetism*, 2nd ed., Berlin: Springer, Chaps. 1–4 and 7.

Paramagnetism of Ions in Solids

Abragam, A., Bleaney, B. (1970), *Electron Paramagnetic Resonance of Transition Ions*, Oxford: Clarendon Press.

Ham, F. S. (1972), "Jahn-Teller Effect in Electron Paramagnetic Resonance Spectra," in: S. Geschwind (Ed.), *Electron Paramagnetic Resonance*, New York: Plenum, pp. 1–119.

Martin, D. H. (1967), *Magnetism in Solids*, Cambridge, MA: MIT Press.

Owen, J., Thornley, J. H. M. (1966), "Covalent Bonding and Magnetic Properties of Transition Metal Ions," *Rep. Prog. Phys.* **29**, 675–730.

Sturge, M. (1967), "The Jahn-Teller Effect in Solids," in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, Vol. 20, New York: Academic, pp. 92–211.

Taylor, K. N. R., Darby, M. I. (1972), *Physics of Rare Earth Solids*, London: Chapman and Hall.

Paramagnetism of Metals

Fawcett, E. (1988), "Spin-Density-Wave Antiferromagnetism in Chromium," *Rev. Mod. Phys.* **60**, 209–283.

Freeman, A. J. (1972), "Energy Band Structure, Indirect Exchange Interactions and Magnetic Ordering," in: R. J. Elliott (Ed.), *Magnetic Properties of Rare Earth Metals*, London: Plenum, pp. 245–333.

Heeger, A. J. (1969), "Localized Moments and Non-moments in Metals: the Kondo Effect," in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, Vol. 23, New York: Academic, pp. 283–411.

Applications

Hudson, R. P. (1972), *Principles and Application of Magnetic Cooling*, Amsterdam: North-Holland.

Schulz-Du Bois, E. O. (1960), "The Three-Level Solid State Maser," in: K. Mendelsohn K (Ed.), *Progress in Cryogenics*, Vol. 2, New York: Academic, pp. 175–231.

PARTICLE ACCELERATORS

See ACCELERATORS, LINEAR; ACCELERATORS, POTENTIAL DROP; BETATRONS; CYCLOTRONS; SYNCHROTRONS

PARTICLE BEAMS

See INTENSE PARTICLE BEAMS

19
Particle Colliders

James W. Rohlf

19.1 Introduction 587
19.1.1 Brief History 587
19.1.2 The Birth of Particle Colliders 587
19.1.3 Ingredients of Colliders 589
19.1.4 Famous Colliders 591
19.1.4.1 Stanford Positron Electron Accelerating Ring 591
19.1.4.2 PETRA 592
19.1.4.3 CERN Proton–Antiproton Collider 594
19.1.4.4 Fermilab Tevatron Collider 597
19.1.4.5 Large Electron–Positron Project 599
19.1.5 The Energy Frontier: CERN Large Hadron Collider 599
19.1.5.1 LHC Basic Parameters 599
19.1.6 Beam Dump 602
19.1.7 LHC Layout 605
19.1.8 Cryogenics 607
19.1.8.1 Injection Scheme 607
19.1.9 Radio Frequency Acceleration 611
19.1.10 Beam Vacuum 613
19.1.11 Super LHC 615
19.1.12 The Next Frontier 615
19.1.12.1 International Linear Collider 615
19.1.12.2 Compact Linear Collider 615
19.1.12.3 Very Large Hadron Collider 615
References 617
Further reading 618

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

19.1
Introduction

19.1.1
Brief History

The first particle accelerator consisted of two parallel plates with voltage applied, which was used by J. J. Thomson in 1897 to accelerate and detect the passage of charged particles (Thomson, 1897). Thomson ingeniously applied a magnetic field that allowed him to determine the charge-to-mass ratio of the mysterious ''cathode rays,'' which when combined with Millikan's measurement of the fundamental charge gave the first measurement of the electron mass. In 1929, John Cockcroft and Ernest Walton managed to accelerate protons to 800 keV, providing enough energy to produce the first artificial disintegration of the nucleus. The more sophisticated accelerators that followed (cyclotrons, linacs, and synchrotrons) were used to discover and measure the properties of a hoard of particles while pushing the frontier of physics to higher and higher energies. During this exciting period, M. Stanley Livingston pointed out that the technology to accelerate particles to higher and higher energies was developing exponentially in time, producing a factor of 10 increase in the highest particle energy achieved, every 7 years or so (Livingston, 1969).

19.1.2
The Birth of Particle Colliders

The prophesy of exponential growth by Livingston (Figure 19.1) was realized with the invention of colliding beam machines, which have the advantage of producing a center-of-mass energy equal to the sum of the energies of the colliding particles, in contrast to a single beam colliding with a stationary target, which produces a center-of-mass energy proportional to the square root of the beam energy. For example, it would take a 36-TeV electron hitting a stationary electron to have the same center-of-mass energy as 3-GeV electrons in a collider. The first colliding beam machines were driven by the physics of quantum electrodynamics (QED), as it was speculated that a breakdown of QED may be observable in electron–electron scattering. The first particle collider was the Princeton–Stanford accelerator at Stanford, CA (Figure 19.2). The project was colead by Gerard K. O'Niell of Princeton and Wolfgang Panofsky of Stanford. The energy was 550 MeV per beam and first physics results were recorded in 1963. Another electron–electron collider, VEP-1 in Novosibirsk, Siberia, followed quickly

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

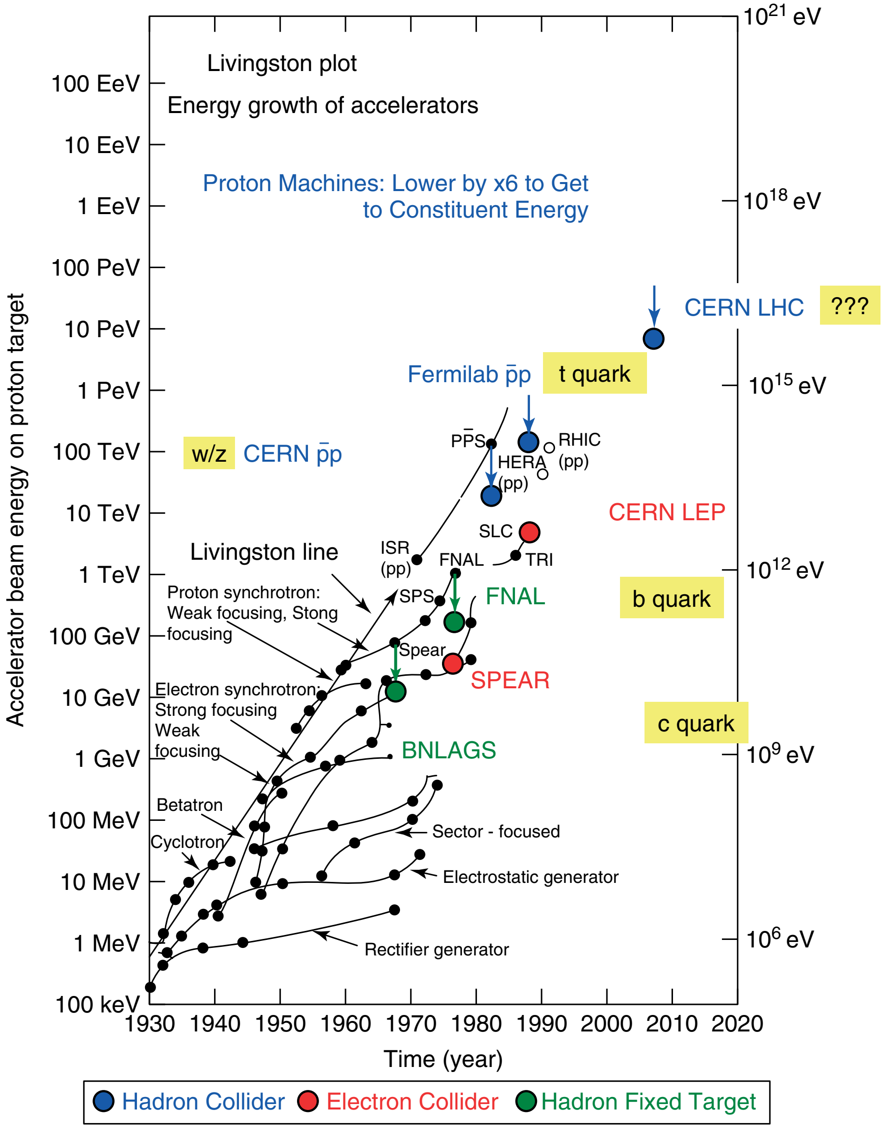

Fig. 19.1 Livingston plot of accelerator energy (fixed target equivalent) versus year of operation. The leading edge of technology has given a 10-fold increase in energy every seven years. (Adapted from ref. Panoksky 1997.)

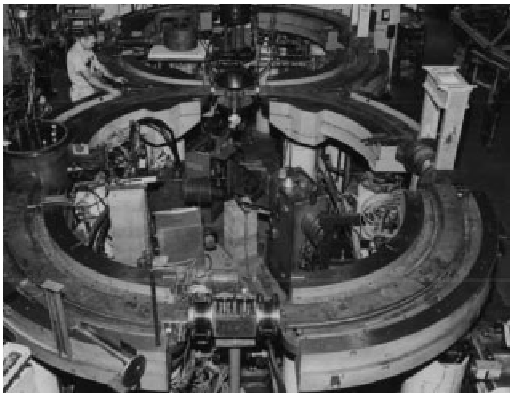

Fig. 19.2 The first particle collider, the Princeton–Stanford accelerator, was a double ring of 550-MeV electrons (Panoksky, 1997).

in 1966 with an energy of 700 MeV per beam. The beam instabilities discovered at these first colliders were overcome in future designs, allowing an increase in luminosity of nearly 2 orders of magnitude (Panoksky, 1997). The development of colliders through the solution of these beam instability problems became an international effort with huge contributions made by the groups led by B. Richter (SLAC), G. I. Budker (Novosibirsk), and F. Amman (Frascati) (Richter, 1976).

It was quickly realized that new dynamics could be explored with an electron–positron (e^+e^-) collider. Demonstration of the technique was made in Frascati, Italy, with the construction of the ADA e^+e^- collider (Bernadini, 2004). This machine was soon followed by other larger e^+e^- machines in Cambridge (MA, USA), Orsay (France), Novosibirsk (Siberia), and DESY (Germany).

19.1.3
Ingredients of Colliders

All particle colliders need an injection system, a ring (or two) of magnets to store the particles, a mechanism for accelerating the particles and keeping them in stable orbit, and a system to focus the particles at one or more collision points. In modern colliders, the storage ring is a synchrotron. The figure of merit for colliders is the luminosity (L), or number of times per square centimeter per second that the colliding particles meet. All physics rates are proportional to L, which may be written as

$$L = f\frac{n_1 n_2}{4\pi\sigma_x\sigma_y} \tag{1}$$

where f is the collision frequency of bunches containing n_1 and n_2 particles, and σ_x and σ_y are the transverse widths of the Gaussian beams. Thus, to be useful, a particle collider needs a large number of well-focused particles colliding at high frequency. The transverse size of the beams can be written in terms of a transverse emittance (ϵ) and an amplitude function (β) (Yao *et al.*, 2006). A particle collider is designed to collide large numbers of particles with low ϵ at places with low β. For electrons, the emittance is an equilibrium condition of quantum fluctuations in radiation with synchrotron radiation damping. For protons, the emittance is dominated by forces exerted during acceleration and storage.

Synchrotron radiation is emitted by relativistic particles traveling in a circular orbit. The power emitted may be written as

$$P = \frac{2ke^2\gamma^4 c}{3r^2} \tag{2}$$

where k is the electric force constant, e is the electric charge ($ke^2 = 1.44\,\mathrm{eV}\cdot\mathrm{nm}$), γ is the Lorentz-gamma factor, c is the speed of light, and r is the radius of the orbit. For electrons, the energy loss per turn around the ring (W) is

$$W = 8.85 \times 10^{-5}\,\frac{E^4/(\mathrm{GeV}^4)}{r/(10^3\,\mathrm{m})}\,\mathrm{MeV} \tag{3}$$

The radiation spectrum is broad but falls rapidly above a critical energy ($3\hbar c\gamma^3/2r$), which is typically in the hard X-ray region. Synchrotron radiation is the reason that high-energy e^+e^- colliders have a large radius. For protons, with their much larger mass, the energy loss per turn is

$$W = 7.8 \times 10^{-3}\,\frac{E^4/(\mathrm{TeV}^4)}{r/(10^3\,\mathrm{m})}\,\mathrm{keV} \tag{4}$$

Synchrotron radiation will be important for the first time in a proton collider at the CERN Large Hadron Collider (LHC).

Table 19.1 summarizes the many colliders that have been built in the last 40 years. The particle colliders fall into two general classes: e^+e^- and hadron machines, which are either proton–proton (pp) or proton–antiproton ($\bar{p}p$). There is but one exception, Hadron-Electron Ring Accelerator (HERA), which is an electron–proton (ep) machine. The energy of hadron colliders is limited by high-field magnet technology, rather than synchroton radiation,

Tab. 19.1 Summary of high-luminosity particle colliders.

Collider	Location	Beams	Energy (GeV)	Year	Luminosity ($\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$)
ISR	CERN	pp	31 × 31	1971–1984	1.4×10^{32}
SPEAR	Stanford	e^+e^-	4 × 4	1972–1990	10^{31}
DORIS	DESY	e^+e^-	5.6 × 5.6	1973–1993	3.3×10^{31}
PETRA	DESY	e^+e^-	23.4 × 23.4	1978–1986	2.4×10^{31}
CESR	Cornell	e^+e^-	6 × 6	1979–2002	1.3×10^{33}
PEP	Stanford	e^+e^-	40 × 40	1980–1990	4×10^{31}
SPS	CERN	$\bar{p}p$	315 × 315	1981–1990	6×10^{30}
TRISTAN	Japan	e^+e^-	15 × 15	1987–1995	6×10^{31}
Tevatron	Fermilab	$\bar{p}p$	980 × 980	1987 to present	1.7×10^{32}
SLC	Stanford	e^+e^-	50 × 50	1989–1998	2.5×10^{31}
LEP	CERN	e^+e^-	104 × 104	1989–2000	10^{32}
BEPC	China	e^+e^-	2.2 × 2.2	1989–2005	1.3×10^{31}
BEPC-II	China	e^+e^-	2.1 × 2.1	1989–2005	10^{33}
DAΦNE	Frascati	e^+e^-	0.7 × 0.7	1989–2008	1.5×10^{32}
HERA	DESY	ep	30 × 920	1992 to present	7.5×10^{31}
VEPP-4 M	Novosibirsk	e^+e^-	6 × 6	1994 to present	2×10^{31}
PEP-II	Stanford	e^+e^-	9 × 3.1	1999 to present	10^{34}
RHIC	Brookhaven	AuAu	100/n × 100/n	2000 to present	1.5×10^{27}
CESR-C	Cornell	e^+e^-	6 × 6	2002 to present	6×10^{31}
KEKB	Japan	e^+e^-	8 × 3.5	2002 to present	1.6×10^{34}
VEPP-2000	Novosibirsk	e^+e^-	1 × 1	2006 to present	10^{32}
LHC	CERN	pp	7000 × 7000	2008 start	10^{34} design

and they tend to be viewed as our "discovery" machines. This has been proved true in the discovery of the W and Z^0 bosons at the CERN Super Proton Synchrotron (SPS) collider and the top quark at the Fermilab Tevatron collider. The reason for this is partly the composite nature of the proton, resulting in a wide variation in center-of-mass energy of the colliding quarks. Thus, a hadron collider cannot be tuned to a specific energy. The e^+e^-, on the other hand, tend to be machines that map out, in great detail, the decays of particles of known mass. This has been true for the strange, charm, and b-quarks as well as the Z^0 boson.

19.1.4
Famous Colliders

A large number of interesting physics results are from the many particle colliders (Table 19.1) as each order of magnitude increase in sensitivity (luminosity), coupled with more sophisticated detectors, reveals even rarer decay modes of known particles. These important results include the observation of high transverse-momentum hadrons in pp collisions (CERN Intersecting Storage Rings (ISR)), detailed studies of strange, charm and b quarks, as well as the tau lepton (e^+e^- colliders), measurements of the proton structure function at high momentum transfer (HERA), and heavy-ion collisions (Relativistic Heavy-Ion Collider (RHIC)). A few discoveries at particle colliders, however, stand out for their particularly important impact on the advancement of the field of high-energy physics and our understanding of the standard model of particle physics. These discoveries are listed in Table 19.2.

Tab. 19.2 Summary of famous discoveries at particle colliders.

Collider	Beams	Discovery	Year
SPEAR	e^+e^-	Charm	1974
PETRA	e^+e^-	Gluon	1979
CERN SPS	$\bar{p}p$	W,Z^0	1983
LEP	e^+e^-	Three neutrinos	1989
Tevatron	$\bar{p}p$	Top	1995

Tab. 19.3 Basic parameters of the LHC.

Circumference	26,659 m
Dipole field	8.33 T
Energy per beam	7 TeV
Number of dipoles	1232
Total number of magnets	9994
Operating temperature	1.9 K
Power consumption	120 MW
Revolution frequency	11.2455 kHz
Revolution period	88.9 μs
Radio frequency (RF)	400.8 MHz
RF electric field	5.5 MV/m

19.1.4.1 Stanford Positron Electron Accelerating Ring

The first particle collider to produce exciting physics results was the Stanford Positron Electron Accelerating Ring (SPEAR), a project led by Richter (1971). The SPEAR design originally called for two asymmmetric rings, but evolved throughout the 1960s becoming ultimately a single ring of magnets about 80 m in diameter, which circulated electrons and positrons at energies up to 4 GeV. The first collisions were recorded in 1972. The codiscovery of charm at SPEAR in 1974, a period still referred to as the *November Revolution*, dramatically changed the field of high-energy physics, providing new insight into the dynamics of both the electroweak and strong interactions. The Schrödinger equation was applied to the

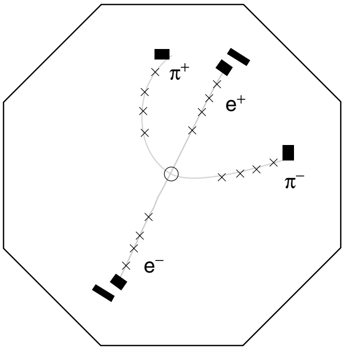

Fig. 19.3 Observation of the decay $\psi' \rightarrow \pi^+\pi^-\psi$, with $\psi \rightarrow e^+e^-$ in the Mark-I detector at SPEAR (Feldman and Perl, 1975). Tne event is the signature of bound states of a new heavy quark (charm) and its antiquark.

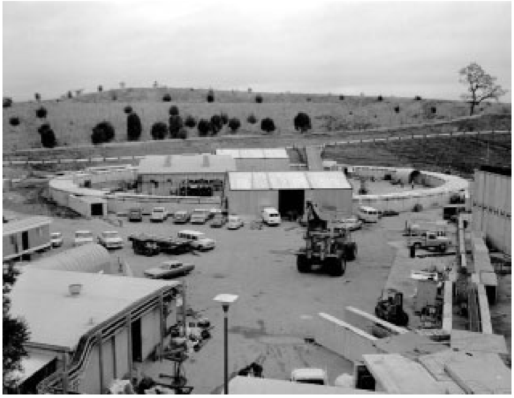

Fig. 19.4 Construction of the SPEAR machine nearing completion in 1971 (Stanford Linear Accelerator Center, 2002).

measured spectra of charmonium providing evidence beyond any reasonable doubt for the existence of quarks in the context of quantum chromodynamics (QCD). Figure 19.3 shows an event recorded at a center-of-mass energy of 3.7 GeV. In addition to charm, a third lepton family (tau) was discovered at SPEAR in 1975.

The SPEAR machine (Figures 19.4–19.5) was located at the end of the 3-km long Stanford Linear Accelerator Center (SLAC) linac, which was used to inject single 5-cm long bunches of positrons first and then electrons. The filling process took 15–30 min. The electrons and positrons collided at midpoint of two straight sections on opposite sides of the machine where quadrupole magnets were used to focus the beams to a transverse size of 1 mm × 100 μm. The data-taking periods lasted about 2 hours in a 5×10^{-9} torr vacuum before the positrons and electrons needed refilling. The synchrotron losses were about 300 kW at the maximum beam energies of 4 GeV (Richter, 1976). The muon components of Mark-I detector used to make the initial physics discoveries at SPEAR are shown in Figure 19.6.

19.1.4.2 **PETRA**

The Positron–Electron Tandem Ring Accelerator (PETRA) at DESY, along with Positron–Electron Project (PEP) at Stanford, was built with strong motivation to search for the top quark, but sadly, the top mass was far beyond the energy reach of these machines. The PETRA machine is basically a higher energy version of SPEAR. In 1979, the experiments at PETRA

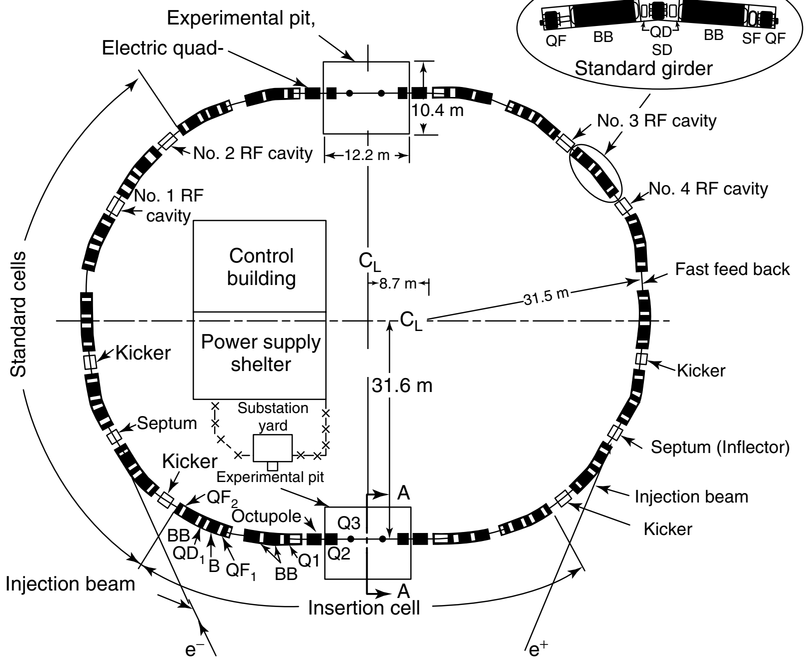
Experimental pit,
Electric quad-
QF
BB
QD
SD
BB
SF QF
Standard girder
No. 3 RF cavity
10.4 m
No. 2 RF cavity
12.2 m
No. 1 RF
cavity
No. 4 RF cavity
Control
building
C_L
8.7 m
Fast feed back
31.5 m
C_L
Standard cells
Kicker
Power supply
shelter
31.6 m
Kicker
Substation
yard
Septum
Septum (Inflector)
Kicker
Experimental pit
Injection beam
QF_2
A
Octupole
Q3
Kicker
BB
QD_1
B
QF_1
BB
Q1
Q2
Injection beam
A
Insertion cell
e^-
e^+

Fig. 19.5 Layout of SPEAR (Richter, 1976).

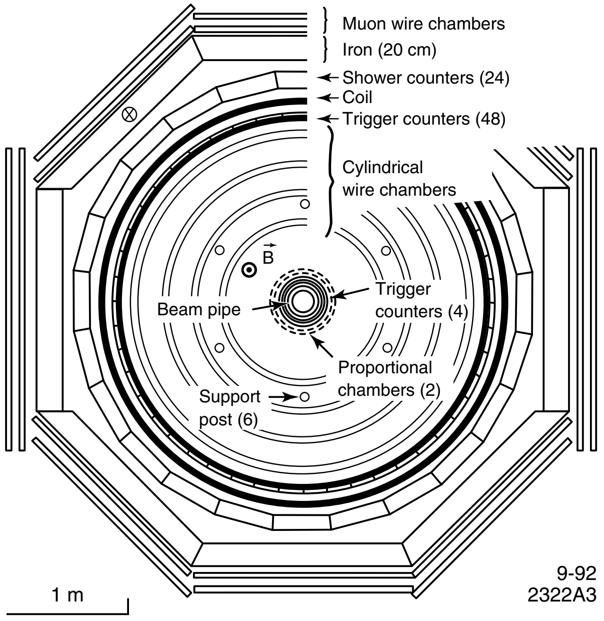
Muon wire chambers
Iron (20 cm)
Shower counters (24)
Coil
Trigger counters (48)
Cylindrical
wire chambers
$\vec{B}$
Trigger
counters (4)
Beam pipe
Proportional
chambers (2)
Support
post (6)
1 m
9-92
2322A3

Fig. 19.6 Components of the Mark-I detector at SPEAR (Perl, 1995).

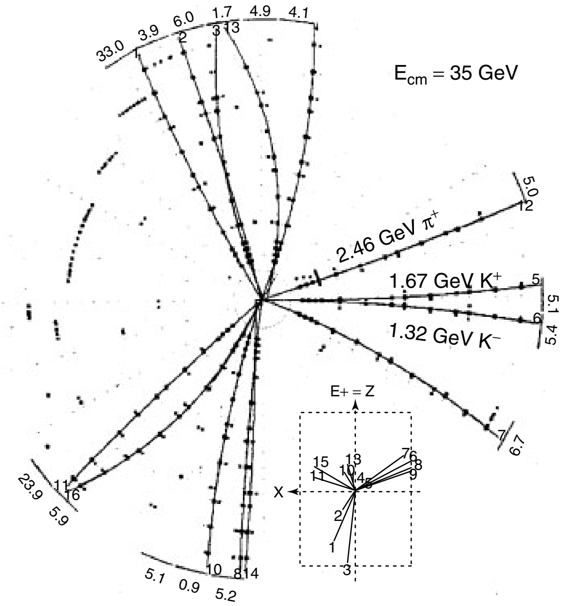

Fig. 19.7 A three-jet event from PETRA signaling gluon radiation.

(TASSO, JADE, MARK-J, and PLUTO) found direct evidence for the existence of the gluon (g) in the form of three-jet events caused by the process $e^+e^- \rightarrow q\bar{q}g$. A three-jet event from TASSO is shown in Figure 19.7.

19.1.4.3 **CERN Proton–Antiproton Collider**

In the late 1970s, an international race was on to observe the weak interaction gauge particles, the W and Z^0 bosons, proposed to exist as free particles at energies of about 100 GeV. The fastest route to such energies was conversion of existing proton synchrotrons into proton–antiproton colliders. The key technical hurdle was that of producing and cooling (reducing the momentum spread) a sufficient number of antiprotons to obtain useful luminosity for the collisions (Rubbia, 1985). This problem was solved by Simon Van der Meer, who applied stochastic cooling to the antiprotons (Van Der Meer, 1985). A proton–antiproton collision showing the decay of a Z^0 boson into e^+e^- is seen in Figure 19.8.

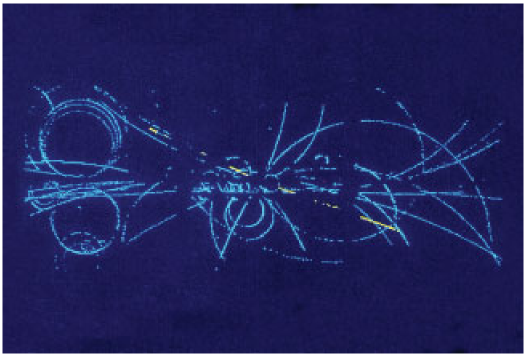

Fig. 19.8 High-energy proton–antiproton collision containing the decay $Z^0 \rightarrow e^+e^-$ (straight tracks) in the UA1 detector at CERN proton–antiproton collider. The events signified the unification of electromagnetic and weak forces.

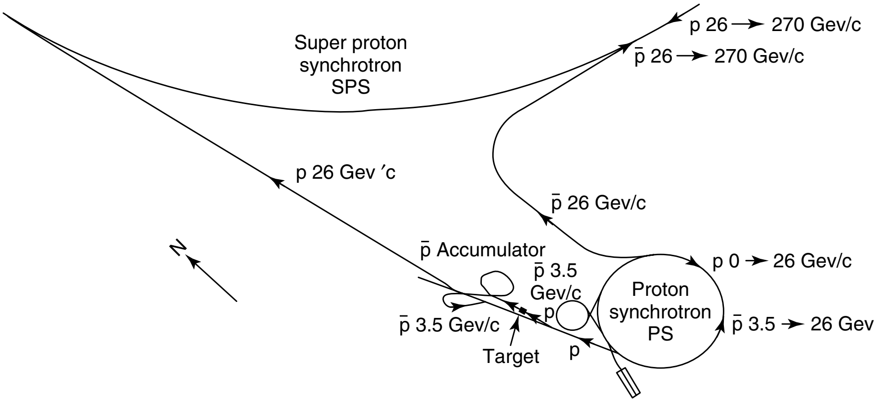
Super proton
synchrotron
SPS
p 26 → 270 Gev/c
p̄ 26 → 270 Gev/c
p 26 Gev ′c
p̄ 26 Gev/c
N
p̄ Accumulator
p̄ 3.5
Gev/c
p
p 0 → 26 Gev/c
Proton
synchrotron
PS
p̄ 3.5 Gev/c
Target
p
p̄ 3.5 → 26 Gev

Fig. 19.9 Schematic of the CERN SPS collider. Beams of protons from the PS were used to make antiprotons, which are accumulated and cooled. The antiprotons are injected into the SPS at an energy of 3.5 GeV and then accelerated to high energy for collisions with protons (Van Der Meer, 1985).

Fig. 19.10 The antiproton accumulator ring (Van Der Meer, 1985).

The successful operation of the CERN SPS as a proton–antiproton collider involved an exercise in accelerator gymnastics (Figure 19.9). Intense beams of protons from the CERN Proton Synchrotron (PS) were sent into a target where antiprotons were occasionally produced. The antiprotons were directed into an antiproton accumulator (Figure 19.10) where they were stochastically cooled. The principle of stochastic cooling involves measurement of the deviation of particles from the nominal orbit and sending a signal to a correction magnet. The correction signal is able to arrive before the particles because the direct path across the ring is shorter than the particle path along the arc (Figure 19.11). A scope trace demonstrating the fast cooling of 6×10^6 antiprotons is shown in Figure 19.12. The cooled antiprotons were injected, together with counter-rotating protons, into the SPS where they were accelerated to 315 GeV and collided.

Besides the anticipated discovery of the W and Z^0 bosons at the CERN SPS,

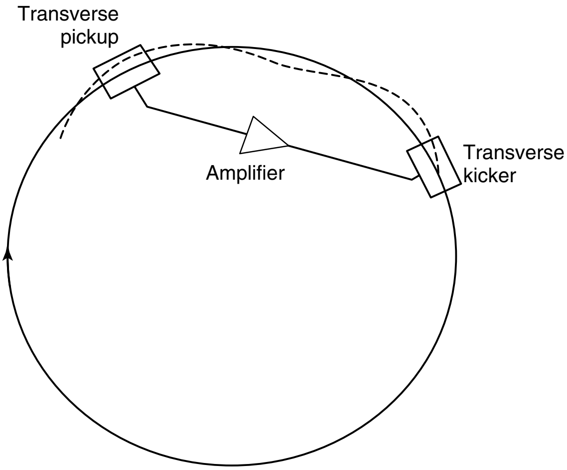

Fig. 19.11 Schematic of the cooling of the horizontal betatron oscillation of a single particle (Van Der Meer, 1985).

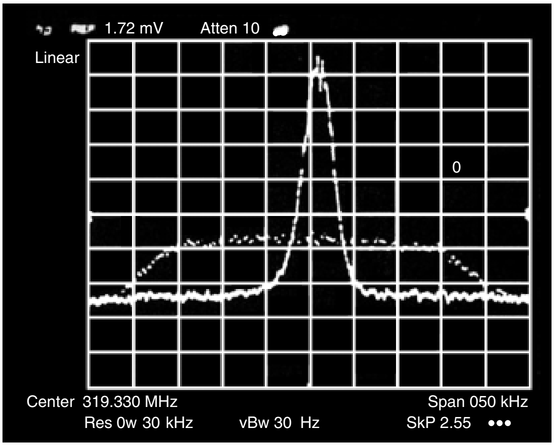

Fig. 19.12 A pulse of 6×10^6 antiprotons is cooled in 2 seconds (Van Der Meer, 1985).

another discovery dramatically changed the planning for future colliders. This was the clear observation of hadron jets from quark scattering at high energies (Figures –19.13, 19.14). Jets had been first observed in e^+e^- collisions (Hanson *et al.*, 1975) and then in low-energy hadron collisions (Bromberg *et al.*, 1980), but the results from the CERN SPS changed the international thrust for the next big accelerator almost overnight from an e^+e^- linear accelerator (linac) to a pp machine with the realization that ‘‘clean’’ physics was possible with a hadron collider (Rohlf, 2005). Sadly, the high-luminosity ISABELLE machine project at Brookhaven was canceled (eventually to become the heavy-ion RHIC machine project) to make room for the Superconducting Super Collider (SSC) in Texas, which eventually was canceled too. In the meanwhile, the rival Large Hadron Collider (LHC) machine at CERN became fully approved and actually built. The entire future of experimental high-energy physics depends on the physics learned (or not) at this machine.

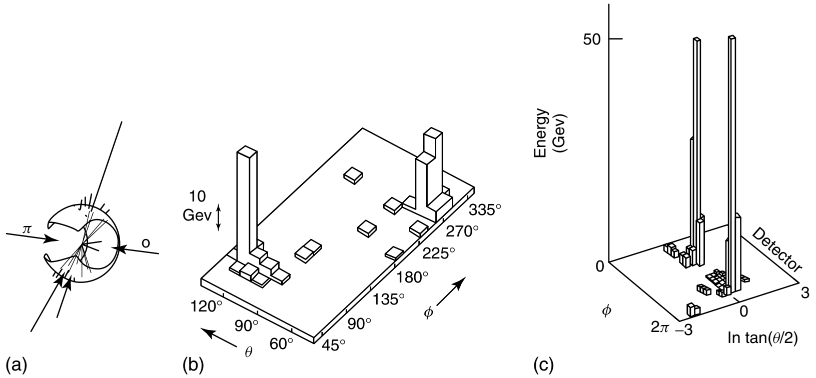

Fig. 19.13 Jet events from quark scattering observed at the CERN SPS collider in the UA2 detector (a and b) and the UA1 detector (c).

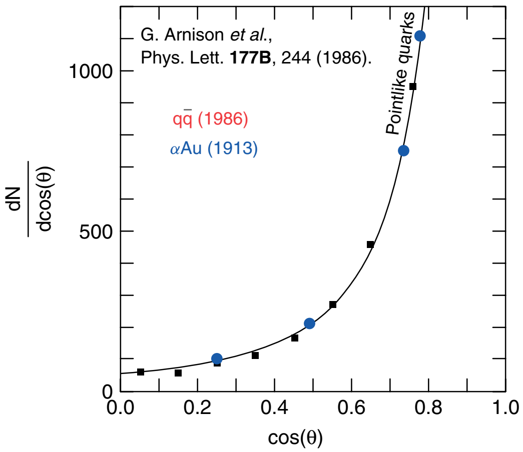

Fig. 19.14 The clean observation of jets at the CERN SPS collider allowed a Rutherford-type scattering experiment to be carried out to prove that the quarks were pointlike down particles to 10^{-18} m, interacting by the exchange of massless spin-1 gluons via a $1/r^2$ force (Arnison *et al.*, 1986).

19.1.4.4 **Fermilab Tevatron Collider**

The Fermilab Tevatron collider *http://www-bdnew.fnal.gov/tevatron/* is conceptually the same as the CERN proton–antiproton collider, except that the more powerful superconducting synchrotron magnets allow beam energies of up to 1 TeV, about three times more than previously available at the SPS collider. The Tevatron accelerator layout is shown in Figure 19.15. Owing to the improvements in the operation of the Tevatron over many years, the luminosity has exceeded that achieved at the CERN ISR (Table 19.1), long considered the benchmark for hadron colliders. The integrated luminosity delivered by the Tevatron collider is shown in Figure 19.16.

A key ingredient for the Fermilab detectors was the development of silicon vertex detectors. These detectors allow the identification of long-lived b quarks, which are the decay products of top quarks. An event with top and antitop quarks from the Collider Detector Facility (CDF) detector is shown in Figure 19.17.

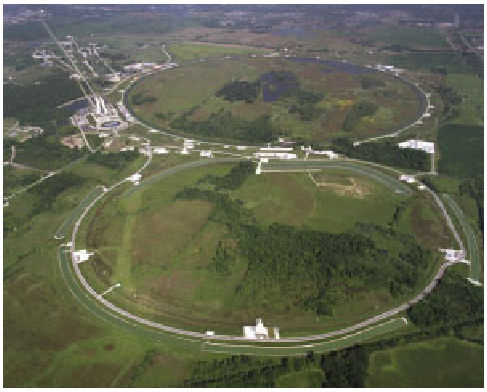

Fig. 19.15 Fermilab site showing the main injector (foreground) and the Tevatron storage ring *http://www-bdnew.fnal.gov/tevatron/*.

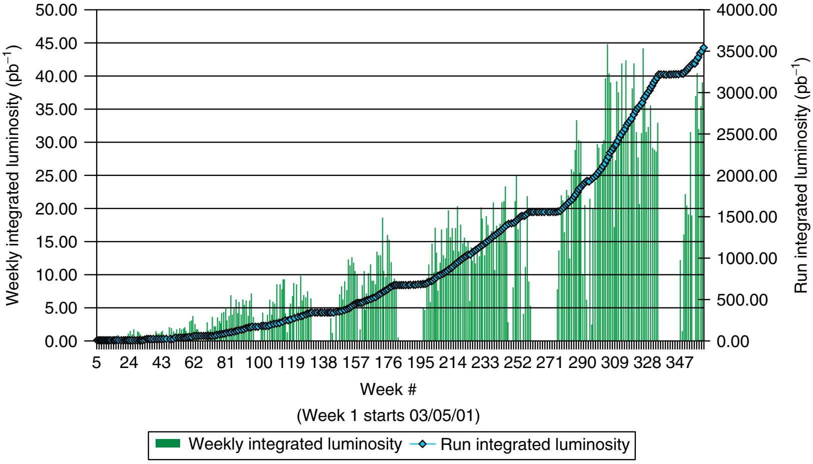

Fig. 19.16 Integrated luminosity delivered by the Tevatron collider in Run II, that began in March 2005 *http://www.fnal.gov/pub/now/tevlum.html*.

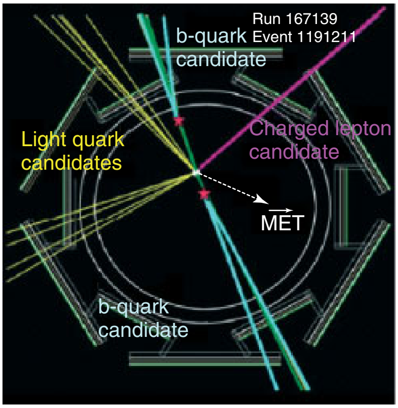
Run 167139
Event 1191211
b-quark candidate
Light quark candidates
Charged lepton candidate
$\overrightarrow{\text{MET}}$
b-quark candidate

Fig. 19.17 Event from the CDF detector at Fermilab showing evidence for the top quark *http://www-cdf.fnal.gov/*.

19.1.4.5 Large Electron–Positron Project

The Large Electron–Positron (LEP) project was designed to study the properties of the Z^0 boson as well as to be able to reach the W^+W^- threshold (Figure 19.18). The LEP accelerator was commissioned in 1989 and ran successfully until the end of 2000 (Assmann, 2001). The LEP collider was distinguished by its unprecedented size, made necessary by the phenomena of synchrotron radiation, whereby particles emit electromagnetic radiation as the fourth power of energy and the inverse square of the radius. The LEP machine was optimized for an energy of 85 GeV per beam, resulting in a circumference of 27 km after balancing the costs of construction and operation. The large radius limits the radiated power from a single electron to the μW level so that large numbers of electrons and positrons can be kept in stable orbits.

At LEP, the collision energy was measured with extraordinary accuracy, about 2 parts per 10^5. This allowed the machine to be sensitive to external forces that could change the diameter of the machine very slightly. An orbit change of 1 mm corresponds to an energy change of about 10 MeV. The machine was thus sensitive to the water level in Lake Geneva and the tides caused by the moon, which produce small distortions in the earth's crust. In addition, trains from Geneva to Paris caused induced currents on the LEP beam pipe that caused observable changes in the magnetic fields of the LEP magnets (Figure 19.19).

Figure 19.20 shows the decay of a Z^0 into quark and antiquark. Detailed measurements of the Z^0 decays proved that three and only three light neutrinos exist in nature. The precise measurement of the Z^0 mass is shown in Figure 19.21.

19.1.5
The Energy Frontier: CERN Large Hadron Collider

The CERN LHC is designed to uncover the secrets of electroweak symmetry breaking, providing a natural progression to access the physics beyond the Tevatron and LEP. The LHC occupies the same tunnel originally built for LEP. It was obvious even before LEP was completed that the radius of the tunnel was approximately 1 TeV per tesla (Amaldi, 1981), implying that if one used the same tunnel to build a proton collider, which is energy limited by magnet technology, 10 TeV per beam could be feasible with 10 T magnets. Figure 19.22 shows the location of the old LEP experiments and the new LHC experiments and services.

19.1.5.1 LHC Basic Parameters

Since the LHC physics demands the highest possible proton energy, the bending magnets are at the limit of present technology. The existing superconducting proton accelerators (Tevatron, HERA, and RHIC) all have fields around 5 T or below using Nb–Ti Rutherford cables cooled to 4.2 K. To make use of the proven cable technology and get increased field strength requires cooling below 2 K with superfluid helium. Unfortunately, the lower temperature reduces the heat capacity of the cables by an order of magnitude such that a magnet quench can be triggered by a substantially lower energy deposit. The tight mechanical control that must be maintained on the movement of the superconducting strands, plus the limited space inside the tunnel together with cost considerations, has lead to a "two-in-one" design with one cold mass containing the apertures and coils for both proton beams. The field design is 8.33 T at a current of

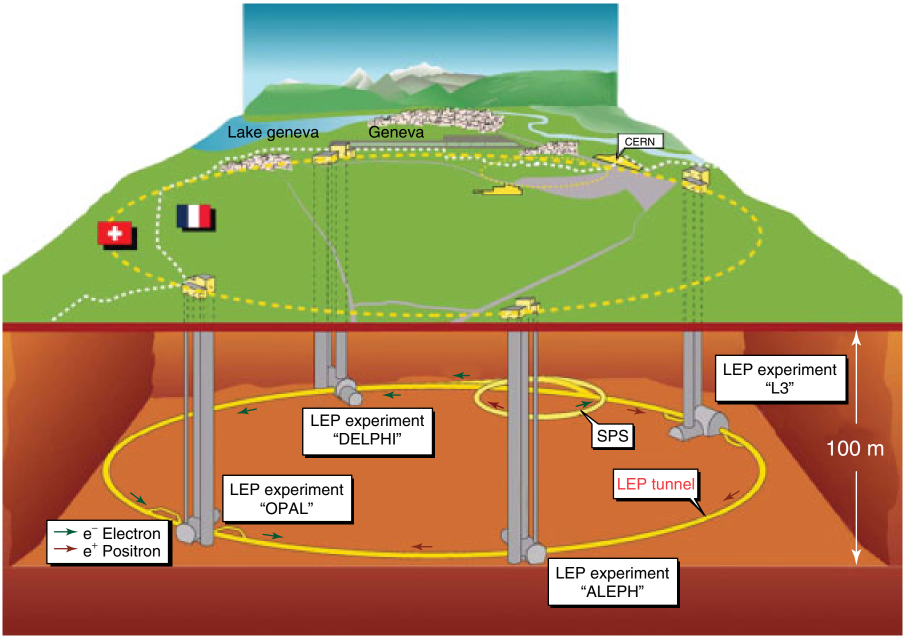
Lake geneva
Geneva
CERN
LEP experiment "L3"
LEP experiment "DELPHI"
SPS
100 m
LEP tunnel
LEP experiment "OPAL"
→ e^- Electron
→ e^+ Positron
LEP experiment "ALEPH"

Fig. 19.18 CERN Layout of the LEP tunnel and location of the four experiments: ALEPH, DELPHI, L3, and OPAL.

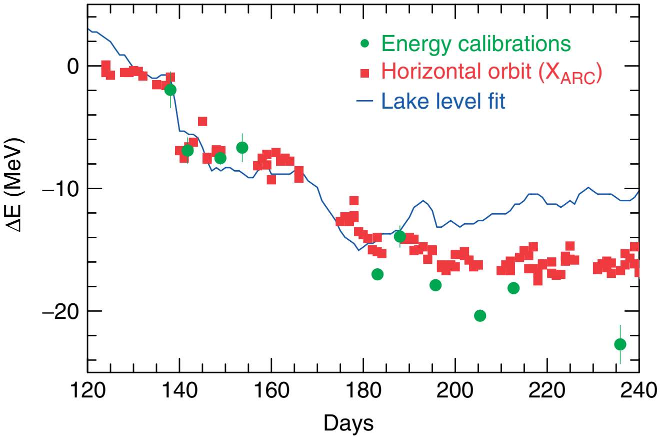

Fig. 19.19 Change in machine energy caused by the water level change in nearby Lake Geneva in 1993. The machine energy is monitored by the precise measurement of the Z^0 mass (*http://delphiwww.cern.ch/*).

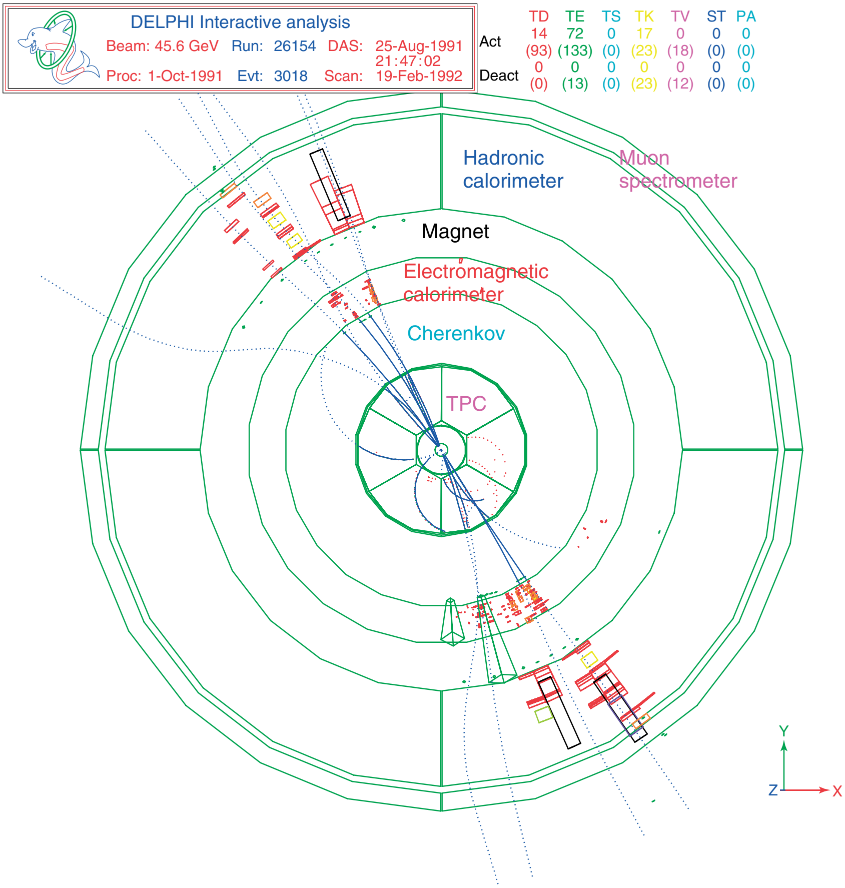
DELPHI Interactive analysis
Beam: 45.6 GeV Run: 26154 DAS: 25-Aug-1991 21:47:02
Proc: 1-Oct-1991 Evt: 3018 Scan: 19-Feb-1992
TD TE TS TK TV ST PA
Act 14 72 0 17 0 0 0
(93) (133) (0) (23) (18) (0) (0)
Deact 0 0 0 0 0 0 0
(0) (13) (0) (23) (12) (0) (0)
Hadronic calorimeter
Muon spectrometer
Magnet
Electromagnetic calorimeter
Cherenkov
TPC
Y
Z
X

Fig. 19.20 A Z^0 decay into quark and antiquark jets observed in the DELPHI detector at LEP (*http://delphiwww.cern.ch/*).

11.8 kA. The two-in-one dipole is shown in Figure 19.23. The integrated dipole magnets and helium cryostats are referred to as *cryodipoles*. Each cryodipole has a mass of 34 tons. The details of a cryodipole are shown in Figure 19.24.

The basic parameters of the machine are given in Table 19.1 (*http://ab-div.web.cern.ch/ab-div/Publications/LHC-DesignReport.html*, 2004). The superconducting main bending dipole magnets fill about two-thirds of the 27-km circumference. The remaining tunnel length is necessary for quadrupole focusing magnets and straight sections to accommodate four interaction regions plus beam injection points, radio frequency (RF) acceleration, the beam dump, and two beam-cleaning sections. The first of 1232 8.33-T dipole bending magnets was installed in 2005. The LHC group at CERN maintains an excellent website (*http://lhc.web.cern.ch/lhc/*) containing a wealth of technical information. Also much technical information is available from the

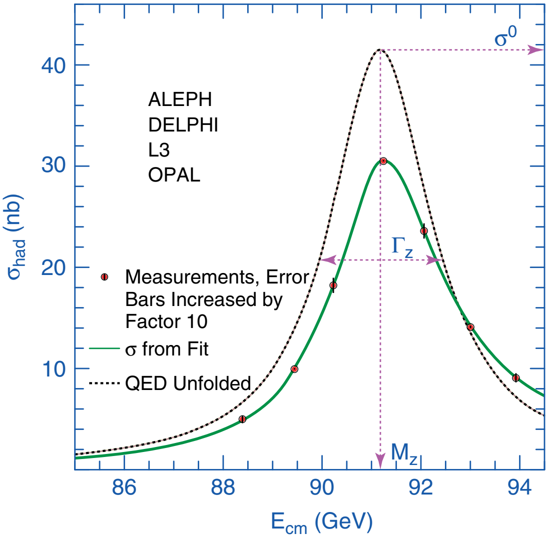

Fig. 19.21 Precision measurement of the Z^0 mass and width provides a powerful constraint to the standard model of particle physics (*http://delphiwww.cern.ch/*).

proceedings of an accelerator workshop held each winter in Chamonix, France (*http://ab-div.web.cern.ch/ab-div/conferences/Chamonix/*).

The amount of energy stored in the circulating LHC beams is unprecedented, having serious consequences for the operation of the accelerator (Schmidt, 2003). A simple calculation shows that the energy per beam is

$$\begin{aligned} E_s &= E_p n_b N_b \\ &= (7\times 10^{12}\,\text{eV})(1.6\times 10^{-19}\,\text{J/eV}) \\ &\qquad (2808)(1.1\times 10^{11}) \\ &= 350\,\text{MJ}. \end{aligned} \tag{5}$$

The LHC beams have the kinetic energy of an aircraft carrier at 10 knots or enough to melt a couple of tons of copper. Figure 19.25 shows the stored energy of previous particle accelerators versus beam momentum together with that of the LHC. The energy stored in the LHC beams is more than 2 orders of magnitude greater compared to any previous accelerator.

19.1.6
Beam Dump

The LHC has interaction region 6 dedicated to dumping the beams. The beam dump is needed for fast extraction of the proton beams into an external absorber. For the assumed 20-year lifetime of the LHC, a total of 2×10^4 beam dumps are foreseen at full energy and intensity. A loss-free extraction requires a gap of about 3 μs in the circulating proton beams to accommodate the rise time of the beam dump kicker magnets. Each proton beam has 15 kicker magnets (MKD), which deflect the beams horizontally into the high-field gaps of 15 septa magnets (MSD), which provide a vertical deflection raising the beams above the entrance to the arc. The beams then travel 750 m through special vacuum chambers passing through 10 dilution kicker magnets (MKB) before finally striking the beam dump graphite core (TDE). Special magnets (TCDS and TCDQ) are inserted to protect the LHC machine from an unplanned beam dump that might occur with the kicker magnets

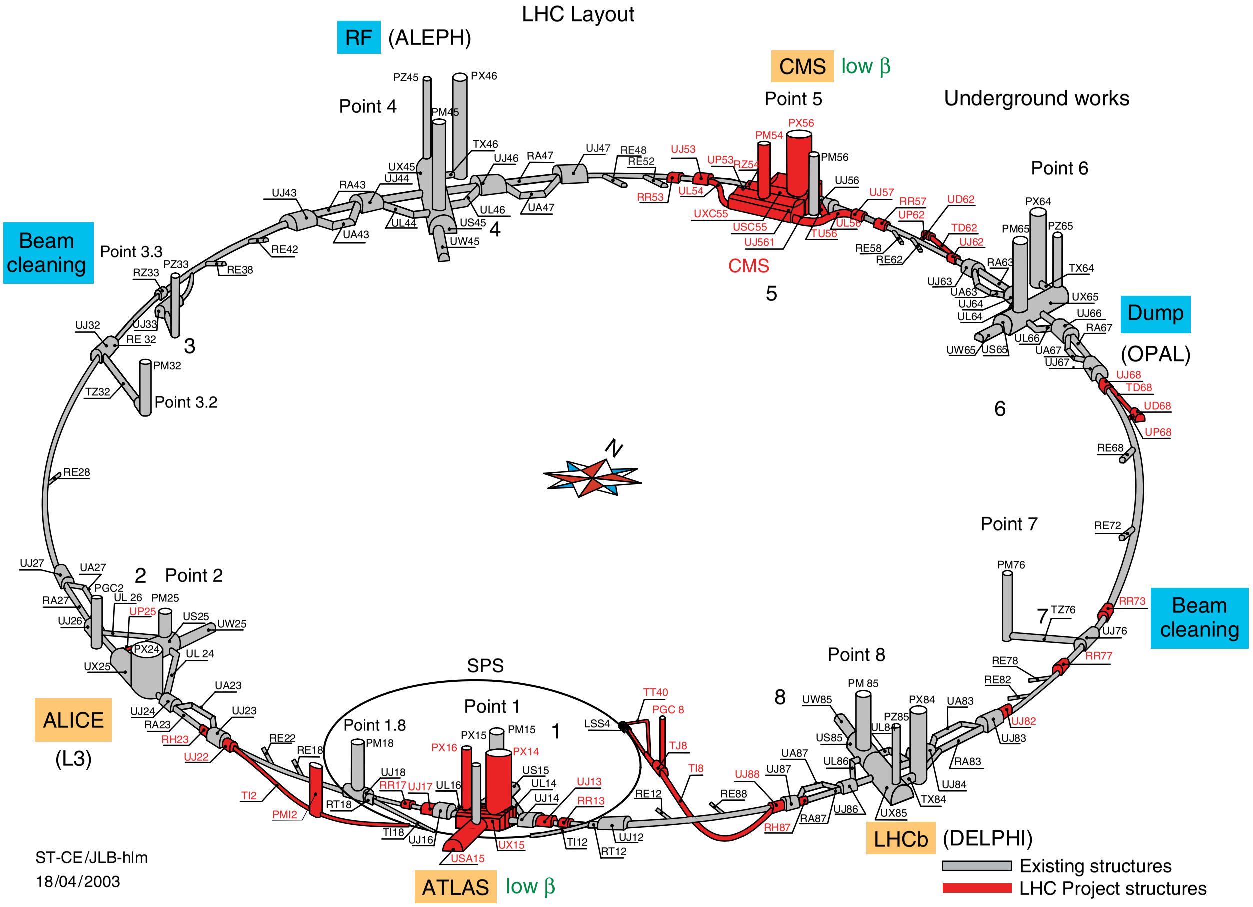
LHC Layout
RF (ALEPH)
CMS low β
Point 5
Underground works
Point 4
Point 6
Beam cleaning
Point 3.3
Point 3.2
3
4
5
CMS
6
Dump
(OPAL)
Point 7
7
Beam cleaning
2 Point 2
ALICE
(L3)
SPS
Point 1
Point 1.8
1
Point 8
8
LHCb (DELPHI)
ATLAS low β
Existing structures
LHC Project structures
ST-CE/JLB-hlm
18/04/2003
PZ45
PX46
PM45
TX46
UJ46
RA47
UJ47
RE48
RE52
UJ53
UP53
RZ54
PM54
PX56
PM56
UJ56
UJ57
RR57
UP62
UD62
TD62
UJ62
UX45
UJ44
RA43
UJ43
UL46
UA47
UL44
US45
UW45
UA43
RE42
RE38
PZ33
RZ33
UJ33
UJ32
RE 32
PM32
TZ32
RE28
RR53
UL54
UXC55
USC55
UJ561
TU56
UL56
RE58
RE62
PX64
PM65
PZ65
RA63
TX64
UJ63
UA63
UJ64
UL64
UX65
UJ66
RA67
UL66
UW65
US65
UA67
UJ67
UJ68
TD68
UD68
UP68
RE68
RE72
PM76
TZ76
RR73
UJ76
RR77
RE78
RE82
UA83
UJ82
UJ83
RA83
UJ84
TX84
PX84
PZ85
UL84
PM 85
UW85
US85
UL86
UA87
UJ87
UJ86
UX85
RA87
UJ88
RH87
RE88
TT40
PGC 8
TJ8
TI8
LSS4
RE12
UJ12
RT12
TI12
UJ13
RR13
UJ14
UL14
US15
PM15
PX15
PX14
PX16
UX15
USA15
UJ16
UL16
UJ17
RR17
UJ18
TI18
RT18
PM18
PMI2
RE18
RE22
TI2
UJ22
RH23
RA23
UJ23
UJ24
UA23
UL 24
UW25
US25
PM25
PX24
UX25
UP25
UL 26
PGC2
UA27
UJ26
RA27
UJ27
N

Fig. 19.22 The location of the old LEP experiments (L3, DELPHI, OPAL, and ALEPH) and the new LHC experiments (ALICE, ATLAS, LHCb, and CMS).

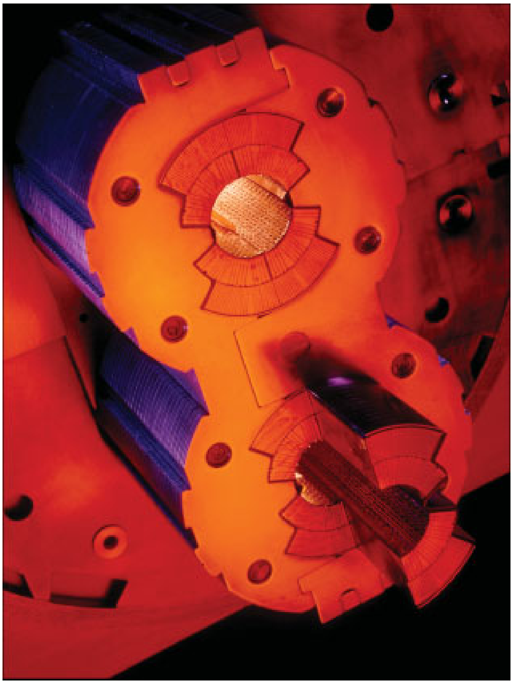

Fig. 19.23 Detailed photo of the two-in-one LHC dipole coils. Each magnet pole has an inner and outer superconducting coil at radii of 28 and 43.9 mm.

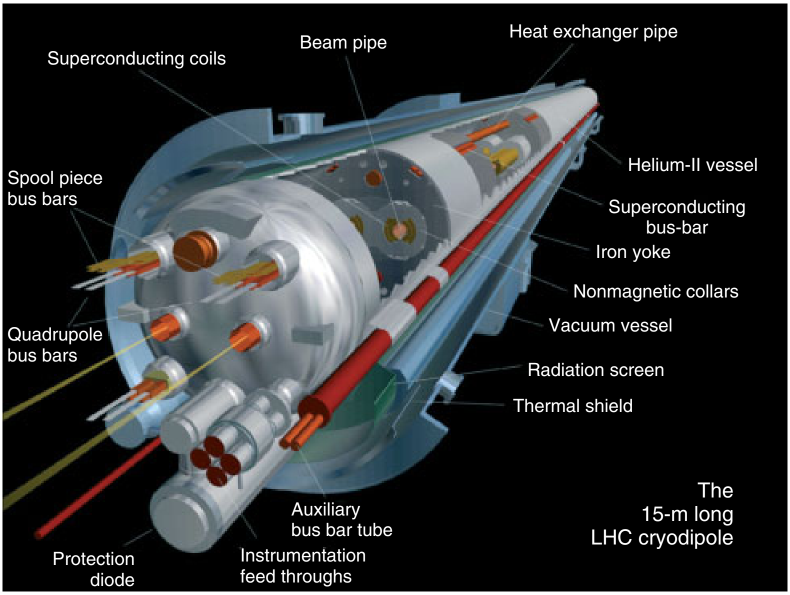
Superconducting coils
Beam pipe
Heat exchanger pipe
Helium-II vessel
Spool piece bus bars
Superconducting bus-bar
Iron yoke
Nonmagnetic collars
Vacuum vessel
Quadrupole bus bars
Radiation screen
Thermal shield
Auxiliary bus bar tube
Instrumentation feed throughs
Protection diode
The 15-m long LHC cryodipole

Fig. 19.24 The LHC bending magnet is a 15-m-long two-in-one coil plus beam pipe embedded in a liquid helium cryostat and is referred to as a *cryodipole*.

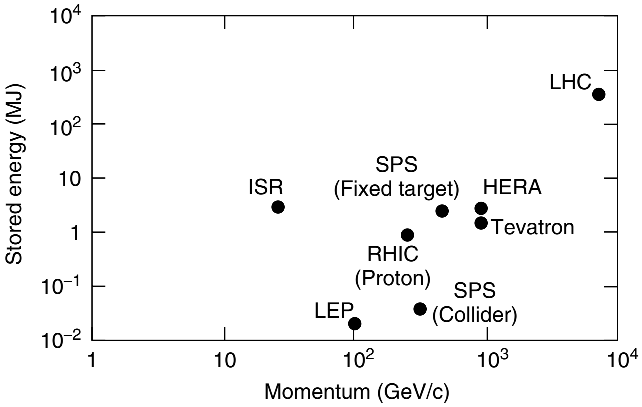

Fig. 19.25 Stored energy versus beam momentum for large accelerators. The unprecedented stored energy of the LHC beams is shown compared with that of the Intersecting Storage Rings (ISR) as the first pp collider at CERN, Super Proton Synchrotron (SPS) at CERN operating in fixed target and proton–antiproton collider modes, Hadron-Electron Ring Accelerator (HERA) ep collider at DESY, the proton–antiproton collider at Fermilab (Tevatron), the Relativistic Heavy-Ion Collider (RHIC) at Brookhaven with protons, and the Large Electron–Positron (LEP) collider at CERN.

ramping out of synchronization with the planned abort gap.

Each beam ends up in a carbon graphite core that is chosen for its high melting temperature and thermal shock resistance. The carbon graphite core is a cylinder 0.7-m diameter and 7.7-m length and is fitted into a water-cooled stainless steel jacket. The core is surrounded by 900 tons of radiation shielding blocks made from recycled ISR magnets filled with concrete. The dilution magnets (Figure 19.26), four horizontal and six vertical, providing kicks of $1.6\,\mathrm{T \cdot m}$ and $1.1\,\mathrm{T \cdot m}$, respectively, are used to sweep the beam into a letter “e” pattern during the dumping process in order to spread out the energy deposition. The dilution magnets are sinusoidally powered by a pulse generator (14.2-kHz horizontal and 12.7-kHz vertical) via a low-impedance transmission line. The resulting pattern at the graphite core is shown in Figure 19.27. The motion of the impact point on the beam dump is at least $10\,\mathrm{mm\mu s^{-1}}$, resulting in a trace that is approximately 1 m in length for the 89-μs orbit period.

19.1.7
LHC Layout

The LHC is divided into octants as indicated in Figure 19.28. Each octant has at its center a 528-m long straight section (LSS). At four of these sections, the beam lines cross and the experiments are located. The general purpose ATLAS experiment is housed at LSS1 or “point 1” adjacent to the CERN main entrance. The neighboring sections, LSS8 and LSS2, house ALICE, a retrofit of the former L3 experiment at LEP to specialize in heavy-ion collisions, and LHCb, a spectrometer specialized for b-quark physics, respectively. The beams are also injected from the SPS into the LHC rings at points 2 and 8 with a vertical momentum kick from below the LHC plane. The other general purpose detector, CMS, is located at LSS5. For ATLAS, the proton beams cross in the vertical plane,

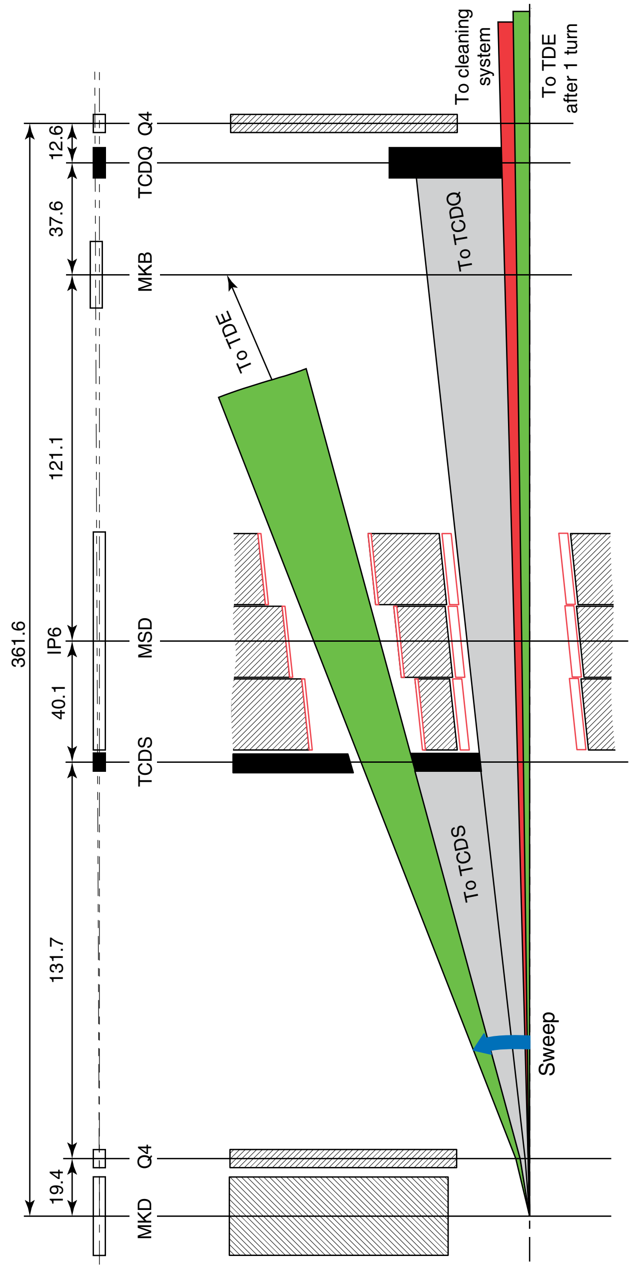
361.6
12.6
37.6
121.1
IP6
40.1
131.7
19.4
Q4
TCDQ
MKB
MSD
TCDS
Q4
MKD
To cleaning system
To TDE after 1 turn
To TCDQ
To TDE
To TCDS
Sweep

Fig. 19.26 Sweeping of the protons by the dilution magnets (LHC Design Report Volume 1, Chapter 17: Beam Dumping System (*http://lhc.web.cern.ch/lhc/*)).

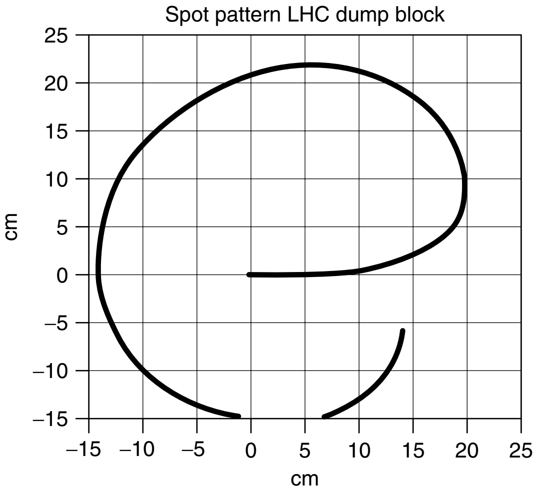

Fig. 19.27 Pattern of the dumped beam as it hits the absorber block (LHC Design Report Volume 1, Chapter 17: Beam Dumping System (*http://lhc.web.cern.ch/lhc/*)). The pattern corresponds to one revolution of the proton beam.

while for CMS they cross in the horizontal plane. Beam collimators are located at LLS3 and LSS7, the beam dump at LSS6, and the RF acceleration of each beam at LSS4. Each LSS is sandwiched between dispersion sections (DSL and DSR).

An arc of the LHC is optimized for maximum integrated field strength to result in the highest possible proton energy. Each arc contains 23 106.9-m cells, which are divided into half cells that each contain a quadrupole embedded in a short straight section (SSS) 5.335-m cold mass in a 6.63-m cryostat plus 3 14.3-m dipole bending magnets. The LHC cell is shown in Figure 19.29.

19.1.8
Cryogenics

Part of the uniqueness of the LHC among superconducting synchrotrons is its operating temperature of 1.9 K in order to maximize the field strength that can be obtained with NbTi coils. This is accomplished by immersing the coils in a pressurized (1.3 bar) bath of superfluid helium. In addition, the effective thermal conductivity of helium is at a maximum at 1.9 K providing a margin of stability over operation at the conventional 4.5 K. The expected temperature profiles of the LHC magnets are shown in Figure 19.30.

The cryogenic system is designed to cool the entire LHC cold mass of 3.7×10^7 kg in 15 days and to rapidly warm up and cool strings of magnets in order to replace a defective unit.

The LHC refrigeration entropic capacity is 144 kW at 4.5 K, making it the world's largest helium system. The design has eight refrigeration stations (one every 3.3 km) including recycled use of the four LEP refrigeration units and their associated infrastructure. The cryogenic system is designed for continuous operation for nine months followed by a three-month shutdown for maintenance.

19.1.8.1 Injection Scheme

The protons for each LHC beam are accelerated by a linac to a kinetic energy of 50 MeV, a booster to 1.4 GeV, the PS to 25 GeV, the SPS to 450 GeV, and finally the LHC to 7 TeV (Figure 19.31). The

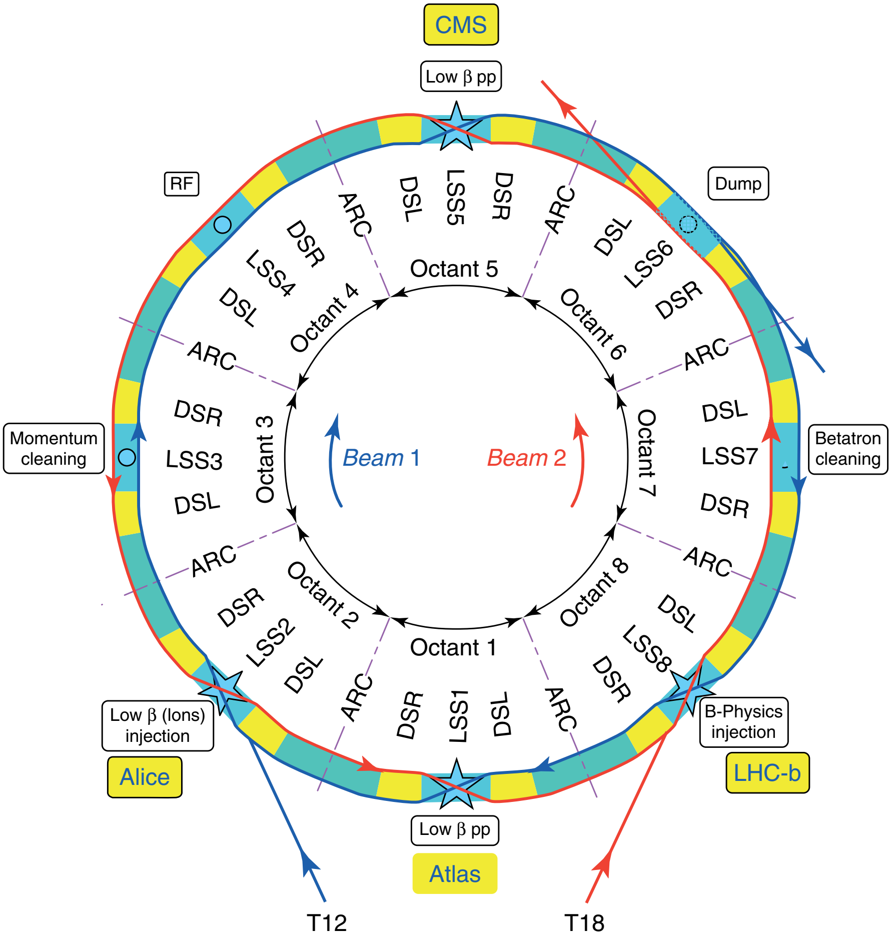
CMS
Low β pp
RF
Dump
ARC
DSL
LSS5
DSR
ARC
DSR
LSS4
DSL
DSL
LSS6
DSR
Octant 5
Octant 4
Octant 6
ARC
ARC
Momentum cleaning
DSR
LSS3
DSL
Octant 3
Beam 1
Beam 2
Octant 7
DSL
LSS7
DSR
Betatron cleaning
ARC
ARC
DSR
LSS2
DSL
Octant 2
Octant 1
Octant 8
DSL
LSS8
DSR
ARC
DSR
LSS1
DSL
ARC
Low β (Ions) injection
Alice
B-Physics injection
LHC-b
Low β pp
Atlas
T12
T18

Fig. 19.28 The overall layout of the LHC (LHC Design Report Volume 1, Chapter 3: Layout and Performance (*http://ab-div.web.cern.ch/ab-div/Publications/LHC-DesignReport.html*, 2004)). A section of the ring consists of a bending arc (ARC) placed between dispersion sections (DSL and DSR) and a long straight section (LSS). The beams collide at the center of octants 1, 2, 5, and 8. Octants 3 and 7 contain collimators and other apparatus to clean up the beams. Octant 4 contains a beam dump and octant 6 contains the radio frequency (RF) acceleration cavities.

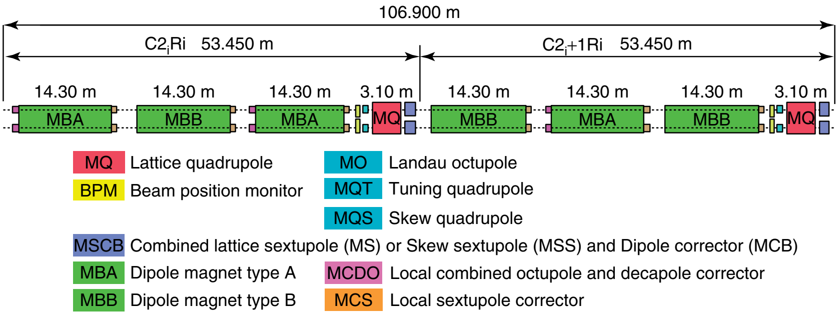
106.900 m
C2iRi 53.450 m
C2i+1Ri 53.450 m
14.30 m
14.30 m
14.30 m
3.10 m
14.30 m
14.30 m
14.30 m
3.10 m
MBA
MBB
MBA
MQ
MBB
MBA
MBB
MQ
MQ Lattice quadrupole
BPM Beam position monitor
MO Landau octupole
MQT Tuning quadrupole
MQS Skew quadrupole
MSCB Combined lattice sextupole (MS) or Skew sextupole (MSS) and Dipole corrector (MCB)
MBA Dipole magnet type A
MBB Dipole magnet type B
MCDO Local combined octupole and decapole corrector
MCS Local sextupole corrector

Fig. 19.29 One cell of the LHC *http://lhc.web.cern.ch/lhc/* contains 3 dipoles (MBA, MBB, and MBA) with their corrector magnets (MCSB, MCDO, and MCS) followed by a short straight section containing the main quadrupole (MQ) plus a number of smaller elements (BPM, MO, MQT, MQS and MSCB). The whole pattern is then repeated with three more dipoles (MBB, MBA, and MBB) and another short straight section.

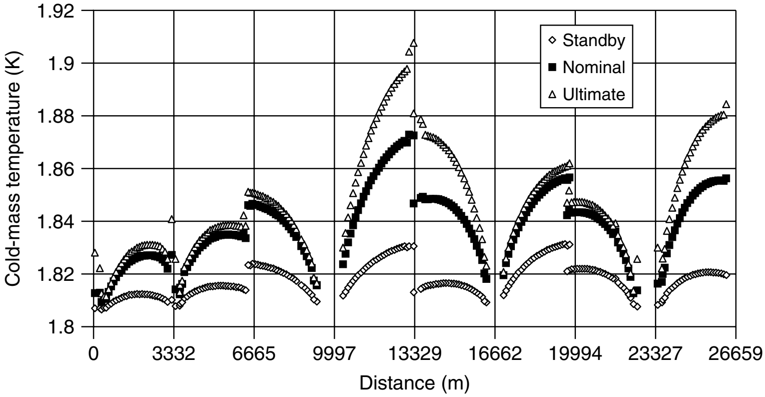

Fig. 19.30 Temperature profiles of the LHC magnets (LHC Design Report Volume 1, Chapter 11: Cryogenics (*http://ab-div.web.cern.ch/ab-div/Publications/LHC-DesignReport.html*, 2004)).

important geometric factors in the injection scheme are that the PS circumference is 1/11 of that of the SPS, which in turn is 7/27 that of the LHC. These factors have important implications for the time structure of the beams.

The baseline design calls for operation of the machine at approximately 40 MHz. This is based on “harmonic 84,” in which the PS is divided into 84 time bins or *buckets*. The geometric factors for the PS and SPS lead to a division of the LHC into

$$(84)\,(11)\left(\frac{27}{7}\right) = 3564 \tag{6}$$

time buckets of size

$$\frac{88924\,\text{ns}}{3564} = 24.95\,\text{ns} \tag{7}$$

resulting in a beam crossing frequency of 40.079 MHz.

There are many factors that limit filling of the LHC buckets with protons. The

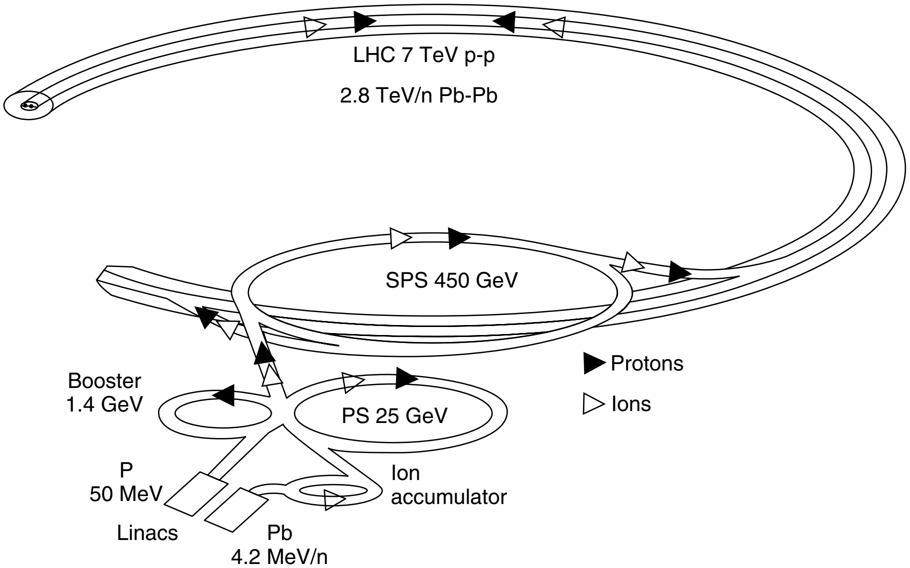

Fig. 19.31 The LHC injection scheme involves a proton linac, a booster ring, the proton synchrotron (PS), and the super proton synchrotron (SPS) accelerators. A second linac is used to accelerate heavy ions where they are collected in an ion accumulator.

main hardware limits come from the SPS injection kicker rise time of 225 nanoseconds, the LHC injection kicker rise time of 975 nanoseconds, and the 3 microseconds needed for the LHC beam dump kicker. The resulting proton bunch disposition is shown in Figure 19.32. The time needed for SPS injection means that only 72 proton bunches of 25-nanosecond spacing (out of 84 possible) are loaded into the PS. The time needed for LHC injection means that only 10 PS fills are loaded into the SPS (of 11 possible). Four SPS fills are needed to load the LHC, but the last SPS injection has only 9 PS fills because of the time needed for the LHC beam dump kicker, necessary to empty the protons from the LHC. The net result is that of the 3564 LHC time buckets only 2808 contain protons, according to the formula

$$\begin{aligned}3564 = \big\{&[[(72\mathrm{b} + 8\mathrm{e}) \times 3] + 30\mathrm{e}] \times 2\\ &+ [[(72\mathrm{b} + 8\mathrm{e}) \times 4] + 31\mathrm{e}]\big\} \times 3\\ &+ \big\{[[(72\mathrm{b} + 8\mathrm{e}) \times 3] + 30\mathrm{e}] \times 3\\ &+ 81\mathrm{e}\} \qquad (8)\end{aligned}$$

where b stands for buckets with protons and e stands for empty buckets. The 8e is due to injection into the SPS, the 8e + 30e and 8e + 31e are due to injection into the LHC, and the 8e + 30e + 81e is due to the LHC beam dump. The resulting bunch structure is extremely fortuitous for the experimenters, because a priori knowledge of when the beams can and cannot collide provides a powerful technique to synchronize the electronics. The large gap of 3.0 microseconds, referred to as the *abort gap*, may be used for fast reset of the electronics and certain calibration triggers.

The first time bucket after the abort gap is referred to as *bunch count zero* (*BC0*).

To synchronize the transfer, the LHC sends two master clock signals via fiber optics to the SPS: (i) a fiducial frequency, which is the LHC revolution frequency divided by 7 (equal to the SPS frequency divided by 27) and (ii) the 400.8-MHz RF that is phase adjustable. The operation of the SPS is locked to these clock signals and the relative phase of the RF is adjusted to place the injected protons at the center of the LHC bucket.

19.1.9
Radio Frequency Acceleration

Since protons cannot be accelerated with magnetic fields, this leaves the electric field as the only technique for particle acceleration. At the beginning of the acceleration stage, this is accomplished with a constant (DC) high voltage. In the PS, SPS, and LHC, the electric field causing the acceleration is sinusoidal (AC) at an RF. RFs are chosen because the electric field can be fed into a cavity whose dimensions are chosen to support a resonant standing wave. The presence of a strong resonance means that the protons are pushed only by the RF electric field and not by any potential noise. The RF cavity contains a gap where voltage is applied such that every time a proton crosses the gap it is accelerated by a small fixed amount. The desired acceleration occurs quickly since the orbit frequency is large. Voltage is fed into the cavity in a suitable fashion.

The RF acceleration in the LHC takes the proton beams from injection energy of 450 GeV to the collision energy of 7 TeV and is also needed to keep the tight bunch structure of the protons. The RF acceleration occurs in the LSS of octant 4 (Figure 19.28). Unlike LEP, which as an electron accelerator was burdened with synchrotron radiation, the main consideration of RF design (Boussard and Linnecar, 1999) for the LHC is determined by the need for precise voltage control in the presence of a high current of protons (0.5 A) and not by the total power delivered to the beam. Use of conventional copper cavities would lead to unacceptable bunch lengths.

The LHC RF must be a multiple of the SPS RF (200.4 MHz) in order to allow rapid transfer of multiple bunches. The bunch length in the SPS at 450 GeV is 1.6 nanoseconds. This means that an RF system operating at 400.8 MHz (2.5-nanosecond period) can be used to capture the protons at injection into the LHC, as well as accelerate them and maintain their energy for storage during collisions. A maximum acceleration voltage of 16 MV is chosen to be compatible with the desired maximum bunch length and emittance at 7 TeV to meet design luminosity. An LHC superconducting RF module is shown in Figure 19.33.

To accommodate the high beam current and the 3-microsecond abort gap, 30-cm-wide aperture superconducting RF cavities are chosen. The RF cavities are based on the successful design used at LEP, copper cavities with a niobium film on the RF surface. (The LEP cavities operated at 352 MHz.) To make room for the RF system for two separate beams, special superconducting dipoles are used to separate the beams by 42 cm at point 4. With the RF systems for the two beams longitudinally staggered, the adjacent beam can pass outside the RF cavity of the beam being accelerated; however, it still passes through the RF cryostat as shown in Figure 19.34. This special section of the beam pipe is designed so as not to interfere with the replacement of a faulty RF cavity.

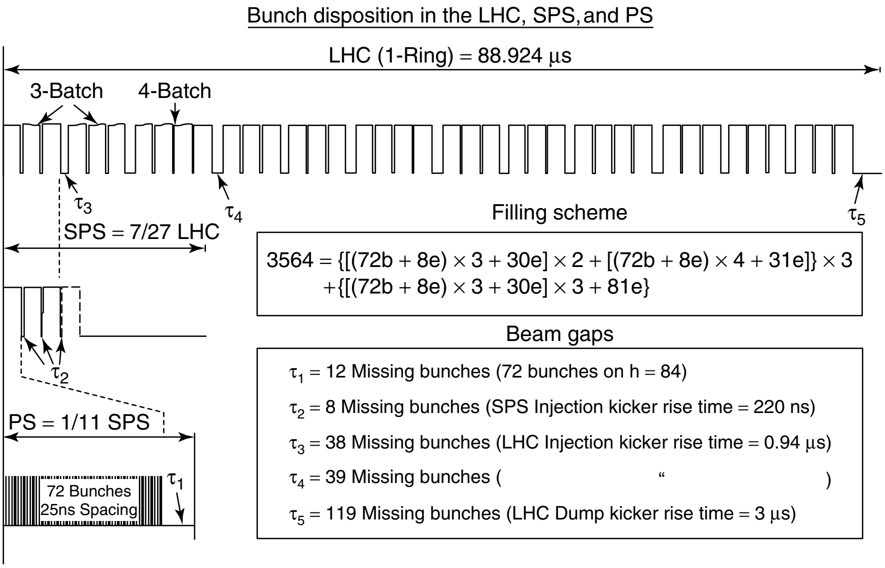
Bunch disposition in the LHC, SPS, and PS
LHC (1-Ring) = 88.924 μs
3-Batch
4-Batch
τ_3
τ_4
τ_5
SPS = 7/27 LHC
τ_2
PS = 1/11 SPS
72 Bunches
25ns Spacing
τ_1
Filling scheme
$3564 = \{[(72b + 8e) \times 3 + 30e] \times 2 + [(72b + 8e) \times 4 + 31e]\} \times 3 + \{[(72b + 8e) \times 3 + 30e] \times 3 + 81e\}$
Beam gaps
τ_1 = 12 Missing bunches (72 bunches on h = 84)
τ_2 = 8 Missing bunches (SPS Injection kicker rise time = 220 ns)
τ_3 = 38 Missing bunches (LHC Injection kicker rise time = 0.94 μs)
τ_4 = 39 Missing bunches (")
τ_5 = 119 Missing bunches (LHC Dump kicker rise time = 3 μs)

Fig. 19.32 Proton bunch structure in the LHC (P. Collier (*http://lhc.web.cern.ch/lhc/*)).

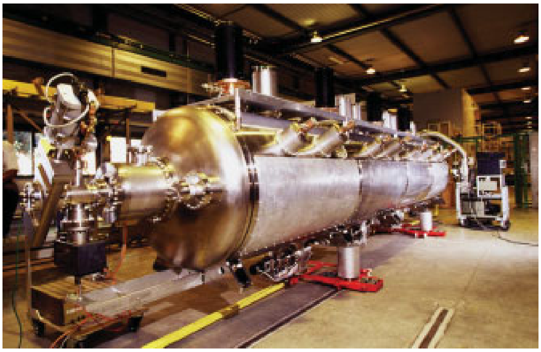

Fig. 19.33 Superconducting RF module.

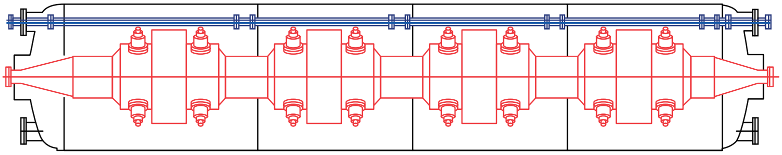

Fig. 19.34 In order to provide space for the RF cavities, the beams are separated from their nominal 194–420 mm in point 4. The beam adjacent to the one being accelerated, however, still must pass through the superconducting RF cryostat. A second beam pipe is installed in the RF cryostat outside of the acceleration cavities as shown.

This is accomplished with eight superconducting RF cavities per proton beam, two groups of four in a single cryostat, giving a total acceleration voltage of 16 MV. Each RF cell provides 2 MV, with a corresponding electric field of 5.5 MeV m^{-1}.

Klystrons are specialized vacuum tubes that have been in existence since 1937. Their function in the LHC is to amplify microwave signals that are sent to the RF cavities through which the proton beams pass. The RF power (4800 kW maximum) is provided by 16 300-kW klystrons, one per cavity, operating at 10 times the crossing frequency, or 400.8 MHz. The klystrons were developed specially for the LHC, although much of the auxiliary equipment is recycled from LEP. The klystrons are installed at a distance of 10 m in a ''klystron gallery'' adjacent to the LHC tunnel.

19.1.10
Beam Vacuum

A high-quality vacuum for the proton beams is needed in the LHC, otherwise the protons will be rapidly depleted through interactions with the gas molecules (beam–gas interactions). This poses a challenge in a superconducting environment where heat must be minimized and large numbers of gas molecules condense on the vacuum chamber. The beam vacuum must be able to handle the dynamic effects of synchrotron radiation (200 mW m^{-1} per beam), image currents (200 mW m^{-1} per beam), energy loss from nuclear interactions (20 mW m^{-1} per beam), plus the energy dissipated from the formation of electron clouds.

The beam vacuum system includes the following parts: the experimental regions, the adjacent long straight sections, and the arcs. The LHC vacuum contains components at both room temperatures in the region of the experiments and at cryogenic temperatures inside the magnets. The beam vacuum is divided into sectors defined by the distance between cryomagnets, except in the arcs, which are single sectors of length 2900 m. Additional vacuum systems are needed to insulate the helium distribution line and the cryomagnets.

The quality of vacuum in an accelerator is measured by the single-beam lifetime, the time taken for the number of particles in stable orbit to be reduced by a factor of $(1 - (1/e))$, *approximately* 0.63, due to beam–gas interactions. The design figure of merit for the LHC vacuum lifetime is 100 h so as not to inhibit the luminosity lifetime of 15 hours, which ideally would be dominated by proton–proton collisions. The 100-hour beam lifetime corresponds to an equivalent hydrogen (H_2) gas density of 10^{15} m^{-3}. In the immediate vicinity of the experiments, this vacuum will be 100 times better at 10^{13} m^{-3} equivalent hydrogen density.

Figure 19.35 shows the vacuum pressure versus temperature for several common molecules (Redhead, Hobson, and Kornelsen, 1993). Superimposed on this graph is the vacuum pressure needed to achieve the LHC design vacuum lifetime of 100 hours. For CO this vacuum pressure corresponds to 10^{-9} torr. An overview of the LHC vacuum system and problems to be faced for maintenance and commissioning of the accelerator are given in (Refs. Baglin [2004–2006]).

The LHC beam vacuum system is equipped with a ''beam screen,'' perforated low magnetic permeability stainless steel, which lines 44 km of magnet walls (Cruikshank, 1997) (see Figure 19.36). The beam screen operates between 5 and 20 K and serves the purpose of reducing the heat load on the superconducting magnet

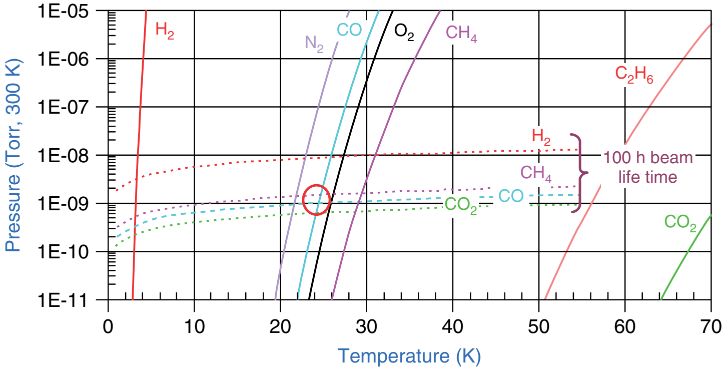

Fig. 19.35 Vacuum pressure of common gases (Redhead, Hobson, and Kornelsen, 1993) with the LHC beam lifetime superimposed (Baglin, 2004).

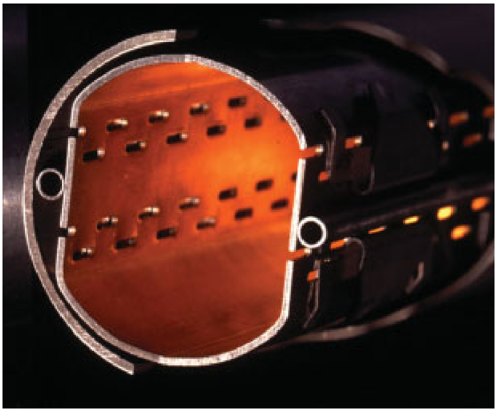

Fig. 19.36 The beam screen is a copper-clad stainless steel liner for the vacuum system. The cross-sectional area is about 44 mm × 34 mm. The beam screen reduces the heat load on the helium, which cools the LHC magnets.

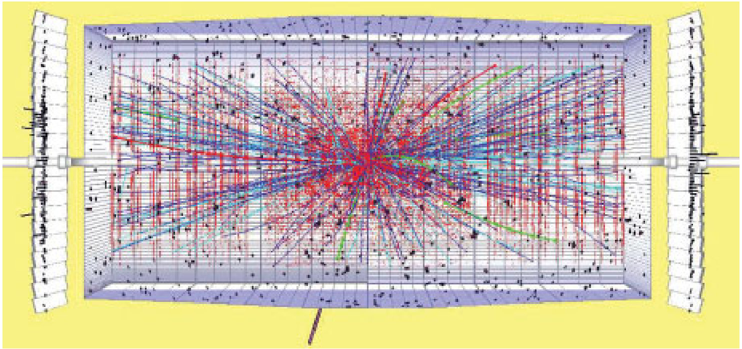

Fig. 19.37 Simulated LHC pp collisions for one beam crossing (25 nanosecond) in the Compact Muon Solenoid detector at design luminosity.

helium bath. The beam screen is lined with 50 μm of high purity copper in order to provide a low-resistance path for the image current of the beam and to resist induced forces from eddy currents during a magnet quench. The beam screen contains holes through which gas molecules may be pumped. The pattern and size of the holes is arranged to minimize impedance and keep RF losses below 1 mW m^{-1}.

The LHC is the largest scientific instrument ever built. The detectors are even more sophisticated than thought possible when the large hadron collider was in the infancy of its planning stage two decades ago. Figure 19.37 shows simulated pp collisions from a single 25-nanosecond beam crossing in the Compact Muon Solenoid detector at design luminosity (compare with Figure 19.3). Individual pp collision occur at 1 GHz.

19.1.11
Super LHC

An upgrade to the LHC is foreseen, which will raise the luminosity by an order of magnitude to $10^{35} cm^{-2} s^{-1}$. This project is known as the Super Large Hadron Collider (SLHC). The increased luminosity would come from increased currents (×5) and squeezing of the beams (smaller β by ×2) at interaction points 1 and 5 (ATLAS and CMS). The increase in luminosity and associated radiation levels near the interaction regions will require extensive upgrades to the detectors (Rohlf, 2003).

19.1.12
The Next Frontier

19.1.12.1 International Linear Collider

The International Linear Collider (ILC) is a proposed 35-km, 500-GeV total energy e^+e^- collider intended to complement the LHC (*http://www.linearcollider.org/cms/*). The design goals call for an integrated luminosity of 50 fb^{-1} during the first four years and an ability to scan in center-of-mass energy from 200 to 500 GeV. An upgrade to 1 TeV is foreseen as a second stage. The ILC design uses 1.3-GHz superconducting RF cavities providing an accelerating gradient of 31.5 MeV m^{-1}. The machine would deliver a 1-microsecond pulse at the rate of 5 Hz.

19.1.12.2 Compact Linear Collider

The Compact Linear Collider (CLIC) is an e^+e^- collider designed for a nominal center-of-mass energy of 3 TeV (*http://clic-study.web.cern.ch/clic%2Dstudy/*). The CLIC design differs from the ILC in that it uses a very high accelerating gradient of 150 MeV m^{-1}. Such a gradient would be achieved using room-temperature traveling wave structures operating at 30 GHz. The CLIC acceleration is a two-beam concept, with the RF power generated by a high-current electron beam (the drive beam) parallel to the main beam. A lower cost option would use an RF of 12 GHz generating 100 MeV m^{-1}. The proposed CLIC layout is shown in Figure 19.38.

19.1.12.3 Very Large Hadron Collider

While a TeV linear collider may well be able to exploit the initial discoveries made at the LHC, it seems that the next big step in energy will require a bigger hadron collider, referred to as the *Very Large Hadron Collider* (VLHC). Such a machine is foreseen to collide protons at a center-of-mass energy of 100 TeV or more (Foster, 2000). While the details are much less defined than the ILC or CLIC, one basic design option for the VLHC is a scaling of the LHC with high-field cold-bore magnets, while other options use larger tunnels with smaller field magnets.

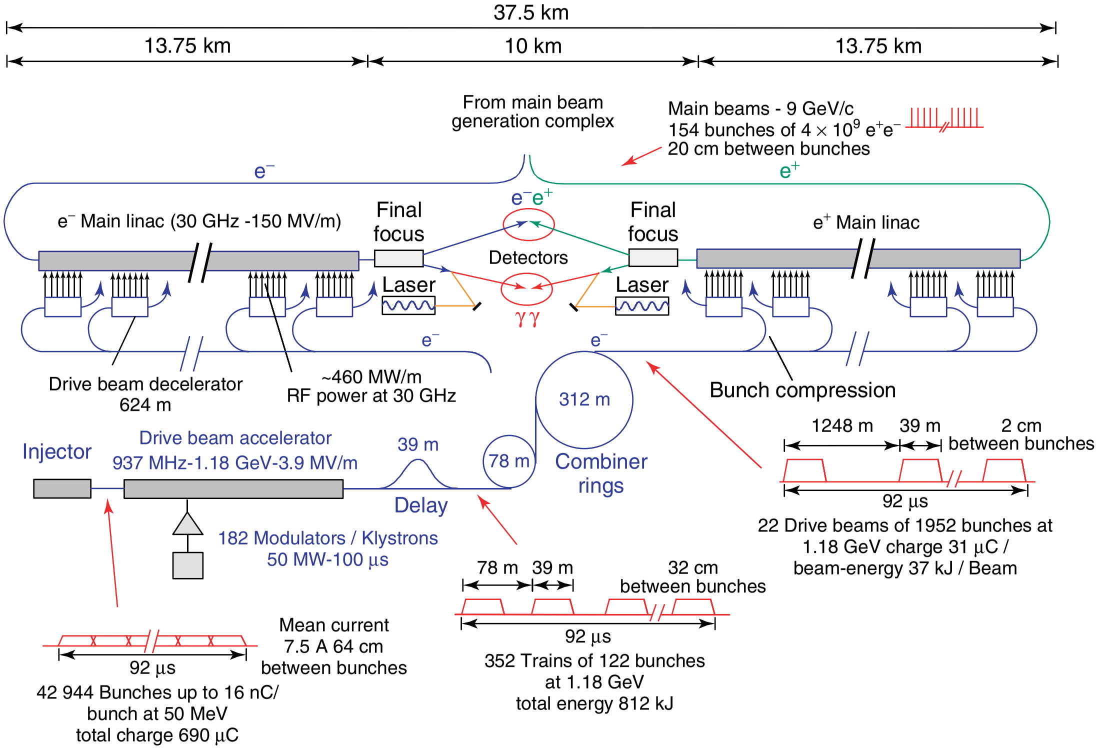
37.5 km
13.75 km
10 km
13.75 km
From main beam generation complex
Main beams - 9 GeV/c
154 bunches of 4 × 10^9 e+e−
20 cm between bunches
e−
e+
e− Main linac (30 GHz -150 MV/m)
Final focus
e−e+
Detectors
Final focus
e+ Main linac
Laser
γγ
Laser
e−
e−
Drive beam decelerator
624 m
~460 MW/m
RF power at 30 GHz
312 m
Bunch compression
Injector
Drive beam accelerator
937 MHz-1.18 GeV-3.9 MV/m
39 m
78 m
Combiner rings
Delay
182 Modulators / Klystrons
50 MW-100 µs
1248 m
39 m
2 cm
between bunches
92 µs
22 Drive beams of 1952 bunches at
1.18 GeV charge 31 µC /
beam-energy 37 kJ / Beam
78 m
39 m
32 cm
between bunches
92 µs
352 Trains of 122 bunches
at 1.18 GeV
total energy 812 kJ
Mean current
7.5 A 64 cm
between bunches
92 µs
42 944 Bunches up to 16 nC/
bunch at 50 MeV
total charge 690 µC

Fig. 19.38 The CLIC accelerator design layout for 3 TeV e^+e^- collisions (*http://clic-study.web.cern.ch/clic%2Dstudy/*).

References

Amaldi, U. (**1981**) *European Projects in High-energy Physics, CERN-EP/81-149*, invited talk given at the AIP Meeting on Particles and Fields, 9–11 Sept. 1981 Santa Cruz, CA, published in DPF: 0415 (QCD161:A6:1981).

An index of useful information from the accelerator workshops held each year in Chamonix, France is maintained by the CERN Accelerators and Beams Department: *http://ab-div.web.cern.ch/ab-div/conferences/Chamonix/*. (**2003**).

Arnison, G. *et al.* (**1986**) Angular distributions for high mass Jet Pairs and a limit on the energy scale of compositeness for quarks from the CERN p anti-p collider. *Phys. Lett. B*, **177**, 244; Rutherford, E. (**1911**) The scattering of α and β particles by matter and the structure of the atom. *Philos. Mag.*, **21**, 669.

For an excellent summary of the history of the LEP machine performance, see Assmann, R. W. LEP Operation and Performance with Electron Positron Collisions at 209-GeV, prepared for 11th Workshop on LEP Performance, 15–19 Jan 2001, Chamonix, France, published in "Chamonix (**2001**), LEP performance" 323–334.

Baglin, V. (**2004**) Vacuum Transients During LHC Operation, prepared for The Chamonix XIII LHC Project Workshop, 19–23 January 2004, Chamonix, France, *http://ab-div.web.cern.ch/ab-div/conferences/Chamonix/chamx 2004/PAPERS/8_06_BV.pdf*

Baglin, V. (**2005**) How to Deal With Leaks in the LHC Beam Vacuum, prepared for The Chamonix XIX LHC Project Workshop, 17–21 January 2005, Chamonix, France, *http://ab-div.web.cern.ch/ab-div/conferences/Chamonix/chamx2005/PAPERS/3_02.pdf*.

Baglin, V. (**2006**) Vacuum Conditions Required, prepared for The Chamonix XV LHC Project Workshop, 23–27 January, 2006, Chamonix, France, *http://ab-div.web.cern.ch/ab-div/conferences/Chamonix/Chamx2006/PAPERS/VB_1_5.pdf*.

Bernadini, C. (**2004**) AdA: the first electron-positron collider. *Phys. Perspect.*, **6**, 156.

Boussard, D. and Linnecar, T. (**1999**) The LHC Superconducting RF System, LHC Project Report 316, Presented at The 1999 Cryogenic Engineering and International Cryogenic Materials Conferece, 12–16 July 1999, Montreal Canada, published in Adv. Cryog. Eng. 45A, 835 (2000) *http://doc.cern.ch/archive/electronic/cern/preprints/lhc/lhc-project-report-316.pdf*.

Bromberg, C. *et al.* (**1977**) Observation of the production of Jets of particles at high transverse momentum and comparison with inclusive single particle reactions. *Phys. Rev. Lett.*, **38**, 1447; (b) Rohlf, J. (**1980**) Caltech Ph. D. thesis, Jet production in high-energy hadron–proton collisions. *Nucl. Phys. B.*, **171**, 1.

The CLIC home page *http://clic-study.web.cern.ch/clic%2Dstudy/*.

Cruikshank P. *et al.*, Mechanical Design Aspects of the LHC Beam Screen. Proceedings of the 1997 Particle Accelerator Conference, Vancouver, Canada, Vol. 3 p. 3386 (**1997**).

The DELPHI experiment, (**1980**) *http://delphiwww.cern.ch/*; The OPAL experiment, *http://opal.web.cern.ch/Opal/welcome.html*; The ALEPH experiment, *http://aleph.web.cern.ch/aleph/*; The L3 experiment, *http://l3.web.cern.ch/l3/*.

Feldman, G. J. and Perl, M. L. (**1975**) Electron-positron annihilation above 2-GeV and the new particles. *Phys. Rep.*, **19C**, 233.

Foster, G. W. (**2000**) Instrumentation Concepts for the Very Large Hadron Collider (VLHC), FERMILAB-Conf-00/189-E.

Hanson, G. *et al.* (**1975**) Evidence for Jet structure in hadron production by $e^+ e^-$ annihilation. *Phys. Rev. Lett.*, **35**, 1323.

http://www-bdnew.fnal.gov/tevatron/ (**2000**).

http://www-cdf.fnal.gov/ (**1986**).

http://www.fnal.gov/pub/now/tevlum.html (**2001**).

The ILC homepage *http://www.linearcollider.org/cms/*. (**2007**).

The LHC Design Report, Vol. 1–3 CERN-2004-003 (**2004**) Available at *http://ab-div.web.cern.ch/ab-div/Publications/LHC-DesignReport.html*.

The LHC home page contains a great deal of technical information about the LHC: *http://lhc.web.cern.ch/lhc/*.

Livingston, M.S. (**1969**) *Particle Accelerators: A Brief History*, Harvard University Press, Cambridge.

Panoksky, W.K. (**1997**) *The Evolution of Particle Accelerators and Colliders*, SLAC Beam Line Vol. **27**, No. 1 *http://www.slac.stanford.edu/pubs/beamline/27/1/27-1-panofsky.pdf*.

Perl, M.L. (**1995**) *Reflections on the Discovery of the Tau Lepton*, Nobel Prize Lecture, Dec. 8.

Redhead, P.A. Hobson, J.P. and Kornelsen, E.V. (**1993**) *The Physical Basis of Ultrahigh Vacuum*, American Institute of Physics, New York.

Richter, B. (**1971**) *Colliding Beams: Present Status; and the SLAC Project*, SLAC-PUB-886.

Richter, B. (**1976**) *From the Psi to Charm - the Experiments of 1975 and 1976*, Nobel Prize Lecture, Dec. 11.

Rohlf, J. (**2003**) *Super-LHC: The Experimental Program*, talk given at the International Workshop on Future Hadron Colliders *http://conferences.fnal.gov/hadroncollider/talks/rohlf.pdf* .

Rohlf, J. (**2005**) Physics with Jets at the LHC. *Acta Phys. Polon.*, **26**, 469.

Rubbia, C. (**1985**) Experimental observation of the intermediate vector Bosons W^+, W^-, and Z^0. *Rev. Mod. Phys.*, **57**, 699.

Schmidt, R. (**2003**) Operating with Unprecidented Stored Beam Energy, prepared for The Chamonix XII LHC Project Workshop, 3–8 March, 2003, Chamonix, France, *http://ab-div.web.cern.ch/ab-div/conferences/Chamonix/chamx2003/PAPERS/005.pdf* .

Stanford Linear Accelerator Center (**2002**) Celebrating Forty Years: A Photo History, SLAC-R-605.

Thomson, J.J. (**1897**) Cathode rays. *Philos. Mag.*, **44**, 293.

Van Der Meer, S. (**1985**) Stochastic cooling and the accumulation of antiprotons. *Rev. Mod. Phys.*, **57**, 689.

Yao, W.-M. *et al.* (**2006**) The review of particle physics. *J. Phys.*, **G 33**, 1.

Further reading

The CLIC Study Team (**2000**) *A 3 TeV* e^+e^- *Linear Collider based on CLIC Technology*, (ed. G. Guignard), CERN 2000-008.

23

Tracking in Particle Detectors

A. Acevski, Thomas C. Meyer, and Dieter Röhrich

23.1 Introduction 713
23.2 Particle Detectors 714
23.2.1 Gaseous Detectors in Proportional Mode 714
23.2.1.1 The Ionization Chamber 715
23.2.1.2 The Proportional Wire Chamber 715
23.2.1.3 The Drift Chamber 719
23.2.1.4 The Parallel-Electrode (Plate) Chamber 720
23.2.1.5 The Multistep Chamber 721
23.2.1.6 The Time Expansion Chamber 722
23.2.2 Proportional Detectors for High-Rate Applications 723
23.2.2.1 Micropattern Gaseous Detectors 723
23.2.2.2 The Straw-Tube Chamber 729
23.2.3 Gaseous Detectors in Nonproportional Mode 730
23.2.3.1 Limited Streamer Mode 730
23.2.3.2 The Gaseous Pixel Chamber 731
23.2.3.3 The Geiger–Müller Counter 732
23.2.3.4 The Spark Chamber 732
23.2.3.5 The Resistive Plate Chamber (RPC) 733
23.2.4 Solid-State Detectors 734
23.2.4.1 The Silicon Sensor 735
23.2.4.2 The Silicon Strip Detector 736
23.2.4.3 The Silicon Drift Chamber 736
23.2.4.4 The Silicon Pixel Detector 737
23.2.4.5 The Gallium–Arsenide (GaAs) Detector 745
23.2.4.6 The Diamond Detector 745
23.2.5 Visual Tracking Detectors 746
23.2.5.1 The Cloud Chamber 746
23.2.5.2 The Bubble Chamber 746
23.2.5.3 The Streamer Chamber 747
23.3 Tracking Systems 748
23.3.1 The Time Projection Chamber 748

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

23.3.1.1 The ALEPH TPC 749
23.3.1.2 TPCs for High-Energy Nucleus–Nucleus Collisions 749
23.3.2 The Silicon Tracker and Vertex Detector 755
23.3.3 Large-Volume Muon Trackers – the ATLAS Muon Spectrometer 757
23.3.3.1 Precision Chambers for Tracking 759
23.3.3.2 Trigger Chambers 760
23.4 Applications of Particle Detectors in the Future 763
23.4.1 Future Trends 765
Glossary 765
References 768

23.1
Introduction

Particles can be emitted in a variety of physical processes. They differ by their energy, the way they interact with other particles, and their mass. In the majority of cases, they are detectable by their electromagnetic interactions, directly if they are charged, and by their decay products if they are uncharged (the latter case applies to the conversion of photons into an electron or electron–positron pair). In a few cases, detectors are directly sensitive to the energy deposited in a medium – for example, the bubble chamber and the superconductive bolometer.

Charged particles are detected when crossing solid, liquid, or gaseous media by the ionization they produce or by the production of photons as in scintillation, Cherenkov light, and transition radiation. This can be exploited in electronic detectors to obtain information on the particle's energy, velocity, and charge. Normally, in the field of high-energy physics, a combination of detectors must be used to identify particles, since each detector can only measure a single parameter.

The types of detectors that the physicist has at his disposal are ionization chambers, wire chambers (in proportional and nonproportional mode), drift chambers, multistep avalanche chambers, micropattern gaseous detectors, scintillation counters, liquid detectors, solid-state detectors, Cherenkov counters, and transition–radiation detectors. While most detectors deliver a signal for a specific particle interaction, it is sometimes necessary to have a full image of particle trajectories in various media. For this, tracking devices play a major role in today's nuclear and subnuclear physics and have been the backbone of large detectors in the past. The development of particle detectors started with the earliest cloud chambers and progressed to nuclear photographic emulsions and then bubble chambers, spark chambers, and streamer chambers, and finally reached the combinations of time projection chambers and silicon trackers used at present.

Most applications of tracking detectors are in magnetic spectrometers. Tracking systems are placed, for example, either in front of/behind or inside dipole magnets, e.g., BRAHMS (Adamczyk *et al.*, 2003), PHOBOS (Back *et al.*, 2003), LHCb (Augusto Alves *et al.*, 2008), or inside large solenoidal fields, e.g., STAR (Ackermann *et al.*, 2003), ATLAS (inner barrel) (Aad *et al.*, 2008), CMS (Chatrchyan *et al.*, 2008), ALICE (Aamodt *et al.*, 2008). Following the trajectories of charged particles in three

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

dimensions inside a known magnetic field determines the momentum of the particle.

With the continuously increasing complexity of particle collisions being examined at present, particle tracking becomes an essential part of the field of particle and nuclear physics, with applications in astrophysics, nuclear imaging in the medical profession, biological analysis, and even archeology.

23.2
Particle Detectors

Particle detectors (with the exception of photographic imaging detectors) deliver an electrical signal after they interact with charged particles. Stable neutral particles such as photons and neutrons are only detectable via an intermediate conversion step, in which their kinetic energy is used to produce secondary charged particles in a nuclear (neutron–proton recoil) or electromagnetic (photon–electron/positron conversion) interaction. These detectors are used to count the number of particles emitted by unstable nuclei or produced in nuclear collisions, to measure their energies or trajectories, and to evaluate their masses. The type of detector depends on the intrinsic properties of the particles to be detected, but generally, it is the electromagnetic interaction of the particle with its surroundings that is observed; typically, the interaction involves the production of electron–ion pairs in a medium. The result may be current pulses in wire chambers, electron-hole production in solid-state devices, photons in a scintillating material, Cherenkov light, or transition radiation.

The electric fields in gaseous detectors are controlled by electrode configurations and are not always of uniform intensity. The nature of the gas filling is strongly dependent on the applications foreseen for a detector. Furthermore, the nature of the interaction or the radiation to be detected may be a factor leading to the choice of some components. For instance, in order to detect slow neutrons with high efficiency, BF_3 or ^{3}He is used. For soft X-rays, xenon or krypton fillings are favorable. For charged particles, the criterion is sometimes the maximum permissible gain in order to simplify electronics. Because of its very low cost, argon is often used as the major component. The additional gases (methane, alcohol, CO_2, etc.) have various functions: quenching of the metastable states of the noble gases that are responsible for UV photon emission, reduction of the resolution time by increasing the drift velocity, enhancement of the visible photon emission when the optical detection of the avalanches is required, reducing the ageing rate due to polymerization of the vapors, and more. The average energy loss of a charged particle in a gas varies between 20 and 35 eV per liberated ion. In large systems, the gas filling is flushed continuously.

Photosensitive gases, used as an additive in a main gas carrier, have extended the field of applications for gaseous detectors. Triethylamine (TEA), for example, has an ionization potential of 6.5 eV. It has a peak quantum efficiency for UV photons larger than the best photocathodes used in photomultipliers, and it permits the detection and imaging of photons emitted by Cherenkov radiation or by scintillators like BF_3, with spatial accuracies close to 1 mm.

23.2.1
Gaseous Detectors in Proportional Mode

When a charged particle crosses a gaseous medium, it can collide with the atoms of that gas, and if it can transfer more than the electron-binding energy, electrons can be liberated from the medium. These can either stay free or be rapidly recaptured by parent ions or atoms with (unwanted) electron affinity, such as oxygen.

The introduction of an electric field in the detection region leads to many other effects that need to be taken into account in the design of detectors or the selection of detector types. These are determined by the nature of the gas and the field intensity. The presence of a field leads to two possibilities:

1. If the field is strong enough, electrons will be sufficiently accelerated between collisions to knock another electron out of the next atom they hit, ultimately causing an electron avalanche.
2. If the field is not sufficiently strong, electrons freed from the material will only collide with atoms and drift along field lines until they reach the field anode, where they are captured.

The average velocity of these drifting electrons is determined by the properties of the gas and the strength of the electric field; Figure 23.1 shows electron drift velocity versus electric field for four different gases (Sauli, 1977). This finite velocity results in a time delay between emission and detection of a given electron, which is used to determine the point at which the electron was liberated (see Section 23.2.1.3).

For further discussion, see Palladino and Sadoulet (1975) and Sauli (1977).

23.2.1.1 **The Ionization Chamber**

In an ionization chamber, a weak homogeneous electric field is used to collect the electrons at the anode, without their undergoing charge multiplication (Figure 23.2). The amount of charge collected gives a measure of the average radiation intensity and the energy loss (dE/dx) that the particles have suffered in traversing the gas gap of the chamber. This instrument is widely used in dosimetry and calorimetry (where a liquid, such as liquid argon (LAr), liquid xenon (LXe), or tetramethyl pentane (TMP) Ankoviak *et al.*, 1989, is often used instead of a gas). As a result of not relying on gas gain, it has a very stable and uniform response and can measure average fluxes with an accuracy of 10^{-4}.

Ionization chambers can detect a single particle if that particle has enough energy and gas volume to liberate many electrons. Depending on the quality of the amplifier used, the noise equivalent can be between 100 and 1000 electrons, so that at least this many must be produced for a detectable signal to be generated by the chamber.

Generally, ionization chambers have little or no spatial granularity (except in calorimetry) and are rarely used in tracking. They often serve as charge-integrating rate meters, such as accelerator beam monitors.

23.2.1.2 **The Proportional Wire Chamber**

The simplest proportional chambers (Figure 23.3) consist of a thin wire held taut between two points, along the axis of a hollow cylinder. The wire acts as the anode, and the cylinder, which is the cathode, is filled with gas that can be maintained at a pressure between 100 Pa and 10 MPa, although it is generally kept at atmospheric pressure for mechanical reasons. The parameters of size and wire potential are chosen to give the correct signal gain (and hence electron multiplication factor) close to the wire. The detector is strictly operated in a regime where the response of the detection cell is proportional to the primary charge produced within the sensitive volume.

The simple cylindrical wire counter is still used for limited applications where the detection of soft X-rays (in the range of 1–50 keV) is required. It delivers pulses

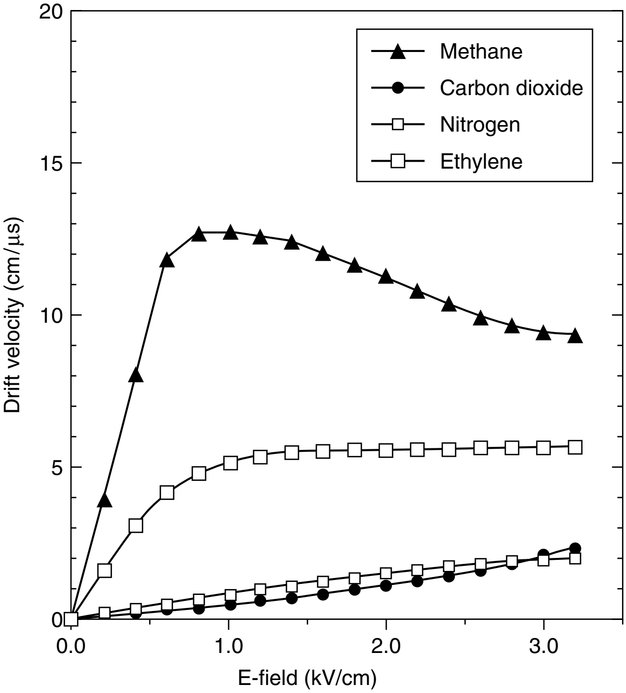

Fig. 23.1 Electron drift velocity for various gases.

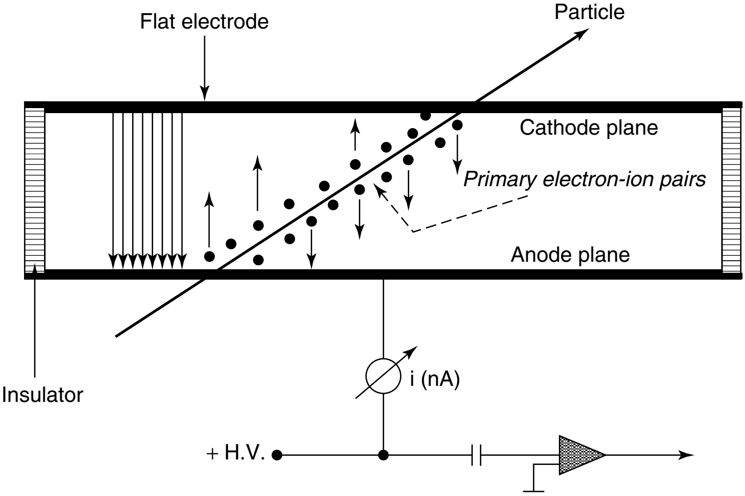

Flat electrode
Particle
Cathode plane
Primary electron-ion pairs
Anode plane
Insulator
i (nA)
+ H.V.

Fig. 23.2 The ionization chamber.

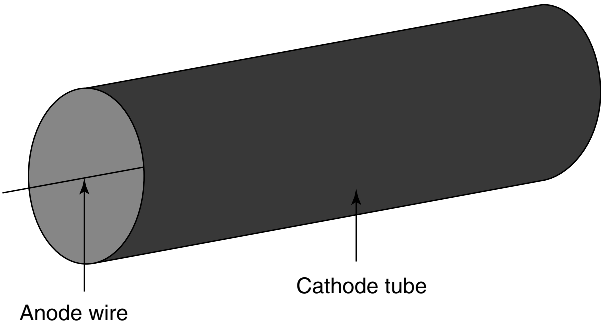

Fig. 23.3 The proportional wire tube.

with an energy resolution of 13% (full width at half maximum, FWHM) at 6 keV, varying like $E^{-1/2}$. This is much poorer than what can now be achieved with solid-state detectors (cf. Section 23.2.4), but the device is less expensive.

More elaborate detectors – for example, the multiwire proportional chamber (MWPC) (Charpak *et al.*, 1979) – have been developed from these simple chambers and are in use at CERN and other research facilities because of their versatility, good resolution, and large surface coverage (Figure 23.4). With the possibility of very accurate localization of ionizing events, they allow precise imaging for high-energy collisions.

In multiwire chambers, there are two electric field regions: one far from the wire, where the field is uniform, and the other close to the wire, where the field is radial. In the latter region, electrons experience inelastic collisions leading to an avalanche effect, which gives the detector its signal gain; that is, one electron outside this region leads to many electrons hitting the wire.

Typically, with mean free paths for the electrons of about 1 μm and proportional gains of $\leq 10^5$, the multiplication process occurs over a distance of about 15 μm. The electrons from the avalanche are rapidly collected by the wire, whereas the positive ions must drift across the full potential gap of the chamber before hitting the

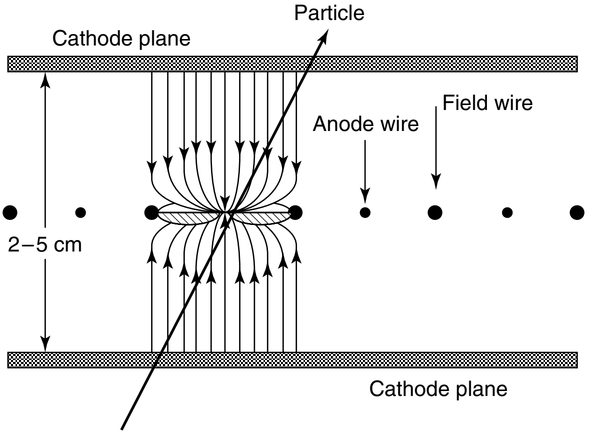

Fig. 23.4 The multiwire proportional chamber (MWPC).

cathode. The electrons generated in the avalanche result in a sharp current pulse followed by a steady decay because of the induced current from the receding positive ions. It is important to notice that the collection time of electrons induces a very small pulse on the wire since the induced charge is proportional to the potential drop experienced by the moving charge and, for electrons, it is only a minute fraction of the total applied potential difference. In contrast, the heavy positive ions have to cross the entire anode–cathode distance, and it is their motion that induces most of the observed pulse. The motion is fast close to the wire, because of the intense electric field, and slow after a distance of a few tens or hundreds of microns.

The resolution time of a proportional counter, that is, the accuracy with which one can ascertain the initial time of the ionization process, is equal to the fluctuations in the time required for an electron liberated in the gas to reach the multiplication region. It may be of the order of 0.5 microseconds for a counter of 10-mm diameter. This resolution time is much larger than that reached by scintillation counters.

In planar geometries, sense (anode) and field (shaping) wires are interspaced at regular distances (in the order of millimeters) between an upper and lower cathode plane (Figure 23.4). This arrangement builds individual and isolated detection segments with a sense wire at the center. The cell size, also known as the *wire pitch*, determines the potential accuracy of the detector. In such planar geometries, an ionizing event can be measured with an accuracy of 10^{-2} times the wire spacing. This is done by using the values of the current pulses measured from neighboring field wires to construct a Gaussian envelope, the peak of which lies at the position of the ionizing event. To obtain information *along* the wire in the same chamber, two methods are most commonly employed in MWPCs:

1. *Cathode segmentation*: if the cathode surface of the detector is divided into strips or pads along the wire, the charges induced on individual cathode segments can be measured as functions of distance from the electron-wire impact. The mutual sharing results in a Gaussian-shaped distribution whose peak position describes the impact point with very high precision (order of tens of microns). This method, however, requires adequate granularity and yields a large number of electronic channels. This property is essential for the imaging of soft radiations, such as X-rays or neutrons, when the two coordinates have to be determined from a single totally absorbing interaction.
2. *Charge division*: if a resistive alloy (several kiloohms per meter) is used for the sense wire, signals detected at *both* ends of the wire will suffer different attenuation depending on the distance traveled. The charge asymmetry can then be used to determine the avalanche impact point on the wire. The precision of this method crucially depends on the gain (signal/noise ratio (S/N)) and is typically 1% of the total wire length.

MWPCs have shaped more than three decades of particle tracking. No other detector has ever acquired the same versatility, experience, and confidence in operational safety, even at present system standards. In maintaining the basic operating principles they have gradually paved the way to more elaborate and sophisticated devices that will be discussed in the following sections.

For further discussion, see Charpak *et al.* (1979) and Walenta (1983).

23.2.1.3 The Drift Chamber

The drift chamber is an extension of the proportional wire chamber in which the position of the ionizing event normal to the plane is found by examining the drift time of the electrons. This is done by using an external triggering device to start a clock and exploiting the fact that the electrons have constant drift velocity until they enter the nonuniform field region close to the sense wires. The signal, which is proportional to the ionization of the track segment as projected on the receiving sense wire cell, stops the clock.

Most drift chambers operate with the drift (E) field oriented transverse or parallel to the direction of a magnetic bending field (B). Therefore, nonparallel E and B fields strongly influence the velocity and direction of drifting electrons, usually leading to loss of efficiency and resolution (Schmidt, 1980; Atac *et al.*, 1986). The time resolution of such a chamber also depends on how much the electrons diffuse away from the ionization cluster while they undergo drifting, which in turn depends on the magnetic field and combined electric and magnetic effects. However, choosing suitable gases (fast or slow) in combination with high drift fields can restore chamber performance.

Planar Drift Chambers A typical drift-chamber cell is shown in Figure 23.5. Using inexpensive electronics, which permits reaching an accuracy of the order of 1 nanosecond in the measurement of the drift time (corresponding to an order of 100 μm for the position of the initial electron), it is easy to reach the same spatial resolution as that the wire planes achieve in the proportional chamber. The drift distance can also be large (of the order of 10 cm), which reduces the need for large numbers of sense wires, helping to cut the cost of such detectors.

Large planar chambers have dimensions of 163 cm × 556 cm (width × length) (Adeva *et al.*, 1990), and large areas of up to 600 m^2 (Aarnio *et al.*, 1991) have been instrumented. The spatial resolution attainable with large area chambers is

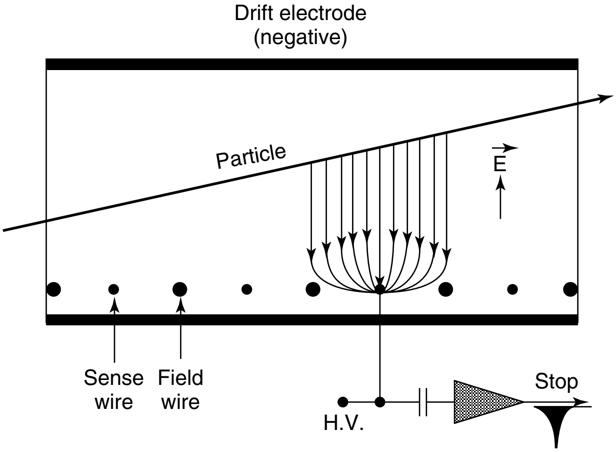

Fig. 23.5 The drift chamber.

mostly limited by mechanical tolerances in positioning wires (100–300 μm) and by the sagging of wires under their own weight.

For further discussion see Breskin *et al.* (1975) and Sauli (1977).

Cylindrical Drift Chambers Solenoidal magnetic fields around the beam pipe are usually used in colliding-beam experiments, which necessitate cylindrically symmetric layers of drift cells. The wires generating the E-field run parallel to the B-field or have a small stereo angle with respect to the B-field axis; field wires that are parallel to each other surround a single sense wire. Cylindrical drift chambers were and still are widely used in colliding-beam experiments like BABAR (Sciolla, 1998) and BELLE (Uno, 1996).

23.2.1.4 The Parallel-Electrode (Plate) Chamber

Parallel-electrode or plate chambers produce uniform electric fields in which electron multiplication can take place nonetheless, but it is more difficult to achieve high gains with this type of detector without the strong radial fields created by wires. The gain of such a detector is proportional to an exponential function of the distance from the ionization point to the anode. The result is that this type of detector has been preferred for particles that do not require high gains (i.e., strongly ionizing particles). However, for average gas gains of 5×10^4 over a 1-mm gap, an efficiency of 90% for minimum-ionizing particles (MIPs) can be maintained. Moreover, these chambers have a fast response and an excellent time resolution of a few nanoseconds, since all electrons from an avalanche will arrive at the anode at the same time, eliminating the time-dispersion characteristics of the wire chamber, in which the charge multiplication is confined to a small region near the wire (Figure 23.6). As a result of this, the parallel-plate detector can sustain very high rates of charged particles (up to $10^7\,\text{cm}^{-2}\,\text{s}^{-1}$) and, because of its highly polished conducting surface, will provide radiation resistance 100 times larger than conventional wire chambers. This, combined with geometrical flexibility and low-production costs, makes the device

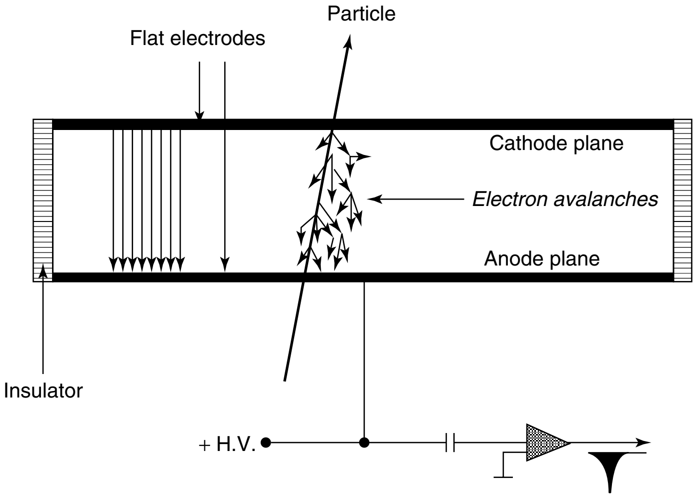

Fig. 23.6 The parallel-plate chamber.

suitable for operation at high luminosity and large backgrounds, such as the forward regions of detectors at hadron colliders.

23.2.1.5 The Multistep Chamber

Parallel-plate counters have experienced a renewal of interest for two reasons: the introduction of multistep structures (Figure 23.7), and the fact that this structure is most favorable for the imaging of avalanches from light emitted by some vapor additives in the counter gas.

The multistep chamber allows the use of more than one stage of gas amplification: a wire grid is used as the anode for the first stage of amplification, where the electrons from the avalanche are extracted and transferred to a second parallel-plate structure where they start a second round of avalanches. This technique can achieve gains of 10^6 or even 10^7, with fewer secondary effects such as sparking at the anode, which produces undesirable ultraviolet radiation or positive ions. This high level of gain easily allows the detection of a single electron, making this type of detector ideal for measuring weakly ionizing particles and imaging vacuum ultraviolet photons emitted by Cherenkov radiators.

For further discussion, see Breskin *et al.* (1980), Bouclier *et al.* (1983), and Charpak and Sauli (1984).

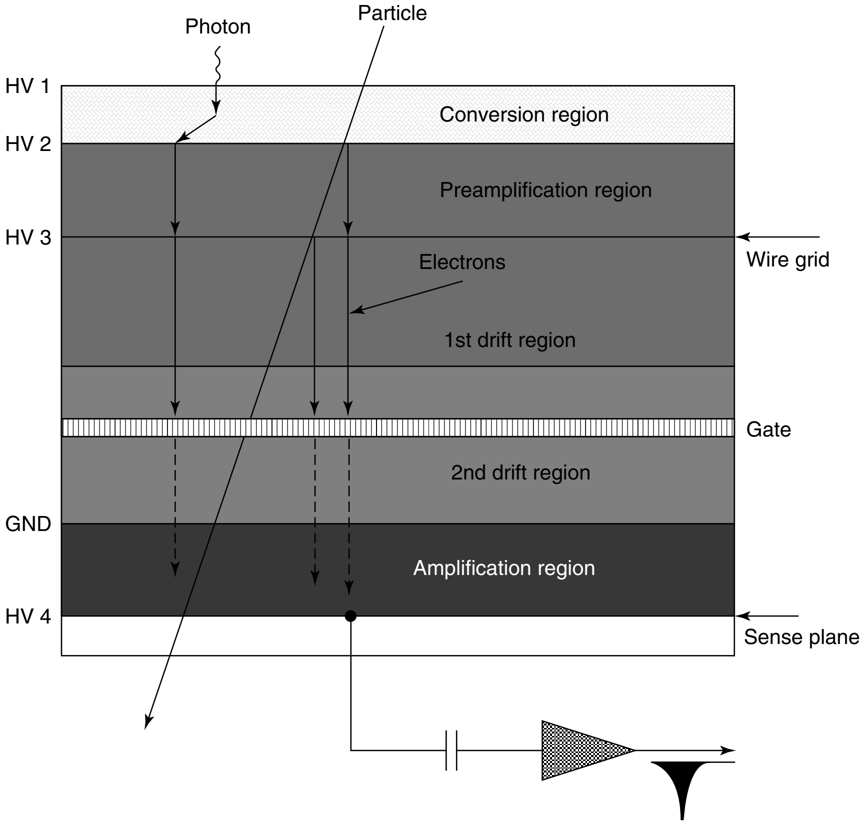
Photon
Particle
HV 1
Conversion region
HV 2
Preamplification region
HV 3
Wire grid
Electrons
1st drift region
Gate
2nd drift region
GND
Amplification region
HV 4
Sense plane

Fig. 23.7 The multistep chamber.

23.2.1.6 The Time Expansion Chamber

The internal structure of the time expansion chamber (TEC) is very similar to that of the MWPC, but the principle of operation is slightly different (Figure 23.8). The electrons drift much more slowly in the drifting region than in the multiplication region. Usually, these two regions are separated by another wire grid held at ground. Using this technique, it is possible to improve the resolution dramatically, since there is a much longer time interval between successive ionized clusters. The mean drift-time interval from subsequent single-ionization clusters of a relativistic particle in the TEC can be made large compared to the width of the anode signal to allow the recording of the clusters separately. In other detectors, these usually appear as one long ionized track and cannot be resolved because they arrive too quickly. Tracks that are parallel to the wire plane are resolved by reducing the wire spacing. The result is that one can distinguish individual ionized clusters as opposed to track segments made up of many clusters. Basic ionization phenomena and drift properties can be measured at the single-electron level.

The problems associated with this type of chamber arise from the relatively long drift time of some of the clusters, which will tend to fade out through diffusion and electron capture by electronegative impurities at the parts per million levels, for example, oxygen. However, in TEC chambers, a "cool" gas is used giving a low drift velocity ($v_d \sim E/p$) and resulting in low diffusion of $a/x^{1/2} \sim 100\,\mu\text{m cm}^{-1/2}$. The effect of diffusion can be further reduced if, instead of measuring the arrival of the first electron (leading-edge discrimination), the centroid of the nearly Gaussian electron cloud of n electrons can be measured. In that case, one ideally obtains $\sigma = \sigma_0/n^{1/2}$,

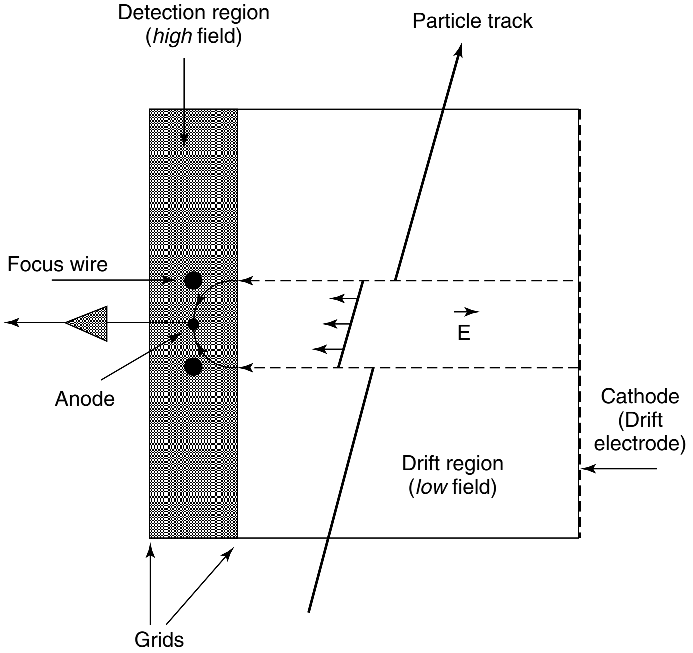

Fig. 23.8 The time-expansion chamber.

where σ_0 is the single-electron diffusion. The transition from the low-field to the high-field region is also a point of concern. The resulting deformation of the arriving electron cloud by the converging field lines in the amplification region can normally be sharpened by altering the choice of gas, chamber geometry, and applied voltages so as to define a narrow amplification region.

For further discussion, see Walenta (1979) and Anderhub *et al.* (1985).

23.2.2

Proportional Detectors for High-Rate Applications

For proportional gas detectors to cope with very high particle rates (i.e., $>10\,\mathrm{MHz\,cm^{-2}}$), the detector cell area must be sufficiently small to avoid signal pileup (Walenta, 1983). This effect is further enhanced by the long drift times of electrons and the very low mobility of positive ions generated in the charge avalanche. Unfortunately, in attempting to achieve very short drift times, the drift volume of gaseous chambers cannot be made arbitrarily thin, because the ionization of gas atoms by charged particles proceeds via a nonuniform clustering process. This limits the efficiency and operation of any gas-filled detector. On the other hand, positive ions copiously produced at the anode migrate to the cathodes at a speed more than a thousand times slower than electrons. Their presence in the drift space reduces the effective fields. The effect of this is twofold: a change of drift velocity for electrons and a reduction of the amplification field near the anode.

The best way to overcome these drawbacks is to reduce the detector cell size to submillimeter dimensions. Fast gases with electron drift velocities of more than $100\,\mathrm{\mu m\,ns^{-1}}$ can help for certain (nonmagnetic) applications as well. However, these gases operated in a magnetic field face the problem of large drift angles for electrons subject to the Lorentz force (Atac *et al.*, 1986; Schmidt, 1980).

23.2.2.1 Micropattern Gaseous Detectors

Modern photolithographic technology has enabled a series of novel micropattern gaseous detector concepts, revolutionizing cell size limits. By using a smaller feature size compared to classical gas counters, these detectors offer an intrinsic high rate capability (fine pitch and fast collection of positive ions) and an excellent spatial resolution ($\sim 30\,\mathrm{\mu m}$) (Titov, 2007).

For further discussion, see Sauli and Sharma (1999).

The Microstrip Gas Chamber (MSGC) Microstrip chambers are, as the name implies, made of very thin, narrow, conducting strips that lie on a thin and flat glass or ceramic substrate similar to printed circuit boards (PCBs). For more ‘‘flexible’’ applications, plastic foils made of Kapton or Tedlar with similar strip designs have been shown to work satisfactorily in nonplanar, say cylindrical, detector geometries (Bouclier *et al.*, 1992b). These strips, whose pitch is $\leq 200\mathrm{\mu m}$, replace the anode and cathode wires in the traditional MWPC, which normally uses much larger wire separation. Similar to the induction drift chamber, the drift region of this detector lies between these strips and the cathode plane (2–5 mm above) as shown in Figure 23.9. The strips are formed by electron-beam or photolithography, which can give a spatial accuracy of $0.1\,\mathrm{\mu m}$ for the position of the strips. Despite a modest potential difference of 200–500 V between the anode and cathode strips, a very high anode field of $100\,\mathrm{kV\,cm^{-1}}$ is produced (Florent *et al.*,

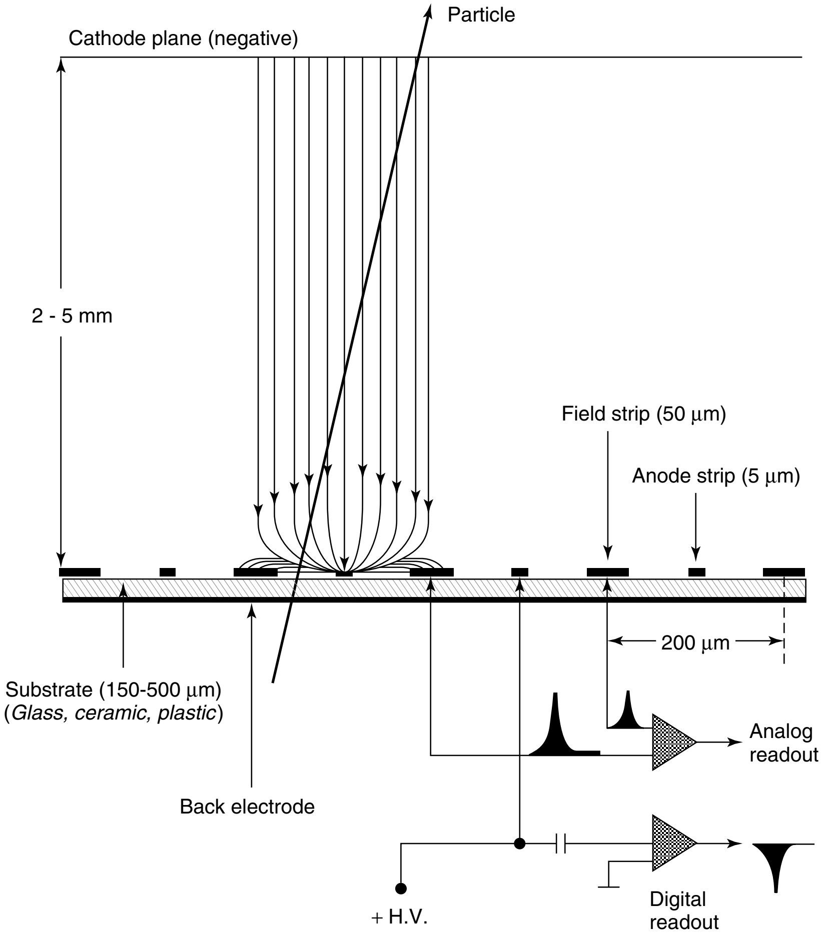
Particle
Cathode plane (negative)
2 - 5 mm
Field strip (50 μm)
Anode strip (5 μm)
200 μm
Substrate (150-500 μm)
(*Glass, ceramic, plastic*)
Back electrode
Analog
readout
+ H.V.
Digital
readout

Fig. 23.9 The microstrip gas chamber.

1993). Apart from the close separation of electrodes, this must be attributed to the edges of the strips and the high dielectric constant of the insulating substrate. Thus, almost the entire gain is established between the anode and cathode strips with almost no effect from the upper cathode plane. The result is a very homogeneous proportional gain ($\leq 10^4$) giving rise to an excellent energy resolution (12% FWHM for 6-keV photons), and a high rate capability ($\geq 10^7$ events/(cm^2 s^{-1})) due to the short drift path of the positive ions.

The small detection cell lends itself to another important feature of this chamber: spatial resolution. In the classical readout mode, only the anode strips are read out, resulting in a modest space resolution comparable to the cell size divided by $\sqrt{12}$, that is, $\approx 60\,\mu$m at 200-μm cell size. A more elaborate method makes use of the induced signal on the field or cathode strips next to the signal-producing anode strip (cf. Section 23.2.1.2). Using carefully calibrated analog electronics, the relative charge sharing between the two cathode strips allows a weighted interpolation inside the cell width with a precision of better than 30 μm. It should be noted that the resolution of this detector is spoiled by angular tracks, where the charge is deposited over several detection cells (Angelini *et al.*, 1992).

To exploit the full measuring potential of microstrip gas chambers (MSGCs), the back electrode of the substrate (see Figure 23.9) can also be segmented into strips or pads of any suitable shape, which, in analogy to silicon double-sided detectors (refer to Section 23.2.4.1), can be used for the readout of the second hit coordinate. The capacitive coupling of the anode signal into the back plane is achieved by placing the back electrodes into close proximity (separated only by an insulating SiO_2 film of ~1 μm) to the front (anode–cathode) plane (Angelini *et al.*, 1993b). In these so-called double-sided MSGCs, the back-plane potential not only ensures an optimum field configuration for avoiding charge buildup (as for single-sided MSGCs) but also plays an essential role in the gain mechanism because of its closeness to the anode.

Gases with high cluster statistics and drift velocity must be chosen for operation in high-rate environments. Two gases that simultaneously satisfy these constraints are dimethyl ether (DME)-CO_2 (60 : 40) and Xe-DME-CO_2 (30 : 30 : 40). Efficiencies are maintained at nearly 100%, even for high discriminator thresholds ($\geq 5 \times$ noise level). Under operating fields close to $10\,\mathrm{kV\,cm^{-1}}$, these gases reach drift speeds of $70\,\mu\mathrm{m\,ns^{-1}}$.

Early test samples operated at low rates have shown significant drops in gain (>30%) within a few hours (Figure 23.10), most likely due to surface charging as a consequence of the high surface and bulk resistivity of the detector substrate (Bouclier *et al.*, 1992a). The applied potentials on the substrate favor the depletion of charge carriers within the bulk, thus further increasing the resistivity of the material. By choosing substrates of low bulk resistivity (i.e., $10^{10}\ \Omega\,\mathrm{cm}$), this effect can be avoided or substantially reduced.

MSGCs exposed to very high rates of ionizing radiation experience flux-dependent dynamic gain modifications. As in the

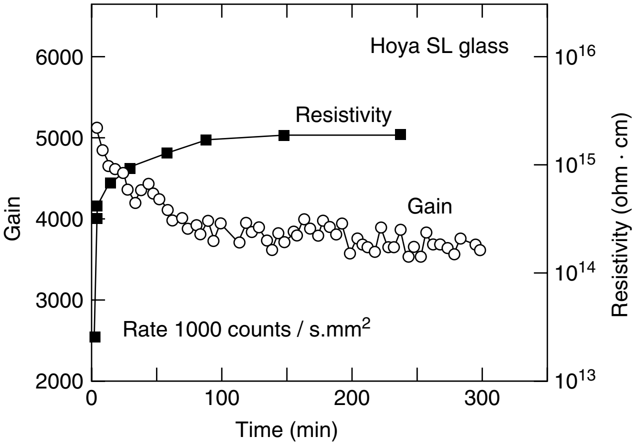

Fig. 23.10 Gain drop in a microstrip gas chamber after turning on the voltage and simultaneous rise in bulk resistivity for a substrate made of Hoya SL glass.

case of spontaneous charge buildup, these can be avoided by a proper choice of a low-resistivity support and suitably chosen bias voltages. Despite gain changes, the detectors can be safely operated at fluxes up to $10^8\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ (Bouclier *et al.*, 1993). More serious is the phenomenon of aging in MSGCs as observed by Bouclier *et al.* (1992a) leading to an unrecoverable loss of performance due to permanent damage of the electrode area. The mechanisms seem to be comparable to those observed in ordinary proportional chambers. Therefore, similar precautions must be taken in the selection of material and counting gases with very low impurity concentrations.

The hadron-electron ring accelerator (HERA)-B experiment has used several hundred MSGC plates as part of the inner tracking system. Despite their impressive performance, the MSGC turned out to be easily damaged by discharges caused by imperfections in the photolithography or induced by heavily ionizing particles destroying the fragile electrode structure (Bagaturia *et al.*, 2002).

For further discussion, see Oed (1988), Angelini *et al.* (1991), Angelini *et al.* (1993a) and Sauli and Sharma (1999).

MICRO-MEsh GASeous structure (Micromegas) The MICRO-MEsh GASeous structure (Micromegas) is miniaturized from a very asymmetric two-stage parallel-plate avalanche chamber (Giomataris *et al.*, 1996). The detector consists of a thin metal grid (micromesh) stretched at a very small distance, 25–150 μm, above a readout electrode (strips/pads of conductor printed on an insulator board), shown in Figure 23.11 (Colas *et al.*, 2004). With a very high field applied across the gap, typically 50–70 $\mathrm{kV\,cm}^{-1}$, electrons released in the upper drift region are collected and multiplied.

Micromegas has many advantages; some of them are a fast signal, no $E \times B$ effect, a high gain, and a good energy resolution. It also has the capability to naturally stop most of the ions produced in the amplification space. Owing to the very large field ratio between the multiplication and the drift regions (as high as 400 or 500) the electric field lines are very much compressed between the two regions (‘‘funnel’’ effect). However, owing to collisions in the gas, electrons do not drift along the field lines. They diffuse, especially in the multiplication space: the transverse extension σ (standard deviation) of the avalanche due to diffusion is of the order of 10–15 μm, depending on the gas mixture, the electric field, and the gap width. A simulation by GARFIELD of the electron diffusion and the multiplication in the drift and the multiplication gaps is shown in Figure 23.12 (Colas *et al.*, 2004). The size of the electron cloud is much larger than that of the funnel end, but owing to their high mass ions are not submitted to diffusion and drift along the field lines. Assuming that they are emitted with the same distribution as the avalanche, most of them are naturally collected by the micromesh and only the fraction of ions created in the small funnel will flow back into the drift volume.

Fast signals (≤ 1 ns) are obtained during the collection of the electron avalanche on the anode microstrip plane. The fast evacuation of positive ions combined with the high granularity of the detector provides a high rate capability. Gas gains of up to 10^5 have been achieved (Giomataris *et al.*, 1996).

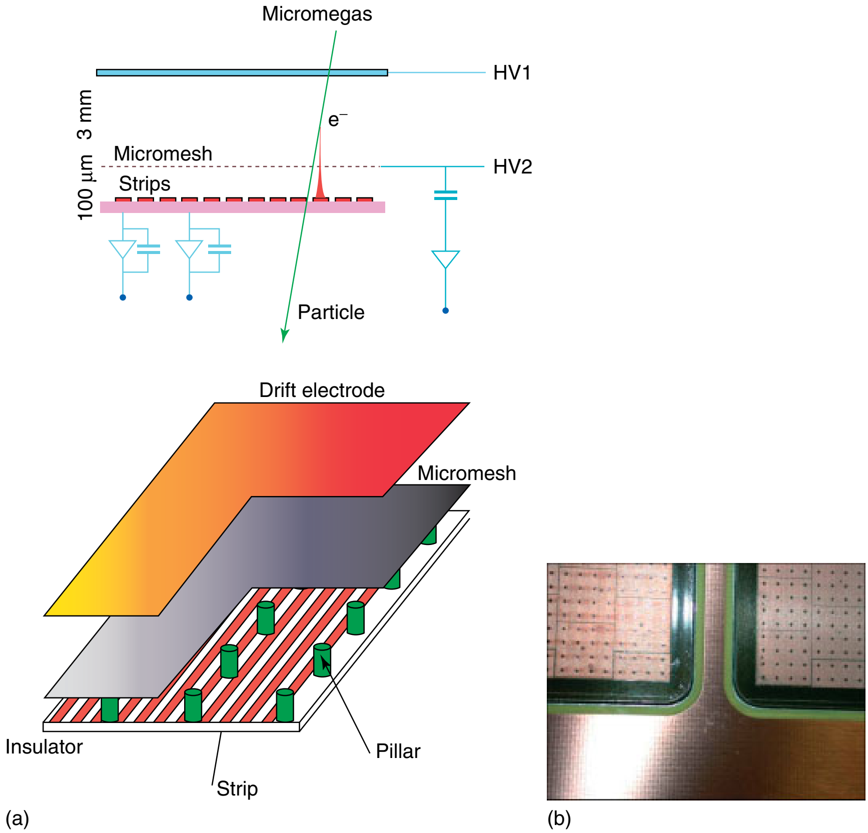

Fig. 23.11 Schematic drawing of the Micromegas detector (a) and photograph of the "Bulk" Micromegas detectors (b) (Giomataris *et al.*, 2006). Pillars of 400-µm diameter every 2 mm are visible (b).

Gas Electron Multiplier (GEM) The gas electron multiplier (GEM), introduced by Sauli in 1996 (Sauli, 1997), consists of a thin, metal-clad polymer foil chemically perforated by a high density of holes, typically 100 mm^{-2} (see Figure 23.13). As shown in panel (b) of the figure, with a suitable choice of voltages, all electrons released by ionization in the overlying gas layer are sucked into the holes, where charge multiplication occurs in the high electric field (50–70 kV cm^{-1}). Most of the electrons generated in the avalanches transfer into the lower region; the GEM foil acts as a charge preamplifier, preserving the original ionization pattern to a large extent. The gain is a property of the GEM structure and is only mildly affected by the external fields, considerably relaxing the mechanical requirements. The GEM manufacturing method, developed at CERN, is a refinement of the double-side printed circuit technology. The metal-clad polymer (Kapton) is engraved on both sides with the desired hole pattern; controlled immersion in a Kapton-specific solvent opens the channels in the insulator.

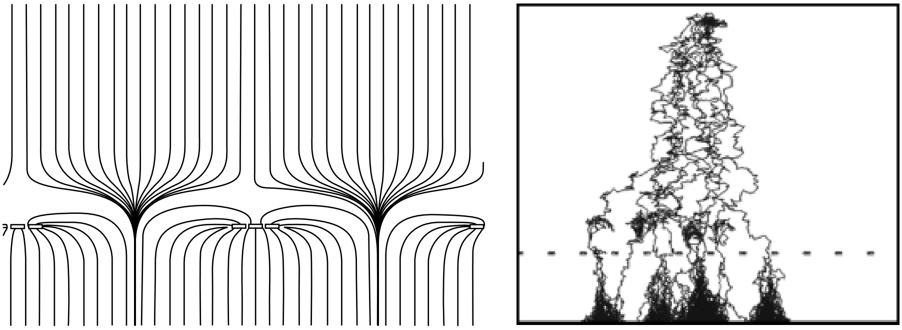

(a) (b)

Fig. 23.12 Field lines in Micromegas (a) and GARFIELD simulation of electron drift and multiplication (b) (Giomataris *et al.*, 2006).

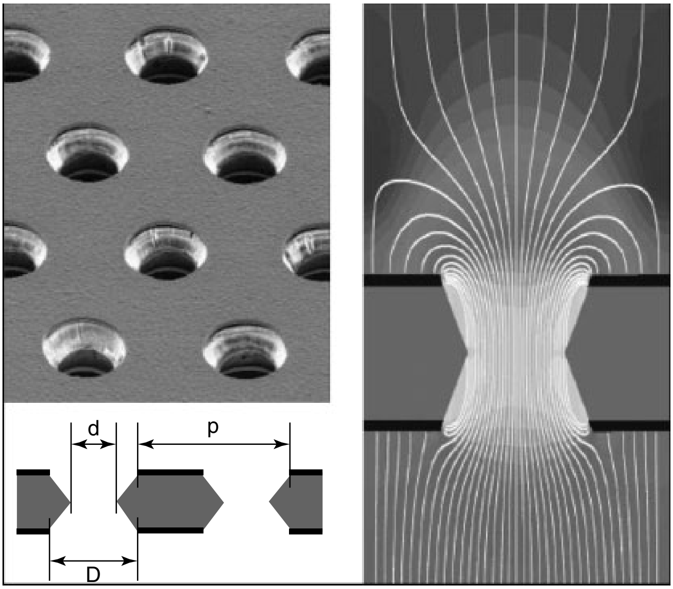

(a) (b)

Fig. 23.13 (a) Electron microscope photograph of a GEM in standard geometry ($p = 140$-μm pitch, $D = 70$-μm outer diameter and $d = 60$-μm inner hole diameter). (b) The plot of the electric field lines illustrates that electrons are guided into the holes and undergo proportional gas amplification in the locally strong electric field (Kappler *et al.*, 2004).

A unique property of the GEM detector is the complete decoupling of the amplification stage (GEM) and a simple PCB readout electrode, which operates at a gain of 1 and serves only as a charge collector. This offers some freedom in the optimization of the anode readout structure, which can be made of pads or strips of arbitrary pattern.

The maximum gain achievable with the GEM electrode depends on the thickness of the polymeric support, the diameter of the holes, the gas mixture, and the applied voltages. Proportional gains up to 10^4 have been achieved for single GEM devices. Cascading amplifying elements in a double or triple GEM (Figure 23.14a) can produce gains well above 10^5 (Figure 23.14b,) (Benlloch *et al.*, 1998a; 1998b, Büttner *et al.*, 1998).

Single and double GEM detectors with PCB readout have been extensively tested in the laboratory and in particle beams (Bressan *et al.*, 1999). Figure 23.15 shows the large efficiency plateau and the position resolution as a function of voltage obtained with a double GEM device. The signal detected on the strips is entirely due to the electrons, without a slow ion tail, and is typically a few tens of nanoseconds wide for a 1-mm-wide gap.

GEMs can also be easily bent to form cylindrically curved ultralight detectors, as preferred for inner tracker applications (Bencivenni *et al.*, 2007). Controlled etching of GEM foils (decreasing the thickness of the copper layer from 5 to 1 μm) allows to reduce the material budget in triple GEMs to $1.5 \times 10^{-3}\, X_0$, which is about one half of a 300-μm-thick Si-microstrip detector (Bondar *et al.*, 2006).

For further discussion, see Sauli (2004).

23.2.2.2 The Straw-Tube Chamber

As the name implies, straw-tube detectors are long (≥1 m) cylindrical chambers with a central wire for charge collection (Figure 23.3). Closely resembling the classical Geiger–Müller tube in basic geometry, they differ in size and operational mode. The radial size of a few millimeters (diameters mostly between 4 and 6 mm) is chosen to employ the detector in high-rate environments. Correct choice

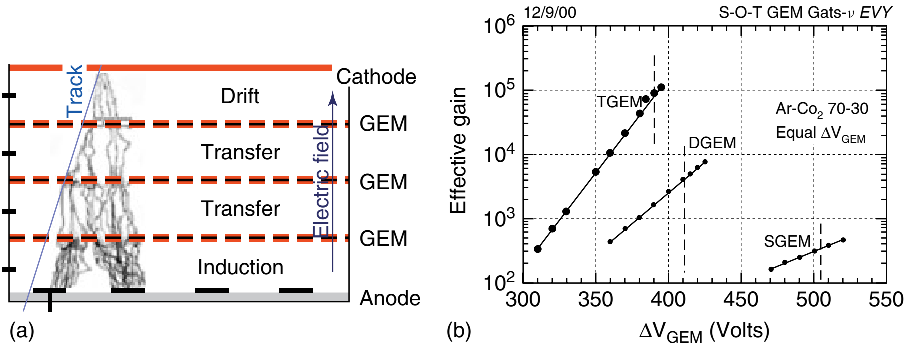

Fig. 23.14 Schematic drawing of the triple-GEM detector (Titov, 2007) (a) and (b) the gain (full curves) and discharge limits on exposure to alpha particles (dashed lines) of multiple GEM detectors (Sauli, 2004).

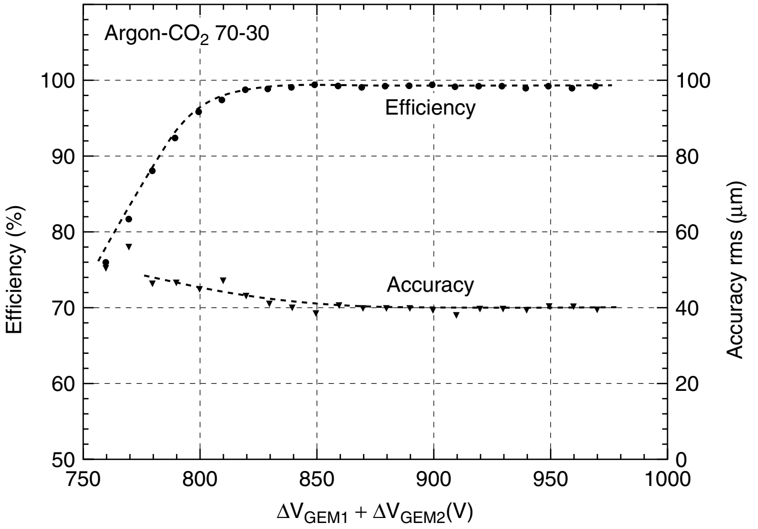

Fig. 23.15 Efficiency and position accuracy for minimum-ionizing particles measured with the double GEM plus printed circuit board detector, as a function of total potential difference applied to the multipliers, $\Delta V_{GEM1} + \Delta V_{GEM2}$.

of gas and electric field makes it work as a proportional counter that is fast, simple, and relatively radiation hard while not sacrificing the good spatial resolution ($\leq$100 μm) derived from electron drift time. Straws can equally well be used to form detector modules or layers in a geometry dictated by other experimental components (Figure 23.16). Typical problems encountered in long (>1 m) straws are the precise positioning of the sense wires and gravitational sag.

Straw-tube chambers have been widely used as vertex detectors (Frautschi *et al.*, 1991) and are being used as tracking detectors at the hadron colliders, Relativistic Heavy-Ion Collider (RHIC) and LHC. In comparison to the previously discussed microcell proportional chambers, straw tubes are still large-surface devices that benefit from high drift velocities and simple construction.

23.2.3
Gaseous Detectors in Nonproportional Mode

If, in a gaseous detector, the electric field is above a certain critical value the ultraviolet photons emitted in an avalanche will have sufficient energy to generate further ionization elsewhere in the chamber. This is determined by the choice of wire diameter and the gas composition in the chamber. If the appropriate gases (i.e., with a reduced concentration of organic quencher) that do not impede the propagation of the photons are used, then secondary avalanches can occur far from the initial one.

23.2.3.1 Limited Streamer Mode

In this mode of operation, avalanches occur at the wires but are not confined and extend *radially* out from them to form

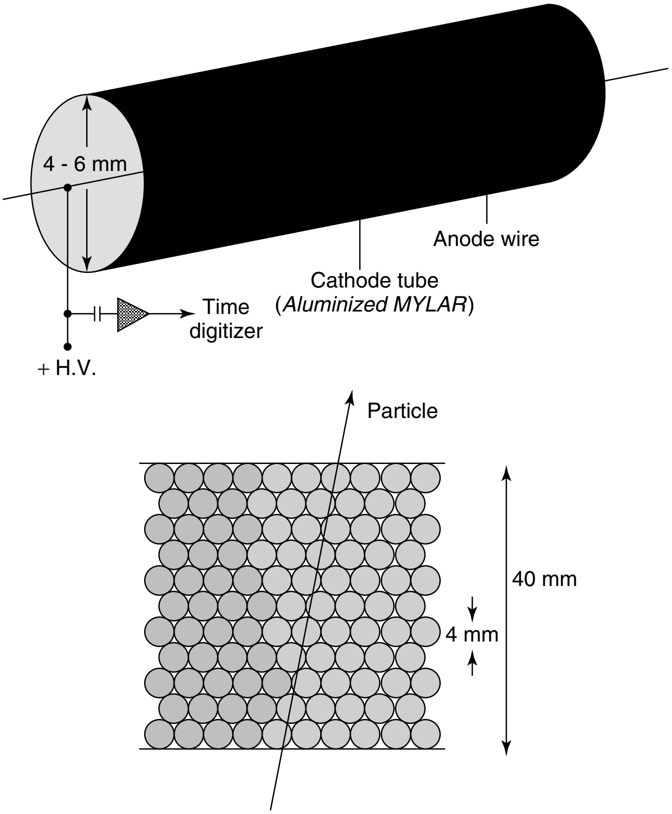

Fig. 23.16 A single straw tube and a straw-tube detector.

"streamers." However, the electric field is not high enough for a complete ionized channel to be formed between anode and cathode, which would lead to an unwanted spark. Depending on the geometry of the chamber, the streamer can finish close to one face or two streamers may be formed on neighboring wires that do not quite meet each other. This method of detection leads to "fast" pulses, which are 100 times larger than in the proportional mode of operation. However, the locally high concentration of positive space charge causes the effective field in the avalanche region of the wire to drop so low that during the migration of the positive ions toward the cathode, charge multiplication cannot occur. These are the typical dead regions formed in limited streamer mode detectors.

See Brehin *et al.* (1975) and Sauli (1977).

23.2.3.2 **The Gaseous Pixel Chamber**

The pixel chamber is based on etched PCBs instead of wires, using a small-scale modular approach, making it useful for moderate rate applications in which detector occupancy would not allow the use of large wire chambers. The accuracy of measurement is determined by the film printing and etching process used on the board; this means that relatively cheap chambers can be constructed with high levels of

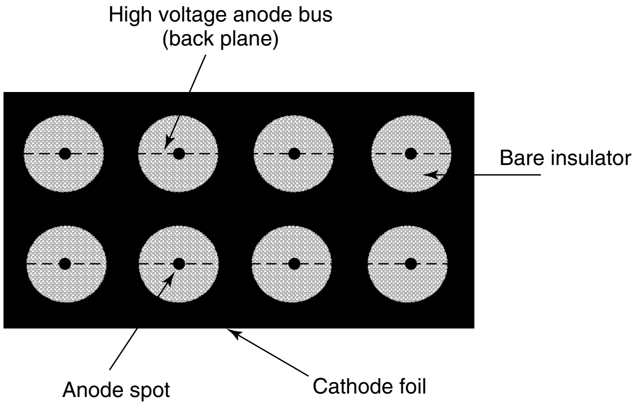

Fig. 23.17 The gaseous pixel detector (top view).

accuracy on a large scale. The chamber is constructed from a thin metal foil, used as a cathode plane, with small holes for the anodes, overlaid on a circuit board as shown in Figure 23.17.

A second foil is employed 8 mm above the anode–cathode plane, on which are mounted cathode strips for field shaping, as shown in Figure 23.18. These strips are linked together in one direction in the detector plane and the anode points are linked in the other direction in order to give a two-dimensional picture of each hit. The third spatial coordinate is obtained from the electron cluster drift time as determined by the choice of chamber gas, as in other types of chambers.

For further discussion, see Antreasyan *et al.* (1990).

23.2.3.3 The Geiger–Müller Counter

When one is using (‘‘counting’’) gases with essentially no photon-absorbing additives, electron avalanches will copiously produce photons that are not suppressed and, consequently, will lead to more charge multiplication along the wire (in contrast to radial streamers). This process will continue until the entire region is saturated with ions, causing the electric field near the wire to drop sufficiently to prevent further avalanches. The result is a charge at the output of the detector of almost constant value for every particle entering the detector. This tends to make counting particles easy, and this type of chamber was used extensively in the past for such measurements. The disadvantage of this counter is the long busy time experienced (1 microsecond approximately), which is why it has been superseded, for the most part, by solid-state devices. The Geiger–Müller counter can also be used in proportional mode. In this mode, the dead time is the same but is limited to a region of about 0.1-mm length instead of extending over the full length of the wire.

For further discussion, see Charpak and Sauli (1971).

23.2.3.4 The Spark Chamber

This chamber is constructed in a parallel-plate configuration, which produces a uniform field throughout the gas volume. A spark is caused as follows: an ionizing particle passes through the chamber and ionizes a track; at some stage during this process, the particle causes a trigger to be

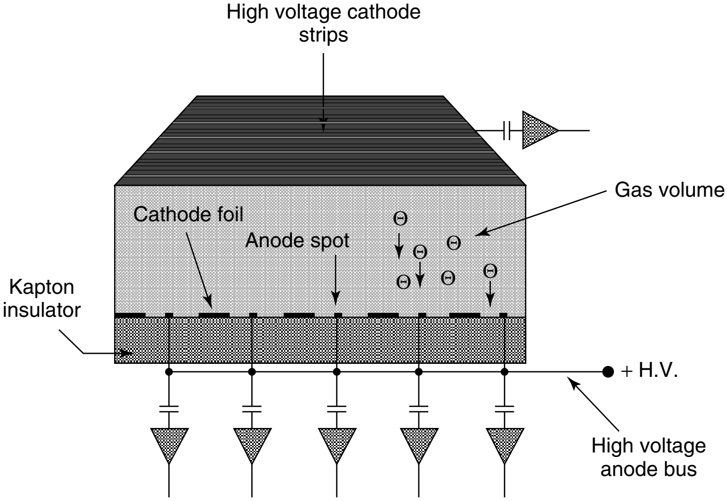

Fig. 23.18 The gaseous pixel chamber.

set that rapidly raises the voltage across the chamber to a breakdown level, suitable for discharges anywhere in the chamber. At this point, a complete ionized channel is formed from anode to cathode leading to a rapid (nanosecond) discharge, or spark. This results in a large number of positive ions in the detector, which are removed by the application of a clearing field while the chamber is waiting to be triggered again. Spark chambers are extremely robust and easy to handle. Because of the high spark energy, simple and low-cost electronics can be used. They are, however, limited for use at very low-rate particle detection.

23.2.3.5 **The Resistive Plate Chamber (RPC)**

The operating principle of the resistive plate chamber (RPC) differs greatly from that of the classical spark chamber although it is one itself. Using flat electrodes with high resistivity allows the operation of this type of spark detector in a continuous, that is, nontriggered, mode. The violent discharge that is produced by an ionizing particle does not propagate nor continue for long because of the instant *local* breakdown of the electric field at the point of impact. The resistivity of the electrode material prevents an immediate charge rebuildup within an area of typically $0.1\,\mathrm{cm}^2$ around the spark center. Furthermore, gases with very high photon-absorption cross section (quenching) must be used to avoid secondary sparks induced by UV photons generated in the first spark.

The immediate advantages of the RPC lie in the ease of construction (costs) and excellent time resolution (<1 nanosecond), pulse rise time (3–4 nanoseconds), and pulse duration (about 10 nanoseconds). The generated pulses are so large that they induce sufficient charge (300–500 mV) on so-called pickup electrodes exterior to the high field electrodes. These pickup electrodes can be arranged in x and y giving the detector spatial granularity and modularity (Figure 23.19). This permits use of low-cost amplifiers for large-scale operation as, for example, outer muon

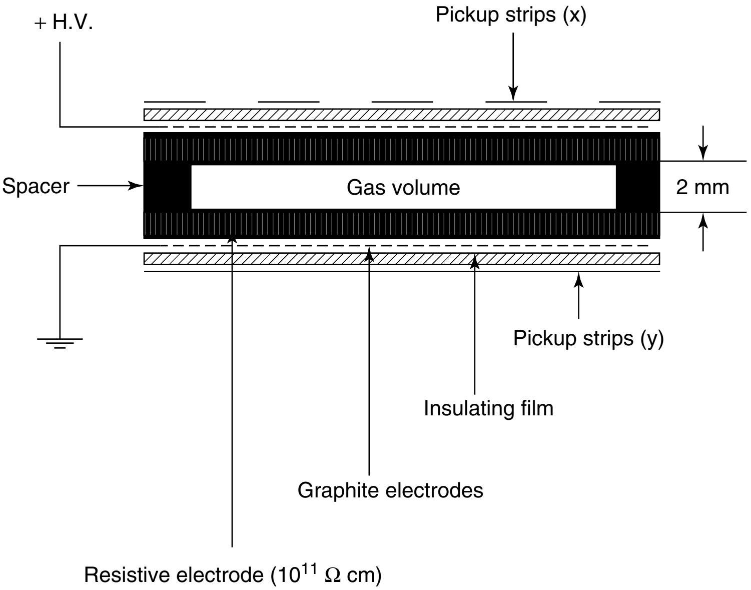

Fig. 23.19 The resistive plate chamber.

detectors in colliding-beam experiments at the LHC (CMS, 1992).

It must be noted, however, that in spite of the fast response of the chamber the rate capability is largely limited by the field restoring time. Resistive plate detectors therefore operate only at rates not exceeding 100 Hz cm^{-2}.

For further discussion, see Cardarelli and Santonico (1981) and Cardarelli *et al.* (1988).

23.2.4
Solid-State Detectors

Solid-state or semiconductor detectors are based on a reverse-biased n–p semiconductor junction in which a carrier-free zone, the depletion zone, is created. In this way, they are used as ionization chambers. Charged particles interacting with the material in this area will liberate electron-hole pairs at the expense of only 3.6 eV (in Si) energy loss per pair, against 30 eV for a gas and 1000 eV for the combination of a scintillator and a photomultiplier. Therefore, compared to gaseous and liquid ionization detectors, larger primary signals and high charge statistics are expected, yielding exceptional performance in energy and position resolution. These detectors have revolutionized fields where the detection of γ-rays played an essential role, like the study of nuclear structure.

Silicon and germanium are the most widely used materials. The latter is mostly used for kiloelectronvolt and megaelectronvolt photon detection, but it has to be operated at the temperature of liquid nitrogen. The energy resolution varies with the energy of the photons. At 6, 60, and 661 keV it reaches 0.15, 3.3, and 4.3 keV (FWHM), respectively. This is 10 times better than what can be obtained with the best scintillator. Silicon can be used at room temperature. It is commonly used in low-energy nuclear physics for the energy loss of particles and their identification (Comiso *et al.*, 1975).

For high-energy physics applications, these detectors can be made very compact, with excellent granularity and precision. They are, therefore, first-choice technologies for today's vertex detection and tracking, predominantly in congested high-rate environments.

23.2.4.1 **The Silicon Sensor**

Silicon has become the most widely used semiconductor material in electronics application today. Its features, quality, and production costs have also made it a nearly ideal candidate for a solid-state detector. The basic detector size is determined by the area available on a 100–150-mm (4–6-in.) diameter wafer. An area less than this is available for the active detector region, which will be subdivided into strips or pads. A convenient choice is dimensions of $\sim 70 \times 56\,\text{mm}^2$ with an active area of $64 \times 51.2\,\text{mm}^2$.

The simplest type of a silicon sensor is shown in Figure 23.20 (Rossi *et al.*, 2006). A large area p^+-implantation is placed in an *n*-bulk. On the backside, an n^+-implantation and an aluminum layer provide an ohmic contact. A positive bias is applied to the backside; the metal electrode on the top is on ground potential. A depletion zone starts to grow from the *pn*-junction into the bulk. This reverse biased *pn*-junction is the basic building block of silicon sensors. It builds up an electric field that collects the signal charge and suppresses the leakage current. An ionizing particle traversing the depletion zone liberates charge, which is collected by the field and can be detected by both electrodes; usually a preamplifier is connected to the electrode on ground potential.

In order to obtain spatial information, one or both electrodes have to be segmented. If the elements of a segmented electrode are approximately rectangular, the detector is called a *pixel detector* (or *pad detector* if the segmentation is coarse): a strip detector consists of largely asymmetric segmented electrodes (e.g., several centimeters long and a few tens of micrometers wide). One of the first examples of pad detectors used for tracking is the UA2 experiment at the CERN SppS Collider (Ansari *et al.*, 1989); H1 at the HERA storage ring at Deutsches Elektronen-Synchrotron (DESY) used large pad detectors in a calorimeter (Fretwurst *et al.*, 1990).

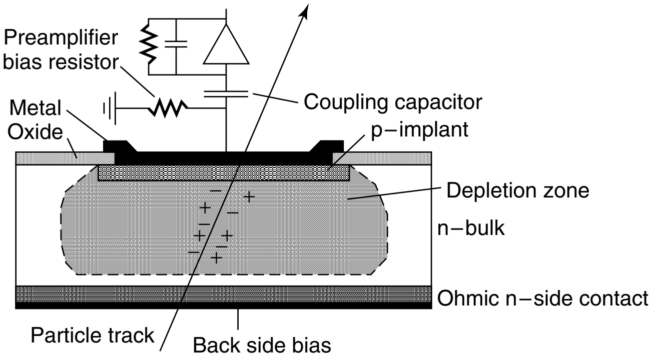

Fig. 23.20 Schematic cross section of a simple silicon pad sensor (Rossi *et al.*, 2006).

23.2.4.2 The Silicon Strip Detector

The most common design is that of a strip detector where thin metal strips are deposited on the p or junction side of the silicon wafer (single-sided readout) (Bencze *et al.*, 1993; Heijne *et al.*, 1980). Ion implantation and photolithographic techniques are used to selectively dope the surfaces of the wafer and to deposit the metallization patterns necessary to extract the signals (Kemmer, 1980). A sketch of a generic single-sided microstrip detector is shown in Figure 23.21. The strips are p^+-implants on n-type silicon, repeated at a pitch p. A voltage is applied between the backside metallization and the strips (Rossi *et al.*, 2006). Perpendicular strips or suitable pad structures on the back plane (n side) of the wafer provide two-dimensional track information (double-sided readout) (Sadrozinski, 1989; O'Shaughnessy *et al.*, 1990; Brenner *et al.*, 1992). Diffusion during charge collection allows interpolation from strip pulse heights. Reading, for example, every strip (20-μm wide) results in a single-track resolution of $\sigma < 3\,\mu\text{m}$, and a double-track resolution of $\sigma < 50\,\mu\text{m}$ (Sadrozinski and Seiden, 1989).

The band-gap energy of silicon is only 1.3 eV. Thus, in a substrate of 300-μm thickness, an MIP generates a signal of ~25 000 electrons, which can easily be interfaced and processed with integrated circuits, sometimes even mounted on the same substrate. Signal collection time is limited by the mobility of the holes, which is roughly one-third of the electron mobility ($0.15\,\text{m}^2\,\text{V}^{-1}\,\text{s}^{-1}$). Most detectors have collected the full charge within tens of nanoseconds. Interstrip capacitance and strip resistance of 6-cm-long strips have little influence on signal quality and signal propagation and are found to be adequate even for operation at the highest particle rates expected at hadron colliders. The effects of radiation on the detector performance (e.g., n- to p-type inversion, bulk damage, and surface charging) impose stringent conditions on the sensor, that is, operating at temperatures in the range of $-5°$ to $-10°$. (Wunstorf *et al.*, 1989; RD-20 1992; Barberis *et al.*, 1993; Aad *et al.*, 2008).

For further discussion, see Lutz (1999).

23.2.4.3 The Silicon Drift Chamber

By making use of the uniform and constant drift velocity of charge in silicon, a simple

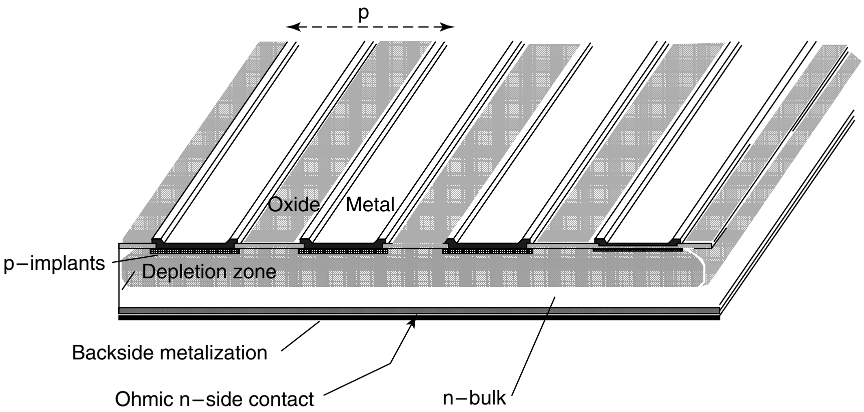

Fig. 23.21 The single-sided silicon strip detector (Rossi *et al.*, 2006).

and low-cost detector can be built from this material providing two-dimensional readout for tracking. This detector employs anode pads for signal readout and junction strips, spaced at 150 μm, for creating a uniform drift field of 150 V cm^{-1} transverse to the bias field (Figure 23.22). The drift velocity within the device is approximately 2 μm ns^{-1}.

One track coordinate is derived from the measurement of the drift time, and the other from anode pad interpolation. Both result in about 20-μm resolution over 4 mm of drift distance. Because of the weak drift field, very pure and uniform doping of the *n*-type Si crystal is needed. Furthermore, this method will result in inferior double-track separation as compared with classical silicon strip detectors with similar strip spacing.

For more information, see Gatti and Rehak (1984, 2005), Gatti *et al.* (1984), Faschingbauer *et al.* (1996), Holl *et al.* (1996), Pandey *et al.* (1999, 2002), Beole *et al.* (2007), and Aamodt *et al.* (2008).

23.2.4.4 **The Silicon Pixel Detector**

Under extreme conditions, such as in high-luminosity hadron colliders, there is such a density of particles that individual tracks cannot be resolved, even using the most granular strip detectors. Particle multiplicities in high-energy jets with opening angles of order 20 mrad are likely to be 30–70 and several thousand charged particles per unit pseudorapidity might be produced in Pb + Pb collisions at the LHC. The pixel detector is a device that precisely detects and recognizes dense tracks in a crowded environment (Campbell *et al.*, 1990; Heijne *et al.*, 1990; Anghinolfi *et al.*, 1992). The sensitive cell size of this detector is reduced to pads or pixels of $100 \times 100\,\mu\text{m}^2$ or less, operated similarly to the silicon strip detector.

Hybrid Pixel Sensor The fabrication of a pixel sensor is very similar to the fabrication of a microstrip sensor, but the implants have a higher segmentation. If the area of the sensing element is below

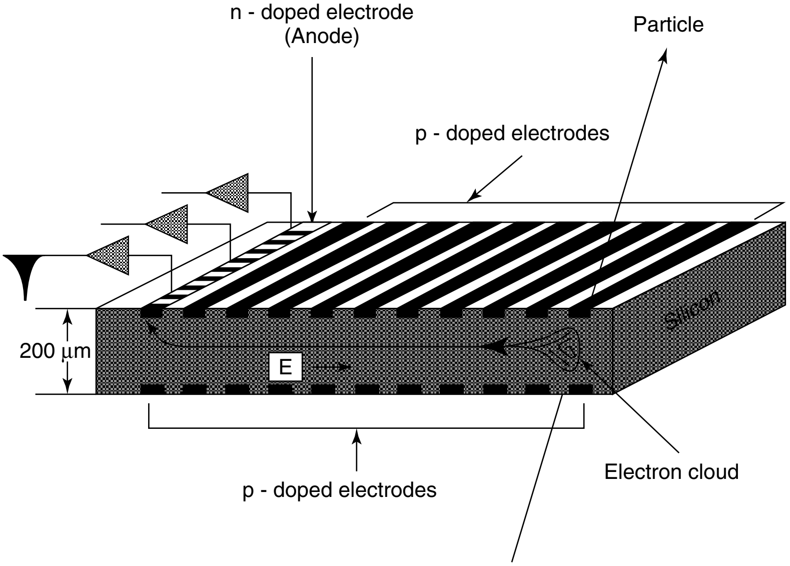

Fig. 23.22 The silicon drift chamber.

1 mm^2 and their number is in the order of 10^3–10^4, the signals cannot be routed to the periphery any longer. This makes a real two-dimensional connectivity to the electronics necessary. The connection between the sensor and the electronics must be vertical and match the sensor dimension and the electronics chip must be very close (10–20 μm) to the sensor (Rossi *et al.*, 2006). Figure 23.23 shows a typical pixel cell, the basic building block of a hybrid pixel detector. The ionizing particle traverses the sensor and generates charges, which will be collected by the electric field and detected on the electrodes. The produced signal is amplified and processed by the electronics chip.

One of the key advantages of hybrid pixel detectors is the low capacitance of the sensing element due to its small dimensions. Each pixel covers a very small area (about 10^{-4} cm^2) over a thin (about 300 μm) layer of silicon, resulting in a very low capacitance (0.2–0.4 pF). Such a low capacitance allows fast signal shaping with very low noise. Noise figures of $200e^-$ for electronics operating at 40 MHz are obtained. The S/N thus exceeds 100 for a fully depleted sensor of 300-μm thickness and an MIP traversing it.

Hybrid pixel sensors are the choice of technology for the vertex and tracking detectors at the LHC. They can be made radiation-hard and provide nonambiguous three-dimensional information with good time resolution and a spatial resolution, which is needed to measure short-lived particles (Aad *et al.*, 2008; Chatrchyan *et al.*, 2008; Aamodt *et al.*, 2008; Augusto Alves *et al.*, 2008).

The Monolithic Active Pixel Sensor (MAPS) The monolithic active pixel sensor is a silicon detector technique that allows to

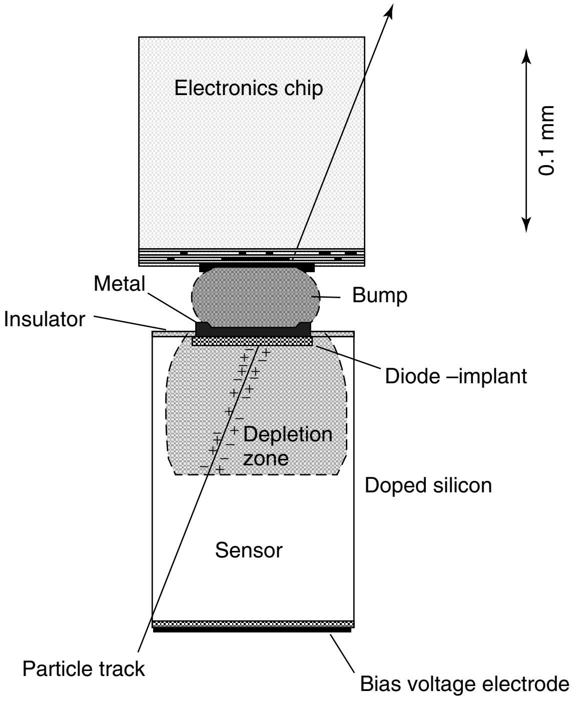

Fig. 23.23 Schematic view of one cell of a hybrid pixel detector (Rossi *et al.*, 2006).

integrate the detector element and the processing electronics on the same substrate (Gornushkin *et al.*, 2002). Commercial CMOS technologies use low-resistivity silicon, which is not suited for charge collection. However, some technologies allow the use of thin (at most 15 μm) epitaxial silicon layers. In a MAPS, a partially depleted thin epitaxial layer of low-resistivity silicon is used as a sensitive detector volume. The charge generated by the impinging MIP, about 80 electron-hole pairs/μm, is collected by an *n*-well/*p*-epi diode, created by *n*-well implantation into the epitaxial layer. The electrons liberated in the epitaxial layer diffuse toward the diode within a typical time of a few tens of nanoseconds. The active volume is underneath the readout electronics, allowing 100% fill factor. Because of the 3 orders of magnitude between the doping levels of the *p*-epitaxial layer and of the p^{++} wells and substrate, potential barriers are created at the region boundaries, which act like mirrors for the excess electrons (Figure 23.24).

This concept has undergone a very rapid development, with much effort concentrated on the MIMOSA (Turchetta *et al.*, 2006) chip, which has completed at least 15 full cycles of chip design since 1999. Noise figures of 10–30 e^- and $S/N > 20$ have been achieved (Dulinski *et al.*, 2003) with spatial position resolutions below 5 μm. The advantage of a fully CMOS monolithic device is its potentially low cost – besides the excellent spatial resolution. The disadvantages are the small signal, the long charge collection time, and the sensitivity to radiation effects, although some progress concerning the radiation hardness of MAPS has been made (Deveaux *et al.*, 2007).

The Depleted Field Effect Transistor (DEPFET) Another approach to build monolithic pixel detectors is to use high-resistivity silicon and to integrate simple electronics circuits (e.g., the first stage of the amplifier of each pixel and some addressing scheme). The signal generation is optimal as large depletion zones and

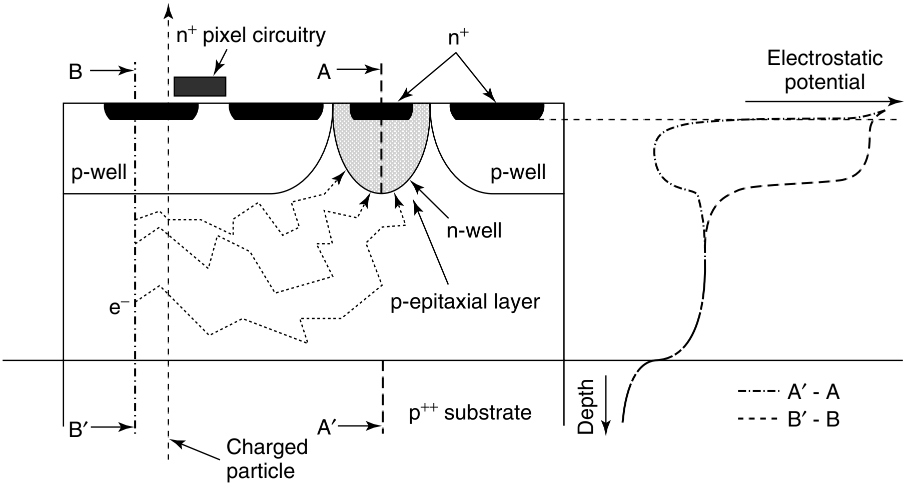

Fig. 23.24 Schematic cross section of a CMOS pixel detector cell (MAPS) (Turchetta *et al.*, 2001).

high electric fields provide fast and efficient charge collection. However, the on-chip signal processing is limited since all design technologies are tailored to low-resistivity silicon. The depleted field effect transistor (DEPFET) structure is one such method to provide both signal detection and amplification (Kemmer and Lutz, 1987; Kemmer *et al.*, 1990; Ulrici *et al.*, 2001). A DEPFET detector and amplification structure is shown in Figure 23.25. A field effect junction is integrated onto a sideward depleted substrate. Sideward depletion provides a parabolic potential (see Figure 23.25b) with a minimum for majority carriers (electrons in the case of n-type silicon) inside the detector volume. This local minimum can be moved toward the top surface right underneath the transistor channel (internal gate) by locally varying the doping concentration (deep n-implant) and by applying appropriate electrical potentials. When a charged particle traverses the detector, it creates electron-hole pairs, the electrons being collected at the internal gate. The transistor channel current, steered by the voltage at the external gate, is now modulated by the collected charge at the internal gate. The collected electrons are eventually removed from the internal gate, for example, by applying a clear pulse to a dedicated contact outside the transistor region. The very low input capacitance ($\approx$ femtofarads) and the *in situ* charge to current conversion result in low-noise figures (a few e^-) (Ulrici *et al.*, 2001). The low-noise feature can be exploited to make very thin sensors ($\approx$50 μm), while still yielding a very good S/N performance and thus allowing a very fast readout.

Silicon-On-Insulator (SOI) A third possibility for a monolithic pixel detector is based on the silicon-on-insulator (SOI) technology. High-resistivity bulk material for charge collection is combined with an electronics layer. By means of a thick buried oxide layer (BOX) in the bulk, the active sensor volume is isolated from layers with CMOS circuitry (see Figure 23.26). Contact through the insulation (BOX) to the sensor substrate is made by vias of p^+-implants, which also form the diodes. The advantages of this pixel detector type are the 100% fill factor, a large signal and a fast charge collection, and the possibility to use NMOS

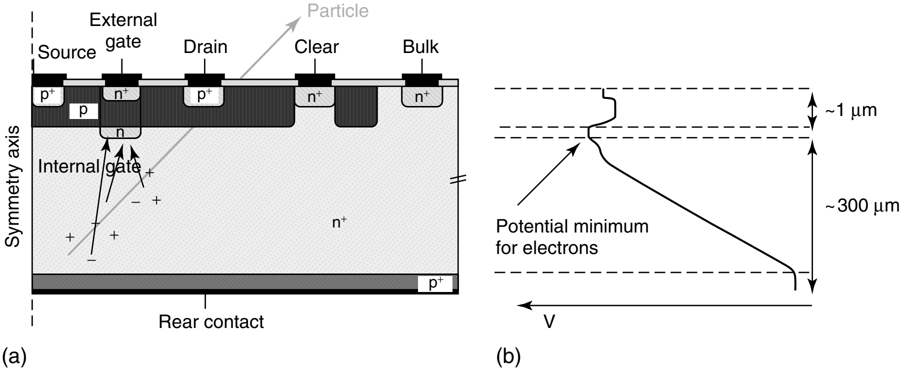

Fig. 23.25 Cross section of a DEPFET pixel (a) and potential between top gate and rear contact as function of depth (b) (Wermes, 2004).

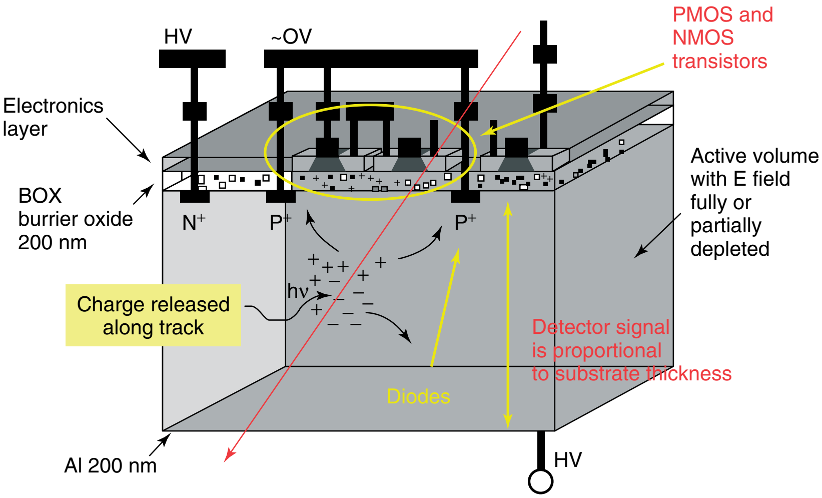
HV
~OV
PMOS and NMOS transistors
Electronics layer
BOX burrier oxide 200 nm
N+
P+
P+
Active volume with E field fully or partially depleted
Charge released along track
hν
Detector signal is proportional to substrate thickness
Diodes
Al 200 nm
HV

Fig. 23.26 Sketch of an SOI monolithic pixel cell (Yarema *et al.*, 2007).

and PMOS transistors for the readout. A first prototype of a 64 × 64 pixel array with 26-μm pitch has been produced (Arai *et al.*, 2007; Deptuch *et al.*, 2007).

Charge-Coupled Device (CCD) Since the invention of the charge-coupled device (CCD) in 1970 (Boyle and Smith, 1970; Amelio *et al.*, 1970), this technology has had and still has applications in the fields of optical imaging, particle tracking, X-ray detection, analog storage devices, and so on. CCDs are large two-dimensional arrays of highly segmented (very small pixel dimensions below $2 \times 2\,\mu\text{m}^2$) silicon sensors with a very thin active layer (10–20 μm). Their unique advantage for tracking lies in their ability to confine the signal charge within an extremely compact collection volume, and then to transfer the signal packet without loss onto the gate of a very low capacitance on-chip sensing transistor. However, this on-detector multiplexing has its disadvantages: the serial sensing of the signals from a large number of pixels by a single output circuit takes time. CCDs are slow detectors in comparison with "real" two-dimensional detectors where the signals are read out in parallel by a separate output circuit for every detector element, that is, the hybrid pixel detector.

In optical imaging, each pixel is typically supplied with a signal of the order of $10^5 e^-$. MIPs generate a signal in the sensitive layer, which is about a factor 100 smaller, which imposes two major challenges on design and operation of CCDs used for particle tracking detectors. First, there is the requirement of extremely efficient, noiseless transport of the tiny charge packet through possibly thousands of transfers within the imaging and readout regions of the device. The second major challenge is the electronic noise performance of the output circuit and associated signal processing.

Figure 23.27 shows the cross section of a typical device. It consists of a lightly doped epitaxial p layer on a heavily doped substrate, with a $\sim 1\,\mu\text{m}$ n^+ layer on the

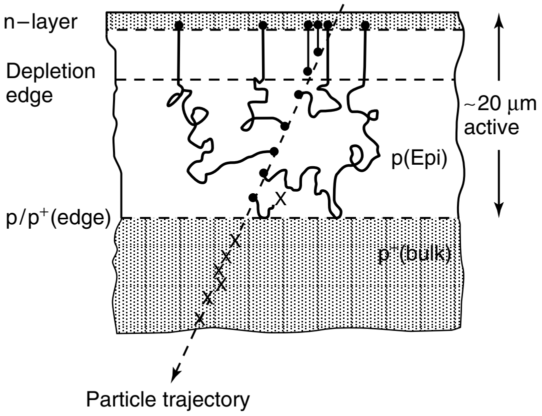

Fig. 23.27 Cross section of a buried-channel CCD structure (Damarell, 1998).

top. The partial depletion of the n^+/p region creates a potential minimum for storage of charge carriers (electrons in this case) just above the depth of the n^+/p edge. Signal electrons generated within the depletion region of thickness of ~5 μm will be transported toward this minimum by drift in the electric field. Minority carriers (electrons) generated in the undepleted p-type substrate will diffuse isotropically. Those that reach the depletion edge will be drawn into the potential energy minimum while those that diffuse to the p/p^+ edge are reflected by a potential barrier due to the thin intrinsic depletion layer and will eventually be collected in the storage region.

Charge is confined in depth by the potential distribution mentioned above. Transversely, charge confinement in the imaging area is achieved by vertical p^+ implants (channel stop regions) and horizontal polysilicon gates, electrically isolated from each other and from the substrate (Figure 23.28). The charge localized in such a way is transferred in parallel down each column toward the output register by manipulating the potentials on the imaging register ($I\Phi$) gates in such a way that the two-dimensional image is preserved. On each I shift, the charge from the bottom row of the image area is transferred into the R register. This row is then read out serially through the output circuit at the end of the R register, after which the next row transfer is made. The horizontal transfer along the R register is achieved by a similar, independent set of readout register ($R\Phi$) gates. The signal charge is then transferred – pixel after pixel – to the output circuit where it modulates the drain current in an FET (Damarell, 1998).

CCDs have been employed as vertex detectors in a fixed target experiment, NA32 (Damerell, 1985), and an e^+e^- collider experiment, the SLD (SLAC Large Detector) experiment (Abe *et al.*, 1997). The use of CCD detectors is ruled out at hadron colliders because the instantaneous rates are excessive and their life expectancy is inadequate owing to their sensitivity to hadronic radiation damage. CCDs are a well-established baseline option for the International Linear Collider (ILC) vertex detector design. However, the readout speed and radiation hardness have to be improved. A major innovation developed for the ILC design is to replace the traditional serial register readout, where the charges from each readout column

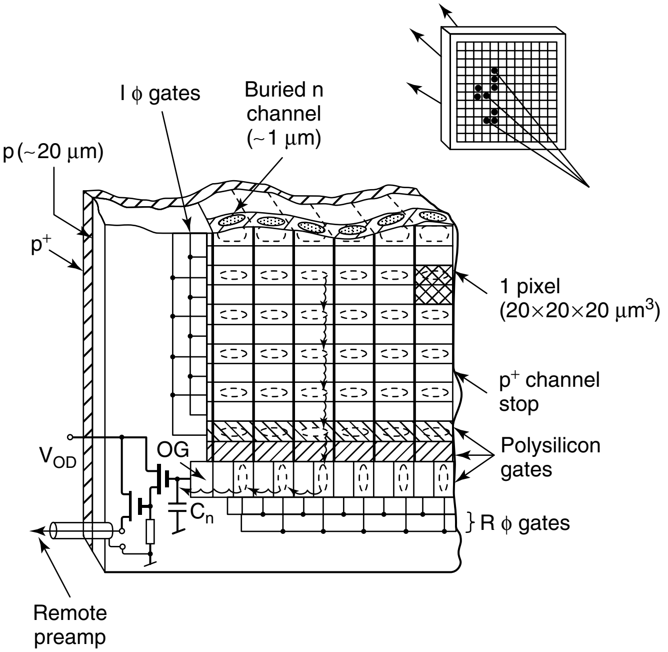

Fig. 23.28 Upper right: Sketch of charge storage in a CCD detector traversed by a number of ionizing particles. Lower left: Corner region of the CCD showing the principal structural features (Damarell, 1998).

are moved toward a readout circuit on the corner, with a column parallel readout, where the readout chip receives analog signals from all columns in parallel (Stefanov, 2005). A recent development of the CCD concept is the *in situ* storage image sensor (ISIS). The design of this detector is tailored to the specific bunch structure of the ILC (1-millisecond-long bunch trains at a rate of 5 Hz) and the beam-related RF pickup (Figure 23.29). The charge is collected at the photogate and transferred to a storage pixel every 50 microseconds during the beam. After 20 samples have been stored, the charge is shifted out for readout like in normal CCD and converted to voltage in the 200-millisecond gap between the bunch trains, completely avoiding the RF pickup. Clock frequency of only 1 MHz, assuming column parallel readout, is required to retrieve all stored charge from the full-scale device (Stefanov, 2006b).

Three-Dimensional Detectors For efficient operation of a conventional (planar) silicon detector, all electrons and holes generated by the incoming radiation have to be collected at the detector electrodes. The detector should be thick enough to generate a sufficient number of electron-hole pairs to get a high S/N and the detector should be fully depleted to collect the generated carriers from the full thickness of the detector. When the detector is exposed to radiation defects are produced in the silicon lattice. Some of these defects are electrically active

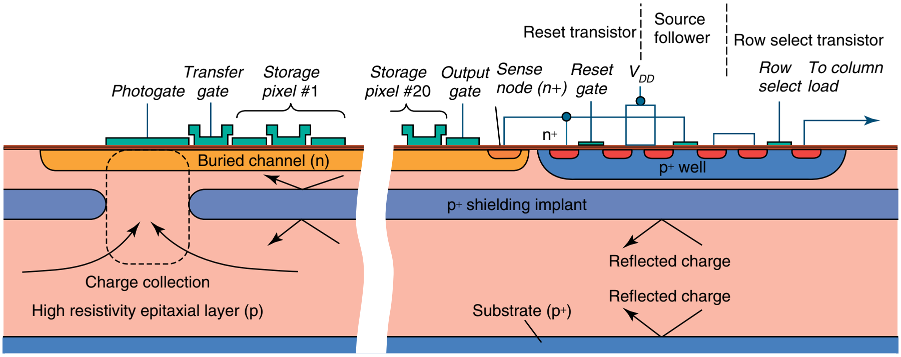
Reset transistor
Source follower
Row select transistor
Transfer gate
Storage pixel #1
Photogate
Storage pixel #20
Output gate
Sense node (n+)
Reset gate
V_{DD}
Row select
To column load
n+
Buried channel (n)
p+ well
p+ shielding implant
Reflected charge
Charge collection
High resistivity epitaxial layer (p)
Substrate (p+)
Reflected charge

Fig. 23.29 Principle of operation of the ISIS sensor (Stefanov, 2006a).

and are very efficient carrier recombination centers. The formation of such centers reduces carrier lifetime, and charge collection efficiency deteriorates. Besides, these electrically active centers influence the equilibrium carrier concentration (effective doping) in Si, which results in increasing full depletion voltage and ultimately renders the detectors inoperable. A radical approach to combat radiation damage is to make the electrodes penetrate into the bulk instead of lying on the surface of the sensor. This design has the great advantage of decoupling the drift distance, which governs the radiation hardness properties, from the sensor thickness, which dictates the amount of signal collected. The small electrode separation gives intrinsically fast charge collection, and low depletion voltages, along with the possibility of removing dead regions at the sensor edges.

Three-dimensional detectors (Parker *et al.*, 1997) have electrodes that go all the way through the bulk of the material (Figure 23.30). This allows the electrodes to be positioned much closer together without the need to reduce the thickness of the detector, and thus the active

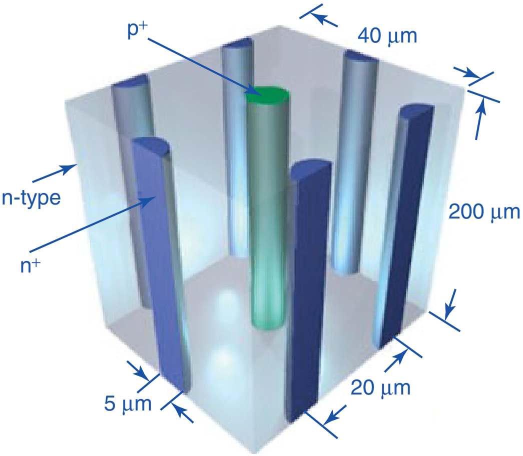
p+
40 µm
n-type
n+
200 µm
5 µm
20 µm

Fig. 23.30 The three-dimensional silicon pixel detector.

volume. The close positioning of the electrodes is beneficial for both the full depletion voltage and charge collection efficiency. Indeed, since the depletion voltage depends quadratically on the distance between the electrodes, the effect is dramatic and the full depletion voltage can be decreased to as low as 1–10 V depending on the Si doping (compared to 50–500 V for conventional detectors). The charge collection efficiency improves simply owing to the fact that the generated carriers have to travel a shorter distance to the electrodes, which decreases the carrier collection time (Kenney *et al.*, 1999) and thus less carriers will recombine before reaching the electrodes. Three-dimensional detectors can be fabricated using established techniques that are widely implemented for Si micromachining and porous Si formation: reactive ion etching (RIE) and electrochemical etching (EE) (Pellegrini *et al.*, 2002). Although some parts of the fabrication process are relatively well characterized, the problem of fabrication of the three-dimensional detector poses serious scientific and technological challenges specific to carrier recombination and charge collection.

Recent testbeam results are very encouraging, with efficiencies reaching 98%, and irradiated devices (single-type-column design, only *n*-type columns etched into a *p*-type substrate (Piemonte *et al.*, 2005)) have been characterized with lasers, showing good charge collection efficiency at relatively modest depletion voltages (Eckert *et al.*, 2007).

23.2.4.5 The Gallium–Arsenide (GaAs) Detector

Solid-state detectors are often subject to high radiation fluences because they are located close to the beam-interaction vertex (Groom *et al.*, 1989; Stevenson, 1992). From the physics point of view, they must be placed close to the decay origin of short-lived particles. Therefore, silicon detectors are particularly prone to radiation damage by charged particles and neutrons, resulting, for example, in high-leakage currents from bulk defects in the crystal (Wunstorf *et al.*, 1989).

Gallium–Arsenide (GaAs), as a semiconducting material operated in a similar fashion as silicon, can tolerate higher radiation levels than silicon before undergoing gradual destruction (Smith, 1990). The signal produced by GaAs detectors is, however, inferior to that from silicon by approximately a factor of 2.

23.2.4.6 The Diamond Detector

For reasons of even higher radiation tolerances, diamond crystals have become an attractive alternative to silicon and GaAs devices. For a long time, natural diamond crystals have been found to function as homogeneous conduction counters offering a few unique features for application as tracking detectors. Their major advantages over comparable detection techniques are survivability after radiation exposure to tens of millirads and more than 10^{14} neutrons/cm^2/year, very fast charge collection ($\sim$1 nanosecond) due to roughly equal electron-hole mobility ($0.18/0.12\,\mathrm{m^2\,V^{-1}\,s^{-1}}$), and lower strip capacitance due to different dielectric properties of the crystal. The high resistivity of the crystal lattice ($>10^{12}\,\Omega\,\mathrm{cm}$) as compared with silicon (10^5) reduces leakage currents to negligible levels at high bias voltages allowing saturation velocity of the charge carriers. Unfortunately, high bulk resistivity also leads to polarization effects that, from accumulation of charges, render the detector inoperable when exposed to high rates. A band-gap energy more than 5 times as high as in silicon, requires higher energy losses per ionizing particle in order to yield comparable signals. Thus, pulse heights of diamond detectors, at present, are roughly 3–4 times lower than silicon signals.

To form a full-scale tracking detector, some of the inherent drawbacks in this technology have yet to be solved. Natural diamond, although produced in quantities of several tons (i.e., millions of karats) per year (Kozlov *et al.*, 1975), is still unaffordable, where sensitive detector areas in the order of several square meters are required. Furthermore, even rather pure crystals contain traces of impurities (these create the different colors) such as nitrogen that trap free carriers and thereby reduce the lifetime of drifting electrons and holes. A commercial process called *chemical vapor deposition* (*CVD*) produces low-cost, thin sheets of artificial diamond that is chemically purer than natural crystals. In the beginning, the process suffered from lattice defects that act as trapping centers for drifting charges. The production of high-quality polycrystalline CVD material made it possible to develop a prototype of a pixel detector. A pixel module (46 000 channels) using diamond from a 12-cm wafer with a charge collection distance exceeding 300 μm was bump-bonded to the radiation-hard electronics of ATLAS. The module was tested in a 120-GeV pion test beam at CERN. This module performed quite well with a noise of $136\,e^-$ and an efficiency of >97% (Trischuk, 2008). High-purity single-crystal CVDs have been grown to sizes larger than $1\,\text{cm}^2$. Owing to the lack of grain boundaries, this CVD material has superior electronic characteristics (Barbero *et al.*, 2007).

23.2.5
Visual Tracking Detectors

23.2.5.1 The Cloud Chamber

This is the simplest and most straightforward form of tracking technique, using visual means to observe the tracks of charged particles. The basic construction is simple: a closed chamber containing a supersaturated vapor and a window to view the interior (see Henderson, 1970). When ionized clouds form inside the chamber, small droplets of liquid condense out of the gas at these points and become visible. The chamber can be used only for the study of reactions that are relatively abundant, because a large chamber of, say, $0.5\,\text{m}^3$ volume can be triggered only once every few minutes. This type of detector has been superseded by the more sophisticated and accurate bubble chamber.

23.2.5.2 The Bubble Chamber

Discovered by Glaser in 1956, the bubble chamber has been the major tool in particle physics for almost two decades. It relies on the properties of a superheated liquid for its operation, although it is basically similar to the cloud chamber (see Henderson, 1970). The liquid in question is held well above its boiling point and under pressure to prevent expansion. To bring the chamber into operation, the pressure is released slightly at the point when the trigger is set and the liquid is free to vaporize. Bubbles preferably form along the paths of charged particles in the chamber and can be recorded visually using photographic emulsion or CCD imaging systems. The further growth of bubbles is stopped by recompression. The repetition cycle is typically of the order of seconds. With separated beams of charged particles, this instrument has permitted the gathering of

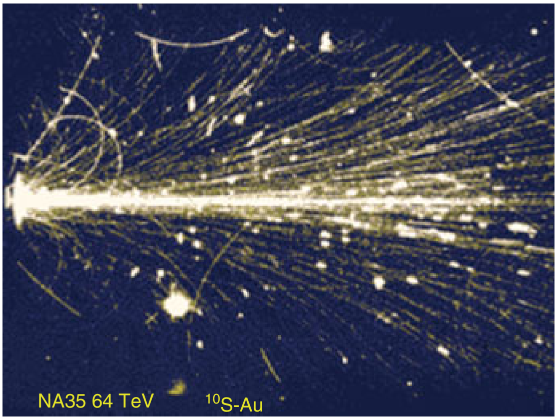

Fig. 23.31 A heavy-ion collision seen by the NA35 streamer chamber (Brinkmann *et al*., 1995).

Tab. 23.1 Summary and special features of the different detector technologies.

Detector technology	**Major application**	**Pros**	**Cons**
Gaseous detectors	Large volume and high-rate tracking, muon triggering	Rugged, cheap, relatively radiation hard, tailored to experimental needs	Sensitive to magnetic field, limited precision
Solid-state detectors	Vertex detection, high-rate tracking	High precision and good energy resolution	Radiation sensitive (Si), expensive
Visual tracking detectors	Vertex measurement (outdated)	High track precision (holographic)	Very slow, dangerous

information on resonant states and new particles, which has been the basis for our present concept of particle structure. The intrinsic accuracy in the measurement of the bubble position is of the order of 100 μm. A great variety of liquids has been used, from liquid hydrogen to xenon.

Extensive data analysis and processing are necessary to obtain three-dimensional track information (a common factor in all types of tracking systems). This kind of chamber is rarely used nowadays, with physicists choosing fast electronic detectors, which are more reliable and easier to control.

23.2.5.3 The Streamer Chamber

A streamer chamber is a large rectangular box with two planar electrodes at its ends and filled with a suitable gas mixture (e.g., Ne/He 80/20). When triggered, a sharp high-voltage pulse (~340 000 V) of short rise time and small duration (~12 nanoseconds) is applied to the electrodes. Every electron released by ionization of a charged particle, which has traversed the gas volume, initiates an avalanche in the intense uniform electric field. Since the electric field is of short duration, the avalanche formation terminates quickly. The excited atoms in the avalanche emit visible light, which forms luminous streamers of about 1-cm length in the direction of the electric field along the trajectory of the charged particle. The streamer chamber events are photographed and processed like bubble chamber data. Nuclear collisions at the SPS have been recorded by such a device (Figure 23.31).

A brief summary of the different detector technologies is given in Table 23.1.

23.3
Tracking Systems

A complete tracking system is composed of various tracking techniques and different types of chambers, combined in such a way as to be complementary in operation. The ultimate aim is to completely pin down a particle in space and time as it passes through the detector. In the following, some of the tracking systems in operation today are described, complete with details of position, momentum, and time resolution, where they are of importance.

23.3.1
The Time Projection Chamber

The time projection chamber (TPC) is a more sophisticated version of the imaging drift chamber. In most applications, the electric and magnetic fields are parallel (Clark *et al.*, 1976). Under the combined action of the fields, it is possible to drift electrons over large distances (up to several meters), while retaining their precise location in space and time (Figure 23.32) (ALEPH, 1990). Thus the detector essentially consists of a large gaseous drift volume with a homogeneous drift field similar to the classical parallel-plate configuration. One end generates the electric field necessary to transport electrons, while the other, at low potential, receives and measures the arriving charge. Similar to a photographic plate, this so-called end cap of the detector records the arrival of information (tracks) in short time slices, so that the full history of an event can be reproduced.

The actual detector elements are proportional chambers shaped into sectors in

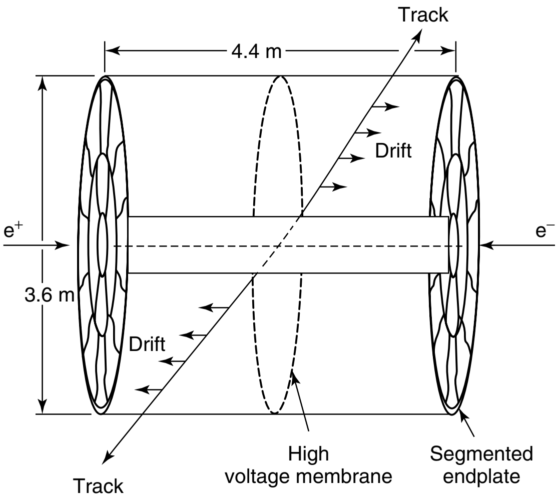

Fig. 23.32 Principle of the time projection chamber (TPC). Long-distance electron drift without lateral diffusion is assured by the parallel arrangement of electric and magnetic field.

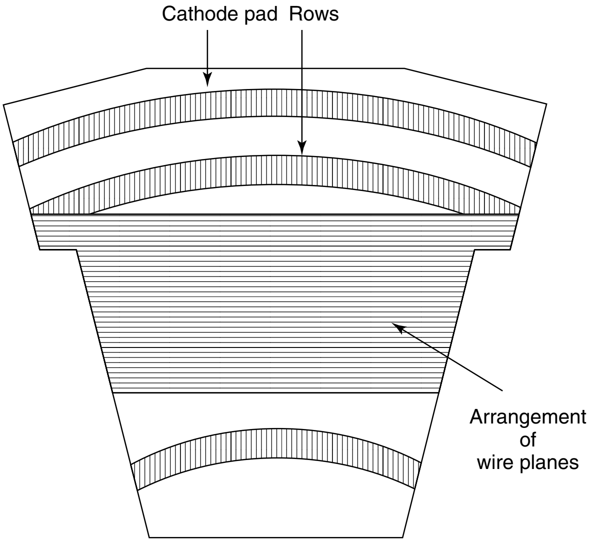

Fig. 23.33 Schematic of an end-cap sector of the ALEPH TPC. Sectors of a TPC are standard multiwire proportional chambers, specially shaped in order to avoid long cracks for particle tracks projected onto the end plates. Cathode pad shape and their circular arrangement are optimized for high spatial and double-track resolution.

order to adapt to the overall geometry of the TPC (Figure 23.33). They have fine-grained cathode segmentation (pad structure) for high-resolution tracking. In analogy to photographic techniques, these sectors can even be equipped with electronic shutters (known as *gates*) to prevent unwanted charge from entering them. Given the size of such a detector (typically 25–88 m^3), an impressive resolution in all three space coordinates can be achieved.

23.3.1.1 The ALEPH TPC

A good example of such a detector is the ALEPH TPC at LEP/CERN. It had a diameter of 3.6 m and a length of 4.4 m. The two end caps were segmented into 18 sectors each, which came in three different shapes, designed to make optimum use of the space and to lose a minimum of data along the sector boundaries (Figure 23.34). This detector had a spatial accuracy in the event projection on the end plates $(x - y) \leq 200\,\mu\text{m}$. This was kept low by having the magnetic and electric fields parallel, so that electrons did not diffuse laterally but were forced to orbit about the center of the ionized cluster. The z-coordinate precision are determined by the longitudinal diffusion of electrons along their drift path and was of the order of 1 mm (Atwood *et al.*, 1991).

The values for the resolution also depend on the angle at which tracks hit the anode wires, since, in the avalanche region, the magnetic field is no longer parallel to the electric field and there is a smearing effect on each track segment along the wire caused by the electromagnetic interaction with the electrons.

This detector concept is especially suited to colliding-beam experiments with sufficiently long (≈10–100 microseconds) time intervals between successive beam interactions. Figure 23.35 illustrates a typical two-jet event from the decay of a Z^0 boson into a quark–antiquark pair as produced at the 100-GeV LEP Electron–Positron Collider (ALEPH, 1991).

23.3.1.2 TPCs for High-Energy Nucleus–Nucleus Collisions

TPCs are the main tracking detectors in collisions of heavy ions at ultrarelativistic

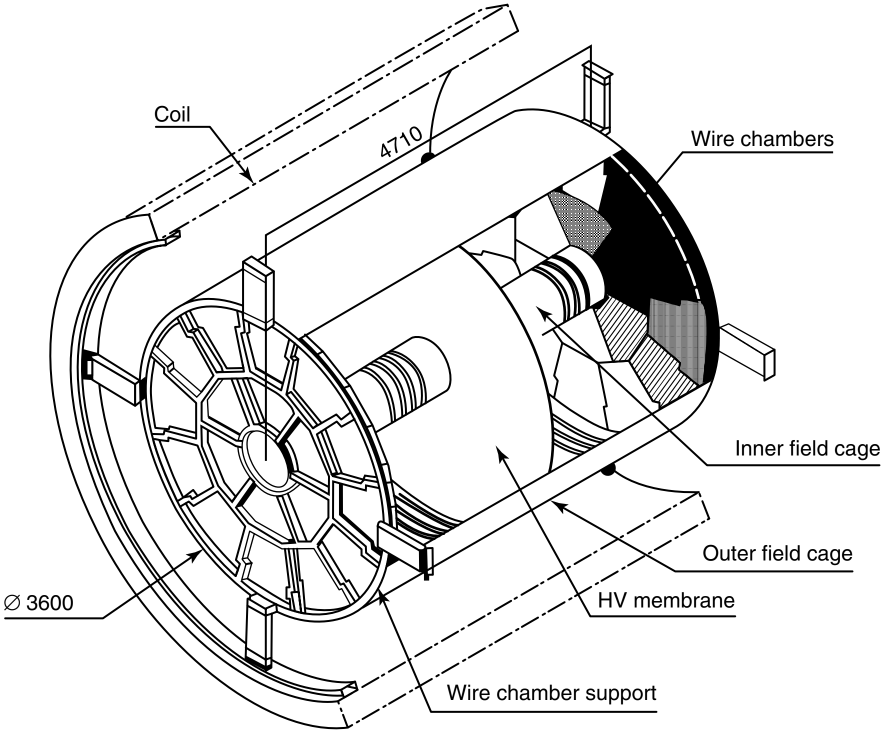

Fig. 23.34 Three-dimensional view of the ALEPH TPC.

energies. They have been used in fixed target experiments at the AGS and the SPS (e.g., NA49 (Afanasiev *et al.*, 1999)). The STAR detector at RHIC at BNL has demonstrated that a large-volume TPC is capable of recording thousands of charged particle tracks per collision at an event rate of up to 1 kHz. Figure 23.36 shows an event display of an Au + Au collision at $\sqrt{s_{NN}} = 200$ GeV (Ackermann *et al.*, 1999).

The ALICE experiment at the LHC is designed to record collisions of Pb nuclei at $\sqrt{s_{NN}} = 5.5$ TeV. Various detectors arranged in cylindrical shells around the interaction point (see Figure 23.37) are designed to determine the identity and precise trajectory of the more than 10 000 charged particles propelled by a lead–lead nuclear collision into the active volume of the apparatus (ALICE, 1995; Braun-Munzinger and Stachel, 2007). The innermost detector is the inner tracking system (ITS), which consists of six layers of silicon detectors surrounding the 1-mm-thick beam pipe that encloses the ultrahigh vacuum of the accelerator. These detectors are capable of high-precision tracking (resolution around 20 μm) in order to determine the decay vertex of short-lived particles carrying strange, charm, or bottom quantum numbers that typically decay within a few millimeters to centimeters of the primary interaction point. ALICE's major tracking device is a TPC with 560 000 readout channels; it provides essentially continuous, three-dimensional tracking of charged particles between radii of 80 and 250 cm from the central interaction point.

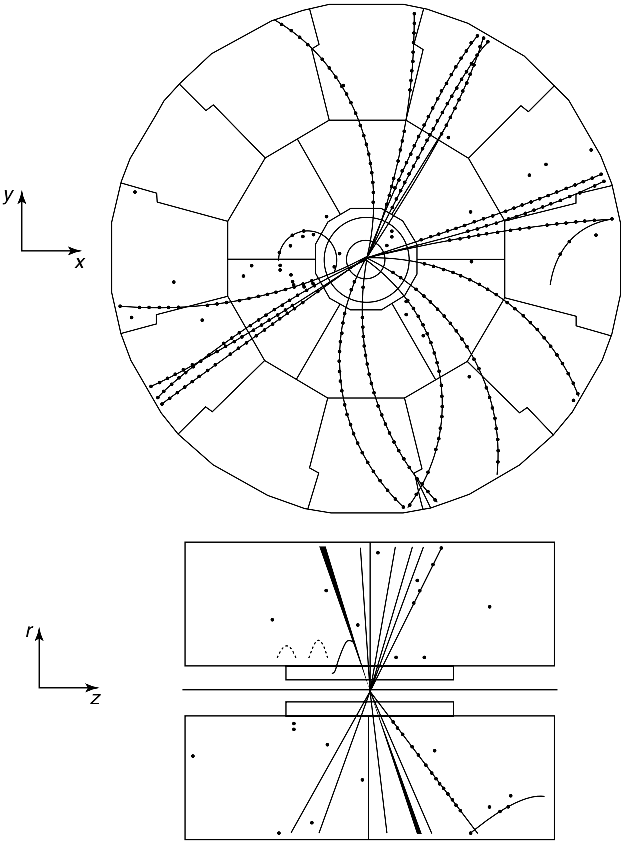

Fig. 23.35 Decay of a Z^0 into hadrons (jets) seen by the ALEPH TPC. Each point on a track corresponds to a pad-row hit (x–y plane).

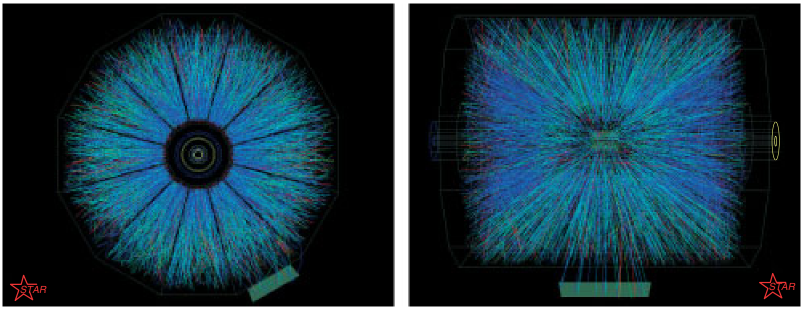

(a) (b)

Fig. 23.36 Event display of charged particle tracks in a Au + Au collision at RHIC energies recorded by the STAR TPC. (a) Front view and (b) side view. Images reproduced with permission from BNL.

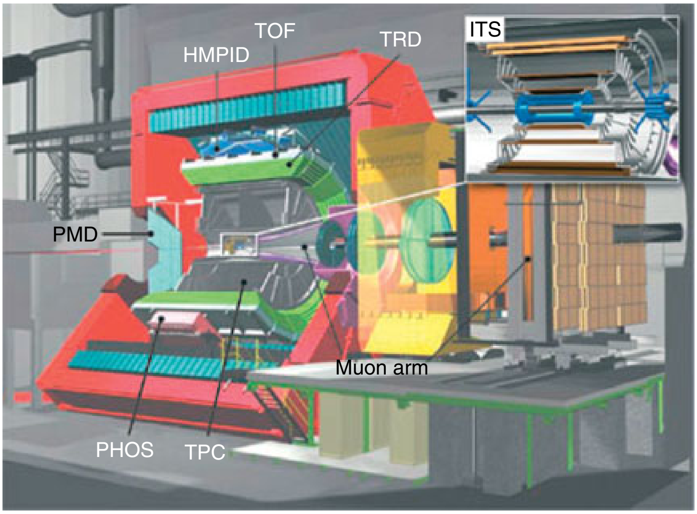

Fig. 23.37 The ALICE detector at the LHC (Braun-Munzinger and Stachel, 2007).

Outside the TPC there are two large particle identification detectors: the transition radiation detector (TRD) and the time-of-flight (TOF). The central barrel of ALICE is completed by dedicated detectors to measure photons (PHOS and EMCal). Separated from the main detector in the forward direction of one of the accelerator beams, and behind a conical absorber that projects into the central barrel, is a muon detector with its own large dipole magnet.

The TPC as the main tracking detector of the ALICE central barrel has to provide – together with the other central barrel detectors – charged-particle momentum measurements with good two-track separation, particle identification, and vertex determination.

At the Pb + Pb design luminosity of the LHC an interaction rate of 8 kHz is expected, of which about 10% are to be considered as central collisions. In the TPC design phase, that is, before RHIC data allowed a better extrapolation to LHC energy, an extreme multiplicity of $dN_{ch}/d\eta = 8000$ was assumed, which would have resulted in 20 000 charged primary and secondary tracks in the acceptance. These extreme multiplicities result in maximum occupancies (defined as the ratio of the number of readout pads and time bins above threshold to all pads and time bins) of about 40% at the innermost radius and 15% at the outermost radius.

It is expected that the TPC can be operated at central collision rates of up to 200 Hz. Simulations have shown that at this rate the space charge due to the ion feedback during gate-open time starts to be comparable to the space charge due to the ionization in the TPC drift volume itself. The true space-charge limit, however, might eventually depend on the background conditions due to the beam quality.

The TPC is cylindrical in shape; the active volume has an inner radius of about 85 cm, an outer radius of about 250 cm, and an overall length along the beam direction of 500 cm. The detector is made of a large cylindrical field cage, filled with 90 m^3 of $Ne/CO_2/N_2$ (90/10/5), which is needed to transport the primary electrons over a distance of up to 2.5 m on either side of the

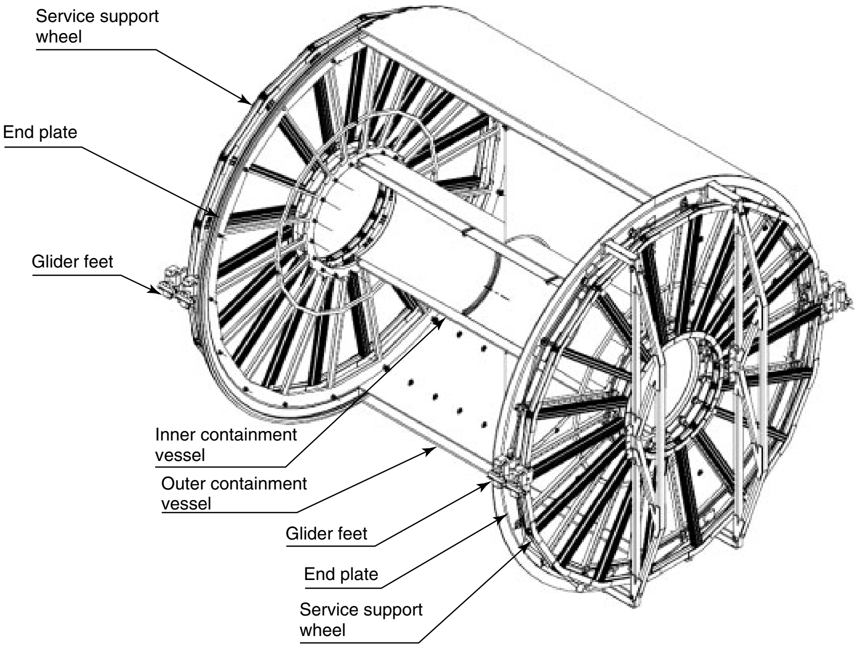

Fig. 23.38 Three-dimensional view of the TPC field cage and service support wheels, without readout chambers (Aamodt *et al.*, 2008).

central electrode to the end plates. MWPCs with cathode pad readout are mounted into 18 trapezoidal sectors of each end plate.

The field cage (Figure 23.38) is based on a design with a central high-voltage electrode and two opposite axial potential dividers, which create a highly uniform electrostatic field in the common gas volume. The central electrode is a stretched Mylar foil to satisfy the requirement of minimal material near 90° relative to the beam direction. The electrical potential in the drift region is defined by aluminized Mylar strips wound around 18 inner and outer support rods, which also contain two resistive potential dividers (inner and outer). Because of the $Ne/CO_2/N_2$ gas mixture used in the TPC, the field cage is operated at high-voltage gradients, of about 400 $V\ cm^{-1}$, with a high voltage of 100 kV at the central electrode, which results in a maximum drift time of about 90 microseconds. The drift gas is optimized for drift speed, low diffusion, low radiation length and hence low multiple scattering, small space-charge effect, and aging and stability properties.

The readout chambers instrument the two end plates of the TPC cylinder with an overall active area of $32.5\,m^2$. The chambers are MWPCs with cathode pad readout. Because of the radial dependence of the track density, the readout is segmented into two readout chambers with slightly different wire geometry adapted to the varying pad sizes. The readout chambers are made of standard wire planes, that is, they consist of a grid of anode wires above the pad plane, a cathode wire plane, and a gating grid. They combine a very tight electrode configuration (anode-to-cathode distance 2

Tab. 23.2 Synopsis of the ALICE TPC parameters (Aamodt *et al.*, 2008).

Pseudorapidity coverage	$-0.9 < \eta < 0.9$ for full radial track length
	$-1.5 < \eta < 1.5$ for 1/3 radial track length
Azimuthal coverage	360°
Radial position (active volume)	$848 < r < 2466$ mm
Radial size of vessel (outer dimensions)	$606.5 < r < 2780$ mm
Radial size of vessel (drift gas volume)	$788 < r < 2580$ mm
Length (active volume)	2×2500 mm
Segmentation in φ	18 sectors
Segmentation in r	2 chambers per sector
Total number of readout chambers	$2 \times 2 \times 18 = 72$
Inner readout chamber geometry	Trapezoidal, $848 < r < 1321$ mm active area
Pad size	4×7.5 mm ($\varphi \times r$)
Pad rows	63
Total pads	5504
Outer readout chamber geometry	Trapezoidal, $1346 < r < 2466$ mm active area
Pad size	6×10 and 6×15 mm ($\varphi \times r$)
Pad rows	$64 + 32 = 96$ (small and large pads)
Total pads	$5952 + 4032 = 9984$ (small and large pads)
Detector gas	$Ne/CO_2/N_2$ 90/10/5
Gas volume	$90\,m^3$
Drift length	2×2500 mm
Drift field	$400\,V\,cm^{-1}$
Drift velocity	$2.7\,cm\,\mu s^{-1}$
Maximum drift time	$92\,\mu s$
Total HV	100 kV
Diffusion	$D_L = D_T = 220\,\mu m\,cm^{-\frac{1}{2}}$
Material budget	$X/X_0 = 3.5\%$ near $\eta = 0$
Front-end cards (FEC)	121 per sector $\times$ 36 = 4356
Readout Control Unit (RCU) scheme	6 per sector, 18–25 FEC per RCU
Total RCUs	216
Total pads–readout channels	557 568
Pad occupancy (for $dN/dy = 8000$)	40–15% inner/outer radius
Pad occupancy (for pp)	$5-2 \times 10^{-4}$ inner/outer radius
Event size (for $dN/dy = 8000$)	~90 MB (conditions see text)
Event size (for pp)	~1–4 MB (depending on pileup and noise threshold)
Total bandwidth	$\sim 30\,GB\,s^{-1}$
Trigger rate limits	200 Hz Pb–Pb central events
	1000 Hz proton–proton events
ADC	10 bit
Sampling frequency	5–10 MHz
Time samples	500–1000
Conversion gain	6 ADC counts/fC
Position resolution (σ) in $r\varphi$	1100–800 μm inner/outer radii
in z	1250–1100 μm
dE/dx resolution, isolated tracks	5.0%
$dN/dy = 8000$	6.8%

and 3 mm for inner and outer readout chamber, respectively) with an operating gain of up to 20 000. To keep the occupancy as low as possible and to ensure the necessary dE/dx, position and two-track resolution, there are about 560 000 readout pads of three different sizes: 4×7.5 mm^2 in the inner chambers, and 6×10 mm^2 and 6×15 mm^2 in the outer chambers. The readout chambers are normally closed by a gating grid for electrons coming from the drift volume and for ions from the amplification region. The gate is opened only by a valid trigger for the duration of one drift-time interval, thus reducing the building up of space charge due to positive ions drifting back from the multiplication region for nontriggered interactions and background. It has been verified that the closed gate cuts the drift of electrons by a factor better than 10^5. The main parameters of the TPC design are summarized in Table 23.2.

23.3.2
The Silicon Tracker and Vertex Detector

Bunches of 10^{11} protons will collide 40 million times per second inside the LHC to provide 14-TeV proton–proton collisions at the LHC design luminosity of 10^{34} cm^{-2} s^{-1}. Around 1000 charged particles will emerge from the collision points every 25 nanoseconds, within a volume defined by $|\eta| < 2.5$, where pseudorapidity η is a logarithmic measure of the polar angle of the emitted particle relative to the beam axis, $\theta : \eta = -\ln(\tan(\theta/2))$.

The main challenges for the inner detector parts are the high particle rates, the radiation tolerance needed, and the control of aging effects. The central trackers of ATLAS and CMS, the two large general-purpose experiments, are similar in size and features. They are located in the central part of the detector surrounding the beam pipe and are about 2 m in

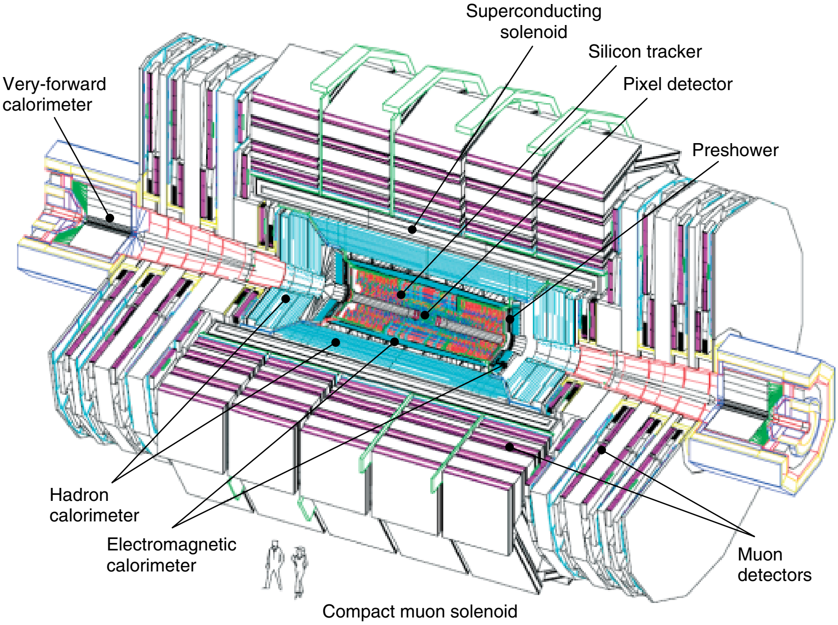

Fig. 23.39 The CMS (Compact Muon Solenoid) detector at LHC (Chatrchyan *et al.*, 2008).

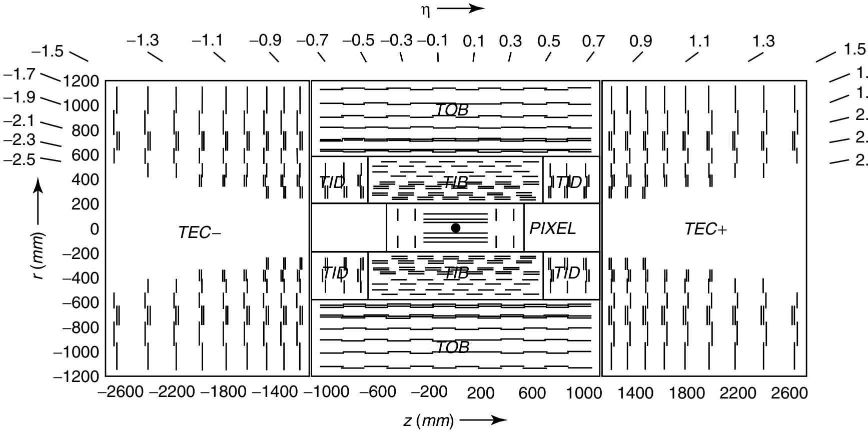

Fig. 23.40 Schematic cross section through the CMS tracker. Each line represents a detector module. Double lines indicate back-to-back modules that deliver stereo hits (Chatrchyan *et al.*, 2008).

diameter and 6 m in length. They cover the angular range $|\eta| < 2.5$ and are both immersed in a solenoidal magnetic field which is 4 and 2 T for CMS and ATLAS, respectively. Figure 23.39 shows the overall layout of the CMS detector (Chatrchyan *et al.*, 2008); the inner tracker is located inside the bore of the magnet coil of the 4 T superconducting solenoid. In both trackers – ATLAS and CMS – the innermost detector layers are built with silicon pixels and the intermediate layers with silicon strip detectors with high strip density (see Figures 23.40 and 23.43a). The pixel detectors that surround the interaction point with a set of three measurements per primary track at small radii have a pixel size of $50\,\mu\text{m} \times 400\,\mu\text{m}$ (ATLAS) and $100\,\mu\text{m} \times 150\,\mu\text{m}$ (CMS) in $R\varphi \times z$. The inner silicon strip detectors provide high-resolution measurements ($\sim 25\,\mu\text{m}$).

The barrel of ATLAS is located between $r = 30$ and $r = 50$ cm and provides eight measurements while that of CMS is located between $r = 20$ cm and $r = 50$ cm and provides six points. The technology used for the outer layers is different as well: CMS uses silicon strip detectors with coarse pitch providing eight precise measurement points ($\sim 50\,\mu\text{m}$) between $r = 60$ and $r = 110$ cm, while ATLAS uses a transition radiation detector (TRT) with 4-mm diameter gas straw tubes providing 35 points with $\sim 170\,\mu\text{m}$ resolution located between $r = 55$ and $r = 105$ cm (Ragusa and Rolandi, 2007 and references therein). The main parameters of the ATLAS and CMS tracking systems are given in Table 23.3.

Figure 23.41 illustrates the tracking capabilities of the ATLAS inner detector. A charged particle of 10 GeV/c transverse momentum, produced at the primary vertex, successively traverses the beryllium beam pipe, the three cylindrical silicon-pixel layers with individual sensor elements of $50 \times 400\,\mu\text{m}^2$, the four cylindrical double layers (one axial and one with a stereo angle of 40 mrad) of barrel silicon-microstrip sensors (SCT) of pitch $80\,\mu\text{m}$, and approximately 36 axial straws of 4-mm diameter contained in the barrel transition–radiation tracker modules.

Tab. 23.3 Main parameters of the ATLAS and CMS tracking systems (Froidevaux and Sphicas, 2006).

Parameter	**ATLAS**	**CMS**
Dimensions (cm)		
-Radius of outermost measurement	101–107	107–110
-Radius of innermost measurement	5.0	4.4
-Total active length	560	540
Magnetic field B (T)	2	4
BR^2 (T_m^2)	2.0–2.3	4.6–4.8
Total power on detector (kW)	70	60
Total weight in tracker volume (kg)	≈4500	≈3700
Total material (X/X_0)		
-At $\eta \approx 0$ (minimum material)	0.3	0.4
-At $\eta \approx 1.7$ (maximum material)	1.2	1.5
-At $\eta \approx 2.5$ (edge of acceptance)	0.5	0.8
Total material (λ/λ_0 at maximum)	0.35	0.42
Silicon microstrip detectors		
-Number of hits per track	8	14
-Radius of innermost measurement (cm)	30	20
-Total active area of silicon (m^2)	60	200
-Wafer thickness (microns)	280	320/500
-Total number of channels	6.2×10^6	9.6×10^6
-Cell size (μm in $R\phi \times$ cm in z/R)	80×12	$80/120 \times 10$
-Cell size (μm in $R\phi \times$ cm in z/R)		and $120/180 \times 25$
Straw drift tubes (ATLAS only)		
-Number of hits per track ($\lvert\eta\rvert < 1.8$)	35	
-Total number of channels	350 000	
-Cell size (mm in $R\phi \times$ cm in z)	4×70 (barrel) 4×40 (end caps)	

The innermost layer(s) of the tracking system nearest to the interaction vertex are usually called *vertex detectors*, even though they provide tracking information. The purpose of such a detector is to measure the starting point of an event with great accuracy and, in addition, to track the decay origin of any short-lived particles that may have been created. Typically, a resolution of $\leq 50\,\mu$m is required of such a detector, measuring particles with lifetimes of around 10^{-13} seconds. Figures 23.42 and 23.43b show examples of such vertex detectors that consist of silicon pixel detectors. Table 23.4 summarizes the main parameters of the ATLAS and CMS pixel systems.

23.3.3
Large-Volume Muon Trackers – the ATLAS Muon Spectrometer

Muon detection in high-energy physics experiments is usually accomplished by using thick absorbers as particle filters followed by MWPCs or planar drift chambers. The dimensions of these absorbers (typically iron) are in the order of meters to prevent “punch through” of hadrons. Therefore, muon detectors tend to be

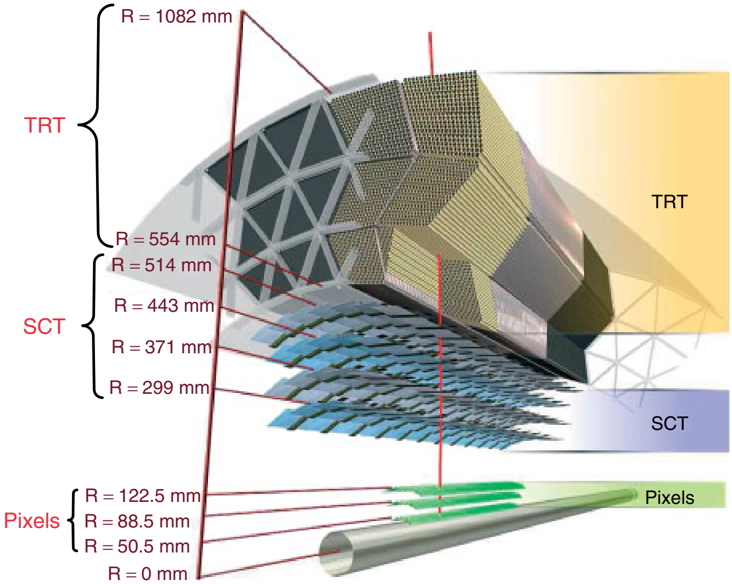
R = 1082 mm
TRT
R = 554 mm
R = 514 mm
SCT
R = 443 mm
R = 371 mm
R = 299 mm
Pixels
R = 122.5 mm
R = 88.5 mm
R = 50.5 mm
R = 0 mm
TRT
SCT
Pixels

Fig. 23.41 The ATLAS inner tracking detector traversed by a charged track of 10-GeV transverse momentum (Aad *et al.*, 2008).

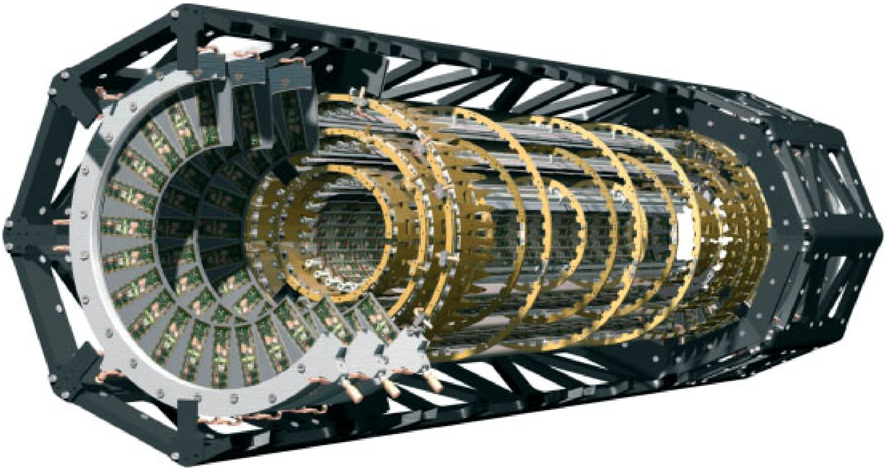

Fig. 23.42 A perspective cut-away view of the ATLAS pixel detector. The view shows individual barrel and end-cap modules, supported with their associated services on staves and disks within an octagonal support frame (Aad *et al.*, 2008).

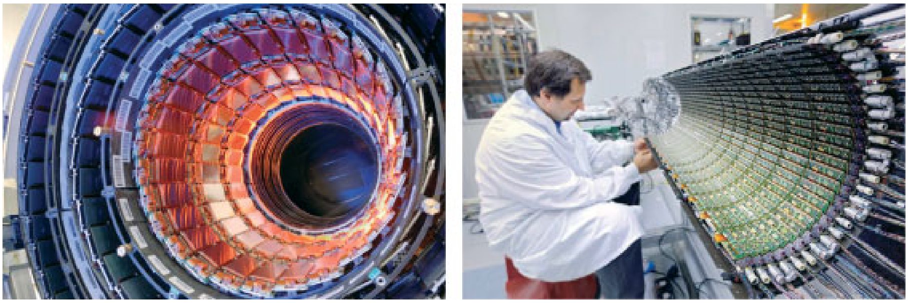
(a) (b)

Fig. 23.43 (a) Silicon microstrip detectors in the CMS barrel region. (b) The ATLAS pixel detector during the final assembly stage. Images reproduced with permission from CERN.

large and rather costly if good coverage in solid angle is required for muon detection.

The choice of magnet in ATLAS was motivated by the measurement method of muons with momenta up to teraelectronvolt scales (Figure 23.44). ATLAS opted for a high-resolution, stand-alone measurement independent of the rest of the subdetectors, resulting in a large volume with low material density. The ATLAS toroidal magnetic field provides a momentum resolution that is essentially independent of pseudorapidity η to a value of 2.7. The muon system consists of three large superconducting air-core toroid magnets, which are instrumented with different types of chambers to provide two necessary functions, namely, high-precision tracking and triggering (Figure 23.45). The central (or barrel) region, $|\eta| < 1.0$, is covered by a large barrel magnet consisting of eight coils that surround the hadronic calorimeter. In this region, tracks are measured in chambers arranged in three cylindrical layers (stations) around the beam axis. In the end-cap region, $1.4 < |\eta| < 2.7$, muon tracks are bent in two smaller end-cap magnets each of which is inserted into one end of the barrel toroid. The intermediate region, $1.0 < |\eta| < 1.4$, is less straightforward because here the barrel and end-cap fields overlap, thus partially reducing the bending power (Stapnes, 2007; Froidevaux and Sphicas, 2006).

23.3.3.1 Precision Chambers for Tracking

The basic element of the precision momentum measurement is the monitored drift tube (MDT) chamber. It combines high measuring accuracy with mechanical rigidity and simplicity of construction. The tubes are operated at an absolute pressure of 3 atm, improving the measuring accuracy by about a factor of $\sqrt{3}$ compared to operation at normal pressure. A resolution of 80 μm per tube has been achieved. In the very forward region ($2 < |\eta| < 2.7$) cathode strip chambers (CSCs) are used because of higher rate capability and time resolution (Aad *et al.*, 2008).

Monitored Drift Tube (MDT) chamber The basic element of the MDT chambers is the pressurized drift tube with a diameter of 30 mm, operating with Ar/CO_2 gas (93 : 7) at 3 bar. Electrons created by ionization are collected at the central tungsten–rhenium

Tab. 23.4 Main parameters of the ATLAS and CMS pixel systems (Froidevaux and Sphicas, 2006).

	ATLAS	**CMS**
Number of hits per track	3	3
Total number of channels	$80\ 10^6$	$66\ 10^6$
Pixel size (μm in $R\phi$ × μm in z/R)	50 × 400	100 × 150
Lorentz angle (degrees), initial to end	12 to 4	26 to 8
Tilt in $R\phi$ (degrees)	20 (only barrel)	20 (only end cap)
Total active area of silicon (m^2)	1.7 (n^+/n)	1.0 (n^+/n)
Sensor thickness (μm)	250	285
Total number of modules	1744 (288 in disks)	1440 (672 in disks)
Barrel layer radii (cm)	5.1, 8.9, 12.3	4.4, 7.3, 10.2
Disk layer minimum to maximum radii (cm)	8.9–15.0	6.0–15.0
Disk positions in z (cm)	49.5, 58.0, 65.0	34.5, 46.5
Signal-to-noise ratio for minimum-ionizing particles (day 1)	120	130
Total fluence at $L = 10^{34}$ ($n_{eq}/cm^2/year$) at radius of 4–5 cm (innermost layer)	3×10^{14}	3×10^{14}
Signal-to-noise ratio (after $10^{15} n_{eq}/cm^2$)	80	80
Resolution in $R\phi$ (μm)	≈10	≈10
Resolution in z/R (μm)	≈100	≈20

wire with a diameter of 50 μm at a potential of about 3080 V. The wire is held in position at the tube ends by a cylindrical element, which guarantees the centricity of the wire with respect to the tube with an accuracy of $\sigma < 10$ μm (Figure 23.46). All regular MDT chambers consist of two packages of tube layers (multilayers) separated by a support structure, where each multilayer consists of three tube layers in the outer and four tube layers in the inner layers of the muon detector. The exact definition of the wire positions in an MDT, due to tight tolerances in production and continued monitoring of global deformations, results in the accuracy of the track segment in the 6 (8) tube layers being only limited by the error of the single tube measurements of about 80 μm. Thus, the resolution of the central point of a track segment in a multilayer of 3 (4) tubes is 50 (40) μm, while the one in a chamber of two multilayers is 35 (30) μm. The limit for safe operation of the MDTs is at counting rates of about 150 Hz cm^{-2}.

Cathode Strip Chamber (CSC) The CSCs are MWPCs with cathodes segmented into strips, the direction in both cathodes being orthogonal to each other. This allows the measurement of both coordinates from the distribution of the induced charge (Figure 23.47). The resolution of the CSCs in the bending plane will be 60 μm. Operation is considered safe up to counting rates of about 1000 Hz cm^{-2}, which is sufficient up to the forward boundary of the muon system at $|\eta| = 2.7$.

23.3.3.2 **Trigger Chambers**

RPCs in the barrel region and thin gap chambers (TGCs) in the end-cap region provide a fast response with good time resolution (inside 15–20 nanoseconds) and can therefore unambiguously identify the

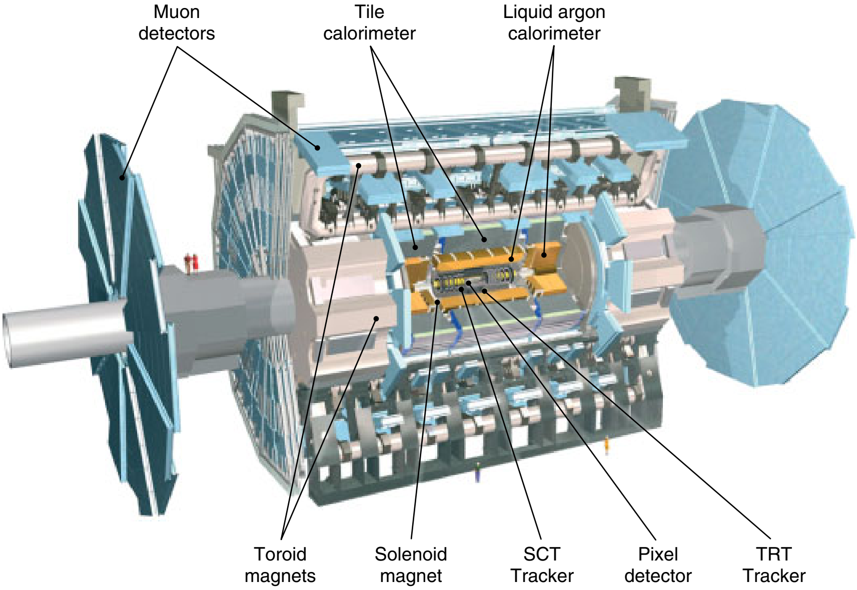
Muon detectors
Tile calorimeter
Liquid argon calorimeter
Toroid magnets
Solenoid magnet
SCT Tracker
Pixel detector
TRT Tracker

Fig. 23.44 Overall layout of the ATLAS detector (Aad *et al.*, 2008).

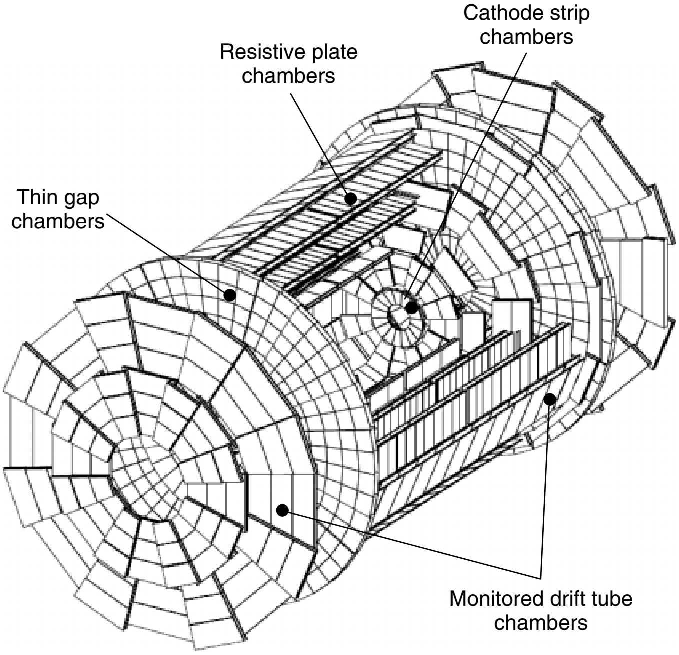
Cathode strip chambers
Resistive plate chambers
Thin gap chambers
Monitored drift tube chambers

Fig. 23.45 Layout of the ATLAS muon system displaying the regions in which the different muon chamber technologies are used (Aad *et al.*, 2008).

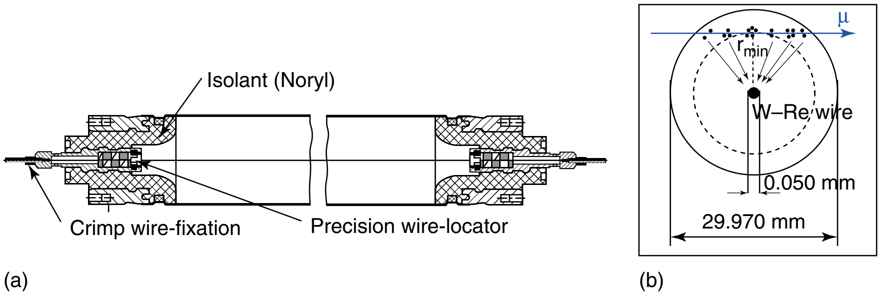

Fig. 23.46 Longitudinal cut (a) and cross section of an MDT tube (b) (Aad *et al.*, 2008).

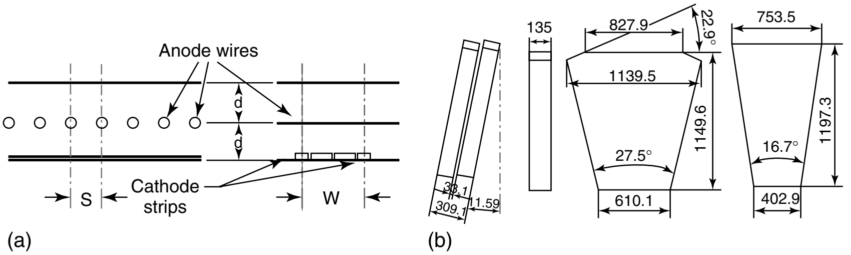

Fig. 23.47 (a) Structure of the CSC-cell and cathode segmentation in the bending plane. (b) Shapes of the CSC modules (Aad *et al.*, 2008).

correct bunch crossing and are invaluable for triggering.

Resistive Plate Chamber (RPC) The trigger system in the barrel consists of three cylindrical layers of RPCs around the longitudinal axis. Two resistive plates made from plastic laminate are kept parallel to each other at a distance of 2 mm by insulating spacers. The electric field between the plates of about $4.9\,\mathrm{kV\,mm^{-1}}$ allows avalanches to form along the ionizing tracks toward the anode. The signal is read out via capacitative coupling to metallic strips, perpendicular to each other, on both sides of the detector, that is, on the outer face of the resistive plates. The gas used is a mixture of $C_2H_2F_4$/Iso–C_4 H_{10}/SF_6, which combines relatively low operating voltage, nonflammability, and low cost, while providing a sufficient plateau for safe operation. An RPC trigger chamber consists of two rectangular detectors, contiguous to each other. Each unit has two independent resistive plate structures, so-called gas volumes, which are each read out by two orthogonal groups of pickup strips (Figure 23.48).

In the high-background environment encountered at the LHC, the RPCs are operated in the avalanche mode that offers the benefit of higher rate capability and rate-independent time resolution. At the nominal operating voltage of 9.7 kV, a signal with a width of about 5 nanoseconds is generated by the track with a streamer probability of less than 1%.

Thin Gap Chamber (TGC) Thin gap chambers provide two functions in the end-cap of the muon spectrometer: the generation

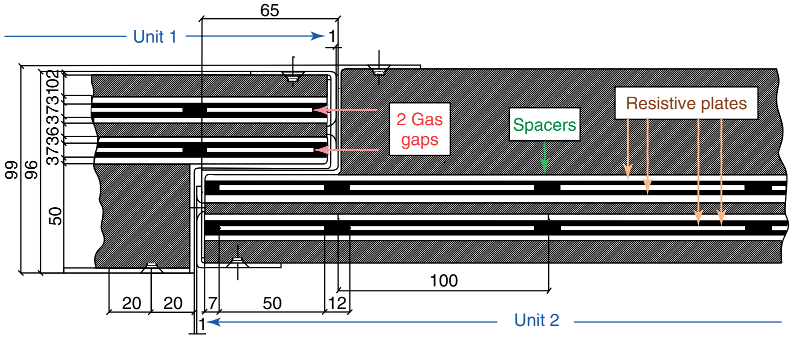

Fig. 23.48 Cross section through an RPC, where two units are joined to form a chamber. Each unit has two gas volumes (gas gaps), supported by spacers, four resistive electrodes and four readout planes, reading the transverse and longitudinal direction. The sandwich structure (hashed) is made from paper honeycomb (Aad *et al.*, 2008).

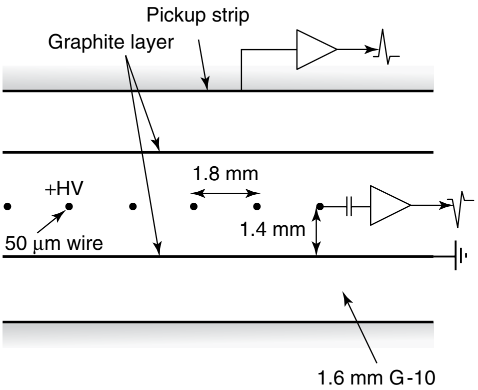

Fig. 23.49 TGC structure showing anode wires, graphite cathodes, laminate (G-10) layers, and a readout strip, orthogonal to the wires (Aad *et al.*, 2008).

of a muon trigger and the determination of the second, azimuthal coordinate to complement the measurement of the MDTs in the bending direction. TGCs are MWPCs with the characteristics that the wire-to-wire distance (1.8 mm) is larger than the wire-to-wall distance (1.4 mm) (see Figure 23.49). With a highly quenching gas mixture of CO_2 and n-C_5 H_{12} (n-pentane), this cell geometry allows for operation in saturated mode, that is, with a very high gas gain ($\sim 10^6$), which makes pulse heights highly uniform and nearly independent of the amount of primary ionization.

23.4
Applications of Particle Detectors in the Future

Particle tracking is, in its true sense, a visualization of complex event structures in nuclear and subnuclear reactions. To disentangle and measure individual tracks with increasingly higher precision has been the aim of every tracking detector in the past and continues to be so at present. As a result of continuous improvement in the hardware technologies combined with powerful pattern-recognition algorithms, tracking systems have acquired state-of-the-art performance in terms of accuracy and speed. Nearly reaching the grain size of particle tracks as ‘‘seen’’ in holographic bubble chambers (in the order of microns), today’s electronic detectors measure, filter, and analyze track data within microseconds or even less.

The R&D activity in the field of tracking detectors is mainly defined by the requirements of detection systems for e^+/e^- Linear Colliders in the TeV or even multi-TeV energy range and the upgrade of LHC detectors, but there are many other detector systems currently being developed for, for example, neutrino experiments, flavor studies at high intensity and lower energy, and heavy-ion collisions at low energy and high intensity.

The proposed luminosity upgrade of the LHC by a factor of 10 will require significant upgrades for the LHC detectors, especially the replacement of the inner tracker systems. New trackers will have also to be redesigned to handle the increased occupancy rates and the increased radiation background. For the innermost layers of the vertex detectors in, for example, ATLAS and CMS, the levels of radiation background are so high that new technology is required. The current favorites are three-dimensional silicon and diamond, but other options are being considered such as micromegas and thin Si planar detectors (Dawson, 2008).

Detector R&D for the ILC is concentrated in three different design studies (ILC, 2008). The requirements for the tracking system and the vertex detector are challenging: robust and highly efficient tracking in a large acceptance, excellent space point resolution (better than 4 μm for the vertex detector), excellent impact parameter resolution, and a very low material budget ($\sim$0.1% X_0 per layer of the innermost layers). The three different detector concepts are conceptually very similar with respect to the tracking and vertex detector, but they pursue different implementations. One design relies completely on silicon sensors, and the other two consider a combination of silicon and gaseous tracking detectors. A number of candidate technologies are being developed for ILC vertex detector applications, all based on the integration of the detecting element with the electronics: DEPFEPs, MAPS, CCDs, and SOI technology (Collins, 2007). A TPC is a candidate for a large and robust tracking system. Traditional TPCs are equipped with MPWCs for gas amplification and a gating grid where two consecutive wires are polarized at opposite voltages, thus creating a transverse field that stops most of the secondary ions produced in the amplification region before they reach the drift space. Especially in high-rate applications and under high-background conditions, like at the ILC, it is very important to have a very limited ion backflow from the secondary ions in order to avoid distortions of the drift electric field. A Micromegas micromesh has the intrinsic property to naturally stop a large fraction of the secondary positive ions and is therefore studied to read out the TPC instead of an MWPC (Colas *et al.*, 2004). Another approach is to use GEMs for the readout (Ledermann *et al.*, 2006). The fine granularity and high-rate capability of micropattern devices can be fully exploited using a high-density pixel readout with a size corresponding to the intrinsic width of the detected avalanche charge. A CMOS pixel chip assembled directly below the GEM or Micromegas amplification structure serves as an integrated charge collecting anode. With this arrangement, avalanche electrons are collected on the top metal layer of the CMOS ASIC and every input pixel is then directly connected to the amplification, digitization, and sparsification circuits integrated in the underlying active layers of the CMOS technology. Using this approach, gas detectors can reach the level of integration typical of solid-state pixel devices and a high spatial resolution that allows to count individual clusters can be achieved (Titov, 2007 and references therein, Visschers *et al.*, 2007).

23.4.1
Future Trends

Technology, especially microelectronics, has and will in the future enable advances in detectors: segmentation, speed, feature size, and new materials. For example, the emerging new technique of three-dimensional integration of thin ASICs makes it possible to stack different layers of sensors and ASICs together to form a sandwich, allowing the choice of the optimal technology for each tier (Umemoto *et al.*, 2004). Integration issues – mechanical stability of very thin and large detector systems, power supply, and cooling will become even more important – not to speak about the readout and transfer of large amounts of data produced by a (self-triggered) many-channel front-end electronics. The boundaries between vertex and tracking detectors will become even more blurred – vertex detectors are stand-alone trackers. Large volume gaseous detectors will be combined with high-granularity CMOS pixel readout. Calorimeters will provide three-dimensional space point information: the particle flow calorimetry (Thomson, 2008; Schätzel, 2008).

Glossary

Affinity: Ability to bind charge carriers to neutral molecules. Oxygen and Freon, for example, are known to capture free electrons in drift gases. These charges are lost in the final signal arriving at the anode.

ALEPH: A large (about 4000 tons) 4π solenoidal detector for the LEP electron–positron storage ring facility.

ALICE: Experiment at the LHC designed especially to study heavy ion collisions.

ATLAS: A large general-purpose experiment at the LHC.

Bunch: In most particle accelerators and storing ring facilities, particles are accelerated in a noncontinuous way. For optimum efficiency and luminosity, particles are injected and accelerated in discrete packages or bunches that, depending on their number, are separated in time from several milliseconds to tens of nanoseconds. The spatial dimensions of a bunch are typically a few micrometers laterally and of order centimeters longitudinally (i.e., along the direction of motion).

Calorimeter: In particle physics, a detector to measure the total energy of neutral (photons) and charged particles by total absorption in a dense medium such as iron, lead, uranium, or others. One distinguishes between two classes of calorimetry: electromagnetic calorimetry for the energy measurement of photons and electrons (positions), and hadronic calorimetry for the interaction of charged hadrons (primarily pions).

CERN: Centre Européenne pour la Recherche Nucléaire: The world's largest laboratory in high-energy particle physics, based in Geneva, Switzerland.

Chemical Vapor Deposition (CVD): A catalytic process to produce thin sheets (up to 100 μm) of diamond from carbon-based raw materials.

Cherenkov Light: Electromagnetic radiation emitted by a particle moving (with constant velocity) through a material medium at a velocity greater than the phase velocity of light in that medium. This phenomenon is exploited for the identification of charged particles passing through matter.

CMOS, NMOS, PMOS: Complementary-metal-oxide-semi-conductor (CMOS) uses complementary and symmetrical pairs of p-type and n-type metal-oxide semiconductor field effect transistors (MOSFETs) for logic functions. PMOS logic uses p-type, NMOS logic n-type MOSFETs to implement logic gates and other digital circuits.

CMS: A large general-purpose experiment at the LHC.

DESY: (Deutsches Elektronen-Synchrotron): A German-based research facility for high-energy particle physics similar to CERN.

Electron-Beam Lithography: A process known in semiconductor technology to produce by means of photographic masks thin metal layers (order micrometers) on solid-state or insulating substrates. The short wavelength of the electron beam allows accuracies below 1 μm.

Flash Analog-to-Digital Converter (FADC): A device that samples and digitizes the time response and pulse height of analog signals (coming from a detector) in time slices comparable to the speed of the signal itself. The reconstruction of the time buckets allows, a recovery of the original signals and their timing history.

FWHM: Full width at half maximum.

Gaussian Curve, Gauss Function: The "normal," bell-shaped distribution, which is symmetric about its mean value.

Hadron Collider: An accelerator primarily for protons (antiprotons) and heavy ions. These machines are in contrast to electron (positron) facilities, which are lower in energy because of energy losses from synchrotron radiation.

HERA (Hadron-Electron Ring Accelerator): Facility for colliding electrons of 30 GeV on protons of 820 GeV at DESY, Germany.

Hodoscope: An arrangement of detectors, usually an array of scintillators with photomultipliers, used to count and measure the position and trajectory of incoming particles to an experiment.

Holographic Bubble Chamber: Photographic method of recording particle tracks in a bubble chamber using interferometry. This improves the spatial resolution (bubble size) significantly.

Jet: In most cases, the fragmentation products of a quark decay into hadrons. Depending on the initial energy of the parent particle, jets appear as collimated particle tracks with high multiplicity. Their high density of tracks makes them particularly difficult to measure in common detectors because of insufficient separation of tracks.

LHC (Large Hadron Collider): This storage ring accelerator will make use of the present underground tunnel built for LEP. Bunches of 10^{11} protons will collide 40 million times per second to provide 14-TeV proton–proton collisions at the LHC design luminosity of $10^{34}\,\text{cm}^{-2}\,\text{s}^{-1}$. The acceleration of nuclei will provide Pb + Pb collisions at $\sqrt{s_{NN}} = 5.5$ TeV; the interaction rate is expected to be 8 kHz.

Luminosity: In analogy to optical terms, luminosity describes the probability of accelerated particles to interact with each other in head-on collisions. The luminosity depends on the particle density (usually confined in a bunch) *and* the interaction cross section. Since rare phenomena have very low production cross sections, accelerators and colliders are designed for maximum luminosity.

LHCb: Experiment at the LHC designed especially to study b-quark systems.

Minimum-Ionizing Particle: Charged particles whose energy loss (depending on their velocity) in a medium is near or inside the minimum of the Bethe–Bloch function. Most charged particles produced in a high-energy physics interaction have a velocity range where the ionization from energy losses (dE/dx) falls into this minimum. This makes them difficult to detect, predominantly in gas-filled detectors, where only about 100 electrons/cm are liberated by these particles.

Occupancy: Term describing the ''busy time'' of a single detector channel due to ionizing radiation, noise, and pickup, thus creating true plus uncorrelated hits. It is commonly expressed in percent and depends primarily on the incoming flux of particles and the intrinsic speed of the detector itself; that is, gaseous detectors with long signal-processing time (e.g., drift time, pulse shaping) suffer from higher occupancies than fast devices such as scintillators. High occupancies (i.e., $>10\%$) result in substantial detector dead time and lead to confusion in the pattern recognition of genuine tracks. To reduce detector occupancy in high-background environments to tolerable levels, small cell sizes must be chosen.

Pattern Recognition: Sophisticated software algorithms to disentangle complex event structures and select physically interesting tracks from confusing backgrounds.

Photomultiplier: A device consisting of a photosensitive surface inside a vacuum tube where electrons are liberated from the cathode. Each electron is multiplied in a succession of dynodes by secondary emission under the influence of the electric field. Photomultipliers are essentially sensitive to single photons and provide very fast signals in the nanosecond or lower range.

Pitch: The distance between adjacent cells in detectors with repetitive cell structures. Also known as *cell width*.

Pseudorapidity: Pseudorapidity η is a logarithmic measure of the polar angle of the emitted particle relative to the beam axis, $\theta : \eta = -\ln(\tan(\theta/2))$. Emission at 90° corresponds to $\eta = 0$; 45° to $\eta \approx 0.9$.

RHIC: Relativistic Heavy-Ion Collider at Brookhaven National Laboratory, New York. Provides beams of heavy ions and polarized protons (Au + Au collision at $\sqrt{s_{NN}} = 200$ GeV).

Scintillation Light: Light produced by quantum effects in atoms excited by charged particles traversing the medium.

SLAC: National Accelerator Laboratory at Stanford is home to a two-mile linear accelerator, the Stanford Linear Collider and the Babar experiment.

Strongly Ionizing Particle: Typically low- (megaelectronvolts) and medium-energy (order 100 MeV) charged particles whose energy loss in matter is large enough to produce a high rate of electron–positive-ion pairs. These are in contrast to minimum-ionizing particles (order gigaelectronvolts), which are more common in high-energy physics.

Tevatron: Proton–Antiproton Collider with 2-TeV center-of-mass energy located at Fermi National Laboratory, Batavia, Illinois.

Vertex: The origin of a particle interaction. This interaction can be the collision of two accelerated beam particles or the location of a parent particle prior to its decay. In the latter case, the position of the decaying particle is reconstructed from the tracks of its decay products.

References

Aad, G. *et al.* The ATLAS Collaboration (**2008**) *J. Instrum.*, **3**, S08003.

Aamodt, K. *et al.* The ALICE Collaboration (**2008**) *J. Instrum.*, **3**, S08002.

Aarnio, P.A. *et al.* The DELPHI Collaboration (**1991**) *Nucl. Phys. A*, **303**, 233–76.

Abe, K. *et al.* (**1997**) *Nucl. Instrum. Methods A*, **400**, 287.

Ackermann, K.H. *et al.* (**1999**) *Nucl. Phys. A*, **661**, 681c–685c.

Ackermann, K.H. *et al.* (The STAR Collaboration) (**2003**) *Nucl. Phys. A*, **499**, 624–32.

Adamczyk, M. *et al.* (The BRAHMS Collaboration) (**2003**) *Nucl. Instrum. Methods A*, **499**, 437–68.

Adeva, B. *et al.* (The L3 Collaboration) (**1990**) *Nucl. Instrum. Methods A*, **289**, 35–102.

Afanasiev, F. *et al.* (The NA49 Collaboration) (**1999**) *Nucl. Instrum. Methods A*, **430**, 210.

ALEPH (**1990**) *Nucl. Instrum. Methods A*, **294**, 127–78.

ALEPH (**1991**) *Phys. Lett. B*, **273**, 181–92.

ALICE (**1995**) Technical Proposal for a Large Ion Collider Experiment at the CERN LHC, CERN-LHCC-95-71.

Amelio, G.F. *et al.* (**1970**) *Bell. Syst. Tech. J.*, **49**, 593.

Anderhub, H. *et al.* (**1985**) *Nucl. Instrum. Methods A*, **235**, 267–78.

Angelini, F. *et al.* (**1991**) *Nucl. Phys. B Proc. Suppl.*, **23A**, 254–60.

Angelini, F. *et al.* (**1992**) *Nucl. Instrum. Methods A*, **323**, 229–35.

Angelini, F. *et al.* (**1993a**) *Nucl. Instrum. Methods A*, **335**, 69–77.

Angelini, F. *et al.* (**1993b**) *Nucl. Instrum. Methods A*, **336**, 106–115.

Anghinolfi, F. *et al.* (**1992**) *IEEE Trans. Nucl. Sci.*, **NS-39**, 654–61.

Ankoviak, K. *et al.* (**1989**) *Nucl. Instrum. Methods A*, **279**, 83–90.

Ansari, R. *et al.* (**1989**) *Nucl. Phys. A*, **279**, 388–95.

Antreasyan, D. *et al.* (**1990**) A Device for Particle Detection at Future Hadron Colliders: The Gaseous Pixel Chamber, CERN–European Committee on Future Accelerators Proceedings. Vol. 3, *CERN Report No.* 90-10, *ECFA Report No.* 90-133, CERN, pp. 230–33.

Arai, Y. *et al.* (**2007**) *SOI Detector R&D: Past and Future*. 1st SOI Detector R&D Workshop, March 6, 2007, KEK.

Atac, M. *et al.* (**1986**) *Nucl. Instrum. Methods A*, **249**, 265–76.

Atwood, W.B. *et al.* (**1991**) *Nucl. Instrum. Methods A*, **306**, 446–58.

Augusto Alves, A. Jr. *et al.* (The LHCb Collaboration) (**2008**) *J. Instrum.*, **3**, S08005.

Back, B.B. *et al.* The PHOBOS Collaboration (**2003**) *Nucl. Instrum. Methods A*, **499**, 603.

Bagaturia, Y. *et al.* (**2002**) *Nucl. Instrum. Methods A*, **490**, 223.

Barberis, E. *et al.* (**1993**) *Nucl. Instrum. Methods A*, **326**, 373–380.

Barbero, M. *et al.* (The RD42 Collaboration) (**2007**) Development of diamond tracking detectors for high luminosity experiments at LHC. CERN/LHCC 2007-002.

Bencivenni, G. *et al.* (**2007**) *Nucl. Instrum. Methods A*, **581**, 221–24.

Bencze, G. *et al.* (**1993**) *Nucl. Phys. B Proc. Suppl.*, **32**, 172–180.

Benlloch, J. *et al.* (**1998a**) *IEEE Trans. Nucl. Sci.*, **45**, 234.

Benlloch, J. *et al.* (**1998b**) *Nucl. Instrum. Methods A*, **419**, 410.

Beole, S. *et al.* (**2007**) *Nucl. Instrum. Methods A*, **570**, 236–40.

Bondar, A. *et al.* (**2006**) *Nucl. Instrum. Methods A*, **556**, 495.

Bouclier, R. *et al.* (**1983**) *Nucl. Instrum. Methods A*, **205**, 403–23.

Bouclier, R. *et al.* (**1992a**) *Nucl. Instrum. Methods A*, **323**, 240–46.

Bouclier, R. *et al.* (**1992b**) *IEEE Trans. Nucl. Sci.*, **39**, 650–53.

Bouclier, R. *et al.* (**1993**) *Nucl. Instrum. Methods A*, **332**, 100–106.

Boyle, W.S. and Smith, G.E. (**1970**) *Bell. Syst. Tech. J.*, **49**, 587.

Braun-Munzinger, P. and Stachel, J. (**2007**) *Nature*, **448**, 302–9.

Brehin, S. *et al.* (**1975**) *Nucl. Instrum. Methods*, **123**, 225–29.

Brenner, R. *et al.* (**1992**) *Nucl. Instrum. Methods A*, **315**, 502–6.

Breskin, A. *et al.* (**1975**) *Nucl. Instrum. Methods*, **124**, 189–214.

Breskin, A. *et al.* (**1980**) *Nucl. Instrum. Methods*, **178**, 11–25.

Bressan, A. *et al.* (**1999**) *Nucl. Instrum. Methods A*, **425**, 262.

Brinkmann, D. *et al.* (**1995**) *Nucl. Instrum. Methods A*, **354**, 419–36.

Büttner, C. *et al.* (**1998**) *Nucl. Instrum. Methods A*, **409**, 79.

Campbell, M. *et al.* (**1990**) *Nucl. Instrum. Methods A*, **290**, 149–57.

Cardarelli, R.R. and Santonico, R. (**1981**) *Nucl. Instrum. Methods*, **1887**, 377–80.

Cardarelli, R.R. *et al.* (**1988**) *Nucl. Instrum. Methods A*, **263**, 20–25.

Charpak, G. and Sauli, F. (**1971**) *Nucl. Instrum. Methods*, **96**, 363–67.

Charpak, G. and Sauli, F. (**1984**) *Nucl. Instrum. Methods*, **225**, 627–35.

Charpak, G. *et al.* (**1979**) *Nucl. Instrum. Methods*, **162**, 405–28.

Chatrchyan, S. *et al.* (The CMS Collaboration) (**2008**) *J. Instrum.*, **3**, S08004.

Clark, A.R. *et al.* (**1976**) Proposal for a PEP Facility Based on the TPC. *SLAC Report No.* PEP-4, SLAC.

CMS (**1992**) The Compact Muon Solenoid. Letter of intent, *CERN Report No.* CERN/LHCC 92-3, CERN.

Colas, P. *et al.* (**2004**) *Nucl. Instrum. Methods A*, **535**, 226–30.

Collins, P. (**2007**) *Nucl. Instrum. Methods A*, **581**, 38–47.

Comiso, J. *et al.* (**1975**) *Phys. Rev. Lett.*, **35**, 13–15.

Damerell, C.J.S. (**1985**) *Proceedings of Physics in Collision*, Vol. **4**, Editions, Frontieres, Gif-Sur-Yvette, France, p. 453.

Damerell, C.J.S. (**1988**) *Rev. Sci. Instrum.*, **69**, 1549–1573.

Dawson, I. (**2008**) *J. Phys.: Conf. Ser.*, **110**, 092008.

Deptuch, G. *et al.* (**2007**) *SOI Pixel Design at FNAL, Counting Pixel for Imaging*. 1st SOI Detector R&D Workshop, March 6, 2007, KEK.

Deveaux, M. *et al.* (**2007**) *Nucl. Instrum. Methods A*, **583**, 134–38.

Dulinski, W. *et al.* (**2003**) *CMOS Monolithic Active Pixel Sensors for High Resolution Particle Tracking and Ionizing Radiation Imaging*. Proceedings of the Frontiers Detectors for Frontier Physics, Elba.

Eckert, S. *et al.* (**2007**) *Nucl. Instrum. Methods A*, **581**, 322–25.

Faschingbauer, U. *et al.* (**1996**) *Nucl. Instrum. Methods A*, **377**, 362.

Florent, J.J. *et al.* (**1993**) *Nucl. Instrum. Methods A*, **329**, 125–132.

Frautschi, M. *et al.* (**1991**) *Nucl. Instrum. Methods A*, **307**, 52–62.

Fretwurst, E. *et al.* (**1990**) *Nucl. Phys. A*, **288**, 1–12.

Froidevaux, D. and Sphicas, P. (**2006**) *Annu. Rev. Nucl. Part. Sci.*, **56**, 375–440.

Gatti I. and Rehak, P. (**1984**) *Nucl. Instrum. Methods A*, **225**, 608.

Gatti I. and Rehak, P. (**2005**) *Nucl. Instrum. Methods A*, **541**, 47–60.

Gatti, I. *et al.* (**1984**) *Nucl. Instrum. Methods A*, **226**, 129.

Giomataris, I. *et al.* (**1996**) *Nucl. Instrum. Methods A*, **376**, 29.

Giomataris, I. *et al.* (**2006**) *Nucl. Instrum. Methods A*, **560**, 405–8.

Gornushkin, Y. *et al.* (**2002**) *Nucl. Instrum. Methods A*, **478**, 311–15.

Groom, D.E. *et al.* (**1989**) *Nucl. Instrum. Methods A*, **279**, 1–6.

Heijne, E. *et al.* (**1980**) *Nucl. Phys. A*, **178**, 331–43.

Heijne, E. *et al.* (**1990**) Development of Silicon Pixel Detectors for the LHC. Vol. 3, *CERN Report No.* 90-10, *ECFA Report No.* 90-133, CERN–European Committee on Future Accelerators Proceedings, pp. 237–40.

Henderson, C. (**1970**) *Cloud and Bubble Chambers*, Methuen, London.

Holl, P. *et al.* (**1996**) *Nucl. Instrum. Methods A*, **377**, 367.

ILC (**2007**) *International Linear Collider Reference Design Report*. ILC-REPORT-2007-001, *http://www.linearcollider.org* (**2008**).

Kappler S. *et al.* (**2004**) *IEEE Trans. Nucl. Sci.*, **51**, 1524–28.

Kemmer, J. (**1980**) *Nucl. Instrum. Methods A*, **42**, 499–502.

Kemmer, J. and Lutz, G. (**1987**) *Nucl. Instrum. Methods A*, **253**, 356–77.

Kemmer, J. *et al.* (**1990**) *Nucl. Instrum. Methods A*, **288**, 92–98.

Kenney, C. *et al.* (**1999**) *IEEE Trans. Nucl. Sci.*, **46**, 1224.

Kozlov, S.F. *et al.* (**1975**) *IEEE Trans. Nucl. Sci.*, **22**, 160–70.

Ledermann, B. *et al.* (**2006**) *IEEE Trans. Nucl. Sci.*, **53**, 2936–43.

Lutz, G. (**1999**) *Semiconductor Radiation Detectors*, Springer, Berlin, Heidelberg, New York.

Oed, A. (**1988**) *Nucl. Instrum. Methods A*, **263**, 351–59.

O'Shaughnessy, K. *et al.* (**1990**) Conceptual Design for a Silicon Tracker at the SSC, CERN–European Committee on Future Accelerators Proceedings. Vol. 3, *CERN Report No.* 90-10, *ECFA Report No.* 90-133, CERN, pp. 234–36.

Palladino, V. and Sadoulet, B. (**1975**) *Nucl. Instrum. Methods*, **128**, 323.

Pandey, S. *et al.* (The STAR Collaboration) (**1999**) *Nucl. Phys. A*, **661**, 686c–689c.

Pandey, S. *et al.* (The STAR Collaboration) (**2002**) *Nucl. Instrum. Methods A*, **477**, 88–92.

Parker S. *et al.* (**1997**) *Nucl. Instrum. Methods A*, **395**, 328.

Pellegrini, G. *et al.* (**2002**) *Nucl. Instrum. Methods A*, **487**, 19.

Piemonte, C. *et al.* (**2005**) *Nucl. Instrum. Methods A*, **546**, 99.

Ragusa, F. and Rolandi, L. (**2007**) *New J. Phys.*, **9**, 336.

RD-20 (**1992**) Status Report. *Report No.* CERN/DRDC/92-28.

Rossi, L., Fischer, P., Rohe, T., and Wermes, N. (**2006**) *Pixel Detectors*, Springer, Berlin, Heidelberg, New York.

Sadrozinski, H.F.-W. (**1989**) Tracking with Silicon Detectors at the SSC. Vol. 1, CERN–European Committee on Future Accelerators Proceedings, *CERN Report No.* 89-10, *ECFA Report No.* 89-124, pp. 297–300.

Sadrozinski, H.F.-W. and Seiden, A. (**1989**) Resolution of a Silicon Tracking Device for the SSC. *Physics Report No.* SCIPP 89/50, Santa Cruz Institute for Particle Physics.

Sauli, F. (**1977**) Principles of Operation of Multiwire Proportional and Drift Chambers. *CERN Report No.* 77-09, CERN.

Sauli, F. (**1997**) *Nucl. Instrum. Methods A*, **386**, 531.

Sauli, F. (**2004**) *Nucl. Instrum. Methods A*, **522**, 93–98.

Sauli, F. and Sharma, A. (**1999**) *Annu. Rev. Nucl. Part. Sci.*, **49**, 341–88.

Schätzel, S. (**2008**) *J. Phys.: Conf. Ser.*, **110**, 092027.

Schmidt, B. (**1980**) Elektronendrift in Zählgasen unter dem Einfluß elektrischer und magnetischer Felder, *Diplomarbeit*, Physikalisches Institut der Universität Heidelberg.

Sciolla, G. (**1998**) *Nucl. Instrum. Methods A*, **419**, 310–14.

Smith, K.M. (**1990**) Gallium Arsenide Detectors. Vol. 3, *CERN Report No.* 90-10, *ECFA Report No.* 90-133. CERN–European Committee on Future Accelerators Proceedings, pp. 244–47.

Stapnes, S. (**2007**) *Nature*, **448**, 290–96.

Stefanov, K. (**2005**) *Nucl. Instrum. Methods A*, **549**, 93.

Stefanov, K. (**2006a**) *Nucl. Instrum. Methods A*, **565**, 157–64.

Stefanov, K. (**2006b**) *Nucl. Instrum. Methods A*, **569**, 48–52.

Stevenson, G.R. (**1992**) *On the Prediction of Radiation Levels in LHC Experiments*, paper presented at the 3rd International Conference on Advanced Technology and Particle Physics, Como, Italy. CERN Report No. CERN/TIS/RP/92-09/CF.

Thomson, M. (**2008**) *J. Phys.: Conf. Ser.*, **110**, 092032.

Titov, M. (**2007**) *Nucl. Instrum. Methods A*, **581**, 25–37.

Trischuk, W. (**2008**) Proceedings of the 34th International Conference on High Energy Physics, Philadelphia, PA.

Turchetta, R. *et al.* (**2001**) *Nucl. Instrum. Methods A*, **458**, 677–89.

Turchetta, R. *et al.* (**2006**) *Nucl. Instrum. Methods A*, **560**, 139.

Ulrici, J. *et al.* (**2001**) *Nucl. Instrum. Methods A*, **465**, 247–52.

Umemoto, M. *et al.* (**2004**) High-performance vertical interconnection for high density 3D chip stocking package. *Proceedings of the IEEE Electronic Components Technology Conference*, Vol. 1 p. 616–623.

Uno, H. (**1996**) *Nucl. Instrum. Methods A*, **379**, 421–423.

Visschers, J.L. *et al.* (**2007**) *Nucl. Instrum. Methods A*, **572**, 203–4.

Walenta, A.H. (**1979**) *IEEE Trans. Nucl. Sci.*, **26**, 73–80.

Walenta, A.H. (**1983**) *Nucl. Instrum. Methods*, **217**, 65–76.

Wermes, A. (**2004**) *IEEE Trans. Nucl. Sci.*, **51**, 1006–15.

Wunstorf, R. *et al.* (**1989**) Radiation Damage in Silicon Detectors. *CERN Report No.* 89-10, *ECFA Report No.* 89-124, Vol. 2, CERN–European Committee on Future Accelerators Proceedings, pp. 321–23.

Yarema, R. *et al.*(**2007**) *PoS* (Vertex 2007) OA.

12
Particle Impact Phenomena

Norman H. Tolk, Michael Albert, Justin Gregory, Glennys Mensing,
Andrew Steigerwald, and Travis Wade

12.1 Introduction 377
12.2 Scattering from a Coulomb Potential 379
12.3 Electron Impact 380
12.3.1 Interaction Mechanisms 380
12.3.1.1 Collisions with Electrons in the Solid 380
12.3.1.2 Scattering from the Lattice 382
12.3.1.3 Bremsstrahlung 382
12.3.2 Phenomena 383
12.3.2.1 Plasmons 383
12.3.2.2 Radiative Transitions 383
12.3.2.3 Auger Processes 384
12.3.2.4 Desorption Induced by Electronic Transitions (DIET) 385
12.3.2.5 Secondary Electron Emission 385
12.3.3 Applications 385
12.3.3.1 Scattered-Beam Electrons 385
12.3.3.2 Electrons from the Target 386
12.3.3.3 Fluorescence Measurements 386
12.4 Ion Impact 386
12.4.1 Interaction Mechanisms 386
12.4.1.1 Elastic Collisions with Target Atoms 386
12.4.1.2 Collisions with Electrons in the Solid 387
12.4.2 Phenomena 388
12.4.2.1 Depth Dependence of Backscattering 388
12.4.2.2 Sputtering 389
12.4.2.3 Channeling 389
12.4.2.4 Damage and amorphization 390
12.4.2.5 Electronic Excitations 390
12.4.3 Applications 390
12.4.3.1 Measuring Beam Particles 390
12.4.3.2 Recoil Target Atoms 390

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

12.4.3.3 Fluorescence Measurements 391

12.5 Photon Impact 391

12.5.1 Interaction Mechanisms 391

12.5.1.1 Electronic Absorption 391

12.5.1.2 Nuclear Scattering 391

12.5.1.3 Coupling to Vibrational and Rotational Modes 391

12.5.2 Phenomena 392

12.5.2.1 Photoelectric Effect 392

12.5.2.2 Compton Scattering 392

12.5.2.3 Raman Scattering 392

12.5.2.4 Two-Phonon Absorption 393

12.5.2.5 Resonant Bond Breaking 393

12.5.2.6 Laser Ablation 393

12.5.2.7 Exciton Creation 393

12.5.2.8 Pockels and Kerr Effects 393

12.5.2.9 Coherent Phonon Generation 393

12.5.2.10 Sum Frequency Generation 393

12.5.3 Applications 394

12.5.3.1 Scattered or Absorbed Incident Light 394

12.5.3.2 Electrons from the Target 394

12.5.3.3 Measuring Scattered Light 395

12.5.3.4 Mode Locking Lasers 395

12.6 Neutron Impact 395

12.6.1 Interaction Mechanisms 395

12.6.1.1 Scattering from Target Nuclei 395

12.6.1.2 Elastic Scattering from a Lattice 395

12.6.1.3 Neutron Activation 396

12.6.2 Neutron-Bombardment Phenomena 396

12.6.2.1 Phonon Emission and Absorption 396

12.6.2.2 Decay-Product Energy Loss 397

12.6.3 Applications 397

12.6.3.1 Elastic Scattering 397

12.6.3.2 Inelastic Scattering 397

12.6.3.3 Reaction Products 397

12.6.3.4 Spallation 397

12.7 Conclusion 398

Glossary 398

References 400

Further Reading 401

12.1
Introduction

In the past century, there has been a revolution in our understanding of the microscopic nature of matter, and much of this new insight was gained by observing the impact of beams of particles and photons on surfaces. Rutherford made his historic discovery of the nucleus by studying the scattering of a beam of α particles as they passed through thin foils. Davisson and Germer confirmed the existence of de Broglie matter waves by observing the diffraction of a beam of electrons by crystalline targets. In fact, many of the most important experiments of the twentieth century relied on impact phenomena.

While high-energy particle beams continue to be the vanguard of research into the fundamental nature of matter, the techniques that played major roles in the development of modern physics have now become indispensable tools for materials analysis in nuclear, atomic, medical, and condensed-matter physics, as well as physical chemistry and materials science (Kittel, 2004). Much of the new technology that defines the modern world resulted from progress in these fields, and, in turn, this progress depends upon a basic understanding of impact phenomena. In fact, impact phenomena play a prominent role throughout this Encyclopedia. This discussion, however, focuses on the rich variety of phenomena associated with beam interactions with condensed matter.

Analysis techniques that employ impact phenomena have in turn relied heavily upon progress made in related fields (Feldman and Mayer, 1986). Useful particle beams can only be produced in vacuum, and surfaces are difficult to characterize in the presence of air at normal temperatures and pressures. Therefore, vacuum technology has been particularly important to the development of impact techniques. Also, the emerging laser technologies and synchrotron-radiation sources are revolutionizing the way materials are characterized. Of course, progress in particle accelerator designs in both the high- and low-energy regimes has been crucial. Finally, the development of a tremendous variety of detectors and measurement devices has greatly expanded the range and sensitivity of experiments.

The nature of the interaction between a particle or photon beam and a target is highly dependent on the wavelength of the incident species. For a particle beam, the de Broglie relation gives the wavelength

$$\lambda = \frac{h}{\sqrt{2mE}} \tag{1}$$

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

where h is Planck's constant, m is the mass of the particle, and E is the energy (Shankar, 1994). The wavelength of a photon is just

$$\lambda = \frac{hc}{E} \quad (2)$$

where c is the speed of light. If the wavelength is much larger than the interatomic distance, then the incident particle wave function interacts with several target atoms at the same time. Consequently, in the infrared, visible, and ultraviolet atomic resolution is not possible. For photons, near-field techniques can circumvent this restriction to some extent. If the wavelength is the same order of magnitude as the interatomic distance in the target material, as it is in the X-ray region of the electromagnetic spectrum, then interference effects occur, allowing diffraction experiments, which can reveal periodic atomic structure. If the wavelength is much smaller, collisions can be characterized as simple two-particle interactions, and true atomic resolution is possible. Table 12.1 shows, for several beam species, the energy required to obtain wavelengths of 1 Å, which is about the size of an atom, and 5 Å, which is typical of interatomic distances in crystals.

As the table indicates, with photons only X-rays have short enough wavelength for diffraction experiments. Electrons, on the other hand, have a very convenient mass so that diffraction and atomic resolution are possible with modest energies easily achieved in the laboratory. In contrast, nucleons have such large mass that interference effects are difficult to produce. Nevertheless, neutrons at thermal energies can be collimated into beams and used for diffraction experiments. For example, the energy for a 1-Å neutron given in the table would correspond to a temperature of 650 °C. This illustrates that each species has inherent difficulties and advantages.

Another central issue is the degree to which the probe itself modifies the target. Almost all experiments damage the target on a microscopic level, and high-energy or high-flux beams can do macroscopic damage. With the recent developments in constructing and characterizing nanostructures, minimizing even microscopic damage from analysis techniques has become crucial. Unfortunately, a useful description of damage requires a very thorough description of the extremely complicated dynamics and is beyond the scope of this chapter.

Section 12.2 describes the basic Coulomb interaction, which is the basis for most simple models of particle impact and is therefore applicable to several of the subsequent sections. The remaining sections are organized by the identity of the impacting beam. Section 12.3 treats electron bombardment. Section 12.4 deals with ion–solid interactions. Section 12.5 discusses photon interactions, and Section 12.6 presents neutron interactions. Although the chapter is organized in this manner, many of the concepts

Tab. 12.1 The energy (in electron volts) required to achieve wavelengths of 1 and 5 Å for various species.

Wavelength (Å)	Photon	Electron	Nucleon	Helium	Argon
1	12 400	150	0.0819	~0.02	~0.002
5	2 480	6.02	0.00328	~0.0008	~0.00008

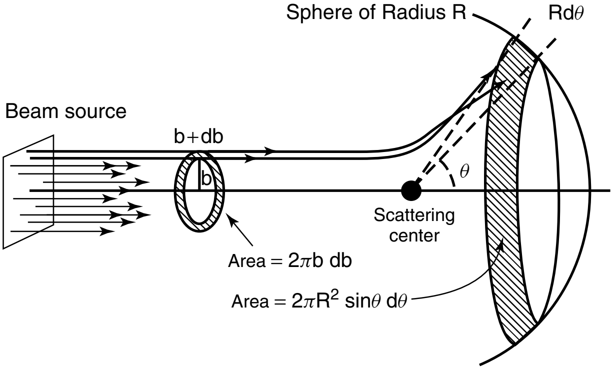

Fig. 12.1 The relationship between differential cross section and impact parameter.

of impact are applicable to more than one beam species. Within each section, the basic mechanisms of interaction are described, the observed phenomena are indicated, and finally, the most common applications are outlined.

12.2
Scattering from a Coulomb Potential

Many of the impact phenomena treated in this chapter begin with two-particle Coulomb scattering (Fetter and Walecka, 1980). Since this model is central to the discussion, a basic description is included. Charged particles traveling through a target material will feel a Coulomb repulsion or attraction to the electrons and nuclei in the material. Because materials are typically macroscopically neutral, it is assumed that long-range interactions are screened, and only individual classical two-particle collisions are significant. Because of the limited validity of this assumption, some of the later treatments will include correction factors that will not be derived.

Low-mass particles scattering from heavy target particles will transfer very little energy. Large scattering angles are possible, and scattering behavior can be determined from the relation of impact parameter to scattering angle. The impact parameter b is defined as the perpendicular distance from the target particle to the initial line of flight. Figure 12.1 indicates the relationship between impact parameter and scattering angle. The differential cross section $d\sigma$ can be defined by

$$d\sigma = 2\pi b\, db \tag{3}$$

From this model, it is seen that the differential cross section per unit solid angle is

$$\frac{d\sigma}{d\Omega} = \frac{b}{\sin\theta}\left|\frac{db}{d\theta}\right| \tag{4}$$

The integration of this over all solid angles yields the total cross section. The scattering problem is now reduced to determining $db/d\theta$, which can be inferred from the equations of motion in a Coulomb potential. In the center-of-mass frame, a particle with mass m and charge ze scattering from a particle of mass M and charge Ze will satisfy the equation

$$\mu\ddot{\mathbf{r}} = zZe^2\frac{\mathbf{r}}{r^3} \tag{5}$$

where μ is the reduced mass, $\mu = mM/(M+m)$, and $\mathbf{r}$ is the vector from the scattering center to the incident particle. The resulting orbits lead to the following relation between b and θ:

$$b = \frac{|zZ|e^2}{\mu v^2} \cot \frac{\theta}{2} \tag{6}$$

where v is the velocity of the incident particle. Note that the sign of the charges is indistinguishable. With reference to Figure 12.1, an attractive potential would draw the particle down rather than push it up, but the scattering angle would be the same. Using this relation, the differential cross section may be written as

$$\frac{d\sigma}{d\Omega} = \left(\frac{zZe^2}{4E \sin^2(\theta/2)} \right)^2 \tag{7}$$

where E is the energy of the incident particle in the center-of-mass frame. If the scattering center is much more massive than the incident particle, the center-of-mass energy is approximately equal to the laboratory energy. Thus, this derivation is extremely useful for describing the scattering of light particles from heavy particles.

On the other hand, a heavy particle scattering from a light particle will change direction only slightly. In this case, the interesting parameter is the amount of energy transferred. Assume that the heavy particle does not deflect at all and that the interaction is sudden so that the target particle cannot move very much. The net impulse in the direction of travel is zero as a result of the symmetry of the incoming and outgoing paths. However, integrating the Coulomb force, the momentum transfer in the perpendicular direction $\Delta; p$ is

$$\Delta p = \frac{2zZe^2}{bv} \tag{8}$$

This then gives the transferred energy directly:

$$\Delta E = \frac{\Delta p^2}{2m} = \frac{mz^2 Z^2 e^4}{Mb^2 E} \tag{9}$$

where the velocity of the incident particle has been written in terms of its kinetic energy:

$$E = \frac{1}{2} mv^2 \Longrightarrow v^2 = \frac{2E}{m} \tag{10}$$

Equation (9) may be inverted to find the impact parameter in terms of the transferred energy. Using this for the impact parameter in Eq. (3), the differential cross section may be written in terms of the energy transfer as

$$d\sigma = \frac{\pi mz^2 Z^2 e^4}{ME\Delta E^2} \, d\Delta E \tag{11}$$

12.3
Electron Impact

12.3.1
Interaction Mechanisms

12.3.1.1 Collisions with Electrons in the Solid

The interaction of a beam of electrons impinging on a solid or liquid target can be modeled to a good approximation by classical two-particle scattering from target electrons. Although the masses of the incident particle and target particle are identical, the impulse approximation applies if the energy of the incoming electron is sufficiently high. Here, the change in momentum Δp and thus the change in energy $\Delta; E$ as a function of impact parameter b reduce to

$$\Delta p = \frac{2e^2}{bv} \Longrightarrow \Delta E = \frac{e^4}{Eb^2} \tag{12}$$

where e is the electron charge, v is the velocity of the incident electron, and E is the energy of the incident electron. Thus the differential cross section $d\sigma$ is

$$d\sigma(E) = -2\pi b\, db = \left(\frac{\pi e^4}{E\Delta E^2}\right) d\Delta E \quad (13)$$

Integrating this gives the cross section for a particular transition with energy E_t:

$$\sigma = \pi \frac{e^4}{E}\left(\frac{1}{E_t} - \frac{1}{E}\right) \quad (14)$$

If the beam energy is much greater than the transition energy, the cross section simplifies to

$$\sigma \cong \frac{\pi e^4}{EE_t} \quad (15)$$

The possible electronic transitions are determined by the electronic structure of the target. The transition corresponding to E_t could be from a bound state, valence-band state, or conduction-band state to any higher energy state including free-particle states, which, if the surface is nearby, may result in the escape of electrons. For example, a valence-band electron could be promoted to the conduction band resulting in an electron–hole pair, or a bound core electron could be ejected from the solid.

Assuming that one transition dominates the absorption, the rate of energy loss as a function of distance traveled in the target may also be estimated by integrating energy times the differential cross section. This yields

$$-\frac{dE}{dx} = \frac{2\pi e^4 n}{E} \ln\left(\frac{E}{E_t}\right) \quad (16)$$

where n is the density of electrons in the initial state. Since energy can only be transferred in discrete amounts, the mean free path can be estimated from the rate of energy loss. For example, the dominant interaction in a metal is plasmon excitation, and so the mean free path λ can be estimated from the plasmon energy $\hbar\omega$:

$$\frac{1}{\lambda} = -\frac{m\omega e^2}{2\hbar E} \ln \frac{4E}{\hbar\omega} \quad (17)$$

where m is the mass of the electron. For a 1-keV electron incident on a typical metal, this initial penetration depth is a few nanometers. This formula applies equally well to the escape depth of ejected electrons in the solid.

Note that the cross section and thus the mean free path depend on the energy of the incident particle. Since the particle loses energy after each collision, λ must be integrated to find the average range of travel of the incident electron in the target. This yields an equation for the path length (range) R of the form

$$R = \left(\frac{K}{\rho}\right) E^{\gamma} \quad (18)$$

where ρ is the density of the material and K and γ are constants. Since this derivation is approximate, the form of the electron range as shown in Eq. (18) is assumed, but the parameters K and γ are fitted empirically. The parameter K is dependent on the material's density and atomic weight, but for R in microns, ρ in grams per cubic centimeter, and E in electron volts the value for K is approximately 0.064 for all materials. The parameter γ depends on the target material and is typically between 1.2 and 1.7. While this estimates the length of the electron's trajectory, each interaction results in a random scattering angle so that the trajectory is generally not straight.

12.3.1.2 Scattering from the Lattice

Because of the wave nature of the incident electron, it may also scatter from the atoms in the target without exciting an electronic state. Quantum mechanically, the wave function interacts with many atoms at the same time (Wichmann, 1971). The scattered wave function interacts with itself, resulting in quantum interference effects, and scattering is enhanced in those directions for which the interference is constructive. Since the mass of the target is so much greater than the mass of the electron, the momentum transfer is negligible, and the electron will lose almost no energy. Electrons are well suited to diffraction studies of crystals because diffraction requires that the wavelength of the incident particle must be near the size of the structure to be investigated. At 150 eV, an electron will have a wavelength of about 11 Å. Electron backscatter diffraction (EBSD) has been used effectively for texture, stored energy, and grain characterization on small scales down to ∼200 nm (Humphreys, 2004).

If the atoms in the target are regularly spaced in a lattice, the interference will also have a regular pattern. This can easily be demonstrated by considering Bragg reflection from adjacent planes of atoms (Figure 12.2). A beam reflecting from parallel planes separated by a distance d will have a path-length difference of $2d\sin\theta$, where θ is the angle of incidence. If these paths are such that constructive interference occurs, the path-length difference must equal an integer number of wavelengths:

$$n\lambda = 2d\sin\theta\,\text{(Bragg condition)} \qquad (19)$$

In a three-dimensional lattice, there are many ways to choose parallel planes of atoms, and each of these facets imposes another Bragg condition. Constructive interference will be greatest in scattering directions that best match all of these Bragg conditions, so that scattering in those directions is highly probable. On a detector screen, the set of favorable scattering directions forms a distinct pattern of points referred to as a *diffraction pattern*. The Bragg condition along with some simple geometry indicates that the distance between adjacent points in the diffraction pattern is inversely proportional to the distance between planes in the lattice.

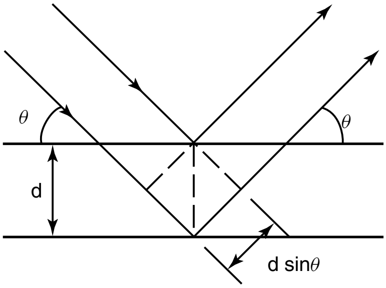

Fig. 12.2 Bragg reflection.

12.3.1.3 Bremsstrahlung

In scattering from the target atoms, the incident electron is decelerated, and so it must emit radiation. While this is true for any charged particle, it is particularly important for electrons whose light mass readily allows them to experience large decelerations. A continuous spectrum of radiation will be produced in which the intensity decreases monotonically as a function of increasing energy up to the energy of the incident electron (Jackson, 1999). The radiation spectrum can be calculated from classical electromagnetic theory. For nuclear scattering of nonrelativistic electrons, the radiation spectrum is characterized by the Bethe–Heitler theory (Bethe and Heitler, 1934)

$$I = \frac{16Z^2e^6}{3m^2v^2c^3} \times \ln\left(\frac{2-(\hbar\omega/E)+2\sqrt{1-(\hbar\omega/E)}}{\hbar\omega/E}\right) \tag{20}$$

where I is the spectral intensity and Z is the atomic number of the target species. Note that the effect is greatest for high atomic number targets. The ratio of bremsstrahlung energy E_B to other losses is approximately

$$\frac{dE_B}{dE_{\text{other}}} \approx \frac{4Z}{3\pi 137}\left(\frac{v}{c}\right)^2 \tag{21}$$

For an energy of 100 keV and a target with $Z \approx 30$, roughly 4% of the beam power goes into bremsstrahlung.

12.3.2
Phenomena

12.3.2.1 Plasmons

A plasmon is a density wave in the quasi-free gas of conduction electrons. It is important to note that electrons interact and produce correlated behavior in insulators as well as in metals; however, the following discussion applies only to metals. In metals, plasmon excitation will dominate the electron interactions. Consider a volume of negative charge embedded in a stationary positive charge of similar volume. Displacing the negative charge creates a dipole field that in turn produces a linear restoring force. The plasmon is thus a harmonic oscillation with frequency

$$\omega = \sqrt{\frac{4\pi ne^2}{m}} \tag{22}$$

where n is the density of electrons. The energy of these oscillations is $\hbar\omega$. For typical metals, this falls in the range of 10–15 eV. Because of the surface boundary conditions, there is a separate surface-plasmon mode that has a characteristic frequency of $\omega/\sqrt{2}$. For very small particles, boundary conditions complicate the plasmon model, and the plasmon frequency will depend on the size and shape of the particle. Plasmons have been proposed for use as data interconnects in nanoscale transistors (Ozbay, 2006).

12.3.2.2 Radiative Transitions

Electrons in the solid excited by impact may relax by emitting a photon. Also, if a core target electron is ejected, other electrons in the solid from either the same atom or neighbors may fall into the vacancy, emitting a photon. For radiative transitions, the energy of the photon will equal the difference in energy of the initial and final states of the radiating electron.

The easiest of these transitions to understand conceptually and interpret experimentally is a transition of a bound electron to fill a vacancy in the same atom. From perturbation theory, assuming only dipole interactions, the rate R for a spontaneous transition from a filled state to an empty state is found to be

$$R = \frac{4e^2}{3\hbar^4c^3}\Delta E^3|\langle f|\mathbf{r}|i\rangle|^2 \tag{23}$$

where ΔE is the transition energy and the labels i and f refer to the initial and final electron states, respectively.

The rate depends heavily on the transition energy, but many transitions are disallowed by the dipole matrix element. Thus, the transition energy dependence is most apparent when comparing similar transitions in atoms with differing atomic number. The dipole matrix element between the initial and final state indicates that transitions between highly overlapping states are preferred. Since the change in angular momentum of the absorber (or emitter) must equal the angular momentum of the photon, the matrix element automatically imposes dipole selection rules. The allowed dipole transitions have $\Delta l = \pm 1$ and $\Delta j = \pm 1$ or 0, where l and j are the angular-momentum quantum numbers. However, other multipole transitions may occur with different restrictions.

12.3.2.3 Auger Processes

An electron making a transition to a core vacancy may also release the excess energy by exciting another electron. This is called an *Auger process*, and the resulting excited electron is called an *Auger electron* (Figure 12.3). The electron that falls into the vacancy can most easily transfer energy to an electron in the same subshell. One falls into the vacancy, and the other carries the excess energy to a higher state, possibly the vacuum state, leaving two new vacancies in the higher shell or band. The transition rate also depends on the overlap of the initial and final states of the electron that fills the vacancy. Unlike radiative transitions, a vacancy may be filled by an electron from a higher orbital substate of the same primary shell as the vacancy. Since the initial and final states for such an intrashell transition are very similar, these transitions have higher rates than might be expected from the previous argument. These intrashell transitions are called *Coster–Kronig transitions.*

To first order, the energy of an escaping Auger electron can be estimated from the binding energies of the initial and final states. If a vacancy in state a is filled by an electron originally in state b, resulting in an electron in state c being ejected, the energy of the liberated electron is

$$E = E_a - E_b - E_c \tag{24}$$

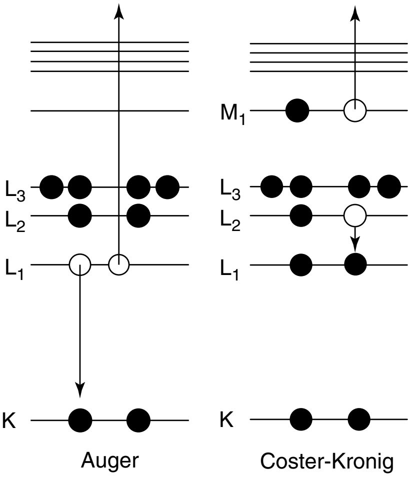

Fig. 12.3 Auger processes.

where these are the normal binding energies of the states. However, there is a significant correction due to the vacancies, which alter the normal binding energies. A fairly good estimate can be constructed from the binding energies of similar states in an atom with one more proton:

$$E = E_a^Z - E_b^Z - E_c^Z - \frac{1}{2}\left(E_c^{Z+1} - E_c^Z + E_b^{Z+1} - E_b^Z\right) \quad (25)$$

12.3.2.4 Desorption Induced by Electronic Transitions (DIET)

When a *localized* excited state is created in a solid, it dramatically changes the local electronic environment. If the state is sufficiently long lived, the electromagnetic forces can alter the position of an atom or molecule in the lattice, resulting in defect formation or even desorption (Tolk *et al.*, 1983). Consider an atom adsorbed on a surface. An electron in some bonding orbital holds the atom to the surface. If that electron is excited to a weaker-bonding or even an antibonding state, the bond may break, resulting in desorption. Somewhat more spectacular is the highly repulsive state that results when the electron of one atom makes a transition into the core vacancy of a neighboring atom. Several other mechanisms exist by which the electronic energy can be channeled into the kinetic energy of the atoms or molecules in the lattice, resulting in damage.

12.3.2.5 Secondary Electron Emission

In addition to directly ejected electrons or Auger electrons, a target will also emit a broad band of electrons with energies less than 100 eV that are relatively insensitive to the energy of the incident beam. These low-energy electrons are called *secondary electrons* because they clearly do not result from a single-step process or even a two-step process like Auger electrons. Although secondary electrons arise from several sources, an argument based on plasmons provides a clue to their origin. One or more plasmons may couple to a single electron, thus giving that electron enough energy to escape from the surface. Since plasmons have fixed energies, the shape of the energy spectrum of emitted electrons would not depend on how the plasmons were created. This and other similar processes result in an energy spectrum of which only the amplitude depends on the incident energy.

12.3.3
Applications

12.3.3.1 Scattered-Beam Electrons

Since the probability for electron scattering depends on the density of electrons in the material, the scattering of electrons for a fixed target thickness will vary with different materials. For example, this effect provides the contrast for transmission electron microscopy (TEM).

The incident electron can only lose energy in discrete amounts determined by the allowed transitions in the target material. Thus, one can learn a great deal about the electronic structure of the target by measuring the energy of beam electrons that pass through the target. This is the basis of electron energy-loss spectroscopy (EELS). If the same measurement is made for the reflected portion of a low-energy beam, the measurement is surface sensitive. This is the principle behind high-resolution electron-energy-loss spectroscopy (HREELS).

Since elastically scattered electrons may interfere, the structure of the lattice can be determined from the spatial distribution of scattered electrons. This is the basis of electron diffraction. For example, the transmitted-beam electrons in a transmission electron microscope demonstrate a diffraction pattern. Low-energy electrons reflected at normal incidence from a crystalline surface will also show a diffraction pattern, which is primarily surface sensitive. This is called *low-energy electron diffraction* (*LEED*). Because the component of energy into the surface is small, a high-energy electron beam reflected at grazing incidence will also produce a surface-sensitive diffraction pattern. This technique, called *reflection high-energy electron diffraction* (*RHEED*), provides a more convenient geometry than LEED in that the surface is more accessible for simultaneous experiments.

12.3.3.2 **Electrons from the Target**

In addition to scattered-beam electrons, Auger electrons may also escape from the target. Measuring the energy of electrons reveals peaks at energies corresponding to the Auger transitions, which are characteristic of the target atoms. Thus Auger-electron spectroscopy (AES) can detect target species down to the 0.1% level. Elaborate techniques are usually required to separate the Auger peaks from the background of scattered-beam electrons.

12.3.3.3 **Fluorescence Measurements**

To avoid the background problem of Auger-electron spectroscopy, the characteristic photons produced by similar radiative transitions may be detected. Since the size of the target volume probed by the electron beam can be reduced to about a micron, this technique is referred to as *electron microprobe analysis* (*EMA*). Characteristic photons are typically in the hard ultraviolet or X-ray region and thus have their own detection challenges.

12.4
Ion Impact

12.4.1
Interaction Mechanisms

12.4.1.1 **Elastic Collisions with Target Atoms**

Particles such as ions or molecules striking a surface will interact primarily by elastic collisions with target atoms. By conservation of energy and momentum, it is easily shown that the fraction of initial energy retained by the incident particle is

$$\frac{E_1}{E_0} = \left[\frac{(M_2^2 - M_1^2 \sin^2\theta)^{1/2} + M_1 \cos\theta}{M_1 + M_2} \right]^2 \qquad (26)$$

where M_1 is the mass of the incident particle, M_2 is the mass of the target particle, and θ is the scattering angle. This ratio is referred to as the *kinematic factor* and is often denoted by K. The kinematic factor as a function of angle for several choices of the ratio M_2/M_1 is plotted in Figure 12.4 The fraction of energy transferred to the target particle is similarly given by

$$\frac{E_2}{E_0} = \frac{4M_1M_2}{(M_1 + M_2)^2} \cos^2\phi \qquad (27)$$

where ϕ is the recoil angle. For particular scattering angles, these relations become quite simple. For example, in the case of backscattering, $\theta = \pi$ and $\phi = 0$, so

$$\frac{E_1}{E_0} = \left(\frac{M_2 - M_1}{M_1} \right)^2 \quad \text{and}$$

$$\frac{E_2}{E_0} = \frac{4M_1M_2}{(M_1 + M_2)^2} \qquad (28)$$

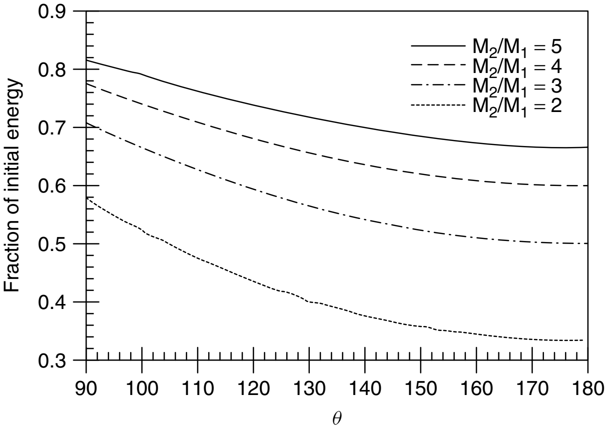

Fig. 12.4 The kinematic factor as a function of scattering angle for several choices of the ratio M_2/M_1.

Electron screening of target nuclei complicates heavy particle scattering. Nevertheless, if the energy of the incident particle is high enough to penetrate well within the electron cloud, the simple scattering model will be approximately correct. This condition will be met if

$$E_0 > \frac{zZe^2}{a_0} \tag{29}$$

where z is the atomic number of the incident atom, Z is the atomic number of the target species, and a_0 is the Bohr radius. In any case, screening will produce some deviation from the simple model, and, in general, a correction factor must be included. Ignoring electron screening, the cross section is as given above:

$$\sigma(\theta) = \left(\frac{zZe^2}{4E\sin^2(\theta/2)}\right)^2 \tag{30}$$

12.4.1.2 **Collisions with Electrons in the Solid**

An incident heavy particle will also collide with electrons in the solid. However, as a result of momentum conservation, a heavy particle can transfer only a small fraction of its momentum to an electron. Thus, as a high-energy particle traverses a target, it will slowly lose energy through numerous electron collisions in an almost continuous fashion. These collisions may also be modeled with the impulse approximation. The derivation of the rate of energy loss is completely analogous with that for electron–electron scattering. Thus the rate of energy loss as the particle traverses the target is

$$-\frac{dE}{dx} = \frac{2\pi z^2 e^4 n}{E}\frac{M}{m_e}\ln\frac{4Em_e}{IM} \tag{31}$$

where M is the mass of the incident particle, z is its atomic number, I is the average excitation energy of electrons in the solid, and n is the density of electrons in the solid. I is complicated, but, for heavier elements, it is approximately equal to $10Z$. This is usually converted to a ‘‘stopping power’’ per atom (Figure 12.5). The stopping power ε per atom is defined as

$$\varepsilon = \frac{dE/dx}{N} \tag{32}$$

where N is the density of the target atoms. The stopping power can then be used to determine dE/dx for compounds.

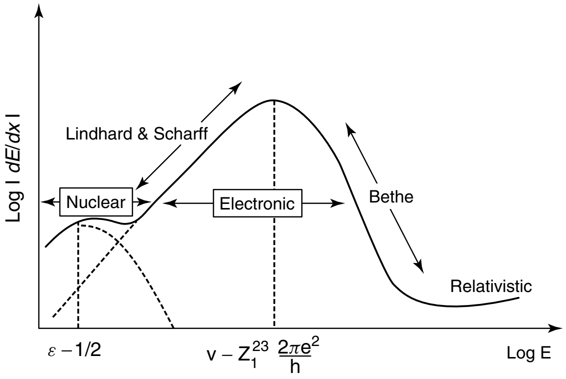

Fig. 12.5 Stopping power.

According to Bragg's rule, for a compound A_nB_m, the stopping power is

$$\varepsilon_{A_nB_m} = n\varepsilon_A + m\varepsilon_B \tag{33}$$

where n and m are the relative atomic abundances of the respective constituents. Thus, multiplying by the molecular density, the rate of energy loss in a compound target may be determined.

Although electronic stopping is the most important contribution to the energy loss for a large range of energies, other factors such as scattering from the atomic cores are also important. The convoluted path resulting from numerous scattering events further complicates this picture. To compute the total depth beneath the surface attained by an incident particle, all of these factors must be considered. Table 12.2 indicates the median penetration depth for various species and energies of particles incident on amorphous alumina.

12.4.2
Phenomena

12.4.2.1 Depth Dependence of Backscattering

Since a heavy particle will lose energy to electrons as it penetrates a target, the electron scattering produces depth dependence in the energy of atomic scattering. For example, consider a light ion with energy E scattering from an atom at a depth d within the target. To find the total energy lost to electron scattering, dE/dx

Tab. 12.2 Penetration depth in μg/cm² for various incident species and energies, derived from implantation of amorphous alumina.[a]

Energy (eV)	Na	K	Kr	Xe
10	4.5	2.6	2.5	2.3
100	35	20	14	11
1000	500	260	170	100

[a] (Jespergard and Davies, 1967).

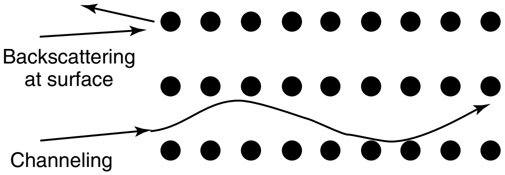

Fig. 12.6 Channeling.

should be integrated to the penetration depth. However, for modest depths, the fractional energy loss will be small, and the total energy loss is approximately

$$\Delta E \cong \frac{dE}{dx} d \tag{34}$$

where dE/dx is evaluated at the incident energy E . The particle with this reduced energy now backscatters from a target atom, and the energy is reduced by the kinematic factor above. The particle then traverses the target again before escaping. Since the particle may have been scattered at an angle to the surface normal, the distance traveled on the way out is $d/\cos\theta$. The energy loss is approximated as above with dE/dx now evaluated at the appropriate backscattered energy. Thus, the total energy lost to the electrons both in entering and exiting the sample is

$$\Delta E = d\left(K\left.\frac{dE}{dx}\right|_{\text{in}} + \frac{1}{|\cos\,\theta|}\left.\frac{dE}{dx}\right|_{\text{out}}\right) \tag{35}$$

where K is the kinematic factor and ''in'' and ''out'' indicate at which energy dE/dx should be evaluated. Note that the total energy lost to electron scattering is directly proportional to the penetration depth.

12.4.2.2 **Sputtering**

Atomic collisions impart significant energy to the target atoms, and, in general, they will in turn collide with other target atoms. The resulting collision cascade can give some atoms sufficient momentum to escape from the target, and so the ion beam erodes the target. This process is called *sputtering* (Behrisch, 1981). The sputtering yield is defined as the number of target atoms emitted per incident particle. This depends not only on the nature of the target material but also on the beam species and energy, and the possible values vary tremendously. Although calculation of sputtering yields is understandably daunting, theory has produced good results. Nevertheless, for practical applications, empirical values are used.

12.4.2.3 **Channeling**

If the target is crystalline, an interesting phenomenon known as *channeling* may occur (Figure 12.6). If the incident particles are aligned along one of the highly oriented directions of the crystal, they cannot collide solidly with atoms in the bulk of the target. Essentially, the atoms on the surface shadow the rest of the atoms. However, the effect is greater than that simple explanation indicates. Consider the regular lattice to be parallel planes of atoms. If the incident particle is moving nearly tangential to these planes, it will merely glance off the planes at a shallow angle. This requires very little energy, and the incident particle will move down this channel, making a series of shallow collisions with the ''walls''. Thus, a particle incident in a channeling direction may penetrate very deeply into the target. The ultimate depth achieved is primarily determined by electron stopping. Furthermore, backscattering will only occur at the surface.

12.4.2.4 Damage and amorphization

For sufficiently energetic ion beams, energy loss can occur through defect creation. Depending on the temperature during ion bombardment, both point and complex defects may be created. Consequences include deteriorated electronic and optical properties, a problem especially relevant in high-energy particle detectors (such as at LHC or GANIL). For increased dosage, enough damage may occur such that the amorphous state becomes energetically favorable. However, this threshold varies with relative target-ion mass difference, temperature, and dosage. Finally, many materials exhibit interesting nonlinear behavior when irradiated with swift (0.1–1 GeV) ions. Such effects include latent tracks and phase changes to metastable lattice configurations.

12.4.2.5 Electronic Excitations

In colliding with the electrons, a heavy particle excites electronic states in the target just as in the case of electron bombardment. Therefore, many of the same phenomena are seen in each case. Electrons will be ejected, electron–hole pairs will be created, plasmons will be stimulated, secondary electrons will be produced, and most of the other previously discussed electronic excitations will occur. Of particular interest are the radiative transitions to fill a core hole created by the incident particle. The origins of the core hole are irrelevant, and the process is exactly as described in the Section 12.3.

12.4.3
Applications

12.4.3.1 Measuring Beam Particles

The energy of a light ion that has been scattered is uniquely determined by the species from which it scattered. If the energy of scattered-beam particles is measured, the composition of the target can be determined. This is the basis of Rutherford backscattering spectrometry (RBS). For simplicity, the measurement is usually made as close to the backscattering direction as possible. Since the energy has a simple dependence on depth due to electron scattering, depth profiling is also possible. Finally, since subsurface atoms cannot readily backscatter when the beam is aligned along a channeling direction, channeling can enhance the surface sensitivity of RBS. For example, RBS can easily detect whether a surface is reconstructed, relaxed, or covered with an adsorbate.

To enhance surface sensitivity, an ion beam with energy in the range of one to a few kiloelectron volts may be used in conjunction with an electrostatic energy analyzer. Since particles that penetrate the surface are almost always neutralized, the electrostatic energy analyzer will detect only particles scattered from the surface. The low-energy beam ensures that most of the ions will be scattered from the surface. This is the basis of low-energy ion scattering (LEIS).

12.4.3.2 Recoil Target Atoms

Backscattering is inefficient for lighter target materials and kinematically impossible if the target material is lighter than the beam species. Rather than measuring backscattered beam particles, measuring the energy of forward recoiling target atoms gives the same kinematic information. This is the basis of forward recoil spectrometry (FRS). The beam is usually incident at a shallow angle to the target surface, and the recoiling atoms are detected at some small angle from the surface in the forward direction.

12.4.3.3 Fluorescence Measurements

Heavy particles may also produce core-hole vacancies in target atoms. The mechanisms for filling the vacancy are exactly the same as described above. In particular, the target may be characterized by measuring the radiation from such transitions. This is the basis of proton- (or particle)-induced X-ray emission (PIXE). Because the larger mass of the incident particle results in far less bremsstrahlung, PIXE has less background noise than electron microprobe analysis.

12.5
Photon Impact

12.5.1
Interaction Mechanisms

12.5.1.1 Electronic Absorption

A photon incident on a target atom or molecule may excite higher electronic states in the target. If the photon is modeled as an oscillating perturbation $H'e^{-i\omega t}$, from first-order perturbation theory, the probability of stimulating a transition is given by Fermi's "golden rule,"

$$P_{i\to f} = t\frac{2\pi}{\hbar}|\langle\langle f|H'|i\rangle|^2\delta(E_f - E_i - \hbar\omega) \tag{36}$$

where t is the length of time that the perturbation is applied, and the labels i and f indicate initial and final states of the *unperturbed* system. The terms E_i and E_f are the energies of the unperturbed initial and final states. Note that the transition energy must be equal to the energy of the photon. This seems to imply that the whole photon must be absorbed, and scattering will not occur. However, the excited state may relax by emitting another photon with either equal or lower energy, resulting in an effective scattering. Also, according to the uncertainty principle, noneigenstates can exist for a short time, so photons may be scattered even when no resonant transition exists. Nevertheless, the probability of interaction is much greater at resonance, that is, when there exists a transition with a difference in energy equal to the energy of the photon. Therefore, light will be strongly absorbed at frequencies that correspond to the allowed electronic transitions. The matrix element indicates the strength of coupling of the photon to the electronic states. In general, the coupling is proportional to the dipole moment. However, as in the case of atomic hydrogen, time dependence allows some coupling where no dipole moment exists. Finally, because a photon has angular momentum, only transitions with $\Delta l = 1$ are allowed with high probability in the dipole approximation.

12.5.1.2 Nuclear Scattering

An incident photon may also scatter from the target atom without leaving the atom in an excited state. The large mass of the nuclei allows momentum to be conserved without significantly changing the energy of the photon. Since the energy does not change, the photon maintains its initial wavelength, and interference may occur. As with electrons, the resulting diffraction pattern is indicative of the long-range order of the target. The diffraction relations are exactly the same as discussed previously for particle interference (see Section 12.3.1.2).

12.5.1.3 Coupling to Vibrational and Rotational Modes

Molecular bonds may have many modes of vibration and rotation, each with one or more natural frequencies. If a given bond has electric dipole characteristics, the electric field of a photon with a frequency resonant to the bond may excite these bonds to higher vibrational and rotational states. A photon is absorbed in the process.

12.5.2
Phenomena

12.5.2.1 Photoelectric Effect

If the incident photon has greater energy than the binding energy of electrons in the target, electrons may be liberated. The maximum energy of the ejected electron will be

$$E = E_b - \hbar\omega \tag{37}$$

where E_b is the binding energy of the electron. Since the final state is a free particle, selection rules can always be satisfied, and all electrons with binding energy less than or equal to the photon energy may be liberated. Applying Fermi's "golden rule" to hydrogenic states yields the cross section for a photon of energy $\hbar\omega$ liberating an electron with binding energy $E_b < \hbar\omega$:

$$\sigma = \frac{128\pi e^2 \hbar E_b^{5/2}}{3mc(\hbar\omega)^{7/2}} = \frac{7.45\text{eV}\text{Å}^2}{\hbar\omega}\left(\frac{E_b}{\hbar\omega}\right)^{5/2} \tag{38}$$

12.5.2.2 Compton Scattering

For photon-electron scattering, if the energy transferred to the electron is much greater than its binding energy, the scattering will be nearly identical to scattering from a free electron. The resulting scattering behavior can be deduced simply from energy and momentum conservation. If a photon initially with energy $\hbar\omega$ scatters at an angle θ from an electron initially at rest, conservation of energy requires

$$\begin{aligned} E_p + E_e = E_p' + E_e' \Longrightarrow \hbar\omega + mc^2 \\ = \hbar\omega' + E' \end{aligned} \tag{39}$$

where the subscripts refer to photon and electron, and unprimed and primed stand for before and after the collision, respectively. The term E' is the subsequent relativistic energy of the electron. Momentum conservation requires

$$\mathbf{p}_p = \mathbf{p}_p' + \mathbf{p}_e' \Longrightarrow (\mathbf{p}_p - \mathbf{p}_p') = \mathbf{p}_e'^2 \tag{40}$$

which may be written as

$$\frac{\hbar^2\omega^2}{c^2} + \frac{\hbar^2\omega'^2}{c^2} - \frac{2\hbar^2\omega\omega'}{c^2}\cos\theta = p_e'^2 \tag{41}$$

Combining Eqs. (39) and (41) yields

$$\omega' = \frac{\omega}{1 + (\hbar\omega/mc^2)(1 - \cos\theta)} \quad \text{or}$$
$$\lambda' = \lambda + \frac{2\pi\hbar}{mc}(1 - \cos\theta) \tag{42}$$

Note that the wavelength of the scattered photon depends on the scattering angle. Also, the wavelength shift is independent of the initial wavelength. The shift only becomes significant for fairly short initial wavelengths, and hard (high-energy) X-rays must be used to see the effect.

12.5.2.3 Raman Scattering

For a photon to couple to one of the collective nonlocalized vibrational modes of a lattice, momentum must be conserved. A photon traversing a material can emit or absorb an optical phonon, but in order to conserve momentum and energy, the frequency and direction of the photon must change. If a phonon is created, the interaction is called a *Stokes process*. If a phonon is absorbed, it is an anti-Stokes process (Figure 12.7). Conservation of energy requires $\omega = \omega' \pm \Omega$ where ω is the frequency of the incident photon, ω' is the frequency of the scattered photon, and ω: is the frequency of the phonon that is created or absorbed by the photon. To conserve momentum, $\mathbf{k} = \mathbf{k}' \pm \mathbf{K}$, where $\mathbf{k}$, $\mathbf{k}'$, **and** $\mathbf{K}$ are the wave vectors for the incident photon, scattered photon, and phonon, respectively.

12.5.2.4 **Two-Phonon Absorption**

Similarly, a photon cannot be completely absorbed into a single phonon because a phonon with the same energy will not have the same wave vector. In order to conserve momentum, the photon must create two phonons (Figure 12.8).

12.5.2.5 **Resonant Bond Breaking**

Desorption induced by electronic transitions (DIET) (see Section 12.3.2.4) is possible with photons as well (Tolk *et al.*, 1983). If the incident photon has the appropriate resonant energy to excite an electron initially in a bonding state to an antibonding state, the bond may be broken.

12.5.2.6 **Laser Ablation**

An incident photon may liberate electrons and desorb ionized target atoms. With a high-intensity laser beam, electrons and ions may be created in sufficient quantity to make a dense plasma near the surface of the target. Once the plasma exists, the laser may couple directly to the plasma, and the superheated plasma then ablates the target by collisions with plasma particles.

12.5.2.7 **Exciton Creation**

If photons are incident on a semiconducting material, and the photon energy is slightly below the energy of the material's band gap, an electron may be excited from the material's valence band to an energy level immediately below the band gap, where it will form, with the resulting hole, a bound pair state very similar to a hydrogen atom. This bound state is called an *exciton* (Kittel, 2004).

12.5.2.8 **Pockels and Kerr Effects**

Under intense photon irradiation, materials may experience a change in their refractive indices, although the effect is generally very weak. There are some materials, however, where the effect is very strong. If λ is the wavelength of light to be refracted, K is a measure of the strength of the Kerr effect in the material (the Kerr constant), and E is the electric field due to the photon irradiation, the change in refractive index is given as

$$\Delta n = \lambda K |E|^2 \tag{43}$$

12.5.2.9 **Coherent Phonon Generation**

Massive excitation of electron energy levels through photon bombardment may result in coherent multiple phonon generation as the excited electrons rapidly decay to the conduction band, shedding phonons in the process. This coherent acoustic phonon (CAP) wave may be used to probe materials and activate interesting events (Xu *et al.*, 2008).

12.5.2.10 **Sum Frequency Generation**

In certain nonlinear media under intense photon irradiation, photons with frequencies higher than the incident photon frequencies may be produced in the material itself. If the material is under irradiation of two or more frequencies, the resulting photon frequency may be any of the sums of the incoming frequencies. For instance, if the incoming frequencies are ω_1 and ω_2, the resulting frequency could be

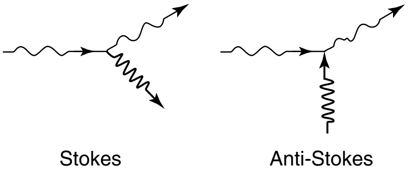

Fig. 12.7 Raman processes.

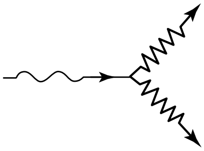

Fig. 12.8 Two-phonon absorption.

$\omega_1 + \omega_2$. This is sum frequency generation. If the material is under monochromatic irradiation with frequency ω, the resulting radiation may have a frequency of 2ω or even 3ω;. These special cases are called *second-* and *third-harmonic generation*, respectively (Shen, 2002).

12.5.3
Applications

12.5.3.1 **Scattered or Absorbed Incident Light**

Photons with energies resonant with electronic transitions in the target will be strongly absorbed. Thus, if the transmission of light as a function of frequency is measured, the electronic structure of the target may be indirectly mapped. This is the basis of absorption spectroscopy. Because the resonances correspond to the difference in energy of two electronic states, rather than the absolute energy of a single electronic state, the information is somewhat convoluted. However, absorption spectra for most materials are well known, and so information about the target can be obtained without a detailed knowledge of the electronic structure. If the target is not transparent, the same information can be inferred from reflection.

12.5.3.2 **Electrons from the Target**

In order to map the electronic states directly, the energy of photoelectrons can be measured. Since the energy of a photoelectron is just the energy of the photon minus the electron-binding energy, the spectrum of photoelectron energies is a direct measure of the electronic states of the target. This is the basis of X-ray photoelectron spectroscopy (XPS). However, this will only indicate filled states. In order to map the normally empty excited states, the excited states can be populated prior to measurement in a pump–probe experiment. For example, the density of states of the conduction band in a semiconductor can be mapped in this manner.

In semiconductors, some of the electronic structure of the target can be directly mapped without reverting to X-rays as in XPS. If a target is biased with a small voltage and the current flow is measured, the current will be enhanced when photons promote electrons to the conduction band. If this photocurrent is measured as a function of photon energy, the derivative of the resulting spectrum is a direct map of the density of states below the band gap. This is the basis of internal photoemission (IPE) spectroscopy. The simplicity of the measurement and the lower photon energy make this method preferable to XPS.

The energy spectrum of photoelectrons will also have some additional structure. This is due to a characteristic energy loss as an electron is liberated from an inner shell, coupled with an interference effect as the photoelectron interacts with target atoms as it escapes from the target. This is the basis of extended X-ray absorption fine-structure (EXAFS) measurements. The advantage of EXAFS is the extremely localized sensitivity due to the short range of electrons in solids. Thus, the primary peak indicates the species of the scattering atom from the characteristic energy loss, and the fine structure indicates the lattice positions of its neighbors.

12.5.3.3 Measuring Scattered Light

Since the phonon-dispersion relation is a unique function of the target material, Raman scattering will also be unique to the target material. The shift in energy of a scattered photon is a direct measure of the energy of the optical phonon with wave vector nearly equal to zero. This is the basis of Raman spectroscopy. If the incident beam is focused to a small spot, the identity of crystallites in a heterogeneous target may be spatially mapped as well.

Exactly as in the electron case, the diffraction pattern of X-rays scattered from a crystalline target is a measure of the long-range order of the target. This is the basis of X-ray diffraction.

12.5.3.4 Mode Locking Lasers

The optical Kerr effect may be used to create self-focusing ultrafast lasers. If a material with a strong Kerr constant (a Kerr lens) is employed in a laser cavity, the stronger standing wave modes in the laser cavity are more tightly focused, and the weaker modes may be therefore filtered out by geometric means, with the result that the strongest modes become stronger and the weaker modes disappear. This is the phenomenon of mode locking.

12.6
Neutron Impact

12.6.1
Interaction Mechanisms

12.6.1.1 Scattering from Target Nuclei

Unlike ions or electrons, Coulomb scattering does not apply to neutrons because of their charge neutrality. A complete discussion of the neutron-scattering process is not possible here, but a few points should be made. A neutron can only scatter from the atomic nuclei in the target, which means neutron scattering is insensitive to the electronic structure of the target material. Since the cross section for neutron scattering is relatively small, neutrons usually penetrate well into the target, and neutron bombardment consequently is not surface sensitive. In addition, the scattering amplitude depends on the relative orientations of the neutron spin and nuclear spin so that a polarized neutron beam is sensitive to magnetization. In many cases, neutron beams may have very low energies of less than 0.1 eV per particle. With other beam particles, interactions could be treated as simple two-particle collisions. For a low-energy neutron beam, this is not the case. In fact, interactions with phonons dominate.

12.6.1.2 Elastic Scattering from a Lattice

A low-energy neutron may also scatter from the nuclei in the target as a whole. As in the case of electrons, this elastic scattering results from the wave nature of the neutron, and as before, quantum interference effects result. For constructive interference, the Bragg condition must be satisfied, and specific scattering angles are preferred, resulting in a diffraction pattern. Note that the Bragg condition depends on the de Broglie wavelength, which in turn depends upon the energy. Thus, while electrons are much preferred for diffraction studies, neutron diffraction is still very important as a means of monochromatizing a neutron beam.

12.6.1.3 Neutron Activation

For many target atoms, energetic neutron bombardment may produce a nuclear reaction, resulting in an unstable isotope (*see* Chapter 10). The unstable isotope will then decay producing radiation. Since the target is now said to be ''radioactive'' or ''activated,'' this process is called *neutron activation*. If N_x is the number of atoms available to be activated in the target and N_γ is the number of activated atoms, then the growth of the number of activated nuclei is given by

$$\frac{dN_\gamma}{dt} = N_x\sigma\phi - N_\gamma\lambda_\gamma \tag{44}$$

where σ is the reaction cross section, ϕ is the flux of neutrons, and λ_γ is the decay constant of the unstable nuclei. The first term in this equation is the rate of activation, while the second term is the rate of decay of already activated atoms. Integrating this gives the number of activated atoms as a function of the time of bombardment t:

$$N_\gamma = \frac{N_x\sigma\phi}{\lambda_\gamma}(1 - e^{-\lambda_\gamma t}) \tag{45}$$

The activity A_t, which is the number of decays per second, is just the number of activated atoms times the decay constant:

$$A_t = N_x\sigma\phi(1 - e^{-\lambda_\gamma t}) \tag{46}$$

It is apparent that after a long bombardment time, the activity saturates. In fact, the activity reaches 50% of saturation in one half-life and 75% in two half-lives. This gives experimental control over other undesired processes. Processes with much longer half-lives than the desired reaction can be avoided by stopping activation after two half-lives. Processes with much shorter half-lives than the desired reaction may be eliminated by waiting to measure the activity until a full half-life after bombardment ceases.

12.6.2
Neutron-Bombardment Phenomena

12.6.2.1 Phonon Emission and Absorption

Excluding nuclear processes, a low-energy neutron will interact with target nuclei primarily by inducing normal-mode vibrations or phonons. In a given direction in a given lattice, only particular phonon energies and momenta are allowed. Since momentum must be conserved, neutrons will scatter most readily in those directions for which the change in momentum of the neutron corresponds to an allowed phonon in the lattice. In addition, during absorption or emission of a phonon, the lattice as a whole may give or take specific quantities of momentum as indicated by the reciprocal lattice vectors **G**. Thus, if **q** is the wave vector of an allowed phonon and **k** is the propagation vector of the neutron, then the change in the propagation vector of the neutron $\Delta\mathbf{k}$ satisfies the following condition:

$$\Delta \mathbf{k} = \pm \mathbf{q} + \mathbf{G} \quad (47)$$

The plus–minus sign indicates that the neutron may either absorb or emit a phonon. The above condition specifically applies to the bulk phonon structure since neutrons penetrate deeply into the target, and neutron bombardment is not sensitive to surface phonons.

12.6.2.2 Decay-Product Energy Loss

If the decay product of an activated nucleus is a charged particle, then it will lose energy as it escapes from the target through scattering from electrons in the target material. This is exactly the same as the electronic energy loss discussed in Section 12.4. Assuming that the fractional energy loss is small and that the target material is fairly uniform, the energy loss is found to be proportional to the distance traveled through the material. Decay-product energy loss differs from the Rutherford backscattering case in one very important way: the charged particle only has to travel through the target material once. Typically, these decay products have energies of a few mega electron volts and can escape from depths up to a few microns beneath the surface.

12.6.3
Applications

12.6.3.1 Elastic Scattering

Diffraction of neutrons by a crystal depends on the energy of the neutrons. In order to produce low-energy neutron beams with well-defined energies, a beam of thermal neutrons is scattered from a crystal and collimated. The diffraction angle determines the energy and can be varied by rotating the crystal and apertures. Neutron monochromators operate according to these principles.

12.6.3.2 Inelastic Scattering

Measuring the scattering angle and energy of neutrons scattered from a crystal maps the phonon structure of the lattice. The change in momentum of the neutron corresponds to the momentum of the phonon, and the change in energy of the neutron is the energy of the phonon. Such measurements are routinely carried out with a triple-axis neutron spectrometer.

12.6.3.3 Reaction Products

If the cross section for activation and the lifetime of the unstable isotope are known, then the concentration of a particular species in a target can be calculated directly from the measured activity. This is the basis of neutron activation analysis. Some care is necessary to reduce the background of other activation processes.

If the reaction product is a charged particle, then a depth profile of the activated species can be obtained by measuring the energy of the escaping particles. This is the basis of neutron activation depth profiling.

12.6.3.4 Spallation

In addition to fusion- and fission-based neutron sources, a process known as *spallation* may be utilized to produce high-energy pulsed neutron beams. Firing 1.0 GeV protons (as opposed to decay product neutrons, as in a chain reaction) into a high atomic mass target releases tens of neutrons per proton as the target is vaporized. This provides an efficient source of pulsed neutrons for research purposes. The Spallation Neutron Source (SNS) at the Oak Ridge National Laboratory recently came online and represents the most powerful neutron source available (Mason *et al.*, 2000).

12.7
Conclusion

Particle impact onto condensed matter results in a great variety of phenomena. In this chapter, the most important effects have been presented, but these are by no means all of the possible results of an impact event. More importantly, the reader should understand that any given impact experiment will produce many of the phenomena described above. Experiments are usually designed to emphasize one particular aspect of the impact event. However, while an electron energy-loss experiment may emphasize plasmon excitation, the experimenter must be aware that the electron beam may also be diffracting, undergoing collision cascades, ejecting core electrons, creating electron–hole pairs, or producing lattice defects. It is this great variety of effects that makes directed particle or photon beams powerful probes for measuring the properties of materials.

Glossary

AES (Auger-Electron Spectroscopy): An experimental technique for measuring the atomic species in the region near the surface of a solid by measuring the energies of Auger electrons.

Anti-Stokes Process: A Raman scattering process in which a photon is scattered by absorbing a phonon.

Auger Electron: An electron excited and possibly even liberated by the energy released when another electron falls into a vacant low-energy state.

Auger Process: When an electron falls into a lower energy vacancy, energy is liberated. If that energy is transferred to another electron, resulting in that electron being excited to a higher state or the vacuum state, the process is called an *Auger process*, and the excited electron is called an *Auger electron*.

Bragg Reflection: The reflection of light or particles from parallel planes that have spacing such that self-interference determines the allowed scattering angles.

Bremsstrahlung: German for ‘‘braking radiation.’’ Bremsstrahlung is the broad spectrum of electromagnetic radiation produced when charged particles are decelerated by impact with a target.

Channeling: Phenomenon in which particles incident on a crystal in a high-symmetry direction reach unusually great depths by traveling through the spaces between the atoms.

Coster–Kronig Transitions: A particular Auger transition in which the vacancy is filled by an electron from the same subshell.

Coulomb Scattering: A model for scattering of charged particles in which the Coulomb repulsion or attraction due to the electric charge is assumed to dominate.

Depth Profiling: A technique in Rutherford backscattering experiments that exploits the nearly uniform deceleration of a charged particle in a solid to map the distribution of atomic species as a function of depth into the target.

DIET (Desorption Induced by Electronic Transitions): The ejection of an atom or molecule from a surface by a repulsive Coulomb force resulting from an electronic excitation.

Diffraction Pattern: The interference pattern produced at a detector or phosphor screen by particles or light elastically scattered from a lattice.

Dipole Approximation: Approximation of the interaction of light with electronic states that only considers the dipole term in the expansion of the electronic state.

EELS (Electron Energy-Loss Spectroscopy): An experimental technique for determining the electronic structure of a material by measuring the change in energy of a beam of electrons as they pass through a material.

EMA (Electron Microprobe Analysis): An experimental technique in which a tightly focused electron beam produces vacancies in the target atoms, and the radiation produced when other electrons fill the vacancies is spectroscopically analyzed.

EXAFS (Extended X-Ray Absorption Fine Structure): An experimental technique for measuring the lattice structure in the neighborhood of a particular impurity by carefully measuring the energy of photoelectrons produced by X-rays.

Exciton: In a semiconducting material, an electron and hole that are bound in a hydrogen-like state.

Fermi's "Golden Rule": An approximation of the probability of a photon producing a particular electronic transition calculated by treating the photon as a short-lived oscillatory perturbation and considering only dipole interactions.

FRS (Forward Recoil Spectroscopy): An experimental technique for measuring light atomic species in a thin film by measuring the energy of particles ejected in the forward direction by a high-energy ion beam.

HREELS (High-Resolution Electron Energy-Loss Spectroscopy): An experimental technique for measuring the electronic structure of the surface of a solid by measuring the change in energy of a low-energy beam of electrons reflected from a surface.

Impact Parameter: The perpendicular distance from the target particle to the initial line of flight of the incident particle.

Impulse Approximation: An approximation based on the assumption that the interaction between a fast-moving beam particle and a stationary target particle happens so quickly that neither particle has time to change its initial motion much while the interaction is occurring.

IPE (Internal Photoemission): An experimental method for mapping the density of states of the valence band of a semiconductor by measuring the photocurrent produced by a tunable light source.

Kinematic Factor: In an elastic two-particle collision, the fraction of energy retained by the incident particle.

LEED (Low-Energy Electron Diffraction): An experimental technique for measuring the near-surface lattice structure of a crystal by diffracting a low-energy beam of electrons from the surface.

LEIS (Low-Energy Ion Spectroscopy): An experimental technique for measuring the atomic species near the surface of a solid by reflecting a low-energy beam of ions from the surface and measuring the ion energy electrostatically.

Mean Free Path: The average distance traveled between interactions.

Mode locking: A laser generation technique where weaker pulses are filtered out of the laser cavity allowing the stronger pulses to become even stronger.

Neutron Activation Analysis: An experimental technique for measuring the concentration of atomic species in a sample by neutron bombardment activation.

Phonon: A collective vibration of a lattice that carries momentum and energy and is therefore treated as a virtual particle.

PIXE (Proton- or Particle-Induced X-Ray Emission): An experimental technique for measuring the atomic species in a target, in which particle bombardment produces vacancies in the target atoms, and the radiation produced when other electrons fill the vacancies is spectroscopically analyzed.

Plasmon: A collective oscillation of the electrons of a material, which carries momentum and energy and is therefore treated as a virtual particle.

Quantum Interference: Interference effects that result from the quantum wave nature of particles.

Raman Scattering: Photon scattering from the lattice either producing or absorbing a phonon.

Reconstruction: The rearrangement of the lattice structure at the surface of a crystal into a symmetry that differs from the bulk solid.

RBS (Rutherford Backscattering): An experimental technique for determining the atomic species in a target by measuring the energy of reflected ions.

RHEED (Reflection High-Energy Electron Diffraction): A convenient technique for determining the surface structure of a sample by diffracting a grazing-incidence high-energy beam of electrons from the surface.

Scattering Angle: The angle between the initial line of flight continued through the scattering center and the subsequent line of flight.

Second Harmonic Generation: The process by which photons with a doubled frequency is produced from a material under irradiation by a single original frequency.

Spallation: The splitting of heavy atoms into smaller nucleons by a proton or ion beam. Provides an efficient source of pulsed neutrons.

Spectral Intensity: The radiative power per unit frequency.

Sputtering: The ejection of target atoms from a surface by collision cascade.

Sputtering Yield: The number of target atoms sputtered from the surface per incident particle.

Stokes Process: A Raman process in which a photon is scattered from the lattice by emitting a phonon.

Stopping Power: The distance rate of energy loss normalized to the number of atoms.

TEM (Transmission Electron Microscopy): The classic electron microscopy in which a high-energy beam of electrons is passed through a very thin sample and the beam is then imaged on a phosphor screen.

XPS (X-Ray Photoemission Spectroscopy): An experimental technique for determining the electronic density of states of a sample by measuring the energy of photoelectrons produced by X-rays.

References

Behrisch, R. (ed.) (**1981**) *Sputtering by Particle Bombardment I*, Springer-Verlag, New York.

Bethe, H. and Heitler, W. (**1934**) On the stopping of fast particles and on the creation of positive electrons. *Proc. R. Soc. London Ser. A*, **146**(856), 83–112.

Feldman, L. and Mayer, J. (**1986**) *Fundamentals of Surface and Thin Film Analysis*, North-Holland, New York.

Fetter, A. and Walecka, J. (**1980**) *Theoretical Mechanics of Particles and Continua*, McGraw-Hill, New York.

Humphreys, F.J. (**2004**) Reconstruction of grains and subgrains from electron backscatter diffraction maps. *J. Microsc.*, **213**(3), 247–56.

Jackson, J.D. (**1999**) *Classical Electrodynamics*, 3rd edn, Wiley-VCH Verlag GmbH.

Jespergard, P. and Davies, J.A. (**1967**) *Can. J. Phys.*, **45**, 2983–94.

Kittel, C. (**2004**) *Introduction to Solid State Physics*, 8th edn, Wiley-VCH Verlag GmbH, New York.

Mason, T.E. *et al.* (**2000**) The spallation neutron source: a powerful tool for materials research, *LINAC 2000* paper FR203.

Ozbay, E. (**2006**) Plasmonics: merging photonics and electronics at nanoscale dimensions. *Science*, **5758**, 189–93.

Shankar, R. (**1994**) *Principles of Quantum Mechanics*, 2nd edn, Springer, New York.

Shen, Y.R. (**2002**) *The Principles of Nonlinear Optics*, Wiley Interscience.

Tolk, N. *et al.* (**1983**), *Desorption Induced by Electronic Transitions*, Springer-Verlag, New York.

Wichmann, E. (**1971**) *Quantum Physics*, McGraw-Hill, New York.

Xu, Y. *et al.* (**2008**) Pump-probe studies of travelling coherent longitudinal acoustic phonon oscillations in GaAs. *Physica Status Solidi (C)*, **5**(8), 2632–36.

Further Reading

General

Prutton, M. (**1994**) *Introduction to Surface Physics*, Oxford Science Publications, Oxford.

Woodruff, D.P. and Delchar, T.A. (**1994**) *Modern Techniques of Surface Science*, 2nd edn, Cambridge Univ. Press, New York.

Electron Impact

Bundle, C.R. and Baker, A.D. (eds) (**1981**) *Electron Spectroscopy: Theory, Techniques and Applications*, Academic Press, New York.

Ibach, H. (ed.) (**1977**) *Topics in Current Physics*, Vol. **4**, Springer-Verlag, Berlin.

Ion Impact

Tesner, J.R. and Natasi, M. (eds) (**1995**) *Handbook of Modern Ion Beam Analysis*, Materials Research Society, Pittsburgh.

Tolk, N.H., Traum, M.M., Tully, J.C., and Madey, T.E. (eds) (**1983**) *Desorption Induced by Electronic Transitions, DIET I*, Springer-Verlag, New York.

Photon Impact

Cohen-Tannoudji, C., DupontRoc, J., and Grynberg, G. (**1992**) *Atom-Photon Interactions*, Wiley-VCH Verlag GmpH, New York.

Kienzle-Focacci, M.N. and Wadhwa, M. (eds) (**2001**) *PHOTON 2001: Proceedings of the International Conference on the Structure and Interactions of the Photon Including the 14th International Workshop on Photon-Photon Collisions*, World Scientific Publishing Co. Ltd, Singapore.

Neutron Impact

Izyumov, Y.A. and Chernoplekov, N.A. (**1992**) *Neutron Spectroscopy*, Consultants Bureau, New York.

Skold, K. and Price, D.L. (**1986**) *Neutron Scattering*, Part A, Vol. **23**, Academic Press, Boston.

6
Particle Physics

Dan Green

6.1 Introduction 171
6.2 The Standard Model – Synthesis and Overview 171
6.2.1 The SM Constituents 171
6.2.2 SM Forces 174
6.2.3 QCD Calculations and Other States of Matter 174
6.2.4 CP Violation 175
6.2.5 Higgs Boson 177
6.2.6 Neutrino Oscillations 178
6.2.7 Frontier Issues in Particle Physics 179
6.2.7.1 Unification 179
6.2.7.2 Dark Energy 179
6.2.7.3 Dark Matter 180
6.3 Accelerators and Other Facilities 180
6.3.1 Accelerator-Based Particle Physics 180
6.3.1.1 Fixed-target Experiments 180
6.3.1.2 e^+e^- Collider Experiments 182
6.3.1.3 ep Collider Experiments 182
6.3.1.4 $p\bar{p}$ Collider Experiments 182
6.3.1.5 Applications 182
6.3.2 Nonaccelerator Particle Physics 183
6.4 Mapping the SM Constituents onto the Detectors 183
6.4.1 Confinement and Jets – Quarks and Gluons 184
6.4.2 Missing Energy and Neutrinos 184
6.4.3 Charged Leptons and EM Interactions 184
6.4.4 W, Z, and τ Decays 184
6.4.5 Displaced Secondary Vertices 185
6.5 A Model $pp/\bar{p}p$ Collider Detector 185
6.5.1 “Soft” Backgrounds 185
6.5.2 Magnetic Analysis 187
6.5.3 Si Inner Tracker 187

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

6.5.4 Outer Tracker 187
6.5.5 Calorimetry 187
6.5.5.1 EM Segment 188
6.5.5.2 Hadronic Segment 188
6.5.6 Muon Detection 188
6.5.7 Trigger and DAQ 188
6.5.8 Analysis 189
6.6 Other Experimental Techniques 189
6.6.1 Ionization Detectors 189
6.6.2 Photomultiplier Tubes, Scintillator 189
6.6.3 Slow-*n* Detection 190
6.6.4 Cherenkov Detectors 190
6.6.5 Transition-radiation Detectors 190
6.7 Future Developments 190
6.7.1 The Energy Frontier at the LHC 190
6.7.1.1 Higgs Boson 190
6.7.1.2 Supersymmetry and DM 192
6.7.2 International Linear Collider (ILC) 192
6.7.3 Heavy Ions at the LHC 193
6.7.4 CP Violation After BaBar and Belle 193
6.7.5 The Future Neutrino Program 193
6.7.6 DE Experiments 194
6.7.7 The Far Future 194
Glossary 194
References 196
Further Reading 196

6.1
Introduction

The goal of particle physics is to discover the most elementary particles and to study their basic interactions. Over the past 30 years, remarkable strides have been made, culminating in what is called the *standard model* (SM) of particle physics. This model, which exhibits a deep simplicity and beauty, along with the classical theory of gravity (general relativity, GR), together explain all phenomena that have been observed to date with a compelling economy and power. The SM is valid at all distances down to $\sim 10^{-16}$ cm, or by uncertainty-principle arguments up to energies $\sim 200\,\text{GeV} = 2 \times 10^{11}$ eV, which corresponds to a time 10^{-12} s after the Universe was born in the ''Big Bang.'' The development of the SM should be ranked as one of the great scientific triumphs of twentieth century science.

The SM is both a stunning success and the goad to further progress. A brief description of the SM is given below, followed by a discussion of the present frontier issues in the field. The experimental program that has been mounted to confront those issues is then described with a subsequent explanation of the experimental tools that are presently in active use in particle physics. Finally, new facilities are mentioned as are the questions that they are designed to explore.

6.2
The Standard Model – Synthesis and Overview

6.2.1
The SM Constituents

All matter is assumed to be made up of pointlike elementary particles. The great complexity and diversity exhibited by the world are assumed to proceed from the interaction of only a few basic entities. At normal energy scales, the world can be described as composed of neutrons and protons, the electron, and the photon, which is the quantum of the familiar electromagnetic (EM) interaction between charged particles. The SM asserts that neutrons and protons are themselves the composite bound states of pointlike u and d (up and down) quarks. That belief is partly attributed to the fact that the properties of proton and neutron are very similar.

At higher energies (shorter distances), the strong interactions that are responsible for binding the nucleus together must be considered. An analogy is perhaps useful. In atomic physics, the EM interaction

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

between charged protons and electrons binds atoms together. A residual van der Waals force exists between neutral atoms, which, in turn binds atoms into molecules. Similarly, quarks strongly interact by means of a quantum number called *color* through a messenger field with quanta called *gluons*. The bound-state ''neutral'' colorless composites – for example, protons and neutrons – interact to form bound nuclei. The fundamental theory is called quantumchromodynamics (QCD), which is the theory of the interaction of colored pointlike quarks via massless gluons (see Chapter 3).

Gluons and quarks are believed to be permanently confined by the strong forces that they experience. However, their existence has been inferred from the observation of pointlike behavior in experiments scattering electrons off protons. These experiments are the direct analog of the famous Rutherford scattering experiment by which the nucleus within the atom was inferred by scattering of atoms (Friedman and Kendall, 1972). Even if atoms could not be ionized, and the nuclei and electrons were separately examined, the Rutherford scattering experiment would allow one to infer the existence of the nucleus. Quarks were also inferred by the observed regularities and multiplet structure exhibited by several hadrons, or strongly interacting colorless quark composites (Gell-Mann, 1964; Zweig, 1964).

The discovery of radioactive decay required the existence of a new force, the weak interaction. The weakness of this interaction can be seen by comparing the lifetime of a typical strongly interacting hadron resonance to that of a neutron, where the lifetime ratio is $\sim 10^{26}$. The decay products of weak decays were found to include neutral, approximately massless neutrinos. For example, the neutron is unstable, decaying into a proton, an electron, and a neutrino. One of the profound insights that led to the formulation of the SM was that the weak and EM interactions are unified (Glashow, 1961; Weinberg, 1967; Salam, 1968) despite their apparent great disparity in strength. The Fermi coupling constant describing the strength of weak interactions, G_F, is describable at a deeper level as due to a dimensionless coupling constant g_W, just as is the EM coupling constant e. The weak and EM interactions have approximately the same basic strength. The ratio is measured by the Weinberg angle, θ_W (Prescott *et al.*, 1978) and M_W is the mass of the weak force carrier, the W boson.

$$\begin{aligned} G_F &\sim \frac{g_W^2}{M_W^2} \\ e &= g_W \sin\theta_W \\ M_W &\sim e/\sqrt{G_F} \sim 90\,GeV \end{aligned} \tag{1}$$

Therefore, electromagnetism and the weak force are unified into the electroweak (EW) theory. This great triumph of the SM is akin to the prior unification of electricity and magnetism by Maxwell. It culminated in the discovery of the charged and neutral quanta of the weak force, the W and Z gauge bosons (Arnison *et al.*, 1983; Banner *et al.*, 1983). The EW unification is evident only at the very high energies probed by accelerators. At lower energies, the unification is masked by the large mass of the W and Z, and the weak and EM forces appear to have vastly different interaction strengths. It seems that Nature becomes simpler and more symmetric at higher energies. The EM and weak potentials differ only slightly in coupling strength, α and α_W, $\alpha = e^2/\hbar c$, $\alpha_W = g_W^2/\hbar c$, but profoundly in the range of the interaction:

$$\begin{aligned} V_{\mathrm{EM}} &= \frac{\alpha}{r} \\ V_W &= \left(\frac{\alpha_W}{r}\right) e^{-r/\bar{\lambda}_W} \\ \bar{\lambda}_W &= \frac{\hbar}{M_W} \end{aligned} \qquad (2)$$

At low energies, much less than M_W, the exponential factor causes V_W to be $< V_{\mathrm{EM}}$, while at high energies, $V_W \cong V_{\mathrm{EM}}$.

The constituents of the SM are given in Table 6.1. The quanta, or force carriers, of the two basic interactions, QCD and EW, are the gluons, the photon, and the massive W and Z weak bosons. There are three dimensionless coupling constants, describing the strength of the strong, EM, and weak forces. All three basic interactions are mediated by spin-1 quanta.

The antiparticle partners that exist for each quark and lepton are not shown in Table 6.1. For example, the positron is paired to the electron. The leptons have only EW interactions, while the quarks also possess strong QCD interactions. Both quarks and leptons are spin-1/2 fermions.

The first-generation quarks and leptons (e, ν_e, u, d) describe our world at low energies, as mentioned before. What has been discovered by experiment is that this basic pattern is repeated in "generations," whose members differ from those of the first largely only in mass. For example, the muon appears to interact exactly as if it were a heavy electron. Why the pattern recurs is unknown. How many generations are there? Recently, both cosmology (primordial helium abundance) and high-energy physics (Z decay width) have converged on the conclusion that there are three and only three generations (Burkhardt and Steinberger, 1991; Steigman, Schrammn, and Gunn, 1977). Thus, the "periodic table" of the SM, as given in Table 6.1, is complete except for the Higgs boson, which is discussed in Section 6.7. In fact, the discovery of the top quark, t, has very recently completed the entries in Table 6.1 (CDF Collaboration, 1995; D0 Collaboration, 1995). Indirect inferences of the top mass had previously been available (Burkhardt and Steinberger, 1991).

Tab. 6.1 The SM constituents.

	Three generations			**Charge[a] Q**	
Matter (spin 1/2)[b]	$\begin{pmatrix} e \\ \nu_e \end{pmatrix}$	$\begin{pmatrix} \mu \\ \nu_\mu \end{pmatrix}$	$\begin{pmatrix} \tau \\ \nu_\tau \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	Leptons
	$\begin{pmatrix} u \\ d \end{pmatrix}$	$\begin{pmatrix} c \\ s \end{pmatrix}$	$\begin{pmatrix} t \\ b \end{pmatrix}$	$\begin{pmatrix} 2/3 \\ -1/3 \end{pmatrix}$	Quarks
	Quanta	Force	Coupling	No. of quanta	
Interactions (spin 1)	g	QCD	g_s	8	Gluons
	γ		e	1	Photon
	W,Z	EW	g_W	2,1	Weak bosons
Mass (spin 0)	H^0	EW	Yukawa	1	Higgs boson

[a] Units are electron charge, e.
[b] Units are $\hbar$.

6.2.2
SM Forces

The unfolding of the SM has also resulted in a much deeper understanding of the basic forces. It is remarkable that all forces now thought to be fundamental appear to be derivable from a single principle. The similarity of all the fundamental forces holds out the hope that they can one day be unified into a grand unified theory (GUT). The gauge principle is one of the great organizing insights brought about by the development of the SM. What exactly is the gauge principle, expressed nonmathematically?

As a familiar example, consider classical relativity. If the global space–time coordinate transformations of special relativity are required to hold at arbitrary local points, then a new field, the gravitational field, is required, and the interaction of that field with matter is largely specified (Utiyama, 1956). Clearly, local gauge invariance is a powerful and compact principle.

In quantum mechanics, the overall phase of the wave function Ψ is irrelevant because only $|\Psi|^2$ is an observable. The requirement that physics be unchanged under an arbitrary local choice of overall phase implies the existence of the EM field A_μ and specifies the interaction of that field with the basic charged fermions (Weyl, 1929). The gauge replacement of the derivative is familiar from both classical Hamiltonian mechanics and introductory quantum mechanics and indicates that gauge invariance requires a new field A_μ:

$$P_\mu \longrightarrow P_\mu - eA_\mu, \quad \partial_\mu \longrightarrow \partial_\mu - ie A_\mu \tag{3}$$

Making this replacement in the free Lagrangian density specifies the interactions of the gauge field with the fermions.

This elegant and simple gauge principle then defines the theory of electromagnetism. Similar considerations allow one to write the gauge theories of the QCD and EW forces. In these cases, the quanta themselves couple together, in contrast to the photon, which is uncharged. This type of behavior is familiar from the prototypical gauge theory, GR. The gravitational field possesses energy, and because all energy has mass ($E = mc^2$), the field itself gravitates. Thus, GR is a nonlinear theory, as is QCD, for example, because the gluons themselves carry color. Nonlinear theories are difficult to treat mathematically but imply a very rich phenomenology – perhaps as rich as the world itself.

6.2.3
QCD Calculations and Other States of Matter

As formulated, QCD has no free parameters except the coupling constant. Thus, for example, one should be able to calculate the bound-state hadron mass spectrum from first principles. However, the fact that the quarks and gluons are permanently confined well explains the strength of the forces involved. Hence perturbation theory, a familiar tool used throughout all areas of physics, is not generally applicable. The strong force becomes strong at an energy scale of about 200 MeV.

At present, the mass spectrum and quantum numbers of the hadrons are found by experiment. Similarly, the distribution of quarks and gluons bound in the proton is determined by scattering experiments. Recently, using strong-coupling techniques adopted from solid-state physics, a cutoff approximation to the full QCD theory has been formulated. QCD is formulated on a mesh with finite spacing

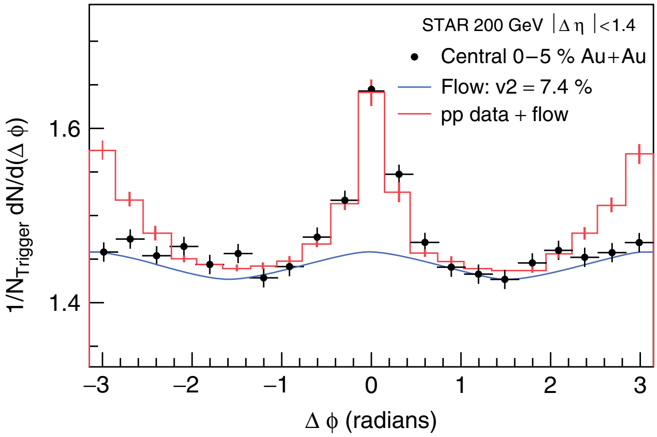

Fig. 6.1 The azimuthal distribution of particles in proton–proton and heavy ion, Au–Au, collisions. Note the depletion of particles "away" from the peak at 0° in the heavy-ion collisions compared to p–p collisions indicating strong absorption.

(limited momentum) in a "lattice gauge theory." The numerical problems are still formidable and massively parallel computing is needed to complete the calculations, but preliminary indications are heartening in that the spectrum of the lowest lying two- and three-quark bound states can now be accurately calculated from first principles.

Given that we possess the complete theory of QCD, it is appropriate to begin to explore its full nonlinear richness. For example, it is an open question whether there exist phase transitions in the quark–gluon system at high temperatures. For example, the propagation of quarks and gluons in high-temperature heavy-ion collisions has recently been studied at the Relativistic Heavy Ion Accelerator (RHIC) located at Brookhaven National Laboratory (BNL). As seen in Figure 6.1, there appears to be strong quark absorption in this new medium. More complete studies are planned in upgraded experiments to run in the near future.

6.2.4
CP Violation

The generations shown in Table 6.1 refer to quantum-number labels of states produced in strong interactions where quark–antiquark pairs are produced. As stated above, decays are due to the weak interactions, whose conservation laws are different from those obeyed by the strong interactions. Thus, the weak eigenstates are not necessarily the same as the strong-interaction eigenstates listed in Table 6.1.

In general, there is a unitary 3×3 (three generations) mixing matrix that connects the strong and weak eigenstates: the Cabibbo–Kobayashi–Maskawa (CKM) matrix. The SM does not give us any guidance about this matrix; its elements are determined purely experimentally (Cabibbo, 1963). For example,

$|V_{du}|^2$ and $|V_{su}|^2$ represent, respectively, the relative strengths of $d \to u$ and $s \to u$ weak transitions. The CKM matrix is almost diagonal. Because there are three generations, the matrix can be complex. If it is complex, then invariance under the operation of charge conjugation followed by parity inversion (CP) is violated in the SM because of the CKM matrix. CP violation has been known for 30 years (Fitch, 1981; Cronin, 1981), but a deeper explanation of the phenomenon still eludes us.

Recently, "B factories" have been built to explore the consistency of the CKM description as the complete explanation of CP violation. The current state of the art is summarized in Figure 6.2. The unitarity of the CKM matrix implies relations between complex matrix elements, in particular, there is a closed triangular constraint. At the limit of experimental accuracy, the "unitary triangle" is indeed closed, which implies that the CKM model of CP violation fully explains the present data. Moreover, CP nonconservation has profound cosmological implications. It is required for the presently observed preponderance of matter over antimatter in the universe and the matter-to-entropy ratio to appear elegantly (Sakharov, 1967). However, the CKM strength of CP violation is insufficient to explain the matter dominance of our Universe, leaving it unexplained in the SM.

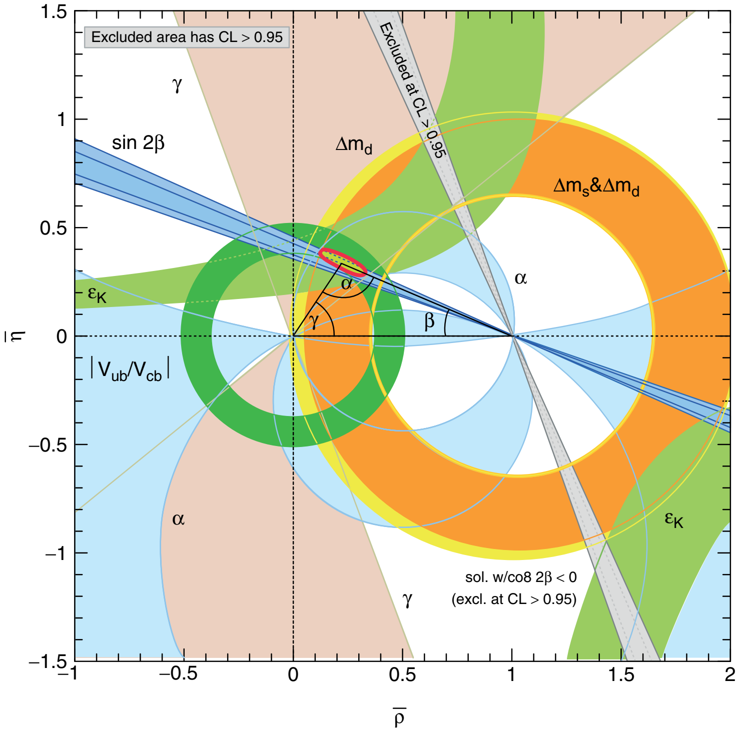

Fig. 6.2 The present constraints on the CKM matrix elements placed by experiment. Unitarity of the matrix requires a triangular relationship among the matrix elements.

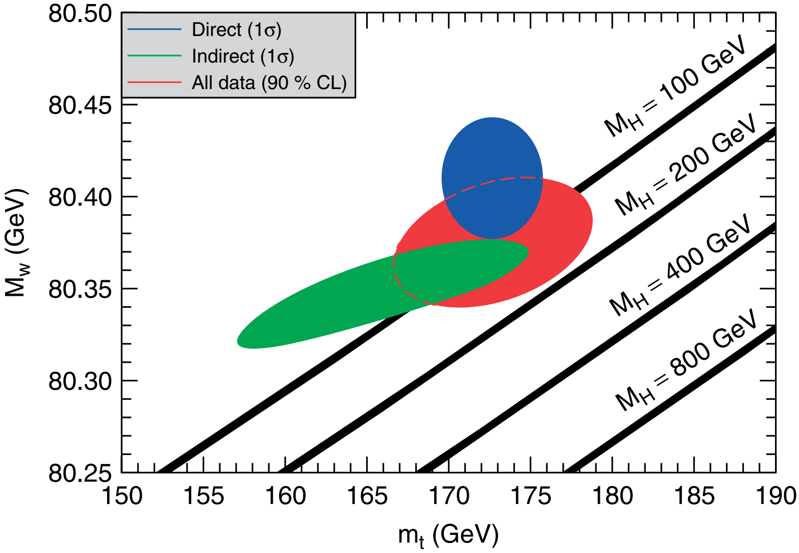

Fig. 6.3 Allowed contour in the *W* and top mass plane for different Higgs masses. Note that a light Higgs boson is weakly favored.

6.2.5
Higgs Boson

The *W* and *Z* are believed to gain their mass by interacting with the hypothesized Higgs boson, H^0. This interaction allows for the evasion of the rule that gauge bosons, for example, the photon and gluon, are massless. The situation is similar to that for a massless photon propagating in a superconductor (Meissner effect), wherein the photon acquires a mass by interacting with the medium. The Higgs ground-state field is considered to be nonzero and to pervade all of space–time. This background field is held responsible not only for the generation of the *W* and *Z* mass, but also for the generation of the masses of all SM particles. A more basic understanding of the Higgs field and its coupling to quarks and leptons should shed light on the regularities of the masses, which range from 0.51 MeV for the electron to 175 GeV for the top quark. Note that the Higgs boson has not yet been discovered and there is no prediction for the Higgs mass in the SM. The mass scale is set by the vacuum value of the Higgs field, $<\phi> \sim 174$ GeV.

The masses of the *W* and top quark acquire contributions from "quantum loops" (emission and virtual reabsorption) containing the Higgs boson. The data for masses allows for indirect constraints on the Higgs mass, shown in Figure 6.3. A light Higgs mass is favored by the data. The Higgs boson is also being directly searched for. A limit of 114 GeV was set at the Large Electron–Positron (LEP) collider at CERN. At the Fermilab Tevatron, searches are being made in several final-state channels where the Higgs boson should decay. The exact search strategy depends on the Higgs mass, which is unknown. Present limits are as not yet sensitive to the cross section of a SM Higgs particle, and so t more data is needed.

6.2.6
Neutrino Oscillations

In the SM, neutrinos are assumed to be massless, although there is no known symmetry that requires this. Indeed, quite recently, neutrinos have been observed to change their generation type (or "oscillate") as they propagate in vacuum (Hosaka *et al.*, 2005).

This behavior requires that the neutrinos have a finite mass. Therefore, we now know that neutrinos have small masses and that they mix their eigenstates in a way analogous to the quarks. This new 3 × 3 unitary neutrino matrix must have its matrix elements determined experimentally just as for the CKM matrix. The oscillation experiments measure the difference of the squares of the neutrino masses. At present, both atmospheric and solar neutrino mass differences are known and the mixings are large, in stark contrast to the quark mixings.

$$\sqrt{\Delta m^2_{\nu\mathrm{atm}}} \sim 0.05\,\mathrm{eV},$$
$$\sqrt{\Delta m^2_{\nu\mathrm{sol}}} \sim 0.008\,\mathrm{eV} \qquad (4)$$

Data from the Fermilab Main Injector Neutrino Oscillation Search (MINOS) experiment are shown in Figure 6.4. The neutrino oscillation is measured by observing a deficit of the produced neutrinos after a long flight path; the produced neutrinos "disappeared" in that they oscillated into a neutrino of another flavor that was not observed.

There is now a program to explore the matrix elements of the "neutrino CKM" matrix. For example, two mixing angles are known to be large, but the third is only known to be small. The implication of the small neutrino mass, m_ν, implies the existence of heavy neutrinos at a very

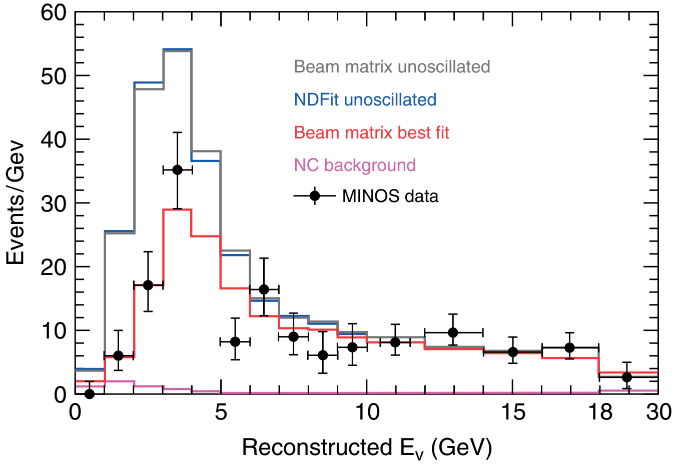

Fig. 6.4 Energy distribution of neutrinos observed after a long flight path subsequent to production at Fermilab. The disappearance of low-energy neutrinos is evident.

large energy scale, M_R by the "seesaw" mechanism relating to the quark mass, m_q:

$$m_\nu \sim \frac{m_q^2}{m_R} \tag{5}$$

Assuming that the actual masses are near the mass differences, taking 0.008 eV for the neutrino mass and 1 GeV for the quark mass yields a heavy neutrino mass of 1.2×10^{13} GeV.

This is an energy scale very far above that associated with the SM forces and is a first experimental indication that the SM is incomplete. Recall that we have ignored gravity so far. A completely naïve extrapolation of Newtonian gravity, which increases in strength as the square of the energy scale to the point where the gravitational coupling is as strong as the EM, yields a scale of 1.05×10^{18} GeV, which is another example of a very high-energy scale.

6.2.7
Frontier Issues in Particle Physics

6.2.7.1 Unification

The SM has been spectacularly successful as a synthesis of our understanding of all fundamental phenomena. However, it is incomplete. For example, why are there only three generations of quarks and leptons? The SM gives us no answer to this question. Why is the top quark so heavy (or why are the other fermions so light?) A basic understanding of the mass spectrum of the SM constituents eludes us.

Note that the SM is mute about any relationship between leptons and quarks, yet the masses of quarks and leptons seem to exhibit a pattern, for example the near equality of masses of u quark and e or strange quark and muon. The situation is akin to that in which atomic spectroscopists found themselves before the advent of quantum mechanics. Clearly, there were patterns in the spectra, but the underlying dynamics was not available.

Dramatic progress in physics in the past has occurred when a unification of forces has taken place, such as the recent EW unification. It is possible that the similarities in mass between quarks and leptons indicate that they are related in some more basic manner. In that case, the heavier quarks would eventually decay into light leptons and, the proton, for example, would not be absolutely stable. Present experiments set a lower limit of $\sim 10^{31}$ years on the proton lifetime. Note that some violation of baryon-number conservation is in fact also needed along with CP nonconservation, if the observed excess of matter in the universe is to be explained. Although the idea of unification of QCD and EW forces is tantalizing, no truly compelling evidence for such unification exists as of now.

6.2.7.2 Dark Energy

Cosmological data have recently been taken, which require that 73% of the energy density of the Universe is an unknown "dark energy (DE)" that is accelerating the expansion of the Universe. This is a new fundamental field that we are completely ignorant of. Is it related to inflation of the early Universe? Is it a cosmological constant? The accuracy of present data is consistent with that hypothesis, but the implied background field is $<\phi> \sim 0.002$ eV, which is 10^{14} times smaller than the Higgs field. Why is the Higgs field gravitationally inert? Clearly, our theories of gravity or the SM are either incomplete or incorrect. The exploration of the Higgs boson, therefore, has a high priority in particle physics with implications outside the SM.

6.2.7.3 Dark Matter

Cosmology tells us that 22% of the energy density of the Universe is ‘‘dark matter (DM),’’ whereas only 4.6% is ‘‘normal’’ baryonic matter. Clearly, there must be new fundamental fields beyond those contained in the SM. DM was first inferred from galactic rotation curves (Fich and Tremaine, 1991). Experiments have been mounted to detect the elastic recoils of the supposed DM particles off laboratory material and limits have been set, as yet without positive detection.

If the DM particles have masses at the 1000 GeV scale, then standard cosmology tells us that the 22% relic density requires that the DM annihilate with a cross section characteristic of weak processes. There is a SM extension with a new symmetry relating particles differing by spin 1/2 called *supersymmerty* (SUSY). Because that model requires doubling the number of fundamental particles, it provides a neutral, stable, and weakly interacting candidate for the DM. The mass scale for SUSY also gives the correct relic density. Perhaps this is a coincidence or a strong hint. Clearly, future searches for SUSY particles are of high priority in particle physics.

6.3
Accelerators and Other Facilities

After this very brief explanation of the state of the art and frontiers of particle physics, we turn to a description of the facilities available to confront those frontier issues. Experimental particle physics is the subject of worldwide study, which takes place in a variety of facilities. The primary focus for experimental particle physics has traditionally been the accelerators of ever-increasing energy. The increase in energy is driven by the desire to produce heavier particles together with the Einstein relation between mass and energy, $E = mc^2$. This growth in energy will, naturally, have some ultimate limits. Hence, other avenues for research may gain in importance in the future.

The program described below constitutes a broad-based attack on the validity of the SM following up on strong hints from Higgs, DE searches, neutrino oscillations, and DM. It is focused on the outstanding problems alluded to in Section 6.2.

6.3.1
Accelerator-Based Particle Physics

The locations and energies of the accelerator facilities operating at the highest energies are shown in Table 6.2. There are two broad categories of particle physics experiments: fixed-target experiments, where a beam of accelerated particles strikes a stationary target, and collider experiments, where beams of particles collide head-on. The former type of experiment is unrivaled in the study of rare interactions, while the latter is the technique of choice in exploring high mass scales because all the beam energy in such a collision is available for particle production. This distinction appears in Table 6.2, where beam energy is distinct from center-of-momentum (CM) energy, which is roughly the energy available to create new particles.

6.3.1.1 Fixed-target Experiments

Fixed-target experiments are currently in progress at several facilities in Russia, Europe, Japan, and the United States. Secondary beams of unstable particles of different types are the exclusive purview of fixed-target experiments. Examples include photon, muon, pion, kaon, and neutrino (γ, μ, π, K, ν) beams. Following the discovery of neutrino oscillations, a strong

Tab. 6.2 Operational experimental facilities at accelerators.

Type	**Location**	**Name/energy**	**CM energy = $\sqrt{s}$ (GeV)**	**Experiments**
Fixed target	KEK, Japan	12 GeV	5.0	Neutrino beams
	BNL, United States	AGS	–	Muon magnetic moment
	–	33 GeV	8.0	–
	SLAC, United States	50 GeV	9.8	–
	Serpukhov, Russia	76 GeV	12.0	Test beams – LHC, ILC
	CERN, Switzerland	SPS	29.2	
	–	450 GeV	–	Kaon Beams – rare decays
	Fermilab, United States	Tevatron, Main injector	38.8	–
	–	800 GeV, 120 GeV	–	–
e^+e^- collider	Beijing, China	BEPC	4	$c\bar{c}$ spectroscopy
	Cornell, United States	CESR – CLEOC	10	$c\bar{c}$ spectroscopy
	KEK, Japan	KEKB – Belle	10	$b\bar{b}$ spectroscopy
	CERN, Switzerland	LEP II	200	W mass, $WW\gamma$ coupling
	SLAC, United States	PEP-II – BaBar	10	$b\bar{b}$ spectroscopy
ep collider	DESY, Germany	HERA – ZEUS, H1	314	Proton structure function c, b production
$\bar{p}p$ collider	Fermilab, United States	Tevatron – CDF, D0	1960	Higgs, SUSY search
Nucleus–nucleus collider	BNL, United States	RHIC	200	Quark–gluon plasma

international program using intense neutrino beams is being mounted.

6.3.1.2 e^+e^- Collider Experiments

Heavy quarks form quark–antiquark bound states, ''onia,'' in analogy to the electron–positron bound states of positronium. These states occur at high enough mass scales that QCD perturbation theory may be expected to apply. The QCD force becomes weaker at high mass, as is discussed in Section 6.7. The explication of the spectroscopy of the bound-state systems is a vital test of QCD, just as positronium spectroscopy was a crucial test for quantum electrodynamics (QED).

At present, an international effort to study CP violation in the b quark system is nearing completion at Belle and BaBar. The beautiful data on the ''unitary triangle'' that comes from these ''B factories'' is shown in Figure 6.2. Searches for rare c quark decays at experiments Cornell (CLEOC) and Beijing (BEPC) are also nearing program completion.

At elevated CM energies, the Z are resonantly formed. Very precise EW tests were performed at these Z ''factories,'' which yield the most accurate determinations of the EW parameters. The LEP accelerator was then upgraded and data was taken at higher energies. That data produced a W mass determination, a determination of the nonlinear self-coupling of the gauge bosons, and the best current limit on the Higgs mass (greater than 114 GeV). The LEP program was terminated in preparation for the Large Hadron Collider (LHC) at CERN.

6.3.1.3 *ep* Collider Experiments

At present, the distribution of quarks and gluons in the proton is determined experimentally. The data come from lepton–proton scattering with fixed proton targets, and neutrino, electron, and muon incident beams. The range of such experiments is limited by the available CM energy.

6.3.1.4 $p\bar{p}$ Collider Experiments

The proton is a composite object and can be considered to consist of valence *uud* quarks, quark–antiquark pairs, and gluons of different momenta. Therefore, $p\bar{p}$ colliders are capable of performing simultaneous studies of many aspects of current investigations in particle physics. Examples include QCD studies in jet phenomena, $W/Z/\gamma$ production and decay properties, heavy-flavor (b) quark production dynamics and decay spectroscopy, studies of the top (t) quark, and searches for new phenomena outside the SM. Most notably, the Tevatron experiments are making precision measurements of the W and t quark mass, searching for the Higgs boson and for evidence of SUSY particles (see Figure 6.3).

6.3.1.5 Applications

Accelerators have also been a fruitful source of applications to areas outside particle physics. The large-scale use of superconductivity was pioneered in accelerators for high-energy physics. Large-scale high-vacuum and distributed-control systems are another outgrowth of accelerator ''technology transfer.'' Electron accelerators are used as synchrotron light sources. These new facilities are heavily subscribed to by materials and solid-state physicists at many locations worldwide.

In addition, there are numerous examples of medical applications. The detectors used in particle physics are widely used in medicine. High-field magnets similar to those built for accelerators are used in magnetic resonance imaging. Finally, accelerators themselves are used in treatment. Examples include electron, neutron, and proton irradiation. For example, proton beams of variable energy allow precise doses to be placed deep in the body with minimal damage to intervening tissues. Electron linear accelerators are, however, simpler and therefore the most widespread therapeutic devices.

6.3.2
Nonaccelerator Particle Physics

As mentioned above, DE is a new fundamental field. Because it operates only over cosmological distance scales, the studies will, of necessity, be made only using ground- or space-based telescopes.

The DM is a second unknown fundamental field that is unexplained in the SM. Direct measurement of the scattering of DM particles is being attempted at very low background underground facilities. These studies are complementary to the searches for SUSY DM candidate particles produced in accelerator experiments. Further, as with DE, the accelerator, and nonaccelerator experiments are synergistic.

Cosmic rays extend up to energies that are still inaccessible at the accelerator facilities given in Table 6.2. Therefore, it is natural to search for new phenomena at the highest cosmic-ray energies. In addition, stimulated by unification speculations, several experiments have been set up to search for proton decay, and all experienced failure to date. However, they have set useful limits, and because they are located deep underground to minimize cosmic-ray backgrounds, they also function as active neutrino ''observatories.'' Detectors of this type – for example, at Kamioka in Japan, Gran Sasso in Italy, and Soudan in the United States – detected the neutrino pulse from supernova SN1987 A. The utility of these ''observatories'' indicates the increasing interconnection between particle physics, astrophysics, and cosmology. A DM search Cryogenic dark matter search (CDMS) is currently operating in the Soudan facility.

There were tantalizing indications that the number of neutrinos arriving from the sun was not in accord with the standard solar model prediction (Davis *et al.*, 1989). This was confirmed by the observation of atmospheric neutrino oscillations (Hosaka *et al.*, 2005) at Kamioka and of electron-neutrino oscillations at reactors (Eguchi *et al.*, 2003). New data from Canada, at the Sudbury Neutrino Observatory (SNO) (Aharmim *et al.*, 2005), has greatly helped in clarifying the neutrino mixing.

Neutrino beams prepared so that they are composed of only one neutrino generation at an accelerator can be directed toward distant underground observatories, where the appearance of a different generation of neutrino can be searched for. Such long-baseline neutrino experiments will improve the determination of neutrino masses and mixing parameters. These positive results from underground laboratories have stimulated a new generation of accelerator experiments in Japan (K2K), Europe (Gran Sasso), and the United States (MINOS).

6.4
Mapping the SM Constituents onto the Detectors

Having described the SM constituents and the present experimental program, one can now briefly describe the specifics of experimental particle physics. Detectors are used to measure the vector position and momentum of all the constituents of the SM emanating from a collision and to distinguish between them (particle identification, "id"). In principle, a perfect detector would measure and identify all the particles in both the initial and final states of a collision. The design of detectors is dictated by the characteristic interactions that the SM constituents participate in. The different categories are shown in Table 6.3.

Tab. 6.3 Detection/identification methods.

Signature	Detector	Particle
Jet of hadrons	Calorimeter	$u, c, t \rightarrow Wb$ d, s, b G
Missing Energy	Calorimeter	ν_e, ν_μ, ν_τ
Electromagnetic shower	Calorimeter	$e, \gamma, W \rightarrow e\nu$
Only ionization interactions	Muon absorber	$\mu, \tau \rightarrow \mu\nu\overline{\nu}$ $Z \rightarrow \mu\mu$
Decay with $c\tau \gtrsim 100\,\mu\text{m}$	Si tracking	s, c, b, τ

6.4.1
Confinement and Jets – Quarks and Gluons

Quarks(q) and gluons(g) carry the color quantum number and interact strongly. It has been found that quarks and gluons cannot be detected as distinct free states. Rather, they appear in detectors as collimated "jets" of colorless hadrons moving nearly in the direction of and with roughly the momentum as that of the parent q or g. Finely segmented energy-measuring detectors (calorimeters) are used to measure the positions and energies of these ensembles of particles. Thus, one may infer the properties of the q or g parent.

6.4.2
Missing Energy and Neutrinos

Leptons, by definition, interact only via the EW interaction. In particular, neutrinos are uncharged, and therefore interact only weakly. Therefore, neutrinos pass right through a calorimeter system that absorbs and detects all strongly and electromagnetically interacting particles, without interacting with the detecting medium. The presence of neutrinos is then signaled by the existence of the "missing energy" that they carry off undetected

6.4.3
Charged Leptons and EM Interactions

Charged leptons interact electromagnetically. Extremely relativistic photons and electrons interact with a calorimeter medium by pair production and bremsstrahlung, respectively. The resulting cascade is a characteristic signature. Therefore, a calorimeter can be used to perform id of e/γ particle by observation of the characteristic cascade.

The muon is unstable, with a lifetime characterized by the decay length $c\tau \sim 658\,\text{m}$. Therefore, it will typically pass through a detector before it decays. Muons are "heavy electrons" – second-generation leptons identical to electrons save for mass. The large mass means that with the identical forces experienced by e and μ, the acceleration of a muon is much reduced with respect to an electron. Because radiation is caused by acceleration, the muon cascade occurs only at much higher energies than the e cascade. For example, muons at energies below about 300 GeV only ionize the detection medium, thereby leaving a minimum-ionizing particle, (MIP), signal in particle detectors. This signature is the basis for the identification of the muon.

6.4.4
W, Z, and τ Decays

The W and Z gauge bosons rapidly decay into lighter final states containing either lepton or quark pairs. The cleanest experimental signatures are the decays into final-state leptons, $W \rightarrow l + \nu$ and $Z \rightarrow l^{+} + l-$ where l denotes charged lepton. Final states containing quarks, such as $W^{+} \rightarrow u + \overline{d} \rightarrow \mathrm{Jet}_1 + \mathrm{Jet}_2$, are subject to large backgrounds.

The τ lepton often decays into lighter final-state leptons – for example, $\tau \rightarrow \mu + \overline{\nu}_{\mu} + \nu_{\tau}$. Because the τ has a measured lifetime of $c\tau \sim 89\,\mu\mathrm{m}$, the experimental signature is the existence of both a secondary decay vertex (see Section 6.4.5) and a primary production vertex.

6.4.5
Displaced Secondary Vertices

All the elementary constituents of matter except electrons, protons, and massless neutrinos are unstable. The leptons decay, for example, $\mu \rightarrow e + \overline{\nu}_e + \nu_{\mu}$, with a decay rate that scales as the available energy to the fifth power. Therefore, the light muon can be considered as stable during its detection, while the heavy τ must be detected by measuring its decay products.

The separation of light (u, d) from heavy (s, c, b) quark states can be accomplished by the observation of a secondary decay vertex. The quarks can decay semileptonically – for example, $s \rightarrow u + l + \overline{\nu}_l$. Hadrons containing s quarks have lifetime distances, $c\tau$, of the order of centimeters, while c and b quark decays lead to hadron lifetimes, $c\tau \sim 200\,\mu\mathrm{m}$.

6.5
A Model *pp/$\overline{p}$p* Collider Detector

The characteristic interactions of the SM constituents are used in their detection as outlined above. In this section, a ''typical'' detector that is used in experimental particle physics is described. Fixed-target detectors are often special-purpose instruments optimized to explore a specific type of physics – for example, neutrino interactions. Collider detectors tend to be general-purpose instruments that are designed to study a large variety of reactions simultaneously as mentioned above. Therefore, a collider detector can be discussed ''generically.'' Generally, the backgrounds and trigger strategies are more challenging at a hadron collider than at an $e^{+}e^{-}$ collider, and therefore, the former is described. A schematic of a typical detector and event is shown in Figure 6.5.

6.5.1
''Soft'' Backgrounds

At a CM energy of 1800 GeV, the *pp* inelastic cross section is $\sim 50\,\mathrm{mb} = 5 \times 10^{-26}\,\mathrm{cm}^2$. The overwhelming majority of these interactions are soft, that is, they consist of produced hadrons with limited transverse momentum, $\langle p_t \rangle \sim 0.5$ GeV, where p_t is the component of secondary particle momentum transverse to the incident beam direction. The secondary particles are distributed in longitudinal momentum, $p_{|}$, according to single-particle phase space:

$$\begin{aligned}\text{single-particle phase space} \\ &= d^4p\delta\,(p^2 - m^2) \\ &= \frac{d^3p}{E} \sim dydp_t^2 \qquad (6)\end{aligned}$$

where m is the particle mass, E is the particle energy, and y is the rapidity, which is defined by $\tanh y = p_{|}/E$. The nonrelativistic limit for one-particle phase space

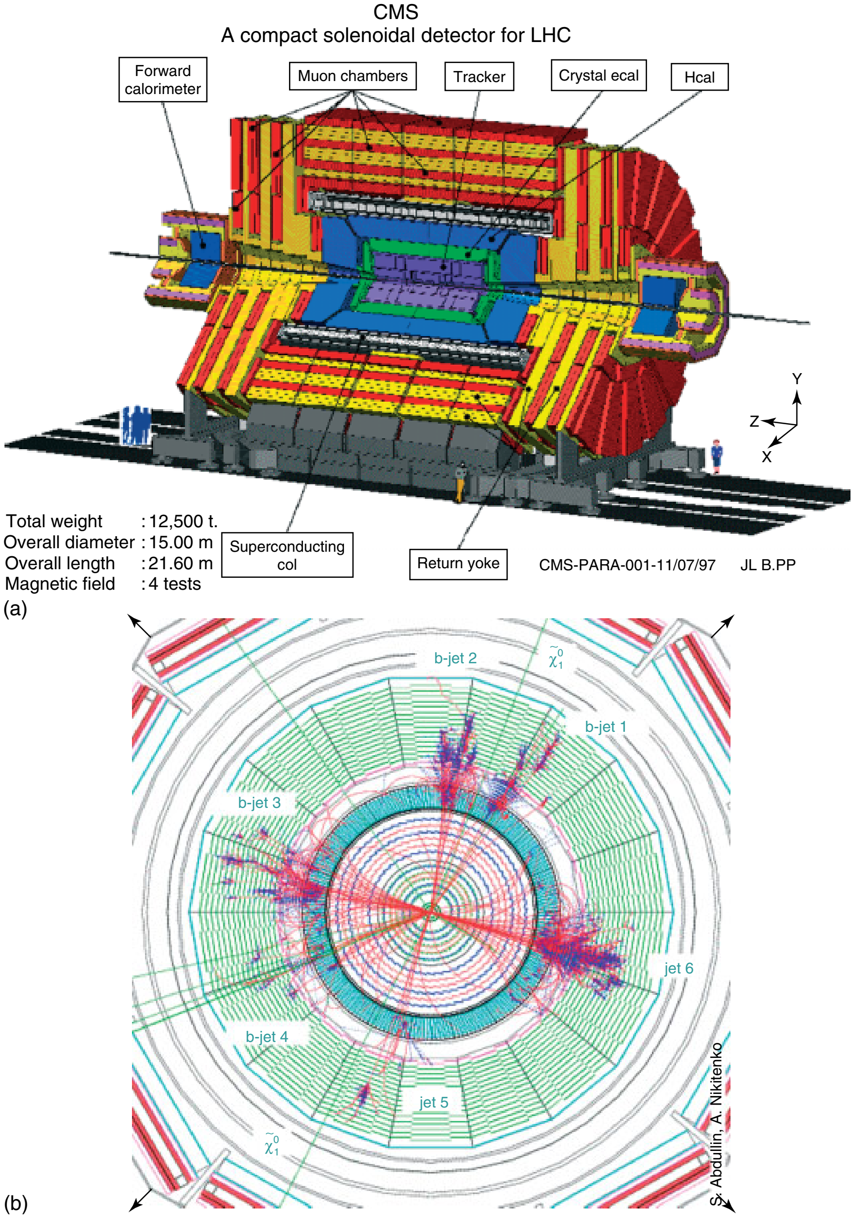
CMS
A compact solenoidal detector for LHC
Forward calorimeter
Muon chambers
Tracker
Crystal ecal
Hcal
Y
Z
X
Total weight : 12,500 t.
Overall diameter : 15.00 m
Overall length : 21.60 m
Magnetic field : 4 tests
Superconducting col
Return yoke
CMS-PARA-001-11/07/97
JL B.PP
(a)
b-jet 2
$\tilde{\chi}_1^0$
b-jet 1
b-jet 3
jet 6
b-jet 4
jet 5
$\tilde{\chi}_1^0$
S. Abdullin, A. Nikitenko
(b)

Fig. 6.5 (a) A typical pp collider detector being constructed for the LHC showing tracking, calorimetry, muon electromagnets, and muon detectors. (b) Charged tracks and EM and HAD energy deposits are shown, indicating the existence of six jets, four jets associated with displaced secondary vertices and missing energy in the simulated SUSY interaction.

is d^3p/m, which is familiar from classical statistical mechanics.

The collisions of greatest interest occur at large transverse momentum because they probe small distances. Therefore, the detection, triggering, and data-acquisition (DAQ) strategy adopted for these detectors puts a strong emphasis on collecting the rare ''hard'' collisions, those containing high-p_t secondary particles.

6.5.2
Magnetic Analysis

Typically, the detector contains a large electromagnet, which provides a volume filled with a magnetic field. A uniform magnetic field causes all charged particles to move on helical paths, with radii of curvature determined by the particle momentum and charge. Thus, determination of the particle trajectories in the field volume determines the charge, vector momentum, and position of all the charged particles produced in the collision.

6.5.3
Si Inner Tracker

Because b and c quarks and the τ leptons have lifetimes of order 200 μm, an inner solid-state detector consisting of strips of Si with a strip spacing of ∼50 μm or pixels with sides of a similar size is often employed as the first detector element encountered by a secondary particle after it exits the vacuum pipe containing the incident beams. The Si detector strip spacing/pixel width is well matched to the decay distances of the particles. The Si detector also serves as the initial element in a tracking system that measures the particle trajectory in the magnetic field.

6.5.4
Outer Tracker

The Si elements are fairly expensive. Therefore, most tracking systems employ a different detector technology at larger radii. Gas-filled wire chambers (see Chapter 23) are the present standard. The size of these outer trackers is typically 1–2 m in radius. However, the high interaction rates of the LHC are pushing detector builders to higher rate capable systems. For example, the CMS detector uses all silicon in the tracking system.

A perfect tracking system in principle allows the experimenter to perform particle id on the c and b quarks and the τ lepton. The e and μ charged leptons are also measured and, by extension, the leptons from τ, W, and Z decays. All charged-hadron trajectories are also determined.

6.5.5
Calorimetry

The next system encountered by the outgoing particles is usually the calorimeter (see Chapter 21). Within the calorimeter, particles cascade in a dense medium and deposit their energy. The calorimeter is segmented transversely so that this system measures, destructively, the position and energy of all particles, both charged and neutral, that possess strong/EM interactions. The calorimetric medium can be either completely active (crystals) or interspersed heavy absorber plates with active detection layers (sampling calorimetry). Particle id is achieved by using the transverse and longitudinal segmentation of the calorimetry to distinguish between the differences in cascade development of the different incident particles – for example, electrons, and hadrons.

6.5.5.1 EM Segment

The first longitudinal segment encountered typically contains Pb or some other element with high atomic number that induces e and photon cascades via the EM interaction. Because hadrons interact differently, the EM segment is used to identify e and γ. The tracking system is then used for e–γ separation because e will leave an ionizing track while the uncharged photon will not.

The depth of the EM segment is sufficient to contain the EM cascade fully. It is typically $\sim$20 radiation lengths or $\sim$12 cm of Pb. By comparison, the ratio of radiation length to hadron interaction length is $\sim$30 in Pb, so that most hadrons will not even begin to produce a cascade in the EM segment. This difference in interaction-length scales, and a similar difference in the transverse size of the EM and hadronic (HAD) cascades, is the basis of clean e/γ particle id with respect to hadrons.

6.5.5.2 Hadronic Segment

The EM segment is followed by a HAD segment. Its task is to absorb all the energy of the hadrons. The scale forHAD interactions defines the size of the calorimeter system. For example, $\sim$10 nuclear absorption lengths (or 167 cm of Fe) are needed to absorb most of the HAD energy. Any particle that passes through the calorimeter unscathed is a candidate for identification as a muon. Muons will deposit ionization energy only in the calorimeter. Thus, a particle that acts like a MIP in the calorimeter can be provisionally identified as a muon.

6.5.6
Muon Detection

Detectors external to the calorimetry are often employed to provide independent muon identification and a redundant measurement of the muon position and momentum. This redundancy is similar to that provided by the two measurements of an electron by tracking in a magnetic field and by the measured energy deposited in the EM segment of the calorimetry. Redundancy is required when looking for rare processes where extremely clean particle id is called for.

Redundant muon information is supplied by a second magnetic analysis of the muon after most of the secondary particles in the interaction are removed by absorption in the calorimeter. This analysis may occur either in air or in a solid steel electromagnet. Wire chambers surrounding this additional magnetic volume typically provide the trajectory.

6.5.7
Trigger and DAQ

Because the inelastic cross section is large and the cross sections for the processes of interest are often small, the trigger for readout must be very selective. The ratio of all events to the events of primary interest may be as small as 10^{-10}. With such a small signal-to-noise value, redundant measurement and particle id are absolutely necessary. The situation for e^+e^- interactions is somewhat more relaxed.

In modern detectors, the trigger is hierarchical. The initial high rate is reduced by specialized electronics built by the experiment. A selection after that is done using digital information, commercial event building switches and ‘‘farms’’ of commercial computers working in parallel.

The trigger hierarchy is designed to reduce the total inelastic rate ($\sim$900 MHz at the LHC) to a rate logged to permanent medium (e.g., magnetic tape) of perhaps 100 Hz. Clearly, the trigger system cannot accomplish this prodigy (reduction by a factor 5×10^5) in one step. Several successively slower but more incisive trigger algorithms are used.

6.5.8
Analysis

Often, crude analysis of complete events consisting of the full set of information supplied by the detector is part of the highest level of triggering. Thus, triggering and DAQ merge smoothly into off-line analysis of the events. Typically, a complete event is assembled from the individual detector systems by the DAQ system and loaded into a high-level workstation where the highest level triggering is performed. Off-line analysis of the events stored on a permanent medium can be more leisurely and more exacting.

Note that these experiments will generate vast amounts of data. In addition, they are operated by large international collaborations with members distributed over the entire globe. Therefore, the task of analyzing and distributing the data is formidable. Indeed, the issues for particle physics are at the frontier of distributed computing and serve as a cutting-edge example for ‘‘grid computing,’’ which features distributed processing and data storage.

6.6
Other Experimental Techniques

Detectors are designed to address specific physics processes. Some detector systems are of a more specialized nature than the generic detectors for collider experiments that were discussed in the previous two sections. All detectors of necessity must utilize the interactions of the particles of interest with a detecting medium. In the previous sections, ionization in the tracking system, EM/HAD cascades in the calorimetric media, and ionization energy loss in the muon system were discussed.

6.6.1
Ionization Detectors

Ionization refers to the interaction of incident charged particles with the atomic electrons in a medium, which leads to ionization of the atoms. The energy loss by ionization varies with incoming particle velocity v as $1/v^2$ for slow incident particles. At high velocity, the ionization loss approaches a minimum and then slowly rises with incident energy.

If the value of the deposited energy is measured several times, the velocity is determined. With the momentum known from the curvature of the path in the magnetic field, the mass of the particle may thus be inferred. Therefore, ionization detectors can be used for slow-particle id if many redundant samples of the ionization are made. The medium used can be gaseous, liquid (e.g., liquid argon), or solid (e.g., silicon). Solid-state detectors using Si or Ge are used to measure ionization in many nonaccelerator experiments. When operated at cryogenic temperatures, the lattice phonon excitation can also be measured, yielding complementary information (e.g., CDMS).

6.6.2
Photomultiplier Tubes, Scintillator

Scintillating plastics coupled to photomultiplier tubes (PMT) can be used to measure the particle time of flight between two sets of counters to subnanosecond (10^{-9} s) accuracy. The measurement of time over a fixed distance is a velocity measurement. Thus, scintillation-counter arrays are also useful in particle id for slow ionizing particles.

6.6.3
Slow-*n* Detection

Neutrons are uncharged and hence do not ionize. Nevertheless, they can be detected by measuring the proton recoils from elastic *np* interactions. The recoil protons take up a significant fraction of the initial neutron energy, about half, on average. Thick organic scintillation counters can both supply the proton targets and detect the recoil protons via their large ionization losses in the plastic.

6.6.4
Cherenkov Detectors

A charged particle with a velocity v that exceeds the velocity of light in a medium of index of refraction n, $v > c/n$, emits a Mach cone of radiation. This is called the *Cherenkov effect*. Cherenkov counters consist, typically, of a gas or liquid-filled vessel outfitted with mirrors to collect the emitted light, and PMT to convert the light into a fast electrical signal. The angle of the emitted light, θ_c, is determined by $\beta = v/c$ and $n : \cos\theta_c = 1/\beta n$. Clearly, a measurement of the angle of the emitted light is a measure of the particle velocity. Such counters are called ring- imaging Cherenkov counters (RICH) (from the Mach cone image).

6.6.5
Transition-radiation Detectors

A transition-radiation detector (TRD) may be considered as utilizing the diffraction pattern arising from the radiation of a Cherenkov detector of finite length. The radiation is confined to very forward angles. In addition, it is in the X-ray frequency range, with an intensity that is proportional to $\gamma = 1/\sqrt{1-\beta^2}$. Therefore, the TRD is a device that is particularly useful for particle id with highly relativistic particles. The TRD is a detector complementary to those, like the Cherenkov detector, which lose discrimination power as $\beta \to 1 (v \to c)$.

6.7
Future Developments

Particle physics is a dynamic and vital field with an extremely broad-based experimental program, as has been shown above. In general, the program seeks to probe the SM either with increasing precision or with higher energies where phenomena appear to become simpler. To date, the SM has withstood this assault and proven itself capable of very precise and testable predictions. To make decisive progress, new facilities are now under construction, as shown in Table 6.4, to press forward on these two complementary fronts. Note also that with the discovery of DM and DE, particle physics is evolving toward a greater emphasis on nonaccelerator experiments.

6.7.1
The Energy Frontier at the LHC

6.7.1.1 Higgs Boson

The SM predicts the existence of a neutral scalar Higgs particle, but does not constrain the mass. One can estimate an upper limit to the requiring Higgs mass by requiring that the EW interactions do not become strong. For example, the *S*-wave amplitude A_0 for the EW process $e^+ + e^- \to W^+ W-$at

Tab. 6.4 New facilities.

Type	**Location**	**CM energy $\sqrt{s}$ (GeV)**	**Experiments**
pp collider	CERN, Switzerland Large Hadron Collider (LHC) – ATLAS, CMS, LHCb	14 000	Higgs, SUSY search t factory
	ALICE		b spectroscopy, CP nonconservation heavy-ion collisions
e^+e^- collider	International Linear Collider (ILC)	500–1000 GeV	Precision Higgs, SUSY studies.
Reactors	Daya Bay, China CHOOZ, France		Third neutrino mixing angle
Neutrino beams	Fermilab, United States -NOVA	Variable – off axis narrow band	Neutrino mixing matrix, CP violation and neutrino mass spectrum
	Japan Italy	Beams	
Underground lab	DUSEL – United States SNO – Canada		DM search, p decay, very large neutrino detectors, $\beta\beta 0\nu$ search
Dark energy survey	Several facilities	telescopic	Equation of state of DE

a CM energy squared of s is

$$A_0(ee \longrightarrow WW) = \left(\frac{\alpha_w}{4\pi}\right)\left(\frac{s}{M_w^2}\right) \qquad (7a)$$

Because A_0 increases with energy, we expect that new phenomena will occur to restore partial-wave unitarity. This restoration guarantees that new physics will be seen if the H^0 mass range below about 1000 GeV can be explored:

$$A_0(ee \longrightarrow WW) < 1 \Longrightarrow M_H < \sqrt{\frac{4\pi}{\alpha_w}} M_w \sim 1600 \text{ GeV} \qquad (7b)$$

The discovery of the Higgs boson is crucial if we are to progress to a deeper understanding. The full exploration of the TeV (1000 GeV) mass scale is the province of the next machine at the energy frontier, the LHC. The LHC is the single new facility to operate at the EW-breaking mass scale (set by $<\phi> = 174$ GeV) so that the LHC is truly a discovery machine.

Note that a one-year run at design luminosity of the LHC provides an exposure of $100 fb^{-1}$. All Higgs masses from the LEPII limit to 1000 GeV will be discovered at the LHC with only 10% of this exposure. When and if the Higgs is discovered, the mass and width and quantum numbers as well as interactions with other SM particles will need to be determined. Note that in the SM, the only unspecified parameter is the Higgs mass, so that completely unambiguous predictions are made for all these quantities after the mass is measured. Thus, very incisive tests of the SM become available after the discovery.

6.7.1.2 **Supersymmetry and DM**

Evidence for a very high mass scale comes from the small neutrino masses and from extrapolating Newtonian gravity. There is a third hint that high masses bring new simplifications beyond the SM indicating new symmetries or interactions as yet unknown. This third scale requires some explanation. It is well known (vacuum polarization) that the EM coupling constant is not a constant but increases with decreasing distance of observation because of the screening effect of the virtual cloud of e^+e^- pairs present in the vacuum polarization. For example, $\alpha(M=0) \sim 1/137$, while $\alpha(M=M_z) \sim 1/128$. We have asserted that the QCD quanta are colored gluons, whereas the EM quanta are uncharged photons. This implies that QCD is weak at large mass and strong at small mass, that is, $\alpha_s(M=0) \to \infty$, which leads to q and g confinement. Therefore, since $\alpha_s(M=M_z) \sim 1/10$, as the mass increases further, α_s will drop and α will rise to meet it at some mass. Because the W and Z self-couple similarly to the gluons, α_w will also drop. It is remarkable that within errors α_s, α_w, and α, roughly meet at a mass scale $\sim 10^{15\pm1}$ GeV.

At this scale, QCD and EW forces are expected to be fully unified (Pati and Salam, 1973; Georgi and Glashow, 1974). Extrapolation by 13 orders of magnitude to the GUT scale leads to a potential unification of all forces, if no new physics intervenes. In fact, precision measurements of the coupling constants may decisively allow the conclusion that new physics does intervene. This is an example of the use of precision studies to probe presently inaccessible mass scales. For example, SUSY leads to a rather better unification than the SM alone and at a higher mass.

SUSY also predicts the value of the Weinberg angle (Eq. 1) rather precisely. It slows down the predicted rate of proton decay to be consistent with experiment. Most importantly, it protects the Higgs mass from radiative corrections coming from particles at the high mass scales contributing to the mass from quantum loops. SUSY provides this protection mechanism naturally if the SUSY particles have masses of about 1000 GeV. SUSY also provides a DM candidate, a stable neutral weakly interacting boson of mass about 1000 GeV.

The LHC provides a great discovery potential for SUSY particles. A one-year exposure at the design capability of the LHC will allow for the discovery of SUSY particles up to ~1000 GeV and thus either decisively disprove or prove the SUSY hypothesis. Therefore, the LHC is also a discovery machine for SUSY.

6.7.2
International Linear Collider (ILC)

As in the past (see Table 6.2), it is important that electron–positron as well as hadron colliders are operating. The ILC is being designed and will soon be formally proposed. It will operate in the crucial 1000 GeV mass range, where we have many hints that new physics beyond the SM will occur and where the LHC will make an initial study. This collider will have a ‘‘cleaner’’ environment with a well-defined fundamental initial state. The initial state polarization can also be prepared, allowing for new and more incisive tests than are available at the LHC. With the new LHC results being available in 2009–2010, the ILC can be built knowing in advance how to design the machine and detectors so as to maximize the impact on the new physics (Dawson and Oreglia, 2004).

6.7.3
Heavy Ions at the LHC

The search for a deconfined quark–gluon plasma will continue at the CERN LHC, which will also take data with nucleus–nucleus (A–A) collisions. The CM energy will be more than 20 times larger than the RHIC experiments. Therefore, at the LHC, higher initial temperature and higher energy probes of the medium will be available, for example, quarks with transverse momenta >200 GeV. There will be three experiments studying A–A collisions, ATLAS and CMS (general-purpose experiments), and ALICE, specifically designed to study heavy-ion collisions.

6.7.4
CP Violation After BaBar and Belle

At a fundamental level, one asks why the universe is composed of matter and not of antimatter. The violation of CP invariance is a necessary ingredient (Sakharov, 1967; Primakoff and Rosen, 1981). CP nonconservation has been extensively studied at ''B factories'' operated at the Stanford Linear Accelerator Center (SLAC – BaBar) and the Japanese National Laboratory for High-Energy Physics (KEK – Belle). To date, there appears no indication of CP violation beyond that allowed by the complex CKM matrix. More data and redundant measurements will overconstrain the ''unitary triangle'' and put ever tighter constraints on extensions to the SM.

The effects to be studied are so small that they may require the higher reaction rates available at hadron colliders. The LHC program contains an experiment dedicated to the study of b decays (LHCb) with smaller branching fractions and with higher statistics than the present experiments. Indeed, if SUSY is found in direct production at the LHC, then CP nonconserving effects would most likely be observable in b decay modes due to virtual quantum loops containing the SUSY particles. Thus the LHC experiments can be quite complementary.

6.7.5
The Future Neutrino Program

With the discovery of neutrino mass and generation mixing, the possibility of CP violation in lepton decays becomes possible. CP violation in heavy neutrino (seesaw mechanism) decays can cause a lepton asymmetry that can then be transmuted into a baryon asymmetry. For this scenario to be valid, the neutrino mixing matrix must have substantial complex elements. Initially, reactor experiments are being prepared in an attempt to detect the third mixing angle of this matrix.

In addition, accelerator neutrino beams directed toward a detector far away (long baseline experiments) are being prepared in the United States, Europe, and Japan. These experiments also aim to explore the third mixing angle, to study the masses of neutrinos themselves and not just the differences, and to start to attack the measurement of CP-violating terms in the mixing matrix.

Underground experiments, which achieve very low cosmic ray background levels, are also being mounted. Searches for neutrinoless double β-decay, $A_Z \rightarrow A_{Z+2} + e^- + e^-$, $\beta\beta 0\nu$, are made because the rate for these processes is proportional to the neutrino mass and because the reaction clearly violates lepton generation conservation. The scale of mass from cosmological and astrophysical experiments is limited to be <0.2 eV. This reaction cannot occur unless the neutrino is its own antiparticle, a situation not found anywhere else in the SM. Smaller backgrounds are needed, leading to a proposed deep underground science and engineering laboratory (DUSEL). This facility could house next generation DM, p decay, $\beta\beta 0\nu$, and neutrino detectors.

6.7.6
DE Experiments

At present, DE appears to be consistent with a classical Einstein ''cosmological term'' with a constant background field. To sharpen that conclusion, many experiments are being contemplated to determine the equation of state of the DE and to find out whether it has evolved in time as the Universe unfolds. These surveys are telescopic, either ground- or space- based. More data on this basic new field, outside the realm of the SM, are very important to achieve.

6.7.7
The Far Future

New fundamental particles and/or interactions are expected to appear at the EW-breaking mass scale of $\sim$1000 GeV or below. It is impossible to predict the shape of this new physics. One possibility is that the objects that are now considered to be elementary are really composites. At each point in the past – atoms to nuclei to nucleons to quarks – this hypothesis has proved to be true. It remains clear that particle physics will maintain its vitality by again pushing to higher energies and more precise experiments. New facilities beyond those now under construction should include a TeV-scale e^+e^- linear collider (ILC). The choice of the appropriate tools clearly must await the insights provided by the new data produced at the facilities presently under construction.

Many of the questions to be answered depend on the success of the SM. A partial list might include these: what is the origin of EW symmetry breaking? Why are there three generations? What explains the pattern of the fermion masses? What is the origin of the CKM matrix? What is the origin of CP nonconservation? What is the origin of the neutrino-mixing matrix? Are the QCD and EW forces unified at some energy? What is the connection of gravity to the other fundamental gauge interactions? What is the correct quantum theory of gravity?

The high mass scales implied by gravity, neutrino masses, and GUT unification are many orders of magnitude above the capabilities of any feasible accelerator. The intellectual challenge of these scales is to design incisive experiments at available energies. This task will require considerable ingenuity.

It is humbling to note that we do not understand the nature of the basic fields that constitutes 95% of the Universe, the DM, and DE. Clearly, much remains to be understood, and the intellectual vitality of particle physics is assured for the foreseeable future.

Glossary

Baryon: Half-integral spin hadron made up of qqq valence quarks.

Boson: Integral-spin particle.

Calorimeter: A device that provides a signal proportional to the energy deposited by a particle incident upon it.

Color: The quantum number carried by quarks and gluons; the "charge" of QCD.

Confinement: The hypothesis that colored quarks and gluons cannot exist as free particles.

CP Symmetry: A symmetry of interactions under the combined parity and charge-conjugation operation.

Fermion: Half-integral spin particle.

Flavor: Approximately conserved quantum number of quarks. Pair production of strange, charmed, bottom, and top ($s\bar{s}$, $c\bar{c}$, $b\bar{b}$, $t\bar{t}$) quarks is required. Different generations of quarks have different flavor labels.

g_s: Strong coupling constant.

g_W: Weak coupling constant; analogous to e in QED.

Gauge Symmetry: A symmetry arising from requiring that physics be independent of the local phase convention chosen at arbitrary space-time points.

Generations: The replication of the basic quark/lepton four elementary fermions, u, d, e, ν_e, at higher masses.

Gluon: The massless quantum of QCD.

Hadron: A strongly interacting, colorless composite of quarks.

Higgs Boson: A hypothesized particle whose nonzero vacuum field gives rise to the masses of all fermions and bosons.

Interaction Length: The characteristic length of material within which a hadron interacts by a strong interaction with a nucleus.

Jet: An ensemble of hadrons that are the final-state remnants of a quark or gluon parent that is permanently confined. This ensemble approximately retains the direction and momentum of the parent.

Kaon: A meson composed of strange (s) and u or d quarks, for example, $K^+ = (u\bar{s})$.

Lepton: Fermion without color (QCD interactions).

Lifetime: A characteristic time for the exponential decay of unstable particles.

Mach Cone: The characteristic shock-wave pattern at fixed angle that appears when an object exceeds the speed of propagation of a wave in the medium.

Meson: Integral-spin hadron made up of $q\bar{q}$ valence quarks.

Neutrino: Uncharged lepton, assumed to be massless.

"onium": Bound state of heavy quark and antiquark, for example, $c\bar{c}$ is charmonium. Analog of positronium $= e^+e^-$.

Photon: The massless quantum of QED.

Pion: The lightest meson composed of u and d quarks, for example, $\pi^+ = (u\bar{d})$

Planck Mass: The mass scale where gravity becomes strong: $M_{PL} = \sqrt{\hbar c / G_N}$ where G_N is Newton's constant.

QCD: Quantum theory of colored quarks and gluons.

QED: Quantum theory of charged fermions and the photon.

Quark: Fermion with color (QCD interactions).

Radiation Length: The characteristic length of material within which an electron interacts by emission of radiation caused by acceleration by a nucleus.

Rapidity: A variable, conventionally denoted by y, that defines longitudinal velocity and is invariant under Lorentz transformations: tanh $y = (p_\parallel / E)$, where E is energy and $p_\parallel$ is momentum along a defined axis. Nonrelativistically $y \to \beta_\parallel = v_\parallel / c$.

Spin: The intrinsic, quantized angular momentum possessed by SM constituents. Units are $\hbar$.

Structure Function: Experimentally determined distribution of momenta of constituents within a composite bound state (e.g., proton).

Uncertainty Principle: The inverse relationship between the possible simultaneous precision of knowledge of momentum and position, $\Delta x \Delta p_x = \hbar$.

Units: The convention $\hbar = c = 1$ is adopted. The dimensions of energy, momentum, 1/length, and 1/time are then all expressible in electron volts (eV). Conversion using $\hbar c = 2 \times 10^{-14}$ GeV cm, $\hbar = 0.66 \times 10^{-25}$ GeV s is straightforward.

Vacuum Polarization: In QED, the observed charge of an electron is modified by the sea of e^+e^- pairs virtually present in the vacuum. Because the neutral sea reduces the effective charge, and the sea extends only over distances $\sim\hbar/m_e c$, the effective charge increases with decreasing distance.

Wire Chamber: A detector with an array of wires maintained at high electric field that localizes, amplifies, and detects the ionization deposited by charged particles traversing a volume of gas.

***W, Z*:** The massive quanta of the weak force.

Yukawa Coupling: The coupling of fermions to massive quanta analogous to EM coupling. Massive quanta give rise to short-ranged forces.

References

Aharmim, B. *et al.* (**2005**) nucl-ex/0502021.

Arnison, G. *et al.* (**1983**) *Phys. Lett.*, **122B**, 103–16.

Banner, M. *et al.* (**1983**) *Phys. Lett.*, **122B**, 322–28.

Burkhardt, H. and Steinberger, J. (**1991**) *Annu. Rev. Nucl. Part. Sci.*, **41**, 55–96.

Cabibbo, N. (**1963**) *Phys. Rev. Lett.*, **10**, 531–33.

CDF Collaboration (**1995**) *Phys. Rev. Lett.*, **74**, 2626–31.

Cronin, J.W. (**1981**) *Rev. Mod. Phys.*, **53**, 373–83.

Davis, R. *et al.* (**1989**) *Proceedings of the 13th International Conference on Neutrino Physics and Astrophysics* (eds J. Schnepps, T. Kafka and W.A. Mann), World Scientific, Singapore.

Dawson, S. and Oreglia, M. (**2004**) *Ann. Rev. Nucl. Part. Prop.*, **54**, 269.

D0 Collaboration (**1995**) *Phys. Rev. Lett.*, **74**, 2632–37.

Eguchi, K. *et al.* (**2003**) *Phys. Rev. Lett.*, **90**, 021892.

Fich, M. and Tremaine, S. (**1991**) *Annu. Rev. Astron. Astrophys.*, **29**, 409–45.

Fitch, V.L. (**1981**) *Rev. Mod. Phys.*, **53**, 367–71.

Friedman, J.I. and Kendall, H.W. (**1972**) *Annu. Rev. Nucl. Sci.*, **22**, 203–54.

Gell-Mann, M. (**1964**) *Phys. Lett.*, **8**, 214–15.

Georgi, H. and Glashow, S.L. (**1974**) *Phys. Rev. Lett.*, **32**, 438–41.

Glashow, S.L. (**1961**) *Nucl. Phys.*, **22**, 579.

Hosaka, J. *et al.* (**2005**) hep-ex/0508053.

Pati, J.C. and Salam, A. (**1973**) *Phys. Rev. D*, **8**, 1240–51.

Prescott, C. *et al.* (**1978**) *Phys. Lett. B*, **77**, 347–52.

Primakoff, H. and Rosen, S.P. (**1981**) *Annu. Rev. Nucl. Part. Sci.*, **31**, 145–92.

(a) Sakharov, A.D. (**1967**) *Zh. Eksp. Teor. Fiz. Pis'ma Red.*, **5**, 31–35; (b) Sakharov, A.D. (**1967**) *JETP Lett.*, **5**, 24–30.

Salam, A. (**1968**) *Elementary Particle Theory*, Nobel Symposium No. 8 (ed. N. Svartholm), Almqvist and Wiksells, Stockholm.

Steigman, G., Schramm, D.N., and Gunn, J. (**1977**) *Phys. Lett. B*, **66**, 202–4.

Utiyama, R. (**1956**) *Phys. Rev.*, **101**, 1597–607.

Weinberg, S. (**1967**) *Phys. Rev. Lett.*, **19**, 1264–66.

Weyl, H. (**1929**) *Z. Phys.*, **56**, 330–52.

Zweig, G. (**1964**) CERN Report 8182/TH.401.

Further Reading

The Standard Model – Synthesis and Overview

Aitchison, I.J.R. and Hey, A.J.G. (**1989**) *Gauge Theories in Particle Physics*, 2nd edn, Adam Hilger, Philadelphia.

Misner, C., Thorne, K., and Wheeler, J. (**1973**) *Gravitation*, W.H. Freeman and Co, San Francisco.

Okun, L.B. (**1984**) *Particle Physics, The Quest for the Substance of Substance*, Harwood, Chur.

Perkins, D. (**1987**) *Introduction to High Energy Physics*, Addison-Wesley, Redwood City.

Quigg, C. (**1983**) *Gauge Theories of the Strong, Weak, and Electromagnetic Interactions*, Benjamin Company, Reading.

http://www.fnal.gov/pub/forphysicists/hepapbook/index.html contains *The Quantum Universe* and *Discovering the Quantum Universe* along with other reference texts that are downloadable. source: *http://www.interactions.org/quantumuniverse/*

Accelerators and Other Facilities

(**2006**) Review of particle physics, *J. Phys. G.*, **33** *http://pdg.lbl.gov/2008/html/authors_2008.html*.

Mapping the SM Constituents onto the Detectors

Barger, V.D. and Phillips, R.J.N. (**1987**) *Collider Physics*, Addison-Wesley, Redwood City.

Eichten, E., Hinchliffe, I., Lane, K., and Quigg, C. (**1984**) Supercollider physics. *Rev. Mod. Phys.*, **56**, 579–707.

Shapiro, M. and Siegrist, J. (**1991**) Hadron collider physics. *Annu. Rev. Nucl. Part. Sci.*, **41**, 97–132.

A Model Collider Detector

Ferbel, T. (**1987**) *Experimental Techniques in High Energy Physics*, Addison-Wesley, Redwood City.

Froidevaux, D. and Sphicas, P. (**2006**) *Ann. Rev. Nucl. Part. Sci.*, **56**, 375.

Kleinknecht, K. (**1985**) *Detectors for Particle Radiation*, Cambridge Univ. Press, Cambridge.

The model detector is described in detail in: *http://www.uscms.org/SoftwareComputing/ProjectManagement/DOE_NSFReviews/2006-02/docs/underptdr-2.pdf*

Other Experimental Techniques

Jackson, J.D. (**1962**) *Classical Electrodynamics*, Wiley-VCH Verlag GmbH, New York.

Future Developments

Dawson, S. and Oreglia, M. (**2004**) *Ann. Rev. Nucl. Part. Sci.*, **54**, 269.

Gunion, J., Haber, H., Kane, G., and Dawson, S. (**1990**) *The Higgs Hunter's Guide*, Addison-Wesley, Redwood City.

Hocker, A. and Ligeti, Z. (**2006**) *Ann. Rev. Nucl. Part. Sci.*, **56**, 501.

Ross, G.G. (**1987**) *Grand Unified Theories*, Cambridge Univ. Press, Cambridge.

Mohapatra, R.N. and Smirnov, A.Y. (**2006**) *Ann. Rev. Nucl. Part. Sci.*, **56**, 569.

PARTICLE SOURCES—ION

IAN G. BROWN, *Lawrence Berkeley Laboratory, University of California, Berkeley, California, U.S.A.*

- **Introduction** ... 213
- 1. **Plasma Formation** ... 213
- 2. **Beam Formation and Related Issues** ... 214
- 3. **Some Representative Ion Sources** ... 217
- 4. **Applications** ... 219
- 5. **Summary** ... 220
- **Glossary** ... 220
- **Works Cited** ... 221
- **Further Reading** ... 221

INTRODUCTION

Ions are formed from neutral atoms by one kind or another of ionization mechanism. The various ways in which the ionization can be done include electron-impact ionization, photoionization, field ionization, surface ionization, and others. When incorporated within an ion source, often the ionization method lends its name to the ion source; thus for example one can have microwave ion sources, field-emission ion sources, vacuum-arc ion sources, and so on. Most ion sources are plasma based. Plasma might be formed from a gas (gaseous plasma) as in microwave, Penning, beam-plasma, and duoplasmatron ion sources, or from a metal (metal plasma) as in vacuum-arc and some laser ion sources. In some ion sources, the ions are formed without going through the plasma state, as in the field-emission or liquid-metal ion source. Because of the predominance of ion sources in which the ions are formed via the plasma state, here we restrict consideration to this kind of source.

Plasma-based ion sources are comprised of two basic parts—a means of plasma production (the "plasma source"), and a means of beam formation (the "extractor"). A simplified schematic of an elementary generic ion source is shown in Fig. 1. The requirements of the plasma source and the extractor may be either trivial or very complicated depending on the required ion-source performance and ion-beam characteristics. For example, the production of a low-energy, low-current (say a few μA at a few keV), singly charged argon-ion beam with no particular demands on beam geometry or divergence is rather easily accomplished; but the generation of a beam of very highly charged ions of low divergence at high energy and beam current is quite a different story.

Note the terminology used. An ion source is a device that produces an ion beam, and it (usually) contains a plasma source. Even though the plasma indeed contains ions and so the plasma source is in a sense a source of ions too, it is more appropriate to refer to the device that produces an extracted beam of more or less energetic ions, i.e., the ion-beam generator, as an ion source, and to use the term plasma source for the generator of charge-neutral plasma.

Here we briefly review the general characteristics of plasma-based ion sources and examine a few specific kinds of sources that have become well known and that are used extensively, including both plasma-based sources as well as one non-plasma-based source, the liquid-metal ion source.

1. PLASMA FORMATION

The plasma formation process is the heart of the ion source. The properties of the ion source and of the extracted ion beam are determined in large part by the plasma source, and the name given to the ion source is often that of the kind of plasma generator used. Some of the most basic ion-beam characteristics, such as ion species, beam current, beam purity, and ion charge state are determined by the equivalent parameters of the plasma

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

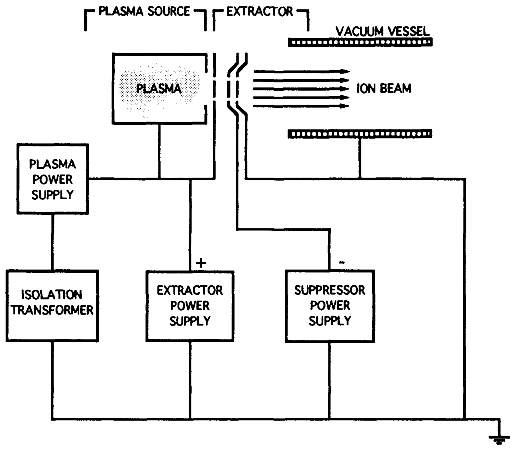

FIG. 1. Simplified schematic of a "generic" ion source.

source—plasma ion species, plasma density, plasma composition, and plasma ion charge-state distribution. For example, in an ECR ion source, the plasma is formed in an ECR (electron cyclotron resonance) plasma discharge and the ion-beam characteristics reflect the properties of the ions that are produced in an ECR plasma device; most notably, they are highly charged. If a high-current ion source is required, the plasma source should be of high density and high ion-production rate. As for all plasma sources, the means by which the plasma is confined is vitally important, and this is no less so when embodied within an ion source. Magnetic field configurations of various kinds have been employed using both electromagnetic field coils and high-field, rare-earth permanent magnets. The "magnetic bucket" geometries that have been used so successfully for the production of large, uniform plasmas have also been used in ion sources for the production of large-area ("broad") ion beams. Such ion sources are often referred to as bucket sources.

Whereas a plasma source as commonly used in plasma physics research is basically at ground potential, in an ion source the plasma and its electrical system are necessarily elevated to extractor potential (see Fig. 1), which might be many tens of kilovolts. This requirement adds complexity and expense to the ion-source configuration. Simple methods of plasma preparation have an advantage in this respect. For high-power dc ion sources, the requirements of the floating power supplies and their control and protection electronics can be a major part of the overall system hardware and cost.

2. BEAM FORMATION AND RELATED ISSUES

The plasma in an ion source is more or less confined within the region in which it is generated except for the place where the beam is to be formed, where the beam-formation electrodes are located. The plasma ion and electron components are separated at the beam-formation electrodes, the ions are accelerated to the required beam energy, and the ion beam is formed. Note that the beam is not really "extracted" from the plasma. What takes place is that the plasma ions drift across a Child–Langmuir sheath at a rate determined by the plasma density and the electron temperature, and not at all by the applied "extraction voltage"; it is only consequent to the ion–electron separation at the sheath that the ions experience the full applied voltage and are thereby accelerated to the requisite beam energy. Thus, the terms "ion extractor," "beam extraction electrodes," etc., are in fact misleading in terms of forming a good conceptual model of the ion source. Nevertheless the terms are used widely, and with the qualifications presented we use this terminology here too.

The ion-beam current that can be extracted from an ion source is given by the Child–Langmuir equation for ion current flow under space-charge–limited conditions (Child, 1911; Langmuir and Compton, 1931; Green, 1974, 1976),

$$I = \tfrac{4}{9}\varepsilon_0 S \left(\frac{2q_i}{M_i}\right)^{1/2} \frac{V^{3/2}}{d^2} \qquad (1)$$

$$= 1.72S \left(\frac{Z}{A}\right)^{1/2} \frac{V^{3/2}}{d^2}, \qquad (2)$$

where ε_0 is the permittivity of free space, S the open extractor area, $q_i = eZ$ the ion charge, e the electronic charge, Z the ion charge state, $M_i = Am_{\text{amu}}$ the ion mass, A the atomic weight (amu), V the extractor voltage, and d the extractor gap (separation between the main beam-forming electrodes), and in the second expression I is in mA, V in kV, S in cm^2, and d in cm. This equation is strictly valid only under special conditions, the most important of which are

1. the assumption of space-charge–limited ion flow, and
2. plane-parallel geometry throughout.

Although condition **1** is often well satisfied, condition **2** is often only poorly met. But nevertheless, the Child–Langmuir equation in its simple form usually provides a good approximation to the ion current that can be expected.

The beam-formation electrodes, or extractor grids, are in the form of aligned apertures in metallic electrodes across which the ion acceleration voltage, or extraction voltage, is applied. A pair of single, circular holes might constitute the extractor, as for example is commonly used in ECR ion sources, or an array of a large number, perhaps thousands, of small holes each producing its own "beamlet," as for example in some rf ion sources, "Kaufman" ion sources, and sources used for space thrusters, or an array of parallel slits as for example in some of the broad-beam ion sources developed for the formation of large, high-power neutral beams used in the controlled thermonuclear fusion research program. A photograph of the large and high-precision extractor used for the 100-keV, 75-A neutral beam source developed at the Lawrence Berkeley Laboratory (Kunkel, 1990) is shown in Fig. 2.

To maximize the beam current produced by the source, the extractor gap needs to be as small as it can be for a given extraction voltage while not allowing breakdown between the grids under a wide range of plasma conditions. Designing for these conditions is where the art and skill come in. The Child–Langmuir equation provides a good estimate of the beam current that can be expected from an ion source when the plasma density and the extraction voltage are optimized with respect to the extraction geometry (also called the perveance match condition). When the mechanical geometry and the plasma and operational parameters are matched to each other in this way, the beam divergence is minimum, determined now by the ion temperature and not the extraction optics, and the beam current density is highest. Calculation of beam formation is a complicated process, and a number of computer codes have been written to predict the extracted ion trajectories (Spädtke, 1989). An example of such a calculation is shown in Fig. 3 (taken from Spädtke, *et. al.*, 1994).

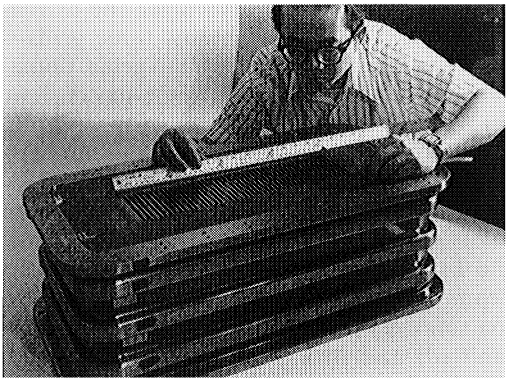

FIG. 2. Beam formation electrodes from a 10 x 40-cm² long-pulse ion source showing the multiple-slot, rail-type plasma grid facing upward.

In the simplest scenario, the ion-beam extractor can be made from just a pair of grids. These are often called the plasma grid and the ground grid, where the plasma grid is in contact with the plasma and the ground grid is held at ground potential. The extraction voltage, typically some tens of kilovolts, is put across the extractor, which means that the whole plasma formation system is now elevated to high positive voltage—ions created in the plasma are accelerated across this high extraction voltage; this is where the beam energy comes from.

It is usual that the ion beam is injected into and transported through a vacuum system in which ionization of the residual gas by the ion beam can occur. The cold electrons so created serve an important purpose, especially for high-current ion beams: they provide neutralization of the positive space-charge potential of the beam due to the ions themselves. In this way, space-charge blowup of the beam is minimized. The tendency of an ion beam to space-charge blowup increases with ion mass and beam current density. An expression has been given by Holmes (1989) for the current density for which the beam space charge of an incompletely neutralized beam would cause a 10% increase in beam diameter, which for the nonrelativistic limit can be written as

$$j_0 = \frac{3.24 \times 10^{-7}\, U^{3/2}\, Z^{1/2}}{A^{1/2}\, z^2\, (1 - h)}, \tag{3}$$

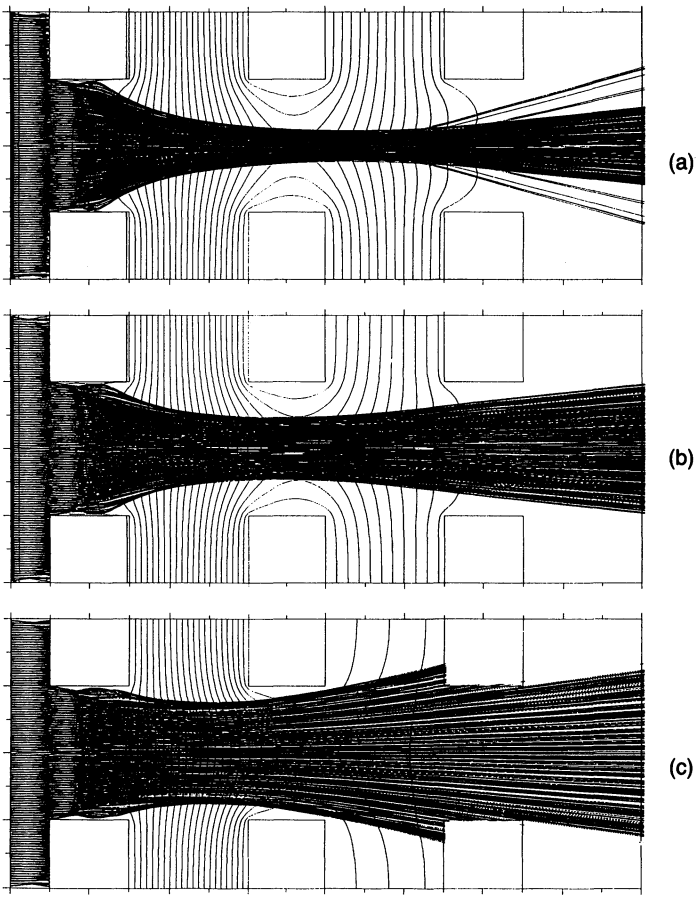

FIG. 3. Ion trajectory plots for extraction from a single hole as calculated by the AXCEL-INP code. Net extractor voltage 5 kV; Ti ions. Suppressor (middle) grid voltage (a) −9 kV, (b) −5 kV, (c) −1 kV.

where j_0 is the beam current density in A/m^2, U is the beam energy expressed in volts, Z is the ion charge state, A is the ion mass in amu, z is the beam propagation distance in meters, and h is the fraction of the beam space charge that is neutralized.

The space-charge–neutralizing electrons form a background sea for the energetic ion beam. The cold electrons can stream backward to the ion-source extractor, where they see an accelerating field in the direction opposite to the extracted ions, i.e., back into the plasma chamber. This electron backflow can sometimes cause damage to the source because of its power deposition, and in any case it shows up as an added electrical load—the measured "extractor current" (power supply drain) is composed not only of the forward ion current but also the backward electron current, and the electron current can easily be several times greater than the true ion current. For these reasons, it is common to make the extractor with a third electrode between the plasma grid and the ground grid, usually called the "suppressor grid," since its function is to suppress the electron backflow. It is biased negatively, say at about 10% or so of the full extraction voltage. The extraction shown in Fig. 3 contains three grids—plasma, suppressor, and ground grids. Sometimes, when design of the beam-ion current flow is particularly critical, yet more grids are added so as to be able to fine-tune the shape of the electric field within the extractor. A cutaway view of an extractor consisting of four rail grids is shown in Fig. 4 (from the Lawrence Berkeley Laboratory neutral-beam development program). In low-current ion sources the suppressor electrode is often omitted; and in ion sources where the extraction is across a magnetic field, the magnetic field serves to suppress electron backflow and so an additional electrostatic suppression is not needed.

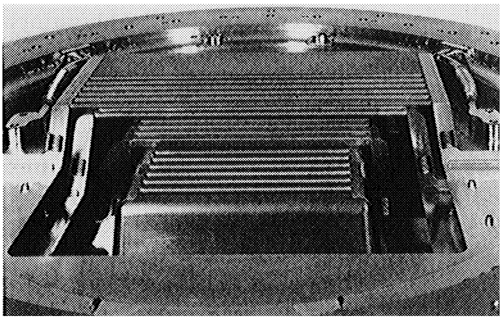

FIG. 4. Cut-away view of a four-grid multiple-slot extractor.

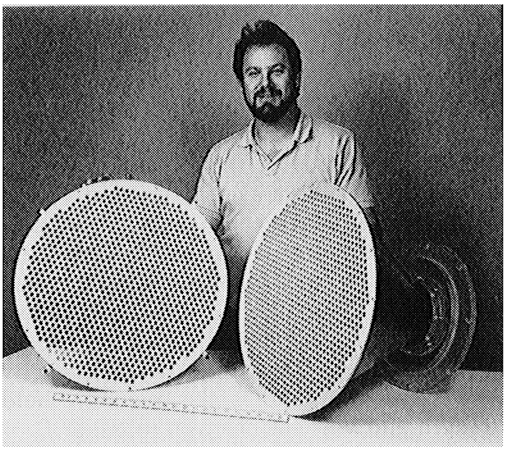

FIG. 5. 50-cm-diam multiple-aperture extractor from a vacuum-arc ion source.

The electron backflow can also give rise to quite substantial x-ray generation when the electron stream is stopped on electrodes within the ion source. This can pose a significant personnel safety hazard. Proper operation of the suppressor can vastly reduce the x-ray intensity. In any case, whenever an ion source is operated at voltages near or above about 20 kV, it is important to monitor carefully the x-ray intensity in the vicinity. Lead shielding is commonly used to reduce the intensity to acceptably low levels.

An ion beam may be focused to submicron dimensions or it may be of very wide area. Ion sources producing small beams often do so by forming the ions at a tiny point, as for example in the liquid-metal ion source where ions are produced by field emission from the pointlike tip of the "Taylor cone" formed electrostatically in the liquid metal. Production of large-area beams is done by ion sources in which the plasma is produced or expanded to large area before being presented to the extractor. The mechanical design of broad-beam extractors is not straightforward, especially when the electrodes can be heated by the beam and suffer thermal expansion while good alignment between the grids must be maintained. The 50-cm extractor from a broad-beam vacuum-arc ion source is shown in Fig. 5. This source produced a pulsed titanium ion beam of energy 100 keV and pulsed current about 10 A (Brown, 1994).

3. SOME REPRESENTATIVE ION SOURCES

Ion sources can be classified in a great many different ways, corresponding to a wide range of source and beam characteristics, parameters, and applications. One can categorize according to beam current intensity, ion charge state, or ion species, or according to whether the source is short pulse, long pulse, or dc, etc. Here we briefly describe several different kinds of ion sources that are broadly used in a number of important fields. These sources are for the most part "mature" in that within their field of use they have been studied and developed by a relatively large number of workers at numerous laboratories around the world. They convey some of the important principles involved in the physics and technology of ion sources. These sources as well as a wide range of other kinds of ion sources also are well described in the Proceedings of the International Conferences on Ion Sources (see Further Reading).

The ECR (electron cyclotron resonance) ion source (Jongen and Lyneis, 1989) has achieved a high degree of prominence because of the very high charge states of the ions produced—species such as fully stripped argon, Ar^{18+}, for example, have been produced (Kutner, 1994). These sources are widely used for accelerator (cyclotron) injection and for atomic physics experiments. They are an outgrowth of the hot-electron plasma source research carried out as part of the world controlled-fusion research effort in the 1960s and 1970s. The early pioneering work to develop these kinds of plasma sources into usable ion sources was led by Geller and his group at Grenoble (Geller, 1976). In the ECR source, a hot-electron plasma is created by the resonant coupling of high-power microwaves to the plasma electrons trapped within a minimum-*B* stabilized magnetic mirror. The electrons can be heated to very high energies, up to hundreds of keV, and it is these hot electrons that provide the ionization energy to strip the plasma ions, also trapped within the magnetic mirror confinement geometry. A small fraction of ions is axially lost from the mirror trap, and it is this plasma loss flux from which the ion beam is then extracted, providing a beam of very highly stripped ions. The electron cyclotron frequency is given by $\omega_c = eB/m$, where B is the magnetic field strength and m is the electron mass, or $f_{ce} = 2.8B$ GHz, where now the field-strength B is in kG; for typical laboratory magnetic fields of, say, 3–10 kG, this means that the resonant frequency is in the low microwave band, around 10 GHz. A schematic of a typical ECR ion source is shown in Fig. 6. The ECR source is a dc source, producing stable beams for long periods of time (up to weeks). Beams of virtually all of the elements of the periodic table can be produced, including the solid elements either by using an oven or by plasma sputtering from an inserted target. Beam currents are typically in the 1- to 100-eμA (electrical microampere) range for a given charge-state component, dropping to fractional microamperes for the highest charge states.

Freeman sources are used extensively in the semiconductor ion implantation field. Their ion beams are particularly quiescent and stable, and they work well with B, P, and As, the ion species most commonly used for semiconductor doping. The source was initially developed at the Harwell Laboratory, U.K., for electromagnetic isotope separation application (Freeman, 1963), and was subsequently adapted for ion implantation when the need arose in the late 1960s. In the Freeman source, electron emission from a thick (several millimeter diameter) filament is used to control the ionization of material admitted to an arc chamber and the plasma so formed is magnetically confined by the field that is the sum of an externally applied homogeneous field parallel to the filament and that due to the filament itself. It is usual for the arc chamber to be fed by vapor from an oven, providing the feed material containing the wanted ion species. Ion extraction is through a long slit, typically several millimeters wide by several centimeters long, in the same direction as the filament and the external field. The extracted beam is thus of rectangular cross section. Typical ion beam current is up to the 10-mA range.

The EBIS, or electron beam ion source, holds the record for the highest charge-state production. These sources were developed initially by Donets and co-workers at Dubna (Donets, 1990). In the EBIS, a high-energy electron beam is injected into a strong, uniform magnetic field so as to compress the beam to high current density. The beam is surrounded by a number of cylindrical "drift tubes" that are electrically biased from an external source. The drift-tube structure is cryogenically cooled to liquid helium temperature so as to keep the background gas pressure very low, in the 10^{-12}-Torr region or less. Neutral gas is injected into the electron beam where a fraction is ionized and trapped

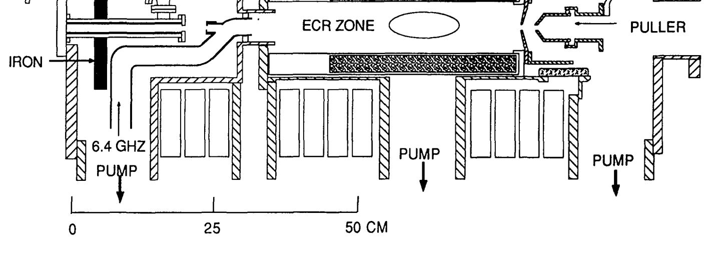

FIG. 6. Simplified schematic of a typical ECR ion source.

within the beam, where the ions are stripped to progressively higher charge states by successive electron impact. After maximal ion charge state is achieved, the drift-tube potentials are switched so as to allow the confined ions to be axially extracted from the device. Fully stripped xenon ions, Xe^{54+}, have been detected (Kutner, 1994).

4. APPLICATIONS

Historically, the first major application of ion sources was for injection into particle accelerators, initially for the relatively simple high-voltage accelerators and later for the growing spectrum of linear and circular machines. To match the capabilities of the accelerators the ion-source parameters required were in general quite modest, for example calling for beam currents only in the microampere range. As the diversity and performance of particle accelerators expanded, so also have the ion-source requirements grown, and sources used in this field today include Penning sources for linear accelerators, ECR sources for cyclotrons, EBIS and laser sources for synchrotrons, and many others. The injection requirements of modern accelerators can be quite severe, calling for very stable and reliable ion beams, often of high charge state and high current. The state of the art of ion sources as relevant to the field of particle accelerators is well presented at regular meetings of the community and can be found in the conference proceedings (see Further Reading).

Electromagnetic isotope separation (the deflection and collection of an ion beam in a magnetic field), in small quantities as an analytical tool and in larger quantities for their production for medical and industrial applications, is a field that also made early use of developments in ion source technology. By far the largest project of this kind was that undertaken in the United States as part of the Manhattan Project in World War II for the separation of uranium isotopes; thereafter these facilities, called calutrons, were used for industrial isotope production. The ion sources first employed were essentially vapor sources with electron-beam ionization. A drawback of the method is the very large investment of electric power required for other than minute quantities. For atomic mass analysis as a research tool, however, the technique is well developed, and the kinds of ion sources employed can be of very low current. The conference series, Electromagnetic Isotope Separators and Techniques Related to their Application, reports well on ion-source developments in this field (see Further Reading).

The field of ion implantation has grown steadily from its introduction in the early 1960s to its wide acceptance as a mature industrial technology for semiconductor material fabrication today. Ion implanters are manufactured and sold by a number of companies throughout the world, and this industrial research community is a well established one. The kind of ion sources used in "mainstream" implanters are the Freeman source and the Bernas source, where the qualities of beam quiescence and stability are paramount; other kinds of ion sources, such as microwave-driven sources, have found niche application. As described above, the beam species of particular application are B, P, and As, and a great deal of development of these sources has been done. The semiconductor ion-implantation community holds a regular biannual conference at which ion-source technology is a significant component; the proceedings of these meetings provide a wealth of detail about ion-source developments in this field (see Further Reading).

Ion implantation is a surface-modification technique with scientific and technological application wider than for semiconductor materials processing alone, and there are active research communities in nonsemiconductor areas such as, for example, metallurgical, ceramic, polymer, superconductor, and electro-optics materials synthesis and modification. For these applications the characteristics of the ion beams required vary widely and so also the kinds of ion sources needed. The field remains largely a research field, having not yet made the transition to widely used industrial application, and the diversity of the experimental configurations investigated and of the ion sources used is great. Relevant major conference series that address ion sources for these applications are those on Ion Beam Modification of Materials and the Surface Modification of Metals by Ion Beams (see Further Reading).

Applications such as ion-beam lithography, micromachining, and materials surface characterization using ion beams call for low-current, finely focused beams, and field-emission liquid-metal ion sources are often used for this purpose. Ion-beam focal spots as small as a few hundred angstroms have been produced.

The quest for propulsion of space vehicles and satellites by ion thrusters has motivated research internationally since the beginning of the space era. Though the size of this effort has decreased with time, work still continues, with emphasis on guidance and steering. A number of different kinds of ion sources have been developed and investigated for this potential application, including rf-driven sources, broad-beam sources, and liquid-metal sources. The requirements of the ion source needed are clear; not surprisingly, a major hurdle is in obtaining the specified ion-beam performance in an adequately small package with adequately high efficiency.

5. SUMMARY

The ion source is the beginning of the ion process, whatever the application is to be. The ion beam is created at the source, transported over distances from a meter or so up to many hundreds of meters, and delivered to the intended device or use. The physics and technology of beam transport is a sibling field to that of ion sources, which attains increasing importance as the required beam transport efficiency and the beam current density increase. The concern of space-charge blowup of the beam due to uncompensated electrostatic repulsive forces between ions is often critical, and as available ion-source beam currents steadily increase, this is becoming more commonly the limiting feature in some ion-beam facilities. In the absence of externally applied electric or magnetic fields, the beam can acquire a high degree of space-charge neutralization from the residual gas in the vacuum system, but the fields that are inherent to beam focusing and bending elements, such as dipole and quadrupole magnets or electrostatic lenses, disrupt the low-energy neutralizing electrons and beam neutralization can be diminished or lost. This concern is more or less limited to the high-current, low-energy regions that can exist in the beam transport line immediately after the ion source. It is usual that as the beam is transported further downstream toward the application, the current becomes lower and often some postacceleration (ion acceleration after initial ion extraction) adds energy to the ion beam; consequently beam space-charge blowup is less of a concern further away from the ion source. Often the ion source provides much more beam, perhaps several orders of magnitude more, than is finally available for the application. Beam loss is a major concern in many cases.

Understanding of the physical principles of ion-source design and operation has increased greatly in recent years, and the technology of their construction and application has similarly flourished. Fields of use both in fundamental science and in applied technology call for increasingly formidable ion-beam parameters and sophisticated ion-source performance, and the field has responded by generating a wide array of ion devices. Whereas in the past it was the case that the field was segmented into distinct and largely noncommunicating application areas, there is now a thriving international community of ion-source researchers with good cross-specialty communication links. The result is a healthy and fertile research area and a steady growth in ion-source technology.

GLOSSARY

ECR: Electron cyclotron resonance. Condition when an applied rf field equals the gyro frequency of electrons in the confining magnetic field. Usually in the microwave band.

Extractor: The electrode structure that forms an energetic, directed ion beam from a plasma.

Magnetic Bucket: Magnetic field geometry formed when an array of permanent bar magnets is distributed around the plasma confinement region.

Magnetic Mirror: Magnetic field geometry formed between two spaced field coils, with maxima in the plane of the coils and a central minimum.

Sheath: Boundary layer between plasma and electrode, where charge neutrality is not preserved.

Taylor Cone: Conical surface shape assumed by a liquid subjected to an electric field.

Works Cited

Brown, I. G. (1994), *Rev. Sci. Instrum.* **65**, 3061.

Child, C. D. (1911), *Phys. Rev.* (*Series 1*) **32**, 492.

Donets, E. D. (1990), *Rev. Sci. Instrum.* **61**, 225.

Freeman, J. H. (1963), *Nucl. Instrum. Methods* **22**, 30.

Geller, R. (1976), *IEEE Trans. Nucl. Sci.* **NS-23**, 904.

Green, T. S. (1974), *Rep. Progr. Phys.* **37**, 1257.

Green, T. S. (1976), *IEEE Trans. Nucl. Sci.* **NS-23**, 918.

Holmes, A. J. T. (1989), in: I. G. Brown (Ed.), *The Physics and Technology of Ion Sources*, New York: Wiley, p. 53.

Jongen, Y., Lyneis, C. M. (1989), in: I. G. Brown (Ed.), *The Physics and Technology of Ion Sources*, New York: Wiley, p. 207.

Kunkel, W. B. (1990), *Rev. Sci. Instrum.* **61**, 354.

Kutner, V. B. (1994), *Rev. Sci. Instrum.* **65**, 1039.

Langmuir, I., Compton, K. T. (1931), *Rev. Mod. Phys.* **3**, 251.

Spädtke, P. (1989), in: I. G. Brown (Ed.), *The Physics and Technology of Ion Sources*, New York: Wiley, p. 107. See also the codes AXCEL-INP and KOBRA-INP: INP, Junkernstr. 99, Wiesbaden D-65205, Germany.

Spädtke, P. *et. al.* (1994), *Rev. Sci. Instrum.* **65**, 1441.

Further Reading

Brewer, G. R. (1970), *Ion Propulsion Technology and Applications*, New York: Gordon and Breach.

Brown, I. G. (Ed.) (1989), *The Physics and Technology of Ion Sources*, New York: Wiley.

Forrester, A. T. (1988), *Large Ion Beams*, New York: Wiley.

Valyi, L. (1977), *Atom and Ion Sources*, New York: Wiley.

Wilson, R. G., Brewer, G. R. (1973), *Ion Beams with Applications to Ion Implantation*, New York: Wiley.

Conference Proceedings

The Proceedings of the biannual International Conferences on Ion Sources (ICIS) are excellent sources of information on ion sources of all kinds. See, for instance: *Rev. Sci. Instrum.* **61** (No. 1, Pt. 2), January 1990, (ICIS '89); *Rev. Sci. Instrum.* **63** (No. 4, Pt. 2), April 1992 (ICIS '91); *Rev. Sci. Instrum.* **65** (No. 4, Pt. 2), April 1994 (ICIS'93).

For ion sources of relevance to the field of particle accelerators, see the proceedings of the biannual IEEE Particle Accelerator Conferences (published by the IEEE as Conference Proceedings), and of the annual Conferences on the Application of Accelerators in Research and Industry (published in *Nuclear Instruments and Methods*).

Progress in ion sources as used for semiconductor ion implantation is reported at the International Conferences on Ion Implantation Technology, held every 2 years, with the proceedings published in *Nuclear Instruments and Methods*. The International Conference on Ion Beam Modification of Materials is focused more on nonsemiconductor implantation; these proceedings are also published in *Nuclear Instruments and Methods*. Another relevant conference series is the biannual Surface Modification of Metals by Ion Beams, proceedings published in *Surface and Coatings Technology*.

The proceedings of the conference series on Electromagnetic Isotope Separators and Techniques Related to their Application are archived in *Nuclear Instruments and Methods*, where ion-source and ion-beam developments and methods in the field are well covered.

PARTICLES, ELEMENTARY

MICHAEL P. SCHMIDT, *Physics Department, Yale University, New Haven, Connecticut, U.S.A.*

	Introduction	223
1.	**Building Blocks of the Standard Model**	224
1.1	Quarks and Leptons	224
1.2	Bound States of Quarks: Baryons and Mesons	225
2.	**The Forces**	226
2.1	The Electromagnetic Force	226
2.2	The Strong Force: Color	227
2.3	The Weak Force: Radioactive Decay	228
2.3.1	Flavor Conservation	229
2.3.2	Conservation of Baryon and Lepton Number	230
2.3.3	Parity and Charge Conjugation	230
2.3.4	Weak Neutral-Current Interactions and Electroweak Unification	231
3.	**Beyond the Standard Model**	231
4.	**Summary**	232
	Works Cited	232
	Further Reading	232

INTRODUCTION

The term "elementary particles" refers to the entities that are considered to be the fundamental constituents of matter. The definition of elementary depends on the distance scale relevant to the physics of interest. The behavior of bulk matter, under normal terrestial conditions, is largely determined by the properties of atoms, in the form of the chemical elements (the elementary particles of chemistry) and combinations of atoms into molecules. Atoms, with a typical size of 10^{-8} cm, have structure and are comprised of electrons orbiting a nucleus, with a typical size of 10^{-12} cm. Atomic nuclei are known to have structure as well and are built up out of protons and neutrons that are an order of magnitude smaller in size. Finally, protons and neutrons are believed to be made of *quarks*, which, along with electrons, appear to be elementary particles, in the sense that they reveal no internal structure down to distances as small as 10^{-16} cm.

Equally important to an understanding of the structure of matter are the forces of nature. Four forces have been identified, each with its domain of major importance. *Gravity* is the most familiar force, being responsible for the formation of stars and planets and the motion of celestial bodies. The *electromagnetic* force is also familiar and among other things leads to the binding of electrons to nuclei and the structure of atoms. On the subnuclear scale, less familiar forces play a fundamental role. The *strong* nuclear force "glues" quarks together into nucleons (protons and neutrons) and the nucleons together into nuclei. The *weak* nuclear force is responsible for the radioactive transmutation of nucleons and nuclei. At the subnuclear scale, it is also important to understand the *quanta*, also elementary particles, that mediate the fundamental forces. The most familiar example is the photon, the quantum of electromagnetic energy and mediator of the electromagnetic interaction.

Remarkable achievements occurring in the study of subnuclear particle physics over the last several decades have led to the development of what is now known as the Standard Model of elementary particle physics. This model is extremely successful in describing and predicting the relationships and interactions among the observed elementary particles and forces. Most notably, the Standard Model includes the unification of the electromagnetic and weak interactions, a discovery akin to that of Maxwell's unification of electricity and magnetism. The development and verification of the Standard Model is a result of the interplay between experimental and theoretical advances.

On the experimental side, the develop-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ment of high-energy accelerators as well as the sophisticated detectors and techniques used to study these particles has been essential. Progress in the field has largely been set by the development of accelerators capable of producing increasingly energetic and intense beams of particles (electrons, protons, etc.). Very high-energy particles when used as projectiles can serve as probes of subnuclear structures by virtue of their small Compton wavelength, $\lambda = h/mc$ (where h is Planck's constant and m is the particle mass); they can also give rise to very energetic collisions in which the energy released in the collision can materialize as massive particles in accordance with Einstein's equation, $E = Mc^2$, relating mass (M), energy (E), and the speed of light (c). Intense beams of particles are needed for precision measurements and the study of rare phenomena.

On the theoretical side, the development of a relativistic quantum field theory for electrodynamics was crucial. Its success in providing a means to calculate detailed properties of atomic structure led theorists to conclude correctly that there was hope in attempting to formulate the other subnuclear interactions in a similar fashion. The exploitation of symmetries and conservation laws also proved to be most valuable. The plethora of new particles discovered with the advent of particle accelerators was initially bewildering. Fortunately, regularities were uncovered in the systematics of the production and decays of these particles, and these regularities were successfully interpreted as resulting from symmetries (sometimes hidden or partially broken) of the physics underlying the properties of the particles and their interactions.

The development of the Standard Model represents a very important step in achieving a comprehensive understanding of subnuclear physics, but it is not yet the ultimate theory of elementary particles. Continued research will be required in order to answer fundamental questions left open by the Standard Model, such as the origin of the masses of the elementary particles. It is also hoped and suspected that further unification of the fundamental forces should be possible, including gravity.

The field of subnuclear physics, which grew out of the study of nuclear physics and cosmic rays, continues to evolve. For example, the field of particle astrophysics is growing in importance, building on the interplay between particle physics, the physics of the early universe, and cosmology. Collisions occurring at high-energy accelerators can create conditions not experienced since shortly after the Big Bang. In addition, the number and type and properties of the species of particles existing in the early moments of the universe affect its evolution.

1. BUILDING BLOCKS OF THE STANDARD MODEL

The elementary particles of the Standard Model can be divided into two broad classes: the particles that carry the properties (e.g., electric charge) that give rise to the forces of interaction between the particles and the particles that serve to communicate these forces between the particles. The quarks and leptons, discussed in the present section, are in the first class. The gauge bosons that mediate the forces are discussed in the following section.

1.1 Quarks and Leptons

The building blocks of the Standard Model are the *quarks* and *leptons*. Quarks are constituents of protons, neutrons, and related particles (previously considered elementary) known as *hadrons* that feel the strong nuclear force. The leptons form a separate group of particles that are not subject to the strong interactions. Examples of leptons are the electron and the neutrino. The neutrino, being electrically neutral, participates only in the weak interactions. The electron as well as the quarks experiences both the electromagnetic and weak interactions.

The quarks come in six *flavors* named up, down, charm, strange, top, and bottom. Leptons come in three pairs, each pair consisting of a charged lepton and an associated neutrino, and they are named after the charged lepton of the pair: electron, muon, and tau. The electric charges carried by the quarks and leptons are given in Table 1 (in units of the electron's charge). Electric charge is one of several *quantum numbers* associated with the elementary particles. These quantum num-

Table 1. The quarks and leptons[a]

Particle	Symbol	Spin	Baryon number	Charge	Mass (GeV)
Up quark	u	1/2	1/3	+2/3	$\approx 5 \times 10^{-3}$
Down quark	d	1/2	1/3	−1/3	$\approx 10 \times 10^{-3}$
Charm quark	c	1/2	1/3	+2/3	≈1.3
Strange quark	s	1/2	1/3	−1/3	≈0.2
Top quark	t	1/2	1/3	+2/3	≈170
Bottom quark	b	1/2	1/3	−1/3	≈4.3
Electron neutrino	ν_e	1/2	0	0	$<5.1 \times 10^{-9}$
Electron	e	1/2	0	−1	0.511×10^{-3}
Muon neutrino	ν_μ	1/2	0	0	$<0.27 \times 10^{-3}$
Muon	μ	1/2	0	−1	0.106
Tau neutrino	ν_τ	1/2	0	0	<0.031
Tau (lepton)	τ	1/2	0	−1	1.78

[a]Extracted from Particle Data Group (1994).

bers are found to be conserved quantities in the interactions of elementary particles. The determination of the conserved quantities has played an important role in establishing the relationships among the subnuclear particles. The significance of other quantum numbers such as lepton number and baryon number will be discussed later.

Each quark and lepton also has a corresponding *antiparticle* that carries equal but opposite quantum numbers; the antielectron, also known as the *positron*, carries +1 unit of electric charge. Particles and antiparticles have equal masses. In Table 1, the particle masses are expressed in terms of an energy (millions of electron volts) using the relation $E = mc^2$. Sufficiently energetic collisions of elementary particles can create conditions in the laboratory in which particle–antiparticle pairs are created. When this occurs through the electromagnetic or strong interaction, the various quantum numbers, including quark or lepton flavor, associated with the particles must be conserved. For example, a positron must be created in association with an electron so that no net electric charge is created from energy alone. However, as discussed later, the weak interactions do not conserve quark flavor.

All of the quarks and leptons are *fermions*, obey Fermi–Dirac statistics, and carry an intrinsic angular momentum, or spin, of 1/2 (in units of $\hbar$, Planck's constant divided by 2π). Angular momentum as well as energy and linear momentum are conserved in the interactions of elementary particles.

1.2 Bound States of Quarks: Baryons and Mesons

Spin plays an important role in determining the properties of the hadrons that are formed out of combinations of quarks and antiquarks by the strong interaction. As discussed more fully below, the strong interaction binds together groups of three of quarks (denoted as qqq) and quark–antiquark pairs (denoted by $q\bar{q}$). In principle, any flavor can be chosen for each of the quarks. The three-quark bound states qqq are known as *baryons*; they have baryon number +1 (from the three quarks, each contributing +1/3). The rules of addition for spin angular momentum result in baryons having odd half-integral spin (1/2, 3/2, etc.) and therefore being fermions. The proton and neutron are the most familiar baryons—the proton consisting of two up quarks and a down quark ($p = uud$) and the neutron consisting of two down quarks and an up quarks ($n = udd$). The electric charge of the proton and neutron follows from the algebraic sum of the electric charges of the constituent quarks. A partial list of observed baryons is given in Table 2, including their quark content, spin, and mass.

The quark–antiquark bound states, known as *mesons*, have baryon number equal to zero. They also have integral spin (0, 1, etc.) and so are *bosons* and obey Bose–Einstein statistics. Again, in principle, any flavor of quark can be chosen in forming the mesons. Some examples of mesons readily produced in high-energy particle collisions include the pi me-

Table 2. Some representative baryons and mesons[a]

Particle	Symbol	Quark content	Spin	Charge	Mass (GeV)
Proton	p	uud	1/2	+1	0.938
Neutron	n	udd	1/2	0	0.940
Lambda	Λ^0	uds	1/2	0	1.12
Sigma	$\Sigma^+, \Sigma^0, \Sigma^-$	uus, uds, dds	1/2	+1, 0, −1	1.19
Delta	$\Delta^{++}, \Delta^+, \Delta^0, \Delta^-$	uuu, uud, udd, ddd	3/2	+2, +1, 0, −1	1.23
Omega	Ω^-	sss	3/2	−1	1.67
Lambda (*c*)	Λ_c^+	udc	1/2	+1	2.29
Pions	π^+, π^0, π^-	$u\bar{d}, u\bar{u} + d\bar{d}, d\bar{u}$	0	+1, 0, −1	0.137
Kaons	$K^+, K^0, \bar{K}^0, K^-$	$u\bar{s}, d\bar{s}, s\bar{d}, s\bar{u}$	0	+1, 0, 0, −1	0.495
Rho	ρ^+, ρ^0, ρ^-	$u\bar{d}, u\bar{u} + d\bar{d}, d\bar{u}$	1	+1, 0, −1	0.770
Phi	ϕ^0	$s\bar{s}$	1	0	1.02
D mesons	$D^+, D^0, \bar{D}^0, D^-$	$c\bar{d}, c\bar{u}, u\bar{c}, d\bar{c}$	0	+1, 0, 0, −1	1.86
	D_s^+, D_s^-	$c\bar{s}, s\bar{c}$	0	+1, −1	1.97
J (Psi)	J/Ψ	$c\bar{c}$	1	0	3.1
B mesons	$B^+, B^0, \bar{B}^0, B^-$	$u\bar{b}, d\bar{b}, b\bar{d}, b\bar{u}$	0	+1, 0, 0, −1	5.28
	$B_s^0, \bar{B}_s^0$	$s\bar{b}, b\bar{s}$	0	+1, −1	5.37
Upsilon	Υ	$b\bar{b}$	1	0	9.46

[a] Extracted from Particle Data Group (1994).

sons, also known as pions. The pions come in three charged varieties: $\pi^+ = u\bar{d}$, $\pi^- = \bar{u}d$, and π^0, a linear combination of $u\bar{u}$ and $d\bar{d}$. It is noted that the π^- is the antiparticle of the π^+, and the π^0 is its own antiparticle. The *K* mesons, or kaons, are also relatively easily produced, being the least massive particles containing the strange quark: $K^+ = \bar{s}u$, $K^- = s\bar{u}$, $K^0 = \bar{s}d$, and $\bar{K}^0 = s\bar{d}$. Table 2 also gives a partial list of some of the observed mesons and their quark contents, spins, and masses.

2. THE FORCES

The interactions among elementary particles are the result of several forces. Three of these forces—the strong, electromagnetic, and weak—have been understood in terms of relativistic quantum field theories and in particular as *gauge* field theories. Each of the forces is described in terms of the exchange of quanta of the gauge field between the interacting particles. The strength of the gravitational force is sufficiently small that gravity can safely be ignored in studying the interactions of small numbers of elementary particles.

2.1 The Electromagnetic Force

The electromagnetic force is the most familiar of the forces relevant to the study of elementary particles. It is the attractive *Coulomb* force between the negatively charged electron and the positively charged proton that gives rise to the bound state known as the hydrogen atom. Ignoring relativistic effects, the potential energy of the bound state can be expressed as $V = e^2/r$, where e is the magnitude of the electron (or proton) charge and r is the distance separating the proton and electron.

The infinite range of the electromagnetic force is in keeping with the masslessness of the photon, the quantum of the electromagnetic field. The electromagnetic field of a charged particle is taken to be the result of the radiative emission and reabsorption of *virtual* photons by the charged particle. Virtual particles are those that do not satisfy the equation $E^2 - p^2c^2 = m^2c^4$; that is, their existence violates energy conservation for a short time interval, as allowed by the uncertainty principle of quantum mechanics: $\Delta E \Delta t \leq \hbar$.

In first approximation, the Coulomb force responsible for the binding of the hydrogen atom is due to the emission of a photon by the proton and subsequent absorption by the electron. The strength of the interaction due to the exchange of one photon is $\alpha = e^2/\hbar c = 1/137$ when expressed in dimensionless form. Here, α is associated with the quantum mechanical probability of one photon being exchanged between two particles with charge e. The full theory, known as quantum electrodynamics (*q.v.*) or QED, is used to determine the contributions to the force arising from the exchange of multiple photons. Perturbative methods can be employed and lead to converging results as a result of the smallness of α.

Even as it is recognized that the definition of what is an elementary particle depends on the distance scale in question, it is also recognized that the definition of the strength of an interaction depends on the distance scale being probed. The strength of the electromagnetic interaction, α, depends on the electric charge e. However, the magnitude of the effective charge can be shown to increase as the distance associated with the measurement of the charge decreases. This is due to the screening of the electron's charge at longer distances by the polarization of the vacuum surrounding the electron. The electron emits virtual photons to give rise to the electromagnetic field, and these photons can themselves create virtual electron–positron pairs. The electron–positron pairs are affected by the presence of the central electron charge and become oriented so as to have the positrons closer to the central electron. The screening effect is measurable, though small, changing α by about 10% for a change in the distance scale by a factor of roughly 100 000.

Quantum electrodynamics (QED) serves as the paradigm in the construction of theories of the strong and weak interactions. The gauge bosons that are involved in these theories are listed in Table 3. The significant differences observed between the interactions result in large part from the different symmetries inherent to the interactions. The properties of quantum electrodynamics result from a local gauge symmetry based on the unitary group U(1). This symmetry requires the invariance of the physics under a redefinition of the quantum mechanical wave functions of particles carrying electric charge by a space-time–dependent complex phase. The U(1) gauge invariance of QED is responsible for the conservation of electric charge.

2.2 The Strong Force: Color

The strong force, also called the nuclear force, is formulated in terms of a gauge field theory known as quantum chromodynamics (QCD) (see STRONG INTERACTIONS). The gauge symmetry for QCD is based on the special unitary group SU(3). In QED, there is one kind of charge, the electric charge, acting as the source for the electromagnetic field. In QCD, there are three kinds of charge, called color; each flavor of quark (u,d,c,s,t,b) can possess one of the three color charges, and the leptons carry no color charge. The three color charges can be named R, G, and B in analogy with the primary colors red, green, and blue. The properties of the color force are such that only color-charge neutral (i.e., colorless) bound states are observed in nature. This has the unusual consequence that free, isolated colored quarks are not observable in nature, a fact that so far has been borne out experimentally. The only color-neutral objects are the baryons and mesons of which more than a hundred different types have been observed.

The baryons are colorless bound states in which each of the three quarks carries a different color charge to obtain equal contributions of R, G, and B color charge. The color charges are independent of the quark flavor and, for example, the lambda baryon, $\Lambda = uds$ is a linear combination of $u_R d_G s_B$, $u_G d_B s_R$, $u_B d_R s_G$, etc. (where the subscript indicates the color charge of the quark). A linear combination of colors is required to assure that the net color charge of the particle is invariant under rotations in the abstract space associated with the color SU(3) symmetry group.

Antiquarks carry "anti"-color charges $\bar{R}$, $\bar{G}$, $\bar{B}$ (or, to follow the visual color theory

Table 3. The gauge bosons[a]

Particle	Symbol	Spin	Force mediated	Electric charge	Color charge	Mass. (GeV)
Photon	γ	1	Electromagnetic	0	0	0
Gluon	g	1	Strong	0	r, g, b	0
W boson	W^+, W^-	1	Charged weak	±1	0	80.2
Z boson	Z^0	1	Neutral weak	0	0	91.2

[a] Extracted from Particle Data Group (1994).

analogy, the complementary colors cyan, magenta, and yellow), just as the positron carries an electric charge opposite to that of the electron. The mesons are colorless objects by virtue of being comprised of a quark and antiquark having "opposite" colors. The positive pi meson is a linear combination of $u_R\bar{d}_R$, $u_G\bar{d}_G$, and $u_B\bar{d}_B$.

The analogue of the photon in QED for QCD is the gluon. The gluon, like the photon, is massless and has spin 1, and color charge is a conserved quantity. However, because of the *non-Abelian* property of the QCD gauge group, SU(3), there are in fact eight distinct types of gluons, and they all possess net color charge: $R\bar{G}$, $B\bar{R}$, etc. Photons do not carry electric charge and can only interact with each other via charged particles. Gluons, on the other hand, can interact directly with each other via their color charges, and this plays a crucial role in determining the properties of the strong force.

In quantum chromodynamics, the chromodynamic fields of quarks are created by the emission and reabsorption of virtual colored gluons. The vacuum can be polarized, as in the case for QED. However, for QCD, the polarization of the vacuum leads to an antiscreening of color charge, a result due to the fact that gluons carry color charge. The antiscreening implies that the strength of the color force decreases as the distance associated with the measurement decreases, and conversely, the force increases with increasing distance. The former result has been proven (by Gross and Wilczek and Politzer; see Wilczek, 1982) and is known as the *asymptotic freedom* property of QCD. The latter result is conjectured and is known as the *confinement* property of QCD. The phenomenological result (consistent with experimental observations) is that quarks are confined in protons and other hadrons at a distance scale of about 1 fm (10^{-13} cm) where the chromodynamic interaction coupling is strong ($\alpha_s > 1$). However, inside the hadrons, the quarks behave relatively freely.

Unfortunately, perturbative methods cannot be used to determine the masses of hadrons because α_s effective at a distance scale of 1 fm is not small. Nonperturbative techniques, such as lattice gauge theory, show promise for addressing these problems. It is expected and has been observed that the masses of hadrons containing the more massive quarks (*c,b,t*) are largely due to the mass of the massive quark.

The strong force that binds protons and neutrons into nuclei (and overcomes the large repulsive electric force existing between protons inside a nucleus) is then related to the structure of the nucleons (and other hadrons), as determined by the more fundamental chromodynamic interaction of quarks and gluons. This is similar to the relation of the forces at play between neutral atoms and molecules and the more fundamental electromagnetic interaction that establishes atomic structure. The nuclear binding force is strong but has a short range; the nucleons must be close enough to detect the distribution of color charge within each other.

Before the advent of the quark model of hadrons, it was suggested that the nuclear binding force could be due to the exchange of mesons between nucleons (protons and neutrons). The finite range of the force could be explained in terms of a potential of the form $V = (\alpha_s^2\hbar c/r)e^{-mc^2r/\hbar c}$, now known as the Yukawa potential, where m is the mass of the exchanged meson. The Compton wavelength, $\lambda = h/mc$, of pi mesons agrees with the observed effective range of the nuclear force between nucleons.

2.3 The Weak Force: Radioactive Decay

Only a few elementary particles are stable; the rest of the particles rapidly decay into other less massive particles. The study of decays has been an essential aspect of elementary particle research and has provided clues to understanding the relationships and interplay among the forces affecting elementary particles. The exponential decay lifetimes observed in the decays of elementary particles can be understood in terms of the forces responsible for the decay. A typical strong-interaction decay time is 10^{-24} seconds. Electromagnetic decays occur over time intervals in the range of 10^{-16}–10^{-20} seconds. Weak-interaction decay times are found in the range of 10^{-13}–10^{-6} seconds. The special properties of the weak interaction were initially revealed in the study of particle decays. The weak interaction was in fact first recognized as being the agent responsible for the radioactive β decay of nuclei. As an example, in the decay $n \rightarrow pe^- \bar{\nu}_c$, a neutron can be transformed by the weak interaction into the less massive proton and an electron (the β ray) plus an anti–electron neutrino.

As in the case of the electromagnetic and strong interactions, the weak interaction is understood in terms of a gauge field theory. The weak interaction is governed by the group SU(2) and mediated by the exchange of three types of spin-1 gauge bosons between the interacting fermions. Two of the gauge bosons, known as the W^+ and W^-, carry electric charge and mediate the *charged-current* weak interaction responsible for nuclear radioactive decay. The third boson, known as the Z^0, has zero electric charge and is responsible for the weak *neutral-current* interactions. Typical decays are given in Table 4.

2.3.1 Flavor Conservation The electromagnetic and strong interactions conserve quark and lepton flavors. The photon can couple with any charged fermion and conserves the fermion flavor: An electron can radiate a photon, or a very energetic photon can materialize as an electron–positron pair or any other particle–antiparticle pair of quarks or charged leptons. Electric charge as well as quark and lepton flavors must be conserved: A muon cannot just radiate a photon and as a result be transformed into an electron. The gluon can couple to any quark, as they carry color charge (leptons do not): A blue up quark can radiate a blue–anti-red gluon and be transformed into a red up quark, or a very energetic green–antiblue gluon can materialize as a particle–antiparticle pair of quarks [e.g., charm (green) plus anticharm (antiblue)]. Color and quark flavor must be conserved: A bottom quark cannot just radiate a gluon and be transformed as a result into a strange quark.

The "weak" charges carried by the quarks and leptons are such that the charged-current weak interactions lead to specific sets of transmutations that do not conserve quark flavor. The weak interactions view the quarks and leptons as coming in doublets:

$$\begin{pmatrix} u \\ d \end{pmatrix}, \begin{pmatrix} c \\ s \end{pmatrix}, \begin{pmatrix} t \\ b \end{pmatrix}; \begin{pmatrix} \nu_e \\ e \end{pmatrix}, \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}, \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}.$$

Referring to Table 1, it is noted that the elements of each doublet differ in electric charge by one unit. The W bosons carry one unit of electric charge and couple in an equivalent way to each of the weak fermion doublets and thereby give rise to transmutations between the fermions of a weak doublet. For example, a top quark (charge +2/3) can decay into a W^+ boson and a bottom quark (charge −1/3): $t \rightarrow bW^+$; the W^+ can then subsequently decay into a positron and an electron neutrino: $W^+ \rightarrow e^+\nu_e$. The process of neutron β decay, $n \rightarrow pe^-\bar{\nu}_e$, can be described in terms of the decay of one down quark in the neutron into an up quark and a virtual W^- that subsequently materializes as an electron plus an

Table 4. Some representative particle decays[a]

Particle	Decay	Interaction	Lifetime (s)
Photon	—	—	Stable
Electron	—	—	Stable
Neutrinos	—	—	Stable
Muon	$\mu^- \rightarrow e^-\bar{\nu}_e\nu_\mu$	Weak	2.2×10^{-6}
Tau	$\tau^- \rightarrow \mu^-\bar{\nu}_\mu\nu_\tau$	Weak	3×10^{-13}
Pion	$\pi^- \rightarrow \mu^-\bar{\nu}_\mu$	Weak	2.6×10^{-8}
Pion	$\pi^0 \rightarrow \gamma\gamma$	Electromagnetic	8.4×10^{-17}
Kaon	$K^+ \rightarrow \pi^+\pi^0$	Weak	1.2×10^{-8}
Rho	$\rho^0 \rightarrow \pi^+\pi^-$	Strong	4×10^{-24}
Neutron	$n \rightarrow pe^-\nu_e$	Weak	887
Lambda	$\Lambda^0 \rightarrow p\pi^-$	Weak	2.6×10^{-10}
Sigma	$\Sigma^0 \rightarrow \Lambda^0\gamma$	Electromagnetic	7.4×10^{-20}
Delta	$\Delta^+ + \rightarrow p\pi^+$	Strong	5×10^{-24}
Omega	$\Omega^- \rightarrow \Lambda^0K^-$	Weak	8.2×10^{-11}
D meson	$D^0 \rightarrow K^-\mu^+\nu_\mu$	Weak	4.2×10^{-13}
Lambda − c	$\Lambda_c^+ \rightarrow pK^-\pi^+$	Weak	2×10^{-13}
B meson	$B^- \rightarrow D^0e^-\bar{\nu}_e$	Weak	1.5×10^{-12}

---

[a] Extracted from Particle Data Group (1994).

electron antineutrino: $n \rightarrow (ud)d \rightarrow (ud)u + W^- \rightarrow pe^-\bar{\nu}_e$.

The weak interaction, like the strong interaction, and in contrast to the electromagnetic interaction, is very short ranged. The effective strength of the weak interaction is smaller than the electromagnetic interaction by a factor of about 10 000. Both of these properties result from the fact that the *W* (and *Z*) are very massive particles. The intrinsic strength of the weak interaction is in fact comparable to the electromagnetic interaction after one accounts for the suppressed probability of creating a virtual *W* (or *Z*) compared to that for creating a virtual photon.

2.3.2 Conservation of Baryon and Lepton Numbers The weak interactions (as well as the electromagnetic and strong interactions) conserve both the baryon and lepton numbers. In fact, separate lepton numbers are conserved, one for each of the three flavors of leptons: electron, muon, and tau. The electron and its companion neutrino (ν_e) each carry one unit of electron lepton number; the positron and electron antineutrino carry minus one unit while the muon and tau and their companion neutrinos carry zero units of electron lepton number. This pattern continues in the obvious way for muon and tau lepton numbers. The origin of baryon and lepton number conservation is not understood. This is in contrast to the case for electric charge conservation that follows from the local U(1) gauge invariance of electromagnetism.

2.3.3 Parity and Charge Conjugation Another essential difference of the weak interaction is that it violates the discrete symmetries of *parity* (*P*) and *charge conjugation* (*C*). Invariance under parity requires the physics of a system to be unchanged for an inversion of all coordinates through the origin. Invariance under charge conjugation requires the physics to be unchanged if all chargelike quantum numbers are reversed (i.e., particles and antiparticles interchanged). The violation of both symmetries by the charged-current weak interaction is manifest in the properties of the apparently massless neutrinos that partake only in the weak interaction.

Neutrinos are found to be left-handed objects, in the sense that their spin angular momentum is always directed in opposition (or antiparallel) to their line of motion. Under the assumption that neutrinos are massless and therefore move at the speed of light, no object can overtake their motion and thereby change their observed handedness. Antineutrinos are right-handed with their spin angular momentum always directed along their line of motion. A parity-reversed universe (left-handed spins interchanged with right-handed spins) or a charge-conjugated universe (lepton number interchanged with antilepton number) would contain right-handed neutrinos and left-handed antineutrinos—particles that are not observed. It should be noted, however, that a combined operation of charge conjugation and parity inversion (*CP*) transforms the observed left-handed neutrinos and right-handed antineutrinos into themselves.

The quarks and the charged leptons, having nonzero masses, occur in both left- and right-handed varieties. However, only the left-handed variety are subject to the weak force; the right-handed fermions do not carry the weak charge. The charged-current weak interactions are then completely left-handed while the electromagnetic and strong interactions are completely ambidextrous.

The weak transitions among quarks reveal another intriguing aspect of the charged-current weak interaction. For reasons that are not yet understood, the mass *eigenstates* and the weak-charge *eigenstates* of quarks are not identical. This leads to weak transitions among charge $+2/3$ quarks of one weak doublet with charge $-1/3$ quarks of another weak doublet. For example, the charm quark is observed to decay mainly to the strange quark, but it is also found less frequently to decay (about one chance in 20) into the down quark. The effect of the mismatch between the mass and weak-charge eigenstates of quarks can be described in terms of a unitary transformation acting on the charge $-1/3$ quark fields. This transformation can be parametrized as a complex rotation with three Euler-type angles and one complex phase. The values of these angles and the phase are fundamental parameters of the Standard Model. The angles govern the relative frequency of all of the possible transitions between the charge $+2/3$ and charge $-1/3$ quarks, and the complex phase is thought to be connected with a small violation of *CP* invariance, so far detected only in certain decays of neutral *K* mesons. It is believed that *CP* noninvariance is required to explain the gross asymmetry represented by the preponderance of matter over antimatter observed in the universe.

2.3.4 Weak Neutral-Current Interactions and Electroweak Unification The remaining agent of subnuclear interactions, the Z^0, is responsible for the weak neutral-current interactions. The Z^0 is massive enough that after being produced in high-energy collisions, it will decay into particle–antiparticle pairs of any flavor of lepton or quark (except for the very massive top quark.) Precision measurements of Z^0 decays have been used to rule out the existence of any additional flavors of massless neutrinos. These measurements have also been used to determine how the strength of the weak neutral-current interaction differs for various classes of fundamental fermions: the neutrinos, the charged leptons, and the charge $+2/3$ and charge $-1/3$ quarks. In addition, it has been found to be neither left-handed nor ambidextrous. These properties of the neutral-current interactions have a common cause: the unification of the electromagnetic and weak interactions, as proposed by Glashow, Salam, and Weinberg.

In the Standard Model of the electroweak interactions, the theory has an underlying symmetry based on the direct product group SU(2) × U(1). This symmetry requires that the electroweak gauge bosons (two electrically charged and two electrically neutral) and the fermions (quarks and leptons) are all massless—in contrast to what is observed in the real world. The symmetry is broken, however, by the solution nature has chosen from the equations governing the theory, and this gives rise to masses for the quarks and leptons as well as for the *W* and *Z*. The photon remains massless as a result of the fact that the U(1) gauge symmetry of electromagnetism remains unbroken. The breakdown of the initial gauge symmetry is often compared with the case of the spontaneous breakdown of rotational symmetry that occurs in a ferromagnet when the long-range correlation among the spins (and magnetic moments) results in the alignment of the spins along a random direction.

The spontaneous breakdown of the SU(2) × U(1) gauge symmetry in the Standard Model occurs in such a manner that the two electrically neutral gauge bosons associated with the original unbroken symmetry become mixed, leaving two linear combinations, one of which is the photon and the other the Z^0. The "peculiar" couplings of the Z^0 to the quarks and leptons thus reflect the specific "charges" associated with the unbroken gauge symmetries that are carried by the right- and left-handed fermions. In this way, three otherwise phenomenologically distinct interactions—electromagnetism and the charged- and neutral-current weak interactions—are unified in the sense of being inextricably intertwined. The masses of the *W* and the *Z* as well as the weak and electric charges are no longer independent but related in a predictable way that has been very precisely confirmed by experiment.

The price of this unification is the introduction of a new force, and the precise understanding of its origin is not yet clear. In the simplest version of the Standard Model, a set of four spin-0 particles (*scalar* fields) are introduced, and it is the pattern of the scalar boson interactions with the fermions, gauge bosons, and themselves that gives rise to the unification. The scalar fields carry the weak and electric charges; two are electrically neutral, and the other two carry $+1$ and -1 unit of electric charge. Under the spontaneous breakdown of the electroweak symmetry, the two charged scalar fields and one of the neutral scalar fields become absorbed in the definition of the massive *W* and *Z* and the final neutral scalar emerges as a detectable elementary particle known as the Higgs particle. The mass of the Higgs particle is not known, but it is expected to be less than 1 TeV. Discovering the Higgs particle, or whatever agent is in fact responsible for the electroweak-symmetry breakdown, is the next crucial step to make in order to reach a more complete understanding of the nature of matter.

3. BEYOND THE STANDARD MODEL

The development of the Standard Model of elementary particle physics is a great achievement. The quark model of hadrons and the unification of the electromagnetic and weak interactions successfully provide a coherent and simplifying framework in which to view the complex spectrum of subnuclear particles and their interactions. The predictive power of the Standard Model has been and continues to be thoroughly tested; so far, no discrepancy has emerged. However, several questions remain unanswered by the Standard Model. These questions include the following:

1. The origin of mass: What is the physics that causes the top quark to be so heavy or the electron to be so light? Are the neutrinos in fact massless?
2. What is the physical significance of there being three types of each class of fundamental fermions: neutrinos, charged leptons, and the charge +2/3 and −1/3 quarks? Why not four types?
3. What is the underlying rationale for the assignment of quantum numbers (for example electric charge) to the fundamental fermions?
4. What physics determines the magnitude of the Euler angles describing the weak quark transitions?
5. What is the origin of *CP*-invariance violation?

Some of these questions are undoubtedly related, and some extensions of the Standard Model have been contemplated that partially address these questions. It is possible that nature will reveal additional symmetries that will lead to a resolution of these issues. It is useful to mention briefly two examples. The first is supersymmetry: a postulated symmetry of nature between fermions and bosons, so that every fermion with a specific set of quantum numbers has a boson partner. While the symmetry has so far not been observed in nature (the supersymmetric partners have not been produced and identified for any of the fundamental fermions), it imposes constraints that make it very attractive in addressing the question of the origin of mass. A second class of conjectures has to do with further unification of the fundamental forces (see UNIFIED FIELD THEORIES). The possibility of unifying the symmetries governing the strong and electroweak interactions, SU(3) and SU(2) × U(1), into a larger group, such as SU(5), seems very attractive in that it can provide a rationalization for the assignment of quantum numbers to the fundamental fermions. The implication of this type of unification is the presence of additional interactions, most notably interactions that can give rise to proton decay ($p \rightarrow \pi^0 e^+$), which violates baryon (and lepton) number conservation and has so far not been observed in spite of very sensitive searches. Finally, there are even grander schemes involving an extension of quantum field theory (of point fields) to a supersymmetric quantum field theory of strings (*superstrings*), which appears to be the most promising path toward including gravity in the unified theory of the fundamental forces.

4. SUMMARY

The Standard Model in a sense plays a role in particle physics that the periodic table of the elements plays in chemistry. The rules of chemistry are established primarily by the electronic structure of atoms. It is pressing on and trying to understand the relative abundance of specific nuclear isotopes (atoms that differ only by the number of neutrons in the nucleus) that leads to the study of nuclear physics, rich in its own way. As research continues in an effort to understand the questions raised by the Standard Model, we can expect new insights into the properties of matter to emerge, whether it be in the form of massive new particles being produced at high-energy accelerators, unexpected effects discovered in precision experiments or searches for rare processes, or from astrophysical considerations and observations.

Works Cited

Particle Data Group (1994), "Review of Particle Properties," *Phys. Rev. D* **50**, 1173–1826.

Wilczek, F. (1982), *Annu. Rev. Nucl. Part. Phys.* **32**, 177–209.

Further Reading

Frauenfelder, H., Henley, E. (1991), *Subatomic Physics*, 2nd ed., Englewood Cliffs, NJ: Prentice Hall.

Pais, A. (1986), *Inward Bound*, New York: Oxford University Press.

Pickering, A. (1984), *Constructing Quarks*, Chicago: University of Chicago Press.

Williams, W. S. C. (1991), *Nuclear and Particle Physics*, New York: Oxford University Press.

PATENTS AND INTELLECTUAL PROPERTY

MARIA REICHMANIS, *Registered Patent Attorney, Aiken, South Carolina, U.S.A.*

	Introduction	233
1.	**Types of Intellectual Property: Patents, Trademarks, Copyrights, and Trade Secrets**	233
2.	**Patents**	235
2.1	The Patent Application	235
2.1.1	The Specification, Claims, and Drawings	236
2.1.2	Examination	237
2.2	Enforcing Patent Rights	237
2.2.1	Assignment vs License	237
2.2.2	Infringement	238
2.3	International Filing	238
2.3.1	International Filing and the Patent Cooperation Treaty	239
3.	**Trademarks, Service Marks, Certification Marks, and Collective Marks**	239
3.1	Enforcement of Trademark Rights	239
4.	**Copyrights**	240
4.1	Copyrightable Works	240
4.2	Exclusive Rights	240
4.2.1	Fair Use	240
4.3	Ownership	240
4.4	Notice, Registration, and Deposit	241
4.5	Enforcement of Copyrights	241
5.	**Trade Secrets**	241
6.	**Appendix**	242
	Glossary	242
	Works Cited	243
	Further Reading	243

INTRODUCTION

"Property" has been defined as the exclusive right to possess, enjoy, and dispose of a thing (Nolan and Connolly, 1979). Thus, property in the legal sense is not so much the thing itself, but rather the aggregate of rights in the thing. Property can be real (land; and whatever is erected on, growing upon, or affixed to the land) or personal (money; goods; stocks and bonds). It can be tangible (a house or an automobile) or intangible (contract rights). Intellectual property—patents, trademarks, copyrights, and trade secrets—is personal property.

The owner of intellectual property has the same rights with respect to the property as the owner of a house has with respect to her property: subject to applicable laws, she may prevent another from using, selling, or otherwise exercising any form of control over the house. However, she may choose to rent the house for a period of time, give it away, or sell it. The owner of intellectual property, like the owner of a house, may assign all her rights in the property to another, or grant a license giving the right to use the property under specified conditions.

1. TYPES OF INTELLECTUAL PROPERTY: PATENTS, TRADEMARKS, COPYRIGHTS, AND TRADE SECRETS

The four generally recognized forms of intellectual property are listed in Table 1. Each protects different aspects of an idea and confers different rights upon its owner.

A patent is a grant from a government that conveys to an inventor the exclusive right to make, use, and sell the patented invention for a specified term, usually a term of years. United States law provides for three types of patents: utility patents, for novel, useful, and nonobvious machines, processes, compositions of matter, and articles of manufacture; plant patents, for new and distinct, asexually reproduced varieties of plant; and design patents, for novel and nonobvious ornamental designs for articles of manufacture.

A trademark or service mark is any word,

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Table 1. Patent, trademark, copyright, and trade secret protection.

	Patent	Trade secret	Copyright	Trademark, service mark
Subject matter	Process, machine, article of manufacture, composition of matter	Something that gives its owner an advantage over competitors that do not know it	Literary works; musical and dramatic works; pantomimes and choreographic works; pictorial, graphic, and sculptural works; motion pictures and other audiovisual works; sound recordings	A word, name, symbol, device, or combination thereof that identifies the product or service
Requirements for protection	Novel, useful, nonobvious	Keep it secret	Original; fixed in a tangible medium of expression	Distinctive; used as a mark
What is protected?	The substance of the idea	The substance or the form of the idea, or both	The expression of the idea	The appearance of the mark
How is it protected?	Prevent others from making, using, or selling it	Prevent others from using it, if obtained unlawfully	Prevent others from making copies or derivative works; prevent others from distributing, performing, or displaying the copyrighted work	Prevent others from using later-registered, confusingly similar marks
When is it protected?	Upon the grant of a patent	Upon creation	Upon fixation in a tangible medium of expression	Upon first use
For how long is it protected?	17 years (utility patent) 17 years (plant patent) 14 years (design patent)	As long as it is a secret	Life of author plus 50 years; if a "work for hire," 100 years after creation or 75 years after publication	As long as it is used

name, symbol, device, or any combination thereof, adopted and used to identify goods or services. Trademarks are used to identify the source of goods; service marks are used to identify the source of services. The owner of a trademark or service mark can prevent others from using it or confusingly similar marks, but trademark or service-mark protection alone does not give him the right to prevent others from providing the same product or service *without* the mark.

Copyright law deals with information and ideas reduced to tangible form. Unlike patents, which protect the substance of the idea, copyrights protect the expression of the idea.

A trade secret is any confidential information that gives its owner an advantage over competitors that do not know it. The secret may be a machine or a chemical process, an abstract idea such as a concept for a new consumer product, or a compilation of data such as a list of customers. Matters of common knowledge, or matters of general knowledge in a particular industry, cannot be protected as trade secrets.

Patent law protects the substance of an idea in order to encourage investment in new technology and invention, whereas trademark law is concerned with the protection of symbols that identify products and services. Copyright law protects the original expression of an idea, and trade secret law protects confidential information. A single product may be the subject of more than one form of intellectual property. For example, the design of an automobile body may be copyrightable or protectible by a design patent; the engine, brakes, seat belts, and other parts may be protected by utility patents; some of the parts may be made by secret processes; and the model name may be a trademark (McCarthy, 1990).

2. PATENTS

The United States Patent and Trademark Office (PTO) issues three types of patents: utility patents, plant patents, and design patents.

Utility patents are issued for novel, useful, and nonobvious processes, machines, articles of manufacture, or compositions of matter. A "nonobvious" invention may be the result of inspiration, routine study and experimentation, or pure chance—there is no single test of obviousness. The owner of the patent need not exploit the invention himself. However, he can *prevent others* from making, using, or selling the invention during the term of the patent regardless of whether or not he, himself, does so.

Fundamental laws of nature ($F = ma$; $E = mc^2$) and mathematical algorithms ($\sin^2\theta + \cos^2\theta = 1$) are not patentable, under the rationale that the result would be a monopoly on a scientific truth. This prohibition includes computer programs that are pure algorithms, although such programs may be the subject of copyright protection. However, a process that uses a mathematical algorithm may be patentable if the process itself, not merely the algorithm, is new and useful. Other nonpatentable categories include naturally occurring substances, methods of doing business, forms, and abstract ideas.

Even though naturally occurring substances are not patentable in themselves, highly purified products and extracts thereof may be patentable. In addition, they may form part of a patentable product or be used in a patentable process. Drugs, vitamins, food additives, and other products that are derived from natural products are patentable; processes for making such products may also be patentable. Genetically engineered life forms are patentable if they meet the statutory requirements of novelty, usefulness, and nonobviousness (Miller and Davis, 1983).

Plant patents are issued for new and distinct asexually reproduced plants, including mutants, hybrids, cultivated sports, and newly found seedlings other than tuber-propagated plants or plants found in an uncultivated state (sexually reproduced plants are protectible under the Plant Variety Protection Act). To be patentable, a plant must be nonobvious as well as new and distinct, but it need not be useful. Patent protection applies only to the plants themselves, not to flowers, fruits, bulbs, or seeds *per se*.

Design patents protect new, original, and ornamental designs for articles of manufacture, that is, the *appearance* of a product rather than its function. A design patent can protect surface ornamentation applied to a product or the overall shape of the product, but not a shape that is dictated by the function of the product. For example, a simple bed frame with four legs and a platform for supporting a mattress would not be a proper subject for design patent protection, since a bed needs a mattress support in order to *be* a bed. However, the shape and size of the legs, the headboard, and the footboard may be patentable, since these aspects of the design are not dictated by their utilitarian purpose (the requirement that the frame be capable of supporting a mattress). All of these features could easily be changed or even eliminated without changing the *function* of the product.

An invention may be the subject of both design and utility patent protection: a utility patent, for its function, and a design patent, for its ornamental appearance. When there is a choice between trademark registration of a nonfunctional design and a design patent, trademark protection may be preferred because of the potentially unlimited term (i.e., as long as the mark is used by the trademark owner).

2.1 The Patent Application

In most countries, patents are obtained by filing an application with the appropriate government agency. In the United States, a patent is obtained by filing an application with the Commissioner of the Patent and Trademark Office (PTO). The PTO examines the application to determine whether the three tests for patentability—novelty, utility, and nonobviousness—are met. If the invention in question satisfies these tests and all relevant fees are paid, a patent issues in due course.

In the United States, a patent application must be filed by the inventor or on behalf of the inventor; in some countries, an application may be filed by the owner of the invention—for example, the inventor's employer. A patent application includes a verbal description of the invention (the "specification"), drawings (where needed to understand the invention), and "one or more claims particularly pointing out and distinctly claiming the subject matter which the applicant regards as his invention." The application process is governed by a set of regulations developed by the PTO (see U.S. Patent and Trademark Office, 1992). Some applicants prepare their own patent applications; however, most work with a registered patent attorney or agent.

2.1.1 The Specification, Claims, and Drawings The specification includes the title of the invention, cross-references to related applications (if any), a discussion of the background of the invention explaining the context of the problem that is being addressed, a brief summary of the invention, a description of the drawings, a detailed description of the invention, the claims, an abstract, and the drawings, and an oath or declaration signed by the inventor(s). When included in a patent application, drawings must show every element of the claims. "Drawings" may include not only line drawings, but graphs, flow charts, and photographs where appropriate. The invention should be described as clearly as possible so that it is understandable by the person "of ordinary skill in the art to which it pertains." Either the specification, the claims, or both jointly must constitute an "enabling disclosure," that is, provide an explanation that is sufficient for others to practice the invention.

The claims provide a concise, formal definition of the invention. The claims define the metes and bounds of the invention, much as a deed defines the boundaries of a parcel of land. A claim contains three parts: the preamble, the transition, and the body. The preamble generically defines the invention, for example, "An apparatus for doing X." The transition relates the preamble to the body, typically with the words "comprising" ("including parts A, B, and C, and maybe D, E, . . . "), "consisting" ("including only A, B, and C"), or "consisting essentially of" ("including A, B, and C, and nothing else that matters"). The body describes elements A, B, and C that make up apparatus X. Because the claims describe the invention in words, they are inherently less precise than the description that appears in a deed to real estate. Any given invention can be described by many different combinations of words, ranging from broad ("a chair, comprising means for supporting a person in a sitting position") to narrow ("a chair, consisting of a round wooden seat, and three equally spaced cylindrical legs attached to the bottom of the seat").

It is sometimes helpful to view an invention as a territory analogous to real estate. Thus, a new, useful, and nonobvious widget is located in a region unoccupied by previously known, prior-art widgets A, B, and C (Fig. 1). A broad claim for the new widget encompasses as much previously unclaimed territory as possible, up to (but not crossing) the boundaries of the territories occupied by A, B, and C (Fig. 2). A narrow claim covers much less territory (Fig. 3). A patent application for the new widget might contain a series of claims, ranging from the broadest possible, covering many different ways of practicing the invention, to the narrowest that will protect the actual manufactured product. A single, broad claim would protect the invention, but narrower claims could protect

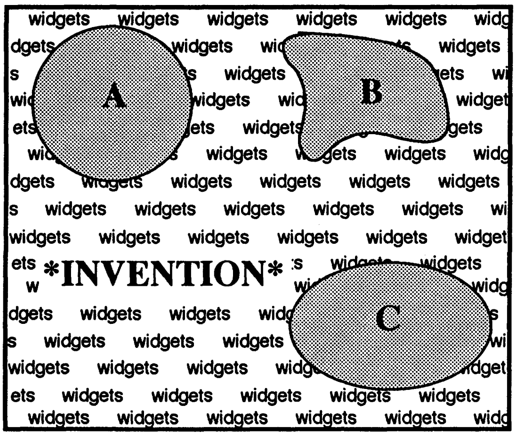

FIG. 1. The field of widgets. Previously known, prior-art widgets (A, B, C) and the invention occupy separate territories. The invention is outside the territories of widgets A, B, and C.

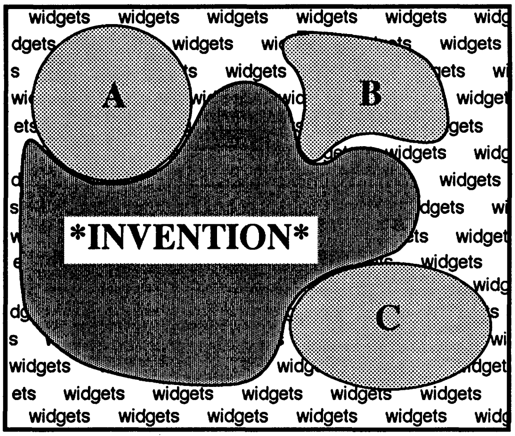

FIG. 2. Broad claim. In addition to the best mode, the claimed territory includes other modes of making the new widget, up to the boundaries of prior art widgets A, B, C.

the actual product if the broader claims are later found invalid.

2.1.2 Examination After a patent application is filed, the PTO conducts a search of prior United States patents, foreign patents, and publications (including the scientific literature) to determine whether the invention is novel and nonobvious. Frequently, an application is rejected and the applicant applies for reexamination. He may submit arguments in response to the examiner's stated reasons for rejection, with or without amending the application to address the reasons for the rejection. This process of examination, rejection, and amendment, known as "prosecution," may be repeated until eventual acceptance or a "final rejection" of the application. Procedures are available for appealing a rejection based on the substance of the application; rejections based on matters of form and procedure must be reviewed by petition to the Commissioner of Patents and Trademarks.

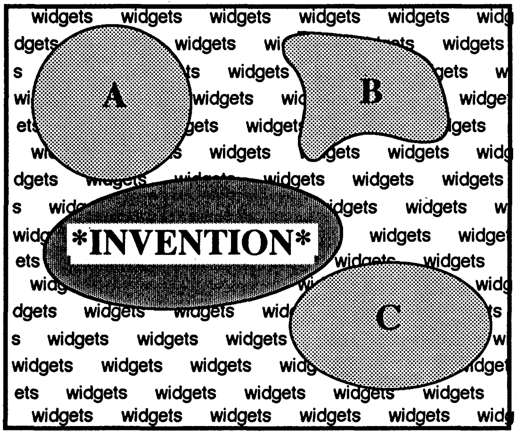

FIG. 3. Narrow claim. The claimed territory includes only the best mode of making the new widget; it does not reach to the boundaries of prior-art widgets A, B, C.

Occasionally an application claims the same invention as is claimed in another pending application or an issued patent. In that case, the PTO may conduct an interference proceeding to determine who is the first inventor and entitled to a U.S. patent on the invention.

If and when the PTO decides that a patent is to be granted, the applicant is sent a Notice of Allowance. If the issue fee is paid, the patent is granted. Thereafter, maintenance fees are due in the 4th, 8th, and 12th years to maintain the patent. If the fees are not paid, the rights lapse and the patented invention enters the public domain.

2.2 Enforcing Patent Rights

An issued patent is intellectual property having all the attributes of other forms of property. Most importantly, the owner has the right to exclude others from making, using, or selling the patented invention for a specified term (at present, 17 years in the United States; to compensate for delays in marketing caused by the regulatory process, the term may be extended for inventions that are subject to review by the Food and Drug Administration). At the end of that term, the invention enters the public domain.

2.2.1 Assignment vs License Although patent rights come into being only when a patent issues, patentable inventions may be treated as property even before issuance. For example, patent applications and notices of allowance may be viewed as property for income tax purposes. An inventor may assign rights to an invention at any time; in general, the further along the inventor has taken the application process, the more valuable his or her property rights in the invention.

The patentee may assign rights in a patent application or a patent, that is, sell all rights that he or she has in the patent. When he does so, he conveys the entire right to exclude others from making, using, or selling the invention. An assignment is equivalent to a sale of real estate. To pass legal title, an assignment must be in writing. While not essential, recording an assignment in the PTO establishes title against subsequent assignees who take from the same assignor.

A patentee may license the right to make, use, and sell the invention. A license may be exclusive or nonexclusive, for all or part of the claimed invention, for the remaining term of the patent or for a lesser term. It may be limited to a certain geographic area ("only in North Carolina and South Carolina"), provide for a fixed payment or royalties based on sales, or indeed any terms that suit the business needs of the parties.

2.2.2 Infringement Patent rights come into being only when the patent issues; others are free to make, use, and sell the invention before that time. The words "patent pending" may offer some psychological protection, since competitors do not know when the patent will issue, or what features of the product the inventor believes are patentable (U.S. patent applications are not available to the public), but afford no legal protection.

Anyone who makes, uses, or sells a patented invention during the term of the patent is a direct infringer. Anyone who actively encourages another to do so is an indirect infringer. Anyone who knowingly sells an item for which the only use is in connection with a patented invention is a contributory infringer.

Determining whether a patent has been infringed involves a thorough analysis of the claims and all the documents relating to the patent, including the original application, the prior art cited and discussed during prosecution of the application, written exchanges between the patent attorney or agent and the examiner, and the record of interference proceedings (if any). The test is whether the allegedly infringing product or process does the same work, in substantially the same way, and accomplishes substantially the same result as the patented invention.

Remedies for patent infringement include injunctions to prevent future infringement, damages to compensate the patent owner, attorneys' fees when appropriate, and costs. Damages may be trebled in the case of willful infringement (Miller and Davis, 1983).

2.3 International Filing

Patent laws vary from country to country, but certain features are common to all. Patents are granted for a specified period of time, and for novel and nonobvious inventions. In some countries, the invention must be "applicable to industry" (roughly equivalent to the "useful" requirement of U.S. law); some require that a patent, once granted, be exploited (i.e., the owner *must* make, use, or sell the invention in order to keep his rights in force).

Depending on the law of the country involved, applications may be filed by the actual inventor, in the name of the inventor, or by the assignee. Different types of patent protection may include patents of improvement, granted for an improvement of another patented invention; dependent patents, for inventions which cannot be practiced without infringing an earlier patent; and a type of patent known as an importation, revalidation, confirmation, or introduction patent, that gives the owner of a foreign patent local protection for a term limited to the unexpired term of the underlying patent. Inventor's certificates constitute recognition of inventorship and a grant of the rights and privileges provided by the national law; however, the exclusive rights are transferred to the State. Precautional patents provide that no patent relating to the subject matter of the patent will be granted without notice to the owner, who is given the opportunity to lodge objections within a stipulated term. Petty patents (for example, the German Gebrauchsmuster and the Japanese utility model) protect the shape or configuration of an object insofar as it serves some useful purpose (Baxter and Sinnott, 1992).

The United States is almost alone in having a "first to invent" patent system; most countries have a "first to file" system, where the first inventor to file a patent application is entitled to a patent, regardless of who actually invented first. In most countries, including the United States, the government has the right to make secret any patent application covering an invention which is of importance to the national defense. Once a secrecy order is given, the application proceeds through the normal examination stages, but is not published, nor is a patent granted, until the order is removed.

2.3.1 International Filing and the Patent Cooperation Treaty Before implementation of the Patent Cooperation Treaty (PCT) in 1978, patent applications were filed in each country where protection was desired. Although it does not provide for an "international patent" as such, the PCT has simplified the procedures that nationals or residents of member countries must follow when filing patent applications in other member countries. The PCT provides for centralized filing and searching of an international application (the "international stage"), followed by filing an application in each country where patent protection is desired (the "national stage"). Substantive examination of the application is carried out by each respective national patent office. The PCT preserves the right of priority of the Paris Union Convention and the Inter-American Convention, that is, an applicant who has filed in any member country may thereafter file a patent application for the same invention in any other member country within 12 months, with the effect that these later-filed applications will be treated as if they were filed on the first filing date.

In addition to the PCT route for obtaining international patent rights, the European Patent Convention (EPC) provides for centralized filing and examination leading to the grant of a "regional" patent, which matures into an individual patent for each of the member countries designated by the applicant. An applicant may, of course, file applications only in the desired countries.

Obtaining patent protection in African countries can be complicated by the variety of different patent systems involved. The African Intellectual Property Organization (OAPI) is responsible for implementing and applying common administrative procedures for the protection of industrial property (Baxter and Sinnott, 1992).

3. TRADEMARKS, SERVICE MARKS, CERTIFICATION MARKS, AND COLLECTIVE MARKS

Trademarks and service marks are believed to have originated with the medieval guilds, whose members affixed the mark of their guild to the goods they sold to identify it as the product of a particular craftsman or group of craftsmen. Like guild marks, modern trademarks serve two purposes: they identify the source of the goods, and they provide a marketing advantage for the owner of the mark (Miller and Davis, 1983).

A trademark is a symbol that allows a purchaser to identify goods or services that have been satisfactory in the past and those that have not. Service marks are identical to trademarks, except that they indicate the source of services, not goods. Trademarks and service marks encourage the production of quality products, and reduce the customer's costs of acquiring information about products and services. Most countries afford some form of trademark and service mark protection as an identification of the source and quality of goods and services (McCarthy, 1990). PTO trademark registration procedures are described in U.S. Patent and Trademark Office (1993).

A certification mark is any word, name, symbol, device, or combination thereof that is used to indicate that goods or services meet certain quality or regional origin standards. For example, the term "Roquefort" identifies a particular type of cheese that originates in a province of France; the "Underwriters Laboratories" and "Good Housekeeping" seals indicate that particular products meet the standards set by the owner of these marks. The mark must be one that buyers are likely to recognize as a symbol of guarantee or certification. Most titles or degrees such as Professor, Professional Engineer, MD, CPA, etc. are not used to certify the services performed by an individual, but rather to convey personal information about the individual performing those services.

Collective marks are used by members of a cooperative, an association, or a collective group or organization. While the organization itself neither sells goods nor performs services, its members do. Use of the mark allows members of the group to identify themselves as such (McCarthy, 1990).

3.1. Enforcement of Trademark Rights

Federal remedies for trademark infringement include injunctions, an accounting for profits, damages (including treble damages where appropriate), attorney's fees, and costs.

4. COPYRIGHTS

Copyright protection is available for original works of authorship. "Original" does not mean "new" or "novel"; it means only that the author did not copy the work from someone else. Consider *Romeo and Juliet*: an author could copyright a story about star-crossed lovers as long as it was not copied from Shakespeare.

4.1 Copyrightable Works

Copyrightable works include all original works of authorship "fixed in any tangible medium of expression, now known or later developed." Copyrightable works include literary works, musical and dramatic works, pantomimes and choreographic works, pictorial, graphic, and sculptural works, motion pictures and other audiovisual works, and sound recordings. Copyright protection does not extend to the idea embodied in the work, just the expression. Architectural plans, technical drawings, diagrams, and models are copyrightable (Nimmer and Nimmer, 1992).

Purely utilitarian works, including forms, contest rules, tests, and similar items are not protected by copyright. Since others are free to use the *idea* embodied in the work, there is very little protection available to the original author of, for example, an accounting system, a layout for a laboratory notebook or a method of doing business. Even though purely utilitarian designs are not copyrightable, fabric designs, jewelry designs, etc. may be copyrightable to the extent that the design is separable from its utilitarian aspect.

To the extent that a work is not "fixed," it is not copyrightable. For example, a dance routine is copyrightable only to the extent that it is filmed or written down in choreographic notation. Typefaces are usually considered to be noncopyrightable, perhaps because the subject matter—the alphabet—is limited and finite. However, the names of typefaces are subject to trademark protection.

Computer programs and databases are copyrightable to the extent that they incorporate the programmer's original expression of ideas, as distinguished from the ideas themselves (whether or not the idea is original). Microcode or firmware, which converts source code into object code, is itself copyrightable as a computer program (Nimmer and Nimmer, 1992). To the extent that a computer program lacks minimal originality or constitutes the only way of accomplishing a particular result, it is not copyrightable. Such a program embodies the idea itself rather than the expression of an idea.

4.2 Exclusive Rights

The owner of a copyright has the exclusive rights to do and to authorize the following: to reproduce the copyrighted work; to prepare derivative works based upon the work; to distribute copies or phonorecords of the work by sale, rental, lease, or lending; to perform the work publicly; and to display the work publicly. The right of reproduction allows the owner to exclude all others from reproducing the work in the form of a copy or phonorecord. Two exceptions allow radio, television, and cable television companies to record works temporarily for later transmission, and the owners of computer programs to make copies to utilize the program effectively.

The right of making derivative works allows the owner to prevent others from creating works based on the copyrighted work, such as translations, arrangements, dramatizations, films, abridgements, etc.

The right of distribution applies only to the first sale of a copy. Once a sale has occurred, the new owner can proceed to sell the work to another, but not copy the work.

4.2.1 Fair Use "Fair use" of a copyrighted work is not infringement. The concept of fair use balances the purpose and character of the use, the nature of the work, the amount and substantiality of the portion used, and the effect of the potential market for the work. One factor is whether the use is commercial in nature, or for nonprofit educational purposes. Educational purposes are traditionally a basis in favor of finding that the use is "fair." Under the current statute, limited photocopying by libraries and archives, and for classroom use, is allowed.

4.3 Ownership

A work may have a single author or joint authors. It may be a collection or a compilation, or a work for hire created for publication by another, either as an employee or an independent contractor. The ownership and duration of the copyright may hinge upon the classification of the work.

Ownership vests initially in the author or authors, or, in the case of a work made for hire, in the employer or other person for whom the work was prepared. Ownership may be transferred in whole or in part, may be bequeathed by will, or may pass as personal property. The individual rights may be transferred separately; thus, the right to make a movie and the right to translate a novel may be separated and transferred to different persons. Coauthors of copyrighted works are much as co-owners of other types of personal property; each has the right to use the property for his own purposes, but neither has the right to exclude the other from using it.

In the case of a single author, the copyright endures for the life of the author plus 50 years (70 years in most Berne Convention countries); for coauthors, for the life of the last surviving author plus 50 years. For works for hire, the term is 75 years from the date of first publication, or 100 years from the date of creation, whichever comes first.

The present U.S. copyright statute allows authors who have sold or otherwise transferred their rights to terminate the transfer by serving notice that they intend to do so during a five-year period beginning at the end of 35 years from the date of execution of the grant. This provision protects authors who have sold their works, but discover later in their careers that the works have become more valuable. If the author is dead, his or her spouse and children may exercise the right.

4.4 Notice, Registration, and Deposit

Under present U.S. law, an author has a copyright in a work as soon as it is fixed in a tangible medium of expression; a notice on published copies is no longer necessary to preserve the copyright. A notice of copyright consists of the symbol ©, the word "Copyright," or the abbreviation "Copr.," followed by the year of first publication and the name of the owner. For phonorecords, notice consists of the letter "P" in a circle, the year of first publication, and the name of the copyright owner.

Omission of notice does not invalidate the copyright, if

1. registration is obtained within five years, and
2. good faith efforts are made to affix notice to all remaining copies.

An innocent infringer incurs no liability for damages for infringing acts committed before receiving actual notice.

There is no need for examination: since others are free to create whatever they want as long as they do not copy the work, there is no need to insure that the work meets any particular standards. However, registration is a condition to filing a lawsuit for infringement.

Deposit of two copies of the work with the Library of Congress is required within three months of publication; however, deposit is not a condition of copyright protection. Registration requires deposit of at least one copy of a work. United States copyright registration procedures are described in U.S. Copyright Office (1984).

4.5 Enforcement of Copyrights

The owner of the copyright is entitled to sue for infringement. Available remedies include injunctions to prevent or restrain infringement, court-ordered licensing fees, and actual damages plus the defendant's profits. Costs and attorney's fees may also be recovered by the prevailing party. The defendant may establish his expenses to reduce the gross profit to a net amount. The plaintiff may choose to ask for statutory damages (a minimum of $250 and a maximum of $10,000) instead of actual damages, and a court may award as much as $50,000 statutory damages for willful infringement.

5. TRADE SECRETS

A trade secret has been described as any commercially valuable information that gives its owner an advantage over competitors that do not know it. A trade secret may be a machine or a process, an abstract idea such as a concept for a new product, or a compilation of data such as a list of customers. Trade secrets are confidential; however, not all confidential business information is a trade secret. Matters of common knowledge, or matters of general knowledge in a particular industry or trade, cannot be protected as trade secrets (McManis, 1988).

In some cases, an inventor may have the option of obtaining patent protection for an invention or relying on trade secret protection. A patent protects the invention for a period of time after it is disclosed; however, trade secret law protects it for as long as it remains secret.

Courts protect trade secrets under several theories. Some consider trade secrets to be a form of intellectual property, analogous to patents, copyrights, and trademarks. Others emphasize the breach of a confidential relationship, impropriety in ascertaining a secret, and use of a trade secret or idea known to have been improperly disclosed by another. For example, industrial espionage and bribery of a competitor's employees are an improper means of obtaining information, whereas "reverse engineering," discovery by one's own independent efforts or by inspection of a publicly distributed product, is usually considered to be proper.

6. APPENDIX

Professional societies involved in intellectual property law and related areas include the Section of Patent, Trademark and Copyright Law of the American Bar Association (ABA), the American Intellectual Property Law Association (AIPLA), the International Trademark Association (INTA, formerly the United States Trademark Association), and the World Intellectual Property Organization (WIPO). These groups sponsor numerous meetings, workshops, and courses throughout the year. Courses on various aspects of intellectual property are also offered by the Patent Resources Institute, Inc. (Charlottesville, VA), the Practicing Law Institute (New York, NY), and several universities.

Treatises, introductory texts, and practice handbooks dealing with intellectual property law, both U.S. and international, are available from Matthew Bender & Co., Inc. (New York, NY), Clark, Boardman, Callaghan (Deerfield, IL), and West Publishing Co. (St. Paul, MN), among others. The *Copyright Law Journal* and the *Trademark Reporter* deal with current developments in the areas of copyright and trademark law, respectively. The *United States Patents Quarterly* and the *Patent, Trademark & Copyright Journal,* both published by the Bureau of National Affairs, Inc. (Washington, DC), provide comprehensive coverage of U.S. court cases. Texts of newly issued patents are published in the *Official Gazette.* Information on court decisions, texts of recently issued patents, periodicals, news reports, etc. is also available on computer databases, including LEXIS (Mead Data Central) and WESTLAW (West Publishing Co.).

United States statutes dealing with intellectual property are found in Titles 15 (commerce and trade), 17 (copyright), and 35 (patents) of the United States Code (U.S.C.). Applicable regulations may be found in the Code of Federal Regulations (C.F.R.). Both are available online (LEXIS, WESTLAW). Trademarks, trade secrets, and unfair trade practices may also be the subject of state law.

The procedural aspects of the U.S. patent application process, federal trademark registration, and copyright registration are detailed in the *Manual of Patent Examining Procedure* (*MPEP*), *Trademark Manual of Examining Procedure* (*TMEP*), and *Compendium of Copyright Office Practices,* available from the Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402.

GLOSSARY

Assignment: A conveyance of all rights in property.

Certification Mark: A word, name, symbol, device, or any combination thereof, used to indicate that goods or services meet certain standards of quality or regional origin. The mark must be such that buyers are likely to recognize it as a symbol of guarantee or certification.

Claim: A numbered, single-sentence paragraph that provides a concise, formal definition of an invention; part of a patent application.

Collective Mark: A word, name, symbol, or device used by members of a cooperative, an association, a collective group, or an organization to identify themselves as such.

Copyright: A statutory right granted to the author of a literary or artistic work.

Design Patent: A patent granted for a new ornamental design.

EPO: European Patent Convention.

Infringement: A violation of the rights secured by a patent, trademark, or copyright; unauthorized use of the subject matter of a patent, trademark, or copyright.

Injunction: An equitable remedy granted by a court, forbidding a party to do some act, or restraining him in continuing to perform the act.

Intellectual Property: Includes patents, trademarks, copyrights, and trade secrets.

License: A conveyance of the right to make, use, or sell a patented article for a limited period or in a limited territory.

Notice of Allowance: A notice issued by the PTO, informing the applicant that a patent is to be granted.

OAPI: African Intellectual Property Organization.

Patent: A grant from a government that conveys to the patentee the exclusive right to make, use, and sell the invention for a specified term.

PCT: Patent Cooperation Treaty.

Plant Patent: A patent granted for a new, distinct, asexually reproduced variety of plant.

Property: The exclusive right to possess, enjoy, and dispose of a thing.

PTO: Patent and Trademark Office.

Service Mark: A word, name, symbol, device, or any combination thereof, adopted and used to identify services.

Specification: A detailed verbal description of an invention; part of a patent application.

Trade Secret: Any commercially valuable information that gives its owner an advantage over competitors that do not know it.

Trademark: A word, name, symbol, device, or any combination thereof, adopted and used to identify goods.

Utility Patent: A patent granted for a machine, process, composition of matter, or article of manufacture.

Works Cited

Baxter, J. W., Sinnott, J. P. (1992), *World Patent Law and Practice*, 3rd ed., New York: Matthew Bender & Co., Inc.

McCarthy, T. J. (1990), *McCarthy on Trademarks and Unfair Competition*, 3rd ed., Deerfield, IL: Clark, Boardman, Callaghan.

McManis, C. R. (1988), *Unfair Trade Practices*, St. Paul, MN: West Publishing Co.

Miller, A. R., Davis, M. H. (1983), *Intellectual Property—Patents, Trademarks and Copyright*, St. Paul, MN: West Publishing Co.

Nimmer, M. B., Nimmer, D. (1992), *Nimmer on Copyright. A Treatise on the Law of Literary, Musical and Artistic Property, and the Protection of Ideas*, New York: Matthew Bender, Inc.

Nolan, J. R, Connolly, M. J. (1979), *Black's Law Dictionary*, 5th ed., St. Paul, MN: West Publishing Co.

U.S. Copyright Office (1984), *Compendium of Copyright Office Practices*. Available from the Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402.

U.S. Patent and Trademark Office (1992), *Manual of Patent Examining Procedure (MPEP)*. Available from the Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402.

U.S. Patent and Trademark Office (1993), *Trademark Manual of Examining Procedure (TMEP)*. Available from the Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402.

Further Reading

Becker, S. A. (1992), *Patent Applications Handbook*, Deerfield, IL: Clark, Boardman, Callaghan.

Boorstyn, N. (1981), *Copyright Law*, Deerfield, IL: Clark, Boardman, Callaghan.

Cooper, I. P. (1982, revised annually), *Biotechnology and the Law*, Deerfield, IL: Clark, Boardman, Callaghan.

Epstein, M. A., Laurie, R. S., Elder, L. E. (1992), *International Intellectual Property: The European Community and Eastern Europe*, Englewood Cliffs, NJ: Prentice Hall.

Epstein, M. A. (1992), *Modern Intellectual Property*, Englewood Cliffs, NJ: Prentice Hall.

McManis, C. R. (1988), *Unfair Trade Practices*, St. Paul, MN: West Publishing Co.

Miller, A. R., Davis, M. H. (1983), *Intellectual Property—Patents, Trademarks and Copyright*, St. Paul, MN: West Publishing Co.

Nimmer, M. B., Geller, P. E. (1988, revised annually), *International Copyright Law and Practice*, New York: Matthew Bender, Inc.

Scott, M. D. (1992), *Scott on Computer Law*, Englewood Cliffs, NJ: Prentice Hall.

PETROLEUM

FRED H. POETTMANN AND ROBERT S. THOMPSON, *Department of Petroleum Engineering, Colorado School of Mines, Golden, Colorado, U.S.A.*

	Introduction	269
1.	**Definition of Petroleum**	270
2.	**Physical Properties of Petroleum**	272
2.1	Gas	273
2.1.1	Dry Gases	273
2.1.2	Wet Gases	273
2.1.3	Retrograde Gas Condensates	274
2.2	Oils	274
2.2.1	Volatile Oils	274
2.2.2	Black Oils	275
3.	**Geology and Occurrence of Petroleum**	275
3.1	Reservoir Traps	275
3.2	Reservoir Rock Characteristics	275
3.2.1	Porosity	275
3.2.2	Permeability	276
3.2.3	Factors Influencing Rock Properties	276
4.	**Drilling and Completion Operations**	277
4.1	Directional and Horizontal Wells	277
4.2	Casing and Wellhead Design	278
4.3	Completion Types	279
4.4	Improving Well Performance	280
5.	**Petroleum Recovery Mechanisms**	280
5.1	Primary Recovery Mechanisms	281
5.1.1	Solution-Gas Drive	281
5.1.2	Gas-Cap Drive	282
5.1.3	Water-Drive Reservoirs	282
5.1.4	Gravity Drainage	282
5.1.5	Combination Drives	282
5.2	Secondary Recovery	282
5.2.1	Water Flooding	282
5.2.2	Gas Injection, Pressure Maintenance	283
5.3	Improved Oil Recovery Mechanisms	284
5.3.1	Polymer Flooding	284
5.3.2	Miscible Flooding	284
5.3.3	Thermal Methods	285
5.3.3.1	Steam Injection	285
5.3.3.2	*In Situ* Combustion	286
5.3.4	Micellar-Polymer Flooding	286
5.3.5	Alkaline-Surfactant-Polymer Flooding	287
6.	**Production Operations**	287
6.1	Natural Flow	287
6.2	Artificial Lift	287
6.3	Field Processing	289
6.3.1	Natural Gas Processing at the Wellhead	289
6.3.2	Wet Gas, Gas Condensates, and Volatile Oil Processing at the Wellhead	289
6.3.3	Black Oils Processing at the Wellhead	289
7.	**Petroleum Reserves and Economics**	289
7.1	Definitions	291
7.2	Classification of Reserves	293
7.3	Petroleum Resource Assessment	293
7.4	Petroleum Economics	293
	Glossary	293
	Works Cited	294
	Further Reading	295

INTRODUCTION

The word "petroleum" is derived from the Latin—"petra" meaning rock and "oleum" meaning oil, "rock oil." Crude oil, natural gas, retrograde gas condensate, and volatile oils are all varieties of petroleum and are referred to as petroleum.

The petroleum industry as we know it today got its start in the mid nineteenth century. The very rapid growth of the industry came about with the development of the in-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ternal combustion engine in the beginning of the twentieth century. However, archaeological evidence indicates that ancient man made use of oil and gas seeps. Asphalt, a weathered oil, was used to cement stones, to waterproof cisterns, and to caulk boats. Chinese references over 3000 years old indicate that the Chinese used petroleum in their lamps. The Chinese while looking for brine to make salt developed a drilling technique similar to modern cable tool rigs. The ancient Chinese also transported natural gas through bamboo tubing for the purposes of heating and light.

In more modern times, the exploration for and the production of petroleum were accelerated by the demand for low-cost energy necessary to satisfy our consumer-oriented society. Not just gasoline, but heating oils, diesel fuels, lubricants, and the raw materials for petrochemical manufacture are produced from petroleum. Fuel oil and natural gas have replaced coal for home heating purposes.

The various categories of petroleum can best be characterized by their temperature–pressure phase diagrams along with the temperature at which the fluid exists in the reservoir. This identifies the petroleum as being a dry natural gas, a wet natural gas, a retrograde gas condensate, a volatile oil, or a black oil.

Petroleum occurs in underground reservoirs consisting of porous and permeable rock into which the petroleum migrated from a source bed. In order to produce this petroleum, wells are drilled to provide a conduit between the reservoir and the surface. The wells serve as a pressure sink into which the petroleum flows. The initial or primary recovery of the petroleum is due to the energy initially present in the reservoir fluids. Gas in solution in the petroleum or hydrodynamic pressure provides this energy. Once the primary energy has been dissipated, a large fraction of oil may still remain in the reservoir as a result of the capillary nature of the reservoir rock. At this stage of production, water or gas may be injected into the reservoir to provide the energy to displace additional oil. This is an immiscible displacement and is referred to as secondary recovery. Improved or enhanced oil recovery results from the injection of special gases at miscibility pressures or of liquid chemicals, or the use of thermal energy. The injected fluids interact with the rock–oil system to create conditions favorable to oil recovery.

When the natural energy of the petroleum in the reservoir is no longer adequate to cause the petroleum to flow to the surface, artificial means are used to lift the petroleum to the surface. This can be done in a number of different ways, depending on the circumstance. Sucker-rod pumps and gas lift are the most common methods used.

Once the petroleum reaches the surface, it is processed before entering a pipeline. The purpose of surface processing of the petroleum is to provide the purchaser with a water-free, hydrocarbon-liquid–free natural gas or, in the case of a condensate or crude oil, to provide a water-free, emulsion-free, and solid-free liquid ready for processing at a refinery or gasoline plant.

The lead time required from project inception to the first production of petroleum may be anywhere from 5 to 10 years and require billions of U.S. dollars of investment. Reservoirs with large initial petroleum fluids in place are needed. The ultimate economics will depend on the rate of petroleum fluids recovery and the recovery efficiency of the petroleum from the reservoir.

1. DEFINITION OF PETROLEUM

Petroleum can be divided into five classifications: dry natural gas, wet natural gas, retrograde gas condensate, volatile oils, and black oils. Although this classification appears to be quite specific, there is a continuity of the phase state between that of a dry gas and a black oil. A dry natural gas consists primarily of methane with perhaps traces of ethane, propane, and heavier components. A wet natural gas also consists primarily of methane; however, it contains a higher percentage of ethane, propane, and butane with traces of pentane through hexane and heavier components. At surface conditions of temperature and pressure, a dry gas remains in a single phase, whereas a wet gas produces condensate (thus wet) consisting of propane through hexane and heavier components. In the reservoir, both dry and wet gases remain in a single gaseous phase.

Retrograde gas condensates most often are found as a single gas phase in the reservoir.

As the reservoir fluid is produced and the pressure in the reservoir drops below its retrograde dew point, liquid condensate precipitates within the pore space of the reservoir rock. Unless the condensate is very rich in intermediate hydrocarbons (i.e., ethane through hexanes), the liquid saturation seldom reaches the critical value where the condensate starts to flow.

Volatile oils are liquids in the reservoir. However, the reservoir temperature is very close to the critical temperature of the volatile oil, and once the reservoir pressure drops below the bubble-point pressure a large gas saturation forms very rapidly in the reservoir. Without doing a phase-behavior study, it is difficult in the field to distinguish a rich retrograde gas condensate from a volatile oil.

Black oils are found as liquid in the reservoir. The composition of the black oil is high in the C_{7+} fraction. The C_{7+} fraction characterizes the black oil. Black oils or crude oils are further broken down as being paraffinic, aromatic, naphthenic, and asphaltic. This means that in the case of a paraffinic crude the C_{7+} fraction consists primarily of paraffinic hydrocarbons. Penn Grade crudes are paraffinic. High–molecular-weight paraffinic hydrocarbons make excellent lubricating-oil stock. A barrel of Penn Grade crude oil may consist of as much as 60% lubricating-oil stock compared with an asphaltic crude, which may have as low as 10% lubricating-oil stock. If the crude is classified as aromatic, naphthenic, or asphaltic, it means that the C_{7+} fraction contains a predominance of aromatic, naphthenic, or asphaltic compounds.

Figure 1 is a plot of the C_{7+} concentration in petroleum against the initial gas/oil ratio (GOR) of the petroleum reservoir fluid (McCain, 1993). The gas/oil ratio is the number of standard cubic meters of gas that is liberated from a reservoir fluid per cubic meter of liquid or stock-tank oil at the surface. A GOR of infinity would be a dry gas whereas a GOR of less than 360 m^3/m^3 would be a black oil. As can be seen from the plot, there is a continuity between the gaseous and liquid states of matter. As the GOR increases, we approach volatile oil and retrograde gas condensates. In fact, these two classifications of reservoir fluid overlap on the plot. As the reservoir fluid or well-stream effluent becomes leaner in the C_{7+} fraction, the GOR increases. Wet gas can have a GOR as low as 9000 m^3/m^3 of oil. As in the case of volatile oils and retrograde gas condensates, retrograde gas condensate and wet gases can overlap. In summary, black oils are repre-

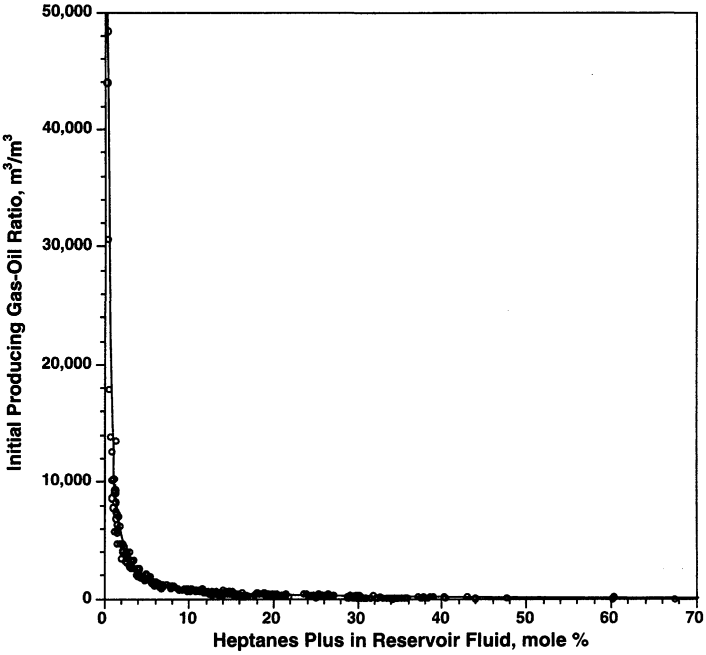

FIG. 1. Relationship between initial producing gas/oil ratio and heptanes-plus fraction in reservoir fluid (McCain, 1993).

Table 1. Classification of petroleum.

Classification	GOR (m^3/m^3)	CH_4 (%)	C_{7+} (%)
Dry gas	Infinity	>90	...
Wet gas	>9000	80–95	Trace
Retrograde gas condensates	530–9000	60–90	<12.5
Volatile oil	360–590	50–70	12.5–20
Black oil	<360	<45	>20

sented in the lower right of the figure and dry gases in the upper left. All other fluids are found in between with no sharp demarcation between these characterizations.

Table 1 is another way of illustrating this point. It shows the composition of petroleum under this classification system. Petroleum may also contain small amounts of nonhydrocarbons such as nitrogen, carbon dioxide, and hydrogen sulfide. Generally speaking, the concentration of nitrogen will range from 0 to 15 mole %, carbon dioxide from 0 to 4 mole %, and hydrogen sulfide from 0 to 6 mole % in petroleum. However, there are always exceptions. There are gas reservoirs that yield very high concentrations of both carbon dioxide and hydrogen sulfide. Table 2 illustrates some typical compositions of reservoir fluids.

2. PHYSICAL PROPERTIES OF PETROLEUM

The physical properties of petroleum reservoir fluids determine the calculation procedures that are used to predict the petroleum in place, the production performance, and the way the reservoir fluids are processed at the surface.

The five categories of petroleum reservoir fluids can best be described by their phase diagrams. The phase state of a reservoir fluid along with a knowledge of the rock and fluid properties is necessary to make reservoir performance calculations. The phase diagrams for these reservoir fluids are plots of pressure vs temperature that show the condition under which the fluid will exist as a gas, liquid, or both. Similarly, as shown in Fig. 1, phase diagrams also show a continuum between a dry gas and a black oil. The range of critical temperatures from that for a dry gas to that of a black oil can vary from

Table 2. Typical composition of reservoir fluids.

	Initial GOR (m^3/m^3)				
	Infinity Dry gas	27 660 Wet gas	1140 Gas cond	347 Volatile oil	145 Black oil
Component			Mole %		
N_2	...	0.30	0.88	1.55	...
CO_2	0.30	2.78	0.53	0.29	0.30
H_2S	...	...	9.42	...	...
C_1	98.50	94.41	67.98	51.97	41.30
C_2	0.87	1.44	6.21	11.72	4.93
C_3	0.17	0.34	2.37	9.23	3.85
C_4's	0.06	0.17	2.07	4.13	3.67
C_5's	0.02	0.08	1.21	1.82	2.36
C_6's	0.08	0.04	1.47	1.59	3.01
C_{7+}	...	0.44	7.88	17.70	40.58
	100.00	100.00	100.00	100.00	100.00
MWC_{7+}[a]	...	160	135	185	228
SGC_{7+}[b]	...	0.8190	0.7025	0.8231	0.8833
Reservoir T (°C)	...	126	120	173	76
Saturation pressure (kPa)	...	39 655	30 448	28 103	19 959

[a] Molecular weight of the C_{7+} fraction.
[b] Specific gravity of the C_{7+} fraction.

as low as −82.6 °C, the critical temperature of methane, to well over +600 °C for a black oil. The reservoir temperatures will range from as low as 32 °C to as high as 170 °C. Thus, in the case of a dry natural gas the reservoir temperature is considerably above the critical temperature of the natural gas, and in the case of a black oil the reservoir temperature is considerably below the critical temperatures of the black oil. In the case of a retrograde gas condensate, the reservoir temperature lies between the critical temperature and the cricondentherm or point of maximum temperature of the phase diagram.

The phase diagrams of natural gas will cover a much smaller temperature range because it consists of only very low–molecular-weight paraffin hydrocarbons when compared with a black oil that has in the mixture very large nonvolatile molecules. In going from a dry gas through wet gas, retrograde gas condensate, and volatile oil to black oil, the temperature range of the phase diagram increases dramatically.

2.1 Gas

2.1.1 Dry Gases Figure 2 shows the phase diagram of a dry gas. Since dry gas is primarily methane, the phase envelope is very small when compared with that of a black oil. The critical temperature of a dry gas will be in the range −82.6 to −20 °C, which is well below the reservoir temperature, line 1–2. The reservoir temperature is also well above the maximum temperature of the phase envelope. The fluid is a single-phase dry gas in the reservoir. Surface conditions are represented by point 3, also a single gaseous phase. As the reservoir is depleted, the pressure in the reservoir will drop from point 1 to point 2, the abandonment pressure.

The viscosity of the gas in the reservoir is low and its ability to flow will be primarily limited by the permeability of the rock matrix. Typically, the viscosity of dry gases at reservoir conditions will range between (0.013 and 0.030) × 10^{-3} Pa s.

2.1.2 Wet Gases The phase diagram of a wet gas is shown in Fig. 3. Here the reservoir temperature line and the pressure depletion path 1–2 in the reservoir lie much closer to the two-phase region of the phase diagram but still do not enter the two-phase region. The reservoir fluid remains in a single phase in the reservoir throughout the life of the field.

However, the surface condition of temperature and pressure lies within the two-phase region, point 3, and a low–specific-gravity liquid will condense at the surface. The viscosity of the wet gas at reservoir conditions of temperature and pressure will be in the range of (0.013 to 0.040) × 10^{-3} Pa s. The

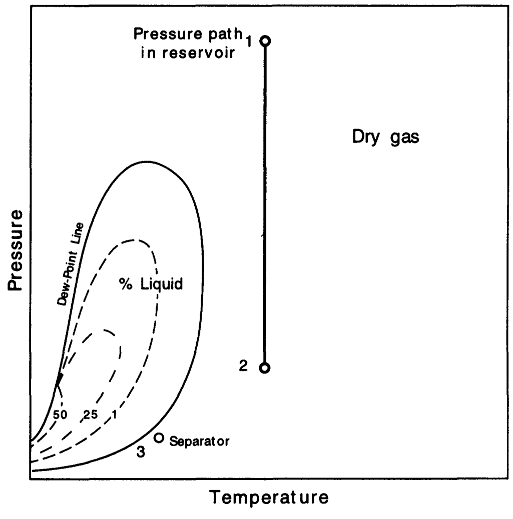

FIG. 2. Pressure–temperature diagram for dry gas reservoir.

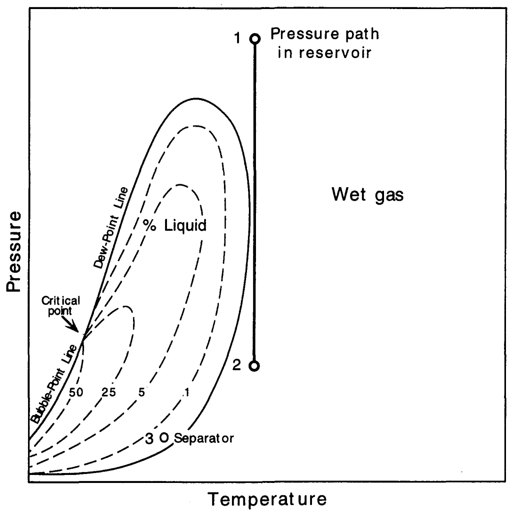

FIG. 3. Pressure–temperature diagram for wet gas reservoir.

stock-tank liquid or condensate produced from a wet gas is usually light yellow to water white and has a very low specific gravity in the range of 0.7389 to 0.8251. The specific gravity is calculated by comparing the density of the condensate with the density of water when both are at 15.56 °C (15.56/15.56 °C). During the 1920s and early 1930s, the condensate from a wet gas was used as gasoline to power oil-field vehicles.

2.1.3 Retrograde Gas Condensates In the case of a retrograde gas condensate, the reservoir temperature lies between the critical point and the point of maximum temperature, Fig. 4. At the original reservoir pressure, the gas exists as a single phase in the reservoir, point 1. As the fluid is produced, the reservoir pressure drops until the retrograde dew-point pressure is reached, point 2. At this point, liquid starts to condense in the reservoir. Generally, the liquid does not reach a saturation at which it flows. Upon further reduction in pressure, some of the liquid begins to revaporize; however, most of it is lost in the reservoir unless some enhanced recovery process is employed.

Point 3 represents the reservoir abandonment pressure. Point 4 represents the surface conditions of the separator. Retrograde gas condensate reservoirs generally range from 20 000 to 55 000 kPa and from 93 °C to as high as 170 °C. Fluid viscosity at reservoir conditions is in the range of (0.010 to 0.060) × 10^{-3} Pa s. The specific gravity of the stock-tank condensate is in the same range as the condensate of a wet gas, i.e., 0.7389 to 0.8251 at 15.56/15.56 °C. However, the condensate can range in color from water white through lightly colored to brown, orange, or green.

2.2. Oils

2.2.1 Volatile Oils In the case of a volatile oil, the reservoir temperature lies just below the critical temperature of the volatile oil. The critical temperature of a volatile oil is considerably lower than that of a black oil. When the reservoir pressure drops below the bubble-point pressure of the oil, a large gas saturation immediately forms in the reservoir. A pressure drop of (5–15) × 10^2 kPa may result in as much as a 50% gas saturation in the reservoir. The gas released in the reservoir is a retrograde gas and, consequently, a large part of the stock-tank liquid is produced from this retrograde gas. The stock-tank liquid has a specific gravity that is less than 0.8251 at 15.56/15.56 °C. The stock-tank liquid is usually orange to brown in color.

Figure 5 is the phase diagram for a volatile oil. Line 1–2–3 is at the reservoir temperature. Since the temperature is below the critical temperature, a volatile oil exists as a liquid in the reservoir, point 1. Point 2 is the

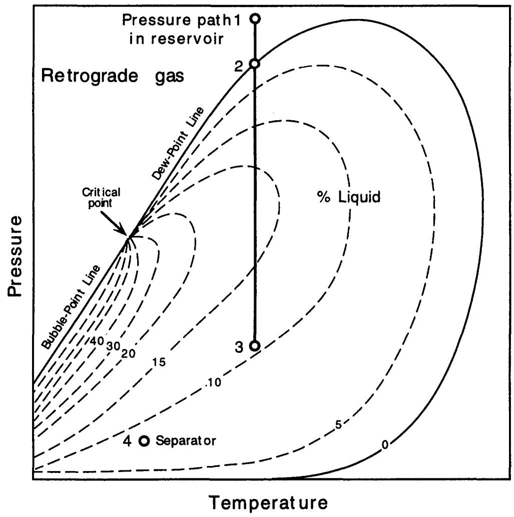
Pressure path 1 in reservoir
Retrograde gas
2
Dew-Point Line
Critical point
% Liquid
Bubble-Point Line
40
30
20
15
.10
3
5
0
4 Separator
Pressure
Temperature

FIG. 4. Pressure–temperature diagram for retrograde gas reservoir.

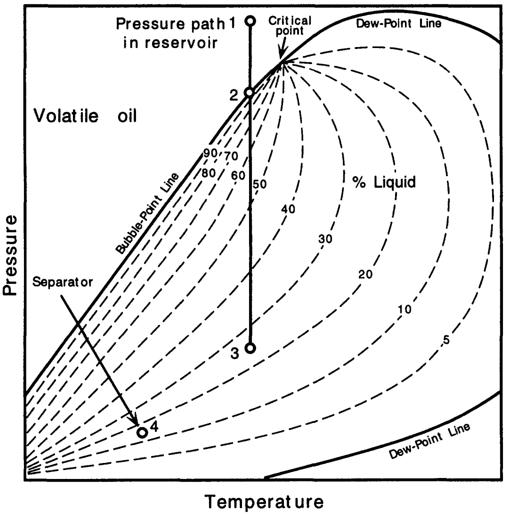
Pressure path 1 in reservoir
Critical point
Dew-Point Line
2
Volatile oil
90
70
80
60
50
40
30
20
10
.5
% Liquid
Bubble-Point Line
Separator
3
4
Dew-Point Line
Pressure
Temperature

FIG. 5. Pressure–temperature diagram for volatile oil reservoir.

bubble point of the oil at reservoir temperature. Point 3 represents the reservoir abandonment pressure. Point 4 represents the separator conditions at the surface. At point 1, the reservoir fluid is undersaturated, that is, no gas evolves as the pressure drops until the bubble-point pressure is reached.

2.2.2 Black Oils The phase diagram of a black oil is shown in Fig. 6. The path 1–2–3 at constant reservoir temperature represents the reduction in pressure that occurs during the production of the reservoir. Point 4 represents the surface separator conditions. When the pressure is above point 2, the reservoir fluid is undersaturated, that is, no gas comes out of the solution as the pressure drops. At the bubble-point pressure, point 2, gas will start to come out of the solution and form a free gas phase in the reservoir. As the pressure drops, more gas comes out of the solution and flows to the surface separator with the oil. The specific gravity of the stock-tank oil is usually greater than 0.8251 at 15.56/15.56 °C. Stock-tank oil viscosity can range from 1.0×10^{-3} Pa s to very heavy viscous crudes having viscosities of many Pa s. The heavier crudes are usually void of the intermediate light hydrocarbons C_4 through C_{15}. The color of the stock-tank liquids usually range from light to dark brown. Pennsylvania Grade crudes, in which the heavy ends are highly paraffinic, are dark fluorescent green in color.

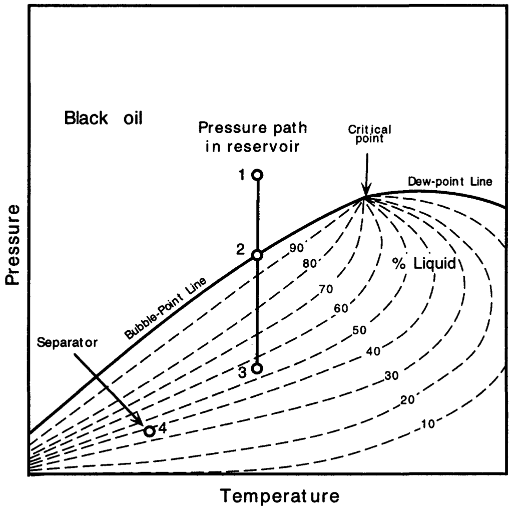

FIG. 6. Pressure–temperature diagram for black oil reservoir.

3. GEOLOGY AND OCCURRENCE OF PETROLEUM

Petroleum occurs in sediments deposited over geologic time. It is generally thought that petroleum originated in source beds consisting of mud and silt and was formed through the decomposition of plant and animal life. As the sediments were buried, the temperature and pressure increased, causing the organic matter to decompose thermally to form petroleum. The petroleum is expelled from the source rocks and migrates into layers of porous and permeable reservoir sands and carbonates. These sedimentary reservoir rocks are capped by an impermeable bed that "traps" the hydrocarbon and prevents further migration of the petroleum. The portion of the trap in which the petroleum occurs is called a petroleum reservoir.

3.1. Reservoir Traps

Reservoirs are generally classified according to the geological features under which they are formed. Cross sections of the different types of reservoir traps are presented in Fig. 7. Generally, traps are classified as structural, stratigraphic, or a combination structural/stratigraphic trap. In Fig. 7, example combination traps are referred to as stratistructural traps. Structural traps are a result of the deformation of the earth. Stratigraphic traps are caused by a change in the permeability and porosity in the reservoir bed.

3.2 Reservoir Rock Characteristics

Generally speaking, reservoir rocks are either sandstones or limestones; however, some reservoirs have been discovered in fractured shale. A knowledge of a reservoir's physical characteristics is of utmost importance in the management of the reservoir for maximum recovery of the petroleum it holds.

3.2.1 Porosity Porosity is the fraction of the total bulk volume of the rock not occupied by rock. Total porosity includes all the pore space in the rock, whereas effective porosity includes only the void space that is interconnected and through which fluids can flow. The pore space in a reservoir is filled with water and petroleum, that is oil and gas. Porosity is measured as a percentage of total

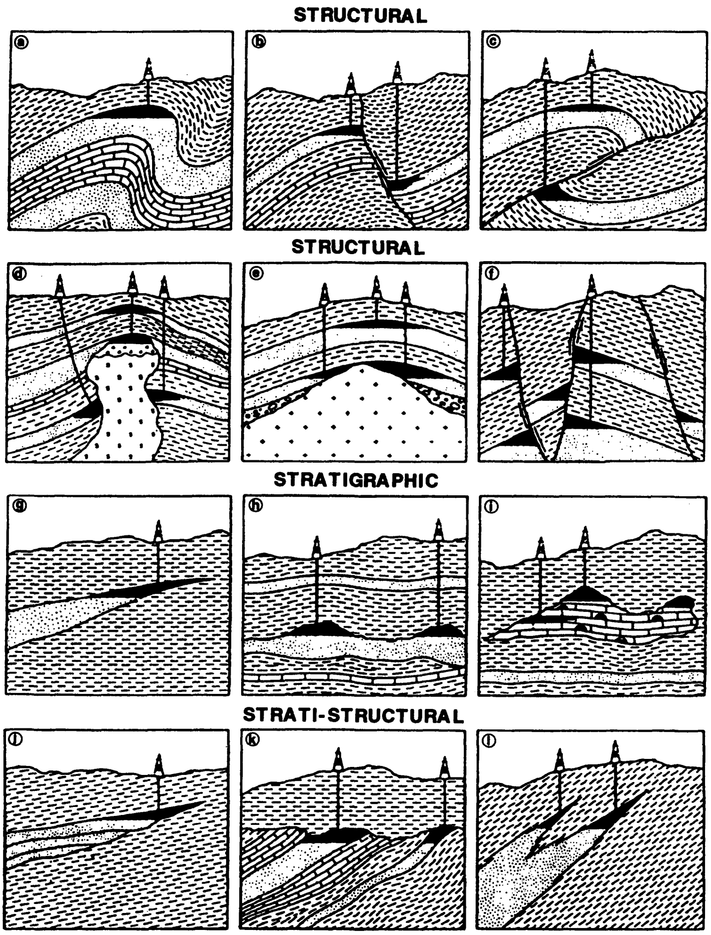

FIG. 7. Reservoir traps (courtesy Colorado School of Mines).

rock bulk volume, e.g., 20%–30% porosity. Porosity can vary from a few percent to as high as 35%–40%. Fluid content (saturation) of the pore space is measured as a percentage of the total pore space, e.g., 60% oil saturation and 40% water saturation.

3.2.2 Permeability Permeability is a measure of the ease with which a fluid can flow through the interconnected pore space of a rock. The unit of permeability is the square micrometer (μm^2), or 10^{-12} m^2. A permeability of 1 μm^2 will permit the flow of 1 m^3/s of fluid of 1 Pa s viscosity through an area of 1 m^2 under a pressure gradient of 10^{12} Pa/m. Or, stated another way, a permeability of 1 m^2 will permit the flow of 1 m^3/s of 1 Pa s viscosity through an area of 1 m^2 under a pressure gradient of 1 Pa/m.

Usually, there is more than one fluid flowing through the reservoir rock. Under these circumstances, we consider the relative permeability of the fluid, that is, the ease with which one fluid (of the two or more fluids present in the rock) will flow through the connecting pore spaces in the presence of the other fluids as compared with the ease with which the fluid will flow if it alone were present in the rock. Relative permeability then is the ratio of the permeability of a fluid flowing in the presence of other fluids to the permeability of the fluid if it alone were present in the rock.

3.2.3 Factors Influencing Rock Properties Sedimentary rocks are classified as sandstones, shales, limestones, and dolomites. The size and shape of the sedimentary materials as they were laid down, the presence of clay, the types of clay, and the presence of other minerals and the manner in which the sedimentary materials are cemented will affect the porosity and permeability of the reservoir rock.

4. DRILLING AND COMPLETION OPERATIONS

Drilling operations are conducted to provide a conduit between the reservoir rock and the surface for the production of oil and gas. Rotary drilling rigs are commonly used to drill wells. Drilling mud is pumped down the drill pipe and is returned to the surface in the annular space between the drill string and the earth, or a string of pipe called casing. The drilling mud must be capable of lifting the cuttings to the surface and controlling subsurface pressures encountered during the drilling process while minimizing damage to the reservoir rock.

The team responsible for the well gathers information during the drilling process to help evaluate the economic merit of the well. Drill cuttings are examined and identified. The drill cuttings may indicate zones of possible hydrocarbon accumulation. These zones are often tested with a tool run in the well on the drill string. This test, called a drill-stem test, gives an estimate of formation pressures, produced fluids (oil, gas, water), and potential flow rates for the well. Sometimes a larger sample of the formation is needed to estimate formation properties such as porosity and permeability. In these cases, a core is cut. A core is a solid, 7- to 10-cm-diameter cylindrical piece of the formation that may be up to 10 m or more in length. The core helps in characterizing the formation. The reservoir characterization helps the development team with reservoir management decisions such as well placement, completion methods, and secondary and enhanced recovery methods.

Well logs are another source of information for the evaluation team. Tools are run into a well on a cable to evaluate a section of the well bore before casing is set opposite this section. Logs run in this manner are called conventional open-hole wireline logs. These logs are run to estimate the porosity of the rock and the water saturation. The estimate of the water saturation permits an estimate of the hydrocarbon content of the formation. The porosity and water saturation of the rock give the evaluation team an estimate of the amount of hydrocarbon that may be present in the formation.

In some cases, it is desirable to have real-time porosity and saturation data rather than waiting for conventional wireline open-hole logs. In these cases, logging tools are run behind the drill bit as the well is drilled. These tools are called MWD (measurement while drilling) tools. When MWD tools are run, conventional wireline logs are still run.

The decision to abandon the well as a dry hole or to complete the well as a producer is made after all the available data are examined: the well logs, cores, and drill-stem tests. Petroleum must be present in an economic quantity as well as producible at economic rates before the decision is made to complete the well.

4.1 Directional and Horizontal Wells

Wells should be placed in the reservoir to maximize the economic recovery of oil and gas. In deep water, offshore fields can often only be economically developed from a single platform with multiple wells because of high platform costs. For this reason, offshore fields require the use of *directional* (or deviated) wells to develop the field efficiently and economically. In this case, the wells are drilled from an offshore structure and "steered" to the desired bottom-hole location selected by the development team of geologists, geophysicists, and petroleum engineers.

Another example of where directional wells are encountered is the Long Beach Unit near Los Angeles, California. Four man-made islands offshore act as "platforms" where over 1200 directional wells have been drilled to develop the Wilmington Field. A portion of the field lies beneath the city of Long Beach, California. The drilling rigs are soundproof and look like apartment buildings. Care is taken to survey each well bore so that the many well-bore paths do not accidently intersect.

Directional wells are drilled onshore whenever the bottom-hole objective cannot be reached by a vertical well.

When the well bore that penetrates the reservoir is horizontal, the well is referred to as a *horizontal* well. Horizontal wells are becoming an important method to improve the recovery from some reservoirs. Advances in reservoir characterization have identified reservoirs where horizontal placement of the well bore will enhance the drainage geometry and the ultimate recovery of oil from the well. Reservoirs having thin beds, vertical fracture systems, and compartmentalization are good candidates for horizontal drilling. Horizontal wells are more expensive to drill than a conventional or vertical well. Accordingly, the productivity from a horizontal well must be greater than a conventional well to justify the incremental cost of drilling. The incremental costs to drill a horizontal well may be as much as 30%–100% over the cost to drill a conventional vertical well. However, when the technology is appropriate, the productivity may be 2–6 times greater than a conventional completion (Fisher, 1988).

4.2 Casing and Wellhead Design

Pipe is placed in a well to provide a safe passage for the produced fluids and to protect formations (fresh water, for example) from contamination by the produced fluid; see Fig. 8. This pipe is referred to as casing and is cemented in the well for these reasons. Conductor pipe, set at 15–30 m, is cemented in the well, and provides a base for the surface wellhead equipment and an annular space for the drilling mud to be circulated to the surface.

After the conductor pipe is set, surface casing is set approximately 300 m below ground level to protect fresh-water sands. The surface casing is usually cemented completely from the surface to the bottom of the surface casing.

In deep wells, several strings of casing may be cemented in the well. The final casing that is set is referred to as the production casing or string. This casing is not set until after well logs are run and the decision is made to complete and produce the well. The production casing string and any intermediate strings may not be cemented all the way to the surface. Production casing is set to control well pressures and to protect formations during production. Figure 8 shows a production liner that has been set instead of running production casing from the surface to the producing formation. This is done to reduce costs.

When the reservoir fluid reaches the surface, it flows through the wellhead; see Fig. 9. Wellheads can vary greatly in design depending on the fluid pressure. The first element of a wellhead is the casing head, which helps support the weight of the casing or outer pipe. If there is more than one casing string, several casing heads may be used. Next is the tubing head, which is used to support the tubing string. Next comes the Christmas Tree; see Fig. 9. The Christmas Tree controls the

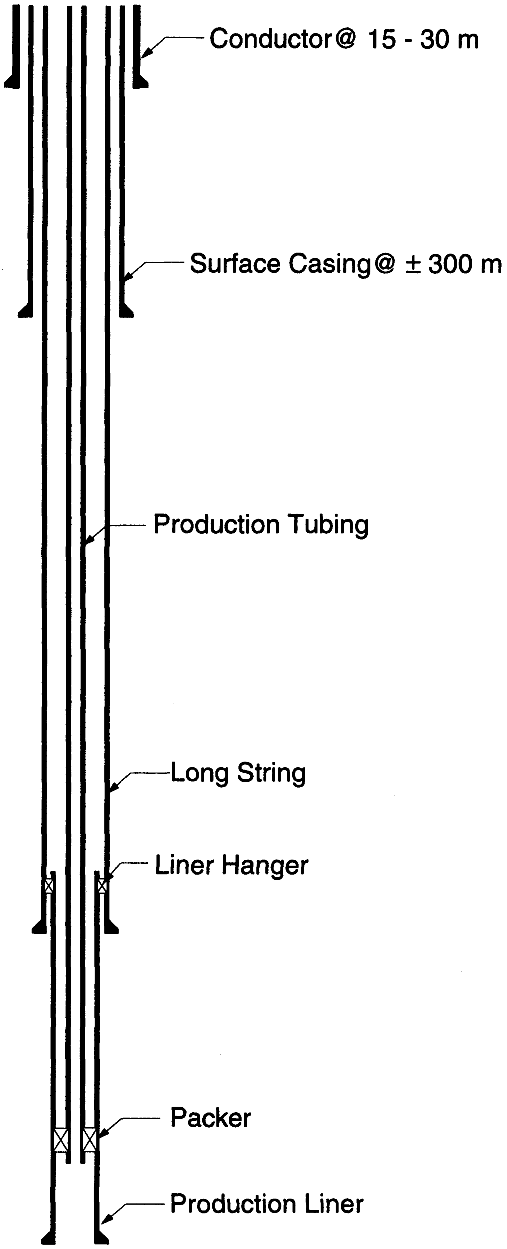

FIG. 8. Well completion design.

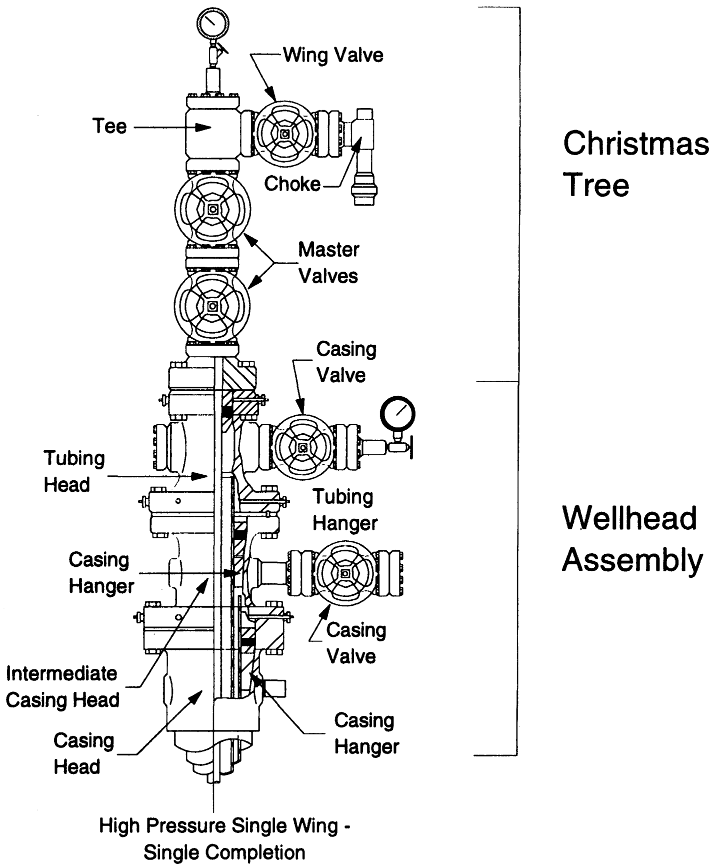
Wing Valve
Tee
Choke
Master Valves
Casing Valve
Tubing Head
Tubing Hanger
Casing Hanger
Casing Valve
Intermediate Casing Head
Casing Head
Casing Hanger
Christmas Tree
Wellhead Assembly
High Pressure Single Wing - Single Completion

FIG. 9. Well-head equipment and Christmas Tree.

flow of gas and oil from the well and is installed on top of the tubing head. The Christmas Tree design can vary depending on the fluid pressure. Pressure gauges are installed on the Christmas Tree. Valves on the "Tree" control the flow from the well. A choke is also part of the Christmas Tree. The choke is an orifice in the line that controls the rate of flow of the well fluid. A choke can be fixed or adjustable.

4.3 Completion Types

Once the decision is made to complete the well, the type of completion must be considered. Completion refers to the manner in which the well bore opposite the producing formation is designed or equipped. Some of the advantages of the various completion methods will be presented.

Perforated production casing is a common method of completing a well in areas where the producing formation is consolidated. Casing is set through the production interval and perforated with a perforating gun; see Fig. 10. The holes in the pipe can be made with bullets or with a shaped charge. The jet from a shaped charge has a velocity of approximately 9500 m/s and can penetrate through the pipe and cement and more than a meter into the formation. The holes are approximately 1.3 cm in diameter and provide the conduit between the formation and the well bore. Production tubing is run from the surface to the perforated interval. The production tubing is anchored in the well by a tool called a packer. The packer also isolates the annulus between the production tubing and the production casing; see Fig. 10.

In some cases, the last string of casing does not run from the surface to the bottom of the well. This last casing string is referred to as a *production liner*; see Fig. 8. A production liner is a piece of pipe that is hung from the next to last casing string. The liner is cemented in the well and perforated. The pri-

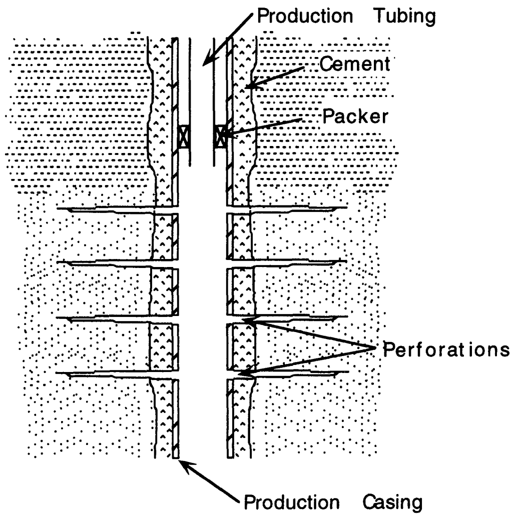

FIG. 10. Perforated casing completion.

mary advantage is a savings in the cost of casing since the last string does not need to be run from the surface to the bottom of the well. The disadvantage is a possible leak at the liner hanger.

In areas where the producing formation is unconsolidated and sand production is a problem, a *gravel-pack* completion is used to hold the sand in place and minimize the resistance to reservoir flow. In a gravel-pack completion, gravel is placed against the formation to hold the formation sand in place followed by a screen to keep the gravel in place. This method of completion is common in the Gulf Coast. New methods of completing unconsolidated formations are being tested today (1995).

A final completion method is an *open-hole* completion where the last string of casing stops above the productive zone. The well is drilled out below the casing into the zone of interest. In an open-hole completion, the restrictions to flow into the well bore are minimal compared with a perforated completion. This type of completion will only work in consolidated formations where the formation will not collapse.

4.4 Improving Well Performance

The value of the oil and gas is a function of how many barrels can be produced and the rate of production (deliverability). There are many formations where economic flow rates are not possible because the natural formation permeability is extremely low. *Hydraulic* fracturing is used to improve the formation flow rates. In these cases, a fluid is pumped down the well at high pressures and rates until the formation fractures. Once the formation has begun to fracture, a proppant (e.g., sand or bauxite beads) is pumped into the fractures. The pressure is then decreased, permitting the fractures to close on the proppant. The proppant-filled fractures provide a high-permeability conduit to the well bore for the produced fluids.

Acidizing a well may also improve the flow performance. Acidizing is the process of pumping an acid into the formation to improve the flow rate of the fluids from the formation into the well bore. Acidizing can be used to remove fine silica materials deposited during the drilling process. This fine silica material if *not* removed will restrict the flow of reservoir fluids. An acid can also be used to react with the formation rock and dissolve some of the rock. The formation may or may not be fractured during an acid job. Limestone and dolomite formations are good candidates for an acid treatment. An acid job in this case may include pumping 15% HCl (hydrochloric acid) that will react with the carbonate rock. There are acids that will react with quartz sandstone formations.

5. PETROLEUM RECOVERY MECHANISMS

Petroleum recovery can generally be classified into three steps: primary recovery, secondary recovery, and improved or enhanced recovery methods.

Primary recovery is that recovery in which the natural energy present initially in the reservoir is the main source of energy causing the displacement of petroleum to the producing well. The petroleum in the reservoir upon discovery is generally under pressure because of gas in solution in the petroleum liquid or hydrodynamic pressure or both. In the case of a gas reservoir, the gas itself is under pressure. When a well penetrates the reservoir, it provides the pressure sink to which the reservoir fluid flows. Because of the capillary nature of the reservoir rock and the flow of both oil and gas in the rock, oil recovery can vary from as low as 5% to as high as 50% of the liquid petroleum initially in place. On the average, primary recovery generally varies around 5%–30% of the oil initially in place.

In the case of a dry gas or wet gas reservoir, the production of the gas depletes the reservoir energy to the point where the production rate is no longer economical and the well is abandoned. In this case, the recovery factor varies from 60% to 90% of the gas initially in place.

In the case of a retrograde gas condensate reservoir, liquid condensate will drop out in the pore spaces of the reservoir rock as the pressure in the reservoir drops. Liquid dropout will reach a maximum and then as the pressure further declines some of the liquid condensate will revaporize. In the case of a rich retrograde gas condensate, the loss of liquid in the reservoir can be quite large when the abandonment pressure is reached. To prevent this liquid dropout, dry gas is quite often recycled into the reservoir to maintain the pressure and thus prevent the liquid from condensing.

In the case of a volatile oil reservoir, a small reduction in reservoir pressure below the bubble-point pressure of the volatile oil creates a large gas saturation. As a result, considerable liquid saturation can remain in the reservoir on abandonment by simple pressure depletion. Similarly, reinjection of dry gas into the reservoir to maintain pressure can reduce the loss of hydrocarbon.

After the primary energy in the reservoir has been dissipated, as much as 70%–95% of the oil may be left in the reservoir. The injection of water or gas into the reservoir to provide the energy to displace the oil to the producing well is referred to as *secondary recovery*. The displacement of the oil by gas or water is an immiscible fluid process, and because of the capillary effects, some of the oil is not displaced and a residual saturation of oil remains in the reservoir.

Improved or *enhanced oil recovery* results primarily from the injection of special gases at miscibility pressures or of liquid chemicals, or the use of thermal energy. The injected fluids interact with the rock–oil system to create conditions favorable to oil recovery by

1. lowering the interfacial tension between the oil and displacing fluids, thus reducing or eliminating capillary effects,
2. swelling the oil,
3. reducing the oil viscosity,
4. modifying the rock wettability, and/or
5. creating favorable phase behavior.

The viability of any improved oil recovery process depends on two factors, the unit displacement efficiency and the volumetric displacement efficiency. Unit displacement efficiency is the effectiveness of the displacing fluid displacing the oil in the *rock volume contacted,* whereas volumetric displacement efficiency refers to the fraction of total reservoir volume contacted by the displacing fluid.

Unit displacement efficiency is determined by the interfacial forces between the displacing fluids and displaced fluids. If the interfacial tension between the displacing fluid and displaced fluid is high, the capillary forces will be high and oil retention will be large. On the other hand, as the interfacial tension between the fluids is lowered the capillary forces will be lower. At an interfacial tension of zero, the fluids become miscible and the capillary forces are completely eliminated. At this point, the unit displacement efficiency is 100%.

The volumetric displacement efficiency is a function of mobility of the displacing and displaced fluids. The lower the viscosity of a fluid, the more mobile the fluid. A fluid of low mobility displacing a fluid of high mobility displaces the fluid in a pistonlike manner, whereas a fluid of high mobility displacing a fluid of low mobility will cause fingering and bypassing of the low-mobility fluid resulting in a lower volumetric displacement efficiency.

Improved oil recovery processes attempt to maximize both the unit displacement efficiency and the volumetric displacement efficiency. It is not always possible to do both at the same time. Application of improved oil recovery processes usually entails substantial risk because of their technological difficulties and front-loaded financial requirements.

5.1 Primary Recovery Mechanisms

5.1.1 Solution-Gas Drive In a solution-gas drive reservoir, the reservoir energy is derived from the light hydrocarbons, primarily methane, that are in solution in the oil in the reservoir. As the oil is produced and the reservoir pressure drops, light hydrocarbons come out of the solution and form a gas phase. The remaining gas in the solution plus the gas phase drives the oil to the well bore as the pressure declines. Since the gas is more mobile than the oil, the producing gas/oil ratio increases as the pressure decreases until a pressure is reached where the gas/oil ratio peaks and then starts declining. Up to this point, the reservoir energy is sufficient for the oil to flow up the tubing of the well. However, as the pressure decreases a pressure will eventually be reached where the well is put on artificial lift. Solution-gas drive reservoirs leave large quantities of oil in the reservoir. Recoveries are of the order of 10%–15% of the original oil in place.

5.1.2 Gas-Cap Drive Oil reservoirs can exist with a cap of gas over the oil. This gas cap is a source of energy along with the gas in solution in the oil to provide energy to move the oil to the well bore and lift it to the surface. As oil production proceeds, the gas cap expands into the reservoir originally containing the oil. The larger the gas cap, the smaller the rate of pressure drop as the oil is produced. This has the advantage of keeping gas in solution in the oil. As the gas cap expands and encroaches on a producing well, the gas/oil ratio may suddenly increase. Oil recovery will normally be in the range 20%–40% of oil originally in place.

5.1.3 Water-Drive Reservoirs In an active water-drive reservoir, the reservoir pressure declines relatively slowly as the oil is withdrawn from the reservoir and is replaced by water from an aquifer underlying the reservoir. The gas/oil ratio remains relatively constant because the pressure decreases very slowly, which keeps the gas from coming out of solution and forming a gas saturation. Water production from a well will increase gradually as the water encroaches into the reservoir. When water production gets too high, the well is abandoned since it becomes uneconomic to operate. Oil recovery from a water-drive reservoir will range from 30% to 75% of the oil originally in place, but is highly dependent on the depositional character of the reservoir as well as the drive mechanism.

5.1.4 Gravity Drainage In the case of gravity drainage, the fluids in a reservoir segregate in the reservoir according to their densities. Gravity drainage can play an effective role in the recovery of oil from a reservoir. By proper control of production rate, wells that are drilled into the bottom of the reservoir structure produce oil by gravity drainage of the oil into the well bore. Gravity drainage can be a very effective production mechanism, and recoveries can reach as high as 80%–90% of the oil originally in place.

5.1.5 Combination Drives Most often a combination of producing mechanisms exists in a reservoir under primary production. A common combination is one in which both water and gas-cap gas displace oil toward a producing well with perhaps some assistance from gas in solution in the oil. The degree to which each mechanism participates in the primary production of the oil is specific to the particular reservoir and the environment in which the reservoir exists. Since the water, oil, and gas are immiscible with one another, the capillarity of the reservoir rock traps large amounts of oil that are left behind once the primary production phase is completed. This oil provides the challenge to improved or enhanced oil recovery processes.

5.2 Secondary Recovery

5.2.1 Water Flooding The process of injecting water into a reservoir to displace oil was called a secondary recovery process because the process produced a second yield of oil after the field was depleted by primary production. Water flooding began accidentally in 1865 in the Pithold City area of western Pennsylvania. In the late 1800s, wells were being abandoned in the Bradford Field of Pennsylvania. In those days, there were no rules and regulations governing the abandonment of wells. Operators walked away from the wells, sometimes pulling casing, without plugging the wells. Fresh water from shallower horizons would enter the producing formation and stimulate production. These operations were named "Dump Floods." When operators realized that water was stimulating production, they deliberately started injecting water. By 1907, water injection had a significant impact on production in the Bradford Field. It was not until the 1930s that water flooding operations were reported in the midcontinent, i.e., Kansas, Oklahoma, and Texas. However, it was not until the late 1940s and early 1950s that water flooding operations became common practice. About 50% of the current oil production in the lower 48 states is attributed to water flooding.

Water flooding is usually conducted in the form of a pattern flood under different spacing conditions. Patterns can vary among line drives, 5-spot, 7-spot, and 9-spot: see Fig. 11. In some cases, a peripheral pattern is used, in which wells around the periphery of the field or lease are used for injection purposes. The type of spacing and pattern used depends on the location of existing wells, geometry of the reservoir, cost of drilling new wells, and the nature of the offset operations.

5.2.2 Gas Injection, Pressure Maintenance The injection of air or natural gas into a reservoir to stimulate oil recovery occurred well before the advent of water flooding. In the early 1900s, most secondary recovery projects involved gas injection. In gas-injection pressure maintenance operations, the gas can be directly injected into the oil-bearing formation. This is usually done through a pattern injection, that is, 5-spot or line-drive pattern, in order to obtain uniform distribution of the injected gas. Reservoir configuration must also be taken into consideration. In oil reservoirs having a gas cap, gas is often injected into the gas cap using wells in the crest of the structure. Such formations having structural relief and high permeabilities also benefit from gravity drainage. The number of wells required will depend on well injectivity and aerial coverage.

In the case of a retrograde gas condensate reservoir, economics may dictate the injection of a dry gas to maintain the reservoir pressure at or above the dew-point pressure to prevent the dropout of retrograde liquid condensate and thus the loss of liquid in the reservoir. If the retrograde gas condensate reservoir has been produced by pressure depletion, cycling dry gas through the reservoir will recover some of the liquid condensed in the reservoir by vaporization into the gas phase. Whether to cycle or not to cycle depends on the economics and the amount of liquid lost in the reservoir. If dry natural gas is unavailable, an inert gas such as a stack gas or nitrogen has been used. Nitrogen has

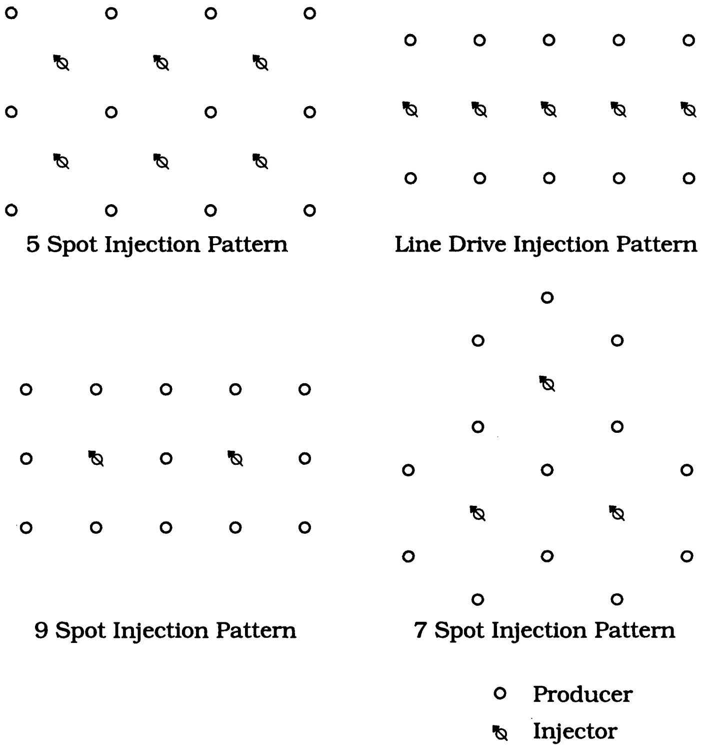
5 Spot Injection Pattern

Line Drive Injection Pattern

9 Spot Injection Pattern

7 Spot Injection Pattern

FIG. 11. Water flood patterns.

also been mixed with dry natural gas if not enough dry natural gas was available to maintain the desired pressure.

5.3 Improved Oil Recovery Mechanisms

Improved oil recovery results primarily from the injection of miscible gases and liquid chemicals or the use of thermal energy. Gases employed are hydrocarbon gases, carbon dioxide, or nitrogen. Liquid chemicals employed are polymer solutions, surfactant solutions, or hydrocarbon solvents. Thermal methods use steam or *in situ* combustion.

The lead time for the implementation of an improved oil recovery process from the time of first inception to the time of the first incremental volume of oil produced is anywhere from two to ten years depending on the complexity of the process and the reservoir to which it is to be applied.

The chemical requirements for improved oil recovery are also quite large, as shown in Table 3.

5.3.1 Polymer Flooding Polymer flooding is often referred to as improved water flooding. A high–molecular-weight polymer is added to water to lower the mobility of the water. Mobility of a fluid is defined as the permeability of the rock to the flowing fluid divided by the viscosity of the flowing fluid. The mobility of the water can be reduced either by increasing the viscosity of the water by adding polymer to the water or by reducing the permeability of the water to the rock by polymer adsorption and retention on the porous matrix of the rock. Improved volumetric sweep efficiency results. Polymer concentrations vary from 250 to 2000 ppm.

Table 3. Chemical requirements for improved oil recovery processes.

Process	Kg chemical/m^3 of oil recovered
Liquid chemicals	
Polymer flooding	
Polymer	5.5–8.5
Micellar polymer flooding	
Polymer	3.0–4.2
Surfactant sulfonates	20.0–48.0
Co-surfactant alcohols	2.8–28
Gases	
Carbon dioxide miscible flooding	1400–2700 standard m^3 CO_2/m^3 of oil recovery
Thermal	
Steam flooding	1–3 m^3 of water as steam/m^3 of oil recovered
In situ combustion	1400–5300 standard m^3 air/m^3 of oil recovered

There are a number of general requirements for the polymers used in polymer flooding. The polymer in solution should

1. reduce mobility of the water solution;
2. propagate through the reservoir rock without plugging;
3. experience minimum loss on the reservoir rock;
4. have adequate mechanical, chemical, thermal, and biochemical stability; and
5. be economical.

Polymer flooding should be initiated as early as possible during the course of a water flood while the oil saturation is still high. The higher the producing water/oil ratio at the start of the polymer flood, the higher the risk of failure. By initiating a polymer flood early, oil production is accelerated such that the cumulative water/oil ratio at the economic limit is much lower than if initiated later in the life of the flood.

There are two types of polymers used in polymer flooding, partially hydrolyzed polyacrylamides (PHPA) and biopolymers, the latter also known as polysaccharides and xanthan gums. PHPA is a synthetic polymer and is the most commonly used polymer for polymer flooding. Biopolymers are produced by fermentation. Biopolymers control mobility by viscosity increase only, whereas PHPA controls mobility by permeability reduction and viscosity increase.

The amount of polymer solution injection into a reservoir varies widely. Generally speaking, successful slug sizes are 15% of pore volume or greater. Incremental oil recovery from polymer floods over that of conventional water flooding may only be of the order of 2%–3% of the oil in place at the end of the water flood. Even so the process is still economic.

5.3.2 Miscible Flooding In miscible flooding, a fluid is injected into a reservoir that at reservoir pressures either is first-contact miscible with the reservoir crude oil or develops miscibility with the reservoir crude oil through a multicontact mixing process. The approach is applicable to medium and light crude oils. The fluids injected can be light hydrocarbon gases, nitrogen, flue gases, and carbon dioxide. Carbon dioxide is by far the most commonly used of the gases because of its unique properties. For a gas to become miscible with a crude through a multicontact mixing process, there is a minimum pressure requirement. This pressure is referred to as the minimum miscibility pressure, MMP. When carbon dioxide is injected into a reservoir at or above the MMP, light to intermediate components in the oil are transferred to the carbon dioxide, and when the enrichment of light to intermediate components in the carbon dioxide becomes large enough the gas at the leading edge becomes miscible with the oil and displaces the oil miscibly. As pointed out previously, the miscible displacement approaches 100% unit displacement efficiency. The MMP can be lowered by enriching the gas with intermediates. If, for example, the existing reservoir pressure is below the MMP for pure carbon dioxide, enrichment of the carbon dioxide with intermediate hydrocarbon will lower the MMP to the point that it will be at or below the reservoir pressure. In this case, there is a transfer of the intermediates into the crude oil, enriching the oil to a point where the injected rich gas becomes miscible with the crude. Although miscible displacement with a gas achieves high unit displacement efficiency, the mobility of the gas is higher than that of the crude oil and results in low volumetric displacement efficiency, i.e., fingering and bypassing of oil. To overcome this problem, alternate injection of water with carbon dioxide to improve the carbon dioxide mobility has been practiced with limited success.

Operating problems with carbon dioxide miscible flooding can be severe if the proper precautions are not taken. Corrosion, scaling, and asphalt precipitation are some of the problems encountered.

Carbon dioxide forms a highly nonideal solution with crude oil, resulting in swelling of the crude and in the lowering of the crude viscosity. Both of these effects enhance oil recovery.

The magnitude of additional oil recovery by CO_2 miscible flooding after water flooding is generally in the range of 8%–15% of the oil in place at the start of the flood; however, there are exceptions.

Large volumes of carbon dioxide are required for oil-field flooding. Gross carbon dioxide/oil ratios have varied from as low as 900 to as high as 4600 standard m^3 of carbon dioxide per m^3 of stock-tank oil recovered. Carbon dioxide cost will determine the cutoff. Sources of carbon dioxide can be varied: that occurring naturally in reservoirs, byproduct of ammonia plants, power-plant stack gas, and manufactured carbon dioxide. Reservoirs that contain carbon dioxide are presently being used to supply carbon dioxide for oil-field flooding, primarily because they can supply the volumes and rates at reasonable costs. Produced gases containing high concentrations of carbon dioxide are often recycled.

Of all the enhanced oil recovery processes, carbon dioxide miscible flooding is one of the most frequently applied.

5.3.3 Thermal Methods Thermal recovery methods can be classified into two broad categories: that in which heat is injected from an external source and that in which heat is generated *in situ*. Of the various possibilities, only steam injection and forward or concurrent combustion have proven viable commercially.

5.3.3.1 Steam Injection. Steam injection can be broken down into two different processes, steam flooding or continuous steam injection and cyclic steam injection or stimulation. About 72 187 m^3 of oil per day were produced in the United States in 1992 through steam injection (Moritis, 1994). The great majority of this oil is produced in California.

When steam is used, thermal expansion, viscosity reduction, and steam distillation are the principal mechanisms by which oil is recovered. Steam distillation is an important mechanism. The lighter the crude oil, the more important the distillation effect becomes. One characteristic that makes steam so useful as an injection fluid is that it carries a large amount of heat per unit mass compared with other fluids.

Steam flooding is generally applied to heavier crude oils ranging in specific gravity from 0.9000 to 1.000 and viscosities from 0.10

Pa s to several thousand Pa s. The viscosities of heavy crude oils are very sensitive to temperature. In continuous steam flooding, the steam moves out into the reservoir until it condenses, heating the rock and oil to steam temperature. The condensed water is produced, and an expanding steam zone efficiently displaces the crude oil. Generally, the equivalent of about one-third of the additional oil recovered is consumed to generate the required steam. However, because of environmental drawbacks, natural gas is now the preferred fuel to generate the steam.

Cyclic steam injection is a single-well operation where steam is injected into a producing well for a limited time. The well is then shut in for a short period of time before it is put on production. The well is produced anywhere from a few months to a year before the cycle is repeated. Each cycle produces a decreasing amount of oil. Sometimes only a few cycles are employed, and other times as many as 15 cycles are employed before the production rate becomes so low that it is not economic. Steam cycling works best for thick sands, viscous oils, and small well spacings. Cyclic steam recoveries are generally 10%–25% of the oil in place compared with 50%–60% by steam flooding.

5.3.3.2 In Situ Combustion The *in situ* generation of heat involves injection of air or an oxygen–nitrogen mixture into a reservoir and burning part of the crude oil in the form of coke deposited on the sand grains just ahead of the moving combustion front. The combustion zone moves in the same direction as the flow of fluids. A clean sand is left behind the combustion front. The temperature in the combustion zone ranges from 315 to 650 °C. The combustion-zone thickness is perhaps 10–20 cm. The *in situ* combustion process is ideal in that an undesirable fraction of the oil (heavy ends) is consumed as fuel in the form of coke, a clean sand is left behind, and the heat generated is used as efficiently as possible. Oil recoveries for successful projects can range from 60%–90% of the oil in place at the time the project was initiated.

Basically, three physical processes are taking place simultaneously, but not independently of one another. These processes are chemical reaction, heat transfer, and fluid flow. *In situ* combustion is initiated by simultaneously injecting air- or oxygen-enriched gas into a well and heating the well bore opposite the formation to the ignition temperature. Occasionally, in the case of a highly reactive crude oil, spontaneous ignition will take place. Well-bore heaters are usually electrical or gas burners.

When applicable, the *in situ* combustion process is conducted in sandstone reservoirs. *In situ* combustion is generally not applicable to limestone or dolomite reservoirs, although there have been a few exceptions.

A variation of *in situ* combustion is wet combustion, also known as COFCAW (Combination of Forward Combustion and Waterflooding). Water is injected either with the air or intermittently. The process takes advantage of the latent heat of vaporization of water to recover heat energy stored behind the combustion front. However, one needs to be careful, as injecting too much water can extinguish the combustion front.

The Tulare formation of the South Belridge Field has undergone both *in situ* combustion and steam drive, and as a consequence an excellent comparison can be made between these two processes in this reservoir (Gates *et al.*, 1978; Gates *et al.*, 1975). *In situ* combustion was much more energy efficient than that of steam, using only about one-quarter of the surface energy compared with steam. From an environmental standpoint, steam emits about 2.5 times as much flue gas as *in situ* combustion. Well failures were as bad for steam as for *in situ* combustion. The question then arises as to why *in situ* combustion was not used more in this formation. The answer lies in the fact that steam was probably applied first and was an established improved oil recovery process in this field.

5.3.4 Micellar-Polymer Flooding Micellar-polymer flooding is a fluid-injection process wherein a phase-stable, optically clear (transparent or translucent) solution of oil, water, and one or more surfactants, along with electrolytes or salts, is injected into a formation and is displaced by a solution of polymer. The polymer solution is in turn displaced by water. Oil and water are displaced ahead of the micellar solution, and a stabilized oil and water bank develops. The micellar solution (microemulsion) is the key to the process. The micellar solution displaces the oil and water either miscibly or at very low interfacial tensions and thus has a high unit displacement efficiency. In addition, the mobility of the micellar solution can be controlled, resulting in high volumetric displacement efficiencies. The process is applicable over a wide range of reservoir conditions. Micellar-polymer flooding is usually applicable if water flooding is applicable because of the mobility relationships. The polymers used are usually partially hydrolyzed polyacrylamide. The surfactants used are petroleum sulfonates and alcohols. The alcohol is referred to as a co-surfactant.

Because of the cost of the micellar solution, only a limited amount is injected. Oil recovery by micellar-polymer flooding can be over 60% of the oil in place at the time the project was initiated.

The current price ($18–$20 per barrel) of crude oil makes micellar-polymer flooding uneconomic. Micellar-polymer flooding is front-loaded financially; that is to say, it takes a large initial investment to initiate a project and the maximum financial exposure is quite high. Unless the price of crude oil goes up to the $35–$45 per barrel range in 1994 dollars, micellar-polymer flooding will not be a factor in improved oil recovery.

5.3.5 Alkaline-Surfactant-Polymer Flooding A relatively new flooding process has been developed in recent years and shows considerable promise as a result of a number of field pilot operations. Many oils are reactive with alkali and produce surfactants that lower the interfacial tension between the oil and brine. However, the optimum alkali concentration to produce this reduction in interfacial tension is quite low. By adding a low concentration of a surfactant to the system, a synergism results wherein the interfacial tension is considerably lower than alkali or surfactant acting by themselves. Polymer is also added to the system to insure the best possible mobility control. Alkalis used range among caustic soda, soda ash, sodium orthosilicates, and trisodium phosphate. The polymer generally used is partially hydrolyzed polyacrylamide. Surfactants used are usually high–molecular-weight organic sulfonates. It has been shown in a limited number of field tests that alkali surfactant polymer (ASP) flooding can recover just as much oil as micellar polymer flooding but at considerably less cost per barrel of oil recovered. In 1994 dollars, the costs range from $2 to $3 per barrel of incremental oil recovered (Surkalo *et al.*, 1993). The design of ASP flooding systems is quite involved, requiring detailed knowledge of reservoir geology, and fluid and rock chemistry.

6. PRODUCTION OPERATIONS

Artificial lift is installed to increase production from naturally flowing wells and to produce from wells when the natural energy in the reservoir is no longer adequate to lift the oil and gas from the well. The following sections discuss natural flow and the four major methods of artificial lift used in the oil and gas industry: gas lift, sucker-rod pumps, electrical submersible pumps, and hydraulic pumps.

6.1 Natural Flow

The ability of a well to flow naturally is a function of the formation pressures and deliverability, well-head pressures, tubing size, and depth. Except in the case of a very strong aquifer that is providing pressure support as fluids are produced from the reservoir, the pressure of the reservoir will decline with production. As the pressure declines, the energy is no longer sufficient to lift the fluids from the well bore. At this time, an external means of lifting the fluids from the reservoir is considered. The goal is to match the ability of the formation to deliver fluids at a specified well-bore pressure with a method of artificial lift that will add sufficient energy to lift the fluids to the surface. Wells may be placed on artificial lift before natural flow ceases. Economics will govern when it is viable to place the wells on artificial lift.

6.2 Artificial Lift

The selection of an artificial lift method depends on several factors. Capital costs, operating costs, reliability, and salvage value are major considerations when comparing the various alternatives for artificial lift. Flexibility to handle changing production conditions over the life of the well should also be considered. A "cradle to grave" economic analysis should be performed when selecting artificial lift methods.

Approximately 10% of the wells on artificial lift in the United States are *gas lifted* (Clegg *et al.*, 1993). Gas lift is a common method of artificial lift offshore. In gas lift, an external source of gas is injected as deep as possible into the production tubing to augment the formation gas: see Fig. 12. In this sense, gas lift is an extension of natural flow. The injection gas lightens the fluid column in the production tubing and lowers the flowing bottom-hole pressure at the formation. The effect is to increase the formation flow and extend "natural flow." In cases where flow rates cannot be established by the continuous injection of an external source of gas, the well may be produced by the intermittent injection of high volumes of gas. In these cases, the well is alternately produced and shut in. During the shut-in periods, formation fluids accumulate in the well bore. High-pressure gas is then injected into the tubing below the slug of formation fluid to displace the formation fluids to the surface.

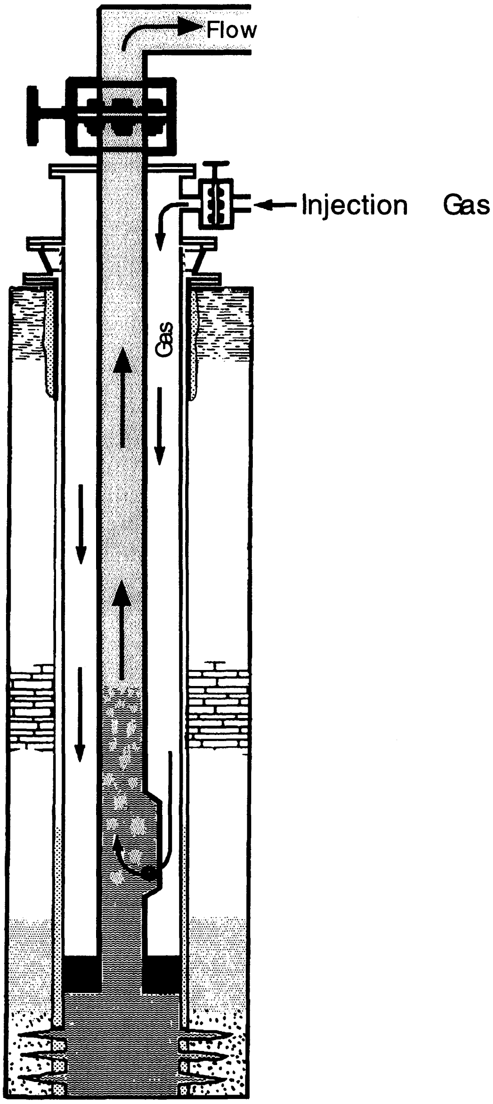

FIG. 12. Continuous gas lift.

The *sucker-rod pump* is the most common method of artificial lift: see Fig. 13. Approximately 85% of the wells on artificial lift in the United States are lifted with sucker-rod pumps. Sucker-rod pumps operate on the principle of positive displacement. The reciprocating motion of the down-hole pump is generated by a beam pumping unit at the surface and transmitted to the down-hole pump through steel rods called sucker rods. Part of the rod string may be made of fiberglass rods.

Sucker-rod pumps are easy to operate, and this is one reason for their popularity. Where space and weight are limiting factors, sucker-rod pumps are not a good choice. Sucker-rod pumps are rarely a good choice for offshore.

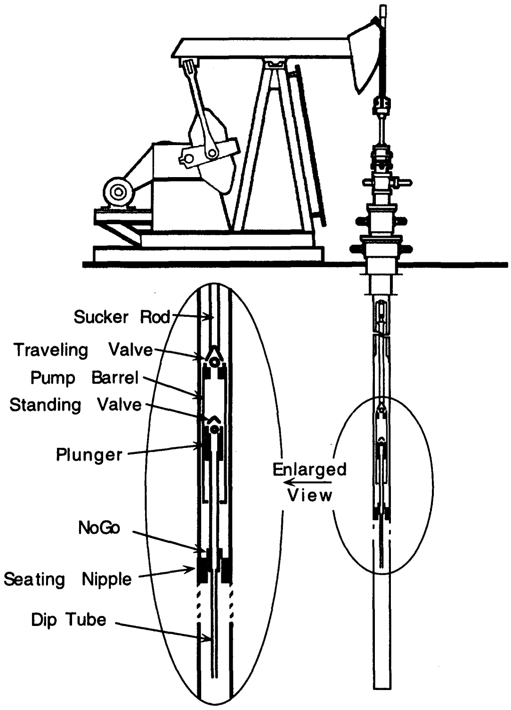

FIG. 13. Sucker-rod pump.

Depth of lift, especially for high-volume wells, is also a limiting factor for sucker-rod pumps.

Electrical submersible pumps (ESPs) are centrifugal pumps installed in the well on the end of the tubing and submerged in the well fluid. The down-hole components consist of an electric motor, motor protector, gas separator, pump, and power cable. The power cable, which is strapped to the tubing, transmits power from the surface to the down-hole motor.

ESPs are best suited for high-volume applications and are a common method of artificial lift where the reservoir is waterflooded. ESPs make up approximately 4% of wells on artificial lift in the United States (Clegg *et al.*, 1993).

There are two basic types of *hydraulic pumps* used in the petroleum industry, the positive displacement pump and the jet pump. In both cases, the pump is installed down hole to add energy to the produced fluid. Hydraulic pumps make up less than 2% of the wells on artificial lift in the United States (Clegg *et al.*, 1993).

6.3 Field Processing

When a reservoir fluid reaches the surface and flows through the wellhead, the reservoir fluid is further processed before entering a pipeline. In the case of a dry gas, the surface processing facilities can be quite simple, and in the case of a gas condensate or black oil they can be quite extensive. The objective of surface processing before the fluid enters the pipeline is to provide the purchaser with a water-free, hydrocarbon-liquid–free natural gas, or a clean condensate or crude oil.

6.3.1 Natural Gas Processing at the Wellhead Natural gas wells can produce both reservoir water and water that is in the form of vapor in the gas. Upon reaching the surface, water vapor condenses when the temperature is lowered. The presence of free water can cause corrosion and form solid hydrates of natural gas that can plug lines and valves. If the water problem is not too severe, the gas may be heated in an indirect heater to prevent water condensation and keep the temperature of the gas above the hydrate-forming conditions. Water produced from the formation is trapped in a surface separator. Gas sold to a pipeline must meet certain specifications in regard to its water content. To meet this specification, the gas is dehydrated, that is, the water is removed in equipment called a dehydrator. There are different types of dehydration equipment. One common unit uses glycol to remove the water from the gas.

6.3.2 Wet Gas, Gas Condensates, and Volatile Oil Processing at the Wellhead In the case of a wet gas, gas condensate, or volatile oil, the reservoir fluid must be separated at the surface into a liquid condensate and a gas stream. The reservoir fluid enters a separator that allows the gas and condensate to separate from one another. If free water is produced, three phases will exist in the separator. Quite often two separators are used and the pressure controlled on the separators to maximize the amount of liquid that enters the stock tank, which is at atmospheric pressure. The gas phase is then treated in a manner similar to that of a dry natural gas well.

6.3.3 Black Oils Processing at the Wellhead When black oils are produced from a reservoir, they may include solids such as sand and scale along with water. This is generally referred to as basic sediments and water. The produced water may occur in two forms, free water or an oil-in-water emulsion. The amount and degree of emulsification depend on the nature and amount of surfactants that are present in the crude oil. Some oil-field emulsions are difficult to break, and heat and chemicals must be used to separate the water from the oil. The well-stream effluent first flows into a separator where the gas, oil and/or emulsion, and free water are separated. The gas exits from the top of the separator, the oil and/or emulsion from the middle, and the free water from the bottom. The separators come in different designs, both horizontal and vertical.

The emulsion layer goes to a heater-treater where both heat and chemicals are applied to the emulsion. The emulsion breaks and water settles to the bottom of the vessel and clean oil floats to the top from where it is then piped to a stock tank. The stock-tank oil is then ready for the pipeline.

7. PETROLEUM RESERVES AND ECONOMICS

Industry and government agencies are interested in estimates of the future availability of petroleum. Unfortunately, consistent

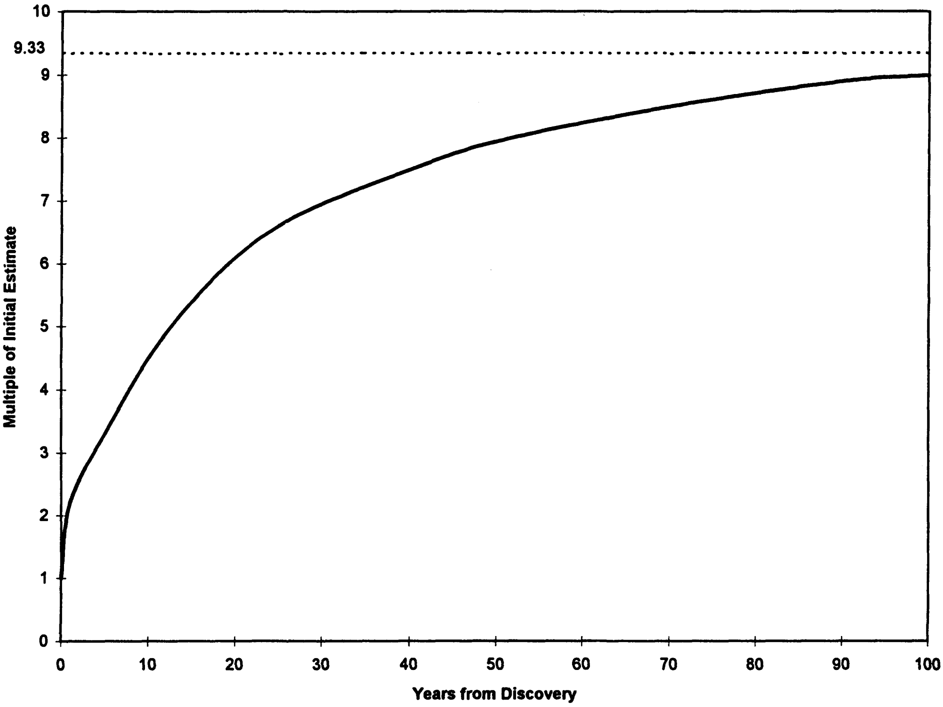

FIG. 14. Reserve appreciation factors for oil (Energy Information Administration, 1990).

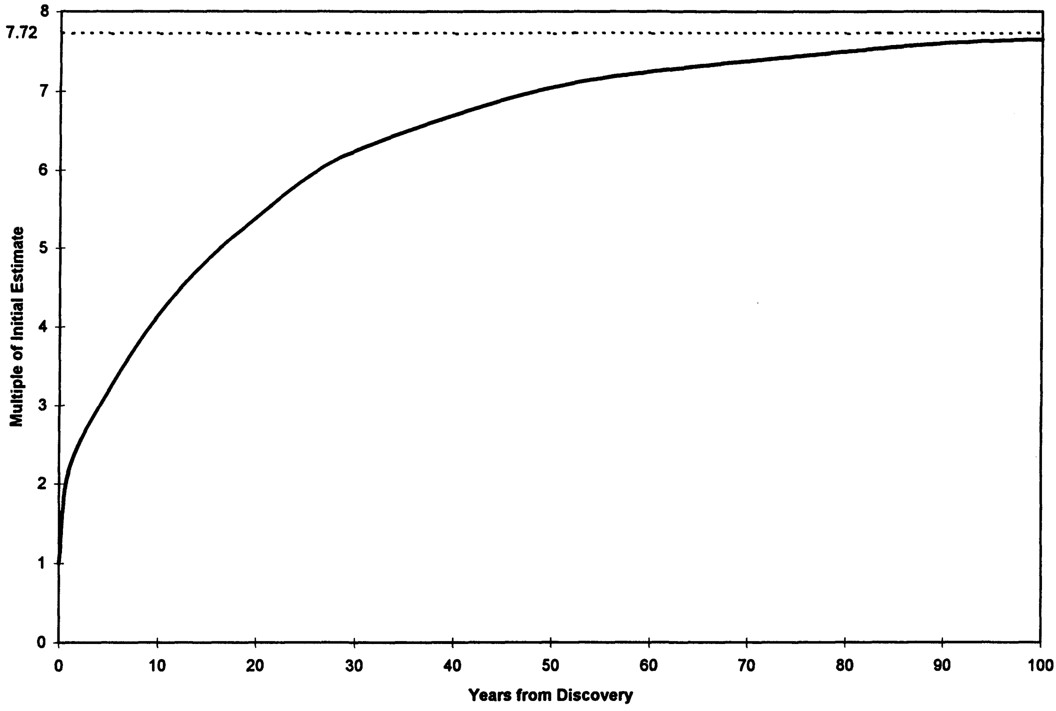

FIG. 15. Reserve appreciation factors for gas (Energy Information Administration, 1990).

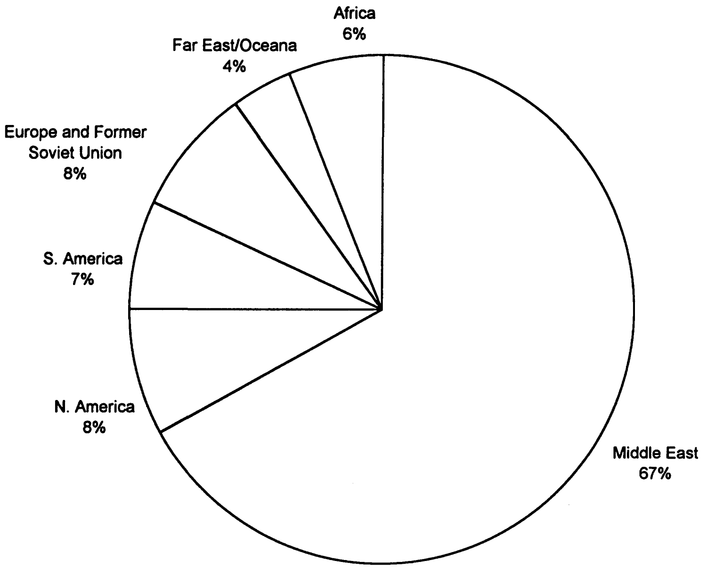

Total World Oil Reserves =15.8 x 10^{10} m^3

FIG. 16. Total world oil reserves.

terminology is not in use worldwide to describe future estimates of petroleum resources.

7.1 Definitions

The distinction between *resource base, resource,* and *reserves* is helpful in understanding resource assessment studies. The petroleum resource base is all the oil and gas in the earth regardless of the economic or technical feasibility of extraction. The term petroleum resource is used to denote the quantity of petroleum that is recoverable under *specified* economic and technological conditions (Lovejoy, 1965).

Petroleum reserves are the oil and gas that *remain* to be recovered under *current* economic conditions using *current* technology. This definition implies certain conditions:

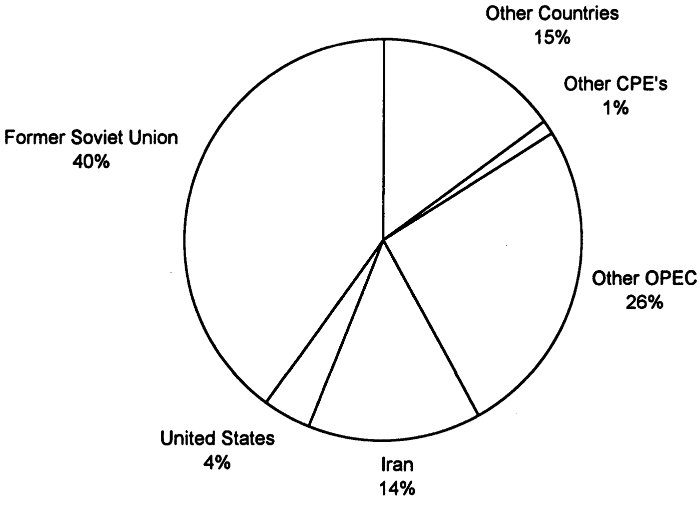

Total World Natural Gas Reserves = 124.0x 10^{12} m^3

Note: Other CPE's includes former, evolving and current Centrally Planned Economies

FIG. 17. Total world gas reserves.

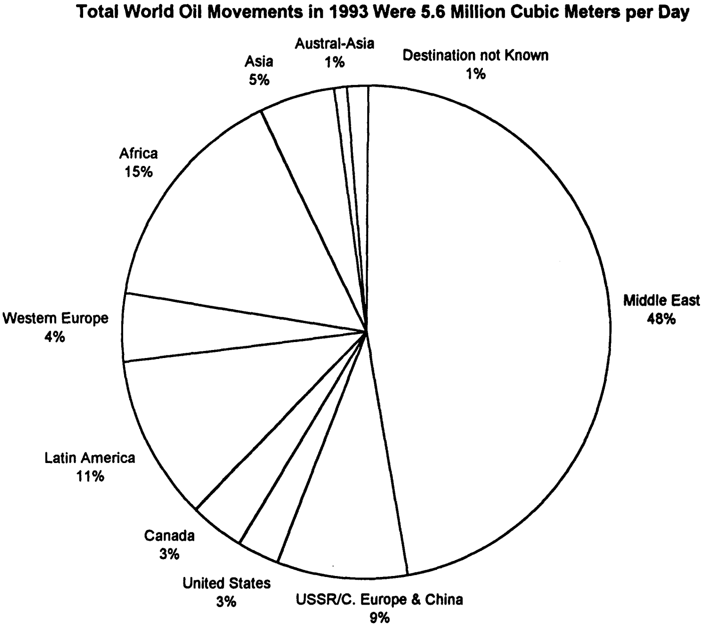

FIG. 18. World oil movements—exports.

First, in order to be called reserves, oil and gas must be physically producible. Second, it must be economic to produce the oil and gas. Third, since the resource has not been produced and is inaccessible, reserves cannot be measured but must be estimated. Finally, since reserves are remaining, there must be a time line associated with the estimate.

Hydrocarbons that have been produced up to a given date are referred to as cumulative recovery or production. The remaining reserves on this date are referred to as esti-

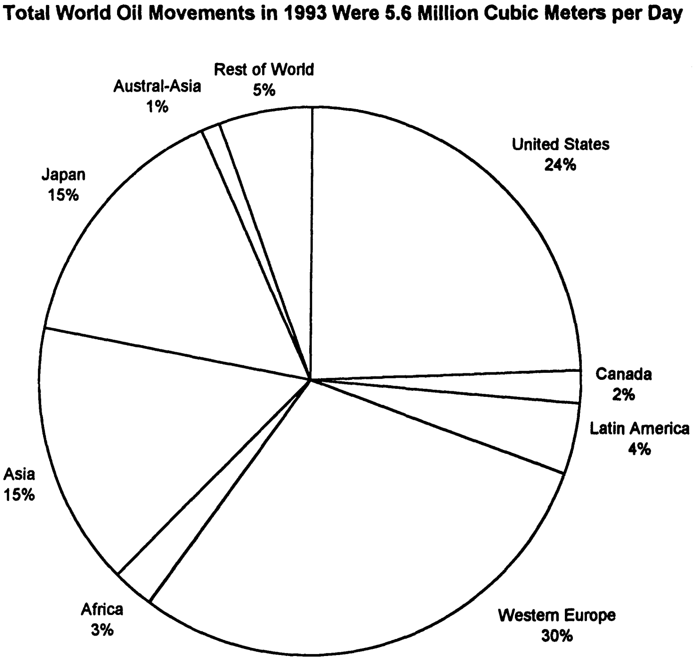

FIG. 19. World oil movements—imports.

mated remaining reserves (ERR). On a given date, the estimated ultimate recovery (EUR) would be equal to the cumulative recovery plus the estimated remaining reserves (Thompson, 1985).

7.2 Classification of Reserves

Classification systems commonly group reserves by the degree of uncertainty associated with the estimate. Three categories are proved, probable, and possible reserves. Proved reserves are those quantities that the estimator is *reasonably certain* are recoverable from known reservoirs under current economic and operating conditions. If the evidence does not support the proved reserves category, the reserves may be classified as probable. In cases where the evidence does not support either the proved or probable categories, the reserves may be classified as possible reserves.

7.3 Petroleum Resource Assessment

Since reserves are estimated, not measured, quantities, at the time of discovery there may be considerable uncertainty in the estimated ultimate recovery for a field. Historically, on an aggregate basis, estimates for the ultimate recovery for fields increase (or appreciate) from the initial estimate. The Energy Information Administration (EIA) has developed a database that contains the year of discovery, the initial estimate for ultimate recovery, and subsequent revisions in reserves on a field basis for the United States (Energy Information Administration, 1990). This information has been used to develop appreciation factors (also referred to as growth factors) that can be multiplied times the initial estimate for ultimate recovery to obtain a better estimate of the final ultimate recovery. The appreciation factors for oil and gas are shown in Figs. 14 and 15. The appreciation factors calculated from this database represent reserves additions from field extensions, new reservoirs, infill drilling, improved technology, enhanced recovery techniques, and longer productive life from an increase in oil and gas prices.

An interesting approach for estimating reserves growth for gas reservoirs has been developed by the Bureau of Economic Geology (Fisher, 1988). This method relates the depositional environment and the reservoir compartmentalization to the amount of reserves growth that may be expected from infill drilling gas reservoirs.

7.4 Petroleum Economics

On a macroeconomic scale, the changes taking place in Eastern Europe and the former Soviet Union and the worldwide distribution and movement of petroleum are examples of the global nature of the petroleum industry. The distribution of world oil and natural gas reserves is shown in Figs. 16 and 17 (Anon., 1991). The movement of oil, imports and exports, around the world is shown in Figs. 18 and 19 respectively (American Petroleum Institute, 1994).

On a microeconomic scale, actual or perceived supply disruptions are reminders of the capital-intensive and risky nature of the petroleum business and the long lead time from project inception to first production. Major projects, for example, often require 5 to 10 years and billions of United States dollars to complete.

The risk is often put into three categories: economic risk, political risk, and technological risk (Garb, 1988). The technical uncertainty includes the estimates of hydrocarbon volumes and the recovery rates. The economic risk includes prices and capital costs for drilling and development. Political risk includes local and national taxes, environmental regulations, and international instability. The petroleum industry is placing more emphasis on decision methods that help quantify the risks, including sensitivity analysis, portfolio analysis, and Monte Carlo simulation.

GLOSSARY

Christmas Tree: The control valves, pressure gauges, and chokes assembled at the top of the well to control the flow of oil and gas after the well has been drilled and completed.

Drilling Mud: The liquid that circulates through the well bore during rotary drilling. It brings cuttings to the surface, cools the bit and drill stem, controls subsurface pressure, and deposits a mud cake on the wall of the borehole to prevent fluid loss to the formation.

Enhanced Oil Recovery (EOR): The injection of special gases at miscibility pressures or of liquid chemicals or the use of thermal energy into a reservoir to interact with the rock–oil system to create conditions favorable to oil recovery. Factors such as lowering interfacial tension, lowering viscosity, changing rock wettability, and creating favorable phase behavior are EOR techniques.

Gas-Oil Ratio (GOR): The ratio of the volume of gas produced with a unit volume of oil measured at standard temperature and pressure.

Permeability: A measure of the ease with which a fluid can flow through the interconnected pore network of a porous rock.

Porosity: The fraction of the total bulk volume of the rock not occupied by rock solids.

Primary Recovery: Oil production from a reservoir in which the natural energy present initially in the reservoir fluids is the main source of energy causing the displacement of petroleum to a producing well.

Reservoir: A subsurface, porous, and permeable rock formation in which oil and/or natural gas have accumulated.

Retrograde Condensation: The condensation of a light hydrocarbon liquid that occurs when the pressure on a gas containing intermediate to heavy hydrocarbons is reduced at constant reservoir temperature.

Saturation Pressure: When a liquid is at its bubble-point pressure or a gas is at its dew-point pressure, the fluids are said to be at their saturation pressure.

Secondary Recovery: After the primary energy in a reservoir has been dissipated, the injection of water or gas into the reservoir to provide energy to displace the oil immiscibly to the producing well is referred to as secondary recovery.

Specific Gravity: For a liquid, the ratio of the weight of a unit volume of the liquid to that of a unit volume of water both measured at 15.56 °C and at atmospheric pressure. For a gas, the ratio of the weight of a unit volume of the gas to that of a unit volume of air under ideal gas conditions. This is equivalent to the ratio of the molal average molecular weight of the gas to the molal average molecular weight of air (28.96).

Well Completion: The activity and procedure necessary to prepare a well for production of oil and gas. The procedure depends on the characteristics of the producing formation. Examples of completion types include open hole, sand exclusion (gravel pack), and perforated casing.

Well Log: A systematic recording of data as a function of well depth such as a driller's log, mud log, electrical well log, or radioactivity log in order to obtain properties of the formations.

Well Stimulation: The treatment of a producing formation to obtain increased production of oil or gas, e.g., hydraulic fracturing, sand control, and acidizing.

Works Cited

American Petroleum Institute (1994), *Basic Petroleum Data Book*, Washington, DC: American Petroleum Institute, Vol. XIV, No. 3, Sec. IX, Tbl. 10j.

Anonymous (1991), "World Wide Production Report," *Oil Gas J.* (December 30, 1991), 43–88.

Clegg, J. D., Bucaram, S. M., Hein, N. W., Jr. (1993), Recommendations and Comparisons for Selecting Artificial Lift Methods, *J. Pet. Tech.* **45**, 1128–1131, 1163–1167.

Energy Information Administration (1990), *The Domestic Oil and Gas Recoverable Resource Base: Supporting Analysis for the National Energy Strategy*, Washington, DC: Energy Information Administration, pp. 15–21.

Fisher, W. L. (1988), *An Assessment of the Natural Gas Resources Base of the United States*, Austin, TX: Bureau of Economic Geology, p. 66.

Garb, F. A. (1988), "Assessing Risk in Estimating Hydrocarbon Reserves and in Evaluating Hydrocarbon-Producing Properties," *J. Pet. Tech.* **40**, 765–778.

Gates, C. F., Jung, K. D., Sierface, R. A. (1978), "Insitu Combustion in the Tulare Formation, South Belridge Field, Kern County, California," *J. Pet. Tech.* **30**, 798–806.

Gates, C. F., Brewer, S. W. (1975), "Steam Injection into the D and F Zone, Tulare Formation, South Belridge Field, Kern County, California," *J. Pet. Tech.* **27**, 343–348.

Lovejoy, W. F., Homan, P. T. (1965), *Methods of Estimating Reserves of Crude Oil, Natural Gas, and Natural Gas Liquids*, Baltimore, MA: Johns Hopkins, pp. 1–5.

McCain, W. D., Jr. (1993), "Chemical Composition Determines the Behavior of Reservoir Fluids," *Pet. Eng. Int.* **65** (20), 18–25.

Moritis, G. (1994), "EOR Dips in U.S. But Remains a Significant Factor," *Oil Gas J.* **92** (39), 51–79.

Surkalo, H., Pitts, M. J., Wyatt, K. (1993), "Design and Field-Implementation of Alkaline-Surfactant-Polymer Chemical Enhanced Oil Recovery Systems," in *Proceedings of the 7th European Symposium on Improved Oil Recovery (October 27–29, 1993), Moscow, Russia,* Vol. 2, Moscow: Russian Interbranch Scientific and Technological Complex "Oil Recovery," pp. 308–316.

Thompson, R. S., Wright, J. D. (1985), *Oil Property Evaluation*, Golden, CO: Thompson-Wright Associates, pp. 4-1–4-6.

Further Reading

Allen, T. O., Roberts, A. P. (1978), *Production Operations*, Vols. 1 and 2, Tulsa, OK: Oil & Gas Consultants International Inc.

Anderson, R. O. (1984), *Fundamentals of the Petroleum Industry*, Norman, OK: Univ. of Oklahoma Press.

Archer, J. S., Wall, C. G. (1986), *Petroleum Engineering—Principles and Practices*, Norwell, MA: Graham and Trotman.

Arnold, K., Stewart, M. (1986), *Surface Production Operations*, Vol. 1, Houston: Gulf Publishing Company.

Arnold, K., Stewart, M. (1989), *Surface Production Operations*, Vol. 2, Houston: Gulf Publishing Company.

Baker, R. (1985), *A Primer of Offshore Operations*, Austin, TX: Petroleum Extension Service, The University of Texas at Austin.

Bradley, Howard B. (Ed.) (1987), *Petroleum Engineering Handbook*, Dallas: Society of Petroleum Engineers.

Brock, J. (1986), *Applied Open-Hole Log Analysis*, Houston: Gulf Publishing Company.

Clark, N. J. (1969), *Elements of Petroleum Reservoirs*, Henry L. Doherty Series, Dallas: Society of Petroleum Engineers.

Craft, B. C., Hawkins, M. (1991), *Applied Petroleum Reservoir Engineering*, 2nd ed., Englewood Cliffs, NJ: Prentice-Hall.

Gerding, Mildred (Ed.) (1986), *Fundamentals of Petroleum*, 3rd ed., Austin, TX: Petroleum Extension Service, The University of Texas at Austin.

Golan, M., Whitson, C. H. (1986), *Well Performance*, Dordrecht, the Netherlands: D. Reidel Publishing Company.

Gray, F. (1986), *Petroleum Production for the Non Technical Person*, Tulsa, OK: PennWell Publishing Co.

Katz, D. L., Cornell, D., Kobayashi, R., Poettmann, F. H., Vary, J. A., Elenbaas, J. R., Weinaug, C. F. (1959), *Handbook of Natural Gas Engineering*, New York: McGraw-Hill.

McCain, W. D., Jr. (1990), *The Properties of Petroleum Fluids*, 2nd ed., Tulsa, OK: PennWell Publishing Company.

Morton-Thompson, Diana, Woods, Arnold M. (Eds.) (1992), *Development Geology Reference Manual*, Tulsa, OK: The American Association of Petroleum Geologists.

North, F. K. (1985), *Petroleum Geology*, Boston: Univin Hyman.

Poettmann, Fred H. (Ed.) (1983), *Improved Oil Recovery*, Oklahoma City, OK: Interstate Oil and Gas Compact Commission.

Skinner, D. R. (1982), *Introduction to Petroleum Production*, Vols. 1, 2, and 3, Houston: Gulf Publishing Company.

PHASE EQUILIBRIA

Arthur D. Pelton, *Centre de Recherche en Calcul Thermochimique, Ecole Polytechnique, Montréal, Québec, Canada*

	Introduction	297
1.	**Phase Diagrams of Single-Component Systems**	298
2.	**Binary Phase Diagrams—Descriptive**	299
2.1	Systems with Complete Miscibility in All Phases	299
2.1.1	Minima and Maxima in Two-Phase Regions	301
2.2	Eutectics	302
2.3	Miscibility Gaps	303
2.4	Polymorphism	304
2.5	Intermediate Compounds	305
2.5.1	Congruently Melting Compounds	305
2.5.2	Incongruently Melting Compounds	305
2.5.3	Nonstoichiometric Compounds	305
2.6	Topology of Binary Phase Diagrams	306
2.7	Phase Diagrams of Systems with Liquid-Crystalline Phases	307
3.	**The Gibbs Phase Rule**	308
4.	**Ternary Phase Diagrams**	309
4.1	Ternary Composition Triangle	309
4.2	Ternary Space Model	309
4.3	Polythermal Projections of Liquidus Surfaces	311
4.4	Ternary Isothermal-Isobaric Sections	311
5.	**Other Kinds of Phase Diagrams**	312
6.	**Thermodynamic Origin of Phase Diagrams**	314
6.1	Models of Solutions—Regular Solution Theory	315
6.2	Thermodynamic-Phase Diagram Optimization	317
7.	**Glass Formation**	318
8.	**Order-Disorder Transitions**	319
9.	**Interfacial Effects**	319
	Glossary	319
	Works Cited	320
	Further Reading	321

INTRODUCTION

In general, a chemical system will consist of several phases, solid, liquid, and gaseous, of different compositions. The study of phase equilibria and phase transformations is central to materials science, chemical engineering, geology, and many other disciplines.

A phase diagram is a graphical representation of the conditions of temperature, pressure, composition, etc. for which equilibrium among the phases of a system is established. A system at equilibrium is one that has achieved a state of minimum free energy and is unchanging even over an infinite amount of time. Although departures from equilibrium will occur in any real system, a knowledge of the equilibrium state under a given set of conditions is the starting point for the understanding of most processes.

We restrict ourselves to phase diagrams with pressure, temperature, composition, and, in a few examples, partial pressures as the axes. However, other variables, such as electrical, magnetic, and gravitational fields, can also affect phase equilibria, and phase diagrams with these as axes could also be drawn.

In Secs. 1 and 2, we describe phase diagrams of one-component and two-component systems without rigorously justifying the rules governing their construction. This justification follows, however, in Sec. 3 on the Gibbs phase rule. The thermodynamic origin of phase diagrams is discussed in more detail in Sec. 6. Some simple examples of the relation between the phase diagram of a system and the structure and bonding of the phases at the atomic level are also presented in Sec. 6.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. PHASE DIAGRAMS OF SINGLE-COMPONENT SYSTEMS

A single-component system consists of only one chemical substance which does not dissociate. The phases present may be solid, liquid, or gaseous. The most common type of phase diagram of a one-component system is a *P-T* diagram in which the axes are temperature and pressure as shown for the system H_2O in Fig. 1. Volume is not an independent variable since, for any phase in a single-component system, the volume is uniquely given, as a function of *P* and *T*, by the equation of state. Figure 1 can be considered to be an orthogonal projection, upon the *P-T* plane, of the nodal lines of the *P-V-T* surface.

In Fig. 1(a), regions in which the entire system is solid, liquid, or gaseous are labeled. The coexistence of two phases occurs along the lines of the diagram. For example, ice transforms to liquid water at a constant temperature of 0 °C at a pressure of 101.3 kPa (1.0 atm), and the liquid and gas coexist in equilibrium at this pressure at 100 °C.

If a previously evacuated container is partially filled with water, sealed, and heated, the pressure in the container will follow the two-phase boundary between the liquid and vapor regions of Fig. 1 as long as the liquid and vapor phases are both present. This line thus gives the *vapor pressure* of H_2O as a function of *T*. Solid, liquid, and vapor can only coexist at one point, the *triple point*, which, for H_2O, occurs at 0.0098 °C and 0.611 kPa. At pressures and temperatures below this point, ice sublimes directly to vapor.

At low pressures, ice has a hexagonal crystal structure. This form of ice is known as ice-I. At very high pressures, other *polymorphs* (i.e., other crystal structures) of ice, known as ice-II, ice-III, etc., are observed. The phase diagram at high pressures is shown in Fig. 1(b). Again, the diagram consists of single-phase areas, two-phase lines, and three-phase triple points.

The (liquid + vapor) two-phase line in Fig. 1(a) terminates at a *critical point*. For H_2O, the *critical temperature*, T_c, and *critical pressure*, P_c, are 374 °C and 22.3 MPa. At this point, all the properties of the two phases become identical. Above the critical point there is no longer a distinction between liquid and vapor. As discussed above, if liquid and vapor are heated in a sealed container, the pressure follows the two-phase line. As the critical point is approached, the densities of the two phases approach each other, becoming identical at the critical point. At the molecular level, the behavior of a system very near its critical point is complex. For a discussion, see CRITICAL PHENOMENA.

The slopes of the two-phase lines in Fig. 1 are given by the *Clapeyron equation*,

$$dP/dT = \Delta H/T\Delta V, \tag{1}$$

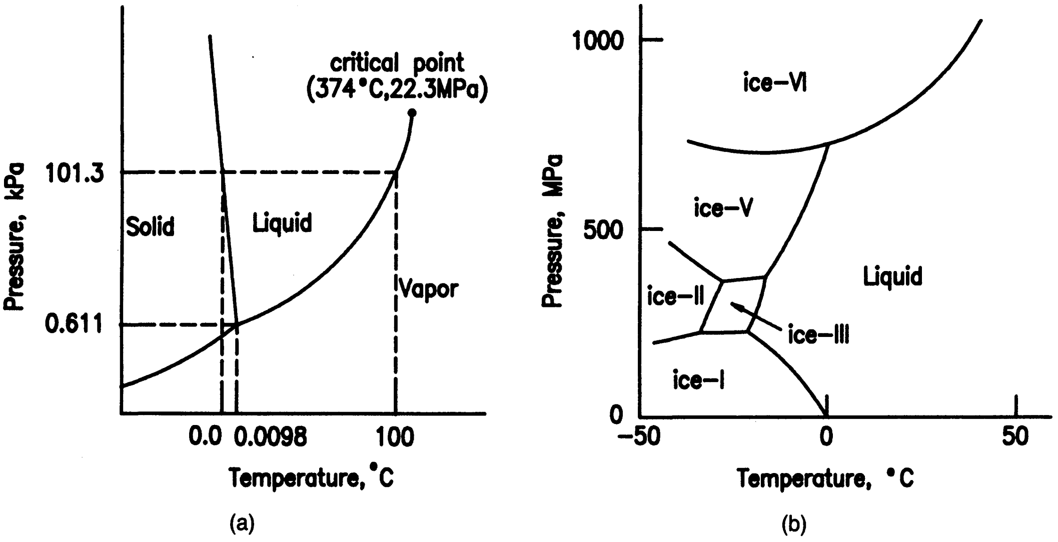

FIG. 1. *P-T* phase diagram of water at (a) low and (b) high pressures.

where ΔH is the heat (enthalpy) change for the transformation of one mole of substance from one phase to the other, ΔV is the corresponding molar volume change, and T is the temperature in kelvins. For the (liquid + vapor) line, ΔH is the heat of vaporization ($\Delta H > 0$) and ΔV is the volume change accompanying the vaporization of one mole of liquid. Hence, ΔH is very large and so, from Eq. (1), the slope of the line is small. For the (solid + liquid) boundary, however, ΔV of transformation from solid to liquid is small. Consequently, this line is very steep. It is interesting to note that H_2O is unusual in that $dP/dT < 0$ for the (solid + liquid) boundary. This is due to a negative ΔV of melting. (Ice floats on water.) For most substances, ΔV of melting is positive, and the (solid + liquid) boundary has a positive slope.

2. BINARY PHASE DIAGRAMS—DESCRIPTIVE

A binary system is one with two components. The precise definition of "component" is given in Sec. 3. However, the components of a system are usually evident. In this section we discuss binary temperature-composition (*T-X*) and pressure-composition (*P-X*) phase diagrams.

2.1 Systems with Complete Miscibility in All Phases

A *T-X* phase diagram at constant pressure $P = 101.3$ kPa for the methanol-water system is shown in Fig. 2. The abscissa is the composition, expressed as mole fraction of water, X_{H_2O}. Note that $X_{CH_3OH} = 1 - X_{H_2O}$. Phase diagrams are also often drawn with the composition axis as weight percent.

At a temperature and overall composition given by a point in the region labeled "vapor" in Fig. 2, the system, at equilibrium, will be entirely in the vapor phase. In the region labeled "liquid," the system, at equilibrium, will consist entirely of one miscible liquid phase. For example, it can be seen from Fig. 2 that pure water transforms from a liquid to a gas at 100.0 °C at $P = 101.3$ kPa (1.0 atmospheres), while the "normal boiling point" of methanol is 64.6 °C.

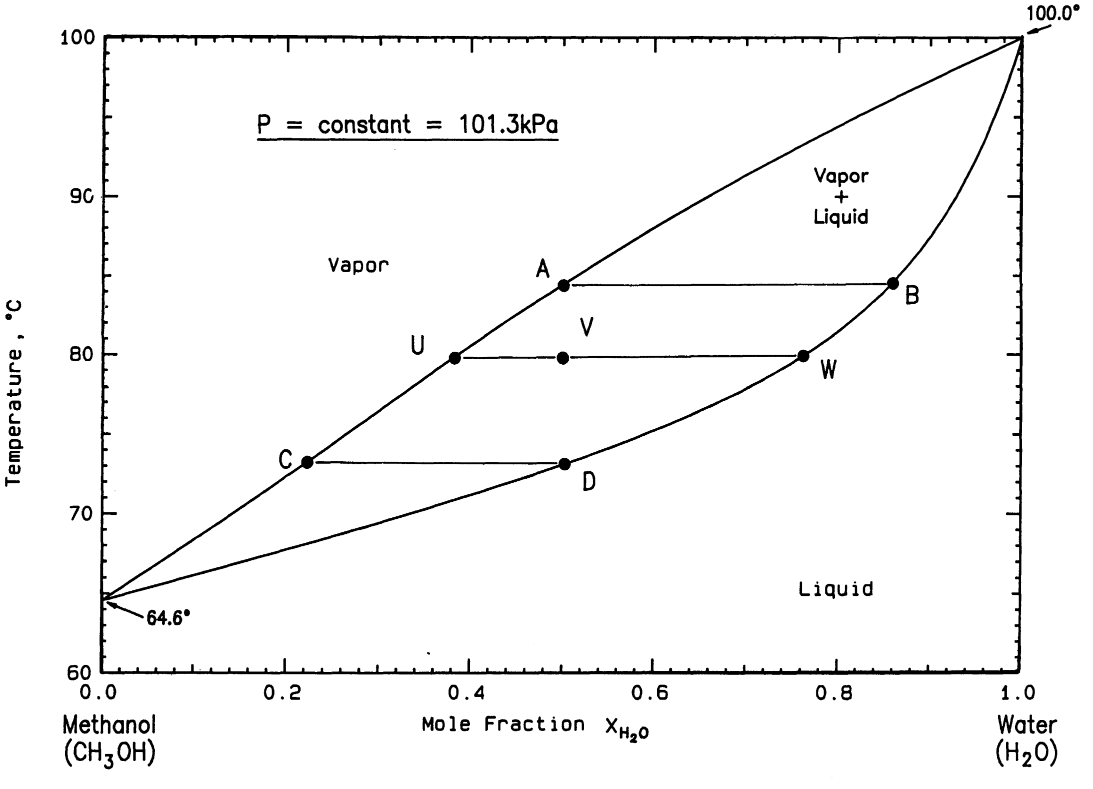

FIG. 2. *T-X* phase diagram at constant $P = 101.3$ kPa of the methanol-water system.

Suppose that a system at an overall composition of $X_{H_2O} = 0.5$ is held at 80 °C in a container at $P = 101.3$ kPa (i.e., at point V in Fig. 2). At equilibrium, two phases will be observed, a vapor phase of composition U ($X_{H_2O} = 0.385$) and a liquid solution of composition $W(X_{H_2O} = 0.76)$. The line UW is called a *tie-line* or *conode*.

If we were to add some water to the system to bring the overall composition to, say, $X_{H_2O} = 0.65$, keeping T constant at 80 °C, we would observe that the compositions of the vapor and liquid phases remain unchanged at points U and W, although the relative amount of the liquid phase would increase. From a simple mass balance, one can derive the *lever rule* for binary systems: (moles of vapor)/(moles of liquid) = VW/UV, where VW and UV are the lengths of the line segments.

Hence, at 80 °C and 101.3 kPa for a system with an overall composition $X_{H_2O} = 0.50$, the amounts of vapor and liquid at equilibrium are in the molar ratio $(0.76 - 0.5)/(0.5 - 0.385) = 2.26$. When the composition axis of a phase diagram is in weight percent, then the lever rule gives the weight ratio of the two phases.

Suppose that a gas with composition $X_{H_2O} = 0.5$ is slowly cooled at a constant pressure of 101.3 kPa. When the temperature has decreased to 84.5 °C the system is at point A on the upper phase boundary in Fig. 2. At this temperature (the "dew point") the first liquid condensate appears with a composition at point $B(X_{H_2O} = 0.86)$ on the lower phase boundary. As the temperature is decreased further, liquid continues to condense with the compositions of the equilibrium vapor and liquid phases at any temperature always lying on the phase boundaries. Condensation is complete at 73 °C, the last vapor to condense having composition $X_{H_2O} = 0.22$ at point C.

At pressures above (or below) 101.3 kPa, the T-X diagram in Fig. 2 will be shifted to higher (or lower) temperatures. It is often useful to plot a P-X phase diagram at constant T as illustrated in Fig. 3 for the methanol-water system. The interpretation of this diagram is analogous to that of Fig. 2. The tie-line UVW at 101.3 kPa at 80 °C in Fig. 3

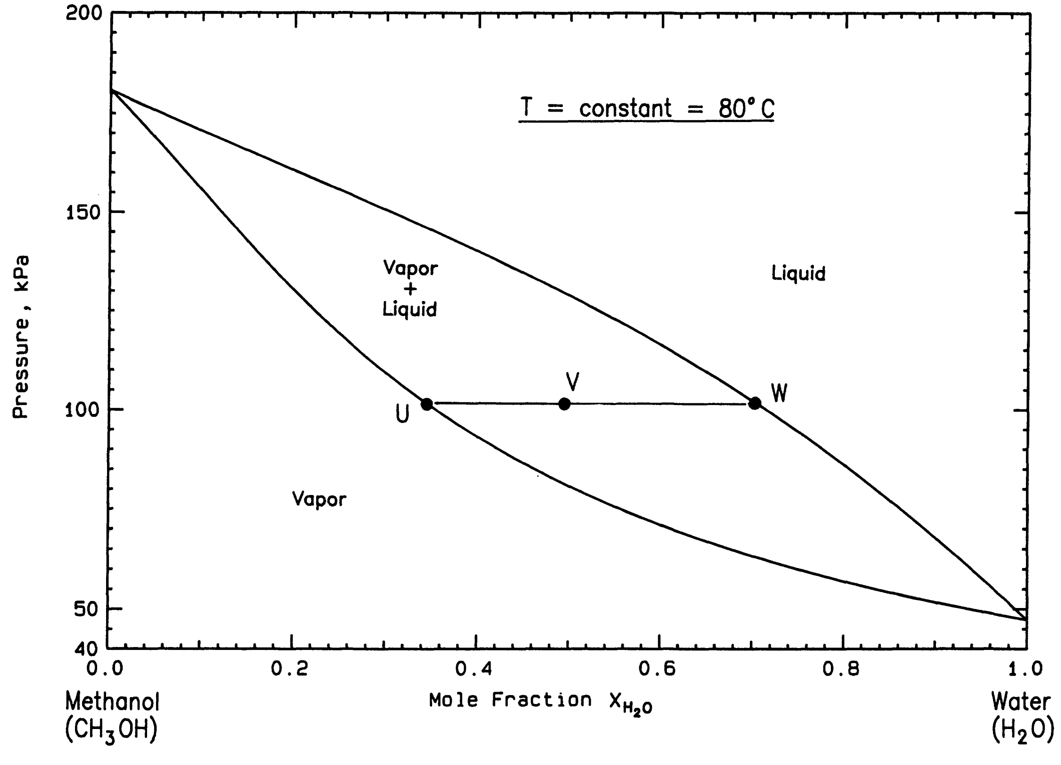

FIG. 3. P-X phase diagram at constant P = 80 °C of the methanol-water system.

is the same as the tie-line *UVW* at the same temperature and pressure in Fig. 2. The lever rule applies to *P-X* diagrams just as to *T-X* diagrams. At temperatures above (or below) T = 80 °C, the *P-X* diagram in Fig. 3 will be shifted to higher (or lower) pressures.

A *T-X* diagram at constant pressure of 101.3 kPa for the CaO-MnO system is shown in Fig. 4. The interpretation of this diagram is similar to that of Fig. 2. Pure CaO and MnO melt at 2572 °C and 1842 °C, respectively. CaO and MnO are completely miscible in the solid as well as the liquid state. The upper phase boundary, which gives the composition of the liquid phase in equilibrium with the solid phase, is called the *liquidus* line, while the lower phase boundary, the *solidus* line, gives the composition of the solid in equilibrium with the liquid.

Suppose that a liquid CaO-MnO solution with composition $X_{\mathrm{MnO}} = 0.60$ is cooled very slowly from an initial temperature of about 2500 °C. When the temperature has decreased to the liquidus temperature 2270 °C (point *B*) the first solid appears, with a composition at point *A* ($X_{\mathrm{MnO}} = 0.28$). As the temperature is decreased further, solid continues to precipitate, with the compositions of the two phases at any temperature being given by the liquidus and solidus compositions at that temperature and with their relative proportions being given by the lever rule. Solidification is complete at 2030 °C, the last liquid to solidify having composition $X_{\mathrm{MnO}} = 0.60$ (point *C*).

The process just described is known as equilibrium cooling. At any temperature during equilibrium cooling the solid phase has uniform (homogeneous) composition. In the preceding example, the composition of the solid phase during cooling varies along the line *APC*. Hence, in order for the solid to have a uniform composition at any temperature, diffusion of CaO from the center to the surface of the growing grains must occur. Since solid-state diffusion is a relatively slow process, equilibrium cooling conditions are only approached if the temperature is decreased very slowly. If a sample of composition X_{MnO} = 0.60 is cooled very rapidly from the liquid, concentration gradients will be observed in the solid grains, with the concentration of MnO increasing toward the surface from a minimum of $X_{\mathrm{MnO}} = 0.28$ (point *A*) at the center. Furthermore, in this case solidification will not be complete at 2030 °C, since at 2030 °C the average concentration of MnO in the solid particles will now be less than X_{MnO} = 0.60.

Since only condensed (i.e., nongaseous) phases are involved in Fig. 4, the phase diagram is very insensitive to pressure. Typically, phase boundaries involving only condensed phases are shifted by the order of one degree per 10 MPa (100 atm) change in pressure.

2.1.1 Minima and Maxima in Two-Phase Regions

A two-phase region may exhibit a minimum or a maximum. For example, in the chloroform-methanol system shown in Fig. 5, a minimum occurs at 35 mol % methanol at 53.4 °C. A liquid of this composition, when heated at constant pressure of 101.3 kPa, will

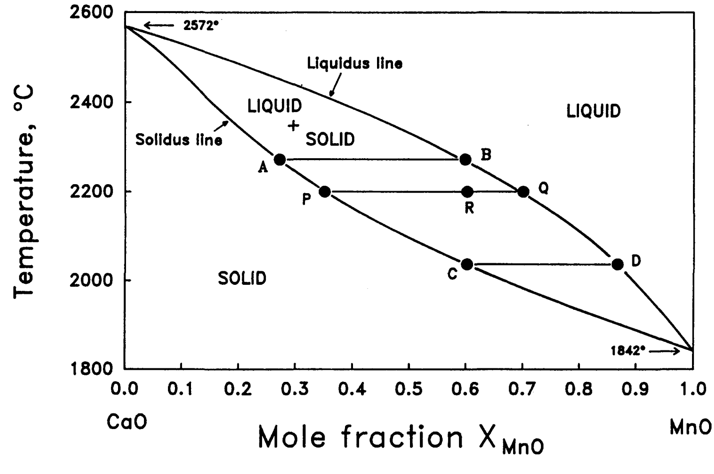

FIG. 4. Phase diagram of the CaO-MnO system at P = 101.3 kPa (after Schenck *et al.*, 1964, and Wu *et al.*, 1993).

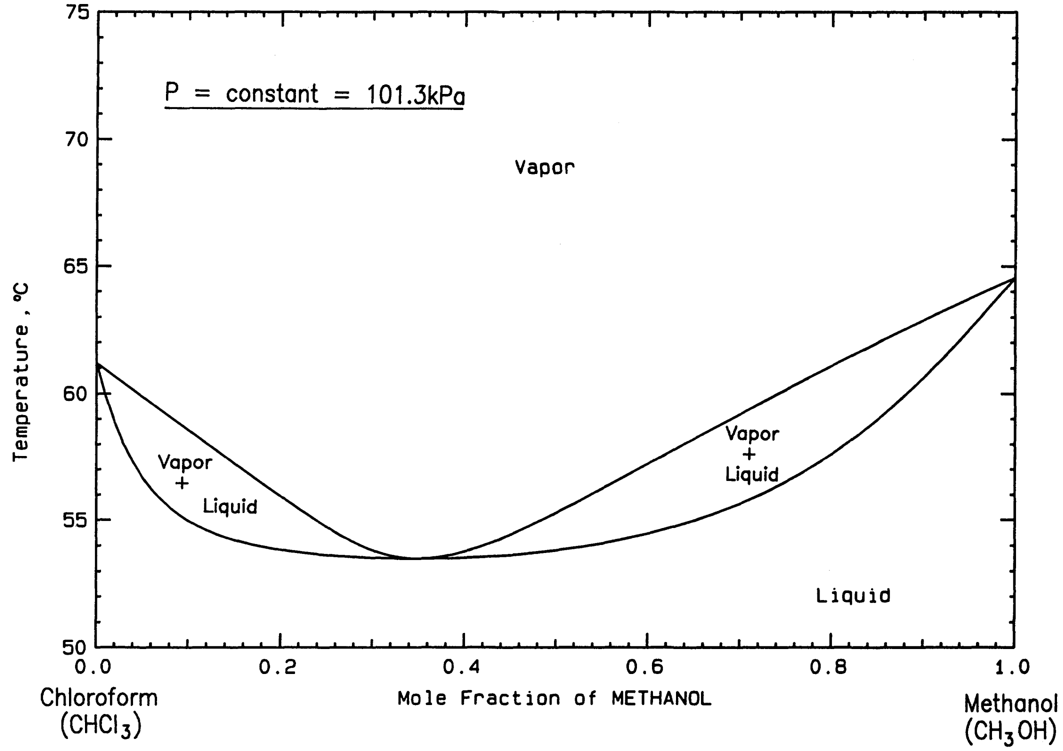

FIG. 5. *T-X* phase diagram at P = 101.3 kPa of the chloroform-methanol system (reprinted from Pelton, 1991).

transform completely to a vapor of the same composition at a constant temperature of 53.4 °C.

An extremum in a (liquid + vapor) two-phase region is called an *azeotrope*. Azeotropes are important in the design of distillation processes. In distillation, the aim is to separate the components of a mixture by virtue of the difference in composition between the liquid and vapor phases. At an azeotrope, however, both phases have the same composition.

Such extrema are not limited to (liquid + vapor) regions. Examples of extrema in (solid + liquid) regions are shown in Figs. 7, 9, and 19. Maxima in (solid + liquid) regions are usually associated with the existence of an intermediate compound as discussed in Sec. 2.5.3. At all such extrema, both phase boundaries must pass through their minimum or maximum at the same point.

2.2 Eutectics

The phase diagram of the Ag-Cu system is shown in Fig. 6. In this system, solid Ag and Cu have limited mutual solubility. At all compositions in the region labeled "solid S_1" a single-phase homogeneous solid solution of Cu in Ag will be observed. The *solvus line* shown in Fig. 6 indicates the solubility limit of Cu in Ag. Similarly, the solvus line on the Cu-rich side indicates the solubility limit of Ag in Cu. At overall compositions lying in the region labeled "2 solids," the system at equi-

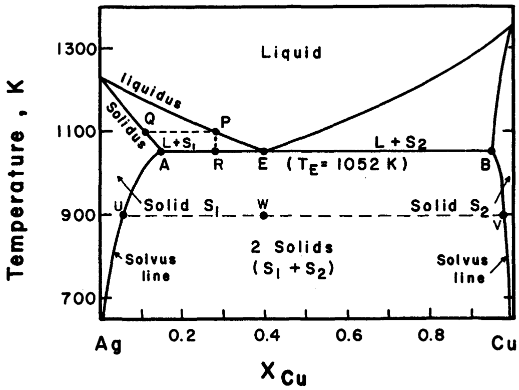

FIG. 6. Phase diagram at P = 101.3 kPa of the Ag-Cu system (after Hultgren *et al.*, 1973).

librium will consist of two phases: Ag saturated with Cu and Cu saturated with Ag. For example, at 900 °C, at an overall composition of 40 mol % Cu (point *W* in Fig. 6), the equilibrium microstructure will consist of two phases with compositions 5 mol % Cu (point *U*) and 97 mol % Cu (point *V*). The relative amounts of the two phases are given by the lever rule: (moles of S_1)/(moles of S_2) = $WV/UW = (0.97 - 0.4)/(0.4 - 0.05) = 1.63$.

The temperature T_E is called the *eutectic temperature*. At T_E, a liquid of composition *E*, the *eutectic composition*, is simultaneously in equilibrium with solid S_1 and solid S_2 of compositions *A* and *B*, respectively. Suppose a liquid Ag-Cu solution of composition $X_{Cu} = 0.29$ is cooled from the liquid state very slowly under equilibrium conditions. At 1100 °C, when the liquid is at point *P*, the first solid appears with composition *Q*. As *T* decreases further, solidification continues with the liquid composition following the liquidus curve from *P* to *E* and the solid composition following the solidus curve from *Q* to *A*. The relative amounts of the two phases at any *T* are given by the lever rule. At T_E, the overall system composition and temperature are given by point *R*. At a temperature just above T_E, two phases are observed, a solid and a liquid of compositions *A* and *E*, respectively, while at a temperature just below T_E, according to the diagram, two solids with compositions *A* and *B* will be observed. Therefore, at T_E, during cooling, the following *binary eutectic reaction* occurs:

$$\text{liquid} \rightarrow \text{solid}_1 + \text{solid}_2. \qquad (2)$$

Under equilibrium conditions, the temperature will remain constant at $T = T_E$ until all the liquid has solidified, and during the reaction the compositions of the three phases will remain fixed at points *A*, *B*, and *E*. For this reason, the eutectic reaction is called an *invariant* reaction.

After the eutectic solidification is completed, if the microstructure is examined at a temperature slightly below T_E, one will observe two types of grains. There will be grains of pure S_1 solid solution of composition *A* that solidified at temperatures above T_E. This is called the *primary* solid or *pro-eutectic* constituent. There will also be grains that formed during the eutectic reaction. These *eutectic* grains will have an overall composition at point *E*, and will consist of an intimate mixture of the phases S_1 and S_2 of compositions *A* and *B*. At high magnification it will be seen that this intimate mixture may consist of alternating lamellae of S_1 and S_2 or of globules or needles of one phase in the other, or of another morphology depending upon the system and upon kinetic factors. Although the pro-eutectic and eutectic *constituents* are not phases (the eutectic grains are two-phase mixtures), the lever rule may nevertheless be applied at a temperature just below T_E as follows: (moles of pro-eutectic constituent)/(moles of eutectic constituent) = RE/AR. If the composition axis is in weight fraction, then this ratio gives the weight ratio of the constituents.

Upon heating an alloy with overall composition between points *A* and *B* ($0.15 < X_{Cu} < 0.96$), the eutectic reaction (2) will occur in the reverse direction at T_E. Hence, 1052 K is the temperature of first appearance of a liquid phase for such alloys. For alloys with compositions outside this range the temperature of first appearance of a liquid phase, under equilibrium conditions, will be the solidus temperature. However, equilibrium conditions may not prevail. For example, an alloy containing 10 mol % Cu, which consists of two solid phases at lower temperatures, will consist of only one solid solution phase when at equilibrium above the solvus temperature of ~1020 °C. Hence, if the alloy is first held somewhat above 1020 °C until equilibrium is attained, before further heating, then the first liquid will appear at the solidus temperature of ~1110 °C. However, if the alloy is heated very quickly from room temperature to T_E, there will not be enough time for the S_2 phase to dissolve by diffusion in the S_1 phase. In this case, the alloy will arrive at T_E as a two-phase mixture, and a liquid phase will appear at T_E.

For further discussion of equilibrium and nonequilibrium heat treatment of alloys and other materials and the resultant microstructures, see MATERIALS TREATMENT and THERMAL ANNEALING.

2.3 Miscibility Gaps

The CaO-SiO$_2$ system is of great importance in the fields of ceramics, refractories, and metallurgy as well as in geology. The phase diagram of this system in Fig. 7 ex-

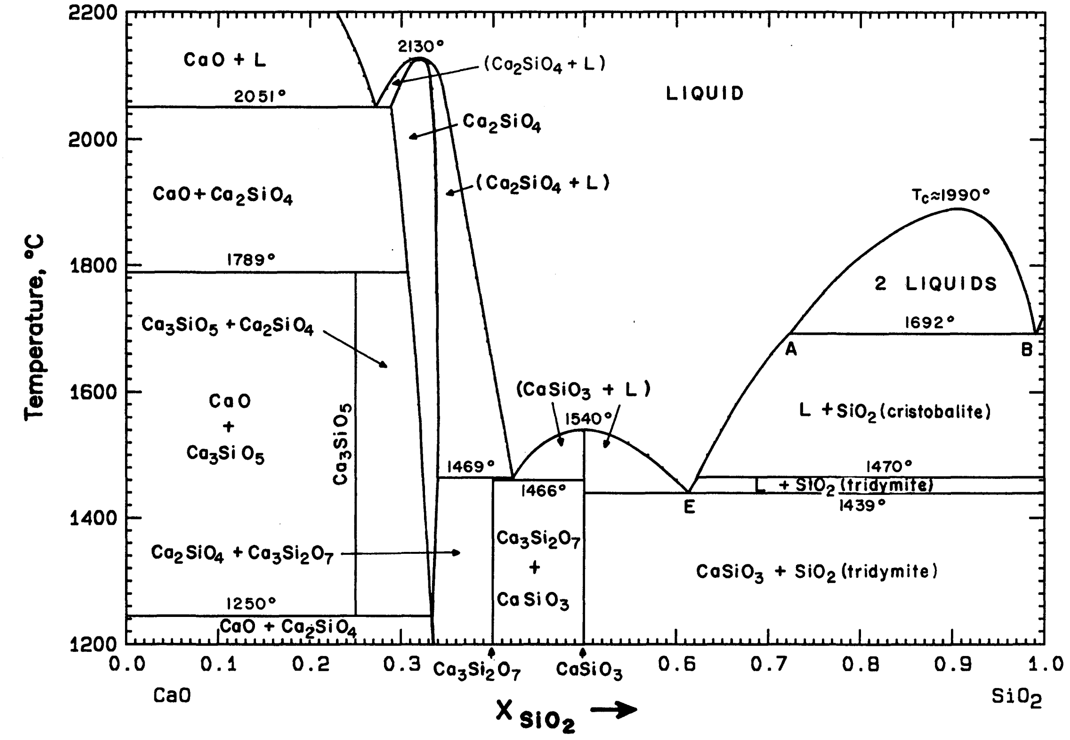

FIG. 7. CaO-SiO$_2$ phase diagram at P = 101.3 kPa (after Eriksson *et al.*, 1994).

hibits many features that are discussed in this section as well as in Secs. 2.4 and 2.5.

A *liquid-liquid miscibility gap* with a *consolute* or *critical point* at T_c = 1990 °C and $X_{c(\mathrm{SiO_2})} \approx 0.91$ is seen in Fig. 7.

Suppose that a liquid of composition $X_{\mathrm{SiO_2}} = 0.80$ is cooled slowly from high temperatures. At ~1820 °C the miscibility gap boundary is crossed and a second liquid layer appears with an initial composition of $X_{\mathrm{SiO_2}} \approx 0.966$. As the temperature is lowered further, the composition of each liquid phase follows its respective phase boundary until, at 1692 °C, the SiO$_2$-rich liquid has a composition of $X_{\mathrm{SiO_2}} = 0.99$ (point B) and in the CaO-rich liquid $X_{\mathrm{SiO_2}} = 0.725$ (point A).

At 1692 °C the following *invariant monotectic reaction* then occurs upon cooling:

$$\text{liquid } B \rightarrow \text{liquid } A + \mathrm{SiO_2(solid)}. \qquad (3)$$

The temperature remains constant at 1692 °C and the compositions of the phases remain constant until all of liquid B is consumed. Cooling then continues with precipitation of solid SiO$_2$, with the equilibrium liquid composition following the liquidus from point A to the eutectic E at 1439 °C.

The critical point in Fig. 7 is similar to the critical point in Fig. 1 in that the properties of the two phases become identical at the critical point and above the critical point there is only one miscible liquid phase. For further discussion, see CRITICAL PHENOMENA.

2.4 Polymorphism

Solid compounds can often exist in two or more crystal structures depending upon the temperature and pressure. The example of water ice has already been given in Sec. 1. As another example, pure SiO$_2$ transforms from the tetragonal tridymite crystal structure to the cubic cristobalite structure at 1470 °C. The transformation involves an invariant reaction, which appears on the phase diagram in Fig. 7. As yet another example, the polymorphism of Fe is seen on the phase diagram in Fig. 9, as is discussed in Sec. 2.6.

In Fig. 7, SiO$_2$ is shown as a *stoichiometric* solid. That is, the solubility of CaO in solid SiO$_2$ is negligible, such that the liquid on the liquidus AE is in equilibrium with pure solid

SiO_2. In reality, no solid is ever exactly stoichiometric (that is, no solubility is ever identically zero), so that there is actually a solidus line and solvus line in Fig. 7 just as in Fig. 6. However, these lines are indistinguishable from the vertical axis at $X_{SiO_2} = 1$.

2.5 Intermediate Compounds

2.5.1 Congruently Melting Compounds In the CaO-SiO_2 system in Fig. 7, there is a compound $CaSiO_3$ that melts at a constant temperature of 1540 °C to form a liquid of the same composition. $CaSiO_3$ is thus called a *congruently melting* compound. There is a eutectic on either side of this compound in Fig. 7. In the eutectic reaction at 1439 °C, for example, liquid of composition E solidifies at constant T to give $CaSiO_3$ and SiO_2. In the eutectic reaction at 1466 °C, liquid at the eutectic composition at $X_{SiO_2} = 0.425$ solidifies to give $CaSiO_3$ and $Ca_3Si_2O_7$.

2.5.2 Incongruently Melting Compounds The compound $Ca_3Si_2O_7$ in Fig. 7 is said to be an *incongruently melting compound*. When heated, $Ca_3Si_2O_7$ decomposes at a constant temperature of 1469 °C to form solid Ca_2SiO_4 plus a liquid of composition $X_{SiO_2} = 0.42$. This is the temperature of first appearance of a liquid. The material will not be completely liquid, however, until it is heated to the liquidus temperature of ~1630 °C. When a liquid containing 40 mol % SiO_2 is cooled at equilibrium, these events occur in reverse. The first solid, Ca_2SiO_4, appears at the liquidus temperature of ~1630 °C. At the *peritectic temperature* of 1469 °C, the following *peritectic reaction* occurs:

$$\text{liquid} + \text{solid}_1 \rightarrow \text{solid}_2, \tag{4}$$

where solid_1 and solid_2 are Ca_2SiO_4 and $Ca_3Si_2O_7$, respectively.

Like the eutectic reaction, this is an invariant reaction, which, under equilibrium conditions, occurs at constant temperature and a fixed compositions of all three phases. However, upon cooling, peritectic reactions generally occur with formation of the product solid ($Ca_3Si_2O_7$) on the surface of the reactant solid (Ca_2SiO_4). This coating can prevent contact between the reactant solid and liquid. Further reaction may thus be greatly retarded so that equilibrium conditions can only be achieved by extremely slow cooling. In the present example, if a liquid containing 40 mol % SiO_2 were cooled rapidly, some unreacted Ca_2SiO_4 would remain at room temperature.

An incongruently melting compound is always associated with a peritectic invariant in a phase diagram.

2.5.3 Nonstoichiometric Compounds In the phase diagram of the CaO-SiO_2 system, Fig. 7, the phase labeled Ca_2SiO_4 is a *nonstoichiometric compound* with nominal composition Ca_2SiO_4 ($X_{SiO_2} = 0.333$). At 2000 °C, for example, the width of the single-phase region extends from $X_{SiO_2} = 0.28$ to $X_{SiO_2} = 0.335$. This is the *range of stoichiometry* over which the compound can exist as a single homogeneous phase. Generally, as in Fig. 7, the range of stoichiometry becomes narrower at lower temperatures.

The nonstoichiometry of Ca_2SiO_4 can also be regarded as a solid solubility of CaO in Ca_2SiO_4 on one side and of SiO_2 in Ca_2SiO_4 on the other (the solubility of SiO_2 being less than that of CaO).

All compounds exhibit some finite range of stoichiometry. Although the compounds $CaSiO_3$ and $Ca_3Si_2O_7$ in Fig. 7 are shown as stoichiometric "line compounds," in reality they have finite, but very narrow, ranges of stoichiometry. That is, a "stoichiometric" compound like $CaSiO_3$ is simply a limiting case of nonstoichiometric compound.

Ca_2SiO_4 is a congruently melting compound with a congruent melting point of 2130 °C. It can be seen that this congruent melting point can also be regarded as the point at which the solidus and liquidus defining the two-phase ($Ca_2SiO_4 + L$) region pass simultaneously through a maximum (see Sec. 2.1.1). In this regard, it can be appreciated that the congruent melting point of the 'stoichiometric' compound $CaSiO_3$ is simply a limiting case of this construction as the range of stoichiometry becomes negligibly small.

As an example of the practical consequences of nonstoichiometry, consider a ceramic with $X_{SiO_2} = 0.31$ in the single-phase region at 2000 °C. If this ceramic is cooled to, say, 1200 °C in the two-phase region and then held at this temperature, fine precipitates of CaO will form. If, on the other hand, it is rapidly quenched to room temperature, there will be insufficient time for precipitation to occur by solid-state diffusion, and the single-phase metastable Ca_2SiO_4 structure will be frozen in.

Generally, nonstoichiometry in a compound results from the presence of lattice *point defects*. Let A_mB_n be the ideal stoichiometry of a compound as given by its crystal structure. An excess of B, $A_{m-x}B_{n+x}$ with $x > 0$, can result from defects such as interstitial B atoms, B atoms occupying lattice sites normally occupied by A atoms, vacant A sites, etc. An excess of A can result from similar defects. The energy of forming such defects will determine the number of defects at equilibrium. If this energy is very high, few defects will be present and the range of stoichiometry will be narrow. A relatively low energy of formation, on the other hand, will result in a large number of defects and a wide range of stoichiometry. In Fig. 7, it can be seen that the energy of forming defects that permit an excess of SiO_2 in Ca_2SiO_4 is larger than the energy of forming defects permitting an excess of CaO.

2.6 Topology of Binary Phase Diagrams

From the preceding examples it can be seen that all binary *T-X* and *P-X* diagrams obey the same topological rules of construction. Scanning horizontally across the diagrams, at constant *T* and *P*, one encounters single-phase regions alternating with two-phase regions. Two-phase regions can pass through maxima or minima as discussed in Sec. 2.1.1, or they can terminate in a critical point as discussed in Sec. 2.3.

There are only two general types of three-phase *invariants* in binary phase diagrams. These are the *eutectic-type* and *peritectic-type invariants* as illustrated in Fig. 8. Let the three phases concerned be called α, β, and γ with β as the central phase as in Fig. 8. The α, β and γ can be solid, liquid, or gaseous phases. At a eutectic-type invariant, the following invariant reaction occurs isothermally as the system is cooled:

$$\beta \rightarrow \alpha + \gamma, \tag{5}$$

while, at a peritectic-type invariant, the invariant reaction upon cooling is

$$\alpha + \gamma \rightarrow \beta. \tag{6}$$

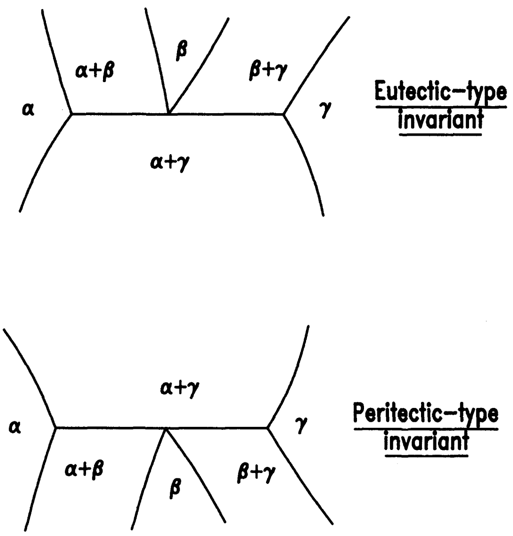

FIG. 8. The two types of invariants in binary *T-X* and *P-X* phase diagrams.

For example, at a *eutectic*, α = solid$_1$, β = liquid, γ = solid$_2$. The eutectic reaction is $l \rightarrow s_1, + s_2$. A *monotectic*, as seen in Fig. 7 at 1692 °C is also a eutectic-type invariant with α = liquid$_1$, β = liquid$_2$, γ = solid. The monotectic reaction is $l_2 \rightarrow l_1 + s$.

Other examples can be seen in the phase diagram of the Fe-Mo system in Fig. 9. *Peritectics*, in which α = liquid β = solid$_1$, γ = solid$_2$, are seen at 1540 °C, 1488 °C, and 1450 °C. The peritectic reaction is $l + s_2 \rightarrow s_l$. At 1235 °C and 1200 °C in Fig. 9 are examples of *eutectoids* (which are eutectic-type invariants) involving three solid phases. With α = solid$_1$, β = solid$_2$, and γ = solid$_3$, the eutectoid reaction is $s_2 \rightarrow s_1 + s_3$. Examples of *peritectoids* are seen at 1370 °C and 950 °C. These are peritectic-type invariants. The peritectoid reaction is $s_1 + s_3 \rightarrow s_2$. In Fig. 7 the invariants at 1250 °C and 1789 °C are, respectively, limiting cases of a eutectoid and a peritectoid where the range of homogeneity of the phase Ca_3SiO_5 is negligibly small.

Many other eutectic-type and peritectic-type invariants are observed such as *syntectics* ($l_1 + l_2 \rightarrow s_1$) and *catatectics* ($s_1 \rightarrow l + s_2$).

In Secs. 2.2 and 2.5.2, the microstructures resulting from heat treatments of systems with eutectics and peritectics were briefly described. Similar considerations apply in the case of eutectoids and peritectoids. For ex-

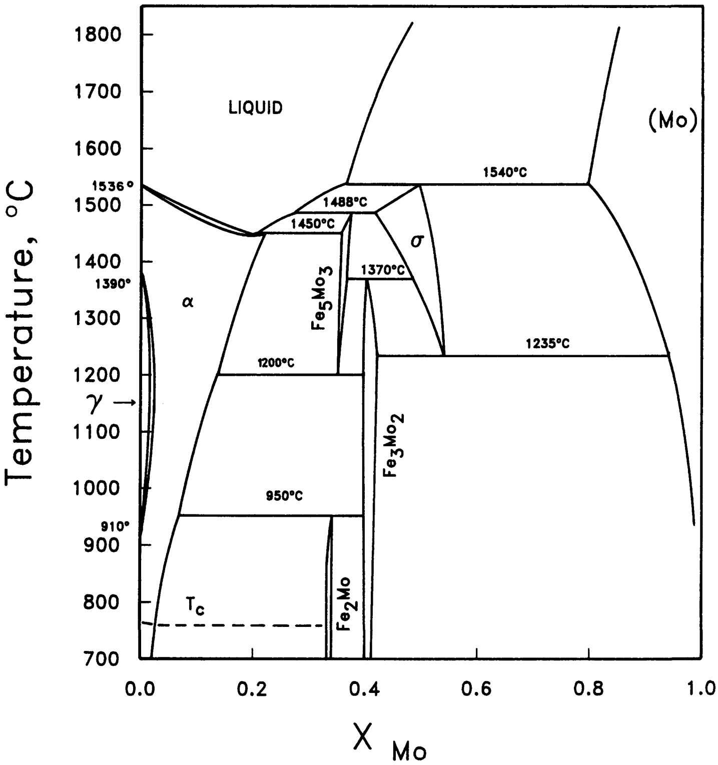

FIG. 9. Fe-Mo phase diagram at P = 101 kPa (Kubaschewski, 1982; reprinted from Pelton, 1991). The α and γ phases have the body-centered-cubic and face-centered-cubic structures, respectively. The σ phase is an intermediate solid phase.

ample, the best-known example of a eutectoid microstructure is that of pearlitic steel. For more information, see MATERIALS TREATMENT; METALLURGY, PHYSICAL; and STEEL AND IRON.

A final point of interest in Fig. 9 is the two-phase ($\alpha + \gamma$) γ *loop* between 910 °C and 1390 °C. Pure Fe adopts the face centered-cubic (fcc) γ crystal structure between these two temperatures, but exists as the body-centered-cubic (bcc) α phase at higher and lower temperatures. Mo, however, is much more soluble in the bcc than in the fcc structure. Hence, small additions of Mo stabilize the bcc structure, thereby producing the γ loop.

2.7 Phase Diagrams of Systems with Liquid-Crystalline Phases

In Sec. 2.4, solid polymorphism was discussed. In certain systems, liquid polymorphism is also observed. For example, the phase diagram of the system EEBC-TBBA is shown in Fig. 10.

All the phases shown with the exception of solid$_1$ and solid$_2$ are liquid phases. The high-temperature "isotropic" phase is a "normal" disordered liquid. In all the other phases there exists, on the average, an orientational order of the molecules. This orientational order may be distinguished from the positional order found in solid crystals. Generally, the molecules of liquids that form liquid crystals are elongated and rigid. If, on the average, the molecular axes are aligned, a *nematic* liquid-crystalline phase results. In other liquid-crystalline phases, the molecules tend to be both aligned and arranged in layers. Such phases are termed *smectic*. Various smectic phases can exist, differing, for example, in the angle between the molecular axes and the planes of the layers. In all respects, the phase diagram in Fig. 10 obeys the topological rules of binary T-X diagrams discussed in Sec. 2.6. For example, pure EEBC and TBBA transform from the smectic A to the nematic liquid crystal structure at 158.3 °C and 199.0 °C, respectively. At intermediate compositions there is a two-phase (smectic A + nematic) region, with a minimum at 156.2 °C. Several eutectic-like and peritectic-like invariants are

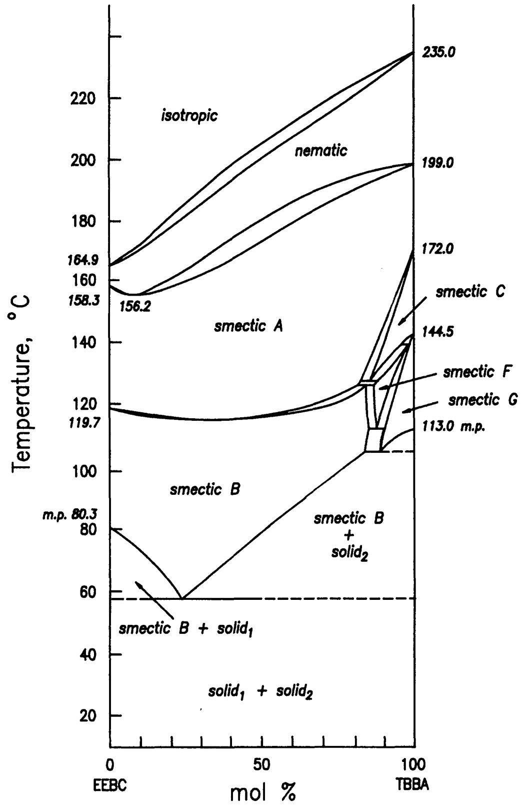

FIG. 10. Equilibrium *T-X* diagram at $P = 101.3$ kPa of the EEBC-TBBA system, which exhibits liquid-crystalline phases (after Richter *et al.*, 1981). (EEBC = ethyl 4-[4-ethoxy benzylidene-amino]-cinnamate: TBBA = terephthalylidene-*bis*-[4-*n*-butyl-aniline].)

seen in Fig. 10. For further discussion see LIQUID CRYSTALS, STRUCTURE OF.

3. THE GIBBS PHASE RULE

In Secs. 1 and 2 the topology of phase diagrams was described with no justification given as to why these topological rules should be obeyed. The present section addresses this question, although space restrictions do not permit a completely rigorous development to be made. For a more complete treatment see THERMODYNAMICS, EQUILIBRIUM.

First, the term *component* should be defined more precisely. At most, the number of components in a system is equal to the number of chemical elements. However, there may be chemical constraints such that certain elements always appear in fixed ratios. If these constraints apply *in every phase* that could possibly appear over the ranges of interest of *T* and *P*, then there will be a consequent reduction of the number of components. For example, in the methanol-water system, Figs. 1 and 2, the elements are always combined as CH_3OH and H_2O in every phase. Hence, this is a two-component system. A system of water and sodium chloride near ambient conditions is a two-component system since the number of Na^+ ions must equal the number of Cl^- ions in every phase. A system prepared by mixing the salts NaCl, KCl, NaF, and KF is a three-component system because the number of cations must always equal the number of anions. (A solution of any desired composition could always be formed by mixing only three of the four salts.)

Next we must determine the number of variables required to define the equilibrium state of a system. We choose these as temperature *T*, total pressure *P*, and the compositions of the phases at equilibrium. In the general case, other variables such as electrical, magnetic, and gravitational fields can also influence phase equilibria. However, we do not consider these here. It must be stressed that the overall (global) composition of the system is NOT a variable in the present context. We are interested in knowing which phases are present and their compositions, but not their relative amounts. For example, in a two-phase region in a binary system, as long as the overall composition lies on the tie-line between the compositions of the two phases, they will both be present and their compositions will not change.

Now, in order to specify the composition of a phase in a system with *C* components, $C - 1$ composition variables are required, since the sum of the mole fractions is unity. Hence, if *P* phases are present at equilibrium. $P(C - 1)$ composition variables are needed. Along with temperature and pressure, the number of variables is thus $P(C - 1) + 2$.

When we say that two phases are at equilibrium we mean that there is no net transfer with time of any component between the phases. Now, there must be some solution property that determines the direction of spontaneous transfer of a component between two phases in contact. This property is called the *chemical potential of component i*, and is given the symbol μ_i (where the components are numbered 1, 2, 3, . . ., *C*). Just as electrical charge is transferred spontaneously from high to low electrical potential, so component i will migrate spontaneously from a phase of higher chemical potential to a phase of lower chemical potential.

Therefore, if two phases, α and β are in equilibrium, then $\mu_i^\alpha = \mu_i^\beta$ for all components $1 \leq i \leq C$. If P phases are present at equilibrium, then there are $C(P - 1)$ such equations. However, μ_i, being a property of a phase, is a function of T, pressure, and the composition of the phase. That is, each of the $C(P - 1)$ equations is an equation relating those variables which specify the equilibrium state of the system.

The difference between the number of variables, $P(C - 1) + 2$, and the number of equations relating the variables, $C(P - 1)$, is called the *variance*, or the *number of degrees of freedom*, F, of the system. This then is the Gibbs phase rule:

$$F = C - P + 2. \tag{7}$$

That is, F is the number of variables which can and must be specified in order to specify completely the state of the system.

In a one-component system as in Fig. 1, $C = 1$. Hence, the phase rule reduces to $F = 3 - P$. The only variables being considered are temperature and pressure. In a single-phase area, $F = 3 - 1 = 2$. That is, T and pressure can be specified independently. These regions are thus called *bivariant*. When two phases are at equilibrium, $F = 3 - 2 = 1$. If, say, T is specified, then the pressure is fixed. For example, if liquid and vapor are in equilibrium at a given T, then the pressure must be the equilibrium vapor pressure. Such two-phase regions are thus called *univariant* and are represented by lines on Fig. 1. When three phases are in equilibrium, $F = 0$, the system is *invariant*, and the state is represented by a triple point in Fig. 1.

In a two-component system, $C = 2$ and $F = 4 - P$. When we consider a T-X phase diagram at a constant pressure, however, we have already eliminated one degree of freedom. Hence, for an isobaric T-X diagram, $F = 3 - P$. The variables of interest are now T and the compositions of the phases. In a single-phase area, $F = 3 - 1 = 2$. Hence, T and composition can be independently specified, and the system is bivariant. When two phases are in equilibrium, $F = 1$. If, say, T is specified, then the compositions of both phases are fixed by the ends of the tie-line. When three phases are at equilibrium, the system is invariant. That is, the compositions of all three phases as well as the temperature are fixed. This gives rise to eutectic-type and peritectic-type invariants.

It can thus be seen how the topology of phase diagrams is determined by the phase rule.

4. TERNARY PHASE DIAGRAMS

Only a very brief introduction to isobaric ternary T-X diagrams can be given here. For further discussion see Pelton (1983, 1991) and Prince (1966).

4.1. Ternary Composition Triangle

In a ternary system with components A, B, C, the sum of the mole fractions is unity: $X_A + X_B + X_C = 1$. Hence, there are two independent composition variables. A representation of composition, symmetrical with respect to all three components, may be obtained with the equilateral "composition triangle" as shown in Fig. 11(a) for the LiF-NaF-KF system. Compositions at the corners of the triangle correspond to the pure components. Along the edges of the triangle are found compositions corresponding to the three binary subsystems LiF-NaF, LiF-KF, and NaF-KF. Lines of constant mole fraction X_{LiF} are parallel to the NaF-KF edge, while lines of constant X_{NaF} and X_{KF} are parallel to the LiF-KF and LiF-NaF edges, respectively. For example, at point *a* in Fig. 11(a), $X_{\mathrm{LiF}} = 0.06$, $X_{\mathrm{NaF}} = 0.23$, and $X_{\mathrm{KF}} = 0.71$.

Similar equilateral composition triangles can be drawn with coordinates in terms of weight percent of the three components.

4.2 Ternary Space Model

A ternary temperature-composition "phase diagram" at constant total pressure may be plotted as a *three-dimensional space model* within a right triangular prism with the equilateral composition triangle as base and temperature as vertical axis. Such a space model for the simple eutectic ternary system LiF-NaF-KF is illustrated in Fig. 11(b). On the three vertical faces of the prism are found the

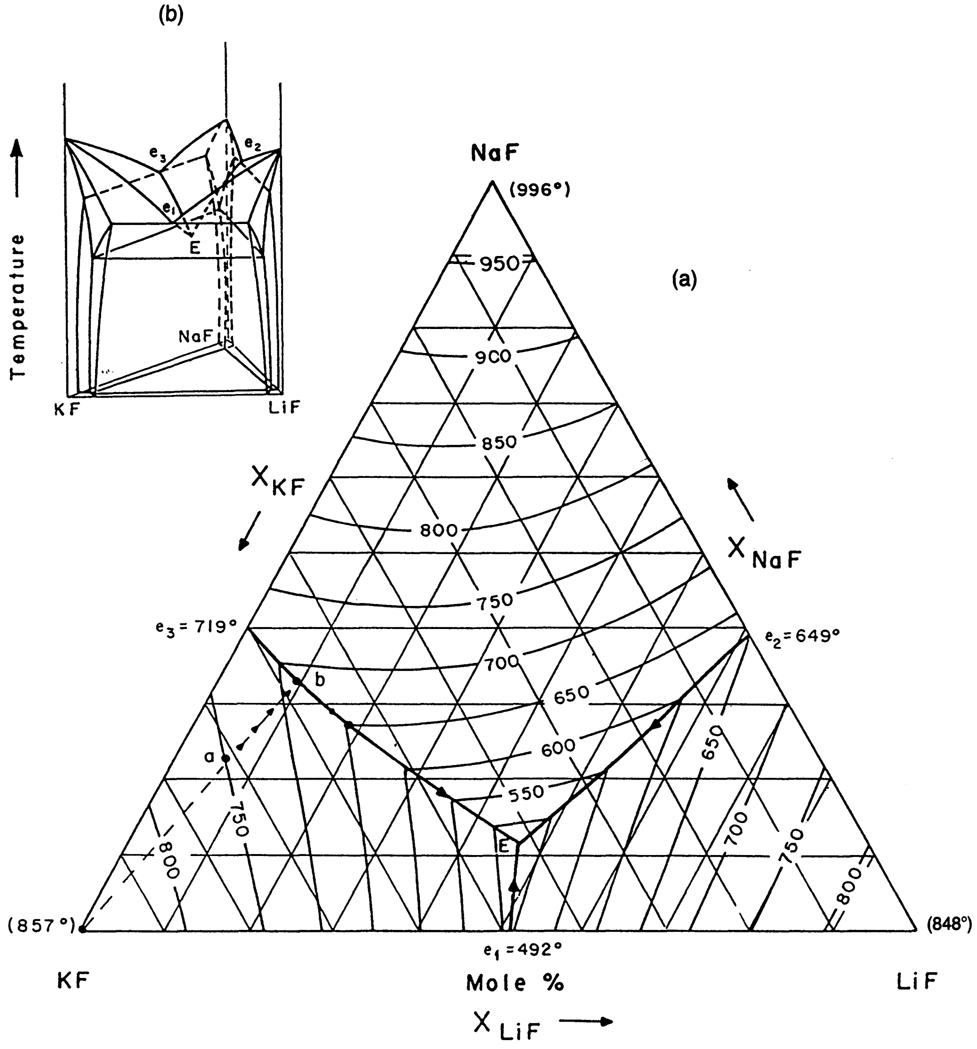

FIG. 11. (a) Polythermal projection of the liquidus surface of the LiF-NaF-KF system. Temperature in °C (Levin *et al.*, 1964.) (b) Perspective view of ternary space model of a simple eutectic system. ϵ_1, ϵ_2, and ϵ_3 are the binary eutectics and *E* is the ternary eutectic. The base of the prism is the equilateral composition triangle. (Reprinted from Pelton, 1983.)

phase diagrams of the three binary subsystems, LiF-NaF, LiF-KF, and NaF-KF, which in this example are all simple eutectic binary systems. The binary eutectic points are at e_1, e_2, and e_3. Within the prism are seen three *liquidus surfaces* descending from the melting points of pure LiF, NaF, and KF. Compositions on these surfaces correspond to compositions of liquids in equilibrium with LiF-, NaF-, and KF-rich solid phases.

The NaF and KF liquidus surfaces in Fig. 1(b) intersect along the line e_3E. Liquids with compositions along this univariant "valley" are therefore in equilibrium with NaF-rich and KF-rich solid phases simultaneously. The three univariant lines meet at the *ternary eutectic point E* at which the temperature and the compositions of all four phases in equilibrium are fixed.

In a 3-component system at constant pressure, the phase rule, Eq. (7), reduces to $F = 4 - P$. Therefore, for a liquid in equilibrium with one solid, $F = 2$. Hence, liquiduses are bivariant surfaces. The composition of a liquid in equilibrium with two solids is represented by a univariant line, and liquid can exist in equilibrium with three solids at a ternary eutectic invariant point.

4.3 Polythermal Projections of Liquidus Surfaces

A two-dimensional representation of the ternary liquidus surface may be obtained as an orthogonal projection upon the base composition triangle. Such a *polythermal projection* of the liquidus of the LiF-NaF-KF system is shown in Fig. 11(a). This is a simple eutectic ternary system. The constant-temperature lines in Fig. 11(a) are called *liquidus isotherms*. The univariant valleys are shown as the heavier lines. By convention, the arrows indicate the directions of decreasing temperature along these lines.

Let us consider the sequence of events that occur during the equilibrium cooling of a liquid of overall composition *a* in Fig. 11(a). Point *a* lies within the *field of primary crystallization* of KF. That is, it lies within the composition region in Fig. 11(a) in which KF-rich solid will be the first solid to precipitate upon cooling. As the liquid is cooled, the KF liquidus surface is reached at $T = 750$ °C. A solid KF-rich phase begins to precipitate at this temperature. Now, in this particular system, NaF and LiF are nearly insoluble in solid KF so that the solid phase is virtually pure KF. [Note that this fact cannot be deduced from Fig. 11(a) alone]. Therefore, as solidification proceeds, the liquid becomes depleted in KF, but the ratio $X_{\mathrm{NaF}}/X_{\mathrm{LiF}}$ in the liquid remains constant. Hence, the composition path followed by the liquid (its *crystallization path*) is a straight line passing through point *a* and projecting to the KF corner of the triangle. This crystallization path is shown in Fig. 11(a) as the line *ab*. (Note that in the general case, in which a solid solution rather than a pure component or stoichiometric compound is precipitating, the crystallization path will not be a straight line.)

When the composition of the liquid has reached point *b* in Fig. 11(a) at $T \approx 680$ °C, the primary crystallization of KF is complete. Upon further cooling, the liquid composition follows the univariant valley from *b* to *E* while LF- and NaF-rich solids coprecipitate as a binary eutectic mixture. When the liquidus composition attains the ternary eutectic composition *E* at $T = 463$ °C, the invariant *ternary eutectic reaction* occurs:

$$\text{liquid} \rightarrow s_1 + s_2 + s_3, \qquad (8)$$

where s_1, s_2, and s_3 are the three solid phases and where the compositions of all four phases, as well as T, remain fixed until all liquid is solidified.

As another more complex example, the polythermal liquidus projection is shown in Fig. 12 for the $CaO\text{-}Al_2O_3\text{-}SiO_2$ system. Liquidus surfaces, with isotherms, are seen for several primary crystalline phases. The compositions of the solid compounds are shown in Fig. 12 by points. For example, $CaSiO_3$ (pseudowollastonite) is a *binary compound* (since its composition is in the $CaO\text{-}SiO_2$ binary system), which melts at 1540 °C. For any liquid with composition in the field labeled "pseudowollastonite," $CaSiO_3$ will be the first solid to precipitate upon cooling. Anorthite, $CaAl_2Si_2O_8$, is a *ternary compound* which melts congruently at 1555 °C.

The liquidus surfaces for the various solid phases in Fig. 12 intersect along univariant lines. The arrows shown on these lines indicate the direction followed by the liquid composition during cooling in equilibrium with both solid phases. The univariant lines meet at invariant points. If the arrows on all three lines converge at such a point, then the point is a *ternary eutectic point*. If one or more arrows lead away from the point, then the point is a *ternary peritectic point*.

4.4 Ternary Isothermal-Isobaric Sections

If the solid phases in equilibrium with the liquid are stoichiometric, or nearly stoichiometric compounds, then a polythermal liquidus projection as in Figs. 11(a) and 12 gives complete information on the ternary space model. However, when solid solutions are present, the liquidus projection does not give information on the compositions of the solid phases. This information can be conveyed by a series of *isothermal sections*, which are constant-temperature slices through the space model.

An isothermal section at 25 °C of the Al-Zn-Mg system is shown in Fig. 13. Such isothermal sections follow strict topological rules. With three components, at constant T and pressure, the phase rule, Eq. (7), reduces to $F = 3 - P$. Single-phase regions are thus

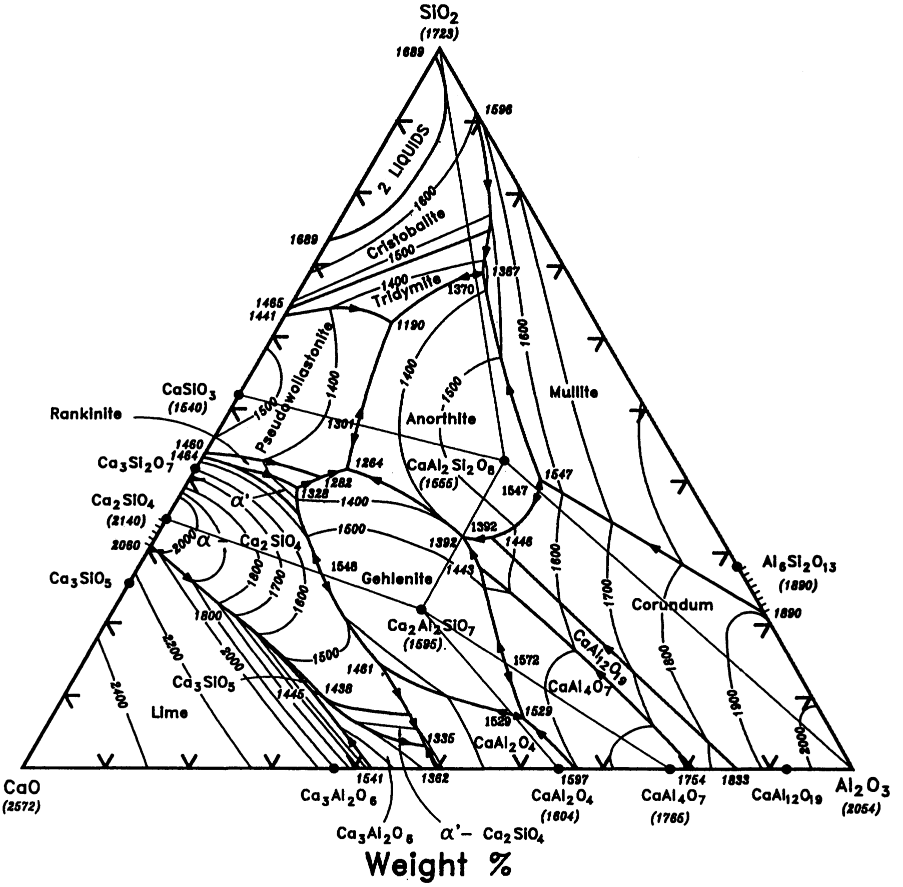
SiO_2
(1723)
1689
1596
2 LIQUIDS
Cristobalite
1600
1500
1400
Tridymite
1370
1367
1190
1689
1465
1441
Pseudowollastonite
$CaSiO_3$
(1540)
Rankinite
1460
1464
$Ca_3Si_2O_7$
Ca_2SiO_4
(2140)
2060
α'
α
Ca_2SiO_4
1548
Gehlenite
Ca_3SiO_5
Lime
Anorthite
$CaAl_2Si_2O_8$
(1555)
1547
Mullite
$Ca_2Al_2SiO_7$
(1595)
1572
Corundum
$CaAl_4O_7$
$CaAl_2O_4$
$Al_6Si_2O_{13}$
(1890)
1890
CaO
(2572)
1541
$Ca_3Al_2O_6$
1362
$Ca_3Al_2O_6$
α'- Ca_2SiO_4
1597
$CaAl_2O_4$
(1604)
1764
1833
$CaAl_4O_7$
(1765)
$CaAl_{12}O_{19}$
Al_2O_3
(2054)
Weight %

FIG. 12. Polythermal projection of the liquidus surface of the CaO-Al_2O_3-SiO_2 system. Temperatures in °C. (Eriksson and Pelton, 1993.)

bivariant ($F = 2$). These are shown as the black regions in Fig. 13. When two phases are in equilibrium, $F = 1$, and a tie-line results. These tie-lines together form two-phase regions. For example, in Fig. 13, a set of tie-lines joins the δ-phase and τ-phase regions. For an overall composition on any one of these tie-lines a δ-phase alloy with composition at one end of the tie-line is in equilibrium with a τ-phase alloy whose composition is at the other end. When three phases are at equilibrium, $F = 0$ and the compositions of all three phases are fixed. Three-phase regions are represented by *tie-triangles* as shown in Fig. 13. For an overall composition inside such a triangle, three phases are present at equilibrium, and their compositions are given by the corners of the triangle.

5. OTHER KINDS OF PHASE DIAGRAMS

The types of phase diagrams discussed in the preceding sections, and exemplified in Figs. 1 through 13, are the most commonly encountered. However, there are many other useful types of phase diagrams.

Vertical sections, known as *isopleths*, through a ternary space model like Fig. 11(b), are often a convenient method of representing ternary equilibria. In systems of four or more components, many techniques of presenting two-dimensional sections or projections have been proposed. For a discussion of these, see Prince (1963), Palatnik and Landau (1964), Prince (1966), and Hillert (1985).

Many types of phase diagrams result when the chemical potentials of components are

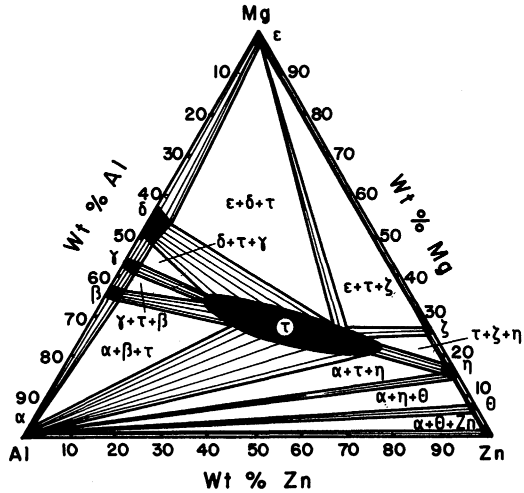

FIG. 13. Ternary isothermal section of the Al-Zn-Mg system at 25 °C (after Koster and Dullenkopf, 1936, reprinted from Pelton, 1983).

used as axis variables or are held constant. The chemical potential of a gas is easily controlled by fixing its partial pressure. Hence, the most useful and common phase diagrams of this class are those with partial pressures of gases as variables.

For example, Fig. 14 shows a *predominance diagram* of the Cu-S-O system at 1000 K. The axes are the equilibrium SO_2 and O_2 partial pressures. The diagram is divided into areas or domains of stability of the various solid Cu-containing compounds. The conditions for coexistence of two and three solids are indicated respectively by the lines and triple points on the diagram. Such diagrams have many applications in hot corrosion, roasting, chemical vapor deposition, etc. It can be seen that these diagrams obey the same topological rules of construction as *P-T* diagrams of one-component systems as in Fig. 1.

Another type of predominance diagram, which again obeys the same rules of construction, is shown in Fig. 15. This diagram indicates which solid Fe-containing phases are stable as a function of equilibrium oxygen partial pressure and temperature when the partial pressure of SO_2 is held constant.

Predominance diagrams, their thermodynamic calculation, and their extension to systems of several components have been discussed by Bale *et al.* (1986) and Bale (1990).

In Fig. 16 is shown a *chemical-potential–composition phase diagram* for the Fe-Cr-O system at 1300 °C. This is an isothermal plot of the equilibrium oxygen pressure vs the molar metal ratio, $n_{Cr}/(n_{Fe} + n_{Cr})$, where n_i = number of moles of *i*. The diagram consists of single-phase areas, two-phase regions with

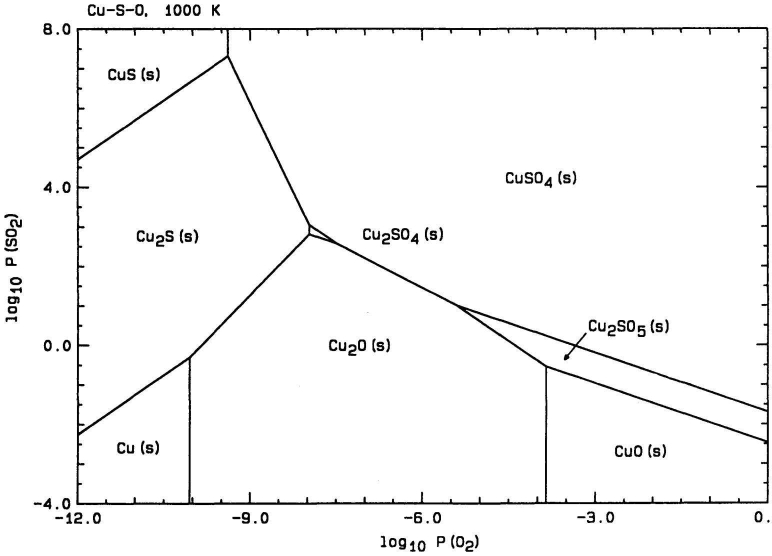

FIG. 14. Predominance diagram, $\log p_{SO_2}$ vs $\log p_{O_2}$ (atm) at 1000 K for the Cu-S-O system (Bale *et al.*, 1986).

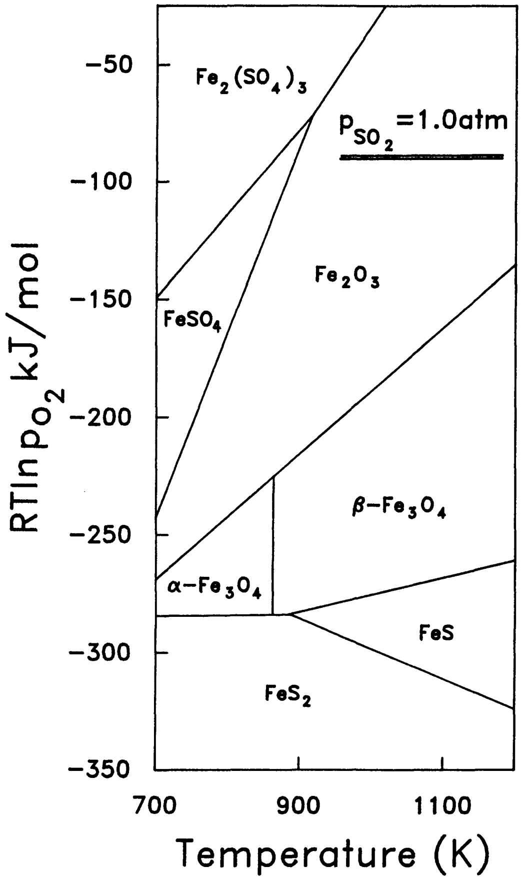

FIG. 15. Predominance diagram, *RT* lnp_{O_2} versus *T* at p_{SO_2} = 1.0 atm for the Fe-S-O system (reprinted from Pelton, 1991).

horizontal tie-lines (to which the lever rule applies), and three-phase eutectic-type and peritectic-type invariants. That is, topologically the diagram obeys the same rules of construction as binary *T-X* or *P-X* diagrams.

Many other phase diagrams can be drawn which have the same topologies as either one-component *P-T* diagrams, binary *T-X* and *P-X* diagrams, or ternary isothermal sections. Such phase diagrams and their classification have been discussed by Pelton and Schmalzried (1973).

6. THERMODYNAMIC ORIGIN OF PHASE DIAGRAMS

In this section we briefly discuss the thermodynamic origin of phase diagrams. This subject is treated in more detail by Pelton (1991). A basic knowledge of solution thermodynamics is required for this section. The principles are covered in THERMODYNAMICS, EQUILIBRIUM and by Pelton (1991).

Every system is characterized by a Gibbs energy function. The equilibrium state of the system is the state with minimum Gibbs energy. The isobaric *T-X* phase diagram of the Ge-Si system is shown in the lowest panel of Fig. 17. In the upper three panels of Fig. 17 the molar Gibbs energies of the solid and liquid phases, g^s and g^l, at three temperatures are shown to scale. As illustrated in the top panel, g^s varies with composition between the standard molar Gibbs energies of pure solid Ge and pure solid Si, $g^{0(s)}_{\mathrm{Ge}}$ and $g^{0(s)}_{\mathrm{Si}}$, while g^l varies between the standard molar Gibbs energies of the pure liquid components, $g^{0(l)}_{\mathrm{Ge}}$ and $g^{0(l)}_{\mathrm{Si}}$.

The quantities Δg^s_m and Δg^l_m are the molar Gibbs energies of mixing of the solutions. Each of these is equal to an ideal mixing term, $RT(X_1 \ln X_1 + X_1 \ln X_2)$, plus an excess term, g^E:

$$\Delta g_m = RT(X_1 \ln X_1 + X_2 \ln X_2) + g^E, \tag{9}$$

where X_1 and X_2 are the component mole fractions. The ideal mixing term is negative and is calculated statistically from the entropy of a random distribution of Ge and Si atoms over the lattice positions (see STATISTICAL MECHANICS, CLASSICAL).

As shown in the top panel of Fig. 17, the difference between $g^{0(l)}_{\mathrm{Si}}$ and $g^{0(s)}_{\mathrm{Si}}$ is equal to the standard molar Gibbs energy of fusion (melting) of pure Si, $\Delta g^0_{f(\mathrm{Si})} = g^{0(l)}_{\mathrm{Si}} - g^{0(s)}_{\mathrm{Si}}$. Similarly for Ge, $\Delta g^0_{f(\mathrm{Ge})} = g^{0(l)}_{\mathrm{Ge}} - g^{0(s)}_{\mathrm{Ge}}$. The Gibbs energy of fusion of a pure component may be written as

$$\Delta g^0_f = \Delta h^0_f - T\Delta s^0_f, \tag{10}$$

where Δh^0_f and Δs^0_f are the standard molar enthalpy and entropy of fusion.

Since to a first approximation, Δh^0_f and Δs^0_f are independent of T, Δg^0_f is approximately a linear function of T. If $T > T^0_f$, then Δg^0_f is negative. If $T < T^0_f$, then Δg^0_f is positive. Hence, as seen in Fig. 17, as T decreases, the g^s curve descends relative to g^l. At 1500 °C, $g^l < g^s$ at all compositions. Therefore, by the principle that a system always seeks the state of minimum Gibbs energy at constant T and

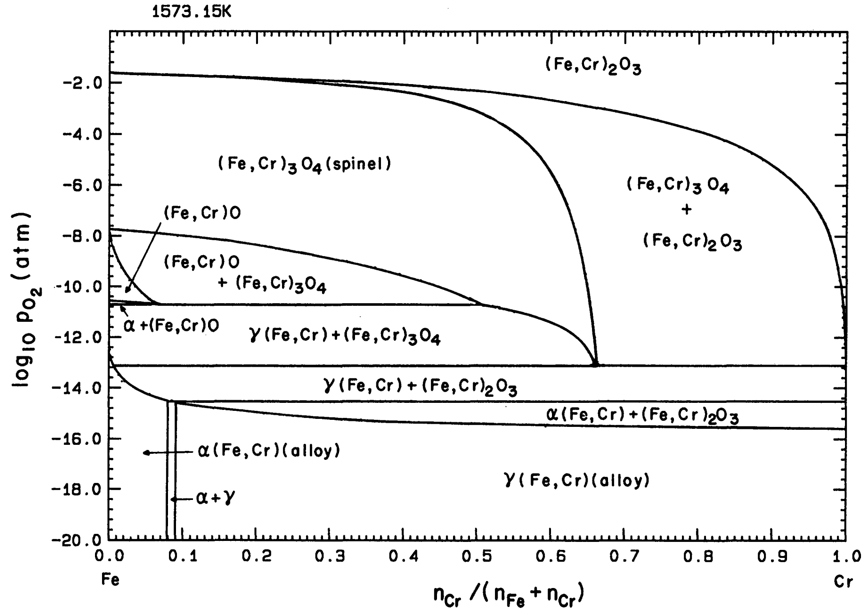

FIG. 16. Phase diagram of $\log p_{O_2}$ vs molar metal ratio for the Fe-Cr-O system at 1300 °C (Pelton and Schmalzried, 1973).

P, the liquid phase is stable at all compositions at 1500 °C.

At 1300 °C, the curves of g^s and g^l cross. The line P_1Q_1, which is the common tangent to the two curves, divides the composition range into three sections. For compositions between pure Ge and P_1, a single-phase liquid is the state of minimum Gibbs energy. For compositions between Q_1 and pure Si, a single-phase solid solution is the stable state. Between P_1 and Q_1, a total Gibbs energy lying on the tangent line P_1Q_1 may be realized if the system adopts a state consisting of two phases with compositions at P_1 and Q_1 and with relative proportions given by the lever rule. Since the tangent line P_1Q_1 lies below both g^s and g^l, this two-phase state is more stable than either phase alone. Furthermore, no other line joining any point on the g^l curve to any point on the g^s curve lies below the line P_1Q_1. Hence, this line represents the true equilibrium state of the system, and the compositions P_1 and Q_1 are the liquidus and solidus compositions at 1300 °C.

As T is decreased to 1100 °C, the points of common tangency are displaced to higher concentrations of Ge. For $T < 937$ °C, $g^s < g^l$ at all compositions. In this way, the lens-shaped Ge-Si phase diagram is generated.

If both Δg_m^l and Δg_m^s are ideal [i.e., if the g^E term in Eq. (9) is zero for both phases], then a lens-shaped two-phase region always results. However, in most systems, even approximate ideal behavior is the exception. Curves of g^s and g^l for a hypothetical system A-B are shown schematically in Fig. 18 at a constant temperature below the melting points of pure A and B such that the solid state is the stable state for both pure components. However, in this system $g^{E(l)} < g^{E(s)}$, so that g^s presents a flatter curve than does g^l and there exists a central composition region in which $g^l < g^s$. Hence, there are two common tangent lines, P_1Q_1 and P_2Q_2. Such a situation gives rise to a phase diagram with a minimum in the two-phase (solid + liquid) region similar to the diagram of Fig. 5.

Other forms of the Gibbs energy curves can give rise to phase diagrams with regions of immiscibility, eutectic-type and peritectic-type invariants, etc. The treatment can be extended to multicomponent systems.

6.1 Models of Solutions—Regular Solution Theory

The phase diagram of a system is determined by the thermodynamic properties of the phases which, in turn, depend upon their structure and bonding. To illustrate this re-

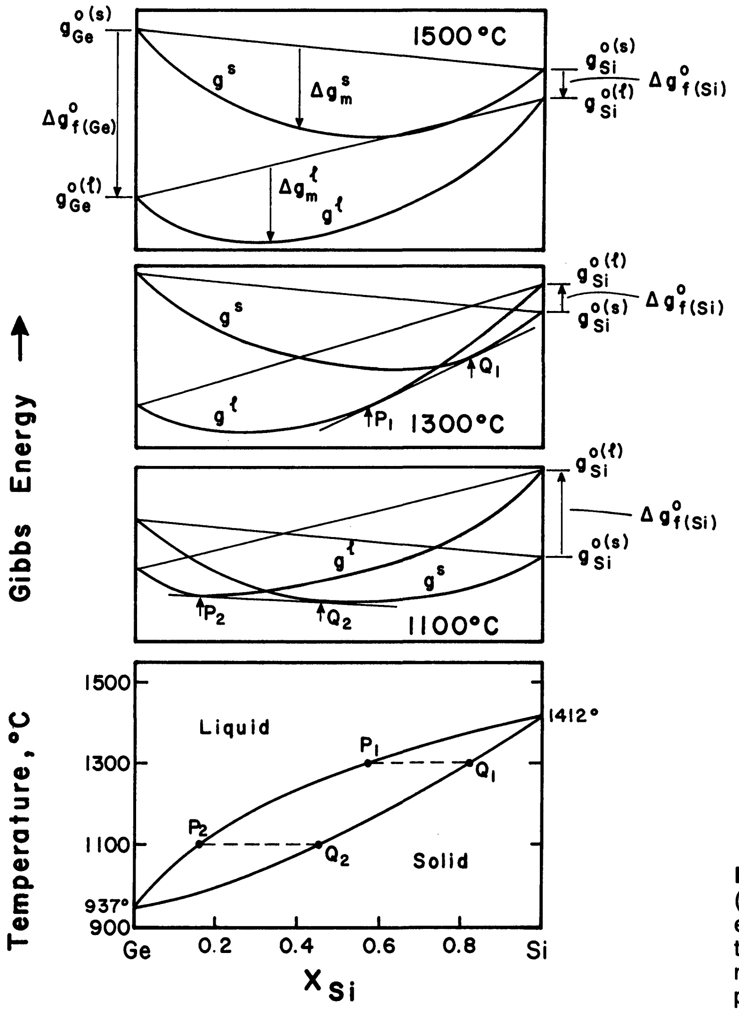

FIG. 17. The Ge-Si phase diagram (after Hansen, 1958) and Gibbs energy-composition curves at three temperatures illustrating the common tangent construction (reprinted from Pelton, 1983).

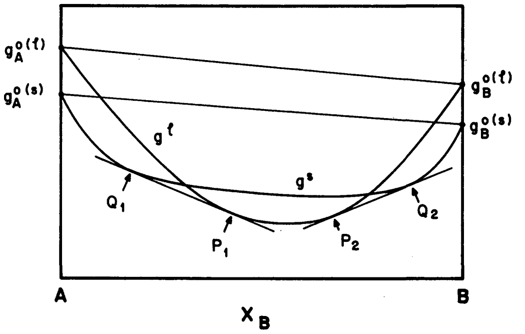

FIG. 18. Isothermal Gibbs energy- composition curves for solid and liquid phases in a system *A-B* in which $g^{E(l)} < g^{E(s)}$. A phase diagram with a minimum in the (solid + liquid) two-phase region results.

lationship, we shall briefly discuss a very simple solution model, that of a "regular" solution.

Consider a binary solution formed of atoms or molecules of the components *A* and *B* that mix substitutionally and nearly randomly. The entropy of mixing will then be nearly ideal and will give rise to the ideal mixing term in Eq. (9).

We now assume that the bond energies ϵ_{AA}, ϵ_{BB}, and ϵ_{AB} of nearest-neighbor pairs of atoms or molecules are independent of temperature and composition and that the average nearest-neighbor coordination number, *Z*, is also constant. Finally, we assume that the excess Gibbs energy of mixing, g^E, results mainly from the change in the total energy of nearest-neighbor pair bonds.

In one mole of solution there are $N^0Z/2$ nearest-neighbor pair bonds, where N^0 is Avogadro's number. Since the distribution is assumed to be random, the probability that a given bond is an A-A bond is equal to X_A^2. The probabilities of B-B and A-B bonds are, respectively, X_B^2 and $2X_AX_B$. The molar excess Gibbs energy of mixing is then equal to the sum of the energies of the nearest-neighbor bonds in one mole of solution, less the energy of the A-A bonds in X_A moles of pure A and the energy of the B-B bonds in X_B moles of pure B:

$$\begin{aligned} g^E &= (N^0Z/2)(X_A^2\epsilon_{AA} + X_B^2\epsilon_{BB} + 2X_AX_B\epsilon_{AB}) \\ &\quad - (N^0Z/2)(X_A\epsilon_{AA}) - (N^0Z/2)(X_B\epsilon_{BB}) \\ &= (N^0Z)[\epsilon_{AB} - (\epsilon_{AA} + \epsilon_{BB})/2]X_AX_B \\ &= \omega X_AX_B, \end{aligned} \tag{11}$$

where ω is a constant, and where the relation $X_A + X_B = 1$ has been used in the simplification of Eq. (11).

If A-B bonds are stronger than AA and BB bonds, then $\epsilon_{AB} < (\epsilon_{AA} + \epsilon_{BB})/2$ and $\omega < 0$, and $g^E < 0$. That is, the solution is rendered more stable. If the A-B bonds are relatively weak, then the solution is rendered less stable since $\omega > 0$ and $g^E > 0$.

In Fig. 19, several phase diagrams are shown, calculated for a hypothetical system A-B containing a solid and a liquid phase with melting points of $T^0_{f(A)} = 800$ K and $T^0_{f(B)} = 1200$ K and with entropies of fusion of both A and B set to 10 J/mol K, which is a typical value for metals. The solid and liquid phases are both regular, with temperature-independent excess Gibbs energies:

$$g^{E(s)} = \omega^s X_AX_B \quad \text{and} \quad g^{E(l)} = \omega^l X_AX_B.$$

The parameters ω^s and ω^l have been varied systematically to generate the various panels of Fig. 19.

In panel (n) both phases are ideal. Panels (l)–(r) exhibit minima or maxima depending upon the sign and magnitude of $g^{E(l)} - g^{E(s)}$. In panel (h) the liquid is ideal, but positive deviations in the solid give rise to a solid-solid miscibility gap. That is, A-B bonds are energetically so unfavorable that $g^{E(s)}$ becomes so positive that the solid solution splits into an A-rich phase and a B-rich phase.

On passing from panel (h) to panel (c), an increase in $g^{E(s)}$ results in a widening of the miscibility gap so that the solubility of A in solid B and of B in solid A decreases. Panels (a)–(c) illustrate that negative deviations in the liquid cause a relative stabilization of the liquid with resultant lowering of the eutectic temperature.

If the negative deviations in the liquid phase are sufficiently large, then the liquid solution will adopt a structure in which the number of A-B nearest-neighbor pairs in greater than the value $2X_AX_B$ for random mixing. Such *short-range ordering* is not predicted by the simple regular solution model, but can be taken into account by more sophisticated models such as the quasichemical model (Pelton and Blander, 1986).

6.2 Thermodynamic-Phase Diagram Optimization

In recent years, the development of solution models, numerical methods, and computer software has permitted a quantitative application of thermodynamics to phase-diagram analysis. For a great many systems it is now possible to perform a simultaneous critical evaluation of available phase-diagram measurements and of available thermodynamic data (calorimetric data, measurements of chemical activities, etc.) with a view to obtaining optimized equations for the Gibbs energies of each phase that best represent all the data.

The phase diagram can be calculated from these thermodynamic equations, and so one set of self-consistent equations can describe all the thermodynamic properties and the phase diagram. This technique of analysis greatly reduces the amount of experimental information needed to characterize a system fully. All data can be tested for internal consistency. The data can be interpolated and extrapolated more accurately, and metastable phase boundaries can be calculated. All the thermodynamic properties and the phase diagram can be represented and stored by means of a small set of coefficients. Finally and most importantly, it is often possible to estimate the thermodynamic properties and phase diagrams of ternary and higher-order systems from the assessed parameters for their binary subsystems. The *Calphad Journal* (Pergamon Press) is an in-

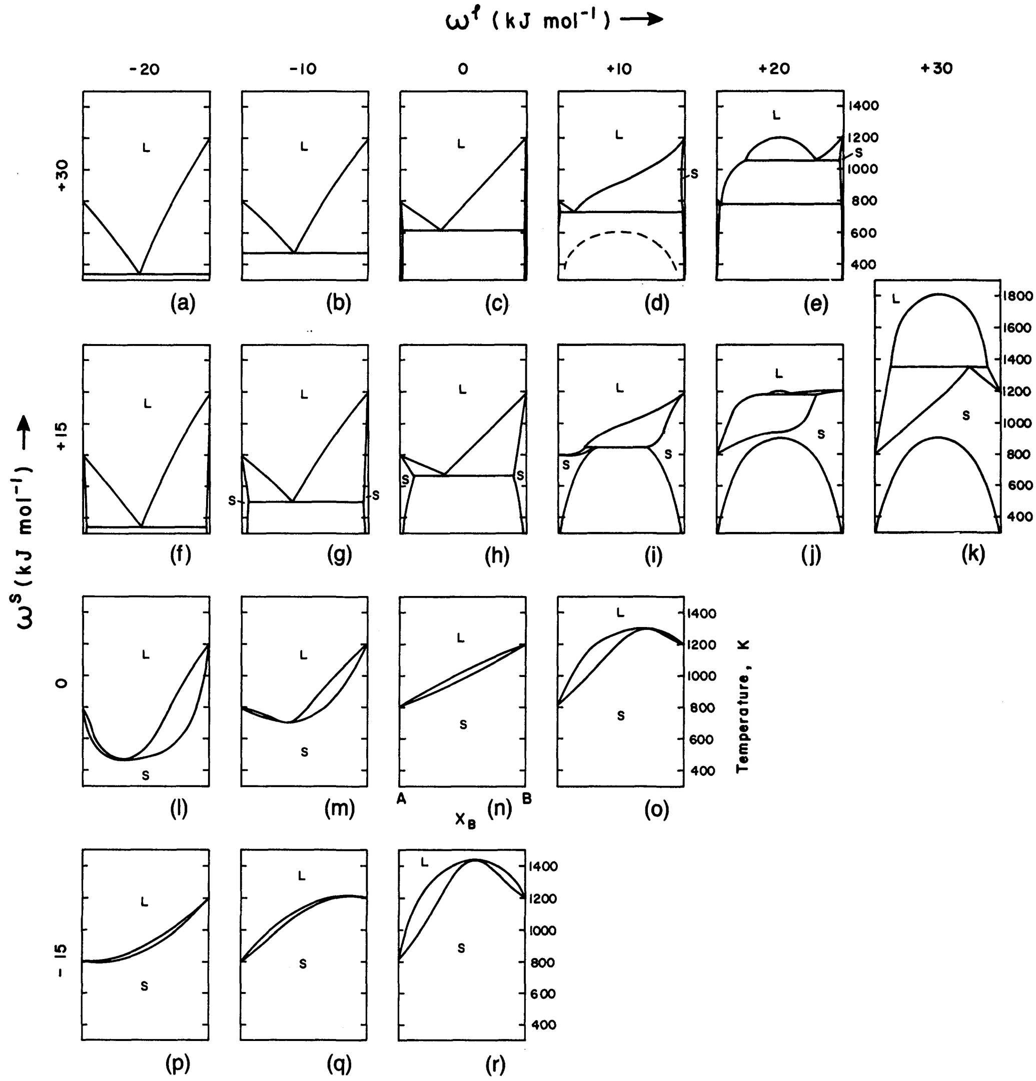

FIG. 19. Topological changes in the phase diagram for a system *A*-*B* with regular solid and liquid phases, brought about by systematic changes in the regular solution parameters ω^s and ω^l. Melting points of pure *A* and *B* are 800 and 1200 K. Entropies of fusion of both *A* and *B* are 10.0 J/mol K (Pelton and Thompson, 1975). Dotted curve in panel (d) is metastable liquid-liquid miscibility gap. (Reprinted from Pelton, 1983.)

ternational journal devoted to publishing this type of work.

7. GLASS FORMATION

In certain liquids, such as molten SiO_2, the nature of the chemical bonds is such that solidification to form an ordered crystal structure upon cooling the melt is a relatively slow process. Consequently, if such melts are cooled quickly, the disordered liquid structure will be frozen in, and the result at room temperature will be an *amorphous, glassy,* or *vitreous* phase, which can be considered to be a supercooled liquid. Glasses, being nonequilibrium phases, do not appear on equilibrium phase diagrams. Nevertheless, the equilibrium phase diagram is very useful in understanding the microstructure of materials with a glassy phase. For example, suppose that a CaO-SiO_2 melt of overall composition $X_{SiO_2} = 0.53$ (see Fig. 7) is annealed at 1500 °C and then quenched. The equilibrium phases at 1500 °C will be $CaSiO_3$ and liquid of composition $X_{SiO_2} = 0.57$ with the relative amounts given by the lever rule: (moles $CaSiO_3$)/(moles liquid) = (0.57 − 0.53)/(0.53 − 0.50) = 1.33. Upon rapid quenching, the entire liquid phase becomes a glass, and at room temperature we have (moles $CaSiO_3$)/(moles glass) = 1.33.

Most liquids, even metals, will form an amorphous phase if the quenching rate is sufficiently rapid. Generally, when a liquid is quenched to form a glass, the physical properties, such as viscosity, change from values typical of a liquid to values typical of a solid over a relatively narrow range of temperature about an average or *glass transition temperature* that depends upon the cooling rate. Glass formation is discussed in AMORPHOUS MATERIALS, STRUCTURE OF, and GLASSES.

8. ORDER-DISORDER TRANSITIONS

In solid solutions, long-range ordering can occur. For example, in the Ag-Mg system, a transformation occurs at approximately 392 °C at the composition Ag_3Mg from a low-temperature form with long-range ordering (superlattice formation) to a high-temperature form in which Ag and Mg are randomly distributed over the lattice sites. Unlike first-order phase transformations such as solidification, or the transformation between two different crystal structures, such *order-disorder transformation* can occur by atomic rearrangement over distances of the order of atomic dimensions.

A type of order-disorder transformation of importance in ferrous metallurgy is the magnetic transformation. Below its *Curie temperature* of 769 °C, Fe is ferromagnetic. Above this temperature it is not. The transformation involves a change in ordering of the atomic spins and is not first order. Additions of alloying elements will change the temperature of transformation as shown in Fig. 9.

For further discussion see ORDER-DISORDER TRANSITIONS and Pitsch and Inden (1991). Magnetic transformations are discussed by Miodownik (1982) and Inden (1982).

9. INTERFACIAL EFFECTS

In Sec. 6 it was shown that equilibrium between phases results from a minimization of the total Gibbs energy of the system. If a homogeneous solution is slowly cooled, a second phase will begin to precipitate at a temperature below which the formation of a second phase lowers the total Gibbs energy. In Sec. 6, however, it was assumed that the Gibbs energy associated with the formation of the interface between the phases is negligible. This is usually a good assumption unless the precipitating particles are very small. In the case of very small precipitates, however, the interfacial area is so large that the interfacial Gibbs energy cannot be ignored. An interface is a region in which the continuity of the structure and the bonding is interrupted. The Gibbs energy associated with the formation of an interface is thus positive. If the precipitates are very small, then, the precipitation temperature will be lowered below the value shown in the equilibrium phase diagram.

A further positive contribution to the Gibbs energy of forming a second phase arises if the interface is *coherent*, that is if there is a matching of the crystal structures across the interface. Since the lattice parameters of the two phases are not identical, such coherency gives rise to a positive strain energy which will also influence the phase equilibrium.

The existence of a strain energy in the region near the interface can result in concentration gradients in this region, since the formation of such gradients will partially decrease the strain energy. An applied stress on the sample can also affect the phase equilibria, as well as the kinetics of precipitation, because of interaction between the applied and coherency stress fields. For discussions of these topics, the reader is referred to Wagner and Kampmann (1991) and Purdy (1991).

GLOSSARY

Azeotrope: An extremum in a liquid + vapor two-phase region.

Binary System: A system of two components.

Bivariant: Having a variance of two.

Component: One of the chemical species required to define the composition of a system.

Congruent Melting: Melting to form a liquid of the same composition as the solid.

Conode: See **Tie-Line**.

Consolute Point: The summit of a miscibility gap. See also **Critical Point**.

Critical Point: A point on a phase diagram at which the properties of two phases become identical and a two-phase region ends.

Critical Pressure: The pressure at a critical point.

Critical Temperature: The temperature at a critical point.

Crystallization Path: The locus followed by the liquid composition during equilibrium cooling.

Equilibrium: A state of minimum free energy. A system at equilibrium is unchanging even over an infinite amount of time.

Eutectic Composition: The (constant) composition of the liquid during a eutectic reaction.

Eutectic Reaction: An invariant reaction during which, upon cooling, a liquid solidifies at constant temperature to give two solids (in a binary system), three solids (in a ternary system), etc.

Eutectic Temperature: The (constant) temperature at which a eutectic reaction occurs.

Eutectoid Reaction: An invariant reaction during which, upon cooling, a solid dissociates, at constant temperature, to two (or more) different solids.

Incongruent Melting: Melting (in a peritectic reaction) to form a liquid plus a solid of another composition.

Invariant: Having a variance of zero.

Isopleth: A constant-composition section through a multicomponent phase diagram.

Lever Rule: The rule by which the relative amounts of two phases existing at equilibrium can be calculated from a phase diagram.

Liquidus: The surface on a phase diagram giving the composition of the liquid phase in equilibrium with a solid phase.

Miscibility Gap: A region on a phase diagram in which two liquid phases (or two solid phases of the same crystal structure) are in equilibrium.

Mole Fraction: A composition variable. The mole fraction X_i of component i is equal to $n_i/\Sigma n_j$ where Σn_j is the sum over all the components.

Monotectic Reaction: An invariant reaction during which, upon cooling, a liquid reacts at constant temperature to give another liquid and one (or more) solids.

Nonstoichiometric Compound: A compound that exists as a single phase over a range of composition.

Peritectic Reaction: An invariant reaction during which, upon cooling, a liquid and one (or more) solids react, at constant temperature, to give one (or more) other solids. See also **Incongruent Melting**.

Phase: A region of a system of uniform physical and chemical properties.

Primary Solid: The first solid phase to crystallize upon equilibrium cooling of a single-phase liquid. Also called *pro-eutectic solid*.

Solidus: The surface on a phase diagram giving the composition of a solid phase in equilibrium with a liquid phase.

Solvus: The surface on a phase diagram giving the composition of a solid phase in equilibrium with another solid phase.

Space Model: The three-dimensional temperature-composition phase diagram of a ternary system.

Stoichiometric Compound: A compound of fixed composition.

Ternary System: A system of three components.

Tie-Line: A straight line on a phase diagram whose ends give the compositions of two phases at equilibrium. Also called *Conode*.

Tie-Triangle: A triangle on a phase diagram whose corners give the compositions of three phases at equilibrium.

Triple Point: At invariant point on a phase diagram at which three phases are at equilibrium.

Univariant: Having a variance of one.

Variance: The number of variables (temperature, pressure, compositions of the phases) which can and must be specified in order to completely specify the state of a system.

Works Cited

Bale, C. W., Pelton, A. D., Thompson, W. T. (1986), *Can. Metall. Quart.* **25**, 107–112.

Bale, C. W. (1990), *Can. Metall. Quart.* **29**, 263–278.

Eriksson, G., Pelton, A. D. (1993), *Metall. Trans. B*, **24B**, 807–816.

Eriksson, G., Wu, P., Blander, M., Pelton, A. D. (1994), *Can. Metall. Quart.* **33**, 13–22.

Hansen, M. (1958), *Constitution of Binary Alloys*, 2nd ed., New York: McGraw-Hill.

Hillert, M. (1985), *Int. Metall. Rev.* **30**, 45.

Hultgren, R., Desai, P. D., Hawkins, D. T., Gleiser, M., Kelly, K. K., Wagman, D. D. (1973), *Selected Values of the Thermodynamic Properties of the Elements and Binary Alloys*, Metals Park, OH: Am. Soc. of Metals.

Inden, G. (1982), *Bull. Alloy Phase Diag.* **2**, 412.

Köster, W., Dullenkopf, W. (1936), *Z. Metallk.* **28**, 309.

Kubaschewski, O. (1982), *Iron-Binary Phase Diagrams*, New York: Springer Verlag.

Levin, E. M., Robbins, C. R., McMurdie, H. F. (1964), *Phase Diagrams for Ceramists*, and (1969) Supplement; Levin, E. M., McMurdie, H. F. (1975) Supplement; Roth, R. S., Negas, T., Cook, L. P. (1981, 1983) Supplements; Roth, R. S., Denis, J. R., McMurdie, H. F. (1987) Supplement; Cook, L. P., McMurdie, H. F. (1989) Supplement; Mysen, B. O. (1990) Supplement; Stringfellow, G. B. (1992) Supplement, Columbus, OH: Am. Ceramic Soc.

Miodownik, A. P. (1982), *Bull. Alloy Phase Diag.* **2**, 406.

Palatnik, L. S., Landau, A. I. (1964), *Phase Equilibria in Multicomponent Systems*, New York: Holt, Rinehart and Wilson.

Pelton, A. D. (1983), in: R. W. Cahn, P. Haasen, (Eds.), *Physical Metallurgy*, Amsterdam: Elsevier, Chap. 7 (revised edition to appear in 1995).

Pelton, A. D., Blander, M. (1986), *Metall. Trans. B*, **17B**, 805–815.

Pelton, A. D., Schmalzried, H. (1973), *Metall. Trans.* **4**, 1395–1404.

Pelton, A. D., Thompson, W. T. (1975), *Prog. Solid State Chem.* **10**, Part 3, 119.

Pelton, A. D. (1991), in: R. W. Cahn, P. Haasen, E. J. Kramer (Eds.), *Materials Science and Technology*, Vol. 5, New York: VCH, Chap. 1.

Pitsch, W., Inden G. (1991), in: R. W. Cahn, P. Haasen, E. J. Kramer (Eds.), *Materials Science and Technology*, Vol. 5, New York: VCH, Chap. 9.

Prince, A. (1963), *Metall. Rev.* **8**, 213.

Prince, A. (1966), *Alloy Phase Equilibria*, Amsterdam: Elsevier.

Purdy, G. R. (1991), in: R. W. Cahn, P. Haasen, E. J. Kramer (Eds.), *Materials Science and Technology*, Vol. 5, New York: VCH, Chap. 5.

Richter, L., Demus, D., Sackmann, H. (1981), *Mol. Cryst. Liq. Cryst.* **71**, 269–291.

Schenck, H., Frohberg, M. G., Nunninghof, R. (1964), *Arch. Eisenhütten* **35**, 269–277.

Wagner, R., Kampmann, R. (1991), in: R. W. Cahn, P. Haasen, E. J. Kramer (Eds.), *Materials Science and Technology*, Vol. 5, New York: VCH, Chap. 4.

Wu, P., Eriksson, G., Pelton, A. D. (1993), *J. Am. Ceram. Soc.* **76**, 2065–2075.

Further Reading

Phase Diagram Compilations

Elliott, R. P. (1965), *Constitution of Binary Alloys*, First Supplement, New York: McGraw Hill. Second volume in the series started by Hansen (1958).

Hansen, M. (1958), *Constitution of Binary Alloys*, 2nd ed., New York, McGraw Hill. Classic compilation in the field of binary alloy phase diagrams.

Journal of Phase Equilibria (formerly *Bulletin of Alloy Phase Diagrams*) (1977–1994), Materials Park, OH: ASM International. Since 1979, the American Society for Metals, in collaboration with the National Institute of Science and Technology and with national organizations in several countries, has undertaken a project whose goal is a complete critical evaluation of all binary and ternary alloy phase diagrams. All available literature on phase equilibria, crystal structures, and often on thermodynamic properties, is cited and critically evaluated in great detail. The evaluations are peer reviewed and the majority are published in this journal.

Kubaschewski, O. (1982), *Iron-Binary Phase Diagrams*, New York: Springer Verlag. A critical compilation.

Levin, E. M., Robbins, C. R., McMurdie, H. F. (1964), *Phase Diagrams for Ceramists* (1969) Supplement; Levin, E. M., McMurdie, H. F. (1975) Supplement; Roth, R. S., Negas, T., Cook, L. P. (1981, 1983) Supplements; Roth, R. S., Denis, J. R., McMurdie, H. F. (1987) Supplement; Cook, L. P., McMurdie, H. F. (1989) Supplement; Mysen, B. O. (1990) Supplement; Stringfellow, G. B. (1992) Supplement, Columbus, OH: Am. Ceramic Soc. Phase Diagrams of over 8000 binary, ternary, and multicomponent ceramic systems, including oxides, halides, carbonates, sulfates, etc. Earlier volumes are noncritical compilations. Recent volumes include critical commentaries.

Massalski, T. (1986–1990), *Binary Alloy Phase Diagrams*, Metals Park, OH: Am. Soc. of Metals. Condensed versions of approximately 5000 of the critical evaluations published in the *Journal of Phase Equilibria*.

Moffatt, W. B. (1979–1993), *Handbook of Binary Phase Diagrams*, Schenectady, NY: General Electric Corp. A noncritical compilation of binary alloy phase diagrams.

Shunk, F. A. (1969), *Constitution of Binary Alloys*, Second Supplement, New York: McGraw Hill. Third volume in the series started by Hansen.

Wisniak, J. (1981). *Phase Diagrams: A Literature Source Book*, Vols. 1 and 2, New York: Elsevier. An extensive bibliography of binary and multicomponent phase diagrams of all types of systems: metallic, ceramic, aqueous, organic, etc.

General Reading

Alper, A. M. (Ed.) (1970–1978), *Phase Diagrams—Materials Science and Technology*, Vols. 1–5, New York: Academic. Many aspects of phase diagrams in materials science are discussed.

Bergeron, C. J., Risbud, S. H. (1984), *Introduction to Phase Equilibria in Ceramics*, Columbus, OH: American Ceramic Society. An introduction to phase diagrams with particular attention to applications in ceramic systems.

CALPHAD Journal (1977–1994), New York, Pergamon. A journal devoted to the thermodynamic calculation of phase diagrams.

Pelton, A. D. (1983), in R. W. Cahn, P. Haasen, (Eds.), *Physical Metallurgy*, Amsterdam: Elsevier, Chap. 7. Revised addition to appear in 1995. Relationships between thermodynamics and phase diagrams are discussed as are techniques of measuring phase diagrams.

Pelton, A. D. (1991), in: R. W. Cahn, P. Haasen, E. J. Kramer (Eds.), *Materials Science and Technology*, Vol. 5, New York: VCH, Chap. 1. Relationships between thermodynamics and phase diagrams are discussed.

Prince, A. (1966), *Alloy Phase Equilibria*, Amsterdam: Elsevier. Contains a detailed treatment of the geometry of multicomponent phase diagrams.

Raynor, G. V. (1970), in: R. W. Cahn (Ed.), *Physical Metallurgy*, 2nd ed., Amsterdam: North Holland, Chap. 7. Contains a discussion of techniques of measuring phase diagrams.

Rhines, F. N. (1956), *Phase Diagrams in Metallurgy*, New York: McGraw Hill. A classical discussion of phase diagrams in metallurgy.

PHASE SEPARATION

PIERRE WILTZIUS, *AT&T Bell Laboratories, Murray Hill, New Jersey, U.S.A.*

Introduction ... 323
1. **Phase Diagrams** ... 324
1.1 One-Component Systems ... 324
1.2 Multicomponent Systems ... 325
1.3 Thermodynamic Principles ... 326
2. **Nucleation and Metastable States** ... 327
2.1 Free Growth ... 329
2.2 The Lifshitz–Slyozov–Ostwald Evaporation–Condensation Mechanism ... 330
2.3 Advanced Issues ... 332
3. **Spinodal Decomposition and Unstable States** ... 333
3.1 The Early Regime ... 335
3.2 The Intermediate Regime ... 337
3.3 The Late-Stage Regime ... 338
4. **Conclusions** ... 340
Glossary ... 340
Works Cited ... 341
Further Reading ... 341

INTRODUCTION

Among the most fundamental notions about the physical world is the existence of matter in different phases. H_2O molecules, for example, can exist as either gas (water vapor), liquid (water), or crystals (ice, snow). The processes by which matter transforms from one phase to another are at the basis of numerous phenomena that either occur naturally or play a technologically important role. Of equal fundamental importance is the notion of phase coexistence (Stanley, 1971; Gunton *et al.*, 1983). The temperature of a block of ice rises as it is heated up until it reaches the melting temperature. Further addition of heat energy does not raise the temperature but is used to transform the ice crystal into water at constant temperature. Only after all the ice is melted does additional added heat raise the temperature further. The coexistence of water and ice is an example of phase coexistence for which numerous examples can be found in everyday life. In addition to the melting and freezing cycle, the evaporation and condensation cycle is commonly observed.

A further kind of phase coexistence can be found when the same chemical adopts different crystalline or amorphous structures, with sometimes very different properties. A well-known example is the element carbon, which can exist as amorphous carbon, diamond, graphite, or the recently discovered fullerenes. These different phases can coexist as evidenced by diamond crystals embedded naturally in carbon.

So far, only single-component systems have been considered, and their phase behavior is described in the first section below. If one considers two-component or multicomponent systems, the phase diagrams become rapidly more complex, and the possible combinations of coexisting phases grow rapidly in number. One example that is treated below is mixtures of two fluids. Crudely, fluid mixtures can be categorized as either miscible, like water and alcohol, or immiscible, like water and oil. There are well-known examples of binary fluids (say *A* and *B*), however, which at high temperatures are miscible and at low temperatures separate into two phases that coexist. One of the two phases is rich in fluid *A* and the other in fluid *B*. The above situation describes a system with an upper critical temperature, but many examples are known where the reverse occurs, i.e., at low temperatures a single phase exists and as the temperature is raised the physical system undergoes phase separation. This is described by a lower critical temperature. The phase separation can take place in a variety of binary mixtures ranging from liquids, polymers (de Gennes, 1979), and complex fluids (Safran and Clark, 1987) to glasses (Cahn and Charles, 1965) and metal alloys (Komura and Furukawa, 1988). In such sys-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

tems, the phase-separation process can be induced by changing thermodynamic variables, such as temperature, pressure, or volume rapidly. How the phase-separation process occurs and how it might be controlled can affect the material properties considerably and be thus of great technological importance. Examples of phase diagrams of multicomponent systems are shown in Sec. 1.2.

A system that has been rapidly quenched from a one-phase equilibrium state to a point in the phase diagram where two phases coexist in a new equilibrium is at first in a state far from equilibrium. The processes by which the system evolves into its thermodynamic state of coexisting phases are called phase separation (Gibbs, 1948). The two main mechanisms are nucleation and growth, and spinodal decomposition. The path between the initial one-phase equilibrium and the final state of coexistence determines which one of these two mechanisms takes place. Under most circumstances, this path goes through a first-order phase-transition point, and nucleation and growth is observed. Such a phase transition is called first order or discontinuous, because even for an infinitesimally small change in parameter space, the two coexisting phases have a finite difference in composition. There are singular points in thermodynamic phase space, however, where a system can enter the two-phase region continuously, and if the final state is close enough to such a critical point the difference in composition between the two phases can become vanishingly small. Here the phase-separation process is called second order.

1. PHASE DIAGRAMS

1.1 One-Component Systems

The phases of a system with a single component can be described by an equation of state that relates thermodynamic variables, such as temperature, volume, and pressure. For a one-component system, three variables uniquely define all the possible phases. The equation of state defines a surface in the three-dimensional parameter space on which each point defines an equilibrium state. In Fig. 1, projections of the pressure/volume/temperature surface are shown schematically for what could be for instance a gas like argon. The projection onto the temperature–pressure plane has three well-known regions separated by lines. The region of the solid is separated from the liquid by the fusion curve and from the gas by the sublimation curve. These curves represent states where two or even three adjoining phases coexist.

Liquid and gas are separated by the vapor-pressure curve. This line terminates at a critical point, which is of great interest. In its

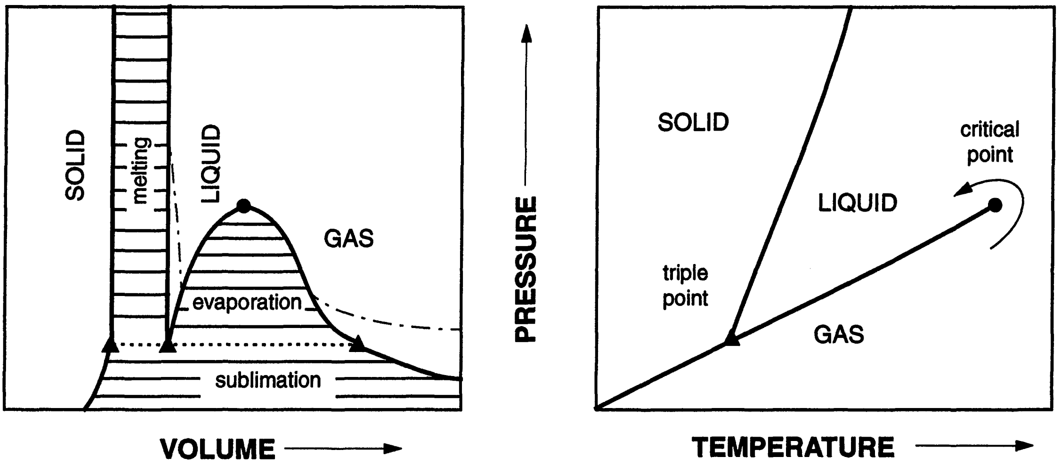

FIG. 1. Projections of the *PVT* surface of a one-component system that can exist in three different states, solid, liquid, and gas. The solid curves represent coexisting phases. The tricritical point is a point where all the phases coexist, and at the critical point two coexisting phases (gas and liquid) become identical.

vicinity, the gas and the liquid phases can be arbitrarily similar, and one can change the state of the system continuously from the gas to the liquid without crossing a phase-transition line, as indicated by the arrow. All the points on the various curves are first-order transition points. The phases coexisting on these lines are clearly distinguishable. The critical point is an exception. On approaching this point along the vapor-pressure curve, the gas and the liquid phases become less and less distinguishable and become ultimately identical. This point is also called a second-order phase-transition point. The search for a similar end point on the fusion curve has been in vain so far, and theoretical arguments have been made precluding its existence.

A further point of interest in this projection is the intersection of the three curves, which is a triple point. At this well-defined temperature and pressure, three phases of the substance coexist.

Another instructive projection is that onto the pressure/volume plane. Here the vapor-pressure curve (the gas-liquid coexistence curve) has the shape of a hump with the critical point at the top. As one reduces the volume of the system at constant temperature, the gas is compressed, and the pressure increases until the gas-liquid curve is reached, as indicated by the dashed-dotted line. Further volume reduction does not increase the pressure. Liquid droplets start to condense, and further reduction of the volume transforms more of the gas phase into the liquid phase. The horizontal lines under the coexistence curve are called tie-lines. After all the gas has been compressed to liquid, the pressure inside the liquid will increase rapidly until the coexistence between liquid and solid is reached.

The phase diagram in Fig. 1 is for illustrative purposes only. Real phase diagrams, even for one-component systems, can be much more complex. The phase diagram for helium has additional phases, such as the superfluid phase and associated with it a special point called λ point where the normal fluid, the superfluid, and the gas coexist. The phase diagram of water and ice is also of great complexity. For this single-component system, eight different crystalline structures and eight different triple points are known. Various ice phases can coexist with each other or with water. Interestingly, only one phase of liquid water is known.

The concepts described above for very familiar states of matter can be found in a multitude of dissimilar physical systems. Changing the temperature of certain materials can change them from being nonmagnetic to magnetic, conducting to superconducting, metallic to insulating. The same liquid-crystalline material can exist in a number of different structures depending on temperature.

1.2 Multicomponent Systems

An example of a two-component system is shown in Fig. 2. It is the schematic phase diagram of a mixture of a polymer and a liquid crystal (Doane *et al.*, 1986; Smith, 1993). At high temperatures, the two mix at all proportions. The pure liquid-crystalline material has a transition between an isotropic and a nematic state at a temperature T_{NI}. In the nematic phase, the liquid-crystalline molecules have orientational order but lack the translational order found in crystalline structures. At temperatures below T_{NI}, there

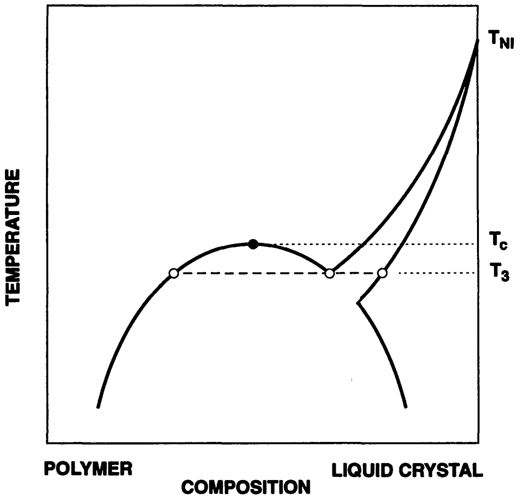

FIG. 2. Phase diagram for a mixture of a polymer and a liquid crystal in the temperature-composition space. T_{NI} is the nematic-to-isotropic transition temperature of the pure liquid-crystalline material. T_c is a critical point where the polymer-rich phase and the liquid-crystalline–rich phase become identical. T_3 is the temperature where three phases coexist, a polymer-rich phase and two liquid-crystalline–rich phases, one where the liquid crystal is isotropic and one where it is nematic.

are several regions of coexistence between two phases. Depending on composition and temperature, the two phases can be both isotropic or one can be isotropic and the other nematic. At one particular temperature T_3, three phases coexist. One is rich in polymer, and the two others are rich in liquid crystal. In the two latter phases, the liquid-crystalline material is in the isotropic or nematic phase, respectively. In addition, this phase diagram shows a critical point at T_c.

Figure 2 displays only a portion of the full phase diagram. At very high temperatures, there is a vapor phase, and at low temperatures the polymer can crystallize and the liquid crystal can assume a number of different structures with increasing degree of crystalline order. In addition, pressure and volume have to be considered as possible degrees of freedom for a full description in this case.

1.3 Thermodynamic Principles

Phase diagrams are based on thermodynamic principles and describe the conditions under which various phases of matter can coexist. The processes by which matter changes from a single-phase to a two-phase state upon sudden changes of parameters are called phase separation.

Of particular interest for the description of two fundamentally different mechanisms by which phase separation can occur are two-component fluid systems. These systems have been studied extensively and lend themselves readily for comparison between experiment and theory. The free energy of a mixture of two fluids, *A* and *B*, consists in its simplest form of two terms, an entropic term that favors disorder or mixing of the molecules *A* and *B*, and an enthalpic term that favors the proximity of similar molecules. The expression for the free energy of a mixture of two polymers has been provided by Flory and Huggins (Flory, 1953):

$$F = \frac{\phi \ln(\phi)}{N_A} + \frac{(1 - \phi)\ln(1 - \phi)}{N_B} + \chi\phi(1 - \phi), \tag{1}$$

where ϕ is the volume fraction of polymer *A* and $1 - \phi$ the volume fraction of polymer *B*. The terms N_A and N_B are the polymerization indices or number of monomers in chains *A* and *B*, and χ is the Flory interaction parameter. It has the following temperature dependence:

$$\chi(T) = a/T + b. \tag{2}$$

The free energy is plotted schematically in Fig. 3 as a function of composition for two different temperatures corresponding to two different values of χ. At temperatures above a certain critical temperature T_c, the free energy has a single minimum and the two components *A* and *B* mix at all compositions. Separating mixture 1 into two mixtures at slightly higher and lower concentrations would increase the total free energy, and the two phases could achieve a state of lower energy by remixing. This is the case for the upper curve in Fig. 3. An example of this is a mixture of water and ethanol at room temperature, for instance. As one lowers the temperature of the *A*/*B* mixture below T_c, however, the free energy has two minima and the state of the mixture depends on its global composition. Three thermodynamically different regions can be distinguished. The phase behavior of the mixture depends on the sign

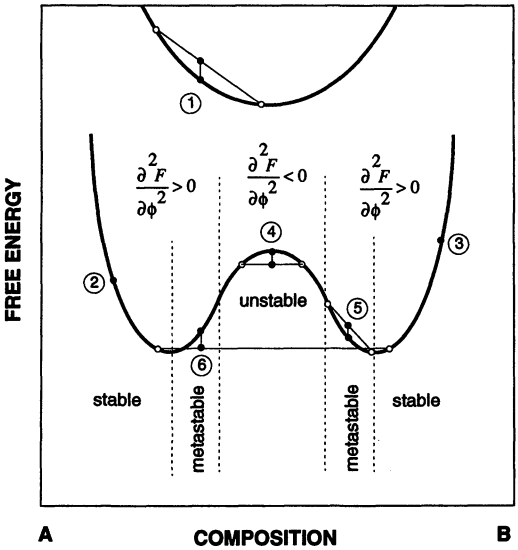
①
$\frac{\partial^2 F}{\partial \phi^2} > 0$
$\frac{\partial^2 F}{\partial \phi^2} < 0$
$\frac{\partial^2 F}{\partial \phi^2} > 0$
④
②
③
unstable
⑤
⑥
stable
metastable
metastable
stable
FREE ENERGY
A
COMPOSITION
B

FIG. 3. Free energy as a function of composition for a two-component mixture. The upper curve is a high-temperature single phase, whereas the lower curve is at low temperatures where two phases coexist. Three regions (stable, unstable, and metastable) as well as various hypothetical samples (1–6) and their stability behavior are described in the text.

of the second derivative of the free energy. Very dilute mixtures, indicated by 2 and 3, behave like the high-temperature mixtures described above. They are miscible and form a single stable phase. Mixtures with compositions in the unstable region, where the second derivative of the free energy is negative, spontaneously separate into two coexisting phases. A mixture prepared at these compositions, as indicated by 4 in Fig. 3, can have a lower global free energy by forming domains of different compositions. One of the phases is rich in component A and the other in component B. Between the unstable and the stable regions is a metastable region. Here the second derivative of the free energy is positive and small separations in composition do not lower the overall free energy, as shown for mixture 5. Formation of a domain on the other side of the maximum in the free-energy curve, however, leads to a globally lower free energy. This is shown schematically for mixture 6. This region of the phase diagram is thus metastable. It takes a triggering event such as nucleation to induce phase separation.

The transition from a single-minimum free-energy curve to one with two minima is common to phase transitions. Even though such free-energy curves help us to understand the cause of the phase transition, they are not very convenient if one is interested in the phase behavior of a given system. A complete description of the temperature dependence would require many of these curves. The important points are the inflection points, defined by the condition

$$\frac{\partial^2 F}{\partial \phi^2} = 0, \tag{3}$$

and the points separating the metastable from the stable compositions, given by

$$\left.\frac{\partial F}{\partial \phi}\right|_{\phi_1} = \left.\frac{\partial F}{\partial \phi}\right|_{\phi_2}. \tag{4}$$

A plot of these points as a function of temperature is shown in Fig. 4. It is important to note that both Figs. 3 and 4 are for model systems, with symmetry about the 50% composition line. For most real systems, the curves are asymmetric, with the maximum coexistence temperature occurring at a composition different from 50%. In Fig. 4, the three regions in the temperature-composition diagram are clearly visible. They are separated by two lines, the coexistence curve, also called binodal curve, and the spinodal curve, or stability limit. Crossing these lines induces phase separation in a sample that was originally in a single-phase state at high temperatures. The process by which the separation occurs is fundamentally different depending on whether the quench point in the temperature/composition is in the metastable region (nucleation and growth) or the unstable region (spinodal decomposition). The volumes of the coexisting phases will not necessarily be identical. An illustration of this is given in Fig. 5. Here three samples were prepared at compositions corresponding to points 1, 2, and 3 in Fig. 4. In the one-phase region, $T > T_c$, no meniscus is observed within the sample. After a quench into the unstable region (sample 1) or the metastable region (samples 2 and 3), phase separation starts. Immediately after the mixtures are quenched, they are in a highly dispersed, nonequilibrium state that has a milky appearance (Fig. 5, sample 1, $T < T_c$, immediately after quench). Under the influence of gravity, the coexisting phases are allowed to settle. In sample 1, the A-rich and B-rich phases occupy equal volumes whereas for the off-critical samples the minority phases occupy a smaller volume. The three samples in Fig. 5 have reached their new equilibrium. Both processes are explained in detail in Secs. 2 and 3. Samples at different compositions will separate into phases with compositions at the same coexistence-curve values, if quenched to the same temperature.

2. NUCLEATION AND METASTABLE STATES

A system quenched into the region between the coexistence curve and the spinodal curve is thermodynamically metastable. For short enough times, it will not demix, but remain as a one-phase system. Very small fluctuations in concentration, such as those due to thermal activity, will decay as if the mixture were in thermodynamic equilibrium. In such a supersaturated mixture, a nucleation process is needed to initiate formation of a region having a composition of the coexisting minority phase. This nucleation process can

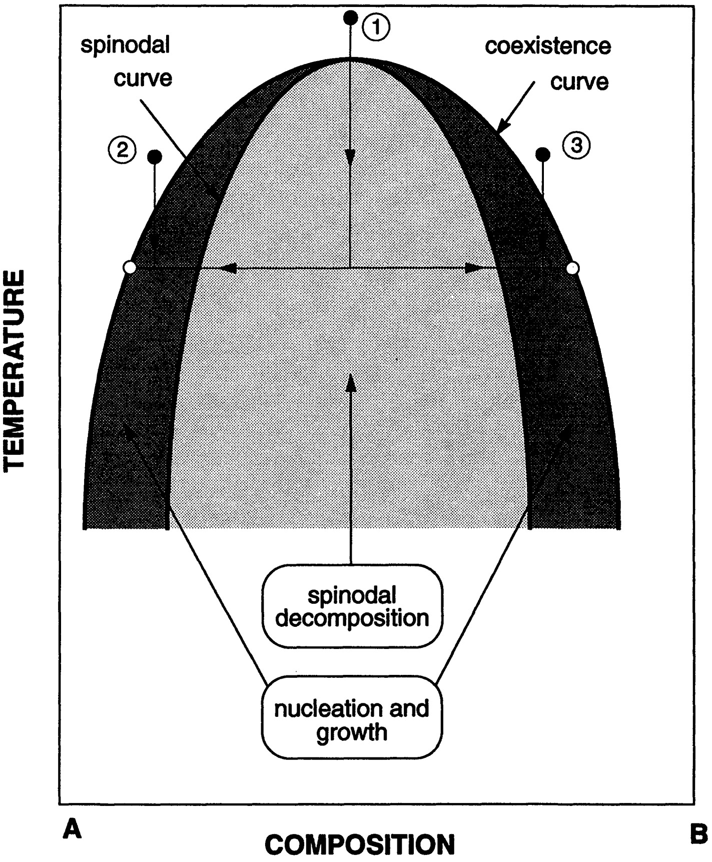

FIG. 4. Coexistence and spinodal curves for a two-component mixture in the temperature-composition space. The phase-separation process of sample 1 is called spinodal decomposition, whereas samples 2 and 3 undergo nucleation and growth.

be either heterogeneous or homogeneous. A heterogeneous process is induced by a third component, such as a contaminant or dirt in the bulk mixture, or a surface. This third component can locally decrease the free energy required for molecules of the minority phase to aggregate and create a domain of composition different from that of the initial mixture. During homogeneous nucleation, thermal fluctuations alone can cause the minority phase to nucleate without the presence of a third component. Far away from the spinodal curve, heterogeneous nucleation will dominate, whereas in the vicinity of the spinodal, the system is close enough to being thermodynamically unstable that thermal fluctuations can cause nucleation.

Once a droplet of the minority phase has been formed, it can evolve in two different ways. It can either grow by accreting more material from its supersaturated environment, or it can shrink and redissolve. The radius of the nucleated droplet determines which route it is likely to take.

Gibbs (1948) was among the first to recognize that a balance of competing effects leads to a critical droplet radius. Deep quenches or large supersaturation of the mixture favor the growth of a droplet. The deeper the quench, the more energy that can be gained by growth of the minority phase. Surface tension (interfacial free energy) is comparatively more costly for small droplets and opposes their existence. For every value of the supersaturation, there is a critical radius R_c. Droplets with a radius greater than R_c grow, and those with a smaller radius shrink. The critical radius depends on temperature, composition, and time elapsed since the system was quenched (see below).

Figure 6 shows a series of photomicrographs of quenched lead (Pb) and tin (Sn) mixtures (Hardy and Voorhees, 1988). The Sn-rich domains were held at 185 °C for varying periods of time. As time elapsed, the average particle size increased and the number of particles decreased, i.e., particle coarsening occurs.

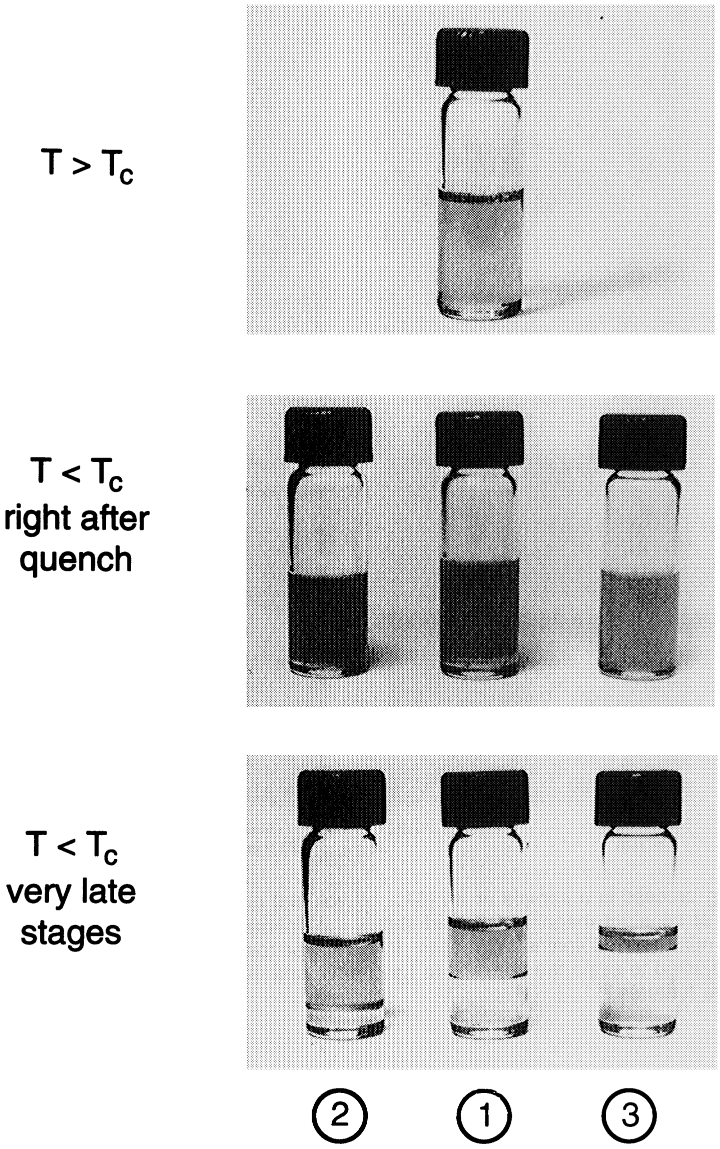

FIG. 5. Mixtures of two liquids at various compositions and temperatures. The mixture in the top vial is at a temperature above the critical temperature T_c and is in a single phase. Samples 1, 2, and 3 are prepared at compositions corresponding to the samples schematically depicted in Fig. 4. Immediately after being quenched to a lower temperature corresponding to a point in the two-phase region of the phase diagram, all three samples have a cloudy appearance. Many small domains of the two phases coexist, and with time these domains coalesce and coarsen. Finally, after the heavier phase has settled under gravity a meniscus appears separating the two coexisting phases. Note that for sample 1 the meniscus is in the middle (equal volumes of the two coexisting phases), whereas in samples 2 and 3 the volumes are unequal. These mixtures correspond to fluid *A* in Fig. 4 having a lower density than fluid *B*.

2.1 Free Growth

Droplets in a supersaturated environment are depicted schematically in Fig. 7. Here the *B* material is the minority component. Immediately after having been nucleated, each droplet is surrounded by the same composition ϕ_0, which is related to the supersaturation Δ by

$$\Delta = (\phi_0 - \phi_A)/(\phi_B - \phi_A). \tag{5}$$

As long as the degree of supersaturation does not change, there will be a concentration gradient, called the depletion layer, favoring diffusion of material to droplet, where it accumulates. This leads to growth of the droplets (first and second frames in Fig. 7). At the same time as the droplets grow, their depletion layers also grow since the total amount of materials *A* and *B* must be conserved. If the depletion layers of individual droplets do not overlap, so-called free growth is observed, with no interactions between droplets. Under these conditions, the radius of any droplet that has been nucleated with a radius greater than R_c will grow asymp-

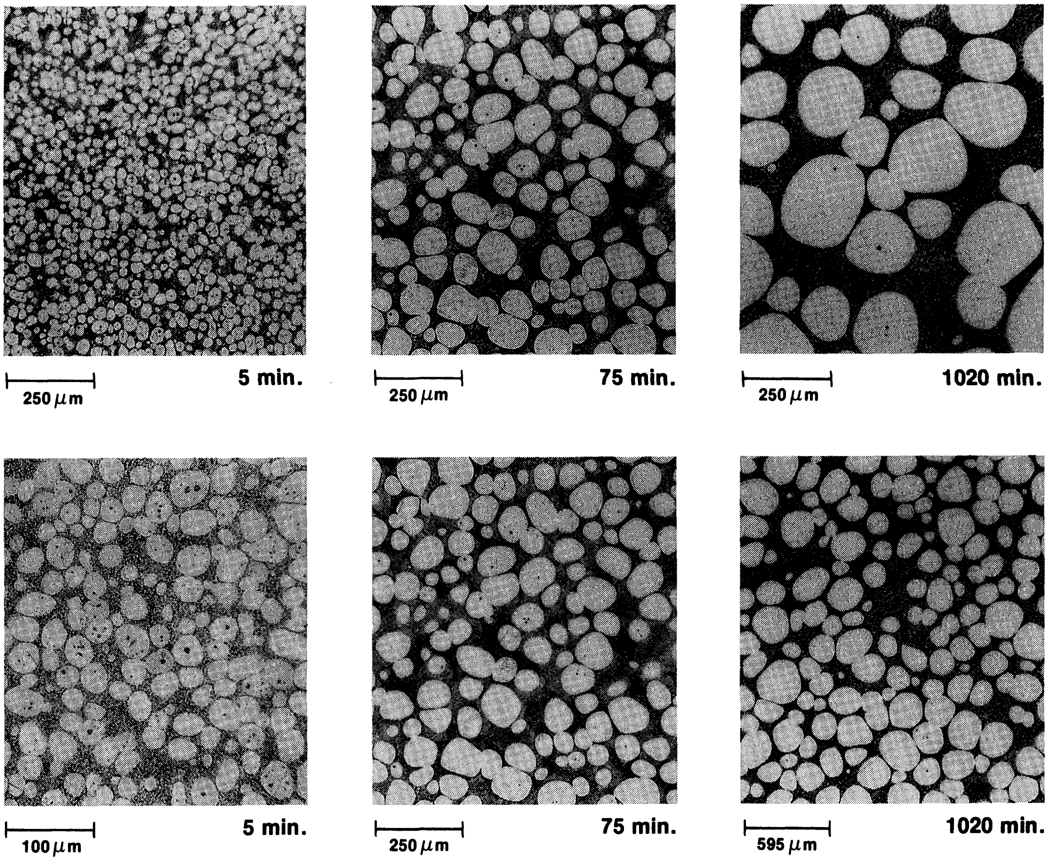

FIG. 6. Micrographs of the coarsening process in a sample of Sn (64% by volume) and Pb over a 17-h period at 185°C. The top row of pictures is at constant magnification and shows the coarsening process, where the average domain size grows and the number of domains decreases. In the lower row, the same micrographs are shown but with a magnification adjusted to scale the domains to have the same average size. Interestingly, the pictures show the same qualitative features.

totically as the square root of time. An example of this is shown in Fig. 8. Here the radii of droplets nucleated in a mixture of two polymers are shown as a function of time. The mixture had been quenched from a temperature in the one-phase region to five different temperatures in the metastable region, and the sizes of the droplets were measured by light scattering. The temperature differences ΔT cited are the depths below the coexistence curve. All the quenches were into the metastable region, since the spinodal curve lies about 1.2° below the coexistence curve for the composition used in the experiment. The radii are given in units of the critical radius, and the time is also in dimensionless units. Note that even though the raw data do not fall on a single curve, this scaled representation of the data supports the idea that a generalized dimensionless description of the growth phenomenon can be found. The slope of the experimental data is in agreement with the above-mentioned square-root growth law.

2.2 The Lifshitz–Slyozov–Ostwald Evaporation–Condensation Mechanism

The free-growth regime lasts only as long as the growing droplets are surrounded by a constant supersaturation layer. As more and more material diffuses through the depletion layers and accretes onto the droplets, the size of the depletion layers grows (see Fig. 7). Eventually, the depletion layers of neighboring droplets overlap, and competition for material from the supersaturated medium sets in, leading to a decrease of the level of supersaturation (third panel of Fig. 7). This, in turn, increases the droplet critical radius R_c. Small droplets stop growing and can even become subcritical in size, leading to their evaporation as indicated for droplet 1 in Fig. 7. This process is called the evaporation–condensation mechanism. This leads to the late-stage regime of coarsening shown in the fourth panel of Fig. 7, where the supersaturation has vanished. Diffusion due to concentration gradients ceases, but processes reducing the surface free energy can lead to further changes in the droplet size distribution. In one such mechanism, diffusion of droplets and their collisions allows them to coalesce, thus reducing the total surface free energy. Coarsening processes that lead to reduction of the interfacial area are also known as Ostwald ripening.

An important feature during all phases of coarsening is that diffusion is normal, i.e., in the direction of a concentration gradient. This is fundamentally different from the case for spinodal decomposition, which is discussed in Sec. 3.

Interestingly, the late-stage coarsening laws do not depend on whether the coarsening occurs via evaporation, diffusion, and condensation of material molecule by molecule (as might happen in a metal alloy) or by global diffusion and collision of macroscopic droplets (as can happen in fluids). A significant

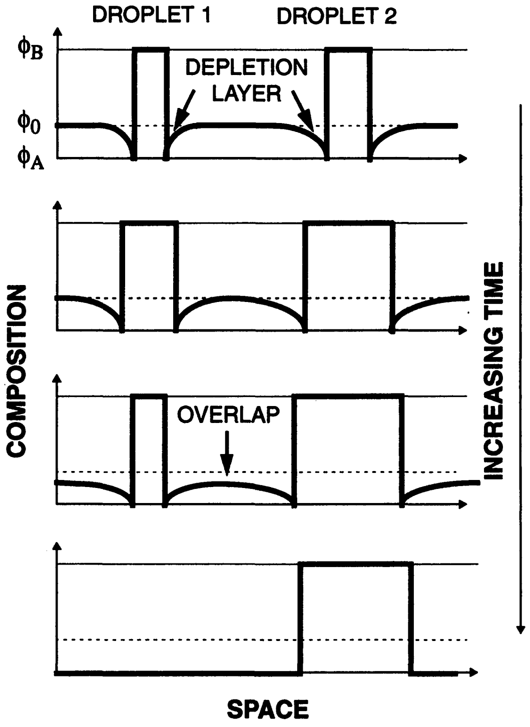

FIG. 7. Growth of droplets as a function of time in a supersaturated medium in the nucleation and growth regime. The average composition of the sample is ϕ_0, and the compositions of the coexisting phases are ϕ_A and ϕ_B, respectively. Droplet 2 is growing as a function of time, whereas droplet 1 is first growing and then shrinking.

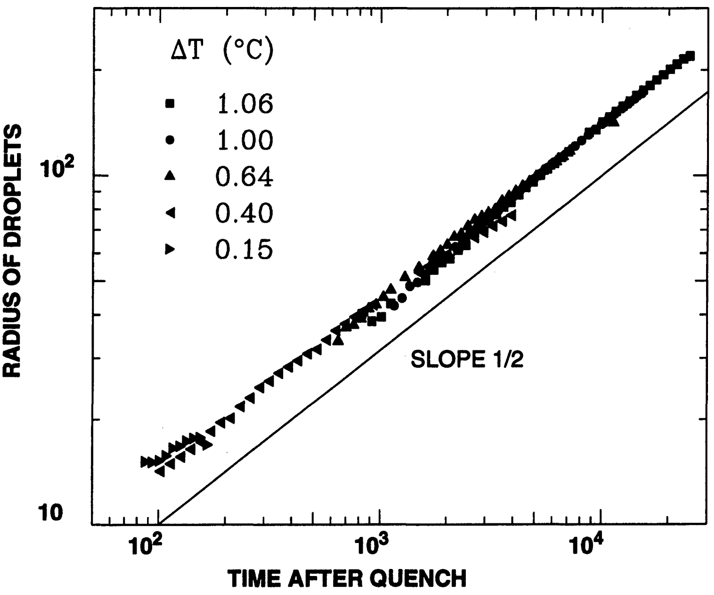

FIG. 8. The average radius of droplets of the minority phase in a phase-separating polymer mixture as a function of time for several quench depths in the metastable region of the phase diagram. A line of slope $\frac{1}{2}$ is shown for comparison.

theoretical body based both on analytical work and on numerical simulations predicts that such a system asymptotically evolves toward a polydisperse distribution of droplets. The breadth of the distribution is about 50%, and the average droplet size grows as the $\frac{1}{3}$ power of time. A full pictorial description of the various phases of the coarsening in the metastable region is shown in Fig. 9 [for details see Cumming *et al.* (1992)]. Here the evolution of the radii of individual droplets that have been "nucleated" at early times has been calculated using the differential equation proposed by Lifshitz and Slyozov (1961):

$$\frac{dR}{dt} = \frac{D}{R}\left[\Delta - \frac{\alpha}{R}\right]. \tag{6}$$

Here D is a diffusion coefficient, and α is the thickness of the interface between the droplet and the surrounding medium. Note that the absence of new droplets after time $t = 6$ is arbitrarily assumed. This case might describe a situation in which droplets are nucleated around nucleation centers of a third material (heterogeneous nucleation). Four regimes are observed. The nucleation process is assumed to be heterogeneous, since no new droplets are nucleated after a reduced time of 6. Only droplets with a radius greater than R_c (assumed to be unity here) are nucleated. After nucleation, the droplets grow as the square root of time during the free-growth regime. An interesting phenomenon is observed in this regime. An originally polydisperse distribution of droplets becomes more monodisperse. This monodispersity disappears, however, after the transition regime to the late-stage Lifshitz–Slyozov–Ostwald coarsening. The duration of the free-growth regime depends on the degree of the original supersaturation. Free growth lasts longer for deep quenches.

During the last phase, the smallest droplets become subcritical and evaporate rapidly, returning material to the majority phase, which is the source of growth of the largest droplets. This ongoing process keeps the size distribution quite broad.

2.3 Advanced Issues

The theoretical framework evoked above has been known to fail to predict correctly the quantitative features of particle size distributions. Figure 10 (Marder, 1987) shows a compilation of a number of experimentally obtained particle size distributions. It is found that the distributions are broader than the distribution predicted by Lifshitz and Slyozov (dashed line).

This failure has been attributed to the neglect of spatial distributions and interactions of the droplets that can play an important role

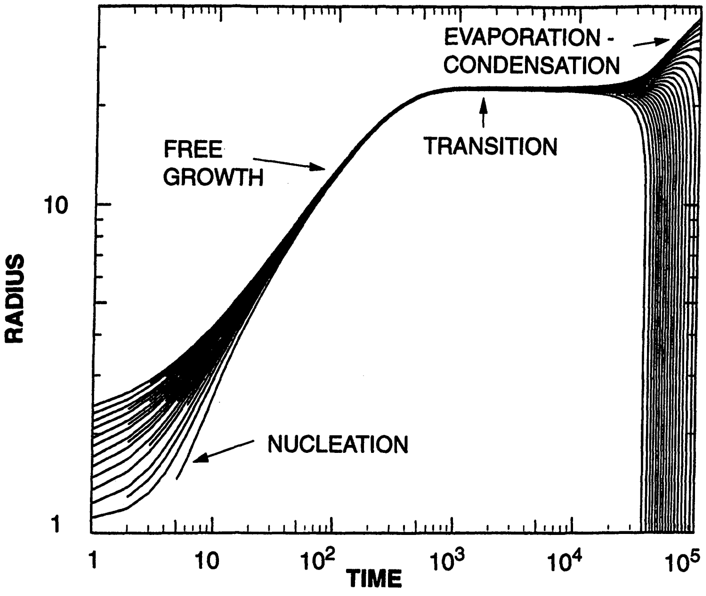

FIG. 9. Model curves obtained from numerical integration of a differential equation [Eq. (6)] proposed by Lifshitz and Slyozov. The various stages of the phase-separation process are shown. The nucleation process was halted artificially after times $t = 6$, which mimics a fast heterogeneous nucleation process at early times. During the free-growth period, the slope of the curves approaches $\frac{1}{2}$, whereas at late times an asymptotic slope of $\frac{1}{3}$ is found.

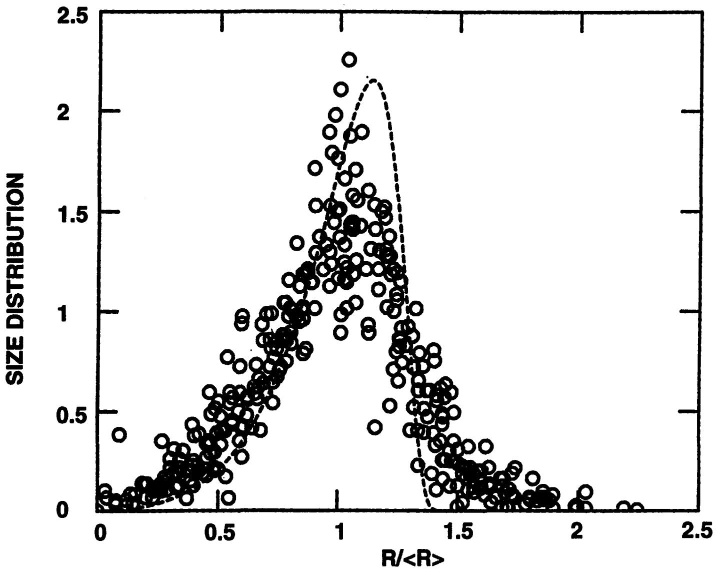

FIG. 10. Experimentally observed size distributions in a variety of phase-separating systems. For comparison, the distribution predicted by Lifshitz and Slyozov for the late-stage evaporation–condensation regime in Fig. 9 is shown (dashed line). [Adapted from Marder (1987).]

during the coarsening mechanism (Voorhees and Glicksman, 1984; Voorhees, 1992). The simplified theory predicts that all particles of a given size grow at the same rate, since only the global, mean supersaturation counts. This mean-field picture may be a good approximation for dilute systems, but in denser systems it is clear that a particle of a given radius will have a different evaporation rate depending on whether it is next to a large or a small sphere. These effects of spatial distributions around given particles are difficult to treat in analytical theories. They are currently investigated by means of large-scale numerical simulations.

3. SPINODAL DECOMPOSITION AND UNSTABLE STATES

Spinodal decomposition is the phase-separation process that occurs when a binary system is quenched from the one-phase region into the thermodynamically unstable two-phase region (see Fig. 4). A convenient way of observing this experimentally is to prepare a mixture at the critical composition where the spinodal curve and the coexistence curve coincide. For off-critical quenches, any trajectory at constant composition from the one-phase region into the two-phase region is always through a region of metastability, where the above-described nucleation and growth mechanisms can occur. In an off-critical quench, phase separation could begin locally, leading to composition inhomogeneities before spinodal decomposition begins.

Thermodynamic considerations dictate that a mixture of molecules A and B that is quenched rapidly into the unstable region separates into two different phases with compositions prescribed by the coexistence curve. Since the interfacial energy between these two phases should tend toward minimal value, the ultimate equilibrium state of a phase-separated fluid mixture in a test tube will have the denser phase at the bottom and the lighter phase on top of it with a meniscus separating the two. The scenario by which this final state is reached can be divided into at least three stages, as shown in Fig. 11. During the early, linear-stage regime, composition fluctuations with long enough wavelength will be amplified spontaneously. During this phase, the wavelength of the fluctuations, $L_m(0)$, is time independent, as explained below. During a second, or intermediate, regime, both the amplitude and the wavelength of the composition fluctuations grow. Finally, domains form with constant compositions given by the coexistence-curve values, ϕ_A and ϕ_B. Thermodynamically, the mixture is at its equilibrium point. Interfacial tension, however, drives further coarsening during the late-stage regime. These three regimes are discussed below in detail.

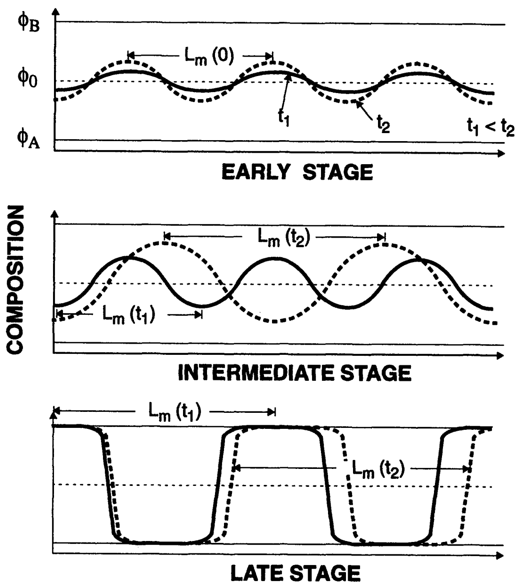

FIG. 11. Schematic description of the various stages of spinodal decomposition. $L_m(0)$ is the maximally unstable wavelength of a composition fluctuation.

The morphologies observed during the two phase-separation processes (nucleation and growth and spinodal decomposition) are quite different. Figure 12 shows photomicrographs of two different polymer mixtures undergoing phase separation. One of them (b) is prepared at an off-critical composition and has been quenched into the metastable nucleation and growth region of the phase diagram. The other (a) is at the critical composition, and it undergoes phase separation in the unstable, spinodal decomposition regime. The droplet morphology in the nucleation and growth region is obvious. Since in this case the sample was thick compared with the focal depth of the microscope, only the spheres close to the focal plane appear bright with a sharp interface. Droplets above or below the focal plane are out of focus and appear gray.

A spinodally decomposing sample (a) appears very different. This micrograph was taken during the intermediate stages where boundaries are not yet defined and only concentration gradients are present. There are

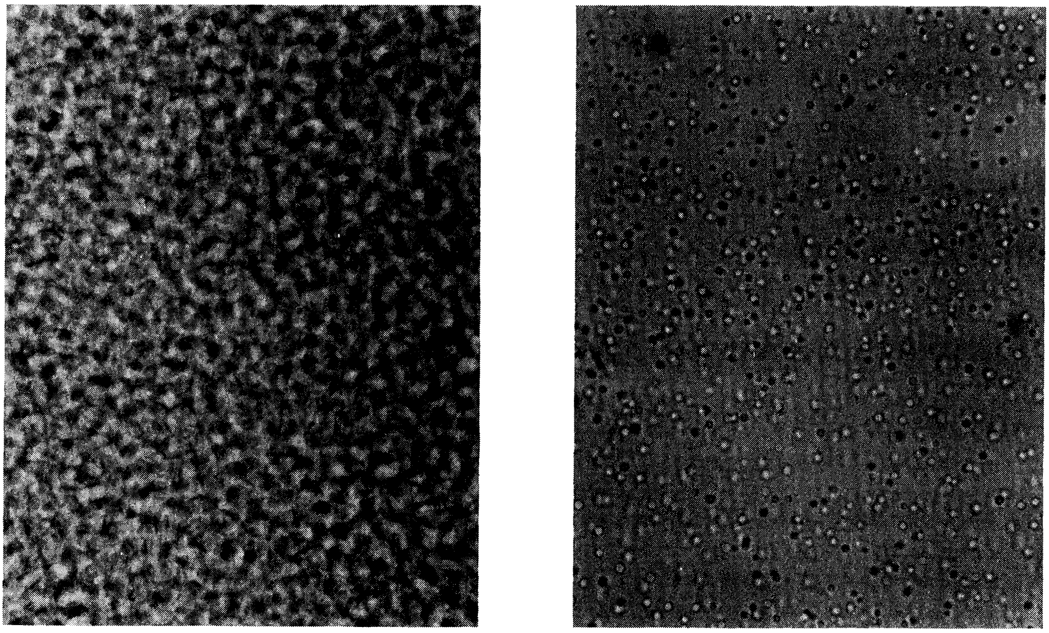

FIG. 12. Photomicrographs of two different polymer mixtures undergoing phase separation. The mixture shown in (a) was prepared at the critical composition and undergoes spinodal decomposition. The mixture shown in (b) was prepared at an off-critical composition and was quenched into the metastable region. Droplets of the minority phase nucleate and grow as a function of time.

no sharp interfaces to focus on, which explains the blurry image. More importantly, a characteristic length scale is apparent. This is even more evident in a Fourier transform, which has a ring pattern, indicative of a dominant length scale in the system. This dominant length is directly related to the maximally unstable wavelength q_m described below.

Since optical images of spinodally decomposing systems are not very easy to interpret, experimental work is typically done using light-scattering techniques (Goldburg, 1983; Hashimoto, 1988; Bates and Wiltzius, 1989; Cumming *et al.*, 1992). Figure 13 shows the time evolution of light scattered by a polymer blend undergoing phase separation. At time = 0, the sample is quenched from the one-phase region into the two-phase region. The 1-mm-thick sample is perfectly clear and does not scatter any light. The central spot is due to the laser beam directly passing through the sample without scattering. After 2 days, a ring of scattered light becomes visible. This ring is caused by fluctuations in the index of refraction (i.e., fluctuations in the concentration in the sample as schematically shown in Fig. 11). The most dominant fluctuations are those with a wavelength $L = \lambda/[2n \sin(\theta/2)]$. Here λ is the wavelength of the light, n is the index of refraction of the material, and θ is the scattering angle of the intensity maximum. As can be seen in Fig. 13, the ring becomes smaller and brighter in intensity as a function of time. This behavior reflects both an increase in the amplitude of the concentration fluctuations and a coarsening of the domains.

3.1 The Early Regime

Immediately after a quench, the thermodynamically favored state of a mixture is phase separation into two coexisting phases. This must occur, however, by transport of molecules A and B to the appropriate regions in space. From a thermodynamic point of view, a very finely divided arrangement of coexisting domains is the same as two macroscopic domains next to each other. The former arrangement of domains, however, has a huge interfacial area separating the two phases and a large energy cost associated with it. John Cahn (Cahn and Hilliard, 1958; Cahn, 1965), in seminal contributions, formalized these concepts and combined the thermodynamic free energy [Eq. (1)] with an interfacial energy into a functional form,

$$\frac{\Delta F}{k_B T} = \int \left(\frac{\phi \ln(\phi)}{N_A} + \frac{(1-\phi)\ln(1-\phi)}{N_B} + \chi(T)\phi(1-\phi) + r_p^2[\nabla\phi]^2 \right) dr, \qquad (7)$$

where r_p is the range of interactions in the

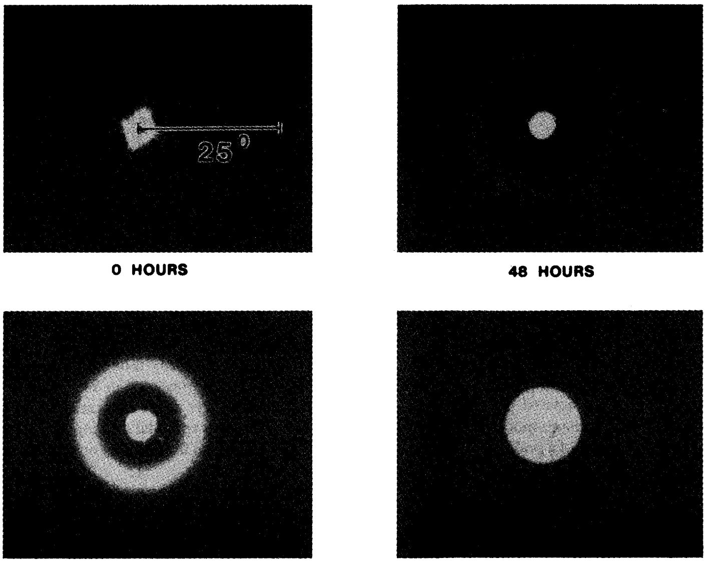

0 HOURS 48 HOURS

80 HOURS 240 HOURS

FIG. 13. Light-scattering patterns on a screen photographed at various times after a polymer mixture had been quenched to a temperature in the unstable two-phase region. (From Bates and Wiltzius, 1989.)

system and for polymer blends is of the size of the polymer chains. The phase separation in fluid system has to obey both a continuity equation,

$$\frac{\partial \phi}{\partial t} + \nabla J = 0, \tag{8}$$

for the mass flux density J, and a transport equation,

$$J = -D\nabla \mu, \tag{9}$$

relating J to the chemical potential μ via a mass diffusion coefficient D. Combining Eqs. (7)–(9) in reciprocal q space leads to the fundamental differential equation describing the phase-separation process in the unstable region,

$$\frac{\partial}{\partial t} \phi(q,t) = -Dq^2 \mu(q,t). \tag{10}$$

Unfortunately, this equation is highly nonlinear and analytically not soluble. Cahn's insight led him to develop a linear stability analysis of the differential equation. Rewriting the composition as

$$\phi(r,t) = \phi_0 + \delta\phi(r,t) \tag{11}$$

with $\delta\phi(r,t) \ll \phi_0$ and keeping only terms linear in $\delta\phi$, one finds a simpler differential equations for small fluctuations in composition:

$$\frac{\partial}{\partial t} \delta\phi = R(q)\delta\phi, \tag{12}$$

which has the solution

$$\delta\phi(q) = e^{R(q)t}. \tag{13}$$

The amplification factor $R(q)$ in the exponential has the following form:

$$R(q) = -Dq^2\left\{\frac{\partial^2 F}{\partial \phi^2} + 2q^2 r_p^2\right\}. \tag{14}$$

In the one-phase region (see Fig. 3), the second derivative of the free energy is always positive, and thus $R(q) < 0$ for all values of q. This means that spontaneous composition fluctuations of arbitrary wavelength decay exponentially, in agreement with our intuitive understanding of a system in thermodynamic equilibrium. In the two-phase region, however, $\partial^2 F/\partial\phi^2$ is negative for mixtures with compositions in the unstable region. For small enough values of q, the right-hand side of Eq. (14) is thus positive, leading to exponential growth of spontaneous fluctuations of those wavelengths as described by Eq. (13). Concentration fluctuations with a wavelength longer than a critical wavelength are unstable and grow as a function of time. Very short-wavelength fluctuations cannot grow, because they are too costly in terms of gradient or interfacial energy. Very long-wavelength fluctuations grow very slowly, because mass has to be transported over long distances. This makes it intuitively clear that there is an intermediate wavelength that grows at a maximal speed, the maximally unstable wavelength $L_m = 2\pi/q_m$. This wavelength depends on materials parameters and on the quench depth. For polymer blends, it can be expressed as follows (Binder, 1983; de Gennes, 1979):

$$q_m^2(t = 0) = \frac{3}{2R_G^2}\left(\frac{\chi - \chi_c}{\chi_c}\right), \tag{15}$$

where R_G is the radius of gyration, or size, of the polymer chains and χ_c is the value of the Flory parameter χ at the critical temperature. During the early stages, where the linear analysis applies, L_m is also independent of time. Part of this scenario is schematically depicted in the first panel of Fig. 11. One spontaneous spatial composition fluctuation with wavelength $L_m(0)$ grows as a function of time. Fluctuations with longer wavelength would also grow as a function of time, albeit more slowly, whereas very short-wavelength fluctuations would decay the same way as they would in the one-phase–equilibrium region of the phase diagram.

It is important to realize that this process is happening by "uphill" diffusion, i.e., against the concentration gradient. Uphill diffusion is the behavior opposite to "downhill" diffusion (in the direction of the concentration gradient as shown in Fig. 7).

Early stages of spinodal decomposition are difficult to observe experimentally. The composition fluctuations are small, and thus the scattering of light is weak. The early stages also evolve very rapidly, leaving little time to study them carefully. Nevertheless, a great deal of work has gone into extrapolating experimental data back to early times. This makes it possible to determine the typical unstable length scale soon after a sample has been quenched. The quench depth dependence of the measured length scales is found to be in good agreement with theoretical predictions. Qualitatively, it is found that the deeper the quench the shorter the unstable wavelengths. In addition to characteristic length scales, such an analysis also yields a quench-depth–dependent diffusion coefficient associated with the speed at which the composition fluctuations grow. The diffusion coefficients, which are formally negative quantities since the diffusion of material is against the concentration gradient, are quantitatively not understood.

3.2 The Intermediate Regime

As can be seen in Fig. 13, the maximally unstable wavelength and thus the characteristic domain size grow as a function of time. Typical data for fluid mixtures are shown in Fig. 14. The inverse of the domain size obtained from light scattering and direct visualization techniques is plotted as a function of time in reduced units. The units used in the scaling procedure are the natural time, $[Dq_m^2(t = 0)]^{-1}$, and length, $q_m^{-1}(t = 0)$, of the early-stage linear analysis. For the experiments shown in this figure, the scaling procedure is appropriate and yields a universal representation of the data. In experiments on spinodally decomposing polymer blends, systematic deviations from such universality have been found (Kedrowski *et al.*, 1993). For the reasons described above, data at very early stages, where the wave vector q_m is independent of time, are not available.

A complete theoretical understanding is still not available. This is a complicated nonlinear problem that poses a serious challenge to analytical work. The solid curve through the data in Fig. 14 is based on an empirical formula given by Furukawa (Komura and Furukawa, 1988). A number of numerical simulations have been performed on two- and three-dimensional models of these systems, and typical results are shown in Fig. 15 (Bortz *et al.*, 1974; Huse, 1986; Amar *et al.*, 1988). These are extensive Monte-Carlo simulations of domain growth in the two-dimensional spin-exchange model of a system of spins that can point either up or down. Such a system has a high-temperature phase where the spins are disordered and a low-temperature phase where they want to be organized in domains. In this particular model, the total number of up and down spins is conserved. The lattice of spins is shown at three different times, and the coarsening of the domains is apparent. The average domain size in this simulation grows as the $\frac{1}{3}$ power of time, which interestingly is the same exponent as that predicted for the disconnected dilute droplets observed in the Lifshitz–Slyozov model. The same power-law

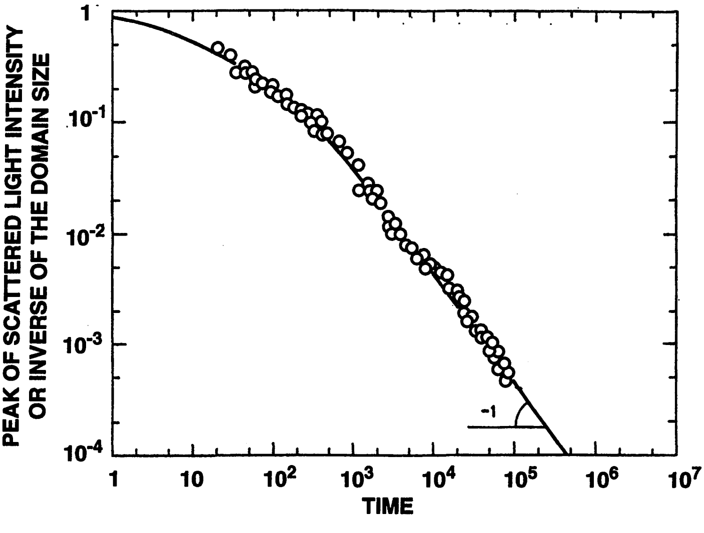

FIG. 14. The inverse domain size is plotted as a function of time for a variety of different liquid mixtures quenched to several temperatures in the unstable region. Both the domain sizes and the time have been scaled with appropriate scale factors. The asymptotic linear growth as a function of time is predicted for a growth of bicontinuous phases, where the growth mechanism happens via hydrodynamic flow. [Adapted from Beysens *et al.* (1988).]

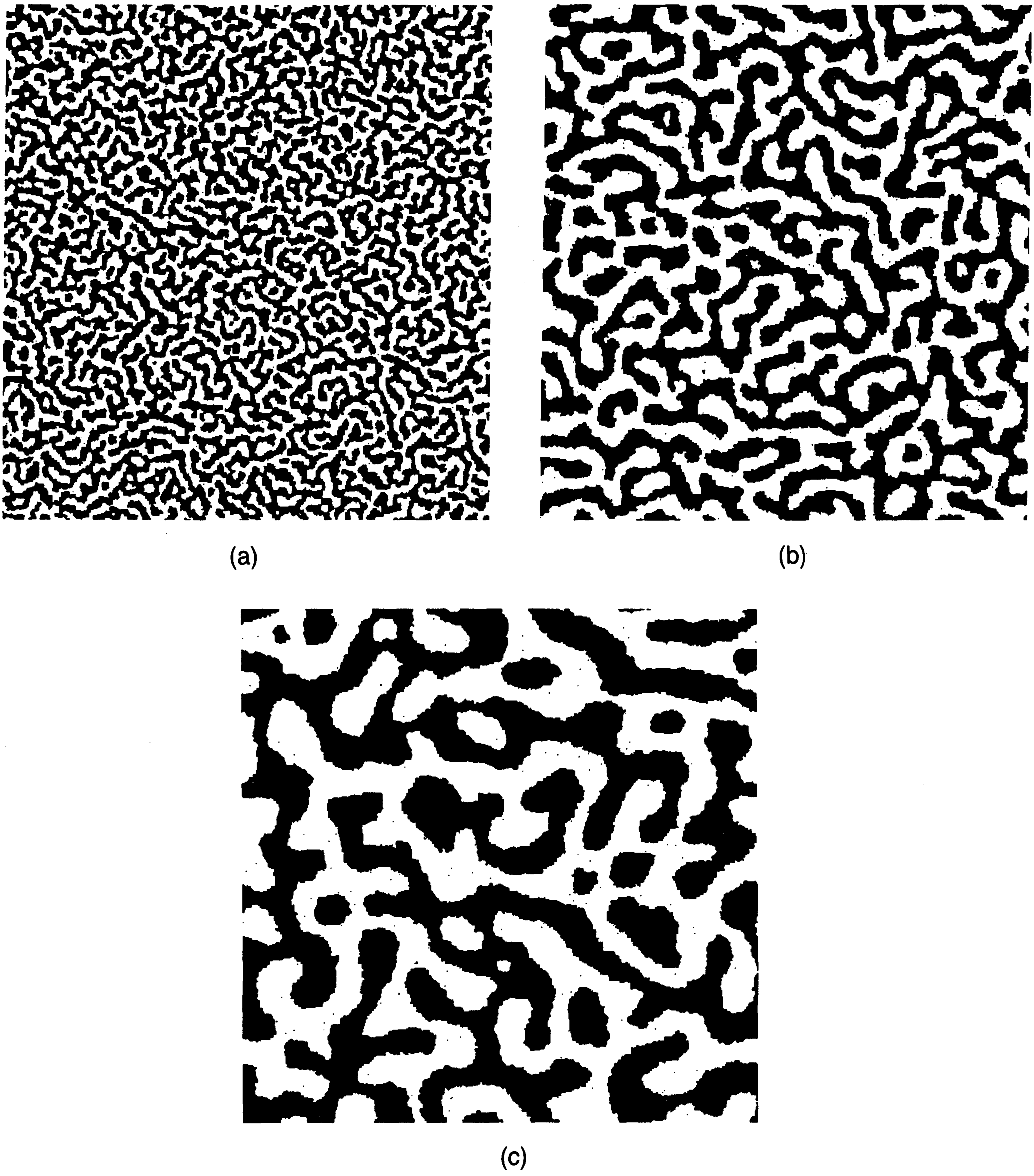

(a) (b)

(c)

FIG. 15. Numerical computer simulations of a two-dimensional spin-exchange model displaying the qualitative features of phase separation in physical systems. Here a given site can be occupied by either a spin up (white) or a spin down (black). The total number of up and down spins is conserved. Parts (a)–(c) are at increasing times. [From Amar *et al.* (1988).]

behavior has also been observed in three-dimensional simulations. This growth law has proven remarkably universal and independent of details of the models.

3.3 The Late-Stage Regime

Once the composition within the phase-separated domains has reached the values given by the coexisting curves (see Fig. 11), the only driving force is reduction of the interfacial energy. Some of the mechanisms described above lead to the $\frac{1}{3}$-power law behavior of the growth. As can be seen in Fig. 14 (Beysens *et al.*, 1987), however, in fluid mixtures at late stages a linear growth law (slope of -1) of domain size as a function of time is observed. In a very insightful paper, Siggia (1979) described the basic physical mechanism leading to this growth law. It hinges on the fact that in phase-separating mixtures with both phases topologically fully connected, fluid flow can occur throughout these bicontinuous phases. Analysis of the underlying hydrodynamic equations leads to the prediction for the linear growth law.

It has been shown experimentally and in simulations that during the late stages of domain coarsening the structures are self similar. That is, properly rescaled, the three panels in Fig. 15 have the same global properties. It is impossible to tell which are at earlier times and which at later. The self similarity is evidenced by structure factors or pair-correlation functions, which have a generalized form that depends only on a scaled spatial variable. In Fig. 16(a), several

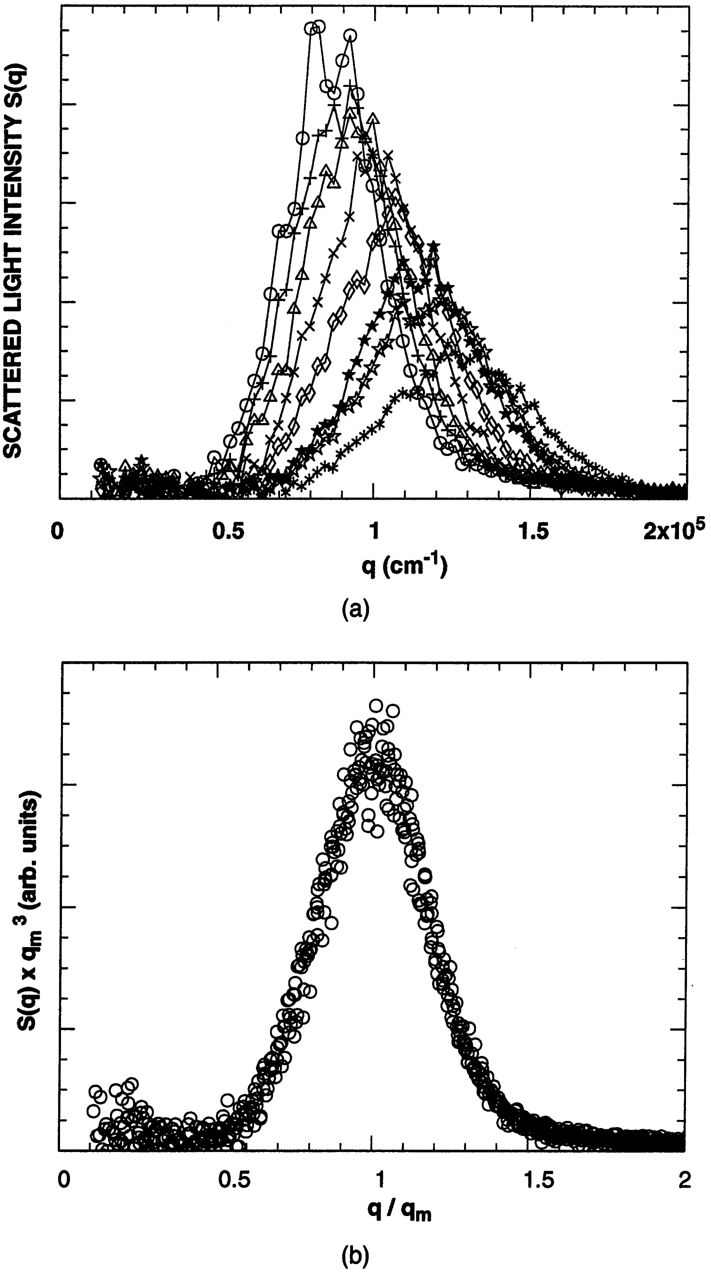

FIG. 16. (a) Light-scattering data for a polymer mixture undergoing spinodal decomposition for various times after the mixture has been quenched into the unstable region. The peak intensity is seen to increase with time (the earliest time is indicated by stars and the latest time by open circles) while its position decreases to smaller wave numbers. This indicates growth of spatial and compositional amplitude fluctuations. (b) The same data rescaled in a way appropriate to demonstrate self similarity of the time-dependent phase-separating structures. [Adapted from Kedrowski *et al.* (1993).]

such structure factors of phase-separating polymer blends are shown in their unscaled form. These are obtained from light-scattering measurements on samples similar to those shown in Fig. 13. In Fig. 16(b), the same structure factors are shown in a rescaled form. Superposition of the structure factors is in agreement with the concept of self similarity, which has been formalized by Furukawa in a scaling hypothesis.

4. CONCLUSIONS

Phase-separation phenomena can be broadly classified into two different categories depending on whether the processes occur in the thermodynamically unstable or metastable region of the phase diagram. Our insight both in nucleation and growth and in spinodal decomposition is based on work spanning well over a hundred years. Although the experimental and theoretical understanding is quite good, there are a number of issues that are not resolved. These include the quantitative understanding of nucleation barriers, the kinetics of phase separation, and finite-volume effects during late-stage coarsening. Most of the interest in the topic of phase separation has been driven by the challenging theoretical and experimental problems. There are some applications worth mentioning. Polymer-dispersed liquid crystals (PDLC) mentioned in Sec. 1.2 (Doane *et al.*, 1986; Smith, 1993) have emerged recently as switchable windows and even as candidates for flat-panel displays. PDLCs are made through a phase-separation process that results in a dispersion of micron-size droplets of liquid crystal trapped in a polymer matrix. An electric field can align the nematic textures inside the droplets, resulting in large changes in their scattering properties. It is possible to switch the optical appearance of a PDLC from transparent to milky white with reasonably small electric fields. The electro-optic properties of PDLCs depend largely on the morphology of the droplets formed during the phase-separation process. Predicting and controlling these morphologies is thus of great interest. In other applications, it can be desirable to tailor, e.g., the mechanical properties of a polymer blend by mixing various materials. Unfortunately, this results quite often in macroscopic phase separation of the components without the desired effect. A better understanding of the phase behavior and the separation processes could lead to better control.

GLOSSARY

Coarsening: Growth process of the minority-phase droplets during the nucleation and growth phase-separation process.

Coexistence Curve: Line separating the thermodynamically stable one-phase region of the phase diagram from the metastable two-phase region. Also known as binodal curve.

Critical Composition: Composition of a two-component mixture equal to the composition where the coexistence curve and the spinodal curve touch.

Critical Radius: Droplets of the minority phase smaller than the critical radius shrink and ultimately vanish from a phase-separating mixture. These droplets are also called subcritical.

Critical Temperature: Phase-separation temperature of a mixture prepared at the critical composition.

Depletion Layer: Region of lower concentration surrounding a droplet of the minority phase.

Liquid Crystal: Material with properties intermediate between those of liquids and crystals. Many different liquid-crystalline structures can be distinguished according to their degree of translational and/or orientational order.

Majority Phase: The phase in a two-component phase-separating mixture that occupies more volume than the minority phase.

Minority Phase: The phase in a two-component phase-separating mixture that occupies less volume than the majority phase.

Nucleation: Creation of a domain of the minority phase in a system undergoing phase separation in the metastable region of the phase diagram. The nucleation process can be heterogeneous, if it is induced by a third component, or homogeneous, if it is due to thermal fluctuations.

Self Similarity: Property of a system that changes, e.g., with time, but has qualitatively the same features if the dimensions are rescaled appropriately.

Spinodal Curve: Line separating the thermodynamically metastable two-phase region from the thermodynamically unstable two-phase region of the phase diagram.

Spinodal Decomposition: Phase-separation process in the unstable region of the phase diagram.

Supersaturation: Measure of the quench depth below the coexistence curve. See Eq. (5).

Works Cited

Amar, J. G., Sullivan, F. E., Mountain, R. D. (1988), *Phys. Rev. B* **37**, 196–208.

Bates, F. S., Wiltzius, P. (1989), *J. Chem. Phys.* **91**, 3258–3274.

Beysens, D., Guenoun, P., Perrot, F. (1987), *Phys. Rev. A* **38**, 4173–4185.

Binder, K. (1983), *J. Chem. Phys.* **79**, 6387–6409.

Bortz, A. B., Kalos, M. H., Lebowitz, J. L., Zendejas, M. A. (1974), *Phys. Rev. B* **10**, 535–541.

Cahn, J. W. (1965), *J. Chem. Phys.* **42**, 93–99.

Cahn, J. W., Charles, R. J. (1965), *Phys. Chem. Glasses* **6**, 181–191.

Cahn, J. W., Hilliard, J. E. (1958), *J. Chem. Phys.* **28**, 258–267; (1959), *J. Chem. Phys.* **31**, 688–699.

Cumming, A., Wiltzius, P., Bates, F. S., Rosedale, J. H. (1992), *Phys. Rev. A* **45**, 885–897.

de Gennes, P. G. (1979), *Scaling Concepts in Polymer Physics*, Ithaca: Cornell University Press.

Doane, J. W., Vaz, N. A., Wu, B.-G., Zumer, S. (1986), *Appl. Phys. Lett.* **48**, 269–271.

Flory, P. (1953), *Principles of Polymer Chemistry*, Ithaca: Cornell University Press.

Gibbs, J. W. (1948), *Collected Works*, Vol. 1, New Haven, CT: Yale University Press.

Goldburg, W. I. (1983), in: H. Z. Cummins, A. P. Levanyuk (Eds.), *Light Scattering Near Phase Transitions*, Amsterdam: North-Holland.

Gunton, J. D., San Miguel, M., Sahni, P. S. (1983), in: C. Domb, J. L. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 8, London: Academic Press.

Hardy, S. C., Voorhees, P. W. (1988), *Metall. Trans. A* **19**, 2713–2721.

Hashimoto, T. (1988), *Phase Transitions* **12**, 47–119.

Huse, D. A. (1986), *Phys. Rev. B* **34**, 7845–7850.

Kedrowski, C., Bates, F. S., Wiltzius, P. (1993), *Macromolecules* **26**, 3448–3454.

Komura, S., Furukawa, H. (1988), *Dynamics of Ordering Processes in Condensed Matter*, New York: Plenum Press.

Lifshitz, I. M., Slyozov, V. V. (1961), *J. Phys. Chem. Solids* **19**, 35–50.

Marder, M. (1987), *Phys. Rev. A* **36**, 858–874.

Safran, S. A., Clark, N. A. (1987), *Physics of Complex and Supermolecular Fluids*, New York: Wiley Interscience.

Siggia, E. D. (1979), *Phys. Rev. A* **20**, 595–605.

Smith, G. W. (1993), *Int. J. Mod. Phys. B* **7**, 4187–4213.

Stanley, H. E. (1971), *Introduction to Phase Transitions and Critical Phenomena*, New York: Oxford Science Publications.

Voorhees, P. W., Glicksman, M. E. (1984), *Metall. Trans. A* **15**, 1081–1088.

Voorhees, P. W. (1992), *Annu. Rev. Mater. Sci.* **22**, 197–215.

Further Reading

Binder, K. (1987), *Colloid Polym. Sci.* **265**, 273–288.

Cahn, J. W. (1968), *Trans. Metall. Soc. AIME* **242**, 166–180.

Gunton, J. D., San Miguel, M., Sahni, P. S. (1983), in: C. Domb, J. L. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 8, London: Academic Press.

Stanley, H. E. (1971), *Introduction to Phase Transitions and Critical Phenomena*, New York: Oxford Science Publications.

PHASE TRANSITIONS: RENORMALIZATION AND SCALING

Franz Schwabl, *Institut für Theoretische Physik, Physik-Department der Technischen Universität München, Garching, Germany*

Uwe Claus Täuber, *Lyman Laboratory of Physics, Harvard University, Cambridge, Massachusetts, U.S.A.*

Introduction 343
1. **Phase Transitions** 344
1.1 Phenomenology and Basic Concepts 344
1.2 Examples 347
1.3 Universality 348
1.4 Scaling Laws 349
2. **Mathematical Representations** 350
2.1 Microscopic Hamiltonians 350
2.2 Ginzburg–Landau Functional 351
2.3 Landau–Ginzburg Approximation 352
2.3.1 Mean-Field Theory 352
2.3.2 Gaussian Fluctuations 353
3. **Renormalization Group** 355
3.1 Introductory Remarks 355
3.2 Real-Space Renormalization Group 355
3.2.1 1D Ising Model 356
3.2.2 2D Ising Model 357
3.2.3 General Real-Space RG Transformations 360
3.2.4 Emergence of Scaling Laws 360
3.3 Momentum-Shell Renormalization Group 361
3.3.1 Wilson's RG Scheme 361
3.3.2 Gaussian Model 362
3.3.3 Perturbation Theory and ϵ Expansion 362
3.3.4 More Sophisticated Field-Theoretical Methods 365
4. **Scaling Laws in Statistical Physics** 366
4.1 Phase Transitions and Properties of Phases 366
4.2 Disordered Systems 366
4.3 Polymers, Membranes 367
4.4 Critical Dynamics and Nonequilibrium Phenomena 368
Glossary 369
Works Cited 370
Further Reading 371

INTRODUCTION

The fundamental laws of nature governing the properties of matter possess distinguished symmetry properties: They are invariant with respect to translation in space and time, to rotations, and to inversions. The states of matter that we observe, however, usually do not display the full symmetries of the underlying basic principles. Rather, the atoms or molecules of a specific substance can be found in certain condensed phases consisting of enormously large numbers (typically 10^{23} cm^{-3}) of constituents, and these phases may be classified according to some specific symmetry transformation under which the system remains invariant (a subgroup of the original symmetry group; see Group Theory). In passing, we note that even the "fundamental" laws themselves, like the electroweak interaction and quantum chromodynamics, may have originated in the transition to a broken-symmetry state of an even more fundamental and more symmetric basic law. In this article, we shall be solely concerned with matter governed by the Schrödinger equation and electromagnetic forces.

The macroscopic properties of the phases of condensed matter, whose stability regimes are determined by the internal interactions of the particles involved, as well as by the

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

external conditions (temperature, pressure, electric and magnetic fields, etc.), can be very different. Upon a change of the external control parameters, the system may undergo a phase transition to a new macroscopic phase, thus lowering its total free energy (see THERMODYNAMICS, EQUILIBRIUM). Frequently, the internal energy of the system as a function of the control variables is minimized by the change of its thermodynamic phase, but there are also important cases where the driving force toward a new equilibrium state is a gain of entropy.

The stability of a particular phase and the implications of phase transitions are of course of prominent importance for applications and technology. Moreover, these issues provide a considerable intellectual challenge for their understanding in the framework of statistical mechanics (see STATISTICAL MECHANICS, CLASSICAL; STATISTICAL MECHANICS, QUANTUM). Accordingly, methods need to be developed in order to deal with a strongly interacting system containing a large number of particles, and near its stability boundary.

Fortunately, it is not necessary to study all the microscopic degrees of freedom in detail; instead it suffices to investigate the statistical mechanics of a few macroscopic entities (which, however, usually still poses a formidable task), namely the order parameter characterizing the specific phase transition and distinguishing the phases involved and its correlation functions, as well as the dependence on the external fields serving as control parameters (see CATASTROPHE THEORY).

Furthermore, near a critical point in the phase diagram the fluctuations of the order parameter become increasingly strong and long-range correlations emerge in the system, although the interactions of the particles are genuinely short ranged. Therefore, any approximate approach that neglects these fluctuations entirely, or merely treats them in some perturbative manner, is at least questionable.

On the other hand, because of the almost entirely collective behavior in the vicinity of a critical point, the microscopic details of the system become irrelevant there, and its long-wavelength (and low-frequency) properties tend to display *universality* (*q.v.*), i.e., in some asymptotic regime they depend solely on the symmetry and number of components of the order parameter, on the space dimension, and possibly on the long-range nature of the interactions. Historically, this emergence of universal behavior has provided one of the essential keys for a mathematical description and profound theoretical understanding of critical phenomena, in the framework of *renormalization-group* (RG) theory. In turn, the RG approach has laid a solid foundation for the striking phenomena of scale invariance and universality near second-order phase transitions.

Originally, RG concepts were developed in the context of elementary-particle quantum field theory. Within the theory of phase transitions and critical phenomena, scaling ideas had been developed before, but their sound justification was only provided when Wilson formulated a version of the RG approach designed to focus on the underlying structure of the nonlinear field theories involved, a procedure that also allowed for systematic and detailed calculations (Wilson and Kogut, 1974; Wilson, 1975). This considerable breakthrough has not only led to an enormous increase of the amount of knowledge and to a deeper understanding of the properties of condensed matter (see, e.g., Chaikin and Lubensky, 1995), but has in turn also provided important insights for the formulation and treatment of quantum field theories, from where it had originated (see UNIFIED FIELD THEORIES). For a thorough account of the historical development of the field, see CRITICAL PHENOMENA.

Remarkably, RG techniques, along with the ensuing notions of fixed points governing universal physical properties, relevant and irrelevant perturbations, the emphasis on the long-wavelength and low-frequency collective behavior, and the emergence of scaling laws have matured into a most important paradigm and tool well beyond the original realm of second-order phase transitions, and have found fruitful applications in a large variety of different fields of intensive current research (see Sec. 4).

1. PHASE TRANSITIONS

1.1 Phenomenology and Basic Concepts

At a phase transition, the free energy F (or some other appropriate thermodynamic potential; see STATISTICAL MECHANICS, CLASSI-

CAL) is a nonanalytic function of a control parameter. According to a standard classification scheme introduced by Ehrenfest, a phase transition is called of *n*th order if at least one of the *n*th derivatives of F is discontinuous or singular at the transition, while all the $(n - 1)$th derivatives are still continuous. In order to distinguish the different phases, a macroscopic *order parameter* ϕ is introduced, which is zero in one phase and finite in the other. If ϕ is given by the first derivative of the thermodynamic potential with respect to its conjugate external field h, its value will jump at a first-order transition; at a second-order transition, ϕ is to be a continuous function, but its derivatives will be singular or discontinuous.

To be specific, one may refer to a ferromagnetic material undergoing a second-order phase transition from the paramagnetic to the ferromagnetic phase at the critical (Curie) temperature T_c. The basic microscopic entities are the magnetic moments $\phi(\mathbf{x})$ at site $\mathbf{x}$, also called local order parameter (in this article, $\mathbf{x}$ denotes a d-dimensional vector referring to a point either in a discrete lattice or in a continuum). The macroscopic order parameter (the spontaneous magnetization) is then given by the thermal average $\phi = \langle\phi(\mathbf{x})\rangle$. In anisotropic crystals, the local order parameter may assume just the two values $\phi(\mathbf{x}) = \pm 1$ (referring to spin up and down, respectively). Accordingly, the (average) order parameter may point up or down. Whichever of the equivalent directions is chosen, the original discrete symmetry is spontaneously broken for $T < T_c$ [see Fig. 1(a)]. In cubic crystals, the order parameter may be a three-dimensional vector $\boldsymbol{\phi}(\mathbf{x})$, and point in an arbitrary direction. In this case, in the ordered phase a continuous (rotational) symmetry is broken.

The singularities in the vicinity of such continuous phase transitions generally assume the form of power laws, with certain *critical exponents* characterizing the transition. In case the temperature T serves as control parameter, with a second-order phase transition occurring at the critical temperature T_c, the most important examples of such power laws, and associated exponents, are the behavior of the order parameter near the critical point T_c [$\tau = (T - T_c)/T_c \ll 1$] [see Fig. 1(a)],

$$\phi \propto (-\tau)^{\beta} \quad \text{for } T < T_c, \tag{1}$$

the divergence of the isothermal susceptibility on approaching the critical point (from both sides of the transition),

$$\chi = \left(\frac{\partial \phi}{\partial h}\right)_T \propto |\tau|^{-\gamma}, \tag{2}$$

the power law for the critical isotherm,

$$h \propto \phi^{\delta} \quad \text{for } T = T_c, \tag{3}$$

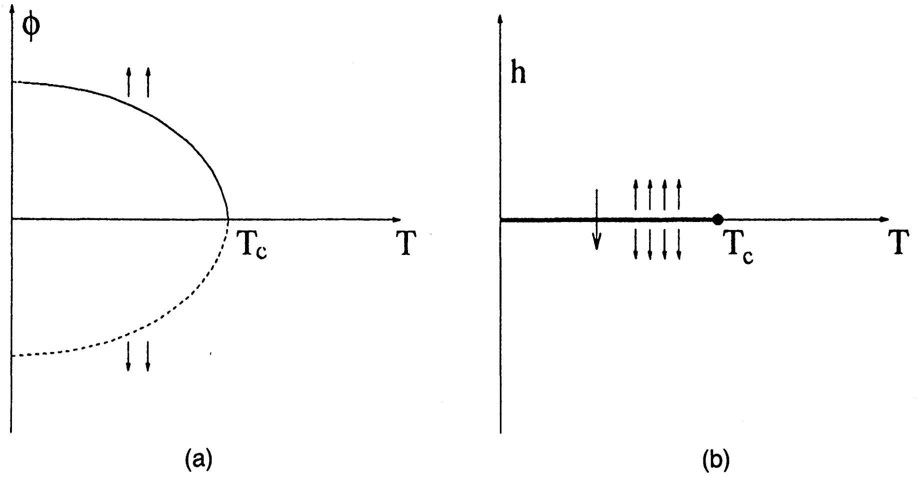

(a) (b)

FIG. 1. (a) The order parameter ϕ of a ferromagnet for zero external field $h = 0$, as a function of the temperature T. ϕ sets in continuously at T_c and has at least two equivalent orientations. (b) Phase diagram (external field h vs temperature T) of a ferromagnet.

and the specific-heat singularity in the vicinity of T_c,

$$C_{h=0} \propto |\tau|^{-\alpha}. \tag{4}$$

Equation (2) leads to a most remarkable behavior of the order-parameter *correlation function,* defined as second cumulant (again, the brackets $\langle \ldots \rangle$ denote the thermal average)

$$G(\mathbf{x} - \mathbf{x}') = \langle \phi(\mathbf{x})\phi(\mathbf{x}') \rangle - \langle \phi \rangle^2 \tag{5}$$

(here translational symmetry has been assumed); namely from the fluctuation-response theorem in d dimensions,

$$\chi = \frac{1}{k_B T} \int d^d x G(\mathbf{x}), \tag{6}$$

it becomes apparent that a diverging susceptibility implies correlations of infinite range. This feature can be described by two additional critical exponents η and ν, for the power-law behavior of the correlation function precisely at the critical point,

$$G(\mathbf{x}) \propto \frac{1}{|\mathbf{x}|^{d-2+\eta}} \quad \text{for } T = T_c, \tag{7}$$

and the temperature dependence of the *correlation length* (measuring the spatial extent of the correlations) in the vicinity of T_c,

$$\xi \propto |\tau|^{-\nu}, \tag{8}$$

respectively. The growth of correlated regions upon approaching the critical point is illustrated in Fig. 2, where the spin-up/down configurations are plotted for a two-dimensional Ising model at different temperatures.

It turns out that for temperatures in the vicinity of T_c, the so-called *asymptotic region,* Eq. (7) may be generalized to the scaling form

$$G(\mathbf{x},\tau) = \frac{1}{|\mathbf{x}|^{d-2+\eta}} \hat{G}_{\pm}(|\mathbf{x}|/\xi), \tag{9}$$

or, by introducing the Fourier transform of the correlation function, $G(\mathbf{k}) = \int d^d x G(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}}$,

$$G(\mathbf{k},\tau) = \frac{1}{|\mathbf{k}|^{2-\eta}} \hat{g}_{\pm}(|\mathbf{k}|\xi). \tag{10}$$

(The suffix "±" distinguishes between the cases $T > T_c$ and $T < T_c$.) The *scaling functions* $\hat{G}_{\pm}(y)$ and $\hat{g}_{\pm}(p)$ approach constant values in the limits $y \to 0$ and $p \to \infty$, respectively (see Sec. 1.4).

Near a critical point, where the only relevant length scale is set by the correlation

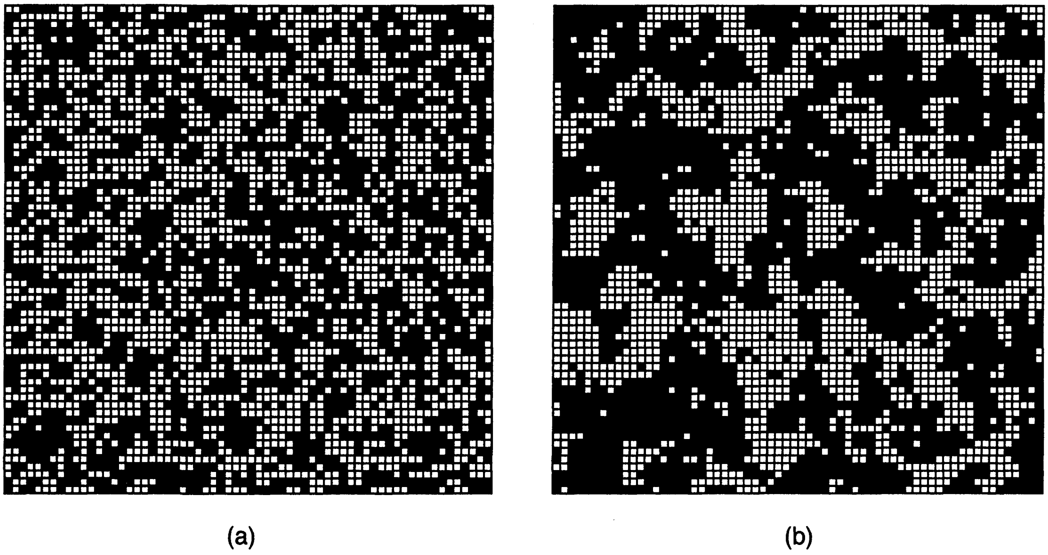
(a) (b)

FIG. 2. Correlated clusters in the two-dimensional Ising model (the black and white points correspond to the two possible states at each site i in the lattice) for two different temperatures: (a) far away from T_c in the disordered phase, and (b) near the critical point.

length ξ, the increasing size of correlated regions also causes a *critical slowing down* of the dynamics of the system. In fact, the characteristic frequency ω_c of the system is expected to obey a scaling law similar to Eq. (10), namely

$$\omega_c(\mathbf{k},\tau) = |\mathbf{k}|^z \hat{\Omega}_{\pm}(|\mathbf{k}|\xi), \tag{11}$$

with $\hat{\Omega}_{\pm}(p) \to \text{const}$ for $p \to \infty$, such that at the critical temperature

$$\omega_c(\mathbf{k},0) \propto |\mathbf{k}|^z \quad \text{for } T = T_c, \tag{12}$$

with the *dynamic critical exponent* z (Hohenberg and Halperin, 1977). For example, in the case of purely relaxational dynamics ω_c may be identified with the inverse relaxation time t_0^{-1}; thus by using Eq. (11) and $\Omega_{\pm}(p) \to p^{-z}$ for $p \to 0$, the temperature dependence of the relaxation time is found to be $t_0 \propto \xi^z \propto |\tau|^{-z\nu}$.

In addition, certain universal features of the thermodynamic phases themselves can sometimes be characterized by power laws also. Specifically, the cost in free energy Σ of creating a domain wall in a phase with broken symmetry is assumed to scale with its linear dimension L according to

$$\Sigma = \sigma L^{\theta}, \tag{13}$$

where σ denotes the interfacial tension; a positive *stiffness exponent* θ means that long-range order is stable in the phase under consideration, whereas for $\theta < 0$ the system would spontaneously break up into domains (see, e.g., Fisher, 1993).

1.2 Examples

We have already referred to the phase transition from a paramagnetic to a ferromagnetic state, occurring at the Curie temperature T_c (in fact, the language used in this article is adopted from that particular physical realization). The order parameter is the macroscopic magnetization, χ is the magnetic susceptibility, and h the external magnetic field. A nonzero order parameter selects a particular direction; thus, depending on the Hamiltonian (see Sec. 2), the broken symmetry may be discrete (inversions) or continuous (rotations). At T_c, the order parameter sets in continuously [Fig. 1(a)], justifying the terminology "continuous transition." In anticipation of the comparison with the liquid–gas transition, we note, however, that reversing the external magnetic field for $T < T_c$ leads to a discontinuous change of sign of the magnetization [Fig. 1(b)]. The full phase diagram containing ϕ, h, and T is depicted in Fig. 3(a).

Probably the most familiar phase transitions are the solid–liquid and the liquid–gas

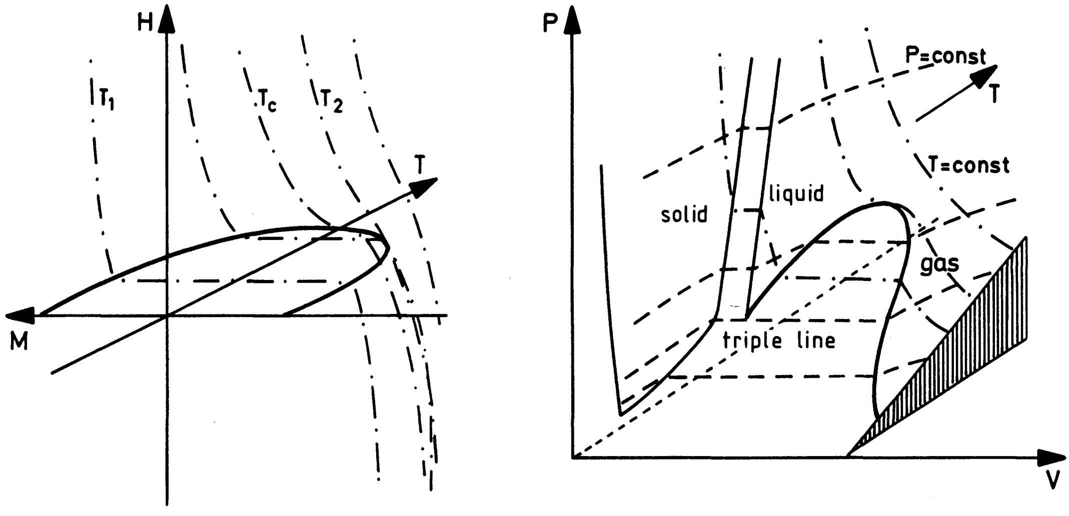

(a) (b)

FIG. 3. Comparison of the phase diagrams (a) of a ferromagnet, and (b) of a one-component gas/liquid/solid.

transitions [Fig. 3(b)]. For the latter, the density at the coexistence line minus the critical density may serve as an order parameter, which thus has a low- and high-density, i.e., a gas and liquid, branch. The associated external field is the pressure P, the (isothermal) compressibility κ would have to be identified with the order-parameter susceptibility, and the surface tension of the liquid defines the stiffness. At the critical point, where the isotherm as a function of P and V has an inflection point, one may observe the critical power laws as stated in Eqs. (1)–(4), (7), and (8) above (e.g., the divergence of κ becomes apparent in the phenomenon of critical opalescence near T_c).

In everyday life, the liquid–gas transition is usually observed at constant pressure well below P_c. As the system is heated up, the density changes discontinuously as a function of the temperature. Consequently, the liquid–gas transition is frequently referred to as being of first order, and the critical point as the end point of the transition line where the distinction between liquid and gas disappears. The analogy between the liquid–gas and the ferromagnetic transitions may be exhibited vividly if one encloses a fluid of critical density ρ_c in a so-called Natterer tube. For $T > T_c$, there is a uniform fluid phase, which splits up into a liquid and a gas phase if the system is cooled down below T_c. These phases appear separated by a meniscus.

At the liquid–solid transition, the translational symmetry is broken, and the order parameter may be chosen as the Fourier component of the density at a reciprocal-lattice vector of the solid; this transition does not terminate at a critical point, however [Fig. 3(b)]. Somewhat more subtle and diverse are structural transformations between different solid phases, where suitable components either of the displacement field or of the strain tensor appear as order parameters (see PHASE TRANSITIONS, STRUCTURAL); and furthermore the various ordered phases of liquid crystals (see LIQUID CRYSTALS, STRUCTURE OF). In the former, the order parameter may be the polarization or staggered polarization at ferroelectric or antiferroelectric phase transitions, or some quantity characterizing the orientation of molecular groups. At elastic transitions, it is a component of the strain tensor, specifying the deformation involved.

It should be noted that the phase diagrams of systems as different as ferromagnets and fluids look remarkably similar near their respective critical points (Fig. 3). Again, there is a very rich variety of magnetically ordered systems (e.g., antiferromagnets, ferrimagnets, helical phases), the order parameters of which may be of a rather complicated structure (see MAGNETIC ORDERING IN SOLIDS). A powerful tool to study the various phases of solids is neutron scattering, which directly measures the Fourier transform of the correlation function [Eq. (5)]. Moreover, time-dependent phenomena can be explored by inelastic scattering of neutrons and light, which renders the critical slowing down [Eq. (12)] of the collective excitations (see COLLECTIVE PHENOMENA IN SOLIDS) directly observable.

Finally, phase transitions to macroscopic quantum states, namely superfluidity and superconductivity, should be mentioned. The order parameter in these cases is a complex field Φ, which may be identified with a macroscopic wave function, and the broken symmetry is a gauge invariance with respect to the phase of Φ (see SUPERFLUIDITY: LIQUID HELIUM SYSTEMS; SUPERCONDUCTIVITY, LOW TEMPERATURE). Nature also allows for most interesting interplays between several kinds of phase transitions, which can lead to so-called multicritical points, and fairly complicated phase diagrams (see, e.g., SUPERCONDUCTIVITY, HIGH TEMPERATURE).

1.3 Universality

Near critical points, the topologies of the phase diagrams of such diverse systems as a liquid–gas mixture and a ferromagnetic material look amazingly similar (Fig. 3). Moreover, both in computer simulations of simplified models and in real experiments the critical exponents for the corresponding phase transitions turn out to be identical for a broad variety of physical systems, and to depend only on the number of components and symmetry properties of the order parameter, the space dimension, and possibly the character of the particle interaction, namely if it is long range (e.g., Coulomb or dipolar forces). This surprising feature is usually termed *universality*. The microscopic details of the strongly interacting many-particle system merely enter the prefactors of the ensuing power laws, and even ratios of such amplitudes are in fact universal numbers.

A clue for this most remarkable fact may be found in the divergence of the correlation length [Eq. (8)] at a continuous phase transition. Upon approaching T_c, ξ will asymptotically become the only relevant length scale in the system, dominating over all microscopic scales. Even though the phase transition itself is caused by some (usually short range) interactions of the microscopic constituents, the ensuing large fluctuations will markedly reduce all dependences on the degrees of freedom acting on short length scales, like the underlying crystal structure, the interaction range (as long as it is short), etc. Therefore, the system's thermodynamic properties near T_c should be entirely determined by global features such as its overall symmetry and large-scale properties (this is why long-range interactions, though possibly weak, may become important). In this regime, the system will behave strongly collectively, and one expects universal laws to apply for its low-wavelength and low-frequency excitations.

1.4 Scaling Laws

Another striking experimental feature near a critical point is the apparent *scale invariance* of physical properties (Fig. 4 illustrates this spectacular effect for the specific heat of helium-4 near the normal–superfluid transition). Again, this may be related to the diverging correlation length. The appropriate mathematical description for this phenomenon of scale invariance is that the free energy and the correlation functions assume the form of generalized homogeneous functions. In the case that the only relevant control parameters are the temperature and the external field, and using the definitions of the critical exponents (also called critical indices), one is thus led to the formulation of the *scaling hypotheses* for the singular part of the free energy:

$$F_s = A_\pm |\tau|^{2-\alpha} \hat{f}_\pm(h/|\tau|^{\beta\delta}), \tag{14}$$

and the correlation function, Eq. (9) [or Eq. (10)] (Fisher, 1974). In Eq. (14), the scaling function $\hat{f}_\pm$ approaches a constant value in the limit $h \to 0$.

There are various ways to make Eq. (14) plausible. First, it can be considered as a generalization of the mean-field results [see Sec. 2.3.1] to general critical exponents. Second, one may start from the observation that singularities are present only at $\tau = 0$ and $h = 0$. To what extent these become apparent depends on the separation from the critical point, τ, and on the value of $h/|\tau|^{\beta\delta}$, i.e., the ratio of the external field to the field equivalent of τ, namely $h_\tau = \phi^\delta = |\tau|^{\beta\delta}$. As long as $h \ll h_\tau$, the system is essentially in zero field, and $F_s \approx A_\pm|\tau|^{2-\alpha}\hat{f}_\pm(0) \propto |\tau|^{2-\alpha}$. On the other hand, when τ gets so small that $|\tau| \leq h^{1/\beta\delta}$, the system is manifestly in a finite field. Any further decrease of τ will then not lead to any changes: F_s remains at the value it acquired for $|\tau| = h^{1/\beta\delta}$, i.e. $A_\pm h^{(2-\alpha)/\beta\delta}\hat{f}_\pm(1)$; that is to say, in the limit $\tau \to 0$ one must have $\hat{f}_\pm(y)$

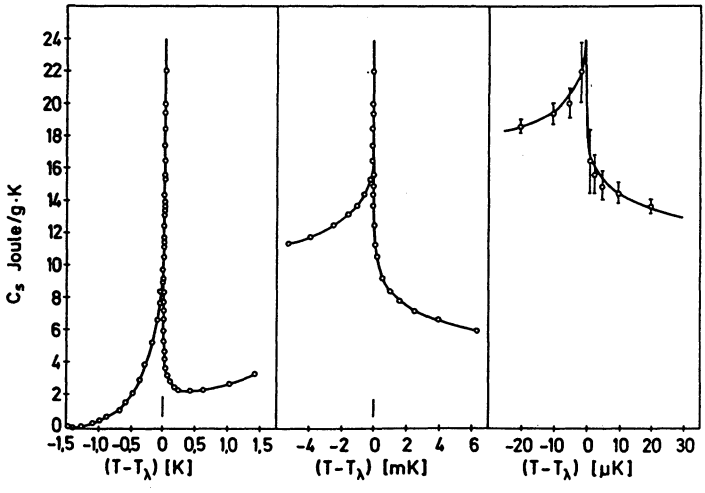

FIG. 4. Temperature dependence of the specific heat of helium-4 at the normal–superfluid phase transition, shown for successively reduced scales in the temperature variable (from Buckingham and Fairbank, 1961, p. 86).

$\rightarrow y^{(2-\alpha)/\beta\delta}$, in order to cancel the singular dependence of F_s on τ.

The scaling forms [Eqs. (14), (9), and (10)] incorporate the power laws [Eqs. (1)–(4), (7), and (8)]; moreover, by taking suitable derivatives with respect to τ and h, several *scaling relations* connecting the different critical indices may readily be obtained, namely

$$\alpha = 2 - \beta(\delta + 1), \tag{15}$$

$$\gamma = \beta(\delta - 1) = \nu(2 - \eta). \tag{16}$$

Relations of the type in Eqs. (14) and (9) are called *scaling laws*, because they are invariant under the following scale transformations:

$$\begin{aligned} &x \rightarrow x/b, \quad \xi \rightarrow \xi/b, \\ &\tau \rightarrow \tau b^{1/\nu}, \quad h \rightarrow h b^{\beta\delta/\nu}, \\ &F_s \rightarrow F_s b^{(2-\alpha)/\nu}, \quad G \rightarrow G b^{(d-2+\eta)/\nu}. \end{aligned} \tag{17}$$

If, in addition, one assumes that underlying these scale transformations there is a microscopic elimination procedure mapping the original system with lattice constant a and N sites to a new one with the same lattice constant a, but a reduced number Nb^{-d} degrees of freedom, one has

$$\frac{F_s(\tau,h)}{N} = b^{-d}\frac{F_s(\tau b^{1/\nu}, h b^{\beta\delta/\nu})}{Nb^{-d}}, \tag{18}$$

which in turn implies the *hyperscaling relation*

$$2 - \alpha = d\nu \tag{19}$$

containing the space dimension d. Equations (15), (16), and (19) leave only *two independent static critical exponents* characterizing the transition.

For dynamic critical phenomena, one arrives at a *dynamic scaling hypothesis* (Hohenberg and Halperin, 1977), originally proposed in the context of the λ transition of helium-4, by generalizing Eq. (10) to

$$G(\mathbf{k},\omega) = \frac{1}{|\mathbf{k}|^{2-\eta}}\,\hat{g}_{\pm}(|\mathbf{k}|\xi, \omega/\omega_c(\mathbf{k})) \tag{20}$$

for the frequency-dependent correlation function. Scaling laws of the form in Eqs. (14), (9), and (20) can be derived by the renormalization-group theory (see Sec. 3.2.4).

2. MATHEMATICAL REPRESENTATIONS

2.1 Microscopic Hamiltonians

In view of the expected universality of critical phenomena, simplified models containing the essential physics, namely the symmetry properties of the order parameter and the nature of the interactions, have provided a useful means for theoretical studies. For example, the *Ising* Hamiltonian describes a scalar order parameter, $\phi = \langle \sigma_i \rangle$, where the spin variables $\sigma_i = \pm 1$ are defined on a d-dimensional lattice, interacting with each other via a pair interaction J_{ij}, and subject to an external field with local strength h_i,

$$H[\sigma_i] = -\frac{1}{2}\sum_{i,j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i. \tag{21}$$

Positive exchange interaction J amounts to a ferromagnetic coupling, tending to align the spins, while an antiferromagnetic $J < 0$ favors alternating spin orientation. Its range is often restricted to nearest-neighbor interaction, but could also be taken as long-range. In the simplest case, the nonzero J_{ij} are chosen as constant, but in order to describe disordered systems they may as well be drawn from a certain random distribution (*random-bond* Ising model). Similarly, for most applications the external field is assumed to be uniform, $h_i = h$, but may in certain circumstances also be considered a random variable (*random-field* Ising model). Once the Hamiltonian is specified, in principle all thermodynamic properties of the model can of course be derived from the partition sum (see STATISTICAL MECHANICS, CLASSICAL)

$$Z = e^{-F/k_BT} = \mathrm{Tr}(e^{-H/k_BT}) \tag{22}$$

$$= \sum_{\{\sigma_i = \pm 1\}} e^{-H[\sigma_i]/k_BT}, \tag{23}$$

and the definition of thermodynamic averages

$$\langle A \rangle = \frac{1}{Z}\mathrm{Tr}(e^{-H/k_BT}A). \tag{24}$$

The short-range Ising model without disorder can be shown to display a second-order finite-temperature phase transition for all dimensions $d \geq 2$, whereupon the discrete symmetry $\sigma_i \leftrightarrow -\sigma_i$ is broken (see CRITICAL PHENOMENA). The statistical mechanics of lattice gases as well as binary liquid mixtures may be mapped onto Ising-type models; this once more illustrates the intimate analogies of phase transitions in very different systems.

Similarly, a three-component vector order parameter is incorporated in the anisotropic *Heisenberg* model,

$$H = -\frac{1}{2}\sum_{i,j} [J_{ij}^{\perp}(S_i^x S_j^x + S_i^y S_j^y) + J_{ij}^{\parallel} S_i^z S_j^z] - \sum_i h_i S_i^z \qquad (25)$$

(we use the notation S_i^α, $\alpha = x, y, z$, for the three-component spin variables here; of course, this may be readily generalized to an n-component order parameter), where the symmetry-breaking field h_i is taken along the z direction. For nearest-neighbor interactions, in the cases $J^{\parallel} > J^{\perp} > 0$ and $J^{\perp} > J^{\parallel} > 0$ the Hamiltonian [Eq. (25)] describes a uniaxial ferromagnet or planar ferromagnet, respectively, while in the special situation $J_{ij}^{\parallel} = J_{ij}^{\perp}$ Eq. (25) yields the isotropic Heisenberg model, and for $J_{ij}^{\parallel} = 0$ the xy model, the latter being appropriate for the description of superfluid helium-4, for example (see Fisher, 1967; Ma, 1976a). As a consequence of the continuous rotational symmetry broken in the ordered phase, the Heisenberg and xy models may have long-range order only for $d > 2$ dimensions. The two-dimensional xy model displays another very interesting scenario, namely the Berezinskii–Kosterlitz–Thouless transition (Nelson, 1983): There appears a phase, although without long-range order, in which the correlation function decays as a power law. The system then disorders at higher temperatures via the creation and unbinding of pairs of topological defects (vortex-antivortex pairs of spin excitations). The relatively simple models introduced above have become of paradigmatic importance in the statistical mechanics of phase transitions. One should keep in mind, however, that in more complicated physical situations more elaborate models may be required.

2.2 Ginzburg–Landau Functional

Near a continuous phase transition, predominantly long-wavelength properties will be of significance, because of the divergence of the correlation length according to Eq. (8). Therefore, a continuum model for just the long-wavelength degrees of freedom should be sufficient and appropriate, at least for the description of the universal features. We remark that this concept of an *effective (field) theory* is of course defined on a certain energy scale, which in this context is set by the transition temperature. On different scales, the corresponding effective models may differ considerably, and one may in fact imagine them to be generated by a sequence of symmetry-breaking phase transitions occurring at successively lower energies.

Such an effective model for the long-wavelength degrees of freedom may either be derived from a microscopic Hamiltonian, such as Eq. (21) or (25), by mapping the partition function on a continuum problem and performing a gradient expansion; or merely be written down by carefully exploiting symmetry arguments. Let us denote the n-component vector order parameter by $\boldsymbol{\phi}$. For the case of a second-order phase transition breaking an O(n) symmetry, the effective free energy, also called Hamiltonian in the context of critical phenomena, which retains only the lowest orders in the gradients and nonlinear terms, is the *Ginzburg–Landau free energy functional* (in d dimensions)

$$\mathcal{F}[\phi] = \int d^d x \left[\frac{r}{2}\boldsymbol{\phi}^2 + \frac{u}{4}(\boldsymbol{\phi}^2)^2 + \frac{1}{2}(\nabla\boldsymbol{\phi})^2 - \mathbf{h}\cdot\boldsymbol{\phi} \right], \qquad (26)$$

where $r = a(T - T_c)$, and a, u are positive coefficients, while $\mathbf{h}$ is the (n-component) external field. The coefficient of the gradient term in Eq. (26) was assumed to be positive, assuring that spatially uniform configurations be energetically preferred to inhomogeneous ones, and has then been absorbed in a redefinition of the field variable $\boldsymbol{\phi}$. For $n = 1$ the Hamiltonian [Eq. (26)] has the discrete symmetry of the Ising model [Eq. (21)] (i.e., for vanishing h), with a scalar order parameter, while for $n = 2$ and $n = 3$ the continuous symmetries of the xy model and the isotropic Heisenberg model are recovered.

Again, the thermodynamics can be retrieved from the partition function, which mathematically assumes the form of a functional integral of the probability distribution

$\mathcal{P} \propto \exp(-\mathcal{F}/k_BT)$ over the vector field $\boldsymbol{\phi}(x)$:

$$Z = Z_0(T)\int \mathcal{D}[\boldsymbol{\phi}(x)]e^{-\mathcal{F}[\boldsymbol{\phi}]/k_BT}. \qquad (27)$$

Here the term Z_0 contains those contributions that are not explicitly stated in the Ginzburg–Landau functional [Eq. (26)] and that should be insignificant near the transition. Figure 5 depicts the probability distribution $\mathcal{P}$ in the case of a uniform one-component order parameter ϕ, containing the first two terms of the integrand in Eq. (26), for $r > 0$, and $r < 0$, respectively (the phase transition occurs at $r = 0$). We note that a very convenient way to calculate correlation functions uses derivatives with respect to the components of $\mathbf{h}$, e.g.,

$$G_{\alpha\beta}(x - x') = \frac{(k_BT)^2\partial^2 \ln Z[\mathbf{h}]}{\partial h_\alpha(x)\partial h_\beta(x')}. \qquad (28)$$

2.3 Landau–Ginzburg Approximation

2.3.1 Mean-Field Theory Evaluating the functional integral [Eq. (27)] is a highly nontrivial task, and in general one is forced to resort to some approximation. Because of the exponential in the integrand, one might expect the probability distribution to be sharply peaked, and the thermodynamics of the system thus to be dominated by the most probable configuration, which is given by the solution to the stationarity equation

$$\begin{aligned} 0 &= \frac{\delta\mathcal{F}}{\delta\boldsymbol{\phi}(x)} \\ &= [r + u\boldsymbol{\phi}(x)^2 - \nabla^2]\boldsymbol{\phi}(x) - \mathbf{h}(x). \end{aligned} \qquad (29)$$

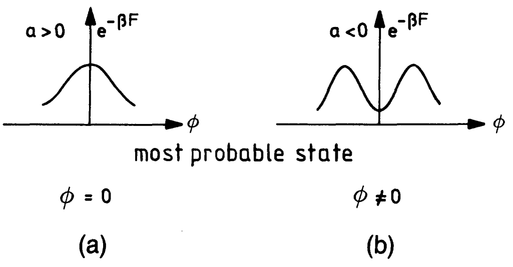

FIG. 5. Probability distribution $\mathcal{P} \propto \exp(-\mathcal{F}/k_BT)$ for the Landau part of the free energy $\mathcal{F}$, Eq. (26), as function of a single-component order parameter ϕ, for (a) $T > T_c$, and (b) $T < T_c$.

In order to determine the homogeneous order parameter [$\nabla\boldsymbol{\phi}(x) = 0$] as a function of T, one must evaluate the Landau–Ginzburg equation [Eq. (29)] for $\mathbf{h} = h\mathbf{e}_n \to 0$ (note the order of limits here: *first* take the thermodynamic limit, and only *then* let the symmetry-breaking field vanish; with the reverse procedure there would be no nonzero order parameter for symmetry reasons, because one would have to average over the degenerate ground states). This yields

$$\boldsymbol{\phi} = \begin{cases} 0, & T \geq T_c, \\ \pm\sqrt{|r|/u}\mathbf{e}_n, & T < T_c; \end{cases} \qquad (30)$$

and inserting Eq. (30) into Eq. (26) one arrives at the following approximate result for the singular part of the free energy F_s:

$$F_s[\boldsymbol{\phi}] = \begin{cases} 0, & T \geq T_c, \\ -r^2L^d/4u, & T < T_c \end{cases} \qquad (31)$$

(here L^d is the volume of a cube in d-dimensional space). By comparison with Eqs. (1) and (4), one thus immediately finds the critical exponents $\beta = \frac{1}{2}$, and $\alpha = 0$; more precisely, the specific heat displays a discontinuity at the transition temperature T_c, the height of which is

$$\Delta C_{h=0} = T_c a^2 L^d/2u. \qquad (32)$$

Similarly, for finite external field Eq. (29) leads to the equation of state in terms of the order-parameter component ϕ_n parallel to the symmetry-breaking field $\mathbf{h} = h\mathbf{e}_n$,

$$h = (r + u\phi_n^2)\phi_n, \qquad (33)$$

from which the exponent for the critical isotherm is readily obtained to be $\delta = 3$. Finally, differentiating Eq. (33) with respect to h yields the isothermal susceptibility

$$\chi = \begin{cases} 1/r = C_+(T - T_c)^{-1}, & T > T_c, \\ 1/2|r| = C_-(T_c - T)^{-1}, & T < T_c, \end{cases} \qquad (34)$$

and hence the critical exponent $\gamma = 1$ (on both sides of the transition), as well as the also universal amplitude ratio $C_+/C_- = 2$. The above results for the critical indices coincide with those of the mean-field approximation for the Ising and Heisenberg models, and with the exponents found for the van der Waals gas near its critical point. (In the mean-field approximation for the Ising model [Eq. (21)], for instance, the exchange field $\Sigma_j J_{ij}\sigma_j$ acting on spin σ_i is replaced by its average $\Sigma_j J_{ij}\langle\sigma_j\rangle$.)

We note that if u in Eq. (26) became negative, which may happen as some external parameter, e.g., the pressure, is varied, additional terms would have to be included in the Ginzburg–Landau functional for stability reasons. In the case of an underlying O(n) symmetry, the leading such nonlinearity would be $\propto (\phi^2)^3$, and the ensuing phase transition of first order. Its transition temperature $T_1 \neq T_c$ is to be defined by the condition that the free energies of the distinct coexisting phases be equal. In the special case that u vanishes, T_c becomes a tricritical point, displaying most interesting and complex scaling behavior.

2.3.2 Gaussian Fluctuations In keeping the most probable state only, fluctuations have been neglected entirely so far. However, the divergence of the susceptibility, Eq. (34), implies that fluctuations must indeed become very strong near the critical point. The next step is therefore to check whether the above results are altered by including fluctuations to lowest order. Thus, we introduce the deviations $\psi(x)$ from the most probable state $\phi(x)$, given by the solution to Eq. (29), by writing the order-parameter field as $\phi(x) + \psi(x)$, and expand the Ginzburg–Landau functional [Eq. (26)] with respect to $\psi(x)$, keeping terms up to second order only. This is most conveniently done in Fourier space, where we define the Fourier transform of $\psi(x)$ as

$$\psi(x) = L^{-d/2} \sum_{k\in \mathrm{B}} \psi_k e^{ikx}; \tag{35}$$

here B denotes the Brillouin zone for the allowed values of the k vectors. The partition function [Eq. (27)] then becomes

$$Z = Z_0(T) \int \prod_k d\psi_k e^{-\mathscr{F}_G[\psi]/k_\mathrm{B}T}. \tag{36}$$

Beginning with the situation above T_c, for $h = 0$, the Fourier expansion renders the Ginzburg–Landau functional diagonal, $\mathscr{F}_\mathrm{G} = \Sigma_k(r + k^2)\psi_k \cdot \psi_{-k}/2$. The functional integrals in the partition function thus reduce to simple Gaussian integrals, yielding for the free energy in *Gaussian approximation*

$$F_s(T,h = 0) = -k_\mathrm{B}T \frac{n}{2} \sum_k \ln \frac{2\pi k_\mathrm{B}T}{r + k^2}, \tag{37}$$

and by transforming the discrete sum over the Brillouin zone into an integral, the most singular part of the specific heat follows as

$$C_s = k_\mathrm{B}L^d \frac{n}{2} (Ta)^2 \int \frac{d^dk}{(2\pi)^d} \frac{1}{(r + k^2)^2}. \tag{38}$$

The result of evaluating the integral is

$$C_s = \begin{cases} A_+(T - T_c)^{-(4-d)/2}, & d < 4, \\ \propto \ln(T - T_c), & d = 4, \\ B - A(T - T_c)^{(d-4)/2}, & d > 4; \end{cases} \tag{39}$$

hence while for $d > 4$ the specific heat shows a (nonanalytical) cusp at T_c, it diverges for $d \leq 4$, with the exponent $\alpha = (4 - d)/2$; here an exponent "0" is to be understood as a logarithmic singularity. Neglecting deviations from the most probable state therefore turns out to be entirely unjustifiable in the latter case, where the thermodynamics is obviously dominated by the order-parameter fluctuations.

Similarly, in the ordered phase one finds

$$C_s = TL^d a^2/2u + A_-(T_c - T)^{-(4-d)/2} \tag{40}$$

for $d < 4$; the critical exponent α is therefore the same above and below the transition. The universal (i.e., depending only on d and n) amplitude ratio for the specific heat is $A_+/A_- = n/2^{d/2}$, where the factor n stems from the fact that the $n - 1$ transverse fluctuations (with respect to the direction of the spontaneous order parameter) do not contribute to the specific heat in the ordered phase.

With the help of the result in Eq. (40), the *Ginzburg–Levanyuk criterion* (see Als-Nielsen and Birgeneau, 1977) may be formulated, stating that fluctuations are negligible as long as the first term ΔC dominates over the second contribution; introducing $\xi_0 = (aT_c)^{-1/2}$ this means that the Landau–Ginzburg approximation gives essentially correct answers, provided

$$|\tau|^{(4-d)/2} \gg \frac{\mathcal{N}}{(\xi_0 L)^d} \Delta C, \tag{41}$$

where $\mathcal{N}$ is a numerical factor. Therefore, for $d > d_c = 4$ fluctuations are not very important [$\alpha_{fl} < 0$, Eq. (39)], and mean-field types of approach yield reliable results. Below this *upper critical dimension*, however, such approximations will break down in the vicinity of T_c. Depending on the value of ξ_0, the nonclassical scaling behavior (with exponents different from the mean-field indices) will be important only very near to the critical point, or already noticeable even far away from it.

For example, for magnets typically $\xi_0 \approx 1$ Å, and fluctuations become important for $|\tau| < \tau_{\mathrm{GL}} \approx 0.01$. For the normal–superfluid transition of helium-4, $\xi_0 \approx 4$ Å, and $\tau_{\mathrm{GL}} \approx 0.3$, which renders this system a prominent candidate for the study of critical phenomena (also, by varying the pressure one can explore a line of critical points and thus examine universality issues in detail). On the other hand, for conventional (low-T_c) superconductors ξ_0 is of the order 1000 Å, and $\tau_{\mathrm{GL}} \approx 10^{-10}$ to 10^{-14}, which means that the Ginzburg–Landau theory is an excellent approximation there. This is in remarkable contrast with the high-T_c materials, where $\xi_0 \approx 1$ Å, and fluctuations play a dominant role; in fact, they account for an enormously rich phase diagram of the cuprate-based superconductors (see Blatter *et al.*, 1994).

For $|\tau| < \tau_{\mathrm{GL}}$, it does not at all suffice to add to the results of the Ginzburg–Landau approximation the Gaussian corrections, since in this temperature range higher-order nonlinear contributions will be even more infrared divergent and eventually exceed those from the harmonic term. This constitutes of course a serious breakdown of the standard perturbational procedure, and signals that an entirely new theoretical approach is required in the case $|\tau| < \tau_{\mathrm{GL}}$.

For finite external field, the partition function [Eq. (36)] becomes above T_c

$$Z[\mathbf{h}] = \exp\left[\frac{1}{2k_{\mathrm{B}}T}\sum_k \frac{\mathbf{h}_k \cdot \mathbf{h}_{-k}}{r + k^2}\right], \tag{42}$$

and the correlation function can be obtained by differentiating this generating functional [see Eq. (28)]. Above T_c, the result in k space is

$$\langle \psi_k^\alpha \psi_{-k'}^\beta \rangle = \frac{k_{\mathrm{B}}T}{r + k^2}\,\delta_{kk'}\delta^{\alpha\beta}; \tag{43}$$

consequently, for $r = 0$ one finds in real space

$$G(x) \propto 1/|x|^{d-2}, \tag{44}$$

and thus the critical exponent defined in Eq. (7) is $\eta = 0$ in the Landau–Ginzburg approximation. Also, the inverse of r defines the square of a characteristic length, which we identify with the correlation length, $\xi_+ = r^{-1/2}$, and hence $\nu = \frac{1}{2}$. In three dimensions, the Fourier transform of Eq. (43) is the *Ornstein–Zernicke correlation function*

$$G(x) = \frac{k_{\mathrm{B}}T}{4\pi|x|}\,e^{-|x|/\xi_+}. \tag{45}$$

In the ordered phase, one has to distinguish between the longitudinal and transverse correlation functions, with respect to the spontaneous order parameter, which are still diagonal in the indices α, β, and for the O(n) symmetric model [Eq. (26)] in Gaussian approximation read

$$G_\parallel(k) = k_{\mathrm{B}}T/(\xi_-^{-2} + k^2) \tag{46}$$

and

$$G_\perp(k) = k_{\mathrm{B}}T/k^2, \tag{47}$$

respectively. Hence, on this level $G_\parallel(k)$ is of the form in Eq. (43), with a different correlation length $\xi_- = (2|r|)^{-1/2}$, however. Note that the amplitude ratio for the correlation lengths above and below the transition is universal again, with the mean-field value $\sqrt{2}$. Remarkably, the transverse correlation function diverges for $k \to 0$, which means that the range for the transverse correlations is infinite in the entire low-temperature phase. This can be shown to be a rigorous consequence of the breaking of the continuous rotational symmetry. For $n \geq 2$ and $d \leq 4$, the large transverse fluctuations [Eq. (47)] via nonlinear effects imply a longitudinal correlation function surpassing the result obtained within the harmonic approximation [Eq. (46)], which for $k \to 0$ has to be replaced by

$$G_\parallel(k) \propto 1/k^{4-d}. \tag{48}$$

Remarkably, Eq. (48) is a universal power law characterizing not a phase transition singularity, but an *entire ordered phase.* Yet the longitudinal correlation length ξ_- still retains some physical meaning; namely, it provides the scale on which the crossover from the critical behavior [Eq. (7)] to the asymptotic form [Eq. (48)] occurs.

In the Ising case ($n = 1$), Eq. (46) is qualitatively correct outside the critical region, and ξ_- sets the characteristic length scale in the ordered phase. For instance, a domain-wall configuration $\phi_K(x)$ connecting the two degenerate ground states $\phi_\pm = \pm\sqrt{|r|/u}$, and centered at x_0, extends over the range ξ_-,

$$\phi_K(x) = \phi_\pm \tanh\frac{x - x_0}{2\xi_-}, \tag{49}$$

as can be seen by solving the Landau–Ginzburg differential equation [Eq. (29)] for $h = 0$, allowing for a spatially varying order parameter with the appropriate boundary conditions.

3. RENORMALIZATION GROUP

3.1 Introductory Remarks

The term "renormalization" of a theory stands for a certain reparametrization, with the goal that the resulting renormalized theory be more tractable than the original one. Historically, renormalization was invented by Stückelberg and Feynman, in order to cure quantum field theories (e.g., quantum electrodynamics) from divergences (see UNIFIED FIELD THEORIES). Instead of the bare parameters (masses and coupling constants), the Lagrangian was expressed in terms of physical masses and coupling coefficients, such that ultraviolet-divergent (UV) virtual processes were systematically absorbed in the relationship between the bare and physical quantities, thus rendering the renormalized theory finite. The renormalization procedure is not unique; the renormalized quantities may for instance depend on a cutoff length scale up to which certain virtual processes are taken into account. The *renormalization-group* (RG) theory studies the dependence on this specific length scale, also called *flow parameter*. The name renormalization group derives from the fact that two successive RG transformations lead to a third such transformation.

In the context of critical phenomena, where one needs to explain the long-distance (or, in Fourier space, small–wave-number) regime, it is natural to implement the renormalization scheme by performing some kind of elimination of the short-wavelength fluctuations. Such a partial evaluation of the partition function will be much easier to achieve than a calculation of the complete partition sum, and also permits the systematic implementation of approximate methods. As a result of the elimination step, the remaining degrees of freedom will experience modified *effective* interactions.

Quite generally, various benefits can be expected from such a RG procedure:

1. The new coupling coefficients might be smaller. By repeating the renormalization process, one may thus arrive at an effectively free theory with no interactions.
2. The successively iterated coupling coefficients, also called the *parameter flow*, may have a finite *fixed point*, at which the system does not change any more under subsequent RG transformations. Since the elimination of some degrees of freedom is accompanied by a change of the underlying lattice spacing, or length scale, one may anticipate that the fixed points are under certain circumstances related to critical points. Furthermore, one may hope that the properties of the flow near these fixed points will yield information about the universal physical quantities in the vicinity of the critical point.

Below, we shall find scenario **1** in the case of the one-dimensional Ising model and **2** for the Ising model in two dimensions, as well as for the Ginzburg–Landau–Wilson effective field theory (ϕ^4 model).

3.2 Real-Space Renormalization Group

The idea of any RG approach in statistical mechanics is to exploit the scale invariance near a critical point in parameter space. In real-space RG transformations, which refers to procedures working in direct space (as opposed to Fourier space), one eliminates specific degrees of freedom that are defined on a lattice, thus performing a partial partition sum. Then the lattice constant of the ensuing system is readjusted, and also the internal variables are renormalized, in order to achieve a new Hamiltonian whose form resembles the original one. By comparison, effective couplings are defined, which thus become scale dependent. The fundamental paradigm is that certain fixed points in parameter space exist, at which these running couplings remain invariant. At these fixed points, the theory is manifestly scale invariant, and the critical exponents can be obtained by studying the linearized flow equations. As this concept of the RG may seem rather abstract at first sight, the method is explained by applying a simplified version to the ferromagnetic Ising model [Eq. (21)] in $d = 1$ and $d = 2$ dimensions (Niemeijer and van Leeuwen, 1976). Then a more general discussion, including an explanation of the emergence of scaling laws, is given. A continuous field-theoretical formulation for the RG in momentum space is presented in Sec. 3.3.

3.2.1 1D Ising Model For a one-dimensional Ising chain with constant ferromagnetic nearest-neighbor exchange coupling $J_{ij} = J > 0$ in zero external field, the Hamiltonian [Eq. (21)] becomes $H = -J \sum_i \sigma_i\sigma_{i+1}$, and introducing the abbreviation $K = J/k_BT$ the partition function [Eq. (23)] reads (for N spins, and periodic boundary conditions $\sigma_{N+1} = \sigma_1$)

$$Z_N = \sum_{\{\sigma_i=\pm 1\}} \exp\left[K \sum_i \sigma_i\sigma_{i+1}\right]. \tag{50}$$

A partial evaluation of Eq. (50) may be performed by summing over every other spin in the chain, marked by an "x" in Fig. 6. The sum over the possible orientations of spin i, for instance, yields

$$\sum_{\sigma_i=\pm 1} e^{K\sigma_i(\sigma_{i-1}+\sigma_{i+1})} = \exp[2g' + K'\sigma_{i-1}\sigma_{i+1}], \tag{51}$$

for the result can only depend on whether the adjacent spins σ_{i-1} and σ_{i+1} are oriented parallel or antiparallel to each other. By comparing the results for these two configurations, $\sigma_{i-1}\sigma_{i+1} = \pm 1$, one readily obtains the two recursion relations

$$K' = \tfrac{1}{2} \ln \cosh 2K, \tag{52}$$

$$g' = \tfrac{1}{2}(\ln 2 + K'). \tag{53}$$

Thus, via decimation of half of the spins in the chain one arrives at another Ising Hamiltonian, whose nearest-neighbor coupling is K'; in addition, a spin-independent contribution g' to the free energy is generated. If this transformation is repeated, the *exact* recursion relations [Eqs. (52) and (53)] quite generally describe how the effective couplings $K^{(k)}$ and $g^{(k)}$ in the kth step result from the corresponding values in the $(k-1)$th elimination step. In Fig. 7, the ensuing flow for the coupling K, starting with some initial value K_0, is depicted for several steps of the procedure. Clearly, the fixed-point equation $K^* = (\frac{1}{2}) \ln \cosh 2K^*$ has two solutions: $K^* = 0$, which (for the original fixed J) corresponds to infinite temperature, or vanishing effective interaction; and $K^* = \infty$, the $T = 0$ fixed point. On starting at some initial value $K_0 > 0$, the parameter flow always drives the system into the $K^* = 0$ fixed point [by expanding Eq. (52) around $K^* = 0$, this fixed point is seen to be stable]. Physically, this means that there can be no ordered state at any finite temperature $T > 0$, because asymptotically the ratio of exchange coupling and tem-

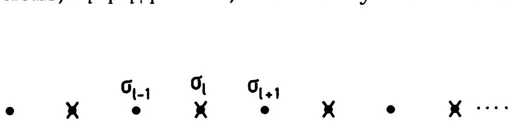

FIG. 6. Decimation transformation for the one-dimensional Ising chain with nearest-neighbor interaction.

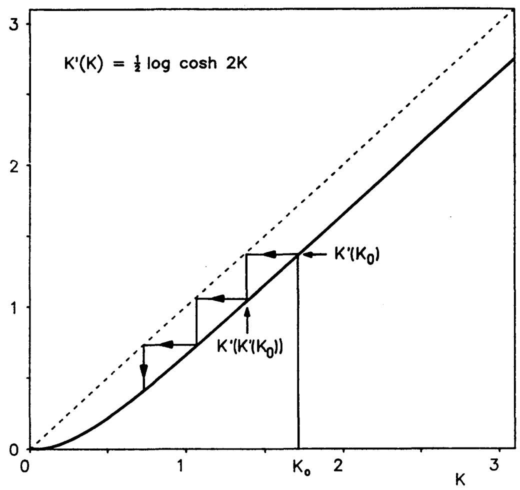

FIG. 7. Recursion relation for the nearest-neighbor coupling of the Ising chain (full; the dashed line marks the line $K' = K$). The iteration steps are indicated by the arrows.

perature vanishes, and the system becomes completely disordered.

At $T = 0$, however, the situation is rather special. Expanding Eq. (52) near $1/K^* = 0$ yields $1/K' = 1/K + \ln b/(2K^2)$, where $b = 2$ is just the factor rescaling all lengths. It now turns out to be useful to introduce $\delta = \ln b$, such that after l steps $b^l = e^{\delta l}$, and then take the limit $\delta \to 0$; thereby the discrete RG steps are mapped onto a continuous flow with the flow parameter l. The resulting differential flow equation for the now scale-dependent coupling $K(l)$ reads to lowest order

$$\frac{dK^{-1}(l)}{dl} = \tfrac{1}{2}[K^{-1}(l)]^2, \tag{54}$$

which is readily integrated to $K(1) - K(l) = l/2$. From this, the temperature dependence of the correlation length ξ can be determined by applying the ("matching") condition that $K(b = \xi)$ be a constant of order 1. This gives

$$\xi \propto \exp(2J/k_BT), \tag{55}$$

i.e., the correlation length diverges exponentially at the *marginal* critical point $T = 0$, a result that may also be readily obtained by more elementary methods.

Using the above recursion relations the partition function may be evaluated:

$$\begin{aligned} Z_N &= e^{Ng(K')}Z_{N/2}(K') = \cdots \\ &= \exp\left[N\sum_{k=1}^{n}\frac{g(K^{(k)})}{2^{k-1}} + \ln Z_{N/2^n}(K^{(n)})\right]. \end{aligned} \tag{56}$$

The factor 2 in Eq. (56) can be interpreted as stemming from a readjustment of the overall scale, as the number of degrees of freedom is being diminished by the elimination procedure. The reduced free energy (in units of k_BT) thus becomes $f = -(1/N)\ln Z_N(K)$. As is apparent from Fig. 7, $K^{(k)}$ rapidly becomes very small, and perturbative methods may be employed. Assuming $K^{(n)} \approx 0$, one arrives at the approximate result

$$f^{(n)}(K) = -\sum_{k=1}^{n}\frac{g(K^{(k)})}{2^{k-1}} - \frac{1}{2^n}\ln 2, \tag{57}$$

for in a noninteracting spin-1/2 system the free energy per spin is $-\ln 2$. Figure 8 shows the rapid approach toward the exact result, $f(K) = -\ln(2\cosh K)$, as the number of elimination steps is increased.

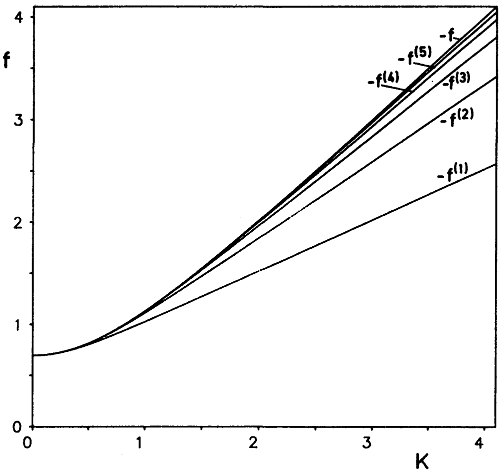

FIG. 8. Reduced free energy f for the nearest-neighbor Ising chain. The $f^{(k)}$ denote the approximation to the exact result f as obtained in the kth step of the RG procedure.

3.2.2 2D Ising Model More interesting is the case of the two-dimensional Ising model, which is known to have a phase transition at a finite temperature $T_c > 0$ (see Critical Phenomena). Here a quadratic lattice with nearest-neighbor ferromagnetic interactions will be considered (Fig. 9). In order to find the recursion relations in this case, one may perform the partial partition sum over those spins

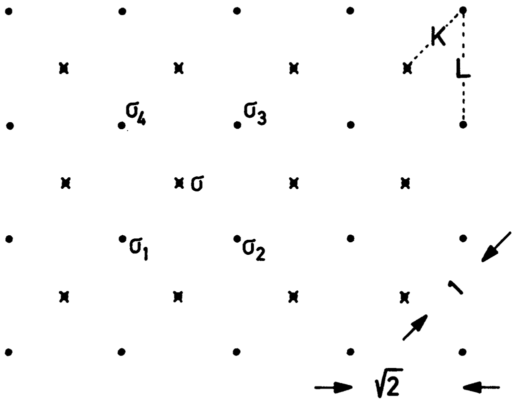

FIG. 9. Ising model on a quadratic lattice. Even if originally there were only nearest-neighbor interactions K present, by application of the decimation procedure next-nearest-neighbor couplings L are generated.

marked with an "x" in Fig. 9, which leads to a new square lattice with a lattice constant larger by a factor of $\sqrt{2}$ than that of the original one. A typical sum is

$$\sum_{\sigma=\pm 1} e^{K\sigma(\sigma_1+\sigma_2+\sigma_3+\sigma_4)}$$
$$= \exp[A' + K'(\sigma_1\sigma_2 + \sigma_2\sigma_3 + \sigma_3\sigma_4 + \sigma_4\sigma_1)] \times \exp[L'(\sigma_1\sigma_3 + \sigma_2\sigma_4) + M'\sigma_1\sigma_2\sigma_3\sigma_4]; \tag{58}$$

thus some new interaction, namely a next-nearest-neighbor coupling L' and a four-spin interaction M', are *generated* by the transformation. In the vicinity of the actual critical value $K_c = J/\kappa_B T_c = 0.4406$, one finds $M' \ll L' \leq K'$. Therefore, M' shall henceforth be neglected here (a more thorough perturbational analysis is given in Niemeijer and van Leeuwen, 1976.) The terms A', K', and L' can readily be expressed by K. If these relations are expanded to second order in K, and if one realizes that a next-nearest-neighbor interaction L in the original Hamiltonian appears as a contribution to the nearest-neighbor interaction in the primed Hamiltonian, one obtains

$$K' = 2K^2 + L, \tag{59}$$

$$L' = K^2. \tag{60}$$

A simple physical picture behind these approximate recursion relations is that they just represent the numbers of paths between the remaining spins, weighted by the product of the bonds involved.

It should be clear, however, that with each step of the elimination new couplings appear, as opposed to the one-dimensional case. Of course, restricting the flow to the subspace (K,L) will not produce quantitatively satisfactory results, but still the structure and typical features of such recursion relations are preserved.

The fixed points of Eqs. (59) and (60) are

1. $K^* = L^* = 0$,
2. $K^* = L^* = \infty$, and
3. $K_c^* = \frac{1}{3}$, $L_c^* = \frac{1}{9}$ (Fig. 10).

As in the 1D Ising model, fixed points **1** and **2** refer to infinite temperature (disordered phase) and $T = 0$, respectively. Yet now the latter describes an entire (ordered) phase. Furthermore, there is an additional nontrivial finite fixed point, which governs the second-order transition separating these two phases. For at each step in the procedure, the effective lengths are reduced by a factor $b = \sqrt{2}$, which applies to the correlation length also, and therefore at a fixed point either $\xi = 0$ (which is true for the $T = 0$ and $T = \infty$ fixed points), or $\xi = \infty$ (for finite values of K^* and L^*), which implies scale invariance. Therefore, the fixed point (K_c^*, L_c^*) with diverging correlation length describes the critical state. Even more, actually all points lying on a trajectory leading into the fixed point represent critical behavior, because on these paths the correlation length is infinite as well.

In order to determine the critical behavior of the model, the recursion relations [Eqs. (59) and (60)] are linearized near the point (K_c^*, L_c^*) in parameter space; with $\delta K_l = K_l - K_c^*$ and $\delta L_l = L_l - L_c^*$ in the lth step of the

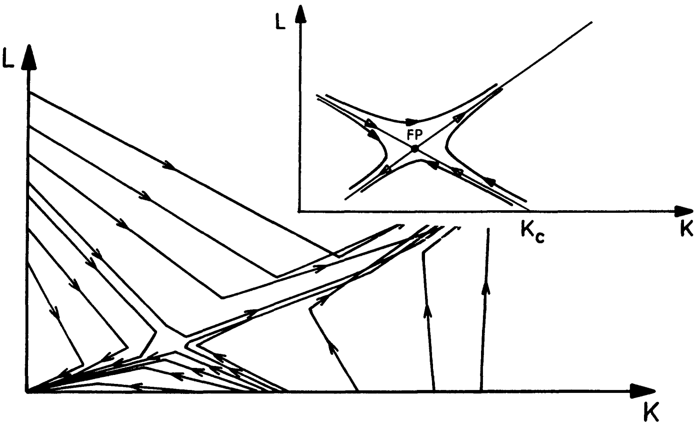

FIG. 10. Flow diagram for the two-dimensional Ising model on a square lattice near its critical point. (Only the results of every other recursion step are shown.) The inset shows the behavior in the vicinity of the critical fixed point, which is described by Eq. (63).

procedure, one finds

$$\begin{pmatrix} \delta K_l \\ \delta L_l \end{pmatrix} = \begin{pmatrix} \frac{4}{3} & 1 \\ \frac{2}{3} & 0 \end{pmatrix} \begin{pmatrix} \delta K_{l-1} \\ \delta L_{l-1} \end{pmatrix}. \tag{61}$$

The eigenvalues of the transformation matrix are $\lambda_{1,2} = (2 \pm \sqrt{10})/3$, and the corresponding eigenvectors $\mathbf{e}_{1,2} = (3, \pm\sqrt{10} - 2)$. Thus, if the original couplings are expanded in the basis of eigenvectors,

$$\begin{pmatrix} K_0 \\ L_0 \end{pmatrix} = \begin{pmatrix} K_c^* \\ L_c^* \end{pmatrix} + c_1\mathbf{e}_1 + c_2\mathbf{e}_2, \tag{62}$$

after l steps one arrives at

$$\begin{pmatrix} K_l \\ L_l \end{pmatrix} = \begin{pmatrix} K_c^* \\ L_c^* \end{pmatrix} + \lambda_1^l c_1\mathbf{e}_1 + \lambda_2^l c_2\mathbf{e}_2. \tag{63}$$

As is depicted in the flow diagram (Fig. 10), if $c_l \neq 0$, the flow will in general run away from the critical fixed point (K_c^*, L_c^*), which is thus unstable toward the *relevant perturbation* $c_1\mathbf{e}_1$, which is characterized by $|\lambda_1| > 1$. On the other hand, if $c_1 = 0$, the flow will always tend toward this fixed point, because $|\lambda_2| < 1$ ($\mathbf{e}_2$ is called the *irrelevant* direction; had we included more coupling parameters initially, additional irrelevant directions would have emerged). The original nearest-neighbor model had $L_0 = 0$. Inserting this in Eq. (62) for $c_1 = 0$, one finds the critical value of K (within the linear approximation), for which the decimation transformation will lead into the critical fixed point, to be

$$K_c = \tfrac{1}{3}\left(1 + \frac{1}{\sqrt{10} + 2}\right) \approx 0.3979, \tag{64}$$

which should be compared to the mean-field and exact values, $K_c = 0.25$ and $K_c = 0.4406$, respectively. Note that $K - K_c = -J(T - T_c)/k_B T T_c$, and hence $c_1 \neq 0$ whenever $T \neq T_c$. In this case, the flow will tend toward either the zero-temperature (for $T < T_c$) or the disordered fixed point (for $T > T_c$).

After a sufficient number of steps λ_2^l will be minute, and

$$(K - K_c)_l = b^{y_1 l}(K - K_c), \tag{65}$$

where $b = \sqrt{2}$, and $y_1 = 2 \ln\lambda_1/\ln 2 \approx 1.566$. As also $\xi_l = b^{-l}\xi$, and the definition of the critical exponent ν for the correlation length translates to $\xi \propto |K - K_c|^{-\nu}$, one finds

$$\nu = 1/y_1 \approx 0.638, \tag{66}$$

which is already larger than the mean-field result $\nu = \frac{1}{2}$, but still far off the exact value $\nu = 1$.

Similarly, an external field h may be included in the scheme. The approximate recursion relation for h turns out to be

$$h' = (1 + K)h, \tag{67}$$

with the fixed point $h^* = 0$, and the eigenvalue for the linearized recursion relation is $\lambda_h = 1 + K_c^* = \frac{4}{3}$. Hence the external field constitutes a second relevant variable besides (essentially) $T - T_c$. Notice that the flows for K and h decouple. The two relevant fields τ and h precisely correspond to the two independent critical exponents of the model.

In addition, the universal properties of the phases, as the stiffness exponent θ in Eq. (13), can be determined in a similar way. For the stiffness, the most convenient way is to study the change in free energy of a finite system as the boundary conditions are changed, e.g., from periodic to antiperiodic (see Fisher, 1993). For an Ising system of linear dimension $L = b$ in d dimensions, the domain wall energy at $T = 0$ turns out to be

$$\Sigma \approx 2Jb^{d-1}, \tag{68}$$

and hence $\theta = d - 1$. Thus, a rescaling of the length by b is equivalent to a recursion relation for the temperature

$$T' = b^{-\theta}T \tag{69}$$

for the temperature, which shows that the stiffness exponent may be identified with the negative eigenvalue $\lambda_T = -\theta$ at the zero-temperature fixed point ($T^* = 0$). Equation (68) demonstrates that the ordered phase is stable for $d \geq 2$ in the Ising model. For Heisenberg models with continuous rotational symmetry, on the other hand, one finds $\theta = d - 2$, and no long-range order is possible in two dimensions. Note that mean-field theory fails to predict such a *lower critical dimension*, at and below which fluctuations destroy the formation of spatially homogeneous long-range order.

3.2.3 General Real-Space RG Transformations A general real-space RG transformation maps a certain spin system $\{\sigma\}$, defined on a lattice, and with Hamiltonian $H\{\sigma\}$, onto a new spin system $\{\sigma'\}$ with new Hamiltonian $H'\{\sigma'\}$ and by a factor of $N'/N = b^{-d}$ fewer degrees of freedom. It may be represented by a transformation $\tau\{\sigma',\sigma\}$, such that ($k_B T$ has been set equal to unity here)

$$e^{-F-H'\{\sigma'\}} = \sum_{\{\sigma\}} \tau\{\sigma',\sigma\} e^{-H\{\sigma\}}, \tag{70}$$

with the conditions

$$\sum_{\{\sigma'\}} H'\{\sigma'\} = 0 \tag{71}$$

and

$$\sum_{\{\sigma'\}} \tau\{\sigma',\sigma\} = 1, \tag{72}$$

which guarantee that $e^{-F}\,\mathrm{Tr}_{\{\sigma'\}} e^{-H'\{\sigma'\}} = \mathrm{Tr}_{\{\sigma\}} e^{-H\{\sigma\}}$.

Important examples for such transformations are decimation transformations, and linear as well as nonlinear block-spin transformations. Several types of approximations, like the cumulant expansion and cluster and finite-lattice approximations, have been applied to various kinds of lattices (quadratic, triangular, etc.), where a large number of interaction coefficients have been included. In two dimensions, critical exponents and thermodynamic properties as a whole have thereby been obtained to an impressive degree of accuracy (Niemeijer and van Leeuwen, 1976). A combination of numerical Monte-Carlo methods and real-space RG transformations leads to the *Monte-Carlo renormalization group* (Swendsen, 1984). Apart from that, however, the more versatile momentum-space methods (see Sec. 3.3 below) have been far more useful in higher dimensions ($d > 2$).

3.2.4 Emergence of Scaling Laws While the decimation procedure employing only a few parameters does not provide quantitatively satisfactory results, and also is inappropriate for the study of correlation functions, it still elucidates the general structure of RG transformations R, whereupon the original Hamiltonian H is mapped onto a new one, $H' = R(H)$ (this includes the rescaling of lengths in the problem, or for the degrees of freedom: $N' = Nb^{-d}$ in d dimensions). Near the fixed-point Hamiltonian $H^* = R(H^*)$, one may expand $R(H^* + \delta H) \approx H^* + L\delta H$, which yields the linearized recursion relation $\delta H' = L\delta H$. The eigenoperators δH_1, $\delta H_2 \ldots$ for this linear transformation are then determined by the eigenvalue equations

$$L\delta H_i = \lambda_i \delta H_i. \tag{73}$$

A given Hamiltonian H that does not differ too much from the fixed-point Hamiltonian may be expanded according to

$$H = H^* + \tau\delta H_\tau + h\delta H_h + \sum_{j\geq 3} c_j \delta H_j, \tag{74}$$

where δH_τ and δH_h denote the two relevant perturbations, $|\lambda_\tau| = b^{y_\tau} > 1$, $|\lambda_h| = b^{y_h} > 1$, connected with the temperature variable $T - T_c$ and the external field h, while $|\lambda_j| < 1$ (and hence $y_j < 0$) for $j \geq 3$ (in the Ising case, $\delta H_h = \Sigma_i \sigma_i$). The coefficients τ, h, and c_j are called *scaling fields*.

After l steps, the free energy per spin, $f(\{c_j\}) = F_N(\{c_j\})/N$, transforms according to

$$f(\tau,h,c_3,\ldots) = b^{-dl} f(\tau b^{y_\tau l}, h b^{y_h l}, c_3 b^{y_3 l},\ldots) \tag{75}$$

in the linear approximation. (Here we have omitted an additive term, which does not affect the following derivation of the scaling law, but would of course be important for the evaluation of the total free energy.)

The scale parameter l may now be chosen conveniently, for example $|\tau| b^{y_\tau l} = 1$, which sets the first argument of f equal to ± 1; this yields

$$f(\tau,h,\ldots) = |\tau|^{d/y_\tau} \hat{f}_\pm (h|\tau|^{-y_h/y_\tau}, c_3|\tau|^{|y_3|/y_\tau},\ldots), \tag{76}$$

where $\hat{f}_\pm(x,y,\ldots) = f(\pm 1,x,y,\ldots)$. Close to T_c, the dependence on the irrelevant fields $c_3, \ldots$ can be neglected, and Eq. (76) then precisely assumes the scaling form [Eq. (14)], with the exponents

$$\beta\delta = y_h/y_\tau \tag{77}$$

and

$$2 - \alpha = d/y_\tau. \tag{78}$$

Thus, the scaling law [Eq. (14)] has been derived within the RG theory for fixed points with precisely one relevant field, aside from the external magnetic field and the irrelevant operators. Furthermore, the dependence of the irrelevant fields $c_3, \ldots$ yields *corrections to scaling*, which may have to be taken into account in situations not yet close enough to T_c to be in the fully asymptotic regime.

In order to relate y_τ to the exponent ν, one has to recall that l iterations reduce the correlation length to $\xi' = b^{-l}\xi$, which implies $(\tau b^{y_\tau l})^{-\nu} = b^{-l}\tau^{-\nu}$, and thus

$$\nu = 1/y_\tau \tag{79}$$

[see Eq. (66) for the 2D Ising model]. Thus, with the basic assumption that there exists a fixed-point Hamiltonian with two relevant operators, the scaling form of the free energy could be derived, and moreover, a prescription for the calculation of the critical exponents is provided. A similar procedure may be followed for the correlation function [Eq. (9)]. At this point, the renormalization of the spin variables, $\sigma' = b^\zeta \sigma$, becomes important; it turns out that the choice

$$\zeta = (d - 2 + \eta)/2 \tag{80}$$

ensures the validity of Eq. (7) at the critical point.

We add some remarks on the generic structure of the flow diagram in the neighborhood of a critical fixed point (Fig. 11). In the multidimensional space of coupling coefficients, there will be one specific direction that leads away from the fixed point (the relevant direction). The other eigenvectors of the linearized RG transformation span the *critical hypersurface*. Further away from the fixed point, the hypersurface is no longer a plane but acquires some curvature. From any point on the critical hypersurface, the trajectories lead into the critical fixed point. If the starting point is close to but not at the critical hypersurface, the corresponding trajectory will first run parallel to the hypersurface until the relevant part becomes magnified sufficiently, so that finally the trajectory leaves the vicinity of the critical hypersurface, and turns off to either the low- or high-temperature fixed point. For a particular physical system (a ferromagnet, a binary mixture, etc.), the parameters will depend on the temperature (the long-dashed curve in Fig. 11). The temperature for which this curve intersects the critical hypersurface constitutes its transition temperature T_c.

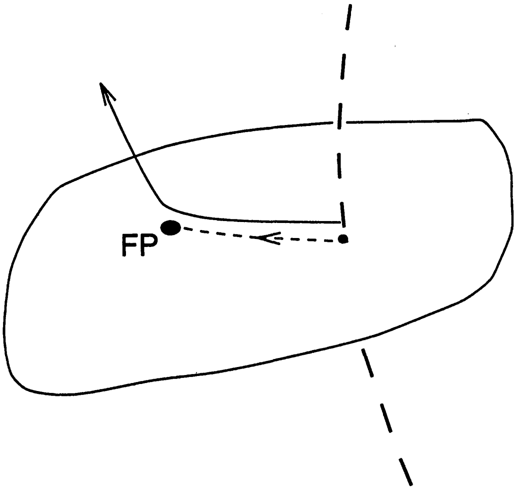

FIG. 11. Critical hypersurface. A trajectory in the critical hypersurface is shown dashed, and a trajectory starting close to the critical hypersurface is displayed as a full curve. The coupling coefficient of a particular physical system as a function of a control parameter is indicated by the long-dashed line.

From the above discussion, the universality property becomes obvious, too. All systems belonging to a particular part of parameter space, namely the region of attraction of a specific fixed point, are governed by the same scaling behavior and are described by identical power laws in the vicinity of the critical hypersurface of the fixed point.

3.3 Momentum-Shell Renormalization Group

3.3.1 Wilson's RG Scheme The RG method shall now be applied to the continuum model [Eq. (26)], called *Landau–Ginzburg–Wilson functional* in this context. A very intuitive approach was proposed by Wilson (Wilson and Kogut, 1974; Ma, 1976a). Basically, in momentum space the trace over the large-k degrees of freedom is performed, and thus recursion relations for the Ginzburg–Landau coefficients are obtained. Because the short-wavelength fluctuations are not expected to contribute essentially, the Brillouin zone may be approximated simply by a d-dimensional sphere with radius (cutoff)

Λ. The momentum-shell RG transformation then consists of the following steps:

1. Performing the trace over all Fourier components ϕ_k with $\Lambda/b < |k| < \lambda$ (Fig. 12) eliminates these short-wavelength modes.
2. By means of a scale transformation

$$k' = bk, \tag{81}$$

$$\phi' = b^{\zeta}\phi, \tag{82}$$

and therefore

$$\phi'_{k'} = b^{\zeta-d}\phi_k, \tag{83}$$

the ensuing effective Hamiltonian is brought to a form resembling the original model, by which then effective scale-dependent coupling parameters are defined.

By successive application of this RG transformation (which constitutes a semigroup, as there is no inverse element), the supposedly universal features of the long-wavelength regime are disclosed. The fixed points of the transformation will, as above, correspond to the distinct thermodynamic phases and the phase transitions between them. The eigenvalues of the linearized RG flow equations near the critical fixed point finally yield the critical exponents [see Eqs. (79), (77), and (78)].

Although a perturbational approach (with respect to u) is not at all justifiable in the critical region, it appears entirely sound in the region far from the critical point, where the fluctuations are negligible. The important observation now is that the RG flow connects these very different regimes, and thus results of the perturbation expansion obtained in the noncritical region can be transported to the vicinity of T_c, keeping track of the nonanalytic singularities on the way in a consistent, controlled, and reliable manner. Perturbative methods may also be employed in the elimination of the short-wavelength degrees of freedom (step **1**).

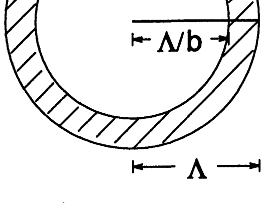

FIG. 12. Momentum-space RG: The partial trace is performed over Fourier components ϕ_k with momenta in the shell $\Lambda/b < |k| < \Lambda$.

3.3.2 Gaussian Model This concept shall now first be applied to the Gaussian model, where $u = 0$ (see Sec. 2.3.2),

$$\mathcal{F}_0[\phi_k] = \int_{|k|<\Lambda} \frac{r + k^2}{2} |\phi_k|^2 \tag{84}$$

[$\int_k = \int d^dk/(2\pi)^d$]. As Eq. (84) is diagonal in the Fourier modes, elimination of the large-k components simply produces constants; the form of the effective Hamiltonian is then unaltered, provided

$$\zeta = (d - 2)/2, \quad \text{i.e., } \eta = 0, \tag{85}$$

$$r' = b^2 r. \tag{86}$$

The fixed points of Eq. (86) are $r^* = \pm\infty$ corresponding to the high- and low-temperature phases, respectively, and the critical fixed point $r^* = 0$. The eigenvalue in the relevant temperature direction at this critical fixed point is obviously $y_\tau = 2$, and therefore the exponent $\nu = 1/2$ is recovered, as in mean-field theory.

3.3.3 Perturbation Theory and ϵ Expansion The nonlinear "interaction" term

$$\mathcal{F}_{\text{int}}[\phi_k] = \frac{u}{4} \int_{|k_i|<\Lambda} \phi_{k_1}\phi_{k_2}\phi_{k_3}\phi_{-k_1-k_2-k_3} \tag{87}$$

can now be treated perturbationally by expanding the exponential in Eq. (27) with respect to u. Separating the field variables into their parts in the outer and inner momentum shells, respectively, $\phi_k = \phi_{k_<} + \phi_{k_>}$, with $|k_<| < \Lambda/b$ and $\Lambda/b < |k_>| < \Lambda$, to first order in u one arrives at terms of the following (symbolically written) forms (henceforth, k_BT shall be set equal to 1):

1. $u \int \phi^4_< e^{-\mathcal{F}_0}$ just needs to be reexponentiated, for these degrees of freedom are not eliminated;

2. any terms with odd numbers of $\phi_<$ or $\phi_>$ like $u\int \phi_<^3\phi_> e^{-\mathcal{F}_0}$ vanish;
3. $u\int \phi_>^4 e^{-\mathcal{F}_0}$ yields a constant, contributing to the free energy; and finally
4. $u\int \phi_<^2\phi_>^2 e^{-\mathcal{F}_0}$, for which the Gaussian integral over the $\phi_>$ may be performed by using Eq. (43) for the *propagator* $\langle \phi_{k_>}^\alpha \phi_{-k_>'}^\beta\rangle_0$, an average calculated with the statistical weight $e^{-\mathcal{F}_0}$.

Quite generally, *Wick's theorem* states that expressions of the form $\langle \Pi_i^m \phi_{k_{i>}}\rangle_0$ factorize into a sum of products of all possible pairs $\langle \phi_{k_>}\phi_{-k_>}\rangle_0$, if m is even, and vanish otherwise. Especially for treating higher orders in perturbation theory, *Feynman diagrams*, with lines symbolizing the propagators, and *interaction vertices* standing for the nonlinear coupling u, provide a very helpful representation for the large number of contributions to be summed in the perturbation series.

With these means, the two-point function $\langle \phi_{k_<}\phi_{-k_<}\rangle$ and the similarly defined four-point function can be evaluated. Using Eq. (80), the following recursions for r and u are then obtained to first nontrivial order ("*1-loop*", a notation stemming from the graphical representation):

$$r' = b^2[r + (n+2)A(r)u], \tag{88}$$

$$u' = b^{4-d}u[1 - (n+8)C(r)u], \tag{89}$$

where A and C denote the integrals

$$\begin{aligned} A(r) &= K_d \int_{\Lambda/b}^{\Lambda} (k^{d-1}/r + k^2)dk \\ &= K_d[\Lambda^{d-2}(1 - b^{2-d})/(d-2) \\ &\quad - r\Lambda^{d-4}(1 - b^{4-d})/(d-4)] + O(r^2), \end{aligned}$$

$$\begin{aligned} C(r) &= K_d \int_{\Lambda/b}^{\Lambda} [k^{d-1}/(r + k^2)^2]dk \\ &= K_d\Lambda^{d-4}(1 - b^{4-d})/(d-4) + O(r), \end{aligned}$$

with $K_d = 1/2^{d-1}\pi^{d/2}\Gamma(d/2)$, and the factors containing the number of components n of the order-parameter field originate from the combinatorics of counting the equivalent ways for "contracting" the fields $\phi_{k_>}$, i.e., performing the integrals over the large momenta. Note that again Eq. (85) is valid here.

By linearizing Eqs. (88) and (89) at the Gaussian fixed point $r^* = 0$, $u^* = 0$, one immediately finds the eigenvalues $y_r = 2$ and $y_u = 4 - d$. Thus for $d > d_c = 4$, the nonlinearity $\propto u$ turns out to be irrelevant, and the mean-field exponents are valid, as has already been anticipated in Sec. 2.3.2. However, for $d < 4$ the fluctuations become relevant, and any initial $u \neq 0$ grows under renormalization. In order to obtain the scaling behavior in this case, we thus have to search for a nontrivial, finite fixed point. This is most readily done by introducing the differential flow again, via $b^l = e^{\delta l}$, and $\delta \to 0$, whereby the number of RG steps effectively becomes a continuous variable, and studying the ensuing differential recursion relations

$$\begin{aligned} \frac{dr(l)}{dl} &= 2r(l) + (n+2)u(l)K_d\Lambda^{d-2} \\ &\quad - (n+2)r(l)u(l)K_d\Lambda^{d-4}, \end{aligned} \tag{90}$$

$$\frac{du(l)}{dl} = (4-d)u(l) - (n+8)u(l)^2K_d\Lambda^{d-4}. \tag{91}$$

Namely, a fixed point is defined by the condition that $dr/dl = 0 = du/dl$. In Fig. 13, the flow of $u(l)$ according to Eq. (91) is depicted; for any initial value $u_0 \neq 0$ one finds that asymptotically, i.e., for $l \to \infty$, the nontrivial fixed point

$$u_c^* K_d = \frac{\epsilon}{n+8}\Lambda^\epsilon, \quad \epsilon = 4 - d, \tag{92}$$

is approached, which therefore describes scale-invariant behavior and should govern the universal critical properties of the model. Here, too, the RG procedure generates new interactions—for instance, terms $\propto \phi^6$ and $\nabla^2\phi^4$, etc., which in successive steps couple back into the recursion relations for r and u. However, it turns out that up to order ϵ^3 these terms may be disregarded.

The original assumption that u be small, justifying a perturbation expansion, now means that the effective expansion parameter here is the deviation from the upper critical dimension ϵ. Inserting this in Eq. (90) and collecting the terms of $O(\epsilon)$, one finds

$$r_c^* = -\frac{n+2}{2}u_c^* K_d\Lambda^{d-2} = -\frac{(n+2)\epsilon}{2(n+8)}\Lambda^2; \tag{93}$$

the physical interpretation for this result is

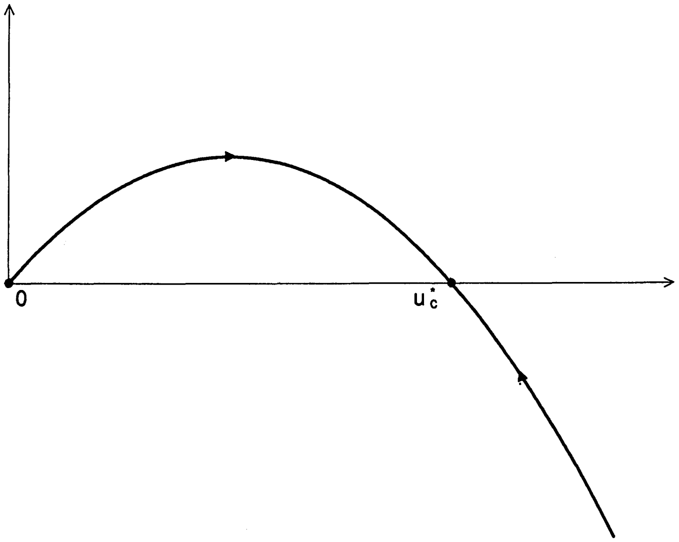

FIG. 13. Flow of the effective coupling $u(l)$, as determined from the right-hand side of Eq. (91). Both for initial values $u_0 > u_c^*$ and $0 < u_0 < u_c^*$, one has $u(l) \to u_c^*$ for $l \to \infty$.

that the fluctuations lead to a downward shift of the transition temperature. Finally, with $\tau = r - r_c^*$, the differential flow equation

$$\frac{d\tau(l)}{dl} = \tau(l)[2 - (n + 2)uK_d\Lambda^{d-4}] \qquad (94)$$

yields the eigenvalue $y_\tau = 2 - (n + 2)\epsilon/(n + 8)$ near the critical fixed point [Eq. (92)]. To one-loop order, $O(\epsilon)$, one therefore finds the critical exponent

$$\nu = \tfrac{1}{2} + \frac{n + 2}{4(n + 8)}\epsilon + O(\epsilon^2). \qquad (95)$$

Using $\eta = O(\epsilon^2)$ and the scaling relations (15), (16), and (19), one arrives at the following results [note the remarkable difference from the result of Eq. (39), obtained within the Gaussian approximation]:

$$\alpha = \frac{4 - n}{2(n + 8)}\epsilon + O(\epsilon^2), \qquad (96)$$

$$\beta = \tfrac{1}{2} - \frac{3}{2(n + 8)}\epsilon + O(\epsilon^2), \qquad (97)$$

$$\gamma = 1 + \frac{n + 2}{2(n + 8)}\epsilon + O(\epsilon^2), \qquad (98)$$

$$\delta = 3 + \epsilon + O(\epsilon^2) \qquad (99)$$

to first order in the expansion parameter $\epsilon = 4 - d$. The first nontrivial contribution to the exponent η turns out to appear at two-loop order,

$$\eta = \frac{n + 2}{2(n + 8)^2}\epsilon^2 + O(\epsilon^3). \qquad (100)$$

The universality of these results becomes manifest by the fact that they depend solely on the space dimension d and the number of order-parameter components n, and not on the original "microscopic" Ginzburg–Landau parameters. (In passing, we note that for long-range power-law interactions, $\propto |x|^{-(d+\sigma)}$, the critical exponents acquire an additional dependence on the parameter σ.) At the upper critical dimension, $d_c = 4$, the solution of Eq. (91) becomes an inverse power law instead of an exponential, which leads to logarithmic corrections to the mean-field exponents.

Differential recursion relations of the form of Eqs. (90) and (91) also constitute the basis for the treatment of more subtle issues, like the calculation of the scaling functions, or even *crossover phenomena*, within the RG approach. For instance, an anisotropic perturbation in the n-component Heisenberg model favoring m directions leads to a crossover from the O(n) Heisenberg fixed point to an m-component fixed point (see Amit, 1984, Chap. 5-3). The instability of the former is described by a crossover exponent. For a small anisotropic perturbation, the RG flow trajectory still comes very close to this unstable fixed point, which implies that further away from T_c the system behaves like an n-component system, before it is finally dominated by the anisotropic critical behavior. The crossover from one RG fixed point to another may be illustrated (and measured) by introducing *effective exponents*, defined as logarithmic derivatives of appropriate physical quantities. Other important perturbations that have been treated in the framework of RG theory are so-called cubic terms reflecting the underlying crystal structure, and contributing a fourth power of the Cartesian components of ϕ, as well as the dipolar interaction that alters the harmonic part of the theory.

For systems characterized by certain spatial anisotropies, the upper critical dimension may be shifted from $d_c = 4$ to lower values as a result of restricted space for fluctuations. For example, for uniaxial dipolar ferromagnets $d_c = 3$, as well as for tricritical points, where $u = 0$ and a sixth-order term in ϕ is required. In the case of elastic structural phase transitions, for instance, one finds $d_c = \frac{5}{2}$ or $d_c = 3$, for a one- or two-dimensional soft sector, respectively (Schwabl, 1985).

3.3.4 More Sophisticated Field-Theoretical Methods Once one has to go beyond the first order in the perturbation expansion, Wilson's k-shell renormalization scheme, despite its very intuitively appealing features, is not the best choice for practical calculations. The technical reason is that the integrals in Fourier space involve nested momenta, and become rather tedious with a finite cutoff Λ. Thus, it appears more convenient to use a field-theoretical renormalization scheme, where $\Lambda \to \infty$ (Amit, 1984). Yet this leads to additional ultraviolet (UV) divergences in the integrals for $d \geq d_c$. Observing that at the critical dimension d_c both the ultraviolet and infrared (IR) singularities appear combined in logarithmic form [$\propto \ln(\Lambda^2/r)$], the idea is to treat the UV divergences with the methods originally developed in quantum field theory, and through this diversion arrive at the correct scaling equations for the opposite, IR limit.

Formally, this is done by taking advantage of the fact that the original unrenormalized theory does not depend on an arbitrarily chosen renormalization point; as a consequence one is led to the *Callan–Symanzik* or *RG equation*. This is a partial differential equation, which, when solved via the method of characteristics, leads to a set of ordinary differential equations that resemble the differential flow equations in Wilson's scheme. This procedure may be pursued provided the theory is renormalizable, which for the ϕ^4 theory is the case for $d \leq d_c$. However, this also means that irrelevant operators cannot, at least without considerable precautions, be included in the scheme. The field-theoretical treatments are therefore confined to the critical manifold in parameter space; yet, for the universal properties of the system under consideration, they yield fully equivalent results as compared with the momentum-shell approach.

ϵ expansions have been carried out up to seventh order; the resulting series is, however, only asymptotically convergent (quite obviously, the convergence radius of the perturbation series in u must be zero, as $u < 0$ corresponds to an unstable theory). Yet, by a combination of these fairly high-order results with the divergent asymptotic behavior and Borel-type resummation techniques, the critical exponents can be obtained with impressive accuracy; see Table 1 (Le Guillou and Zinn-Justin, 1985).

Near the lower critical dimension of the O(n) ($n \geq 2$) symmetric Heisenberg model, $d = 2$ (for the Ising case, $n = 1$, the lower critical dimension is $d = 1$; see Sec. 3.2.1), the ϵ expansion does not provide very accurate results, as might be expected. Near two dimensions, the investigation of the O(n)-symmetric nonlinear σ model by means of a $2 + \epsilon$ expansion, with a nontrivial fixed point of order ϵ corresponding to T_c, has proven to be more valuable (see Amit, 1984).

Besides the ϵ expansion, an expansion in terms of powers of $1/n$ is possible, where the limit $n \to \infty$ corresponds to the exactly solvable spherical model (Ma, 1976b). While this approach may help to clarify certain general issues, it is usually less accurate numerically, though, for the small values of n that are of practical interest.

Table 1. Best estimates for the static critical exponents γ, ν, β, and δ for the O(n)-symmetric ϕ^4 model in $d = 2$ and $d = 3$ dimensions, as obtained from large-order ϵ expansion in combination with Borel resummation techniques (from Le Guillou and Zinn-Justin, 1985). For comparison, the exact Onsager results for the 2D Ising model are also listed. The limit $n = 0$ describes the statistical mechanics of polymers.

		γ	ν	β	η
$d = 2$	$n = 0$	1.39 ± 0.04	0.76 ± 0.03	0.065 ± 0.015	0.21 ± 0.05
	$n = 1$	1.73 ± 0.06	0.99 ± 0.04	0.120 ± 0.015	0.26 ± 0.05
2D Ising	(exact)	1.75	1	0.125	0.25
$d = 3$	$n = 0$	1.160 ± 0.004	0.5885 ± 0.0025	0.3025 ± 0.0025	0.031 ± 0.003
	$n = 1$	1.239 ± 0.004	0.6305 ± 0.0025	0.3265 ± 0.0025	0.037 ± 0.003
	$n = 2$	1.315 ± 0.007	0.671 ± 0.005	0.3485 ± 0.0035	0.040 ± 0.003
	$n = 3$	1.390 ± 0.010	0.710 ± 0.007	0.368 ± 0.004	0.040 ± 0.003

4. SCALING LAWS IN STATISTICAL PHYSICS

4.1 Phase Transitions and Properties of Phases

Phenomenological Ginzburg–Landau expansions, scaling approaches, and the language and methods of RG methods have become fundamental tools in the study of phase transitions and of other universal properties of condensed-matter phases, in a large variety of systems. A list (incomplete) includes structural phase transitions of various kinds, namely distortive, ferroelectric, order-disorder, and elastic transitions (see PHASE TRANSITIONS, STRUCTURAL; COLLECTIVE PHENOMENA IN SOLIDS), liquid crystals (see LIQUID CRYSTALS, STRUCTURE OF), magnetic systems (see MAGNETIC ORDERING IN SOLIDS), magnetic impurities in a lattice (*Kondo problems*; see Wilson, 1975), superfluidity (see SUPERFLUIDITY: LIQUID HELIUM SYSTEMS), and superconductivity (see SUPERCONDUCTIVITY, LOW-TEMPERATURE); here specifically the numerous phases of type-II superconductors subject to an external magnetic field have aroused considerable interest lately (see SUPERCONDUCTIVITY, HIGH-TEMPERATURE, and Blatter *et al.*, 1994). Complex scaling and crossover behavior may ensue when either anisotropies or additional long-range interactions are included.

Finite-size scaling studies how the asymptotic power laws of the thermodynamic limit are approached as the system size grows, and is thus an important tool for numerical studies of continuous phase transitions (Privman, 1990). Furthermore, the fact that all real systems are finite leads to additional interesting phenomena. At surfaces, new critical exponents may come into play (Diehl, 1986). Also, interfacial structures may change drastically as external parameters are varied; an important example is the wetting transition of liquid drops on a substrate (Dietrich, 1988).

Scaling and RG concepts have also been applied to the theory of symmetry breaking in the fundamental field theories of matter (see UNIFIED FIELD THEORIES), which also have important implications for cosmology (see GALAXIES AND COSMOLOGY). (We note that in elementary-particle physics a Compton wavelength $\lambda_c = \hbar c/m$ may be assigned to a particle of rest mass m. For large momenta $\hbar k \gg mc$, the theory becomes scale invariant, and the propagator behaves as $k^{-2+\eta}$. The correlation length ξ thus corresponds to the Compton wavelength or inverse mass.)

4.2 Disordered Systems

In many physical systems, disorder plays an important role. The basic question therefore is, under what circumstances can quenched randomness affect the nature of a phase transition and its properties. Let the disorder induce a relevant perturbation $g(x)$ in the effective Hamiltonian, e.g., changing the harmonic part of the Landau–Ginzburg–Wilson functional (*random* T_c), the scaling dimension of which is $[g] \propto L^{\Delta-d}$, with $\Delta \leq d$. The *Harris criterion* then states that the character of the pure transition may be considerably affected by the disorder, provided $\Delta < d/2$ (Itzykson and Drouffe, 1989). If defects of this kind are correlated only over short distances, this criterion says that the critical behavior is changed if the specific-heat exponent of the pure system is positive ($\alpha > 0$).

On the other hand, in the case of *random fields*, both the upper and lower critical dimensions can be changed as compared with the pure case. The *Imry–Ma argument* (see Nattermann and Villain, 1988) compares the interfacial energy $\propto L^\theta$ [Eq. (13)] with the gain in bulk free energy $\propto L^{d/2}$ by splitting into domains of size L (the mean square order-parameter fluctuations due to the random field add up to a contribution $\propto L^d$ for each domain). Thus, for $\theta \leq d/2$, the system will always break into domains and no long-range order can emerge. For Ising systems ($\theta = d - 1$), this will be true for $d \leq 2$, while in the case of continuous symmetry ($\theta = d - 2$) the lower critical dimension becomes $d = 4$.

An example for such disorder-dominated systems is *spin glasses*, e.g., an Ising spin glass with bond disorder, where the couplings J_{ij} in Eq. (21) are drawn from some stochastic distribution. Similarly, disorder can be introduced in Heisenberg models [Eq. (25)]. Because spin glasses are model systems for phase transitions in highly disordered materials, a considerable amount of effort has been devoted to the examination of such spin glasses (Binder and Young, 1986). In addition to the thermal average, all physical quantities have to be averaged over all possible configurations of the quenched disorder. A frequently used method in analytical calculations is the *replica trick*, which replaces the computation of the mean free energy by the average of the nth power of the partition function, and then entails taking both the derivative with respect to n and the limit $n \to 0$:

$$\bar{F} = -\lim_{n\to 0} \frac{\partial}{\partial n} \overline{Z^n}. \qquad (101)$$

Extensions of these studies have led to the investigation of manifolds in random media, a direct application of which may be found in biological membranes, or flux lines in high-T_c superconductors, and the design of neural networks (see NEURAL NETWORKS).

Disorder may also play a prominent role in the quantum regime ($T \to 0$). While near a phase transition with finite T_c quantum fluctuations are of minor importance (although quantum effects may provide the fundamental mechanism for the transition itself, as in magnets, superfluids, and superconductors), at $T = 0$ the interplay between disorder and quantum mechanics can lead to very interesting novel phases with unusual phase transitions. Examples of such quantum critical phenomena (see Continentino, 1994) are electron localization (see ELECTRON STATES, LOCALIZED), metal-insulator transitions (see METAL-INSULATOR TRANSITIONS; Belitz and Kirkpatrick, 1994), helium-4 on disordered substrates (disordered bosons), and quantum antiferromagnets, which are of relevance in the microscopic theory of high-temperature superconducting materials and as one phase of solid helium-3. For a Fermi system, the interesting singularities and the universal behavior are expected in the vicinity of the Fermi surface, instead of the point $k = 0$ in Fourier space, and therefore the RG approach has to be modified accordingly (Shankar, 1994).

The mathematical theory of percolation also has immediate applications for the structure of certain disordered systems, and their transport properties (see FRACTAL GEOMETRY; Bunde and Havlin, 1991). The percolation threshold may be understood as a critical point, and the scaling behavior in its vicinity can be obtained by mapping onto the so-called Potts model (Wu, 1982), and with the help of certain continuum models of the same universality class.

4.3 Polymers, Membranes

An important ingredient to the physics of macromolecules like polymers is the interplay and competition of energetic and entropic optimization. De Gennes has demonstrated that the partition function of a polymer, or the statistics of a self-avoiding random walk, can be mapped to the $n \to 0$ limit of the n-component Heisenberg model (see Table 1). As a consequence, scaling and RG approaches to the static and dynamic properties of polymers have been successfully employed (see POLYMERS, STRUCTURE OF; De Gennes, 1979; Freed, 1987). The inverse separation from the critical point $|\tau|^{-1}$ formally corresponds to the degree of polymerization (chain length) N of a long flexible polymer, and, e.g., in a dilute solution in a good solvent the average end-to-end distance of the polymer scales as $R_e \propto N^\nu$, where $\nu \approx 0.588$ (in three dimensions) is the $n = 0$ result for the correlation length exponent of the O(n) model. Naturally, similar attempts have been used for biological systems like muscular fibers and membranes (see STATISTICAL PHYSICS OF MEMBRANES AND LAMELLAR SYSTEMS).

4.4 Critical Dynamics and Nonequilibrium Phenomena

In Sec. 1.4, the extension of the above scaling concepts to dynamical phenomena was mentioned. The order-parameter dynamics displays critical slowing down near a second-order phase transition, described by the critical exponent z [Eq. (11)]. One very powerful and common mathematical description of critical dynamics rests on *Langevin-type equations of motion* (see STOCHASTIC PROCESSES) for the order-parameter fluctuations and the conserved quantities, which constitute the *slow modes* of the system (Hohenberg and Halperin, 1977; Enz, 1979). For, as a consequence of the continuity equation related to a conservation law, the corresponding density shows diffusive behavior (inverse relaxation time $t_0^{-1} \propto k^2$). In a phase with broken continuous symmetry, in addition the fluctuations transverse to the order parameter have to be included, for their static susceptibility is also singular for $k \to 0$ [Eq. (47)], and thus the corresponding mode has frequency $\propto \chi^{-1/2}$, and damping coefficient $\propto \chi^{-1}$. Collecting all these slow modes in a multicomponent field $\Phi(x,t)$, the equations of motion assume the general form

$$\frac{\partial \Phi^\alpha(x,t)}{\partial t} = K[\Phi] - \Gamma^\alpha \frac{\delta \mathcal{F}[\Phi]}{\delta \Phi^\alpha(x,t)} + \zeta^\alpha(x,t). \tag{102}$$

Here, $K[\Phi]$ comprises the reversible forces, as obtained from the microscopic equations of motion. The second term provides a damping term, which after some small perturbation drives the system back into the equilibrium state, where $\delta\mathcal{F}/\delta\Phi = 0$; the damping coefficients Γ^α are constants for the order-parameter fluctuations, if the latter is not conserved, and take the form $\lambda\nabla^2$ for each conserved quantity. Finally, all "fast" modes have been subsumed in the stochastic forces $\zeta^\alpha(x,t)$, whose mean is taken to be zero, $\langle\zeta^\alpha(x,t)\rangle = 0$, and whose second moments are, for simplicity, assumed to be drawn from a Gaussian distribution with a variance corresponding to "white noise":

$$\langle\zeta^\alpha(x,t)\zeta^\beta(x',t')\rangle = 2k_B T\Gamma^\alpha\delta^{\alpha\beta}\delta(x - x')\delta(t - t'). \tag{103}$$

The *fluctuation-dissipation theorem (Einstein relation)* [Eq. (103)] assures that for $t \to \infty$ the probability distribution is given by the equilibrium expression $\exp(-\mathcal{F}/k_B T)$.

The set of Eqs. (102) with (103) can then be very elegantly mapped onto a dynamic functional, and the ensuing field theory can be treated with precisely those perturbation expansions and RG methods explained above for the statics. In some cases, the underlying symmetries and/or conservation laws lead to so-called *Ward identities*, which yield a connection between the dynamic exponent z and already known static indices, prominently η. For example, for a model merely consisting of a conserved order-parameter field (model B according to the classification of Hohenberg and Halperin, 1977), the result is $z = 4 - \eta$; for the isotropic ferromagnet (model J) one finds $z = (d + 2 - \eta)/2$, and for planar ferromagnets as well as isotropic antiferromagnets (models E and G, respectively), $z = d/2$. In Table 2, a list of dynamical systems, with the corresponding values of z, and the physical quantities for which the critical slowing down can be observed in experiment, is presented. Note that although the range of the dynamical critical exponent is limited to $\frac{3}{2} \le z \le 4$ in three dimensions, the ensuing behavior of the transport coefficients can yet be as different as the diverging magnetization diffusion $D_M \propto |\tau|^{-1/3}$ for an isotropic antiferromagnet, and the vanishing order-parameter relaxation rate $\Gamma \propto |\tau|$ in the same system, or the vanishing magnetization diffusion constant $D_M \propto |\tau|^{1/3}$ for an isotropic Heisenberg ferromagnet.

An alternative route to finding these exponents, and probably superior for the issue of precisely calculating the scaling functions, is the *mode-coupling* approach, which at least in some cases may be represented as a self-consistent one-loop theory for the generalized (i.e., wave-number– and frequency-dependent) transport coefficients (Frey and Schwabl, 1995).

Another physical situation, where equations of motion of the type in Eq. (102) and a RG approach have been used, is the problem of quenching a system from above T_c into the low-temperature phase, and observing

Table 2. Dynamic critical exponents z in d dimensions, for various dynamical systems, along with a list of the physical quantities, and their temperature dependence in the vicinity of T_c, for which the ensuing critical slowing down may be observed experimentally (classification according to Hohenberg and Halperin, 1977). The approximate numerical results are for the three-dimensional case, using $\nu \approx \frac{2}{3}$, $\eta \approx 0$, and $\alpha \approx 0$.

System	Model	Dynamic exponent	Physical quantity
Anisotropic magnet	A	$z = 2 + c\eta,\ c = O(1)$	Relaxation rate $\Gamma \propto \lvert\tau\rvert^{z\nu} \approx \lvert\tau\rvert^{4/3}$
Uniaxial ferromagnet	B	$z = 4 - \eta$	Magnetization diffusion $D_M \propto \lvert\tau\rvert^{(z-2)\nu} \approx \lvert\tau\rvert^{4/3}$
Structural phase transitions	C	$z = 2 + \alpha/\nu$	Relaxation rate $\Gamma \propto \lvert\tau\rvert^{z\nu} \approx \lvert\tau\rvert^{4/3}$
Planar ferromagnet	E	$z = d/2$	Magnetization diffusion $D_M \propto \lvert\tau\rvert^{z\nu} \approx \lvert\tau\rvert$
Superfluid helium-4	F	$z = \frac{1}{2}(d + \tilde{\alpha}),\ \tilde{\alpha} = \max(\alpha,0)$	Thermal conductivity $\lambda_T \propto \lvert\tau\rvert^{(z-2)\nu} \approx \lvert\tau\rvert^{-1/3}$ (+ logarithmic corrections)
Isotropic antiferromagnet	G	$z = d/2$	Order-parameter relaxation $\Gamma_N \propto \lvert\tau\rvert^{z\nu} \approx \lvert\tau\rvert$ Magnetization diffusion $D_M \propto \lvert\tau\rvert^{(z-2)\nu} \approx \lvert\tau\rvert^{-1/3}$
Isotropic Heisenberg ferromagnet	J	$z = \frac{1}{2}(d + 2 - \eta)$	Magnetization diffusion $D_M \propto \lvert\tau\rvert^{(z-2)\nu} \approx \lvert\tau\rvert^{1/3}$

how the ordered domains grow with time (Bray, 1993).

Upon relaxing the condition of Eq. (103), the same methods may also be applied to phase transitions in nonequilibrium systems. A prototypical example might be the Burgers–Kardar–Parisi–Zhang equation, which is supposed to describe the kinetic roughening of growing surfaces (Family and Vicsek, 1991), but is also intimately connected to the problem of directed lines in random environment, and to the hydrodynamics of randomly stirred fluids. Related issues are the depinning transitions of lines and interfaces, and the dynamics of charge-density waves. The behavior of interfaces in disordered systems in general poses a number of most interesting issues and requires certain new conceptual approaches (see Forgacs *et al.*, 1991).

The properties of systems very far from equilibrium, well within the scope of scaling approaches and RG methods, turn out to be remarkably rich. On the one hand, there is the vast field of nonequilibrium pattern formation (Cross and Hohenberg, 1994). On the other hand, self-similar and self-affine structures emerge in many dynamical processes (see FRACTAL GEOMETRY), which are hence good candidates for the application of the above methods. The ubiquity of fractals and $1/f$ noise has led to the paradigm of *self-organized criticality (SOC)* (reviewed in Bak and Creutz, 1994). SOC refers to the tendency of large dissipative systems to drive themselves into a self-similar critical state without the special adjustment of parameters. SOC has been suggested for avalanches in sand piles, earthquake statistics, evolution, spreading of diseases, forest fires, and many others (Bak and Creutz, 1994). In the so-called forest-fire model, one finds SOC if a certain double separation of time scales holds (which represents still a large region in parameter space), while in other regions of the parameter space spiral structures emerge (Drossel and Schwabl, 1994).

At any rate, scaling ideas and the language of the renormalization group, including such concepts as effective Hamiltonians, relevant and irrelevant perturbations, fixed points, diverging correlation lengths, self-similarity and power laws, and universal long-wavelength and low-frequency behavior of collective modes, as well its powerful mathematical tools, have provided considerable new insight into the physics of many important research areas, and it seems likely that more interesting applications will be discovered in the future.

GLOSSARY

Asymptotic Regime: Region near the critical point where the relevant perturbations dominate, leading to pure power laws.

Corrections to Scaling: Contributions of irrelevant operators to thermodynamic quantities.

Correlation Length: Length scale for the spatial decay of correlations, usually a function of the external control parameters (temperature, magnetic field, etc.).

Critical Exponents: Exponents describing the power laws at continuous phase transitions.

Critical Slowing Down: Decrease of the order-parameter decay or oscillation rate ω_c, characterized by the dynamic exponent z; cf. Eq. (11).

Critical Surface: Hypersurface in parameter space with the property that all RG trajectories starting on it lead into the critical fixed point.

Crossover: Change in scaling behavior from power laws characterized by a certain unstable fixed point to different power laws governed by the asymptotically stable fixed point.

Effective Theory: Mathematical model describing the long-wavelength and low-frequency behavior of a physical system, at a certain energy scale.

Fixed Point: Point in the space of coupling coefficients that remains invariant under the RG transformation.

Lower Critical Dimension: Borderline dimension below which a certain model or physical system does not have a phase transition leading to long-range order.

Order Parameter: Certain macroscopic quantity that may be used to characterize different phases.

Relevant and Irrelevant Perturbations: Eigenoperators of the linearized RG transformation that grow and decrease, respectively, under successive RG steps.

Renormalization-Group (RG) Transformation: Operation by which a Hamiltonian is transformed into a new one, with a reduced number of degrees of freedom, accompanied by a subsequent scale transformation.

Scaling Law: Mathematical expression of the fact that at a critical point the free energy and correlation functions are scale invariant, i.e., are generalized homogeneous functions.

Scaling Relations: Equations that relate certain critical indices, thus reducing the number of independent exponents.

Universality: The fact that critical properties depend only on global properties such as the dimension and the symmetry and number of components of the order parameter.

Upper Critical Dimension: Borderline dimension above which the critical exponents of a certain model or physical system are correctly given by their mean-field values. Below the upper critical dimension, fluctuations are relevant and affect the scaling behavior.

Works Cited

Als-Nielsen, J., Birgeneau, R. J. (1977), *Am. J. Phys.* **45**, 554–560.

Amit, D. J. (1984), *Field Theory, the Renormalization Group, and Critical Phenomena*, 2nd ed., Singapore: World Scientific.

Bak, P., Creutz, M. (1994), in: Bunde, A., Havlin, S. (Eds.), *Fractals in Science*, Berlin: Springer-Verlag, pp. 26–47.

Belitz, D., Kirkpatrick, T. R. (1994), *Rev. Mod. Phys.* **66**, 261–380.

Binder, K., Young, A. P. (1986), *Rev. Mod. Phys.* **58**, 801–976.

Blatter, G., Feigel'man, M. V., Geshkenbein, V. B., Larkin, A. I., Vinokur, V. M. (1994), *Rev. Mod. Phys.* **66**, 1125–1388.

Bray, A. J. (1993), in: Riste, T., Sherrington, D. (Eds.), *Phase Transitions and Relaxation in Systems with Competing Energy Scales*, Dordrecht: Kluwer Academic Publishers, pp. 405–436.

Buckingham, M. J., Fairbank, W. M. (1961), in: Gorter, C. J. (Ed.), *Progress in Low Temperature Physics*, Vol. III, Amsterdam: North Holland Publishing Company, pp. 80–112.

Bunde, A., Havlin, S. (1991), in: Bunde, A., Havlin, S. (Eds.), *Fractals and Disordered Systems*, Berlin: Springer-Verlag, pp. 51–149.

Chaikin, P. M., Lubensky, T. C. (1995), *Principles of Condensed Matter Physics*, Cambridge, U.K.: Cambridge Univ. Press.

Continentino, M. A. (1994), *Phys. Rep.* **239**, 179–213.

Cross, M. C., Hohenberg, P. C. (1994), *Rev. Mod. Phys.* **65**, 851–1112.

de Gennes, P. G. (1979), *Scaling Concepts in Polymer Physics*, Ithaca: Cornell Univ. Press.

Diehl, H. W. (1986), in: Domb, C., Lebowitz, J. L. (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 10, London: Academic Press, pp. 75–267.

Dietrich, S. (1988), in: Domb, C., Lebowitz, J. L. (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 12, London: Academic Press, pp. 1–218.

Drossel, B., Schwabl, F. (1994), *Physica A* **204**, 212–229.

Enz, C. P. (Ed.) (1979), *Dynamic Critical Phenomena and Related Topics*, Lecture Notes in Physics Vol. 104, Berlin: Springer-Verlag.

Family, F., Vicsek, T. (Eds.) (1991), *Dynamics of Fractal Surfaces*, Singapore: World Scientific.

Fisher, D. S. (1993), in: Riste, T., Sherrington, D. (Eds.), *Phase Transitions and Relaxation in Systems with Competing Energy Scales*, Dordrecht: Kluwer Academic Publishers, pp. 1–25.

Fisher, M. E. (1967), *Rep. Prog. Phys.* **30**, 615–730.

Fisher, M. E. (1974), *Rev. Mod. Phys.* **46**, 597–616.

Forgacs, G., Lipowsky, R., Nieuwenhuizen, T. M. (1991), in: Domb, C., Lebowitz, J. L. (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 14, London: Academic Press, pp. 135–363.

Freed, K. F. (1987), *Renormalization Group Theory of Macromolecules*, New York: Wiley.

Frey, E., Schwabl, F. (1995), *Adv. Phys.* (in press).

Hohenberg, P. C., Halperin, B. I. (1977), *Rev. Mod. Phys.* **49**, 435–479.

Itzykson, C., Drouffe, J.-M. (1989), *Statistical Field Theory*, Vols. 1, 2, Cambridge, U.K.: Cambridge Univ. Press.

Le Guillou, J. C., Zinn-Justin, J. (1985), *J. Phys. Lett.* **46**, L-137–L-141, and references therein.

Ma, S. (1976a), *Modern Theory of Critical Phenomena*, Reading, MA: Benjamin.

Ma, S. (1976b), in: Domb, C., Green, M. S. (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 6, London: Academic Press, pp. 249–292.

Nattermann, T., Villain, J. (1988), *Phase Transitions* **11**, 5–51.

Nelson, D. R. (1983), in: Domb, C., Lebowitz, J. L. (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 7, London: Academic Press, pp. 1–99.

Niemeijer, T., van Leeuwen, J. M. J. (1976), in: Domb, C., Green, M. S. (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 6, London: Academic Press, pp. 425–505.

Privman, V. (Ed.) (1990), *Finite Size Scaling and Numerical Simulation of Statistical Systems*, Singapore: World Scientific.

Schwabl, F. (1985), *J. Stat. Phys.* **39**, 719–737.

Shankar, R. (1994), *Rev. Mod. Phys.* **66**, 129–192.

Swendsen, R. H. (1984), *J. Stat. Phys.* **34**, 963–973.

Wilson, K. G., Kogut, J. (1974), *Phys. Rep. C* **12**, 76–199.

Wilson, K. G. (1975), *Rev. Mod. Phys.* **47**, 773–840.

Wu, F. Y. (1982), *Rev. Mod. Phys.* **54**, 235–268.

Further Reading

General

Binney, J. J., Dowrick, N. J., Fisher, A. J., Newman, M. E. J. (1993), *The Theory of Critical Phenomena*, 2nd ed., New York: Oxford Univ. Press.

Brey, J., Jones, R. B. (Eds.) (1976), *Critical Phenomena*, Lecture Notes in Physics Vol. 54, Berlin: Springer-Verlag.

Domb, C., Green, M. S. (Eds.) (1972–1976), *Phase Transitions and Critical Phenomena*, Vols. 1–6, London: Academic Press.

Domb, C., Lebowitz, J. L. (Eds.) (1983–1992), *Phase Transitions and Critical Phenomena*, Vols. 7–15, London: Academic Press.

Goldenfeld, N. (1992), *Lectures on Phase Transitions and the Renormalization Group*, Reading, MA: Addison-Wesley.

Patashinskii, A. Z., Pokrovskii, V. L. (1979), *Fluctuation Theory of Phase Transitions*, New York: Pergamon Press.

Stanley, H. E. (1971), *Introduction to Phase Transitions and Critical Phenomena*, Oxford, U.K.: Clarendon Press.

Yeomans, J. M. (1991), *Statistical Mechanics of Phase Transitions*, Oxford, U.K.: Clarendon Press.

Field Theory

Parisi, G. (1988), *Statistical Field Theory*, Redwood, CA: Addison-Wesley.

Zinn-Justin, J. (1989), *Quantum Field Theory and Critical Phenomena*, Oxford, U.K.: Clarendon Press.

PHASE TRANSITIONS, STRUCTURAL

E. K. H. SALJE, *Department of Earth Sciences, University of Cambridge, Cambridge, United Kingdom*

	Introduction	373
1.	**Entropic Effects for Temperature-Driven SPTs**	374
1.1	Displacive Phase Transitions	375
1.2	"Almost" Displacive Transitions	376
1.3	The Order-Disorder Limit	376
2.	**The Order Parameter**	376
3.	**Ferroelasticity and Ferroelectricity**	379
4.	**Coupling between Order Parameters**	381
5.	**Influence of Defects**	382
6.	**Twins and Antiphase Boundaries as Fingerprints of SPTs: Surface Relaxations**	383
7.	**Kinetics of Site Ordering**	384
8.	**Conclusion**	386
	Glossary	387
	Works Cited	387

INTRODUCTION

When crystalline materials are subjected to changes of temperature *T*, pressure *P*, magnetic field *H*, etc., they will change their atomic arrangements. If such changes involve topological modifications of the crystal structure and/or changes of symmetry, the various structures represent phases with structural phase transitions (SPTs) between them.

A typical example for a phase transition between a hexagonal structure (at high temperatures) and a trigonal structure (at low temperatures) is shown for quartz (SiO_2) (Fig. 1). Small rotations of SiO_4 tetrahedra destroy the hexagonal symmetry and lead to a denser arrangement of tetrahedra at low temperatures. The amplitudes of the various polyhedra tilts increase with decreasing temperature, whereas no such tilts exist at temperatures above the transition point. This transition mechanism, with gradual and mostly very subtle atomic shifts, is termed "displacive." (A more precise definition of the term "displacive" is based on the appearance of soft phonons as fundamental excitation of such systems; a simple static picture is given later.)

Symmetry breaking can occur via many other mechanisms. A common mechanism involves the ordering of different sorts of atoms on various positions in a crystal structure. This process is often seen in alloys, which may form a random distribution of metal atoms such as Cu and Au at high temperatures but an ordered structure at low temperatures. Another example, more similar to that of quartz, is the feldspar structure. Feldspars are most common materials: ca. 60% of the outer crust of the planet Earth is made from feldspars and related materials. In Fig. 2, the framework of feldspar at high and low temperatures is depicted. In each case, the structure is made up from oxygen tetrahedra with a metal position inside. This metal atom can be Al or Si. At high temperatures, Al and Si are disordered; at low temperatures, they are ordered whereby the order starts at a well-defined transition temperature. The degree of order increases when cooled further. Note that the AlO_4 and SiO_4 tetrahedra have different sizes so that the ordering of Al and Si is accompanied by local strain in the structure, which accommodates for the appropriate atomic ordering. SPTs in which the essential mechanism is related to changes of atomic order are usually called "order-disorder" (or simply od).

Many other transition mechanisms exist, and it is not just temperature that acts as the relevant variable to generate SPTs but also pressure, variations of the chemical composition, etc. Furthermore, all these mecha-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

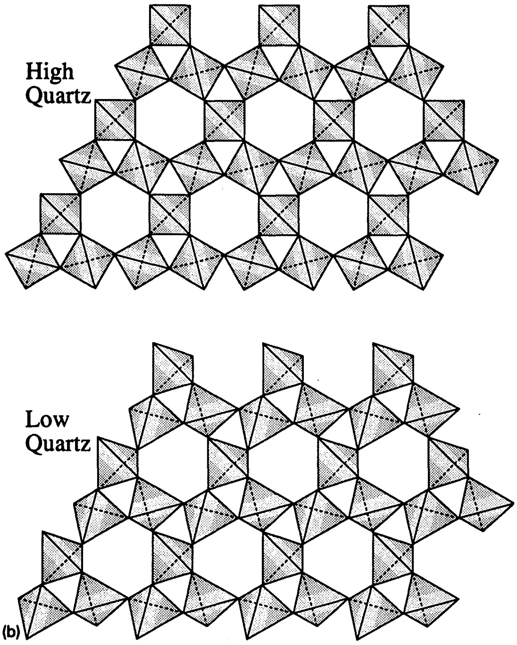

FIG. 1. Crystal structures of quartz at temperatures above and below the transition temperature. Rather rigid SiO_4 tetrahedra build up a framework that reduces its specific volume at low temperature via tilts of the tetrahedra. Courtesy A. Putnis, Cambridge.

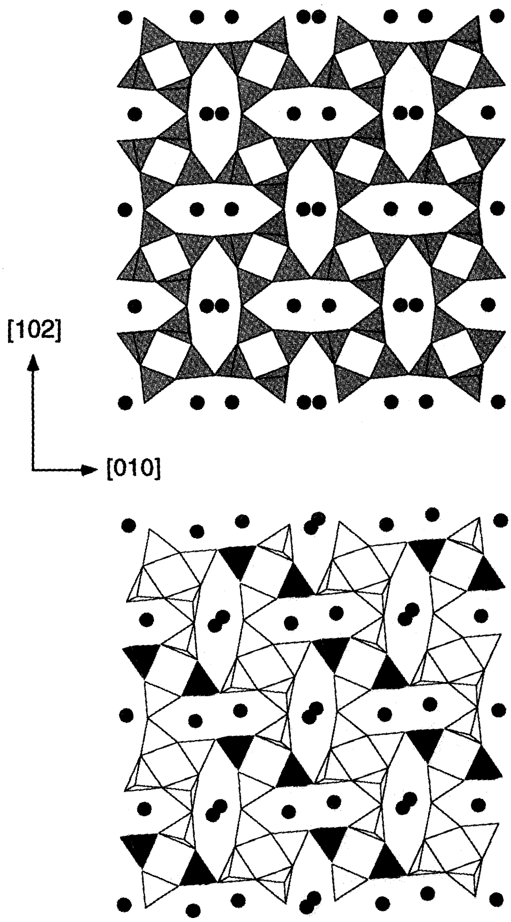

FIG. 2. Structural changes in alkali feldspars (Na,K)-$AlSi_3O_8$: the high-symmetry form is monoclinic with random distribution of Al and Si on tetrahedral sites. The alkali atom reflects the overall monoclinic symmetry (top). The triclinic, low-symmetry form shows Al ordered on the black tetrahedra whereas Si is inside the unshaded tetrahedra. The monoclinic symmetry is broken during the STP. The locations of alkali atoms are indicated by the circles. Courtesy D. Palmer, Cambridge.

nisms can interplay, so that materials often show several SPTs, each of which influences all others. It is an essential first step for research on SPTs to disentangle the various contributions (Landau and Lifshitz, 1980; Bruce and Cowley, 1981; Khatchaturyan, 1983; Levanyuk and Sigov, 1988; Salje, 1993b).

1. ENTROPIC EFFECTS FOR TEMPERATURE-DRIVEN SPTs

A SPT occurs when the enthalpic energy H is compensated by entropy S so that the Gibbs free energy $G = H - TS$ changes sign. A simple atomistic picture illustrates this mechanism (Salje *et al.*, 1991). Let us first consider local potentials that are anharmonic so that the midpoint of the potential is not necessarily at an energy minimum (Fig. 3). The second energy contribution is the coupling between the relevant atomic positions. In SPT, this communication is mainly achieved by elastic deformation of the structure (structural stress) or via electric dipolar forces. In both cases, the effective interaction decays as $1/R^3$ with increasing distance R. This long-ranging effective interaction together with its strong directional dependencies (in contrast to nearest-neighbor Ising models) allows us to summarize the effect of all interactions on each atom as a field ("mean field theory," MFT), and we may depict the interatomic coupling as a simple spring (Fig. 3). The effective Hamiltonian of this model is then the sum over the kinetic energy, the local potentials, and the coupling energy:

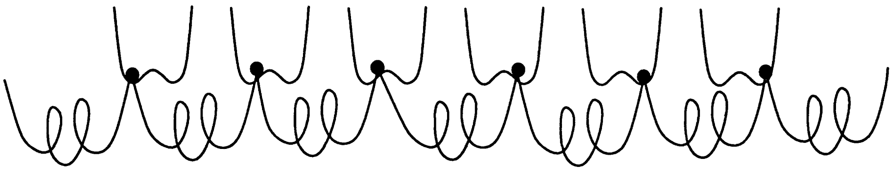

FIG. 3. Simple model representing a physical system with nonharmonic local potentials and harmonic intersite interactions.

$$H = \sum_l \left(\frac{1}{2M}P_l^2 + \frac{1}{2}M\Omega_{ol}^2 Q_l^2 + \frac{1}{2}uQ_l^4 + \cdots\right) - \frac{1}{2}\sum_{l\neq l'} v_{ll'}Q_lQ_{l'}. \tag{1}$$

Here M is the effective atomic mass, P is the momentum, Ω_{ol} is the local frequency for small oscillations around the local energy minimum at site l if it were purely parabolic in the absence of interatomic interactions, u gives the strength of the anharmonicity, and $v_{ll'}$ are the coupling constants that form the amplitudes Q_l on site l and $Q_{l'}$ on site l'. We can now define an order parameter Q as the time- and space-averaged value of the position coordinates Q_l on all sites l. This average value is denoted by angular brackets $Q = \langle Q_l \rangle$. If the average position is situated in the middle of the potential, we find $Q = 0$. If the average position is shifted to the right or the left, we find $Q > 0$ or $Q < 0$. One sees already in this simple model that $Q = 0$ represents a high-symmetry phase, whereas $Q > 0$ and $Q < 0$ represent a low-symmetry phase. If the low-symmetry phase contains macroscopic areas with $Q > 0$ and others with $Q < 0$, these areas are called "domains" with "domain walls" between them. The order parameter is a one-dimensional parameter in this simple model; higher dimensional order parameters are needed in 3D structures with high crystallographic symmetries. If the site ordering is not uniform, the order parameter is defined as the relevant Fourier component of $\langle Q_l \rangle$:

$$\langle Q_l \rangle = Q = Q_0 \exp(-i\mathbf{q}\cdot\mathbf{r}).$$

Thermal fluctuations occur around the order parameter, with

$$\sigma = \langle (Q_l - Q)^2 \rangle,$$

which depend on temperature via the dissipation-fluctuation theorem as

$$\sigma = \frac{\hbar}{2M\Omega}\coth\frac{\hbar\Omega}{2k_\mathrm{B}T}. \tag{2}$$

The Gibbs free energy G can then be calculated in MFT at the wave vector q at which long-range order with the order parameter (OP) $Q = \langle Q_l \rangle$ occurs. The SPT occurs when G changes sign with $Q = 0$ in the high-symmetry phase and a gradual increase of $Q \neq 0$ in the low-symmetry phase.

Approximate solutions exist for three regimes, which are characterized by a parameter $\Delta = v/M\Omega_0^2 - 1$ that vanishes when the intersite coupling compensates the parabolic part of the on-site potential. We furthermore define two temperatures. Firstly, T_c defines the temperature below which the high-symmetry phase becomes unstable. Secondly, thermal fluctuations saturate at temperatures below $\theta_s = \hbar\Omega_0/2k_\mathrm{B}$, whereas at $T \gg \theta_s$ fluctuations increase linearly with temperature [Eq. (2)].

The three solutions mostly used for the analysis of experimental observations are the following.

1.1 Displacive Phase Transitions

These have $0 \leqslant \Delta \ll 1$, and

$$G = \tfrac{1}{2}A\theta_s\left[\coth\frac{\theta_s}{T} - \coth\frac{\theta_s}{T_c}\right]Q^2 + \tfrac{1}{4}BQ^4 + \tfrac{1}{6}CQ^6, \tag{3}$$

which becomes for $\theta_s \ll T$ the standard Landau potential

$$G = \tfrac{1}{2}A(T - T_c)Q^2 + \tfrac{1}{4}BQ^4 + \tfrac{1}{6}CQ^4. \tag{4}$$

Here Q follows from the minimum condition

$\partial G/\partial Q = 0$ (see below). Note that the standard Landau potential is a high-temperature approximation and does not hold at temperatures at which the quantum-mechanical treatment of phonons becomes relevant.

1.2 "Almost" Displacive Transitions

These also have $\Delta \leqslant 1$, and

$$\sigma = \sigma_c\left(\frac{\eta\Omega_0}{\Omega}\right)\coth\left(\frac{\eta\Omega}{\Omega_0}\right), \tag{5}$$

$$\Omega^2 = \Omega_0^2[1 + \Delta(3 - 2\sigma/\sigma_c)], \tag{6}$$

where

$$\eta = \hbar/2M\Omega_0\sigma_c, \quad \sigma_c = (v - M\Omega^2)/3u,$$
$$Q^2 = 3(\sigma_c - \sigma). \tag{7)–(9}$$

These "self-consistency equations" have to be solved numerically in order to obtain the value of Q for each temperature T. The resulting curves of Q vs T show smaller values of Q at low temperatures compared with the equivalent solution [Eq. (3)].

1.3 The Order-Disorder Limit

The order-disorder (od) limit is reasonably well described by the Bragg–Williams equation

$$Q = \tanh(QT_c/T), \quad \Delta \gg 1. \tag{10}$$

In Fig. 4, a sequence of measured OPs are compared with theoretical predictions for $\Delta < 1$. In all cases, Landau theory is a good approximation for $T \gg \theta_s$, and we list the analytical expressions for the temperature dependence of the OP for the Landau potential [Eq. (4)]:

1. a "second-order" transition:

$$Q = \left(\frac{T_c - T}{T_c}\right)^{1/2} \tag{11}$$

for $T < T_c$, $C = 0$, $B > 0$;

2. a "tricritical" transition:

$$Q = \left(\frac{T_c - T}{T_c}\right)^{1/4} \tag{12}$$

for $T < T_c$, $C > 0$, $B = 0$;

3. a "first-order" transition

$$Q^2 = -\frac{1}{2C}[B - \sqrt{B^2 - 4AC(T - T_c)}]$$

for

$$T < T_{\text{trans}} = T_c + B^2/4AC, \tag{13}$$

$$B < 0, C > 0.$$

Other Landau potentials including odd-order terms (when compatible with the symmetry of the crystal) or higher-order terms lead to similar expressions for $Q(T)$.

Experimental observations often follow the predicted T dependence of Q very well—in particular for displacive phase transitions (Fig. 4). In fact, the applicability of the generalized Landau expression in Eq. (3) is much wider than originally anticipated and includes thermal regimes with OPs close to unity (Salje, 1993b; Salje *et al.*, 1991). Furthermore, structural phase transitions for which the order parameter disappears continuously on heating through the transition point (cases **1** and **2**) do not necessarily show criticalities as in case of phase transitions with short-range interactions such as described in the nearest-neighbor Ising model. In most cases, critical fluctuations are suppressed by the long-ranging and strongly anisotropic elastic interactions in the crystal structure (Cowley, 1976; Folk *et al.*, 1979; Marais *et al.*, 1991).

2. THE ORDER PARAMETER

The order parameter, OP, was so far introduced as a measure for the deviation of the structural state in the low-symmetry phase ($Q \neq 0$) from that of the high-symmetry phase ($Q = 0$). Its meaning stems from thermodynamic requirements, and it is not always obvious how the order parameter is related to structural quantities that maintain the general structural topology of the high-symmetry phase. One finds that all symmetry elements of the low-symmetry structure are already present in the high-symmetry phase. Some symmetry elements of the high symmetry, on the other hand, disappear during the phase transition, i.e., specific symmetry elements are "broken." This requires that an

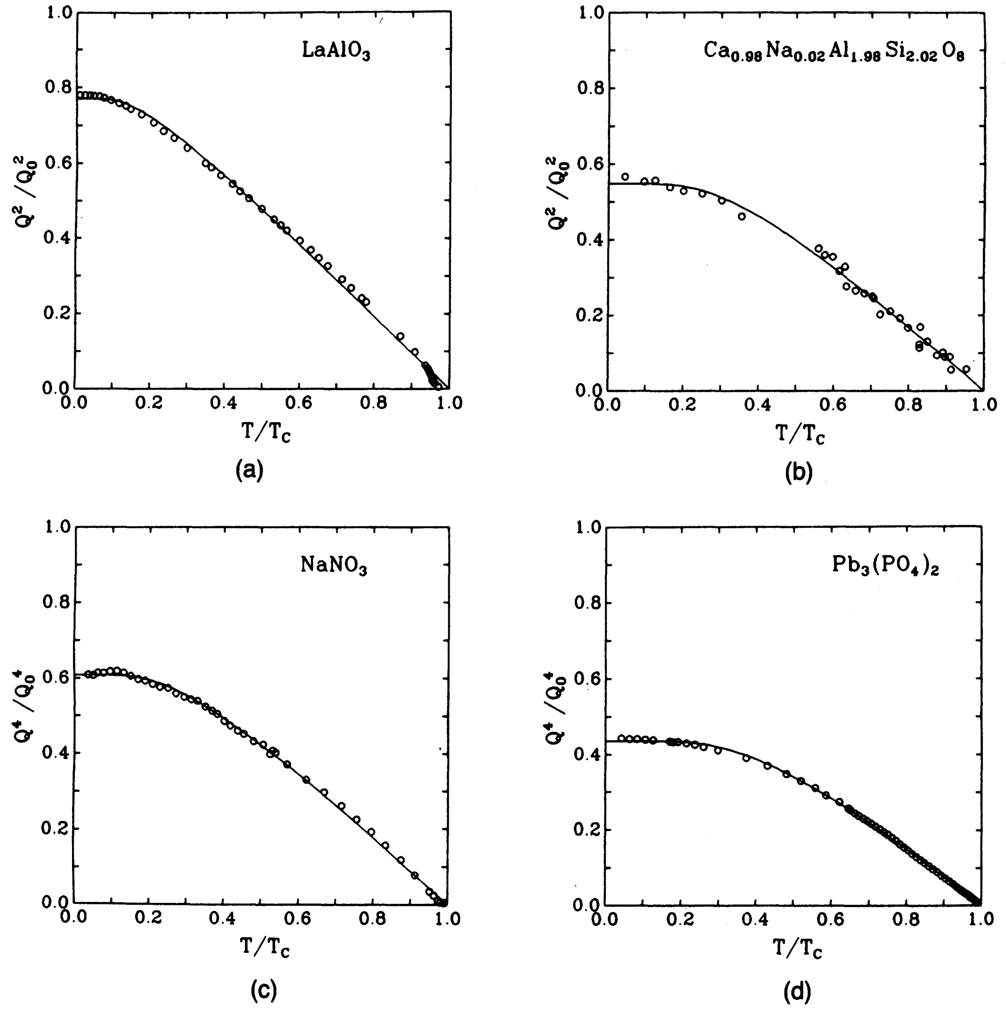

FIG. 4. Temperature evolution of the OP for four typical SPTs. In each case, the curves of a self-consistent solution [Eqs. (5)–(9)] are shown. The relevant parameters are, for (a) $LaAlO_3$, $T_c = 800$ k, $\theta_s = 95$ k, $\Delta = 0.116$; for (b) $Ca_{0.98}Na_{0.02}Al_{1.98}Si_{2.02}O_8$ (anorthite), $T_c = 513$ k, $\theta_s = 226$ k, $\Delta = 0.082$; for (c) $NaNO_3$, $T_c = 552$ k, $\theta_s = 165$ k, $\Delta = 0.174$; and for (d) $Pb_3(PO_4)_2$, $T_c = 455$ k, $\theta_s = 259$ k, $\Delta = 0.087$. In all cases $\Delta \ll 1$, so that the data points follow approximately the predictions of classic Landau theory.

additional parameter exists that specifies the thermodynamic state of the low-symmetry structure (which is already completely defined in the high-symmetry structure).

Let us illustrate this requirement using the example of quartz in Fig. 1. In the high-symmetry form, the space- and time-averaged orientation of the tetrahedra is fully determined by the symmetry of the crystal structure. In the low-symmetry structure, this is not the case because the symmetry is reduced so that small rotations of the tetrahedra are compatible with the symmetry constraints. In order to characterize the low-symmetry structure, we need to specify "how much" the tetrahedra rotated and "in which direction."

In a similar manner, the high-symmetry phase of the feldspar structure requires that Al and Si are randomly distributed over the relevant tetrahedra positions. In the low-symmetry structure, this distribution is no longer random with some structural sites being occupied preferentially by Al rather than Si. Again, one needs to quantify to what degree this site ordering has taken place.

As the basic process of breaking a symmetry constraint is always the same, it makes sense to introduce a thermodynamic quantity that describes the structural state of the low-symmetry phase in a universal manner. This thermodynamic parameter is called the order parameter Q. In order to connect Q with the thermodynamic properties of the material, we simply note that the Gibbs free energy G has to depend on $T, P, \ldots$ *and* Q in $G(T,P, \ldots, Q)$. With $Q = 0$ in the high-symmetry phase, we can then define $\Delta G = G - G(Q = 0)$ as the excess Gibbs free energy. Similarly, it is customary to define all other physical quantities as excess values with respect to the state defined by $Q = 0$. Once the dependence of ΔG on Q is either determined experimentally or anticipated by phenomenological models (such as in Sec. 1), the dependence of Q on $T, P \ldots$ simply follows from the equilibrium condition that G must have a minimum at the equilibrium value of Q, i.e., $dG/dQ = 0$, $d^2G/dQ^2 > 0$. All other physical parameters are then determined via their relationship with Q.

It is essential for the definition of an order parameter that a close structural relationship exists between the two phases. This allows us to describe a "difference structure" as the true crystal structure minus the high-symmetry structure as reference structure, i.e., we calculate the difference of atomic densities $\rho(\mathbf{r}) = \rho_{Q\neq 0}(\mathbf{r}) - \rho_{Q=0}(\mathbf{r})$. The pattern of this difference structure describes then directly the atomic displacements, changes in site occupancies, etc. The OP is then the amplitude of this reference structure and tells us "how much happens," whereas the pattern itself describes "what happens." The OP is usually normalized to unity for the maximum change and is zero when no change occurs, i.e., in the high-symmetry phase.

The required close structural relationship may not exist, however. If one crystal structure transforms into another without any topological relationship between them, the transition is called "reconstructive" and no order parameter besides a trivial normalized volume ratio of the two phases can be defined. If the two structures transform into each other with a well-defined three-dimensional topological relationship, the transition is called "topotactic." For two-dimensional relationships, the structural transition is called "epitactic." Several materials show layer-type structures. If the relative stacking of the layers changes during a phase transition, the transition is called "polytypic." The various structures that can be related by such polytypic phase transitions (hypothetical or real) are usually called "polytypes"; they are distinguished by the different stacking of identical (or almost so) slabs as the basic structural unit. The order parameter Q describes in this case the amount of material in a given stacking sequence.

Group-theoretical methods have been used to derive the symmetry properties of Q. The OP is a scalar quantity for phase transitions between structures with triclinic, monoclinic, or orthorhombic symmetry. For transitions from phases with higher symmetry, the order parameter can have up to three components $Q = \{Q_i, i \leq 3\}$ representing degeneracies of the symmetry constraints. Further degeneracies stem from the fact that the order parameter is a complex quantity $Q = Q_0 e^{-i\mathbf{q}\cdot\mathbf{r}}$, if the structural change is not uniform. In this case, the translational symmetry is broken with the excess wave vector $\mathbf{q}$ ($\mathbf{q} = 0$ if the translational symmetry is maintained). Degeneracies can occur if the relevant wave vector $\mathbf{q}$ describes several symmetry-equivalent transition mechanisms.

Similar symmetry constraints also exist for the terms in a polynomial expression of the excess Gibbs free energy. As this energy is invariant under all symmetry operations, only those terms Q^n or $Q_i^n Q_j^m$ are symmetry allowed that are invariant under all symmetry operations. The various possible combinations are listed in a great number of publications.

Let us consider a typical geometrical example. A structural unit such as an octahedron is embedded in a cubic crystal structure (Fig. 5). Now let us consider a movement of the structural unit such as a rotation, an elongation, etc. The movement is topologically identical if we interchange the three cubic axes, so that the crystal axes cannot depend on the choice of the Gibbs free energy. The OP must be threefold degenerate, therefore. The symmetry-allowed form of the Gibbs free energy includes then three components Q_1, Q_2, and Q_3 with

$$\begin{aligned} G = G_0 &+ \tfrac{1}{2}A'(Q_1^2 + Q_2^2 + Q_3^2) \\ &+ \tfrac{1}{4}B(Q_1^4 + Q_2^4 + Q_3^4) \\ &+ \tfrac{1}{4}B'(Q_1^2Q_2^2 + Q_2^2Q_3^2 + Q_1^2Q_3^2) + \cdots . \end{aligned} \tag{14}$$

For $B' = 0$, there is no correlation between

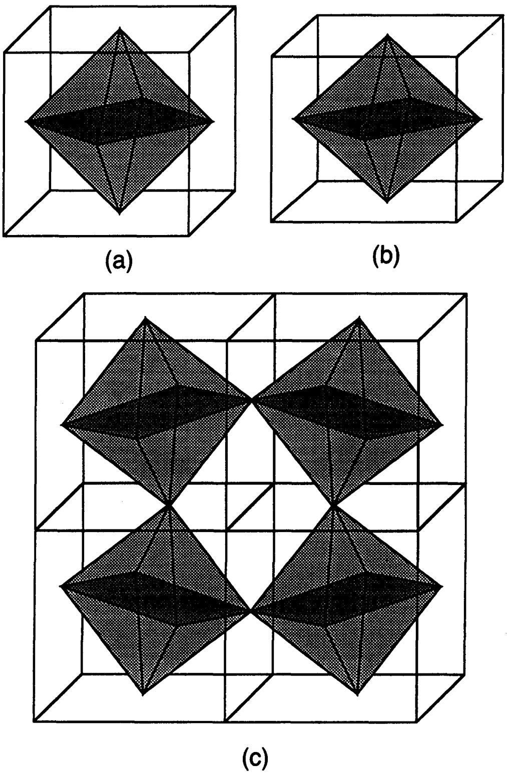

FIG. 5. Phase transitions generated by the distortion of (a) an octahedron in a cube-shaped cage. Two basic deformations can occur depending on the energy contribution of B'; (b) an elongation or shrinking along one of the cube axes (with three degenerate orientations) or along the cube diagonal (with four degenerate orientations). (c) If the translational symmetry is broken via rotation of the octahedra but not the cubes, tetragonal and trigonal structures are commonly formed.

Q_1, Q_2, and Q_3 with each component Q_i generating a phase transition. The term $Q_i^2Q_k^2$ couples the three components of the order parameter. With the condition $\delta G = 0$, $\delta^2 G > 0$ leads to only two types of solutions ($B' \neq 0$) (Slonczewski and Thomas, 1970):

$$\text{(a)}\ Q = (\pm Q,0,0), (0,\pm Q,0), (0,0,\pm Q), \tag{15}$$

$$\text{(b)}\ Q = \frac{1}{\sqrt{3}}(\pm Q,\pm Q,\pm Q). \tag{16}$$

The solution (a) represents a tetragonal phase with the fourfold axis along the cubic a, b, or c axis. The solution (b) describes a trigonal phase with the triad along the cubic [1 1 1], [$\bar{1}$ 1 1], [1 $\bar{1}$ 1], or [1 1 $\bar{1}$] axis. Such phase transitions between cubic and tetragonal or trigonal symmetries are often encountered, in particular in perovskite type crystal structures.

If the wave vector **q** of the order parameter Q is zero, the translational symmetry of both phases is equal. Such phase transitions are sometimes called "ferrodistortive." If **q** leads to a doubling of the translational unit, the transition is "antiferrodistortive." All simple multiplications of the size of the translational unit lead to a "commensurate" phase because the translational periodicity of the low-symmetry phase is commensurate with that of the high-symmetry phase. If this is not the case, the low-symmetry phase is called "incommensurate" (IC). Most IC phases exist only for small temperature intervals although notable exceptions occur (e.g., Na_2CO_3).

3. FERROELASTICITY AND FERROELECTRICITY

Atomic ordering may involve the reshuffling of atoms of different size or with different polyhedral environments. In such cases, the interchange of atoms will lead to a local distortion of the crystal structure. Local distortions lead to stresses that penetrate far into the structure and influence other atomic-ordering processes even when they happen at large distances from the original event. It is this long-ranging character of the elastic interactions that makes them so important. Experimental work has shown that in many SPTs a strong coupling between the OP and elastic degrees of freedom is observed (Salje, 1993b; Cowley, 1976; Folk *et al.*, 1979; Marais, 1991). As a consequence, the shape of the structural unit changes by typically 1%–6% with a smaller change of the specific volume. Even larger changes are seen in some metallic alloys. The excess deformation is called the "spontaneous strain." The generation of the spontaneous strain contributes significant elastic energy to the excess Gibbs free energy and leads to temperature dependences of elastic constants. In many crystals, a SPT was first discovered by the experimental observation of such anomalous thermal expansions using x-ray or neutron-diffraction methods. Materials with elastic anomalies are called coelastic. A specific case exists if the

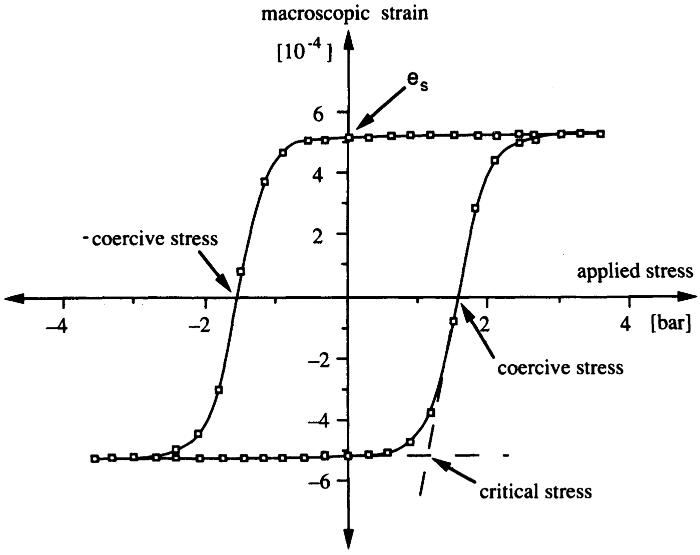

FIG. 6. Ferroelastic hysteresis for $Pb_3(PO_4)_2$.

spontaneous strain can be switched between symmetry-equivalent orientations (ferroelasticity). In Fig. 6, a ferroelastic hysteresis related to $Pb_3(PO_4)_3$ is shown for a macroscopic sample (Salje, 1993b). Mechanical shear stress of a monoclinic sample allows switching of the monoclinic axis by 120°, so that the macroscopic shape of the crystal changes accordingly. This switch is fully reversible.

Similar hysteresis phenomena exist for ferroelectric materials in which the reorientation of the order parameter involves the change of the spontaneous polarization (Lines and Glass, 1977). In Fig. 7, the spontaneous polarization in $BaTiO_3$ is shown as the dipole moment of the positively charged Ti and the negatively charged O with less contribution from Ba. The centers of gravity of the posi-

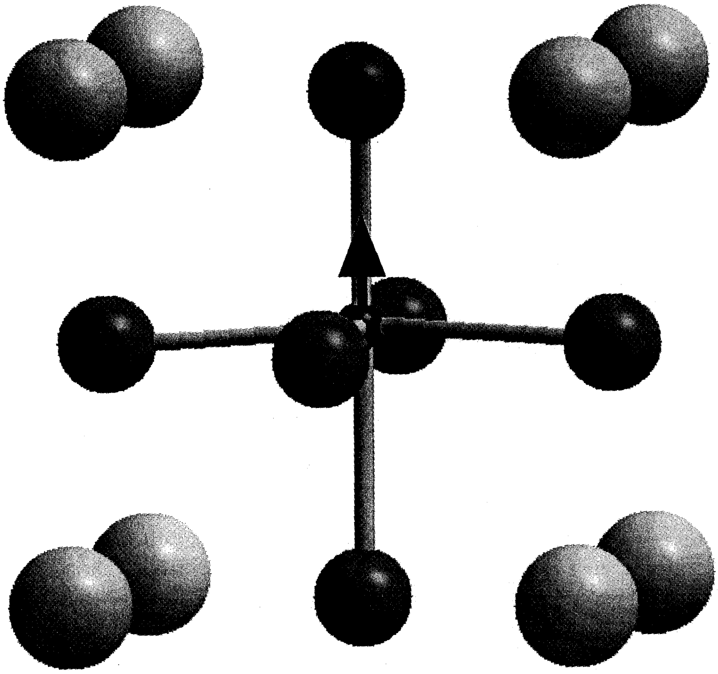

FIG. 7. Crystal structure of ferroelectric, tetragonal $BaTiO_3$. During the SPT from the paraelectric cubic phase, the black Ti cations are slightly shifted with respect to the darker gray oxygen anions. The charge separation generates a spontaneous dipole moment.

tive and negative changes are not identical, so that an electrical dipole moment builds up in the low-symmetry phase (i.e., the spontaneous polarization). Ferroelasticity and ferroelectricity sometimes occur simultaneously with comparable contributions of the electrical and mechanical energy to the excess Gibbs free energy (e.g., in KH_2PO_4).

4. COUPLING BETWEEN ORDER PARAMETERS

Several SPTs may occur in the same material if temperature or pressure is varied over large intervals. For example, almost all incommensurate transitions lead to a second, commensurate (lock-in) transition. Other examples include cascades of three or more transitions each of which breaks a different symmetry element of the same high-symmetry phase. If such elements are restored in a subsequent transition, the SPT is called "reentrant."

Even when the T and P conditions for the various SPTs are very different, the OPs may still interact. The total excess Gibbs free energy is then the sum of the individual contributions and the interaction energy G_{coupling}:

$$G = G_1 + G_2 + G_{\text{coupling}}.$$

Three types of coupling occur commonly whereby the dependence of the coupling energy on the OPs is different (Salje, 1993b):

1. If the coupled order parameters have the same symmetry (but, say, different transition points), $G_{\text{coupling}} = \lambda Q_1 Q_2$ is the coupling energy in lowest order of the OPs. The term λ is a measure for the coupling strength.
2. The coupling energy $\lambda Q_1^2 Q_2^2$ may occur under all symmetry conditions and is often considered as the "standard" coupling term.
3. Specific systems show $G_{\text{coupling}} = \lambda Q_1 Q_2^2$ or $\lambda Q_1^2 Q_2$, whereby this product of OPs must be invariant under the relevant symmetry operations.

The stability of phases has now to depend on both order parameters, which are commonly displayed in the order-parameter vector diagram (Fig. 8). Such diagrams contain all combinations of order parameters that represent thermodynamic equilibrium under the various T and P conditions. The origin is always $Q_1 = Q_2 = 0$, i.e., the high-symmetry phase. When T and P are changed, the structural states run along the trajectories in the diagram. Structural states outside the trajectories represent no minima of the Gibbs free energy and are, thus, not equilibrium states.

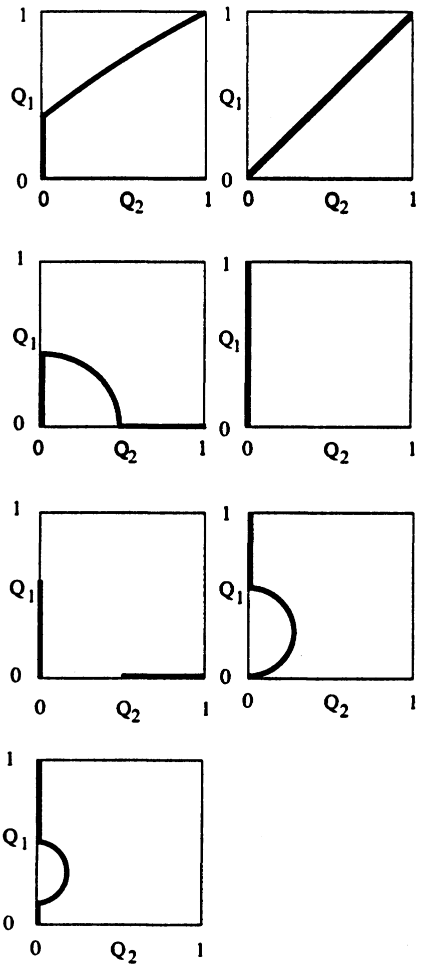

FIG. 8. The typical order parameter vector diagrams for two OPs coupled biquadratically ($\propto \lambda Q_1^2 Q_2^2$). An equilibrium structural state is always characterized by two OPs, which define a point in the diagram. All equilibrium points together form lines of equilibrium states. Changing T, P, etc. moves the equilibrium point along such equilibrium lines as shown in the figures. The high-symmetry phase is represented by the origin ($Q_1 = Q_2 = 0$). Additional phase transitions occur when one of the OPs vanishes [e.g., between the states ($Q_1 = 0$, $Q_2 \neq 0$) and ($Q_1 \neq 0$, $Q_2 \neq 0$)].

Bilinear coupling **1** leads to bending of the equilibrium line, which is the locus of all stable configurations. Topological changes occur for nonlinear coupling with the creation of additional phase transitions that stem from the competition between the OPs and are *not* due to an independent, third, transition mechanism. All quantitative analyses of SPTs with several OPs have to account for these coupling effects, which greatly modify the T and P dependences of the OPs. A typical example of the experimentally observed T dependence of two coupled OPs is shown in Fig. 9.

5. INFLUENCE OF DEFECTS

SPTs are influenced by structural defects. Defects often couple with thermal fluctuations. Defects can generate anomalies such as excess heat capacities, softening of elastic constants, and parasitic short-range order in the high-symmetry phase. These anomalies tend to mimic closely the behavior of critical systems with short-range interactions (such as magnetic phase transitions). It makes sense, therefore, to describe these anomalies with the same effective exponents, α for the specific heat, β for the order parameter, and γ for the susceptibility. Note that the observed anomalies are not intrinsic to the SPT but they are generated by the lattice relaxations around the defects. Common exponents for α are $\frac{1}{2}$ for unpolarized defects and $\frac{3}{2}$ for polarized defects (in contrast to $\alpha = 0$ for the defect-free SPT in the MFT approximation and as observed in many ferroelastic crystals). Other values exist depending on the details of the relaxation (Levanyuk and Sigov, 1988).

At temperatures small compared with T_c, the OP increases (reduces) when the defects stabilize (destabilize) the low-symmetry phase. The extrapolated transition temperature is accordingly changed. These changes are linear for small defect concentrations ($\Delta G_{\text{defect}} \propto \zeta N Q^2$ changes T_c to $T_c + N\zeta$). For larger defect concentrations N, strongly nonlinear effects lead to a "plateau" of the extrapolated T_c vs N curves.

If the defects stabilize the low-symmetry phase, they may destroy the high-symmetry phase altogether. A typical example is that a point defect leads to a local lattice deformation that has a common component with the spontaneous strain. Both contributions generate the total spontaneous strain, which will not disappear (because of the defect contribution) at $T > T_c$ thereby preventing the high-symmetry phase from appearing under heating. A common example is that the defect acts as fields H_d conjugated to the OP with $\Delta G_{\text{defect}} \propto -\lambda H_d Q$. In this case, Q decreases near the defect-free transition point but then forms a tail at $T > T_c$ (Fig. 10). The structural phase with $Q = 0$ is no longer a minimum of

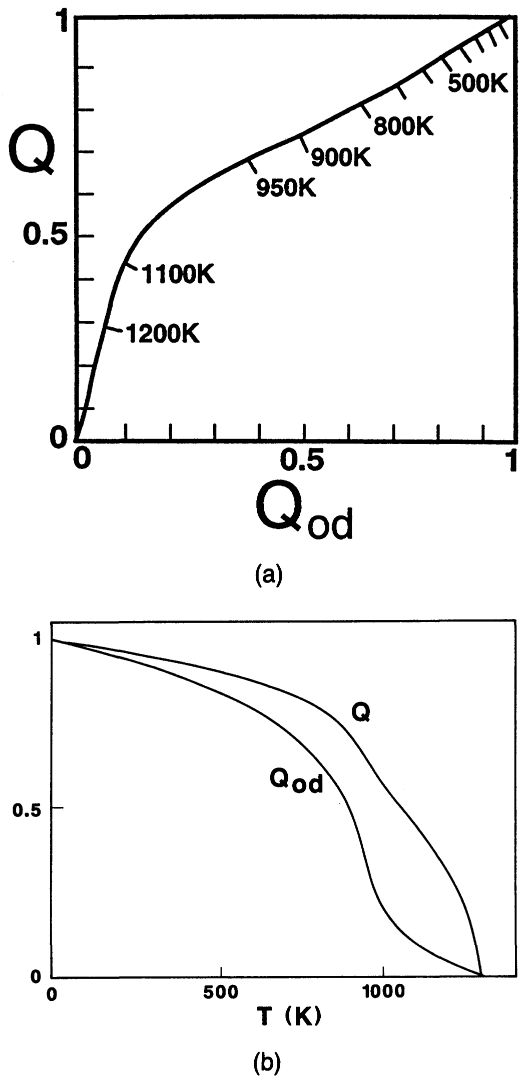

(a)

(b)

FIG. 9. Example for an experimentally determined order-parameter vector diagram of two coupled OPs (bilinear coupling $\propto \lambda Q_1 Q_2$) for Na feldspar in Fig. 2. Q_{od} represents the OP for the od process of ordering Si and Al on various lattice sites, Q represents an OP of a displacive SPT. (a) Vector diagram, (b) temperature evolution of both OPs [after Salje (1993b)].

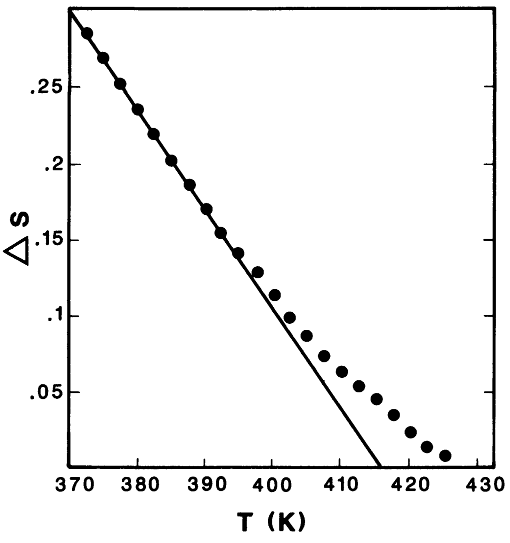

FIG. 10. Defect-related tail in displacive SPT in (Na,K) feldspar. The excess entropy ΔS is proportional to the square of the OP; the pure material has a SPT at 416 K. Defects generate a tail of the OP at higher temperatures.

the total Gibbs free energy (i.e., $\delta G = 0$ always leads to $Q \neq 0$). If the defect field is intrinsic to the crystal structure, this behavior is called a nonconvergent SPT, indicating that the high-symmetry form is asymptotically approached but never reached. Strictly speaking, a nonconvergent SPT is not a SPT because the actual transition does not take place, although the thermodynamic and structural features of nonconvergent behavior are appropriately described within the same theoretical framework as that of SPTs (including the intrinsic field term).

6. TWINS AND ANTIPHASE BOUNDARIES AS FINGERPRINTS OF SPTs; SURFACE RELAXATIONS

The generation of the OP may allow the crystal to become heterogeneous with different values of the OP in different domains. In ferrodistortive SPTs if two domains have OPs with opposite sign, the domains are called twins with a twin boundary between them. Such OPs usually carry large spontaneous strain e_{ik}. The locus of the boundary X_i is determined by the compatibility between the two twins S_1 and S_2, i.e.,

$$[e_{ik}(S_1) - e_{ik}(S_2)]X_iX_k = 0, \tag{17}$$

leading for traceless e_{ik} to two types of walls. These two types of walls would be strictly orthogonal in the metric of the high-symmetry phase. In the low-symmetry phase, the angle between the walls is $\pi/2 \pm \omega$ where the small angle ω is numerically identical to the relevant tensor component of the spontaneous strain ($\omega \lesssim 10^{-2}$). In many experimental observations, this angle ω is too small to be seen and it is customary to describe the wall pattern with the two types of wall as "right-angle domain walls." Walls that are fully localized by symmetry constraints (e.g., those along mirror planes of the high-symmetry phase) are called W walls. Those which have orientational degrees of freedom are called W' walls. W' walls occur, e.g., when the only symmetry constraint is that a twofold axis of the high-symmetry phase has to be contained in the wall. The wall can then rotate around the axis in order to minimize its energy. The rotation angle will then depend explicitly on the control parameter (e.g., T) (Salje, 1993b).

The internal structure of twin walls is usually derived from the Euler–Lagrange equation with $G = G(Q,T) + \frac{1}{2}g(\nabla Q)^2$:

$$\frac{dQ}{dt} \propto \frac{\partial G}{\partial Q} + g\frac{d^2Q}{dx^2}, \tag{18}$$

where x is the distance from the twin wall. The gradient energy describes the energy needed to modify the OP spatially. The solutions depend on the explicit form of $G(Q,T)$; in most cases, they are identical with or close to the solution for the simplest Landau potential with

$$Q = Q_0 \tanh(x/w), \tag{19}$$

where Q_0 is the equilibrium OP and w is the wall thickness. The term w increases close to the transition point as

$$w = \sqrt{\frac{2g}{A|T - T_c|}}, \tag{20}$$

where $\sqrt{2g/AT_c}$ is in the order of lattice repetition units.

The internal structure of twin walls is

modified when several order parameters couple (Houchmandzadeh *et al.*, 1991). Instead of having a rather smooth bend of the OP from one equilibrium value on one side of the wall to another equilibrium value on the other side of the wall, new structural states may occur inside the wall. These new states are described by the second order parameter. The true crystal structure inside the wall and close to the wall is then characterized by the space dependence of the two OPs. Within the spirit of order-parameter vector diagrams introduced in Fig. 8, we now depict solutions of the Euler–Lagrange equation for two OPs as a graph of one OP vs the second OP and obtain the "trajectory" of the domain wall in the order-parameter space. Two typical trajectories for walls with biquadratic coupling between two OPs are shown in Figs. 11(a) and 11(b). The first OP is Q_1 and shows a modified hyperbolic tangent profile. The second OP, Q_2, is zero in the uniform state [Fig. 12(a)] and is generated only inside the wall. In the spatial evolution of the order-parameter vector diagram [Fig. 12], the effective OP "rotates" in the wall, so that this wall is called chiral (similar to magnetic Bloch walls). Parts of walls with right-handed and left-handed chirality meet in lines of vanishing chirality. The second solution in Figs. 11(b) and 12(b) shows a right-handed chirality on one side of the wall and a left-handed chirality on the other side. Such walls are called bichiral.

Antiferrodistortive phase transitions can lead to the formation of antiphase boundaries (APBs) in which the translational symmetry is broken. In cases with weak strain coupling, APBs wiggle through the crystal without orientational preference. Increasing strain coupling leads to preferential orientations and lock-in phenomena along the optimum direction. The thickness of APBs is of the order of interatomic distances.

Structural relaxations close to surfaces and interfaces are enhanced by SPTs. In case of interfaces, such relaxations appear as ripples of the OP and decrease exponentially with distance from the interface. The characteristic length increases close to the transition point. Surfaces show additional relaxations in the high- and low-symmetry phases. Such relaxations act as precursor phenomena although the relaxational modes are not necessarily identical with the deformation pattern of the actual phase transitions. Theoretical work predicts a dominance of incommensurate relaxations and quasiperiodic pattern formation even when the phase transition is (anti)ferro-distortive.

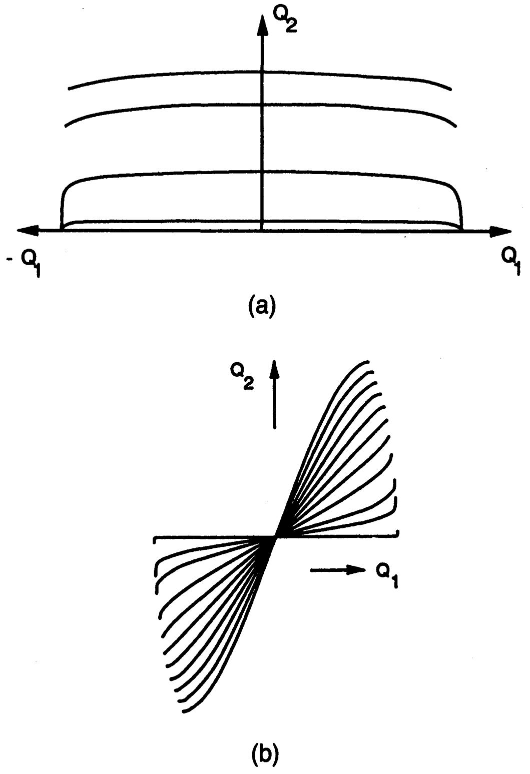

FIG. 11. Vector diagrams of biquadratically coupled OPs together with trajectories of the domain boundaries. Each line represents a trajectory with the stable OPs of the bulk at either end of the trajectory. In (a), only Q_1 changes sign in the boundary but not Q_2. Trajectories in (b) show Q_1 and Q_2 changing sign in the domain boundary.

7. KINETICS OF SITE ORDERING

A system out of equilibrium (i.e., $\partial G/\partial Q \neq 0$) will tend to lower its energy via kinetic processes. In displacive SPTs, equilibrium is reached via phonon processes so that the time scale is fast. Much slower processes occur in od systems, which involve the activation of site exchanges that is necessary to interchange atomic sites. Two principal processes dominate od SPTs. Firstly, local site ordering lowers the energy so that local state parameters equivalent to OPs can be used to describe the kinetic rate equation. If only two

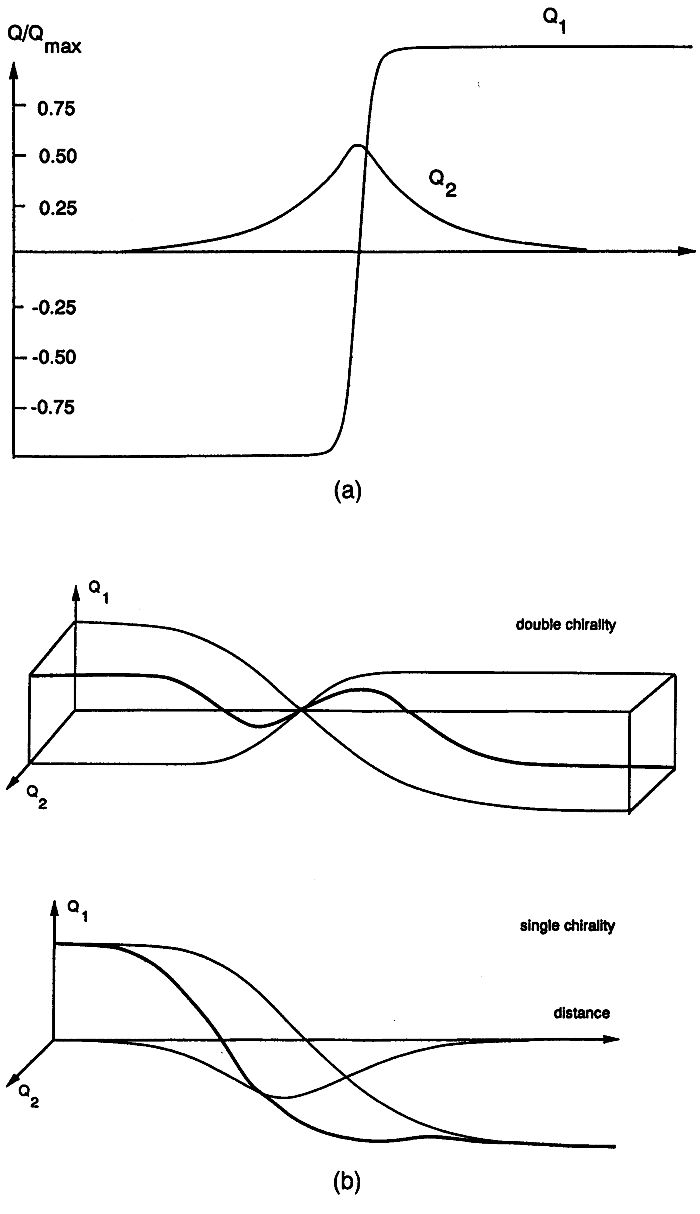

(a)

(b)

FIG. 12. Typical wall profiles for both types of trajectories shown in Fig. 11. (a) Q_1 changing sign but not Q_2 whereby the case $Q_2 = 0$ in the bulk is shown. Q_2 exists only in the wall indicating a SPT created in a domain boundary but not in the bulk. (b) The two types of wall profiles with both OPs changing sign (top) and only one changing sign (bottom). The latter is similar to a chiral Bloch wall in magnetism whereas the former self-compensates the chirality in opposite sign on either side of the wall. The abscissa is the distance from the center of the wall; typical wall thicknesses are measured in tens of ångstroms.

local states exist (e.g., like in spin systems), the rate is determined by the random choice of sites and subsequent state changes according to Boltzmann statistics [the so-called "Glauber" kinetics, often in the MF limit because of long-ranging correlations (Glauber, 1963)]. In more complex framework structures, many local exchange schemes exist and compete for the kinetic mechanism. Several structural states intermediate between the nonequilibrium starting configuration and the final (equilibrium) configurations have energy states between the initial and final states. These intermediate states can then serve as "stepping stones" for the system progressing kinetically toward equilibrium. If a sufficient number of such intermediate states exist, the kinetic rate law can be described in the continuum approximation (and in MFT due to long-ranging interactions) by the generalized time-dependent Ginzburg–Landau equation $dQ/dt \propto \partial G/\partial Q$, where G is the equilibrium Gibbs free energy.

The second process involves the transport of atoms (e.g., via interstitial sites). This part of the ordering requires chemical neutrality and is "conserved" so that each local change of state is compensated by an equivalent change in the opposite direction somewhere else in the structure. This process dominates exsolution kinetics and follows for two-state systems the so-called "Kawasaki" rate law, i.e., now pairs of sites are chosen for an exchange of state variables according to Boltzmann statistics (Kawasaki, 1966; Langer, 1971). This kinetic mechanism is in contrast to Glauber statistics where each site is chosen individually so that it is only in the Kawasaki case that the global order parameter remains unchanged. In SPTs, both processes may exist simultaneously and a general rate equation for continuous state parameters can be formulated (Salje, 1993a):

$$\frac{dQ}{dt} \propto \left(1 - \frac{\xi_c^2}{\xi^2}\frac{\sinh(\xi\nabla)}{\xi\nabla}\right)\frac{dG}{dQ}, \tag{21}$$

where ξ is the Kawasaki length and ξ_c scales the amount of conservation in the individual rate process: With $\xi_c = 0$ we recover the time-dependent Ginzburg–Landau equation, and $\xi_c = 1$ defines in lowest order the so called "Cahn–Hilliard rate equation" (Langer, 1971; Amengual *et al.*, 1993; Salje, 1993a).

These rate laws have been used successfully to describe the kinetics of atomic ordering in framework structures and exsolution processes in alloys and other materials. The dependences of G and the kinetic operator $L = 1 - (\xi_c^2/\xi^2)\sinh(\xi\nabla)/\xi\nabla$ on the gradient operator ∇ lead to heterogeneous patterns for the OP. The exact form of these patterns and their time dependence (e.g., coarsening of domain structures) is still a matter of debate. Experimental observations often show that exsolution patterns ($\xi_c = 1$) show stripe patterns of (approximately) parallel domains with strong variations of the OP at the boundary between two domains. The internal structure of the boundaries appears to be similar to twin boundaries in ferroelastic crystals. SPTs involving atomic ordering show characteristic sequences of mesoscopic domain structures. The early stages contain tweed patterns consisting of patches of ordered material with different OPs (e.g., opposite signs in different patches) and wall-like boundaries between them. In the course of kinetic ripening of such microstructures, the tweed transforms into stripe patterns with well-developed domain walls. These stripe patterns then coarsen via an increase of the domain width (Plate 1).

Kinetic development of such mesoscopic microstructures with characteristic lengths of some 100 Å has been simulated by computer experiments (Parlinski *et al.*, 1993; Semenovskaya and Khatchaturyan, 1992). In most cases, the effective interaction between the sites is assumed to be elastic; the entropic part is determined by pseudospin variables at the sites, which describe the occupancy by one or the other atomic species. The standard Hamiltonian of such a system is (Marais *et al.*, 1991)

$$H_{\text{eff}} = \tfrac{1}{2}\mathbf{uAu} - \tfrac{1}{2}\mathbf{uFs}, \tag{22}$$

where $\mathbf{u}$ is the displacement vector, $\mathbf{A}$ is the dynamical matrix, $\mathbf{s}$ is the pseudospin variable, and $\mathbf{F}$ describes the spin-lattice coupling. This Hamiltonian can be renormalized to spin coordinates at sites R and R':

$$H_{\text{eff}} = -\sum_{\mathbf{RR}'} J(\mathbf{R} - \mathbf{R}')S(\mathbf{R})S(\mathbf{R}'). \tag{23}$$

The spin-spin coupling is now long-ranging with a strong directional dependence and nonanalytical behavior in the long-wavelength limit:

$$J(R) \propto \frac{1}{R^3}\,[A_2Y_{2m}(\theta,\varphi) + A_4Y_{4m}(\theta,\varphi)] + \frac{Z}{N} \tag{24}$$

with the constant term Z/N where N is the number of particles and Z is the Zehner–Eshelby constant describing the macroscopic relaxation of the sample. Using this Hamiltonian in computer simulation, the characteristic domain structures of tweed and stripes have been reproduced. Such mesoscopic structures are shown in Plate I for a simulation of 64×64×8 unit cells with thermodynamic conditions similar to those for the experimental results in Fig. 13.

8. CONCLUSION

Structural phase transitions are dominated by long-ranging interactions, which makes their description by mean field theories possible. Much work has been domi-

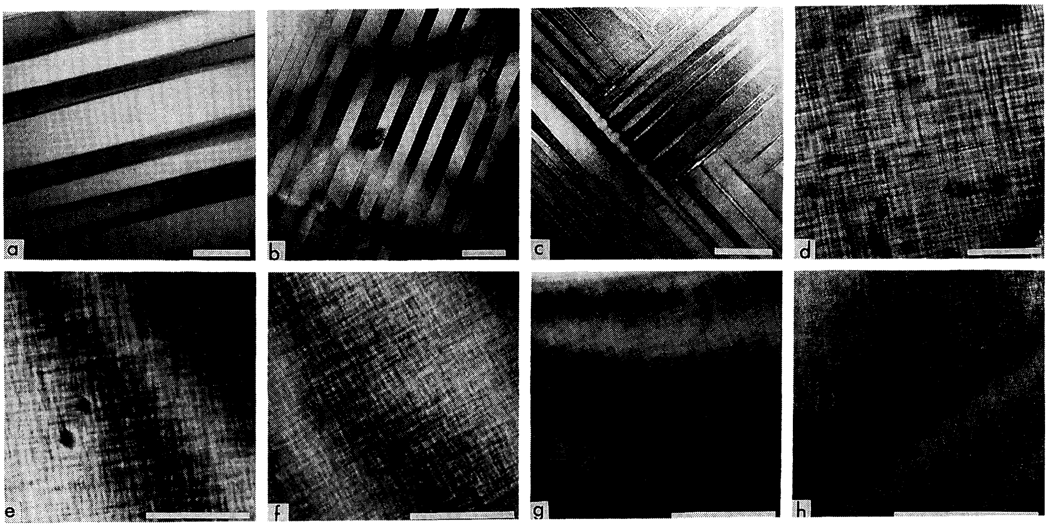

FIG. 13. Characteristic pattern for nonequilibrium SPTs. Quenched samples of Co-doped $YBa_2Cu_3O_7$ show the typical stripe (a)–(c) and tweed (d)–(h) pattern. Co-concentrations are (a) 0%, (b) 1%, (c) 2%, (d) 2.5%, (e) 2.8%, (f) 3%, (g) 5%, and (h) 7%. The scale bar is 0.1 μm. Note the formation of junctions between orthogonal stripes in (c). Courtesy A. Putnis, Cambridge.

nated by the application of Landau theory, including the coupling between several order parameters and their saturation at low temperatures. Future work is likely to concentrate on the time dependence (kinetics) of structural phase transitions.

GLOSSARY

Excess Gibbs Free Energy *ΔG*: Difference between the Gibbs free energies of the low-symmetry phase $G(Q)$ and the high-symmetry phase $G(Q = 0)$, also called the excess Gibbs free energy of the SPT.

Order Parameter (*Q*): Variable that describes the structural state, with $Q = 0$ for the high-symmetry phase and $Q \neq 0$ for the low-symmetry phase.

Phase Transition Temperature: (T_c for continuous phase transitions, T_{trans} for stepwise transition.) Temperature at which the Gibbs free energy of the high-symmetry form is identical to that of the low-symmetry form.

Saturation Temperature (θ_s): Temperature below which the temperature dependence of the order parameter diminishes rapidly.

Spontaneous Strain (e_{ik}): Strain of the crystal generated by the phase transition. The spontaneous strain is measured as change of lattice parameters with respect to the high-symmetry phase.

Works Cited

Amengual, A., Hernandez Garcia, E., San Miguel, M. (1993), "Ordering and Finite Size Effects in the Dynamics of One-Dimensional Transient Pattern," *Phys. Rev. E* **47**, 4151–4160.

Bruce, A. D., Cowley, R. A. (1981), *Structural Phase Transitions*, London: Taylor and Francis.

Cowley, R. A. (1976), "Acoustic Phonon Instabilities and Structural Phase Transitions," *Phys. Rev. B* **13**, 4877–4885.

Folk, R., Iro, H., Schwabl, F. (1979), "Critical dynamics of Elastic Phase Transitions," *Phys. Rev. B* **20**, 1229–1244.

Glauber, R. J. (1963), "Time-Dependent Statistics of the Ising Model," *J. Math. Phys.* **4**, 294–307.

Houchmandzadeh, B., Lajzerowicz, J., Salje, E. (1991), "Order Parameter Coupling and Chirality of Domain Walls," *J. Phys. Condensed Matter* **3**, 5163–5169.

Kawasaki, K. (1966), "Diffusion Constants Near Critical Points for Time Dependent Ising Models," *Phys. Rev.* **145**, 224–230.

Khatchaturyan, A. G. (1983), *Theory of Structural Transformations in Solids*, New York: Wiley.

Landau, L. D., Lifshitz, E. M. (1980), *Statistical Physics*, Oxford: Pergamon Press.

Langer, J. S. (1971), "Theory of Spinodal Decomposition in Alloys," *Ann. Phys. (New York)* **65**, 53–86.

Levanyuk, A. P., Sigov, A. S. (1988), *Defects and Structural Phase Transitions*. New York: Gordon and Breach.

Lines, M. E., Glass, A. M. (1977), *Principles and Applications of Ferroelectrics and Related Materials*, Oxford: Clarendon Press.

Marais, S., Heine, V., Nex, C., Salje, E. (1991) "Phenomena Due to Strain Coupling in Phase Transitions," *Phys. Rev. Lett.* **66**, 2480–2483.

Parlinski, K., Heine, V., Salje, E. (1993), "Origin of Tweed Texture in the Simulation of a Cuprate Superconductor," *J. Phys. Condensed Matter* **5**, 497–518.

Salje, E., Wruck, B., Thomas, H. (1991), "Order Parameter Saturation and a Low-Temperature Extension of Landau Theory," *Z. Phys. B Condensed Matter* **82**, 399–404.

Salje, E. (1993a), "On the Kinetics of Partially Conserved Order Parameters: A Possible Mechanism for Pattern Formation," *J. Phys. Condensed Matter* **5**, 4775–4784.

Salje, E. K. H. (1993b), *Phase Transitions in Ferroelastic and Co-elastic Crystals*, Cambridge: Cambridge Univ. Press.

Semenovskaya, S., Khatchaturyan, A. G. (1992), "Structural Transformations in Nonstoichiometric $YBa_2Cu_3O_{6+\delta}$," *Phys. Rev. B* **46**, 6511–6534.

Slonczewski, J. C., Thomas, H. (1970), "Interaction of Elastic Strain with Structural Transition of Strontium Titanite," *Phys. Rev. B* **1**, 3599–3608.

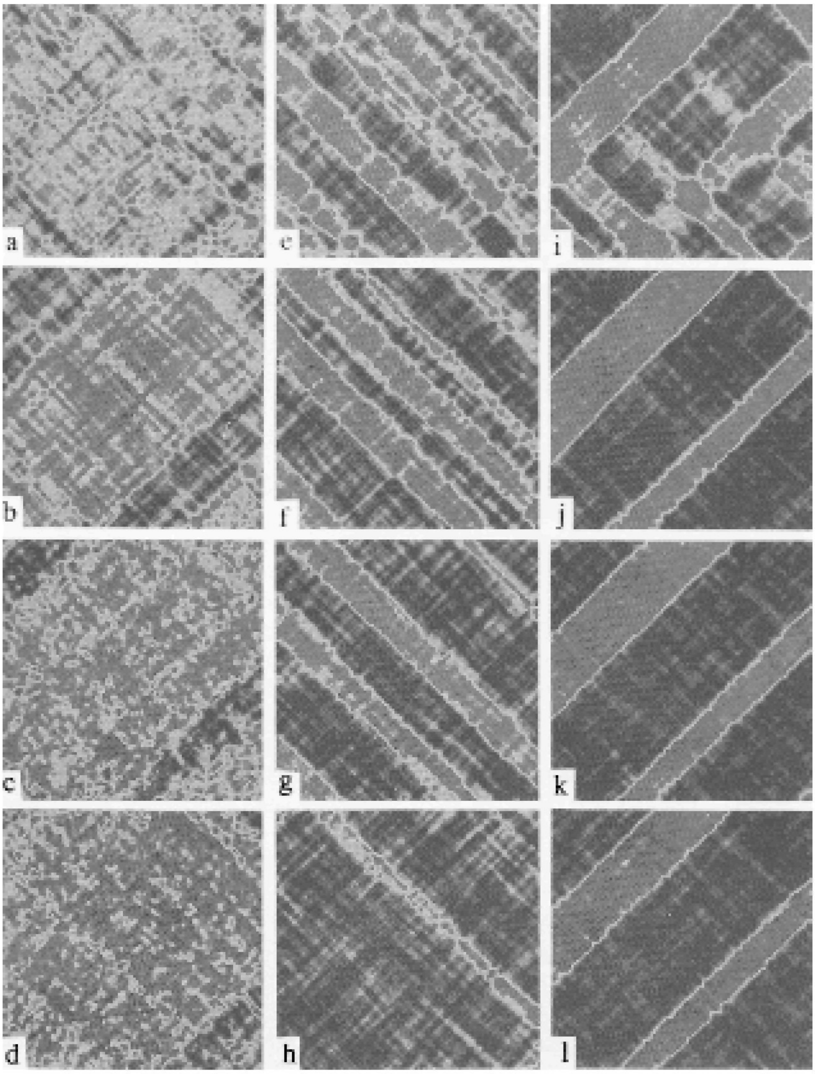

Vol. 13 – PLATE 1. Results of computer calculations of domain structures of a quench through a SPT. The system has been equilibrated at 1.75 T_c; the annealing temperatures are (a)–(d) 0.93 T_c, (e)–(h) 0.83 T_c, and (i)–(l) 0.73 T_c.

PHILOSOPHY OF PHYSICS

MAX JAMMER, *Bar-Ilan University, Ramat-Gan, Israel*

	Introduction	389
1.	**The Philosophy of Preclassical Physics**	390
1.1	The Early Development of the Philosophy of Physics	390
1.2	Philosophy and Experimentation in Preclassical Physics	391
2.	**The Philosophy of Classical Physics**	392
2.1	Philosophical Problems of Classical Mechanics	392
2.2	The Philosophy of Space and Time in Classical Physics	393
2.2.1	The Dimensionality Problem of Space	394
2.2.2	The Directionality Problem of Time	396
3.	**Antimechanistic Theories**	397
3.1	Dynamism	397
3.2	The Electromagnetic World Picture	398
3.3	Energetism	399
4.	**The Philosophy of Modern Physics**	400
4.1	The Philosophical Background of Modern Physics	400
4.2	Philosophical Problems of the Special Theory of Relativity	400
4.3	Philosophical Problems of the General Theory of Relativity	403
4.4	Philosophical Problems of Quantum Mechanics	407
	Works Cited	414

INTRODUCTION

Philosophy of physics, as currently understood, studies the conceptual foundations and philosophical implications of physics, and examines the interdependence between observation and theory as well as the structure and interrelation of physical theories. It also deals with issues traditionally classified as "metaphysical," such as the nature of space, time, matter, and energy. Last, but not least, it discusses the question of whether physical theories lead to, or are compatible with, a coherent and intellectually satisfying picture of the world.

Philosophy of physics differs from physics with respect to both the objectives aimed at and the techniques or tools employed. The objective of physics is the construction of a physical theory of nature by discovering new data, embedding them in a consistent scheme, and testing the consequences of the theory by experimental methods. The objective of the philosophy of physics is the study of the nature of a physical theory by examining its evidential and inferential foundations, its logical consistency, and its metaphysical assumptions and implications. The work of the philosopher of physics is therefore usually, in contrast to that of the physicist, *ex post facto*. The tools of the physicist are the laboratory and the computer, the tool of the philosopher of physics is the method of conceptual analysis.

This contradistinction does not, however, imply that a great physicist could not be also a philosopher of physics—a combination that, as we shall see, has been rather the rule than the exception, especially at periods of revolutionary innovations.

The historical development of physics can be divided into three major periods:

1. preclassical physics, which began in the sixth century B.C. with the natural philosophy of the school of Miletus (Thales and his disciples) and lasted until the scientific revolution of the 17th century;
2. classical or Newtonian physics, which came to an end at about 1900; and
3. modern physics, which began with the advent of the special theory of relativity and of quantum mechanics.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Because of the perpetual interaction between physics and its philosophy, it will prove useful to adhere to this historical division also in our discussion of the philosophy of physics.

1. THE PHILOSOPHY OF PRECLASSICAL PHYSICS

1.1 The Early Development of the Philosophy of Physics

The first to use the expression "philosophy of physics" as a *terminus technicus* was Aristotle. He used this term, in his treatise *Metaphysics* (Book 6, Chap. 1, 1026 a 18)*, in order to distinguish its subject matter from those of other branches of philosophy and science. In fact, at his time philosophy consisted already of several subdivisions, ranging from metaphysics to ethics and aesthetics. But prior to his time, when philosophy was born, when Thales and his school, the φυσικοι in Aristotle's terminology, emancipated themselves from mythological beliefs, rational reflexion was almost exclusively a philosophy of nature or φυσις, the Greek term for nature. These "physikoi" tried to explain the multifarious phenomena of nature by conceiving them merely as modifications of one and the same substratum or as manifestations of a single all-embracing principle. Their quest for a universally valid explanatory principle has ever since become one of the guidelines of physical research. The search for a unified theory of physics in our days is but a modern version of this quest.

In the fifth century B.C., the Pythagoreans claimed to have found this principle in the relation between numbers or in the structure of geometrical patterns. On the other hand, Empedocles, convinced that qualitative differences cannot be accounted for by only one principle, posited four principles, which he identified with the four elements earth, water, air, and fire; in addition, he introduced the two principles of Love and Strife, anticipating thereby the dynamics of attractive and repulsive forces in physics. Atomism, inaugurated by Leucippus and Democritus, conceived the universe to consist only of atoms and the void; devoid of qualitative properties except impermeability, atoms may differ in size, shape, arrangement, and position, and are in perpetual motion through the void. Although rejected by Aristotle and anathematized in the Middle Ages, atomism exerted a profound influence on the development of physics. True, the "atom" of Democritus is not that of Boyle, Dalton, or Rutherford, or the atom as conceived in the modern theory of elementary particles. But some of the fundamental tenets of ancient atomism persisted almost unchanged throughout the ages. Leucippus' dictum "Nothing comes into being at random, but everything comes from reason and by necessity" was the earliest enunciation of the principles of causality and determinism that dominated classical physics. Democritus' statement "Things appear to be colored, sweet or bitter, but in reality there are only atoms and the void" anticipated the distinction between primary and secondary qualities; the former, like solidity, extension, or motion, are properties inherent in bodies and inseparable from them, while the latter, like color, sound, or taste, are "nothing in the objects themselves" (Locke, 1690), but merely sensations in our consciousness. This distinction led Galileo, Descartes, Hobbes, and Locke to confine the task of physics to the study of only quantitative relations. It also raised, however, the philosophical question of whether macroscopic entities, like tables or trees, have an existence on their own, not only as aggregates of atoms—that is, the problem of whether the composite is more than the sum of its components and has a reality of its own.

An answer to this problem was given by Plato, who declared that the objects perceived by our senses are but imperfect copies or crude imitations of their ideas or "forms," which exist in a supersensible world beyond space and time and which are accessible to us only by pure thought. Influenced by Pythagoras, Plato contended that true knowledge of physical objects can be obtained only by mathematical studies of the ideas or forms of the objects. Consistent with his philosophy, Plato proposed in his influential dialogue *Timaeus* a purely geometrical theory of the structure of atoms as regular polyhedra.

Aristotle rejected the Platonic transcendentality of forms and argued that the form or idea of an object inheres in its substance: Every object is a union of matter and form; matter endows it with its individuality, form makes it a member of a class of similar objects. On this metaphysical assumption, he based his inductive-deductive method, according to which physical research has to progress from sense experience and to derive by induction generalizations that serve as premises for the deduction and explanation of observed phenomena. To avoid an infinite regress of explanations, self-evident premises have to be posited without demonstration. In accordance with the rule that "demonstrations must be based on premises prior to and better known than the conclusions," as stated in his *Posterior Analytics* (Book 1, Chap. 3, 72 b 25), Aristotle laid the foundations of the axiomatic approach to the formulation of physical theories and developed a rather elaborate proof-theory, stating, e.g., that "that proof is the better which proceeds from the fewer postulates" (*ibid.*, Book 1, Chap. 25, 86 a 33). His invention of the axiomatic method had an incisive influence not only on mathematics, where it was first applied to geometry by Euclid in his *Elements*, but also on the development of physics. Archimedes, for example, adopted the axiomatic method in his presentation of the science of statics when he began his treatise *On the Equilibrium of Planes* with the postulate that "equal weights balance at equal distances," implying thereby that a symmetrically loaded lever is in equilibrium.

---

*All references to works of Aristotle use what is known as the Bekker pagination (Bekker, 1831–1870).

Already Aristotle himself axiomatized, though not yet in a syntactical formalism, his theory of enforced motion by postulating that "any object that is in motion must be moved by some agent" (256 a 3) and that the mover must be in direct contact with the moved. This statement requires some explanation. In his treatise *Τα Φυσικα* (the Greek for "nature," from which the discipline of "physics" derived its name), Aristotle defined "physics" as the science of things that move or change, and he distinguished enforced from natural motion. The motion of a vehicle, when drawn by a horse, or of a stone, when thrown by a hand, is an enforced motion, while the motion of an apple falling off a tree is a natural one. The Aristotelian theory of natural motion is intimately related to the doctrine of the natural places of the elements, which doctrine in turn is the basis of Aristotle's cosmological world picture. For he accepted the traditional theory of the four elements earth, water, air, and fire, the natural places of which determine the sublunary spheres with Earth at the center; the celestial spheres that enclose the finite universe consist of a fifth element, the incorruptible ether. Now, the natural motion of an object in the sublunary world is but the quest of the element of which it consists to return by the shortest way to its natural place; and the natural motion of celestial matter is the eternal and self-contained circular motion. This ontological and nomological differentiation between a physics of the heavens and a physics of the Earth complied later very well with the teachings of medieval theology and contributed, therefore, to the survival of Aristotelian physics throughout the Middle Ages.

1.2 Philosophy and Experimentation in Preclassical Physics

Ancient and medieval physicists based their science upon observation and speculation and not, as a rule, on experimentation. Sociologists explain this fact by pointing out that in those times the institution of slavery provided cheap labor and thus retarded the development of energy-saving technologies and hence also the construction of scientific instruments. Another no less important reason was undoubtedly the thesis of the ancient philosophy of physics that nature is endowed with a life of its own (hylozoism) and that things can behave naturally only within their natural habitat. An experiment, in contrast to passive observation, was regarded as an interference with the course of nature or even as an act of violence. In fact, simple devices, such as the lever, the wheel and axle, or the pulley, were not called "physical" instruments but rather "mechanical" contrivances, for the Greek term μεχανη, the root of the words "machine" and "mechanics," meant "trick" or "deceit"; by using these appliances, one supposedly bilked or cheated nature to behave contrary to itself. "When we have to produce an effect contrary to nature . . .," says the pseudo-Aristotelian treatise *Mechanical Problems*, "we call it mechanical" (*Mechanica*, 847 a 10). In short, since the performance of an experiment imposes artificial conditions, "contrary to nature," an experiment, it was argued, could never increase our knowledge about nature. This idea, though slightly modified, reappeared in modern physics, where according to certain interpretations of quantum mechanics the Heisenberg indeterminacies result from disturbances caused by the measuring process.

As a result of the invention of mechanical devices that, like the waterwheel, were widely applied to spare human effort, these arguments began to lose much of their cogency already in the early Middle Ages. During the Renaissance, the development of mechanical craftmanship in architecture, engineering, and the fine arts led gradually to an appreciation of experimentation, as illustrated by Leonardo da Vinci's experiments with small-scale models to solve architectural problems. Finally, though not yet for the purpose of discovering new physical effects or laws, Galileo Galilei established the method of controlled experimentation in order to test mathematically derived conclusions. Since, as Galileo (1623) put it, "The book of nature is written in mathematical language" and since an experiment *qua* quantitative test is likewise described in mathematical terms, the experiment was no longer regarded as an extraneous interference with nature.

With his experimental study of free fall, Galileo refuted the teleological conception of "natural motion" and replaced it by the purely descriptive notion of a uniformly accelerated motion satisfying the famous formula $s = \frac{1}{2}gt^2$, the earliest physical equation to contain "time" as the independent variable. Galileo also introduced the use of thought experiments, as, e.g., in his refutation of Aristotle's thesis that heavy bodies fall faster than light ones, by simply conceiving a heavy body as a combination of light components and showing that the Aristotelian thesis thus leads to a self-contradiction.

2. THE PHILOSOPHY OF CLASSICAL PHYSICS

2.1 Philosophical Problems of Classical Mechanics

The most devastating attack against Aristotelian physics was, of course, the announcement of the principle of inertia, the very antithesis of Aristotle's first axiom of motion. Historians differ on who was the first to announce the law of inertia, whether it was Galileo who applied it only to horizontal motions along the surface of the Earth, whether it was the Dutch physicist Isaac Beeckman who derived it from the properties of space, or whether it was Descartes who deduced it from the conservation of linear momentum (Jammer, 1990). In his *Philosophiae Naturalis Principia Mathematica*, which has been called "the most important work ever published in the physical sciences" (S. W. Hawking), Isaac Newton (1687) formulated the principle of inertia as the first law of motion as follows: "Every body continues in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed upon it." Although Newton's first law has ever since been regarded as a cornerstone of classical physics, its philosophical status is still a matter of dispute. Opinions differ on whether it should be regarded as the result of experimental experience (Frank, 1946), whether it is an idealized construct, inaccessible to empirical realization, and hence merely a convention (Poincaré, 1902), or whether it should serve as a definition of the absence of forces or as a statement of the existence of inertial systems. As the Preface of the second edition of the *Principia* clearly indicates, for Newton the first law was firmly rooted in experience. True, Newton regarded physics, as the title of his work clearly shows, as a part of philosophy; but his philosophy was not a kind of speculative metaphysics but rather "experimental philosophy" in the sense of Francis Bacon's *philosophia experimentalis* (Bacon, 1620). Newtonians were convinced that the laws of motion express undeniable truths. In fact, according to Leonhard Euler (1750), these laws could even be used to reverse the traditional relation between physics and philosophy, for "knowledge of the truths of mechanics can serve as a guide in the thorny researches of philosophy . . . and it would be justified to reject all the reasons and all the ideas of metaphysics, however well founded they otherwise be, if they lead to conclusions contrary to these truths of mechanics." This statement proclaimed not only the independence of physics but even its epistemological priority over philosophy and was a protest against the traditional view that had been dominating physical thought in the past. Let us not forget in this context that philosophical, teleological, and theological thought had played an important role in the development of physics. Descartes, for example, derived the conservation laws of mass and of linear momentum from the immutability of God; for Newton, absolute space was an attribute of divine omnipresence; Maupertuis interpreted the principle of least action, from which can be derived the Lagrangian and Hamiltonian equations, the most frequently used equations in theoretical physics, as a teleological maxim, according to which nature strives to attain a maximum of parsimony. It took quite some time until it was understood that these extremum principles are merely integral formulations of differential equations and as such not more indicative of purposiveness in nature than any other law of physics.

In the 19th century, physics and philosophy were regarded as two independent and autonomous disciplines. Their alienation led, at the close of that century, to the program of a systematic elimination of all metaphysical ingredients from physics. The way to this positivistic enterprise had been prepared already by the sensationalistic school founded by George Berkeley, a contemporary of Newton. It claimed that physical objects in their *knowable* aspects are reducible to sensible properties (*"esse est percipi"*). But it did not completely exempt itself from metaphysics since the physical object continued to lurk behind the scenes as the source of its sensible appearances. The radical step of identifying the physical object with its sensible appearances was made by Ernst Mach. His positivistic replacement of metaphysical notions by logical constructions and operational definitions was instrumental for the rise of modern physics, as we shall see in Sec. 4.

2.2 The Philosophy of Space and Time in Classical Physics

Most of the philosophical problems of classical physics are conceptually related to Newton's laws of motion. It will have been noted, for example, that the first law, as stated by Newton, does not specify the reference system relative to which the state of rest or uniform motion is to be determined. The question of the identification and ontological status of this reference system became one of the major problems in the philosophy of space and time. Newton thought to have solved this problem with his theory of absolute space, absolute time, and absolute motion, the existence of which he claimed to have proved experimentally. A water-filled bucket, suspended by a twisted rope, was allowed to rotate with the unwinding of the rope; analyzing the different phases of this bucket experiment, Newton argued that the deformation of the rotating water surface could not be due to the motion of the water relative to the bucket but was rather the result of a centrifugal force and hence, according to the second law of motion, of an acceleration, which, since not relative to the bucket, must be relative to something else that he identified with absolute space.

Newton also maintained the absoluteness of time. Herein he followed his teacher Isaac Barrow (1670) who in his *Lectiones Geometricae* had already rejected the Aristotelian-scholastic conception of time as "the number (or measure) of motion" and argued that since God's creation of the world is itself a process that takes place in time, time must have existed prior to the act of creation. Influenced by Barrow and the Cambridge Platonists, like Henry More, whose notions of space and time were theologically biased, Newton declared in the famous Scholium of the second edition of the *Principia* that God, being eternal and infinite, and existing always and everywhere, "constitutes duration and space." In addition to these arguments, Newton also resorted—a point rarely mentioned—to purely epistemological considerations in support of his theory of absolute space and absolute time. Thus he wrote, for example, anticipating somewhat Immanuel Kant's critical approach in the "Transcendental Aesthetic" of the *Critique of Pure Reason* (Kant, 1781), "One can imagine that there is nothing in space, yet we cannot think that space does not exist, just as we cannot think that there is no duration, even though it would be possible to suppose that nothing whatever endures" (Hall and Hall, 1962).

Newton's theory of space and time was strongly opposed by G.W. Leibniz's relational theory according to which space and time are merely systems of relations between things or events. In a posthumously published essay, Leibniz (1715) presented a deductive construction of relational time and space that became the prototype of modern causal theories of space-time. Leibniz defined the temporal relations of simultaneity and priority in terms of causal relations and space as the order of simultaneous things or events. In a famous epistolary debate with Newton, who was represented by Samuel Clarke, Leibniz argued that absolute space offers no sufficient reason why physical objects have the particular position or orientation they happen to have and are not "placed the quite contrary way, for instance by changing East into West" (Clarke, 1717). Neither succeeded in convincing the other. Leibniz failed to counter Newton's dynamical arguments, while Newton, aware of the fact that mechanics obeys Galilean relativity so that only accelerations but not velocities can be absolute, failed to identify the privileged reference frame. He could only conjecture that "in the remote regions of the fixed stars, or perhaps beyond them, there may be some body absolute at rest," an idea taken up later by Carl Neumann (1869) in his theory of the "Alpha-Body" (Jammer, 1954).

Because of the spectacular success of Newtonian mechanics in its applications to terrestrial and astronomical problems, the absolutist theory dominated classical physics until the end of the 19th century. In the course of that century, the ancient idea of the ether, which had played an important role in the natural philosophy of Descartes, Huygens, Hooke, Euler, and others, gained renewed prominence as an explanatory principle for electrical, magnetical, and optical phenomena. With the general acceptance of Augustin Fresnel's theory of the ether as an all-pervasive stationary substratum, in which light propagates and through which the Earth is moving, the ether became the representative of the ultimate system of reference and assumed the role of absolute space. On this assumption, the time required for a light wave to pass between two points on the surface of the Earth would have to depend on whether the light wave is moving in the direction of the Earth's motion or opposite to this motion.

As is well known, the meticulously performed ether-drift experiments, like the famous Michelson–Morley interferometer experiment, showed with their "null results" the untenability of these ideas and led, with the rise of the special theory of relativity, to the abolishment of the ether and with it, presumably, of absolute space. This does not mean, however, that a non-null result of these experiments would have left Newtonian space-time as the only surviving alternative or that the issue "absolute vs relational space" has been definitely settled thereby. In fact, quite a few modern philosophers of space and time abstain from taking sides or even declare that "relativity theory, in either its special or general form, is more inimical to a relational conception ... than is classical physics" (Earman 1989; Jammer, 1993).

2.2.1 The Dimensionality Problem of Space One of the first to realize that the three-dimensionality of space, which had been usually taken for granted, poses a physical problem was Immanuel Kant. In the very first of his published writings, he claimed that "the threefold dimension of space is due to the law according to which the forces in substances act upon another" and "that the strength of their action holds inversely as the square of the distance" (Kant, 1747). Kant's reasoning can be explained most briefly by using the notation of modern vector analysis as follows. If **P** denotes the force produced by a substance of amount M which is enclosed by a surface S of radius r, then by the divergence theorem of Gauss M is equal (up to certain constants) to $\mathbf{F} \cdot d\mathbf{S}$; hence, if **F** varies like r^{-2}, S must be two-dimensional and space consequently three-dimensional.

What Kant had in mind was of course Newton's inverse-square law of gravitation. The discoveries, after 1760, of Tobias Mayer and of Charles Augustin Coulomb that also magnetic and electric forces obey the same inverse-square law seemed to support the objection, raised by Kant's critics, that the inverse-square dependence is a consequence of the three-dimensionality of space rather than vice versa as Kant maintained. It is remarkable that recent results in the study of the dimensionality problem of physical space, which Kant inaugurated in 1747, agree on this issue more with Kant than with his opponents. In order to understand this issue, let us divide the problem into three questions:

1. What, precisely, is meant by saying that physical space has a certain number of dimensions?
2. How do we know that it really has, as generally claimed, three dimensions (or, equivalently, that space-time has four dimensions)?

3. Is there a *raison d'être*, an inherent reason, for this dimensionality?

The first question is a problem of mathematics and has been dealt with by Poincaré, Brouwer, Menger, Urysohn, and other mathematicians. The second question has been answered by physicists who, like Paul Ehrenfest (1917), showed that the stability of planetary orbits, the stability of atoms and molecules, the unique properties of wave operators and many other phenomena are manifestations of the three-dimensionality of space (perhaps only after compactification if the recent Kaluza–Klein–type theories are admitted). Concerning the third question, which from the philosophical point of view is perhaps the most interesting one, it is generally agreed that there is no fully satisfactory answer. There are, however, some indications that, as Hermann Weyl (1949) phrased it, the "physical laws may cease to be indifferent to the number of dimensions on some deeper level that has hardly been touched by the physics of today. There is thus good reason to hope that our problem will one day find a cogent solution along such lines."

For a number of reasons, a promising candidate along these lines is perhaps the electromagnetic field. Referring to the fact that the Dirac equation of quantum electrodynamics, the best-established theory of elementary particles, makes sense only in a (3 + 1)-dimensional space-time, already Cornelius Lanczos (1962) declared that "we can hardly doubt that electricity has something to do with the specifically four-dimensional nature of the universe." The Maxwell equations, which can be established without "any appeal whatsoever to the notion of metric or length" (Misner and Wheeler, 1957), relate the electric field vector **E**, which as a polar vector has in an n-dimensional space n components, with the magnetic field vector **B**, which as an axial vector, or more precisely an antisymmetric second-rank tensor, has $\frac{1}{2}n(n-1)$ components. The necessary equality of the number of their components holds in an n-dimensional space only if $n = 3$, that is, only in a coordinate system in which the number of the coordinate axes equals the number of planes spanned by the axes. Only in an $(n + 1)$-dimensional space-time can Maxwell's theory be constructed upon an invariant integral form of the action and be conformally invariant. Only there do there exist two fields related as "duals," as Weyl once pointed out. Only in a (3 + 1)-dimensional space-time do Maxwell's equations determine the Maxwell field with the same "strength" as the Einstein equations determine the metric of the gravitational field (Einstein, 1956) or as the Weyl linear first-order matrix differential equation of the free neutrino determines the neutrino field (Penney, 1965). Furthermore, the Maxwell equations are fully compatible with quantum mechanics; if one calculates, for instance, the angular momentum in a multipole light-wave field, a quantity that must be an integral number of h, Maxwell's equations neatly avoid the impasse by yielding only ratios in which Planck's quantum of action h has cancelled out (Biedenharn, 1969). Finally, all information about the physical world reaches us only via the electromagnetic field. If then, in view of this truly astonishing all-pervasive reign of Maxwell's equations, we concede that the electromagnetic field is the prior datum of experience, then clearly the *raison d'être* of the (3 + 1)-dimensionality of space-time is the existence of the electromagnetic field and "that which constitutes the spatial character of reality is then simply the four-dimensionality of the field" (Einstein, 1953). But then also Kant's derivation of the three-dimensionality of space from the laws "according to which the forces in substances act upon another" is not so erroneous as his critics contend.

In contrast to the dimensionality of space, the dimensionality of time has scarcely ever been the subject of philosophical inquiry. It has played some role in certain scholastic debates about the Aristotelian doctrine of the unity and uniqueness of time. For if time, as defined by Aristotle, is the measure of motion, the obvious multiplicity of motion, it was argued, should imply a multiplicity of times ("*tot tempora quot motus*"). It was in this context that, e.g., Roger Bacon in his *Liber Primus Communium Naturalis* emphasized in 1245 that time "has only one dimension like length." A noteworthy exception is also a passage in Kant's *Inaugural Dissertation* in which time is represented as a one-dimensional line, intersected by surfaces (today we would say "hypersurfaces") of simultaneity, anticipating modern graphical representations of space-time (Kant, 1770). The controversial speculations of accounting for precognition or extrasensory perception, such as J. W. Dunne's (1927) theory of multidimensional time or C. D. Broad's thesis of two-dimensional time, had little relevance to physics. A serious attempt at rigorously deriving the one-dimensionality of time was made only within the context of the theory of relativity (Dorling, 1970).

2.2.2 The Directionality Problem of Time Just as Newton's first law, the law of inertia, by not specifying the reference system relative to which it holds, gave rise to problems in the philosophy of space (as explained in Sec. 2.2), so did Newton's second law give rise to problems in the philosophy of time. The second law, expressed in Euler's formulation, states that "force equals mass times acceleration" or $F = md^2x/dt^2$ and, being a second-order differential equation, is invariant under the time-reversal transformation $t \rightarrow -t$. Hence, if $f(t)$ is a solution of the equation of motion, $f(-t)$ is a solution too, so that, if any process occurs in accordance with the laws of mechanics, so does the reverse process. The same invariance holds for the Lagrangian or Hamiltonian equations of motion, and for the Maxwell equations of the electromagnetic field, as well as for the Lorentz force formula. Moreover, it also holds for the fundamental equations of modern physics, such as Einstein's field equations of the gravitational field in general relativity, the Schrödinger equation in quantum mechanics, and even the equations of quantum electrodynamics (Watanabe, 1951). And yet, experience shows that the time-reversed versions of processes such as the fall of an apple from the tree or of the emission of an electromagnetic wave or the heat flowing from a hotter to a colder body never occur in nature. The most elementary example of irreversibility is the case of two particles in a Newtonian universe. If initially they are approaching each other, they will eventually be moving apart; but if they are initially moving apart, they will never be approaching each other.

Already Aristotle mentioned in his *Physics* instances of such an irreversibility, and referring to the so-called "destructive power of time," he pointed out that "we are wont to say that time crumbles things" (221 a 32); but he emphasized that "after all, it is not really time itself that destroys things in this way, but the changes that do destroy take place in time" (222 b 26). Two thousand years later, physicists reaffirmed that the apparent asymmetry does not result from the properties of time itself but is rather due to an asymmetry inherent in the physical processes. They thus faced the question of how it is possible that, for instance, the changes or motions of a macroscopic body are not time reversible, whereas the laws of motion pertaining to its constitutive parts are.

The earliest attempt to answer this question was made in the context of thermodynamics and the statistical theory of entropy. In 1854, Rudolf Clausius announced what became known as the second law of thermodynamics, namely that the entropy of an isolated system never decreases. Ludwig Boltzmann, who had studied Maxwell's probability distribution of the velocities of molecules, made use of them in the construction of his famous H function, whose negative value behaves like entropy. By showing in his H theorem that on the molecular level a nonequilibrium state evolves toward equilibrium with an increase of entropy corresponding to an increase of randomness, he thought to have solved the problem of how individual time-reversible motions can account for collective time-irreversible motions.

However, his proof of the monotonic increase in time of the negative value of the H function contained a probabilistic postulate, the assumption of the molecular chaos (*Stoßzahlansatz*) according to which the velocity of a molecule is uncorrelated with its position. This probabilistic component of the proof was not only the ultimate reason of Josef Loschmidt's "reversibility paradox" and Ernst Zermelo's "recurrence paradox," it also led Boltzmann to admit that the entropy law itself has only probabilistic validity, so that there might exist in the vast expanses of space regions in which the entropy decreases and time "runs backward."

Although the conception of time "running backward" has sometimes been used in physics, as, e.g., in Richard P. Feynman's interpretation of an antiparticle as a particle moving "backward in time" for a fraction of a microsecond, in order to explain pair production and pair annihilation, there is little agreement among physicists and philosophers on what precisely it means that the direction of time is reversed throughout cosmic expanses envisaged by Boltzmann. Could intelligent beings, if they existed in such regions, regard their own time as "forward" and merely claim that the entropy decreases, as the philosopher Hans Reichenbach (1956) suggested? If it is true, as Norbert Wiener (1958) and F. Russell Stannard (1966) contended, that any exchange of information or any physical interactions between regions of opposing time directions are impossible, such considerations would always be doomed to be mere speculations.

Boltzmann's search for a reduction of thermodynamics to Newtonian mechanics was motivated primarily by his desire to save the mechanistical world picture and the unity of physics. His awareness of having ultimately failed in accomplishing this task may well have been one of the causes of his suicide. It is therefore noteworthy that more recently the incompetence of classical mechanics to describe irreversible processes has been challenged and it has been pointed out that there exist irreversible processes that do not involve any increase of entropy (Popper, 1956). In fact, time-irreversible solutions of Newton's equations actually have been constructed, albeit only by resorting to forces the existence of which is not vouchsafed by classical mechanics or by any other well-established theory in physics. Nonentropic, boundary-dependent, so-called *de facto* (in contrast to nomologically) irreversible processes have become a widely discussed topic in the philosophy of time (Grünbaum, 1963).

It has also been pointed out that it is a widespread error to invoke the second law of thermodynamics as an objective criterion of the direction of the "flow" of time from the past to the future, or what Arthur Stanley Eddington (1929) called "the arrow of time," to express the one-way property of time that has no analog in space. Since the entropy law has no factual content in the absence of an independent definition of the direction of increasing time, such a criterion would only debase this law to a mere tautology.

Alternative criteria for the definition of the direction of time, the so-called "nonentropic time arrows," have been proposed. The most important of these are the "electromagnetic time arrow," based on the exclusive acceptance of the retarded solutions of the electromagnetic wave equation, the "microscopic time arrow," based on the fact that a rare mode of the neutral K-meson decay can be explained only as a violation of the time-reversal invariance, the "cosmological time arrow," which derives the unidirectionality of time as a result of the expansion of the universe, the "quantum-mechanical time arrow," expressing the preparation-registration asymmetry of quantum states (Bohm *et al.*, 1994), and the "quantum-gravitational arrow of time" (Wald, 1980).

It has, however, also been maintained that the idea of a "time arrow" has no objective reality in nature nor that it is even required in theory, for "throughout the sciences of physics and chemistry, symmetrical or two-way time everywhere suffices . . . we can no longer regard effect as subsequent to cause. If we think of the present as pushed into existence by the past, we must in precisely the same sense think of it as pulled into existence by the future" (Lewis, 1930). This theory of "symmetrical time" does not necessarily conceive of time as merely a psychological phenomenon, as has been done by numerous philosophers in the past; it only denies the existence of a preferential and structurally distinguished direction of time. It resembles, therefore, but is not identical with, the conception of spatialized time that characterizes the "block universe" interpretation of the space-time of relativity, which will be dealt with in the sequel.

3. ANTIMECHANISTIC THEORIES

3.1 Dynamism

The purely mechanistic world picture, most consistently proposed by Descartes according to whom only space (or extension, as he called it) and motion are the ultimate ingredients of physical reality, and subsequently modified in Boyle's kinetic theory of matter, suffered already a serious blow as soon as Newton ascribed to the notion of "force" physical reality. The "dynamization" of physics, which characterizes the renunciation of the mechanistic world picture in favor of a dynamical one, was a gradual and slow-paced process that began with Leibniz, who, in fact, introduced the term "dynamics" in his treatise *Specimen Dynamicum* (Leibniz, 1695). In opposition to Descartes, Leibniz declared that "agere est character substantiarum," "To act is the characteristic of substance; extension, on the other hand, is nothing but the continuation or diffusion of the substance already presupposed, which thrives, withstands and resists. . . ." This statement may well serve as the motto of "dynamism," the doctrine that "force" (in Greek δυναμις), which at that time denoted also what later was called "energy," has ontological supremacy in nature and is the ultimate element of physical reality. In contrast to the scholastic dictum "*operari sequitur esse*" ("action follows existence"), dynamism declares "*quod non agit, non existit*" ("what does not act, does not exist"). That in spite of such declarations Leibniz had to resort in his treatment of mechanical problems to the corpuscular theory of particle collisions had methodological reasons.

Physical arguments for the vindication of dynamism against the corpuscular-kinetic approach have been put forward especially by Roger Joseph Boscovich (1758) on the basis of the principle of continuity, which admits only continuous changes in nature in accordance with the proverbial "*natura non facit saltus.*" The ultimate particles of matter or atoms cannot be spatially extended, Boscovich argued, because this would imply a discontinuous change of density at their boundaries. Furthermore, in a collision of two particles their velocities would have to change abruptly when they rebound from each other. Hence, he concluded, atoms must be merely unextended or pointlike sources of force. To account for the impenetrability and gravitation of matter, he assumed that these forces are repulsive at small distances and become attractive in accordance with Newton's law of gravitation. Using modern terminology, one is tempted to say that the Boscovichian atom is a singularity of a field of forces.

Boscovich's ideas had a profound influence on many philosophers, amongst them Christian Wolff, Moses Mendelssohn, Immanuel Kant, Gustav Theodor Fechner, and Hermann Lotze, as well as on physicists like André Ampère, Augustin Cauchy, Siméon Poisson, and, most importantly, Michael Faraday. For Faraday, under the influence of Boscovich, was the first to propose experimental evidence for the physical reality of the notion of force. As a mathematical device that assigns to each point in a region of space some quantity, the value of which is a function of the coordinates of space and time, this notion had been used before by Lagrange, Laplace, and others. But if the criterion of the physical reality of a theoretical construct lies in the demonstration that this construct possesses detectable properties other than those for the sake of which it has been introduced, Faraday's famous experiments, described in his *Experimental Researches in Electricity and Magnetism* (1839), fully satisfy this criterion with respect to the notion of field. Although Faraday's philosophical interest focused primarily on the problem of whether electric and magnetic interactions propagate continuously through space or are actions at a distance, his work and its elaborations by James Clerk Maxwell and Heinrich Hertz led to a change in the conception of physical reality that Einstein (1934) called "the most profound and fruitful one that has come to physics since Newton." On another occasion, Einstein went even further when he wrote, "Under the influence of the ideas of Faraday and Maxwell the notion developed that the whole of physical reality could perhaps be represented as a field whose components depend on four space-time parameters. If the laws of this field are in general covariant, then the introduction of an independent (absolute) space is no longer necessary" (Einstein, 1953).

3.2 The Electromagnetic World Picture

Still after the introduction of the field conception the mechanistic world picture had a strong ally to keep it alive, the advantage of its psychological pictorability. The statement of William Thomson (Lord Kelvin) that we could really "understand a particular topic in physics" only "if we can make a mechanical model of it" characterizes this psychological need and explains why even a mathematically minded physicist, like Maxwell, described electromagnetic phenomena, and in particular the ether of optics, in terms of mechanical models. But Maxwell was fully aware that such pictorial representations should not be taken too seriously. "I wish merely to direct the mind of the reader to mechanical phenomena which will assist him in understanding the electrical ones. All such phrases in the present paper are to be considered as illustrative, not as explanatory," he wrote in his famous article "A Dynamical

Theory of the Electromagnetic Field" (Maxwell, 1865).

Only two years after his death, an electromagnetic field theory of dynamics made its debut with a study by Joseph John Thomson that envisaged the possibility of explaining mechanical inertia in terms of electromagnetism. A new research program intended to abolish the hegemony of mechanics by reducing it to electromagnetism rather than to reduce electromagnetism to mechanics. The central difficulty of explaining ordinary mass seemed to have been solved by showing that a moving charged particle has an effectively greater mass than its rest mass as a consequence of self-induction in accordance with Faraday's laws. Oliver Heaviside, George F. C. Searle, Wilhelm Wien, and most persistently Max Abraham, who declared in 1902 that "the mass of the electron is of purely electromagnetic nature," believed that ultimately all processes in nature would be reduced to convection currents and their radiations and that the "stuff" of the world would be stripped of its material substantiality. The metaphysical maxim underlying these ideas, namely that matter no longer does what it does because it is what it is, but is what it is because it does what it does, attracted much attention even on the part of more conservative physicists. Thus for example Boltzmann (1897), who had tried to reduce thermodynamics to mechanics, gave his blessings to this new program when he wrote, "... The advantage of deriving the whole science of mechanics from conceptions which anyhow are indispensable for the explanation of electromagnetism would be as important as if conversely electromagnetic phenomena were explained on the basis of mechanics. May the former succeed."

3.3 Energetism

In its attempt to renounce the supremacy of the mechanistic philosophy of nature, the electromagnetic doctrine had a serious competitor in the energetic monism or energetism, which claimed that the "stuff" of the physical universe is energy. Energetism referred to the empirical evidence that the various types of physical energy can be transformed one into another in accordance with the law of energy conservation. It interpreted these phenomena in terms of a single substratum that manifests itself as mechanical, thermal, chemical, electromagnetic, and possibly still other forms of energy. The metaphysical tenets of energetism can be traced back to the romantic *Naturphilosophie* of Friedrich W. J. Schelling, Lorenz Oken, and Wolfgang von Goethe, who criticized the mechanico-atomistic approach of Newtonian physics as incapable of giving a true account of nature, which they regarded as a vitalistic evolutionary system. Encouraged also by remarks made by Hans C. Oersted and Faraday, and searching for a scientific foundation of a nonmaterialistic world view, the energetists replaced the mechanistic-corpuscular conception of matter, motion, and the ether by the notion of energy.

One of the earliest energetists was the evolutionist Herbert Spencer, mostly known for his application of Darwinism to social developments. In his *First Principles,* using still "force" as synonymous with "energy," he wrote "We come down, then, finally to Force, as the ultimate of ultimates. Though Space, Time, Matter, and Motion, are apparently all necessary data of intelligence, yet a psychological analysis shows us that these are either built up of, or abstracted from, experience of Force," (Spencer, 1862). The most prominent advocate of energetism, however, was the chief founder of modern physical chemistry, Friedrich Wilhelm Ostwald. His energetism was based not only on his agreement with Ernst Mach that the atom is merely a convenient methodological device, but on much deeper metaphysical deliberations. What we sense, he claimed, is not the dubious and in itself entirely indefinite "matter" that classical physics thought to have brought within the compass of exact quantitative determination by its concept of inertial mass, it is rather a source of activities or energies that affect our senses. Ostwald claimed to have deduced Newton's second law of motion from the law of energy conservation, a deduction that would have decisively changed the course of modern physics. But as Boltzmann (1896) has shown, Ostwald's reasoning was mathematically flawed. Still, the idea of a "dissolution of matter into energy" became a topic of great interest. It was widely publicized, e.g., by Gustave Le Bon (1905) in a book published in the same year in which Einstein wrote his first paper on relativity and derived the mass-energy equivalence ($E = mc^2$) that prompted

Bertrand Russell (1948) to declare that "mass is only a form of energy . . . it is energy, not matter, that is fundamental in physics."

4. THE PHILOSOPHY OF MODERN PHYSICS

4.1 The Philosophical Background of Modern Physics

Twentieth-century physics extended the reach of scientific exploration to areas that previously had been inaccessible to research, such as high-speed phenomena or the structure of atoms. It soon became evident that, in order to cope with the newly discovered phenomena, not only the laws but even the conceptual foundations of traditional physics had to be subjected to a drastic revision. Questions concerning the precise meaning and the philosophical implications of the new conceptions confronted the physicist with the need of adopting a specific philosophical stance for his work. From the epistemological point of view, he could choose between *realism*, which maintains that the physical theory reflects, strictly or only structurally, the behavior of real objects, the existence of which was affirmed without doubt, and *positivism*, according to which the task of physics consists only in describing relations between measurable quantities or results of measurements, without specifying whether anything, yielding these results, does actually exist. From the metaphysical point of view, he has the option between *determinism*, according to which, like in classical physics, the final state of a system is unambiguously determined or predictable, provided the initial conditions and the laws are known, and *probabilism*, which admits chance as an unavoidable and inherent element at least in the realm of microscopic phenomena. Each of the four compatible combinations among these alternatives has a representative among the founders of modern physics.

4.2 Philosophical Problems of the Special Theory of Relativity

Special relativity was the first fully elaborated physical theory the conceptual foundations of which deviated fundamentally from those of classical physics. Traditionally, it is said to be based on two major principles; the principle of relativity, which for purely mechanico-kinematical processes had been announced already by Galileo (Galilean relativity), and the principle of the invariance of the velocity of light, which explained the null results of the well-known ether drift experiments, such as the Michelson–Morley experiment, performed at the end of the 19th century. Aware of the apparent contradiction between these two principles and puzzled by certain asymmetries in the explanations of electromagnetic phenomena, Einstein resolved these difficulties in 1905 by a drastic revision of the traditional conceptions of space and time. The critical step in these revisions was Einstein's operational definition of the simultaneity of spatially separated events. In fact, the only topic of the first section of his revolutionary paper "On the Electrodynamics of Moving Bodies" (Einstein, 1905) was precisely the concept of simultaneity from which, in the sequel of the paper, the relativistic effects, the relativity of simultaneity, the time dilatation, the length contraction, and even the Lorentz transformations are derived (in accordance with the two principles) as logical consequences. Taking into consideration that this procedure, as documentary evidence proves, was instrumental for the development also of quantum mechanics and modern physics as a whole, "It is hard to overestimate the impact of Einstein's definition of simultaneity" not only on the physics but also on the "philosophy of this century," as a contemporary philosopher of science expressed it (Nerlich, 1982).

Although the notion of simultaneity had been the subject of numerous discussions in the past (Jammer, 1986), Einstein realized that it had never been given a physically meaningful definition. For according to Einstein (1917a), "The concept does not exist for the physicist until he has the possibility of discovering whether or not the concept is applicable in a given case. We thus require a definition of simultaneity such that this definition supplies us with the method by means of which, in the present case, one can decide by experiment whether or not two events occur simultaneously." Einstein's approach, at that time, was obviously that of a staunch positivist, similar to that which had prompted Ernst Mach, 40 years earlier, to his well-known operational definition of the concept of mass (Jammer, 1961).

Einstein defined simultaneity in terms of clock synchronization as follows. Let e_A and e_B denote events that occur respectively at the points A and B of an inertial reference frame. A clock stationary at A, indicates "A time" and an identically constructed clock, stationary at B, indicates "B time." "Common time" can now be defined as follows. A light signal is emitted from A at "A time" t_A, is reflected from B at "B time" t_B, and returns to A at "A time" t'_A. If $t'_A - t_B = t_B - t_A$, then the two clocks are said to indicate "common time" and are called synchronized; and if they, thus synchronized, indicate the same time at the occurrences of e_A and e_B, these two events are by definition simultaneous.

Assuming that this procedure can be generalized for any number of points and that simultaneity is an equivalence relation, Einstein called the thus-generalized common time "the time of the reference system" or simply "coordinate time." In the sequel, he proved the relativity of simultaneity and of time by showing that two events simultaneous in one reference system are not so in another, and that the coordinate time of a given event differs in different reference systems.

Einstein's definition of simultaneity and the conclusions he drew from it gave rise to a number of logical, epistemological, and metaphysical problems. It was asked, for instance, whether Einstein's reasoning was not circular in view of the fact that his definition of time used the notion of an inertial system which, if defined as a system in which Newton's first law holds, obviously presupposes the notion of time. Another logical problem dealt with the question of whether the equivalence of simultaneity, rather than being postulated, can be obtained as a logical consequence of fundamental assumptions.

Interested in clarifying such problems and convinced that only an axiomatic formalism reveals the precise logical structure of a theory, Hans Reichenbach (1924) reformulated the theory of relativity as a rigorously deductive axiomatized system. Faced with the choice of what to posit as axioms and what as definitions, Reichenbach distinguished sharply between the empirical or factual components and the definitional or conventional components of the theory. Calling a definition, or more generally any theoretical construction, conventional if it can be replaced by an alternative that, though logically incompatible with the former, has the same observational consequences, he realized that Einstein's simultaneity definition is conventional. More precisely, he argued, the empirical facts of the theory of relativity allow replacing the Einsteinian condition $t_B = t_A + \frac{1}{2}(t'_A - t_A)$ by the more general alternative $t_B = t_A + \epsilon\,(t'_A - t_A)$, where (for causality reasons) $0 < \epsilon < 1$, without ever leading to conflict with experience.

The thesis of conventionality, namely that no facts of nature preclude the choice of an arbitrary value of ϵ between 0 and 1, gave rise to heated debates and is still today a controversial issue. The traditional method to establish simultaneity or, equivalently, to synchronize distant clocks, by sending a light signal—or any other signal for that matter—from one clock to the other and taking into account the time lapse due to the transmission velocity of the signal, was countered by Reichenbach by pointing out that "to determine the simultaneity of distant events we need to know a velocity, and to measure a velocity we require knowledge of the simultaneity of distant events. The occurrence of this circularity proves that simultaneity is not a matter of knowledge" but a matter of convention. In fact, if c denotes the two-way velocity of light that has been measured by Fizeau, Foucault, Michelson, and others with the help of light-reflecting mirrors, etc., that is, in the above-mentioned notation, twice the distance between A and B divided by $t'_A - t_A$, where only the clock in A has been used showing that c is convention free, then Reichenbach's alternative equation yields for the one-way velocity of light from A to B the value $c/2\epsilon$. Hence, knowledge of ϵ and knowledge of the one-way velocity imply each other as Reichenbach contended in his *The Philosophy of Space and Time* (Reichenbach, 1928), which has been called "the greatest work in the philosophy of physics in the 20th century" (Salmon, 1977).

The professional literature abounds in ingenious proposals for convention-free measurements of the one-way velocity of light and for other very sophisticated contrivances to disprove the conventionality thesis. A historical example is the "twin-rod experiment" of two rods of equal rest length, moving in opposite directions with equal speeds; the meetings of their extremities have been claimed to be simultaneous events. This argument, based on what is one of the most simple experiments conceivable, has misled, among others, even Wolfgang Pauli (1958), one of the shrewdest critics physics ever had, to declare that it "shows that the determination of the simultaneity of spatially separated events . . . can be carried out entirely with the help of measuring rods, without the use of clocks." The fallacy of this and other proposals to this effect follows also from the fact that John A. Winnie (1970) and others succeeded in foregoing deliberately any choice of ϵ and in deriving from premises not containing any hidden one-way velocity assumptions the equations of relativistic kinematics with ϵ as a free parameter, and yet in obtaining all experimentally testable or "factual" results of the theory. However, David Malament's (1977) challenge of Grünbaum's (1963) profound elaboration of the conventionality thesis within the context of the causal theory of time has been regarded by many, though not by all, experts in this field as a serious blow to the thesis.

The metaphysical import of this issue can be gauged by the following considerations. The concept of the "present" or "now," defined as the totality of simultaneous events, or more precisely according to Minkowski as the simultaneity hypersurface through the "here-now" of the observer, depends not only, as relativity implies, on the state of motion of the observer or on the inertial system in which he or she is at rest, but also on the arbitrary choice of ϵ. Moreover, if only the present is real and the past and future unreal, as, e.g., Arthur N. Prior (1970), one of the founders of modern tense logic, persuasively contended, then reality itself is not only relative but also a conventional conception, which sounds like a *contradictio in adjecto*.

The relativization of simultaneity and time was criticized by many philosophers, and most severely by Henri Bergson, on the following grounds. Relativization may be applicable, perhaps, to "spatialized time" as used by Minkowski in his geometrical representation of space-time, it was argued; but since real time, which exhibits itself in dynamical processes and in the unfolding of states of consciousness, loses its uniqueness if joined with, or substituted by, any spatial representation, the time of relativity is but a misleading fiction.

This dynamical point of view stands in sharp contrast to the static "block universe" conception of time, according to which "the objective world simply *is*, it does not *happen*. Only to the gaze of my consciousness, crawling upward along the lifeline of my body, does a section of this world come to life as a fleeting image in space that continuously changes in time" (Weyl, 1949). As has been objected, however, even were the block-universe conception successful in eliminating temporal succession in the physical world, it could not avoid acknowledging the progress of time at least in the movement of our consciousness to the future, and every attempt to exclude dynamical time introduces somewhere at least surreptitiously the dynamical reality of time.

Nevertheless, the idea of the static four-dimensional universe found support in certain chronogeometrical analogs to the scholastic logical arguments for predestination and fatalism, which can be traced back to Aristotle's treatment of determinism in his *De Interpretatione*. The logical problem was this: Since according to the law *tertium non datur* a proposition is either true or false, and is so at all times, one of two contradictory propositions, like "tomorrow there will be a sea-battle" and "tomorrow there will not be a sea-battle," must be true today; hence, future contingencies are determinate before they occur. The chronogeometrical analogs in question are Cornelius W. Rietdijk's (1966) "rigorous proof of determinism from the special theory of relativity" and Hilary Putnam's (1967) similar proof of his claim that future events are real already in the present. Both proofs are based on the fact that for any two events e_1 and e_2 in Minkowski space-time there always exists a third event e_3 in space-like separation from them and hence simultaneous with e_1 in one inertial frame and with e_2 in another. If in particular e_1 is my "here-now" and e_2 any event in my future, then e_3 is happening in my "now" and also in the "now" of e_2. But if what happens now is determinate and real, and if determinateness and reality are frame independent, then e_2 is just as determinate and real as e_1. It follows that the future is determinate and real already in the present and that the notions of an "open future" and "free will" are fictitious.

Both proofs have been rebutted by Howard Stein (1968) and Roberto Torretti (1983) on the grounds that both arguments resort surreptitiously to the prerelativistic temporal structure of using the tenseless present for the pronouncement of their ontological assertions.

The "twin-rod experiment," mentioned above, was originally designed by Einstein (1911) to decide between two interpretations of the relativistic length contraction, whether it was a purely metrological effect, resulting from the measurement procedure or the "perspective of observation," like the change of a rod when viewed from different angles, or whether it is an intrinsic phenomenon as real as "a change of the length of a rod, caused by the change of temperature" (Podlaha, 1975). That the ether-theoretic FitzGerald–Lorentz contraction, which had been proposed to account for the null result of the Michelson–Morley experiment, had to be regarded, within the context of the Lorentz theory of electrons, as a real phenomenon, has never been questioned since it was explained as a result of the variation of the intermolecular electromagnetic forces due to the motion through the ether. Nevertheless, it became the paradigm of an *ad hoc* hypothesis in the sense of not being testable by experiments that differ in principle from the Michelson–Morley experiment (Grünbaum, 1963). It thus initiated numerous discussions on the logical and epistemological status of *ad hoc* hypotheses in physics and science in general. The Lorentz–Larmor time dilatation, on the other hand, was hardly referred to in this context, probably because the relativistic time dilatation, in contrast to the length contraction, was generally regarded as empirically well established by the Ives–Stilwell transversal Doppler effect, the lifetime measurements of muon decay in cosmic rays, and the comparison of moving atomic clocks, even though the strength of this evidence is sometimes called into question (Huang, 1993).

4.3 Philosophical Problems of the General Theory of Relativity

A philosophical analysis of a physical theory has first of all to define the subject matter of the theory. That this is not a trivial issue is shown by the fact that the special theory of relativity has been variously classified as a geometry of light, as a theory of space and time, or as part of the electromagnetic theory of Maxwell. In fact, the German periodical *Fortschritte der Physik* used to list its abstracts of papers on relativity under the heading "Elektrizität und Magnetismus" until Max von Laue (1913) pointed out that the relativity theory deals with all branches of physics and their theories and serves "as a criterion for the admissibility of all of them." More recently, Eugene P. Wigner (1962) compared it with a "superlaw" that restricts the possible forms of laws to satisfy an invariance or symmetry principle.

In order to find out what the general theory of relativity is about, it should be recalled that Einstein's 1905 theory is named "special" or "restricted" because it deals with physical phenomena in inertial systems, a restriction that from the purely kinematical point of view seems unwarranted and philosophically unsatisfactory. The question of how to formulate the laws of physics for arbitrary reference systems motivated Einstein's search for a "general" theory, the "clearest logical characterization" of which is according to Einstein (1955) the statement that it is "the theory which avoids the introduction of the 'inertial system'." Clearly, for Einstein the general theory was an extension or "generalization" of the special theory. Studying at first the simplest noninertial motion, the uniformly accelerated motion of a freely falling body in a homogeneous gravitational field, and being aware of the experimentally well-established equality between inertial and gravitational mass, Einstein postulated the (strong) "principle of equivalence" according to which every freely falling, nonrotating reference system constitutes a local inertial frame. This principle, in turn, led him to the revolutionary discovery that gravitational phenomena are reducible to, and explainable by, the Riemannian geometry of curved spacetime. The fact that the ensuing theory of gravitation was generally convariant, i.e., valid in arbitrary reference systems, justified in Einstein's view calling it the "general theory of relativity."

That this term is misleading and confusing was argued by Vladimir Fock (1959); for already in Newtonian mechanics, he pointed out, "One deals with the generally covariant Lagrange equations of the second kind, so that one would have to say that Newtonian mechanics contains in itself 'general' relativity." Fock's proposal of calling Einstein's 1916 theory a relativistic "theory of gravitation" has been accepted by many physicists.

If one tries to resolve this controversy by investigating the asymptotic approximations of the theory, one will find that for vanishing gravitation the theory reduces to the special theory of relativity, justifying Einstein's point of view, but that for slow motions and weak static fields it reduces to the classical gravitation theory of Newton and Poisson, justifying Fock's point of view. But Newtonian gravitation, based on the notion of actions at a distance, and special relativity, excluding such interactions and even superluminal velocities, are irreconcilable. The general theory exhibits therefore the philosophically interesting feature, rarely encountered elsewhere, of having not only different but even incompatible asymptotic approximations. This poses a problem since quite frequently the meaning of a high-level scientific concept is given its intuitive clarification in terms of its application in an asymptotic approximation of the theory.

No less unsatisfactory than the restriction to inertial systems and "contrary to the spirit of science" was the causal asymmetry between matter and space: Space affects matter, as in the aforementioned bucket experiment performed by Newton, but matter does not affect space. In the general theory of relativity, space and matter, or more precisely mass-energy, become mutually interacting partners: Mass-energy determines the geometry of space-time, and the latter determines the motion of matter. The mathematical expression of this interdependence is provided by the famous covariant field equations, which have been called "one of the greatest achievements of the human mind." Reducing the whole mechanics of gravitation to the solution of a single system of covariant partial differential equations, these field equations proved very successful: Whenever their predictions deviated from those of classical physics, observation confirmed their predictions. The geometry defined by these equations changes constantly in accordance with the changing mass-energy distribution, while its curvature, i.e., the deviation of its space-time metric from the flatness of Euclidean geometry, accounts for the physical phenomena usually ascribed to a gravitational field. General relativity dispelled not only the previously prevailing doctrine of the Euclidicity of physical space—which Kant regarded as an *a priori* necessary condition of experience—but also the erroneous identification of abstract logically formalized geometry with the geometry of physical space, which latter is the product of empirical investigation. Referring to this differentiation, Einstein (1921) declared, "As far as the propositions of geometry refer to reality, they are not certain; and as far as they are certain, they do not refer to reality."

The question of whether geometry is a corollary of physics or physics is a corollary of geometry has been a subject of philosophical discussions long before the advent of geometrodynamics according to which empty curved space-time is the only "stuff" of physical reality (Wheeler, 1980a). Even the fundamental problem of how the geometrical structure of a region of space or space-time can be empirically determined by means of measuring devices that themselves are physical objects and hence subjected to the deforming influence of gravitation, a problem dealt with already by Poincaré, is still a controversial issue. The epistemological holism claims that geometry and physics form an inseparable whole; the validity or nonvalidity, say, of Euclidean geometry cannot be proved by empirical observation, because it is only the whole composed of physics and geometry that can be tested empirically. "We require this whole network of theoretical considerations to arrive at an experimental means of verifications" (Weyl, 1992). This argumentation has been criticized, e.g., on the ground that "it would then have to be acknowledged that there is a sense in which physical geometry itself does not provide a geometric characterization of physics reality" (Grünbaum, 1963).

To generalize the special to the general theory, Einstein (1916) formulated "the postulate of general covariance," demanding that "the general laws of nature are to be expressed by equations which hold good for all systems of coordinates, that is, are covariant with respect to any substitution [coordinate transformation] whatever." This postulate implies, he claimed, what he called "the postulate of general relativity" according to which "the laws of physics must be of such a nature that they apply to systems of reference in any kind of motion." To prove that the latter principle is a logical consequence of the former, he pointed out that "the sum *of all* substitutions [coordinate transformations] in any case includes those which correspond to all relative motions of three-dimensional systems of coordinates." Einstein's argument was soon criticized by Erich Kretschmann (1917) on the ground that all physical theories, independently of the physical content of their laws, are expressible in generally covariant form and hence would have to satisfy the general relativistic principle; this principle could therefore not lead to a new physics entailing factual consequences not encountered before. This controversy initiated extended discussions on the nature and status of the principles of relativity, covariance, invariance, and symmetry, and led to the distinctions between coordinate systems and reference frames, between laws and metalaws (Bunge, 1961), and between "dynamical objects" and "absolute objects" (Anderson, 1967), all designed to resolve the conceptual muddle.

Another principle, that has also played an important, though only heuristic, role in the attempt to extend the relativity principle to all frames of reference, and which has likewise intrigued philosophers of physics, is what Einstein has called "Mach's principle." Briefly stated, it claims that the inertial mass of a body results from its interaction with other bodies in the universe. Because of the vagueness of its formulation and its cosmic connotations, it has never lost its philosophical appeal. True, the field equations of the theory partly incorporate this principle. For according to them the metric tensor, i.e., the geometrical structure of the inertial or gravitational field, is determined by the energy-stress tensor, that is to say, by the distribution of matter and energy in the universe. Since, however, these differential field equations are of the second order, boundary conditions have to be imposed in order to obtain solutions. But the very idea of boundary conditions is contrary to the spirit of the Machian program, according to which the observed universe should self-referentially supply the terms of its own description. Some authors, like Grünbaum (1963), regard the need of boundary conditions as a substitute for absolute space, claiming that the general theory has not at all "exorcised the ghost of Newton's absolute space." Others, like Helmut Hönl (1953), interpret Mach's principle as a constraint to serve as a selection rule to separate admissible from inadmissible solutions of the field equations. Still other philosophers, following Lawrence Sklar (1974), dispose of the principle by simply accepting the existence of inertial and noninertial systems as "just a brute, inexplicable fact about the world."

The question of how to save the general theory from the need of imposing anti-Machian boundary conditions was precisely the issue that led to the establishment of modern cosmology. For it changed the cognitive status of traditional cosmology by shifting it from the realm of poetical, theological, and metaphysical speculations to that of a rigorously established physical science. The pivotal idea, first conceived by Paul Ehrenfest in 1916, but fully elaborated only by Einstein, was to eliminate boundary conditions by conceiving the four-dimensional universe as a spherically closed manifold, finite with respect to its spatial dimensions, and to let matter turn space back on itself—in analogy to the two-dimensional unbounded, but finite, surface of an ordinary three-dimensional sphere.

In his foundational paper "Cosmological Considerations in the General Theory of Relativity," Einstein (1917b) constructed on the basis of the field equations (modified by the addition of the so-called "cosmological constant") an isotropic, homogeneous, unbounded, spatially finite, and static model of the universe. Its material content was thought to exist of a pressureless fluid or "dust," uniformly smeared over all space-time. Surprisingly, not much later, Willem de Sitter found a non-flat solution of the same field equations devoid of all matter. Clearly, the existence of two disparate solutions of the same field equations was a fatal blow to Mach's principle. After strenuous efforts to save the principle, Einstein rejected it and its underlying positivist philosophy and declared that scientific theories are free creations of the human mind; they are not derivable from sense experience but have to be confirmed by it.

The subsequent development of relativistic cosmology, from Aleksandr A. Friedmann's construction of models of an expanding universe as solutions of the field equations (without the cosmological constant), their subsequent observational confirmation by Edwin P. Hubble, until, say, Kurt Gödel's solutions yielding an absolutely rotating universe, disproved not only the validity of Mach's principle, but also that of Einstein's cosmological model of 1917. But the same development also shows that sometimes it is not the answer but the question and method of tackling it that are of lasting importance.

The promotion of traditional cosmology to the status of a respectable physical science with the help of general relativity was greatly enhanced when in the recent decades relativity, in cooperation with particle physics, investigated not only the structure of the universe but also its physical history or what is traditionally called cosmogony. By developing, for example, the by now widely accepted Big Bang theory and studying baryosynthesis and nucleosynthesis, as well as the formation of galaxies and stars, cosmology could scientifically address the age-old problem of how the universe began and how it will end. Like other branches of physics, cosmology made specific quantitative predictions that subsequently have been borne out by observation as, e.g., in the case of the well-known 3-K microwave radiation.

However, the object of these studies is the universe as a whole and not, as in all other branches of physics, a spatially and temporally small-sized portion of it. This raises the question of whether this peculiarity of cosmology allows free use of the research methodology employed in ordinary physics. For if the universe is by its very definition necessarily unique in the sense of not being a member of an ensemble and *a fortiori* not being reproducible, then clearly inductive, statistical, and probabilistic methods, employed in ordinary physics, are inapplicable.

Some philosophers distinguish between essential or necessary and accidental or contingent uniqueness, the former being the case if any imaginable alternatives to the actual universe are necessarily impossible; the latter, if it is conceivable that different kinds of universes might have been formed. In this latter case, the additional question arises: Why is the actual universe just the one that has come into existence?

It is precisely this last question that has recently become a much-debated topic in the philosophy of physics. For modern physics has shown that the universe in which we live is "finely tuned" to the conditions of human life in the sense that, if the values of the fundamental physical constants, like the charge of the electron, the strength of the strong-nuclear force, or the gravitational constant, differed even only slightly from their actual values, life would have been impossible. One may, of course, answer that this state of affairs is simply a matter of pure coincidence, or the result of a purpose or design.

Eschewing teleological explanations, the proponents of what Brandon Carter (1974) called the "anthropic principle" (or "cognizability principle") have given a different answer, which departs, however, from the traditional method of scientific reasoning. It says "that what we can expect to observe must be restricted by the conditions necessary for our presence as observers." More precisely, Carter distinguished between the basic and less controversial *weak* version of the anthropic principle, according to which "our spatio-temporal location in the universe is necessarily privileged to the extent of being compatible with our existence as observers," and the philosophically much more audacious *strong* version, according to which the universe "must be such as to admit the creation of observers within it at some stage."

Historically viewed, both versions generalize Descartes' famous "Cogito, ergo sum" ("I think, therefore I exist") to "Cogito, ergo mundus talis est" ("I think, therefore the world is as it is"), and both versions qualify the so-called (but misnamed) "Copernican principle," which denies any privilege to man's location in the universe, by asserting man's epistemological centrality. Both versions maintain an ontological and nomological compatibility relation between man and the universe as it is; but the strong version affirms in addition that any counterfactual variation of the laws or initial conditions of the universe would exclude the possibility of man's existence. A still stronger version is John A. Wheeler's (1980b) "participatory anthropic principle," "Observers are necessary to bring the universe into being," implying that in a certain sense man "participates" in the creation of the universe. The epistemological basis of Wheeler's self-referential cosmogony, namely that "no elementary phenomenon is a phenomenon until it is observed or registered," is a logical consequence of his interpretation of quantum mechanics, which will be discussed in due course.

4.4 Philosophical Problems of Quantum Mechanics

Since its inception in 1925, modern quantum mechanics, established primarily by Werner Heisenberg, Erwin Schrödinger, Max Born, and Paul A. M. Dirac, has been stupendously successful in accounting for virtually all phenomena in atomic and molecular physics, in solid-state physics, and chemistry, as well as in its contributions to nuclear physics and to the physics of elementary particles. But in no other theory is there such an irreconcilable disagreement among physicists and philosophers as to what the theory tells us about physical reality, and are physics and philosophy so inextricably entangled, as in quantum mechanics. In spite of an unprecedented sharpening of the logical, mathematical, and semantical tools for the philosophical analysis of physical theories, the ontological and epistemological problems of quantum mechanics have not only not been resolved but have even been aggravated in the past few decades. These difficulties are partly the result of the fact that, unlike the situation in any other theory of physics, the formalism of quantum mechanics preceded its interpretation.

Opinions already disagree on the referent of the theory, that is, on what precisely the theory or its formalism is about. According to the standard version, the so-called Copenhagen interpretation, the mathematical formalism in terms of the state vector or Schrödinger wave function provides a description of the individual quantum system, whereas the ensemble interpretation refers the state vector to a statistical collection of systems.

Nor is there any consensus on whether the immediate referent is physical reality itself or merely knowledge of it. Those who agree with Einstein (1949) that "physics is an attempt conceptually to grasp reality as it is thought independently of its being observed," and that "the belief in an external world independent of the perceiving subject is the basis of all natural science," insist that quantum mechanics should be a realist theory. The opposing instrumentalist or subjectivist position is taken by those who agree with Bohr that "there is no quantum world; there is only an abstract quantum physical description; it is wrong to think that the task of physics is to find out how nature *is*, physics concerns what we can say about nature" (Petersen, 1963), or with Heisenberg (1958), that "the subject of research is no longer nature itself, but man's conception of nature." It should be noted that some positivistically inclined critics refuse to attach any philosophical significance to a controversy like that between Einstein and Bohr and regard it rather, as Ernest Nagel (1961) called it, as "a conflict over preferred modes of speech." But, as we shall see at the end of our discussion, the endorsement or rejection of a philosophical position may well be motivated by physical arguments rather than by the predilection for a certain mode of speech.

If the question concerning the referent of quantum mechanics is, indeed, a matter of physical considerations, it may be tempting to search for the answer by way of applying the method of asymptotic approximations, as it has been applied above in the case of general relativity. Although there exist, in fact, different and—surprisingly—nonequivalent approximation methods (Planck's method of $h \to 0$, Bohr's method of $n \to \infty$, where n is the principal quantum number, and Ehrenfest's method of taking the expectation values of the observables), a closer analysis (Liboff, 1984; Ajanopon, 1987) shows that this method cannot supply any answer.

The tenability of the ontology of classical physics and hence also the maintainability of its realism for microphysics have been questioned already by the "old quantum theory." This term denotes the microphysics of the period between 1900 and 1924, i.e., from Planck's explanation of the black-body radiation in terms of the quantum of action h, and Einstein's subsequent interpretation of the photoelectric effect, through Bohr's theory of the hydrogen atom until Louis de Broglie's thesis of the wavelike behavior of material particles. In spite of its name, the "old quantum theory" has never become a full-fledged scientific theory. Its principal fault was its indiscriminate conflation of classical and nonclassical conceptions, making a coherent interpretation impossible.

The development of a conceptually coherent and logically consistent quantum theory began in 1925 with the advent of Heisenberg's matrix mechanics and Schrödinger's wave mechanics, and reached its maturity about 1930 with John von Neumann's rigorous proof that both formalisms, in spite of their apparent disparities, are but different representations of the same operator calculus in Hilbert space. In view of the unvisualizability of matrix mechanics, wave mechanics with its presentation of particles in terms of wave packets gained greater popularity. But the "spread" or instability of these wave packets and other peculiarities caused serious difficulties. To resolve these difficulties, Max Born (1926) proposed the statistical interpretation of the wave function as a "probability wave," the amplitude of which specifies the probability of the particle's location. The wave function thus becomes a catalog of possible probability distributions, where the word "possible" refers to the choice of the observer as to what observable he wishes to measure, while the word "probability" refers to the observer-independent relative frequency of obtaining a specific value of the chosen observable. Born's introduction of the notion of probability into quantum mechanics is generally regarded as marking the demise of the determinism of classical physics. True, classical physics has made ample use of the probability concept, as for instance in the theory of errors in measurement procedures or in statistical mechanics; but classical probability was always thought to be due to human ignorance of a more refined specification of the physical data. In quantum mechanics, however, at least according to Born's interpretation, probability became an objectively existing, intrinsic, and irreducible component of physical ontology and the source of an uneliminable indeterminism of the theory.

Laws of nature no longer, as they do in classical physics, determine the occurrence of an event but only the probability of its occurrence. This raises the problem of the precise ontological status of this probability. Its temporal evolution and spatial propagation in accordance with the Schrödinger equation seem to indicate that it is not merely a mathematical construct. Since probability waves do not transmit energy or momentum and hence can hardly be regarded as physically "real," Heisenberg (1961) compared their ontological status with that of the Aristotelian notion of *potentia*.

Heisenberg's (1927) indeterminacy relations, which exclude the possibility of simultaneously specifying the values of two canonically conjugate observables, like position and momentum, to within arbitrary degrees of accuracy, have also been regarded as a confutation of determinism. For determinism states that *if* the initial state, e.g., position and momentum, of a system is exactly known, then the state of the system at any other time can be predicted. But Heisenberg's principle denies the possibility of ascertaining the initial state. The objection that the falsity of the premise of an implication does not necessarily entail the falsity of the implication was rebuffed by Heisenberg on the ground that any conception of a deterministic universe, hidden behind the statistical universe of perception, is a meaningless speculation since physics deals only with the relationships between perceptions.

The indeterminacy relations are intimately related to Bohr's (1927) complementarity principle and generally regarded as its quantitative aspect. Complementarity requires that the quantum object be described in two different, even contradictory, ways, which are complementary in the sense that both are necessary to exhaust the possibility of obtaining information about the object, even though the two ways are experimentally not simultaneously performable. Complementarity is in Bohr's view the only possibility of generalizing the classical mode of description without ignoring the quantum of action. His belief in the unity of knowledge led him to extend the epistemology of complementarity to other branches of science, like psychology or biology, and even to consider it as a means to reach greater mutual understanding among nations and cultures.

The best-known example of complementary descriptions is the wave-particle duality; and the easiest way to demonstrate it is the well-known two-slit experiment which, according to Richard Feynman (1951), "has in it the heart of quantum mechanics . . . and contains the *only* mystery" of this theory. In fact, this experiment, which, incidentally, has recently been actually performed with neutrons, implies a wave description because of the interference effect if both slits are uncovered, and a particle description if only one slit is uncovered. Moreover, since in any case, the interference case included, the photographic plate records pointlike and not spread-out events, the experiment may be regarded in its final stage also as a measurement of position and as illustrating thereby the so-called "collapse of the wave packet" or "reduction of the wave function," a characteristic feature of every measurement process in quantum mechanics.

The very notion of a quantum-mechanical measurement raises a number of philosophical problems not encountered in classical physics. Every measurement or observation requires an interaction between the measured or observed system and the measuring device or observer. In classical physics, in which the action of the latter on the former is assumed to be negligible or eliminable by correction, this interaction is a one-way process; but in quantum mechanics, because of the indivisibility of the quantum of action, it is a two-way process with the dire consequence that, in general, a measurement of an observation changes the quantum state of the system measured or observed. Furthermore, since such measurements necessarily involve macroscopic measurement devices to yield pointer readings, and since quantum mechanics should include classical physics as a special case, the ensuing logical circularity between quantum and classical concepts questions the self-consistency and conceptual autonomy of quantum mechanics.

Aware of this difficulty, Bohr regarded the measurement outcome as depending upon the entire experimental setup, combining object and apparatus, and conceived the object-apparatus interaction as epistemologically irreducible. In order to eliminate this irreducibility, it has been deemed necessary to include the measurement process into the formal account of quantum mechanics. How to do this is the task of the "measurement theory of quantum mechanics."

Such theories, most of them modifications of a proposal made by von Neumann (1932), face two difficulties, namely the questions of how and where the so-called "collapse of the wave packet" takes place, questions that are controversial issues still today. For if the measuring device itself is treated as a quantum system, it adds to each component of the superposition, which describes the initial state of the object, an additional term instead of reducing it to an eigenstate corresponding to the eigenvalue of the observable that is to be measured. In other words, object and measuring device form a combined system for the measurement of which a second measuring apparatus is required. This process obviously leads to an infinite regress.

Among the numerous suggestions to resolve these difficulties, the proposal made by Fritz London and Edmond Bauer (1939) attracted much attention because of its relevance to the mind-body problem. They suggested that it is the human "faculty of introspection" ("*connaissance immanente*") that actualizes the reduction of the superposition in the consciousness of the sentient observer. The London-Bauer measurement theory has far-reaching philosophical implications. Since according to quantum mechanics every physical interaction between particles involves a superposition, a world without a sentient observer, who reduces the superposition, would be a limbo of mere potentialities that never actualize into factualities; it would be a world in which nothing whatever happens. It was precisely this consideration that led Wheeler to the formulation of his aforementioned "participatory anthropic principle."

Critics objected that this measurement theory, although free from logical inconsistencies, is hardly compatible with the intersubjective agreement among independent observers. Furthermore, they pointed out, the notion of a mental act should not serve as a primitive concept in physics, for too little is known about the mind-body problem or about the brain, which is supposedly the seat of consciousness and probably the most complex system in the whole universe. This observer-dependent measurement theory in their view simply explains a riddle by a mystery.

To avoid this impasse, alternative measurement theories have been proposed, some of them resorting to the irreversibility of macroscopic systems, others to a hypothetical micro-to-macro amplification process, and still others to a modification of the very structure of quantum mechanics into a nonlinear theory. But so far no satisfactory solution seems to have been obtained. The well-known "paradoxes" of "Schrödinger's cat" (1935a) and of "Wigner's friend" (1962) illustrate the bizarre conclusions of such measurement theories.

An attempt to dispense altogether with the reduction of the wave function has been made by Hugh Everett III (1957) and elaborated by Bryce S. DeWitt and Wheeler in the ontologically most daring and epistemologically most provoking theory ever made in physics. By regarding the quantum object, the measuring apparatus, and the observer as one single quantum system, described by a state vector that never collapses, their theory eliminates the need for an extraneous observer or consciousness and affirms that physical reality is rigorously deterministic. The theory implies a continual splitting of the universe into a multitude of noncommunicating worlds, each corresponding to a component of the superposition, and it is therefore called "the many-world interpretation." Since no experiment performed in one branch of the splitting universe can detect the outcome of a measurement obtained in another branch, appearances are saved. The theory has been criticized as violating Ockham's rule that entities should not be multiplied without necessity, as ambiguous in its determination of the branches of the splitting universe because a given state function can be expanded in different sets of functions, and as not providing an explanation of why the observer is just in that branch in which he happens to be.

As is well known, the orthodox interpretation of quantum mechanics has been strongly opposed by physicists, like Einstein and de Broglie, as well as by philosophers, like Karl Popper and Mario Bunge. Most noteworthy in this context is the long-standing debate between Einstein and Bohr (Bohr, 1949), in which Einstein challenged the probability interpretation and the indeterminacy relations by a number of thought experiments that Bohr successfully rebutted on the basis of his complementarity interpretation (Jammer, 1974). One of the highlights of this debate was the argument in which Einstein presented his thought experiment with a photon box in order to disprove the time-energy indeterminacy relation, and which Bohr refuted by referring Einstein to his own general theory of relativity (Bohr, 1949).

Although defeated by Bohr, Einstein did not lose interest in the photon box. To eliminate the troublesome gravitational effect, he regarded the box and the photon as two interacting systems that then move apart along a horizontal line. This led him to the physical scenario of the Einstein–Podolsky–Rosen (EPR) argument against the completeness of quantum mechanics. The published version (Einstein *et al.*, 1935) owes its central idea to Einstein, its mathematical elaboration to Nathan Rosen, and its final formulation to Boris Podolsky. Although Einstein, as we know from his correspondence with Schrödinger, did not fully agree with the printed text, it was, of course, the published version that affected the subsequent conceptual development of the theory.

The paper begins with a necessary condition for the completeness of a physical theory: "Every element of the physical reality must have a counterpart in the physical theory," and defines "physical reality" by the sufficient condition: "If, without in any way disturbing a system, we can predict with certainty . . . the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity." Accordingly, physical reality is not to be defined by *a priori* philosophical considerations, "but must be found by an appeal to results of experiments and measurements." In addition to these two explicitly stated stipulations of completeness and reality, a third premise is assumed that is mentioned only *en passant*: If two systems no longer interact, "no real change can take place in the second system in consequence of anything that may be done to the first system." For Einstein this premise or "principle of separation," as he called it in a letter to Schrödinger, was an essential feature of his realist philosophy. "But on one supposition," he wrote (Einstein, 1949), "we should, in my opinion, absolutely hold fast: the real factual situation of the system S_2 is independent of what is done with the system S_1, which is spatially separated from the former." Actually, this supposition is a conflation of two different principles, the "separability principle" and the "locality principle." The former states that the spatio-temporal separation is a sufficient condition for the individuation of a system or its state, and the latter, that the state of a system at a given space-time location can be affected only by events within or on the past light cone of that location. The separability principle is a metaphysical maxim; the locality principle is a physical feature of the special theory of relativity and excludes, in particular, any simultaneous action at a distance.

EPR consider a system of two particles S_1 and S_2 that are widely separated in space after having interacted with each other. They are supposed to be in a state in which their total momentum $p_1 + p_2$ and their relative position $q_1 - q_2$ have definite known values. The existence of such eigenstates is assured by the fact that the operators for this sum and difference commute. From these assumptions, EPR infer that, if they measure the momentum of S_1, they can predict with certainty the momentum of S_2 without disturbing it; "if, instead of this, [they] had chosen" to measure the position of S_1, they could have predicted with certainty the position of S_2 without disturbing it. Hence, the reality criterion implies that the momentum as well as the position of S_2 has reality. Were quantum mechanics a complete theory, it should therefore, in accordance with the completeness criterion, ascribe definite values to the momentum and position of S_2. But quantum mechanics does not do this because the respective operators do not commute. Hence EPR conclude "The quantum mechanical description of physical reality given by the wave function is not complete." The possible objection that this conclusion would be valid only if the momentum and position of S_2 are simultaneously predictable is countered by pointing out that this condition implies simultaneity of the corresponding measurement processes on S_1; but then the reality of p_2 and q_2 "would depend upon the process of measurement carried out on the first system [S_1], which does not disturb the second system in any way. No reasonable definition of reality could be expected to permit this." Obviously, EPR resort to the principles of separability and locality to meet this objection and the related criticism of making use of the counterfactual definiteness, i.e., of the controversial assumption that unperformed experiments have definite results, an assumption ascribed to their above-quoted statement beginning with the words "if, instead . . ."

As might be expected, the EPR paper became the target of numerous criticisms, the most important of which was Bohr's. Faithful to his positivistic philosophy, Bohr (1935) rejected the argument by claiming that in order to speak validly of real properties of particles "it is necessary to take into account the experimental conditions that define the possible types of prediction about the system since these conditions are an inherent element in the description of any phenomenon to which the expression 'physical reality' can meaningfully be attached." In particular, this requirement illegalizes the use of counterfactual implications for defining physical reality. Bohr was particularly interested in refuting the EPR argument for the following reason. If Heisenberg's indeterminacy relations are interpreted not merely as a prohibition of the simultaneous measurability of conjugate observables but rather, in the spirit of the Copenhagen interpretation, as a denial of the very existence of sharp values of such observables, the EPR argument disproves even the consistency of quantum mechanics.

Schrödinger (1935b), in contrast to Einstein and Bohr, saw the importance of the EPR argument in its disclosure of the existence of the nonseparability, or entanglement ("*Verschränkung*") as he called it, of the subsystems or components S_1 and S_2 in a composite system S. This means that a measurement carried out on one of the components, whatever its distance from the other, interacts with S and affects thereby the other component; and it implies holistically that the state of S is not fully described in terms of the states of S_1 and S_2, a situation often epitomized by saying that "the whole is more than its parts." This entanglement, Schrödinger declared, is *the* characteristic feature of quantum mechanics, "the one that enforces its entire departure from classical lines of thought."

Schrödinger's notion of entanglement was eagerly seized by parapsychologists as scientific support for their theses of ESP (extrasensual perception), clairvoyance, psychokinesis, and other paranormal phenomena, as well as by interpreters of quantum mechanics as a revival of Eastern mysticism. Although some reputable physicists, like the Nobel laureate Brian Josephson or Richard Mattuck, expressed some sympathy with such ideas, the majority of physicists strongly reject them. Even Wheeler, whose name is often quoted by parapsychologists because of his participatory anthropic principle, vehemently opposed such ideas with the words ". . . Let no one use the EPR experiment to claim . . . any 'quantum interconnectedness' between separate consciousnesses," and he condemned those ideas as "baseless" and "moonshine."

Certain disadvantages of the published EPR argument have been eliminated in its reformulation by David Bohm (1951) in terms of the states of a pair of spin-$\frac{1}{2}$ particles in the singlet state. Bohm's formulation became the standard presentation of the EPR argument in the philosophical literature on this issue.

The EPR incompleteness argument with its emphasis on "elements of physical reality" was soon regarded as an incentive for the construction of hidden-variable theories "to complete" quantum mechanics, although—as still unpublished documentary evidence shows—the argument was definitely not intended to serve as such an incentive. Attempts to construct such theories had been made already in the late 1920s for the purpose of reinstating classical determinism in quantum mechanics in quantum physics. But as a result of von Neumann's (1932) "proof" of the impossibility of hidden variables, interest in such theories had waned until it was revived by the publication of the EPR paper. Bohm (1952) formulated a logically consistent hidden-variable theory that ascribes to every particle definite values of position and momentum as well as a continuous trajectory by adding to the usual potential a "quantum mechanical potential." By averaging over the values of the hidden variables, he retrieved all empirically testable results of the standard theory. Since in Bohm's theory of measurement the hidden variables depend not only on the state of the observed object but also on that of the measuring device and thus differ from those considered by von Neumann, the latter's impossibility "proof" did not apply.

Still, the status of Bohm's theory and of other hidden-variable theories and their relation to von Neumann's argument were fully understood only when John Stewart Bell clarified the situation. His first contribution to these problems was his proof that von Neumann's reasoning, though mathematically impeccable, rests on a certain premise, the so-called additivity postulate, which need not hold for hidden variables (Bell, 1964a). Furthermore, he drew attention to the fact that in Bohm's theory the quantum potential exhibits the "curious feature of a grossly nonlocal character."

Studying the question of whether local deterministic hidden-variable theories, i.e., theories in which the hidden variables determine unambiguously the result of every single act of measurement and satisfy the locality principle, are possible, Bell made what Josephson called "the most important recent advance in physics" and Henry P. Stapp even "the most profound discovery of science" (Stapp, 1975). Using Bohm's EPR version, Bell showed that every deterministic hidden-variable theory that is local involves a certain inequality that became known as "Bell's inequality." This inequality—and this is the important point—is *not* satisfied by quantum physics. This result, "Bell's theorem," namely that no local deterministic hidden-variable theory can reproduce all the predictions of quantum mechanics, has been formulated by Bell as follows: "In a theory in which parameters are added to quantum mechanics to determine the results of individual measurements, without changing the statistical predictions, there must be a mechanism whereby the setting of one measuring device can influence the reading of another instrument, however remote" (Bell, 1964b).

Bell (1971) subsequently generalized his theorem for the case of stochastic hidden variables. Philippe H. Eberhard (1977) derived a Bell-type inequality without using hidden variables and without the assumption of determinism. Michael Redhead (1987) showed that on this assumption counterfactual conditionals are innocuous. Generalizations of Bell's inequality are still today a widely studied subject in the foundational research of quantum mechanics. But so also are experiments performed to test whether quantum mechanics satisfies these inequalities. Practically all these experiments, and among them the very sophisticated one performed by Alain Aspect (1981) and his team, confirm the predictions of quantum mechanics and violate Bell's inequality. These results imply that the classical world picture of local realism has to be given up irretrievably. For even should ever quantum mechanics turn out to be wrong, its successor must comply with these experimental facts. Moreover, these experiments can be regarded as inaugurating what has been called "experimental metaphysics," the existence of which strikingly disproves the aforementioned thesis that a disagreement between philosophical doctrines about the nature of physical reality is but "a conflict over preferred modes of speech."

As mentioned above, locality is a fundamental requirement of special relativity. The experimentally verified nonlocality of quantum mechanics may therefore be thought to imply the incompatibility of quantum mechanics with special relativity and, moreover, the possibility of sending messages with superluminal velocities. A detailed analysis (Jarrett, 1984) shows, however, that the locality principle used in the proof of Bell's theorem is a logical conjunction of two other conditions, one of which does not violate the principles of special relativity. Since quantum mechanics satisfies only this latter condition, it enjoys a peaceful coexistence with relativity.

The conceptual problems dealt with so far have been discussed in the language of the traditional framework of quantum mechanics. However, there exist other theoretical frameworks in which quantum mechanics, especially if axiomatized, is studied, each of them having its advantages and its deficiencies. Philosophically most remarkable is the so-called quantum logical approach according to which quantum mechanics should be formulated in a logic different from that applied in the description of classical physics.

Proponents of quantum logic (*q.v.*) justify their approach by comparing the relation between logic and quantum physics with the relation between geometry and physics in Poincaré's philosophy of geometry, referred to above in Sec. 4.3. Since always only an inseparable combination of two such components is empirically testable, one of them (e.g., logic) can be altered if also the other component (the formulation of the physical theory) is changed without affecting the factual situation; but the modified physical theory, it is claimed, may turn out to be conceptually more acceptable than its original unmodified version. Garrett Birkhoff and von Neumann (1936) were the first to propose a nonclassical quantum logic. In their logic, the distributivity law with respect to the logical connectives "and" ($\wedge$) and "or" ($\vee$) is not valid. In other words, if p, q, and r denote propositions, then the classical distributive law $p \wedge (q \vee r) = (p \wedge q) \vee (p \wedge r)$ does not always hold. If, e.g., it is known that the spin of a spin-$\frac{1}{2}$ particle is "up" in the sense that the proposition p, "$\sigma_z = +1$," in units of $\frac{1}{2}\hbar$, is true, and if q and r denote, respectively, the propositions "$\sigma_x = +1$" and "$\sigma_x = -1$," then it is easy to confirm that—at least in accordance with the realist interpretation of quantum mechanics—this distributivity law does not hold. A more technical exposition that does justice to the Birkhoff–von Neumann logic would have to specify it as a propositional calculus having the structure of an orthocomplemented nondistributive (modular) lattice, in contrast to the Boolean lattice of classical logic (see QUANTUM LOGIC).

Other versions of quantum logic soon followed. Reichenbach (1944) developed a three-valued logic that contains, in addition to the truth values "true" and "false" of classical logic, a third truth value "indeterminate" that applies to statements that were regarded as "meaningless" in the Copenhagen interpretation: but in Reichenbach's logic the question of whether a system has a certain value for a given observable has three answers: "yes," "no," and "maybe," where "maybe" represents a genuinely irreducible state of affairs. Carl Friedrich von Weizsäcker (1955) proposed a "many-valued complementarity logic" in which truth values correspond to probabilities.

Quantum logic has become an immensely rich and diverse field of new mathematical and philosophical ideas. It has stimulated the development of lattice theory and, more generally, of the theory of partially ordered sets and has even contributed to various mathematical generalizations of existing theories. Pavičić's (1992) bibliography on quantum logic, compiled for the *International Quantum Logics Association* and containing about 2000 entries, is indicative of the immensity of work done in this field.

Nevertheless, there are quite a few critics who challenge the status of quantum logic on the ground that it does not provide satisfactory answers to fundamental questions such as the following: Has quantum logic convincingly shown that the genuine and universally valid logic is non-Boolean and that the traditional logic is merely a kind of asymptotic approximation to it? Can logic be really empirical in the sense of "being read off" from empirical data, in contrast, e.g., to Einstein's famous remark to Heisenberg (1971) that "it is the theory which decides what we can observe"? How can quantum logic differ from classical logic if the mathematics used in quantum mechanics is necessarily based on classical logic? And, after all, does quantum logic truly resolve any of the problematic issues of quantum mechanics instead of explaining a riddle by mystery, as the observer-dependent measurement theory is said to have done?

In conclusion, it should be pointed out that this review of the philosophy of physics has said nothing about the philosophical implications of some physical theories that are at the forefront of current research. In fact, no philosophical comments have been made about quantum field theory or the S-matrix approach with its bootstrap conjecture, or about the standard model of elementary-particle physics, based on gauge theories of the electroweak and strong interactions, or about superstring theories designed to incorporate the standard model with general relativity in a logically consistent way and to unify all known fundamental forces in nature. Nor have any remarks been made on the prospects and philosophical significance of the so-called "Final Theory," which is to explain everything in nature from a minimum of basic assumptions.

These topics are widely publicized in popular magazine articles. But they are also rigorously discussed in professional writings of such competent physicists and philosophers of science as Cushing (1983), Redhead (1983), Stöckler (1984), Brown and Harré (1988), Kanitscheider (1989), Weinberg (1992), and Teller (1994). Since, however, research in these areas is at present making enormously rapid progress, we believe that it is still too early to grasp the full philosophical significance of these developments and that it is still premature to draw well-founded philosophical conclusions. We agree therefore with the statement, made by Teller, that "as yet we have no grip on what these theories are telling us about how to conceive the world."

Works Cited

Ajanopon, P. (1987), *Am. J. Phys.* **55**, 159–163.

Anderson, J. L. (1967), *Principles of Relativity Physics*, New York: Academic Press.

Aspect, A., Grangier, P., Roger, G. (1981), *Phys. Rev. Lett.* **47**, 460–463.

Bacon, F. (1620), *Novum Organum*, London: Billius.

Barrow, I. (1670), *Lectiones Geometricae*, London: Godbid.

Bekker, I. (Ed.) (1831–1870), *Aristotelis Opera*, Berlin: Reimer.

Bell, J. S. (1964a), *SLAC-PUB-44*; (1966), *Rev. Mod. Phys.* **38**, 447–452.

Bell, J. S. (1964b), *Physics (Long Island City)* **1**, 195–200.

Bell, J. S. (1971), "Introduction to the Hidden Variable Question," in: B. d'Espagnat (Ed.), *Foundations of Quantum Mechanics*, New York: Academic Press, pp. 171–181.

Biedenharn, L. C. (1969), Private communication.

Birkhoff, G., von Neumann, J. (1936), *Ann. Math.* **37**, 823–843.

Bohm, A., Antoniou, I., Kielanowski, P. (1994), *Phys. Lett. A* **189**, 442–448.

Bohm, D. (1951), *Quantum Theory*, Englewood Cliffs, NJ: Prentice-Hall.

Bohm, D. (1952), *Phys. Rev.* **85**, 166–193.

Bohr, N. (1928), *Nature Suppl.* **121**, 580–590.

Bohr, N. (1935), *Phys. Rev.* **48**, 696–702.

Bohr, N. (1949), "Discussion with Einstein," in: P. A. Schilpp (Ed.), *Albert Einstein: Philosopher-Scientist*, Evanston, IL: Library of Living Philosophers, pp. 199–241.

Boltzmann, L. (1896), *Ann. Phys. (Leipzig)* **57**, 39–71; **58**, 595–598.

Boltzmann, L. (1897), *Vorlesungen über die Prinzipien der Mechanik*, Leipzig: Barth.

Born, M. (1926), *Ann. Phys. (Leipzig)* **37**, 863–867; **38**, 803–827.

Boscovich, R. J. (1758), *Theoria Philosophiae Naturalis*, Vienna: Remondidiana.

Brown, R. H., Harré, R. (1988), *Philosophical Foundations of Quantum Field Theory*, Oxford: Clarendon Press.

Bunge, M. (1961), *Am. J. Phys.* **29**, 518–524.

Carter, B. (1974), "The Anthropic Principle and Large Number Coincidences," in: M. Longair (Ed.), *Confrontations of Cosmological Theories with Observation*, Dordrecht: Reidel.

Clarke, S. (1717), *A Collection of Papers which passed between the late Learned Mr. Leibniz and Dr. Clarke*, London: Knapton.

Cushing, J. T. (1983), *Philos. Sci. Assoc.* **2**, 31–56.

Dorling, J. (1970), *Am. J. Phys.* **38**, 539–540.

Dunne, J. W. (1927), *An Experiment with Time*, London: Black.

Earman, J. (1989), *World Enough and Space-Time*, Cambridge MA: MIT Press.

Eberhard, P. A. (1977), *Nuovo Cimento B* **38**, 75–80.

Eddington, A. S. (1929), *The Nature of the Physical World*, New York: Macmillan.

Ehrenfest, P. (1917), *Proc. K. Ned. Akad. Wet.* **20**, 200–209.

Einstein, A. (1905), *Ann. Phys. (Leipzig)* **17**, 891–921.

Einstein, A. (1911), *Phys. Z.* **12**, 509–510.

Einstein, A. (1916), *Ann. Phys. (Leipzig)* **49**, 769–822.

Einstein, A. (1917a), *Über die Spezielle und die Allgemeine Relativitätstheorie*, Braunschweig: Vieweg.

Einstein, A. (1917b), *Sitzungsber. K. Preuss. Akad. Wiss.*, 142–152.

Einstein, A. (1921), *Sitzungsber. K. Preuss. Akad. Wiss.*, 123–130.

Einstein, A. (1934), *Essays in Science*, New York: Philosophical Library.

Einstein, A., Podolsky, B., Rosen, N. (1935), *Phys. Rev.* **47**, 777–780.

Einstein, A. (1949), "Autobiographical Notes," in: P. A. Schilpp. (Ed.), *Albert Einstein: Philosopher-Scientist*, Evanston, IL: Library of Living Philosophers, pp. 1–95.

Einstein, A. (1953), "Preface," in: M. Jammer, *Concepts of Space*, Cambridge, MA: Harvard Univ. Press.

Einstein, A., Kaufman, B. (1955), *Ann. Math.* **62**, 128–138.

Einstein, A. (1956), *The Meaning of Relativity*, 5th ed., Princeton, NJ: Princeton Univ. Press.

Euler, L. (1750), *Hist. Acad. R. Sci. Belles Lett.* **4**, 324.

Everett, Hugh, III (1957), *Rev. Mod. Phys.* **29**, 454–462.

Faraday, M. (1839), *Experimental Researches in Electricity and Magnetism*, London: Quaritch.

Feynman, R. (1951), *The Feynman Lectures in Physics*, Vol. 2, Reading, MA: Addison-Wesley.

Fock, V. (1959), *The Theory of Space Time and Gravitation*, London: Pergamon Press.

Frank, P. (1946), *International Encyclopedia of Unified Science*, Vol. 1, Chicago, IL: Univ. of Chicago Press.

Galileo, G. (1623), *Il Saggiatore*, Rome: Mascardi.

Grünbaum, A. (1963), *Philosophical Problems of Space and Time*, New York: Knopf; (1973), 2nd ed., Dordrecht: Reidel.

Hall, A. R., Hall, M. B. (Eds.) (1962), *Unpublished Scientific Papers of Isaac Newton*, Cambridge: Cambridge Univ. Press.

Heisenberg, W. (1927), *Z. Phys.* **43**, 172–198.

Heisenberg, W. (1958), *The Physicist's Conception of Nature*, London: Hutchinson.

Heisenberg, W., Born, M., Schrödinger, E., Auger, P. (1961), *On Modern Physics*, New York: Collier Books.

Heisenberg, W. (1971), *Physics and Beyond*, London: George, Allen and Unwin.

Hönl, H. (1953), *Z. Naturforsch. A* **8**, 2–6.

Huang, Y. S. (1993), *Helv. Phys. Acta* **66**, 346–360.

Jammer, M. (1954), *Concepts of Space*, Cambridge, MA: Harvard Univ. Press; (1993), 3rd ed., New York: Dover Publications.

Jammer, M. (1961), *Concepts of Mass in Classical and Modern Physics*, Cambridge, MA: Harvard Univ. Press.

Jammer, M. (1974), *The Philosophy of Quantum Mechanics*, New York: Wiley.

Jammer, M. (1986), "The History of the Concept of Distant Simultaneity," *Rend. Accad. Naz. Sci. 40 Pt. 2: Mem. Sci. Fis. Nat.* **9**, 169–184.

Jammer, M. (1990), "The Program and Principles of Descartes' Physics," in: G. Belgioioso (Ed.), *Descartes—Il Metodo e i Saggi*, Rome: Istituto della Enciclopedia Italiana, pp. 303–334.

Jarrett, J. (1984), *Nous* **18**, 569–589.

Kanitscheider, B. (1989), "Cosmology and Particle Physics—The Philosophical Aspects of a New Scientific Symbiosis," in: P. Weingartner, G. Schurz (Eds.), *Philosophy of the Natural Sciences, Proceedings of the 13th International Wittgenstein Symposium, Kirchberg/Wechsel, Austria, 1988*, Vienna: Hölder, Pichler, Tempsky, pp. 60–70.

Kant, I. (1747), "Gedanken von der Wahren Schätzung der Lebendigen Kräfte," in: *Kant's Werke* (Akademie Ausgabe), Vol. 1 (1910), Berlin: Reimer, pp. 1–181.

Kant, I. (1770), "De Mundi Sensibilis atque Intelligibilis Forma et Principiis," in: *Kant's Werke*, Vol. 2 (1912), Berlin: Reimer, pp. 385–419.

Kant, I. (1781), *Kritik der Reinen Vernunft*, Riga: Hartknoch.

Kretschmann, E. (1917), *Ann. Phys. (Leipzig)* **83**, 575–614.

Lanczos, C. (1962), *Rev. Mod. Phys.* **34**, 379–389.

Le Bon, G. (1905), *L'Évolution de la Matière*, Paris: Flammarion.

Leibniz, G. W. (1695), "Specimen Dynamicum," in *Opera Omnia* (1768), Geneva: Dutens.

Leibniz, G. W. (1715), "Initium Rerum Mathematicarum Metaphysica," in: C. I. Gerhardt (Ed.) (1962), *Mathematische Schriften*, Hildesheim: Olms; Loemker, L. E. (Ed.) (1956), *Philosophical Papers and Letters*, Chicago, IL: Univ. of Chicago Press.

Lewis, G. N. (1930), *Science* **71**, 569–577.

Liboff, R. L. (1984), *Phys. Today* **37** (2), 30–55.

Locke, J. (1690), *An Essay on Human Understanding*, London: Bassett.

London, F., Bauer, E. (1939), *La Théorie de l'Observation en Mécanique Quantique*, Paris: Hermann.

Malament, D. (1977), *Nous* **11**, 293–300.

Maxwell, J. C. (1865), "A Dynamical Theory of the Electromagnetic Field," *Philos. Trans. R. Soc.* **155**, 459–508; reprinted (1990) in: W. D. Niven (Ed.), *The Scientific Papers*, Vol. 1, Cambridge, U.K.: Cambridge Univ. Press, pp. 526–597.

Misner, C. W., Wheeler, J. A. (1957), *Ann. Phys. (New York)* **2**, 525–603.

Nagel, E. (1961), *The Structure of Science*, New York: Harcourt.

Nerlich, G. (1982), "Simultaneity and Convention in Special Relativity," in: R. McLaughlin (Ed.), *What? Where? When? Why?*, Dordrecht: Reidel.

Neumann, C. (1869), *Über die Prinzipien der Galilei-Newton'schen Theorie*, Leipzig: Teubner.

Newton, I. (1687), *Philosophiae Naturalis Principia Mathematica*, London: Streater; (1947), *Mathematical Principles of Natural Philosophy*, Berkeley, CA: Univ. of California Press.

Pauli, W. (1958), *Theory of Relativity*, New York: Pergamon Press.

Pavičić, M. (1992), *Int. J. Theor. Phys.* **31**, 373–461.

Penney, R. (1965), *J. Math. Phys.* **6**, 1607–1611.

Petersen, A. (1963), *Bull. At. Sci.* **19**, 8–14.

Podlaha, M. F. (1975), *Indian J. Theor. Phys.* **23**, 69–75.

Poincaré, H. (1902), *La Science et l'Hypothèse*, Paris: Flammarion.

Popper, K. (1956), *Br. J. Philos. Sci.* **1**, 117–133, 173–195.

Prior, N. (1970), *Studium Generale* **23**, 245–248.

Putnam, H. (1967), *J. Philos.* **64**, 240–247.

Redhead, M. (1983), *Philos. Sci. Assoc.* **2**, 57–99.

Redhead, M. (1987), *Incompleteness, Nonlocality and Realism*, Oxford: Clarendon Press.

Reichenbach, H. (1924), *Axiomatik der Relativistischen Raum-Zeit-Lehre*, Braunschweig: Vieweg; *Axiomatization of the Theory of Relativity*, Berkeley, CA: Univ. of California Press.

Reichenbach, H. (1928), *Philosophie der Raum-Zeit-Lehre*, Berlin: Gruyter; (1958), *The Philosophy of Space and Time*, New York: Dover Publications.

Reichenbach, H. (1944), *Philosophical Foundations of Quantum Mechanics*, Berkeley, CA: Univ. of California Press.

Reichenbach, H. (1956), *The Direction of Time*, Berkeley, CA: Univ. of California Press.

Rietdijk, C. W. (1966), *Philos. Sci.* **33**, 341–344.

Russell, E. (1948), *Human Knowledge—Its Scope and Limits*, New York: Simon and Schuster.

Salmon, W. C. (1977), "Introduction," in: H. Reichenbach (Ed.), *Gesammelte Werke*, Vol. 1, Braunschweig: Vieweg.

Schrödinger E. (1935a), *Naturwissenschaften* **23**, 807–812, 824–828, 844–849.

Schrödinger E. (1935b), *Proc. Cambridge Philos. Soc.* **31**, 555–562.

Sklar, L. (1974), *Space, Time, and Spacetime*, Berkeley, CA: Univ. of California Press.

Spencer, H. (1962), *First Principles*, London: Williams and Norgate.

Stannard, F. R. (1966), *Nature* **211**, 693–695.

Stapp, H. P. (1975), *Nuovo Cimento B* **29**, 270–276.

Stein, H. (1968), *J. Philos.* **65**, 5–23.

Stöckler, M. (1984), *Philosophische Probleme der relativistischen Quantenmechanik*, Berlin: Duncker & Humblot.

Teller, P. (1983), *Philos. Sci. Assoc.* **2**, 100–111.

Teller, P. (1994), *An Interpretive Introduction to Quantum Field Theory*, Princeton, NJ: Princeton Univ. Press.

Torretti, R. (1983), *Relativity and Geometry*, New York: Pergamon Press.

von Laue, M. (1913), *Das Relativitätsprinzip*, Braunschweig: Vieweg.

von Neumann, J. (1932), *Mathematische Grundlagen der Quantenmechanik*, Berlin: Springer; (1958), *Mathematical Foundations of Quantum Mechanics*, Princeton, NJ: Princeton Univ. Press.

von Weizsäcker, C. F. (1955), *Naturwissenschaften* **42**, 521–529.

Wald, R. A. (1980), *Phys. Rev. D* **21**, 2742–2755.

Watanabe, S. (1951), *Phys. Rev.* **84**, 1008–1025.

Weinberg, S. (1992), *Dreams of a Final Theory*, New York: Pantheon.

Weyl, H. (1922), *Space-Time-Matter*, London: Methuen.

Weyl, H. (1949), *Philosophy of Mathematics and Natural Science*, Princeton, NJ: Princeton Univ. Press.

Wheeler, J. A. (1980a), *Epistemologia* **3**, 59–98.

Wheeler, J. A. (1980b), "Beyond the Black Hole," in: H. Woolf (Ed.), *Some Strangeness in the Proportion*, Reading, MA: Addison-Wesley.

Wiener, N. (1958), *Cybernetics*, Cambridge, MA: MIT Press.

Wigner, E. (1962), "Remarks on the Mind-Body Question," in: I. J. Good (Ed.), *The Scientist Speculates*, New York: Basic Books, pp. 284–302.

Wigner, E. (1964), *Phys. Today* **17** (3), 34–40.

Winnie, J. (1970), *Philos. Sci.* **37**, 81–89, 223–238.

PHILOSOPHY OF TECHNOLOGY

ERWIN MARQUIT, *School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota, U.S.A.*

Introduction 417
1. **Technology and Societal Development** 419
1.1 Role of Technology in Human Evolution 419
1.2 Hunting and Foraging Cultures 420
1.3 Transition from Mobile to Sedentary Cultures: Crop Cultivation, Animal Domestication, and Social Differentiation 420
1.4 Occupational Specialization in the Bronze and Iron Ages 421
1.5 Reckoning Technology, Mathematics, and the Origins of Writing in Mesopotamia and the Americas 422
1.6 Technological Development and Blockage in Classical Greece 423
1.7 Technology and Society in China 423
1.8 The Middle Ages 424
1.9 Beginnings of the Technological Revolution in Feudal Europe 425
1.10 From Feudalism to the Industrial Revolution in Europe 426
1.11 From the Industrial Revolution to Modern Technology 428
2. **Origins of Philosophy of Science and Philosophy of Technology** 429
3. **Contemporary Views on Philosophy of Technology** ... 431
4. **Unresolved Problems of Technological Development** 434
Works Cited 436
Further Reading 437

INTRODUCTION

Philosophy of technology, which only emerged in the nineteenth century as a field of philosophy, investigates interconnections among the human needs that invoke a technology for their satisfaction, the determinants of these needs, and how people come to know them.

The term *technology* is a combination of the Greek *technē*, meaning art, and *logos*, meaning word or discourse. In English, it first (1615) denoted a discourse on an art or arts (*Shorter Oxford English Dictionary*, 1933), later embracing the scientific study of the practical and industrial arts and their methods. In philosophical literature, the terms *technology, technic* (or *technics*), and *technique* are differentiated. Hood gives the traditional definition of technology originating with Aristotle as "a human arrangement of technics—tools, machines, instruments, materials, sciences, and personnel—to make possible and serve the attainment of human ends" (Hood, 1983, p. 347). *Technique* is usually used to convey the method of using *technics*, but because usage varies among different authors and languages, readers may find the terms used interchangeably. Martin Heidegger's concept is broader: "The manufacture and utilization of equipment, tools, and machines, the manufactured and used things themselves, and the needs and ends that they serve, all belong to what technology is." He adds that according to the current conception, technology is "a means and a human activity" (Heidegger, 1977, pp. 4–5).

Heidegger here incorporates two contemporary connotations of *technology*, the most common being the means and methods used to produce and deliver goods and services for industrial, commercial, and military use. An-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

other, or anthropological, sense is "the body of knowledge available to a civilization that is of use in fashioning implements, practicing manual arts and skills, and extracting or collecting materials" (*American Heritage Dictionary*, 3rd edition, 1993). This second sense excludes aspects of modern technology, which depends more on the integration of science with technological applications and less on the manual arts and skills. Hidden in the differences among these definitions is a philosophical question: does technology deal with a body of knowledge, only its application, or both?

Technology refers in some usage to "spiritual," religious, and aesthetic needs also. In 1897, Alfred Espinas, noting that the Greeks applied *technē* to both the fine and the applied arts, argued that "we should give the applied arts the name techniques in order to distinguish them from the arts that tend to produce aesthetic emotion" (Espinas, 1897, p. 7). The case of a culture that regards a rain dance as necessary for agricultural production, however, shows that such a distinction is not always clear.

Ferré (1988, p. 26) suggests limiting the term *technology* to the "practical implementation of intelligence." He considers the purpose of "practical intelligence" to be "to survive or thrive," and that of "theoretical intelligence" as "to know or understand" (p. 33). This distinction, however, may obscure the social connections that form the essence of technology and constitute the principal stimulus for its development.

The following definition of *technology* is suggested for our discussion here: *Technology denotes the systematic, conscious application of knowledge by a society to meet its culturally conditioned real and perceived physical, biological, and spiritual needs and to guarantee its continued social existence. This knowledge is the socially accumulated understanding of how human labor and available material resources and artifacts are integrated by a society; the aspect of the society primarily involved here is its particular set of historically conditioned social relations of production.*

The association of technology here exclusively with human activity is deliberate to stress the central role that tool use and toolmaking play in human life, despite the fact that animals use and even make tools. The qualitative difference between these activities by animals on the one hand, and tool use and toolmaking in a deeply social context by humans on the other, is so sharp, however, that one scholar even suggests the term *utensil* be used for objects used or produced by animals, while the term *tool* be limited to objects produced by humans (Tobach, 1993). Only humans consciously create artifacts. Early examples are primitive tools made by combining or reshaping natural objects to be held in the hand for physical protection or for procuring or processing food. Shaped stone objects in the vicinity of fossil bones identify the remains as protohuman. Toolmaking has been a criterion to distinguish humans from other animals at least since the time of Benjamin Franklin, to whom James Boswell attributed the definition that man is "a toolmaking animal" (Boswell, 1934, p. 245). Thomas Carlyle's remark that "Man is a Tool-using Animal" has also been used to distinguish between humans and other animals (Carlyle, 1921, p. 37). Charles Darwin (1981) attributed to tool use a distinctive role in human evolution.

Since technology involves the conscious application of knowledge by a society, ontological and epistemological questions arise: What is this knowledge and how do we come to know it? This body of knowledge, in turn, involves three areas: material objects (frequently referred to in philosophical literature as "artifacts"), skill and knowledge in employing them, and the technological relations of production, that is, the way human labor is organized in the technological process.

Our species emerged in a process of biological evolution. The technology by which we provide our material needs was an integral part of this process. After the emergence and biological stabilization of our species, technological development continued. We begin this article with a discussion of the role of technology in human evolution, then go on to consider the interconnections between technological development and the evolution of human societies through their various stages. Readers are referred to PHYSICS AND TECHNOLOGY, HISTORY OF, for a detailed discussion of that history. Here, we focus our attention on stages of technological development that represent turning points in human history. We consider some theoretical and practical areas in which science and technology interact and conclude the article with a discussion of viewpoints being presented in contemporary philosophical literature on problems arising from technology today.

1. TECHNOLOGY AND SOCIETAL DEVELOPMENT

A principal concern of philosophy of technology is how technological development influences the course of societal interrelationships at different stages of societal organization, as well as how specific stages of societal organization and culture affect technological development. Among philosophical approaches, technological determinism sees technological development as a spontaneous evolutionary process requiring a given society to organize itself so as to make efficient use of the technologies becoming available. It also seizes upon individual technological innovations as immediate and direct causes for fundamental social change. A second approach views technological advance as a consequence of the development of human spirit and culture. The most generally held view is that technological change is the cause of profound societal change at certain points in human history, while at other times societal change stimulates technological development. Followers of this third approach attach widely varying weights to the two sides of this interaction.

The discussion in Sec. 1 of this interaction between technology and society is organized around the historical sequence of key stages of human societal development. The aim here is not to unfold a chronology of technology, but to examine the intricacies of the interrelationships between technology on the one hand, and the complex of science, culture, and socioeconomic organization on the other. (It will be noted that this section and PHYSICS AND TECHNOLOGY, HISTORY OF, are thus very different in content and focus.)

1.1 Role of Technology in Human Evolution

The interconnections between technology and society begin with the evolution of the human species. It is therefore appropriate to begin this section with the three-way interaction of technological, societal, and biological interactions that have led to the emergence of the modern human species *Homo sapiens*.

The unique character of human toolmaking and tool use along with the uniqueness of bipedality suggests these two factors as the primary stimulus for the evolution of the modern human being.

The evolutionary line of the modern human species, *Homo sapiens*, can be traced backward through at least two earlier species of the genus *Homo*, namely *Homo habilis* and *Homo erectus*, to the bipedal genus *Australopithecus*, originating in Africa some five million years ago, branching off from the family of the primates from which the modern chimpanzee descends (Hood, 1993). Bipedality frees the hands for many activities, including using tools. The first fossil evidence for tools appears later—some 2.5 million years ago—stone choppers, scrapers, and sharp flakes, probably used for stripping meat and skin from bones (Napier, 1993, pp. 43–44; Howells, 1993, p. 118; Isaac, 1976, p. 487), leading some scholars to suggest that the development of the brain was a more important factor in human evolution (Lovejoy, 1993). On the other hand, the tools that we observe among pongids like the chimpanzee today such as stripped branches for prodding termite nests or stone anvils for cracking open hard fruits (Candland, 1987; Goodall, 1968) are not likely to leave a fossil trail, so that it is not unreasonable to assume (even without evidence) that the early bipedal hominids not only did not lag behind the pongids in toolmaking, but also had begun to advance in their use of tools.

Apart from structural changes favoring bipedal locomotion, the fossil trail from the australopithecines to *Homo sapiens* is marked by significant increases in brain size (Falk, 1993, p. 64); changes in the bone structure of the hand (Napier, 1993); a lowering of the larynx so as to increase the space between the larynx and the back of the nasal cavity, thereby enhancing the possibility for articulate speech (Laitman, 1993); and a decrease in the size of the canine teeth (Skelton *et al.*, 1993). Toolmaking, development of the hand, socialization of food consumption, increasing differentiation in vocal communication, and mental development move forward in the evolutionary process as an integrative whole, a development in one area contributing to the development of the others.

1.2 Hunting and Foraging Cultures

The term *hunting and foraging* is generally used to characterize both the level of technological development and form of societal organization within which this technology is applied. The first evidence of hunting, probably by *Homo sapiens*, appears some 100 000 years ago in northwestern Germany, where a fossil elephant with a wooden spear nearly eight feet long between its ribs was found. Apart from the weapons themselves, the earliest symbolic representation of their use are the cave paintings with hunting scenes in northwestern Spain and southwestern France painted perhaps some 10 000 to 30 000 years ago.

To overcome the lack of information about the technological culture of ancient hunting and foraging societies, it is necessary to consider recent and contemporary hunting and foraging cultures, such as the Mbuti people in the Ituri rain forest in Zaire. These people lived until the 1980s much as they had thousands of years ago, with a cultural form that is essentially pre–Stone Age. The primary economic standard is one of adequacy, the mobile lifestyle mitigating against accumulation of private property. The adult men, women, and unmarried youth all participate in the hunt. Mbuti net hunters establish their symbiotic relationship with the forest ecology by a culture of minimal hunting, spending a daily average of only four to five hours away from camp (Turnbull, 1983, p. 138).

The close cooperation of women and men in the hunt and the division of labor between them are reinforced in their hunting songs, which are always sung in round form, utilizing a technique that recreates the cooperative patterns required by the hunt. The other major types of song, the foraging song and the death song, also reinforce the appropriate patterns for the corresponding activity. Except for the lullaby, none of the songs can be produced by a single singer. The ritualistic character of the singing suggests that this art form is viewed as a necessary component of the hunt itself.

The egalitarian character of Mbuti society is characteristic of most hunting and foraging societies, generally marked by the absence of hierarchical structures. Group members make decisions collectively or through councils of elders. The prestige of certain individuals is based on respect for their achievements as outstanding hunters or foragers. The individual function of the chief stands out; he gives leadership in dealing with conflict with other groups of people. A second individual function is that of the shaman, who is believed to have special powers in connection with the group's religious beliefs or mythology, and a third is one with skills in rituals practiced by the group, such as a dancer or singer (Coudert, 1991, p. 398). Although a certain degree of social division of labor, partly based on levels of skill, emerges in hunting and foraging societies, the society is essentially free of social conflicts; the disputes and tensions that do arise do so as conflicts among individuals and are frowned upon by the other members of the society. It is for this reason that many social philosophers reject the sociobiological views asserting that humans are genetically conditioned to be aggressive and acquisitive.

1.3 Transition from Mobile to Sedentary Cultures: Crop Cultivation, Animal Domestication, and Social Differentiation

The first major revolution in social organization associated with technological developments appears to have occurred toward the end of the Ice Age in the period 12 000 to 10 000 B.C. in the Near East. Sedentary settlements and the subsequent development of crop cultivation and domestication of animals were accompanied by relatively rapid technological changes in comparison with the preceding millennia, resulting not only in the increase in production above the subsistence level, but also in the emergence and deepening of social stratification within the population. This process, culminating in the formation of great city-states, occurred in Mesopotamia, Egypt, China, India, Mesoamerica, and the Andes.

Donald O. Henry, on the basis of suggestions from Binford (1980) and Gould (1982), sees the first signs of technological change as involving a transition from *simple foraging* to *complex foraging*. Henry (1989, p. 4) summarizes simple foraging as "a high mobility and movement to resources" and *complex* *foraging* as a "collecting strategy in which greater residential permanence is maintained through storage and logistical acquisition of resources."

The increased productivity that resulted from this transition led to both population increase and social differentiation. The new technology involved not only a new level of foraging activity, but also qualitative changes in the handling of the products as a result of the increased quantity of goods and people. Administrative problems seem to arise when more than two hundred people engage in coordinated activity. Structures that appear to have been communal storehouses suggest the need for administering and redistribution of the goods stored. Similarly, ceremonial structures suggest leaders (Henry, 1989, p. 212), and some central management may have been needed for housing the larger population.

A widely accepted model of social development through several stages of the social reorganization that began with the production of a food surplus has been put forward by Fried (1967). Fried suggests that the domestication of animals and plants initiated a series of changes in social organization from the egalitarian societies of hunters and foragers to the formation of what he calls rank societies, then into stratified societies, and finally to the formation of the state.

The rank society is characterized by the emergence of an individual, such as a chief, whose principal function is the redistribution of the products derived from the technological activity. According to Fried, in a stratified society, essential basic resources become the private property of a ranked segment of the community. Fried makes a distinction between what might be called the personal property of an individual or family, that is, objects of immediate use or consumption, and the basic resources that are needed to produce them (pp. 193–196).

The state emerges as an institution that stabilizes the stratification through control over the productive resources and products. Haas (1987, p. 2) defines the state society "as a type of society in which rulers have control over production or procurement of basic resources and as a result exercise coercive power over their respective populations."

Stone-walled dwellings and storage pits are evidence of sedentary settlements; stable storage of food was necessary beyond that immediately consumed (Henry, 1989, p. 19). The production of such a surplus is needed since life cycles of plants and, to a large extent, animals are tied to the seasons of the year. Stored supplies can be exchanged among different communities or even be produced specifically for exchange.

Once the technology to produce a surplus is available, the possibility for continual accumulation of a wealth of products arises, although this potential has not been realized uniformly everywhere, since subsistence-level economies persist to this day. Therefore, Fried's model of social evolution must be viewed as a tendency rather than a specific sequence of stages through which every coherent community of people must pass.

The change from mobile to sedentary cultures involves a division of labor in two distinct areas: the organization of production and the associated infrastructures for production and storage on the one hand, and the administration of the process of redistribution for consumption on the other. The managerial/administrative functions are often combined with the role of leader in ritualistic functions, since the rituals to a large extent are considered necessary components of the technology. The increase in the storable surplus also creates the conditions for the separation of the product from the producers through increasing bureaucratization of the administrative/managerial function. Those appropriating the product ultimately must establish state institutions based on a morality justifying the social stratification and also set in place means of physical force to enforce the new social morality.

1.4 Occupational Specialization in the Bronze and Iron Ages

Bronze metallurgy was introduced about 3300 B.C. in Mesopotamia and Egypt. This technological change was not a revolution in the sense of the speed with which it spread, but in the economic, political, and social consequences of the use of metal in weapons and tools. The new technologies significantly increased the level of agricultural production and thereby stimulated larger-scale irrigation projects. The agricultural surplus made possible an increase in the number of people who could separate themselves from agriculture, thus contributing to the growth of urban settlements. The rank societies turned into stratified state societies by the fourth century B.C., with the political and economic control centered in temple-cities ruled by priest-kings. Iron metallurgy appeared about 2200 B.C. and was in widespread use by 1100 B.C.

During the same period, the plow, the horse, and wheeled vehicles were introduced. The workmen involved in the mining, transport, and smelting of ores and in producing tools may have been the first specialists separated from agriculture who did not raise their own food and were not part of the ruling bureaucracy. It therefore appears that by the third millennium B.C. technological development in Mesopotamia and Egypt had reached a level that could sustain the principal types of social stratification we now find in contemporary populations except, perhaps, for a strata of intellectuals. The first intellectuals, the mathematicians of Babylon, may have also emerged in this period, but evidence for their presence does not appear until the second millennium B.C.

1.5 Reckoning Technology, Mathematics, and the Origins of Writing in Mesopotamia and the Americas

Along with first evidence of the cultivation of cereals in Syria and Iran, the harvesting or hoarding of grain, and the domestication of animals occur finds of small clay objects of various geometrical shapes—spheres, cones, disks, and cylinders. Each type of token served as a counter for a definite quantity of a particular food or other basic commodity stored or delivered. These tokens have been found throughout the Middle East, with the earliest dated about 8000 B.C. in the Zagros Mountains of Iran (Schmandt-Besserat, 1992, pp. 40, 168).

After the emergence of city-states in the fourth millennium B.C., accounting acquired a new importance. The tokens, originally sealed in small clay containers (envelopes) were now impressed on the clay envelopes into which they were placed so that the contents were visible on the envelopes themselves. Ultimately, the tokens became unnecessary, as their shapes and markings were reproduced on clay tablets. A conceptual leap toward abstract thinking then occurred. Quantitative and qualitative aspects of the goods became symbolically separated, so that a token of one shape represented a given commodity, while a second set of tokens represented the quantity. The wedge-shaped impressions developed into the system of writing known as cuneiform (Latin *cuneus*, a wedge).

In the Peruvian Andes in the eighth or ninth century, a period of major state formation, a system of knotted strings called *khipus* (Patterson, 1991, p. 43) was used to keep accounts. The *khipus* did not lead to writing, however. The Mayans later adopted a system of writing from other peoples and refined it with both ideographic and phonetic elements. They used it to record events and dates, deeds involving the nobles, the position of the stars and planets, religious rituals, and calendar reckoning, but there is no evidence of its use for accounting. The failure to develop written accounting and, except perhaps for calendars, the largely noneconomic use of writing were manifestations of different paths of technological development as compared, say, to Mesopotamia. The absence of large domesticated animals restricted the transport of goods, and the resulting fragmentation of the economies further limited technological development (Gille, 1986, p. 424). The Mayan, Inca, and Aztec rulers never developed the technological infrastructure that could tightly link the conquered peoples into vast, geographically expanding empires. The extraction of tribute and the continual pillaging of weaker neighbors by the stronger city-states in the Andes and Mesoamerica evoked resistance that could not be permanently repressed, ultimately allowing a small number of Spanish invaders, with their superior military technology, to form local alliances in order to destroy the Inca, Aztec, and Mayan states (Patterson, 1991).

The first sign of activity that can be called *scientific* (according to our current understanding of scientific knowledge) in antiquity is the development of mathematics in Babylon, known to us from tablets written between 1800 and 1500 B.C. The Babylonian tablets contain lists of conversions of measurement units, multiplication tables, and mathematical exercises. The calculations concerned practical problems, such as work quotas, food requirements, and wages for workers; materials for wells and building construction; weights and measures; and inheritance and compound interest (Hooke, 1954, pp. 786–791). Ancient Egyptian mathematics was more geometrically oriented than was Babylonian, its greater sophistication attributable to greater economic complexity and more diverse technological activity. The availability of mathematical methods coordinating elaborate economic projects made it possible to undertake construction projects of increasing complexity (pp. 790–792).

The tokens, *khipus*, and the types of calculations represented on the tablets seem to indicate that mathematics did not arise as an aesthetic intellectual pastime, but was an integral part of the application of technology to the economy. Subsequently, the Greek philosophers of the fourth to first centuries B.C. separated mathematics from technology and integrated it with logic, a process that culminated with Euclid's hypothetico-deductive system of geometry.

1.6 Technological Development and Blockage in Classical Greece

Despite the emergence of Greek science concurrent with the technological advances between the sixth and third centuries B.C., no revolution in production occurred on the scale that might be expected.

The usual explanations for this "technological blockage" are based on the economic and social structures of the Greek city-states. Most Athenians were peasants, and played no direct role in the political life. *Citizens* lived in the cities and neighboring villages, were generally landowners, and did not engage in manual labor or crafts. Their economic well-being was based on extensive domestic, craft, and field labor by slaves taken as spoils of war (Anderson, 1974, pp. 29–52; Lilley, 1965, pp. 32–33). It was simpler and cheaper to put slaves to heavy work than to construct machines to do it (Lilley, 1965, p. 32). The mechanical arts, increasingly being left to slave labor, could not be respected by the ruling elite. According to Aristotle, since artisans, historically, were slaves or foreigners, the "best form of state will not admit them to citizenship" (*Politics*, III.5.1278a.5–10). "Some men are by nature free, and others slaves, and that for these latter slavery is both expedient and right" (*Politics*, I.5.1255a.39–41). "For that some should rule and others be ruled is a thing not only necessary, but expedient; from the hour of their birth, some are marked out for subjection, others for rule" (*Politics*, I.4.1254a.20–25). "No man can practise virtue who is living the life of a mechanic or labourer" (*Politics*, III.5.1278a.20–25).

Free men, according to Aristotle, would best realize their nature through the sciences—above all, through politics, "truly the master art" (*Ethics*, I.2.1094a.25). Such ideological blockage could only have led to technological blockage. It has also been suggested that the Greek view that mechanics could be reduced to a rational, closed system based on the five simple machines (lever, pulley, screw, winch, and wedge) failed to leave room for the role of experiment in the development of mechanics, which in effect had been limited to statics (Vernant, 1957, p. 217).

The stagnation discussed here is relative to the far more rapid developments in Europe in the second millennium A.D. The gradual improvement in the technological system that characterized Greek technology after the third century B.C. was continued by the Romans, who acquired through their conquests the technological traditions of Greece, the Near East, and Egypt, absorbing whatever was useful for them from other peoples in their expanding empire. Individual Romans did gather information about technological practices in various parts of the empire, but little attempt to coordinate or codify it was made (Gille, 1986, pp. 320–326). Although the towns grew, the propertied classes maintained their traditional disdain for trade (Anderson, 1974, p. 81). While the Romans did not introduce revolutionary changes in the technological system, they did make important improvements in hand tools, agriculture, architecture, city planning, the use of water-powered grain mills on an industrial scale, and the building of aqueducts, cisterns, irrigation canals, and roads.

1.7 Technology and Society in China

The history of Chinese technology paralleled in time and in some fields even excelled the early technological developments of the classical civilizations of the Mediterranean and the Near East and, later, of Europe (Needham, 1954–1967). Chinese interest in philosophy and science unfolded at about the same time as in Greece. The Chinese had already learned to write by the fourteenth century B.C., well before the Greeks (Haudricourt and Needham, 1963, p. 162). Yet by the second half of the second millennium A.D. European technology moved decisively ahead of Chinese technology.

Needham discusses in detail what he sees as the reasons why China took an early lead in so many areas of technology and then failed to maintain it. The country's topography and agriculture required a vast network of waterworks for flood protection and irrigation, especially for wet rice cultivation, and a far-flung canal system whereby the tax grain could be brought to granary centers and to the capital. A highly centralized system emerged in which the emperor assumed the role of one great feudal lord governing and collecting taxes through a gigantic bureaucracy that became known as the "mandarinate." The mandarinate collected taxes and mobilized unpaid labor (similar to corvée labor of European feudalism in which a serf is required to perform a day of unpaid labor by a feudal lord) for the water projects. They enjoyed no hereditary succession, and were recruited afresh in each generation through state examinations based on literary and cultural subjects that rarely included what might be called scientific subjects (1969, pp. 195–197). Under this system, referred to by some Chinese scholars as "bureaucratic feudalism" (pp. 177–179), the economy was based on taxation in kind, so that money was not widely used. The mandarinate was not salaried by the state, but supported itself through its administrative powers over agricultural production and whatever wealth it could garner from its bureaucratic functions in the towns.

The value system that favored the development of merchant wealth in the feudal towns of Europe found no parallel in China. While capital accumulation in Chinese society was possible, "the application of [capital] in permanently productive industrial enterprises was constantly inhibited by the scholar-bureaucrats, as indeed any other social action which might threaten their supremacy" (Needham, p. 197). "There is nothing in Chinese history resembling the conception of a mayor or burgomaster, alderman, councillors, masters and journeymen of guilds, or any of those civic individuals who played such a large part in the development of city institutions in the west" (p. 185).

Since Chinese philosophy was not capable of allowing the mercantile mentality a leading place in the civilization, it was not capable of fusing together the techniques of the higher artisanate with the methods of mathematical and logical reasoning that the scholars had worked out, so that the passage from the Vincian to the Galilean stage in the development of modern natural science was probably not possible.

1.8 The Middle Ages

The scientific and technological heritage of the Egyptian, Greek, Roman, and Mesopotamian civilizations of antiquity was in part preserved and continued by the Byzantine and Arabic scholars. *Arabic* here includes Arab and other scholars of the Near and Middle East, Northern Africa, and Moslem Spain, where Arabic was the scientific language. From its beginnings, Islam encouraged the study of the sky and the earth to find proofs of one's faith (Arnaldez and Massignon, 1963, p. 385). "Arabian schools were organized on the pattern of the Greek, publishing commentaries, encyclopaedias, dictionaries and scientific manuals. The scholars versed in more than one science if not in all ... were the rule rather than the exception" (p. 386). The best known of these encyclopedic minds were the Persian physician and philosopher Avicenna (980–1037 A.D.) and al-Biruni (973–1048 A.D., probably born in Uzbekistan). The Moslem scholars elaborated algebra and passed on Hindu–Arabic numerals from India (p. 407). Like the Romans, the Arabs absorbed the technological knowledge of the peoples they conquered and disseminated this knowledge over their empires, innovating, but not generating a technological revolution of their own.

In Europe proper, the period from the end of classical times following the collapse of Rome to beginning of the eleventh century is commonly referred to as the Early Middle Ages or the Dark Ages. The term *Dark Ages* suggests a darkening relative to Greek and Roman antiquity and in many ways is misleading. The so-called barbaric invaders had already achieved an iron-age culture and had access to the technology of the Roman empire. The principal regression relative to Greece and Rome was the collapse of an elaborate city-state culture. With the collapse of Greece and Rome, literacy in Europe outside of Byzantium retreated significantly.

The principal developments in technology during the early Middle Ages, between the collapse of Rome and the eleventh century, appear to have been the introduction of the heavy iron plow in northern and western Europe; the horse harness, which allowed the replacement of oxen by horses for plowing; and triennial crop rotation and fertilization with marl. These innovations caused a rapid increase in population and an agricultural surplus that permitted the growth of towns (Hilton and Sawyer, 1963; White, 1962; Anderson, 1974, pp. 182–183).

A technological determinist explanation for the specific direction of the development of feudalism in Europe was put forth by White, in his influential *Medieval Technology and Social Change* (1962). White asserts that the introduction of the foot-stirrup in the Frankish kingdom of the eighth century immediately changed the nature of warfare. A mounted knight could deliver a strong blow with a lance, and cavalry replaced foot soldiers as the main fighting force. Since "mounted warriors could only be maintained in large numbers by landed endowment . . . [the] estates of the Church were . . . seized and handed over to an enlarged body of followers on condition that they serve him on horseback. . . . Protofeudal and seignorial elements had, of course, saturated the very fluid Celtic, Germanic, late Roman, and Merovingian societies; but the need for cavalry . . . precipitated and crystallized these anticipations to form medieval feudalism" (pp. 4–5).

Hilton and Sawyer (1963) criticize White's technological determinism as being methodologically at fault, the conclusions based on it not standing up to the historical evidence. They show that the stirrup was not a new arrival in Europe but reached Europe well before the eighth century, and that a complete change in Frankish weapons did not take place at that time—despite the stirrup, mounted shock troops did not become the rule in the eighth century.

Cultural, economic, and environmental factors all interact with technology to affect the rate of historical change. The African Mbuti, for example, satisfied with their traditional hunting methods, rejected the use of guns for hunting despite clear advantages over nets and spears, while the differing circumstances of the Eskimos led them to accept guns.

1.9 Beginnings of the Technological Revolution in Feudal Europe

Toward the end of the ninth century, signs appeared that a new period of technological development was about to begin in Europe. The political disarray of constant invasions, marauding, and remapping of boundaries gave way to greater stability as central governments established authority. In 789, Charlemagne issued edicts to establish schools for the largely illiterate clergy. A ninth century mini-Renaissance produced educational advances, including adoption of a legible script and restoration of a good Latin, but with little or no scientific content (Beaujouan, 1963, p. 471).

Increasing artisan and merchant activity fueled rapid expansion of urban commercial centers in the tenth to thirteenth centuries. The growing use of metals—silver, gold, copper, lead, tin, and iron and steel—put mining, metallurgy, and metalworking in the center of industrial activity and trade. Grain mills powered by water wheels had already appeared throughout Europe, not all concentrated in the towns. They became a traditional source of revenue for the lords, who required that the peasants take their grain to be ground at the lord's mill. The Domesday Book, an English economic census completed in 1086, reported 5624 water mills in 3000 communities south of the Trent and Severn (cited by Forbes, 1956, p. 611). Water-powered cloth-fulling and grain mills and forges (many in monasteries) began in the mountainous regions of southern and central Europe, later spreading into France and northern Europe. By the fourteenth century, water-powered hammers, stamps, and bellows were fairly general in the entire Alpine region. "Such machines," writes Forbes, "entailed heavier capital expenditure than formerly, and the rise of western capitalism was largely tied up with the development of mining and metallurgy." The feudal lords promoted the rise of industrial–commercial towns (pp. 62–69). Manufacturing was at the heart of the urban economy, and initially a symbiotic relationship existed between the social environment of urban industry and trade and the feudal agrarian economy. Although more or less self-governing, the towns were still under the aegis of the feudal lords, who cornered markets or scooped off profits from long-distance trade and obtained their domestic or imported luxury goods from the towns. Production was regulated by the artisan guilds, which emerged in the tenth century.

The first great mathematician of the thirteenth century, and perhaps the most outstanding in the Middle Ages, Leonardo Fibonacci (1170–1250, also known as Leonardo Pisano or Leonardo of Pisa), came from the rising commercial social stratum (Smith, 1923, pp. 214–218). Italian merchants acquired computational methods from their North African counterparts. Double-entry bookkeeping first appeared in 1340 in Genoa (Beaujouan, 1963, p. 516). In parallel with the commercial interest in practical mathematics, the growing application of a variety of mechanical devices to manufacturing contributed to an interest in the representation of spatial relationships. The contributions of Leonardo da Vinci in this connection are well known.

A close relationship thereby developed between the artists and the artisans. This bond among those engaged in the arts and crafts continued to spread through Western Europe in the fifteenth and sixteenth centuries. Consisting of artist-engineers, painters, sculptors, architects, builders of locks and canals and fortifications, and toolmakers, this group of practitioners, who themselves did not study questions of a scientific character, stimulated others to do so. A related group included the instrument makers (who supplied navigators, geodesists, astronomers, and musicians with the instruments they needed), the clock-makers, the cartographers, and the military technicians. "Empirical knowledge did not have to be sought deliberately, but arose naturally from the pursuit of technical trades; the waiting was only for theoretical reflection, which, however, was helped by the fact that there is no single department of physics which calls more urgently for mathematical treatment and lends itself more naturally to it than mechanics. The first essential element of classical physics, the mathematical approach, thus came into its own as spontaneously as the second, the empirical foundation" (Dijksterhuis, 1969, p. 243).

The center of scholarship initially rested in the monasteries, later moving into the universities, themselves Church institutions. Interest in mathematics developed among Church scholars because the Church calculated the date of Easter by combining the lunar cycle with the solar year on the basis of the Ptolemaic geocentric model of the heavens. In the tenth century, the French monk Gerbert of Aurillac (who later became Pope Sylvester II) brought Hindu–Arabic numerals to Christian Europe from Spain. Gerbert also greatly simplified the methods for arithmetic computations. Within the Church, however, conflicts arose between the traditions of classical scholarship and the new approaches needed for the burgeoning crafts and trade.

The English monk Roger Bacon (1214?–1294) was among the first Europeans to recognize the need for empirical verification of theoretical knowledge: "Reasoning does not make a conclusion certain, unless the mind discover it by the method of experiment." He also recognized the need to deal with nature quantitatively: "Nature cannot be known without mathematics" (Beaujouan, 1963, p. 492).

Francis Bacon (1561–1626) may be considered the first to have seen a direct dependence of science on technology. He called for scholars to place technology on a scientific foundation by familiarizing themselves with empirical knowledge accumulated in the crafts (Dijksterjuis, 1969, p. 401). The encyclopedia Bacon proposed was accomplished by Diderot and d'Alembert. By including descriptions of all tools and the methods of their fabrication and utilization, it conferred the status of knowledge and social dignity on that which up to that time was a matter of utility, of spontaneous empiricism (Fontenay, 1982).

1.10 From Feudalism to the Industrial Revolution in Europe

In the historical shift that initiated the most rapid growth in material production in history, the interaction among factors that stimulate or are stimulated by technological development may be seen clearly. Small-scale manufacturing continued to grow throughout Europe and North America, but developments in Britain took on a distinctive character that led to what became known as the Industrial Revolution.

In the middle of the eighteenth century, Britain's manufacturing and commerce, especially the slave-based sea trade in colonial products such as tea, sugar, and tobacco, were prospering. Perhaps one-tenth of all fixed investments other than real estate were in the maritime trade. Britain had several times the number of ships as France, its nearest competitor. The rural economy of Britain developed into a substantial system of cash incomes and sales; tenant farmers employed large numbers of farm laborers, unlike the predominantly subsistence-farming peasant households that persisted on the continent (Hobsbawm, 1968, pp. 10–19). Manufacturing spread throughout rural areas, the typical workers being village artisans or smallholders in their cottages making specific products, such as cloth, hosiery, or metal goods. With rapid urbanization and wood being in short supply (Gille, 1986, p. 590), the growth in the production of coal, primarily for heating and cooking, formed a substantial preindustrial base for subsequent industrial expansion (Galloway, 1969, Chap. 4–5). Coal was the first industrial commodity to be produced at a rate of millions of tons a year (Hobsbawm, 1968, p. 31). Unlike on the continent, the large landowners, rather than the monarch, drew royalties from mining operations and had a direct interest in supporting village manufacturing, as well as canals and turnpikes to facilitate the sale of the goods produced. The British economy was thereby in a unique position to benefit from a major expansion in goods produced cheaply enough to be purchased by the middle classes, instead of the production of luxury goods for the wealthier strata as on the continent.

The main advantage of the preindustrial home market was its size and steadiness. "The domestic market may not have provided the spark, but it provided fuel and sufficient draught to keep it burning" (Hobsbawm, 1968, p. 32). The spark, according to Hobsbawm, was the export market. "Between 1700 and 1750 home industries increased their output by seven per cent, export industries by seventy-six percent; between 1750 and 1770 (which we may regard as the runway for the industrial 'take-off') by another seven per cent and eighty per cent respectively. Industrial production increased—but foreign demand multiplied."

Cotton manufacture was responsible for the bulk of the export trade. All raw material had to be imported, and by the end of the eighteenth century, about two-thirds of the industry's output was exported. Prior to the 1760s, the spinning wheels could not supply enough yarn to keep pace with looms now employing the flying shuttle, introduced in 1733. A breakthrough came with the spinning jenny in 1765. Further improvements in spinning followed, leading finally to the fully mechanized spinning mule in 1777 (Gille, 1986, p. 624). Subsequent improvements in the machines increased the yarn yield per worker 100-fold in relation to the spinning wheel. As the textile industry expanded, its machinery was increasingly driven by the steam engines made possible by James Watt's invention, although water power continued to be used when convenient. The widespread use of steam power made possible industrial activities in regions in which water power was not available. The entire technological basis of British manufacturing thereby shifted from water and wood to iron, coal, and steam.

The rapid advances of science and technology at this time were quickly absorbed by the various branches of manufacturing. The relatively simple technology required for cotton production, however, propelled the Industrial Revolution forward at an unprecedented rate without much interaction with science. The majority of the technological innovations were made by people closely involved with production—foremen, entrepreneurs, and skilled workers (Gille, 1986, p. 594). Scientific advance in eighteenth-century Britain came largely from societies established by craftsmen and industrialists, such as the Lunar Society of Birmingham and the Manchester Literary and Philosophical Society (Hill, 1967, p. 198). James Watt was an instrument maker at the University of Glasgow, where he learned about Joseph Black's discoveries of latent heat. Bernoulli's and Euler's work on ships no doubt influenced ship construction. It should not be forgotten, however, that Carnot's theoretical introduction of the cyclical engine came much later, in 1824, while the distinction between force and energy was not theoretically explained until the 1860s. In general, the sciences during the early years of the Industrial Revolution did not lead to direct results in design, but pinpointed quantitative relations that inventors and innovators could take into account in their empirical work.

The immediate result of the Industrial Revolution was great economic benefit for entrepreneurs and harm for workers. Mechanization of the textile industry threw large numbers of weavers out of work. The shift of the workplace into factories from rural cottages and homes destroyed this source of supplemental income. Wages were kept at subsistence levels. The shift to the cities was precipitous. In 1750, only London and Edinburgh had populations over 50 000. In 1851, there were twenty-nine cities with populations that big and nine with over 100 000. "And what cities! It was not merely that smoke hung over them and filth impregnated them, that the elementary public service—water supply, sanitation, street cleaning, open spaces, etc.—could not keep pace with the mass migration of men into the cities, thus producing, especially after 1830, epidemics of cholera, typhoid, and an appalling constant toll of the two great groups of urban killers—air pollution and water pollution, or respiratory and intestinal disease" (Hobsbawm, 1968, p. 67). As the industrial economy reached new heights quantitatively, qualitative changes in market relationships ushered in an era of cyclical economic crises of unprecedented depth. The first deep crises occurred in the 1810s, in response to which displaced textile workers directed their anger against the machines that replaced them. Crises occurred again in the 1830s, and again in the 1840s, leading to large-scale bankruptcies and high unemployment. The conditions of the laboring population were so harsh that advocates of slavery in North America argued that chattel slavery in the U.S. South was far more humane than wage slavery in Britain. In the 1850s, economists began to conclude that higher wages would put more money in circulation and thereby stimulate industrial production. Growing restiveness among the workers and their reliance on trade unions finally led to higher wages and reform in labor legislation.

Although the Industrial Revolution replaced many skilled workers with unskilled workers, a need arose for a literate segment in the work force able to read and follow somewhat complex instructions, read technical drawings, and perform calculations. Public primary education was therefore initiated in the first half of the nineteenth century in Britain, France, and the United States.

1.11 From the Industrial Revolution to Modern Technology

In the wake of the Industrial Revolution, manufacturing moved into a central position in the dynamics of the national economies of Western Europe, Canada, and the United States, spreading to Japan late in the nineteenth century, bringing with it an infrastructure of transportation, raw material sources, fuel, and finance. Raw materials were secured domestically, from other industrialized lands, or from colonial and semicolonial possessions. The role of the national governments in securing, maintaining, and regulating these infrastructures increased greatly. As the size of the enterprises increased, the pattern of ownership also changed from single owners and partnerships to the corporate form. Toward the end of the nineteenth century, investment banks formed, acquired, or merged with industrial firms. These developments took place simultaneously in the United States, Britain, France, Germany, and Japan; the United States overtook Britain in industrial potential around 1900. Competition for sources of raw materials and foreign markets grew intense. The economic, political, and military subjugation of almost all of Africa, Asia (outside of Russia and Japan), Central and South America, the Pacific, and the Caribbean was essentially completed by 1910 in an atmosphere of great political and military rivalry among the industrial powers, a rivalry that sharpened the interest of national governments in military technology.

Industrial managers near the end of the nineteenth century were not as closely connected personally with technological development as formerly. Research and development laboratories arose within the individual enterprises (see PHYSICS AND TECHNOLOGY, HISTORY OF). The universities remained centers of theoretical research and of the experimental research that necessarily preceded it.

The introduction of the assembly line at the Ford factory in 1913 based on the time and motion studies of Frederick W. Taylor and Frank and Lillian Gilbreth had a profound social and economic impact. The organization of work on the assembly line broke up complex manufacturing operations into simple jobs that could be performed by a worker with little or no training. By 1926, 79% of the Ford workers needed less than one to no more than eight days training (Gille, 1986, p. 1017). The economic success of this method of production led to its use wherever technologically feasible. The dehumanizing reduction of labor to essentially mindless tasks, already foreshadowed in the textile mills, reduced workers to appendages of machines, alienating their spirits from their hands and from the object of production. Assembly-line production continues to this day, but with increasing involvement of the workers in setting up and rotating tasks. Many big manufacturing plants are now experimenting with abandoning assembly lines altogether.

2. ORIGINS OF PHILOSOPHY OF SCIENCE AND PHILOSOPHY OF TECHNOLOGY

From the seventeenth century on, scientific research tended to concentrate in academic institutions, while technological research and development was conducted largely in private industry, but this has become less true since the mid-twentieth century. As the time between scientific discovery and technological innovation decreased, technology became increasingly dependent on science for its innovations. Toward the end of the twentieth century, the dominant view is that science is the primary stimulant to technological development and not the reverse.

Prior to the eighteenth century, no fundamental distinction was made between scientific knowledge and philosophy. The advances in physics, especially Newton's mechanics, stimulated the separation. The theoretical structure of Newton's mechanics is that of the hypothetico-deductive model employed by Euclid. The clearly empirical character of Euclid's postulates and axioms had not been understood, even in Newton's time. The simplicity of the axioms and their fewness of number continually stimulated the quest to find a logical basis for their origin, somewhat reminiscent of today's quest for a grand unification.

Kant (1966) proposed a compromise. Certain kinds of knowledge need not be demonstrated; logic, geometry, and time, for example, exist prior to and independently of material things. Knowledge about the material world is acquired through our senses. How does one identify *a priori* knowledge, however, since if it cannot be proved, it must be asserted? John Stuart Mill straddled the issue: "Axioms are but a class, the most universal class, of inductions from experience; the simplest and easiest cases of generalization from the facts furnished to us by our senses and by our internal consciousness" (Mill, 1904, p. 187).

Two logical problems immediately arise with what on the surface appears to be the experimental or empirical basis of Mill's view. One is the logical basis of facts furnished by our internal consciousness—this can only be a return to the Kantian *a priori*. The other is the source of the sensual stimuli that produce the "facts furnished to us by our senses."

Francis Bacon had already recognized the need to put technology on a scientific basis by incorporating technological experiences into scientific investigation. The early nineteenth-century philosopher Hegel (1969) saw (without pursuing its logical consequences in detail) that properties of space and time are not independent of, but originate with, matter. Riemann, confronted first with the non-Euclidian geometry of Lobachevsky and Bolyai, and then with his own, concluded that only experiment would establish the validity of a geometrical system.

Though known, perhaps, as the leading nineteenth-century proponent of philosophical idealism, Hegel uncharacteristically took a very earthy approach to technology by asserting the importance of means in relation to the ends: "The *means* is *superior* to the *finite* ends of *external* purposiveness: the *plough* is more honourable than are immediately the enjoyments procured by it and which are ends. The *tool* lasts, while the immediate enjoyments pass away and are forgotten. In his tools man possesses power over external nature, even though in respect of his ends he is, on the contrary, subject to it" (Hegel, 1976, p. 747). Thus, in Hegel's view, the ends are established through human reason, but this reason is transcended by the technology created to procure the ends and which can continue to be employed for other ends. Hegel here begins what might be called the praxis philosophy, according to which practical activity of humans serves as the ultimate source of their technology and science.

Karl Marx, the first to incorporate technology philosophically into a world outlook, never explicitly identifies the philosophy of technology in what came to be known as his materialist conception of history. In Marx's scheme, the primary motor of societal evolution is the development of the *productive forces*, by which he means natural resources, raw materials, implements and tools, and human labor. Implied, although not mentioned explicitly, is the technological and scientific knowledge associated with productive processes. Technological development, that is, the development of the productive forces, proceeds under more or less stable relations of production, which Marx equates with the dominant form of property relations associated with the use of the productive forces. He then projects that an incompatibility ultimately arises between the productive forces and the property relations so that property relations cannot be kept stable. Social revolution, to Marx, is simply the resolution of this incompatibility through a relatively rapid change in the dominant property relations. Thus, Marx (1987) explains the evolution of society from its early forms of communal property through slavery, feudalism, capitalism, and ultimately communism. A critical point of Marx's theoretical position is that "it is not the consciousness of men that determines their existence, but their social existence that determines their consciousness." The technological implications of this are profound, because production relations are not limited to property relations. People engaged in production also have technological relations of production to the tools and the object of production, and to one another when several people work together. It is therefore the productive activity of human beings that ultimately shapes their understanding of their relations to the material world around them and of the entire complex of social relations that they have to one another.

Such a view of the relationship between existence and consciousness leads Leonard Goldstein to attribute the changes in the way the Renaissance artists viewed space to the need to understand the spatial relationships for the design and use of large machines driven by natural power and for manufacturing processes involving the coordinated labor of a group of workers. Goldstein notes that linear perspective depicts the view grasped by the eye of an individual observer. He identifies the source of this one-point view as a reflection of the individualism that is necessarily shaped by vigorous activity of the ownership of a successful enterprise (Goldstein, 1988).

A propensity to mechanical innovation has sometimes been seen as a character trait common in some nations; Marx, however, would look rather to the conditions motivating people to engage in production as the stimulus for innovation. One source of motivation is production to meet one's own needs for direct consumption; another is the production of commodities for exchange for other commodities for direct consumption. Still another, and this is the principal characteristic of the capitalist economy, is the production of commodities for exchange at a profit on a continuing basis. Since the goal is accumulation, not consumption, the appetite for profit is unsatiable and technological improvements will help feed it. Marx regards technological development as intrinsically neutral, but leading to property relationships that provide the social dynamic. "The hand-mill gives you society with the feudal lord; the steam-mill society with the industrial capitalist" (Marx, 1935, p. 355). Fundamentally, Marx was not a critic of the application of technology or of the various types of social inequality that he associated with given levels of technological development. His radicalism was associated with his conviction that the level of the productive capacity under capitalism had reached a point where the market relations produced a conflict between distribution and consumption that manifested itself in periodic economic crises, that these crises would become deeper and deeper, and that public ownership, as distinct from private ownership, of capital resources was the next stage of the socioeconomic evolutionary process that began with the appearance of the surplus product. He is often incorrectly characterized as a technological determinist. If this were the case, he would not have had to assume the role of a political radical. He did so because he rejected Laplacian determinism in regard to the human spirit, so that he did not consider that the radical social consciousness necessary to bring about this next social transformation would emerge spontaneously, but required a political movement to bring it about, even if history was on its side.

In 1877, the term *philosophy of technology* appears first in the title of a German work, *Grundlinien einer Philosophie der Technik* by Ernst Kapp, who was strongly influenced by Hegel's concept of history being the realization of the Idea. He wrote that human beings using their first tools, their hands, in producing their first works became historical human beings progressing toward self-consciousness (Kapp, 1877, p. 39).

The goal of technology as a field of human endeavor is, like science, not the same for all its practitioners. For some, the "tinkerers," it can be an end in itself, just as mathematicians will sometimes say that they are interested only in the beauty of the logic and do not care if their work has any practical use. On the other hand, it is rare for successful tinkerers to neglect patenting the results of their tinkering.

The empiricist view of John Stuart Mill was further refined by the Austrian physicist Ernst Mach and, subsequently, further developed by a group of scientists and philosophers known as the Vienna Circle into what became the most influential school of philosophy of science in the first half of the twentieth century, *logical positivism*. According to this view, in its application to physics, "quantities such as the charge, temperature, mass, and length of a body are *defined* as the objective results of certain prescribed operations that can be carried out in the laboratory. . . . This *operational viewpoint* is the basis of this philosophy." Further, "physical *laws* are relationships between operationally defined quantities that always occur when certain experiments are performed" (Halliday, 1955, pp. 4–5).

In the logical positivist view, the question of the source of our sense impressions or observations is meaningless, since there is no operational way of establishing existence of objects external to us. The attractiveness of logical positivism to the scientist is its apparent rejection of speculative activity. The practical task of scientific theory is to provide correlations between observational data.

3. CONTEMPORARY VIEWS ON PHILOSOPHY OF TECHNOLOGY

By the middle of the twentieth century, the logical problems connected with operational definitions began to lead to their demise. One objection, for example, was that an operational definition of a field would require an infinite set of measurements. Another objection was that no two instruments are identical and no two measurements by the same instrument of infinite precision would be identical; therefore each act of measurement would produce its own unique operational definition. It can be argued that operational definitions are means for standardizing scales of measurement of physical properties, the meaning of which arise from the laws that embrace them.

With the collapse of logical positivism, scientific realism rose to the fore. Realism itself is not one particular philosophical current but a category of approaches to what is often called the mind–body problem. Realism, in general, recognizes the relative independence of our ideas from that which exists outside our minds. At the same time, realism does not necessarily exclude the dependence of one on the other. Some realist philosophies may give primacy to the world of ideas in the sense that it is ideas that ultimately engender the world outside our minds, while others can argue that it is the world outside our minds that ultimately gives rise to our ideas, allowing for various degrees of this dependence.

According to Don Ihde, philosophy of science has been historically linked to the three prejudices that he associates with a *theory bias* that affects the attitude toward technology. The first goes back to Greek antiquity, according to which "the aim of highest humanity was, in effect to achieve the deepest knowledge and, once attained, meditate upon it." The second is more recent and holds that modern technology is both essentially different from all ancient or traditional technologies, and therefore in some fundamental sense, "better." The third prejudice is the belief that modern technology differs from all previous technology in that it is derived from modern science (Ihde, 1993, p. 20). Ihde cites the Platonic view, reflected by Aristotle, which holds that anything associated with the body and the material is inferior to the realm of ideas and forms, the latter being associated with the mind or soul.

Philosophical attention to technological concerns grew slowly during the first half of the twentieth century. Among the important commentators on the question were Friedrich Dessauer (1927), Ernst Jünger (1932), and José Ortega y Gasset (1941). Renewed interest developed after World War II, stimulated at first by the social, political, and economic concerns resulting from nuclear weapons. Simultaneous with the introduction of nuclear technology was the beginning of the computer age, large-scale application of electronic computers having begun during World War II in connection with coupling artillery (naval and land) to radar.

One work that attracted immediate attention was an essay by Martin Heidegger (1953), "The Question Concerning Technology." Heidegger begins his discussion of technology by pointing to the inadequacies today of the earlier instrumental and anthropological definitions of technology, the instrumental definition being the means to achieve particular ends, while the anthropological definition refers to the human activities employing these means. He thinks that it is necessary to go beyond these definitions to arrive at a deeper understanding of the relationship between humans and the technology they employ. This is because as humans attempt to control technology and bring it spiritually within their power, it continually slips through their fingers no matter how hard they try. It is necessary, above all, to distinguish modern technology from past technologies. What is fundamentally different in modern technology is that it views both nature and human labor as a continually available and extendible standing reserve whose possibilities confront humanity as a challenge for further disclosure or uncovering. Modern technology challenges and stockpiles the sun's warmth stored in coal "for heat, which in turn is ordered to deliver steam whose pressure turns the wheels that keep a factory running" (Heidegger, 1977, p. 15). The windmill of classical technology is of a different character. "Its sails do indeed turn in the wind; they are left entirely to the wind's blowing. But the windmill does not unlock energy from the air currents in order to store it" (1977, p. 14). "The gathering together of that setting-upon that sets upon man, i.e., challenges him forth, to reveal the real, in the mode of ordering, as standing reserve" is the way of revealing "which holds sway in the essence of modern technology and which is itself nothing technological" (p. 20).

Modern and classical technology are both ways of disclosure (or uncovering), but only in modern technology are humans challenged to this uncovering. This already occurred with the development of physics as an exact science, on which modern technology is based (p. 21). Heidegger, however, does not see science as the primary stimulant of technology. He points out that modern physics involves experiment and is therefore dependent on the development and application of technical apparatus. Since the essence of technology lies in the uncovering process that constitutes technological activity, technology must be seen as applying natural science, rather than being considered as applied natural science (1977, p. 14).

Heidegger sees the main danger from modern technology not so much in the threats of environmental disasters resulting from its applications but in human beings surrendering themselves to what is technologically revealed and being absorbed into the revealed technology, all the while believing that they are the masters and in full control of the earth. This is the illusion generated by looking at the technological world merely as a collection of objects and means to be set in motion by human activity, and it is this viewpoint that threatens to reduce humans themselves to a mere standing reserve. It is not enough to acquaint oneself with external objects and the relations among them, including their relationships to humans themselves. It is necessary to be conscious of how our own being is grounded in the complex of relationships among ourselves and the objects around us. An airplane is not merely an object of a given shape and construction. When revealed, it is a standing reserve for the transportation of people and goods, and to fulfill this role, it must be fully appropriate in structure and readied for flight. It needs airports from which to take off and land, and these airports must be planned for and constructed by persons who do not necessarily travel in the planes, etc. Heidegger argues that this type of understanding is necessary before we can do anything to dispel the illusion that humans are the masters of the world, and uncover the paths to reverse the domination that technology maintains.

By his repeated stress on the concept of uncovering, Heidegger in essence projects an *a priori* character onto that which can be uncovered. Moreover, he regards continual involvement in technological uncovering as human destiny, but not in a deterministic way. The danger of technology that he sees is therefore that people see themselves challenged to uncover technology in a one-sided way, a way that contributes to their virtual enslavement by technology and obliterates any free will. He sees human destiny to be the confrontation of the challenge for technological uncovering and thus has to reject the possibility of abandoning technological development. Rather, he calls for the implementation of human free will to proceed along a path of technological uncovering "that in no way confines us to a stultified compulsion to push on blindly with technology or, what comes to the same thing, to rebel hopelessly against it and curse it as the work of the devil" (1977, p. 26). Heidegger does not, however, see this enactment of human "free claim" as a purely human undertaking. That which is there to be uncovered was not put there by human beings. This last point underscores the essentially mystical character of Heidegger's philosophy, since the human response to an objectively existing situation will be indifferent to the process by which this objective situation arose if no knowledge of its supposed purpose, assuming there was one, is available.

Heidegger's *a priori* concept of technology was disputed by Herbert Marcuse in *One-Dimensional Man* (1964). "The technological *a priori* is a political *a priori* inasmuch as the transformation of nature involves that of man, and inasmuch as the 'man-made creations' issue from and re-enter a society ensemble" (p. 154). Marcuse saw the inherent rationality of modern technology as a source of human enslavement. "Scientific management and scientific division of labor vastly increased the productivity of the economic, political, and cultural enterprise. Result: the higher standard of living. At the same time and on the same ground, this rational enterprise produced a pattern of mind and behavior which justified and absolved even the most destructive and oppressive features of the enterprise. Scientific–technical rationality and manipulation are welded together into new forms of social control" (1964, p. 146). Marcuse criticizes the harm done by relying on positivist criteria in assessing technological performance. What is quantitatively measurable is profit and efficiency, not the toll in human suffering.

The views of another important writer on technology, Jacques Ellul, continue to be influential today. According to Ellul's summary of his own views (1983), first expressed in his book *The Technological Society* (1964), technology replaces nature by creating the specific milieu in which humans must exist. Technology thus becomes artificial, autonomous with respect to values, ideas, and the state, self-determining, growing in a process that is causal, but directed to ends. "It is incorrect to say that economics, politics, and the sphere of the cultural are influenced or modified *by* Technique; they are rather situated *in* it, a novel situation modifying all traditional social concepts. Politics, for example, is not modified by Technique as one factor among others which operate upon it; the political world is today defined through its relation to the technological society. Traditionally, politics formed a part of a larger social whole; at the present the converse is the case" (1983, p. 86). Ellul argues that technological progress is in itself neither good nor bad. "The fact is that, viewed objectively, technological progress produces values of unimpeachable merit, while simultaneously destroying values no less important. As a consequence it cannot be maintained that there is absolute progress or absolute regress" (p. 98). In summary, Ellul maintains that all technological progress exacts a price, raising more problems than it solves, pernicious and generally unforeseeable effects being inseparable from its favorable effects. The technological society cannot not be a humanist one. There is a contradiction between technological perfection and human development because technological perfection "is only to be achieved through quantitative development and necessarily aims exclusively at what is measurable. Human excellence, on the contrary, is of the domain of the qualitative and aims at what is not measurable" (p. 90). Ellul sees no escape from this situation. "There is no possibility of turning back, of annulling, or even arresting technical progress. What is done is done. It is our duty to find our place in our present situation and in no other" (p. 91).

The most widely read contemporary philosopher of technology is Landon Winner (1977, 1986), who also focuses on the theme of the autonomous character of technology, viewing it as self-moving and self-evolving according to its own laws. Megatechnical systems like the electric power system dominate humanity by exerting demands for their own maintenance. Humans can monitor and guard the system against destruction and deterioration, or introduce innovations to improve its efficiency, but are unable to replace it by another. "We indeed 'use' telephones, automobiles, electric lights, and computers in the conventional sense of picking them up and putting them down. But our world soon becomes one in which telephony, automobility, electric lighting, and computing are forms of life in the most powerful sense: life would scarcely be thinkable without them" (1986, p. 11).

In the 1950s and 1960s, the environmental hazards associated with the unrestricted development of nuclear and chemical technologies became evident. Problems also arose in connection with the disposal of nuclear and toxic chemical waste. The growing quantity of nonbiodegradable plastic waste created additional problems. The growing integration of electric power systems into national and even international networks generated the threat of a national and international catastrophe from a major breakdown in the electrical grid. With the development of supertankers, ecological disasters produced by oil spills became a reality. Standing above all other threats facing humanity was that of nuclear war. These developments stimulated a number of philosophers to focus their attention on the scientific and societal implications of contemporary technological development. Philosophy of technology, which hitherto was generally limited to sporadic study by philosophers primarily interested in other areas, emerged as a field of specialization within philosophy, with international organizations and publications.

One set of strategies put forward, according to Ihde (1993, pp. 119–27), to deal with environmental issues involves returning to some previous lower level of technology or to minimalist technologies. Ihde argues that such a strategy would produce social disruption as well as an immediate drastic worldwide decline in population comparable to times of vast warfare and plagues.

An *ameliorationist* set of strategies accepts high-technology developments, but turns them into regional and decentralized modes for such activities as food and energy production. Ihde believes that such strategies can only have limited success, since the rate of technological change outpaces the possibilities of such alternative strategies.

The third set of strategies Ihde calls *hi-tech reformist*. Solutions to the problems of continued technological development are to be provided by miniaturization, clean-up technologies, and alternative energy production. These strategies require a centralized administration and planning on an international scale and continual technological development of "clean" solutions. Efforts to deal with global problems must be approached with a multicultural perspective and, according to Ihde, a holistic view of the interconnections of technology, environment, and nature.

4. UNRESOLVED PROBLEMS OF TECHNOLOGICAL DEVELOPMENT

Modern technology is now capable of eradicating human life on earth through a number of separate paths. The most obvious has been through nuclear war. Although the principal sources of political tension that posed this threat were removed at the beginning of the 1990s, production and maintenance of nuclear weapons continues. New sources of political tension may emerge and pose the threat of nuclear war once again. Further development of chemical and biological weapons may bring them to the point that they pose the same scale of danger to human existence as nuclear weapons.

Scientists in the United States are divided over the question of nuclear power, although a majority of physicists probably believe that current technologies carry an unacceptable risk of nuclear disaster. The potential danger from commercial application of genetic engineering, which began on a significant scale in the 1980s, is twofold. Unanticipated effects of scale may occur—that is, harmful effects appearing only when artificially created genetic strains are available in much larger quantities and concentrations than in the research and development stage. The second danger is of accidental release of harmful organisms. Richard Leakey has pointed out how fortunate it is for humanity that the AIDS virus is not transmitted through sneezing. What if an accidental release occurred of some potentially lethal experimental microorganism capable of such rapid transmission?

The most commonly acknowledged threat to human existence at present is from industrial chemicals. International agreements designed to halt the release of chemical agents that destroy ozone in the upper atmosphere are now in effect. No international agency, however, has the authority to act independently to meet other dangers as they arise; each hazard requires a separate international agreement. Differences in the assessment of dangers arising among different countries, based on political and economic pressures rather than scientific considerations, may prevent averting the ultimate disaster.

Between 1700 and the end of World War II, industrialized countries pursued economic policies denying the export of technology. The breakup of the colonial systems, increasing resistance in the Third World to such imperialistic policies, and export of capital seeking lower wage scales are now bringing technological development to previously underdeveloped regions. Global resource and environmental crises loom, such as the destruction of the Brazilian rain forests. As the peoples of these countries aspire to the level of commodity consumption that they see in the industrialized countries, the potential for international conflicts is obvious.

An area of crisis limited to the industrialized countries is the growth of unemployment as a result of automation, robotization, and computerization of industrial and office work. The still-growing replacement of metal and wood by the less labor-intensive polymers in manufacturing and construction exacerbates this problem. In the early 1990s, over a period that is not usually characterized as a time of deep economic crisis, unemployment in the industrialized countries of Europe was often above 10%, while in the United States (with less strict statistical measures for unemployment), unemployment held at about 6%. While the financial consequences of this unemployment can be offset in part by social insurance and social welfare programs, the shifting of skilled labor into unskilled labor at lower pay as well as threatened future displacement for those not yet affected has spread insecurity among employees in all income levels.

The source of these problems lies outside of technology itself, in the driving force behind the application of technology. The application of technology to production is not stimulated directly by the need for products, but by the profit that can be made from the sale of products that satisfy real or imagined needs. The "unseen hand" that classical laissez-faire economics saw balancing economic life seems no longer to function. Various degrees of economic regulation and intervention by the state have been introduced into otherwise free-market economies. Marx suggested that such measures would not eliminate the business cycle with its growing swings of prosperity and depression and that the solution would lie in replacing production for profit by production for need, displacing private ownership of the capital resources by public ownership and planned utilization of these resources. The effort to establish such a system on a wide scale beginning in the U.S.S.R. in 1917, extending to Eastern Europe and China and some other countries in the 1940s and later, collapsed completely in the U.S.S.R. and Eastern Europe. It has survived in modified form in the less-industrialized countries of China, Vietnam, North Korea, and Cuba, where various mixtures of market and planned economies are being tried. Chinese, Vietnamese, and Cuban leaders attribute the European collapse to the extreme bureaucratization of political and economic life and excessive centralization of the economic planning mechanisms. Whether they can produce different results is still to be seen.

As long as economies function within a capitalist framework, technological development must proceed, for the inanimate corporation or its animate principal stockholders cannot consume all profits from a successful year of activity, but must continually reinvest them on an expanding scale. Global environmental and resource problems must ultimately be solved by limiting reinvestments in material production to levels that can be technologically tolerated. Expansion of economic activity in directions that minimally absorb material resources offers at present a partial solution to the problem. Commodities and services that take largely nonmaterial forms, such as computer software and artistic production, are of this character. Perhaps the welfare of humanity in the immediate future lies in the expansion of the range of such activities.

Works Cited

Anderson, P. (1974), *Passages from Antiquity to Feudalism*, London: Verso.

Aristotle (1947), *Politics* and *Ethics*, in: R. McKeon (Ed.), *Introduction to Aristotle*, New York: Modern Library.

Arnaldez, R., Massignon, L. (1963), in: R. Taton (Ed.), *Ancient and Medieval Science: From Prehistory to AD 1450*, London: Thames and Hudson, pp. 385–410.

Beaujouan, G. (1963), in: R. Taton (Ed.), *Ancient and Medieval Science: From Prehistory to AD 1450*, London: Thames and Hudson, pp. 468–532.

Binford, L. R. (1980), *American Antiquity* **45**, 4–20.

Boswell, J. (1934), *Boswell's Life of Johnson*, Vol. 3, Oxford, U.K.: Clarendon.

Candland, D. K. (1987), in: G. Mitchell, J. Erwin (Eds.), *Comparative Primate Biology*, Vol. 2B, *Behavior, Cognition, and Motivation*, New York: Alan R. Liss, pp. 85–103.

Carlyle, T. (1921), *Sartor Resartus*, New York: Harcourt.

Coudart, A. (1991), in: S. A. Gregg (Ed.), *Between Bands and States*, Occasional Paper No. 9, Center for Archaeological Investigations, Southern Illinois University at Carbondale, pp. 395–420.

Darwin, C. (1981), *The Descent of Man and Selection in Relation to Sex*, London: John Murry.

Dessauer, F. (1927), *Philosophie der Technik: Das Problem der Realisierung*, Bonn: Friedrich Cohen.

Dijksterhuis, E. J. (1969), *The Mechanization of the World Picture*, London: Oxford Univ. Press.

Ellul, J. (1964), *The Technological Society*, New York: Knopf.

Ellul, J. (1983), in: C. Mitcham, R. Mackey (Eds.), *Philosophy and Technology: Readings in the Philosophical Problems of Technology*, New York: Free Press, pp. 88–105.

Espinas, A. V. (1897), *Les Origines de la Technologie*, Paris: F. Alcon.

Falk, D. (1993), in: R. L. Ciochon, J. G. Fleagle (Eds.), *The Human Evolution Source Book*, Englewood Cliffs, NJ: Prentice-Hall, pp. 61–70.

Ferré, F. (1988), *Philosophy of Technology*, Englewood Cliffs, NJ: Prentice-Hall.

Fontenay, E. de (1982), *Diderot, Reason and Resonance*, New York: G. Braziller.

Forbes, R. J. (1956), in: C. Singer, E. J. Holmyard, A. R. Hall (Eds.), *A History of Technology*, Vol. 2, Oxford, U.K.: Clarendon, pp. 41–80, 589–628.

Fried, M. H. (1967), *The Evolution of Political Society*, New York: Random House.

Galloway, R. L. (1969), *A History of Coal Mining in Great Britain*, Newton Abbot, Devon, U.K.: David and Charles Reprints.

Gille, B. (1986), *The History of Techniques*, New York: Gordon and Breach, 2 vols.

Goldstein, L. (1988), *The Social and Cultural Roots of Linear Perspective*, Minneapolis: MEP Publications.

Goodall, J. (1968), *Anim. Behav. Monographs* **1**, 165–311.

Gould, R. (1982), in: N. M. Williams, E. S. Hienn (Eds.), *Resource Managers: North American and Australian Hunter-Gatherers*, AAAS Selected Symposium No. 67, Boulder, CO: Westview, pp. 69–92.

Haas, J. (1987), in: J. Haas, S. Pozorski, T. Pozorski (Eds.), *The Origins and Development of the Andean State*, Cambridge, U.K.: Cambridge Univ. Press, pp. 1–4.

Halliday, D. (1955), *Introductory Nuclear Physics*, 2nd ed., New York: Wiley.

Haudricourt, A., Needham, J. (1963), in: R. Taton (Ed.), *Ancient and Medieval Science: From Prehistory to AD 1450*, London: Thames and Hudson, pp. 161–77.

Hegel, G. W. F. (1969), *Hegel's Science of Logic*; translation by A. V. Miller, London: Allen & Unwin; reprint (1976), New York: Humanities.

Heidegger, M. (1977), *The Question Concerning Technology and Other Essays*, translation by W. Lovitt, New York: Harper & Row.

Henry, D. O. (1989), *From Foraging to Agriculture*, Philadelphia: Univ. of Pennsylvania Press.

Hill, C. (1967), *Reformation to Industrial Revolution: A Social and Economic History of Britain 1530–1780)*, London: Weidenfeld & Nicolson.

Hilton, R. H., Sawyer, P. H. (1963), *Past & Present*, No. 24.

Hobsbawm, E. J. (1968), *Industry and Empire: An Economic History of Britain since 1750*, London: Weidenfeld and Nicolson.

Hood, B. (1993), in: R. L. Ciochon, J. G. Fleagle (Eds.), *The Human Evolution Source Book*, Englewood Cliffs, NJ: Prentice-Hall, pp. 319–331.

Hood, W. F. (1983), in: C. Mitcham, R. Mackey (Eds.), *Philosophy and Technology: Readings in the Philosophical Problems of Technology*, New York: Free Press, pp. 347–363.

Hooke, S. H. (1954), in: C. Singer, E. J. Holmyard, A. R. Hall (Eds.), *A History of Technology*, Vol. 1, Oxford, U.K.: Clarendon, pp. xlviii–lv.

Howells, W. (1993), *Getting Here: The Story of Human Evolution*, Washington, DC: Compass.

Ihde, D. (1993), *Philosophy of Technology: An Introduction*, New York: Paragon House.

Isaac, G. L., McCown, E. R. (Eds.) (1976), *Hu-*

man Origins: Louis Leakey and the East African Evidence, Menlo Park, CA: Benjamin.

Jünger, E. (1932), *Der Arbeiter: Herrschaft und Gestalt*, Hamburg: Hanseatische Verlagsanstalt.

Kant, I. (1966), *Critique of Pure Reason*, Garden City, NY: Doubleday.

Kapp, E. (1877), *Grundlinien einer Philosophie der Technik: Zur Entstehungsgeschichte der Kultur*, Braunschweig: G. Westermann.

Laitman, J. (1993), in: R. L. Ciochon, J. G. Fleagle (Eds.), *The Human Evolution Source Book*, Englewood Cliffs, NJ: Prentice-Hall, pp. 56–60.

Lilley, S. (1965), *Men, Machines and History*, London: Lawrence & Wishart.

Lovejoy, C. O. (1993), in: R. L. Ciochon, J. G. Fleagle (Eds.), *The Human Evolution Source Book*, Englewood Cliffs, NJ: Prentice-Hall, pp. 230–243.

Marcuse, H. (1964), *One-Dimensional Man*, Boston: Beacon.

Marx, K. (1935), *The Poverty of Philosophy*, in: E. Burns (Ed.), *A Handbook of Marxism*, London: V. Gollancz.

Marx, K. (1987), Preface to *A Contribution to the Critique of Political Economy*, in: *Karl Marx, Frederick Engels: Collected Works*, Vol. 29, New York: International Publishers, pp. 257–265.

Mill, J. S. (1904), *A System of Logic: Ratiocinative and Inductive*, New York: Harper.

Mitcham, C., Mackey, R. (Eds.), *Philosophy and Technology: Readings in the Philosophical Problems of Technology*, New York: Free Press.

Napier, J. (1993), in: R. L. Ciochon, J. G. Fleagle (Eds.), *The Human Evolution Source Book*, Englewood Cliffs, NJ: Prentice-Hall, pp. 61–70.

Needham, J. (1954–1967), *Science and Civilisation in China*, Cambridge, U.K.: Cambridge Univ. Press, 7 vols.

Needham, J. (1969), *The Grand Titration: Science and Society in East and West*, Toronto: Univ. of Toronto Press.

Ortega y Gasset, J. (1941), *Toward a Philosophy of History*, New York: Norton.

Patterson, T. C. (1991), *The Inca Empire: The Formation and Disintegration of a Pre-Capitalist State*, Oxford, U.K.: Berg.

Schmandt-Bessarat, D. (1992), *From Counting to Writing*, Vol. 1, Austin: Univ. of Texas Press.

Skelton, R. R., McHenry, H. M., Drawhorn, G. M. (1993), in: R. L. Ciochon, J. G. Fleagle (Eds.), *The Human Evolution Source Book*, Englewood Cliffs, NJ: Prentice-Hall, pp. 112–134.

Smith, D. E. (1923), *History of Mathematics*, Vol. 1, Boston: Ginn.

Tobach, E. (1993), in: L. W. Martin, K. A. Nelson, E. Tobach (Eds.), *Social-Cultural Psychology: Theory and Practice of Knowing and Doing*, Cambridge, U.K.: Cambridge Univ. Press.

Turnbull, C. M. (1983), *The Mbuti Pygmies*, New York: Holt.

Vernant, J.-P. (1957), *Rev. Histoire Sci. Appl.* **10**, 205–225.

White, L., Jr. (1962), *Medieval Technology and Social Change*, Oxford, U.K.: Oxford Univ. Press.

Winner, L. (1977), *Autonomous Technology: Technics-Out-of-Control As a Theme in Political Thought*, Cambridge, MA: Massachusetts Institute of Technology Press.

Winner, L. (1986), *The Whale and the Reactor*, Chicago: Univ. of Chicago Press.

Further Reading

Anderson, P. (1974), *Passages from Antiquity to Feudalism*, London: Verso.

Ciochon, R. L., Fleagle, J. G. (Eds.) (1993), *The Human Evolution Source Book*, Englewood Cliffs, NJ: Prentice-Hall.

Dijksterhuis, E. J. (1969), *The Mechanization of the World Picture*, London: Oxford Univ. Press.

Durbin, P. (1994), "Philosophy of Science, Technology, and Medicine," in: C. Mitcham, W. Williams (Eds.), *The Reader's Adviser*, 14th ed., Vol. 5, New York: Bowker, pp. 59–104. Contains valuable bibliography.

Ellul, J. (1964), *The Technological Society*, New York: Knopf.

Gille, B. (1986), *The History of Techniques*, New York: Gordon and Breach, 2 vols.

Heidegger, M. (1977), *The Question Concerning Technology and Other Essays*, translation by W. Lovitt, New York: Harper & Row.

Henry, D. O. (1989), *From Foraging to Agriculture*, Philadelphia: Univ. of Pennsylvania Press.

Hobsbawm, E. J. (1968), *Industry and Empire: An Economic History of Britain since 1750*, London: Weidenfeld and Nicolson.

Ihde, D. (1993), *Philosophy of Technology: An Introduction*, New York: Paragon House.

Marcuse, H. (1964), *One-Dimensional Man*, Boston: Beacon.

Mitcham, C., Mackey, R. (Eds.) (1983), *Philosophy and Technology: Readings in the Philosophical Problems of Technology*, New York: Free Press.

Mumford, L. (1934), *Technics and Civilization*, New York: Harcourt and Brace.

Needham, J. (1954–1967), *Science and Civilisation in China*, Cambridge, U.K.: Cambridge Univ. Press, 7 vols.

White, L., Jr. (1962), *Medieval Technology and Social Change*, Oxford: Oxford Univ. Press.

Winner, L. (1977), *Autonomous Technology: Technics-Out-of-Control As a Theme in Political Thought*, Cambridge, MA: Massachusetts Institute of Technology Press.

PHONONS IN CRYSTAL LATTICES

A. G. EVERY, *Physics Department, University of the Witwatersrand, Johannesburg, South Africa*

Introduction ... 439
1. **Theoretical Background** ... 440
1.1 Quantization; the Phonon ... 442
1.2 Statistical Properties of Phonons ... 443
2. **Phonon–Phonon Interactions** ... 444
3. **Scattering Methods for Determining Phonon Dispersion Relations** ... 446
3.1 Inelastic X-Ray Scattering ... 446
3.2 Coherent Inelastic Neutron Scattering ... 447
3.3 Helium-Atom and Electron Scattering; Surface Phonons ... 448
4. **Infrared Absorption and Light Scattering** ... 449
4.1 Raman Scattering ... 450
4.2 Brillouin Scattering ... 450
5. **Electron–Phonon Interaction** ... 451
5.1 Metals ... 451
5.2 Metallic Superconductors ... 452
5.3 Semiconductors ... 453
5.4 Insulators ... 453
6. **Phonon–Defect Interactions** ... 453
6.1 Local Modes ... 454
6.2 Electronic States ... 455
6.3 Spin-Lattice Relaxation ... 455
6.4 Mössbauer Effect ... 456
7. **Closing Remarks** ... 456
Works Cited ... 457
Further Reading ... 458

INTRODUCTION

Crystal-lattice vibrations are quantized in much the same way that electromagnetic waves are, with the counterpart of the photon being the *phonon*. As one might expect, there are a number of parallels between the two classes of phenomena. For example, the phonon energy flux radiated into the interior of a cold crystal from a heated surface varies as the fourth power of the absolute temperature, just as for blackbody electromagnetic radiation. The blue of the sky has its counterpart in that small crystal inhomogeneities give rise to Rayleigh scattering of phonons, which varies with the fourth power of the frequency. It is common for an atom undergoing a transition between two quantum levels to conserve energy by absorbing or emitting a phonon rather than a photon.

There are, however, limits to how far the analogy can be taken. In contrast to electromagnetic waves in vacuum, lattice vibrations have a lower bound to their wavelength, equal to twice the lattice spacing in principal crystallographic directions, and associated with this is an upper cutoff in the spectrum of vibrational frequencies. For wavelengths approaching this limit, there is pronounced dispersion. Nonlinearity is important for lattice vibrations, leading in particular to phonon–phonon scattering.

The main emphasis of this article is on the interaction between phonons and other excitations or particles, on how these interactions influence the physical properties of crystals, and on how the scattering of x rays, neutrons, and other particles can be used to determine phonon dispersion relations. The concept of the phonon is introduced in Sec. 1, which also contains a synopsis of the main *lattice dynamics* results that are drawn on in this article. For a fuller discussion of lattice dynamics, the reader is referred to LATTICE VIBRATIONS, STATISTICS OF (hereafter cited as LV). Subsequent sections deal with scattering and absorption processes involving phonons, γ rays, x rays, optical and infrared photons, neutrons, electrons, neutral atoms, and localized spin and other excitations. A major aspect of phonon physics that is touched on

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

only very briefly in the present article is *phonon transport*. This subject is treated extensively in LV and in HEAT PULSE PROPAGATION IN SOLIDS (HP).

1. THEORETICAL BACKGROUND

We start by recalling that in the harmonic approximation the Hamiltonian for the *vibrations of a crystal lattice* takes the form (LV)

$$H_L = \sum_{l\kappa\alpha} \frac{p_\alpha^2(l\kappa)}{2M_\kappa} + \tfrac{1}{2} \sum_{l\kappa\alpha} \sum_{l'\kappa'\beta} \Phi_{\alpha\beta}(l\kappa;\, l'\kappa') u_\alpha(l\kappa) u_\beta(l'\kappa'), \qquad (1)$$

where $u_\alpha(l\kappa)$ are the displacements of the atoms from their equilibrium positions, with α denoting the Cartesian component, l the cell index, and κ ranging over the r atoms in the unit cell; $p_\alpha(l\kappa) = M_\kappa \partial u_\alpha(l\kappa)/\partial t$ are the momenta of the atoms, where M_κ are the atomic masses; and $\Phi_{\alpha\beta}(l\kappa;\, l'\kappa')$ are the *atomic force constants*, given by the derivatives

$$\Phi_{\alpha\beta}(l\kappa;\, l'\kappa') = \frac{\partial^2 \Phi}{\partial u_\alpha(l\kappa) \partial u_\beta(l'\kappa')} \qquad (2)$$

of the lattice potential energy Φ at equilibrium. The equations of motion derived from this Hamiltonian are

$$M_\kappa \frac{\partial^2 u_\alpha(l\kappa)}{\partial t^2} = -\sum_{l'\kappa'\beta} \Phi_{\alpha\beta}(l\kappa;\, l'\kappa') u_\beta(l'\kappa'). \qquad (3)$$

Because of the *translational invariance* of the lattice, the normal modes can be taken as plane waves, and the displacements and momenta expressed in terms of *normal coordinates* $Q(\mathbf{k}j)$ and their *conjugate momenta* $P(\mathbf{k}j)$ thus:

$$u_\alpha(l\kappa) = \frac{1}{\sqrt{NM_\kappa}} \sum_{\mathbf{k}j} e_\alpha(\kappa|\mathbf{k}j) Q(\mathbf{k}j) e^{i\mathbf{k}\cdot\mathbf{x}(l)}, \qquad (4)$$

$$p_\alpha(l\kappa) = \frac{1}{\sqrt{NM_\kappa}} \sum_{\mathbf{k}j} e_\alpha(\kappa|\mathbf{k}j) P(\mathbf{k}j) e^{i\mathbf{k}\cdot\mathbf{x}(l)}. \qquad (5)$$

The factor N comes from invoking *cyclic boundary conditions* over a macrocell of $N = L \times L \times L$ unit cells. The sum extends over a regular grid of N values of the wave vector $\mathbf{k}$ within the first Brillouin zone and $3r$ values of the branch index j. The $e_\alpha(\kappa|\mathbf{k}j)$ for each value of $\mathbf{k}$ are the $3r$ normalized eigenvectors of the *dynamical matrix*

$$D_{\alpha\beta}(\kappa\kappa'|\mathbf{k}) = \frac{1}{\sqrt{M_\kappa M_{\kappa'}}} \sum_{l'} \Phi_{\alpha\beta}(l - l', \kappa;\, 0\kappa') \times e^{-i\mathbf{k}\cdot[\mathbf{x}(l)-\mathbf{x}(l')]}. \qquad (6)$$

These eigenvectors satisfy the linear equations

$$\omega^2(\mathbf{k}j) e_\alpha(\kappa|\mathbf{k}j) = \sum_{\kappa'\beta} D_{\alpha\beta}(\kappa\kappa'|\mathbf{k}) e_\beta(\kappa'|\mathbf{k}j), \qquad (7)$$

with the normal-mode frequencies $\omega(\mathbf{k}j)$ being determined by the secular equation

$$|D_{\alpha\beta}(\kappa\kappa'|\mathbf{k}) - \omega^2 \delta_{\alpha\beta}\delta_{\kappa\kappa'}| = 0. \qquad (8)$$

Since $D^*_{\alpha\beta}(\kappa\kappa'|\mathbf{k}) = D_{\alpha\beta}(\kappa\kappa'|-\mathbf{k})$, we may choose the phases of the eigenvectors such that

$$e^*_\alpha(\kappa|\mathbf{k}j) = e_\alpha(\kappa|-\mathbf{k}j). \qquad (9)$$

It follows from Eqs. (4), (5), and (9) and the requirement that $u_\alpha(l\kappa)$ and $p_\alpha(l\kappa)$ are real that

$$Q^*(\mathbf{k}j) = Q(-\mathbf{k}j), \qquad (10)$$

$$P^*(\mathbf{k}j) = P(-\mathbf{k}j). \qquad (11)$$

Three of the solutions $\omega(\mathbf{k}j)$ of Eq. (8) go to zero in the limit $\mathbf{k} \to 0$ and represent the *acoustic branches* of the *dispersion relation.* They pertain in the long-wavelength continuum limit to longitudinal and transverse acoustic waves. The remaining $3r - 3$ solutions $\omega(\mathbf{k}j)$ approach finite limiting values as $\mathbf{k} \to 0$ and represent the *optic branches.* In these latter modes near $\mathbf{k} = 0$ the sublattices vibrate against one another, and this motion can generally be detected by infrared absorption or Raman scattering.

In order to solve Eq. (8), the atomic-force constants are required. There are some successful *ab initio* theories, but most lattice dynamics calculations are based on a parametrized model of some kind (see Bilz and Kress, 1979). Over the years, many different models have been devised. The *valence force field* model is one of the simplest and is based on intuitive ideas that have a foundation in molecular vibrational studies. One pictures the atoms in the crystal as joined together by bonds and expresses the interatomic forces in terms of the forces and torques required to stretch, bend, and twist these bonds. This model has been applied with some success to covalent crystals such as diamond, silicon, and germanium. By and large though, most successful models incorporate long-range Coulomb forces in some form. This is obviously necessary for treating ionic crystals, but even in the case of covalent crystals the number of parameters required for a successful fit tends to be less when Coulomb forces are incorporated explicitly in the model in some way. The *rigid-ion model* is the simplest approach to ionic crystals, and there are a number of more sophisticated models that allow for deformation of the electron wave functions of the atoms as they move relative to one another. These include dipole models, various shell and deformation-shell models, and the bond-charge model. In metals, the interaction between the ions and the conduction electrons has to be taken into account. In the de Launay (1953) model for instance, the electron gas is attributed a finite bulk modulus and is assumed to adjust adiabatically to the motion of the ions, screening their long-range Coulomb interactions in the process. This coupling has a pronounced effect on the longitudinal lattice modes, while pure transverse modes are unaffected.

Whatever model is adopted for a particular application, it commonly features a number of parameters such as bond force constants, shell charges, atomic polarizabilities, or whatever. These parameters have to be determined from experiment. The phase velocities $v(j) = \omega(\mathbf{k}j)/k$ of long-wavelength acoustic modes are a function of the *bulk elastic constants* of the crystal, and this provides as many constraints on the parameters of the model as the crystal has independent elastic constants. Measured Raman frequencies, dielectric tensor, and optical absorption often provide further constraints on the parameters. If the model is simple enough, this information may be sufficient to determine all the parameters. The ultimate test of a model, though, is that it should fit the dispersion relations measured by neutron scattering or one of the other techniques discussed in Sec. 3. As an example, Fig. 1(a) shows dispersion curves for silicon in a number of principal crystallographic directions

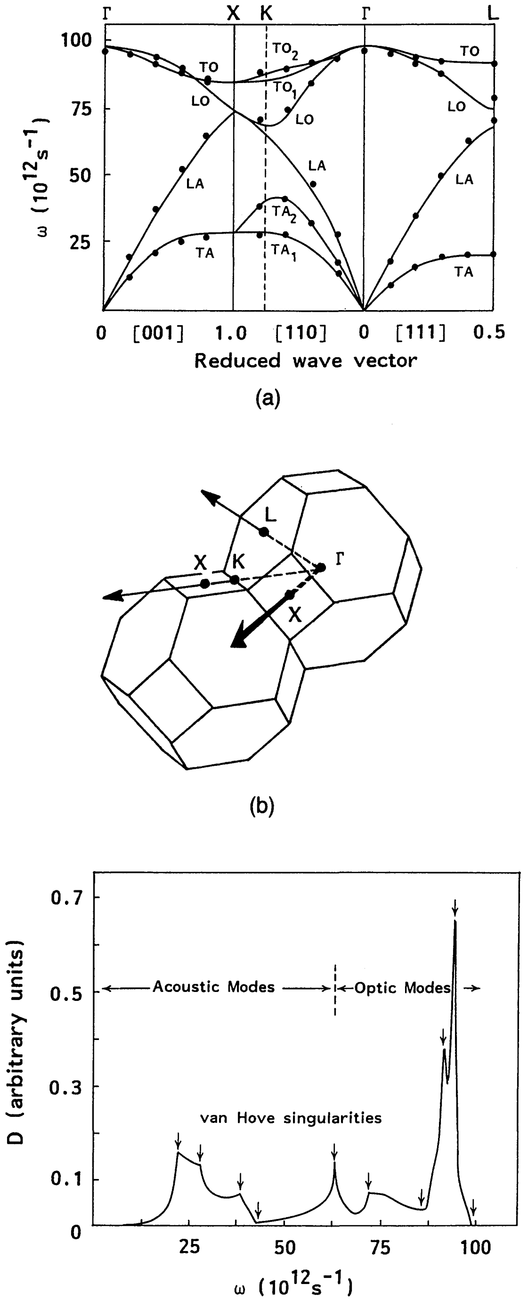

FIG. 1. (a) Phonon dispersion curves for silicon (adapted from Bilz and Kress, 1979). The directions are referred to the face-centered cubic (fcc) Brillouin zone shown in (b). The branches are distinguished as longitudinal acoustic (LA), longitudinal optic (LO), transverse acoustic (TA), and transverse optic (TO). The two TA and two TO branches are degenerate in the [001] and [111] directions because of symmetry. (c) Calculated frequency-distribution function $D(\omega)$ for Si.

[see Fig. 1(b)]. The points denote neutron-scattering data, and the lines represent calculations on the basis of a four-parameter *bond-charge model.*

Many thermodynamic quantities can be calculated from a knowledge of the *frequency distribution function* or *density of states $D(\omega)$*, defined as the number of normal-mode frequencies per unit frequency interval at ω. Figure 1(c) shows $D(\omega)$ calculated for Si. The region below about $20 \times 10^{12}\ \mathrm{s}^{-1}$ corresponds to long-wavelength acoustic modes. Here $D(\omega)$ is roughly proportional to ω^2, in accordance with the Debye model. At higher frequencies, $D(\omega)$ is punctuated by a number of sharp features known as *van Hove singularities* (Maradudin *et al.*, 1971). These correspond to critical points at the zone boundaries (and elsewhere) where the dispersion relation has an extremum. To one side of the critical frequency ω_0, $D(\omega) \sim \sqrt{|\omega - \omega_0|}$, giving a divergence in the derivative of $D(\omega)$ at ω_0, while on the other side there is no singular behavior. These singularities carry over to other spectral quantities such as defect-activated infrared absorption, and this provides scope for experimentally determining zone-boundary frequencies.

1.1 Quantization; the Phonon

Quantization of the lattice vibrations is effected by regarding the normal coordinates $Q(\mathbf{k}j)$ and their conjugate momenta $P(\mathbf{k}j)$ as operators satisfying the commutation relations (LV)

$$[P(\mathbf{k}'j'), Q(\mathbf{k}j)] = (\hbar/i)\,\delta_{jj'}\Delta(\mathbf{k} - \mathbf{k}'), \tag{12}$$

where $\hbar$ is Planck's constant, $\delta_{jj'}$ is the Kronecker delta, and

$$\Delta(\mathbf{k}) = \begin{cases} 1 & \text{if } \mathbf{k} = \mathbf{G}, \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

where $\mathbf{G}$ is a reciprocal lattice vector. The Hamiltonian is given in terms of these conjugate operators by

$$H_L = \tfrac{1}{2} \sum_{\mathbf{k}j} [P(\mathbf{k}j)P^*(\mathbf{k}j) + \omega_j^2(\mathbf{k})Q(\mathbf{k}j)Q^*(\mathbf{k}j)] \tag{14}$$

and is in the form of a sum of $3rN$ independent harmonic oscillator Hamiltonians. The eigenstates of H_L can thus be represented as products of eigenstates of the individual oscillators.

The usual approach to the harmonic oscillator, found in most elementary textbooks on quantum mechanics (see, e.g., Schiff, 1955), is to solve Schrödinger's equation in the coordinate or momentum representation and thereby obtain the eigenstates and eigenenergies. In the present context, of more interest than the form of eigenstates are the energy eigenvalues and the matrix representations of operators such as Q, Q^2, and P, etc., in the energy basis. This information is required for calculating expectation values and thermal averages, carrying out perturbation calculations, and many other purposes. The terms P and Q both appear quadratically in the Hamiltonian, and this symmetry allows the energy spectrum and required matrix representations to be obtained by purely algebraic means.

The trick, which goes by the name of *second quantization*, is to introduce operators $a^\dagger(\mathbf{k}j)$ and $a(\mathbf{k}j)$ for each mode $\mathbf{k}j$ by means of the equations

$$Q(\mathbf{k}j) = \sqrt{\frac{\hbar}{2\omega(\mathbf{k}j)}}\,[a^\dagger(-\mathbf{k}j) + a(\mathbf{k}j)], \tag{15}$$

$$P(\mathbf{k}j) = i\sqrt{\frac{\hbar\omega(\mathbf{k}j)}{2}}\,[a^\dagger(\mathbf{k}j) - a(-\mathbf{k}j)]. \tag{16}$$

It is easily verified (see, e.g., Ghatak and Kothari, 1972) that these operators satisfy the commutation relations

$$[a(\mathbf{k}'j'), a^\dagger(\mathbf{k}j)] = \delta_{jj'}\Delta(\mathbf{k} - \mathbf{k}') \tag{17}$$

and that the Hamiltonian is given in terms of them by

$$H_L = \sum_{\mathbf{k}j} \hbar\omega(\mathbf{k}j)\,[a^\dagger(\mathbf{k}j)a(\mathbf{k}j) + \tfrac{1}{2}]. \tag{18}$$

Considering an individual oscillator, it may be shown using the commutation relations that $a^\dagger(\mathbf{k}j)$ and $a(\mathbf{k}j)$ operating on an eigenstate of the Hamiltonian yield other eigenstates that are higher and lower in energy respectively by $\hbar\omega(\mathbf{k}j)$, that the complete set of eigenstates can thereby be generated, and that there is a ground state. The eigenstates can thus be denoted $|n(\mathbf{k}j)\rangle$, and the associated eigenvalues are given by

$$E(n(\mathbf{k}j)) = [n(\mathbf{k}j) + \tfrac{1}{2}]\hbar\omega(\mathbf{k}j), \tag{19}$$

where the quantum number $n(\mathbf{k}j)$ takes on all positive integer values and is zero for the ground state. When the oscillator undergoes a transition, energy is added or removed in multiples of $\hbar\omega(\mathbf{k}j)$. This quantum of energy is called the *phonon*, and $n(\mathbf{k}j)$ is known as the *phonon occupation number* for the state. The operators $a^\dagger(\mathbf{k}j)$ and $a(\mathbf{k}j)$ are called the *phonon creation* and *annihilation* (or *destruction*) *operators* respectively, and raise and lower the phonon occupation number of a mode in accordance with

$$a^\dagger(\mathbf{k}j)|n(\mathbf{k}j)\rangle = \sqrt{n(\mathbf{k}j) + 1}\,|n(\mathbf{k}j) + 1\rangle, \tag{20}$$

$$a(\mathbf{k}j)|n(\mathbf{k}j)\rangle = \sqrt{n(\mathbf{k}j)}\,|n(\mathbf{k}j) - 1\rangle. \tag{21}$$

The product $a^\dagger(\mathbf{k}j)a(\mathbf{k}j)$ is called the number operator because it extracts the phonon occupation number of a state:

$$a^\dagger(\mathbf{k}j)a(\mathbf{k}j)|n(\mathbf{k}j)\rangle = n(\mathbf{k}j)|n(\mathbf{k}j)\rangle. \tag{22}$$

Eigenstates of the combined system of all the oscillators can be represented as products of the individual oscillator states: $|\cdots n(\mathbf{k}j)\cdots n(\mathbf{k}'j')\cdots\rangle = |\cdots\rangle|n(\mathbf{k}j)\rangle|\cdots\rangle|n(\mathbf{k}j)\rangle \times |\cdots\rangle$, with the energy being the sum of the energies of the individual oscillators, i.e.,

$$E(\{n(\mathbf{k}j)\}) = \sum_{\mathbf{k}j} [n(\mathbf{k}j) + \tfrac{1}{2}]\hbar\omega(\mathbf{k}j). \tag{23}$$

The phonon can in many respects be regarded as a particle. If one thinks of it as occupying a wave packet centered on some value of $\mathbf{k}$, rather than an infinitely extended plane wave, then the phonon can be ascribed an instantaneous position in space. The more precisely that point is specified, the greater of course is the spread of $\mathbf{k}$'s in the wave packet, which is a manifestation of the Heisenberg uncertainty principle. The rate at which the position of the phonon changes with time is its group velocity, which is given by $\mathbf{V} = \nabla_{\mathbf{k}}\omega(\mathbf{k}j)$. The quantity $\hbar\mathbf{k}$ is called the *pseudomomentum* or *crystal momentum* of the phonon. Although not momentum in the true sense, it nevertheless plays a role very similar to that of physical momentum. For instance, when a photon of light is scattered inelastically from a crystal with the loss of energy being taken up in the creation of a phonon, the change in momentum of the photon is equal to the value of $-\hbar\mathbf{k}$ for the phonon, which seems to signify the conservation of total momentum. Momentum conservation is a consequence of translational invariance, and a crystal lattice is only invariant under a discrete set of translations, not all translations. The upshot of this is that in general a scattering event involving phonons takes place with conservation of real plus crystal momentum, but only to within $\hbar\mathbf{G}$, where $\mathbf{G}$ is a reciprocal-lattice vector. An inelastic neutron-scattering event, for instance, might involve a combination of Bragg scattering by $\mathbf{G}$ together with the creation of a phonon of wave vector $\mathbf{k}$, so that the momentum transfer to the lattice, $\hbar(\mathbf{G} + \mathbf{k})$, differs from the phonon momentum by $\hbar\mathbf{G}$.

1.2 Statistical Properties of Phonons

Collectively, phonons can be ascribed a phase-space distribution function $n(\mathbf{k}j;\ \mathbf{x})$, from which quantities such as energy density, heat current, and momentum flux can be derived. Phonon scattering can lead to a rearrangement of the distribution, the process being governed by the Boltzmann equation (LV). Phonons obey Bose statistics, with the chemical potential being zero. The average number of phonons in any mode in thermal equilibrium depends on the local temperature T and is given by the *Planck distribution function* (LV)

$$\bar{n}(\mathbf{k}j) = \{\exp[\hbar\omega(\mathbf{k}j)/k_{\mathrm{B}}T] - 1\}^{-1}, \tag{24}$$

where k_{B} is Boltzmann's constant. It follows that

$$\begin{aligned}\langle Q(\mathbf{k}j)Q^*(\mathbf{k}j)\rangle &= \frac{\hbar}{2\omega(\mathbf{k}j)}[2\bar{n}(\mathbf{k}j) + 1]\\ &= \frac{\hbar}{2\omega(\mathbf{k}j)}\coth[\hbar\omega(\mathbf{k}j)/2k_{\mathrm{B}}T].\end{aligned} \tag{25}$$

The evaluation of thermodynamic quantities such as the Helmholtz free energy, the internal energy, heat capacity, and entropy with the aid of the partition function is discussed in LV.

The thermal expectation values of operators of the form $\exp[isu_\alpha(l\kappa)]$, where s is a parameter, are met in many situations. In treating the scattering of x rays and neutrons from a vibrating crystal, for instance, the phase factor for each atom is multiplied by $\sim\exp[isu_\alpha(l\kappa)]$ to take account of the instantaneous displacement of that atom from equilibrium, and the result has to be averaged over the ensemble of vibrational states of the crystal.

The evaluation of this expectation value entails expanding the exponential in powers of $u_\alpha(l\kappa)$, calculating the diagonal matrix elements, and then averaging over all oscillator states with their Boltzmann weighting factors. The details of the calculation can be found in Maradudin *et al.* (1971), and the result is

$$\langle\exp[isu_\alpha(l\kappa)]\rangle = \exp[-\tfrac{1}{2}s^2\langle u_\alpha^2(l\kappa)\rangle]. \tag{26}$$

This quantity exhibits a reduction with temperature as the mean square displacement of the atom increases and there is a growing degree of phase cancellation. Equation (26) is the basis of the *Debye–Waller factor* for the decrease in the intensity of x-ray Bragg peaks with temperature, related effects for neutron and other types of scattering, the temperature dependence of the recoil-free fraction in the Mössbauer effect, and zero-phonon line intensities in color center spectra. These phenomena are discussed individually in subsequent sections.

2. PHONON–PHONON INTERACTIONS

For any real crystal, the Hamiltonian considered thus far represents only part of the full picture. To arrive at Eq. (1), the lattice potential energy has been expanded in powers of the displacements of the atoms from equilibrium, and only the quadratic terms retained. Moreover, interactions between the atoms and other particles or excitations that might be present in the crystal have not been considered. These additional terms account for many interesting physical phenomena. They constitute the coupling to the probes such as light, neutrons, and x rays that are used to study lattice vibrations, they are the means by which thermal equilibrium is established, and they give rise to scattering processes that play a central role in phonon transport phenomena. In this and subsequent sections, we consider the forms that these interactions take, and some of their consequences.

Continuing the expansion of the lattice potential energy beyond second order, the terms of third order can be grouped together as $\Phi^{(3)}$, those of fourth order as $\Phi^{(4)}$, and so on. Expressed in terms of the previously defined normal coordinates, $\Phi^{(3)}$ takes the form (LV)

$$\Phi^{(3)} = \sum_{\mathbf{k}j}\sum_{\mathbf{k}'j'}\sum_{\mathbf{k}''j''} V^{(3)}(\mathbf{k}j,\mathbf{k}'j',\mathbf{k}''j'') \times Q(\mathbf{k}j)Q(\mathbf{k}'j')Q(\mathbf{k}''j''), \tag{27}$$

where $V^{(3)}(\mathbf{k}j,\mathbf{k}'j',\mathbf{k}''j'')$ are a set of coefficients that preserve the translational invariance of the lattice. A number of effects can be handled by treating $\Phi^{(3)}$ by perturbation theory. In the first place, the vibrational states form a quasicontinuum, and Fermi's "golden rule" can be used to calculate transition rates. It follows from Eqs. (15) and (27) that a matrix element of $\Phi^{(3)}$ between eigenstates of the harmonic Hamiltonian will be zero unless precisely three phonon occupation numbers in the one state differ by $+1$ or -1 from the corresponding occupation numbers in the other state. The three occupation numbers cannot all increase or all decrease in a transition, since that would violate energy conservation. A transition in which, say, $n(\mathbf{k}j)$ decreases and $n(\mathbf{k}'j')$ and $n(\mathbf{k}''j'')$ both increase, i.e., phonons $\mathbf{k}'j'$ and $\mathbf{k}''j''$ are created whose combined energies equal that of the original phonon $\mathbf{k}j$, has transition probability

$$p = \frac{2\pi}{\hbar^2}|V^{(3)}(\mathbf{k}j,-\mathbf{k}'j',-\mathbf{k}''j'')|^2 \times \frac{\hbar^3}{8\omega(\mathbf{k}j)\omega(\mathbf{k}'j')\omega(\mathbf{k}''j'')}n(\mathbf{k}j)[n(\mathbf{k}'j')+1] \times [n(\mathbf{k}''j'')+1]\delta(\omega(\mathbf{k}j)-\omega(\mathbf{k}'j')-\omega(\mathbf{k}''j'')). \tag{28}$$

Because of the translational symmetry of the lattice, $|V^{(3)}(\mathbf{k}j,-\mathbf{k}'j',-\mathbf{k}''j'')|^2$ is proportional to $\Delta(\mathbf{k}-\mathbf{k}'-\mathbf{k}'')$, and so there is conservation of crystal momentum to within a reciprocal-lattice vector times $\hbar$. The transition can be represented diagrammatically as shown in Fig. 2(a). Transitions represented by Fig. 2(b), in which two phonons are annihilated and one created, are calculated in a similar way. The total *decay rate*, or *inverse lifetime* of a phonon against two-phonon decay under thermal equilibrium conditions, is then obtained by summing p over all decay channels, using the thermal expectation values for the n's.

Phonon–phonon scattering processes play a central role in phonon transport. Transitions in which $\mathbf{k} - \mathbf{k}' - \mathbf{k}'' = \mathbf{G} \neq 0$ are termed *Umklapp* (U) processes and require that the participating phonons have wave vectors that are comparable with the zone-boundary values. Crystal momentum is not conserved in these processes, and they are the dominant source of thermal resistance in most insulating crystals at room temperature, where large–wave-vector phonons are present in appreciable numbers. Transitions in which $\mathbf{k} - \mathbf{k}' - \mathbf{k}'' = 0$ are termed *normal* (N) processes and conserve crystal momentum. They will rearrange a nonequilibrium distribution of phonons without changing the total momentum, and on their own do not represent a source of thermal resistance.

As a general observation, in thermal equilibrium at a temperature T well below the Debye temperature Θ_D, the average phonon lifetime against two-phonon decay is proportional to T^{-5} or $\bar{\omega}^{-5}$. The requirements of energy and momentum conservation, though, place constraints on the channels by which a specific phonon can decay, and there are considerable differences in lifetime depending on the particular phonon branch. The phonon dispersion curves for silicon shown in Fig. 1(a) are fairly representative of a large class of insulating crystals, in that the acoustic branches curve downward with increasing $|\mathbf{k}|$ (soft-mode materials discussed below provide exceptions to this rule). This being the case, a phonon cannot decay into two other phonons of the same branch with collinear momenta, since in order for momentum to be conserved the combined energies of the two phonons would exceed that of the original one. For an isotropic crystal, or in a slow direction in an anisotropic crystal, noncollinear decay would leave an even greater energy shortfall. Elaborating on this idea, Lax *et al.* (1981) have proved that a phonon cannot decay by anharmonic processes of any order into phonons each of whose phase velocities are higher than that of the original phonon. This is true for U as well as N processes. Slow transverse phonons can therefore, in the absence of other forms of scattering, be very long lived. This has important implications for *ballistic transport.*

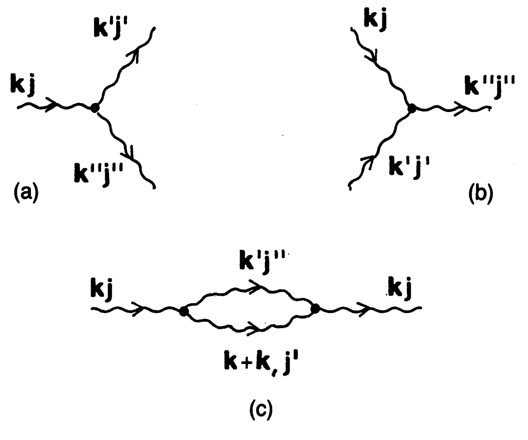

FIG. 2. Anharmonic processes involving $-\Phi^{(3)}$ in (a,b) first order and (c) second order.

Perturbation theory predicts a mixing of phonon states, and in second order a *renormalization of the phonon energies*, which can be represented by diagrams such as the one shown in Fig. 2(c). The phonon dispersion relations, phonon velocities, and thermodynamic quantities such as the elastic constants thereby become temperature dependent. In *quantum crystals* such as solid He, the effects of renormalization are pronounced, even at the absolute zero of temperature (Guyer, 1969). The anharmonic terms in the lattice potential account for thermal expansion, which in the Grüneisen approach is treated by considering the volume dependence of the phonon frequencies and minimizing the total free energy (Ashcroft and Mermin, 1981).

Phonon–phonon interactions play an important role in second-order *structural phase transitions* (see PHASE TRANSITIONS, STRUCTURAL). Many of these are accompanied by a *soft mode* in the high-temperature symmetrical phase, which shows up as a pronounced dip centered on a particular value $\mathbf{k}_0$ in the phonon dispersion relation. The soft-mode $Q(\mathbf{k}_0)$ has a negative harmonic restoring force constant $\omega_0^2(\mathbf{k}_0)$ but is dynamically stabilized through interactions with all the other modes.

The stabilization is most easily accounted for in terms of the quartic terms in the lattice potential, specifically those with $\mathbf{k} = \mathbf{k}' = \mathbf{k}_0$ and $\mathbf{k}'' = \mathbf{k}'''$. On taking a statistical average over the latter modes and summing, the mode $Q(\mathbf{k}_0)$ comes out to have a quasiharmonic frequency $\omega(\mathbf{k}_0)$ given by (Dorner and Comes, 1977)

$$\omega^2(\mathbf{k}_0) = \omega_0^2(\mathbf{k}_0) + \sum_{k''} V^{(4)}(\mathbf{k}_0,-\mathbf{k}_0,\mathbf{k}'',-\mathbf{k}'') \times \langle Q^*(\mathbf{k}'')Q(\mathbf{k}'')\rangle. \tag{29}$$

In the high-temperature phase, the second term in Eq. (29) is positive and exceeds the first in magnitude, and so $\omega(\mathbf{k}_0)$ is real. The soft-mode frequency is usually fairly small, though, and the mode heavily damped. As the temperature is lowered, $\omega(\mathbf{k}_0)$ approaches zero, the lattice becomes unstable, and a transition to a lower-symmetry structure takes place. In simple cases, the order parameter for the transition corresponds fairly closely to the soft-mode normal coordinate.

Varying the isotopic composition of a substance changes the phonon frequencies and zero-point motion amplitudes, and coupled with anharmonicity this results in changes to the lattice constant, elastic constants, and related physical quantities. The isotope effect in superconductivity is a classic example that is discussed later. Recently, the properties of isotopically enriched diamond have attracted much attention (see Hass *et al.*, 1992, and references contained therein).

3. SCATTERING METHODS FOR DETERMINING PHONON DISPERSION RELATIONS

Thermal diffuse scattering of x rays provided the earliest means of measuring the bulk phonon dispersion relations of crystals. Coherent inelastic neutron scattering arrived later on the scene but is a more powerful technique and has largely supplanted x-ray scattering in this task. Neutron scattering is also a valuable source of information on the displacement eigenvectors $e_\alpha(\kappa|\mathbf{k}j)$, phonon lifetimes, vibrations at defects, mode softening, and many other physical phenomena. Helium-atom and electron scattering are widely used in the study of lattice vibrations at surfaces.

The above-mentioned methods share certain common features. To permit measurement of the phonon dispersion relation over the entire Brillouin zone, the wavelength of the probe (de Broglie wavelength in the case of a particle) has to be comparable to the lattice spacing. The scattering is consequently dominated by elastic Bragg scattering from the crystal planes. Information about lattice vibrations is gleaned from the much weaker diffuse background scattering. This background is caused by inelastic scattering in which some of the energy of the probe is given over to the creation or annihilation of one, two, or more phonons. In the case of x rays, the limited wavelength resolution does not permit the energy shift to be measured directly, and the phonon frequencies are inferred from the temperature dependence of the diffuse intensity. For inelastic neutron and atom scattering, the energy shift is measured as a matter of course, and the phonon frequencies are obtained directly. As the temperature is raised, the vibrational amplitudes increase and the crystal appears less and less ordered to the probing particle. This results in a reduction in the intensities of the Bragg peaks and a compensatory increase in the inelastic scattering.

3.1 Inelastic X-Ray Scattering

In treating x-ray scattering from a vibrating crystal, the frequency shift is ignored. This change is of the order of the phonon frequency, $\sim 10^{12}$ Hz, which is small compared with the x-ray frequency, $\sim 10^{18}$ Hz. For simplicity, we will consider scattering from a monatomic Bravais lattice, which entails dropping the index κ. The scattering from any point within an atom is proportional to the local electron probability density, which is assumed to accompany the atom as it vibrates. The resulting scattering amplitude for an instantaneous configuration of all the atoms, taking into account the phase relationships between the scattering amplitudes for the individual atoms, is

$$\Psi \sim f(\mathbf{S}) \sum_{l} \exp\{i\mathbf{S}\cdot[\mathbf{x}(l) + \mathbf{u}(l)]\}, \tag{30}$$

where $f(\mathbf{S})$ is the *atomic structure factor*, $\mathbf{S} = \mathbf{q} - \mathbf{q}'$ is the scattering vector, and $\mathbf{q}$ and $\mathbf{q}'$ are the wave vectors of the incoming and outgoing radiation. The time-averaged *scattering intensity*

$$I = \langle \Psi^* \Psi \rangle = f^2 \sum_{ll'} \langle \exp\{i\mathbf{S}\cdot[\mathbf{x}(l) - \mathbf{x}(l') + \mathbf{u}(l) - \mathbf{u}(l')]\}\rangle \tag{31}$$

is obtained as an ensemble average over all vibrational states. Making use of Eq. (26) and expressing the atomic displacements in terms of the normal coordinates, and then expanding the exponential in powers of the averaged displacements, one obtains (see Maradudin *et al.*, 1971)

$$I = f^2 e^{-2W} [I_0 + I_1 + \ldots], \tag{32}$$

where

$$2W = \frac{2}{NM} \sum_{\mathbf{k}j} [\mathbf{S}\cdot\mathbf{e}(\mathbf{k}j)]^2 \langle Q^2(\mathbf{k}j)\rangle, \tag{33}$$

$$I_0 = \sum_{ll'} \exp\{i\mathbf{S}\cdot[\mathbf{x}(l) - \mathbf{x}(l')]\} = N^2 \Delta(\mathbf{S}), \tag{34}$$

$$I_1 = \frac{\hbar}{M} \sum_{j=1}^{3} [\mathbf{S}\cdot\mathbf{e}(\mathbf{K}j)]^2 \frac{\coth[\hbar\omega(\mathbf{K}j)/2k_B T]}{\omega(\mathbf{K}j)}, \tag{35}$$

with $e(\mathbf{k}j)$ standing for the mode eigenvector.

The scattering for the static lattice is $f^2 I_0$, which is zero except in the direction of the *Bragg peaks* where the *Laue condition* $\Delta(\mathbf{S}) = 1$ is satisfied. As a result of zero-point and thermal motion, the intensities of the Bragg peaks are reduced by the *Debye–Waller factor* e^{-2W}. The second term, I_1, describes the diffuse background due to one-phonon scattering, with $\mathbf{K}$ being the distance of $\mathbf{S}$ from the nearest reciprocal-lattice point. The term I_1 increases rapidly as $\mathbf{S}$ approaches the reciprocal-lattice point, and its behavior is determined by the three lattice modes of wave vector $\mathbf{K}$. For $\mathbf{K}$ along a symmetry axis, $\mathbf{S}\cdot\mathbf{e}(\mathbf{K}j)$ is nonzero except for the longitudinal mode, and so a measurement of the diffuse intensity as a function of temperature yields $\omega(\mathbf{K}j)$ for that mode. Measurements in other directions in reciprocal space yield data on the transverse branches. A drawback of this technique is that corrections have to be made for other sources of diffuse scattering. These include second-, third-, and higher-order processes, whose contributions to the diffuse background grow relative to the one-phonon scattering the further from the reciprocal-lattice point.

3.2 Coherent Inelastic Neutron Scattering

Coherent inelastic neutron scattering represents the most effective method for measuring phonon dispersion relations. The source is a collimated beam of monochromatic neutrons of wavelength comparable to the interatomic spacing. The crystal specimen is placed in the beam and gives rise to Bragg scattering, as well as inelastic scattering in which the change in energy and momentum of the neutron is taken up in the creation or annihilation of one or more phonons. Neutron scattering differs from x-ray scattering in the important respect that for given momentum transfer $\hbar\mathbf{k}$ the energy shift $\hbar\omega(\mathbf{k}j)$ is measured directly, usually with an accuracy of a few percent.

In neutron scattering, one measures the *differential cross section* $d^2\sigma/dEd\Omega$, which is the probability per unit solid angle and energy interval that an incoming neutron of wave vector $\mathbf{q}$ and energy $E = \hbar^2 q^2/2m$ is scattered into wave vector $\mathbf{q}'$ and energy $E' = \hbar^2 q'^2/2m$. For a single stationary nucleus, this is equal to $b^2\delta(E - E')$, where b is a quantity called the scattering length, which is independent of the neutron wavelength or angle of scattering.

For scattering from a vibrating monatomic crystal, in which there may be a change in the magnitude of the wave vector, the ensemble-averaged cross section taken over all possible initial vibrational states i of the crystal, with their Boltzmann probabilities p_i and final vibrational states f, is given by (Lovesey, 1977)

$$\frac{d^2\sigma}{dEd\Omega} = b^2 \frac{q'}{q} \sum_{if} p_i \left| \left\langle i \left| \sum_{l} \exp\{i\mathbf{S}\cdot[\mathbf{x}(l) + \mathbf{u}(l)]\} \right| f \right\rangle \right|^2 \times \delta(\hbar\omega + E_i - E_f), \tag{36}$$

where $\mathbf{S} = \mathbf{q} - \mathbf{q}'$ is the scattering vector and $\hbar\omega = E - E'$ is the energy transfer to the crystal.

The evaluation of Eq. (36) proceeds along similar lines to that of Eq. (31) for x-ray scattering. There are the Bragg diffraction peaks, whose intensities are reduced by the Debye–Waller factor e^{-2W}. The inelastic scattering is obtained by expanding the exponential in Eq. (36) in powers of the $\mathbf{u}(l)$ and expressing the $\mathbf{u}(l)$ in terms of the normal-mode phonon creation and annihilation operators. The linear terms yield energy- and momentum-conserving one-phonon creation and annihilation processes. The scattering cross section for these one-phonon processes is (Cowley, 1969)

$$\frac{d^2\sigma}{dE d\Omega} = b^2 \frac{q'}{q} \frac{N\hbar e^{-2W}}{2M\omega(\mathbf{k}j)} [\mathbf{S}\cdot\mathbf{e}(\mathbf{k}j)]^2[\bar{n}(\mathbf{k}j) + \tfrac{1}{2} \pm \tfrac{1}{2}] \times \delta(\hbar\omega \mp \hbar\omega(\mathbf{k}j))\Delta(\mathbf{S} - \mathbf{k} - \mathbf{G}), \tag{37}$$

with the upper sign corresponding to energy loss by the neutron through the creation of a phonon $\mathbf{k}j$, and the lower sign to energy gain through annihilation of that phonon.

The dispersion relations of a great many crystals have been measured with neutron scattering (see Bilz and Kress, 1979, for a survey of results). The measurements are fairly time consuming and are usually confined to principal crystallographic directions. Figure 1(a) shows the measured phonon dispersion relation of Si.

3.3 Helium-Atom and Electron Scattering; Surface Phonons

In recent years, a great deal has been learned about lattice vibrations at crystal surfaces using helium-atom scattering and *electron energy-loss spectroscopy* (EELS). The He atoms in these experiments are given energies of the order of 0.02 eV, which permits them to approach no closer than about 3–4 Å from the crystal surface, and they thus effectively sense only the motion of the top layer of atoms. In EELS, on the other hand, the electrons have energies of the order of 100 eV, which allows them to penetrate two to three atomic layers into the crystal and sense the motion of all of those atoms. The surfaces that are studied are usually parallel to principal crystallographic planes such as (001) and (111), and the scattered particles are generally measured in the plane of incidence, which is located in another principal plane (Toennies, 1991).

A bounded crystal retains the translational symmetry of the lattice parallel, but not perpendicular, to the boundary, and so it is only the component of the wave vector parallel to the surface, $\mathbf{k}_\parallel$, that serves for labeling the vibrational modes. There has to be another quantum number taking on a quasi-continuum of values, but it does not represent the perpendicular component of the wave vector. Surface-phonon dispersion relations thus have the general appearance of that of platinum (111) shown in Fig. 3. Corresponding to each value of $\mathbf{k}_\parallel$ there is a continuous band of modes, which far from the surface match up with the subset of bulk modes of the crystal whose wave vectors have $\mathbf{k}_\parallel$ as their in-plane component. There may be resonances in the continuum, for which the displacements of the associated modes become very large near the surface. In addition, there is a discrete Rayleigh branch, which is separated from the band modes from below. This corresponds to true surface modes for which the displacements fall off more or less exponentially with distance from the surface.

Both He atoms and electrons undergo elastic Bragg scattering in which the change in the in-plane component of the momentum of the projectile is equal to $\hbar\mathbf{G}$. The Bragg intensities are, as usual, reduced by a Debye–Waller factor. In addition, there is inelastic scattering accompanied by the creation or destruction of a surface phonon. The kinematic conditions are conservation of energy and of momentum in the surface to within $\hbar\mathbf{G}$.

The experimental data in Fig. 3 show res-

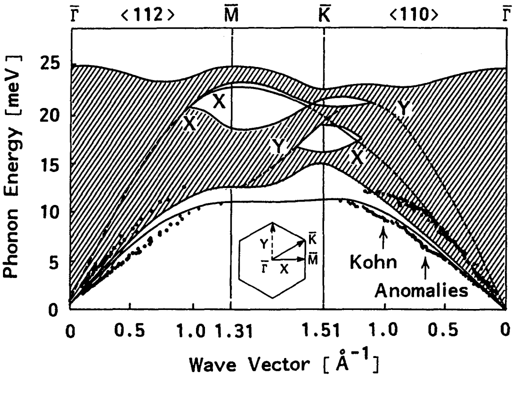

FIG. 3. Surface-phonon dispersion relation for Pt(111) (adapted from Toennies, 1991). Dots represent experimental data: curves and shaded area are theoretical. The letters X and Y indicate the polarization at the edge of the bulk bands.

onances that have been measured with He-atom scattering. These show a number of interesting features. In the ⟨110⟩ direction, there are two dips in the Rayleigh branch that have been ascribed to *Kohn anomalies* (see Sec. 5). In both the ⟨110⟩ and ⟨112⟩ directions, there is a resonance in the continuum due to a longitudinal mode concentrated near the surface by a weakening of the interatomic force constants there.

Helium-atom scattering has revealed a soft mode in tungsten (001) that precedes *surface reconstruction* as the temperature is lowered to 300 K. This is the first observation of a soft mode in a two-dimensional phase transition. Kress (1991) contains an extensive compilation of results obtained with He-atom scattering and EELS.

4. INFRARED ABSORPTION AND LIGHT SCATTERING

Phonons feature importantly in the infrared absorption and reflectivity of crystals, and in the Raman and Brillouin scattering of light.

The infrared properties of polar crystals are dominated by motion of the charged sublattices against each other. In the case of the alkali halides, for example, the near–zone-center longitudinal optical (LO) mode gives rise to an alternating electric field, which causes this branch to split away from the degenerate transverse optical (TO) branches, which are not accompanied by such a field. The well-known Lyddane–Sachs–Teller theory has these zone-center frequencies related by (Kittel, 1986)

$$\omega_L^2/\omega_T^2 = \epsilon(0)/\epsilon(\infty), \tag{38}$$

where $\epsilon(0)$ and $\epsilon(\infty)$ are the values of the dielectric constant at frequencies well below and well above ω_T, respectively. Electromagnetic waves, being transversely polarized, couple very strongly to the TO modes, but not to the LO mode. A coupled-mode polariton approach is required to describe wave propagation in the crystal, and this yields a band gap between ω_T and ω_L, where the crystal ideally should be perfectly reflecting to electromagnetic radiation. High values of reflectivity are indeed measured in this frequency range.

In nonpolar diamond-lattice–type crystals, one-phonon absorption is forbidden because there is a center of inversion midway between each pair of neighboring atoms, and there is therefore no electric dipole moment associated with the vibration of the two sublattices against each other. The presence of defects in the crystal breaks the symmetry, and all modes become infrared active, though not all to the same extent.

Two-phonon absorption is allowed in perfect crystals subject to the kinematic constraints of energy and momentum conservation. An infrared photon with the right energy to be absorbed will have a wavelength that is extremely large compared with the lattice spacing, and hence its wave vector $\mathbf{q}$ can be taken to be zero. The absorption is mediated mainly by the quadratic dependence of the electric dipole moment on the atomic displacements, which consists of momentum-conserving terms of the form $b(0)[a^\dagger(-\mathbf{k}j) + a(\mathbf{k}j)][a^\dagger(\mathbf{k}j') + a(-\mathbf{k}j')]$, where $b(0)$ is the photon-destruction operator. The absorption consists of a difference band, in which one phonon is created and another destroyed, and a summation band, in which two phonons are created. The absorptivity for the latter is given by (Reissland, 1973)

$$\alpha(\omega) = \sum_{\mathbf{k}jj'} \frac{A(\mathbf{k}jj')}{\omega(\mathbf{k}j)\omega(-\mathbf{k}j')}\{1 + n(\mathbf{k}j) + n(-\mathbf{k}j')\} \times \delta(\omega - \omega(\mathbf{k}j) - \omega(-\mathbf{k}j')), \tag{39}$$

where $A(\mathbf{k}jj')$ are not very easily determined coupling coefficients. As a crude first approximation, taking the combination of factors preceding the delta function as a constant, the absorption comes out to be proportional to the joint frequency distribution function $D_2(\omega)$. One can think of $D_2(\omega)$ as pertaining to a dispersion relation with $r(r + 1)/2$ branches representing $\omega(\mathbf{k}j) + \omega(-\mathbf{k}j')$ as a function of $\mathbf{k}$ for all combinations of j and j'. As with $D(\omega)$, $D_2(\omega)$ possesses critical points, but in much greater profusion, and it is a complex combinatorial exercise to extract the critical points of $D(\omega)$ from those of $D_2(\omega)$. Figure 4 shows the two-phonon infrared absorption spectrum of silicon and identifies some of the critical-point features.

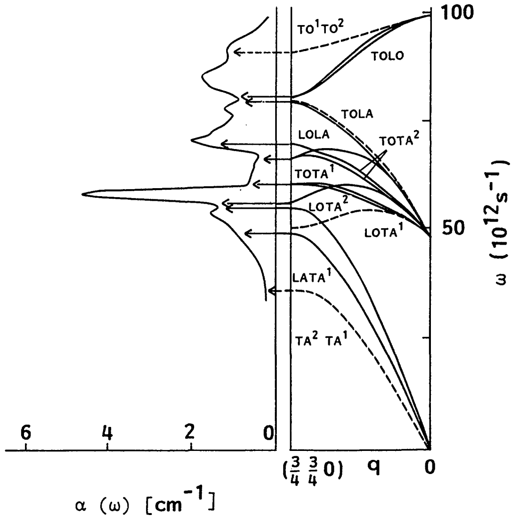

FIG. 4. Two-phonon infrared absorption $\alpha(\omega)$ in silicon at 20 K and the identification of some of the critical points in terms of the two-phonon dispersion relation in the [110] direction (adapted from Bilz, 1966).

4.1 Raman Scattering

In Raman scattering, a photon of light is scattered inelastically from a crystal, with the change in energy and momentum being taken up in the creation or destruction of one or more phonons. In the case where a single acoustic phonon is involved, the process is termed Brillouin scattering. Raman scattering comes about through the modulation of the frequency-dependent electronic polarizability and hence refractive index by the lattice modes. In first order, the interaction terms take the form $b(\mathbf{q}_i)b^\dagger(\mathbf{q}_f)[a^\dagger(-\mathbf{k}j) + a(\mathbf{k}j)]$. For all practical purposes, the wave vectors $\mathbf{q}_i \approx \mathbf{q}_f \approx 0$ for the incident and scattered photons, and so it is only the near–zone center optical modes, whose frequencies are independent of $\mathbf{k}$ to first order, that are affected.

The spectrum of the inelastically scattered light consists of two sets of lines with frequency shifts

$$\Delta\omega = \omega_i - \omega_f = \pm\omega(\mathbf{k}j), \quad \mathbf{k} \approx 0, \tag{40}$$

where ω_i and ω_f are the incident and scattered photon frequencies. The + sign denotes the *Stokes component*, which corresponds to phonon creation and has intensity proportional to $n(\mathbf{k}j) + 1$, and the − sign denotes the *anti-Stokes component*, which corresponds to phonon destruction and has intensity proportional to $n(\mathbf{k}j)$.

The quadratic dependence of the polarizability on the lattice displacements gives rise to second-order Raman scattering comprising Stokes and anti-Stokes summation and difference bands. In the Stokes summation band, for instance, the scattered intensity is given by (Hayes and Loudon, 1978)

$$\begin{aligned} I(\omega_f) = \sum_{\mathbf{k}jj'} & \frac{B(\mathbf{k}jj')}{\omega(\mathbf{k}j)\omega(-\mathbf{k}j')} \\ & \times \{1 + n(\mathbf{k}j)\}\{1 + n(-\mathbf{k}j')\} \\ & \times \delta(\omega_i - \omega_f - \omega(\mathbf{k}j) - \omega(-\mathbf{k}j')), \end{aligned} \tag{41}$$

where $B(kjj')$ are a set of coupling coefficients. Much the same comments apply here as earlier to two-phonon infrared absorption.

4.2 Brillouin Scattering

Brillouin scattering involves scattering by acoustic modes whose frequencies are proportional to k, and so the frequency shifts depend sensitively on the scattering geometry and incident photon frequency ω_i. Since these shifts are small compared with ω_i, it follows that $|\mathbf{q}_i| \approx |\mathbf{q}_f|$, and momentum conservation implies that $k = 2q_i \sin(\theta/2)$, where θ is the scattering angle (see Fig. 5). By changing θ, the direction of $\mathbf{k}$ can be varied. For each scattering geometry, there are in general three Stokes and three anti-Stokes components with frequency shifts (Reissland, 1973)

$$\Delta\omega = v(\mathbf{k}j)k = \frac{2v(\mathbf{k}j)\eta\omega_i}{c}\sin(\theta/2), \tag{42}$$

where η is the refractive index of the me-

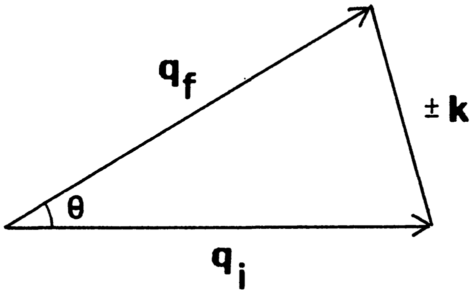

FIG. 5. Momentum-conservation condition for Brillouin scattering.

dium, c is the velocity of light in vacuum, and $v(\mathbf{k}j)$ are the three directionally dependent acoustic phase velocities. These velocities are functions of the bulk elastic constants, and this has led to Brillouin scattering being one of the more important methods for measuring the anisotropic elastic constants of crystals (see Every, 1994a). Such measurements, carried out under variable conditions of temperature and pressure, are a valuable source of information on phase transitions (see, e.g., Dil, 1982; Comins *et al.*, 1990).

There is considerable interest at present in *surface Brillouin scattering* (Nizzoli and Sandercock, 1990). The kinematic conditions are similar to those of He-atom scattering, except that because of the long wavelengths involved, the crystal behaves like an elastic continuum and there are only normal (N) processes to consider. One of the advantages of surface Brillouin scattering is that it can be applied to opaque materials. A drawback of the technique is that because the scattering volume is small, the scattered intensity is weak and sophisticated detection techniques are required.

5. ELECTRON–PHONON INTERACTION

Electrons are an inseparable part of any solid and lie at the very heart of lattice dynamics. They determine the crystal-lattice spacing and contribute in large part to the effective interatomic potential Φ. The problem of the coupled motion of the electrons and the nuclei is made manageable by the fact that the electrons are 2000 or more times less massive than the nuclei, move very fast in comparison, and are by and large able to adjust adiabatically to the instantaneous positions of the nuclei. These ideas are formalized in the *adiabatic* or *Born–Oppenheimer approximation* (see Maradudin, 1974). For many purposes, as we have seen earlier, the contribution of electrons to lattice vibrations can be treated phenomenologically by picturing them in bonds, or rigidly attached to atoms, or in shells that can be displaced relative to ion cores. Our concern in this section is with the more subtle effects that arise from the interaction between phonons and mobile electrons located within $\sim k_B T$ of the Fermi surface in metals and in semiconductors. These interactions come from small residual terms that are neglected in the Born–Oppenheimer approximation. This is a very extensive field, and we will merely touch on some of the many interesting effects that there are.

5.1 Metals

In a perfect metallic crystal, we may as a first approximation consider the electrons to be noninteracting and occupying Bloch states $\psi_{\mathbf{q}\sigma}$, which are characterized by crystal momentum $\hbar\mathbf{q}$ and a band index σ. The electrons are not scattered by the ions in their lattice positions, since the periodic potential of the lattice is already taken into account in constructing the Bloch states. Crystal-lattice vibrations represent departures from perfect periodicity and therefore do cause the electrons to be scattered, and this is the dominant source of temperature-dependent electrical resistance in most metals.

For small displacements of the ions from their equilibrium positions, the change in the potential the electrons experience at any point $\mathbf{x}$ is linear in the displacements. The electron-lattice interaction can thus be expressed as (Ziman, 1965)

$$H_{eL}(\mathbf{x}) = \sum_{\mathbf{k}j} U_{\mathbf{k}j}(\mathbf{x})(a^{\dagger}_{-\mathbf{k}j} + a_{\mathbf{k}j}), \tag{43}$$

where the form of the functions $U_{\mathbf{k}j}(\mathbf{x})$ depends on the specific model being considered, subject to preserving the symmetry of the lattice. In first order of perturbation theory, H_{eL} gives rise to the scattering processes depicted in Figs. 6(a) and 6(b), in which an

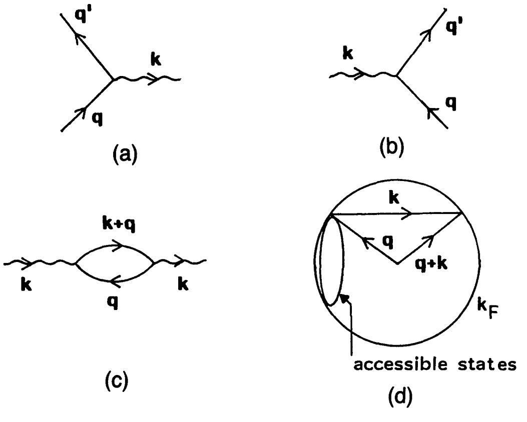

FIG. 6. The electron–phonon interaction in metals.

electron undergoes a transition from one Bloch state $\psi_{\mathbf{q}\sigma}$ to another $\psi_{\mathbf{q}'\sigma'}$, with the simultaneous creation or destruction of a phonon. Conservation of energy and momentum require that

$$E_{\mathbf{q}\sigma} - E_{\mathbf{q}'\sigma'} = \pm\hbar\omega(\mathbf{k}j), \tag{44}$$

$$\Delta(\mathbf{q} - \mathbf{q}' - \mathbf{k}) = 0. \tag{45}$$

The scattering probability is proportional to $n(\mathbf{k}j) + \frac{1}{2} \pm \frac{1}{2}$.

Electron–phonon scattering limits the electron mobility and is responsible for the *ideal electrical resistance* of metals, which at temperatures well below the Debye temperature is proportional to T^5 (Ziman, 1965). Likewise, phonons have greatly reduced mean free paths because of their coupling to the electrons, and so contribute minimally to thermal conduction in metals, and ballistic phonon transport is not observed. At low temperatures, where phonon U processes are rare, the displaced momentum distribution of the electrons associated with an electrical current is communicated to the phonons, so that they too acquire a skewed momentum distribution. This is known as the *phonon drag effect* and is observed as a heat flux in the direction of the current.

In second order of perturbation, H_{eL} leads to a renormalization of the phonon frequencies and electron effective masses. The change in phonon frequencies comes from the type of process depicted in Fig. 6(c), in which the intermediate state is that of the phonon **k** destroyed and an electron scattered from **q** to **q** + **k**. Figure 6(d) depicts these wave vectors in relation to a spherical Fermi surface, near which **q** and **q** + **k** must be located in order to make a significant contribution. Evidently, it is only when k is less than the diameter of the Fermi sphere that there is an effect. A singularity appears in the phonon dispersion relation as k crosses this threshold. The precise form and location of this singularity, which is known as a *Kohn anomaly*, depends on the geometry of the Fermi surface (Ziman, 1965). There may be more than one anomaly in a particular direction in the Brillouin zone, as is the case with the surface-phonon dispersion relation of Pt in Fig. 3.

5.2 Metallic Superconductors

The electron–lattice interaction is the origin of the attractive potential that is responsible for Cooper pair formation and superconductivity in metals. As the first of the two electrons in a *Cooper pair* moves through the lattice, it pulls the positive ions inward, and the second electron is then drawn toward the slightly compacted positively charged region. The net result is an attraction between the two electrons.

The effective interaction between the electrons is characterized by the matrix elements representing scattering of one electron from **q** to **q** − **k**, and a second electron from **q**′ to **q**′ + **k**. The scattering is mediated by the exchange of a *virtual phonon* of wave vector **k**. The process is second order in H_{eL} and involves intermediate states in which one of the electrons is scattered with the creation or destruction of a momentum-conserving phonon. The two processes, which have energy denominators $E_{\mathbf{q}} - E_{\mathbf{q}-\mathbf{k}} \mp \hbar\omega(\mathbf{k})$, are additive, and the resulting scattering matrix element therefore has the form (Ziman, 1965)

$$\begin{aligned}&\langle \mathbf{q} - \mathbf{k}, \mathbf{q}' + \mathbf{k}|V|\mathbf{q}, \mathbf{q}'\rangle \\ &\quad = \frac{|(U_{-\mathbf{k}j})_{\mathbf{q},\mathbf{q}-\mathbf{k}}|^2 2\hbar\omega(\mathbf{k})}{\{E_{\mathbf{q}} - E_{\mathbf{q}-\mathbf{k}}\}^2 - \{\hbar\omega(\mathbf{k})\}^2}.\end{aligned} \tag{46}$$

For

$$|E_{\mathbf{q}} - E_{\mathbf{q}-\mathbf{k}}| < \hbar\omega(\mathbf{k}), \tag{47}$$

the effective interaction is negative, corresponding to an attractive force between the two electrons. The details of how this coupling leads to the formation of an energy gap and superconductivity lie beyond the scope of this article, but certain salient features of superconductivity are already evident in the form of Eq. (46). One is that high electrical resistance in the normal metal and a high superconducting transition temperature T_c go hand in hand, since they both depend on the electron lattice interaction H_{eL} being large. Another is that because of the factor $2\hbar\omega(\mathbf{k})$ in the numerator of Eq. (46), which is proportional to $M^{-1/2}$, where M is the atomic mass,

$$T_c \propto M^{-1/2}. \tag{48}$$

The observation of the dependence of T_c on the isotopic composition has been one of the crucial tests of the Cooper pairing model of superconductivity in metals.

5.3 Semiconductors

The electron–phonon interaction plays an important role in the dynamics of electrons and holes in semiconductors. At low temperatures and in the absence of external excitation, it is the coupling to long-wavelength acoustic modes that is most important, and this is usually modeled by the *deformation potential* (see, e.g., Challis *et al.*, 1987). To illustrate the basic idea, let us consider electrons in a conduction band that has an energy minimum at $\mathbf{q} = 0$. For small lattice displacements, the band is assumed to be shifted vertically in proportion to the lattice dilatation $\Delta = \Sigma_\alpha \, \partial u_\alpha / \partial x_\alpha$, while the band curvature, which determines the electron effective mass m^*, is unaffected to first order. The electron–lattice interaction is thus given by

$$
\begin{aligned}
H_{eL}(\mathbf{x}) = \Xi\Delta = \Xi \sum_{\mathbf{k}j} i\mathbf{k}\cdot\mathbf{e}(\mathbf{k}j) \\
\times \sqrt{\frac{\hbar}{2NM\omega(\mathbf{k}j)}}\, e^{i\mathbf{k}\cdot\mathbf{x}} \\
\times [a^\dagger(-\mathbf{k}j) + a(\mathbf{k}j)], \qquad (49)
\end{aligned}
$$

where Ξ is the deformation-potential constant.

As in the case of metals, H_{eL} gives rise to electron scattering with the creation or annihilation of a phonon. This limits the carrier mobility and is a determining factor in the electrical resistance. There has been considerable work done at low temperatures on phonon emission and absorption and phonon drag by two-dimensional electron gases (2DEGs) in the inversion layer of Si MOSFETs, in GaAlAs heterostructures, and elsewhere (Challis *et al.*, 1987). The converse of phonon drag, known as the *phonon wind effect*, has been very beautifully demonstrated in germanium (Markiewicz *et al.*, 1980). Under intense optical excitation at low temperatures, a condensed state of electrons and holes known as electron-hole liquid is formed. Droplets of this liquid are propelled into the crystal by the capture of momentum from phonons generated in the region of excitation, and this motion is revealed by the spreading recombination luminescence (for a review, see Keldysh and Sibeldin, 1986).

Interband optical absorption in semiconductors involves predominantly direct or "vertical" transitions, in which the wave vector of the electron is unchanged. *Indirect band-gap transitions* take place with much lower probability, and do so because of the slight mixing of different electron–phonon product states of the same total momentum by H_{eL}. Conservation of momentum requires that a transition involving a change in electron momentum by $\hbar\mathbf{q}$ is accompanied by the creation or annihilation of a phonon of momentum $-\hbar\mathbf{q}$ or $\hbar\mathbf{q}$ respectively (Ziman, 1965). The optical absorptivity near an indirect band-gap threshold is thus much weaker than for direct band-gap transitions, and the lifetime of carriers against radiative recombination is much larger.

5.4 Insulators

In polar insulators such as the alkali halides, the effective mass of conduction electrons is increased by the electron–lattice interaction by as much as a factor of 2 or more (Kittel, 1986). An electron pulls the nearby positive ions toward itself while pushing the negative ions away, and this adds to the inertia of the electron, thereby increasing its effective mass. The combination of the electron and ion displacements is known as a *polaron*.

6. PHONON–DEFECT INTERACTIONS

Lattice imperfections such as vacancies, interstitial and substitutional impurities, isotopic differences, dislocations, and grain boundaries, etc., have a marked effect on the vibrational properties of crystals. They destroy the translational symmetry of the lattice, thereby giving rise to momentum-nonconserving phonon scattering, defect-activated infrared absorption, and various other phenomena. They modify the local dynamical response of the lattice, resulting under certain conditions in the appearance of in-band resonances or local vibrational modes. Local electronic and spin states couple to the vibrations of nearby atoms, allowing an exchange of energy between these states and the phonons. This gives rise to *energy relaxation* to the phonon bath and provides a mechanism for doing *phonon spectroscopy* on defects. When the coupling is strong, the local electronic and vibrational states need to be considered as a composite system of *vibronic states*, and optical transitions occur in which the electronic and vibrational states simultaneously change. An extensive review covering most of these topics is Maradudin (1966).

6.1 Local Modes

The vibrational properties of crystal lattices containing defects are usually treated by the *dynamical response-function* approach (Maradudin, 1966; Elliott, 1966). The method involves expressing the dynamical response of the defective crystal in terms of the response of the perfect crystal and the local modification to the dynamical matrix brought about by the presence of the defect, thus obviating the need to determine from scratch the eigenvalues and eigenvectors of $3rN$ coupled oscillators.

The method can be illustrated in a simple way with the example of an isotopic defect. Consider an otherwise perfect silicon crystal in which the atom at the origin is replaced by one of a different mass without altering any of the force constants. For the unperturbed crystal, the dynamical response of the silicon atom to a sinusoidal force $\exp(i\omega t)$ is given by

$$G_0(\omega) = \frac{1}{6NM}\sum_{\mathbf{k}j}\left\{\frac{1}{\omega^2(\mathbf{k}j) - \omega^2} + \frac{i\pi}{2\omega(\mathbf{k}j)}\delta(\omega - \omega(\mathbf{k}j))\right\}, \quad \omega > 0. \tag{50}$$

The terms in the sum represent the real and imaginary parts of the response function of an individual harmonic oscillator, and the factor $1/6\,NM$ comes from projecting the force acting on the atom in question onto each normal coordinate of the lattice, and then projecting the displacement response of that mode back onto that atomic coordinate. More generally, there would be a factor of the form $e_\alpha^*(\kappa|\mathbf{k}j)e_\beta(\kappa'|\mathbf{k}j)$ inside the sum, but this comes out to be equal to $\frac{1}{6}$ because of the symmetry of the Si lattice.

If the atom at the origin is now replaced by one of different mass $M + \delta M$, then the applied force has the additional inertia of δM to contend with, and it is readily shown that the perturbed lattice response function is given by (Elliott, 1966)

$$G(\omega) = \frac{G_0(\omega)}{1 - \delta M\omega^2 G_0(\omega)}. \tag{51}$$

Force-constant changes can be handled in a similar way. If the impurity is charged, it will give rise to one-phonon infrared absorption proportional to $\omega Im(G(\omega))$. For $\delta M = 0$, it follows from Eqs. (50) and (51) that the absorptivity is proportional to the lattice frequency distribution function $D(\omega)$. For $\delta M \neq 0$, the absorption still extends over the entire band of frequencies of the perfect lattice but is modified in intensity. In the case of a light impurity for which $\delta M < 0$, there may in addition be absorption outside the perfect-lattice spectrum at a discrete frequency ω satisfying (Elliott, 1966)

$$1 - \delta M\omega^2 G_0(\omega) = 0. \tag{52}$$

This corresponds to a local mode in which the displacement is mainly confined to the impurity atom and its first few shells of neighbors. Conversely, for a heavy impurity there may be an in-band resonance. These resonances feature in infrared absorption and also in the resonant scattering of bulk phonons. Figure 7 shows calculated infrared absorption spectra for various substitutional

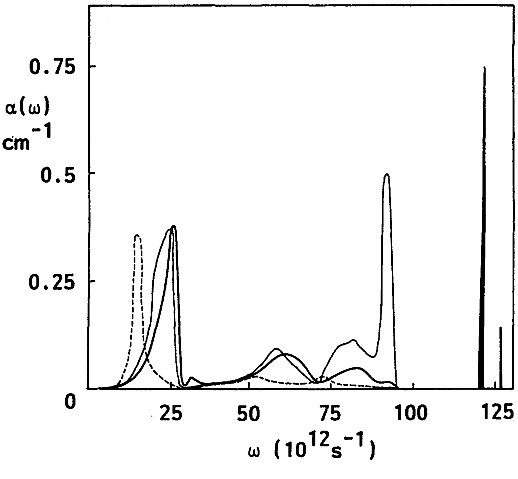

FIG. 7. Calculated infrared absorption spectra for substitutional boron and antimony in silicon (adapted from Elliott, 1966).

impurities in silicon. The full curves are for ^{10}B and ^{11}B in their natural abundances, which are lighter than the host lattice Si atom and give rise to the local modes above 10^{14} s^{-1}. The dashed curve is for Sb, which is heavier than Si and gives rise to the low-lying *in-band resonance* at $\sim 15 \times 10^{12}$ s^{-1}. Quantities such as correlation functions and mean square displacements, etc., are readily calculated once the modified response functions are known. For long-wavelength bulk acoustic modes well below any resonance, the scattering rate from point defects is proportional to $1/\lambda^4$, in accordance with the *Rayleigh scattering* law.

6.2 Electronic States

Electronic states localized at a defect are perturbed by the motion of neighboring atoms, and this provides a mechanism for the exchange of energy. In the case where the electronic energy level splitting lies within the band of vibrational frequencies, transitions can occur with the emission or absorption of a single phonon. The Cr^{3+} laser impurity center in ruby is a case in point that has a transition at 29 cm^{-1} between two excited states (see Renk, 1986; Kaplyanskii and Basun, 1986). By means of optical pumping into the upper of the two excited states and subsequent phonon emission, very high occupation numbers for 29-cm^{-1} phonons can be achieved, leading to *stimulated phonon emission*. The coupling also gives rise to the resonant scattering and *trapping* of 29-cm^{-1} phonons.

The effect of coupling between the localized electronic states and phonons is also seen in the optical absorption and fluorescence spectra of crystals. A grossly oversimplified model, which nevertheless conveys some of the essential features, is to assume that there is a single lattice mode of frequency ω_0 that is linearly coupled to the electronic states. The equilibrium positions about which the oscillations occur in the ground and excited electronic states are thus displaced relative to one another, by ΔQ, say. There are now a multitude of transitions possible, determined by the matrix of overlap integrals between the two sets of oscillator states, in which the optical and vibrational states simultaneously change. At $T = 0$, the probability of a transition occurring in which p phonons are created is proportional to (Fitchen, 1968)

$$W_p = e^{-S_0} S_0^p/p!, \tag{53}$$

where

$$S_0 = M\omega_0(\Delta Q)^2/2\hbar. \tag{54}$$

The $p = 0$ transition is called the *zero-phonon line*, and it remains sharp even when a band of coupled modes rather than a single mode is considered and the temperature is finite, whereas the $p > 0$ transitions are broadened. The temperature dependence of the zero-phonon line intensity is dominated by a Debye–Waller factor $\sim\exp(-S_0 \coth[\hbar\omega_0/2k_BT])$, known as the *Huang–Rhys factor*. Many such *vibronic spectra* have been observed in solids (see Fitchen, 1968; Maradudin, 1966).

In fluorescence, the mirror-image spectrum is observed, with p phonons being created, and the remaining energy being carried away by the photon. Figure 8 shows the absorption and emission spectra of the 4150-Å system in diamond, which is associated with a nitrogen impurity complex (Dyer and Matthews, 1957). *Nonradiative transitions* in which all the excitation energy is dumped as phonons occur at an increasing rate as the temperature is raised, thereby decreasing the luminescence efficiency.

6.3 Spin-Lattice Relaxation

Relaxation processes play an important role in paramagnetic resonance. A nuclear or unpaired electron spin belonging to an atom

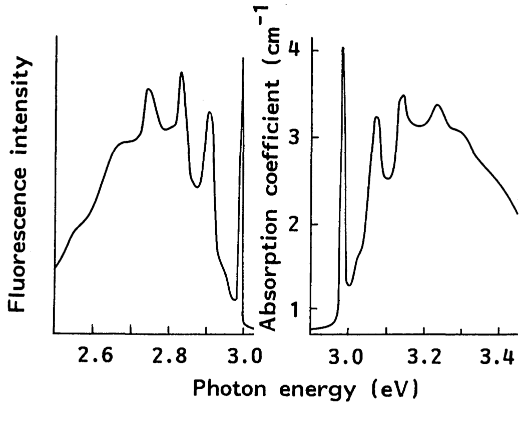

FIG. 8. The fluorescence and absorption spectra of the 4150-Å system in diamond (adapted from Dyer and Matthews, 1957).

in a crystal can exist in various states whose level splittings depend on the crystal field and the external magnetic field. A microwave or radio-frequency field is used to excite the spin system, and various relaxation processes then come into play to return the spin population to equilibrium. Of the many different relaxation processes that there are, we only discuss two that have a direct bearing on phonons.

An important relaxation process for isolated paramagnetic spins in the case where the level splitting is fairly large (which by and large excludes nuclear paramagnetism) is single-phonon emission into long-wavelength acoustic modes (Stevens, 1967). The parameters of the spin Hamiltonian depend on the precise location of the neighboring atoms around the paramagnetic atom and are thus functions of the strain field accompanying acoustic waves in the medium. The linear coupling to the strain yields a spin-lattice interaction Hamiltonian H_{SL} of the form

$$H_{SL} = \sum_{\mathbf{k}j} H_{\mathbf{k}j}(\mathbf{S})|\mathbf{k}| \sqrt{\frac{\hbar}{2NM\omega(\mathbf{k}j)}} \times [a^\dagger(-\mathbf{k}j) + a(\mathbf{k}j)], \tag{55}$$

where $H_{\mathbf{k}j}(\mathbf{S})$ is a spin operator. In first order of perturbation theory, the transition probability p between two spin levels separated in energy by $E = \hbar\omega_0$ is proportional to $D(\omega_0)\bar{n}(\omega_0)\,|\mathbf{k}|^2/\omega_0$. For $k_B T \gg \hbar\omega_0$, it follows that $p \propto \omega_0^2 T$.

If the rate at which phonons are pumped into the narrow band of frequencies around ω_0 exceeds the ability of the lattice to thermalize them, then there is a *phonon bottleneck* effect, in which repeatedly the phonons emitted by one paramagnetic spin are absorbed by another (Stevens, 1967).

For nuclear spins, where ω_0 tends to be relatively small, the number of lattice modes in that frequency region is very small, and the direct relaxation rate is extremely slow. Under these conditions, Raman processes, in which a phonon is absorbed and another of slightly higher frequency emitted, become important. The coupling for this process comes from the quadratic dependence of the spin Hamiltonian on the local strain. All phonons from $\omega = 0$ to $\omega \sim k_B T/\hbar$ contribute, and for $T \ll \Theta_D$ the relaxation rate is independent of the splitting and has a temperature dependence $p \propto T^7$.

6.4 Mössbauer Effect

The Mössbauer effect can be thought of as the nuclear analog of vibronic spectra, in which it is the nucleus rather than the electrons of the atom that undergoes a transition, and the photon that is absorbed or emitted is that of a γ ray rather than of visible light. Corresponding to the zero-phonon line there is a certain fraction of the γ rays that are emitted recoil-free, i.e., without any energy being absorbed by the vibrations of the lattice. The recoil momentum from these γ photons is taken up by the crystal as a whole, and because the crystal is so massive, the associated kinetic energy is totally negligible. The *recoil-free fraction* is given by the Debye–Waller factor (Maradudin, 1966)

$$f = \exp - \langle [\mathbf{K} \cdot \mathbf{u}(l\kappa)]^2 \rangle, \tag{56}$$

where K is the wave vector of the γ ray and $\mathbf{u}(l\kappa)$ is the displacement coordinate of the resonant nucleus. With increasing temperature, as the mean square displacement of the atom increases, f decreases. At higher temperatures, anharmonicity plays an important role in this temperature dependence (Kolk, 1984). There has been considerable theoretical speculation concerning the single- and multi-phonon sidebands, which has greatly outstripped the technical means for measuring them. Part of the problem is that to measure directly a phonon energy shift of ~0.01 eV in a ~10 keV γ ray requires the velocity of the moving stage to be $\sim 10^{-6}c = 300$ m/s.

7. CLOSING REMARKS

This article has mainly concentrated on the interaction between phonons and other "-ons", the physical properties that these interactions impart to crystals, and how the scattering of x rays, neutrons, and other probes can be used to determine phonon dispersion relations. In an article of this nature, it is not possible to give complete coverage to the whole of phonon physics. There has in particular been very little discussion here of phonon transport, which is covered extensively in the articles LV and HP. The first of these two articles contains an extensive discussion of the interplay between boundary scattering, defect scattering, and anharmonic processes in limiting thermal conduction in dielectric crystals. The latter article contains a lot of recent material on *ballistic phonon transport* and its investigation by means of *phonon imaging*. This has shed light on various phenomena, including phonon transmission through surfaces, the scattering of phonons by dislocations, and electron–phonon scattering in semiconductors (for reviews on phonon imaging, see Wolfe, 1989; Every, 1994b). HP also discusses second sound, which is a collective mode of the phonon gas that comes about when *N* processes dominate over *U* and other momentum-nonconserving scattering processes.

A development that is attracting considerable attention at the present time is that of *phonon-mediated elementary-particle detection* (for a review, see Cabrera, 1990). This offers the possibility of much greater energy resolution than can be achieved by other available means, and the prospect of being able to detect the hypothetical particles that are invoked as an explanation of the missing dark matter in our galaxy. While much progress has been made, the technical difficulties are formidable, and the final word on this subject is some way off.

ACKNOWLEDGMENTS

This article was partly written while the author was on leave at the Institute of Applied Mechanics, National Taiwan University. The National Science Council of the Republic of China and the South African Foundation for Research Development are thanked for their financial support. J. Wesner is thanked for critical reading of the manuscript.

Works Cited

Ashcroft, N. W., Mermin, N. D. (1981), *Solid State Physics*, Philadelphia: Saunders.

Bilz, H. (1966), in: R. W. H. Stevenson (Ed.), *Phonons*, Edinburgh: Oliver & Boyd, pp. 208–234.

Bilz, H., Kress, W. (1979), *Phonon Dispersion Relations in Insulators*, Berlin: Springer.

Cabrera, B. (1990), in S. Hunklinger, W. Ludwig, G. Weiss (Eds.), *PHONONS 89*, Singapore: World Scientific, pp. 1373–1382.

Challis, L. J., Toombs, G. A., Sheard, F. W. (1987), in: T. Paszkiewicz (Ed.), *Physics of Phonons*, Berlin: Springer, pp. 348–374.

Comins, J. D., Ngoepe, P. E., Catlow, C. R. A. (1990), *J. Chem. Soc. Faraday Trans.* **86**, 1183–1192.

Cowley, R. A. (1969), in: R. H. Enns, R. R. Haering (Eds.), *Modern Solid State Physics*, Vol. 2, New York: Gordon and Breach, pp. 43–105.

de Launay, J. (1953), *J. Chem. Phys.* **21**, 1975–1986.

Dil, J. G. (1982), *Rep. Prog. Phys.* **45**, 285–334.

Dorner, B., Comes, R. (1977), in: S. W. Lovesey, T. Springer (Eds.), *Dynamics of Solids and Liquids by Neutron Scattering*, Berlin: Springer, pp. 127–196.

Dyer, H. B., Matthews, I. G. (1957), *Proc. R. Soc. A* **243**, 320–335.

Elliott, R. J. (1966), in: R. W. H. Stevenson (Ed.), *Phonons*, Edinburgh: Oliver & Boyd, pp. 377–402.

Every, A. G. (1994a), *NDT&E Int.*, **27**, 3–10.

Every, A. G. (1994b), in: T. Paszkiewicz, T. Rapcewicz (Eds.), *Die Kunst of Phonons*, New York: Plenum, pp. 55–72.

Fitchen, D. B. (1968), in: W. B. Fowler (Ed.), *Physics of Color Centers*, New York: Academic, pp. 293–350.

Ghatak, A. K., Kothari, L. S. (1972), *Lattice Dynamics*, London: Addison-Wesley.

Guyer, R. A. (1969), in: F. Seitz, D. Turnbull, H. Ehrenreich (Eds.), *Solid State Physics*, Vol. 23, New York: Academic, pp. 413–499.

Hass, K. C., Tamor, M. A., Anthony, T. R., Banholzer, W. F. (1992), *Phys. Rev. B* **45**, 7171–7182.

Hayes, W., Loudon, R. (1978), *Scattering of Light by Crystals*, New York: Wiley.

Kaplyanskii, A. A., Basun, S. A. (1986), in: W. Eisenmenger, A. A. Kaplyanskii (Eds.), *Nonequilibrium Phonons in Nonmetallic Crystals*, Amsterdam: North Holland, pp. 373–453.

Keldysh, L. V., Sibeldin, N. N. (1986), in: W. Eisenmenger, A. A. Kaplyanskii (Eds.), *Nonequilibrium Phonons in Nonmetallic Crystals*, Amsterdam: North Holland, pp. 455–686.

Kittel, C. (1986), *Introduction to Solid State Physics*, New York: Wiley.

Kolk, B. (1984), in: G. K. Horton, A. A. Maradudin (Eds.), *Dynamical Properties of Solids*, Vol. 5, Amsterdam: North-Holland, pp. 1–332.

Kress, W. (1991), in: W. Kress, F. W. de Wette (Eds.), *Surface Phonons*, Berlin: Springer, pp. 209–292.

Lax, M., Hu, P., Narayanamurti, V. (1981), *Phys. Rev. B* **23**, 3095–3097.

Lovesey, S. W. (1977), in: S. W. Lovesey, T. Springer (Eds.), *Dynamics of Solids and Liquids by Neutron Scattering*, Berlin: Springer, pp. 1–75.

Maradudin, A. A. (1966), in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, Vol. 18, New York: Academic, pp. 273–420.

Maradudin, A. A., Montroll, E. W., Weiss, G. H., Ipatova, I. P. (1971), *Theory of Lattice Dynamics in the Harmonic Approximation*, in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, Supplement 3, New York: Academic.

Maradudin, A. A. (1974), in : G. K. Horton, A. A. Maradudin (Eds.), *Dynamical Properties of Solids*, Vol. 1, Amsterdam: North-Holland, pp. 1–82.

Markiewicz, R. S., Greenstein, M., Wolfe, J. P. (1980), *Solid State Commun.* **35**, 339–344.

Nizzoli, F., Sandercock, J. R. (1990), in: G. K. Horton, A. A. Maradudin (Eds.), *Dynamical Properties of Solids*, Vol. 6, Amsterdam: North-Holland, pp. 281–335.

Reissland, J. A. (1973), *The Physics of Phonons*, New York: Wiley.

Renk, K. F. (1986), in: W. Eisenmenger, A. A. Kaplyanskii (Eds.), *Nonequilibrium Phonons in Nonmetallic Crystals*, Amsterdam: North-Holland, pp. 317–372.

Schiff, L. I. (1955), *Quantum Mechanics*, New York, McGraw-Hill.

Stevens, K. W. H. (1967), *Rep. Prog. Phys.* **30**, 189–226.

Toennies, J. P. (1991), in: W. Kress, F. W. de Wette (Eds.), *Surface Phonons*, Berlin: Springer, pp. 111–166.

Wolfe, J. P. (1989), *Festkörperprobleme* **29**, 75–105.

Ziman, J. M. (1965), *Principles of the Theory of Solids*, Cambridge, U.K. Cambridge Univ. Press.

Further Reading

Ghatak, A. K., Kothari, L. S. (1972), *Lattice Dynamics*, London: Addison-Wesley.

Maradudin, A. A., Montroll, E. W., Weiss, G. H., Ipatova, I. P. (1971), *Theory of Lattice Dynamics in the Harmonic Approximation*, in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, Supplement 3, New York: Academic.

Reissland, J. A. (1973), *The Physics of Phonons*, New York: Wiley.

Srivastava, G. P. (1990), *The Physics of Phonons*, Bristol: Adam Hilger.

PHOTOACOUSTIC SPECTROSCOPY

See SPECTROSCOPY, PHOTOACOUSTIC

PHOTOCHEMISTRY

See REACTIONS INDUCED BY LIGHT

PHOTODETECTORS*

SUSAN R. SLOAN, *Hewlett-Packard Company, Lightwave Operation, Santa Rosa, California, U.S.A.*

Introduction 459
1. **Physics of Photodetectors** 459
2. **Materials, Design, and Properties of Photodetectors** 460
2.1 Semiconductor Material Systems 460
2.2 Properties of Photodetectors 461
2.2.1 Dark Current 461
2.2.2 Photocurrent 462
2.2.3 Dynamic Resistance 462
2.2.4 Capacitance 462
2.2.5 Responsivity and Quantum Efficiency 463
2.2.6 Reflectivity 463
2.2.7 Response Time 464
2.2.8 Noise 465
3. **Types of Photodetectors** 467
3.1 Photoconductors 467
3.2 *p-n* Junction Photodiodes 467
3.3 *p-i-n* Photodetectors 468
3.4 Metal–Semiconductor–Metal (Schottky Barrier) Photodetectors 469
3.5 Avalanche Photodetectors 470
3.6 Arrays 471
3.7 Waveguide Photodetectors 471
3.8 Quantum-Well Photodetectors 472
3.9 High-Speed Detectors 472
3.9.1 Device Types and Trends 472
4. **Applications in Optical Receivers** 473

INTRODUCTION

An optical communications system requires both a source—laser or light-emitting diode (LED)—and a receiver—a photodetector. Historically, photomultiplier tubes and vacuum photodiodes were available as receivers, but smaller, less expensive, wide-bandwidth semiconductor devices are now widely used. The gas-filled tube devices are limited in bandwidth to around 10 kHz (Melchoir, 1972), and in the case of the vacuum photodiode to relatively high minimum detectable optical powers. The wavelength range over which a device functions is determined by its material structure. Choice of a photodetector is governed by the operating wavelengths of the system, the response to these wavelengths, the speed or frequency response requirements, and the dark current or noise performance of the device.

This article discusses the basic operation of a photodetector, the parameters commonly used for its characterization, semiconductor material systems, design and fabrication techniques for different types of photodetectors, and applications in optical receivers. This first section concentrates on basic photodetector physics. Section 2 covers semiconductor material choices and the properties of photodetectors that need to be characterized to understand their operation. Section 3 focuses on the operation of different types of photodetector devices, the semiconductor fabrication techniques for producing them, and high–speed-device trends. Finally, Sec. 4 examines applications for photodetectors in optical receivers.

1. PHYSICS OF PHOTODETECTORS

The function of a photodetector is to convert incoming modulated light energy into modulated electrical current. Light incident on a semiconductor photodetector is absorbed, and if the photon energy is greater than the band-gap energy, electron–hole pairs

---

*Material in this article is excerpted from Sloan, Susan R. (1994), "Photodetectors," in: Robert G. Huntsperger (Ed.), *Photonic Devices and Systems*, New York: Marcel Dekker, pp. 171–246.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

are generated. These carriers, created in a region of high electric field within the depletion region of the device, are swept out by an applied reverse bias resulting in a photocurrent.

In photomultipliers, vacuum photodetectors, and phototubes, electrons are excited by photons over the energy barrier, or work function, of the material and become free electrons in a process called the external photoeffect (Melchoir, 1972). Semiconductor photodiodes operate on a principle known as the internal photoeffect, where the electrons and holes excited by the photon to higher energy levels remain within the material and contribute to photocurrent. The incident photon absorbed by the semiconductor generates a free electron, which is excited from the valence to the conduction band while a hole is generated in the valence band. The carriers created in the depletion region drift in opposite directions as they are separated by the electric field, and carriers created within a diffusion length of the depletion region diffuse to the edges of the region and are collected. Under reverse bias, the carriers' drift in the electric field induces a photocurrent in the load, R_L. The photocurrent is given by

$$I_P = \frac{P_0 q(1 - r')}{h\nu}(1 - e^{-\alpha d}), \qquad (1)$$

where P_0 is the incident optical power, q is the electrical charge, r' is the Fresnel reflection at the semiconductor–air interface, $h\nu$ is the photon energy, α is the absorption coefficient of the material, and d is the width of the depletion layer.

Figure 1 illustrates a photon incident on a double-heterojunction *p-i-n* photodiode. The photon passes through the larger–band-gap cap material, shown here as *p* type, and is absorbed in the active layer, the *i* region, below, creating the electron–hole pair. As reverse bias is applied, the electron moves toward the *n* side of the device (cathode) and the hole toward the *p* side (anode), creating a photocurrent.

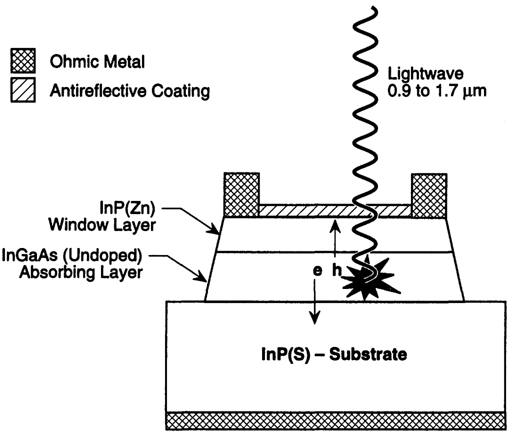

FIG. 1. A cross-sectional view of a *p-i-n* double-heterojunction photodiode illustrating the absorption of the photon, the creation of the electron–hole pair, and their respective motion under influence of the electric field (Sloan, 1989).

2. MATERIALS, DESIGN, AND PROPERTIES OF PHOTODETECTORS

2.1 Semiconductor Material Systems

The photodetector properties are strongly influenced by the choice of the semiconductor material system. For the device active layer, the material band gap is chosen to absorb photons in the wavelength range of interest. Figure 2 illustrates properties of various materials used for optical devices. The left vertical scale shows the energy gap or band gap E_g of materials, in electron volts (eV), the horizontal axis gives lattice constant values in ångströms, and the right vertical scale shows the usable wavelength range in microns. For example, the telecommunications market wavelengths of 1.3 and 1.55 μm can be detected by Ge or InGaAs active-region photodetectors. When InGaAs active regions are used, the III-V material InGaAs is lattice matched to InP substrate material by using 53% In and 47% Ga as the group III element, and As as the group V element, written as $In_{0.53}Ga_{0.47}As$. Germanium is also used for the telecommunications wavelengths, but its response to longer wavelengths, beyond 1500 nm, drops off sharply whereas the curve of response vs wavelength for InGaAs extends to 1700 nm.

The maximum wavelength for which a material will absorb is determined by its band-gap energy—the energy gap between the bottom of the conduction band and the top of the valence band. For wavelengths

$$\lambda_0 > hc/E_g, \qquad (2)$$

the material will not absorb, but becomes transparent. The light absorption is depen-

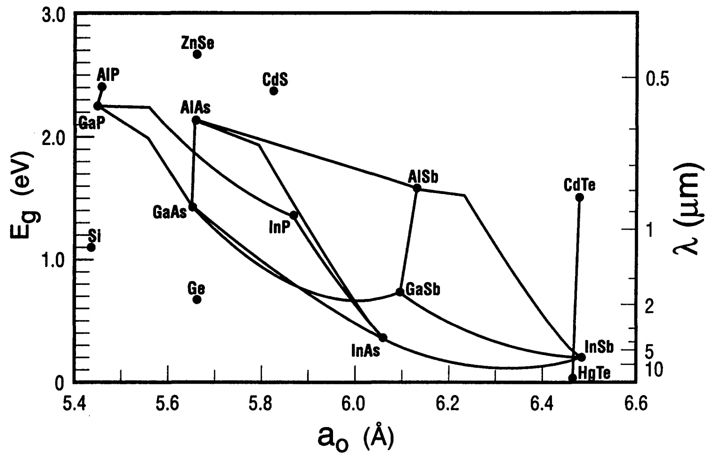

FIG. 2. Common optical-device materials showing their energy band gaps in eV along the left vertical scale, the lattice constants on the horizontal axis, and the usable wavelength range in microns along the right vertical axis. Compositions of the tertiary and quaternary III-V materials may be changed in order to change their detection range as shown by the curves joining the single-III/single-V element compounds (Manasreh, 1993).

dent upon the material absorption coefficient α. In direct–band-gap materials, transitions can occur between the conduction-band minimum and the valence-band maximum with the same wave vector. Indirect materials, where the conduction-band minimum is not at the same value of the wave vector as the valence-band maximum, require a phonon to conserve momentum in addition to the photon of energy $>E_g$. The direct transitions are more probable. The active layer must be thick enough, on the order of $1/\alpha$ (Lucovsky *et al.*, 1964), to absorb the light and create electron–hole pairs. The short-wavelength response cuts off because the absorption coefficient becomes very large at shorter wavelengths; thus absorption occurs near the surface and recombination time is short, and the carriers recombine before they are collected (Sze, 1981).

2.2 Properties of Photodetectors

This section describes important test parameters for photodetectors. A detector's operation is characterized by the wavelength range over which it operates, the response to these wavelengths, the speed of the response, and its noise performance. Low dark current for devices operated under reverse bias and high dynamic resistance for devices operated at zero bias are important for low noise in a receiver system. This section examines sources of noise in photodetectors. For the greatest signal-to-noise performance, responsivity or quantum efficiency should be maximized and reflections minimized. Capacitance, series resistance, and active-layer thicknesses are important considerations for determining the speed of response of the detector.

2.2.1 Dark Current Under reverse bias, the leakage current present in the device in the absence of illumination is referred to as dark current. Low dark current is important for low-noise performance of photodetectors in optical receivers. Dark current of the photodetector is given by the following equation:

$$I_d = I_{d\text{-}G\text{-}R} + I_{d\text{-diffusion}} + I_{d\text{-surface}}, \qquad (3)$$

where $I_{d\text{-}G\text{-}R}$ is the dark current resulting from generation and recombination of carriers, $I_{d\text{-diffusion}}$ results from carrier diffusion, and $I_{d\text{-surface}}$ is dark current contributed by surface leakage (Muller, 1981). By testing devices of various sizes, the dominant dark-current mechanism can be determined for a particular device. A current proportional to the circumference of the device, $I_d \propto r$, where r is the radius of the device (assumed to be round), indicates that surface leakage is dominant; current proportional to the area, $I_d \propto r^2$, is indicative of bulk leakage mechanisms (Carey *et al.*, 1989).

Dark-current measurements are performed by probing or making other electrical connection to the device and applying bias voltage while keeping the device in the dark. Mesa-etched devices, where the active area is defined by removing the surrounding epitaxial material and the *p-n* junction is exposed (such as the InGaAs *p-i-n* structure shown in Fig. 11), typically are dominated by surface leakage. Dark currents in diffused-junction diodes are more likely due to bulk currents.

The following equations describe the dark-current terms in more detail (Forrest *et al.*, 1980; Sze, 1981). The generation-recombination current is given by

$$I_{d\text{-}G\text{-}R} = \frac{qn_iAd}{\tau_{\text{eff}}}\left[1 - \exp\left(\frac{-qV_r}{2kT}\right)\right], \tag{4}$$

where q is the charge, A is the photodetector area, d is the width of the depletion region, τ_{eff} is the effective generation-recombination lifetime, V_r is the applied reverse bias, k is Boltzmann's constant, T is the temperature in kelvins, and n_i, the intrinsic carrier concentration, is given by

$$n_i = 2\left(\frac{4\pi^2 m_e m_h k^2}{h^4}\right)^{3/4} T^{3/2} \exp\left[-\frac{E_g}{2kT}\right], \tag{5}$$

where m_e and m_h are the effective masses of the electron and holes, respectively, and h is Planck's constant. The diffusion current is given by

$$I_{d\text{-diffusion}} = \frac{qn_i^2(D_n/\tau_m)^{1/2}A}{N_a}\left[1 - \exp\left(\frac{-qV_r}{kt}\right)\right], \tag{6}$$

where the additional terms in this equation are D_n, the minority-carrier diffusion constant, τ_m, the minority-carrier lifetime, and N_a, the acceptor concentration. When the photodetector is operated under a reverse bias, the temperature dependence of the generation-recombination and diffusion currents is dominated by the n_i terms. That means that

$$I_{d\text{-}G\text{-}R} \propto \exp[-E_g/2kT] \tag{7}$$

and

$$I_{d\text{-diffusion}} \propto \exp[-E_g/kT]. \tag{8}$$

Plots of the logarithm of the dark current vs $1/T$ in kelvins show that the surface leakage-current behavior is due to generation currents believed to be due to traps near the mesa edge (Carey *et al.*, 1989; Forrest *et al.*, 1980).

2.2.2 Photocurrent A photodiode is generally operated under reverse bias in order to increase the width of the depletion region, thereby lowering the capacitance, and making a stronger electric field to increase the drift velocity of the electrons and holes and minimize the carrier transit time (Saleh and Teich, 1991). The photocurrent produced, I_p, is given by

$$I_p = \frac{P_0q(1 - r')}{h\nu}(1 - e^{-\alpha d}) = \eta q\frac{P_0}{h\nu}, \tag{9}$$

where η is the quantum efficiency discussed in Sec. 2.2.5. The current–voltage (I-V) characteristic for a photodiode is given by

$$I = I_d\left[\exp\left(\frac{eV_0}{n_0kT}\right) - 1\right] - I_p, \tag{10}$$

including the photocurrent portion I_p, and the normal diode diffusion current discovered by Shockley (1949) where I_d is the dark current, n_0 is the diode ideality factor (Sze, 1981)

$$n_0 \equiv \frac{q}{kT}\frac{\partial V}{\partial(\ln I)}, \tag{11}$$

and V_0 is the junction voltage (Lee and Li, 1979).

2.2.3 Dynamic Resistance For photodetectors operated at 0 V bias, the parameter of interest becomes the dynamic resistance, sometimes called shunt resistance, instead of dark current. The zero bias allows for low-noise receiver designs because little or no dark current is present. Dynamic resistance R_d is measured by examining the I-V curve of a device and taking the slope of the curve ($\Delta V/\Delta I$) around the zero-volt region. R_d is controlled by the steepness of the forward voltage characteristic of the diode vs current and the reverse leakage current of the device. For the case of only thermal (Johnson) noise,

$$\langle i_t^2\rangle = 4kTB/R_d, \tag{12}$$

where i_t is the thermal noise current and B is the bandwidth of the measurement. For low-noise operation, R_d should be maximized. Typical dynamic resistance values for 500-μm–active-area InGaAs *p-i-n* photodiodes are ~100 MΩ.

2.2.4 Capacitance Measuring a photodetector's capacitance vs voltage is a way to look at the degree to which the active region of the device is depleted of carriers. For a photodiode, the junction capacitance is given by

$$C = A\left[\frac{2}{q\epsilon}(\Phi + V_a)\left(\frac{1}{N_a} + \frac{1}{N_d}\right)\right]^{-1/2} = \frac{\epsilon A}{d} \tag{13}$$

(Lee and Li, 1979), where A is the junction area, q is the charge, ϵ is the dielectric constant of the i-layer material, Φ is the built-in voltage, V_a is the applied voltage, N_a and N_d are the acceptor and donor concentrations, respectively, and d is the width of the depletion region. Capacitance can be reduced by lowering the doping in the i layer or by increasing the applied voltage. Minimizing capacitance for a *p-i-n* photodetector means minimizing device active area and using a thick, low-doped i layer. The RC time constant of the device, series resistance times capacitance, is key in determining the frequency response of the device. This must be traded off against the increased carrier transit time due to a thicker i layer.

2.2.5 Responsivity and Quantum Efficiency Responsivity is a measure of photodetector sensitivity to incoming light energy. To be most efficient, the photodetector should turn as many of the incoming photons as possible into detectable photocurrent. Responsivity is measured by illuminating the device with a known optical power at a known wavelength, and measuring the magnitude of the resulting photocurrent at a particular bias value; the units are amperes/watt.

The thickness of the *p-i-n* device i layer determines the response. If the i layer is not thick enough to absorb the incoming photons, the response will be less than the maximum calculated value. The absorption coefficient is given as a function of wavelength. For InGaAs at the wavelengths of 1.06 and 1.3 μm, the absorption coefficients are $\alpha(1.06\ \mu\text{m}) = 1.8\ \mu\text{m}^{-1}$ and $\alpha(1.3\ \mu\text{m}) = 1.15\ \mu\text{m}^{-1}$. At a thickness of $1/\alpha$, the light intensity falls to $1/e$ of the original value. The fraction of power absorbed in a depth d is given by

$$P_{\text{abs}} = P_0[1 - \exp(-\alpha d)]. \tag{14}$$

The photon energy must be greater than the semiconductor band gap in order for absorption to occur and carriers contributing to photocurrent to be produced.

The number of electron–hole pairs created by a single incident photon is called the quantum efficiency,

$$\eta = (I_p/q)(h\nu/P_0), \tag{15}$$

where I_p is the photocurrent, q is the magnitude of the charge on a carrier, and P_0 is the optical power at the wavelength corresponding to the photon energy given by $h\nu$. Some authors and manufacturers quote responsivity values while others quote quantum efficiency as the measure of the device optical sensitivity. Quantum efficiency is related to responsivity R by

$$\eta = \frac{1}{q}\frac{hc}{\lambda}R = \frac{1.24}{\lambda}R. \tag{16}$$

By definition, the maximum value of η is 1. Maximum response values for a particular wavelength may be found by setting $\eta = 1$ and finding R.

Responsivity or quantum efficiency vs wavelength is measured using a spectrometer and large–wavelength-range sources. The power is monitored by using an optical coupler to sample a portion of the light energy, which is incident on a power meter. The photocurrent from the photodetector is measured and recorded at each wavelength. The resulting plot for two different InGaAs devices is shown in Fig. 3. One is a conventional *p-i-n* with an InP cap layer, and one has been modified to extend its range to shorter wavelengths. Here, the response was measured from 600 to 1700 nm. Figure 4 shows responses for a variety of material systems (Melchoir, 1972; Sze, 1981) allowing the receiver designer to select the material appropriate for her application. Maximizing response in a receiver system is a way of achieving high signal-to-noise performance.

2.2.6 Reflectivity If the surface of the photodetector is highly reflective, a large part of the incoming light will not be absorbed and not result in the creation of electron–hole pairs, or photocurrent. This will limit the device responsivity. The reflected light may also be coupled back into the light source, perturbing it. This is highly undesirable for laser sources. To prevent this, an antireflective

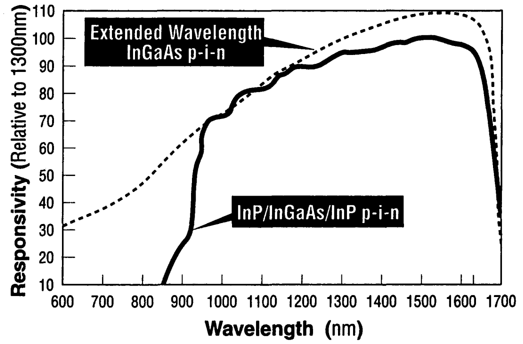

FIG. 3. Plot of dc responsivity vs wavelength for two InGaAs devices. The solid line illustrates a standard InP/InGaAs/InP response, and the dashed line shows an extended-wavelength InGaAs device. Both are used for the long telecommunications wavelengths (measurements courtesy of Mike McClendon and Norm Brewer, Hewlett-Packard Lightwave Operation Standards Lab).

coating is applied to the device surface. The goal is to minimize the percentage of optical power reflected or maximize the optical return loss (ORL). Reflection is measured as a percentage of incident light reflected or in optical receivers as the figure of merit optical return loss (ORL), which is in units of dB and is given by

$$\mathrm{ORL} = 10 \log(P_0/P_r), \tag{17}$$

where P_0 is the incident power and P_r is the reflected power. With no coating, a semiconductor surface reflects about 30% of the light that is incident upon it.

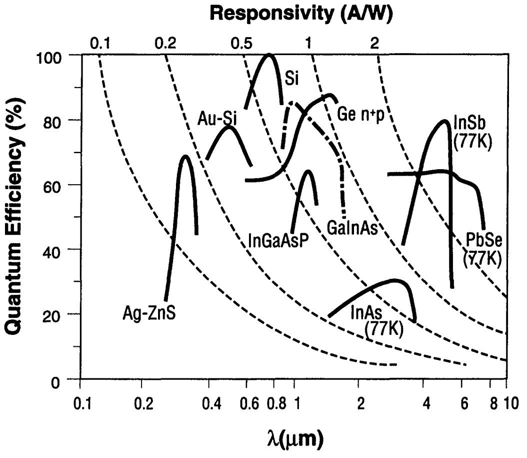

FIG. 4. Response curves for various detector materials (Sze, 1981).

Antireflective coatings are usually made of dielectric materials deposited on the surface of the device. To first approximation, a quarter wavelength of dielectric material is deposited. A plot of the modeled power reflected vs wavelength is shown in Fig. 5 for a single-layer antireflection coating of silicon nitride on an InP/InGaAs/InP photodetector. If optimized for InP/InGaAs/InP devices, dielectric coatings can lower the reflected light to on the order of 0.5% at 1300 and 1550 nm (Braun, 1988). To improve the coating further, multiple layers are used. To prevent the reflected light from the detector from coupling back to the laser source, the device is usually tilted. A 6- to 8-degree angle of the device relative to the normal will prevent light from being coupled back into an optical single-mode fiber from the device.

2.2.7 Response Time Photodetector response time is governed by the transit time of carriers in the active region and the *RC* time constant of the device. The carrier transit time is the distance the carriers travel across the *i* layer divided by the velocity at which the carriers travel. Both electrons and holes are present, and each has a saturation velocity at which it travels:

$$t = d/v_{sat}. \tag{18}$$

To maximize the speed of response, the transit time needs to be minimized. This means making the active layer thin. By making the active region thin, the capacitance of the de-

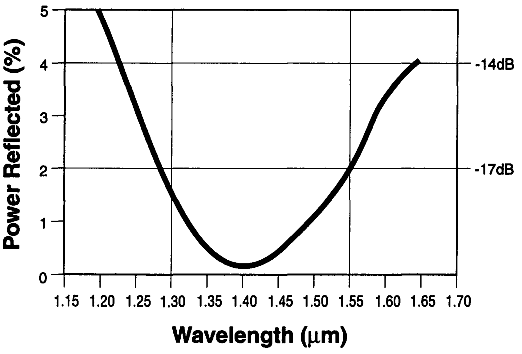

FIG. 5. Reflection vs wavelength characteristic for modeled single-layer antireflection coating for the telecommunications wavelength range (Beck, 1993).

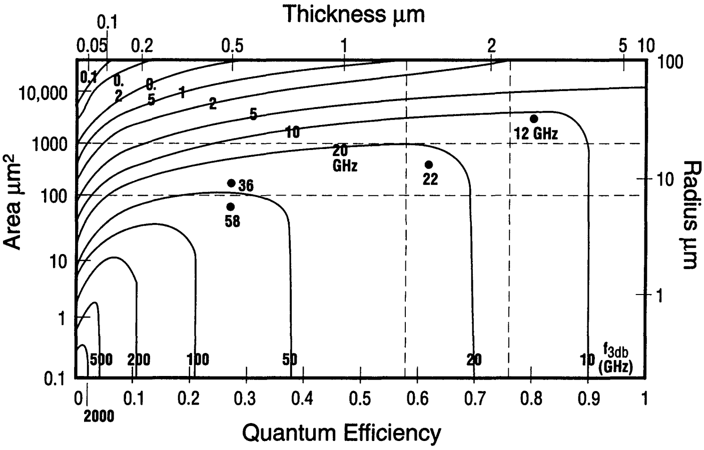

FIG. 6. Frequency-response contours vs active-area diameter, active area, and quantum efficiency (Bowers *et al.*, 1986a).

vice is increased and the quantum efficiency of the device is decreased. The increasing capacitance causes the *RC* time constant to rise and the speed is slowed. Tradeoffs must be made between transit time and *RC* considerations in order to optimize the photodetector speed of response. These tradeoffs have been modeled for InP/InGaAs/InP p-i-n devices and are plotted in Fig. 6 (Bowers *et al.*, 1986) where the device area, *i*-layer thickness, and active-area diameter are examined and frequency contours are established.

2.2.8 Noise Low-noise performance is important for photoreceivers where low power levels need to be detected. The minimum detectable optical signal is determined by the photodetector quantum efficiency and the noise sources present in the receiver. Sources of noise in photodetectors include shot noise and thermal or Johnson noise. Shot noise consists of noise due to dark current, surface leakage current, and quantum effects (Lee and Li, 1979). Assuming dark current and surface leakage are negligible by design, shot noise is due to the fluctuations in the photoexcitation of carriers. It is present whenever current is flowing, as a result of fluctuations in the arrival rate of electrons which in turn lead to random changes in the current flow (Wilson and Hawkes, 1989). This is the lowest noise achievable in the system. For devices with gain such as avalanche photodetectors (APDs), excess shot noise exists due to the multiplication process. Thermal noise from the load resistor in a receiver limits noise in a device without internal gain.

A simple model of a photodetector receiver is shown in Fig. 7. The photodiode is modeled as a current source with a small series resistance R_s due to the resistance of the *p* contacting layer and the metal–semiconductor resistance, a junction capacitance C_j, and a load resistance R_L. The amplifier capacitance and resistance are labeled as C_a and R_a respectively. When a time-varying optical power is incident, the resulting photocurrent is

$$i_{ph}(t) = \eta q p(t)/h\nu, \tag{19}$$

where η is the quantum efficiency, q the electrical charge, and $p(t)$ the modulated incident power, and $h\nu$ is Planck's constant times the frequency (Lee and Li, 1979). The term $i_{ph}(t)$ has a dc component

$$I_p = \eta q P_0/h\nu \tag{20}$$

(Sze, 1981), and a time-varying current, or signal component, i_p. For a device with gain,

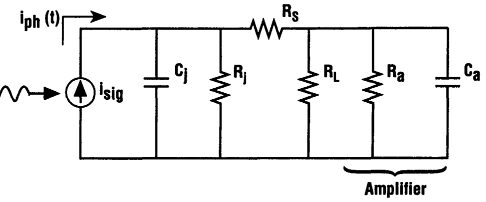

FIG. 7. Model of a photodetector as a current source with a junction resistance, junction capacitance, series resistance, resistance due to the load, R_L, and the resistance and capacitance due to the amplifier following the photodiode.

the signal current is multiplied by the internal avalanche gain. The average value of the signal current is then

$$\langle t_{sig}\rangle = i_p\langle m\rangle, \tag{21}$$

where $\langle m\rangle$ is an average of the varying multiplication factor. For devices without gain such as *p-i-n*, $\langle m\rangle = 1$.

The shot-noise current is determined by the dc component of the photocurrent, I_p, and is given by

$$\langle i_{sp}^2\rangle = 2qI_p\langle m^2\rangle B, \tag{22}$$

where $\langle m^2\rangle$ is the mean square multiplication factor and B is the bandwidth of the noise measurement. Noise-current terms are commonly written as mean squares ($\langle i^2\rangle$). The shot noise from photocurrent and dark current can be combined into a single expression. Dark currents can arise from different sources. Electrons and holes thermally generated in the *p-n* junction of a device contribute to bulk dark current (Keiser, 1983) and currents originating from surfaces due to mesa conditions, trapping, or surface recombination are referred to as surface leakage currents. The bulk currents I_{db} are subject to the avalanche gain mechanism, while the surface currents I_{ds} are not. The shot-noise expression becomes

$$\langle i_{spd}^2\rangle = 2q(I_p + I_{db})\langle m^2\rangle B + 2qI_{ds}B. \tag{23}$$

Random current fluctuations due to background radiation, stray light in the system striking the device active area and being collected by the photodetector, denoted here as I_b, can also be included in the shot-noise term (Sze, 1981):

$$\langle i_{spdb}^2\rangle = 2q(I_p + I_{db} + I_b)\langle m^2\rangle B + 2qI_{ds}B. \tag{24}$$

Thermal noise arises from all resistances present in the system. The series resistance is usually small compared with the other resistance values in the system and will be neglected. The equivalent resistance for the parallel combination is given by (Sze, 1981)

$$\frac{1}{R_{eq}} = \frac{1}{R_j} + \frac{1}{R_L} + \frac{1}{R_a}, \tag{25}$$

where R_j is the junction resistance. The mean square thermal, or Johnson, noise is

$$\langle i_T^2\rangle = 4kTB/R_{eq} \tag{26}$$

(Sze, 1981), where k is Boltzmann's constant, T is the temperature in kelvins, and B is the bandwidth of the measurement.

In a receiver, the signal-to-noise ratio determines its sensitivity and low-power detection capabilities. The signal-to-noise ratio is defined as the signal due to the photocurrent divided by the mean squared noise components from the photodetector and the amplifier. It is important to have high quantum efficiency in the detector and low noise in the amplifier in order to maximize receiver sensitivity. The signal-to-noise ratio is then

$$\frac{S}{N} = \frac{\langle i_p^2\rangle\langle m^2\rangle}{2q(I_p + I_{db} + I_b)\langle m^2\rangle B + 2qI_{ds}B + 4kTB/R_{eq} + \langle i_{amp}^2\rangle}, \tag{27}$$

where i_{amp} is the amplifier current. The minimum detectable power associated with a particular S/N ratio can be found [for $m = 1$, $I_{ds} + I_{db} = I_d$, i_{amp} negligible, and $\langle i_p^2\rangle = \frac{1}{2}(nqP_0/h\nu)^2$] (Sze, 1981) as

$$\frac{S}{N} \approx \frac{\frac{1}{2}(q\eta P_0/h\nu)^2}{2q(I_p + I_d + I_b)B + 4kTB/R_{eq}} \tag{28}$$

and

$$P_{omin} = \frac{2h\nu B}{\eta}\left(\frac{S}{N}\right)\left[1 + \left(1 + \frac{I_{eq}}{qB(S/N)}\right)^{1/2}\right], \tag{29}$$

where

$$I_{eq} = I_d + I_b + 2kT/qR_{eq}. \tag{30}$$

If the receiver is in the shot-noise–limited regime, the dark-current shot noise sets the minimum detectable power limit. For $m = 1$, this is

$$P_{omin} = \frac{2qI_dB}{R} = \frac{2hcI_dB}{\eta\lambda} \tag{31}$$

in units of watts.

If the background radiation or thermal noise is the dominant noise source, a figure of merit for the receiver is the noise-equivalent power (NEP). The NEP is the rms optical power ($P_0/\sqrt{2}$). To calculate NEP, Eq. (29) is used where the bandwidth is assumed to be 1 Hz, $S/N = 1$, and $I_{eq}/qB(S/N)$ is assumed to be $\gg 1$:

$$\mathrm{NEP} = (\sqrt{2}h\nu/\eta)(I_{eq}/q)^{1/2} \tag{32}$$

in units of W/cm^2 Hz$^{1/2}$.

3. TYPES OF PHOTODETECTORS

The choice of a photodetector for a specific optical receiver application depends on the performance and cost tradeoffs for the design. The wavelength range of interest will determine the material used. More mature technologies with simple fabrication processes in Si, and Ge, and GaAs usually mean lower-cost devices with adequate performance. High-performance devices for the telecommunications wavelengths of 1300 and 1550 nm typically have more involved fabrication processes and more difficult-to-control dark currents. Speed of response, noise, quantum efficiency, and gain are parameters to consider in the choice of a photodetector.

3.1 Photoconductors

A photoconductor is fabricated by placing Ohmic contacts onto a semiconductor material. When the material is illuminated, electron–hole pairs are produced and the electrical conductivity of the material increases. The greater the photon flux, the greater the increase in conductivity (Saleh and Teich, 1991; Sze, 1981). When voltage is applied to the contacts, the electrons and holes move in opposite directions resulting in a photocurrent. The anode and cathode may be placed on opposite ends of the semiconductor or both on top in an interdigitated pattern as illustrated in Fig. 8.

Photoconductive detectors have photoconductive gain defined as the flow of electrons per second from the device divided by the rate of generation of electron–hole pairs within the device (Wilson and Hawkes, 1989). The device gain is expressed as

$$G = \tau_m/\tau_e, \tag{33}$$

where τ_m is the excess carrier recombination lifetime and τ_e is the electron transit time ($\tau_e = 1/v_e$) (Saleh and Teich, 1991). Gain may be larger than unity.

Photoconductive detectors have the advantages of photoconductive gain and ease of fabrication over *p-n* or *p-i-n* devices, but lack their speed capabilities. The transit time in a photoconductor is set by the peak velocity of the fastest carrier, as contrasted with the case of *p-i-n*'s where it is limited by the slowest carrier. In III-V materials where the velocity of electrons is greater than the velocity of holes, and low-field electron velocity is greater than the high-field peak saturation velocity, high gain–bandwidth products can be obtained, but the *p-i-n* (with no gain) has much higher bandwidth (Forrest, 1988).

3.2 *p-n* Junction Photodiodes

The *p-n* junction photodiode is formed by doping layers of semiconductor material *n* type and *p* type. When a photon is absorbed, it creates an electron–hole pair as in the photoconductive device. Here, the electric field is formed only in the depletion region of the

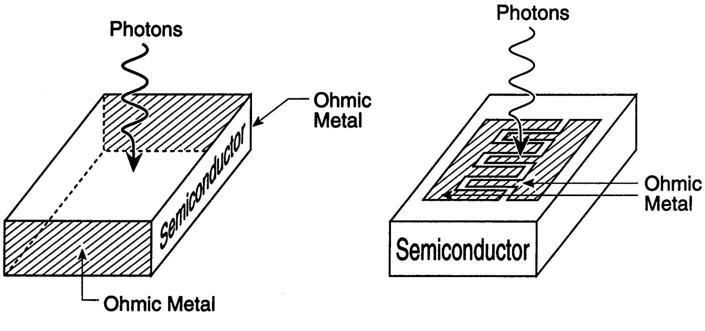

FIG. 8. Two designs for photoconductors showing Ohmic contacts on the ends of a semiconductor slab, and interdigitated contacts on the surface of a semiconductor.

device, and only carriers generated there, or near there (within a diffusion length), have the ability to be moved in a specific direction. The device is operated under reverse bias and the electrons are moved toward the *n* region while the holes drift to the *p* region under the influence of the electric field.

Response time of a *p-n* junction device is faster than a photoconductive detector as a result of the strong electric field in the depletion region giving carriers a larger velocity. The wider depletion region of the *p-i-n* photodetector makes it a better choice for high-speed devices.

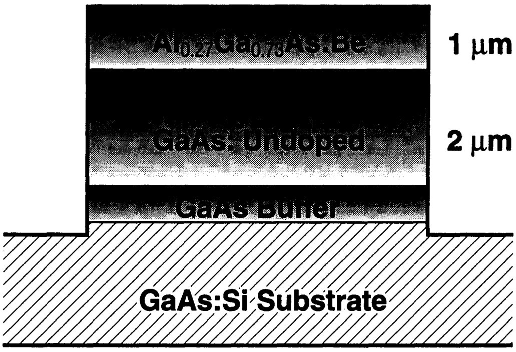

FIG. 9. The epitaxial layer structure for an AlGaAs/GaAs heterojunction *p-i-n* detector.

3.3 *p-i-n* Photodetectors

A *p-i-n* detector is typically formed by growing an intrinsic (lightly doped) layer on top of a highly doped n^+ substrate and growing a p^+ layer on top of the intrinsic layer or diffusing the *p* dopant into the top layer. The detector structure will also work in reverse, that is, growing an *i* layer and an n^+ layer on a p^+ substrate. The *p-i-n* has advantages over the *p-n* junction device in that it has a wider depletion region (thickness of *i* layer) and, therefore, a lower capacitance and lower *RC* time constant. The transit time, however, will increase (Saleh and Teich, 1991). High bandwidths, in excess of 100 GHz (Wey, 1993), are achievable with *p-i-n* photodetectors.

When the *p-i-n* device is formed out of a single semiconductor material, as in silicon *p-i-n*'s, the device material is homogeneous and it is referred to as a homojunction device. A photodetector composed of two different semiconductor materials is referred to as a heterojunction device. An example of this structure is the $Al_{0.27}Ga_{0.73}As/GaAs$ device for use from about 680 to 880 nm. This structure, which can have high bandwidths (>22 GHz) and high external quantum efficiency (73%) (Johnsen *et al.*, 1991), is shown in Fig. 9 and can be grown using molecular beam epitaxy (MBE). The higher–band-gap material sits on top where light incident at

$$\lambda_0 > hc/E_g \qquad (2)$$

is not absorbed, but passes through to the intrinsic layer, which, under bias, is depleted of carriers as they are continuously swept out. This top transparent layer, the heterojunction window layer of higher–band-gap material, allows for the photons to pass through and be collected in the *i* layer below. A double heterojunction is formed when there are two junctions of differing semiconductor materials as in the $InP/In_{0.53}Ga_{0.47}As/InP$ detector material system (Wang *et al.*, 1987).

p-n or *p-i-n* junction photodiodes can be formed using a planar diffusion process. These diffused devices can have high response and very low dark current or high dynamic resistance for low-noise receiver performance.

The fabrication process for InP/InGaAs/InP diffused heterojunction devices will be used to illustrate a typical planar device process (Beck, 1993). The epitaxial structure used is an InP buffer layer on the InP:S (InP sulfur doped) substrate, followed by the InGaAs active layer, which has a low (1×10^{15} cm^{-3}) background doping, and then the window layer of InP, which will be diffused with the *p*-type dopant zinc. To form the diffusion barrier, silicon nitride is deposited on the wafer, patterned using a photoresist mask, and removed from the areas where Zn is diffused into the semiconductor. The diffusion depth is targeted to be through the InP window layer and into the InGaAs layer beneath. An antireflection coating is used to provide low reflections from the top surface and also to act as passivation for the device. The *p* contact is formed by etching away a small region of the antireflection coating and depositing Ti/Pt/Au metallization by a liftoff process. A bonding pad can then be formed by gold plating. The wafer is lapped and polished to provide the desired chip thickness and polish on the back side. The *n*-side metal contact (AuGe alloy) is deposited on the back side of the wafer on the InP:S substrate. A cross sec-

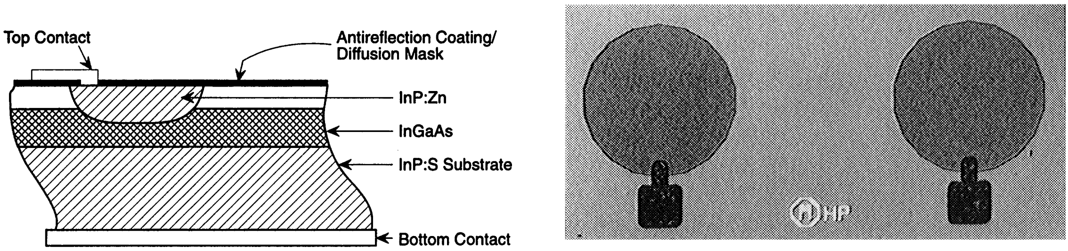

FIG. 10. (Left) Cross section of a diffused InGaAs detector showing the double-heterojunction material structure, Si_xN_y diffusion mask, AR coating, *p*-metal bondpad, and back-side *n* metals. (Right) Photograph of a pair of diffused detectors fabricated in this way (Sloan, 1989).

tion of the device and a photograph of a pair of diffused detectors are shown in Fig. 10.

Photodetector mesa structures provide a low-capacitance design for high-frequency performance. Mesa devices can be front-side or back-side illuminated. They can have grown-in doped layers or diffused layers (Ohnakea *et al.*, 1987) with mesas etched outside the diffused region. Mesa devices often have higher dark currents than diffused-junction devices because of the exposed junction surface.

A typical process for a *p-i-n* structure (Sloan, 1989) with a grown-in *p* layer involves deposition of the metal contact on the *p* layer by a lift-off process, deposition of an antireflection coating, etching the mesa with wet-chemical etches or by reactive-ion etching, deposition of the *n*-metal contact, and passivation of the mesa surface. A cross section and scanning electron micrograph of the device are shown in Fig. 11.

3.4 Metal–Semiconductor–Metal (Schottky Barrier) Photodetectors A thin metal layer deposited on a semiconductor can form a Schottky barrier (see SCHOTTKY BARRIERS). This metal–semiconductor contact can act as a *p*- or *n*-type layer forming a *p-n* junction photodetector. If made thin enough, the metal is transparent to the incident light, and a depletion region is formed at the semiconductor surface. An example of a Schottky or metal–semiconductor device fabricated in the GaAs material system is given as follows (Wang and Bloom, 1983). An n^+-GaAs layer and an *n*-GaAs layer are grown on a semi-insulating GaAs substrate. The area under the Schottky and Ohmic contacts is left active and conductive, while the surrounding area is isolated by proton bombardment, which destroys the ordered atomic lattice. This also minimizes the capacitance, which is necessary for high-speed performance. The Ohmic contact is placed on the n^+ layer after etch-

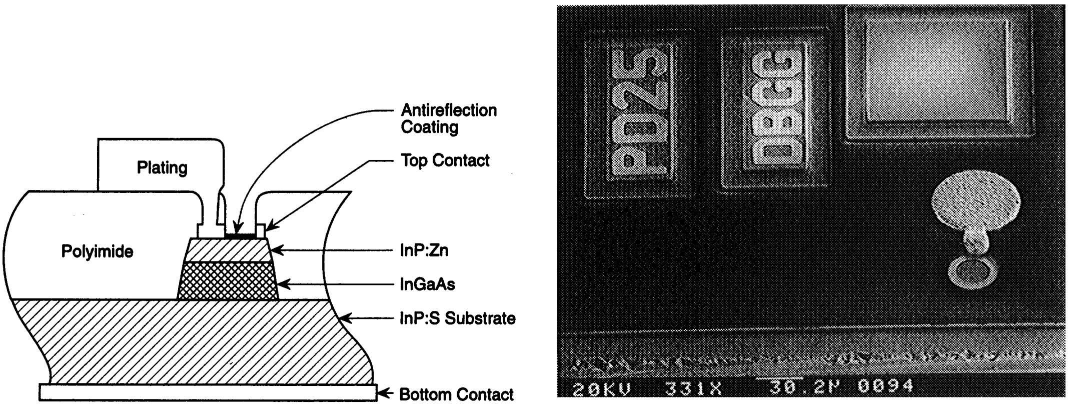

FIG. 11. (Left) Cross section of a mesa InGaAs detector showing the fabrication process steps—*p* metal, AR coating, wet-etched mesa, polyimide passivation, bond pad, and back-side metal. (Right) Photograph of a mesa *p-i-n* using this process (Sloan, 1989).

ing away the n layer of GaAs. A AuGe-type Ohmic contact to the n^+ layer is deposited using a liftoff process. The transparent contact to the n-GaAs layer is made by depositing a thin layer (100 Å) of platinum and patterning using photoresist and etching with an Ar ion mill. The light-absorbing area is covered with a dielectric passivation. For a small-area device, 5×5 μm^2, a pulse response of 5.4 ps full-width-half-maximum (FWHM) or >100 GHz 3-dB frequency response has been demonstrated (Wang and Bloom, 1983).

Metal–semiconductor–metal (MSM) photodetectors commonly use interdigitated metal fingers to form the two electrical contacts for the device. These devices are attractive because of the low capacitance of the small active areas and the short transit time of the carriers between the electrodes. MSMs typically have lower responsivity or quantum efficiency than *p-i-n* devices because of the large amount of the incident light area covered by metal. Response can also be reduced by using thin *i* layers to try to minimize the transit-time effects of carriers needed to achieve high bandwidths. This can be a problem for *p-i-n* devices as well. The broad bandwidth possibilities and the ease and compatibility of processing with existing field-effect transistors (FETs) make MSMs a good choice for integrated optical devices. MSMs are being fabricated in GaAs and InGaAs material systems (Klingenstein *et al.*, 1992; Zirngibl *et al.*, 1989; Van Zeghbroeck *et al.*, 1988).

3.5 Avalanche Photodetectors

The avalanche photodetector (APD) operates under a strong reverse bias so that the electric field strength across the region between the p-type and n-type layers is large. When an incoming photon is absorbed, a cascading of carriers is created by a process known as impact ionization. This avalanche process of carriers produces carrier multiplication, or gain, in the device.

The process of impact ionization or multiplication that takes place in an APD is illustrated in Fig. 12 (Saleh and Teich, 1991). An incoming photon absorbed at point A creates an electron–hole pair, where the electron resides in the conduction band and the hole in the valence band. Under the strong electric field provided by the high reverse bias, the electrons are accelerated toward the n side of the device and the holes toward the p side. As the electron is accelerated, it gains energy with respect to the bottom of the conduction band and undergoes collisions with the semiconductor lattice, losing some of its energy. The electron reaches an average saturation velocity balancing the increase in velocity due to the acceleration in the electric field and decrease in velocity experienced in the random lattice collisions. An electron with energy greater than the band-gap energy E_g can generate another electron–hole pair as shown at point B in Fig. 12 by kicking an electron into the conduction band and leaving a hole in the valence band (Pearsall and Pollack, 1985), giving up energy in the process. The energy required is greater than E_g because of the energy used in the three-body collision (Capasso, 1985). This process can now continue with the two electrons each having the possibility of creating more electron–hole pairs and causing the cascading effect. The holes are likewise accelerated toward the p side of the device and can undergo impact ionization spawning more electron–hole pairs (see point C in Fig. 12).

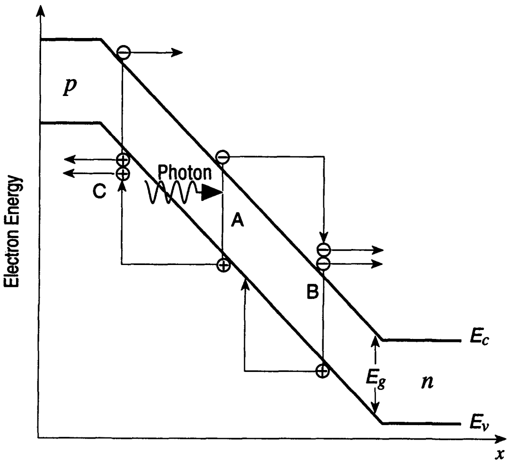

FIG. 12. The process of multiplication by impact ionization in an APD. A photon incident at point A creates an electron in the conduction band and a hole in the valence band. In a high electric field, the electron gains enough energy with respect to the bottom of the conduction band to generate another electron–hole pair at point B. Holes can likewise generate carrier pairs as shown at point C (Saleh and Teich, 1991).

Novel device designs are used to optimize APD performance. To increase quantum efficiency, the absorption layer of the APD needs to be thick enough to absorb the majority of the carriers incident, but to minimize the chance for the inclusion of an area of potential instability or area of uncontrolled avalanche (microplasmas), the multiplication region needs to be as small as possible. Various device structures have been designed in an attempt to reconcile these competing tendencies. The separate-absorption-multiplication (SAM) APD uses separate semiconductor layers to perform the necessary device functions and is able to optimize the features of each by maintaining spatially separate regions. Guard-ring structures are a way of lessening leakage currents, which can limit multiplication, by having a ring of low-doped, high-breakdown material around the active region of the device. Guard-ring structures are used in Si, Ge, and III-V devices.

For wavelengths from 0.8 to 1.0 μm, silicon is the material used for APDs. In the longer-wavelength range of 1.0–1.55 μm, germanium is used as well as III-V materials. In III-V semiconductors, because of the high leakage due to tunneling in homojunction APDs, heterojunction devices are used. Gain-bandwidth products of 100–130 GHz have been reported for various configurations.

Response times in APDs are limited by the same factors affecting *p-i-n*'s—transit times in the depletion region, *RC* time constants, diffusion times of carriers in the nearby undepleted regions—and in addition by avalanche buildup time.

In addition to the noise contributions listed in Sec. 2.2.8 for shot noise and thermal noise, another parameter of interest for APDs is multiplication noise described by the excess noise factor.

3.6 Arrays

Photodetector arrays are used in imaging and in spectroscopy. Detectors are placed close together, and light strikes many of them simultaneously. In spectroscopy, the light is dispersed into its various wavelengths before it is incident on the detector array. A current resulting from illumination of a particular array element indicates that light of that wavelength is present corresponding to a chemical signature of a substance under test. Such arrays have been fabricated in the InGaAs material system having over 500 elements sensitive to wavelengths from 600 to 1700 nm. A type of two-dimensional array called a charge-coupled device (CCD) is formed using a series of Schottky photodiodes and can be used to do visible or infrared imaging (see CHARGE-COUPLED DEVICES).

3.7 Waveguide Photodetectors

Detectors where the input light is coupled to the edge of the device are known as waveguide photodetectors. In conventional top-illuminated *p-i-n* detectors, a tradeoff exists between the speed of response and the quantum efficiency of the device. The *i* layer is made thin to minimize transit time and maximize frequency response, but thinning the *i* layer means there may not be enough absorption depth for a large fraction of the light to be absorbed and the quantum efficiency or responsivity is not optimized. In an edge-illuminated device, the current flow and the light absorption occur in perpendicular directions and both the speed and quantum efficiency can be high (Bowers and Burrus, 1986). Internal quantum efficiencies are determined by the wavelength of light and the absorption coefficient, the *i*-layer thickness, and the length of the chip. External quantum efficiency takes the coupling efficiency into account. Careful light coupling and packaging schemes are needed to get the light focused into the *i* layer of the device. Waveguide devices with 3-dB electrical bandwidths of 50 GHz and quantum efficiencies of 68% for antireflection-coated devices have been fabricated and tested (Kato *et al.*, 1992).

Fabrication processes have used both n^+ and semi-insulating substrates. The epitaxial structure is similar to a laser structure in that it contains waveguiding layers on either side of the active layer, though here the active layer absorbs rather than emits light. A typical epitaxial structure for the InP/InGaAs material system consists of a contact layer (either the n^+ substrate or an InP *n*-type grown layer), an InGaAsP *n*-type cladding layer to aid in the guiding of the mode, an InGaAs absorbing layer, an InGaAsP top guiding layer (p^+ or undoped), and *p* contacting layers (InP *p* type with an InGaAs and/or InGaAsP contacting layer on top to lower the contact resistance). A mesa or ridge is etched down through the active layer, and the mesa is passivated with polyimide. After a via hole is etched to contact the *p* layer, *p* contacts are formed on the top of the device by liftoff techniques. The wafers are thinned by lapping to facilitate cleaving. On the semi-insulating substrate, devices are etched down to the *n* layer and the *n* metal is deposited, and on the n^+-substrate devices the *n* metal is placed all over the back side of the device. Devices are cleaved and antireflection coated.

3.8 Quantum-Well Photodetectors

Quantum-well photodetectors take advantage of band-gap engineering in order to control their wavelength detection range and response. They depend on precisely controlled, often thin, alternating epitaxial layers grown by solid-source MBE, organometallic vapor-phase epitaxy, or gas-source MBE (Ritter *et al.*, 1991). Rapid switching between materials in order to achieve abrupt interfaces on an atomic scale (layers on the order of tens of ångströms) with well-characterized composition and doping profiles is necessary in order to grow the quantum wells (Manasreh and Brown, 1993). Electron confinement in these thin layers, caused by the conduction- or valence-band offset due to the difference in band-gap energies at the heterojunction interfaces, shows quantum effects where continuous energy levels become discrete (Dingle, 1987). A thin layer of small–band-gap material is placed between layers of higher–band-gap material forming a potential well where the electron or hole energy states are quantized (particle in a box). These discrete energy levels can be varied by adjusting the thickness and depth (band discontinuity) of the wells. More bound states with closer energy levels are found in wider quantum wells. Multiple quantum-well (MQW) devices are formed by growing stacked wells with barrier layers between them on the wafer. As the barrier layers become thinner than about 50 Å, the structure is referred to as a superlattice. Materials that are grown lattice mismatched to change the band structure of the device purposely are known as strained layers (Manasreh and Brown, 1993).

Detection at the long infrared wavelengths (3–14 μm) is achieved by using such quantum-well devices. Applications at these long wavelengths are for imaging with linear arrays, medical imaging, global and space surveillance, and astronomy. Information on other types of infrared detectors can be found in the article SENSORS, INFRARED.

3.9 High-Speed Detectors

As optical data rates increase in both the data communications and telecommunications markets, high-speed detectors will be necessary for the transportation of that information. 10-Gbit and 20-Gbit systems are now being designed in research environments. 2.488-Gbit SONET (Synchronous Optical Network) systems are already being implemented. Detectors designed for high speed that trade off thin *i* layers for short transit times and thick *i* layers for low capacitance will be needed. Designs that involve low capacitance, low series resistance, and short transit times and still maintain reasonable responsivities will be needed (Bowers and Burrus, 1987). *p-i-n* detectors with bandwidths of 100 GHz have been made in the GaAs and InGaAs material systems (Wang and Bloom, 1983; Wey, 1993). APDs with gain-bandwidth products of 120 GHz have also been demonstrated (Watanabe *et al.*, 1993).

3.9.1 Device Types and Trends As systems push to higher data rates, detector areas are shrunk in an effort to reduce junction capacitance, eventually making the devices difficult to fabricate and illuminate. Parasitic capacitances and inductances become very important to the speed of response of the devices, influencing the flatness and rolloff of the frequency response. Mesa *p-i-n* devices as small as 5 × 5 μm^2 have been fabricated for research applications (Wey, 1993), but would be difficult to reproduce repeatably in a device manufacturing environment. The small mesa size formed by wet etching means that the undercutting would need to be carefully monitored and traded off against the surface etches necessary to achieve low dark currents. Back-side–illuminated devices that can be attached by a solder bump technique known as flip-chip (Katsura *et al.*, 1990; Makiuchi *et al.*, 1993; see also ELECTRONIC COMPONENTS AND ASSEMBLIES: RELIABILITY AND TESTING) are a means of lowering device capacitance by removing the need for a front-side bondpad and lowering parasitics since bond wires are replaced with solder interconnects. In addition to the rf response increase, dc responsivity also increases for the back-side–illuminated flip-chip device over the front-side–illuminated device, where the benefit comes from using a double pass of the light through the *i* layer of the device, thus absorbing a larger fraction of the incident photons and converting them into usable photocurrent.

Optoelectronic integrated circuits (OEICs) are another way of forming interconnections between the devices by fabricating them on a single substrate, often using multiple material regrowths to achieve the layer structures necessary for the varied devices. OEICs using photodiodes as one of their elements have also been fabricated. Performance of the OEICs has not yet met the performance of the individual optical and electrical components, but improvements have been made and performance is beginning to approach that of the hybrid circuits, making OEICs useful for high-speed applications and lower-cost alternatives.

4. APPLICATIONS IN OPTICAL RECEIVERS

An optical receiver converts incoming optical information into an electrical signal that can then be analyzed by conventional electronic means. In long-haul communications systems, the optical signal incident on the photodetector may be weak and distorted, and so many receivers involve the photodetector, or optical-to-electrical converter (O/E), followed by an amplifier to boost the electrical signal to levels where it can be processed by the subsequent electronics. Some applications for receivers, however, involve looking at the source characteristics in the time domain on an oscilloscope where cleanly reproduced pulse response is a major concern. For these applications and those where the optical and electrical signals are at an adequate level, unamplified receivers are the choice (King *et al.*, 1993). Figure 13 shows a photograph of two such *p-i-n*–only photoreceivers attached to the test set of a high-speed oscilloscope. Desired characteristics of optical receivers are high sensitivity, low noise, wide dynamic range [ratio of maximum to minimum input signals (Senior, 1992)], and adequate bandwidth. These are necessary for long-haul systems where long repeater spacing and high capacity are important and, along with low cost, for local-area networks, local loops, cable TV, and data links (Williams, 1991).

In a digital system, a stream of data, or bits, is sent electrically as a series of ones and zeros. An electric pulse is applied to a laser or light-emitting diode (LED), and a series of optical pulses is transmitted down an optical fiber where it is attenuated and distorted somewhat before it impinges upon the pho-

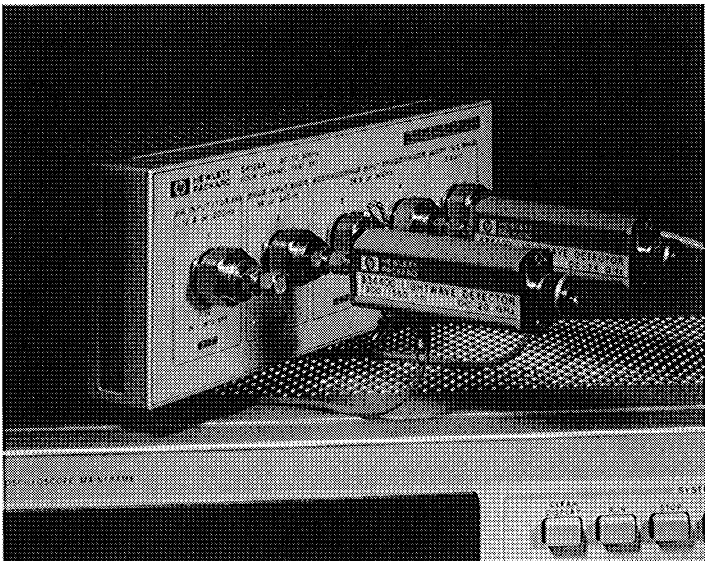

FIG. 13. Photograph of Hewlett-Packard 83440 series lightwave receivers.

todetector. An electrical wave form results that contains the fiber distortions plus any noise due to the photodetector. This is amplified and picks up any noise associated with the amplifier, and the pulses pass to a decision circuit where the ones and zeros must be discerned (Keiser, 1983). If the voltage is greater than some threshold voltage, then the conclusion is that a light pulse was present and a one is sensed. If the voltage falls below that threshold, than no light was present and a zero is detected. Detection can be done by asynchronous or synchronous methods. The bit error rate (BER) of the digital system is defined as the number of errors occurring in a given time interval N_e by the number of pulses transmitted during that time interval N_t (Keiser, 1983)

$$\mathrm{BER} = N_e/N_t. \tag{34}$$

Typical BERs for optical communications systems are 10^{-6}–10^{-10} (Keiser, 1983). Most systems now require BERs of 10^{-9} (Williams, 1991).

Noise, variable or undesirable components of the electrical signal, limits receiver performance. Noise can arise from several sources. Random fluctuation of currents in resistors is called Johnson or thermal noise, and in diodes it is called shot noise. Shot noise comes from photodiode dark current from both the semiconductor bulk and surface. APDs have additional noise due to the random nature of the gain process. Quantum noise is associated with the random process of converting light into electrical current. Interference of adjacent signals is noise referred to as cross talk or intersymbol interference. In a *p-n* or *p-i-n* photoreceiver, the thermal noise tends to dominate, and in an APD receiver, the random noise of the multiplication process adds additional shot noise (due to fluctuations in dark current and quantum noise) and this is the limiting factor (Senior, 1992).

Optical-receiver amplifier designs are usually of two types—high-impedance and transimpedance designs (Hullett and Muoi, 1976; Kasper, 1988). The high–input-impedance amplifier provides good sensitivity, but does not have wide bandwidth characteristics because of the effective integration of the detector output over a large time constant (Senior, 1992; Personick, 1973). The transimpedance amplifier uses a low-noise, high-impedance design where the high impedance is reduced by negative feedback provided by the feedback resistor (Senior, 1992). It has higher bandwidth performance than the high-impedance case. Transimpedance amplifiers are less sensitive than high-impedance designs (Williams, 1991). Transimpedance amplifier designs are being used in hybrid schemes and also in monolithic versions where an APD is integrated with an amplifier or a *p-i-n* diode is integrated with a FET to form a *p-i-n*-FET (Akahori *et al.*, 1991; Uchida *et al.*, 1991). *p-i-n*'s are generally used when speed is the major concern, and APDs can provide increased sensitivity as a result of their gain (Forrest, 1985).

Monolithic integration of photodetectors and amplifiers has been demonstrated using *p-i-n*'s and HBTs (heterojunction bipolar transistors) (Chandrasekhar *et al.*, 1990), MSMs and HEMTs (high–electron-mobility transistors) (Sasaki *et al.*, 1989; Yano *et al.*, 1990). These OEICs have been grown using InP/InGaAs heterostructures or InAIAs/InGaAs and fabricated on a single substrate. Photoreceivers that will respond to data rates of 1–4 Gbit/s have been made (Chandrasekhar *et al.*, 1990; Yano *et al.*, 1990).

As data rates increase, photodetector speeds will increase and OEICs will become necessary to overcome the parasitics associated with hybrid designs and interconnections. Photodetectors will continue to evolve as even lower-capacitance, lower–transit-time, high-response designs are in demand.

Works Cited

Akahori, Y., Ikeda, M., Uchida, N., Kohzen, A., Temmyo, J., Yoshida, J., Kokubun, T., Suto, K. (1991), *IEEE Photon. Technol. Lett.* **3**, 378–380.

Beck, P. A. (1993), *Hewlett-Packard J.* **44**, 49–51.

Bowers, J. E., Burrus, C. A. (1986), *Electron. Lett.* **22**, 905–906.

Bowers, J. E., Burrus, C. A. (1987), *J. Lightwave Technol.* **LT-5**, 1339–1350.

Bowers, J. E., Burrus, C. A., Mitschke, F. (1986), *Electron. Lett.* **22**, 633–635.

Braun, D. M. (1988), *Appl. Opt.* **27**, 2006–2011.

Capasso, F. (1985), "Physics of Avalanche Photodiodes," in: W. T. Tsang (Ed.), *Semiconductors and Semimetals*, Vol. 22, *Lightwave Communications Technology*, Pt. D, *Photodetectors*, Orlando, FL: Academic, pp. 2–172.

Carey, K. W., Wang, S. Y., Chang, J. S. C., Nauka, K. (1989), *J. Cryst. Growth* **98**, 90–97.

Chandrasekhar, S., Dentai, A. G., Joyner, C. H., Johnson, B. C., Gnauck, A. H., Qua, G. J. (1990), *Electron. Lett.* **26**, 1880–1882.

Dingle, R. (1987), "Applications of Multiquantum Wells, Selective Doping and Superlattices," in: R. K. Willardson, A. C. Beer (Eds.), *Semiconductor and Semimetals*, Vol. 24, San Diego: Academic Press.

Forrest, S. R., Leheny, R. F., Nahory, R. E., Pollack, M. A. (1980), *Appl. Phys. Lett.* **37**, 322–325.

Forrest, S. R. (1985), "Sensitivity of Avalanche Photodetector Receivers for High-Bit-Rate Long-Wavelength Optical Communication Systems," in: W. T. Tsang (Ed.), *Semiconductors and Semimetals*, Vol. 22, *Lightwave Communications Technology*, Pt. D, *Photodetectors*, Orlando, FL: Academic, pp. 329–387.

Forrest, S. R. (1988), "Optical Detectors for Lightwave Communication," in: S. E. Miller, I. P. Kaminow (Eds.), *Optical Fiber Telecommunications II*, San Diego: Academic, pp. 569–599.

Hullett, J. L., Muoi, T. V. (1976), *IEEE Trans. Commun.* **COM-24**, 1180–1185.

Johnsen, C., Sloan, S., Braun, D., Russell, J. L., Zurakowski, M., Lightner, M., Kellert, F., Patterson, G., Koo, R., Derrickson, D., Bowers, J. E. (1991), *IEEE Trans. Electron. Devices* **ED-38**, 1968–1970.

Kasper, B. L. (1988), "Receiver Design," in: S. E. Miller, I. P. Kaminow (Eds.), *Optical Fiber Telecommunications II*, New York: Academic, pp. 689–722.

Kato, K., Hata, S., Kawano, K., Yoshida, J., Kozen, A. (1992), *IEEE J. Quantum Electron.* **QE-28**, 2728–2735.

Katsura, K., Hayashi, T., Ohira, F., Hata, S., Iwashita, K. (1990), *J. Lightwave Technol.* **8**, 1323–1326.

Keiser, G. (1983), *Optical Fiber Communications*, New York: McGraw-Hill.

King, R., Braun, D. M., Hinch, S. W., Shubert, K. (1993), *Hewlett-Packard J.* **44**, 83–86.

Klingenstein, M., Kuhl, J., Notzel, R., Ploog, K., Rosenzweig, J., Moglestue, C., Hulsmann, A., Schneider, J. (1992), *Appl. Phys. Lett.* **60**, 627–629.

Lee, T. P., Li, T. (1979), "Photodetectors," in: S. E. Miller, A. G. Chynoweth (Eds.), *Optical Fiber Telecommunications*, New York: Academic.

Lucovsky, G., Schwarz, R. F., Emmons, R. B. (1964), *J. Appl. Phys.* **35**, 622.

Makiuchi, M., Norimatsu, M., Sakurai, T., Kondo, K., Yano, M. (1993), *IEEE Photon. Technol. Lett.* **5**, 518–520.

Manasreh, M. O., Brown, G. J. (1993), "Introduction to Long-Wavelength Infrared Quantum Detectors," in: M. O. Manasreh (Ed.), *Semiconductor Quantum Wells and Superlattices for Long-Wavelength Infrared Detectors*, Boston: Artech House.

Melchoir, H. (1972), "Demodulation and Photodetection Techniques," in: F. T. Arecchi, E. O. Schulz-Dubois (Eds.), *Laser Handbook*, Vol. 1, Amsterdam: North-Holland, pp. 703–835.

Muller, J. (1981), "Photodiodes for Optical Communication," in: L. Maron, C. Marton (Eds.), *Advances in Electronics and Electron Physics*, Vol. 55, New York: Academic.

Ohnaka, K., Kubo, M., Shibata, J. (1987), *IEEE Trans. Electron. Devices* **ED-34**, 199–204.

Pearsall, T. P., Pollack, M. A. (1985), "Compound Semiconductor Photodiodes," in: W. T. Tsang (Ed.), *Semiconductors and Semimetals*, Vol. 22, *Lightwave Communications Technology*, Pt. D, *Photodetectors*, Orlando, FL: Academic, pp. 174–245.

Personick, S. D. (1973), *Bell Syst. Tech. J.* **52 (6)**, 843–887.

Ritter, D., Hamm, R. A., Panish, M. B., Vandenber, J. M., Gershoni, D., Ganapala, S. D., Levine, B. F. (1991), *Appl. Phys. Lett.* **59**, 552–554.

Saleh, B. E., Teich, M. C. (1991), *Fundamentals of Photonics*, New York: Wiley.

Sasaki, G., Koike, K., Kuwata, N., Ono, K. (1989), *J. Lightwave Technol.* **7**, 1510–1514.

Senior, J. M. (1992), *Optical Fiber Communications: Principles and Practice*, 2nd ed., Englewood Cliffs, NJ: Prentice-Hall.

Shockley, W. (1949), *Bell Syst. Tech. J.* **28**, 435.

Sloan, S. R. (1989), *Hewlett-Packard J.* **40**, 69–75.

Sze, S. M. (1981), *Physics of Semiconductor Devices*, New York: Wiley.

Uchida, N., Akahori, Y., Ikeda, M., Kohzen, A., Yoshida, J., Kokubun, T., Suto, K. (1991), *IEEE Photon. Technol. Lett.* **3**, 540–542.

Van Zeghbroeck, B. J., Patrick, W., Halbout, J. M., Vettiger, P. (1988), *Electron. Device Lett.* **9**, 527–529.

Wang, S. Y., Bloom, D. M. (1983), *Electron. Lett.* **19**, 554–555.

Wang, S. Y., Carey, K. W., Kolner, B. H. (1987), *IEEE Trans. Electron. Devices* **ED-34**, 938–940.

Watanabe, I., Sugou, S., Ishikawa, H., Anan, T., Makita, K., Tusji, M., Taguchi, K. (1993), *IEEE Photon. Technol. Lett.* **5**, 675–677.

Wey, Y. G. (1993), *ECE Technical Report 93-04*, Department of Electrical and Computer Engineering, University of California at Santa Barbara.

Williams, G. F. (1991), "Lightwave Receivers," in: Tingye Li (Ed.), *Topics in Lightwave Transmission Systems*, New York: Academic, pp. 79–149.

Wilson, J., Hawkes, J. F. B. (1989), *Optoelectronics*, Englewood Cliffs, NJ: Prentice-Hall.

Yano, H., Aga, K., Kamei, H., Sasaki, G., Hayashi, H. (1990), *J. Lightwave Technol.* **8**, 1328–1333.

Zirngibl, M., Bischoff, J. C., Theron, D., Ilegems, M. (1989), *Electron. Device Lett.*, **10**, 336–338.

Further Reading

Lee, T. P., Li, T. (1979), *Photodetectors*, in: S. E. Miller, A. G. Chynoweth (Eds.), *Optical Fiber Telecommunications*, New York: Academic.

Melchoir, H. (1972), "Demodulation and Photodetection Techniques," in: F. T. Arecchi, E. O. Schulz-Dubois (Eds.), *Laser Handbook*, Vol. 1, Amsterdam: North-Holland, pp. 703–835.

Muller, J. (1981), "Photodiodes for Optical Communication," in: L. Maron, C. Marton (Eds.), *Advances in Electronics and Electron Physics*, Vol. 55, New York: Academic.

Pearsall, T. P., Pollack, M. A. (1985), "Compound Semiconductor Photodiodes," in: W. T. Tsang (Ed.), *Semiconductors and Semimetals*, Vol. 22, *Lightwave Communications Technology*, Pt. D, *Photodetectors*, Orlando, FL: Academic, pp. 174–245.

Saleh, B. E., Teich, M. C. (1991), *Fundamentals of Photonics*, New York: Wiley.

9
Photoemission and Photoelectron Spectroscopy

Franz J. Himpsel and Ingolf Lindau

9.1 **Introduction 285**
9.1.1 History 285
9.1.2 Phenomena 286
9.1.3 Theory 288
9.2 **Instrumentation 293**
9.2.1 Light Sources 293
9.2.2 Detectors 295
9.3 **Photoelectric Yield and Absorption Spectroscopy 298**
9.3.1 Mechanism 298
9.3.2 Unoccupied Atomic and Molecular Orbitals 298
9.3.3 Solids 299
9.3.4 Magnetic Dichroism 300
9.3.5 Microspectroscopy 300
9.4 **Core-Level Spectroscopy 301**
9.4.1 Origin of Core-Level Shifts 301
9.4.2 Atoms, Shake-Up, and Shake-Off Satellites 302
9.4.3 Adsorbates and Surfaces 303
9.4.4 Solids 305
9.5 **Valence Electron Spectroscopy 305**
9.5.1 Occupied Atomic and Molecular Orbitals 307
9.5.2 Surfaces and Quantum Wells 308
9.5.3 Band Mapping 309
9.5.4 Time-Resolved, Pump-Probe Measurements 311
9.5.5 Inverse Photoemission 311
9.6 **Structural Methods 312**
9.6.1 EXAFS and Photoelectron Diffraction 312
9.6.2 Photoelectron Diffraction 313
9.7 **Summary and Outlook 314**
9.8 **Databases 315**
Glossary 315
References 317

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

9.1
Introduction

9.1.1
History

Photoemission, that is, the emission of electrons by photons, was discovered as an unwanted side effect in the study of electromagnetic waves by Hertz (1887). At that time, neither the electron nor the photon was known as *particles*. Einstein (1905) explained the photoelectric effect as a quantum phenomenon and derived the correlation

$$E_{\max} = h\nu - \Phi \tag{9.1}$$

between the maximum kinetic energy $E_{\max}$ of the photoelectrons, the photon energy $h\nu$, and the work function Φ of the emitting solid (compare Figure 9.1). Many of the following decades were spent in finding clear experimental evidence for or against the Einstein formula (Millikan, 1916). The time was not ripe for photoelectron spectroscopy as we know it today, because photoemission turned out to be a rather surface-sensitive effect. The mean free path of photoelectrons in a solid limits the probing depth to a few atomic layers. The prerequisites for the preparation of reproducible, clean, and well-ordered surfaces were not in place until the sixties, when the necessary ultrahigh vacuum could be produced and measured on a routine basis. At a residual gas pressure of 10^{-6} torr (1.3×10^{-4} Pa), a surface atom is being hit by a residual gas molecule about once in every second, thus requiring pressures in the 10^{-10} torr range for controlled photoemission studies. When surface characterization techniques, such as low-energy electron diffraction (LEED) and Auger spectroscopy became generally available, photoelectron spectroscopy took a big leap forward. Photoelectron spectroscopy is now an indispensable technique for the determination of the electronic structure of solids, surfaces, and gas-phase species. Tunable and polarized synchrotron radiation became the light source of choice in the 1970s, and gave rise to the construction of a series of storage rings dedicated to the production of light from the ultraviolet to the X-ray regime. In the 1980s, the measurement of a complete set of quantum numbers became feasible for electrons in solids, which involves measuring the momentum and spin polarization of photoelectrons as a function of the photon energy and photon polarization. In the 1990s, measurements with an energy resolution

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

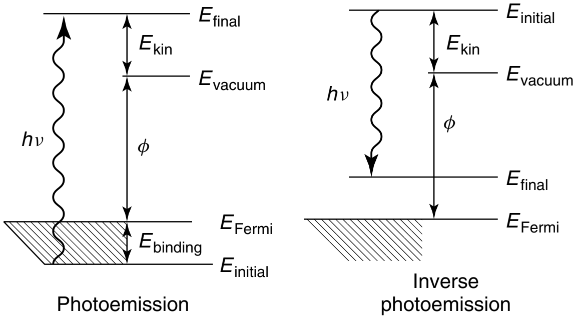

Fig. 9.1 Energy diagram for photoemission (photon in, electron out) and inverse photoemission (electron in, photon out). The energy of the lower state is determined by subtracting the photon energy $h\nu$ from the energy of the upper state. The latter corresponds to the kinetic energy of the emitted or the incoming electron.

better than the thermal energy became practical, which provided an opportunity to examine the electronic states that are relevant to magnetism, superconductivity, and other transport properties. Most recently, temporal resolution was introduced with pump–probe experiments for observing dynamics down to tens of femtoseconds, and spatial resolution down to tens of nanometers was achieved in photoelectron microscopes.

9.1.2
Phenomena

The basic energy diagram of the photoemission process is given in Figure 9.1. An electron is lifted from an occupied initial state into an unoccupied final state. Energy conservation allows it to obtain the desired energy E_{initial} of the initial state from the measured energy E_{final} of the photoelectron in the final state by simply subtracting the photon energy $h\nu$. As reference energies, one can use either the vacuum level E_{vacuum}, corresponding to a photoelectron with zero velocity, or the Fermi level E_{Fermi}, corresponding to the highest occupied state. From there one obtains energy differences, such as the kinetic energy E_{kin}, the work function Φ, and the binding energy E_{binding}. The Einstein relation for the maximum kinetic energy $E_{\max} = \text{h}\nu - \Phi$ follows directly for photoemission from the highest occupied state at E_{Fermi}.

Figure 9.1 also shows the energy diagram for the time-reversed process, that is, an incident electron dropping down into a lower unoccupied state by emission of a photon. This process is equivalent to the quantum description of bremsstrahlung, but in the low-energy regime of typically 10-eV kinetic energy, it is known today as *inverse photoemission*. It characterizes unoccupied states in a fashion complementary to the information provided by photoemission about occupied states. The counterpart to the Einstein relation is the Duane–Hunt equation

$$h\nu_{\max} = E_{\text{kin}} + \Phi \tag{9.2}$$

for the highest photon energy of the bremsstrahlung emitted by electrons with kinetic energy E_{kin}. Inverse photoemission as a surface technique got a late start compared to photoemission because its cross section is down by 4–5 orders of magnitude, requiring extremely sensitive photon detectors.

There are other electron and photon emission phenomena related to photoemission. As shown in Figure 9.2, a core hole created in the photoemission process can decay by an Auger process (a), which is most likely for shallow core levels, or by fluorescence (b), which is prevalent for deep core levels (Keski-Rahkonen and Krause, 1974). In addition, the photoelectron can lose energy by creating electron–hole pairs or plasmons during its escape. These processes actually limit the escape depth. The end result is a spectrum of emitted photoelectrons that contains several secondary features in addition to the directly emitted photoelectrons (Figure 9.3). For valence spectroscopy (often labeled as ultraviolet photoelectron spectroscopy (UPS)), the main disturbance is a tail of secondary electrons at low kinetic energies, caused by energy losses to electron hole pairs (Figure 9.3a). For core-level spectroscopy (often labeled as X-ray photoelectron spectroscopy (XPS), or electron spectroscopy for chemical analysis (ESCA)), there are additional Auger peaks and plasmon loss satellites complicating the spectra (Figure 9.3b). With tunable synchrotron radiation, it is straightforward to separate out the Auger peaks, because they do not change their kinetic energy with increasing photon energy, in contrast to elastic core lines and their plasmon loss satellites.

An important consideration in photoemission experiments is the probing depth, which is governed by the mean free path of the photoelectrons (Figure 9.4). Electrons penetrate solids farther at high kinetic energies, but their mean free path increases again at very low energies, because they do not have enough energy to excite plasmons (in metals) or electron–hole pairs (in insulators). In the 30- to 70-eV range, the mean free path reaches a minimum of only a few atomic distances, making it possible to separate surface

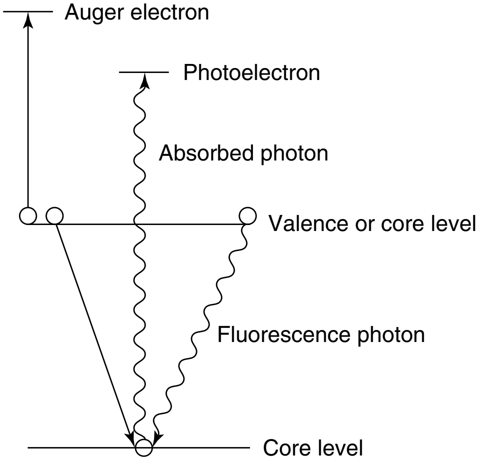

Fig. 9.2 Energy diagram for core-level photoemission, including the decay of the core hole by an Auger process (a) or by fluorescence (b).

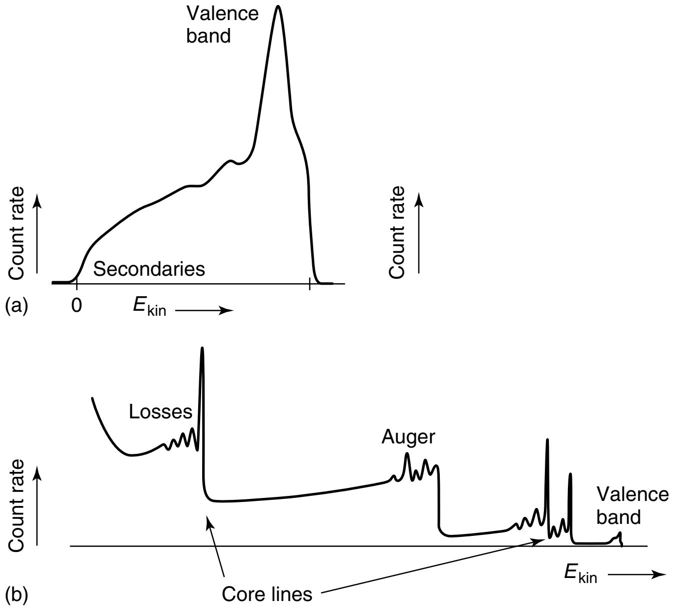

Fig. 9.3 (a) Typical valence photoelectron spectrum, showing photoelectrons directly emitted from the valence band and secondary electrons emitted after electron–hole pair production. The two energy cutoffs are determined by the vacuum level (zero kinetic energy E_{kin}) and the Fermi level (i.e., excitation of the highest occupied state). (b) Typical core-level photoelectron spectrum, showing core-level peaks, their plasmon loss satellites, and Auger electron features (Feuerbacher, Fitton and Willis, 1978).

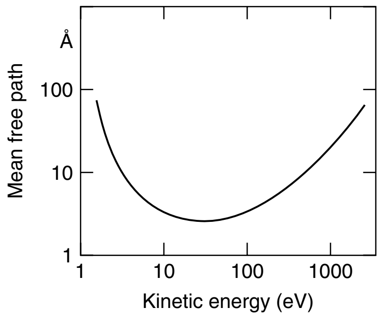

Fig. 9.4 Schematic dependence of the mean free path of photoelectron on their kinetic energy. The minimum at 30–70 eV can be used to probe surface or interface layers.

atoms from the bulk. The tunability of synchrotron radiation provides the ability of producing photoelectrons with the desired kinetic energy, which has spawned the field of surface core-level spectroscopy.

9.1.3
Theory

Photoemission is unique in providing the complete set of quantum numbers that characterize electrons in solids.

This made photoemission the preferred technique for investigating electrons in a solid and led to Philip Anderson's famous characterization of photoemission as "smoking gun" that might give away the secrets of electrons in high-temperature superconductors (Anderson and Schrieffer, 1991). For electrons in a periodic crystal lattice, these quantum numbers are energy E, momentum $\mathbf{p} = \hbar\mathbf{k}$, spin, and angular symmetry (represented by the point group in a solid). Polycrystalline solids can be characterized by the same quantum numbers, because each crystallite typically contains thousands to millions of unit cells. However, when the size of a crystallite reaches dimensions in the single digit nanometer regime (tens of atom diameters), the situation changes: instead of a continuum of energies, one obtains quantized states with discrete energies. This alters the fundamental electronic properties of a material, such as the band gap of a semiconductor, and leads to the capability of tailoring the electronic properties of materials in nanotechnology (Himpsel *et al.*, 1998).

The quantum numbers are obtained by using conservation laws during the photoemission process (see the general reviews Himpsel (1983), Kevan (1992) and Hüfner (2007)). We have already depicted the energy conservation law in Figure 9.1. A similar conservation law holds for the momentum parallel to the surface $\hbar k_{||}$, where $k_{||}$ is the corresponding wave vector. The wave vector of the photon is usually negligible compared to that of the electron for the energy range in question, because its wavelength is very much longer. The wavelength λ and wave vector $k = 2\pi/\lambda$ of the photon are given by the photon energy $h\nu$ via

$$\left(\frac{\lambda}{\text{Å}}\right) = \frac{12399}{(h\nu/\text{eV})} \tag{9.3}$$

and the parallel wave vector of the electron by its kinetic energy E_{kin} and polar escape angle ϑ via

$$\left(\frac{k_{||}}{\text{Å}^{-1}}\right) = 0.51 \times \left(\frac{E_{\text{kin}}}{\text{eV}}\right)^{1/2} \sin\vartheta \tag{9.4}$$

(1 Å = 0.1 nm). The momentum component perpendicular to the surface is not conserved during the passage of the photoelectron across the surface energy barrier, but it can be varied by tuning the photon energy. Angular symmetry can be inferred from polarization selection rules for the point group at a particular point in momentum space. Spin is conserved.

The complete set of quantum numbers is usually displayed in the form of energy-versus-momentum band dispersion, labeled with the appropriate symmetry and spin symbols (see Figure 9.17 in Section 9.5.3). Another way to represent the quantum numbers is in the form of constant energy surfaces with two momentum components as variables (see Figure 9.18 in Section 9.5.3). The Fermi surface is particularly important, because electrons near the Fermi level determine most electronic properties of a solid, such as conductivity, superconductivity, magnetism, and electronic phase transitions. The relevant electron energies are within a thermal energy range around the Fermi level, which is given by the width 3.5 $k_{\text{B}}T$ of the Fermi–Dirac distribution (k_B = Boltzmann constant, T = absolute temperature, $k_BT \approx 25$ meV at room temperature).

The theoretical description of photoemission is provided by an expression similar to Fermi's Golden Rule for a differential cross section (for details see Schattke *et al.* (2003)). It contains the absolute square of a matrix element M_{if} for the optical transition between the initial state $|\psi_i\rangle$ and the final state $\langle\psi_f|$ of the photoelectron, an average over all possible initial states, and a sum over all final states that are detected. The matrix element takes the following form:

$$M_{\mathrm{if}} \propto \langle\psi_{\mathrm{f}}|\boldsymbol{A}(\mathbf{r}) \cdot \mathbf{p}|\psi_{\mathrm{i}}\rangle \tag{9.5}$$

$\mathbf{A}(\mathbf{r})$ is the vector potential of the photon, which becomes a constant in the widely used dipole approximation.

The initial state is $|\psi_i\rangle$ is a Bloch function in a solid or at a surface. For core levels and gas-phase photoemission, the initial state simply becomes an atomic/molecular wave function. In many-body theory, the initial state is not an electron but a hole. This seems to be at odds with the fact that the initial state was occupied by an electron before the photoemission process. However, to measure the binding energy of this electron, one needs to eject it by photoemission. That leaves a hole in the solid. In the following, we keep using the notion of electrons for both initial and final states with the understanding that initial state electrons should really be considered as holes.

The final state $\langle\psi_f|$ is often called *a time-reversed LEED state* (LEED is low energy electron diffraction), because an energy- and angle-resolving detector accepts only photoelectrons with well-defined momentum. Those are similar to the incident electrons in a diffraction experiment, except for the fact that the $\langle\psi_f|$ is a complex conjugate with an inverted propagation direction. The diffracted waves in LEED correspond to incoming electron waves in the final state of photoemission, which is somewhat counterintuitive but mathematically correct. Codes for such "single-step" photoemission calculations are now widely available (see Section 2.10 in Schattke *et al.* (2003)). The earlier "three-step" model separates the Golden Rule expression into three simpler but approximate steps, that is, optical excitation, scattering on the way out, and transmission/refraction at the surface.

For a quick interpretation of photoemission spectra, one has to mainly consider the various selection rules inherent in the optical matrix element M_{if}, such as the conservation laws for energy and momentum parallel to the surface (modulo a reciprocal surface lattice vector, see Himpsel (1983) and Kevan (1992)). The momentum perpendicular to the surface is not conserved, because the periodicity of a solid is destroyed in this direction. In the three-step model, one can keep perpendicular momentum conservation for the first step, which takes place completely inside the solid. However, transitions from bulk states to evanescent final states are not covered by this approximation. Another useful selection rule holds for parity of the wave functions with respect to a mirror plane. The product of the parities of the initial, the operator $\mathbf{A} \cdot \mathbf{p}$, and the final state has to be even.

Although selection rules provide clear yes–no decisions, there are also more subtle effects due to the wave functions of the electronic states involved in the matrix element. The atomic symmetry character determines the energy dependence of the cross section (Yeh and Lindau, 1985), allowing a selection of specific orbitals by varying the photon energy. For example, the s, p-states in transition and noble metals dominate the spectra near the photoelectric threshold, while the d-states turn on at 10 eV above threshold. It takes photon energies of 30 eV above threshold to make the f-states in rare earths visible. Resonance effects at a threshold for a core-level excitation can also enhance particular orbitals. Conversely, the cross section for states with a radial node exhibits so-called Cooper minima, where they become almost invisible.

The wave functions of the electronic states in solids and at surfaces are usually approximated by the ground state wave functions obtained from a variety of schemes, for example, tight binding, local density approximation (LDA), or Hartree–Fock. Strictly speaking, one should use the excited-state wave functions and energies that represent the hole created in the photoemission process, or the extra electron added to the solid in the case of inverse photoemission. Such excited-state quasiparticle calculations have now become feasible, and provide the most accurate band dispersions to date. Particularly in the case of semiconductors, the traditional ground state methods are unable to determine the fundamental band gap from first principles, with Hartree–Fock overestimating it and LDA underestimating it, typically by a factor of 2.

Photoemission is able to provide even more information than the complete set of quantum numbers. In many-body theory, an electron is described by the Green's function G(A, B), which describes the propagation of the electron from point A to point B – obviously a very fundamental description of an electron. Photoemission essentially measures the imaginary part of the Green's function in momentum space. From the Greens function, one can extract the self-energy Σ of an electron, whose imaginary part quantifies the scattering of electrons in a solid. The real part of Σ describes an energy shift due to many-body interactions, which is not contained in single-electron band calculations. The real and imaginary parts of Σ are related to each other by an integral transform, the Kramers–Kronig relation.

Experimentally, the information about the Green's function is contained in the lineshape of the transitions observed in angle-resolved photoemission. Measuring the intrinsic lineshape requires high energy and momentum resolution together with high-quality samples. This has become possible in recent years (see Figures 9.5 and 9.6 and Hüfner (2007)). The energy width ΔE of a spectral peak determines the scattering time τ and the momentum width Δp determines the scattering length l:

$$\tau = \frac{\hbar}{\Delta E} \qquad l = \frac{\hbar}{\Delta p} \tag{9.6}$$

For example, one can alloy 20% Fe into the Ni crystal to form permalloy, a key material for magnetoelectronics. Taking spectra similar to those in Figure 9.7, one finds that the momentum width greatly increases for the spin-down electrons but not for the spin-up electrons (not shown; see Altmann *et al.* (2001)). This leads to a scattering length l of 0.6 nm for spin-down electrons, which agrees with the result of magnetoresistance measurements.

The real part of the self-energy manifests itself in a transition from "dressed" electrons close to the Fermi level to "bare" electrons farther away from it. For example, an electron can be "dressed" by a phonon cloud due to electron–phonon interaction (Plummer *et al.*, 2003). If the

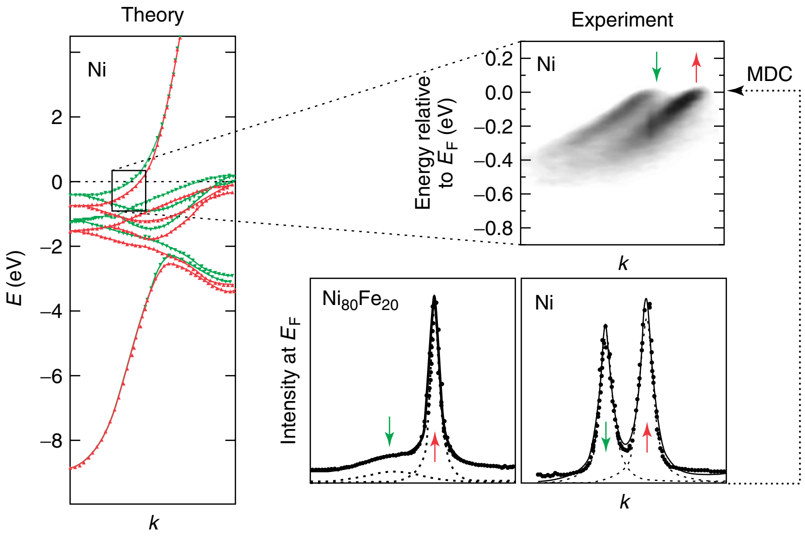

Fig. 9.5 Energy-versus-momentum band dispersions of ferromagnetic nickel, showing two sets of bands for the two spin directions (calculations on the left, data on the right). Modern electron spectrometers with energy and angle multidetection can zoom into the bands close to the Fermi level and image the band dispersion (upper right, with high photoemission intensity shown in dark shade). A momentum distribution curve (MDC) reveals the momentum broadening due to the finite scattering length (lower right). After Altmann *et al.* (2001).

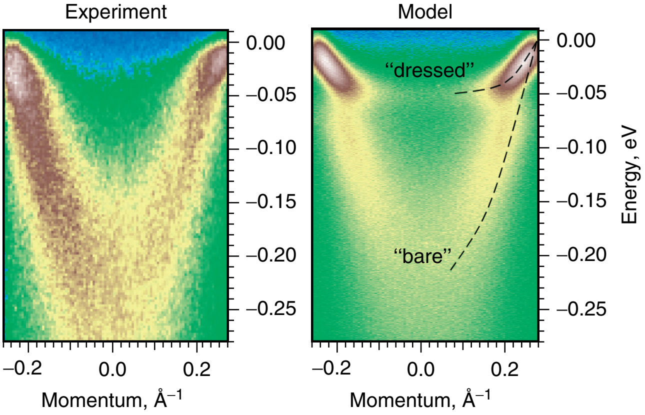

Fig. 9.6 Experimental and calculated plots of the photoemission intensity versus E and k for a high-temperature superconductor. There is a kink at about −0.04 eV, which indicates the transition from "dressed" to "bare" electrons. From Inosov *et al.* (2007).

electron moves too fast, the phonon cloud cannot follow and leaves a bare electron. This transition from dressed to bare electrons happens about one phonon energy away from the Fermi level. It manifests itself as a kink in the

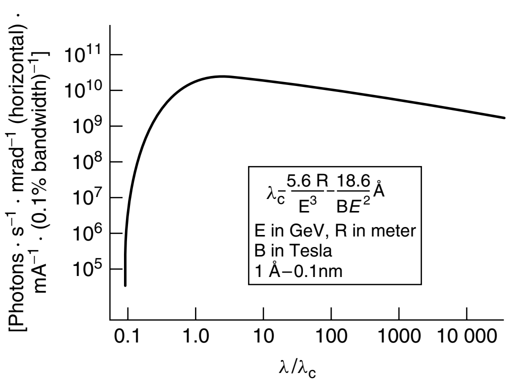

Fig. 9.7 The universal curve for the synchrotron radiation flux, plotted with λ/λ_c on the horizontal axis. λ_c is the critical wavelength, given by the electron energy E, the radius R of the electron orbit, or the magnetic field of the bending magnet.

$E(k)$ band dispersion where band curvature changes. The curvature is related to the inverse effective mass, and the mass changes from heavy for dressed electrons to light for bare electrons. An example is shown in Figure 9.6. This effect is particularly interesting for superconductors. Electrons are paired by exchanging phonons in conventional superconductors. The big question in high-temperature superconductors is the pairing mechanism, that is, the nature of the particle exchanged between pairs. The search for this particle has led to an intense study of kinks in the bandstructure of high-temperature superconductors. The example in Figure 9.6 shows how to model such a kink self-consistently in terms of the real and imaginary parts of the self-energy (Inosov *et al.*, 2007).

9.2
Instrumentation

9.2.1
Light Sources

To excite photoelectrons, a minimum photon energy of 5–10 eV is required. This is one of the reasons why conventional lasers have been of limited applicability in photoelectron spectroscopy. A detailed discussion of the possible sources for photoelectron excitation is given in Siegbahn and Karlsson (1982). In the ultraviolet radiation region, the most common source is based on a capillary glow discharge. The most convenient line is the so-called He I, which provides monochromatic radiation with a narrow line-width at 21.2 eV. This radiation originates from the 2p to 1s transition in neutral He atoms. Emission can also be produced from the He ions, the primary emission line being He II at an energy of 40.8 eV. Similarly, emission lines of Ne I at 16.8 eV and Ne II at 26.9 eV have frequently been used in photoelectron spectroscopy. Recent designs produce very efficient and highly monochromatic radiation by electron cyclotron resonance (ECR) in low-pressure He gas.

When a target is bombarded with an electron beam of sufficient energy and intensity, X-rays will be emitted via core-level fluorescence or bremsstrahlung. In principle, a vast amount of X-ray emitting materials should be available, but there are very few practical X-ray sources with high intensity, narrow energy width, and without satellite structures. The dominant radiation sources are Mg K_α at 1253.6 eV and Al K_α at 1486.7 eV. Because both magnesium and aluminum are easy to handle and their inherent line-width is as narrow as 1 eV, most commercial X-ray anodes are based on Al and/or Mg as the target material. It is also possible to produce monochromatized Al K_α and Mg K_α radiation by using, for instance, a quartz crystal monochromator. An excitation line-width of a few tenths of an electronvolt can be achieved, but with considerable loss of intensity. These technical approaches of achieving higher resolution have found their way into commercial instruments.

During the last two decades, synchrotron radiation has emerged both as a powerful and convenient excitation source in photoelectron spectroscopy (Winick and Doniach 1980; Koch, 1983; Winick *et al.*, 1989). The term *synchrotron radiation* is usually associated with the radiation emitted by relativistic electrons accelerated in a circular orbit. The basic properties of synchrotron radiation were first presented in an elegant paper by Schwinger (1949), although they had been treated much earlier by Schott (1912). With E being the electron energy and R the radius of the circular orbit, the total synchrotron radiation energy emitted per revolution by one electron is given by

$$\Delta E(R, E) = \frac{4\pi}{3} \times \frac{e^2}{R} \times \left(\frac{E}{m_0 c^2}\right)^4 \quad (9.7)$$

The spectral radiation distribution is characterized by a critical energy ε_c, given by

$$\varepsilon_c = \frac{3hc}{2R} \times \left(\frac{E}{m_0 c^2}\right)^3 \quad (9.8)$$

The physical significance of this equation is that half of the total power is radiated above the critical energy and half below. The photon intensity rapidly decreases for photon energies above the critical energy ε_c. A universal curve for the synchrotron radiation photon flux is shown in Figure 9.7. Synchrotron radiation has a number of desirable properties. It provides (i) a continuous spectral distribution from the infrared region into the X-ray region, (ii) high intensity, (iii) high degree of collimation, (iv) high degree of polarization (completely linearly polarized in the plane of the orbital and elliptically out of the plane), and (v) a pulsed time structure given by the orbital frequency of the circulating electron beam. An important development occurred with the realization that magnetic structures inserted in straight sections of the storage ring, so-called undulators and wigglers, could drastically improve the radiation characteristics of conventional bending magnets. Presently, severalz facilities are in operation and under construction that are optimized for these insertion devices, so-called third-generation sources. Figure 9.8 (Winick, 1994) shows the brightness of X-ray sources as a function of time since the X-ray tube was first introduced in 1895. Synchrotron radiation provides a few orders of magnitude higher brightness than conventional sources. Intense efforts are presently under way for the next step of source developments, that is, vacuum ultraviolet (VUV) and X-ray free-electron lasers (XFEL), also based on undulators but using linear accelerators instead of storage rings and using seed lasers. In addition to the properties listed above, such sources will provide femtosecond time resolution, full transverse coherence, and about ten orders of

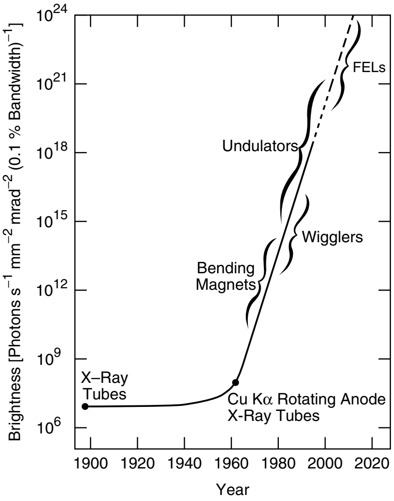

Fig. 9.8 Spectral brightness of X-ray sources as a function of time. From Winick (1994).

magnitude increase in peak brightness. Coupled with the rapid developments of conventional lasers toward even shorter wavelengths, the next decade will undoubtedly see spectacular improvements in light sources for photoemission (Chattopadhyay *et al.*, 2001).

An important advancement in laser-based sources for photoelectron spectroscopy has recently been achieved by developing a UV laser that can reach a photon energy of 6.994 eV while producing 10^{15} photons per second (Kiss *et al.*, 2005). It operates in a quasi-continuous-wave (quasi-CW) mode at a repetition frequency of 80 MHz. The high photon energy of 6.994 eV is achieved via the second harmonic of a frequency-tripled Nd : YVO_3 laser by using an optically contacted prism-coupled KBBF ($KBe_2BO_3F_2$) crystal (Togashi *et al.*, 2003). The setup is shown schematically in Figure 9.9(a). The electrons emitted from the sample have low kinetic energies, typically a few electron volts, and their energy and angle distributions are determined with a high-precision hemispherical electron analyzer. The laser source with unprecedented performance characteristics in combination with a high-resolution spectrometer has made it possible to measure spectra with a total resolution down to 360 μeV for electrons plus photons (Figure 9.9b). This points to the importance of developing novel sources in parallel with improvements in detector technology (as discussed in Section 9.2.2). The impact of these new laser sources is also related to the basic photoemission phenomena discussed in Section 9.1.2. The mean free path of photoelectrons increases with the decrease in kinetic energy, with the implication that the laser source discussed here makes it possible to study bulk phenomena. This is of particular importance for materials with large unit cells. The combination of ultrahigh energy resolution and bulk sensitivity has opened up new avenues to study complex phenomena, such as high-temperature superconductivity.

9.2.2
Detectors

The most common type of photoelectron spectrometers has been the hemispherical analyzer, which is of the electrostatic type. A key feature of efficient spectrometers is parallel detection. Energy

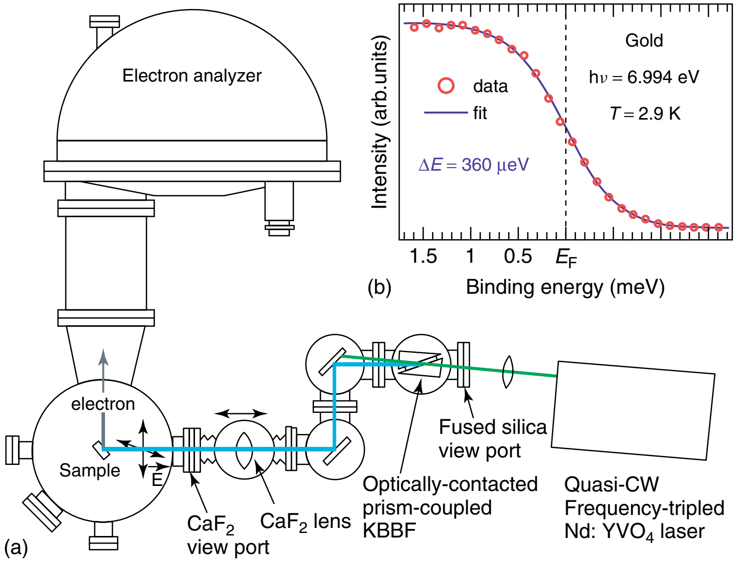
Electron analyzer
electron
Sample
E
CaF$_2$ view port
CaF$_2$ lens
Optically-contacted prism-coupled KBBF
Fused silica view port
Quasi-CW Frequency-tripled Nd: YVO$_4$ laser
(a)
Intensity (arb.units)
data
fit
$\Delta E = 360\ \mu$eV
Gold
hν = 6.994 eV
T = 2.9 K
1.5
1
0.5
E_F
(b)
Binding energy (meV)

Fig. 9.9 (a) Schematic of a laser-based setup for high-resolution photoemission measurements. (b) Fermi edge of gold at low temperature, demonstrating 360 μeV overall resolution. From Kiss *et al.* (2005).

multidetection in hemispherical analyzers has been widely used for core-level spectroscopy (Section 9.4). There exist custom designs with more than one variable being detected simultaneously, such as energy and one angle in toroidal analyzers and two angles in display spectrometers. Simultaneous energy and angle multidetection has become available commercially with hemispherical spectrometers (Martensson *et al.*, 1994), which has led to a renaissance of angle-resolved photoemission from valence states. The states very close to the Fermi level E_{Fermi} can now be mapped with an energy resolution comparable to the thermal energy and a momentum resolution comparable to the inverse mean free path. This is illustrated in Figure 9.5 for ferromagnetic nickel, in which the energy-versus-momentum band dispersion is plotted (Altmann *et al.*, 2001). The $E(k)$ relation can be viewed directly on a television monitor, and the data in Figure 9.5 (upper right) are simply the time average of such a screen image. Currently, a third dimension is added to the multidetection method by incorporating time-of-flight energy detection into a spectrometer that displays both in-plane momentum components. Combined with short laser pulses, one can achieve very high energy resolution.

ZEKE (zero kinetic energy spectroscopy) was invented about 25 years ago (Schlag, 1998) and has developed rapidly since then (Cockett, 2005). Its main impact has been on high-resolution studies of molecular ions where it has yielded information on transition states in chemical reactions. ZEKE can provide energy resolutions 2–3 orders of magnitude higher than conventional photoelectron spectroscopy. The ZEKE technique is based on laser ionization of the molecule followed by a delayed

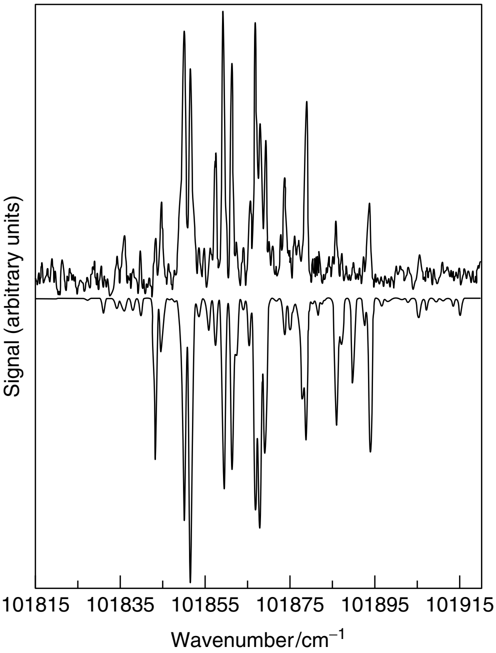

Fig. 9.10 Top: Rotationally resolved ZEKE photoelectron spectrum of a single vibrational structure in deuterated methane (CD_2H_2). Bottom: A simulated spectrum (inverted), broadened by a Gaussian line shape (0.7 cm^{-1} = 87 μeV full width half maximum). From Signorell and Merkt (2000).

electric pulse. In this way, electrons with zero kinetic energy can be separated out from the rest of the emitted electrons. Thus in a time-of-flight arrangement, the zero kinetic energy electrons can be collected separately. The physics behind ZEKE is intimately connected to Rydberg states. Rydberg states have hydrogenic characteristics with a very large principal quantum number n. The electrons in Rydberg states reside in very large orbitals and thus are much less sensitive to decay processes. This is the key for achieving high resolution. The ZEKE method is based on field ionization from the large-n Rydberg states. Even though conventional photoelectron spectroscopy with HeI radiation has been used routinely to study the vibrational fine structure of molecular energy levels, ZEKE makes it possible to resolve the entire set of vibrational levels, not only for the electronic ground state but also for the first excited state (Signorell and Merkt, 2000). It has also been demonstrated that fully resolved rotational spectra provide unprecedented information on the structure of molecular ions, for example, in a number of isotopomers of the methane ion (Keski-Rahkonen and Krause, 1974) (see Figure 9.10). Most of them exhibited such a small Jahn–Teller effect that structural changes can be ascribed to tunneling phenomena. For a normal Jahn–Teller effect, by contrast, the structure will be frozen to a specific minimum.

For detecting the spin of photoelectrons, one typically measures the right/left asymmetry in the electron scattering cross section at heavy nuclei (Mott scattering), after accelerating photoelectrons to 100 keV. A less bulky detector uses scattering at the electron cloud surrounding a heavy nucleus, typically at 150-eV electron energy. It lacks the absolute calibration of a Mott detector, however. In general, spin detection costs 3–4 orders of magnitude in detection efficiency.

9.3
Photoelectric Yield and Absorption Spectroscopy

9.3.1
Mechanism

This section starts with the simplest measurement in photoemission, which consists of determining the total photocurrent versus photon energy. Subsequent sections will deal with energy- and angle-resolved differential cross sections.

Before sophisticated measurements became routine, the determination of the quantum yield near the photoelectric threshold dominated the field. Such measurements were used to confirm the Einstein formula (Millikan, 1916), and to obtain the work function Φ of various materials by finding the photon energy for the onset of photoemission. Accurate determinations involved various extrapolation schemes, predicting certain power laws for the photoelectric yield near threshold (Fowler, 1931; Cardona and Ley, 1978). The photoelectric yield increases roughly like the square of the excess energy, because photoemission is dominated near threshold by the narrowing of an "escape cone" in momentum space that contains the electronic states with sufficient momentum perpendicular to the surface to overcome the surface potential step. Quantum yield measurements had technological consequences in the development of photocathodes for photomultipliers and night vision goggles.

The bulk of today's yield measurements is performed in a regime far above the photoelectric threshold, that is, at core-level absorption edges. In this case, the total electron yield (TEY) is closely related to the absorption coefficient, as long as the escape depth of the photoelectrons is short compared to the absorption length of the photons (Gudat and Kunz, 1972). Essentially, the number of photoelectrons is proportional to the number of core holes generated, which in turn is proportional to the absorption coefficient in the linear absorption regime near the surface. In the following, we will discuss applications of near edge X-ray absorption fine structure (NEXAFS) measurements. They involve transitions from a core level to the lowest unoccupied states. These are typically atomic/molecular orbitals that characterize the local chemical bonding and the magnetism at a specific atom. In semiconductors one observes the lowest conduction bands. Electron yield measurements over a wider photon energy range (hundreds of electronvolts) are usually labeled as EXAFS (extended X-ray absorption fine structure). The related technique of photoelectron diffraction is discussed in Section 9.6.1.

9.3.2
Unoccupied Atomic and Molecular Orbitals

Many molecules or polymers have characteristic unoccupied orbitals that dominate the absorption fine structure within the first tens of electron volts above a core-level threshold, most notably the lowest unoccupied orbital (LUMO), which serves as the electron acceptor in chemical reactions. (Its occupied counterpart, the highest occupied molecular

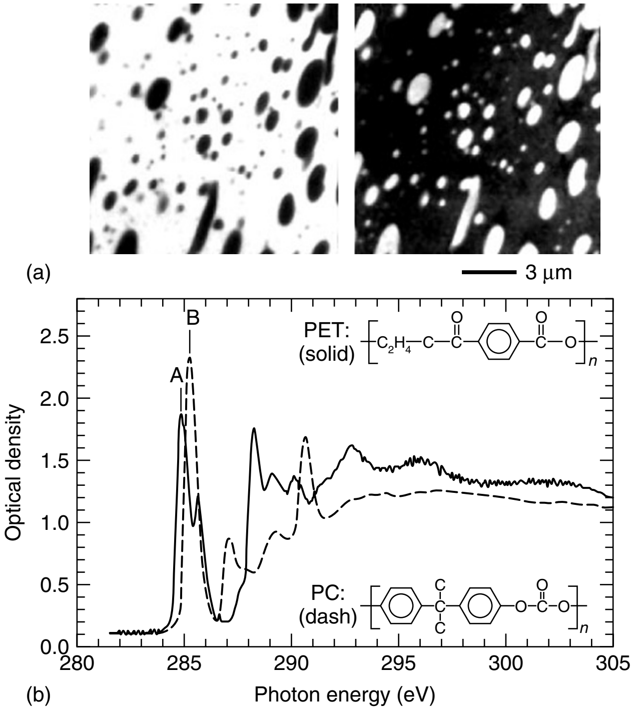

Fig. 9.11 Identification of different polymers in a blend by their π^* orbitals. The carbon 1s core-level absorption spectra (a) are used to identify the C1s to π^* transition energies of two polymers. By tuning into the respective π^* resonances, the two components can be distinguished in the micrograph (b). From Ade *et al.* (1992).

orbital (HOMO), can be measured by valence electron spectroscopy, which is discussed in Section 9.5.1.) The pattern of antibonding π^* and σ^* orbitals often serves as a fingerprint of a molecule. A large database of gas-phase spectra (Hitchcock and Mancini, 1994) helps in identifying surface species in NEXAFS (Stöhr, 1992). The polarization dependence of the absorption from different orbitals provides the orientation of adsorbed molecules. Similar applications exist with polymers. Figure 9.11 shows that two different polymers in a blend can easily be distinguished by their π^* orbitals (Ade *et al.*, 1992).

9.3.3
Solids

Absorption spectroscopy in solids provides information about the unoccupied states, like in the gas phase. In contrast to molecular systems, in which the Coulomb interaction between the excited electron and the core hole significantly distorts the energy scale, there is enough screening in many solid state systems to use the absorption spectrum directly as a measure of the density of unoccupied states. By using different core levels, it is possible to project the density of states onto a particular atomic species and to select states with

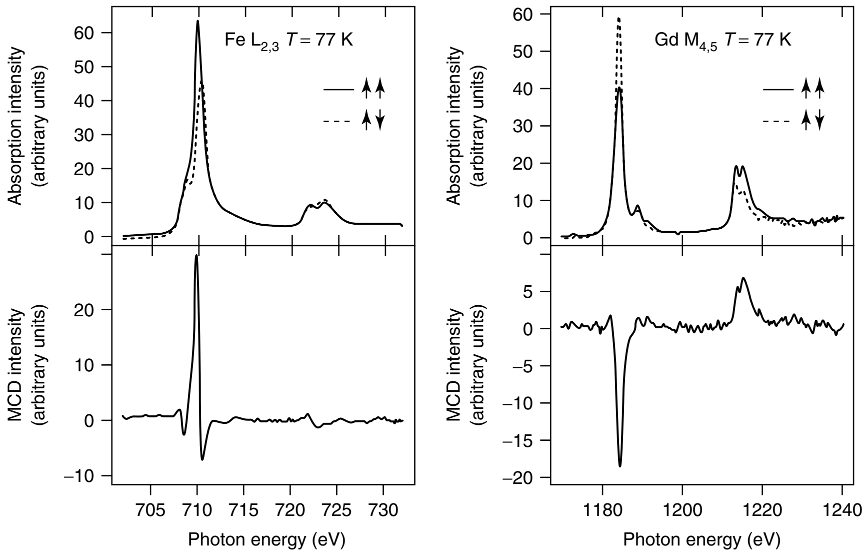

Fig. 9.12 Use of magnetic circular dichroism for determining the magnetization of the elemental species Fe (a) and Gd (b) in a garnet. Two absorption spectra for opposite magnetization directions are shown on top, and their difference at the bottom. From Rudolf *et al.* (1992).

a specific angular momentum l, using the $l \rightarrow (l \pm 1)$ selection rule. A more thorough analysis has to include the electron–hole interaction, which lowers the transition energy and leads to a characteristic multiplet structure, particularly for the localized 2p → 3d and 3d, 4d → 4f transitions.

9.3.4
Magnetic Dichroism

Using circularly polarized light with magnetic samples, it is possible to determine magnetic properties of specific atoms in magnetic alloys, at surfaces, and in molecules. Atomic sum rules produce an estimate of both the spin and the orbital contribution to the local magnetic moment. The two pieces of independent information come from the dichroism of the two spin-orbit partner lines of a core level, that is, the change in absorption induced by flipping the helicity of the light relative to the magnetization of the sample. In addition, one can play all the other selection rules, such as projecting the magnetic moment onto an atomic species or separating magnetism of d-electrons from that of f- and s-, p-electrons. Figure 9.12 gives an example, in which the magnetization of different constituents in a gadolinium iron garnet is determined (Rudolf *et al.*, 1992). This ability of separating the magnetism of different elements is very helpful for tailoring magnetic materials in magnetic nanostructures (Himpsel *et al.*, 1998).

9.3.5
Microspectroscopy

The recent development of third-generation synchrotron light sources with high brilliance makes it possible to perform absorption measurements with a spatial resolution down to tens of nanometers. The two pictures of a polymer blend in Figure 9.8 were taken at two different photon energies, where the two components absorb differently because of their different π^* orbital energies (Ade *et al.*, 1992). In this case, the absorption coefficient was measured directly in a scanning transmission X-ray microscope (STXM). Such experiments are frequently performed with an imaging photoemission electron microscope (PEEM). Microspectroscopy promises to become a versatile method that identifies not only the elemental species in a sample by the choice of the core-level absorption edge but also their chemical state by selecting transitions into individual unoccupied valence states. There are many applications in magnetism, biology, geology, tribology, and environmental science.

9.4
Core-Level Spectroscopy

9.4.1
Origin of Core-Level Shifts

The photoelectric effect shown in Figures 9.1 and 9.2 applies not only to valence electrons but also to the deeper core levels, provided the incoming photons have sufficient energy to overcome the core-level binding energy. Analogous to Eq. (9.1), it is possible to obtain the binding energy of a core electron from the photon energy and the kinetic energy. In a simplified picture, the electrons will thus appear at discrete energies characteristic for each element. XPS is thus element-specific. The binding energy is also sensitive to the chemical environment of an atom, which led to the acronym ESCA.

Quantitative calculations of core-level shifts need to take into account both the initial state electronic structure (the chemical shift) and the relaxation in the final state (which is caused by the removal of an electron) (Fadley, 1978). In the most simplified picture, the magnitude of the chemical shift is correlated with the amount of charge transfer in the initial state. For instance, the binding energy of a core level increases with the oxidation state of an element, because the increasing positive charge on the atom binds a core electron more strongly. This is demonstrated for the oxidation states of silicon in Figure 9.13 (Himpsel, 1990). All oxidation states of silicon are resolvable, with the three peaks between Si and SiO_2 originating from intermediate oxidation states, Si^{1+}, Si^{2+}, and Si^{3+}. An electropositive ligand gives up electrons to silicon, causing a chemical shift in the opposite direction, as shown on the right side of Figure 9.13 for calcium. It is interesting to note that the chemical shift in SiO_2 increases slightly with increasing oxide thickness (not shown), a manifestation of decreased screening in the final state. The intensity of the different oxidation state peaks in Figure 9.13 can be used to estimate the amount of suboxides in the interfacial $Si-SiO_2$ region (Himpsel *et al.*, 1988). This information has been quite valuable as a diagnostic tool for improving the gate oxide in silicon-base field-effect transistors and for developing high-k dielectrics to replace SiO_2 in the most recent generation of silicon microchips.

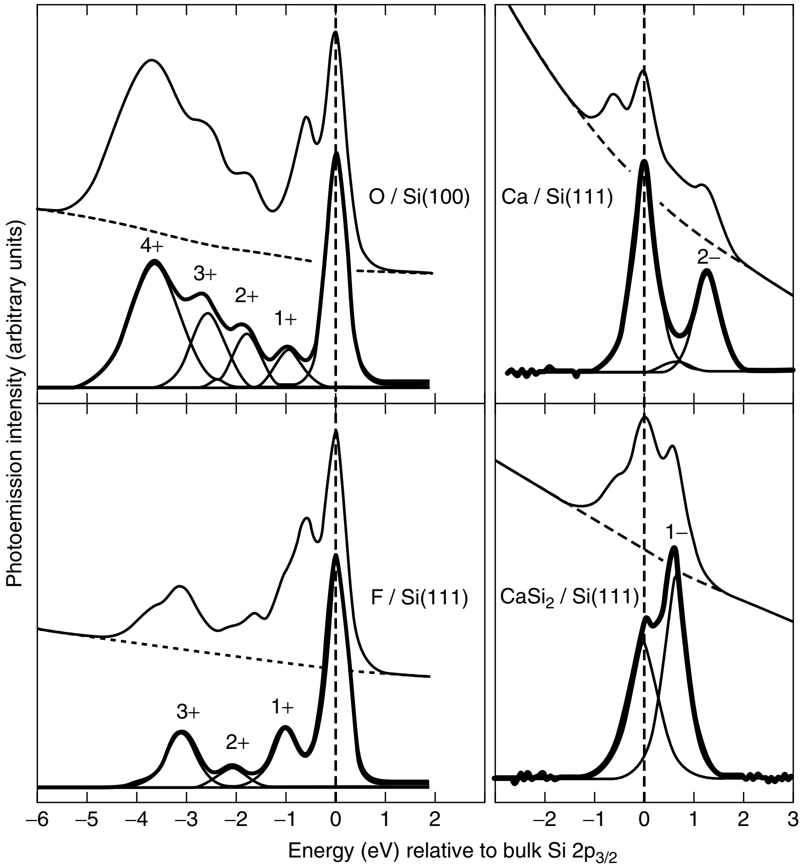

Fig. 9.13 Discrete chemical shifts, observed for different oxidation and reduction states of silicon. Electronegative ligands (O, F) give rise to chemical shifts toward higher binding energy and electropositive ligands (Ca) induce shifts to lower binding energy. Raw data are given on top of each panel. At the bottom, the secondary electron background has been subtracted and the Si $2p_{1/2}$ component removed. After Himpsel *et al.* (1988).

9.4.2
Atoms, Shake-Up, and Shake-Off Satellites

The core-level spectra from a closed-shell noble gas atom should in principle be very simple with discrete peaks corresponding to the different atomic subshells. However, final-state effects give rise to multielectron excitations, and atomic core-level spectra are characterized by a complex satellite structure, first discussed in Carlson (1967). The most likely multielectron process is a two-electron transition. However, higher order transitions and configuration interactions also contribute to the satellite spectra. It is common to distinguish between shake-up states and shake-off states, in which the second electron is excited to a higher bound state or an unbound continuum, respectively. The

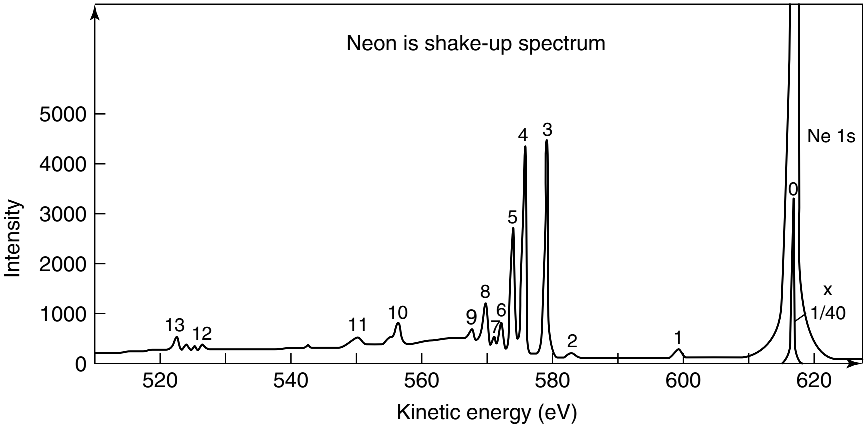

Fig. 9.14 The Ne 1s core-level photoelectron spectrum with accompanying satellites, labeled 1 through 13. From Gelius (1974).

Ne 1s core-level spectrum and the magnified satellite structure of the one-electron 1s line are shown in Figure 9.14. A detailed discussion and compilation of atomic core-level spectra and satellites can be found in Siegbahn and Karlsson (1982) and Sonntag and Zimmermann (1992).

9.4.3
Adsorbates and Surfaces

The chemical core-level shift of molecules adsorbed at surfaces provides information about the nature of chemisorption, dissociative or non-dissociative, as well as structural information. Figure 9.15 illustrates the results for carbon monoxide on Pt(111) (Björneholm *et al.*, 1993) by showing spectra for all relevant core levels (O 1s, C 1s, and Pt 4f). The clean Pt 4f spectrum contains two components, one for the topmost atomic layer, and the other for the underlying bulk Pt (Johansson and Mårtensson, 1980). CO forms a number of ordered phases on Pt (111), depending on the coverage. All of these have distinct arrangements of the CO molecule at the surface (on top, bridge site, and so on), and each of these structures leads to a characteristic chemical shift of the O 1s and C 1s levels. Chemically shifted peaks can thus provide information about both the bonding and the structure of adsorbed systems (Andersen *et al.*, 1991). They are used extensively by companies involved in semiconductor materials processing, developing catalysts, tribology, and surface coating. Tunable synchrotron radiation makes it possible to optimize the surface sensitivity of core-level spectroscopy by detecting photelectrons near the escape depth minimum (Figure 9.4).

Third-generation synchrotron radiation sources have opened up new avenues in core-level studies thanks to higher brightness and improved energy resolution. Time-resolved core-level spectroscopy is now feasible, as demonstrated for CO overlayers on Mo(110) in Figure 9.16 (Jaworowski *et al.*, 2001). As the exposure to CO is increased, the C 1s level first shifts toward higher

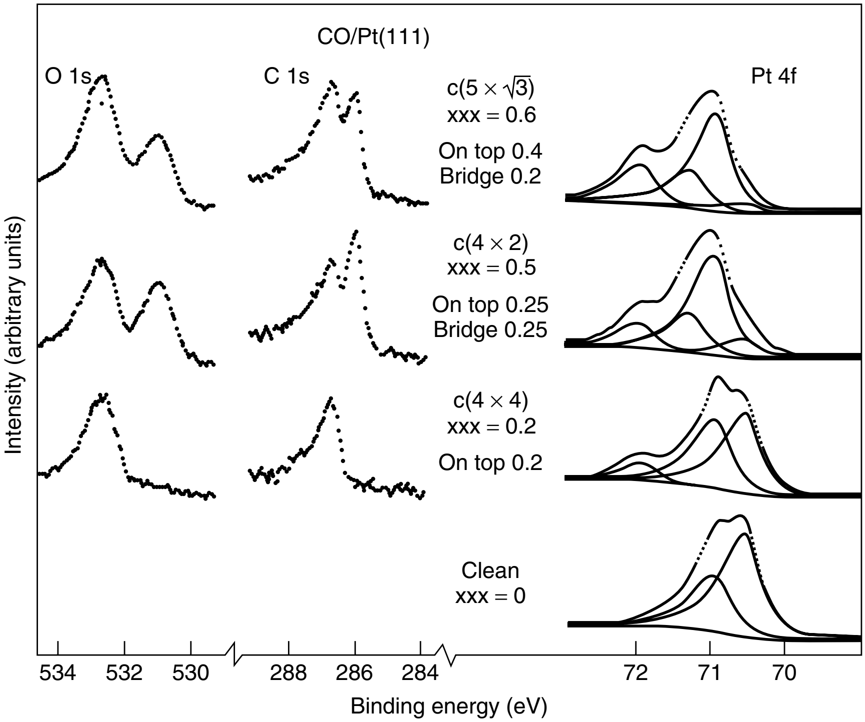

Fig. 9.15 The O 1s, C 1s, and Pt 4f core-level photoelectron spectra for various structures of CO chemisorbed on Pt(111). Adsorbate and substrate core-level shifts provide bonding site information. From Björneholm *et al.*, (1993).

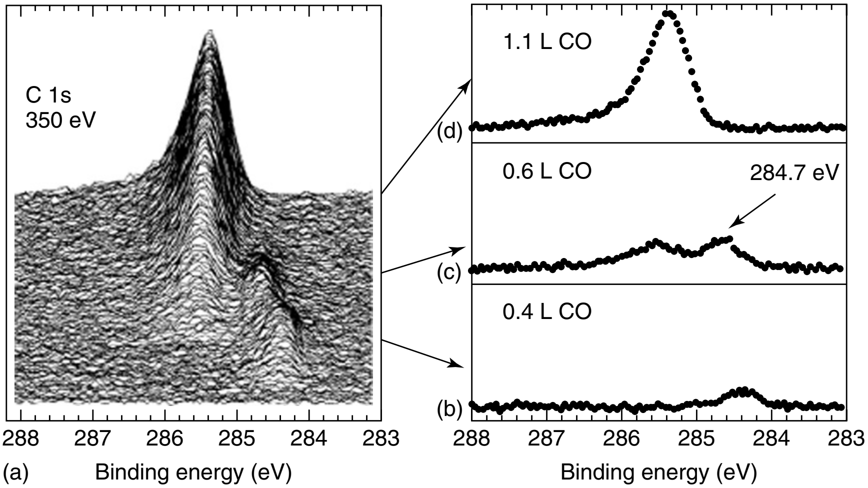

Fig. 9.16 Time-resolved core-level spectra from the C 1s region as a Mo(110) surface is exposed to CO. The recording time for each spectrum is 16 s. From Jaworowski *et al.* (2001).

binding energy and then develops a second component. A further increase of the CO coverage reduces the first component until it eventually disappears at saturation coverage. This evolution can be understood by a change in the tilt angle of the CO molecules at the surface. Such studies are valuable for understanding the precursor state before the molecular dissociates at higher temperature.

9.4.4
Solids

In Section 9.4.1, it was pointed out how core-level spectroscopy can be used for both qualitative (elemental and chemical specific) and quantitative analysis. Here, two other aspects of studies of solids are illustrated. The first example relates to the growth of thin films and their interfaces, a topic of particular importance for semiconductor structures (for an overview, see Weaver (1988)). The growth of a II-VI compound on a III-V semiconductor, that is, CdTe on InSb(100) is characterized by quantitative core-level spectroscopy using synchrotron radiation in Figure 9.17 (Mackey *et al.*, 1986). Two different growth conditions are shown, that is, room temperature (a) and 500 K (b). The core-level spectra show the evolution of the intensity of the 4d levels of In and Sb (substrate), and Cd and Te (overlayers) with increasing CdTe deposition. At room temperature, the overlayer appears to be stoichiometric (no relative change in the Cd and Te core- level intensities), and the In–Sb 4d core levels are attenuated monotonically. The interface is obviously quite abrupt because the substrate core levels are covered up by 7–15 Å of CdTe. The situation is dramatically different for deposition at 500 K. The In 4 d signal does not decrease as much as Sb 4d, the Te 4d is broadened, and the Cd signal is barely detectable. In this case, chemical reactions take place at the interface and an interfacial layer with indium telluride forms with segregated Sb. It can be concluded that the interface is neither abrupt nor stoichiometric on an atomic scale.

A second example deals with core-level spectra, which are completely dominated by atomic final-state effects and are very insensitive to the chemical environment. The most prominent representatives can be found among rare-earth compounds. The 4f core-level spectra for a few rare-earth antimonides are shown in Figure 9.18 (Campagna, Wertheim and Bucher, 1976). All the three spectra can be described very well with final-state multiplets without any consideration of the effects from differences in the chemical environment. From the multiplet structure, one can directly read off the valence of the rare-earth atoms.

9.5
Valence Electron Spectroscopy

Valence states are more complex than core levels, because their wave functions extend out to neighbor atoms and interact with them. Consequently, the set of quantum numbers characterizing valence states is richer, containing the three momentum components as additional quantum numbers (in solids) and forming bonding–antibonding orbital combinations (in molecules). The extra quantum numbers are determined by using additional variables, such as the angular pattern of the photoelectron

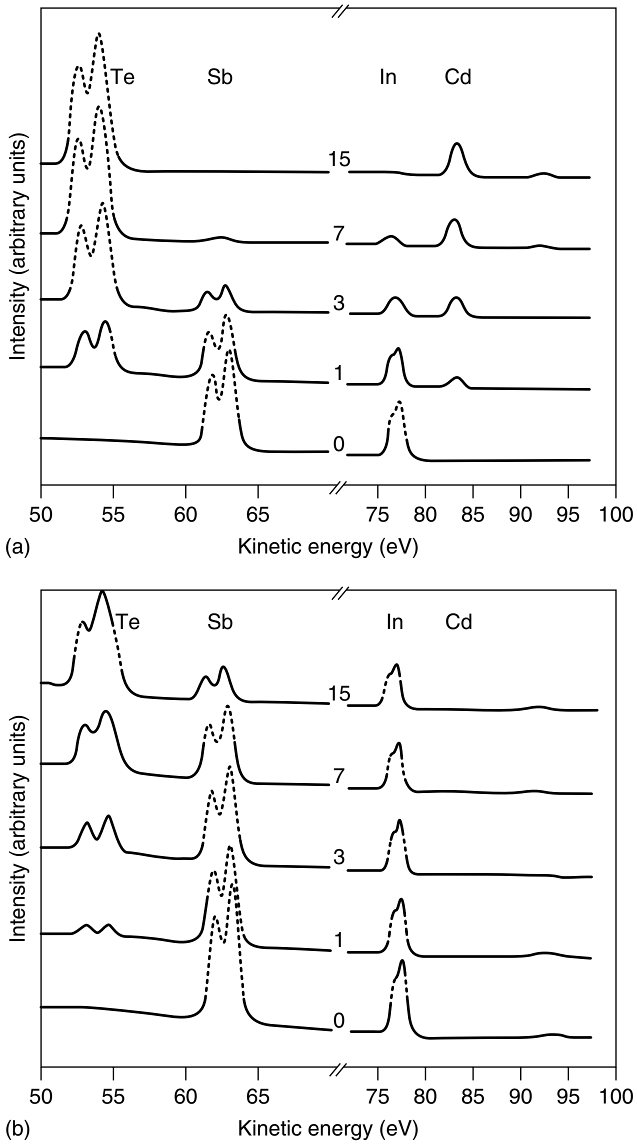

Fig. 9.17 Growth of CdTe on InSb(100), studied via the intensities of the 4d core levels (film thickness given in angstroms). At room temperature (a) the overlayer is stochiometric with a sharp interface. At 500 K (b) an interfacial layer of InTe is formed, together with segregated Sb. From Mackey *et al.* (1986).

and the photon polarization. The optimum energy range for mapping valence states has an upper limit of about 50 eV, because the cross section of valence states decays rapidly at higher photon energies.

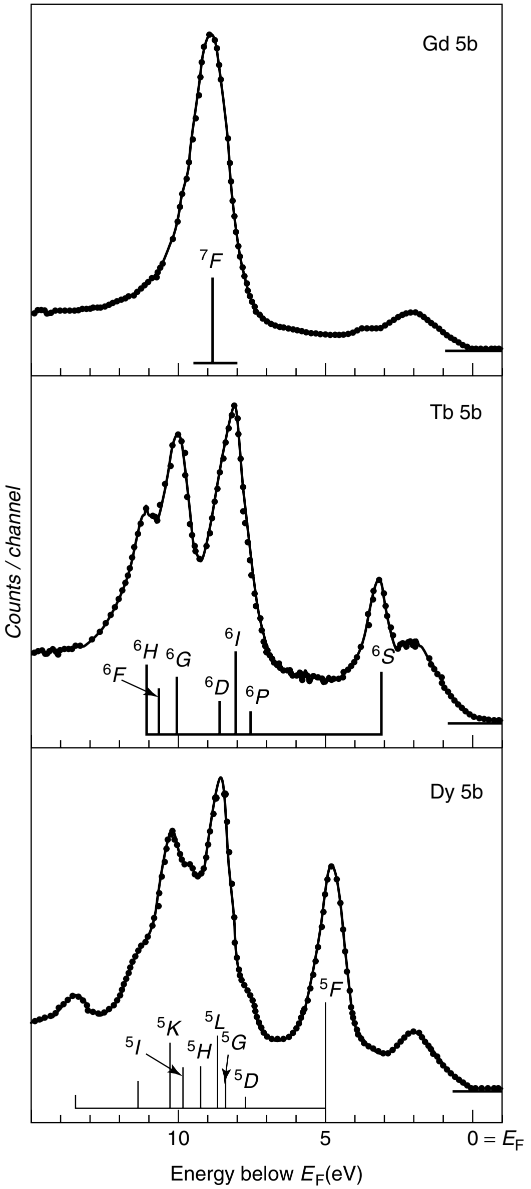

Fig. 9.18 Core-level spectra of the 4f levels for rare-earth antimonides. The agreement with calculated $4f^{n-1}$ multiplets shows that the multiplet structure reflects the number of 4f electrons in the final state. From Campagna, Wertheim and Bucher (1976).

9.5.1
Occupied Atomic and Molecular Orbitals

The valence orbitals of free molecules can be resolved in great detail by UPS (Turner *et al.*, 1970), including vibrational and rotational fine structure. During adsorption at a surface, for example, in catalysis-oriented studies, the molecular orbitals can be used for fingerprinting the adsorbate (Figure 9.19 (Demuth and Eastman, 1974)). Orbitals

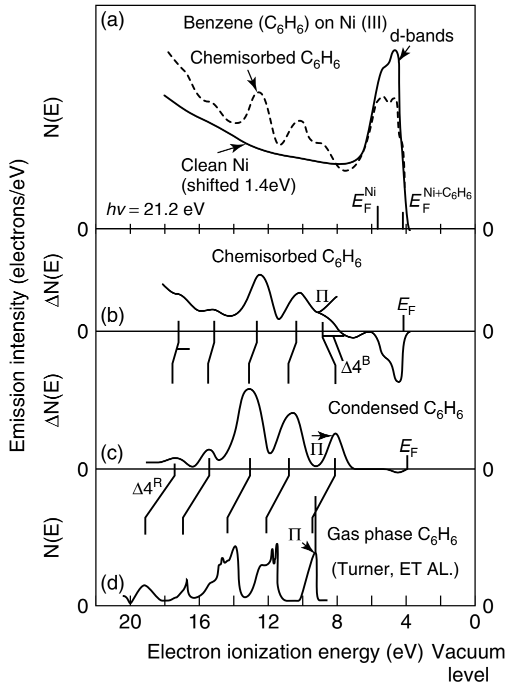

Fig. 9.19 Fingerprinting of adsorbed molecules by photoelectron spectroscopy. Benzene in the gas phase exhibits a characteristic pattern of valence orbitals (Turner *et al.*, 1970), which is shifted rigidly in energy after condensing it at a surface. The stronger chemisorption bond on a nickel surface causes the uppermost bonding orbital to fall out of place. From Demuth and Eastman (1974).

not involved in the surface-to-adsorbate bond experience only a rigid, upward energy shift, which is due to dielectric screening at the surface. Therefore, the pattern of orbital energies at a surface is often sufficient to identify the adsorbate, to decide whether or not a molecule has decomposed, and to pinpoint the orbitals that are involved in the surface bond by their additional energy shift.

Angular effects are averaged out in randomly oriented molecules except for only a single parameter β, which determines the angular distribution $d\sigma/d\Omega$ of photoelectrons in terms of the polar emission angle θ relative to the polarization of the light:

$$\frac{d\sigma}{d\Omega} = \left(\frac{\sigma}{4\pi}\right) \times (1 + \beta P_2(\cos\vartheta)) \qquad (9.9)$$

$P_2(x) = (3x^2 - 1)/2$ is the $l = 2$ Legendre polynomial. The asymmetry parameter β can take values between -1 and 2. At the magic angle $\vartheta = 54°$ the emission intensity is independent of β, making this a useful geometry for quantitative measurements of the cross section σ of a given orbital.

Molecules adsorbed at surfaces can be oriented by a crystalline substrate. In such a case, it becomes possible to determine their orientation by changing the polarization of the light and using dipole selection rules. For example, a p_z orbital is excited with the electric field vector along the z-direction, and a $p_{x,y}$ orbital with the component along the x, y- direction.

9.5.2
Surfaces and Quantum Wells

The surface disrupts the potential for electrons in a solid enough to give rise to separate surface states. They can have localized, broken-bond character, such as in semiconductors, or delocalized, evanescent wave character, such as s, p-states in metals. The d-states at metal surfaces lie somewhere in between. Surface states can be ideally probed with photoelectron spectroscopy when working at kinetic energies of 20–70 eV for the photoelectrons, where their escape depth goes through a minimum that is only a few atomic layers deep. Another way to become more sensitive to surface states is by placing the momentum and the kinetic energy of the photoelectrons in a band gap of bulk states. Both methods require tunable synchrotron radiation.

There are several criteria that help in distinguishing surface states from bulk states. The "crud" test is based on the assumption that a surface state is quenched by an adsorbate, which is true in most, but not all, cases. A cleaner method utilizes the two-dimensional nature of surface states, which implies that the energy of a surface state does not vary with the momentum perpendicular to the surface. This is tested by varying the photon energy at fixed parallel momentum. The most thorough test for a surface state consists of mapping all the bulk bands and ensuring that the state falls into a symmetry gap of the band structure. There are many surface resonances, however, that do not pass this last test but still exhibit strong localization in the outermost layer.

Quantum well states are two-dimensional, like surface states, but they are confined on both sides. Typically they are observed in highly perfect thin films where they are confined on the outer side by vacuum and on the inner side by a step in the inner potential. These films may be viewed as electron interferometers similar to a Fabry–Perot interferometer in optics. The standing electron waves inside the inerferometer correspond to quantum well states. Magnetic quantum well states play a role in magnetoelectronics, for example in the effect of giant magnetoresistance (GMR) that is used in reading heads for magnetically-stored data (compare Himpsel *et al.* (1998) for electronic states in magnetic nanostructures).

Quantum well states can be confined laterally in quantum wires and quantum dots, leading to one- and zero-dimensional electronic states. Currently, it is only possible to measure an ensemble of wires and dots where size inhomogeneities lead to an inhomogeneous broadening of the electronic states. Instruments for angle-resolved photoemission with a spatial resolution in submicrometer regime are under development that eventually will make it possible to study the sharp levels of a single nano object.

9.5.3
Band Mapping

The electronic structure of a crystalline solid is characterized mainly by the quantum numbers E and $\mathbf{p} = \hbar\mathbf{k}$, plus spin and point group symmetry (see Section 9.1.3). This information is traditionally compounded into $E(\mathbf{k})$ band dispersion plots with the appropriate symmetry labels for each band (Figures 9.5 and 9.20). The experimental techniques that can probe the complete set of quantum numbers are photoemission and inverse photoemission, with photoemission probing occupied states and inverse photoemission unoccupied states (see Figure 9.1). There are just

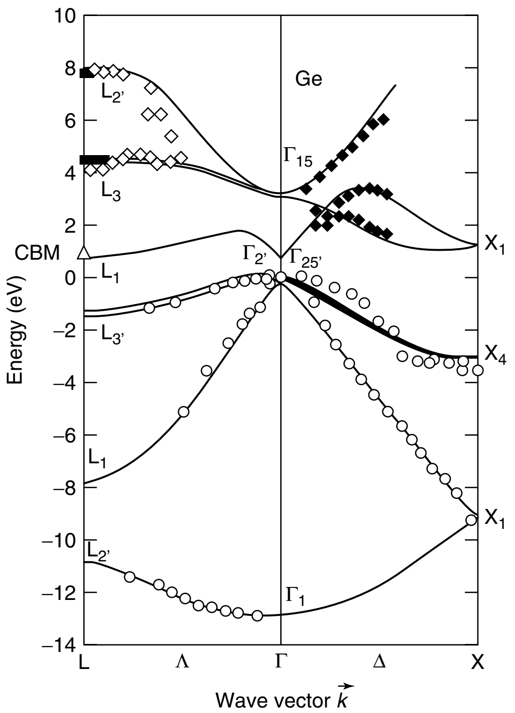

Fig. 9.20 Energy-versus-momentum band dispersions for germanium obtained from angle-resolved photoemission, inverse photoemission, and quasiparticle band calculations. Such band dispersions provide the complete information about electrons in solids. The labels denote the point group symmetry. From Wachs *et al.* (1985), Ortega and Himpsel (1994), and Hybertsen and Louie (1986).

enough variables in these experiments to cover all the unknowns. Roughly speaking, E is determined from the kinetic energy of the electrons, the two momentum components parallel to the surface, $k_{||}$, from the polar and azimuthal angles of the electrons, the perpendicular momentum from the photon energy, spin from the spin of the electrons, and the point group symmetry from the photon polarization. A more thorough analysis, particularly concerning the perpendicular momentum, can be found in various reviews (Himpsel, 1983; Kevan, 1992). Figure 9.20 shows the results from band mapping for germanium, both experimentally and theoretically. Angle-resolved photoemission and inverse photoemission results are compared with state-of-the-art quasiparticle calculations (Wachs *et al.*, 1985; Ortega and Himpsel, 1994; Hybertsen and Louie, 1986). Traditional local density calculations would not give the correct band gap for germanium, and in fact predict it to be metallic. A substantial body of experimental and theoretical band dispersions has been accumulated, and is being compiled as the basic set of electronic structure data for solids (Goldmann and Koch, 1989).

Multidetection allows it to determine the photoemission intensity versus E and $\mathbf{k}$ in quasi-continuous fashion and thereby make a connection to the imaginary part of the Greens function that describes electronic states in the presence of many-electron interactions (see Section 9.1.3). The photoemission intensity $I(E, \mathbf{k})$ can be displayed versus E, k_x (as shown in Figures 9.5 and 9.20) or versus k_x, k_y (as shown in Figure 9.21;

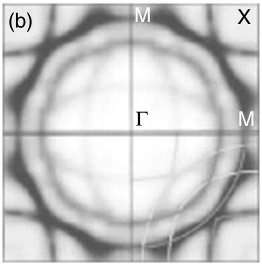

Fig. 9.21 Fermi surface of the two-dimensional material strontium ruthenate. From Damascelli *et al.* (2000).

(Damascelli *et al.*, 2000)). The former represents the $E(k_x)$ band dispersion, the latter the Fermi surface.

Current developments are directed toward high-energy resolution, to map the electronic states at the Fermi level that determine transport phenomena and phase transitions, such as superconductivity (Damascelli *et al.*, 2000; Inosov *et al.*, 2007) and magnetism (Altmann *et al.*, 2001).

9.5.4
Time-Resolved, Pump-Probe Measurements

The lifetime of electronic excitations and their decay mechanisms are at the heart of many technologies, such as optoelectronics and solar cells where long lifetimes are desirable. While optical experiments determine joint density of states, time-resolved photoemission is able to resolve the momentum of the states. The lifetime of electrons and holes strongly increases when approaching the Fermi momentum. It can also become quite long near local band minima in momentum space.

The experiments are typically performed with pulsed lasers in a pump–probe mode, which leads to the name two-photon photoemission (for an overview see Fauster (2003). Typically (but not exclusively), low-energy photons pump electrons from the valence band into long-lived intermediate states and higher energy ultraviolet photons ionize them from there (Höfer *et al.*, 1997; Fauster, 2003). These photons originate from the same laser, either directly or after frequency multiplication, and thus can be synchronized within tens of femtoseconds.

An example is provided in Figure 9.22, in which electrons are pumped into image states and then ionized by the probe pulse (Höfer *et al.*, 1997). Image states are electrons bound to a metal surface by their own image potential, which generates a hydrogen-like series of states converging to the vacuum level, but with 16 times smaller binding energies than in hydrogen. Coherent pumping of several image states with an ultrashort pulse causes quantum oscillations between different image states. These show the time dependence of the photoemission intensity as in Figure 9.19.

9.5.5
Inverse Photoemission

Inverse photoemission or bremsstrahlung isochromat spectroscopy (BIS) may be viewed as the time-reversed photoemission process (see Figure 9.1), with an incident electron and an emitted photon. It complements photoemission by probing unoccupied states, and in

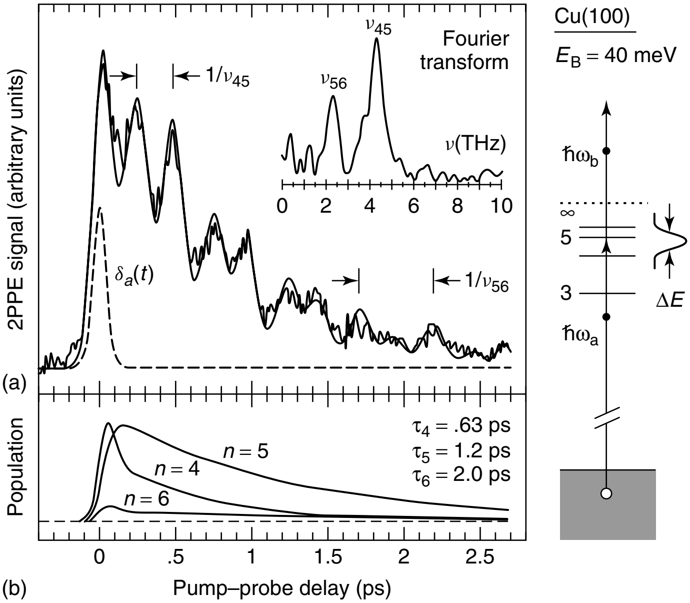

Fig. 9.22 Quantum beats in time-resolved photoemission. A femtosecond laser excites several quantum well states coherently, leading to an oscillatory population change between different states. From Höfer *et al.* (1997).

particular, the unoccupied states between the Fermi level and the vacuum level that are inaccessible to photoemission, neither as initial nor as final state. This gap contains many interesting states, for example, minority spin states in ferromagnets, broken-bond orbitals at semiconductor surfaces, and antibonding orbitals of adsorbed molecules. To visualize the territory covered by photoemission and inverse photoemission, one may examine the analogy between occupied and unoccupied states, bonding versus antibonding states, or donor versus acceptor states. Inverse photoemission poses experimental challenges, because its cross section is 4–5orders of magnitude lower than that for photoemission, due to a phase space factor in the cross section. Typical elastic photoelectron yields are in the range of 10^{-3} electrons per photon, while typical inverse photoemission yields are 10^{-8} photons per electron. Nevertheless, this technique has become a routine complement to photoemission work, as exemplified by the conduction band data in Figure 9.17. Results on semiconductors are reviewed by Himpsel (1990), results on metals and adsorbates by Dose (1985) and Smith (1988).

9.6
Structural Methods

9.6.1
EXAFS and Photoelectron Diffraction

As mentioned in Section 9.3.1, EXAFS provides structural information. The ejected photoelectron from an absorbing atom can be described as an outgoing spherical wave. The outgoing photoelectron wave will be backscattered by neighboring atoms, which thereby will create incoming waves, one from each atom. The interference between the outgoing and incoming waves gives rise to modulation of the absorption coefficient for photon energies above the threshold. The periodicity of this modulation contains information about the distance between the absorbing atom and its neighbors, and the amplitude of the modulation is a measure of the number of neighbors, that is, the coordination number. The EXAFS technique is used extensively for structural determinations of solids, liquids, and biomolecules (Prins and Koningsberger, 1985; Teo and Jay, 1981; Teo, 1986). A surface-sensitive version of EXAFS, with the acronym SEXAFS, surface extended X-ray absorption fine structure can be applied to determine the surface structure of monolayers of adsorbates on surfaces (Prins and Koningsberger, 1985). Experimentally, two parameters are measured: adsorbate-substrate bond lengths and absorption sites.

9.6.2
Photoelectron Diffraction

Although an EXAFS experiment integrates over all outgoing electron energies and emission angles, there is additional information to be gained from the energy and the angular distribution of the diffracted photoelectrons. The outgoing waves of photoelectrons emitted from a specific core-level scatter at nearby atoms and produce intricate diffraction patterns. This information is used in photoelectron diffraction (PhD) (Fadley, 1993; Fadley *et al.*, 1994; Woodruff, 2001). Experimentally, the intensity of a core level is measured either as function of the direction, or as the energy of the photoelectron, providing what is termed as *scanned-angle or scanned-energy data.* Considerable theoretical effort has been made into establishing accurate models relating photoelectron diffraction data to the surface structure (Fadley *et al.*, 1994). Compared to X-ray diffraction, the analysis is complicated by multiple scattering, which is inherent in the strong interaction of electrons with matter. That, on the other hand, provides high surface sensitivity. For photoelectrons with kinetic energies above 500 eV, it turns out that the scattering amplitude is highly peaked in the forward direction (forward focusing), which provides a direct determination of the bond directions for absorbed molecules and epitaxial relationships in thin films.

In addition to measure the momentum parallel to the surface via the angular distribution, it is also possible to vary the magnitude of the momentum by varying the photon energy, like in EXAFS. This requires the access to synchrotron radiation so that the photoelectron energy can be tuned continuously. This method is illustrated by a structure determination for the acetate species, CH_3COO-on the Cu (110) surface (Woodruff, 2001) (see Figure 9.23). Because the two carbon atoms are in different chemical environments, they can be distinguished by their chemical shift. The intensity modulations of the two chemically shifted peaks were measured in a well-defined emission direction as a function of the photoelectron kinetic energy. Such data can be analyzed by a multiple scattering model using trial structures (Woodruff, 2001), analogous to the determinaton of surface structures in LEED.

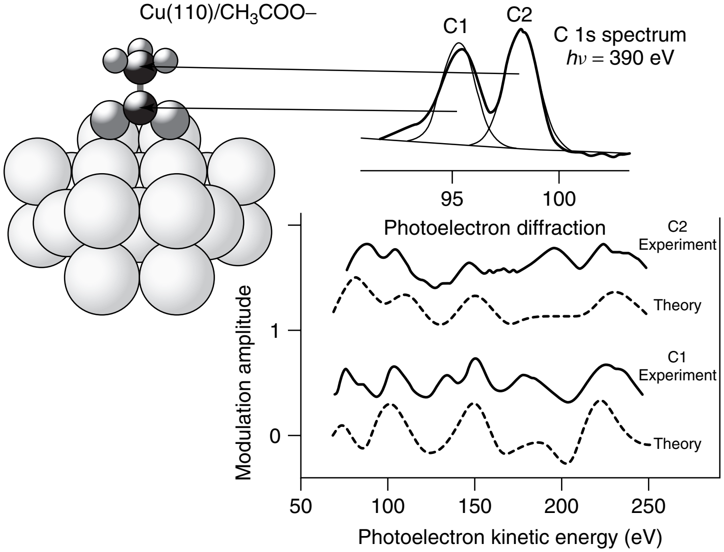

Fig. 9.23 Adsorption geometry of acetate (CH_3COO-) on the Cu(110) surface determined by photoelectron diffraction (PhD). From Woodruff (2001).

9.7
Summary and Outlook

Photoelectron spectroscopy has made significant contributions to the basic understanding of the electronic structure of atoms, molecules, solids, and surfaces. The technique has also been adopted into commercial instrumentation for routine qualitative and quantitative chemical analysis and materials characterization. Presently, thousands of research papers are published each year involving photoelectron spectroscopy in basic research or applications.

The development of the field of photoelectron spectroscopy, both for basic research and applications, has been closely tied to advancements in instrumentation, that is, better light sources and detectors. This trend will certainly continue in the foreseeable future. The third-generation synchrotron radiation sources are presently being implemented and will provide orders of magnitude improvements in brightness. It will be possible to study atoms, molecules, and solids with unprecedented spectral resolution. This will have a major impact for a better theoretical understanding of the complex photoelectron spectra of atoms and molecules. For solids and surfaces, major advances can be expected for studies of the electronic states around superconducting gaps in high-T_c materials, of charge density waves, surface transitions, and surface magnetism. Major breakthroughs should be anticipated in photoelectron microscopy in which all the assets of photoelectron spectroscopy will be applied with a spatial resolution down to a few tens of nanometers. Therefore, it is reasonable to predict that the tremendous growth of photoelectron spectroscopy during the last 40 years will continue well into the future.

9.8
Databases

Core-Level Binding Energies: Cardona and Ley (1978).

Atomic Core-Level Spectra: Siegbahn and Karlsson (1982).

Absorption and Core-Level Spectra of Metal Atoms: Sonntag and Zimmermann (1992).

Molecular Core-Level Absorption: Hitchcock and Mancini (1994). Avialable online at: http://unicorn.chemistry.mcmaster.ca/corex/cedb-title.html

Molecular Core-Level Energies: Bakke, Chen and Jolly (1980).

Molecular Valence Spectra: Turner *et al.* (1970).

Solid State Core and Valence Spectra: Goldmann and Koch (1989).

Cross sections: Yeh and Lindau (1985) and Keski-Rahkonen and Krause (1974).

Various databases are published regularly in Surface Science Spectra, published by the American Institute of Physics for the American Vacuum Society.

Glossary

Al K_α: X-ray line at 1.49-keV photon energy, emitted from the 1s level of aluminum. It is frequently used for core-level photoelectron spectroscopy (see Section 9.2.1).

ARPES: Angle-resolved photoelectron spectroscopy.

Auger Spectroscopy: Use of the Auger electrons emitted during the decay of core holes for elemental analysis at surfaces (see Figure 9.2).

BIS: Bremsstrahlung isochromat spectroscopy. A form of inverse photoemission with constant photon energy and variable electron energy, originally used in the X-ray regime (see Section 9.5.5).

ESCA: Electron spectroscopy for chemical analysis. Mapping of the core-level energies and their chemical shifts, typically using the K_α lines of Al or Mg at 1.49 keV and 1.25 keV, respectively (Siegbahn *et al.*, 1967). Synonym: XPS.

EXAFS: Extended X-ray absorption fine structure, determination of nearest-neighbor distances by interference effects in the core-level absorption spectrum (see Section 9.6.1).

GMR: Giant magnetoresistance, a strong variation of the electrical resistance in a magnetic field induced in sandwiches of magnetic and nonmagnetic metal layers.

GW: Calculation method for quasiparticles (see below), based on solving the Dyson equations with the full electron and photon propagators G and W (therefore the GW acronym), and a free particle approximation for the vertex. It goes beyond the LDA in determining the electronic structure for the excited state, which is measured in photoemission and inverse photoemission (see Section 9.1.3).

He I: Resonance line of helium gas at 21.2-eV photon energy, frequently used for photoelectron spectroscopy of valence states (see Section 9.2.1).

HOMO: Highest occupied molecular orbital in molecules. Together with the LUMO, the HOMO is responsible for most chemical reactions of a molecule, in particular charge transfer reactions, such as in a battery or in a solar cell.

LDA: Local density approximation, representing one of the most common methods for determining electronic states in solids and at surfaces (see Section 9.1.3).

LEED: Low-energy electron diffraction, often used in conjunction with photoelectron spectroscopy to characterize the ordering at surfaces.

LUMO: Lowest unoccupied molecular orbital in molecules. Together with the HOMO, the LUMO is responsible for most chemical reactions of a molecule.

Many-Body Theory: Many-body theory stands for a theory dealing with an infinite number of interacting particles. This is mathematically simpler than dealing with 10^{20} electrons in a macroscopic solid.

MCD: Magnetic circular dichroism, change of absorption between left- and right-handed, circularly polarized light, used for the determination of the magnetization at specific atomic sites (see Section 9.3.4).

NEXAFS: Near Edge X-ray absorption fine structure, mapping of unoccupied states via transitions from core levels (see Section 9.3). Synonyms: XANES, XAS, Partial Yield Spectroscopy.

PhD: Photoelectron diffraction, determination of the local atomic structure of surfaces (see Section 9.6.2).

Propagator: The propagator (or Green's function) describes the propagation of a particle from point A to point B in many-body theory. Mathematically, it is the expectation value of a particle creation operator at point A by an annihilation operator at point B.

Quasiparticle: In many-body theory, an electron loses its simple particle character by dragging other particles alon, such as phonons or other electrons in a solid. This whole particle cloud moves together and becomes a quasiparticle.

SEXAFS: Surface extended X-ray absorption fine structure, determination of nearest-neighbor distances at surfaces by interference effects in the core-level absorption spectrum (see Section 9.6.1).

Synchrotron Radiation: Versatile source of ultraviolet and X-ray light, emitted by high-energy electrons in a storage ring deflected by a magnetic field (see Section 9.2.1).

TEY: Total electron yield. One of the detection modes in XAS/NEXAFS/XANES spectroscopy where all electrons originating from the decay of the core hole are collected (see Section 9.3).

UPS: Ultraviolet photoelectron spectroscopy, mapping of the valence band structure of solids and surfaces with photon energies in the 5 to 50-eV range (see Figure 9.3).

XANES: X-ray absorption near edge structure, mapping of unoccupied states via transitions from core levels (see Section 9.3). Synonyms: NEXAFS, Partial Yield Spectroscopy.

XAS: X-ray absorption spectroscopy. Related terms: NEXAFS, XANES, Partial yield spectroscopy (see Section 9.3).

XPS: X-ray photoelectron spectroscopy. Mapping of the core-level energies and their chemical shifts, typically using the K_α lines of Al or Mg at 1.49 keV and 1.25 keV, respectively. See also ESCA.

XSW: X-ray standing wave, determination of the local atomic structure of surfaces (see Section 9.6.1).

ZEKE: Zero kinetic energy spectroscopy, a high-resolution spectroscopy for studies of molecular ions (see Section 9.2.2).

References

Ade, H., Zhang, X., Cameron, S., Costello, C., Kirz, J., and Williams, S. (1992) *Science*, **258**, 972.

Altmann, K.N., Gilman, N., Hayoz, J., Willis, R.F., and Himpsel, F.J. (2001) *Phys. Rev. Lett.*, **87**, 137201.

Andersen, J.N., Qvarford, M., Nyholm, R., Sorensen, S.L., and Wigren, C. (1991) *Phys. Rev. Lett.*, **67**, 2822.

Anderson, P.W. and Schrieffer, R. (1991) *Phys. Today*, **44**(June), 54.

Bakke, A.A., Chen, H.-W., and Jolly, W. (1980) *J. Electron. Spectrosc.*, **20**, 333.

Björneholm, O., Nilsson, A., Tillborg, H., Bennich, P., Sandell, A., Hernnäs, B., Puglia, C., and Mårtensson, N. (1993) in *MAX-Lab Activity Report 1992* (eds J.N. Andersen, L.S.O. Johansson, and R. Nyholm), National Laboratory, Lund, pp. 58–59.

Campagna, M., Wertheim, G.K., and Bucher, E. (1976) *Struct. Bond.*, **30**, 99.

Cardona, M. and Ley, L. (1978) *Photoemission in Solids I, Topics in Applied Physics,* Vol. 26. Springer, Berlin.

Carlson, T.A. (1967) *Phys. Rev.*, **156**, 142.

Chattopadhyay, S., Cornacchia, M., Lindau, I., and Pellegrini C. (eds) (2001) *Physics of and Science with the X-Ray Free-Electron Laser,* Vol. 581, American Institute of Physics, New York, Conference Proceedings.

Cockett, M.C.R. (2005) *Chem. Soc. Rev.*, **34**, 933.

Damascelli, A., Lu, D.H., Shen, K.M., Armitage, N.P., Ronning, F., Feng, D.L., Kim, C., Shen, Z.-X., Kimura, T., Tokura, Y., Mao, Z.Q., and Maeno, Y. (2000) *Phys. Rev. Lett.*, **85**, 5194.

Demuth, J.E. and Eastman, D.E. (1974) *Phys. Rev. Lett.*, **32**, 1123.

Dose, V. (1985) *Surf. Sci. Rep.*, **5**, 337.

Einstein, A. (1905) *Annn. Phys.*, **17**, 132.

Fadley, C.S. (1978) in *Electron Spectroscopy: Theory, Techniques and Applications,* Vol. 2 (eds C.R. Brundle and A.D. Baker), Academic Press, New York, pp. 1–156.

Fadley, C.S., Thevuthasan, S., Kaduwela, A.P., Westpal, C., Kim, Y.J., Ynzunza, R., Len, P., Tober, E., Zhang, F., Wang, Z., Ruebush, S., Budge, A., and Van Hove, M.A. (1994) *J. Electron. Spectrosc. Relat. Phenom.*, **68**, 19.

Fadley, C.S. (1993) in *Synchrotron Radiation Research: Advances in Surface Science* (ed. R.Z. Bachrach), Plenum, New York.

Fauster, Th. (2003), in *Solid State Photoemission and Related Methods*, Chapter 8 (eds W. Schattke and M.A. Van Hove), Wiley-VCH Verlag Gmbh, Weinheim.

Feuerbacher, B., Fitton, B., and Willis, R.F. (1978) *Photoemission and the Electronic Properties of Surfaces,* John Wiley & Sons, Inc., New York.

Fowler, R.F. (1931) *Phys. Rev.*, **38**, 45.

Gelius, U. (1974) *J. Electron. Spectrosc. Relat. Phenom.*, **5**, 985.

Goldmann, A. and Koch, E.-E. (1989) *Numerical Data and Functional Relationships in Science and Technology, New Series, Group III, Electronic Structure of Solids: Photoemission Spectra and Related Data,* Vols. 23a,b,c, Springer, Berlin.

Gudat, W. and Kunz, C. (1972) *Phys. Rev. Lett.*, **29**, 169.

Hertz, H. (1887) *Ann. Phys.*, **32**, 983.

Himpsel, F.J. (1983) *Adv. Phys.*, **32**, 1.

Himpsel, F.J. (1990) *Surf. Sci. Rep.*, **12**, 1.

(a) Himpsel, F.J., McFeely, F.R., Taleb-Ibrahimi, A., Yarmoff, J.A., and Hollinger, G. (1988) *Phys. Rev. B*, **38**, 6084; (b) Himpsel, F.J., Meyerson, B.S., McFeely, F.R., Morar, J.F., Taleb-Ibrahimi, A., and Yarmoff, J.A. (1990) in *Photoemission and Absorption Spectroscopy of Solids and Interfaces with Synchrotron Radiation* (eds M. Campagna and R. Rosei), North Holland, Amsterdam, pp. 203–236.

Himpsel, F.J., Ortega, J.E., Mankey, G.J., and Willis, R.F. (1998) **47**, 511.

Hitchcock, A.P. and Mancini, D.C. (1994) *J. Electron. Spectrosc.*, **67**, 1.

Höfer, U., Shumay, I.L., Reuss, Ch., Thomann, U., Wallauer, W., and Fauster, Th. (1997) *Science*, **277**, 1480.

Hybertsen, M.S. and Louie, S.G. (1986) *Phys. Rev. B*, **34**, 5390.

Hüfner, S. (ed.) (2007) *Very High Resolution Photoelectron Spectroscopy, Lecture Notes in Physics, Vol.715,* Springer, Heidelberg.

Inosov, D.S., Borisenko, S.V., Eremin, I., Kordyuk, A.A., Zabolotnyy, V.B., Geck, J., Koitzsch, A., Fink, J., Knupfer, M., Büchner, B., Berger, H., and Follath, R. (2007) *Phys. Rev. B*, **75**, 172505.

Jaworowski, A.J., Smedh, M., Borg, M., Sandell, A., Beutler, A., Sorensen, S.L., Lundgen, E., and Anderson, J.N. (2001) *Surf. Sci.*, **492**, 185.

Johansson, B. and Mårtensson, N. (1980) *Phys. Rev. B*, **21**, 4427.

Keski-Rahkonen, O. and Krause, M.O. (1974) *At. Data Nucl. Data Tables*, **14**, 139.

Kevan, S.D. (ed.) (1992) *Angle-Resolved Photoemission*, Elsevier/North Holland, Amsterdam.

Kiss, T., Kanetaka, F., Yokoya, T., Shimojima, T., Kanai, K., Shin, S., Onuki, Y., Togashi, T., Zhang, C., Chen, C.-T., and Watanabe, S. (2005) *Phys. Rev. Lett.*, **94**, 057001.

Koch, E.E. (ed.) (1983) *Handbook on Synchrotron Radiation*, North Holland, Amsterdam.

Mackey, K.J., Allen, P.M.G., Herrenden-Harker, W.G., Williams, R.H., Whitehouse, C.R., and Williams, G.M. (1986) *Appl. Phys. Lett.*, **49**, 354.

Martensson, N., Baltzer, P., Bruhwiler, P.A., Forsell, J.O., Nilsson, A., Stenborg, A., and Wannberg, B. (1994) *J. Electron. Spectrosc. Relat. Phenom.*, **70**, 117.

Millikan, R.A. (1916) *Phys. Rev.*, **7**, 18 and 355.

Ortega, J.E. and Himpsel, F.J. (1994) *Phys. Rev. B*, **47**, 2130.

Plummer, E.W., Shib, J., Tanga, S.-J., Rotenberg, E., and Kevan, S.D. (2003) *Prog. Surf. Sci.*, **74**, 251.

Prins, R. and Koningsberger, D. (eds) (1985) *X-Ray Absorption: Principles, Applications, Techniques of EXAFS, SEXAFS and XANES*, John Wiley & Sons, Inc., New York.

Rudolf, P., Sette, F., Tjeng, L.H., Meigs, G., and Chen, C.T. (1992) *J. Magn. Magn. Mater.*, **109**, 109.

Schattke, W., Van Hove, M.A., Garcia de Abajo, F.J., Diez Muino, R., and Manella, N. (2003) in *Solid State Photoemission and Related Methods*, Chapter 2 (eds W. Schattke and M.A. Van Hove, Wiley-VCH Verlag GmbH.Weinheim.

Schlag, E.W. (1998) *ZEKE Spectroscopy*, University Press, Cambridge.

Schott, G.A. (1912) *Electromagnetic Radiation*, Cambridge University Press, Cambridge.

Schwinger, J. (1949) *Phys. Rev.*, **75**, 798, 1912.

Siegbahn, H. and Karlsson, L. (1982) in *Encyclopedia of Physics*, Vol. XXXI (ed. S. Flügge), Springer, New York, pp. 215–468.

Siegbahn, K., Nordling, C., Fahlman, A., Nordberg, R., Hamrin, K., Hedman, J., Johansson, G., Bergmark, T., Karlsson, S.-E., Lindgren, I., and Lindberg, B. (1967) *ESCA: Atomic, Molecular, and Solid State Structure Studied by Means of Electron Spectroscopy*, Almquist and Wiksells, Stockholm.

Signorell, R. and Merkt, F. (2000) *Faraday Discuss.*, **115**, 205.

Smith, N.V. (1988) *Rep. Prog. Phys.*, **51**, 1227.

Stöhr, J. (1992) *NEXAFS Spectroscopy, Springer Series in Surface Sciences, Vol. 25*, Springer, Heidelberg.

Sonntag, B. and Zimmermann, P. (1992) *Rep. Prog. Phys.*, **55**, 911.

Teo, B.K. (1986) *EXAFS: Basic Principles and Data Analysis*, Springer, New York.

Teo, B.K. and Jay, D.C. (eds) (1981) *EXAFS Spectroscopy: Techniques and Applications*, Plenum, New York.

Togashi, T., Kanai, T., Sekikawa, T., Watanabe, S., Chen, C.-T., Zhang, C., Xu, Z., and Wang, J. (2003) *Opt. Lett.*, **28**, 254.

Turner, D.W., Baker, C., Baker, A.D., and Brundle, C.R. (1970) *Molecular Photoelectron Spectroscopy*, Wiley-Interscience, London.

Yeh, J.J. and Lindau, I. (1985) *At. Data Nucl. Data Tables*, **32**, 1.

Wachs, A.L., Miller, T., Hsieh, T.C., Shapire, A.P., and Chiang, T.C. (1985) *Phys. Rev. B*, **32**, 2326.

Weaver, J.H. (1988) in *Analytical Techniques for Thin Films, Treatise on Materials Science and Technology*, Vol. 27 (eds K.N. Tu and R. Rosenberg), Academic Press, New York, pp. 15–63.

Winick, H. (ed.) (1994) *Synchrotron Radiation Sources – A Primer*, World Scientific, Singapore.

Winick, H. and Doniach, S. (Eds) (1980) *Synchrotron Radiation Research*, Plenum Press, New York.

Winick, H., Xian, D., Ye, M.-H., and Huang, T. (eds) (1989) *Applications of Synchrotron Radiation*, Gordon & Breach, New York.

Woodruff, D.P. (2001) *Surf. Sci.*, **485**, 49.

PHOTOLUMINESCENCE

Claude Benoit à la Guillaume, *Groupe de Physique des Solides, Université Denis Diderot et Université Pierre et Marie Curie, Paris, France.*

Introduction 497
1. **Physical Principles** 498
1.1 General Scheme 498
1.2 Photoluminescence Quantum Efficiency 498
1.3 Main Types of Luminescence Materials 498
1.4 Specificity of Photoexcitation 499
2. **Methods and Equipment** 499
2.1 Excitation Sources 499
2.2 Spectrometers and Detectors 499
2.3 Control of Sample Temperature 500
2.4 Photoluminescence Kinetics and Excitation Spectrum 500
2.5 Time Resolution 500
2.6 Spatial Resolution 501
2.7 External Perturbations 501
3. **Applications** 502
3.1 Fundamental Applications ... 502
3.1.1 Organic Crystals 502
3.1.2 Luminescent Centers in Transparent Solids 502
3.1.3 Bulk Semiconductors 503
3.1.3.1 Intrinsic Luminescence 503
3.1.3.2 Effect of Impurities and Defects 504
3.1.4 Low-Dimensional Semiconductor Structures ... 506
3.1.5 Optical Pumping and Hot Luminescence 507
3.1.6 Exciton Transport 508
3.2 Excitation–Medium Interaction 508
3.2.1 Electron–Phonon Coupling, Polarons, Self-Trapping 508
3.2.2 Electron–Magnetic-Ion Exchange Interaction, Magnetic Polaron 508
3.3 High Excitation Effects 508
3.3.1 Saturation, Optical Gain, Laser 509
3.3.2 Many-Body Effects: Biexcitons, Electron–Hole Plasma or Liquid 509
3.3.3 Higher-Order Emission Processes 509
3.4 Practical Applications 509
3.4.1 Nondestructive Material Testing 509
3.4.2 Devices: Optically Pumped Solid-State Lasers and Amplifiers 510
Glossary 510
Works Cited 511
Further Reading 511

INTRODUCTION

Luminescence is described as light emitted by a *cold* excited body, in contrast with the light emitted by incandescent bodies. There is a variety of excitation processes: In photoluminescence, the excitation results from light absorption. Because of the quantum nature of light and considering only low excitation conditions, one absorbed photon can induce at best one luminescence photon whose energy is necessarily smaller than (or equal to) that of the exciting photon (Stokes's law). Historically, luminescence was restricted to the visible part of the spectrum, but nothing prevents an extension toward the ultraviolet or infrared regions. By comparison with other excitation processes like electron beam (see CATHODOLUMINESCENCE), electric field (see ELECTROLUMINESCENCE), injection in *p-n* junctions (see LIGHT-EMITTING DIODES), chemiluminescence, and bioluminescence, photoluminescence allows the best control over the excitation conditions and for that reason is the most useful process for studying the fundamental properties of luminescent materials. Development of luminescent materials for fluorescent tubes and cathode-ray

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

tubes took place in the 1940s. Study of photoluminescence in organic and semiconductor materials started at the end of the 1950s and developed very fast with the advent of lasers. The tremendous improvements of semiconductor quality linked to microelectronic applications, in terms of purity and doping control, allowed the identification of well-established luminescence mechanisms. Conversely, photoluminescence has become a very common nondestructive test of semiconductor structures.

1. PHYSICAL PRINCIPLES

1.1 General Scheme

Photoluminescence can be described as a three-step process:

1. absorption of one photon of the excitation source creating in the material a primary excitation;
2. evolution of this primary excitation by energy relaxation or energy transfer toward secondary excitation at lower energy; and
3. coupling of this secondary excitation to light, giving rise to the spontaneous emission of one luminescence photon.

Steps **1** and **3** require, as in the field of optical properties (see Optical Properties of Solids), the knowledge (a) of the electronic energy spectrum of the material and (b) of the strength of optical transition between any pair of energy levels. Traditionally, the energy spectrum of condensed matter is described in the framework of the *Born-Oppenheimer approximation*, which relies upon the fact that the movements of the nuclei are much slower than those of the electrons, because of the large mass ratio. Actually, electronic excitations and lattice vibrations are coupled with a strength depending on the material under consideration. Luminescence is linked mainly with electronic excitations; coupling between electronic and lattice excitations enters into play mainly in step **2** to allow for some energy relaxation. Let us note that Stokes's law is strictly valid only at temperature $T = 0$; at finite T, there is a finite probability to see the excitation jumping to a higher energy by phonon absorption: This leads to anti-Stokes emission.

1.2 Photoluminescence Quantum Efficiency

The photoluminescence quantum efficiency η is defined as the ratio of the number of photons emitted n_l to the number of excitation photons absorbed n_a: $\eta = n_l/n_a$. The fact that $\eta < 1$ is due to nonradiative processes that allow complete energy dissipation of the exciting photon. Among these are the multiphonon processes where an electronic excitation ΔE is dissipated into n lattice vibrations. Since the probability of such process decreases very fast with n, an energy gap large enough between the emitting state and the ground state is required in order to have a sizable quantum efficiency. This explains why metals, which exhibit a continuum of electronic excited states, are not luminescent materials.

1.3 Main Types of Luminescence Materials

To be more specific, we will consider three main types of luminescent materials:

1. Molecular organic materials (see Semiconductors, Organic—Electronic Properties). In that case, the building block is the isolated molecule (see Molecules), which is only weakly perturbed by its environment, e.g., a solvent or its inclusion in a molecular solid, tied by van der Waals interactions.
2. Luminescent centers in large-gap (hence transparent) insulators, like oxides, tungstates, silicates, or glasses (see Electron States, Localized). The most studied are transition-metal or rare-earth ions; the isolated ions have usually a rich spectrum of excitations linked to their partially filled d or f shell, providing opportunities to obtain luminescence in the visible. Perturbation of these ions by the crystalline matrix is usually treated in the crystal-field approximation. Another example is the F center in alkali halides, where an electron is trapped in an anion vacancy.
3. Semiconductors. They present an energy gap E_g between a filled valence band and an empty conduction band. In that case, luminescence takes place at an energy close to the band gap, which spans a range of values from near ultraviolet to infrared, depending on the material. The building blocks are now free electrons in the conduction band and free holes in the valence

band, characterized by their effective masses (see ELECTRONIC STRUCTURES OF SOLIDS). When Coulomb interaction between these particles and with impurities is taken into account, weakly bound states are built, which can be described in the effective-mass approximation: free excitons (see EXCITONS) or bound excitons; they dominate the luminescence spectrum in pure semiconductors at low temperature.

1.4 Specificity of Photoexcitation

Through the use of monochromatic excitation, very specific energy levels can be excited; this can be used either to optimize the luminescence quantum efficiency or for fundamental research in order to get a better understanding of step **2**. In addition, one can take advantage of light polarization, which allows additional selectivity in anisotropic materials and is at the roots of a technique known as "optical pumping" (see Sec. 3.1.5).

Depending upon the energy difference $E_a - E_l$ between excitation and luminescence, one can distinguish three cases (applicable mainly to semiconductors):

1. Resonance luminescence, when $E_a = E_l$. This is sometimes called Rayleigh scattering.
2. Hot luminescence, when E_l spans a region between E_a and the main luminescence band.
3. Thermalized luminescence corresponding to the main luminescence band, close to the band gap.

The relative importance of these three contributions is governed by a system of rate equations describing such an out-of-equilibrium situation.

2. METHODS AND EQUIPMENT

A typical experimental photoluminescence setup is shown in Fig. 1. It is composed of an excitation source, optical elements to focus excitation on the sample and collect the luminescence light, a spectrometer (or filter) to analyze the luminescence, and a detector followed by a data acquisition system.

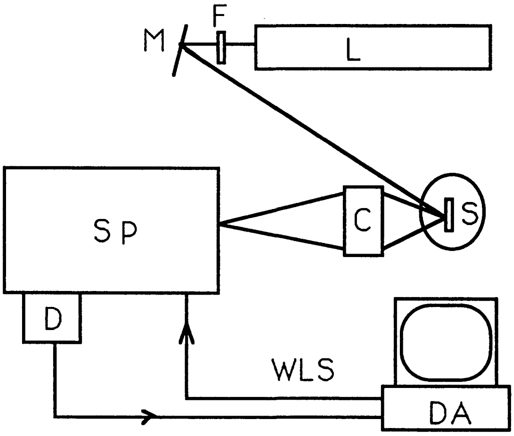

FIG. 1. Typical photoluminescence setup. *L*, laser; *F*, filter; *M*, mirror; *S*, sample; *C*, condenser; *SP*, spectrometer; *D*, detector; *DA*, data acquisition system; WLS, wavelength scanning.

2.1 Excitation Sources

Before the advent of lasers, the most useful sources were high-pressure mercury or xenon lamps. Both have a broad continuous spectrum extending from 300 nm to near infrared. Appropriate filters have to be inserted in order to cut source emission in the spectral region of sample luminescence. Lasers (see LASERS, GAS; LASERS, SOLID STATE) are excitation sources much more convenient for several reasons:

1. They are much more powerful.
2. Since they are well collimated, it is easy to use 100% of their power, as compared to 15% at best for conventional sources.
3. Their emission is essentially monochromatic.

Notice, however, that a filter may be required to suppress plasma emission lines in gas lasers. A wide range of wavelengths is now available from the infrared (CO_2 near 10 μm, CO near 5 μm, and Nd solid-state laser at 1.06 μm), through gas lasers in the visible (He-Ne, Ar, Kr), and recently gas lasers with intracavity second-harmonic generation for the 250-nm region.

2.2 Spectrometers and Detectors

If in some cases a filter could be enough to select a luminescence line, in most cases a spectrometer is needed to get the luminescence spectrum (see SPECTROMETERS, OPTICAL). The most widely used are grating spectrometers, either single or double. The light emitted by the sample should be focused on the entrance slit of the spectrometer by an objective or a combination of mirrors in the infrared to avoid absorption and dispersion.

In the ultraviolet, visible, and near infrared down to 1 μm, the most useful detectors are photomultipliers (see DETECTORS, RADIATION). Semiconductor detectors (photovoltaic, photoconductor, avalanche *p-n* junction) cover the visible and infrared range down to 10 μm (see DETECTORS, SEMICONDUCTOR). Chopping of the excitation at a frequency f followed by synchronous detection is common practice. Let us point out that this is mandatory only when using detectors exhibiting $1/f$ noise, since the detector contributes to the noise in a window δf around the frequency f. With photomultipliers, the best is to use a discriminator followed by a counter. Detector arrays, featuring up to 2048 (or even more) elements, provide a very efficient detection method, since wavelength scanning is suppressed. Fourier transform spectrometers (see INTERFEROMETERS AND INTERFEROMETRY; SPECTROMETERS, INFRARED) have also been used in the infrared; for a given spectral resolution, they can accommodate a luminescence beam of larger extension.

2.3 Control of Sample Temperature

Sample temperature is an important parameter in photoluminescence experiments, since, in general, line narrowing is expected on cooling. Moreover, new ground or excited states (see Sec. 3.2) can appear at low temperature, like some type of condensation (see Sec. 3.3.2). Optical cryostats (see CRYOGENICS) allow easy operation down to 1.5 K, the best cooling conditions being achieved by immersion of the sample in superfluid liquid helium. In peculiar cases, operation down to about 0.1 K has been obtained with a dilution refrigerator; in that case, excitation and luminescence light are transported through optical fibers, in order to minimize heat dissipation. Experiments between helium and room temperature are also easily performed—for example, using commercial closed-cycle refrigerators.

2.4 Photoluminescence Kinetics and Excitation Spectrum

It is often useful to study photoluminescence *kinetics*, i.e., how the luminescence intensity depends upon the excitation intensity, taking care of the fact that different luminescence lines may exhibit different kinetics. Another powerful way to obtain information about the photoluminescence process is the study of the dependence upon excitation wavelength of the luminescence spectrum. This is performed usually by recording the *photoluminescence excitation spectrum*, the excitation wavelength being scanned at fixed luminescence wavelength. A tunable source is required; this can be done either with a broad-band source and a monochromator or, more conveniently, with a tunable laser, like a dye laser (see ORGANIC DYE LASERS) or a solid-state tunable laser (titanium–sapphire or *F*-center). Often, the excitation spectrum is close to the absorption spectrum of the sample. Light *polarization* is an additional useful feature: by inserting a polarizer (either linear or circular) in the excitation and/or the luminescence beam, one can gain information about the symmetries of the primary excited state and the final emitting state, when the sample presents some kind of anisotropy, either natural or induced by external perturbations (see Sec. 2.7). Selection of the luminescence photon wave vector, realized by inserting a slit out of focus to limit the solid angle of the luminescence beam, can also be useful with anisotropic samples. Even with isotropic samples, a polarized excitation can induce some degree of polarization of the luminescence, a technique known as *optical pumping* (see Sec. 3.1.5).

2.5 Time Resolution

Time resolution adds new capabilities to directly measure characteristic times in the sample, like lifetime of states or transfer time between states. It requires first a pulsed excitation. For decay time longer than about 10 ns, modulation by electro-optical (see QUANTUM-OPTOELECTRONIC DEVICES) or acousto-optical (see MODULATORS AND DEMODULATORS, ACOUSTICAL) devices can be used. For shorter time scales down to 0.1 ns, mode-locked lasers can be used. Synchronously pumped dye (or titanium–sapphire) lasers can provide pulses reaching the subpicosecond range (see ULTRAFAST PULSE PHENOMENA), tunable in a substantial wavelength range.

Detection was done first using fast oscilloscopes or boxcar techniques, which require a scanning of the delay between excitation and gate defining the measurement window. A better method, in conjunction with photomultiplier detectors, is the so-called *single-photon counting* technique: the time delay between the excitation pulse and the first luminescence photon detected is measured with a time-to-amplitude converter, and the data are stored in the multichannel mode. This method is very efficient at low levels of excitation since essentially all luminescence photons are detected. With fast discriminators, time resolution below 50 ps is achievable. A second method relies upon *up-conversion* in a nonlinear medium (see OPTICS, NONLINEAR): Part of the excitation pulse is combined with the luminescence beam in a nonlinear crystal under phase-matching conditions; photons at an energy that is the sum of excitation and luminescence photon energies are detected. This method, less sensitive than the previous one, allows a time resolution down to the subpicosecond range; moreover, it can be used for infrared luminescence. A third method uses a *streak camera*: The luminescence light focused on a transparent photocathode creates electrons focused on a fluorescent screen after deflection by a sawtoothlike electric field synchronized with the excitation pulse. The bright trace on the screen, which is an image of the time-dependent luminescence, is read by an array of detectors. This method is fairly sensitive with a typical time resolution of about 10 ps and a spectral domain of wavelength below 850 nm. In all cases where the expected time resolution is below 10 ps, time broadening due to the passage on one grating has to be suppressed with the use of a double-grating spectrometer in the substractive mode.

Detection of *quantum beats* occurs when two nearby energy states separated by ΔE have been both excited by a short pulse; during their coherent decay, beats with a period $\Delta t = \hbar/\Delta E$ can be detected on the luminescence signal, under appropriate polarization conditions, thus allowing measurement in the time domain of small energy splitting.

2.6 Spatial Resolution

Using well-focused excitation in a spot of a few microns, any extension of the luminescence spot gives information about the diffusion of the excitations. This can be done by inserting a slit in the plane of an enlarged image of the emitting region. Such measurements can be performed in the time-resolved mode, allowing the measurement of a signal $S(\Delta x,\Delta\tau)$, where Δx is the distance to the exciting spot and $\Delta\tau$ is the delay after the exciting pulse. Under typical conditions, the analysis of $S(\Delta x,\Delta\tau)$ provides access to the lifetime and the diffusion coefficient of the excitations.

Another more powerful technique is provided by *four-wave mixing* experiments (see OPTICS, NONLINEAR), but this is outside the scope of this article.

2.7 External Perturbations

The purpose of external perturbations applied to the sample is to induce shift or splitting (see ELECTRON LEVEL SPLITTING) of the energy levels in order to help their theoretical identification. On solid samples, *hydrostatic pressure* up to 10 GPa can be applied by use of a diamond anvil cell: This induces shifts and in some cases phase transitions. *Uniaxial stress* of much lower intensity can be applied; it can induce splitting associated with polarization effects. It can also be applied as a modulation, using piezoelectric elements.

Magnetic fields up to about 10 T can be applied by use of split-coil superconducting magnets to allow light propagation perpendicular or parallel to the field. Different effects are induced, like Zeeman splitting, diamagnetic shift, or Landau level splitting; they can be detected on the photoluminescence spectrum, mainly at low temperature. Photoluminescence and electron paramagnetic resonance (see ELECTRON PARAMAGNETIC RESONANCE) can be combined into a technique known as *optical detection of magnetic resonance* (Cavenett, 1981): The change in spin populations at resonance is detected through modification of the intensity of some polarized components of photoluminescence lines. This method is quite sensitive and provides a correlation between different magnetic resonance components (i.e., different g factors) and different luminescence components. A related method, *level-crossing spectroscopy*, does not require microwaves: When two magnetic sublevels happen to cross, transfer between these two levels is enhanced and changes in polarized photoluminescence intensity can be recorded as bump or dip as a function of the field strength.

To conclude this section, let us point out that several extensions of photoluminescence as described in Secs. 2.3 to 2.7 can be combined in a given experiment, provided they are not mutually exclusive.

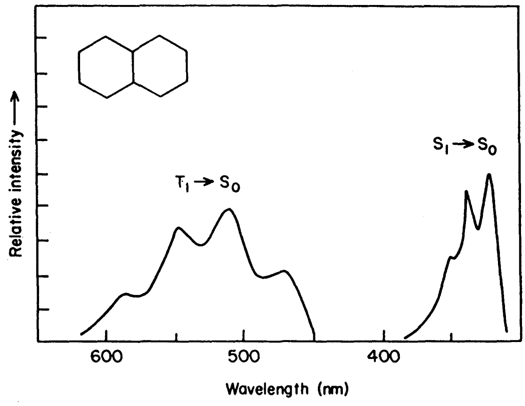

FIG. 2. Fluorescence ($S_1 \rightarrow S_0$) and phosphorescence ($T_1 \rightarrow S_0$) spectra of naphthalene at room temperature. The $T_1 \rightarrow S_0$ spectrum was obtained using a rigid polystyrene solution of naphthalene. The two spectra are plotted on different intensity scales.

3. APPLICATIONS

3.1 Fundamental Applications

In general, photoluminescence spectroscopy provides information about energy levels, as do other kinds of optical spectroscopy (see SPECTROSCOPY OF SOLIDS), but with special emphasis upon the lowest excited states. In addition, increasing the excitation level may induce new collective states (see Sec. 3.3.2). We give first examples of low-level photoluminescence in the three kinds of luminescence materials defined in Sec. 1.3. Particular emphasis is put on semiconductors.

3.1.1 Organic Crystals As an example, we will consider solid naphthalene or anthracene. Looking first into the lowest energy states of the isolated molecule, one finds a singlet ground state denoted S_0, and a first excited singlet state S_1 with a corresponding triplet state T_1 more than 0.5 eV below. These states are coupled to vibrational modes of the molecule. The optical transition $S_0 \rightarrow S_1$ is allowed whereas $S_0 \rightarrow T_1$ is forbidden. So the luminescence is dominated by the $S_1 \rightarrow S_0$ transition, accompanied on its low-energy side by several molecular vibration sidebands: This is called "prompt" fluorescence. The T_1 state is also populated from S_1 by intersystem crossing; it gives rise to slow phosphorescence (Fig. 2).

In the molecular crystal, where binding is due to van der Waals forces, the prompt fluorescence is not very different from that of the molecule; there is a moderate red shift due to the binding. The main difference is the possibility of rapid *energy transfer* to neighbors: At low temperature, this takes place through the formation of Frenkel excitons (see EXCITONS), which are delocalized excitations characterized by a wave vector **k**. When the temperature increases, the mean free path of the Frenkel exciton can decrease below the intermolecular distance. Then one has an incoherent hopping regime for these excitations. That rather fast energy transfer has several consequences: If the crystal is doped with a parent molecule, like naphthacene in anthracene, S_1 excitons can be trapped by the impurities; then the matrix prompt fluorescence is quenched at the benefit of naphthacene emission at lower energy. Another effect comes from collisions between two triplet excitons resulting in regeneration of a singlet S_1 (called also triplet fusion); this is the origin of "delayed" fluorescence.

Let us add that more complicated molecules, the organic dyes, are used in liquid solution to achieve tunable lasers (see ORGANIC DYE LASERS).

3.1.2 Luminescent Centers in Transparent Solids Most of the examples concern rare-earth or transition-metal ions in large-gap matrices (see SOLID, LIQUID, AND GASEOUS INSULATOR MATERIALS). Going from rare-earth elements to transition metals, as a result of the smaller diameter of *f* orbitals as compared with *d* orbitals, rare-earth ions are less perturbed by the crystal field and less coupled to lattice vibrations. Thus, emission and absorption lines of rare-earth ions are quite narrow. Some inhomogeneous broadening occurs in glasses (Er^{3+} in silica glass fiber, Nd^{3+} in glass). Strong coupling with lattice vibrations can lead to broad absorption and emission bands, like in the case of titanium-doped sapphire. Usually, ion concentration is in the few percent range so that samples of length in the centimeter range are required for laser emission. Concentration quenching is observed mainly for transition-metal ions (above 0.3% in ruby: Cr^{3+} in Al_2O_3). For rare-earth ions, stoichiometric material like $PrCl_3$ exhibits no luminescence quenching. Such material can give a micron-size optically pumped laser.

Another type of defect in ionic crystals has been widely studied: the *F* center, which corresponds to an electron trapped in an anion vacancy. Since the electron wave function extends toward the neighboring cations, strong electron–phonon coupling induces a large Stokes shift between absorption and emission (Fig. 3).

A nonlinear effect giving rise to "up-conversion" occurs in some transparent crystals doped with two types of luminescent ions, the first emitting at an energy E and the second at an energy E' smaller than but close to $2E$. If a lot of ions of the first type are excited, an energy transfer from two such ions to one of the second kind may occur, followed by the emission of a photon at energy E'. If E is in the near infrared, then an IR–visible conversion results.

3.1.3 Bulk Semiconductors

3.1.3.1 Intrinsic Luminescence. High-temperature photoluminescence in pure semiconductors is dominated by the band-to-band process: Free electrons recombine with free holes. This is a typical case of thermalized luminescence where electrons and holes are in thermal equilibrium with the lattice. However, since their densities n and p exceed the values of their densities n_0 and p_0 at thermal equilibrium in the unexcited semiconductor, their statistical distributions are described by separate *quasi-Fermi* levels, one for the electrons F_e and one for the holes F_h, with $F_e > F_h$. At vanishing excitation intensity, F_e and F_h both coincide with the Fermi level at equilibrium.

Another approach is given by the use of the principle of *detailed balance*, to get a relationship between light absorption and emission in a semiconductor at temperature T (Van Roosbroeck and Shockley, 1954). At thermodynamic equilibrium, at every photon energy, the number of photons emitted (through spontaneous and stimulated emission) is equal to the number of blackbody photons absorbed. Thus, a luminescence spectrum at temperature T can be deduced from the absorption spectrum, as well as a radiative lifetime (i.e., the lifetime if the only existing recombination process were the radiative one).

In fact, the effect of electron–hole Coulomb interaction is already present in the ex-

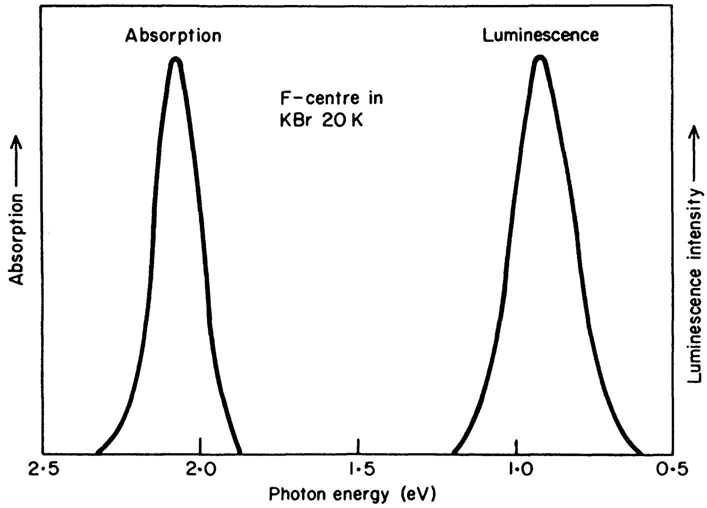

FIG. 3. Absorption and luminescence transitions between the ground state and the lowest electronic excited state of the *F* center in KBr.

perimental absorption spectrum and, thus, also in the luminescence. This interaction between an electron and a hole induces the formation of bound states, analogous to those of the hydrogen atom, called *Wannier excitons* (see EXCITONS). Their binding energy $\mathfrak{R}^*$ is scaled from the hydrogen Rydberg $\mathfrak{R}$ by the electron-hole reduced mass μ and the relative permittivity ϵ according to $\mathfrak{R}^* = (\mu/\epsilon^2)\mathfrak{R}$.

On the other hand, the optical band gap of a semiconductor can be either direct or indirect in reciprocal space. Since the photon wave vector is very small at the Brillouin-zone scale, first-order optical transitions take place vertically in **k** space. In direct-gap semiconductors, the extrema of conduction and valence bands are at the same point in the Brillouin zone; so first-order optical transitions are in general allowed (except in special cases). In indirect-gap semiconductors, band extrema are at different points $\mathbf{k}_v$ and $\mathbf{k}_c$ in the Brillouin zone; in that case, first-order optical transitions near the extrema are always forbidden. The next higher-order process, in pure semiconductors, has to involve the absorption or emission of a phonon of wave vector $\mathbf{k}_c - \mathbf{k}_v$. As a consequence, optical absorption near threshold for indirect-gap material is weak, and the radiative lifetime of the lowest exciton is large (in the millisecond range for Ge). In contrast, in direct-gap material, the optical absorption near threshold is strong; it exhibits well-resolved exciton lines, and the photoluminescence spectrum is in a region of strong optical absorption. This has several consequences. At high temperature, when electron–hole recombination dominates, once a photon has been emitted in the material, it has a good probability to be reabsorbed, creating again an electron–hole pair before escaping the material: Such a process, called *photon recycling*, enhances the diffusion of the electron–hole pairs and modifies the radiative lifetime, especially if the quantum efficiency is good. At lower temperature, when free excitons dominate the luminescence spectrum, a free exciton of small wave vector **k** will be absorbed, creating a photon with the same **k** that later on will create an exciton **k**, and so on, unless it makes a collision with a phonon or a lattice defect. This occurs if the exciton–photon coupling is strong enough; the result of this coupling is a mixed exciton–photon excitation that has been called *polariton*. Under such conditions, a luminescence photon has a chance to be created when a polariton hits the sample surface. In fact, photoluminescence is not the best method to study polaritons; resonant Brillouin scattering is the best way to determine the polariton dispersion relation.

The tendency of low-temperature photoluminescence to accumulate the excitations in the lowest possible excited state is also valuable to detect weak transitions related to *triplet* states, which are only weakly coupled to light; this allows the measurement of electron–hole *exchange* interaction. As an example, consider the case of exciton–polaritons in CdS, a crystal of uniaxial symmetry; the exciton is built from a spin-$\frac{1}{2}$ electron and a spin-$\frac{3}{2}$ hole, quantized along the *c* axis of the crystal. So excitons of total angular momentum $J = 1$ or $J = 2$ are obtained. The first are strongly coupled to light of polarization $\perp c$ and show a strong polariton effect; the second, our triplet, are weakly coupled to light. By using polarization and wave-vector selection in a photoluminescence experiment on pure CdS platelets, it was possible to measure accurately the energy difference between the $J = 2$ excitons and the longitudinal $J = 1$ polariton (see Fig. 4).

3.1.3.2 Effect of Impurities and Defects (see ELECTRON STATES, LOCALIZED). At 2 K, low-level photoluminescence in the near–band-gap spectral region is usually dominated by the effect of *shallow impurities*, donors and acceptors that are well-known dopants. Under typical conditions, unintentionally doped semiconductors present some compensation; under weak photocreation of electron–hole pairs, ionized donors and acceptors capture respectively free electrons and free holes. The first recombination process comes from pairs D^0-A^0 of neutral donor and acceptor (Dean, 1973). At somewhat higher temperature, a fraction of electrons may remain free, thus allowing electron-to-neutral acceptor transitions, noted e-A^0. When all donors and acceptors are neutral, additional electron–hole pairs can form free excitons, and these excitons can form bound states with different impurities or defects. Among these are an exciton bound to a neutral donor, denoted D^0X, or to a neutral acceptor, denoted A^0X (Bebb and Williams, 1972). Depending on the electron-to-hole mass ratio, binding to an ionized impurity is possible, e.g., D^+X in CdS. These

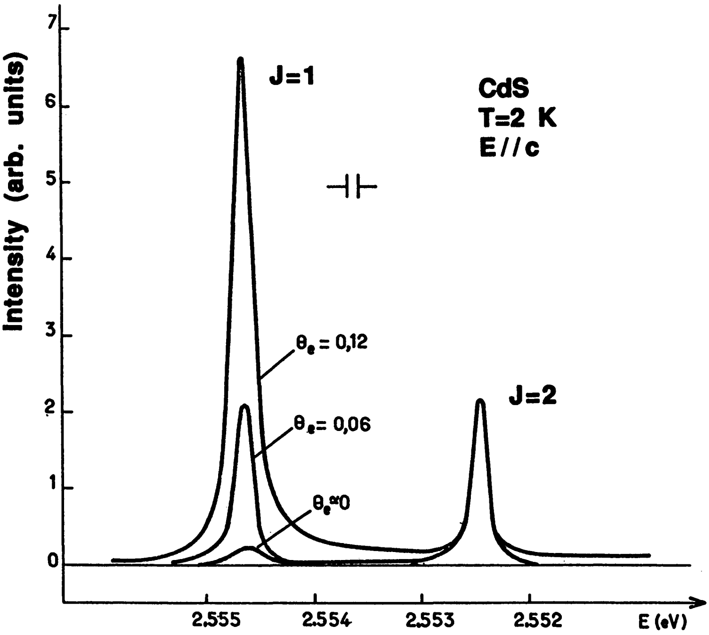

FIG. 4. Exchange splitting of free exciton in CdS. The photoluminescence spectrum is obtained at 2 K with polarization selection ($E \parallel c$) and photon wave-vector selection (**k** nearly normal to c), c being the axis of this hexagonal crystal: θ_e is the angle between **k** and a plane normal to c. For $\theta_e \to 0$, the $J = 1$ exciton becomes purely longitudinal, and the coupling with a photon $\parallel c$ vanishes.

complexes, in direct- as well as indirect-gap materials, give narrow luminescence lines that can be split in a magnetic field of a few tesla. If one considers the complex D^0X, its main decay process gives a photon and a neutral donor D^0 in its ground $1S$ state; however, there is less probable decay process, named *two-electron transition*, which leaves the neutral donor in its nS state (nP states are forbidden by symmetry considerations). So, from the energy difference between the two luminescence lines, the energy difference between the $1S$ and $2S$ states of the donor D^0 can be measured, which is very helpful for its identification. The same occurs for A^0X (Fig. 5).

When one atom of a semiconductor is replaced by an atom of the same column of the Mendeleev table, if they are not too different, an *alloy* can be obtained in the whole composition range. Examples of such behavior are $Ge_{1-x}Si_x$, $Ga_{1-x}Al_xAs$, and $GaAs_{1-x}P_x$. As a first approximation, they behave just like ordinary semiconductors whose parameters are now x dependent. However, specific features appear in the low-temperature photoluminescence spectra of these alloys. Since they are random alloys, statistical fluctuations of local concentration are unavoidable. They produce a tail of bound exciton states readily observable in the luminescence spectrum. In addition, time resolution shows the existence of energy transfer toward lower-energy states.

If the electronegativity of the impurity is too different from that of the host, in general, the solubility will be small. An impurity less electronegative can bind a hole; in contrast, it can bind an electron if it is more electronegative. Then, an additional carrier can neutralize the complex, which is called an *isoelectronic exciton*. This happens, for example, in GaP when Bi or N is substituted for P. Such complexes exhibit an electron–hole exchange splitting, readily observable

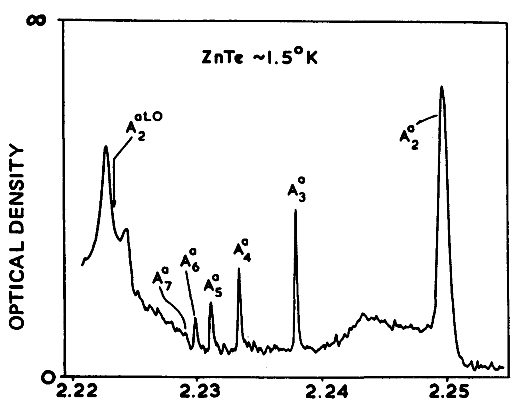

FIG. 5. Two-hole transitions in pure as-grown ZnTe at 1.5 K. The main line A_1^a, related to an exciton bound to a neutral acceptor, is at 2.375 eV; the lines A_n^a correspond to transitions starting from the ground state of A^0X, leaving A^0 in an nS excited state.

in low-temperature photoluminescence. Excitons can also be bound more strongly to pairs of isoelectronic impurities at different separation. Let us stress a strong apparent difference between GaAs and GaP. In GaAs at atmospheric pressure, it is impossible to detect any N-related luminescence signal. Under the action of a hydrostatic pressure larger than 2.2 GPa, photoluminescence lines related to N appear. This is because the energy levels of the N isoelectronic exciton are linked mainly to the conduction-band minima of X symmetry, at the Brillouin-zone edge, the lowest in GaP. In contrast, these minima are above the Γ minimum, at the Brillouin-zone center, in GaAs. The effect of pressure is to push up the Γ minimum and push down the X minima. With increasing pressure, GaAs becomes GaP-like, and the N-related lines appear.

Among the variety of *deep centers*, let us distinguish the *transition metals*. When they are incorporated in large-gap semiconductors, one recovers the situation encountered in Sec. 3.1.2. Mn in ZnS is such an example of a traditional luminescent material. In fact, Mn incorporated in most II-VI compounds with band gap larger than 2.2 eV produces an analogous yellow-orange luminescence band coming from the excitation of the ^5d shell of Mn^{2+}. Hybridization of transition-metal d orbitals with valence-band orbitals of the host crystal is somewhat larger in semiconductors. Another feature of interest is the magnetic moment carried by some of these transition-metal ions; consequences for the luminescence are examined in Sec. 3.2.2. Another class of deep centers may be called *lattice defects*, like vacancies, interstitials, antisites defects in compounds, and complex defects involving impurities and lattice defects. Photoluminescence is just a tool, among a lot of others, to help defect identification. Let us just mention the famous *EL*2 defect in GaAs; it is a native defect that exists in three charged states (0, +, and ++), giving energy levels near midgap and providing semi-insulating GaAs substrates. It involves likely an arsenic antisite As_{Ga} that might dissociate into a gallium vacancy V_{Ga} and an interstitial arsenic As_i in some of its possible charge states; this model is still under debate.

3.1.4 Low-Dimensional Semiconductor Structures With the advent in the seventies of refined growth methods [see CVD (CHEMICAL VAPOR DEPOSITION) and MOLECULAR BEAM EPITAXY], the growth of artificial layered structures became possible, with an accuracy approaching the atomic-layer limit. Other methods were developed later to obtain one-dimensional wires and zero-dimensional boxes (or dots) with somewhat poorer accuracy. Photoluminescence turned out to be a very efficient tool to study these structures since it requires no further preparation of the sample, in contrast with optical absorption measurements, which require substrate removal. A widely used substitute for absorption spectrum is the photoluminescence excitation spectrum. Figure 6 shows results for a GaAs quantum well; the photoluminescence excitation spectrum exhibits peaks that are related to the *quantization* of electron and hole movement along the growth axis (see ELECTRON STATES IN ZERO-, ONE-, AND TWO-DIMENSIONAL STRUCTURES). Excitonic effects are enhanced in two-dimensional structures, in agreement with theory. Another interesting feature is the Stokes shift between the lowest excitation peak and the photoluminescence peak. This shift, which depends on the well width and on the growth conditions, is related to the fact that it is impossible to grow perfect interfaces: There are always some steps or *interface roughness* and also some amount of interdiffusion. This specific kind of disorder induces a weak localization of the excitons, readily observed at 2 K.

Among the very large number of fundamental applications of photoluminescence in low-dimensional structures, let us single out the following, taken under extreme conditions: the study of a two-dimensional free-electron system in modulation-doped heterostructures in a transverse magnetic field. Through magnetotransport measurements, the integer quantum Hall effect and later the more unexpected fractional quantum Hall effect were discovered in such structures. Photoluminescence performed down to 0.1 K and in magnetic field up to 25 T did provide new insight into the nature of the ground state of this electron system: A plane of acceptors was introduced in GaAs at some distance from the interface; a weak photoexcitation neutralizes these acceptors, which can then recombine radiatively with the electrons, thus providing a precise measurement of the energy $E(H)$

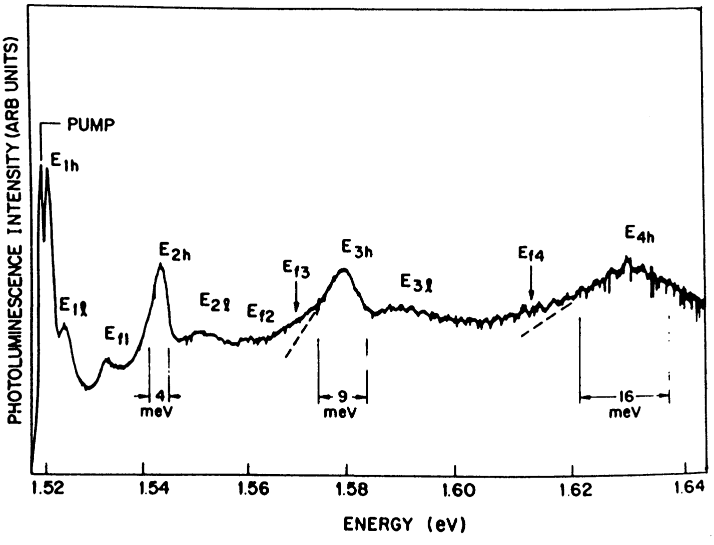

FIG. 6. Excitation spectrum of the exciton luminescence of a 26-nm GaAs quantum well in GaAlAs barriers. The strongest E_{nh} transitions ($n = 1$ to 4) are related to electron and heavy hole in their nth quantized states; transitions noted E_{nl} involve electron and light hole.

of the electron system as a function of the magnetic field H. Indeed, anomalies in the $E(H)$ curve were observed at the field values where the fractional quantum Hall effect is detected. This gives strong support to the theory, based on electron-electron Coulomb interaction, which predicts a new kind of condensation into an incompressible electron liquid for these particular fractional Landau level-filling factors. Experiments are still in progress to determine under what conditions the electron system could crystallize, as predicted by Wigner.

3.1.5 Optical Pumping and Hot Luminescence Absorption of circularly polarized light allows us to excite unequally sublevels in a degenerate set of levels. This *optical pumping* scheme, first developed on atoms, leads, through spin–orbit coupling, to *spin orientation.* The degree of spin orientation, $p_e = (n^+ - n^-)/(n^+ + n^-)$ where n^+ and n^- are the populations of the two spin states, can be monitored by measuring the polarization of the luminescence. This was done first in p-type GaSb (or GaAs). Because of spin–orbit effect near the top of the valence band, absorption of circular photons close to the band gap creates electrons with a degree of spin orientation $p_e = 0.5$; when they recombine with neutral acceptors, the degree of luminescence polarization, $p_L = (I^+ - I^-)/(I^+ + I^-)$ where I^+ and I^- are the intensities of plus or minus circular components of the luminescence, has a maximum value of 0.25. This value is obtained only if the electron lifetime τ is much shorter than the electron spin relaxation time τ_s. In general, one finds $p_L = 0.25/(1 + \tau/\tau_s)$. Then, using a magnetic field H perpendicular to the excitation beam, which is also the quantization axis for the spins, one can induce spin precession at the Larmor frequency $\Omega = g^* \mu_B H/\hbar$, where g^* is the effective gyromagnetic factor of the electron and μ_B is the Bohr magneton. This causes a depolarization of the luminescence known as the *Hanle effect*, the field-dependent polarization being given as

$$p_L(H) = p_L(0)/[1 + (\Omega T_s)^2]$$

where $1/T_s = 1/\tau + 1/\tau_s$. So, from these two measurements, the lifetime τ and the spin relaxation time τ_s can be determined, provided g^* is known (Parsons, 1969); in this way, fairly short times can be measured (down to the 10^{-11}-s range) with dc method; the trick was to use the Larmor frequency Ω as an internal clock.

Orientation of *excitons* is also possible: Under appropriate conditions, a polarized photon will be transmitted at the vacuum–crystal interface as an exciton (polariton) of the *same* polarization. In particular, one can create linearly polarized excitons; when submitted to a magnetic field parallel to their wave vector, the plane of their polarization will rotate, an effect that has been called resonant Faraday effect.

In a typical *hot photoluminescence* experiment, an excitation about 0.5 eV above the band gap is used; the very weak luminescence in the region between the excitation and the main near–band-gap luminescence was at the origin of a new effect: the *optical alignment of electron momenta*. For example, pumping between the heavy-hole band and the conduction band in GaAs with a linear photon of electric polarization along z creates preferentially electrons with wave vector perpendicular to z. Here again, alignment can be demonstrated by measuring the rate of polarization of hot luminescence; alignment can be relaxed by elastic collisions and destroyed by the action of a magnetic field, which induces a cyclotron motion in $\mathbf{k}$ space and introduces a new internal clock, the cyclotron frequency.

Let us add that analogous parameters can be measured in the time-resolved mode, but at the expense of more sophisticated equipment.

3.1.6 Exciton Transport Using a sharply focused pump beam, it is possible to detect the diffusion of the excitations by measuring the diameter of the luminescence spot, or even better using a time-resolved mode. For example, it was possible to show that excitons are indeed trapped at 2 K in type-II GaAs/AlAs superlattices, but become mobile at increasing temperature.

A new kind of exciton transport experiment occurs with the use of a nonuniform stress: Thus, a gradient of band gap is created so that an exciton experiences a force proportional to this gradient. With an excitation pulsed and localized, one can follow the movement of an exciton (or electron–hole) cloud by looking at the time- and space-resolved luminescence. The so-called "phonon wind" effect has been demonstrated (Wolfe, 1985).

3.2 Excitation–Medium Interaction

The excitation introduced into the semiconductor may interact with other (intrinsic) excitations of the medium; we consider two different cases.

3.2.1 Electron–Phonon Coupling, Polarons, Self-Trapping The excitation (electron–hole pair or exciton) can induce a lattice distortion, an effect described in terms of electron–phonon coupling. The strongest coupling occurs generally with longitudinal optical (LO) phonons in polar materials. Such an LO phonon carries with it an electric field to which any charge carrier couples. This coupling, known as Fröhlich interaction, increases in the series GaAs, ZnSe, CuBr. Such coupling has two effects on the luminescence: At moderate coupling strength, sidebands appear, separated from the no-phonon line by 1, 2, 3, . . . times the LO phonon energy; this occurs for bound states, like the D^0A^0 luminescence (e.g., the well-known "green edge" luminescence in CdS) as well as for free excitons. If the coupling is strong enough, self-trapping may occur: The exciton is localized in a strongly distorted region, which acts as a potential well for the exciton. This is observed in KI and NaI; in fact, free-running excitons and self-trapped excitons can coexist because these two states are separated by a potential barrier.

3.2.2 Electron–Magnetic-Ion Exchange Interaction, Magnetic Polaron Another type of interaction of the excitation with the medium is found in magnetic semiconductors: this is the carrier–ion exchange interaction. Starting from a paramagnetic medium, like a diluted magnetic semiconductor (see SEMICONDUCTORS, DILUTED MAGNETIC), a carrier (or an exciton) can lower its energy by the orientation of the ion spins with respect to the spin of the carrier. This can take place already in the ground state, e.g., for localized carriers like neutral donors or acceptors. For example, the increase of neutral acceptor binding energy due to exchange was determined by a shift of the D^0A^0 luminescence with temperature. Such shifts affect also the recombination of excitons weakly localized by disorder, either in bulk or in low-dimensional structures based on diluted magnetic semiconductors. Here again, as in Sec. 3.2.1, if the coupling between the exciton and the ions is strong enough, self-trapping (or free magnetic polaron) may occur. It should be favored in low-dimensional structures; experiments to detect true free magnetic polarons in low-dimensional structures are going on actively.

3.3 High-Excitation Effects

Increasing the excitation level induces two kinds of effects.

3.3.1 Saturation, Optical Gain, Laser We have already shown in Sec. 3.1.3.2 an example of luminescence saturation when all donors and acceptors are neutralized: the D^0A^0 luminescence is saturated. But another effect takes place when the population n_e of the excited state is larger than the population n_g of the ground state: Photons at the energy difference can be amplified rather than absorbed; absorption is proportional to $n_g - n_e$ and gain is just a negative absorption. With band-to-band transition, population inversion can be realized near the band gap because electrons and holes are fermions. When gain occurs, the luminescence is modified in a way that depends upon the geometry of the excited region and the existence of an optical cavity that provides feedback: Laser emission may result see (LASER PHYSICS).

3.3.2 Many-body Effects: Biexcitons, Electron–Hole Plasma or Liquid When the intensity of excitation increases, especially in semiconductors, the density of electrons, holes, and excitons may increase to a point where their mutual Coulomb interaction can shift the energy of the excited states, or even produce new kinds of excited states. At low temperature, with increasing exciton density in pure materials, two excitons can form a bound state, the *exciton molecule* (or *biexciton*), in close analogy with the hydrogen molecule. These have been detected as specific luminescence lines first in direct-gap material: CuCl and later CdS; then in indirect-gap materials like AgBr and Si. At a total density of electrons and holes exceeding the Mott density, the excited medium is described as an electron–hole plasma. In such a state, the mean energy of an electron–hole pair is modified, with respect to the band gap, by a positive term, the Fermi energy of the particles, and a negative one coming from the Coulomb interaction between the particles. As a consequence, the luminescence band due to an electron–hole plasma exhibits a broadening related to the Fermi energy and a shift related to the Coulomb term, both functions of the density. In direct-gap materials, a strong optical gain is also obtained; it may distort the luminescence band and lead to laser action. In indirect-gap materials, something special may occur: because one kind of carrier occupies several equivalent valleys, there is a reduction of the Fermi energy; then the mean energy of a pair exhibits, for some density n_0, a minimum that can stay below the free exciton energy. This phase at density n_0, which is the most stable one at $T = 0$, has been called *electron–hole liquid*; its existence was predicted by Keldysh and well studied mainly using photoluminescence in the 1970s, first in Ge and Si (Fig. 7), later in GaP and SiC. At finite temperature, thermal equilibrium between the electron–hole liquid and the free-exciton gas is nothing but a new example of liquid–gas coexistence. Another kind of condensation was considered in the case of free excitons, which are bosons, being made of two fermions: this is the Bose–Einstein condensation, which is likely occurring in Cu_2O.

3.3.3 Higher-Order Emission Processes A first kind may be called partial Auger process. For example, an exciton interacting with another particle (electron or exciton) can recombine with emission of a photon, some energy being transferred to the second particle. This was well studied in direct-gap materials like CdS. Such a process is likely playing a major role in the mechanism of blue laser emission in ZnSe-based heterostructures, a domain of intense activity.

In a second kind, two electron–hole pairs can recombine with emission of a single photon. For example, a weak green luminescence at about two times the band gap was detected from electron–hole liquid and biexcitons in Si.

3.4 Practical Applications

3.4.1 Nondestructive Material Testing Photoluminescence turns out to be an easy test of a class of materials since it requires

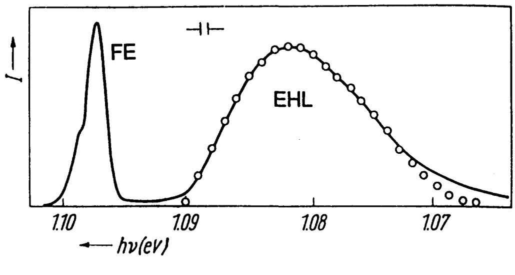

FIG. 7. Photoluminescence spectrum of pure Si at 4.2 K, showing the coexistence of free excitons (FE) and electron–hole liquid (EHL). Si being an indirect-gap material, both luminescence bands are phonon assisted with the same phonons. Open circles correspond to a theoretical fit of the EHL band.

generally no special treatment of the sample and is clearly nondestructive. It finds applications mainly with semiconductors.

In the case of bulk samples, it gives an indication about impurity content and crystalline perfection. It is especially useful for fairly pure samples, where some impurities can be identified from their specific narrow luminescence lines at low temperature. However, precise determination of impurity density is not possible, but comparison with calibrated samples may be useful. Indication about nonradiative centers may be obtained from luminescence decay.

Testing of low-dimensional structures is even more interesting. From the luminescence energy and the photoluminescence excitation spectrum, a precise determination of the energy spectrum of the structure is obtained, which is compared with a model using as inputs the nominal parameters of the growth process. Low-temperature line broadening and Stokes shift give also indications about the quality of the interfaces, to be correlated with the growth conditions.

3.4.2 Devices: Optically Pumped Solid-State Lasers and Amplifiers Whereas most practical luminescent sources use electrical excitation (e.g., semiconductor lamps or electroluminescent panels), optical excitation is used to produce optical gain in the medium, and this is applied to get light amplifiers and lasers. As an amplifier, one can cite Er^{3+}-doped silica fiber, which operates near 1.5 μm; the most efficient pump was found to be a GaAsP injection laser. Such an amplifier is now the best candidate for fiber-optics high-rate communications all around the world. Most solid-state lasers like Nd^{3+} in glass or crystals are optically pumped. Here again, the best overall efficiency is obtained using as a pump semiconductor injection lasers. In the case of a luminescent material like titanium-doped sapphire, the broad gain spectrum leads to a laser with wide tunability, from 680 to 1000 nm; in addition, since the broadening due to phonon coupling is homogeneous, very short laser pulses can be obtained, down to 100 fs.

GLOSSARY

Boxcar: Detection mode using a gate to select in a time window an electronic signal.

Brillouin Scattering: Light scattering by acoustical phonons in crystals.

Brillouin Zone: The polyhedron in reciprocal space over which the wave vector **k** is allowed to vary.

Effective Mass: In a crystal near a band extremum, the dispersion relation for an electron is a quadratic function of its wave vector **k**, written as $E = \hbar^2k^2/2m^*$. The effective mass m^* is the quantity governing the response of the electron in the crystal to an external perturbation, in the same way as m_0, the free electron mass, for an electron in vacuum.

Exchange Interaction: Interaction between particles that satisfy the Pauli exclusion principle, electrons among them.

Frenkel Exciton: In a crystal, coherent propagation of an excitation primarily localized on a molecule or an ion.

Landau Level: Orbital quantized state of a free electron in a magnetic field.

Mode-Locked Laser: By forcing all the longitudinal modes of a laser to be in phase near an end mirror, a pulsed operation is obtained, with a delay between pulses equal to the round-trip time of a photon in the laser cavity.

Modulation-Doped Heterostructure: By introducing donors in the barrier material (e.g., GaAlAs) at some distance from the interface with the well (e.g., GaAs), part of the donor electrons transfer to the well, creating a high-mobility two-dimensional electron gas.

Mott Density: At increasing electron density, a transition from a collection of isolated impurities (or exciton gas) to a metallic state (or an electron–hole plasma) takes place at the Mott density.

Phase Matching: In a nonlinear process, conditions upon the propagation angles of the different beams such that, inside the crystal, wave-vector conservation is insured.

Phonon Wind: In semiconductors, a cloud of acoustical phonons created, e.g., at the focus of an optical beam can push electrons, holes or excitons, thus contributing to an anomalous carrier diffusion.

Quantum Hall Effect: In a two-dimensional electron gas, a plateau in the Hall effect and a dip in the longitudinal resistance are observed at integer (or fractional) Landau-level filling factors.

Rayleigh Scattering: Elastic scattering of light by fixed defects or surface roughness.

Reciprocal Space (k Space): A mathematical space, with dimension of inverse length, over which wave vectors are defined.

Synchronous Detection: Detection of an electronic signal involving the product with a sinusoidal reference signal at frequency *f*. The component of the signal at frequency *f* is extracted.

Subtractive Mode: In a double-grating spectrometer, geometry that insures constant optical path length for any ray.

Wannier Exciton: In crystalline semiconductors, excitation made of an electron–hole pair tied by Coulomb interaction, free to move in the crystal.

Works Cited

Bebb, H. Barry, Williams, E. W. (1972), "Photoluminescence I: Theory," in: R. K. Willardson, A. C. Beer, (Eds.), *Semiconductors and Semimetals*, Vol. 8, New York and London: Academic Press, pp. 181–320.

Cavenett B. C. (1981), "Optical Detection of Magnetic Resonance (O. D. M. R.) investigations of recombination processes in semiconductors," *Adv. Phys.* **30**, 475–538.

Dean P. J. (1973), in: J. O. McCaldin, G. Somorjai (Eds.), *Progress in Solid State Chemistry*, Vol. 8, Oxford: Pergamon, pp. 1–126.

Parsons, R. R. (1969), "Band-to-band optical pumping in solids and polarized photoluminescence," *Phys. Rev. Lett.* **23**, 1152–1154.

Van Roosbroeck, W., Shockley, W. (1954), "Photon-radiative recombination of electrons and holes in germanium," *Phys. Rev.* **94**, 1558–1560.

Wolfe J. P. (1985), "Transport of degenerate electron-hole plasma in Si and Ge," in: M. H. Pilkuhn (Ed.), *High Excitation and Short Pulse Phenomena*, Amsterdam: North-Holland, pp. 82–113.

Further Reading

Jeffries, C. D., Keldysh, L. V. (Eds.) (1983), *Electron-Hole Droplets in Semiconductors*, Amsterdam: North-Holland.

Lumb, M. D. (Ed.) (1978), *Luminescence Spectroscopy*, New York: Academic Press.

Meier, F., Zakharchenya, B. P. (Eds.) (1984), *Optical Orientation*, Amsterdam: North Holland.

Pankove, J. I. (1971), *Optical Processes in Semiconductors*, Englewood Cliffs, NJ: Prentice Hall.

Pope, M., Swenberg, C. E. (1982), *Electronic Processes in Organic Crystals*, New York: Oxford University Press.

Rashba, E. I., Sturge, M. D. (Eds.) (1982), *Excitons*, Amsterdam: North-Holland.

PHOTOSYNTHESIS

JOHN WHITMARSH, *Agricultural Research Service/USDA, Department of Plant Biology, University of Illinois, Urbana, Illinois, U.S.A.*

GOVINDJEE, *Department of Plant Biology, University of Illinois, Urbana, Illinois, U.S.A.*

Introduction 513
1. **Brief History** 514
2. **Classification of Photosynthetic Organisms** ... 515
2.1 Oxygenic Photosynthetic Organisms 516
2.2 Anoxygenic Photosynthetic Organisms 516
3. **General Principles of Photosynthetic Energy Transformation in Plants** 516
4. **Oxygenic Photosynthesis in Plants** 518
4.1 Chloroplasts—Structure and Organization 518
4.2 Light Absorption—The Antenna System 519
4.3 Primary Photochemistry—Photosystem II and Photosystem I Reaction Centers 520
4.4 Electron Transport 522
4.5 Creation of a Proton Electrochemical Potential Across the Photosynthetic Membrane 523
4.6 Synthesis of ATP by the ATP-Synthase Enzyme 524
4.7 Synthesis of Carbohydrates from Atmospheric CO_2 by the Calvin Cycle Enzymes 524
4.8 Photosynthetic Quantum Yield and Energy Conversion Efficiency 525
5. **Oxygenic Photosynthesis in Algae** 526
6. **Oxygenic Photosynthesis in Bacteria** 526
6.1 Cyanobacteria 526
6.2 Prochlorophytes 527
7. **Anoxygenic Photosynthesis in Bacteria** 527
7.1 Purple Bacteria 527
7.2 Green Sulfur Bacteria 528
7.3 Green Gliding Bacteria 528
7.4 Heliobacteria 529
8. **Control of Intraprotein Electron Transfer** 529
9. **Global Photosynthesis and the Atmosphere** 530
Glossary 530
Works Cited 532
Further Reading 532

INTRODUCTION

Photosynthesis is the physicochemical process by which plants, algae, and photosynthetic bacteria use light energy to drive the synthesis of organic compounds. In plants, algae, and certain types of bacteria, the photosynthetic process results in the release of molecular oxygen and the removal from the atmosphere of carbon dioxide, which is used to synthesize carbohydrates. Other types of bacteria use light energy to create organic compounds but do not produce oxygen. Photosynthesis provides the energy and reduced carbon required for the survival of virtually all life on our planet, as well as the molecular oxygen necessary for the survival of oxygen-consuming organisms. In addition, the fossil fuels currently being burned to provide energy for human activity were produced by ancient photosynthetic organisms. Although photosynthesis occurs in cells or organelles that are typically only a few microns across, the process has a profound impact on the earth's atmosphere and climate. Each year more than 10% of the total atmospheric carbon dioxide is reduced to carbohydrate by photosynthetic organisms. Most, if not all, of

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

the reduced carbon is returned to the atmosphere as carbon dioxide by microbial, plant, and animal metabolism, and by biomass combustion. In turn, the performance of photosynthetic organisms depends on the earth's atmosphere and climate. Over the next century, the large increase in the amount of atmospheric carbon dioxide created by human activity is certain to have a profound impact on the performance and competition of photosynthetic organisms. Knowledge of the physicochemical process of photosynthesis is essential for understanding the relationship between living organisms and the atmosphere and the balance of life on earth.

1. BRIEF HISTORY

In the 1770s, Joseph Priestley, an English chemist and clergyman, performed experiments showing that plants release a type of air that allows combustion. He demonstrated this by burning a candle in a closed vessel until the flame went out. He placed a sprig of mint in the chamber and after several days showed that the candle could burn again. Although Priestley did not know about molecular oxygen, his work showed that plants release oxygen into the atmosphere. It is noteworthy that over 200 years later, investigating the mechanism by which plants produce oxygen is one of the most active areas of photosynthetic research. Building on the work of Priestley, Jan Ingenhousz, a Dutch physician, demonstrated that sunlight was necessary for photosynthesis and that only the green parts of plants could release oxygen. During this period, Jean Senebier, a Swiss botanist and naturalist, discovered that CO_2 is required for photosynthetic growth, and Nicolas-Théodore de Saussure, a Swiss chemist and plant physiologist, showed that water is required. It was not until 1845 that Julius Robert von Mayer, a German physician and physicist, proposed that photosynthetic organisms convert light energy into chemical free energy.

By the middle of the nineteenth century, the key features of plant photosynthesis were known, namely, that plants could use light energy to make carbohydrates from CO_2 and water. The empirical equation representing the net reaction of photosynthesis for oxygen-evolving organisms is

$$CO_2 + 2H_2O + \text{light energy} \rightarrow [CH_2O] + O_2 + H_2O, \qquad (1)$$

where $[CH_2O]$ represents a carbohydrate (e.g., glucose, a six-carbon sugar). The synthesis of carbohydrate from carbon and water requires a large input of light energy. The standard free energy for the reduction of one mole of CO_2 to the level of glucose is +478 kJ/mol. Because glucose, a six-carbon sugar, is often an intermediate product of photosynthesis, the net equation of photosynthesis is frequently written as

$$6CO_2 + 12H_2O + \text{light energy} \rightarrow C_6H_{12}O_6 + 6O_2 + 6H_2O. \qquad (2)$$

The standard free energy for the synthesis of glucose is +2870 kJ/mol.

Not surprisingly, early scientists studying photosynthesis concluded that the O_2 released by plants came from CO_2, which was thought to be split by light energy. In the 1930s, comparison of bacterial and plant photosynthesis lead Cornelis van Niel to propose the general equation of photosynthesis that applies to plants, algae, and photosynthetic bacteria (discussed by Wraight, 1982). Van Niel was aware that some photosynthetic bacteria could use hydrogen sulfide (H_2S) instead of water for photosynthesis and that these organisms released sulfur instead of oxygen. Van Niel, among others, concluded that photosynthesis depends on electron donation and acceptor reactions and that the O_2 released during photosynthesis comes from the oxidation of water. Van Niel's generalized equation is

$$CO_2 + 2H_2A + \text{light energy} \rightarrow [CH_2O] + 2A + H_2O. \qquad (3)$$

In oxygenic photosynthesis, $2A$ is O_2, whereas in anoxygenic photosynthesis, which occurs in some photosynthetic bacteria, the electron donor can be an inorganic hydrogen donor, such as H_2S in which case A is elemental sulfur, or an organic hydrogen donor such as succinate in which case A is fumarate.

The biochemical conversion of CO_2 to carbohydrate is a reduction reaction that involves the rearrangement of covalent bonds between carbon, hydrogen, and oxygen. The energy for the reduction of carbon is provided by energy-rich molecules that are produced by the light-driven electron-transfer reactions. Carbon reduction can occur in the dark and involves a series of biochemical reactions that were elucidated by Melvin Calvin, Andrew Benson, and James Bassham in the late 1940s and 1950s. Using the radioisotope carbon-14, most of the intermediate steps that result in the production of carbohydrate were identified. Calvin was awarded the Nobel Prize for Chemistry in 1961 for this work.

In 1954, Daniel Arnon and co-workers discovered that plants, and A. Frenkel discovered that photosynthetic bacteria, use light energy to produce ATP, an organic molecule that serves as an energy source for many biochemical reactions. During the same period, L. N. M. Duysens showed that the primary photochemical reaction of photosynthesis is an oxidation/reduction reaction that occurs in a protein complex (the reaction center). Over the next few years, the work of several groups, including those of Robert Emerson, Bessel Kok, L. N. M. Duysens, Robert Hill, and Horst Witt, combined to prove that plant, algae, and cyanobacteria require two reaction centers, photosystem II and photosystem I, operating in series.

In 1961, Peter Mitchell suggested that cells can store energy by creating an electric field or a proton gradient across a membrane. Mitchell's proposal that energy is stored as an electrochemical gradient across a vesicular membrane opened the door for understanding energy transformation by membrane systems. He was awarded the Nobel Prize in Chemistry in 1978 for his theory of chemiosmotic energy transduction.

Most of the proteins required for the conversion of light energy and electron transfer reactions of photosynthesis are located in membranes. Despite decades of work, efforts to determine the structure of membrane-bound proteins had little success. This changed in the 1980s when Johann Deisenhofer, Hartmut Michel, and Robert Huber determined the structure of the reaction center of the purple bacterium *Rhodospeudomonas viridis*. They were awarded the Nobel Prize for Chemistry in 1988 for their work, which provided insight into the relationship between structure and function in membrane-bound proteins.

A key element in photosynthetic energy conversion is electron transfer within and between protein complexes and simple organic molecules. The electron-transfer reactions are rapid (as fast as a few picoseconds) and highly specific. Much of our current understanding of the physical principles that guide electron transfer is based on the pioneering work of Rudolph A. Marcus, who received the Nobel Prize in Chemistry in 1992 for his contributions to the theory of electron-transfer reaction in chemical systems.

The overall equation for photosynthesis is deceptively simple. In fact, a complex set of physical and chemical reactions must occur in a coordinated manner for the synthesis of carbohydrates. To produce a sugar molecule such as sucrose, plants require nearly 30 distinct proteins that work within a complicated membrane structure. Research into the mechanism of photosynthesis centers on understanding the structure of the photosynthetic components and the molecular processes that use radiant energy to drive carbohydrate synthesis. The research involves several disciplines, including physics, chemistry, structural biology, biochemistry, molecular biology, and physiology, and serves as an outstanding example of the success of multidisciplinary research. As such, photosynthesis presents a special challenge in understanding several interrelated molecular processes. From a physicist's viewpoint, the reactions that transform energy are of particular interest, and it is on these processes that this description will focus.

2. CLASSIFICATION OF PHOTOSYNTHETIC ORGANISMS

All life can be divided into three domains, Archaea, Bacteria, and Eucarya, which originated from a common ancestor (Woese *et al.*, 1990). Historically, the term photosynthesis has been applied to organisms that depend on chlorophyll (or bacteriochlorophyll) for the conversion of light energy into chemical free energy (Gest, 1993). These include organisms in the domains Bacteria (photosynthetic bacteria) and Eucarya (algae and higher plants). The most primitive domain, Archaea, includes organisms known as halobacteria, which convert light energy into chemical free energy. However, the mechanism by which halobacteria convert light is fundamentally different from that of higher organisms because there is no oxidation/reduction chemistry and halobacteria cannot use CO_2 as their carbon source. Consequently, some biologists do not consider halobacteria as photosynthetic (Gest, 1993). This article will follow the historical definition of photosynthesis and omit halobacteria.

2.1 Oxygenic Photosynthetic Organisms

The photosynthetic process in all plants and algae, as well as in certain types of photosynthetic bacteria, involves the reduction of CO_2 to carbohydrate and the removal of electrons from H_2O, which results in the release of O_2. In this process, known as oxygenic photosynthesis, water is oxidized by the photosystem II reaction center, a multisubunit protein located in the photosynthetic membrane. Years of research have shown that structure and function of photosystem II is similar in plants, algae, and certain bacteria, so that knowledge gained in one species can be applied to others. This homology is a common feature of proteins that perform the same reaction in different species. The importance of this homology at the molecular level is shown by the fact that there are an estimated 300 000–500 000 species of plants. If different species had evolved diverse mechanisms for oxidizing water, research aimed at a general understanding of photosynthetic water oxidation would be hopeless.

2.2 Anoxygenic Photosynthetic Organisms

Some photosynthetic bacteria can use light energy to extract electrons from molecules other than water. These organisms are of ancient origin, presumed to have evolved before oxygenic photosynthetic organisms. Anoxygenic photosynthetic organisms occur in the domain Bacteria and have representatives in four phyla: Purple Bacteria, Green Sulfur Bacteria, Green Gliding Bacteria, and Gram Positive Bacteria.

3. GENERAL PRINCIPLES OF PHOTOSYNTHETIC ENERGY TRANSFORMATION IN PLANTS

The energy that drives photosynthesis originates in the center of the sun, where mass is converted to heat by the fusion of hydrogen. Over time, the heat energy reaches the sun's surface, where some of it is converted to light by blackbody radiation that reaches the earth. A small fraction of the visible light incident on the earth is absorbed by plants. Through a series of energy-transducing reactions, plants are able to transform light energy into chemical free energy in a stable form that can last for hundreds of millions of years (e.g., fossil fuels). A simplified scheme describing how plants transform energy is presented in this section. The focus is on the structural and functional features essential for the energy-transforming reactions. For clarity, mechanistic and structural details are omitted. A more highly resolved description of oxygenic and anoxygenic photosynthesis is given in the remaining sections.

The photosynthetic process in plants and algae occurs in small organelles, known as chloroplasts, that are located inside cells. The photosynthetic reactions are traditionally divided into two types: the "light reactions," which consist of electron- and proton-transfer reactions, and the "dark reactions," which consist of the biosynthesis of carbohydrates from CO_2. The light reactions occur in a complex membrane system (the photosynthetic membrane) that is made up of protein complexes, electron carriers, and lipid molecules. The photosynthetic membrane is surrounded by water and can be thought of as a two-dimensional surface that defines a closed space, with an inner and an outer water phase. A molecule or ion must pass through the photosynthetic membrane to go from the inner space to the outer space. The protein complexes embedded in the photosynthetic membrane have a unique orientation with respect to the inner and outer phases. The asymmetrical arrangement of the protein complexes allows some of the energy released during electron transport to create an electrochemical gradient of protons across the photosynthetic membrane.

Photosynthetic electron transport consists of a series of individual electron-transfer steps from one electron carrier to another. The electron carriers are metal ion complexes and aromatic groups. The metal ion complexes and most of the aromatic groups are bound within proteins. Most of the proteins involved in photosynthetic electron transport are composed of numerous polypeptide chains that lace through the membrane, providing a scaffolding for metal ions and aromatic groups. An electron enters a protein complex at a specific site, is transferred within the protein from one carrier to another, and exits from the protein at a different site. The protein controls the pathway of electrons between the carriers by determining the location and environment of the metal ion complexes and aromatic groups. By setting the distance between electron carriers and controlling the electronic environment surrounding a metal ion complex or an aromatic group, the protein controls pairwise electron transfer reactions. Between proteins, electron transfer is controlled by distance and free energy, as for intraprotein transfer, and by the probability that the two proteins are in close contact. Protein association is controlled by a number of factors, including the structure of the two proteins, their surface electrical and chemical properties, and the probability that they collide with one another. Not all electron carriers are bound to proteins. The reduced forms of plastoquinone and nicotinamide adenine dinucleotide phosphate (NADPH) act as mobile electron carriers operating between protein complexes. For electron transfer to occur, these small molecules must bind to special pockets in the proteins known as binding sites. The binding sites are highly specific and are a critical factor in controlling the rate and pathway of electron transfer.

The light reactions convert energy into several forms (Fig. 1). The first step is the conversion of a photon to an excited electronic state of an antenna pigment molecule located in the antenna system. The antenna system consists of hundreds of pigment molecules (mainly chlorophyll and carotenoids) that are anchored to proteins within the photosynthetic membrane and serve a specialized protein complex known as a reaction center. The electronic excited state is transferred over the antenna molecules as an exciton. Some excitons are converted back into photons and emitted as fluorescence, some are converted to heat, and some are trapped by a reaction-center protein. Excitons trapped by a reaction center provide the energy for the primary photochemical reaction of photosynthesis—the transfer of an electron from a donor molecule to an acceptor molecule. Both the donor and acceptor molecules are attached to the reaction-center protein complex. Once primary charge separation occurs, the subsequent electron-transfer reactions are energetically downhill. In oxygenic photosynthetic organisms, two different reaction centers work in series: photosystem II and photosystem I. Electrons are transferred from photosystem II to the photosystem I reaction center by intermediate carriers. The net reaction is the transfer of electrons from a water molecule to NADP$^+$, producing the reduced form, NADPH. In the photosynthetic process, much of the energy initially provided by light energy is stored as redox free energy (a form of chemical free energy) in NADPH, to be used later in the reduction of carbon. In addition, the electron-transfer reactions concentrate protons inside the membrane vesicle and create an electric field across the photosynthetic membrane. In this process, the electron-transfer reactions convert redox free energy into an electrochemical po-

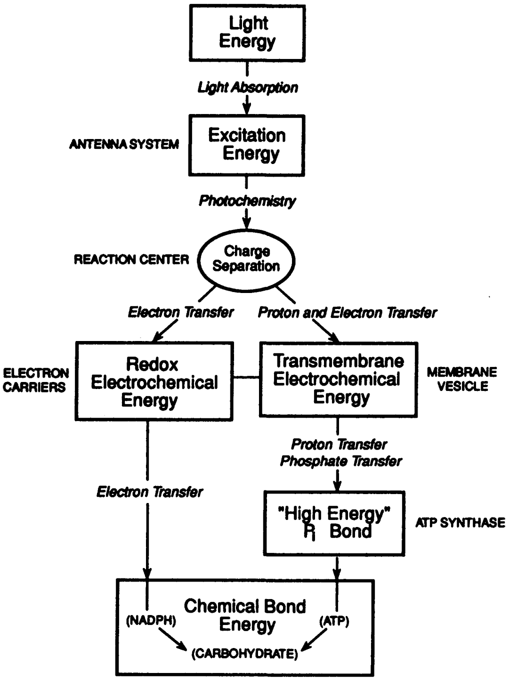

FIG. 1. Photosynthesis shown as a series of reactions that transform energy from one form to another. The different forms of energy are shown in boxes, and the direction of energy transformation is shown by the arrows. The energy-transforming reaction is shown by italics in the arrows. The site at which the energy is stored is shown in capital letters outside the boxes. The primary photochemical reaction, charge separation, is shown in the oval. Details of these reactions are given in the text.

tential of protons. The energy stored in the proton electrochemical potential is used by a membrane-bound protein complex (ATP synthase) to attach a phosphate group covalently to adenosine diphosphate (ADP), forming adenosine triphosphate (ATP). Protons pass through the ATP-synthase protein complex that transforms electrochemical free energy into a type of chemical free energy known as phosphate group-transfer potential (or a high-energy phosphate bond) (Klotz, 1967). The energy stored in ATP can be transferred to another molecule by transferring the phosphate group. The net effect of the light reactions is to convert radiant energy into redox free energy in the form of NADPH and phosphate group-transfer energy in the form of ATP. In the light reactions, the transfer of a single electron from water to $NADP^+$ involves about 30 metal ions and seven aromatic groups. The metal ions include 19 Fe, 5 Mg, 4 Mn, and 1 Cu. The aromatics include quinones, pheophytin, NADPH, tyrosine, and a flavoprotein.

The NADPH and ATP formed by the light reactions provide the energy for the dark reactions of photosynthesis, known as the Calvin cycle or the photosynthetic carbon reduction cycle. The reduction of atmospheric CO_2 to carbohydrate occurs in the aqueous phase of the chloroplast and involves a series of enzymatic reactions. The first step is catalyzed by the protein Rubisco (D-ribulose 1,5-bisphosphate carboxylase/oxygenase), which attaches CO_2 to a five-carbon compound. The reaction produces two molecules of a three-carbon compound. Subsequent biochemical reactions involve several enzymes that reduce carbon by hydrogen transfer and rearrange the carbon compounds to synthesize carbohydrates. The carbon reduction cycle involves the transfer and rearrangement of chemical bond energy.

4. OXYGENIC PHOTOSYNTHESIS IN PLANTS

4.1 Chloroplasts—Structure and Organization

Photosynthesis occurs inside chloroplasts, which are small organelles found in plant cells. Chloroplasts provide the energy and reduced carbon needed for plant growth and development, while the plant provides the chloroplast with CO_2, water, nitrogen, organic molecules, and minerals necessary for the chloroplast biogenesis. Most chloroplasts are located in specialized leaf cells, which often contain 50 or more chloroplasts per cell. Each chloroplast is defined by an inner and an outer envelope membrane and is shaped like a meniscus convex lens that is 5–10 μm in diameter (Fig. 2), although many different shapes and sizes can be found in plants. The inner envelope membrane acts as a barrier, controlling the flux of organic and charged molecules in and out of the chloroplast. Water passes freely through the envelope membranes, as do other small neutral molecules like CO_2 and O_2. There is convincing evidence that chloroplasts were once free-living bacteria that invaded a nonphotosynthetic cell long ago. They have retained some of the DNA necessary for their assembly, but much of the DNA necessary for their biosynthesis is located in the cell nucleus. This enables a cell to control the biosynthesis of chloroplasts within its domain.

Inside the chloroplast is a complicated membrane system, known as the photosynthetic membrane (or thylakoid membrane), that contains most of the proteins required for the light reactions. The proteins required for the fixation and reduction of CO_2 are located outside the photosynthetic membrane in the surrounding aqueous phase. The photosynthetic membrane is composed mainly of glycerol lipids and protein. The glycerol lipids are a family of molecules characterized by a polar head group that is hydrophilic and two fatty acid side chains that are hydrophobic. In membranes, the lipid mol-

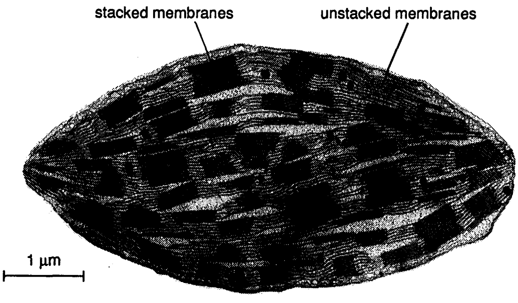

FIG. 2. An electron micrograph of a plant chloroplast. Micrograph by A. D. Greenwood, courtesy of J. Barber.

ecules arrange themselves in a bilayer, with the polar head toward the water phase and the fatty acid chains aligned inside the membrane forming a hydrophobic core (Fig. 3). The photosynthetic membrane is vesicular, defining a closed space with an outer water space (stroma) and an inner water space (lumen). The organization of the photosynthetic membrane can be described as groups of stacked membranes (like stacks of pita bread, with the inner pocket representing the inner aqueous space), interconnected by nonstacked membranes that protrude from the edges of the stacks (Staehelin, 1986). Experiments indicate that the inner aqueous space of the photosynthetic membrane is likely continuous inside of the chloroplast. It is not known why the photosynthetic membrane forms such a convoluted structure. To understand the energetics of photosynthesis, the complicated structure can be ignored and the photosynthetic membrane can be viewed as a simple vesicle.

4.2 Light Absorption—The Antenna System

Plant photosynthesis is driven primarily by visible light (wavelengths from 400 to 700 nm) that is absorbed by pigment molecules (mainly chlorophyll *a* and *b* and carotenoids). Plants appear green because of chlorophyll, which is so plentiful that regions of the earth appear green from space. The absorption spectra of chlorophyll *a* and *b* and of a chloroplast are shown in the article EXCITONS. Light is collected by 200–300 pigment molecules that are bound to light-harvesting protein complexes located in the photosynthetic membrane. The three-dimensional structure of the light-harvesting complex (Kühlbrandt *et al.*, 1994) shows that the protein determines the position and orientation of the antenna pigments. The light-harvesting complexes surround the reaction centers and serve as antennae. Photosynthesis is initiated by the absorption of a photon by an antenna molecule, which occurs in

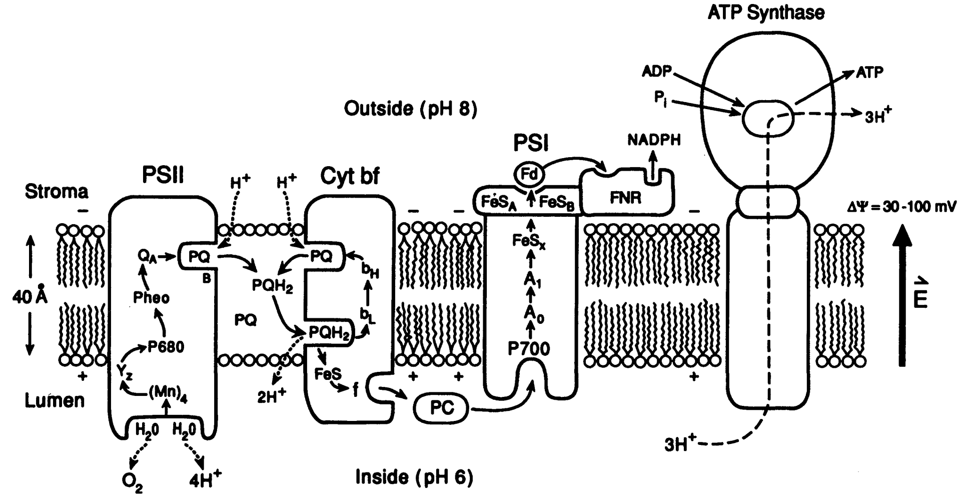

FIG. 3. Model of the photosynthetic membrane of plants showing the electron-transport components and the ATP-synthase enzyme (cross-sectional view). The complete membrane forms a vesicle. The pathways of electrons are shown by solid arrows. The membrane-bound electron-transport protein complexes involved in transferring electrons are the photosystem II and I reaction centers (PSII and PSI) and the cytochrome *bf* complex (Cyt *bf*). Abbreviations: Y_z, tyrosine; P680 and P700, the reaction-center chlorophyll of photosystem II and photosystem I, respectively; Pheo, pheophytin; Q_A, bound plastoquinone; PQ, free plastoquinone (oxidized form), PQH_2, free plastoquinone (reduced form); b_L and b_H, different forms of *b*-type cytochromes; FeS, iron-sulfur centers; *f*, cytochrome *f*; PC, plastocyanin; A_0, chlorophyll; A_1, phylloquinone; Fd, ferredoxin; FNR, ferredoxin/NADP$^+$ oxidoreductase; NADPH, nicotinomide adenine dinucleotide phosphate (reduced form); ADP, adenosine diphosphate; ATP, adenosine triphosphate; P_i, inorganic phosphate; H^+, protons; $\Delta\Psi$, the light-induced electrical potential across the membrane. The light-harvesting protein complexes are not shown. Details are given in the text.

10^{-15} s and causes a transition from the electronic ground state to an excited state. Within 10^{-13} s, the excited state decays by vibrational relaxation to the first excited singlet state. The fate of the excited-state energy is guided by the structure of the protein. Because of the proximity of other antenna molecules with the same or similar energy states, the excited-state energy has a high probability of being transferred by resonance energy transfer to a near neighbor. Exciton energy transfer between antenna molecules is due to the interaction of the transition dipole moment of the molecules. The probability of transfer is dependent on the distance between the transition dipoles of the donor and acceptor molecules ($1/R^6$), the relative orientation of the transition dipoles, and the overlap of the emission spectrum of the donor molecule with the absorption spectrum of the acceptor molecule (for a discussion of resonance energy transfer, see EXCITONS). Photosynthetic antenna systems are very efficient at this process. Under optimum conditions, over 90% of the absorbed quanta are transferred within a few hundred picoseconds from the antenna system to the reaction center, which acts as a trap for the excitons.

4.3 Primary Photochemistry—Photosystem II and Photosystem I Reaction Centers

Photosystem II uses light energy to drive two chemical reactions: the oxidation of water and the reduction of plastoquinone. The photosystem II complex is composed of more than 15 polypeptides, and at least nine different redox components (chlorophyll, pheophytin, plastoquinone, tyrosine, Mn, Fe, cytochrome b559, carotenoid, and histidine) have been shown to undergo light-induced electron transfer (Debus, 1992). However, only five of these redox components are known to be involved in transferring electrons from H_2O to the plastoquinone pool—the water-oxidizing manganese cluster $(Mn)_4$, the amino acid tyrosine, the reaction center chlorophyll (P680), pheophytin, and the plastoquinone molecules, Q_A and Q_B. Of these essential redox components, tyrosine, P680, pheophytin, Q_A, and Q_B have been shown to be bound to two key polypeptides that form the heterodimeric reaction center core of photosystem II (*D*1 and *D*2). Recent work indicates that the *D*1 and *D*2 polypeptides also provide ligands for the $(Mn)_4$ cluster. The three-dimensional structure of photosystem II is not known. Our knowledge of its structure is guided by the known structure of the reaction center in purple bacteria and biochemical and spectroscopic data. Figure 4 shows a schematic view of photosystem II that is consistent with current data.

Photochemistry in photosystem II is initiated by charge separation between P680 and pheophytin, creating $P680^+/Pheo^-$. Primary charge separation takes about 3 ps (Fig. 5). Subsequent electron-transfer steps have been designed through evolution to prevent the primary charge separation from recombining. This is accomplished by transferring the electron within 200 ps from pheophytin to a plastoquinone molecule (Q_A) that is permanently bound to photosystem II. Although plastoquinone normally acts as a two-electron acceptor, it works as a one-electron acceptor at the Q_A site. The electron on Q_A- is then transferred to another plastoquinone molecule that is loosely bound at the Q_B site. Plastoquinone at the Q_B site differs from Q_A in that it works as a two-electron acceptor, becoming fully reduced and protonated after two photochemical turnovers of the reaction center. The full reduction of plastoquinone requires the addition of two electrons and two

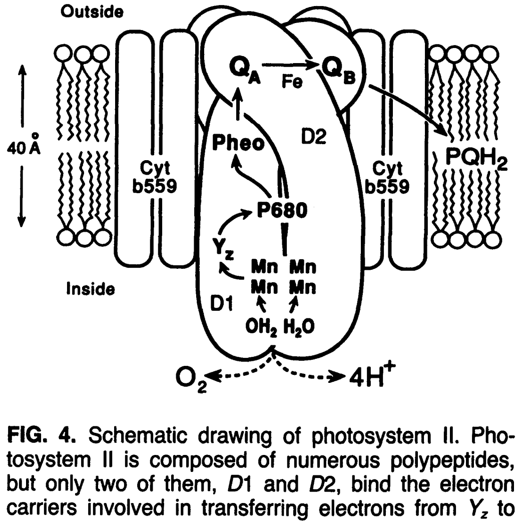

FIG. 4. Schematic drawing of photosystem II. Photosystem II is composed of numerous polypeptides, but only two of them, *D*1 and *D*2, bind the electron carriers involved in transferring electrons from Y_z to plastoquinone. Abbreviations: Y_z, tyrosine; P680, reaction-center chlorophyll (primary electron donor); Pheo, pheophytin; Q_A and Q_B, bound plastoquinone; PQH_2, reduced plastoquinone; Cyt *b*559, *b*-type cytochrome. Details are given in the text.

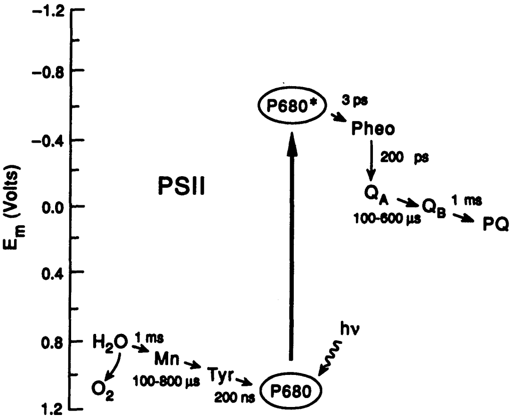

FIG. 5. Photosystem II electron-transport pathways and rates. The vertical axis shows the midpoint potential of the electron carriers. The heavy vertical arrow shows light absorption. P680* is the electronically excited state of P680. The abbreviations are given in the legend of Fig. 4.

protons, i.e., the addition of two hydrogen atoms (Fig. 6). The reduced plastoquinone then debinds from the reaction center and diffuses into the hydrophobic core of the membrane, then an oxidized plastoquinone molecule finds its way to the Q_B-binding site, and the process is repeated. Because the Q_B site is near the outer aqueous phase, the protons added to plastoquinone during its reduction are taken from the outside of the membrane.

Photosystem II is the only known protein complex that can oxidize water, resulting in the release of O_2 into the atmosphere. Despite years of research, little is known about the molecular events that lead to water oxidation. Energetically, water is a poor electron donor. The oxidation-reduction midpoint potential ($E_{m,7}$) of water is +0.82 V (pH 7). In photosystem II, this reaction is driven by the oxidized reaction center, P680$^+$ (the midpoint potential of P680/P680$^+$ is estimated to be +1.2 V at pH 7). How electrons are transferred from water to P680$^+$ remains a mystery (Govindjee and Coleman, 1990). It is known that P680$^+$ oxidizes a tyrosine on the $D1$ protein and that Mn plays a key role in water oxidation. Four Mn ions are present in the water-oxidizing complex. X-ray absorption spectroscopy shows that Mn undergoes light-induced oxidation. Water oxidation requires two molecules of water and involves four sequential turnovers of the reaction center. Each photochemical reaction creates an oxidant that removes one electron. The net reaction results in the release of one O_2 molecule, the deposition of four protons into the inner water phase, and the transfer of four electrons to the Q_B site (producing two reduced plastoquinone molecules).

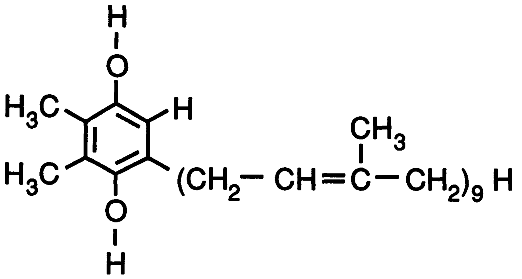

FIG. 6. Structure of plastoquinone (reduced form), an aromatic molecule that carries electrons and protons in photosynthetic electron transport.

The photosystem I complex catalyzes the oxidation of plastocyanin, a small soluble Cu protein, and the reduction of ferredoxin, a small FeS protein (Fig. 7). Photosystem I is

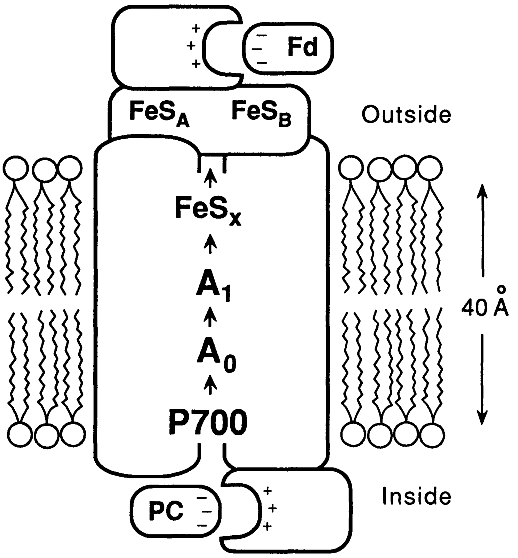

FIG. 7. Schematic drawing of photosystem I. Photosystem I is composed of numerous polypeptides, but only three of them bind the electron carriers. Abbreviations: PC, plastocyanin; P700, reaction-center chlorophyll (primary electron donor); A_0, chlorophyll; A_1, phylloquinone; FeS, FeS centers; Fd, ferredoxin. Details are given in the text.

composed of a heterodimer of proteins that act as ligands for most of the electron carriers (Krauss *et al.*, 1993). The reaction center is served by an antenna system that consists of about two hundred chlorophyll molecules (mainly chlorophyll *a*), and primary photochemistry is initiated by a chlorophyll *a* dimer, P700. In contrast to photosystem II, many of the antenna chlorophyll molecules in photosystem I are bound to the reaction-center proteins. Also, FeS centers serve as electron carriers in photosystem I, and so far as is known, photosystem I electron transfer is not coupled to proton translocation. Primary charge separation occurs between a primary donor, P700, a chlorophyll dimer, and a chlorophyll monomer (A_0). The subsequent electron transfer events and rates are shown in Fig. 8.

4.4 Electron Transport

Electron transport from water to NADP$^+$ requires three membrane-bound protein complexes operating in series: photosystem II, the cytochrome bf complex, and photosystem I (Fig. 3). Electrons are transferred between these large protein complexes by small mobile molecules. Because these small molecules carry electrons (or hydrogen atoms) over relatively long distances, they play a unique role in photosynthetic energy conversion. This is illustrated by plastoquinone (PQ), which serves two key functions. Plastoquinone transfers electrons from the photosystem II reaction center to the cytochrome *bf* complex and carries protons across the photosynthetic membrane. It does this by shuttling hydrogen atoms across the membrane from photosystem II to the cytochrome *bf* complex. Because plastoquinone is hydrophobic, its movement is restricted to the hydrophobic core of the photosynthetic membrane. Plastoquinone operates by diffusing through the membrane until, as a result of random collisions, it becomes bound to a specific site on the photosystem II complex. The photosystem II reaction center reduces plastoquinone at the Q_B site by adding two electrons and two protons, creating PQH_2. The reduced plastoqinone molecule debinds from photosystem II and diffuses randomly in the photosynthetic membrane until it encounters a specific binding site on the cytochrome *bf* complex. The cytochrome *bf* complex is a membrane-bound protein complex that contains four electron carriers, three cytochromes, and an FeS center. In a complicated reaction sequence that is not fully understood, the cytochrome *bf* complex removes the electrons from reduced plastoquinone and facilitates the release of the protons into the inner aqueous space. The electrons are eventually transferred to the photosystem I reaction center. The protons released into the inner aqueous space contribute to the proton chemical free energy across the membrane.

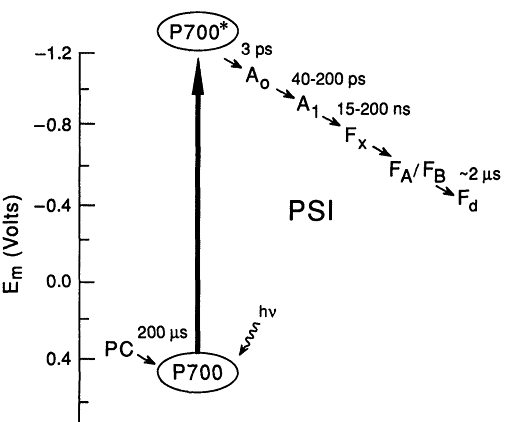

FIG. 8. Photosystem I electron transport pathways and rates. The vertical axis shows the midpoint potential of the electron carriers. Abbreviations are given in the legend of Fig. 7.

Electron transfer from the cytochrome *bf* complex to photosystem I is mediated by a small Cu protein, plastocyanin (PC). Plastocyanin is water soluble and operates in the inner water space of the photosynthetic membrane. Electron transfer from photosystem I to NADP$^+$ requires ferredoxin, a small FeS protein, and ferredoxin-NADP oxidoreductase, a peripheral flavoprotein that operates on the outer surface of the photosynthetic membrane. Ferredoxin and NADP$^+$ are water soluble and are found in the outer aqueous phase.

As discussed in Sec. 3, the pathway of electrons is largely determined by the energetics of the reaction and the distance between the carriers. The electron affinity of the carriers is represented in Fig. 9 by their midpoint potentials, which show the free energy available for electron-transfer reactions under equilibrium conditions. (It should be kept

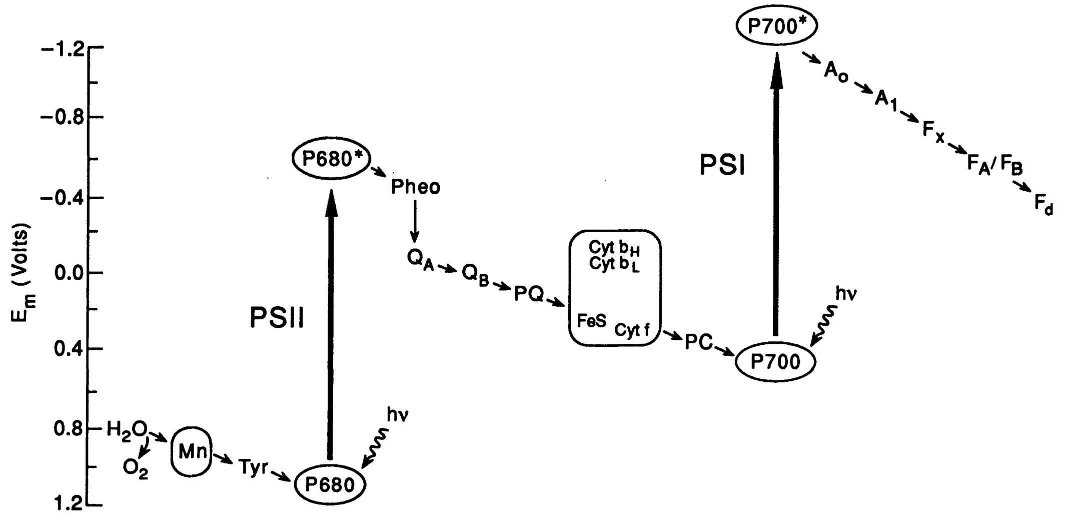

FIG. 9. The electron transport pathway of plants (oxygenic photosynthesis). Abbreviations are given the legend of Fig. 3. Details are given in the text.

in mind that reaction conditions during photosynthesis are not in equilibrium.) Subsequent to primary charge separation, electron transport is energetically downhill (from a lower to a higher redox potential). It is the downhill flow of electrons that provides free energy for the creation of a proton chemical gradient.

Photosynthetic membranes effectively limit electron transport to two dimensions. For mobile electron carriers, limiting diffusion to two dimensions increases the number of random encounters (Whitmarsh, 1986). Furthermore, because plastocyanin is mobile, any one cytochrome *bf* complex can interact with a number of photosystem I complexes. The same is true for plastoquinone, which commonly operates at a stoichiometry of about six molecules per photosystem II complex.

4.5 Creation of a Proton Electrochemical Potential

Electron transport creates the proton electrochemical potential of the photosynthetic membrane by two types of reactions.

1. The release of protons during the oxidation of water by photosystem II and the translocation of protons from the outer aqueous phase to the inner aqueous phase by the coupled reactions of photosystem II and the cytochrome *bf* complex in reducing and oxidizing plastoquinone on opposite sides of the membrane creates a concentration difference of protons across the membranes ($\Delta p\mathrm{H} = p\mathrm{H}_{\mathrm{in}} - p\mathrm{H}_{\mathrm{out}}$).
2. Primary charge separation at a reaction center drives an electron across the photosynthetic membrane, which creates an electric potential across the membrane ($\Delta\Psi = \Psi_{\mathrm{in}} - \Psi_{\mathrm{out}}$).

Together, these two forms of energy make up the proton electrochemical potential across the photosynthetic membrane ($\Delta\mu_{\mathrm{H}^+}$), which is related to the pH difference across the membrane and the electrical potential difference across the membrane by the following equation:

$$\Delta\mu_{\mathrm{H}^+} = F\Delta\Psi - 2.3\mathrm{R}T\,\Delta p\mathrm{H},$$

where F is the Faraday constant, R is the gas constant, and T is the temperature in kelvins. Although the value of $\Delta\Psi$ across the photosynthetic membrane in chloroplasts can be as large as 100 mV, under normal conditions the proton gradient dominates. For example, during photosynthesis, the outer pH is typically near 8 and the inner pH is typically near 6, giving a pH difference of 2 across the membrane that is equivalent to 120 mV. Under these conditions, the free energy for proton transfer from the inner to the outer aqueous phase is −12 kJ/mol of protons.

4.6 Synthesis of ATP by the ATP-Synthase Enzyme

The conversion of proton electrochemical energy into chemical free energy is accomplished by a single protein complex known as ATP synthase. This enzyme catalyzes a phosphorylation reaction, which is the formation of ATP by the addition of inorganic phosphate (P_i) to ADP:

$$ADP^{-3} + P_i^{-2} + H^+ \rightarrow ATP^{-4} + H_2O.$$

The reaction is energetically uphill (ΔG = +32 kJ/mol) and is driven by proton transfer through the ATP-synthase protein. The ATP-synthase complex is composed of two major subunits, CF_0 and CF_1 (Fig. 10). The CF_0 subunit spans the photosynthetic membrane and forms a proton channel through the membrane. The CF_1 subunit is attached to the top of the CF_0 on the outside of the membrane and is located in the aqueous space. The CF_1 subunit is composed of several different protein subunits, referred to as α, β, γ, δ, and ϵ. The top portion of the CF_1 subunit is composed of three $\alpha\beta$ dimers that contain the catalytic sites for ATP synthesis. The molecular processes that couple proton transfer through the protein to the chemical addition of phosphate to ADP are poorly understood. It is known that phosphorylation can be driven by a *p*H gradient, a transmembrane electric field, or a combination of the two. Experiments indicate that three protons must pass through the ATP-synthase complex for the synthesis of one molecule of ATP. However, the protons are not involved in the chemistry of adding phosphate to ADP. Paul Boyer and co-workers have proposed an alternating binding site mechanism for ATP synthesis (Boyer, 1993). One model based on their proposal is that there are three catalytic sites on each CF_1 that cycle among three different states. The states differ in their affinity for ADP, P_i, and ATP. At any one time, each site is in a different state. Initially, one catalytic site on CF_1 binds one ADP and one inorganic phosphate molecule relatively loosely. Through a conformational change of the protein, the site becomes a tight binding site, which stabilizes ATP. Next, proton transfer induces an alteration in protein conformation that causes the site to release the ATP molecule into the aqueous phase. In this model, the energy from the proton electrochemical gradient is used to lower the affinity of the site for ATP, allowing its release to the water phase. The three sites on CF_1 act cooperatively; i.e., the conformational states of the sites are linked. It has been proposed that protons affect the conformational change by driving the rotation of the top part (the three $\alpha\beta$ dimers) of CF_1. This revolving-site mechanism would require rates as high as 100 rev/s. It is worth noting that flagella that propel some bacteria are driven by a proton pump and can rotate at 60 rev/s.

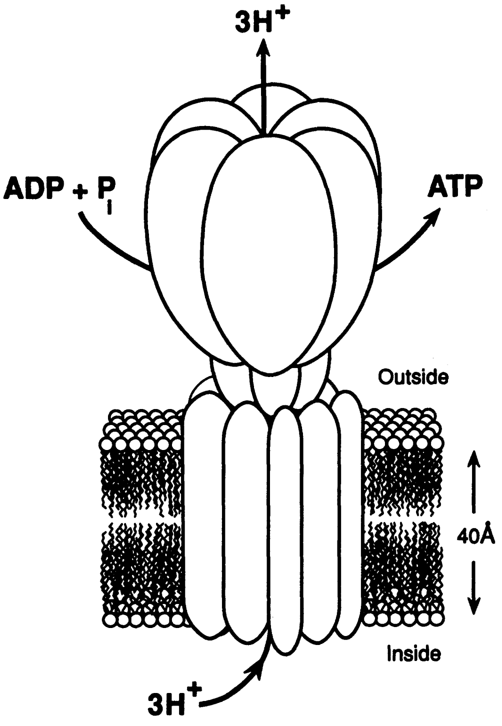

FIG. 10. Schematic drawing of the ATP-synthase enzyme embedded in the membrane. Proton transfer through the ATP synthase provides the energy for the creation of ATP from ADP and P_i. Abbreviations are given in the legend of Fig. 3. Details are given in the text.

4.7 Synthesis of Carbohydrates from Atmospheric CO_2 by the Calvin-Cycle Enzymes

All plants and algae remove CO_2 from the environment and reduce it to carbohydrate by the Calvin cycle. The process is a sequence of biochemical reactions that reduce carbon and rearrange bonds to produce carbohydrate from CO_2 molecules. The first step is the addition of CO_2 to a five-carbon compound (ribulose 1,5-bisphosphate) (Fig. 11). The six-carbon compound is split, giving two molecules of a three-carbon compound (3-phosphoglycerate). This key reaction is catalyzed by Rubisco, a large water-soluble protein complex. The carboxylation reaction is energetically downhill. The main energy input in the Calvin cycle is the phosphorylation by ATP and subsequent reduction by NADPH of the initial three-carbon compound forming a three-carbon sugar, triose phosphate. Some of the triose phosphate is exported from the chloroplast and provides the building block for synthesizing more complex molecules. In a process known as regeneration, the Calvin cycle uses some of the triose phosphate molecules to synthesize the energy-rich ribulose 1,5-bisphosphate needed for the initial carboxylation reaction. This reaction requires the input of energy in the form of one ATP. Overall, 13 enzymes are required to catalyze the reactions in the Calvin cycle. The energy-conversion efficiency of the Calvin cycle is approximately 90%. The reactions do not involve energy transduction but rather the rearrangement of chemical energy. Each molecule of CO_2 reduced to a sugar $[CH_2O]_n$ requires two molecules of NADPH and three molecules of ATP.

Rubisco is a bifunctional enzyme that, in addition to binding CO_2 to ribulose bisphosphate, can also bind O_2. This oxygenation reaction produces the 3-phosphoglycerate that is used in the Calvin cycle and a two-carbon compound (2-phosphoglycolate) that is not useful for the plant. In response, a complicated set of reactions (known as photorespiration) are initiated that serve to recover reduced carbon and to remove phosphoglycolate. The Rubisco oxygenation reaction appears to serve no useful purpose for the plant. Some plants have evolved specialized structures and biochemical pathways that concentrate CO_2 near Rubisco, which serves to decrease the fraction of oxygenation reactions.

4.8 Photosynthetic Quantum Yield and Energy Conversion Efficiency

The theoretical minimum quantum requirement for photosynthesis is eight quanta for each molecule of oxygen evolved (four quanta required by photosystem II and four by photosystem I). Measurements in leaves under optimal conditions (e.g., low light) give

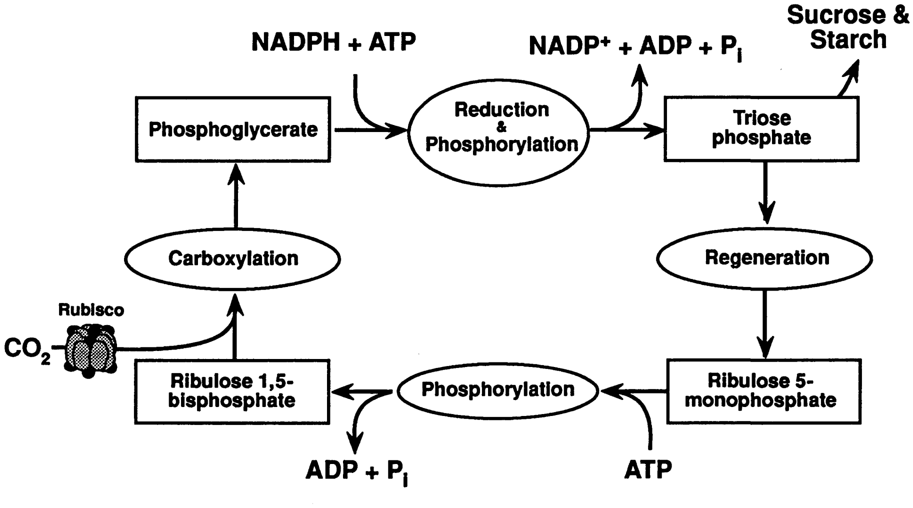

FIG. 11. An abbreviated scheme showing reduction of carbon dioxide by the Calvin cycle. The first step is carboxylation, in which ribulose 1,5-bisphosphate carboxylase/oxygenase (Rubisco) catalyzes the addition of CO_2 to the five-carbon compound ribulose 1,5-bisphosphate, which is subsequently split into two molecules of the three-carbon compound 3-phosphoglycerate. Next are reduction and phosphorylation reactions that form the carbohydrate triose phosphate. Some of the triose phosphate molecules are used to form the products of photosynthesis, sucrose and starch, while the rest is used to regenerate ribulose 1,5-bisphosphate needed for the continuation of the cycle. Details are given in the text.

quantum requirements of 8–10 photons per oxygen molecule released. These quantum yield measurements show that the quantum yields of photosystem II and photosystem I reaction centers under optimal conditions are near 100%. These values can be used to calculate the theoretical energy conversion efficiency of photosynthesis (free energy stored as carbohydrate/light energy absorbed). If eight red quanta are absorbed (8 mol of red photons are equivalent to 1400 kJ) for each CO_2 molecule reduced (480 kJ/mol), the theoretical maximum energy efficiency for carbon reduction is 34%. Under optimal conditions, plants can achieve energy conversion efficiencies within 90% of the theoretical maximum. However, under normal growing conditions, the actual performance of the plant is far below these theoretical values. The factors that conspire to lower the quantum yield of photosynthesis include limitations imposed by biochemical reactions in the plant and environmental conditions that limit photosynthetic performance. One of the most efficient crop plants is sugar cane, which has been shown to store up to 1% of the incident visible radiation over a period of 1 year. However, most crops are less productive. The annual conversion efficiency of corn, wheat, rice, potatoes, and soybeans typically ranges from 0.1% to 0.4% (Odum, 1971).

5. OXYGENIC PHOTOSYNTHESIS IN ALGAE

Algae are photosynthetic eukaryotic organisms that evolve O_2 and reduce CO_2. They represent a diverse group that include dinoflagellates, euglenoids, yellow-green algae, golden-brown algae, diatoms, red algae, brown algae, and green algae. The photosynthetic apparatus and biochemical pathways of carbon reduction of algae are similar to plants. Photosynthesis occurs in chloroplasts that contain photosystems II and I, the cytochrome bf complex, the Calvin-cycle enzymes, and pigment-protein complexes containing chlorophyll *a* and other antenna pigments (*e.g.*, chlorophyll *b* in green algae, chlorophyll *c* and fucoxanthol in brown algae and diatoms, and phycobilins in red algae). Green algae are thought to be the ancestral group from which land plants evolved. Algae are abundant and widespread on the earth, living mainly in fresh and sea water. Some algae live as single-celled organisms, while others form multicellular organisms, some of which can grow quite large, such as kelp and seaweed. Phytoplankton in the ocean is made up of algae and oxygenic photosynthetic bacteria. Most photosynthesis in the ocean is due to phytoplankton, which is an important source of food for marine life.

6. OXYGENIC PHOTOSYNTHESIS IN BACTERIA

6.1 Cyanobacteria

Cyanobacteria are photosynthetic prokaryotic organisms that evolve O_2. Fossil evidence indicates that cyanobacteria existed over 3 billion years ago, and it is thought that they were the first oxygen-evolving organisms on earth. Cyanobacteria are presumed to have evolved in water in an atmosphere that lacked O_2. Initially, the O_2 released by cyanobacteria reacted with ferrous iron in the oceans and was not released into the atmosphere. Geological evidence indicates that the ferrous Fe was depleted around 2 billion years ago, and earth's atmosphere became aerobic. The release of O_2 into the atmosphere by cyanobacteria has had a profound affect on how life evolved.

The photosynthetic apparatus of cyanobacteria is similar to that of chloroplasts. The main difference lies in the antenna system. Cyanobacteria depend on chlorophyll *a* and specialized protein complexes called phycobilisomes that gather light energy. They do not contain chlorophyll *b*. As in chloroplasts, the chlorophyll *a* is located in membrane-bound proteins. The phycobilisomes are bound to the outer side of the photosynthetic membrane and act to funnel exciton energy to the photosystem II reaction center. They are composed of phycobiliproteins, protein subunits that contain covalently attached open ring structures known as bilins that are the light-absorbing pigments. Primary photochemistry, electron transport, phosphorylation, and carbon reduction occur much as they do in chloroplasts. Cyanobacteria have a simpler genetic system than plants and algae, which enables them to be easily modified genetically. Because of this, cyanobacteria have been used as a model to understand photosynthesis in plants. By genetically altering photosynthetic proteins, researchers can investigate the relationship between molecular structure and function.

6.2 Prochlorophytes

Over the past three decades, several types of oxygenic bacteria known as prochlorophytes (or oxychlorobacteria) have been discovered that have light-harvesting protein complexes that contain chlorophyll *a* and *b* but do not contain phycobilisomes (Palenik and Haselkorn, 1992). Because prochlorophytes have chlorophyll *a*/*b* light-harvesting proteins like chloroplasts, they are being investigated as models for plant photosynthesis.

7. ANOXYGENIC PHOTOSYNTHESIS IN BACTERIA

Anoxygenic photosynthetic bacteria differ from oxygenic organisms in that each species has only one type of reaction center. In some photosynthetic bacteria, the reaction center is similar to photosystem II, and in others, it is similar to photosystem I. However, neither of these two types of bacterial reaction center is capable of extracting electrons from water, and so they do not evolve O_2. Many species can only survive in environments that have a low concentration of O_2. To provide electrons for the reduction of CO_2, anoxygenic photosynthetic bacteria must oxidize inorganic or organic molecules available in their environment. For example, the purple bacterium *Rhodobacter sphaeroides* can use succinate to reduce NAD^+ by a membrane-linked reverse electron transfer that is driven by a transmembrane electrochemical potential. Although many photosynthetic bacteria depend on Rubisco and the Calvin cycle for the reduction of CO_2, some are able to fix atmospheric CO_2 by other biochemical pathways.

Despite these differences, the general principles of energy transduction are the same in anoxygenic and oxygenic photosynthesis. Anoxygenic photosynthetic bacteria depend on bacteriochlorophyll, a family of molecules similar to chlorophyll that absorb strongly in the infrared between 700 and 1000 nm. The antenna system consists of bacteriochlorophyll and carotenoids that serve a reaction center where primary charge separation occurs. The electron carriers include quinones (e.g., ubiquinone, menaquinone) and the cytochrome *bc* complex, which is similar to the cytochrome *bf* complex of oxygenic photosynthetic apparatus. As in oxygenic photosynthesis, electron transfer is coupled to the generation of an electrochemical potential that drives phosphorylation by ATP synthase. The energy required for the reduction of CO_2 is provided by ATP and NADH, a molecule similar to NADPH.

7.1 Purple Bacteria

There are two divisions of photosynthetic purple bacteria: the nonsulfur purple bacteria (e.g., *Rhodobacter sphaeroides* and *Rhodospeudomonas viridis*) and the sulfur purple bacteria (e.g., *Chromatium vinosum*). Nonsulfur purple bacteria typically use an organic electron donor, such as succinate or malate, but they can also use hydrogen gas. The sulfur bacteria use an inorganic sulfur compound, such as hydrogen sulfide, as the electron donor. The only pathway for carbon fixation by purple bacteria is the Calvin cycle. Sulfur purple bacteria must fix CO_2 to live, whereas nonsulfur purple bacteria can grow aerobically in the dark by respiration on an organic carbon source.

The determination of the three-dimensional structures of the reaction center of the nonsulfur purple bacteria, *Rhodopseudomonas viridis* and *Rhodobacter sphaeroides,* provides an unprecedented opportunity to understand the structure and function of photosynthetic reaction centers (Deisenhofer *et al.*, 1984, 1985; Feher *et al.*, 1989). The positions of the electron-transfer components in the reaction center of *Rhodobacter sphaeroides* are shown in Fig. 12 (Norris and van Brakel, 1986). The reaction center contains four bacteriochlorophyll and two bacteriopheophytin molecules. Two of the bacteriochlorophyll molecules form the primary donor (P870). At present, there is controversy over whether a bacteriochlorophyll molecule is an intermediate in electron transfer from the P870 to bacteriopheophytin. However, there is agreement that the remaining steps involve two quinone molecules (Q_A and Q_B) and that two turnovers of the reaction center result in the release of reduced quinone (QH_2)

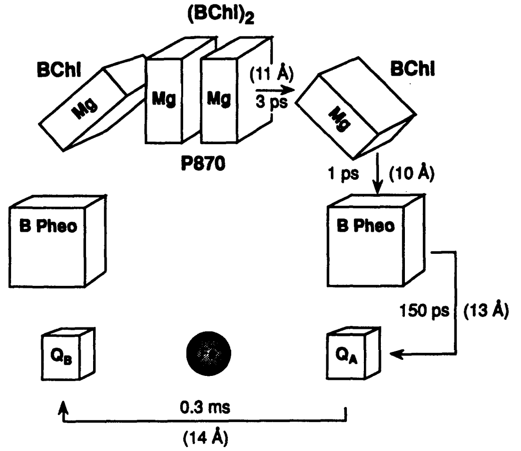
(BChl)$_2$
BChl
Mg
Mg
Mg
(11 Å)
3 ps
BChl
Mg
P870
1 ps
(10 Å)
B Pheo
B Pheo
150 ps
(13 Å)
Q$_B$
Q$_A$
0.3 ms
(14 Å)

FIG. 12. Relative positions of the chromophores of the reaction center of *Rhodobacter sphaeroides* (from Norris and van Brakel, 1986). Abbreviations: P870, reaction center bacteriochlorophyll (primary electron donor); B Pheo, bacteriopheophytin; Q_A and Q_B, bound ubiquinone. Details are given in the text.

into the photosynthetic membrane. Although there is a nonheme Fe between the two quinone molecules, there is convincing evidence that this Fe is not involved directly in transferring an electron from Q_A to Q_B. Because the primary donor (P870), bacteriopheophytin, and quinone acceptors of the purple bacterial reaction center are similar to the photosystem II reaction center, the bacterial reaction center is used as a guide to understand the structure and function of photosystem II.

Light-driven electron transfer is cyclic in *Rhodobacter sphaeroides* and other purple bacteria (Fig. 13). The reaction center produces reduced quinone, which is oxidized by the cytochrome *bc* complex. Electrons from the cytochrome *bc* complex are transferred to a soluble electron carrier, cytochrome c_2, which reduces the oxidized primary donor P870$^+$. The product of the light-driven electron transfer reactions is ATP. The electrons for the reduction of carbon are extracted from an organic donor, such as succinate or malate, or from hydrogen gas, but not by the reaction center. The energy needed to reduce NAD$^+$ is provided by light-driven cyclic electron transport in the form of ATP. The energy transformation pathway is complicated. Succinate is oxidized by the membrane-bound enzyme succinate dehydrogenase that transfers the electrons to quinone—the source of electrons for the reduction of NAD$^+$. However, electron transfer from reduced quinone to NAD$^+$ is energetically uphill. By a mech-

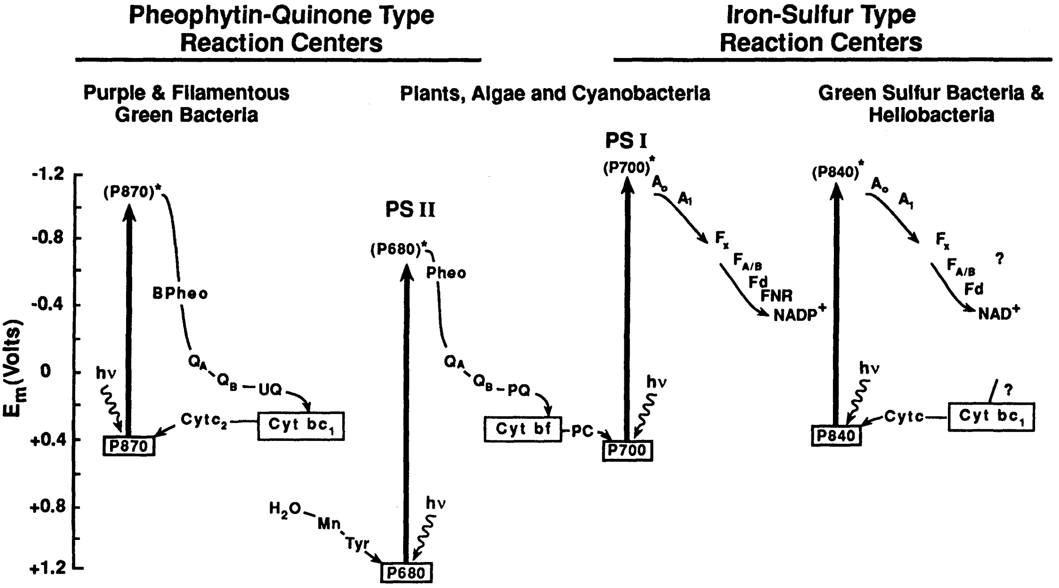
Pheophytin-Quinone Type Reaction Centers
Iron-Sulfur Type Reaction Centers
Purple & Filamentous Green Bacteria
Plants, Algae and Cyanobacteria
Green Sulfur Bacteria & Heliobacteria
E$_m$ (Volts)
-1.2
-0.8
-0.4
0
+0.4
+0.8
+1.2
(P870)*
BPheo
Q$_A$
Q$_B$ – UQ
hv
Cytc$_2$
Cyt bc$_1$
P870
PS II
(P680)*
Pheo
Q$_A$
Q$_B$ - PQ
Cyt bf
PC
H$_2$O
Mn
Tyr
hv
P680
PS I
(P700)*
A$_0$
A$_1$
F$_x$
F$_{A/B}$
Fd
FNR
NADP$^+$
hv
P700
(P840)*
A$_0$
A$_1$
F$_x$
F$_{A/B}$
?
Fd
NAD$^+$
hv
Cytc
?
Cyt bc$_1$
P840

FIG. 13. Comparison of electron-transport pathways in oxygenic and anoxygenic organisms (from Blankenship, 1992). Abbreviations: Cyt bc_1, cytochrome *bc* complex; P840, reaction center bacteriochlorophyll; other abbreviations are given in the legends of Figs. 3 and 12.

anism that is poorly understood, a membrane-bound enzyme is able to use energy stored in the proton electrochemical potential to drive electrons from reduced quinone to NAD^+.

7.2 Green Sulfur Bacteria

Green sulfur bacteria (e.g., *Chlorobium thiosulfatophilum* and *Chlorobium vibrioforme)* can use sulfur compounds as the electron donor as well as organic hydrogen donors. As shown in Fig. 13, the reaction center of green sulfur bacteria is similar to the photosystem I reaction center of oxygenic organisms. The FeS centers in the reaction center can reduce NAD^+ (or $NADP^+$) by ferredoxin and the ferredoxin-$NAD(P)^+$ oxidoreductase enzyme; therefore green sulfur bacteria are not necessarily dependent on reverse electron flow for carbon reduction. The antenna system of the green sulfur bacteria is composed of bacteriochlorophyll and carotenoids and is contained in complexes known as chlorosomes that are attached to the surface of the photosynthetic membrane. This antenna arrangement is similar to the phycobilisomes of cyanobacteria. Green sulfur bacteria can fix CO_2 without Rubisco. It has been proposed that they accomplish this by using the respiratory chain that normally oxidizes carbon (known as the Krebs cycle), resulting in the release of CO_2. With the input of energy, this process can be run in the reverse direction, resulting in the uptake and reduction of CO_2.

7.3 Green Gliding Bacteria

Green gliding bacteria (e.g., *Chloroflexus aurantiacus*), also known as green filamentous bacteria, can grow photosynthetically under anaerobic conditions or in the dark by respiration under aerobic conditions. Like the green sulfur bacteria, green gliding bacteria harvest light using chlorosomes. The green gliding bacteria appear to have reaction centers similar to those of the purple bacteria (Fig. 13), but there are several notable differences. For example, instead of two monomer bacteriochlorophyll molecules, *C. aurantiacus* has one bacteriochlorophyll and one bacteriopheophytin, and the metal between the two quinones is Mn rather than Fe. *C. aurantiacus* appears to fix CO_2 by a scheme that does not involve the Calvin cycle or the reverse Krebs cycle (Ivanovsky *et al.*, 1993).

7.4 Heliobacteria

Heliobacteria (e.g., *Heliobacterium chlorum* and *Heliobacillus mobilis*) are in the phylum Gram Positive Bacteria that are strict anaerobes. Although the heliobacterial reaction center is similar to photosystem I in that it can reduce NAD^+ (or $NADP^+$), it contains a different type of chlorophyll known as bacteriochlorophyll *g*.

8. CONTROL OF INTRAPROTEIN ELECTRON TRANSFER

The three-dimensional structure of the reaction center of *Rhodopseudomonas viridis* and *Rhodobacter sphaeroides* reveals the distances between the electron donors and acceptors (Deisenhofer *et al.*, 1984,1985; Norris and van Brakel, 1986; Feher *et al.*, 1989) and has had an important influence on biophysical and molecular genetics studies designed to identify the factors that control the rate of electron transfer within proteins. There is currently a controversy concerning the importance of the amino acid composition of the protein on the rate of intraprotein electron transfer. In part, the disagreement centers on whether the protein between the donor and acceptor molecules can be treated as a uniform material or whether the specific amino acid composition of the protein significantly alters the rate. For example, it has been proposed that aromatic amino acids may provide a particular pathway that facilitates electron transfer between a donor and acceptor pair. This is the case in the photosystem II reaction center, where a tyrosine residue on one of the reaction center core proteins donates an electron to the primary donor chlorophyll, $P680^+$. However, in other cases, replacement of an aromatic residue by another nonaromatic residue results in relatively minor changes in the rate of electron transfer. L. Dutton and co-workers (Moser *et al.*, 1992) have analyzed electron-transfer reactions in biological and chemical systems in terms of electron tunneling theory developed by R. Marcus and others (DeVault, 1984). Dutton and co-workers argue that protein provides a uniform electronic barrier to electron tunneling and a uniform nuclear characteristic frequency. They suggest that the specific amino acid residues between an electron transfer pair are generally of less importance than the distance in determining the rate of pairwise electron transfer. In their view, protein controls the rate of electron transfer mainly through the distance between the donor and acceptor molecules, the free energy, and the reorganization energy of the reaction. The importance of distance is demonstrated by electron-transfer data from biological and synthetic systems showing that the dependence of the electron transport rate on the edge-to-edge distance is exponential over 12 orders of magnitude when the free energy is optimized (Moser *et al.*, 1992). Increasing the distance between two carriers by 1.7 Å slows the rate of electron transfer tenfold. The extent to which this view is generally applicable for intraprotein electron transfer remains to be established (Williams, 1992). One of the challenges in understanding pairwise electron-transfer rates from first principles is illustrated by the reaction center of *Rhodobacter sphaeroides*, in which the redox components are arranged along a twofold axis of symmetry that extends from the primary donor (P870) to the Fe. Despite the fact that the reaction center presents two spatially similar pathways for electron transfer from P870 to quinone, nearly all electrons are transferred down the right arm of the reaction center, as shown in Fig. 12. The same holds true for the reaction center of *Rhodopseudomonas viridis*, in which it is estimated that electron transfer down the left arm is less than 1:100 (Kellogg *et al.*, 1989). The challenge to theorists is to explain the surprisingly high probability of electron flow down the right arm. Since the distances are similar, it has been suggested that electron transfer down the left arm is less probable because of an endothermic free-energy change (Parson *et al.*, 1990) or to an unfavorable rearrangement energy for the reaction (Moser *et al.*, 1992).

9. GLOBAL PHOTOSYNTHESIS AND THE ATMOSPHERE

The amount of CO_2 removed from the atmosphere each year by oxygenic photosynthetic organisms is massive. It is estimated that photosynthetic organisms remove 100×10^{15} g/yr of carbon (Houghton and Woodwell, 1989). This is equivalent to 4×10^{18} kJ of free energy stored in reduced carbon, which is roughly 0.1% of the incident visible radiant energy incident on the earth per year. Each year the photosynthetically reduced carbon is oxidized, either by living organisms for their survival or by combustion. The result is that more CO_2 is released into the atmosphere from the biota than is taken up by photosynthesis. The amount of carbon released by the biota is estimated to be $(1–2) \times 10^{15}$ g/yr of carbon. Added to this amount is carbon released by the burning of fossil fuels, which amounts to 5×10^{15} g/yr of carbon. The oceans mitigate this increase by acting as a sink for atmospheric CO_2. It is estimated that the oceans remove about 2×10^{15} g/yr of carbon from the atmosphere. This carbon is eventually stored on the ocean floor. Although these estimates of sources and sinks are uncertain, the net global CO_2 concentration is increasing. Direct measurements show that each year the atmospheric carbon content is currently increasing by about 3×10^{15} grams. Over the past 200 years, CO_2 in the atmosphere has increased from about 280 parts per million (ppm) to its current level of 360 ppm. On the basis of predicted fossil fuel use and land management, it is estimated that the amount of CO_2 in the atmosphere will reach 700 ppm within the next century. The consequences of this rapid change in our atmosphere are unknown. Because CO_2 acts as a greenhouse gas, some climate models predict that the temperature of the earth's atmosphere may increase by 2–8 °C. Such a large temperature increase would lead to significant changes in rainfall patterns. Little is known about the impact of such drastic atmospheric and climatic changes on plant communities and crops. Current research is directed at understanding the interaction between global climate change and photosynthetic organisms.

GLOSSARY

ATP: Adenosine triphosphate, a small water-soluble molecule that acts as an energy currency in cells.

ATP Synthase: A membrane-bound protein complex that uses the energy stored across the photosynthetic membrane to add inorganic phosphate to ADP, thus creating ATP. (Also known as coupling factor.)

Calvin Cycle: The biochemical reactions, initiated by Rubisco, that result in the reduction of CO_2 to a carbohydrate (also known as the photosynthetic carbon reduction cycle).

Cytochrome: Heme-containing protein.

Cytochrome *bc* Complex: A membrane-bound electron-transfer protein complex, found in all anoxygenic photosynthetic organisms, that oxidizes reduced quinone and reduces a *c*-type cytochrome. The complex contains a *c*-type cytochrome, two *b*-type cytochromes, and an FeS center.

Cytochrome *bf* Complex: A membrane-bound electron-transfer protein complex, found in all oxygenic photosynthetic organisms, that oxidizes reduced plastoquinone and reduces plastocyanin (or cytochrome *c*). The complex contains a *c*-type cytochrome, two *b*-type cytochromes, and an FeS center.

Free Energy: The amount of energy in a reaction available to do work. Because most biochemical reactions occur at a constant temperature and pressure, the free energy is frequently the Gibbs energy.

Light-Harvesting Complex: A protein complex that harvests light energy and converts it to exciton energy that can migrate to a reaction center. The light is absorbed by pigment molecules (e.g., chlorophyll, bacteriochlorophyll, carotenoids, phycobilin) that are attached to the protein.

NADPH: Reduced form of nicotinamide adenine dinucleotide phosphate, a small water-soluble molecule that acts as a hydrogen carrier in biochemical reactions.

$NADP^+$: Oxidized form of nicotinamide adenine dinucleotide phosphate.

Oxidation: The removal of one or more electrons from an atom or molecule. In the case of a molecule, protons may be involved as well, resulting in hydrogen being removed.

Phosphorylation: The covalent attachment of a phosphate group to a molecule.

Photorespiration: The removal of O_2 from the atmosphere by Rubisco and the subsequent biochemical reactions that serve to recycle some of the reduced carbon.

Photosynthesis: The physical-chemical process by which certain chlorophyll- (or bacteriochlorophyll-) containing organisms use light energy for the biosynthesis of organic molecules.

Photosynthetic Membrane: A bilayer of lipid molecules in which are embedded proteins that transform light energy into chemical free energy. (Also known as the thylakoid membrane.)

Photosystem I: A protein complex located in the photosynthetic membrane. Photosystem I is one of two types of reaction centers found in higher plants, algae, and cyanobacteria. The photosystem I reaction center uses light energy to transfer an electron from a mobile electron-transfer protein (plastocyanin or cytochrome *c*) on one side of the photosynthetic membrane to a mobile electron-transfer protein (ferredoxin) on the opposite side of the photosynthetic membrane.

Photosystem II: A protein complex found in the photosynthetic membrane. Photosystem II is one of two types of reaction centers found in higher plants, algae, and cyanobacteria. The photosystem II reaction center uses light energy to transfer electrons from water to plastoquinone. Photosystem II is the source of the molecular oxygen in the atmosphere.

Plastoquinone: A small organic molecule involved in electron and proton transfer in photosynthesis.

Protein: A chemical structure composed of one or more polypeptides. In photosynthesis, proteins serve as the scaffolding that hold the cofactors that gather light energy, transfer electrons, and catalyze biochemical reactions.

Reaction Center: A protein complex that uses light energy to create a stable charge separation by transferring a single electron energetically uphill from a donor molecule to an acceptor molecule, both of which are located in the reaction center.

Reduction: The addition of one or more electrons to an atom or molecule. In the case of a molecule, protons may be involved as well, resulting in hydrogen being added.

Rubisco (D-ribulose 1,5-bisphosphate carboxylase/oxygenase): A water-soluble protein complex responsible for the removal of CO_2 from the atmosphere. The enzyme works by attaching CO_2 to a five-carbon compound (1,5 ribulose bisphosphate) that is split into two identical three-carbon compounds (phosphoglycerate). In addition to catalyzing the removal of CO_2 from the atmosphere, Rubisco also catalyzes the removal of O_2 from the atmosphere (less efficiently). The removal of O_2 is thought to be a consequence of poor design and leads to a complex set of compensatory reactions known as photorespiration.

Works Cited

Blankenship, R. (1992), *Photosynth. Res.* **33,** 91–111.

Boyer, P. D. (1993), *Biochim. Biophys. Acta* **1140,** 215–250.

Debus, R. (1992), *Biochim. Biophys. Acta* **1102,** 269–352.

Deisenhofer, J., Epp, O., Miki, K., Huber, R., Michel, H. (1984), *J. Mol. Biol.* **180,** 385–398.

Deisenhofer, J., Epp, O., Miki, K., Huber, R., Michel, H. (1985), *Nature* **318,** 618–624.

DeVault, D. (1984), *Quantum Mechanical Tunnelling in Biological Systems,* Cambridge, U.K.: Cambridge University Press.

Feher, G., Allen, J. P., Okamura, M. Y., Rees, D. C. (1989), *Nature* **339,** 111–116.

Gest, H. (1993), *FEMS Microbiol. Lett.* **112,** 1–6.

Govindjee and Coleman, W. J. (1990), *Sci. Am.* **262** (2), 50–58.

Houghton, R. A., Woodwell, G. M. (1989), *Sci. Am.* **260** (4), 36–44.

Ivanovsky, R. N., Krasilnikova, E. N, Fal, Y. I. (1993), *Arch. Microbiol.* **159,** 257–264.

Kellogg, E. C., Kolaczkowski, S., Wassielewski, M. R., Tiede, D. M. (1989), *Photosynth. Res.* **22,** 47–59.

Klotz, I. M. (1967), *Energy Changes in Biochemical Reactions,* New York: Academic Press, Chap. 6.

Krauss, N., Hinrichs, W., Witt, I., Fromme, P. Pritzkow, W., Dauter, Z. Betzel, C., Wilson, K. S., Witt, H. T., Saenger, W. (1993), *Nature* **361,** 326–331.

Kühlbrandt, W., Wang, D. N., Fujiyoshi, Y. (1994), *Nature* **367,** 614–621.

Moser, C. C., Keske, J. M., Warncke, K., Farid, R. S., Dutton, P. L. (1992), *Nature* **355,** 796–802.

Norris, J. F., van Brakel, G. (1986), in: Govindjee, J. Amesz, D. C. Fork (Eds.), *Light Emission by Plants and Bacteria,* Orlando: Academic Press, pp. 35–36.

Odum, E. P. (1971), *Fundamentals of Ecology,* Philadelphia: W.B. Saunders Co.

Palenik, B., Haselkorn, R. (1992), *Nature* **355,** 265–267.

Parson, W. W., Chu, Z. T., Warshel, A. (1990), *Biochim. Biophys. Acta* **1017,** 251–272.

Staehelin, L. A. (1986), *Encyclopedia Plant Physiol.* **19,** 1–84.

Whitmarsh, J. (1986), *Encyclopedia Plant Physiol.* **19,** 508–527.

Williams, R. J. P. (1992), *Nature* **355,** 770.

Woese, C. R., Kandler, O., Wheeler, M. L. (1990), *Proc. Natl. Acad. Sci. USA* **87,** 4576–4579.

Wraight, C. (1982), in: Govindjee (Ed.), *Photosynthesis,* New York: Academic Press, pp. 17–61

Further Reading

Cramer, W. A., Knaff, D. B. (1991), *Energy Transduction in Biological Membranes,* New York: Springer-Verlag.

Barber, J. (Ed.) (1992), *The Photosystems: Structure, Function and Molecular Biology,* New York: Elsevier.

Huzisige, H., Ke, B. (1993), "Dynamics of the History of Photosynthesis," in: *Photosynthesis Res.* **38,** 185–209.

Lawlor, D. W. (1993), *Photosynthesis: Molecular, Physiological and Environmental Processes,* Essex, U.K.: Longman Scientific & Technical.

Taiz, L., Zeiger, E. (1991), *Plant Physiology,* Redwood City, CA: Benjamin/Cummings Publishing Co.

Walker, D. (1992), *Energy, Plants and Man,* East Sussex, U.K.: Oxygraphics Limited.

PHOTOVOLTAIC DEVICES

Joseph J. Loferski, *Division of Engineering, Brown University, Providence, Rhode Island, U.S.A.*

Introduction ... 533
1. **Physical Principles Underlying the PV Effect** ... 535
1.1 Light Absorption in Semiconductors ... 535
1.2 Charge Separation in PV Cells ... 536
1.3 Lifetime of Light-Generated Carriers ... 538
2. **Equations Governing the PV Effect** ... 539
2.1 Diode Equation ... 539
2.2 Calculation of the Short-Circuit Current I_{SC} ... 541
2.3 The Reverse Saturation Current I_0 and the Junction Quality Factor A ... 543
2.4 Maximizing $\eta_{\max}$ ... 544
2.5 Dependence of $\eta_{\max}$ on E_G ... 545
2.6 Ultimate Efficiency Calculation ... 547
2.7 Higher Efficiencies through Spectrum Splitting ... 547
2.8 Dependence of $\eta_{\max}$ on Temperature ... 549
3. **Semiconductor Materials Used in PV Cells: History and Current Status** ... 549
3.1 Silicon Monocrystalline and Polycrystalline Cells ... 549
3.2 Gallium Arsenide Cells ... 550
3.3 Cadmium Sulfide/Copper Sulfide ... 550
3.4 *p*-Copper Indium Selenide/*n*-CdS Cells ... 551
3.5 Cadmium Telluride Cells ... 551
3.6 Amorphous Silicon Solar Cells ... 551
3.7 Cells Based on Other Semiconductors ... 552
4. **Effects of High-Energy Radiation on PV Cells** ... 552
5. **Environmental Aspects of PV Cells** ... 553
6. **Applications of PV Power Supplies** ... 553
6.1 Satellite Power Supplies ... 554
6.2 Remote Power ... 554
6.3 Other Applications ... 554
7. **Economics of PV Power Plants for Large-Scale Electricity Generation** ... 554
8. **Summary and Conclusions** ... 555
Glossary ... 556
Works Cited ... 556
Further Reading ... 557

INTRODUCTION

The photovoltaic effect (PVE) is the generation of an emf as a result of absorption of ionizing radiation. The absorbing medium can be a gas, a liquid, or a solid. For the PVE to occur, an electrostatic potential-difference site must be present in the absorbing medium. The positive and negative charges resulting from the ionization must then be separated by this internal electrostatic field. The separated charges can pass through an external resistive load connected across the potential barrier, thus transforming a portion of absorbed ionizing energy directly into electricity.

The first paper describing the PVE was published by E. Becquerel (1839). He observed an emf when he directed light onto platinum or gold electrodes immersed in acidic or alkaline solutions. The first report of a PVE observed in a solid was published by W. G. Adams and R. E. Day in 1877. They observed the effect in a small selenium rod into whose ends they fused platinum electrodes. During the following 30 years, research and development resulted in an im-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

proved selenium PV cell, which eventually found practical application as a light meter for photography and as a light sensor in optical instruments.

The next important step in the development of PV science and technology was the invention/discovery during the period 1930–1933 of the PV effect in copper/copper oxide devices by Lange (1930), Schottky (1930), and Grondahl (1933). Schottky's efforts to explain this PV effect made him a pioneer in the then fledgling science of semiconductors. A very important contribution by Schottky was his introduction of the concept of the metal–semiconductor electrostatic potential barrier, the Schottky barrier, to explain the PV effect in both the Cu/CuO_x and the metal/selenium cells. Because of simplicity of fabrication, the Cu/CuO_x cell became a competitor of the selenium cell in the gradually increasing commercial market for such devices.

However, a revolution in the application of the PV effect began with the invention of the transistor (Brattain and Bardeen, 1948; Shockley, 1949). The first transistors utilized as semiconductors the elements silicon and germanium. These new semiconductors could be grown as large single crystals whose electronic properties could be controlled by the addition of minute quantities of other elements. Furthermore, electric current in these new semiconductors could be carried by either positive or negative charges. The positive charges proved to be as mobile as the negative charges (known to be electrons). These mobile positive charged carriers are called "holes" because they result from holes in the electron bonding structures of these materials.

Germanium and silicon are elements from the fourth column of the periodic table; they have a valence of 4. Each atom shares its four valence electrons with its four nearest neighbors, thus forming a crystal consisting of two interlocking face-centered cubic lattices—the diamond structure. At absolute zero, all of the electrons are tied up in covalent bonds; there are no electrons free to move through the lattice if an electric field is applied across the semiconductor. As the temperature increases, some of the electrons are thermally excited and thus liberated from their covalent bonds, leaving behind "holes" in the bonding structure. The liberated electron and the hole can now move through the lattice; they are no longer bound to a specific atom. Generation of such free electrons and holes requires an external source of energy (heat, light, etc.). The minimum energy needed to generate a free electron-hole pair is the difference in energy between the highest valence band and the lowest conduction band in the semiconductor, i.e., the forbidden energy gap E_G.

In an intrinsic semiconductor, the concentrations of free holes and electrons are equal since a free hole is formed whenever a free electron is generated. If elements having a valence different from 4 are incorporated into the crystal, the concentrations of electrons and holes need not be equal any more. There can be an excess of electrons or an excess of holes, depending on the valence of the impurity atom. A semiconductor whose conductivity is controlled by such "extrinsic" elements is referred to as an extrinsic semiconductor. If the dominant impurity element has a valence of 5 (like phosphorus, arsenic, or antimony), then it can replace silicon or germanium atoms in the crystal lattice, forming four covalent bonds with its neighbors. The extra valence electron is not bound to the impurity atom as tightly as the electrons participating in the bonding with the column-four host element; it can move freely through the lattice at temperatures around room temperature. A semiconductor in which the dominant impurity contributes such free electrons to the conduction band is called an *n*-type semiconductor. If the dominant impurity element has a valence of 3 (like boron, aluminum, or indium), when it replaces a silicon or germanium atom in the lattice, it lacks one electron to satisfy the covalent bonding with its neighboring atoms, i.e., there is a "hole" in the bonds around the trivalent atom. This hole arises from the absence of an electron and, therefore, moves through the lattice as if it had a positive charge. A semiconductor in which the dominant impurity contributes free holes to the valence band is called a *p*-type semiconductor.

Shockley developed an elegant theory that explained what happens if a piece of semiconductor has a variable concentration of holes and electrons so that a transition from *p* type to *n* type occurs, thus forming a *p-n* junction. He showed that the *p-n* junction produced an internal electrostatic potential barrier whose properties were far more predictable and reproducible than the properties of Schottky barriers. Rectifiers based on such *p-n* junctions in germanium and silicon proved to be much better than those based on the metal-semiconductor junctions in metal-selenium and Cu/CuO$_x$ rectifiers.

They also proved to be much better photovoltaic cells, as the scientific world was to learn from three brief papers published in journals of the American Physical Society in 1954 (Rappaport, 1954; Chapin *et al.*, 1954: Reynolds *et al.*, 1954). These papers, which marked the beginning of the modern photovoltaic age, described the electron voltaic and photovoltaic effects in *p-n* junction devices made from single crystals of germanium, silicon, or cadmium sulfide. The new PV cells, more specifically those based on silicon and cadmium sulfide, were able to convert the radiant energy of sunlight directly into electricity with efficiencies in the 4% to 6% range, about an order of magnitude higher than the highest solar-energy conversion efficiencies measured on Cu/CuO$_x$ or metal-selenium devices of that time.

Subsequent developments in theory and in the experimental science and technology of PV cells are the subjects addressed in the balance of this article. These developments have led to the fabrication and study of PV cells made from many semiconductors besides silicon; to PV cells having solar-energy conversion efficiencies up to 35%; and to widespread application of PV cells as power supplies for earth satellites, pocket calculators, remote sites, etc. If continued—and it shows no signs of abating—this long-term trend toward improved performance accompanied by reduced cost will result in large-scale, environmentally benign PV power plants displacing conventional fossil fuel and nuclear plants as preferred sources of electricity.

1. PHYSICAL PRINCIPLES UNDERLYING THE PV EFFECT

To produce a PV effect, three events must occur. First, excess positive and negative carriers must be produced by the absorption of ionizing radiation. The ionizing radiation can be electromagnetic (infrared, visible, ultraviolet, x rays, γ rays) or charged-particle radiation (electrons, protons, α particles, β particles). The photons of the electromagnetic radiation must have an energy equal to or in excess of E_G, the forbidden energy gap of the semiconductor that is the internal ionization energy of the semiconductor. As energetic charged particles pass through the semiconductor, they produce internal ionization by liberating bonding electrons. The average energy expended by a charged particle to produce one hole-electron pair in silicon and germanium is about $3.5E_G$. When high-energy electrons are the ionizing radiation, the effect is called the electron-voltaic effect with similar nomenclature for other charged particles. All such effects are essentially the same and the photovoltaic effect can be viewed as a generic term describing them. Second, the excess charges of opposite signs must be separated by an internal electrostatic potential barrier like a *p-n* junction or a metal-semiconductor junction. Third, the mobile carriers generated by the ionizing radiation must "live" long enough to reach the charge-collecting or -separating barrier. For further details, see the following subsections and Loferski (1979).

1.1 Light Absorption in Semiconductors

Let a monochromatic beam of light containing photons whose energy exceeds the forbidden energy gap of a particular semiconductor be incident on that semiconductor. If the photon flux incident on the plane $x = 0$ is $N_{\mathrm{ph}}(0)$, then the photon flux remaining after traversing a thickness x of the semiconductor, $N_{\mathrm{ph}}(x)$, is given by the relation

$$N_{\mathrm{ph}}(x) = N_{\mathrm{ph}}(0)\exp(-\alpha x), \qquad (1)$$

where α is the absorption constant for that particular energy photon. The photon flux lost from the beam, $N_{\mathrm{ph}}(0) - N_{\mathrm{ph}}(x)$, is absorbed in the semiconductor and generates at least one hole-electron pair for each photon absorbed. The absorption constant α is a function of the wavelength λ or photon energy $h\nu$. The dependence of α on $h\nu$ is a fundamental characteristic of a semiconductor. Figure 1 shows curves of α vs $h\nu$ for a few semiconductors that have been used in PV cells. The threshold energy for the onset of absorption is the energy gap E_G of the semiconductor, which is a first-order parameter in determining the efficiency with which a PV cell can

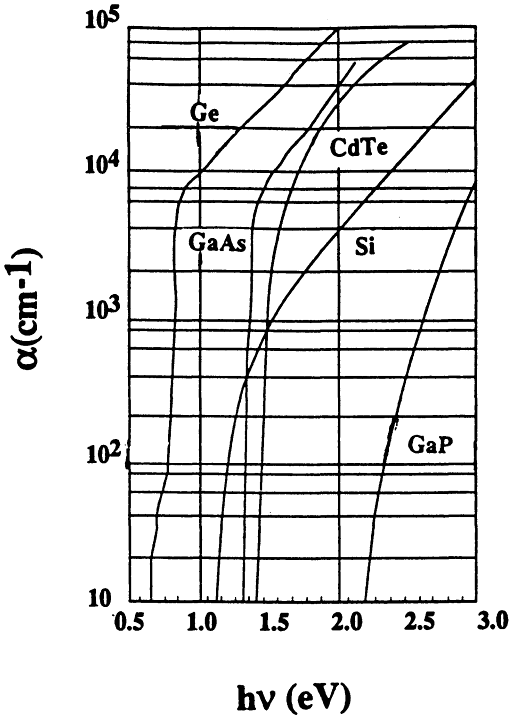

FIG. 1. Absorption constant α versus photon energy $h\nu$ for selected semiconductors of interest as photovoltaic cell materials.

convert the energy of ionizing energy into electricity. Closer examination of these curves of α vs $h\nu$ leads to the conclusion that they can be divided into two groups:

1. those in which α rises very rapidly from very low values, say 10 cm^{-1}, to values in excess of 10^4 cm^{-1} (the GaAs and CdTe curves are in this category); and
2. those in which α rises more gradually (the Si, Ge, and GaP curves are in this category).

Semiconductors with rapidly rising curves of α vs $h\nu$ are called direct-gap semiconductors; those with gradually rising curves of α vs $h\nu$ are called indirect-gap semiconductors.

A practical consequence of this difference for PV cells used as solar-energy converters can be seen in Fig. 2. Here we compare the fraction of photons in the solar spectrum that have an energy greater than the forbidden energy gap E_G and are absorbed in a given thickness of the indirect-gap semiconductor silicon with the same fraction in the direct-gap semiconductor gallium arsenide (GaAs).

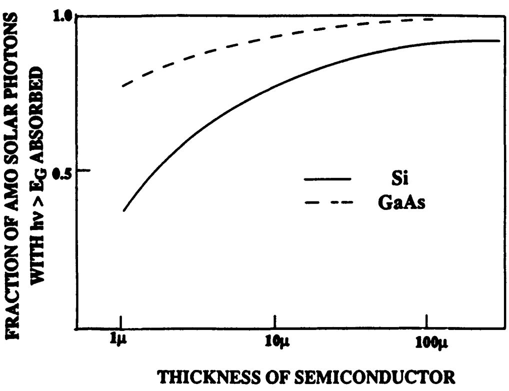

FIG. 2. Comparison of fraction of solar photons with energy greater than their energy gaps absorbed in a given thickness of the direct-gap semiconductor GaAs and the indirect-gap semiconductor silicon.

For example, to absorb 90% of the photons that silicon is capable of absorbing, we need a thickness of approximately 100 μm, whereas for GaAs, the corresponding thickness is only 10 μm. This difference is very significant for large-scale conversion of sunlight into electricity, which would require covering tens of thousands of square kilometers with solar cells, because the amount of direct-gap semiconductor required is much smaller and, therefore, should ultimately be less expensive than the amount of an indirect-gap semiconductor required for the same purpose. Of course, the relative abundance of the materials of which the semiconductor is composed has a strong effect on the cost. For example, silicon is an extremely abundant element, whereas gallium is not. Nevertheless, the amount of gallium available is sufficient to make direct-gap semiconductors in which it is a constituent strong candidates as building blocks for large-scale PV applications.

1.2 Charge Separation in PV Cells

As has been noted, charge separation requires the presence of an electrostatic potential barrier inside the PV cell. Figure 3 shows the electron energy-band diagrams of the three basic kinds of charge-separation barriers. In this figure, E_C is the electron energy at the bottom of the conduction band, E_V is the electron energy at the top of the valence band, and E_F is the electron energy at the Fermi

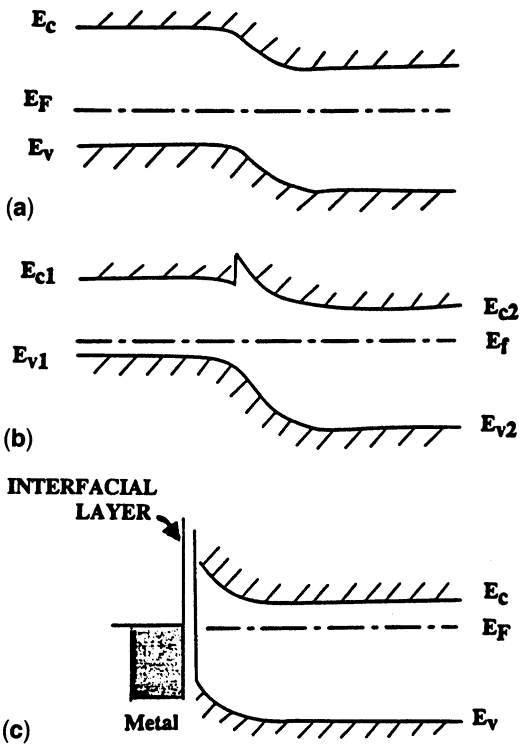

FIG. 3. Energy-band diagrams of (a) a *p-n* homojunction, (b) a *p-n* heterojunction, and (c) a metal-semiconductor (Schottky barrier) junction.

level. If the temperature across the barrier is uniform and there is no radiant energy input, i.e., the PV cells are in the dark, the Fermi energy is the same throughout the material including the ends of the junction. The energy differences $E_F - E_V$ and $E_C - E_F$ are determined by the temperature and the ionized-impurity (dopant) concentrations in the various regions of the semiconductor. The energy-band diagrams of Fig. 3 correspond to the case of uniform temperature and constant but different dopant concentrations in the p and n regions. When $E_F - E_V$ is less than $E_C - E_F$, the material is p type and the current carriers are predominantly holes. When $E_F - E_V$ is greater than $E_C - E_F$, the material is n type and electrons are the dominant charge carriers. So band diagrams (a) and (b) of Fig. 3 have p-type material on the left side of the barrier and n-type material on the right side of the barrier, i.e., they are the band diagrams of a p-n junction.

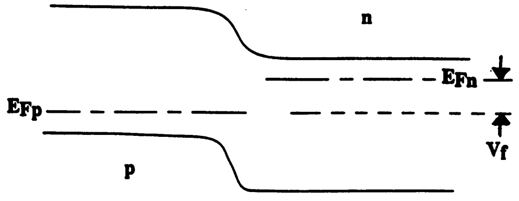

FIG. 4. Energy-band diagram of a *p-n* homojunction under forward bias.

Figure 3(a) is the band diagram of a p-n homojunction because the p-n junction spans a single semiconductor. Figure 3(b) is the band diagram of a p-n heterojunction because the p side (left side) consists of one semiconductor and the n side (right side) consists of another semiconductor. In Fig. 3(b), the energy gaps $E_C - E_V = E_G$ are substantially different on the p and n sides.

Figure 3(c) is the band diagram of a metal-semiconductor or Schottky barrier junction in which the work functions of metal and semiconductor bear such a relation to each other that the band edges of the n-type semiconductor bend upward at the junction. The sketch includes a thin interfacial layer (which may or may not be present) between the metal and the semiconductor. If this interfacial layer is thin enough, carriers can tunnel through it as they pass from the metal to the semiconductor or vice versa.

If a voltage V is applied across the p-n homojunction of Fig. 3(a), the band diagrams change to the forms shown in Figs. 4 and 5, depending on whether the n side is more negative than the p side (Fig. 4) or the n side is more positive than the p side (Fig. 5). The voltage bias of Fig. 4 is referred to as a forward bias, while the voltage bias of Fig. 5 is referred to as a reverse bias. When an open-circuited p-n junction absorbs ionizing radiation, it acquires a forward bias like that depicted in Fig. 4. This occurs because the elec-

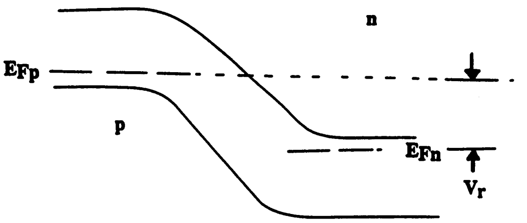

FIG. 5. Energy-band diagram of a *p-n* homojunction under reverse bias.

trostatic barrier associated with a *p-n* junction like that depicted in Fig. 3(a) is a "sink" for excess minority electrons generated on the *p* side and for excess minority holes generated on the *n* side. A flow of excess electrons from the *p* side to the *n* side and of excess holes from the *n* side to the *p* side would constitute an electric current flowing in the "reverse" direction. But if the junction is open circuited, by definition there can be no current flowing through it. So the junction reacts by biasing itself in the forward direction, i.e., it reduces the height of the electrostatic barrier just enough to produce a forward current equal in magnitude to the reverse current generated by the ionizing radiation. Thus, an illuminated *p-n* junction that is not open circuited acquires a forward bias and the current flows through it in the reverse direction, as shown in the *I–V* diagram of an illuminated solar cell of Fig. 6. The figure shows the open-circuit (forward) voltage (V_{oc}); the short-circuit (reverse) current I_{sc}; the maximum power point, P_{max}; the load voltage at maximum power, V_{mp}; and the load current at maximum power, I_{mp}.

1.3 Lifetime of Light-Generated Carriers

As noted above, the excess carriers generated by the absorbed ionizing radiation must remain mobile long enough to allow them to migrate to the charge-separating junction. Their lifetime is terminated when positive and negative carriers recombine with each other. Recombination can be intrinsic or extrinsic in origin; it can occur in the bulk or at the surface. The lifetime of the excess carriers is inversely proportional to the recombination rate.

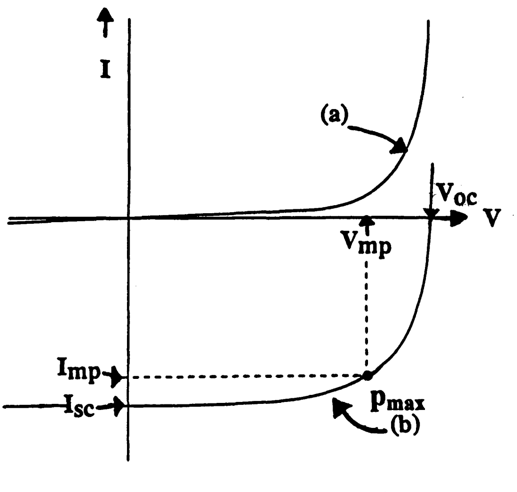

FIG. 6. Current–voltage characteristic of a solar cell (a) in the dark and (b) subjected to ionizing radiation, showing the open-circuit voltage V_{oc}; the short-circuit current I_{sc}, the maximum power point, P_{max}; the current at maximum power, I_{mp}; and the voltage at maximum power, V_{mp}.

For the case of bulk intrinsic recombination, the excess minority carriers generated in one of the bands by the ionizing radiation recombine with carriers of the opposite sign in the other band. Such band-to-band recombination sets an upper limit on the minority-carrier lifetime. Its value can be as high as 10^{-2} s in an indirect-gap semiconductor like Si and as low as 10^{-8} s in a direct-gap semiconductor like GaAs.

Indirect recombination requires the presence of recombination centers in the semiconductor. Recombination centers have energy levels that are closer to the middle of the forbidden energy gap than the energy levels of dopant impurities. In the case of silicon, typical energy levels E_r of recombination centers are located between 0.15 and 0.25 eV from one of the band edges. If the concentration of recombination centers N_r is high enough (10^{13} centers/cm^3 is high in silicon doped with 10^{16}/cm^3 donors or acceptors), the lifetime is determined by these centers. It is given by the expression

$$1/\tau_{\text{ext}} = N_r \sigma_r v f(E_r - E_{\text{F}}), \qquad (2)$$

where τ_{ext} is the extrinsic bulk minority-carrier lifetime; σ_r is the minority-carrier capture cross section of the recombination center (10^{-12} cm^2 < σ_r < 10^{-16} cm^2); v is the thermal velocity of the carriers (about 10^7 cm/s at 300 K); and $f(E_r - E_{\text{F}})$ is the probability that the recombination center is occupied by a majority carrier, i.e., it is the Fermi–Dirac statistics occupancy factor for the center. If $E_{\text{F}} - E_r > kT, f = 1$, and the range of values for $N_r\tau_{\text{ext}}$ is

$$10^5 \text{ s/cm}^3 \le N_r\tau_{\text{ext}} \le 10^9 \text{ s/cm}^3.$$

Therefore, with $N_r = 10^{13}$/cm^3 the range of values of extrinsic lifetime would be

$$10^{-8} \text{ s} \le \tau_{\text{ext}} \le 10^{-4} \text{ s}. \qquad (3)$$

The actual lifetime τ is related to τ_{ext} and the intrinsic lifetime τ_{int} by the relation

$$1/\tau = 1/\tau_{\text{int}} + 1/\tau_{\text{ext}}. \tag{4}$$

In an indirect-gap semiconductor, like silicon, the extrinsic lifetime dominates. In direct-gap material like GaAs, the intrinsic lifetime (which is low anyway) can dominate unless the concentration of recombination centers is greater than about $10^{15}/\text{cm}^3$.

If a PV cell is to achieve its optimum performance, its minority-carrier lifetime must be as high as possible. This in turn means that the lifetime should be close to the intrinsic lifetime, i.e., the concentration of recombination centers N_r must be as low as possible. For an indirect-gap semiconductor like silicon in which τ_{int} is large (10^{-2} s), N_r must be of the order of $10^{11}/\text{cm}^3$. For a direct-gap semiconductor like GaAs, N_r can be as high as $10^{15}/\text{cm}^3$ without affecting τ.

2. EQUATIONS GOVERNING THE PV EFFECT

2.1 Diode Equation

The derivation of equations for the PVE begins with the ambipolar diffusion equation formulated by Shockley (1951). This equation governs the transport of excess carriers (like those produced by absorption of ionizing radiation) in a semiconductor. The appropriate one-dimensional form of this equation for the light-generated excess minority-carrier population Δn on the p side of a p-n junction is

$$D_e \frac{d^2 \Delta n}{dx^2} + \alpha N_0 e^{-\alpha x} - \frac{\Delta n}{\tau} = 0. \tag{5}$$

An analogous equation applies to the excess minority hole population on the n side. In this equation, D_e is the free-electron diffusion constant, α is the absorption constant for the wavelength of the incident light, N_0 is the photon flux of that wavelength incident on the surface of the planar PV cell, and τ is the minority-carrier lifetime.

These same equations with $N_0 = 0$ were used by Shockley to derive the I–V relations for a p-n junction diode under forward and reverse bias conditions. He showed that the current flowing through the junction I_j with a voltage V_j applied across it was

$$I_j = I_0[\exp(qV_j/kT) - 1]. \tag{6}$$

The value of I_0 depends on semiconductor properties, junction parameters, and the boundary conditions appropriate to the particular cell. If the width w of each side of the junction is greater than the respective minority-carrier diffusion lengths (L_e and L_h), then

$$\begin{aligned} I_0 &= I_{0n1} + I_{0p1} \\ &= q[D_e n_p/L_e + D_h p_n/L_h], \end{aligned} \tag{7}$$

where I_{0n} and I_{0p} are the contributions to I_0 from the n and p sides, respectively; n_p and p_n are the minority electron concentration on the p side and minority hole concentration on the n side, respectively; and $L_e \equiv D_e\tau_p$ and $L_h = D_h\tau_n$. The term I_0 is proportional to n_i^2 because $n_p p_p = p_n n_n = n_i^2$, and therefore

$$I_0 = qn_i^2[D_e/N_A L_h + D_h/N_D L_e], \tag{8}$$

because in extrinsic semiconductors, $N_A \sim p_p$ and $N_D \sim n_n$; N_A is the acceptor concentration on the p side and N_D the donor concentration on the n side of the junction. Since $n_i^2 \propto \exp(-E_G/kT)$ we can write

$$I_0 = K(T)\exp(-E_G/kT), \tag{9}$$

where E_G is the energy gap and $K(T)$ is a weak function of absolute temperature T.

As we have already shown, when a PV cell is exposed to ionizing radiation, it acquires a forward bias (Fig. 6). The current generated by the ionizing radiation, I_{sc}, flows in the direction of a reverse current. From the principle of superposition, the equation for the current I_L through a load resistor R connected across the PV cell as shown in Fig. 7 is

$$I_L = I_j - I_{\text{sc}}, \tag{10}$$

where I_j is the junction current and I_{sc} is the current that flows through the short-circuited PV cell and is of course the current generated by the ionizing radiation. Substitution of Eq. (6) into Eq. (10) leads to

$$I_L = I_0[\exp(\Lambda V_L - 1)] - I_{\text{sc}}, \tag{11}$$

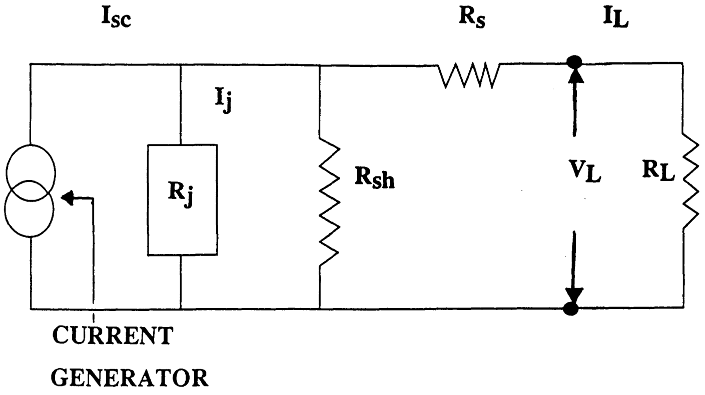

FIG. 7. Equivalent circuit of an illuminated solar cell. R_j is the nonlinear junction impedance, R_{sh} and R_s are the internal shunt and series resistances, respectively, and R_L is the load resistance.

where $\Lambda \equiv q/AkT$. The parameter A is equal to unity for the ideal Shockley diode [Eq. (6)]. However, in general $2 < A \leq 1$.

Figure 6(a) shows the I–V characteristic of a PV cell in the dark and Fig. 6(b) shows the I–V characteristic of the same cell under illumination. The dark I_j–V_j curve has a forward branch in the first quadrant where both voltage and current are positive and a reverse branch in the third quadrant where both the voltage and current are negative. Power (the product of current and voltage) is therefore consumed by the junction in either the forward or reverse directions. When the cell is illuminated, part of the characteristic drops into the fourth quadrant where the voltage is positive (a forward voltage) and the current is negative (a reverse current). Their product (the power) is negative, which means that the device is generating power. Figure 6(b) also shows the open-circuit voltage V_{oc}, the short-circuit current I_{sc}, the maximum power point P_{max}, the current at maximum power, I_{mp}, and voltage at maximum power, V_{mp}.

The expression for the open-circuit voltage V_{oc} is derived by setting $I_L = 0$ in Eq. (11). The result is

$$\exp(\Lambda V_{oc}) = (I_{sc} + I_0)/I_0. \tag{12}$$

The equivalent circuit of the PV cell (Fig. 7) must include an internal series (R_s) and an internal shunt (R_{sh}) resistance so that the complete equation for I_L assumes the form

$$\begin{aligned} I_L = {} & I_0[\exp\Lambda(V_L - I_LR_s) - 1] \\ & - I_{sc} - (V_L - I_LR_s)/R_{sh}. \end{aligned} \tag{13}$$

Observations on "good" PV cells lead to the conclusion that R_{sh} is usually large enough so that its effect can be neglected and R_s is small enough so that it can be neglected in a first-order calculation of the maximum power parameters. With these approximations ($R_s = 0$, $R_{sh} = \infty$) and applying the impedance-matching condition that must be satisfied for maximum power transfer from the PV cell to the load, i.e.,

$$R_{Lmp} = \left(\frac{\partial V}{\partial I}\right)_{mp}, \tag{14}$$

where R_{Lmp} is the load resistance for maximum power (mp) transfer and $(\partial V/\partial I)_{mp}$ is the dynamic impedance at the maximum power point, we can calculate the "mp" parameters. Algebraic manipulation leads to the following expressions for the load current at maximum power I_{Lmp} and the load voltage at maximum power V_{Lmp}:

$$\begin{aligned} (1 + \Lambda V_{Lmp}) \exp(\Lambda V_{Lmp}) &= I_{sc} + I_0/I_{sc} \\ &= \exp(\Lambda V_{oc}), \end{aligned} \tag{15}$$

$$I_{Lmp} = I_{sc}\Lambda V_{Lmp}/(1 + \Lambda V_{Lmp}). \tag{16}$$

These equations contain only three parameters that the PV cell fabricator can hope to control, i.e., I_{sc}, I_0, and the factor A (recall that $\Lambda \equiv q/AkT$). Their values are determined by the choice of semiconductor and PV cell parameters.

The maximum conversion efficiency η_{max}, i.e., the efficiency for matched load conditions, is given by

$$\begin{aligned}\eta_{\max} &= \frac{\text{Maximum Power Out}}{\text{Power In}} \\ &= \frac{I_{L\mathrm{mp}}V_{L\mathrm{mp}}}{P_{\mathrm{in}}} && (17) \\ &= \frac{\Lambda V_{L\mathrm{mp}}}{(1+\Lambda V_{L\mathrm{mp}})} \\ &\quad \times \frac{\Lambda V_{L\mathrm{mp}}}{\Lambda V_{L\mathrm{mp}} + \ln(1+\Lambda V_{L\mathrm{mp}})}\,\frac{I_{\mathrm{sc}}V_{\mathrm{oc}}}{P_{\mathrm{in}}} && (18) \\ &\equiv (CF)I_{\mathrm{sc}}V_{\mathrm{oc}}/P_{\mathrm{in}}, && (19)\end{aligned}$$

where CF is defined as the curve factor. The curve factor is the ratio of the area of the power rectangle $I_{L\mathrm{mp}}\ V_{L\mathrm{mp}}$ to the area of the $I_{\mathrm{sc}}\ V_{\mathrm{oc}}$ rectangle (see Fig. 6). Obviously it is always less than unity. In "good" cells, it can attain values close to 0.90.

2.2 Calculation of the Short-Circuit Current I_{sc}

The short-circuit current generated as a result of the absorption of ionizing radiation is calculated (Loferski and Wysocki, 1961) from the ambipolar diffusion equation [Eq. (5)] with two boundary conditions:

1. The junction is a sink for carriers, and therefore the excess carrier concentration at the edge of the space-charge region (SCR) is zero, i.e.,

$$\Delta n_{\mathrm{SCR}} = 0. \qquad (20)$$

2. At the outer surfaces of the PV cell,

$$(D_e\,\mathrm{grad}\Delta n)_{\mathrm{surf}} = s n_{\mathrm{surf}}, \qquad (21)$$

where s is the surface recombination velocity.

The surface recombination velocity is a parameter that is inversely proportional to the recombination lifetime of minority carriers at the surface. If $s = 0$, the minority carriers do not recombine at the surface and are "reflected" back into the bulk region. When $s = 0$, the surface is a minority-carrier "mirror." The parameter s will be zero if there are no recombination centers at the surface or if there are surface charges that generate an electrostatic field of such a polarity that the minority carries are repelled from the surface. When the surface recombination velocity is infinite, all minority carriers arriving at the surface recombine with carriers of the opposite sign. For s to be infinite, the surface concentration of recombination centers must be infinite or else there are surface charges of such a polarity that they give rise to an electrostatic field that drives minority carriers toward the surface where they recombine at surface recombination centers. From Eq. (21), it is clear that when $s = 0$, $\mathrm{grad}\Delta n$ at the surface is zero. As s approaches infinity, n_{surf} approaches zero, and $\mathrm{grad}\Delta n$ at the surface remains finite. For intermediate values of s, both n_{surf} and $\mathrm{grad}\Delta n$ at the surface have finite values.

The solution of the ambipolar diffusion equation for an n "skin" on a p-base cell in which ionizing radiation is incident on the $x = 0$ surface, the p-n junction is located at $x = l$, and the cell thickness is b is divided into two parts. The first of these is the contribution to I_{sc} from the n region and has the form

$$\begin{aligned}I_p(l) = \frac{qN_0(1-R)}{\alpha(1-\nu_p^2/\alpha^2)}&[\beta_p\nu_p\exp(\nu_p l) \\ &- \gamma_p\nu_p\exp(-\nu_p l)\alpha e^{-\alpha l}], \qquad (22)\end{aligned}$$

where N_0 is the flux of photons, characterized by having an absorption constant α, incident on the cell surface $x = 0$ and R is the fraction of these photons which are reflected, and the following definitions apply:

$$\beta_p\Delta_p = \exp(-\nu_p l)(h+\alpha) - e^{-\alpha l}(h+\nu_p); \qquad (23)$$

$$\gamma_p\Delta_p = e^{-\alpha l}(h-\nu_p) - \exp(\nu_p l)(h+\alpha); \qquad (24)$$

$$\Delta p = \exp(\nu_p l)(h+\nu_p) - \exp(-\nu_p l)(h-\nu_p); \qquad (25)$$

$$h = s/D_p, \qquad (26)$$

and

$$\nu_p = (D_p\tau_p)^{-1/2} = L_p^{-1}. \qquad (27)$$

Similarly, the contribution to the current from the p base is

$$\begin{aligned}I_n(l) = \frac{qN_0(1-R)}{\alpha(1-\nu_n^2/\alpha^2)}&[\beta_n\nu_n\exp(\nu_n l) \\ &- \gamma_n\exp(-\nu_n l) - \alpha e^{-\alpha l}], \qquad (28)\end{aligned}$$

$$\begin{aligned}\beta_n\Delta_n &= \alpha\exp[-(\nu_n l + \alpha b)] \\ &\quad - \nu_n\exp[-(\nu_n b + \alpha l)]; \qquad (29)\end{aligned}$$

$$\gamma_n \Delta_n = -\alpha \exp(\nu_n l - \alpha b) - \nu_n \exp(\nu_n b - \alpha l); \tag{30}$$

$$\Delta_n = 2\nu_n \cosh(\nu_n b - \nu_n l); \tag{31}$$

and

$$\nu_n = (D_n \tau_n)^{-1/2} = L_n^{-1}. \tag{32}$$

The total short-circuit current is

$$I_{sc} = I_p(l) + I_n(l). \tag{33}$$

It is convenient to define a collection efficiency for each value of α (i.e., for each wavelength of the incident light) as follows:

$$Q_{total}(\alpha) = \frac{I_p(l, \alpha)}{I_{max}(\alpha)} + \frac{I_n(l, \alpha)}{I_{max}(\alpha)} = Q_p(\alpha) + Q_n(\alpha), \tag{34}$$

where $I_{max}(\alpha)$ is the current that would have been produced if every photon with absorption constant α contributed one carrier to $I_{sc}(\alpha)$. Thus,

$$I_{max}(\alpha) = qN_0(\alpha)[I - R(\alpha)]. \tag{35}$$

Figures 8 and 9 show the results of calculations of $Q_p(\alpha)$ and $Q_n(\alpha)$ for various values of the cell parameters. The results of the calculations shown in Fig. 8 correspond to the case of a low surface recombination (about 100 cm/s) on the face $x = 0$. The parameter $\nu_p l = l/L_p$, where L_p is the hole diffusion length, is varied from 0.1, which corresponds to a cell in which the junction depth l is small compared to the minority-carrier diffusion length L, to $\nu_p l = 6.0$, a cell in which the inverse is true. If the calculation is conducted for higher values of $h = s/D$, i.e., higher values of surface recombination velocity, the values of Q_p are reduced for all values of α. Figure 9 shows the results of similar calculations for base collection efficiency Q_n as a function of α for various values of l/L_n with a fixed value for b (=0.5 cm). As expected, Q_n increases as the diffusion length in the base increases. Figure 10 shows $Q_{total} = Q_n + Q_p$ for a particular set of values of h, νl, and b.

Data like those presented in Figs. 8, 9, and 10 can be transformed into spectral response curves for specific semiconductors by transforming the abscissa to wavelength or photon energy with the help of curves of α vs λ like those shown in Fig. 1. If this is done for a Si cell having specific values of cell parameters, the result is a family of curves like those shown in Fig. 11, where νl in the diffused "skin" is varied. In the particular case chosen for Fig. 11, the surface recombination velocity $s = 0$ and so, for small values of l/L_p,

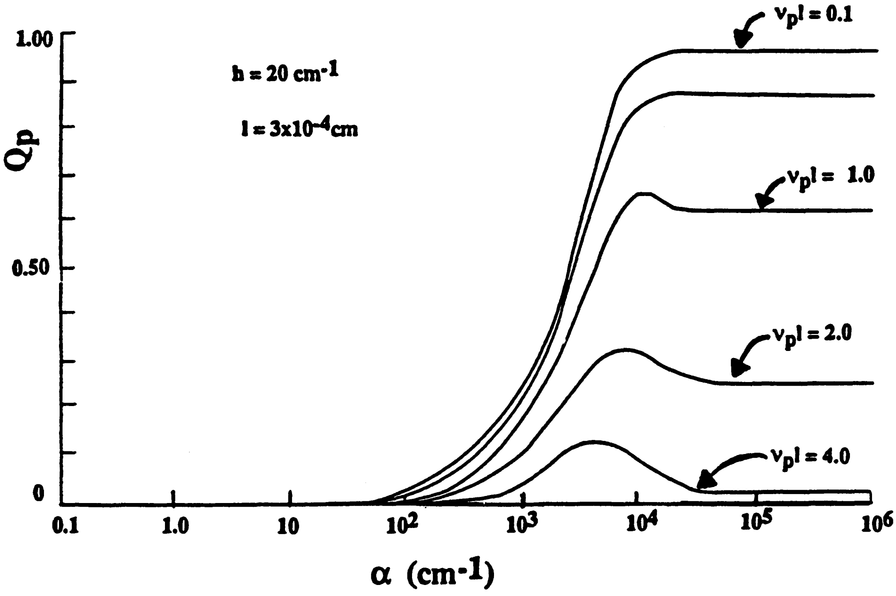

FIG. 8. Collection efficiency Q_p of the diffused *p*-type skin of a PV cell vs α for different values of νl ($n = L^{-1}$).

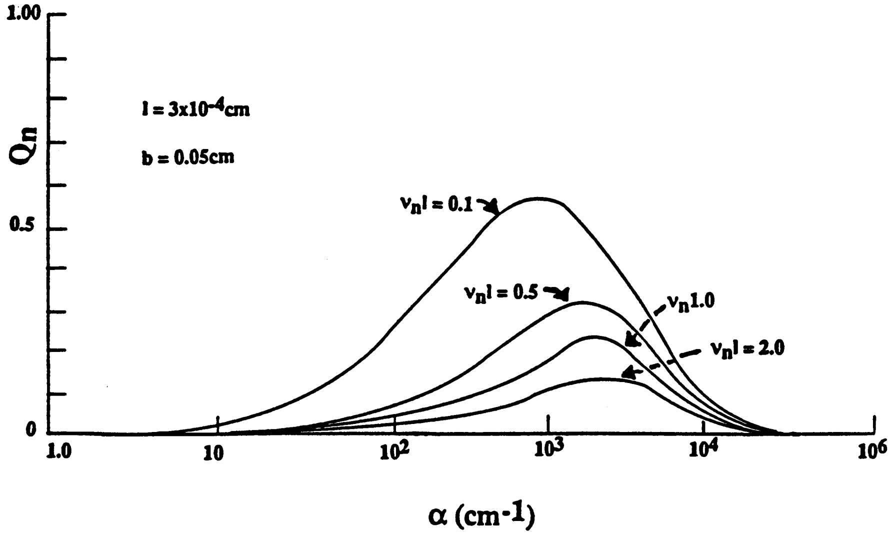

FIG. 9. Collection efficiency Q_n of the n-type base region of a PV cell vs α for different values of nl.

the total collection efficiency approaches unity value for short-wavelength radiation.

2.3 The Reverse Saturation Current I_0 and the Junction Quality Factor A

Most real junctions do not behave as predicted by the ideal Shockley diode equation, i.e.,

$$I_j = I_0[\exp(qV_j/kT) - 1], \tag{36}$$

where I_j is the current through the junction and V_j is the voltage across it. Rather their I_j–V_j characteristic can be approximated by the following expression:

$$I_j = \Sigma_i I_{0i}[\exp(\Lambda_i V_j) - 1], \tag{37}$$

where

$$\Lambda_i = q/A_i kT, \tag{38}$$

and A_i is the diode quality factor (Loferski, 1979). For the ideal diode ($i = 1$), $A_1 = 1$, and I_0 is given by Eq. (8).

A fundamental assumption underlying the derivation of the ideal diode equation and of the corresponding expression for I_0 is that there is no loss of carriers as they traverse the space-charge region. If this restriction is

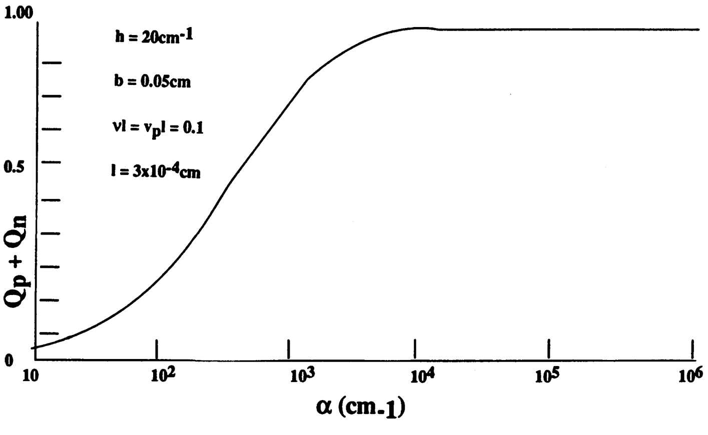

FIG. 10. Total collection efficiency $Q_p + Q_n$ vs α for $\nu l = 0.1$ in both the skin and base regions and with a low surface recombination velocity on the illuminated surface of the p-type skin.

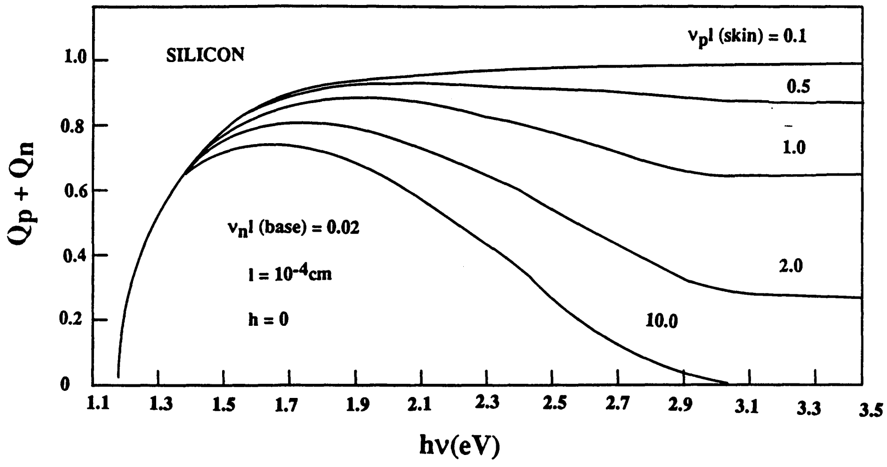

FIG. 11. Family of silicon PV-cell spectral response curves ($Q_p + Q_n$) vs photon energy for different values of $l/L_p = \nu_p l$ in the skin and a large diffusion length in the base.

removed and recombination and generation of carriers is assumed to occur in the space-charge region at recombination centers having their energy levels in the middle of the forbidden energy gap, then the dependence of I_0 on n_i can be shown to have the form

$$I_{02} = K_2(T)n_i. \tag{39}$$

Therefore,

$$I_{02} \propto \exp(-E_G/2kT) \tag{40}$$

and

$$I_{j2} = I_{02}[\exp(qV/2kT) - 1]. \tag{41}$$

More generally,

$$I_{0i} = K_i(T)\exp(-E_G/B_ikT) \tag{42}$$

and

$$I_{ji} = I_{0i}[\exp(qV/A_ikT) - 1]. \tag{43}$$

In the ideal diode case $A_1 = B_1 = 1$, while in the recombination–generation case $A_2 = B_2 = 2$. In real diodes, $2 \le A \le 1$ and $2 \le B \le 1$.

It is easy to show that a PV cell that obeys the ideal diode equation ($A = B = 1$) should have a higher value of $\eta_{\max}$ than a PV cell with $A = B = 2$. The short-circuit current I_{sc} is not affected by the values of A and B and, therefore, will be the same in either cell. There is, however, a difference in V_{oc} values for the two cases, namely

$$V_{oc2} - V_{oc1} = (kT/q)\ln(I_{01}/I_{02}). \tag{44}$$

Since I_{02} is always greater than I_{01}, V_{oc1} is always greater than V_{oc2}. Consequently, according to Eq. (19), $\eta_{\max 1} > \eta_{\max 2}$. It is, therefore, a goal of PV cell technology to produce junctions whose I_j–V_j characteristics are governed by the ideal diode equation.

There is evidence that ideal diode behavior is more likely to be observed in *p-n* junctions made from smaller-bandgap semiconductors like germanium (0.7 eV) than in larger-bandgap semiconductors like GaAs (1.35 eV). The bandgap of silicon (1.1 eV) places it in a transition region between low- and high-E_G materials; consequently, it has been difficult but not impossible to produce silicon cells having ideal diode behavior and, therefore, exhibiting the highest $\eta_{\max}$ of which silicon is capable.

2.4 Maximizing $\eta_{\max}$

The equations governing the PV effect provide guidance on extracting the highest efficiency from a PV cell made from any semiconductor. Most important of all, they show that both surface and bulk recombination rates must be minimized. Minimizing surface losses means passivation of the surface. In the case of silicon, passivation is achieved by covering the outer surfaces with a silicon oxide that has a low interface state density and/or that produces a band bending at the surface in such a direction that light-generated minority carriers are repelled from the surface region.

A more generally applicable method of reducing front-surface recombination losses is to employ the wide-bandgap window concept. The idea is to cover the photovoltaically active semiconductor (PVAS), i.e., the absorber of the ionizing radiation, with a different semiconductor whose bandgap is so large that none of the photons in the light source are absorbed in it. This can result in a heterojunction device like that shown in Fig. 12. There is an added small benefit to efficiency in that the contribution from the wide-E_G semiconductor to I_0 will be orders of magnitude lower than the contribution to I_0 from the PV-active material. Lower I_0, of course, leads to somewhat higher V_{oc} and η_{max}.

Minimizing bulk recombination means minimizing the concentration of bulk recombination centers throughout the light-absorbing volume of the PV cell. In the case of an indirect-gap material like silicon, the concentration of elements like Cu and Fe and defects like vacancies and interstitial atoms, all of which are effective recombination centers, must be minimized. Concentration of these defects must be held to values in the range 10^{12}–$10^{13}/\text{cm}^3$. Such defects can be introduced during crystal growth, during growth of thin films, or during preparation of the cell. As noted above, highest efficiency requires A values of unity. This in turn requires minimizing recombination-center concentration in the space-charge region.

If the surfaces that serve as boundaries of the PV-active region of the cell have zero surface recombination velocity, the value of I_0 calculated from the one-dimensional ambipolar diffusion equations for a cell in which $w \ll L$ has the value (Loferski, 1988)

$$I_0 = I_\infty \tanh(w/L), \tag{45}$$

where w is the distance between the edge of the space-charge region and the edge of the cell and I_∞ is the value of reverse saturation current for a cell in which $w \gg L$.

Since $\tanh x \rightarrow x$ as $x \rightarrow 0$,

$$I_0 \rightarrow I_\infty(w/L). \tag{46}$$

Thus, I_0 decreases as the thickness w decreases. But as w decreases, the photon-absorbing volume decreases and so does I_{sc}. This situation can be remedied by resorting to light trapping, i.e., by making the internal surfaces of the cell optically reflecting. Figure 13 shows how the cell efficiency depends on w for $s = 0$ and $s = \infty$ with 100% efficient optical reflection $R = 1$ and with zero internal reflection $R = 0$. Clearly the goal of the designer is to fabricate cells that approach the case $s = 0$ and $R = 1$, and indeed the highest-performance solar cells approach such values for s and R.

2.5 Dependence of η_{max} on E_G

Up to this point, the discussion has been focused on the performance of PV cells made from a given semiconductor. Clearly, the forbidden energy gap E_G has a first-order effect on the maximum efficiency that can be extracted from a semiconductor (Loferski, 1956). As E_G increases, I_{sc}, which is given by the

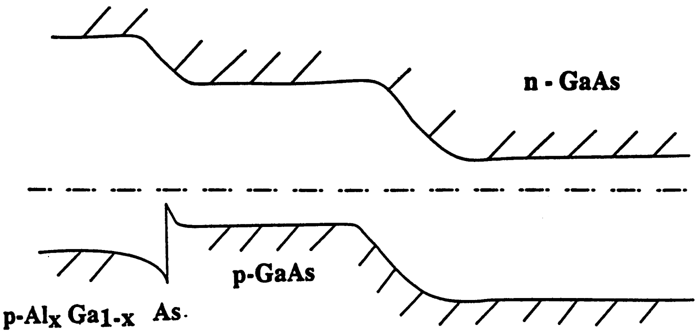

FIG. 12. Energy-band diagram of the p-Al$_x$Ga$_{(1-x)}$As, p-GaAs/n-GaAs solar cell.

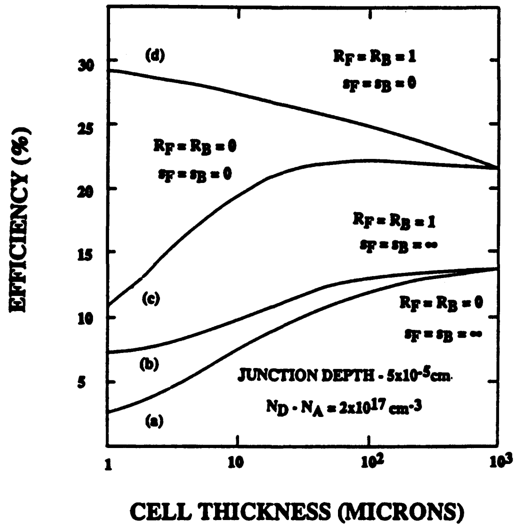

FIG. 13. Solar-energy conversion efficiency vs cell thickness calculated for a silicon solar cell for the four combinations of two extreme values of surface recombination velocity ($s = 0$ and $s = \infty$) and of internal reflectance ($R = 0$ and $R = 1$).

equation

$$I_{sc} = \int_{E_G}^{\infty} Q(h\nu)n_{ph}(h\nu)d(h\nu), \tag{47}$$

decreases. For the solar spectrum, it turns out that

$$I_{sc} = I_{sc0}\exp(-K_E E_G). \tag{48}$$

The value of I_{sc0} is the value of I_{sc} if every photon in the incident solar spectrum contributed a carrier to I_{sc}. The term K_E has a value between 1.5 and 2, depending on the distribution of photons in the spectrum.

The value of I_0 also decreases with E_G because

$$I_0 = K_i(T)\exp[-(E_G/B_i kT)]. \tag{49}$$

At $T = 300$ K and with $B = 1$,

$$I_0 \sim K_1(300\text{ K})\exp(-40E_G). \tag{50}$$

Now, Eq. (12) can be written in the form

$$V_{oc} = \frac{A_i kT}{q}\ln\left(\frac{I_{sc}}{I_0}\right). \tag{51}$$

For $A = 1$ and 300 K, and with $K_E = 2$,

$$V_{oc}(E_G) \sim 0.025[38E_G + \ln(I_{sc0}/K_1)]. \tag{52}$$

So V_{oc} increases as E_G increases. Therefore, since

$$\eta_{max} = I_{sc}V_{oc}(CF)/P_{in}, \tag{53}$$

η_{max} passes through a maximum.

Figure 14 shows η_{max} vs E_G for the Air Mass Zero (AM0) solar spectrum for two cases: $A = B = 1$, i.e., the ideal diode case, and $A = B = 2$, i.e., the case of recombination-generation in the SCR. (The AM0 solar spectrum is the spectrum above the earth's atmosphere.) In generating these curves, the value

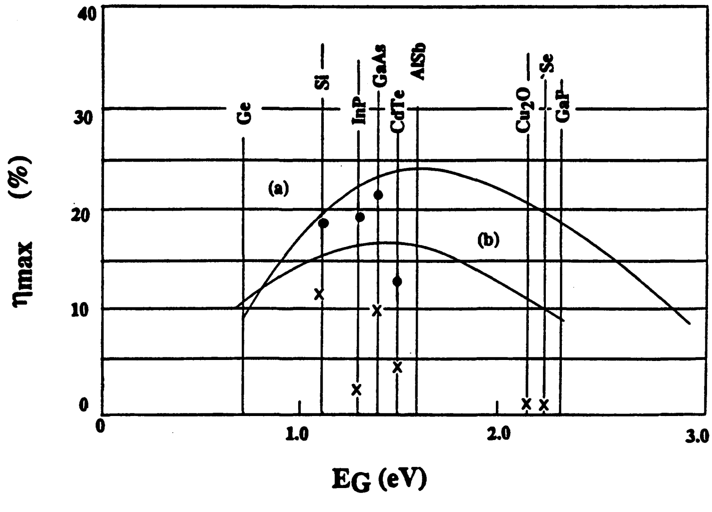

FIG. 14. Efficiency η_{max} vs energy gap E_G for solar cells exposed to AM0 solar spectrum for (a) an ideal p-n junction ($A = B = 1$) and (b) a p-n junction including recombination generation in the space-charge region ($A = B = 2$). The crosses indicate the best experimental values of η_{max} when this theoretical curve was constructed (Loferski, 1956) and the solid dots show the best experimental values in 1994.

of I_0 was calculated for Si for these two cases and it was assumed that the values of K_i are the same as in Si for all other semiconductors. This of course is not true; a separate calculation of K_i should be made for each semiconductor. However, the values of K_i do not vary by large amounts from semiconductor to semiconductor and so the curves of $\eta_{\max}$ vs E_G based on the indicated assumptions correctly predict the gross dependence of $\eta_{\max}$ on E_G.

The major conclusions of this exercise are as follows:

1. The maximum efficiency for solar energy conversion for a single-junction solar cell is around 26%.
2. The best semiconductors for PV solar cells have bandgaps between 1.0 and 2.2 eV with the highest $\eta_{\max}$ occurring around 1.5 eV for AM0 illumination.
3. Cells in which $A = B = 1$ have much higher potential efficiencies than cells with $A = B = 2$.
4. Although the data of Fig. 14 do not show it, the shape of the curves of $\eta_{\max}$ vs E_G changes as the solar spectrum changes.

2.6 Ultimate Efficiency Calculation

The $\eta_{\max}$ vs E_G calculations presented above are based on empirical values of the parameters of semiconductors and of the *p-n* junctions made from them. An important result of these calculations is that qV_{oc} can never be equal to E_G and the curve factor is less than unity. If, however, one assumes that $qV_{oc} = E_G$, which is the maximum imaginable value qV_{oc} can attain, that $CF = 1$, and that every absorbed photon contributes one carrier to I_{sc}, then the efficiency for a cell made from a semiconductor with energy gap E_G would be

$$\eta(E_G) = E_G N_G / E_{\text{total inc}}, \tag{54}$$

where

$$N_G = \int_{h\nu_G = E_G}^{\infty} n(h\nu)d(h\nu); \tag{55}$$

i.e., N_G is the total number of photons having energy equal to or greater than E_G and $n(h\nu)d(h\nu)$ is the photon distribution in the spectrum. The total incident energy is given by

$$E_{\text{total inc}} = \int_0^{\infty} h\nu n(h\nu)d(h\nu). \tag{56}$$

If the solar spectrum is approximated by the spectrum of a 6000-K blackbody, the peak in the curve of η vs E_G calculated from Eq. (54) lies at $E_G \sim 1.1$ eV with $\eta(1.1 \text{ eV}) = 44\%$ (Trivich and Flinn, 1955).

2.7 Higher Efficiencies through Spectrum Splitting

The maximum imaginable energy that a light-generated carrier can deliver to a load in a PV cell is the energy gap E_G. The energy in excess of E_G delivered to the cell by photons with $h\nu > E_G$ is converted into heat inside the cell. This loss of photon energy can be reduced by splitting the spectrum among a stack of cells made from semiconductors having different energy gaps. For example, suppose the smallest bandgap of an absorber semiconductor in a PV cell stack is E_{G1}, and suppose further that the solar spectrum incident on this cell produces a current I_{sc1}. Construct a stack of N cells whose energy gaps are chosen so that each cell in the stack absorbs such a number of photons that the value of $I_{scN} = I_{sc1}/N$. Then the conversion efficiency of the stack will be higher than the conversion efficiency of any one cell in the stack acting alone because the fraction of energy deposited by absorbed photons that is converted into heat (i.e., all the energy of the photon in excess of E_G) is less in the stack than in a single cell. For a three-cell stack with $E_G = 0.94$, 1.34, and 1.91 eV and the AM0 solar spectrum, the theoretical efficiency is about 37%, a value considerably higher than the 26% calculated for a single cell made from the semiconductor having the "optimum" bandgap.

The upper limit on the conversion efficiency of such a stack of PV cells has been the subject of much discussion in the literature. Calculations of the efficiency of a stack of cells with absorber bandgaps varying from 0 to ∞ has shown that for the case of maximum possible concentration of sunlight with lenses or mirrors, namely a concentration ratio of 46 747, the limiting efficiency is $\eta_L \sim 86\%$. The concentration ratio C is the ratio of radiant power per unit area delivered to the PV cell by the concentrator (a lens or a mirror) to the radiant power per unit area that the PV cell receives from the sun if no concentrator is used, in which case $C = 1$. The concentrator focuses all of the sun's rays that it intercepts into an image of the sun. If the cell is a disk and it is placed in the focal plane and the image of the sun just covers the cell, then C is equal to the ratio of the geometrical area of the concentrator or aperture to the area of the solar image it produces, which, in this example, is equal to the area of the PV cell. The maximum attainable value of C is the ratio of the maximum solid angle aperture of a concentrating mirror that is focusing the solar image onto a cell of the same dimension and shape as the image (this solid angle is π steradians) to the solid angle subtended by the solar disc [$\pi \sin^2(\frac{1}{2}\theta)$ steradians; $\frac{1}{2}\theta = 15.9$ minutes]. In this case, $C_{\max} = 46\,747$. For the case of no concentration, $\eta_L \sim 68\%$ (Landsberg *et al.*, 1991). The efficiency is higher as the concentration is increased because the value of I_{sc} increases with light intensity while the voltage increases as the logarithm of the light intensity. Figures 15 and 16 show the results of empirical calculations of the efficiency of a finite number of cells based on the model of PV cells used to calculate the curves of $\eta_{\max}$ vs E_G of Fig. 14 for the case $A = B = 1$ (Gocsen and Loferski, 1979). For this calculation, the bandgaps of the cells were selected so that each of them would absorb the same number of photons and, therefore, generate the same I_{sc}. Figure 15 shows the results for unconcentrated AM0 sunlight and Fig. 16 for a concentration C of 1000. For example, the efficiency of a 26-cell optimized stack at 300 K is about 48% for $C = 1$ and about 59% for $C = 1000$. Note that as T increases, η decreases.

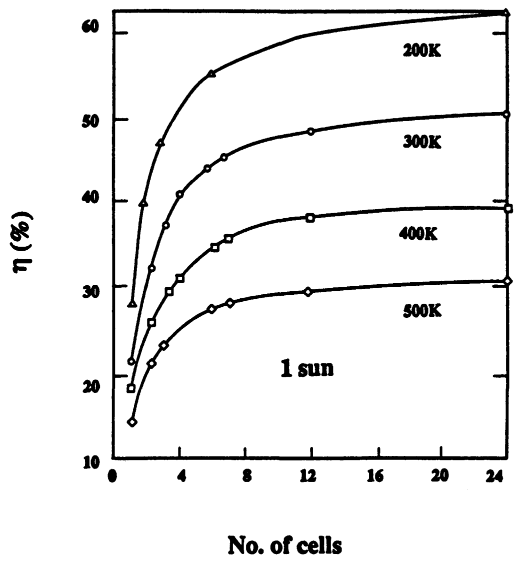

FIG. 15. Calculated optimized efficiencies at various temperatures for tandem PV cell stacks consisting of N solar cells with bandgaps of PV active semiconductors selected so that $I_{sc1} = I_{sc2} = I_{scN}$. AM0 spectrum, no concentration, i.e., $C = 1$.

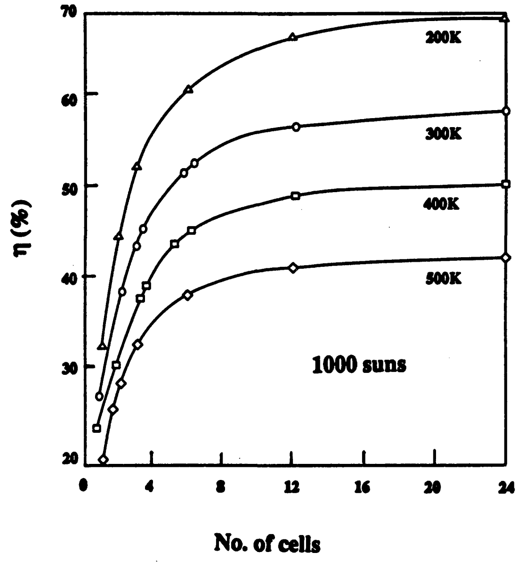

FIG. 16. Same as Fig. 15 except that concentration $C = 1000$.

These calculations do not take into account losses, which are unavoidable in tandem cell systems or for that matter in any PV cell. For example, there will always be an internal resistive loss ρ, a grid transmissivity loss γ, and a reflection loss β. In a three-cell stack, if we assume that each of these loss factors is only 2%, a very optimistic value, then the accumulated fractional loss F_L for a three-cell stack would be (Bennett and Olsen, 1978)

$$F_L = \prod_{1}^{3} (1-\rho)(1-\gamma)(1-\beta) = 0.83, \qquad (57)$$

i.e., the efficiency of the third cell in the stack would be 83% of the lossless value. For more realistic loss factors of 0.05, $F_L = 0.63$. Such losses would limit the number of cells in a stack to perhaps four cells.

2.8 Dependence of $\eta_{\max}$ on Temperature

The open-circuit voltage V_{oc} is a strong function of temperature and, therefore, so is $\eta_{\max}$ (Wysocki and Rappaport, 1960). From Eqs. (12), (42), and (48),

$$V_{oc} = \frac{AkT}{q}\left[\frac{E_G}{BkT} - E_G K_E + \ln(I_{so0}/K_i)\right] \quad (58)$$

$$\approx \frac{AE_G}{Bq} - \frac{AkT}{q}[K_E E_G + \ln(I_{so0}/K_i)]. \quad (59)$$

It turns out that I_{sc0} is always larger than K_i and, therefore, V_{oc} always decreases as T increases. Figure 17 shows a family of curves of $\eta_{\max}$ vs E_G calculated for the AM0 solar spectrum using the same assumptions as those used to establish the curves of $\eta_{\max}$ vs E_G of Fig. 14 with $A = B = 1$. The curves of Fig. 17 show that $\eta_{\max}$ of all cells decreases as T increases and that the peak in the curves of $\eta_{\max}$ vs E_G shifts toward larger E_G as T increases.

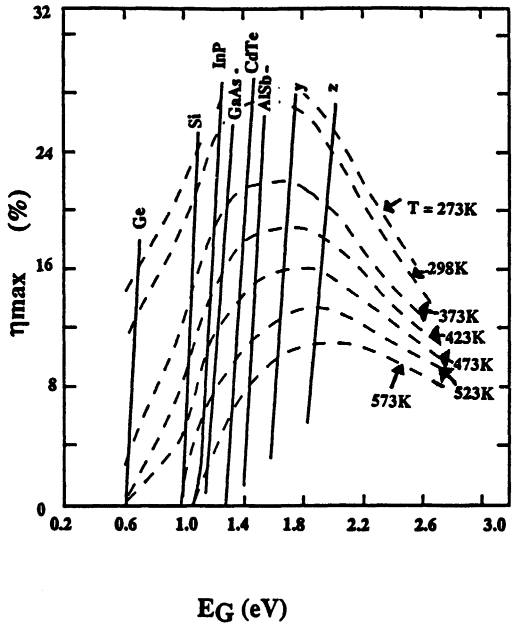

FIG. 17. Efficiency $h_{\max}$ vs energy gap E_G as a function of temperature.

3. SEMICONDUCTOR MATERIALS USED IN PV CELLS: HISTORY AND CURRENT STATUS

The calculations of $\eta_{\max}$ vs E_G show that the best semiconductors for PV cells have energy gaps in the range 1.0 eV $\leq E_G \leq$ 2.2 eV and that the maximum efficiency to be expected is around 26%. To attain the highest efficiencies, the minority-carrier diffusion lengths in the photovoltaically active volume (PVAV) of the cell must be as large as possible. Furthermore, the surface recombination velocity must be vanishingly small on the surfaces of the PVAV.

Photovoltaic cells have been fabricated from many materials having bandgaps in the interesting range. The most successful of these materials have crystal lattice structures related to the diamond lattice, which consists of two interpenetrating face-centered cubic lattices displaced with respect to each other by one-fourth of a lattice constant in each of the three mutually perpendicular directions. These materials include silicon; the III-V semiconductor compounds [gallium arsenide (GaAs), indium phosphide (InP)]; III-V semiconductor alloys thereof like $Ga_xAl_{1-x}As$, $GaAs_yP_{1-y}$, etc.; certain II-VI semiconductor compounds [cadmium telluride (CdTe); cadmium sulfide (CdS); zinc telluride (ZnTe)] and their alloys; and the chalcopyrite semiconductors [copper indium diselenide (CISe), copper gallium diselenide (CGSe), copper indium disulphide (CISu)] and their alloys. Most of the cited materials have served as the constituent of the PVAV, but some of them, notably CdS, have served mainly as wide-bandgap windows or have other functions to perform in the cells. Other materials having different crystal lattices, like ZnP_2, WSe_2, CuO, and CuO_2, have also been explored, but none of them has performed as well as the best of the diamond-lattice materials. There is another important class of PVAV materials that holds promise, namely certain amorphous semiconductors, notably hydrogenated amorphous silicon and its "alloys" with germanium and carbon. The next few paragraphs are devoted to brief commentaries on solar cells made from these materials.

3.1 Silicon Monocrystalline and Polycrystalline Cells

The silicon cell has been the workhorse of the modern PV age. The most successful, most utilized, and oldest of the modern PV cells is the single-crystal silicon cell. In its original form (Chapin *et al.*, 1954), it consisted of a diffused boron-doped p-type "skin" region on a phosphorus-doped, n-type single-crystal base wafer. The p-n junction was located about 1 μm below the surface. As shown in Fig. 18, a gridded metallic Ohmic contact was deposited over the diffused region and the back surface was covered with a metallic Ohmic contact. The area of individual cells was (and continues to be) controlled by the diameter of silicon single-crystal ingots, which had a maximum value of about 3 cm in 1954 and about 15 cm and even 20 cm in 1994. The solar-energy conversion efficiency of the 1954 cell was around 10% for terrestrial sunlight. In 1993, the highest efficiency of crystalline cells having an area of 4 cm^2 was about 23% (Green *et al.*, 1987; Green and Emery, 1993). The cells that produce such efficiencies are of a design radically different from that of the simple 1953 cell. These high-performance cells incorporate surface passivation to reduce surface losses; light trapping to improve utilization of the incident photons; and high minority-carrier lifetime silicon material and cell fabrication techniques that preserve the lifetime. The PVAV has a relatively high resistivity; the heavily doped p and n regions are both located on the side opposite the one on which light is incident.

Special-design, high-efficiency single-crystal silicon cells have also been fabricated for concentrator applications. The highest efficiency reported for such cells is about 28% with $C \sim 200$. These cells had a small area (3×5 mm^2); they had a high-resistivity PVAV and "point-contact" p^+ and n^+ collector regions on the back of the cell (Verlinden *et al.*, 1993).

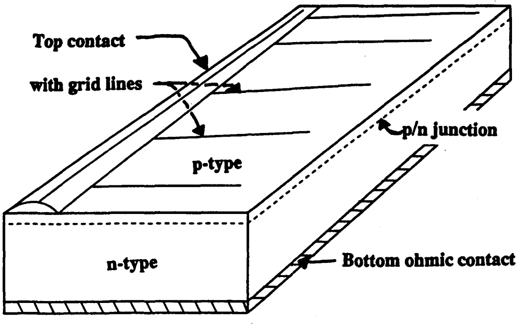

FIG. 18. Schematic representation of standard planar solar cell.

These concepts have been applied to polycrystalline silicon having grain volumes of at least several cubic millimeters. The best of the cells made from such polycrystalline silicon have exhibited solar-energy conversion efficiencies in excess of 17% in cells having an area of 4 cm^2 (Green and Emery, 1993).

3.2 Gallium Arsenide Cells

The first GaAs solar cells were reported in 1956 (Jenny *et al.*, 1956). They were prepared by diffusing cadmium into n-type single-crystal wafers. Their area was small (between 0.1 and 6.0 mm^2) because only small single crystals were available at that time, and the highest efficiency was 6.5%. By 1961, the size of GaAs single-crystal ingots had increased and processing had improved so that 10% efficient cells having areas up to 2 cm^2 were being fabricated.

A significant breakthrough in efficiency was achieved in 1972 by the invention of the wide-bandgap window concept and its application to GaAs (Woodall and Hovel, 1972). Analyses of extant GaAs cells showed that, in this direct-gap material, surface recombination losses exacted a high toll on excess carriers generated in the diffused skin. Incorporation of a wide-bandgap p-Ga$_x$Al$_{1-x}$As window (see Fig. 12) reduced surface losses dramatically and pushed the efficiency up to 15%. Further refinement of this type of cell has resulted in GaAs cells with efficiencies up to 25% in 1993.

Another approach to reducing surface losses is to make the diffused surface layer so thin that a negligible fraction of photons are absorbed in it (Gale *et al.*, 1984). This also has resulted in 23% efficient cells by 1993 (Green and Emery, 1993).

3.3 Cadmium Sulfide/Copper Sulfide

The 1954 CdS cell was made by depositing a thin layer of copper onto a cadmium sulfide single crystal; its actual solar-energy conversion efficiency was only about 1.5% (Reynolds *et al.*, 1954). By 1956, the efficiency of such cells had increased to about 5%. The first thin-film CdS cell was reported in 1954 and the attention of researchers was thereafter directed to the thin-film cell (Nadjakov *et al.*, 1954). At first it was thought that the cell was based on a p-n junction in CdS since it was observed that incorporation of copper into *n*-type CdS eventually resulted in a *p*-type material. However, by the mid-1960s it was clear that this cell was based on a heterojunction between *n*-CdS having a bandgap of 2.4 eV and a *p*-type copper sulfide having a bandgap around 1.0 eV. The copper sulfide was the photovoltaically active material; the CdS was serving as a wide-bandgap window. The Cu-S binary system includes many compounds having various crystal structures. Only the chalcocite form, whose chemical composition is close to Cu_2S, could produce good PV cells. However, the compound is not stable. Efficiencies up to 10% were achieved in *n*-CdS/*p*-Cu_xS cells, but the efficiency was not stable. Consequently, interest in this type of cell waned and the cell was practically abandoned.

3.4 *p*-Copper Indium Selenide/*n*-CdS Cells

CdS continues to play an important role in PV cell science and technology because of its role as the wide-bandgap window in the *p*-$CuInSe_2$ (CISe)/*n*-CdS cell. The first of these cells was made in 1974 by depositing a thin layer of CdS onto a small-area CISe crystal; its efficiency was 12% and it was stable (Shay *et al.*, 1975). By 1976, thin-film CISe/CdS cells with efficiencies up to 6% were made by vacuum deposition of the constituents (Kazmerski *et al.*, 1976). In 1981, a 1-cm^2-area, 10% efficient, stable CISe/CdS cell was reported (Mickelsen and Chen, 1980). It was made by simultaneous controlled-rate evaporation of Cu, In, and Se onto a substrate heated to a high enough temperature (about 500 °C) to react the elements so that CISe is formed. These cells have been further improved by using alloys of CISe with $CuGaSe_2$ and alloys of CdS with ZnS. In 1993, the efficiency of 1-cm^2-area cells of this type had reached 16.9% active area, 14.8% total area (Hedstrom *et al.*, 1993, 1994). In 1994, a total-area efficiency of 16.4% was reported (Gabor *et al.*, 1994). At the time of this writing, this is the highest confirmed efficiency for a thin-film polycrystalline cell measured under standard conditions.

CISe cells have been produced from electrodeposited films, by reactively sputtered films, and by sputtering from a previously synthesized CISe target. These deposition methods are less expensive than the simultaneous evaporation technique, but the cells produced by these methods have not as yet been as efficient as cells made by three-source evaporation.

3.5 Cadmium Telluride Cells

Cadmium telluride has a bandgap of about 1.45 eV, which is close to the peak of the curve of $\eta_{\max}$ vs E_G. It can be made *n* or *p* type so that *p*-*n* junction CdTe cells are possible. The first of these cells was made in 1954 but its efficiency was only around 2%. In 1963, a thin-film *p*-Cu_xTe/*n*-CdTe cell having $\eta \sim 6\%$ was reported (Cusano, 1963). Subsequent research showed that the most successful CdTe cells were heterojunction devices with wide-bandgap window semiconductors including CdS, ZnO, ZnS, and alloys of these materials. Values of $\eta_{\max}$ up to 15.8% (total area) have been achieved (Chu and Chu, 1993).

The films for CdTe cells have been made by close spaced sublimation ($\eta \sim 15.8\%$), electrodeposition (12%), screen printing (11.3%), vacuum evaporation (11%), and chemical spray pyrolysis (12.7%). These low-cost production methods make CdTe cells among the most promising candidates for large-scale PV power plants.

3.6 Amorphous Silicon Solar Cells

The invention of the amorphous Si solar cell at RCA Laboratories in 1974 unleashed a worldwide research effort on this promising device. The first 1974 cells had $\eta \sim 0.2\%$, but by 1976 devices using *a*-Si had reached $\eta \sim 5.5\%$ (Carlson and Wronski, 1976).

The amorphous silicon layers were prepared by electrodeless rf discharge in silane (SiH_4) so that the resulting layers were actually "alloys" between Si and H_2. The best solar cells were made from hydrogenated *a*-Si (*a*-Si:H) films containing between 10% and 20% H_2. The H_2 passivates the dangling bonds in the Si atoms with the result that the "forbidden energy gap" has a much lower density of states than a pure amorphous Si layer contains and, therefore, the material can be doped substitutionally with elements from column 3 or 5 of the periodic table. To make *p*-type layers, some borane (B_2H_6) is added to the SiH_4; to make *n*-type layers, some phosphene (PH_3) is added to the SiH_4. Consequently it is possible to make *p*-*n* junctions in *a*-Si:H. The best solar cells utilize a *p-i-n* structure.

The optical bandgap E_o of *a*-Si:H varies with H_2 content between 1.5 and 1.8 eV. The absorption constant rises rapidly from E_o to values around 10^5 cm^{-1} so that the layer thicknesses of the order of 1 μm absorb all the photons with energies in excess of E_o.

The optical bandgap can be increased by incorporating carbon and decreased by incorporating germanium. Therefore, tandem cell structures can be made from *a*-Si:H. By 1993, the best *a*-Si:H single-junction cells had initial η values of about 11%. Multijunction cells reached initial η values of about 12% (Green and Emery, 1993).

Unfortunately, *a*-Si:H cells have exhibited degradation during illumination. In the best cases, the decay of efficiency from its initial value is about 10%. However, in 1994, it was reported that after several thousand hours of operation under constant illumination by simulated sunlight, the efficiency of certain *a*-Si:H modules stabilized at a value slightly in excess of 10%.

3.7 Cells Based on Other Semiconductors

Many semiconductors have been explored as the PVA material in PV cells. They include InP, WSe_2, ZnSe, Zn_3P_2, $GaAs_xP_{(1-x)}$, organic semiconductors, etc.

The most notable in this group is indium phosphide (InP), which has a bandgap around 1.25 eV. The first InP cells were made in 1956; their efficiency was only about 2%. For the next few decades, InP was neglected because GaAs with its larger energy gap was a more promising material. Research on InP was revived in the late 1980s and cells with efficiencies up to 22% were prepared by epitaxial growth techniques in 1990. These InP cells proved to be extremely radiation-resistant high-efficiency cells, and, therefore, they may play a role in power supplies of future satellites.

There may be other materials which could produce high-performance solar cells, but the research effort that must be expended to realize the potential of each material is formidable. The dominance of Si, GaAs, and InP results, to a considerable degree, from the large effort expended on these materials for more commercially lucrative applications like integrated circuits and optoelectronics devices. The amount of research expended on materials whose only application appears to be in PV cells (like CISe and *a*-Si:H) is very modest and this impedes progress in the field.

4. EFFECTS OF HIGH-ENERGY RADIATION ON PV CELLS

From Eqs. (23) through (33), it is clear that the short-circuit current I_{sc} of an illuminated PV cell is a strong function of the minority-carrier diffusion lengths L_n and L_p in the *n* and *p* sides of the junction. The diffusion lengths in turn are functions of the respective minority-carrier lifetimes τ through relations of the type

$$L^2 = D\tau, \tag{60}$$

where D is the respective minority-carrier diffusion constant. From Eq. (2), τ is inversely proportional to the product of the concentration of recombination centers of interest, N_r; the capture cross section σ_r of the centers; and $f(E_r - E_F)$, their Fermi–Dirac occupancy factor, where E_r is the energy level in the forbidden energy gap associated with the particular recombination centers under consideration.

When a solid is exposed to irradiation by high-energy particles (electrons, protons, α particles, neutrons, γ rays, β rays, etc.), collisions between the high-energy particles and lattice atoms can lead to dislodging of lattice atoms resulting in the generation of vacancy-interstitial pairs (Frenkel defects), provided that the energy transferred to the lattice atom during the collision exceeds the threshold energy for the displacement, E_{th}. For silicon and germanium, E_{th} has a value of about 15 V (Loferski and Rappaport, 1958). This amount of energy can be delivered to a Si atom by an electron having an energy of about 130 keV or a proton having an energy of about 500 eV. The corresponding quantities for Ge, which has a higher mass than Si, are about 350 keV for electrons and about 1100 eV for protons.

In a semiconductor, radiation-generated vacancies and interstials act as recombination centers having energy levels lying between about 0.10 and about 0.25 eV from the band edges. Sometimes the vacancies and interstials form complexes with impurity atoms already present in the semiconductor. The resulting complexes have E_r values in the same 0.10- to 0.25-eV range, and therefore they too can serve as recombination centers. In either case, τ decreases as the concentration of recombination centers introduced by bombardment N_{rB} increases. If the lifetime before irradiation is τ_0, then after bombardment by N_{rB} particles

$$1/\tau = 1/\tau_0 + N_{rB}\sigma_{cB}f_B V. \tag{61}$$

The reason why these irradiation effects are important for PV solar cells is that the cells are used as primary sources of power on artificial earth satellites. In this application, they are exposed to irradiation by electrons and protons trapped in the radiation belts surrounding the earth. Indeed, the discovery of the so-called Van Allen radiation belts by the first satellite launched by the U.S. in 1960 raised questions about the longevity and, therefore, the usefulness of solar cells in space. The cells used on that first satellite were silicon *p* skin on *n*-base cells. Experiments with electron beams from accelerators showed that these cells would lose about 25% of their power output after about 1 year of exposure to the electrons in the Van Allen belts of low Earth orbits. When these irradiation experiments were repeated on silicon *n* skin on *p*-base cells, it was found that they were more tolerant of electron irradiation by almost an order of magnitude. Consequently, *n*-on-*p*-base cells became the silicon cell of choice for use in satellite power supplies. The search for even more radiation-resistant cells eventually led to the discoveries that GaAs and InP cells are substantially more tolerant of radiation than any of the cells made from silicon. They are also more efficient converters of AM0 radiation than silicon, and therefore they are strong candidates for eventually replacing silicon in space applications.

5. ENVIRONMENTAL ASPECTS OF PV CELLS

PV solar cells are completely silent, nonpolluting generators of electricity. They produce no smoke, no carbon dioxide, and no radioactive wastes and are very attractive from an environmental point of view.

One objection to their large-scale use is, however, that generation of large amounts of electricity would require covering large areas with solar panels because sunlight, the primary source of power, is so diffuse. At high noon at the equator, the density of sunlight normally incident on a receiver is about 1 kW/m^2. Over the course of 1 year, such a square meter of surface area would receive about 2 MWh of solar energy. If this solar energy is converted into electricity by 15% efficient solar PV panels, each square meter covered by the panels would generate about 0.3 MWh of electrical energy per annum. For comparison, a 1000-MW(e) fossil fuel or nuclear power plant generates about 8.7×10^6 MWh(e) per annum. A solar power plant based on 15% efficient PV panels would need to cover about 30 km^2 to equal the capacity of one such conventional electrical power plant. This is a large area; however, the solar power plant need not be concentrated at a single site. In fact, it may be preferable to distribute the PV panels over a wide geographical area so that the power would be produced close to where it will be consumed.

There are other potential causes of environmental concern. For example, the production of solar cells requires the use of toxic substances, like silane, phosphene, arsene, etc. Furthermore, the major constituents of some solar cells can be hazardous in themselves, elements like cadmium, selenium, and tellurium. However, studies of these potential problems have lead to the conclusion that they will not be serious, provided that standard safe industrial procedures are followed (Goetz *et al.*, 1994; Patterson *et al.*, 1994).

6. APPLICATIONS OF PV POWER SUPPLIES

The current (1994) annual production of PV cells is such that if all the cells were gathered in some one place and illuminated by solar radiation having an intensity of 1 kW/m^2, their total power output would be about 40 MW. This amount of solar cells is referred to as 40 MW peak or 40 MW$_p$. By 1994, the world production had been at this level for about 15 years so that the accumulated production of PV cells was about 600 MW$_p$ in 1994. About 70% of these cells have been single-crystal or polycrystalline silicon cells; most of the balance have been thin-film amorphous silicon cells. These cells have been used in satellite power supplies, for remote applications, and in electronic products.

6.1 Satellite Power Supplies

Satellite PV power supplies have always utilized the highest-performance solar cells available because the penalty for failure of the power supply is unacceptably high; it is difficult or impossible to repair or replace cells in satellites. The cells have therefore always been single-crystal cells, overwhelmingly silicon, although gallium arsenide cells have begun to be used on some satellites, partly because GaAs cells are more radiation resistant than silicon and even more importantly because they have a higher conversion efficiency. The cost of solar cells is not as important in space applications as in most terrestrial applications because the cost of the power supply for the satellite is a small fraction of the combined cost of the satellite and of launching it. Space applications account for only a few hundred kilowatts of the annual solar-cell production.

6.2 Remote Power

When there is a need for up to a few kilowatts at a remote site in a developed country, it is often more economical to install a PV power supply than to extend the electric grid. Such sites include power for microwave repeater stations, for cathodic protection of oil and gas pipelines, for railroad crossing signals, etc.

In developing countries, where the electric power grid covers only small fractions of the country, PV power is the cost-effective way to provide a few kilowatts for villages far from the grid. The power can be used for educational television receivers, night lighting, refrigerators for medical supplies, etc. The cells used in such remote applications are based either on single-crystal or on polycrystalline silicon.

6.3 Other Applications

Most hand-held calculators use a small solar-cell power supply. These power supplies usually utilize amorphous silicon cells because they cost less and are more suited to indoor artificial lighting where the calculators are commonly used. Amorphous silicon cells have also been used in toys, clocks, battery trickle chargers, etc.

7. ECONOMICS OF PV POWER PLANTS FOR LARGE-SCALE ELECTRICITY GENERATION

The principal impediment to the utilization of PV cells to produce electricity on a large scale has been their cost, even though the cost has decreased dramatically over the past 40 years. In 1955, the price of single-crystal silicon PV cells was about \$2/cm^2 (in 1955 dollars) for 10% efficient cells. In 1994, the price of 12% efficient single-crystal cells has dropped to about \$0.04/cm^2 (in 1994 dollars).

Economic studies of large-scale PV power generation lead to the conclusion that the price must still drop by a factor of about 5, if electricity produced by PV cells is to be competitive with electricity produced by fossil-fuel and nuclear power plants (Wolf, 1981). Furthermore, these studies conclude that there is a minimum efficiency that these low-cost solar PV cells and panels made from them must exhibit if the PV plant is to produce electricity at a competitive price. This is because the solar power plant is not built from PV cells alone. The other components of the system can be classified into two groups:

1. components whose cost is proportional to the area occupied by the PV plant, which in turn is inversely proportional to the PV panel efficiency η (e.g., these components include the site and its preparation, panel support structures, wiring, etc.), and
2. components whose cost is proportional to the power generated but not to the area (e.g., dc-to-ac converters, etc.).

The cost C_T of a PV plant per unit power produced will therefore be given by

$$C_T = A(C_A + C_{PV}) + C_{PW},$$

where A is the area of the PV plant required to produce unit power when the incident solar power is I; C_A and C_{PV} are the costs per unit power produced per unit area of the area-proportional components and of the PV panels, respectively, and C_{PW} is the cost per unit power of the power-proportional components. The total power produced by the PV panels is given by

$$P_0 = A_0 \eta I,$$

where η is the panel efficiency. From the definition of A, $P_0 = 1$ unit when $A_0 = A$. Therefore, $A = 1/\eta I$ per unit power produced. Consequently,

$$C_T = C_{PV}/\eta I + C_A/\eta I + C_{PW}.$$

The intended price of unit power at the output of the plant sets the value of C_T. If PV panels are free, i.e., $C_{PV} = 0$, then for given values of C_A and C_{PW}, there exists a minimum efficiency η that the panels must have if the target value of C_T is to be achieved. Figure 19 shows the results of one such calculation for a solar power plant with specific (reasonable) costs for C_A and C_{PW}. The calculation includes a "planning target" for the price of electricity at the output busbar of the solar PV power plant; it was set at 6 cents (U.S.) per kWh. The figure shows that if the solar-energy conversion efficiency of the PV modules used in the plant is 10%, then the allowable module cost (C_{PV}) will be about \$10/m^2. Since 10% cells produce about 100 W/m^2 when illuminated by 1 kW/m^2 sunlight, this translates to \$0.10/peak watt. For 15% modules, the numbers are \$45/m^2 and \$0.30/peak watt; for 20% modules, \$75/m^2 and \$0.37/peak watt; and for 25% modules, \$120/m^2 and \$0.48/peak watt. PV modules with efficiencies of 15% have been fabricated from silicon cells. Individual tandem cells having efficiencies in excess of 30% and an area of 1 cm^2 have also been made. Thus, cells with the efficiencies required for large-scale power plants have been made but their price remains too high.

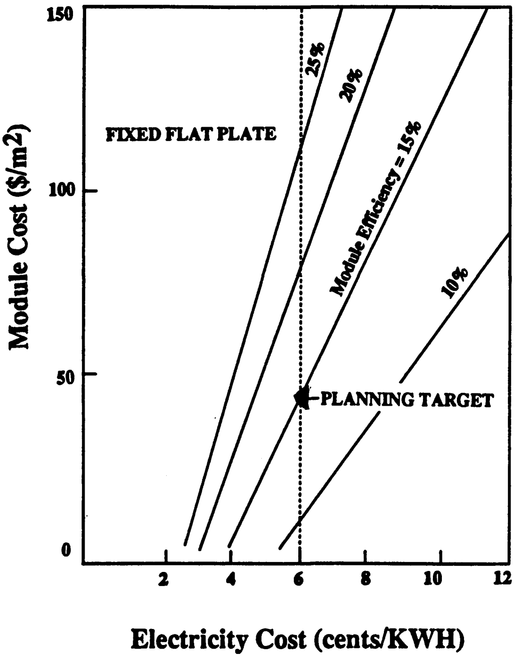

FIG. 19. Allowable cost/unit area of solar-cell panels vs busbar cost of electrical energy produced by a PV power plant.

Note also that if the "planning target" cost of electricity is increased to say 12 cents (U.S.)/kWh, then the allowable module cost for 10% efficient modules rises to about \$70/m^2 and \$0.70/peak watt. For 15% modules, the corresponding numbers are \$160/m^2 and \$1.10 peak watt. For orientation purposes, note that the cost of 10% efficient modules in 1994 is about \$4/peak watt so that reaching the allowable cost goals for 12 cents (U.S.)/kWh electricity is in sight, but the requirements for 6 cents (U.S.)/kWh module cost goals are formidable indeed. It can be expected that the price of PV modules will continue to decrease and that PV power plants will eventually become economically competitive.

8. SUMMARY AND CONCLUSIONS

1. Although the PV effect was first reported in 1839, its first limited applications occurred around 1910 and its first significant applications occurred only after 1953, which date marks the beginning of the modern PV age.
2. By 1955, theoretical analyses of the new PV cells led to the conclusion that solar-energy conversion efficiencies of about 26% were possible in single-junction PV cells. Multijunction PV cell systems can have efficiencies in excess of 30%, according to theory.
3. By 1955, single-crystal silicon cells with efficiencies of about 10% were being fabricated. By 1985, the gap between the 1955 theory and experimentally observed Si efficiencies was closed. The same was true

for cells made from single crystal of GaAs. Multicell efficiencies in excess of 30% have also been reported by the mid-1980s.

4. Thin-film cells have attracted attention since the earliest times because of the anticipated lower cost of such cells. By 1994, stable polycrystalline thin-film cells having efficiencies up to 16.4% were reported.
5. Amorphous silicon-hydride PV cells were invented in 1974 and continue to attract attention because of ease of fabrication and the promise of low cost. Their highest stabilized efficiencies in 1994 were around 10%.
6. Applications of PV cells include satellite power supplies, remote power, hand-held calculators, etc. The technology for large-scale generation of electricity from PV power plants exists, but its application awaits reduction in cost of high-performance PV cells.

GLOSSARY

Air-Mass Zero (AM0) Solar Spectrum: The solar spectrum just outside the Earth's atmosphere, which has an intensity of about 1.4 kW/m^2 and a specific, never changing spectral composition.

Air-Mass One Solar Spectrum: The solar spectrum at the surface of the Earth when the sun is at the zenith. This is necessarily an idealized spectrum since water vapor and other impurities in the atmosphere can change the intensity and spectral content of the sunlight. The AM1 intensity is about 1.0 kW/m^2, and the spectral composition is a best estimate of what it would be if water vapor and other impurity content were negligible. It can be approximated by the spectrum of a 5800-K blackbody.

***p-n* Junction:** An interface between an n-type and a p-type region in a semiconductor structure. It is a homojunction if the n and p sides are differently doped regions of the same semiconductor. It is a heterojunction if the n-type semiconductor is a different material than the p-type semiconductor.

Peak Watt Definition: An amount of solar cells that deliver 1 watt of electrical power to a matched load when the cells are illuminated by sunlight having a specified intensity and spectrum at normal incidence. The appropriate intensity for terrestrial solar cells is 1 kW/m^2; for space cells, 1.4 kW/m^2. Thus, one peak watt of 10% efficient solar cells would have an area of 100 cm^2, whereas one peak watt of 20% efficient solar cells would have an area of 50 cm^2.

Photovoltaically Active Volume (PVAV): That part of a PV cell in which absorption of light leads to the generation of excess carriers that contribute to the PV effect. Wide-bandgap windows are not part of the PVAV.

PV Solar-Energy Efficiency: The ratio of electrical power output to the solar power input. If comparisons of efficiency are to be made, the efficiency must be measured under comparable conditions. These conditions have been standardized so that the spectral composition and intensity of the solar radiation is fixed (usually AM1.5 for terrestrial PV cells) and the temperature of the cell is kept at some specified value (usually room temperature). The power delivered to a load whose resistance is equal to the resistive part of the impedance of the PV cell is measured. This is the maximum power that can be transferred to a load and the efficiency is called the maximum efficiency $\eta_{\max}$.

Semiconductor: A material whose electrical conductivity is higher than that of an insulator but lower than that of a conductor. The room-temperature conductivities of semiconductors lie between about 100 S/cm and 10^{-5} S/cm. The conductivity of semiconductors increases with temperature, whereas the conductivity of metals decreases with increasing temperature.

Works Cited

Becquerel, E. (1839), *Comptes Rendues* **9**, 561–565.

Benett, A., Olsen, C. (1978), *Conf. Rec. 13th IEEE PVSC.*, New York; IEEE Press, 968–975.

Brattain, W., Bardeen, J. (1948), *Phys. Rev.* **74**, 230–233.

Carlson, D. E., Wronski, C. R. (1976), *Appl. Phys. Lett.* **28**, 671–673.

Chapin, D. M., Fuller, C. S., Pearson, G. L. (1954), *Phys. Rev.* **96**, 533.

Chu, T. L., Chu, S. S. (1993), *Prog. Photovoltaics*, **1**, 31–42.

Cusano, D. A. (1963), *Solid State Electron.* **6**, 217–232.

Gabor, A., Tuttle, J., Contreras, M., Alban, D., Tennant, A., Qu, Y., Du, D., Carapella, J., Noufi, R. (1994), *Proc. 12th European PV Solar Energy Con-*

ference, Amsterdam, 1994, Dordrecht, The Netherlands: Kluwer Science Publishers, pp. 939–943.

Gale, R. F., Fan, J. C. C., Turner, G. W., Chapman R. L. (1984), *Conf. Rec 17th IEEE PV Spec. Conf.*, New York: IEEE Press.

Gocsen, N., Loferski, J. J. (1979), *Solar Energy Mater.* **1**, 271–286.

Goetz, M., Vaucher, S., Tscharner, R., Shah, A. (1994), *Proc. 12th European PV Solar Energy Conference, Amsterdam, 1994,* Dordrecht, The Netherlands: Kluwer Science Publishers, pp. 1229–1231.

Green, M. A., Blakers, A. W., Wenham, S. R., Jianhua, Z., Chong, C. M., Taouk, M., Narayanan, S., Willison, M. R. (1987), *Conf. Rec. 19th IEEE PVSC, May 1987, New Orleans, LA,* New York: IEEE Press, pp. 49–52.

Green, M. A., Emery, K. (1993), *Prog. Photovoltaics* **1**, 225–227.

Grondahl, L. O. (1933), *Rev. Mod. Phys.* **5**, 141–160.

Hedstrom, J., Ohlsen, H., Bodegard, M., Kylner, A., Stolt, L., Hariskos, D., Ruckh, M. Schock, H.-W. (1993), *Conf. Rec. 23rd IEEE PVSC May 1993, Louisville, KY,* New York: IEEE Press, pp. 364–371.

Hedstrom, J., *et al.* (1994), *Proc. 12th European PV Solar Energy Conference, Amsterdam, 1994,* Dordrecht, The Netherlands: Kluwer Science Publishers, pp. 1743–1746.

Jenny, D. A., Loferski, J. J., Rappaport, P. (1956), *Phys. Rev.* **101**, 1208–1209.

Kazmerski, L. L., White, F. R., Morgan, G. K. (1976), *Appl. Phys. Lett.* **29**, 268–270.

Landsberg, P. T., DeVos, A., Baruch, P. (1991), *J. Condens. Matter* (U.K.) **3**, 6415–6421.

Lange, B. (1930), *Z. Phys.* **31**, 139–145.

Loferski, J. J. (1956), *J. Appl. Phys.* **27**, 777–784.

Loferski, J. J., Rappaport, P. (1958), *RCA Rev.* **19**, 536–554.

Loferski, J. J., Wysocki, J. J. (1961), *RCA Rev.* **22**, 38–68.

Loferski, J. J. (1979), *Materials Science in Energy Technology,* New York: Academic Press, 219–253.

Loferski, J. J. (1988), *Proc. of Int. Elec. Dev. and Mat. Symp., Taiwan, ROC,* pp. 221–226.

Mickelsen, R. A., Chen, W. S. (1980), *Appl. Phys. Lett.* **36**, 371–373.

Nadjakov, G., Andreitchine, R., Borissov, M. (1954), *Izv. Bulg. Aka. Nauk.* **4**, 10–15.

Patterson, M. H., Turner, A. K., Sadeghi, M., Marshall, R. I., (1994), *Proc. 12th European PV Solar Energy Conference, Amsterdam, 1994,* Dordrecht, The Netherlands: Kluwer Science Publishers, pp. 951–953.

Rappaport, P. (1954), *Phys. Rev.* **93**, 246–247.

Reynolds, D. C., Leis, G., Antes, L. L., Marburger, R. E. (1954), *Phys. Rev.* **96**, 533–534.

Schottky, W. (1930), *Z. Phys.* **31**, 913–920.

Shay, J. L., Wagner, S., Kasper, M. (1975), *Appl. Phys. Lett.* **27**, 89–90.

Shockley, W. (1949), *Bell Syst. Tech. J.* **28**, 435–450.

Shockley, W. (1951), *Electrons and Holes in Semiconductors,* New York: Van Nostrand.

Trivich D., Flinn, P. A. (1955), in F. Daniels, J. Duffie (Eds.), *Solar Energy Research,* Madison, WI: Univ. of Wisconsin Press, pp. 143–145.

Verlinden, P. J., Swanson, R. M., Sinton, R. A., Crane, R. A., Tilford, C., Perkins, J., Garrison, K. (1993), *Conf. Rec. 23rd IEEE PVSC, 1993, Louisville, KY,* New York: IEEE Press, pp. 58–64.

Wolf, M. (1981), *Solar Cells* **3**, 327–336.

Woodall, J. M., Hovel, J. J. (1972), *Appl. Phys. Lett.* **21**, 379–381.

Wysocki, J. J., Rappaport, P. (1960), *J. Appl. Phys.* **32**, 571–578.

Further Reading

A detailed history and current status of PV science and technology are to be found in the conference records and proceedings of the three sets of international photovoltaic conferences. The oldest continuous set is the Institute of Electrical and Electronic Engineers Photovoltaic Specialists Conferences, the first of which was held in 1956 and the twenty-third in 1993. The conference proceedings are published by the IEEE Press, New York. The second set is the European Community Photovoltaic Solar Energy Conferences, the first of which was held in 1977 and the 12th in 1994. Their proceedings are published by Kluwer Academic Publishers, Dordrecht, Germany. The third set of conferences is held in the countries of the Eastern Pacific and is entitled the Photovoltaic Science and Engineering Conferences, the first of which was held in 1980 and the seventh in 1993. Their proceedings are published as separate books. These three sets of conferences are so synchronized that one of them takes place every 6 months.

There are two archival journals devoted to PV, namely *Solar Cells and Solar Energy Materials,* Elsevier, which appears monthly, and *Progress in Photovoltaics,* John Wiley & Sons, which appears quarterly.

Fahrenbruch, A., Bube, R. H. (1983), *Fundamentals of Solar Cells,* New York: Academic.

Green, Martin A. (1992), *Solar Cells,* Englewood Cliffs, NJ: Prentice-Hall.

PHYSICAL GEOLOGY

See STRUCTURAL AND PHYSICAL GEOLOGY

PHYSICS AND ENGINEERING ORGANIZATIONS

H. William Koch, *Department of Physics, University of Denver, Denver, Colorado, U.S.A.*

F. Karl Willenbrock, *Washington, DC, U.S.A.*

Introduction 560
1. **Physics Organizations** 560
1.1 Science Disciplines in Natural and Engineering Sciences 560
1.2 Chronology and Operational Tenets of Physics Organizations 561
2. **Precursers and Physics Societies** 562
2.1 National Academies of Science 563
2.1.1 ICSU and IUPAP 563
2.1.2 National Academy of Sciences, Washington 563
2.1.3 The Royal Society, London 564
2.2 Associations for the Advancement of Science 564
2.2.1 American Association for the Advancement of Science 564
2.2.2 British Association for the Advancement of Science 564
2.3 Physics Societies 565
2.3.1 Overview of Physics Societies in Three Regions of the World 565
2.3.2 Details of Selected Societies 568
2.3.2.1 The American Physical Society 568
2.3.2.2 American Institute of Physics 569
2.3.2.3 Other U.S. Physics and Related-Science Societies 571
2.3.2.4 American Center for Physics 572
2.3.2.5 The Institute of Physics, London 572
2.3.2.6 German Physical Society 572
2.3.3 Physics Societies in the Rest of the World 573
3. **Engineering Organizations** 573
3.1 General 573
3.1.1 Organization of This Section 574
3.1.2 International Counterparts 574
3.2 Inherent Characteristics of Engineering 575
3.2.1 Integration of Knowledge 575
3.2.2 Relation to Mathematics and Science 575
3.2.3 Societal Impact 576
3.3 Technically Oriented Organizations 576
3.3.1 Services of Professional Societies 576
3.3.2 Founder Societies 577
3.3.3 Organizational Characteristics 577
3.3.4 Unique Character of the IEEE 578
3.3.5 Nontechnical Activities 578
3.4 Educationally Oriented Organizations 578
3.4.1 Studies of Engineering Education 578
3.4.2 Accreditation of Academic Programs 579
3.5 Industrially Oriented Organizations 580
3.5.1 Standards-Writing Organizations 580
3.5.1.1 International Standardization 581
3.5.2 Trade Associations 581
3.6 Public Policy Oriented Organizations 582
3.6.1 National Academy of Engineering 582
3.6.2 National Society of Professional Engineers 582
3.6.3 Umbrella Organizations 582
3.7 Challenges for Engineering Organizations 583
Appendix A. Engineering's "Founder Societies" 584
Appendix B. Comparison of Two Leading U.S. Computer Societies 586
Works Cited 587
Further Reading 587

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

The foundation of physical sciences is physics, which emphasizes fundamental understanding of physical processes in nature. Physics is based on theoretical concepts supported by experimental data. In contrast to the physical sciences, engineering is concerned with the man-made world. It applies the understanding gained from all the sciences and also generates new knowledge in support of the creation of products and services useful to mankind. The principal engineering disciplines have been civil, mechanical, electrical, and chemical engineering. These names alone suggest that the central sciences applied by engineering are the physical sciences.

The professional institutions that serve physicists and engineers are physics and engineering organizations, respectively. This article summarizes the national and international histories and roles of these organizations, primarily from a U.S. perspective.

1. PHYSICS ORGANIZATIONS

Physics organizations provide the institutional framework that enables physicists and those in related fields to develop and distribute their knowledge; thus, they advance science and technology to meet the needs of government, of industry, and of society at large. They do so by arranging scientific meetings, publishing peer-reviewed research and education journals, and doing other profession-related services.

From the earliest days of the study of physics, the body of fundamental knowledge grew slowly, nourished by physicists working almost exclusively in academia. That slow growth changed dramatically in the twentieth century. With the exigencies of wars, scientific advances brought the world into the nuclear and the space ages.

Critical contributions by physicists in meeting national needs led to recognition of physics worldwide as a science to be encouraged by governments to obtain technological and economic advancements. The result was a 25-year period beginning in 1946 when physics and physicists prospered and basic physics received an increasing share of national funds in many industrialized countries. Then in the 1970s, with worsening economic conditions, due in part to the world oil shortage and increasing national debts, abundant financial support for basic physics research slackened. The emphasis for the next two decades or so was increasingly on applications of physics.

As physics expanded in the twentieth century, so, too, was there a need for expansion within existing physics organizations and for different types of organizations. What are those physics organizations? How do they serve the science of physics and related disciplines, the nation, and physicists themselves? This section and Sec. 2 aim to answer those questions.

1.1 Science Disciplines in Natural and Engineering Sciences

The characteristics of physics organizations are determined by the science that requires them. An understanding of that science can be obtained from the diagram in Fig. 1 that schematically relates the "things" in our natural universe and the science disciplines in the natural and engineering sciences that study them.

The shaded, outer ring of Fig. 1 identifies the four categories of things in our natural universe. They are nonliving materials, living materials, man-made systems, and natural systems.

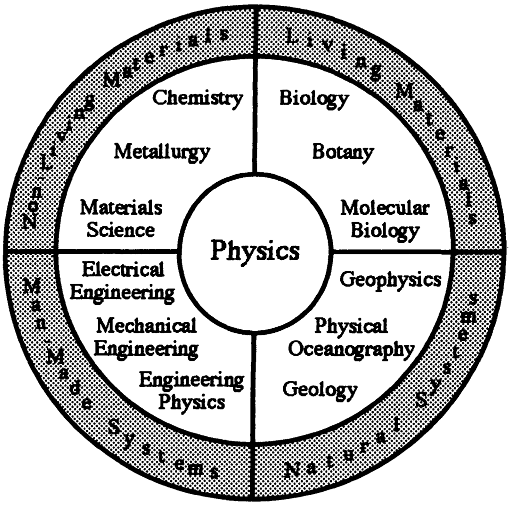

FIG. 1. "Things" in our natural world (shaded areas) and their relation to science disciplines in natural and engineering sciences.

Within the ring are four quadrants containing examples of the names of four groups of science disciplines. The sciences listed in the quadrant study "things" named in the part of the outer ring closest to the quadrant. For example, chemistry, metallurgy, and materials science study nonliving materials, while biology, botany, and molecular biology study living materials. Similarly for the parts of the ring and quadrants that deal with systems. Engineering physics and other sciences fundamental to electrical, mechanical, civil, or chemical engineering, for example, are usually involved with man-made systems, while geophysics, physical oceanography, and geology, for example, study natural systems.

The sciences in the four groupings are largely observational in character. Their scientists observe materials and systems as found in our natural world.

Physics is a science that is in its own grouping in the central small circle of Fig. 1. It is separate from, but coupled to, the other four. Partly observational in character, physics usually confines its observations to basic materials and systems that are simple and sometimes idealized. Physicists also use observational data obtained by others in other science disciplines, and provide the analysis and interpretation that clarify the underlying simplifications. Thus, physics is primarily regarded as a conceptual science because its involvement is principally with concepts of matter, motion, and energy that are fundamental to all of the natural and engineering sciences. Theoretical concepts form a significant part of all work done in physics. Consider, for example, the roughly 30% of all U.S. doctoral theses in physics that are devoted to theoretical topics.

Because of its coupling to all the other sciences, physics is very interdisciplinary. Evidence of this characteristic can be found in the interdisciplinary nature of subdisciplines, organizations, conferences, and journals in which physicists routinely participate. Examples of such subdisciplines are chemical physics, biophysics, geophysics, and physical oceanography. Other evidence to be discussed later is the unusual interdisciplinary mix of primary science disciplines represented in the members of physics membership societies in the U.S.

The diagram of Fig. 1 is not to be taken too literally, even though it does give some useful understanding of the way sciences relate to one another. As an example of such understanding, the relationship among the sciences is perhaps confirmed from the number of graduate doctorate degrees awarded each year by the physics graduate schools in the U.S.A. Data on these degrees have been collected by the U.S. National Academy of Sciences (NAS) since 1920. For each year since that time, the ratio of the number of degrees awarded in physics to the total in natural and engineering sciences has been remarkably constant at 11% to within one percentage point. "Rough statistics from Canada and the United Kingdom, with rather different educational systems, show this same fraction and constancy" (Physics Survey Committee, 1972). Thus, the science enterprises in the U.S., Canada, and the United Kingdom (at least) invest a fixed fraction (11%) of their respective resources each year in the development of concepts in physics to maintain the "health" of their natural science communities.

1.2 Chronology and Operational Tenets of Physics Organizations

By definition, the science of physics has benefited from physics organizations, ever since physics was established as an important branch of science in the mid 1800s. The precursors of physics organizations were science organizations, in which a physics component developed.

Thus, the history of physics organizations really began with the formation of an all-science, honorific society in 1660, the Royal Society of London. The Berlin Academy of Science followed in about 1720 (Evan, 1981). Membership in these organizations and counterparts elsewhere was limited because of the selection criteria, and they were followed by all-science, open-membership societies called associations for the advancement of the sciences. The British Association for the Advancement of Science (BAAS) was founded in 1831, while the American Association (AAAS) followed in 1848. These organizations satisfied the need for meetings and journals of scientific record, even in specialty sciences such as physics, until the German Physical Society (GPS) was founded in 1845, the Physical Society in London (PS) in 1873, and The American Physical Society (APS) in

1899, to cite only a few examples. The chronological development of physics organizations in Europe and the U.S. was consistent with the increases in the numbers of trained physicists.

As the numbers increased, so did the understanding of the basic tenets under which most physics organizations believe they must operate. The tenets are based on the demonstration that physics, with its heavy emphasis on pure physics and original ideas, requires for its healthy development strong interdisciplinary coupling plus free, open, voluntary interchanges of scientific information among all participating scientists. Thus, the important operational tenets are these:

1. All physics publishers should subject proposed contributions for archival journals to peer review.
2. A single language—English—should be accepted internationally for communicating physics information.
3. National policies should allow free, open, voluntary interchanges of scientific information among scientists.
4. National policies should honor intellectual property rights by use of patents and copyrights.

Examples of the application of these tenets during the last 50 years are the following:

1. "Rochester" conferences, named after the first high-energy-physics conference at Rochester University, Rochester, New York, have been held annually since 1952 in different parts of the world (including Moscow). Soviet scientists were able to convince their authorities that open communication with physicists in the U.S. and Europe was essential for their participation in high-energy research. Permission by the communist regimes in the USSR (before its dissolution) was given to Soviet high-energy physicists to participate openly in international conferences.
2. An important requirement for an international conference to be endorsed by the International Union of Pure and Applied Physics (IUPAP) is that participation in the conference must be open to physicists independent of their home country's politics.
3. English-language journals of American physics societies were sent most articles on pure physics of high quality by Japanese, Chinese, and Russian physicists during the last two decades. The selection of American journals is motivated by the desire to reach as large an audience of readers doing similar research as possible. Thereby, these foreign physicists attempt to establish their scientific credentials in the western world and to communicate more effectively with their American peers.
4. In about 1976, a decision was made by the German Physical Society (GPS) and the German government to encourage journal articles and abstracting services published by GPS to be published in the English language, because German physicists were reluctant to publish in journals read by limited audiences. One indirect result of that decision led to the creation of an English-language abstracting service called PHYSICS BRIEFS that replaced a service in the German language called PHYSIKALISCHE BERICHTE, which had been started in 1885.

2. PRECURSORS AND PHYSICS SOCIETIES

The physics organizations and their precursors to be treated in this article are those that now serve effectively the international community of physicists. They are

1. national academies of sciences,
2. the International Union of Pure and Applied Physics (IUPAP) of the International Council of Scientific Unions (ICSU),
3. associations for the advancement of science, and
4. physics societies.

A general discussion of each type in the order listed follows, with details for several outstanding examples. Many examples are from the United States of America (U.S.) and England, because here the information was readily available and because societies in the U.S. and England are large with interesting evolutionary characteristics. The information for both the U.S. and England will show the interrelationships of the four types of organizations, including their roles in the development of strong national physics programs.

2.1 National Academies of Sciences

National academies of sciences are defined as honorary organizations that afford membership and preferential treatment to those who have made outstanding contributions to the advancement of science. An example is the Japan Academy, which was established in 1877 and which now has two sections. Section 1 in the Humanities and Social Sciences has 59 members, none of whom is a physicist; Section 2 in the Pure Sciences and their Applications has 68 members, of whom 12 are physicists (Anon., 1994). Similarly, the Royal Societies in the separate territories within Australia and New Zealand are also honorary only.

Some national academies not only have an honorific role but also officially provide scientific advice to their federal governments. The U.S. National Academy of Sciences (NAS) is an example.

In addition to providing honor and advice, some academies are also the chief coordinating bodies of scientific research in their countries. Two important examples are in the Soviet Union (before its breakup) and in the People's Republic of China.

2.1.1 ICSU and IUPAP An important international role of many national academies of sciences is their adherence to the International Council of Scientific Unions (ICSU). ICSU was founded in 1919 "largely through the representatives of the National Academy of Sciences, Washington, and of the Royal Society, London, to coordinate international efforts in the different branches of sciences, under whose aegis international associations or unions in different branches of science could be formed" (International Union of Pure and Applied Physics Secretariat, 1992). Ninety-five countries are now members of ICSU, headquartered in Paris.

According to ICSU principles, physicists at the General Assembly in 1922 decided that an International Union of Pure and Applied Physics (IUPAP) should be formed. National physics committees or groups of physicists from 43 countries adhered to IUPAP in 1990. IUPAP "promotes international cooperation in physics and the use of international symbols, units, nomenclature, and standards; provides assistance in organizing committees and meetings; encourages research and publication of papers and tables; and fosters free circulation of scientists" (Irvin, 1993). General assemblies are held triennially.

The list of 18 IUPAP scientific and technical commissions in Table 1 (International Union of Pure and Applied Physics Secretariat, 1992) shows the major subdisciplines regarded by IUPAP to be active in the field of physics.

2.1.2 National Academy of Sciences, Washington A Congressional charter approved by President Abraham Lincoln created the National Academy of Sciences (NAS) in 1863, not only to obtain scientific advice regarding the execution of the Civil War, but also as an honorific, private, nonprofit, self-governing membership body responsible for advising the federal government, upon request and without fee, on questions of science and technology. Under this charter, the National Academy of Engineering (NAE) was established in 1964 and the Institute of Medicine (IOM) in 1970.

Specific examples of three of the NRC committees are the Space Science Board that advises on the activities of the National Aeronautics and Space Administration (NASA), the Defense Science Board that advises the Department of Defense (DOD), and the Visiting Committees to the National Institute of Standards and Technology (NIST) in the U.S. Department of Commerce.

Surveys of Physics in 1964 by a committee chaired by George Pake, in 1972 chaired by D. Allan Bromley (Physics Survey Committee, 1972), and in 1984 chaired by William Brinkman are examples of *ad hoc* assignments to the NRC.

Table 1. IUPAP scientific and technical commissions.

1. (Administrative)
2. Symbols, Units, and Nomenclature
3. Thermodynamics and Statistical Mechanics
4. Cosmic Rays
5. Very Low Temperature
6. Biological Physics
7. Acoustics
8. Semiconductors
9. Magnetism
10. Condensed Matter
11. Particles and Fields
12. Nuclear Physics
13. Physics for Development
14. Physics Education
15. Atomic and Molecular Physics and Spectroscopy
16. Plasma Physics
17. Quantum Electronics
18. Mathematical Physics
19. Astrophysics

2.1.3 The Royal Society, London The Royal Society of London (RS) for Improving Natural Knowledge was founded as The Royal Society in 1660. RS was granted a Royal Charter under the full title in 1663 (Anon., 1976).

Most academies provide honor and advice and/or coordinate scientific research largely for their own nations, and they tend to pursue their scholarly activities in like manner. The single exception is the Royal Society, London. The Russian physicist, Peter Kapitza, has referred to "Newton's active role in the affairs of the Royal Society, observing that the Royal Society 'in contrast to all academies, has still the character of a non-state society. . . . Since Newton's time, the Royal Society has maintained a lively contact with scientific societies the world over'" (Evan, 1981).

The Royal Society has 763 Fellows and 74 Foreign members. Fellows are elected, on the basis of distinction in science, at each annual meeting of the Society. The Society has a long tradition in publishing with *Philosophical Transactions* (Series A and B) and *Proceedings* (Series A and B).

2.2. Associations for the Advancement of Science

Because academies are highly selective in choosing members, scientists in most countries require organizations that reach out also to the rest of the scientific community. Associations for the advancement of science are one such organizational type.

These associations attempt to be broadly inclusive of all sciences, interested individuals, other societies, companies, and audiences. Examples are the Australian and New Zealand Association for the Advancement of Science (founded in 1886 and now with 1000 members), the China Association for Science and Technology (founded in 1958), the British Association for the Advancement of Science (founded in 1831 and now with 2000 members), the French-Canadian Association for the Advancement of Science (founded in 1923 and now with 23 societies as members), and the American Association for the Advancement of Science (AAAS). The founding, history, and current status are treated next.

2.2.1 American Association for the Advancement of Science The American Association for the Advancement of Science (AAAS) was founded in 1848 with about 400 members. Then, physics was still not fully established as a separate discipline, but as part of natural philosophy or science. It was not until 1876 that the first Professor of physics was appointed at a U.S. university. Henry A. Rowland of the Johns Hopkins University had that distinction (Phillips, 1990). Physics received further recognition in the AAAS Constitution of 1874 that established physics in the name of a new Section: Section A—Mathematics, Physics, and Chemistry; Section B was Natural History. Nine Sections date from 1882, and they include a Section B, Physics. By 1895, there were 12 Sections, and the number has now expanded to 21. Section B in 1899 decided that their growing numbers required them to have a separate society. Thus was formed The American Physical Society, with Professor Rowland as the first president. Currently, the AAAS has 140 000 individual members and 300 affiliated societies and regional and state academies.

The AAAS holds large national meetings, publishes a weekly news magazine—*Science*—, publishes books, and manages many advisory committees. Some of the committees serve important sociopolitical roles. They address matters that influence science and technology generally, such as secrecy in government, civil rights of scientists, and scientific information exchange.

As the membership of AAAS grew, particularly after 1950, management by a very large governing council became unwieldy. At that time, AAAS was an association in which representatives of member societies on a large Council (then 500 members) had direct involvement in governance and which was advised by a Board of Directors (then 13 members). In 1970, it changed to an association in which the council was the group advising the Board (Wolfle, 1989).

2.2.2 British Association for the Advancement of Science The British Association (BAAS) was established at York in 1831 by Royal Charter. Membership was and still is open to all interested in science without proposal or election. 2000 members are active.

The annual meeting, held by invitation in a different city each year, is the main activity to achieve the ends of advancing science. It is generally held in provincial cities of the United Kingdom and the Republic of Ireland and in the past has been held in certain cities of the Commonwealth. It lasts for a week, and the program of the meeting is arranged by scientists from the major disciplines of modern science (Anon., 1976).

2.3. Physics Societies

Science societies were founded by small groups of scientists who wanted to share their ideas and scientific discoveries. They originated "in Renaissance Italy in the sixteenth century, spreading across Europe in the seventeenth, and over the American continents in the eighteenth and nineteenth century, and now in the twentieth century, rapidly invading the cities of Asia . . ." (Bates, 1965).

Thus, the concept of a science society was well established before the concept of a science of physics was established in the mid nineteenth century. Once the German physicists founded a German Physical Society in 1845, others followed in a relatively short time in France in 1873, England in 1873, Switzerland in 1887, Hungary in 1891, Netherlands in 1921, and Poland in 1921. The spread to the Pacific Rim began first with astronomical societies in China and Japan, both in 1908, and then later with physical societies in China in 1932 and Japan in 1946. In the case of the Americas, the Canadians and the Mexicans founded their astronomical societies in 1890 and 1878, respectively, followed in the mid 1900s by physical societies. In the U.S., physicists and astronomers founded both of their independent societies, The American Physical Society and the American Astronomical Society, respectively, in 1899 (Anon., 1994). Incidentally, in this article, physical and astronomical societies are both regarded as physics societies, because of the close relationship the subjects of physics and astronomy have in academic departments and because the American Astronomical Society (AAS) elected to be affiliated with other U.S. physics societies in 1966, as is discussed later.

Physics societies are now organized in many industrialized nations of the world. They currently establish the standards of their profession. They do so by arranging open scientific meetings, specialized review conferences, and peer-reviewed scientific journals. These societies usually are open to anyone interested in joining who meets certain minimum qualifications. Other societies can sometimes join if permitted by the bylaws of the parent societies. Physics societies are outstanding examples of organizations that accomplish their purposes of advancing and distributing knowledge in physics.

They may elect to focus only on science-related matters, profession-related matters (such as employment and career information), or education, or any combination of these. They may function in a national sphere only, or they can be international as well. They may also be business or commercially oriented. If they limit themselves to cultural matters, such as education, science, literature, philosophy, and charity, the societies have little in the way of resources with which to operate. To help such societies, their federal governments often provide direct or indirect financial subsidies. In some countries, the subsidies consist of free housing; in the U.S., the subsidies consist usually of freedom from the payment of certain federal and state taxes. Physics societies vary enormously in size. They range from the 20 or so physicists in a school's physics club to the 40 702 members of the world's largest physics society, The American Physical Society (APS).

For the purposes of this article, the highlights of physics societies will be covered by

1. an overview of the principal physics societies in three key regions of the world for which information can be found in the published literature;
2. a description of The American Physical Society headquartered near Washington, DC, in College Park, Maryland;
3. a description of a membership corporation of U.S. physics societies that includes The American Physical Society, called the American Institute of Physics;
4. comments on other U.S. physics and related-science societies;
5. a report on a new American Center for Physics at the APS address; and
6. comments on the Institute of Physics, London, on the German Physical Society (GPS), and on a few other societies in other regions of the world.

2.3.1 Overview of Physics Societies in Three Regions of the World The data in Tables 2 through 4 give an overview of some

Table 2. Principal physics societies in Europe (including Nation Group, Society, Year Founded, and Total Current Members). The 29 societies listed are incorporated as members in a membership society, the European Physical Society. EPS, founded in 1968, currently has 3700 individual members.

Nation group	Society	Year founded[a]	Members
Austria	Austrian Physical Society	1950	779
Belgium	Belgian Physical Society	NA	496
Bulgaria	Bulgarian Academy of Science	1971	105
Czechoslovakia	Physics Section, Mathematics/Physics Union	NA	1 000
Denmark	Danish Physical Society	NA	274
England	The Institute of Physics	1873/1970	14 500
Finland	Finnish Physical Society	NA	915
France	French Physical Society	1873	2 405
Germany	German Physical Society	1845	18 400
Greece	Hellenic Physical Society	NA	55
Hungary	Eotvos Lorand Physical Society	1891	628
Iceland	Icelandic Physical Society	NA	45
Ireland	Royal Irish Academy	NA	60
Israel	Israeli Physical Society	1954	250
Italy	Italian Physical Society	1897	1 100
Netherlands	The Netherlands Physical Society	1921	2 653
Norway	Norwegian Physical Society	NA	709
Poland	Polish Physical Society	1921	1 968
Portugal	Portuguese Physical Society	NA	200
Romania	Romanian Physical Society	NA	235
Spain	Spanish Royal Society of Physics	1903	300
Sweden	Swedish Physical Society	1920	1 043
Switzerland	Swiss Physical Society	1887	1 370
Turkey	Turkish Physical Society	NA	302
USSR	Department of General Physics & Astronomy	NA	2 000
Yugoslavia	Assoc. of Mathematics, Physics, & Astronomy	1950	200
Zagreb	Ruder Boskovic Institute	NA	95

[a] NA = Not available.

Table 3. Principal physics societies in Pacific Rim (including Nation, Society, Year Founded, and Total Current Members).

Nation	Society	Year	Members
Japan	Physical Society of Japan	1946	17 022
	Japan Society of Applied Physics	1932	9 000
	Crystallography Society of Japan	1950	10 007
	Acoustical Society of Japan	1936	3 600
	Astronomical Society of Japan	1908	2 640
China	Chinese Physical Society with 9 subsocieties in: 1. High Energy Physics 2. Nuclear Physics 3. Liquid Crystals 4. Electron Microscopy 5. Mass Spectroscopy 6. Luminescence 7. Particle Accelerators 8. Gravitational/Relativistic Astrophysics 9. Phase Diagrams	1932	27 095
	Chinese Optical Society	1979	9 018
	Acoustical Society of China	1988	2 879
	Chinese Astronomical Society	1908	2 400
Australia	Australian Institute of Physics	1963	2 400
	Australian Acoustical Society	1971	390
	Astronomical Society of Australia	1966	250

Table 4. Principal physics societies in North America (including Nation, Society, Year Founded, and Total Current Members). The 10 U.S.A. societies listed are incorporated as members in a membership society, the American Institute of Physics. The dates of affiliation with AIP are given in parentheses.

Nation	Society	Year[a]	Members
U.S.A.	The American Physical Society (1931)	1899	40 702
	Optical Society of America (1931)	1916	11 634
	Acoustical Society of America (1931)	1929	6 900
	The Society of Rheology (1931)	1929	1 269
	American Assoc. of Physics Teachers (1932)	1930	11 178
	American Crystallographic Association (1966)	1949	2 100
	American Astronomical Society (1966)	1899	6 119
	Amer. Assoc. of Physicists in Medicine (1973)	1958	3 678
	American Vacuum Society (1976)	1953	5 455
	American Geophysical Union (1986)	1919	29 766
Canada	Canadian Association of Physicists	1945	2 000
	Royal Astronomical Society of Canada	1890	3 500
	Spectroscopy Society of Canada	1967	500
Mexico	Mexican Physical Society	NA	NA
	Mexican Astronomical Society	1902	500

---

[a] NA = Not available.

of the principal physics societies in the world today (Anon., 1994; Burek, 1993; Irvin, 1993).

In Europe in 1968, 29 nation groups organized their physics societies (but not including their astronomical societies) in the European Physical Society (EPS) (see Table 2). These societies have grown in the 1990s to a total of more than 50 000 members (European Physical Society, 1992). Although EPS is a membership corporation of societies similar to the American Institute of Physics (AIP), the EPS selected at its formation a name that was analogous to The American Physical Society (rather than to the AIP). The reason is presumed to be that APS is the U.S. physics organization that European physicists wanted the EPS most to emulate.

Before the founding of the EPS, a major problem for European physicists had been the many individual societies in Europe. These societies tended to arrange local or regional meetings and local or regional journals. Very few Pan-European physics meetings had been available. The result was that European physicists had to come to meetings in the U.S. arranged by the APS, and others, to participate in free, large, and multidisciplinary exchanges of useful scientific information. Besides the meaningful choice of name, the EPS has identified appropriate operations for itself such as *Europhysics Journals*, and for individual nation groups who are qualified to contract to do publishing on behalf of the EPS community. Also, the EPS has begun holding triennial congresses, the most recent of which was held in 1993. It is evident that the European Physical Society represents a big step forward in meeting the organizational needs of European physicists.

In the Pacific Rim, three major sets of physics societies are to be found in the nations of Japan, Australia, and China (see Table 3).

Of these three, the nation that has had considerable investments in technology, but small investments before 1980 in basic physics, is Japan. The result has been that Japan has not generally developed a strong tradition for publishing journal articles in pure physics, although their journal *Progress in Theoretical Physics* has been an important exception since the 1950s. They have emphasized applications of physics. It is to be expected that the investments in the 1990s will be more directed to pure, long-range research in physics, in order to improve Japan's scientific and technological competitiveness. The general physics society is the Physical Society of Japan (PSJ), founded in Tokyo in 1946. PSJ has 15 859 members and publishes *Butsuri* (monthly, in Japanese), *Journal* (monthly), *Progress in Theoretical Physics* (monthly), and, jointly with the Japan Society of Applied Physics (JSAP), *Japanese Journal of Applied Physics* (monthly). JSAP was founded in 1932 and presently has 9000 members. The Crystallographic Society of Japan (founded 1950—now 10 007 members), the Acoustical Society of Japan (founded 1936—now 2600 members), and the Astronomical Society of Japan (founded 1908—now 2400 members) are the other Japanese physics societies.

Australians have made significant contributions to physics, in spite of having had limited resources of funds and personnel. Their general physics society is the Australian Institute of Physics. Their Institute was founded in Victoria in 1963, has 2400 members, and publishes *The Australian Physicist* (monthly). Two other Australian physics societies are the Australian Acoustical Society (founded 1971—now 390 members) and the Astronomical Society of Australia (founded 1966—now 250 members).

The final nation in this grouping is China. China has been slow to develop technologically; nevertheless, there is tremendous potential that will eventually be tapped when the Chinese accept the importance of free and open exchange of scientific and technological information. Their general physics society is the Chinese Physical Society (CPS) founded in Beijing in 1932. The CPS has 25 000 members and publishes various journals, mostly in the Chinese language, such as *Advances in Physics, Chinese Physics Letters, Wuli (Physics), University Physics,* and *Physics Education.* The CPS has nine subsocieties: the Chinese High Energy Physics Society, the Chinese Nuclear Physics Society, the Chinese Liquid Crystal Society, the Chinese Electron Microscopy Society, the Chinese Mass Spectrometry Society, the Chinese Luminescence Society, the Chinese Particle Accelerator Society, the Chinese Gravitation and Relativity Astrophysics Society, and a Special Society on Phase Diagrams (Anon., 1990). The Chinese Optical Society (founded 1979—now 9019 members), the Acoustical Society of China (founded 1985—now 2879 members), and the Chinese Astronomical Society (founded 1908—now 2400 members) are three other Chinese physics societies.

Of the three nations that make up North America, Canada has two physics societies, Mexico has two, and the U.S.A. has ten member societies of the American Institute of Physics (AIP) (see Table 4). Representatives of the Canadian Association of Physicists, the Mexican Physical Society (Sociedad Mexicana de Fisica), and The American Physical Society have met to identify feasible steps to strengthen physics in North America. Joint meetings, long-term exchange agreements, and reciprocal memberships are some of those steps, some of which have been carried out.

2.3.2 Details of Selected Societies

2.3.2.1 The American Physical Society. The American Physical Society (APS) is the oldest and largest of the U.S. physics societies (Phillips, 1990). The APS was formed in 1899, the same year the American Astronomical Society (AAS) was formed. The two organizations went their independent ways until 1966.

The original intent of the APS was to focus on meetings. They did so by organizing joint meetings with AAAS's Section B until 1940. Journals joined meetings in 1903 as important communication vehicles. At that time, the APS began "cooperation" in the publication of *The Physical Review* with Cornell University editors who had initiated the journal in 1893. *The Physical Review* was subsequently published by the APS beginning in 1913, published for the APS by the AIP from 1932, and, as of 1975, is now published by the APS through the AIP.

The meetings and journals arranged by the APS displayed that pure physics, very narrowly defined, was to be the province of the APS. "APS was elitist from the beginning. It was expedient, and undoubtedly wise, to persuade the most prestigious physicists in the country to take leadership positions in the new society.... Members might raise questions of applications and of pedagogy, but the decisions of the APS Council did not reflect these concerns.... Applied physics and even fundamental physics related to applications suffered much the same neglect" (Phillips and Wilson, 1989).

Obviously, the most prestigious physicists who were leaders of the APS in the early 1900s recognized their responsibilities to the profession. They were defining the very nature of physics by controlling the scientific content of APS meetings and journals. As some unknown seer has observed, "physics is what physicists do." Therefore, the contents of the APS meetings and journals defined what physicists did and so defined physics.

The role of APS leaders in establishing strict standards for defining physics during the formative years of the APS was appropriate. Now, a century later, when the membership has grown to more than 40 000, the society leadership is no longer autocratic. The APS has become a democratically run society that is much influenced by the current needs and interests of its membership. The Governing Council of the APS now consists of present and past officers, 23 Divisional Councillors, 12 Councillors-at-large, and two committee chairs. The total is 44. The Council meets twice a year.

Any 200 or more members who wish to advance the knowledge of a subfield of physics may petition the Council to establish a Topical Group. If a Topical Group attains a size specified in the By-laws for two consecutive years, it can become a division. Continuation of a division depends on continuation of membership interests in that division according to stated criteria. If at least 200 members wish to advance knowledge regarding the interrelation of physics with matters not exclusively in physics, they may petition the council to establish a forum.

The names of divisions, forums, and topical groups listed in Table 5 show the wide breadth of coverage of what is today considered physics appropriate for the meetings and journals of the APS. These organizational units plus 19 bylaw committees, three constitutional committees, and 31 prize committees provide the membership input to two annual general meetings, plus numerous subunit meetings and a journal program in which some 20 000 articles are published annually.

Table 5. Divisions, forums, and topical groups of the APS.

APS Divisions
1. Astrophysics
2. Atomic, Molecular, & Optical Physics
3. Biological Physics
4. Chemical Physics
5. Computational Physics
6. Condensed Matter Physics
7. Fluid Dynamics
8. High Polymer Physics
9. History of Physics
10. Materials Physics
11. Nuclear Physics
12. Particles & Fields
13. Physics of Beams
14. Plasma Physics

APS Forums
1. Education
2. International Physics
3. Physics and Society

APS Topical Groups
1. Few-Body Systems
2. Fundamental Constants
3. Instrument and Measurement Science
4. Laser Science
5. Shock Compression of Compressed Matter

Several measures of the success of The American Physical Society in achieving its purpose of "advancement and diffusion of the knowledge of physics" are available. One measure is the continued, considerable growth in the total membership of the society as plotted as a function of time in Fig. 2: from 30 000 in 1980 to 40 702 in 1993 (American Institute of Physics, 1993). Another measure is the continued growth in attendance at meetings and in the total number of pages published annually in APS journals. A final measure is the many interdisciplinary activities sponsored by the APS.

2.3.2.2 American Institute of Physics. During the first 25 years of the 1900s, in which the Council of the APS exerted strict management and editorial control, other scientific and educational interests became so important to some physicists as to threaten the formation of several new societies and the fragmentation of the APS. For example, the Optical Society of America (OSA) and the Acoustical Society of America (ASA) were founded in 1916 and 1929, respectively, as societies independent of the APS. This dispersal of physicists was of great concern to the APS. However, "unlike the American Chemical Society, which embraced all aspects of chemistry, the APS continued to take a very narrow view of its role" (Phillips and Wilson, 1989).

An APS Committee on Applied Physics in 1930 took steps to stop the organizational disintegration as well as to overcome the financial distress in their and other societies. They did so by proposing a federation of societies to be called the American Institute of Physics (Barton, 1956). The societies supporting the proposal were the APS, OSA, and ASA. Invited to join also was the Society of Rheology (SOR); and when the American Association of Physics Teachers (AAPT) was incorporated in 1932, it was invited also. Table 4 gives the name, year founded, year of affiliation with the AIP, and total current membership for each of the five founder societies together with the same data for the five societies that have joined the AIP in the more than 50 years since.

The original assignments to the AIP were the complete responsibility, including journal ownership, for the publishing programs

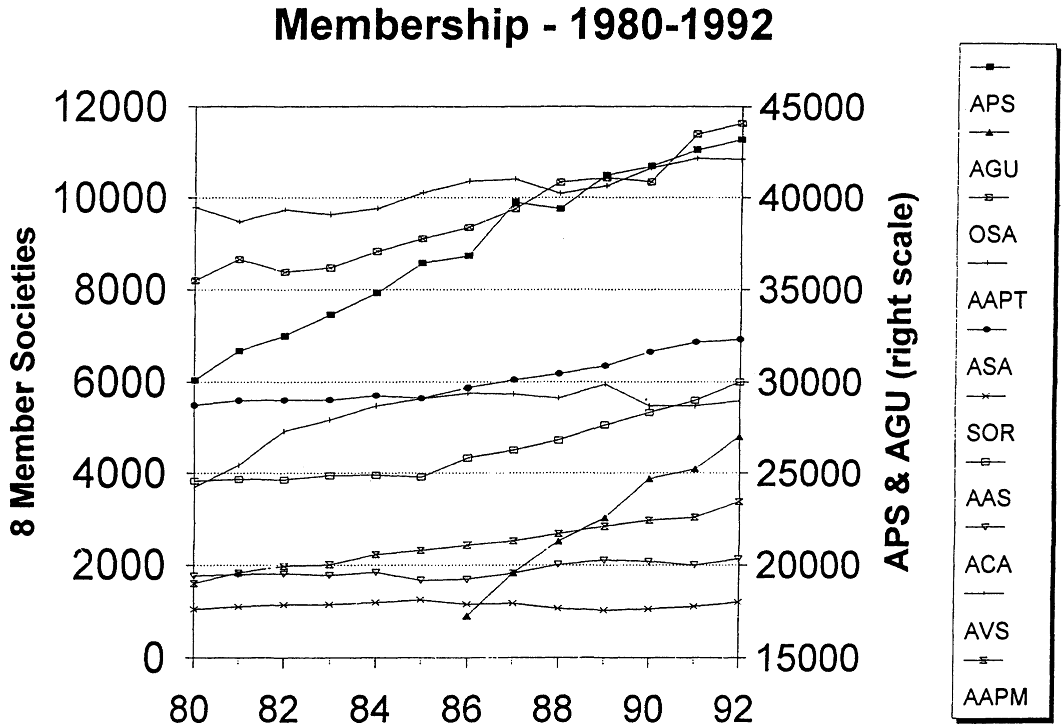

FIG. 2. AIP member society growth chart: the logarithms of total society membership plotted as a function of time.

of the societies and also some limited scientific responsibility. More specifically, the AIP was not to be a mere joint publishing department, but was "to be the joint instrument of the societies for the gaining of any of their collective ends in which success would be gained by combining" (Barton, 1956). One area of scientific responsibility was the applications of physics not covered by any other member society. Thus, several applications journals were begun by the AIP and continue to be the sole responsibility of the AIP.

Although the AIP was formed to be a "federation," in which some autonomy of the member society was to be transferred to the AIP, the development over the years has been to increase the independence of each of the AIP societies from each other and from the AIP. For this reason, the change of the publishing arrangement between the APS and the AIP cited above was made in 1975. For this reason, also, the AIP is referred to as a membership corporation and not a federation, since it is now clear that the members of the AIP are principally the societies and not their individual members.

In retrospect, the organizational relationship of the AIP and its member societies has been of great mutual benefit. The societies, who have relied on volunteer help from their memberships, have been able to focus on scientific matters, while assigning some business matters, including the development of new facilities as required, to the permanent professional staff at the AIP. Also, the early fixation on "pure" physics has made it possible for the APS to spawn new societies in close liaison with the rest of the physics community.

The first 25 years in the history of the APS must have been difficult years for their members interested in "pure" physics as well as for physicists about to form new societies focusing on physics applications and education. The difficulties were little indication of the flourishing success that has resulted in the 1990s both for the APS and for the other U.S. physics societies that have been formed separate from the APS.

Several measures of success in the development of all U.S. physics societies affiliated in the AIP are available. The most obvious is

Table 6. Type of employer for each society, 1990.

	AAPT	ACA	AAS	APS	AGU	ASA	SOR	AAPM	OSA	AVS
Academic	89	53	51	45	38	37	35	29	26	17
Industry/self	7	24	16	30	26	40	59	16	56	71
Federal	3	16	28	23	31	16	4	5	15	12
Nonprofit	1	7	5	2	5	7	2	50	3	0

the healthy growth in the memberships of all of the member societies as shown in the graph of Fig. 2. The total number of individuals represented by the 10 member societies is about 110 000 (American Institute of Physics, 1993).

Another measure is the mix of the types of employers represented among the physicist members of the member societies as shown by the percentages in Table 6 (American Institute of Physics, 1990). Note particularly that the AAPT has the largest percentage in academic institutions and that the AVS has the largest percentage in the industry/self category.

A similar measure is the unusual interdisciplinary mix of professions represented by the memberships of AIP member societies as shown by the data in Table 7 (American Institute of Physics, 1990). The mixtures of professions and of employers should lead to excellent exposures of ideas, methods, and applications in meetings and journals operated by the various member societies.

A final measure of success of the AIP created by its societies is the variety of programs operated by the AIP for the benefit of the member societies and their members that are exemplified by the listing in Table 8. Further details on publishing by U.S. physics societies can be found in the related article PHYSICS LITERATURE AND PUBLICATIONS.

2.3.2.3 Other U.S. Physics and Related-Science Societies. An additional class of membership in the AIP is the Affiliated Society that is any local, regional, or national organization interested in physics. Eighteen societies have so affiliated. They include such societies as the American Nuclear Society, the American Meteorological Society, the Geological Society of America, and the Materials Research Society. The total membership in the Member and Affiliated Societies is more than 200 000 and can be assumed to be most of the U.S. physics organizations (except for the NAS, the U.S. Liasion Committee to IUPAP, and the AAAS) of interest in this article.

Each of the Member and Affiliated Societies successfully serves important and well-defined constituencies. Several of them have organizational and historical characteristics that are particularly noteworthy in understanding U.S. physics organizations, as follows:

1. The American Astronomical Society has a history equivalent to that of the APS in that they were both founded in 1899. The AAS presently includes most U.S. astronomers though its membership is only 5900. Although physicists and astronomers are closely affiliated in U.S. academic departments, the AAS was not organizationally affiliated with other physics societies until 1966, when it became a Member Society of the AIP.
2. The American Geophysical Union, now a

Table 7. Professional self-identification for each society (1990).

Prof	AAPM	APS	OSA	AAPT	AVS	ASA	AGU	AAS	ACA	SOR
Physicist	86	73	48	44	31	21	14	12	9	4
Engineer	2	11	36	5	41	42	12	4	4	68
Chemist	1	8	4	4	15	···	4	···	57	22
Geologist	···	···	···	1	···	···	26	1	2	···
Astronomer	···	12	1	1	···	···	4	72	1	···
Geophysicist	···	···	···	···	···	1	21	1	1	···
Computer scientist	1	2	1	1	···	1	1	2	1	···
Other scientist	3	3	6	3	8	22	16	3	20	4
Other	7	3	4	42	6	13	5	6	7	3

Table 8. AIP publishing and professional programs for member societies.

Publishing	Professional
Publishing of research/ educ. journals	Studies of physics history
Monthly magazine: *Physics Today*	Press and public information programs
Computerized publishing facility	Projects on physics education
PINET electronic network for physicists	Corporate associates programs
Cooperation on *Phys. Abstracts* database	Job placement and career studies
	Society of Physics Students
	Management

very large society with 29 766 members, is like AAS in that it developed independently of the APS. Although the AGU was founded in 1919, it did not elect to join other Member Societies in the AIP until 1986.

3. The American Vacuum Society is a society that splintered off from the APS in 1953 and that elected to become an AIP Member Society in 1976.
4. The Materials Research Society is a relatively new society (founded in 1973) that has mostly physicists or chemists as members. The society considered affiliation with other Member Societies of the AIP as a Member in 1986. Instead, it elected to become an Affiliated Society. Thereby, it has developed strong ties to the other Societies in the AIP without being characterized as either a physics or chemistry society, because it indeed is both.

2.3.2.4 American Center for Physics. A common and highly visible home for the physics community in the U.S. called the American Center for Physics has been built jointly by The American Physical Society, the American Association of Physics Teachers, and the American Institute of Physics. The Center is built on a 24-acre tract of land in College Park, Maryland, 12 miles from the nation's capital in Washington, DC. The building contains 120 000 square feet of floor space, enough initially to house the headquarters of the three founder societies and at least one other member society of the AIP, the American Association of Physicists in Medicine. Most of the AIP professional programs itemized in Table 8 have been relocated to the Center. A featured element in the new building is the Niels Bohr Library, used for studies of the history and philosophy of physics.

2.3.2.5 The Institute of Physics, London. The Physical Society (founded in 1873 and since having developed into a society primarily for university physicists) and the original Institute of Physics (founded in 1918 and since having developed into a society primarily for industrial physicists) were amalgamated in 1970 to form the current Institute of Physics (IOP). Both organizations were based in London, and, upon their amalgamation, they received a Royal Charter.

The object of the Institute is to promote the advancement and dissemination of a knowledge of and education in the science of physics, pure and applied. The total membership is over 16 000, including 2400 Fellows of the Institute and 820 Fellows of the former Physical Society.

The Institute makes a great effort to be a "democratic organization which needs contributions from physicists with wide experiences if it is to be fully representive. By working through the Boards, Committees, Branches, Divisions, and Groups, members can actively participate in the work of the Institute influencing its activities and policies and representing its views to official bodies" (Institute of Physics Secretariat, 1993).

There are currently five divisions of the Institute (see Table 9) that act as forums for reviewing research and development in their respective fields of physics. The divisions collaborate with 32 subject and professional groups in arranging meetings and conferences.

IOP Publishing is a wholly owned subsidiary company of the Institute and one of the largest publishers of physics information in the world. The publishing activity has its origins in the *Proceedings of the Physical Society*, first published over a century ago. Today, it publishes over 70 new books and proceedings a year and over 30 journals.

2.3.2.6 German Physical Society. Germany has probably the longest tradition for educating physicists of any nation in the world. The universities in Germany were the schools of choice for most of the nineteenth century for many students who wanted to enter careers in physics and for many U.S. physicists. Also, the German Physical Society

Table 9. Divisions and groups of IOP.

IOP Divisions
1. Applied Optics
2. Atomic, Molecular, and Optical Physics
3. Condensed Matter and Materials Physics
4. Engineering Physics
5. Nuclear and Particle Physics

Subject and Professional Groups
1. Atomic Collisions in Solids
2. Carbon
3. Combustion Physics
4. Computational Physics
5. Consultancy
6. Education
7. Electronics
8. Electron Microscopy and Analysis
9. Environmental Physics
10. High Energy Particle Physics
11. History of Physics
12. Instrument Science and Technology
13. Low Temperature
14. Magnetism
15. Materials and Testing
16. Mathematical Physics
17. Neutron Scattering
18. Nuclear Physics
19. Optical
20. Physical Acoustics
21. Physical Crystallography
22. Plasma Physics
23. Polymer Physics
24. Printing, Packaging, and Papermaking
25. Quantum Electronics
26. Semiconductor Physics
27. Spectroscopy
28. Static Electrification
29. Stress Analysis
30. Thin Films and Surfaces
31. Tribology
32. Vacuum

(GPS) (Deutsche Physikalische Gesellschaft) founded in 1845 represents the oldest physics society in the world as well as the largest in Europe with 18 400 members.

Thus, it is useful to compare this tradition and history with the relatively small size of the publishing program of GPS. In 1992, they published only two primary publications, called *Physikalische Blätter* (monthly) and *Verhandlungen der DPG* (eight issues per year). They also had an abstracting service called *Physikalische Berichte* (every two weeks). After a study of the large number of members and the small number of authors in their publications, they concluded that their solution would be to encourage authors in their publications to switch to the use of the English language, not only for their primary journals but also for their abstracting service. The English-language abstracting service is now called *Physics Briefs* and is highly successful because it is available worldwide on a computer network called STN International (Scientific and Technical Network International).

2.3.3 Other Physics Societies in the Rest of the World Many industrialized nations may not have physics societies, though they may have chemical, geology, mineralogy, meteorology, or other natural-science societies. The small number of physicists that are educated each year compared with the other natural-science professionals in all countries is the reason. Therefore, the occurrence of physics societies may be spotty in any given region of the world.

A few of the outstanding physics societies not covered by this article are the following:

1. The Indian Physics Association (IPA) was founded in Bombay in 1970. IPA has 30 chapters in different cities of India and has a total membership of about 2500.
2. The Malaysian Institute of Physics was founded in Malaysia in 1973 and has 60 members.
3. The Institute of Physics, Singapore, was founded in 1972. It has 237 members.
4. The African Union of Physics was founded in Oyo, Nigeria, in 1982.

3. ENGINEERING ORGANIZATIONS

3.1 General

Engineering has strongly influenced the development of human society. As the capability of engineers has increased, many technologies have been created that have extended and are continuing to extend human abilities. Examples are the ability to travel more widely and rapidly, to communicate globally and almost instantaneously, to control large amounts of energy, to make many new types of materials, and to modify environmental conditions. Engineers have assisted in effecting these changes and, in turn, have been affected by them. A primary means by which engineers interact with society is through the many organizations the profession has developed. An examination of these engineering organizations can provide insight on the functioning of the engineering profession and how it interacts with society.

In the U.S.A., engineers have created a number of nongovernmental and not-for-profit organizations during the last 150 years during which the modern profession emerged. These organizations range from those concerned with advancing the knowledge base of the profession and educating future engineers to those concerned with industrial practice and public policy on engineering-related issues. They reflect the common interests of engineers as a community and represent these interests to other elements of society. They operate in local, regional, national, and international contexts.

3.1.1 Organization of This Section Engineering-related organizations may be considered in four categories. The first includes the technically oriented organizations whose major concern is the advance of engineering knowledge and know-how—the research and development that make advances possible. The major group in this category, the engineering professional societies, have developed along both disciplinary and industrial sector lines.

In a second category are the education-oriented societies. Although the professional societies and some other engineering-related organizations carry on educational activities, there are several organizations primarily devoted to education-related activities. One such is the Accreditation Board for Engineering and Technology, which accredits engineering and engineering technology postsecondary educational programs.

The third category includes industrially oriented organizations that carry on engineering-related activities. In many industrial sectors, trade associations have been organized that support the engineering interests of their member companies. Another type of organization in this category comprises the standards-writing organizations, which develop the standardized procedures and processes used in many commercial transactions.

A fourth category includes organizations with a major focus on public policy related to engineering. Principal among these are the National Academy of Engineering, the National Society of Professional Engineers, the American Association of Engineering Societies, and the American Association for the Advancement of Science.

These four categories are in no sense mutually exclusive. Most engineering organizations have activities in several categories. Each organization is described here in the category that represents its major activity and the one in which the organization is most widely recognized.

3.1.2 International Counterparts There are counterpart organizations to these U.S.-based organizations in other nations. In the highly industrialized nations, there are usually well-developed organizations with a broad range of activities similar to those in the U.S.A. The national professional societies are typically organized along disciplinary lines, with the holding of meetings, publication of technical periodicals, and recognition of outstanding technical achievements as their principal functions.

However, there are also significant differences among the developed countries. In the United Kingdom, for example, the certification of professional competence, or registration, of engineers is linked to the membership category of the relevant engineering professional society. Thus, "chartered engineer" is a title conferred by a professional society and is used to designate someone recognized as qualified for the public practice of a particular branch of engineering. In the U.S.A., registration of engineers is a responsibility of the state governments, which they carry out with the assistance of the various professional societies. Such national differences occur in many of the industrialized countries.

Strong professional organizations are not the norm in most developing countries. Professional engineers in these countries often lack the benefits of good technical periodicals and regular meetings and conferences. As a consequence, they are not as able to develop effective technical networks with fellow specialists. Having recognized this weakness, some developing countries are in the process of organizing a professional society infrastructure to compensate.

It is beyond the scope of this article to survey the numerous engineering organizations that have been established throughout the world. But it is noteworthy that, in general, engineers everywhere have found that professional societies provide them with services essential to the well-being and health of their professional communities.

There are a number of international organizations in which U.S. engineering societies are members; among these are the World Federation of Engineering Organizations and the Union of International Technical Associations. Typically, their purpose is to facilitate collaboration among engineers throughout the world. While these two organizations operate on a global basis, others operate on a regional basis in a major geographical area, such as Latin America or Southeast Asia.

Some exceptions to the norm of engineering professional societies being organized on a national basis can be found. In the U.S., the most notable is the Institute of Electrical and Electronics Engineers (IEEE), which has a transnational organizational structure. It is described in detail in Sec. 3.3.4. Also, the United Kingdom system of "chartered engineers" is recognized in many other countries that are or have been part of the British Commonwealth.

3.2 Inherent Characteristics of Engineering

3.2.1 Integration of Knowledge Engineers both generate and apply knowledge. Thus, the activities of engineers range from research and development, in which the creation of new knowledge is the principal objective, to manufacturing, construction, and other operations in which knowledge is applied to deliver products and services. To encompass this wide range of functions, engineers draw on and integrate a variety of sources of information and knowledge. The process of integration, drawing on knowledge and experience from many fields, is an inherent characteristic of engineering. The services and products that are the output of engineering must meet the requirements of customers and competition in the marketplace.

3.2.2 Relation to Mathematics and Science Mathematics and the natural sciences provide much of the knowledge needed by engineers. As these disciplines advance, they give engineers the ability to improve on existing products and services and to create new ones. However, as the complexity of the products and services has increased, engineers have had to undertake fundamental research to improve the performance of the systems they have created and to understand their limitations. For example, as communication systems became more powerful and complex, a deeper understanding of their characteristics was needed. This need was met by the development of information theory at the Bell Telephone Laboratories in the 1940s. It enabled design engineers to determine the information-carrying–capacity limitations of existing systems and to improve their performance. While engineers, particularly in academic environments, carry on fundamental research driven by individual initiative as well as research in strategic areas, engineering research in industrial laboratories is primarily driven by needs of a strategic nature. The traditional engineering disciplines are linked closely to the most relevant science: chemical engineers to chemistry, electrical and mechanical engineers to physics, and civil and mining engineers to geology.

As the complexity of engineering-designed systems has increased, these traditional knowledge sources have been complemented by others. New analytical techniques have been created to handle the greater complexity. The field of systems engineering has gradually evolved to enable highly complex systems to be designed, produced, and operated. To optimize the performance of systems that involve both complex instrumentation and human operators requires improved analysis, modeling, and simulation techniques. Engineers who specialize in the design and production of such complex systems are no longer closely linked to one of the natural sciences.

The characteristics of engineering research and an identification of some research fields in which there are fruitful opportunities are discussed in a report written by a panel of the National Research Council Engineering Research Board (National Research Council, 1987). Figure 3, from this report, illustrates the relationship between

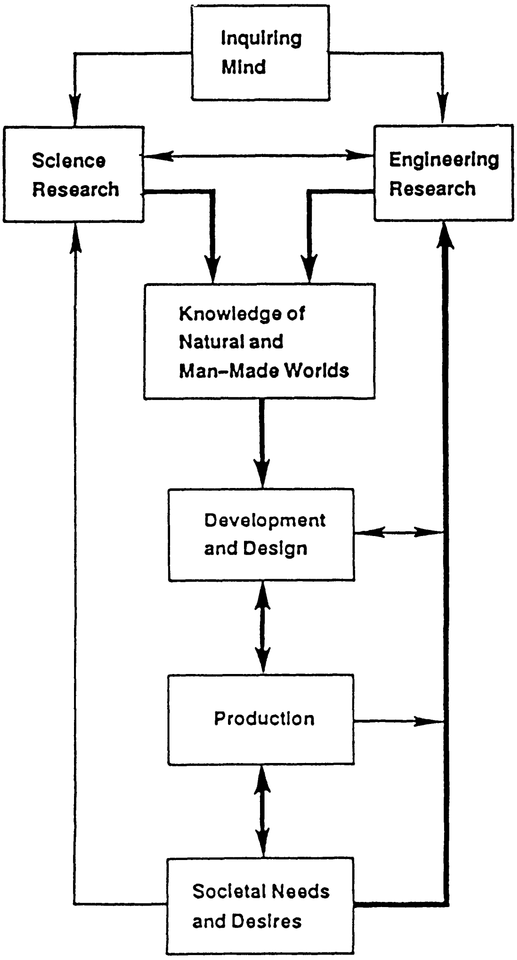

FIG. 3. Relationship between engineering and scientific research (from National Research Council, 1987).

engineering and scientific research in the perception of this panel. The broad distinction is made that scientific research seeks to gain more knowledge about the natural universe, while engineering research seeks more knowledge of the man-made world. Thus, technological advances can open up new branches of engineering, which become areas in which new research efforts are launched. One such branch of engineering, which is still in an early stage of development, is software engineering (*q.v.*). Its fundamental concepts are still in the process of being identified (Gibbs, 1991).

3.2.3 Societal Impact As technologies of increasing power have been developed, engineering's effect on human society has become more far-reaching, and the knowledge sources that engineers must draw on have widened. As medical technology has advanced, engineers have worked more closely with physicians and biologists. The need to control pollution and to protect the environment has required engineers to work closely with behavioral scientists as well as government officials. To meet competition in the global marketplace, engineers must work with economists and international experts. To meet changing conditions, engineers must integrate not only quantitative knowledge in their design and production of products and services but also the nonquantitative knowledge that characterizes many nonscientific disciplines.

The existing engineering societies have responded to these new demands by creating new publications and jointly sponsoring meetings. In some cases, new societies have been formed to facilitate the development of new fields of knowledge. The new modes of handling information have resulted in computer experts and other engineers working with specialists in medicine, education, organizational behavior, and governmental affairs as well as the natural sciences. Expanding frontiers have put new requirements on both individual engineers and the organizations that serve their interests.

3.3 Technically Oriented Organizations

3.3.1 Services of Professional Societies

The engineering professional societies, organized to provide services needed by the developing engineering professions, have, in turn, influenced the actual development of engineering. Their primary service is to provide an efficient and cost-effective way for their members to communicate with other specialists in their fields. Through meetings and publications, the societies bring geographically dispersed engineers together intellectually, to learn from each other both the theory-based and the practice-based knowledge needed to advance the engineering state of the art. New research results as well as information about new instrumentation and techniques are rapidly disseminated to those working in the field or those interested in new applications. In addition, they provide the crucially important archival function of recording and preserving the evolving knowledge base in their discipline.

Second, they provide their members an opportunity to address issues of common concern, whether related to education of future engineering practitioners, to industrial practice, or to governmental policies. In a number of these arenas, professional societies provide the most effective means of developing, presenting, and implementing positions that reflect the judgment of the engineering community.

Third, they give their members recognizable credentials. By requiring those seeking society membership to meet specified professional criteria as determined by peers, the societies give their members professional recognition. By identifying the outstanding work of individual engineers, they encourage and reward above-average professional achievements. Through a system of membership grades and awards, individual accomplishments are publicized to peer groups and to external communities.

3.3.2 Founder Societies In the U.S.A., the first engineering professional societies were organized in the latter half of the nineteenth and early twentieth centuries. These first societies later became known as the Founder Societies.

The first, the American Society of Civil Engineers founded in 1852 (Wiseley, 1974), had the task of establishing the concept of civil as distinct from military engineering. Prior to the 1850s, the Military Academy at West Point was the major source of U.S. engineering graduates; there was little recognition of engineering as a civilian activity crucial to the construction of bridges, railroads, canals, and water supply and sewage disposal systems. Within the next few decades, the American Institute of Mining Engineers (1871), the American Society of Mechanical Engineers (1880), and the American Institute of Electrical Engineers (1884) were formed. As a result of the frequent explosions of poorly designed boilers during the late 1800s and the consequent danger to the public, the new mechanical engineering society was immediately involved with the development of safety standards for boilers, an activity that persists to the present day. Also during this period, there were major new technical developments in electrical power generation, lighting, and telegraphic and telephonic communications. The new electrical engineering society had a major influence in these fields by initiating the publication of technical papers on topics in which secrecy to protect commercial advantages had previously prevailed.

In the early 1900s, chemical engineering developed as a discipline distinct from both chemistry and mechanical engineering. The American Institute of Chemical Engineers (Reynolds, 1983) played a major role in defining the field and designing appropriate undergraduate curricula. Also at this time, the invention and use of wireless communications resulted in the formation of a new but rapidly growing society, the Institute of Radio Engineers.

These Founder Societies combined their efforts to build facilities in New York City to house their headquarters. A gift of more than $1 million in 1903 from Andrew Carnegie, an engineer who became a successful businessman, provided impetus. Appendix A includes detailed information about the organizational structure, budget, membership, publications and meetings, and other activities of the Founder Societies.

3.3.3 Organizational Characteristics As the various engineering disciplines matured, the societies developed internal structures with both geographical and technological components. The geographical units typically feature local and regional meetings and support student activities at local schools and universities; the technological units focus on national conferences and technical publications. Both entities provide useful membership services. The volunteer officers, elected by the members, typically reflect both of these structural components.

The membership requirements of the Founder Societies as well as the more recently formed engineering societies are usually similar. The typical requirement is a baccalaureate degree for the entering member grade, while senior grades are given to more experienced engineers. Student grades are used to encourage those preparing for engineering careers to develop university-based engineering activities and to facilitate their joining the appropriate society after graduation. The top membership grade is frequently restricted to a small percentage of the members elected in recognition of outstanding accomplishments.

3.3.4 Unique Character of the IEEE Many of the U.S. engineering societies have the term "American" in their title and consider themselves national organizations. Although they accept overseas members, they do not directly compete with their counterpart societies in other countries. A major exception is the Institute of Electrical and Electronics Engineers (IEEE), which was formed by the merger of the American Institute of Electrical Engineers and the Institute of Radio Engineers in 1962 (McMahon, 1984).

Although headquartered in the United Engineering Center in New York with the other Founder Societies, the IEEE organized and structured its geographical and technological activities on a global basis. It has attracted approximately 50 000 overseas members, one-sixth of its total membership of more than 300 000. Many meetings and conferences are held outside the U.S.A. These meetings and the Institute's publications are widely considered the most prestigious in their fields.

The IEEE has developed into a federation of more than 30 societies in the various specialties into which the electrical, electronics, and computer fields have evolved. The Institute's internal structure readily permits the organization of additional specialty groups as new technological fields emerge (Ryder and Fink, 1984). For example, the explosive growth of the computer field after World War II was accompanied by the rapid expansion of the newly organized Computer Society of the IEEE. It now has a membership of more than 70 000 and publishes 17 technical periodicals.

The IEEE Computer Society growth has been paralleled by the growth of another professional society in the computer field, the Association for Computing Machinery (ACM). In their early development, these societies were primarily focused on hardware and software, respectively. But as the computer technology advanced, this separation was no longer valid; both societies currently have publications in software engineering. The emergence of this latter field of engineering is being studied by joint committees of the two societies. With strongly overlapping memberships, the two societies have both flourished and have cooperated. Appendix B gives information about these societies' membership and activities and illustrates some of their complementarity.

3.3.5 Nontechnical Activities While the primary activity of the engineering professional societies has focused on the advance of the discipline, there has also been concern about engineering careers. For example, are enough students being attracted to the field? What are the salary ranges and fringe benefits available to practicing engineers? What are the legal obligations of engineers in public practice? The societies have responded by undertaking professional—in contrast to technical—activities; many standing committees and boards have been established to address such issues.

Currently, almost all major U.S. engineering societies have offices or representatives in Washington, DC, and devote both staff and volunteer time to relations with the federal government. On one hand, they serve as a valuable resource to federal agencies and to the Congress in supplying specialized technical information. On the other, they seek to avoid legislation or regulations that are considered to be inimical to the engineering profession. In this arena, the societies operate differently; most are still exploring how to respond to changes in the environment in which engineering is practiced.

3.4 Educationally Oriented Organizations

3.4.1 Studies of Engineering Education The engineering professional societies have had a strong interest in academic programs in engineering and have had considerable effect on their development. In addition to the educational activities of the disciplinary societies, another society, the American Society for Engineering Education (ASEE) devoted solely to engineering education, was founded in 1893. It has served as the sponsor for a number of major reports on engineering education that have influenced the philosophy and content of engineering education at the undergraduate and graduate levels. Its membership includes primarily engineering faculty members with strong interest in undergraduate and continuing education and teaching methods (Reynolds and Seely, 1993). In addition to engineering faculty members, it also includes engineering technology faculty members who teach in two- and four-year colleges offering degrees in engineering technology. Most ASEE members also belong to at least one disciplinary society.

There have been a number of major ASEE studies of engineering education spanning more than 70 years. Although none of them has had an effect as great as the Flexner report on medical education in the 1920s, they have served to highlight the problems that engineering educators have faced and to point to new directions. In the post–World War II era, one report written in 1952–1955 was notable for involving a large segment of the academic engineering community in gathering data and in stimulating discussions. The wartime experience at laboratories such as the MIT Radiation Laboratory had demonstrated that engineers educated before the War did not have the academic background needed to contribute to such new engineering developments as radar and microwave communications. The fact that the authors of the 25-volume series of books published by the MIT Radiation Laboratory in the late 1940s were primarily physicists and mathematicians had a profound effect on engineering educators. As a result, much of the handbook-based material in undergraduate curricula was replaced with more advanced courses in calculus, physics, and chemistry. The "Report on the Evaluation of Engineering Education" by an ASEE committee chaired by L. E. Grinter (American Society for Engineering Education, 1955) reflected this trend. It urged increased emphasis on what were labeled the engineering sciences and less emphasis on current engineering practice. Thus, postwar engineering curricula differed from the prewar in requiring engineering students to gain much greater depth of understanding in mathematics and the physical sciences.

Another major ASEE study resulted in the publication of a report entitled "The Goals of Engineering Education" (American Society for Engineering Education, 1968). By drawing an analogy between engineering and medical education, this report sought to replace the four-year undergraduate programs as preparation for an engineering career and require a master's degree as the first "professional" degree. Such proposals, although frequently discussed, have never gained the endorsement of either the academic community or industrial companies, the major employers of baccalaureate-degree engineers. Thus, engineering remains one of the few professions for which the baccalaureate is considered to be a professional degree. In some engineering specialties, such as sanitary engineering, the first professional degree is at the master's level. However, students undertaking such programs frequently complete a four-year civil engineering program first and so already have a professional degree.

A 1987 ASEE report, "A National Action Agenda for Engineering Education" (American Society for Engineering Education, 1987), chaired by Edward E. David, Jr., endorsed the four-year baccalaureate programs as the appropriate "launching platform" for an engineering career. The report also recommended a practice-oriented master's degree as a follow-on for students interested in technical specialties where greater academic depth is needed. This degree would complement the research-oriented Ph.D. degree. Much current thinking in engineering education is reflected in this report.

3.4.2 Accreditation of Academic Programs The disciplinary societies are also active in various aspects of engineering education. They frequently support programs concerned with the continuing education of their members as well as for students majoring in their respective specific disciplines. They exert significant influence on the content of undergraduate curricula through their participation in the accreditation process.

The Accreditation Board for Engineering and Technology (Accreditation Board for Engineering and Technology, 1993) is a creation of the engineering professional societies. ABET accredits programs in the various engineering disciplines in schools of engineering and of engineering technology. Both federal and state agencies recognize ABET as the accepted accrediting body for these fields. The ABET board of directors includes representatives of the professional societies that have curricular responsibilities, as well as some other engineering-related organizations. The professional societies develop the criteria used by accreditation teams that visit engineering schools and inspect their academic programs at three- or six-year intervals. Team members, selected from lists of engineers submitted by the disciplinary societies, typically include engineers employed by industry as well as by universities.

Essentially, the accreditation process sets a floor for the academic quality of the approved programs. While the accreditation process keeps pressure on all the more than 300 U.S. engineering schools and 120 engineering technology schools to keep their courses and laboratories up-to-date, many academicians feel that the accreditation process stultifies innovation. Currently, a number of innovative undergraduate programs are being developed as a result of the availability of curriculum development funding from the National Science Foundation. These programs are testing the ability of ABET to respond to nontraditional programs.

ABET has a powerful position with respect to the engineering schools because of its link to the legal process of registration. Registration, which is required for engineers in the public practice of engineering, is controlled in the U.S.A. by state governments. The state registration boards recognize graduation from an ABET-accredited school as completing a major requirement for registration. Since schools want their graduates to qualify for registration, they must retain ABET-accredited programs. Some schools offer both accredited and nonaccredited programs to allow some students greater flexibility in course selection. The National Council of Engineering Examiners, the organization of the state engineering examiners, is also represented on the ABET board of directors.

3.5 Industrially Oriented Organizations

In this category are two major types of engineering-related organizations: those that develop standards and the industrial trade associations. While these latter organizations typically engage in many nontechnical activities, some of their programs have significant engineering content.

3.5.1 Standards-Writing Organizations Much engineering practice depends directly on the use of standards, that is, concepts and procedures systematically developed and approved by experts in a technical specialty. Standards development is primarily an industry-based activity since industrial products and processes must usually meet standardized requirements set by customers, regulatory authorities, and other external bodies. The term *standards* is used to include methods of test, technical specifications, definitions and terminology, and recommended practices.

In the U.S.A., in contrast to most other industrialized countries, the writing of voluntary standards other than those related to national defense is largely a private-sector activity (Toth, 1991). Committees of engineers are convened from a variety of companies, testing laboratories, public agencies, and universities to draft voluntary standards. These documents are then subjected to wide review and comment before being accepted and put into general use. Since a consensus is sought among divergent interests, the time needed to develop a standard can vary from about six months to several years. Standards were first widely used in the specification of commercially important materials such as metals and plastics and hardware devices but have now also become an important means of specifying processes and procedures in many industrial sectors such as telecommunications and computers.

The U.S. government is involved in the standards-writing process through technical support provided by federal agencies such as the National Institute for Standards and Technology (NIST), formerly the National Bureau of Standards, and through the participation of federal engineers on standards-writing committees. Of great importance in both domestic and international commerce, standards can serve to facilitate or impede trade. Some privately developed voluntary standards adopted by public agencies attain the force of law. A prime example is building construction: Most mandatory building codes cite many voluntary standards that have been developed in the private sector.

The U.S. standards-writing community currently maintains some 94 000 standards. While about a third of these relate to the Department of Defense, more than a half were developed and maintained by the private sector. Among the leading standards-writing professional societies are the Society of Automotive Engineers, the American Society of Mechanical Engineers, and the Institute of Electrical and Electronics Engineers. Among the leading trade associations active in standards-writing are the Aerospace Industries

Association, the Association of American Railroads, and US Pharmacopeia.

There are also several other organizations whose primary function is standards development. Principal among these are the American Society for Testing and Materials, the National Fire Protection Association, and Underwriters Laboratories, which also provides certification services. The private-sector American National Standards Institute (ANSI) serves as a coordinator for domestic standards and represents the U.S.A. internationally in standards-writing activities.

3.5.1.1 International Standardization. Complementing the domestic U.S. activity are the international standards-writing organizations and national standards-writing societies in other countries. In most countries, standards-writing is considered both a private- and a public-sector activity, and public funding supplements private-sector resources. Such countries are better organized than the U.S.A. to participate systematically in international activities since they are not as totally dependent on private company personnel. However, when the industrial sector is well organized and U.S. companies are in a technological leadership position, the U.S.A. is very effective in international standards activities.

There are several hundred regional and international organizations conducting standards-related activities (Breitenberg, 1989); the U.S.A. participates in almost 200. Some organizations have primarily scientific concerns in fields such as physics, chemistry, and geology and devote their efforts to developing consistent terminology and agreed-on procedures for data handling. There are also engineering-related groups that cover the entire range of technical subjects with which U.S. domestic groups are concerned.

The major umbrella-type international organizations for engineering-related issues, the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), establish standards that are increasingly accepted internationally. The absence of agreed-on international standards can be an effective nontariff barrier to trade. The U.S. member in these organizations is ANSI. The major federal agency concerned with standards, the Department of Commerce, has no direct role in representing U.S. interests in ISO or IEC. There is general agreement in both the private and public sectors that this arrangement is far from optimum.

3.5.2 Trade Associations The 1993 issue of the Encyclopedia of Associations lists more than 23 000 national associations in the U.S.A. Among these are trade associations that represent the interests of their industrial sectors in political and public relations as well as in technical arenas. These arenas are not always separable in issues such as environmental protection, where there are strong political forces as well as scientific and engineering factors involved.

Trade associations in some major industrial sectors such as chemical processing, semiconductors, gas technology, and electric power have undertaken research programs to solve problems generic to their technology. Most such efforts have been funded from private sources; but increasingly, public funding is being provided to support generic technology development. Since changes in antitrust legislation have encouraged industrial companies to carry on cooperative research programs, it is probable that even more industrial associations will undertake engineering projects in the future.

Trade associations are also active in many other industrialized countries. In the United Kingdom, for example, there are trade associations in nearly every industrial sector. Their activities include not only presenting the point of view of their industry to the government but also the dissemination of information about issues such as health and safety, environmental protection, and other topics of general concern. They also recommend "best practices" for use by their member companies. In other European countries, such associations are also active. Currently, these associations are developing links throughout the European Community to bring their influence to bear throughout the Community.

In Japan, trade associations are very important actors in both the technical and the public policy arenas (Bloom, 1990). Associations can be licensed by the Ministry of International Trade and Industry (MITI) and in some cases receive modest funding. They have the significant function of facilitating communication between public policy makers and the various industrial sectors and providing a forum where consensus positions are developed. On the technical side, they are sometimes organized on a temporary basis to carry out a specific research program. On a continuing basis, many associations provide services for their members including collecting and disseminating information from both domestic and international sources, development of standards, and educational programs.

3.6 Public Policy Oriented Organizations

3.6.1 National Academy of Engineering In varying degrees, all the engineering-related organizations discussed are concerned with public policy issues. Most of them operate at the federal level, although some have significant operations at the state and local levels. Their ability to affect legislative action or executive branch program decisions varies. Among the most influential is the National Academy of Engineering (NAE). The Academy was organized in 1964 under the congressional charter of the National Academy of Sciences, granted in 1863, which charges the Academy to respond to requests from federal entities for advice and recommendations. As part of the National Academy complex, the NAE participates in the governance and operation of the National Research Council, which produces about 300 reports a year in response to requests from executive and legislative federal agencies. These reports, which cover a wide variety of scientific and technological topics, involve some 10 000 scientists and engineers serving as volunteers. The NAE has initiated a number of its own studies on broad issues such as sustainable technology and national strategies in an era of global technologies. As an honorary society of approximately 1700 leading engineers from industry, academia, and government, the NAE has become increasingly effective in its relatively brief span of existence (Edson, 1989).

3.6.2 National Society of Professional Engineers The National Society of Professional Engineers (NSPE) is a society that restricts its membership to registered engineers (Robbins, 1984). It was established more than 50 years ago to consolidate the activities of a number of engineering associations operating at the state level. The state societies are primarily concerned with the registration process by which engineers are certified as competent to practice engineering publicly. These state societies frequently bring considerable influence to bear on engineering-related issues with which state governments are involved. These issues range from policies of state agencies concerned with engineering projects to budgetary allocations for engineering schools in state universities. NSPE's headquarters in Washington is primarily concerned with federal legislation and monitors federal programs that involve the interests of registered engineers. The society is active in promoting the recruitment of young engineers, in surveying salaries of practicing engineers, and in analyzing ethical and legal issues confronting engineers.

3.6.3 Umbrella Organizations To increase its effectiveness in influencing public policy, the engineering community has made several attempts since the 1920s to form a single organization that speaks for the entire engineering profession. The current umbrella society is the American Association of Engineering Societies (AAES), which includes in its membership the engineering professional societies and associations listed in Table 10. The combined membership of its constituent societies is some 700 000 U.S. engineers. The positions it develops on public policy issues must be compatible with those of the major member societies, which are generally unwilling to abdicate power on issues vital to their members. Since the member societies possess technical expertise in their disciplinary fields, they usually prefer to speak to such issues directly before congressional or executive branch committees.

However, there are issues on which most engineers can agree. Examples are more portable pension plans, improving engineering's public image, and forging better links with international engineering organizations. In such arenas, the AAES serves as the voice and surrogate for U.S. engineering societies. It is the U.S. member in international engineering organizations such as the World Federation of Engineering Organizations (WFEO/FMOI) and also in regional groups such as the Panamerican Federation of Engineering Societies (UPADI), a federation of engineering societies in Latin and North America. In general, these groups have not been as effective as the International Council

Table 10. List of member, associate, and regional societies in the AAES as of 1993–1994.

- Member Societies
  - American Academy of Environmental Engineers
  - American Indian Science and Engineering Society
  - American Institute of Chemical Engineers
  - American Institute of Mining, Metallurgical and Petroleum Engineers
  - American Institute of Plant Engineers
  - American Nuclear Society
  - American Society for Engineering Education
  - American Society of Agricultural Engineers
  - American Society of Civil Engineers
  - American Society of Mechanical Engineers
  - Institute of Electrical and Electronics Engineers
  - National Institute of Ceramic Engineers
  - National Society of Professional Engineers
  - Optical Society of America
  - Society of Fire Protection Engineers
  - Society of Women Engineers
- Associate Societies
  - American Council of Independent Laboratories
  - Association for International Practical Training
  - Federation of Materials Societies
  - National Action Council for Minorities in Engineering
  - National Association of Corrosion Engineers
  - Operations Research Society of America
  - Tau Beta Pi Association
- Regional Societies
  - Engineering Society of Detroit
  - Engineers' Society of Western Pennsylvania
  - Washington Society of Engineers
  - Western Society of Engineers

of Scientific Unions (ICSU) in launching programs that involve technical specialists and financial resources from many nations.

Within the U.S.A., the American Association for the Advancement of Science (AAAS) (see Sec. 2.2.1) has developed activities that affect both science and technology policy. While most engineers regard the AAAS as primarily a spokesman for the sciences, many engineers have been prominent in its activities. The major engineering professional societies are affiliates of the AAAS. They are active participants in the AAAS Congressional Fellows program, which has supplied the Congress with fellows drawn from the ranks of midcareer scientists and engineers to serve as members of congressional staffs. The AAAS annual analysis of the proposed federal research and development budget has involved several of the engineering societies. The editorial board of the AAAS's weekly publication *Science* usually includes engineers.

Although engineers constitute one of the largest professional groups in the U.S.A., they have not been a powerful influence on public policies. This lack of effectiveness may be traced to the reluctance of individual engineers to participate actively in politics. Engineers who work in industrial companies may also feel obligated to defer to their employer's views on public policy issues. Few engineering organizations have developed the skilled staffs or devoted the necessary resources to become forceful advocates of engineering interests in public policy arenas.

3.7 Challenges for Engineering Organizations

Engineering organizations are challenged to respond effectively to both rapid technological changes and worldwide changes in the industrial, political, and social environment. Engineers who are at the forefront of technological advances must be able to adapt to both types of change and still retain their operational effectiveness. Cutting-edge engineers need appropriate means to communicate with colleagues throughout the world. Those professional societies that provide such domestic and global communication services efficiently and cost-effectively will continue to grow. Others may well disappear.

Engineers will also require educational services beyond those currently available. Education is being increasingly recognized as a career-long activity for practicing engineers. The content of such educational programs should be determined cooperatively by active engineers in industry, government, and universities. The engineering organizations can contribute by providing educational materials and also the means for delivering them. While the problem of designing and delivering career-long educational programs is largely unsolved, the need is growing and engineering organizations have many assets to offer.

As the marketplace for engineering-intensive products and services becomes more global, engineering operations are being carried on in many cultural contexts. Many multinational corporations currently operate with relative ease in many different nations with little concern about national boundaries. Both individual engineers and the organizations they have created must learn how to operate effectively in this changing environment.

As new technologies affect more and more the way industry, government, and academia operate, engineers must not only add to their technical competence but must also be concerned with the nontechnical implications of what they create. The broader issues can no longer be considered someone else's responsibility. If engineering organizations help impart such broader perspectives, they will continue to serve as well as influence the future development of the profession.

APPENDIX A. ENGINEERING'S "FOUNDER SOCIETIES"

The first organized of the major engineering professional societies are known as the Founder Societies. These societies have played a crucial role in the development of the United Engineering Center in New York City, where many engineering societies maintain their headquarters. This Appendix presents data about each of these five societies: the number of members, their publications, and their stated objectives. The Founder Societies are listed in the order of their formation: American Society of Civil Engineers (ASCE) in 1852; American Institute of Mining, Metallurgical and Petroleum Engineers (AIME) in 1871; American Society of Mechanical Engineers (ASME) in 1880; Institute of Electrical and Electronics Engineers (IEEE) in 1884; and American Institute of Chemical Engineers (AIChE) in 1908. The AIME has undergone several organizational changes since its founding and is now a federation of four societies, with the Society for Mining, Metallurgy and Exploration (SME) as the continuing society.

The data about these societies have been excerpted from the *International Directory of Engineering Societies and Related Organizations 1993*, published by the American Association of Engineering Societies, and are used with permission.

- **American Society of Civil Engineers—ASCE**
  345 East 47th Street, New York, NY 10017,
  U.S.A.

MEMBERSHIP:

Individual	56 736
Student	14 177
Honorary	26 574

OBJECTIVE: The advancement of the science and profession of engineering to advance the welfare of mankind.

PUBLICATIONS: Publications catalog is available on request. The following is a sampling of periodicals published by the Society: *ASCE News*, monthly; *Civil Engineering*, monthly; *ASCE Publications Information*, bimonthly; journal *Hydraulic Engineering*, monthly; journal *Environmental Engineering*, bimonthly; journal *Structural Engineering*, monthly; journal *Engineering Mechanics*, monthly; journal *Irrigation and Drainage Engineering*, quarterly; journal *Geotechnical Engineering*, quarterly; journal *Surveying Engineering*, quarterly; *Transactions*, annual; journal *Aerospace Engineering*, quarterly; *Combined Index*, annual; *Cumulative Index*, quinquennial. Approximately 70 books are published annually, in addition to periodicals.

- **American Institute of Mining, Metallurgical and Petroleum Engineers, Inc.—AIME**
  345 East 47th Street, New York, NY 10017,
  U.S.A.

MEMBERSHIP:

Societies	4

AIME is a separate corporation controlled jointly and severally by its four Member Societies: Society of Mining Engineers, The Metallurgical Society, Iron and Steel Society, and Society of Petroleum Engineers. Individual nonvoting membership consists of AIME Honorary Members and Legion of Honor Fifty-Year Members.

OBJECTIVES: The advancement of knowledge of the arts and sciences involved in the production and use of useful minerals, metals, energy sources, and materials and to record and disseminate significant developments in these areas of technology for the benefit of mankind.

- **Society for Mining, Metallurgy and Exploration, Inc.—SME**

MEMBERSHIP:

Individual	18 702
Student	1 095

OBJECTIVES: To furnish a common ground to members of AIME interested in the fields of mining engineering, minerals exploration, geological engineering, rock mechanics, and minerals benefication; to promote the advancement of the profession through the exchange of information, the preparation and presentation of technical papers, their discussion and circulation; to advance education in the fields covered above and to promote high standards among the members of the profession.

PUBLICATIONS: *Mining Engineering*, monthly; *Minerals and Metallurgical Processing*; various proceedings and textbooks.

- **The American Society of Mechanical Engineers—ASME**

345 East 47th Street, New York, NY 10017,
U.S.A.

MEMBERSHIP:

Individual	96 100
Student	20 000
Honorary	100
Emeritus	1 800

OBJECTIVES: To promote the arts and sciences connected with engineering and mechanical construction for scientific purposes, and to that end to meet and associate together to read and discuss professional papers, and to circulate by means of publication among its members the information thus obtained, and for the purpose of maintaining a library. Also, in 1971, the following Goal was adopted: to move vigorously from what is now a Society with essentially technical concerns to a Society that, while serving the technical interests of its members ever better, is increasingly professional in its outlook, sensitive to the engineer's responsibility to the public interest, and dedicated to a leadership role in making technology a true servant of man.

PUBLICATIONS: *Mechanical Engineering*, monthly; *ASME News*, monthly; *Transactions Journals* (18), quarterly; *Applied Mechanics Reviews*, monthly; Technical Preprints, approximately 1200/yr; Symposia and Proceedings, approximately 165/yr; *Codes and Standards*, approximately 270/yr.

- **The Institute of Electrical and Electronics Engineers—IEEE**

345 East 47th Street, New York, NY 10017,
U.S.A.

MEMBERSHIP:

Individual	270 800
Student	49 500

OBJECTIVES: Its purposes are: (1) Scientific and educational, directed toward the advancement of the theory and practice of electrical engineering, electronics, radio, and the allied branches of engineering and the related arts and sciences. Means to these ends include, but are not limited to, the holding of meetings for the reading and discussion of professional papers and the publication and circulation of works of literature, science, and art pertaining thereto. (2) Professional, directed toward the advancement of the standing of the members of the professions it serves. Means to this end include, but are not limited to, the conduct and publication of surveys and reports on matters of professional concern to the members of such professions, collaboration with public bodies and with other societies for the benefit of the engineering profession as a whole, and the establishment of standards of qualification and ethical conduct. The IEEE shall not engage in collective bargaining on such matters as salaries, wages, benefits, and working conditions customarily dealt with by labor unions.

PUBLICATIONS: Sixty-seven periodicals, including *IEEE Proceedings, Spectrum, Transactions & Journals*, plus more than 50 000 pages annually in nonperiodicals (conference records, reports, standards, etc.).

MEMBER SOCIETIES: Aerospace and Electronic Systems Society (AES), Antennas and Propagation Society (AP), Broadcast Technology Society (BT), Circuits and Systems Society (CAS), Communications Society (COM), Components Hybrids and Manufacturing Technology Society (CHMT), Computer Society (C), Consumer Electronics Society (CE), Control Systems Society (CS), Dielectrics and Electrical Insulation Society (DEI), Education Society (Ed), Electromagnetic Compatibility Society (EMC), Electron Devices Society (ED), Engineering Management Society (EM), Engineering in Medicine and Biology (EMB), Geoscience and Remote Sensing (GRS), Industrial Electronics Society (EI), Industry Applications Society (IA), Information Theory (IT), Instrumentation and Measurement Society (IM), Lasers and Electro-Optics Society (LEO), Magnetics Society (Mag), Microwave Theory and Techniques Society (MTT), Nuclear and Plasma Sciences Society (NPS), Oceanic Engineering Society (OE), Power Electronics Society (PEL), Power Engineering Society (PE), Professional Communication Society (PC), Reliability Society (R), Robotics and Automation Society (RA), Signal Processing Society (SP), Social Implications of Technology Society (SIT), Systems Man and Cybernetics Society (SMC), Ultrasonics Ferroelectrics and Frequency Control Society (UFFC), Vehicular Technology Society (VT), Neural Networks Council (NN), and Solid-State Circuits Council (SSC).

- **American Institute of Chemical Engineers—AIChE**

345 East 47th Street, New York, NY 10017, U.S.A.

MEMBERSHIP:

Individual	40 800
Student	6 500
Honorary	800
Emeritus	3 900

OBJECTIVES: The objectives of the Institute are the advancement of chemical engineering in theory and in practice, the maintenance of a high professional standard among its members, and to serve society, particularly where chemical engineering can contribute to the public interest.

PUBLICATIONS: *Chemical Engineering Progress*, monthly; *AIChExtra*, monthly; *Chapter One*, 3/yr; *AIChE Journal*, bimonthly; *International Chemical Engineering*, quarterly; *Energy Progress*, quarterly; *Environmental Progress*, quarterly; *Plant/Operations Progress*, quarterly; *Biotechnology Progress*, bimonthly; *Symposium Series*, bimonthly; *The Directory of Chemical Engineering Consultants*, annual; *Chemical Engineering Faculties*, annual. Also, process safety procedures, education series, symposium proceedings.

APPENDIX B. COMPARISON OF TWO LEADING U.S. COMPUTER SOCIETIES (FROM *INTERNATIONAL DIRECTORY OF ENGINEERING SOCIETIES, 1993*).

- **Association for Computing Machinery Inc.—ACM**

1515 Broadway, 17th Floor, New York, NY 10036, U.S.A.

MEMBERSHIP:

Individual	65 000
Student	16 000
Corporate	800

OBJECTIVES: To advance the sciences and arts of information processing including, but not restricted to, the study, design, development, construction, and application of modern machinery, computing techniques and appropriate languages for general information processing, for scientific computation, for the recognition, storage retrieval, and processing of data of all kinds, and for the automatic control and simulation of processes; to promote the free interchange of information about the sciences and arts of information processing both among specialists and among the public in the best scientific and professional tradition; to develop and maintain the integrity and competence of individuals engaged in the practices of the sciences and arts of information processing.

PUBLICATIONS: *Communications of the ACM (CACM)*, monthly; *Computing Reviews*, monthly; *Journal of the ACM (JACM)*, quarterly; *Computing Surveys*, quarterly; *CALGO*, quarterly; *Transactions on Database Systems*, quarterly; *Transactions on Mathematical Software*, quarterly; *Transactions on Programming Language and Systems (TOPLAS)*, quarterly; *Transactions on Office Information Systems (TOOIS)*, quarterly; *Transactions on Computer Systems (TOCS)*, quarterly; *Transactions on Graphics (TOG)*, quarterly; *Transactions on Software Engineering and Methodology (TOSEM)*, quarterly; *ACM Guide to Computing Literature*, annual.

- **Computer Society of the IEEE**

IEEE Computer Society, 1730 Massachusetts Avenue, Washington, DC 20036, U.S.A.

MEMBERSHIP:

Individual	69 756
Student	9 995

OBJECTIVES: To advance the theory, practice, and application of computer and information-processing science and technology and to maintain a high professional standing among its members. The society shall promote cooperation and exchange of technical information among its members and to this end shall hold meetings for the presentation and discussion of technical papers, shall publish technical journals, and shall through its organization and other appropriate means provide for the needs of its members. The scope of the Society shall encompass all aspects of theory, design, practice, and application relating to computer and information-processing science and technology.

PUBLICATIONS: *Annals of the History of Computing*, quarterly; *Computer Magazine*, monthly; *Transactions on Knowledge and Data Engineering*, bimonthly; *Transactions on Parallel and Distributed Systems*, bimonthly; *Transactions on Pattern Analysis and Machine Intelligence*, monthly; *Transactions on Software Engineering*, monthly; *Computer Graphics and Applications Magazine*, bimonthly; *Design and Test of Computers Magazine*, quarterly; *Expert Magazine*, bimonthly; *Micro Magazine*, bimonthly; *Software Magazine*, bimonthly; *Transactions on Computers*, monthly.

Works Cited

Accreditation Board for Engineering and Technology (1993), *ABET Accreditation Yearbook*, New York: ABET.

American Institute of Physics (1990), *Society Membership Profile: Employment Mobility and Career Change*, New York: American Institute of Physics.

American Institute of Physics (1993), *AIP Handbook for Member Society Officers, 1993–1994*, New York: American Institute of Physics.

American Society for Engineering Education (1955), "Report on Evaluation of Engineering Education (1952–1955)," *J. Eng. Educ.* **46**, 25–60.

American Society for Engineering Education (1968), "Final Report: Goals of Engineering Education," *J. Eng. Educ.* **58**, 367–446.

American Society for Engineering Education (1987), *A National Action Agenda for Engineering Education*, Washington, DC: ASEE.

Anon. (Ed.) (1994), *World of Learning*, London: Europa Publications.

Barton, H. A. (1956), *Phys. Today* **9** (1), 93–97.

Bates, Ralph S. (1965), *Scientific Societies in the United States*, Cambridge, MA: MIT Press, p. 1.

Bloom, J. L. (1990), *Japan as a Scientific and Technical Superpower*, NTIS PB 90-234923, Springfield, VA: National Technical Information Service.

Breitenberg, M. (Ed.) (1989), *Directory of International and Regional Organizations Conducting Standards-Writing Activities*, NIST Special Publication No. 767, Washington, DC: U.S. G.P.O.

Burek, Deborah M. (Ed.) (1993), *Encyclopedia of Associations: National Organizations of the U.S., 1993*, Detroit: Gale Research.

Edson, L. (1989), *The Making of the NAE, the First 25 Years*, Washington, DC: National Academy of Engineering.

European Physical Society (1992), *Europhys. News* **23**, 121 (Directory issue).

Evan, William M. (Ed.) (1981), *Knowledge and Power in a Global Society*, Beverly Hills, CA: Sage Publications.

Gibbs, N. E. (1991), "Software Engineering and Computer Science: the Impending Split," in: *Education and Computing*, Amsterdam: North-Holland, pp. 111–117.

Institute of Physics Secretariat (1993), *IOP Diary, Membership Brochure*, London: Institute of Physics.

International Union of Pure and Applied Physics Secretariat (1992), *IUPAP 1922–1992, Anniversary Brochure*, Geneva: IUPAP.

Irvin, Linda (Ed.) (1993), *Encyclopedia of Associations: International Organizations, 1993*, Detroit: Gale Research.

McMahon, A. M. (1984), *The Making of a Profession: A Century of Electrical Engineering in America*, New York: IEEE Press.

National Research Council (1987), *Directions in Engineering Research: An Assessment of Opportunities and Needs*, Washington, DC: National Academy of Sciences.

Phillips, Melba (1990), *Am. J. Phys.* **58**, 219–227.

Phillips, Melba, Wilson, Jack M. (1989), *AAPT Announcer Suppl.*, Sept.

Physics Survey Committee (1972), *Physics in Perspective*, NAS–NRC Report, Washington, DC: National Academy of Sciences.

Reynolds, T. S. (1983), *75 Years of Progress—A History of the American Institute of Chemical Engineers, 1908–1983*, New York: AIChE.

Reynolds, T. S., Seely, B. E. (1993), "Striving for Balance: A Hundred Years of the American Society for Engineering Education," *J. Eng. Educ.* **82**, 136–151.

Robbins, P. H. (1984), *Building for Professional Growth*, Washington, DC: National Society of Professional Engineers.

Ryder, J. D., Fink, D. G. (1984), *Engineers and Electrons, a Century of Electrical Progress*, New York: IEEE Press.

Tell, Norman, Chadwick, John (Eds.) (1976), *Professional Organizations in the Commonwealth*, London: Hutchinson.

Toth, R. B. (Ed.) (1991), *Standards Activities of Organizations in the United States*, NIST Special Publication No. 806, Washington, DC: U.S. G.P.O.

Wang Yulong (Ed.) (1990), *Directory of Chinese Learned Societies*, Berlin: World Publishing.

Wisely, W. H. (1974), *The American Civil Engineer, 1852–1974*, New York: American Society of Civil Engineers.

Wolfle, Dael (1989), *Renewing a Scientific Society*, Washington, DC: American Association for the Advancement of Science.

Further Reading

There is not a well-defined body of literature on the various types of engineering organizations discussed in this article. A description of the historical context in which the engineering societies emerged is given by Haber, Samuel (1991), *The Quest for Authority and Honor in the American Professions, 1750–1900*, Chicago: Univ. of Chicago Press. He devotes a chapter to "The Engineers: A Profession in a Business Bureaucracy."

One of the series of reports issued by the Carnegie Commission on Science, Technology, and Government (1993), *Facing Toward Governments, Nongovernmental Organizations and Scientific and Technical Advice*, Washington, DC: Carnegie Commission, views the activities of the many types of nongovernmental organizations (NGOs) as policy advisers to government. The engineering organizations discussed in this article fit under the broad NGO rubric.

Most of the engineering professional societies, typically on a special anniversary of their founding, have commissioned the writing of a history of the organization. Many of these are long on organizational aspects and short on technical development of their field.

There are a large number of directories that provide detailed information about organizations. Some contain lists of engineering-related organizations. There are two such directories that give detailed information about the engineering-related organizations discussed in this article. One is the *International Directory of Engineering Societies and Related Organizations 1993*, 14th ed., Washington, DC: American Association of Engineering Societies. This edition includes updated data about 950 organizations in 70 countries. The second is a more general directory of some 7000 entries including trade associations and labor unions as well as professional and scientific societies: *National Trade and Professional Associations of the United States*, Washington, DC: Columbia Books, published annually. By scanning such directories, a reader can gain a feeling for the number, diversity, and scope of currently existing organizations related to engineering.

PHYSICS AND TECHNOLOGY OF ION AND ELECTRON SOURCES

VADIM G. DUDNIKOV, *Budker Institute of Nuclear Physics, Novosibirsk, Russia*
JOHN H. WHEALTON, *Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A.*

Introduction 1
1. **Physics of Electron Emission** 2
2. **Technology of Efficient Electron Sources** 7
3. **Physical Principles of Ion Production** 13
3.1 Positive-Ion Production 13
3.2 Negative-Ion Production 17
4. **Actual Ion Sources** 19
4.1 Positive-Ion Sources 19
4.2 Negative-Ion Sources 26
5. **Conclusion** 30
Acknowledgments 31
Glossary 31
Works Cited 32
Further Reading 32

INTRODUCTION

Electron (ion) sources are most commonly used to obtain electron (ion) beams of limited transverse size, which are often necessary for focusing into very narrow beams ($\approx 10^{-6}$ m wide, and in the near future 10^{-10} m wide), but wide electron or ion beams (1 to 3 m wide) are also often used. Transport of beams as far as 10^2 m is rather common. In circular accelerators and storage rings, they cover distances of about 10^{10} to 10^{12} m, but the most widely used beams are those no more than 0.5 m long, as used for electronic displays, indicators, microwave generators, and electronics. Before the discovery of the electron, gas-discharge tubes with low gas density and high-voltage discharge were used as the first sources of electrons (cathode rays), as demonstrated by E. Goldstein in 1876. Electrons were emitted by a cold flat (or concave) cathode as it was bombarded with gas-discharge plasma ions (ion–electron emission); then the electrons were accelerated by cathode surface fields and moved along the tube to strike the opposite end, causing luminescence of the glass. Through an aperture in the opposite end, the electron beam was let out of the gas-discharge tube into another vacuum reservoir for further use. In 1883, T. A. Edison observed the flow of electric current in a vacuum between a closely placed negative electrode (cathode) and anode. Since that time, the solid sources of thermal-electron emission have been extensively improved, and they have become widespread. However, as late as 1927, in studying electron diffraction, G. P. Thomson used a gas-discharge electron source because narrow beams were necessary for that purpose, while C. Davisson used a burning hot tungsten emitter in his work. Both were awarded the Nobel prize in 1937. Gas-discharge electron sources were used for high-voltage installations such as x-ray generators. In recent years, new kinds of plasma sources have been developed for accelerators, especially for pulsed high-current accelerators. Thermionic cathodes became mass produced as components of electron tubes and, later, in a great number of other products of emission electronics. After tubes were replaced by semiconductor devices, the electron-beam devices with thermionic cathodes became most widespread.

Electron and ion fluxes are important components of a wide and growing variety of applications in many areas of fundamental scientific research, in sophisticated control systems, in large-scale industrial manufacturing, and in everyday life. Indeed, much of modern high technology is closely related to sophisticated techniques that enable us to manipulate atoms and their components—electrons and ions. The sources of electrons

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

and ions are the real sources of technology power. The starting point for the development of this basis of modern technology was J. J. Thomson's discovery of the electron in 1897, which initiated regular work with electrons, ions, and atoms. The first Nobel prize winner was K. Röntgen, who discovered x rays generated by electrons in a high-voltage discharge (together with ion generation). Most of the early scientific works that led to Nobel prizes included the generation of electrons and ions as an important component. In modern experiments, the efficiency of the generation of polarized Z bosons (elementary particles 97 times as massive as protons) in the Stanford electron–positron Linear Collider beams is determined by the operation efficiency of a small photoemitting source of polarized electrons and of multimegawatt electron tubes—klystrons.

Practical application of electron sources and their development was initiated by the manufacture of x-ray tubes and electron tubes for radio communication. The physics and technology of emission electronics was developed for this application. Electron accelerators, a variety of microscopes, television tubes, displays, oscilloscopes, radio-frequency and microwave generators, gas-discharge tubes and lasers, focused electron beams, and electron-projection microlithography are the basis of modern microtechnology; these represent only a small part of the modern applications of electron sources. The development of field-emission sources producing electron and ion beams of high brightness has made it possible to carry out control and treatment on nanometer scales. The discovery of precisely controlled tunneling emitters has initiated the development of the technique that enables controlled manipulations with single atoms. Electron beams with high brightness and superhigh-frequency modulation have been obtained with the help of efficient photocathodes irradiated by lasers.

The development of ion-beam production was stimulated by the needs of mass spectrometry and accelerators for nuclear physics. The first essential application—electromagnetic separation of isotopes—stimulated the development of the physics and technology of highly intense ion sources. These developments and further improvements provided the ion sources that are widely used for ion implantation in the semiconductor industry. Many variants of ion sources have been optimized for different applications. Progress in plasma physics helped to transform the development of ion sources from a fine art to an applied science; however, in obtaining optimal parameters, the art of an experimenter still plays an important role. Ion sources with up to 100-A beam intensity have been developed for controlled nuclear fusion. New ways of using a negative ion instead of a positive one were revealed by the discovery of significant enhancement of negative-ion generation in gas discharges due to the addition of small admixtures of catalysts, such as cesium or other substances with low ionization potential. The stable generation of ion fluxes with an emission current density of 10^8 A/cm^2 was achieved because of the development of electrohydrodynamic emitters (made of liquid metals and dielectrics). The great increase in source brightness has initiated new applications of beams, such as colliding-beam experiments, long-distance transportation of energy, focused-beam and ion-beam projection lithography, and implantography with nanometer resolution.

The development and use of new materials, such as high-temperature alloys, ceramics and composite materials, strong permanent magnets, new microwave generators, power supplies, and control systems have been very important in the improvement of the source characteristics.

Particle sources, as far as size and cost are concerned, are only a small part of the device in which they are used; but some of the most important characteristics of the devices are often determined by the properties of the sources, and the improvement of these characteristics is an essential physical and technological problem. Recently, electron and stochastic cooling have been developed to increase beam brightness, but obtaining high-quality beams right from the sources appears to be a very attractive method.

1. PHYSICS OF ELECTRON EMISSION

Negative electrons and a positive nucleus are bound in neutral atoms by electromagnetic forces. An electron with charge $e = 1.6 \times 10^{-19}$ C in a potential difference $U = 1$ V attains the energy $eU = 1$ eV $= 1.6 \times 10^{-19}$

J. At the atomic-scale distance from the proton $r \sim 10^{-10}$ m, the electric field is $E = -e/4\pi\epsilon_0 r^2 \sim -10^{10}$ V/m, and the binding energy of the electron (ionization potential I) has a value of several electron volts (from $I = 3.89$ eV for cesium to $I = 24$ eV for helium). Such energy must be transmitted to the electron to make it leave the atom.

Bound electrons in atoms and molecules can occupy allowed discrete levels of energy, while free particles can have a continuous spectrum of positive energies. An electron in a hydrogen atom can occupy states with energies $E = -2\pi me/h^2n^2 = -I/n^2$, where $m = 0.9 \times 10^{-30}$ kg is the mass of the electron, $h = 6.6 \times 10^{-34}$ J is the Planck constant, $I = 13.6$ eV, $n = 1$ is the ground state, $n = 2, 3, \ldots$ are the excited states. In such complex systems as molecules, allowed electron states transform into bands of levels with less energy difference, and the bands of allowed energies in a solid (as in a very big molecule) have a pseudocontinuous distribution of electron states. The potentials and electron states in different systems are shown schematically in Fig. 1.

When electrons are drawn apart from atoms or molecules, there appear free electrons and atomic (A^+, A^{++}, . . .) or molecular (M^+, M^{++}, . . .) ions with different charges. More complex cluster ions, consisting of many weakly bonded atoms and molecules, can be formed. Electrons can be attached to any atom, molecule, or cluster, with negative ions A^-, A^{--}, . . ., M^-, M^{--}, . . . being formed if the system has any electron affinity, S (binding energy).

Usually systems have allowed states with very low ionization energy, but the probability of being in these states is low. The energy quantum that is necessary for liberation of electrons is several electron volts. This energy can be transmitted to an electron in a system in collision with other particles such as an electron, an ion, or a neutral particle, or by a photon of light, or by heating complex systems.

The emissive properties of a solid are determined by an electron's behavior, and simple models (as in the Sommerfeld approximation) can be used for better understanding (Blakemore, 1985). The interaction of particles in a solid creates a boxlike potential well of average depth $-W_0$ for electrons. The form of the potential barrier at the solid–vacuum boundary, which is important for emissive properties, can be approximated by the "image force potential" $U(x) = -e^2/16\pi\epsilon_0 z$. It is shown in Fig. 1(c).

In the periodic potential of a crystalline lattice, the wave functions and movement of electrons are much like those of free electrons (Bloch theorem). The interaction with the lattice changes the effective mass m^* of electrons and divides the phase space of motion into bands of allowed energies and forbidden gaps (Brillouin zones of the Fermi surface). The phase-space distribution of electrons is determined by Fermi–Dirac statistics. The density of states in the allowed energy band is proportional to the phase volume $d(p^3/h^3)$, and for free particles it is proportional to the square root of the distance W from a minimum or maximum energy of

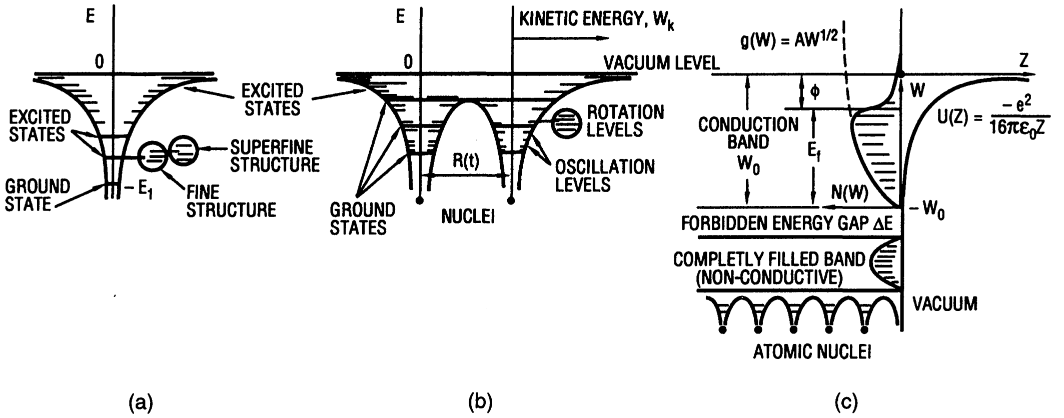

FIG. 1. The potentials and electron states in different systems.

the allowed band:

$$g(W) = A'W^{1/2}. \tag{1}$$

The probability of these states being occupied at temperature T is

$$f(W,T) = \{\exp[(W - E_F)/kT] + 1\}^{-1}, \tag{2}$$

where E_F is the Fermi energy (the thermodynamic electrochemical potential) determined by the electron density and equal to several electron volts for metals, k is the Boltzmann constant, $f(W,0) = 1$ when $W \leq E_F$, and $f(W,0) = 0$ when $W > E_F$. It means that no more than one electron can be in any of the allowed states. In a completely filled band, electrons cannot change the states of motion and cannot create conductivity. Only excitation to the allowed states of the next upper band can create electricity carriers, but it requires a high energy ΔE (the forbidden gap). If the highest energy band is empty and ΔE is high, we have an insulator (thermal excitation cannot create conductivity). If ΔE is low, we have a semiconductor, and rapid exponential increase of conductivity as temperature rises. If the highest energy band is not completely filled with electrons, as in Fig. 1(c), we have a metal.

The Fermi energy E_F is less than the potential-box depth W_0. The difference $\phi = W_0 - E_F$ is the work function (the evaporation energy), which is the analog of the ionization potential for atoms. This energy must be transmitted to the electron for its liberation from the Fermi level, where the electron density in phase space is high. For $T > 0$ the electron distribution is

$$\begin{aligned} N(W,T) &= g(W)f(W,T) \\ &= A'W^{1/2}/\{\exp[(W - E_F)/kT] + 1\}, \end{aligned} \tag{3}$$

and some electrons can occupy the levels with kinetic energy $W > E_F + \phi = W_0$, as is shown in Fig. 1(c). These electrons are able to escape from the solid (to evaporate) if their velocity v_z (normal to the surface) is high enough (normal kinetic energy $mv_z^2/2 > W_0$). The emission current density j_e is defined by the density of electron flux j_w (flux per unit area of solid–vacuum boundary) and the probability of penetrating the barrier, $D = 1 - R$, where R is the reflection coefficient:

$$j_w = e\int_0^\infty N(W,T)v_z dv_z; \tag{4}$$

$$j_e = e\int_0^\infty D(W)N(W,T)v_z dv_z. \tag{5}$$

For thermionic emission of electrons when the electric field applied to the surface of the emitter is weak but intense enough to extract emitted electrons onto the anode collector, $D = 0$ when $S < W_0$, and $D(W) > 0$ when $W > W_0$. In cylindrical coordinates $v_\perp$, v_z,

$$\begin{aligned} j_e &= eA' \int_{v_z>(2W_0/m)^{1/2}} \\ &\times \frac{D(v_z)\, 2\pi v_\perp dv_\perp v_z dv_z}{\exp[mv_\perp^2/2kT]\exp[(mv_z^2/2 - E_F)/kT] + 1} \\ &= BT^2 \int_{\phi/kT}^\infty dz \int_0^\infty \frac{dy}{e^y e^z + 1} \\ &= BT^2 \int_{\phi/kT}^\infty \ln(1 + e^{-z})\, dz \\ &\approx A'D_{ef}T^2 e^{-\phi/kT}. \end{aligned} \tag{6}$$

Equation (6), the Richardson equation, provides a proper description of thermionic emission of electrons. The constant $A' \sim 10^2$ A/cm² K² for metals. For pure surfaces with a smooth potential barrier, the effective transmission coefficient $D_{ef} \sim 1$; however, it falls to $D_{ef} \ll 1$ because of the absorption of impurity films on the surface of the emitter, which creates surface barriers with potential fluctuations. The work function ϕ is the most essential characteristic of an electron emitter. Higher emission current density is obtainable at lower temperature by the lowering of ϕ. Refractory metals such as tungsten or tantalum have ϕ from 4.5 to 4 eV, while the work function of some specially effective emitters may approach 1 eV. For the latter, $e^{-\phi/kT} \approx 0.01$–$0.1$ at $T \approx 10^3$ °C and the emission current density reaches values of 1–10 A/cm². For real emitters, the values of ϕ and D_{ef} vary greatly from one point to another, and therefore Eq. (6) should involve spatially averaged values.

The space charge of the emitted electrons induces an electric field that limits the emission current (Forrester, 1988). The density of the space-charge–limited current in a vacuum diode is given by the Child–Langmuir law:

$$j_{\mathrm{CL}} = (4\epsilon_0/9)(2e/m)^{1/2}U^{3/2}/d^2, \tag{7}$$

where d is the cathode–anode separation and U is the voltage difference between the flat cathode and the anode. The change of the current in the diode is dictated by this law if the emission current density in Eq. (6) exceeds j_{CL} of Eq. (7).

An increase of U causes the extraction of electrons onto the anode, and the limitation of emission by space charge is eliminated. This leads to saturation of the emission and the creation of fields on the surface of the cathode, intensifying the emission of electrons (Blakemore, 1985; Forrester, 1988).

If a strong electric field E is set up, the emission of electrons is intensified, and the surface-potential profile changes as is shown in Fig. 2. The superposition of the image potential $U(z) = -e^2/16\pi\epsilon_0 z$ and the uniform field potential $U_1(z) = -eEz$ has its maximum at $z = z_0 = (e/16\pi\epsilon_0 E)^{1/2}$. The surface-barrier height is lowered by an amount

$$\begin{aligned}\Delta\phi &= (e^3E/4\pi\epsilon_0)^{1/2} \\ &= (3.8 \times 10^{-5}\ \mathrm{eV\ m^{1/2}\ V^{-1/2}})E^{1/2}.\end{aligned} \tag{8}$$

This reduction of the work function results in an increase in emission that is referred to as the Schottky effect. The value of the emission current is defined by Eq. (6) when ϕ is replaced by $\phi - \Delta\phi$. This effect is more essential for low–work-function cathodes, functioning at low temperatures, and pointed cathodes with a very strong field.

As the electric field increases further (up to $E \approx 10^9$ V/m), another kind of emission, field emission (autoelectronic, tunnel, cold emission), rapidly increases (Blakemore, 1985). R. Wood was the first to observe field emission of electrons, in 1897. In 1992 Lilienfeld used the emission in a strong field inside an x-ray tube. Milliken and Lauritsen obtained the voltage–current characteristics of field emission for a wide range of parameters and confirmed the weak dependence of emission on temperature ($\ln j = A' - B/E$). In 1929, P. Fowler and L. Nordheim gave the theoretical explanation of field emission, which was based on the tunneling effect.

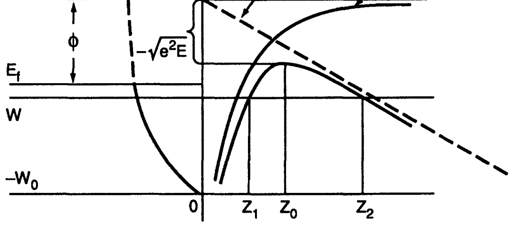

FIG. 2. The surface-potential profile including an external electric field **E**.

When field emission takes place, electrons penetrate the potential barrier at the boundary of the emitter by tunneling through it rather than passing over it as in the case of thermionic emission or photoemission. A strong electric field at the surface (Fig. 2) creates a potential well for electrons, into which they can penetrate from the occupied (filled) band in the emitter. It also makes the potential barrier lower and narrower and the probability of tunneling sufficiently higher.

To calculate the current density in the general equation [Eq. (5)], it is necessary to use the quantum-mechanical value of the transmission coefficient $D(W)$ at the potential of $U(z)$ when the energy level of the particle intersects the barrier line at the points z_1 and z_2 (Fig. 2):

$$\begin{aligned}D(W) &= \exp\left\{-\int_{z1}^{z2} [(8m/h^2)(U(z) - W)]^{1/2}dz\right\} \\ &= \exp\left\{-\int_{z1}^{z2} [(8m/h^2)(-eEz + |W| \right. \\ &\qquad \left. - e^2/16\pi\epsilon_0 z)]^{1/2}dz\right\},\end{aligned} \tag{9}$$

where

$$z_1 = |W|/2eE[1 - (1 - e^3E/W^2)^{1/2}],$$

$$z_2 = |W|/2eE[1 + (1 - e^3E/W^2)^{1/2}].$$

Integration similar to that performed in Eq. (6) gives the Fowler–Nordheim formula for metals:

$$j_{fc} = C_1E^2 \exp(-C_2/E), \tag{10}$$

where $C_1 = e^3/8\pi h t^2(y)\phi$, $C_2 = 8\pi(2m)^{1/2} \times \phi^{3/2}\vartheta(y)/3he$; t and ϑ are tabulated functions of $y = e^{3/2}\epsilon^{1/2}/\phi$; $t(y) \approx 1$, and $\vartheta(y) \approx 1 - y^2$. A good approximation to the exact formula is provided by the expression

$$j_{fe} \approx (1.4 \times 10^{-6}E^2/\phi) \times 10^a,$$

where

$$a = 4.4\phi^{-1/2} - 2.8 \times 10^7\phi^{3/2}/E, \tag{11}$$

where j is in units of A/cm^2, E in V/cm, and ϕ in eV. The Fowler–Nordheim formula agrees well with experiment at densities up to 10^{10} A/cm^2, as is illustrated in Fig. 3. Field emission produces a narrow energetic spectrum of electrons. The half-width ΔW of the energy distribution at $T = 0$ K is determined by the equation

$$\Delta W = 6.7 \times 10^{-9}E/\phi^{1/2}t(y). \tag{12}$$

When $\phi = 4.4$ eV, $j = 1 \times 10^7$ A/cm^2, and ΔW varies from 0.08 to 0.2 eV, while at $T = 300$ K, $\Delta W = 0.17$–0.3 eV. The adsorption of impurity films complicates the emission and distorts the energy spectra.

At low temperatures, field emission causes the emitter to heat because the energy of the electrons emitted from the occupied part of the band is $<E_F$, and the energy of the incoming conduction electrons is $>E_F$ (Nottinheim effect). At higher temperatures, the emission of hard electrons increases, and cooling of the emitter prevails.

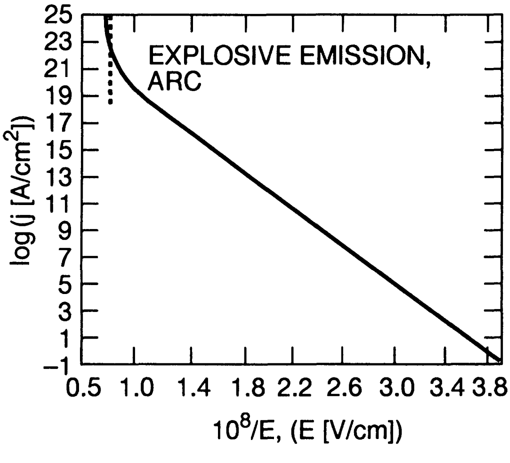

FIG. 3. Graph showing agreement of the Fowler–Nordheim formula with experiment at densities up to 10^{10} A/cm^2.

Field emission from semiconductors possesses some characteristic features (Blakemore, 1985):

1. An electric field penetrates deep into a semiconductor, which results in the displacement of the energetic bands and in the concentration of charge carriers and their spectrum.
2. When the density of electrons is much lower, j_{fe} becomes limited and greatly sensitive to the influences of light, temperature, etc.
3. Field emission depends on surface states.
4. Voltage–current characteristics and energy spectra are determined by the band structure of semiconductors.
5. The current in the emitter substantially redistributes the potential and changes the spectra of the electrons.

All the previously mentioned facts require many effects to be taken into account, which complicates the calculations and makes the interpretation of the results more difficult, but at the same time the said phenomena provide the basis for many important applications.

A gradual increase in the tunneling current up to the level of 10^{10} A/cm^2, together with a voltage increase, usually gives rise to the impulse of explosive emission (Bystritskiy and Didenko, 1989). The consequent development of vacuum breakdown and arc discharge leads to the destruction of the emitter.

Photoelectric emission is caused by the transmission of energy from photons to electrons. For the photoemission of an electron to occur, it must receive energy from a photon $h\omega > \phi = h\omega_0$ (Fig. 1), where ω_0 is the threshold frequency of light and $\lambda = 2\pi c/\omega_0$ is the threshold wavelength. Some electrons possessing this energy, which is greater than that of electrons above the barrier in vacuum, will move toward the surface of the emitter and may be emitted into the vacuum (or another medium). While moving toward the surface, the high-energy electron transmits energy to the conduction electrons and loses energy with high probability before reaching the surface of the emitter. The basic characteristic of photoelectric emission is the quantum yield Y, representing the number of emitted electrons per photon falling onto the photoemitter surface. The value of

Y depends on the surface state of the emitter, the energy, the angle of incidence, and the polarization of the photon.

For pure surfaces of most metals, $\phi > 3$ eV, and so photoemission from metals is observed when $\lambda < 0.4$ μm. Photoemission from alkali metals and barium can be caused by visible light, and from other metals only by ultraviolet radiation. For most metals, $Y \approx 10^{-4}$ electron per photon near the threshold. This light penetrates into the metal 10^{-7} cm and is absorbed there, but high-energy electrons are not able to escape from this depth because of the intense interaction with the conduction electrons. The electrons that have received sufficient energy from the photons at the depth of 10^{-9} cm from the surface are emitted with high probability. However, the probability of getting sufficient energy is low. The more tangent the light falling on the surface, the higher the probability of reflection of the photons. As the energy of the photons becomes higher, Y increases ($Y_{\mathrm{Bi}} = 4 \times 10^{-2}$, $Y_{\mathrm{Al}} = 1.5 \times 10^{-2}$ when $h\omega = 12$ eV). When $h\omega > 15$ eV, the absorption of photons rapidly increases (the reflection coefficient R falls to 5%) and Y rises to 0.1–0.2. On adsorption of admixtures, ϕ becomes less, and Y becomes greater because of the photoemission threshold being displaced into the visible region of the spectrum. Submonolayer films of alkali and alkaline-earth (electropositive) elements improve the effect of photoemission best of all.

In semiconductors and dielectrics, photoemission is provided by the excitation of electrons so that they get to the upper region of the conduction band (to the part of the band above the barrier) from the occupied valence band or from impurity levels [the donor band at the bottom of the conduction band or the acceptor band at the top of the valence band (Blakemore, 1985)]. In this case, the threshold energy $h\omega = \Delta E + \chi$ includes the width of the forbidden band ΔE, the electron affinity χ, and the height of the potential barrier above the bottom of the conduction band (the analog of the work function ϕ). In emitters with low density of electrons in the conduction band and with $\chi > \Delta E$, the electrons spend their energy on collision ionization of the valence-band levels to form free electrons in the conduction band and vacancies in the valence band. As a result of this process, the depth from which electrons can escape is lowered to 10^{-9} m, while photons can excite electrons at distances of no more than 10^{-6} m from the surface. Only a small number of excited electrons escapes into the vacuum ($Y \approx 10^{-6}$–10^{-4}). In substances with $\chi < \Delta E$, the energy of the photoelectrons is transmitted to the lattice oscillations (phonons) in small portions (10^{-2} eV); the probability Y for the electrons to escape into the vacuum increases and when $\chi > 1$ eV ($Y \approx 10^{-2}$) remains low.

The special treatment of p-type semiconductors using the adsorption of cesium and oxygen layers in an alloy is used in the manufacture of emitters with a negative electron affinity, from which most electrons excited by photons into the conduction band are emitted into the vacuum. In this process Y goes from 0.1 to 0.5 right above the threshold for photoemission. Initiation of photoemission by circularly polarized light enables one to obtain bunches of polarized electrons.

The emission of electrons into a vacuum is also initiated by other mechanisms of transmission of energy in a solid—for example, by electron or ion bombardment or by bombardment with excited (metastable) and accelerated atoms and molecules. We can observe a higher yield of electrons due to a greater concentration of energy release at the surface of the emitter.

2. TECHNOLOGY OF EFFICIENT ELECTRON SOURCES

Electron sources are most commonly used to obtain electron beams of limited transverse size. Gas-discharge electron sources were used for high-voltage installations such as x-ray generators. In recent years, new kinds of plasma sources have been developed for accelerators, especially for pulsed high-current accelerators. Thermionic cathodes became mass produced as the components of electron tubes and, later, in a great number of other products of emission electronics. After tubes were replaced by semiconductor devices, the electron-beam devices with thermionic cathodes became most widespread.

The simplest electron source is a tungsten filament heated to a high temperature T by

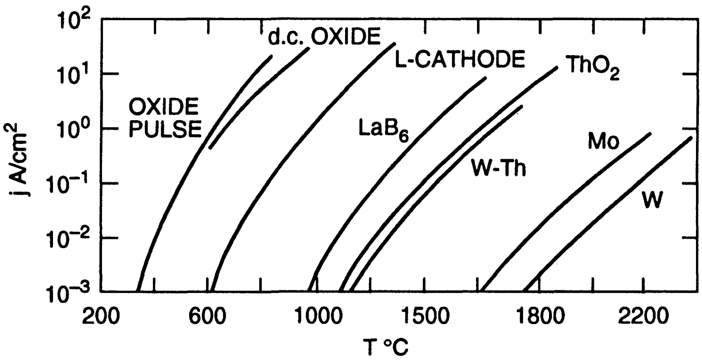

FIG. 4. The comparative characteristics of various thermoelectric emitters.

applying a certain voltage to it. At $T \approx 2600$ K, specific emission may reach $j \approx 0.5$ A/cm^2, the lifetime being as long as 10^4 h under good vacuum conditions (Kuznetsov, 1994; Forrester, 1988). Tantalum possesses similar characteristics at $T \approx 2400$ K. Other pure metals evaporate too quickly at the temperatures that provide proper emission of electrons. Admixtures, especially of rhenium, improve mechanical properties of tungsten and its longevity, and they prevent it from brittle destruction resulting from recrystallization. Directly heated tungsten cathodes in the form of studs are still being widely used for sources of bright electron beams in electron microscopes, and in the form of loops or various spirals in gas-discharge devices and ion sources, among others. Pure tungsten has a work function of $\phi = 4.5$ eV, causing a rather low efficiency of emission. Nevertheless, pure tungsten withstands ion bombardment and poisoning with residual gases and admixtures; it is easy to manufacture and can be used with demountable devices. The comparative characteristics of various thermoelectron emitters are given in Fig. 4.

Optimal lowering of the work function and a substantial increase in thermionic emission are achieved by covering the surfaces of emitters with thin films of electropositive elements (with lower ionization potentials) (Kuznetsov, 1994; Forrester, 1988). The typical dependence of the work function on the atomic density of cesium film on tungsten is shown in Fig. 5. The work function reaches its minimum when the atomic density is less than one monolayer. The film can sustain the emission temperature, and its slow evaporation may be balanced by atoms arriving by means of diffusion from the bulk of the substance or migration through pores. Thus, for example, if optimal conditions hold, the addition of thorium oxide to tungsten makes the work function ϕ as low as 2.63 eV. Although the beam brightness A' goes down to 3 A/cm^2 K^2, the emission current density, $j = 3$ A/cm^2, holds at the temperature of only 2000 K and at a low rate of thorium evaporation. At $T = 1800$ to 1900 K and the specific emission $j \approx 0.3$–0.8 A/cm^2, the efficiency is 35–50 mA/W. Carbonizing (carbon saturation) improves the characteristics of these cathodes, which are used in high-voltage devices. At present, highly tolerant thermionic emitters are being made of alloys of noble metals (platinum, iridium, os-

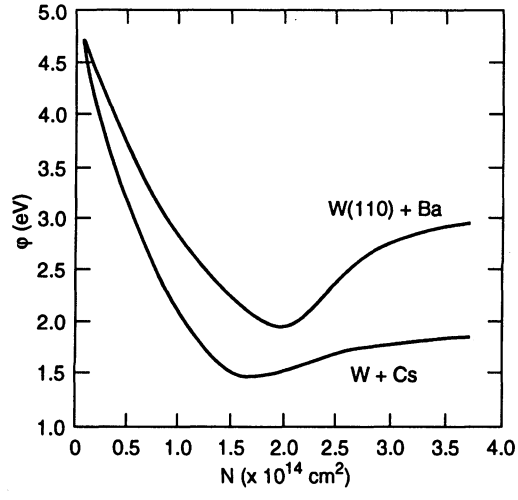

FIG. 5. Typical dependence of the work function on the atomic density of cesium film on tungsten. The work function reaches its minimum when the atomic density is less than one monolayer.

mium, etc.) with rare-earth elements (lanthanum, yttrium, scandium, etc.). Activation by means of optimization of coverings turns these alloys into efficient emitters (Kuznetsov, 1994).

Thermionic emitters with lower temperatures (<1800 K) are manufactured mainly in the form of knots with indirect heating. Tungsten spirals with high-temperature ceramic insulation serve as heaters. The usual form of emitters is a cap at one end with emitting surface at the opposite end (for electron-beam devices) or on the side surface (for electronic tubes, magnetrons, etc.). The heater is inside the cap. Power is transmitted to the emitter by irradiation or through heat conductivity. Screens are put against the nonemitting surfaces to minimize losses from irradiation. Optimal barium film on tungsten lowers the work function to ϕ = 1.66 eV. Under these conditions, $A' \approx 2$ A/cm^2 K^2, j = 3–50 A/cm^2 at T = 1320–1540 K, and the evaporation rate is 4×10^{-10} g/cm^2 s. The evaporation of the active substance is counterbalanced by the barium (oxide, aluminate, etc.) from the reserve designed into the cathode. In L cathodes, barium migrates onto the emitting surface through a porous wafer of baked tungsten powder from a hollow that contains a reserve of the active substance. In the initial state, it contains a mixture of barium and strontium carbonates. When activated in a vacuum, these carbonates decompose to form BaO, SrO, and CO_2, which is pumped out during this process. Then in the process of the reaction $6BaO + W \rightarrow Ba_3WO_6 + 3Ba$, free barium is formed. At $T \approx 950$ °C, j = 1.3 A/cm^2, and the lifetime is up to 8000 h. The L cathode is tolerant of adverse conditions, but evaporation of barium is rather substantial. By the use of the more regular structure of clutched and baked tungsten filaments (bundle cathode) instead of the porous wafer, it is possible to improve the characteristics of emitters (Le Duff, 1990; Kuznetsov, 1994).

The impregnated cathode consists of a porous wafer impregnated with an active substance (barium aluminate). Such a structure improves its manufacturability in comparison with L cathodes. For the former, it is typical to have ϕ = 1.67 eV, A' = 1.25 A/cm^2 K^2, and j = 1.5 A/cm^2 at T = 1020 °C. Pressed cathodes are produced by means of pressing mixed barium carbonates or aluminates together with molybdenum and tungsten powder with addition of restoring agents (WC, ZrH, TiH). At T = 1000 °C, j = 1 A/cm^2 and t = 5000 h, while t = 300 h at T = 1070 °C, and j = 3 A/cm^2. These cathodes are tolerant of ion bombardment and of gas poisoning even in bad vacuum conditions, $\approx 10^{-4}$ Torr. They also serve as good electron sources for gas discharges in hydrogen and noble gases.

Impregnated (dispenser) cathodes have operating times of 10 h at an operating temperature of 1100 °C. Many versions of impregnated cathodes have been developed during the last decades of investigation. The B-type cathode has been the most widely used in electron guns for commercial equipment since 1957. It consists of porous tungsten impregnated with BaO and Al_2O_3. These mixtures react chemically to liberate free barium atoms. CaO is also added to the impregnate. It reduces the barium sublimation rate and enhances emission properties. The names of B-type cathodes reflect the relative chemical compositions of the various compounds. As in high-temperature superconductors, a 532-type cathode contains five parts of BaO to three parts of CaO and two parts of Al_2O_3, and a 411-type cathode is $4BaO + CaO + Al_2O_3$.

The M-type cathode is a B-type cathode with the deposition of a thin film of ruthenium–osmium on the emission surface. The film must be very thin to permit the passage of barium. The possible film thickness is 0.1–1 μm. For a given temperature, the presence of a surface layer has increased the emission current density by a factor of 2. The M cathodes have a longer operation time relative to the B type since they have lower temperature for the same emission property. But damage to the thin film by ion bombardment leads to degradation of its emission characteristics.

A scandate cathode has a ScO admixture to the standard impregnation compound of a B-type cathode. The current density of a scandate cathode is almost identical to that of an M-type cathode. Since the scandate cathode does not rely on a thin surface film, and it is not as susceptible to mechanical damage and poisoning, it has some distinct advantages over the M type. Until now, scandate cathodes have not been used in long-term operating equipment.

Oxide cathodes consisting of thick (30–

200 μm) layers of the oxides BaO, SrO, and CaO on nickel or tungsten backing have become the most widespread. In the initial state, a mixture of the carbonates in suspension with organic compounds as a binder is applied to the base by spraying or by means of cataphoresis.

An oxide cathode is activated by heating at T = 800–1100 °C so that the carbonates decompose to form oxides while CO_2 is pumped out. Maintaining 1000 °C with the current removed leads to a partial restoration of barium and, as a consequence, to the transformation of an insulating mixture of oxides into an electronic semiconductor with donor barium admixture with levels near the bottom of the conduction band. The position of its Fermi level is between the bottom of the conduction band and the donor level. The optimized work function of an oxide cathode is ϕ = 1–1.2 eV, $A' = 10^1$–10^{-3} A/cm^2 K^2. At T = 800 °C, j = 0.5 A/cm^2, and t = 500 h; when j = 5–10 mA/cm^2, $t \approx 10^5$–10^6 h. In the impulse regime, j is 150 A/cm^2 at T = 1100 K. Oxide cathodes are 25 times as efficient as tungsten cathodes. The emission of electrons takes the major part of the power expenditure. To obtain high-brightness beams, it is necessary that emitters have smooth and uniform surfaces. For emitters without strong restrictions implied on their geometry, backings with developed surfaces are used, namely sponges, grids, sets of tubes, and bars.

Hexaborides of rare-earth elements are used as efficient cathodes. LaB_6 with, typically, ϕ = 2.66 eV and A' = 29 A/cm^2 K^2 (melting temperature 2210 °C) is most commonly used. LaB_6 is a polycrystalline substance or monocrystals of high hardness. Warming up and degassing at T = 1500–1600 °C leads to quick degasification and activation. At T = 1680 K, j = 1 A/cm^2, and $t \approx 10^4$ h; the evaporation rate is 10^{-9}–10^{-11} g/cm^2 s. Boron diffuses rapidly from LaB_6 into tungsten and molybdenum, which are in surface contact with it, the diffusion into tantalum being slower. To prevent boron diffusion, $MoSi_2$ is used. It is preferable to use graphite holders for fixing emitters made of LaB_6. Cathodes made of LaB_6 are most commonly used in dismountable systems such as accelerators, welding guns, etc. Pointed emitters in the form of straight-channel "studs" with the pointed end of the filament loop made of monocrystalline LaB_6 are used as thermoemitters in electron microscopes and electron-beam lithography.

More sophisticated configurations of thermionic emitters are in use: hollow cathodes in the form of hollow spheres or cylinders with a higher intensity of emission from the edges of the outlet aperture, and ribbon and circular cathodes.

In thermionic emission converters and other applications, the work function of tungsten emitters (and colder anode collectors) is lowered by the adsorption of cesium. In the presence of oxygen or fluorine, the optimal cesium film on refractory metals makes the work function as low as ϕ = 0.75–1 eV, and $j \approx 10$ A/cm^2 at $T \approx 1200$ K. These systems are used for sustaining the discharges in ion sources and gas lasers (Forrester, 1988; Dudnikov, 1992).

The class of cold cathodes includes autoelectronic (field) emitters, film-coated cathodes with a strong electric field in dielectric layers that are close to the surface, and secondary cathodes where emission is provided by interaction with gas-discharge plasma. To provide stable field emission in autoelectronic emitters, it is necessary to apply a stable electric field of $E \approx 5 \times 10^6$ to 5×10^7 V/cm to its surface and to provide high stability of the form and the work function. The latter is particularly hard to carry out because of the strong influence of the electric field and ion bombardment on the migration processes of surface atoms and on their adsorption–desorption.

At moderate voltages of $U \approx 1$–10 kV, the field strength on the surface of a field-emission cathode reaches its highest values on the points, with small radii of curvature $r \approx 10^{-4}$–10^{-5} cm. Such cathodes are pointed emitters with high-brightness beams. The emission current density $j \approx 10^7$ A/cm^2 dc, and j is up to 10^8 A/cm^2 in the pulse regime. The full current range is $I \approx 10^{-3}$ to 10^3 A. The stability and uniformity of the emission zone are provided by selective adsorption of substances lowering the work function toward the end of the point (barium on tungsten, zirconium on tungsten). Monocrystalline emitters with a certain stable face of the crystal sticking out of the point are commonly used. Ultrahigh vacuum, degassing, and higher temperature are used to reduce the effect of adsorption. Tungsten points function reliably in a vacuum of 10^{-9} Torr and ZrC with LaB_6 at 10^{-7} Torr. Functioning in the regime of thermal field emission is more stable at higher temperature but with a brightness higher than that of thermionic cathodes.

Recently, matrix field emitters have been developed with densely packed microemitters. The gap between the emitter and the anode being small ($d \approx 10^{-4}$ cm), the voltage $U \approx 10^2$ V is enough to cause field emission, and the average current density is $j \approx 1$ A/cm² dc. Because of the specific configuration of the unit (Fig. 6), ion bombardment of the point is excluded.

Although the interaction of electrons in a beam of high density increases the angular and energetic spread, with autoelectronic emitters it is possible to obtain beams of record brightness that are applicable mainly in high-speed electron lithography.

A strong electric field can accelerate electrons in a solid and thus cause some to escape into the vacuum (the emission of "hot" electrons). Favorable conditions for the emission of "hot" electrons may appear in thin dielectric films on conductive backings that have either "conductor–dielectric" or "conductor–dielectric–conductor" structure.

A strong electric field is created either by applying a positive charge to the vacuum surface of the dielectric film (by means of bombardment with slow ions or the removal of electrons from the surface in the process of secondary emission or photoemission) or by applying positive voltage to the very thin conductive film covering the dielectric. It induces field emission of electrons from the conductor into the dielectric film. Free electrons accelerate in the dielectric film and get sufficient energy for avalanche-type multiplication and for escaping into the vacuum. Some electrons penetrate the potential barriers and pass through the layers of the substance into the vacuum. The dissipation of energy being rather low because of the small film thickness, the emission of electrons is not accompanied by a considerable rise in temperature, but is usually accompanied by luminescence resulting from the free-carrier recombination. In producing "cold" sources of "hot" electrons, a metal backing (e.g., nickel with a dead surface) is coated with a thin dielectric film (MgO, 10 μm, often porous), then degassed in vacuum and subjected to heating in oxygen at $t \approx 850$ °C for 10 min. Self-maintained emission is excited by electron, ion, and photon impulse irradiation, the emitted electrons being extracted by 200 to 300 V at the collector. Average current density is 10 mA/cm², the effective temperature of the "hot" electrons is $T \approx 10$ eV, and the lifetime is 10 h.

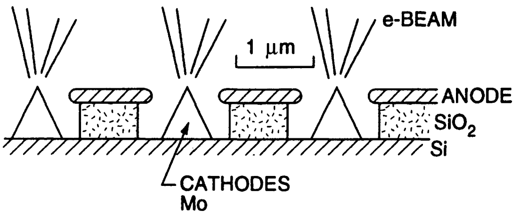

FIG. 6. Multipoint autoelectronic field emitters for high-speed electronic lithography applications.

After the deposition of a very thin conductive film (Au, ≈100 °A, islandlike films), transparent for the "hot" electrons, over the dielectric film, the controlled emission is initiated by applying 100–200 V to this film relative to the backing. In the impulse regime, higher voltage excites emission of high current density without inertia.

The emission of electrons from the surface of ferroelectrics is excited in a similar way. In a polarized plate of a ferroelectric, the polarization field is balanced by the surface charges. After the quick application of an inverse voltage to the surface electrodes (one of which is transparent to electrons), there is an unbalanced surface charge left on the surface of the ferroelectric, which induces a strong electric field and initiates the emission of electrons into the vacuum.

In recent years, efficient photocathodes subjected to laser irradiation have become important sources of electrons (Le Duff, 1990). Photocathodes with negative electron affinity made of p-type semiconductors having epitaxial layers (GaAsP, GaAs, InAsP + Cs, O) and activated by multilayer structures of cesium and oxygen are in common use. Values of $Y \approx 0.2$–0.5 of quantum efficiency of these photocathodes are reached at the minimum excess of photon energy over the threshold. It enables intense electron beams to be obtained with minimum spread of energy. Quick modulation of laser irradiation (up to 10^{-10} s) provides electron-beam modulation at the same rate, which is important for the generation of superhigh-frequency radiation, for free-electron lasers, and for carrying out a series of unique experiments. Application of circularly polarized radiation provides the photoemission of electrons with a definite spin orientation (obtaining polarized electrons).

Other ways of obtaining polarized electrons are based on the polarization of atoms in a magnetic field by circularly polarized radiation (optical pumping) and further ionization of polarized atoms with polarized electrons being freed. It is efficient to use metastable helium atoms in the 2^3S state with forbidden optical transition to the ground state. With optical pumping, it is possible to achieve a high density of polarized metastable helium atoms that can generate free polarized electrons in collisions with other molecules through Penning ionization (the expenditure of the energy of excitation of the metastable state on ionization).

Polarized electrons are pulled out of the plasma by the extracting voltage through the emission aperture.

The most generally used source of electrons is the electron projector, which produces the narrow electron beam in electron-beam (cathode-ray) tubes. The typical design of the electron projector is shown in Fig. 7. It comprises the heating cathode K, the Venelt electrode B, the first anode A_1, and the second anode A_2 connected with the conductive bulb coating. The oxide cathode applied to the end of the cap emits electrons. The size of the emitting zone and beam intensity are regulated by a rather low negative potential on the Venelt electrode, which creates a short-focus lens for the emitted electrons. The next electrodes accelerate and shape the beam, producing an image of the beam crossover formed at the cathode. A similar design is used for the electron sources in electron microscopes, but the emitter used is either a small spot on a straight-channel tungsten "stud" or the point of an autoelectronic or field emitter.

In the electron guns of electron-beam microwave generators (klystrons, traveling-wave tubes, inverse-wave tubes, generators of diffraction radiation, etc.), electron beams of high current density are formed by Pierce electrodes establishing the distribution of the potential at the beam boundary, which is close to the distribution of the potential in beams with space charge in flat, cylindrical, or spherical diodes. Typical configurations of electrodes are shown in Fig. 8. Indirectly heated cathodes, namely oxide cathode, L cathode, impregnated cathode, etc., are used as emitters. Computer simulation enables the optimization of the configurations of electrodes to derive the required beam parameters and to work out electron sources possessing the required characteristics. The example of an electron gun with a real beam compression of up to 3000 times is given in Fig. 9.

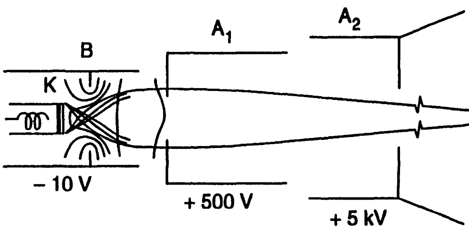

FIG. 7. The typical design of an electron projector. It comprises the heating cathode K, the Venelt electrode B, the first anode A_1, and the second anode A_2 connected with the conductive bulb coating.

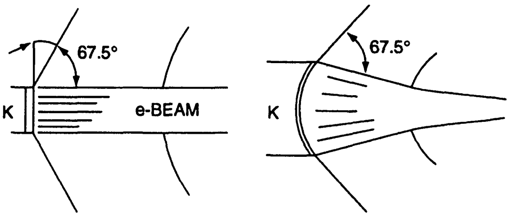

FIG. 8. Typical configurations of electrodes for a highly compressed electron-beam source.

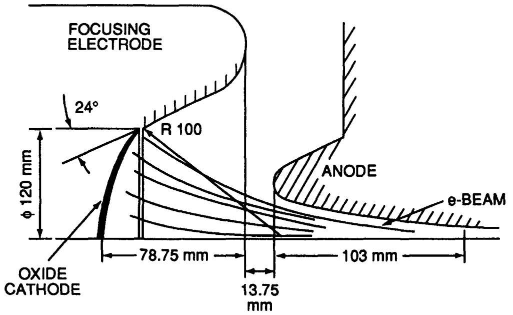

FIG. 9. An example of an electron gun with a real beam compression of up to 3000 times.

Plasma sources of electrons with the design close to that of plasma ion sources, but with changed extractor voltage polarity, are used for technological applications.

Impulse electron beams (5–10 μs) with intensities up to 10^6 A are obtained in sources of explosive emission (Bystritskiy and Didenko, 1989). Rapidly rising voltage (10^5 to 10^6 V) is set up between a flat cathode and an anode made of thin metal foil. High electric field strength at the cathode micropoints initiates field emission with rapid heating and explosion of the micropoints, which produces dense plasma on the surface. Electrons are extracted from this plasma, accelerate in the gap, and pass through the foil into the drift space. To maintain uniform emission, a high density of microscopic whiskers is achieved on the cathode. Stable emission is produced by cathodes made of graphite or carbon material, velvet, and velvetlike materials. The magnetic field of the electron flux focuses the electrons on the center into a small spot, and the space charge of the electrons suppresses the emission from the central part of the cathode. To maintain uniform emission on the entire surface of the cathode and to form a parallel electron beam, a strong longitudinal magnetic field is applied (0.1–1 T). The length of electron impulse is limited by the time necessary for the cathode plasma to reach the anode, and it is 10 μs when the gap width is $d \approx 20$ cm.

3. PHYSICAL PRINCIPLES OF ION PRODUCTION

3.1 Positive-Ion Production

The method of electron ionization of atoms and molecules in the gaseous phase still remains the basic one for positive-ion production. The ion flux coming through holes in a gas-discharge tube cathode (canal rays) was described by E. Goldstein in 1886. Atomic and molecular electron ionization cross sections increase as the energy of electrons W rises from the threshold of the ionization potential I up to $W \approx (3\text{–}4)I$. The magnitude of the cross section reaches a maximum value of $\sigma \approx 10^{-16}\text{–}10^{-14}$ cm^2 and then decreases as W^{-1} after the maximum point. At a gas pressure of 1 Torr, an electron produces up to 1 ion per cm in He, up to 10 ions in N_2 or O_2, and up to 20 ions in Hg. Electrons may be captured by ions in collisions with atoms and molecules; thus they experience charge exchange. The charge-exchange cross sections are very large (up to 10^{-14} cm^2) for resonant charge exchange of the same atom and ion at low collision energies (Green, 1984; Forrester, 1988).

Ions disappear (recombine) when they capture free electrons, electrons of negative ions, or electrons in solids. In the process of recombination, the energy expended upon ionization is released again by emission of a photon, transmission of energy to other particles, or excitation. In a plasma with an electron velocity distribution $f(v)$, the rate of ion production is proportional to the gas density n, the electron density $n_e = n_i$, and the ionization rate

$$\langle \sigma v \rangle = \int_0^\infty \sigma(v) f(v) v dv. \tag{13}$$

The last increases as the temperature of the electrons, T_e, rises. The electron and ion formation rate is in equilibrium with recombination in wall collisions. To suppress this effect, a magnetic field and special electrode configurations are used, namely electron and ion traps. To extract ions or electrons and form beams, a potential difference $U_0 \approx 1\text{–}10^2$ kV accelerating ions toward the extractor is established. This potential is established between the region of ion generation and the extractor (an electrode with an aperture that lets beams through). To sustain a strong electric field within this gap d without breakdown, the gas density in the gap should be low enough ($p < 10^{-4}$ Torr). This circumstance limits the maximum gas density n in the ion source. In ion sources for mass spectrometers, the gas is ionized by the electron beam generated by an ion gun ($I \approx 1$ mA, $W \approx 10^2$ eV).

In more powerful ion sources, the gas discharge produces a dense plasma, $n \approx 10^{10}\text{–}10^{14}$ cm^{-3}. In sources containing a hot cathode but no gas, the flow of electrons from the cathode is limited by space charge to a low level. With gas being delivered, the charge of the formed ions and the space charge of the electrons compensate each other, and inside the source a quasineutral plasma is formed. The ion flux with mass M coming onto the electrode with negative potential U_0 in the plasma is determined by the ion density $n_i = n_e$ and the temperature T_e of the electrons according to the Bohm formula,

$$j^+ = 0.4 n_i e (kT_e/M)^{1/2}. \tag{14}$$

The variation of thickness of the space-charge sheet with the directional flux j^+ is determined by the formula

$$d = (2/3)(2e/M)^{1/2}(\epsilon_0 U_0^3/j^+)^{1/2}. \tag{15}$$

If we make an opening $2r < d$ in size, we will have the elementary scheme of ion-beam production. Equations (14) and (15) provide elementary estimations of the beam production characteristics: the emission current density j, the plasma density n_i, the extraction voltage U_0, the extraction gap d, and the size r of apertures in the plasma electrode, which separates the extractor and the reservoir of the source with high gas density inside. An elementary scheme of an ion source with extractor is shown in Fig. 10.

Improved extracting systems included several electrodes that provide beam production with minimum divergence and aberrations. An example of such a system is shown in Fig. 11. Computer simulation using the Vlasov equation and the particle method enables one to design perfect systems (Whealton, 1981; Whealton *et al.*, 1988, 1989). The local divergence (angular width) of a beam, $\Delta\alpha$, is determined by the temperature of the ions, $T_i \approx 1$ eV, and the accelerating voltage U_0 $[\Delta\alpha \approx (T_i/eU_0)^{1/2}]$, the minimum beam emittance (the measure of the volume in phase space occupied by the beam) being close to $\epsilon = \pi r \Delta\alpha$ and the normalized emittance $\epsilon_n = \epsilon\beta = \epsilon v/c$.

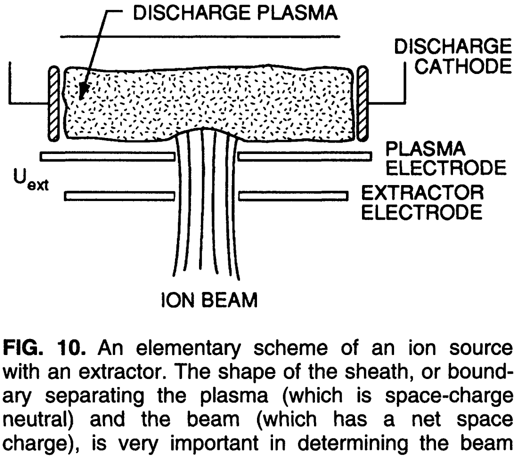

FIG. 10. An elementary scheme of an ion source with an extractor. The shape of the sheath, or boundary separating the plasma (which is space-charge neutral) and the beam (which has a net space charge), is very important in determining the beam properties.

The emission of electrons from the cathode is determined by the ion current in the direction of the cathode. In discharges with only a cold cathode, secondary emission of electrons is initiated by bombardment with ions, $j_e = \gamma j^+$, and characterized by the generalized ion–electron emission coefficient γ. For pure metals when the energy of the ions is as much as 10^3 eV, $\gamma \approx 0.1$ to 0.01. For materials covered by insulator films, such as Al_2O_3 and MgO, secondary emission significantly increases as a result of the strong electric fields within the film. The efficiency of secondary emission (caused by bombardment with ions, atoms, electrons, and photons) is provided by the activation of the cathode surface by alkali-metal and alkaline-earth elements. The space charge begins to limit the electron current from the cathode at the level $j_e \approx j^+(M/m)^{1/2}$. Efficient thermionic emitters are capable of emitting such a current; they are equivalent to cold cathodes having $\gamma \approx (M/m)^{1/2}$. The discharge voltage is determined by the following condition: Having been emitted from the cathode, an electron with energy eU_d should make γ^{-1} ions return to the cathode in the discharge and cause a new electron to be emitted into the discharge. The minimum energetic cost of ion production in the discharge is $p \approx 3I$. For discharges with cold cathodes, $\gamma^{-1} \approx 10$–10^2 and $U_d \approx 10^3$ V; for discharges in thermionic emitters, $U_d \approx 10^2$ V. In ion sources using electron oscillation, the electrons remain in the reservoir until they expend their energy on ionization and excitation, ions being efficiently returned onto the cathode by the magnetic field. The discharge voltage U_d is weakly dependent on the discharge current I_d. In efficient ion sources, a significant portion of the ions (~10%) is extracted and forms the beam. In plasma sources having high beam intensity $I^+ \approx 1$–100 A, the characteristic energy expense is $p \approx 1$–10 kW/A; these expenses are substantially higher in sources with lower intensity.

Repeated and successive ionization followed by the production of multiply charged

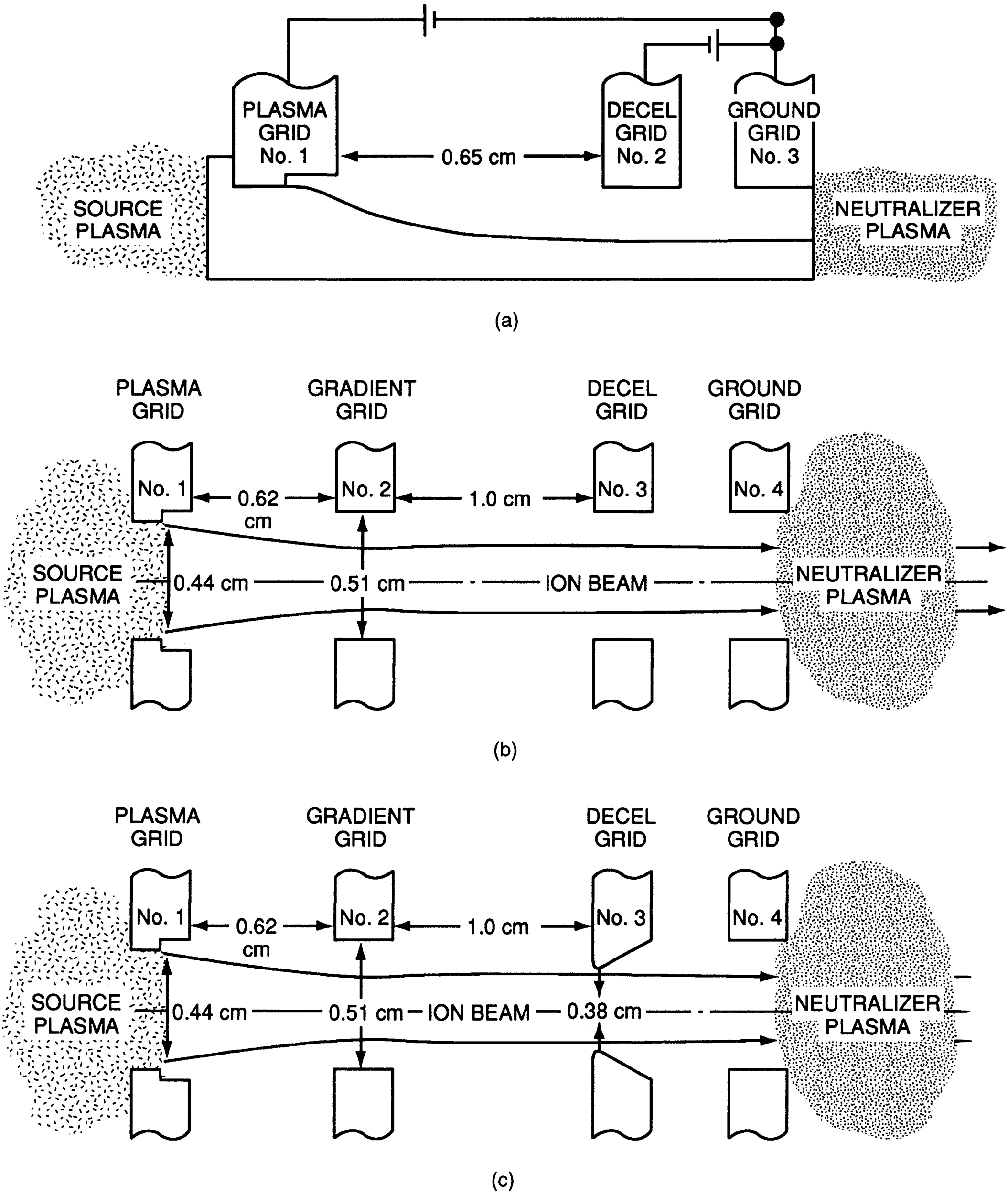
PLASMA GRID No. 1
0.65 cm
DECEL GRID No. 2
GROUND GRID No. 3
SOURCE PLASMA
NEUTRALIZER PLASMA
(a)
PLASMA GRID
GRADIENT GRID
DECEL GRID
GROUND GRID
No. 1
0.62 cm
No. 2
1.0 cm
No. 3
No. 4
SOURCE PLASMA
0.44 cm
0.51 cm
ION BEAM
NEUTRALIZER PLASMA
(b)
PLASMA GRID
GRADIENT GRID
DECEL GRID
GROUND GRID
No. 1
0.62 cm
No. 2
1.0 cm
No. 3
No. 4
SOURCE PLASMA
0.44 cm
0.51 cm
ION BEAM
0.38 cm
NEUTRALIZER PLASMA
(c)

FIG. 11. Example of an improved extraction system, which includes several electrodes that provide beam production with minimum divergence and aberrations. Computer simulation using the kinetic Vlasov equation and the particle method enables one to design perfect systems.

ions is performed in the ion sources with efficient confinement of ions in the magnetic field and high energy of electrons. The charge Z is determined by the electron current density j_e and the ion lifetime in the source, τ. In the latest sources with $j_e \approx 10^3$ A/cm^2 and $\tau = 1$ s, ions having $Z = 80$ may be obtained, but the intensity of such beams is not high.

Alkali metals have an ionization potential I less than the work function of pure metals ($\phi \approx 5$ eV). During the adsorption of an atom having $I < \phi$, a valence electron tunnels into the metal to the Fermi level and the particle resides on the surface in the form of a positive ion. If the temperature rises, these particles will evaporate in the form of ions, because the evaporation heat of atoms, l_0, exceeds that of ions, l_+, by $\phi - I$. The surface ionization is characterized by the degree of surface ionization α^+, the ratio of the density of the flux of evaporated ions to that of the flux of evaporated atoms. This flux is given by the Saha–Langmuir formula,

$$\alpha^+ = j^+/j_0 = A_+ \exp[(\phi - 1)/kT], \qquad (16)$$

where $A_+ = \gamma^+/\gamma_0$, the ratio of the statistical weights of ions and atoms at temperature T. The ionization coefficient $\beta^+ = \alpha^+/(1 + \alpha^+) = j^+/(j^+ + j_0)$ is also used.

At low temperatures, the adsorption of alkali metals reduces the work function ϕ to $\phi < I$, the fraction of the evaporated ions being small. If the temperature exceeds the threshold (usually $T \gtrsim 10^3$ K), the atom's lifetime on the surface and the thickness of the adsorbate lessens, while the work function goes up to $\phi > I$ and the surface ionization coefficient discretely increases to $\beta^+ \approx 1$. On the surface of W with $\phi = 4.3$–5.3 eV, we obtain surface ionization with $\beta^+ \approx 1$ for Cs ($I = 3.89$ eV) or Rb ($I = 4.18$ eV). Significant surface ionization is observed for K, Na, Mg, Ca, Sr, Ba, Al, In, Tl, U, Th, etc. For efficient ionization, it is necessary that the evaporated ions be removed by an electric field so that the space charge should not hamper the evaporation of ions.

When the electric field E at the surface of the emitter becomes higher, it leads to a decrease in the ion's evaporation energy according to the Schottky formula, Eq. (8). Also it causes an increase in surface ionization at E from 10^4 to 10^7 V/cm,

$$\alpha_E^+ = \alpha_0^+ \exp[e(eE/4\pi\epsilon_0)^{1/2}/kt], \qquad (17)$$

and a lowering of the threshold temperature at which surface ionization is initiated.

When very high fields are applied, the ion evaporation becomes considerable at low temperatures (field evaporation) because of the tunneling of ions from the adsorptional states on the surface into the potential well of the external field, $U(z) = -eEz$ (see FIELD EMISSION AND FIELD IONIZATION). Cesium ions on the surface of tungsten evaporate at room temperature in the field $E \approx 10^7$ V/cm. Stronger fields are created at points with a very small radius of curvature, $r \approx 10^{-7}$–10^{-8} cm, to achieve ion evaporation with high ionization potential. However, such points grow dull quickly because of field evaporation of the substance. Self-sharpening points having such extremely small radii of curvature are formed when a liquid film, either on a solid point or in a capillary in the presence of a very high electric field, becomes electrohydrodynamically unstable. When the field strength is $E \approx 10^8$ V/cm, the ion field evaporation produces an ion flux with current density reaching up to 10^8 A/cm^2 dc. The flow of the liquid in the film offsets the loss of the substance in the form of ions. Comparatively low temperatures of the liquid point make it possible to obtain a beam of very high brightness (Orloff, 1993; Prewett and Mair, 1991). Collective processes in such dense beams raise the temperature of the ions to 10 eV. But among all the sources developed, such electrohydrodynamic (EHD) emitters made of liquid metals, semiconductors, or alloys and dielectrics remain brightest (Belchenko *et al.*, 1990; Dudnikov and Shabalin, 1992). EHD sources have a beam current up to 10^{-3} A; but when the intensity is $I \approx 10^{-4}$ A, instabilities appear that substantially worsen the beam characteristics. Ions of hydrogen and noble gases are produced in the same way as a result of field ionization at a cryogenic needle. The intensity of such beams is significantly lower, but the brightness may be very high.

Valence electrons in atoms that are closer to the surface have higher energy levels, and so for the adsorbed particles the condition $I < \phi$ is easier to satisfy than for the free particles. The degree of positive ionization of the adsorbate is usually higher than could be expected if we used the Saha–Langmuir formula. Because of the quick nonequilibrium desorption initiated by the impact of ions, photons, and electrons we can observe the emission of particles having higher degrees of ionization. Previously this process was used to diagnose the surface composition.

Illumination of the surface by strong laser radiation fluxes and intense electron and ion beams initiates intense ion emission.

For production of nuclear polarized positive- and negative-ion beams, various processes are being used (Schiemenz, 1992):

1. Nuclear polarized atoms are produced by means of sextupole focusing of slow atoms with one projection of electron spin; conversion of electron polarization into a nuclear one by radio-frequency transition or Sona field reversal; transformation of nuclear polarized atoms into negative or positive ions; focusing them into a beam and accelerating. Slow atoms with electron polarization may be obtained by some other methods: by accumulation in liquid helium film in the presence of a strong magnetic field, by chemical polarization of nuclei, by optical pumping, and by spin exchange with optically polarized atoms of alkali metals. The transformation of polarized atoms into ions is performed in electron-beam ionizers, in rf and microwave discharge ionizers, and by charge exchange with atom and ion beams. The beams of highest intensity (6 mA of H^+ ions and 1 mA of H^-) are obtained by means of resonant charge exchange on positive and negative ions.

2. In sources using the Lamb shift, a beam of metastable hydrogen atoms in the 2*S* state is formed through charge exchange in the proton beam (deuteron beam) passing through a target of cesium vapor. The binding energy of an electron in cesium is close to that in hydrogen and deuterium in the 2*S* state. When the proton energy is $W = 0.4$ keV, the yield of the 2*S* state reaches its maximum. In a transverse electric field, the particles with one electron-spin projection make the transition from the 2*S* state to the 2*P* state and immediately emit photons. By a quick change of the direction of the accompanying magnetic field, the electron polarization is converted into the nuclear one. After that, the metastable nuclear polarized atoms transform selectively into positive or negative ions in iodine or argon vapor targets. These sources are small in size and have much power economy, but their intensity is not high (10^{-6}–10^{-4} A). The methods of selective lasing and ionization of metastable atoms by means of resonant laser irradiation, creating no limits for intensity, are very promising.
3. Beams of accelerated atoms with electron polarization are obtained by means of proton charge exchange in polarized targets. Polarization of atoms in an alkali-metal target is achieved by optical pumping by means of circularly polarized laser radiation in a magnetic field. A high magnetic field helps to keep a high degree of polarization. Then the electron polarization is converted into the nuclear one by magnetic field reversal. After that, atoms having nuclear polarization turn into positive ions in the process of charge exchange in helium or into negative ions in alkali-metal targets (cesium or sodium).

3.2 Negative-Ion Production

Active research on techniques for negative-ion production is one of the directions for source development. An electron has a low binding energy in a negative ion ($S = 0.75$ eV in H^-), which therefore dissociates easily in collisions with electrons, ions, or neutral particles. Positive ions, however, neutralize only in wall collisions. The high destruction probability for negative ions makes it very difficult to produce them. Negative ions are formed in collisions of electrons with molecules. As a rule, an electron is bound by a molecule temporarily to form an autoionized molecular negative ion that later falls apart in different ways (Green, 1984; Forrester, 1988):

$$H_2 + e \rightarrow H_2^- \rightarrow \begin{cases} H_2(v,J) + e & \text{(vibrational–rotational excitation)} \\ H^- + H & \text{(dissociative attachment)} \\ H^- + H^+ + e & \text{(polar dissociation)} \end{cases}$$

The cross section for negative-ion formation from the ground molecular state is 10^{-20} cm^2, increasing to 10^{-16} cm^2 in collisions of electrons with molecules excited by vibration and rotation ($v > 6$, $J > 6$).

The cross section for negative ions recombining with positive ions is large: $\sigma_r \approx 10^{-14}$ cm^2. The cross section for H^- destruction by electrons is $\sigma_e \approx 10^{-14}$ cm^2 with a broad maximum close to $W \approx 10$ eV and a threshold at $W = 0.75$ eV. The vibrational–rotational excitation of molecules decreases the energy of those electrons that form negative ions efficiently, at the same time causing a decrease in the rate of negative-ion destruction. Even when optimal electron-energy distributions are used, the cross section for negative-ion formation is much smaller than that for their destruction. In plasma H^-

sources, the emission current density is 0.01–0.1 A/cm^2. Up to 3% of the gas delivered may be transformed into H$^-$ ions. A considerable problem with negative-ion sources concerns the extraction of the large beams of accompanying electrons and negative-ion destruction in the gas coming from the source.

Previously great attention was paid to the charge-exchange method for negative-ion-beam production that involves the capture of two electrons by a positive ion. This method is most universal. At present it is used when other methods lack efficiency. In charge-exchange sources in the initial state, a positive-ion beam possessing the required energy and intensity is generated. In the next stage, positive ions are passed through an appropriate target, attaching two electrons and turning into negative ions (Green, 1984; Forrester, 1988; Belchenko *et al.*, 1990).

The optimal negative-ion yield, F^-, is determined by the ratio of their formation and destruction cross sections. For protons in hydrogen, $F^- = 0.02$ at the energy of $W = 10$ keV; in sodium, $f^- = 0.1$ at $W = 10$ keV; in cesium, $F^- = 0.5$ at $W = 0.4$ keV. Helium atoms produce He$^-$ ions when in the metastable 2S state; only the charge-exchange method can provide efficient He$^-$ ion production. Beam charge exchange is used for the production of positive and negative ions with nuclear spin polarization.

On surfaces with lowered work function ϕ, the particles having high electron affinity S are adsorbed in the form of negative ions (the work function increases during this process). Such particles evaporate in the form of negative ions. Negative surface ionization is characterized by the degree of surface ionization α^-, which is the ratio of the negative-ion current density to the atom flux. In equilibrium conditions, α^- is determined by the Saha–Langmuir formula,

$$\alpha^- = A_- \exp[(S - \phi)/kT], \tag{18}$$

where A_- is the ratio of the statistical sums of negative ions and atoms.

Only halogens can be ionized with considerable efficiency on the surfaces of refractory metals ($\phi \approx 4.3$ eV). High ionization efficiency is achieved for many substances on the surface of LaB$_6$, which has $\phi = 2.6$ eV and is highly tolerant of poisoning. On the surfaces of tungsten or molybdenum coated with cesium in optimal concentration, the formation of H$^-$ and D$^-$ is observed. As a particle approaches the surface, the level of its electron affinity lowers by $\Delta S = e^2/16\pi\epsilon_0 z$; therefore, $S + \Delta S > \phi$ for many substances on surfaces with low work function, and the electrons from the solid fill up the electron-affinity levels. However, when the particle moves away slowly, the electron affinity rises above the Fermi level again, and electrons once more tunnel into the solid body. Finally, only an exponentially small portion of the particles is evaporated in the form of negative ions according to the Saha–Langmuir formula.

On the other hand, quick removal of the particles from the surface by sputtering, reflection, or stimulated desorption with electrons or photons may leave no time for an electron to tunnel into the metal, and therefore it is "carried away" from the surface onto the electron-affinity level at the expense of the initial kinetic energy of the particle $W_\perp = mv_\perp^2/2$. The negative ionization coefficient b is defined by the expression

$$\beta^- = \frac{j^-}{j^{\mp}j_0} = \exp\left[-\int_{z0}^{\infty} \frac{w(z)dz}{v_\perp(z)}\right], \tag{19}$$

where $w(z)$ is the frequency of electron tunneling from the electron-affinity level at the distance z from the surface. The probability $\beta^-(v)$ of the fast particles being negatively ionized rapidly increases for particles with energy higher than the threshold energy $W_\perp > \phi - S$ and with the ability to escape from the surface in the form of negative ions (Belchenko *et al.*, 1990; Dudnikov, 1992). For sputtered and reflected hydrogen particles, the capture of the surface electrons intensifies dramatically as ϕ decreases to $\phi < 1.5$ eV through the adsorption of optimally concentrated cesium on the surface. Other alkali-metal and alkaline-earth admixtures catalyze negative-ion production this same way; however, cesium catalysis is much more efficient (Dudnikov, 1992).

In the sources using sputtering, a target having a specific form is subjected to cesium-ion beam bombardment with the energy of 1–10 keV and a current of 1 mA. To minimize the work function, cesium is deposited on the surface. Desorbed and sputtered negative ions with a current up to 10^{-3}–10^{-4} A are accelerated by the extracting voltage and focused into a beam. The sources with sputtering allow us to obtain a wide set of atomic and molecular negative ions (Alton, 1993).

Significant improvement of ion-source characteristics and the enhancement of negative-ion production efficiency have been achieved in surface-plasma sources in which cesium catalysis is used for negative-ion production (as discovered by Dudnikov in 1971). In surface-plasma sources, an electrode with negative potential is immersed into a plasma, and then the cesium admixture is delivered. Cesium ions are adsorbed on the surface, lowering the work function, and the extraction of ions from the plasma provides intense bombardment of the electrode surface, initiating the emission of intense negative-ion fluxes.

Because the space charge is canceled out, the limitations for the negative-ion beam intensities are eliminated, while rapid acceleration in the layer close to the electrode minimizes the negative ions' destruction when they move through the plasma toward the emission aperture. Fast negative ions coming from the electrode may be converted into cold negative ions as a result of resonant charge exchange with cold atoms. It enables beams of high brightness to be obtained. The emission current density for H^- ions reaches the level of 8 A/cm^2 dc.

Intense laser and electron irradiation of the surfaces with minimized work function initiates the negative-ion emission. Accompanying electron fluxes are suppressed by using specific configurations of extracting systems and a magnetic field (Dudnikov, 1992).

Self-consistent ion-extraction optics are very important for the production of the brightest beams. Such calculations are difficult because of the high degree of nonlinearity of the formulation, and the large number of dimensions that needs to be considered (Whealton, 1981). For negative-ion modeling, the situation is exacerbated by the determination of a formulation that is both reasonably self-consistent and tractable (Whealton *et al.*, 1988, 1989).

When applied to the surface, a high electric field causes an increase in the probability of particle desorption in the form of negative ions. In practice, such a process is performed in self-sharpening EHD emitters with the points covered by films of ion salt melts (Dudnikov and Shabalin, 1992). The surface-plasma sources are used as ionizers in the production of negative ions with nuclear polarization.

4. ACTUAL ION SOURCES

4.1 Positive-Ion Sources

The considered principles of ion production are used for the creation of ion sources with required characteristics: beam intensity I and ion energy W, composition of ions, required emittance $\epsilon = \Delta r \Delta \alpha$ and brightness $B = 2I/\epsilon^2 = j/T_i$, acceptable energy ($I/P$) and gas ($I/Q$) efficiency, required operational time and reliability. The achievement of these necessary characteristics is a result of ingenuity and deep optimization.

In real ion sources, efficient ionization must take place in a condition of low gas density, when the ionization free path length of electrons significantly exceeds the dimension of the gas-discharge chamber. The classical solution of this problem is an electron-oscillation gas discharge (PIG, from its origin in the Penning ion gauge), in which the electrons are oscillated along the magnetic field inside an anode tube with reflection from the opposite cathode plates at negative potential.

In a strong magnetic field, it is possible to initiate a gas discharge with a cold cathode and low gas density, but for the discharge transformation to the low-voltage, high-current mode (to the superdense glow discharge) with $U_d = 0.4$ kV and $I = 1$–100 A, a gas pressure of 10^{-2} Torr is required. The ion beam can be extracted from the gas-discharge plasma through a hole in the cathode plate along the magnetic field or through a slit in the anode tube perpendicular to the magnetic field. The use of a thermionic cathode is enhanced by the transition into the low-voltage ($U_d = 50$ V) discharge with low gas density and with an acceptably long operational time. The structures of Penning ion sources are extremely multifarious (Alton, 1993; Kutner, 1994). They have various electrode configurations and materials, extraction systems, and gas-injection systems. Some versions have electromagnets; others have permanent magnets. An example of a Penning-discharge ion source with self-heated LaB cathodes is presented in Fig. 12. The distinctive features of this source are low U_d = 80 V, with I_d = 0.8 A and j^+ = 0.6 A/cm^2, high proton content (up to 60%), high gas efficiency (up to 60%), and long operational time (10^2 h).

Another famous version of an ion source with electron oscillation is the Calutron, developed for high-productivity isotope separation. The scheme of this source is shown in Fig. 13. A thin plasma column is created by a thermionic cathode discharge in a strong magnetic field. The electron oscillation is provided by reflection of electrons from an oppositely located anticathode (a reserved cathode). As a thermocathode is used, a dc-heated pin from a tungsten or tantalum filament or massive tungsten block is heated by electron flux. An ion beam is then extracted through a slit (up to 1 × 20 cm^2) in the graphite or metallic body of the source anode. A three-electrode extraction accelera-

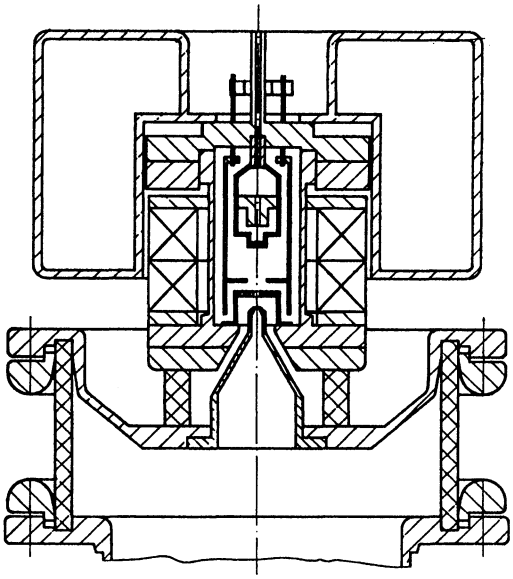

FIG. 12. An example of a Penning-discharge ion source with self-heated LaB cathodes.

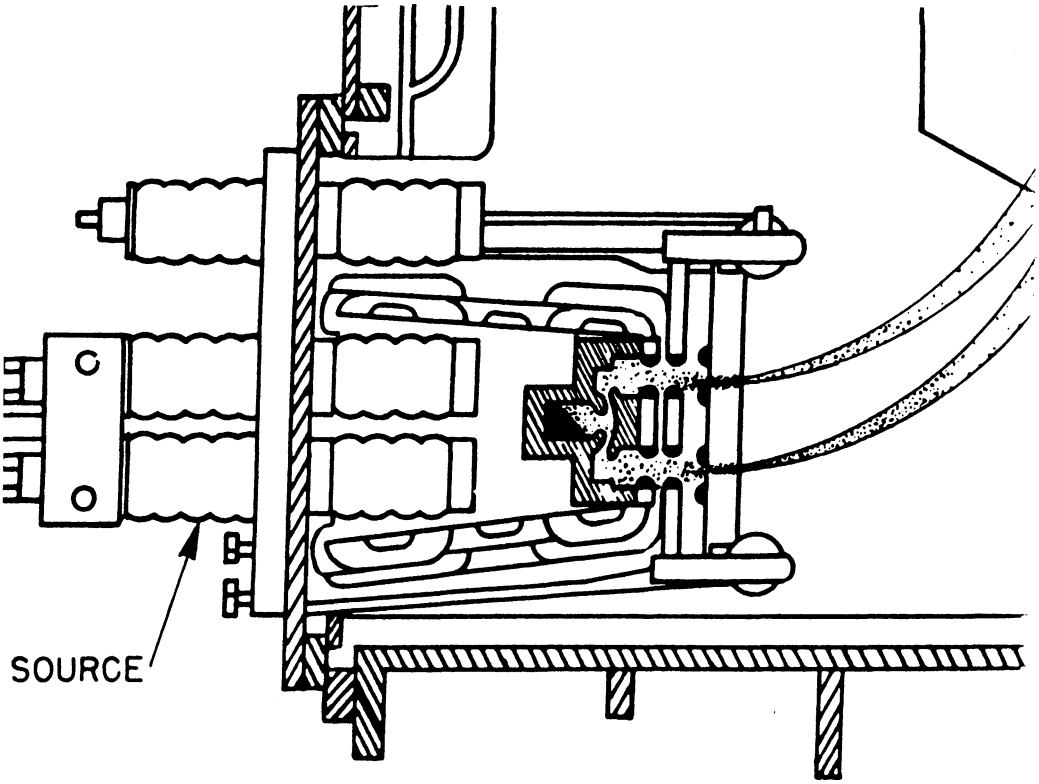

FIG. 13. Another famous version of an ion source with electron oscillation, the Calutron, developed for high-productivity isotope separation. A thin plasma column is created by the thermionic cathode discharge in a strong magnetic field. The electron oscillation is provided by reflection of electrons from an oppositely located anticathode (a reserved cathode).

tion–deceleration system with specific profiles of electrodes is used for forming a good ribbon beam convenient for separation in a magnetic field (Forrester, 1988; Kutner, 1994).

Gases of separated elements, compound vapors, and sputtered solids are used as working substances. A sputtered electrode is immersed into the plasma column opposite the extraction slit and biased negatively with hundreds of volts. Ion sources can be heated up to high temperature, (1–2) × 10^3 °C, for creating the required density of a working-substance vapor. Emission current density is 0.1 A/cm^2 for a light ion and decreases as $M^{-1/2}$ for heavy ions. An increase of the magnetic field enhances the efficiency of ion generation, but this increases the fluctuation of the plasma parameters and increases beam emittance. Several versions of Calutrons have been developed for many applications including use in ion implantation (Alton, 1993; Kutner, 1994).

A successful structure of an ion source with magnetron electrode configuration was proposed and developed by Freeman in 1947. The layout of the Freeman ion source is shown in Fig. 14. A heated cathode of tungsten or tantalum wire located along a magnetic field inside the anode cylinder is used. The magnetic field is not too high, B = 10–30 mT. Freeman source characteristics are close to those of Calutrons, but the discharge is really noiseless. This is very important for precise beam formation and for good separation of the beam from impurities. The Freeman source is widely used for ion implantation (Chivers, 1992).

Very high efficiencies of ion generation have been reached in a duoplasmatron, having many applications (Forrester, 1988; Alton, 1993). In a duoplasmatron, discharge between a heated cathode and a small anode with an emission hole is used. The discharge column is compressed by a hole in a floating-potential (isolated) intermediate electrode and a strong magnetic field between this electrode and the anode (Fig. 15). The compressing of the discharge by the intermediate electrode and the magnetic field is created by a double sheath with a spherical surface; accelerated and focused electrons are near the emission hole. The magnetic field is compressed by the plasma and iso-

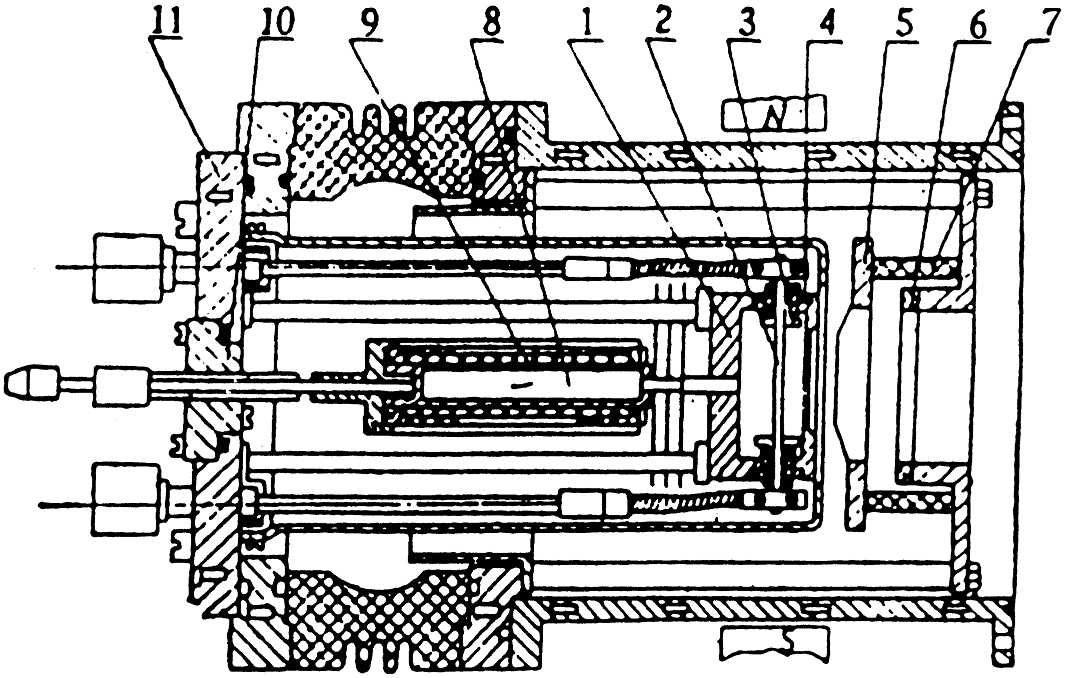

FIG. 14. Layout of the Freeman ion source. A heated cathode of tungsten or tantalum wire located along a magnetic field inside the anode cylinder is used.

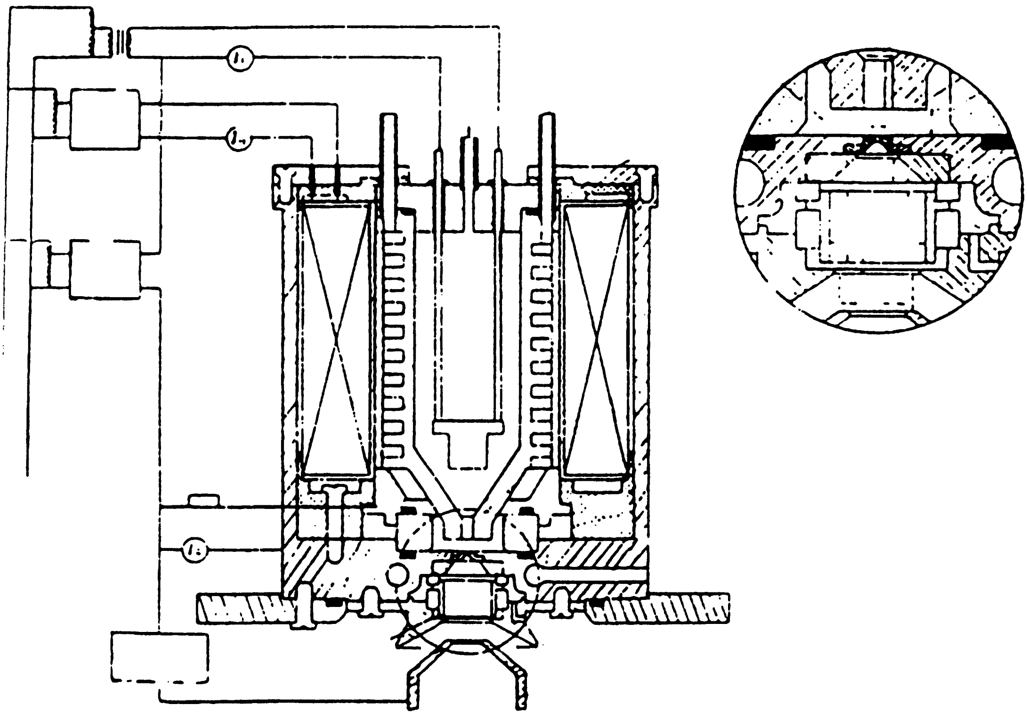

FIG. 15. Duoplasmatron ion source, where the discharge column is compressed by the hole in the floating-potential (isolated) intermediate electrode and the strong magnetic field between this electrode and the anode. The compression of the discharge by the intermediate electrode and magnetic field is created by the double sheath with a spherical surface.

lates it from the wall. Axial extraction of ions is used. It is important to use a nonmagnetic anode insert with an emission hole. Through the emission hole of 0.1 to 1 mm diameter, a plasma jet with very high density (up to 10^{-14} cm^{-3}) is emitted. An ion beam is formed from an expanded plasma flux with lower density and adjusted for beam extraction by a high-voltage electric field. Some duoplasmatrons are operated with a cold hollow-cathode arc discharge or a superdense glow discharge. The energy spread of ions is 6 to 8 eV when ions are extracted from the expander and up to 60 eV for ion extraction from the anode hole. In a duoplasmatron with a strong magnetic field, the ion generation is efficient, but the discharge fluctuation is very high. High levels of fluctuation decrease the brightness B to

$$j/T_{ef} \sim 10^{-2}\ \text{A/cm}^2\ \text{eV} \quad (T_{ef} \sim 5\text{–}10\ \text{eV}). \qquad (20)$$

By the addition of an additional anode between the intermediate electrode and the electrode with the emission hole, a duoplasmatron with electron oscillation is created. By this transformation, increased aperture can be used for intense beam generation. This modified DuoPIGatron with a multiaperture extraction system has intensities up to 70 A. Versions with a multicusp magnetic field near the cylindrical wall have low-noise plasma and better ion-beam characteristics than versions with longitudinal magnetic field in the plasma.

Also, for generation of intense ion beams (50 to 100 A), periplasmatron modification of the duoplasmatron with radial injection of plasma from the hot-cathode toroidal chamber to the central chamber, with a multiaperture four-electrode extraction system, is used (Green, 1984; Forrester, 1988).

After 1978, ion sources were developed with a hot-cathode discharge chamber with a periphery multicusp magnetic wall for suppression of plasma loss. dc-heated tungsten filaments are used as cathodes for the discharge, and the chamber wall or special electrodes as anodes. Good homogeneity of the plasma can be reached on the big surface of the multiaperture four-electrode extraction system. The typical emission current density *j* is 0.25 A/cm², and *I* is 100 A in long (up to 100 s) pulses. Small multicusp sources with *I* ~ 0.1 A have been developed for industrial application. This ion source has a low level of fluctuations and high beam quality (Green, 1984; Forrester, 1988).

In the ion sources considered, the operation time is limited by destruction of the hot cathode to about 10 to 50 h (Forrester, 1988). For overcoming this defect, electrodeless rf and microwave discharges have been developed. In a widely used version of rf sources, the discharge is initiated in a quartz or Pyrex vessel in gas at a pressure of 10^{-2} to 10^{-3} Torr. A capacitor connection with the discharge is used with rf voltage between two rings around the vessel. An inductive connection uses an rf coil of copper tube around an insulating vessel immersed into a metallic chamber. The weak magnetic field (3–10 mT) is optimized by the plasma-generating rf of 2–10 MHz, and power of 10^2–10^3 W is used for plasma production. In the rf source with ion extraction along the inductance coil or capacitor discharge, the energy spread is very large (up to 10^2 eV). Better energy distributions have been produced with a perpendicular inductance coil and extraction along the magnetic field (Fig. 16). It is convenient to use a moving permanent magnet for resonant magnetic field production. By special processing of the vessel wall,

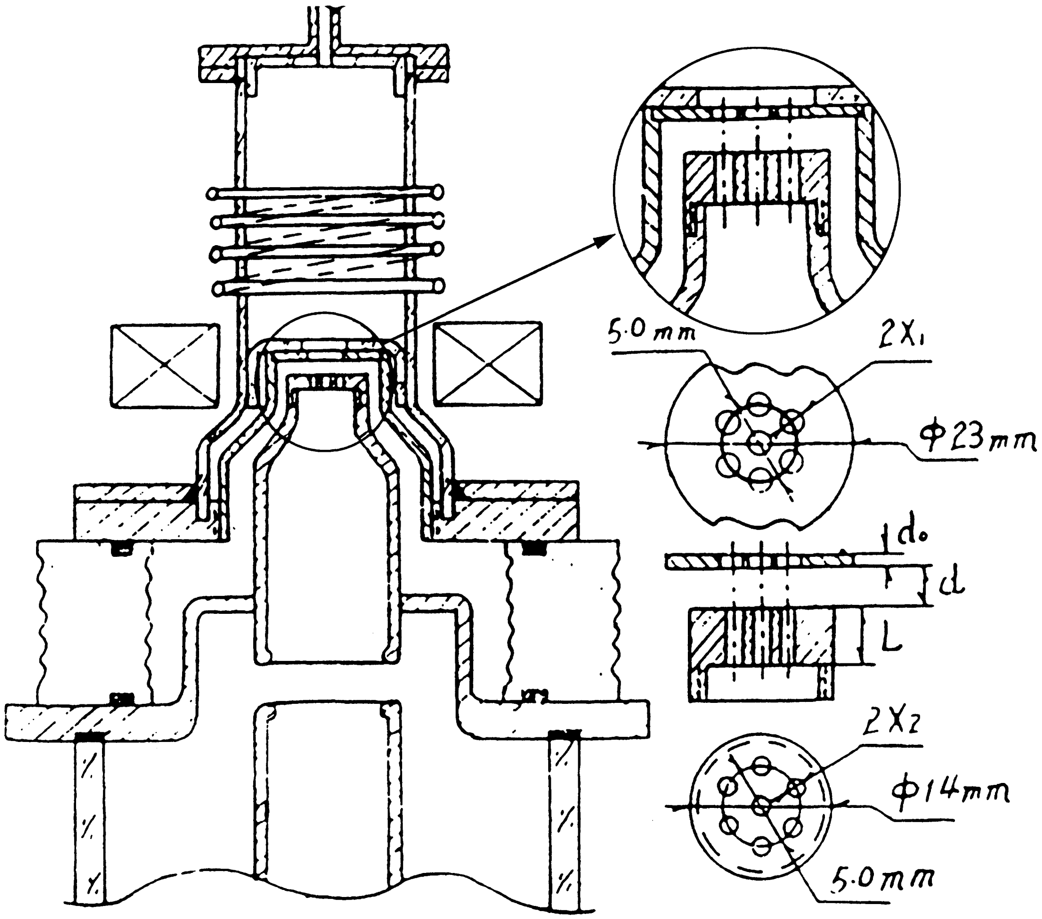

FIG. 16. rf ion source. The inductive connection uses a rf coil of copper tubing around an insulating vessel immersed in a metallic chamber. The weak magnetic field (3–10 mT) is optimized by the plasma-generating rf of 2–100 MHz, and power of 10^2–10^3 W is used for plasma production.

it is possible to reduce the recombination of the atoms and have a high atomic percentage (up to 80%) with low discharge power. Similar rf generators are used for atomic-flow production in polarized sources. The ions escape from the plasma through a hole at the plasma electrode. In a source with probe extraction, a thin positively biased probe is immersed in the plasma. By this probe, the plasma is biased positively relative to the plasma electrode, and ions are extracted and accelerated to the plasma electrode by a potential difference $U = 1$–10 kV. The screening of the plasma electrode by a dielectric diaphragm allows the ion flux to be focused on the emission holes and formed into a beam. Multiaperture probing extraction systems can be used also.

In rf sources with the usual extraction system, the plasma has a potential close to that of the plasma electrode, and a high potential difference has been applied between the plasma electrode and the extractor. Three- and four-electrode extraction systems can be used also.

In recent times, highly efficient plasma generation has been reached in rf discharges with a helical antenna having powerful resonance in strong magnetic fields (10–40 mT).

Microwave sources with electron-cyclotron resonance (ECR) have very high-efficiency plasma generation but a complex power supply (Alton, 1993). For widely used microwave generators with a frequency of 2.45 GHz, the ECR needs a magnetic field of 80 mT. The microwave power is guided into the discharge-chamber cavity by a waveguide and ceramic window and has created high-density plasma near the emission holes. For the production of H^+ of 0.1 A with a current density of 0.3 A/cm^2, ECR power of 0.5 kW is required.

Ion beams of different elements are generated in ECR sources for implantation. A slit extraction system is used for separation of beam components. A dielectric inert material such as BN is used for coating the wall of the discharge-chamber cavity. It is very important for long-term operation.

An improved version of ECR source with enhanced plasma electron confinement is used for high-charge ion production. A magnetic mirror trap with a high mirror ratio ($R = 3$ to 6) and an octupole or sextupole minimum-B system from strong permanent magnets has improved the stability of the plasma and increased the plasma lifetime. Krypton ions with charge Z up to 25 and thorium to 35 are formed in an ECR source ($I \sim 10^{-7}$ A) by increasing the ECR frequency to 16 GHz.

Successful generation of ions was obtained in a high-current arc discharge with a cold cathode (Belchenko *et al.*, 1990). By use of a freely expanding plasma jet and a precise multiaperture extraction system, intense beams of H^+ with very high brightness can be produced. The scheme of this arc-discharge source is shown in Fig. 17. Some fea-

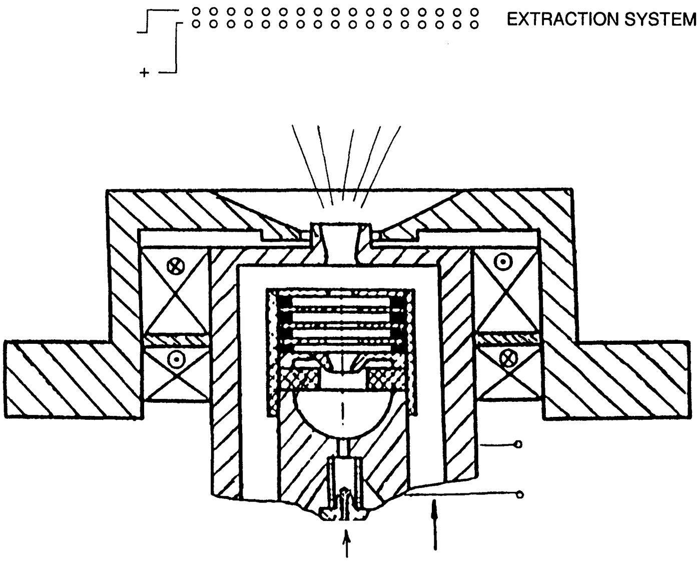

FIG. 17. Dimov arc-discharge ion source.

tures of the source are very high current density in a washed channel, $j = 3 \times 10^3$ A/cm^2; this helps to reach a high ionization coefficient (99.9%) and very high proton percent (95%). A plasma flux with a proton current up to 100 A can be reached. Plasma-jet expansion decreases the local transverse-ion temperature to the level of $T = 0.1$ eV. Normalized brightness $B = 3$ A/mm^2 mrad^2 was reached in an intense beam with energy $W = 1$–30 keV.

Intense generation of ions in the cathode spot of a vacuum arc is the basis for the metal vapor vacuum arc source (MEVVA; Brown, 1989). In the cathode spot with a diameter 10^{-3} mm and an arc current $I_d = 10^2$ A, the electron flux with a current density of $j = 10^8$ A/cm^2 produces intense ionization of the high-density cloud of vapor near the spot. The high-density cloud of plasma has a potential distribution with a maximum at the same place as the maximum of plasma density. By this potential distribution, part of the ions are accelerated to the cathode spot but another part to the anode. The anode flux of ions is formed into a beam by a multiaperture extraction system. The vacuum arc is ignited by the breakdown of an insulator surface and is supported in the vacuum during a pulse of 0.1–1 ms. Many modifications of MEVVA have been developed for industrial applications (Wolf, 1994).

Ions with charge of more than 1 ($Z = 2,3$) are formed in almost all plasma sources; but for generation of ions with significantly higher Z, it is necessary to have plasma electrons with higher energy (some tens of keV) and a long ion lifetime in the electron flux. Ions with $Z = 5,6,7$ are created in ion sources with electron oscillation by increasing the discharge voltage and by reducing the gas density (Kutner, 1994). It is easier to have higher Z in a pulsed discharge. In the electron beam ion source (EBIS), the electron beam from a thermocathode in a strong homogeneous magnetic field, with a current density $j \sim 10^3$ A/cm^2, is used for high-charge ion formation (Drentje, 1994). An ultrahigh vacuum (10^{-10} Torr), the proper quantity of ion injection into the electron beam, and programmable potential barriers near the cathode and near the exit of the electron beam to the recuperator help to retain ions for the stripping of the 35 electrons from krypton. By eliminating the potential barrier near the recuperator, multicharged ions are extracted from the EBIS. The cooling of ions by the addition of light ions improves the ionization level. In multicharged ion sources with ECR, an increase of charge is attained at the end of pulses.

Irradiation of a solid surface by a powerful laser pulse is used for multicharged ion production. Ions with $Z \sim 10$ to 15 are formed in the laser plasma jet. An ion beam is formed by a multiaperture extraction system from the plasma jet. For a CO_2 laser (10.6 μm) and pulse of 0.1 μs, an ion pulse up to 10 μs and a current of 1 A can be produced.

The brightest ion beam is produced in an EHD or liquid-metal ion source by field evaporation of ions ($E \sim 10^8$ V/cm, $j \sim 10^8$ A/cm^2) (Orloff, 1993; Prewett and Mair, 1991). For strong evaporation field formation, a sharp needle with a film of melted substance is used, and positive voltage of 3–15 kV is applied between the needle and extractor plate, which has a hole for the extracted beam. A typical distance for a needle spot-extractor is 1 mm. In simple EHD sources, the needle is welded to the filament pin and heated by dc, and a drop of melt is kept there by surface tension. The ion current is emitted after the threshold level of voltage is reached and after the melting of the working substance.

In an improved version of EHD sources, a special container for the working substance is used, and melted film flows through a small slit between the needle and the container wall (Dudnikov and Shabalin, 1992). This design can be used with a working substance of high vapor pressure and with a melted dielectric with ionic conductivity. The container-type EHD source is shown in Fig. 18. Needles from porous impregnated materials or from filaments are used for EHD ion-beam production. Ion beams from an EHD source can be focused to 0.1–0.03-μm diameter with a current density up to 5 A/cm^2. Full intensity from 0.1 to 1 mA can be reached, but ion-optic aberration for a wide-angle beam is large.

Ion sources with surface ionization evaporate as ions up to 100% of Cs, Rb, and K; up to 20% of Li; about 1% of Sr and In; and about 0.1% of Al, Ga, Tl, and Ca. For beam production, an electrode configuration similar to an electron gun is used. The ionizing vapor of elements or salts is guided to the

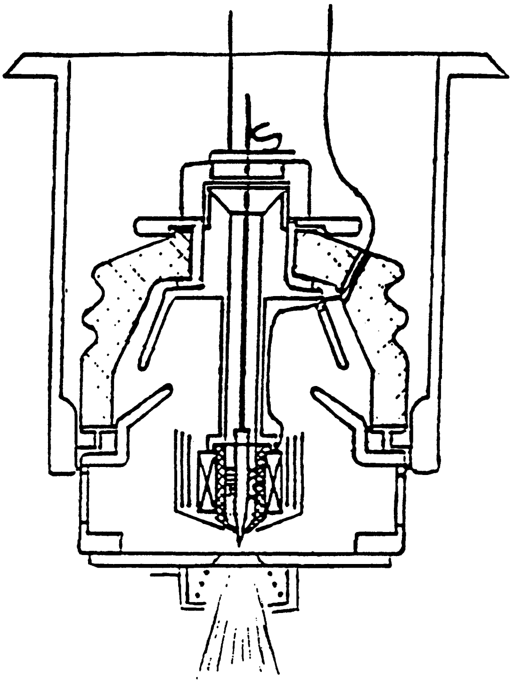

FIG. 18. Container-type EHD source showing the needles from porous impregnated materials or from filaments.

hot (>1000 °C) emitter surface of W, Ta, Mo, Ni, Rt, or Ir. Emitters of porous impregnated materials are used (Forrester, 1988; Alton, 1993).

In the emitters of aluminosilicates of alkali-metal or alkaline-earth elements, the functions of ionizer and source of ionizing substance are combined. Emission current density up to 10 mA/cm^2 and a long operation time can be reached in sources with surface ionization. Beam brightness increases with increasing current density, but operation with a strong electric field on the emitter surface increases the angular spread through the nonhomogeneity of the surface.

4.2 Negative-Ion Sources

Beams of negative ions (H^-, O^-) were registered on the first parabolic mass spectrograms of canal rays produced by J. J. Thomson in 1911. The negative ions were formed by charge exchange of positive ions, accelerated in the near-cathode potential drop of the gas discharge. The purposeful development of negative-ion sources began after a proposal by L. Alvarez in 1951 to use negative ions for multistage acceleration with charge exchange in tandem accelerators. Since that time, many versions of charge-exchange negative-ion sources with double-electron capture by positive ions have been developed; for example, see Fig. 19. As sources of positive ions are used, the following versions of sources have been modified for this application: rf and microwave, duoplasmatron, Penning-type, arc discharges, etc. As charge-exchange targets, tubes are used and a supersonic flow of gas or metallic vapor. The gas or vapor is condensed and recirculated.

In the first versions of negative-ion sources (1956), intensities were reported as high as 1–2 μA. Now the best charge-exchange sources have an H^- beam, acceptable for accelerators, with dc intensity up to 1 mA. The energy of a primary ion optimal for conversion into a negative ion is low (0.3–10 keV). For this reason, sources of primary beams with high perveance (a measure of the ratio of current to transverse space-charge–limit current) have been developed. With an arc-discharge source and a high-perveance multislit extraction system (Fig. 17), pulsed (0.2-ms) beams of H^- up to 15 mA (charge exchange in H_2, 1968) or 0.1 A (charge exchange in Na, 1972), or of He up to 10 mA (in Na) with high beam quality have been produced (Belchenko *et al.*, 1990). Later by use of a wide-aperture neutral-beam injector, H^- and D^- beams up to 10 A and He up to 1 A were obtained, but the current density of these beams is low.

Now charge-exchange ion sources are used for production of positive and negative ions with nuclear spin polarization (Schiemenz, 1992). The capture of polarized electrons from an atomic target by positive ions is used. The electron polarization of an alkali-metal atom target is produced by optical orientation (optical pumping), using circularly polarized resonant light in a strong magnetic field. The electron polarization of an accelerated atomic beam is converted into nuclear polarization by rf transition and by fast reversal of a longitudinal magnetic field (Sona transition). Nuclear polarized fast atoms are converted into negative ions by the capture of a second electron in a second target. The efficiency of conversion of H^+ into polarized H^- is $<10^{-2}$. For high-quality po-

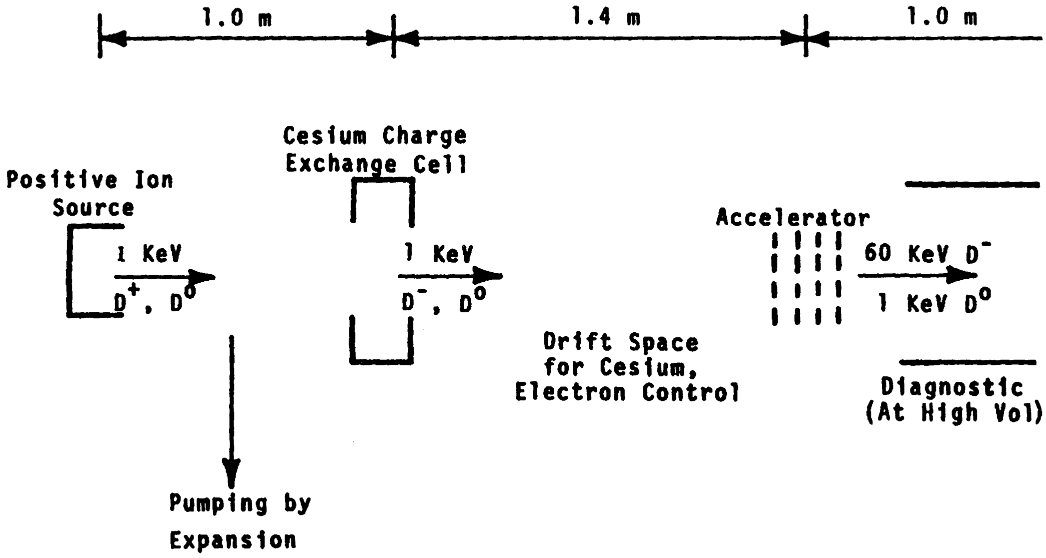

FIG. 19. Double–charge-exchange negative-ion source.

larized H^- beam production, it is necessary to have very high brightness of the primary H^+ beam.

In a plasma source, the negative ions are formed on the periphery of the discharge with reduced electron temperature (Green, 1984; Forrester, 1988). This reduces the negative-ion destruction significantly. In electron-oscillation sources, the emission slit is located a longer distance from the discharge column than in positive-ion sources. Higher gas density is required for increased negative-ion production (Forrester, 1988). Extraction perpendicular to the magnetic field is used to separate negative ions from electrons and to help reduce the escort electron flux. From a Calutron modified by K. Ehlers in 1965, up to 5 mA of H^- with $j^- \sim 20$ mA/cm² were produced. Later the H^- intensity was increased to 40 mA.

In earlier versions of a duoplasmatron, the enhancement of negative-ion generation has been achieved by a shift of the emission hole to the periphery of the discharge (I^- up to 0.2 mA, $i_e \sim 15$ mA). A further development of this idea uses a tubelike discharge around an emission hole (up to 17 mA of H^-).

Modifications of multicusp ion sources with a peripheral magnetic field are used for negative-ion production. A dipole magnetic barrier is created near the extraction system for reflection of energetic electrons from the hot-cathode discharge. This magnetic field helps to reduce the flux of escort electrons. The deflection magnetic field is created in a multiaperture, multielectrode extraction system for reducing the electron flux. From a big multicusp plasma source, up to 3 A of H^- with $j \sim 10$ mA/cm² have been extracted. The destruction of extracted negative ions in collisions with gas escaping from the source is a problem of plasma sources.

Surface ionization of negative-ion sources is similar to that for positive-ion sources, but a material with low work function ϕ, such as LaB_6, is used for negative surface ionization (Alton, 1993). The flux of the working substance is guided to the high-temperature (1000–1500 °C) ionizer surface with a low work function. The evaporated negative ions are extracted by an electric field. A magnetic field parallel to the ionization surface is used for deflection of the escort electrons to a low-potential electrode. The halogens, halides, and other elements and compounds with a high electron affinity S have been converted to negative ions by surface-ionization sources with good efficiency if $S > \phi$. The typical construction of a surface-ionization negative-ion source is shown in Fig. 20. The increase of surface electric field is enhanced by the ionization coefficient. In an EHD source with a very strong electric field

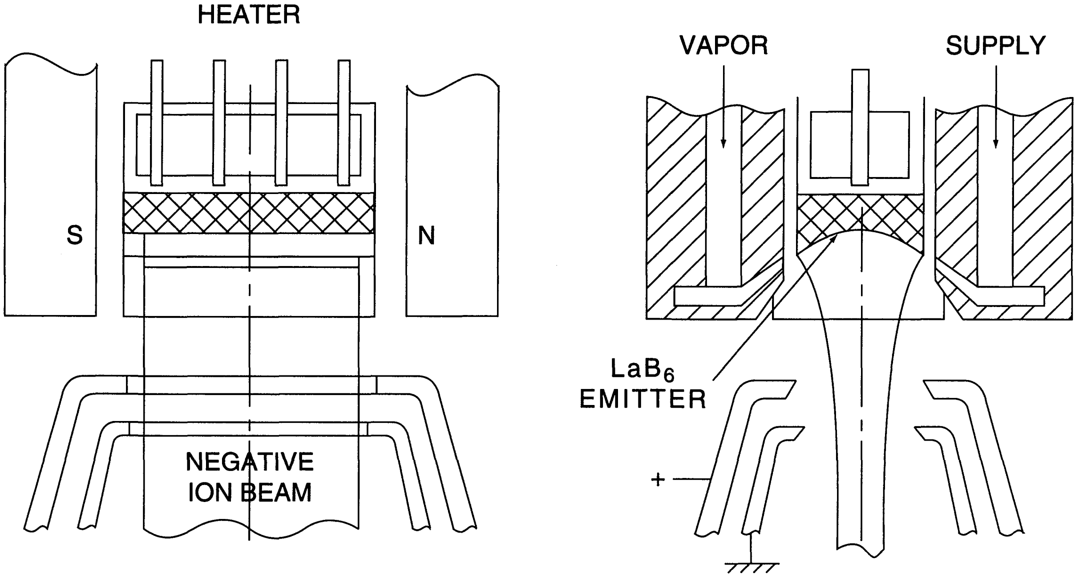

FIG. 20. Typical construction of surface-ionization negative-ion source, in two orthogonal transverse views, the left having the magnetic field in the plane of the drawing and the right having it perpendicular to the paper.

(10^8 V/cm), high–current-density negative-ion beams have been extracted from a low-melting eutectic halide compound (Dudnikov and Shabalin, 1992).

In the more universal sputtering ion sources, the emitter surface with the working substance is bombarded by ion beams of 1–5 mA and energy 5–20 keV (Alton, 1993). Bombardment by Cs^+ ions is convenient, because a compact surface-ionization Cs^+ source is commonly used. The sputtered negative ions are extracted by an electric field and focused by an emitter surface and an ion-optic system. The negative ionization of sputtered particles is catalyzed by optimal cesium deposition on the emitter surface. A large variety of sputtering negative-ion sources has been developed for application in tandem accelerators with intensity of 0.1 mA. Negative ions of almost all elements and compounds with electron affinity are produced in sputtered sources.

The most efficient negative-ion production has been achieved in surface plasma sources (SPS) with cesium catalysis (Dudnikov, 1992). The addition of a small admixture of cesium or other substances with low ionization potential decreases the surface work function and increases significantly the negative-ion emission from all types of plasma sources. But the highest efficiencies have been reached in surface-plasma sources with an electrode configuration optimized for surface-plasma production of negative ions.

The planotron (flat magnetron) was the first version of surface-plasma source (Institute for Nuclear Physics, Novosibirsk, 1971), and it is still used for accelerators and other applications. Figure 21 shows distinctive features of the planotron. This SPS consists of a spool-like cathode and racetrack anode with a small (up to 0.5 mm) gap. A magnetic field is created by magnetic poles along the gap. Plasma is generated by a glow discharge in crossed fields. Along the magnetic field, electron motion is limited by cathode reflectors. Plasma density up to 10^{14} cm^{-3} can be achieved in long-pulse discharges. The emission slit in the anode is oriented perpendicular to the magnetic field and separated from the plasma column by a gap. Cylindrical focusing of emitted negative ions by the concave surface of the cathode is used with a slit extractor, and spherical focusing is used with a circular extraction aperture. A small admixture of cesium increases the negative-ion emission up to 100 times through the reduction of the work function. An emission current density of 5 A/cm^2 was reached. Different versions of the planotron have been developed at Fermi National Laboratory, Brookhaven National Laboratory, SSCL, Ar-

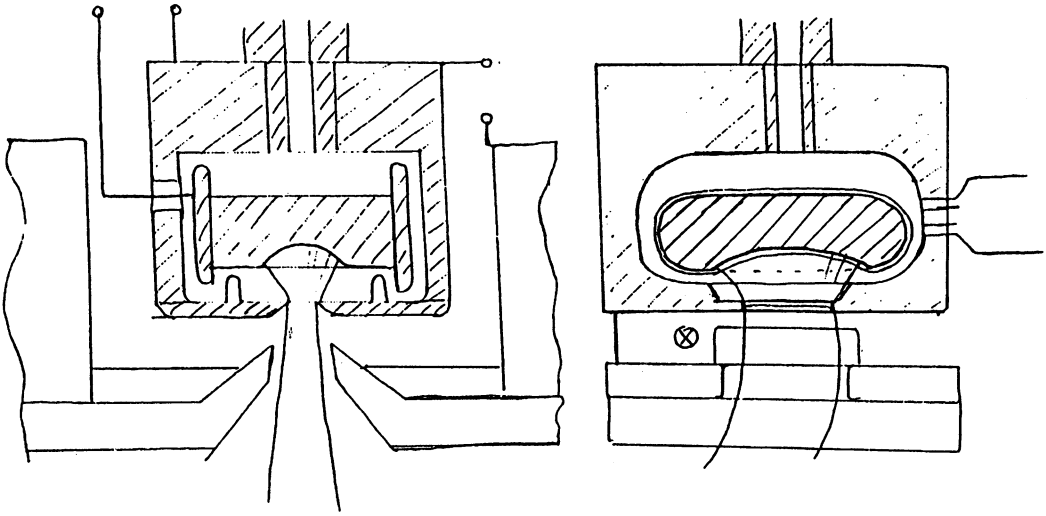

FIG. 21. The planotron (flat magnetron) showing a spool-like cathode and racetrack anode with a small (up to 0.5 mm) gap and a magnetic field created by poles along the gap, and a plasma generated by the glow discharge in crossed fields.

gonne National Laboratory, Deutsches Elektronen-Synchrotron Laborator, and others. It is important to use a high-speed gas valve for pulsed SPS operation. Precisely controlled cesium admixing is very important for long-term reproducible operation. An H^- beam of 80 mA and $\epsilon_{n(90)} = \pi$ mm mrad (normalized emittance of 90% of the beam) can be delivered to a preaccelerator by a planotron.

Another modification of the SPS, the semiplanotron (Fig. 22), is very similar to the planotron, but it uses only the front part of the discharge where plasma is drifted between the concave cathode and the part of the anode with the emission slit. Features of the semiplanotron are increased efficiency of negative-ion production and convenient cathode cooling. The semiplanotron is very good for intense, high-brightness, high–duty-factor negative-ion production.

The Penning SPS (Dudnikov-type SPS) is produced from a planotron by removal of the central plate of the spool-like cathode. A glow discharge, in a crossed field between cathode reflectors and anode, is used for plasma production. Primary negative ions are formed by secondary emission from the cathode surface bombarded by plasma particles. Thermal negative ions are extracted

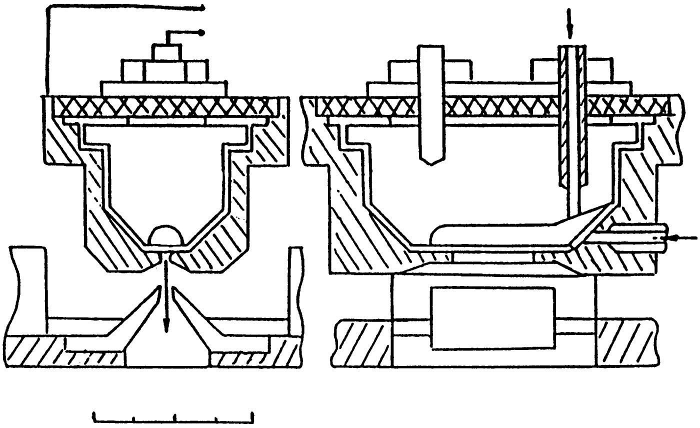

FIG. 22. A modification of the SPS-semiplasmatron, which is very similar to the planotron but uses only the front part of the discharge, where plasma is drifted between the concave cathode and the part of the anode with the emission slit.

through an emission slit after resonant charge exchange of fast negative ions with thermal atoms near the emission aperture. By this process, the negative ions are cooled to a very low temperature (less than 1 eV), and high enhancement of beam brightness can be achieved. In the Penning SPS and semiplanotron, an SPS noiseless mode of discharge can be achieved. It is very important for good negative-ion–beam space-charge compensation and high-brightness preservation during transport.

An independently developed embodiment of the cesium catalysis principle was made in 1977 by W. L. Stirling of Oak Ridge National Laboratory. It ran a $\frac{1}{2}$-A negative hydrogen beam for 30 s. This embodiment is a variant of the Calutron ion source mentioned above.

Many variants of the Penning SPS and semiplanotron with multiaperture slits and circular-hole extractor systems have been developed for production of negative-ion beams with intensity up to 11 A. Small-scale sources with spherical geometrical focusing and hollow-cathode discharge have been developed for long-term dc operation. Negative-ion beams with $j^- = 0.3$ A/cm^2 ($I = 3$ mA) were extracted during 10 h of operation.

A multicusp magnetic-wall SPS was developed at Lawrence Berkeley Laboratory (1982). This large-volume, low-pressure SPS can operate at high duty factor or dc. The source chamber is a large box or cylinder (~20 cm) with alternating polarity of permanent magnetic dipoles on the wall surface to produce a multicusp magnetic field. A negatively biased (~0.3–0.6 kV) emitter convertor is immersed in plasma. Under bombardment by plasma particles, negative ions are formed and emitted from the emitter surface. The emitter surface is curved to give geometrical focusing of negative ions to the extraction aperture. Negative-ion beams with intensity up to 1 A and $j^- \sim 10$ mA/cm^2 can be extracted from this type of SPS with emitter–extractor gap $d = 8$ cm. Cesium admixture is used for catalysis of negative-ion formation. A barium emitter surface has been used for this purpose, but with less efficiency.

A smaller version multicusp SPS has been adapted for 20-mA negative-ion–beam formation for an accelerator. This version of the SPS (Fig. 22) was optimized for heavy negative-ion production. For this application, the emitter surface must have material of heavy negative ions and heavy gas, such as Xe, Kr, or Ar, used for discharge support. Heavy negative-ion beams of up to 0.1 A in pulses and up to 0.01 A dc can be extracted. Different negative ions of elements and compounds can be produced.

Different versions of gas-discharge ion sources can be used for negative-ion production with good efficiency after any modification and addition of cesium admixture for catalysis of surface-plasma formation of negative ions.

Like an intense pulsed electron beam, an intense positive or negative pulsed ion beam can be generated by the application of a fast-rising high-voltage (near 1 MV) pulse to a diode (Bystritskiy and Didenko, 1989). For ion acceleration, a strong magnetic field is used for magnetic insulation, which inhibits the movement of electrons across the magnetic field. The first explosive emission of electrons from the cathode is initiated by electron field emission from small whiskers and evaporation of whiskers. Positive ions are emitted by the anode plasma layer, which is generated by electron bombardment of the anode. The negative-ion emission is initiated by bombardment of the cathode surface by ions from the diode gap. A prepulse is used for the generation of ions in the gap. Ion intensity up to 10^6 A with current density $(0.5–5) \times 10^3$ A/cm^2 can be reached in a pulse of 10^{-7} s.

5. CONCLUSION

The first hundred years since the discovery of the electron have witnessed a significant development in electron and ion sources. The growth of sources and beam parameters has been very active and accelerating during this time. Unexpected discoveries and inventions have opened new possibilities for improvement of electron- and ion-beam characteristics and for new applications of beams. Real nanofabrication and manipulation of single atoms are now coming into practice. The highest concentration of power densities has been achieved by use of ion and electron beams. Further development of this activity will do much to make the presently impossible become a reality.

ACKNOWLEDGMENTS

The authors would like to thank Russell Cowan and Ron McLachlan for their help in preparing this paper.

GLOSSARY

Brightness: The specific current of particles emitted by unit surface into unit solid angle.

Brillouin Zone: The polyhedron in reciprocal space over which the electron wave vector $\mathbf{k} = \mathbf{p}/\hbar$ is allowed to vary.

Calutron: A powerful plasma ion source with hot cathodes, gas discharge with electron oscillation in a strong magnetic field, and transverse extraction of the ion beam, developed by Lawrence for isotope separation.

Cathode Rays: The flux of electrons emitted from the cathode in low-pressure discharges.

Cesium Catalysis: The significant enhancement of negative-ion formation in gas discharges by the addition of a small admixture of cesium or other substances with low ionization potential, decreasing the electrode work function.

Child–Langmuir Law: The equation for space-charge–limited current density in a vacuum diode.

Cluster: A compound particle consisting of several atoms or molecules weakly bonded together.

Electron Affinity: The binding energy of the additional electron in a negative ion.

Electron Cyclotron Resonance (ECR): Resonance absorption of electromagnetic radiation by electrons in a static magnetic field when the electromagnetic frequency is equal to the cyclotron frequency of the electrons. It is a most powerful method of heating electrons to a high temperature.

Emittance: A measure of the volume in phase space occupied by a beam, indicative of the reciprocal of the quality of the beam.

Explosive Emission: The emission of a high-density electron flux from a high-density plasma layer generated by whisker explosion through field emission in a strong electric field.

Fermi Level, Fermi Energy: The value of the energy that separates occupied from empty states in the conduction zone in the limit as $T = 0$.

Fermi Surface: The surface of constant energy in electron momentum space, corresponding to the Fermi level in a crystal.

Field Emission (Autoelectronic Emission, Tunneling Emission, Cold Emission, Field Evaporation): The emission of charged particles by tunneling from the solid surface to the potential well created by a strong electric field.

Forbidden Zone (Gap): The difference in energy between the top of the occupied band and the bottom of the conduction (partially occupied or empty) band.

Ionization Potential: The smallest potential difference necessary for electron acceleration to produce atom or molecule ionization. Equal to the binding energy of an electron in the atom or molecule.

Optical Pumping: The transfer of spin polarization from circularly polarized photons to the electrons in atoms or a solid.

Oxide Cathodes: The most widely used cathodes, consisting of thick layers of oxides—BaO, CrO, or CaO—on a nickel base.

Penning Discharge: A low-pressure gas discharge with oscillation of electrons along the anode tube in a magnetic field between two cold cathodes.

Perveance: Beam current divided by $\frac{3}{2}$ power of acceleration potential; used as a measure of the ratio of current to the transverse space-charge–limited current.

Photoelectric Emission (Photoeffect): The emission of electrons initiated by the absorption of photons.

Polarized Electrons, Polarized Atoms or Ions: Particles with an increased probability of electron or nuclear spin projection along a particular direction.

Quantum Yield: The average number of photoelectrons emitted on absorption of one photon.

Schottky Effect: The increase of thermionic electron or ion emission by a strong electric field at the surface of the emitter.

Secondary Emission: The emission of electrons or ions initiated by bombardment with electrons, ions, excited or fast atoms, etc.

Surface-Plasma Sources (SPS): Ion sources with negative-ion formation by the interaction of plasma particles with electrode surfaces in gas discharges.

Thermionic Emission: The emission of electrons or ions from a solid heated to a high temperature, taking place by the "evaporation" of the particles.

Thermionic Emission Converter: A power supply for conversion of heat into electricity by emission of electrons from a cathode with higher work function and collection of them by an anode with lower work function.

Work Function: The energy difference between the vacuum energy level of an electron and the Fermi level in a solid; the evaporation energy of an electron.

Works Cited

Alton, G. D. (1993), *Nucl. Instrum. Methods B* **73**, 221–303.

Belchenko, Yu. I., Davydenko, V. I., Derevyankin, G. E., Dimov, G. I., Dudnikov, V. G., Morozov, I. I., Roslyakov, G. V., Shabalin, A. L. (1990), *Rev. Sci. Instrum.* **61**, 378–384.

Blakemore, J. S. (1985), *Solid State Physics,* Cambridge, U.K.: Cambridge Univ. Press, Chaps. 3, 4.

Brown, Ian (Ed.) (1989), *The Physics and Technology of Ion Sources,* New York: Wiley.

Bystritskiy, M., Didenko, A. N. (1989), *High Power Ion Beams,* New York: American Institute of Physics.

Chivers, D. J. (1992), *Rev. Sci. Instrum.* **63**, 2501–2506.

Drentje, A. G. (1994), *Rev. Sci. Instrum.* **65**, 1045–1049.

Dudnikov, V. G. (1992), *Rev. Sci. Instrum.* **63**, 2260–2668.

Dudnikov, V. G., Shabalin, A. L. (1992), *Rev. Sci. Instrum.* **63**, 2460–2463.

Forrester, A. T. (1988), *Large Ion Beams: Fundamentals of Generation and Propagation,* New York: Wiley.

Green, T. S. (1984), in: H. S. W. Massey, E. W. McDaniel, B. Bederson (Eds.), *Applied Atomic Collision Physics,* Vol. 2, *Plasma,* Orlando, FL: Academic, pp. 339–380.

Kutner, V. B. (1994), *Rev. Sci. Instrum.* **65**, 1039–1041.

Kuznetsov, G. I. (1994), *Nucl. Instrum. Methods A* **340**, 204–208.

Le Duff, Joel (Ed.) (1990), *Short Pulse High Current Cathodes,* Paris: Editions Frontiéres.

Orloff, J. (1993), *Rev. Sci. Instrum.* **64**, 1105–1130.

Prewett, P. D., Mair, G. L. (1991), *Focused Ion Beams from Liquid Metal Ion Sources,* New York: Wiley.

Schiemenz, P. (1992), *Rev. Sci. Instrum.* **63**, 2519–2525.

Whealton, J. H. (1981), *Nucl. Instrum. Methods Phys. Res.* **189**, 55–70.

Whealton, J. H., Bell, M. A., Raridon, R. J., Rothe, K. E., Ryan, P. M. (1988), *J. Appl. Phys.* **64**, 6210–6226.

Whealton, J. H., Meszaros, P. S., Raridon, R. J., Rothe, K. E., Bacal, M., Bruneteau, J., Devynck, P. (1989), *Rev. Phys. Appl.* **24**, 945–949.

Wolf, B. H. (1994), *Rev. Sci. Instrum.* **65**, 1248–1251.

Further Reading

Physics of Electron Emission

Blakemore, J. S. (1985), *Solid State Physics,* Cambridge, U.K.: Cambridge Univ. Press, Chaps. 3, 4.

Humphries, S. (1990), *Charged Particle Beams,* New York: Wiley, Chap. 7.

Technology of Efficient Electron Sources

Humphries, S. (1990), *Charged Particle Beams,* New York: Wiley, Chap. 7.

Kuznetsov G. I., (1994), *Nucl. Instr. Methods A* **340**, 204–208.

Le Duff, Joel, (Ed.) (1990), *Short Pulse High Current Cathodes,* Paris: Editions Frontiéres.

Physical Principles of Ion Production

Alton, G. D. (1993), *Nucl. Instrum. Methods B* **73**, 221–303.

Brown, Ian (Ed.) (1989), *The Physics and Technology of Ion Sources,* New York: Wiley.

Bystritskiy, M., Didenko, A. N. (1989), *High Power Ion Beams,* New York: American Institute of Physics.

Dudnikov, V. G. (1992), *Rev. Sci. Instrum.* **63**, 2260–2668.

Dudnikov, V. G., Shabalin, A. L. (1992), *Rev. Sci. Instrum.* **63**, 2460–2463.

Forrester, A. T. (1988), *Large Ion Beams: Fundamentals of Generation and Propagation,* New York: Wiley.

Green, T. S. (1984), in: H. S. W. Massey, E. W. McDaniel, B. Bederson (Eds.), *Applied Atomic Collision Physics,* Vol. 2, *Plasma,* Orlando, FL: Academic, pp. 339–380.

Orloff, J. (1993), *Rev. Sci. Instrum.* **64**, 1105–1130.

Prewett, P. D., Mair, G. L. (1991), *Focused Ion Beams from Liquid Metal Ion Sources,* New York: Wiley.

Schiemenz, P. (1992), *Rev. Sci. Instrum.* **63**, 2519–2525.

Actual Ion Sources

Alton, G. D. (1993), *Nucl. Instrum. Methods B* **73**, 221–303.

Belchenko, Yu. I., Davydenko, V. I., Derevyankin, G. E., Dimov, G. I., Dudnikov, V. G., Morozov, I. I., Roslyakov, G. V., Shabalin, A. L. (1990), *Rev. Sci. Instrum.* **61**, 378–384.

Chivers, D. J. (1992), *Rev. Sci. Instrum.* **63**, 2501–2506.

Drentje, A. G. (1994), *Rev. Sci. Instrum.* **65**, 1045–1049.

Green, T. S. (1984), in: H. S. W. Massey, E. W. McDaniel, B. Bederson (Eds.), *Applied Atomic Collision Physics*, Vol. 2, *Plasma*, Orlando, FL: Academic, pp. 339–380.

Kutner, V. B. (1994), *Rev. Sci. Instrum.* **65**, 1039–1041.

Sakudo, N. (1994), *Rev. Sci. Instrum.* **65**, 1284–1288.

Wolf, B. H. (1994), *Rev. Sci. Instrum.* **65**, 1248–1251.

PHYSICS AND TECHNOLOGY, HISTORY OF

David F. Channell, *The University of Texas at Dallas, Richardson, Texas, U.S.A.*

	Introduction	35
1.	**Early Interactions between Science and Technology**	36
1.1	Greek Mechanics	36
1.2	Chinese and Muslim Mechanics	37
1.3	Medieval Mechanics	37
2.	**The Early Modern Period**	38
2.1	The Renaissance	38
2.2	The Scientific Revolution	38
2.2.1	Mechanics	39
2.2.2	Pneumatics and Steam	40
2.2.3	Optics	40
3.	**The Eighteenth Century**	41
3.1	The Industrial Revolution	41
3.2	Rationalism and Experimentalism	41
3.2.1	Waterwheels	41
3.2.2	The Steam Engine	41
3.2.3	Electricity	42
3.3	Institutions	42
4.	**The Nineteenth Century**	43
4.1	The Rise of Engineering Science	43
4.2	Applied Mechanics	44
4.2.1	Strength of Materials	44
4.2.2	Structures	44
4.2.3	Machines	45
4.2.4	Hydrodynamics	45
4.3	Thermodynamics	45
4.4	Electricity and Magnetism	46
5.	**The Twentieth Century**	47
5.1	Applied Physics and Industrial Research	47
5.2	Atomic Theory and Quantum Mechanics	48
5.2.1	The Atomic Bomb	49
5.3	Postwar Developments	49
5.3.1	Nuclear Weapons and Reactors	49
5.3.2	The Space Race	49
5.3.3	Computers	50
5.3.4	Solid State	51
	Works Cited	51
	Further Reading	53

INTRODUCTION

In the modern world, science and technology have developed a symbiotic relationship with one another. It is usually assumed that all scientific discoveries will eventually lead to some new or improved technologies. It is also commonly assumed that the successful development of technology requires a knowledge of science and mathematics. There is also some awareness that developments in technology have led to new advances in science. But this continual interaction between science and technology, which has come to characterize the modern industrial world, is of relatively recent origin, dating back only some two hundred and fifty years.

Throughout most of history, science and technology followed different social and intellectual traditions and, with a few exceptions, had little impact on each other (Layton, 1974). Since classical times, science was closely associated with philosophy. As a form of natural philosophy, science was concerned with asking and answering philosophical questions such as, what is the ultimate nature of reality? With the goal of generating new knowledge about the world, science put little or no emphasis on using such knowledge to solve practical problems. Scientists used mathematical, geometrical, and other rational methods in order to idealize problems so that their solutions could be universally applied to the description of nature as a whole. Scientific knowledge, the result of contemplation, was also associated with the elite, and after the establishment of universities during the Middle Ages, most scientists had university educations. Finally, knowledge of science was gained and disseminated through a written tradition of treatises, textbooks, and journals.

Historically, technology was practical ac-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

tivity concerned with asking and answering questions about what should be done in particular situations. The goal of technology was to develop workable solutions to practical problems. Engineers traditionally used cut-and-try empiricism and rule-of-thumb techniques that, while applicable to specific situations, could not be generalized to a wide class of problems. Before the late eighteenth century, few mechanics or engineers had formalized educations or university training; technology was the province of an artisan class. Knowledge of technology, which resulted from observation and experience, was gained through an apprenticeship system and disseminated through direct contacts between master and pupil. Very little technological knowledge was written down.

During the eighteenth and nineteenth centuries, new research and pedagogy arose in such areas as applied mechanics, the strength of materials, the theory of machines, thermodynamics, heat transfer, hydrodynamics, electromagnetism, and optics. This led to a new scientific approach to technology that could not be easily classified in terms of the traditional categories of science/technology or theory/practice. By the nineteenth century, this new research and pedagogy had become institutionalized in engineering schools, technical institutes, research laboratories, and professional societies, and by the second half of the century, it was becoming more and more common to refer to this range of activities under the terms engineering science or applied physics.

1. EARLY INTERACTIONS BETWEEN SCIENCE AND TECHNOLOGY

1.1 Greek Mechanics

Although a sustained and institutionalized scientific approach to technology dates back only some two hundred and fifty years, there was a long but sporadic history of interactions between science and technology that would pave the way for the emergence of engineering science and applied physics (Clagett, 1963). While the Greeks' interest in ideal philosophies, such as that of Plato, and their dependence on slavery led them to emphasize the creation of scientific and mathematical theories over the practical development of technology, the Greeks did make some important contributions in applying their scientific knowledge to technology. Beginning as early as the fourth century B.C. the Greeks were undertaking a systematic study of mechanics based on mathematical principles. Such studies would establish a scientific basis for understanding technology. The pseudo-Aristotelian work *Mechanical Problems* (ca. 280 B.C.) described the law of the lever and used the law to explain such practical things as rowing a ship and throwing a rock from a sling. There is also evidence that Euclid provided one of the earliest geometrical proofs of the law of the lever in his *Treatise on the Balance* (ca. 300 B.C.). He also conducted some early studies on optics, dealing with the problem of perspective.

Some of the most important work in the application of science to technology was done by Archimedes, who combined an interest in mathematics with a background as a practical inventor (Drachmann, 1962). Through his mathematical studies of the lever, including the case of unequal arms, his work determining the centers of gravity of plane and solid figures, and his geometric study *On Floating Bodies* (ca. 250 B.C.), which led to the famous "principle of Archimedes" concerning the weight of submerged bodies, he provided a new understanding of the basic theory of machines by reducing statics and hydrostatics to rigorous propositional disciplines. He also conducted some studies on optics and mirrors.

During the Hellenistic period, work on applied mechanics continued (Drachmann, 1948). Ctesibius experimented with air and water pressure, leading to his inventions of a forced pump, a water organ, and a water clock. Philo of Byzantium developed empirical ballistics formulas for artillery and described an early thermoscope that could represent differences in temperature of a fluid. The most extensive treatises on applied mechanics were produced by Hero of Alexandria, who probably derived many of his ideas from Ctesibius and Philo. In his book *Mechanics* (first century A.D.), Hero analyzed the five simple machines (wheel and axle, lever, pulley, wedge, and screw) in terms of a given weight moved by a given force. Hero's book *Pneumatics* described a number of inventions and devices operated by water, air, or steam (Hall, 1973). Most of these devices were simple playthings or automata, but they were based on a rather sophisticated theory that argued that small portions of vacuum could be interspersed between small pieces of matter. In a book entitled *On the Dioptra,* Hero described a surveying instrument combining elements of the theodolite and the level, and he showed how it could be used in engineering problems such as boring a tunnel.

1.2 Chinese and Muslim Mechanics

Both the Chinese and the Muslims made important contributions to the advancement of science and technology. While in most cases the two areas were independent of one another, there were some instances where the barriers between theoretical and empirical studies were overcome. During the Warring States period in China (480–221 B.C.), the philosophy of Mo Ti led his followers to conduct experimental studies that led to empirical rules concerning the reflection of light from concave and convex mirrors and rules concerning the functioning of levers and pulleys. Chinese interest in alchemy may have contributed to the casting of iron, studies of the magnet, and the invention of paper during the Han dynasty (202 B.C.–220 A.D.), as well as the creation of gunpowder during the Thang dynasty (618–906 A.D.). Studies of astronomy undoubtedly led to Su Song's creation of one of the earliest astronomical clocks in the 1080s (Needham, 1954, 1970).

The Muslim world benefited from contacts with both Greece and China. Arab scholars in Baghdad translated and extended Greek and Hindu scientific and philosophical works and made significant advancements in the area of mathematics (Nasr, 1968). Research into pneumatics, stimulated by the translation of Hero's works, probably contributed to the development of various mechanical devices such as fountains and automata. The astronomical work of Ptolemy influenced the development of astronomical clocks about the same time as they were developed in China, althought it is unclear if there was any direct connection between the two inventions (Hill, 1984).

1.3 Medieval Mechanics

During the Middle Ages, much more emphasis was placed on technological development than had been during the period of the Greeks. But the bulk of the Greek science and mathematics that had been taken over and extended by Islamic scholars only began to become available to the Latin West after the twelfth century. With the translation of Greek scientific and philosophical texts, particularly the works of Aristotle, science began to play more of a role in technological development (Clagett, 1959). Aristotle had argued that nature and *techne* function in similar ways. In fact, he believed that nature provided the forms that technology was to imitate.

In the first half of the twelfth century, there were attempts to connect the practical arts with the seven liberal arts. Hugh of St. Victor, Director of the School of St. Victor in Paris, argued that there were seven mechanical arts—weaving, weapon foundry, navigation, agriculture, hunting, medicine, and acting—that paralleled the seven liberal arts of grammar, dialectic, rhetoric, arithmetic, geometry, astronomy, and music.

The translation of Greek science from Arabic and the establishment of universities led to a new spirit of experimentation. At Oxford, Robert Grosseteste argued that natural philosophy had to be based on mathematics and experimentation. His treatise on optics suggested that lenses could be used to magnify small objects. Roger Bacon, who was Grosseteste's successor, continued the experimental tradition. He conducted experiments on the magnifying effects of lenses and the optical properties of parabolic mirrors. His experimental studies led Bacon to suggest that telescopes, self-propelled boats, submarines, and flying machines could be constructed.

One of the leading experimenters of the thirteenth century was Pierre de Maricourt, who argued that manual skill was important in order to correct errors, which could not be done with a knowledge of natural philosophy alone (Gimpel, 1976). In his *Epistolae de Magnete,* he described an experimental study of the magnet and suggested several practical applications. This work paralleled work that had been done earlier in China, but it is unclear if Maricourt knew of such work, either directly or indirectly. Using a spherical lodestone and small iron wires, he discovered magnetic meridians, and he described the use of a compass to find directions. He also noted the laws of magnetic attraction and repulsion and used that knowledge to attempt the construction of a perpetual motion machine in which a magnetic arm would be forced to revolve in a circular path through the alternate attraction and repulsion of other magnets.

The combination of an experimental with a rational spirit could also be seen in two major technological achievements of the Middle Ages—the Gothic cathedral and the weight-driven mechanical clock. Beginning in the twelfth century with the building of St. Denis near Paris, the new Gothic style of cathedral began to spread throughout Europe (Mark, 1982). The application of rational and experimental techniques in these buildings can be seen in the notebooks of the thirteenth century engineer-architect Villard de Honnecourt, who outlined a practical geometry for stone cutting, carpentry, surveying, construction, and bridge building. During the thirteenth and fourteenth centuries, the development of a weight-driven mechanical clock, indirectly influenced by earlier Chinese and Muslim devices, by such people as Giovanni di Dondi and Richard of Wallingford reflected a desire to imitate the motion of the heavens by mechanical means (Price, 1964).

2. THE EARLY MODERN PERIOD

2.1 The Renaissance

Although there were examples of the interaction between science and technology during the medieval period, these instances were limited by the small amount of Greek science that had been translated into Latin, the dogmatic position of the medieval Catholic Church when it came to conflicts between faith and reason, and the rather rigid social distinctions that existed between the scholar and the craftsman that arose from barriers between the universities and the guild systems. During the period of the Renaissance in the fifteenth and sixteenth centuries, many of these limitations on the interaction of science and technology began to disappear. New translations of Greek texts, including the works of Plato, Euclid, Archimedes, and the Greek atomists, began to appear. The Protestant Reformation began to undermine the authority of the Catholic Church, and the barriers between the scholarly and craft traditions began to break down.

Renaissance artist-architect-engineers provided new connections between theory and practice. For example Filippo Brunelleschi, the designer of the dome on Santa Maria del Fiore in Florence, combined interests in hydraulics, mechanics, construction, gold smithing, sculpture, and perspective (Prager and Scaglia, 1970). Leon Battista Alberti, who helped to introduce fixed-point perspective into painting, developed a theory of construction that combined the use of mathematical proportion with experimentation (Ferguson, 1977). He showed that in designing bridges, floating material could be used to determine what parts of the river flowed most swiftly so that the piers could avoid those locations. Franceso di Giorgio Martini, another well-known artist, conducted experiments on guns and cannons to discover the relationship between the amount of gunpowder and the weight of the projectile (Reti, 1963). One of the most important artist-engineers was Leonardo da Vinci who used both theory and experiment in his study of engineering. He conducted experiments on small-scale structures in order to discover formulas that connected the strength of pillars and beams with their thickness and length. The recent discovery of the *Madrid Codices* of Leonardo indicates that he was attempting a systematic study of machines by distinguishing and analyzing the function of individual mechanisms within different types of machines, rather than treating each different machine as unique (Reti, 1974).

2.2 The Scientific Revolution

During the period of the Scientific Revolution in the sixteenth and seventeenth centuries, new scientific theories and discoveries led to further interactions between science and technology (Hall, 1983; Jacob, 1988; Lindberg and Westman, 1990). In universities such as Padua, the medical faculty came into a position of dominance over the divinity faculty, and with this change more attention was paid to the scientific works of Aristotle rather than his logical works. This led to a new interest in observation, experimentation, and practical applications of science. On the other hand, the translation of Plato during the fifteenth century contributed to a new interest in the development of mathematical theories of science and technology. These two different approaches were brought together in the emergence of a magical or Hermetic approach to nature (Yates, 1979). Although rejected by most scientists and engineers today, Hermeticism strongly influenced many people such as Copernicus, Kepler, and Newton. This approach to nature led to the view that knowledge in general, and science in particular, was a tool that could be used to manipulate the forces of nature for some practical purpose. For example, in his preface to the first English translation of Euclid, John Dee, a possible model for Shakespeare's Prospero, argued that science and such practical activities as architecture, hydraulics, mechanics, and navigation were all governed by principles of geometry, which he saw as a form of mathematical magic (Dee, 1975). Giambattista della Porta, author of *Natural Magick* (1658), saw magic as explaining both scientific and practical activities. His book included discussions of astrology, alchemy, distillations, optics, metallurgy, magnetism, and mechanics.

The recognition of a new relationship between science and technology was put forward in the seventeenth century by Francis Bacon, who argued that the goal of science was not just knowledge, but knowledge that would allow humans to exert power and control over nature (Rossi, 1970). Bacon foresaw a new technological form of science. In his *New Atlantis* (1628), he described an institution called Solomon's House in which group research was conducted and aimed at improving engines, machines, cannons, clocks, ships, and other elements of technology. Although Bacon's goal of a technological science was not realized in his lifetime, his view influenced the founding of the Royal Society of London in 1662, and it foreshadowed the emergence of the industrial research laboratory in the twentieth century.

While some new theories emphasized how science could be applied to practice, other theories focused on how technology could be used to explain science. In the sixteenth century, beginning with René Descartes and continuing with Pierre Gassendi, Thomas Hobbes, Robert Boyle, G.W. Leibniz, and Isaac Newton, there emerged a mechanical philosophy in which all of nature was seen to function like a gigantic machine or clockwork mechanism. This encouraged scientists to turn their attention to technology if they were to understand how the natural world functioned (Dijksterhius, 1961).

2.2.1 Mechanics These new mathematical-experimental theories led to some significant interactions between science and technology (Drake and Drabkin, 1969). One of the most important was in the area of mechanics. During the sixteenth century, following on the work of some artist-engineers, Nicola Tartaglia showed how impetus and gravity acted together on a projectile causing it to follow a curved path. He was also able to show, through a mixture of experiment and mathematics, that the maximum range of a cannon would take place when it was elevated at an angle of 45°. Another leading contributor to the science of mechanics was Simon Stevin, who was one of the earliest people to discover the resultant action of two noncollinear or nonparallel forces. He also contributed to a new understanding in ship building by extending Archimedes's principle of buoyancy, showing that the center of gravity of a floating body was in the same line as the center of gravity of the displaced fluid.

One of the leading figures in the creation of modern mechanics and its connection to the practical arts was Galileo Galilei, who studied at the University of Padua and gained practical experience in the Venetian Arsenal. Although best known for his support of the Copernican system and his subsequent trial, Galileo (1633) also helped to establish modern mechanics in his work *Discourses on Two New Sciences*. Galileo began this work with a discussion of the scale effect by showing that if a machine were doubled in size by simply doubling all of its dimensions, it was likely to fail since its weight would increase eight times while the strength of the individual parts would increase only four times. He continued with a discussion of how to calculate the load that a beam fixed at one end could support, if loaded at the other end. Following Tartaglia, Galileo was able to show that a projectile fired from a cannon would follow a parabolic path.

Galileo also conducted experimental and mathematical studies of the pendulum and suggested that a clock regulated by a pendulum would be more precise than clocks regulated by the verge and foliot. A few years later, Christian Huygens, working with a clockmaker, constructed the first pendulum clock. Through his work on the pendulum, Huygens was able to deduce the law of centripetal force. Such a law played a significant role in Isaac Newton's formulation of a universal theory of gravitation (Westfall, 1980), in his *Principia Mathematica* (1687). Although Newton's work was primarily theoretical, his formulation of mechanics, based on his mathematical definition of force, provided the basis for a rational study of technology during the eighteenth century.

2.2.2 Pneumatics and Steam A second area in which there was a significant interaction between science and technology during the Scientific Revolution was pneumatics, which was stimulated by a renewed interest in the work of Hero of Alexandria and the practical problem of pumping water from mines. Miners had noted for some time that there was a limit to the distance that water could be raised by suction. Such figures as Giambattista della Porta, Giovanni Branca, and Salomon de Caus wrote books on pneumatics describing pumps, fountains, and steam-powered devices. In the middle of the seventeenth century, Gasparo Berti, Evangelista Torricelli, and Blaise Pascal conducted experiments with columns of water and mercury that established that the weight of the atmosphere exerted a pressure that could support a column of mercury. About the same time, Otto von Guericke in Germany began a series of experiments using his newly invented air pump to create a vacuum (Guericke, 1962). By creating a vacuum in a cylinder fitted with a piston, he was able to show that the pressure of the atmosphere could do mechanical work by moving the piston to the bottom of the cylinder.

Guericke's work influenced work in England where Robert Boyle and Robert Hooke made improvements on the air pump and used it to determine that the pressure of a given amount of air varied inversely with its volume. In Paris, the Dutch scientist Christian Huygens and his assistant Denis Papin constructed an apparatus similar to von Guericke's, but instead of an air pump it used a small amount of gunpowder in an attempt to force air out of the cylinder. Later Papin moved to England where he served for a time as Boyle's assistant. While in England, Papin experimented with steam and suggested that condensing steam could be used to create a vacuum in a cylinder. Using a small apparatus he was able to demonstrate that steam at high pressure could be used to raise a piston while condensing steam could lower it again. By the beginning of the eighteenth century these ideas would come together in the invention of the steam engine.

2.2.3 Optics The third area in which there was a significant interaction of science and technology during the Scientific Revolution was in the field of optics (Lindberg, 1976). Renaissance artists' use of perspective had encouraged a study of optical phenomena. At the beginning of the seventeenth century lens makers in Holland, such as Hans Lippershey and Zacharias Jansen, began to develop more powerful instruments by combining lenses in order to form telescopes and microscopes. The telescope became famous as an astronomical instrument after Galileo used it to observe the sun and planets providing him with empirical evidence to support the Copernican theory. Galileo also used a microscope to study insects. This work was followed by Robert Hooke in England and by Antoni Leeuwenhoek in Holland.

The invention of the telescope and microscope led to new theoretical and experimental studies in the area of optics. Johann Kepler, who designed some telescopes, conducted experimental studies on the laws of refraction. A mathematical formulation of the law of refraction was put forward by Willebrord Snell in 1621. During the seventeenth century, different theories were put forward concerning the nature of light. Descartes believed that light consisted of small particles, while Huygens put forward a wave theory of light. The experimental and theoretical studies on light culminated in the work of Isaac Newton and the publication of his *Opticks* (1704). Defects with telescopes led him to begin a study of optics that resulted in his design of a reflecting telescope and later to experiments with prisms in which he was able to resolve white light into the spectrum. Newton adopted a corpuscular theory of light, but he also maintained some aspects of a wave theory by assuming that moving particles of light could cause vibrations in the surrounding aether.

3. THE EIGHTEENTH CENTURY

3.1 The Industrial Revolution

During the eighteenth century, there were significant advances in the development of technology, particularly in Great Britain, that have been called the Industrial Revolution. The production of iron using coke rather than charcoal, the invention of the steam engine, the mechanization of the production of textiles, the expansion of the network of canals, and the centralization of production in the factory system dramatically transformed eighteenth-century society. Although most of the technological advances that contributed to the Industrial Revolution owed little dependence to science, there were an increasing number of interactions between science and technology that set the stage for a formalized relationship between the two areas.

Developments that had taken place during the Scientific Revolution did have some impact on eighteenth-century technology. At the very end of the seventeenth century, Thomas Savery of Dartmouth, England, designed a steam-powered pump based on an idea put forward by della Porta. The so-called miner's friend used condensing steam to create a vacuum that raised water from a mine by suction. By 1712, Thomas Newcomen, an English ironmonger and blacksmith, brought together the ideas of von Guericke, Huygens, and Papin to construct the first workable steam engine, which used condensing steam to raise and lower a piston (Cardwell, 1963). There is still considerable debate concerning how much Newcomen's invention owed to science, but it does seem possible that he knew of Papin's work and that he was in correspondence with Robert Hooke.

3.2 Rationalism and Experimentalism

The greatest impact of science on eighteenth-century technology was in the emergence of a rationalistic–experimental approach that drew heavily on the scientific method. In France, Antoine Parent developed a theory of the efficiency of machines through the use of Galilean mechanics, and he was one of the first to apply differential calculus to engineering problems (Parent, 1700). Although his application of his theory to an analysis of waterwheels contained several errors, it stimulated a new approach to the study of machines.

3.2.1 Waterwheels The new approach to machines was continued in England by John Smeaton, who has been called by some the first modern engineer. He was influenced by Newton's *Principia* and combined its rules of reasoning with an experimental approach to technology. Smeaton's most famous work was his study of waterwheels. With the increasing demands for water power brought about by the growth of the textile industry, there was a debate concerning whether the overshot or undershot waterwheel was more efficient (Reynolds, 1983). During the middle of the century, Smeaton constructed a small model waterwheel that could function as either an overshot or an undershot wheel. By varying the speed of the water and the speed of the wheel, he was able to show that an overshot wheel was the most efficient.

At the same time, several scientists, such as Jean and Daniel Bernoulli and Jean d'Alembert, began to develop a theory of hydrostatics and hydrodynamics and applied these theories to practical problems. Bernard Forest de Bélidor, in his widely influential *Architecture hydraulique* (1737–1753), attempted to calculate the efficiency of waterwheels but wrongly concluded that undershot wheels were more efficient. Not long after, Johann Albrecht Euler theoretically established, through the use of rational mechanics and the use of *vis viva* (mv^2), that the overshot wheel was more efficient than the undershot (Euler, 1754). Many of these same scientists, such as Daniel and Jean Bernoulli, Pierre Bouger, Henri-Louis Duhamel du Monceau and Leonhard Euler applied mechanics and higher mathematics to the study of ship design, propulsion, and stability (Prichard, 1987).

3.2.2 The Steam Engine The new rationalistic–experimental approach to technology also led to significant improvements in the steam engine during the second half of the eighteenth century (Cardwell, 1971). After his work on the waterwheel, Smeaton conducted an experimental study of the Newcomen engine by building small model engines in which he could vary different factors such as the size of the cylinder, the size of the boiler, and the number of strokes per minute. Through these tests, he was able to improve dramatically the efficiency of the engine.

The most significant improvements in the steam engine were made by James Watt. As instrument maker at the University of Glasgow, he began a study of the inefficiencies of a small model Newcomen engine. Through a series of experiments on the engine and on the properties of steam, he realized that for the engine to work efficiently the cylinder had to be hot when the steam entered and cold when the steam was condensed. It was not efficient for one cylinder to be at two temperatures but if a second cylinder were added, a separate condenser, the engine would be much more efficient (Channell, 1974). There has been much speculation that Joseph Black's concept of latent heat was crucial in the improvement of the steam engine, but Watt, although a friend of Black's, denied that his work played any important role. Most historians now agree that it was Watt's broad knowledge of the laws of heat and his use of the scientific method, rather than the concept of latent heat, that led to the improvement of the steam engine.

3.2.3 Electricity The design of new devices and instruments led to scientific and technological advances in the area of electricity (Heilbron, 1979). In the seventeenth century, Otto von Guericke constructed a revolving sulfur globe that would build up a static electrical charge when it rubbed against a dry hand or piece of cloth. Static electrical machines were improved during the eighteenth century, using disks of glass revolving between carpetlike material. The study of static electricity advanced when the Dutch scientist Pieter van Musschenbroek invented the Leyden jar, which served as a condenser for static electricity.

Using these new devices, Benjamin Franklin, in colonial America, began a series of electrical experiments that led him to reject the old notion that there were two different types of electrical fluid and to argue instead for a single fluid in which an excess of fluid would result in a positive charge and a deficit of fluid would result in a negative charge (the exact opposite of the modern definition). One of Franklin's most significant accomplishments was his famous kite experiment in 1752 that demonstrated that atmospheric lightning was an electrical phenomenon. This knowledge led Franklin to the important invention of the lightning rod, which was able to protect structures against destruction from lightning.

During the middle of the eighteenth century John Mitchell at Cambridge used a magnet suspended from a string to show that magnetic repulsion obeyed an inverse square law. A few years later Joseph Priestley concluded that electrical force followed an inverse square law, on the basis of the fact that a charge inside a cavity was not affected by an electrical charge on the outside of the hollowed body. Between 1785 and 1789 Charles A. de Coulomb used a torsion balance to show that the attractive and repulsive forces of electricity and magnetism followed an inverse square law and were similar to Newton's gravitational force.

3.3 Institutions

One of the most important developments in the interaction between science and technology was the emergence of the idea that a scientific approach to technology could be institutionalized in a new form of scientific and technical education (Artz, 1966). As early as the seventeenth century, Jean-Baptiste Colbert, chief finance minister to Louis XIV, argued that French manufacturers and mercantilists would benefit from educational reform. At a time when higher education emphasized theology and the classics, Colbert proposed establishing a number of academies to teach the scientific and mathematical basis of such practical subjects as road and bridge building, fortification, design, navigation, and manufacturing. Although few of his plans were actually realized during his lifetime, the establishment of the *Académie des Sciences* being one of the few exceptions, his ideal continued and came to fruition in the eighteenth century.

Throughout Europe during the eighteenth century, government involvement in mercantile and military projects led to the establishment of new schools to provide engineers for the state. In France, the *Corps des Ponts et*

Chaussées was created in 1716 and led to the establishment of the *École des Ponts et Chaussées* in 1747, as a school to train engineers for the Corps (Brunot and Coquand, 1982). At about the same time, military schools at La Fère and Mézières were established as was the *École des Mines,* all of which taught engineering subjects. In Hungary, the state helped to create a mining school at Schemnitz in 1735, which later became the Royal Hungarian Mining Academy, while in Saxony the *Bergakademie* at Freiberg was established in 1765. These schools combined practical studies in technology with courses on chemistry and geology, Freiberg becoming particularly well-known because of the geology lectures of Abraham Werner. Although England lagged behind the Continent in establishing engineering education in the eighteenth century, John Smeaton, concerned over the developments in France, helped to form the Society of Civil Engineers in 1771 in order to give some formal structure to the training of engineers. Also in England, industrialists and scientists, such as James Watt and Joseph Priestley, came together in the Lunar Society of Birmingham in order to exchange ideas (Schofield, 1963).

These institutions reflected an attempt to develop a new scientific approach to technology. In France, Bernard Forest de Bélidor, a professor at the artillery school at La Fère, published a book entitled *La science de ingénieures* (1729), which was notable as one of the first uses of the term engineering science. It was in the engineering schools of France that some of the first elements of what would be called engineering science emerged. Jean Rodolphe Perronet, director of the the *École des Ponts et Chaussées,* pioneered a scientific approach to bridge building by designing a testing machine in order to determine the strength of stone for use in bridges (Perronet, 1782–1789). A short time later, Charles de Coulomb, a graduate of the military school at Mézières, used mathematical studies of maxima and minima to solve problems of bending beams, the construction of vaults, and the pressure against earthwork walls (Heyman, 1972). These experimental and theoretical studies culminated in Pierre-Simon Girard's *Traité analytique de la résistance des solides* (1798), which was one of the first texts on the subject. Also at the same time the mathematician Gaspard Monge, while teaching at Mézières, developed his system of descriptive geometry, which allowed for the graphic solution to a wide range of engineering problems (Monge, 1799). Many of the graduates of these engineering schools, such as Jean Charles Borda, Lazare Carnot, and J. V. Poncelet, would make significant contributions to science and technology during the nineteenth century.

4. THE NINETEENTH CENTURY

4.1 The Rise of Engineering Science

Although an important and institutionalized interaction between science and technology began to emerge in the second half of the eighteenth century, it did not have a significant impact until the nineteenth century. Much of the technology that contributed to the Industrial Revolution was, for the most part, based on traditional skills and craftsmanship rather than on engineering science. But the new technologies, and the increased scale of traditional technologies, that resulted from the Industrial Revolution helped to establish engineering science as a necessary part of scientific and technological development, and in the process helped to bring about its autonomy from pure science (Layton, 1971).

With the development of the steam engine, railroads, ocean-going iron-hulled ships, large-scale suspension bridges, and the electrical telegraph, it became impractical and uneconomical for engineers to rely on traditional rule of thumb or trial and error. At the same time scientists, who were interested in the practical applications of science, were beginning to discover—what most engineers had already known—that many of the laws of science were not directly applicable to technology. Newtonian mechanics might explain the forces acting between two atoms but it did not help to determine how an iron beam might act under a complex load. Boyle's law explained the relationship between pressure and volume in an ideal gas but was of little use in describing how steam acted in a working steam engine. The Bernoulli equation or the Navier-Stokes equations of classical hydrodynamics had limited application in describing real fluids undergoing nonlaminar flow. Finally, Ohm's law or Maxwell's equations, in their scientific form, were of limited use in the study of real electrical circuits, telegraph cables, and dynamos.

One of the most important developments in the emergence of engineering science took place at the end of the eighteenth century with the establishment in Paris of the *École polytechnique* in 1794–1795 (Fourcy, 1987; Shinn, 1980). The school, designed to train both military and civil engineers, represented a new recognition that these two branches of engineering rested on the same principles. At the *École polytechnique,* all areas of technology were based on a core curriculum of geometry, trigonometry, physics, chemistry, mechanics, laboratory, and shop work. Some of the most famous mathematicians, scientists, and engineers, including J. L. Lagrange, Simon Laplace, A. L. Cauchy, G. Coriolis, J. M. C. Duhamel, G. Lamé, G. Monge, C. L. M. H. Navier, S. Poisson, J. V. Poncelet, G. de Prony, and Lazare Carnot, taught at the school. Through their work, a theoretical approach to technology emerged based on mathematical techniques and a scientific methodology that could be formalized in written treatises and articles.

The *École polytechnique,* by providing a new model for scientific and engineering education, helped to spread a scientific approach to technology to other countries. The school's organization and curriculum influenced Austrian polytechnics at Vienna and Prague, and German polytechnic schools at Karlsruhe, Munich, Dresden, Stuttgart, and Hanover. In the United States, the Military Academy at West Point and Rensselaer Polytechnic Institute, two of the earliest American engineering schools, introduced methods and textbooks that had originated at the *École polytechnique.* Many of the leading engineer-scientists of the nineteenth century, including Ferdinand Redtenbacher, Otto Mohr, Franz Grashof, Friedrich Clebsch, August Föppl, and Ludwig Prandtl, were associated with these polytechnics.

4.2 Applied Mechanics

4.2.1 Strength of Materials

During the nineteenth century, the development of engineering science led to significant new interactions between science and technology. In the area of applied mechanics, a great deal of new work was done in the strength of materials (Timoshenko, 1953; Todhunter, 1886–1893). In England, Peter Barlow and Thomas Tredgold conducted experimental tests on timber. The introduction of cast and wrought iron as building materials stimulated more experimental studies. In order to test iron, new machines capable of larger loads had to be designed. In 1833, the Royal Navy installed a one-hundred-ton machine at the dockyard at Woolwich that was used to test iron chains. Barlow used this machine to conduct tests on iron rails. About the same time, William Fairbairn and Eaton Hodgkinson constructed a testing machine to do compression tests on full-size cast-iron pillars. On the Continent, Johann Bauschinger created one of the earliest university materials testing laboratories at the Polytechnic Institute of Munich. While in America, Robert Thurston founded laboratories at Stevens Institute of Technology and Cornell University.

While material testing was undergoing development in England, Germany, and America, there were important advancements in the theory of elasticity, especially in France. The work of Augustin Cauchy, Gabriel Lamé, C. L. M. H. Navier, Siméon Poisson, and Adhemar Barré de Saint-Venant led to a new theoretical understanding of elasticity. Particularly important was Cauchy's formulation in 1822 of the concept of stress. In Germany, Johann Bauschinger and Hermann Zimmermann extended the theory with the application of graphical methods to stress analysis.

4.2.2 Structures

Applied mechanics also led to new experimental and theoretical analyses of structures, particularly in response to demands created by the construction of railroad bridges and suspension bridges (Billington, 1983). In England, William Fairbairn, Eaton Hodgkinson, and Robert Stephenson conducted experimental studies that contributed to the design of the Conway and Britannia tubular railway bridges (Vincenti and Rosenberg, 1978), while W. J. M. Rankine and James Clerk Maxwell developed theoretical techniques of parallel projection and reciprocal figures, which allowed new ways to study the forces acting in a structure (Maxwell, 1864). In

America, Squire Whipple discovered graphical and mathematical methods for the analysis of truss bridges, which could not be analyzed by other methods. In France, C. L. M. H. Navier published a mathematical study of suspension bridges while Jean V. Poncelet developed a mathematical study of the stability of revetments and their foundations.

4.2.3 Machines The development of applied mechanics also resulted in new theoretical approaches to the study of machines. At the end of the eighteenth century and beginning of the nineteenth century, Lazare Carnot, C. L. M. H. Navier, J. V. Poncelet, and Baron Prony generalized the concept of *vis viva* (mv^2) and applied it to the study of machines. Later, G. G. Coriolis, William Whewell, and Henry Moseley employed the idea of work to treat a machine as a device in which transmitted work was divided into useful and wasteful work. This functional approach to a machine provided a new method to measure efficiency.

During the nineteenth century, a theory of machines began to focus on a study of the kinematics of mechanisms (Ferguson, 1962; Hartenberg and Denavit, 1964). Jean Nicolas Pierre Hâchette, Agustin de Betancourt, and Phillipe Lanz, using a program outlined by Gaspard Monge, created a system of classification of mechanism based on the conversion of one type of motion into another. André-Marie Ampère extended the work of Monge and Lanz and coined the term *cinématique* (Ampère, 1838–1843). By the middle of the century, Robert Willis, Jacksonian Professor at Cambridge, put forward a theory based on the relationship of motions created by a mechanism. By the second half of the century, Franz Reuleaux, director of the Royal Technical School at Charlottenburg, further extended the theory by moving away from isolated mechanisms and toward the idea of mechanisms as part of an integrated system in which a mechanism could be inverted into another mechanism (Reuleaux, 1895). Particularly important work was done in 1868 by James Clerk Maxwell in his analysis of mechanical governors, which led to the development of an early theory of feedback control (Maxwell, 1868; Mayr, 1970).

4.2.4 Hydrodynamics During the nineteenth century, applied mechanics led to new advances in hydrodynamics (Rouse and Ince, 1957). In the first half of the century, C. L. M. H. Navier and G. G. Stokes put forward mathematical theories on the motion of fluids. In 1842, Stokes introduced the concept of lines of motion, which was developed by W. J. M. Rankine in 1870 into a theory of stream lines. During the second half of the century, hydrodynamic theories were extended by Hermann von Helmholtz, Lord Kelvin, Lord Rayleigh, and Osborne Reynolds. Many of these theories led to improvements in naval architecture and hydraulic machines. At mid-century, John Scott Russell conducted experimental and theoretical studies on bodies moving through a liquid and concluded that objects shaped like a wave offered the least resistance (Emmerson, 1977). At the same time, William Froude and Edward James Reed conducted studies on the stability of ships. In 1834, Benoît Fourneyron, influenced by the work of Lazare Carnot, developed the Fourneyron water turbine. At mid-century, James B. Francis conducted experiments at Lowell, Massachusetts, that led to a set of design principles for water turbines (Layton, 1979).

4.3 Thermodynamics

The emergence of engineering science also led to significant interactions between science and technology in the area of thermodynamics (Cardwell, 1971). After the invention of the steam engine in the eighteenth century, there was a great deal of interest in the nature of heat. At the end of the eighteenth century, while superintendent at an arsenal in Munich, an American, Benjamin Thompson (Count Rumford), raised questions concerning the prevailing caloric theory of heat by showing that more heat was generated while boring a cannon with a blunt tool than with a sharp one. Although both Humphrey Davy and Thomas Young put forward mechanical theories of heat in the early part of the nineteenth century, such theories were not widely accepted until the second half of the century. But even the caloric theory led to significant developments. Jean Fourier in his *Théorie analytique de la chaleur* (1822) used the idea of caloric as the basis for a new mathematical analysis concerning the flow of heat through a material body. Sadi Carnot, an army engineer and son of Lazare, published his *Réflexions sur la* *puissance motrice du feu* (1824), in which he used an analogy between a waterwheel extracting work from falling water and a heat engine extracting work from heat falling from a higher to a lower temperature to analyze the factors that determined the production of work in a heat engine, and he put forward what has come to be known as the Carnot cycle. After Carnot's death, Emile Clapeyron rediscovered Watt's idea of an indicator (or *P-V*) diagram and introduced it into Carnot's theory (Clapeyron, 1834).

During the 1840s, several scientists, including Robert Mayer, James Prescott Joule, and Hermann von Helmholtz, began theoretical and experimental studies based on a mechanical theory of heat. This work led to the formulation of the conservation of energy (Elkana, 1974). The mechanical theory of heat was applied to the theory of heat engines by Rudolf Clausius in Germany, and Lord Kelvin and W. J. M. Rankine in Scotland. Through a study of transformations between heat energy and mechanical energy they were led to the concept of entropy and formulations of the second law of thermodynamics. During the second half of the century, the first and second laws of thermodynamics were interpreted in terms of an atomistic, kinetic theory of heat by Clausius, James Clerk Maxwell, and Ludwig Boltzmann. In order to treat the large number of particles in a gas, Maxwell and Boltzmann resorted to the use of probabilities, establishing the foundations for the creation of statistical mechanics.

Although most of the developments in thermodynamics in the nineteenth century were associated with the steam engine, there were advances associated with other heat engines. Both Joule and Rankine conducted studies of hot-air engines, a few of which were built by John Ericsson. In 1860, Étienne Lenoir constructed the first commercially successful engine powered by illuminating gas. This led to a new interest in the application of thermodynamics to internal combustion engines (Bryant, 1967). In 1862, Alphonse Beau de Rochas showed that air must be compressed to gain maximum efficiency and he set out the conditions that would become the four-stroke cycle. Although he did not know of Beau de Rochas's work, Nicolaus Otto produced a workable four-stroke engine in 1876 with the help of Franz Reuleaux's insight into the importance of compression. At about the same time, Rudolf Diesel, while a student at the Munich Technische Hochschule, realized that an engine could increase its efficiency if there was a very high compression at the beginning of the cycle (Thomas, 1987). Soon after, he published a book on his "rational heat-engine," but it was not until 1897 that he was able to build a workable engine.

4.4 Electricity and Magnetism

Some of the most significant developments in the interaction between science and technology took place in the area of electricity and magnetism. The discovery of current electricity by Luigi Galvani and Alessandro Volta at the end of the eighteenth century led to a new range of electrical experimentation and with it new scientific discoveries concerning the nature of electricity and magnetism. In 1820, Hans Christian Oersted, who had been searching for a connection between electricity and magnetism, discovered that an electric current generated a magnetic effect. Only a week after Oersted's discovery of electromagnetism, André-Marie Ampère suggested the possibility of creating an electrical telegraph. During the 1820s and 1830s, several scientists, including Pavel Schilling in Russia, Carl Gauss and Wilhelm Weber in Germany, Charles Wheatstone and William Cooke in England, and Joseph Henry in America, worked on various electrical telegraphs. The most successful telegraph was the system invented by Samuel F. B. Morse, whose code made his system more practical than many of the others. During the second half of the century, the idea of the telegraph was extended by Elisha Gray, who showed how several messages could be sent through one wire, by Alexander Graham Bell (Bruce, 1973), who showed how electricity could carry the human voice, thus inventing the telephone, and by Oliver Heaviside, who developed a theory that led to long-distance transmissions.

Oersted's discovery of electromagnetism led researchers to search for the possibility that magnetism could produce electrical effects. In 1831, Michael Faraday, in England, discovered that moving a coil of wires in the presence of a magnet could induct an electrical current in the coil (Williams, 1971). By the second half of the century, Faraday's discovery of electromagnetic induction served as the basis for the creation of electrical dynamos by William Siemens in 1856 and Z. T. Gramme in 1867. Early in the century, Humphrey Davy had discovered both arc and incandescent lighting, but dependence on batteries limited their use. During the 1870s and 1880s, new dynamos made possible Charles Brush's practical development of arc lighting and Thomas Edison's invention of incandescent lighting (Friedel and Isreal, 1986). Edison's development of an entire electrical lighting system at his laboratory in Menlo Park, New Jersey, reflected a new approach to invention—one based on systematic, planned research by an interdisciplinary team of scientists and engineers. His work would provide one model for the emergence of the industrial research laboratory in the twentieth century. The invention of the dynamo also led to improvements in the electrical motor. Both Michael Faraday and Joseph Henry built small electrical motors as part of their research, and several inventors succeeded in building direct-current motors (Molella, 1976). The development of electrical motors advanced in 1888 with Nikola Tesla's proposal for a motor powered by alternating current.

One of the most significant developments in the area of electricity and magnetism was the discovery that light was an electromagnetic wave (Buchwald, 1989). At the beginning of the nineteenth century, Thomas Young put forward a wave theory of light. This theory raised serious questions concerning the properties of the medium that was undergoing vibration, and different models were put forward by Augustin Fresnel, Siméon Poisson, George Stokes, James MacCullagh, Lord Kelvin, George Green, and Augustin Cauchy. The most successful approach began in 1846, when Michael Faraday suggested that light might be a vibration of electric and magnetic lines of force, on the basis of his experimental discovery that the plane of polarization of light could be affected by a strong magnetic field. Faraday's work led James Clerk Maxwell to put forward an electromagnetic theory of light in 1862. Using a mechanical model of the aether, Maxwell derived a set of laws that governed electricity and magnetism. By 1865, he abandoned his mechanical model and focused his attention on a set of equations that expressed how electric currents could give rise to magnetic fields and how changing magnetic fields could give rise to electrical fields.

Maxwell's theory predicted the existence of electromagnetic waves that would travel at the speed of light. Such waves were experimentally detected by Heinrich Hertz in 1888 (Buchwald, 1994). Between 1894 and 1896, Guglielmo Marconi, using some ideas of Oliver Lodge, demonstrated that Hertz's waves could be used to transmit wireless signals. By the beginning of the twentieth century, the creation of radio and Edison's observation of the "Edison effect" in the 1880s led to John Ambrose Fleming's invention of the diode and Lee DeForest's invention of the triode.

5. THE TWENTIETH CENTURY

5.1 Applied Physics and Industrial Research

At the end of the nineteenth century and beginning of the twentieth century, the interactions between science and technology brought about by the rise of engineering science contributed to what some people have called a second industrial revolution (Mumford, 1934). Unlike the first industrial revolution, this one was heavily dependent upon scientific advances, especially advances in physics. It was during the nineteenth century, particularly after 1870, that physics began to emerge as an autonomous discipline with its own journals, university departments, and research institutes (Jungnickel and McCormmach, 1986). More and more by the beginning of the twentieth century this new physics was becoming associated with industrial concerns.

One of the most important developments in the emergence of applied physics was the establishment of industrial multidisciplinary research laboratories, based on the idea of group research (Hughes, 1989). Although some industrial research evolved in the German chemical industry during the 1880s, it was not multidisciplinary and made limited use of group research. In America, Thomas Edison's Menlo Park Laboratory provided one model for industrial research. Here physicists, chemists, mathematicians, and engineers were brought together in teams to form an "invention factory." Another model for industrial research emerged at the beginning of the twentieth century in science-based companies such as American Bell Telephone and the General Electric Company. While the industrial research laboratories established by these companies had some similarities to Edison's laboratory, they emerged at a time when some of their fundamental patents were expiring, and their primary purpose was to maintain and defend the companies' dominance over a particular technology. Although the original intent of these laboratories was to block competition, they conducted a great deal of fundamental research, and in exploring all possible routes a competitor might follow, they made many discoveries that led to new commercial products. For example, the Bell Laboratories' focus on the control of long-distance telephony led George Campbell to develop the wave filter and Harold Arnold to perfect the triode (Brittain, 1970). The General Electric Laboratory's focus on the light bulb led to William Coolidge's development of the tungsten filament and development of the x-ray tube, and later to Irving Langmuir's Nobel Prize–winning work on atomic physics (Wise, 1985).

By the period of World War II, a great deal of applied physics research was being done in government-sponsored research laboratories. Although some work, such as aerodynamic studies of airplanes and chemical research for gas warfare, had been conducted during World War I, it was World War II that brought about what Dwight Eisenhower called the military-industrial complex. Particularly in the United States, Great Britain, Germany, and the Soviet Union, university, industrial, and governmental laboratories conducted wartime applied research directed and sponsored by the military. This work led to such developments as radar, sonar, computers, jet aircraft, rockets, operations research, and the atomic bomb. With the end of World War II, this model of research continued into the era of the Cold War and the Space Race, and still plays a significant role in applied research today.

5.2 Atomic Theory and Quantum Mechanics

During the twentieth century, the emergence of applied physics in university, industrial, and governmental research laboratories led to significant new developments in science and technology. Much of this work was based on the revolution that was taking place in physics at the end of the nineteenth century and beginning of the twentieth century. Of particular importance was the new model of the atom that emerged during this period. Since the time of the Greeks, it had been assumed that atoms were the smallest units of matter, indivisible, immutable, and eternal. But in the 1890s, experimental discoveries began to challenge this view. Scientists discovered that electricity applied to an evacuated tube produced cathode rays. In 1895, W. Roentgen discovered that when cathode rays struck a metal target, other rays, termed x rays, were produced. In 1897, J. J. Thomson proved that cathode rays were a stream of particles (later called electrons) whose weight was less than that of the smallest atom. At the same time, other scientists, including Henri Becquerel and Pierre and Marie Curie, discovered that certain elements, such as uranium and radium, emitted a type of radiation that indicated the change of an atom of one chemical substance into another.

At the same time that questions were being raised about the billiard-ball model of matter, scientists were also discovering problems with the idea of continuous fields. The difficulty centered on describing the nature of blackbody radiation (Kuhn, 1978). Theories predicted an infinite amount of radiation in the ultraviolet region, while experiments had shown radiation with a maximum wavelength below the ultraviolet region. The solution to this "ultraviolet catastrophe" was suggested by Max Planck in 1899, when he argued that radiation was not emitted and absorbed continuously but in discrete packets of energy, or quanta. In 1905, Albert Einstein extended Planck's theory and introduced the idea that light consisted of a stream of photons with their own energy and momentum.

During the first decades of the twentieth century, a new model of the atom was formulated. Ernest Rutherford developed a model of the atom in which electrons orbited around a nucleus (Wilson, 1983). His student, Niels Bohr, used quantum theory to extend the model by arguing that electrons were confined to certain discrete orbits (Moore, 1985). The Bohr model provided scientists with a new understanding of chemical combinations and the atomic spectrum of hydrogen, but it could not explain other spectra or the relative intensities of spectral lines. In 1925, Werner Heisenberg suggested abandoning mechanical models and using instead a matrix mechanics (Van der Waerden, 1967). About the same time, Erwin Schrödinger used Louis de Broglie's idea that matter could have wavelike properties to formulate a wave mechanics, which was shown to be equivalent to matrix mechanics. Heisenberg's and Schrödinger's formulation of quantum mechanics combined with the Bohr model of the atom provided the basis for some important new technological advances.

5.2.1 The Atomic Bomb One of the most significant developments of the twentieth century was the creation of nuclear weapons (Rhodes, 1986). The new atomic theory predicted that a great deal of energy was stored in the nucleus of an atom and might be released if a heavy atom was split into two smaller atoms. With the discovery of the neutron by James Chadwick in 1932, physicists had a possible way to split the atom. The next year, Leo Szilard speculated on the possibility that a chain reaction could release significant amounts of energy. In December of 1938, Otto Hahn and Fritz Strassmann in Germany conducted an experiment in which uranium nuclei were split by bombarding them with neutrons.

With the beginning of World War II, America, Great Britain, Germany, and the Soviet Union (Holloway, 1994) all began research in an attempt to exploit nuclear energy. By 1940, Otto Frisch and Rudolf Peierls in England had calculated that the critical mass needed to sustain a chain reaction was small enough to consider building a bomb. In December of 1942, a team of scientists led by Enrico Fermi at the University of Chicago succeeded in creating the first sustained nuclear reaction. During the next few years, the American atomic bomb project focused on the engineering problems associated with separating uranium-235 from uranium-238 and producing plutonium-239 from uranium-238 in order to create enough weapons-grade materials. Through the development of an electromagnetic separation by Eastman Kodak, gaseous diffusion by Union Carbide, and nuclear reactors by Dupont and Westinghouse, enough material was available to test the world's first atomic bomb on 16 July 1945 and for the bombs used a few weeks later against Japan.

5.3 Postwar Developments

5.3.1 Nuclear Weapons and Reactors After World War II, research continued on nuclear energy. Further development of nuclear weapons culminated in 1951–1952 with the creation of hydrogen bombs under the leadership of Stanislaw Ulam and Edward Teller in the United States and Andrei Sakharov in the Soviet Union. About the same time, work continued on the development of nuclear reactors for military and civilian purposes. Several experimental reactors were built during the late 1940s in the United States, England, France, and the Soviet Union. In 1946, Captain Hyman Rickover of the U.S. Navy began development of a nuclear reactor to power a submarine, in conjunction with Argonne National Laboratory, Westinghouse, and General Electric (Hewlett and Duncan, 1974). By the early 1950s, the nuclear-powered submarine Nautilus was launched and the first civilian nuclear power plant, using Rickover's designs, was built at Shippingport, Pennsylvania.

5.3.2 The Space Race World War II also stimulated interactions between science and technology in the area of aeronautics and space research (McDougall, 1985). While some work on rockets was done before World War II by Konstantin Tsiolovsky and Sergi Korolev in Russia and Robert H. Goddard in America, it was during the war that Germans under Wernher von Braun did significant research and development of rockets resulting in the V-1 and the first ballistic missile, the V-2. After World War II, the United States and the Soviet Union placed a great deal of effort into the development of missile technology both for military purposes and for the scientific exploration of space. Although Stalin's purges almost destroyed the Soviet rocket program, Korolev's and Valentin Glushko's work allowed the Soviets to launch the first earth-orbiting satellite, Sputnik I, on 4 October 1957. This was followed quickly by Sputnik 2, which placed a dog into orbit in November of 1957; Lunik I, which missed the moon but discovered the solar wind; Lunik II, which landed on the moon; and Lunik III, which photographed the dark side of the moon. The Soviet program culminated on 12 April 1959 with Yuri Gagarin becoming the first man to orbit the earth.

In America, Wernher von Braun, who was brought from Germany after the war, developed the Army's Redstone rocket and played a leading role in America's space program including the launching of Explorer I, the first U.S. satellite; the Mercury program, which led to John Glenn's being the first American to orbit the earth; and the Apollo program, which culminated with Neil Armstrong's and Edwin (Buzz) Aldrin's moon landing in July of 1969. At the same time, America sent several unmanned spacecraft on missions to Mars, Venus, and eventually Jupiter, Saturn, and Neptune, opening the era of the scientific exploration of space. These developments provided scientists with new ways to conduct both astronomical and geological observations and experiments, and at the same time they led to significant technological developments in the areas of communication and navigation.

5.3.3 Computers Another area of interaction between science and technology that was stimulated by World War II was the development of the computer (Goldstine, 1972). While some elements of the computer date back as far as the Greeks, and later several mathematicians, including Blaise Pascal and G. W. Leibniz, designed mechanical calculating machines, the basic aspects of the modern computer, including the ideas of a memory, central processing unit, and ability to be programmed, were put forward in the nineteenth century by the English mathematician Charles Babbage, but he was never able to build such a machine. By the end of the nineteenth century, Herman Hollerith designed a tabulating machine, incorporating elements of Boolean algebra, to process data from the 1890 U.S. Census, but these early machines were limited by mechanical problems and were not general programmable computers. The modern computer developed in response to the need to calculate ballistic equations that resulted from the introduction of much more powerful guns during World War I, and to the need to break enemy codes during World War II.

By the beginning of World War II, several projects were under way to develop a computer to carry out scientific calculations and to break the German Enigma code. One of the first successes, the Mark I, was created by a group at Harvard University and IBM led by Howard Aiken in 1944. Although it was electrically powered, it performed calculations through a series of essentially mechanical telephone relay switches. As early as 1937, John V. Atanasoff began work on a machine that used vacuum tubes, but other war duties prevented him from completing his work. The first truly electronic computer, known as ENIAC (Electronic Numerical Integrator and Computer) and using eighteen thousand radio tube switches, was created by a group under J. Presper Eckert and John Mauchly in 1946 at the University of Pennsylvania.

While computer hardware was being developed, there were also significant advancements in software. In 1936, the British mathematician Alan Turing had described an ideal computer known as a Turing machine. A year later, the American Claude Shannon demonstrated that a set of electronic switches could be used to express logical statements using Boolean Algebra. Beginning in 1945, a group, including mathematician John von Neumann, attempted to design an improved version of the ENIAC, called the EDVAC (Electronic Discrete Variable Calculator), by incorporating some of the ideas of Turing and Shannon. In his *First Draft of a Report on the EDVAC* von Neumann outlined the design for a general all-purpose computer, what has come to be called a von Neumann machine. The major advance of the EDVAC was its ability to store a program along with data. This allowed the computer to do a new set of tasks without having to rewire circuits physically as had been the case with the ENIAC, and it allowed the computer to modify its own program. The creation of the general programmable computer opened up research in new areas including the development of cybernetics by Norbert Wiener in 1948, the development of artificial intelligence by such people as John McCarthy, Marvin Minsky, Herbert Simon, and Allen Newell, the idea of parallel processing and neural networks by Warren McCullock, Walter Pitts, Donald Hebb, Frank Rosenblatt, Seymour Papert, and W. Daniel

Hillis, and the creation of artificial life by Fred Cohen and J. Doyne Farmer (Pratt, 1987). The computer has also revolutionized many fields of science by its ability to bridge the gap between theory and experiment with the concept of computer modeling.

5.3.4 Solid State The advancement of computer science and technology was stimulated by the postwar revolution in solid-state technology. On the basis of research that involved linkages between quantum mechanics and electrical engineering, William Shockley, Walter Brattain, and John Bardeen invented the transistor at Bell Laboratories on 19 December 1947 (Weiner, 1973). The success of the transistor led to further miniaturization when Jack Kilby of Texas Instruments and Robert Noyce of Fairchild Electronics each developed an integrated circuit in the late 1950s.

At the same time, the interaction between quantum mechanics and electrical engineering, stimulated by the scientific and military interest in short-wave radiation, led Joseph Weber and Charles Townes in the United States and N. G. Brasov and A. M. Prokhorov in Russia to conceive of the idea of microwave amplification by stimulated emission of radiation. By 1954, Townes and James P. Gordon had constructed a functioning ammonia-beam maser at Columbia University. This success led to the development of solid-state masers by a group at Bell Laboratories. These devices found almost immediate applications in the areas of military radar, communications, and radio astronomy.

By 1957, Townes at Columbia and Arthur Schawlow at Bell Laboratories had the idea of creating a similar device for light. At the same time, R. Gordon Gould, a Columbia graduate student, followed a somewhat independent line of research, which also led to the idea of a laser. In 1960, Theodore H. Maiman, of the Hughes Research Laboratories, created the first operating laser. The invention of the laser led to significant research into spectroscopy, chemistry, and semiconductors, and to a wide range of commercial applications in such areas as communications, material processing, and fusion (Bromberg, 1991). As with other postwar developments in the interaction of science and technology, the solid-state revolution reflected a symbiotic relationship in which knowledge was flowing as much from technology to science as it was flowing from science to technology.

Works Cited

Ampère, A.-M. (1838–1843), *Essai sur la Philosophie des Sciences,* Paris: Bachelier.

Artz, F. B. (1966), *The Development of Technical Education in France, 1500–1850,* Cambridge, MA: MIT Press.

Bacon, F. (1628), *The New Atlantis,* London.

Bélidor, B. F. (1729), *La science des Ingénieurs,* Paris: C. Jombert.

Bélidor, B. F. (1737–1753), *Architecture Hydraulique,* 4 Vols., Paris: C. A. Jombert.

Billington, D. P. (1983), *The Tower and the Bridge: The New Art of Structural Engineering,* New York: Basic Books.

Brittain, J. E. (1970), *Technol. Culture* **11**, 36–57.

Bromberg, J. (1991), *The Laser in America, 1950–1970,* Cambridge, MA: MIT Press.

Bruce, R. V. (1973), *Bell: Alexander Graham Bell and the Conquest of Solitude,* Boston: Little, Brown.

Brunot, A., Coquand, R. (1982), *Le Corps des Ponts et Chaussées,* Paris: Centre National de la Recherche Scientifique.

Bryant, L. (1967), *Technol. Culture* **8**, 178–198.

Buchwald, J. (1989), *The Rise of a Wave Theory of Light,* Chicago: Univ. of Chicago Press.

Buchwald, J. (1994), *The Creation of Scientific Effects: Heinrich Hertz and Electrical Waves,* Chicago: Univ. of Chicago Press.

Cardwell, D. S. L. (1963), *Steam Power in the Eighteenth Century: A Case Study in the Application of Science,* London: Sheed & Ward.

Cardwell, D. S. L. (1971), *From Watt to Clausius: The Rise of Thermodynamics in the Early Industrial Age,* Ithaca: Cornell Univ. Press.

Carnot, S. N. L. (1824), *Réflexions sur la Puissance Motrice du Feu,* Paris: Bachelier; translation (1986), *Reflexions on the Motive Power of Fire: A Critical Edition with the Surviving Manuscripts,* Manchester: Manchester Univ. Press.

Channell, D. F. (1974), *Am. J. Phys.* **42**, 65–67.

Clagett, M. (1959), *The Science of Mechanics in the Middle Ages,* Madison: Univ. of Wisconsin Press.

Clagett, M. (1963), *Greek Science in Antiquity,* New York: Collier.

Clapeyron, E. (1834), *J. École Polytech.* **14**, 153–190.

Dee, J. (1975), *The Mathematical Preface to Euclid (1570),* New York: Science History Publ.

Dijksterhius, E. J. (1961), *The Mechanization of the World Picture,* Oxford: Oxford Univ. Press.

Drachmann, A. G. (1948), *Ktesibios, Philon and Heron: A Study in Ancient Pneumatics*, Copenhagen: Ejnar Munksgaard.

Drachmann, A. G. (1962), *The Mechanical Technology of Greek and Roman Antiquity*, Madison: Univ. of Wisconsin Press.

Drake, S., Drabkin, I. E. (Eds.) (1969), *Mechanics in Sixteenth Century Italy*, Madison: Univ. of Wisconsin Press.

Elkana, Y. (1974), *The Discovery of the Conservation of Energy*, Cambridge, MA: Harvard Univ. Press.

Emmerson, G. S. (1977), *John Scott Russell: A Great Victorian Engineer and Naval Architect*, London: John Murray.

Euler, J. A. (1754), *Enadatio Qvaestionis*, Göttingen: Svmtibvs Dan. Frid. Kvbleri.

Ferguson, E. S. (1962), *U.S. Nat. Museum Bull.* No. 228.

Ferguson, E. S. (1977), *Science* **197**, 827–836.

Fourcy, A. (1987), *Histoire de L'École Polytechnique*, Paris: Belin.

Fourier, J. (1822), *Théorie Analytique de la Chaleur*, Paris: F. Didot.

Friedel, R., Israel, P. (1986), *Edison's Electric Light: Biography of an Invention*, New Brunswick, NJ: Rutgers Univ. Press.

Galilei, G. (1638), *Discorsi e Dimostrazioni Matematiche, Intorno a due Nuoue Scienze*, Leyden: Elzevir; translation (1954), *Two New Sciences*, New York: Dover.

Gimpel, J. (1976), *The Medieval Machine: The Industrial Revolution of the Middle Ages*, New York: Holt, Rinehart & Winston.

Girard, P. (1798), *Traité Analytique de la Résistance des Solides*, Paris: F. Didot.

Goldstine, H. (1972), *The Computer: From Pascal to von Neumann*, Princeton, NJ: Princeton Univ. Press.

Guericke, O. (1962), *Experimenta Nova (1672)*, Aalen, Germany: O. Zellar.

Hall, A. R. (1983), *The Revolution in Science, 1500–1750*, 3rd ed., London: Longman.

Hall, M. B. (Ed.) (1973), *The "Pneumatics" of Hero of Alexandria: A Facsimile of the 1851 Woodcroft Edition*, New York: American Elsevier.

Hartenberg, R., Denavit, J. (1964), *Kinematic Synthesis of Linkages*, New York: McGraw-Hill.

Heilbron, J. L. (1979), *Electricity in the 17th and 18th Centuries: A Study of Early Modern Physics*, Berkeley, CA: University of California Press.

Hewlett, R., Duncan, F. (1974), *Nuclear Navy, 1946–1962*, Chicago: University of Chicago Press.

Heyman, J. (1972), *Coulomb's Memoir on Statics: An Essay in the History of Civil Engineering*, Cambridge, U.K.: Cambridge Univ. Press.

Hill, D. (1984), *A History of Engineering in Medieval and Classical Times*, La Salle, IL: Open Court.

Holloway, D. (1994), *Stalin and the Bomb*, New Haven, CT: Yale Univ. Press.

Hughes, T. P. (1989), *American Genesis*, New York: Viking.

Jacob, M. (1988), *The Cultural Meaning of the Scientific Revolution*, Philadelphia, PA: Temple Univ. Press.

Jungnickel, C., McCormmach, R. (1986), *Intellectual Mastery of Nature*, 2 vols., Chicago: Univ. of Chicago Press.

Kuhn, T. S. (1978), *Black-Body Theory and the Quantum Discontinuity, 1894–1912*, Chicago: Univ. of Chicago Press.

Layton, E. T. (1971), *Technol. Culture* **12**, 562–580.

Layton, E. T. (1974), *Technol. Culture* **15**, 31–41.

Layton, E. T. (1979), *Technol. Culture* **20**, 64–89.

Lindberg, D. (1976), *Theories of Vision from al-Kindi to Kepler*, Chicago: Univ. of Chicago Press.

Lindberg, D., Westerman, R. (1990), *Reappraisals of the Scientific Revolution*, Cambridge, U.K.: Cambridge Univ. Press.

McDougall, W. (1985), *The Heavens and the Earth: A Political History of the Space Age*, New York: Basic Books.

Mark, R. (1982), *Experiments in Gothic Structures*, Cambridge, MA: MIT Press.

Maxwell, J. C. (1864), *Philos. Mag.* **27**, 250–261.

Maxwell, J. C. (1868), *Proc. R. Soc. London* **16**, 270.

Mayr, O. (1970), *The Origins of Feedback Control*, Cambridge, MA: MIT Press.

Molella, A. P. (1976), *Proc. IEEE* **64**, 1273–1278.

Monge, G. (1799), *Géométrie Descriptive*, Paris: Baudouin.

Moore, R. (1985), *Niels Bohr: The Man, His Science, and the World They Changed*, Cambridge, MA: MIT Press.

Mumford, L. (1934), *Technics and Civilization*, New York: Harcourt Brace.

Nasr, S. (1968), *Science and Civilization in Islam*, Cambridge, MA: Harvard Univ. Press.

Needham, J. (1954–), *Science and Civilization in China*, Cambridge, U.K.: Cambridge Univ. Press.

Needham, J. (1970), *Clerks and Craftsmen in China and the West*, Cambridge, U.K.: Cambridge Univ. Press.

Newton, I. (1687), *Philosophiae Naturalis Principia Mathamatica*, London: Streater; translation (1971), *Mathematical Principles of Natural Philosophy*, Berkeley, CA: Univ. of California Press.

Newton, I. (1704), *Opticks*, London: Smith & Walford; reprint (1952), New York: Dover.

Parent, A. (1700), *Éléments de Mécanique et de Physique*, Paris: F. & P. Delaulne.

Perronet, J. R. (1782–1789), *Description des*

Projets et de la Construction des Ponts de Neuilly, 2 vols., Paris: L'Imprimérie royale.

Porta, G. (1658), *Natural Magick,* London: Thomas Young.

Prager, F., Scaglia, G. (1970), *Brunelleschi: Studies of His Technology and Inventions,* Cambridge, MA: MIT Press.

Pratt, V. (1987), *Thinking Machines: The Evolution of Artificial Intelligence,* Oxford: Blackwell.

Price, D. (1964), *Technol. Culture* **5**, 9–23.

Prichard, J. (1987), *Technol. Culture* **28**, 1–25.

Reti, L. (1963), *Technol. Culture* **4**, 287–298.

Reti, L. (Ed.) (1974), *The Unknown Leonardo,* New York: McGraw-Hill.

Reuleaux, F. (1895), *The Kinematics of Machinery: Outlines of a Theory of Machines,* London: Macmillan.

Reynolds, T. (1983), *Stronger than a Hundred Men: A History of the Vertical Water Wheel,* Baltimore: Johns Hopkins Univ. Press.

Rhodes, R. (1986), *The Making of the Atomic Bomb,* New York: Simon & Schuster.

Rossi, P. (1970), *Philosophy, Technology and the Arts in the Early Modern Era,* New York: Harper & Row.

Rouse, H., Ince, S. (1957), *History of Hydraulics,* Ames, IA: Iowa Institute of Hydraulic Research.

Schofield, R. E. (1963), *The Lunar Society of Birmingham,* Oxford: Oxford Univ. Press.

Shinn, T. (1980), *Savoir Scientifique et Pouvoir Social: L'École Polytechnique, 1794–1914,* Paris: Fondation National des Sciences Politiques.

Thomas, D. E. (1987), *Diesel: Technology and Society in Industrial Germany,* Tuscaloosa, AL: Univ. of Alabama Press.

Timoshenko, S. (1953), *History of the Strength of Materials,* New York: McGraw-Hill.

Todhunter, I. (1886–1893), *A History of the Theory of Elasticity and of the Strength of Materials from Galileo to the Present,* 2 vols., Cambridge, U.K.: Cambridge Univ. Press.

Van der Waerden, B. L. (Ed.) (1967), *Sources of Quantum Mechanics,* New York: Dover.

Vincenti, W., Rosenberg, N. (1978), *The Britannia Bridge: The Generation and Diffusion of Technical Knowledge,* Cambridge, MA: MIT Press.

Weiner, C. (1973), *IEEE Spectrum* **10** (1), 24–33.

Westfall, R. S. (1980), *Never at Rest: A Biography of Isaac Newton,* Cambridge, U.K.: Cambridge Univ. Press.

Williams, L. P. (1971), *Michael Faraday: A Biography,* New York: Simon & Schuster.

Wilson, D. (1983), *Rutherford: Simple Genius,* Cambridge, MA: MIT Press.

Wise, G. (1985), *Willis R. Whitney and the Origins of U.S. Industrial Research,* New York: Columbia Univ. Press.

Yates, F. (1979), *Giordano Bruno and the Hermetic Tradition,* Chicago: Univ. of Chicago Press.

Further Reading

Brush, S., Belloni, L. (1983), *The History of Modern Physics: An International Bibliography,* New York: Garland Publishing.

Burke, J. G. (Ed.) (1987), *Science and Culture,* Scottsdale, AZ: Gorsuch Scarisbrick.

Cardwell, D. S. L. (1995), *The Norton History of Technology,* New York: W. W. Norton.

Channell, D. F. (1989), *The History of Engineering Science: An Annotated Bibliography,* New York: Garland Publishing.

Finn, B. (1991), *The History of Electrical Technology: An Annotated Bibliography,* New York: Garland Publishing.

Home, R. W., Gittins, M. (1984), *The History of Classical Physics: A Selected and Annotated Bibliography,* New York: Garland Publishing.

Kevles, D. (1979), *The Physicists,* New York: Vintage.

Krohn, W., Layton, E., Weingart, P. (Eds.) (1978), *The Dynamics of Science and Technology,* Dordrecht: D. Reidel.

Laudan, R. (Ed.) (1984), *The Nature of Technological Knowledge: Are Models of Scientific Change Relevant?,* Boston: D. Reidel.

Olby, R. C., Cantor, G. N., Christie, J. R. R., Hodge, M. J. S. (Eds.) (1990), *Companion to the History of Modern Science,* London: Routledge.

Olson, R. (1982), *Science Deified and Science Defied: The Historical Significance of Science in Western Culture from the Bronze Age to the Beginnings of the Modern Era,* Berkeley, CA: Univ. of California Press.

Pacey, A. (1992), *The Maze of Ingenuity,* 2nd ed., Cambridge, MA: MIT Press.

PHYSICS APPLICATIONS OF COMPUTERS—EXPERIMENTAL

DAVID A. KNAPP, *Lawrence Livermore National Laboratory, Livermore, California, U.S.A.*

Introduction 55
1. **Computers and the Real World** 55
2. **Devices and Equipment** 56
3. **Computers in Experiments** . 57
3.1 Experimental Design 57
3.2 Experimental Control and Data Acquisition 57
3.2.1 Logged Values 57
3.2.2 Wave Forms 58
3.2.3 Scans 58
3.2.4 Histograms 58
3.2.5 Event Data 58
3.3 Data Reduction 58
3.4 Data Analysis 58
3.5 Presentation of Results 59
4. **Computer Requirements for Experimental Applications** .. 59
4.1 Hardware Requirements 59
4.1.1 Processing Speed 59
4.1.2 Input/Output Speed 59
4.1.3 Memory 59
4.1.4 Mass Storage 59
4.1.5 Display and Printing Capabilities 60
4.2 Software Requirements 60
4.2.1 Operating System 60
4.2.2 User Interface 60
4.2.3 Speed 60
4.2.4 Flexibility 60
4.2.5 Extensibility 61
5. **Evaluation of Computer Use in Experimental Physics** 61
5.1 Advantages 61
5.2 Disadvantages 61
5.3 Example 62
6. **Future Trends** 66
Works Cited 66

INTRODUCTION

Experimental physics is heavily dependent on computers. It would be difficult to conceive of many modern experiments without computers, and computers are an indispensable and integral part of most experimental apparatus in physics. They are used in every stage of the experiment. Computer systems used for these applications range from simple microprocessors embedded in "smart instruments," to distributed processors able to acquire data at high speed, to supercomputers used to model experimental results.

It is nearly impossible to give a comprehensive picture of the myriad ways computers are used in physics experiments. Here the basic concepts involved in these applications are presented, and examples of their use in real-life experimental situations given, concentrating primarily on the role of computers in the experimental process as opposed to the details of the hardware and software used.

A typical physics experiment goes through several phases: physics design, apparatus design, construction, data acquisition, data reduction, data analysis, and, finally, presentation of results. In most experiments, computers are used in every phase. Although the same could be said for many other sciences, in physics there is a particularly close relationship between physical results and mathematical models, which has resulted in a central role for computers in experiments.

1. COMPUTERS AND THE REAL WORLD

The physical sciences describe the world in mathematical terms. Computers are designed to perform mathematical operations, and so it is natural that they are used to model and measure the physical world. Computers allow physicists to obtain quantitative values that would be impossible to calculate

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

using traditional mathematical techniques. This capability has had a profound impact on physics. The precision to which values can be calculated using computers has resulted in much more stringent tests of theories.

Computers have also allowed experimental physicists to measure the real world more accurately and precisely. Computer-based data-acquisition systems allow experimentalists to run complicated experimental setups and to store extremely detailed data.

Despite the successes of computers in modeling and measuring the real world, there are important differences between computer calculations and reality (see, for example, Kent, 1978). One such difference is the distinction between *continuous* real-world values and *discrete* representations them.

Many physical quantities such as voltages, positions, times, etc., can take on any real value. A device that can produce any real value as the result of an operation is known as an *analog device* (Goldstine, 1972, pp. 140–147). Amplifiers, integrators, and filters are examples of analog devices. Analog computing devices exist; one example is a position computer that takes several inputs from a position-sensitive detector and produces voltages corresponding the position of a detected particle. However, analog computers are not flexible; they are usually designed for a single calculation. The vast majority of computing devices in use today are *binary digital computers*. These computers perform operations on numbers represented in base 2. This scheme allows general-purpose computational devices and much simpler input and output of results; however, the finite number of base-2 digits, or bits, used means that not every real value can be represented. For most practical purposes, this difference is negligible, but in some applications it can be critical.

Another difference between computer models of the real world and reality is the limited storage and speed of a computer. Real-world processes may be very complex, making their simulation impractical without simplifying assumptions. These assumptions often reduce the accuracy of a model in order to make it computationally tractable. Thus, the results of computer models and calculations often do not exactly correspond to reality.

2. DEVICES AND EQUIPMENT

The overwhelming majority of computers used in physics today are binary digital computers (Dasgupta, 1989). Most digital computers are built around at least one *central processing unit* (CPU). A CPU performs basic arithmetic and logical operations on binary data. It communicates with the rest of the computer via a *data bus*, an electronic path over which it sends and receives data. The data bus connects the CPU to memory and other devices, such as mass storage, keyboards, displays, and external communications (for more information, see COMPUTER HARDWARE).

Four primary computer configurations are used by experimental physicists today: single-primary-user *workstations*, multiple-user *mainframe computers*, *embedded processors*, and *distributed processors*. By far the majority of computers in use today are workstations, such as personal microcomputers or graphics workstations (Grubb and Borchers, 1991). The popularity of multiple-user mainframe computers has waned as the hardware costs of computers have plummeted in recent years; nonetheless, this configuration is still useful for applications requiring supercomputing power or shared resources, such as long-term storage. Many of the functions of traditional mainframe computers are being displaced by networks, in which individual workstations communicate and share resources.

Embedded processors are dedicated computers in instruments that control the instrument and allow communication with the outside world. These processors are usually not available for general-purpose computing, but they are an important part of many physics experiments.

More complex setups, such as high-energy physics detectors, may use a system of distributed processors to process huge amounts of data efficiently and quickly. These systems often use special-purpose processors to evaluate and preprocess the data in order to send only necessary information to a coordinating computer system.

An important computer configuration used by physicists is the wide-area network. Many experimental groups have members from widely scattered institutions. The use of a network (such as the Internet) for communication enables these groups to function efficiently, pass data back and forth, and even run experiments remotely. This use of computers is presently widespread in high-energy physics, and is becoming more so in other disciplines.

Many physics experiments require data communication between a computer and external measurement devices. These devices are frequently accessed by use of a standard communications bus, such as CAMAC, VME bus, Fastbus, or IEEE-488 (also known as the General Purpose Interface Bus or GPIB) (see BUS SYSTEMS AND COMPUTER INTERFACING). A wide array of devices for experimental control and data acquisition is available for these standard buses.

3. COMPUTERS IN EXPERIMENTS

A typical experiment goes through a variety of stages, from experimental design to presentation of results. Today, computers are commonly used in every stage (Matey, 1993).

3.1 Experimental Design

The first phase of any experiment is the design of the experiment itself, which usually incorporates several parts. First, the scientific problem the experiment is meant to address must be considered. Once an experimental method to address the scientific problem is devised, the apparatus is designed and an experimental protocol set. Specific predictions about the outcome of the experiment under various circumstances are often needed before it ever begins.

The physics design of any experiment requires a quantitative understanding of the physical effect the experimentalist wishes to study. While many of the necessary predictions may be calculated with theoretical models using computers (see SIMULATION AND MODELING USING COMPUTERS), experimental design involves a great deal more than predictions from a specific theory. For example, the strength and quality of the observed signal from the actual detectors must often be predicted, as must the noise or background signal in the same detectors. In this stage, the experimentalist may also use computer-based techniques such as *Monte-Carlo modeling* (see MONTE-CARLO METHODS) to design the experiment to optimize the statistical significance of the results.

Once the conceptual design of the experiment is complete, the apparatus itself must be designed. Here computers also play an important role, as they are widely used in computer-aided design (CAD), engineering (CAE), and machining (CAM) applications. In addition, computers are used extensively in the detailed design of the physical characteristics of the apparatus. For example, experimentalists may use a computer to calculate exact electric and magnetic fields in an experimental setup.

3.2 Experimental Control and Data Acquisition

Computers lend themselves ideally to the control of instruments, and to the task of collecting data for experiments. They are extensively used in both these roles in most physics experiments. Experimental physicists are able to draw upon a considerable industry in computer control and data-acquisition products for both these tasks. Computers control many devices in a typical physics laboratory such as power supplies, frequency synthesizers, stepping motors, valves, switches, timers, vacuum pumps, lasers, and, of course, other electronic devices. Using such devices, the computer can be programmed to operate virtually the entire experimental apparatus.

Today, most experimental data are acquired via computers. The most ubiquitous devices employed for this purpose are *analog-to-digital converters* (ADCs), which convert an analog voltage into a digital number, *time-to-digital converters* (TDCs), which convert a time interval into a digital number, and *scalers,* which count electronic pulses. These basic devices can be packaged in a variety of instruments, and communicate to the computer using a variety of communication standards. Experimental data often fall into one of the following categories.

3.2.1 Logged Values *Logged data* represent values that are recorded periodically, with a low recording frequency and fairly loose requirements on the exact times the values are recorded. For example, a record of the temperature of a device taken every few minutes would be logged data.

3.2.2 Wave Forms *Wave-form data* consist of values recorded at regular, precise intervals, generally at a high frequency. The result is a digitized representation of the variation of the values with time. A digital oscilloscope is the most common form of wave-form digitizer.

3.2.3 Scans A *scan* is formed by varying one parameter and recording the value of another. For example, varying the current to a device and recording its temperature at each current value would create a data scan.

3.2.4 Histograms A *histogram* is a plot of the frequency of detection of a signal as a function of some attribute of that signal. One common use of the histogram is in spectroscopy; an energy scale is divided into a number of bins, and the value of each bin represents the number of photons detected with the corresponding energy.

3.2.5 Event Data *Event data* consist of records of measurements taken by a number of instruments, often triggered by the detection of a signal in one or more of them. Each set of associated measurements is called an *event*. For example, a high-energy physics experiment may have many detectors looking at a target. When a particle hits the target, it gives off a shower of particles from the reaction in the target. The time of arrival and energy of each detected particle would comprise an event. The recording of individual events allows the data to be reconstructed and analyzed in detail after the experiment is over. It is especially useful for complex multiparameter experiments.

The above data types are not mutually exclusive; for example, a multichannel scaler, which records the number of detections observed as a function of time, is a mixture of wave-form and histogrammed data.

3.3 Data Reduction

The widespread use of computers in experiments has coincided with an explosion in the volume of data produced. These data must be managed and reduced to a comprehensible form during a phase of the experiment often called *data reduction*. Depending on the experiment, data reduction may occur at the same time as data acquisition, later during data analysis, or both. In the data-reduction phase, raw numbers collected during data acquisition are put into a form from which the data analysis may be performed. At this point of the experiment, the data are also usually archived in a computer-readable format for future reference.

A typical data-reduction task is the combination of spectra taken with different detectors into a single spectrum representing the experimental observations. Another might be the adjustment of several data sets to some common process so they can be directly compared. In the case of event data, data reduction usually involves calculating physically meaningful values from the data and applying constraints or *cuts* to the events to select only events of interest. For example, in a high-energy physics experiment, the paths of the detected particles might be reconstructed, and only those paths that converge on a selected point chosen for further analysis.

These data-reduction processes would be essentially impossible without modern computers, as they consist of huge numbers of repetitive tasks.

3.4 Data Analysis

Once the data have been reduced to a manageable size, the analysis begins. It is difficult to imagine the analysis and interpretation of any nontrivial data set today without the use of computers.

The final results of physics experiments are almost always quantitative values for physical parameters. Of those experiments that do not provide such parameters, most report either the observation or the nonobservation of a physical effect. Either way, the results of physics experiments must be presented in a quantitative manner. Computers are frequently indispensable tools in the determination of the results from the experimental data.

Data analysis takes as many forms as there are experimental designs. However, a number of techniques are widely used. Statistical determination of physical parameters from experimental data via *curve fitting,* in which the data are compared with a theoretical function, and *direct statistical estimation,* in which the values of parameters are directly determined from the statistical properties of the measurements, are two such techniques. Such analyses often require a large number of calculations, and are almost invariably performed by computer. Another technique often used in data analysis is the *Fourier transform,* by which the frequency components of a wave form are calculated. Yet another commonly used analysis technique is *Monte-Carlo* modeling. In this process, the results of a complicated experiment are modeled by simulating the measurements using random-number generators to determine the effects of complex physical processes. The modeled results produced via a number of assumptions are then compared with the actual experimental results.

3.5 Presentation of Results

The final step in the life of an experiment is the presentation of the results obtained from the data analysis. Physicists report these results mostly via published papers and talks. Computers are widely used in the preparation of both, and are beginning to be used in their presentation as well.

Many results are presented in graphical form; today, essentially all charts and graphs are computer generated. Papers are usually written using computer-based word processors, and can be submitted to journals electronically. Indeed, figures that once were hand-drawn for journals are now regularly generated on commonly available desktop printers.

4. COMPUTER REQUIREMENTS FOR EXPERIMENTAL APPLICATIONS

4.1 Hardware Requirements

The computer hardware requirements for all physics applications have generally placed physicists among the class of so-called "power users" who use all the computer power they can obtain and constantly demand more. This generalization applies to both experimental and theoretical physicists, although the specific hardware requirements are somewhat different for the two cases.

Some of the many computer hardware considerations for experimental physicists are the following.

4.1.1 Processing Speed The speed with which problems can be solved and calculations performed is, naturally, a prime consideration for experimental applications of computers. The complexity of calculations that can be performed in a reasonable time frame is directly related to the processing speed. In Monte-Carlo techniques, for example, the precision of the result is determined by the number of events that are simulated. Experimental physicists have an apparently insatiable appetite for computer processing speed; the greater the speed of the computer, the more data that can be analyzed and the more detailed the analysis can be.

4.1.2 Input/Output Speed Data-acquisition and experimental control applications often require very high data input and output rates. For example, a wave-form may be digitized at 500 MHz, or 2 ns per measurement, with each measurement consisting of an 8-bit number, for a total data rate of 4 billion bits per second. Typical data-transfer rates today in general-purpose computers are a few hundred million bits per second, lower by about a factor of 10 than the requirements for such a wave-form. Moving the data to and from the computer is often the primary bottleneck in the data-acquisition speed.

4.1.3 Memory As the number of data samples increases, the memory required to hold them also increases. Thus, the memory requirements for experimental applications increase as the available computing power and input/output speed increase. Additionally, event recording allows many-dimensional representations of the data, and calculations and displays of these representations require large amounts of available memory.

4.1.4 Mass Storage The considerations for mass storage are similar to those for memory. As the volume of data that can be processed increases, the requirements for mass storage to archive the data also increase. However, mass storage also involves a different consideration: that of retrieval speed for data.

Traditionally (that is, until a few years ago) physicists stored data primarily on magnetic tapes. This technology allowed relatively cheap, high-capacity storage, but the time required to access an individual piece of data could be extremely long. Tape technology has improved greatly, but new technologies, such as magnetic and optical disk drives, have arisen to challenge it. The time required to access data stored on an optical disk is several orders of magnitude smaller than that required to access data on a tape, and the time to access data on a magnetic disk is smaller still. However, new technologies allow tapes to record data at a much lower cost than optical disks. Tapes are primarily useful for voluminous data that do not need to be accessed in a random order, such as archival storage, while more compact data that need to be frequently accessed are stored on disks.

4.1.5 Display and Printing Capabilities The increasing complexity of physical data and the increasing use of computers for presentation of the data have resulted in increased demands upon the display and printing capabilities of computers. For example, the complexity of representations of data and results has led to an increased use of color in displays and printed articles. The display capabilities required for experimental applications, both resolution and color, have likewise increased.

4.2 Software Requirements

The evaluation of software requirements for experimental physicists is more difficult than that of hardware requirements. Many high-quality commercial programs exist that perform a wide variety of tasks; however, the market for experimental physics applications, which frequently require specialized features that can be found in no commercial product, is small, and many physicists must write their own programs to obtain the optimization and flexibility required.

Nonetheless, here is an enumeration of some of the software considerations that must be evaluated by the experimental physicist.

4.2.1 Operating System The *operating system* of a computer is the software that allows other programs access to the computer's resources. It determines the specific software packages that the experimentalist can use, because presently most software applications run under only one operating system. All hardware systems offer a limited selection of operating systems; many offer only one.

The determination of the best operating system to use for a specific application can be complex. For example, an experimentalist may want to make use of a multitasking operating system that can run more than one process at a time. Programming such a system for a real-time application, such as data acquisition, may be very difficult, however. Today many operating systems have *graphical user interfaces,* which produce more consistent displays to the user and are considered easier to use than *text-based interfaces.*

Multiuser operating systems are commonly used today, though their influence is much less than it once was. These operating systems offer several advantages in simultaneous use of shared resources, but have disadvantages stemming from limitations that must be imposed on each user.

4.2.2 User Interface The *interface* of a software product is the way the user interacts with it and how it displays information to the user. The interface often has a dramatic effect on the utility of the software (Card *et al.,* 1983). Physics applications have a miserable track record in this respect; many programs developed specifically for use in physics applications have horrendous user interfaces. Newer physics software has offered improvements, especially in the use of *graphical user interfaces,* or GUIs, but the quality of user interfaces in commercially available programs for physics applications still lags industry standards.

On the other hand, general-purpose software has undergone dramatic improvements in user interfaces in the past several years. General-purpose word-processing packages, for example, now routinely offer easy-to-use equation editors, enhancing their use for scientific publication.

4.2.3 Speed The power and speed of the software are, of course, important considerations in the choice of any program. In data-acquisition applications, speed may become critical, as it determines the maximum data rate that the system can handle.

4.2.4 Flexibility The *flexibility* of a software program is its utility in a variety of situations. For example, a program that can read and display data from a wide variety of possible formats is more flexible than one that requires data in a single format; a data-acquisition program that can acquire data from a wide variety of instruments is more flexible than one that acquires data from a specific configuration. Traditionally, it has been conceded that there generally exists a tradeoff between flexibility and ease of use. However, some new programs seem to be defying that rule.

4.2.5 Extensibility An *extensible* program is one to which the user can add new functions. Language compilers are the most extensible programs that exist, as they allow the user to make a new application. However, many more specialized programs now incorporate some degree of extensibility or programmability. In these situations, the display and interface may already be in place, allowing the physicist to concentrate on the physics alone.

5. EVALUATION OF COMPUTER USE IN EXPERIMENTAL PHYSICS

5.1 Advantages

The advantages of computers for experimental physics are clearly manifold. Without computers, it would have been impossible to obtain most of the significant experimental results of the last thirty years. Computers have made experimental physics more productive, more sensitive, and more comprehensive.

There are several specific but less obvious advantages that computers bring to experimental physics. First, computers allow the optimization of experimental design, both in equipment and in the design of the data-acquisition process itself. Such optimization can be used to obtain more from an existing apparatus or to design a new experiment to be more powerful for the same cost.

Second, experiments can now be automated to a great degree, significantly reducing the amount of time the experimentalist must spend monitoring the equipment.

These first two advantages can be combined into a third: Carefully designed experiments can now be carried out in a manner that minimizes the possibility that the experimentalist will introduce a bias into the results. If the entire data-acquisition phase of the experiment is planned ahead and automated, there is less tendency on the part of the scientist to abort a partially completed measurement in order to "optimize" some parameter based upon inadequate statistics.

Finally, computers have enabled scientists to search through enormous quantities of data for subtle effects, and to redo analyses of old data in the light of new discoveries. These capabilities may be important in the future, and may allow searches for experimental evidence of new effects without the need for new data.

5.2 Disadvantages

The disadvantages of computer use in experimental physics are more difficult to discern than the advantages. The most apparent is a tendency by many to hold the results of computer calculations above question. Much analysis of data requires a great deal of judgment; curve-fitting routines, for example, can often return incorrect results when not used judiciously. The same problem can plague modeling calculations.

The ability of computers to process data quickly can lead to the unwitting introduction of experimenter bias into results. If an experiment is not carefully planned and the data acquisition automated, a scientist may be tempted to modify the apparatus or the plan too quickly based on preliminary results produced as the experiment runs.

Another disadvantage arises because of the number of ways the same data can be analyzed. Statistical analysis programs exist that allow researchers to search through huge volumes of data for correlations. Unfortunately, the significance of a correlation discovered in this way is much less than that of a correlation hypothesized before the data were taken. For example, a gamma-ray spectrum might be searched for peaks corresponding to gamma-ray emission lines. The statistical significance of any lines found is quite dependent on whether the energies were known before the search was started. In short, it is important to understand statistics and probability well before using a computer to calculate probabilities despite the apparent simplicity of using the computer programs.

Finally, computer models are sometimes assigned a presumed accuracy that far exceeds their true accuracy. Great care must be used in applying computer models to the real world without a detailed understanding of the modeling process.

5.3 Example

A brief example of the use of computers throughout the experimental process is provided by a recent experiment (Knapp, 1995). Its purpose was the investigation of x-ray production from impacts between electrons and highly charged uranium ions, with particular emphasis on two processes: dielectronic recombination, a resonant process, and radiative recombination, a nonresonant process.

The experiment took place in an electron-beam ion trap (EBIT), a device that traps and strips ions in a high–current-density electron beam. The EBIT was designed using computer calculations and simulations. For example, the magnetic field along the electron-beam path had to be designed to very precise tolerances in order to achieve the required current density. For this, the magnetic field from a given magnet configuration was calculated. Then the electron trajectories in that magnetic field were calculated and analyzed, and a new magnet configuration produced that would give better results. The process was iterated many times before the optimal design was reached. The resulting optimized trajectories are shown in Fig. 1. Once the magnet configuration was optimized, the magnets themselves were designed by computer to correspond as closely as possible to the ideal magnets used in the optimization process. Finally, the mechanical parts and assembly were drawn using a CAD program.

The experiment was designed to measure the intensity of x rays produced by the two processes as a function of the electron-beam energy. It was performed by "sweeping" the electron-beam energy across the dielectronic recombination resonance energies. Both physical processes cause the ions to recombine with electrons, changing their ionization state. If the energy was not swept quickly enough, too many ions would recombine as data were obtained, making it difficult to interpret. If the energy was swept too quickly, the power supplies controlling it would be unable to keep up and it could not be measured properly. The rate at which the sweeps should be performed was estimated by fairly simple computer simulations.

The data-acquisition process consisted of recording the x-ray energy, electron-beam energy, and time for each observed x ray. The operation of the EBIT was computer controlled, and the electron-beam energy sweeps were generated by a programmable function generator with a built-in microprocessor, which communicated with the main computers over a GPIB bus. The data were taken using a CAMAC-based data-acquisition system with a total processing time of 20 μs per event. The events were stored on a read-write optical disk for later playback. They were also displayed on a computer screen so that the progress of the measurements could be followed. A total of about 50 million events were recorded in two weeks of running.

For data reduction, the data were replayed from the optical disk and displayed as a *scatter plot* of x-ray energy vs electron-beam energy, as shown in Fig. 2. Each event was displayed as a point on the screen, with the horizontal position determined by the electron-beam energy and the vertical position determined by the x-ray energy. As more events fell at the same point on the plot, the color of the displayed point was changed.

The resulting points formed patterns that allowed determination of the processes that gave rise to them. Radiative recombination resulted in x rays along angled bands, while dielectronic recombination gave rise to intense spots. The data reduction was continued by selecting only the events falling on the angled L_{12} radiative recombination band and projecting them onto the electron-beam energy axis, forming a histogram, as shown in Fig. 3. In this plot, the radiative recombination appears as a gently sloping continuum, while the dielectronic recombination produces sharp peaks.

To analyze the results, the histogrammed data were fitted by theoretical models using an iterative nonlinear maximum-likelihood algorithm running on a personal computer. The initial fit, shown in Fig. 4(a), was very poor; the theoretical curve did not match the data at all. Although the computer program that performed the fit worked well, it could

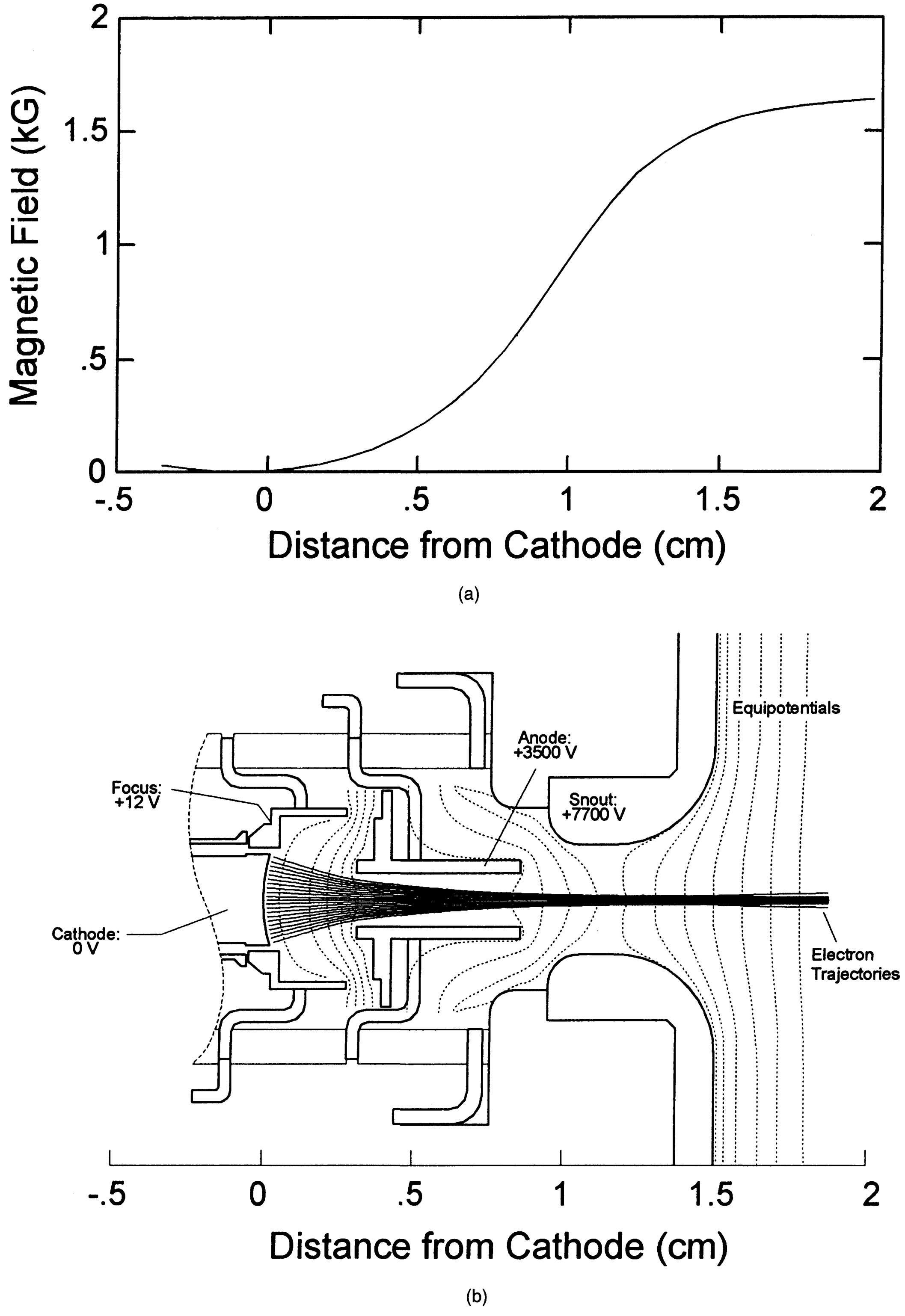

FIG. 1. Results of computer simulations for the final design of the EBIT. (a) The magnetic field on axis calculated in the vicinity of the electron gun. (b) The calculated electron trajectories and voltages in the electron gun.

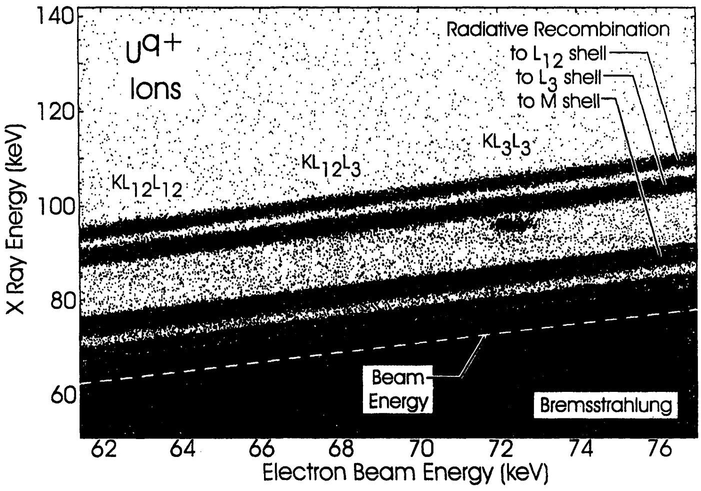

FIG. 2. Scatter plot of the data. Each point represents an event; the horizontal position corresponds to the electron-beam energy, and the vertical position to the x-ray energy. The angled bands are caused by radiative recombination, while the intense spots result from dielectronic recombination.

not reveal what was wrong. Analysis of the residuals [shown in Fig. 4(b) and defined in the caption] indicated that the shape of the peaks used in the fit was incorrect. The final fit, shown in Fig. 4(c), uses a theoretical model that includes quantum-mechanical interference between the two processes. The need to include the interference for this case was unexpected and very interesting.

A paper describing the experiment and the results was written and submitted electronically. The figures used in the paper were generated using computers.

Thus, computers were applied to this experiment from its inception through its publication. It illustrates many of the ways computers are used in physics experiments today.

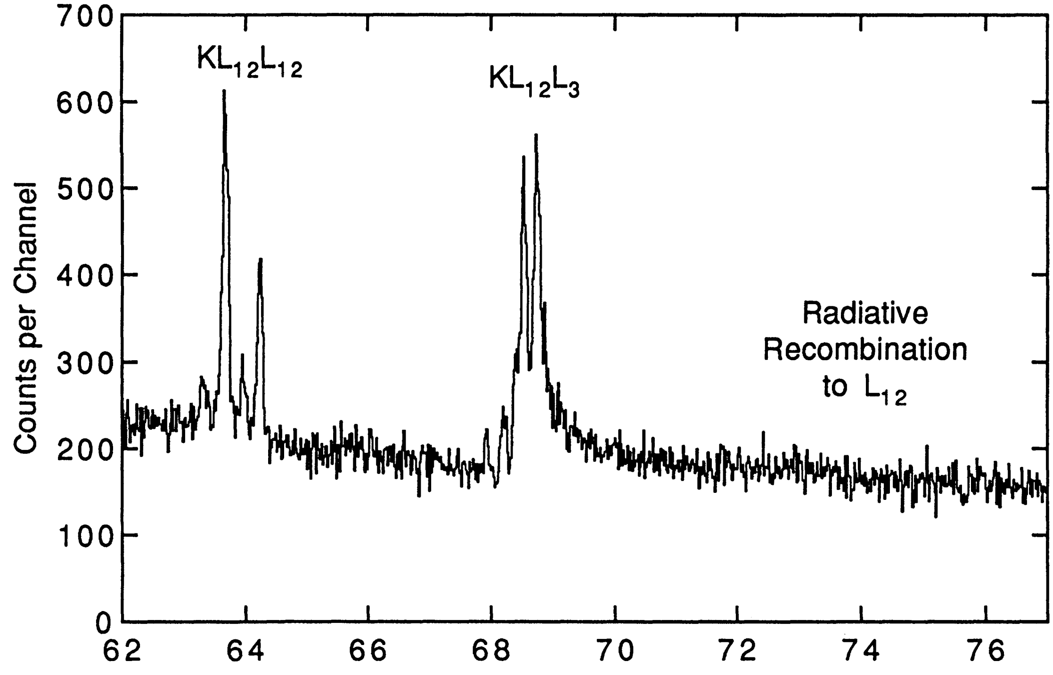

FIG. 3. Histogram of events falling on the L_{12} radiative recombination band projected onto the electron-beam energy axis. The dielectronic recombination resonances are easily visible as sharp peaks in this view.

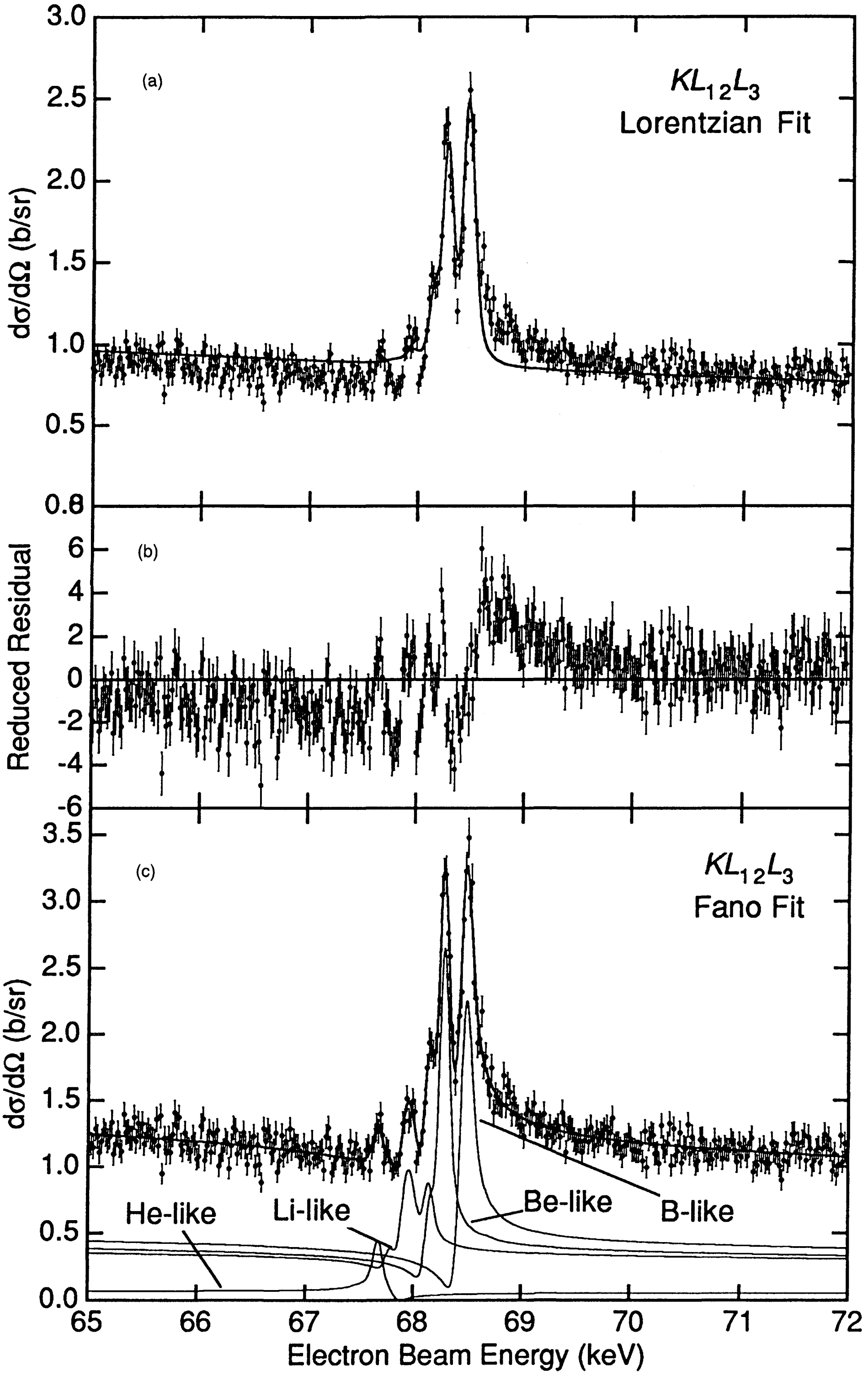

FIG. 4. Fits of the $KL_{12}L_3$ dielectronic recombination resonances by two competing models. (a) Fit by theory without quantum interference. (b) Reduced residuals (difference between data and fit) for (a). (c) Fit by theory with quantum interference.

6. FUTURE TRENDS

Predicting the future is always a difficult affair; however, the future role of computers in experimental physics is relatively easy to predict. As computers become cheaper and more powerful, they will become even more tightly integrated into experimental physics than they are today. More computer-controlled smart instruments that can communicate with a main controlling computer will become available.

The current trend of decentralization of computing resources is likely to continue, as individual processors are used for each task in an experimental environment. It seems likely that these individual processors will be networked to a greater degree than they are today. The use of wide-area networks for communications and data transfer will continue to increase, probably dramatically.

One particularly interesting trend is that of electronic and on-line publishing. The role of traditional print journals will certainly change in the next few years as the ability to post experimental results quickly to a wide audience grows.

The biggest problem today will likely continue into the near future; it is that of development of the specialized software required for experimental applications. Nonetheless, the sophistication and complexity of experimental physics software will continue to increase.

GLOSSARY

Analog-to-Digital Converter (ADC): Digital device that converts an analog voltage into a digital number.

Analog: A value represented by a continuous quantity, such as a voltage.

Binary: Numbers expressed in base 2: each binary digit (*bit*) can have a value of 0 or 1.

Bit: Binary digit.

Byte: Shortest binary number that a computer processes. In the vast majority of computers today, a byte is made up of 8 bits.

Central Processing Unit (CPU): A device in a computer that performs the basic arithmetic and logical operations for the computer.

Continuous: Describing a function or operation whose value can correspond to any real number.

Data Bus: The electronic data path via which components in a computer communicate with each other and with external devices.

Digital: A representation of numbers by digits in some base. For example, most numbers are commonly written in base 10, while computers primarily work with numbers in base 2. Both of these are digital representations of the numbers.

Discrete: Describing a function or operation whose value cannot correspond to any real number, but only to a set of separate numbers.

Embedded Processor: A computer processor that is integrated into an instrument or device.

Graphical User Interface: An interface consisting of graphical elements such as pictures.

Interface: The means by which a user interacts with a program running in a computer.

Mainframe: A large computer intended primarily to be accessed by many users simultaneously.

Monte Carlo: A technique for studying mathematical problems that consists of generating many random simulations of the problem.

Scaler: A device that counts events.

Text-Based User Interface: An interface consisting primarily of text, with few or no pictures.

Time-to-Digital Converter (TDC): A device that converts the value of a time interval into a digital number.

Workstation: A compact computer that is configured to be used primarily by one (or a few) users.

Works Cited

Card, S., Moran, T. P., Newell, A. (1983), *The Psychology of Human-Computer Interaction,* Hillsdale, NJ: Lawrence Erlbaum Associates.

Dasgupta, S. (1989), *Computer Architecture: A Modern Synthesis,* New York: Wiley.

Goldstine, H. (1972), *The Computer: from Pascal to von Neumann,* Princeton, NJ: Princeton Univ. Press.

Grubb, D. P., Borchers, R. R. (1991), "Workstations Emerge as Major Scientific Resource," *Comput. Phys.* **5**, 571.

Kent, W. (1978), *Data and Reality: Basic Assumptions in Data Processing Reconsidered,* Amsterdam: North-Holland Publishing.

Knapp, D. A., *et al.,* (1995), "Observation of Interference between Dielectronic Recombination and Radiative Recombination in Highly Charged Uranium Ions," *Phys. Rev. Lett.* **74**, 54.

Matey, R. (1993), "General Principles for Application of Computers in Experimental Physics," *Comput. Phys.* **7** (12).

PHYSICS APPLICATIONS OF COMPUTERS, THEORETICAL

See SIMULATION AND MODELING BY COMPUTER

PHYSICS IN ARCHAEOLOGY

Robert S. Sternberg, *Department of Geosciences, Franklin and Marshall College, Lancaster, Pennsylvania, U.S.A.*

Introduction 69
1. **Dating** 69
1.1 Radiocarbon Dating 70
1.2 Other Radiometric Methods .. 71
1.3 Archaeomagnetic Dating 73
2. **Prospecting** 75
2.1 Magnetics 75
2.2 Resistivity 76
2.3 Other Electrical Methods 77
3. **Artifact Characterization** 78
3.1 Provenance Studies 78
3.2 Ancient Technology 78
4. **Paleoenvironment** 80
4.1 Paleoclimate 80
4.2 Paleodiet 80
4.3 Organic Residues 82
Glossary 82
Works Cited 83
Further Reading 84

INTRODUCTION

Archaeology adds to our understanding of human society through the systematic excavation and analysis of material culture. Traditional archaeological methodology focuses on controlled excavation, classification of features and artifacts, social interpretation of what is found at a particular site, and integration of results from multiple sites into regional cultural frameworks.

The application of physics and other natural sciences to archaeology adds tremendous power to the archaeologist's toolkit. This multidisciplinary and interdisciplinary field of endeavor has been named *archaeometry*, or archaeological science. Archaeometric methods were initially used largely by scientists with an interest in archaeology (Leute, 1987, pp. 2–4; Craddock, 1991). In recent years, archaeometry has started to become more central in the education of the archaeologist and in the planning and execution of archaeological projects (Taylor and Payen, 1979; Olin, 1982).

The diverse methods of archaeometry are most easily classified according to the major problems they can help the archaeologist to solve: dating, prospecting, characterization, and paleoenvironment. Absolute or *chronometric* dating is necessary to examine contemporaneity and change within an individual site, within a culture region, or between culture areas. Geophysical prospecting is used to map out paleogeography favorable for site location, to locate sites, to delineate site boundaries or work areas, or to detect features and artifacts within a site. Characterization of features and artifacts is used to infer the sources (also called *provenances*) of raw materials or the technology that was used to manufacture artifacts. Paleoenvironmental studies suggest the landscape, climate, and biological (including dietary) resources that were available. It is impossible to review all archaeometric methods comprehensively herein, but general principles and applications are described for several of the most important techniques, with some bias in coverage toward the author's areas of expertise. References given are not in general to seminal articles but mostly to review articles and books. In addition to the references listed at the end of this paper, many articles concerning archaeometry can be found in the following journals: *Journal of Archaeological Science, Journal of Field Archaeology, Archaeometry, Archaeomaterials*, and *Geoarchaeology*.

1. DATING

Absolute, or chronometric, dating determines an age or interval of time in years, while relative dating assesses the contemporaneity or sequence of events. The full panoply of chronometric methods vary consider-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ably in the amount and type of material required, the range of archaeological time they can be used for, and the precision of the dates (Parkes, 1986; Aitken, 1990). Sternberg (1990) classifies these methods according to two dimensions:

1. whether they are "clocks" or involve matching of patterns, and
2. whether they are applicable globally or regionally/locally.

Dean (1978) provides a useful theoretical framework for chronometric dating, which clarifies the dependence of a meaningful date upon both proper laboratory work and consideration of archaeological context and relationships. The event being dated and the target cultural event of interest may be separated by an additional bridge of time, which must be archaeologically inferred.

1.1 Radiocarbon Dating

Radiocarbon (^{14}C) dating (Taylor, 1987) has had a more profound effect on the development of archaeology than any other archaeometric method. It has also played an important role in Quaternary geology, geochemistry, geophysics, hydrology, environmental studies, and medicine (Taylor *et al.*, 1992). The discovery of ^{14}C dating in the years after World War II by Libby (1955) earned him the Nobel Prize in chemistry in 1960.

Radiocarbon is formed in the upper atmosphere when neutrons generated by collisions of cosmic rays with atmospheric constituents are captured by nitrogen-14. The radioactive isotope beta decays back to ^{14}N with a half-life of 5730 years. However, the ^{14}C rapidly oxidizes to CO_2 and becomes distributed through the carbon geochemical system involving atmosphere, biosphere, and oceans. Modern carbon is about 98.9% ^{12}C (stable isotope), 1.1% ^{13}C (stable), and 1×10^{-10} % ^{14}C. Living objects will maintain the ambient ratio of $^{14}C/^{12}C$, but upon death, this ratio begins to decrease at a rate commensurate with the half-life. Thus, the radiocarbon age can be derived from the equation for radioactive decay,

$$A = A_0 e^{-t/\tau}, \tag{1}$$

where A is the present ^{14}C activity of the sample, A_0 is the initial activity of the sample at $t = 0$, t is the age in years, and τ is the mean life. The relation between mean life and half-life is

$$\tau = t_{1/2}/0.693, \tag{2}$$

which is equal to 8267 years for ^{14}C. Radiocarbon dating is an example of a "global clock" that "ticks" at a uniform rate anywhere in the world at any time. The technique is useful from historic times back to about 50 000 years ago, when the amount of ^{14}C becomes too low to measure.

For the first three decades of ^{14}C dating, activity was measured by counting the beta particles emitted by the decaying ^{14}C atoms. Libby's original method was to coat the inside of a Geiger counter with solid carbon. This was soon replaced by proportional gas counting by converting the ^{14}C back into CO_2 or a hydrocarbon gas, which is used as a counting gas within a proportional counter. Liquid-scintillation counting methods combine a liquid scintillator with a benzene solution that contains the radiocarbon. With these conventional methods, a few grams of carbon are typically needed. High-precision conventional dates can be as precise as ±20 years, but counting time may be several months. Common sources of error are background counts from sources other than the beta decay of ^{14}C and counting errors due to the random nature of the decay process.

Since 1980, a new measurement technique has emerged that has caused a revolution in ^{14}C dating: direct counting of ^{14}C atoms with an accelerator mass spectrometer (AMS), using either a cyclotron or an electrostatic accelerator. In the electrostatic accelerator, a negative-ion source and stripping of C^- ions to form C^{3+} ions separates ^{14}C from the more abundant but short-lived ^{14}N negative ions. The ions are accelerated in a tandem (two-stage) electrostatic accelerator. Mass spectrometry is used to separate and count ^{12}C, ^{13}C, and ^{14}C atoms. The ^{14}C ions are measured in a conventional counter while the more numerous stable isotopes are measured with a Faraday cup. The key impact of AMS dating is not in being able to date older material but in being able to date smaller amounts of material. Only milligrams of material are required, as opposed to the grams required by conventional methods. This allows more selectivity to be used in the choice of sample material. The actual counting time is also reduced from several days in beta counting to under an hour in direct counting.

In the late 1950s, it was first realized that there were secular changes in the amount of ^{14}C in the atmosphere and other geochemical reservoirs; in other words, the A_0 term in Eq. (1) is not a constant. It is now generally agreed that there is a long-term variation in the atmospheric radiocarbon concentration, with an amplitude of about ±5% and a period of 10 000 years, and short-term fluctuations with periods of decades and amplitudes of about ±1% (Olsson, 1970; Stuiver and Kra, 1986). The details of these variations, as shown in Fig. 1, have been deciphered by radiocarbon dating tree rings of known dendrochronological age. The long-term changes are probably geomagnetic in origin; when the earth's magnetic field is stronger (or conversely, weaker), fewer (more) cosmic rays can penetrate into the atmosphere and less (more) ^{14}C is produced. The short-term changes are probably heliomagnetic in origin; when the sun's magnetic field is stronger (weaker), fewer (more) cosmic rays can penetrate into the inner solar system. Climatic changes can also alter the carbon geochemical cycle and affect the radiocarbon concentration of the various reservoirs (Stuiver *et al.*, 1991). These changes necessitate the calibration of ^{14}C dates, which is the conversion of the laboratory-generated raw dates given in radiocarbon years to calibrated ages given in calendar years (Stuiver and Kra, 1986). During the past 10 000 years, the discrepancy between radiocarbon and calendar ages is as high as 1000 years, occurring ca. 5000 B.C.

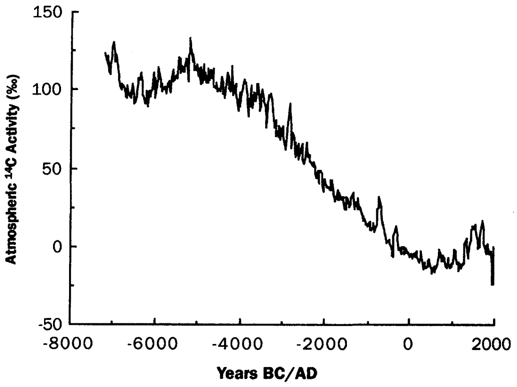

FIG. 1. Radiocarbon activity (per mil deviation from the standard) vs time, from measurements on dated tree-ring samples. Data from Stuiver and Reimer (1986).

Many materials have been used for radiocarbon dating, including wood and charcoal, plant materials, textiles, shells, and bone, involving a variety of sampling considerations and pretreatments. Dating of individual items occasionally attracts great public interest. Damon *et al.* (1989) describe the dating of the linen known as the *Shroud of Turin*, which many people believe was used to wrap the body of Jesus. This study was only made possible because of the small samples allowed by AMS dating. Three laboratories (Arizona, Oxford, Zurich) participated in the study, obtaining consistent results that yielded an overall calibrated calendar age of A.D. 1260–1390 at the 95% confidence level, demonstrating that the linen is medieval. Most problems involve a greater number of dates from an individual archaeological site or a group of sites. For example, Haynes (1992) reviews the contributions of radiocarbon dating to the continuing debate about when people entered the New World. Radiocarbon dating made it possible to first date North American Paleoindian sites by means other than association with glacially related geologic deposits. Radiocarbon dates from the Clovis Paleoindian culture in the American Southwest cluster from 11 200 to 10 900 years before present. Haynes claims that earlier ^{14}C dates from other sites, such as Meadowcroft Rockshelter in Pennsylvania (ca. 14 000 BP) and Monte Verde in Chile (ca. 12 800 BP), are as yet ambiguous, because of the possibilities of sample contamination, complicated stratigraphy, erroneous identification of human artifacts, and/or unclear associations between dated material and human artifacts. The controversy concerning the peopling of the New World is discussed by Marshall (1990). Other published applications of radiocarbon dating abound, especially in the proceedings of the triennial International Radiocarbon Conference, published in the journal *Radiocarbon.*

1.2 Other Radiometric Methods

Another conventional radioactive dating method useful in archaeology is the *potassium-argon* (K-Ar) method (Aitken, 1990).

The decay of ^{40}K to the gas ^{40}Ar with a half-life of 1250 million years still allows precision of ±1000 years for materials as young as 1000–30 000 years old if there is sufficient potassium. Volcanic rocks associated with archaeological sites are the actual materials dated; the thermal volcanic event releases the argon gas and resets the radiometric clock. The traditional approach determined potassium content chemically, but the newer argon-argon method converts some of the potassium into ^{39}Ar by neutron irradiation, after which it can be measured by mass spectrometry along with the ^{40}Ar. Potassium-argon dating is useful in studies of hominid evolution (Bishop and Miller, 1972), as discussed below for magnetic reversals.

The decay of the radioactive isotope ^{238}U provides the basis for *uranium-series dating* (Aitken, 1990). The chain of daughter radioisotopes includes three relatively long-lived isotopes: ^{238}U, ^{234}U, and ^{230}Th. The buildup of ^{230}Th is controlled by its 75 400-year half-life. Measurements in the conventional method are made by alpha spectrometry, following chemical extraction of the elements involved. The age range covered by this method is from 5000 to 350 000 years, again extending into the time frames of hominid development and the Paleolithic period. The materials that can be dated include material associated with archaeological sites—calcite in cave speleothems, and travertine deposits from springs—as well as archaeological materials of bone, teeth, coral, and shells.

Other dating methods are possible because of the effects of radiation. *Thermoluminescence dating* (Aitken, 1985; Aitken, 1990) has been useful as a method for directly dating the manufacture of pottery or the heating of other materials such as baked clay, burned flint and stones, volcanic rocks, slag, and glass. Certain minerals in pottery, mostly quartz and feldspar, will emit light in excess of the normal red-hot glow when heated above 350 °C. This extra light is the thermoluminescence (TL). The origin of the TL can be described via a four-step process:

1. Irradiation of the sample from radioactive sources within the soil and the sample itself will free electrons in the sample from their parent nuclei. Over archaeological time, the irradiation is constant, and the number of freed electrons is proportional to time.
2. These electrons can then become trapped in crystalline defects, the number of which remains constant in the sample over archaeological time.
3. Heating of the sample, in nature or in the laboratory, expels electrons from their traps.
4. Nearly instantaneous recombination of these electrons with ions at luminescence centers generates the thermoluminescence.

The age is determined by calibrating the natural TL against a TL generated by a known laboratory dose of radiation. The TL can be measured by a sufficiently sensitive photomultiplier to a precision of ±5% of the age. The age range for this method extends back beyond the origin of pottery, ca. 7000 B.C. in the Near East. TL is a "regional clock" in that it is sensitive to the radioactive environment at each site.

Besides dating *per se*, TL has also become important as a method of detecting forgeries. For example, forgery was successfully detected for a group of stylistically suspect ceramics from Haçilar, Turkey, by the Research Laboratory for Archaeology and the History of Art at Oxford (Jones, 1990). Authenticity dating differs from regular dating in two important respects: only a small sample size is normally possible, and the radiation history during the life of a museum piece is so variable that it is basically unknown.

Another technique related to TL is *electron-spin resonance* (ESR) dating, also known as electron paramagnetic resonance (EPR) (Ikeya and Miki, 1985; Aitken, 1990). This is another method of evaluating the dosage of radiation accumulated by a sample. With this method, electrons are not ejected from their traps. Instead, they are detected by their response to high-frequency electromagnetic radiation while also in the presence of a slowly varying dc magnetic field. At a given frequency, a condition of resonance is established for a certain magnetic field strength. The absorption of the electromagnetic radiation is proportional to the concentration of trapped electrons. The same materials used for TL dating are amenable to ESR; dating of speleothem calcite is the most common application, with an age range of about 10^4–10^6 years.

1.3 Archaeomagnetic Dating

Archaeomagnetism is a subfield of *paleomagnetism* that uses archaeological materials or materials that have been magnetized during archaeological time (Tarling, 1983; Eighmy and Sternberg, 1990). The materials most commonly used are baked clays from fireplaces, clays from floors or walls baked during conflagration, and ceramics. If the date of the magnetization is known, these materials can be used to construct the pattern for secular variation for a geographic region. Once this has been done, the pattern can be used as a master curve for absolute dating.

Two major phenomena are involved in archaeomagnetic dating. The first is that the geomagnetic field changes over time. This secular variation is due to fluid motions in the Earth's outer core, where the geomagnetic dynamo originates. Geomagnetic directions change at rates of typically 1° per 25 years, whereas the dipole moment of the magnetic field has decreased by about 5% since the time of Gauss, who first measured absolute field strength.

The other major phenomenon is the ability of certain minerals in rocks to acquire a permanent (remanent) magnetization. The most important such minerals are magnetite (ferrimagnetic) and hematite (antiferromagnetic). Rocks can become magnetized parallel to the magnetic field by several mechanisms, the most important occurring when they are heated above (or close to) the Curie temperature (580 °C for magnetite; 680 °C for hematite), thereby acquiring a thermoremanent magnetization (TRM). Even when the field changes, stably magnetized rocks can retain most of their remanence parallel to the direction of the ancient field. Archaeomagnetism, like TL, has an advantage over ^{14}C in that the event being dated, the firing of the clay feature, is usually a cultural event due to the activity of people living at a site.

Archaeomagnetic direction studies require samples oriented in the field. Samples of about 2 cm^3 are collected within an oriented plastic or plaster-of-paris cube. From 6 to 12 oriented specimens might be collected from a single feature, such as a hearth. In the laboratory, magnetic moments of these samples are measured with either a spinner or cryogenic magnetometer. Typical moments of magnetization are 3 A m^2. For studies of magnetic direction, the stability of magnetization must be assessed by demagnetizing the sample, using either thermal or alternating-field demagnetization. These analyses are used to remove unstable secondary components of magnetization in order to retrieve the primary magnetic direction of interest. Directions are analyzed using spherical statistics, which were originally developed largely for the analysis of paleomagnetic results.

Unoriented as well as oriented materials can be used for paleointensity studies, an advantage because potsherds are very abundant at many archaeological sites. Paleointensity studies have most often been done with samples carrying a TRM, which theory and experiment indicate is linearly proportional to the strength of the magnetizing field. Thus,

$$\mathrm{TRM}_{\mathrm{paleo}}/\mathrm{TRM}_{\mathrm{lab}} = B_{\mathrm{paleo}}/B_{\mathrm{lab}}, \qquad (3)$$

where $\mathrm{TRM}_{\mathrm{paleo}}$ is the strength of thermoremanent magnetization acquired in the paleomagnetic field having paleointensity B_{paleo} and $\mathrm{TRM}_{\mathrm{lab}}$ is the strength of magnetization acquired in the laboratory field B_{lab}.

If the secular variation pattern is not yet well established for a region, archaeomagnetism can be useful as a relative dating method. A wide variety of dating methods have been applied to the destruction of Thera, or Santorini (Hardy and Renfrew, 1989). Downey and Tarling (1984) used archaeomagnetic results to argue that the destruction of Minoan civilization on Crete ca. 1450 B.C. occurred in two phases. Archaeomagnetic directions and paleointensities in central Crete were consistent with each other and paleomagnetic results from ashfall deposits on Thera, 120 km to the north. Archaeomagnetic directions and paleointensities in eastern Crete were consistent with each other and paleomagnetic results from stratigraphically younger volcanic deposits on Thera. The difference in directions between these two groups, along with typical rates of secular variation, suggests a time difference of 20 ± 10 years. Downey and Tarling (1984) argued that if the first eruption did not eradicate civilization throughout Crete simultaneously, then different earthquakes are more likely causes than volcanic eruptions for the two phases of destruction.

However, archaeomagnetic dating has most often been used as an absolute dating method by comparing the direction of the feature to be dated against a pre-established secular variation curve. Because the nondipole field of the earth constitutes regional anomalies of the dipole field, archaeomagnetic dating becomes a "regional pattern-matching" dating method. Secular variation data from different regions are geophysically important in understanding the behavior of the magnetic field (Creer *et al.*, 1983). In the American Southwest, archaeomagnetism has become a routine dating method (Eighmy and Sternberg, 1990); the secular variation curve can be calibrated by tree-ring dating at the higher elevations of the Colorado Plateau, but archaeomagnetic dating can be applied in the nearby desert where tree rings are too complacent for dendrochronology. The accuracy and precision of archaeomagnetic dating in this region are comparable to ^{14}C dating (Sternberg *et al.*, 1991). Paleomagnetic results for recent volcanic rocks can also be compared with and dated against archaeomagnetic results. For example, as shown in Fig. 2, paleomagnetic data for the Sunset Crater volcanic rocks in northern Arizona indicate that the flow from Gyp Crater could indeed have been erupted in A.D. 1066 (Sternberg and McGuire, 1990), as is suggested by stunted tree rings dating to that year.

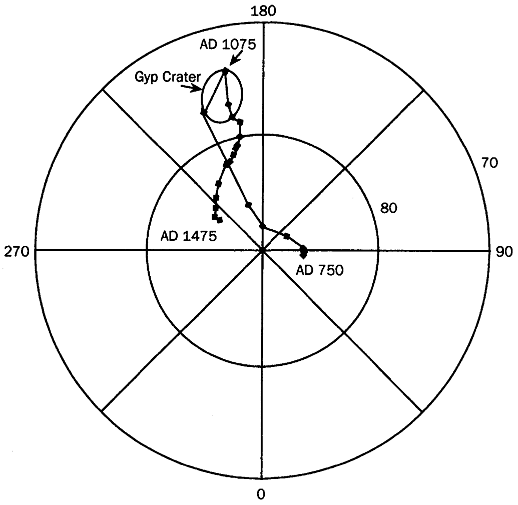

FIG. 2. Archaeomagnetic dating of igneous rocks from the Sunset Crater volcanic field, northeastern Arizona (Sternberg and McGuire, 1990). View is looking down on the north geographic pole. The archaeomagnetic directions are represented as geomagnetic poles. The points on the secular variation curve cover the time period A.D. 750–1475; each point on the curve represents a 50- or 100-year average of archaeomagnetic directions. The paleomagnetic direction and 95% oval of confidence for the lava flow from Gyp Crater yields an archaeomagnetic date of A.D. 1000–1175 by comparison to the secular variation curve.

The paleointensity experiment is tedious and difficult, and so, although it can ideally resolve ambiguities in directional dating for oriented samples (Downey and Tarling, 1984), it has been rarely used for dating alone. One example was the use of paleointensities to examine the ceramics from the Glozel site in France. It has been suggested that these very unusual ceramics may be fakes; one element of this argument is the inconsistency of different dating methods, including ^{14}C, TL, and paleointensity (Jones, 1990). On the basis of additional results from other well-dated ceramics, Shaw (1979) concluded that the paleointensity in Europe varied more rapidly between 100 B.C. and A.D. 300 than was previously thought, yielding a revised Glozel paleointensity date of about 50 B.C.–A.D. 100 that would be consistent with thermoluminescence dates. Although it did not resolve the Glozel controversy, this study emphasized the important archaeomagnetic principle that if the secular variation curve is revised, the original basic paleomagnetic results on the dated object can be reinterpreted to give a new date in accordance with the new pattern.

Another type of field variation that is useful in archaeology is the geomagnetic reversal: Occasionally, the magnetic field reverses polarity as the north and south magnetic poles exchange positions. There have been at least 23 reversals in the past 5 million years, the last major reversal being at 730 000 years ago. The recent reversal chronology pattern was established through paleomagnetic studies and potassium-argon dating of igneous rocks. Magnetic reversal stratigraphy has been used to calibrate the evolution of hominids. Brock and Isaac (1974) used reversals recorded via depositional remanent magnetization in sediments from the Lake Rudolf area of the African Rift Valley to examine the chronology of hominid fossils and associated artifacts and to compare these dates with other potassium-argon and paleontologic ages. Dating by use of globally synchronous magnetic reversals is a "global pattern-matching" dating method.

2. PROSPECTING

Archaeological excavation is a time-consuming and laborious undertaking. Geophysical prospecting (also known as prospection, surveying, exploration, and field geophysics) can help archaeologists optimize their available time and resources for field work. Geophysical prospecting at archaeological sites utilizes the same principles of geophysics used to prospect for geologic features and resources (Wynn, 1986), but with an emphasis on shallow and small-scale targets (Burger, 1992).

All geophysical surveying methods depend on an "operative physical property" (Kearey and Brooks, 1984, Chap. 1). Contrasts in these properties between different rocks and different archaeological materials are necessary to generate detectable anomalies in the quantity being measured. Different methods are sensitive to different properties, and so the technique used for a particular survey depends on the features being sought, their expected properties relative to the surrounding material, the area to be covered, the desired resolution, and the ability to see through cultural noise (fences, power lines, *etc.*) (see GEOPHYSICS).

2.1 Magnetics

The operative physical property in magnetic surveying is magnetization. Archaeological materials may carry a remanent magnetization, such as that used for archaeomagnetic studies, which is considerably stronger than the induced magnetization of surrounding material. The magnetic susceptibility and induced magnetization of archaeological features and soils can also vary, largely depending on the magnetite content. These variations in the magnetic properties of materials at archaeological sites generate anomalies in the earth's magnetic field.

Magnetic surveys differ from archaeomagnetic studies in that the magnetization of samples is not directly measured; rather, the geomagnetic field, including its perturbation by the field due to these magnetized materials, is studied. Two types of field magnetometers are in common use today. In proton precession magnetometers, a field applied by the instrument torques the nuclear magnetic moments of the proton-rich sensor fluid into alignment; when the applied field is turned off, the protons precess into alignment with the ambient field at a frequency equal to the product of the proton's gyromagnetic ratio and the strength of the ambient field. Counting the frequency thus gives a measure of the total field strength. The cycle time for the proton precession magnetometer is typically several seconds, limiting the spatial resolution or speed with which surveys can be carried out. The fluxgate magnetometer uses a sensor of two soft ferromagnetic cores wound such that a signal is generated with an amplitude proportional to the strength of the magnetic field along the axis of the sensor. The fluxgate magnetometer is generally used to measure the vertical component of the magnetic field. The measurement precision of these instruments is better than 1 nT, out of a typical midlatitude geomagnetic field strength of 50 000 nT. The method of data treatment and display affects the ability of the user to identify meaningful anomalies. Various methods are available for the great number of points that are measured during magnetic surveys, 2500–10 000 for a typical European site (Scollar *et al.*, 1986).

The geomagnetic field undergoes a diurnal variation of up to several tens of nanoteslas due to ionospheric currents. This variation may be an order of magnitude greater than anomalies of interest. One way to account for this is to establish a base station at which readings are periodically or continuously monitored. Both types of magnetometer can also be operated in the vertical gradient mode using two sensors. This has the dual advantage of eliminating the effects of diurnal variation, and of emphasizing the effects of near-surface (archaeological) over deeper seated (perhaps geological) features.

Breiner and Coe (1972) discussed an unusual application of magnetic prospecting for individual features: carved basalt monuments, weighing up to 40 tons, of the Olmec civilization that inhabited southern Mexico from 1200 to 400 B.C. These impressive features were individually identifiable through their anomalies of over 100 nT, presumably due to the remanence of these features. Magnetic susceptibility contrast in disturbed soil is responsible for the 10–20-mT anomalies shown in Fig. 3 (Clark, 1986), from a site in Oxfordshire, U.K. The two large circles represent ditches of the Bronze Age burial mounds; other circular and linear features are also visible.

2.2 Resistivity

Magnetic surveying is "passive" in that it reads a potential field that already exists. Gravity is a similar passive geophysical method but has only seen limited application in archaeology because of the limited resolution of conventional gravity surveying. Several other geophysical methods are "active," in that currents or waves are sent out, and the response of the earth is then measured.

Electrical resistivity has been the most commonly used active method for archaeological prospecting. The operative physical property in this technique is electrical resistivity,

$$\rho = RA/l, \tag{4}$$

where ρ is resistivity in ohm-meters, R is electrical resistance in ohms, A is the cross-sectional area of the current flow path, and l is the length of this path. Resistivities of earth materials can vary by over six orders of magnitude. Resistivity depends not only on the material itself but also on porosity, fraction of the pore space filled with water, and the resistivity of the pore water. Thus, resistivity surveying is often used to locate relatively low-porosity building structures in a matrix of higher-porosity soil.

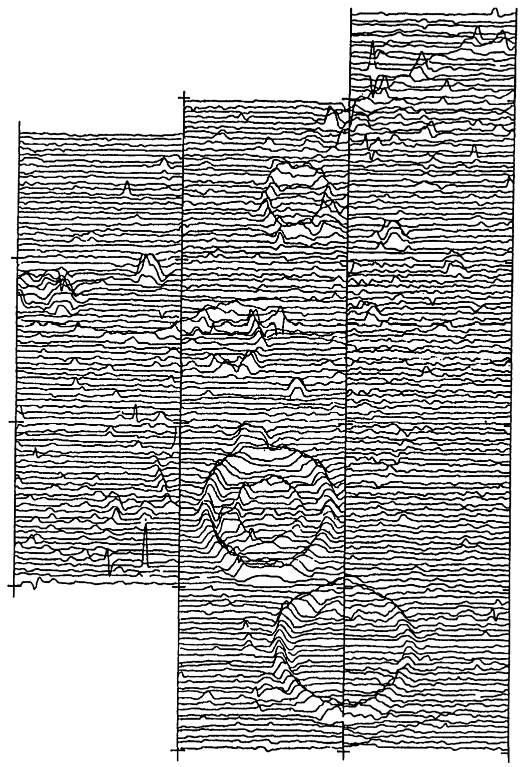

FIG. 3. Magnetic anomalies due to Bronze-age burial mounds at Oxfordshire, England (Clark, 1986; reproduced by permission of the Society of Exploration Geophysicists), shown by individual traces from a fluxgate gradiometer survey. Sensitivity is 7.5 nT/line interval.

Resistivity surveying is commonly done with a four-electrode setup, with two electrodes used to inject dc or low-frequency ac current into the ground, and the other two electrodes being used to measure the resulting voltage. A typical current is 30 mA. For the Wenner electrode array, the electrodes are collinear, equally spaced, with two outer current electrodes and two inner potential electrodes. For a uniform earth with resistivity r, the resistivity will be

$$\rho = 2\pi aV/I, \tag{5}$$

where a is the electrode spacing, I is the current, and V is the measured voltage. However, when the resistivity varies laterally or horizontally, this becomes an apparent resistivity ρ_a, which integrates the varying resistivity of the ground below. The effective depth of penetration for the Wenner array is about $1.5a$.

Resistivity surveys are carried out in two modes: sounding and profiling. In the sounding mode, using the Wenner array as an example, the spacing a is increased, with the center of the array remaining at the same point. As the spacing a increases, more of the current penetrates to greater depth, and a vertical electrical stratigraphy can be inferred. Various methods of interpretation can be used to infer the resistivities and thicknesses of horizontal, planar, electrically homogeneous layers. When profiling, the entire array is moved forward between measurements, maintaining the same spacing a.

The goal is now to find lateral changes in resistivity. These two modes can be combined to generate a pseudosection, where apparent resistivities are plotted horizontally at the midpoint of the array, and at a depth corresponding to the spacing *a*, to give an apparent cross section. Figure 4(a) shows contoured results from an electrical survey at historic Fort Wilkins, Michigan (Young and Droege, 1986). Privies were among the targets of this study; privies become convenient dumping grounds for one-time garbage that has now become archaeological artifacts. Resistivity anomalies A and B mark the site of known privies, whereas C is an anomaly of unknown origin. The pseudosection [Fig. 4(b)] indicates that the anomalies are strongest at depths between 1.0 and 1.5 m.

2.3 Other Electrical Methods

Two other methods that are receiving increased attention also involve electrical response of earth materials (Wynn, 1990). Electromagnetic conductivity (conductivity = 1/resistivity) surveying uses waves rather than electrical currents, transmitted and received via antennae with frequencies of typically 10–100 Hz. Anomalies are due to secondary magnetic fields at the receiver from eddy currents generated in the ground. This method is less useful than resistivity for sounding because antenna separation is fixed in most instruments, but faster for profiling because there are no electrodes to insert into the ground. Ground-probing radar (GPR) is a rapidly developing method, where dual antennae send radar waves between 80 and

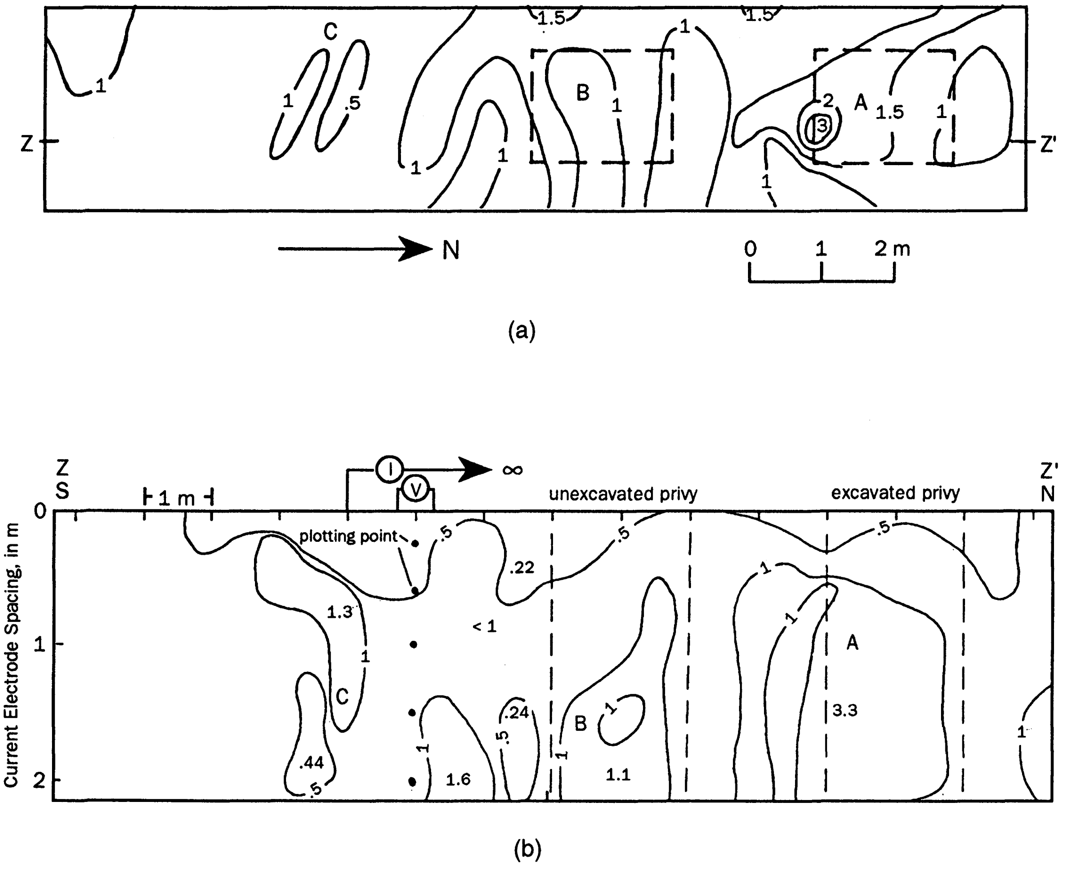

FIG. 4. Resistivity results from historic Fort Wilkins, Michigan (after Young and Droege, 1986; by permission of the Society for Exploration Geophysicists). Resistivities are contoured in units of kΩ-m. (a) Map view showing resistivity high-low anomaly pairs over two known privy locations at A and B. The source of anomaly C is unknown. (b) A resistivity pseudosection along the line of profile Z-Z′; maximum values are shown within contours.

1000 MHz into the ground and continuously map the reflections as the antennae are pulled across the ground. Both discrete parabolic reflectors and layered planar reflectors can be seen by radar when there is sufficient contrast in electrical impedances. Highly conductive soils can greatly limit the depth of penetration of the radar waves.

3. ARTIFACT CHARACTERIZATION

Studies that characterize artifacts involve almost every type of archaeological material, including pottery, stone, glass, and metals. Parkes (1986) provides a useful overview of the goals and methods of these studies. A wide variety of analytical methods has been used to determine the source location (or provenance) of the raw materials, to follow the trade routes of the materials or manufactured artifacts, to infer how the objects were manufactured, and to suggest how these objects were used. The method to be used depends upon the problem to be solved (e.g., which elements are of interest for chemical composition), whether quantitative (e.g., content) or qualitative (e.g., presence vs absence) results are needed, whether the artifact can be destroyed or not, and whether the surface or bulk properties of the artifact are of interest.

3.1 Provenance Studies

Provenance studies involve chemical characterization of the materials of interest to determine their source, or place of origin. Chemical analyses may focus on the major elements (≥2%), minor elements (0.1–2%), or trace elements (≤0.1%). Most chemical analyses are done using some type of spectroscopy. Besides the choice of analytical method, the other major methodological problem in provenance investigations is having analyses available from all possible source deposits and using appropriate statistical methods for comparing the artifact and potential source compositions (Rapp and Gifford, 1985).

Lead isotopes have been particularly useful in provenance studies of metallic artifacts (Gale, 1989), including objects made of lead, copper and copper alloy, iron, and silver. Of the four isotopes of lead, ^{204}Pb is nonradiogenic, and ^{206}Pb, ^{207}Pb, and ^{208}Pb are in part due to the radioactive decay of uranium and thorium. Thus, the lead isotope composition, which varies by 5% between deposits, depends upon the primeval lead content and the radiogenic history of the body. The lead isotope content shows little variation within an ore body and is essentially unaffected by smelting, two prerequisites for provenance analysis. The same cannot be said for chemical composition of ores and artifacts. The most common analytical method is thermal ionization mass spectrometry, which can measure ratios to ±0.1% on 100 ng of lead. Among the applications cited by Gale (1989) is the question of the source of copper for the Mycenaeans and Minoans in Crete ca. 1700–1500 B.C. The conventional assumption is that Cyprus (the root of the word copper) was the major source area, because Cyprus was a known copper producer, and Mycenaean and Minoan pottery was found in Cyprus. However, the lead isotopes indicate instead that the Laurion field in Attica on the Greek mainland was the likely source area.

Among nonmetals, stable isotopes of the lighter elements carbon and oxygen have also been useful for provenance studies of marble (Herz, 1985). The variations of the isotopes ^{13}C and ^{18}O are expressed as δ^{13}C and δ^{18}O, where

$$\delta(‰) = \left(\frac{R_{\text{sample}} - R_{\text{standard}}}{R_{\text{standard}}}\right) \times 1000 \qquad (6)$$

where $R = {}^{13}\text{C}/{}^{12}\text{C}$ or ${}^{18}\text{O}/{}^{16}\text{O}$ and the per mil variations are relative to the PDB (Pee Dee belemnite) fossil standard. Variations in these isotopes in the carbonate fraction of marble are due to the chemistry of the parent sedimentary rock, the composition of waters involved in the history of the rock, the temperature of metamorphism, and the later weathering of the rock. Measurement by mass spectrometry yields a precision of 0.2‰ or better. Within the Carrara district in Italy, the variation is less than 0.5‰ within a block, and less than 2‰ within an outcrop, whereas the variation between quarries is as much as 10‰ (Fig. 5).

3.2 Ancient Technology

Much archaeometric work on that most important of archaeological materials, pottery (Olin and Franklin, 1982), has involved

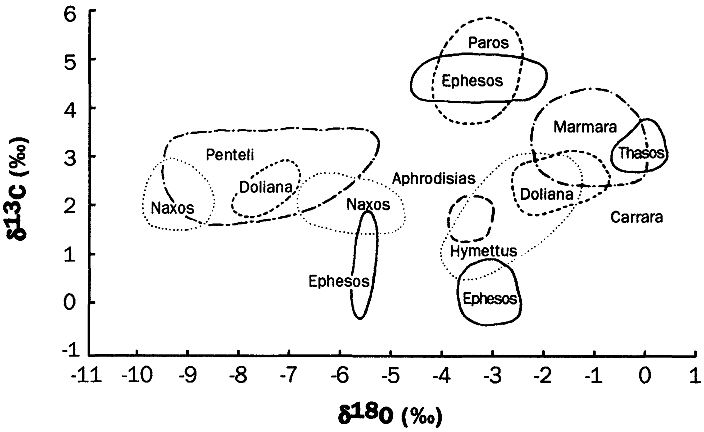

FIG. 5. Carbon and oxygen isotope fields for classical marble quarries (after Herz, 1985; by permission of Yale University Press).

sourcing the clay or temper according to chemical "fingerprints" (Perlman and Asaro, 1970; Hughes, 1991). Other ceramic studies have focused on the technology of production. One of the key aspects of production is firing temperature, which will depend on the method of firing and the fuel used. The final pottery product will depend upon the original raw materials as well as the temperature, time, and the atmosphere of firing. Three of the more traditional methods of assessing firing temperature are color evaluation, mineralogical analysis (because different minerals are stable over different temperature ranges), and measurement of thermal expansion (which changes markedly when the original temperature of firing is exceeded).

Newer, more technical, methods for inferring firing temperature include the use of scanning electron microscopy (SEM) and Mössbauer spectroscopy (Heimann, 1982; Maniatis *et al.*, 1982; Tite *et al.*, 1982). SEM uses a beam of electrons to scan the surface of the object under investigation. Backscattered electrons are detected in a density proportional to the atomic weight and hence to the elemental composition of the surface of the material. On the other hand, secondary electrons are emitted by the surface material and detected at a concentration depending on the angle at which the primary electrons strike the surface, thus giving an image of the surface's microstructure. In Mössbauer spectroscopy, a radioactive source, ^{57}Co, emits a γ ray as it decays to ^{57}Fe. These γ rays can be reabsorbed by ^{57}Fe in the sample. Because the source and target are both embedded in solids, the recoilless emitted and absorbed γ rays have small differences in their energies. Thus, the γ rays will only be absorbed if the source is moving relative to the sample, creating a Doppler effect. A measure of γ-ray absorption vs source speed indicates the transition energies of the sample's ^{57}Fe, which depends on the oxidation state (valence) of iron and its mineralogic form in ceramics. The iron mineralogy depends in turn on the temperature and atmosphere of firing as well as on the raw material. Thus, the Mössbauer results give clues to these aspects of the technology of pottery manufacture. Mössbauer spectra will vary as the sample is reheated in the laboratory (Fig. 6) because the iron oxidation state and mineralogy change. As an example of the joint application of these techniques, Maniatis *et al.* (1982) describe measurements on a small group of potsherds from a large Iraqi ceramic collection. The SEM patterns show the characteristic microstructure of calcareous clays fired between 850 and 1050 °C. Although microstructure does not vary within this temperature range, Mössbauer spectra can provide additional resolution. The sherds were refired to 950 °C, in the middle of the possible temperature range, in an oxidizing atmosphere. The spectra before and after refiring are the same, so the firing temperature for these sherds is narrowed to 950–1050 °C.

Glass, though technically a ceramic, is in many ways different from pottery. Henderson (1989) discusses the scientific analysis of ancient glass. One aspect of these studies is the nature and source of colorant for glasses. In the first detailed technical study of ancient manufactured materials, Sir Humphrey

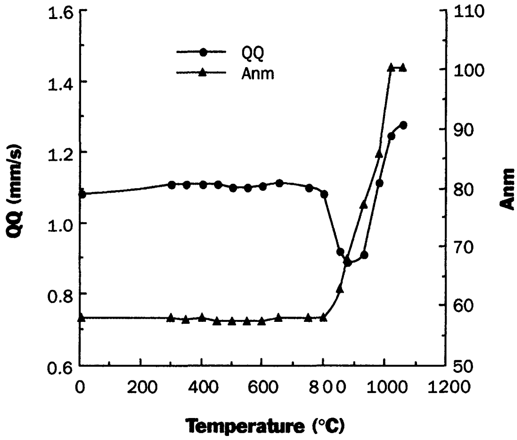

FIG. 6. The effects of laboratory reheating on Mössbauer spectra parameters for a brick from the Spandau Citadel, dating from A.D. 1540 (after Wagner *et al.*, 1986; by permission of Smithsonian Institution Press). The increase in the nonmagnetic Fe^{+3} quadrupole splitting (QQ) and the fractional intensity of the nonmagnetic Fe^{+3} and Fe^{+2} contributions (A_{nm}) indicate an original firing temperature of about 800 °C.

Davy, in 1815, suggested that cobalt was used to color ancient blue glass (Craddock, 1991). During the later Iron Age in Celtic Europe, from 500 B.C. to the Roman conquest, cobalt blue was one of the most common glass colors. Henderson (1989) cites several studies that considered whether cobalt-blue glass was actually made in late Iron Age Europe, or imported from the Mediterranean and then worked and shaped into the final product on site. This has technological implications: the temperature required to fuse silica and produce glass is 300–400°C higher than temperatures needed to work glass that has already been made. Many glass beads and armlets are found at the large European settlements of this era, and a high proportion of them are colored blue in the characteristic Celtic La Tène art style, suggesting local sources for the cobalt. Chemical analyses at several sites have also suggested that local ores could indeed provide the raw material for this colorant.

4. PALEOENVIRONMENT

Many paleoenvironmental questions concerning archaeological sites can be addressed through archaeological geology (Rapp and Gifford, 1985). These include studies of climate, ancient landscape, site formation processes, plant and animal resources, and diet. Paleoenvironmental problems are often studied through analysis of floral and faunal remains and geological studies of sediments and landforms. After brief comments on isotope paleoclimatology, this discussion will focus more narrowly on isotopic studies of paleodiet, rather than on a more general discussion of how the overall physical and biological environment can be reconstructed.

4.1 Paleoclimate

Aside from biological and geological reconstructions of climate, the major physicochemical methods used in paleoclimatology during the time period of interest to archaeology are oxygen and hydrogen isotope studies (Friedman, 1983; Herz, 1990). Isotope fractionation between ^{16}O and ^{18}O and between ^{1}H and ^{2}H is temperature dependent so that these isotope pairs can be used as paleothermometers. Oxygen isotopes in carbonates from harvested shells have been used in this way. Variations of $\delta^{18}O$ within seasonal bands of mollusc shells have been used to infer the seasonal occupation of sites. Isotope paleothermometry has also been applied to nonarchaeological materials that date to archaeological time, in particular sedimentary cores, glacial ice, cave speleothems, and tree rings. Complicating factors in isotope paleothermometry include varying isotopic composition of the original atmosphere and water (especially due to glaciation), isotopic composition of various shell minerals, and postdepositional chemical change (diagenesis).

4.2 Paleodiet

The most important light elements for dietary studies have been carbon and nitrogen, with the useful stable isotopes being ^{12}C, ^{13}C, ^{14}N, and ^{15}N. Nitrogen fractionation is presented as a per mil difference relative to the AIR (ambient inhalable reservoir) standard. Bone is the archaeological material commonly used for isotopic studies of diet because it is most often preserved; most analyses are made on the bone protein collagen because it is relatively stable.

Carbon isotopes were the first light element used for paleodiet studies (van der Merwe, 1982). Terrestrial plants metabolize carbon dioxide into more complex molecules by two different photosynthetic pathways that are significant in archaeology. The C_3 pathway, followed by most North American and European species, fractionates the stable isotope ^{13}C to an average value of $\delta^{13}C = -26.5‰$. Plants using the C_4 pathway, including maize and sugarcane, have less depleted $\delta^{13}C$ values of $-12.5‰$ on the average. Additional fractionation occurs as these foodstuffs move up the food chain, into the human diet, and are incorporated into human tissue. The fractionation between diet and human bone collagen for carbon seems to be about 3–5‰. This does not preclude inferences about major changes in diet but does complicate attempts to determine the proportions of different dietary components. Nonetheless, carbon isotopic ratios in collagen do appear to be directly related to the consumer's diet. Van der Merwe (1982) presents an example of how carbon isotopes can be used: $\delta^{13}C$ values in human bone collagen from peoples of the North American woodlands remained at about $-21.5‰$ from 4000 B.C. to 1 A.D., then increased to $-12‰$ by A.D. 1200. Thus, the proportion of carbon in the collagen from C_4 plants went from 0 to over 70%, indicating the introduction of and dramatic increase in the importance of the cultivation of maize in the diet of these people. This dietary change was accompanied by other major cultural changes, including increases in population and changes in settlement patterns.

Analysis of a single isotope does not necessarily yield a unique diet since different combinations of foodstuffs can yield the same isotope values. Use of carbon and nitrogen isotopes together extends the range of problems that can be addressed (DeNiro, 1987; Herz, 1990; Schoeninger and Moore, 1992). For nitrogen isotopes, the $^{15}N/^{14}N$ ratios vary between legumes and nonlegumes; legumes can directly fix atmospheric nitrogen, while nonlegumes must derive their nitrogen from the soil. Prehistoric legumes probably had $\delta^{15}N$ values of $+1‰$, while nonlegumes were most likely about $+9‰$.

DeNiro (1987) describes how isotopic dietary information can be "scraped" from a potsherd. In one of his studies, carbon and nitrogen isotopes were used together to show that the dietary importance of maize and legumes varied over time for some sites in the central Peruvian Andes. Residues from cooking ware indicated that the pottery was not used to cook legumes. Measurements in that study were made on the plant remains themselves.

The isotopic composition of carbon and nitrogen isotopes in human bones has also been used. In both terrestrial and marine environments, passing up the food chain to higher trophic levels increases fractionation of $\delta^{15}N$ values. Fractionation into human bone collagen typically increases $\delta^{15}N$ values by 3‰ relative to diet. Terrestrial C_3 and C_4 agricultural products have $\delta^{15}N$ values about 5‰ lower than marine foods. Nitrogen isotopic signatures in bone thus appear to be useful as paleodiet indicators in some, but not all, ecosystems. The simpler relationships seem to break down, for example, in arid environments, where water and/or protein stress may be factors. Figure 7 illustrates how carbon and nitrogen isotopes are useful in distinguishing between several diets from North American cultures (Schoeninger and Moore, 1992). The marine diet from the northwest coast is reflected in distinctly higher $\delta^{15}N$ values than the other diets shown. Some of the variation in this cluster of data is due to a temporal change in diet, with increasing reliance on marine foods through the late prehistoric period. The high $\delta^{13}C$ values for Pecos Pueblo are due to the dietary importance of maize. The Ontario population shows slightly elevated $\delta^{15}N$ values due to fish in the diet, and lower premaize $\delta^{13}C$ values. Bones from the people of St. Catherine's Island, Ontario, show intermediate $\delta^{13}C$ values due to the mix of C_4 values from maize and C_3 from deer meat and varying $\delta^{15}N$ values because of some fish in the diet.

Trace elements in bone and teeth have also been used to study paleodiet (Roberts, 1991). Strontium, like calcium, is an alkaline-earth element, and so it can substitute for calcium in trace amounts. Strontium can be a useful dietary discriminant because its concentration is high in plants, nuts, and marine invertebrates, but low in meat. It can thus be used to indicate the main food source for protein. Strontium concentrations, typically about 1000 ppm, can be measured by chemical methods such as inductively

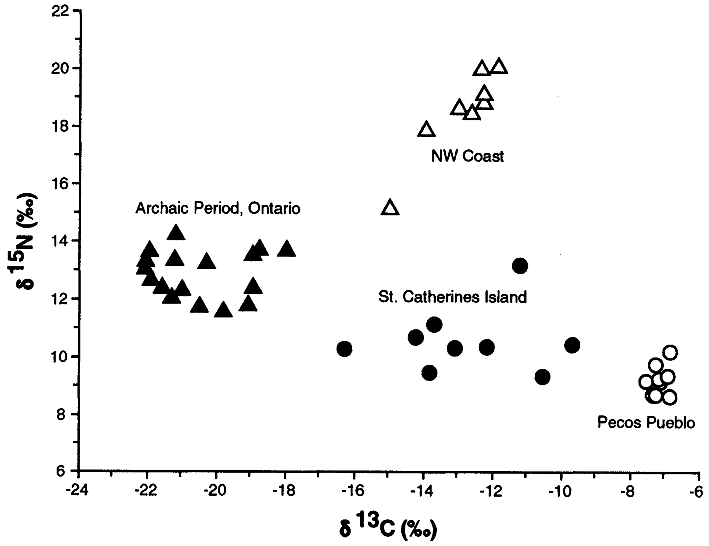

FIG. 7. Nitrogen and carbon isotope fields for several North American diets (after Schoeninger and Moore, 1992; by permission of Plenum Press). The major components of the various diets are as follows: NW coast, nonmaize plants and salmon; Archaic period, Ontario, nonmaize plants and freshwater fish; St. Catherine's Island, maize and coastal marine; Pecos Pueblo, maize and mule deer.

coupled plasma spectroscopy (ICPS). As with most chemical methods, contamination is a potential problem, in this case due to diagenesis from leaching or deposition by groundwater. Various cleaning methods are being investigated to minimize this overprint, a process quite similar in principle to demagnetization in archaeomagnetic studies.

4.3 Organic Residues

Analysis of other organic residues from pottery and tools is becoming a very promising area of research (DeNiro, 1987; Loy, 1991; Roberts, 1991). Materials recovered include blood, paint, tissue, plant remains, and hair. Identification of these residues and consideration of their preservation has accounted for much of this research so far. Some of these trace residues can also be dated by AMS ^{14}C dating. In the future, DNA extracted from skeletons of varying ages will be compared for genetic similarity or distance due to evolution or for the presence or absence of genetic diseases.

GLOSSARY

Accelerator Mass Spectrometry (AMS): Mass spectrometry after accelerating the ions to be measured, used for direct counting of ^{14}C in radiocarbon dating.

Anomaly: Unusually high or low measurement in geophysical prospecting due to a change in the subsurface material, indicating for example the presence of an archaeological feature.

Archaeomagnetism: The paleomagnetic study of archaeological materials, or rocks dating to archaeological time, providing the basis of archaeomagnetic dating.

Archaeometry: The application of the natural sciences to archaeology.

Chronometric Dating: A dating method using archaeometric techniques.

Electron-Spin Resonance: The measurement of electrons trapped in crystalline defects by their absorption of microwaves in the presence of a magnetic field, used as the basis for ESR dating.

Geophysical Prospecting: The use of the methods of physics to make measurements on the earth's surface in order to infer subsurface materials or properties, including the presence of archaeological features.

Mass Spectrometry: The measurement of isotopic masses after separating moving ions in a magnetic field.

Mössbauer Spectroscopy: A resonance technique utilizing the spectrum of γ rays absorbed from a moving source by a recoilless target; used to infer the iron mineralogy and temperature of firing in archaeological materials.

Paleoenvironment: The environment (e.g., climate, landscape, resource availability) at some time past.

Paleomagnetism: The study of the history of the earth's magnetic field, including the subfield of archaeomagnetism.

Provenance: The place of origin of a raw material.

Radiocarbon Dating: Dating of geological or archaeological materials using the radioactive isotope ^{14}C.

Scanning Electron Microscopy: The study of a surface's composition or microstructure by measuring backscattered or emitted secondary electrons resulting from a beam of incident electrons.

Spectroscopy: The study of matter by the wavelengths of radiation emitted or absorbed by atoms or nuclei, forming the basis for most chemical analysis techniques in archaeology.

Thermoluminescence (TL): The emission of light when electrons are ejected from traps in crystalline defects by heating an artifact, used as the basis for TL dating.

Thermoremanent Magnetization (TRM): Permanent magnetization acquired when a rock or archaeological object is heated above its Curie temperature.

Works Cited

Aitken, M. J. (1985), *Thermoluminescence Dating*, London: Academic Press.

Aitken, M. J. (1990), *Science-based Dating in Archaeology*, London: Longman.

Bishop, W. W., Miller, J. A. (Eds.) (1972), *Calibration of Hominoid Evolution*, Edinburgh: Scottish Academy Press.

Breiner, S., Coe, M. D. (1972), *Am. Sci.* **60**, 566–575.

Brock, A., Isaac, G. L. (1974), *Nature* **247**, 344–348.

Burger, H. R. (1992), *Exploration Geophysics of the Shallow Subsurface*, Englewood Cliffs, NJ: Prentice Hall.

Clark, A. J. (1986), *Geophysics* **51**, 1404–1413.

Craddock, P. (1991), in: S. Bowman (Ed.), *Science and the Past*, Toronto: University of Toronto Press, pp. 11–15.

Creer, K. M., Tucholka, P., Barton, C. E. (Eds.) (1983), *Geomagnetism of Baked Clays and Recent Sediments*, Amsterdam: Elsevier.

Damon, P. E., Donahue, D. J., Gore, B. H., Hatheway, A. L., Jull, A. J. T., Linick, T. W., Sercel, P. J., Toolin, L. J., Bronk, C. R., Hall, E. T., Hedges, R. E. M., Housley, R., Law, I. A., Perry, C., Bonani, G., Trumbore, S., Woelfli, W., Ambers, J. C., Bowman, S. G. E., Leese, M. N., Tite, M. S. (1989), *Nature* **337**, 611–615.

Dean, J. S. (1978), in: M. B. Schiffer (Ed.), *Advances in Archaeological Method and Theory*, Vol. 1, New York: Academic Press, pp. 223–265.

DeNiro, M. J. (1987), *Am. Sci.* **75**, 182–191.

Downey, W. S., Tarling, D. H. (1984), *Nature* **309**, 519–523.

Eighmy, J. L., Sternberg, R. S. (Eds.) (1990), *Archaeomagnetic Dating*, Tucson: The University of Arizona Press.

Friedman, I. (1983), in: S. C. Porter (Ed.), *Late-Quaternary Environments of the United States: Vol. 1, The Late Pleistocene*, Minneapolis: University of Minnesota Press, pp. 385–389.

Gale, N. H. (1989), in: Y. Maniatis (Ed.), *Archaeometry: Proceedings of the 25th International Symposium*, Amsterdam: Elsevier, pp. 469–502.

Hardy, D. A., Renfrew, A. C. (Eds.) (1989), *Thera and the Aegean World III: Chronology: Proceedings of the Third International Conference*, London: The Thera Foundation.

Haynes, C. V. (1992), in: R. E. Taylor, A. Long, R. S. Kra (Eds.), *Radiocarbon After Four Decades: An Interdisciplinary Perspective*, New York: Springer-Verlag.

Heimann, R. B. (1982), in: J. S. Olin, A. O. Franklin (Eds.), *Archaeological Ceramics*, Washington, DC: Smithsonian Institution Press, pp. 89–96.

Henderson, J. (1989), in: J. Henderson (Ed.), *Scientific Analysis in Archaeology*, Oxford: Oxford University Committee for Archaeology and UCLA Institute of Archaeology, pp. 30–62.

Herz, N. (1985), in: G. Rapp, J. A. Gifford (Eds.), *Archaeological Geology*, New Haven: Yale University Press, pp. 331–351.

Herz, N. (1990), *Geologic. Soc. Am. Centennial Special Volume* **4**, 585–595.

Hughes, M. (1991), in: S. Bowman (Ed.), *Science and the Past*, Toronto: University of Toronto Press, pp. 99–116.

Ikeya, M., Miki, T. (Eds.) (1985), *ESR Dating and Dosimetry (Proceedings of the First International Symposium on ESR Dating)*, Tokyo: Ionics.

Jones, M. (Ed.) (1990), *Fake? The Art of Deception*, University of California Press: Berkeley.

Kearey, P., Brooks, M. (1984), *An Introduction to Geophysical Exploration*, Oxford: Blackwell Scientific Publications.

Leute, U. (1987), *Archaeometry: An Introduction to Physical Methods in Archaeology and the History of Art*, Weinheim, Germany: VCH.

Libby, W. F. (1955), *Radiocarbon Dating*, Chicago: University of Chicago Press.

Loy, T. H. (1991), in: E. Pernicka, G. A. Wagner (Eds.), *Archaeometry '90*, Basel: Birkhäuser Verlag, pp. 645–656.

Maniatis, Y., Simopoulos, A., Kostikas, A. (1982), in: J. S. Olin, A. O. Franklin (Eds.), *Archaeological Ceramics*, Washington, DC: Smithsonian Institution Press, pp. 97–107.

Marshall, E. (1990), *Science* **249**, 738–741.

Olin, J. S. (Ed.) (1982), *Future Directions in Archaeometry: A Round Table*, Washington, DC: Smithsonian Institution.

Olin, J. S., Franklin, A. O. (Eds.) (1982), *Archaeological Ceramics*, Washington, DC: Smithsonian Institution Press.

Olsson, I. U. (Ed.) (1970), *Radiocarbon Variations and Absolute Chronology*, Nobel Symposia, Stockholm: Almqvist and Wiksell.

Parkes, P. A. (1986), *Current Scientific Techniques in Archaeology*, New York: St. Martin's Press.

Perlman, I., Asaro, F. (1970), in: R. Berger (Ed.), *Scientific Methods in Medieval Archaeology*, Berkeley: University of California Press, pp. 389–408.

Rapp, G., Gifford, J. A. (Eds.) (1985), *Archaeological Geology*, New Haven: Yale University Press.

Roberts, C. A. (1991), in: P. Budd, B. Chapman, C. Jackson, R. Janaway, B. Ottaway (Eds.), *Archaeological Sciences 1989*, Oxford: Oxbow Books, pp. 373–385.

Schoeninger, M. J., Moore, K. (1992), *J. World Prehistory* **6**, 247–296.

Scollar, I., Weidner, B., Segeth, K. (1986), *Geophysics* **51**, 623–633.

Shaw, J. (1979), *Geophys. J. R. Astron. Soc.* **58**, 107–116.

Sternberg, R. S. (1990), in: J. L. Eighmy, R. S. Sternberg (Eds.), *Archaeomagnetic Dating*, Tucson: University of Arizona Press, pp. 5–28.

Sternberg, R. S., Lange, R. C., Murphy, B. A., Deaver, W. L., Teague, L. S. (1991), in: E. Pernicka, G. A. Wagner (Eds.), *Archaeometry '90*, Basel: Birkhäuser Verlag, pp. 597–606.

Sternberg, R. S., McGuire, R. H. (1990), in: J. L. Eighmy, R. S. Sternberg (Eds.), *Archaeomagnetic Dating*, Tucson: University of Arizona Press, pp. 199–225.

Stuiver, M., Braziunas, T. F., Becker, B., Kromer, B. (1991), *Quaternary Res.* **35**, 1–24.

Stuiver, M., Kra, R. (Eds.) (1986), *Radiocarbon: Calibration Issue, Proceedings of the 12th International ^{14}C Conference*, New Haven: American Journal of Science.

Stuiver, M., Reimer, P. J. (1986), *Radiocarbon* **28**, 1022–1030.

Tarling, D. H. (1983), *Palaeomagnetism: Principles and Applications in Geology, Geophysics, and Archaeology*, London: Chapman and Hall.

Taylor, R. E. (1987), *Radiocarbon Dating: An Archaeological Perspective*, London: Academic Press.

Taylor, R. E., Long, A., Kra, R. S. (Eds.) (1992), *Radiocarbon After Four Decades: An Interdisciplinary Perspective*, New York: Springer-Verlag.

Taylor, R. E., Payen, L. A. (1979), in: M. Schiffer (Ed.), *Advances in Archaeological Method and Theory*, Vol. 2, New York: Academic, pp. 239–283.

Tite, M. S., Freestone, I. C., Meeks, N. D., Bimson, M. (1982), in: J. S. Olin, A. O. Franklin (Eds.), *Archaeological Ceramics*, Washington, DC: Smithsonian Institution Press, pp. 109–120.

van der Merwe, N. J. (1982), *Am. Sci.* **70**, 596–606.

Wagner, U., Wagner, F. E., Riederer, J. (1986), in: J. S. Olin, M. J. Blackman (Eds.), *Proceedings of the 24th International Archaeometry Symposium*, Washington DC: Smithsonian Institution Press, pp. 129–142.

Wynn, J. C. (1986), *Geophysics* **51**, 533–537.

Wynn, J. C. (1990), in N. P. Lasca, J. Donahue (Eds.), *Archaeological Geology of North America*, Boulder: Geological Society of America, pp. 603–617.

Young, C. T., Droege, D. R. (1986), *Geophysics* **51**, 568–575.

Further Reading

Aitken, M. J. (1974), *Physics and Archaeology*, Oxford: Clarendon Press.

Bradley, R. S. (1985), *Quaternary Paleoclimatology: Methods of Paleoclimatic Reconstruction*, Boston: Allen & Unwin.

Brothwell, D., Higgs, E. (Eds.) (1969), *Science in Archaeology*, London: Thames and Hudson.

Budd, P., Chapman, B., Jackson, C., Janaway, R., Ottaway, B. (Eds.) (1991), *Archaeological Sciences 1989*, Oxford: Oxbow Books.

Clark, A. (1990), *Seeing Beneath the Soil: Prospecting Methods in Archaeology*, London: B. T. Batsford Ltd.

Damon, P. E., Long, A., Kra, R. S. (Eds.) (1992), *Radiocarbon: Proceedings of the Workshop, Paleoastrophysics and Natural Variations of Cosmogenic Isotopes, 14th International Radiocarbon Conference*, Tucson: The University of Arizona.

Lamb, H. H. (1985), *Climatic History and the Future*, Princeton: Princeton University Press.

Long, A. (Ed.) (1989), *Radiocarbon: Proceedings of the 13th International Radiocarbon Conference*, Tucson: The University of Arizona.

Long, A. (Ed.) (1992), *Radiocarbon: Proceedings of the 14th International Radiocarbon Conference*, Tucson: The University of Arizona.

Maniatis, Y. (Ed.) (1989), *Archaeometry: Proceedings of the 25th International Symposium*, Amsterdam: Elsevier.

Mommsen, H. (1986), *Archäometrie: Neure naturwissenschaftliche Methoden und Erfolge in der Archäologie*, Stuttgart: B. G. Teubner.

Olin, J. S., Blackman, M. J. (Eds.) (1986), *Proceedings of the 24th International Archaeometry Symposium*, Washington, DC: Smithsonian Institution Press.

Pernicka, E., Wagner, G. A. (Eds.) (1991), *Archaeometry '90: Proceedings of the 27th International Symposium*, Basel: Birkhäuser Verlag.

Scollar, I. (1990), *Archaeological Prospecting and Remote Sensing*, Cambridge, UK: Cambridge University Press.

Stuiver, M., Kra, R. (Eds.) (1986), *Radiocarbon: Proceedings of the 12th International ^{14}C Conference*, New Haven: American Journal of Science.

Stuiver, M., Reimer, P. J. (1986), *Radiocarbon* **28**, 1022–1030.

Tite, M. (1972), *Methods of Physical Examination in Archaeology*, Academic Press.

Weymouth, J. W. (1986), *Geophysics* **51**, 538–552.

PHYSICS LITERATURE AND PUBLICATIONS

Rita G. Lerner,* *American Institute of Physics (ret.), College Park, Maryland, U.S.A.*

	Introduction	87
1.	**Types of Literature**	88
1.1	Journals	88
1.1.1	Journal Costs	89
1.1.2	Bibliometrics	89
1.2	Abstracts	90
1.3	Review Literature	90
1.4	Books	90
1.5	Conference Proceedings	91
1.6	Report Literature	91
1.7	Information Exchange Groups	91
2.	**Data Compilations**	92
2.1	Information Analysis Centers	92
2.2	CODATA	92
2.3	Symbols, Units, and Nomenclature	93
3.	**Electronic Publishing**	93
3.1	Computer Composition	93
3.2	Electronic Journals	93
4.	**Document Delivery**	94
5.	**Copyright**	94
6.	**Summary**	95
	Works Cited	95
	Further Reading	95

INTRODUCTION

Books on physics are quite old as a means of communication; they date as least as far back as Aristotle. Journals, on the other hand, are a relatively recent development. The first scientific and technical journals are only a little more than 300 years old. The *Journal des Sçavans* (later called the *Journal des Savants*) was founded in 1665 by Denys de Sallo in France; it provided summaries of important books and reports on philosophical, scientific, and literary research. The *Philosophical Transactions of the Royal Society* in England also first appeared in 1665; in 1887 it was split into two series, with Series A containing the physics and mathematics articles, and is still being published. In Germany, the *Acta Eruditorum*, founded in 1682, published contributions from Leibniz. The first Russian periodical was *Yezhemyesyatchnüya Sochineniya*, published by the Academy of Sciences from 1755 to 1764. Ever since, the formal communication of research information has been carried on through the medium of the printed journal.

The journal literature of physics serves several important functions. It is a public archive and it establishes priority for the authors. It is also a means of controlling quality through the referee system (as the late Simon Pasternack, an editor of *The Physical Review*, remarked, the information store is like a gigantic grocer's warehouse, filled with everything from gourmet goodies to rotten materials that will make one ill if ingested). The referee system for journals, which relies on volunteers, is a means of ensuring the quality of the archive. (Book publishers usually pay an honorarium for a review.)

It was not until after the middle of the nineteenth century, when the first physical societies were established, that specialized journals devoted exclusively to physics appeared. *The Physical Review* was published at Cornell University in 1893 by E. L. Nichols and Ernest Merritt. It was published with "the cooperation of the American Physical Society" in 1903, and published by The American Physical Society (founded in 1899) only starting in 1913.

As the number of physicists grew, and publication expanded, societies formed special interest groups and journals began to specialize by property, techniques, phenomena, etc. Journals and journal articles proliferated, and the need for abstract journals became apparent. One of the first abstract journals, the *Journal du Journal*, was published in 1670. The number of abstracting and indexing services grew as journal publi-

---

*Deceased.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

cation increased. By 1975, there were 330 abstracting and indexing services in the United States alone, covering all fields and processing nearly 3 000 000 items per year.

1. TYPES OF LITERATURE

The literature of physics can be divided into many formats: journals and reprints, abstracts, reviews, monographs, reference books, conference proceedings, report literature, and preprints (information exchange groups).

1.1 Journals

It has been estimated (King and Roderer, 1982) that there were as of that date about 450 journals devoted exclusively to physics, of which 150 are published in the United States; about 23% of the U.S. physics journals are published by scientific societies, 57% are published by commercial publishers, and 20% are published by other organizations, such as educational institutions, government agencies, and companies. In a study of the 2000 journals scanned for *Physics Abstracts* (Singleton, 1976), it was found that almost half of the journals were published by learned societies and somewhat more than half by commercial publishers. Publishers in the United Kingdom and the Netherlands dominate the European scene, particularly Elsevier/North Holland and the Institute of Physics, while the not-for-profits in the United States and Japan account for most of their journals. Commercial organizations publish a slightly larger number of journals, but the not-for-profit organizations publish a far larger number of articles, partly as a result of the publishing programs of the American Institute of Physics and its member societies, the Institute of Physics in the United Kingdom, and the Japanese physical societies. The American Institute of Physics and its member societies, with their 80 journals, now publish more than one-quarter of the world's research literature in all fields of physics.

Most journals are now available both in printed form and on microfilm or microfiche. Although these are not popular formats with readers, libraries find them useful as a means of saving scarce library storage space, and also saving on the cost of binding back copies. CD-ROM is a more expensive process, since it requires that the manuscript be in computer-readable form; the advantage is the capability of browsing through large collections of articles.

A major difference between commercial and society publishers is in the area of growth and capital investment. The not-for-profit societies usually have no risk capital for new ventures, and are frequently hard put to find money for research, development, and investment in new equipment to replace that outmoded by advancing technology, without government support in the form of grants and contracts. Another difference between society and commercial publishers is that learned societies tend to create new journals by dividing journals of broad scope into more specialized ones. Commercial publishers are more active in seeking wholly new journals, either by starting them or by publishing journals for the smaller societies.

Table 1 shows the number of all U.S. scientific and technical periodicals by type of publisher in 1977 as estimated (King *et al.*, 1981).

Among the largest journals in terms of the number of articles published per year are the *Physical Review* (American Physical Society), the *Journal of Chemical Physics* and the *Journal of Applied Physics* (American Institute of Physics), the *Journal of Physics* (Institute of Physics), and *Nuclear Physics* and *Nuclear Instruments and Methods* (North Holland), all of which have become so large that they have been split into sections. As journal size increased, the concomitant delay in publication gave rise to letters journals such as *Physical Review Letters*, *Applied Physics Letters*, *Physics Letters*, etc.; these offered rapid publication of short reports of important results.

Table 1.

	Journals	Other periodicals (bulletins, etc.)	Total
Societies	1734	1653	3387
Commercial publishers	1557	1072	2629
Educational institutions	1247	800	447
Other	356	1296	1652

An analysis of the journal literature of physics (Keenan and Atherton, 1964) based on 1961 *Physics Abstracts* found that 20 287 abstracts appeared in 405 journals published in 39 countries. Of these, 13 874 were in English, 3087 in Russian, 1472 in French, and 1304 in German. Articles in the major Japanese journals were published in English. The remainder of the articles were scattered among other languages, including Spanish, Polish, Ukrainian, and Greek. By 1992, the number of abstracts published had escalated to 151 677, reflecting the increase in the number of physicists doing research as well as the proliferation of primary journals.

1.1.1 Journal Costs Table 2 shows the revenue and expense of publishing two journals of 6000–8000 pages and 3000–4000 subscribers; Journal A has page charges and Journal B has advertising income but no page charges.

The first four expense items are fixed costs, which are incurred no matter how many copies of a journal are printed. The miscellaneous category includes microfilm and microfiche, storage, insurance, back numbers, and so forth. The main source of income is from subscriptions, although some journals have additional income from advertising. Voluntary page charges were introduced by The American Physical Society for *The Physical Review* in February 1930 (Barton, 1963), as a means of keeping subscription costs low for members. At first, they were only $2 per page, but rising costs have brought them to $70 or more per page. Page charges were requested from the institutions supporting the reported research, and were intended to cover the fixed costs of preparing a manuscript for publication; member subscribers pay only for the runoff cost of printing and distributing their copies of the journal, and do not contribute to the editorial and composition costs. Page charges are more common in the United States, where they are used by professional societies and nonprofit organizations, than elsewhere in the world.

Table 2.

	Journal A	Journal B
Revenues		
Page charges	15%	—
Advertising income	—	3%
Reprint sales	6	6
Subscriptions	77	89
Miscellaneous income	2	2
Expense		
Composition	18%	22%
Editorial mechanics	14	17
Editorial management	17	11
Administration	15	12
Printing and binding	9	14
Reprint print and bind	6	4
Paper	7	6
Mailing, postage	11	10
Fulfillment	1	1
Marketing	2	3

Since the late 1960s, most publishers have experienced a steady decline in the number of subscriptions to their printed journals. As subscriptions decrease, the unit cost of the remaining subscriptions increases, since the fixed expenses are averaged over a smaller number of subscriptions. Several factors are responsible for the decline. Libraries are canceling multiple subscriptions and are relying on interlibrary loan for the less frequently read journals. Physicists are relying less on personal subscriptions and more on library subscriptions, preprints, reprints, and photocopies. The number of pages published has increased for many journals; *Applied Physics Letters* published 1600 pages in 1977 and 4300 pages in 1987, and the *Physical Review* grew by 75% in 10 years. Some journals have reduced or eliminated page charges. To compensate, publishers have been forced to raise prices, and by 1987 nonmember subscriptions to physics journals, per 1000 characters published, varied from 0.39 to 31 cents for 1987 journals (Barschall, 1988).

Publishers have attempted to control production costs by accepting "compuscripts" or machine-readable manuscripts, and by optical character recognition systems, which scan manuscripts. The American Physical Society has developed a guide for submission of compuscripts (Sutherland, 1989), and 20% of *Physical Review* manuscripts are now submitted in REVTEX.

1.1.2 Bibliometrics The Institute for Scientific Information publishes an annual listing of the impact factors for all of the major scientific journals (except for translation journals), taken from the *Science Citation Index*. The "impact" of a journal is defined as the average number of citations in a given year to articles published in that journal in the previous 2 years.

In 1972 and 1973, the National Science Foundation supported a series of studies (Narin *et al.*, 1972; Narin and Carpenter, 1973) to develop a model of the interrelationship of scientific journals, based on citations to and from journals in the *Journal Citation Index*, a file derived from the *Science Citation Index*. Seventy-nine physics journals were in the database. The first study found that the journal cited most frequently was *The Physical Review*. Journals listed as "preferentially cited" were the *Astrophysical Journal*, the *Journal of Applied Physics*, the *Journal of Chemical Physics*, and the *Journal of Geophysical Research*. "Highly cited" journals were *Applied Physics Letters, Nuclear Physics, Proceedings of the Royal Society A, Physics of Fluids, Physics Letters, Physical Review Letters*, and *Soviet Physics—JETP*. The largest journals were found to be cited out of proportion to their size, perhaps reflecting a perception of quality. The second study found that a group of 81 physics journals formed distinct clusters based on citations to and from.

1.2 Abstracts

There are two comprehensive abstracting services in English. *Physics Abstracts*, a section of INSPEC, is published by the Institution of Electrical Engineers in Great Britain. *Physics Briefs* (formerly published in German as *Physikalische Berichte*) is compiled by the Fachinformationszentrum Energie, Physik, Mathematik and printed and distributed by VCH Publishers in Germany. Each of these services published approximately 150 000 items in 1992 in printed and electronic formats. *Bulletin Signaletique* and *Referativni Zhurnal*, which have physics sections, are the corresponding French and Russian services. *General Physics Advance Abstracts* covers research to be published in 46 primary journals of the American Institute of Physics and its member societies; it is available on PINET. The SPIN (Searchable Physics Information Notices) database includes abstracts from the AIP and member society journals as well as other American physics journals; it is available on the PINET and DIALOG information systems. *Current Physics Index* is a quarterly abstracts journal containing approximately 8000 articles from 53 journals of the American Institute of Physics and its member societies.

There are, of course, other services that overlap physics, such as *Chemical Abstracts Service* from the American Chemical Society and *Compendex* from Engineering Information, Inc., as well as specialized services such as *Solid State Abstracts*. Another form of secondary literature are titles journals, such as *Current Contents* from the Institute of Scientific Information, which lists the table of contents of journals. Many research laboratories circulate their own collections of tables of contents and abstracts.

1.3 Review Literature

Reviews are essential for the evaluation, synthesis, and consolidation of the literature. Prominent review journals include *Reviews of Modern Physics, Reports on Progress in Physics*, and *Uspekhi Fizicheski Nank*. There are a number of serials for different fields of physics such as "Advances in . . ." and "Progress in . . .". Reviews can vary from critical, to exhaustive, to bibliographies.

1.4 Books

> "The book, however it is formed, has been entertaining human beings for 3,000 years, is doing so now, and will do so 5,000 years from now, if we're still here."—Isaac Asimov (1920–1992)

Asimov added that the book is a portable device, and the ultimate technology for thought transfer. It is easily transported, private for the end user, kinetically activated but requiring no energy source, easily filed, and browsable so that the end user can go back and forth.

Research is disseminated through the journals, summarized in abstracts, and distilled into reviews, eventually appears in a monograph, and finally is captured in handbooks and encyclopedias. At one time, monographs were almost exclusively published by commercial publishers, such as John Wiley & Sons and Plenum Publishers in the United States, Elsevier in the Netherlands, and Springer-Verlag and VCH Publishers in Germany. Other major publishers of physics books include Cambridge University

Press and Oxford University Press in England, and Mir and Nauka in the former Soviet Union. Recently, the American Institute of Physics and the Institute of Physics in England have expanded their book publishing programs in order to provide books for members at a lower cost.

Dover Publications has had a long-standing program of reprinting some of the classic books of physics that were out of print. These include, among many others, works by Sir Lawrence Bragg, Paul Ehrenfest, Werner Heisenberg, Sir James Jeans, James Clerk Maxwell, Sir Isaac Newton, Max Planck, Erwin Schrödinger, and Sir Edmund Whittaker.

Two recent one-volume encyclopedias of physics are those edited by Besançon (Van Nostrand Reinhold) and by Lerner and Trigg (VCH Publishers). Multivolume works include the McGraw-Hill *Encyclopedia of Science and Technology* and the *Encyclopedia of Physical Science and Technology* from Academic Press. Handbooks of physics include the *AIP Handbook of Physics*, edited by Gray, and the *Handbook of Physics* edited by Condon and Shortley, although there have been no new editions of these two. The *Physicist's Desk Reference*, edited by Anderson, has been recompiled and is a concise collection of useful formulas, numerical data, definitions, and references. Springer-Verlag is the publisher of two well-known series of multivolume reference works: the *Handbuch der Physik/Encyclopedia of Physics* edited by Flügge and the *Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology*.

Dictionaries of physics have been compiled by Thewlis and by Isaacs; there is also an IEEE *Dictionary of Electronics* which is useful for physicists, and a *Dictionary of Space Technology* by Williamson.

Other useful general reference sources include the 1990 *AIP Style Manual* by Scott, the *Energy Source Book* edited by Howes, directories of members of various physics organizations, a *Dictionary of Scientific Quotations* edited by MacKay, and *Database Systems in Science and Engineering* by Rumble and Smith.

1.5 Conference Proceedings

In recent years, the number of scientific meetings has escalated so that there are many meetings every week in all parts of the world. Many of these meetings result in conference proceedings, which include congresses, workshops, symposia, seminars, and summer schools. In general, these papers are not refereed; they are reports of work in progress and may subsequently appear, in revised form, in a refereed journal or as a journal supplement. Most proceedings are produced from camera-ready copy supplied by the authors. Some proceedings are published as books, while others appear in journals.

Conferences that are held on a regular basis frequently change publishers, making it difficult to locate and obtain a copy of the proceedings. Conferences in physics are listed every two months in *Physics Today* with the name of a contact person; *Chemical and Engineering News* publishes a listing of conferences in chemistry and related fields every 6 months. Some journals, such as the *Review of Scientific Instruments*, publish announcements of conferences in their field. Published proceedings are listed in the Science, Engineering, Medicine, Technology section of the *Directory of Published Proceedings* (Interdok). One study (Barschall and Haeberli, 1992) found that one publisher produced one-quarter of the physics proceedings listed by Interdok; 10% of the total listed were proceedings of NATO workshops. They also found that the cost of individual proceedings ranged from free volumes to participants to 50 cents per page, with an average of about 15 cents per page.

1.6 Report Literature

Reports are sometimes referred to as the "grey literature" because of the difficulty of obtaining copies. They are unrefereed publications. In the United States, reports resulting from research supported by the U.S. government are distributed in print form or microfiche by the Clearinghouse for Scientific and Technical Information. The National Aeronautics and Space Administration supports STAR, Scientific and Technical Aerospace Reports.

1.7 Information Exchange Groups

Information exchange groups are usually informal arrangements among research groups to exchange preprints and reports. A well-known group is the SLAC Preprint Exchange, which issues lists of preprints received and an "anti-preprint" list of articles that have been published indicating which preprints can be removed from the collection. An "e-print" (electronic preprint) archive is hosted by the Los Alamos National Laboratory, the Scuola Internazionale Superiore di Studi Avanzati in Trieste, the Supercomputer Computations Research Institute at Florida State University, and Duke University. Scientists can acquire abstracts and preprints by e-mail or ftp, have title and author abstracts of any new papers automatically sent to their e-mail mailbox, and contribute their own papers to the mailing list of their choice. The first of this series of electronic bulletin boards was Hep-th (High Energy Particle Theory) set up by Ginsparg at the Stanford Linear Accelerator Center.

At one time, the U.S. government (particularly the National Institutes of Health) supported some information exchange groups, but the groups dissolved when support was withdrawn.

2. DATA COMPILATIONS

Numerical/factual data have been defined by Kaufman (1993) as "quantitative data with the supporting textual facts; it is not . . . limited to numeric data, but . . . it includes additional factual information in whatever format needed to describe the usefulness or limitations of the numeric part of the data." The additional information may include the significance level of the data, and the conditions and methods of measurement. Quantitative scientific and technical data usually appear in tables and graphs, and come in ranges as well as discrete values and a variety of units, making it difficult to capture the information in a form suitable for computer-searchable databases.

Earlier in this century, data compilations had broad coverage. Publications included the *International Critical Tables of Numerical Data*, the *Landolt-Börnstein Tabellen*, and the *Tables Annuelles de Constantes et Données Numériques*. The amount of effort that went into these compilations is not likely to be repeated; as the amount of data grew, it became necessary to develop specialized tables and databases.

2.1 Information Analysis Centers

There are now a number of information centers, dedicated to critically evaluating data and compiling specialized numerical databases. Several of them are located within the Office of Standard Reference Data (OSRD) of the National Institute for Science and Technology, or are sponsored by the OSRD. OSRD, jointly with the American Chemical Society and the American Institute of Physics, publishes the *Journal of Physical and Chemical Reference Data*, which collects the results of critically evaluated data in various fields and also makes articles available as separates; the journal also occasionally publishes supplements that are too long for a journal article, such as the JANAF Thermochemical Tables. Some of these data collections are also available as computer databases. Other data collections include the *Nuclear Data Sheets* and occasional publications such as the *Particle Properties Data Booklet* and the *Review of Particle Properties* compiled at the Lawrence Berkeley Laboratory.

Other information analysis centers include those at Oak Ridge National Laboratory, the Center for Information and Numerical Data Analysis and Synthesis (CINDAS) at Purdue University, the Crystallographic Data Centre in Cambridge, England, and the Canadian Scientific Numeric Database Service (CAN/SND). STN International, an online database service, offers two data compilations: the Materials Property Data Network and the Chemical Property Data Network.

2.2 CODATA

The work of information centers in all fields is promoted and encouraged by the Committee on Data for Science and Technology (CODATA), established in 1966 by the International Council of Scientific Unions. CODATA's mission is to improve the quality, reliability, and accessibility of data of importance to science and technology. In 1986, CODATA released its new recommended values for the fundamental physical constants (Cohen and Taylor, 1986). The reevaluation was made necessary by the extraordinary amount of experimental and theoretical work published since the previous evaluation in 1973, such as the redefinition of the speed of light, the discovery of the quantum Hall effect, and the measurement of the electron anomalous magnetic moment.

2.3 Symbols, Units, and Nomenclature

The General Conference of Weights and Measures, established by international treaty, in 1960 founded the Système International d'Unités (International System of Units), known as SI. The system is based on seven units: length, mass, time, electric current, thermodynamic temperature, amount of substance, and luminous intensity; all other units are defined from these. The cgs system, based on the centimeter, gram, and second, is still used in some fields of physics. A revised set of symbols, units, and nomenclature was published by the International Union of Pure and Applied Physics in 1987 (IUPAP, 1987).

3. ELECTRONIC PUBLISHING

The term "electronic publishing" is used both for production of publications by computer composition and for electronic versions of publications.

3.1 Computer Composition

Many publishers are using computer composition to produce journals and books, and some are encouraging authors to submit computer-readable manuscripts. As equipment prices have come down, computers are also being used to scan and digitize figures and graphs, and to make up pages. These developments will improve the linkage between primary and secondary publishers, and between primary journals and numerical data centers.

The Standard Generalized Markup Language (SGML) has made it easier to exchange data. SGML is an international standard, implemented by a number of system vendors. It is platform independent, and provides the ability to link graphics and other nontext data to the text.

3.2 Electronic Journals

The electronic journal and the electronic library are logical consequences of the ability to create machine-readable full text. An electronic journal can be established by a group of authors with local or network access to a computer. An author would keyboard his article into a computer file, and authorize the editor to look at the file and copy it. The editor would select the referees, who are given access to the file. The referees would evaluate the article, entering their comments and suggestions into the file. The author may revise his article, and the editor enters it into the journal archive, which is a public file available to subscribers. The electronic journal would have no publication deadlines, and could publish new articles as they are accepted by the editor.

There have been several attempts to establish electronic journals. The best-known of these in the *Online Journal of Current Clinical Trials*, started in July 1992 by the American Association for the Advancement of Science and the Online Computer Library Center, Inc. It was the first peer-reviewed electronic journal to support graphs, tables, illustrations, and mathematical equations. Articles are published within 48 hours of acceptance. Authors have the option of having a short version of their article considered for publication in *Lancet*, the British medical journal. Beginning in 1994, abstracts from this journal will appear in the Index Medicus database.

After a pilot study in 1993, Springer-Verlag, AT&T Bell Laboratories, and the University of California at San Francisco (UCSF) have announced a 2-year joint experiment called Red Sage, which will deliver journals to departments at UCSF through the campus computer system, using AT&T's RightPages technology (Hoffman *et al.*, 1993). Copyright fees for printed copies will be recorded and processed by RightPages; each publisher will set its own prices for copying and collect fees through the system.

Other experimental scientific journal collections include the American Chemical Society's Chemical Journals Online (text only, no figures) and the Chemical Online Retrieval Experiment (CORE) at Cornell University. The University Licensing Program (TULIP), an Elsevier Science Publishers project, is another major electronic library.

There is even an online journal about online journals: the *Public-Access Computer Systems Review*, published by the University of Houston Libraries, and distributed free on the Internet.

4. DOCUMENT DELIVERY

Document delivery services have become the focus of much attention and concern among publishers and librarians because of the possible impact on the economics of journal publishing. Providing copies of documents has become a service provided not only by libraries but by online vendors and service organizations.

The Commission of European Communities explored the feasibility of a document delivery system, named ARTEMIS for Automatic Retrieval of Text from Europe's Multinational Information Service. ARTEMIS would accept a document in digital form (either text or facsimile), store it in a computerized database, and transmit it by telecommunications to the requester's terminal.

The European Space Agency conducted a pilot project, named APOLLO for Article Procurement with Online Ordering, based on the ARTEMIS recommendations. A consortium of American and European publishers is experimenting with a system similar to ARTEMIS, called ADONIS for Article Delivery Over Network Information Services. A commercial version of ADONIS began in 1991 with approximately 350 titles on CD-ROM. The American Institute of Physics and the American Chemical Society are also conducting experiments in online delivery of full text of journals, both facsimile and fully digitized.

5. COPYRIGHT

In 1976, the United States Congress passed the first major revision of the Copyright Act since 1909; the new law went into effect in 1978. The changes were necessary because of new communication technologies, including computers, photocopiers, television, and videotape recorders. Statutory copyright now rests automatically with the author as soon as the manuscript is "fixed in any tangible medium of expression, now known or later developed, from which [it] can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device" (Title 17, U.S. Code). There are two exceptions: "work made for hire," in which the employer holds the copyright, and works by federal government employees as part of their official duties (although in the latter case, copyright may exist outside the United States). Since 1989, when the United States joined the Berne Copyright Convention, copyright owners are no longer required to put a copyright notice on the work or register the work to claim copyright, although registration is necessary at a later time in case of infringement. Since the United States became a signatory to the Berne Copyright Convention and the Universal Copyright Convention, the same rights are extended to foreign authors as to domestic authors; this applies to all the countries adhering to these conventions.

The new law provides several basic rights to the copyright holder:

1. the right to authorize reproduction of the copyrighted material;
2. the right to prepare or authorize derivative works, such as abstracts;
3. the right to distribute copies of the work; and
4. the right to display the work publicly, e.g., in print or on a cathode-ray tube.

Transfer of copyright to a publisher is no longer automatic, but requires a written transfer. This transfer permits the publisher to distribute reprints and microform versions of the article, allow abstracting and indexing services to use material from the journals and books, authorize photocopying, and produce computer tapes of the material.

The development of microfilm as a substitute for interlibrary loan of printed publications raised copyright issues which led, in 1935, to the "Gentlemen's Agreement" between librarians and publishers on fair use. The subsequent proliferation of photocopies has exacerbated the continuing controversy on the limits of fair use, balancing the rights of the copyright holder against the right of access to information.

Most publishers belong to the Copyright Clearance Center in the United States, or similar organizations in other countries. A line is printed at the bottom of the first page of a journal article identifying the journal and giving a price for copying the article; the money is then sent to the Center, which distributes it to the publishers. Some of the larger commercial organizations have annual contracts that permit them to make lump-sum payments rather than keeping track of individual articles copied. National copyright offices may still have to determine whether delivery of documents on demand over computer networks constitutes a publication or a public performance.

6. SUMMARY

"They that weave networks shall be confounded." Isaiah, chapter 19, verse 9.

The worldwide growth of computer networks will make information readily available to those who have access to computer terminals. All of the necessary technology exists today to create an electronic library, but there are still legal, regulatory, and management issues to be resolved. These include problems of copyright, transborder data flow, telecommunications regulations, and standards. It remains to be seen whether an electronic library is economically viable—whether publishers can afford the startup and input costs, and whether users will be willing to pay the delivery costs.

Although some have predicted the demise of printed publications, they are not likely to be completely replaced by the electronic library. Printed journals and books have their own advantages, including browsability, portability, the ability to underline and write in margins, and accessibility to those without computers. There is certainly room for both modes of communication. The automobile may have replaced the horse, but the airplane has not replaced the automobile!

Works Cited

Barschall, H. H. (1988), *Phys. Today* **41** (7), 56–59.

Barschall, H. H., Haeberli, W. (1992), *Coll. Res. Lib.* **1992** (Nov.), 563–566.

Barton, H. A. (1963), *Phys. Today* **16** (6), 45–57.

Cohen, E. R., Taylor, B. N. (1986), The 1986 Adjustment of the Fundamental Physical Constants, *CODATA Bull.* No. 63.

Hoffman, M. M., O'Gorman, L., Story, G. A., Arnold, J. Q., Macdonald, N. H. (1993), *J. Am. Soc. Inf. Sci.* **44**, 446–452.

IUPAP (1987), Symbols, Units, Nomenclature and Fundamental Constants in Physics, *Physica* **164A**.

Kaufman, J. G. (1993), *J. Chem. Inf. Comput. Sci.* **33**, 2–5.

Keenan, S., Atherton, P. (1964), *The Journal Literature of Physics*, New York: American Institute of Physics (AIP/DRP PA1).

King, D. W., Roderer, N. K. (1982), *Phys. Today* **35** (10), 43–47.

King, D. W., McDonald, D. D., Roderer, N. K. (1981), *Scientific Journals in the United States*, Stroudsburg, PA: Hutchinson, Ross Publishing Co., p. 90.

Narin, F., Carpenter, M., Berlt, N. C. (1972), *J. Am. Soc. Inf. Sci.* **23**, 323–331.

Narin, F., Carpenter, M. (1973), *J. Am. Soc. Inf. Sci.* **24**, 425–436.

Singleton, A. (1976), *Phys. Bull.* **1976** (Sept.), 399–403.

Sutherland, P. C. (1989), *Physical Review Input Guide for TEX Author-Prepared Compuscripts*, New York: American Physical Society.

Further Reading

King, D. W., McDonald, D. D., Roderer, N. K. (1981), *Scientific Journals in the United States*, Stroudsburg, PA: Hutchinson, Ross Publishing Co.

Report of the APS Task Force on Electronic Information Systems (1991), *Bull. Am. Phys. Soc.* **36** (4), 1105–1107, 1119–1151.

Schauder, D. (1994), *J. Am. Soc. Info. Sci.* **45** (2), 73–100.

PHYSICS OF FLEXIBLE MEMBRANES

See STATISTICAL PHYSICS OF MEMBRANES AND LAMELLAR SYSTEMS

PHYSICS OF PHOTOGRAPHY

See TECHNOLOGY AND PHYSICS OF PHOTOGRAPHY

PHYSICS RESEARCH FACILITIES

Louis Ianniello, *Gaithersburg, Maryland, U.S.A.*

Introduction 97

1. **Condensed Matter Physics** .. 98

1.1 Neutron Sources 98

1.2 Synchrotron Radiation Sources 99

2. **Particle Physics** 100

2.1 CERN, European Laboratory for Particle Physics 101

2.2 Fermi National Accelerator Laboratory 101

2.3 Stanford Linear Accelerator Center 102

2.4 Alternating-Gradient Synchrotron 102

2.5 KEK, National Laboratory for High-Energy Physics 102

3. **Nuclear Physics** 102

3.1 Argonne Tandem/Linear Accelerator (ATLAS) 102

3.2 Tandem/Alternating Gradient Synchrotron 103

3.3 Relativistic Heavy-Ion Collider (RHIC) 103

3.4 Continuous Electron-Beam Accelerator Facility (CEBAF) . 103

3.5 Clinton B. Anderson Meson Physics Facility (LAMPF) 103

3.6 Bates Linear Accelerator Center 103

3.7 Indiana University Cyclotron Facility 103

3.8 National Superconducting Cyclotron Laboratory 104

3.9 Tri-University Meson Facility (TRIUMF) 104

3.10 UNILAC/SIS/ESR Accelerator Facility 104

3.11 Nuclear Structure Facility 104

4. **Plasma Physics/Fusion Research** 104

4.1 Princeton Plasma Physics Laboratory (PPPL) 105

4.2 Joint European Torus (JET) .. 105

Further Reading 105

INTRODUCTION

Many physics experiments today are required to be done on state-of-the-art, complex facilities that are very expensive and centrally located. The researcher, wherever located, travels to the facility, which is operated by a resident permanent staff. It is these facilities that will be discussed here. The host laboratory where the facility is located also has a staff that may conduct research at the facility, but the arrangements for research include the availability of the facility to qualified researchers regardless of their affiliation. In this category of physics research facilities are perhaps less than a hundred worldwide, depending on the definitions. This article will focus on the most important ones, representative of four different categories of physics research: condensed matter physics, nuclear physics, particle physics, and plasma physics.

A facility includes the machine or device that produces the neutrons, protons, photons, electrons, or other particles, together with the associated detectors and instruments. The basic device is originally constructed with only a few instruments and detectors. With time, the experimental program matures and new instruments are added by the research community. Many of the experiments are conducted by teams of researchers from many different institutions. Typically, frontier facilities under construction cost many hundreds of millions of dollars to construct and will require tens of millions of dollars to operate annually. The annual operating budget includes the electrical power, materials and supplies, and the staff to run the facility. The research to be done at the facility usually is funded separately. A researcher generally has to submit a proposal to obtain approval and an allocation of time

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

in these facilities. Every facility has its own set of rules and the prospective experimenter should contact the individual facility management to obtain specific guidelines on how to seek approval for research in any of the facilities discussed here.

1. CONDENSED MATTER PHYSICS

In the field of condensed matter physics, neutrons and photons are extremely useful as probes to study the structure and dynamics of condensed matter. Facilities built to produce neutrons and photons for condensed matter physics are at the same time useful for other fields of science such as materials research, biology, chemistry, and others. Neutron sources are reactors or spallation sources, and photon sources are called light sources. Each facility usually provides many beam lines individually instrumented for a particular class of experiments.

1.1 Neutron Sources

Modern high-flux neutron sources for beam research are of two types: (1) fission reactors, which produce neutrons through the use of fissionable fuel, and (2) accelerator-driven devices which produce neutrons through a spallation process when high-energy particles hit a heavy-element target such as tungsten.

When a new material is discovered, often its basic atomic and magnetic structures as well as the dynamics of the atoms and magnetic moments need to be established. Neutrons, with their unique capability to penetrate materials and scatter from magnetic moments and their ability to discern light elements, are extremely useful as probes to investigate these materials. An example of the use of neutrons in such investigations is the crucial experiments done with neutrons on high-temperature superconducting oxides. Neutrons were used to determine the oxygen positions and content, the role of phase stability and impurity phases, as well as the nature of the magnetic structure and lattice dynamics. Neutrons have also been important in studies of heavy-fermion systems, in obtaining information on their magnetic properties; of fullerenes, in determining the structure, the lattice and rotational dynamics, and phase transitions; and of quasicrystals, in determining the atomic and magnetic structure as well as the vibrational frequencies.

More common materials are also investigated with neutrons using their unique characteristics for probing special effects. For example, the characterization of the structural consequences of bone mineralization has provided insight on the role of calcification on the structure and strength of bone. Neutrons continue to provide unique information on hydrogen effects in metals. Important neutron experiments have been performed on a variety of glasses to understand the domain sizes and the dynamics of spin glasses and disordered ferromagnets, as well as dipolar glasses.

Neutrons are used in some of the most fundamental studies vital to understanding matter in the condensed state. One of the most fundamental systems that has been illuminated with neutron scattering is the liquid and solid phases of helium. Experiments have provided important information on the structures and parameters of the various phases of this simple system. Neutrons have also been essential to understanding low-dimensional (one- and two-dimensional) systems. Neutron scattering has shown that one-dimensional spin systems are excellent realizations of spin liquids, with qualitatively different properties depending on whether the spin is an integer or a half-odd-integer. In the area of layered and multilayered structures, neutron scattering has provided detailed information about the magnetic structures on the nanometer scale, showing, for example, that magnetic order can propagate through nonmagnetic layers and that giant magnetoresistance is not necessarily associated with antiferromagnetic couplings.

Other areas of research using neutrons include chemistry- and biology-related investigations. In chemistry, neutron diffraction is used to determine the structure of molecules and of molecular assemblies. The uniqueness of neutron diffraction lies in the particular distribution of neutron cross sections across the periodic table and with different isotopes. This aspect differs from x-ray diffraction and allows investigations of particular atoms. Inelastic neutron scattering in chemistry permits the study of the dynamics of molecules and molecular assemblies. Biological investigations using neutron probes are

Table 1. Reactor neutron sources.

Reactor	Location	Number of spectrometers	Neutron flux (10^{14} neutrons/cm^2 s)
High Flux Beam Reactor (HFBR)	Brookhaven National Laboratory (U.S.A.)	13	5
High Flux Isotope Reactor (HFIR)	Oak Ridge National Laboratory (U.S.A.)	9	11
National Institute of Standards and Technology (NIST) Reactor	NIST (U.S.A.)	12	4
High Flux Reactor Institute Laue Langevin (ILL)	Grenoble, France	32	12

increasing for many reasons. Neutrons are complementary probes and sometimes essential in the study of complex biological molecules. Isotope effects can be exploited, especially between hydrogen and deuterium with labeling experiments since hydrogen is the most abundant element in biological systems.

The importance of neutron diffraction and neutron scattering was recognized with the awarding of the 1994 Nobel Prize in physics to two of the developers of these techniques, Clifford Shull and Bertram Brockhouse. More than 40 years previously, the two researchers pioneered methods to investigate the properties of materials using neutrons, independently. Shull worked at Oak Ridge National Laboratory on techniques of neutron diffraction to determine the atomic structure in solids and then joined the staff at the Massachusetts Institute of Technology where he continued his research with neutron beams. Brockhouse did his early work at Chalk River, Canada, where he developed the technique of inelastic neutron scattering to study the vibrational structure of atoms in solids, and later moved to McMaster University.

There are advantages for each type of neutron source: reactor type or spallation type. Small-angle scattering is best done at a reactor cold source. The higher-energy neutrons readily available at a spallation source provide advantages for experiments at high pressure and with glass and liquid materials. Diffuse scattering to display planes in reciprocal space is well suited for pulsed spallation sources. However, small-crystal diffuse-scattering studies, especially near Bragg peaks, are more suited for a reactor source. Residual stress measurements and time-resolved studies can be done equally well at either source. Table 1 provides a sample listing of some of the most powerful reactor neutron sources. Table 2 provides a similar listing of the spallation neutron sources.

1.2 Synchrotron Radiation Sources

Photons emitted from synchrotron radiation sources have become extremely attractive probes for experiments, not only in physics, but also in biology, materials science, chemistry, and other technological fields. Synchrotron radiation results from charged particles traveling in curved paths. The electromagnetic radiation given off can be controlled by controlling the particle energy and other characteristics of the particle beam. By constraining a particle beam with a magnetic field in a storage ring, the radiation produced can be made extremely in-

Table 2. Spallation neutron sources.

Source	Location	Number of spectrometers	Proton beam power (kW)
Los Alamos Neutron Scattering Center (LANSCE)	Los Alamos National Laboratory (U.S.A.)	8	38
Intense Pulsed Neutron Source (IPNS)	Argonne National Laboratory (U.S.A.)	12	6
Pulsed Spallation Source (ISIS)	England	13	77

tense over a broad range of wavelengths. Radiation emitted from such storage rings ranges from the infrared through the visible and ultraviolet into the soft and hard x-ray regions. The advantage of synchrotron radiation over other sources of photons lies in the combination of high intensity, broad spectral range, collimation, pulsed time structure, and partial coherence for the wavelengths produced.

Some photon sources were designed for purposes other than synchrotron-radiation research devices and are called first generation sources. The sources specifically designed and optimized for dedicated synchrotron radiation research are referred to as second generation sources. Those designed later and optimized for dedicated synchrotron radiation research using insertion devices are termed third generation sources.

Many experimental stations are possible at a storage ring. Electrons or positrons are circulated in a closed orbit within a storage ring. Radiation is emitted from all curved parts of the orbit. Radiation from bending magnets and from specially designed magnetic insertion devices leave the ring through tangential ports called beam lines on which experimental stations are located. Usually the storage rings are optimized to produce radiation in either the ultraviolet or the x-ray region of the spectrum. Storage-ring sizes range from a few meters in circumference operating at a few hundred million electron volts to large rings, 1500 meters in circumference, operating at 8 GeV.

In addition to the usual diffraction, microscopy, and chemical analysis experiments, which can be done with greater precision, quicker, and on smaller samples, new experiments are possible on synchrotron radiation sources. Additionally, new technological applications such as x-ray lithography, micromachining, and x-ray angiography are possible.

For certain experiments, photon flux [photons/s·mrad·(unit bandwidth)] hitting the sample is all that is important. For others, brightness [photons/s·(unit solid angle)(unit source area)(unit bandwidth)] is increasingly becoming important. Time-lapse experiments require a certain time structure to the radiation. Electrons in a storage ring are bunched, causing light to be emitted in pulses. Each bunch may be typically 0.5 to 5 cm long, resulting in a pulse duration of 100 ps for a 3-cm-long bunch, as an example. The spacing, typically, between bunches can be adjusted to result in light pulses from a few picoseconds to a microsecond apart.

There are more than 30 rings operating worldwide. Table 3 provides a sample listing of some of the most powerful x-ray rings. Table 4 contains a similar list for ultraviolet/soft x-ray rings.

2. PARTICLE PHYSICS

Experimental facilities to accelerate, collide, and detect nature's smallest particles have enabled the scientific community to study and understand the basic constituents of matter. Particle accelerators, colliders, and detectors have increased in size, energy, and sophistication in order to peel back successive layers of the building blocks of matter from the atom to the electron, proton, and neutron to the subnuclear particles to the quarks. Observing these fundamental constituents has enabled physicists to put together a picture or model of the known basic constituents of matter, the forces that control their behavior, and the carrier particles

Table 3. Synchrotron radiation sources (x-rays).

Ring	Location	Energy (GEv)	Ring generation
SPEAR	Stanford, CA (U.S.A.)	3.5	First
CESR	Ithaca, NY (U.S.A.)	5.5	First
SRS	Daresbury, England	2	Second
Photon Factory	Tsukuba, Japan	3	Second
NSLS II	Brookhaven, NY (U.S.A.)	2.8	Second
APS	Argonne, IL (U.S.A.)	7	Third (under construction)
ESRF	Grenoble, France	6	Third
SPring-8	Nishi Harima, Japan	8	Third (under construction)

Table 4. Synchrotron radiation sources (UV).

Ring	Location	Energy (GeV)	Ring generation
BESSY I	Berlin, Germany	0.8	Second
CAMD	Baton Rouge, LA (U.S.A.)	1.2	Second
ALADDIN	Stoughton, WI (U.S.A.)	1	Second
NSLS I	Brookhaven, NY (U.S.A.)	0.75	Second
Super ACO	Orsay, France	0.8	Third
ALS	Berkeley, CA (U.S.A.)	1.5	Third
ELETTRA	Trieste, Italy	1.5	Third
SRRC	Hsinchu, Taiwan	1.3	Third

which communicate the forces. This so-called Standard Model, however, has many arbitrary parameters, leaving still-unanswered questions about the origin and basic composition of mass. Some of the important contemporary experiments focus on efforts to understand fully and detect the postulated quark structures, to detect the postulated carrier particles such as the Higgs particle, and to understand the matter–antimatter asymmetry. To continue these experiments and to move to additional frontiers, the following facilities are among the most important.

2.1 CERN, European Laboratory for Particle Physics

CERN is a large scientific laboratory built through a collaborative effort of 15 member states from Western Europe. It straddles the French-Swiss border just west of the city of Geneva. CERN has a 27-km electron–positron collider ring (LEP) fed by a 20-GeV Super Proton Synchrotron (SPS). The electrons and positrons are first taken to 600 MeV in a 100-m-long preinjector, then accelerated to 3.5 GeV before passing into the SPS. In addition to electron–positron collisions, CERN can supply researchers with proton beams of both high and low energy, nuclear beams, proton–antiproton collisions, and low-energy antiproton beams. When LEP is operating, the injectors can continue to handle proton beams, interleaved with electrons and positrons destined for LEP. At various LEP intersections, detectors are located to pick up the particles resulting from the collisions and measure their energy and momentum. LEP has four large detectors built of modules, each to carry out a specific detector task.

Since the Higgs particle is likely to be present only at higher energies than is presently available, CERN is planning an upgrade. A new facility, the Large Hadron Collider (LHC), will utilize the existing LEP tunnel, replacing the existing magnets with high-field superconducting magnets. The plan is to collide proton beams with 7 TeV energy each.

2.2 Fermi National Accelerator Laboratory

Known as Fermilab, this laboratory is located about 35 miles west of Chicago, Illinois (U.S.A.). The main facility is the Tevatron which is the final of five separate accelerators working in tandem to accelerate protons to an energy of nearly 1 TeV. A tunnel that contains the main ring, which accelerates protons to 150 GeV, also contains the superconducting magnets of the Tevatron proton synchrotron. With the energies available at Fermilab, experiments are under way to confirm the detection of the top quark. An upgrade of the injector system is under way and will permit the production of antiprotons to be increased fivefold and will enable the delivery of three times as many protons to the Tevatron. Both colliding and fixed-target experiments with protons and antiprotons are conducted at Fermilab. Two large colliding-beam detectors are presently available.

In April of 1994, after a year of experimentation at Fermilab, a research paper was presented giving the first direct evidence for the top quark. The bottom quark was discovered at Fermilab in 1977, leaving the top quark the last undiscovered quark of the six predicted by the Standard Model. Although more experimental data are needed to confirm the preliminary results, the researchers believe they have detected the top quark and measured its mass using the collider detector, CDF. Because of the background and the statistics for confirmation, plus some discrepancies concerning another detector, the D-zero, which could not detect any particles, the researchers are gathering more data in order to definitely confirm these results.

2.3 Stanford Linear Accelerator Center

The Stanford Linear Accelerator Center (SLAC) is located near the campus of Stanford University at Menlo Park, California (U.S.A.). This 30-GeV linear accelerator facility includes experimental areas for fixed-target and colliding-beam experiments with electrons and positrons. Two storage rings are located at the facility, the larger positron–electron storage ring (PEP), and the smaller ring SPEAR which is now used exclusively for synchrotron radiation research. The present focus at the laboratory is to complete a recently initiated construction project to build a facility called a B-Factory that will produce copious quantities of b-mesons. The objective of the facility will be to understand why the universe is made up of mostly matter and not equal quantities of matter and antimatter. Several other laboratories around the world are also pursuing this objective; the KEK facility in Japan, for example, is undertaking to use the same approach of building a B-Factory.

2.4 Alternating-Gradient Synchrotron

The Alternating-Gradient Synchrotron (AGS) is located at Brookhaven National Laboratory, New York (U.S.A.). It is a high-intensity proton accelerator with a maximum energy of 30 GeV. Although not the highest-energy proton accelerator in the world, it has been continually upgraded since it first began operation in 1960. Kaon decay, neutrino experiments, and polarized-proton physics are among the areas of current interest. Shortly, it will take on a new role as the injector for Brookhaven's newest accelerator, the Relativistic Heavy Ion Collider (RHIC), to be completed about 1998.

2.5 KEK, National Laboratory for High-Energy Physics

The KEK-PS proton synchrotron is a 12-GeV machine built for high-energy physics that is now being used for both particle and nuclear-physics experiments. Begun in 1976, the accelerator is located at Tsukuba, Japan. Ten secondary beam lines in two experimental halls deliver antiprotons and pion and kaon beams in various momentum ranges and intensities. A polarized-proton beam of 40% polarization is also available up to 3.5 GeV.

3. NUCLEAR PHYSICS

The goal of nuclear physics research is to understand the structure and forces within that part of the elemental universe called the nucleus. In the nucleus, matter is densely packed and accelerator facilities are needed to study this strongly interacting matter together with the forces that govern its structure and dynamics. The topics in nuclear physics include the electroweak force, which plays an important role in nuclear stability, precision tests of fundamental interactions, large-scale collective nuclear behavior, motions of individual nucleons and mesons in atomic nuclei, and the underlying distribution of quarks and gluons. Because the science of the nucleus involves extremes of temperature, pressure, and density, the research has a strong relationship to astrophysics and cosmology. In the pursuit of nuclear physics research, it is necessary to have accelerators that produce high-quality beams of electrons, hadrons, and heavy ions over a very large energy range together with detectors and targets of novel design. In view of the variety of experiments needed, there are many facilities, worldwide, that are used to conduct nuclear-physics research. The following list contains a sampling of these facilities.

3.1 Argonne Tandem/Linear Accelerator System (ATLAS)

Located at Argonne National Laboratory, Illinois (U.S.A.), this facility is the world's first superconducting accelerator for heavy ions. The accelerator consists of a 40-MV superconducting linac and a 9-MV Tandem Van de Graaff injector. ATLAS is used for high-precision nuclear research using beams of heavy ions such as uranium.

3.2 Tandem/Alternating Gradient Synchrotron

This facility is located at Brookhaven National Laboratory, New York (U.S.A.), and provides beams for relativistic heavy-ion research. Ions up to gold have been accelerated. The facility consists of two linked machines, the Alternating-Gradient Synchrotron (AGS) and the Tandem accelerator. A Booster, a smaller accelerator that began operation in 1991, accelerates particles before they enter the AGS, where they are accelerated to 15 GeV per nucleon. The complex of Tandem/Booster/AGS will provide a full range of heavy-ion beams for fixed-target experiments and also constitute the complete injector system for the Relativistic Heavy-Ion Collider (RHIC), under construction as of 1995.

3.3 Relativistic Heavy-Ion Collider (RHIC)

To be completed in 1998, RHIC will be a dedicated research facility capable of accelerating two beams of nuclei to energies of 100 GeV per nucleon and colliding them. The collider consists of two 3.8-km rings of superconducting magnets for accelerating and storing beams of ions. The collisions at high energies are expected to create a quark–gluon plasma for study. This hot, dense matter is postulated to resemble matter shortly after the Big Bang.

3.4 Continuous Electron-Beam Accelerator Facility (CEBAF)

Experiments at this new facility are expected to begin in 1995. The facility is located in Newport News, Virginia (U.S.A.). The accelerator is designed to deliver high-quality electron beams in the energy range 0.4 to 4.0 GeV with an intensity of 200 A. The racetrack-shaped accelerator consists of two parallel superconducting linear accelerators joined at the ends by recirculation arcs, with three experimental areas. Its main goal is to understand the structure of the nucleon and of nuclei at the quark-gluon level, especially the nature of the low-energy, quark-confining strong interactions.

3.5 Clinton P. Anderson Meson Physics Facility (LAMPF)

LAMPF is located at Los Alamos National Laboratory, New Mexico (U.S.A.). The facility consists of a high-intensity proton linac and associated experimental areas. The 800-MeV linac produces three beams:

1. used to create secondary beams of pions, muons, neutrinos and spallation products,
2. for nucleon-nucleon and nucleon-nucleus studies and production of polarized neutron beams, and
3. to fill a storage ring used to produce spallation neutrons.

The main emphasis in the nuclear physics research program has been to study the interaction of pions with target nuclei; proton and neutron reactions with nuclei; basic proton-proton, neutron-neutron, and pion-nucleon interactions; and fundamental aspects of the weak interaction.

3.6 Bates Linear Accelerator Center

The Bates facility is located in Middleton, Massachusetts (U.S.A.), and is capable of providing very high-quality electron beams with energies up to 1 GeV. The facility contains an energy-loss spectrometer system for high-resolution electron scattering, a 180° scattering facility, an out-of-plane magnetic spectrometer, and other experimental spectrometers used to study a variety of nuclear elementary excitations.

3.7 Indiana University Cyclotron Facility

Located in Bloomington, Indiana (U.S.A.), this facility utilizes two coupled cyclotrons to accelerate polarized and unpolarized light ions. The beams are used directly in experiments and also injected into a synchrotron accelerator (Cooler Ring) that allows electron cooling of the stored beam. After proton energies of 200 MeV are achieved, the protons are extracted and steered into one or more of the various experimental areas. The Cooler Ring is used if higher energies or a finer beam quality is needed. The synchrotron ring is capable of accelerating the beam to 500 MeV and the beam can circulate for long periods of time while being used for in-ring nuclear experiments. In order to produce a higher-quality narrower beam at these higher energies, the ring has a short segment over which the proton beam is accompanied by a precisely tuned, concentric, renewable electron beam. This electron beam damps out the normal oscillatory behavior of the proton beam. The electron beam in this short ring segment runs parallel to the proton beam, is cylindrical with a diameter of about 25 mm, and causes the proton beam to be reduced in diameter from 6 mm to about 1.5 mm.

3.8 National Superconducting Cyclotron Laboratory (NSCL)

The NSCL is located at Michigan State University in East Lansing, Michigan (U.S.A.), and is dedicated to research with heavy nuclear beams with intermediate energy. Two superconducting cyclotrons are available for research.

3.9 Tri-University Meson Facility (TRIUMF)

Located at Vancouver, British Columbia (Canada), TRIUMF operates a six-sector isochronous cyclotron accelerating protons to 520 MeV. Two proton beams are extracted simultaneously to feed two experimental areas, the proton hall and the meson hall. Fields of research include decay modes and other fundamental properties of pions and muons, strong interactions of pions with nucleons and nuclei, elastic and inelastic scattering of nucleons by nuclei, and symmetries of the nucleon-nucleon system.

3.10 UNILAC/SIS/ESR Accelerator Facility

This heavy-ion accelerator facility is located in Darmstadt, Germany and consists of a linac (UNILAC), a heavy-ion synchrotron (SIS), and a storage cooler ring (ESR). Research under way includes nuclear reactions up to relativistic energies, and exotic nuclei far off stability.

3.11 Nuclear Structure Facility

The Nuclear Structure Facility is a 20-MV, vertical Tandem Van de Graaff accelerator, located in Daresbury, England. The research program includes nuclear spectroscopy at high angular momentum, nuclear moment measurements, and low-energy heavy-ion reactions and scattering.

4. PLASMA PHYSICS/FUSION RESEARCH

The energy produced in the sun and the stars is the result of the fusion of light atoms. This type of energy production is believed by many to be the ultimate energy source for Earth if it can be controlled and economically developed. For fusion to occur, an extremely high temperature must be attained with the fuel, this resulting in the fuel material being in the plasma state. Experimentation on heating and confining these hot plasmas is ongoing and necessary in order to develop fusion energy.

There are many different nuclear fusion reactions possible. However, the most readily attainable fusion process on Earth, that is, the one occurring at the lowest temperature, is the combination of a deuterium nucleus with one of tritium. Most fusion technology developmental programs aim to use deuterium–tritium fuel, but a considerable amount of plasma physics research is done on nonradioactive systems. In order to release energy at a level needed for practical use in the production of electricity, the gaseous deuterium–tritium fuel must be heated to about 10^8 K. Attaining these temperatures is not as difficult as the problem of confining the fuel under these extreme conditions. Since at these temperatures the fuel is in a plasma state, magnetic fields can be used to confine the electrons and ions, keeping them from striking the walls of a containment vessel. In order to achieve fusion, a minimum temperature, energy confinement time, and plasma density is necessary. Plasma instabilities of many types are a major obstacle to achieving the proper conditions for net fusion energy release to be greater than the energy needed for heating. In addition to plasma physics research, research is under way on technology development for economical energy production.

Of all the many concepts for containment that have been studied, the one receiving the most attention today is the tokamak concept. The tokamak is a toroidal device which has a hollow, doughnut-shaped vessel through which magnetic fields twist to confine the enclosed plasma. Magnetic fields are produced by current passing through wires or the plasma. Plasma heating can be achieved by a combination of resistive heating, high-energy neutral beam injection, magnetic compression, and radio-frequency heating.

Two examples of major fusion research facilities follow. The experiments to achieve an economical fusion reactor are so large currently that cooperation among countries is necessary. The International Thermonuclear Experimental Reactor (ITER) is an international project to demonstrate the scientific and technological feasibility of fusion energy for peaceful purposes and is being designed by the United States, Russia, Japan, and the European Community.

4.1 Princeton Plasma Physics Laboratory (PPPL)

Located in Princeton, New Jersey (U.S.A.), the major experimental facility at the PPPL is the Tokamak Fusion Test Reactor (TFTR). TFTR is the latest in a series of experimental devices located at PPPL and designed to study high-temperature fusion plasmas. The facility has a major radius of 2.5 m and a minor radius of 0.9 m. The toroidal magnetic field strength is 5.2 T. In December 1993, the TFTR achieved a world record of 6.2×10^6 W of controlled fusion power using a mixture of deuterium and tritium fuel. A new facility, the Tokamak Physics Experiment (TPX), is currently in the design/construction phase and will be built at PPPL in the test cell now occupied by the TFTR. It will be operated as a national user research facility. The purpose of the TPX is to develop the scientific basis for a compact and continuously operating tokamak fusion reactor. TPX is designed for continuous operation, with both high plasma pressure and enhanced plasma confinement. The TPX will be the first tokamak with fully superconducting magnetic field coils. In addition to superconducting technology, TPX will pioneer internal shielding and robotics for maintenance.

4.2 Joint European Torus (JET)

The Joint European Torus (JET), located at Abingdon, England, is jointly funded and staffed by 14 European countries. The overall objective of the JET facility is to study plasmas in conditions and with dimensions close to those that would be needed in a fusion reactor. The experimental facility is a D-shaped tokamak toroid with a major radius of 3.0 m, a minor radius of 1.3 m, and a toroidal field of 3.5 T. In preliminary tritium experiments, JET has produced 1.7×10^6 W of fusion power. The four main areas of research at JET are studies of scaling of plasma behavior as parameters approach the reactor range, studies of plasma heating, studies of plasma–wall interactions, and studies of alpha-particle production, confinement, and consequent plasma heating. Additionally, key technologies such as tritium handling and remote maintenance and repair are being actively investigated. JET will soon implement a new Pumped Divertor Phase of its operation, which will establish the effective control of plasma impurities prior to full deuterium–tritium operations.

Further Reading

BNL Public Affairs Office (1993), *Brookhaven Report on Research*, Upton, NY: Brookhaven National Laboratory.

CERN Public Information Service (1991), *CERN, an International Collaboration in Science*, Geneva: CERN Publications.

Department of Energy (1984), *User's Guide to DOE Facilities*, DOE/ER-0174, Springfield, VA: National Technical Information Service.

Department of Energy (1991), *Nuclear Physics Accelerator Facilities of the World*, DOE/ER-0532T, Springfield, VA: National Technical Information Service.

Dupen, D. W. (1966), *The Story of Stanford's Two-Mile Long Accelerator*, Springfield, VA: National Technical Information Service.

Fermilab Public Information Office (1991), *Focus on Fermilab*, Batavia, IL: Fermilab.

Joint European Torus (1993), *JET Joint Undertaking Annual Report, 1992*, EUR 15080-NE-C, Abingdon, U.K.: JET Publications Office.

Price, D. L., Rush, J. J. (Eds.) (1994), *Neutron Sources and Applications*, report of a review held at Oak Brook, IL, for the Basic Energy Sciences Advisory Committee Panel on Neutron Sources, 8–10 September 1994, DOE/ER-0607P, Springfield, VA: National Technical Information Service.

Princeton Plasma Physics Laboratory (1992), *Annual Report, Princeton Plasma Physics Laboratory, Fiscal Year 1991*, Princeton, NJ: PPPL Information Services.

Winick, H. (Ed.) (1995), *Synchrotron Radiation Sources—A Primer*, Singapore: World Scientific.

PHYSICS, TECHNOLOGY, AND SOCIETY

RUTH H. HOWES, *Department of Physics and Astronomy, Ball State University, Muncie, Indiana, U.S.A.*

Introduction 107
1. **Basic Physics Research and the Development of Technology** 108
2. **Physics, Technology, and Public Policy** 108
2.1 Shaping of Policy by Physics and Related Technologies 108
2.1.1 The Arms Race and the Cold War 109
2.1.2 The Communications Revolution and the End of the Cold War 110
2.1.3 Energy Policy 111
2.2 Assessing Technologies 111
2.2.1 The Cold-Fusion Issue 112
2.2.2 Nuclear Reactors and Electric Power Generation 112
2.2.3 How Much Verification? 113
2.3 How Policy Makers Get Technical Information 114
2.3.1 Scientists in the Executive Branch 114
2.3.2 Classification and Physics Research 114
2.3.3 Advice to Congress 115
2.3.4 The Nongovernmental Organizations—NGOs 115
3. **The Impacts of Physics on Society** 116
3.1 Economics 116
3.1.1 Benefits of Research 116
3.1.2 Funding for Research and Development 118
3.2 Social Impacts of Physics and Related Technologies 119
3.3 Making Science Policy 120
4. **Public Knowledge of Physics** 120
4.1 Formal Education 120
4.1.1 The General Public 120
4.1.2 College Graduates 121
4.1.3 Professional Physicists 121
4.2 Informal Sources of Information 122
4.2.1 Magazines, Journals, and Newspapers 122
4.2.2 Science Museums 122
4.2.3 The Electronic Media and Films 122
5. **Three Post–World War II Case Studies Contrasted With the U.S.** 123
5.1 Western Europe 123
5.2 The Former Soviet Union 124
5.3 Japan 125
6. **Summary: Mutual Influences** 126

INTRODUCTION

The development of twentieth-century physics has been closely tied both to the rise of modern industry and to the building of technologically advanced weapons systems. Traditionally, many governments have funded physics research as a means to a better national defense. During the last decade, this perception has changed as the industrialized nations have adopted the position long held by developing nations that a strong basic science infrastructure forms an essential building block for an economically competitive industrialized society. The importance of a scientifically literate work force is reflected by a global call for investment in human capital. While these trends are worldwide, they can be examined by looking at the evolution of the interaction between physics and society in the United States and comparing it to the development of physics in other societies.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. BASIC PHYSICS RESEARCH AND THE DEVELOPMENT OF TECHNOLOGY

The distinction between basic physics and applied physics is generally easier to draw in principle than in practice. Ideally, basic research asks questions about the nature of the universe that arise only from the curiosity of the investigator. For example, physicists are searching for the top quark to see whether it exists and whether their current theoretical models of the structure of matter and the evolution of the universe are correct. No current plans exist to develop any specific technology on the basis of the "standard model."

Applied physics seeks to solve a specific practical problem using the results of basic physics research. For example, an applied physicist might develop a new photoelectric cell with a higher efficiency for use in portable solar-powered electric devices or study the interaction of sound waves with defects in metals in order to build a detector for locating metal fatigue before the metal structure fails. In these cases, the physics experiment is driven by the need for a technology.

In practice, distinguishing an applied physics laboratory from one engaged in basic research can be impossible. Workers in both discuss the same theories and experiments. In many cases, they use identical equipment. Applied physicists may make observations or develop theories that advance fundamental understanding, while physicists doing basic research may develop technologies. The search for new materials that exhibit high-temperature superconductivity can belong to basic or applied research, depending only on the goal of the research group. Physics research lies on a continuum stretching from basic research through applied research and on into the development of specific technologies. The lines are not clearly drawn.

Physics and technology feed one another. New discoveries in physics frequently lead to new technologies. Without physicists' attempt to confirm Einstein's prediction of stimulated emission of radiation, lasers would not exist, along with applications ranging from compact disc players to lecture pointers. Conversely, new technologies make possible advances in basic physics. The use of inexpensive, powerful computers in data collection allows modern nuclear- and particle-physics experiments to be done in a finite time. Available high-vacuum, high-voltage systems for construction of cathode-ray tubes led directly to the discovery of the electron and the beginning of modern physics.

2. PHYSICS, TECHNOLOGY, AND PUBLIC POLICY

Just as new physics promotes new technologies and new technologies give rise to new physics, so the availability of new technologies can shape public policy and public policy can influence the development of new technologies and encourage or discourage new fundamental physics. The influence of new technologies on societies has long been recognized. The steam engine triggered the industrial revolution. For further examples, life in the United States has been shaped by inexpensive private automobiles, and new computer technologies are rapidly changing the culture of today's work place. Policy makers must understand the capabilities and limitations of new technologies in order to respond to their social impacts. Because few policy makers have studied science, they rely on science advisers.

Recent concerns over the "competitiveness" of U.S. industries stress the role of public policy in shaping the development of both basic physics and technologies. Funding policies have allowed the growth of "big science" (Weinberg, 1967), that is, experimental facilities involving groups of tens to hundreds of physicists. By directing funds, the government can emphasize particular subdisciplines in physics. Industry supports research areas where managers see promise of commercially successful technologies.

2.1 Shaping of Policy by Physics and Related Technologies

Because of its role in the development of radar and the atomic bomb during World War II, physics has been particularly closely associated with the making of public policy in the United States. In contrast, agricultural research has dominated public policy discussions in much of the developing world. To understand the interplay between physics, technology, and public policy in the United

States, which mirrors the interplay between science and public policy in other nations, several case studies are presented.

2.1.1 The Arms Race and the Cold War During World War II, physicists applied basic results from modern physics to developing military technologies. Although enormous advances were made in fields such as the use of vacuum electronics for radar and understanding of sound propagation for sonar, the use of nuclear physics to develop the first atomic bomb was probably the most spectacular and certainly the best publicized use of physics in weapons development (Rhodes, 1986). Politicians and physicists alike recognized that nuclear weapons changed the way wars of the future would be fought. Because the physics on which the nuclear weapon is based was commonly known in the international community, nations that wished to be considered major contenders on the world stage rushed to develop their own nuclear arsenals. Nations such as India, Pakistan, and Iraq pulled technical talent and money away from industrial and agricultural development to support the development of nuclear technology.

Nuclear weapons offered a cheap substitute for large deployments of military manpower. The United States and the Soviet Union grappled for the control of Europe and developed powerful nuclear arsenals. The awesome destructive power of nuclear weapons meant that each side realized that a nuclear war would destroy not only the enemy but itself also. Physicists and other scientists and engineers were hired by the military to develop more capable nuclear warheads. The two great nuclear powers were effectively at war, but neither side dared actually to attack the other for fear of the terrible destructive power of the nuclear arsenals (Schroeer, 1984; Barash, 1987).

In the meantime, new technologies increased the accuracy of long-range nuclear delivery systems (Mackenzie, 1990). This opened the possibility of attacking the other side's arsenal rather than its population centers. New guidance technologies effectively drove strategic policies (Scowcroft, 1983). Physicists played prominent roles in analyzing strategic options (U.S. Congress Office of Technology Assessment, 1981).

On 23 March 1983, in a nationally televised address, President Reagan announced a new technological initiative designed to make nuclear weapons "impotent and obsolete." The Strategic Defense Initiative (SDI), popularly called "Star Wars," called for research into new defensive military technologies based on recent developments in the physics of directed-energy weapons and the possibility of directing an attack on missiles in the early phases of their trajectories from weapons based in space. Controversy erupted on whether or not such a defensive shield would be technically possible and politically desirable (U.S. Congress Office of Technology Assessment, 1985; Broad, 1985). Physicists featured prominently on both sides of the controversy. Even the nonpolitical American Physical Society produced a technical report on the state of the art in particle and radiation beam weapons (Study Group on Science and Technology of Directed Energy Weapons, 1987).

The resulting debate illustrated that physicists' results can be influenced by their politics. Those who felt that SDI would offer an effective defense against modern nuclear delivery systems generally agreed that the arms race could find a technological solution. Recent evidence suggests that the political beliefs of some actors were sufficiently strong to influence their perceptions of the results of their experiments and calculations (Broad, 1992). Physicists who believed that there could be no effective defense in a nuclear war generally held that the arms race should be solved through the traditional political means of arms control. A consensus gradually emerged that, although SDI could not provide a large-scale defense for a population against nuclear weapons, some of its technologies could provide defenses for hardened military targets or against very limited nuclear attacks (U.S. Congress Office of Technology Assessment, 1988). Physicists defined the cost of these limited and practical systems, leaving a political decision concerning whether the military advantages of the defenses justified the cost of deploying them.

In the case of nuclear deterrence, the cold war, and SDI, technologies based on developments in physics shaped possible military and foreign policies. Options available to political leaders were defined and limited by the technology available. Physicists acquired a role in government as judges of what was and was not technically possible.

At the same time, the utility of physics in the development of military technologies and the resulting support of physics by the federal government, defense contractors, and the military shaped the way in which physics research was conducted. The massive national laboratories established to support the Manhattan Project developed into models for today's large accelerators. Today's research groups in physics mimic the group of physicists and supporting engineers and other scientists who designed the atomic bomb. Physicists specialized, with some individuals doing only theoretical work and others, experimental work. Some became specialists in the design and construction of apparatus. The individual who did theoretical work and then designed, constructed, and carried out his own experiment nearly disappeared.

Finally, physicists have learned the limitation of their ability to control the use of devices they create. Following their success in creating an atomic bomb, many physicists were shocked by the destruction at Hiroshima and Nagasaki. They quickly learned that their technical expertise bought them no ability to control the use of the bomb. Science advisers may influence policy, but power in our democratic society belongs to those elected through the political process (see articles in Mendelsohn *et al.*, 1988). As more and more nations acquire the technical capability to construct nuclear weapons, the role of physicists in shaping policies for their use seems likely to decline further. For example, sales of nuclear warheads from the former Soviet Union on the world market have become a real possibility.

2.1.2 The Communications Revolution and the End of the Cold War Although the military uses communications technologies, industry has been the primary supporter of basic research in the field, and technological developments have been driven by consumer demands. Even the first computers were developed for civilian applications (Cohen, 1988, pp. 119–120). Development of the transistor and other breakthroughs in fundamental physics came from Bell Telephone Laboratories and other industrial research centers. The first high-temperature superconductors were found in an IBM laboratory. These major corporations have supported basic research in condensed-matter physics because they see the results of this research leading directly to devices that can be applied to communications technologies.

The industrial support of basic physics has proved profitable for industry and productive for physics. The social impact of the new communications technologies based on developments in condensed-matter physics has reached almost all aspects of daily life, from the ready availability of dedicated computers in microwave ovens and automobiles to the business use of personal computers. An extensive literature documents the impact of computer technologies on everything from the changing nature of the workplace to the question of protecting personal data stored in databanks from unauthorized users (see Brunn and Leinbach, 1991; Ganley and Ganley, 1989; Lundstedt, 1990; Pool, 1990).

Perhaps the most dramatic evidence of the impact of the communications revolution is its role in the breakup of the Soviet Union. During the August 1991 takeover of the Kremlin by a hard-line communist faction, the rebels controlled traditional means of communication within the Soviet Union. In the absence of radio, television, or newspapers, the people of Moscow were still able to communicate with the West and with other parts of the Soviet Union via FAX and e-mail. Citizens of Moscow itself received information from Yeltsin and his advisors thanks to desktop publishing and the availability of copy machines. Instant communications kept the coup public and mobilized local, national, and international support that helped lead to its overthrow. Among other effects, sophisticated communications links tend to open closed societies.

Sophisticated communication links form an essential element of a competitive industrialized society. Not only the republics of the former Soviet Union but also developing nations are attempting to develop computer-based communication links. For example, India currently seeks both to increase the number of computers throughout the subcontinent and to network them (Rahman, 1990). The presence of the basis for a nationwide communications network is one of the signs that distinguish the newly industrialized nations of Asia from the undeveloped nations of sub-Saharan Africa (Forje, 1989).

Physicists never controlled the applications of their research in the industrial laboratories. Unlike the social impacts of military technologies brought about by physics research, the social impacts of the communications technologies were a side effect of industrial policies motivated by profit. Physics research was seen as valuable in its ability to produce profit-making devices. Unlike physicists in military development, physicists in communications development neither expected nor attempted to control the social effects of their research.

Because industry has funded basic physics work as an investment in future profits, physicists in industrial laboratories are liable to lose their jobs if management does not perceive their products as being useful. Thus, many industries demand shorter-term profits in hard economic times and many have reduced their staffs in recessions. The type of basic research supported by industry tends to show a clear connection to a possible product and is generally not long term or highly speculative. Like research supporting the military, physics done in industrial laboratories often uses large groups, expensive equipment, and many support personnel.

2.1.3 Energy Policy Basic physics research has long played a role in the development of energy technologies from the fossil-fuel electric generator to the nuclear-reactor electric power plant to photoelectric cells. Most physicists agree that finite supplies of fossil fuels will force reliance on new technologies for energy production within the next few decades and that energy conservation measures are inexpensive means for stretching valuable natural resources. Few expect any technology to be without some adverse side effects. Beyond these limited areas of consensus, physicists disagree strongly as to the best energy sources to replace fossil fuels, the relative importance of seeking new energy sources and implementing conservation measures, and the probable long-range environmental consequences of using various energy technologies (Howes and Fainberg, 1991).

Despite the fact that physicists fail to agree in advocating any single energy policy, individuals have been visible and active as advocates of a number of different policies. The profession has been strongly associated in the public eye with making energy policy.

Funding for basic research to be applied to the development of energy technology has undergone wide swings as concern over U.S. energy supplies has varied with the price of oil. When funding levels were high, as during the mid seventies, energy research showed the same tendencies to large-group work noted in basic physics research related to military technologies. Reduced funding levels have caused researchers to return to the small-group model of research.

Although physicists had long warned that nuclear reactors were dangerous if not handled carefully and that nuclear wastes required development of storage technologies, physicists were seen as supporting the nuclear power industry for their own profit. Since World War II, physicists had been popularly perceived as nearly infallible workers of technical miracles. The accidents at Three Mile Island and Chernobyl badly injured their public credibility. Although physicists were granted a special voice in setting military policy, their public differences over energy reduced their privileged role in policy making. As in the case of communications, physics produced research that could lead to commercial results. Unlike their role in communications, physicists exhibited a strong interest in influencing energy policy, but they found that they would need to work inside the political process as citizens rather than be granted a special role in the policy process.

2.2 Assessing Technologies

Critical issues in energy policy, communications, and military policy center on whether or not technologies work and what their side effects will be on society and the environment. Although these issues are primarily technical, science may not provide unambiguous answers. Policy makers demand guidance on what science does and does not know. In sending technical aid to developing nations, industrialized governments face the complex task of examining the effect of the technology on the recipient society and the ability of that society to make the imported system work (Carnegie Commission, 1992a, 1992b, 1993). In technology assessment, physicists are asked to clarify those issues where science can provide answers and outline the decisions that are essentially political.

2.2.1 The Cold-Fusion Issue In 1989, the University of Utah called a press conference to announce that Dr. Stanley Pons and Dr. Martin Fleischmann had discovered a revolutionary form of nuclear fusion at room temperature that promised an almost unlimited source of inexpensive energy. Unfortunately, room-temperature or "cold" fusion violated the established laws of nuclear physics. The physics community reacted to the news with skepticism and excitement. All over the country, physics laboratories attempted to duplicate the Utah results. The Utah group had not published its results and refused to release details of its experiments pending patent application procedures. The physics community gathered information from the pages of *The Wall Street Journal* and buzzing e-mail networks. For three turbulent weeks, reports confirming and denying cold fusion jostled one another in the press. Dr. Steven Jones of Brigham Young University, who had long studied muon-catalyzed fusion, reported observing cold fusion, although his effect was many orders of magnitude weaker than that in the Utah report.

The press played up the dispute raging within the scientific community, billing it as a feud between all chemists and all physicists. Sober debates on the statistics and technology of neutron counting got lost in colorful name calling and interviews with graduate students who listed pizzas as part of their expenses in seeking cold fusion. Politicians, particularly in Utah, began campaigning for large research grants to support cold fusion. Serious technical issues played a small role in popular discussion.

The early stages of the cold-fusion debate indicate how technology assessment can be done in a way that damages society. Hope of technological breakthroughs and personalities jockeying for professional advancement and public recognition distort the public debate and polarize the scientific community. Fortunately physicists (and chemists) recognized their public responsibility to clarify the technical possibilities for political decision makers. A series of papers gradually began to appear reporting detailed experimental results and offering theoretical explanations for results and lack thereof. The early Pons and Fleischmann results were shown to be overoptimistic. On the other hand, the debate on a possible relationship between nuclear physics and condensed-matter physics was probably healthy for both communities once the physics began to dominate the political debate. Funding returned to a level consistent with the physics issues involved, and a textbook example of how not to conduct technology assessment came to a quiet end (Huizenga, 1993; Pool, 1990; Mallove, 1991).

The cold-fusion question provides an excellent illustration of major elements in technology assessment: scientific knowledge and uncertainty, political and economic ramifications, and public pressure.

2.2.2 Nuclear Reactors and Electric Power Generation Like cold fusion, the use of nuclear fission reactors to generate electric power provides a case study in technology assessment featuring technical uncertainty, economic pressures, and strong public feelings. The technical issues involved in nuclear fission for electric power generation center on the risk of a major nuclear accident and its consequences for the public. Reactors are complex systems, and analysis of their possible failure modes has established the state of the art for risk assessment in large systems (Vogelsang and Barschall, 1991). Despite the excellent work that has been done in the field, risk assessment yields at best a probability that a major accident could occur. The magnitude of the probability and the consequences of a particular accident are a function of individual reactor design and the geography of its location. Consequences also depend on the biological effects of low-level radiation, which, again, are just beginning to be understood and are necessarily presented as a statistical measure, such as the probability of the occurrence of a particular type of cancer at a particular exposure level.

The economic stakes in constructing nuclear reactors are much larger than those involved in cold fusion. The expense of a nuclear power plant can bankrupt an electric utility if the reactor does not generate useful electric power. Public concern over safety has increased the cost and complexity of both the licensing and the construction process to the point where many utility companies cannot afford to borrow capital to construct a new nuclear plant. There are currently no nuclear reactors on order in the United States. Insurance companies cannot bear the risk of giant claims following a nuclear accident, and so the federal government assumes a major portion of the risk under the Price–Anderson Act. The public questions whether or not the government should absorb the risks of a technology that earns profits for private corporations.

Finally, nuclear power plants have been sold to the public in a manner that had little to do with the technical issues of their safety. In the 1950s, the power industry oversold reactors as a safe, cheap way to generate electric power. Throughout the sixties and seventies, few outside the physics community questioned the safety of nuclear reactors. With the accidents at Three Mile Island and Chernobyl, the public became aware of the dangers of the technology and reacted strongly against nuclear power generation.

The physics community generally prefers to compare the risks and benefits of nuclear reactors to the coal-fired plants that are the other major choice for electric power generation. Public opposition to nuclear reactors has become so great that nuclear power generation has been assessed negatively, not by physicists and the rest of the technical community, but by public opinion. Unlike the case of cold fusion, the nuclear power issue in the United States has been settled politically rather than on technical merit.

Other nations conduct risk assessment on a considerably different basis. Japan and France are currently developing and using plutonium-fueled reactors, and Britain has announced the intention of constructing a major facility to process plutonium reactor fuel. The republics of the former Soviet Union continue to operate nuclear power plants that are known to be less safe than western systems, despite their experience with the accident at Chernobyl. In each of these cases, the national government has determined that the economic benefits of nuclear power outweigh the risk to the population.

2.2.3 How Much Verification? In arms control, verification of treaties refers to the use of technical means to ascertain that all parties to the treaty are complying with the provisions of the treaty. Physicists, the military, and politicians have long accepted the fact that no treaty verification can produce certainty that all parties are in total compliance with the provisions of the treaty. Thus, verification regimes are designed to insure that compliance is good enough so that no party to the treaty can gain a militarily significant advantage by violating the provisions of the treaty (Krass, 1985).

Physicists can analyze the capabilities of various monitoring technologies, such as satellite imagery or overflight. On the basis of known properties of detectors and electronics, scientists can tell politicians how likely it is that a particular property of a missile system or a radar can be measured and to what accuracy this can be done. In many cases, treaty provisions are written to limit precisely those properties of weapons systems that can be measured. For example, the Anti-Ballistic Missile Treaty limits the apertures of radar transmitters and their orientation, both of which can be readily measured from satellite imagery.

As in the case of nuclear-power risk assessment, analysis of the capabilities of monitoring systems yields at best a probability of detecting a violation. The acceptance of a verification regime becomes a political decision based on how certain the decision maker feels the U.S. should be of treaty compliance. Physics provides a baseline but cannot make the final judgment call.

Public opinion plays a role in verification, but it has usually been a minor one since the arcana of treaty verification are too complex to be of enormous public interest. In this arena, the political arguments are generally among different constituencies within the executive branch and Congress, who are driven by their perceptions of the threat posed to the U.S. by other treaty parties and their own agendas (for example, the military may wish to build a particular weapon and not wish it to be limited). Unlike the cold-fusion issue, which was primarily resolved by the physics community, and the nuclear-fission power-generation issue, which was primarily resolved by public opinion, the question of how much verification is enough is resolved by a group of specialists, including physicists, and presents still another model of technology assessment.

2.3 How Policy Makers Get Technical Information

Governments have long recognized the need of high-quality scientific advice. To provide it, the National Academy of Science was established during the Civil War. European academies of science, which today support ministries of science, were established much earlier to promote and reward scientific research. Developing nations have generally established government agencies to promote science and advise government officials. In many cases, these agencies have been so involved in the political process that they have been ineffective in providing sound technical advice.

The stakes for science advice to policy makers can be very high. Mistakes are costly both in economic terms, as for example in funding of projects that are not technically feasible, and politically, as in holding up negotiations for a treaty based on a technical misperception or in negotiating a treaty whose provisions cannot be enforced. Because jobs depend on technical issues, such as which weapon system to build, and because issues like nuclear power touch public emotion, pressure can be brought to bear on science advisors. Thus, good science advisors not only must possess technical expertise and sound scientific judgement, but they must have the character to withstand a considerable amount of personal and professional pressure and be adept in politics as well as science.

2.3.1 Scientists in the Executive Branch Almost all the executive branch agencies have staffs with technical expertise. Physicists are located prominently on the staffs of the U.S. Departments of Energy, Defense, and State. The Department of Energy is responsible for the operation of the National Laboratories, which are a repository of scientific expertise and experience. In addition, agencies such as NASA, the U.S. Arms Control and Disarmament Agency, the CIA, the Agency for International Development, the National Institute for Standards and Technology, and the National Oceanographic and Atmospheric Agency have large technical staffs and may be headed by people with technical training.

At the highest level, the Science Advisor to the President must be confirmed by the Senate and heads the White House Office of Science and Technology Policy. The Science Advisor does not hold cabinet rank and his influence depends upon his personal relationship with the President. Models have ranged from the strong influence of Kistiakowsky on President Eisenhower to the relatively weak influence of Keyworth and Graham on President Reagan. Physicists have frequently served as Science Advisor (Golden, 1980, 1988).

On 23 November 1993, President Clinton issued an executive order forming the National Science and Technology Council (NSTC), which he will chair personally. The Cabinet-level body will function in parallel to the National Security Council and the National Economic Council. Beneath the NSTC, there will be a system of nine or ten committees with interagency representation at the assistant secretary level. The areas addressed by the committees have not yet been completely determined, but a preliminary list includes such topics as health, safety, and food research, civilian industrial technology, national security research and development, and fundamental science and engineering research. The intent of the new administrative structure is to coordinate the $76 billion research and development budget and to increase cooperation among the various agencies that conduct research. Each of the major committees will be supported by a system of subcommittees. The administration expects to produce a new statement on science policy during 1994 that is analogous to the statement on Technology issued in February 1993 (Clinton and Gore, 1993).

Most other nations have long had formal science advice at the cabinet level or immediately below. It is also common for nations to issue formal statements of science or industrial policy. The United States has been relatively late in promoting science to a formal position in the executive branch and developing science policy, probably because of a national reluctance to control any aspect of industrial operations.

2.3.2 Classification and Physics Research Physics advice to the military and foreign policy makers frequently involves classified information. From the days of the Manhattan Project, a fundamental tension has existed between the freedom to tap broad-based knowledge for new ideas and the military need for secrecy. In physics literature, not to cite one's sources constitutes plagiarism. In the world of classified policy making, physicists are warned that to cite their sources is to lose them. From Rabi's legendary World War II refusal to design a radar tube unless he was told its purpose, an uneasy compromise has existed between the two communities.

Classified research contracts flow to a number of major corporations and "think tanks." Not infrequently, the contract monitors are very honest, capable young scientists and military officers who hold advanced degrees in technical fields. They are asked to judge the work of senior scientists in fields outside of their own, an impossible task. No form of peer review exists within the classified system, and even library research is limited. The flow of information depends on private contacts and individual physicists' knowledge of what experiments have been conducted by whom.

Most physicists agree that some circumstances demand secrecy. Certainly no one wishes to publicize the technical details of the construction of nuclear weapons. Physicists who work within the classified community must accept security clearance. By their acceptance, they agree not to discuss their work with colleagues who are outside of the loop of cleared personnel and to publish in the field only with permission of security personnel. The system is closed, which is probably good for keeping technical secrets necessary to the national security, but is bad for the quality of scientific advice received by policy makers. This problem probably extends to other nations, but, for obvious reasons, cannot be documented.

2.3.3 Advice to Congress Congress receives technical advice from four sources (Golden, 1988; Kuehn and Porter, 1981). First, the staffs of the Congressional committees and individual members increasingly contain people with scientific training. Programs such as the Congressional Fellowship Program of the American Association for the Advancement of Science, the American Institute of Physics, and The American Physical Society have successfully encouraged physicists to become involved in the legislative process. A handful of members have scientific training themselves.

Second, the Congressional Budget Office and the General Accounting Office, which report to Congress on matters touching the budget, contain staff members with training in several sciences, including physics. Although the agency reports are primarily oriented towards financial issues, technical issues become critical to budget estimates and are discussed in the reports. Third, the Office of Technology Assessment was established chiefly to provide members of Congress with impartial analysis of technical issues that separates uncertainties in scientific information from aspects that are primarily political. The Congressional Budget Office, the General Accounting Office, and the Office of Technology Assessment explore issues only at the request of members of Congress. These agencies have developed a reputation for separating political issues from technical ones and making concise, clear presentations of what is and is not scientifically understood. Finally, the Library of Congress has a growing research division that prepares reports at the request of members. Division specialists include physicists.

Like physicists who advise the executive branch, those working for Congress are limited by their personal relationships with members and the bounds of classification. However, the number of members of Congress allows for a much wider variety of interactions than does the vertical structure of the executive branch agencies. The number of physicists advising Congress has grown steadily as staff sizes have increased and more members have recognized the importance of physics to issues, such as energy policy, that directly touch their districts.

2.3.4 The Nongovernmental Organizations—NGOs Since the turbulent years of the Vietnam War, many physicists have chosen to present their advice to government through organizations outside of government. The federal government had long used groups of civilian experts for advice. Experts were paid by some federal agency for part-time service as advisors. The most famous group is the Jasons, who are a group of physicists advising the Department of Defense on weapons of the future. Perhaps their best-known project is the electronic defensive MacNamara Line constructed during the Vietnam War. The Air Force funded the

Rand Corporation to examine technical issues of interest, and the Department of Defense supported professors at a variety of universities in the Institute for Defense Analysis. Each of these groups and others like them worked closely with the agency that funded it.

The campus uprisings of the late 1960s and early 1970s convinced universities that agencies like the Institute for Defense Analysis should not be a part of their academic mission. Thus, these agencies became not-for-profit corporations that have full-time employees and stables of consultants. These think tanks continue to perform a major portion of the analysis of physics issues that is conducted by the government.

In addition to the independent corporations, which developed from groups of advisors established by the federal government, a second group of organizations grew from the interest of physicists and other scientists in particular policy issues. They are supported by public contributions and grants from foundations. In contrast to the think tanks, they receive no federal money. Some of these organizations, such as the Sierra Club and the Federation of American Scientists, grew from the changing interests of established lobbying groups. Others, such as the Union of Concerned Scientists and the Natural Resources Defense Council, were formed by groups of scientists, with physicists playing prominent roles. The founders of these lobbies felt that the U.S. government had misled the pubic on issues such as the possibility of a nuclear test ban treaty and nuclear fission energy. They collected small, dedicated staffs of highly trained scientists, which were supplemented by volunteers.

During the 1980s, the lobby groups played an increasingly significant role in presenting technical issues to the public and, hence, Congress. Professional scientific organizations, such as The American Physical Society and the American Association for the Advancement of Science, have occasionally submitted reports to their memberships clarifying technical discussions with policy impacts. These reports also served to inform the public, the press, and policy makers.

Although neither the NGOs nor the think tanks appear on any organizational chart, they provide much of the physics input to the making of policy in the United States. Their variety and lack of formal structure allow creative thinkers to present their ideas to policy makers, and they are important to the role that physics plays in shaping public policy.

In the developing world, technical advice provided by the NGOs may play a critical role in providing science advice. The various agencies of the United Nations have provided most of the technical advice available to many African governments (Forje, 1989). Among the less-developed nations, regional cooperation for science has become increasingly important since no single nation can afford to develop all aspects of science. In Latin America, nations recently increased their emphasis on international collaboration (Stann, 1993).

3. THE IMPACTS OF PHYSICS ON SOCIETY

Although physics has played a significant role in shaping U.S. policy, particularly on defense, energy, and the environment, it also affects American society in other less formal ways. While the precise effects of developments in basic physics on such areas as the economy and the structure of the workplace are difficult to quantify, there seems little doubt that advances in basic physics do change these important aspects of society.

3.1 Economics

Developments in basic physics frequently lead to new technologies that can be marketed. The justification for industrial funding of physics research has been its long-range economic benefit. On the other hand, commercial exploitation of breakthroughs in physics requires years of applied research and development. Thus, the connection between economic development and basic physics research can seem very tenuous (Smith, 1990).

3.1.1 Benefits of Research American technological heros have generally been portrayed as practical men of genius who lacked formal education. Contrary to this image, many American inventors were students of basic physics. Benjamin Franklin not only published *Poor Richard's Almanac* but also a number of serious scientific papers advancing his theory of two types of electric charge. Samuel Morse had studied physics at Yale, and even Thomas Edison had a solid understanding of the basic research in electricity and magnetism of his day (Smith, 1990, pp. 17, 25–26).

The Manhattan Project represented an unusually rapid advance from the basic physics of nuclear fission announced in early 1939 to the first nuclear reactor (1942) and the atomic bomb (1945). Less-publicized applications of basic physics made radar and other vital weapons systems possible. American leaders perceived scientific research, particularly physics, as a key to the industrial and military might of the country. Their philosophy was codified in Vannevar Bush's report, *Science, the Endless Frontier* (1945).

The launch of Sputnik in 1957 convinced remaining doubters that basic science research must be a key to national prestige. The glamour of the Apollo Program of the 1960s persuaded talented students to study physics. Thus, for 25 years following World War II, U.S. policy makers and citizens accepted the vital role of physics research in promoting the economic growth of the country without serious question. Funding for physics research was plentiful, and the community thrived. Physics held the key not only to defense technology but to space and to the development of high-technology industry. American technology dominated the world's market and was transferred to other countries through a profitable trickle-down mechanism of licensed manufacture.

By 1970, policy makers and the public had begun to doubt the efficacy of basic physics research in promoting economic success. The Defense Department had produced a study that seemed to indicate that basic research had little relationship to new ideas that came in the development phase of projects (Chalmers and Isenson, 1966). The Apollo Program came to an abrupt end, and American industries began to lose their total domination of the global economy. The environmental side effects of uncontrolled technological development had an impact on public consciousness, and public belief in technological miracles and science as a cure for all ills disappeared.

Although the physics research community pointed out that the DOD study simply had not looked back far enough at the research basis for new ideas and that the Apollo shutdown was for good technical reasons, funding agencies began to demand clearly defined links between economic products and basic research. Conventional wisdom indicates that the payback time that the American public is willing to accept is no more than three years. On that time scale, concrete links between basic physics and economic development are difficult to demonstrate. Public funding of huge demonstration projects that were supposed to set models for private development during the oil crises of the 1970s failed to meet national energy needs. Distrust of these projects was transferred to basic physics research. The research infrastructure, particularly in the universities, lost funding. Meanwhile, aging American industries continued to lose market share in the global economy.

In the mid 1980s, policy makers realized that neglect of basic physics research damaged the development of high-technology industries, which produced the most lucrative employment opportunities. The time scale for the process might be too long to expect individual companies to support this research, but the support was necessary for future developments. The end of the cold war and the decline of the need for new weapons meant that the power of the national laboratories should be focused on supporting civilian industry. The U.S. recognized a need for a science policy that both fostered basic research and helped companies utilize research results in industrial development. The policy took two forms: encouragement of technology transfer from universities and the national laboratories to industry and changes in policies toward the formation of research consortia.

Technology transfer encouraged the formation of industry–university partnerships. Universities developed policies to allow the conduct of proprietary research on campus. Similarly, policies were changed so that the national laboratories could enter into cooperative research and development agreements (CRADAs) with industrial partners. Industry would be able to use the supercomputers and other advanced capabilities of the national laboratories to support research with immediate applications in industry. The system of CRADAs has proven difficult to implement because the laboratories and industry, not to mention the universities, are used to very different research objectives. University researchers worry about freedom to publish, industries about proprietary information, and the laboratories about leaks of classified information. The CRADAs themselves involve complex negotiations over intellectual property rights, the time scales for anticipated results, and even favoritism in selection of firms with which to develop them. Despite a slow start, the system appears to be gathering momentum. The Clinton administration and Congress appear to share a commitment to streamline the creation of industry, university, and national laboratory partnerships. Legislation to accomplish this is under consideration.

U.S. antitrust laws have discouraged cooperation between competing companies for technology development. The government has stayed carefully away from supporting any research that might benefit a private concern. To support industrial development, new interpretations of the antitrust laws hold that they do not apply to basic research. Furthermore, the government now should support basic research to aid industry as long as its results are shared by all U.S. firms in the arena. In 1987, the government and the semiconductor industry formed a partnership to create Sematech, a research effort to improve the manufacturing technology for semiconductor devices. The semiconductor industry had been losing its majority market share to the Japanese. Encouragingly, recent trade figures show that the U.S. firms have begun to regain a lead in the field. It seems likely that Sematech will set a pattern for other government aid to basic research. Certainly, cooperation among government agencies including the national laboratories, the universities, and industry will increase in the future support of basic physics research.

3.1.2 Funding for Research and Development Traditionally, federal funding and university employment have dominated physics research. This pattern seems to be shifting. In 1990, 46.1% of R & D funding in the United States came from the federal government about 20% from industry, and the remaining funding from states, private sources, and universities (Hirano, 1992, p. 582). During 1990–1991, of graduates with Ph.D.s in physics who found employment that might be permanent, 27% entered higher education, 42% entered industry, and 25% joined the government or federally funded research centers. Of bachelor's degree recipients, 42% entered industry, 12% entered education at the college or secondary level, and 23% either joined the military or entered government service. Twenty-one percent of bachelors found only part-time work or no work at all, and 70% of new Ph.D.s either accepted postdoctoral temporary positions or were unemployed (Ellis and Mulvey, 1992). Clearly, the traditional academic career pattern of a research physicist is no longer a dominant model in the U.S. and is rapidly declining in other countries.

Federal grants for research have always existed in the United States. However, following World War II, the magnitude of support for research increased dramatically because policy makers saw research as an engine for economic growth. Despite efforts by Vannevar Bush and others to establish an umbrella organization to oversee funding in all areas of research, the major funding channel became large federal agencies, such as the Departments of Defense and Energy and NASA, which supported development and basic research that had promise of specific application to their missions. The National Science Foundation assumed responsibility for science in general, science education, and those areas of basic research not supported by the agencies (Smith, 1990).

Research funding emphasized support of individual projects rather than umbrella grants to laboratories. In part, funding patterns respected the individual genius, and in part, they made a political compromise to avoid supporting universities on the coasts where postwar research was well established. Research grants were awarded following a review by a committee of scientific peers to establish the merit of proposals. Congress entered the process occasionally to establish programs to target researchers in specific geographical areas or with specific demographic characteristics, such as gender, race, or age, but the majority of grants were awarded on the peer review system.

The review system has been widely criticized for supporting an "old boys" network in the prestigious research universities at the expense of talented researchers elsewhere.

Peer review may also foster conservatism among authors of proposals, since radical ideas are unlikely to attract funding. On occasion, university researchers have bypassed the review system by appealing directly to Congress for research support. These "pork barrel" projects have been criticized for lacking scientific merit. The ultimate political battle for research funding involved the siting of the multibillion dollar Superconducting Super Collider. States submitted proposals offering various financial incentives to the Department of Energy, and the selection process certainly considered these bids and the clout of state representatives in Congress as well as the technical merits of proposed sites. In 1993, Congress discontinued funding for the SSC because of the perception that it offered little potential for economic benefit in comparison with its cost. Legislators were also distressed by management problems that surfaced on inspection of the project.

The space station represents the largest of the "big science" projects that have dominated funding debates in recent years. The mission-oriented agencies have long supported projects involving entire laboratories so that funding might come from an umbrella grant rather than on a project-by-project basis. Even the NSF has begun to support science and technology centers, such as the controversial Supercomputing Centers. Physicists apply to use the centers on a competitive basis since equipment is too expensive for duplication at individual universities. Thus, research physicists increasingly travel to research sites and work with large groups. In high-energy physics, group sizes range up to several hundred physicists. Critics fear that funding for big science will cut off funding for individuals and small groups, particularly those working in universities. Lack of funding can destroy individual creativity and discourage physics research that does not fall in the mainstream of current thinking, as well as damaging the education of graduate students.

Individual states have recently begun to fund some physics research in efforts to promote economic development. State funding has become important in attracting federally funded science and technology centers. Because many states mandate a balanced budget, recessions force immediate cuts in the support of physics research. Similarly, physics research in high-technology industries has been cut during economic hard times. The link between physics research and economic development that produced comfortable funding levels during periods of economic growth and led to the expansion of the physics community has led to job crises during economic recessions.

3.2 Social Impacts of Physics and Related Technologies

Physics discoveries and related technologies have influenced almost every aspect of American life. Information flows around the country by electronic networks, and desktop publishing allows literally any private citizen to develop a newsletter. The flood of information available swamps individual abilities to absorb and process it. The way in which business is done and policy made depends increasingly on information exchange. In many ways, the value of information as a commodity in its own right has vastly increased.

Electronic communications link people in different countries instantly. National borders pose no barrier to information flow. Pictures of atrocities in Bosnia and starvation in Somalia reach American television screens and change national policy. Citizens actually watch wars live and in color on television signals relayed by satellites. All aspects of life have become encoded in various databases. Marketing can be done by targeting a very narrow group with special interests. Access to databases provides information on such private matters as driving record, credit rating, or even medical history. Even white-collar criminals take advantage of today's communications technologies.

The computer technologies that grew from developments in condensed-matter physics have allowed industry to automate its production processes. Automated production requires fewer workers with higher technical skills. Automated manufacturing has displaced unskilled workers from manufacturing into the service sector. Even a new crop of injuries arises from computer use.

Finally, global communications links and satellite imagery of the Earth from space present the Earth as a small planet whose problems are indeed global in scope. Emphasis in foreign policy has shifted from the interaction among nation states to a discussion of international organizations designed to implement world peace and solve worldwide environmental problems, which themselves are partially defined by physics research. Developments in basic physics thus have an impact on both the mechanics of day-to-day life and the philosophical assumptions that underlie national policy.

3.3 Making Science Policy

As yet, the United States has no entity called science policy although one seems to be under development. For many years, the federal government has carefully avoided even the appearance of interfering in private competition. On the other hand, policies have been implemented to fund research through specialized agencies. The best example of this is the large-scale reorganization of energy research during which the Department of Energy was created. Legislation fostered the development of large-scale technology demonstrations and research directed at new sources of energy, such as the Synfuels project. Similarly, Congress controls the direction of basic research by funding or not funding initiatives, such as the space station or the SSC.

Like other legislation, proposals for science legislation come from members of Congress and the committees they form, as well as from a variety of agencies within the executive branch. No single agency or congressional committee has been charged with making a unified science policy. Outside of a general consensus that national well-being requires support of basic and applied research, there has been little coherent attempt to develop a long-range science policy. It should be noted that while a typical physics experiment requires times on the order of 5 years from conception to publication, the term of a Representative is only 2 years. The time scale for industrial development is at least 10 years, and a weapons system requires 15 years from conception to deployment. Funding cycles run on an annual basis, and elected officials prefer tangible results before the next election.

The problem of long-range planning is particularly acute in physics education, since a graduate physicist typically requires at least 20 years of education. Policies changing elementary science education will bear fruit only after 15 years or so. The NSF and the Department of Education have attempted long-range planning, but political changes happen too quickly to allow a follow through on these policies.

4. PUBLIC KNOWLEDGE OF PHYSICS

One problem in making science policy is public ignorance of science. The work force in a technologically sophisticated society needs a basic knowledge of physics that most American workers lack. Education of physicists and of the general public about physics has become a major concern of a large segment of the physics community.

In the next decade, women, African- and Hispanic-Americans, and other minorities will constitute the majority of new members of the U.S. labor force. These groups, who are presently conspicuously missing from the physics community and college physics courses, will have to be attracted into physics if the profession is to remain healthy.

4.1 Formal Education

Most citizens encounter physics as a part of their formal education in either high school or college. While physical science is part of the middle-school and elementary curriculum, it is hidden in general science courses, which often center on the biological sciences.

4.1.1 The General Public About 20% of high-school graduates have completed a course in physics. These students tend to be those who plan to attend college. High-school courses depend on the quality of the teacher instructing them. Although there are many excellent, creative, and dedicated high-school physics teachers, there are also many teachers who have been recruited from other fields to offer a single course in physics and lack preparation. Thus, a high-school physics course provides widely varying background.

Students who do not enroll in high-school physics must depend on physics taught as part of general science courses for the background they need in their jobs. In the elementary classroom, all science tends to be neglected in favor of teaching basic reading and mathematics. Mathematics may also be neglected in favor of reading. In part, this is due to time pressure in the classroom. In part, it is due to fear on the part of teachers and lack of experience with hands-on physical science. Many teachers complete college training and remain afraid of, and confused by, basic mathematics. They see physical science as an extension of the dreaded math. Elementary science texts tend to be lists of vocabulary words with little stress on experiment. Students are forced to memorize words that they do not comprehend. Understandably, they find science boring.

In middle school, much physics is taught as a vocabulary list. Teachers have limited laboratory space and preparation time, if indeed they have any. Although there are outstanding exceptions, middle-school science remains isolated from the rest of the curriculum, and a book and worksheet course.

One exciting trend in middle-school and high-school education comes through technology education, the old shop teachers. Technology curriculum today can be hands-on and oriented toward solving real world problems and working in groups. The students are prepared to enter the labor force and work in the milieu of high-technology industries. Such students will need to acquire an understanding of the rudiments of physics, which can be presented to these students in such a way as to tie directly to their career goals.

4.1.2 College Graduates Almost all colleges and universities require a laboratory science course as part of their general education curriculum. Few who do not plan to major in the sciences or engineering elect to take physics because physics is popularly perceived as both difficult and dull. Majors in other disciplines, such as premedicine, all too frequently take specially designed courses. They perceive these courses as obstacles in their chosen career paths in which they badly need a grade of A. Courses in physics designed for liberal arts majors struggle to teach physics without mathematics. At best, they provide a strong overview of the most important concepts of physics but may be so superficial as to be misleading. Many introductory physics classes are taught in large lecture sections with laboratories and problem sessions led by graduate assistants. This instructional style may repel bright students who otherwise could succeed in basic physics (Tobias, 1990).

If basic courses in physics do little to increase the physics literacy of non-major college graduates, few college lecture series correct the problem. Distinguished physicists rarely appear on the lecture circuit. Thus, college graduates frequently have little more knowledge of physics than is provided by their high-school educations.

4.1.3 Professional Physicists Issues in training professional physicists center on maintaining a supply of physicists for the research laboratories of the next century. In addition to the problems of attracting women and minority students to physics in the first place, the first college course in physics discourages many potential physics majors who leave the field.

A second major attenuation of the supply of physicists in the educational pipeline occurs at the entrance to graduate school. Nearly half the students in physics Ph.D. programs are not U.S. citizens. Should the supply of foreign students decrease, physics departments could not maintain the current supply of physicists. Universities, critics claim, are using graduate students to support their research efforts without realistic expectations that these students can eventually find jobs. Others argue that the current poor job market is largely due to the recession, which has been particularly damaging to high-technology industry that employs the majority of physicists who work in industry.

Industries expect physics graduates to be experienced in use of state-of-the-art technology, which is not available in many university laboratories. The DOE, NSF, and even DOD have initiated programs to upgrade university instrumentation, but the problem remains serious and is exacerbated by the rising cost of modern sophisticated systems.

Although funding for research has increased, funding per physicist in academia has declined. Thus, junior faculty are caught in a cycle of grant applications and hasty experimentation that certainly damages their teaching performance. The extreme pressure to publish in order to get tenure at a major university may inhibit creativity, as young physicists pursue established lines of inquiry that are more likely both to attract outside funding and to be published in established journals.

4.2 Informal Sources of Information

Since most Americans gain a very limited knowledge of physics during their formal education, they rely on a variety of media and experiences to increase their knowledge of current events in physics. Many are fascinated by breaking events in high-energy physics and cosmology, which basic courses rarely cover.

4.2.1 Magazines, Journals, and Newspapers Most people interested in physics rely on popular magazines such as *Scientific American* or *Science News*. These provide solid coverage of major science issues and are available on newsstands or by subscription. Leading physicists prepare articles for these journals aimed at an audience of nonspecialists.

Few interested citizens read technical journals, which are geared for specialists. These journals provide an avenue for communication within the field and are generally subject to a peer review process. They are criticized for belonging to the same old boys' network doing grant proposal reviews. Also, many journals have increased their cost so that libraries find it difficult to maintain their collections. In addition, much information travels through preprints, which are sent only to people whose names appear on selected lists, thus tightening an elite network. Various suggestions to these problems are being discussed, including a variety of ways to use electronic mail for publishing.

Major newspapers employ reporters who specialize in science writing. Coverage by these papers is accurate and interesting but limited in scope, as physics must compete with medicine or political analysis for space. Local papers rarely have science reporters on their staffs and inform their readers on technical issues only when those issues make headlines, as during the Chernobyl reactor accident or the cold-fusion incident. Local coverage of physics issues varies widely in quality and does not provide a major source of public information.

4.2.2 Science Museums Following the lead of the Chicago Museum of Science and Industry and San Francisco's Exploratorium, many cities have established museums dedicated to hands-on science exhibits. These museums use considerable ingenuity in designing exhibits that entice their visitors, both children and adults, to perform simple experiments. They successfully introduce participants to the basics of physics and convince them that physics can be fun. Such museums have experienced a growing popularity, and their numbers are increasing, but they are still limited to major metropolitan areas.

National laboratories and universities also participate in community outreach programs that provide opportunities for hands-on experiences with science. Such programs frequently target such special groups as students from the inner cities or especially talented high-school students. In areas without museums, college or high-school physics teachers frequently provide a sort of traveling road show to carry physics demonstrations to younger students.

4.2.3 The Electronic Media and Films Radio and television provide some coverage of physics. National Public Radio carries news stories related to physics on such programs as "All Things Considered." Public television has taken a lead in presenting programs like "Nova" and "Newton's Apple," which are designed to educate the interested public. Other cable networks such as the Arts and Entertainment Network also produce excellent physics programs. Such media presentations do an excellent job of making physics accessible to the viewing public.

The major networks provide less regular coverage of developments in physics. Network news shows present developments related to physics, but the coverage is limited by time and depth. Occasionally a science-related issue will surface on one of the news magazine programs, but the networks provide little if any programming at the level of "Nova." In general, physicists who appear in television programs are either mad scientists or absent-minded professors (Weart, 1988). The public image of physics and physicists presented on television and in films generally lacks the glamour afforded either lawyers or doctors. Thus, the public not only fails to gain information from the commercial electronic media but also is given little reason to want to learn about physics.

5. THREE POST–WORLD WAR II CASE STUDIES CONTRASTED WITH THE U.S.

Physics has traditionally been an international enterprise. Although the criteria and goals of basic research cross national borders, even a cursory examination shows that funding patterns and research philosophies differ widely from country to country. In particular, different countries have defined widely different roles for national governments and industry in directing and funding basic physics research, which in turn has played differing roles in fostering industrial development.

The three case studies presented below study nations or groups of nations with well-developed science establishments. European science experienced a long, slow growth without the sudden impetus that World War II provided for the U.S. The Confederation of Independent States (CIS) provides a model for sudden growth of science under government central planning. The Japanese model describes the role played by an initial partnership between government and industry that continues to dominate research.

None of these models applies to most of the developing nations. Even such relatively industrialized countries as Australia carry a legacy of colonial domination that still hampers their scientific development (Rowayne and Boag, 1989). In much of Africa and Latin America, colonial powers oriented any scientific development toward the agricultural and mineral resources they wanted the colony to produce. They trained the native population only sufficiently to ensure production. Newly independent colonies found themselves with little scientific infrastructure and almost no technically trained personnel. Attempts to develop human capital and scientific infrastructure have been inhibited by political instability and the tendency of scientifically trained personnel to try to return to the colonial institutions. Universities and laboratories are poorly staffed and equipped, and industries fund little research despite elaborate declarations of science policy (Forje, 1989; Stann, 1993). Many of these nations have attempted to follow the model of centralized development used by the Soviet Union.

India has managed to develop an indigenous scientific establishment and several sophisticated industries, including space-launch capability and a nuclear explosive. In this case, the colonial pattern was overcome by the leadership of Nehru, who stressed the importance of developing an educated population (Rahman, 1990). The Indian model followed that of Britain. The newly industrialized nations of Southeast Asia have followed a still different pattern in copying the ideas used by Japan. Thus, the three examples below, in addition to the U.S., show the basic patterns of scientific development that have been copied by other nations.

5.1 Western Europe

World War II devastated European economies and physics laboratories. Before the war, major developments in physics had been driven by European scientists working in a small-science model rather than in the big-science model that developed during World War II. Following the war, nations faced the problems of rebuilding their scientific enterprises (Anderson, 1992; Franklin, 1988; Merton and Gaston, 1977).

Many countries established a system of national research institutes, which provided direct funding to laboratories rather than to individuals. Universities funded basic research equipment. Thus, physicists working on small-science projects were supported over the long term without the constant need to seek funding by writing grants. This steady, relatively noncompetitive funding has encouraged innovative lines of research reflected in the Nobel prizes awarded for the discoveries of high-temperature superconductivity, the quantum Hall effect, and scanning tunneling microscopy. On the other hand, the model does not encourage competitive exploitation of the new discoveries and may induce a more leisurely research pace than is found in U.S. laboratories. Current financial hard times have led to funding cutbacks and debates over the proper role of research and teaching, such as the discussion of the abolition of tenure in the U.K. Funding for small science is still done on a national basis, and one of the problems facing the European Community is encouraging collaborations that cross national borders.

In arenas such as high-energy physics requiring large, expensive equipment, European nations could not afford individual efforts. The tradition of establishing a European center of excellence that would collect the best scientific minds in Europe and provide equipment for their experimental work began with the European Laboratory for Particle Physics, CERN. CERN is located in Geneva, Switzerland, and funded by annual contributions from 17 member nations. The laboratory works closely with the national research institutes and universities in member states. The political issues of national "ownership" have been handled by rotating high-level administrative positions and taking decisions by consensus. Experimental resources are allocated strictly on the basis of merit, and all members have equal access to facilities for training their students.

Despite its success in leading research in high-energy physics, CERN has remaining problems. The recession in Europe has caused member states to question the value they obtain for their contributions. Germany, which is paying the expenses of reunification, and the United Kingdom, which has severe economic problems, have debated reducing their commitments to CERN. Spain, which is a major contributor, has very few physicists actually working at CERN and questions whether it is getting its money's worth. Nevertheless, CERN continues to produce exciting results, and plans to build the Large Hadron Collider (LHC) in the existing tunnel housing the Large Electron Positron storage ring. The LHC was designed as a rival for the larger SSC, but CERN believes it can be built faster and cheaper in the existing tunnel, although critics point out that it is more technically challenging. Like SSC, LHC has been the subject of a major funding debate. Other areas of physics have begun international cooperation including the European synchrotron x-ray source at Grenoble, the Joint European Torus fusion laboratory in Culham, England, the Institut Laue-Langevin at Grenoble, and the European Southern Observatory.

Industrial research in European nations has been closely linked to government support. The strength of the ties varies from country to country. In some cases, such as the development of the Concorde or Airbus, government support from several countries has assisted industrial development. European industries are so strongly supported by governments that the U.S. blocks mergers with European firms in industries that are considered critical to national defense. Unlike funding in the U.S., national funding for specific industrial projects has been the rule rather than the exception in Western Europe.

5.2 The Former Soviet Union

Marxist doctrine held that science should dominate the growth of the communist state. Thus, the communist government of the Soviet Union stressed the growth of science and its direct role in advancing the planned economy. All science funding was channeled through the All-Union Academy of Sciences, which also served as the employer of most of the scientists in the Soviet Union. The Academy was composed of 365 institutes, including such prestigious physics research centers as the Lebedev Institute. Except for a few university departments in large cities, the Academy performed all basic physics research. Ranking physicists in the Soviet nuclear weapons program were members of the Academy. Almost all the institutes were concentrated in Moscow and Leningrad (St. Petersburg), except for a few science centers like Novosibirsk. National science academies formed in individual republics were dominated by the All-Union Academy (Dickman, 1991; Fortescue, 1986; Hutchings, 1976; Medvedev, 1978).

Basic science research was more or less tightly controlled by the Kremlin. Although physics escaped the political disaster associated with Lysenko, many physicists were isolated from their field and denied foreign travel for political "errors." Equipment tended to be poor in basic research laboratories, although Soviet physicists are famous for work in areas such as mathematical models. Certainly, basic physics principles were readily applied to the Soviet military research and development programs, although they seem to have found less application in civilian industries. The 1957 launch of Sputnik demonstrated Soviet determination to become a technological superpower.

The Soviet approach to military development concentrated on incremental improvements using stable design bureaus rather than on the radical new designs favored by the U.S. Universities, with a handful of notable exceptions such as Moscow State University, concentrated primarily on teaching.

Public education in the Soviet Union stressed physics literacy, and high-school graduates had far greater exposure to physics and mathematics than in the U.S. Popular physics magazines such as *Quantum*, now available in the U.S., enjoyed wide circulation among both students and citizens. The state-controlled media educated nonspecialists about physics. Scientists were held in great social esteem and enjoyed high salaries and perquisites such as good apartments, cars, and *dachas*.

With the collapse of the Soviet Union following the 1990 coup attempt, Soviet physicists found themselves faced with severe problems. Inflation reduced salaries, which had been generous, and scientists were forced to struggle for survival in very basic terms. In addition, research money dried up. Institute directors bartered for food supplies. Many physicists accepted appointments abroad, and many areas of science suffered a brain drain, particularly of younger people, in areas where Soviet science had made outstanding progress. Western governments worried about Soviet weapons designers accepting employment with third-world nations and funded programs for them, but western help has been slower to arrive for other physicists. Finally, the All-Union Academy has become the Russian Academy of Science and is embroiled in political strife. The other national academies are asserting their newfound independence (Dickman, 1991).

The centrally controlled model of funding and directing basic physics research and policies creating an elite social position for scientists produced an outstanding scientific enterprise in the former Soviet Union. It is anybody's guess whether physics can maintain its leading position in the turbulent political and economic times that have followed the fall of communism.

5.3 Japan

Western scientific tradition conflicts sharply with such Japanese values as respect for seniors, refusing to criticize others in public, or even lifelong employment. Physics developed more slowly in Japan than other sciences such as seismology and biology because it was seen as uniquely western in its values (Bartholomew, 1989). By the time of World War II, Japanese physics had matured to world-class status and produced its first Nobel prizes (Yukawa in 1949, Tomonaga in 1965, and Esaki in 1973), but the structure of research funding and policy still reflected traditional values. Modern Japanese physics continues to be a mix of traditional and western approaches (Anderson *et al.*, 1992; Maddox and Swinbanks, 1992; Shigeru *et al.*, 1974).

Faculty in the 98 national universities in Japan elect their presidents; however, Monbusho, the powerful Ministry of Education, appoints the rest of the administrative staff, who actually run the campus. Monbusho finances the national universities, including all research, and oversees and subsidizes the private universities. Unlike American professors, Japanese professors depend mostly on the university for their research funds. Although the government has made several recent efforts to provide sources of outside grant funding, these are still limited.

Universities are structured around chairs or *koza*, which are allocated to disciplines. The allocations are political decisions and may not be driven by student demand or need of the university. Each *koza* has one professor, one assistant professor, and two assistants and is allocated about $65,000 per year for expenses. After the university and department have collected overhead, the *koza* unit has about $8000 per year for research. As each member of the unit has a guaranteed lifetime job, pressure to do cutting-edge research is not great. University facilities tend to be run down and equipment poor.

Japanese industries perceive basic research as a key to retaining a lead in high-value–added, technology-based industries. They support private laboratories, where researchers are encouraged to design their own research programs within very broad guidelines established by the management. The Canon Research Institute, for example, automatically grants $8000 to employees to explore their new ideas, and grants further funds if the idea looks interesting. At present, Japanese industry supports 39% of the basic research done in the country, although industrial managers designate their work as applied to the development of products, while only universities do basic research.

The research efforts are coordinated, at least to some extent, by the Ministry for International Trade and Industry (MITI), which also jointly funds some industrially oriented research. The Science and Technology Agency (STA) also supports laboratory work in such fields as materials science and condensed-matter physics, which are of interest to industry. This research partnership between the government and the industrial laboratories has worked very effectively in using research results quickly to develop new commercial products. Other nations, notably the U.S., are studying the system to facilitate technology transfer from research labs into industry.

The government, usually through MITI, also funds a system of national laboratories devoted to various sorts of research. These laboratories pursue projects of interest to industry and also support "blue sky" ideas through grants targeted at frontier fields and young researchers. Many experimental programs are funded for a definite limited time, typically five years, and reviewed at the end with a possibility that they will be canceled if they do not show promising developments. Like the industrial laboratories, government laboratories are well equipped and well funded. Major physics projects, including the SuperKamiokande neutrino detector, the JT-60 tokomak, and KEK, the National Institute for High Energy Physics, are government institutes. In some cases, the government laboratories encourage participation by university faculty.

In the past, Japanese science has discouraged foreign participation, although Japanese students have always traveled abroad. Today, industry and the government have programs to encourage foreign participation in Japanese research. These programs are new and starting slowly. Similarly, Japanese research papers have gradually become of greater interest to the outside world. Japanese companies support this new stress on basic research and foreign integration in Japanese labs. They speak of doubling their research budgets even though the country is in an economic recession. It remains to be seen whether their efforts will move Japanese research into the mainstream of world science.

6. SUMMARY: MUTUAL INFLUENCES

In all cultures, physics research leads to the development of new technologies. These technologies, in turn, change the society that uses them.

The same cycle operates equally effectively in reverse. Societies determine the technologies they desire to have and encourage their development by funding and regulating basic research. Technologies make possible new types of physics research. Technology developers demand particular basic research results.

The lines of communication among physics, technology, and society are difficult to trace. Even the boundaries between these areas become blurry. Physics is not done in a vacuum, and so the interaction among physics and technology and society forms a support for the creation of and a background for the analysis of basic results in physics.

GLOSSARY

Applied Research: Research designed to produce a specific product or technology.

Basic Research: Research designed to answer fundamental questions about the nature of the universe.

Science Policy: Legislation and executive orders designed to control what research is done and how it is conducted through mechanisms such as funding or regulation.

Technology Assessment: Evaluation of the scientific merit of a new technology and its possible effects on society and the environment.

Technology Transfer: Incorporation of basic research results into commercially useful products.

Works Cited

Anderson, A. (Ed.) (1992), *Science* **256**, 457–488.

Anderson, A., Keiner, D., Voss, D. (Eds.) (1992), *Science* **258**, 561–592.

Barash, D. P. (1987), *The Arms Race and Nuclear War*, Belmont, CA: Wadsworth Publishing Company.

Bartholomew, J. R. (1989), *The Formation of Science in Japan: Building a Research Tradition*, New Haven, CT: Yale Univ. Press.

Broad, W. J. (1985), *Star Warriors*, New York: Simon and Schuster.

Broad, W. J. (1992), *Teller's War: The Top-Secret Story Behind the Star Wars Deception*, New York: Simon and Schuster.

Brunn, S. D., Leinbach, T. R. (Eds.) (1991), *Collapsing Space and Time: Geographic Aspects of Communications and Information*, London: Harper Collins Academic.

Bush, V. (1945), *Science, the Endless Frontier*, Washington, DC: National Science Foundation; reprinted in 1960.

Carnegie Commission on Science Technology and Government (1992a), *Science and Technology in U.S. International Affairs*, New York: Carnegie Corporation.

Carnegie Commission on Science Technology and Government (1992b), *Partnerships for Global Development: The Clearing Horizon*, New York: Carnegie Corporation.

Carnegie Commission on Science Technology and Government (1993), *Science, Technology and Government for a Changing World*, New York: Carnegie Corporation.

Chalmers, W. S., Isenson, R. (1966), *First Interim Report on Project Hindsight*, Washington, DC: Office of the Director of Defense Research and Engineering.

Clinton, W. J., Gore, A. (1993), *Technology for America's Economic Growth, A New Direction to Build Economic Strength*, Washington, D.C.: U.S. G.P.O.

Cohen, B. (1988), in: E. Mendelsohn, M. R. Smith, P. Weingart (Eds.), *Science, Technology, and the Military*, Dordrecht: Kluwer Academic.

Dickman, S. (1991), *Science* **254**, 1716–1719.

Ellis, S. D., Mulvey, P. J. (1992), *Employment Survey 1991*, New York: American Institute of Physics.

Forje, J. W. (1989), *Science and Technology in Africa*, Burnt Mill, Harlow, Essex, U.K.: Longman.

Fortescue, S. (1986), *The Communist Party and Soviet Science*, Baltimore, MD: The Johns Hopkins Univ. Press.

Franklin, M. N. (1988), *The Community of Science in Europe*, Aldershot, UK: Gower.

Ganley, O. H., Ganley, G. (1989), *To Inform or To Control? The New Communications Networks*, 2nd Ed., Norwood, NJ: Ablex Publishing Corporation.

Golden, W. T. (Ed.) (1980), *Science Advice to the President*, New York: Pergamon.

Golden, W. T. (Ed.) (1988), *Science and Technology Advice to the President, Congress, and Judiciary*, New York: Pergamon.

Huizenga, J. R. (1993), *Cold Fusion: The Scientific Fiasco of the Century*, New York: Oxford Univ. Press.

Hirano, Y. (1992) in: A. Anderson, D. Keiner, D. Voss (Eds.), *Science* **258**, 561–592.

Howes, R., Fainberg, A. (Eds.) (1991), *The Energy Sourcebook: A Guide to Technology, Resources, and Policy*, New York: American Institute of Physics.

Hutchings, R. (1976), *Soviet Science, Technology, Design: Interaction and Convergence*, London: Oxford Univ. Press.

Krass, A. S. (1985), *Verification: How Much Is Enough?*, Lexington, MA: Lexington Books.

Kuehn, T. J., Porter, A. L. (Eds.) (1981), *Science, Technology, and National Policy*, Ithaca, NY: Cornell Univ. Press.

Lundstedt, S. B. (Ed.) (1990), *Telecommunications, Values, and the Public Interest*, Norwood, NJ: Ablex Publishing Corporation.

MacKenzie, D. (1990), *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance*, Cambridge, MA: MIT Press.

Maddox, M., Swinbanks, D. (1992), *Nature* **359**, 573–582.

Mallove, E. F. (1991), *Fire from Ice: Searching for the Truth Behind the Cold Fusion Furor*, New York: Wiley.

Medvedev, Z. A. (1978), *Soviet Science*, New York: Norton.

Mendelsohn, E., Smith, M. R., Weingart, P. (Eds.) (1988), *Science, Technology and the Military*, Dordrecht: Kluwer Academic.

Merton, R. K., Gaston, J. (Eds.) (1977), *The Sociology of Science in Europe*, Carbondale, IL: Southern Illinois Univ. Press.

Pool, I. de S. (1990), *Technologies Without Boundaries*, Cambridge, MA: Harvard Univ. Press.

Pool, R. (1990), *Science* **250**, 754–755.

Rahman, A. (1990), *Science and Technology in India, Pakistan, Bangladesh and Sri Lanka*, Burnt Mill, Harlow, Essex, U.K.: Longman.

Rhodes, R. (1986), *The Making of the Atomic Bomb*, New York: Simon and Schuster.

Rowayne, J., Boag, C. (1989), *Science and Technology in Australasia, Antarctica and the Pacific Islands*, Burnt Mill, Harlow, Essex, U.K.: Longman.

Schroeer, D. (1984), *Science, Technology and the Arms Race*, New York: Wiley.

Scowcroft, B. (1983), *Report of the President's Commission on Strategic Forces*, Washington, DC: Department of Defense.

Shigeru, N., Swain, D., Eri, Y. (Eds.) (1974), *Science and Society in Modern Japan: Selected Historical Sources*, Cambridge, MA: MIT Press.

Smith, B. L. R. (1990), *American Science Policy Since World War II*, Washington, DC: The Brookings Institution.

Stann, E. J. (Ed.) (1993), *Science and Technology in the Americas: Perspectives on Pan American Collaboration*, Washington, DC: American Association for the Advancement of Science.

Study Group on Science and Technology of Directed Energy Weapons (1987), *Rev. Mod. Phys. Suppl.* **59**, S1–S202.

Tobias, S. (1990), *They're Not Dumb, They're Different: Stalking the Second Tier*, Tucson, AZ: Research Corporation.

U.S. Congress Office of Technology Assessment (1981), *MX Missile Basing*, Washington, DC: U.S. G.P.O.

U.S. Congress Office of Technology Assessment (1985), *Ballistic Missile Defense Technologies*, Washington, DC: U.S. G.P.O.

U.S. Congress Office of Technology Assessment (1988), *SDI: Technology, Survivability, and Software*, Washington, DC: U.S. G.P.O.

Vogelsang, W.F., Barschall, H.H. (1991), in: R. Howes, A. Fainberg (Eds.), *The Energy Sourcebook: A Guide to Technology, Resources, and Policy*, New York: American Institute of Physics.

Weart, S. (1988), *Physics Today* **41** (6), 28–37.

Weinberg, A. M. (1967), *Reflections on Big Science*, Cambridge, MA: MIT Press.

Further Reading

Golden, W. T. (Ed.) (1988), *Science and Technology Advice to the President, Congress, and Judiciary*, New York: Pergamon. A collection of essays by participants that illuminates all aspects of providing science advice to policy makers.

Kuehn, T. J., Porter, A. L. (Eds.) (1981), *Science, Technology, and National Policy*, Ithaca, NY: Cornell Univ. Press. Essays providing a variety of views on the interaction of science, technology and society.

Mendelsohn, E., Smith, M. R., Weingart, P. (Eds.) (1988), *Science, Technology and the Military*, Dordrecht: Kluwer Academic. A very readable two-volume series of papers, including several case studies of the interaction of physics and technology with one of the most conspicuous aspects of modern society.

Smith, B. L. R. (1990), *American Science Policy Since World War II*, Washington, DC: The Brookings Institution. A brief, comprehensive, and readable history of science and society in the United States.

PHYSIOLOGICAL ACOUSTICS

See ACOUSTICS, PHYSIOLOGICAL

PHYSIOLOGICAL OPTICS

See OPTICS, PHYSIOLOGICAL

PIEZOELECTRIC DEVICES

ARTHUR BALLATO AND JOHN R. VIG, *U.S. Army Research Laboratory, Fort Monmouth, New Jersey, U.S.A.*

Introduction 129

1. **Fundamental Properties of Piezoelectric Materials** 130

1.1 Piezoelectricity—Definition and Mathematical Description 130

1.2 Electromechanical Coupling Factors 131

1.3 The Piezoelectric Effect and Crystal Symmetry 131

1.4 Pyroelectricity 132

1.5 Ferroelectricity 132

1.6 Piezoelectric Polymers 133

1.7 Piezoelectric Semiconductors . 133

1.8 Numerical Values of Coupling Factors 134

2. **Applications of Piezoelectricity** 134

2.1 Oscillators, Clocks, and Filters 134

2.1.1 Introduction 134

2.1.2 Equivalent Circuit of Crystal Unit 135

2.1.3 Oscillators 136

2.1.4 Factors Affecting Frequency Stability 137

2.1.4.1 Temperature 137

2.1.4.2 Aging and Drift 138

2.1.4.3 Short-Term Stability 139

2.1.4.4 Thermal Hysteresis and Retrace 139

2.1.4.5 Drive Level 139

2.1.4.6 Acceleration, Vibration, and Shock 139

2.1.4.7 Radiation 140

2.1.5 Filters 140

2.2 Sensors and Transducers 140

2.3 Actuators, Motors, Smart Materials and Devices 141

2.4 Optical and Acousto-optical Devices 142

2.5 Underwater Sound Systems, Ultrasonic Cleaners, Medical Ultrasonics, and Nondestructive Evaluation 142

2.6 Other Piezoelectric Devices ... 143

Glossary 143

Works Cited 144

Further Reading 145

INTRODUCTION

Piezoelectricity was discovered by the brothers Curie in 1880. The Curies found that in certain materials such as zincblende, tourmaline, cane sugar, topaz, and quartz, mechanical stresses were accompanied by the production of electric surface charges. The following year, Lippmann, from thermodynamic considerations, predicted the converse effect: an imposed voltage produces mechanical deformations. The name "piezo" derives from the Greek, meaning "to squeeze"; in more modern terminology, we say that the effect couples the electric and elastic field variables (Cady, 1964).

The piezoeffect remained a curiosity until the early 1920s when its presence in quartz was utilized to realize crystal resonators for the stabilization of oscillators, thereby launching the field of frequency control. With the introduction of quartz control, timekeeping moved from the sun and stars to small, man-made sources that exceeded astronomy-based references in stability. Since then, the applications of devices based on piezoelectricity have expanded dramatically. The quartz resonator has continued to evolve to become a device capable of precision one million times greater than the original. It also serves as the "flywheel" in atomic frequency standards, which make frequency the most accurate entity known.

Piezoelectric devices are additionally used as transducers in applications from telephone speakers to sonar arrays; newer and

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

rapidly burgeoning areas of utilization are those of sensors, actuators, and microelectromechanical structures (MEMS). In many traditional applications, applied voltages conform to the norms for electronic circuits; accompanying mechanical displacements are then usually nanometers or less. Newer configurations available today for actuators often have much larger amplitudes at comparable voltages.

One of the most appealing aspects of piezoelectricity for modern applications is the compelling immediacy and simplicity of the transduction mechanism. Micro- and nanoelectronics are built upon the behavior of charged species subjected to electric fields; the extreme miniaturization of these structures is owed in large measure to their ability to take on a capacitorlike form. Bulky inductors are banished and done without, and thin, planar electrodes introduce the electric fields to operate the circuits. These fields provide exactly the forces required to drive mechanical motions in a piezoelectric device. Elastic field/mechanical considerations are therefore incorporated into the operations of modern electronic components in a clean, efficient, and very direct manner, by making use of the voltages resident on the chips.

1. FUNDAMENTAL PROPERTIES OF PIEZOELECTRIC MATERIALS

1.1 Piezoelectricity—Definition and Mathematical Description

Formally stated, piezoelectricity couples the mechanical stress (T) and strain (S) fields with the electric intensity (E) and displacement (D) fields (Nye, 1957, 1985; Mason, 1966; IEEE, 1978). Depending on variable choice, four sets of material constitutive relations are defined for the linear case of insulating dielectric materials; these are, in compressed matrix form,

$$[T] = [c^E][S] - [e]'[E];$$
$$[D] = [e][S] + [\epsilon^S][E];$$

$$[T] = [c^D][S] - [h]'[D];$$
$$[E] = -[h][S] + [\beta^S][D];$$

or

$$[S] = [s^E][T] + [d]'[E];$$
$$[D] = [d][T] + [\epsilon^T][E];$$

$$[S] = [s^D][T] + [g]'[D];$$
$$[E] = -[g][T] + [\beta^T][D].$$

Because the electric variables $[E]$ and $[D]$ are tensors of rank 1 (i.e., three-dimensional vectors), and the mechanical variables $[T]$ and $[S]$ are tensors of rank 2 that have been converted to six-dimensional vectors, $[c^E]$, $[c^D]$, $[s^E]$, and $[s^D]$ are (6×6) symmetric matrices representing the elastic stiffnesses and compliances at constant E and D; $[\epsilon^S]$, $[\epsilon^T]$, $[\beta^S]$, and $[\beta^T]$ are symmetric (3×3) dielectric permittivity and impermeability matrices at constant S and T, while $[e]$, $[h]$, $[d]$, and $[g]$ are (3×6) piezoelectric matrices. A prime denotes transpose.

The material constitutive relations given may be further broadened to include couplings to thermal, magnetic, and other fields. Figure 1 depicts the phenomenology arising from the mutual couplings among the electric, elastic, and thermal fields. The outer quantities are the thermodynamic intensive variables (generalized forces); the inner quantities are the extensive variables (generalized displacements).

Piezoelectricity is a linear effect; reversal of the electric field reverses the mechanical deformation. Conditions for its existence in crystals are given below. Electrostriction, on the other hand, exists in all dielectric solids. It is a deformation quadratic in the applied electric field. Biased electrostriction, where small electric field variations are superimposed on a constant component, is phenomenologically equivalent to linear piezoelectricity; this artifice may be used with nonpiezoelectric crystals such as silicon, but the coupling depends upon the bias, and is often not large.

Resonance measurements on thickness-mode plate vibrators and pulse-echo (transit-time) measurements determine values for the matrix elements $[c^E]$, $[e]$, and $[\epsilon^S]$, from which the other material parameter sets may be computed if desired. Resonance measurements on low-frequency bar and rod specimens lead alternatively to determination of the set $[s^E]$, $[d]$, $[\epsilon^T]$. These dynamic techniques are both more accurate and more precise than static measurements.

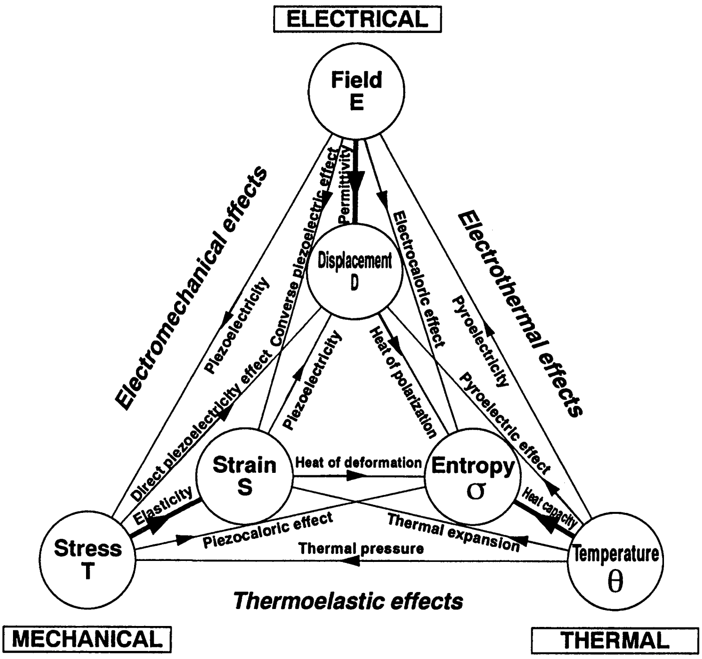
ELECTRICAL
Field
E
Permittivity
Displacement
D
Electromechanical effects
Electrothermal effects
Converse piezoelectric effect
Direct piezoelectricity effect
Piezoelectricity
Piezoelectricity
Electrocaloric effect
Pyroelectricity
Heat of polarization
Pyroelectric effect
Strain
S
Heat of deformation
Entropy
σ
Elasticity
Heat capacity
Stress
T
Piezocaloric effect
Thermal expansion
Thermal pressure
Temperature
θ
Thermoelastic effects
MECHANICAL
THERMAL

FIG. 1. The phenomenology arising from the mutual couplings among the electric, elastic, and thermal fields.

1.2 Electromechanical Coupling Factors

Electromechanical coupling factors k ($0 \leq k < 1$) are dimensionless scalar measures of efficacy of piezoelectric transduction, and are far more important than the piezoelectric constants taken by themselves. They appear in considerations of bandwidth and insertion loss in transducers and signal-processing devices, in location and spacing of critical frequencies of resonators, etc. For high-frequency plate resonators, they have the generic form $k = e/\sqrt{\epsilon c}$; for low-frequency bar and rod vibrators, the form becomes $k = d/\sqrt{\epsilon s}$; alternative forms using the h and g piezo parameters are used as well. Here, e, d, ϵ, c, and s are scalars, contracted from the material tensors according to the geometry at hand. These quantities are also called piezocoupling factors; representative values are listed below.

1.3 The Piezoelectric Effect and Crystal Symmetry

Crystallographers have, over the centuries, catalogued the various geometrical arrangements of atoms that can exist in crystal structures. The starting point is the unit cell, which normally contains the smallest grouping of atoms representing the chemical formula of the substance. In a crystal, this cell, with its particular size, shape, and internal arrangement of atoms, is repeated over and over in three dimensions. The result is an orderly stacking of cells that has translational invariance, meaning that moving one's observation point by an integral number of cells in any direction yields a completely equivalent view. The shapes of the unit cells suitable for such stacking fall into seven general types, or systems. These are triclinic, monoclinic, orthorhombic (also called rhombic), tetragonal, trigonal, hexagonal, and cubic (also called isometric). When the actual arrangements of atomic species possible within the unit cell are compounded with the possible shapes of unit cells, the total number of independent geometrical arrangements available in three dimensions is 230. These are the space groups.

In most practical situations, one deals with a continuum approximation, where considerations of the exact placement of atoms in the cells are not important, and the situation simplifies considerably. For example, elasticity is described by the few phenomenological stiffness and compliance constants used in the equations above. These numbers arise from a "smearing-out" of the effects of the individual interatomic potentials; only important are the directional aspects stemming from the atomic arrangements, not the arrangements themselves. Variation with direction of continuum properties such as acoustic wave speed is known as anisotropy. All the requisite geometric information specifying the varieties of anisotropy resides in the classification that the crystallographers call the point groups. These are only 32 in number, and arise simply from the symmetry elements such as two-, three-, four-, or six-fold rotational symmetry, mirror planes, inversion axes, etc. The point groups are described by their Hermann-Mauguin symbols; these are shorthand listings of the symmetries. Alternative notation, due to Schoenflies, is also current.

The crystallographic point-group symmetry elements completely determine the form and symmetries of the elastic, piezoelectric, and dielectric matrices that appear in the constitutive relations given earlier. The specific values of the individual components result, of course, from the particular substance considered.

Of the 32 point groups, 11 have a center of symmetry, and therefore cannot be piezoelectric, because piezoelectricity is a polar effect. Class 432 lacks a center of symmetry, but has other symmetries that, taken together, preclude the existence of piezoelectricity. The 20 piezo classes are triclinic class 1; monoclinic classes 2 and *m*; orthorhombic classes 222 and *mm*2; tetragonal classes 4, $\bar{4}$, 422, 4*mm*, and $\bar{4}2m$; trigonal classes 3, 32, and 3*m*; hexagonal classes 6, $\bar{6}$, 622, 6*mm*, and $\bar{6}m2$; and cubic classes 23 and $\bar{4}3m$.

As examples, we give the [*e*] forms of the piezo matrices pertinent to classes 3*m* (the strong piezoelectrics lithium niobate and lithium tantalate), 32 (quartz), $\bar{4}3m$, and 6*mm* (poled polycrystalline ceramics). The last two classes, $\bar{4}3m$ and 6*mm*, categorize nearly all of the piezoelectric semiconductors, such as silicon carbide and gallium arsenide. For class 3*m*,

$$[e] = \begin{pmatrix} 0 & 0 & 0 & 0 & e_{15} & -e_{22} \\ -e_{22} & e_{22} & 0 & e_{15} & 0 & 0 \\ e_{31} & e_{31} & e_{33} & 0 & 0 & 0 \end{pmatrix};$$

for class 32,

$$[e] = \begin{pmatrix} e_{11} & -e_{11} & 0 & e_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & -e_{14} & -e_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix};$$

for class $\bar{4}3m$,

$$[e] = \begin{pmatrix} 0 & 0 & 0 & e_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & e_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & e_{14} \end{pmatrix};$$

for class 6*mm*,

$$[e] = \begin{pmatrix} 0 & 0 & 0 & 0 & e_{15} & 0 \\ 0 & 0 & 0 & e_{15} & 0 & 0 \\ e_{31} & e_{31} & e_{33} & 0 & 0 & 0 \end{pmatrix}.$$

Two effects that are associated with piezoelectric materials are pyroelectricity and ferroelectricity.

1.4 Pyroelectricity

Of the 20 piezoelectric point groups, ten contain a unique polar axis; crystals in these classes possess electric dipole moments along their polar axes. When the crystals are subjected to a uniform change of temperature, electric charges accumulate on the ends of the polar axes. This phenomenon is known as pyroelectricity; because a temperature change (a scalar number) produces a charge separation, and consequently an electric field (a vector), the pyroelectric coefficients form a tensor of rank 1, i.e., a vector, and would be represented as a (3 × 1) matrix if we were to extend the constitutive equations given earlier to include this effect.

The crystal symmetry classes exhibiting pyroelectricity are 1, 2, *m*, *mm*2, 4, 4*mm*, 3, 3*m*, 6, and 6*mm*. Thus, α-SiC and lithium niobate are pyroelectric, while β-SiC, quartz, and gallium arsenide are not. The presence of piezoelectricity is necessary but not sufficient for a crystal to be pyroelectric.

1.5. Ferroelectricity

The name ferroelectricity is given to those polar crystals capable of having the spontaneous polarization along the polar axis reversed or reoriented, usually in response to an applied electric field. The name is derived by analogy with ferromagnetism, but in the magnetic case the very atoms themselves possess permanent magnetic moments; in the electric case, the permanent electric moment is a property of the placement of the atoms within the unit cells of the structure.

The presence of the ferroelectric state requires that the crystal be of a pyroelectric class; the converse is, however, not true. One cannot conclude *a priori* that a pyroelectric crystal is ferroelectric; one must settle this issue experimentally. The ferroelectric state usually exists only below a certain critical temperature, called the Curie temperature or Curie point. Above this point, the material may belong to a class that is not pyroelectric, or even piezoelectric. Upon cooling through the Curie point, the crystal undergoes a phase transition (which may be either first- or second-order), where the atoms of the structure shift their positions slightly and the crystal changes its symmetry, usually from a nonpolar form, to the symmetry of a pyroelectric class. Since the ferro state has reversible dipoles, upon cooling it is a matter of chance which polarity will be assumed; near the Curie point the structure is very susceptible to disturbance. Because the lattice is relatively unstable around this point, small forces can produce large effects; this is the case with the dielectric constant. This number is the measure of electric displacement produced in response to an applied electric field. Near the Curie point, the dielectric constant assumes very large values; this behavior ("dielectric anomaly") led to the initial discovery of this class of materials. Very often ferroelectrics have quite large piezoelectric coupling values.

Piezoelectric ceramics, such as lead zirconate titanate ($Pb(Ti,Zr)O_3$), are solid solutions comprised of microscopic crystallites, each of which is ferroelectric; the material originally is macroscopically isotropic at all temperatures because of the random orientations of the crystallites. When cooled through the Curie point in the presence of a static electric field (a process called "poling"), the domains are preferentially oriented and frozen into a configuration where the body is macroscopically piezoelectric. It remains isotropic perpendicular to the field axis, but the isotropy is destroyed in the parallel direction. The result is cylindrical polar symmetry, represented by the symbol ∞mm; as far as the elastic, piezoelectric, and dielectric matrices are concerned, the scheme of coefficients is identical to that for the class $6mm$.

Above the Curie point, most (but not all) ferroelectrics belong to one of three classes: 222 (e.g., Rochelle salt), $\bar{4}2m$ (e.g., potassium dihydrogen phosphate, KDP), and $m3m$ (e.g., $BaTiO_3$). In the ferroelectric state, most of the practically important compounds likewise fall into three classes: $4mm$, $3m$, and $mm2$.

1.6 Piezoelectric Polymers

Certain polymers, such as polyvinyl chloride (PVC), polyvinyl fluoride (PVF), and polyvinylidene fluoride (PVF_2 or PVDF), can be made piezoelectric by a combination of mechanical stretching and electric field poling at elevated temperatures. The induced anisotropy varies with the polymer and the subsequent treatment; polymethyl glutamate, for example, has symmetry $\infty 22$, which is equivalent to class 622 for the properties discussed here. PVDF sheet, poled along the thickness, belongs to class ∞mm (equivalent to $6mm$); uniaxial stretching or rolling converts it to orthorhombic class $mm2$, while a biaxial stretching again results in class ∞mm.

1.7 Piezoelectric Semiconductors

In the case of semiconductors, the piezoelectric equations are broadened to incorporate the further coupling to charge transport mechanisms. It is apparently not well appreciated how widespread is the phenomenon of piezoelectricity in semiconductors. The single-element semiconductors silicon and germanium have, of course, no piezo effect because of the inherent symmetry by which the atoms are arrayed in their crystals; they share with diamond the perovskite structure of point group $m3m$. Binary semiconductors, however, because of the dissymmetry of the two atomic species, are generally piezoelectric as well. Table 1 provides a partial listing of these materials, grouped according to the columns of the periodic table. The compounds fall into two categories: those marked "c" belong to the cubic point group $\bar{4}3m$, the sphalerite or zincblende structure, while those indicated with "h" belong to the hexagonal point group $6mm$, the wurtzite structure.

Table 1. Piezoelectric binary semiconductors.

3-5	N	P	As	Sb
B	c	c		
Al	h	c	c	c
Ga	h	c	c	c
In	h	c	c	c

2-6	O	S	Se	Te
Be	h			
Zn	h	$\begin{pmatrix}\alpha\text{-h}\\ \beta\text{-c}\end{pmatrix}$	c	c
Cd		h	h	c

1-7	Br	Cl	I
Cu	c	c	c
Ag			h

4-4	C
Si	$\begin{pmatrix}\alpha\text{-h}\\ \beta\text{-c}\end{pmatrix}$

h = hexagonal, 6mm
(wurtzite)

c = cubic, $\bar{4}3m$
(sphalerite; zinc blende)

Piezosemiconductors have an encouraging future because of the appealing promise of monolithic, integrated, mixed-effect devices; for example, electronic drivers modulating photonic processors via submicroscopic mechanical actuators containing optical components such as lenses, mirrors, etc.

Zinc oxide, ZnO, is one of the most useful piezosemiconductors as a consequence of its interesting set of properties. It is photoconducting, and is transparent in the visible region; it has a broad band gap, fairly high refractive indices, large piezoelectric constants, and strong acousto-optical, electro-optical, and nonlinear optical coefficients. Chemisorption at its surface results in an electron transfer from the ZnO to the chemisorbed molecule. ZnO thin films, usually deposited by sputtering, have a wide range of applications. Examples are as transducers in surface-acoustic-wave (SAW) filters and delay lines, in integrated optical waveguides and acousto-optic Bragg deflectors, as transparent electrode materials in optoelectronic devices such as light-emitting diodes, and as sensors of oxidizing and reducing gases.

1.8 Numerical Values of Coupling Factors

Table 2 shows the piezocoupling factors of a few representative materials from various crystal classes (Hellwege and Hellwege, 1979, 1981, 1982a, 1982b; Madelung, 1982; Gualtieri *et al.*, 1993).

2. APPLICATIONS OF PIEZOELECTRICITY

2.1 Oscillators, Clocks, and Filters

2.1.1 Introduction The piezoelectric material on which most frequency-control, timing, and frequency-selection devices depend is quartz. A quartz crystal can act as a stable mechanical resonator, which, by its

Table 2. Piezocoupling factors of representative materials.

Material/Cut	Class	Coupling factor
Lithium niobate crystal	$3m$	
X cut		$k_{15} = 68.9\%$
36° rotated *Y* cut		$k_{22} = 48.7\%$
Quartz crystal	32	
AT cut		$k_{26'} = 8.8\%$
BT cut		$k_{26'} = 5.0\%$
PZT ceramic solid solutions	∞mm	
Z cut		$k_{33} = 30\%$ to 60%
Polyvinylidene fluoride, PVDF	$mm2$	
Z-normal sheet		$k_{31} = 16\%$
Gallium Arsenide	$\bar{4}3m$	
(110) cut		$k_{26'} = 6.1\%$
(111) cut		$k_{11'} = 4.3\%$
Silicon Carbide, α-SiC	$6mm$	
X or *Y* cut		$k_{15} = 2.5\%$
Z cut		$k_{33} = 3.5\%$
Zinc Oxide	$6mm$	
Z-normal, thin film		$k_{15} = 26\%$
		$k_{33} = 29\%$
Aluminum Nitride	$6mm$	
Z-normal, thin film		$k_{33} = 15.5\%$
		$k_{33} = 24.2\%$

piezoelectric behavior and high *Q*, determines the frequency generated in an oscillator circuit. Bulk-acoustic-wave (BAW) resonators are available in the frequency range from about 1 kHz to 300 MHz. Surface-acoustic-wave (SAW) and surface-transverse-wave (STW) devices can be made to operate at well above 1 GHz.

About 2×10^9 BAW quartz resonators are manufactured annually for oscillator, clock, and filter applications (Hellwig, 1993; Vig, 1993). Hundreds of millions of SAW devices, mostly filters of various types, are manufactured annually (Ruppel *et al.*, 1993). A major SAW device application is TV IF filters. Every TV set produced currently (i.e., in 1994) contains at least one SAW filter. SAW filters play important roles in communication systems, test and measurement instruments (e.g., synthesizers and spectrum analyzers), and remote-control systems. High-frequency oscillators are another important application of SAW devices.

In the manufacturing of quartz resonators (Gerber and Ballato, 1985; Stein and Vig, 1992), wafers are cut from the mother crystal along precisely controlled directions with respect to the crystallographic axes. The properties of the device depend strongly on the angles of cut. After shaping to required dimensions, metal electrodes are applied to the quartz wafer, which is mounted in a holder structure. The assembly is usually sealed hermetically. A BAW resonator is called a crystal unit (or crystal or resonator). SAW devices are excited by interdigital transducers, which consist of two sets of interleaved metal electrodes, called fingers. The resonance frequency of a BAW device depends on the dimensions of the plate. A SAW device's frequency depends on the spacing of the fingers. Much of the remaining discussion will refer to BAW devices, as these are used in the highest production volumes.

To cover the wide range of frequencies, different cuts, vibrating in a variety of modes, are used. The various cuts usually have two-letter names. A *T* in the cut's name indicates a temperature-compensated cut; e.g., the *AT* cut was the first temperature-compensated cut named. The *BT* cut, *CT* cut, *DT* cut, *IT* cut, and *SC* cut are other temperature-compensated cuts. The locus of zero-temperature-coefficient cuts in quartz is shown in Fig. 2. The *X*, *Y*, and *Z* directions are the crystallographic axes, chosen to make the description of properties as simple as possible. The *Z* axis is an axis of threefold symmetry in quartz; i.e., the physical properties repeat every 120° as the crystal is rotated about the *Z* axis. The *AT* cut and the *SC* cut are the most popular cuts for high-stability applications. The *SC* cut has several important advantages over the *AT* cut, as is discussed below.

AT-, *BT*-, and *SC*-cut crystals can be manufactured for fundamental-mode operation at frequencies up to about 200 MHz. *AT*-, *BT*-, and *SC*-cut crystals vibrate in the thickness shear mode. Above 60 MHz, overtone crystals are generally used. Such crystals usually operate at the third or fifth overtone (e.g., a 20-MHz fundamental mode crystal excited on its fifth overtone is a 100-MHz crystal). Below 1 MHz, tuning forks, *X*-*Y* and *NT* bars (flexure mode), +5° *X* cuts (extensional mode), or *CT* and *DT* cuts (face-shear mode) can be used. Several hundred million 32.8-kHz tuning-fork crystals are produced annually for the world watch market.

2.1.2 Equivalent Circuit of Crystal Unit

The circuit designer can treat the behavior of a quartz-crystal unit by considering its (ap-

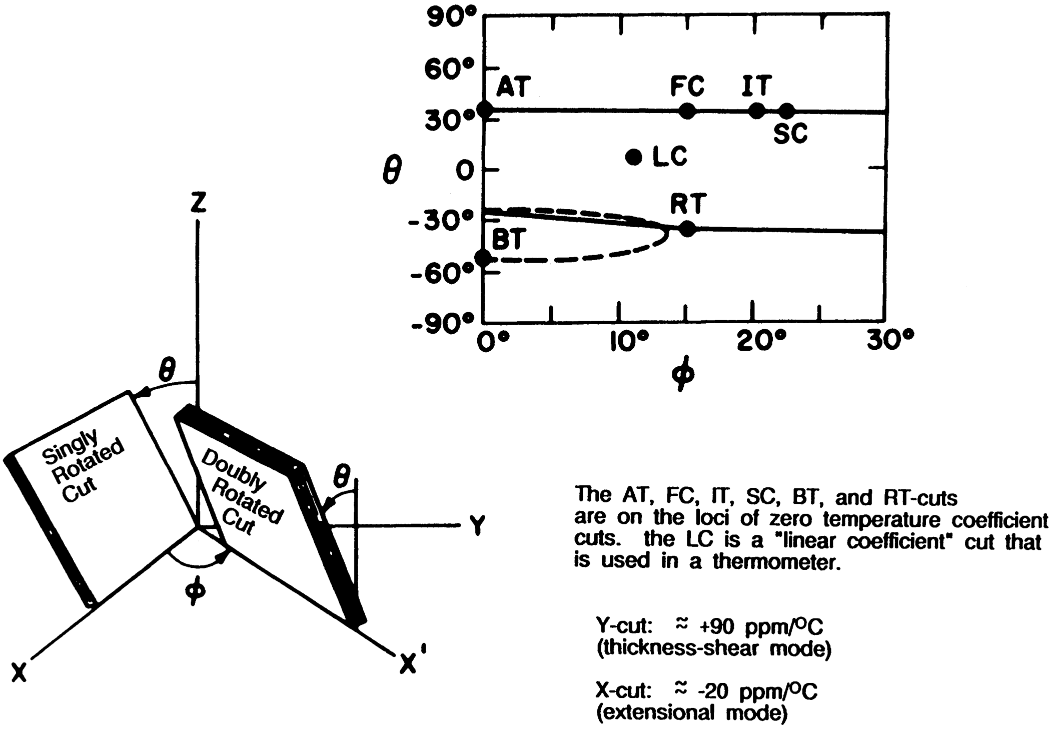

The AT, FC, IT, SC, BT, and RT-cuts are on the loci of zero temperature coefficient cuts. the LC is a "linear coefficient" cut that is used in a thermometer.

Y-cut: ≈ +90 ppm/°C (thickness-shear mode)

X-cut: ≈ -20 ppm/°C (extensional mode)

FIG. 2. The locus of zero-temperature-coefficient cuts in quartz.

proximate) equivalent circuit (Gerber and Ballato, 1985) (Fig. 3). The mechanical resonance in the crystal is represented by L_1, C_1, and R_1. Because it is a dielectric with electrodes, the device also displays an electrical capacitance C_0. The parallel combination of C_0 and the "motional arm," C_1-L_1-R_1, represents the equivalent circuit to a good approximation. As shown in Fig. 4, the reactance of this circuit varies with frequency.

The Q values ($Q^{-1} = 2\pi f_s R_1 C_1$) of quartz-crystal units are much higher than those attainable with other circuit elements. In general-purpose units, the Q is usually in the range of 10^4 to 10^6. The intrinsic Q of quartz is limited by internal losses. For *AT*-cut crystals, the intrinsic Q has been experimentally determined to be 1.6×10^7 at 1 MHz; it is inversely proportional to frequency. At 5 MHz, for example, the intrinsic Q is 3.2×10^6.

2.1.3 Oscillators The commonly used crystal oscillator circuits fall into two broad categories (Gerber and Ballato, 1985). In series-resonance oscillators, the crystal operates at series resonance, i.e., at f_s. In parallel-resonance or antiresonance oscillators, the crystal is used as a positive reactance; i.e., the frequency is between f_s and f_A, where f_A is the antiresonance frequency. In this latter mode of operation, the oscillator circuit provides a load capacity to the crystal unit. The oscillator then operates at the frequency where the crystal unit's reactance cancels the reactance of the load capacitor. When the load capacitance is changed, the oscillator frequency changes. An important parameter in this connection is the capacitance ratio $r = C_0/C_1$. Typically, the value of r is a few hundred for fundamental-mode crystals. It is larger by a factor of n^2 for nth-overtone crystals.

When a load capacitor C_L is connected in series with a crystal, the series-resonance frequency of the combination is shifted from f_s by a Δf which is related to the other parameters by

$$\frac{\Delta f}{f_s} \approx \frac{C_1}{2(C_0 + C_1)}.$$

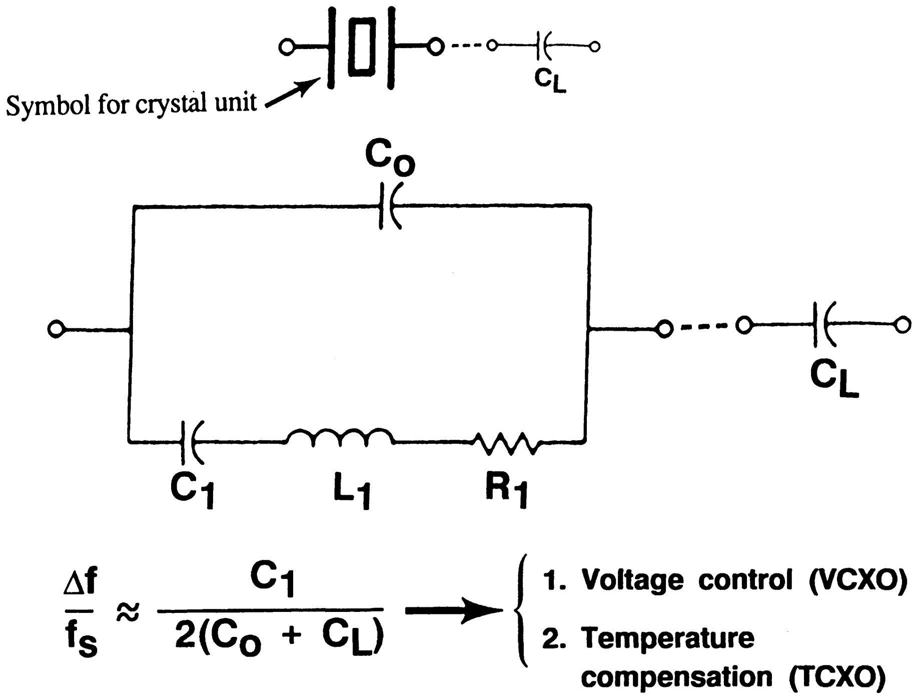

FIG. 3. The equivalent circuit of quartz-crystal unit.

For a typical fundamental-mode *AT*-cut crystal unit with $r = 250$, $C_0 = 5$ pF, and $C_L = 30$ pF, the shift is 286 ppm. If such a crystal unit is to remain stable to, for example, 1×10^{-9}, the load reactance due to C_L and the other circuit components must remain stable to within 1.2×10^{-4} pF. The frequency can also be "tuned" by intentionally changing C_L.

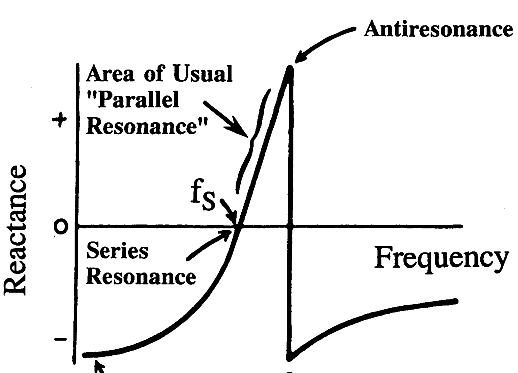

FIG. 4. The reactance of a quartz-crystal unit as a function of frequency.

For the above example, a change of 1 pF in C_L shifts the frequency by nearly 10 ppm.

2.1.4 Factors Affecting Frequency Stability Except for the atomic frequency standards, quartz devices exhibit the highest available frequency stability. However, all frequency sources exhibit frequency instabilities. The main instabilities are as follows (Gerber and Ballato, 1985; Stein and Vig, 1992; Vig and Walls, 1994).

2.1.4.1 Temperature. The frequency-vs-temperature characteristics are determined primarily by the angles of cut of the crystal plates with respect to the crystallographic axes of quartz (Gerber and Ballato, 1985; Stein and Vig, 1992). Typical characteristics are shown in Fig. 5. The points of zero temperature coefficient, the turnover points, can be varied over a wide range by varying the angles of cut.

The frequency–temperature characteristics of *AT*- and *SC*-cut crystals follow an approximately third-order law, as shown in Fig. 5 for the *AT* cut. A slight change in the orientation angle (7 minutes in the example shown in Fig. 5) greatly changes the frequency–temperature characteristic. Curve 1

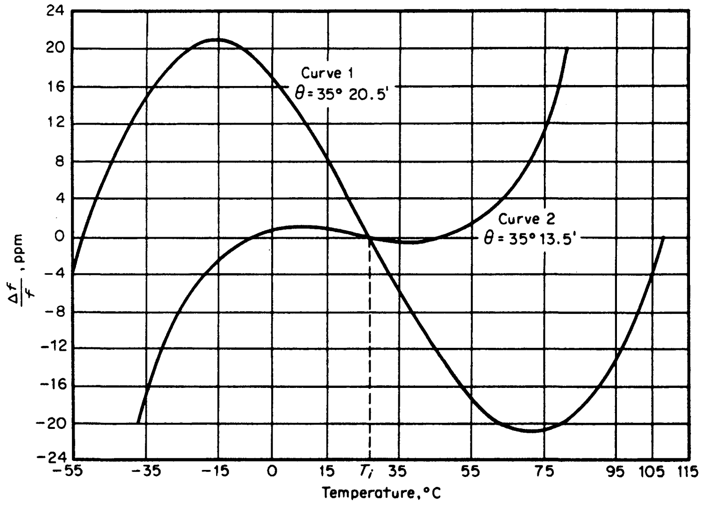

FIG. 5. Typical frequency-vs-temperature characteristics of quartz-crystal units, illustrating how a small change in the angle of cut can significantly affect the characteristic.

is optimal for a wide temperature range (−55 to 105 °C). Curve 2 gives minimum frequency deviation over a narrower range near the inflection temperature T_i. The frequency-vs-temperature characteristic of *SC*-cut crystals is similar to the curves shown in Fig. 5 except that T_i is shifted to about 96 °C. The *SC* cut is a doubly rotated cut; i.e., two angles must be specified and controlled during the manufacturing process. Several of the crystal cuts have a parabolic frequency-vs-temperature characteristic in the range −60 to +100 °C, e.g., the *BT, CT,* and *DT* cuts. The turnover temperatures of these cuts can also be shifted up or down by changing the angles of cut.

To achieve the highest stability, oven-controlled crystal oscillators (OCXO) are used. In OCXOs, the crystal unit and the temperature-sensitive components of the oscillator are placed in a stable oven, the temperature of which is set to the crystal's turnover temperature. OCXOs are bulkier and consume more power than other types. In addition, when an oven-controlled oscillator is turned on, one must wait several minutes for the oscillator to stabilize. During warmup, the thermal stresses in the crystal can produce significant frequency shifts. This thermal-transient effect causes the typical warmup time of an oscillator to be several minutes longer than the time it takes for the oven to stabilize. The thermal-transient effect is absent in *SC*-cut crystals. In oscillators that use *SC*-cut crystals, the warmup time can therefore be much shorter.

In temperature-compensated crystal oscillators (TCXOs) the crystal's frequency-vs-temperature behavior is compensated by varying a load reactance. For example, the output signal from a temperature-sensing network can be used to generate the correction voltage applied to a varactor. Digital techniques are capable of providing better than 1×10^{-7} frequency stability from −55 to +85 °C. TCXOs are smaller and consume less power than OCXOs and require no lengthy warmup times. A major limitation on the stabilities achievable with TCXOs is the thermal hysteresis exhibited by crystal units (see below).

2.1.4.2 Aging and Drift. Aging is the systematic variation of frequency with time due to internal changes in the oscillator, i.e., while all environmental parameters are held constant. Drift is the systematic change in frequency with time of an oscillator, i.e., drift is due to aging plus changes in the environment, power supply, etc. Aging can be due to several factors (Vig and Meeker, 1991). The main causes are mass transfer to or from the resonator surfaces (due to adsorption and desorption of contamination) and stress relief within the mounting structure or at the interface between the quartz and the electrodes. The observed aging is the sum of the aging produced by the various mechanisms and may be positive or negative.

The aging rate of a crystal unit is highest when it is new. As time elapses, stabilization occurs and the aging rate decreases. The aging observed at constant temperature often follows an approximately logarithmic dependence on time. When the temperature of a crystal is changed, however, a new aging cycle starts.

High-quality general-purpose crystals may typically age 1 ppm per year. If a lower aging rate is desired, overtone crystals in clean glass, metal, or ceramic enclosures should be used. Advanced surface-cleaning, packaging, and ultrahigh-vacuum fabrication techniques have resulted in units which age less than 1×10^{-10} per day after a few days of stabilization, or 1×10^{-8} per year.

2.1.4.3 Short-Term Stability. Crystal oscillators' frequency and phase also exhibit random fluctuations, i.e., short-term instabilities (Gerber and Ballato, 1985; Stein and Vig, 1992). Short-term instability, or "noise," in the time domain is usually expressed as the two-sample deviation of the fractional frequency fluctuations for a specified averaging time τ, and is denoted by $\sigma_y(\tau)$. The averaging times τ over which $\sigma_y(\tau)$ is specified generally range from 10^{-3} to 10^3 s. For a good oscillator, $\sigma_y(\tau)$ may range from 1×10^{-9} for $t = 10^{-3}$ s to 1×10^{-12} for $t = 10^3$ s. For $t > 10^4$ s, the stability is usually referred to as long-term stability or aging. For $t < 1$ s, the short-term instabilities are generally attributed to noise in the oscillator circuitry, although the crystal itself can also be a significant contributor.

When measured in the frequency domain, short-term stability is denoted by $\mathcal{L}(f)$, $S_\phi(f)$, or $S_y(f)$—phase noise, the spectral density of phase fluctuations, and that of frequency fluctuations, respectively, at a frequency separation f from the carrier frequency ν, and y is the normalized frequency $\Delta\nu/\nu$. The three quantities are related by $f^2 S_\phi(f) = S_y(f) = 2f^2\mathcal{L}(f)$. For a low-noise 5-MHz oscillator, $\mathcal{L}(f)$ may be -115 dBc at 1 Hz from the carrier, and -160 dBc at 1 kHz from the carrier, where dBc denotes decibels below the carrier.

2.1.4.4 Thermal Hysteresis and Retrace. When the temperature of a crystal unit is changed and then returned to its original value, the frequency will generally not return to its original value. This phenomenon, called thermal hysteresis for TCXOs and retrace for OCXOs, can be caused by the same mechanisms as aging, i.e., mass transfer due to contamination and stress relief (although the mechanisms are not fully understood) (Kusters and Vig, 1991).

For a given oscillator, the magnitude of the effect depends on the magnitude and direction of the temperature excursion, on the thermal history of the unit, and on the design and construction of both the crystal unit and the oscillator circuit. The effect tends to be smaller in OXCOs, where the operating temperature of the crystal is always approached from below, than in TCXOs, where the operating temperature can be approached from either direction. Thermal hysteresis can be minimized through the use of clean, ultrahigh-vacuum fabrication techniques (which minimize the mass-transfer contribution) and through the use of properly mounted *SC*-cut crystals (which minimize the stress-relief contributions). The magnitude of the effect typically ranges from several ppm in general-purpose crystals to less than 1×10^{-9} in the highest-stability crystals operated in OXCOs.

2.1.4.5 Drive Level. The amplitude of vibration is proportional to the current through a crystal. Because of the nonlinearities of quartz, the frequency of a crystal and the incidence of interfering modes are functions of the drive level. Therefore, the drive level must be stated when a crystal's parameters are specified. The oscillator designer's ability to improve signal-to-noise ratios by using a higher drive level is limited. In *SC*-cut crystals, the drive level effects can be significantly below those observed in *AT*- and *BT*-cut units. When the drive level is increased, the frequency of *AT*- or *SC*-cut crystals increases.

2.1.4.6 Acceleration, Vibration, and Shock. The frequency of a crystal unit is affected by stresses. Even the acceleration due to gravity produces measurable effects. When an oscillator using an *AT*-cut crystal is turned upside down along the worst direction, the frequency typically shifts about 4×10^{-9} because the acceleration sensitivity of the crystal is typically 2×10^{-9} per g. The sensitivity is the same when the crystal is subjected to vibration; i.e., the time-varying acceleration due to the vibration modulates the frequency at the vibration frequency with an amplitude of 2×10^{-9} per g. In the frequency domain, the vibration sensitivity manifests itself as vibration-induced spectral lines that appear at plus and minus the vibration frequency away from the carrier frequency (Vig *et al.*, 1992). The acceleration sensitivity of *SC*-cut crystals can be made to be less than that of comparably fabricated *AT*- or *BT*-cut crystals.

Shock places a sudden stress on the crystal. During shock, the crystal's frequency changes due to the crystal's acceleration sensitivity. If the elastic limits in the crystal's support structure or in its electrodes are exceeded during shock, the shock can produce a permanent frequency change. Crystal units made with chemically polished crystal plates can withstand shocks in excess of 30 000*g*. Such crystals have been successfully fired from howitzers.

2.1.4.7 Radiation. Ionizing radiation affects the frequency and resistance of a crystal unit. The degree to which high levels of ionizing radiation affect the frequency and resistance of a crystal unit depends primarily on the quality of quartz used to fabricate the device and on the dose (Gerber and Ballato, 1985; Stein and Vig, 1992). The frequency change per radian is higher at lower doses. When crystals made of natural quartz are subjected to ionizing radiation, their frequencies are changed permanently by approximately a few parts in 10^{11} per rad at a 1 Mrad dose. The resistance can increase to the point where oscillation stops because of insufficient gain in the oscillator. To minimize the effects of ionizing radiation, high-purity cultured quartz should be used. Pulse irradiation produces a transient frequency shift due to the thermal-transient effect. This effect can be minimized by using *SC*-cut crystals. Energetic neutrons change a crystal's frequency by about a few parts in 10^{21} per neutron per cm^2.

2.1.5 Filters Quartz-crystal units are used as selective components in crystal filters (Gerber and Ballato, 1985). With the constraint imposed by the equivalent circuit of Fig. 3, filter design techniques can provide bandpass or bandstop filters with a range of characteristics (see PIEZOELECTRIC FILTERS AND RESONATORS). Crystal filters exhibit low insertion loss, high selectivity, and excellent temperature stability.

Filter crystals are designed to have only one strong resonance in the region of operation, with all other responses (unwanted modes) attenuated as much as possible. The application of energy-trapping theory (Gerber and Ballato, 1985) can provide such a response. If electrode size and thickness are selected in accordance with that theory, the energy of the main response is trapped between the electrodes, whereas the unwanted modes are untrapped and propagate toward the edge of the crystal resonator, where their energy is dissipated. It is possible to manufacture *AT*-cut filter crystals with greater than 40-dB attenuation of the unwanted modes relative to the main response.

Filters made of piezoelectric ceramics, usually of PZT compositions, cover the range of about 50 Hz to above 10 MHz. As is the case for quartz filters, these devices range from simple resonators to multielectrode resonators to coupled sets of devices. PZTs, having a high coupling coefficient and moderate *Q*, allow the use of these devices in medium- to high-bandwidth applications. Ceramic filters have poorer frequency vs temperature and long-term stability than quartz crystal filters.

SAW bandpass filters are used in applications such as television-receiver intermediate-frequency circuits. Such filters can provide stop-band rejection of better than 60 dB and in-band response flat to 0.1 dB. SAW devices can also perform signal processing functions, such as pulse compression and side-lobe suppression that can be performed by dispersive filters and reflective-array-compressor (RAC) filters, and a chirp transform can provide the Fourier transform of an input signal.

2.2 Sensors and Transducers

Whereas in frequency control and timing applications of quartz-crystal devices the components are designed to be as insensitive to the environment as possible, quartz crystals and other piezoelectric materials can also be designed to be highly sensitive to environmental parameters such as temperature, mass changes, pressure, force, and acceleration (EerNisse *et al.*, 1988). Quartz-crystal transducers can exhibit unsurpassed resolution and dynamic range. For example, one commercial "quartz pressure gage" exhibits a 1-ppm resolution, i.e., 60 Pa at 76 MPa (0.01 lb/in.^{2} at 11,000 lb/in.^{2}), and a 0.025% full-scale accuracy. Quartz thermometers can provide millidegrees of absolute accuracy over wide temperature ranges. Quartz sorption detectors can detect a change in mass of 10^{-12} g. Quartz accelerometer/force sensors are capable of resolving 10^{-7} to 10^{-8} of full scale.

Piezoelectric ceramics are also used for sensors (Newnham and Ruschau, 1991). In some applications, ceramics (e.g., various compositions of lead zirconate titanate, PZT) are preferred over quartz and other single-crystal piezoelectrics because of their higher piezoelectric coupling, which can result in a larger signal when a force is applied. For example, ceramics are used in some accelerometers wherein the acceleration of a mass applies a compressional or shear force to the piezoelectric ceramic. The signal-conditioning circuitry, which must amplify what is usually a small signal from a high-impedance source, is often part of the device.

2.3 Actuators, Motors, Smart Materials and Devices

Piezoelectric actuators, devices that are capable of moving something electromechanically via electric-field–induced strain, can be classified into two categories, based on the type of driving field applied to the device:

1. rigid displacement devices, in which a dc field is applied, and
2. resonating strain devices, in which an ac field excites the device (Newnham and Ruschau, 1991).

Piezoelectric materials have conventionally been used in devices of the second type, e.g., as vibrators in oscillators, filters, buzzers, speakers, and ultrasonic generators. For rigid displacement devices, large electric fields (~1 kV/mm) have been needed in order to achieve sufficiently large stresses and strains. Therefore, the materials have had to be strong mechanically and electrically (i.e., with high insulation strength). Another problem has been the fact that when an electric field is applied to a piezoelectric ceramic such as PZT, at low fields the strain is proportional to the field, but at high (e.g., >100 V/mm) fields, the relationship becomes nonlinear, and significant hysteresis is exhibited. The hysteresis has limited the applications in areas that require a linear, nonhysteretic response.

In recent years, the need for precision displacement devices has increased in fields such as optics and semiconductor processing. Submicron-level precision is required in several of these applications, which has renewed interest in the use of improved piezoelectric (and electrostrictive) materials for fine positioning control.

Rigid displacement devices have been used for precision optical control, e.g., in deformable mirrors, and for mechanical control, e.g., in dot-matrix printer heads, piezoelectric relays, and cutting-error–correction devices. Deformable mirrors allow changing the focal length and adjustments for curvature imperfections, and can even compensate for telescope image distortions resulting from fluctuating airflow. Other applications in optics include variable diffraction gratings, microangle optical element adjustments, and precision microscope stages. Cutting-error–correction devices have been used in, for example, cutting tools used for the forming of large radius-of-curvature aspheric surfaces. A piezoelectric actuator synchronized to the rotation of the cutting tool can reduce the imperfections of the cutting tool 20-fold, to an accuracy of about 0.1 mm.

In dot-matrix printers, a printing ribbon is impacted by a multiwire array to form, e.g., a 9 × 9 dot matrix. In some such printers, the actuator consists of about 100 layers of a thin piezoelectric ceramic sheet. The displacement of the actuator is magnified via levers so that the total displacement is about 300 μm. The arrangement provides higher printing speeds than printers using conventional electromagnetic actuators, with lower energy consumption and less noise (as a result of easier shielding).

Resonating displacement devices include ultrasonic knives and scapels, piezoelectric fans, ultrasonic humidifiers, piezoelectric pumps, ultrasonic wire bonders, components in active vibration-damping systems, and ultrasonic motors. Ultrasonic wire bonders are used, for example, for bonding gold wire leads to semiconductor devices. The ultrasonic vibration of the bonding tip disperses the surface contamination at the bond interface and, thereby, improves the strength and reliability of the bond.

Ultrasonic motors of various rotary and linear types have been constructed. In a rotary device, the rotation of a rotor is caused by piezoelectric vibration of a stator instead of the conventional electromagnetic interaction. The driving force is transmitted directly from the stator to the rotor by friction. One linear stepping motor ("inchworm"), for example, consists of a pair of piezoelectric clutches, each at the end of a piezoelectric ceramic tube. The clutches alternately grab, move, and release a metal rod in a manner that moves the rod linearly, in steps. Such a motor can provide submicron resolution, and large (although slow) displacements. Both linear and rotational motors have been produced by means of surface-wave vibrations. A traveling elastic wave is induced by a piezoelectric driver. A slider in contact with the rippled surface of a body moves the body in the direction of the wave. Longitudinal, torsional, and bending multimode vibrators have also been employed in piezoelectric motors.

The name "smart material" has been used to describe some actuators (Newnham and Ruschau, 1992). Smart materials have the ability to be both sensors and actuators. Passively smart devices can sense a change in the environment and respond in a useful manner, automatically, without any assistance. Actively smart devices contain a feedback loop that allows them to recognize a change and initiate an appropriate response through an actuator. A passively smart material differs from an actively smart material in that no external fields, forces, or feedback systems are used to enhance its behavior (Newnham, 1992). "Very smart" materials have the additional ability to learn and alter their property coefficients in response to a changing environment. An example of an actively smart device is a videotape head positioner in which a PZT bilaminate bender (i.e., bimorph) with a segmented electrode pattern is used. The bender vibrates at 450 Hz. Voltage across the sensing electrodes is processed and applied to the positioning electrodes so as to follow the tape track path and keep the head perpendicular to the track. Modern "high-tech" automobiles use several piezoelectric sensor–actuator combinations. For example, in one system, a piezoelectric raindrop sensor senses the amount of rain falling and adjusts the windshield wipers to the optimum speed. Piezoelectric ceramics are also used in active noise cancellation and in smart shock absorbers (a bump in the road produces a voltage, which produces a displacement that changes the damping produced by the hydraulic system).

2.4 Optical and Acousto-optical Devices

Applying a strain to a solid can produce a deformation of the lattice that can change the solid's optical properties (Kino, 1987). Sound waves and light can thereby interact in a solid or fluid. The sound wave creates a periodic variation of the properties of the medium, which can result in scattering and diffraction. The coupling between light and sound can result from the photoelastic (or piezo-optic) effect, or the converse electrostrictive effect. In the photoelastic effect, strains due to the sound waves change the atomic lattice spacings, resulting in localized changes in the dielectric constant. In the converse electrostrictive effect, the presence of the electric field due to light causes a strain that is proportional to the square of the electric field.

When a light beam is deflected by sound waves, the amplitude of the deflected beam depends on the amplitude of the sound wave, and the angle of deflection depends on the frequency of the sound wave. The light pattern formed by the deflected beam can thereby provide the Fourier transform of the modulating acoustic beam. Using the acousto-optic interactions with two acoustic beams adds further possibilities that lead to a wide variety of signal-processing functions, such as the real-time wideband convolution and correlation of electrical signals.

2.5 Underwater Sound Systems, Ultrasonic Cleaners, Medical Ultrasonics, and Nondestructive Evaluation

Water is an excellent medium in which to transmit (compressional) sound waves, and piezoelectric ceramic transducers, e.g., PZTs and barium titanates, are excellent materials for generating and detecting such waves. The generator is usually called a source or projector, the detector is called a hydrophone, and the term sonar (derived from SOund Navigation And Ranging) is a major topic in the field of underwater acoustics. Arrays of transducers can be used to achieve directionality; the arrays can be steered mechanically or electrically, similar to the techniques used in radar. The functions performed by sonar are similar to those performed by microwave systems above water, i.e., detection, tracking, classification, communication, navigation, etc.

In ultrasonic cleaners, the vibration of an ultrasonic transducer produces compression (high pressure) during the half-cycle when the transducer moves toward the liquid, and rarefaction (low pressure) when it moves away from the liquid. If the amplitude of the ultrasonic vibration is high enough, then during the low-pressure part of the cycle, the pressure on the liquid will be below the liquid's vapor pressure at the temperature of operation, and a large number of minute bubbles are formed. One-half cycle later, the high pressure causes the vapor to recondense to liquid, thus forming tiny partial vacuum bubbles that implode with great force. Such a formation and implosion of vapor bubbles is called cavitation. The cavitation provides the ultrasonic scrubbing of surfaces. Use of a cleaning solution in the cleaners can enhance the cleaning action.

Medical ultrasonic (i.e., ultrasound) systems apply the principles of sonar to the visualization of internal organs. Lithotripsy is a treatment method wherein a patient is immersed in a liquid and very intense and highly directional ultrasonic waves are directed, through the liquid, at internal organs, e.g., to break apart kidney stones.

The reflection, scattering, and diffraction of ultrasonic waves at the boundaries of different materials (e.g., different tissue densities in medical ultrasonics) is also used in a nondestructive evaluation of materials and devices.

2.6 Other Piezoelectric Devices

That a dc potential can produce a displacement and an alternating electrical potential can produce vibration in a piezoelectric material has been exploited in a wide variety of devices. For example, piezoelectric ceramics (often in the form of bimorphs) and piezoelectric polymers are used in microphones, telephony, intruder alarm systems, audio tone generators in signaling devices, range finders, sensors for door openers, remote-control devices, loudspeakers, phonograph pickups, atomization devices, pulverization devices, welders (for ultrasonic welding of plastics and metals), ultrasonic soldering, ultrasonic drilling and machining of brittle materials, valves, pumps, optical choppers and scanners, grippers, cooling fans, and actuators in pinball games, and as transducers that produce sonochemical reactions.

That a force on a high-coupling piezoelectric ceramic material can produce a high dc potential has been exploited in, for example, high-voltage generators used for ignition purposes—for lighting a fire in gas appliances, in cigarette lighters, and in charcoal grill lighters; fuses for explosives; for triggering flash bulbs, in push buttons, and in keyboards.

That the velocity of ultrasonic waves is far smaller than the velocity of electromagnetic waves in solids has been exploited in a variety of subminiature BAW and SAW delay lines for replacement of electromagnetic delay lines.

Highly sensitive uncooled infrared detector arrays are based on ferroelectric materials (Dereniak and Sampson, 1992; Andresen and Shepherd, 1993). In one detector array, for example, bolometer elements are made of ferroelectric barium strontium titanate ($BaSrTiO_3$). Near the ferroelectric phase transition, i.e., near the Curie point, the pyroelectric coefficient is large, and the dielectric constant varies rapidly. In the dielectric bolometer, as the temperature of the elements changes on account of incoming infrared radiation, the capacitance changes are sensed. In the pyroelectric mode, the temperature-change–induced polarizations are sensed. A new type of uncooled infrared detector array of potentially superior performance consists of an array of quartz microresonators. This recent invention utilizes the steep frequency-vs-temperature characteristics of resonators made of certain cuts of quartz, the low noise characteristics of quartz crystal oscillators, and the small thermal mass and high thermal isolation capability of microresonators (Vig, 1995).

GLOSSARY

Aging: The systematic variation of frequency with time due to internal changes in an oscillator, i.e., while all environmental parameters are held constant.

Crystal Unit: An electrical component consisting of a quartz-crystal plate of appropriate orientation with respect to the crystallographic axes of quartz, electrodes on or near the quartz plate, a mounting structure, and an enclosure. The component is sometimes also referred to as a "crystal" and as a "resonator."

Drift: The systematic change in frequency with time of an oscillator, i.e., drift is due to aging plus changes in the environment, power supply, etc.

Electromechanical Coupling Factor: A number describing conversion efficiency of energy between electric and elastic forms in a piezoelectric material.

Ferroelectric: The description given to those polar crystals capable of having the spontaneous polarization along the polar axis reversed or reoriented, usually in response to an applied electric field.

Frequency Deviation $y(t)$: The instantaneous normalized frequency deviation at time t from a mean frequency.

Oven-Controlled Crystal Oscillator (OCXO): A type of oscillator in which the crystal unit and the temperature-sensitive components of the oscillator are placed in a stable oven, the temperature of which is set to the crystal's turnover temperature.

Phase Noise $\mathscr{L}(f)$: The measure of short-term stability in the frequency domain (see IEEE Standard 1139-1988 for definition and discussion).

Piezoelectricity: The phenomenon in which an electric polarization couples to a mechanical strain. In the direct piezoelectric effect, the electric polarization results from mechanical strain; in the converse effect, a mechanical strain is produced in a material by a polarizing electric field. The polarization is proportional to the strain, and changes direction with it.

Point Group: One of 32 possible geometric symmetry arrangements of atoms in a crystalline material determining which properties (e.g., piezoelectricity) exist, and their relations with the crystal axes.

Pyroelectricity: The phenomenon in which a temperature change produces a charge separation, and consequently an electric field. When the crystals are subjected to a uniform change of temperature, electric charges accumulate on the ends of the polar axes.

Quartz Crystal: Single-crystal silicon dioxide, SiO_2. Alpha quartz is the crystalline form of quartz at temperatures below 573 °C. Cultured quartz is quartz produced in a factory, as opposed to natural quartz, which is mined.

Stability: Sensitivity to a frequency change as a result of changes of parameters such as time, temperature, shock, and the like. Stability is usually expressed as a dimensionless, normalized frequency change from a nominal or mean frequency.

Temperature-Compensated Crystal Oscillator (TCXO): A type of oscillator in which the crystal unit's frequency-vs-temperature behavior is compensated by varying a load reactance in the oscillator circuit.

Two-Sample Deviation $\sigma_y(\tau)$: The square root of the two-sample variance. Also called the Allan deviation, or the square root of the Allan variance.

Two-Sample Variance $\sigma_y^2(\tau)$: The time average over the sum of the squares of the differences between successive readings of the frequency deviation sampled over a sampling time τ, under the assumption that there is no dead time between the frequency deviation samples. Also called the Allan variance.

Ultrasonic: A modifier to indicate a device or system employing, or pertaining to, frequencies above the audio range.

Works Cited

Andresen, B. F., Shepherd, F. D. (Eds.) (1993), *Infrared Technology XIX*, SPIE Proceedings No. 2020, Bellingham, WA: SPIE.

Cady, W. G. (1946), *Piezoelectricity*, New York: McGraw-Hill; reprint (1964), New York: Dover.

Dereniak, E. L., Sampson, R. E. (Eds.) (1992), *Infrared Detectors and Focal Plane Arrays*, SPIE Proceedings No. 1685, Bellingham, WA: SPIE.

EerNisse, E. P., Ward, R. W., Wiggins, R. B. (1988), "Survey of Quartz Bulk Resonator Sensor Technologies," *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **35**, 323–330.

Gerber, E. A., Ballato, A. (Eds.) (1985), *Precision Frequency Control*, New York: Academic.

Gualtieri, J. G., Kosinski, J. A., Ballato, A. (1993), "Piezoelectric Materials for Acoustic Wave Applications," *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **41**, 53–59.

Hellwege, K.-H., Hellwege, A. M. (Eds.) (1979), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, New Series, Group III, Vol. 11, *Elastic, Piezoelectric, Pyroelectric, Piezooptic, Elektrooptic Constants, and*

Nonlinear Dielectric Susceptibilities of Crystals, Berlin: Springer.

Hellwege, K.-H., Hellwege, A. M. (Eds.) (1981), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, New Series, Group III, Vol. 16, *Ferroelectrics and Related Substances*, Part a, *Oxides*, Berlin: Springer.

Hellwege, K.-H., Hellwege, A. M. (Eds.) (1982), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, New Series, Group III, Vol. 16, *Ferroelectrics and Related Substances*, Part b, *Non-Oxides*, Berlin: Springer.

Hellwig, H. (1993), "Time and Frequency Applications," *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **40**, 538–543.

IEEE (1978), *IEEE Standard on Piezoelectricity*, IEEE Standard No. 176-1978, New York: IEEE.

Kino, G. S. (1987), *Acoustic Waves: Devices, Imaging, and Analog Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall.

Kusters, J. A., Vig, J. R. (1991), "Hysteresis in Quartz Resonators," *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **38**, 281–290.

Madelung, O. (Ed.) (1982), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, New Series, Group III, Vol. 17, *Semiconductors*, Part a, *Physics of Group IV Elements and III-V Compounds*, Berlin: Springer.

Mason, W. P. (1966), *Crystal Physics of Interaction Processes*, New York: Academic.

Newnham, R. E. (1992), "Piezoelectric Sensors and Actuators: Smart Materials," in: *Proceedings of the 1992 IEEE Frequency Control Symposium*, New York: IEEE, pp. 513–524.

Newnham, R. E., Ruschau, G. R. (1991), "Smart Electroceramics," *J. Am. Ceram. Soc.* **74**, 463–480.

Nye, J. F. (1957), *Physical Properties of Crystals*, Oxford: Clarendon; (1985), Oxford: Oxford Univ. Press.

Ruppel, C. C. W., *et al.* (1993), "SAW Devices for Consumer Communication Applications," *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **40**, 438–451.

Stein, S. R., Vig, J. R. (1992), "Communications Frequency Standards," in: F. E. Froehlich, A. Kent (Eds.), *The Froehlich-Kent Encyclopedia of Telecommunications*, Vol. 3, New York: Marcel Dekker, pp. 445–500.

Vig, J. R. (1993), "Military Applications of High Accuracy Frequency Standards and Clocks," *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **40**, 522–527.

Vig, J. R., Walls, F. L. (1994), "Fundamental Limits on the Frequency Instabilities of Quartz Crystal Oscillators," in: *Proceedings of the 1994 IEEE International Frequency Control Symposium*, New York: IEEE, pp. 506–523.

Vig, J. R., Meeker, T. R. (1991), "The Aging of Bulk Acoustic Wave Resonators, Filters, and Oscillators," in: *Proceedings of the 45th Annual Symposium on Frequency Control*, New York: IEEE, pp. 77–101.

Vig, J. R., *et al.* (1992), "Acceleration, Vibration, and Shock Effects—Guidelines for the Measurement of Environmental Sensitivities of Precision Oscillators, IEEE Standards Project P1193," in: *Proceedings of the 1992 IEEE Frequency Control Symposium*, New York: IEEE, pp. 763–791.

Vig, J. R., *et al.* (1995), "Microresonator Sensor Arrays," in: *Proceedings of the 1995 IEEE International Frequency Control Symposium*, New York: IEEE, to be published.

Further Reading

Gerber, E. A., Ballato, A. (Eds.) (1985), *Precision Frequency Control*, New York: Academic.

Jespersen, J., Hanson, D. W. (Eds.) (1991), Special Issue on Time and Frequency, *Proc. IEEE* **79** (7).

Kino, G. S. (1987), *Acoustic Waves: Devices, Imaging, and Analog Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall.

Continuing Series

The *Proceedings of the International IEEE Ultrasonics Symposium* is published annually by the IEEE. These volumes contain research and review papers on a wide variety of piezoelectric ultrasonic devices.

The *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* reports on topics dealing with piezoelectric materials, nondestructive evaluation, biological and medical ultrasonics, imaging, ferroelectricity, sensors, optical interactions, underwater sound, measurement and control applications, ultrasonic motors, and frequency control.

The *Proceedings of the IEEE International Frequency Control Symposium*, and its predecessor, the *Proceedings of the Annual Symposium on Frequency Control*, are the primary fora for publishing research results and review papers on piezoelectric materials for frequency control, and on resonators and oscillators for frequency control, selection, and timing applications. These have been published since the Tenth Symposium in 1956. The latest volumes are available from IEEE, 445 Hoes Lane, Piscataway, NJ 08854. The earlier volumes are available from the National Technical Information Service, Sills Building, 5285 Port Royal Road, Springfield, VA 22161.

PIEZOELECTRIC RESONATORS AND APPLICATIONS

THRYGVE R. MEEKER *Allentown, Pennsylvania, U.S.A.*

Introduction 148

1. **A Quasistatic Linear Model for Piezoelectric Resonator Modes** 148
1.1 Basic Variables and Equations 148
1.2 The Elasto-Piezo-Dielectric Matrix 150
1.3 Coordinate-System Transformations 150
1.4 Comments on Piezoelectric Resonator Theories 150

2. **Bulk Acoustic-Wave Resonator Modes** 151
2.1 Shear and Dilatation Thickness Modes in Plates and Disks 151
2.1.1 Infinite Length, Width, and Radius: Completely Electroded 151
2.1.2 Finite Length, Width, and Radius: Completely Electroded 152
2.1.3 Infinite Length, Width, and Radius with Finite Electrode Dimensions: Partially Electroded 152
2.1.4 Thickness-Mode Resonators Using Contoured Plates 153
2.2 Contour Modes in Bars, Cylinders, Plates, and Disks 154
2.2.1 Extension in Bars and Cylinders 154
2.2.2 Extension in Plates and Disks 155
2.2.3 Flexure in Bars and Cylinders 155
2.2.4 Flexure in Plates and Disks 156
2.2.5 Face Shear in Plates 156
2.3 Torsional Modes 157
2.3.1 Torsion in Bars and Cylinders 157
2.3.2 Torsion in Disks 158

3. **Bulk Acoustic-Wave Resonators** 158
3.1 Quartz-Crystal Resonators 158
3.2 Resonators Using Other Materials (Ceramic, Crystals, Films) 158

4. **Composite Electromechanical Resonators** 159

5. **Surface Acoustic-Wave Resonator Modes on Plates** 159
5.1 Rayleigh Waves 160
5.2 Gulyaev–Bleustein Waves 161
5.3 Devices 162

6. **Equivalent Electrical Circuits** 162

7. **Materials** 163

8. **Applications of Piezoelectric Resonators** 163
8.1 Piezoelectric Filters 163
8.1.1 Discrete Crystal Filters 164
8.1.2 Monolithic Crystal Filters 164
8.1.3 Electromechanical Filters 164
8.1.4 Surface Acoustic-Wave Filters 165
8.2 Piezoelectric Crystal or Electromechanical Oscillators 165
8.3 Piezoelectric Sensors 165

Appendix A: Thickness-Excited Thickness Modes in Plates 165
A.1 Infinite Length and Width 165
A.2 Electrical Impedance per Unit Area 166
A.3 Resonators with Loaded Surfaces 167
A.3.1 Frequency Shifts due to Electrical Loading 167
A.3.2 Frequency Shifts due to Mass Loading 167

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

A.3.3 Frequency Shifts due to Nonuniform Mechanical Motion 167
Glossary 167
Works Cited 167
Further Reading 169

INTRODUCTION

Early theoretical treatments and experimental observations of the electrical properties of piezoelectric resonators have been reviewed. Also reviewed were early publications on potential applications of piezoelectric resonators in oscillators, filters, and circuit couplers (Cady, 1964, Vol. 1, Chap. 1, pp. 1–8).

An important purpose of the present article is to review some developments in the understanding of the theory and application of piezoelectric resonators. Some possible mode types in piezoelectric resonators and associated mode characteristics are described in this article. Although some applications of piezoelectric resonators can be based on a quasistatic linear theory, many important applications need to use a nonlinear theory (Gagnepain and Besson, 1975; Tiersten and Sinha, 1978). In some cases, an approximate linear theory can be used in place of the more accurate nonlinear theory. The nonlinear behavior of the piezoelectric resonator often produces undesired resonator performance, which needs to be understood (at least approximately) in order to optimize the use of resonators in particular applications. In Appendix A, a quasistatic theory of thickness-excited thickness-mode piezoelectric resonators is described as an example. Three general kinds of applications of piezoelectric resonators described in this article are frequency control (i.e., filters), frequency generation (i.e., oscillators), and measurement of various environmental conditions (i.e., sensors). Filters, oscillators, and sensors using crystal and ceramic resonators all have considerable commercial significance.

Most sections of this article are based on published surveys and review articles. References to specific topics are included if suitable reviews are not available. Except in special cases, references found in recent papers or reviews are not repeated in this article. Throughout this article, differentiation with respect to a system variable is indicated by the comma notation, i.e., $u_{1,3} = du_1/dx_3$ or $u_{2,t} = du_2/dt$. The summation convention (Jeffreys, 1931) is used for repeated indices. Quantities are called tensors if they change like tensors when the coordinate system is rotated (Nye, 1957, 1960).

1. A QUASISTATIC LINEAR MODEL FOR PIEZOELECTRIC RESONATOR MODES

Useful background material for this section can be found in the literature (Tiersten, 1969, Part I, pp. 11–61; Part II, pp. 65–93).

1.1 Basic Variables and Equations

The stress differential equations that govern the mechanical motion of elastic or piezoelectric materials are

$$T_{ij,i} + f_j = \rho u_{j,tt}, \tag{1}$$

where the T_{ij} are the nine components of the mechanical stress tensor, the f_j are the three components of the body force vector, ρ is the material density, and the u_j are the three components of the mechanical displacement vector. This equation of motion can be derived from the conservation of linear momentum for the system. It is also Newton's second law that describes the mechanical motion of elastic or piezoelectric materials.

The electrical differential equations that govern the polarization of electric charge-free dielectrics or piezoelectric materials are

$$D_{i,i} = 0, \tag{2}$$

where the D_i are the three components of the electrical displacement vector. Equation (2) is one of Maxwell's equations that describes the electrostatic behavior of charge-free dielectrics.

For linear piezoelectric materials, the components of mechanical stress and electrical displacement are proportional to the components of mechanical strain and electric field:

$$T_{kl} = c^E_{klmn}S_{mn} - e_{mkl}E_m \tag{3}$$

and

$$D_j = e_{jkl}S_{kl} - \epsilon^S_{jk}E_k, \tag{4}$$

where $S_{ij} = 0.5(u_{i,j} + u_{j,i})$ are the nine components of the mechanical strain tensor, $E_k = -\phi_{,k}$ are the three components of the electric field vector, the c^E_{klmn} are material constants called the 81 components of the stiffness tensor at constant E, the e_{jkl} are material constants called the 27 components of the piezoelectric constant tensor, and the ϵ_{jk} are material constants called the nine components of the dielectric permittivity tensor at constant S.

Equations (3) and (4) are called constitutive equations. The definitions of the quantities in the constitutive relations allow expressing the components of the mechanical stress and electric displacement in terms of the electric potential and the components of the mechanical displacement. Other forms of the constitutive equations may be useful in describing particular systems (Berlincourt *et al.*, 1964; Meeker, 1985, pp. 60–61).

Under some conditions, it is useful to allow the material constants in a linear theory to depend on environmental conditions such as temperature, pressure, acceleration, etc. These environmental conditions are more accurately treated as nonlinear effects. For some external conditions, such as inhomogeneous heating, the linear constitutive equations are not accurate, and approximate or nonlinear theories are needed (Tiersten and Sinha, 1978).

Boundary and initial conditions that insure a unique solution of Eqs. (1) and (2) have been derived. A specific example states boundary specification of three mechanical stress components and the electric potential on all surfaces for all time and initial specification of the components of mechanical velocity, mechanical displacement, and electric field at all points in the body (Tiersten, 1969, Chap. 5, pp. 33–39). Practical resonator designs are usually based on zero values for the mechanical stress components and a specified electric potential as the boundary conditions. Components of mechanical velocity and the electrical current at the surfaces can be calculated directly from the final solution.

After a desired direction in the crystal lattice has been selected, four types of mechanical stress and two types of electrical boundary conditions lead to nearly all practical resonator designs. Many other types of boundary conditions, which might lead to useful resonator designs for particular applications, will not be considered in this article.

For thickness-mode transducers (as contrasted to resonators), useful boundary conditions are the specification of three mechanical stress components, three mechanical velocity components, and an electric potential and electric current at each of the two surfaces of the transducer. Details for a single-mode thickness transducer with infinite length and width have been reported (Meeker, 1985, Chap. 2.1, pp. 47–110, and Chap. 2.2, pp. 110–119).

Expressing the differential equations, boundary conditions, and initial conditions in other coordinate systems is often useful in the development of theories for resonators with special symmetries, such as circular plates or cylindrical rods (Meitzler *et al.*, 1973).

For the quasistatic case, the free electric charge within the body is zero, and resonator properties do not depend on magnetic effects. The material density, electric charge, and electric potential are scalars or zero-order tensors; the mechanical displacement, the acceleration, the electric field, and the electric displacement are vectors or first-order tensors; and the strain and stress are second-order tensors. Physical quantities are called tensors because of the way they change when the coordinate axes are rotated (Nye, 1957, 1960).

The coupled second-order differential equations are solved, subject to the boundary conditions, for the dependence of the distribution of mechanical motion on time and position within the resonator body (Tiersten, 1969, Chap. 8, pp. 65–80; Chap. 9, pp. 81–93). The general solution process would yield mechanical displacement distributions for all possible modes of motion of the resonator. Except for some idealized resonators, the solutions are not easily expressed as analytical functions and have to be determined numerically for particular boundary conditions on the surfaces of the resonator body (Holland and EerNisse, 1969).

Other choices of dependent and independent field quantities can be useful in developing theories for particular resonator geometries (Berlincourt *et al.*, 1964; Meeker, 1985). The superscripts on ϵ_{ik}^{S} and c_{klmn}^{E} can only signify a particular choice of independent and dependent field quantities, because the constants defined for other choices can be calculated directly from those defined above; i.e., all alternative sets of these constants pertain to the resonator material at all times and at all places and are not related to specific boundary condition choices (Meeker, 1985).

1.2 The Elasto-Piezo-Dielectric Matrix

A compressed notation (Halliburton *et al.*, 1985) allows the tensor constitutive equations in Sec. 1.1 to be rewritten as

$$T_m = C_{mn}^{E} S_n - \epsilon_{km} E_k \tag{5}$$

and

$$D_i = e_{in} S_n + \epsilon_{ik}^{S} E_{ik}, \tag{6}$$

where $m = 1$ to 6, $n = 1$ to 6, $k = 1$ to 3, and $i = 1$ to 3. The 9×9 symmetric matrix of the coefficients of S_n and E_k is called the *elasto-piezo-dielectric matrix.*

1.3 Coordinate-System Transformations

By selecting a proper orientation for a resonator blank in the material coordinate system, the temperature coefficient, acceleration sensitivity, or some other desired property can be optimized. The arbitrary orientation of a plate or bar with respect to the material coordinate system can be described by three rotations, starting with a plate or bar thickness, edge, or length along one of the standard axes. This system for specifying the orientations of rectangular plates (or of plates with other shapes with orientation marks on the surfaces or edges) is described in standards (Hafner, 1985, Chap. 16, pp. 297–304), to provide better communication between theoreticians, designers, suppliers, and customers on plate and bar specifications. Effects of plate or bar orientation in the crystalline lattice and the symmetry of the crystal lattice on the elements of the elasto-piezo-dielectric matrix can be found in the literature (Halliburton *et al.*, 1985).

1.4 Comments on Piezoelectric Resonator Theories

The deceptively simple equations in Sec. 1.1 have never been solved in closed form for an actual physically realizable piezoelectric resonator. Exact solutions for idealized resonators such as infinitely wide and long thickness-mode plate resonators or infinitely narrow and thin length-mode bar resonators have been reported (Cady, 1964, Chap. 5, pp. 84–115; Berlincourt *et al.*, 1964). Approximate solutions for thickness shear-mode plate resonators with finite width and length and for extensional and flexural length-mode bar resonators with finite thickness and width have been reported (Sykes, 1982, pp. 214–219; Meeker, 1985, Chap. 2.1, pp. 47–110, and Chap. 2.2, pp. 110–119). In Secs. 2 and 3, some of the more useful modes of piezoelectric resonators are briefly described. For each resonator, a few modes (usually only one or two) are desired, and the other modes are unwanted. Both desired and unwanted mode activities must be controlled by the design of the resonator. This means that the designer must understand the properties of all the active modes in the resonator of interest in order to optimize the performance and cost of the resonator.

To optimize some important resonator properties, such as the electrical impedance level of unwanted resonator modes, it is necessary to have adequate models for each mode of interest. For particular applications, these models may have to include couplings between various modes. Until the required models are developed, considerable experience is required to design resonators with desired low activities of unwanted modes.

Many piezoelectric resonators made of crystalline quartz have been named with one or two letters to designate the desired mode (Sykes, 1982, pp. 211–214). Some characteristics of several of these quartz-crystal resonators are described in Sec. 3.

Piezoelectric resonators can be classified in terms of mode type (i.e., bulk or surface acoustic) and resonator material (i.e., crystal, polarized ceramic, polarized film, etc.). Each mode type refers to a particular distribution of the motion (direction and magnitude) of the individual volume elements of the piezoelectric resonator body. All volume-element motions have the same frequency and phase (Cady, 1964, Vol. 1, Chap. 5, pp. 86–87; Hildebrand, 1965). In some cases, the actual motion is a complicated combination of several simple types of motion.

Each mode type may also be characterized by a resonance frequency, an antiresonance frequency, an electrical impedance level, and a dissipation factor (Hafner, 1985, Chap. 7, pp. 1–44). These mode characteristics depend on the shape, dimensions, and crystal orientation of the piezoelectric body and on the electrode material, thickness, orientation, and spatial distribution. The electrodes may be inside the body, on the body surfaces, or outside the body in the surrounding space.

Some details of using the linear theory to derive electrical impedances per unit area for thickness mode resonators are given in Appendix A.

In Secs. 2 through 5, many of the modes that can exist in a piezoelectric resonator are discussed briefly.

2. BULK ACOUSTIC-WAVE RESONATOR MODES

2.1 Shear and Dilatation Thickness Modes in Plates and Disks

2.1.1 Infinite Length, Width, and Radius: Completely Electroded Several theories of piezoelectric thickness-excited thickness-mode resonators with infinite width and length, using only exact solutions for the equations in Sec. 1.1, have been reviewed (Meeker, 1985, Chap. 2.1, pp. 47–110, and Chap. 2.2, pp. 110–119). Electrical and mass surface-loading effects are included in some of these theories. By use of the appropriate mechanical displacement components and elasto-piezo-dielectric matrix elements, this theory can be used for both shear and dilatational mode resonators (Meeker, 1985, Chap. 2.1, pp. 47–110, and Chap. 2.2, pp. 110–119). As an example of these theories, some detailed results are given in Appendix A for thickness-mode resonators. Appendix A also includes a brief discussion of the effects of nonuniform spatial distributions of mechanical motion in physically realizable thickness-mode resonators.

Figure 1 shows mechanical displacement components for the fundamental lowest-frequency thickness shear and dilatational

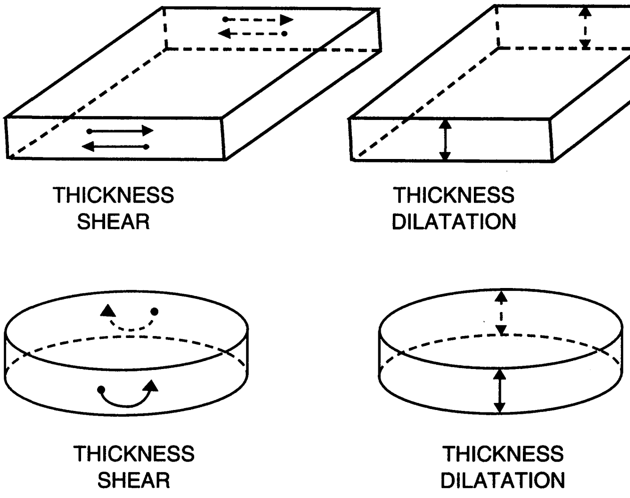

FIG. 1. Mechanical displacement components for thickness modes in plates and disks with infinite length, width, and radius.

modes in plates and disks with infinite lateral dimensions. The lateral boundaries shown in Fig. 1 only serve to draw attention to a localized region of the infinitely extended plate or disk. An infinite number of overtone higher-order thickness modes have more than one phase reversal of mechanical displacement between the major surfaces of the plate or disk. The plate or disk major surfaces are either completely electroded or completely unelectroded. In Fig. 1, the circular thickness shear is a torsion (Sec. 2.3.2).

For well-designed thickness-mode resonators, the modes described by Eq. (7) in Appendix A will be the resonator modes with the lowest values of Z. Many other modes not described by Eq. (7) of Appendix A will have impedances that are higher and are called weaker modes, because in most circuits they will have a lower amplitude. In some cases, coupling between the stronger modes and the weaker modes produces unacceptable resonator properties.

Applications of thickness-mode piezoelectric resonators depend on resonator electrical impedances, as described by equations like Eq. (7) in Appendix A. Some general and specific applications are discussed in Sec. 8.

For some crystals at certain orientations, thickness modes in plates or disks can be excited by electric fields along a direction in the plane of the plate (Meeker, 1985, Chap. 2.1, pp. 47–110, and Chap. 2.2, pp. 110–119; Ballato *et al.*, 1986).

2.1.2 Finite Length, Width, and Radius: Completely Electroded For the discussion in Sec. 2.1.1, the thickness is the only finite dimension of the plate or disk. Each of the three possible modes (one quasidilatational and two quasishears) has an infinite series of thickness modes, each characterized approximately by the number of half-wavelengths of mechanical motion through the plate thickness. The antiresonance frequencies of these modes are evenly spaced (harmonic); the resonance frequencies are not evenly spaced (anharmonic).

Each finite lateral dimension of the completely electroded or completely unelectroded plate or disk leads to an infinite series of anharmonic thickness modes with frequencies that are not evenly spaced and that are associated with the number of half-wavelengths of mechanical motion along the lateral dimension. Thus, three numbers signify the number of half-wavelengths of mechanical motion along the three dimensions of the resonator plate. For the circular disk, the three dimensions are the thickness, radius, and a rotation angle in the plane of the disk. The relationships among the resonance frequencies of these three sets of thickness modes are shown in Fig. 2, in which dispersion relations for plate mode propagation have been converted into frequency equations by a wave-number approximation (Meeker, 1985). There are relations, as shown in Fig. 2, for each of the two finite lateral dimensions of the plate. Figure 3 shows how the plate deforms in shear for three combinations of half-wavelengths in two dimensions (Sykes, 1982, Fig. 6.5, p. 211). A low-order mode not shown in Fig. 3 has two half-wavelengths in the plate thickness direction. Lateral anharmonic modes are found on fundamental and overtone resonance and antiresonance mode branches.

2.1.3 Infinite Length, Width, and Radius, with Finite Electrode Dimensions: Partially Electroded When finite-area

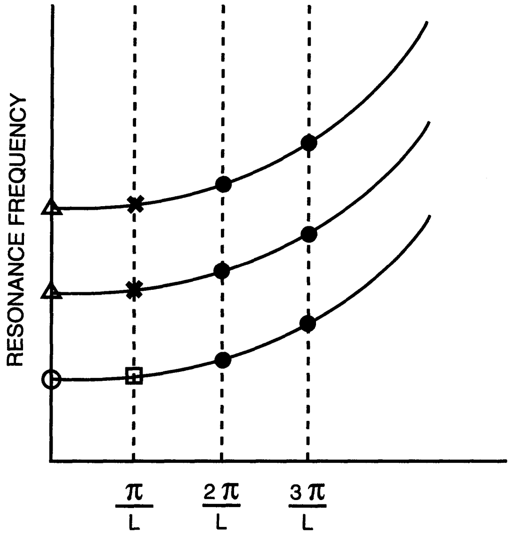

FIG. 2. Thickness-mode resonance frequencies in plates with infinite width. Open circles, fundamental thickness-mode frequency (infinite length). Triangles, overtone thickness-mode frequencies (infinite length). Squares, fundamental thickness-mode frequency (finite length). Crosses, overtone thickness-mode frequencies (finite length). Solid circles, two lowest anharmonic mode frequencies of thickness fundamental modes and overtone modes (finite length).

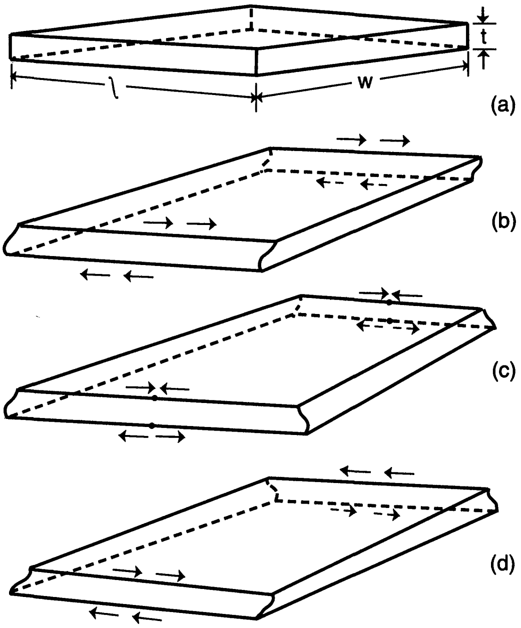

FIG. 3. Thickness-shear motion of plates: (a) undeformed plate, (b) fundamental thickness mode, (c) lowest-frequency length anharmonic thickness mode, (d) lowest-frequency width anharmonic thickness mode. From Sykes (1982); reproduced by permission of Van Nostrand Reinhold Company.

electrodes are placed on a plate or disk with infinite lateral dimensions, the mechanical motion depends on the frequency lowering of the electroded region relative to the unelectroded region (the relative frequency lowering is called plateback) and on the lateral electrode dimensions. The frequency lowering depends on the electrode mass and on the external electrical elements connecting the electrodes (open or short circuited, etc.). Electrodes with a given plateback and with lateral dimensions larger than a critical value (different for fundamental and each overtone) excite mechanical displacement components under the electroded area that are similar to those shown in Fig. 1, except that the motion amplitude will not be uniform across the electroded region. The motion amplitude in the unelectroded regions of the plate decays exponentially with distance away from the edges of the electrodes. This situation is called energy trapping (Holland and EerNisse, 1969, Chap. 5, pp. 179–244; Spencer, 1972; Smythe, 1985). For anharmonics of fundamental thickness modes and for electrode lateral dimensions smaller than the critical value, the entire plate has mechanical displacement components similar to those shown in Fig. 1. In this case, the motion amplitude is more uniform everywhere in the plate than the motion amplitude for the completely energy-trapped case. Trapped-energy thickness-mode resonators have higher quality factors and unwanted modes with reduced amplitudes (Meeker, 1985).

When partially electroded plates or disks have finite lateral dimensions, the mechanical motion is not completely trapped in the region of the electrodes and becomes a sinusoid with decaying amplitude in the unelectroded region of the plate. Resonance and antiresonance frequencies and the extent of the trapping depend on the lateral dimensions of the electrode and plate or disk, as well as on the electrode mass (Spencer, 1972).

These effects of lateral boundaries (edges of plate or electrodes) on the solutions for the piezoelectric resonator can be understood in terms of waves propagating in the plane of piezoelectric plates (Lee and Syngellakis, 1975; Wagers, 1977). Important basic concepts can be obtained from the theory of mechanical wave propagation in the plane of elastic plates (Meeker and Meitzler, 1964).

Important and widely used theories of elastic and piezoelectric plates and electrodes with both length and width finite have been developed (Tiersten, 1969). Coupling between the thickness and contour (flexure, extension, face shear) modes of the plate is included in some of these theories. These approximations have had considerable success in providing an understanding of lateral boundary conditions on the properties of piezoelectric resonators (Meeker, 1985).

2.1.4 Thickness-Mode Resonators Using Contoured Plates Properly designed finite electrodes offer one way of confining the vibrational motion of a thickness-mode resonator to a desired region of the plate. Another way of confining the motion is to make the plate thicker in the desired region of the plate (usually the central region). Resonators with very high quality factor and reduced undesired modes can be made by properly contouring the plate in this way. Many references to the desirable properties of contoured thickness mode resonators can be found in the literature (*Proceedings of the*

Annual Symposium on Frequency Control). Further discussion of the properties of contoured thickness mode resonators is beyond the scope of this article.

2.2 Contour Modes in Bars, Cylinders, Plates, and Disks

In this section, some of the characteristics of modes in plates, bars, disks, and cylinders with resonance frequencies that depend mostly on length, width, and radius are reviewed. For rectangular plates, a wave-number–type approximate theory that couples thickness, extension, flexure, and face-shear modes has been published (Meeker, 1985).

2.2.1 Extension in Bars and Cylinders

For length extension in bars or cylinders, the word "longitudinal" is sometimes used in place of the word "extensional." Bars with small enough width and thickness and cylinders with small enough radii have modes with resonance frequencies that depend mostly on the length dimension. The associated wave motions are one dimensional but not plane. Electric field excitation along the length or along one of the lateral dimensions can be used. Excitation along diameters of cylinders requires special electrode geometries (Meeker, 1985). Figure 4 shows the mechanical displacement components associated with extensional modes. An infinite number of resonance frequencies are associated with mechanical displacement component phase reversals along the length. Figure 5 shows how bars with two free ends deform for the fundamental length mode and the next two overtone length modes (Sykes, 1982, Fig. 6.3, p. 208).

The mechanical motion has components parallel and perpendicular to the length direction. A neutral surface separates regions with oppositely directed motions perpendicular to the length direction. Motion components parallel to the length direction are in the same direction on both sides of the neutral surface. Both the cross section (rectangular, circular, elliptical, etc.) and the lateral dimensions (length, width, diameter, etc.) have a small effect on the resonance frequency. The effect of the shape of the cross section is usually small; the effects of lateral dimensions increase with increasing lateral dimensions as bars or cylinders become plates or disks (Holland and EerNisse, 1969, Chap. 3, pp. 96–124; Chap. 4, pp. 125–178). These modes can be excited by electric fields along the directions of the length, width, radius, and thickness.

Many characteristics of these resonators depend on the direction of the electric field excitation (Berlincourt *et al.*, 1964; Meeker, 1985). For thickness electric fields, this is the desired mode in the quartz *E* resonator and

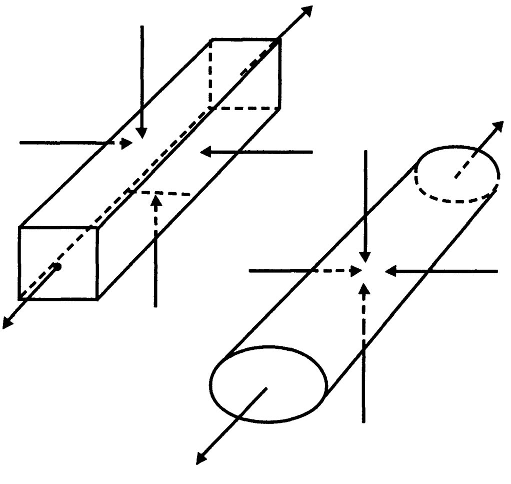

FIG. 4. Mechanical displacement components for fundamental extensional mode in bars and cylinders.

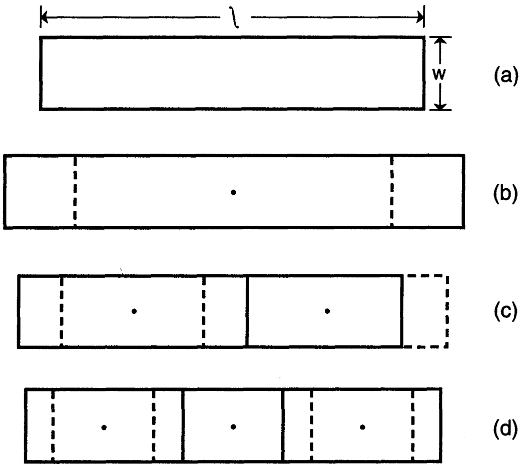

FIG. 5. Extensional deformations for three lowest-frequency modes of a bar with stress-free ends: (a) undeformed bar, (b) fundamental mode, (c) first overtone mode, (d) second overtone mode. From Sykes (1982); reproduced by permission of Van Nostrand Reinhold Company.

the −18°-rotated *X*-cut quartz resonator (Sykes, 1982, pp. 211–214).

2.2.2 Extension in Plates and Disks

The word "longitudinal" is sometimes used in place of the phrase "plate extensional." The mechanical motion is an extension in the direction of one lateral dimension accompanied by a contraction in the direction of the other lateral direction. Bars and cylinders with larger lateral dimensions have mechanical displacement components along the two lateral dimensions (Sykes, 1982, pp. 214–219). Phase reversals of mechanical displacement components along both lateral directions are associated with mechanical motions identified by three numbers. For the cylinder the three numbers denote the number of phase reversals along length, radius, and a rotation angle in a plane perpendicular to the length axis. A review of the wave motions for isotropic plates and cylinders with infinite lateral dimensions (Meeker and Meitzler, 1964) shows the detailed mechanical displacement components for four independent mode types called longitudinal, flexural, symmetric shear, and antisymmetric shear. For longitudinal and flexural modes, the mechanical displacement components lie in a sagittal plane defined by the propagation direction and the normal to the plate surfaces. The two shears have mechanical displacement components in the plane of the plate. For cylinders, the four mode types are longitudinal, flexural, symmetric torsion, and antisymmetric torsion. For piezoelectric plates, the four types of waves are similar to those in the isotropic plate for some materials and for some propagation directions. In the general case, the four simple motion types are not independent, and the actual motion can be considered to have contributions from each of the four simpler types.

Longitudinal modes in isotropic plates have mechanical displacement components in the sagittal plane, as shown in Fig. 6. When the shear modes contribute to the actual motion, the mechanical motion is not planar. Higher-order longitudinal modes can be named with three numbers denoting the approximate number of phase reversals of mechanical displacement along all three coordinate directions. For isotropic cylinders, families of longitudinal modes are characterized by phase reversals of mechanical displacement along the length, radius, and a rotation in the plane normal to the length axis.

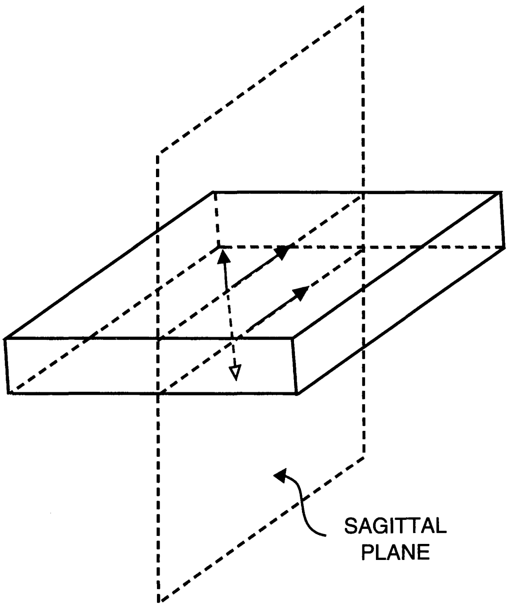

FIG. 6. Mechanical displacement components for longitudinal modes in isotropic plates.

For nearly square plates, coupled two-dimensional extension is the mechanical motion of quartz *GT* elements, for which lengths and widths can be adjusted to give a very small dependence of resonance frequency over a wide range of temperatures (Sykes, 1982, pp. 211–214). Extension is the desired mode in the *X*-cut rotated quartz *MT* element (Sykes, 1982, pp. 211–214). Extension in a plate with circular lateral shape is called a radial mode (Berlincourt *et al.*, 1964).

2.2.3 Flexure in Bars and Cylinders

Flexural mechanical displacement components produce a rotation of the cross section of the bar or cylinder length direction about an axis that is perpendicular to the length direction, as shown in Fig. 7. The resulting bending of the bar or cylinder is accompanied by an expansion on one side of the neutral surface and a contraction on the other. Figure 8 shows deformations for the three low-order flexural modes of a bar with two free ends (Sykes, 1982, Fig. 6.1, p. 206). Because of the antisymmetry, these modes must be excited by specially placed electrodes, by using a composite material, or by

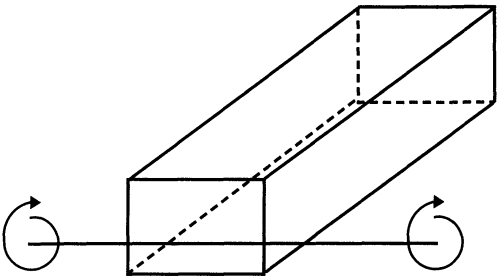

FIG. 7. Flexural deformations in bars.

modifying the material (viz., in ceramics, using inhomogeneous polarization) to provide the necessary asymmetry to couple the motion to the electric field. The quartz *J* element uses a thickness-excited composite quartz material (two extensional bars glued together so that one bar expands and the other contracts with the same direction of electric field). Examples using special electrode distributions to excite these modes are the +5° *X*-cut and the rotated quartz *NT* element (Sykes, 1982, pp. 211–214; Meeker, 1985).

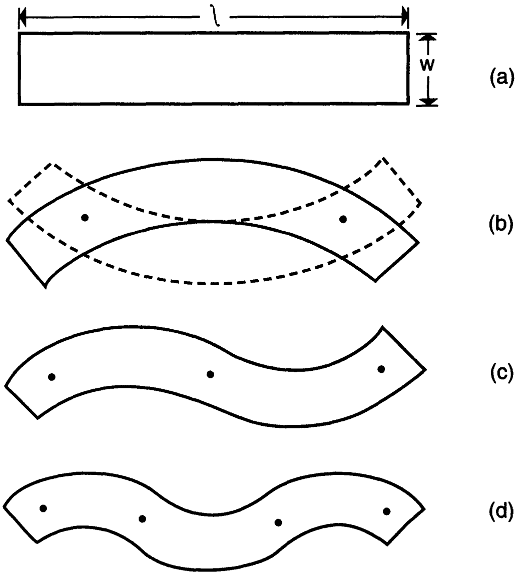

FIG. 8. Flexural deformations for the three lowest frequency modes of a bar with stress-free ends: (a) undeformed bar, (b) fundamental mode, (c) first overtone mode, (d) second overtone mode. From Sykes (1982); reproduced by permission of Van Nostrand Reinhold Company.

2.2.4 Flexure in Plates and Disks For isotropic plates, mechanical displacement components lie in the sagittal plane, as shown in Fig. 9. The mechanical motion has components parallel to and perpendicular to the length direction. A neutral surface separates regions with parallel motion components in opposite directions. Perpendicular motion components are in the same direction on both sides of the neutral surface. Figure 10 shows deformations of a flexural plate with four free edges (Sykes, 1982, Fig. 6.2, p. 207). The mechanical motion is in the direction of the plate length, width, or radius. A small thickness motion, antisymmetric about the width-length plane, accompanies the motion (Sykes, 1982, pp. 214–219). The motion is a twist of a cross section about an axis in the plane of the plate and normal to the direction of the plate length, width, or radius. This is the desired mode in the *X*-cut rotated quartz *NT* element (Sykes, 1982, pp. 211–214).

2.2.5 Face Shear in Plates Two-dimensional motions of thin face-shear plates with different numbers of phase reversals along the length and width directions are shown in Fig. 11 (Sykes, 1982, Fig. 6.4, p. 210). The mechanical motion is a shear in

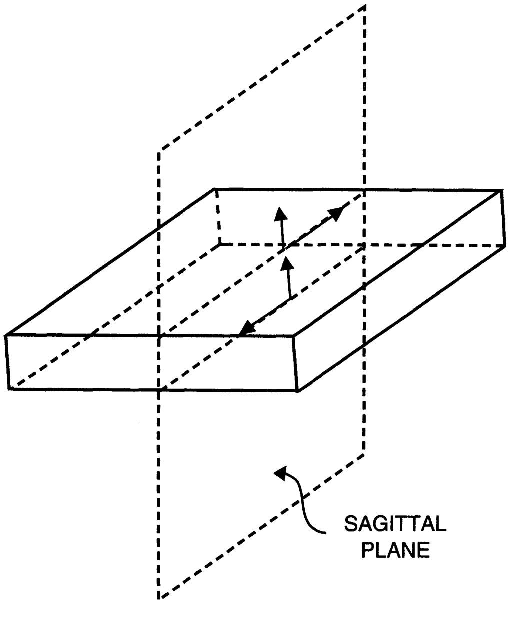

FIG. 9. Mechanical displacement components for flexural modes in isotropic bars and plates.

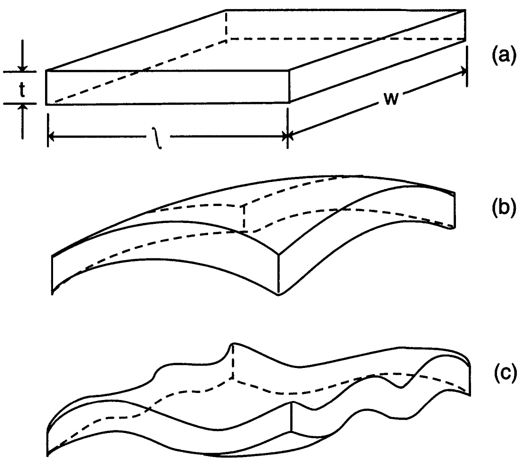

FIG. 10. Deformations for two flexural modes of a plate with four stress-free edges. (a) Undeformed plate, (b) fundamental plate mode with no phase reversals along length or width, (c) overtone plate mode with one phase reversal along the length and three phase reversals along the width. From Sykes (1982); reproduced by permission of Van Nostrand Reinhold Company.

the plane determined by the directions of the length, width, or radius of the plate. Face-shear modes can be excited by electric fields in the direction of one of the lateral dimensions or the thickness (Meeker, 1985). For thickness excitation, this is the desired mode in the quartz *CT* and *DT* resonators (Sykes, 1982, pp. 214–219). The corresponding mode in circular disks is a torsion (Sec. 2.3.2).

2.3 Torsional Modes

2.3.1 Torsion in Bars and Cylinders

The only mechanical displacement component is a twist around the length axis in a plane normal to the length axis, as shown in Fig. 12. For simple torsion, phase reversals along the length characterize an infinite number of modes. For higher-order modes and larger lateral dimensions, mechanical displacement components have phase reversals along the length and radius (Mason, 1964a, Chap. 6, pp. 78–113; Meeker and Meitzler, 1964; Timoshenko *et al.*, 1974; Meeker, 1985). The bar cross section can be circular, elliptical, or rectangular.

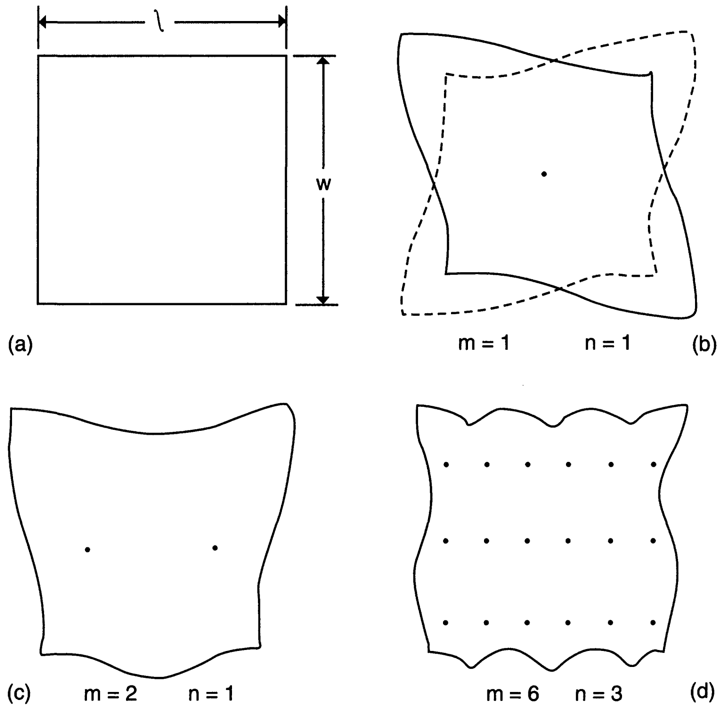

FIG. 11. Face-shear deformations in a plate: (a) undeformed plate, (b) fundamental plate mode with one phase reversal along the length and one phase reversal along the width, (c) overtone plate mode with two phase reversals along the length and one phase reversal along the width, (d) overtone plate mode with three phase reversals along the width and six phase reversals along the length. From Sykes (1982); reproduced by permission of Van Nostrand Reinhold Company.

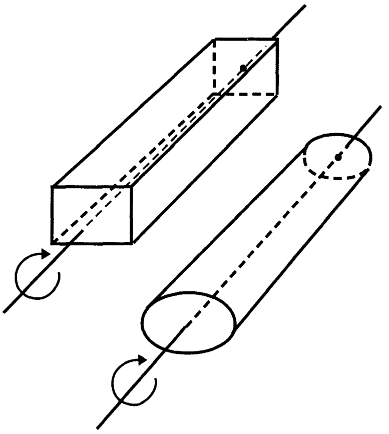

FIG. 12. Torsional motions in bars and cylinders.

2.3.2 Torsion in Disks Thickness torsion is a circular shear or twist in the plane of the plate, which is normal to the direction of the plate thickness. The plate lateral shape can be circular, elliptical, rectangular, etc. This mode corresponds to plate face shear in plates. Phase reversals of mechanical displacement along the thickness direction define overtone modes. Phase reversals of mechanical displacement along the radius and in rotation angle produce circular deformations similar to those shown in Fig. 11 for plates.

3. BULK ACOUSTIC-WAVE RESONATORS

3.1 Quartz-Crystal Resonators

As practical examples of the properties of resonators using the modes discussed in Sec. 2, several properties of quartz resonators have been tabulated (Meeker, 1985). Many piezoelectric crystalline quartz resonators have had a short name (usually one or two letters) assigned to designate the mode. These names, orientations of the resonator in the quartz crystalline lattice, modes of vibration, and useful resonance frequency ranges are shown in Table 1. Table 2 shows the approximate dependence of resonance frequency, series motional capacitance, and series motional inductance on the dimensions of quartz resonators. Also shown in Table 2 are ratios of capacitances (clamped dielectric capacitance to motional capacitance) and series motional resistances. The ratios of capacitances, series motional capacitances, series motional inductances, and series motional resistances refer to electrical elements of the simple lumped-element equivalent circuit for the various resonator types. This simple electrical equivalent circuit is shown in Fig. 13. Equivalent electrical circuits are described in Sec. 6. For quartz-crystal resonators, Figs. 14 and 15 show the dependence of resonance frequency on mode type and temperature. Resonator orientations in the right-hand alpha-quartz crystal lattice are shown for several types of quartz resonator modes in Fig. 16.

3.2 Resonators Using Other Materials (Ceramics, Crystals, Films)

Piezoelectric ceramics and crystals are selected for use in resonators for reasons of cost and performance. Ceramics and some crystals have high piezoelectric coupling factors that are useful in resonators for wide-bandwidth or low-impedance filters. Al-

Table 1. Mode names and properties of quartz-crystal resonators. From Meeker (1985); reproduced by permission of Academic Press.

Element name	Reference	Vibration mode	Frequency range (MHz)
A	*AT*	T shear	0.5–100
B	*BT*	T shear	5–15
C	*CT*	Plate shear	0.3–1
D	*DT*	Plate shear	0.2–0.5
E	−5° *X*-cut	Longitudinal	0.06–0.3
F	+18.5° *X*-cut	Longitudinal	0.06–0.3
G	*GT*	Longitudinal	0.1–0.55
H	−5° *X*-cut	L-W flexure	0.01–0.1
J	−5° *X*-cut 2 plates	Duplex L-T flexure	0.0012–0.01
M	*MT*	Longitudinal	0.01–0.1
N	*NT*	L-W flexure	0.01–0.1

Table 2. Typical properties of quartz-crystal resonators. All linear dimensions are in meters: t is thickness, w is width, l is length, and n is overtone number. In the simple equivalent circuit of Fig. 13, C_0 is clamped dielectric capacitance, C_1 is series motional capacitance, L_1 is series motional inductance, and R_1 is series motional resistance. From Meeker (1985); reproduced by permission of Academic Press.

Element name	Resonance frequency f (kHz)	C_0/C_1	C_1 (pF)	L_1 (H)	R_1 (kΩ)
A (n)	$1.6n/t$	$250n^2$	$0.097wlf/n^3$	$2.62 \times 10^5 n^3/lwf^3$	0.1
B	$2.56/t$	650	$0.0242wlf$	$10.5 \times 10^5/lwf^3$	0.1
C	$3.07/l$	350	$1.08/tf^2$	$23\,300t$	1.0
D	$2.07/l$	400	$0.43/tf^2$	$59\,000t$	1.0
E	$2.82/l$	125	$0.383/tf^2$	$66\,000t$	1.0
F	$2.56/l$	130	$0.301/tf^2$	$84\,000t$	1.0
G	$3.37/l$	350	$1.52/tf^2$	$16\,700t$	0.1
H	$5.00w/l^2$	190	$0.0179/tf^2$	$1.42 \times 10^6 t$	10.0
J	$5.60t/l^2$	200	$2.54/10^4 tf^2$	$1.0 \times 10^6 t$	10.0
M	$2.80/l$	190	$0.664/tf^2$	$38\,200t$	1.0
N	$5.60w/l^2$	900	$2.42/10^3 tf^2$	$1.05 \times 10^7 t$	10.0

though piezoelectric films are flexible and not likely to break during processing into high-frequency resonators, most of the reported work involves transducer applications (Meeker, 1988). Examples of ceramic materials are lead zirconate titanate and lead titanate (Fujishima *et al.*, 1984). Examples of crystal materials are lithium tantalate, lithium niobate, and berlinite (Sheahan and Johnson, 1977; Smythe, 1985). Studies on many other useful ceramic and crystal materials have been reported (Meeker, 1985).

4. COMPOSITE ELECTROMECHANICAL RESONATORS

For many applications, it is desirable to make resonators from nonpiezoelectric materials such as glasses or metals, which can be excited by piezoelectric transducers attached to a particular place on the resonator body. The thickness asymmetry required to excite flexural-mode resonators can be achieved by fastening two properly oriented plates of piezoelectric material together. Some applications of composite electromechanical resonators are discussed in Sec. 8. Proprietary iron–nickel alloys with small amounts of other metals, such as chromium or molybdenum, can have compositions adjusted to give a low-temperature coefficient of resonance frequency and a sufficiently low dissipation for particular modes of interest (Sheahan and Johnson, 1977). Film transducers have potential applications as composite resonators (Linville, 1988; Meeker, 1988).

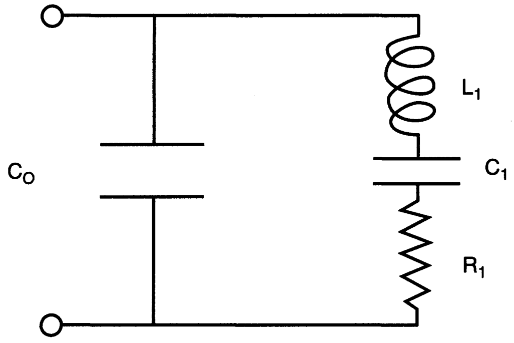

FIG. 13. Simple electrical circuit of single-mode piezoelectric resonator. From Meeker (1985); reproduced by permission of Academic Press.

5. SURFACE ACOUSTIC-WAVE RESONATOR MODES ON PLATES

The theory for a single-surface semi-infinite half-space is useful for the practical case of a plate that is many wavelengths thick. When the propagation direction of the assumed mechanical wave motion is parallel to a stress-free surface, solutions to the equations of Sec. 1.1 have been found for several different cases. Depending on the direction of the mechanical displacement and the orientation and nature of the surface, two different types of surface wave motion can be identified. Rayleigh and Gulyaev–Bleustein Waves are discussed in Secs. 5.1 and 5.2.

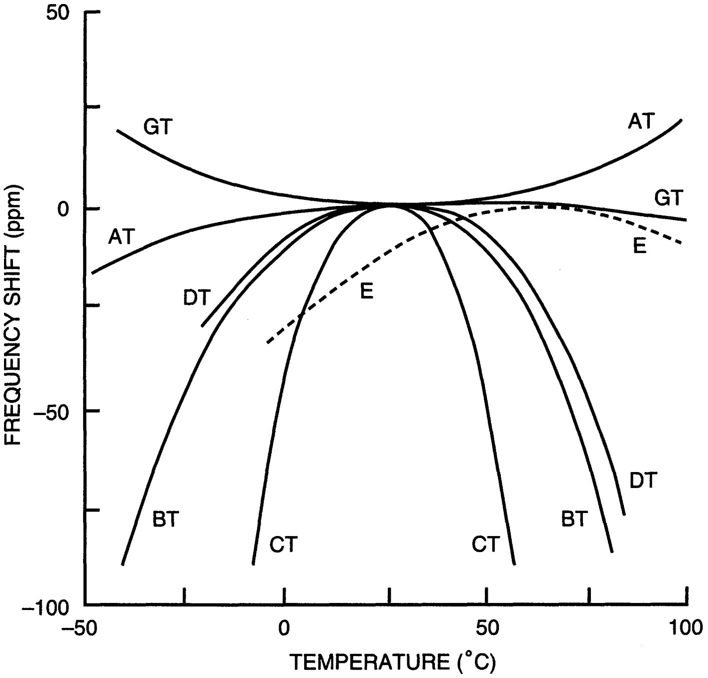

FIG. 14. Dependence of resonance frequency on temperature for quartz-resonator types *GT*, *AT*, *E*, *DT*, *CT*, and *BT*. From Meeker (1985); reproduced by permission of Academic Press.

5.1 Rayleigh Waves

For isotropic elastic materials, Rayleigh surface waves (Meeker and Meitzler, 1964) have mechanical displacements entirely in the sagittal plane that contains the directions of the wave propagation and the normal to the surface. The mechanical displacement has components along the directions of the propagation and the normal to the surface. There is no coupling to modes with displacement components normal to the sagittal plane. The wave motion has a propagation vector with a real component in the propagation direction and an imaginary component in the plate normal direction. The surface-wave mechanical displacement amplitude decays exponentially away from the surface. The imaginary wave number along the direction perpendicular to the surface means that no acoustic energy is propagated into the elastic medium; the major

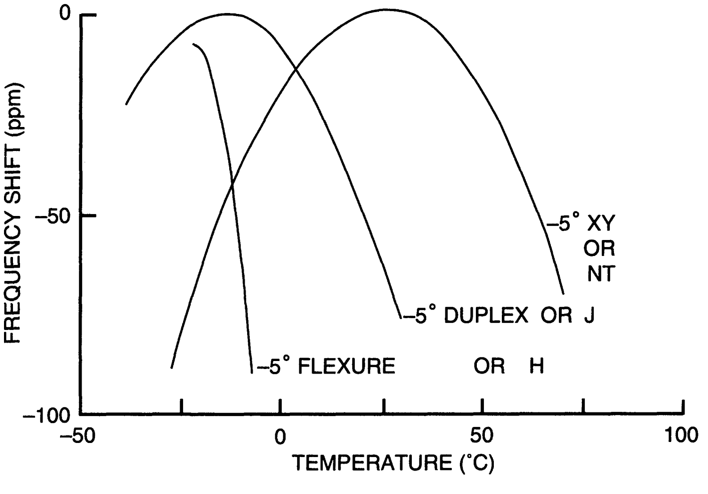

FIG. 15. Dependence of resonance frequency on temperature for quartz-resonator types *F*, *XD*, and *NT*. From Meeker (1985); reproduced by permission of Academic Press.

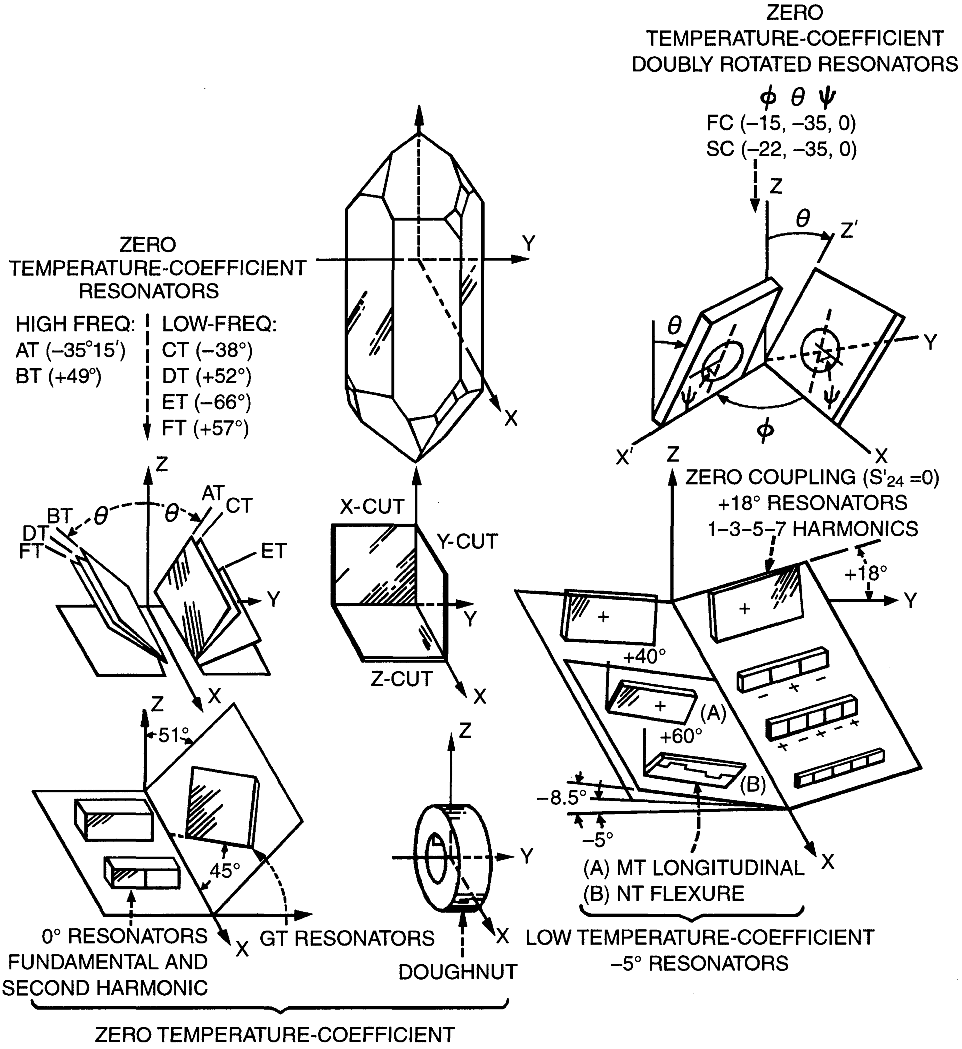
ZERO
TEMPERATURE-COEFFICIENT
DOUBLY ROTATED RESONATORS
φ θ ψ
FC (−15, −35, 0)
SC (−22, −35, 0)
Z
θ
Z'
θ
Y
ψ
ψ
φ
X'
X
ZERO
TEMPERATURE-COEFFICIENT
RESONATORS
HIGH FREQ:
AT (−35°15')
BT (+49°)
LOW-FREQ:
CT (−38°)
DT (+52°)
ET (−66°)
FT (+57°)
Y
X
Z
BT
DT
FT
θ
θ
AT
CT
ET
Y
X-CUT
Y-CUT
Y
Z-CUT
X
Z
ZERO COUPLING (S'24 =0)
+18° RESONATORS
1-3-5-7 HARMONICS
+18°
Y
+
+
+40°
+
(A)
+60°
(B)
−8.5°
−5°
(A) MT LONGITUDINAL
(B) NT FLEXURE
X
Z
X
51°
45°
Z
Y
0° RESONATORS
FUNDAMENTAL AND
SECOND HARMONIC
X
GT RESONATORS
DOUGHNUT
X
LOW TEMPERATURE-COEFFICIENT
−5° RESONATORS
ZERO TEMPERATURE-COEFFICIENT

FIG. 16. Crystal orientations of piezoelectric resonators made from right-hand alpha quartz. From Meeker (1985); reproduced by permission of Academic Press.

portion of the mechanical energy of the surface wave is within a few wavelengths of the surface.

For anisotropic piezoelectric materials, the existence and velocity of the mechanical surface wave depends on the orientations of the surface in the crystal lattice, on the propagation direction in the plane of the surface, and on the mechanical and electrical boundary conditions at the surface. The electrical boundary conditions can involve a relationship between the electric displacement and the electric potential at the surface. Furthermore, transverse surface or bulk modes may be coupled to the Rayleigh wave motions. The coupling to the transverse modes may also depend on the orientations of the surface and the wave propagation in the crystal lattice and on the mechanical and electric boundary conditions on the surface (Farnell, 1970).

5.2 Gulyaev–Bleustein Waves

Transverse or shear-type surface acoustic waves cannot exist alone for a piezoelectric medium unless a layer of material with a lower shear wave velocity is placed on the surface. This lower-velocity layer produces the required condition for the existence of

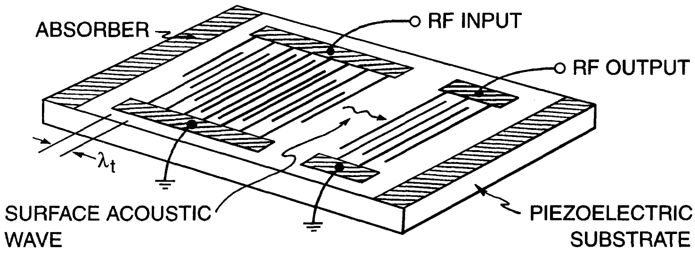

FIG. 17. Schematic of a surface acoustic wave filter with two interdigital transducers. From Shreve and Cross (1985); reproduced by permission of Academic Press.

exponential decay of the transverse wave motion along the direction of the surface normal. The mechanical displacements are in the direction of the plane of the surface and normal to the propagation direction (Farnell and Adler, 1972).

5.3 Devices

Surface acoustic-wave devices became much more useful when the interdigital transducer and reflector were invented (White, 1970; Shreve and Cross, 1985, Chap. 2.3, pp. 119–145; Wagers, 1985). The interdigital transducer has alternate electrodes spaced one half-wavelength apart and driven 180° out of electrical phase. The interdigital transducer can be made of metal deposited on the surface or in grooves etched into the surface. The interdigital reflector can be made of metal strips or other material deposited on the surface or in grooves in the surface; empty grooves can also be used. These elements preserve the planar advantages of surface-wave devices and allow the design and implementation of transversal filters, resonators, and resonator filters. These filters and resonators have found uses in many different applications, which are discussed in Sec. 8.

Figure 17 shows two interdigital transducers used in a transversal filter with a bandwidth mostly determined by the reciprocal of the number of electrode elements. The position and length of the electrode elements can be changed to optimize the performance of the filter (Shreve and Cross, 1985, Chap. 2.3, pp. 119–145; Wagers, 1985). The transversal filter can also be used in oscillator applications. Figure 18 shows how a surface acoustic-wave resonator can be made by exciting the surface wave with a broadband interdigital transducer with a small number of electrode elements and by providing surface-wave reflectors on each side of the transducer to define a resonant cavity (Shreve and Cross, 1985, Fig. 2.3-2, p. 122). The reflectors can be carefully prepared plate edges, grooves, or metal (or other material) elements on the surface. Useful resonator *Q*'s can be attained with a large number (>1000) of reflectors. A two-port resonator can be made from two reflectors and two properly positioned interdigital transducers, each with a small number of elements. Many different electrode configurations have been reported as ways to optimize various performance criteria of the surface-wave device (Shreve and Cross, 1985, Chap. 2.3, pp. 119–145). These resonators can be used in oscillators and filters. Some applications of surface acoustic-wave devices are discussed in Sec. 8.

6. EQUIVALENT ELECTRICAL CIRCUITS

Impedance equations [such as Eq. (7) in Appendix A for the thickness-excited thickness-mode resonator] for the particular mode of interest can be represented approximately by a very useful and intuitive equiv-

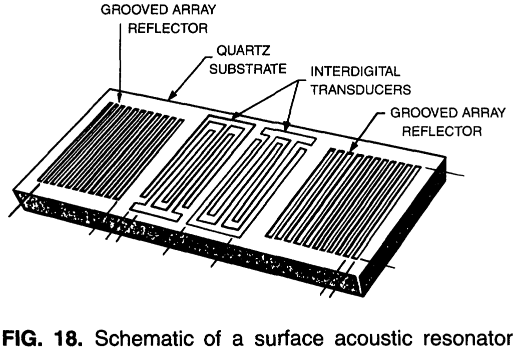

FIG. 18. Schematic of a surface acoustic resonator with two interdigital transducers and two grooved reflectors. From Shreve and Cross (1985); reproduced by permission of Academic Press.

alent electrical circuit with four physically realizable electrical elements (a capacitor in shunt with a series resonant circuit), as shown in Fig. 13 (Cady, 1964, Vol. 1, Chap. 14, pp. 333–383; Hafner, 1985, Chap. 7, pp. 1–44). An important advantage of this approximate equivalent circuit is its simplicity. For many applications, this representation is accurate enough for design purposes or even for final system analyses. This representation can be placed directly in most standard circuit-analysis computer programs, so that the circuit–resonator interfaces can be modeled conveniently. Most piezoelectric resonator specifications include a resonance frequency, an inductance, a ratio of capacitances, and a quality factor (Q) or a series resistance. These terms relate directly to the simple equivalent circuit and are readily understood by the circuit-design users of the piezoelectric resonator. For many applications, the advantages of expressing the characteristics of the piezoelectric resonator in terms of a small number of simple electrical elements far outweigh any disadvantages associated with the reduced accuracy of the representation. For some purposes, the simple equivalent circuit is not accurate enough, and exact equivalent circuits or physical electrical impedance relations must be used to represent resonator characteristics (Ballato, 1977, pp. 126–129; Meeker, 1985).

7. MATERIALS

The selection of a material for a piezoelectric resonator to be used in a particular application involves a number of complex and sometimes contradictory ways in which the material characteristics affect the performance of the resonator. For a given frequency, the size and shape of the desired resonator depend on the choice of resonance mode and on the associated material constants. The dependence of resonance frequency and other resonator characteristics on temperature, pressure, acceleration, drive level, time, etc. are controlled by the resonance mode, by the orientation of the resonator body in the crystal lattice, and by the dependence of the relevant elasto-piezo-dielectric material constants on the appropriate environmental variables. The mechanical and dielectric dissipations (Berlincourt *et al.*, 1964; Ballato, 1977, Chap. 5, pp. 115–181) and material nonlinearities vary considerably from material to material. Zeros in the elasto-piezo-dielectric matrix for a particular orientation of the resonator body in the crystal lattice depend on the choice of material and can be used to control the activity of particular unwanted mode resonances. For applications requiring very low dissipation and narrow bandwidth, quartz has been the material of choice because of the general resonator stability that can be achieved over temperature and time. Other applications, in which wide bandwidth and low resonator impedance are the most important resonator properties, use piezoelectric ceramics such as lead zirconate titanate (PZT) (Fujishima *et al.*, 1984). Other materials, such as berlinite, lithium niobate, and lithium tantalate have properties that lie between those of quartz and the ceramics (Meeker, 1985). Piezoelectric plastic film resonators have advantages for some applications (i.e., higher material linearity and mechanical flexibility), but the scope of applications has not been fully defined (Meeker, 1988). Composite electromechanical resonators (Sec. 4) can be made of metal with bonded piezoelectric transducers (Sheahan and Johnson, 1977) or deposited piezoelectric films. These resonators have been used in many commercial applications (Sec. 8). The material properties of deposited metal film electrodes, metal or glass enclosures or seals, and gases packaged with the device all affect the performance of the resonator or filter.

8. APPLICATIONS OF PIEZOELECTRIC RESONATORS

Table 3 gives some reasons for the popularity of piezoelectric resonators in certain applications. Some disadvantages of piezoelectric resonators are also shown in Table 3.

8.1 Piezoelectric Filters

A common method of designing a piezoelectric filter is to design an electric filter with inductors, capacitors, and resistors to meet the desired requirements. Lattice and ladder filter designs are often used (Mason, 1964b; Smythe, 1985). The lattice filter has the topology of a bridge and can be balanced

Table 3. Comparison of some characteristics of piezoelectric and electrical (inductance–capacitance) resonators.

	Piezoelectric	Electrical
Electromagnetic radiation and absorption	Smaller	Larger
Temperature coefficient of frequency	Smaller	Larger
Time drift and aging	Smaller	Larger
Resonance frequencies	Many	One
Antiresonance frequencies	Many	None
Unwanted resonance modes	Many	None
Useful frequencies	Higher	Lower

for stop-band attenuation (reflection) and unbalanced for pass-band transmission. Ladder filters have circuit sections in series; more circuit sections provide greater stop-band attenuation. In the next four sections, discrete crystal filters, monolithic crystal filters, electromechanical filters, and surface acoustic-wave filters are discussed.

8.1.1 Discrete Crystal Filters After a suitable electric filter is designed, piezoelectric resonators are substituted for the electric resonators formed from the appropriate inductor–capacitor combinations. The resonator couplings are left in electrical form. This kind of filter is called a discrete crystal filter, since the resonant elements of the filter are discrete crystal resonators. The piezoelectric resonator is equivalent to an inductor–capacitor electrical resonator over a narrow range of frequencies. The frequency difference between resonance and antiresonance frequencies of the piezoelectric resonator restricts the bandwidth of simple piezoelectric discrete crystal filters. Electrical inductances can be used in parallel with the piezoelectric resonator to widen the useful electrical bandwidth of the filter, especially at lower frequencies. For smaller electrical bandwidths at higher frequencies, the quality factor of the piezoelectric resonator can be much higher than the quality factor of the inductor–capacitor electrical resonator. Higher-frequency discrete crystal filters can have lower midband insertion losses than filters using discrete electrical components. The filter can be called a discrete crystal filter, a discrete ceramic filter, a discrete composite filter, or a discrete film filter, depending on the type of resonator used. Both lattice and ladder discrete-component piezoelectric filters have been widely used in numerous applications for many years (Smythe, 1985). These applications include voice channel and signaling filters in analog communication systems and intermediate-frequency filters for radio and television.

Filters that contain at least one coupled resonator pair on a piezoelectric plate and have no electrical coupling elements are called monolithic crystal filters because popular examples use crystal plates. Monolithic crystal filters are discussed in Sec. 8.1.2.

8.1.2 Monolithic Crystal Filters When two or more trapped-energy (Sec. 2.1.3 for thickness shear) piezoelectric resonators occupy the same piece of piezoelectric material and the resonator coupling is a mechanical interaction between the resonators, the filter is called a monolithic crystal filter. A popular type of monolithic crystal filter uses more than one mechanically coupled monolithic crystal filter section with electrical coupling between the sections. Sometimes, parasitic electrical elements between sections can be used to provide the electrical coupling required by the design (Spencer, 1972; Sheahan and Johnson, 1977; Smythe, 1985).

These filters have been used in voice channel and signaling in communication systems, as front-end filters for cellular and mobile radio, as intermediate filters for radio and television, and for clock-recovery filters for timing extraction in repeatered digital transmission systems (Sheahan and Johnson, 1977; Fujishima *et al.*, 1984; Smythe, 1985).

8.1.3 Electromechanical Filters Electromechanical filters are mechanical filters for which input and output resonators are electrically (electrostrictively or piezoelectrically) or magnetically coupled to the use system. Only piezoelectrically coupled electromechanical filters are considered in this article. Except for the input and output resonators, the interior resonators are usually mechanical resonators. The input and output resonators are often composite resonators made from mechanical resonators with an attached (cemented, deposited, etc.) piezoelectric film or transducer (Sec. 7). Resonators can be rods, bars, plates, disks, tuning forks, etc. Couplers can be wires, ribbons, bars, etc. Resonator and coupler modes can be flexure, torsion, extension, etc. Piezoelectric ceramic has usually been the material selected for the piezoelectric part of the composite input and output resonators.

Although other materials, such as glass, could be used for the resonators, metal seems to offer advantages of lower processing cost with adequate performance for most applications. Several useful reviews contain many details of design, manufacture, and application for these filters (Mason, 1964b; Sheahan and Johnson, 1977; Smythe, 1985). Other filters are mechanical filters with mechanical resonators and mechanical coupling.

8.1.4 Surface Acoustic-Wave Filters Reviews of several consumer communication applications discuss filters for radar pulse compression, television intermediate-frequency control, television video–audio signal separation, and mobile communication radio frequency control. Detailed design and performance evaluations are presented (White, 1970; Ruppel *et al.*, 1993).

8.2 Piezoelectric Crystal or Electromechanical Oscillators

A crystal or electromechanical oscillator is a closed-loop circuit that contains an amplifier and a crystal or electromechanical resonator (Cady, 1964, Vol. 2, Chap. 19, pp. 489–509). The frequency at which the total loop phase shift is an integral number of cycles is the oscillator frequency. At the oscillator frequency, the amplitude grows until the circuit nonlinearities reduce the total loop gain to unity. If the starting low-amplitude circuit gain is less than unity at the oscillator frequency, oscillation will not occur. Oscillation will not occur if the loop gain and phase requirements are not satisfied simultaneously. Usually, the loop gain and phase shift of the oscillator circuit are so sensitive that computer modeling of a simple oscillator circuit is not sufficient to arrive at a useable design (Frerking, 1978; Parzen and Ballato, 1983; Frerking, 1985; Smith, 1985), and some experimental adjustments are necessary.

Applications include clocks, computers, household appliances, analog and digital communications systems, radios, and televisions.

8.3 Piezoelectric Sensors

Many sensor applications use a piezoelectric resonator to convert some environmental change into a change of resonance frequency. Two very important advantages of using a piezoelectric-resonator sensor are the large frequency shifts associated with small changes in the resonator environment and the availability of very accurate frequency-measurement equipment at a reasonable cost. In some cases, hysteresis in the characteristic frequencies limits the applicability of the sensor. For all cases, the availability of relatively inexpensive microprocessors allows rapid and on-line calculations using the measurements and a display of the desired information.

Sensor applications (Walls and Gagnepain, 1985; EerNisse *et al.*, 1988) include temperature measurement, pressure measurement (Besson *et al.*, 1993), thin-film mass measurement, thin-film stress measurement, bulk mass measurement, gas or vapor identification and concentration measurement (Frye *et al.*, 1991; Watson *et al.*, 1991), acceleration measurement, force measurement, and measurement of liquid viscosity and liquid–solid phase transitions (Hoummady *et al.*, 1991).

APPENDIX A: THICKNESS-EXCITED THICKNESS MODES IN PLATES

A.1 Infinite Length and Width

The condition of infinite plate length and width is met by setting all derivatives in the differential equations (Sec. 1.1) with respect to the length and width directions equal to zero for all time and for all points in space. Trial plane-wave mechanical displacement solutions of the differential equations are expressed in terms of $\cos(\text{arg}^+)$, $\sin(\text{arg}^+)$, $\cos(\text{arg}^-)$, and $\sin(\text{arg}^-)$, where

$$\text{arg}^+ = \omega t - m_r x_r \quad \text{and} \quad \text{arg}^- = \omega t + m_r x_r,$$

with m_r as the wave number for the selected propagation direction x_r and ω the angular frequency. Substituting the trial wave solutions into the differential equations converts them into four algebraic equations.

The quasistatic approximations in Sec. 1.1

allow a separation of the four coupled algebraic equations into three coupled algebraic equations in the mechanical displacements and a fourth equation for the electric potential in terms of the same mechanical displacement components. The three coupled equations can be solved for the mechanical displacement components, which can then be used to calculate the electric potential from the fourth equation. The three coupled homogeneous algebraic equations have wave-type solutions when the determinant of the coefficients is zero. There are three values of wave velocity (the products $v_r^{(n)} \times v_r^{(n)}$ are the eigenvalues of the coefficient matrix) for which the determinant is zero. Associated with each of the three eigenvalues is an eigenvector, with a component along each of the coordinate axes. Normalized eigenvector components are $U_{qr}^{(n)}$. Each value of $v_r^{(n)}$ gives a linear dependence of the frequency on the associated wave number m_r. These linear relations between frequency and wave number are called dispersion relations.

The three eigenvectors are orthogonal and can be normalized. One of the eigenvectors has the largest component along the chosen propagation direction (quasidilatational), and the other two eigenvectors have largest components normal to the chosen propagation direction and to each other (quasishears). Of these two, the one with the smaller v_r can be called the slow quasishear, and the other can be called the fast quasishear. The four homogeneous equations are the same for trial solutions that are antisymmetric and symmetric in the plate thickness coordinate. The electrical equation can be integrated over x_r in terms of two integration constants.

The general solution is the sum of the six solutions (three antisymmetric and three symmetric) with arbitrary coefficients, along with the integral of the electrical equation. There are eight coefficients to be determined from the three zeros of T_{ir} and one ϕ on each of the two surfaces of the plate. Components of mechanical displacement, velocity, strain, and stress and of electric field, displacement, and current can be calculated directly from the general solution.

When the boundary conditions are either symmetric or antisymmetric, the symmetric solution set and the antisymmetric solution set are independent of each other and can be obtained separately. Furthermore, the solution set that has symmetric mechanical displacement components is not driven by the applied electric potential.

If the boundary conditions are neither symmetric nor antisymmetric, i.e., different layer mass per unit area on the two surfaces, then the symmetric and the antisymmetric solution sets are dependent and the symmetric solution modes are excited directly by a thickness electric field.

A.2 Electrical Impedance per Unit Area

For the resonator with zero thickness electrodes, the surface stresses are zero and the antisymmetric solution can be expressed as an electrical impedance as

$$Z = \frac{1}{jB_0}\left\{1 - \sum_n \left[2(k_r^{(n)})^2 \frac{T_h^{(n)}}{m_r^{(n)}H}\right]\right\}, \tag{7}$$

where n identifies the three eigensolutions, j is the imaginary number identifying the phase angle or reactive nature of the impedance, $B_0 = \omega C_0$ is the susceptance of the resonator when no mechanical motion is allowed (clamped), ω is the angular frequency,

$$C_0 = \frac{WL}{H}\epsilon_{rr}^S \tag{8}$$

is the clamped capacitance, W is the plate width, L is the plate length, H is the plate thickness,

$$m_r^{(n)} = \omega / v_r^{(n)}, \tag{9}$$

$$k_r^{(n)} = \frac{\sum_q U_{qr}^{(n)} e_{rqr}}{v_r^{(n)}[\rho\epsilon_{rr}^S]^{1/2}}, \tag{10}$$

and

$$T_h^{(n)} = \tan(0.5Hm_r^{(n)}). \tag{11}$$

At any given frequency, the impedance of Eq. (3) is defined by C_0, three $k_r^{(n)}$, and three $m_r^{(n)}$. This impedance has three sets of modes, each with a resonance frequency ($Z = 0$) and an antiresonance frequency ($Y = 1/Z = 0$). These three mode sets are one quasidilatational mode and two quasishear modes. For each set, there is a fundamental and an infinite number of thickness overtones. The separation between corresponding resonance and antiresonance frequencies depends on the associated $k_r^{(n)}$, which are called piezoelectric coupling factors.

A.3 Resonators with Loaded Surfaces

A.3.1 Frequency Shifts due to Electrical Loading The open-circuit (or high-driving-impedance or small-electrical-current) resonance frequency is nearly equal to the antiresonance frequency. The short-circuit (or low-driving-impedance or small-electrical-voltage) resonance frequency is nearly equal to the resonance frequency.

A.3.2 Frequency Shifts due to Mass Loading The mechanical loading of two electrode layers that are thin enough can be represented by a mass per unit area. For this case, the effect of the electrode mass is to change $[k_r^{(n)}]^2$ in Eq. (7) to

$$\frac{[k_r^{(n)}]^2}{1 - (\rho_e H_e/\rho \mathrm{H}) \times 0.5 H m_r^{(n)} T_h^{(n)}}, \tag{12}$$

where ρ_e is the density of the electrode material and H_e is the combined thickness of the two electrode layers. For small H_e, both resonance and antiresonance frequencies are shifted downward (Ballato, 1977, pp. 126–129).

A.3.3 Frequency Shifts due to Nonuniform Mechanical Motion Thickness-mode resonance frequencies and piezoelectric coupling depend on the uniformity of the spatial distribution of mechanical motion. Antiresonance frequencies do not depend on the uniformity of mechanical motion. This means that the effects of nonuniform distribution of mechanical motion can be incorporated into the theory by multiplying the coupling {$[k_r^{(n)}]^2$ in Eqs. (7) and (10)} by a factor that depends on the spatial distribution of mechanical motion (Kosinski *et al.*, 1993).

GLOSSARY

Antiresonance Frequency: The high-impedance frequency of zero reactance.

Discrete Crystal Filters: Filters with discrete crystal resonators for resonant elements.

Electrical Differential Equations: $D_{i,i} = 0$, where D_i are the three components of the electrical displacement.

Ladder Filters: Filters with sections in series; more sections provide greater stop-band attenuation.

Lattice Filters: Filters with the topology of a bridge that can be balanced for stop-band attenuation (reflection) and unbalanced for pass-band transmission.

Linear Constitutive Equations: $T_{kl} = c_{klmn}^E S_{mn} - e_{mkl} E_m$ and $D_j = e_{jkl} S_{kl} - \epsilon_{jk}^S E_k$, where $S_{ij} = 0.5(u_{ij} + u_{j,i})$ are the nine components of the mechanical strain; $E_k = -\phi_{,k}$ are the three components of the electric field, with ϕ the electric potential; c_{klmn}^E are the 81 components of the stiffness at constant E; e_{jkl} are the 27 components of the piezoelectric constant; ϵ_j^S are the 9 components of the dielectric permittivity at constant S; and ω is the angular frequency, equal to $2\pi \times$ frequency.

Piezoelectric Effect: Direct: Electric displacement produced by mechanical strain. Converse: Mechanical strain produced by electric displacement.

Resonance Frequency: The low-impedance frequency of zero reactance.

Stress Differential Equations: The equations $T_{ij,t} + f_j = \rho u_{j,tt}$, where T_{ij} are the nine components of mechanical stress, f_j are the three components of the body force, ρ is the material density, and u_j are the three components of the mechanical displacement.

Works Cited

Ballato, A. (1977), in: W. P. Mason, R. N. Thurston (Eds.), *Physical Acoustics—Principles and Methods*, Vol. 13, New York: Academic.

Ballato, A., Hatch, E. R., Lukaszek, T., Mizon, M. (1986), *IEEE Ultrasonics Symposium Proceedings*, New York: IEEE, pp. 339–342.

Berlincourt, D. A., Curran, D. R., Jaffe, H. (1964), in: W. P. Mason (Ed.), *Physical Acoustics—Principles and Methods*, Vol. 1, Part A, New York: Academic, Chap. 3, pp. 169–270.

Besson, R. J., Boy, J. J., Glotin, B., Junzaki, Y., Sinha, B., Valdois, N. (1993), *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **40**, 584–591.

Cady, W. G. (1964), *Piezoelectricity*, Vols. 1–2, New York: Dover; reprint of 1946 publication by McGraw-Hill.

EerNisse, E. P., Ward, R. W., Wiggins, R. B. (1988), *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **35**, 323–330.

Farnell, G. W. (1970), in: W. P. Mason, R. N.

Thurston (Eds.), *Physical Acoustics—Principles and Methods*, Vol. 6, New York: Academic, pp. 109–166.

Farnell, G. W., Adler, E. L. (1972), in: W. P. Mason, R. N. Thurston (Eds.), *Physical Acoustics—Principles and Methods*, Vol. 9, New York: Academic, Chap. 2, pp. 35–127.

Frerking, M. E. (1978), *Crystal Oscillator Design and Temperature Compensation*, New York: Van Nostrand Reinhold.

Frerking, M. E. (1985), in: E. A. Gerber, A. Ballato (Eds.), *Precision Frequency Control—Oscillators and Standards*, Vol. 2, New York: Academic, Chap. 9, pp. 99–111.

Frye, G. C., Martin, S. J., Cernosek, R., Pfeifer, K. B., Anderson, J. S. (1991), *IEEE Ultrasonics Symposium Proceedings*, New York: IEEE, pp. 311–316.

Fujishima, S., Merlina, J., Miyasaki, J. (1984), in: *Proceedings of the 38th Annual Symposium on Frequency Control*, Springfield, VA: National Technical Information Service, pp. 184–189.

Gagnepain, J. J., Besson, R. (1975), in: W. P. Mason, R. N. Thurston (Eds.), *Physical Acoustics—Principles and Methods*, Vol. 11, New York: Academic, Chap. 5, 245–288.

Hafner, E. (1985), in: E. A. Gerber, A. Ballato (Eds.), *Precision Frequency Control—Oscillators and Standards*, Vol. 2, Orlando: Academic, Chap. 7, pp. 1–44.

Halliburton, L. E., Martin, J. J., Koehler, D. R. (1985), in: E. A. Gerber, A. Ballato (Eds.), *Precision Frequency Control—Acoustic Resonators and Filters*, Vol. 1, Orlando: Academic, Chap. 1, pp. 1–45.

Hildebrand, F. B. (1965), *Methods of Applied Mathematics*, 2nd ed., Englewood Cliffs, NJ: Prentice-Hall, Sec. 2.14, pp. 165–172.

Holland, R., EerNisse, E. P. (1969), *Design of Resonant Piezoelectric Devices*, Cambridge, MA: MIT Press.

Hoummady, M., Hauden, D., Bastien, F. (1991), *IEEE Ultrasonics Symposium Proceedings*, New York: IEEE, pp. 345–348.

Jeffreys, H. (1931), *Cartesian Tensors*, New York: Cambridge Univ. Press.

Kosinski, J. A., Ballato, A., Lu, Y. (1993), *IEEE Ultrasonics Symposium Proceedings*, New York: IEEE, pp. 537–540.

Lee, P. C. Y., Syngellakis, S. (1975), *Proceedings of the 29th Annual Symposium on Frequency Control*, New York: IEEE, pp. 65–70.

Linville, J. G. (1988), in: T. T. Wang, J. M. Herbert, A. M. Glass (Eds.), *The Application of Ferroelectric Polymers*, New York: Chapman and Hall, Chap. 12, pp. 274–304.

Mason, W. P. (1964a), *Piezoelectric Crystals and Their Application to Ultrasonics*, Sixth Printing, Princeton, NJ: Van Nostrand.

Mason, W. P. (1964b), in: W. P. Mason (Ed.), *Physical Acoustics—Principles and Methods*, Vol. I, Part A, New York: Academic, Chap. 5, pp. 335–416.

Meeker, T. R., Meitzler, A. H. (1964), in: W. P. Mason (Ed.), *Physical Acoustics—Principles and Methods*, Vol. I, Part A, New York: Academic, Chap. 2, pp. 111–167.

Meeker, T. R. (1985), in: E. A. Gerber, A. Ballato (Eds.), *Precision Frequency Control—Acoustic Resonators and Filters*, Vol. 1, New York: Academic, Chap. 2.1, pp. 47–110; Chap. 2.2, pp. 110–119.

Meeker, T. R. (1988), in: T. T. Wang, J. M. Herbert, A. M. Glass (Eds.), *The Application of Ferroelectric Polymers*, New York: Chapman and Hall, Chap. 13, pp. 304–328.

Meitzler, A. H., O'Bryan, H. M., Jr., Tiersten, H. F. (1973), *IEEE Trans. Son. Ultrason.* **SU-20**, 233–239.

Nye, J. F. (1957), *Physical Properties of Crystals*, Part 1, London: Oxford Univ. Press.

Nye, J. F. (1960), *Physical Properties of Crystals*, Part 2, London: Oxford Univ. Press.

Parzen, B., Ballato, A. (1983), *Design of Crystal and Other Harmonic Oscillators*, New York: Wiley.

Ruppel, C. C. W., Dill, R., Fischerauer, A., Fischerauer, G., Gawlik, W., Machui, J., Muller, F., Reindl, L., Ruile, W., Scholl, G., Schropp, I., Wagner, K. C. (1993), *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **40**, 438–452.

Sheahan, D. F., Johnson, R. A. (1977), *Modern Crystal and Mechanical Filters*, IEEE Press Selected Reprint Series, New York: IEEE.

Shreve, W. R., Cross, P. S. (1985), in: E. A. Gerber, A. Ballato (Eds.), *Precision Frequency Control—Acoustic Resonators and Filters*, Vol. 1, Orlando: Academic.

Smith, W. L. (1985), in: E. A. Gerber, A. Ballato (Eds.), *Precision Frequency Control—Oscillators and Standards*, Vol. 2, Orlando: Academic, Chap. 8.1, 45–66, and Chap. 8.3, pp. 79–98.

Smythe, R. C. (1985), in: E. A. Gerber, A. Ballato (Eds.), *Precision Frequency Control—Acoustic Resonators and Filters*, Vol. 1, Orlando: Academic, Chap. 5.1, pp. 185–188, and Chap. 5.2, pp. 188–228.

Spencer, W. J. (1972), in: W. P. Mason, R. N. Thurston (Eds.), *Physical Acoustics—Principles and Methods*, Vol. 9, New York: Academic, Chap. 4, pp. 167–220.

Sykes, R. A. (1982), in: R. A. Heising (Ed.), *Quartz Crystals for Electrical Circuits—Their Design and Manufacture*, Washington, DC: Electronic Industries Association.

Tiersten, H. F. (1969), *Linear Piezoelectric Plate Vibrations*, New York: Plenum.

Tiersten, H. F., Sinha, B. K. (1978), *Proceedings of the 32nd Annual Symposium on Frequency Control*, Washington, DC: Electronics Industries Association, pp. 155–161.

Timoshenko, S., Young, D. H., Weaver, D. H., Jr. (1974), *Vibration Problems in Engineering*, 4th ed., New York: Wiley.

Wagers, R. S. (1977), in: W. P. Mason, R. N. Thurston (Eds.), *Physical Acoustics—Principles and Methods*, Vol. 13, New York: Academic, Chap. 3, pp. 49–78.

Wagers, R. S. (1985), in: E. A. Gerber, A. Ballato (Eds.), *Precision Frequency Control—Acoustic Resonators and Filters*, Vol. 1, Orlando: Academic, Chap. 5.3, pp. 230–269.

Walls, F. L., Gagnepain, J. J. (1985), in: E. A. Gerber, A. Ballato (Eds.), *Precision Frequency Control—Oscillators and Standards*, Vol. 2, Orlando: Academic, Chap. 15, pp. 287–296.

Watson, G., Staples, E., Horton, W. (1991), *IEEE Ultrasonics Symposium Proceedings*, New York: IEEE, pp. 305–309.

White, R. M. (1970), *Proc. IEEE* **58**, 1238–1276.

Further Reading

Auld, B. A. (1973), *Acoustic Fields and Waves in Solids*, Vols. 1 and 2, New York: Wiley-Interscience.

Cady, W. G. (1946), *Piezoelectricity*, New York: McGraw Hill; reprint (1964), Vols. 1–2, New York: Dover.

Gerber, E. A., Ballato, A. (Eds.), *Precision Frequency Control—Oscillators and Standards*, Vols. 1–2, Orlando: Academic.

Heising, R. A. (Ed.) (1946), *Quartz Crystals for Electrical Circuits: Their Design and Manufacture*, New York: Van Nostrand; reprint (1982), Washington, DC: Electronic Industries Association.

IEEE Ultrasonics Symposium Proceedings, New York: IEEE.

Love, A. E. H. (1927), *A Treatise on the Mathematical Theory of Elasticity*, 4th ed., London: Macmillan; reprint (1944), New York: Dover.

Mason, W. P. (1948), *Electromechanical Transducers and Wave Filters*, 2nd ed., Princeton, NJ: Van Nostrand.

Mason, W. P. (1964), *Piezoelectric Crystals and Their Application to Ultrasonics*, Princeton, NJ: Van Nostrand.

Proceedings of the Annual Symposium on Frequency Control, Springfield, VA: National Technical Information Service.

Lord Rayleigh (1877), *The Theory of Sound*, 2nd ed., Vol. 1, London: Macmillan; reprint (1945), New York: Dover.

Lord Rayleigh (1878), *The Theory of Sound*, 2nd ed., Vol. 2, London: Macmillan; reprint (1945), New York: Dover.

Tiersten, H. F. (1969), *Linear Piezoelectric Plate Vibrations*, New York: Plenum.

PLASMA ETCHING

Eray S. Aydil, *Department of Chemical Engineering, University of California, Santa Barbara, Santa Barbara, California, U.S.A.*

Introduction 171
1. **Plasmas and Plasma Etching** 172
1.1 Reasons for Plasma Etching and Applications 172
1.2 History of Plasma Etching 173
1.3 Plasma Physics and Chemistry 173
1.4 Key Steps in Plasma Etching 177
1.5 Ion-Assisted Etching 177
2. **Plasma Etching Considerations** 178
2.1 Process Parameters and Reactor Choice 178
2.2 Etch Rate and Selectivity 180
2.3 Uniformity 180
2.4 Profile Control 181
3. **Plasma Etching Reactors** 182
3.1 Barrel Reactor 182
3.2 (Parallel Plate) Diode Reactor 183
3.3 Magnetically Enhanced Reactive Ion Etching Reactor (MERIE) 184
3.4 Hexode Reactor 184
3.5 Downstream (Remote) Plasma Reactor 184
3.6 Triode Reactor 185
3.7 High-Density Plasma Reactors 185
3.7.1 Electron Cyclotron Resonance Plasma Reactor 186
3.7.2 Helicon Plasma Reactors 187
3.7.3 Helical Resonators 187
3.7.4 Inductively Coupled Plasma (ICP) Reactors—Radio-Frequency Induction (RFI) or Transformerly Coupled Plasma (TCP) Reactors 188
4. **Modeling of Plasma Etching Reactors and Plasmas** 188
4.1 Electron-Kinetics Modeling 190
4.2 Continuum Glow-Discharge Models 191
4.3 Neutral Transport and Reaction Models 191
4.4 Sheath Models 191
4.5 Particle-in-Cell and Monte Carlo Simulations 191
5. **Specific Plasma Etching Applications** 192
6. **Plasma Process Monitoring: Plasma Diagnostics** 192
6.1 Optical Emission and Optical Diagnostic Techniques 193
6.2 Mass Spectroscopy 194
6.3 Langmuir Probes 194
6.4 *In Situ* Film-Thickness Measurements: Laser Interferometry and Reflectometry 194
6.5 Surface Diagnostic Techniques 194
7. **SUMMARY** 195
Glossary 195
Works Cited 195
Further Reading 197

INTRODUCTION

A plasma is an electrically energized gas that consists of electrons, ions, neutral atoms, molecules, and molecular fragments. Etching is selective removal of material from a surface by chemical reactions and/or physical sputtering. Plasma etching is a process wherein chemically reactive species and ions produced in a gas discharge (plasma) are used to remove selectively a film from a substrate or portions of such a film or substrate. Plasma etching has also come to be known as dry etching because the chemically reactive medium is a gas instead of a liquid such as an acid or a base solution. Plasma etching is widely used in manufacturing of electronic and optoelectronic devices and integrated

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

circuits, which are made layer by layer through patterning of thin films on the surface of a semiconductor substrate such as a Si or a GaAs wafer (Manos and Flamm, 1989; Gottscho, 1993).

A plasma (or gas discharge) is obtained by applying an electromagnetic field to a volume of gas under vacuum. Although extremely high temperatures and gases at atmospheric pressure can be used to produce a plasma, these methods do not find applications in plasma etching. Electrons in the plasma are accelerated by an electric field induced by the electromagnetic excitation and gain energy from the applied field. These energetic electrons collide with the neutral-gas species and dissociate them into reactive radicals and molecular fragments. Plasma etching makes use of the chemical reactivity of the various molecular fragments produced in the plasma combined with energetic action of the electrons and the ions to etch various materials. This combined action of charged particles and reactive neutral species provides the means to etch very small features on thin films and enables the integrated-circuit (IC) manufacturers to pack more solid-state electronic devices into smaller areas. This article is intended to provide an overview of plasma etching and to serve as an introduction to the field, as well as a reference to some more advanced topics.

1. PLASMAS AND PLASMA ETCHING

1.1 Reasons for Plasma Etching and Applications

Plasma etching is used mainly for two purposes:

1. to transfer a pattern from a mask into an underlying layer of material by selectively removing the unmasked portions of that layer, and
2. to remove a layer completely from on top of a substrate.

The use and advantages of plasma etching in patterning thin films are illustrated in Fig. 1, which shows a schematic of a film to be etched on top of a substrate. The film is unmasked in places where it is to be removed. A mask is a material that is chemically and physically resistant to the etchant in order to protect the desired parts of the film. Common mask materials include polymers (photoresist), oxides (e.g., SiO_2), nitrides (Si_3N_4), and metals. The unmasked part of the film can be removed by exposing the surface to a chemically reactive medium such as an acid solution or a plasma. Wet etching (i.e., removal of the film by using an acidic solution) and plasma etching are the two major techniques to transfer the pattern defined by the mask to the underlying film. At present, plasma etching is the method of choice for etching thin films of electronic materials in the microelectronics industry.

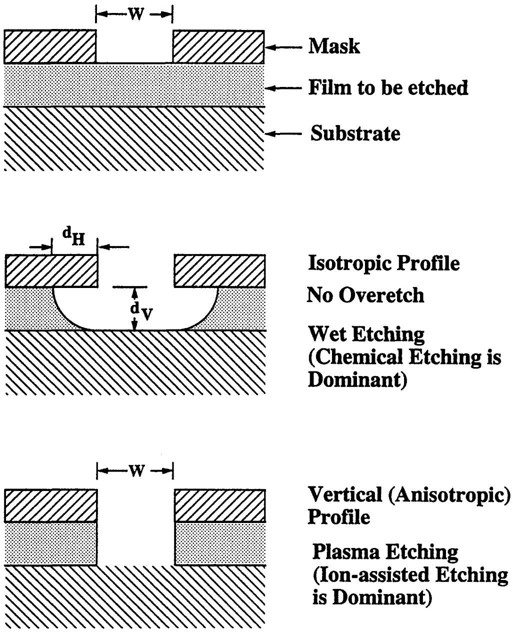

FIG. 1. (a) Cross section of film to be etched, with mask. (b) Cross section of film etched using wet etching, showing the profile resulting from etching rates equal in vertical and horizontal directions. (c) Cross section of film etched using plasma etching, for which the etch rate in the horizontal direction is negligible compared with the rate in the vertical direction.

One of the major reasons for preferring plasma (dry) etching over wet etching is depicted in Figs. 1(b) and 1(c), which compare the resulting patterns in the films when wet and plasma etching techniques are used, respectively. When wet etching techniques are used, the etching reactions proceed at equal rates in vertical and horizontal directions and the profile shown in Fig. 1(b) results.

Since the film is removed isotropically (i.e., independent of direction), the resulting profile has the characteristic quarter-circle shape. Isotropic etching in this manner results in an etched feature that is larger than the original feature in the mask. This undesirable and unintentional enlargement of the original feature size in transferring the pattern from the mask to the underlying film is called linewidth loss. The horizontal ingression as a result of etching under the mask is called mask undercutting. In order to pack more or smaller features in same area, linewidth loss or undercutting should be minimized. In fact, in order to replicate the mask pattern onto the film exactly and to reduce linewidth loss, the etch profile depicted in Fig. 1(c) is desired. The etched feature in Fig. 1(c) has straight vertical sidewalls and is the preferred profile in microelectronics manufacturing.

The etch profile shown in Fig. 1(c) can be obtained when the etch rate in the horizontal direction is zero or negligible compared to the etch rate in the vertical direction. Using a plasma provides the means to enhance the etch rate in the vertical direction relative to the etch rate in the horizontal direction thereby enabling one to obtain etch profiles like the one shown in Fig. 1(c). This type of etch profile is called an anisotropic etch profile and the etching method that produces such profiles is called anisotropic etching. The directionality in etching arises from anisotropic transport of ions from the plasma to the wafer surface. The reasons behind directional ion transport from the plasma to the surface are discussed later. Plasma etching has become the dominant technique for patterning thin films in microelectronics manufacturing because of its ability to etch anisotropically in a dry environment. This ability enables the integrated-circuit manufacturers to pack more solid-state devices such as memory elements into smaller areas, which reduces the unit production cost and increases processor speeds and memory capacity.

1.2 History of Plasma Etching

Glow discharges and reactions in plasmas have been studied for many years. It is difficult to pinpoint the first use of plasmas for etching. However, credit is generally given to Irving (1968), who suggested that an oxygen plasma can be used to remove photoresist films that are used as masks in production of microelectronic devices. Shortly thereafter fluorine-containing plasmas were used to etch silicon films (Coburn and Winters, 1983). Positive results from these experiments combined with the advantages of plasma etching over wet chemical etching methods stimulated the commercialization of plasma etching reactors. As of the early 1990s, 30% of the integrated-circuit manufacturing equipment had become plasma based (Proud and Gottscho, 1991).

Advances in the IC manufacturing technology are a direct result of advances in plasma etching technology. Features with sizes (e.g., the width W of the trench shown in Fig. 1) 5–10 μm could be etched during the initial years of development. It is now possible to etch 0.5–1-μm size features routinely with plasmas, and the goal is to reduce this dimension below 0.25 μm by the turn of the century. The submicron electronic device fabrication was made possible by plasma etching. Plasma etching played and continues to play an important role in the electronics and communication revolution.

1.3 Plasma Physics and Chemistry

A plasma is a partially ionized gas produced by applying an electric field to a volume of gas at low pressure (10^{-4}–100 Torr). Thus, it is composed of electrons, ions, and a variety of neutral species. Plasmas used for thin-film etching are low-pressure glow discharges excited usually by radio-frequency (0.1–100 MHz) or microwave fields (usually 2.45 GHz). The gas pressure and temperature in typical plasma etching reactors range from 10^{-4} to 10 Torr and 300 to 500 K, respectively. Although temperatures as high as 800 K have been reported, plasma etching is not a high-temperature process, and for most etching processes the gas temperature remains near ambient.

The electrons are accelerated by the applied field and can only lose energy when they collide with the neutral gas species elastically or inelastically. Elastic collisions preserve the internal energy of the colliding partner. Inelastic collisions, on the other hand, can leave the colliding partner in an excited state. Electrons lose only a small fraction (10^{-4}) of their energy during elastic collisions because of the three orders of magnitude mass difference between the electrons and typical neutral species. However, electrons can lose energy in inelastic collisions. Compared to elastic collisions, inelastic collisions occur relatively infrequently at low energies and occur in sufficient frequency only when the electron energies are high (more than a few electron volts). In addition, at low pressures, the mean free path of the electrons is large. Because of the low energy-loss rate and long mean free paths, the electrons can reach high energies compared to the more massive ions and neutral species. Typically, the average electron energies are between 1 and 10 eV, which correspond to temperatures of 10^4–10^5 K! On the other hand, ion and neutral temperatures are closer to the ambient temperature. Thus, the glow discharge is a highly nonequilibrium environment characterized by species that can have drastically different velocity-distribution functions and temperatures. The electron and ion densities range from 10^8 to 10^{13} cm^{-3} depending on the excitation method employed, power, pressure, etc. However, the ratio of ion or electron density to the neutral-species density remains in the range of 10^{-7}–10^{-1}; the upper ratio is reached only in high-density discharges, and for most processes the discharge consists of relatively few electrons and ions in a sea of neutral species. The basic physics of glow-discharge processes are discussed by Chapman (1980) and by Lieberman and Lichtenberg (1994) in detail.

The characteristics listed above imbue glow discharges with unique and useful properties. Energetic electrons are capable of breaking chemical bonds via collisions with gas molecules, thereby overcoming kinetic limitations or energy barriers that exist in thermal dissociation processes. In this way, very reactive chemical species can be produced by dissociating the process gasses even at low temperatures (e.g., at ambient temperature). The low process temperature allows etching of temperature-sensitive materials in a plasma environment.

Since the first use of a gas discharge for etching, different types of plasma reactors have been utilized (Manos and Flamm, 1989; Flamm, 1991; Lieberman and Gottscho, 1993; Mucha *et al.*, 1994; Graves, 1994). These reactors differ in method of excitation, reactor geometry, or other operational parameters (e.g., power, pressure, and so forth). Today, a wide variety of plasma etching reactors exists. Despite the variations in reactor configuration and operation, the key physicochemical phenomena that enable plasma etching remain the same. The plasma performs two basic functions. First, chemically reactive species are generated by electron-impact collisions. Second, the discharge is a source of energetic particles—ions, electrons, and photons—that bombard the surfaces immersed in the plasma and alter the surface chemistry during etching (Hess, 1990). The bombardment of the film or substrate surface by energetic ions, primarily in the direction normal to the surface, increases the etch rate in the vertical direction and is responsible for anisotropic etching.

In the following discussion of plasma physics and chemistry, the parallel-plate single-wafer plasma etching reactor shown in Fig. 2 is taken as an example to describe the key phenomena that make plasmas useful for etching. In the parallel-plate plasma reactor (also known as the diode reactor), the plasma is sustained between two electrodes, as shown in Fig. 2, by applying radio-frequency power to either one or both of the electrodes. The process gases are usually introduced uniformly through a showerhead, which also serves as the upper electrode. The gases are pumped out from the periphery of the electrodes. This type of reactor is discussed in detail in Sec. 3. Table 1 lists typical operating conditions and dimensions of a parallel-plate plasma reactor.

Electrons are the heart of the plasma and are responsible for ionizing and dissociating the process gasses. Free radicals, metastable species, ions, and electrons themselves are generated by homogeneous gas-phase collisions such as the ones shown in Fig. 2 and Table 2. Depending upon the electron energy, an electron–neutral collision can result in a number of different reactions. Table 2 gives a representative but by no means a complete list of possible electron-impact reactions in a plasma; in this list chlorine is taken as an example but analogous reactions can be written for other species as well. Of these reactions, dissociation is one of the most important reactions for etching. In the

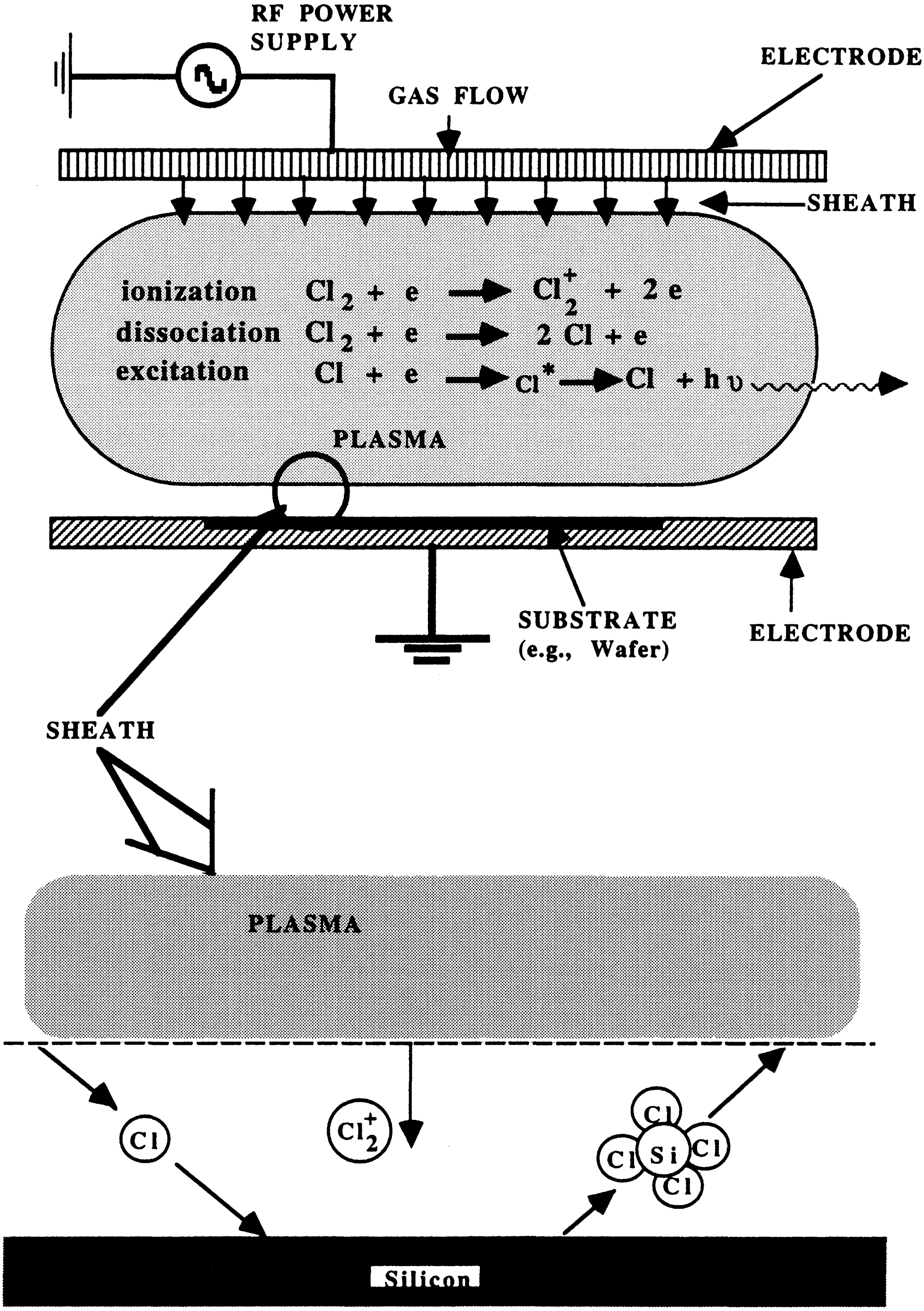
RF POWER SUPPLY
GAS FLOW
ELECTRODE
SHEATH
ionization $Cl_2 + e \rightarrow Cl_2^+ + 2e$
dissociation $Cl_2 + e \rightarrow 2Cl + e$
excitation $Cl + e \rightarrow Cl^* \rightarrow Cl + h\nu$
PLASMA
SUBSTRATE (e.g., Wafer)
ELECTRODE
SHEATH
PLASMA
Cl
Cl_2^+
Cl
Cl Si Cl
Cl
Silicon

FIG. 2. Schematic of a parallel-plate plasma reactor showing the sheath and important phenomena occurring in plasmas of chemically reactive gas. Chlorine plasma etching of silicon is depicted as an example. The plasma is sustained between two electrodes by application of radio-frequency power to either or both of the electrodes. Gases are usually introduced through the top electrode using a shower head and pumped out from the periphery of the electrodes.

Table 1. Typical operating parameters and properties of a rf parallel-plate (diode) plasma etching reactor and a high-density plasma reactor.

Parameter	rf parallel-plate reactor	High-density plasma reactor
Pressure	0.001–10 Torr	0.0001–0.1 Torr
Power	50–1 000 W	100–5 000 W
Frequency	0.1–100 MHz	0.1–20 MHz or 2.45 GHz
Gas flow rate	10–3 000 sccm	10–200 sccm
Electron temperature	1–10 eV	1–10 eV
Plasma density	10^8–10^{11} cm^{-3}	10^{10}–10^{13} cm^{-3}
Fractional ionization	10^{-7}–10^{-4}	10^{-4}–10^{-1}
Ion bombarding energy	50–1000 eV	10–500 eV
Ion bombarding flux	0.01–5 mA/cm^2	1–50 mA/cm^2
Magnetic field	0 (except in magnetically enhanced reactive ion etching, MERIE)	0–1200 G
Plasma volume (for 8-in.-diameter wafers)	300–5000 cm^{-3}	1000–10 000 cm^{-3}

case of Cl_2, this reaction produces the etchant, reactive Cl radical. Via this reaction, Cl_2 gas can be dissociated close to 100% in a plasma at room temperature. Without the energetic electrons, even 10% dissociation would require a gas temperature above 1000 K. The excitation reaction and subsequent deexcitation of the excited species creates photons. Thus, electromagnetic radiation in the visible, ultraviolet, infrared, or even the x-ray region is emitted from the excited molecules and ions as they relax to a lower energy state. This is what makes a plasma glow. The rates of the reactions in Table 2 depend on the electron energy-distribution function, the density of the electrons, the density of the colliding partner in the reactor, and the collision cross section for the reaction (Mucha *et al.*, 1994; Lieberman and Lichtenberg, 1994).

Table 2. A list of representative electron impact reactions in a Cl_2 plasma.

Electron-impact ionization	$e^- + Cl_2 \rightarrow 2e^- + Cl_2^+$
Dissociative ionization	$e^- + Cl_2 \rightarrow 2e^- + Cl^+ + Cl$ $e^- + Cl_2 \rightarrow e^- + Cl^- + Cl^+$
Dissociative attachment	$e^- + Cl_2 \rightarrow Cl^- + Cl$
Electron-impact dissociation	$e^- + Cl_2 \rightarrow e^- + 2Cl$
Electron-impact excitation (rotational, vibrational, electronic)	$e^- + Cl_2 \rightarrow e^- + Cl_2^*$

The energy and trajectory of the ions bombarding the substrate surface are determined by the potential distribution established in the plasma, particularly near the substrate surface. Since electrons are lighter and more mobile than heavier ions, it is expected that they would be lost to the walls of the plasma reactor in larger numbers than ions. This exodus of electrons would in turn deplete the plasma of electrons, leaving a space charge behind. In fact, this is exactly what happens during the first moments of plasma initiation. However, within a short period of time electrons are depleted next to the walls and the electrodes and a space-charge layer forms next to the surfaces in contact with the plasma. In order to stop the electron depletion, an electric field is established to repel the electrons away from the walls. At steady state, the ion flux and electron flux equilibrate to satisfy the overall current continuity and charge balance. The majority of the electrons are repelled and excluded from the region near the walls by the high electric field created by the space charge. Thus, the plasma is separated from the walls of the vessel and the electrodes (in fact, any surface that is immersed in a plasma) by a space-charge region called the sheath. The sheath appears relatively dark as a result of lack of electrons and hence electronic excitation in this region. One consequence of this space-charge formation is that the potential distribution in the discharge is such that the plasma is at a higher potential than all surfaces it is in contact with. In other words, an electric field near the walls always points toward the electrodes and repels the electrons from the sheath. The same electric field that repels the electrons away from the walls accelerates the oppositely charged ions toward surfaces. The ions are accelerated by the sheath electric field and bombard the film and substrate surfaces predominantly at normal incidence. This bombardment is responsible for enhancing the etch rate in the vertical direction and for inducing anisotropic etching, which produces etch profiles such as the one shown in Fig. 1(c). The etch rate with simultaneous bombardment with ions and reactive neutrals is higher, sometimes by more than an order of magnitude, than the sum of the etch rates with ions and reactive neutrals alone (Coburn and Winters, 1979). This synergistic enhancement and resulting etching is called ion-assisted etching.

1.4 Key Steps in Plasma Etching

The mechanism of plasma etching can be broken down into six steps, as depicted in Fig. 3 (Mucha *et al.*, 1994). First, reactive species are generated by electron-impact reactions in the plasma. This step is crucial in etching because many of the gases do not react spontaneously with the film to be etched without dissociation. The reactive etchant species are transported to the surface by gas flow and diffusion and adsorb onto a surface site on the film that is being etched. In general, reactive radicals (Cl) have higher adsorption and reaction probabilities than the parent molecule (Cl_2) even at low surface temperature. Details of what happens on the surface are often unclear and provide an active area for research. Ion-assisted surface reactions take place and volatile reaction products are formed on the surface. Product formation is followed by desorption, another crucial step in the etch process. A free radical can react rapidly with a solid surface, but unless the product species has a appreciable vapor pressure so that desorption occurs, no etching takes place. Desorbed products diffuse into the bulk and are pumped out of the reactor.

1.5 Ion-Assisted Etching

Depending on the reactor and operating conditions the energy of the ions incident onto the surface can range from a few to

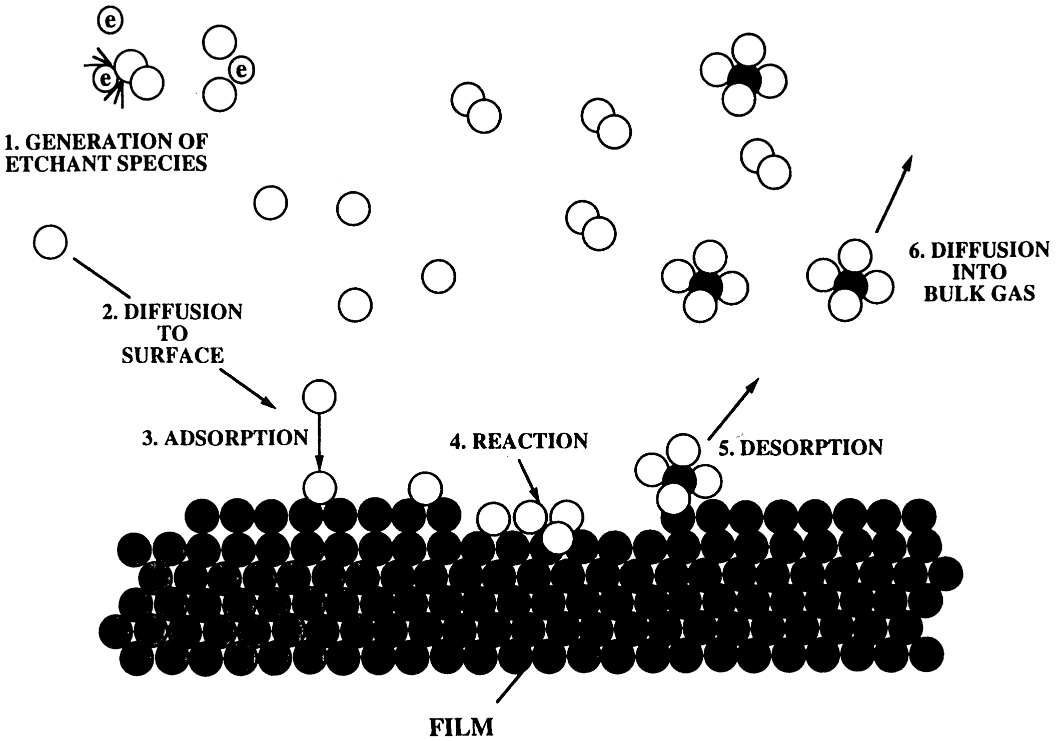

FIG. 3. General mechanism of plasma etching (reprinted from Mucha *et al.*, 1994, with permission).

hundreds of electron volts, more than enough to break bonds and effect surface reactions. The enhancement of etch rate due to particle bombardment can be envisaged by referring to Figs. 3 and 4. Steps 3–5 are heterogeneous reactions with kinetics that depend on the temperature and the nature of the surface. Increasing the temperature is one way of increasing the energy of the reactants and surface species to facilitate reactions. Particle bombardment is another way that energy can be imparted to the surface. Particle bombardment can break chemical bonds on the surface, thereby creating reactive sites for adsorption or creating reactive surface species that can react to form volatile products. Much like the electrons in the gas phase, ion bombardment can cause new reaction pathways that the surface chemistry can evolve through. Ion bombardment can also assist product desorption, facilitating step 5 in Fig. 3 in cases where the etching product is not very volatile. Which of these steps is responsible for enhanced etch rates in the vertical direction is not clear and most likely depends on the specific case.

Ion-assisted etching was demonstrated by Coburn and Winters in 1979 in a classic experiment. In their experiment, Coburn and Winters observed that Si could be etched chemically with XeF_2 at $\approx$5 Å/min and could be physically sputtered by Ar^+ at $\approx$2 Å/min. However, when both XeF_2 and Ar^+ were used, an etch rate of 55 Å/min, an order of magnitude higher than both the chemical and physical etching rates, was observed. This experiment demonstrated the existence of a synergy between chemical etching and ion bombardment. It turned out that these effects are very general in plasma etching and are encountered in etching of other materials with other gases.

Two simple mechanisms for ion-assisted anisotropic etching are depicted in Fig. 4. In the ion-induced damage mechanism shown in Fig. 4(a), the ion bombardment breaks the bonds on the crystal surface and leaves the film more reactive and more accessible to the chemical etchant. For example, this could happen via formation of a reactive layer that is mixed and removed by ion bombardment. The sidewalls, on the other hand, do not receive ion bombardment and etching is unaccelerated. Consequently, material removal proceeds faster in the ion flux direction, normal to the film surface, resulting in anisotropic etching. In the surface inhibitor mechanism, the second scenario shown in Fig. 4(b), the sidewalls are protected from etching by an inhibitor. This protecting layer can be a thin layer of polymer film or oxide, which can readily form in a plasma environment. The inhibitor may also form at the bottom surface but will be removed quickly by ion bombardment, exposing the bare surface to etchants that in turn can remove the film.

2. PLASMA ETCHING CONSIDERATIONS

The general goal of any plasma etching process is to achieve high etch rate, high degree of anisotropy (vertical sidewalls), good uniformity on a single wafer as well as from wafer to wafer, and selective etching of the film relative to the mask and the underlying film or substrate (Sawin, 1994) with minimal radiation damage to the devices on the wafer. Unfortunately, many of these goals are mutually exclusive. For example, the ion bombardment of the surface that provides the anisotropic etching also causes damage. The relative importance of these goals and objectives depends on the application. Given the numerous combinations of process gases, process conditions, and plasma reactor configurations, the difficulty of designing, developing, and optimizing a process to achieve a given set of objectives is the major limitation of plasma etching processes. In this section, important factors and considerations that affect the design and development of plasma processes are described briefly.

2.1 Process Parameters and Reactor Choice

Plasma reactors are notoriously complex systems because of strong coupling of physical and chemical phenomena occurring in the gas phase and on surfaces. At present, it is impossible to predict theoretically the optimum operating parameters for a given plasma etching application. For this reason, reactor and process parameters are chosen empirically, after laborious trial-and-error experimentation and optimization.

One can distinguish between two different

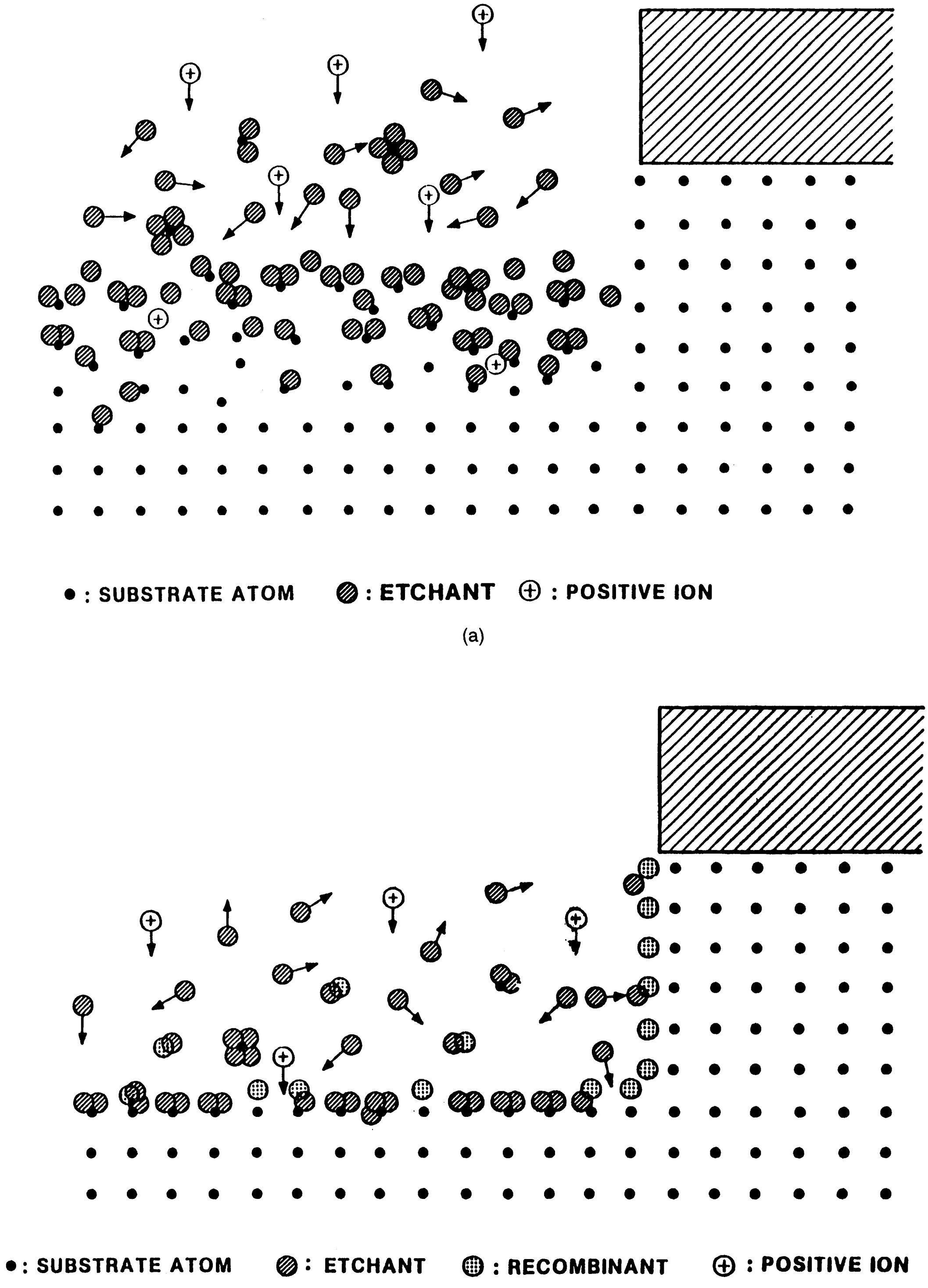
● : SUBSTRATE ATOM
▨ : ETCHANT
⊕ : POSITIVE ION
● : SUBSTRATE ATOM
▨ : ETCHANT
▦ : RECOMBINANT
⊕ : POSITIVE ION

(a)

(b)

FIG. 4. (a) Surface damage and (b) inhibitor mechanisms for ion-assisted etching (adapted from Flamm and Donnelly, 1981a).

kinds of parameters in plasma etching: externally controlled variables and internal plasma variables. The externally controlled variables are those that can be manipulated directly by turning a knob or pushing a button. These parameters include power, pressure, excitation frequency, flow rate, and gas composition. These operating parameters dictate the outcome of a plasma process, the "figures of merit," through their effect on the internal properties of the plasma. These internal properties include the electron energy distribution function, reactive species concentrations, ion flux, ion-bombardment energy and angular distribution, and their spatial variations. The difficulty in controlling and developing plasma etching processes arises from the fact that these key internal properties cannot be controlled independently but are determined by the externally set operating parameters. Often, the relationship between the external parameters and the internal properties is not known. Furthermore, changing one external parameter may affect more than one internal property. For example, increasing the power may result in an increase of ion bombardment flux and energy, as well as etchant concentration.

In summary, there are no simple guidelines for chosing process parameters and designing reactors. Usually, choices are made on the basis of scientific as well as economic reasons, which are different for each application. Thus, each individual process should be considered on the basis of its own objectives. At present, process complexity and the large number of parameters that affect the process are the most serious impediments to routine use of plasma etching.

2.2 Etch Rate and Selectivity

Etch rates determine the processing time in the plasma, and the throughput. Process economics and the thicknesses of the films to be etched dictate the magnitude of an acceptable etch rate. Typically, etch rates should be and are on the order of several thousand angstroms to a few microns per minute for a manufacturing process. Wafer throughput rates on the order of 20 wafers/h are typical. The ratio of the etch rates for two different materials immersed in the same plasma determines the selectivity of removing one film without significantly etching the mask or the selectivity of removing one film without etching an underlying film or the substrate. Selectivities in excess of 1:10 are generally desirable. High etch rates and selectivity can be achieved by judicious selection of feed gases to a plasma reactor. For a given material, the gas selection is generally made by trial and error and by using guidelines based on previous experience with similar materials. Numerous materials have been etched using plasmas. Table 3 is a list of materials frequently encountered in IC manufacturing and the etchants used to pattern these films.

2.3 Uniformity

Uniformity refers to uniformity of etching not only from wafer to wafer but also over a single wafer as well. For example, if the etch rate is higher at one position on the wafer than at another position, the film will be removed in that location before complete removal of the film in locations where the etch rate was not as high. However, the wafer must remain in the reactor to allow for complete removal of the film everywhere. This results in overexposure of some areas to the plasma. This is called overetching. Overetching can result in undercutting and etching or damage of the underlying films and structures. Uniformity on a single wafer is particularly important because uniform etching over a wafer is required in order to minimize the necessity to overetch.

Nonuniform etching can be on a macroscopic or a microscopic scale. Macroscopic uniformities are on the length scale of the wafer dimensions and are caused by radial gradients in the etchant concentration (Economou *et al.*, 1991; Aydil and Economou, 1993), ion-bombardment flux and energy, and/or wafer temperature. Uniformity is usually characterized by reporting the maximum percentage deviation of the etch rate from an average etch rate or from the etch rate at the center. Today's plasma etching processes require a uniformity that is better than ±5%.

Etchant concentration gradients can exist at the wafer edge where two surfaces with different reactivities meet. The etchant is depleted by etching reactions and the etchant concentration is lowered over the wafer with respect to the surrounding electrode. This

Table 3. List of process gases used in plasma etching of materials encountered in microelectronic-device fabricatiom.

Material	Process gases	Review reference
Si, polysilicon	CF_4, CF_4/O_2, C_2F_6/O_2, C_3F_8/O_2, F_2, Cl_2, Br_2, CCl_4, CF_2Cl_2, CF_3Cl, Cl_2/C_2F_6, CF_3Br, SF_6/O_2, NF_3, ClF_3, SiF_4/O_2, Cl_2/O_2, Cl_2/N_2, and various mixtures of these gases with each other and noble gases such as Ar and He	a, b
SiO_2	CF_4/H_2, C_2F_6, C_3F_8, CHF_3, and various mixtures of these gases with each other and noble gases such as Ar and He	a, b
Si_3N_4	$CF_4/O_2/CBrF_3$, CH_2F_2, CHF_3, SF_6, Cl_2/NF_3	a, b
Al	BCl_3, CCl_4, $CHCl_3$, and mixtures with Cl_2 and $SiCl_4$; BBr_3 and mixtures with Br_2, Cl_2, $Cl_2/BCl_3/CH_4$, and $Cl_2/BCl_3/N_2$	a, b
Au	CCl_2F_2 and $CClF_3$	b
Cr	Cl_2 and CCl_4, mixtures of each with O_2	b
Cu	CCl_4/Ar, $SiCl_4$, N_2, and $SiCl_4/Ar$	b
Ti	CF_4, $CClF_3$, and $CBrF_3$	b
W	CF_4, SF_6, SF_6/O_2, SF_6/Cl_2 SF_6/N_2, SF_6/O_2, CHF_3 and $CBrF_3$ mixtures with SF_6 and CHF_3/O_2	b, c
Ti, Ta, Mo and W silicides	CF_4/O_2, NF_3/CHF_3, other F- or Cl-containing plasmas, gases used in etching of Si and the metal	b, c
Photoresists and organic films	O_2	b, d
GaAs and other III-V semiconductors and alloys	Cl_2, $COCl_2$, CH_4/H_2, CCl_2F_2, CCl_4, PCl_3, HCl, and mixtures of these gases with Ar and other noble gases	e

[a]Flamm and Donnelly (1981a).
[b]Mucha *et al.* (1994).
[c]Hess (1988).
[d]Hartney *et al.* (1989).
[e]Burton *et al.* (1984).

concentration gradient will cause the etch rate to decrease from the edge of the wafer toward the center, causing the film to be removed from the edges first.

Since the surface reactions are ion-bombardment driven, the radial uniformity of ion-bombardment flux and energy also affect the uniformity of etching. In this case, the important factors that determine the uniformity are the power absorption profile and ion transport in the reactor (Stevens *et al.*, 1992; Economou *et al.*, 1991; Aydil and Economou, 1991).

Another factor that may affect macroscopic uniformity is the temperature profile across the wafer. Chemical reactions depend on the temperature exponentially, and even very minute temperature differences on the surface can lead to unacceptable uniformity. Nonuniform etching may result if the wafer is not cooled or is in poor thermal contact with the cooled electrode (Aydil and Economou, 1993).

Microscopic uniformity refers to uniformity of etching on the scale of the feature size (Gottscho *et al.*, 1992). For example, features with different shapes and sizes are know to etch at different rates. This creates a situation where the etch rate varies from hole to hole or from trench to trench even when they are next to each other (within several micrometers). This phenomenon has been called microloading, RIE (reactive ion etching) lag, and aspect-ratio–dependent etching. Generally, it is observed that smaller features etch at a slower rate than larger ones as a result of transport limitations for ions and neutrals. There are several possible mechanisms that have been put forth to explain microscopic nonuniformities. Generally, microscopic nonuniformities result from differences in ion and neutral transport rates within the feature. The topic has been reviewed by Gottscho *et al.* (1992).

2.4 Profile Control

The goal of pattern-etching processes is to transfer the exact replica of the features on the mask to the underlying film. The most useful metric for profile control is the ratio of lateral undercutting, d_H, to vertical etching, d_V, as depicted in Fig. 1. The ratio d_H/d_V is inversely related to the quality of replication. In anisotropic etching, $d_H/d_V = 0$, and high-quality replication of the mask pattern is obtained. On the other hand, for isotropic etching, $d_H/d_V = 1$, which results in linewidth loss and poor pattern replication. Anisotropic etching is desired for high-density, fine line patterning.

It should be clear from the discussion in Sec. 1 that the degree of anisotropy depends critically on the ion energy and angle distribution. For anisotropic etching, a very narrow angular distribution about the normal to the surface is desired. A narrow angular distribution is obtained when the ions are accelerated through the sheath without undergoing collisions that will randomize their velocity distribution function, and when the velocity acquired in the sheath is much larger than the random velocity that the ions possess before entering the sheath (Liu *et al.*, 1990). Since the mean free path of the ions increases with decreasing pressure, anisotropic etching is usually obtained at low pressures (<50 mTorr). The desired energy distribution is less clear. The ions must be energetic enough to cause anisotropic etching but not so energetic that they damage the substrate.

It has been shown that deliberate sidewall inhibition can be used as a strategy to achieve anisotropic etching (Donnelly and Flamm, 1981). This strategy relies on the sidewall-inhibition mechanism described previously. Polymer-forming halocarbons are added to the plasma to form protecting layers on the sidewalls. Ion bombardment maintains a cleared bottom surface where the etching can take place. Sidewall passivation layers can be formed in plasmas even without the polymer-forming species. For example, a silicon dioxide inhibitor layer can be formed on the sidewalls by adding O_2 to the etching mixture (e.g., BCl_3/Cl_2 or BCl_3/HCl) for etching silicon.

Another strategy for controlling anisotropy is to mix two etchant gases where one etchant is known to etch anisotropically and the other isotropically. Control of anisotropy by changing the composition of the plasma has been demonstrated in this manner by mixing ClF_3 with Cl_2 (Flamm *et al.*, 1982).

Other major factors in profile control in etching are the mask material and the rate of mask erosion. The masks shown in Fig. 1 represent an ideal case where the mask is not eroded by the etchant or the ion bombardment. Inevitably, mask erosion, particularly near the edges, can result in linewidth loss.

Finally, it should be noted that vertical walls are not always desirable. For example, if the etched pattern is going to be filled with another material, using a deposition method, a slightly tapered profile may be desired to reduce shadowing during deposition.

3. PLASMA ETCHING REACTORS

A general plasma etching system has several basic components: a vacuum chamber and pumping system to maintain the desired low process pressures, a gas handling system to deliver, meter, and control the flow of reactant gases, and a power supply to create a glow discharge. Pressure in the reactor is measured and controlled using a pressure transducer and a throttling valve. The electrode or the platen that the wafer is placed on is usually cooled to prevent it from heating by ion bombardment. In addition to basic equipment, commercial plasma reactors include automated wafer and process handling systems. Some reactors also employ plasma diagnostic equipment to measure etching rates and etchant concentrations. Detailed assembly of a plasma etching system is subject to variations depending on the desired operating parameters and application. On account of the large variety of existing plasma etching reactor configurations, a simple categorization is not possible. The objective in the following subsections and the associated figures is to describe briefly the most common reactor configurations and to emphasize differences and similarities among them. References are given for more detailed information. Summaries of different reactor configurations are also given in Mucha *et al.* (1994), Lieberman and Lichtenberg (1994), and Lieberman and Gottscho (1993).

3.1 Barrel Reactor

The barrel reactor is the earliest plasma apparatus constructed for etching. In this reactor, several wafers are placed on a "boat,"

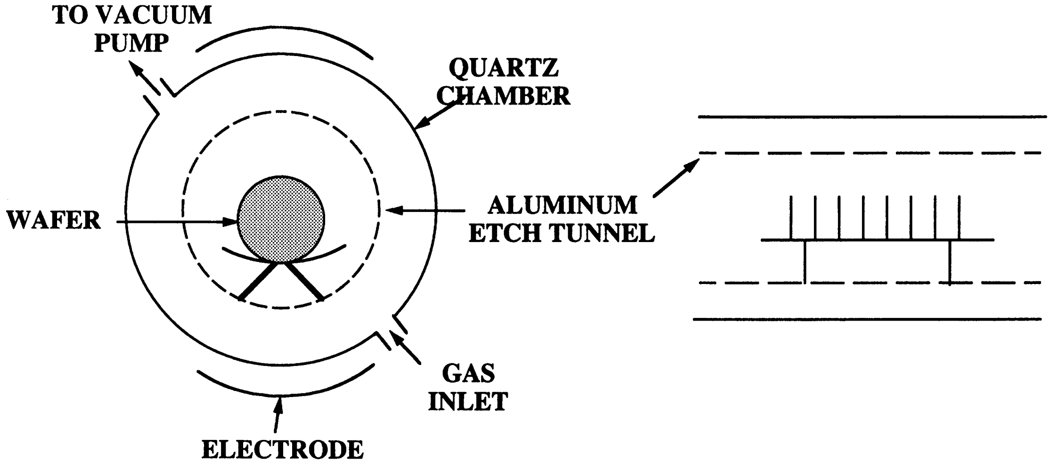

FIG. 5. Barrel reactor (reprinted with permission from Mucha *et al.*, 1994).

which in turn is placed in a cylindrical evacuated chamber. Usually the chamber is made out of quartz and a perforated metal shield is placed around the wafers to protect them from radiation damage, heating, and energetic ion bombardment. The plasma is maintained in the annulus by applying rf power to coils or electrodes placed outside the chamber as shown in Fig. 5. The wafers receive little or no ion bombardment because of the shield: only reactive radicals reach the wafer, resulting in purely chemical etching and isotropic etching profiles. For this reason, barrel systems are used in noncritical etching steps where undercutting can be tolerated, or in applications such as photoresist mask removal where anisotropic etching is not needed.

3.2 (Parallel Plate) Diode Reactor

The diode reactor, already mentioned in Sec. 1, is one of the most widely used plasma reactors. The plasma is created by applying a rf field between two electrodes. Usually 13.56 MHz is used in order not to affect communications. However, frequencies ranging from 100 kHz to 100 MHz have been employed. The electrodes can be large to accommodate many wafers or small to process a single wafer (Mathad, 1985). At present, most parallel-plate reactors are designed to etch one wafer at a time in order to provide better wafer-to-wafer control. As discussed earlier, this geometry results in ion bombardment of the substrate. Etching is by both physical (ion bombardment) and chemical (reactive radicals) processes, which makes it possible to etch high–aspect-ratio (ratio of depth to width, d_V/W) features such as holes and trenches.

The diode reactor is usually operated in two modes. In one configuration, the wafers are placed on the grounded electrode while the opposite electrode is powered. The processes using this configuration are termed "plasma etching." When substrates are placed on the powered electrode, a reactive ion etching (RIE) mode results (Coburn, 1982). In the RIE configuration (see Fig. 6), the chamber walls are grounded and act as the counter electrode. Although no exact de-

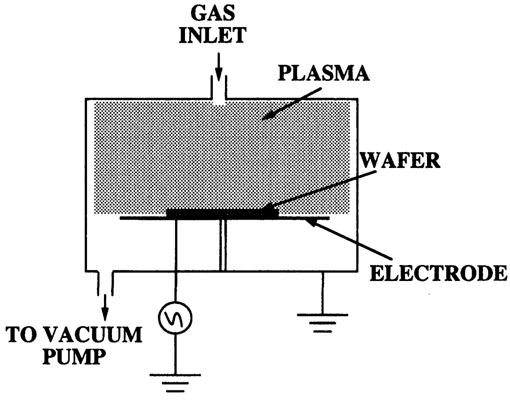

FIG. 6. Diode reactor in RIE configuration (reprinted with permission from Mucha *et al.*, 1994).

marcation in pressure exists, RIE is typically conducted at pressures less than 100 mTorr (typically 1–100 mTorr) and plasma etching is carried out at pressures above 100 mTorr (typically 0.1–5 Torr). The RIE arrangement exposes the wafers to higher-energy ion bombardment than in the plasma etching mode because higher potentials are established next to the powered electrode when compared to the sheath next to the grounded electrode. Another reason for higher-energy ion bombardment in RIE is the lower operating pressure. Ion bombardment in RIE mode is more energetic and more directional (narrower incident angle distribution about the normal to the surface) than in plasma etching because the ions do not collide with the neutrals and lose energy and directionality as they are accelerated in the sheath.

3.3 Magnetically Enhanced Reactive Ion Etching Reactor (MERIE)

The magnetically enhanced reactive ion etching reactor (see Fig. 7) is one of the variations of the RIE configuration (Leahy and Kaganowicz, 1987). The MERIE reactor is an RIE reactor that employs a static magnetic field applied parallel to the powered electrode. The purpose of the magnetic field is to confine the electrons and increase the plasma charge density by reducing the electron loss rate.

3.4 Hexode Reactor

The hexode reactor is another variation of the RIE reactor. In this reactor, the wafers are mounted around the periphery of a hexagonal electrode as shown in Fig. 8. The hexagonal electrode is powered and the reactor walls are grounded to serve as the counter-electrode. The plasma is sustained in the annular volume and etches the substrates placed on the hexagonal electrode. The region around each wafer locally resembles a diode reactor. The hexode increases the capacity of RIE configuration by extending the powered electrode into "the third dimension."

3.5 Downstream (Remote) Plasma Reactor

In a downstream plasma etching reactor (see Fig. 9), the plasma is sustained away from the wafer to be processed (Cook, 1987). For this reason, this reactor configuration is also referred to as the remote plasma or the afterglow reactor. As in the barrel reactor, etching is isotropic because only long-lived reactive radicals transported via diffusion and gas flow make it to the wafer surface. Depending on the distance between the plasma and the wafer, very few or almost no ions and electrons reach the surface. The plasma is sustained in a quartz tube either by external electrodes or coil as in the barrel

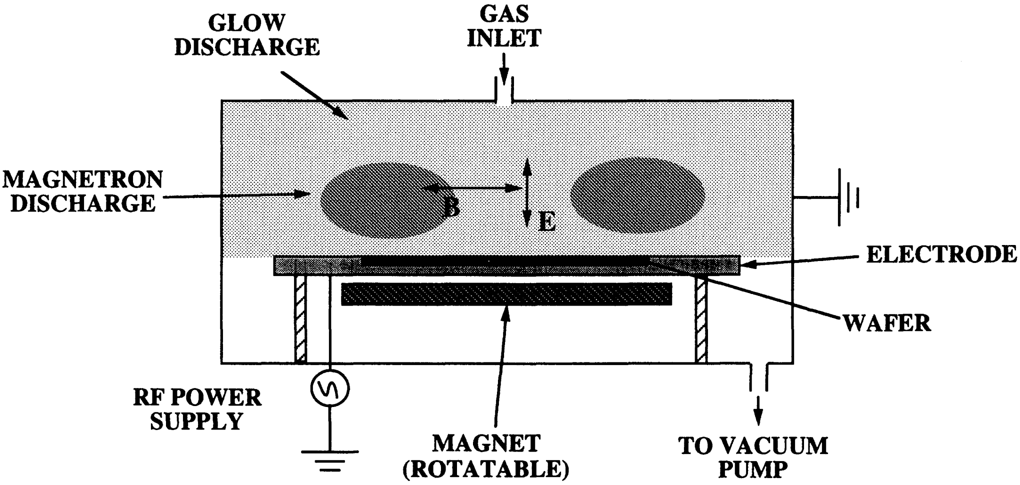

FIG. 7. Magnetically enhanced reactive ion etching reactor (MERIE) (reprinted with permission from Mucha *et al.*, 1994).

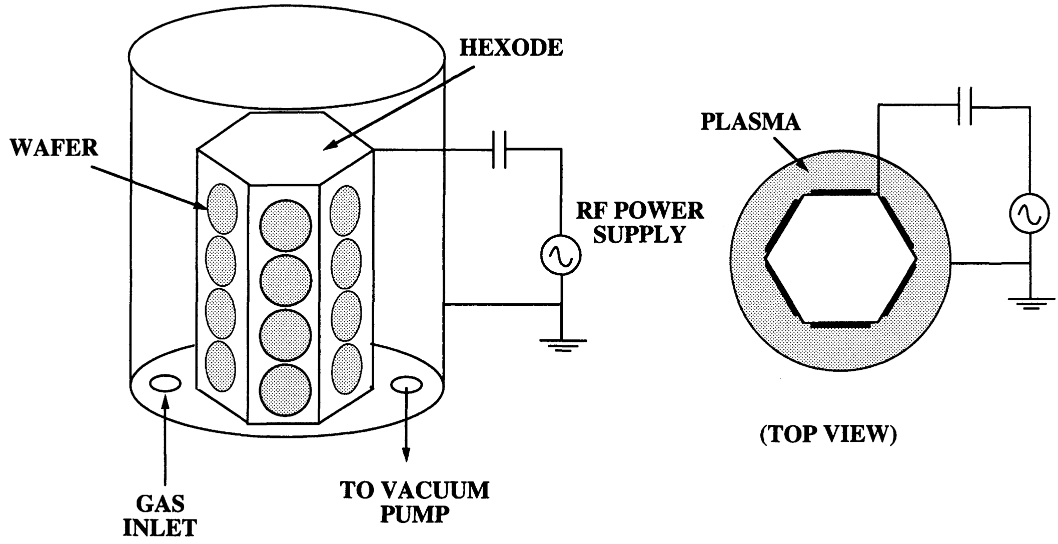

FIG. 8. Hexode plasma reactor (reprinted with permission from Mucha *et al.*, 1994).

reactor or by enclosing the plasma tube in a microwave cavity. Often the tube is bent to minimize line-of-sight radiation incident onto the wafer.

3.6 Triode Reactor

The triode reactor, shown in Fig. 10, is another variation of the diode configuration but employs three electrodes. As in the diode reactor, two of the electrodes are used to create ionization and to sustain the plasma by applying rf power to one of them. However, the wafer is placed on a third electrode, which is biased by an independent, and often different (lower) frequency, power supply in an effort to provide independent control of the ion-bombardment energy from the ion flux. A major limiting feature of the diode reactor is that the ion-bombardment flux and energy cannot be controlled independently from each other. For instance, an attempt to increase the ion flux by increasing the rf power can cause an increase in the sheath potential and thus the ion-bombardment energy in addition to an increase in the ion flux. To a certain extent, the triode arrangement decouples the power needed to sustain the plasma from the power needed to accelerate the ions across the sheath toward the wafer.

3.7 High-Density Plasma Reactors

These plasma sources were developed in the late 1980s to surmount the problems of the diode etcher and its variations, particularly to enable the independent control of the ion-bombardment flux and energy over a wide range of operating parameters. Thus, these reactors all operate like the triode reactor, where the plasma generation is decoupled from the substrate biasing so that the ion-bombardment energy can be controlled by separately biasing the workpiece. The difference between the high-density reactors and the triode reactor is that two of the electrodes used to energize the plasma have been replaced by very efficient high-power plasma sources. In fact, aside from hardware, the major differences among the different high-density plasma sources are in the mechanisms by which the applied field is coupled to electrons in the plasma. The high-density sources provide high ion densities (10^{11}–10^{13} cm^{-3}) at low pressure (1–10 m-Torr) that enable anisotropic etching at high rates. In addition, high-density plasma sources use efficient noncapacitive coupling of power to the electrons, thereby providing high ion fluxes (1–50 mA/cm^2) at low energies. In the next sections, each of the high-density sources is briefly described. Typical operating parameters for high-density discharges are listed in Table 1. A comprehen-

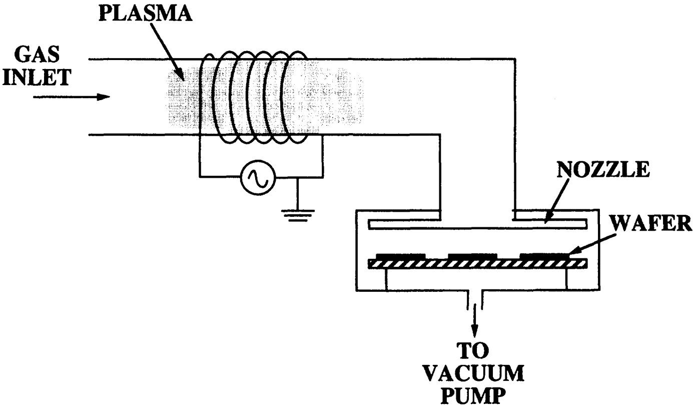

FIG. 9. Downstream (remote) plasma reactor (reprinted with permission from Mucha *et al.*, 1994).

sive review of the high-density sources is given by Lieberman and Gottscho (1993).

3.7.1 Electron Cyclotron Resonance Plasma Reactor ECR reactors (Asmussen, 1989) are conceptually separated into two regions: a source region where the power is absorbed by the electrons and a reactor region downstream from the source. However, ECR plasma reactors are not downstream reactors. The wafer is placed in the reactor region on a substrate platen that can be biased with a rf power supply. The process gases are injected into the source region. The plasma extends out of the source region into the reactor region and both reactive species and energetic ions impinge onto the wafer. The plasma is maintained by coupling microwave power at 2.45 GHz to the electrons in presence of a static magnetic field B, such that the microwave frequency ω is in resonance ($\omega = \omega_{ce}$) with the electron cyclotron frequency, $\omega_{ce} = eB/m_e$, where e and m_e are the electron charge and mass, respectively ($B = 875$ G for resonance at $\omega = \omega_{ce} = 2.45$ GHz). Electrons gain energy from the microwave electric field as they rotate in phase with the microwaves. Microwaves generated by a magnetron are guided to the reactor and launched through a dielectric window into the source region. One or more electromagnets surround the source and the reactor

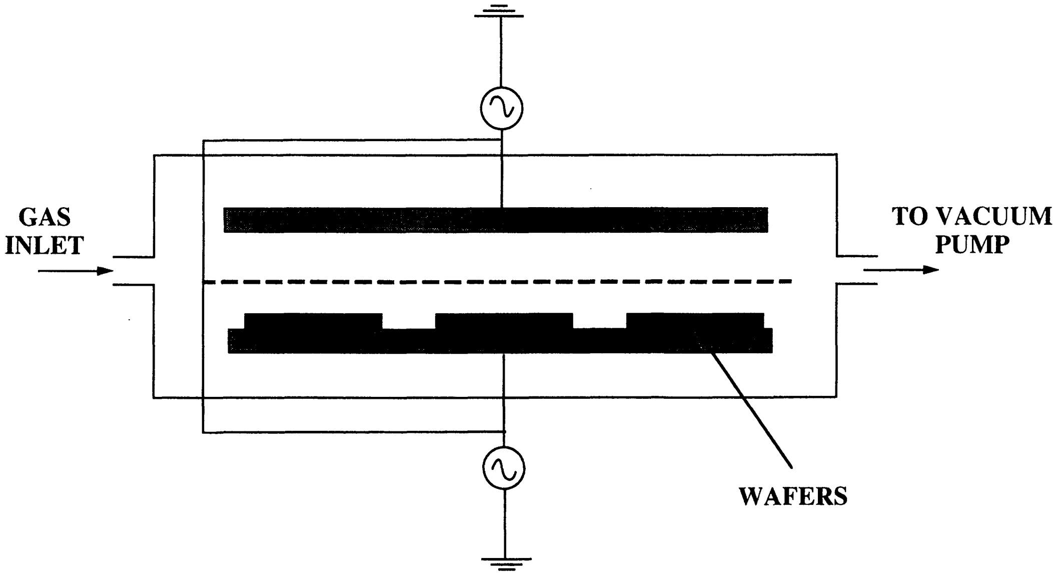

FIG. 10. Triode reactor (reprinted with permission from Mucha *et al.*, 1994).

to create the magnetic field necessary for resonance. ECR plasma reactors operate at pressures 0.1–100 mTorr and utilize 0.5–2 kW of power to generate a very dense (10^{11}–10^{13} ions/cm^3) plasma. There are different types of ECR plasma reactors. They differ in geometry, in the manner in which the microwaves are introduced into the reactor, and in the magnetic field profiles used to achieve resonance. One is shown in Fig. 11.

3.7.2 Helicon Plasma Reactors As in ECR reactors, helicon reactors can also be separated into a source region and a reactor region. The reactor region is similar to that of an ECR reactor. However, helicon plasma reactors utilize helicon waves to sustain a plasma in the source region rather than microwaves (Boswell, 1970, 1975; Chen, 1992). Helicon waves are right-hand circularly polarized whistler waves that can propagate in axially magnetized plasma columns. The helicon waves have complex modes characterized by radial, axial, and azimuthal variation, and they are called helicon waves because the electric and magnetic field components of the wave trace out a helical locus in space. Application of helicon-wave–excited plasmas to etching is relatively recent (Boswell and Henry, 1985; Perry *et al.,* 1991). Helicon waves are excited in the plasma by an appropriately shaped rf-driven (13.56 MHz) antenna placed concentrically outside the quartz tube that contains the discharge (see Fig. 12). There are several different helicon antenna shapes that can be used to excite helicon waves. A static magnetic field is also necessary to produce helicon waves and is created by coils placed around the source. As the helicon waves are excited and propagate through the plasma, energy is thought to be absorbed by the electrons through collisional or Landau damping. Process gases are injected from the top of the quartz tube and are ionized and dissociated as they flow through the source. Helicon sources produce a plasma with parameters that are comparable to those produced by ECR sources. Ion densities are in the range of 10^{11}–10^{13} ions/cm^3. However, the hardware required for helicon plasma excitation is much simpler than ECR sources.

3.7.3 Helical Resonators As in ECR and helicon plasma reactors, helical resonators can also be separated into a source region and a reactor region. Helical resonators use an entirely different method to excite the plasma. A helical resonator consists of a coil surrounded by a grounded shield as shown in Fig. 13. The plasma is sustained inside the dielectric tube placed inside the coil. One end of the coil is grounded while the other is left floating. Power is applied to a position in between the two ends. The structure becomes resonant when an integral number of quarter waves of the rf field fit between the two ends of the helix (Cook *et al.,* 1990). Large currents flowing through the coil at resonance create high fields inside the dielectric tube, which in turn accelerate the electrons in the discharge. Compared to the ECR

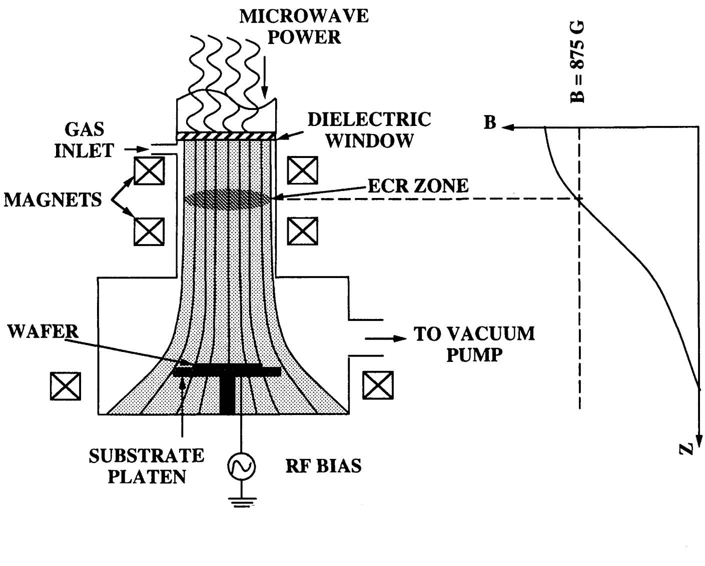

FIG. 11. High-profile electron cyclotron resonance (ECR) plasma reactor (reprinted with permission from Mucha *et al.,* 1994).

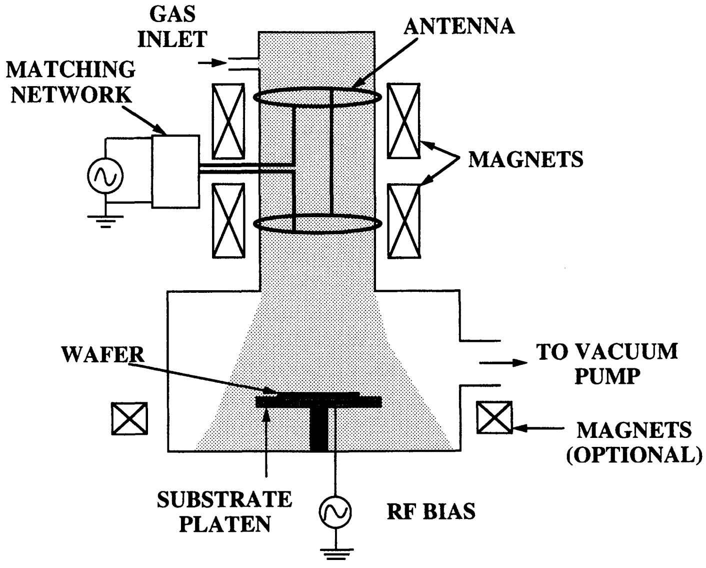

FIG. 12. Helicon plasma reactor (reprinted with permission from Mucha *et al.*, 1994).

and helicon reactors, helical resonators require very simple hardware but create a plasma with similar properties.

3.7.4 Inductively Coupled Plasma (ICP) Reactors—Radio-Frequency Induction (RFI) or Transformerly Coupled Plasma (TCP) Reactors In its simplest form, an inductively coupled plasma consists of a coil wrapped around a tube very much like the helical resonator. However, the name inductively coupled plasma reactor is generally associated with a new class of sources invented in the late 1980s that utilized the planar configuration shown in Fig. 14 (Hopwood, 1993). The planar coil used to excite the plasma is a cylindrical spiral that resembles a stove-top heater element. Radio-frequency (13.56 MHz) power is applied to the outer end of the coil, which is grounded at the center. The coil is placed on top of a quartz window, which forms one of the walls of the reactor. Radio-frequency inductive electric fields created inside the reactor sustain the plasma. ICP sources create high-density plasmas with properties similar to those of ECR, helicon, or helical resonator plasmas. As in other high-density sources, the wafer is placed on a platen that can be biased separately. The hardware required for TCP sources is as simple as helical resonators. Hopwood (1992) has reviewed the design and principles of RFI reactors.

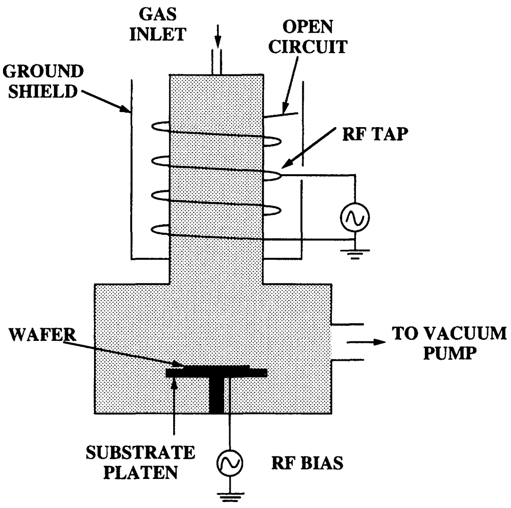

FIG. 13. Helical resonator (reprinted with permission from Mucha *et al.*, 1994).

4. MODELING OF PLASMA ETCHING REACTORS AND PLASMAS

There has been a growing interest in modeling of chemically reactive plasmas in order to improve our understanding of gaseous discharges and to develop more rational procedures for plasma reactor design. Modeling can give insight into many aspects of plasma etching. The models developed so far range from microscopic scale to reactor scale. One of the goals of plasma reactor

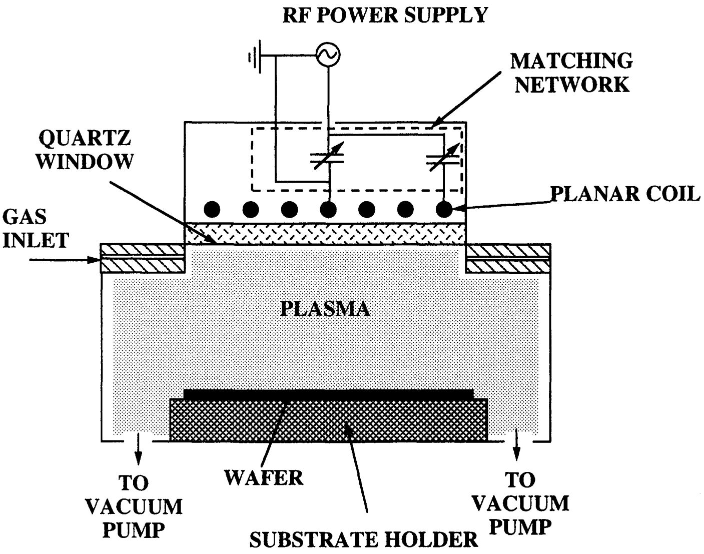

FIG. 14. Radio-frequency induction plasma reactor.

modeling has been to reduce the dependence of the technology on costly trial-and-error process development methods by developing predictive plasma reactor simulators and computer-aided design tools. A chemical reactor model fulfilling this need requires modeling of both the discharge physics and the discharge chemistry. The discharge-physics models can predict the charged-particle densities and energies, the electric field strength, and their distributions in the reactor. These variables are important in generating chemically reactive species and altering surface reactions primarily through the ion bombardment. Discharge-chemistry models include gas-phase and surface reactions of neutral species along with the transport of these species by convection and diffusion.

A common approach to plasma reactor modeling is to subdivide the reactor model into more manageable, simpler submodels as depicted in Fig. 15. Initial models of the plasma etching reactors concentrated on studying each submodel in isolation, despite the fact that the submodels are strongly and nonlinearly coupled to each other. Several investigators have conducted modeling studies of low-pressure chemically reacting plasmas in this manner. In each of the investigations, different aspects of the problem were emphasized. Some researchers focused on understanding the discharge physics (Graves and Jensen, 1986) while others emphasized neutral transport and reaction phenomena (Economou and Alkire, 1988). For detailed reviews of the plasma reactor literature the reader is referred to articles by Graves (1989), Lister (1992), and Kline and Kushner (1989), and review chapters by Meyyappan (1994) and Mucha *et al.* (1994). In this section, a brief overview of the plasma etching reactor modeling literature is given.

Herein, plasma models in the literature are summarized by separating them into five categories according to approach and focus:

1. electron-kinetics modeling—Boltzmann transport equation;
2. continuum glow-discharge models;
3. neutral transport and reaction models;
4. sheath models; and
5. particle-in-cell and Monte Carlo simulations.

Clearly there is an overlap between the different topics. There are two approaches to discharge modeling: the fluid (continuum) approach and the kinetic (particle simulation) approach. In the fluid approach, the plasma is treated as a continuum and as a mixture of neutral, ion, and electron fluids. In particle or Monte Carlo simulations, temporal behavior of large numbers of electrons, ions, or neutrals is computed as they move under the influence of the existing electric and/or magnetic fields and undergo collisions. Hybrid models where fluid and particle simulation approaches are combined have also been used.

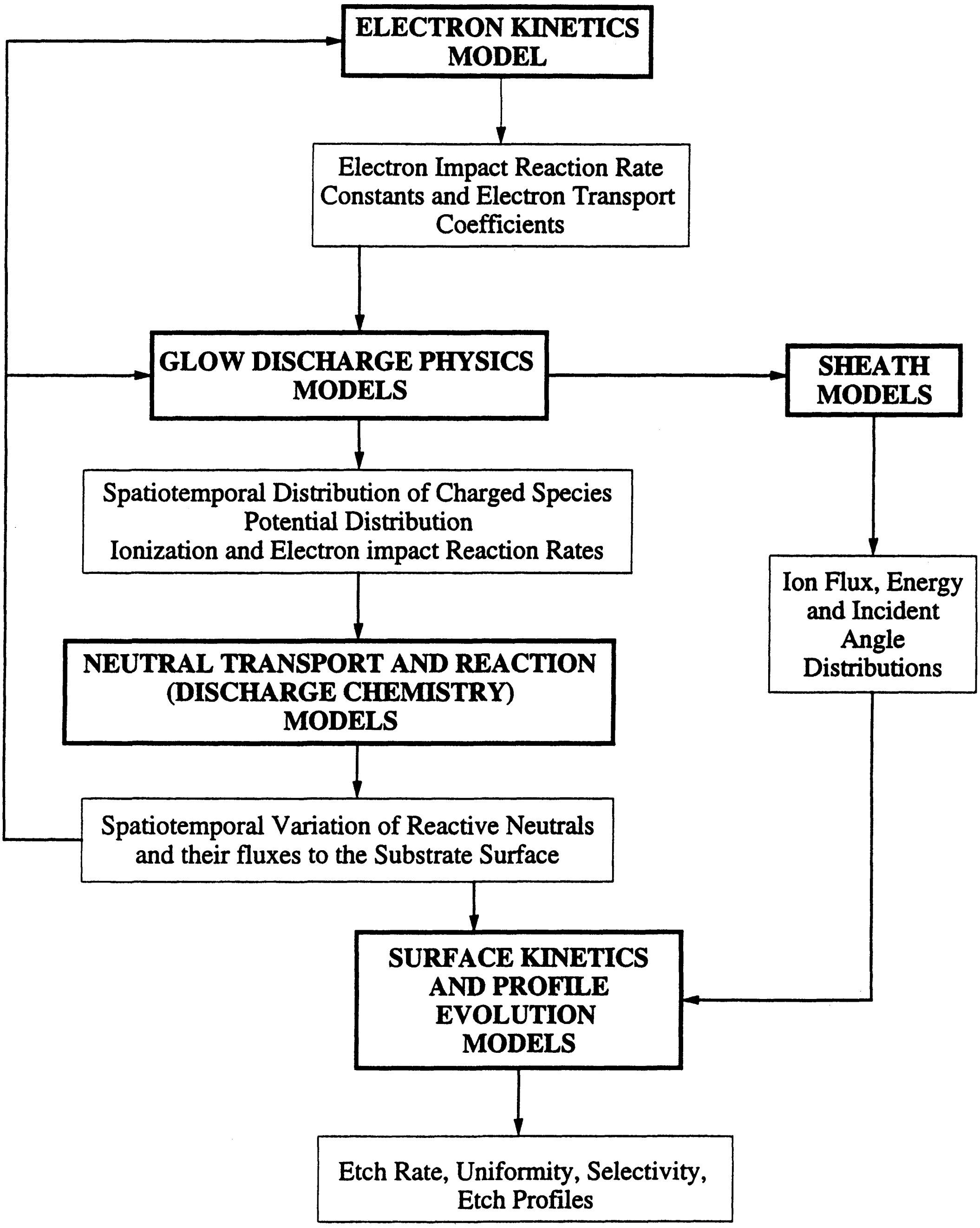
ELECTRON KINETICS MODEL
Electron Impact Reaction Rate Constants and Electron Transport Coefficients
GLOW DISCHARGE PHYSICS MODELS
SHEATH MODELS
Spatiotemporal Distribution of Charged Species
Potential Distribution
Ionization and Electron impact Reaction Rates
Ion Flux, Energy and Incident Angle Distributions
NEUTRAL TRANSPORT AND REACTION (DISCHARGE CHEMISTRY) MODELS
Spatiotemporal Variation of Reactive Neutrals and their fluxes to the Substrate Surface
SURFACE KINETICS AND PROFILE EVOLUTION MODELS
Etch Rate, Uniformity, Selectivity, Etch Profiles

FIG. 15. Schematic representation of a common way of dividing the plasma etching reactor model into several submodels and the interrelationship between these submodels.

4.1 Electron-Kinetics Modeling

The first step in plasma modeling is the computation of the electron energy-distribution function (EEDF), which is done either by solving the Boltzmann transport equation (Cherrington, 1979) or by resorting to Monte Carlo techniques. The product of the EEDF with the collision cross section for any of the electron-impact reactions shown in Table 3 determines the rate of that reaction. EEDF also determines the transport properties such as the mobility and diffusivity of the electrons in the plasma. A recent article by Jiang and Economou (1993) is an example of such a computation and contains a brief review of the literature related to solving the EEDF.

4.2 Continuum Glow-Discharge Models

Many investigators have simulated plasmas by using continuum models with emphasis on the discharge structure and physics. It is assumed that the mean free paths of the particles that make up the plasma are much larger than the reactor dimensions. In these models, electron- and ion-transport coefficients are treated phenomenologically, preferably by using functions determined after computing the EEDF. Transport and reaction of the charged particles are modeled by solving the material balances for the electrons, positive ions, and negative ions, the energy balance for the electrons, and Poisson's equation with the appropriate boundary conditions. In continuum models of the discharge, the focus is the discharge structure, and usually gas and surface chemistry are de-emphasized. The output of the continuum models comprises the self-consistent spatial and temporal distributions of all the charged species, ionization and electron-collision reaction rates, fluxes to the surfaces, and the electric field (potential) distribution.

4.3 Neutral Transport and Reaction Models

Neutral transport and reaction models emphasize the species transport, and gas-phase and surface reactions, although the surface reactions are treated phenomenologically. The purpose is to predict the spatial distribution of the reactive radicals in the reactor and their fluxes to the substrate. The neutral transport models are coupled to the continuum discharge models through the radical generation rates by the electron-impact reactions, which in turn are coupled to the EEDF. On the other hand, the calculation of the EEDF requires the knowledge of the plasma composition. This shows the intricate coupling between different phenomena occurring in plasma reactors.

In neutral transport models, the emphasis is on understanding the phenomena that control the spatial distribution of reactive radicals by solving the diffusion–convection–reaction equations (i.e., differential material balances) for the neutral species. The gas-flow field is obtained by solving the Navier–Stokes equations. The output of the neutral transport models includes the spatial variation of the etchant and its flux to the wafer surface. If the etching kinetics are known, the etchant flux can be used in conjunction with the ion flux to calculate the etch rate and its spatial distribution across the wafer.

4.4 Sheath Models

The goal of modeling the sheath is to predict the flux and the energy and angular distributions of ions bombarding the substrate. Ion transport in the sheath has been studied using both continuum and Monte Carlo models. Details of collision processes occurring in the sheath are difficult to incorporate into continuum models. The Monte Carlo models can treat details of ion–neutral collisions and conveniently yield the energy and angular distribution of ions impinging onto the film surface (Liu *et al.*, 1990). For this reason, Monte Carlo models are preferred over continuum models in modeling the sheath. In Monte Carlo techniques, the trajectory of an ion in the sheath is determined by solving the equations of motion. The calculations are repeated for many thousands of ions and the desired distributions are determined from the histograms.

4.5 Particle-in-Cell and Monte Carlo Simulations

The Monte Carlo simulation techniques used in modeling the sheath can also be used to model the plasma itself. This approach is an alternative to the continuum models of the discharge. Discharge physics and even chemistry can be modeled using the so-called particle-in-cell Monte Carlo (PIC-MC) simulations. The particle-in-cell simulations involve solving the equations of motion for thousands of electrons, ions, and neutrals with each particle representing millions of particles in reality. These particles used in the simulations are called superparticles. Particle density distributions and currents are determined on a grid and Maxwell's equations are solved simultaneously for the electric and magnetic fields. Collisions are modeled by allowing the superparticles to collide with other species or among themselves during their motion. Birdsall and Zangdon (1985) and Birdsall (1991) give an introduction to PIC-MC simulation methods and a comprehensive list of references.

5. SPECIFIC PLASMA ETCHING APPLICATIONS

In plasma etching, many materials may be present in the reactor and on surfaces simultaneously. Even in the simplest system, consisting of a mask and a film to be etched, there are two different materials on the substrate exposed to the plasma. In choosing etchants, one must be aware of the important aspects of the chemistry of each material. Over the past three decades many materials have been etched using plasmas. Guidelines for choosing etchants for plasma etching of various materials, particularly those encountered in IC manufacturing, have been reviewed numerous times. Table 3 is a summary of etchants used in plasma etching for different materials. References are also listed for more information.

Isotropic and anisotropic etching of silicon and polycrystalline silicon has been studied the most. Plasmas that produce F, Cl, and Br radicals can etch Si at rates ranging from 0.01 to several μm/min depending on the plasma density, pressure, etc. In fact, halogen atoms can etch many materials including metals, oxides, nitrides, silicides, III–V semiconductors, and III–V alloys, with a few exceptions. The difficulty in choosing an etchant is encountered when it is desired to etch one material selectively over another material. For example, F-containing plasmas can be used to etch silicon dioxide and nitride. However, if one desires to etch these materials selectively over silicon, plasma chemistry must be adjusted. Although adjusting plasma chemistry and conditions to achieve good selectivity remains an art, high selectivity is achievable and selective etching of oxides and nitrides over Si has been demonstrated. It has been empirically established that high selectivities can be achieved in F-deficient and apparently CF_3-rich plasmas. It is believed that Si etching is inhibited by formation of a polymer film on Si while the nitride and the oxide remain uncovered and etched.

Plasma etching of metals requires more care than etching of other materials since metals are readily oxidized with O_2 and water vapor to form stable oxides, which may act as an inhibitor as in the case of Al. In some cases where the metal oxide forms volatile chlorides or fluorides (Cr, W), O_2 may be intentionally added to the plasma to promote formation of an oxide. In etching some metals—for example, aluminum—an initiation or "lag time" is encountered at the beginning of plasma etching. This slow etching is due to a native oxide layer that must be removed before the underlying metal can be etched. The inhibiting properties of metal oxides can be used to one's advantage to promote anisotropic etching of holes and trenches. For example, oxidation can be delicately balanced by small additions of O_2 so that an inhibitor oxide forms on the sidewalls but not on the bottom of the features where the oxide can be removed by ion bombardment.

Polymeric organic films are usually present in almost all plasma etching processes because they are widely used as mask materials. The organic films are readily etched by O_2 or O_2-containing plasmas. Since O_2 is also used as an additive to plasmas for other purposes, selective etching of the film with respect to the mask may become an issue. Too much O_2 addition can result in mask etching. Plasma etching of organic films has been reviewed in detail by Hartney *et al.* (1989).

6. PLASMA PROCESS MONITORING: PLASMA DIAGNOSTICS

Discussions in the previous sections have established that plasma etching involves a wide variety of feed-gas compositions, and operating variables such as rf power, pressure, gas-flow rates, and reactor configurations. Although the black-box approach of changing these variables and observing the effects can be used in developing etching processes, it is far better and far more cost effective to identify and understand fundamental processes occurring in a gas discharge. To this end, a variety of plasma diagnostic techniques have been developed and used in determining the key plasma properties. These diagnostic techniques include various optical spectroscopy methods and mass spectroscopy to monitor gas-phase species, laser interferometry to measure, *in situ* and in real time, thicknesses of the films being etched, and Langmuir probes to determine the ion density and electron temperature. Some of these techniques are simple and robust enough to be used in manufacturing environment and have been used for process control. Because of its utility, plasma diagnostics in itself is a large field. Plasma diagnostic methods have been reviewed by Auciello and Flamm (1989). Below, some of the common diagnostic methods are summarized briefly. Of the diagnostic methods reviewed here, very few are actually used in manufacturing. This, however, has been changing rapidly in the last few years as the diagnostic equipment hardware becomes more inexpensive and robust. Still, some of the more complex diagnostics, such as laser-induced fluorescence and attenuated total-reflection Fourier-transform infrared spectroscopy, are only used in process development and research but not in manufacturing.

6.1 Optical Emission and Optical Diagnostic Techniques

Probably the most widely used diagnostic technique is optical-emission spectroscopy. In this technique, plasma-induced emission from the discharge is collected and analyzed in a spectrometer. Spectra from ultraviolet to near infrared have proven most useful for identifying the species in the plasma. Optical emission is well suited for species identification because the spectra, complex as they may be, provide a fingerprint of each atomic or molecular species. In addition, optical emission is a nonintrusive method and does not disturb the plasma. Optical emission can be collected as a function of spatial position and time, thus allowing one to follow spatiotemporal changes in the plasma. Obtaining concentrations of species in the plasma is more difficult since emission intensity depends not only on species concentration but also on electron temperature and density. An approximate technique called actinometry (Coburn and Chen, 1980; d'Agostino *et al.*, 1981) has been discovered to overcome this difficulty. Although actinometry does not provide absolute species concentrations, it has been used to map relative etchant concentrations as a function of position and operating parameters in a variety of reactors.

Optical emission can also be used to determine the end of etching of a film. This is commonly called "end-point" detection. The technique makes use of the fact that the etchant or etching-product concentrations undergo abrupt changes when the film being etched is removed. For example, in etching of a polysilicon film, Cl is depleted in the discharge by etching reactions. When the Si film is totally removed from the surface, a rise in Cl concentration will be observed because the reactions that eliminate Cl have been completed. Thus, by monitoring Cl concentration in Si etching, one can detect the end of the etching process. This technique is not unique to Cl and has been used in many plasma etching processes to detect the end point. Optical emission and other end-point detection techniques have been reviewed by Roland *et al.* (1985).

Optical emission has also been used to map out etchant concentrations in a plasma during etching (Economou *et al.* 1991). Since uniformity of etching is related to etchant distribution over the wafer, this method provides an *in situ* method for monitoring the uniformity.

Another optical technique that has been used to obtain the spatiotemporal variation of species concentration in plasma reactors is laser-induced fluorescence (LIF). In this technique, ground-state species are excited to a higher energy level by a laser and emission arising from the subsequent deexcitation to a lower energy level is detected. Since one has a control over the excitation and emission wavelengths, spectra obtained are simpler than plasma-excited emission and highly selective with respect to the species being monitored. However, hardware for these experiments is less robust and more expensive than in optical emission spectroscopy. Thus, LIF is not a suitable candidate for monitoring plasmas in a manufacturing environment but can be and has been used as a process development tool. Optical emission, laser-induced fluorescence, and other optical diagnostic techniques have been reviewed by Gottscho and Miller (1984) and by Donnelly in the book edited by Auciello and Flamm (1989).

6.2 Mass Spectroscopy

Mass spectroscopy, like optical emission, also provides a method for identification of species in a discharge and products in the effluent from a reactor. Mass spectroscopy is most commonly used to monitor vacuum integrity and process-gas purity. Mass spectroscopy does not give spatially resolved information but can be used for end-point detection in a similar manner to optical emission.

One use of mass spectroscopy in plasma etching research is in identification of species and ions bombarding the surfaces immersed in a plasma. This is done by making a sampling orifice on the substrate platen. The species incident onto the electrode pass though the sampling orifice and are analyzed by a mass spectrometer. There are many subtleties involved with these types of experiments; the technique is reviewed by Coburn (1989) and discussed in Auciello and Flamm (1989).

6.3 Langmuir Probes

A Langmuir probe is a thin wire or wires immersed in a plasma and biased with respect to a reference electrode. The current *I* collected by the probe as a function of the bias voltage *V* carries information about the electron and ion density, the electron temperature, and the plasma potential. Extraction of these values is done by using a model for current collection by the probe. For this reason, Langmuir-probe data interpretation may be difficult. Furthermore, Langmuir-probe measurements involve placing the probe into the plasma and therefore are intrinsically intrusive. To minimize the disturbance fast-injection probes have been invented in which the probe is pneumatically injected into the plasma and remains in the discharge only for a limited amount of time (≈100 ms). Langmuir probes can be used to collect spatially resolved information and are mostly used for determining the uniformity of ion density in a plasma. Langmuir-probe operation and theory have been reviewed in a number of publications (Auciello and Flamm, 1989; Swift and Schwar, 1969; Chen, 1965).

6.4 *In Situ* Film-Thickness Measurements: Laser Interferometry and Reflectometry

Monitoring plasma etching using optical reflection from surfaces has proven to be a very simple and useful technique. Optical reflection can be used for *in situ,* real-time film-thickness measurements and end-point detection. In these techniques, light from a source, such as a laser, is reflected off the etching surface and the reflected photons are collected using a detector. If the film being etched is an absorbing material such as metal, a change in the reflectivity will be observed at the interface between the film and the substrate because of a change in the refractive index. Although this is a useful method of determining the end point, films of known thickness must be etched in order to determine the etch rate.

Reflectivity measurement becomes even more powerful if the film is a nonabsorbing dielectric such as SiO_2 or Si_3N_4. In this case, the reflected intensity varies periodically as a function of the film thickness because of constructive and destructive interference produced by the reflections from the front and the rear (substrate–film interface) surfaces of the film. For this reason, this technique is called interferometry. As the film is etched and its thickness decreases, the reflected intensity undergoes oscillations with a period related to the film thickness. The etch rate can be calculated from the period of oscillations and the refractive index of the film. Interferometry and end-point detection using reflection-based techniques have been reviewed by Busta *et al.* (1979), Marcoux and Foo (1981), and Roland *et al.* (1985). More recently, interferometry has been extended to measure uniformity of etching in plasma reactors (Economou *et al.,* 1991; Dalton *et al.,* 1994) as well as substrate temperature (Donnelly and McCaulley, 1990).

6.5 Surface Diagnostic Techniques

Standard surface analysis techniques such as Auger and x-ray photoelectron spectroscopy provide a wealth of information on surfaces after etching but cannot be used *in situ* during plasma processing. Two *in situ* techniques have emerged to monitor surfaces: these are ellipsometry and attenuated total-reflection Fourier-transform infrared spectroscopy (ATR-FTIR). ATR-FTIR spectroscopy is one of the possible arrangements for performing surface infrared spectroscopy and is particularly well suited for studying surface adsorbates on nonabsorbing substrates with relatively high index of refraction such as semiconductors (Chabal, 1988). In ATR-FTIR spectrometry, infrared radiation from a Michaelson interferometer is focused onto one of the beveled edges of a semiconductor sample that is shaped as a trapezoid. The infrared beam is trapped inside the sample and undergoes multiple total internal reflections at the vacuum–semiconductor interface. The beam finally exits the sample from a second beveled edge. The transmitted beam is collected and Fourier analyzed. In this manner, an evanescent electric field is created at the vacuum–semiconductor interface and the interaction of this field with the surface adsorbates and thin films can be studied. ATR-FTIR spectroscopy provides the means to study the infrared absorption spectra of surface species and thin films. Use of ATR-FTIR in plasma processing has been reviewed by Aydil *et al.* (1994).

In ellipsometry, the substrate surface is illuminated by light with a known polarization and the change in polarization upon reflection is measured. The change in polarization provides information about the substrate and its overlayers. The optical properties can in turn be related to the nature of the overlayers and thus ellipsometry can provide information on surface processes. *In situ* ellipsometry has been used from the ultraviolet to the infrared region of the spectrum to diagnose surfaces exposed to plasmas (Drevillon, 1988). Principles of ellipsometry are reviewed in a collection of papers by Azzam (1991) and Azzam and Bashara (1977).

7. SUMMARY

Plasma etching is a crucial process for patterning thin films, particularly in manufacturing of integrated circuits and electronic devices. This article describes the basic principles behind plasma etching and properties of glow discharges as they relate to thin-film etching. Plasma etching is a very fast-evolving field and any introductory article is bound to be dated. This article is an attempt to assimilate the information currently available on plasma etching and present a framework within which plasma etching can be viewed. This article and the literature cited should serve as an entry point to the field of plasma etching.

GLOSSARY

Actinometry: An approximate calibration technique used to determine relative concentration of species in a plasma using optical emission spectroscopy.

Anisotropic Etching: Etching that proceeds at a higher rate in a direction normal to the substrate surface.

End Point: End of etching of a film.

Etchant: Collective name given to the reactive radicals, molecular fragments, and ions that participate in etching in a plasma.

Glow Discharge: A term used synonymously with low-pressure plasmas. It implies nonequilibrium among the ions, electrons, and neutrals.

Integrated Circuits (ICs): A collection of solid-state electronic devices built on a semiconductor substrate and interconnected together to form a working unit capable of performing complex electronic functions.

Ion-Assisted Etching: Etching that results from the combined synergistic action of bombardment by ions and reactive radicals in a plasma.

Isotropic Etching: Etching that proceeds at equal rates in all directions into the film or substrate.

Loading: A decrease in etchant concentration in a plasma due to depletion by etching reactions.

Overetching: Etching for a time longer than required to remove a given thickness.

RIE: Reactive ion etching.

Selective Etching: Preferential etching of one material over another.

Selectivity: Ratio of etch rates of two materials. Usually one material is the material to be etched and the other is a mask.

Sheath: A thin, dark, space-charge region that forms next to any surface that is brought in contact with a plasma.

Works Cited

Asmussen, J. (1989), *J. Vac. Sci. Technol. A* **7**, 883–893.

Auciello, O., Flamm, D. L. (Eds.) (1989), *Plasma Diagnostics: Discharge Parameters and Chemistry,* Orlando, Florida: Academic Press.

Aydil, E. S., Economou, D. J. (1991), *J. Appl. Phys.* **69**, 109–114.

Aydil, E. S., Economou, D. J. (1993), *J. Electrochem. Soc.* **140**, 1471–1481.

Aydil, E. S., Gottscho, R. A., Chabal, Y. J. (1994), *Pure Appl. Chem.* **66**, 1381–1388.

Azzam, R. M. (1991), *Selected Papers on Ellipsometry,* Bellingham, Washington: SPIE Optical Engineering Press.

Azzam, R. M., Bashara, N. M. (1977), *Ellipsometry and Polarized Light,* Amsterdam: North Holland.

Birdsall, C. K., Langdon, A. B. (1985), *Plasma Physics via Computer Simulation,* New York: McGraw-Hill.

Birdsall, C. K. (1991), *IEEE Trans. Plasma Sci.* **PS-19**, 65–85.

Boswell, R. W. (1970), *Phys. Lett.* **33**, 457–458.

Boswell, R. W. (1975), *Nature* **258**, 58–60.

Boswell, R. W., Henry, D. (1985), *Appl. Phys. Lett.* **47**, 1095–1097.

Burton, R. A., Gottscho, R. A., Smolinsky, G. (1984), "Dry Etching of Group III–Group V Semiconductors," in: R. A. Powell (Ed.), *Dry Etching for Microelectronics,* Chap. 3, Amsterdam: Elsevier.

Busta, H. H., Lajos, R. E., Kiewit, D. A. (1979), *Solid State Technol.* **22**, 61–64.

Chabal, Y. J. (1988), *Surf. Sci. Rep.* **8**, 211–357.

Chapman, B. (1980), *Glow Discharge Processes,* New York: John Wiley.

Chen, F. F. (1965), in: R. A. Huddlestone, S. L. Leonard (Eds.), *Plasma Diagnostic Techniques,* Orlando, Florida: Academic Press, Chapter 4, pp. 113–200.

Chen, F. F. (1992), *J. Vac. Sci. Technol. A* **10**, 1389–1401.

Cherrington, B. E. (1979), *Gaseous Electronics and Gas Lasers,* New York: Pergamon Press.

Coburn, J. W., Winters, H. F. (1979), *J. Appl. Phys.* **50**, 3189–3196.

Coburn, J. W., Chen, M. J. (1980), *J. Appl. Phys.* **51**, 3134–3136.

Coburn, J. W. (1982), *Plasma Etching and Reactive Ion Etching,* AVS Monograph Series, New York: American Vacuum Society.

Coburn, J. W., Winters, H. F. (1983), *Annu. Rev. Mater. Sci.* **13**, 91–116.

Coburn, J. W. (1989), *Thin Solid Films* **171**, 65–80.

Cook, J. M. (1987), *Solid State Technol.* **32** (4), 147–151.

Cook, J. M., Ibbotson, D. E., Flamm, D. L. (1990), *J. Vac. Sci. Technol. B* **8**, 1–4.

d'Agostino, R., Cramarossa, F., Debenedictis, S., Ferraro, G. (1981), *J. Appl. Phys.* **52**, 1259–1265.

Dalton, T. J., Conner, W. T., Sawin, H. H. (1994), *J. Electrochem Soc.* **141**, 1893–1900.

Donnelly, V. M., Flamm, D. L. (1981), *Solid State Technol.* **26**, 161–166.

Donnelly, V. M., McCaulley, J. A. (1990), *J. Vac. Sci. Technol. A* **8**, 84–92.

Drevillon, B. (1988), *Thin Solid Films* **183**, 157–163.

Economou, D. J., Alkire, R. C. (1988), *J. Electrochem. Soc.* **135**, 2786–2794.

Economou, D. J., Aydil, E. S., Barna G. (1991), *Solid State Technol.* **36**, 107–111.

Flamm, D. L. (1991), *Solid State Technol.* **34**, 47–50.

Flamm, D. L., Donnelly, V. M. (1981a), *Plasma Chem. Plasma Proc.* **1**, 317–363.

Flamm, D. L., Donnelly, V. M. (1981b), *Solid State Technol.* **24**, 161–166.

Flamm, D. L., Wang, D. N., Maydan, D. J. (1982), *J. Electrochem. Soc.* **129**, 2755–2760.

Gottscho, R. A. (1993), *Phys. World* **6**, 39–45.

Gottscho, R. A., Miller, T. A. (1984), *Pure Appl. Chem.* **56**, 189–208.

Gottscho, R. A., Jurgensen, C. W., Vitkavage, D. J. (1992), *J. Vac. Sci. Technol. B* **10**, 2133–2147.

Graves, D. B. (1989), *AIChE J.* **35**, 1–29.

Graves, D. B. (1994), *IEEE Trans. Plasma Sci.* **22**, 31–42.

Graves, D. B., Jensen, K. F. (1986). *IEEE Trans. Plasma Sci.* **PS-14**, 78–91.

Hartney, M. A., Hess, D. W., Soane D. S. (1989), *J. Vac. Sci. Technol. B* **7**, 1.

Hess, D. W. (1988), *Solid State Technol.* **31**, 97–103.

Hess, D. W. (1990), *J. Vac. Sci. Technol. A* **8**, 1677–1684.

Hopwood, J. (1992), *Plasma Sources Sci. Technol.* **1**, 109–116.

Hopwood, J., Guarnieri, C. R., Whitehair, S. J., Cuomo, J. J. (1993), *J. Vac. Sci. Technol. A* **11**, 147–151.

Irving, S. M. (1968), in: *Proceedings of the Kodak Photoresist Seminar,* Vol. 2, Syracuse, NY: Eastman Kodak, p. 26.

Jiang, P., Economou, D. J. (1993), *J. Appl. Phys.* **73**, 8151–8160.

Kline, L. E., Kushner, M. J. (1989), *Crit. Rev. Solid State Mater. Sci.* **16**, 1–35.

Leahy, M. F., Kaganowicz, G. (1987), *Solid State Technol.* **32**, 99–103.

Lieberman, M. A., Lichtenberg, A. J. (1994), *Principles of Plasma Discharges and Materials Processing,* New York: John Wiley & Sons Inc.

Lieberman, M. A., Gottscho, R. A. (1993), in: M. Francombe, J. Vossen (Eds.), *Physics of Thin Films,* Orlando, Florida: Academic Press.

Lister, G. G. (1992), *J. Phys. D* **25**, 1649–1680.

Liu, J., Huppert, G. L., Sawin, H. H. (1990), *J. Appl. Phys.* **68**, 3916–3934.

Manos, D. M., Flamm, D. L. (Eds.) (1989), *Plasma Etching: An Introduction,* New York: Academic Press Inc.

Marcoux, P. J., Foo, P. D. (1981), *Solid State Technol.* **24**, 115–122.

Mathad, G. (1985), *Solid State Technol.* **28**, 221–225.

Meyyappan, M. (1994), "Plasma Process Modeling," in: M. Meyyappan (Ed.) *Computational Modeling in Semiconductor Processing,* Chap. 5, New York: Artech House, pp. 231–323.

Mucha, J. A., Hess, D. W., Aydil, E. S. (1994), "Plasma Etching," in: L. F. Thompson, C. G. Willson, M. J. Bowden, (Eds.), *Introduction to Microlithography,* Chap. 5, Washington, DC: American Chemical Society, pp. 376, 377–507.

Perry, A. J., Vender, D., Boswell, R. W. (1991), *J. Vac. Sci. Technol. B* **9**, 310–317.

Proud, J. M., Gottscho, R. A. (1991), *Plasma Processing of Electronic Materials: Scientific Opportunities and Technical Challenges,* National Research Council, Panel on Plasma Processing of Materials, Washington, DC: National Academy Press.

Roland, J. P., Marcoux, P. J., Ray, G. W., Rankin, G. H. (1985), *J. Vac. Sci. Technol. A* **3**, 631–636.

Sawin, H. H. (1994), *Microelectronic Eng.* **23**, 15–21.

Stevens, J. E., Huang, Y. C., Jarecki, R. L., Cecchi, J. L. (1992), *J. Vac. Sci. Technol A* **10**, 1270–1275.

Swift, J. D., Schwar, M. J. (1969), *Electric Probes for Plasma Diagnostics,* New York: Elsevier.

Further Reading

Auciello, O., Flamm, D. L. (Eds.) (1989), *Plasma Diagnostics: Discharge Parameters and Chemistry,* Orlando, Florida: Academic Press.

Birdsall, C. K., Langdon, A. B. (1991), *Plasma Physics Via Computer Simulation I,* New York: Adam Hilger.

Chapman, B. (1980), *Glow Discharge Processes,* New York: John Wiley.

Coburn, J. W., Winters, H. F. (1979), *J. Vac. Sci. Technol.* **61**, 391–403.

Coburn, J. W. (1982), *Plasma Chem. Plasma Proc.* **2**, 1–41.

Gottscho, R. A. (1993), *Phys. World* **6**, 39–45.

Graves, D. B. (1989), *AIChE J.* **35**, 1–29.

Graves, D. B. (1994), *IEEE Trans. Plasma Sci.* **22**, 31–42.

Jensen, K. F. (1987), *Chem. Eng. Sci.* **42**, 923–958.

Kline, L. E., Kushner, M. J. (1989), *Crit. Rev. Solid State Mater. Sci.* **16**, 1–35.

Lieberman, M. A., Lichtenberg, A. J. (1994), *Principles of Plasma Discharges and Materials Processing,* New York: John Wiley & Sons Inc.

Manos, D. M., Flamm, D. L. (Eds.) (1989), *Plasma Etching: An Introduction,* New York: Academic Press Inc.

Meyyappan, M. (1994), "Plasma Process Modeling," in: M. Meyyappan (Ed.), *Computational Modeling in Semiconductor Processing,* Chap. 5, New York: Artech House, pp. 231–323.

Mucha, J. A., Hess, D. W., Aydil, E. S. (1994), "Plasma Etching," in: L. F. Thompson, C. G. Willson, M. J. Bowden, (Eds.), *Introduction to Microlithography,* Chap. 5, Washington, DC: American Chemical Society, pp. 376, 377–507.

Proud, J. M., Gottscho, R. A. (1991), *Plasma Processing of Electronic Materials: Scientific Opportunities and Technical Challenges,* National Research Council, Panel on Plasma Processing of Materials, Washington, DC: National Academy Press.

Sze, S. M. (1985), *Semiconductor Devices: Physics and Technology,* New York: John Wiley.

Winters, H. F., Coburn, J. W. (1992), *Surf. Sci. Rep.* **14**, 161–269.

PLASMA PHYSICS

Francis F. Chen, *Electrical Engineering Department, University of California, Los Angeles, California, U.S.A.*

Introduction 200
1. **History** 201
2. **Basic Plasma Phenomena** ... 202
2.1 Methods of Treatment 202
2.1.1 Fluid Description 202
2.1.2 Kinetic Description 203
2.1.3 Computer Simulation 204
2.2 Characteristics of Plasma Behavior 205
2.2.1 Quasineutrality and Debye Shielding 205
2.2.2 Characteristic Frequencies and Velocities 205
2.2.3 Particle and Fluid Drifts 206
2.2.4 Adiabatic Invariants 207
2.2.5 Resistivity and Diffusion 207
2.2.5.1 Weakly Ionized Plasmas 207
2.2.5.2 Fully Ionized Plasmas 208
2.2.6 Collisionless Damping 209
2.3 Plasma Waves and Instabilities 209
2.3.1 Electrostatic Waves 210
2.3.2 Electromagnetic Waves 210
2.3.3 Electrostatic Instabilities 211
2.3.4 Hydromagnetic Instabilities ... 211
2.4 Radiation in Plasmas 212
2.4.1 Bremsstrahlung 212
2.4.2 Synchrotron Radiation 212
2.4.3 Scattering of Radiation 212
2.4.4 Diagnostics with Radiation 212
2.5 Nonlinear Phenomena 213
2.5.1 Nonlinear Waves and Wave–Particle Interactions 213
2.5.2 Sheaths and Double Layers ... 213
2.5.3 Solitons, Cavitons, and Collapse 213
2.5.4 Turbulence and Chaos 214
2.6 Nonneutral and Strongly Coupled Plasmas 214
2.6.1 Single-Species Plasmas 214
2.6.2 Liquid and Solid Plasmas 215
3. **Physics of Magnetic Confinement** 215
3.1 Effects of Magnetic Fields on Plasmas 215
3.2 Magnetic-Confinement Fusion 215
3.3 Particle and Energy Transport 216
3.4 The Energy Principle 216
3.5 Toroidal Confinement 216
3.5.1 Rotational Transform 217
3.5.2 Instabilities 217
3.5.3 Stabilization Methods 217
3.5.4 Banana Orbits 217
3.5.5 Neoclassical Diffusion 217
3.5.6 Bohm and Gyro-Bohm Diffusion 218
3.6 Tokamak Physics 218
3.6.1 Fields and Currents 218
3.6.2 Sawteeth and Disruptions 218
3.6.3 Auxiliary Heating 218
3.6.4 H-Mode Confinement 219
3.6.5 Current Drive 219
3.7 Magnetic Mirrors 219
3.7.1 Loss Cone 219
3.7.2 Instabilities 219
3.7.3 Tandem Mirrors 219
4. **Physics of Inertial Confinement** 220
4.1 Laser Fusion 220
4.2 Parametric Instabilities 220
4.3 Rayleigh–Taylor Instability 221
4.4 Driver Technology 221
5. **Plasmas in Space** 221
5.1 Plasmas in the Solar System . 221
5.2 Plasmas in the Sun 222
5.3 Magnetic-Field Reconnection 222
5.4 Plasma Astrophysics 223
6. **Light and Particle Beams** ... 224
6.1 Coherent Radiation Generators 224
6.1.1 Microwave Sources 224
6.1.2 Free-Electron Lasers 224
6.2 Pulsed Power 224
6.2.1 Relativistic Electron Beams ... 224
6.2.2 Light-Ion Beams 225
6.2.3 Pinches, Switches, and Exploding Wires 225
6.3 Plasma Accelerators 225

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

6.3.1 Beat-Wave Accelerator 225
6.3.2 Wake-Field Accelerator 225
6.3.3 Plasma Lenses 225
6.4 X-Ray and Short-Pulsed Lasers 226
6.4.1 X-Ray Lasers 226
6.4.2 Femtosecond Lasers 226
7. Industrial Applications 226
7.1 Plasma Sources 226
7.2 Plasma Etching and Deposition 227
7.3 Gaseous Electronics 228
7.4 Implantation and Polymerization 228
7.5 Thermal Plasmas 229
7.6 Isotope Separation 229
7.6.1 Laser Isotope Separation 229
7.6.2 Ion-Cyclotron Isotope Separation 229
Glossary 229
Works Cited 234
Further Reading 236

INTRODUCTION

The most common type of plasma is a gas of such high temperature that it is ionized; that is, an appreciable number of atoms have been stripped of at least one electron and have become positive ions (Plates 1 and 2). Because of its free electrical charges, a plasma differs from ordinary gases because it is subject to electric and magnetic forces. Indeed, the science of plasma physics mostly concerns those plasmas in which the usual molecular forces are negligibly small compared with the electromagnetic forces. Plasma has been called the "fourth state of matter," after solid, liquid, and gas. All matter goes into this electrified state above temperatures of 10 000 K or so, at which point the fastest particles in the thermal distribution undergo collisions energetic enough to ionize an atom.

Plasmas can also exist at room temperature in metals and semiconductors in which there are free charge carriers. At the other extreme, in the vacuum of outer space, hydrogen atoms are ionized by photoionization by starlight. Once ionized, the ions and electrons do not recombine easily because, at a density of about 1 ion/cm^3, collisions are rare. Consequently, perhaps 99% of all the matter in the universe is in the plasma state. On Earth, however, the temperatures and atmospheric densities that are consistent with life cannot support plasmas, and these can be studied only in laboratory vacuum chambers.

Within the framework of physics and engineering, plasma physics is an interdisciplinary science that affects many other fields. It is an extension of electromagnetics that can treat conducting gases that exist in stellar interiors and in nuclear-fusion devices constructed to reproduce the conditions there. It describes the motion of charge carriers in semiconductor devices and in the machines used to make them. It is used in the design of gas lasers and microwave generators. Plasma physics is essential to the understanding of the behavior of matter under conditions of high energy density, such as in quasars and pulsars, and in targets bombarded by intense laser or particle beams. It has applications even to chemistry through the use of electrons to form new compounds and polymers. The development of femtosecond lasers to probe chemical and surface reactions in real time will be aided by the knowledge of how intense radiation interacts with matter in the plasma state.

Plasmas are more complicated than ordinary fluids for several reasons. First, they are subject to long-range electromagnetic forces. Second, plasmas have a particulate nature, unlike fluids, which are continuous media. At high temperatures, collisions are so rare that plasmas can be treated as collisionless fluids; and in the absence of randomizing collisions, the individual particles of the fluid must be taken into account. Third, since the charged-particle motions are greatly affected by magnetic fields, the plasma is a highly anisotropic dielectric and supports a rich variety of possible wave motions.

Though plasmas have been known for a much longer time, plasma physics as a science did not begin to advance rapidly until the late 1950s, when the search for an infinite energy source through controlled hydrogen fusion began. The principles of plasma science are now considered to be indispensable in many other disciplines besides controlled fusion: in astrophysics and space physics, in particle accelerators and radiation generators, in solid-state devices and lasers, in the understanding of turbulence and chaos in nonlinear dynamics, and, most recently, in manufacturing sciences.

1. HISTORY

Plasma physics had its beginnings in two disparate disciplines: gaseous electronics and cosmic electrodynamics. In the 1920s, Irving Langmuir and his colleagues, in their studies of gas discharges used in electron tubes, laid out the principles of plasma measurements with the electrostatic probes that bear his name. In the 1930s, Chapman and Cowling (1939) worked out the theory of nonuniform gases, and Mott and Massey (1933) contributed the theory of atomic collisions. During World War II, research on isotope separation using plasma discharges at Los Alamos led David Bohm to invent the infamous "Bohm diffusion" law whose origin is still shrouded in mystery. Gaseous electronics took a leap after the war as electrical engineers at Massachusetts Institute of Technology and Stanford University competed to understand beam–plasma interactions and to develop better klystrons and other microwave generators for radar. The theory of radio-wave propagation in the ionosphere was worked out simultaneously by V. L. Ginzburg (1964) and others in Russia. Experiments were still limited to Langmuir-probe measurements in weakly ionized discharges. It was also in the late 1950s and early 1960s that the other stem of plasma physics began its growth. H. Alfvén (1950), S. Chandrasekhar (1961), and other astrophysicists became aware of the role of plasmas in deep space and worked out the principles of magnetohydrodynamics, now known as hydromagnetics, including the discovery of "Alfvén waves." Astronomers tended to view plasmas as conductive fluids like mercury, whereas the gaseous electronics community treated plasmas from the particle point of view. These opposing views of the dual nature of plasmas were not merged until the era of nuclear-fusion research.

Following World War II, dense, hot plasmas were produced by the "pinch" effect at the Harwell military laboratories in England; these experiments foreshadowed the scramble for fusion power. Attempts to produce thermonuclear reactions in the laboratory and thus to "tame the H-bomb" were started in England, in Russia, and at four sites in the United States: at the national laboratories at Livermore, Los Alamos, and Oak Ridge, and at Princeton University. All of these programs were classified. The project at Princeton was code-named Project Matterhorn by its mountain-climbing leader, astrophysicist Lyman Spitzer, Jr. The leader of the Los Alamos Project was James Tuck, and this connection to Friar Tuck of Robin Hood fame led to the code name "Project Sherwood" for the U.S. fusion program. Classification was removed and international cooperation began at the 1958 Atoms for Peace conference in Geneva, where each nation displayed its best results. The United States unveiled two mirror-confinement programs (at Livermore and Oak Ridge), a pinch program at Los Alamos, and a "Stellarator" program at Princeton. The United Kingdom exhibited a large toroidal pinch called "Zeta," which produced copious neutrons, but which was later discredited when the neutrons were found to be of nonthermonuclear origin. The Russians exhibited an unglamorous piece of hardware that they dubbed a "tokamak," a Russian acronym for "toroidal magnetic chamber." This concept not only survived but has become the dominant one by far.

Spurred by the race for controlled fusion, plasma physicists progressed rapidly on both theoretical and experimental fronts. To generate power from fusing heavy hydrogen into helium requires heating a plasma to over 10^8 K and confining it for longer than 1 s. The principles of confining plasmas with the magnetic walls needed to withstand such temperatures were worked out on the basis of classical physics concepts for both closed (toroidal) and open (mirror) systems, but the unexpectedly fast leakage out of such magnetic "traps" or "bottles" was a huge problem. To solve it required going back to basic principles and establishing a new science, which barely resembles its precursors. The instabilities that a magnetized plasma can support required new theoretical and experimental methods. An elegant variational-principle formulation of the problem in a paper by Bernstein *et al.* (1958) is generally credited with establishing the respectability of plasma physics as a new and viable field. Also in 1958, the *Physics of Fluids* was established as the plasma journal, and the Division of Plasma Physics was founded in The American Physical Society to give plasma physics a home.

When the exploration of space started in the early 1960s, a need for plasma diagnostics arose, and Langmuir probes were quickly adapted for the Earth's ionosphere and the interplanetary environment. Exploration of the plasmas in the magnetospheres of the planets, as well as of the solar wind and other solar phenomena, have greatly expanded our knowledge of the solar system in the past three decades. The advent of radio astronomy has led to observations of plasma phenomena in deep space. Most recently, the importance of dusty plasmas has been realized in the understanding of comet tails and planetary rings.

In the early 1970s, high-power pulsed lasers were invented, and the concept of inertial confinement was born. The idea was to do away with magnetic fields altogether, using laser beams to compress a small pellet of deuterium–tritium fuel to 10^4 times solid density, and hence to speed up the fusion reaction time to the order of picoseconds. In that case, the inertia of the imploding pellet would confine the fuel long enough for fusion to occur. Unfortunately, the unexpected effects of plasma physics could not be avoided. When intense laser light strikes a solid object, the surface material is immediately ionized into a plasma cloud. A new type of instability, called a parametric instability, was driven by the laser beam; and this had the effect of reflecting the laser light before it could reach the solid surface of the pellet. Inertial confinement spawned a new branch of plasma physics, which required extensive development of nonlinear and computational techniques. Simultaneously, it was found that intense electron and proton beams of order 1 MA at 1 MV could be produced by storing electrical energy in capacitor banks and discharging it into a diode in a short pulse. Pulsed power became yet another branch of plasma physics, permitting the study of matter subject to terawatts of power, and leading to the invention of new plasma switches and x-ray generators.

Until recently, most plasma-physics research fell into one of the four areas described above: magnetic fusion, inertial fusion, space plasma physics, and pulsed power. In the 1990s, the use of partially ionized plasmas in manufacturing began to expand rapidly, along with the prevalent use of semiconductor ULSI (ultralarge-scale integrated) circuits and personal computers. Plasma etching and deposition are essential in the fabrication of computer chips with millions of internal features. It is anticipated that the production of flat-panel displays will ultimately also depend on plasma processing. Surface hardening of turbine blades, tools, construction materials, medical prostheses, and so forth is already a large commercial application of plasma physics. Plasma treatment of plastics and fibrous materials can be used, for instance, to make plastic containers more leakproof; to make paint adhere to plastic bags, clothing, currency, and automobile parts; to make fabrics absorb or repel water; and to tailor the refractive indices of optical fibers. In summary, though the original impetus for developing plasma physics was the lofty goal of an inexhaustible energy source for mankind, fusion power plants will not be commercially viable for several more decades. In the near future, however, plasma physics is likely to have a large impact on our daily lives because of its utility in the manufacture of a diverse variety of common materials.

2. BASIC PLASMA PHENOMENA

2.1 Methods of Treatment

2.1.1 Fluid Description At the simplest level, a plasma can be treated as an electrified fluid, without regard for its particulate nature. A common approach is to consider the positive ions and negative electrons of the plasma to be two interpenetrating fluids, which interact via the electric and magnetic field that they each generate in their motions (Chen, 1983). The equation of motion for fluid elements of each species j can be written

$$m_j n_j\left(\frac{\partial \mathbf{v}_j}{\partial t} + \mathbf{v}_j \cdot \nabla \mathbf{v}_j\right) = q_j n_j(\mathbf{E} + \mathbf{v}_j \times \mathbf{B}) - \nabla p_j - m_j n_j \nu_{jk}(\mathbf{v}_j - \mathbf{v}_k), \quad (1)$$

where q_j and m_j are the charge and mass of species j; n_j, p_j, and $\mathbf{v}_j$ are the fluid density, pressure, and velocity of that species; and ν_{jk} is its average collision frequency with species k. The pressure p_j is given by $p_j = \gamma_j n_j K T_j$, where K is Boltzmann's constant, T_j is the temperature of the fluid in kelvins, and γ_j is an adiabatic constant equal to 3, 2, or $\frac{5}{3}$ for one-, two-, or three-dimensional plasmas, respectively. For constant T_j, γ_j is unity. A plasma can have a different temperature and even a different value of γ for each species. In a fully ionized plasma, j takes on the values i and e, while k takes the opposite value. In a partially ionized plasma with less than about 10% ionization fraction, collisions with the neutral atoms are dominant; then ν_{jk} is ν_{io} or ν_{eo}, the collision rate with the neutral gas. Each species also follows the equation of continuity,

$$\frac{\partial n_j}{\partial t} + \nabla\cdot(n_j\mathbf{v}_j) = 0. \tag{2}$$

A set of Eqs. (1) and (2) can be written for each species of positive or negative ions. The electric and magnetic fields **E** and **B** are governed by the vacuum Maxwell's equations:

$$\epsilon_0\nabla\cdot\mathbf{E} = \sigma, \tag{3}$$

$$\nabla\times\mathbf{E} = \frac{\partial\mathbf{B}}{\partial t}, \tag{4}$$

$$\nabla\cdot\mathbf{B} = 0, \tag{5}$$

$$\nabla\times\mathbf{B} = \mu_0\left(\mathbf{J} + \epsilon_0\frac{\partial\mathbf{E}}{\partial t}\right). \tag{6}$$

Together with the equation of state for p_j given above, the system of equations is closed by defining the plasma charge and current as

$$\sigma = \sum_j n_j q_j, \quad \mathbf{J} = \sum_j n_j q_j \mathbf{v}_j. \tag{7}$$

The charged-particle motions generate the internal electromagnetic fields, and the fields move the particles; thus, Eqs. (1)–(7) must be solved self-consistently. It is also possible to describe the plasma as a dielectric medium with a susceptibility χ and a permittivity $\epsilon = \epsilon_0(1 + \chi)$ that incorporate the effects of the plasma charges and currents, as given by Eqs. (1) and (2). The terms **J** and σ are then omitted from Eqs. (3) and (6), and ϵ_0 replaced by ϵ. Plasmas in a magnetic field are anisotropic, and ϵ is then a rather complicated dielectric tensor. It is a characteristic of plasmas that χ is often negative. The electromagnetic quantities **D** and **H** need not be defined and are generally not used in plasma physics.

An alternative description of a plasma is that of a single conducting fluid. The single-fluid, or MHD (magnetohydrodynamic), equations of motion are linear combinations of Eq. (1) for electrons and Eq. (1) for a single species of ions with charge number $Z = 1$, with a number of terms neglected:

$$\rho\frac{\partial\mathbf{v}}{\partial t} = \mathbf{J}\times\mathbf{B} - \nabla p, \tag{8}$$

$$\mathbf{E} + \mathbf{v}\times\mathbf{B} = \eta\mathbf{J}, \tag{9}$$

where the single-fluid density ρ, velocity $\mathbf{v}$, pressure p, and resistivity η are defined by

$$\rho = \sum_{j=i,e} m_j n_j, \quad \mathbf{v} = \sum_{j=i,e} n_j M_j \mathbf{v}_j/\rho,$$
$$J = \sum_{j=i,e} n_j q_j \mathbf{v}_j, \quad p = p_i + p_e, \quad \eta = m_e\nu_{ei}/n_e e^2. \tag{10}$$

Equation (9) is known as the *generalized Ohm's law.* Equations (2) for ions and electrons are replaced by

$$\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}) = 0, \quad \frac{\partial\sigma}{\partial t} + \nabla\cdot\mathbf{J} = 0. \tag{11}$$

Note that the quasineutrality condition $n_i = n_e$ used here is almost always a good approximation because only a small difference $\sigma = e(n_i - n_e)$ is enough to create a sizable field **E**. The single-fluid equations are convenient for resistive plasmas, in which the dominant collisions are between ions and electrons, but the equations are less accurate than the two-fluid equations (for instance, the Hall current is neglected in the generalized Ohm's law), and multiple species are not easily handled.

2.1.2 Kinetic Description The fluid description is valid when collisions are frequent enough to maintain a Maxwell–Boltzmann velocity distribution for each species, so that a temperature T can be defined. Fluid theory is also a good approximation for motions perpendicular to a strong magnetic field even when collisions are rare. In the general case, deviations from a Maxwellian distribution are possible, and one has to solve for the distribution function $f_j(\mathbf{r},\mathbf{v},t)$ for each species. A function of seven independent scalar variables, $f_j(\mathbf{r},\mathbf{v},t)$ is governed by the *Boltzmann equation*

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla f + \frac{q}{m}(\mathbf{E} + \mathbf{v} \times \mathbf{B})\cdot\frac{\partial f}{\partial \mathbf{v}} = \left(\frac{\partial f}{\partial t}\right)_c, \tag{12}$$

where $(\partial f/\partial t)_c$ is a collision term. "Collisionless" plasmas are hot enough that this term can be set equal to zero, in which case Eq. (12) is called the *Vlasov equation.* Such "Vlasov plasmas" have been studied intensively by plasma theorists (Nicholson, 1983). Other forms of the collision term are the *Fokker–Planck* term, which describes binary Coulomb collisions, and the *Krook* term, which is used to treat collisions with a neutral species. A simplified form of Eq. (12) in which the Larmor gyrations (cf. Sec. 2.2.3) have been averaged and replaced by the motion of a guiding center is called the *drift-kinetic equation.*

Since the integral of $f(\mathbf{v})$ over all velocities is just the density, the continuity equation [Eq. (2)] can be recovered by integrating the Vlasov equation. The first moment, found by multiplying Eq. (12) by $\mathbf{v}$ and then integrating, yields Eq. (1). The second moment gives the heat-flow equation, and so on; but the process is usually truncated after the first moment. Solutions of Eq. (12) with Maxwell's equations reveal two main features not found in Eqs. (1) and (2). First, electric fields can distort $f(\mathbf{v})$ away from a Maxwellian in the direction parallel to a magnetic field or when there is no magnetic field. Second is the finite–Larmor-radius (FLR) effect, in which the thermal motions of the particles cause them to react to the average field that they experience during a Larmor, or cyclotron, orbit. The fluid equation, Eq. (1), correctly describes the Larmor orbits of *cold* fluids driven by electric fields, but the pressure term ∇p_j does not include all the thermal effects; in particular, the FLR effect is contained in the viscosity tensor, which has been omitted from Eq. (1).

Though plasmas can often be treated as fluids, plasma physics differs from hydrodynamics because non-Maxwellian distributions sometimes occur, which require the kinetic treatment described above; it differs also from electromagnetics because the details of the dielectric tensor ϵ are treated from the particulate point of view. Indeed, the complex motions of charged particles in electric and magnetic fields support a rich variety of wave phenomena that do not occur in ordinary fluids or dielectrics.

2.1.3 Computer Simulation Small-amplitude waves and instabilities can be analyzed by linearizing the plasma equations given above. That is, each fluctuating quantity—for instance, the fluid density n—is divided into an undisturbed, time-independent part and an oscillating part: $n = n_0 + n_1$, where n_1 is assumed to be small relative to n_0. When higher-order terms containing more than one power of fluctuating quantities are discarded, the resulting set of linear equations can be solved to give the possible wave modes. These linear modes can be superimposed without affecting one another. Large-amplitude waves, and instabilities that have grown large, however, are difficult to treat with these equations. In the early years of plasma-physics history, there was a gap between the small-amplitude phenomena accessible by theory and the large-amplitude phenomena observed experimentally. The advent of high-speed computers provided the Rosetta Stone, particle simulation (Dawson and Lin, 1984), needed to relate experiment to theory.

In computer simulation (Hockney and Eastwood, 1988; Birdsall and Langdon, 1991), a large number of particles—typically 10^4 but as large as 10^6—are placed in a "box," and the position and velocity of each ion and electron are recorded in the memory. The **E** and **B** fields caused by these particles are then found from Maxwell's equations by fast Fourier transform. The particles are then allowed to move in response to these fields during a short time step, and the fields are then recalculated with the new positions and velocities. Thus, the response of the plasma to applied forces or unstable initial conditions can be followed step by step from threshold to turbulence regardless of the amplitudes of the perturbations. The boundaries of the "box" can be made to absorb particles, to reflect particles, or to be reentrant or periodic so that particles leaving one side reappear on the opposite side. To facilitate computation, the box is divided into a grid of cells, and a particle that ends up in any particular cell is placed at the center of that cell for the next time step. Such a scheme is called a particle-in-cell, or PIC, code. Since each computer particle represents many real particles, it cannot be given the full electric charge of all those particles, for then the Coulomb collisions between particles would be unduly magnified. These collisional effects have to be properly scaled.

The number of particles, cells, and time steps in the simulation is limited only by the speed and memory of the computer. Programs range from those that can be run on personal computers (PCs) to those that require many hours or days on the most advanced supercomputers. Initially, computer simulation was limited to one-dimensional systems, in which each "particle" was merely a charge sheet. Later, two-dimensional grids could be used, with axes corresponding, for instance, to (x,y) in a Cartesian system or (r,z) in a cylindrical system. A common compromise is a so-called $2\frac{1}{2}$D computation, in which two components of position and all three components of velocity are retained. Fully three-dimensional simulations are now possible. Once a computational run has been finished, it is easy to do diagnostics on the plasma. For instance, one can cull from the data the distribution functions as a function of space or time, the frequency or wavenumber spectrum of the fluctuations, the conservation of energy and momentum, or the direction of the Poynting vector. The results of such computer "experiments" can be used to compare with real measurements, but plasma simulations can be formulated and interpreted only with the insight gained from an intimate knowledge of analytic theory (see SIMULATION AND MODELING BY COMPUTER).

2.2 Characteristics of Plasma Behavior

2.2.1 Quasineutrality and Debye Shielding Plasmas are characterized by an overriding tendency to be neutral. The numbers of positive and negative charges per unit volume cannot be exactly equal, of course; else there would be no electric fields at all. Since the opposing charge densities need differ only by, say, one part per million to create the electric fields normally observed in a plasma, for all intents and purposes these densities can be treated as being equal. This approximation, known as "quasineutrality," is an important concept in plasma physics. The consequences of quasineutrality may not be apparent to physicists from other disciplines. For instance, at frequencies low enough that the heavy and light particles can move together, **E** cannot be calculated from Poisson's equation, Eq. (3), since $n_i - n_e$ is not known accurately enough; rather, Eqs. (1) and (2) determine the value of **E** because the ions and electrons must move in such a way as to make their densities track each other. The density n_i or n_e is simply known as the plasma density n.

A plasma shields itself from external fields by *Debye shielding*. When a positively charged object is placed in a plasma, the mobile electrons are attracted to it and form an electron cloud covering the charge; similarly, a negative object will repel electrons, leaving an ion cloud for shielding. The thickness of these clouds is of the order of the Debye length λ_D, defined by

$$\lambda_D^2 = \epsilon_0 KT_e/ne^2. \tag{13}$$

The shielding is not perfect, however, since electrons near the edge of the cloud can escape because of their thermal energies; consequently, electric potentials of order $\frac{1}{2}KT_e$ can exist in the plasma. [The temperatures of a plasma are specified by the value of KT in electron volts (eV), where 1 eV corresponds to 11 600 K. Thus, the average thermal energy of the electrons in a plasma with $KT_e = 10$ eV is $\frac{3}{2}KT_e$, or 15 eV.] Debye shielding also occurs at the walls confining a plasma, where an ion-rich sheath of about $5\lambda_D$ thickness forms to repel electrons. The Coulomb barrier of the sheath is just high enough to make the escaping electron flux equal to the ion flux, so that no net charge is lost. Unless the plasma is extremely tenuous, these sheaths are much thinner than the dimensions of the plasma, so that the interior is quasineutral.

2.2.2 Characteristic Frequencies and Velocities When the electrons of a plasma are displaced from their equilibrium positions, the ions left behind draw them back with an electric field. The electrons then overshoot and oscillate about their equilibrium positions with a characteristic frequency ω_p, called the *plasma frequency,* which depends on the density of electrons as well as their charge and mass:

$$\omega_p = (ne^2/\epsilon_0 m)^{1/2}\ \text{rad/s}, \quad f_p = \omega_p/2\pi\ \text{Hz}. \tag{14}$$

An *ion plasma frequency* ω_{pi} is similarly defined with the ion charge and mass but is not a natural oscillation frequency because the ions are shielded by Debye shielding.

When a background magnetic field B is present, the ions and electrons gyrate in circular orbits, called Larmor orbits, in a plane perpendicular to the field lines. The frequency of these gyrations is the cyclotron frequency ω_c, given for species j by

$$\omega_{cj} = |q_jB/m_j|, \quad f_{pj} = |q_jB/2\pi m_j|, \tag{15}$$

where q_j is $-e$ for electrons and Ze for ions. The radius of the orbit—the *Larmor radius*—is given by

$$r_L = v_\perp/\omega_c = mv_\perp/|q|B, \tag{16}$$

where $v_\perp$ is the velocity component perpendicular to **B**. For a Maxwellian distribution, the root mean square value of $v_\perp$ is used to evaluate the average r_L, where

$$\langle v^2\rangle = 3KT/m, \quad \langle v_\perp^2\rangle = 2KT/m. \tag{17}$$

Charged particles are constrained from moving freely across **B** and act as if they were tied to the field lines. When these mass-loaded field lines are displaced, the disturbance propagates along the field with a characteristic velocity known as the *Alfvén speed* v_A:

$$v_A = B/\sqrt{\mu_0\rho}. \tag{18}$$

This behavior can occur only if the field B is strong enough to constrain the particles—that is, if the plasma pressure is sufficiently small. The critical quantity is the ratio between plasma pressure and magnetic field pressure, known as "beta":

$$\beta = \frac{p_i + p_e}{B^2/2\mu_0}. \tag{19}$$

Laboratory plasmas have $\beta \ll 1$, but gravitationally confined plasmas in space may have $\beta \geq 1$.

Sound waves in a plasma, called *ion-acoustic waves,* travel at a characteristic velocity v_s given by

$$v_s^2 = (ZKT_e + \gamma_i KT_i)/m_i. \tag{20}$$

Unlike sound waves in air, this velocity does not vanish when $KT_i = 0$. The wave motion is sustained not by pressure fluctuations but by the electric fields, of order KT_e, which extend beyond the Debye clouds shielding the ion-density fluctuations.

2.2.3 Particle and Fluid Drifts In a strong magnetic field **B**, the parallel motions of the particles are unaffected by **B**, while the perpendicular motions can be considered to be circular Larmor gyrations about a moving center, called a *guiding center.* In crossed electric and magnetic fields, the guiding center drifts with a velocity $\mathbf{v}_E$ given by (Kruskal, 1965, p. 67)

$$\mathbf{v}_E = \mathbf{E} \times \mathbf{B}/B^2. \tag{21}$$

This velocity is the same for all species, and this is true to the extent that the FLR effect can be neglected. If we replace the force $q\mathbf{E}$ used in deriving Eq. (21) with a general force **F**, the resulting guiding-center drift $\mathbf{v}_f$ would be in opposite directions for ions and electrons:

$$\mathbf{v}_f = \mathbf{F} \times \mathbf{B}/qB^2. \tag{22}$$

If the magnitude of B is nonuniform, the guiding center suffers a *gradient-B drift* $\mathbf{v}_{\nabla B}$:

$$\mathbf{v}_{\nabla B} = \pm v_\perp r_L\, \mathbf{B} \times \nabla B/2B^2, \tag{23}$$

where ± indicates the sign of the charge. If the direction of **B** is not constant, one can define a local vector radius of curvature $\mathbf{R}_c$ pointing toward the center of curvature; the guiding center then has a *curvature drift* $\mathbf{v}_R$ given by

$$\mathbf{v}_R = \frac{mv_\parallel^2}{q}\frac{\mathbf{R}_c \times \mathbf{B}}{R_c^2 B^2}. \tag{24}$$

When the electric field varies in time, the Larmor gyrations are disrupted, and the guiding center suffers a *polarization drift* $\mathbf{v}_p$ given by

$$\mathbf{v}_p = \pm\frac{1}{\omega_c B}\frac{d\mathbf{E}}{dt}. \tag{25}$$

In fluid theory, the drifts of the fluid elements are similar to the guiding-center drifts, but with some differences. The drifts $\mathbf{v}_E$, $\mathbf{v}_R$, and $\mathbf{v}_p$ are the same; but the gradient-B drift $\mathbf{v}_{\nabla B}$ cancels out when summed over the particles in a fluid element. The summation process gives rise to a fluid drift, called the *diamagnetic drift* $\mathbf{v}_D$, which does not exist for particles (Spitzer, 1967):

$$\mathbf{v}_D = \mathbf{B} \times \nabla p/qnB^2. \tag{26}$$

Since $\mathbf{v}_D$ is in a different direction for ions and electrons, there is a net *diamagnetic current* $\mathbf{j}_D$ whose $\mathbf{j} \times \mathbf{B}$ force is just sufficient to keep the ∇p force from pushing the plasma across $\mathbf{B}$, as seen from Eq. (8).

Because of the electrons' high mobility along $\mathbf{B}$, the net parallel force on the electron fluid must be close to zero except at high frequencies. The dominant terms in Eq. (1) are then

$$0 = en_e\nabla_\parallel\phi - KT_e\nabla_\parallel n_e, \tag{27}$$

where we have taken $E_\parallel = -\nabla\phi$ and $\gamma_e = 1$. For $n_e = n_0 + n_1$, integration of Eq. (27) gives the *Boltzmann relation*

$$n_1 = n_0 \exp(e\phi/KT_e), \tag{28}$$

which has wide applicability and is useful for relating density to potential fluctuations.

2.2.4 Adiabatic Invariants

In addition to energy and momentum, the action of a particle in periodic motion is conserved, permitting a simple way to describe the motion of gyrating particles in slowly varying fields (Northrop, 1963; Kruskal, 1965, p. 91). The first such *adiabatic invariant* is the magnetic moment μ, defined by

$$\mu = mv_\perp^2/2B. \tag{29}$$

This quantity, equal to the magnetic flux enclosed in a Larmor orbit, remains constant if B changes in time or if the guiding center moves to a region of different B. The invariance of μ is broken if the field seen by the particle changes on the time scale $1/\omega_c$ or if B varies spatially over a Larmor radius. The second adiabatic invariant J, defined by

$$J = \int v_\parallel ds, \tag{30}$$

concerns the motion of a gyrating particle between two "magnetic mirrors" (Fig. 1). When such a particle moves into a region of stronger B, conservation of μ causes $v_\perp$ to increase. Since $v_\parallel$ must then decrease to conserve energy, the particle is reflected. If the bounce frequency of a particle trapped between two regions of strong field, such as the polar regions of the Earth's magnetic field, is larger than the rate of change of B, the integral J evaluated between turning points is conserved. Other adiabatic invariants involve slower periodic motions and are rarely used.

2.2.5 Resistivity and Diffusion

2.2.5.1 Weakly Ionized Plasmas. Driven by a pressure gradient or an electric field, electrons and ions in a weakly ionized discharge have their velocities along $\mathbf{B}$ limited by collisions with the constant neutral-gas background. For the isothermal case $\gamma_j = 1$, solution of Eq. (1) with the $\mathbf{v} \times \mathbf{B}$ term and the left-hand side set equal to zero gives the particle flux (Chen, 1983):

$$n_j v_{\parallel j} = \mu_j n_j E_\parallel - D_{\parallel j}\nabla_\parallel n_j, \tag{31}$$

where

$$\mu_{\parallel j} = q_j/m_j\nu_{0j}, \quad D_{\parallel j} = KT_j/m_j\nu_{0j} \quad (j = i,e), \tag{32}$$

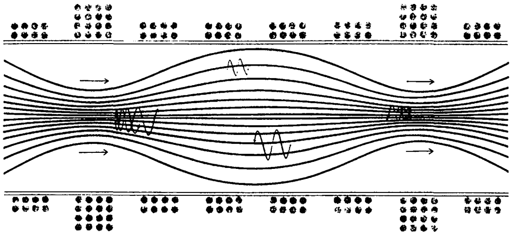

FIG. 1. Charged particles gyrate around magnetic field lines and are trapped between strong-field regions by the magnetic mirror effect, which results from the adiabatic invariant μ. The second adiabatic invariant, J, controls the frequency of the longitudinal bounce motion between mirrors (Fowler and Post, 1966).

μ is the mobility, and D the diffusion coefficient. In the direction perpendicular to **B**, inclusion of the $\mathbf{v} \times \mathbf{B}$ term in Eq. (1) results in

$$n_j\mathbf{v}_{\perp j} = \mu_j n_j \mathbf{E}_\perp - D_{\perp j}\nabla_\perp n_j + \frac{\mathbf{v}_E + \mathbf{v}_{Dj}}{1 + \nu_{0j}^2/\omega_{cj}^2}, \qquad (33)$$

where

$$\mu_{\perp j} = \frac{\mu_{\parallel j}}{1 + \omega_{cj}^2\tau_j^2}, \quad D_{\perp j} = \frac{D_{\parallel j}}{1 + \omega_{cj}^2\tau_j^2}, \quad \tau_j = \frac{1}{\nu_{0j}}, \qquad (34)$$

and $\mathbf{v}_E$ and $\mathbf{v}_D$ are given by Eqs. (21) and (26). We see that $\mathbf{v}_E$ and $\mathbf{v}_D$ are slowed by collisions, whereas cross-field mobility and diffusion exist only by virtue of collisions.

These relations have a simple physical meaning. Ordinary diffusion is a random walk with a step length $L = \lambda_m$, the mean free path, and mean time $\tau = 1/\nu$ between steps. For particles of velocity v, $\lambda_m = 1/n_n\sigma(v) = v\tau$, where n_n is the neutral density. The collision frequency ν_0 averaged over velocities is thus $\nu_0 = n_n\langle\sigma v\rangle$. *The diffusion coefficient is then* $\mathbf{D} = \mathbf{L}^2/\tau = v^2\tau^2/\tau \approx KT/m\nu_0$. Mobility is like diffusion except that the force $qn\mathbf{E}$ takes the place of the pressure force $KT\nabla n$, so that the coefficient of **E** should be q/KT times the coefficient of $\nabla n/n$, as in Eq. (32). For perpendicular diffusion across a magnetic field, a collision displaces the guiding center a distance of order r_L or λ_m, whichever is smaller. Taking $L^{-1} = r_\text{L}^{-1} + \lambda_m^{-1}$ gives the reduction factors in Eq. (34) for perpendicular transport.

Since ions and electrons have different transport coefficients, an *ambipolar electric field* will develop to retard the faster species and accelerate the slower species. In the direction along **B**, electrons will escape rapidly, leaving behind a positive charge the field of which equalizes the electron and ion escape rates to maintain quasineutrality. The net diffusion rate, found from Eqs. (31) and (32), is described by an *ambipolar diffusion coefficient* D_a:

$$D_a = \frac{\mu_i D_e + \mu_e D_i}{\mu_i + \mu_e}. \qquad (35)$$

For diffusion across **B**, the same formula obtains if the definitions in Eq. (34) are used. In this case, the ambipolar field is reversed, since ions generally escape faster than electrons across **B**. In the general case, diffusion in a given density gradient is described by the equation $\nabla\cdot(n\mathbf{v}_i) = \nabla\cdot(n\mathbf{v}_e)$, with the parallel and perpendicular fluxes given by Eqs. (31) and (33), respectively. Though this equation is not separable into perpendicular and parallel components, in principle the solution will yield the ambipolar potential ϕ, which is the only unknown in the problem. Unfortunately, the problem is not well posed because of the *short-circuit effect.* In plasmas whose length along **B** is not extremely large relative to the radius, electrons escaping along field lines can short-circuit the radial ambipolar field through the sheaths at the ends. The ambipolar fields in the axial and radial directions then depend on the details of the end conditions.

2.2.5.2 Fully Ionized Plasmas. Because of detailed balance, Coulomb collisions between like particles—for instance, electrons with electrons—cause neither a drag of the fluid motion along **B** nor a net diffusion of the fluid across **B** unless there is a large gradient on the scale of r_L. Collisions between electrons and ions cause friction between the two fluids, giving rise to a resistivity $\eta_\parallel$ for current flow along **B**. The value of $\eta_\parallel$, called the *Spitzer resistivity* (Spitzer, 1967), to be used in Eq. (9) for $E_\parallel$, is approximately

$$\begin{aligned}\eta_\parallel &= 5.2 \times 10^{-5} Z \ln\Lambda/T_e^{3/2}(\text{eV})\ \Omega\text{m},\\ \ln\Lambda &\approx 31 - \ln(n_e^{1/2}/T_e).\end{aligned} \qquad (36)$$

Here Z is the ion charge number, n_e is in m^{-3}, and T_e is in electron volts. The quantity $\ln\Lambda$, called the *Coulomb logarithm,* is unique to plasma physics (Spitzer, 1967). In averaging the scattering of electrons by ions over all velocities and impact parameters h, one chooses the minimum value of h to be that for a 90° deflection (to avoid the singularity at $h = 0$), and the maximum value of h to be λ_D, to account for Debye shielding. Λ is the ratio $h_\text{max}/h_\text{min}$ and $\ln\Lambda$ varies so slowly that it can be taken as 10 in most cases. The Spitzer resistivity is nearly independent of density and falls rapidly with temperature. At temperatures of 1 keV and above, the value of $\eta_\parallel$ is so small that, according to Eq. (9), $E_\parallel$ must vanish. An ideal, collisionless plasma is a superconductor along **B** and a superinsulator across **B**.

Particles in a fully ionized plasma can move across a magnetic field only by electron–ion collisions. At each collision, the guiding center can be displaced a distance of order $r_L \propto v_\perp/B$. In the random walk, the net displacement is proportional to r_L^2, or $\propto KT/B^2$. Because of momentum conservation at each collision, ions and electrons move across **B** at the same rate; hence, there is a transverse diffusion but no transverse mobility. Similar considerations govern the transport of heat (Braginskii, 1965). The flux of either species is given by

$$nv_\perp = -\mathbf{D}_\perp \nabla \mathbf{n}, \quad \mathbf{D}_\perp = \frac{\eta_\perp n}{B^2} \sum_j KT_j \equiv D_c, \qquad (37)$$

where $\eta_\perp \approx 2\eta_\parallel$, differing from $\eta_\parallel$ only in the way different velocities are weighted. The *classical diffusion coefficient* D_c depends on density, making the diffusion problem nonlinear, and decreases with temperature because of the $T_e^{-3/2}$ dependence of $\eta_\perp$. The value of D_c is so small that it is rarely observed; plasmas normally escape from magnetic traps by *anomalous diffusion* caused by instabilities.

2.2.6 Collisionless Damping Plasmas are the only substances in which dissipation of energy can occur without collisions. This phenomenon, discovered by Landau (1946) in the course of a rigorous solution of the Vlasov equation for the case of linear electron plasma waves, is called *Landau damping*. An apparent damping of these waves is found when Eq. (12) is Fourier analyzed and solved by contour integration, the damping appearing as an imaginary part of the complex frequency ω resulting from the residue around a pole in its complex plane (Jackson, 1960). The physical mechanism was subsequently explained by Dawson (1961). Some electrons in the distribution (the *resonant electrons*) have a thermal velocity close to the phase velocity of the wave, and those that are in the accelerating part of the wave electric field can stay in synchronism long enough to gain energy from the wave. There are also electrons in the decelerating phase of the wave, but for a Maxwellian distribution these are fewer in number; and, hence, there is a net gain of particle energy and a loss of wave energy. In the absence of collisions, the electrons accelerated by this wave–particle interaction maintain their energies. In carefully designed echo experiments, this energy gain can be reversed, and an echo of the wave can be seen after the wave has damped away.

The plasma dispersion function $Z(\zeta)$ is defined by (Fried and Conte, 1961)

$$Z(\zeta) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{\exp(-s^2)}{s - \zeta}\, ds, \quad \mathrm{Im}(\zeta) > 0, \qquad (38)$$

where the integral is to be taken along the real s axis but passing below all the poles of the integrand. For linear plasma waves of frequency ω and wave number k, solution of the Vlasov equation for $B = 0$ gives the dispersion relation

$$2k^2\lambda_D^2 = Z'(\zeta), \quad \zeta = \omega/kv_{\mathrm{th}}, \qquad (39)$$

where $Z'(\zeta)$ is the derivative of $Z(\zeta)$ with respect to its argument, and the electron thermal velocity v_{th} is $(2KT_e/m)^{1/2}$. Equation (39) gives the Landau damping rate

$$[\mathrm{Re}(\omega)]^2 \approx \omega_p^2 + \tfrac{3}{2}k^2v_{\mathrm{th}}^2,$$
$$\mathrm{Im}(\omega) \approx -\omega_p \sqrt{\frac{\pi}{e^3}} \left(\frac{\omega_p}{kv_{\mathrm{th}}}\right)^3 \exp[-(\omega_p/kv_{\mathrm{th}})^2], \qquad (40)$$

which has been confirmed experimentally.

2.3 Plasma Waves and Instabilities

A rich variety of coherent wave motions is possible in a plasma because of its electromagnetic, anisotropic, and particulate nature. When a source of free energy is available, some of these waves can grow spontaneously and are called *instabilities* (Fig. 2). Most waves can be treated with fluid theory, but *kinetic* waves and instabilities involve distortions of the velocity distribution and must be treated with kinetic equations. The most common waves in a plasma with electrons and a single species of singly charged ions are summarized below. These simple results are modified by multiple species, finite boundaries, and damping (Allis *et al.*, 1963; Swanson, 1989; Stix, 1992). For further details, see PLASMA WAVES AND INSTABILITIES.

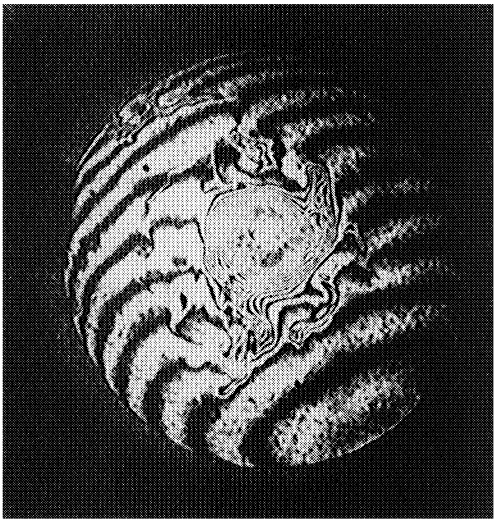

FIG. 2. A plasma instability is revealed by laser interferometry of an initially circular cylinder of dense plasma (Fowler and Post, 1966).

2.3.1 Electrostatic Waves When there is no magnetic field, two electrostatic waves exist:

1. a high-frequency *electron plasma (Bohm-Gross) wave* with a frequency given by Eq. (40), reducing for $T_e \to 0$ to a *Langmuir oscillation,* Eq. (14); and
2. a low-frequency ion-acoustic wave with a constant velocity $\omega/k = v_s$, given by Eq. (20).

These waves can also propagate *along* a background magnetic field $\mathbf{B}_0$; in addition, there are three new waves propagating *across* $\mathbf{B}_0$. At high frequencies, electrons undergo an upper-hybrid oscillation, with $\omega = \omega_{\text{uh}}$; at low frequencies there is an electrostatic ion cyclotron wave at ω_{esic}; and at intermediate frequencies, there is a lower-hybrid oscillation at ω_{lh}:

$$\omega_{\text{uh}}^2 = \omega_p^2 + \omega_c^2, \quad \omega_{\text{esic}}^2 = \omega_{ci}^2 + k^2c_s^2,$$
$$\omega_{\text{lh}}^{-2} = \omega_{pi}^{-2} + (\omega_{ce}\omega_{ci})^{-1}. \tag{41}$$

The hybrid frequencies are commonly occurring combinations of intrinsic frequencies. For propagation in a direction $\mathbf{k}$ at an angle θ relative to $\mathbf{B}_0$, the two high frequencies are given by the solutions of

$$\omega^2(\omega^2 - \omega_{\text{uh}}^2) + \omega_p^2\omega_c^2 \cos^2\theta = 0. \tag{42}$$

When these waves are confined to a conducting cylinder of radius a, the oscillating potentials of the eigenmodes have the form $\varphi_1 = AJ_m(p_{mn}r/a)e^{i(m\theta+\beta z-\omega t)}$, and their frequencies are given by the frequently used *Trivelpiece–Gould* formula:

$$\beta a = \pm p_{mn}\left[\frac{\omega^2(\omega_{\text{uh}}^2 - \omega^2)}{(\omega^2 - \omega_p^2)(\omega^2 - \omega_c^2)}\right]^{1/2}, \tag{43}$$

where p_{mn} is the nth root of the Bessel function J_m. As the axial wave number β is increased, the frequency of one branch increases from 0 to the lower of ω_p and ω_c, while that of the other branch decreases from ω_{uh} to the higher of ω_p and ω_c.

Examples of waves not describable by fluid equations are the electron and ion *Bernstein waves,* whose frequencies are near the harmonics of ω_{ce} and ω_{ci}, respectively. Since these depend on the FLR effect, they are found in solutions of the Vlasov equation.

2.3.2 Electromagnetic Waves Light waves and microwaves propagate in an unmagnetized plasma with a frequency

$$\omega^2 = \omega_p^2 + c^2k^2, \tag{44}$$

corresponding to a dielectric constant

$$\epsilon = c^2k^2/\omega^2 = 1 - \omega_p^2/\omega^2. \tag{45}$$

Transmission is cut off for $\omega < \omega_p$. In the presence of a field $\mathbf{B}_0$, light waves propagating *across* $\mathbf{B}_0$ polarized with $\mathbf{E} \parallel \mathbf{B}_0$ also obey Eq. (44) and are called *ordinary* waves. Those with $\mathbf{E}\perp\mathbf{B}_0$ have the dispersion relation

$$\epsilon = 1 - \frac{\omega_p^2}{\omega^2}\frac{\omega^2 - \omega_p^2}{\omega^2 - \omega_{\text{uh}}^2} \tag{46}$$

and are called *extraordinary* waves. These waves have a "resonance," where $\epsilon \to \infty$, at $\omega = \omega_{\text{uh}}$, and two "cutoffs," where $\epsilon \to 0$, at the *right-* and *left-hand cutoff frequencies* ω_R, ω_L given by

$$\omega_R = \tfrac{1}{2}[(\omega_c^2 + 4\omega_p^2)^{1/2} + \omega_c],$$
$$\omega_L = \tfrac{1}{2}[(\omega_c^2 + 4\omega_p^2)^{1/2} - \omega_c]. \tag{47}$$

Extraordinary waves are reflected at the cutoffs and have forbidden bands for $0 < \omega < \omega_L$ and $\omega_{\text{uh}} < \omega < \omega_R$. The absorption of these waves at the resonance is a method for heating a plasma.

Light waves propagating *along* $\mathbf{B}_0$ have normal modes that are circularly polarized, with dielectric constants

$$\epsilon = 1 - \frac{\omega_p^2/\omega^2}{1 \pm \omega_c/\omega}. \tag{48}$$

Since these two components have different phase velocities, plane-polarized electromagnetic waves along $\mathbf{B}_0$ exhibit *Faraday rotation.* The left-hand circularly polarized mode (upper sign) is cut off below ω_L, and the right-hand mode is cut off between ω_c and ω_R. Below ω_c, the right-hand mode has a low-frequency propagation band in which the waves are called *whistler waves,* named after their gliding tones after passing through the earth's ionosphere.

Low-frequency electromagnetic waves *along* $\mathbf{B}_0$ are called *Alfvén* waves and have the velocity v_A given by Eq. (18). Such hydromagnetic waves propagating *across* $\mathbf{B}_0$ are called magnetosonic waves, or simply "fast" waves and have a velocity

$$v_{ms}^2 = c^2 \frac{v_s^2 + v_A^2}{c^2 + v_A^2} \approx v_s^2 + v_A^2. \tag{49}$$

This wave is useful for heating a plasma. Another wave used for this purpose is the *electromagnetic ion-cyclotron* wave, with a frequency given by

$$\omega_{\text{emic}}^2 = \omega_{ci}^2 \left[1 + \frac{\omega_{pi}^2}{c^2}\left(\frac{1}{k_\parallel^2} + \frac{1}{k^2}\right)\right]^{-1}, \tag{50}$$

where k is the total wave number and $k_\parallel$ is its component along $\mathbf{B}_0$.

2.3.3 Electrostatic Instabilities All the waves of Sec. 2.3.1 can be excited by beams of electrons or ions whose velocity component in the direction of the wave is near the phase velocity (Mikhailovskii, 1974). Beam–plasma interactions can involve any combination of beams and waves. The growth rate of the instability varies typically as the cube root of the beam density. The Buneman instability occurs when the electron and ion fluids of a plasma are both cold and drift through each other.

Current-driven instabilities, on the other hand, involve drifts that are small relative to the thermal velocity. In the *ion-acoustic instability,* an ion wave is excited by the relative motion of a warm electron fluid with a threshold drift velocity that varies from c_s when $T_i \ll T_e$ to v_{the} when $T_i \approx T_e$. In the presence of a magnetic field, the *electrostatic ion-cyclotron instability* [cf. Eq. (41)] has a lower threshold. An intermediate case occurs when a warm beam is added to a thermal-electron population, resulting in a *bump-on-tail* distribution. Resonant electrons can then excite plasma oscillations by *inverse Landau damping.* In addition to this particle effect, there is also the basic hydrodynamic effect, of which the Buneman instability is the zero-temperature limit.

In crossed electric and magnetic fields, electrons and ions can have slightly different $\mathbf{E} \times \mathbf{B}$ drifts due to small effects such as collisions with neutrals [cf. Eq. (33)]. When $\mathbf{E}$ and ∇n_0 are in the same direction, this drift causes an *E × B instability* that is important in the laboratory and in space (Simon, 1963). As in ordinary fluids, velocity shear gives rise to *Kelvin–Helmholtz* instabilities (D'Angelo and von Goeler, 1966; Jassby, 1972); in plasmas, these are modified by the magnetic field. Sources of free energy on which instabilities feed can occur without beams or drifts. In plasmas confined by magnetic mirrors, the velocity distribution can be anisotropic, leading to *velocity-space* or *loss-cone* instabilities (Hasegawa, 1975). Any confined plasma must deviate from complete thermal equilibrium, since it must have a pressure gradient, at least at the edge. This gradient drives drift-wave instabilities with a velocity given by v_{De} [Eq. (26)], and in the collisionless case these are called *universal instabilities* (Kadomtsev, 1965).

2.3.4 Hydromagnetic Instabilities The principal hydromagnetic instabilities (Bateman, 1978; Freidberg, 1987), in which the magnetic field fluctuates, are the *Rayleigh–Taylor instability* and the *kink instability.* These are the primary ones to be considered in designing a fusion confinement device. As in hydrodynamics, a Rayleigh–Taylor instability occurs when a heavy fluid is supported by a light fluid. In plasma physics, the heavy fluid is the plasma, and the light fluid is the massless magnetic field. The role of gravity is played by the centrifugal force on particles following curved magnetic-field lines in their thermal motions. This instability is also called the *gravitational* or *interchange* instability. The ripples that grow have a velocity of the order of v_R [Eq. (24)] and allow the plasma to escape across **B**. At low densities, this instability is electrostatic in nature. The kink instability is driven by currents along $\mathbf{B}_0$. These currents themselves generate a magnetic field $\mathbf{B}_1$ that is perpendicular to $\mathbf{B}_0$, causing the total field to twist into a helix, which then grows. The *sausage instability* is a special case of the kink instability that has azimuthal symmetry; the field lines are pinched into constrictions resembling the links of a sausage.

2.4 Radiation in Plasmas

Most plasmas in the laboratory and in space have densities below 10^{21} m^{-3}, or less than 10^{-4} of atmospheric density. At these densities, the optical depth for most radiation is much larger than the plasma. Consequently, though the particles of a plasma can be in thermal equilibrium, radiation in a plasma is rarely in equilibrium with the plasma and will pass through it freely unless the frequency strikes a particular resonance with a characteristic frequency of the plasma or with an atomic or molecular transition.

2.4.1 Bremsstrahlung The acceleration of electrons when they collide with ions generates *bremsstrahlung* photons, which have energies of the order of thermal-electron energies (Bekefi, 1966). In plasmas with KT_e in the 100–10 000-eV range, bremsstrahlung is in the soft x-ray regime and passes readily through the plasma. X-ray pinhole cameras are often used for imaging the internal features of the plasma. Further information on this appears under FUSION, MAGNETIC CONFINEMENT.

2.4.2 Synchrotron Radiation In a magnetic field, gyrating electrons emit synchrotron radiation at ω_c and its harmonics, in the submillimeter-wave regime. Since the fundamental frequency ω_c is resonant with the gyrofrequency of other electrons of any energy, this radiation is easily reabsorbed before it leaves the plasma. The higher harmonics, however, depend on distortions of the circular orbits due to relativistic mass increase or field inhomogeneities, and they can escape reabsorption. Thus the plasma radiates like a black body at ω_c and its first few harmonics but is transparent to the high harmonics. The transition harmonic number depends on the details (Dawson, 1981).

2.4.3 Scattering of Radiation *Thomson scattering* of laser light occurs when the electric field of the light beam accelerates the electrons of a plasma, and the electrons reradiate in other directions and at different frequencies (Bekefi, 1966). The ratio of scattered power to incident power is proportional to $r_0^2 n_e$, where $r_0 = e^2/mc^2$ is the classical electron radius and n_e is the electron density. The frequency spectrum of the scattered light depends on the parameter $\alpha = 1/k\lambda_D$, k being the magnitude of the wave vector **k** doing the scattering in the plasma. For incident ($\mathbf{k}_0$) and scattered ($\mathbf{k}_s$) waves at a relative angle θ, **k** is given by $\mathbf{k} = \mathbf{k}_0 - \mathbf{k}_s$, so that $k^2 = k_0^2 + k_s^2 - 2k_0k_s \cos\theta$. Incoherent scattering from individual electrons occurs when $\alpha \ll 1$, since the scattering wavelength is smaller than the Debye length λ_D and does not average over a Debye cloud. Coherent scattering occurs at small angles θ when $\alpha \geq 1$; then the scattering wavelength extends over plasma structures, and the spectrum yields the frequencies of these macroscopic motions

2.4.4 Diagnostics with Radiation In spite of the small Thomson cross section of 6.65×10^{-29} m^2, the large intensities available with Nd-glass or CO$_2$ lasers make Thomson scattering a commonly used, nonperturbative diagnostic for temperatures and densities inside a plasma (Lochte-Holtgreven, 1968; Griem and Lovberg, 1970; Sheffield, 1975). In this technique, the laser wavelength is much shorter than the wavelengths being probed, so that $|\mathbf{k}_0| \approx |\mathbf{k}_s| \gg |\mathbf{k}|$. The scattering parameter defined above is then approximated by $\alpha \approx 1/2k_0\lambda_D \sin(\theta/2)$. For $\alpha \ll 1$, the Doppler-shifted light yields the temperature or, more generally, the velocity distribution of the electrons. For $\alpha \gg 1$, the light probes the cooperative motions of the plasma, including electron plasma waves, whose frequency yields the density, and ion-acoustic waves, whose frequency and damping yield the electron and ion temperatures.

Microwave interferometry (Huddlestone

and Leonard, 1965) provides a reliable way to measure plasma density; it is based on the change of the index of refraction given by Eq. (45). Since the waves cannot propagate for $\omega_p > \omega$, densities above 10^{19} m^{-3} cannot be measured by centimeter waves; and, in fact, millimeter waves must be used for such densities because of refraction. To measure higher densities, far-infrared lasers are used for interferometry with submillimeter waves (Fig. 2). The reflection of far-infrared radiation at the cutoff frequency ω_p or ω_R [Eq. (47)] can be used in reflectometers for density profiles and fluctuations. Emission of electron-cyclotron harmonic radiation gives information on electron temperature. Detection of optical emission gives information based on the intensity ratios of spectral lines and on Stark and Doppler broadening (Lochte-Holtgreven, 1968; Griem and Lovberg, 1970; Selwyn, 1993). For the intense plasmas of laser fusion, x-ray spectrometry is used. Finally, the sophisticated method of laser-induced fluorescence permits measurements of ion velocities and of plasma motions by following tagged metastable ions.

2.5 Nonlinear Phenomena

Plasma methods devised for studying nonlinear phenomena been found useful in other fields of applied physics also. A few examples of interesting results are given below.

2.5.1 Nonlinear Waves and Wave–Particle Interactions When ion-acoustic waves reach large amplitude, their sinusoidal wave forms steepen into sawteeth, as predicted by the *Korteweg–de Vries* equation (Davidson, 1972). Ion-acoustic shock waves have a steep front followed by a wave train and have velocities larger than v_s (Fig. 3). Single pulses called *solitons* are also possible and have been studied experimentally. Plasmas can support collisionless shock waves of both ion-acoustic and Alfvén-wave nature. Nonlinear electron plasma waves tend to form *envelope solitons,* or wave packets, described by the nonlinear Schrödinger equation (Hasegawa, 1975). The envelopes can shrink in length and grow in amplitude as a result of a *modulational instability.* Waves and particles can affect each other through nonlinear interactions (Sagdeev and Galeev, 1969), resulting in, for instance, the generation of harmonics.

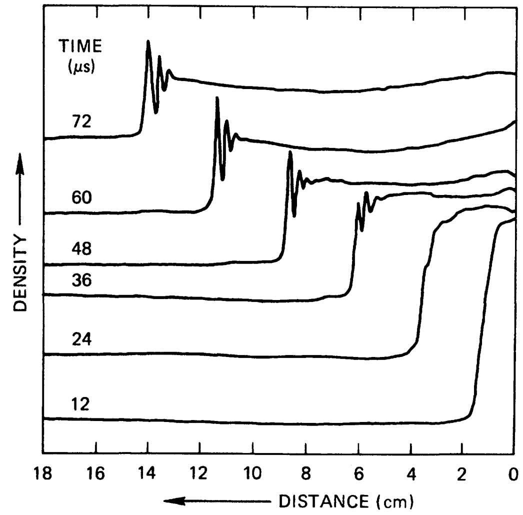

FIG. 3. Ion-acoustic shock waves, as produced and detected in the laboratory. Many such nonlinear plasma phenomena are well described by theory (Chen, 1983).

2.5.2 Sheaths and Double Layers The ion-rich sheath at the wall of a discharge has an electron density that falls exponentially, so that the description of such *Debye sheaths* is intrinsically a nonlinear problem. A negatively biased electrode has, in addition, a *Child–Langmuir sheath,* which is well known to be nonlinear. In space plasma physics, it was found that free-standing sheaths, called *double layers,* can be formed away from all boundaries if at least one of the species has a streaming velocity. Such double layers have been reproduced in the laboratory.

The oldest and simplest plasma diagnostic is the electrostatic probe, whose current–voltage characteric gives information on density and temperature, and sometimes even the nonthermal electron components. Based on sheath theory, the theory of probes was first worked out by I. Langmuir. Though Langmuir probes cannot be used in plasmas so hot and dense that they would destroy anything inserted into them, probes are especially valuable for localized, rather than integrated, measurements (Griem and Lovberg, 1970).

2.5.3 Solitons, Cavitons, and Collapse The study of ion-acoustic solitons has led to light-wave solitons in optical fibers, useful for long-range communications. A *caviton* is a density cavity filled with radiation that is trapped because the plasma around it is overdense; that is, $\omega_p > \omega$. The radiation pressure, which is called *ponderomotive force* when modified by a plasma, creates the density cavity. A caviton can also be filled with electron plasma waves, which also exert a ponderomotive force. Turbulent Langmuir waves are predicted to precipitate into a number of such cavitons by a process called *collapse* (Zakharov, 1984; Hasegawa, 1975).

2.5.4 Turbulence and Chaos The growth of unstable waves into nonlinear waves and finally into turbulence was studied first analytically (e.g., Galeev and Sagdeev, 1983), and more recently by plasma simulation on high-speed computers (e.g., White, 1983). Plasma turbulence is quite different from turbulence in fluid dynamics, in which the "inertial range" spectrum postulated by Kolmogoroff to vary as $k^{-5/3}$ has been observed, k being the reciprocal of the eddy size. Plasma turbulence in strong magnetic fields tends to be two-dimensional, in which case small eddies tend to coalesce into larger ones, instead of vice versa. If an inertial range exists for magnetized plasmas, the energy spectrum can be shown to vary as k^{-5} (Chen, 1965), but many other spectra have also been predicted depending on the assumed origin of the turbulence (e.g., Zakharov, 1984; Sagdeev *et al.*, 1991). For treating drift-wave turbulence, many theoretical techniques have been devised because of its perceived importance in magnetic fusion.

The approach to a state of disorder is not always gradual. Mechanisms such as period doubling or mode coupling can cause oscillations to jump suddenly into or out of a chaotic state with a small change in parameters. Such discontinuous changes are mathematically predicted in the field of *chaos* (Ott, 1993) (see CHAOTIC PHENOMENA), but it is not clear whether or not these processes are the same as those observed in real plasmas. Abrupt changes from order to disorder were first found in computational studies of the behavior of particle orbits and magnetic-field lines in fusion devices (Lichtenberg and Lieberman, 1983) (Fig. 4). This modern field of nonlinear dynamics was initiated in large part by plasma physicists.

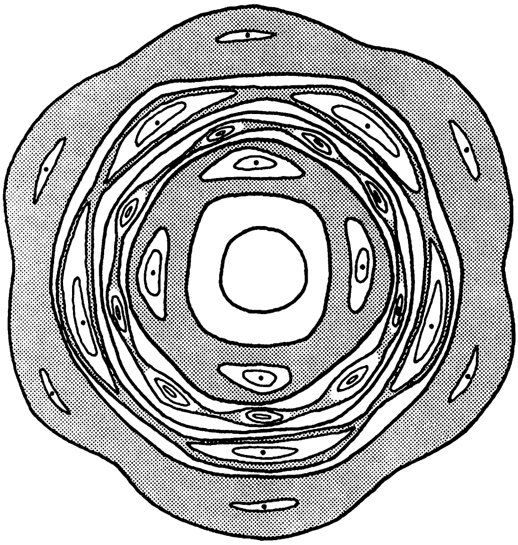

FIG. 4. Numerical mapping of a perturbed Hamiltonian reproduces the magnetic island structure of the field lines in a toroidal tokamak fusion machine. In this cross section, the magnetic surfaces break up into "islands" at those radii where the field lines return to the same position after a rational number of turns around the torus. Between these islands, the field lines lie in KAM (Kolmogoroff–Arnol'd–Moser) surfaces and cover the space ergodically (Lichtenberg and Lieberman, 1983.)

2.6 Nonneutral and Strongly Coupled Plasmas

2.6.1 Single-Species Plasmas Pure electron plasmas can be trapped for days in a uniform magnetic field with electrostatic confinement at the ends (a *Penning trap*), there being no unlike-particle collisions to cause diffusion. Since the electron charge is not neutralized by ions, the resulting electric field causes the plasma to rotate, but a stable equilibrium is possible (Davidson, 1974). Many of the features of normal plasmas, such as Debye shielding, are retained in *single-species*, or *one-component*, plasmas. Because of their simplicity, they have been successfully used to study basic phenomena such as conservation laws and the details of Coulomb collisions. Single-species plasmas can also be made with ions or positrons. Ions can be confined also by rf in *Paul traps*. Of particular interest are plasmas of extremely low temperature. Electrons can be cooled by synchrotron radiation, but ions can be cooled much more efficiently with lasers. In laser cooling, a laser is tuned to the Doppler-shifted wavelength of ions that are moving rapidly toward the laser. The radiation pressure of the beam is then preferentially applied to slowing down these fast ions.

2.6.2 Liquid and Solid Plasmas When a plasma is cooled sufficiently, the number of particles in a Debye sphere, $N = n\lambda_D^3$, falls to the order of unity or below, and Debye shielding can no longer occur. This is the regime of *strongly coupled plasmas* (Ichimaru, 1994). The controlling parameter Γ, proportional to the ratio of Coulomb energy to thermal energy, is defined by

$$\Gamma = q^2/a_s KT, \quad \tfrac{4}{3}\pi a_s^3 n = 1, \tag{51}$$

where a_s is the average interparticle distance. When $\Gamma \geq 2$, the thermal energy of the particles is too weak to overcome interparticle Coulomb forces, and the plasma exhibits the short-range order of a liquid. When Γ exceeds 100 or so, the ions fall into a solid lattice. This has been seen in a pure ion plasma. An easier method to form a solid plasma is to use negatively charged particulates or powders, with a typical grain size of 1 μm. Since these particles can have as many as 10^4 electron charges each, the condition $n\lambda_D^3 \leq 1$ can easily be achieved. Such plasma lattices have been produced in the laboratory. Since $n\lambda_D^3$ varies as $n^{-1/2}$, strongly coupled plasmas also occur at very high densities, such as in white dwarf and neutron stars and in compressed inertial-fusion targets (see FUSION, INERTIAL CONFINEMENT). The equation of state of matter under such conditions is an active field of study.

3. PHYSICS OF MAGNETIC CONFINEMENT

3.1 Effects of Magnetic Fields on Plasmas

Magnetic confinement of plasmas in space and in small laboratory experiments is usually only partially effective: Diffusion across field lines is reduced to the rates given by Eq. (34) for partially ionized plasmas and Eq. (37) for fully ionized plasmas. The physics involved with reduced motion across straight magnetic fields is relatively straightforward and contains few surprises. To trap a plasma so that it cannot escape *along* the field lines, however, is considerably more difficult. This requires specially shaped magnetic fields and elegant methods to suppress the instabilities that such fields engender. Total confinement in complex geometries entails a large number of new concepts. For this reason, the extensive body of knowledge on magnetic confinement is connected primarily with controlled fusion, for which plasma containment in all directions is required.

Two types of magnetic containment geometries are possible: *open* and *closed.* Open systems consist of two or more regions of strong magnetic field separated by a weak-field region. Charged particles in the weak-field region are reflected from the strong-field regions by the magnetic mirror effect (Fig. 3). Particles with small magnetic moment [Eq. (29)], however, can escape through the mirrors at too fast a rate for efficient fusion, and methods for stoppering the ends had to be devised. Closed, or toroidal, geometries have magnetic field lines that do not reach the walls, so that plasma cannot escape along field lines. However, the curvature of the field induces a number of instabilities that destroy the confinement. The problem in closed systems is to devise methods for suppressing these instabilities. (Many books on fusion have been written; several are listed under Further Reading.)

3.2 Magnetic-Confinement Fusion

Controlled fusion energy is generated when deuterium and tritium ions in a plasma combine to form a helium nucleus and a neutron. Because of Coulomb repulsion, directed beams of ions will scatter elastically, and undergo a fusion reaction only in rare head-on collisions. In a plasma with a Maxwellian distribution of ion energies, however, the mean energy will not be changed by elastic collisions, and fusion events will eventually occur. These are called *thermonuclear* reactions. An ion temperature of about 10 000 eV is usually required. The electron fluid may be cooler; but since the temperatures will equilibrate in about the same time as needed for the fusion reactions, the electrons will also be hot enough to maintain the high conductivity of the plasma.

The helium "ash" of the D-T fusion reaction consists of 3.5-MeV α particles, which are confined by the magnetic field. *Ignition* is achieved when the transfer of energy from the α's sustains the plasma temperature against losses via radiation and particle escape. Electrical energy is derived from the other reaction product, the 14-MeV neutrons, which are captured in a *blanket* and thermalized to produce heat for a conventional power plant. For net energy gain, the *Lawson criterion* $n\tau_E \geq 10^{20}$ s/m^3 must be satisfied, where n is the plasma density and τ_E is the energy confinement time given by the plasma energy content divided by the power input. For toroidal systems, n is typically 10^{20} m^{-3} and τ_E is typically 1 s at this *breakeven* condition. Since high ion temperature is also required, the triple product $T_i n\tau_E$ should exceed 10^{21} keV-s/m^3. References and details on this subject can be found under FUSION, MAGNETIC CONFINEMENT and FUSION TECHNOLOGIES.

3.3 Particle and Energy Transport

In principle, plasma can escape from a closed system only at the classical diffusion rate given by Eq. (37). Under fusion conditions the *particle confinement time* τ_p would then be of the order of hours. In practice, confinement times as long as 1 s are difficult to achieve because of *anomalous diffusion* caused by instabilities and field errors. Kinetic energy is also transported across **B** in Coulomb collisions. Classically, the heat flow q_e of electrons, for instance, should be proportional to $n_e\nabla T_e$:

$$q_e = \chi_e n_e \nabla T_e, \tag{52}$$

where the heat diffusivity χ_e has the same $n_i T_e^{-1/2} B^{-2}$ dependence as does $D_\perp$ [Eq. (37)]. In tokamaks, the measured value of χ_i agrees approximately with this prediction when it is modified by the neoclassical effects described below, but χ_e is highly anomalous. Not only is it large, but it can depend on ∇T_e itself, and its dependences on T_e and B can be inverted. In view of this rapid transport, the loss of energy by radiation can be neglected. The energy confinement time τ_E, which should then be proportional to $1/\chi_e \propto T_e^{1/2}B^2/n_i$, is actually found to increase with n and decrease with B. Much of magnetic fusion physics concerns the explanation of such anomalies through the study of instability and turbulence.

3.4 The Energy Principle

To enumerate the possible instabilities in the complicated magnetic geometries required for confinement would be an impossible task. Fortunately, the discovery of an energy principle for hydromagnetic stability (Bernstein *et al.*, 1958) reduces the distinction between stability and instability to evaluating the sign of a quantity δW, which represents the change of the potential energy of the magnetic-field–plasma system when the fluid is given a displacement $\xi(\mathbf{r})$. If the field **B** and pressure p are known everywhere in the volume V, δW in its simplest form is given by the integral

$$\delta W = \tfrac{1}{2}\int [\mathbf{Q}^2 + (\nabla \times \mathbf{B})\cdot(\xi \times \mathbf{Q}) + (\nabla\cdot\xi)(\xi\cdot\nabla p) + \gamma p(\nabla\cdot\xi)^2]d^3V, \tag{53}$$

where $\mathbf{Q} = \nabla \times (\xi \times \mathbf{B})$. Stability requires $\delta W > 0$ for all trial functions $\xi(\mathbf{r})$.

3.5 Toroidal Confinement

Bending straight magnetic-field lines into circles, thus forming a torus (pictured here as a horizontal doughnut), would prevent plasma from escaping along **B**. Unfortunately, the field would then be nonuniform, and the gradient-B drifts [Eq. (23)] would separate the ions and electrons, causing a vertical electric field that drifts the plasma outward by the drift v_E [Eq. (21)]. Toroidal devices must be designed to have twisted, or helical, **B** lines so that the particles, in their thermal motion along **B**, experience an oscillating gradient-B drift with zero average. There are three general types of toroidal devices. Tokamaks employ a plasma current to add a *poloidal B* field (around the minor axis) to the basic *toroidal* field generated by external coils. Helical devices employ currents in windings outside the plasma to generate the poloidal component. *Stellarators, torsatrons, heliacs,* and *spherators* are examples of helical devices. If a closed configuration has no hole in the center and all the fields are generated by currents in the plasma, it is a *compact torus.*

3.5.1 Rotational Transform Let θ be the azimuthal angle at which a given helical field line crosses a given cross section of the torus. After going once around the torus, this field line will have an azimuth angle $\theta + \Delta\theta$. The rotational transform ι, defined as the average value of $\Delta\theta$ over many traverses, is a measure of the amount of twisting. For tokamaks, the reciprocal quantity $q = 2\pi/\iota$ is known as the *quality factor.* Kink instabilities will occur if the plasma current is large enough to give $\iota > 2\pi$, or $q < 1$, a condition known as the *Kruskal–Shafranov limit.* The value of q generally varies with minor radius r. At most radii, q is an irrational number, and the field lines cover an inner torus, called a *magnetic surface,* without ending. In steady state, a perfectly conducting plasma will have $\nabla_{\parallel} p = E_{\parallel} = 0$, according to Eqs. (8) and (9), so that such magnetic surfaces are surfaces of constant pressure and potential. Particle drifts [Eqs. (21) and (23)] cannot cross these nested surfaces. However, at radii such that $q(r)$ is a rational number, the field lines close upon themselves and form *magnetic islands* (Fig. 6). Overall hydromagnetic stability of toroidal configurations depends on the existence and positions of these *rational surfaces,* and hence on the profile of $q(r)$.

3.5.2 Instabilities In a perfectly conducting plasma, in which $E_{\parallel} = 0$, the gravitational and kink instabilities discussed in Sec. 2.3.4 can be controlled by shaping the magnetic field. Unfortunately, a small resistivity η suffices to invalidate the assumption $E_{\parallel} = 0$ and to allow growth of waves localized within a magnetic surface. The primary resistive instability is the *tearing mode.* Modes localized to the outside regions of a torus where the effective "gravity" is strong are called *ballooning* modes. The drift-wave instability discussed in Sec. 2.3.3 is believed to be the cause of electrostatic turbulence, leading to anomalous diffusion. Drift waves can be made unstable not only by finite resistivity but also by the trapping of particles in banana orbits (Sec. 3.5.4); these are called *trapped particle* instabilities. The η_i instability is a drift wave driven by ion temperature gradients. The α-particle product of fusion reactions is predicted to drive Alfvén waves unstable. The number of known instabilities is large (Wesson, 1978), but fortunately only a few in practice are deleterious to plasma confinement.

3.5.3 Stabilization Methods The two primary methods for suppression of instabilities are *shear* and *minimum-B.* Magnetic shear occurs when $q(r)$ varies with radius, so that the **B** lines have different pitch angles on each magnetic surface. This has the effect of scrambling unstable displacements and suppressing them. A minimum-B configuration is one in which the magnitude $|\mathbf{B}|$ is larger on the outside of the plasma than on the inside, forming a potential hill that unstable motions must overcome. These field lines bulge inward toward the plasma rather than outward. A similar benefit accrues if only the average of $|\mathbf{B}|$ along a field line has an interior minimum; such configurations have minimum-average-B, or minimum-$\bar{B}$.

3.5.4 Banana Orbits Let a magnetic surface intersect a cross section of the torus in a circle. As a particle's guiding center moves along a field line on this magnetic surface, the points at which it passes a given cross section should eventually cover the entire circle. However, particles with large magnetic moments will be reflected from regions of strong field by the magnetic mirror effect (cf. Sec. 2.2.4) and will not be able to reach the inside part of the cross section, near the major axis, where the field is strong. The guiding center will then trace out not a circle but a banana-shaped path localized to the outside of the cross section. The particle is then a *trapped particle.* We define the aspect ratio of a torus as $A = R/a$, where R is the major radius and a the minor radius, and the inverse aspect ratio as $\epsilon = a/R$. The fraction of trapped particles is proportional to $\epsilon^{1/2}$. The width of the banana orbits is $(B_t/B_p)r_L$, where B_t and B_p are respectively the toroidal and poloidal field components, or approximately the Larmor radius evaluated with the poloidal field. The step length for diffusion is therefore increased by the factor $\epsilon^{1/2}(B_t/B_p)$ (Kadomtsev and Pogutse, 1971).

3.5.5 Neoclassical Diffusion Because of the toroidal effects described above, transport by classical Coulomb collisions is increased even without instabilities. The resulting "neoclassical" diffusion coefficient has three ranges as the collision frequency ν is increased (Hinton and Hazeltine, 1976). In the banana diffusion regime (small ν), $D_\perp$ is given by

$$D_\perp = D_c q^2/\epsilon^{3/2} = q^2 r_L^2 \nu/\epsilon^{3/2} \tag{54}$$

In the large-ν limit, diffusion is governed by the Pfirsch–Schlüter coefficient

$$D_{PS} = (1 + 2q^2)D_c \cong q^2 r_L^2 \nu. \tag{55}$$

Here the enhancement over D_c [Eq. (37)] is caused by the added electron–ion friction as electrons flow along **B** to cancel their gradient-B drifts, as described above. At intermediate collisionalities, $D_\perp$ is independent of ν; in this *plateau* regime, $D_\perp$ is given by

$$D_\perp = q r_L^2 v_{th}/R. \tag{56}$$

3.5.6 Bohm and Gyro-Bohm Diffusion Observed losses are much faster than neoclassical. In the presence of electrostatic turbulence, the guiding centers of ions and electrons suffer random $\mathbf{E} \times \mathbf{B}$ drifts, and the resulting diffusion can be shown to scale as KT_e/B. The infamous Bohm diffusion coefficient contains an empirical factor of $\frac{1}{16}$:

$$D_B = \tfrac{1}{16} KT_e/eB\ . \tag{57}$$

This scaling holds for both random *diffusion* and steady *convection* as long as the eddy sizes are scaled to the plasma radius. Improvements in stabilization have reduced $D_\perp$ to below this value, at least for electrons. If the correlation lengths are scaled instead to the ion Larmor radius, as is predicted for certain instabilities such as the *ion temperature gradient instability*, the diffusion coefficient would have a smaller value, called the *gyro-Bohm* coefficient:

$$D_{GB} = \frac{r_{Li}}{a}\frac{KT_e}{eB} \tag{58}$$

Electron transport has been observed to follow this scaling law, which recovers the $1/B^2$ dependence of classical diffusion but is still much larger in magnitude.

3.6 Tokamak Physics

3.6.1 Fields and Currents A tokamak (Furth, 1981) has at least three sets of coils: TF (toroidal field) coils, which encircle the doughnut and produce the primary toroidal field; VF (vertical field) coils, which provide a $\mathbf{j} \times \mathbf{B}$ force to prevent expansion of the torus; and OH (Ohmic heating) coils (or a transformer) for inducing the toroidal current. In addition, there are often field-shaping coils. The all-important q profile is determined by the radial distribution of the toroidal current $j(r)$. The Ohmic dissipation of this current heats the plasma in the absence of auxiliary heating. Since the resistivity is sensitive to electron temperature [Eq. (36)], $q(r)$, $j(r)$, and $T_e(r)$ are related to one another. The value of q typically varies from below 1 on the minor axis to above 3 at the periphery. The current tends to be concentrated on reentrant field lines, for instance on the $q = 2$ or 3 surfaces, generating localized poloidal fields. These form magnetic islands that degrade the confinement.

3.6.2 Sawteeth and Disruptions Sawtooth-shaped fluctuations in n and T_e are invariably seen near the $q = 1$ surface. Inside this surface, the plasma is unstable and turbulent because the Kruskal–Shafranov limit is not satisfied. Heat and density are redistributed by the sawtooth oscillations; this is a mechanism for the plasma to readjust $j(r)$ to regain stability. This self-regulation mechanism is not available to stellarators, where $\iota(r)$ is primarily fixed by external currents. In contrast to sawteeth, disruptive instabilities occur unpredictably and are believed to be caused by the growth and overlapping of magnetic island chains. A major disruption stops the plasma current completely and brings the plasma to the wall.

3.6.3 Auxiliary Heating Ohmic heating can be supplemented in four main ways. In ECRF (electron-cyclotron range of frequencies) heating, microwaves from powerful gyrotrons are used to generate the extraordinary wave [Eq. (46)], which is absorbed at the upper hybrid resonance. In ICRF (ion-cyclotron range of frequencies), radio frequency is applied from antennas to generate the fast wave [Eq. (49)]. In lower hybrid heating (LHH), an intermediate frequency is applied by phased grid arrays to excite the lower hybrid resonance [Eq. (41)]. The most successful method has been neutral-beam injection (NBI) heating, in which 100–200-keV beams of neutral deuterium atoms are injected across the magnetic field. The atoms are then ionized by the plasma and trapped. The resulting fast ions transfer their energy first to electrons, and then to the ions. In fusion reactors, these fast ions enhance the fusion rate, since they pass through the region of maximum fusion cross section as they slow down. The current of these fast ions significantly modifies the plasma equilibrium, depending on *coinjection* (in the OH current direction) or *counterinjection* (opposite), and can cause new instabilities.

3.6.4 H-Mode Confinement With neutral-beam injection, it was found that the plasma sometimes makes a spontaneous transition to a new mode of confinement in which the density and potential gradients at the edge of the discharge become extremely steep. When this happens, the oscillation level drops near the edge, presumably because instabilities are suppressed by the large, sheared electric field there. There is then a transport barrier at the edge, which increases the overall confinement time by about a factor of 2. The details of the edge physics are still under study.

3.6.5 Current Drive Since the toroidal current in a tokamak must be induced by a transformer, with or without an iron core, the length of a current pulse is limited by the number of volt-seconds available. Helical systems do not require an internal current and can be operated cw. Currents can be driven without transformers by using the principle of Landau damping (Sec. 2.2.6), by which electrons are accelerated by radio-frequency waves excited in the plasma. In toruses in which plasma is escaping with a radial velocity v_r, there is an automatic current drive due to the $-ev_r \times B_p$ force on the electrons. Since this is always in the direction of the current creating B_p, it is called the *bootstrap current.* Steady-state operation of tokamaks may be possible using the combination of bootstrap current and rf current drive.

3.7 Magnetic Mirrors

3.7.1 Loss Cone A simple magnetic mirror field can be produced by two separated circular coils with currents in the same direction. The field B along the common axis is strongest at the coils ($B = B_{\max}$) and weak at the midplane between them ($B = B_{\min}$). The mirror ratio R_m and the loss-cone angle θ_m are defined by

$$R_m = B_{\max}/B_{\min}, \quad \sin\theta_m = 1/R_m. \tag{59}$$

Particles with $v_\perp/v_\parallel < \tan\theta_m$ at the midplane cannot be reflected at the mirror throats and are lost. Particles diffuse in velocity space into the loss cone by collisions. Since electrons have a large collision frequency, they are lost rapidly, creating an ambipolar potential that accelerates ions through the mirror. Thus, the ions are driven out with a finite drift velocity. In the limit of large R_m, the confinement time is proportional to $\ln R_m$. The loss rate from a simple mirror is too rapid to satisfy the Lawson criterion (Sec. 3.1).

3.7.2 Instabilities Since the field lines of a simple mirror curve outward instead of inward, the configuration is subject to the Rayleigh–Taylor gravitational instability (Sec. 2.3.4). Early experiments were stabilized by *line tying,* the short-circuiting of electrostatic potentials by conducting plates on which the field lines end. Mirrors can easily be stabilized by the addition of a quadrupole field applied by *Ioffe bars,* which are four equally spaced current-carrying conductors parallel to the axis, placed outside the plasma. Above a critical current, these bars convert the mirror into a minimum-B configuration, with a dramatic increase in stability (Post, 1981). It is common to combine the Ioffe bars with the mirror coils into a single conductor, called a *baseball coil,* which has the shape of the seam on a baseball. For ease of assembly, the baseball coil can be split into two halves, which together are called a *yin-yang* coil. In addition to hydromagnetic instabilities, magnetic mirrors are subject to velocity-space instabilities due to the absence of particles in the loss cone. The most dangerous of these is the drift cyclotron loss cone (DCLC) instability, which can be stabilized by adding a cool but isotropic plasma to the hot plasma (Baldwin, 1977).

3.7.3 Tandem Mirrors Even after instabilities have been brought under control, magnetic mirrors still suffer from large end losses. To minimize these losses, it is necessary to build a *tandem mirror* (Fowler, 1981). This has a large central cell with a weak, uniform B field connected to a mirror with strong B field at each end. These "plugs" have yin-yang coils to provide MHD stability and must be filled with hot, dense plasma. Their volumes, however, are relatively small. In addition, the injection of ion beams and ECRF power into the plugs produces a specially tailored potential distribution that, in one section, electrostatically confines the ions and, in another section, forms a thermal barrier for the electrons. These devices for plugging the end losses have been successfully demonstrated, but they detract from the basic simplicity of the mirror concept.

4. PHYSICS OF INERTIAL CONFINEMENT

4.1 Laser Fusion

In laser fusion, a small pellet of frozen deuterium–tritium fuel is suddenly heated by laser or particle beams impinging symmetrically on the surface, turning it into a hot plasma. As the plasma expands, its momentum is transferred to the solid part of the pellet, compressing it in a violent implosion to a density between 10^3 and 10^4 times solid density (Duderstadt and Moses, 1982). The fusion rate at these densities is high enough to give a net energy gain in the short time, of the order of picoseconds, during which this compressed state is maintained by inertia. The Lawson criterion of Sec. 3.1 can be expressed here as $\rho R > 3$ g/cm^2, where ρ is the density and R the radius in the compressed state. Here, the range of the α particles produced is so small that their energy is efficiently used in keeping the plasma at fusion temperature. In practice, the pellets are hollow shells covered with various layers of absorbers; a small piece of fuel at the center is ignited first, and the α particles from this are used to ignite the rest of the fuel in a propagating burn. In laser fusion, there are two approaches. The laser energy can be directed at the fuel pellet (*direct drive*), or it can be directed through the ends into a tiny high-Z canister, in the center of which the fuel pellet is mounted (*indirect drive*) (Yamanaka, 1991). The laser light is converted into x rays, and the latter are used for heating the pellet symmetrically. *Inertial confinement* is a generalization of the compression–ignition process to include drivers other than lasers (see Sec. 4.4). A more detailed treatment of this subject can be found under FUSION, INERTIAL CONFINEMENT.

4.2 Parametric Instabilities

The hope that inertial fusion would be free from the problems of plasma instabilities that plagued magnetic confinement was not realized: Laser fusion had instabilities of its own, called *parametric instabilities* (Nishikawa *et al.*, 1976). Ideally, laser light of frequency ω_0 should penetrate the plasma cloud up to the critical density $n_c = \epsilon_0 m \omega_p^2 / e^2$ where $\omega_0 = \omega_p$ [Eq. (14)]. At that point, the laser energy would be absorbed by collisional damping (*inverse bremsstrahlung*) and by the generation of plasma waves that are subsequently Landau damped (*resonance absorption*). Unfortunately, reflective instabilities occur in the underdense ($n < n_c$) region, preventing the absorption of the light. Furthermore, these instabilities generate fast electrons that can move forward into the solid fuel and preheat it, making it harder to compress. In *stimulated Brillouin scattering* (SBS), the laser light excites an ion-acoustic wave, which then acts as a grating to reflect the laser beam at a red-shifted frequency. *Stimulated Raman scattering* (SRS) is similar except that an electron plasma wave is generated, and the redshift is of order ω_p. If the plasma wave has $k\lambda_D \leq 1$ so that it is strongly Landau damped, the induced scattering from individual electrons is called *stimulated Compton scattering.* The laser light can also decay into two plasma waves in a *two-plasmon decay* instability occurring at the *quarter-critical layer,* where $n = n_c/4$. The *filamentation* instability can divide the laser beam into smaller beamlets, thus reducing the uniformity of power deposition. Fast electrons accelerated out from the pellet have such large current densities that they can generate multimegagauss magnetic fields, which degrade the smoothing effect of heat conduction. These parametric effects can be minimized by increasing ω_0, since the critical layer is then moved to higher densities, where classical collisions are more dominant. Nd-glass lasers are used for the 100-TW power levels required, the 1.06-μm (red) fundamental being frequency doubled to 0.53 μm (green) for diagnostics and tripled to 0.35 μm (blue) for compression. Though the conversion efficiency in crystals can be of order 75%, frequency-quadrupled light at 0.25 μm (UV) would require unusual optical materials. In indirect drive, the x rays have such high frequency that parametric instabilities are not a problem, but the light still has to pass through a low-density plasma before reaching the metal canister.

4.3 Rayleigh–Taylor Instability

After parametric instabilities, the most important remaining instability is the Rayleigh–Taylor instability (Bodner, 1991). As the pressure of the ablating plasma pushes on the solid wall of the pellet, small irregularities in the surface or the driving force create growing ripples, which prevent symmetric compression. Large-aspect pellets that have an outer shell that is thin compared with its radius are more easily imploded, but they are also more susceptible to the R-T instability. A beneficial effect not available in magnetic fusion is *vorticity shedding,* in which the escaping plasma carries away some of the instability energy. In direct drive, the laser energy must be divided into many beams to spread out the driving force uniformly over a sphere. In indirect drive, the x-ray *hohlraum* is supposed to provide a uniform pressure, but the laser illumination of the canister and the blockage of x rays escaping from the ends of the cylinder have to be adjusted carefully.

4.4 Driver Technology

Glass lasers are exceptionally inefficient and cannot be pulsed rapidly because of heat dissipation in the glass. For fusion power production, therefore, other drivers are necessary. The KrF laser, with 0.3-μm wavelength, is a possible alternative, being a gas laser that can run cw, but its development is in a relatively early stage. Beams of light ions, such as Li, can be produced by pulsed power (cf. Sec. 6.2 below) and directed toward a larger target than in laser fusion, heating the surface in the same way. The main problem is the focusing and transport of the ion beam from the diode structure in which it is produced to the target through an evacuated space that must also accommodate the neutron blanket. In *light-ion fusion,* the pulse length, energy per pulse, target size, and fusion energy per shot are all scaled upward from the laser case. The difficulties with transport are overcome in heavy-ion drivers. In *heavy-ion fusion,* beams of ions such as uranium are accelerated to MeV energies by particle-accelerator techniques. As in high-energy physics, the beams can be transported and focused to small radii. The energy deposition in the target is sharply concentrated at the end of the ions' range. New types of linear accelerators are being developed for these drivers.

5. PLASMAS IN SPACE

5.1 Plasmas in the Solar System

That plasmas exist naturally in outer space has long been apparent from displays of the Aurora Borealis and from disruptions of short-wave radio communications, including the mysterious glide tones now called *whistler waves* (Helliwell, 1965). The exploration of the environments of the Earth and other planets with spacecraft has brought not only an understanding of these phenomena (see IONOSPHERE) but also the realization that the sun and planets form an electromagnetic system controlled by the laws of plasma physics. The dipole magnetic field of the Earth traps ions and electrons coming from the sun by the magnetic mirror effect (Sec. 3.1.5), forming the *van Allen radiation belts.* The plasma density in this ionosphere rises from zero in the atmosphere, where the neutral gas cools the plasma so that it recombines readily, to a peak of the order of 10^5 cm^{-3}, and then falls to the order of 1–10 cm^{-3} in interplanetary space. Occasionally, electromagnetic disturbances drive energetic electrons into the atmosphere, exciting the oxygen atoms to emit the light of the aurora. A thunderstorm in the southern hemisphere can generate electromagnetic plasma waves that are channeled along the Earth's magnetic field to give whistling sounds detected by radio near the north pole, since the higher frequencies travel faster. A stream of charged particles from the sun, called the *solar wind* (Brandt, 1970), populates the ionosphere and controls its behavior. A solar storm can cause the ionosphere to fluctuate and disrupt its normal reflection of communications signals. The ionospheric plasma can even be manipulated by man, using powerful radio-frequency generators at the plasma frequency.

Outside the ionosphere, the plasma is nearly collisionless and is therefore frozen to the magnetic field. Since this is a high-beta plasma [beta is defined in Eq. (19)], the magnetic field cannot hold the plasma; rather, the field lines are shaped by the plasma motion. In particular, the plasma of the solar wind pushes the dipole field of the Earth into a long tail on the night side, called the *magnetotail.* The field lines of the Earth's field are joined to those of the interplanetary magnetic field discontinuously at a boundary called the *magnetopause,* which encloses the *magnetosphere* (*q.v.*). The discontinuity is possible because of sheet currents in the plasma. As the Earth orbits the sun, it can be viewed as plowing through the interplanetary magnetic field, dragging its own plasma and magnetic field with it, the field lines adjusting themselves at every instant by the process of *reconnection.* Since the earth's velocity is supersonic in the interplanetary plasma, a collisionless shock front, called the bow shock, forms ahead of the Earth. For further information and references, see IONOSPHERE and MAGNETOSPHERE.

Each planet has a different plasma environment depending on its magnetic field and ionosphere. The magnetosphere of Jupiter has been explored extensively and is particularly interesting because of its major satellites, which have their own magnetic fields and can even inject particles into Jupiter's magnetosphere through volcanic action. The rings around Saturn have features that may depend on plasma effects. For instance, it has been proposed that the radial spokes in these rings may be caused by charged dust that orbits over the poles of the planet. Comets can often be seen to have two tails, one blown by the sun's photons and the other by the solar wind. The dust cloud that is ejected from a comet as it nears the sun can be ionized if it collides with an existing plasma at greater than the *critical ionization velocity.* Dusty plasmas are of importance not only in comets and planetary rings, but also in plasma processing (see Sec. 7.2).

5.2 Plasmas in the Sun

The sun is a fusion reactor with gravitational plasma confinement. The heat that is generated at the core is convected out to the surface by turbulent motions. In a manner not yet understood, these motions generate complicated magnetic fields, which control the appearance of the sun and its interaction with the Earth via the solar wind. Some magnetic lines leave the sun's surface and loop back, forming the north and south poles of a sunspot pair. Solar prominences reveal the presence of magnetic loops further out. Some field lines do not loop back but reach out into space through holes in the corona, carrying the solar wind with them. The corona itself is extremely hot (Fig. 5). The mechanism of coronal heating and the dynamo that creates the magnetic fields are unsolved problems. For further information and references, see SOLAR RADIATION.

5.3 Magnetic-Field Reconnection

Since the particles in a collisionless plasma gyrate around magnetic-field lines and are effectively "frozen" to them, it is not possible to change the magnetic configuration without a mechanism for particles to jump from one field line to another. This problem of *magnetic reconnection* or *field annihilation* arises both in space plasma physics (Fig. 6) and in magnetic-confinement fusion (Sec. 3.1.4.2). If two oppositely directed

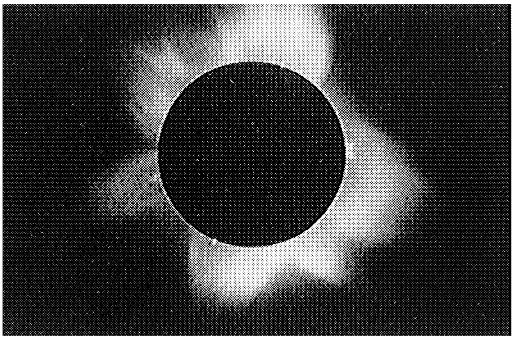

FIG. 5. The sun's corona, prominences, and sunspots are examples of magnetized plasmas. The solar wind carries some of this plasma to the Earth and controls its plasma environment. Fusion reactions in the interior of the sun generate its energy, as in all stars.

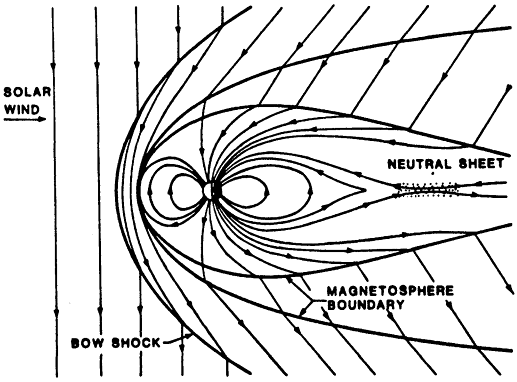

FIG. 6. The magnetic environment of the Earth, showing the neutral sheet in which field lines of opposite polarity coalesce and magnetic reconnection occurs. The transition between the interplanetary magnetic field and the Earth's field is controlled by the presence of plasma from the solar wind that is entrained by the magnetic fields.

field lines come together to cross each other at an X point, entraining plasma with them, eventually the plasma is squeezed into a thin layer in which the density is high enough for collisions to occur. Collisions permit particles to cross field lines, and so the field lines can lose their identity momentarily and reconnect. In this process, some of the magnetic energy is converted into kinetic energy, and acceleration of particles can occur. Reconnection is independent of collisionality, because the thickness of the interface layer can adjust itself to guarantee a sufficiently high density. A peculiarity of plasmas is that reconnection can occur even if the collision rate is identically zero, because collisionless processes such as Landau damping or instabilities can provide an effective resistivity.

5.4 Plasma Astrophysics

Hydrogen atoms are photoionized in interstellar space by starlight, and the plasma does not readily recombine when the density is of order 1/cm^3. The Debye length is nonetheless much shorter than the dimensions of the system, so that interstellar and intergalactic space is filled with true plasmas. The motions of magnetic-field lines are subject to the same constraints as in the solar system. The presence of magnetic fields can be seen in the Crab Nebula, in which the organization of gas into filaments can only be due to magnetic fields (Fig. 7). The directions of these fields can be measured from the polarization of the synchrotron radiation emitted by the magnetized electrons.

Quasistellar objects such as neutron stars, quasars, pulsars, and black holes involve plasmas of extremely high density with magnetic fields as high as 10^{12} G. Radioastronomy has revealed the presence of extended radio sources and astrophysical jets, which also emit plasma radiation. When an object with a large magnetic field spins, possibly with plasma streaming from the poles, there will be a radius at which the field lines will be rotating at the speed of light. Relativistic astrophysics is the name of the science of such phenomena, but the concepts of plasma physics are deeply imbedded in this science. The acceleration of cosmic rays can be attributed to these objects, but also to other phenomena such as field annihilation and magnetic mirroring.

On a larger scale, the formation of galaxies (Plate 3) has been treated as a plasma problem, with the particles of a plasma replaced by stars and the electromagnetic force replaced by gravity. Spiral arms can thus be considered to grow as a result of an instability. The formation of stars from gas and dust is a problem in plasma hydrodynamics involving dusty plasmas, which also occur in the solar system and in industrial plasmas (Secs. 5.1 and 7.2). For further information and references, see ASTROPHYSICS.

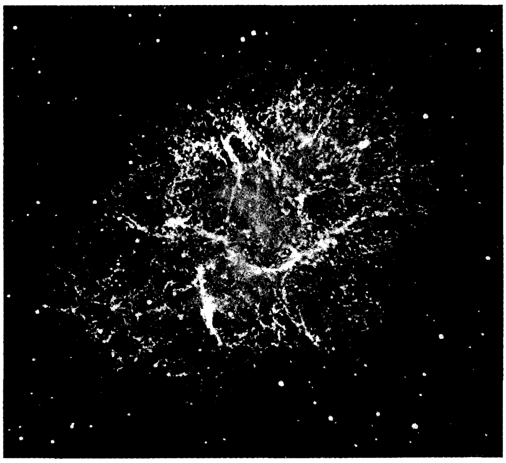

FIG. 7. The filamentary structure of the Crab Nebula reveals the presence of plasma constrained by magnetic fields.

6. LIGHT AND PARTICLE BEAMS

6.1 Coherent Radiation Generators

6.1.1 Microwave Sources Originally developed for radar, microwave generators are now used by everyone in personal communications, television, and microwave ovens. For the scientist, microwave sources are useful in particle accelerators, controlled fusion, and spectroscopy of organic substances. Coherent radiation is generated by electron beams, which are basically single-species plasmas; and their motions, space-charge effects, and instabilities are described by plasma-theoretical techniques. In a *klystron,* an electron beam is shot through two resonant cavities. The electric field in the first cavity modulates the electrons' velocities so that they become bunched. The bunched beam then passes through the second cavity and excites the microwaves in it. A portion of this energy is fed back to drive the first cavity, and the rest is the output energy. In a traveling-wave amplifier (TWT), a slow-wave structure such as a coil surrounds the beam and propagates electromagnetic waves with approximately the electrons' velocity. A beam–plasma instability occurs, with the plasma replaced by the slow-wave structure, and the wave on it is amplified. In both cases, radiation is generated at the expense of electron energy. In a magnetron, a magnetic field causes the electrons to drift at the $E \times B$ velocity. Magnetrons at 2.45 GHz are mass produced economically and are used in microwave ovens. In a gyrotron, the electron beam travels both perpendicular and parallel to a magnetic field, and the electron are bunched in their gyro-motions. The radiation is near the cyclotron frequency. To produce higher frequencies at a reasonable magnetic field, harmonic gyrotrons operate at high multiples of ω_c. Gyrotrons can produce prodigious power in the megawatt range for use in heating fusion plasmas. The technology of gyrotrons has also been used in isotope separation and materials processing. A recent variant is the cyclotron autoresonant maser (CARM), in which the electrons' parallel velocity is larger than their perpendicular velocity, providing an energy source for maintaining the proper relativistic mass to stay in cyclotron resonance.

6.1.2 Free-Electron Lasers The microwave sources mentioned above are generally not tunable over a large range of frequencies. Tunability is the distinguishing feature of the free-electron laser (FEL) (*q.v.*) (Roberson and Sprangle, 1989). In a FEL, a relativistic electron beam and a beam of electromagnetic radiation are injected together into a channel in which there is a spatially alternating magnetic field (a *wiggler* or *undulator*). The combined fields cause the electron beam to oscillate and generate radiation at higher frequency and higher power. Tunability is achieved by varying the wiggler wavelength or the electron energy, since the frequency upshift is proportional to γ^2, where γ is the relativistic Lorentz factor. Thus, FELs are useful not only for applications where tunability is essential, such as spectroscopy, but also for generating short-wavelength radiation. Powerful FELs have been proposed for defense purposes. When the density of electrons is so large that $k\lambda_D \leq 1$ in their rest frame, microwave devices are said to operate in the *Raman regime,* but no such devices have yet been made practical.

6.2 Pulsed Power

6.2.1 Relativistic Electron Beams Extremely intense beams of electrons can be produced by charging a room-sized array of capacitors immersed in oil (a *Marx bank*) and switching the voltage to a pulse-forming transmission line (a *Blumlein*) that ends in a diode (Miller, 1982). By charging the capacitors in parallel and connecting them in series via spark gaps before discharging, voltages of 0.1–10 MV can be achieved. The Blumlein is usually insulated by pure water, which can maintain a high dielectric constant (≈89) for short pulses, and can deliver ≈80-ns pulses of order 1–20 MA, depending on whether the device is designed for high or low impedance. The relativistic electron beams (REBs) emitted by the diode can carry powers up to the order of 10^{13} W. The beam density is large enough to prevent propagation in a vacuum; a gas fill is normally used to provide the ions needed for space-charge neutralization. The beam current also has to be neutralized, since it is beyond the *Alfvén limit,* at which the Larmor radius of the electrons in their self-generated magnetic field is smaller than the beam radius. A return current of dense but slow electrons flows backward in the beam to cancel the magnetic field. REBs can be used to study the behavior of matter under extreme energy conditions, such as in nuclear weapons tests, and to generate multi-megagauss magnetic fields (Turchi, 1980).

6.2.2 Light-Ion Beams By reversing the polarity of the diode in a pulsed-power machine and introducing a source of light ions such as plastic fibers, it is possible to produce intense ion beams up to the order of 1 MA. Magnetic insulation can be used to prevent the electrons from short-circuiting the diode. Light-ion beams are studied as alternate drivers for inertial-confinement fusion. For this purpose, circular arrays of tens of pulsed-power generators have been built to concentrate the power onto a fusion target (cf. FUSION, INERTIAL CONFINEMENT).

6.2.3 Pinches, Switches, and Exploding Wires The advent of pulsed power opened up research on different types of single-shot devices. A *Z* pinch is a linear, high-current discharge whose self magnetic field is sufficient to confine, or even pinch, the plasma radially. Such a plasma is known to be unstable to kink or sausage instabilities, but useful radiation, for instance for x-ray lithography, can be created before breakup. A Marx bank discharged through a solid deuterium fiber or an imploding metal cylinder has been used to study the possibility of a simple fusion device that requires no external coils. Pulsed discharges through solid wires or wire arrays are used to generate short pulses of light or pressure. Fast-opening switches that can transfer large currents have a number of uses including the charging of an inductor for inductive energy storage (Früngel, 1976). A dense plasma generated by pulsed power can be used for this purpose. (See also SPARKS, ARCS, AND OTHER ELECTRIC DISCHARGES.)

6.3 Plasma Accelerators

New ideas from plasma physics have been applied to the technology of high-energy accelerators used for fundamental particle-physics research and for synchrotron light sources. For instance, the *modified betatron* adds a toroidal magnetic field to the betatron accelerator, and the *inverse free-electron laser* (IFEL) accelerates a beam of electrons using a laser beam and an undulator. Conventional accelerators driven by microwave sources are limited to acceleration rates of the order of 100 MeV/m by voltage breakdown. Since a plasma is already ionized, plasma-based schemes do not suffer from this limit and in principle can achieve gradients of order 100 GeV/m (Chen, 1991a).

6.3.1 Beat-Wave Accelerator In this concept, two infrared laser beams of different frequencies are sent into a plasma whose plasma frequency matches the lasers' difference frequency. The beat between the laser beams then resonantly excites an electrostatic plasma wave with a velocity equal to the group velocity of the light (Fig. 8). Electrons injected into this wave at the right phase will be continuously accelerated by the electrostatic field. The maximum electron energy in eV is roughly equal to $n^{1/2}$, where n is the density per cm^{-3}, so that $n = 10^{18}$ cm^{-3} would give 1 GeV/cm. Such linear accelerators would be a factor of 1000 shorter than conventional ones and make the *in situ* production of medical isotopes much easier. Preliminary tests have shown acceleration to 30 MeV, in agreement with theory.

6.3.2 Wake-Field Accelerator Another method to generate a fast plasma wave is to inject a short pulse of electrons into a plasma. A "wake" oscillating at ω_p then follows the pulse like the wake of a boat. If another pulse of electrons is injected into this wake, it can be accelerated in the same way as in the beat-wave accelerator. The wake-field accelerator is effectively a transformer, turning a large bunch of electrons of low energy—say, 100 MeV—into a smaller bunch of electrons at a higher energy—say, 1 GeV. A related concept employs a short laser pulse to generate the wake. In any case, these methods all require a dense, uniform plasma source and either a large laser or a large accelerator.

6.3.3 Plasma Lenses To maximize the rate of high-energy reactions, particle beams must be focused to submicron diameters. Plasmas can provide a focusing force orders of magnitude larger than conventional magnets. A charged relativistic beam can propa-

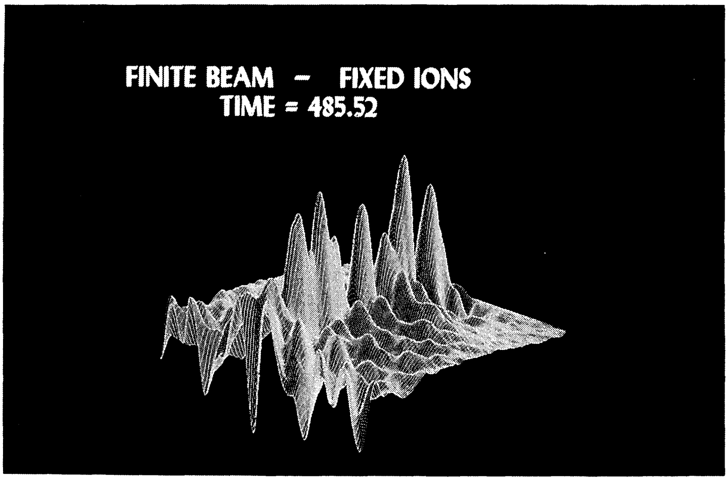

FIG. 8. A two-dimensional (r,z) particle simulation of the beat-wave accelerator mechanism. The laser beams enter from the right (along z), and the ponderomotive force at their beat frequency bunches the electrons into the density fluctuations of a plasma wave. The electric potential of the wave is plotted vs. z and radius r. The pattern travels along z at the laser beams' group velocity. Beam electrons injected into the wave can be trapped in the potential wells and accelerated to the wave's phase velocity (Mori, 1990).

gate without spreading because its electrostatic repulsion is balanced by the inward force of the self magnetic field. When sent through a plasma, the beam electrons replace the thermal electrons, and the ions cancel the electrostatic field. The pinch force is then able to focus the beam.

6.4 X-ray and Short-Pulsed Lasers

6.4.1 X-ray Lasers The dream of generating coherent light in the x-ray regime has been realized by the use of the powerful pulsed lasers developed at the Lawrence Livermore Laboratory for laser fusion. Inner-shell transitions in multiply ionized atoms can be induced by irradiating a plasma several centimeters in length with intense laser pulses (Kauffman, 1991). The medium for x-ray lasers is necessarily in the plasma state. X rays have also been produced by laser heating of carbon fibers and with Z pinches. In each case, special methods are used to depopulate the lower level of the transition. So far, no efficient x-ray reflectors have been developed, and all x-ray lasers are superradiant.

6.4.2 Femtosecond Lasers Pulse-compression techniques on glass lasers have been developed to the extent that tens of terawatts can be delivered in 100 fs (1 fs = 10^{-15} s), with peak intensities exceeding 10^{18} W/cm^2. At these intensities, the electric field of the light wave exceeds that which holds the orbital electrons in an atom, and the ponderomotive force is measured in megabars. Under such conditions of high energy density, almost any material would be turned into a plasma. The behavior of matter in this plasma state would be highly nonlinear; for instance, the electrons moving in the light wave would change their mass relativistically in each half cycle. Production of even shorter pulses will permit the observation of ultrafast biological and chemical phenomena, such as the generation of nerve impulses or the formation of a molecule.

7. INDUSTRIAL APPLICATIONS

Plasmas are extensively used in manufacturing and in the production and improvement of materials. These applications call for low-temperature, partially ionized plasmas, which have properties different from the hot, collisionless plasmas used in fusion and present in space. Industrial plasmas are likely to have many species of atoms and molecules, including negative ions, and are usually collisional, involving simultaneously many different scattering, excitation, and ionization processes (Lieberman and Lichtenberg, 1994).

7.1 Plasma Sources

In some manufacturing procedures, the type of plasma is not crucial; only the existence of electrons is needed. In the "barrel etcher," for instance, a simple discharge between two electrodes is used to create a low degree of ionization. For the fabrication of large-scale integrated semiconductor circuits, however, the plasma performs several functions and must be carefully controlled. By far the most common and the best developed plasma generator for this purpose is the parallel-plate capacitor discharge, called the RIE (reactive ion etching) source (Fig. 9). The RIE plasma is made by applying a rf voltage at 13.56 MHz between two parallel electrodes about 5–10 cm apart. The gas is a mixture containing Cl or F compounds, which are dissociated by the plasma electrons; it is the neutral Cl or F atoms that chemically attack the target in etching. A sheath forms on both electrodes to confine the electrons, that on the negative electrode being thicker. During the rf cycle, these sheaths oscillate, and the plasma "sloshes" back and forth between electrodes. On average, the ions experience a sheath potential that accelerates them toward the substrate, or object to be treated, which is mounted on one electrode. That electrode can be grounded through a separate rf power supply to adjust the rectified sheath voltage. Since the sheath drop, rf oscillation, and plasma density all depend on rf power, the degree of control over the plasma is restricted.

New plasma sources have been proposed to be more flexible than the RIE source and to produce higher plasma densities (Lieberman and Gottscho, 1994). (See also PLASMA ETCHING.) The leading contenders are the electron-cyclotron resonance (ECR) source, inductively coupled plasmas (ICP), the radiofrequency inductive (RFI) or transformer-coupled plasma (TCP) source, and the helicon source. The ECR source, driven by 2.45-GHz microwaves, is a spinoff from ECRF technology used for plasma heating (cf. Sec. 3.1.4.3). This source can produce high densities at low pressures but requires a magnetic field of 875 G at the resonance layer. The other sources operate at the industrial frequency of 13.56 MHz or its harmonics. ICPs utilize a helical winding around a cylindrical tube for applying the rf power, usually with an electrostatic shield to eliminate direct electrostatic coupling. RFI and TCP sources are similar, each with a pancake-shaped rf antenna lying on a quartz plate, which separates it from the plasma chamber. These rf sources do not require a magnetic field and do not resonate with a natural frequency of the plasma. In the helicon source, an external antenna launches a low-frequency whistler wave, called a helicon wave, which ionizes the plasma as it is damped. A magnetic field of 0.1–1 kG is required. When the helicon resonance is struck, ionization is particularly efficient. In all these sources, the plasma is often drifted into a chamber with magnetic confinement at the walls provided by arrays of permanent magnets, called a *magnetic bucket*, to smooth out density variations before the plasma arrives at the substrate.

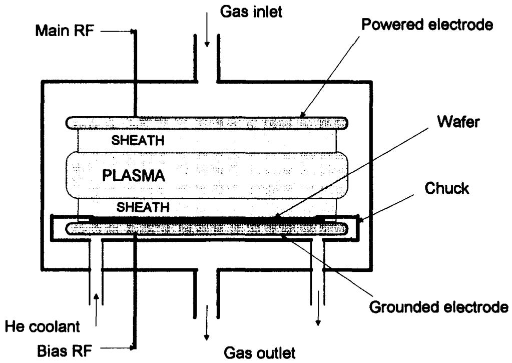

FIG. 9. Diagram of a RIE (reactive ion etching) plasma source commonly used for etching and deposition of silicon wafers in semiconductor fabrication. A 13.56-MHz radio-frequency source applied between the parallel-plate electrodes creates the plasma, which impinges on the wafer mounted on the bottom plate.

7.2 Plasma Etching and Deposition

Plasma processing is essential in the production of semiconductor integrated circuits, in which the individual features can be smaller than 0.5 μm. The manufacture of flat-panel displays, such as the active-matrix liquid-crystal displays (AMLCDs) used in portable computers, also benefits from plasma processing, though it is not essential, since the elements here are on a larger scale. Integrated circuits are made, hundreds at a time, on a polycrystalline silicon wafer, typically eight inches or larger in diameter. The various components of a chip—transistors, capacitors, conductors, etc.—are made in a series of steps involving deposition, masking, stripping, and etching (Manos and Flamm, 1989). Different processes are used for handling semiconductor material, oxide or nitride insulators, conductors, and photoresist. The patterns are of such fine scale that photolithography with short-wavelength x rays must be used; the x-ray sources are often plasma devices. Since the gases, pressures, and even plasma sources are generally different for each step, the wafers are shuttled under vacuum from one "reactor" to the next in a transfer mechanism known as a *cluster tool.*

The role of plasma in the etching process is twofold. First, the electrons produce the active species, and ion bombardment prepares the surface so that chemical etching occurs more readily. The combined effect is to increase the etch rate by over an order of magnitude over chemical etching without plasma or plasma sputtering without chemicals. Second, the sheath electric field accelerates ions so that they strike the substrate at right angles, thus guiding the etching process to proceed in a straight line, so that sharply defined features can be chiseled. Low-pressure operation is desirable for reducing ion scattering in the sheath. The plasmas used for etching are designed for the following attributes:

1. *uniformity* of density over an entire wafer;
2. directionality, or *anisotropy* of ion orbits;
3. high density, and hence high *etch rate;*
4. high *selectivity* in etching the desired material; and
5. compactness, simplicity, and low cost.

Examples of problems that occur in plasma etching and deposition are contamination by particulates and damage to oxide insulators. Micron-size dust particles are formed from the ambient gases, and, being charged negatively by electron bombardment, they are trapped in local maxima of electric potential. When the plasma is turned off, the particles are driven into the wafer and cause defects in some of the chips. Device damage occurs when energetic electrons or ions imbed themselves into an insulating layer and cannot be removed. To prevent damage to delicate circuits, one can use a *downstream reactor,* in which the plasma produces the active species but is swept away before any charged particles can reach the wafer. More detailed information on this subject may be found in PLASMA ETCHING.

7.3 Gaseous Electronics

Traditionally, gaseous electronics is the study of plasmas in gas-filled electronic tubes and switches and of atomic processes such as collisions, ionization, and radiation. Subjects such as positive columns, glow discharges, arcs, coronas, field emission, breakdown, and cathode spots have now been replaced by the microwave and radio-frequency discharges used for plasma processing. The most common efficient light source, the fluorescent light, retains its eminent position in gaseous electronics; but new light sources with higher brightness and efficiency and with no internal electrodes are being developed using microwave and rf technology. For more information on this subject, see SPARKS, ARCS, AND OTHER ELECTRIC DISCHARGES.

7.4 Implantation and Polymerization

Ion implantation with energetic ion beams is used to improve the surface hardness, for instance, of metals or to create special semiconductor materials (Conrad and Sridharan, 1994). In this process, a three-dimensional object has to be rotated to expose its surfaces to the ion beam. In plasma-source ion implantation (PSII), the object is immersed in a plasma and is given a negative pulse of tens of kilovolts. If the Debye length is sufficiently small, a sheath is formed on all surfaces, and the sheath voltage drives ions at near-normal incidence into all surfaces regardless of their irregular shapes. This method is used for implanting nitrogen into cutting tools and other metallic objects to increase hardness and reduce corrosion. Medical prostheses can be hardened and smoothed by PSII. Eventually, PSII might be applied to the manufacture of magnetic and optical disks.

Strong plastic coatings can be created with plasma discharges in organic compounds such as methyl methacrylate (d'Agostino, 1990). When polymerized in the presence of electrons, a plastic material becomes highly cross linked, resulting in a strong and resistant layer. Organic plasmas can be used for a number of practical purposes: for barrier coatings in gasoline tanks, soft drink bottles, and medical capsules; for making textile fibers more absorbent to dyes; for increasing the water resistance of papers and wood; for cleaning and sterilization of biomedical containers; for producing optical fibers with graded indices of refraction; and eventually for producing integrated optics chips for optical computing.

7.5 Thermal Plasmas

Thermal plasmas are plasmas at near-atmospheric pressure that are so collisional that they are in kinetic, though not radiative, equilibrium (Boulos *et al.*, 1994). Plasma spray treatment of aircraft and automobile parts such as turbine blades is a widespread industrial application of plasma physics. Thermal plasma jets are capable of growing diamond coatings at a much faster rate than with low-pressure plasmas. Development of this process can lead to diamond coatings on all cutting tools, and even to diamond substrates for semiconductors. Thermal plasmas can also be used to develop new types of ultrahard materials. Further information can be found in THERMAL PLASMA DEVICES.

7.6 Isotope Separation

The production of U^{235} by gas-diffusion methods is notoriously slow, and many ideas for faster processes have been proposed. The use of plasmas for isotope separation started during World War II with the ill-fated Calutron project, named after the University of California, in which instabilities were so strong that the plasma was lost at what is now called the Bohm diffusion rate. Though new types of plasma centrifuges are still being researched, the major avenues for advanced isotope separation use lasers or radio-frequency generators.

7.6.1 Laser Isotope Separation With narrow linewidth, lasers can selectively excite an isotopic species either as an atom or in a molecule. In atomic-vapor laser isotope separation (AVLIS), the naturally occurring isotopic mixture of an element is vaporized and introduced into a long chamber. A laser beam tuned to a particular line is directed into the gas and selectively ionizes the desired isotope. A high voltage applied to the walls of the chamber then causes the ions to be collected at the cathode (Vitello *et al.*, 1992). This method is limited by the rate at which the ions can drift across the chamber, and to those isotopes for which a suitable laser line can be found.

7.6.2 Ion-Cyclotron Isotope Separation In this scheme, the raw material is vaporized, then ionized by microwaves from a gyrotron, and is confined radially by a uniform magnetic field (Chen, 1991b). A radio-frequency signal near the cyclotron frequency of the desired isotope is then applied to the plasma with a helical antenna. Those atoms that are in cyclotron resonance are selectively accelerated to large perpendicular energies and, because of their large Larmor radii, can be "scraped off" onto suitable collectors. Since both the frequency and the magnetic field can be tuned, this method will obviously work for any isotope; for this reason, it is particularly useful for producing medical isotopes.

GLOSSARY

Adiabatic Invariant: A physical quantity that stays constant under slowly changing conditions.

Alfvén Speed: The propagation speed of displacements of magnetic-field lines imbedded in a plasma.

Alfvén Wave: A low-frequency wave in a magnetized plasma traveling at the Alfvén speed.

Ambipolar Diffusion: Collisional diffusion of ions and electrons to the wall at rates that are made equal by an electrostatic field.

Ballooning Mode: A plasma instability localized to a region in which the magnetic field has unfavorable curvature (see **Magnetic Curvature).**

Banana Orbit: The curve, in a cross-sectional plane of a toroidal confinement device such as a tokamak, traced by an ion or electron as it moves along the magnetic-field lines. These orbits are caused by the magnetic mirror effect, which reflects some particles before they can reach the strong-field region on the inside of the torus.

Bernstein Wave: Electrostatic waves near the ion- and electron-cyclotron frequencies and their harmonics.

Beta: Ratio of plasma pressure to magnetic-field pressure.

Blanket: A meter-thick region surrounding the plasma in a fusion reactor, containing compounds of lithium and used to absorb the product neutrons and convert their energy into heat.

Blumlein: A high-voltage transmission line, filled with water or oil, used for pulse shaping in a pulsed-power device.

Bohm Diffusion: Instability- or convection-driven transport of plasma across a magnetic field, which not only is much faster than collisional diffusion but also scales more weakly with field strength.

Bohm–Gross Wave: An **Electron Plasma Wave.**

Boltzmann Distribution: The Gaussian velocity distribution of a gas in thermal equilibrium.

Boltzmann Relation: Relation between the local electric potential and the density of a charged species with a Boltzmann velocity distribution.

Bootstrap Current: In a closed fusion device, a self-induced toroidal current driven by the momentum loss of radially escaping plasma.

Breakeven: In fusion reactors, the condition for the nuclear energy release to exceed the energy used to produce the plasma.

Bremsstrahlung: X radiation emitted by the electrons of a plasma by the accelerated charge effect when they collide with ions.

Bremsstrahlung, Inverse: Classical collisional absorption of laser light.

CARM: Cyclotron AutoResonant Maser: one of several types of microwave generator based on the cyclotron motion of electrons in a magnetic field.

Caviton: A density cavity in a plasma caused by the radiation pressure of electromagnetic radiation trapped in it.

Chaos: The mathematically random behavior of a variable subject to a small perturbation from a well-organized, periodic state.

Child–Langmuir Sheath: The layer surrounding a charged electrode, which contains only one charge species and in which the potential varies according to the theory of space-charge-limited diodes.

Collapse: The ultimate fate of Langmuir turbulence, according to theory: Langmuir waves become trapped in density cavities that they themselves create, and these cavities become smaller and deeper until the wave energy is dissipated by the acceleration of fast electrons.

Confinement Time: Particle confinement time is the average time an ion–electron pair resides in a plasma before it is lost to the walls; energy confinement time is the ratio of the heat content of a plasma to the power input in steady state.

Coulomb Collision: Deflection of an ion or electron by the electric field of another charged particle.

Coulomb Logarithm ($\ln\Lambda$): A numerical factor of order 10 that accounts for the cumulative effect of many small-angle deflections; here Λ is the ratio of the maximum to minimum impact parameters used in the calculation.

Current Drive: The induction of a steady toroidal current in a closed fusion device by the entrainment of electrons in a toroidally propagating wave.

Cutoff Frequency: A characteristic frequency of a wave at which its phase velocity becomes infinite and beyond which it cannot propagate.

Cyclotron Frequency: The frequency at which an ion or electron executes circular orbits in a magnetic field.

Cyclotron Radiation: Radiation at the cyclotron frequency and its harmonics emitted by electrons gyrating in a magnetic field; also called **Synchrotron Radiation.** The origin of signals received from space, as well as a cause of energy loss from fusion plasmas.

Debye Length: A characteristic length below which a plasma behaves like individual particles without long-range cooperative motions.

Debye Sheath: A thin layer, of the order of several Debye lengths, separating a plasma from walls and electrodes. The sheath forms a self-adjusting potential barrier to equalize the escape rates of the positive and negative charge carriers.

Debye Shielding: The formation of sheaths around externally imposed electrodes to prevent strong electric fields from appearing in a highly conductive plasma.

Debye Sphere: A sphere with radius equal to a Debye length.

Diamagnetism: The tendency for a material to generate magnetic fields opposite to those imposed on it; plasmas are diamagnetic because of the field generated by gyrating particles.

Diffusion, Anomalous: Transport across magnetic fields at a rate exceeding that expected from classical Coulomb collisions, and caused by such effects as instabilities, convection in dc electric fields, and magnetic errors.

Diffusion, Bohm: See **Bohm diffusion.**

Diffusion, Classical: Transport of particles or heat via ordinary Coulomb collisions.

Diffusion, Gyro-Bohm: An anomalous diffusion rate smaller than the Bohm rate by the ratio of Larmor radius to plasma size.

Diffusion, Neoclassical: Classical diffusion in a torus, embodying no anomalous processes other than predictable toroidal effects such as the trapping of particles in **Banana Orbits.**

Direct Drive: In laser fusion, the use of laser light to compress the fuel pellet without first converting the radiation to x rays.

Disruption: A sudden termination, unpredictable and sometimes violent, of the discharge in a tokamak, resulting in deposition of the stored energy into the walls.

Double Layer: A region inside a plasma that contains alternating layers of excess positive and negative charge and that is sustained by the presence of non-Maxwellian velocity distributions.

Drift Instability: Unstable growth of drift waves driven by a pressure gradient; also called a **Universal Instability** in the collisionless limit.

Drift Wave: A low-frequency wave, with a phase velocity near the diamagnetic drift velocity, which exists only in nonuniform plasmas.

ECRH: Electron-cyclotron resonance heating, a method for heating plasmas with microwave power at the electron-cyclotron frequency. Also called ECRF: electron-cyclotron range of frequencies.

Electron Plasma Wave: A Langmuir oscillation that propagates by virtue of finite electron temperature; also called **Langmuir Wave** or **Bohm–Gross Wave.**

Energy Principle: A formulation, based on the calculus of variations, of the criterion for stability or instability of a plasma in a magnetic field of arbitrary shape.

Extraordinary Wave: An electromagnetic wave propagating in a plasma in a direction normal to the background magnetic field.

Faraday Rotation: Rotation of the plane of polarization of a light wave traveling through an anisotropic medium, such as a magnetized plasma.

Favorable Curvature: Concavity of magnetic-field lines that bulge toward the center of the plasma rather than in the normal, convex direction; this sign of the curvature stabilizes the **Gravitational Instability.**

FEL: Free-electron laser: a generator of coherent radiation based on the excitation of waves of upshifted frequency when an electromagnetic wave is passed through a periodically varying magnetic field in the presence of an electron beam.

Field Annihilation: See **Reconnection.**

Filamentation: In laser–plasma interactions, the tendency for an intense laser beam to break up into smaller beamlets when passing through a plasma; in magnetic fusion, the tendency for a toroidal current to do likewise, thus creating **Magnetic Islands.**

Fusion: Release of nuclear energy by the transmutation of light elements into heavier ones; the conversion of heavy hydrogen into helium requires a hot plasma in near-thermal equilibrium and is then called **Thermonuclear Fusion** or **Controlled Fusion.**

Gravitational Instability: A hydrodynamic instability of a plasma confined by a magnetic field; the centrifugal force of particles following curved field lines acts like a gravitational force in pushing small ripples in the plasma surface into large bulges.

Gyrotron: One of several types of microwave generator based on the cyclotron motion of electrons in a magnetic field.

H-Mode: A regime of operation of tokamaks characterized by High confinement, small internal gradients, and an edge layer with highly sheared electric fields that forms a transport barrier.

ICRH: Ion-cyclotron resonance heating: a method for heating plasmas with radio-frequency power at the ion-cyclotron frequency. Also called ICRF: ion-cyclotron range of frequencies.

Ignition: Maintenance of fusion conditions in a thermonuclear plasma by the retained heat of the charged-particle products of the reactions.

Indirect Drive: In laser fusion, compression of the fuel pellet by a uniform bath of x rays, which are generated when laser beams strike the interior of a metal canister surrounding the pellet.

Inertial Confinement: Attainment of fusion conditions by using radiation or particle beams to compress the fuel to more than 1000 times solid density; at such densities, the fusion reactions take place during the instant when the fuel is held together by its own inertia.

Interchange Instability: Same as **Gravitational Instability.** The plasma and magnetic-field regions are imagined to be interchanged, with a decrease in potential energy.

Inverse Bremsstrahlung: Classical collisional absorption of laser light.

Ion Wave: Plasma analog of an acoustic, or sound, wave in air. Only one of the species in the plasma needs to have a finite temperature for the wave to exist.

Ionosphere: The region surrounding the Earth or other planet that contains a plasma.

Kink Instability: The bending or coiling of a current channel in a magnetized plasma.

Landau Damping: Reversible damping of a wave in a collisionless plasma, caused by the loss of wave energy to particles that surf on the wave.

Langmuir Oscillation: The high-frequency oscillation of electrons in a plasma when displaced from their normal positions relative to the relatively immobile ion background. These oscillations do not propagate in an infinite, cold plasma, but they become propagating **Electron Plasma Waves** in a warm plasma and propagating **Trivelpiece–Gould Modes** in a bounded plasma.

Langmuir Probe: A small electrode that, when inserted in a plasma, gives information on local temperatures, densities, and potentials via its current–voltage characteristic.

Langmuir Turbulence: The disorganized state of a plasma subject to large-amplitude Langmuir waves.

Langmuir Waves: Same as **Electron Plasma Waves.**

Larmor Orbit: The circular path of a charged particle in a magnetic field.

Lawson Criterion: The value of the product of plasma density and confinement time above which energy breakeven is achieved in a fusion reactor.

Line-Tying: In an open confinement device, the stabilizing effect of currents in conductors on which magnetic-field lines terminate.

Loss Cone: The range of pitch angles for which particles are not reflected by a magnetic mirror.

Magnetic Bottle: Popular term for fusion confinement device.

Magnetic Curvature: Technical term for the convex (unfavorable) or concave (favorable) curvature of magnetic-field lines, with reference to the stability properties of the plasma in a fusion confinement device; see **Favorable Curvature.**

Magnetic Island: One of a chain of azimuthally distributed regions in the cross section of a torus that contain an isolated group of magnetic-field lines that do not reach other parts of the cross section.

Magnetic Mirror: A region of increasing magnetic field that reflects a fraction of the impinging plasma particles; also refers to a fusion confinement system consisting of a set of two magnetic mirrors.

Magnetic Trap: Same as **Magnetic Well.**

Magnetic Well: A magnetic configuration containing a point from which the field strength increases in all directions.

Magnetosphere: The region surrounding a planet in which the planet's magnetic field is dominant over the interplanetary magnetic field.

Marx Bank: A capacitor bank that is charged in parallel but discharged in series, in order to obtain high voltages.

MHD: MagnetoHydroDynamics: a discipline in which the plasma is treated as a conducting fluid without regard for individual particle motions or deviations from thermal equilibrium.

Minimum-B: Characterizing an absolute **Magnetic Well**; a minimum-$\bar{B}$ configuration is an average magnetic well, in which the average **Magnetic Curvature** experienced by a particle is in the favorable direction.

Mirror Ratio: The ratio between maximum and minimum field strengths in a magnetic mirror.

Modulational Instability: In nonlinear plasma theory, an instability in which the unstable waves develop a growing ripple in the envelope of their amplitudes.

NBI: Neutral Beam Injection: the leading method for heating the plasma in controlled fusion experiments.

Nonlinear: Beyond the scope of validity of small-amplitude, linear theory, in which the behavior of a system can be described by a simple sum over Fourier components.

Non-neutral Plasma: A plasma containing only positive or only negative charges.

Parametric Instability: A phenomenon occurring in laser–plasma interactions in which one wave is excited by another (usually the laser beam), but only in the presence of a third wave. The name comes from *parametric amplifiers* in electronics.

Particulate (adj.): Possessing the characteristics of individual particles, rather than of a continuous fluid.

Particulate (n.): A charged "dust" grain, of the order of 1 μm to 1 mm in size, usually negative in industrial plasmas and positive in space plasmas.

PECVD: Plasma-Enhanced Chemical Vapor Deposition.

Pinch: An electrical discharge in which the plasma is squeezed by the magnetic-field pressure generated by currents within the plasma.

Plasma Dispersion Function: A complex function of a complex argument, found by contour integration, arising in the theory of collisionless plasmas.

Plasma Etch: A step in the manufacture of semiconductors in which material is removed chemically with the aid of a plasma.

Plasma Frequency: The characteristic electron oscillation frequency in a plasma; see **Langmuir Oscillation.**

Plasma Instability: A growing perturbation or wave in a plasma.

Plasma Processing: The use of plasmas in manufacturing.

Plasma Torch: An atmospheric-pressure discharge used for treatment of materials or wastes.

Plasma Wave: Any wave in a plasma, but usually referring to an **Electron Plasma Wave.**

Ponderomotive Force: The force exerted on a plasma by a wave with a gradient in its amplitude; radiation pressure modified by the presence of plasma.

PSII: Plasma-Source Ion Implantation.

Pulsed Power: A discipline involving the generation of terawatt bursts of energy by means of **Marx Banks** and **Blumleins,** usually resulting in an intense electron or ion beam.

q: A quality factor of tokamaks, inversely proportional to the degree of twist of the magnetic-field lines.

Quarter-Critical Layer: In laser–plasma interactions, the density region in which the plasma frequency is $\frac{1}{4}$ of the laser frequency.

Quasineutrality: A characteristic trait of plasmas in which the positive and negative charges move so as to maintain almost equal numbers everywhere, but with small deviations to permit the existence of electric fields.

Rayleigh–Taylor Instability: The hydrodynamic instability of a heavy fluid supported by a light fluid, which, in plasma physics, is the magnetic field or laser beam; see **Gravitational Instability.**

REB: Relativistic Electron Beam generated by **Pulsed Power.**

Reconnection: The joining of one group of magnetic-field lines to another in the presence of plasma. Since plasma particles are entrained by magnetic fields, reconnection can occur only if the particles can cross field lines, and reconnection can occur only in a region, no matter how thin, in which dissipation, including Landau damping, is important.

Resonance Absorption: Absorption of laser radiation by the generation and subsequent damping of plasma waves.

RIE: Reactive Ion Etching, a common process in the fabrication of semiconductors; also a type of plasma discharge used for this.

Resistive Instability: a hydromagnetic instability that occurs only if the plasma resistivity is nonzero.

Rotational Transform: In a torus, the number of turns a magnetic-field line encircles the minor axis for each time around the major axis; the reciprocal of ***q***.

Sawtooth: A temperature or density fluctuation with triangular wave form observed in tokamaks; thought to be caused by small **Kink Instabilities** and to be the mechanism by which a tokamak plasma adjusts its internal current distribution.

SBS: Stimulated Brillouin Scattering: a **Parametric Instability** in which a laser beam excites an ion-acoustic wave and a reflected light wave.

Shear: The change of pitch angle of magnetic-field lines on different surfaces; the radial variation of **Rotational Transform** or ***q***.

Simulation, Computer: Numerical calculations, stepwise in time, following the development of plasma motions.

Simulation, Particle: A computer simulation in which the particle motions, rather than fluid elements or distribution functions, are followed.

Solar Wind: A stream of ions and electrons ejected from the sun.

Soliton: A wave with only one maximum; a solitary wave.

Spitzer Resistivity: Resistivity of a fully ionized plasma accounting only for binary Coulomb collisions of electrons with ions.

SRS: Stimulated Raman Scattering: A **Parametric Instability** in which a laser beam excites an electron plasma wave and a reflected light wave.

Strongly Coupled Plasma: A plasma of such high density and low temperature that the number of particles in a Debye sphere is of order unity or below.

Synchrotron Radiation: Same as **Cyclotron Radiation.**

Tandem Mirror: A magnetic mirror fusion device plugged at both ends with short mirror sections containing a hot, dense plasma.

Tearing Mode: A type of **Resistive Instability** driven by current and involving **Reconnection.**

Thermonuclear: Pertaining to nuclear reactions in a Maxwellian plasma.

Theta Pinch: A **Pinch** in which the current is in the azimuthal direction.

Thomson Scattering: A plasma diagnostic technique in which laser light is scattered by plasma particles or density fluctuations.

Tokamak: A toroidal fusion device in which the magnetic configuration is determined by the distribution of toroidal current within the plasma.

Toroidal Confinement: Trapping of plasma in a closed magnetic-field configuration.

Trapped Particle: In tokamaks, ions or electrons that are prevented from circulating around the minor axis by the magnetic mirror effect of the strong fields near the major axis.

Turbulence: A statistically steady state of random fluctuations outside the realm of linear theory.

Universal Instability: A **Drift Wave Instability** occurring in collisionless plasmas.

Van Allen Belts: Plasma trapped in the Earth's magnetic field by the magnetic mirror effect.

Vlasov Equation: An equation describing the evolution of velocity distribution functions in the absence of collisions.

Whistler Wave: An electromagnetic wave in a plasma propagating along magnetic-field lines; first detected in radio reception as whistling sounds from the ionosphere.

Z Pinch: A **Pinch** in which the current is in the axial direction.

Works Cited

Alfvén, H. (1950), *Cosmical Electrodynamics*, Oxford: Clarendon Press.

Allis, W. P., Buchsbaum, S. J., Bers, A. (1963), *Waves in Anisotropic Plasmas*, Cambridge, MA: M.I.T. Press.

Baldwin, D. E. (1977), *Rev. Mod. Phys.* **49**, 317–379.

Bateman, G. (1978), *MHD Instabilities*, Cambridge, MA: M.I.T. Press.

Bekefi, G. (1966), *Radiation Processes in Plasmas*, New York: Wiley.

Bernstein, I. B., Frieman, E. A., Kruskal, M. D., Kulsrud, R. M. (1958), *Proc. R. Soc. (London) A*, **244**, 17–40.

Birdsall, C. K., Langdon, A. B. (1991), *Plasma Physics via Computer Simulation*, Bristol: Adam Hilger.

Bodner, S. E. (1991), in: A. Rubenchik, S. Witkowski (Eds.), *Physics of Laser Plasma*, Vol. 3 of: M. N. Rosenbluth and R. Z. Sagdeev (Eds.), *Handbook of Plasma Physics*, Amsterdam: North-Holland, p. 247.

Boulos, M. I., Fauchais, P., Pfender, E. (1994), *Thermal Plasmas*, Vol. 1, New York: Plenum.

Braginskii, S. I. (1965), *Transport Processes in a Plasma*, Vol. 1 of: M. A. Leontovich (Ed.), *Reviews of Plasma Physics*, New York: Consultants Bureau.

Brandt, J. C. (1970), *Introduction to the Solar Wind*, San Francisco: W. H. Freeman.

Chapman, S., Cowling, T. G. (1939), *Mathematical Theory of Non-Uniform Gases*, Cambridge, UK: Cambridge Univ. Press.

Chandrasekhar, S. (1961), *Hydrodynamic and Hydromagnetic Stability*, Oxford: Clarendon.

Chen, F. F. (1965), *Phys. Rev. Lett.* **15**, 381–383.

Chen, F. F. (1983), *Introduction to Plasma Physics and Controlled Fusion*, 2nd ed., Vol. 1, *Plasma Physics*, New York: Plenum.

Chen, F. F. (1991a), in: A. Rubenchik, S. Witkowski (Eds.), *Physics of Laser Plasma*, Vol. 3 of: M. N. Rosenbluth and R. Z. Sagdeev (Eds.), *Handbook of Plasma Physics*, Amsterdam: North-Holland, p. 483.

Chen, F. F. (1991b), *Double Helix: the Dawson Separation Process*, in: T. Katsouleas (Ed.), *From Fusion to Light Surfing*, Redwood City, CA: Addison-Wesley.

Conrad, J. R., Sridharan, K. (Eds.), Papers from the 1st International Workshop on Plasma-Based Ion Implantation, *J. Vac. Sci. Technol. B* **12**, 807–1006.

d'Agostino, R. (1990), *Plasma Deposition, Treatment, and Etching of Polymers*, New York: Academic.

D'Angelo, N., von Goeler, S. (1966), *Phys. Fluids* **9**, 309–313.

Davidson, R. C. (1972), *Methods in Nonlinear Plasma Theory*, New York: Academic.

Davidson, R. C. (1974), *Theory of Nonneutral Plasmas,* Reading, MA: W. A. Benjamin.

Dawson, J. M. (1961), *Phys. Fluids* **4**, 869–874.

Dawson, J. M. (1981), *Advanced Fusion Reactors,* in: E. Teller (Ed.), *Magnetic Confinement,* Vol. 1, Part B, *Fusion,* New York: Academic.

Dawson, J. M., Lin, A. T. (1984), in: A. A. Galeev, R. N. Sudan (Eds.), *Basic Plasma Physics II,* Vol. 2, Part 7 of: M. N. Rosenbluth and R. Sagdeev (Eds.), *Handbook of Plasma Physics,* Amsterdam: North-Holland.

Duderstadt, J. J., Moses, G. A. (1982), *Inertial Confinement Fusion,* New York: Wiley.

Fowler, T. K. (1981), in: E. Teller (Ed.), *Magnetic Confinement,* Vol. 1, Part A, *Fusion,* New York: Academic, p. 291.

Fowler, T. K., Post, R. F. (1966), *Sci. Am.* **215**, (6), 21–31.

Fried, B. D., Conte, S. D. (1961), *The Plasma Dispersion Function,* New York: Academic.

Freidberg, J. P. (1987), *Ideal Magnetohydrodynamics,* New York: Plenum.

Früngel, F. B. A. (1976), *High Speed Pulse Technology,* New York: Academic.

Furth, H. P. (1981), *The Tokamak,* in: E. Teller (Ed.), *Magnetic Confinement,* Vol. 1, Part A, *Fusion,* New York: Academic.

Galeev, A. A., Sagdeev, R. Z. (1983), in: A. A. Galeev, R. N. Sudan, *Basic Plasma Physics I,* Vol. 1, Part 4 of: M. N. Rosenbluth and R. Z. Sagdeev (Eds.), *Handbook of Plasma Physics,* Amsterdam: North-Holland.

Ginzburg, V. L. (1964), *Propagation of Electromagnetic Waves in Plasmas,* Oxford: Pergamon.

Griem, H. R., Lovberg, R. H. (Eds.) (1970), *Methods of Experimental Physics,* Vol. 9, *Plasma Physics,* Part A, New York: Academic.

Hasegawa, A. (1975), *Plasma Instabilities and Nonlinear Effects,* New York: Springer-Verlag.

Helliwell, R. A. (1965), *Whistlers and Related Ionospheric Phenomena,* Stanford: Stanford Univ. Press.

Hinton, F. L., Hazeltine, R. D. (1976), *Rev. Mod. Phys.* **48**, 239–308.

Hockney, R. W., Eastwood, J. W. (1988), *Computer Simulation Using Particles,* Bristol: Adam Hilger.

Huddlestone, R. H., Leonard, S. L. (Eds.) (1965), *Plasma Diagnostic Techniques,* New York: Academic.

Ichimaru, S. (1994), *Condensed Plasmas,* Vol. 2 of: D. Pines (Ed.), *Statistical Plasma Physics,* Reading, MA: Addison-Wesley.

Jackson, J. D. (1960), *Plasma Phys.* **1**, 171–189.

Jassby, D. L. (1972), *Phys. Fluids* **15**, 1590.

Kadomtsev, B. B. (1965), *Plasma Turbulence,* New York: Academic.

Kadomtsev, B. B., Pogutse, O. P. (1971), *Nucl. Fusion* **11**, 67–92.

Kauffman, R. (1991), in: A. Rubenchik, S. Witkowski (Eds.), *Physics of Laser Plasma,* Vol. 3 of: M. N. Rosenbluth and R. Z. Sagdeev (Eds.), *Handbook of Plasma Physics,* Amsterdam: North-Holland, p. 111.

Kruskal, M. D. (1965), in: *Plasma Physics, Lectures Presented at the International Centre for Theoretical Physics, Trieste, 1964,* Vienna: International Atomic Energy Agency.

Landau, L. D. (1946), *J. Phys. (USSR),* **10**, 25–34.

Lichtenberg, A. J., Lieberman, M. A. (1983), *Regular and Stochastic Motion,* New York: Springer-Verlag.

Lieberman, M. A., Gottscho, R. A. (1994), *Design of High-Density Plasma Sources for Materials Processing,* Vol. 18 of: M. Francomb, J. Vossen (Eds.), *Physics of Thin Films,* New York: Academic, pp. 1–119.

Lieberman, M. A., Lichtenberg, A. J. (1994), *Principles of Plasma Discharges and Materials Processing,* New York: Wiley.

Lochte-Holtgreven, W. (Ed.) (1968), *Plasma Diagnostics,* Amsterdam: North-Holland.

Manos, D. M., Flamm, D. L. (1989), *Plasma Etching,* New York: Academic.

Mikhailovskii, A. B. (1974), *Theory of Plasma Instabilities,* Vols. 1 and 2, New York: Consultants Bureau.

Miller, R. B. (1982), *Intense Charged Particle Beams,* New York: Plenum.

Mori, W. (1990), private communication.

Mott, N. F., Massey, H. S. W. (1933), *Theory of Atomic Collisions,* Oxford: Clarendon.

Nicholson, D. R. (1983), *Introduction to Plasma Theory,* New York: Wiley.

Nishikawa, K., Liu, C. S., Kaw, P. K., Kruer, W. L. (1976), in: A. Simon and W. B. Thompson (Eds.), *Advances in Plasma Physics,* Vol. 6, New York: Wiley.

Northrop, T. G. (1963), *The Adiabatic Motion of Charged Particles,* New York: Wiley.

Ott, E. (1993), *Chaos in Dynamical Systems,* Cambridge, UK: Cambridge Univ. Press.

Post, R. F. (1981), in: E. Teller (Ed.), *Magnetic Confinement,* Vol. 1, Part A, *Fusion,* p. 358.

Roberson, C. W., Sprangle, P. (1989), *Phys. Fluids B* **1**, 3–42.

Sagdeev, R. Z., Galeev, A. A. (1969), *Nonlinear Plasma Theory,* New York: W. A. Benjamin.

Sagdeev, R. Z., Shapiro, V. D., Shevchenko, V. I. (1991), in: A. Rubenchik, S. Witkowski (Eds.), *Physics of Laser Plasma,* Vol. 3 of: M. N. Rosenbluth and R. Z. Sagdeev (Eds.), *Handbook of Plasma Physics,* Amsterdam: North-Holland, p. 296.

Selwyn, G. S. (1993), *Optical Diagnostic Techniques for Plasma Processing,* New York: American Vacuum Society.

Sheffield, J. (1975), *Plasma Scattering of Electromagnetic Radiation,* New York: Academic.

Simon, A. (1963), *Phys. Fluids* **6**, 382–388.

Spitzer, L., Jr. (1967), *Physics of Fully Ionized Gases,* 2nd ed., New York: Wiley.

Stix, T. H. (1992), *Waves in Plasmas,* Woodbury, NY: AIP.

Swanson, D. G. (1989), *Plasma Waves,* Boston: Academic.

Turchi, P. J. (1980), *Megagauss Physics and Technology,* New York: Plenum.

Vitello, P., Cerjan, C., Braun, D. (1992), *Phys. Fluids B* **4**, 1447–1456.

Wesson, J. A. (1978), *Nucl. Fusion* **18**, 87–132.

White, R. B. (1983), in: A. A. Galeev and R. N. Sudan (Eds.), *Basic Plasma Physics I,* Vol. 1 of: M. N. Rosenbluth, R. Z. Sagdeev (Eds.), *Handbook of Plasma Physics,* Amsterdam: North-Holland, p. 611.

Yamanaka, C. (1991), in: A. Rubenchik and S. Witkowski (Eds.), *Physics of Laser Plasma,* Vol. 3 of: M. N. Rosenbluth, R. Z. Sagdeev (Eds.), *Handbook of Plasma Physics,* Amsterdam: North-Holland, p. 1.

Zakharov, V. E. (1984), in: A. A. Galeev and R. N. Sudan (Eds.), *Basic Plasma Physics II,* Vol. 1, Part 5 of: M. N. Rosenbluth, R. Z. Sagdeev (Eds.), *Handbook of Plasma Physics,* Amsterdam: North-Holland.

Further Reading

Bishop, A. S. (1958), *Project Sherwood,* Reading, MA: Addison-Wesley.

Chen, F. F. (1983), *Introduction to Plasma Physics and Controlled Fusion,* 2nd ed., Vol. 1, *Plasma Physics,* New York: Plenum.

Franklin, R. N. (1976), *Plasma Phenomena in Gas Discharges,* Oxford: Clarendon.

Galeev, A. A., Sudan, R. N. (Eds.) (1983), *Basic Plasma Physics I,* Vol. 1 of: M. N. Rosenbluth, R. Z. Sagdeev (Eds.), *Handbook of Plasma Physics,* Amsterdam: North-Holland.

Galeev, A. A., Sudan, R. N. (Eds.) (1984), *Basic Plasma Physics II,* Vol. 2 of: M. N. Rosenbluth, R. Z. Sagdeev (Eds.), *Handbook of Plasma Physics,* Amsterdam: North-Holland.

Jackson, J. D. (1962), *Classical Electrodynamics,* New York: Wiley, Chap. 10.

Krall, N. A., Trivelpiece, A. W. (1973), *Principles of Plasma Physics,* New York: McGraw-Hill.

Miyamoto, K. (1976), *Plasma Physics for Nuclear Fusion,* Cambridge, MA: M.I.T. Press.

National Research Council (1986), *Physics through the 1990s: Plasmas and Fluids,* Washington: National Academy Press.

Rubenchik, A., Witkowski, S. (Eds.) (1991), *Physics of Laser Plasma,* Vol. 3 of: M. N. Rosenbluth, R. Z. Sagdeev (Eds.), *Handbook of Plasma Physics,* Amsterdam: North-Holland.

Schmidt, G. (1979), *Physics of High Temperature Plasmas,* 2nd ed., New York: Academic.

Sturrock, P. A. (1994), *Plasma Physics, an Introduction to the Theory of Astrophysical, Geophysical, and Laboratory Plasmas,* Cambridge, UK: Cambridge Univ. Press.

Teller, E. (Ed.) (1981), *Magnetic Confinement,* Vol. 1, Parts A and B, *Fusion,* New York: Academic.

Vol. 14 – PLATE 1. The *Aurora Borealis*, an example of a plasma naturally occurring on earth. The glow is excited by electrons precipitating from the van Allen radiation belts.

Vol. 14 – PLATE 2. Electrical discharges in the form of lightning produce plasmas at atmospheric pressure.

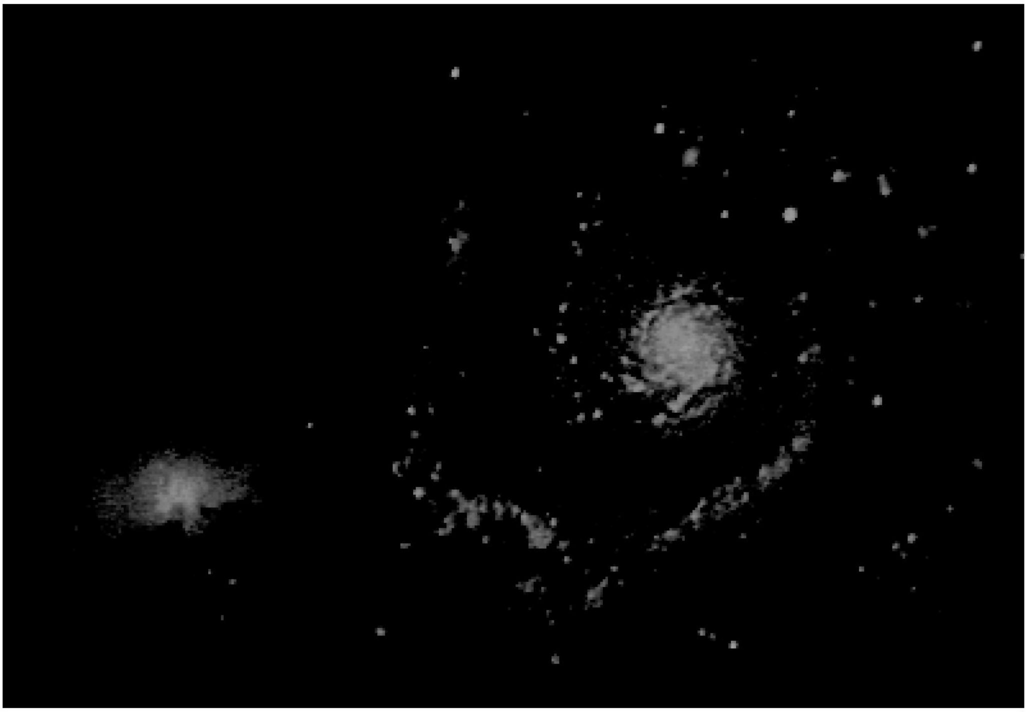

Vol. 14 – PLATE 3. Ionized gases, dust, and magnetic fields are the substances of galaxies as well as of plasmas used in industry. The spiral arms are thought to be caused by an instability of the "plasma", the particles of which, in this case, are stars.

PLASMA WAVES AND INSTABILITIES

D. G. SWANSON, *Physics Department, Auburn University, Auburn, Alabama, U.S.A.*

Introduction 237
1. **Unmagnetized Plasmas—an Overview** 238
1.1 Plasma Oscillations 238
1.2 Electromagnetic Waves 240
2. **Cold-Plasma Waves** 240
2.1 Cold-Plasma Dispersion Relation 241
2.2 The CMA Diagram 242
2.3 Phase and Group Velocity 243
2.4 Ray Tracing 244
2.5 Polarization 244
2.6 Electrostatic Waves 244
2.6.1 Resonance Cones 245
2.7 Beams and Instabilities 245
2.7.1 Convective and Absolute Instabilities 246
2.7.2 Beam–Plasma Instability 246
2.7.3 Two-Stream Instability 246
3. **Warm-Plasma Waves** 247
3.1 Fluid Models 247
3.1.1 Bohm–Gross Dispersion Relation 247
3.1.2 Low-Frequency Dispersion Relation 248
3.1.3 MHD Region 248
3.1.4 Collisions and Resistivity 249
3.1.5 Drift Waves 249
3.1.6 Drift Resistive Instability 250
3.2 Kinetic Models 250
3.2.1 Landau Damping 250
3.2.2 Ion-Acoustic Wave 251
3.2.3 Hot-Plasma Dispersion Relation 251
3.2.4 Cyclotron Damping 252
3.2.5 Electrostatic Waves 253
4. **Plasma-Filled Waveguide** 254
4.1 Boundary Conditions 254
4.2 Unmagnetized Plasmas 255
4.3 Magnetized Plasmas 256
4.3.1 Wave Fields 256
4.3.2 MHD Approximation 256
4.3.3 Loop Antenna (MHD) 257
4.3.4 Electrostatic Case 257
5. **Emission and Absorption** 258
5.1 Simple Resonance 258
5.2 Tunneling Layers 258
5.3 Cyclotron Harmonics 259
5.3.1 Ion Cyclotron Case 259
5.3.2 Electron Cyclotron Case 260
5.4 Cyclotron Emission 260
6. **Nonlinear Waves** 261
6.1 Weak Turbulence 261
6.1.1 1-D Theory 262
6.2 Parametric Instabilities 264
6.2.1 Coupled-Mode Case 265
6.2.2 Unmagnetized Plasma 266
6.2.2.1 Case 1—The Pump Is a Plasma Wave 266
6.2.2.2 Case 2—The Pump Is an Electromagnetic Wave 266
6.2.3 Brillouin Scattering 267
6.2.4 Raman Scattering 267
6.3 Shock Waves 267
6.4 Solitons 268
6.4.1 The Korteweg–de Vries (KdV) Equation 268
6.4.2 Ion-Acoustic Solitons 269
6.4.3 Alfvén-Wave Solitons 269
6.4.4 Nonlinear Schrödinger Equation 269
6.4.5 General Wave Collapse 270
Glossary 271
Works Cited 271
Further Reading 272

INTRODUCTION

Plasma waves are fundamental to the very definition of a plasma, since a plasma is first a collection of charged particles that is essentially neutral, but it must also be a minimum size. For any finite temperature, there is a characteristic thermal speed $v_{th} \equiv (2\kappa T/m)^{1/2}$; and plasmas are characterized by the plasma frequency, $\omega_p = (ne^2/m_e\epsilon_0)^{1/2}$, which is the frequency of oscillation when the plasma density is locally disturbed from equilibrium, creating a local charge imbalance, and the

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Debye length, $\lambda_D = (\epsilon_0 \kappa T/ne^2)^{1/2}$, is roughly the distance a thermal particle travels in a plasma period. This is a fundamental length in a plasma since it represents an effective distance beyond which fields from individual particles are essentially shielded. A plasma must generally have many particles in a Debye sphere to prevent binary collisions from being dominant, and must be many Debye lengths across for it to be dominated by the long-range electrostatic forces between particles rather than external fields. Near the edges of plasmas, this latter condition invariably fails, so that when static external electric fields are applied to a plasma, the fields inside are typically small compared with those near the edge. It is possible to generate large oscillating fields inside, however, if one couples to one of the characteristic plasma waves.

Plasmas support a rich variety of waves and are subject to a wide range of instabilities. Even in a cold plasma, there are generally two types of waves at every frequency, but the variation in angle between the wave vector and the magnetic field in a magnetized plasma adds significant complications, and there are cutoffs and resonances at virtually all angles of propagation. When thermal effects are added, many new but very important phenomena are added, including collisionless damping and numerous waves that have no cold-plasma counterpart. The addition of plasma gradients and nonlinearities adds instabilities and other new phenomena, leading to a bewildering array of effects to be considered. While some of these waves can be examined with a minimal number of effects, some require a significant level of sophistication to explore. In this article, a description of the broad variety of effects is first given for an unmagnetized plasma with few details. For the more general case with a magnetic field, cold-plasma theory is then presented as a foundation. Warm-plasma effects are then treated through both fluid models with finite plasma pressure and kinetic models. The examples are chosen to be illustrative rather than exhaustive. Some effects of boundaries are included, as well as some effects of inhomogeneity, especially in terms of instabilities and in terms of absorption and emission. Finally, there is some discussion of weak plasma turbulence but little of strong plasma turbulence. Several fundamentally nonlinear plasma phenomena are described, but the full range of nonlinear waves of practical interest is beyond the scope of this article.

For every topic covered, the literature is too vast to give all of the appropriate references. In the books listed under Further Reading, two are textbooks on plasma waves and give additional details and references on virtually every topic covered here.

1. UNMAGNETIZED PLASMAS—AN OVERVIEW

In the absence of a magnetic field and with no thermal motions of the particles, the motion of particles in response to an oscillating wave field is relatively trivial. From the equation of motion, assuming plane-wave dependence $[v_j, E \sim \exp(i\mathbf{k}\cdot\mathbf{r} - i\omega t)]$, one finds $v_j = iq_jE/\omega m_j$ for species j and the current by $J = \Sigma_j q_j n_j v_j$, where the sum is typically over $j = i, e$ for the ions and electrons, although multiple ion species are possible. The vector wave equation then becomes

$$\mathbf{k}(\mathbf{k}\cdot\mathbf{E}) - k^2\mathbf{E} + (\omega^2/c^2)K\mathbf{E} = 0, \tag{1}$$

where $K = 1 - (\omega_{pe}^2 + \omega_{pi}^2)/\omega^2$ with $\omega_{pj}^2 = n_j q_j^2/m_j\epsilon_0$. The two independent solutions are $K = 0$ when $\mathbf{k} \parallel \mathbf{E}$, $n^2 = K$ when $\mathbf{k}\cdot\mathbf{E} = 0$, where $n = kc/\omega$ is the index of refraction.

1.1 Plasma Oscillations

In the first case, with $K = 0$, there is no wave since k (or n) is undefined, and the plasma simply oscillates undamped at the *plasma frequency*, ω_p ($\simeq \omega_{pe}$ since $m_e \ll m_i$), and these are called *plasma oscillations*. This is unrealistic, however, since with any finite temperature, this result is modified.

The first generalization is to describe the plasma by a Boltzmann equation for each species that describes the evolution of the distribution function. By taking moments of the Boltzmann equation and truncating with a simple pressure term, the frequency is found to be

$$\omega^2 = \omega_{pe}^2(1 + 3v_e^2/2v_p^2), \tag{2}$$

where $v_e^2 = 2\kappa T_e/m_e$ (κ is Boltzmann's constant) and v_e is called the *thermal speed*, and

$v_p = \omega/k$ is called the *phase velocity*. Since there is now a relation between ω and k, this describes a propagating wave. If higher moments of the Boltzmann equation are taken, the relation becomes more complicated,

$$\omega^2 = \omega_{pe}^2\left(1 + \frac{3v_e^2}{2v_p^2} + \frac{15v_e^4}{4v_p^4} + \cdots\right), \qquad (3)$$

and so it is apparent that the series expression is useful for $v_e \ll v_p$. When this is not satisfied, we need yet another method of analysis.

Instead of taking moments of the Boltzmann equation, we may solve it directly along with the Maxwell equations assuming an initial-value problem in time with wavelike periodicity in space. This *kinetic theory* analysis leads to a dispersion relation [a *dispersion relation* is an equation relating ω to k, whose general form is $D(\omega,k) = 0$] of the form

$$1 = \sum_j \frac{\omega_{pj}^2}{k^2v_j^2} Z'\left(\frac{\omega}{kv_j}\right), \qquad (4)$$

where $Z(\zeta)$ is the *plasma dispersion function* (Fried and Conte, 1961) defined by

$$Z(\zeta) \equiv \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} \frac{e^{-z^2}}{z - \zeta}\,dz, \quad \mathrm{Im}(\zeta) > 0, \qquad (5)$$

and $Z'(\zeta)$ is the derivative of $Z(\zeta)$ with respect to its argument. This function is complex for real argument, and when $\zeta \gg 1$, the dispersion relation reduces to Eq. (3) plus an exponentially small imaginary term [$\sim \exp(-\zeta^2)$], which leads to damping. The damping becomes strong when the phase velocity is near the thermal speed.

The kinetic-theory solution is much more accurate than either the cold-plasma or moment-expansion models, but it is possible to analyze the damping as being due to those particles in the distribution that travel at very nearly the same speed as the phase velocity, in which case they see a virtually stationary electrostatic potential well. As they oscillate in this potential well, they exchange energy with the wave and wave damping results. For any oscillation, of course, there must be a *finite* well depth, which implies that the damping is fundamentally *nonlinear* in character, whereas the kinetic-theory analysis is linear and assumes an infinitesimal amplitude. For truly large amplitudes, nonlinear effects become strong, leading to several different kinds of behavior. First, one finds that the number of particles trapped in the potential well increases with amplitude, and while the damping following an initial disturbance begins at the linear rate, as the particles begin to bounce and complete a full cycle, they bounce on both sides of the well and begin to give energy back to the wave and the wave amplitude may cease falling and begin to rise again. The wave amplitude may oscillate for a number of cycles, eventually settling down to a steady state where there is a group of trapped particles traveling with the wave, which maintains a steady amplitude. Below a certain threshold, there is no finite-amplitude steady-state solution, but the decay may be slower than indicated by the linear theory.

This same type of analysis indicates that if there were an unusual case where there were more trapped particles going slightly faster than the phase velocity than were going slower, whereas the usual case would have more traveling slower (the distribution function is normally a monotonically decreasing function of velocity), the wave amplitude would grow. This is characteristic of a *plasma instability* and occurs whenever the distribution is not a monotonically decreasing function of velocity, and the growth occurs over a range of phase velocities where $df/dv > 0$. As the amplitude grows, particle trapping becomes stronger and eventually saturates the amplitude, and velocity-space diffusion eventually modifies the zero-order distribution function until it is monotonic.

Whenever the amplitude is large enough to lead to a saturated amplitude, another kind of plasma instability arises, since a group of trapped particles may have a bounce frequency ω_b that is an integral submultiple of the wave frequency, so that $\omega = n\omega_b$. This leads to parametric instabilities at frequencies $\omega \pm \omega_b$, and the growth of these unstable waves eventually destroys the saturated wave until everything decays away. When there is a variety of unstable waves, the situation is described as *weak turbulence*.

Finally, for large enough amplitudes, slight variations of amplitude in space (due either to focusing or to random fluctuations) lead to ponderomotive forces that tend to expel plasma from the regions of highest wave amplitude. The creation of a region of lower density (sometimes called a *caviton*) is unstable and leads to collapse of the localized wave structure. This is a strongly nonlinear process, and the collapse of many such cavitons may lead to *strong turbulence*.

Ignoring boundaries and effects of plasma inhomogeneity, we already see the enormous range of processes that occur in plasmas, from linear to strongly nonlinear, and all of these scenarios are known in experiments. The purpose of this section is to illustrate the rich variety of effects that occur and to indicate the possible difficulty in describing accurately all of the phenomena, since many may overlap and occur simultaneously.

1.2 Electromagnetic Waves

Some, but not all, of the effects described above apply also to the transverse-wave solution where $n^2 = K$, which describes an electromagnetic wave traversing a medium with an effective dielectric constant. In this context, electromagnetic waves are those whose wave fields are primarily transverse to the direction of propagation, while the wave electric field for an electrostatic wave is essentially parallel to the propagation direction, as was the case for the plasma wave above. The transverse wave with $n^2 = K$ is called the *ordinary wave* (O wave) and also occurs in a magnetized plasma if the electric-field vector is parallel to the magnetic field. For this case, it is possible that $K < 0$, which occurs whenever $\omega < \omega_p$, and the electromagnetic wave cannot propagate. The cold-plasma model is frequently adequate for this case, whereas it is rarely adequate for the electrostatic case above since thermal effects usually dominate (with no thermal effects, there is no wave). An example of this occurs for radio waves, which propagate in free space, but upon encountering the ionosphere, they *reflect* at low frequencies and transmit at high frequencies, accounting for some of the differences between AM ($\omega < \omega_p$) and FM ($\omega > \omega_p$) propagation. This also accounts for the communications blackout during shuttle reentry and for the transparency of many insulators (low density of electrons implies low ω_p) and the shiny appearance of metals (with high electron densities).

The fluid equations are of little benefit for these waves unless collisions dominate the damping, and the kinetic-theory analysis leads to a dispersion relation of the form

$$n^2 = 1 + \sum_j \frac{\omega_{pj}^2}{\omega k v_j} Z\left(\frac{\omega}{k v_j}\right). \tag{6}$$

This appears to imply that there is damping of this wave also, since the plasma dispersion function is always complex for real argument, but this is not as obvious as before since here $v_p > c$ so that there are *no trapped particles* since no particles see a stationary potential. In fact, to treat this case properly, one must use a *relativistic* analysis so that while the momentum may vary over an infinite range in the relativistic distribution function, the velocity never exceeds the velocity of light.

An interesting example of nonlinear effects on the O wave is in the case of intense laser irradiation of a solid. After the plasma is formed, the incident wave is effectively shielded from the material as the plasma frequency rises to exceed the wave frequency. In extreme cases, the wave amplitude may be so strong that an electron is accelerated to relativistic energy in a single quarter cycle. With the increased effective mass, the effective plasma frequency is lowered, allowing the wave to penetrate slightly deeper into the plasma.

2. COLD-PLASMA WAVES

While waves in a cold plasma are the simplest to treat, the inclusion of a uniform magnetic field already makes this topic one of considerable complexity. Historically (Gillmor, 1982), the analysis is due to Wilhelm Altar and E. V. Appleton, who together found the dispersion relation for electrons only (assuming infinitely massive ions), including collisions (Appleton, 1932). The principal assumptions of cold-plasma theory are that the plasma is uniform and charge neutral and that every particle of the same species moves only in response to the wave electric field and the static magnetic field (taken to be in the z direction).

The calculation begins with the equations of motion for a single charged particle of species j in an electromagnetic field,

$$m_j \frac{d}{dt}\mathbf{v}_j = q_j(\mathbf{E} + \mathbf{v}_j \times \mathbf{B}); \tag{7}$$

the expression for the total current,

$$\mathbf{J} = \sum_j n_j q_j \mathbf{v}_j, \tag{8}$$

where the sum is over the species; and the Maxwell equations. With the wave fields and the velocity varying as $\exp(i\mathbf{k}\cdot\mathbf{r} - i\omega t)$, the wave equation may be written

$$\mathbf{k} \times (\mathbf{k} \times \mathbf{E}) + (\omega^2/c^2)\,\mathbf{K}\cdot\mathbf{E} = 0, \tag{9}$$

where the equivalent dielectric tensor $\mathbf{K}$ is

$$\mathbf{K} = \begin{pmatrix} K_1 & K_2 & 0 \\ -K_2 & K_1 & 0 \\ 0 & 0 & K_3 \end{pmatrix}, \tag{10}$$

with dielectric tensor elements given by

$$\begin{aligned} K_1 &\equiv S = 1 - \sum_j \frac{\omega_{pj}^2}{\omega^2 - \omega_{cj}^2}, \\ iK_2 &\equiv D = \sum_j \frac{\epsilon_j \omega_{cj}\omega_{pj}^2}{\omega(\omega^2 - \omega_{cj}^2)}, \\ K_3 &\equiv P = 1 - \sum_j \frac{\omega_{pj}^2}{\omega^2}, \end{aligned} \tag{11}$$

where $K_1 + iK_2 \equiv R = S + D$ and $K_1 - iK_2 \equiv L = S - D$ along with the definitions of ω_{pj}, the plasma frequency for species j given by $\omega_{pj}^2 = n_j q_j^2/m_j\epsilon_0$ where n_j is the density of species j and q_j the corresponding charge; of ω_{cj}, the cyclotron frequency for species j given by $\omega_{cj} = |q_j|B_0/m_j$; and of $\varepsilon_j \equiv q_j/|q_j| = \pm 1$. The two equivalent sets of tensor elements in Eq. (10) are the common S, D, R, L, and P notation of Stix (1962) and the cold-plasma limit of the more general dielectric tensor elements to be described later. The Stix labels are mnemonics for the Sum $[S = \frac{1}{2}(R + L)]$, Difference $[D = \frac{1}{2}(R - L)]$, Right or Left ($n^2 = R$ or L for a right- or left-handed circularly polarized wave), and Plasma terms, respectively.

2.1 Cold-Plasma Dispersion Relation

We define the vector index of refraction $\mathbf{n} = \mathbf{k}c/\omega$, whose direction corresponds to the direction of the phase velocity, and choose $\mathbf{n}$ to lie in the x-z plane so that $n_x = n_\perp = n \sin\theta$, $n_y = 0$, $n_z = n_\parallel = n\cos\theta$, where $\perp$ and $\parallel$ are with respect to the static magnetic field. A nontrivial solution of Eq. (9) requires that the determinant of coefficients vanish, which leads to

$$An^4 - Bn^2 + C = 0, \tag{12}$$

where

$$\begin{aligned} A &= S\sin^2\theta + P\cos^2\theta, \\ B &= RL\sin^2\theta + PS(1 + \cos^2\theta), \\ C &= PRL. \end{aligned} \tag{13}$$

The solutions of Eq. (12) may be written in either of two forms: as a quadratic in n^2,

$$n^2 = \frac{B \pm F}{2A}, \quad F^2 = B^2 - 4AC \geq 0, \tag{14}$$

or in terms of the angle between the phase velocity and the magnetic field,

$$\tan^2\theta = -\frac{P(n^2 - R)(n^2 - L)}{(Sn^2 - RL)(n^2 - P)}. \tag{15}$$

The general condition for a resonance, where $n^2 \to \infty$, or where $A \to 0$, is given by Eq. (13) or Eq. (15) as

$$\tan^2\theta = -P/S, \quad \text{Resonance}; \tag{16}$$

and the general cutoff condition, where $n = 0$, or $A \to \infty$, is given by Eq. (12) as

$$C = PRL = 0, \quad \text{Cutoff}. \tag{17}$$

The *principal wave types* are defined for $\theta = 0$ or for $\theta = \pi/2$.

1. Parallel propagation: $\mathbf{k} \parallel \mathbf{B}_0$.
   (a) $P = K_3 = 0$, plasma oscillations, described in Sec. 1.1;
   (b) $n_R^2 = R = K_1 + iK_2$, right-handed circularly polarized wave (R wave);
   (c) $n_L^2 = L = K_1 - iK_2$, left-handed circularly polarized wave (L wave),
2. Perpendicular propagation: $\mathbf{k} \perp \mathbf{B}_0$.
   (a) $n_O^2 = P = K_3$, ordinary wave (O wave) described in Sec. 1.2;
   (b) $n_X^2 = RL/S = (K_1^2 + K_2^2)/K_1$, extraordinary wave ($X$ wave), a transverse wave whose wave vector is perpendicular to both the magnetic and the electric field.

In ionospheric and space applications, the names for these waves sometimes vary, with the *R* wave called the ordinary wave and the *L* wave the extraordinary wave. At low frequencies, both the *R* wave and *L* wave are magnetohydrodynamic (MHD) waves, where both dispersion relations become $n^2 = 1 + c^2/V_A^2$ where V_A is the Alfvén speed given by $V_A^2 = B_0^2/\mu_0 n_i m_i$. The low-frequency *L* wave, which has a resonance at the ion cyclotron frequency, is sometimes called the torsional Alfvén wave since the magnetic field lines twist as the wave propagates. The *R* wave is often called the compressional Alfvén wave below and near the ion cyclotron frequency as the magnetic field lines alternately compress and expand, and is commonly called the whistler mode in the range $\omega_{ci} \ll \omega \ll \omega_{ce}$, and has a resonance at the electron cyclotron frequency. Above their high-frequency cutoffs, both the *L* wave and the *R* wave become simple electromagnetic waves with $n \rightarrow 1$ at very high frequencies. Plots of both the *R* wave and the *L* wave dispersion relations are given in Fig. 1 for a low-density case (and artificial mass ratio) to show the character of the cutoffs and resonances and show the range of variation of n^2 and its asymptotic limits. The *X* wave has resonances at the lower and upper hybrid resonance frequencies, given by

$$\omega_{LH}^2 = \omega_{ce}\omega_{ci}\left(\frac{\omega_{pe}^2 + \omega_{ce}\omega_{ci}}{\omega_{pe}^2 + \omega_{ce}^2}\right), \tag{18}$$

$$\omega_{UH}^2 = \omega_{pe}^2 + \omega_{ce}^2. \tag{19}$$

These are called *hybrid resonances* because they depend upon both the density (ω_{pe}) *and* the magnetic field (ω_{ce} and ω_{ci}). The upper hybrid resonance is always above ω_{ce}, and the lower hybrid resonance is always between the two cyclotron frequencies. A plot of the *X* wave dispersion relation is given in Fig. 2 and shows two cutoffs and resonances.

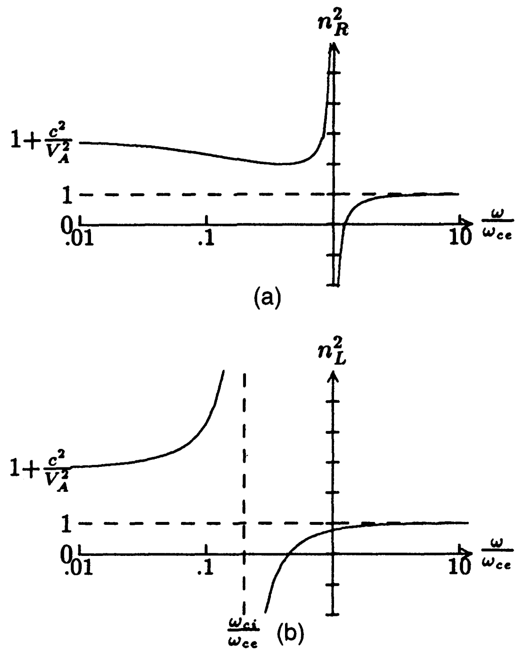

FIG. 1. Dispersion relations for (a) n_R^2 and (b) n_L^2 for $m_i/m_e = 5$, $\omega_{pe}^2/\omega_{ce}^2 = 0.3$.

2.2 The CMA Diagram

The regions of propagation and evanescence for even the principal wave types are nontrivial, and when intermediate angles are included, the picture becomes very complex. The Clemmow–Mullaly–Allis (CMA) diagram (Allis, 1959) is useful in determining which wave types are present in various density and magnetic-field regions, presenting both the number of propagating waves and the transitional boundaries where resonances and cutoffs occur. The boundaries in the CMA diagram are the resonances and cutoffs of the principal wave types, and in each region wave normal surfaces are shown where the radius of the figure is a topological sketch of the phase velocity vs angle, $v_p(\theta)$. If a wave propagates at all angles in some region, it is represented by a spheroid. The axes are effectively the magnetic-field

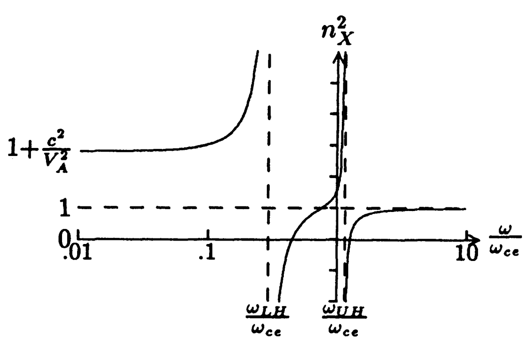

FIG. 2. Dispersion relation for the *X* wave.

strength ($\omega_{ce}/\omega \propto B_0$) vs the density ($\omega_{pe}^2/\omega^2 \propto n_e$). With $\theta = 0$ at the top of each wave normal surface, the CMA diagram is shown in Fig. 3 along with labels of the boundaries and wave types at $\theta = 0$ and $\theta = \pi/2$. If $P/S < 0$ in any bounded volume of the diagram, then there is a resonance at angle θ_r given by $\tan^2\theta_r = -P/S$. In such cases, there is a single lemniscoid in that region. The lightly dotted curve in Fig. 3, which corresponds to $RL = PS$, indicates where the X wave and the O wave cross [n^2 the same for both since $F = 0$ in Eq. (14)], but there is no resonance or cutoff on this surface. Figures 1 and 2 follow the heavy dotted trajectory from the upper right to the lower left.

2.3 Phase and Group Velocity

The wave normal surfaces represented topologically by the spheroids and lemniscoids of Fig. 3 are sometimes referred to as a "plasma pond," since a localized disturbance will exhibit waves propagating in either spheroids or lemniscoids. The surfaces display the phase velocity, however, which must be distinguished from the group velocity. Whereas the phase velocity is given by $v_p(\theta) = \omega/k(\theta)$, the group velocity is given by

$$\mathbf{v}_g = \nabla_k \omega = \left.\frac{\partial\omega}{\partial k}\right|_\theta \hat{e}_k + \frac{1}{k}\left.\frac{\partial\omega}{\partial\theta}\right|_k \hat{e}_\theta, \tag{20}$$

where $\hat{e}_k$ is a unit vector in the $\mathbf{k}$ direction, etc. The angle α between the phase and group velocities is given by

$$\tan\alpha = \left(\frac{1}{k}\left.\frac{\partial\omega}{\partial\theta}\right|_k\right)\left(\left.\frac{\partial\omega}{\partial k}\right|_\theta\right)^{-1} = -\frac{1}{k}\left.\frac{\partial k}{\partial\theta}\right|_\omega .$$

A good example of the effects of the group velocity is the guidance of the whistler wave, which is the R wave when well above the ion cyclotron frequency and well below the electron cyclotron frequency. In this case, the dispersion relation is approximately given by

$$\omega \simeq k^2c^2\omega_{ce}\cos\theta/\omega_{pe}^2, \tag{22}$$

from which $\tan\alpha = -\frac{1}{2}\tan\theta$ so that $\mathbf{v}_g$ always lies between $\mathbf{B}_0$ and $\mathbf{v}_p$. This is illustrated in Fig. 4(a) where the wave normal surface is two spheres joined at $\theta = \pi/2$, while the group velocity $v_g(\theta) \simeq kc^2\omega_{ce} \times \sqrt{1 + \cos^2\theta/\omega_{pe}^2}$ from Fig. 4(b) is a pair of separated spheres. The maximum deviation of the group velocity from the direction of

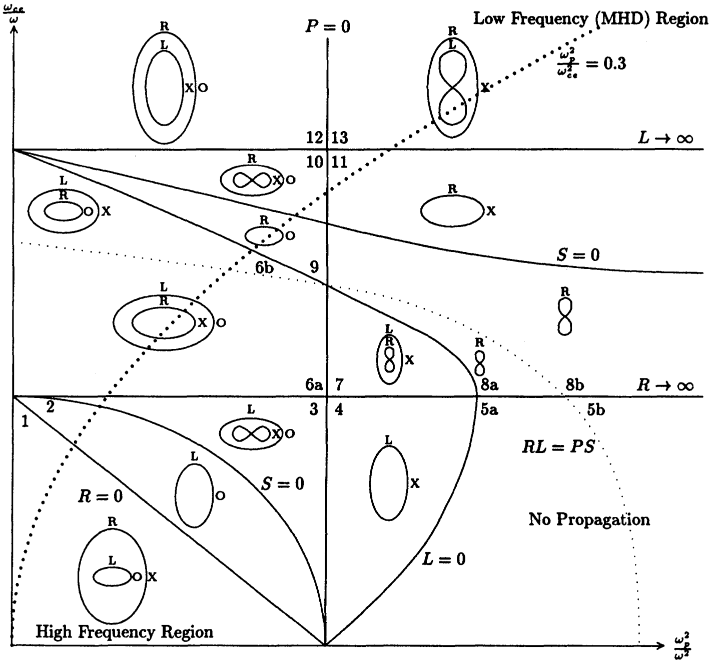

$\frac{\omega_{ce}}{\omega}$
P = 0
Low Frequency (MHD) Region
$\frac{\omega_p^2}{\omega_{ce}^2} = 0.3$
R
L
X
O
12
13
10
11
L → ∞
6b
9
S = 0
6a
7
8a
8b
R → ∞
1
2
3
4
5a
5b
RL = PS
S = 0
R = 0
No Propagation
L = 0
High Frequency Region
$\frac{\omega_p^2}{\omega^2}$

FIG. 3. CMA diagram for plasma with $m_e/m_i \ll 1$ (not to scale).

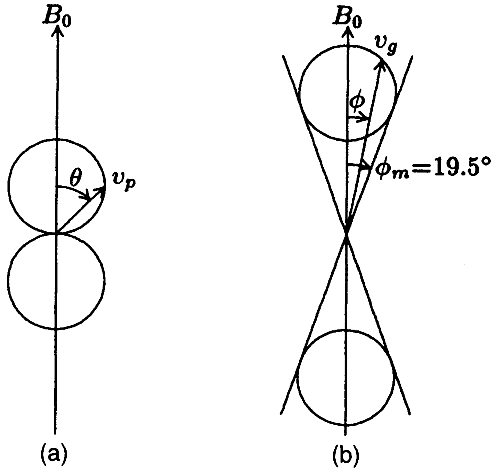

FIG. 4. (a) $v_p(\theta)$ and (b) $v_g(\phi)$ where $\phi = \theta + \alpha$ for the whistler mode.

the magnetic field is seen to be $(\theta + \alpha)_{\max} = \tan^{-1}8^{-1/2} = 19.5°$.

2.4 Ray Tracing

When a plasma is slowly varying in space (or time), these simple expressions for group velocity may not suffice. If one expresses a dispersion relation in the form $D(\mathbf{k},\omega,\mathbf{r},t) = 0$, the direction of power flow may be determined by integrating along a ray, called *ray tracing*, governed by the equations (Weinberg, 1962)

$$\frac{d\mathbf{k}}{d\tau} = \nabla D, \tag{23}$$

$$\frac{d\mathbf{r}}{d\tau} = -\nabla_k D, \tag{24}$$

$$\frac{d\omega}{d\tau} = \frac{\partial D}{\partial t}, \tag{25}$$

where τ is a parameter along the ray given by $\partial D/\partial\omega = dt/d\tau$. The energy density evolves along the ray according to

$$\frac{dU}{dt} = -U\nabla\cdot\mathbf{v}_g - \tfrac{1}{2}\omega\epsilon_0\mathbf{E}^*\cdot\mathbf{K}^A\cdot\mathbf{E}, \tag{26}$$

where U is the energy density from the Hermitian portion of $\mathbf{K}$ and $\mathbf{v}_g = \nabla_k\omega$, so that the first term is related to focusing, and $\mathbf{K}^A$ is the anti-Hermitian portion of $\mathbf{K}$, and relates to dissipation.

2.5 Polarization

We define the terms right-handed and left-handed in terms of the rotation of the electric-field vector as a wave propagates parallel to the magnetic field. If the electric vector rotates clockwise as it propagates along the B-field direction, then this is a right-handed wave, and the left-handed wave rotates counterclockwise. To see what this implies, we consider a wave with complex E_x and E_y that represent a circularly polarized R wave by requiring

$$\begin{aligned}\mathrm{Re}(E_x) &= a\cos(-\omega t) = \mathrm{Re}[a\exp(-i\omega t)],\\ \mathrm{Re}(E_y) &= -a\sin(-\omega t) = \mathrm{Re}[ia\exp(-i\omega t)],\end{aligned} \tag{27}$$

from which it is clear that the measurable fields represented by Eq. (27) with real a rotate clockwise. Thus, it follows that the phases of the waves are given by

$$\begin{aligned}iE_x &= E_y, \quad R\text{ wave},\\ iE_x &= -E_y, \quad L\text{ wave}.\end{aligned} \tag{28}$$

In order to connect this result with plasma waves, we note that the y component of Eq. (9) leads to

$$\frac{iE_x}{E_y} = \frac{n^2 - S}{D} = \pm 1 \tag{29}$$

by letting $n^2 = R, L$. Even this labeling is not always consistent for $\theta \neq 0$, since if the angle given by $\sin^2\theta_c = P/S$ is a real angle, then $n^2 = S$ there and the polarization changes direction at that angle.

2.6 Electrostatic Waves

When the phase velocity of a wave is very slow, the wave magnetic field component is typically small and if neglected, the wave may be obtained from an electrostatic potential, so that $\mathbf{E} = -\nabla\phi = -i\mathbf{k}\phi$. This leads to the electrostatic dispersion relation

$$\mathbf{k}\cdot\mathbf{K}\cdot\mathbf{k} = 0. \tag{30}$$

In cold plasmas, this reduces to $k_\perp^2 K_1 +$

$k_z^2 K_3 = 0$. For parallel propagation, this requires $K_3 = P = 0$, and so plasma oscillations are an example of an electrostatic wave. The sufficient condition for the validity of the electrostatic approximation is simply that $n^2 \gg |K_{ij}|$ for all i and j, or that the index of refraction should be large compared to all of the dielectric tensor elements. The large value of n justifies their frequently being called slow waves, since $v_p \ll c$.

2.6.1 Resonance Cones If one excites a wave in a plasma from a point source, then the radiation pattern is generally not symmetric in space, especially in those regions of parameter space where the wave normal surfaces are lemniscoids, since waves do not even propagate in all directions. In fact, in plasmas that have resonance angles, the propagation is strongly influenced by the resonance and was shown (Kuehl, 1962) to give rise to resonance cones, so called because virtually all of the wave energy is propagated from the point source in cones with a fixed angle with respect to the magnetic field. To see how this occurs, we consider electrostatic waves near the lower hybrid resonance.

We begin with an oscillating point source described by $\rho(\mathbf{r},t) = \delta(\mathbf{r}) \exp(-i\omega t)$ so that the spatial Fourier transform of the source is given by $\rho(\mathbf{k}) = 1$. As long as the source extent is small compared to the distances where measurements are to be made, this provides an adequate description of the source. The Fourier transform of the resulting potential is given by Poisson's equation so that

$$\Phi(\mathbf{k}) = \rho(\mathbf{k})/\epsilon_0 D(k_\perp^2, k_z), \tag{31}$$

where $D(k_\perp^2, k_z) = \mathbf{k}\cdot\mathbf{K}\cdot\mathbf{k}$. In cylindrical coordinates in k space with the axis aligned with the magnetic field, the inverse Fourier transform gives

$$\varphi(\mathbf{r}) = \int_{-\infty}^{\infty} \frac{dk_z e^{ik_z z}}{4\pi^2\epsilon_0} \int_0^{\infty} \frac{k_\perp J_0(k_\perp \rho) dk_\perp}{D(k_\perp^2, k_z)}, \tag{32}$$

where ρ and z are in cylindrical coordinates. The integral over $k_\perp$ has poles at $k_{\perp 0} = k_z\sqrt{-K_3/K_1}$, resulting in

$$\varphi(\mathbf{r}) = \frac{i}{8\pi\epsilon_0 K_1} \int_{-\infty}^{\infty} dk_z e^{ik_z z} H_0^{(1)}(k_{\perp 0}\rho). \tag{33}$$

For large $|\rho|$, the integral becomes a delta function, evaluating the integrand at $\rho = z \times \tan\phi$. Now ϕ is the angle of the group velocity at resonance, and so the only disturbance far from the source occurs at this particular cone angle. Using lower hybrid waves as an example of this, the radiated energy far from the source is found only on cones whose angle is given by

$$\theta_c = \tan^{-1}\sqrt{-K_1/K_3}. \tag{34}$$

These cones are illustrated in Fig. 5 where the geometry is evident. Thermal effects tend to broaden the cone and generate thermal interference fringes just inside the resonance cone angle (Fisher and Gould, 1971).

2.7 Beams and Instabilities

If one adds a drifting plasma component or beam that is moving at uniform velocity v_0 parallel to the magnetic field, the dielectric tensor elements are modified so that

$$\begin{aligned} K_1 &= K_{10} + \frac{\omega_{pb}^2}{[\omega_{cb}^2 - (\omega - k_\parallel v_0)^2]}, \\ K_2 &= K_{20} - \frac{i\epsilon_b \omega_{pb}^2 \omega_{cb}(\omega - k_\parallel v_0)}{\omega^2[\omega_{cb}^2 - (\omega - k_\parallel v_0)^2]}, \\ K_3 &= K_{30} - \omega_{pb}^2/(\omega - k_\parallel v_0)^2. \end{aligned} \tag{35}$$

This additional term adds a source of free energy in the plasma and permits unstable plasma waves with complex roots. If k is assumed complex, there are decaying and amplifying waves, or if ω is taken complex, waves grow or decay in time everywhere. Which, in fact, occurs is not immediately ap-

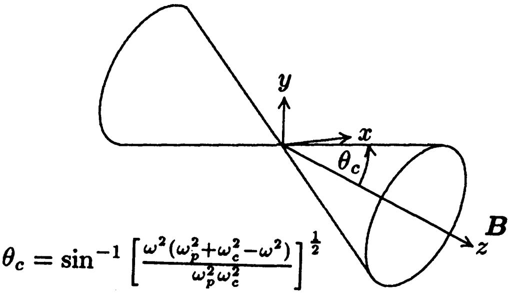
y
x
θ_c
$\mathbf{B}_0$
z
$\theta_c = \sin^{-1}\left[\frac{\omega^2(\omega_p^2+\omega_c^2-\omega^2)}{\omega_p^2\omega_c^2}\right]^{\frac{1}{2}}$

FIG. 5. Geometry of resonance cones, showing cone angle θ_c for a point source at the origin. The z axis is in the direction of $\mathbf{B}_0$.

parent without some criteria to distinguish them.

2.7.1 Convective and Absolute Instabilities The criteria for distinguishing the kinds of instabilities were first given by Sturrock (1958), but the form here is due to Briggs (1964). Waves are classified as *unstable* if for real k, ω has a positive imaginary part; a wave with complex k for real ω is called *amplifying* if it grows in space in the direction of the group velocity and *evanescent* if it decays in that direction. Furthermore, if a finite source leads to growth in time at *every* point in space, it is an *absolute* instability. Finally, a *convective* instability occurs if at a fixed point in space, a growing wave eventually decays in time as the disturbance "convects" away. These are related to amplifying waves.

For a general dispersion relation $D(\omega,k) = 0$, where a general wave is assumed to vary as $\exp(ikz - i\omega t)$, the criteria for distinguishing between amplifying and evanescent waves at a fixed real frequency ω_0 are determined by the motion of the roots of the dispersion relation as the imaginary part of ω is varied. To decide whether a given wave with a complex $k = k_r + ik_i$ for some real ω is amplifying or evanescent, determine whether or not k_i changes sign when the frequency takes on a large positive imaginary part. If it does, then the wave is amplifying; otherwise, it is an evanescent wave. This rule has a relatively simple interpretation based on the concept of a source that is growing in time exponentially. If it grows fast enough, *causality* requires that the wave decay away from the source uniformly. This means that the growth rate of an amplifying wave should *change sign* when the imaginary part of ω becomes large and positive, signifying a growing source.

An absolute instability is obtained whenever there is a double root of k for some complex ω in the upper-half ω plane for which the two merging roots come from different halves of the complex k plane (upper and lower) when ω has a large positive imaginary part. For this case, we can understand the criterion by considering a point source in space that is again growing faster than any instability so that the wave must decay away from the source from causality. When we consider the wave to the right ($z > 0$, k_+) along with the wave traveling to the left ($z < 0$, k_-), in general one has a discontinuity in matching either k or its derivative that is provided by the source. If, however, the increasing imaginary part of ω should cause the two values of $k_\pm$ to *coincide*, then the match at $z = 0$ can be accomplished without any source and this is a normal mode of the system, and if growing, then it is growing everywhere independent of the source.

These criteria for absolute and amplifying instabilities are for infinite systems and are not guaranteed for bounded systems.

2.7.2 Beam–Plasma Instability As a first example, we consider a 1-D problem with an electron beam passing through an unmagnetized plasma, so that the dispersion relation is given by

$$K_3 = 1 - \frac{\omega_p^2}{\omega^2} - \frac{\omega_{pb}^2}{(\omega - kv_0)^2} = 0. \tag{36}$$

The roots of this dispersion relation are

$$kv_0 = \omega(1 \pm \omega_{pb}/\sqrt{\omega^2 - \omega_p^2}), \tag{37}$$

and when $\omega < \omega_p$, the roots are complex and one lies above the real k axis and the other below. When ω has a large imaginary part, $|\omega| \to \infty$, and both roots are given by $k \sim (\omega \pm \omega_{pb})/v_0$, so that both roots are in the upper half plane. Since one had to cross, there is one amplifying wave. This is driven by the beam particles traveling close to the phase velocity which give up energy to the beam.

2.7.3 Two-Stream Instability Generalizing to two beams in opposite directions, the dispersion relation is

$$1 - \frac{\omega_p^2}{\omega^2} - \frac{\omega_{p1}^2}{(\omega - kv_1)^2} - \frac{\omega_{p2}^2}{(\omega + kv_2)^2} = 0. \tag{38}$$

if we let $\omega_{p1} = \omega_{p2} = \omega_0$ and $v_1 = v_2 = v_0$, some analytic results are possible. Defining $P = 1 - \omega_p^2/\omega^2$, the roots are

$$(kv_0)^2 = \omega^2 + \frac{\omega_0^2}{P} \pm \frac{\omega_0}{P}\sqrt{\omega_0^2 + 4P\omega^2}, \tag{39}$$

so there is a double root or saddle point where $\omega_0^2 = -4P\omega^2$ or $\omega^2 = \omega_p^2 - \omega_0^2/4$. For $\omega_p < \omega_0/2$, this leads to complex roots and an absolute instability. When we plot the locus of k^2 for pure imaginary ω in the complex k^2 plane, the saddle point lies on the real k^2 axis and branches off above and below the axis as in Fig. 6(a) for large ω_i and branches out along the real k^2 axis for small ω_i, ending with both roots at the origin. This means there are two symmetric saddle points, one with $k_r > 0$ and one with $k_r < 0$, and the branches for variable ω_i go either along the real k axis or above and below the axis, as shown in Fig. 6(b), which is the condition for the absolute instability.

3. WARM-PLASMA WAVES

When one goes beyond the cold-plasma model, effects of finite temperature may be inserted at several levels of approximation, from simply including a pressure term, or possibly by using higher moments of the Boltzmann equation, through kinetic-theory effects that may be nonrelativistic or relativistic. The first level of approximation is usually based either on one- or two-fluid models, which include MHD theory and indicate the direction that finite-temperature effects will take us from the cold-plasma model. The full kinetic-theory model, however, introduces both quantitatively and qualitatively different behavior, with a new class of waves and collisionless damping.

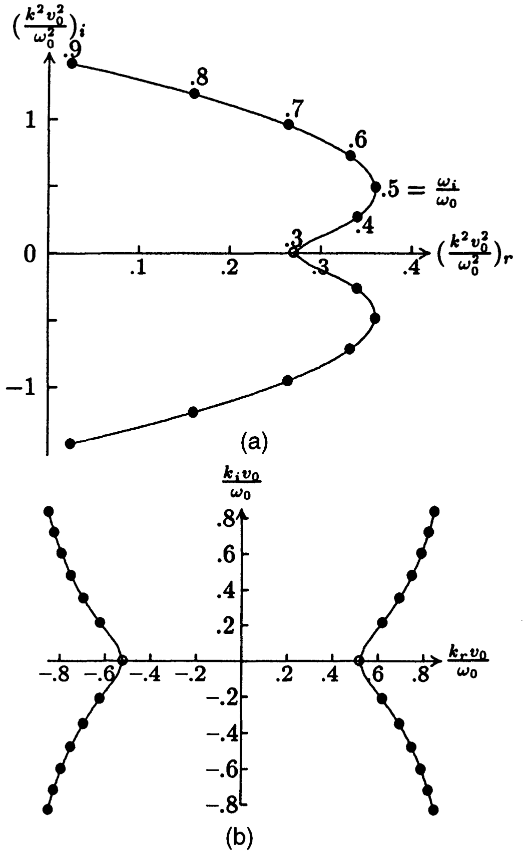

FIG. 6. Locus of (a) $k^2(\omega_i)$ in complex k^2 plane and (b) $k(\omega_i)$ in complex k plane (same points) from Eq. (39) for $\omega_p = 0.4\ \omega_0$.

3.1 Fluid Models

Fluid equations are based on moment expansions of the Boltzmann equation

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla f + \mathbf{a}\cdot\nabla_v f = \left.\frac{df}{dt}\right|_{\text{coll.}}, \tag{40}$$

where $\mathbf{a}$ is the acceleration due to electric and magnetic fields through the Lorentz force. If $\mathbf{u}$ is the mean velocity for a particular species, then the 0th moment of the Boltzmann equation, with the average density given by $n(\mathbf{r}) = \int f d^3v$, is

$$\frac{\partial n}{\partial t} + \nabla\cdot(n\mathbf{u}) = 0, \quad \text{Continuity equation.} \tag{41}$$

The corresponding first moment is

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \frac{\nabla\cdot\boldsymbol{\Psi}}{nm} - \frac{q}{m}(\mathbf{E} + \mathbf{u}\times\mathbf{B}) = 0, \tag{42}$$

where $\boldsymbol{\Psi}$ is the stress tensor. The evaluation of the stress tensor depends on even higher-moment equations, which introduce even higher-order terms. The method of solution is to truncate the hierarchy, in which case the simplest interpretation of $\boldsymbol{\Psi}$ is simply the pressure, which is obtained from an equation of state. Equation (42) represents the conservation of momentum while the continuity equation conserves particles.

3.1.1 Bohm–Gross Dispersion Relation

If we restrict ourselves to a one-dimensional problem, the higher moments are not so complicated; and if we examine the special case of plasma oscillations, where the magnetic field is not involved, then, assuming a Maxwellian distribution and truncating above the third moment, the dispersion relation for plasma oscillations is given by Eq. (2), which is called the Bohm–Gross dispersion relation (Bohm and Gross, 1949). Keeping terms through the fifth moment, the result is Eq. (3), so it is apparent that *the moment expansion is an expansion in the ratio of the thermal velocity to the phase velocity*. The Bohm–Gross dispersion relation resolves the ambiguity in cold-plasma theory where there was no phase or group velocity for plasma oscillations since k was not involved, and supports the notion that $P = 0$ in cold plasma is a cutoff rather than a resonance, since this dispersion relation describes a wave with a cutoff at ω_{pe} that propagates at the electron thermal speed for high frequencies.

3.1.2 Low-Frequency Dispersion Relation If we consider the linearized fluid equations for electrons and one species of ions along with the Maxwell equations, using the adiabatic equation of state so that the pressure $p_j = c_j^2\rho_j$ and c_j is the sound speed for species j given by $c_j^2 \equiv \gamma_j \kappa T_j/m_j$, then we may write the *low-frequency dispersion relation* (Stringer, 1963)

$$\left(1 - \frac{\omega^2}{k^2V_A^2} - \frac{\omega^2}{\omega_{ce}\omega_{ci}} + \frac{k^2c_s^2\sin^2\theta}{\omega^2 - k^2c_s^2}\right) \times \left(\cos^2\theta - \frac{\omega^2}{k^2V_A^2} - \frac{\omega^2}{\omega_{ce}\omega_{ci}}\right) = \frac{\omega^2\cos^2\theta}{\omega_{ci}^2}, \tag{43}$$

where θ is the angle between the direction of propagation and the static magnetic field and c_s is the acoustic speed given by $c_s^2 = (\gamma_e\kappa T_e + \gamma_i\kappa T_i)/(m_e + m_i)$.

3.1.3 MHD Region For the very low-frequency case, where $\omega \ll \omega_{ci}$, the right-hand side of the low-frequency dispersion relation may be neglected, along with $\omega^2/\omega_{ce}\omega_{ci}$, so that it reduces to the two separate factors,

$$1 - \frac{\omega^2}{k^2V_A^2} + \frac{k^2c_s^2\sin^2\theta}{\omega^2 - k^2c_s^2} = 0, \tag{44}$$

$$\cos^2\theta - \omega^2/k^2V_A^2 = 0. \tag{45}$$

Equation (45) gives the cold torsional-wave dispersion relation,

$$\omega^2 = k^2V_A^2\cos^2\theta, \tag{46}$$

which is independent of temperature. This is the torsional Alfvén wave, whose group velocity is parallel to the magnetic field.

Equation (44) leads to the quadratic

$$\omega^4 - k^2(V_A^2 + c_s^2)\omega^2 + k^4V_A^2c_s^2\cos^2\theta = 0, \tag{47}$$

with approximate roots (for $c_s \ll V_A$)

$$\omega^2 = k^2(V_A^2 + c_s^2\sin^2\theta) \tag{48}$$

and

$$\omega^2 = k^2c_s^2\cos^2\theta. \tag{49}$$

This last dispersion relation describes an acoustic wave whose phase velocity is equal to the acoustic speed and whose group velocity is parallel to the magnetic field but otherwise independent of the magnetic field.

The first root, from Eq. (48), is more interesting, since now the cold compressional wave is coupled to sound-wave terms. For this reason, this wave is often called the magnetoacoustic or magnetosonic wave since it is coupled to acoustic modes in the low-frequency limit. It is clear from the dispersion relation that the magnetoacoustic wave is equivalent to the cold-plasma compressional mode when $V_A \gg c_s$ and that the phase and group velocities are parallel in that limit; but when V_A is comparable to c_s, then the phase and group velocities are no longer parallel, since the acoustic-wave component of the dispersion relation tends to guide the wave away from the direction of the magnetic field. This same branch is thus partially guided *across* the field at very low frequencies if acoustic effects are important, isotropic if acoustic effects are negligible, and then at intermediate frequencies is guided *along* the field as the whistler wave, and carries a variety of names, depending on the approximation ($V_A \gg c_s$ or $V_A \ll c_s$), the frequency ($\omega \ll \omega_{ci}$ or $\omega_{ci} \ll \omega \ll \omega_{ce}$), and the angle ($R$ wave for $\theta = 0$, X wave for $\theta = \pi/2$ up to ω_{LH}).

In addition to the MHD region, there are numerous regions in parameter space that have distinctive characteristics, and the low-frequency dispersion relation couples these in a variety of transition regions. Unfortunately, most of these transitions occur in regions where the phase velocity is comparable to a thermal speed, and so the moment expansions upon which the analysis is based are unreliable in precisely those regions. This means that some solutions, and the transitions in particular, require a full kinetic treatment in order to obtain a proper description.

3.1.4 Collisions and Resistivity The effects of collisions on waves may be included relatively easily by adding a drag term in the momentum equation. For example, if Eq. (42) were for electrons (so that $\mathbf{u} \rightarrow \mathbf{v}_e$), then one would add a term $-\nu_{en}\mathbf{v}_e$ on the right for electron-neutral collisions. For the cold plasma, this kind of collision term modifies the dielectric tensor elements so that

$$\begin{aligned} K_1 &\equiv S = 1 - \sum_j \frac{\omega_{pj}^2(\omega + i\nu_j)}{\omega[(\omega + i\nu_j)^2 - \omega_{cj}^2]}, \\ iK_2 &\equiv D = \sum_j \frac{\epsilon_j \omega_{cj} \omega_{pj}^2}{\omega[(\omega + i\nu_j)^2 - \omega_{cj}^2)]}, \\ K_3 &\equiv P = 1 - \sum_j \frac{\omega_{pj}^2}{\omega(\omega + i\nu_j)}. \end{aligned} \tag{50}$$

For electron-ion collisions, the momentum gained by one species must be balanced by a corresponding loss for the other species, and this is generally incorporated through the introduction of plasma resistivity through a form of Ohm's law. The effective collision times from a one-fluid model with $\mathbf{u} = \mathbf{v}_e - \mathbf{v}_i$ are (Braginskii, 1965)

$$\begin{aligned} \tau_e &= \frac{3(2\pi)^{3/2}\sqrt{m_e}\,\epsilon_0^2(\kappa T_e)^{3/2}}{e^4 Z n_e \ln\Lambda} \\ &= \frac{3.4 \times 10^{11} T_e^{3/2}}{Z n_0 \ln\Lambda}, \end{aligned} \tag{51}$$

where the temperature is expressed in eV in the simplified expression with $n_e = n_0 = Zn_i$. The quantity Λ is approximately given by $\Lambda \simeq 9N_\mathrm{D}$ where N_D is the number of particles in a Debye sphere. The drag term is

$$\mathbf{R}u = -ne(\mathbf{j}_\parallel/\sigma_\parallel + \mathbf{j}_\perp/\sigma_\perp), \tag{52}$$

where the perpendicular and parallel conductivities are given by

$$\sigma_\perp = n_e e^2 \tau_e / m_e, \quad \sigma_\parallel = 1.96\sigma_\perp. \tag{53}$$

The inclusion of resistivity in the low-frequency dispersion relation, Eq. (43), modifies only the terms in $\omega^2/\omega_{ce}\omega_{ci}$, changing the first such term so that $\omega^2 \rightarrow \omega(\omega + i\nu_\perp)$ and the second so that $\omega^2 \rightarrow \omega(\omega + i\nu_\perp \cos^2\theta + i\nu_\parallel \sin^2\theta)$, where $\nu_\perp = ne^2/m_e\sigma_\perp$ and $\nu_\parallel = 0.51\nu_\perp$. The reason one ω is modified by collisions and the other not is that the first came from the Maxwell equations (not modified) and the other from the moment equations where collisions modify the time evolution of the distribution function.

3.1.5 Drift Waves Whenever there is a gradient in one of the equilibrium parameters, it generally leads to crossed electric and magnetic fields that allow the propagation of drift waves. The literature is extensive in the kinds of instabilities that may arise from these gradients (see Krall, 1968; Mikhailovskii, 1983). We will examine here only the basic principles and illustrate with an example. The literature is difficult to survey because there are examples of collisionless drift instabilities from a kinetic equation, there are collisional or dissipative drift instabilities from the fluid equations, and some use a mixture of both.

In order to see some of the nature of the problem without extensive analysis, we begin with a case of electrons and ions in the long-wavelength limit ($k^2\lambda_\mathrm{D}^2 \ll 1$ so that $n_e = n_i$) using the fluid equations and examine low-frequency electrostatic waves. If we neglect electron inertia ($m_e \rightarrow 0$) in the momentum equation, then we have the balance between the pressure gradient and the electric field as

$$e\frac{\partial\varphi}{\partial x} = \frac{\kappa T_e}{n_e}\frac{\partial n_e}{\partial x}, \tag{54}$$

with solution $n_e = \exp(e\varphi/\kappa T_e)$. The potential is assumed to vary as $\varphi \sim \exp(ik_y y - i\omega t)$, and the density as

$$n_i = n_e = n_0 \exp(e\varphi/\kappa T_e) \simeq n_0 + n_0 e\varphi/\kappa T_e. \tag{55}$$

The equation of motion for ions is

$$-i\omega m_i \mathbf{v}_i = e(\mathbf{E} + \mathbf{v}_i \times \mathbf{B}_0) \simeq 0, \tag{56}$$

since $\omega \ll \omega_{ci}$. Hence, $v_{ix} \simeq E_y/B_0 = -ik_y\varphi/B_0$. The ion continuity equation gives

$$-i\omega n_{1i} + v_{ix}\frac{dn_0}{dx} = 0, \tag{57}$$

and using the expressions for n_{1i} and v_{ix}, this may be solved for ω as

$$\omega = -\frac{k_y \kappa T_e}{eB_0}\frac{1}{n_0}\frac{dn_0}{dx} \equiv \omega_e^*. \tag{58}$$

Usually $n_0(x)$ decreases with increasing x, and so normally $\omega_e^* > 0$. The wave travels in the direction of $\nabla n_0 \times \mathbf{B}_0$ with phase velocity $c_s\rho_s/L_n$ where c_s is the ion-acoustic speed, $\rho_s = c_s/\omega_{ci}$ is like an ion Larmor radius, except that it depends on T_e rather than T_i, and L_n is the scale length of the density variation.

One way of thinking about the source of free energy for this instability is to realize that the density gradient tends to establish a radial electric field to balance the diffusion rates, and this radial field sets the plasma into rotation as a result of the $\mathbf{E} \times \mathbf{B}_0$ drift. In this case, a cylindrical plasma leads to waves rotating with an azimuthal mode number $k_y \simeq 2\pi m/r$, and so the waves may be localized about a radius where m is an integer. These drifting low-frequency waves are susceptible to a variety of instabilities.

3.1.6 Drift Resistive Instability For this example, we shall generalize somewhat the elementary analysis above, but still use the long-wavelength approximation and the electrostatic approximation, but with $\mathbf{E} = \mathbf{E}(x)\exp[i(k_y y + k_z z - \omega t)]$. This allows the wave to propagate parallel to the magnetic field in addition to its drift perpendicular to both the field and the gradient. Then using the fluid equation for electrons with collisions, and the electron continuity equation, the perturbed electron density is

$$\frac{n_e}{n_0} = \frac{e\varphi}{\kappa T_e}\frac{(1 - i\omega_e^* m_e \nu_{ei}/k_z^2\kappa T_e)}{(1 - i\omega m_e \nu_{ei}/k_z^2\kappa T_e)}. \tag{59}$$

For the ions, we assume $\omega \ll \omega_{ci}$ but keep first-order terms, so that along with the ion continuity equation, one finds

$$\frac{n_i}{n_0} = \frac{e\varphi}{\kappa T_e}\left(\frac{\omega_e^*}{\omega} + \frac{k_z^2 c_s^2}{\omega^2} - k_\perp^2\rho_s^2\right). \tag{60}$$

Equating Eq. (59) to Eq. (60) leads to the dispersion relation, which may be written as a quadratic,

$$\omega^2 + i\omega_s\left(1 + k_\perp^2\rho_s^2 - \frac{k_z^2 c_s^2}{\omega^2}\right)\omega - i\omega_e^*\omega_s = 0, \tag{61}$$

where

$$\omega_s = \frac{\omega_{ce}\omega_{ci}}{\nu_{ei}}\left(\frac{k_\perp^2}{k_z^2} - \frac{\omega_{ci}^2}{\omega^2}\right)^{-1}. \tag{62}$$

There is an unstable root whose maximum growth rate occurs when $\omega_e^* \simeq \omega_s(1 + k_\perp^2\rho_s^2 - k_z^2c_s^2/\omega^2)^2$, at which point

$$\omega = \frac{\omega_e^*(0.618 + 0.300i)}{1 + k_\perp^2\rho_s^2 - k_z^2c_s^2/\omega^2}. \tag{63}$$

This example exemplifies both the complexities and many approximations that are common to drift-type instabilities, and others based on temperature gradients, etc., are numerous.

3.2 Kinetic Models

Kinetic effects in plasma waves are included through the kinetic equation

$$\frac{\partial f_j}{\partial t} + \mathbf{v}\cdot\nabla f_j + \frac{q_j}{m_j}(\mathbf{E} + \mathbf{v} \times \mathbf{B})\cdot\nabla_v f_j = 0 \tag{64}$$

for each species, and the Maxwell equations, with

$$\rho = \sum_j q_j \int d^3v\, f_j, \tag{65}$$

$$\mathbf{J} = \sum_j q_j \int d^3v\, \mathbf{v} f_j \tag{66}$$

as the sources. This set, which neglects correlations, is commonly called the Vlasov equations. They form the basis for virtually all of kinetic theory except when correlations (or collisions) must be included.

3.2.1 Landau Damping The solution of the linearized set of equations without any magnetic field was first attempted using Fourier transforms in both time and space, resulting in a dispersion relation for plasma oscillations of the form

$$K_3 = 1 - \frac{\omega_{pi}^2}{k^2v_i^2}Z'\left(\frac{\omega}{kv_i}\right) - \frac{\omega_{pe}^2}{k^2v_e^2}Z'\left(\frac{\omega}{kv_e}\right), \tag{67}$$

where the prime denotes a derivative with respect to the argument, and the $v_j^2 = 2\kappa T_j/m_j$ are thermal velocities. The difficulty with this simple analysis was that in integrating over the velocity, it must be assumed that ω has a positive imaginary part in order to obtain the result above, while using the asymptotic expression for $Z(\zeta)$, one finds that ω has a negative imaginary part. Conversely, assuming a negative imaginary part to begin with, one finds a positive imaginary part, and so the method is not self-consistent. This difficulty was resolved by Landau who instead took Laplace transforms in time, assuming an initial perturbation at time 0+, and then the result above follows in a natural way, and the resulting damping is called Landau damping (Landau, 1946).

In the weak-damping limit ($v_p \gg v_e$), the dispersive solution for $\omega(k)$ gives the same result as the moment expansion, which to lowest nontrivial order gave the Bohm–Gross dispersion relation of Eq. (2). The corresponding Landau damping rate is given by

$$\omega_i = -\frac{\sqrt{\pi}\,\omega_{pe}^2\omega_r^2}{k^3 v_e^3}\exp\left(-\frac{\omega_r^2}{k^2 v_e^2}\right). \tag{68}$$

This damping is due to those particles in the distribution that travel near the phase velocity, and results in damping instead of growth because there are more traveling slightly slower than traveling faster, drawing energy from the wave as they gain or lose energy to stay with the wave.

3.2.2 Ion-Acoustic Wave In the example above, ion motions were neglected, since they are usually unimportant. In the cold-plasma limit this is because $m_e/m_i \ll 1$. Near the ion plasma frequency, however, it is possible to consider a case where the ions are cold ($v_i \ll v_p$), but the electrons are hot ($v_e \gg v_p$). When the electron and ion temperatures are comparable, there is generally no oscillation, since when one moves some of the ions to form a slight charge imbalance that would induce the ions to oscillate, the period is so slow that electrons, being so much more mobile, rush in to fill the charge imbalance before the ions can respond. When the electron temperature is much higher than the ion temperature, however, the electrons have too much momentum to slow down and fill in the potential depressions, and so some ion oscillations may occur.

Using the same dispersion relation for this case, but assuming $\zeta_i = \omega/kv_i \gg 1$, and $\zeta_e = \omega/kv_e \ll 1$, the dispersion relation becomes

$$1 + \frac{1 + i\sqrt{\pi}\zeta_e}{k^2\lambda_{\mathrm{D}}^2} = \frac{\omega_{pi}^2}{\omega^2}\left[1 - 2i\sqrt{\pi}\zeta_i^3\exp(-\zeta_i^2)\right]. \tag{69}$$

Setting the real and imaginary parts equal to zero separately, we obtain

$$\omega_r^2 = k^2c_s^2/(1 + k^2\lambda_{\mathrm{D}}^2) \tag{70}$$

and

$$\frac{\omega_i}{\omega_r} = -\sqrt{\pi}\zeta_i^3\left[\exp(-\zeta_i^2) + \frac{T_i v_i}{T_e v_e}\right], \tag{71}$$

where $\omega = \omega_r + i\omega_i$ and $c_s = \sqrt{\kappa T_e/m_i}$ is the ion-acoustic speed. We note that if $k\lambda_{\mathrm{D}} \ll 1$, then all waves will travel at the ion-acoustic speed. In order for the damping to be weak, there are constraints on both the electron and ion temperatures so that it requires $T_e \gg T_i$.

3.2.3 Hot-Plasma Dispersion Relation The derivation of the hot-plasma dispersion relation with a static magnetic field is nontrivial. Even if we restrict the results to isotropic Maxwellian distributions without any drifts, then all nine components of the dielectric tensor are nonzero, and if we again take $k_y = 0$, then the *changes* in the structure of the tensor relative to the cold-plasma dielectric tensor may be summarized by the relations

$$K_{yy} = K_1 + K_0, \tag{72}$$

$$K_{xz} = K_{zx} = K_4, \tag{73}$$

$$K_{yz} = -K_{zy} = -K_5, \tag{74}$$

where now $\zeta_{nj} = (\omega + n\omega_{cj})/k_z v_j$ in

$$K_0 = 2\sum_j \frac{\omega_{pj}^2 e^{-\lambda_j}}{\omega k_z v_j}\sum_{n=-\infty}^{\infty}\lambda_j(I_n - I_n')Z(\zeta_{nj}), \tag{75}$$

$$K_1 = 1 + \sum_j \frac{\omega_{pj}^2 e^{-\lambda_j}}{\omega k_z v_j}\sum_{n=-\infty}^{\infty}\frac{n^2 I_n}{\lambda_j}Z(\zeta_{nj}), \tag{76}$$

$$K_2 = i \sum_j \frac{\epsilon_j \omega_{pj}^2 e^{-\lambda_j}}{\omega k_z v_j} \sum_{n=-\infty}^{\infty} n(I_n - I_n')Z(\zeta_{nj}), \tag{77}$$

$$K_3 = 1 - \sum_j \frac{\omega_{pj}^2 e^{-\lambda_j}}{\omega k_z v_j} \sum_{n=-\infty}^{\infty} I_n \zeta_{nj} Z'(\zeta_{nj}), \tag{78}$$

$$K_4 = \sum_j \frac{k_\perp \omega_{pj}^2 e^{-\lambda_j}}{2k_z \omega \omega_{cj}} \sum_{n=-\infty}^{\infty} \frac{nI_n}{\lambda_j} Z'(\zeta_{nj}), \tag{79}$$

$$K_5 = i \sum_j \frac{k_\perp \epsilon_j \omega_{pj}^2 e^{-\lambda_j}}{2k_z \omega \omega_{cj}} \sum_{n=-\infty}^{\infty} (I_n - I_n')Z'(\zeta_{nj}), \tag{80}$$

where the I_n are modified Bessel functions of order n with argument $\lambda_j = \frac{1}{2}k_\perp^2 \rho_j^2$ where $\rho_j = v_j/\omega_{cj}$ is the Larmor radius, and the I_n' are derivatives with respect to the argument. Generally, $k_\perp^2 = k_x^2 + k_y^2$, but here we have taken $k_y = 0$, so that $k_\perp = k_x$.

The appearance of the parameter λ_j, which is effectively the square of the ratio of the Larmor radius to the perpendicular wavelength, adds an important new feature to the hot-plasma effects. When the wavelength is long (λ_j small), a gyrating particle sees no overall effect near integral multiples of the cyclotron frequency. When the lengths are comparable, however, the integrated effect about the orbit is not small, and significant dispersion is the result. There are thus two distinct kinds of parameters that affect the deviations from cold plasma, namely, the ratio v_j/v_p and λ_j.

Using these components, the hot-plasma dispersion relation may be written as

$$\begin{aligned}&[\gamma(\gamma - \kappa_0 + k_\perp^2) + \kappa_2^2]\kappa_3 \\ &\quad + k_\perp^2[(\gamma - \kappa_0 + k_\perp^2)\kappa_1 - \kappa_2^2] \\ &\quad = \kappa_4(\gamma - \kappa_0 + k_\perp^2)(2k_\perp k_z + \kappa_4) \\ &\quad - \kappa_5[\gamma\kappa_5 + 2\kappa_2(k_\perp k_z + \kappa_4)] = 0,\end{aligned} \tag{81}$$

where $\kappa_j \equiv (\omega^2/c^2)K_j$ and $\gamma = k_z^2 - \kappa_1$. Although the dielectric tensor elements are more complicated with $k_y \neq 0$, the dispersion relation is the same as long as $k_\perp$ is used. This dispersion relation and the more complicated dielectric tensor components introduce two new distinct effects: cyclotron damping and finite–Larmor-radius effects. The first is analogous to Landau damping, only one must be in a rotating reference frame with angular frequency $n\omega_{cj}$ to observe the similarity. The effects of finite Larmor radius are exhibited through $\lambda_j \neq 0$, since this parameter is the square of the ratio of the Larmor radius to the perpendicular wavelength.

3.2.4 Cyclotron Damping For parallel propagation, $k_\perp = \lambda_j = 0$, and the tensor elements simplify since the infinite sums reduce to either one or two terms. In addition, we find $K_0 = K_4 = K_5 = 0$, and so the dispersion relation reduces to

$$(\gamma^2 + \kappa_2^2)\kappa_3 = 0, \tag{82}$$

whose roots are $\kappa_3 = 0$ and $k_z^2 = \kappa_1 \pm i\kappa_2$. The first root is the plasma wave, which is unaffected by the magnetic field. The other two roots are the R wave and the L wave, whose dispersion relations reduce to

$$\begin{aligned}n_{R,L}^2 = 1 + \sum_j \frac{\omega_{pj}^2}{\omega k_z v_{lj}} \Bigg[&\left(\frac{1 \pm \epsilon_j}{2}\right)Z(\zeta_{1j}) \\ &+ \left(\frac{1 \mp \epsilon_j}{2}\right)Z(\zeta_{-1j})\Bigg].\end{aligned} \tag{83}$$

The structure of this dispersion relation confirms the general character we had observed with the cold-plasma waves except that now resonance has a different meaning. The R wave is here seen to be a function of $Z(\zeta_{1i})$ and $Z(\zeta_{-1e})$, and assuming that $\omega \gg k_z v_{lj}$, then there is virtually no ion damping associated with the R wave since $|\zeta_{1i}| \gg 1$ and the ion damping is exponentially small. Near the electron cyclotron resonance, however, $|\zeta_{-1e}| \simeq 0$, and in this limit $Z(\zeta_{-1e}) \simeq i\sqrt{\pi}$ so that there is no longer any resonance at the electron cyclotron frequency, but now there is strong damping. A little away from the resonance, assuming weak damping, $\omega_i \ll \omega_r$, the damping is approximately given by

$$\frac{\omega_i}{\omega_r} \sim -\frac{\sqrt{\pi}\omega_{pe}^2 \exp(-\zeta_{-1e}^2)}{\omega_r k_z v_e[2 + \omega_{pe}^2 \omega_{ce}/\omega_r(\omega_r - \omega_{ce})^2]}, \tag{84}$$

which indicates that the damping is exponentially small far from resonance, but as resonance approaches, the exponential term grows. The denominator also grows, suggesting a maximum damping rate, but this approximation fails near the resonance. For $|\zeta_{-1e}| \gg 1$, the damping is weak, and so the cold-plasma dispersion relation is relevant except near resonance.

3.2.5 Electrostatic Waves The effects of finite Larmor radius are very apparent for hot electrostatic waves. In the electrostatic limit, the dispersion relation reduces greatly in complexity, resulting in

$$k_\perp^2 K_1 + 2k_\perp k_z K_4 + k_z^2 K_3 = 0. \qquad (85)$$

Combining the terms in the expressions for K_1, K_3, and K_4, one obtains the hot-plasma electrostatic dispersion relation

$$1 + \sum_j \frac{e^{-\lambda_j}}{k^2\lambda_{Dj}^2} \sum_{n=-\infty}^{\infty} [1 + \zeta_{0j} Z(\zeta_{nj})] I_n(\lambda_j) = 0. \qquad (86)$$

An important special case is when $k_z \to 0$, in which case there is no damping at all and the dispersion relation reduces to

$$k_\perp^2 = \sum_j 2k_{Dj}^2 e^{-\lambda_j} \sum_{n=1}^{\infty} I_n(\lambda_j) \frac{n^2}{\nu_j^2 - n^2}, \qquad (87)$$

where $k_{Dj} = 1/\lambda_{Dj}$ is the Debye wave number and $\nu_j = \omega/\omega_{cj}$. This dispersion relation describes what are known as Bernstein waves and has a resonance at every harmonic of both ion and electron cyclotron frequencies, but the "strength" of the resonance indicated by $I_n(\lambda_j)$ becomes small for large n. We note that there is no absorption to damp out the wave at resonance. Within the framework of the collisionless theory outlined in this section, these resonances remain unresolved.

Examining the behavior for small λ_i (which means $\lambda_e \ll 1$ since $\lambda_e/\lambda_i = T_e m_e/T_i m_i \ll 1$), we approximate $I_n(\lambda) \simeq (\lambda/2)^n/n!$ and consider cold electrons. In this case, the dispersion relation can be approximated by

$$\nu = n - \frac{(n^2 - 1)\lambda_i^{n-1}}{2^n(n-1)![1 - (n^2 - 1)V_A^2/c^2]}, \qquad (88)$$

which indicates that Bernstein waves propagate just below each harmonic while the slope of the dispersion relation near the origin for λ_i is linear for $n = 2$, quadratic for $n = 3$, etc. This behavior changes above the lower hybrid resonance, where the wave propagates just above each harmonic. This behavior is illustrated in Fig. 7(a) below the lower hybrid resonance and in Fig. 7(b) above the resonance.

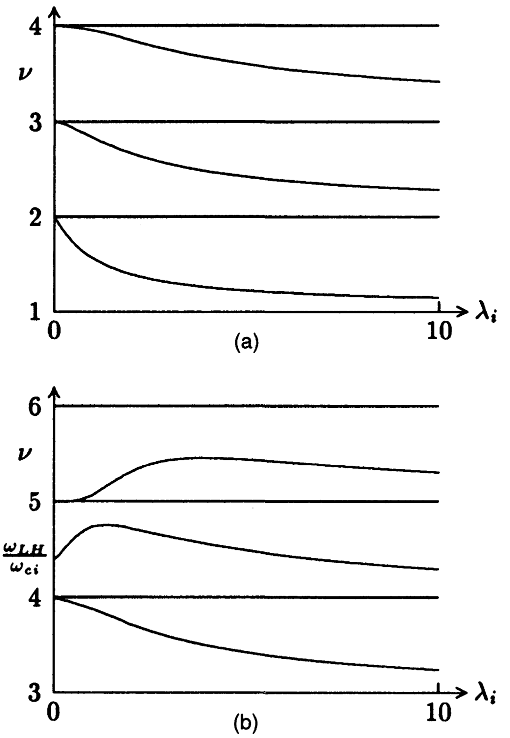

FIG. 7. Ion–Bernstein-wave dispersion relations. (a) High density ($\omega_{pe}^2/\omega_{ce}^2 = 10$). (b) Low density near ω_{LH} ($\omega_{pe}/\omega_{ce} = 0.1$).

There are additional warm-plasma effects when the plasma is relativistic, even if only weakly so. These effects are apparent when one examines the quantity $\zeta_n = (\omega - n\omega_c/\gamma)/k_\parallel v_j$ where γ is the relativistic factor. This argument appears as a constant in the nonrelativistic case, but when $\gamma = 1$ and $k_\parallel \to 0$, $\zeta_n \to \infty$ for any $\omega \neq n\omega_c$, but $\zeta_n \to 0/0$ when $\omega = n\omega_c$. With only weakly relativistic corrections, however, the numerator never vanishes, and there is collisionless damping even for $k_\parallel = 0$. In the weak relativistic limit, where $\mu = mc^2/\kappa T \gg 1$, the dominant dielectric tensor elements with $k_\parallel = 0$ are

$$K_1 = 1 - \frac{\omega_p^2}{\omega^2}\mu \sum_{n=1}^{\infty} \frac{n^2 I_n}{\lambda} \times [F_{n+3/2}(z_+) + F_{n+3/2}(z_-)], \qquad (89)$$

$$iK_2 = \frac{\omega_p^2}{\omega^2}\mu \sum_{n=1}^{\infty} n[I_n'[F_{n+3/2}(z_+) - F_{n+3/2}(z_-)] - I_n[F_{n+5/2}(z_+) - F_{n+5/2}(z_-)]], \qquad (90)$$

$$K_3 = 1 - \frac{\omega_p^2}{\omega^2}\mu\Bigl\{I_0F_{5/2}(\mu) + \sum_{n=1}^{\infty} I_n \times [F_{n+5/2}(z_+) + F_{n+5/2}(z_-)]\Bigr\}, \qquad (91)$$

where $z_\pm = \mu(\omega \pm n\omega_c)/\omega$. The weakly relativistic plasma dispersion function is related to the usual plasma dispersion function through the relation $i\sqrt{z}F_{1/2}(z) = Z(i\sqrt{z})$. The higher-order functions may be obtained from the recursion formula

$$(q - 1)F_q(z) = 1 - zF_{q-1}(z). \qquad (92)$$

From the tensor components above, the X-wave dispersion relation, written as

$$n^2 = 2RL/(R + L), \qquad (93)$$

may be written to first order in λ from the relations

$$R, L = 1 - \frac{\omega_p^2}{\omega^2}\mu[F_{5/2}(z_{\mp 1}) - \lambda F_{7/2}(z_{\mp 1}) + \lambda F_{7/2}(z_{\mp 2})]. \qquad (94)$$

For $z < 0$ (and small), the extraordinary wave will be damped, and this occurs for ω just below the fundamental or the harmonic. The maximum damping typically occurs for $z_n \simeq -\frac{7}{2}$, so that even with κT much less than 1% of the electron rest energy ($\mu > 1000$), nontrivial absorption can occur within a few percent of the cyclotron resonance or its harmonic. It is to be noted that this is with $n_\parallel = 0$ where the nonrelativistic theory indicated no damping at all.

4. PLASMA-FILLED WAVEGUIDE

4.1 Boundary Conditions

The usual boundary conditions encountered are either a conducting wall or a boundary between the plasma and either vacuum or a dielectric (such as glass). These boundary conditions are generally taken as abrupt, and we use jump conditions to relate the fields on opposite sides of the discontinuity. This is never precise, since real plasmas do not have sharp boundaries. In general, as the plasma approaches either a conductor or a dielectric, the density either falls gradually, dominated by a variety of transport phenomena such as electron and ion diffusion, ionization, and recombination, or it falls off rather abruptly, so that a large change occurs in a few Debye lengths. We restrict our attention primarily to waves whose wavelength is long compared to the thickness of the boundary layer (abrupt boundary). The simplest case is the conducting boundary because the condition in a cold plasma is that the tangential electric field must vanish at a perfect conductor. This may be written as

$$\mathbf{E}_t(a) = 0, \qquad (95)$$

where the conducting wall is at radius a (laboratory plasmas are more likely to be cylindrical than rectangular). The boundary conditions used at a plasma–vacuum interface, or at an interface between a plasma and an isotropic, insulating ideal dielectric (such as glass or vacuum), are the usual ideal-dielectric boundary conditions, and with $\hat{\mathbf{e}}_n$ as a unit vector normal to the surface, pointing out from the plasma, are

$$\hat{\mathbf{e}}_n \times (\mathbf{E}_v - \mathbf{E}_p) = 0, \qquad (96)$$

$$\hat{\mathbf{e}}_n \cdot (\mathbf{D}_v - \mathbf{D}_p) = 0, \qquad (97)$$

where we denote either vacuum or an isotropic dielectric by the subscript v. This is not a trivial boundary condition, since for a single wave in the plasma (incident plus reflected, or standing wave normal to the surface, characterized by a single value of $k_\perp$), the conditions Eqs. (96) and (97) imply that in general the vacuum transverse electric and transverse magnetic modes are coupled. The second wave is characterized by the same wave number parallel to the surface but distinct in $k_\perp$. The second value, which we shall denote $k_{\perp 2}$, is derived from the cold-plasma dispersion relation, which is quadratic in $k_\perp^2$, so that choosing $k_{\perp 1}$ or k_z determines the value of $k_{\perp 2}$. If $k_{\perp 1}^2 > 0$ (standing wave radially), the second root is typically evanescent radially, and so it is localized near the surface, in which case a surface wave is required to satisfy boundary conditions.

For electrostatic waves, we simply take the potential to be continuous at the boundary, along with Eq. (97). We note that in calculating the normal component of the displacement vector, the dielectric tensor elements are the ones given previously *only if the magnetic field is parallel to the interface and oriented in the z direction.* When the magnetic-field direction is oblique to the interface, then the usual dielectric tensor must be rotated, but this is rarely necessary, since plasma–vacuum interfaces are typically magnetic surfaces.

4.2 Unmagnetized Plasmas

When an electromagnetic wave scatters from a plasma column, it will generally excite plasma waves in the column, and whenever those plasma waves fit the plasma column so as to represent standing waves, the scattering amplitude exhibits plasma column resonances, called the *Tonks–Dattner resonances*, which were shown to have a dipole character (dominantly linear rather than radial). In addition to verifying the general representation of a plasma by a dielectric constant, the rather widely separated series of resonances also led to the verification of the Bohm–Gross dispersion relation of Eq. (2), which included the lowest-order thermal corrections to the cold unmagnetized plasma.

The plasma wave may be described by the electrostatic approximation, $\mathbf{E} = -\nabla\varphi$, where the potential satisfies Poisson's equation,

$$\nabla \cdot \epsilon \nabla \varphi = -\rho_{\text{ext}}. \tag{98}$$

Here, $\epsilon = \epsilon_0 K_3$ is the uniform plasma dielectric constant and the plasma is assumed to have no *external* charge, and so φ must satisfy Laplace's equation inside and outside the plasma column. Assuming no z dependence, the potentials inside and outside the plasma are

$$\varphi_i = \sum_{m=1}^{\infty} A_m r^m \sin m\theta, \tag{99}$$

$$\varphi_o = \sum_{m=1}^{\infty} B_m r^{-m} \sin m\theta + E_{in} r \sin\theta, \tag{100}$$

where E_{in} is the incident wave field amplitude. The values of A_m and B_m must be obtained by matching boundary conditions at the plasma surface at $r = a$,

$$K_3 \left.\frac{\partial \varphi_i}{\partial r}\right|_a = \left.\frac{\partial \varphi_o}{\partial r}\right|_a, \tag{101}$$

$$\varphi_i(a) = \varphi_o(a). \tag{102}$$

Using the orthogonality relations, these lead to the relations $A_1 = 2E_{in}/(1 + K_3)$ and $(1 + K_3)A_m = 0,\ m \geq 2$. The dominant mode ($A_1$) becomes large (resonant) as $1 + K_3 \to 0$ so that the resonant frequency is given by

$$\omega = \omega_p/\sqrt{2}. \tag{103}$$

The fact that the resonance occurs at $\omega_p/\sqrt{2}$ rather than at the plasma frequency is a consequence of the finite geometry and the cylindrical coordinate system. For a spherical plasma, the resonance occurs at $\omega = \omega_p/\sqrt{3}$.

The existence of such a resonance in the column leads to several conclusions. First, near the resonance, the amplitude of the electric field in the plasma will be large and hence the motions of the plasma particles will be correspondingly large. Secondly, these moving charges will radiate, or scatter the incident radiation. Finally, this scattering of radiation will lead to reflection, and so the reflection peak will correspond to the column resonance. It is generally the peak in the reflected power that is measured in experiments.

While this simple analysis was sufficient to describe the main resonance, which has the largest reflection coefficient, there appeared in the experiments a series of resonances that remained unexplained for many years. The multiplicity of resonances was shown to be due to thermal effects, and the wide spacing was shown to be due to the density variation across the plasma column (Parker *et al.*, 1964), so that an abrupt boundary model was inappropriate for this case.

One can also investigate electromagnetic waves in an unmagnetized plasma column where the waves are generally surface waves that propagate *along* the column. For a case where a plasma column of radius a is surrounded by vacuum, the dispersion relation is given by

$$1 - \frac{\omega_p^2}{\omega^2} = \frac{\tau a}{\tau_0 a} \frac{K_0'(\tau_0 a)}{K_0(\tau_0 a)} \frac{I_0(\tau a)}{I_0'(\tau a)}, \tag{104}$$

where $\tau_0^2 = k_z^2 - \omega^2/c^2$, $\tau^2 = \tau_0^2 + \omega_p^2/c^2$, and the I_0 and K_0 are the regular and irregular modified Bessel functions of order zero. In the limit as $k_z \rightarrow 0$, $\omega \rightarrow \omega_p/\sqrt{2}$, which is the Tonks–Dattner limit. In the high-density limit, the waves are surface waves and are called Trivelpiece–Gould modes (Trivelpiece and Gould, 1959).

4.3 Magnetized Plasmas

4.3.1 Wave Fields When one considers a cold magnetized plasma in a waveguide, the Maxwell wave equation is quadratic in $\nabla_\perp^2$ and may be written in the form

$$(\nabla_\perp^2 + k_{\perp 1}^2)(\nabla_\perp^2 + k_{\perp 2}^2)E_z = 0, \tag{105}$$

where the two values of $k_\perp^2$ are the quadratic pair of roots of the cold-plasma dispersion relation. Since B_z satisfies the same equation, $B_z \propto E_z$. The longitudinal field components may be written as

$$\begin{aligned} B_{zm} &= A_m[J_m(k_{\perp 1}r) + \tau_m J_m(k_{\perp 2}r)], \\ E_{zm} &= \frac{\omega\kappa_1 A_m}{k_z\kappa_2\kappa_3}[\beta_1 J_m(k_{\perp 1}r) + \tau_m\beta_2 J_m(k_{\perp 2}r)], \end{aligned} \tag{106}$$

where the common factor $\exp[i(k_z z + m\theta - \omega t)]$ is suppressed, and the other tangential electric-field component is

$$\begin{aligned} \frac{E_{\theta m}}{A_m} = &-\frac{m\omega}{r\kappa_2}\left[\delta_1 \frac{J_m(k_{\perp 1}r)}{k_{\perp 1}^2} + \tau_m\delta_2 \frac{J_m(k_{\perp 2}r)}{k_{\perp 2}^2}\right] \\ &- i\omega\left[\frac{J_m'(k_{\perp 1}r)}{k_{\perp 1}} + \tau_m \frac{J_m'(k_{\perp 2}r)}{k_{\perp 2}}\right], \end{aligned} \tag{107}$$

where $\beta_j = \gamma + k_{\perp j}^2 - \kappa_2^2/\kappa_1$ and $\delta_j = \gamma + k_{\perp j}^2$ for $j = 1, 2$ and $\gamma = k_z^2 - \kappa_1$.

The boundary conditions in this case are very simple when the plasma completely fills a conducting waveguide, since $\mathbf{E}_t(a) = 0$. This leads to the result

$$\begin{aligned} &\frac{im\delta_1}{k_{\perp 1}^2 a\kappa_2\beta_1} - \frac{1}{k_{\perp 1}\beta_1}\frac{J_m'(k_{\perp 1}a)}{J_m(k_{\perp 1}a)} \\ &\quad = \frac{im\delta_2}{k_{\perp 2}^2 a\kappa_2\beta_2} - \frac{1}{k_{\perp 2}\beta_2}\frac{J_m'(k_{\perp 2}a)}{J_m(k_{\perp 2}a)} \end{aligned} \tag{108}$$

4.3.2 MHD Approximation For frequencies well below the ion cyclotron frequency, κ_1 and κ_2 are approximately related as $\kappa_2 \simeq (\omega/\omega_{ci})\kappa_1 \rightarrow 0$ as $\omega/\omega_{ci} \rightarrow 0$, while $\kappa_1 \rightarrow \omega^2/V_A^2$, and so we can neglect κ_2. In this case, the dispersion relation [Eq. (81) with $\kappa_0 = \kappa_4 = \kappa_5 = 0$ since cold] can be factored into

$$(\gamma + k_\perp^2)(\gamma + k_\perp^2\kappa_1/\kappa_3). \tag{109}$$

Because the dispersion relation factors, the wave equation separates into two separate second-order equations, so that

$$d\nabla_\perp^2 B_z - i\omega B_z = 0, \tag{110}$$

$$q\nabla_\perp^2 E_z + \frac{i\kappa_3}{\omega}E_z = 0, \tag{111}$$

where $d = i\omega\gamma/(\gamma^2 + \kappa_2^2) \rightarrow i\omega/\gamma$, and $q = (\kappa_1\gamma - \kappa_2^2)/i\omega(\gamma^2 + \kappa_2^2) \rightarrow \kappa_1/i\omega\gamma$. The transverse-wave number from Eq. (110) is $k_\perp^2 = -\gamma$ so that $\gamma + k_\perp^2 = 0$ is the dispersion relation for this case. The waves are compressional and transverse electric modes, since there is no E_z component. The boundary condition is

$$E_\theta(a) \propto J_m'(k_\perp a) = 0, \tag{112}$$

so $k_\perp a = p_{m\nu}'$ where $p_{m\nu}$ and $p_{m\nu}'$ are the νth zeroes of the Bessel function of order m and its first derivative, respectively. The dispersion relation can then be written in the form

$$(k_z a)^2 = (\omega a/V_A)^2 - p_{m\nu}'^2, \tag{113}$$

where the approximation $K_1 \simeq \omega_{pi}^2/\omega_{ci}^2 = c^2/V_A^2$ has been used. This dispersion relation is identical to that of electromagnetic waves in an empty cylindrical waveguide except that the effective dielectric constant is $K_{\text{eff}} = c^2/V_A^2$. The cutoff frequencies are given by

$$\omega_{co} = k_\perp V_A = p_{m\nu}' V_A/a. \tag{114}$$

For circularly symmetric modes, $J_0'(x) = -J_1(x)$, and so $p_{0\nu}' = p_{1\nu} = 3.83, 7.01, 10.17, \ldots$, etc. For the lowest cutoff frequency, however, one must choose $m = \pm 1$, since the lowest root of $J_1'(x) = 0$ is $p_{11}' = 1.84$.

The other solution corresponds to Eq. (111), and so the waves are torsional and transverse magnetic modes (no B_z component) and the boundary condition is

$$E_z(a) \propto J_m(k_\perp a) = 0; \tag{115}$$

so $k_\perp a = p_{m\nu}$, and the dispersion relation is

$$k_z^2 \simeq \frac{\omega^2}{V_A^2}\left[1 + \left(\frac{p_{m\nu}c}{\omega_{pe}a}\right)^2\right] \tag{116}$$

if $\omega_{pi} \gg \omega_{ci}$. It should be noted that Eq. (113) has no cutoff, so that transverse magnetic modes propagate down to zero frequency. These waves are the slow Alfvén waves that exhibit the ion cyclotron resonance. Their absence of any waveguide cutoff is due to their torsional character and the fact that the group velocity is parallel to the waveguide, minimizing any wall interactions. It should be noted that near ω_{ci}, $k_\perp a \to p'_{m\nu}$, and so the MHD approximation fails in that range.

4.3.3 Loop Antenna (MHD) In the MHD case for the compressional mode, the response from a loop antenna in a waveguide is calculated by considering the jump in the magnetic field due to a current in the loop, which we take to be of radius b. This jump condition is

$$[B_r] = \mu_0 I \delta(r - b). \tag{117}$$

The dispersion relation from Eq. (113) may be written

$$k_{\parallel\nu}^2 = \omega^2/V_A^2 - k_{\perp\nu}^2. \tag{118}$$

The principal fields are B_r, B_z, and E_θ. In particular,

$$B_{r\nu} = -(ik_{\parallel\nu}/k_{\perp\nu})A_\nu J_1(k_{\perp\nu}r). \tag{119}$$

Then the jump condition gives

$$\begin{aligned}[B_r] &= -2\sum_\nu \frac{ik_{\parallel\nu}A_\nu}{k_{\perp\nu}} J_1(k_{\perp\nu}r) \\ &= \mu_0 I\delta(r - b).\end{aligned} \tag{120}$$

Using orthogonality relations for $J_1(k_{\perp\nu}r)$, the excitation coefficient is given by

$$A_\mu = \frac{i\mu_0 Ibk_{\perp\mu}}{k_{z\mu}a^2}\frac{J_1(k_{\perp\mu}b)}{J_0^2(k_{\perp\mu}a)}. \tag{121}$$

It is possible to adjust b to eliminate one radial mode, but in general there are an infinite number of modes.

The time dependence of the response for an impulse is obtained by an inverse Fourier transform in ω, with the result

$$B_z(r,z,t) = \frac{\mu_0 Ib}{a^2}\sum_\nu \frac{k_{\perp\nu}J_1(k_{\perp\nu}b)}{J_0^2(k_{\perp\nu}a)} J_0(k_{\perp\nu}r) \times \begin{cases} J_0(\omega_\nu\sqrt{t^2 - z^2/V_A^2}), & t > z/V_A, \\ 0, & t < z/V_A \end{cases}. \tag{122}$$

Each radial mode has similar time dependence, and the total B_z is the sum of such terms. We can understand the response by noting first that $v_g \leq V_A$ always, so that nothing can appear before the Alfvén transit time, $t = z/V_A$. Then the high frequencies appear first and give the sharp initial rise as illustrated in Fig. 8, and then, since $v_g \to 0$ as the frequency nears the cutoff frequency, the long-time behavior exhibits an oscillation that approaches the cutoff frequency since for large arguments, $J_0(\omega_\nu t) \sim \exp(i\omega_\nu t + i\phi)/\sqrt{\omega_\nu t}$. For realistic waveguides where there is some loss going down the guide, the higher modes damp more quickly and round off the sharp rise, and the long-time behavior is dominated by the cutoff frequency of the lowest mode. At higher frequencies, the group velocity of the higher-frequency components increases, and so the impulse response begins with a high-frequency oscillation, the maximum of the envelope arrives at the Alfvén transit time, and the final oscillation is the lowest cutoff frequency.

4.3.4 Electrostatic Case The electrostatic approximation can be obtained either by assuming a narrow waveguide (so that $k_\perp$ is very large) or by assuming that the wave

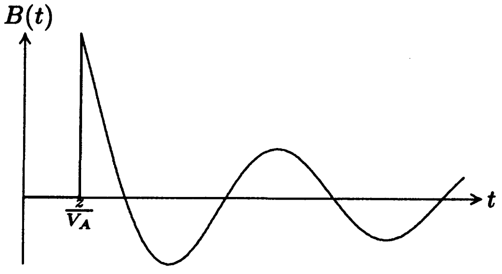

FIG. 8. Impulse response for a single radial mode for the compressional Alfvén wave.

is slow, but either case greatly simplifies the analysis. Assuming charge neutrality in a cold plasma, Poisson's equation becomes

$$\nabla \cdot \epsilon_0 \mathbf{K} \cdot \nabla \varphi_1 = 0, \tag{123}$$

with the cold magnetized-plasma dielectric tensor. Assuming a wavelike potential, i.e., $\varphi_1 = \varphi_1(\mathbf{r}) \exp(ik_z z - i\omega t)$, then Poisson's equation becomes

$$(K_1 \nabla_\perp^2 - K_3 k_z^2)\varphi_1 = 0, \tag{124}$$

where $\nabla_\perp^2 = \nabla^2 - \partial^2/\partial_z^2$. Solutions in cylindrical coordinates are

$$\varphi_1 = \sum_{m=0}^{\infty} A_m J_m(k_\perp r) \exp(ik_z z + im\theta - i\omega t), \tag{125}$$

with $k_\perp^2 = -k_z^2 K_3/K_1$. For a conducting waveguide, we require the tangential field to vanish at the wall at $r = a$; so we must have $J_m(k_\perp a) = 0$ and so $k_\perp a = p_{m\nu}$.

The electrostatic approximation is valid near the electron plasma frequency and the electron cyclotron frequency for sufficiently small a, and so we will neglect the ion motion in K_1 and K_3. The dispersion relation for the space charge waves is then given by

$$\left(\frac{p_{m\nu}}{k_z a}\right)^2 = -\frac{1 - \omega_{pe}^2/\omega^2}{1 + \omega_{pe}^2/(\omega_{ce}^2 - \omega^2)}. \tag{126}$$

The propagation is in two frequency bands, the upper one beginning at the upper hybrid resonance and propagating down to the greater of ω_{ce} or ω_{pe}, while the lower branch propagates from low frequency up to the lesser of ω_{ce} or ω_{pe}. As $B_0 \to 0$, both regions collapse, and no propagation is possible. With an additional vacuum or dielectric layer, however, propagation is again possible as surface waves (Trivelpiece and Gould, 1959).

5. EMISSION AND ABSORPTION

In order to estimate the absorption for a wave in a uniform plasma, one generally needs only to evaluate the imaginary part of k to find the damping rate. Near resonances, however, it is not this simple, since even weak inhomogeneities require a more detailed treatment that considers the possibility of tunneling through a finite resonance layer along with the possibilities of absorption and mode conversion. Mode conversion is the phenomenon of wave energy converting through a linear process from one wave type to another (such as a cold electromagnetic wave to a warm electrostatic wave). After all of the scattering coefficients, which quantify the transmission, reflection, and conversion coefficients, are determined, then the remainder is absorbed. The evaluation of these coefficients is described by mode-conversion theory, which also gives detailed solutions for the fields in the resonance layer as well as the power deposition profiles and the emission along each branch. These problems generally divide into two main classes, one where there is only a simple resonance (no nearby cutoff) with no transmission and one where there is a back-to-back cutoff and resonance with tunneling, where absorption effects may be either weak or strong.

5.1 Simple Resonance

An example of a simple resonance is the lower hybrid resonance, where the X mode approaches the resonance in an inhomogeneous plasma. In this case, the incident wave is totally mode converted from a cold wave to a warm electrostatic wave, and the eventual fate of the energy carried by the wave must be determined by following the wave (generally by ray tracing) until it is damped by electron Landau damping or high-harmonic ion cyclotron damping. This was demonstrated by Stix (1965) by converting the local dispersion relation near the resonance into a differential equation whose solution established the total conversion of wave energy.

5.2 Tunneling Layers

Tunneling layers are not so simple, but are more common, as they occur at every harmonic of the electron and ion cyclotron frequencies and at the two-ion hybrid resonance. This problem was first analyzed by Budden (1955) who treated a wave equation with a cutoff and resonance of the form

$$y'' + k_0^2(1 + x_0/x)y = 0, \tag{127}$$

which has a pole at the origin and a zero at $x = -x_0$, and k_0 is the asymptotic value of k. He showed that the amplitude transmission coefficient is given by $T = e^{-\eta}$, where $\eta = \frac{1}{2}\pi k_0 x_0$, and is the same from either side. He also demonstrated that the reflection coefficient $R = 0$ from the positive-x side, and $R = 1 - T^2$ from the opposite side. In neither case is energy conserved.

The more nearly complete solution includes warm-plasma effects, where cold-plasma waves typically couple to Bernstein modes, and the resonance is resolved. The form of the warm-plasma dispersion relation near $\omega \simeq 2\omega_{ci}$ is shown in Fig. 9, where three wave branches are identified and the tunneling layer where k^2 has an imaginary part is apparent. This dispersion relation may be cast into a differential equation of the form

$$y^{iv} + \lambda^2 z y'' + (\lambda^2 z + \gamma)y = 0, \qquad (128)$$

where $\lambda^2 > 0$ is a constant, the coordinate z is normalized so that the fast wave varies as $e^{\pm iz}$, and γ is a constant. Any fourth-order differential equation derived from a dispersion relation in k^2 with constant and linear coefficients can be cast into this form if the highest coefficient is constant. For the ion case, $\gamma > -1$ and the scattering coefficients are given in Table 1 with the upper signs, while for the electron case, $\gamma < -1$ and the dispersion relation is shown in Fig. 10, and the lower signs in the table are used. For all cases, the tunneling factor is $\eta = \pi|1 + \gamma|/2\lambda^2$ and $\epsilon \equiv 1 - e^{-2\eta}$. There is reciprocity for power between each branch, which means the fraction of incident power from any branch to any other branch is the same in either direction. The general analysis and many special results may be found in a review by Swanson (1985).

Table 1. Amplitude-scattering parameters for three branch problems without absorption.

Transmission	Reflection	Conversion
$T_1 = e^{-\eta}$	$R_1 = 0$	$C_{13} = \mp\epsilon$
$T_2 = e^{-\eta}$	$R_2 = -\epsilon$	$C_{23} = \mp e^{-\eta}\epsilon$
	$R_3 = -e^{-2\eta}$	$C_{31} = \mp 1,$
		$C_{32} = \mp e^{-\eta}$

5.3 Cyclotron Harmonics

5.3.1 Ion Cyclotron Case For ion cyclotron harmonics, without any absorption, the tunneling factor is given for the nth harmonic by

$$\eta = \frac{\pi}{4}\frac{\omega_{pi}L}{c}\frac{n_A^{2n+1}(n^2\beta_i/4)^{n-1}}{(n-1)![1/(n-1)+p^2]^2}, \qquad (129)$$

where $p = k_{\parallel}V_A/\omega$,

$$n_A^2 = \frac{[1/(n+1)-p^2][1/(n-1)+p^2]}{1/(n^2-1)+p^2}, \qquad (130)$$

and $\beta_i = 2\mu_0 n_i \kappa T_i/B_0^2$ if $\mathbf{B}\cdot\nabla B = 0$ with $B = B_0(1 + x/L)$ so that x (which is proportional to the dimensionless z) is in the direction of increasing magnetic field. From Fig. 9 and Table 1, we conclude that there is no reflec-

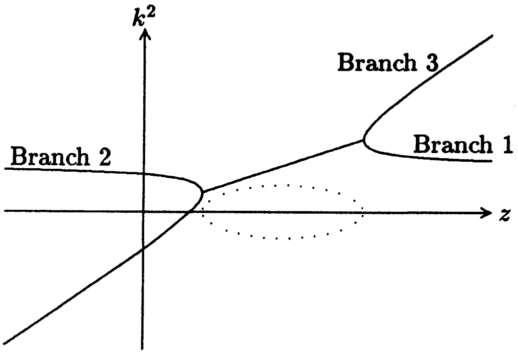

FIG. 9. Dispersion relation for tunneling equation for $\gamma > -1$.

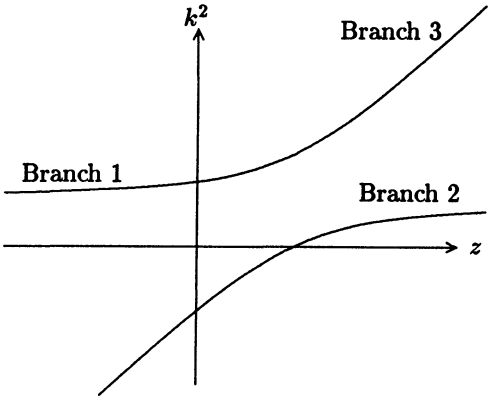

FIG. 10. Dispersion relation for tunneling equation for $\gamma < -1$.

tion from the high-magnetic-field side and transmission is symmetric, but not absorption, since the absorbed fractions on each branch are given by

$$A_1 = 1 - e^{-2\eta} - |C_{13}|^2/\epsilon, \tag{131}$$

$$A_2 = 1 - e^{-2\eta} - |R_2|^2 - |C_{23}|^2\epsilon, \tag{132}$$

$$A_3 = 1 - |R_3|^2 - \varepsilon|C_{31}|^2 - \varepsilon|C_{32}|^2, \tag{133}$$

where in this case, absorption is included in the reflection and conversion coefficients. With $p = 0$, there is no absorption, and Table 1 applies. Even with absorption, the transmission coefficient is unchanged, and there is still no reflection from the high-field side. The other coefficients, however, are reduced by absorption, so that, for example, numerical analysis shows that $R_2 \simeq R_{20}e^{-\kappa^2}$ where $\kappa = n_{\mathrm{A}}\sqrt{\beta_i}k_{\|}L$. The conversion coefficients also become exponentially small with strong absorption, with C_{23} falling much faster than C_{13}.

5.3.2 Electron Cyclotron Case For the second harmonic of the electron cyclotron frequency, $\gamma < -1$ and the tunneling factor for the X mode is given by

$$\eta = \frac{\pi}{4}\frac{\omega_{ce}L}{c}\frac{n_{\perp}^5}{(1 - n_{\|}^2 - \alpha_e/2)^2}\beta_e, \tag{134}$$

where $\alpha_e \equiv \omega_{pe}^2/\omega_{ce}^2$, $\beta_e = 2\mu_0 n_e \kappa T_e/B_0^2$, and

$$n_{\perp}^2 = \frac{(1 - n_{\|}^2 - \alpha_e/2)(1 - n_{\|}^2 - \alpha_e/6)}{(1 - n_{\|}^2 - \alpha_e/3)}. \tag{135}$$

There is also mode conversion for the O mode at the fundamental of the same order such that the tunneling factor is

$$\eta = \tfrac{1}{8}\pi(\omega_{ce}L/c)n_{\perp}\beta_e. \tag{136}$$

For the electron cyclotron cases, $z \propto -x$ so that the slow wave always propagates on the positive-z side. This means that the branch labels for the fast (cold plasma) modes are reversed and again there is no reflection from the high-magnetic-field side. The expressions above include $n_{\|} \neq 0$, where in general there is damping for either relativistic or nonrelativistic cases, but the relativistic absorption is always so much stronger that when considering the effects of absorption the relativistic analysis is nearly always needed. In this case, there is absorption even with $n_{\|} = 0$, and the expressions in Table 1 are never exact, except for the transmission coefficient, and $R_2 \simeq R_{20}\exp(-\frac{7}{2}\kappa^2)$, with $\kappa = \omega L n_{\perp}/\mu c$.

5.4 Cyclotron Emission

From Kirchhoff's law, there is an intimate relationship between emission and absorption. For the emission at cyclotron harmonics, one may assume that there is incident radiation on each branch of Fig. 9 or Fig. 10 of magnitude I_k and emission along each branch of E_k, $k = 1, 2, 3$. For steady state, the incoming and outgoing power must balance on each branch, and if the plasma and walls are assumed to be in thermal equilibrium, then $I_1 = I_2 = I_3 = I_{BB}$, where in the classical limit of $\kappa T \gg \hbar\omega$, $I_{BB} = \omega^2\kappa T/8\pi^3c^2$, with the result

$$E_k = A_k I_{BB}, \tag{137}$$

which is called the *generalized Kirchhoff's law* (Swanson and Shvets, 1992). If the walls are removed to infinity, then this is a local result and independent of the walls. This result differs from some classical calculations that describe emission in terms of the optical depth and the opacity such that

$$E = (1 - e^{-\tau})/I_{BB}, \tag{138}$$

where the optical depth is given by

$$\tau = 2\int \mathrm{Im}k(s)ds, \tag{139}$$

and the opacity is given by $1 - e^{-\tau}$. These analyses were based on the assumption that $k(s)$ is slowly varying, which is valid only in the limit when absorption is very strong so that both reflection and mode conversion are unimportant. The optical depth from Eq. (139) is equivalent to $\tau = 2\eta$, which in fact is determined entirely from the asymptotic parameters λ^2 and γ and is unrelated to absorption in any way. For the electron cyclotron case, we almost recover the classical formula, however, by considering that the emission along branch 3, which is typically the strongest-emitting branch, is totally absorbed close to the resonant layer via either

Landau or cyclotron damping. If a totally absorbing layer is close by, then it is also a blackbody emitting layer, which means there is an incident slow wave that is partially mode converted back into the fast waves, and from the reciprocity relations, this exactly cancels the converted-wave contribution to the absorption, with the result that for branches 1 and 2, which are the only branches that propagate out of the plasma, we have

$$E_1' = (1 - e^{-2\eta})I_{BB}, \tag{140}$$

$$E_2' = (1 - e^{-2\eta} - |R_2|^2)I_{BB}, \tag{141}$$

where the prime denotes the inclusion of both the direct and indirect emission. The result in Eq. (140) is exactly the classical result, and so the emission from the high–magnetic-field side follows the classical opacity formula, but the emission on the low–magnetic-field side is invariably less except in the limit of strong absorption where $R_2 \to 0$. From these arguments, one must redefine opacity.

For ion cyclotron emission, the absorption of the slow wave may occur a considerable distance from the cyclotron layer, so that the indirect emission may be from a region of different temperature and the cancellation of the mode-converted term is no longer exact, and the more general expressions for the A_k must be used. There is also the probability that plasma instabilities may excite a slow wave that will mode convert and mix with the thermal emission, and so the interpretation of ion cyclotron emission is more complicated.

6. NONLINEAR WAVES

There are a variety of nonlinear effects in plasmas that imply that the plasma is changed by the wave as it passes through the medium, and this can modify the wave propagation as well as produce whole new classes of phenomena. Examples are the coupling of waves of different frequencies and wave numbers, the production of very localized disturbances, and, in the strongly nonlinear limit, plasma turbulence. At the lowest level of nonlinearity, the change in the distribution function due to a finite-amplitude wave is the first effect, and the description of the effects of such a modification is called weak-turbulence theory.

6.1 Weak Turbulence

For some externally excited waves, where the wave amplitude can be maintained low, the neglect of finite-amplitude effects is often adequate, but for unstable waves, where the linear theory indicates wave-amplitude growth, the amplitude will eventually become large enough to affect the zero-order distribution function and couple the primary wave to other waves. When this leads to a large number of finite-amplitude waves, or a broad spectrum of waves, the plasma has become turbulent and nonlinear effects may dominate.

In hydrodynamic theory, the turbulent state is particularly difficult because the various nonlinear couplings are characteristically strong, so that it is difficult to isolate the various nonlinear couplings. In a plasma, however, this is frequently not the case, since the nonlinearly induced waves are frequently excited with small amplitudes, and unless these waves are also unstable, we may have a case of *weak turbulence*. As a first approximation, then, we may consider the different modes to be independent, and then add the coupling at various levels and focus our attention on one phenomenon at a time.

In the first approximation, the nonlinear coupling coefficients between various linear modes change slowly in time, and this is the domain of weak-turbulence theory. The fundamental smallness parameter in this expansion is the ratio of the wave energy density to the plasma energy density. More complete treatments are given elsewhere (Davidson, 1972; Galeev and Sagdeev, 1983).

The weak-turbulence theory is generally broken into three types of interactions:

1. *Quasilinear wave–particle interactions.* This is generally a resonant wave–particle process where $\omega = \mathbf{k}\cdot\mathbf{v}$ without any magnetic field, or where $\omega - n\omega_c = k_z v_z$ in a magnetic field. This interaction is due to a group of particles that remain in phase with the wave over many cycles and have time to exchange energy with the wave, resulting in slow changes in the distribution function that may be attributed to a

diffusion process called quasilinear diffusion. Because of the resonant nature of the interaction, this is a kinetic process, and gives rise to phenomena such as plateau formation on the distribution function, and general distortions of the distribution function such as tail formation.

2. *Nonlinear wave–wave interactions.* This process is resonant among the waves such that

$$\sum_i \omega_i = 0, \quad \sum_i \mathbf{k}_i = 0, \tag{142}$$

where ω_i and $\mathbf{k}_i$ are the frequencies and wave vectors of the participating waves. For this case, there are no particle resonances involved, and so a fluid approach may suffice. One of the most important examples of this kind of interaction is the class of parametric instabilities where two waves beat together to interact with a third, and above a certain threshold, the interaction goes unstable.
3. *Nonlinear wave–particle–wave interactions.* This process is also resonant, but here the particles keep constant phase relative to the beat frequency of two waves, so that the resonance condition is $\omega_1 - \omega_2 = (\mathbf{k}_1 - \mathbf{k}_2)\cdot\mathbf{v}$. This process is a kinetic process, and gives rise to phenomena such as plasma-wave echoes.

All of these resonance conditions can be viewed as arising from the conservation of energy and momentum, and so it will not be surprising when we find these conservation laws built into the various nonlinear formalisms. There are also quantum analogs to each of these three processes, but the quantum approach is generally much more complicated than a direct perturbation expansion.

6.1.1 1-D Theory Quasilinear theory was developed to treat weak turbulence observed in nonequilibrium plasmas. Because of the non-Maxwellian nature of a turbulent plasma, a spectrum of electromagnetic waves is generated from local charge separations or particle currents. This spectrum of waves, known as fluctuations, can then interact with the plasma through collisionless damping processes. On a short enough time scale, such that $t \ll \tau_b = \sqrt{m/ek^2\varphi_0}$, where τ_b is the bounce time for a particle in the electrostatic potential well of the wave, the linear Landau theory is adequate. However, the kinetic energy gain of the particles is proportional to the slope of the zero-order distribution function, $\partial f_0/\partial v|_{v=v_p}$. This implies that the damping or growth rate of the wave is independent of time even on the longer time scale over which the first-order distribution function changes, since the zero-order distribution is unaffected by the growth of the first-order distribution function. At some time, this clearly fails to describe the physics of wave–particle interactions if the first-order distribution function grows to be a non-negligible perturbation on the zero-order distribution.

It is the purpose of quasilinear theory to address the time evolution of the zero-order distribution function f_0. Moreover, this theory conserves energy and momentum between the wave and the particle distribution function. Another useful aspect of quasilinear theory is the interaction of the particles with an externally imposed wave such as in wave heating of a plasma or driving a macroscopic current in a plasma with waves. We should be cautious in applying these results, since these wave damping or wave growth processes are noncollisional so that the final state of the distribution function will not be Maxwellian. The treatment of only one nonlinear effect at a time for a non-Maxwellian distribution may occasionally lead to a serious error, especially as the amplitude grows, since somewhere one crosses the boundary from weak turbulence to strong turbulence. One rather surprising experimental observation is that the quasilinear description of the wave–particle interaction is quite good even for fairly large-amplitude waves.

For simplicity, we first restrict our discussion to the case of an unmagnetized plasma and develop the more general case later. This will enable us to bring out the physical processes more clearly without the mathematical complexity added by the magnetic field. Furthermore, the 1-D problem will be applicable to wave–particle interactions along a magnetic field where the Lorentz force is negligible.

The basic procedure of developing the 1-D quasilinear equations is the same as the treatment of the linear Landau case given in Sec. 3. Here, however, the quasilinear approximation implies that the distribution function is allowed to evolve nonlinearly while the wave is calculated using the linear method of Sec. 3. The distribution function is assumed to be locally uniform in space but to approach zero as one tends toward infinity, or $f(x \to \pm\infty, v, t) = 0$. We again assume there are no zero-order electric or magnetic fields and use the Vlasov equations for electrons only (assuming a uniform, stationary ion background to cancel all zero-order electric fields). Thus, we have

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} - \frac{e}{m_e}E\frac{\partial f}{\partial v} = 0, \tag{143}$$

where the distribution function is assumed to have the properties

$$f(x,v,t) = f_0(v,t) + f_1(x,v,t), \tag{144}$$

$$f_1(x,v,t) = \int_{-\infty}^{\infty}\frac{dk}{2\pi}\tilde{f}_k(k,v,t)e^{ikx}, \tag{145}$$

$$E(x,t) = \int_{-\infty}^{\infty}\frac{dq}{2\pi}\tilde{E}_q(q,t)e^{iqx}, \tag{146}$$

where now the zero-order distribution function depends on time and the tilde denotes the Fourier-transformed quantity. We will denote the average over space by

$$\langle f\rangle = \lim_{L\to\infty}\frac{1}{L}\int_{-L/2}^{L/2} f\,dx, \tag{147}$$

and so the first-order perturbations average to zero so that $\langle f\rangle = f_0(v,t)$ and $\langle E\rangle = 0$. We note that the average of the second term of Eq. (143) is eliminated by the boundary condition since $f(x \to \pm\infty) = 0$, and so we have a nonlinear equation for the time evolution of the zero-order distribution function from averaging Eq. (143):

$$\frac{\partial}{\partial t}f_0(v,t) = \frac{e}{m_e}\frac{\partial}{\partial v}\langle Ef_1\rangle. \tag{148}$$

Integrating over the space coordinate first and over q, we have nothing unless $q = -k$ so that

$$\langle Ef_1\rangle = \frac{1}{L_\infty}\int_{-\infty}^{\infty}\frac{dk}{2\pi}\tilde{f}_k(t)\tilde{E}_{-k}(t), \tag{149}$$

where we have introduced the abbreviated notation $\lim_{L\to\infty}1/L \equiv 1/L_\infty$.

From taking Fourier transforms of the kinetic equation, Eq. (64), in both space and time, the linear solution for $\tilde{f}_k$ is

$$\tilde{f}_k(v,t) = \frac{ie\tilde{E}_k}{m_e(\omega - kv)}\frac{\partial f_0}{\partial v}, \tag{150}$$

and so f_0 evolves as

$$\frac{\partial f_0}{\partial t} = \frac{ie^2}{m_e^2}\frac{\partial}{\partial v}\frac{1}{L_\infty}\int_{-\infty}^{\infty}\frac{dk}{2\pi}\frac{\tilde{E}_k\tilde{E}_{-k}}{(\omega - kv)}\frac{\partial f_0}{\partial v}. \tag{151}$$

The form of Eq. (151) is that of a diffusion equation with the spatial coordinate replaced by the velocity, or

$$\frac{\partial f_0}{\partial t} = \frac{\partial}{\partial v}D\frac{\partial}{\partial v}f_0, \tag{152}$$

where

$$D = \frac{e^2}{m_e^2 L_\infty}\int_{-\infty}^{\infty}\frac{dk}{2\pi}\left[i\mathfrak{p}\left(\frac{\tilde{E}_k\tilde{E}_{-k}}{\omega - kv}\right) + \pi\delta(\omega - kv)\tilde{E}_k\tilde{E}_{-k}\right], \tag{153}$$

where $\mathfrak{p}$ denotes the principal part of the integral. The two terms in Eq. (153) contribute differently to the wave–particle interaction. Since the principal part does not involve the particle with velocity $v = \omega/k$, it is nonresonant and has only a weak effect, but this effect is felt over the entire distribution function. The resonant term affects the distribution in the immediate neighborhood of the resonant particles, and if the distribution function is unstable, it tends to modify the distribution function toward a stable configuration, at which time it becomes unimportant. Then the nonresonant term dominates and gradually modifies the remaining distribution function to smooth it out. The $\tilde{E}_k\tilde{E}_{-k}$ terms indicate the nonlinear nature of the quasilinear diffusion coefficient, which depends on the wave energy density.

A particular example of these effects is given in Fig. 11 where an originally stable distribution has a gentle (broad) bump on the tail added as a result of a beam. Wherever $\partial f_0/\partial v > 0$, the distribution is unstable, and the wave amplitude grows rapidly. In the resonant region where the distribution is unstable, resonant velocity diffusion quickly flattens the distribution into the plateau re-

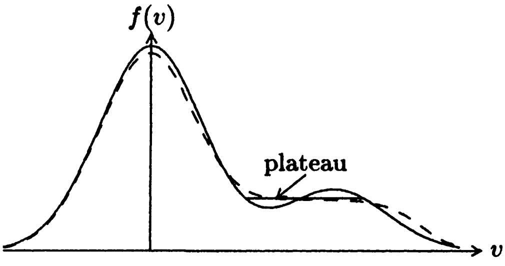

FIG. 11. Distribution function showing effects of quasilinear diffusion. The original distribution (solid) shows an unstable bump on the tail. The straight line shows the plateau that forms quickly through resonant diffusion. The nonresonant diffusion then modifies the entire distribution (dashed).

gime, as shown, and then the nonresonant diffusion begins to thermalize the entire distribution, passing through stages like that shown in the dashed curve.

6.2 Parametric Instabilities

Parametric instabilities are not unique to plasmas, as they relate to any oscillatory system where one of the parameters (hence the term parametric) is modulated at an appropriate frequency. A simple system is that of a child's swing, whose period depends on the length of the swing from its support, and the parametric instability occurs when the child lengthens or shortens the effective length once each period. This modulation at the natural frequency leads to growth of the fundamental oscillation, through what is commonly called "pumping."

Parametric instabilities in plasmas, including magnetized and unmagnetized plasmas, inhomogeneous and uniform plasmas, for linear and nonlinear electrostatic and electromagnetic waves, have been reviewed extensively (Kaw *et al.*, 1976). This section is intended to be an introduction to some of these topics in order to establish the fundamentals of parametric instabilities with a few examples, and as such abstracts some of the topics from this review.

Quite generally, parametric excitations require a minimal set of common characteristics:

1. *Matching condition:* The modulation and the natural oscillation should satisfy a phase-matching condition, such as $\omega T = n\pi$, $n = 1, 2, \ldots$, where ω is the natural frequency and T is the period of the modulation.
2. *Threshold:* Instability or amplification occurs only when the amplitude of the modulation exceeds a critical value.
3. *Frequency locking:* The frequencies of the amplified oscillations are determined by the modulation frequency rather than the natural frequency. For examples with $n = 1$, amplification is at frequency π/T (the natural frequency), but for $n = 2$, amplification occurs at $2\pi/T$ (the natural frequency again) *and* at zero frequency.

The matching and frequency-locking conditions follow from the intrinsic nonlinearity of the multiple-frequency system, and can be viewed either as coming from conservation of energy and momentum or as coming from the resonance conditions of the weak-turbulence analysis such that

$$\omega_0 = \omega_i + \omega_s, \tag{154}$$

$$\mathbf{k}_0 = \mathbf{k}_i + \mathbf{k}_s, \tag{155}$$

where the subscripts 0, i, and s stand for pump, idler, and signal, respectively. The instability occurs when the pump exceeds a certain threshold and the idler and signal waves grow. In general, we assume that the idler and signal represent waves that satisfy a dispersion relation in a plasma and are weakly damped modes, but this is not a strict requirement. When one of the waves is highly damped, we call it a quasimode, and numerous important applications involve quasimodes. From the nonlinear nature of the coupling, it is apparent that energy can be drawn from the pump wave and diverted to the idler and signal waves, or daughter waves. This process can become so efficient that as a pump wave propagates, it loses energy to the daughter waves until it is depleted to the extent that it falls below threshold. Among the various effects that limit this efficiency of coupling are the following: finite-wavelength effects, where the phase-matching conditions cannot be satisfied everywhere; finite pump extent, where the daughter waves grow only to a finite value because they propagate beyond the extent of the pump; and nonlinear effects, where saturation may occur due to quasilinear effects changing the decay and amplification rates, or to nonlinear frequency shifts that modify the matching condition.

In order to sort out this somewhat bewildering array of possible effects, we will examine one model that illustrates the fundamental phenomenon and then apply the principles thus discovered to a few specific plasma problems.

6.2.1 Coupled-Mode Case As plasmas are characterized by many natural frequencies, we shall choose as an example a case where two of the natural frequencies are far apart, with the pump frequency close to the higher frequency. When the sum of the two natural frequencies nearly equals the pump frequency, then both damped and locked excitations may be excited.

When the pump has a finite wavelength, the natural oscillations are described by $\mathcal{L}_s x_s = 0$ for $s = 1, 2$, where

$$\mathcal{L}_s \equiv \frac{d^2}{dt^2} + 2\gamma_s \frac{d}{dt} + (\omega_s^2 + \gamma_s^2), \tag{156}$$

where we assume that the frequencies ω_s and damping rates γ_s are both constant. We choose to denote the lower frequency as ω_1. The pump field is described by $Z(\mathbf{r},t) = 2\epsilon \times \cos(\mathbf{k}_0 \cdot \mathbf{r}\omega_0 t)$, with ϵ constant. This pump field couples $x_1(\mathbf{r},t)$ and $x_2(\mathbf{r},t)$ through the nonlinear interaction. The assumed coupling is of the form

$$\mathcal{L}_1 x_1(\mathbf{r},t) = \lambda_1 Z(\mathbf{r},t) x_2(\mathbf{r},t), \tag{157}$$

$$\mathcal{L}_2 x_2(\mathbf{r},t) = \lambda_2 Z(\mathbf{r},t) x_1(\mathbf{r},t), \tag{158}$$

where the λ_s are constants. We can imagine the pump and the high-frequency wave producing a modulation at their beat frequency, and if this forced oscillation resonates with the low frequency, then we expect a resonant energy transfer when $\omega_0 \simeq \omega_1 + \omega_2$. From the Fourier transforms of the coupled equations in space and time, the coupled equations take the form

$$\begin{aligned} D_1(\omega,\mathbf{k})\tilde{x}_1(\omega,\mathbf{k}) = \epsilon\lambda_1[&\tilde{x}_2(\omega + \omega_0,\mathbf{k} + \mathbf{k}_0) \\ &+ \tilde{x}_2(\omega - \omega_0,\mathbf{k} - \mathbf{k}_0)], \end{aligned} \tag{159}$$

$$\begin{aligned} &D_2(\omega \pm \omega_0,\mathbf{k} \pm \mathbf{k}_0)\tilde{x}_2(\omega \pm \omega_0,\mathbf{k} \pm \mathbf{k}_0) \\ &\quad = \epsilon\lambda_2\tilde{x}_1(\omega,\mathbf{k}), \end{aligned} \tag{160}$$

from which we obtain the dispersion relation

$$\begin{aligned} 1 = \frac{\epsilon^2\lambda_1\lambda_2}{D_1(\omega,\mathbf{k})} &\left[\frac{1}{D_2(\omega + \omega_0,\mathbf{k} + \mathbf{k}_0)}\right. \\ &\left. + \frac{1}{D_2(\omega - \omega_0,\mathbf{k} - \mathbf{k}_0)}\right], \end{aligned} \tag{161}$$

where the $D_s(\omega,\mathbf{k})$ are the dispersion relations from the Fourier transform of Eq. (156) associated with the linear waves. Defining $\omega \equiv \omega_r + i\gamma$ and separating the real and imaginary parts, we obtain

$$\begin{aligned} &\omega_r(\gamma + \gamma_1)[\Delta^2 - (\omega_r - \mu)^2 + (\gamma + \gamma_2)^2] \\ &\quad = (\omega_r - \mu)(\gamma + \gamma_2)[\omega_r^2 - \omega_1^2 - (\gamma + \gamma_1)^2], \end{aligned} \tag{162}$$

$$\begin{aligned} \frac{\Delta\kappa}{\langle\omega_2\rangle} &= \omega_r(\omega_r - \mu)(\gamma + \gamma_1)(\gamma + \gamma_2) \\ &\quad \times \left\{4 + \frac{[\omega_r^2 - \omega_1^2 - (\gamma + \gamma_1)^2]^2}{\omega_r^2(\gamma + \gamma_1)^2}\right\}, \end{aligned} \tag{163}$$

where

$$\mu = \tfrac{1}{2}[\omega_2(\mathbf{k} + \mathbf{k}_0) - \omega_2(\mathbf{k} - \mathbf{k}_0)], \tag{164}$$

$$\Delta = \omega_0 - \langle\omega_2\rangle, \tag{165}$$

$$\langle\omega_2\rangle = \tfrac{1}{2}[\omega_2(\mathbf{k} + \mathbf{k}_0) + \omega_2(\mathbf{k} - \mathbf{k}_0)], \tag{166}$$

and $\kappa = \epsilon^2\lambda_1\lambda_2 > 0$. There are three distinct roots for ω_r with $\mu > 0$: mode I with $\omega_r > \mu$, mode II with $\mu > \omega_r > 0$, and mode III with $\omega_r < 0$. Since $\kappa > 0$, Eq. (163) indicates that both mode I and mode III occur when $\Delta > 0$, and mode II corresponds to $\Delta < 0$. The uniform-pump case ($\mu = 0$) shows that mode II is a purely growing mode and the other two are decay–type modes. For this case with $\mu \neq 0$, however, there is no purely growing mode. Secondly, when $\mu > \omega_1$, then the frequency of mode I becomes nonresonant and we have a quasimode. For $\mu \neq 0$, the threshold can be substantially reduced for $\omega_r \sim \omega_1$, and the growth rate well above threshold can be significantly increased. For the resonant decay case, when $D_1(\omega,k) \simeq 0$, then the growth rate and threshold are given by

$$\gamma = \tfrac{1}{2}[-(\gamma_1 + \gamma_2) + \sqrt{(\gamma_1 - \gamma_2)^2 + \kappa/\omega_r\langle\omega_2\rangle}] \tag{167}$$

and

$$\kappa_0 = 4\omega_r\langle\omega_2\rangle\gamma_1\gamma_2. \tag{168}$$

6.2.2 Unmagnetized Plasma In the unmagnetized plasma, the principal waves are the ordinary wave, where the dispersion relation describing *photons* is

$$D(\omega,k) = \omega_{pe}^2 + k^2c^2 - (\omega - i\nu_e)^2, \tag{169}$$

where ν_e is the collision frequency; the plasma wave from Eq. (2), which may be written as

$$D(\omega,k) = \omega_{pe}^2 + \tfrac{3}{2}k^2v_e^2 - (\omega - i\gamma_p)^2, \tag{170}$$

which describe what are called *plasmons*, where

$$\gamma_p = \sqrt{\frac{\pi}{8}}\frac{\omega_{pe}}{k^3\lambda_D^3}\exp\left[-\frac{1}{2k^2\lambda_D^2} - \frac{3}{2}\right]; \tag{171}$$

and the ion-acoustic wave from Eq. (69), written as

$$D(\omega,k) = \frac{k^2c_s^2}{1 + k^2\lambda_D^2} - (\omega - i\gamma_a)^2, \tag{172}$$

which describe what are called *phonons*, where $k\lambda_D \ll 1$ and

$$\gamma_a = \sqrt{\frac{\pi}{8}}\left[\sqrt{\frac{m_e}{m_i}} + \left(\frac{T_e}{T_i}\right)^{3/2}\exp\left(-\frac{T_e}{2T_i}\right)\right] + \nu_i, \tag{173}$$

which includes a collisional term.

6.2.2.1 Case 1—The Pump Is a Plasma Wave. For this case, ω_0 satisfies Eq. (170) with wave vector $\mathbf{k}_0$. If $\mathbf{k} \parallel \mathbf{k}_0$, then

$$\mu \simeq \frac{3kk_0v_e^2}{2\omega_0}, \quad \Delta \simeq -\frac{3k^2v_e^2}{4\omega_0}. \tag{174}$$

With $\Delta < 0$, this is mode II where $\mu > \omega_r > 0$. We note two special cases:

1. When $\mu \gg \omega_1 \simeq \omega_r$, then Eq. (162) requires that $\Delta^2 \sim \mu^2$ so that $k \sim 2k_0 \sim 2(k - k_0)$. In this case, a plasmon has decayed into another plasmon [ω_2 from Eq. (170)] and a phonon [ω_1 from Eq. (172)]. This is an example where an electron plasma wave is *backscattered* by an ion wave since $k_0 - k \sim -k_0$.
2. Another limiting case is when $\mu \simeq \omega_1 \simeq \omega_r$, in which case Eq. (162) leads to $\Delta^2 \sim 0$, so k^2 must be small. In this limit, $k_0 \pm k \sim k_0$, so this case represents *forward scattering*.

6.2.2.2 Case 2—The Pump Is an Electromagnetic Wave. Here we note four cases:

1. When the pump frequency is just above the plasma frequency, it may decay into a plasmon (ω_2) and a phonon ($\omega_1 = \omega_0 - \omega_2 \ll \omega_0$). Since all three waves are different (photon → plasmon + phonon), this is a decay process that damps the pump.
2. Instead of decaying into a plasmon and a phonon when the pump frequency is just above the plasma frequency, it can also decay into another photon (a scattered electromagnetic wave) and a phonon. This is called *Brillouin scattering*.
3. When the pump frequency is higher such that $\omega_0 > 2\omega_p$, then it may decay into another photon (a scattered electromagnetic wave) and a plasmon. This type of scattering is called *Raman scattering*.
4. For the special case of $\omega_0 = 2\omega_p$, the photon may decay into two plasmons.

A number of these cases with an electromagnetic wave pump can be summarized by a few general relations if we take the driven-wave amplitude to be

$$\mathbf{E}_0^{\mp} = E_0\hat{e}_0\exp[\pm i(\mathbf{k}_0\cdot\mathbf{r} - \omega_0 t)], \tag{175}$$

so that the corresponding electron velocity is $\mathbf{v}_0 = \mathbf{v}_0^+ + \mathbf{v}_0^-$ where $\mathbf{v}_0^{\pm} = \mp ie\mathbf{E}_0^{\pm}/m_e\omega_0$. This driven velocity of the electrons times the density fluctuation of a daughter wave generates a current density that drives waves at the sum and difference frequencies of the velocity and density fluctuations. The dispersion relation then takes the form

$$\mathbf{E}^{\pm}\cdot\mathbf{E}_0^{\mp} = \frac{\omega_{pe}^2\tilde{n}_e(\omega,\mathbf{k})}{k_{\pm}^2 n_0} \times \left(\frac{(\mathbf{k}_{\pm}\times\mathbf{E}_0^{\mp})^2}{D_{2t}^{\pm}} - \frac{(\mathbf{k}_{\pm}\cdot\mathbf{E}_0^{\mp})^2}{D_{2l}^{\pm}}\right), \tag{176}$$

where D_{2t} and D_{2l} are the transverse (photon) and longitudinal (plasmon) dispersion functions given by

$$D_{2t}^{\pm} = k_{\pm}^2 c^2 - \omega_{\pm} K_1(\omega_{\pm}, \mathbf{k}_{\pm}), \tag{177}$$

$$D_{2l}^{\pm} = -\omega_{\pm} K_3(\omega_{\pm}, \mathbf{k}_{\pm}). \tag{178}$$

By introducing the wave coupling through the ponderomotive force, the dispersion relation can be put into the form of Eq. (161) where D_1 is given by

$$D_1(\omega,\mathbf{k})\tilde{n}_e(\omega,\mathbf{k}) = -\frac{\omega_{pe}^2 k^2 \epsilon_0}{m_e \omega_0^2} f(\mathbf{E},\mathbf{E}_0) \tag{179}$$

or

$$D_1(\omega,\mathbf{k})\tilde{n}_e(\omega,\mathbf{k}) = -\frac{\omega_{pi}^2 k^2 \epsilon_0 f(\mathbf{E},\mathbf{E}_0)}{m_e \omega_0^2 (1 + k^2 \lambda_{\mathrm{D}}^2)}, \tag{180}$$

where $f(\mathbf{E},\mathbf{E}_0) = (\mathbf{E}_0^+ \cdot \mathbf{E}^- + \mathbf{E}_0^- \cdot \mathbf{E}^+)$, which represent a plasmon or a phonon, respectively. From these results and Eq. (176), we find that the threshold parameters for several cases may be expressed in terms of $\mathbf{v}_0 = e\mathbf{E}_0/m_e\omega_0$ as in Table 2.

6.2.3 Brillouin Scattering For this case, D_2 is a photon dispersion function and D_1 is a phonon dispersion function, so that

$$\kappa = (k/k_{\pm})^2 \omega_{pi}^2 (\mathbf{k}^- \times \mathbf{v}_0)^2/(1 + k^2\lambda_{\mathrm{D}}^2), \tag{181}$$

since $\mathbf{k} \sim 2\mathbf{k}_0$, making $\mathbf{k}^-$ the resonant mode. In the resonant scattering case, $\omega \simeq \omega_1$, and so the threshold condition from Eq. (163) or Eq. (168) with $\gamma = 0$ is

$$k^2 v_0^2 \sin^2\theta \simeq \frac{4\gamma_a \gamma_2 \omega_1 \omega_0}{\omega_{pi}^2}(1 + k^2\lambda_{\mathrm{D}}^2), \tag{182}$$

where θ is related to the angle between $\mathbf{k}_-$ and $\mathbf{v}_0$. The corresponding growth rate just above threshold from Eq. (167) is

$$\gamma = -\gamma_2 + \frac{\omega_{pi}^2 k^2 v_0^2 \sin^2\theta}{4\omega_1\omega_0\gamma_a(1 + k^2\lambda_{\mathrm{D}}^2)}, \tag{183}$$

for $\gamma \gg \gamma_2 \sim \nu_e$.

6.2.4 Raman Scattering In this example, a photon is scattered off a plasmon rather than a phonon. Hence, D_1 represents a plasma wave from Eq. (170), and D_2 an electromagnetic wave from Eq. (169). Again for this case, $k - k_0$ is the resonant root, so that $\kappa = (k/k_-)^2\omega_{pe}^2(\mathbf{k}_- \times \mathbf{v}_0)^2 = k^2 v_0^2 \omega_{pe}^2 \times \sin^2\theta'$ from Table 2 where θ' again represents an angle between $\mathbf{k}$ and $\mathbf{v}_0$. The threshold is then given by Eq. (168) as

$$k^2 v_0^2 \sin^2\theta' = 4\gamma_p \gamma_2 \langle\omega_2\rangle/\omega_{pe}, \tag{184}$$

and the minimum threshold is for $k \sim 2k_0$ (backscattering) and $\theta' = \pi/2$, when $k_0^2 v_0^2 = \gamma_p\gamma_2\langle\omega_2\rangle/\omega_{pe}$. The gain just above threshold from Eq. (167) (resonant case) is

$$\gamma = -\gamma_2 + \frac{\omega_{pe} k^2 v_0^2 \sin^2\theta'}{4\langle\omega_2\rangle\gamma_p}. \tag{185}$$

6.3 Shock Waves

Shock waves are characteristic of a medium when the wave amplitude gets large enough to begin to modify the medium to the extent that the velocity of the wave is affected. Typically, a shock is related to a sound speed, and in a plasma, a classic example is the ion-acoustic shock wave (or solitary wave). It is assumed in the analysis that the wave shape is stationary in the frame moving with the wave, though not sinusoidal, as one usually assumes a single wave front so that far from the front, all derivatives vanish.

We shall assume the ions to be cold, but the electrons will have a finite temperature. We begin with the fluid equations of continuity and momentum,

Table 2. Threshold parameter κ for several combinations of daughter waves for an electromagnetic pump wave.

Case	D_1	D_2	κ
3 Raman	plasmon	photon	$k^2\omega_{pe}^2(\mathbf{k}_{\pm} \times \mathbf{v}_0)^2/k_{\pm}^2$
2 Brillouin	phonon	photon	$k^2\omega_{pi}^2(\mathbf{k}_{\pm} \times \mathbf{v}_0)^2/(1 + k^2\lambda_{\mathrm{De}}^2)k_{\pm}^2$
4 decay	plasmon	plasmon	$k^2\omega_{pe}^2(\mathbf{k}_{\pm} \cdot \mathbf{v}_0)^2/k_{\pm}^2$
1 decay	phonon	plasmon	$k^2\omega_{pi}^2(\mathbf{k}_{\pm} \cdot \mathbf{v}_0)^2/(1 + k^2\lambda_{\mathrm{De}}^2)k_{\pm}^2$

$$\frac{\partial n}{\partial t} + \frac{\partial}{\partial z} nv = 0, \tag{186}$$

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial z} = -\frac{q}{m}\frac{\partial \varphi}{\partial z} - \frac{1}{nm}\frac{\partial p}{\partial z}, \tag{187}$$

and Poisson's equation,

$$\frac{\partial^2 \varphi}{\partial z^2} = -\frac{e}{\epsilon_0}(n_i - n_e). \tag{188}$$

For electrons, m_e is taken to be so small that we can neglect the left-hand side of the momentum equation, and then letting $p_e = n_e \kappa T_e$, we obtain

$$n_e = n_0 \exp(e\varphi/\kappa T_e). \tag{189}$$

Assuming now a wavelike solution so that all functions may be written as functions of $\xi = z - Ut$, then $\partial/\partial t \to -U\partial/\partial\xi$ and $\partial/\partial z \to \partial/\partial\xi$, so that the continuity and momentum equations for ions become ordinary differential equations with solution

$$n_i = n_0/\sqrt{1 - 2e\varphi/m_i U^2}. \tag{190}$$

Then Poisson's equation may be written as

$$\frac{d^2\varphi}{d\xi^2} = \eta\frac{d}{d\varphi}\eta = \frac{d}{d\varphi}\left(\frac{\eta^2}{2}\right), \tag{191}$$

with the change of variable $\eta = d\varphi/d\xi$, and the solution is

$$\frac{\eta^2}{2} = \frac{n_0}{\epsilon_0}\left\{m_i U^2\left(\sqrt{1 - \frac{2e\varphi}{m_i U^2}} - 1\right)\right.$$
$$\left. + \kappa T_e\left[\exp\left(\frac{e\varphi}{\kappa T_e}\right) - 1\right]\right\}, \tag{192}$$

where the constant terms are chosen so that $\eta \to 0$ as $\varphi \to 0$. Now solitary-wave solutions do not exist for all values of φ and U. It is clear from Eq. (192) that $e\varphi \leq m_i U^2/2$ for a meaningful solution.

If we consider $v(\varphi) = -\eta^2/2$ to be a "pseudopotential" with φ the "coordinate" and ξ the "time," then Eq. (191) has the form of an equation of motion for a particle moving in a "pseudopotential well,"

$$\frac{d^2\varphi}{d\xi^2} = -\frac{dv(\varphi)}{d\varphi}, \tag{193}$$

and we need a potential well for $v(\varphi)$ with two zero crossings for a "bound" (localized) solution. From Eq. (192) we have $v(\varphi) = 0$ at $\varphi = 0$, and so we require another zero crossing for a solitary wave. Using the parameter $\alpha \equiv U^2/c_s^2$ with $c_s^2 = \kappa T_e/m_i$, then for $\alpha < 1$, v increases monotonically with φ, so that there is no second zero crossing. For $\alpha_c > \alpha > 1$ with $\alpha_c = 2.51$, there is a well with two zero crossings, while for $\alpha > \alpha_c$ there is a well but without a second zero. Thus there exist solitary-wave solutions of this type only for Mach numbers $1 < \mathcal{M} < \sqrt{\alpha_c}$ where the Mach number is defined as $\mathcal{M} = U/c_s$ and c_s is the ion-acoustic speed.

6.4 Solitons

In describing shock waves or solitary waves, it is assumed that the wave shape is stationary in the wave frame, but this is not always true in practice. A subclass of solitary waves does have a special kind of stability in the shape and also describes the interactions of two or more individual localized waves that survive collisions with one another, and these are called *solitons*. In this section, we describe only two general types of solitons, although many others are known.

6.4.1 The Korteweg–de Vries (KdV) Equation Although there are many different nonlinear partial differential equations (PDEs) that give rise to solitons, there is one soliton equation that is generic to many different kinds of waves. The Korteweg–de Vries equation, which was shown in 1896 to describe shallow-water waves, provided the key to solving nonlinear PDEs by the method of inverse scattering. The general form of the KdV equation is

$$u_\tau + auu_\xi + bu_{\xi\xi\xi} = 0, \tag{194}$$

where a and b are constants (the classical form has $a = \pm 6$ and $b = 1$) and the subscripts denote partial derivatives. The nonlinearity is evident from the middle term. We may note that the first two terms typically come from the first two terms of the momentum equation so that τ and ξ usually represent normalized time and space variables, respectively. The third term of Eq. (194) is a dispersive term, and we motivate its form and indicate its generality by examining the form of a very simple dispersion relation.

Neglecting damping, the simple dispersion relation for the ion-acoustic wave is given by Eq. (69), which we may write as

$$\omega = \pm kc_s(1 - k^2\lambda_D^2/2 + \cdots), \tag{195}$$

which to the order indicated is representative in form of the dispersion relations of a great many waves that are characterized by a constant phase velocity in the long-wavelength limit and some dispersion. Neglecting the higher-order terms, we can identify the term on the left as coming from $\partial/\partial t$, and the first term on the right as coming from $\partial/\partial x$ [with the characteristic speed being amplitude dependent if we compare term by term with Eq. (194)]. The second term on the right is proportional to k^3, and so it represents a third-order spatial derivative as we find in the KdV equation. The importance of this dispersive term must not be underestimated, since if it were neglected in Eq. (194), then the peak amplitude would grow and speed up, eventually leading to wave breaking where the peak overtakes the trough, and the wave becomes multiple valued. The dispersive term limits this growth, and the tradeoff between the tendency to peak from the nonlinear term and the tendency to spread due to the dispersive term leads to the stable character of solitons.

The single-soliton solution of Eq. (194) may be simply found by assuming that a stationary solution exists in some rest frame moving at velocity U. Assuming then that $u = u(\xi - U\tau)$ leads to an ordinary differential equation, and the boundary conditions that $\lim_{|\xi|\to\infty} u = 0$ along with all its derivatives, allows the solution to be found.

6.4.2 Ion-Acoustic Solitons It has been shown that ion-acoustic waves may be KdV solitons (Washimi and Taniuti, 1966) beginning with the same equations as in Sec. 6.3. Using the normalized variables $n = n_i/n_0$, $u = v_i/c_s$, $x = z/\lambda_D$, $\tau = c_st/\lambda_D$, and $\Phi = e\varphi/\kappa T_e$, along with the variables $\eta = \omega(x - \tau)$, $\xi = \omega^3 x$, and a smallness parameter $\epsilon = \omega^2$, and expanding in the parameter ϵ, one obtains to first order

$$\frac{\partial n^{(1)}}{\partial \eta} = \frac{\partial u^{(1)}}{\partial \eta} = \frac{\partial \Phi^{(1)}}{\partial \eta}, \tag{196}$$

which leads to $n^{(1)} = u^{(1)} = \Phi^{(1)} \equiv \phi$. To second order, the result leads to

$$\phi_\xi + \phi\phi_\eta + \tfrac{1}{2}\phi_{\eta\eta\eta} = 0, \tag{197}$$

which is the Korteweg–de Vries (KdV) equation with $a = 1$ and $b = \frac{1}{2}$.

The KdV equation is more than an ordinary solitary-wave equation, since it has been proved that individual solitary waves survive collisions, and hence these are called solitons to indicate their particlelike behavior. The KdV equation and other soliton equations may be solved by the method of inverse scattering, which is a method of solving the nonlinear partial differential equations for *arbitrary initial conditions* by *linear* methods. The solution via inverse scattering for the KdV equation is discussed by many authors (e.g., see Drazin and Johnson, 1989). Through this formalism, it has been possible to show that the solitary waves are indeed *normal modes* and that arbitrary perturbations invariably relax to these solitary waves and the remainder "radiates" away. Analytic N-soliton solutions are also known, and there are an infinite number of conservation laws associated with the KdV equation and other soliton equations. The one-soliton solution of Eq. (197) is given by

$$\phi = A\,\mathrm{sech}^2[k(\eta - U\xi)], \tag{198}$$

and in terms of the original variables, the result is $\Phi = \Phi_m\,\mathrm{sech}^2[K(x - vt)]$ with $v = c_s/(1 - \Phi_m/3)$, $K = \sqrt{\Phi_m/6}(1 - \Phi_m/3)/\lambda_D$.

6.4.3 Alfvén-Wave Solitons In addition to ion-acoustic solitons, which have been experimentally observed to survive collisions, nonlinear Alfvén waves also are characterized by the KdV equation. This is because they propagate near the Alfvén speed at low frequency and long wavelength and begin to show dispersion at higher frequencies. Both fast and slow Alfvén waves have been shown to satisfy the KdV equation.

6.4.4 Nonlinear Schrödinger Equation The nonlinear Schrödinger equation describes a nonlinear wave where the wave amplitude is large enough to modify the dispersion relation by changing the parameters of the medium. Its solution is an envelope soliton where the wave-packet envelope has the soliton character where the medium is modified, and the complex phase term propagates at the wave phase velocity.

We begin with the assumption of a plane wave of the form $A \exp(i\mathbf{k}\cdot\mathbf{r} - i\omega t)$ + c.c. and a nonlinear dispersion relation

$$D(\omega,\mathbf{k},|A|^2) = 0, \tag{199}$$

where it has been assumed that the nonlinearity depends on the amplitude of the wave and not on the phase. This kind of nonlinearity is quite typical of the ponderomotive force. We also presume that the medium is homogeneous and isotropic in the absence of the wave, so that D and A are scalar quantities. We assume that the nonlinear effects will result in a slow modulation of the wave amplitude about an average amplitude A_0, and so we expand ω and $\mathbf{k}$, which represent the time and space differential operators, about the fast-variation values, so that Eq. (199) becomes a differential equation for the slowly varying amplitude,

$$D\left[\omega + i\frac{\partial}{\partial t}, \mathbf{k} - i\nabla, |A(\mathbf{r},t)|^2\right] A = 0. \tag{200}$$

Assuming $|A|^2 - |A_0|^2 \ll |A_0|^2$, then this may be written as the 3-D nonlinear Schrödinger equation

$$i\frac{\partial A}{\partial t} + \mathbf{P}\colon \nabla_\xi^2 A + Q(|A|^2 - |A_0|^2)A = 0, \tag{201}$$

where we have shifted to the wave frame moving with the group velocity, so that $\xi = \mathbf{r} - \mathbf{v}_g t$, and

$$\mathbf{P} = \tfrac{1}{2}\frac{\partial \mathbf{v}_g}{\partial \mathbf{k}} = \tfrac{1}{2}\frac{\partial^2 \omega}{\partial \mathbf{k}\partial \mathbf{k}}, \tag{202}$$

$$Q = \frac{\partial D}{\partial |A_0|^2}\left(\frac{\partial D}{\partial \omega}\right) = -\frac{\partial \omega}{\partial |A_0|^2}. \tag{203}$$

We can see that Q represents the nonlinear frequency shift due to the finite wave amplitude. We note a difficulty with names at this point, since the 1-D version of Eq. (201),

$$i\frac{\partial A}{\partial t} + P\frac{\partial^2 A}{\partial \xi^2} + Q(|A|^2 - |A_0|^2)A = 0, \tag{204}$$

is commonly called the nonlinear Schrödinger equation, but only the special case with $A_0 = 0$ is a soliton equation that goes by the same name. With $A_0 \neq 0$, Eq. (201) is generally unstable with the maximum growth rate

$$\gamma_{\max} = Q|A_0|^2. \tag{205}$$

The implication of this linear instability is that there exists a range where the wave amplitude will grow. This leads to a collapse where the amplitude will grow until all of the plasma has been expelled as a result of the ponderomotive force, resulting in a density cavity, or *caviton*. For a driven wave, then, it is expected that the wave will collapse, but the severe modification of the plasma may decouple the driven pump wave, so that effectively $A_0 \to 0$, in which case the equation becomes the soliton equation that is stable. These general characteristics are borne out in numerical simulations and experiments.

6.4.5 General Wave Collapse In addition to the caviton collapse above, it has been shown quite generally that in higher dimensional systems, i.e., with either cylindrical or spherical symmetry, where the nonlinear Schrödinger equation assumes the form

$$iu_t + \frac{\partial^2 u}{\partial r^2} + \frac{d-1}{r}\frac{\partial u}{\partial r} + |u|^s u = 0, \tag{206}$$

where d is the dimension and s indicates the degree of nonlinearity, the solutions are stable for $sd < 4$, but unstable for $sd \geq 4$ (Shvets, 1993). In each unstable case, the volume collapses to zero in a finite time, and the classification depends on the assumption of some absorption. For $sd = 4$, the collapse is called *strong*, as the energy absorbed is virtually independent of the absorption strength. For $4 < sd < 2s + 2$, the collapse is called *weak*, since the energy absorbed tends toward zero along with the absorption coefficient. On the other hand, for $sd \geq 2s + 2$, the collapse is *superstrong*, as the presence of absorption slows the collapse as the volume continues to absorb energy from the surrounding medium. Important examples are plasma-wave collapse, both with and without a magnetic field, and lower-hybrid wave collapse, both of which have been shown to convert wave energy to energetic particles as they generate strong turbulence.

GLOSSARY

Absolute Instability: An instability that grows in time everywhere in space.

Alfvén Waves: Waves whose velocity is proportional to the magnetic field strength. The torsional Alfvén wave is strongly guided along the field, while the compressional Alfvén wave is nearly isotropic.

Bernstein Wave: An electrostatic wave that propagates perpendicular to the magnetic field in bands between harmonics of the ion and electron cyclotron frequencies.

Convective Instability: An instability that grows in time in the local rest frame but that decays in time in some rest frame.

Cutoff: Any case where the index of refraction vanishes. In bounded plasmas, cutoff is when the index of refraction along the guide vanishes.

Electrostatic Wave: Any wave where the wave magnetic-field component is negligible. A sufficient condition is that the index of refraction be large compared with any of the dielectric tensor components.

Extraordinary Wave: A wave whose direction and wave electric field are both perpendicular to the magnetic field.

Ion-Acoustic Wave: An ordinary wave propagating close to the ion plasma frequency, whose speed is given by $c_s = \sqrt{\kappa T_e/m_i}$. Requires $T_e \gg T_i$.

Landau Damping: Collisionless damping due to phase mixing of particles traveling near the phase velocity of the electrostatic wave. In a rotating frame, accounts also for collisionless cyclotron damping.

Lower-Hybrid Resonance: A plasma resonance for waves propagating perpendicular to the magnetic field, which lies between the ion and electron cyclotron frequencies.

Mode Conversion: That process near a cold-plasma or cyclotron harmonic resonance that when resolved by thermal effects leads to energy being converted linearly to a warm-plasma wave. Frequently associated with tunneling, reflection, and absorption.

Ordinary Wave: A plasma wave propagating perpendicular to the magnetic field whose electric-field component is parallel to the magnetic field.

Parametric Instability: A nonlinear instability where several waves are coupled together so that one large-amplitude wave (pump wave) gives rise to daughter waves at different frequencies.

Quasilinear Diffusion: The diffusion in velocity space due to the weakly nonlinear effect of plasma waves on the particle distribution function.

Resonance: Any frequency where the index of refraction is unbounded. Cold-plasma resonances are typically resolved by warm-plasma effects.

Shock Waves: Large-amplitude waves characterized by a solitary wave as opposed to a wave train. Characteristically, the speed of a shock wave is amplitude dependent.

Solitons: A special subclass of nonlinear solitary waves whose shapes are unaffected by collisions with one another.

Upper-Hybrid Resonance: A plasma resonance for waves propagating perpendicular to the magnetic field, which lies above both the electron cyclotron and the electron plasma frequencies.

Weak Turbulence: The condition when a nonequilibrium velocity distribution leads to a narrow to modest spectrum of weak instabilities.

Whistler Wave: The right-handed wave well above the ion cyclotron frequency and well below the electron cyclotron frequency.

Works Cited

Allis, W. P. (1959), *Sherwood Conference on Controlled Fusion*, Gatlinburg, TN, p. 32; see also Allis, W. P., Buchsbaum, S. J., Bers, A. (1963), *Waves in Anisotropic Plasmas*, Cambridge, MA: MIT Press.

Appleton, E. V. (1932), *J. Inst. Electr. Eng.* **71**, 642–650.

Bohm, D., Gross, E. P. (1949), *Phys. Rev.* **75**, 1851–1876.

Braginskii, S. I. (1965), in: M. A. Leontovich (Ed.), *Review of Plasma Physics*, Vol. 1, New York: Consultants Bureau, pp. 205–311.

Briggs, R. J. (1964), *Electron-Stream Interaction with Plasmas*, Cambridge, MA: MIT Press.

Budden, K. G. (1955), *Radio Waves in the Ionosphere*, Cambridge, U.K.: Cambridge Univ. Press.

Davidson, R. C. (1972), *Nonlinear Plasma Theory*, New York: Academic Press.

Drazin, P. G., Johnson, R. S. (1989), *Solitons: an Introduction*, Cambridge, U.K.: Cambridge Univ. Press.

Fisher, R. K., Gould, R. W. (1971), *Phys. Fluids* **14**, 857–867.

Fried, B. D., Conte, S. D. (1961), *The Plasma Dispersion Function*, New York: Academic Press.

Galeev, A. A., Sagdeev, R. Z. (1983), in: A. Galeev, R. N. Sudan (Eds.), *Handbook of Plasma Physics*, Vol. 1, Amsterdam: North-Holland, pp. 677–731.

Gillmor, C. Stewart (1982), *Proc. Am. Philos. Soc.* **126**, 395–440.

Kaw, P. K., Kruer, W. L., Liu, C. S., Nishikawa, Kyoji (1976), in: A. Simon, W. B. Thompson (Eds.), *Advances in Plasma Physics*, Vol. 6, New York: Interscience, pp. 1–269.

Krall, N. A. (1968), in: A. Simon, W. B. Thompson (Eds.), *Advances in Plasma Physics*, Vol. 1, New York: Interscience, pp. 153–199.

Kuehl, H. H. (1962), *Phys. Fluids* **5**, 1095–1103.

Landau, L. D. (1946), *J. Phys. (U.S.S.R.)* **10**, 25–34.

Mikhailovskii, A. B. (1983), in: A. Galeev, R. N. Sudan (Eds.), *Handbook of Plasma Physics*, Vol. 1, Amsterdam: North-Holland, pp. 587–610.

Parker, J. V., Nickel, J. C., Gould, R. W. (1964), *Phys. Fluids* **7**, 1489–1500.

Shvets, V. F. (1993), *Wave Collapse in Nonlinear Media*, Alexandria, VA: Global Consultants, Inc.

Stix, T. H. (1962), *The Theory of Plasma Waves*, New York: McGraw-Hill.

Stix, T. H. (1965), *Phys. Rev. Lett.* **15**, 878–881.

Stringer, T. E. (1963), *Plasma Phys. (J. Nucl. Energy, Pt. C)* **5**, 89–107.

Sturrock, P. A. (1958), *Phys. Rev.* **112**, 1488–1503.

Swanson, D. G. (1985), *Phys. Fluids* **28**, 2645–2677.

Swanson, D. G., Shvets, V. F. (1992), *Phys. Rev. Lett.* **68**, 3036–3039.

Trivelpiece, A. W., Gould, R. W. (1959), *J. Appl. Phys.* **30**, 1784–1793.

Washimi, H., Tanuiti, T. (1966), *Phys. Rev. Lett.* **17**, 996–998.

Weinberg, S. (1962), *Phys. Rev.* **126**, 1899–1909.

Further Reading

A. Galeev, R. N. Sudan (Eds.) (1983), *Handbook of Plasma Physics*, Vols. 1, 2, Amsterdam: North-Holland.

Stix, T. H. (1992), *Waves in Plasmas*, New York: AIP.

Swanson, D. G. (1989), *Plasma Waves*, Boston: Academic Press.

PLATE TECTONICS

Joann M. Stock, *Seismological Laboratory, California Institute of Technology, Pasadena, California, U.S.A.*

Introduction 273
1. **General Theory of Plate Tectonics** 274
1.1 History 274
1.2 Types of Plate Boundaries 275
1.3 Location and Characteristics of Active Plate Boundaries 277
1.3.1 Earthquakes 277
1.3.2 Magmatism 277
1.3.2.1 Midocean Ridge Basalts 277
1.3.2.2 Island-Arc and Convergent-Margin Magmatic Rocks 277
1.3.2.3 Ocean Island Basalts 278
1.4 Geometry of Instantaneous Plate Motions on a Sphere 278
1.5 Past Plate Motions 279
2. **Composition and Structure of Plates** 280
2.1 Composition 280
2.1.1 Oceanic Crust and Mantle 281
2.1.2 Continental Crust and Mantle 281
2.1.3 Isostasy 282
2.2 Thermal and Seismic Velocity Structure of Plates 282
2.2.1 Temperature Profiles of Oceanic Plates, and the "Thermal Boundary Layer" 283
2.2.2 Temperature Profiles of Continental Plates 284
2.2.3 Seismic Velocity Structure of Plates 284
2.3 Stress and Driving Forces on the Plates 285
3. **Rheology of Plates and Underlying Asthenosphere** 285
3.1 Rheology of Plates 285
3.1.1 Elastic Behavior of Plates 285
3.1.2 Strength of Plates 286
3.1.3 Viscous Behavior of Zones within the Plates 286
3.2 Rebound from Loads, and Flow beneath the Plates 286
4. **Relationship of Plate Motions to Processes in the Earth's Mantle and Core** 288
4.1 Interaction between Plates and Mantle Convection 288
4.1.1 Present Day 288
4.1.2 Plate Tectonics during Early Earth History 288
4.2 Hotspot Volcanism and the Record of Plate Motions Relative to the Hotspot Source Regions 289
4.3 Connections between Core Processes and Plate Motions 290
5. **Limitations of Plate Tectonic Theory** 290
5.1 Vertical Motions 290
5.2 Deformation in Plate Interiors 291
5.3 Complexities of Deformation at Plate Boundaries 291
5.4 Limitations on the Sizes of Plates 292
5.5 Application to Other Planets 292
6. **Summary** 293
Glossary 293
Works Cited 294
Further Reading 294

INTRODUCTION

Earthquakes, volcanoes, active faults, and mountains tend to occur in long, linear belts on the Earth's surface, bounding large regions that are relatively quiescent. The theory of plate tectonics views the quiescent regions as rigid "plates" and these deformational features, located at the boundaries between the plates, as the inevitable consequences of relative motion among the plates. Since the plates are confined to the surface of a spherical Earth, their relative motion is geometrically constrained. The great power of plate-tectonic theory lies in its ability to

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

provide testable kinematic predictions regarding the expected direction and rate of motion at the boundaries between the plates. In addition, it provides a framework for interpreting past geological events in terms of plate-tectonic processes observable today.

Section 1 of this article describes the geometrical aspects of motion of rigid plates on a sphere, and presents rates and directions of motion between major plates, as determined from the relevant data. Section 2 describes the differences between continental and oceanic parts of plates, constraints on plate thickness, and the vertical variations of temperature and strength within the plates. Section 3 describes the elastic behavior of plates, isostasy, and rebound due to viscous flow beneath the plates. Section 4 discusses plate motions as an indication of processes at greater depths in the earth. Finally, Sec. 5 discusses some of the limitations of plate-tectonic theory.

1. GENERAL THEORY OF PLATE TECTONICS

1.1 History

Plate tectonics is the unifying theory that the Earth's outermost solid layer is composed of a number of rigid plates, moving relative to one another; all tectonic activity on the Earth's surface results from interactions among these plates. These plates form the outer rigid layer of the Earth, called the "lithosphere," which reaches a maximum thickness of perhaps 100–150 km. The lithosphere overlies a less rigid layer, the "asthenosphere," so that plates can be envisioned as thin spherical caps floating on top of a region of less viscous material. Fourteen major lithospheric plates exist at present (Fig. 1); the number and size of these plates has varied with time. The relative motion of these plates follows the behavior that might be ex-

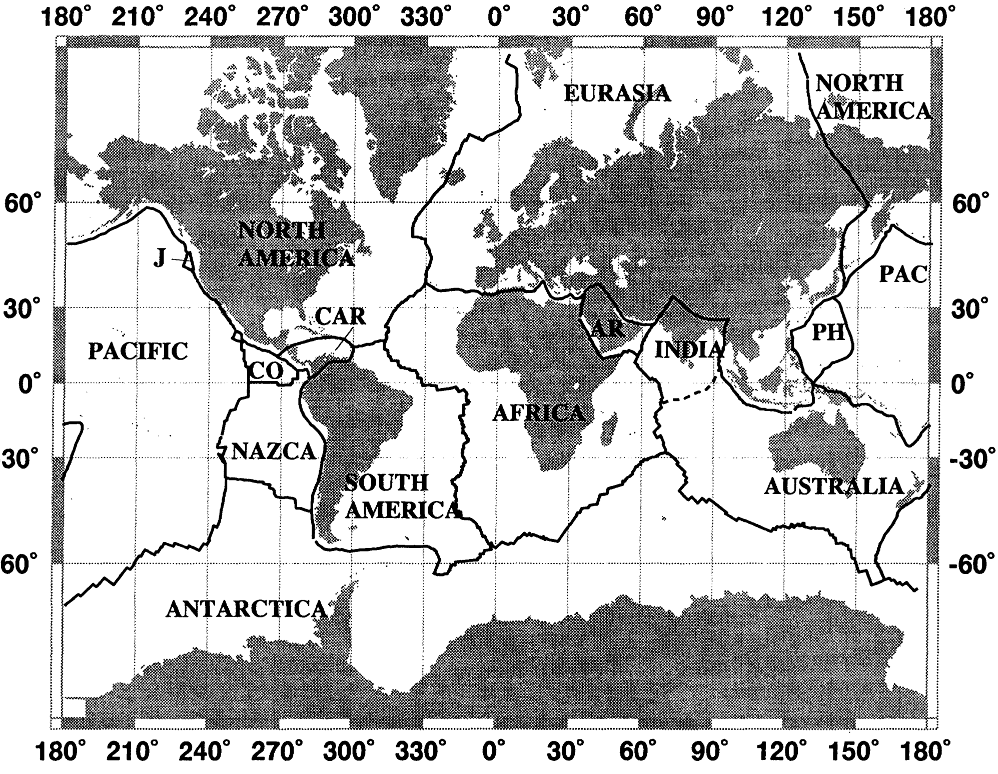

FIG. 1. Map of the Earth's major plates. AR = Arabia, CO = Cocos, J = Juan de Fuca, CAR = Caribbean, PH = Philippines. The India–Australia boundary (dashed) is a zone of diffuse deformation.

pected of rigid caps on the surface of a sphere. Under the theory of plate tectonics, the deformation at the Earth's surface is primarily confined to the boundaries between the plates.

Plate-tectonic theory was first proposed in the mid-1960s on the basis of an accumulated body of geophysical observations collected since the Second World War (see MARINE GEOPHYSICS). These observations included, first, a knowledge of the relief of the ocean floor, including belts of undersea mountains ("midocean ridges") that encircled the globe and corresponded to bands of seismic activity; trenches along continental margins or island arcs; and long undersea escarpments (fracture zones) across which the ridges were offset. Second, bands of positive and negative magnetic anomalies had been found in the oceans, forming symmetric patterns about these ridges; it was discovered that the Earth's magnetic field has changed polarity randomly in the past; and it was discovered that molten rock (lava), as it cools, can acquire a permanent thermal remanent magnetization parallel to the extant direction of the Earth's magnetic field. A key element was the hypothesis of Dietz and Hess that new oceanic crust was forming at the midocean ridges by a process of "seafloor spreading," whereby the two halves of the ocean basin were moving apart, and new molten rock was filling the gap and cooling. Under this theory, the earthquakes occurring along the ridges were due to divergent motions at the plate boundary, and the oceanic material was carried laterally by mantle convection currents until it was thrust back down into the mantle at trenches. Additional support for this model came from the work of Vine and Matthews (1963–1965) who proposed that this new seafloor could then become alternately normally or reversely magnetized according to the direction of the Earth's field at the time the rock was cooling. Continued seafloor spreading would form a linear pattern of magnetic anomalies symmetric about the ridge crest, much as observed. Both the age of the seafloor and the maximum age of the sediments deposited on it should increase with distance from the spreading center; and the magnetic anomaly patterns in the various oceans, once corrected for differences in rate of seafloor spreading, should give a single history of magnetic-field reversals.

The basic framework of plate-tectonics theory was confirmed in 1967–1968 by several lines of observation. Drilling results from the South Atlantic ocean showed that the age of the oldest sediments on the seafloor increased in both directions away from the ridge crest. Observations of the slip directions during earthquakes on the midocean ridge system showed directions of separation and lateral motion consistent with plate-tectonic theory. The age constraints on the seafloor, combined with the pattern of magnetic reversals observed in marine magnetic anomalies, were used to calibrate a "magnetic reversal time scale" with the basic reversal history still in use today. Anomalies numbered from Chron 1, at the present ridge crest, back to Chron 33, 84 Ma ago, correspond to a period of frequent polarity changes (the "Cretaceous-Tertiary-Quaternary Mixed Superchron"); from ~120 to 84 Ma ago was a period of constant magnetic polarity known as the "Cretaceous Normal Superchron"; before 120 Ma ago was an earlier period of mixed-polarity intervals.

The theory of plate tectonics also provided a kinematic framework for explaining earlier geological observations that had prompted Wegener's theory of "continental drift" in the early part of the twentieth century—that is, that continents with apparent similarities of rock deposits, fossils, and structural evolution, were once adjacent and are now widely separated by ocean basins because they have "drifted" apart.

1.2 Types of Plate Boundaries

Plate boundaries are divided into three basic types: divergent, convergent, and transcurrent (or strike-slip) (Fig. 2). The nature of the boundary depends on its orientation with respect to the direction of relative motion of the two plates. Divergent boundaries are midocean ridges, or "spreading centers," where two plates are moving apart. They can also be "back-arc basins," which represent divergent motion between smaller plates near a convergent boundary. In these regions, "seafloor spreading" occurs: upwelling of asthenosphere in the zone between the plates produces volcanism, which creates new seafloor in the zone between the plates, ultimately adding area to the lithospheric plates on each side. (This growth in plate

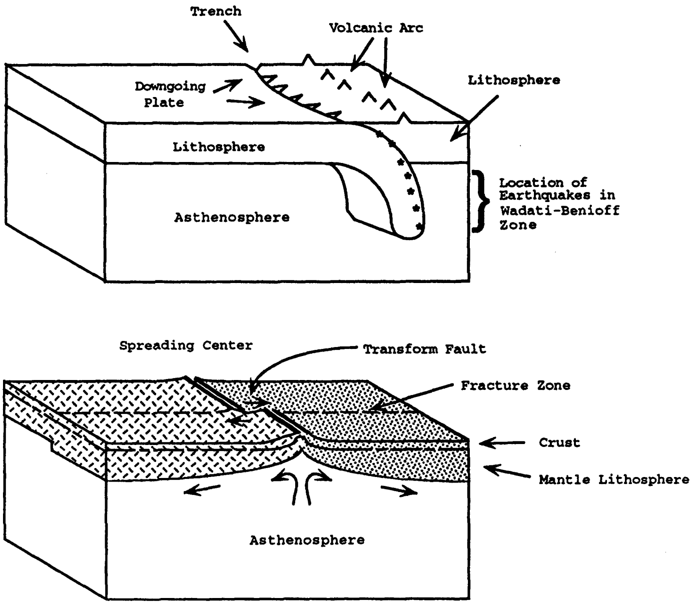

FIG. 2. Block diagrams illustrating the three types of plate boundaries: divergence at oceanic spreading centers (double line), convergence at trenches (line with teeth on upper plate) and earthquakes along the subduction zone, and lateral motion along transform faults (single line). Note that the thickness of the oceanic plate increases with age as a result of cooling and underplating of mantle material to the base of the plate (see Sec. 2.2.1).

area must be compensated by a decrease in plate area at convergent zones elsewhere on Earth, so that a constant surface area is maintained.) Such zones often originate during breakup of continental plates, and are preceded by continental extension. Thus, at very young divergent boundaries such as the Red Sea or the Gulf of California, a narrow rift system containing true seafloor spreading structures is seen to pass gradually into a broader rift system still dominated by continental extensional processes.

At transcurrent boundaries, two plates slide past one another without the creation or destruction of plate area, along fault zones known as "transform faults." Oceanic transform faults are linear bathymetric lows separating the ends of two different segments of spreading ridge and marking a discontinuity in the pattern of seafloor magnetic anomalies. The inactive trace of a transform fault is called a "fracture zone"; the magnetic anomaly discontinuity is still recognizable, and the morphological valley may be still preserved. In continental regions, transform faults may be long fault zones with multiple strands (such as the San Andreas fault system in California or the Alpine fault zone in New Zealand). They may also be topographically distinctive "rift valleys," dominated by lateral motion rather than rifting, such as the Gulf of Aqaba–Dead Sea transform system between the Africa and Arabia plates.

At convergent boundaries, two plates are moving toward one another and plate area is being destroyed. Here, if at least one of the plates is oceanic, an oceanic plate moves down beneath the other plate, into the asthenosphere, in the process of "subduction." Subduction zones are characteristically associated with an inclined zone of seismic activity (within the downgoing plate), a deep oceanic trench at the plate boundary, and volcanism above the zone of inclined seismicity, in the overriding plate. The volcanic region is called an "island arc" (if the overriding plate is oceanic) or an "active continental margin" (if the overriding plate is continental). However, at convergent boundaries involving continental material on both sides, the convergent zone may be much broader and a true subduction zone may be lacking, because continental material is too buoyant to subduct easily (see Sec. 5.3).

Along the length of any plate boundary, there may be a combination of plate-boundary features. Midocean ridges (divergent boundaries) usually comprise a series of spreading segments separated by transform faults. A convergent boundary may have oblique motion (for instance, strike-slip plus convergence). A plate boundary can change type along its length as the direction of relative plate motion changes relative to the orientation of the boundary (Fig. 1). For example, the boundary between the Pacific and North America plates consists of a series of oceanic spreading centers and transform faults (in the Gulf of California), a continental strike-slip fault zone (the San Andreas fault zone) in southern and central California, and an ocean-continent subduction zone in Alaska and the Aleutian Islands.

1.3 Location and Characteristics of Active Plate Boundaries

1.3.1 Earthquakes If plates are rigid, active deformation should be restricted to the zones between the plates. By seeking evidence for active deformation (faulting and folding) from relevant geological and geophysical data, the plate boundaries can be located. Earthquakes, indicating active faulting, are much more common at the edges of plates and along plate boundaries than in the stable interiors of plates. In a map of global shallow seismicity (earthquakes down to 70 km depth), earthquakes are concentrated in bands corresponding to the plate boundaries. Earthquakes at spreading centers are typically in the depth range 0–10 km; earthquakes along transform faults may be in the range 0–30 km; and earthquakes in the depth range from 0 to as much as 670 km may be found in regions of convergence. Intermediate (70–300 km depth) and deep seismicity (300–700 km depth) is generally confined to the interiors of the downgoing plates at subduction zones, and indicates penetration of these "slabs" into the asthenosphere. The zone of seismicity extending downward, from the oceanic trench along the interior of the downgoing plate, is known as the "Wadati–Benioff zone" (see SEISMOLOGY).

The inclinations of Wadati–Benioff zones vary widely, ranging from steep or nearly vertical in much of the western Pacific to only 10°–30° from horizontal at the Nazca–South America boundary beneath the Andes Mountains. The depth extent of the Wadati–Benioff zone can also vary. Where old oceanic plate is being subducted at high rates, the Wadati–Benioff zones may reach 670 km depth. Where very young oceanic plate is being subducted at slow rates, the zone may extend to less than 100 km depth. Some Wadati–Benioff zones have a continuous depth range of seismicity, whereas others have either no deep seismicity, or a gap in seismicity in the depth range 300–500 km. These differences are attributed to the thermal state of the downgoing plate, which is a function of the age of downgoing plate and the rate of subduction. They also seem related to the processes that occur at about 670 km depth in the mantle, which may represent a partial barrier to continued sinking of the slab and the greatest depth at which phase transitions occur in mantle minerals (see Sec. 4.1 and also EARTH, INTERIOR STRUCTURE OF THE).

1.3.2 Magmatism Magmatism includes volcanism (surface eruptions) and plutonism (intrusion and crystallization of magmas at depth). It is characteristically associated with convergent and divergent plate boundaries: 90% of the world's volcanic rocks are produced in these settings. Magmatic processes thus dominate the creation of new oceanic plates and the addition of material to continental plates (see also VOLCANOLOGY).

1.3.2.1 Midocean Ridge Basalts. New seafloor created at midocean ridges represents about 70% of the volume production of global volcanism. This material, midocean ridge basalt (MORB), is very uniform in major-element composition (olivine tholeiitic basalt, with SiO_2 in the range 44%–53% by weight, and Al_2O_3 in the range 12%–16%). Trace elements indicate that MORB is depleted in incompatible elements (those elements that preferentially enter the melt), and isotopic ratios (low $^{87}Sr/^{86}Sr$ and high $^{143}Nd/^{144}Nd$, relative to bulk silicate earth) also indicate depletion relative to the inferred composition of bulk silicate earth.

1.3.2.2 Island-Arc and Convergent-Margin Magmatic Rocks. Magmatism at convergent margins occurs in the upper plate above subduction zones, where the Wadati–Benioff zone is at a depth of 100–125 km. This volcanism dominantly comprises andesitic (SiO_2 55%–63% by weight) stratovolcanoes (often in lines or arcs such as those of Japan, Alaska, or the Andes). A variety of other compositions are present in smaller amounts. The production of magma at this distance from the plate boundary is attributed to upward movement of volatiles released from the downgoing plate by thermally driven dehydration reactions. These volatiles react with the material in the overlying "asthenospheric wedge" and produce melts that rise through the overlying plate, causing surface volcanism and subsurface intrusion of large masses of granitic rocks at depth.

Volcanic rocks in this setting follow a calc-alkaline differentiation trend, with 12%–16% Al_2O_3. Isotopic and trace-element compositions of these rocks are generally enriched in incompatible elements, although with some variations (e.g., low Nb and Ta) that may be due to retention of these elements in the source region, or interactions of the magmas with the continental crust. The $^{87}Sr/^{86}Sr$ ratios are extremely high, because of contamination of low-$^{87}Sr/^{86}Sr$ parental magma with very high-$^{87}Sr/^{86}Sr$ continental crust.

1.3.2.3 Ocean Island Basalts. A third class of volcanic rocks, ocean island basalts (OIB), occurs in a variety of settings not correlated with plate boundaries. These are generally attributed to hotspot volcanism (see Sec. 4.2). They have higher $^{87}Sr/^{86}Sr$, lower $^{143}Nd/^{144}Nd$, and higher radiogenic lead than MORB, and are enriched in incompatible trace elements. They are isotopically variable, requiring a mixture of several end members to explain the range of compositional variation. They are derived from a different, and more highly enriched, mantle reservoir than MORB. One of their end members may be old oceanic crust that has been recycled into the mantle.

1.4 Geometry of Instantaneous Plate Motions on a Sphere

The mathematics describing the motion of rigid plates on a sphere is used to predict the rates and directions of plate-boundary motions. Euler's theorem states that the most general displacement of a sphere with a fixed center is a rotation about an axis through the center of the sphere. Hence the displacement of a rigid spherical cap on the surface of a sphere (i.e., a plate) is also a rotation. Following Euler's theorem, the instantaneous relative motion of any two plates can be described by an angular velocity vector $\boldsymbol{\omega}$ passing through the center of the Earth and projecting through a point on the Earth's surface known as the "Euler pole." The vector $\boldsymbol{\omega}$ is specified as a Cartesian vector or by its components (λ,ϕ,ω) where λ = latitude of the Euler pole, ϕ = longitude of the Euler pole, and ω = rotation rate. The linear velocity $\mathbf{V}$ of plate motion at any point on a plate boundary is related to the angular velocity $\boldsymbol{\omega}$ of the two plates by the equation

$$\mathbf{V} = \boldsymbol{\omega} \times \mathbf{R}, \therefore |V| = \omega R \sin\theta, \qquad (1)$$

where θ is the angular distance from the point of measurement to the Euler pole, and $\mathbf{R}$ is the vector from the center of the Earth to the point of measurement.

The location of the Euler pole is constrained by observations of rate and direction of recent plate motion along the plate boundary. These data are easily obtained for divergent plate boundaries, because the magnetic anomaly pattern constrains the rate of separation of the two plates, and the detailed morphology of transform faults, obtained from seafloor imaging, can constrain their direction of relative motion. Thus, the direction of motion can usually be constrained to within a few degrees of azimuth. The average rate of spreading over the last few million years before present can also usually be well determined.

For convergent boundaries, the primary observations of "instantaneous" plate motions are earthquake focal mechanisms, often of the world's "great" earthquakes which occur at shallow depths (<100 km) in the subduction zone. The amount and direction of displacement in each earthquake can be determined. However, the average rate of plate motion may not be well represented by the observed seismicity, if some of the slip is aseismic, or if the instrumental and historical record of earthquakes is too short to give a valid long-term displacement rate. Finally, if there is deformation of the upper plate, slip in the subduction zone earthquakes will not be parallel to relative plate motion (see Sec. 5.3).

For all these reasons, at convergent boundaries, the time-averaged rate of plate convergence and the direction of plate convergence may be difficult to obtain directly. However, using data from a series of plate boundaries that are mutually interconnected, a solution can be constrained, and the assumption of plate rigidity can be tested. The simplest example of this is the situation at a "triple junction" (the location where three plates meet at a point) (Fig. 3). At this location, plate rigidity requires that the sum of the displacements across two boundaries be consistent with the amount of motion across the third boundary. This can be written in the more general form, for any system of three rigid plates,

$$\boldsymbol{\omega}_{ab} + \boldsymbol{\omega}_{bc} + \boldsymbol{\omega}_{ca} = \boldsymbol{\omega}_{aa} \equiv 0, \tag{2}$$

where $\boldsymbol{\omega}_{ij}$ is the angular velocity of plate i relative to a fixed plate j.

By successively including other plates in the set under consideration, one obtains a set of linear equations that can be solved by linear inversion techniques. Several studies have used the observed rates and directions on all of the major plate boundaries, averaged over the last 3 Ma, to invert for a global set of mutually consistent relative plate velocities (e.g., Minster and Jordan, 1978; DeMets *et al.*, 1990). Such studies give relative angular velocities of the major plates of up to 2° per Ma (or about 0.035 μrad/yr). The corresponding linear velocities at plate boundaries reach up to 16 cm/yr. These results compare well with rates of plate motion estimated over much shorter time scales from satellite geodesy and Very Long Baseline Interferometry (see GEODESY).

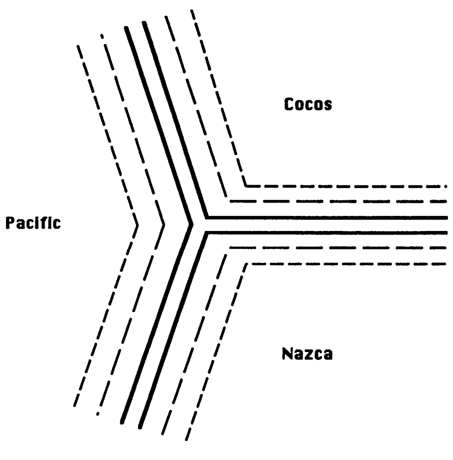

Ridge-Ridge-Ridge Triple Junction

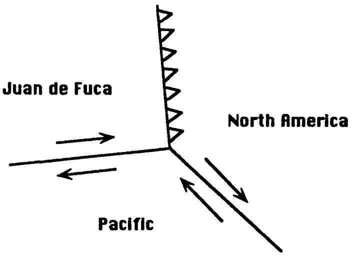

Trench-Fault-Fault Triple Junction

FIG. 3. Two triple junctions in map view. The Pacific, Cocos, and Nazca plates meet at a ridge-ridge-ridge triple junction, with a pattern of symmetric magnetic anomalies shown by the dashed lines. The Juan de Fuca, North America, and Pacific plates meet at a trench-fault-fault triple junction, with the sense of slip along the transform fault shown by arrows.

Some of the plate boundaries with the fastest relative motions at present include the Pacific–Nazca spreading center (with spreading rates from 13 to 16 cm/yr), the Cocos–Pacific spreading center (7–13 cm/yr), and the subduction of the Pacific plate beneath northern Japan at ~10 cm/yr. Boundaries with intermediate rates include the Tonga trench (with subduction of the Pacific plate beneath the Australia plate at 9–6 cm/yr), the Pacific–Antarctic ridge (10–6 cm/yr), the Southeast Indian Ridge between Australia and Antarctica (6–7.5 cm/yr), the subduction of the Nazca plate beneath South America (8–9 cm/yr), and the subduction of the Cocos plate beneath North America (9–6 cm/yr). Well-developed boundaries with a long history of slow motion include the Africa–Antarctica spreading ridge (~1.5 cm/yr), the North America–Africa boundary on the central part of the Mid-Atlantic ridge (2–2.5 cm/yr), and the Africa–South America boundary on the southern part of the Mid-Atlantic ridge (3–4 cm/yr).

1.5 Past Plate Motions

By Euler's theorem, the past relative position of any two plates can be described by a finite rotation of one of them, holding the other plate fixed as a reference frame. Thus, the past relative position of two plates separated by a spreading center can be given as a finite rotation, determined from the best-fit match of the equivalent magnetic anomaly lineations and fracture zones on either side of a spreading center between the two plates. The finite rotation Ω is often parameterized by latitude, longitude of the finite rotation pole, and the angle of rotation, in a fashion analogous to the parameterization of the Euler vector.

The mean age of the Earth's seafloor is, at present, about 100 Ma. The oldest known seafloor, found in the far western Pacific ocean, is about 200 Ma in age. Oceanic crust has presumably been created over many hundreds of millions of years prior to this time in Earth's history, but the older seafloor has subsequently been subducted and returned to the asthenosphere. For times older than those for which seafloor magnetic anomalies are available, the past positions of plates are determined using two alternative but less precise techniques: hotspot tracks (see Sec. 4.2) and paleomagnetic data constraining the past position of the plate relative to an axially dipolar magnetic field (see PALEOMAGNETISM).

Here we give some examples of applications of the finite rotations used in plate reconstructions. The past relative position of the North America plate relative to Africa, for a time 10 Ma in the past, can be described as a rotation, ${}_{\mathrm{NAm}}\Omega_{\mathrm{Af}}$, of 2.41° about a pole at 79°N, 78°E (counterclockwise rotations are positive). If Africa is held fixed and North America is rotated by this amount, it will be closer to Africa after this rotation, and the positions of magnetic anomaly 5 (which is 10 Ma in age) should be brought into coincidence on the two plates. Note that these finite rotations just relate the present positions of the plates to their past relative positions at a particular time; if the plate boundary has experienced changes in direction of relative motion then this finite rotation will not represent a "flow line," or path of actual relative motion, of the two plates.

Plate rigidity requires that, analogous to Eq. (2), a series of finite rotations around a triple junction or closed plate circuit sum to zero, e.g.,

$${}_{\mathrm{Ind}}\Omega_{\mathrm{Ant}}^{10\ \mathrm{Ma}} + {}_{\mathrm{Ant}}\Omega_{\mathrm{Af}}^{10\ \mathrm{Ma}} + {}_{\mathrm{Af}}\Omega_{\mathrm{Ind}}^{10\ \mathrm{Ma}} = 0. \tag{3}$$

If such finite rotations can be independently determined from the data from each plate boundary, then the assumption of plate rigidity can be tested. In addition, series of finite rotations can be combined to yield the past relative motions of pairs of plates for which no direct data are available (e.g., for two plates that are separated by a subduction zone). For example, the following sequence of three rotations, determined from marine magnetic anomalies and fracture zones bounding spreading centers, yields the past relative positions of the converging Nazca and South America plates, at 5 Ma before present:

$${}_{\mathrm{Naz}}\Omega_{\mathrm{Ant}}^{5\ \mathrm{Ma}} + {}_{\mathrm{Ant}}\Omega_{\mathrm{Af}}^{5\ \mathrm{Ma}} + {}_{\mathrm{Af}}\Omega_{\mathrm{SAm}}^{5\ \mathrm{Ma}} = {}_{\mathrm{Naz}}\Omega_{\mathrm{SAm}}^{5\ \mathrm{Ma}}. \tag{4}$$

Information derived from such a plate circuit can be used to estimate past positions of specific geographic features such as volcanic plateaus or fracture zones on the downgoing Nazca plate, or in determining the past rates of convergence through time at this subduction zone for correlation with deformational episodes or volcanic events in the overriding South America plate.

2. COMPOSITION AND STRUCTURE OF PLATES

2.1 Composition

As rigid caps on the surface of the Earth, the plates are rheologically, not compositionally, defined. However, any compositional variations affect their density and rheological structure and the response to stress. The main compositional divisions of the Earth are the crust, mantle, and core (Fig. 4), which are composed of primarily quartz (SiO_2)-rich rocks, primarily olivine ($[Mg,Fe]_2SiO_4$)- and pyroxene ($[Mg,Fe]SiO_3$)-rich rocks, and liquid metal (Fe + Ni), respectively. The base of the plate is below the crust, in the upper mantle (see Sec. 2.2). Thus, all plates contain a major compositional boundary, the crust–mantle boundary or Mohorovičić discontinuity, within them. This boundary also corresponds to a density change and a seismic velocity contrast. Its depth, determined by seismic refraction and/or reflection, ranges from ~5 to 7 km (beneath oceanic regions) to an average of 35 km (beneath stable continental interiors) to >50 km (beneath some high mountain ranges). (See also EARTH, INTERIOR STRUCTURE OF THE.)

From the world plate-boundary map (Fig. 1) it is clear that plates may have varying proportions of continental and oceanic area, ranging from primarily oceanic plates (Cocos, Nazca), through plates composed of

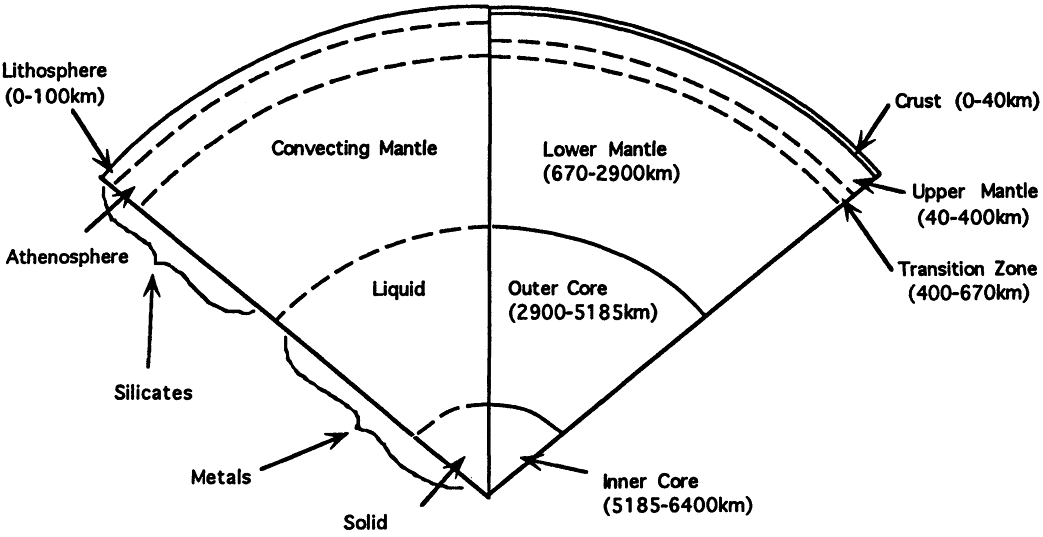

FIG. 4. Main compositional divisions of the Earth (right half of figure) compared to the major rheological divisions of the Earth (left half of figure). It is not known whether convection in the mantle is a single layer (whole mantle) or two layers (upper mantle and lower mantle separately).

both continental and oceanic regions (the majority of plates), to those with virtually no oceanic component, such as Arabia. Because plate material is continually being created and destroyed, the proportions of oceanic and continental material on a given plate can change with time.

In addition, the composition of oceanic crust is very different from continental crust. Thus, regions within plates have different compositional profiles depending on whether they are oceanic or continental (see also STRUCTURAL AND PHYSICAL GEOLOGY). These differences may extend into the mantle.

2.1.1 Oceanic Crust and Mantle An oceanic plate consists of both crust and mantle lithosphere. The crust is formed by magmatic processes at spreading centers (see Sec. 1.3.2) followed by sedimentation and rare additional volcanism at later times after formation. Beneath the oceanic crust, an oceanic plate consists of colder material of the mantle lithosphere (depleted harzburgite, containing olivine and orthopyroxene; and peridotite, which contains these two minerals plus garnet and clinopyroxene). This mantle material accretes or underplates onto the base of the plate as it cools with age and distance from the spreading center.

The magmatic structure of the oceanic crust (Fig. 5) is deduced from a combination of seismic refraction and reflection studies, dredging and drilling, direct observation, gravity and magnetics measurements, and geological observations of old ocean rocks now exposed on land ("ophiolites"). These data lead to models of a 5–7-km-thick oceanic crust with shallow (1–2 km depth) magma chambers, producing basalts that erupt on the seafloor ("pillow basalts") and intrude at shallow levels ("sheeted dikes"), as well as material that crystallizes adjacent to the magma chamber ("massive gabbro") or settles out of the magma chamber by gravity ("layered peridotite"). The magma is formed by partial melting of depleted mantle material (MORB source) at depths of 20–70 km, and then rises to reside in the shallow-level magma chamber. Up to a few kilometers of sediments may accumulate over the igneous crust and form the uppermost layer of the oceanic crustal section.

2.1.2 Continental Crust and Mantle The continental crust is more SiO_2 rich than the oceanic crust, hence less dense. It is usually thicker (30–40 km instead of 5–7 km), and it is strongly enriched in incompatible trace elements. Rather than forming in a single process of seafloor spreading, it forms over millions or billions of years of volcanism, deformation, partial remelting, and tectonic amalgamation. Thus, the continental

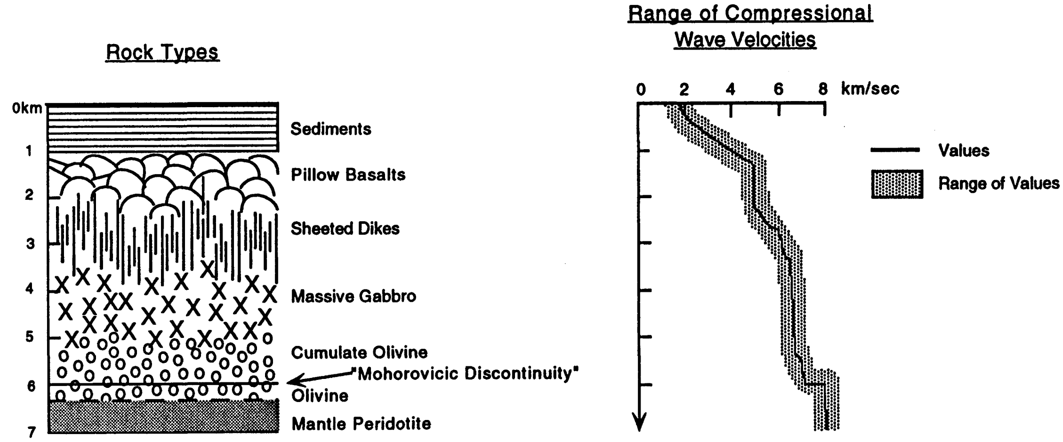

FIG. 5. Oceanic crustal structure, showing rock types inferred from drilling, dredging, and geological study of oceanic rocks preserved on land (ophiolites). Thickness of the sedimentary layer may vary. Boundaries between pillow basalts, sheeted dikes, and massive gabbros are gradational. The crust/mantle boundary (Mohorovičić discontinuity) is chosen on the basis of an increase in compressional seismic wave velocity to greater than 8 km/s.

crust varies in composition, age, and thickness from place to place. Beneath the continental crust we again find mantle lithospheric material (olivine-rich rocks).

The mantle lithosphere of the continental regions may be compositionally slightly heterogeneous compared to the mantle lithosphere of oceanic regions. Thickness of oceanic lithosphere is similar to the thickness of young (<2.5 Ga old) continental lithosphere, but the lithosphere beneath the Archean (>2.5 Ga old) crust seems to be thicker, about 150–200 km thick. These regions of thicker continental lithosphere also are sources for anomalously low-Fe and low-density xenoliths of mantle peridotite, which are brought up as inclusions in volcanic rocks.

2.1.3 Isostasy Continental areas have an average elevation of 0.8 km above sea level, whereas the oceanic regions have an average elevation of 3.7 km below sea level. Gravity measurements indicate that in general these regions are isostatically compensated—positive topography of continents is compensated by an equivalent negative mass anomaly at depth, and negative topography of oceanic regions is compensated by an equivalent positive mass anomaly at depth. The principle of isostasy states that at a certain depth (d_c, the "depth of compensation") the pressures due to the weight of the overlying materials, with surface elevation h, are equal, i.e.,

$$\int_{z=-d_c}^{z=h} \rho(z) g\,dz = \text{constant}, \tag{5}$$

where ρ is density, and g is the acceleration of gravity (Fig. 6). Below d_c the material would be in hydrostatic equilibrium. The depth of compensation appears to vary from place to place. In continental regions, it is often taken to be within the mantle lithosphere, because constraints on crustal thickness are sufficient to yield the observed gravity and topography; in this case, the mechanism of compensation is "Airy compensation"—variations in crustal thickness at constant crustal density. In oceanic regions, d_c is often assumed to lie below the plate, and the mechanism of compensation is "Pratt compensation"—material within the lithospheric column has lateral density variations due to lateral thermal contrasts.

2.2 Thermal and Seismic Velocity Structure of Plates

The "thickness" of the lithosphere is defined in various ways depending on the physics of the problem being considered (e.g., thermally, seismically, or rheologically). Each type of definition leads to different es-

Pratt Isostasy

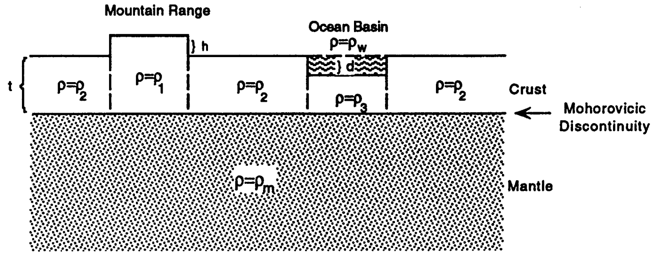

Equation of Isostatic Equilibrium: $g\rho_2 t = g\rho_1(t+h) = g[\rho_3(t-d) + \rho_w d]$

$$\rho_3 > \rho_2 > \rho_1$$

Airy Isostasy

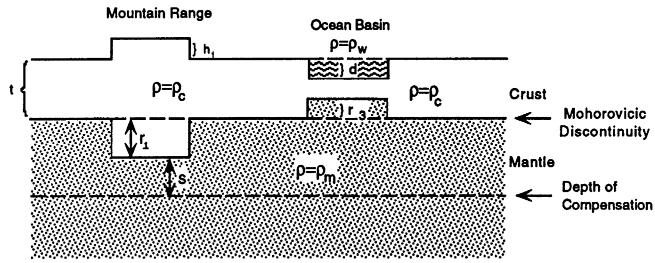

Equation of Isostatic Equilibrium: $g[\rho_c(h_1 + t + r_1) + \rho_m s] = g[\rho_c t + \rho_m(s + r_1)] = g[\rho_c(t - r_3 - d) + \rho_w d + \rho_m(s + r_1 + r_3)]$

FIG. 6. Two mechanisms of isostatic compensation of regions within plates: Pratt compensation (variable crustal density, constant depth to Mohorovičić discontinuity) and Airy compensation (constant crustal density ρ_c, variable depth to Mohorovičić discontinuity). Density of mantle is ρ_m and of water is ρ_w. Depth of ocean basin is d. Equation of isostatic equilibrium is from Eq. (5).

timates for the thickness of the plates. Here we discuss thermal and seismological constraints on plate thickness, which are generally in the range of 100–150 km for old plates. Elastic thicknesses can be much less than this (see Sec. 3.1).

2.2.1 Temperature Profiles of Oceanic Plates, and the "Thermal Boundary Layer" The average depth of mid-ocean ridge spreading centers is about 2500 m below sea level; with distance from the spreading center, the depth of the seafloor increases. In global data, for ages less than about 80 Ma, the depth of the seafloor, h, is observed to be proportional to the square root of the age t of the plate:

$$h\ (\text{km}) = 2.5 + 0.35\sqrt{t\ (\text{Ma})}. \qquad (6)$$

In addition, the surface heat flux, Q_0 is observed to decrease as about $t^{-1/2}$:

$$Q_0(10^{-3}\ \text{W/m}^2) = 473/\sqrt{t\ (\text{Ma})}. \qquad (7)$$

These observations can be explained if the oceanic plate is a rigid thermal boundary layer with no internal heat generation, which thickens during conductive cooling as it moves away from a heat source at the mid-ocean ridges. This thickening process corresponds to underplating of cooled asthenospheric material to the base of the plate. This type of "cooling half-space" model gives functional relations of depth, heat flux, and age in agreement with those observed (Fig. 7). The relationship between depth and square root of age can be obtained theoretically if one assumes isostatic equilibrium at depth beneath the plates; as a plate cools and thickens, its average density increases, and so it sinks further into the asthenosphere below. The variation of the observed depth-age relationships, and ridge-crest depths, among various spreading centers suggests that the temperature variations at depth among the different spreading centers may be on the order of 100 °C.

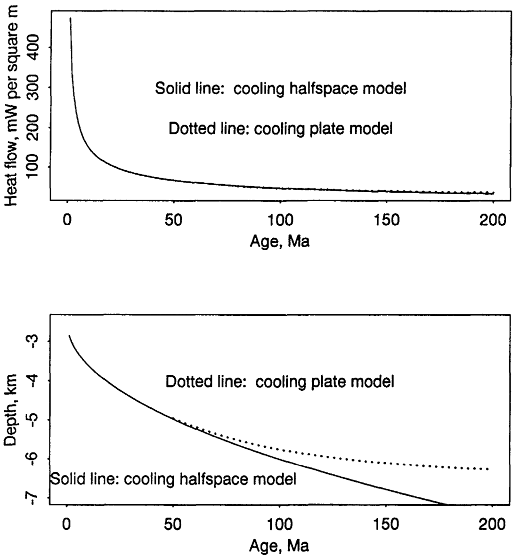

FIG. 7. Theoretical curves relating depth to the seafloor, and oceanic heat flow, to age of the oceanic plate. A "cooling half-space" model fits observations well for young times, but for older times, a "cooling plate model" is required to explain the nearly constant seafloor depth.

For seafloor ages more than about 80 Ma, seafloor depth appears to reach a constant value of about 5.5 km below sea level. This suggests that some dynamics of asthenospheric flow or heat transfer beneath the oceanic plates retard further thickening. Suggested mechanisms include requiring the base of the lithosphere to be isothermal, addition of heat via mantle "plumes," or perturbation of the depths due to pressure differences associated with the asthenospheric flow beneath the plates ("dynamic" effect). These lead to an alternative parameterization of Q_0 and h for $t < 70$ Ma, the "cooling plate model" (with t given in Ma):

$$\begin{aligned} h\,(\mathrm{km}) &= 6.4 - 3.2e^{-t/62.8} \\ Q_0(10^{-3}\,\mathrm{W\,m^{-2}}) &= 33.5 + 67e^{-t/62.8}. \end{aligned} \tag{8}$$

Note that if asthenospheric flow can cause deviations from the expected depth-vs-age curve at old ages, it can also do so at young ages, and may be one major cause of the scatter about the average relationship (residual depth anomalies) present in the ocean basins. The calculations of depth vs age traditionally have not taken this into account.

The half-space model corresponds to an error-function distribution of temperature T with depth z for seafloor of age t:

$$T(z,t) = T_a \mathrm{erf}(z/2\sqrt{\alpha t}), \tag{9}$$

where α, the thermal diffusivity, is typically 10^{-6} m²/s, and T_a, the asthenospheric temperature, is about 1300 °C. This theoretical distribution of temperature with depth in the plate cannot be confirmed by direct measurement because of our present inability to sample more than a few kilometers down into the top of the plates. However, from the surface heat flow Q_0 and measured values of thermal conductivity k one can determine the temperature gradient at the surface:

$$\frac{Q_0}{k} = -\left.\frac{dT}{dz}\right|_{z=0}. \tag{10}$$

This temperature gradient should be approximately linear and about 10 °C/km for old oceanic plates (higher for younger oceanic plates).

The temperature within the plate increases gradually with depth until asthenospheric temperature T_a is reached. Thus, the choice of plate "thickness" as defined by any given isotherm is somewhat arbitrary. It is often chosen to be 90% of T_a. This allows a comparison of the "thickness" of the plate with age.

2.2.2 Temperature Profiles of Continental Plates The structure and thermal profiles of continental plates differ from those of oceanic plates, because of the more complicated process of formation of continental material, and because the upper 15–30 km of continental crust incorporates a significant heat source: radioactive isotopes such as ^{40}K, ^{238}U, and ^{232}Th, which are very rare in oceanic crust. Thus, at a given depth, a continental plate may be warmer than an old oceanic plate. Typical surface thermal gradients for continental regions are 15–25 °C/km.

2.2.3 Seismic Velocity Structure of Plates The thickness of continental plates is not easy to calculate from any simple model, because a significant proportion of the continental material has been present at the surface for billions of years. However, seismic wave velocities (see SEISMOLOGY) indicate that the higher velocities and low seismic attenuation typical of continental and/or oceanic lithosphere extend to about 150–200 km depth beneath the main continental masses. Thus, if the plate thickness is defined seismically, the continental parts of the plates may be somewhat thicker than oceanic parts of the plates, consistent with the greater mean age of the continental lithosphere. In addition, continental regions may have older lithospheric material present beneath them at depth.

2.3 Stress and Driving Forces on the Plates

The major forces driving the plates are thought to be gravitational in nature: sinking of cold, dense plates at trenches (negative buoyancy or "slab pull," perpendicular to the trench) and gravitational sliding of thin, young plates off the thermal high above the upwelling zone at mid-ocean ridges ("ridge push," perpendicular to the ridge). Coupling between the plates and the underlying asthenosphere may cause an effective frictional drag force at the base of the plate.

The absolute levels of stress within the plates are not well known, but the directions of stress can be constrained from observations of borehole ellipticity, hydraulic fractures, overcoring, earthquake focal mechanisms, volcanic alignments, and slip directions of active faults (Zoback, 1992). These data represent stresses at different depths, but generally show good agreement within a given region over distances of hundreds to 5000 km. This broad-scale consistency in stress directions is compatible with a stress field controlled by compressive plate boundary forces (ridge push and continental collision) rather than resistive or driving basal tractions. Some second-order variations in the lithospheric stress field are caused by lithospheric flexure, lateral density contrasts, and lateral strength contrasts (Zoback, 1992).

3. RHEOLOGY OF PLATES AND UNDERLYING ASTHENOSPHERE

3.1 Rheology of Plates

The overall rheology of the Earth is, to first order, that of a fluid body; its gravity field shows that it is close to hydrostatic equilibrium. The plates are the major exception to this observation: they are the most rigid part of the Earth, and least likely to flow on geological time scales. In the theory of plate tectonics, the rigid plates are solid objects "floating" on a more fluid layer (the asthenosphere). Thus, they should be in isostatic equilibrium. Departures from local isostatic equilibrium may occur, if supported by elastic stresses within the plates or local variations of stress within the plates.

3.1.1 Elastic Behavior of Plates Plates respond elastically, by bending, when subjected to applied loads. Such loads include mountain chains, volcanic islands, ice sheets, large lakes, large sediment loads (e.g., sedimentary basins), and changes in volume of water in the ocean basins. From observations of lithospheric loading near trenches and near point or line loads (e.g., large volcanic islands such as Hawaii), the flexural rigidity, or an equivalent parameter such as the effective elastic thickness of the plate, can be determined.

For a line load such as the Hawaiian volcanic chain, the general equation governing the equilibrium deflection of an elastic plate as a function of horizontal distance x from the center of the load is

$$D\frac{d^4w}{dx^4} = V(x) - H\frac{d^2w}{dx^2}, \tag{11}$$

where $w(x)$ is the vertical deflection of the plate, V is the vertical force per unit length applied to the plate, H is a constant horizontal force per unit length applied to the plate, and D is the flexural rigidity of the plate, given by

$$D = \frac{Eh^3}{12(1 - \nu^2)}, \tag{12}$$

where E is Young's modulus, ν is Poisson's ratio, and h is the effective elastic thickness.

For a plate being bent at a subduction zone, or in a zone of continental collision (e.g., the India plate being subducted beneath Asia along the Himalayas, the Pacific plate being subducted beneath the Philippine plate at the Mariana trench), an equation for a plate with an end load and a horizontal bending moment is more appropriate in order to calculate the deflection. Such analyses yield values of effective elastic thicknesses ranging from 24 to 28 km for oceanic plates and 80 to 100 km for continental plates.

In some cases, the flexure induced by a known load is only poorly fitted by using an elastic model, and an elastic–perfectly plastic rheology may be more appropriate. A material with this rheology behaves elastically until it reaches a critical shear stress, at which it suffers permanent deformation.

3.1.2 Strength of Plates The fundamental characteristic of lithosphere is its great strength (rigidity), compared to the underlying asthenospheric material. In the simplest sense, then, the base of the plate could be defined as a particular low value of material strength. Assuming simple compositional layering (crust/mantle) and the temperature gradients within the plates suggests that brittle failure controls the strength near the surface, and temperature-dependent, grain-size–dependent, and strain-rate–dependent processes of crystal plastic flow control the strength at depth. Thus, strength increases with depth in the brittle layer and then decreases with depth in the ductile regime (Fig. 8). The position of greatest strength should lie at the brittle/ductile transition, somewhere within the mantle lithosphere; the base of the plate (as defined rheologically) will be at some greater depth. The definition of the base of the plate in this case can be chosen to correspond to material that will deform at a certain strain rate when subjected to a certain stress level (whatever values of strain rate and stress are thought to be appropriate to the problem under consideration).

The base of the plate does not correspond to a compositional change—the material both above and below it is olivine-rich mantle material in both cases. It is a Maxwell material and can exhibit both elastic and viscous behavior during deformation. Its overall level of shear stress, σ, is a combination of elastic stress and viscous stress:

$$\sigma = E\epsilon + \mu\dot{\epsilon}, \qquad (13)$$

where σ = shear stress, ϵ = strain, $\dot{\epsilon}$ = strain rate, μ = viscosity, and E = Young's modulus. If the ratio $E\epsilon/\mu\dot{\epsilon}$ is large, then elastic stresses dominate and the material can be considered to behave rigidly. If this ratio is $\leqslant 1$, then viscous stresses dominate and the material suffers permanent deformation, i.e., is not rigid. Thus, the time scale of observation (ratio $\epsilon/\dot{\epsilon}$), and the values of the material constants E and μ, influence whether the plate can be considered rigid or not. For most applications in plate tectonics, the time scales of interest are the time scales of large-scale motions of the plates, e.g., 10^7–10^8 years.

3.1.3 Viscous Behavior of Zones within the Plates From the expected temperature gradients within the plates, and the results of laboratory experiments on the mechanical behavior of the dominant minerals in the continental crust (quartz) and in the mantle lithosphere (olivine), we can calculate how the strength of a plate (that is, the maximum shear stress that can be supported by that plate) varies with depth within the plate. Simple models with a quartz-rich crust and olivine mantle lithosphere suggest that there is often a weak zone below ~20 km depth in the lower continental crust, because at the temperature and pressure conditions expected there, quartz can flow viscously. This weak zone should not exist in oceanic lithosphere, because oceanic crust is much thinner than the continental crust, and at equivalent pressure and temperature conditions the olivine of the oceanic mantle lithosphere is much stronger than quartz (Fig. 8). This lower crustal layer may flow laterally in response to loads induced by tectonic activity. In regions such as the Basin and Range province of western North America, where rocks from the middle crust have been brought to the surface by faulting, and yet the crust–mantle boundary appears to be at a constant depth, lateral flow of lower crustal material is invoked as a possible explanation for these observations.

3.2 Rebound from Loads, and Flow beneath the Plates

If plates and the underlying mantle asthenosphere were always elastic in their behavior, or if mantle asthenosphere had very low viscosity, then the removal of a load would cause immediate uplift, a complete and instantaneous elastic adjustment of the shape of the plate. However, geological observations indicate that often the uplift con-

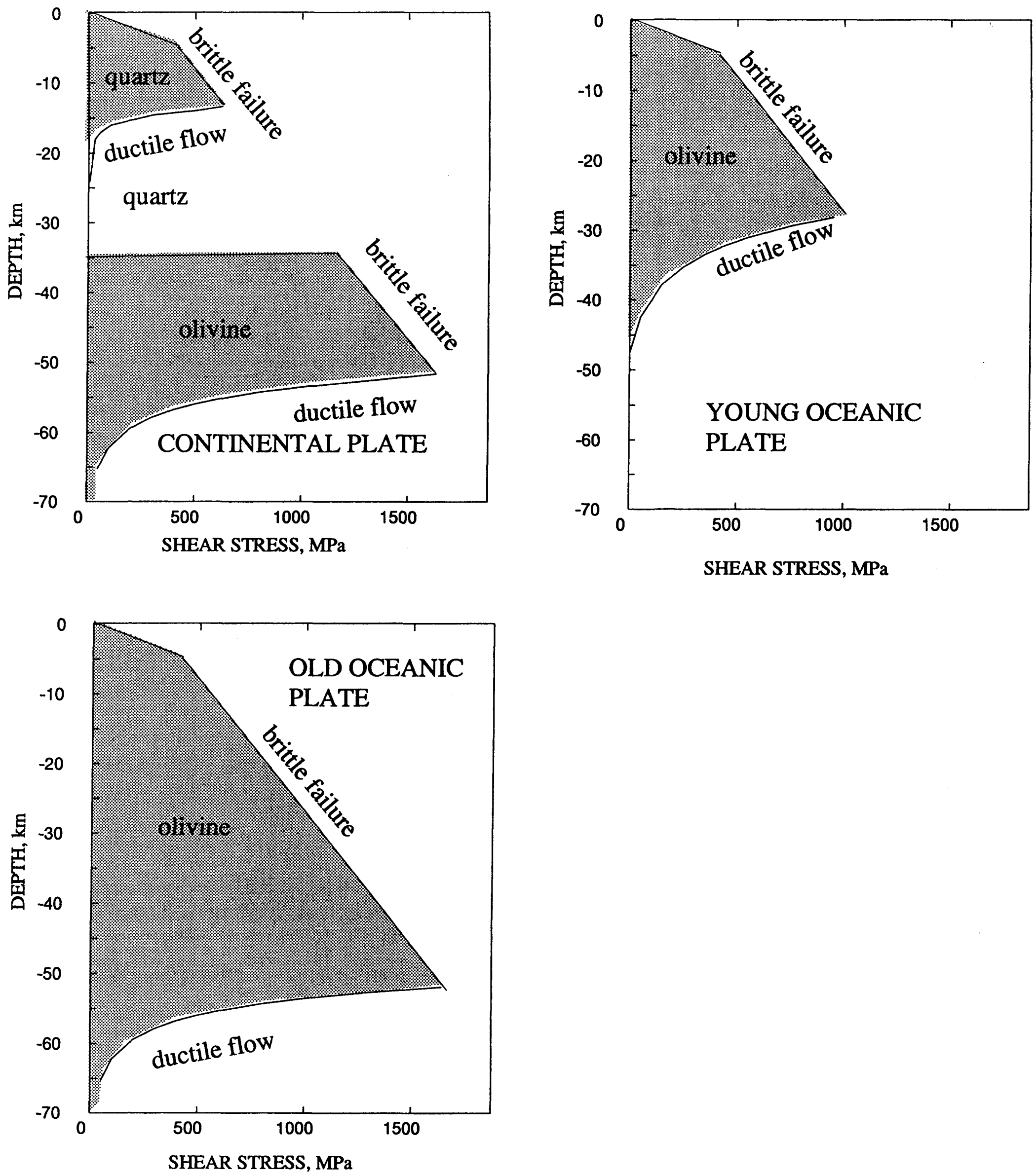

FIG. 8. Theoretical strength of plates as a function of dominant mineral (quartz or olivine) and depth. Ductile strength decreases with increasing temperature; thus, old oceanic plates can support higher shear stress than young oceanic plates because of their lower geothermal gradients. Many continental plates are in regions of low geothermal gradients but may have a weak zone in the lower crust if it is quartz rich, because quartz can flow under lower pressures and temperatures than can olivine. The brittle–ductile transition is formed by the intersection of the brittle failure and ductile flow lines. These curves move up for higher strain rates and down for lower strain rates; a strain rate of 10^{-15} was used here.

tinues after the load is removed. This time-varying behavior is attributed to viscous flow beneath the plates. Postglacial rebound is one such process; although the last Fennoscandian ice sheet melted 10 000 years ago, the crust in Fennoscandia is still slowly rising relative to sea level. Maximum uplift rates at present exceed 0.9 cm/yr. A similar rebound process has been observed, on a smaller scale, in the draining of Lake Bonneville, a large postglacial lake (ancestral to the Great Salt Lake of Utah, U.S.A.). The rebound is attributed to viscous flow in the asthenosphere, below the lithospheric plates, and indicates mantle (asthenospheric) viscosity values of about 10^{19}–10^{20} Pa s.

4. RELATIONSHIP OF PLATE MOTIONS TO PROCESSES IN THE EARTH'S MANTLE AND CORE

4.1 Interaction between Plates and Mantle Convection

4.1.1 Present Day The plates are an integral part of the Earth's convective system, because they are the rigid layer through which convectively transported heat is conducted out of the mantle, and because the cold subducted slabs help to cool the mantle. In addition, motion of surface plates seems to require a convecting mantle beneath the plates. If the subducted slabs penetrated the mantle, they would cause downward flow of asthenospheric material, requiring some compensating pattern of upwellings ("return flow") elsewhere in the mantle. The plate motions are believed to reflect, to a considerable degree, the convective motions in the mantle beneath them. They may also affect these convective motions. For instance, a thick continental layer may insulate the mantle material below it, creating a broad temperature anomaly that ultimately might give rise to convective motions which then drive plate motions.

The 3-D state of the mantle convective system is most often inferred from long-wavelength geoid anomalies and from variations in velocity of seismic waves within the mantle, observed with the technique of seismic tomography (see EARTH, INTERIOR STRUCTURE OF THE; SEISMOLOGY). Within the convecting mantle, lateral velocity and density anomalies are thought to be primarily temperature controlled, rather than indicative of compositional variations. For example, a few 100 °C increase in temperature would be expected to cause a detectable (1%) decrease in compressional seismic wave velocity. Thus, the regions of slow and fast seismic velocity anomalies are used to infer the patterns of upwelling and downwelling, respectively, in the mantle. At shallow depths in the mantle (<700 km depth), the slow and fast regions coincide with the spreading ridges and the downgoing slabs; at greater depth, these velocity anomalies and hence the inferred convection patterns vary laterally and are no longer as obviously correlated with the geometry of the surface plates. In particular, the lower seismic velocities beneath the mid-ocean ridges seem to be confined to the top 350–400 km of the mantle. The surface plate motions also exhibit a significant component of toroidal (strike-slip) power, which is not expected in a convecting fluid and does not appear to persist to depth.

The state of mantle convection is still controversial. There may be two separate convecting layers (upper mantle and lower mantle), separated by a barrier at about 670 km depth. Alternatively, there may be a single convective system ("whole-mantle convection"). Evidence for whole-mantle convection includes the observation that lower-mantle material with fast seismic velocities is found below, and apparently continuous with, the subducted slabs, suggesting that the slabs pass through 670 km depth and continue aseismically into the lower mantle. Evidence for a boundary at 670 km depth includes the cessation of Wadati–Benioff zone seismicity at this depth; cosmochemical evidence that "bulk" mantle is likely to have a different major-element composition than the olivine-rich compositions thought to characterize the upper mantle; and evidence, from the sign of geoid anomalies associated with subduction zones, that there is a substantial (factor of 100) viscosity increase at about this depth in the mantle. The isotopic differences between mid-ocean ridge and ocean island lavas (Sec. 1.3.2) suggest the presence of distinct geochemical reservoirs in the mantle.

4.1.2 Plate Tectonics during Early Earth History The time range over which plate tectonic processes have operated on Earth is important as a constraint on the history of the convection system. Unfortunately, the record of plate motions is less well defined as one goes back in time, because of the lack of remaining seafloor greater than 200 Ma in age. Paleomagnetic data on past continental positions also become more scarce for older times. The geologic time of particular concern, the Precambrian, comprises the Proterozoic (from about

2.5 Ga to 600 Ma ago) and the Archean (older than 2.5 Ga). During early Archean time, heat production in the Earth was considerably higher than at present, possibly resulting in more rapid convective motions and thinner plates than those present now. In addition, it has been proposed by some authors that the early lithosphere may have been too hot to be subducted, so that surface plate motions had a fundamentally different form involving buckling and shearing of small surface plates but no subduction.

Although direct data on the existence of lithospheric plates and on relative plate motions are lacking for these times, geologic data suggest that at least during the Proterozoic time, many of the tectonic settings were similar to the modern ones: arc systems, oceanic ridges, continental cratons, and continental rifts. In addition, paleomagnetic data from Proterozoic rocks suggest the existence of several supercontinents in relative motion. For Archean time, there are fewer similarities with modern settings. Cratonic assemblages and arc assemblages can still be identified but rock types that might represent oceanic rifts or continental rifts have not been recognized in rocks older than about 2 Ga. If subduction was occurring as far back as 2 Ga, which would be consistent with the observations, then another driving mechanism must be sought, because density contrasts of subducted slabs would probably have been insufficient to drive subduction. Many Archean orogenic belts contain a high percentage of komatiite lavas, which are more dense than basaltic lavas and were presumably derived from a less depleted upper mantle with a high percentage of partial melting. If Archean oceanic crust were composed of komatiite, the extra density may have been sufficient to drive subduction and plate tectonics starting before 4 Ga ago.

4.2 Hotspot Volcanism and the Record of Plate Motions Relative to the Hotspot Source Regions

One principal component of the convection system, mantle "hotspots" (sources of anomalous volcanism, often distant from plate boundaries), appears somewhat decoupled from surface plate motions. The best-defined "hotspot tracks" are lines of chemically distinctive, OIB composition (see Sec. 1.3.2.3), volcanoes on the seafloor (Fig. 9). These show an age progression along the chain (e.g., the Hawaiian-Emperor seamount chain in the central Pacific Ocean, formed by the Hawaiian hotspot). Other major hotspots include Iceland, associated with the Iceland–Faeroes ridge; Tristan da Cunha, associated with the Walvis Ridge and Rio Grande Rise in the south Atlantic Ocean; Reunion, associated with the Chagos–Laccadive Ridge in the Indian Ocean; and Kerguelen, associated with the Kerguelen Plateau and the 90°E ridge in the Indian Ocean. Other oceanic ridges or alignments of seamounts, such as the Louisville Ridge in the SW Pacific Ocean, are considered to be hotspot tracks although there is no evidence of modern volcanism or a presently active hotspot. The tracks of continental hotspots may be difficult to identify, because significant interaction of the hotspot magmas and the continental plate complicates both the geometry of the volcanic trace and the compositions of the volcanic products. Interpretations of continental hotspots vary, but Yellowstone (in North America) is one of the most active and well known.

The hotspots appear to originate from sources that are well beneath the plates, from the lower mantle or from the core–mantle boundary. They are generally located away from subduction zones. Their sources appear to lie in a reference frame with very slow relative motions, compared to plate motions. The age progression of the hotspot volcanoes on a single plate indicates the relative motion of the plate over the hotspot source. If the hotspot sources are assumed to be fixed to one another, then the hotspot tracks can be used to define the relative motion of the plate they lie on, relative to the "hotspot reference frame" (sometimes also considered to be an "absolute reference frame"). This provides an alternative method of determining past relative plate positions, e.g.,

$$ {}_{\mathrm{NAm}}\Omega_{\mathrm{hotspots}} + {}_{\mathrm{hotspots}}\Omega_{\mathrm{Pac}} = {}_{\mathrm{NAm}}\Omega_{\mathrm{Pac}}. \qquad (14) $$

There is not yet a consensus on the depth at which hotspots originate. There is also no known theoretical reason why the hotspot sources must be fixed to each other. Until further work is done on these topics and on detailed hotspot and plate kinematics, the validity of the hotspot reference frame cannot be unequivocally established.

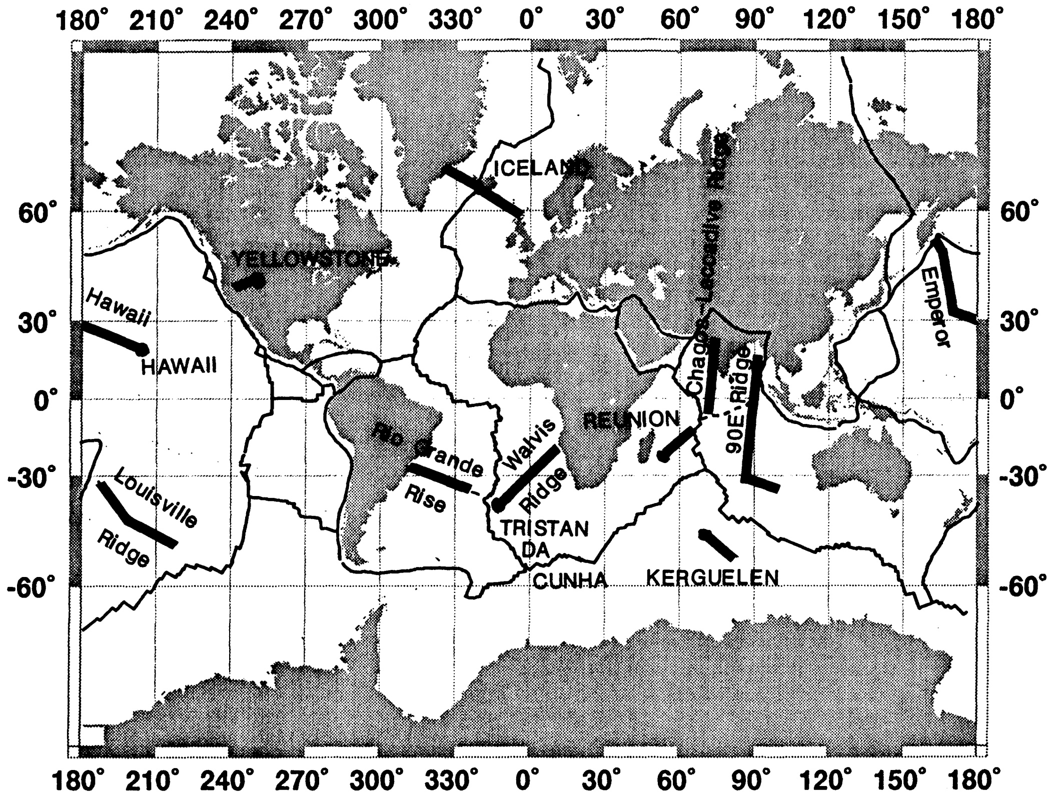
180° 210° 240° 270° 300° 330° 0° 30° 60° 90° 120° 150° 180°
60°
30°
0°
-30°
-60°
ICELAND
YELLOWSTONE
Hawaii
HAWAII
Louisville Ridge
Rio Grande Rise
Walvis Ridge
TRISTAN DA CUNHA
REUNION
KERGUELEN
Chagos–Laccadive Ridge
90E Ridge
Emperor

FIG. 9. Principal hotspot tracks in the world's ocean basins.

4.3 Connections between Core Processes and Plate Motions

A possible link between core processes and surface tectonics has been recently suggested, on the basis of the correlation between the late Cretaceous interval of constant polarity of the magnetic field, and unusually fast seafloor spreading. The frequency of magnetic reversals may directly indicate the state of convection in the Earth's outer core, although it is not known whether periods of constant magnetic field polarity correspond to increased convective activity or decreased convective activity in the outer core. The state of convection in the outer core would certainly exert strong control on the heat input to the base of the mantle and eventually to the heat flux through the Earth's surface (as evident from seafloor spreading and intraplate volcanism). These variables may be related either instantaneously or with some time lag, depending on the time scale and processes of heat transfer between the core–mantle boundary and the lithospheric plates.

5. LIMITATIONS OF PLATE TECTONIC THEORY

As mentioned in Sec. 1.1, there are exceptions to the rigid-plate assumptions of plate tectonic theory: deformation within plates, deformation in broad zones adjacent to the boundaries, and slow vertical motions ("epeirogenic movements") of large regions within plates, without associated internal deformation. In addition, smaller pieces of crust sometimes exhibit independent motion, raising the question of the smallest scale at which plate tectonics is applicable.

5.1 Vertical Motions

The predictive value of plate tectonics theory lies in its constraints on the horizontal components of plate motion. It does not address the question of vertical motions, other than via the implied upwellings and downwellings within the asthenosphere beneath the plates (see Sec. 4.1). There are vertical motions of rocks at the Earth's surface, on both short and long time scales, that are not dealt with by plate tectonic theory. Uplift commonly results from crustal and/or mantle lithospheric thickening in zones of convergent motion. The Tibetan Plateau of Asia, and the Altiplano (high plateau) of South America, are both attributed to convergence-related uplift. Uplift can also be due to removal of a load (Sec. 3.2), to heating and thermal expansion of formerly cold lithosphere, or to removal of cold mantle lithosphere from the base of a plate (e.g., by delamination).

Subsidence can be an isostatic response to lithospheric thinning (e.g., in rift basins) or to cooling of the lithosphere, as discussed for the case of cooling oceanic lithosphere (see Sec. 2.2.1). It can also be an isostatic and flexural response to an applied load. In addition, either uplift or subsidence can be caused dynamically, as a result of density differences caused by flow within the sublithospheric mantle.

Eustatic (worldwide) changes in sea level may give the same sedimentological results on the continental margins as do more local uplift or subsidence events. However, such global events are attributed to variations in the volume of water in the ocean basins (due to melting or formation of ice sheets) or to variations in the volume of the ocean basins themselves. This latter type of variation may be strongly controlled by plate tectonic processes. If seafloor spreading has been rapid for a while, and the mean age of the seafloor is young, then the mean depth of the seafloor will also be shallower, displacing some water up onto the continental shelves and causing a global sea-level rise. For example, a time of rapid seafloor spreading in Late Cretaceous time (~120–80 Ma ago) correlates with, and is believed to be the principal cause of, a corresponding period of high sea levels worldwide.

5.2 Deformation in Plate Interiors

Worldwide, most of the release of energy as earthquakes occurs in plate boundary zones, consistent with plate tectonic theory, which suggests that most of the deformation should be concentrated there. Nevertheless, there are rare large earthquakes in the interiors of plates. Such earthquakes presumably relieve accumulated stresses in the plates, but do not account for much relative motion or internal strain of the plates. In addition, broad zones of intraplate seismicity and youthful fault scarps (e.g., the Basin and Range province of the western U.S.A., or the Tibetan Plateau region of Asia) indicate broad internal zones of extensional or contractional strain within the plates. These are usually associated with the presence of hot or weak continental material adjacent to a plate boundary. (See also STRUCTURAL AND PHYSICAL GEOLOGY.)

Some continental plates have active internal rift zones (e.g., the Rio Grande Rift in western North America, the Baikal Rift in Asia, the East African Rift). Although these narrow zones are prominent topographic features, the rate of extension across them is so low that they are considered to represent internal deformation of plates, rather than discrete plate boundaries. If extension rates increase, these may evolve into divergent plate boundaries. Their structural development is considered analogous to the very early development of divergent plate boundaries between continental plates.

Most oceanic plates seem to behave quite rigidly, and are less likely to exhibit significant internal deformation than are the continental plates. Tests of closure of oceanic triple junctions suggest that in most cases the observed rates and directions of motion fit a rigid three-plate scenario quite well. There are two major exceptions to this. Firstly, in oceanic regions where two different oceanic plates are undergoing very slow relative motions, such as along the North America–South America plate boundary in the Atlantic Ocean or along the Australia–India plate boundary in the Indian Ocean, the deformation may occur over a broad zone within the plate. Secondly, very young oceanic plates, which are thin and weak (e.g., the South Gorda region of the Juan de Fuca plate), may exhibit regions of diffuse seismicity and internal deformation manifested by active faulting and folding.

5.3 Complexities of Deformation at Plate Boundaries

Ideally, in plate tectonic theory, all deformation should be concentrated in narrow boundaries between the plates. The boundary zones between oceanic plates are generally narrow (<20 km wide), but they can be much wider in continental plates. For example, the Pacific–North America plate boundary in southern California comprises the San Andreas fault zone (with about 75% of the plate boundary displacement) and other faults in the continental material on both sides, over a zone a few hundred kilometers in width. All of this zone is characterized by high seismicity and complex patterns of active faulting driven by the motions of the more rigid plates on either side. Wide zones of continental deformation also occur in the upper plate in zones of oblique subduction: a zone of strike-slip faulting may develop in the volcanic arc, tens to hundreds of kilometers from the trench, taking up some of the oblique component of plate motion.

Because continental material is more buoyant than oceanic material, continental material does not appear to be easily subducted at subduction zones. At locations where continental material is in the downgoing plate at a subduction zone, there may be substantial geological deformation in a broad area adjacent to the plate boundary. For example, the attempted subduction of continental India plate beneath Eurasia has caused a region of broad deformation in the Himalayas and Tibetan Plateau.

5.4 Limitations on the Sizes of Plates

Although the major plates of the Earth's surface have areas in the range 10^8–10^6 km^2, some smaller oceanic plates (or "microplates") are found along mid-ocean ridges. These are usually bounded by spreading centers and transform faults, and may have dimensions as small as 100×100 km^2. These "microplates" grow in area through seafloor spreading on both sides of them, and they may be short-lived as independent plates, if one of the spreading centers ceases activity and its net motion is transferred to the other spreading center. In this case, the microplate will become accreted to one of the adjacent major oceanic plates. Microplates are also present in subduction settings, where large oceanic plates entering subduction zones have smaller fragments break off their edges, with dimensions of a few hundred kilometers. These microplates may move independently for a while prior to being subducted or they may accrete to one of the nearby major plates.

Continental material is generally not considered to form distinct microplates, perhaps because of its greater tendency to suffer pervasive deformation via geological faulting. In deforming continental regions, individual fault-bounded blocks can have different directions of relative motion; however, these blocks do not appear to extend down through the entire lithosphere, and therefore are not considered to be plates in their own right.

5.5 Application to Other Planets

Plate tectonic processes are a consequence of the formation of a relatively rigid thermal boundary layer above a convection system within the planet. One might expect plate-tectonic processes to occur on any planetary body within which convection is occurring and which has a sufficiently thin and rigid thermal boundary layer (lithosphere). Calculated temporal variations in planetary temperatures suggest that the Moon, Mercury, and Mars would have been too cold to sustain plate tectonics since about 4 Ga ago. However, on Venus, temperatures that might permit plate-tectonic processes could have persisted until much more recently.

Radar images and gravity and altimetry measurements from various spacecraft missions (most recently, the Magellan mission) show evidence for considerable tectonic activity on Venus, including features that may be analogous to transform faults, trenches, zones of continental compression or extension, and possible mantle downwellings. Although there are rift systems hundreds to thousands of kilometers long, analogues to Earth's spreading centers have not been identified. Despite these similarities, the surface of Venus is not broken up into large, relatively rigid plates, separated by narrow deformation zones, as is the Earth. Rather, Venus has broad deformation zones (hundreds of kilometers wide) surrounding smaller rigid blocks (also hundreds of kilometers wide), much more reminiscent of the broad deforming zones in some of Earth's continental regions (Sec. 5.2). Crater density suggests that the average age of plate-boundary–type features on Venus is 0.4 Ga, but no strong age variations are visible. Thus, ongoing lithospheric recycling due to plate-tectonic processes, such as that occurring in the Earth's ocean basins, seems unlikely. It is thought that the lithosphere of Venus is less rigid than that of the Earth, and differently coupled to the underlying convection system. The degree of modern activity of surface tectonic processes on Venus is not well constrained.

6. SUMMARY

The theory of plate tectonics, in which the surface of the Earth is thought to consist of different lithospheric plates in relative motion, works well for explaining many geophysical observations of the Earth. The Earth's outer rheological layer (~100–150 km thick), or lithosphere, behaves rigidly on time scales of 10^7–10^9 years, overlying a much less viscous layer (asthenosphere) that flows on these time scales. The lithosphere comprises both crust and upper-mantle material. It consists of plates whose interiors are rigid. The deformation at the Earth's surface, including earthquakes and volcanism, is concentrated at plate boundaries. The type and depth of earthquakes, and the composition of the volcanic rocks, varies with the type of plate boundary (convergent, divergent, or transcurrent).

The relative motion of these plates must follow Euler's theorem and therefore can be described in terms of rotations about axes through the center of the Earth. Where two plates diverge, the process of seafloor spreading accretes new oceanic plate material to their edges and the plate area increases. When two plates converge, one will sink below the other one, in the process of subduction. Surface area of the downgoing plate decreases and the subducted portion is resorbed into the underlying mantle. A third type of plate boundary (transform fault) is characterized by lateral slip, so that plate area is neither created nor destroyed. Present rates of relative plate motion range from 1 to 16 cm/y. Past relative plate positions are obtained by matching magnetic anomalies generated by seafloor spreading or by using paleomagnetic directions in rocks, and by assuming that hotspots (persistent sources of volcanism which generate linear volcanic traces on the Earth's surface) have sources in a relatively fixed reference frame to which plates can be reconstructed.

Oceanic and continental parts of plates have different properties because of their different processes of formation, crustal thicknesses, and viscosity structures. As oceanic plates cool with time, their strength and thickness increase. They can behave elastically, with effective elastic thicknesses of tens of kilometers, thinner than thicknesses determined by thermal considerations or by seismic wave propagation. Rebound of plates from applied loads involves elasticity of the plate and viscous flow of the underlying asthenosphere.

The lithosphere insulates the hotter convection system in the Earth's mantle. Plate motion is driven by buoyancy forces at plate edges and opposed by basal viscous drag. Upwellings of the convective system in the upper mantle (<670 km depth) coincide with divergent plate boundaries at shallow depths, and downwellings with convergent boundaries (subduction zones), but at greater depths the mantle convective system is not well correlated with plate motions. The source region of hotspots is thought to lie within the mantle or at the core–mantle boundary, well below the plates, distinct from the major convective upwellings and downwellings.

Plate-tectonic theory cannot account for many second-order features of deformations on the Earth's surface. Minor internal deformation of plates occurs in continental regions and in young and weak oceanic plates. Vertical motions may be controlled dynamically by convection in the mantle. Boundaries of continental plates may be broad zones of deformation due to the lower strength of continental crust. Nevertheless, the concept of plate tectonics successfully explains major first-order features of the Earth's surface deformation.

GLOSSARY

Asthenosphere: Part of the Earth's mantle, immediately below the plates, and of much lower viscosity than the lithosphere. This part of the mantle flows, undergoing plastic deformation on the time scale of plate motions.

Continental drift: Relative motion of continents on the Earth's surface, due to seafloor spreading.

Epeirogenic Movements: Broad, gradual uplift of one continental region relative to another (~200 m max).

Euler Pole: Latitude and longitude at which an angular velocity vector (Euler vector) projects through the Earth's surface.

Euler Vector: Angular velocity vector representing the relative motion of two plates on the surface of the Earth. The origin is at the center of the Earth.

Euler's Theorem: Any displacement of a sphere with a fixed center can be described as a rotation.

Fracture zone: Inactive trace of a transform fault; apparently offsets magnetic anomalies in the ocean basins.

Hotspot: Long-lived volcanic source, often not near a plate boundary, which produces an age-progressive line of volcanic centers and is thought to be caused by an upwelling "plume" of deeper mantle material.

Isostasy: Archimedes' principle applied to the lithosphere "floating" on weaker asthenosphere.

Lithosphere: Rigid outer layer of the Earth, comprising the different plates.

Marine Magnetic Anomaly: Difference between the observed and theoretical magnitude of the magnetic field measured over the oceans. These anomalies often occur in stripes parallel to the spreading ridges, corresponding to periods of alternately reversely magnetized and normally magnetized seafloor.

Seafloor Spreading: The process by which new igneous crust is accreted by volcanism at the mid-ocean ridges, while the plates on either side of the ridge move apart.

Spreading Center: Midocean ridge that lies along a plate boundary and has active volcanism and seafloor spreading occurring along it.

Strike-Slip Motion: Slip along a fault in which the relative motion of the blocks on either side is horizontal in the plane of the fault.

Subduction: Process by which an oceanic plate sinks beneath another plate into the mantle and the two plates can converge.

Transcurrent Motion: Strike-slip motion.

Transform Fault: An active strike-slip fault that forms the principal fault along a plate boundary.

Triple Junction: Position where three plates meet.

Volcanic Arc: Series of active volcanoes that form an arc in map view, and lie within the upper plate of a subduction zone.

Wadati–Benioff Zone: Inclined zone of earthquakes at a subduction zone, believed to lie within the subducting slab.

Works Cited

DeMets, C., Gordon, R. G., Argus, D. F., Stein, S. (1990), "Current Plate Motions," *Geophys. J. Int.* **101**, 425–478.

Minster, J. B., Jordan, T. H. (1978), "Present-Day Plate Motions," *J. Geophys. Res.* **83**, 5331–5354.

Zoback, M. L. (1992), "First and Second-Order Patterns of Stress in the Lithosphere," *J. Geophys. Res.* **97**, 11703–11728.

Further Reading

Brown, G. C., Hawkesworth, C. J., Wilson, R. C. L. (Eds.) (1992), *Understanding the Earth*, Cambridge, U.K.: Cambridge University Press. A collection of review articles in the earth sciences, including three discussing aspects of plate tectonics, and related articles on the interior of the Earth, structure of the lithosphere, and magmatism.

Condie, K. C. (1989), *Plate Tectonics and Crustal Evolution*, 3rd ed., New York: Pergamon Press. A textbook reviewing basic aspects of plate tectonics and geophysics of the Earth, which relates these concepts to the physical and chemical evolution of the Earth's crust and mantle.

Cox, A. V. (Ed.) (1973), *Plate Tectonics and Geomagnetic Reversals*, New York: Freeman & Co. Collection of classic articles chronicling the development of ideas on ocean floor morphology, marine magnetic anomalies and the Earth's magnetic field, and the geometry of plate tectonics, including how the plate tectonics concept relates to heat flow, seismicity, gravity anomalies, and geological observations.

Cox, A. V., Hart, B. (1986), *Plate Tectonics: How It Works*, Cambridge, MA: Blackwell Scientific Publications. Good summary of the details of the vector mathematics and rotations needed to describe plate motions on a sphere.

Fowler, C. M. R. (1990), *The Solid Earth: An Introduction to Global Geophysics*, Cambridge, U.K.: Cambridge University Press. A general textbook in geophysics, with a few chapters devoted specifically to plate tectonics and its geological effects.

Kearey, P., Vine, F. J. (1990), *Global Tectonics*, Cambridge, MA: Blackwell Scientific Publications.

An up-to-date textbook summarizing many aspects of plate tectonics.

Moores, E. M. (Ed.) (1990), *Shaping the Earth: Tectonics of Continents and Oceans* (readings from *Scientific American* magazine), New York: Freeman & Co. A collection of *Scientific American* articles including some subtopics within plate tectonics: continental and oceanic crust, hotspots, breakup of continents, and structure of fracture zones.

PLATES AND FILMS, MAGNETIC

E. SCHLOEMANN AND H. JERROLD VAN HOOK,* *Raytheon Research Division, Lexington, Massachusetts, U.S.A.*

	Introduction	297
1.	**Fabrication of Plates and Films of Magnetic Materials**	298
1.1	Metals and Semimetal Fabrication Techniques	298
1.2	Formation of Metal Oxides in Planar Shapes	299
1.2.1	General Remarks	299
1.2.2	Thick-Film Techniques	300
1.2.3	Thin-Film Techniques	300
1.3	Deposition of Magnetic Films	300
2.	**Properties of Magnetic Metal Films**	300
2.1	High-Permeability Thin Films	302
2.2	Permanent-Magnet Film Properties	302
3.	**Properties of Insulating Magnetic Films**	303
3.1	Dielectric Loss Properties	303
3.2	Magnetic Losses in Insulating Thin Films	303
3.3	Other Properties of Insulating Magnetic Films	303
3.4	Properties of Single-Crystal Ferrite Films	303
4.	**Domains and Domain Walls in Magnetic Films**	304
4.1	Formation of Magnetic Domains	304
4.2	Structure of Stationary Domain Walls	306
4.3	Structure of Moving Domain Walls	306
4.4	Bloch Lines, Bloch Points	307
5.	**Applications of Magnetic Films and Planar Structures**	307
5.1	Magnetic Storage Applications	307
5.1.1	Magnetic Storage Techniques Providing Serial Access	307
5.1.1.1	Mobile Storage Medium	307
5.1.1.2	Stationary Storage Medium	308
5.1.2	Magnetic Storage Techniques Providing Random Access	308
5.2	Microwave Applications	308
5.2.1	Linear Devices	308
5.2.1.1	Filters and Delay Lines	308
5.2.1.2	Nonreciprocal Devices	310
5.2.2	Nonlinear Devices	311
	Glossary	313
	Works Cited	314
	Further Reading	315

INTRODUCTION

In many technological applications of magnetic materials, it is advantageous to use them either as a film deposited on a substrate of some other material, or as a thin platelet that can readily be incorporated into a planar device structure. Such magnetic films and platelets are the subject of the present article.

The advantage of magnetic films over bulk magnetic material of the same chemical composition derives primarily from the need to integrate the magnetic devices with other electronic devices, which are typically based on semiconductor technology. A complete "monolithic" integration is very difficult to achieve, but various types of partial or "hybrid" integration have been used successfully.

Magnetic materials can broadly be classified into metallic and nonmetallic materials, the latter being predominantly oxides. Both of these classes can further be subdivided into magnetically "soft" materials and magnetically "hard" materials. The former are characterized by a relatively low coercivity,

---

*Now with Arc Technologies, Inc., Amesbury, Massachusetts, U.S.A.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

i.e., they require only small magnetic fields to become magnetized or demagnetized, whereas the latter have a high coercivity and therefore resist magnetization and demagnetization. Table 1 summarizes the primary applications of the various materials as films and planar structures. It should be understood that the classification scheme described here may be ambiguous in some cases, since materials exist that may lie near the boundaries between the various classes.

Magnetically soft metallic materials are best suited for applications at low frequencies (<1 kHz), where losses due to eddy currents are relatively small. These losses can be further reduced by using multilayer laminates of the magnetic materials, with each layer being electrically insulated from the adjacent layers. Magnetically soft nonmetallic (oxide) materials are primarily used at higher frequencies, often as bulk material, since eddy-current losses are completely eliminated by the high electrical resistivity of these materials. The advantage of thin-film materials of this type lies in their suitability for devices that integrate magnetic and electronic (semiconductor) materials.

Magnetically hard materials, both metallic and nonmetallic, are used extensively in thin-film form as recording media for magnetic recording (Speliotis and Lynch, 1991). Memory devices such as "bubble" memories (Dong *et al.*, 1988), "plated-wire" memories (Luborsky *et al.*, 1974), "magnetoresistive" memories (Pohm *et al.*, 1988, 1989), and "Bloch-line" memories (Heidmann *et al.*, 1988) also use thin-film magnetic materials as essential elements. Microwave components that use high-coercivity ferrites and hence do not require a bias magnet have been demonstrated using bulk materials (Weiss *et al.*, 1989), and efforts are under way to develop similar "self-biased" microwave components using films instead of bulk material (Dorsey *et al.*, 1994).

A pervasive characteristic of nonmetallic magnetic materials is that they usually form as crystalline materials only at relatively high temperatures, typically 800 to 1200 °C. The exceptions to this rule are the oxide magnetite (Fe_3O_4) and related spinel ferrites ($Me_xFe_{3-x}O_4$, with Me a transition metal and $x < 1$), which can be formed at temperatures below 100 °C (Abe and Tamaura, 1984). Magnetite is also known to be formed as a result of biological processes (Blakemore, 1975; see also BIOMAGNETIC AND GEOMAGNETIC DETECTION BY ORGANISMS). Other spinel-type ferrites with more complex compositions are not readily crystallized at low temperatures and, as a result, do not (yet) have all the desired properties (such as low dielectric loss) that would make them attractive for high-frequency applications (Lubitz *et al.*, 1994). Thus, the deposition of ferrite films with acceptable microwave properties still requires a relatively high deposition temperature or annealing at a high temperature. This, unfortunately, presents a major obstacle for the use of certain semiconductor materials (such as GaAs) as substrates for the ferrite film deposition.

1. FABRICATION OF PLATES AND FILMS OF MAGNETIC MATERIALS

1.1 Metals and Semimetal Fabrication Techniques

According to their methods of fabrication, magnetic metals can be placed in one of three categories:

Table 1. Applications of magnetic materials as films and planar structures.

	Metallic	Nonmetallic
Magnetically "soft"	Transformer cores Inductors Electromagnets Electric motors Metal read/write heads for magnetic recording	Nonreciprocal microwave components Microwave filters Microwave phase shifters Magneto-optic devices Ferrite read/write heads for magnetic recording
Magnetically "hard"	Permanent magnets Recording media Memory devices	Permanent magnets Recording media Memory devices "Self-biased" microwave components

1. crystalline metal alloys such as Ni–Fe (Permalloy) and Si–Fe that are typically formed as sheets by standard metallurgical rolling or stamping processes (Cullity, 1972);
2. amorphous metal alloys produced by rapid cooling techniques (Grossinger and Turtelli, 1994); and
3. the permanent-magnet semimetallic compounds such as $SmCo_5$ and $Nd_2Fe_{14}B$, brittle, relatively undeformable materials that require powder processing, magnetic alignment, and sintering to make finished products (Durst and Kronmüller, 1987).

The techniques used to cast and shape sheets and plates of the Fe-based cubic metals have been well documented in the past (Bozorth, 1951; Jacobs, 1969; Parker, 1990) and are discussed only briefly here. The desired end product for high-permeability metals is a uniformly coarse-grained polycrystalline microstructure, free of defects such as pores or secondary phases that would impede the mobility of magnetic domain walls. When grain boundaries are free of secondary phases and stress, the crystallographic discontinuities from one grain to the next are not a significant impediment to domain-wall motion and permeabilities of 40 000 or more are achievable (Prutton, 1964).

The desirable characteristics of permanent-magnet materials are quite the reverse. Here the emphasis in fabrication is on a polycrystalline microstructure that will minimize rather than maximize domain-wall motion. This forces magnetic reversal to take place by a magnetization rotation process within individual grains, which requires greater energy than reversal by domain-wall motion. To achieve this goal, the crystallites must be well oriented and of a size small enough such that each grain can support only a single magnetic domain. This critical grain size is the order of a micron (Cullity, 1972) in the permanent-magnet compositions mentioned earlier.

An oriented fine-grain microstructure is accomplished by a combination of processing steps that produce magnetic alignment of micron-size particles by applying pulsed dc fields of several tesla to the powder during compaction. To keep the compacted mass of oriented particles from separating because of mutual repulsion after this alignment step, a reversal pulse is applied while the compact is within the die. The magnitude of the pulse is adjusted so as to reverse the moment in half the particles without changing the crystallographic orientation, thereby demagnetizing the part. The material is removed from the die and is then sintered to a dense form in an oxygen-free atmosphere. To minimize grain growth and simultaneously achieve a high density, the final firing is often carried out under pressure.

1.2 Formation of Metal Oxides in Planar Shapes

1.2.1 General Remarks The fabrication of magnetic metal oxides, the most important of which are compounds of iron oxide collectively known as ferrites, is very similar to the powder processing–sintering route that was described for permanent-magnet materials. One important difference in the processing of polycrystalline ferrites is the requirement for an oxidizing atmosphere, in contrast to the permanent-magnet alloys, which must be processed in an oxygen-free environment. The brittle nature of oxides again precludes the metallurgical processes of shaping from billets, requiring powder processing and the firing of bulk shapes, followed by abrasive machining and grinding.

Bulk methods are practical for the production of large pieces, but can be expensive and wasteful of material, when thin plates are needed. Alternatively, the ferrite powder can be formed and fired in shapes more closely matched to the final planar dimensions. Using circular steel dies, parts up to about 10 cm in diameter with thicknesses to about 3 mm can be formed in the green (unfired) state. For larger diameters and thinner parts, die forming becomes impractical, both from the viewpoint of handling unfired parts and because of problems with warping and cracking during firing. For ceramic shapes with larger ratios of lateral dimensions to thickness, it is more practical to employ the techniques of ceramic thick-film and thin-film processing. By convention, "thick" films are defined as being 10 μm or thicker in the plate-normal dimension, whereas thin films are limited to a few microns or less.

1.2.2 Thick-Film Techniques Of the thick-film processes, the most important is tape casting, a process developed primarily for multilayer capacitors (Tummala, 1989) and, more recently, multilayer ceramic packages (Schwartz, 1988). Ceramic plates as thin as 0.5 mm and hundreds of square centimeters in area can be produced by this technique. The tape-casting slurry is a suspension of micron-size powder in a viscous fluid mixture of solvents and organic additives such as dispersants and binders that will vaporize and disappear in the final firing. Casting is performed using a "doctor blade," which controls the thickness of the slurry film. The slurry must have sufficient viscosity to maintain its original shape during drying. The partially dried film, known as the "tape," is cut into the required planar shapes and fired, first at moderate temperature to remove volatiles and then at a higher temperature for sintering.

Tape casting is probably the most cost-effective method for producing planar ferrite materials in quantity because of the savings in grinding cost. The process is adaptable to making composite structures as well, by the alternate layering of tapes of differing composition or by punching holes in the tape and inserting plugs of differing composition before final firing.

1.2.3 Thin-Film Techniques The fabrication of magnetic materials in the form of thin films on dense substrates introduces a significant change from the method used for making free-standing metal sheets or bulk ceramics. Processes can no longer be optimized for the synthesis of a single material but must be tailored to a composite of two different compositions; a nonmagnetic substrate and a magnetic film. These two materials must coexist in intimate contact, withstanding temperature cycles and other process conditions without interfacial reactions or stresses that would cause failure or significantly degrade the magnetic properties. On the other hand, there are magnetic properties obtainable in thin films that are unachievable in bulk magnetic materials, as, for example, the uniaxial anisotropy in magnetic garnets that makes bubble memory films possible.

1.3 Deposition of Magnetic Films

The deposition of crystalline metal films from aqueous solution onto metallic or insulating substrates by electroplating is well known (Lowenheim, 1963). The Permalloy compositions have been plated from chloride and sulfate baths to thicknesses of 10–100 μm. Nickel plating forms very uniform dense deposits. The Fe-rich alloys are more difficult to plate, partly because of the tendency of the iron salts to oxidize unless strongly acidic conditions are maintained and partly because the Fe-rich compositions must be deposited at elevated temperatures. This accelerates both corrosion of the plating equipment and the oxidation of iron in solution.

Although Fe and Fe–Ni show useful characteristics for magnetic memory applications, the use of these conductors in rf devices is greatly limited by the eddy-current losses mentioned previously. As a result, there has been a concerted effort in recent years to develop techniques for depositing insulating magnetic films on appropriate substrates for high-frequency uses, particularly in the microwave range. The deposition techniques for ferrite films can, in general, be placed in one of three categories as shown in Table 2. The high-vacuum techniques such as evaporation and molecular-beam epitaxy (MBE) are generally characterized by a low deposition rate. Somewhat higher deposition rates are achieved using the lower-vacuum techniques of sputtering, pulsed laser deposition (PLD), jet vapor deposition (JVD), and metal-organic chemical vapor deposition (MOCVD). The largest rates are feasible using the atmospheric techniques of liquid-phase epitaxy (LPE), atmospheric plasma deposition (Wang *et al.*, 1992), spin-spray deposition (Abe and Tamaura, 1984; Williams *et al.*, 1994; Lubitz *et al.*, 1994), and the various plating techniques. Most of these methods are quite new, and so their relative strengths and weaknesses cannot yet be critically evaluated.

2. PROPERTIES OF MAGNETIC METAL FILMS

Many of the physical deposition techniques such as evaporation, sputtering, and pulsed laser deposition produce metal thin

Table 2. Methods of depositing magnetic thin films.

Class of methods	Individual methods	Typical deposition rate in nm/s
High-vacuum deposition	Evaporation	0.1–100 material dependent
	Molecular-beam epitaxy	0.1–0.5 material dependent
Low-vacuum deposition	Sputtering	0.1–2
	Ion-beam sputtering	0.1–1 material dependent
	Chemical vapor deposition	10–30
	Metal-organic CVD	0.1–1.5 precursor dependent
	Pulsed laser deposition	0.1–20 material/area dependent
	Jet-vapor deposition	50–150
Atmospheric-pressure deposition	Liquid-phase epitaxy	10–50
	Arc plasma spraying	100–2000
	Spin-spray deposition	0.2–30
	Ferrite plating	0.2–30
	Laser-enhanced ferrite plating	2350*
	Electroplating	5–20
	Electroless plating	1–5

*Result for an experimental technique, short deposition time.

films that are amorphous when formed at low substrate temperatures. Rather large internal stresses are typical in the as-deposited films, the magnitude and sign depending on the process used, the rate of deposition, and the relative density of the resultant film. This growth-induced stress is generally the dominant contributor to overall stress. Under extreme conditions, it may be large enough to cause visible distortions in the shape of the film–substrate composite or to produce physical separation at the interface. In other instances, the film may appear normal, but there may be substantial changes in stress-sensitive properties. These stress effects, when undesirable, can sometimes be reduced by thermal treatment, particularly when time and temperature are sufficient for recrystallization to occur.

The primary technique for following the recrystallization process and attendant relaxation of film stress is x-ray diffraction. As temperature is increased, either during deposition or during subsequent annealing, distinct diffraction peaks develop as indicators of the atomic rearrangement process taking place. If the films are dense as deposited, certain properties—for example, saturation magnetization—will remain unchanged with the transformation from the amorphous to crystalline state. Other properties, however, particularly those associated with magnetic loss, are significantly reduced as the recrystallization proceeds.

Another source of stress that occurs in single-crystal films grown on single-crystal substrates is that due to a mismatch in lattice spacing between film and substrate. The mismatch introduces an elastic strain, which can arise during growth or during the subsequent cooling to ambient temperature if there are differences in thermal expansion between film and substrate. Stresses due to thermal-expansion mismatch occur, of course, also in film–substrate pairs where there is no correlation in atomic dimensions between film and substrate. The thin-film geometry implies an intimate and continuous join between dissimilar materials. The linear thermal-expansion coefficient α of the ferrous metal films (α = 11–13 ppm K^{-1}) is larger than that of typical substrates such as Si (α = 4 ppm K^{-1}), GaAs (α = 7 ppm K^{-1}), or alumina (8.5 ppm K^{-1}). Any heat treatment that is intended to relieve stresses introduced during film deposition will necessarily introduce compressive stress in the film during heating. If the strain is relieved at the annealing temperature, the differences in thermal contraction during cooling will introduce strain of the opposite sign (tensile) in the metal film after cooling. Stresses produced in this way interact with the magnetization via the magnetostriction coefficients and can affect properties such as permeability and coercive field (H_c). Note that residual stresses could be annealed with no difficulty in free-standing magnetic materials because there is no substrate to introduce a thermal-expansion mismatch.

2.1 High-Permeability Thin Films

Crystalline thin films of the Ni–Fe and Si–Fe alloys can be made with very high permeabilities. Values approaching those obtainable for free-standing metal bodies are achievable if a uniform composition with grain boundaries free of pores and secondary phases can be produced. As with bulk alloys, the polycrystalline grain size and grain orientation do not materially affect permeability because domain-wall motion is unimpeded by grain boundaries or changes in crystallographic orientation. There is, however, a stress sensitivity in thin films that is more difficult to deal with than is true of free-standing magnetic metals. This arises from the different thermal-expansion characteristics of the film and substrate and the fact that the two, once joined, cannot be separated.

Metal films can be grown epitaxially on single-crystal substrates if there is a close match between lattice spacing of the metal and the periodicity of the substrate lattice in the chosen direction. The growth of pure Fe films epitaxially onto single-crystal GaAs substrates is an example of the type of integration that is possible between magnetic metal films and semiconductor circuits. This growth had first been demonstrated by molecular-beam epitaxy (Prinz and Krebs, 1981) and was repeated using ion-beam sputtering (Tustison *et al.*, 1987). Despite a difference of about 2% in lattice spacing between GaAs (100) and a doubled Fe unit cell, films up to a micron in thickness were grown epitaxially on GaAs. A broadened diffraction peak (Fig. 1) indicated the introduction of some strain.

Metal films lose their usefulness as high-permeability materials at high frequencies, on account of induced eddy currents. Thus, magnetic insulators, such as ferrites, are required when the frequency is sufficiently high. The highest frequency at which metals have useful magnetic properties depends on the film thickness (as well as other parameters), but is generally less than about 1 GHz.

2.2 Permanent-Magnet Film Properties

Thin films of permanent-magnet materials have significantly greater constraints on microstructure and grain orientation than the magnetically soft metals just discussed. As with bulk permanent magnets, the need is to eliminate domain-wall motion entirely because this is the mechanism by which the film can be demagnetized by fields that are significantly smaller than those predicted from intrinsic properties. If domain-wall motion is not allowed, the magnetization reversal must proceed by a magnetization rotation process at a magnetic field strength equal to the anisotropy field of the compound. In the RE–Co alloys, this is the order of several hundred thousand amperes per centimeter. The observed values of H_c are about 10% of this magnitude, suggesting that magnetization reversal by domain-wall motion is only partially controlled by microstructure.

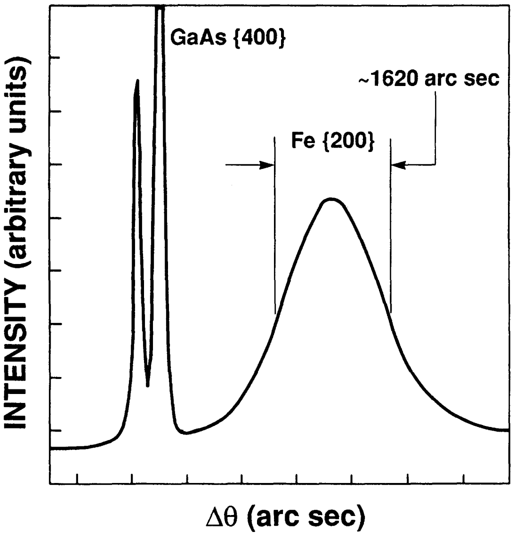

FIG. 1. X-ray diffraction rocking curve for an Fe film on GaAs in the (100) orientation. Film thickness is 570 nm.

In order to attain a high saturation induction $\mu_0 M_s$ (μ_0 = vacuum permeability, M_s = saturation magnetization), the magnetic film must be dense, single phase, and fully grain oriented. The RE–Co compositions have an intrinsic induction of 1.1 T. For the Nd–Fe–B alloy, $\mu_0 M_s = 1.3$ T. If $H_c > 0.5\ M_s$, the energy product, which is a measure of magnet performance, equals $0.25\mu_0 M_s^2$, giving values of 3.0 and 4.1 J cm^{-3}, respectively. Data obtained on sputtered films are somewhat below these maximum values, primarily because crystallite orientation is imperfect and M_s is reduced; the energy product values quoted in recent literature are 2.02 J cm^{-3} for $SmCo_5$ alloys (Cadieu *et al.*, 1990) and 1.16 J cm^{-3} for $Nd_2Fe_{14}B$ alloy (Cadieu *et al.*, 1986).

3. PROPERTIES OF INSULATING MAGNETIC FILMS

3.1 Dielectric Loss Properties

Dielectric loss, represented as the loss tangent $\tan\delta = \epsilon''/\epsilon'$, has no useful function in high-frequency devices and must be minimized to reduce the insertion loss. In polycrystalline ceramics, the spinel ferrites exhibit loss-tangent values in the range of 10^{-3} to 5×10^{-4} with higher losses generally arising from Fe^{+2} in the structure, either stabilized by impurities or introduced by processing. The magnetic garnets generally have smaller values of dielectric loss tangent, 2×10^{-4}, than the spinel ferrites because of a greater thermal stability of the phase and a more limited solubility of Fe^{+2} in garnets. Hexagonal ferrite compositions presently have dielectric losses equal to or somewhat higher than the spinel ferrites.

The dielectric loss in magnetic thin films has not been widely reported. It is inherently more difficult to measure loss properties accurately because of the low filling factor of films compared with the much larger dimensions of bulk ceramic test pieces. There is also a problem in separating the contribution from the substrate, which is considerably more massive than the film.

3.2 Magnetic Losses in Insulating Magnetic Films

An important measure of magnetic loss at microwave frequencies is the width ΔH of the ferromagnetic resonance (FMR) peak, when measured at constant frequency as a function of applied field. For single-crystal films of epitaxially grown yttrium iron garnet, ΔH is of the order of 1 A/cm. For polycrystalline films, ΔH is typically much larger, of the order of 100 A/cm or more, on account of a broadening resulting from the random orientation of crystallites in combination with crystalline anisotropy. This line-broadening effect can be significantly reduced in thin films if grain orientation can be introduced either by choice of single-crystal substrate or by control of the growth conditions, which can introduce a degree of grain orientation.

3.3 Other Properties of Insulating Magnetic Films

Certain intrinsic properties of the ferrite thin films, such as saturation magnetization (M_s) and dielectric constant (ϵ'), are essentially unchanged from their bulk-property values if the deposit is dense and compositionally uniform. The composition dependence of these properties can be obtained from standard handbooks on microwave ferrite materials (von Aulock, 1965; Wohlfarth, 1982, Vol. 2).

Other magnetic properties are dependent on thin-film microstructure and the degree of stress. Such properties are the FMR linewidth, discussed previously, and the hysteresis-loop properties of remanent magnetization (M_r) and, to a lesser extent, coercive field (H_c). In a perfectly grain-oriented, dense, and compositionally uniform thin film, the FMR linewidth would ideally be equal to the single-crystal value. In properly oriented material, M_r would be equal to M_s and would be independent of variations in average grain size. The coercive field, on the other hand, would vary inversely with the average grain size and would also be affected by the magnitude and sign of film stress through the magnetostriction coefficients. The hysteresis properties are of particular importance in microwave applications such as phase shifters and switchable circulators.

3.4 Properties of Single-Crystal Ferrite Films

One very successful film-growth technique has been the epitaxial growth of magnetic garnet films on single-crystal nonmagnetic garnet substrates, notably $Gd_3Ga_5O_{12}$ (gadolinium gallium garnet, GGG). Under growth conditions where there is a perfect lattice-parameter match between the growing film and substrate, as with YIG ($Y_3Fe_5O_{12}$) on GGG, the garnet film remains cubic with properties identical to bulk single-crystal values. However, where a controlled mismatch in lattice parameters would arise on cooling to ambient temperature, the film is placed under stress and a uniaxial anisotropy field develops from the stress working in conjunction with magnetostriction. Additionally, a lattice-parameter mismatch between film and substrate at the growth temperature can introduce cation site ordering, so called growth-induced uniaxial anisotropy, which adds to any stress-induced component. The result is a garnet thin film that has a noncubic characteristic of uniaxial magnetic anisotropy. The planar film strain forces the magnetic domains to lie in the film normal directions, thereby stabilizing magnetic bubble domains. The properties of these garnet films and the effects of induced anisotropy have been well documented by Winkler (1981).

Spinel ferrite films are more difficult to grow than the garnet ferrites. The film surfaces tend to show faceting and there is more flux inclusion. As a result, the magnetic properties of these films have not been well documented. Moreover, an ideal substrate material, such as GGG for garnet film growth, has not yet been found for the spinel ferrites. This has limited the ability of researchers to grow films of reasonable thickness and uniformity for property evaluation.

Hexagonal ferrite thin films have been grown by liquid-phase epitaxy (LPE) in sufficient size and quality on small single-crystal substrates of $ZnGa_2O_4$ to measure FMR linewidth. Values of ΔH as small as 24 A/cm (30 Oe) have been reported (Truedson *et al.*, 1993) along with x-ray diffraction data that confirm the M-type hexaferrite crystal structure. Ba-ferrite films grown by LPE onto basal-plane sapphire (0001) have too large a lattice mismatch for epitaxy but do show highly oriented (0001) polycrystalline deposits. Recent work (Dorsey *et al.*, 1994) on oriented Ba-ferrite film growth on sapphire by PLD (see Table 2) shows that deposition techniques other than LPE are also feasible.

4. DOMAINS AND DOMAIN WALLS IN MAGNETIC FILMS

4.1 Formation of Magnetic Domains

Magnetic thin films, as well as bulk magnetic materials, tend to become "demagnetized" when they are not exposed to a sufficiently strong magnetic field. In the demagnetized state, the saturation magnetization at any point within the material has the same value as in the fully magnetized state (at least to a good approximation), but the direction of the magnetization varies from place to place. The direction is substantially uniform over certain regions, known as "domains," and varies rapidly, but continuously, from any given domain to its neighbors. The borders between domains corresponding to different orientations of the saturation magnetization are known as "domain walls." The thickness of these walls is typically of the order of 0.1 μm, but varies from material to material depending upon parameters such as the anisotropy constant, the exchange constant, and the film thickness. Since the wall thickness is generally much smaller than the domain size, it is frequently appropriate to neglect the thickness altogether and consider the domain wall as a discontinuity of the local orientation of the magnetization vector.

The general principle that determines the domain pattern in demagnetized thin magnetic films and bulk magnetic materials is that all variable parameters of the pattern, such as wall positions and wall orientations, seek those values at which the total energy is a minimum. In the application of this principle to practical materials, the presence of local material imperfections must be taken into account. Since such detailed information is not usually available, it is rarely possible to predict the domain pattern with any accuracy.

An important consideration for thin films is the magnetostatic energy associated with any component of the magnetization vector that is normal to the film plane. This contribution to the total energy frequently dominates all other contributions, particularly for materials with a high saturation magnetization. Because of this magnetostatic energy, the magnetization of thin films tends to lie within the film plane, as shown in the typical domain patterns illustrated in Fig. 2. Here the top diagram corresponds to a material with a uniaxial anisotropy and an in-plane easy axis (Permalloy films are generally of this type). The local magnetization is aligned with this anisotropy axis. The orientation of the domain walls is also parallel to the anisotropy axis, because any other orientation would give rise to a discontinuity of the nor-

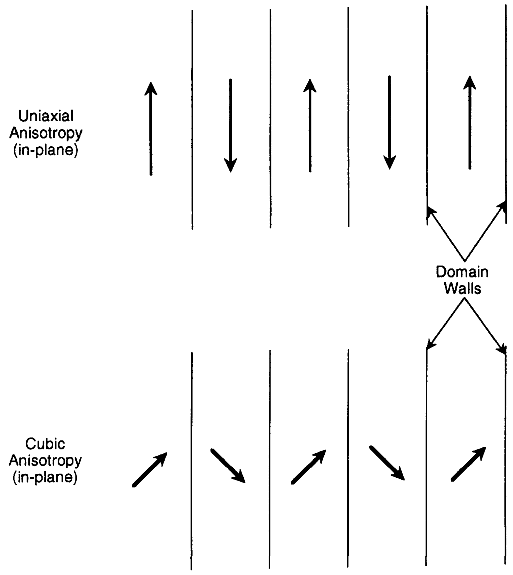

FIG. 2. Typical domain patterns in films having uniaxial (Permalloy) and cubic in-plane anisotropy (Fe on [100] GaAs). Arrows indicate the direction of magnetization.

mal component of the magnetization vector, i.e., to an effective magnetic charge density. Such a charge would raise the total energy. The bottom half of Fig. 2 corresponds to a material with cubic anisotropy, such as Fe on [100] GaAs. In this case, the anisotropy energy is a minimum for four directions of the magnetization, which differ from one another by 90°. In this case, the normal component of the magnetization is again continuous at the domain wall (as in the top illustration of Fig. 2), but the net magnetization in the horizontal direction does not vanish.

The width of the domains in the patterns illustrated in Fig. 2 is also determined by the principle of minimum energy, stated above. The domain width is dependent upon the width of the film in the vertical direction.

A very different domain pattern can be realized with magnetic films having a strong uniaxial anisotropy with an easy axis normal to the film plane. In this case, the local magnetization also tends to be perpendicular to the film plane. Typical domain patterns are illustrated in Fig. 3 for two values of the bias field, which is assumed to be applied along the vertical direction. For a relatively weak bias field, the stripe pattern at the top in Fig. 3 has minimum energy. The width of the domains with magnetization parallel to the bias field is larger than that of domains with the opposite orientation. For a stronger bias field, a pattern of "bubble domains" (as illustrated in the bottom half of Fig. 3) is stable. The bubble domains can be moved about in

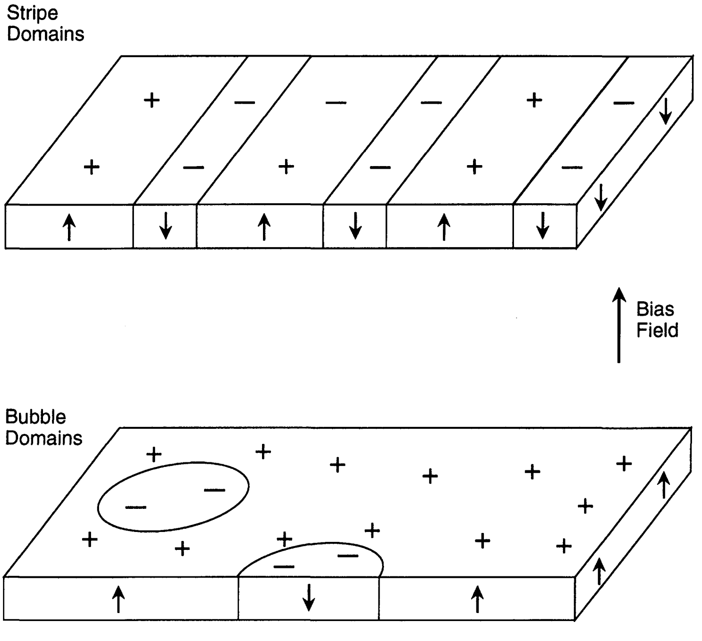

FIG. 3. Typical domain patterns in films having a strong uniaxial anisotropy with easy axis parallel to film normal. Plus and minus signs indicate the polarity of the effective magnetic surface charge.

the film and have been used as the basis of a memory technology.

4.2 Structure of Stationary Domain Walls

In the previous section, the domain wall was treated as a discontinuity in the orientation of magnetization. Consider now the internal structure of the wall itself. For films having an in-plane easy axis, the magnetizations on opposite sides of the wall point in opposite directions as shown in Fig. 4. The magnetization can be continuously taken from one orientation to the other by rotating it either out of the plane of the film (top of Fig. 4) or in the plane of the film (bottom of Fig. 4). These two types of walls are known as "Bloch" and "Néel" walls. Calculations have shown that the Bloch walls are energetically favored for relatively thick films, whereas the Néel walls are energetically favored for very thin films. For intermediate film thicknesses, a more complicated wall structure occurs, which consists of portions of Bloch and Néel walls connected by suitable transition regions known as "cross ties" and "Bloch lines."

For films with a strong uniaxial anisotropy and an easy axis normal to the film plane (the films suitable for bubble memories), the stationary wall structure is illustrated in Fig. 5. In a bulk material, the magnetization would simply be rotated in the plane of the wall as illustrated in the top diagram of Fig. 5. This "untwisted" wall may be used as an approximate model of the actual wall structure, but the model fails to take account of the magnetic surface charge present at the film surface on both sides of the domain wall. These surface charges create a magnetic field that twists the magnetization in the wall interior in the manner indicated in the lower half of Fig. 5. For further details, the reader is referred to the original literature (Slonczewski, 1973; Schloemann, 1973a,b).

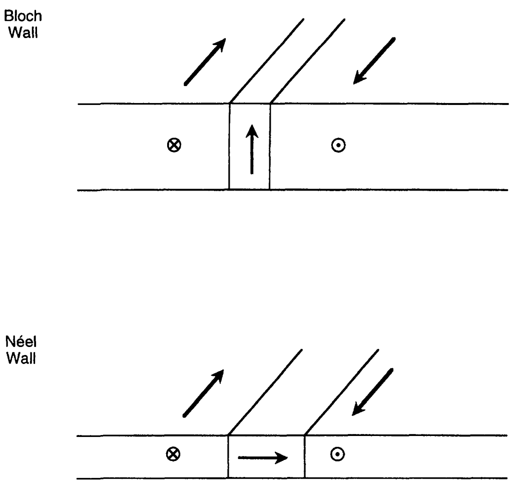

FIG. 4. Domain-wall types for films with an in-plane easy axis.

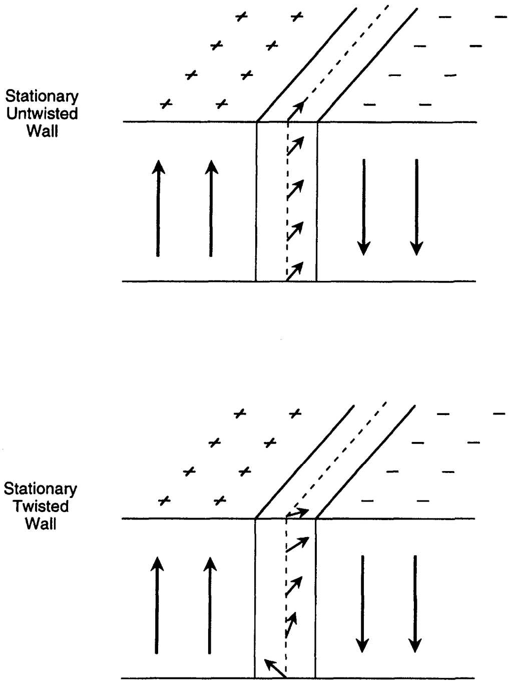

FIG. 5. Structure of untwisted and twisted stationary walls in bubble films. Arrow indicates the direction of the local magnetization vector. Also shown is the sign of the effective magnetic surface charge.

4.3 Structure of Moving Domain Walls

In all preceding discussions, it was assumed that the domain wall is stationary. Moving walls have a similar structure, except that the wall magnetization has a component orthogonal to the domain wall for Bloch walls and orthogonal to the film plane for Néel walls. This component of magnetization is generated by the driving field (applied parallel to the anisotropy axis). The orthogonal component of wall magnetization in turn gives rise to a demagnetizing field in the wall interior that rotates the wall magnetization into the direction of the driving field. For further details, the reader is referred to the original literature (Slonczewski, 1973; Schloemann, 1976).

4.4 Bloch Lines, Bloch Points

The domain-wall structures illustrated in Figs. 4 and 5 represent only one of two possible structures. The alternative structure, not shown in these figures, is the one in which the wall magnetization has the opposite polarity to that shown. It is therefore also possible to have a domain-wall structure in which both types of wall magnetization occur in different parts of the wall. The transition region is known as a "Bloch line." Its width is typically considerably smaller than the wall width. The principal types of Bloch lines are "horizontal" (parallel to the film plane) and "vertical" (orthogonal to the film plane).

Horizontal Bloch lines can be generated when plane domain walls in films suitable for bubble memories are forced to move at high velocity. Similarly, when bubble domains are forced to move at high velocity, a more complicated Bloch-line structure is generated, which consists in part of horizontal Bloch lines and in part of vertical Bloch lines. Bubble domains containing Bloch lines in their domain walls are generally referred to as "hard bubbles." Many different types of hard bubbles exist depending on the type and number of Bloch lines in their domain walls (O'Dell, 1974). Since the hard bubbles are harder to move along a well-defined track (such as used in bubble memories) than regular bubbles, their presence interferes significantly with the desired performance of bubble memories.

Vertical Bloch lines existing in a domain wall can in principle be used for information storage in much the same way as bubble domains, and would have the advantage of much greater storage density.

Bloch lines have a line magnetization analogous to the wall magnetization of the domain walls. Since the line magnetization can have either of two polarities, a composite Bloch line can be formed that consists of sections having opposite polarities. The quasidiscontinuous transition region between the two sections is known as a "Bloch point."

5. APPLICATIONS OF MAGNETIC FILMS AND PLANAR STRUCTURES

The technological applications of magnetic films and planar structures are primarily in the fields of magnetic storage and microwave technology. Since magnetic storage is described in detail in the article MAGNETIC STORAGE, it is covered only very briefly here, with most of the space devoted to microwave applications.

5.1 Magnetic Storage Applications

Data-storage techniques can generally be classified as providing either serial access or random access to the data. In the former case, one can further distinguish between techniques in which the recording medium moves relative to the read/write head and techniques in which the medium and the head are stationary and the stored data move around inside the medium on a well-defined track.

5.1.1 Magnetic Storage Techniques Providing Serial Access

5.1.1.1 Mobile Storage Medium. The storage medium moves relative to the read/write head. This category comprises magnetic recording, which is certainly the most important application of magnetic films. The substrates used for depositing the magnetic layers can be either rigid ("hard disks") or flexible ("floppy disks," tapes). The recording layer itself may consist of small particles of magnetic oxide (such as barium ferrite) in a magnetically inert matrix, or of a continuous layer of high-coercivity magnetic metal. The information is written into the medium by the application of a highly concentrated magnetic field, which may be either parallel or perpendicular to the recording layer. The information is read either inductively (a voltage is induced in a small coil by the motion of the stored bit of information through the sensitive region of the read head), magnetoresistively (the stored bit of information creates a magnetic field that causes a small change in the resistance of the read head), or magneto-optically (sensing the direction of the local magnetization by the Faraday or Kerr effect at optical frequencies).

5.1.1.2 Stationary Storage Medium. Storage medium and read/write head are stationary, stored data move inside medium. Magnetic storage technologies of this type are the "bubble memory" and the "vertical Bloch-line memory." In most (but not all) types of bubble memories, the bubbles propagate around a closed track defined by a thin, patterned layer of Permalloy on the surface of the bubble film. A rotating magnetic field is applied in the plane of the film. This field together with the patterned Permalloy layer creates a series of traveling energy minima for the bubbles on the track. The bubbles therefore propagate around the track in a controlled fashion and can be read magnetoresistively. Bubble memories received much attention in the 1960s and early 1970s, but have found few real applications because of competition from conventional magnetic recording and from semiconductor memories.

In vertical Bloch-line memories, a domain wall serves as a track for propagating vertical Bloch lines (see Sec. 5). At the present time, this type of memory is the subject of considerable academic study, but little practical development. Its advantage over bubble memories is the potential for considerably larger storage density (Konishi, 1983; Wu and Humphey, 1985; Asada *et al.*, 1993).

5.1.2 Magnetic Storage Techniques Providing Random Access The earliest magnetic storage technology providing random access was based on small ferrite toroids ("cores") that were threaded by multiple windings and arranged in a rectangular array. Interrogation occurs by simultaneously sending a current pulse through one vertical and one horizontal wire of this array. The magnetic field generated by a single wire is not large enough to reverse the magnetization in any of the ferrite cores. However, at the intersection of the horizontal and vertical current-carrying wires, the field is sufficient to reverse the magnetization. Thus, the state of the ferrite core at the intersection of the two wires can be set and also sensed by this technique.

A storage technology based on magnetic films and similar to the ferrite-core technology described above is the so-called "plated-wire memory." This technology uses thin CuBe wire (typically 100 or 50 μm in diameter), on which a thin layer of Permalloy ($Ni_{0.8}Fe_{0.2}$) is deposited by electroplating (Luborsky *et al.*, 1974). These wires are placed in a set of parallel grooves and combined with a set of conductive strips that are oriented at right angles to the plated wires. Operation of the plated-wire memory is similar to that of the ferrite-core memory described above. The storage density is, of course, considerably higher. Plated-wire memories were used, for instance, in the Lunar Lander vehicle of Project Apollo, but are now considered obsolete, because they cannot readily be integrated with the semiconductor devices needed for operation of the memory.

In order to provide the high degree of integration with semiconductor devices required for practical high-capacity random-access memories, several techniques are being explored that use magnetic films deposited on a semiconductor substrate. Part of the same substrate can then be used to house the required amplifiers and other devices. Magnetoresistive memories on a silicon substrate with Permalloy storage cells that have a storage density of up to 10^8 bit/cm^2 have been demonstrated (Pohm, 1988, 1989; Ranmuthu *et al.*, 1993). An alternative approach that utilizes films of single-crystal iron on a GaAs substrate has also been described (Schloemann *et al.*, 1988).

5.2 Microwave Applications

The microwave applications of magnetic films may generally be classified as applications involving linear devices, such as filters, delay lines, circulators, and isolators, or nonlinear devices, such as limiters and signal-to-noise enhancers.

5.2.1 Linear Devices

5.2.1.1 Filters and Delay Lines. The use of magnetic films in filters and delay lines generally involves the phenomenon of "ferromagnetic resonance": When a static magnetic field of strength H and an oscillatory field are applied to the material, it exhibits a resonant behavior at a frequency that is approximately given by

$$f_r = \gamma H.$$

Here $\gamma \simeq 2.8$ MHz/Oe $= 3.52$ MHz cm/A is approximately the same for all materials, and is known as the "gyromagnetic ratio." The equation is directly applicable only to spherical samples of the magnetic material and must be modified for film as further detailed below.

Microwave filters based on magnetic materials are generally tunable with the application of a magnetic field. Such filters, based on spherical resonators of YIG, have been used in many practical applications since the late 1950s. More recently, the trend has been toward filters based on magnetic films, because of the ease of fabrication and the greater compatibility with semiconductor components. These filters generally use films of YIG deposited by liquid-phase epitaxy on a substrate of GGG. The technique for growing such films was developed largely as a by-product of the very intensive effort to develop materials for bubble memories, which require similar materials.

For a detailed explanation of filters and delay lines based on magnetic films, the concept of magnetostatic waves (MSW) is important. MSWs are electromagnetic waves in ferromagnetic (or ferrimagnetic) materials, for which the phase velocity is very much smaller than the velocity of light in vacuum. The energy of such waves is almost entirely magnetic in origin. A small electric field is associated with these waves, but it can usually be neglected by way of approximation. The waves can be excited by a current-carrying strip stretched along the surface, and can be received by a similar transducer at a different location on the film surface, as shown schematically in Fig. 6. According to the orientation of the bias magnetic field relative to the film surface and relative to the transmission path, one can distinguish three different MSW propagation modes:

1. Magnetostatic forward volume waves (MSFVW): In this case, the bias field is orthogonal to the film plane.
2. Magnetostatic backward volume waves (MSBVW): In this case, the bias field is in the film plane and parallel to the propagation direction.
3. Magnetostatic surface waves (MSSW): In this case, the bias field is in the film plane and orthogonal to the propagation direction.

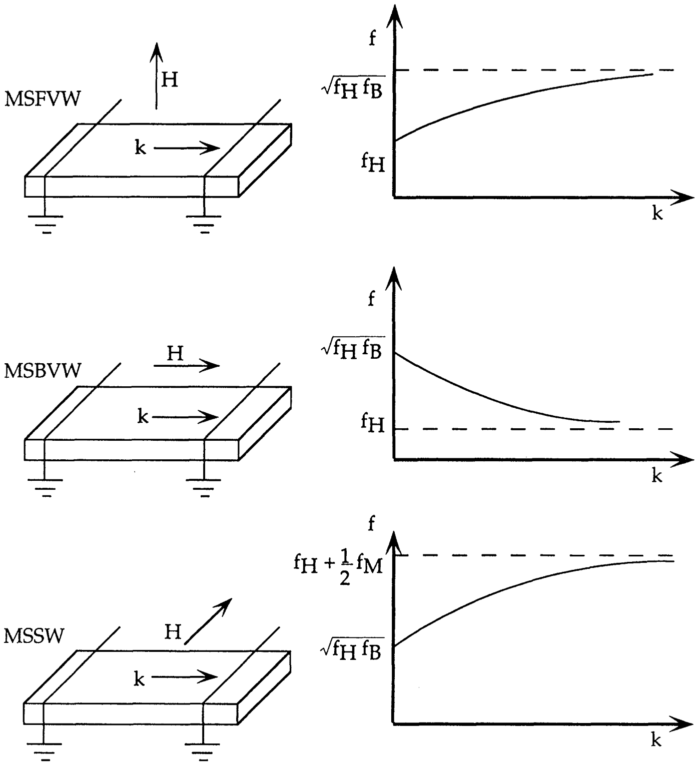

FIG. 6. Principal MSW propagation modes and dispersion diagrams.

A schematic device structure and the corresponding dispersion diagrams showing frequency (*f*) vs wave number (*k*) for these principal propagation modes are illustrated in Fig. 6. As the names suggest, the two types of volume waves exist with equal amplitude throughout the film thickness, whereas the surface-wave type has different amplitudes on the two sides and decays exponentially with distance from the side having the large amplitude. In addition to these principal modes, a large number of mixed modes may exist in which the bias field is oriented at some intermediate angle.

In Fig. 6, $f_H = \gamma H$, $f_B = \gamma(H + M_s)$, $f_M = \gamma M_s$, where γ is the gyromagnetic ratio, H is the internal magnetic field, and M_s is the saturation magnetization (corresponding to $4\pi M_s$ in the cgs system of units). For very thin films, the internal magnetic field equals the externally applied field H_{ext} when the field is in the film plane (cases **2** and **3**, above), and is equal to $H_{ext} - M_s$ when the field is orthogonal to the film plane (Case **1**).

MSWs are generally strongly dispersive, i.e., the phase velocity is a function of wave number or frequency. This feature can be exploited for the design of filters with desirable dispersion characteristics, such as used for pulse compression. The relatively high insertion loss of MSW devices is a significant disadvantage for such applications.

An alternative device structure for magnetic filters based on YIG films is illustrated in Fig. 7 (Ishikawa *et al.*, 1992). This filter uses two YIG films on GGG as separate chips. The two chips are in separate metal enclosures that are connected through a small hole by a copper wire. This device structure is quite similar to YIG filters based on sphere resonators, except that each YIG sphere is replaced by a YIG film. Without further refinement, such filters have a high level of spurious responses. Suppression of these spurious responses has been achieved by creating multiple metal rings on both surfaces of the YIG/GGG chips used in the filter. A design theory based on MSW propagation is used for optimizing the ring configuration with regard to the suppression of spurious responses.

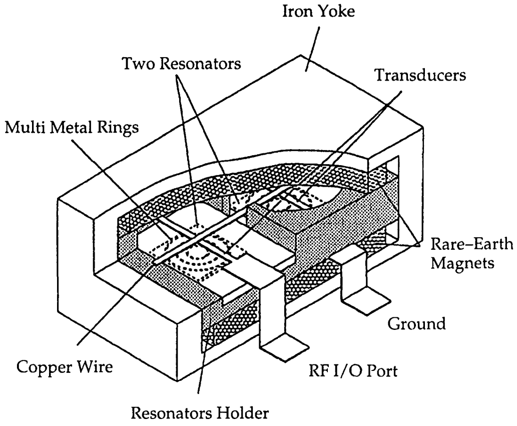

FIG. 7. Construction of the MSW filter (courtesy of Murata Manufacturing Company).

Another application of YIG films on GGG substrates is for filter banks that split an incoming broad-band signal into many narrow-band components at high dynamic range. This work has resulted in a 13-channel magnetostatic-wave filter bank that uses an array of narrow-band MSW delay lines that have a common microstrip input transducer and 13 separate output transducers (Adam *et al.*, 1988). The magnetic field bias is supplied by a permanent magnet with a linear field gradient.

YIG films have also been used to achieve useful tunable band-stop filter characteristics. Such "notch" filters are expected to be used in phased-array antennas as a means of suppressing unwanted signals that might otherwise interfere with the performance of the antenna. A notch filter has been demonstrated that is tunable from 8 to 10 GHz and shows a rejection notch of 40 dB in a bandwidth of over 20 MHz (Daniel and Adam, 1992). This performance is achieved in the $1 \times 1 \times 1\text{-cm}^3$ package illustrated in Fig. 8.

5.2.1.2 Nonreciprocal Devices. The major use of bulk ferrites at microwave frequencies is for nonreciprocal devices, such as circulators and isolators. A monolithic integration of such devices with semiconductor devices such as amplifiers, switches, and phase shifters would be highly advantageous from the point of view of avoiding expensive hand labor and increasing the reproducibility of subsystems, such as phased-array modules (Schloemann, 1988). The use of high-quality ferrite films on semiconductor substrates or on mutually compatible substrates may eventually lead to this kind of monolithic integration. For the nearer term, it appears likely that ferrite nonreciprocal devices will usually

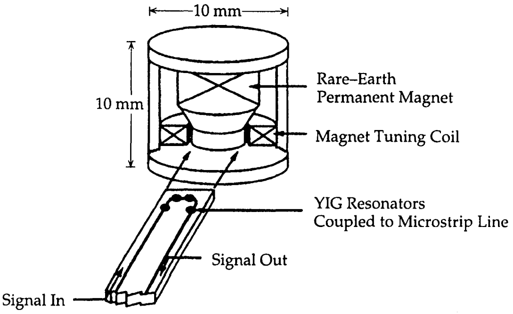
10 mm
10 mm
Rare-Earth Permanent Magnet
Magnet Tuning Coil
YIG Resonators Coupled to Microstrip Line
Signal Out
Signal In

FIG. 8. Schematic drawing of the tunable YIG notch filter and its associated miniature permanent magnet with tuning coil (courtesy of Westinghouse Electric Corp.).

be made as separate components, with a high degree of compatibility with semiconductor technology achieved by the selective use of magnetic films.

5.2.2 Nonlinear Devices For the linear devices discussed in the preceding section, a monochromatic input signal generates a monochromatic response. For nonlinear devices, on the other hand, the response to a monochromatic input signal involves components at frequencies other than the input frequency. These frequencies are usually either multiples of the input frequency ("harmonic generation") or such that the sum of two of the generated frequencies equals the input frequency. Especially important is the case in which the generated frequency equals half the input frequency ("subharmonic generation").

For the devices discussed in the preceding section, the response is, of course, "linear" only at sufficiently low power levels and will inevitably become nonlinear when the power level is raised. Thus, the devices themselves are not linear; they are merely designed to function in the linear regime. By contrast, the devices discussed in the present section are designed to function in the nonlinear regime.

The most important nonlinear microwave device involving magnetic materials is the "limiter." An ideal limiter would transmit a low-level monochromatic signal without attenuation and a high-level monochromatic signal with just enough attenuation to make the output power independent of input power beyond a certain threshold. Actual limiters based on magnetic materials only approximate this ideal behavior. They have a finite insertion loss at low power levels; and the attenuation at high power levels, although much higher than the low-power attenuation, is not enough to keep the output power constant. Another significant shortcoming of magnetic limiters is their transient response. At the start of a high-power pulse, the incident power is transmitted without substantial attenuation for a short time (of the order of 1 μs). Only after this initial "leakage spike" is the power transmitted through the limiter significantly reduced.

The operation of ferrite limiters relies on "parametric excitation" of certain modes of oscillation that can exist in the ferrite material. These modes, generally known as "spin waves," are analogous to elastic waves (lattice vibrations) in solids, except that the magnetic moment associated with each atom, its "spin," rather than the atom's position is varying with time. All magnetic materials can support a huge number of such magnetic oscillating modes, much the same as all solids can support a huge number of lattice vibrations. In the absence of any externally imposed oscillating magnetic field, all spin waves are excited to a small extent to a level corresponding to thermal equilibrium. When an oscillating magnetic field is applied to the ferrite, certain spin waves can be excited above their thermal level by either the "direct" or the "parametric" process. The direct process excites spin waves whose resonant frequency equals the frequency of the oscillating magnetic field, increasing the amplitude of these waves in proportion to the strength of the externally imposed oscillating magnetic field. The parametric process, on the other hand, excites primarily spin waves whose resonant frequency equals one-half the frequency of the oscillating magnetic field (although other frequency combinations can also occur). In this case, the excitation is characterized by a sharp threshold: Virtually nothing happens until the strength of the oscillating field reaches a critical value, at which the excitation level suddenly increases substantially.

Parametric excitation of spin waves can occur under many different experimental conditions. The high-power effects observed in ferromagnetic resonance (FMR) (Bloembergen and Damon, 1952; Bloembergen and

Wang, 1954; Damon, 1953) were explained (Suhl, 1956, 1957) as arising from parametric excitation of spin waves. In these experiments, and in Suhl's theory, a high-frequency magnetic field was applied perpendicular to a constant bias field, as required in FMR. This leads to a rather complicated theoretical analysis, which is not conducive to a simple physical explanation.

A much simpler analysis applies when the oscillating magnetic field is applied parallel to the dc bias field (Schloemann *et al.*, 1960; Schloemann, 1962). Parametric excitation of spin waves occurs under these conditions, known as "parallel pumping," at a threshold comparable to that realized for "perpendicular pumping" (except when the ferrite is biased close to resonance). The parallel-pumping configuration also lends itself to the operation of ferrite limiters, although one type of limiter can only be realized in the perpendicular-pumping configuration as discussed below.

For a simple physical explanation of spin-wave instability under parallel pumping, consider the precessional motion of the magnetization vector of a thin disk of ferrite that is magnetized in its plane by a dc bias field. The spin precession follows an elliptical rather than a circular path in this case, because a strong demagnetizing field is generated when the magnetization vector has a component perpendicular to the disk plane. The elliptical spin precession that occurs in a thin disk of ferrite magnetized by a dc field in its plane is shown as the solid line in Fig. 9.

Consider now the influence of an rf magnetic field $h(t)$ applied parallel to the dc field. We assume that the frequency ω of this "pump field" equals twice the precession frequency ω_0, and that the precession motion is established before the pump field is switched on. We shall then investigate how the pump field affects this motion. In this discussion, the phase relation between the pump field and the precession motion will be important.

Consider the case in which the phase relation is such that the pump field adds to the dc field when the magnetization vector is at points 1 and 3 of the precession cycle and subtracts from the dc field at points 2 and 4 as indicated in Fig. 9. The resultant contributions $\delta(dM/dt)$ to the rate of change of the magnetization vector arising from the pump field are proportional to the vector product of the pump field and the magnetization vector. In Fig. 9, $\delta(dM/dt)$ is indicated by arrows for each of the four points under consideration. It should be noticed that in every case the arrow points toward the exterior of the precession ellipse. This implies that the pump field tends to increase the precession angle if the assumed phase relation is in fact realized.

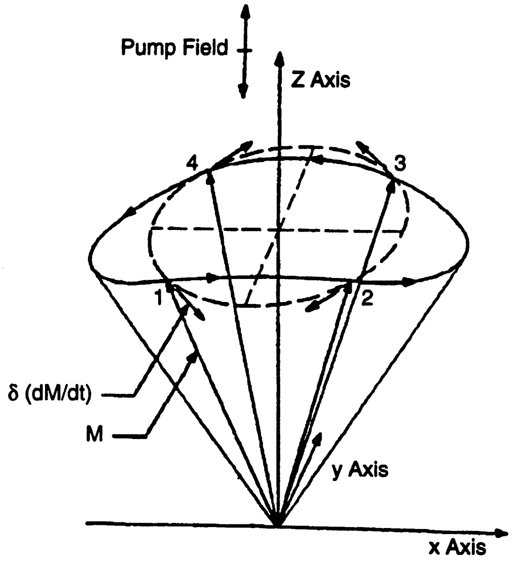

FIG. 9. Physical explanation of spin-wave instability in an rf magnetic field applied parallel to the dc field.

If the phase of the pump field is reversed, a picture very similar to Fig. 9 applies. In this case, all arrows point toward the interior of the precession ellipse, indicating that the pump field tends to decrease the precession angle if this phase relation is realized.

Consider now the case in which the pump field tends to increase the precession angle. In actuality, this tendency is at least partially compensated by relaxation processes, which always tend to decrease the precession angle. Instability occurs when the pump field is sufficiently large to overcome the relaxation processes.

In the above discussion, it was assumed that the ferrite sample has the shape of a thin disk. Actually, the results apply to samples of any shape, because spin waves that propagate through bulk ferrite at right angles to the dc bias field have the same behavior as the uniform precessional mode of a thin disk.

The spin-wave instability threshold h_{crit} calculated from the theoretical analysis of parallel pumping is given by

$$h_{\text{crit}} = \Delta H_k f_p / f_M.$$

Here f_p is the pump frequency, f_M is the magnetization frequency ($\approx$5 GHz for YIG), and ΔH_k is the so-called spin-wave linewidth, a parameter proportional to the relaxation rate of the spin waves. Further details on this subject may be found in the original literature (Schloemann, 1962; Rezende and de Aguiar, 1990).

The parametric excitation process just described has two significant features that are important in its application to microwave limiters:

1. It is effective only for high-amplitude signals. Low-amplitude signals are transmitted without significant loss.
2. It is "frequency selective" in the sense that a high-power signal at any frequency f_1 makes the ferrite lossy only in the vicinity of this frequency. A weak signal at some other frequency f_2 can still be transmitted without significant loss as long as the difference between f_1 and f_2 is substantially larger than the linewidth Δf of the resonant modes.

Ferrite limiters that have been described in the literature can broadly be classified as waveguide limiters, sphere resonator limiters, or microstrip limiters.

Waveguide limiters typically consist of a rectangular waveguide in which part of the cross section is occupied by a suitable ferrite. A dc magnetic field is applied that biases the ferrite into the so-called "subsidiary absorption" region. These devices have frequency-selective limiter characteristics over a broad frequency band. While a strong signal at some frequency f_1 is strongly attenuated, a weak signal at some other frequency f_2 (not too close to f_1) can be transmitted substantially without loss. The limiting threshold of waveguide limiters is relatively high.

Sphere resonator limiters have the same basic structure as sphere resonator filters (YIG filters). These devices have much lower limiting thresholds than the waveguide limiters, but the performance is restricted to a relatively narrow frequency band (which can be tuned if desired).

Microstrip limiters consist of a microstrip line on a thin ferrite substrate, which is exposed to a dc bias field. Stripline versions of the same type of device have also been described in the literature. These devices have the same kind of broad-band performance characteristics as waveguide limiters. The limiting threshold depends sensitively upon the substrate thickness, as well as on other parameters such as the spin-wave linewidth (Adam and Stitzer, 1993).

Although magnetic limiters can be based on polycrystalline bulk ferrites, the use of single crystals has significant advantages, because the threshold power level is generally much lower for single crystals. Since large single crystals of YIG grown by conventional bulk growth techniques are quite expensive, the use of YIG films grown by liquid-phase epitaxy on GGG is very attractive. Recent progress is exemplified by a microwave limiter using a 100-μm-thick film of YIG originally grown on GGG (Adam and Stitzer, 1993). For optimal performance, the GGG substrate was ground away, and the YIG film was used as a free-standing plate, with a conductive ground plane on one side and a microstrip transmission line on the other. The limiting threshold was about 1 mW in this case for frequencies between 2.5 and 5.3 GHz.

GLOSSARY

Atmospheric Plasma Spraying: A method in which particular material is drawn into a plasma discharge that provides the thermal energy to melt and/or vaporize the entrained particles.

Bloch Line: Boundary line between two parts of a Bloch wall in which the wall magnetizations have opposite directions.

Bloch Wall: Domain-wall type. The wall magnetization is parallel to the plane of the wall.

Delay Line: Device that delays a signal by a certain amount of time.

Doctor Blade: Device used for making slurry layers of constant thickness for use in tape casting.

Domain, Magnetic: Region inside a magnetic material in which the magnetization vector has a uniform orientation.

Domain Wall: Boundary between adjacent domains.

FMR: Ferromagnetic resonance.

Jet Vapor Deposition (JVD): A low-vacuum deposition technique in which evaporated metals from a heated wire are entrained in a high-velocity, low-pressure gas stream and rapidly deposited on a substrate. Oxygen gas can be introduced in the carrier gas to yield metal oxides (ferrites) as required.

Liquid-Phase Epitaxy (LPE): A high-temperature atmospheric-pressure film-growth process in which the desired phase is dissolved in a high-temperature solvent and is grown epitaxially onto a (usually) nonmagnetic substrate of closely matched crystallography.

Magnetostatic Approximation: Simplified approach to electromagnetic theory in which the effects of sources of the magnetic field are treated as if they were static.

Magnetostatic Wave (MSW): A type of electromagnetic wave that is predominantly magnetic in character. Its wavelength is much smaller than that of "light" of the same frequency.

Metal-Organic Chemical Vapor Deposition (MOCVD): A thin-film deposition process in which metal-organic compounds are separately vaporized and drawn into a carrier gas that impinges on a heated substrate, where decomposition and final compound formation occur.

Microstructure: The arrangement of crystalline and amorphous phases in a solid. Includes considerations such as crystallite size, size range, random or nonrandom orientation, and the presence of secondary phases and/or pores.

Nonreciprocal Microwave Device: Microwave device that has different propagation characteristics for opposite directions of wave propagation.

Parametric Excitation: Nonlinear excitation process in which the sum of the resonant frequencies of two or more fundamental modes of excitation equals the frequency of the exciting field.

Phase: A homogeneous solid that is distinguishable by optical microscopy, x-ray diffraction, or other techniques as an identifiable part of the microstructure.

Pulsed Laser Deposition (PLD): A film-deposition process in which a laser beam ablates material of the desired film composition from a target. A gaseous/particulate plume of this material is transported to the substrate, where deposition and growth take place.

Spin-Spray Deposition: A low-temperature, atmospheric-pressure technique in which aqueous solutions of soluble salts and precipitation-inducing agents are sprayed onto a substrate and rapidly evaporated by heat sources. The precipitation/growth process can be accelerated and localized by secondary sources, such as laser beams.

Spin Wave: Magnetic wave propagating in a ferromagnetic, ferrimagnetic, or antiferromagnetic medium; an elementary excitation analogous to a lattice vibration in a crystal lattice.

Tape Casting: Method used for making relatively thin plates of ceramic materials.

Wall Magnetization: Magnetization associated with a domain wall.

Works Cited

Abe, M., Tamaura, Y. (1984), *J. Appl. Phys.* **55**, 2614.

Adam, J. D., Daniel, M. R., Talisa, S. H. (1988), "A 13-Channel Magnetostatic Wave Filter Bank," *IEEE MTT-S Digest* 879–882.

Adam, J. D., Stitzer, S. N. (1993), "Frequency Selective Limiters for High Dynamic Range Microwave Receivers," *IEEE MTT-S Digest* 477–478.

Asada, H., Matsuyama, K., Gamachi, M., Miyoshi, H., Taniguchi, K. (1993), "Chip Organization of Bloch Line Memory," *IEEE Trans. Magn.* **MAG-29**, 2572.

Blakemore, R. P. (1975), *Science* **190**, 377.

Bloembergen, N., Damon, R. W. (1952), "Relaxation Effects in Ferromagnetic Resonance," *Phys. Rev.* **85**, 699.

Bloembergen, N., Wang, S. (1954), "Relaxation Effects in Para- and Ferromagnetic Resonance," *Phys. Rev.* **93**, 72–83.

Bozorth, R. M. (1951), *Ferromagnetism,* New York: Van Nostrand.

Cadieu, F. J., Cheung, T. D., Wickramaselcara, L., Kampoatt, N. (1986), *IEEE Trans. Magn.* **MAG-22**, 752.

Cadieu, F. J., Hodge, H., Chan, K. (1990), *J. Appl. Phys.* **67**, 4969.

Cullity, B. D. (1972), *Introduction to Magnetic Materials,* Reading, PA: Addison Wesley.

Damon, R. W. (1953), "Relaxation Effects in Ferromagnetic Resonance," *Rev. Mod. Phys.* **25**, 239–245.

Daniel, M. R., Adam, J. D. (1992), "Magnetostatic Wave Notch Filter," *IEEE MTT-S Digest* 1401–1402.

Dong, J. L., Artman, J. O., Charap, S. H. (1988), *J. Appl. Phys.* **63**, 4609–4615.

Dorsey, P. C., Chrisey, D. B., Horwitz, J. S., Lubitz, P., Auyeung, R. C. Y. (1994), *IEEE Trans. Magn.* **MAG-30**, 4512–4517.

Durst, K. D., Kronmüller, H. (1987), *J. Magn. Magn. Mater.* **68**, 63–75.

Grossinger, R., Turtelli, S. R. (1994), *IEEE Trans. Magn.* **MAG-30**, 455–460.

Heidmann, J., Krause, H., Engemann, J. (1988), *J. Appl. Phys.* **63**, 3159–3161.

Ishikawa, Y., Okada, T., Shimmura, S., Kanaya, F., Wakino, K., Nishikawa, T. (1992), "A Miniaturized Low Spurious 1.9 GHz Band-Pass Filter Using YIG Resonators with Multi-Metal Rings," *IEEE MTT-S Digest* 1403–1406.

Jacobs, I. S. (1969), *J. Appl. Phys.* **40**, 917–928.

Konishi, S. (1983), "A New Ultra High Density Solid State Memory: Bloch Line Memory," *IEEE Trans. Magn.* **MAG-19**, 1838.

Lowenheim, F. A. (1963), "Modern Electroplating," 2nd ed., New York: Wiley.

Lubitz, P., Lawrence, S. H., Rachford, F. J., Rappoli, B. J. (1994), *IEEE Trans. Magn.* **MAG-30**, 4539.

Luborsky, F. E., Drummond, B. J., McCary, R. O., Fahy, W., Sheerin, W. (1974), *IEEE Trans. Magn.* **MAG-10**, 560–563.

O'Dell, T. H. (1974), *Magnetic Bubbles,* New York: Macmillan.

Parker, R. J. (1990), *Advances in Permanent Magnetism,* New York: Wiley.

Pohm, A. V., Huang, J. S. T., Daughton, J. M., Krahn, D. R., Mehra, V. (1988), "The Design of a One Megabit Non-Volatile M-R Memory Chip Using 1.5 × 5 μm Cells," *IEEE Trans. Magn.* **MAG-24**, 3117–3119.

Pohm, A. V., Comstock, C. S., Daughton, J. M. (1989), "Analysis of M-R Elements for 10^8 Bit/cm^2 Arrays," *IEEE Trans. Magn.* **MAG-25**, 4266–4268.

Prinz, G. A., Krebs, J. J. (1981), *Appl. Phys. Lett.* **39**, 397.

Prutton, M. (1964), *Thin Ferromagnetic Films,* London: Butterworths.

Ranmuthu, K. T. M., Pohm, A. V., Daughton, J. M., Comstock, C. S. (1993), *IEEE Trans. Magn.* **MAG-29**, 2593.

Rezende, S. M., deAguiar, F. M. (1990), "Spin-wave Instabilities, Auto-Oscillations, and Chaos in Yttrium-Iron-Garnet," *Proc. IEEE* **78**, 893–908.

Schloemann, E., Green, J. J., Milano, U. (1960), "Recent Developments in Ferromagnetic Resonance at High Power Levels," *J. Appl. Phys.* **31**, 386S–395S.

Schloemann, E. (1962), "Longitudinal Susceptibility of Ferromagnets in Strong rf Fields," *J. Appl. Phys.* **33**, 527–537.

Schloemann, E. (1973a), *J. Appl. Phys.* **44**, 1837–1849.

Schloemann, E. (1973b), *J. Appl. Phys.* **44**, 1850–1854.

Schloemann, E. (1976), *J. Appl. Phys.* **47**, 1142–1150.

Schloemann, E. (1988), "Circulators for Microwave and Millimeter Wave Integrated Circuits," *Proc. IEEE* **76**, 188–200.

Schloemann, E., Tustison, R., Weissman, J., Van Hook, H. J., Varitimos, T. (1988), "Expitaxial Fe Films on GaAs for Hybrid Semiconductor-Magnetic Memories," *J. Appl. Phys.* **63**, 3140–3141.

Schwartz, B. (1988), "Ceramic Packaging of Integrated Circuits," in: L. L. Levinson (Ed.), *Electronic Ceramics,* New York: Marcel Dekker, Chap. 1.

Slonczewski, J. C. (1973), *J. Appl. Phys.* **44**, 1759–1770.

Speliotis, D. E., Lynch, W. (1991), *J. Appl. Phys.* **69**, 4496–4498.

Suhl, H. (1956), "The Nonlinear Behavior of Ferrites at High Microwave Signal Levels," *Proc. IRE* **44**, 1270–1293.

Suhl, H. (1957), *J. Phys. Chem. Solids* **1**, 209.

Truedson, J. R., McKinstry, K. D., Kabos, P., Patton, C. E. (1993), *J. Appl. Phys.* **74**, 2705–2718.

Tummala, R. R. (1989), "Multilayer Ceramic Capacitors," in: R. R. Tummala, E. J. Rymaszewski (Eds.), *Microelectronic Packaging Handbook* New York: Van Nostrand, Chap. 7.

Tustison, R., Varitimos, T., Van Hook, H. J., Schloemann, E. (1987), *Appl. Phys. Lett.* **51**, 285.

von Aulock, W. H. (1965), *Handbook of Microwave Ferrite Materials,* New York: Academic.

Wang, X. W., Kudesia, R., Lou, J., Hao, J., Snyder, R. L., Duan, H. M., Hermann, A. M. (1992), "Deposition of Electronic Films by Atmospheric RF Plasma Aerosol Spray Techniques," in: *Thermal Spray: International Advances in Coatings Technology,* Metals Park, OH: ASM International, pp. 567–574.

Weiss, J. A., Watson, N. G., Dionne, G. F. (1989), "New Uniaxial Ferrite Millimeter-Wave Junction Circulators," *IEEE MTT-S Digest* 145–148.

Williams, C. M., Abe, M., Itoh, T., Lubitz, P. (1994), *IEEE Trans. Magn.* **MAG-30**, 4896–4899.

Winkler, G. (1981), *Magnetic Garnets,* Braunschweig: F. Vieweg.

Wohlfarth, E. P. (Ed.) (1982), *Ferromagnetic Materials,* Amsterdam: North-Holland.

Wu, J. C., Humphey, F. B. (1985), "Vertical Bloch Line Memory," *IEEE Trans. Magn.* **MAG-21**, 1773.

Further Reading

General

Burke, H. E. (1986), *Handbook of Magnetic Phenomena,* New York: Van Nostrand Reinhold (excellent, nonmathematical discussion of magnetic effects).

Cullity, B. D. (1972), *Introduction to Magnetic Materials,* Reading, PA: Addison-Wesley.

Wohlfarth, E. P. (Ed.) (1982), *Ferromagnetic Materials,* Amsterdam: North-Holland. A handbook on the properties of magnetically ordered substances:

Vol. 1 Iron alloys, rare-earth alloys, magnets, strictive compounds;

Vol. 2 Ferromagnetic insulators, microwave ferrites, recording materials; and

Vol. 3 Hexagonal ferrites, oxide spinels, sulfo spinels.

Yamaguchi, T., Abe, M. (Ed.) (1992), *Ferrites,* Tokyo: Japan Society of Powder and Powder Metallurgy.

Magnetic Bubble Films

Jouve, H. (Ed.) (1988), *Magnetic Bubbles,* London: Academic. Discussions on properties of crystalline oxide, amorphous rare-earth alloy, and Permalloy metal films.

Winkler, G. (1981), *Magnetic Garnets,* Braunschweig, Germany: F. Vieweg.

Permanent Magnetism

Lanotte, L. (Ed.) (1993), *Magnetoelastic Effects and Applications,* Proceedings of International Conference on Magnetoelastic Effects, Naples, Italy, May 1993, Amsterdam: Elsevier.

McCaig, M., Clegg, A. G. (1987), *Permanent Magnets in Theory and Practice,* 2nd ed., New York: Wiley. General text on bulk materials.

Parker, R. J. (1990), *Advances in Permanent Magnetism,* New York: Wiley.

POINT AND EXTENDED DEFECTS IN CRYSTALS

W. Beall Fowler, *Department of Physics and Sherman Fairchild Laboratory, Lehigh University, Bethlehem, Pennsylvania, U.S.A.*

Rob Phillips, *Division of Engineering, Brown University, Providence, Rhode Island, U.S.A.*

Anders E. Carlsson, *Department of Physics, Washington University, St. Louis, Missouri, U.S.A.*

Introduction 317
1. Point Defects 318
1.1 Definition 318
1.2 Types of Point Defects 318
1.3 Origins of Point Defects 319
1.3.1 Thermodynamics 319
1.3.2 Irradiation 320
1.3.3 Growth 320
1.3.4 Implantation 320
1.3.5 Molecular-Beam Epitaxy (MBE) 320
1.4 Effects of Point Defects in Metals 320
1.4.1 Electrical Effects 320
1.4.2 Magnetic Effects 320
1.4.3 Diffusion 321
1.5 Effects of Point Defects in Semiconductors 321
1.5.1 Shallow Donors and Acceptors 321
1.5.2 Deep Defects 322
1.5.3 Hydrogen Passivation 323
1.5.4 Further Electrical Properties 323
1.5.5 Optical Properties 324
1.6 Effects of Point Defects in Insulators 325
1.6.1 Color Centers 325
1.6.2 F Center in Alkali Halides 325
1.6.3 F-Center Creation by UV Irradiation 326
1.6.4 Energy Transfer: Vibrational Luminescence of the F-Center–CN$^-$ Pair 326
2. Extended Defects 327
2.1 Definition 327
2.2 Types of Extended Defects 327
2.2.1 One-Dimensional or Line Defects 327
2.2.2 Two-Dimensional or Planar Defects 328
2.2.3 Cracks 329
2.3 Origins of Extended Defects 329
2.4 Structure 330
2.5 Motion and Response to External Forces 334
2.6 Effects on Materials Properties 335
2.6.1 Mechanical Properties 335
2.6.2 Mass Transport 336
2.6.3 Acoustic Attenuation 336
2.6.4 Electrical Properties 336
2.6.5 Magnetic Effects 337
2.6.6 Interaction Effects 337
3. Concluding Remarks 337
Glossary 337
Works Cited 339
Further Reading 339

INTRODUCTION

The perfect crystal is a conceptual framework that can be closely approximated under careful preparation conditions. Simply stated, a perfect crystal is a periodically repeated array of identical building blocks such as atoms or arrays of atoms. For example, a perfect sodium chloride (NaCl) crystal can be thought of as an infinite array of cubes in which alternating corners are Na^+ and Cl^- ions.

Perfect crystals are both less interesting and much less important technologically than crystals that contain imperfections, or defects. Defects in crystals are responsible for many of their most important properties,

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

such as the tunable electrical behavior of semiconductors and plastic deformation and fracture of materials. Defects are thus the basis for much of modern technology. This article discusses various types of defects, dividing the material along a few main lines. Point defects have rather simple structures and by themselves have relatively little impact on mechanical properties. Their electronic and optical properties, on the other hand, are very important both scientifically and technologically and have been widely studied; hence these properties are emphasized in this article.

The additional complexity of extended defects makes their structural properties considerably more varied than is the case for point defects. Furthermore, extended defects play a key role in determining the mechanical properties of solids, and hence these aspects are emphasized in this article. Finally, a synthetic view is taken that points out the crucial role played by interactions between these various types of defects.

1. POINT DEFECTS

1.1 Definition

The point defect is the simplest class of crystal defects by virtue of its reduced dimensionality. Point defects are of the order of atomic size, in contrast, for example, to an interface between two chemically distinct materials, which is of macroscopic size and two dimensional. Implied in the definition of point defect is the notion that these defects occur at sufficiently low concentration that they do not interact with each other. In other words, their effects on crystal properties are simply additive. The defect concentration must be small for this condition to be satisfied. For example, if the point-defect concentration is 1 in 10^6, the average separation is 100 atomic spacings. With a random arrangement, many will be closer, but the fraction that interact is expected to be small. However, if the concentration is much larger, interaction effects may become important.

Even with this very dilute concentration it turns out that major optical and electrical properties can be profoundly affected by the presence of point defects. These properties may be desirable or undesirable. Hence, one may deliberately foster the presence of particular point defects, or, conversely, pains may be taken to eliminate others.

1.2 Types of Point Defects

The most common point defects in chemically pure crystals are vacancies and interstitials. A vacancy is the absence of an atom or ion from its perfect-crystal site, while an interstitial is an atom or ion at a position other than a perfect-crystal site. Vacancies and interstitials may occur in pairs, if an atom is removed from its normal site and placed elsewhere within the crystal, or individually, as for example if an atom is removed from its normal site and leaves the crystal. A vacancy-interstitial pair is called a Frenkel defect, while an isolated vacancy is called a Schottky defect. Figure 1 illustrates these defects schematically.

Another class of point defect is the chemical impurity—an atom, ion, or (small) molecule that is chemically different from the perfect-crystal constituents. A chemical impurity may replace a perfect-crystal atom, in which case it is called substitutional, or it may occur at an interstitial site.

The antisite defect can occur in a crystal that contains more than one type of chemical constituent. Consider, for example, the case in which the perfect crystal consists of a regular array of atoms of types A and B. If instead an A atom is on a B site, or vice versa, an antisite defect is said to exist. A

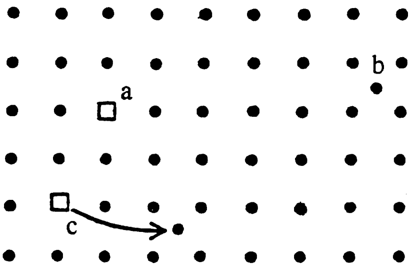

Key:
a = vacancy (Schottky defect)
b = simple interstitial
c = vacancy-interstitial pair (Frenkel defect)

FIG. 1. Schematic illustration of simple defects.

specific example is an arsenic atom on a gallium site in gallium arsenide (GaAs).

These descriptions of point defects barely hint at what in fact is a wide range of structures and properties. Some of these will be treated as examples, but it should be pointed out here that in general the atoms and ions that make up crystals are not inert "billiard balls" but rather are chemically reactive species. Consequently, when an atom is removed, those atoms in the vicinity of the vacancy experience chemical and physical forces that can lead to large, possibly symmetry-lowering relaxations. And when an interstitial is formed, that atom or ion may interact chemically with its neighbors in a way that leads to relaxations and rather different local chemistry. Finally, small molecules may form substitutional or interstitial defects, and their properties can reflect the extra degrees of freedom that they possess. Thus, a CN^- or an OH^- may replace a Cl^- in KCl, for example.

A type of defect that is intermediate between point and extended defects is the cluster or nanocrystal. A number of examples exist. In alkali halide crystals such as KCl, a small cluster of alkali atoms may form into a metallic colloid. The silver halide photographic process involves small metallic silver clusters formed upon exposure to light. Other articles in the Encyclopedia address such topics; for example, see NANOPHASE MATERIALS.

1.3 Origins of Point Defects

1.3.1 Thermodynamics One feature that distinguishes point defects from typical extended defects is that they can exist purely as a consequence of thermal equilibrium. We imagine a particular perfect-crystal structure to represent the lowest-energy arrangement of a large number of atoms or ions at a temperature of 0 K. Any other arrangement, such as a Schottky or a Frenkel defect, will require the addition of energy to the system. If the system is in thermal equilibrium at a finite temperature, a fraction of such defects will exist as determined by Boltzmann statistics:

$$N_D = BN\exp(-AE/kT). \qquad (1)$$

Here N_D is the number of defects, N is the number of atomic sites, A and B are factors of order 1, E is the energy required to create the defect, k is Boltzmann's constant, and T is the Kelvin temperature. Since chemical energies are often of order 1 eV, and since at room temperature kT is about 0.025 eV, the atomic fraction of defects at room temperature may be of order 10^{-18}, a number much lower than can be detected. But if the creation energy is only 0.4 eV, the fraction can be of order 10^{-7}, clearly large enough to have noticeable effects. Figure 2 shows data on the number of vacancies in sodium as a function of temperature, which follows Eq. (1) very closely.

Thermodynamics can play other roles. A given defect may have more than one possible structure. In thermal equilibrium, thermodynamics will determine the fraction of defects in each structure. Also, crystal processing may involve high-temperature operations. If cooling occurs faster than thermal equilibrium can be established, defects can be "frozen in" at concentrations appropriate to the higher temperature, and this higher

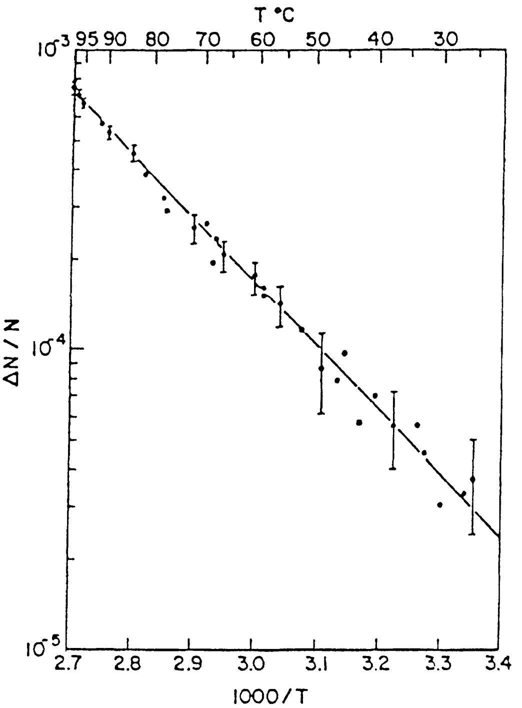

FIG. 2. Fractional number of sodium vacancies in sodium vs temperature. The solid line is a fit by Eq. (1). After Feder and Charbnau (1966).

nonequilibrium concentration can then persist at lower temperatures.

1.3.2 Irradiation Point defects may be introduced by irradiation. The mechanisms and details of various processes will depend on the nature of the crystal. It is easy to see how heavy particles (such as neutrons) of sufficient energy can displace atoms in the crystal by momentum- and energy-conserving "billiard-ball" collisions. For lighter particles such as electrons or gamma rays, the requirements of both energy and momentum conservation require very high energies for the lighter particle; there is a threshold for collision processes such that, for example, electron energies must typically be hundreds of electron volts to induce collisional damage.

In some crystals, irradiation can introduce point defects by simply ionizing atoms or ions, in which case the energy required is just the band gap, of order several electron volts. In such cases, the defect creation occurs via a chemical reaction between the ionized atom, or the atom in an excited state, and other atoms in its vicinity.

1.3.3 Growth Chemical impurities may be introduced into a crystal during the growth process. This is called doping. In the simplest case, one begins with a chemically pure constituent, melts it, and adds a definite amount of the desired impurity to the melt. Then, by one of a variety of growth processes, one cools the melt and ends up with a crystal that contains the impurity, suitably dispersed throughout and in the proper location and chemical state.

1.3.4 Implantation Growth is not always a suitable way to introduce desired impurities. For example, some systems are chemically immiscible, and the impurity segregates from the crystal as the melt is cooled. In other cases, the impurities do not form the desired type of defect during the cooling process. One way to avoid such problems is by ion implantation. Ion implantation (*q.v.*) involves firing a beam of impurities into a crystalline sample under controlled conditions. When this process works as desired, the result is a dispersed set of impurity defects of the proper type. Implantation works best with thin samples, such as those used in planar microelectronics technology. An undesirable side effect is often the creation of defects due to collisions of the implanting ions with atoms in the host crystal. Often most of these undesired defects can be removed by annealing.

1.3.5 Molecular-Beam Epitaxy (MBE) Molecular-beam epitaxy (*q.v.*) (MBE) is a technique that is widely used to grow planar crystals with desired chemical and physical characteristics. In its simplest form, it involves depositing atoms from molecular beams, layer by layer, onto a substrate, thereby forming thin films. This is a very powerful technique that is used to form precisely engineered layered compounds. We mention it here only to point out that chemical impurities can be introduced as part of the process, and so doping can be achieved.

1.4 Effects of Point Defects in Metals

1.4.1 Electrical Effects In most cases, point defects contribute to the electrical resistivity of metals. A perfect metallic crystal such as copper at low temperatures has a high conductivity because the conduction electrons move as Bloch waves, plane waves modulated by functions that have the periodicity of the lattice, without scattering through large regions of the crystal. Point defects break the symmetry and serve as scattering centers for the Bloch waves. This scattering tends to be rather weak, however, because of the screening of the impurity potential by the rest of the conduction electrons. Considering a zinc impurity substituting for a copper atom in copper, for example, there is an extra positive charge on the zinc and therefore the possibility of Coulomb scattering. The conduction electrons, which are free to respond to the Coulomb field, act to screen it and consequently the scattering potential is relatively weak.

1.4.2 Magnetic Effects The magnetic effects of defects depend on the magnetic characteristics of both the host material and the defect. This leads to a wide variety of effects in metals, some of which are very complicated. We mention only one situation here, a magnetic impurity in a nonmagnetic host. A magnetic impurity could be an atom or ion with a partially filled *d* shell, such as vanadium, while a nonmagnetic host could be copper. The interaction between the impurity electrons and the conduction electrons of the host leads to broadening of the defect levels and may render the system nonmagnetic, as in the case of vanadium in aluminum. Further discussion of magnetic defects is beyond the scope of this article.

1.4.3 Diffusion Atomic diffusion is a general property of defects in solids (see SOLIDS, DIFFUSION IN), but it is particularly important in metals where the diffusion of point defects to extended defects can have important effects on the mechanical properties of the system. Many detailed mechanisms of defect diffusion exist, depending on the defect, the host, and other conditions such as temperature and external fields. All share the feature that motion of a defect occurs over a macroscopic distance as the result of a series of individual microscopic motions, which are statistical in nature.

A simple model will suffice to present ideas of diffusion. Consider an impurity A at an interstitial site. The number of equivalent sites will be approximately equal to the number of host atoms, i.e., it will be a very large number. There will be a small number of equivalent nearest-neighbor sites. For A to move to one of these nearest-neighbor sites will require a thermal event, motion over a potential barrier. We assume that the temperature T is high enough that such motions can occur reasonably rapidly. If there is just one A defect and no additional fields the net motion of A will be a "random walk", its net displacement will be hindered by its tendency to jump back, or sideways, in addition to away from the initial site, and it will not tend to move in any particular direction. Its net displacement will vary as the square root of time. If, however, there is an external field that exerts a net force in one direction, then the probability of jumps in that direction will be greater than the probability of jumps in the reverse direction, and so there will be a net motion of A in the direction of the net force.

An additional feature of diffusion is that it can take place in the absence of external fields if there is a concentration gradient of impurity atoms A. Atoms will tend to hop from regions of high concentration to regions of lower concentration, since in a direction of higher concentration the neighboring interstitial sites are likely to already be occupied. This effect is described by Fick's first law:

$$J = -D\frac{\partial C(x)}{\partial x}. \qquad (2)$$

Here J is the particle flux, the number of particles passing through a unit area per unit time, C is the concentration, and D is the diffusion coefficient. The diffusion coefficient D contains all the material properties, and it is also a function of temperature. In fact, since the diffusion process involves thermal motion over a barrier, the diffusion coefficient often has a variation with temperature whose form is similar to that of Eq. (1):

$$D = D_0\exp(-\Delta E/kT). \qquad (3)$$

Here, ΔE is a measure of the height of the potential barrier and k is Boltzmann's constant. Figure 3 shows experimental data for the diffusion coefficient of carbon in iron as a function of temperature, which fits Eq. (3) very closely.

1.5 Effects of Point Defects in Semiconductors

1.5.1 Shallow Donors and Acceptors Semiconductors are important in technology because of the existence of impurity atoms

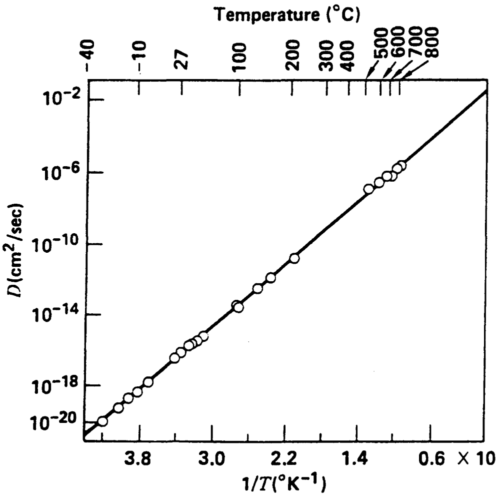

FIG. 3. Temperature dependence of the diffusion coefficient D of carbon in iron. The solid line is a fit by Eq. (3). After Borg and Dienes (1988).

that serve as so-called shallow donors or acceptors and provide electrical carriers. An example of a shallow donor is substitutional phosphorus in silicon. Phosphorus is a group V element, one column to the right of silicon, which is group IV. Silicon has four valence electrons, and in the diamond lattice these form tetrahedral hybrid orbitals and make strong covalent bonds with four silicon neighbors that lie in tetrahedral positions. Phosphorus has five valence electrons. Four of these form tetrahedral hybrid covalent bonds with neighboring silicons. The fifth electron, redundant for bonding and attracted by the positive charge of the phosphorus core, goes into a modified hydrogen-like state that is weakly bound to the core. This state is characterized by an effective mass, associated with the silicon conduction band, and a large dielectric constant, also a silicon characteristic. The electron is bound by 0.044 eV with respect to ionization into the conduction band, a small enough number that by Eq. (1) a substantial fraction of the phosphorus atoms will be ionized, donating their weakly bound electron to the conduction band at room temperature. The number of conduction electrons will be proportional to the number of donor impurities; hence the electrical properties of the system may be customized. Figure 4 shows schematically the energies of several types of semiconductor defects.

A similar argument leads to an understanding of shallow acceptors. A group III impurity such as aluminum has one fewer valence electron than does silicon. It then accepts one of the electrons from the filled valence band, becoming negatively charged. The empty valence state, or "hole," is positively charged and is attracted to the aluminum core. Again, a shallow effective-mass state is formed and the hole is bound, this time to the valence band, by 0.057 eV. At room temperature a large fraction of ionized holes will be present; these conduct electricity much as excess electrons in the conduction band do, so that again the electrical properties may be customized.

A semiconductor with an excess of donors is called "n-type"; one with an excess of acceptors is called "p-type."

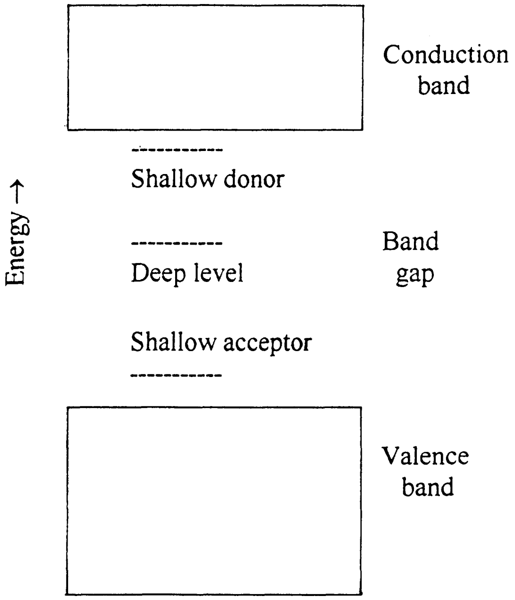

FIG. 4. Schematic illustration of the energies of several types of semiconductor defects.

The formation of shallow effective-mass states, while logical, is by no means guaranteed. It depends on the small effective masses and large dielectric constants that are characteristic of semiconductors. In the insulating crystal silicon dioxide in the alpha quartz structure, silicon also has four neighbors (which happen to be oxygens) in a nearly tetrahedral arrangement. But in this case, when phosphorus or aluminum substitutes for silicon, the resulting defect level is "deep," far away in energy from the conduction or valence band edges, and the defect electron or hole is localized near the impurity. Thus, in silicon dioxide the impurities do not readily provide charge carriers, as they do in silicon.

1.5.2 Deep Defects Deep-level defects also exist in semiconductors. Such defects are generally characterized by wave functions that are localized near the defect, and by energy levels that are not close enough to the conduction or valence band to be thermally ionized in large numbers at room temperature. Deep defects typically have another interesting characteristic, namely, the ability to occur in several charge states. Deep defects exist in great variety, and we mention for illustration here only two: the vacancy in silicon and the anion antisite in gallium arsenide.

The silicon vacancy features "dangling

bonds," hybrid orbitals from four neighbors pointing toward the vacancy. In the case of the neutral vacancy, these neighboring silicons relax toward each other, pairwise, to form extended Si-Si bonds. As electrons or holes are trapped at the vacancy, further symmetry-lowering atomic displacements occur, explainable in large part by a molecular-orbital picture of the defect. Five charge states of the vacancy, ranging from -2 to $+2$, are found to exist within the band gap of silicon. One of the most interesting characteristics of the silicon vacancy is a "negative-U" ordering that means, in effect, that two vacancies in the $+1$ charge state are unstable with respect to one vacancy in the neutral and the other one in the $+2$ charge state. Here U is an "effective" Coulomb interaction between electrons or holes; it can, in principle, be negative (attractive) if the actual repulsive Coulomb interaction is weak and if sufficient energy can be gained by atomic relaxations in particular charge states to offset the weak Coulomb interaction.

Arsenic on a gallium site, labeled $\mathrm{As_{Ga}}$, is of particular interest because of its likely connection to a defect called EL2 in gallium arsenide. EL2 is a deep defect that is desirable in certain cases; as a charge trap it makes gallium arsenide semi-insulating, an important feature for device substrates. EL2 is, however, metastable; its stable configuration is a deep donor, and optical excitation converts EL2 into an alternative metastable atomic configuration, from which there is a 0.3-eV activation energy to return to the stable configuration. EL2 has been extensively studied, and the most likely model is an isolated $\mathrm{As_{Ga}}$, whose metastable state consists of an adjacent As interstitial–Ga vacancy pair.

1.5.3 Hydrogen Passivation

Hydrogen is a common impurity in solids. Its chemical activity makes hydrogen of particular importance in semiconductors, where its presence can drastically change the nature of both shallow and deep defects. For example, an isolated dangling-bond defect, which otherwise might trap electrons or holes, can be passivated by attachment of a hydrogen atom. This is a particularly important effect in amorphous silicon, where otherwise the large number of dangling silicon bonds leads to so many deep traps that the material is electrically useless. But the Si-H combination is not a deep trap, and so hydrogenated amorphous silicon is electrically important. There are many other examples where hydrogen can eliminate deep traps.

Hydrogen can, however, also passivate shallow traps and reduce electrical activity. For example, in boron-doped silicon a hydrogen atom is trapped in a bond-centered site between B and Si. This leads, in a first approximation, to a 3-coordinated boron and a Si-H bond and removes the effective-mass acceptor state associated with boron. This situation is shown in Fig. 5. The model for donors is different, and the situation changes in detail as one considers other semiconductors, but the effect is the same: shallow traps can be passivated and charge carriers can be removed by hydrogen.

1.5.4 Further Electrical Properties

The preceding sections have outlined the essential electrical properties of semiconductor defects. Shallow donors or acceptors provide electrical carriers in numbers proportional to the defect concentrations. Deep defects can trap electrical carriers, also in numbers proportional to their concentrations. Hydrogen can passivate both types of defects and so remove their effect.

As in the case of metals, defects can also scatter conducting electrons and holes and thereby contribute to the electrical resistivity of the solid. Thus, for example, while every

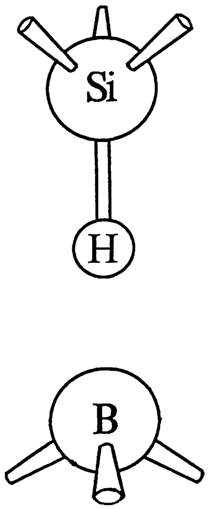

FIG. 5. Illustration of the structure of hydrogen-passivated boron in silicon.

phosphorus impurity can contribute to the conductivity by donating an electron to the conduction band, every phosphorus is also a potential scattering defect for those very electrons. The scattering effects are, in general, larger in semiconductors than in metals, because the much smaller concentrations of conducting electrons (or holes) are less effective in screening the scattering potential. In some device applications, the scattering problem is overcome by arranging that the conducting particles move in a region of the crystal that is defect free. This is possible in planar geometries, where the current flows in a shallow region of the semiconductor near an interface while the donor or acceptor atoms that provide the carriers are engineered to lie in the bulk of the semiconductor.

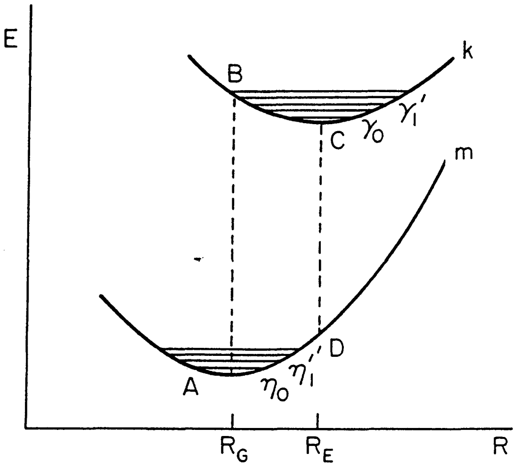

FIG. 6. Configuration-coordinate diagram for an idealized system.

1.5.5 Optical Properties Point defects in nonmetallic crystals will, if they have energy levels within the gap between the valence and conduction bands, lead to characteristic optical absorptions that involve the excitation of electrons or holes from a ground to an excited state. In some cases, there is also luminescence as the excited defect returns to the ground state.

The general theory of optical absorption, as well as some aspects of luminescence, is similar for semiconductor and insulator defects, although details tend to differ somewhat. The simplest conceptual model is adapted from molecular physics and involves the electron–phonon interaction. The defect in its ground state has neighboring atoms that vibrate about certain equilibrium sites. When the defect is excited the neighboring atoms, responding to a new defect wave function, relax to new equilibrium positions. This may be represented by a configuration-coordinate diagram (Fig. 6), which in an idealized case consists of two displaced parabolas [representing the harmonic energy of the vibrating atoms in ground (*m*) and excited (*k*) states]. The vibrational levels are quantized, and hence quantum levels η and γ are shown in association with the parabolas. An optical transition, in general, involves a change of both electronic and vibrational quantum numbers, and the results in two limits are instructive.

In the small-relaxation limit, where the excited-state parabola is nearly directly above that of the ground state, the strongest transitions involve no change in vibrational quantum number. Thus, at low temperature, if only the zero-point vibrational level of the ground state is occupied, the strongest transition will be a sharp "zero-phonon" transition from A to C to the zero-point level of the excited state. In the large-relaxation limit, with both horizontal and vertical displacements of the parabolas, the zero- and few-phonon transitions are weak because the wave functions of the low-lying displaced oscillators do not overlap strongly, and the optical absorption is characterized by a broad multiphonon transition A to B, which is approximately vertical in the configuration coordinate *R*.

In the absence of nonradiative channels, the excited defect will luminesce, emitting a photon as it returns to the ground state. In general, the radiative lifetime is at least several orders of magnitude longer than the vibrational relaxation rates, and so the excited defect will relax to its lowest vibrational state; e.g., at low temperature it will relax to the zero-point state C before the defect radiates. In the weak-coupling limit the zero-phonon line C to A will lie congruent with that for absorption, while few-phonon lines will lie at lower photon energy (in contrast to absorption, where they lie at higher energies). In the strong-coupling limit, the broad multiphonon emission C to D will be Stokes-shifted from that for absorption by a relaxation energy, as shown in Fig. 6. (The difference in energy between the absorption and the emission bands is referred to as the Stokes shift.)

The above discussion is a highly truncated and simplified version of actual situations. It in fact works well to describe, at least qualitatively, many defects in insulators, as discussed in the next section. It also works well in describing some deep-level semiconductor defects such as the carbon–phosphorus pair in silicon. However, for many semiconductor defects, especially shallow ones, the situation is somewhat different. For example, for shallow defects the excited electronic states lie very close to the respective bands and optical transitions into them are rather weak. Thus, rather than observe an isolated transition between two states one sees the superposition of many ground-to-excited-state transitions blending into ground-to-unbound (or continuum) transitions. The corresponding luminescence may or may not be observed; the excited state is likely to be thermally ionized and so one might observe continuum-to-ground-state transitions. However, the conduction electron or hole may well journey to a surface or other defective portion of the crystal at which some nonradiative decay process may take place, or the electron or hole may simply be trapped away from the original defect.

1.6 Effects of Point Defects in Insulators

1.6.1 Color Centers Because many point defects in insulators generate optical activity in the visible, they have been given the general name of color centers. Given the variety of possible hosts and defects, the number of examples of color centers is enormous. The study of simple color centers has been of great importance in elucidating basic physics, while more complicated color centers have proven to be of technological importance in, for example, solid-state laser systems.

As discussed in Sec. 1.5.5, defects in insulators and deep-level defects in semiconductors share many basic concepts. However, defects in insulators generally tend to lie more in the strong-coupling limit than those in semiconductors, although examples in both strong- and weak-coupling limits are found.

In this section, we present several topics as examples. First, we discuss the F center in alkali halides as the prototype color center. Next, we discuss F-center creation by UV irradiation as an example of defect chemistry. Finally, we discuss vibrational luminescence of the F-center–CN^- complex as an example of energy transfer.

1.6.2 F Center in Alkali Halides The simplest color center is the F center in alkali halides, an electron trapped at a halogen$^-$ vacancy. As a quasi one-electron defect, it is the "hydrogen atom" of defect centers in insulators. The site at which the electron is trapped has cubic symmetry, with the nearest-neighbor alkali ions equally spaced along $\pm x$, $\pm y$, $\pm z$ directions. This makes the analogy with hydrogen even stronger since the lowest-lying electronic wave functions are approximately *s*- and *p*-like. The potential-energy function more closely resembles a square well than a hydrogenic potential, however, and this leads among other things to a removal of degeneracy in the angular momentum quantum number; e.g., 2*s* and 2*p* states will not be degenerate.

The absorption transition may then be thought of as follows. The trapped electron is in a "1*s*" state whose energy and wave function are consistent with the positions of neighboring ions. Optical absorption involves a transition to a "2*p*" state, whose energy and wave function are consistent with the positions of the ions before the transition took place: this is the Franck–Condon principle. The peak position is the 2*p*-1*s* energy difference associated with these equilibrium positions, while the width of the band reflects the vibrations of the ions and the consequent instantaneous variation in the transition energy. In KCl, the peak position for absorption at liquid helium temperature (4 K) is 2.31 eV, and the full width at half-maximum is 0.16 eV.

After absorption there is a high probability in many alkali halides at low temperature that luminescence will follow. Now the lattice relaxes and the system comes to equilibrium in an excited state and undergoes a Franck–Condon transition to the ground state. In KCl, the peak position for luminescence is 1.22 eV and the full width at half-maximum is 0.26 eV. Thus, nearly half the absorption energy is lost to heat in the form of lattice relaxation as the F center goes through a complete optical cycle.

1.6.3 F-Center Creation by UV Irradiation It is found that F centers may be created by light that excites an electron from the valence to the conduction band (in fact, light of slightly lower energy that creates an exciton also leads to F-center generation). Thus, defects are created by orders of magnitude less energy than required by familiar collision processes. This is an example of defect chemistry and is analogous to known chemical processes in other systems in that a reaction may be made to go if the system is raised to the appropriate excited state.

A highly simplified explanation of F-center creation is the following. An excited electron–hole pair (a self-trapped exciton, or STE) in an alkali halide relaxes into a configuration in which the hole is shared by two adjacent halogen ions, which in turn move together to form a molecule–ion configuration. The electron is somewhat weakly bound to the self-trapped hole. An alternative configuration for the STE is one in which the halogen molecule-ion moves away from one of the halogen sites along the molecular axis. This creates an adjacent halogen-vacancy-like region, into which the weakly bound electron can be trapped into an F-center-like ground state. It turns out that this process is energetically favorable, at least in some cases, and furthermore the hole can continue to move along the first axis, one step at a time. Each step it moves, the hole is retrapped by a new pair of halogen ions, which then relax. Eventually, the self-trapped hole finds itself some distance from the electron, and one has relatively isolated pair defects consisting of an F center and a so-called F_H center. Following through this process, one sees that an F_H center consists of a halogen molecule-ion at a site normally occupied by a halogen ion.

1.6.4 Energy Transfer: Vibrational Luminescence of the F-Center–CN$^-$ Pair Transfer of excitation energy from one defect to another is of both scientific and technological interest. The scientific interest involves understanding the mechanisms of energy transfer, and from that being able to predict its occurrence. The technological interest arises from a number of possible and achieved applications, one of the oldest of which is sensitized luminescence. Sensitized luminescence involves two different defects, one of which (the sensitizer) absorbs light in a desired wavelength range, the other of which (the activator) emits light in a different wavelength range. Typically, the activator's optical transitions are forbidden, which means that it does not absorb light readily. The sensitizer has allowed transitions and so absorbs light. Subsequent lattice relaxation brings the sensitizer's emission energy difference into resonance with the activator's absorption energy difference (more specifically, the broad bands overlap), which means that an energy-conserving transfer from sensitizer to activator may occur. Its actual occurrence requires that the transfer rate be greater than the rate for the sensitizer to return to the ground state by other means.

There are two transfer mechanisms that have been explored in detail. One is a dipole–dipole process and is analogous to a classical nonradiative transfer of energy between two antennas. The other is a quantum-mechanical exchange process. It should be noted that similar energy transfer also occurs in photosynthesis, and some of the early theoretical work involved biological systems.

The F-center–CN$^-$ pair is a particularly interesting example of energy transfer in which the detailed mechanism is not yet known. In this case, electronic energy from the excited F center is converted to internal vibrational energy, the stretch mode of the CN$^-$ molecule ion. The prototype system utilizes a CsCl host and has an F center and a CN$^-$ ion on nearest-neighbor halogen sites. Absorption into the F band leads to very efficient vibrational luminescence within the CN$^-$ ion, starting from a high vibrational state such as the fourth excited state. The luminescence occurs with a long lifetime, deexciting one vibrational state at a time, so that a series of lines is observed as the excited CN$^-$ works its way to the ground state. Calculations of transfer probabilities show that a dipole–dipole energy transfer is orders of magnitude slower than the observed radiative decay rate of the F center, and an exchange mechanism would not be appropriate in relating an electronic to a vibrational process. Thus, another mechanism must exist, whose nature at present eludes investigators.

2. EXTENDED DEFECTS

2.1 Definition

Extended defects are those that extend through a substantial part of the crystal in at least one direction. These defects, like point defects, can have a profound effect on the properties of a material. In fact, some material properties arise almost exclusively from the presence of extended defects. For example, plastic deformation would be almost entirely absent in metals at low temperatures if it were not for the presence of dislocations. Similarly, a large density of grain boundaries can cause the hardness of a material to increase by 100% or more, can alter the conductivity of semiconductors, and can result in significant enhancements in diffusion rates. Interfaces between different semiconductors can give rise to undesirable electronic effects that can greatly reduce the performance of devices. The growth of small cracks into macroscopic ones can cause an otherwise well-behaved material to fail catastrophically by fracture.

Each of these examples illustrates the role of extended defects in governing real material properties. Although here our primary emphasis is on extended defects in crystalline solids, we note that there are many analogies between the effects discussed here and those in related fields involving defects in other types of hosts. Examples range from the creep of vortex lattices in superconductors to the role of extended defects in dictating the crumpling of membranes. In this part of the article, we give a brief exposition of the definition and structure of extended defects, their response to external forces, and their effects on materials properties.

2.2 Types of Extended Defects

The goals of the study of extended defects within applied physics are to establish their basic structure and character, to explain their origins, to understand their effects on material properties, and eventually to control these effects. We begin with an overview of the basic types of extended defects that commonly occur in materials. The broadest classification is via dimensionality, with line and planar defects as the primary cases, and cracks treated separately.

2.2.1 One-Dimensional or Line Defects

The major types of line defects are dislocations and steps at surfaces or interfaces. Within crystals, the dominant line defects are dislocations. The nature of a dislocation is best illustrated via a picture. Figure 7 shows the two basic kinds of dislocations that are present in a three-dimensional crystal. The edge dislocation [Fig. 7(a)] results from slip of the upper half crystal relative to the lower half by a lattice translation vector, although the slip is confined to only a part of the plane. The translation vector, known as the Burgers vector, is a vector of the lattice, so that in the bulk of the half-plane the bonding is still perfect. For an edge dislocation, the Burgers vector is perpendicular to

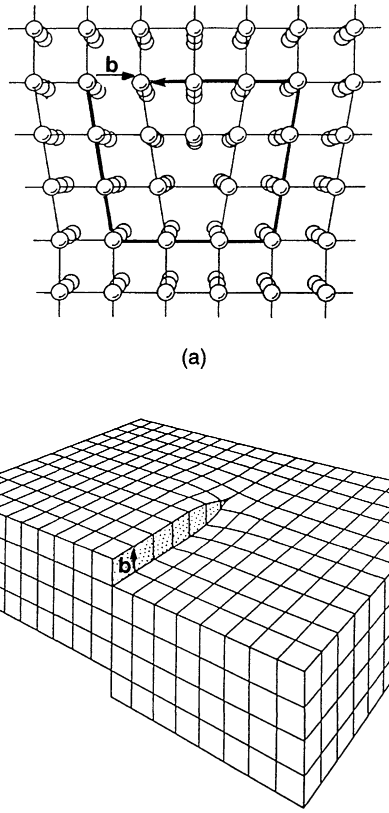

FIG. 7. Schematic drawings of edge and screw dislocations. (a) An edge dislocation viewed along the line direction. An extra plane of atoms is evident in the upper half of the crystal. (b) Illustration of a screw dislocation. The slip associated with the dislocation is indicated by the arrow, which denotes the Burgers vector.

the dislocation line and lies in a plane commonly referred to as the slip plane. Motion in the slip plane is called "glide"; motion out of the slip plane is known as "climb," and is typically more difficult than glide. A different type of structure, the screw dislocation [Fig. 7(b)], is obtained when the Burgers vector is parallel to the dislocation line. A dislocation will in general have mixed character with the Burgers vector neither parallel nor perpendicular to the dislocation line. Ultimately, a strict definition of a dislocation is that it is a boundary between slipped and unslipped regions of the crystal. From the standpoint of continuum mechanics, this is indicated by the fact that in traversing a contour around the dislocation line, the displacement field is found to be multivalued.

In general, crystal surfaces are not strictly planar, but rather include inhomogeneities such as steps between adjacent terraces. In Fig. 8, we show a stepped Si(100) surface as revealed by scanning tunneling microscopy. These steps on such a surface may be treated as one-dimensional defects. In some cases, the density of these steps is related to the slight deviation of a surface from its optimal orientation, with the steps accommodating this geometric mismatch. On the other hand, such steps may also arise as a result of thermal fluctuations at elevated temperatures.

2.2.2 Two-Dimensional or Planar Defects These occur as surfaces of materials, interfaces between regions of a crystal having different orientations (grain boundaries), and interfaces between distinct materials (heterojunctions).

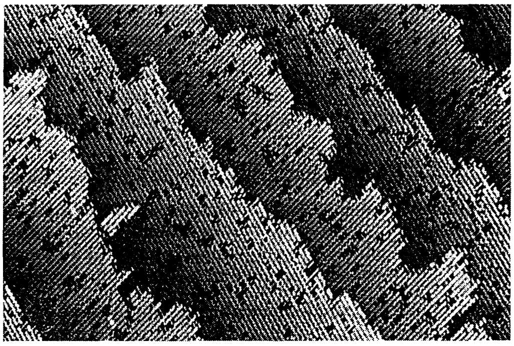

FIG. 8. Scanning tunneling microscopy image of a stepped Si(100) surface tilted by 0.4°. Steps separate flat terraces. Note the imperfections in the steps themselves, which are clearly jagged. After Alerhand *et al.* (1990).

A surface is essentially defined by its orientation and in some cases the type of atom (in compound materials) that terminates the surface. A grain boundary is defined in terms of the orientations of two adjacent parts of the crystal (grains) relative to the boundary, as well as possible translations of the grains relative to each other (see GRAIN BOUNDARIES). In Fig. 9, the change in directions of the lines of atoms, corresponding to grain misorientation, is clearly seen. One often distinguishes between low-angle boundaries, in which much of the interface material is perfectly bonded, and high-angle boundaries, in which the bonding is severely disrupted, though even for certain high angles there are special orientations resulting in a high degree of registry. There are special misorientations and translations for which the bonding at the boundary is not disrupted at all, which are known as stacking faults. Figure 10 shows a typical stacking sequence in which the stacking fault corresponds to a rearrangement of the stacking order. Planar defects are characterized by a hierarchy of energies with stacking faults having the lowest energies, high-angle grain-boundary en-

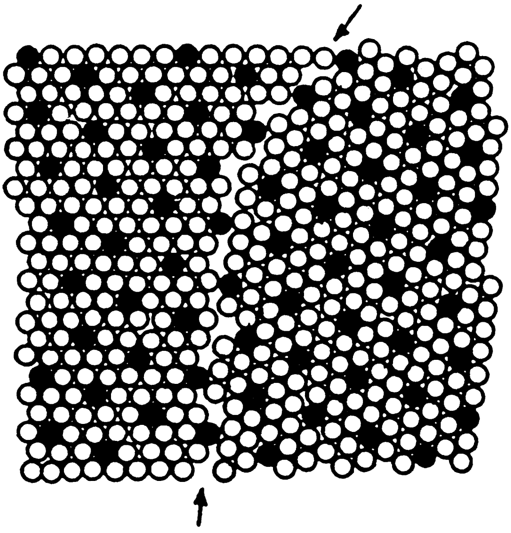

FIG. 9. A two-dimensional model of a pure tilt grain boundary. The arrows indicate the grain-boundary surface, which is perpendicular to the page. The coincident-site lattice is indicated by dark atoms—these form a lattice of sites common to both grains. After Murr (1975).

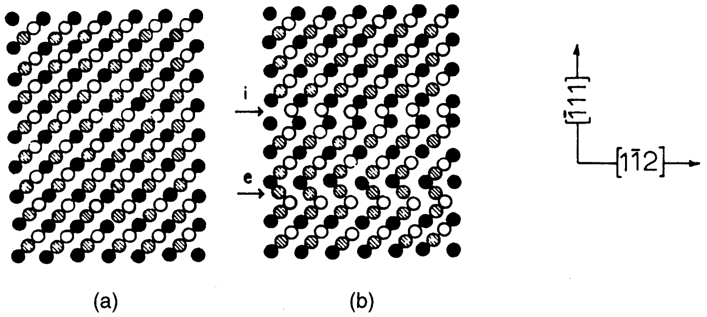

FIG. 10. (a) Regular stacking sequence (ABCABC . . .) of a perfect fcc lattice along a (111)-type direction. (b) Stacking sequence illustrating extrinsic (*e*) and intrinsic (*i*) stacking faults. The intrinsic stacking fault has a stacking sequence (ABCABABC . . .) and corresponds to a missing C-like layer. The extrinsic stacking fault has a stacking sequence (ABCBABC . . .) and corresponds to the addition of a B layer. After Murr (1975).

ergies falling in an intermediate range, and surfaces having the highest values. Typical stacking-fault energies are of the order of 100 mJ/m^2, while grain-boundary energies are 5 to 10 times higher. Surface energies typically fall in the range from 1500 to 3000 mJ/m^2. In compounds, one has defects known as antiphase boundaries or complex stacking faults, which are like stacking faults in that the interfacial bonding is not disrupted. However, the ordering between different atom types is disrupted across the interface. For example, in a material with only unlike-neighbor bonds, an antiphase boundary can lead to energetically unfavorable like-atom bonds.

2.2.3 Cracks Cracks may be thought of as having partly one-dimensional and partly two-dimensional character, as they involve the creation of surfaces. In the ideal case, the crack tip itself is defined by a line. An idealized crack is defined by the crack plane, the crack line, which bounds this plane, and the mode of loading that holds the crack open. In Fig. 11, we show an example of a crack illustrating the three major categories of loading geometry. Cracks are often found along interfaces or grain boundaries, probably because the energy of breaking the solid along these preexisting defects is less than that of breaking the perfect solid.

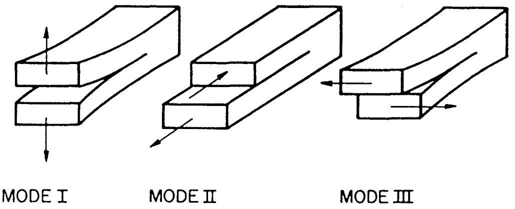

FIG. 11. Various pure loading geometries for an ideal planar crack. Mode-I loading corresponds to a pure tensile stress, while modes II and III correspond to loading by an applied shear stress. Loading is illustrated schematically by arrows. After Thomson (1986).

2.3 Origins of Extended Defects

Unlike many point defects, extended defects such as dislocations do not arise spontaneously in thermal equilibrium. The entropy gain associated with a thermal population of point defects suffices to compensate for typical point-defect formation energies, which are of the order of 1 eV. On the other hand, the formation energies of extended defects are essentially infinite, rendering the entropy contribution to their free energy insufficient to stabilize them. Thus, kinetic factors, possibly combined with work from external sources, must be responsible for their formation. In the case of dislocations, typical densities range from 10^4 cm^{-2} or less in high-quality semiconductors to 10^{12} cm^{-2} in highly deformed metals. The density of grain boundaries in polycrystalline materials is determined by the grain size, which is typically in the range of 5 to 10 μm.

Extended defects often occur directly as a result of the synthesis of a material. For example, after annealing, soft metals invariably have a density of 10^6 to 10^8 cm^{-2} of dislocations. An example of the origins of grain boundaries and interfaces is as a result of the impingement of particles that spontaneously nucleate out of a melt. Dislocations can occur when layered materials are synthesized by, for example, molecular-beam epitaxy, with the mismatch between the two materials leading to the formation of dislocation networks.

Various types of postsynthesis processing can also lead to the formation (and in some cases destruction) of extended defects. Processing methods involving deformation, such as rolling or cold working, lead to internal stresses that can activate dislocation sources such as "Frank–Read" sources. Figure 12 shows such a Frank–Read source in Si, where the contrast of the dislocation loops emerging from the source has been heightened by the use of "chemical decoration." From the central region, dislocation loops can be seen to emerge quite regularly. The operation of these sources results from a dislocation being locked to two points, which in turn causes the dislocation to bow out into a loop under the action of an applied stress. This loop eventually "ties off" and grows into a new dislocation. These sources can operate at very low stresses if the pinning-point separation is large. Deformation eventually leads to increased stress levels, which can activate progressively more and more sources.

Annealing of alloys can result in the formation of heterojunctions at interfaces between the nuclei of new phases and the host. Alternatively, extended defects can be destroyed by annealing processes. For example, dislocations can migrate to the surface of the material and leave it there. Grain boundaries can migrate, resulting in the coalescence of grains and thus reduced grain-boundary area.

2.4 Structure

The structure of extended defects is usually described at large distances in terms of a long-ranged elastic strain field and at short distances in terms of the atomic arrangements in some core region. The elastic field can often be obtained analytically, but the atomic-level bonding requires the use of computer simulations.

The elastic distortions associated with dislocations are characterized by strain fields that decay as $1/r$, where r is the distance from the defect. In the case of straight screw dislocations, the strain field is cylindrically symmetric, but for edge dislocations it is more complicated. Using isotropic elasticity theory, these strain fields lead to an elastic strain energy per unit length of the form

$$E/L = C(\mu \mathbf{b}^2/4\pi)\ln(R/r_0), \qquad (4)$$

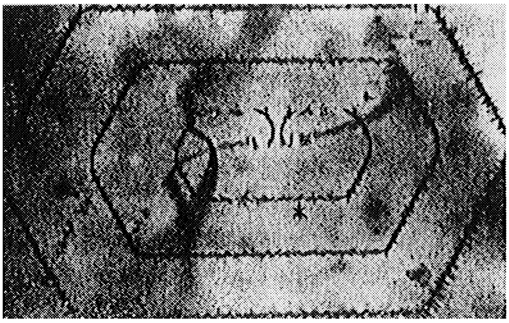

FIG. 12. Series of expanding dislocation loops emerging from a Frank–Read source in Si. The two pinning points about which the dislocation segments bow out are in the center of the figure. After Hull (1965).

where C is a prefactor determined by Poisson's ratio, μ is the shear modulus, $\mathbf{b}$ is the Burgers vector of the dislocation, R is a large-distance cutoff corresponding to the radius of a cylindrical sample containing the defect region, and r_0 is a core cutoff parameter; both R and r_0 are necessary in order to obtain a finite value of E/L. Further details may be found in Hirth and Lothe (1982). One critical atomic-level feature of dislocations is the nature of their dissociation, if any, into partial dislocations (see Fig. 13). Dissociation is driven by the energetic gain

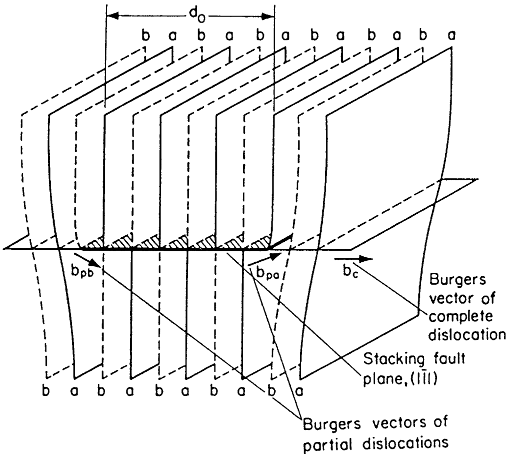
d_0
b a b a b a b a b a b a
b_{pb}
b_{pa}
b_c
Burgers vector of complete dislocation
Stacking fault plane,$(1\bar{1}1)$
b a b a b a b a b a
Burgers vectors of partial dislocations

FIG. 13. Illustration of the geometry of partial dislocations in a fcc crystal. The Burgers vectors of the partial dislocations are related to that of the full dislocation by $\mathbf{b}_c = \mathbf{b}_{pb} + \mathbf{b}_{pa}$. Note that $\mathbf{b}_{pb}$ and $\mathbf{b}_{pa}$ both involve out-of-plane displacements. After Hull (1965).

associated with having two dislocations with small Burgers vectors as opposed to a single one with a large Burgers vector, according to the $\mathbf{b}^2$ term in Eq. (4) (although the estimates of elasticity become more complex when the elastic anisotropy of the medium is taken into consideration). The two dislocations then have Burgers vectors that are not lattice vectors, which results in the terminology "partial dislocations." Partial dislocations are of necessity terminations of planar faults, which are stacking faults in elemental materials and antiphase boundaries or complex stacking faults in compounds. The energy penalty for these planar faults tends to reduce the separation between the partial dislocations. The elastic interaction between the partials is repulsive, on the other hand, resulting in a balance of forces at the equilibrium separation. More complex types of core dissociation can also occur. For example, it is believed that screw dislocations in bcc metals are spread onto three distinct planes, which results in their being hard to move.

Dislocations in real materials are not straight, but rather contain "kinks" and "jogs" (see Fig. 14). A kink is a perturbation of the dislocation line in the slip plane; a jog is a perturbation out of the slip plane. As will be discussed below, kinks are believed to be very important in the low-temperature mobility of dislocations in their slip planes, while jogs become relevant to dislocation motion out of their slip planes at high temperatures.

More complex features of dislocation core structure involve possible "reconstructions" (cf. Fig. 15). Reconstruction means that there are distortions of the atomic geometry in the core region that are parallel to the line direction of the dislocation. These distortions result in the formation of a repeat distance along the line direction that exceeds the repeat distance associated with the unreconstructed structure. The simplest example is demonstrated in Fig. 15, where in Fig. 15(a) the unreconstructed structure is shown and in Fig. 15(b) the "dimerization" is shown. This reconstruction involves alternating pairs of atoms, which pull closer to and farther away from each other. Such reconstructions are thought to occur in Si.

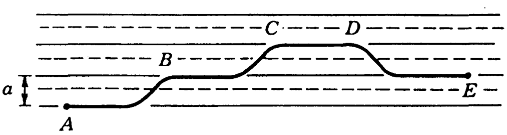

FIG. 14. Schematic illustration of the formation of kinks in a dislocation line. B, C, and D are points at which kinks have formed, allowing segments of the dislocation to move from one potential well to another. After Hirth and Lothe (1982).

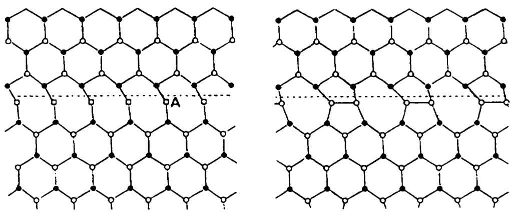

(a) (b)

FIG. 15. Projections of the {111} slip plane for the diamond cubic structure illustrating (a) unreconstructed and (b) reconstructed core structures for the 30° partial dislocation. Note that the reconstruction restores the conventional tetrahedral bonding configuration. Adapted from Duesbery and Richardson (1990).

In general, the elastic fields of interfaces are short ranged. When viewed as an array of dislocations, a simple low-angle grain boundary may be characterized by stress fields that decay exponentially away from the boundary. The atomic-scale structure of interfaces is best attacked by starting with grain boundaries. Ideal planar grain boundaries are commonly categorized into "tilt" and "twist" boundaries according to whether the axis of the rotation that takes one grain's orientation into the other is parallel or perpendicular to the boundary. Just as edge and screw dislocations are the simple limits of dislocation character, tilt and twist character to grain boundaries are special cases of more general boundary character that have both twist and tilt components. A small-angle tilt grain boundary, as indicated in Fig. 9, can be thought of as an array of edge dislocations. Above each edge dislocation is an extra plane between the two grains; the increasing number of planes between the grains as one goes up in the picture corresponds to a tilt between the two grains, with the number of such dislocations per unit length dictating the misorientation between the grains. Similarly, a small-angle twist boundary may be thought of as an array of screw dislocations.

The coincident-site lattice (CSL) is a useful conceptual structure for analyzing commensurate boundaries in which a unit cell (possibly large) can be repeated to form the boundary structure. The number of atoms in the unit cell is known as the "Σ" number, so that, for example, a $\Sigma5$ grain boundary has five atoms per interface unit cell. A high Σ number corresponds to a complex boundary structure. The coincident-site lattice (CSL) is the collection of the lattice points that both of the grains have in common, with $1/\Sigma$ giving the fraction of sites common to both lattices. If the boundary unit cell is large, then the coincident-site lattice has a large lattice constant and correspondingly a low density. In Fig. 9, the coincident-site lattice is indicated by the filled circles.

For analyzing more general high-angle boundaries, a "structural-unit" model has been found to be useful. In this model, the structure of a number of high-Σ boundaries (and incommensurate ones as well) is created from two or more distinct small building blocks that are derived from the structures of low-Σ boundaries. The complexity in the high-Σ structures then comes from the arrangements of these building blocks into larger unit cells. Atomistic simulations have shown that many complex grain-boundary structures can be modeled quite well in this fashion. Grain-boundary structure is more completely discussed in GRAIN BOUNDARIES.

At heterojunctions, mismatch between the lattice constants of the constituents leads to disruption of the interfacial bonding. If the mismatch is small, this disruption may take the form of a network of misfit dislocations having at least some edge character, where the extra planes are inserted on the side with the smaller lattice constant. Figure 16 shows such misfit dislocations at a CdTe–GaAs interface; the dislocations appear as light spots.

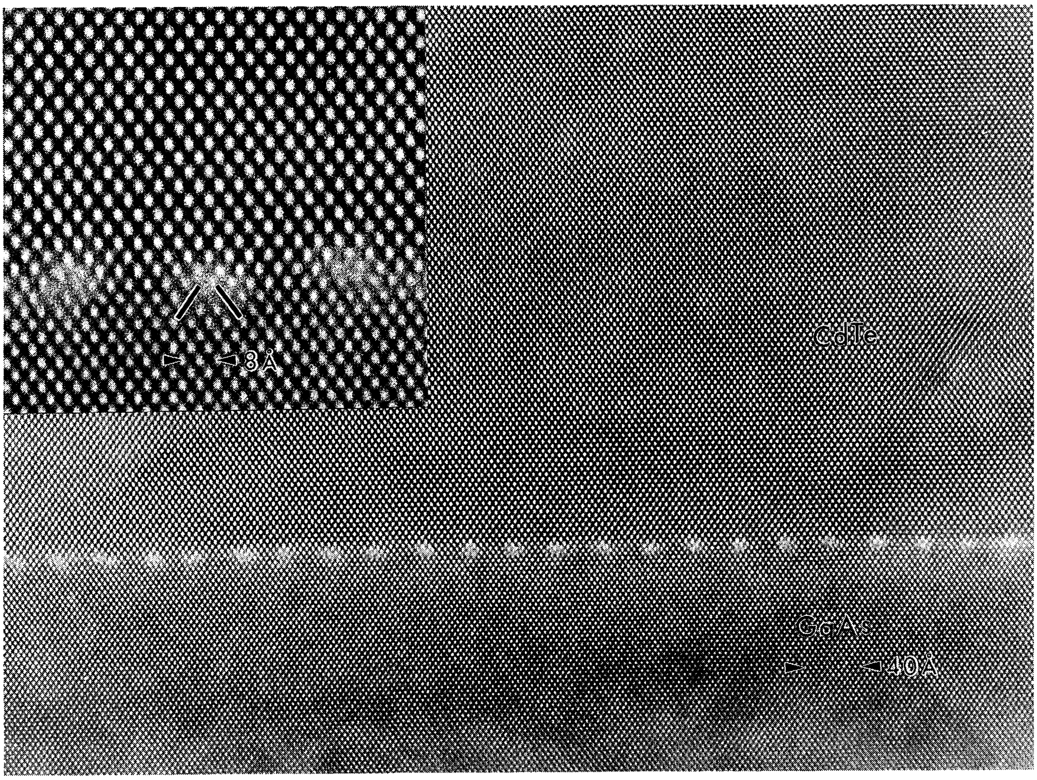

FIG. 16. High-resolution electron micrograph showing a periodic array of Lomer misfit dislocations for an annealed CdTe/GaAs interface. Inset magnifies the two half-planes associated with the perfect pure edge Lomer dislocation (courtesy of A. F. Schwartzman).

At surfaces, the perturbation of the bonding is much greater than at grain boundaries or interfaces. This leads, in addition to the expected relaxation of surface layers into or away from the bulk, to symmetry-breaking reconstructions that can be quite complex. For example, the Si (111) surface displays a "7×7" reconstruction (see Fig. 17) in which both dimensions of the surface unit cell are seven times larger than they would be in the absence of reconstruction. Some types of reconstruction can be explained via simple bond-healing effects, but others involve more complex electronic effects such as the splitting of bands of electronic surface states or environmental dependence of the interatomic forces.

At large distances, the structure of cracks, like dislocations, is best described in terms of associated continuum fields. In elasticity theory, the stress field (which is proportional to the strain) $\sigma(r)$ of a straight, sharp, planar crack has a $1/\sqrt{r}$ behavior, where r is the distance (measured in two dimensions) from the crack line. This is conveniently described in terms of a stress-intensity factor K, where

$$\sigma(r) \propto K/\sqrt{2\pi r}.$$

For details, see Thomson's review article on fracture (Thomson, 1986). For real cracks, this analysis is valid far from the crack tip. Closer to the crack tip, the stress field can often be described by a local stress-intensity factor that accounts for the effects of surrounding dislocations. The continuum description certainly must break down at distances of the order of interatomic spacings. In addition, the crack may well not be sharp on an atomic scale. Some brittle materials can sustain atomically sharp cracks but many others invariably have blunt cracks. Figure 18(a), for Si, shows no dislocation activity indicative of crack-tip deformation. Figure 18(b), for Fe, shows a crack that is blunted by the emission of dislocations. In addition, for real cracks the crack front is not straight, and the crack itself is nonplanar. Indeed, the deviation of the crack surface from strict planarity can be used to quantify the intrinsic ductility of a given material. Some of these effects can be described in terms of kinks in the crack line, which may serve as dislocation sources. These "de-

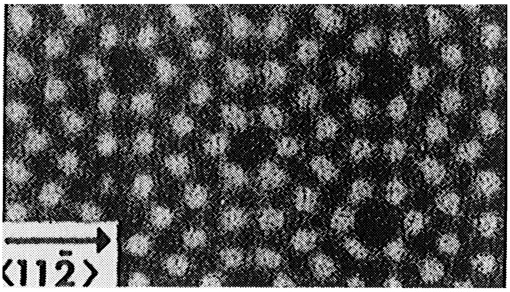

FIG. 17. Scanning tunneling microscopy image of the Si(111) (7×7) reconstructed surface. After Tromp *et al.* (1986).

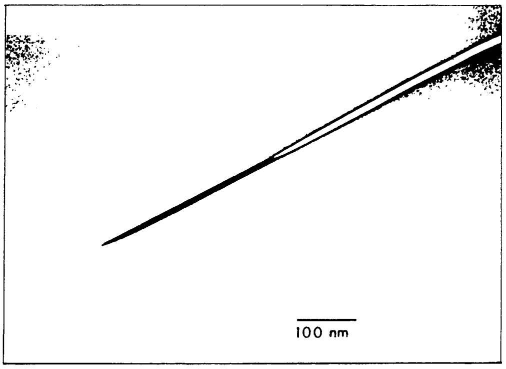

(a)

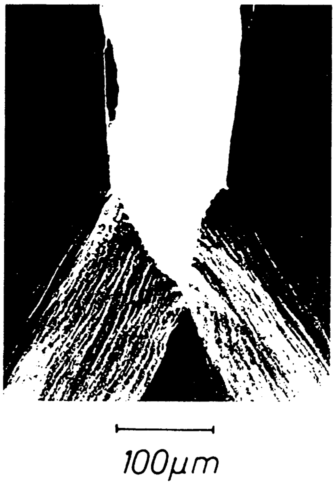

(b)

FIG. 18. (a) Sharp crack tip in Si. After Lawn (1975). (b) Blunted crack in Fe. After Vehoff and Neumann (1980).

fects with a defect" must certainly be included in a complete analysis of crack stability and propagation.

2.5 Motion and Response to External Forces

Although defects are not physical objects in the same sense as atoms themselves, it is still possible to define forces on defects precisely. The defect force is defined as

$$F = -\nabla U, \tag{5}$$

where U is the free energy of the entire system including the defect. The position of the defect does not vary continuously, but can only move by lattice translations. However, provided that U varies slowly on a length scale corresponding to the lattice vectors, the gradient in Eq. (5) can still be defined in a discrete sense, by calculating the energy change corresponding to a virtual displacement equal to a lattice vector.

The force on a dislocation per unit length due to a stress field (the Peach–Koehler force) is the following:

$$\mathbf{F}/L = (\mathbf{b}\cdot\boldsymbol{\sigma}) \times \boldsymbol{\xi}, \tag{6}$$

where $\mathbf{b}$ is the Burgers vector, $\boldsymbol{\sigma}$ is the stress tensor, and $\boldsymbol{\xi}$ is the unit vector parallel to the dislocation line. This equation is discussed in more detail in Hirth and Lothe (1982). Since the stress field of a dislocation decays as $1/r$, this implies that the force between dislocations also decays as $1/r$. The sign of the force is repulsive between parallel straight dislocations having the same Burgers vector, and attractive if the Burgers vectors are opposite. Counteracting the Peach–Koehler force is the frictional or "Peierls" force, which arises because the dislocation must go over an energy barrier in going from a local free-energy minimum to a neighboring one. This means that a minimum "Peierls stress" is required to move the dislocation. In close-packed metals, the Peierls stress is generally very small, typically on the order of $10^{-4}\mu$ or $10^{-5}\mu$, where μ is the shear modulus. In covalently bonded materials, it can be a few orders of magnitude larger; bcc metals typically have values intermediate between the close-packed metals and covalently bonded materials. In general, at finite temperatures the motion of a dislocation is not uniform, but rather consists of the activated motion of kinks parallel to the dislocation line (cf. Fig. 14). These are nucleated as kink–antikink pairs. The motion of a single kink along the whole dislocation line leads to the forward motion of the dislocation line by one Burgers vector. In covalently bonded materials, and bcc metals, the kink nucleation and migration energies are high enough to give rise to a strong exponential dependence of the dislocation velocity on temperature.

The motion of dislocations according to Eq. (6) gives rise to plastic deformation in the direction of the applied stress. For example, for a nonvanishing σ_{xz} component of the stress tensor, a dislocation parallel to the y axis, with a Burgers vector in the x direction, will experience a force in the x direction. Its motion will move an extra half-plane in that direction, resulting in plastic deformation that relieves the applied stress.

Another mechanism for plastic deformation that is related to the migration of dislocations is that of interfacial migration. The motion of interfaces may be broadly classed into two categories: grain-boundary migration, which involves a motion of the boundary perpendicular to the plane of the interface, and grain-boundary sliding, which results upon relative translation of adjacent grains along a direction parallel to their common interface. The relation between grain-boundary and dislocation migration is argued on the basis of models that hold that grain boundaries may be thought of as being built up of dislocations. A host of experiments on both migration and sliding reveal the dependence of these processes on a variety of parameters ranging from geometrical quantities such as the angle of misorientation to chemical effects such as the concentration of impurities.

The force on a crack is most easily understood in the mode I geometry of Fig. 11. There are two contributions to the force. The first is the elastic strain energy. Under a virtual displacement of the crack tip to the right, the elastic strain energy is reduced because the surface moves in the direction of the applied stresses, and thus the energy associated with the external forces is lowered. Thus, the elastic strain energy tends to expand the crack. The corresponding force per length of crack line is proportional to $\sigma^2 a/\mu$, where a is the crack length. The second contribution to the force comes from the surface energy. As the crack tip moves to the right a distance δx, there is an energy increase by an amount $2\gamma_s\delta x$, corresponding to the creation of new free surface, where γ_s is the surface energy per unit area (the prefactor of two arises because two surfaces are created). Thus, the force from this contribution is simply $2\gamma_s$, to the left. The equilibrium crack length is determined by the competition between these two forces. In terms of the stress-intensity factor K discussed above, the equilibrium condition becomes (for mode I loading) the "Griffith criterion"

$$K_c = \sqrt{4\gamma_s\mu}, \qquad (7)$$

where μ is the shear modulus. The criteria for modes II and III are similar, but with prefactors that differ slightly from unity.

Under a given applied stress σ the stress-intensity factor for a crack of length a has the form

$$K \propto \sigma\sqrt{a}. \qquad (8)$$

This means that a long crack will have an increased stress-intensity factor, and thus will be more likely to grow.

2.6 Effects on Materials Properties

The effect of extended defects on the properties of materials is an extremely rich and complex topic, and we can only give some representative examples here.

2.6.1 Mechanical Properties The crucial importance of extended defects in explaining these properties is demonstrated by the failure of models based on perfect crystalline solids. In these models, the ideal shear strength of a solid is computed by evaluating the energy to shift the upper half of a crystal rigidly relative to the lower half along a shearing plane. The energy of such a process is periodic as a function of displacement. It was first modeled by assuming a sinusoidal form for the periodic energy profile. The stress predicted by this model to shear a perfect crystal is $\tau_{\mathrm{ideal}} = \mu b/2\pi d$, where μ is the shear modulus, d is the distance between the planes being sheared, and b is the period length in the slip direction. A comparison of the ideal shear strength with the observed shear strength for a variety of materials indicates, however, that this idealized model of plastic deformation drastically overestimates the stress required to deform a crystal permanently. The "Peierls" stress required to move a dislocation, as noted above, is much smaller.

Since dislocations are the "carriers" of plastic deformation, one might be tempted to think of them as softening agents. However, the hardness of materials usually increases with the dislocation density. This is because the dislocations give rise to spatially varying stress fields, which create energy barriers that inhibit the motion of dislocations. This is the origin of the phenomenon of work hardening. As a material initially containing a small density of dislocations is deformed, progressively more and more dislocation sources become active and the dislocation density increases. This increases the energy barriers that must be traversed for dislocation motion, and the material becomes harder. As a result, the shear strength of the material becomes roughly inversely proportional to the average dislocation spacing.

A material can also be strengthened by having a small grain size, which inhibits the motion of dislocations. In addition, grain boundaries can make a material either more or less brittle, depending on the grain-boundary chemistry and the grain size. Several impurity elements, such as H and S, can segregate to grain boundaries in metals and thereby weaken them, resulting in brittleness. Much of the discussion of this effect has been based on the assumption that the grain-boundary bond strength is reduced by the impurities, but this scenario is by no means proven. A small grain size can reduce stress concentrations in a material, thereby making it less brittle. Finally, fine-grained materials can deform at high temperatures via a mechanism known as superplastic deformation. In this mechanism, grains slide over each other, and mass transport along the grain boundaries results in changes in grain shape, which corresponds to deformation of the material.

The presence of small cracks in a material can have the effect of making the material more brittle. As with dislocations, the necessity for preexisting cracks in explaining brittle fracture is shown by the failure of models based on defect-free solids. In these models, the ideal tensile strength is computed analogously to the ideal shear strength. Estimates of the tensile strength derived in this way again severely overestimate the fracture stress as determined experimentally. Cracks act as "concentrators" of applied or internal stresses, so that in the region near a crack tip these stresses may reach values comparable to the ideal tensile strength. The key issue here is whether the cracks grow under stress or not. As discussed above, at a given stress level, large cracks tend to grow more than small ones. Thus, one hopes that the cracks in a material are below the critical length for growth under the intended conditions of use. It is often observed that the critical stresses for crack growth significantly exceed the Griffith value given above [cf. Eq. (7)]. This is because the work done on the crack can be dissipated via toughening mechanisms such as plastic deformation, whose energy can greatly exceed the surface energy. Inherently brittle materials can be improved via toughening mechanisms such as stress-relieving structural transformations ahead of the crack, toughening fibers, or even subcritical microcracks that shield the applied stresses.

2.6.2 Mass Transport Surfaces, dislocations, and grain boundaries all provide low-energy paths for atomic migration. This effect is known as "pipe diffusion" in the case of dislocations. In part, this effect may be attributed to reduced point-defect formation and migration energies in the core region of the extended defect. In most cases, high-temperature diffusion is dominated by bulk effects, because of the great number of possible paths. Diffusion along extended defects becomes more important at low temperatures, because the higher-energy bulk diffusion paths are "frozen out." The spectra of observed diffusivities for bulk lattices and along extended defects are contrasted in Fig. 19. This figure not only illustrates the wide range of observed diffusivities, but also reveals the variety of activation energies associated with diffusion at extended defects.

2.6.3 Acoustic Attenuation Extended defects (in addition to point defects) provide a variety of "internal friction" mechanisms that result in the dissipation of acoustic energy. Point defects as well as linear and extended defects give rise to acoustic attenuation in characteristic temperature and frequency regimes. For example, in the case of dislocations, acoustic energy is converted to the energy of forming kink–antikink pairs along the dislocation. This dissipation mechanism gives rise to the so-called Bordoni peak in the spectrum of acoustic attenuation.

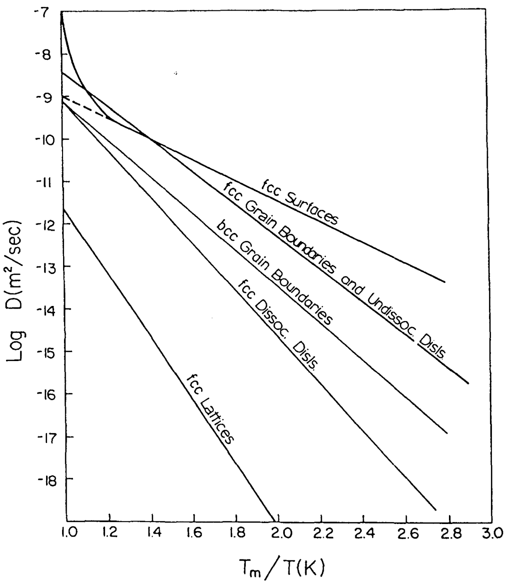

FIG. 19. Diffusivity spectrum for metals as deduced by Gjostein. This schematic illustrates the enhancement in diffusion rates seen at surfaces, grain boundaries, and dislocations in comparison with perfect lattice diffusivity. Taken from modified figure of Balluffi (1984).

2.6.4 Electrical Properties The conductivity effects are particularly pronounced in semiconductors. Extended defects can scatter carriers by reason of their deviation from bulk crystalline geometry. They can also cause defect-localized states, due either to the intrinsic defect geometry, or to the segregation of impurities and other imperfections to the defect. Such states can provide electron–hole recombination paths, which greatly reduce the carrier lifetime. In addition, the occupation of the defect states by charged carriers can result in electrostatic fields that locally shift the valence and conduction bands relative to the Fermi level, and can create barriers for charge transport. For example, grain boundaries in Si can have barriers of several tenths of an electron volt, which strongly inhibit transport. These effects are particularly important in samples with small grain sizes. Similar effects occur at interfaces between different semiconductors. Dislocations in several types of semiconductors have been shown to cause characteristic defects of acceptor character, which can locally change the character of the material from *n*-type to *p*-type. Defect states associated with dangling bonds in semiconductors can be passivated by hydrogen, thereby removing states from the gap and reducing charge trapping.

2.6.5 Magnetic Effects Grain boundaries can have a significant impact on magnetic properties, via their interaction with domain walls. The magnetic permeability of polycrystalline materials is mainly due to the motion of domain walls. This motion is generally inhibited by grain boundaries, resulting in reduced permeabilities. In very fine-grained materials, the grain size is smaller than the domain size, so that there are no domain walls (or the domain walls coincide with the grain boundaries). In this case, the permeability is reduced to the contribution from the spins themselves.

2.6.6 Interaction Effects Many properties are also strongly affected by the interaction of point defects with extended defects. As mentioned above, the interaction of point defects with extended defects can lead to dramatic enhancements in the rates of diffusion in materials. In addition, a wide range of other phenomena are affected as well. In semiconductors, doping by either *n*- or *p*-type point defects results in a change in the Fermi level, which in turn can alter the dislocation mobility (as measured by their velocities). A second relation between point and extended defects relates to the mechanism of dislocation motion. Besides the conventional dislocation glide described earlier, dislocations can climb out of their slip plane through the absorption or emission of point defects. This mechanism is generally observed to begin closer to the melting temperature, and has been invoked as a dominant mechanism in explaining creep phenomena. The elastic interactions between point defects and dislocations can lead to the pinning of dislocations. Such pinning has the consequence that the stress needed to activate dislocation motion is larger than that in the point-defect–free case. This effect is referred to as solution hardening. Finally, point defects are known to segregate to extended defects. In particular, segregation to grain boundaries can either strengthen or weaken an interface, leading to an overall change in the mechanical response of a material.

3. CONCLUDING REMARKS

The understanding of the full complexity of real materials is ultimately predicated upon the ability to predict and control the properties of both point and extended defects. Though our treatment has emphasized point defects, line defects, and planar defects each separately, it is only when taken in conjunction that the full implications of such defects can be realized. For example, we have noted the rich interplay between the diffusion of point defects and the presence of extended defects such as dislocations and grain boundaries that provide low-energy pathways for point-defect migration. Dramatic enhancements in the rate of point-defect migration were seen to be related to the presence of extended defects such as dislocations and grain boundaries. Conversely, the motion of dislocations is strongly inhibited by the presence of point defects. Similarly, realistic treatments of processes near a crack tip must involve a description of the dislocations that accompany plastic deformation in the near-tip region. Finally, the electronic properties of planar semiconductor devices are strongly affected by the interplay of defects and interfaces; for example, the defect known as a P_b center is a threefold-coordinated silicon atom at the silicon–silicon dioxide interface. In each of these cases, it is clear that the behaviors of point and extended defects are inextricably linked. The synthesis of our understanding of these defects and their interaction will undoubtedly remain an important part of applied physics.

GLOSSARY

Antisite Defect: An atom on the "wrong" site in a crystal that contains more than one type of chemical constituent.

Atomic Diffusion: The motion of an atom over a macroscopic distance as the result of a series of individual microscopic motions that are statistical in nature.

Bloch Waves: Plane waves in a crystal modulated by functions that have the periodicity of the lattice.

Bordoni Peak: A feature in acoustic attenuation associated with the formation of kink–antikink pairs.

Burgers Vector: The translation vector associated with a dislocation.

Chemical Impurity: An atom, ion, or (small) molecule that is chemically different from the perfect-crystal constituents.

Climb: Dislocation motion out of the plane of the Burgers vector (the slip plane).

Cluster or Nanocrystal: A type of defect that is intermediate between point and extended defects.

Coincident-Site Lattice (CSL): The collection of the lattice points that both of the grains in a grain boundary have in common.

Color Centers: Point defects in insulators that generate optical activity in the visible.

Configuration-Coordinate Diagram: Plot of energy vs atomic position for a molecule or defect.

Cracks: May be thought of as having partly one-dimensional and partly two-dimensional character, as they involve the creation of surfaces. An idealized crack is defined by the crack plane. In the ideal case, the crack tip itself is defined by a line, the crack line, which bounds this plane.

Deep Defects: Defects in semiconductors or insulators in which the electron or hole is strongly bound and is not likely to be unbound at room temperature.

Dislocation: A line defect in a crystal that represents the boundary between the slipped and the unslipped parts of the crystal.

Doping: The introduction of chemical impurities into a crystal during the growth process.

Energy Transfer: Transfer of excitation energy from one defect to another.

Extended Defects: Defects that extend through a substantial part of the crystal in at least one direction.

F Center in Alkali Halides: An electron trapped at a halogen$^-$ vacancy.

Franck–Condon Principle: Optical absorption or emission in which the energies and wave functions are consistent with the positions of the ions before the transition took place.

Frank–Read Source: Source of dislocation loops under the action of an applied stress.

Frenkel Defect: A vacancy-interstitial pair.

Glide: Dislocation motion in the plane of the Burgers vector (the slip plane).

Grain Boundary: A planar defect that is defined in terms of the orientations of two adjacent parts of the crystal.

Heterojunction: An interface between two materials, or two phases of the same material.

Hydrogen Passivation: The change of the nature of the electrical properties of defects by the incorporation of hydrogen into the defect.

Interstitial: An atom or ion at a position other than a perfect-crystal site.

Ion Implantation: The introduction of chemical impurities by firing a beam of impurities into a crystalline sample under controlled conditions.

Jog: A perturbation of a dislocation line out of the slip plane.

Kink: A perturbation of a dislocation line in the slip plane.

Metastable Defect: A defect that has more than one arrangement of atoms.

Molecular-Beam Epitaxy (MBE): Deposition of atoms from molecular beams, layer by layer, onto a substrate, thereby forming thin films.

***n*-Type:** A semiconductor with an excess of donors.

Negative-*U*: A situation in which the "effective" Coulomb interaction between electrons or holes is negative.

***p*-Type:** A semiconductor with an excess of acceptors.

P_b Center: A threefold-coordinated silicon atom at a silicon–silicon dioxide interface.

Peach–Koehler Force: The force on a dislocation per unit length due to a stress field.

Peierls Force: The frictional force that arises because a dislocation must go over an energy barrier in going from a local free-energy minimum to a neighboring one.

Perfect Crystal: A periodically repeated array of identical building blocks such as atoms or arrays of atoms.

Pipe Diffusion: Atomic migration along a dislocation.

Plastic Deformation: Deformation of a crystal under applied stress. This deformation remains after the stress is removed.

Point Defect: An imperfection in a perfect crystal that is of the order of atomic size.

Reconstruction: Distortions parallel to the dislocation line that result in the formation of a unit cell in this direction that is larger than the original one.

Schottky Defect: An isolated vacancy in an otherwise perfect crystal.

Self-Trapped Exciton, or STE: An electron–hole pair in a nonmetallic crystal that is fixed in space by lattice relaxation.

Sensitized Luminescence: A situation involving two different defects, one of which (the sensitizer) absorbs light in a desired wavelength range, the other of which (the activator) emits light in a different wavelength range.

Shallow Donors and Acceptors: Defects in semiconductors in which the electron (donors) or hole (acceptors) is weakly bound and is likely to be unbound at room temperature.

Stacking Fault: A planar defect in which the regular stacking sequence (ABCABC . . .) is interrupted.

Stokes Shift: The energy difference between absorption and emission bands.

Substitutional Defect: A defect that occurs on a perfect-crystal atom site.

Tilt Grain Boundary: The axis of the rotation that takes one grain's orientation into the other is parallel to the boundary.

Twist Grain Boundary: The axis of the rotation that takes one grain's orientation into the other is perpendicular to the boundary.

Vacancy: The absence of an atom or ion from its perfect-crystal site.

Zero-Phonon Transition: An optical transition from the zero-point level of one state to the zero-point level of another state.

Works Cited

Alerhand, O. L., Berker, A. Nihat, Joannopoulos, J. D., Vanderbilt, David, Hamers, R. J., Demuth, J. E. (1990), *Phys. Rev. Lett.* **64**, 2406–2409.

Balluffi, R. W. (1984), in: G. E. Murch, A. S. Nowick (Eds.), *Diffusion in Crystalline Solids*, New York: Academic.

Borg, Richard J., Dienes, G. J. (1988), *An Introduction to Solid-State Diffusion*, San Diego: Academic.

Duesbery, M. S., Richardson, G. Y. (1990), *Crit. Rev. Solid State Mater. Sci.* **17**, 1–46.

Feder, R., Charbnau, H. P. (1966), *Phys. Rev.* **149**, 464–471.

Hirth, J. P., Lothe, J. (1982), *Theory of Dislocations*, New York: Wiley.

Hull, D. (1965), *Introduction to Dislocations*, New York: Pergamon.

Lawn, B. (1975), *Fracture of Brittle Solids*, New York: Cambridge.

Murr, L. E. (1975), *Interfacial Phenomena in Metals and Alloys*, London: Addison-Wesley.

Thomson, R. (1986), in: H. Ehrenreich, D. Turnbull (Eds.), *Solid State Physics*, Vol. 39, New York: Academic.

Tromp, R. M., Hamers, R. J., Demuth, J. E. (1986), *Phys. Rev.* **B34**, 1388–1391.

Vehoff, H., Neumann, P. (1980), *Acta Metall.* **28**, 265–272.

Further Reading

Ashcroft, N. W., Mermin, N. D. (1976), *Solid State Physics*, New York: Holt, Rinehart and Winston.

Davies, G., DeLeo, G. G., Stavola, M. (Eds.) (1992), *Proceedings of the 16th International Conference on Defects in Semiconductors*, Zurich: Trans Tech.

Flynn, C. P. (1972), *Point Defects and Diffusion*, Oxford: Clarendon.

Fowler, W. Beall (Ed.) (1968), *Physics of Color Centers*, New York: Academic.

Fowler, W. Beall, Itoh, Noriaki (Eds.) (1990), *Atomic Processes Induced by Electronic Excitation in Non-Metallic Solids*, Singapore: World.

Hayes, W., Stoneham, A. M. (1985), *Defects and Defect Processes in Nonmetallic Solids*, New York: Wiley.

Kittel, C. (1986), *Introduction to Solid State Physics*, 6th ed., New York: Wiley.

Luty, Fritz, Dierolf, Volkmar (1993), in: O. Kanert, J. M. Spaeth (Eds.), *Defects in Insulating Materials*, Singapore: World, pp. 17–38.

Nabarro, F. R. N. (Ed.) (1979–), *Dislocations in Solids* (series), Amsterdam: North Holland.

Wolf, Dieter, Yip, Sidney (Eds.) (1992), *Materials Interfaces, Atomic-Level Structure and Properties*, London: Chapman and Hall.

POLAROGRAPHY, VOLTAMMETRY, AND RELATED TECHNIQUES

Lubomír Pospíšil, *J. Heyrovský Institute of Physical Chemistry, Academy of Sciences of the Czech Republic, Prague, Czech Republic*

	Introduction	371
1.	**General Principles**	371
2.	**Polarography in Its Classical Form**	372
2.1	Electrochemical Kinetics in Polarography	374
3.	**Derived Polarographic Techniques**	376
3.1	Pulse Polarography	376
3.2	ac Polarography	377
4.	**Voltammetry**	378
5.	**Experimental Equipment**	380
5.1	Working Electrodes, Solvents, Indifferent Electrolytes	380
5.2	Instrumentation	381
6.	**Applications**	381
	Works Cited	381
	Further Reading	382

INTRODUCTION

The electrochemical method called "polarography" was introduced in 1922 by a Czech scientist, Jaroslav Heyrovský, professor of physical chemistry at Charles University in Prague. In the early 1920s, Heyrovský studied the surface tension of mercury in aqueous solutions of electrolytes as a function of applied electrical voltage (so called "electrocapillarity"). In one of the experimental arrangements, the surface tension was evaluated from intervals between falling drops of mercury that slowly flowed out of a vertical glass capillary connected with a mercury reservoir above it. At this point, an ideally reproducible electrochemical sensor was discovered that featured a periodically renewable, clean, and smooth electrode surface made of metallic material permitting measurements at extreme negative potentials. Having this tool in hand, Professor Heyrovský attempted to explain certain anomalies observed on curves of the surface tension vs potential. Later he began to measure the electric current flowing through the mercury electrode as a function of externally applied potential. Measurement of current-potential curves with the dropping mercury electrode immersed in a solution containing dissolved reducible or oxidizable species is the principle of the polarographic method (Heyrovský, 1922). The discovery of polarography as a powerful analytical and research method for investigation of the most abundant reactions in Nature, redox reactions, was appreciated by the scientific community by the award to its inventor of the Nobel prize for chemistry in 1959.

1. GENERAL PRINCIPLES

The description of the virtues of polarography assumes some essential knowledge of principles of electrochemical reactions and cells, which we review first. This article deals with the transfer of charge across the interface of a metallic electrode and a liquid solution containing dissolved ions or neutral molecules. Such a transfer of charge is observed as the flow of electrical current, and the goal is to relate the observed current to the elementary processes at the interface and in the solution (Vetter, 1961; Thirsk, 1972; Albery, 1975). For principal reasons we cannot handle an isolated boundary of a single electrode because the electric current can flow only in a closed circuit. We have to place in the solution at least another electrode (often called the counter electrode, or the auxiliary electrode), forming thus an electrochemical cell. The two electrodes are connected to appropriate external circuitry that controls the potential difference be-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

tween the electrodes and monitors the current resulting from electron-exchange processes at the two interfaces. The overall process taking place in the cell then consists of two half-reactions representing actual chemical changes invoked by the charge transfer and occurring at each of the two interfaces. To make the situation simpler, we choose as a counter electrode an interface that will maintain a constant composition during the current flow. According to international conventions (Ivez and Janz, 1961) the counter or reference electrode should be the standard hydrogen electrode. For practical reasons, the most widely used electrodes are silver or mercury covered with a layer of insoluble chlorides, immersed in a solution of chlorides of a specified concentration. The latter is called the calomel electrode—according to the concentration of chloride solution one encounters most often, either the *saturated calomel electrode (SCE)* or the *normal calomel electrode (NCE).* The reference electrode is made in such a manner that, as a result of a surplus of insoluble chloride, its composition remains unaffected by the passage of current and hence its potential is well defined and constant. The voltage applied to an electrochemical cell will change exclusively the potential of the other electrode, called the *working electrode.* For the sake of comparison of data in every electrochemical experiment, the type of reference electrode with respect to which the potentials are measured has to be specified.

The voltage applied to an electrochemical cell controls the energy of electrons in the working electrode. When the energy of free electrons in the metallic phase is made sufficiently high, the electrons can occupy vacant orbitals in substances present in the solution near the electrode surface giving rise to a reduction current. In contrast, when the electrode potential is driven toward positive values, the lower energy of electrons in the metal can initiate an electron transfer from the substance to the metal, yielding an oxidation current of the opposite sign. Electric currents caused by the heterogeneous exchange of electrons obey Faraday's law and are termed *faradaic currents.* The exchange of electrons between the electrode and the species in the solution is confined to a very short distance. The concentration of an electroactive substance near the electrode quickly becomes depleted and mass transport resupplies the reaction zone (Heyrovský and Kůta, 1966; Bard and Faulkner, 1980).

An electrode in contact with a solution of electrolyte is surrounded by a very thin layer of solution where there is an enhanced concentration of ions of one kind attracted by the charge of the electrode surface. The layer is known as the electrode double layer (Delahay, 1965, Delahay and Tobias, 1961), and the potential difference between the electrode and the solution is confined within this space. The mass transport of electroactive ions, if present as the only ionic species, would necessarily involve migration, change of the double layer in time and space, and other complications. For these reasons, a strong electrolyte that is not electroactive in the given potential range (called the *indifferent* or *supporting electrolyte*) is added to the electrolyzed solutions in order to assure their good electrical conductivity, the elimination of migration phenomena, and relative simplicity of the double-layer composition. The electric current due to the charging of the double layer is called the *charging current* or *nonfaradaic current.* The theory of polarography and other electrochemical techniques solves the interplay of the electron-transfer rate (governed by the potential, by properties of the reacting substance, etc.) and the mass-transport rate. The instrumentation of electrochemistry solves the main problem of separation of faradaic and charging currents.

2. POLAROGRAPHY IN ITS CLASSICAL FORM

The polarographic experiment is rather a simple one: the objective is to measure the dependence of the electric current i flowing through the mercury drop/solution interface as a function of the potential E applied to it (the "polarographic curve") (Bard and Faulkner, 1980). Since the days of Heyrovský, this is done in an automatic manner, and the "polarograph" ranks among the first automatic scientific instruments [Fig. 1(a)]. A slowly and linearly changing dc voltage (originally derived from a car battery and a ten-turn motor-driven potentiometer) is applied to the cell containing the solution of a sample. Solutions must be deoxygenated by passing a stream of an inert gas (nitrogen or

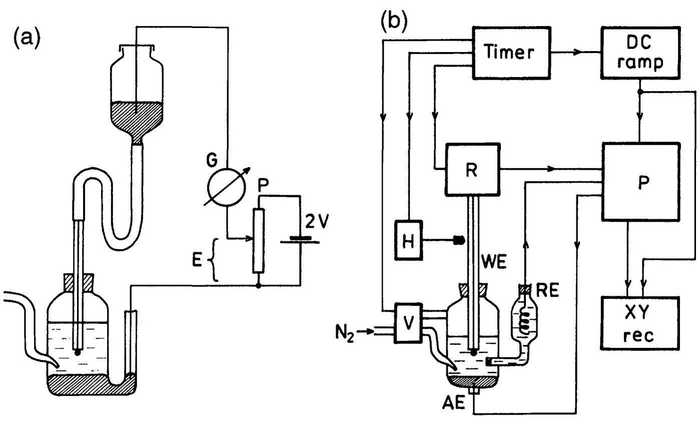

FIG. 1. Experimental setup for polarography: (a) Historical arrangement, where *P* is a ten-turn potentiometer and *G* is a galvanometer. (b) A contemporary polarograph consisting of a potentiostat *P* (see Fig. 6), a timer/synchronizer *T*, a dc ramp voltage generator, a mercury reservoir *R* with a solenoid valve, a mechanical hammer *H* for displacement of a drop, a gas valve *V*, and an *xy* recorder. The electrodes are denoted as WE, the working electrode; AE, the auxiliary electrode; and RE, the reference electrode.

argon) prior to the measurement because dissolved atmospheric oxygen itself gives a large reduction current. The range of 2 V or less is scanned in about 5 minutes. Every 2 to 6 seconds a mercury drop falls off (when its mass cannot be held by surface tension) and a clean new drop electrode pops out of the glass capillary. The current, of the order of microamperes, is amplified and an *xy* recorder is used for the desired chart of the current vs potential. Heyrovský's polarograph produced the record by focusing a beam of light reflected from a mirror galvanometer on the slit of a cassette with a rotating photographic paper coupled to the same motor as the potentiometer. The invention of solid-state devices profoundly changed the design of modern polarographic instrumentation [Fig. 1(b)] and enabled the development of many related methods. Advanced techniques differ from classical polarography by using a more complex shape of the potential wave form (superimposed small pulses, sine waves, and other perturbations) and by detecting only certain components of the resulting current response. A contemporary polarograph is a computer-controlled, multipurpose, electrochemical instrument that offers a variety of methods. Also, the dropping-mercury electrode is often replaced by a stationary mercury drop hanging on a more compact short column of mercury where the outflow is controlled by a magnetic valve at the upper orifice of the glass capillary. A sound understanding of the principles of classical polarography is still a prerequisite for a successful application of more advanced techniques.

Current-voltage curves, or polarographic curves, have shapes given by basic thermodynamic and kinetic principles. Let us consider an aqueous solution of a simple reducible cation like Cd^{2+} in a surplus of 0.1 M potassium chloride. In the potential range from 0 V to about -0.5 V, only a very small residual current flows corresponding to the charging of the electrode double layer. When the applied potential approaches the value of the standard potential $E^0 = -0.64$ V of the redox couple Cd^{2+}/Cd-amalgam, the electrolysis of cadmium salt begins and a reduction (or cathodic) current is observed. The magnitude of the current is given by Faraday's law,

$$i = nFA\frac{dN}{dt}, \qquad (1)$$

where n is the number of exchanged electrons ($n = 2$ for Cd^{2+}), F is the Faraday constant relating the electrolyzed amount to the passed electrical charge, and A is the surface area. The term dN/dt is the flux of particles reaching the surface, which in a general case can be expressed by the Nernst–Planck equation, including diffusion, migration, and convection terms. In the case of pure diffusion control, the electrolytic, or *faradaic*, current is proportional to the concentration gradient at the electrode surface, which can be derived on the basis of Fick's law expressed for an expanding spherical electrode. The solution of the corresponding partial differential equations yields proportionality of i to $t^{1/6}$ and to the bulk concentration. The seemingly unusual time exponent is the result of the mass transport decay ($t^{-1/2}$ according to the Cottrell equation) and the time dependence of the growing mercury-drop electrode (expanding with $t^{2/3}$). For a very fast reversible electron-transfer reaction, the concentrations at the electrode surface must obey the Nernst equation, which relates the potential E of an electrode with the concentration of oxidized and reduced forms (Ox and Red, respectively):

$$E = E^0 + \frac{RT}{nF} \ln \frac{[\mathrm{Ox}]}{[\mathrm{Red}]}. \tag{2}$$

The reduction current thus increases toward more negative potentials according to an exponential law. As a result of the depletion near the surface at potentials sufficiently more negative than E^0, the surface concentration of reducible substance (like Cd^{2+}) finally reaches zero. The further current increase levels off [Fig. 2(a)]. For the remaining part of the current-voltage curve, one observes a constant current, independent of the applied potential, which is called the *limiting current* or the *diffusion-limited current.* An uncomplicated fast–electron-transfer reaction is characterized by a polarographic curve in the form of an S-shaped dependence centered around E^0 with the limiting diffusion current proportional to the bulk concentration of electroactive substance (Cd^{2+} in our example). The magnitude of the limiting diffusion current is given by the Ilkovic equation (Heyrovský and Kůta, 1966), which gave the polarographic method a rigorous background:

$$i_d = 708nD^{1/2}\, C\, m^{2/3}\, t^{1/6}, \tag{3}$$

where C is the bulk concentration (in mol/cm^3), m is the flow rate of mercury (in mg/s), D is the diffusion coefficient (in cm^2/s), and the numerical term includes F and other constants. Already, from this simple introduction, we can deduce that the polarographic curves can be used for determination of the following parameters:

1. the bulk concentration of an electroactive substance,
2. the number of electrons involved in the redox reaction,
3. the standard redox potential,
4. the diffusion coefficient, and
5. the identity of an unknown substance on the basis of its E^0.

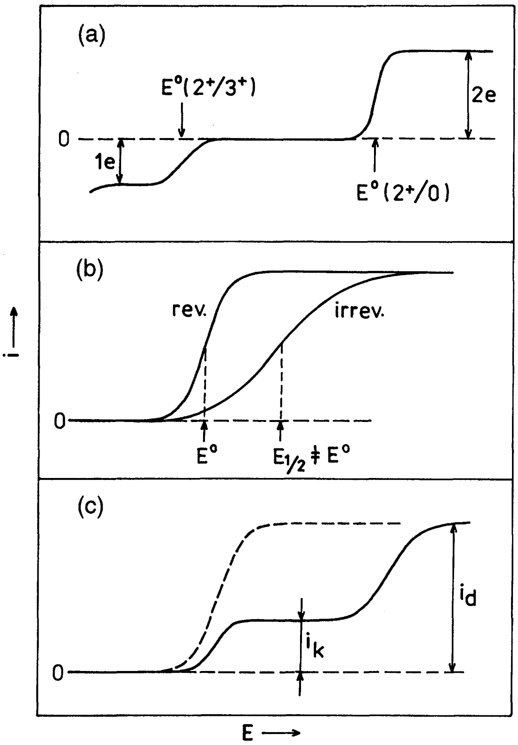

FIG. 2. Selected typical shapes of polarographic curves of current vs potential. (a) Example of a divalent metallic cation which is reduced to metal at $E^0(2^+/0)$ in a two-electron wave and oxidized to a trivalent state at $E^0(2^+/3^+)$. (b) Position of the half-wave potential of a reversible (rev.) and an irreversible (irrev.) wave with respect to E^0. (c) Limiting kinetic current (i_k) of a reversibly reduced form controlled by a preceding chemical reaction of an irreversibly reduced form (i_d). The dashed curve would be observed for a chemical reaction much faster than the diffusion.

The last two cases require a precise knowledge of other parameters and have only a limited scope.

2.1 Electrochemical Kinetics in Polarography

The previous section considered an ideal case of a simple electrode process with an infinitely fast electron-transfer rate. However, numerous redox reactions proceed with an energetic barrier resulting in a finite rate of the electron-transfer step

$$\mathrm{Ox} + ne \underset{k_b}{\overset{k_f}{\rightleftharpoons}} \mathrm{Red}, \tag{4}$$

where k_f and k_b are potential-dependent heterogeneous rate constants for reduction and oxidation, respectively. Kinetic hindrance in the reaction scheme (4) has a profound consequence on faradaic currents. The polarographic waves differ in shape and position in comparison with the previously described reversible case [Fig. 2(b)]. The wave is less steep, and the displacement from E^0 toward higher potentials (more negative for reduction and more positive for oxidation) reflects the kinetic hindrance. Furthermore, the observed currents depend on the time constant of the electrochemical method being used, or, in other words, on a time interval that a particular method affords to the mass transport (Macdonald, 1977). The time constant of polarography is given by the interval between two consecutive mercury-drop falls. Technical obstacles do not allow change of the drop time by more than one order of magnitude, and this limitation was one of the reasons for the introduction of other related polarographic techniques. The quantitative measure of a finite rate of electron transfer (Hush, 1971) is the value of the *standard heterogeneous rate constant* k^0 which has the dimensions $\mathrm{cm \cdot s^{-1}}$. The magnitude of k^0 determines the deviation of the *i-E* curve from the Nernstian shape, or the deviation from a fast equilibrium between oxidized and reduced forms of the substance under investigation. The rate of the electrochemical reduction at a given potential depends exponentially on E according to the Butler–Volmer formalism (Vetter, 1961),

$$k_f = k^0 \left[\frac{\alpha n F}{RT} (E - E^0) \right], \tag{5}$$

where α is the *transfer coefficient* (having values between 0 and 1, but in most cases approximately 0.5). A full explanation of the concept of the slow electrode reaction is a complex subject, and we will only mention that α is connected with the symmetry of the energy barrier for conversion between oxidized and reduced forms. The quantum mechanical theory of k^0 led to numerous correlations between homogeneous and heterogeneous electron-transfer rates, and studies of structural effects and of the influences of external fields, the solvation shell, etc. (Hush, 1971). The upper limit for the estimation of k^0 by the polarographic method is about 0.01 $\mathrm{cm \cdot s^{-1}}$, which implies that the faster reactions appear to be controlled solely by diffusion. In order to measure faster redox rates, numerous faster relaxation techniques have been developed. Unfortunately, the influence of diffusion diminishes in proportionality to $t^{-1/2}$, and hence the lowering of the time constant of a method by two orders of magnitude allows one to measure k^0 higher by only one order.

Another type of kinetic effects in electrode reactions involves the coupling of a chemical reaction and an electron-transfer reaction (Heyrovský and Kůta, 1966). From a long list of various reaction schemes (Rossiter and Hamilton, 1986), we will give here only a few representative mechanisms:

1. A preceding chemical reaction (called also CE mechanism, which is an abbreviation for the order of reactions "Chemical-Electrochemical"):

$$A \pm X \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{Ox} + ne \rightleftharpoons \mathrm{Red}. \tag{6}$$

The kinetic hindrance in this case is due to the finite rate of a chemical reaction between substances A and X, quantitatively expressed by the rate constants k_1 and k_{-1} (which are independent of E because the process is a homogeneous one). Typical examples are protonations $A + H^+$ and ligand dissociations $AL_n \rightarrow A + nL$. Chemical reactions can be coupled with both reversible and irreversible electron-transfer reactions. The reaction scheme (6) describes a situation when the solution contains substances A and X that are electroinactive (or active at much higher potentials), but their reaction yields a reducible product Ox participating in the electrode reaction. By analogy one can write the scheme for generation of an oxidizable substance Red. The consumption of Ox at the electrode surface shifts the equilibrium between $A + X$ and Ox. When the rate of that particular chemical reaction becomes the rate-determining step, the reduction current of Ox will be limited by a homogeneous reaction instead of mass transport [Fig. 2(c)]. Kinetic currents have been used for estimation of homogeneous rate constants of systems that would be difficult to tackle by other techniques.

2. Follow-up chemical reactions (abbreviated sometimes as "EC mechanism" for the sequence "Electrochemical-Chemical" reaction) have an order of reactions the reverse of that for the scheme (6):

$$\mathrm{Ox} + ne \rightleftharpoons \mathrm{Red} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{Product}. \qquad (7)$$

   The product of a follow-up chemical reaction can originate from a simple monomolecular decomposition of Red or from a bimolecular coupling of electrogenerated radicals, or the reaction mechanism can involve other substances present in the solution.

3. Catalytic regeneration of the electroactive substance is a special case of follow-up reaction in which the product is the starting material; in the case of Eq. (7) it is the Ox form. The observed currents then exceed the diffusion-limited values (given by the concentration in the solution) because Red is "recycled." This class of mechanism is the focus of interest for many reasons. For example, the analytical applications take advantage of catalytic enhancement of the faradaic current to achieve a substantial extension of the detection limits.

3. DERIVED POLAROGRAPHIC TECHNIQUES

The previous section pointed to certain restrictions of classical polarography in kinetic studies, but this is not the only limitation. So far we did not consider the circumstance that the electrode–solution interface consists of the electrode double layer having a fairly large electrical capacity, of the order of 10–100 $\mu\mathrm{F}\cdot\mathrm{cm}^{-2}$. This capacitor must be recharged with any change of the electrode area and during the formation of a fresh electrode surface. The ever-present double-layer charging current imposes a lower limit for detection of faradaic processes, which at trace concentrations become hidden in a much larger charging current. Fortunately, the charging current and the faradaic current differ in their time decay, a fact that is the basis for a better resolution of more sophisticated techniques. The electrode double layer behaves as a condenser connected to the voltage source via a solution resistance R_s. Upon application of a constant potential the charging current decays exponentially, with the time constant R_sC (where C is the double-layer capacitance), and after the time lapse of $5R_sC$ its contribution is negligible. The faradaic current also decays with time, but more slowly ($t^{-1/2}$ decay according to the Cottrell equation). Sampling and subtraction of the experimental current at different time instances is the essence of pulse polarography. Different time lags of the two current contributions leads to different phase relations when transformed to the frequency domain, and this property is the main virtue of ac polarography. The methods described below tackle the same problem: the discrimination between charging and faradaic phenomena.

3.1 Pulse Polarography

The applied potential in classical polarography is practically constant during the life of each drop of the dropping-mercury electrode (more precisely, its linear change with time is so slow that it becomes negligible). The pulse polarographic techniques are based on the application of a short rectangular pulse (10–100 ms) superimposed on the dc potential shortly before the end of the drop life. Three related techniques (Greef *et al.*, 1985; Bard and Faulkner, 1980) are called "normal pulse polarography" (NPP), "differential pulse polarography" (DPP), and "square-wave polarography" (SWP), and all differ solely in the form of the pulse shapes and the current sampling modes. The recent revival of electrochemistry in environmental control is due to the high sensitivity and the low cost of equipment for DPP and SWP. DPP uses a polarization regime in which a linearly increasing potential ramp is applied to the electrode (as in the classical polarography); however, close to the end of the drop life the current is sampled, a rectangular pulse of an amplitude of 10–50 mV is triggered, and prior to the end of the pulse another current sampling is made [Fig. 3(a)]. The recorded signal is the difference of the two current samples, which gave this method its name "differential pulse." The difference signal suppresses substantially the

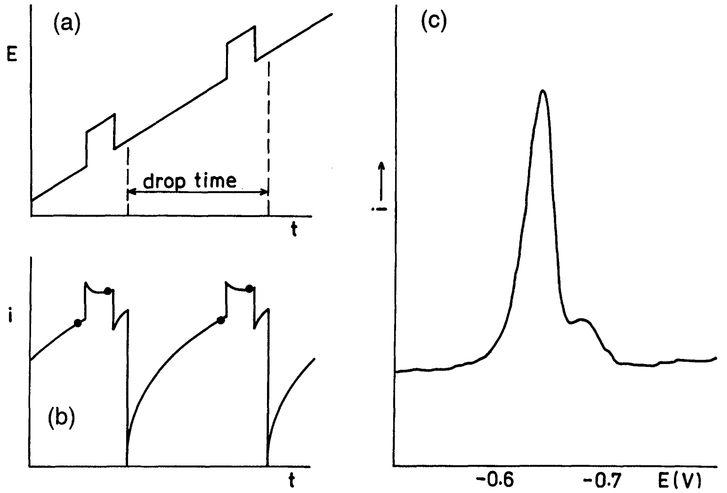

FIG. 3. Differential pulse polarography: (a) The shape of the applied potential; (b) current during the drop life; (c) example of a DPP curve of arsenic at concentration 50 μg/l.

contribution of the unwanted charging current, and the recorded curve has the shape of the first derivative of the corresponding classical polarographic wave. Closely spaced faradaic peaks are hence better resolved than the merging waves. A small drawback of the DPP is a nonlinear background line, which comes into question only at the highest sensitivity levels. DPP is often used in combination with pre-electrolysis or other means of accumulation of products followed by the reverse reaction. This type of sensitivity enhancement is called *stripping* and finds application in the trace analysis either of heavy metals, which form amalgams with the mercury electrode (anodic stripping), or of substances that form insoluble compounds with mercury ions (cathodic stripping). Analogously, *adsorptive stripping* makes use of adsorption accumulation, which saturates the electrode surface with the substance to be determined by its adsorption at a certain constant potential. A voltage scan that follows the accumulation (often accelerated by stirring the solution) changes the electrode potential to values where desorption from the surface takes place. The desorption process yields a sharp current maximum proportional to the bulk concentration provided that the accumulation time, stirring, and other parameters are identical to those in the calibration experiments.

3.2 ac Polarography

ac polarography is a technique based on the measurement of the electrode admittance (the reciprocal of the impedance) as a function of the dc potential. The method is implemented by perturbing the slowly scanned dc potential by a small-amplitude sine-wave form. The signal processing involves filtration of the ac current component corresponding to the ac perturbation followed by amplification and phase-sensitive detection (Smith, 1966; Sluyters-Rehbach and Sluyters, 1970). The ac current is a vector having a phase relation to the ac potential determined by the impedance of the electrochemical cell and by the frequency of the ac signal. The most common mode of ac polarography records two vector components of the ac current (the in-phase and the out-of-phase components, sometimes called the "real" and "imaginary" components, respectively) as a function of the dc potential. The ac polarogram of a sample containing only the supporting electrolyte yields a real component equal to zero and an imaginary current component proportional to the double-layer capacity and the applied frequency [Fig. 4(a)]. Phase-sensitive ac polarography is often used for investigation of adsorption processes, evaluation of adsorption isotherms, and studies of the growth of adsorbed films and of the influence of adsorption of inhibitors on the rate of electrode reactions.

The presence of an electroactive substance in the solution gives peak-shaped maxima on both vector components [Fig. 4(b)], which can be used for analytical purposes. The positions and ratio of the two faradaic maxima may be used for diagnostics of the mechanism of electrode processes and

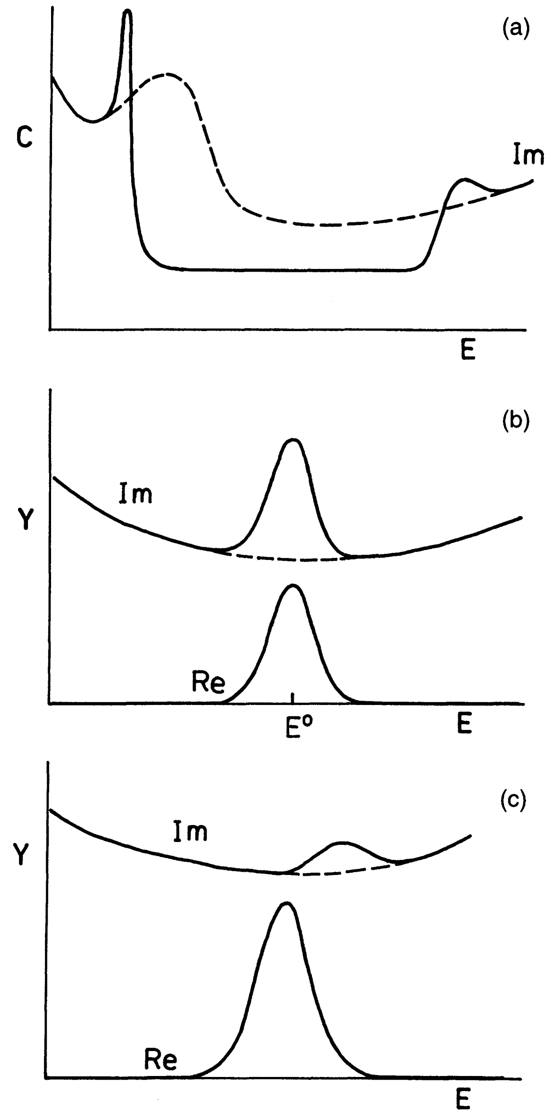

FIG. 4. Selected typical shapes of phase-sensitive ac polarograms (Re and Im denote the real and imaginary components, respectively): (a) The double-layer capacitance of an indifferent electrolyte (dashed curve) and in the presence of an adsorbable surfactant. The two maxima frame the potential region of adsorption. (b) Faradaic maxima of a reversible redox system. (c) Faradaic maxima of a system with kinetic control. Such a picture can be obtained from case (b) by choosing a sufficiently high frequency.

for the estimation of kinetic parameters (Macdonald, 1977). The time constant of ac polarography is given by the frequency of the ac perturbation, and hence it could be varied over a much wider range than for previously described techniques. The impedance data extracted from a set of ac polarograms are often analyzed in the complex plane and interpreted in terms of an equivalent circuit or a transmission-line model. A faster collection of impedance spectra is achieved by a broadband perturbation combined with the fast Fourier transform or similar procedures. Other derived ac techniques, like second-harmonic ac polarography and the faradaic rectification method, have been developed, but at present are rarely used.

4. VOLTAMMETRY

The mercury electrode can be used in a wide potential range limited at negative potentials by the electrolytic decomposition of the solvent (−2 V in water and about −3.2 V in dry nonaqueous solvents) and at positive potentials by the anodic oxidation of mercury (+0.6 V or less). Platinum, gold, and carbon are the most common materials used for the investigation of oxidation reactions proceeding at positive potentials not attainable at mercury electrodes. In experiments performed with solid electrodes, the surface concentration cannot be renewed (as for liquid mercury drops) and the current–voltage curves have shapes characterized by depletion of the electroactive substance near the surface. Instead of the potential-independent plateau of limiting currents in polarography, the voltammetric limiting current steadily decays, reflecting the exhaustion of concentration. Otherwise both polarography and voltammetry are techniques that measure the same quantity, the faradaic current as a function of the applied potential. Since the electrode surface and the initial concentrations at the interface are not periodically renewed (as in polarography), the observed *i*-*E* curve combines the effects of an *i*-*t* decay and a linear *E*-*t* change. For voltammetry a more accurate name, *linear potential sweep chronoamperometry,* is sometimes used, but most of the applications refer to this method as *linear sweep voltammetry.* The change of the potential with time is $E(t) = E_i - vt$, where the initial potential is E_i and v is the sweep or scan rate. The sweep rate is an important variable, which determines the time constant of the voltammetric method. In contrast with polarography, the sweep rate can be chosen in the range of v from 0.05 to

1000 V/s (or even to MV/s), and this feature makes the method very suitable for the detection of short-lived intermediates of complicated redox reactions. The original design of a voltage source for variable scan rates used solid-state operation amplifiers generating the desired linear voltage ramp on the basis of an analog integrator. Currently, many instruments implement the voltage ramp by means of computer-controlled digital-to-analog converters, which results in a staircase voltage. The two methods are theoretically equivalent only for very small step increments. When differences are noticeable, the latter method is called *staircase voltammetry* and a corresponding theoretical treatment is available. We recall that commercial oscillators or function generators are not suitable for a voltammetric experiment. Voltammetry requires the ability to vary the asymmetrical amplitude (the initial and final potentials) and at the same time maintain an unchanged slope (the scan rate), while the change of frequency is irrelevant.

In the course of a potential-sweep experiment, there develops a concentration gradient of generated products of electrode reactions. By reversing the direction of the potential sweep at a certain potential we can record the corresponding reverse redox processes (reoxidation or rereduction). This reversal technique is called *cyclic voltammetry* and has proven to be one of the basic tools for analysis of an unknown complicated electrode mechanism (Noel and Vasu, 1990; Smyth and Vos, 1992). Unfortunately, there is always a price to be paid for an advantage, and in the case of voltammetry it is the complicated mathematics originating from the time dependence of $E(t)$. Many important reaction schemes cannot be easily solved in a closed form, which led to the development of sophisticated simulation methods (Feldberg, 1969; Britz, 1981).

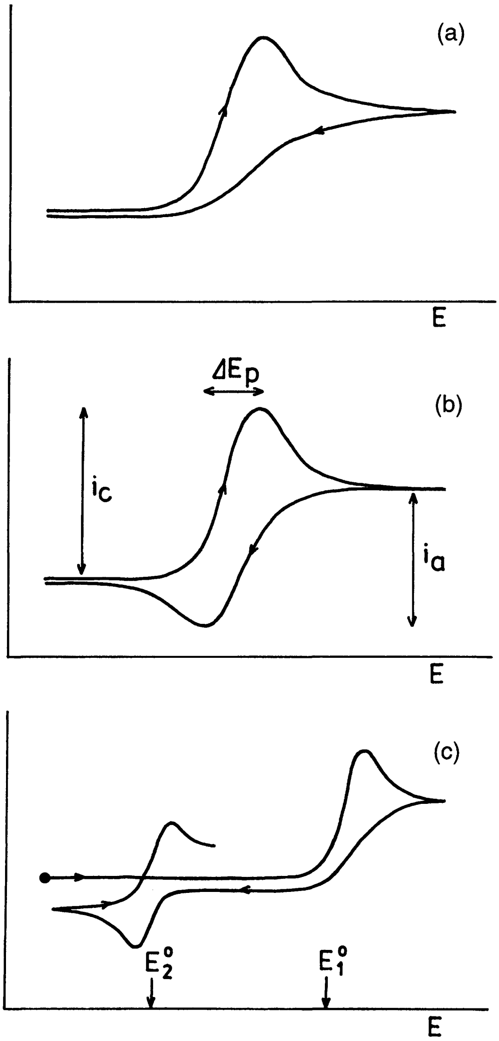

FIG. 5. Selected typical shapes of cyclic voltammograms (arrows mark the scan direction): (a) Reduction coupled with a follow-up chemical decomposition of the product measured at a slow scan rate; (b) the same system as in (a) measured at a high scan rate yielding reversible behavior. Important diagnostic parameters are the ratio of cathodic and anodic currents i_c/i_a, and the separation of peak potentials ΔE_p. (c) Example of a system with E_1^0 yielding in a follow-up reaction a new reversible redox system with E_2^0.

An uncomplicated reversible system yields a cyclic voltammogram that features the forward and reversed wave shapes as mutual mirror images of one another with respect to E^0. Any kind of irreversibility of the electron-transfer reaction is sensitively reflected in the differences of the currents recorded during the forward and backward scanning. A typical area of application is the detection of a coupled chemical decomposition of the primary product (Fig. 5). While the reverse scan fails to yield the mirror wave form at slow scan rates v (the conditions allow a complete decomposition of the product), at a sufficiently high scan rate one can observe again the electrochemical reversibility of the investigated system because the reoxidation (or rereduction) takes place much faster than the chemical decomposition. The limited space of this article does not allow the author to go into details of this powerful method, which has been worked out in substantial detail both in theory and in the experimental approach. For the evaluation of experiments, pattern recognition methods were also applied.

The latest improvement of voltammetry is the use of ultramicroelectrodes, which are made of a glass-sealed platinum wire or carbon fiber having a diameter of the order of microns or even less. These electrodes have much smaller interfacial area than the conventional ones. Both the double-layer capacitance and the current generated at these electrodes are correspondingly smaller, which results in a signal considerably less distorted by the Ohmic resistance of the solution. Also, the smaller time constant of the cell makes possible the achievement of faster perturbation rates. The measurements can be made even in the absence of an indifferent electrolyte, making microelectrodes an attractive tool for sensing pollutants in water or for other applications in high-resistance media.

5. EXPERIMENTAL EQUIPMENT

Basic electrochemical equipment must deal with various problems specific for a given application. The most common questions are the choice of a suitable electrode material for the working electrode, elimination of a high Ohmic resistance of some samples, especially in nonaqueous solvents or at low temperatures, and an improvement of the signal-to-noise ratio at ultralow concentrations of electroactive material.

5.1 Working Electrodes, Solvents, Indifferent Electrolytes

For investigation of reduction processes in water or nonaqueous solvents containing strong acids, the electrode of choice is the mercury electrode. A very slow hydrogen evolution reaction on mercury in comparison with noble metals in concentration of acids as high as 1 $\text{mol} \cdot \text{l}^{-1}$ enables the measurements up to -1 V. Using low-proton media extends the range of accessible potentials to -2.5 V. The indifferent electrolytes are chosen according to the solubility in the given solvent. In aqueous solutions, common salts are the alkali metal halides or perchlorates (when complexation by halides has to be avoided) and a variety of *p*H-regulating buffer solutions. Nonaqueous electrochemistry uses mainly tetra-alkylammonium perchlorates, tetrafluoroborates, and hexafluorophosphates as indifferent electrolytes. Voltammetric measurements have been performed also in liquid gases (SO_2 and NH_3) and in low-temperature melts.

The facile oxidation of metallic mercury limits the useful potential range on the positive side to $+0.2$ to $+0.6$ V depending on the medium. For this reason noble metals, platinum and gold, are used in oxidation studies; however, in aqueous solutions both these metals become covered with a layer of oxides. An inert noble-metal electrode performs quite satisfactorily in nonaqueous solvents. The recently developed technique of preparing ultramicroelectrodes of diameters 0.5 to 20 μ solves the problems of the Ohmic drop of potential in the sample but requires a more sensitive detection of low-level currents. Microelectrodes enable one to perform measurements in solutions without indifferent electrolyte, which makes them promising tools in solvents of a low dielectric constant. Other popular electrodes are various forms of carbon: the glassy-carbon disk electrode and the carbon-paste electrode. All solid electrodes are used also in the form of a rotating disk embedded in an insulating Teflon rod. A special class is constituted by the *modified electrodes* designed in such a way that a solid electrode is covered by a layer of a polymer containing electroactive centers. Surface modifying techniques in most cases are based on oxidative polymerization, but surface casting is used as well. Modified layers containing zeolites, enzymes, or ion-selective polymers are described as special-purpose sensors. The main problem of modified electrodes is good stability and adhesion of a modifier to a metal substrate. The last type of electrodes to be mentioned here is optically transparent electrodes needed for a combination of electrochemical and spectroscopic techniques. The design is based either on a fine micromesh (about 1000 wires/inch) of Au or Pt electrodes, or on a thin transparent layer of conducting tin oxide or indium oxide on glass or quartz. Such electrodes make possible the identification of redox products by their spectral characterization *in situ.*

5.2 Instrumentation

As we said in the introduction, the instrumentation that sufficed for the Nobel prize experiment was the simplest that one can imagine. Low investment cost for electrochemistry still holds, and even now many laboratories rely on their own design of electronic equipment. The heart of the measuring system is a *potentiostat,* which is an analog feedback-controlled circuit performing an automatic compensation of the Ohmic drop of the potential in the cell in such a way that the working electrode is polarized to a desired value regardless of the amount of current that flows (Bard and Faulkner, 1980). The potentiostat in its simple form consists of three solid-state operational amplifiers (Fig. 6). An adder sums all the potential wave forms needed for a given method and a negative feedback loop senses the actual potential of the working electrode against the reference electrode. A current follower maintains the working electrode on the potential of a virtual ground and yields at its output a signal proportional to the measured current that is suitable for further conditioning or a direct recording. The third operational amplifier serves as an impedance buffer (a voltage follower), preventing the reference electrode from undesired loading. Programmable potential-wave forms are usually generated by a computer by means of a digital-to-analog converter having at least a 12-bit resolution. More sophisticated techniques like ac polarography or very fast voltammetry use correspondingly more expensive signal-conditioning instruments (lock-in amplifiers, transient recorders, digital oscilloscopes).

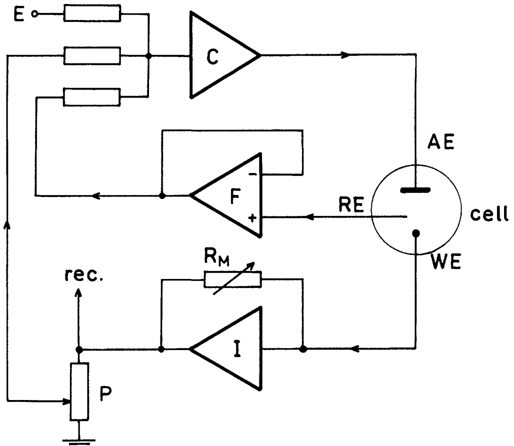

FIG. 6. Schema of a potentiostat: The potential *E* that has to be maintained is fed to a controlling amplifier *C* and to the auxiliary electrode *AE*. The working electrode *WE* is kept at the potential of a virtual ground; the current is measured by a current-to-voltage converter *I* (R_M is the measuring resistor) and led to a recorder. The voltage follower *F* senses the actual applied voltage vs the reference electrode *RE* and forms the negative-feedback loop. An additional compensation of the solution resistance can be achieved by a positive-feedback loop via a ten-turn potentiometer *P*. Passive elements stabilizing the circuit are not shown.

6. APPLICATIONS

The exceedingly large amount of electrochemical applications cannot be classified in a short paragraph. We can only refer to available encyclopedic or special literature (Baizer, 1973; Bard and Lund, 1973–80; Meites and Zuman, 1974). Analytical applications allow the determination of almost all elements that are required for monitoring either in industrial processes or in pollution control (Florence, 1989). Organic electroanalysis finds its place in food control, clinical tests, pharmacology, agrochemistry, pesticide control, and numerous other fields (Smyth and Vos, 1992). An equally important role is played by electrochemical techniques in research praxis because the redox characteristics of a new synthetic compound is information as important as its spectral, crystallographic, and other parameters.

Works Cited

Albery, W. J. (1975), *Electrode Kinetics,* Oxford, U.K.: Clarendon.

Baizer, M. M. (1973), *Organic Electrochemistry,* New York: Dekker.

Bard, A. J., Faulkner, L. R. (1980), *Electrochemical Methods, Fundamentals and Applications,* New York: Wiley.

Bard, A. J., Lund, H. (Eds.) (1973–80), *Encyclopedia of the Electrochemistry of Elements,* Vols. 1–14, New York: Dekker.

Britz, D. (1981), *Digital Simulation in Electrochemistry,* Berlin: Springer-Verlag.

Delahay, P. (1965), *Double Layer and Electrode Kinetics,* New York: Wiley-Interscience.

Delahay, P., Tobias, C. W. (Eds.) (1961), *Advances in Electrochemistry and Electrochemical Engineering,* Vol. 1, *Electrochemistry,* New York: Wiley-Interscience.

Feldberg, S. W. (1969), in: A. J. Bard (Ed.), *Electroanalytical Chemistry,* Vol. 3, New York: Dekker.

Florence, T. M. (1989), in: G. E. Batley (Ed.) *Trace Elements Speciation: Analytical Methods and Problems,* New York: CRC Press Inc., p. 17.

Greef, R., Peat, R., Peter, L. M., Pletcher, D., Robinson, J. (1985), *Instrumental Methods in Electrochemistry,* Chichester: Ellis Horwood.

Heyrovský, J., Kůta, J. (1966), *Principles of Polarography,* New York: Academic.

Heyrovský, J. (1922), *Chem. Listy* **16,** 256.

Hush, N. S. (1971), *Reactions of Molecules at Electrodes,* New York: Wiley-Interscience.

Ivez, D. J. G., Janz, G. J. (Eds.) (1961), *Reference Electrodes, Theory and Praxis,* New York: Academic.

Macdonald, D. D. (1977), *Transient Techniques in Electrochemistry,* New York: Plenum.

Meites, L., Zuman, P. (Eds.) (1974), *Electrochemical Data,* New York: Wiley.

Noel, N., Vasu, K. I. (Ed.) (1990), *Cyclic Voltammetry and the Frontiers of Electrochemistry,* London: Aspect.

Rossiter, B. W., Hamilton, J. F. (Eds.) (1986), *Physical Methods of Chemistry,* Vol. II, *Electrochemical Methods,* New York: Wiley.

Sluyters-Rehbach, M., Sluyters, J. H. (1970), in: A. J. Bard (Ed.) *Electroanalytical Chemistry,* Vol. 4., New York: Dekker, p. 1.

Smith, D. E. (1966), in: A. J. Bard (Ed.) *Electroanalytical Chemistry,* Vol. 1, New York: Dekker, p. 1.

Smyth, M. R., Vos, J. G. (Eds.) (1992), *Analytical Voltammetry,* Comprehensive Analytical Chemistry, Vol. 27, Amsterdam: Elsevier.

Thirsk, H. R. (1972), *A Guide to the Study of Electrode Kinetics,* New York: Academic.

Vetter, K. J. (1961), *Electrochemical Kinetics,* New York: Academic.

Further Reading

A complete background of electrochemical techniques can be found in the book of Bard and Faulkner (1980) and in the monographic literature cited above. Several multivolume series bring periodically the newest developments in the field:

Bard, A. J. (Ed.), *Electroanalytical Chemistry. A Series of Advances,* New York: Dekker.

Conway, B. E., Bockris, J. O'M. (Eds.), *Modern Aspects of Electrochemistry,* New York: Plenum.

Original papers are published in periodicals specializing in electrochemistry: *Journal of Electroanalytical Chemistry, Electrochimica Acta, Journal of the Electrochemical Society,* and in more general journals like *Analytical Chemistry, Electroanalysis, Langmuir, Journal of Physical Chemistry, Inorganic Chemistry,* and others.

POLARONS

J. T. DEVREESE, *Universiteit Antwerpen, Antwerp, Belgium, and Technische Universiteit Eindhoven, Eindhoven, The Netherlands*

Introduction 383
1. **The Polaron Concept and the Fröhlich Polaron** 383
1.1 The Concept 383
1.2 Standard Fröhlich-Polaron Theory. Self-Energy and Effective Mass of the Polaron 385
1.2.1 Strong Coupling 385
1.2.2 Weak Coupling 386
1.2.3 All-Coupling Theory. Feynman Path Integral 387
1.2.4 Response Properties of Fröhlich Polarons. Mobility and Optical Properties 388
1.3 Fröhlich Polarons in 2D 389
1.3.1 Introduction 389
1.3.2 Scaling Relations 390
1.4 Fröhlich Polarons in a Magnetic Field 390
1.4.1 In 3D 390
1.4.1.1 Level Crossing, Pinning 390
1.4.1.2 Static and Dynamic Properties of Polarons in a Magnetic Field 391
1.4.1.3 Cyclotron-Resonance Spectra 392
1.4.2 Cyclotron Resonance of Polarons in 2D 393
1.4.3 Formal Developments 394
1.5 The Bound Polaron 394
2. **The Small Polaron** 395
2.1 The Small-Polaron Concept. Role of Localization 395
2.2 Standard Small-Polaron Theory 397
2.3 Experimental Evidence 398
3. **Bipolarons and Polaronic Excitons** 400
3.1 Fröhlich Bipolarons 400
3.2 Polaronic Excitons 402
3.3 Localized Bipolarons 403
4. **Spin Polarons** 403
5. **Bipolarons and High-Temperature Superconductivity** 404
6. **Further Developments of the Polaron Concept** 407
6.1 Polarons and Bipolarons in Polymers. Solitons 407
6.2 Modeling Systems Using the Polaron Concept 407
Acknowledgments 408
Glossary 408
Works Cited 409
Further Reading 412

INTRODUCTION

The polaron concept is of interest, not only because it describes the particular physical properties of an electron in polar crystals and ionic semiconductors, but also because it is an interesting field-theoretical model consisting of a fermion interacting with a scalar boson field.

The early work on polarons was concerned with general theoretical formulations and approximations, which now constitute the standard theory, and with experiments on cyclotron-resonance and transport properties.

Because of the more recent interest in the two-dimensional electron gas (2DEG), the study of the polaron in two dimensions became important. Cyclotron resonance, and therefore the behavior of polarons in magnetic fields, was a key issue.

1. THE POLARON CONCEPT AND THE FRÖHLICH POLARON

1.1 The Concept

A conduction electron (or hole) together with its self-induced polarization in a polar semiconductor or an ionic crystal forms a

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

quasiparticle, which is called a polaron (Fig. 1). The physical properties of the polaron differ from those of the band electron. In particular, the polaron is characterized by its binding (or self-) energy, its effective mass, and its response to external electric and magnetic fields (e.g., mobility and impedance). The general polaron concept was introduced by Landau (1933) in a paper of about one page. Subsequently, Landau and Pekar (see Pekar, 1951) investigated the self-energy and the effective mass of the polaron, for what was shown by Fröhlich (1954) to correspond to the adiabatic or strong-coupling regime.

The early work on polarons was devoted to the interaction between a charge carrier (electron, hole) and the long-wavelength optical phonons. The (now standard) field-theoretical Hamiltonian describing this interaction was derived by Fröhlich:

$$\mathrm{H} = \frac{p^2}{2m_b} + \sum_k \hbar\omega_{\mathrm{LO}} a_{\mathbf{k}}^{\dagger} a_{\mathbf{k}} + \sum_k (V_k a_k e^{i\mathbf{k}\cdot\mathbf{r}} + H.c.), \tag{1a}$$

where $\mathbf{r}$ is the position coordinate operator of the electron with band mass m_b, $\mathbf{p}$ is its canonically conjugate momentum operator, and $a_{\mathbf{k}}^{\dagger}$ and $a_{\mathbf{k}}$ are the creation (and annihilation) operators for longitudinal optical phonons of wave vector $\mathbf{k}$ and energy $\hbar\omega_{\mathrm{LO}}$. The V_k are Fourier components of the electron–phonon interaction:

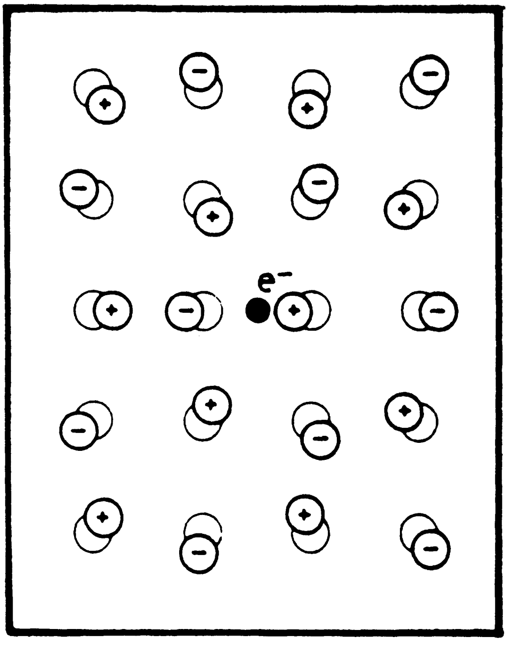

FIG. 1. A conduction electron (or hole) together with its self-induced polarization in a polar semiconductor or an ionic crystal forms a quasiparticle: a polaron.

$$V_k = -i\frac{\hbar\omega_{\mathrm{LO}}}{k}\left(\frac{4\pi\alpha}{V}\right)^{1/2}\left(\frac{\hbar}{2m_b\omega_{\mathrm{LO}}}\right)^{1/4}, \tag{1b}$$

$$\alpha = \frac{e^2}{\hbar c}\sqrt{\frac{m_b c^2}{2\hbar\omega_{\mathrm{LO}}}}\left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_0}\right). \tag{1c}$$

Here α is called the Fröhlich coupling constant; ϵ_∞ and ϵ_0 are respectively the electronic and the static dielectric constant of the polar crystal. In Table 1, a list is given of the coupling constants for a number of crystals.

In deriving the form of V_k, Eqs. (1b) and (1c), it was assumed that

1. the spatial extension of the polaron is large compared with the lattice parameters of the solid ("continuum" approximation);
2. spin and relativistic effects can be neglected;
3. the band electron has parabolic dispersion; and
4. in conjunction with the first approximation it is also assumed that the LO phonons of interest for the interaction are the long-wavelength phonons with constant frequency ω_{LO}.

The original concept of the polaron, as discussed above, has been generalized over the years to include polarization fields other than the LO-phonon field: the acoustical-phonon field, the exciton field, etc. For some materials, the continuum approximation is

Table 1. Fröhlich coupling constants.

Material	α	Material	α
CdTe	0.31	KI	2.5
CdS	0.52	RbCl	3.81
ZnSe	0.43	RbI	3.16
AgBr	1.53	CsI	3.67
AgCl	1.84	TlBr	2.55
CdF_2	3.2	GaAS	0.068
InSb	0.02	GaP	0.201
KCl	3.5	InAs	0.052
KBr	3.05	$SrTiO_3$	4.5

not appropriate insofar as the polarization is confined to a region of the order of a unit cell; the so-called "small polaron" is a more adequate quasiparticle in that case (see Sec. 2).

It is customary to use the term "Fröhlich polaron" or "large polaron" for the quasiparticle consisting of the electron (or hole) and the polarization due to the LO phonons. The term Landau or Pekar or Landau–Fröhlich polaron would be more appropriate.

In this first section, the properties of Fröhlich polarons are reviewed; other realizations of the polaron concept (small polarons, spin polarons, etc.) are discussed in later sections. For reviews on polarons, see Kuper and Whitfield (1963), Appel (1968), Devreese (1972), and Devreese and Peeters (1984).

1.2 Standard Fröhlich-Polaron Theory. Self-Energy and Effective Mass of the Polaron

Historically, the first studies on polarons [the "Russian work": Landau (1933), Pekar (1951)] were based on a *Produkt-Ansatz* for the polaron wave function,

$$|\Phi\rangle = |\Psi(\mathbf{r})\rangle|\text{field}\rangle = |\Psi(\mathbf{r})\rangle|f\rangle, \tag{2a}$$

where $|\Psi(\mathbf{r})\rangle$ is the electron wave function. The field wave function parametrically depends on the electron wave function. Fröhlich showed that the approximation (2a) leads to results that are only valid as $\alpha \to \infty$ (the strong-coupling regime). It should be pointed out that a more systematic analysis of the strong-coupling polarons based on canonical transformations of the Hamiltonian (1a) was performed in pioneering investigations by Bogolubov (1950), Bogolubov and Tyablikov (1949), and Tyablikov (1951).

Fröhlich also found that Eq. (2a) is a poor *Ansatz* to represent actual crystals, which have α values typically ranging from $\alpha = 0.02$ (InSb) to $\alpha \sim 3$ to 4 (alkali halides). This work showed the need for a weak-coupling theory of the polaron, which in fact was provided originally by Fröhlich (1954). However, it turns out that for $\alpha \approx 3$ perturbation expansions in powers of α are not always sufficient; therefore an intermediate-coupling or—better—an all-coupling theory was necessary.

In what follows, the key concepts of the standard Fröhlich-polaron theories are reviewed.

1.2.1 Strong Coupling The *Produkt-Ansatz* (2a)—or Born–Oppenheimer approximation—implies that the electron adiabatically follows the motion of the atoms (Pekar, 1951). With the use of the canonical transformation

$$S = \exp\left[-\sum_{\mathbf{k}}\left(\frac{V_k^* \rho_{\mathbf{k}}^*}{e\hbar\omega_{\mathrm{LO}}} a_{\mathbf{k}} - H.c.\right)\right], \tag{2b}$$

where

$$\rho_{\mathbf{k}} = e\langle\Psi(\mathbf{r})|e^{i\mathbf{k}\cdot\mathbf{r}}|\Psi(\mathbf{r})\rangle, \tag{2c}$$

it leads to $|f\rangle = S|0\rangle$. The ket $|0\rangle$ describes the vacuum state. With Eqs. (2) and a Gaussian trial function for $|\Psi(\mathbf{r})\rangle$, the ground-state energy of the polaron E_0 (calculated with the energy of the uncoupled electron–phonon system as zero energy) takes the form

$$E_0 = -(\alpha^2/3\pi)\hbar\omega_{\mathrm{LO}} = -0.106\alpha^2\hbar\omega_{\mathrm{LO}}. \tag{3a}$$

At strong coupling, the polaron is characterized by Franck–Condon (FC) excited states, which correspond to excitations of the electron in the potential adapted to the ground state. The energy of the lowest FC state is, within the *Produkt-Ansatz*,

$$E_{\mathrm{FC}} = (\alpha^2/9\pi)\hbar\omega_{\mathrm{LO}} = 0.0354\alpha^2\hbar\omega_{\mathrm{LO}}. \tag{3b}$$

If the lattice polarization is allowed to relax or adapt to the electronic distribution of the excited electron (which itself then adapts its wave function to the new potential, etc., leading to a self-consistent final state), the so-called relaxed excited state (RES) results (Pekar, 1951). Its energy is (Evrard, 1965; Devreese and Evrard, 1966)

$$E_{\mathrm{RES}} = -0.041\alpha^2\hbar\omega_{\mathrm{LO}}. \tag{3c}$$

In fact, both the FC state and the RES lie in the continuum* and, strictly speaking, are resonances (Fig. 2).

The strong-coupling mass of the polaron,

---

*The RES lies in the continuum of so-called scattering states, not indicated in Fig. 2.

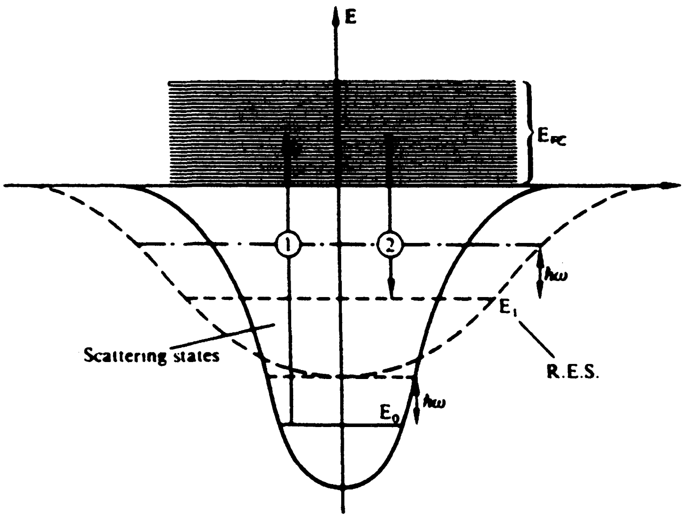

FIG. 2. Internal excitations at strong coupling: E_0, ground state; E_1, first relaxed excited state; E_{FC}, Franck–Condon state.

resulting again from the approximation (2), is given as

$$m^*/m_b = 1 + 0.0200\alpha^4. \tag{3d}$$

More rigorous strong-coupling expansions for E_0 and m^* have been presented in the literature (Miyake, 1975):

$$E_0/\hbar\omega_{\mathrm{LO}} = -0.108\,513\alpha^2 - 2.836, \tag{4a}$$

$$m^*/m_b = 1 + 0.022\,7019\alpha^4. \tag{4b}$$

The main significance of the strong-coupling theory is that it allows a test of "all-coupling" theories in the limit $\alpha \to \infty$. From the formal point of view, a deeper study of the status and uniqueness of the strong-coupling solutions was undertaken by Lieb (1977).

1.2.2. Weak Coupling Fröhlich (1954) has provided the first weak-coupling perturbation-theory results:

$$E_0 = -\alpha\hbar\omega_{\mathrm{LO}} \tag{5a}$$

and

$$m^* = \frac{m_b}{1 - \alpha/6}. \tag{5b}$$

Inspired by the work of Tomonaga, Lee *et al.* (1953) (usually cited as LLP) have derived Eq. (5a) and $m^* = m_b(1 + \alpha/6)$ from an elegant canonical transformation formulation. In fact, they perform two successive canonical transformations: first,

$$S_1 = \exp\left[\frac{i}{\hbar}\left(\mathbf{P} - \sum_{\mathbf{k}} \hbar\mathbf{k}a^\dagger_{\mathbf{k}}a_{\mathbf{k}}\right)\cdot\mathbf{r}\right], \tag{6a}$$

where $\mathbf{P}$ is the polaron total-momentum operator. The canonical transformation (6a) formally eliminates the electron operators from the Hamiltonian.

The second canonical transformation is of the "displaced-oscillator" form:

$$S_2 = \exp\left[\sum_{\mathbf{k}} (a^\dagger_{\mathbf{k}}f_{\mathbf{k}} - a_{\mathbf{k}}f^*_{\mathbf{k}})\right]. \tag{6b}$$

The $f_{\mathbf{k}}$ are treated as variational functions. The physical significance of Eq. (6b) is that it "dresses" the electron with the virtual phonon field that describes the polarization.

The LLP approximation has often been called "intermediate-coupling theory." However, its range of validity is in principle not larger than that of the weak-coupling approximation. The main significance of the LLP approximation is in the elegance of the canonical transformations S_1 and S_2, together with the fact that it puts the Fröhlich result on a variational basis.

In his basic article on polarons, Feynman (1955) found the following higher-order weak-coupling expansions:

$$E_0/\hbar\omega_{\mathrm{LO}} = -\alpha - 0.0123\alpha^2 - 0.00064\alpha^3 - \cdots \ (\alpha \to 0), \tag{7a}$$

$$m^*/m_b = 1 + \frac{\alpha}{6} + 0.025\alpha^2 + \cdots \ (\alpha \rightarrow 0). \tag{7b}$$

Since then a lot of theoretical work has been devoted to obtaining more exact coefficients in this expansion. Höhler and Müllensiefen (1959) calculated the coefficients for α^2 to be -0.016 in the energy and 0.0236 in the polaron mass. Röseler (1968) found the analytical expressions for the above-mentioned coefficients: $2\ \ln(\sqrt{2} + 1) - \frac{3}{2}\ \ln 2 - \frac{1}{2}\sqrt{2} \approx -0.015\,919\,62$ and $\frac{4}{3}\ \ln(\sqrt{2} + 1) - \frac{2}{3}\ \ln 2 - 5\sqrt{2}/8 + \frac{7}{36} \approx 0.023\,627\,63$, respectively. At present, the following most accurate higher-order weak-coupling expansions are known: for the energy (Smondyrev, 1986; Selyugin and Smondyrev, 1989),

$$E_0/\hbar\omega_{\mathrm{LO}} = -\alpha - 0.015\,919\,6220\alpha^2 - 0.000\,806\,070\,048\alpha^3 - \cdots, \tag{7c}$$

and for the polaron mass (Röseler, 1968),

$$m^*/m_b = 1 + \tfrac{1}{6}\alpha + 0.023\,627\,63\alpha^2 + \cdots, \tag{7d}$$

which are a striking illustration of the heuristical value of Feynman's path-integral approach in the polaron theory as follows from comparison of Eqs. (7a) and (7b) with Eqs. (7c) and (7d), respectively.

1.2.3 All-Coupling Theory. Feynman Path Integral In the early 1950s, H. Fröhlich gave a seminar at Caltech. In this seminar, he discussed the weak-coupling polaron mass as derived by him: $m^* = m_b/(1 - \alpha/6)$. He suggested that if the electron–phonon coupling could be accurately treated for intermediate coupling (in particular around $\alpha \approx 6$) this might lead to new insights in the theory of superconductivity (this was before BCS) (Feynman, 1973). Feynman was among the audience. He went to the library to study one of Fröhlich's papers on polarons (Fröhlich, 1954). There he got the idea to formulate the polaron problem into the Lagrangian form of quantum mechanics and then eliminate the field oscillators, "... in exact analogy to QED ... (resulting in) ... a sum over all trajectories. ..." The resulting path integral is of the form (Feynman, 1955)

$$\langle 0,\beta|0,0\rangle = \int \mathfrak{D}\mathbf{r}(\tau)\exp\left[-\tfrac{1}{2}\int_0^\beta \dot{\mathbf{r}}^2 d\tau + \frac{\alpha}{2^{3/2}}\int_0^\beta\int_0^\beta \frac{e^{-|\tau-\sigma|}}{|\mathbf{r}(\tau) - \mathbf{r}(\sigma)|}\,d\tau d\sigma\right], \tag{8}$$

where $\beta = 1/k_{\mathrm{B}}T$. This path integral (8) has a great intuitive appeal: It shows the polaron problem as an equivalent one-particle problem in which the interaction, nonlocal in time or "retarded," is between the electron and itself. Subsequently, Feynman showed (in fact to M. Baranger) how the variational principle of quantum mechanics could be adapted to the path integral and introduced a quadratic trial action (again nonlocal in time) to simulate Eq. (8). It may be noted that the elimination of the phonon field (or the boson field in general) introduced through Eq. (8) by Feynman has found many applications, e.g. in the study of dissipation phenomena.

Applying the variational principle for path integrals resulted in an upper bound for the polaron self-energy at all α, which at weak and strong coupling gave quite accurate limits. Feynman obtained the smooth interpolation between weak and strong coupling (for the ground-state energy). It is worth while to give the asymptotic expansions of Feynman's polaron theory. In the weak-coupling limit, they are given above by Eqs. (7a) and (7b). In another—strong-coupling—limit Feynman found for the energy

$$\frac{E_0}{\hbar\omega_{\mathrm{LO}}} \equiv \frac{E_{3\mathrm{D}}(\alpha)}{\hbar\omega_{\mathrm{LO}}} = -0.106\alpha^2 - 2.83 - \cdots \quad (\alpha \rightarrow \infty), \tag{9a}$$

and for the polaron mass

$$\frac{m^*}{m_b} \equiv \frac{m^*_{3\mathrm{D}}(\alpha)}{m_b} = 0.0202\alpha^4 + \cdots \quad (\alpha \rightarrow \infty). \tag{9b}$$

Over the years, the Feynman model for the polaron has remained in many respects the most successful approach to this problem. Only recently an equivalent Hamiltonian formulation of this path-integral approach has been realized ($\omega_c = 0$) by Devreese and Brosens (1992). [In one case (Yamazaki, 1983), the formal structure of the theory was reobtained in a Hamiltonian formulation—be it very artificial—but no variational principle leading to an upper bound for the energy could be found.]

The theory of polarons in semiconductors with degenerate bands and of statistical ensembles of polarons, related to many-polaron problems, is reviewed from the unified point of view of the path integration over the Wick symbols by Fomin and Pokatilov (1988).

1.2.4 Response Properties of Fröhlich Polarons. Mobility and Optical Properties

The transport properties of polar and ionic solids are influenced by the polaron coupling. Intuitively one expects that the mobility of large polarons will be inversely proportional to the number of real phonons present in the crystal:

$$\mu \sim \exp(\hbar\omega_{\mathrm{LO}}/k_{\mathrm{B}}T). \tag{10a}$$

The first to point out the typical behavior (characteristic for weak coupling) of Eq. (10a) was Fröhlich (1937). Kadanoff (1963) provided a derivation of the weak-coupling, low-temperature mobility starting from the Boltzmann equation; his result is

$$\mu = \frac{4}{3\pi^{1/2}}\,\frac{e}{m^{*}\alpha\omega_{\mathrm{LO}}}\,G(z)\exp\!\left(\frac{\hbar\omega_{\mathrm{LO}}}{k_{\mathrm{B}}T}\right)\left(\frac{k_{\mathrm{B}}T}{\hbar\omega_{\mathrm{LO}}}\right)^{1/2}, \tag{10b}$$

where $G(z)$, defined by Howarth and Sondheimer (1953), is of order 1.

Feynman *et al.* (1962) (usually referred to as FHIP) have elaborated a framework allowing the analysis of the response properties of a system, using path integrals. For low temperatures, they obtain the following expression for the dc mobility:

$$\mu \equiv \mu_{3\mathrm{D}}(\alpha) = \frac{e}{2m^{*}\omega_{\mathrm{LO}}\alpha}\,\frac{3}{2}\left(\frac{w}{v}\right)^{3} \times \exp\!\left(\frac{v^{2}-w^{2}}{w^{2}v}\right)\exp\!\left(\frac{\hbar\omega_{\mathrm{LO}}}{k_{\mathrm{B}}T}\right)\frac{k_{\mathrm{B}}T}{\hbar\omega_{\mathrm{LO}}}, \tag{10c}$$

where w and v are functions of α (also adequate for large α) deriving from the Feynman polaron model.* For the impedance of a polaron $Z(\nu) \equiv Z_{3\mathrm{D}}(\alpha,\nu)$ describing its response to the ac electric field of frequency ν, see Feynman *et al.* (1962).

Equation (10c) and related approximations allowed explanation of the experimental results of Brown and his co-workers (see Brown, 1963) on alkali halides and silver halides (see, e.g., Fig. 3) in the temperature region where the electron–LO-phonon scattering is the dominant process (at low temperatures $T < 50$ K the impurity scattering starts to prevail). Equation (10c) presupposed a "drifted-Maxwellian" velocity distribution of the polarons; this limits the validity of the treatment theoretically. However, because other scattering mechanisms than LO-phonon scattering might "conspire" to induce a Maxwellian distribution, the range of validity of Eq. (10c) seems to be larger than one would expect from electron–LO-phonon interaction only. A review on the mobility of Frölich polarons, written from a unified point of view, was presented by Peeters and Devreese (1984).

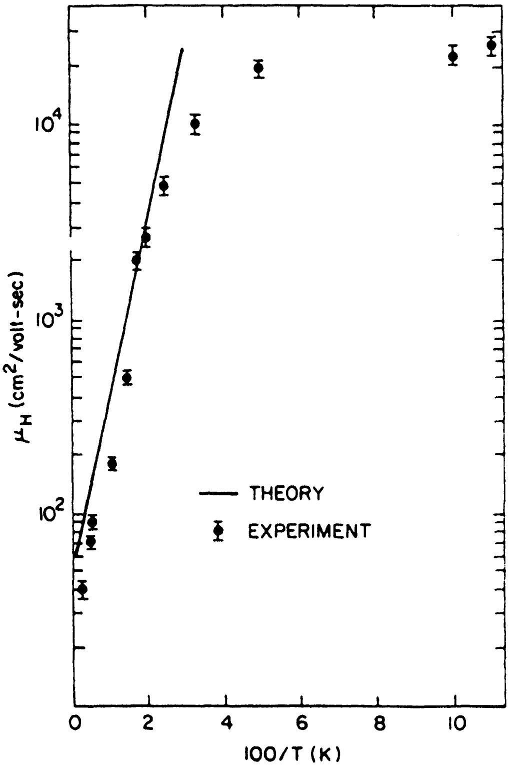

FIG. 3. Polaron mobility in AgBr (solid line) according to the low-temperature self-consistent approach of Devreese and co-workers (Kartheuser *et al.*, 1979) compared with the Hall data, from Brown (1981).

---

*It is noted that a misprint occurred in this formula in the *Encyclopedia of Physics* (Lerner and Trigg, 1991).

It was shown by the author and his coauthors (see Devreese, 1972) how the optical absorption of Fröhlich polarons, for all coupling, can be calculated starting from the FHIP scheme (which gives a derivation for the impedance function). This work led to the following expression for the optical absorption:

$$\Gamma(\nu) \sim \frac{\mathrm{Im}\Sigma(\nu)}{[\nu - \mathrm{Re}\Sigma(\nu)]^2 + [\mathrm{Im}\Sigma(\nu)]^2}, \qquad \text{(10d)}$$

where ν is the frequency of the incident radiation and $\Sigma(\nu)$ is the so-called "memory function" that contains the dynamics of the polaron and depends on α and ν. As an example, in Fig. 4 the optical-absorption spectrum of Fröhlich polarons for $\alpha = 6$ is shown.

It is remarkable that from Eq. (10d), in Devreese (1972) the three different kinds of polaron excitations, studied before only in asymptotic limits, are seen to appear in the spectra:

1. scattering states where, e.g., one real phonon is excited (the structure starting at $\nu = 1$);
2. RES; and
3. FC states.

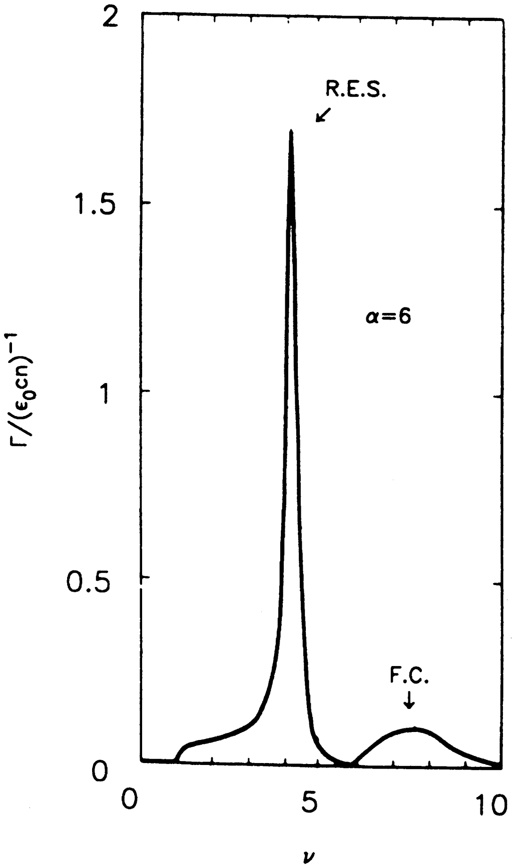

FIG. 4. Optical absorption Γ of polarons at $\alpha = 6$ as a function of frequency ν, expressed in units ω_{LO}, from Devreese (1972).

2 and **3** can be seen in the works of the author and co-workers [Kartheuser *et al.*, 1969 (KED); Devreese and Evrard, 1966; Evrard, 1965]. Experimentally only the "scattering states" have been observed for free polarons (Finkenrath *et al.*, 1969); the RES play a role for bipolarons (Devreese *et al.*, 1994). However, the full structure of Eq. (10d) has been revealed through cyclotron-resonance measurements for which $\nu - \mathrm{Re}\Sigma(\nu)$ is replaced by $\nu - \omega_c - \mathrm{Re}\Sigma(\nu)$ so that the resonance conditions can be tuned by changing ω_c (Peeters and Devreese, 1986).

The weak-coupling limit of Eq. (10d) coincides with the results by Gurevich *et al.* (1962), whereas the structure of the strong-coupling limit confirms the identification of the internal polaron excitations by Devreese and collaborators (KED).

In pioneering experimental studies, Brown and co-workers (see Brown, 1963) have combined mobility experiments and cyclotron-resonance measurements to demonstrate clearly the polaron effect. From a theoretical plot of mobility vs band mass in AgBr compared with experimental mobility data at a given temperature, they estimate the band mass. This allows them to calculate α and the polaron mass m^*. This value of m^* can then be compared with the measured cyclotron mass. The experimental value $m^*/m_b = 0.27 \pm 0.01$ is obtained for AgBr at 18 K, to be compared with a theoretical value $m^*/m_b = 0.27 \pm 0.05$.

1.3 Fröhlich Polarons in 2D

1.3.1 Introduction Today electron systems in reduced dimensions—e.g., in two dimensions like in GaAs–AlGaAs or MOSFETS—are of great interest. Also the electron–phonon interaction and the polaron effect in such systems receive much attention. For one polaron, confined to two dimensions, but interacting with a 3D phonon gas, the Fröhlich Hamiltonian remains of the form (1a) with the following modification for the V_k (Peeters *et al.*, 1986a):

$$V_k = -i\hbar\omega_{\mathrm{LO}}\left(\frac{\sqrt{2}\pi\alpha}{Ak}\right)^{1/2}\left(\frac{\hbar}{m_b\omega_{\mathrm{LO}}}\right)^{1/4}, \tag{11}$$

valid because the electron–polarization interaction is of the standard form $1/r$ in an arbitrary number of space dimensions. A possible dependence of the electron–phonon interaction on a concrete physical mechanism of the electron confinement to two dimensions was considered by Fomin and Smondyrev (1994). The self-energies $\Delta E/\hbar\omega_{\mathrm{LO}}$ for a polaron in 2D for $\alpha \to 0$ and $\alpha \to \infty$ were derived by Devreese and co-authors (Xiaoguang *et al.*, 1985, XPD) and Das Sarma and Mason (1985):

$$\frac{\Delta E}{\hbar\omega_{\mathrm{LO}}} \equiv \frac{E_{\mathrm{2D}}(\alpha)}{\hbar\omega_{\mathrm{LO}}} = -\frac{\pi}{2}\alpha - 0.063\,97\alpha^2 + O(\alpha^3) \quad (\alpha \to 0), \tag{11a}$$

$$\frac{\Delta E}{\hbar\omega_{\mathrm{LO}}} \equiv \frac{E_{\mathrm{2D}}(\alpha)}{\hbar\omega_{\mathrm{LO}}} = -0.4047\alpha^2 + O(\alpha^0) \quad (\alpha \to \infty). \tag{11b}$$

The corresponding results for the polaron mass in 2D are (XPD)

$$\frac{m^*}{m_b} \equiv \frac{m^*_{\mathrm{2D}}(\alpha)}{m_{\mathrm{2D}}} = \frac{\pi}{8}\alpha + 0.127\,2348\alpha^2 + O(\alpha^3) \quad (\alpha \to 0), \tag{11c}$$

$$\frac{m^*}{m_b} \equiv \frac{m^*_{\mathrm{2D}}(\alpha)}{m_{\mathrm{2D}}} = 0.733\alpha^4 + O(\alpha^2) \quad (\alpha \to \infty). \tag{11d}$$

The result $\Delta E/\hbar\omega_{\mathrm{LO}} = -(\pi/2)\alpha$ was first obtained by Sak (1972).

1.3.2 Scaling Relations Peeters and Devreese (1987) have derived several scaling relations connecting the polaron self-energy, the effective mass, the impedance Z, and the mobility μ in 2D to the same quantities in 3D. Those relations were derived on the level of the Feynman approximation and are listed here:

$$E_{\mathrm{2D}}(\alpha) = \tfrac{2}{3}E_{\mathrm{3D}}(3\pi\alpha/4), \tag{12a}$$

$$\frac{m^*_{\mathrm{2D}}(\alpha)}{m_{\mathrm{2D}}} = \frac{m^*_{\mathrm{3D}}(3\pi\alpha/4)}{m_{\mathrm{3D}}}, \tag{12b}$$

$$Z_{\mathrm{2D}}(\alpha,\nu) = Z_{\mathrm{3D}}(3\pi\alpha/4),\nu), \tag{12c}$$

where ν is the frequency of the external electromagnetic field, and

$$\mu_{\mathrm{2D}}(\alpha) = \mu_{\mathrm{3D}}(3\pi\alpha/4). \tag{12d}$$

Except for the investigations on polarons at the surface of liquid He (Jackson and Platzman, 1981; Peeters, 1987), the experimental studies performed at present are related to systems with weak electron–phonon coupling. Those studies have addressed a variety of physical properties including magnetophonon anomalies, optical absorption, plasmon–LO-phonon mode coupling, cyclotron resonance, mobility, many-body effects, etc. The reader is referred to Devreese and Peeters (1987) for more details and additional references. In what follows, some of these studies will be further discussed.

1.4 Fröhlich Polarons in a Magnetic Field

1.4.1 In 3D Fröhlich polarons have been most clearly manifested by investigations of their properties in magnetic fields. Therefore a special section is devoted to the study of polarons in magnetic fields.

1.4.1.1 Level Crossing, Pinning. In interpreting the (low field) cyclotron-resonance experiment of Brown *et al.* (1963), it was supposed that the zero-magnetic-field polaron mass is observed. Of course, there exists no *a priori* guarantee that this supposition is true, and theoretical studies of polarons in magnetic fields are necessary. Larsen (1972, 1991) has made important contributions to the theoretical study of polarons in magnetic fields. In particular, he was the first to point out the level repulsion close to the crossing of levels at $\omega_c = \omega_{\mathrm{LO}}$ (ω_c is the cyclotron-resonance frequency) and the pinning of Landau levels to the phonon continuum as $\omega_c \to \infty$ (see Fig. 5). Measurements on InSb (Johnson and Larsen, 1966) and CdTe (Waldman *et al.*, 1969) provided first indications for these level-crossing and pinning phenomena. Detailed line-shape studies for weak coupling

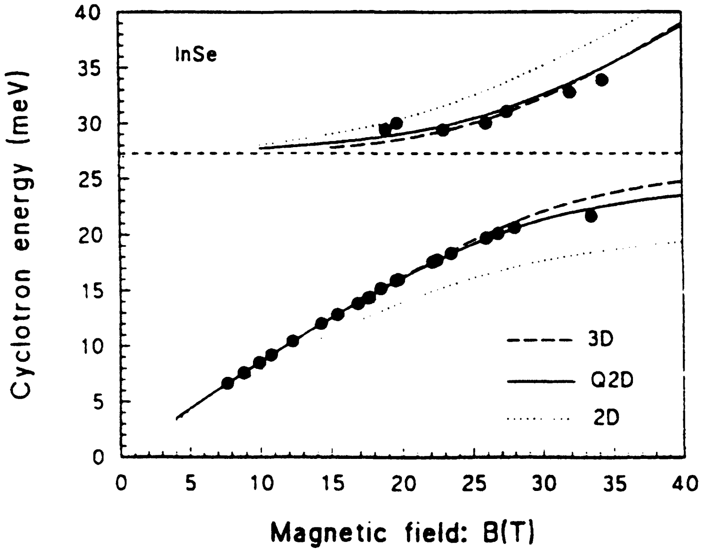

FIG. 5. The cyclotron-resonance position plotted as a function of magnetic field for InSe, from Nicholas *et al.* (1992).

of the cyclotron resonance, revealing a double-peak structure close to $\omega_c = \omega_{\mathrm{LO}}$, are displayed by Vigneron *et al.* (1978) and independently by Devreese and his co-workers (see Van Royen *et al.*, 1977).

1.4.1.2 Static and Dynamic Properties of Polarons in a Magnetic Field. Peeters and Devreese (1982) have generalized the Feynman model of the polaron to the case where a static external magnetic field is applied. The calculation is valid for all α, ω_c, and temperature described by the parameter $\beta = 1/k_BT$. The starting point is the expression of the free energy of the polaron as a path integral,

$$F = F_{\mathrm{ph}} - \frac{1}{\beta}\ln\left\{\int d\mathbf{r}\int_{\mathbf{r}(0)=\mathbf{r}}^{\mathbf{r}(\beta)=\mathbf{r}} \mathcal{D}\mathbf{r}(u)\exp(S[r(u)])\right\}. \tag{13a}$$

The "electron" contribution to the action S is

$$S_e = -\tfrac{1}{2}\int_0^\beta du\{\dot{\mathbf{r}}(u)^2 + i\omega_c[x(u)\dot{y}(u) - y(u)\dot{x}(u)]\}, \tag{13b}$$

where $\mathbf{r}$ is the position vector of the electron, with components x, y in the plane perpendicular to the magnetic field.

A quadratic, retarded model interaction was introduced (Peeters and Devreese, 1982) to simulate the polaron (retarded Coulomb) interaction and, in analogy to the zero-magnetic-field case, the Jensen–Feynman inequality was assumed to be valid:

$$F \le F_{\mathrm{ph}} + F_m - \frac{1}{\beta}\langle S - S_m\rangle_m. \tag{13c}$$

Here F_{ph} and F_m stand for the free energy of the phonon bath and the quadratic model, respectively, while $\langle\cdot\rangle_m$ denotes an evaluation of the corresponding average with

$$e^{S_m}\Big/\int d\mathbf{r}\int_{\mathbf{r}(0)=\mathbf{r}}^{\mathbf{r}(\beta)=\mathbf{r}} \mathcal{D}\mathbf{r}(u)e^{S_m} \tag{13d}$$

as weight factor. For details, the reader is referred to Peeters and Devreese (1982), where results for the energy and related properties of the polaron in a magnetic field for all α, ω_c, and β were derived both numerically and—in a variety of limiting cases—analytically.

A question of considerable significance has been raised about the validity of the inequality (13c) in the presence of a magnetic field. Feynman suggested that in a magnetic field this inequality remains valid, or might need a slight modification only (Feynman and Hibbs, 1965). Several works have been devoted to this challenging problem; see Brosens and Devreese (1988), Sec. 3.4 in Fomin and Pokatilov (1988), and Larsen (1991) for detailed reviews. Larsen (1985) revealed that the ground-state levels of a 2D polaron obtained "variationally" on the basis of Eq. (13c) for sufficiently high magnetic fields lie below those found within the framework of the adiabatic strong-coupling theory or the fourth-order–perturbation weak-coupling approach. Interpreting the latter ground-state level as the exact one, Larsen came to the conclusion that it would seem difficult to attach any particular sense to the variational ground-state level. However, it was pointed out that the adiabatic strong-coupling theory does not provide a rigorous asymptotic limit of high magnetic fields; see discussion in Devreese and Brosens (1991). It is also worth while to note that the perturbation ground-state energy can, however, itself be treated in some cases (see Fomin and Pokatilov, 1988, *loc. cit.*) as an upper bound for the exact energy. But when comparing various upper bounds for the exact energy, one should recall Feynman's (1955) warning that attempts to improve an upper bound by calculating the higher-order correction terms may indeed deprive the treatment of its variational nature! Using a model calculation, Brosens and Devreese (1988) rigorously demonstrated that for sufficiently small electron–phonon coupling the presence of a magnetic field is prohibitive for the application of the Jensen–Feynman inequality.

A generalization of the Jensen–Feynman inequality, which remains valid in the case of a nonzero magnetic field, was derived by Devreese and Brosens (1992) starting from the ordered-operator calculus. On these grounds, the conditions were determined to be imposed on the variational parameters in the model action S_m such that the Feynman upper bound in its original form of the inequality (13c) remains valid for a polaron in a magnetic field. Although so far it has not been conclusively established that a choice of the parameters in the trial action made in Peeters and Devreese (1982) limits them to the domain determined by the conditions derived in the above-cited work, it is interesting to note that most of the existing theories of polarons in a magnetic field (Hellwarth and Platzman, 1962; Marshall and Chawla, 1970; Evrard *et al.*, 1970; Lépine and Matz, 1976) are obtained as special cases of the results by Peeters and Devreese (1982) who simply accepted the inequality (13c) as a working hypothesis. A prediction by Peeters and Devreese (1982) is that some quantities characterizing the internal structure of the polaron [called, e.g., $(\nu_\perp/\omega_\perp)^2$ in the mentioned paper] undergo a drastic change for a well-defined magnetic field ("stripping transition"). Although high magnetic fields are needed (e.g., ~42 T in AgBr), it would be interesting to investigate this point experimentally.

1.4.1.3 Cyclotron-Resonance Spectra. The observation of the cyclotron mass of electrons in AgBr in the low-field case ($\omega_c \to 0$) gave evidence for the occurrence of polaron effects. However it concerns only one number (m^*/m_e) and involves the combination of two measurements (m_e is the mass of the electron in vacuum). It was therefore useful to analyze the magnetic-field dependence of the polaron mass in order to gain quantitative insight into the validity of the polaron picture.

An excellent occasion to realize such an analysis is provided by the more recent precise cyclotron-mass measurements (in AgBr and AgCl) by Hodby *et al.* (1987). These measurements, performed with the 5-cm-bore hybrid magnet at Oxford, cover the range from zero magnetic field to 16 T. These measurements are precise enough to distinguish between various polaron theories. Several theories were compared in analyzing the experimental data of Hodby *et al.* (1987).

First, the variational calculation of Larsen (1972) was considered. This approach is a so-called intermediate-coupling theory to calculate the energy levels (modified Landau levels) of a polaron in a magnetic field. The polaron mass is then defined from the energy differences between the polaron (Landau) energy levels.

In principle, it is better to calculate the magneto-optical absorption spectrum of the polaron (the quantity that is actually measured) and to define the polaron mass, in the same way as the experimentalist, from the peak positions in the spectrum. Starting from the results of Peeters and Devreese (1982), the magneto-optical absorption of polarons for all α and ω_c at $T = 0$ was calculated by Peeters and Devreese (1986). They used the memory-function formalism to generalize the study of the response of a Feynman polaron in a magnetic field. The magneto-absorption is then obtained from

$$\lim_{\epsilon \to 0} \frac{1}{\nu - \omega_c - \Sigma(\nu + i\epsilon)}, \tag{14a}$$

where ν is the frequency of the incident radiation and $\Sigma(\nu + i\epsilon)$ is the memory function. The key ingredient of $\Sigma(\nu + i\epsilon)$ is the space Fourier transform of the density–density correlation function:

$$\langle e^{i\mathbf{k}\cdot\mathbf{r}(t)} e^{-i\mathbf{k}\cdot\mathbf{r}(0)} \rangle, \tag{14b}$$

where $\langle \cdot \rangle$ can be expressed as a path integral. $\Sigma(\nu + i\epsilon)$ is an intricate function that takes into account all the polaron internal states and all the Landau levels. It turns out that the magneto-absorption calculated by Peeters and Devreese (1986) leads to the best quantitative agreement between theory and experiment as was analyzed for AgBr and AgCl (Hodby *et al.,* 1987). It should be pointed out that the weak-coupling theories (Rayleigh–Schrödinger perturbation theory, Wigner–Brillouin formulation and its improvements) fail (and are all off by at least 20% at 16 T) to describe the experimental data for the silver halides. The analysis of Hodby *et al.* (1987) based on the theory of magnetoabsorption by Peeters and Devreese (1986) provides a confirmation of the Fröhlich description of the polaron in a case where weak-coupling approximations are not adequate.

An interesting case is provided by the cyclotron-resonance data for CdTe (Johnson and Larsen, 1966). The early analysis (Larsen, 1972) of these experiments as well as subsequent cyclotron-resonance measurements of the polaron effective mass in *n*-CdTe (Litton *et al.,* 1976) seemed to suggest that a Fröhlich coupling constant as large as $\alpha = 0.4$ could explain the data for the polaron mass as a function of magnetic field. This gave rise to a long-standing challenge about the adequacy of the dielectric continuum model for the magneto-bound polarons in CdTe, which maintains the value $\alpha \sim 0.3$ of the coupling constant. But it turns out that the experimental data on the Zeeman splitting in the $1s \to 2p$ shallow-donor–impurity transitions in CdTe at high magnetic fields obtained by Cohn *et al.* (1972) can be precisely described in the framework of a second-order perturbation theory with band nonparabolicity taken into account using the value $\alpha = 0.286$ for the Fröhlich coupling constant following from Eq. (1c) (Shi *et al.,* 1995). It has been revealed in the latter work that Cohn *et al.* (1972) had to use as a fitting parameter the value $\alpha = 0.4$, higher than the above-mentioned Fröhlich coupling constant, in order to compensate for the underestimation of the polaron effects in the calculation of the transition energies.

1.4.2 Cyclotron Resonance of Polarons in 2D Cyclotron-resonance experiments have been performed on the 2DEG, e.g., in InSb inversion layers and in GaAs–$\text{Al}_x\text{Ga}_{1-x}$As heterostructures (Scholz *et al.,* 1983; Seidenbusch *et al.,* 1984; Sigg *et al.,* 1985; Merkt, 1985; Hopkins *et al.,* 1989). Several theoretical studies have been presented to analyze these experimental results for the 2DEG. In those works, the cyclotron mass was obtained from the positions of the energy levels (Das Sarma, 1984; Larsen, 1984a,b; Peeters and Devreese, 1985; Peeters *et al.,* 1986b,c).

The theory of cyclotron resonance in the 2DEG for cases where the electron–phonon interaction plays a significant role is reviewed by Devreese and Peeters (1987). Like in 3D (Peeters and Devreese, 1986), the theory is expressed in terms of the memory-function formalism. In this treatment, the magneto-optical absorption itself is calculated and the transition frequencies (rather than the individual energy levels) are obtained directly.

Some results of this work are the following:

1. First the magneto-absorption spectrum was calculated, at weak coupling, for one polaron in 2D. A Landau-level broadening parameter is introduced phenomenologically in order to remove the divergences in the magneto-optical absorption spectrum. The effect of the nonzero width of the 2DEG is incorporated along with nonparabolicity. The experimental data for *p*-InSb inversion layers can be adequately explained by this theory.
2. To account for the cyclotron-mass data in GaAs–$\text{Al}_x\text{Ga}_{1-x}$As heterostructures it is essential to include many-body effects. Both the "occupation effect" (Pauli principle) and the effects of screening were included, on top of the effect included in the one-polaron studies under **1**.

It is also worth mentioning that, although for one electron the polaron effect is enhanced by the 2D confinement, in reality—e.g., in GaAs–Al$_x$Ga$_{1-x}$As with $n_e \geq 1.4 \times 10^{11}$ cm^{-2}—screening helps to reduce the polaron effect in 2D so that it becomes smaller than its 3D counterpart. For sufficiently large densities, Peeters *et al.* (1988a,b) revealed that the polaron mass in 2D is not a monotonically increasing function of ω_c but shows structure where the filling factor becomes an integer.

It is obvious, in both the case of InSb and that of GaAs–Al$_x$Ga$_{1-x}$As, that polaron effects do occur, even if these are weak-coupling materials.

A nice example, clearly demonstrating the polaron coupling, is provided using the memory-function approach by the author and his colleagues (Peeters and Devreese, 1985, 1986c; Devreese and Peeters, 1986) through the analysis of the cyclotron resonance in a 2DEG that naturally occurs in InSe where $\alpha \approx 0.3$ (Nicholas *et al.*, 1992). One clearly sees, over a wide range of magnetic fields, the two distinct polaron branches separated by as much as $0.4\omega_{\mathrm{LO}}$ at resonance (Fig. 5). Polaron cyclotron resonance has even been observed in n-type ZnS up to 220 T by Miura *et al.* (1994).

A quantitative interpretation of the cyclotron-resonance measurements in n-GaAs and AlGaAs–GaAs heterojunctions was obtained by Peeters *et al.* (1988a,b) on the grounds of the polaron theory with three factors taken into account: dimensionality, band nonparabolicity, and screening. Impurity-bound resonant magnetopolarons have been clearly observed in bulk GaAs and GaAs–Al$_x$Ga$_{1-x}$As multiple quantum wells (Cheng *et al.*, 1993).

1.4.3 Formal Developments Also to treat Fröhlich polarons in a magnetic field the Feynman path integral proved to be a most powerful method extending the possibilities of perturbation theory and Hamiltonian variational calculations.

From the methodological point of view, the introduction of the total angular-momentum operator of the polaron,

$$\hat{L}_z = \hat{l}_z + i\hbar \sum_{\mathbf{k},\mathbf{k}'} a^\dagger_{\mathbf{k}'} a_{\mathbf{k}} \left(k_x \frac{\partial}{\partial k_y} - k_y \frac{\partial}{\partial k_x} \right) \delta_{\mathbf{k},\mathbf{k}'} \tag{15}$$

(with $\hat{l}_z$ denoting the electron angular momentum), is useful for Hamiltonian treatments of polarons in a magnetic field performed by Evrard *et al.* (1970). Later this operator was used in studies on excitons.

An operator-algebra method was developed by Larsen (1984a), useful to study higher-order effects for polarons in magnetic fields, when expanding in powers of α.

1.5 The Bound Polaron

A polaron can be bound to a charged vacancy or to a charged interstitial. To a first approximation, this system can be approximated by adding the Coulomb potential energy operator ($-e^2/\epsilon_\infty|\mathbf{r}| \equiv -\beta/|\mathbf{r}|$, where $\mathbf{r}$ is the vector operator characterizing the electron position with respect to the center of the vacancy or of the interstitial) to the Fröhlich Hamiltonian.

Intuitively one expects that the weak-coupling polaron spectrum for the bound polaron is approximately given by a Bohr formula adapted to take into account the polaron mass:

$$E_n = -\alpha\hbar\omega_{\mathrm{LO}} - \frac{m_b e^4}{2\hbar^2 n^2 \epsilon_0^2} - \frac{\alpha}{12}\frac{m_b e^4}{\hbar^2 n^2} + O(\alpha^2)\ (n = 1, 2, \cdots). \tag{16}$$

Expansions refining Eq. (16) were derived first using approximate schemes for perturbation theory (Bajaj, 1972); later a rigorous result for the binding energy E_n up to order α was obtained (Engineer and Tzoar, 1972).

Further schemes to treat the ground-state energy of the bound polaron (and approximations for some of the excited states) have been developed by Platzman (1962), Larsen (1969), Devreese *et al.* (1982), and Adamowski (1985).

Brandt and Brown (1969) have interpreted some structure in their infrared optical absorption spectra of AgBr as caused by the bound polaron; in particular the 168-cm^{-1} absorption line has been analyzed as a transition between a 1s and a 2p state (modified by the polaron interaction). Also higher excited states and LO-phonon sidebands play a role in this spectrum (see e.g., Bajaj, 1972) (Fig. 6). The bound polaron is related to the F center. In Tables 2 and 3, some energy levels of the bound polaron are tabulated.

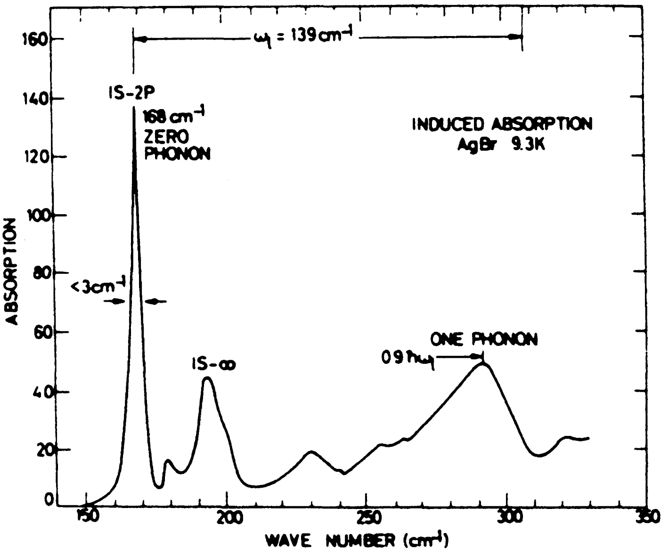

FIG. 6. Induced absorption in AgBr at 9.3 K from 150 to 320 cm^{-1}. The exciting light was in the region of the direct-absorption edge. From Brandt and Brown (1969), reproduced by courtesy of The American Physical Society.

Table 2. Upper bounds for the ground-state energy (in units $\hbar\omega_{\text{LO}}$) of the bound polaron for several values of α and $\tilde{\beta}$ as obtained by Devreese *et al.* (1982) compared with the results by Larsen (1969).

	$\tilde{\beta} = 6.32$		$\tilde{\beta} = 4.47$	
α	E (Devreese *et al.*, 1982)	E (Larsen, 1969)	E (Devreese *et al.*, 1982)	E (Larsen, 1969)
2	−14.66	−14.69	−8.60	−8.64
5	−23.0	−23.0	−15.21	−15.30
7	−29.41	−29.47	−20.52	−20.62
11	−44.6	−44.6	−33.4	−33.4

2. THE SMALL POLARON

2.1 The Small-Polaron Concept. Role of Localization

An electron or a hole trapped by its self-induced atomic (ionic) displacement field in a region of linear dimension ("radius") that is of the order of the lattice constant is called a *small polaron* (Fröhlich, 1957; Sewell, 1958; Fröhlich and Sewell, 1959; Holstein, 1959; Emin and Holstein, 1976). An excellent survey of the small-polaron physics relevant to conduction phenomena in noncrystalline materials and to metal–insulator transitions has been given by Mott (1987, 1990). As distinct from large polarons, small polarons appear as a result of overall short-range forces. Thus, in certain materials, in particular in some oxides, the induced lattice polarization is essentially localized in a volume of the order of a unit cell. Hence, the charge carrier is localized on an individual lattice site during a time that can become large compared with the *localization time* describing the relaxation of the lattice to the small-polaron state. The localization time is of the same order of magnitude as the period of a lattice vibration, ω_{LO}^{-1}.

Because for small polarons the lattice polarization is mostly confined to one unit cell, the atomicity of the solid is felt by the car-

Table 3. Calculated energy of the first excited $2p$ state (in units $\hbar\omega_{\text{LO}}$) of the bound polaron for several values of α and $\tilde{\beta}$ as obtained by Devreese *et al.* (1982).

	E_{2p}		
α	$\tilde{\beta} = 0$	$\tilde{\beta} = 1$	$\tilde{\beta} = 2$
1	−0.7626	−0.8775	−1.113
3	−1.971	−2.237	−2.628
5	−3.207	−3.644	−4.205
7	−4.600	−5.214	−5.955
9	−6.199	−6.996	−7.922
11	−8.029	−9.014	−10.13

rier; a complete treatment of small polarons should therefore start from an *ab initio* calculation that takes into account the detailed local structure of the solid; the Fröhlich continuum approximation would not be adequate. Nevertheless, actual small-polaron theories as developed, e.g., by Yamashita and Kurosawa (1958), Holstein (1959), and others are based on analytical approximations as a starting point. Thus, the adiabatic eigenstates of an electron placed in a deformable continuum were shown to depend drastically on the character of the electron–lattice interaction as well as on the dimensionality of the system (Emin and Holstein, 1976). As distinct from the case of the long-range interaction with a stable large-polaron state, for the short-range interaction in a three-dimensional system there exist two stable states, namely, an unbound electron in an undeformed continuum and an electron collapsed in an infinitesimally localized self-induced potential well. The former is analogous to the band-electron state in a rigid lattice; the latter models a small-polaron state. It is also worth while noting that the common action of a long-range and of a short-range force was found to yield always a small-polaron–like state and—in a certain region of the interaction strengths—a large-polaron state. Thus, even from the early analysis the possibility of *coexistence* of both types of polarons can be distinctly deduced.

In the modern theories (see, e.g., Alexandrov and Mott, 1994), the small-polaron energy is regarded to consist of the following parts:

1. the kinetic energy of the charge carrier in a rigid lattice, which, as distinct from large polarons, is considered to originate from the intersite transfer due to tunneling;
2. the energy of the atomic (ionic) displacements field, describing the lattice distortion; and
3. the potential energy of the charge carrier in the potential well formed by these displacements.

The interaction of the localized electron (hole) with the lattice vibrations then induces the charge carrier to jump from one atom (or ion) to a neighboring one. This process is called *hopping*. The detailed physical picture of hopping (Lang and Firsov, 1962) suggests a sequence of the acts of small-polaron disintegration and reappearance as follows. At sufficiently high temperatures $k_BT > \hbar\omega_{LO}/4$, the typical time interval between jumps Δt satisfies the inequalities $t_0 \ll \Delta t \ll t_p$, where $t_0 \sim \hbar/(W_H k_B T)^{1/2}$ (with W_H the thermal activation energy for hopping) denotes the *jump-over* time and $t_p \sim \hbar/\Delta E_p$ (with ΔE_p the small-polaron bandwidth) is the tunneling time. Hence, an electron remains most of the time at a site, undergoing a hopping transition from site to site rather rarely, but on average earlier than a tunneling occurs. If the jump-over time is much shorter than the period of a lattice vibration, $t_0 \ll \omega_{LO}^{-1}$, under a hopping transition the electron "jumps out of" the old potential well due to its self-induced lattice deformation, thus initiating a multiphonon process due to the lattice relaxation: The small-polaron state disappears. But the time interval between jumps is much larger than the localization time: $\Delta t \gg \omega_{LO}^{-1}$. This inequality describes the *anti-adiabatic limit:* the atoms (ions) can adiabatically follow the motion of an electron, in contrast with the case of a large polaron (see Sec. 1.2.1). Thus, a new potential well due to the lattice deformation adapted to the new position of the electron is formed. A small-polaron state reappears.

Also in the case of small polarons the relevant phonons are commonly the LO phonons. A simple estimate of the temperature dependence of the mobility can be obtained starting from the following. The larger the number of LO phonons n_{ph} present in the solid, the larger the mobility μ_{SP} of the small polaron:

$$\mu_{SP} \sim n_{ph}. \tag{17}$$

For sufficiently low temperatures $n_{ph} = \exp(\hbar\omega_{LO}/k_BT)$ and, as a consequence, we find

$$\mu_{SP} \sim \exp(\hbar\omega_{LO}/k_BT). \tag{18}$$

It should be emphasized that the mobility of small polarons is therefore *thermally activated* and its temperature dependence is totally different from that of Fröhlich polarons [compare, e.g., Eq. (18) with Eq. (10a)].

A more detailed theoretical treatment of the small-polaron mobility (see Lang and Firsov, 1962; Reik, 1972) leads to the following formula, valid for $T > \theta_D/2$ (θ_D is the Debye temperature of the crystal):

$$\mu_{\mathrm{SP}} = \frac{ea^2\omega_{\mathrm{LO}}}{6k_{\mathrm{B}}T}\exp\left(-\frac{W_H}{k_{\mathrm{B}}T}\right), \tag{19}$$

where a is the lattice constant of the crystal in which the small polaron occurs; W_H is the thermal activation energy for hopping and is given by $\frac{1}{2}$ the small-polaron binding energy.

The Arrhenius-type activated behavior of the form (19) of the mobility has been used as a fingerprint to identify small-polaron behavior in solids. The earliest studies concerned one of the uranium oxides, UO_{2+x} (Devreese, 1963; Nagels *et al.*, 1964). Subsequently, small polarons were invoked to interpret the conductivity in many oxides (Mott and Davis, 1979), in particular in transition-metal oxides. It should be mentioned that often the measured quantity is the Hall mobility rather than the drift mobility. The theory of the Hall mobility of small polarons due to hopping (Friedman and Holstein, 1963; Austin and Mott, 1969) leads to

$$\mu_{\mathrm{Hall}} \propto T^{-1/2}\exp\left(-\frac{W_H}{3k_{\mathrm{B}}T}\right). \tag{20}$$

The relation between Hall and drift mobility is not simple, and the Hall mobility depends, e.g., on the interference between several hopping processes [see a comprehensive review by Austin and Mott (1969)]. For s carriers, the Hall coefficient turns out to be always negative.

2.2 Standard Small-Polaron Theory

As shown above, small polarons—at sufficiently high temperature—are characterized by diffusive motion and the band picture with its Bloch states breaks down; in the low-temperature limit the band picture reappears in the theoretical description, although experimental evidence for band conduction has been limited.

The fact that the Bloch-band picture breaks down is connected with the narrowing of the band gap that develops as the carrier becomes more and more localized, resulting in an increasing effective mass and, in the limit, in self-trapping. For example, in KCl the material characteristics are such that a valence-band hole gets self-trapped because of its polaron interaction with the lattice (Stoneham, 1979). The self-trapping of the hole is a subtle process, but the evidence is that it is related to the polaron formation in interplay with the Jahn–Teller effect (Stoneham, 1979).

For the quantitative treatment of small polarons, the so-called molecular-crystal model of Holstein (1959) is perhaps most illuminating. Without going into the mathematical details of this model, we mention its basic ingredients: a linear (1D) chain is considered with N diatomic molecules in which an excess electron is moving. With this model, the occurrence of two regimes, separated by a characteristic temperature typically of order (0.4 to 0.5)$\hbar\omega_{\mathrm{LO}}/k_{\mathrm{B}}$, is established theoretically: (a) hopping induced by phonons and (b) Bloch-type band motion. For the hopping motion of small polarons, Holstein derived the following expression for the mobility:

$$\mu = \frac{ea^2}{k_{\mathrm{B}}T}\frac{J^2}{\hbar^2\omega_{\mathrm{LO}}}\left[\frac{\pi}{\gamma\cosh(\hbar\omega_{\mathrm{LO}}/4k_{\mathrm{B}}T)}\right]^{1/2} \times \exp\left[-2\gamma\tan\left(\frac{\hbar\omega_{\mathrm{LO}}}{4k_{\mathrm{B}}T}\right)\right], \tag{21}$$

where J is a two-center overlap integral and γ is a measure for the electron–phonon coupling strength for small polarons, to be distinguished from the large-polaron coupling constant α ($2J$ corresponds to the width of the electronic Bloch band, which is supposed to be relatively small in small-polaron theory).

An important role in small-polaron theory belongs to the distinction between adiabatic and nonadiabatic hopping transitions [roughly speaking, the adiabatic regime is characterized by the fact that the electron follows the atomic (ionic) motion instantaneously]; see for the details the monograph by Klinger (1979).

The formation of small bipolarons by coupling of electrons to acoustic phonons and in disordered media was examined by Cohen *et al.* (1984). The study of small-polaron properties has been extended and studied in depth by Mott (1990), who identified and analyzed many instances of small-polaron transport including variable-range hopping, in which electrons hop over a range of distances and not only between nearest neighbors. He revealed the role of small polarons in amorphous semiconductors. The coherence and dynamics of small polarons in the presence of disorder were represented in terms of two characteristic energies: The polaron bandwidth specifies the energy scale of disorder at which the polarons become localized as composite particles, while the bare-electron bandwidth defines the energy scale at which the polaron ceases to be a composite particle (Spicci *et al.*, 1994).

In the theory of small polarons there are still some open fundamental questions that have been a subject of considerable recent investigation. Among the urgent issues, we note

1. the problem of the relevance of the Bloch-like states for a single small polaron in spite of the retardation, and
2. the study of the nature and properties of quasiparticles in many-polaron systems, which are of especial interest for the polaron, bipolaron, and hybrid polaron–bipolaron pictures of high-T_c superconductivity (see Sec. 5).

2.3 Experimental Evidence

Experimentally small-polaron effects have been analyzed, e.g., in UO_{2+x}, KCl, LiF, NiO, MnO, TiO_2, $BaTiO_3$, $SrTiO_3$, $LaCoO_3$, etc. We refer to the reviews by Appel (1968) and Firsov (1975) for more details. More recently, de Jongh (1988), Micnas *et al.* (1990), Alexandrov and Krebs (1992), and Alexandrov and Mott (1994) surveyed in detail both the principles and the main results of the small-polaron theory in the context of the (bi)polaronic approach in the physics of high-T_c superconductors and tried to interpret some experimental data in different materials in terms of small polarons and small bipolarons.

The study of the optical absorption for small polarons is complex. A representative example is shown in Fig. 7, where the real part of the ac conductivity, describing the small-polaron absorption, as derived from the Kubo formula, is shown for various Γ ($\Gamma = \hbar/4\tau_0 k_B T$, where τ_0 is $\frac{1}{4}$ times the "hopping time") (Reik and Heese, 1967). Note the completely different character of the optical absorption for small polarons as compared with large polarons (see Fig. 4).

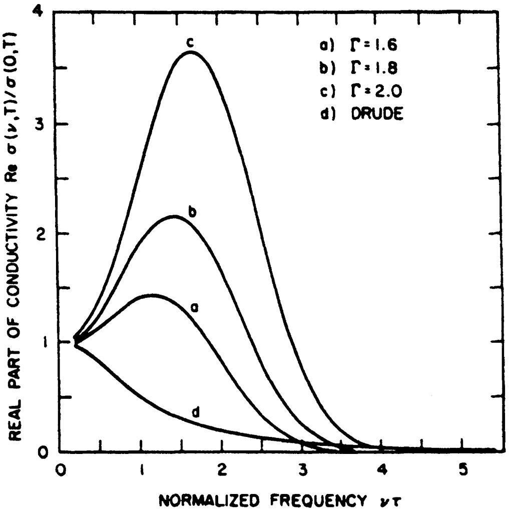

FIG. 7. Real part of the conductivity vs normalized frequency $\nu\tau$ for different values of the parameter Γ; from Reik and Heese (1967), reproduced by courtesy of Pergamon Press Ltd.

In analyzing experimental transport data also thermoelectric-power measurements are used; the theoretical study of the thermopower for small polarons has revealed that no polarization energy is transferred by the polaron motion.

Mobile polarons, observed in WO_{3-x} by Gehlig and Salje (1983), were shown to exhibit at 130 K a transition from a regime of hopping conductivity, characterized by a constant activation energy, to a regime of band conductivity, in which the process is not activated. With increasing carrier density, small polarons are formed up to a density that is equal to the concentration of the sites at which they can be localized. At this critical density, the dc electrical conductivity shows a phase transition, which these authors interpret as an Anderson-type transition: A change from a thermally activated small-polaron behavior to a metallic temperature dependence occurs. The critical density was found to be about 3.7×10^{21} cm^{-3} in WO_{3-x} (Salje and Güttler, 1984) and 1.7×10^{21} cm^{-3} in $NbO_{2.5-x}$ (Rüscher *et al.*, 1988). At higher densities, two type of carriers are suggested to coexist: small polarons, on the one hand, and, on the other hand, conducting carriers that can be regarded as

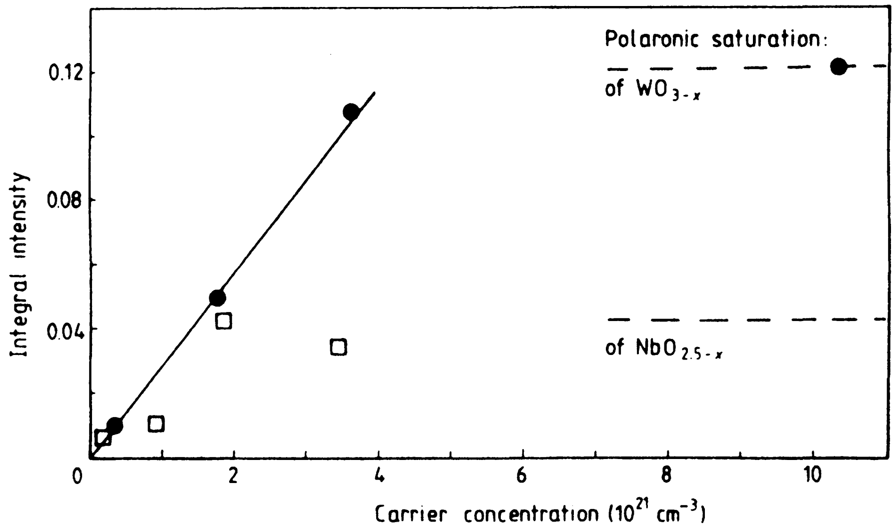

FIG. 8. Integral intensities of the polaronic absorption in dependence on the degree of reduction of $NbO_{2.5-x}$ (●) and WO_{3-x} (□); from Rüscher *et al.* (1988), reproduced by courtesy of Institute of Physics Publishing.

large polarons. The crossover from small-polaronic to metallic temperature dependence of the conductivity is consistently demonstrated by the Arrhenius plot for five different chemical compositions $NbO_{2.5-x}$. A thermally activated small-polaron conductivity seems to occur at lower degrees of reduction than that of $NbO_{2.49}$. This crossover scenario is strongly supported by the fact that close to the same critical densities as mentioned above, a saturation of the "integral intensity" of the polaronic absorption occurs; see Fig. 8.

Quite recently, the dc electrical conductivity of a slightly hyperstoichiometric sample of polycrystalline UO_{2+x} was interpreted (Casado *et al.,* 1994) in the framework of small-polaron theory, where some discrepancies between the semiempirical values of the small-polaron self-energy and the thermal activation energy, from the one side, and those obtained as a result of a fully microscopic calculation, from the other side, are revealed.

The recent measurements of the Seebeck coefficient of $BaBi_{0.25}Pb_{0.75}O_{3-\delta}$ with an oxygen deficiency by Hashimoto *et al.* (1995) suggest the coexistence between band charge carriers of high mobility, on the one side, and localized charge carriers of low mobility, on the other side (Fig. 9). Namely, in the low-temperature region, $T < 200$ K, the decrease of the (negative) Seebeck coefficient with T is supposed to be due to large electron polarons, while a temperature-activated behavior of the (positive) Seebeck coefficient above room temperature is attributed by the authors to small-hole polarons. A nonphenomenological interpretation of such experi-

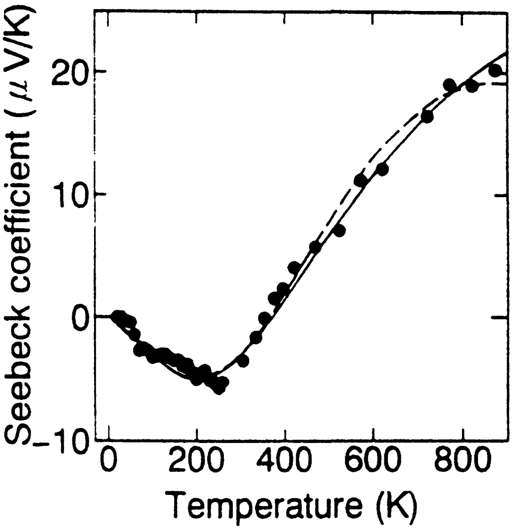

FIG. 9. Seebeck coefficient of $BaBi_{0.25}Pb_{0.75}O_{3.00}$: measured data (●) and phenomenological estimates (solid and dashed lines); from Hashimoto *et al.*, 1995, reproduced by courtesy of The American Physical Society.

ments requires a theoretical approach that would combine the large- and the small-polaron concepts.

A complex of experimental results on dielectric relaxation and ac and dc conductivities of $La_{1-x}Sr_xFeO_3$ with $0.05 \leq x \leq 0.3$ obtained by Jung and Iguchi (1995) was self-consistently explained in terms of small polarons (see, e.g., the Arrhenius plot for conductivity in Fig. 10).

3. BIPOLARONS AND POLARONIC EXCITONS

When two electrons (or two holes) interact with each other simultaneously through the Coulomb force and via the electron–phonon–electron interaction, either two independent polarons can occur or a bound state of two polarons—the *bipolaron* (Pekar, 1951)—can arise (Vinetskii, 1961; Hiramoto and Toyozawa, 1985; Adamowski, 1989). Whether bipolarons originate or not depends on the competition between the repulsive forces (direct Coulomb interaction) and the attractive forces (mediated through the electron–phonon interaction).

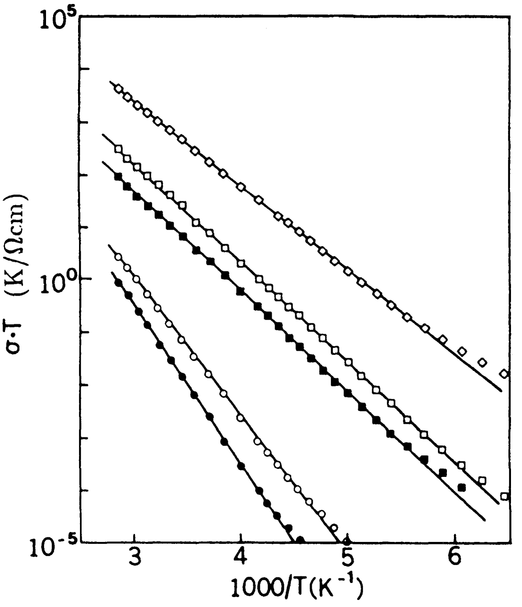

FIG. 10. Arrhenius plot for conductivity in $La_{1-x}Sr_xFeO_3$ with $x = 0.05$ (●), $x = 0.10$ (○), $x = 0.20$ (solid boxes), $x = 0.25$ (□), $x = 0.30$ (◇) in comparison with small-polaron theory (solid lines); from Jung and Iguchi (1995), reproduced by courtesy of Institute of Physics Publishing.

The bipolaron can be *free* and characterized by translational invariance, or it can be *localized.* According to Alexandrov and Ranninger (1981a,b), the many-electron system on a lattice coupled with any bosonic field turns out to be a charged Bose liquid, consisting of small bipolarons in the strong-coupling regime.

Similarly to the case of a bipolaron, an electron and a hole in a polarizable medium interacting with each other simultaneously both through the Coulomb force and via the electron–phonon–hole interaction form a quasiparticle—the *polaronic exciton.*

3.1 Fröhlich Bipolarons

In this section, the case of free bipolarons for electrons or holes interacting with longitudinal-optical phonons is discussed for the case of the continuum limit. They are referred to as Fröhlich bipolarons.

Fröhlich bipolarons are described by the following Hamiltonian:

$$\mathrm{H} = \sum_{j=1,2}\left[\frac{p_j^2}{2m_b} + \sum_{\mathbf{k}}\left(V_k a_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}_j} + V_k^* a_{\mathbf{k}}^\dagger e^{-i\mathbf{k}\cdot\mathbf{r}_j}\right)\right] + \sum_{\mathbf{k}} \hbar\omega_k a_{\mathbf{k}}^\dagger a_{\mathbf{k}} + U(\mathbf{r}_1 - \mathbf{r}_2), \tag{22}$$

where $\mathbf{p}_j$, $\mathbf{r}_j$ characterize the jth electron ($j = 1,2$), the potential energy for the Coulomb repulsion equals

$$U(\mathbf{r}) = e^2/\epsilon_\infty|\mathbf{r}| \equiv U/|\mathbf{r}|, \tag{23}$$

ϵ_∞ is the high-frequency dielectric constant, and the other symbols in Eq. (22) are the same as those in Eq. (1a). Note that one always has

$$U > \sqrt{2}\alpha \tag{24}$$

($\hbar = \omega_{LO} = m_b = 1$): This inequality expresses the obvious fact that $\epsilon_0 > \epsilon_\infty$.

In the discussion of bipolarons, often the ratio

$$\eta = \epsilon_\infty/\epsilon_0 \tag{25}$$

of the electronic and static dielectric constant is used ($0 \leq \eta \leq 1$). It turns out that bipolaron formation is favored by smaller η.

To estimate the order of magnitude of the quantities involved, one may express α and U as follows:

$$\alpha = \sqrt{\lambda}(1/\epsilon_\infty - 1/\epsilon_0), \tag{26a}$$

$$U = \sqrt{2\lambda}(1/\epsilon_\infty), \tag{26b}$$

with

$$\lambda = \frac{m_b}{m_e}\frac{R_y^*}{\hbar\omega_{\text{LO}}}. \tag{26c}$$

The effective Rydberg is characterized by the electron (hole) band mass m_b:

$$R_y^* = m_b e^4/2\hbar^2. \tag{27}$$

Devreese and associates (Verbist *et al.*, 1990, 1991) (VPD) analyzed the Fröhlich bipolaron using the Feynman path-integral formalism. Quite analogously to the above-discussed relations (12a) to (12d), a scaling relation was derived between the free energies F in two dimensions, $F_{2\text{D}}(\alpha,U,\beta)$, and in three dimensions, $F_{3\text{D}}(\alpha,U,\beta)$:

$$F_{2\text{D}}(\alpha,U,\beta) = \tfrac{2}{3}F_{3\text{D}}\left(\frac{3\pi}{4}\,\alpha,\ \frac{3\pi}{4}\,U,\beta\right). \tag{28}$$

This is the generalization to bipolarons of the scaling relation for a single Fröhlich polaron (Peeters and Devreese, 1987). Physically the scaling relation implies that bipolaron formation will be facilitated in 2D as compared with 3D. (The critical value for bipolaron formation α_c will be scaled with a factor $3\pi/4 \approx 2.36$; or $\alpha_c^{(2\text{D})} = \alpha_c^{(3\text{D})}/2.36$).

Smondyrev *et al.* (1995) derived an analytical strong-coupling asymptotic expansion in inverse powers of the electron–phonon coupling constant for the large bipolaron energy at $T = 0$:

$$E_{3\text{D}}(\alpha,u) = -(2\alpha^2/3\pi)A(u) - B(u) + O(\alpha^{-2}), \tag{29a}$$

where the coefficients are closed analytical functions of the ratio $u = U/\alpha$:

$$A(u) = 4 - 2\sqrt{2}u\left(1 + \frac{u^2}{128}\right)^{3/2} + \tfrac{5}{8}u^2 - \frac{u^4}{512}, \tag{29b}$$

and for $B(u)$ see the above-cited paper.

The scaling relation (28) allows one to find the bipolaron energy in two dimensions as

$$E_{2\text{D}}(\alpha,u) = \tfrac{2}{3}E_{3\text{D}}\left(\frac{3\pi}{4}\,\alpha,u\right). \tag{30}$$

A "phase diagram" for the polaron–bipolaron system was introduced by VPD. It is based on the generalized trial action. This phase diagram is shown in Fig. 11 for the 3D case. A Fröhlich coupling constant as high as 6.8 is needed to allow for bipolaron formation. No definite experimental evidence has been provided for the existence of materials with such high Fröhlich coupling constant. (One of the highest α's reported is for RbCl where $\alpha \approx 3.8$ and for CsI where $\alpha \approx 3.7$; see Table 1.)

Materials with sufficiently large α for Fröhlich bipolaron formation in 3D might exist, but careful analysis (involving, e.g., the study of cyclotron resonance), like the one executed for AgBr and AgCl (Hodby *et al.*, 1987), is in order to confirm this. Presumably some modifications to the Fröhlich Hamiltonian are also necessary to describe such high coupling because of the more localized character of the carriers in this case, which makes the continuum approximation less valid.

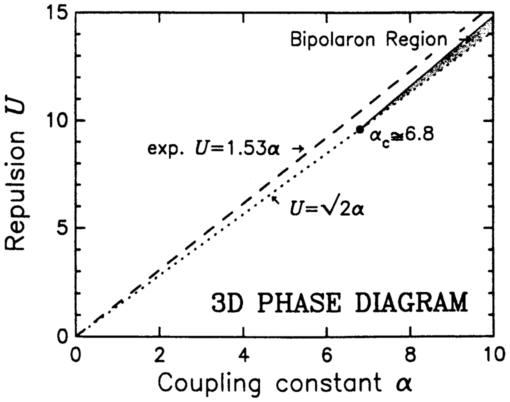

FIG. 11. The stability region for bipolaron formation in 3D from Verbist *et al.* (1990). The dotted line $U = \sqrt{2}\alpha$ separates the physical region ($U \geq \sqrt{2}\alpha$) from the nonphysical ($U \leqslant \sqrt{2}\alpha$). The stability region lies below the full curve. The shaded area is the stability region in physical space. The dashed line is determined by $U = \sqrt{2}\alpha/(1 - \epsilon_\infty/\epsilon_0)$ where we took the experimental values $\epsilon_\infty = 4$ and $\epsilon_0 = 50$. The critical point $\alpha_c = 6.8$ is represented as a full dot.

The confinement of the bipolaron in 2D facilitates bipolaron formation at smaller α. From Fig. 12 it is seen that bipolarons can now be stable for $\alpha \geq 2.9$, a domain of coupling constants that is definitely realized in several solids. Intuitive arguments suggesting that bipolarons are stabilized in going from 3D to 2D had been given before, but the quantitative analysis based on the path integral was presented by VPD.

The stability of bipolarons has also been examined with the use of operator techniques where the center-of-gravity motion of the bipolaron was approximately separated from the relative electron (hole) motion (see Bassani *et al.*, 1991). The results by Bassani *et al.* (1991) and by VPD tend to confirm each other.

The bipolaron was also approached (Hiramoto and Toyozawa, 1985) in the path-integral representation using a special case of the trial action of VPD; in this work the combined effect of LO phonons, acoustic phonons, and deformation potential was analyzed.

Early work on bipolarons had been based on strong-coupling theory, in which case the bipolaron stability can be expressed with η as the sole parameter. ($\eta < 0.08$ is a typical strong-coupling result for bipolaron stability.) It turns out that the numerical stability criteria of VPD can be adequately formulated analytically for all α, for which the bipolaron is stable.

A very clear representation of experimental evidence for bipolarons, e.g., from the data on magnetization and electric conductivity in Ti_4O_7, as well as in $Na_{0.3}V_2O_5$ and polyacetylene, given by Mott (1990) is to be mentioned.

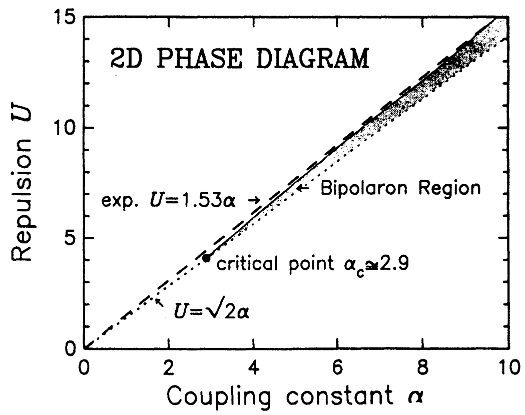

FIG. 12. The same as Fig. 11, but now for 2D, where the critical point is $\alpha_c = 2.9$; from Verbist *et al.* (1990).

Recently, the effect of a high magnetic field on bipolarons was examined both for singlet and triplet states (Fomin and Devreese, 1995).

3.2 Polaronic Excitons

Excitons (*q.v.*) constitute very interesting physical entities, which in polarizable media are relevant to polarons: In polar systems they can be conceived as two interacting polarons of opposite charges (Haken, 1956; Toyozawa, 1963, 1964; Knox, 1963; Bassani and Baldereschi, 1973; Bassani and Pastori Parravicini, 1975; Adamowski *et al.*, 1981; Wallis and Balkanski, 1986).

The exciton ground-state energy in a polar crystal was determined by Pollmann and Büttner (1975, 1977) and by Aldrich and Bajaj (1977) taking into account the fact that the potential energy of the electron–hole interaction depends on the quantum state of the interacting particles as a result of the polaronic effect. Later on, many works were devoted to this problem. Similar considerations were applied by Petelenz and Smith (1981) to explain the dependence of the binding energy of an exciton-ionized donor complex on the electron-to-hole mass ratio in CdS and TlCl, and by Larsen (1981) to show that the ratio of the binding energy of the D^- centers to that of neutral donors in AgBr and AgCl is as much as one order of magnitude larger than in a nonpolar crystal. The binding enhancement is due to the *attraction* between the electrons and the static polarization charge induced by them in the central part of the ion. Experimental data on the spectral photoconductivity in the systems Ca–Sr–Bi–Cu–O (Masumi *et al.*, 1988a) and Ba–Pb–Bi–O (Masumi *et al.*, 1988b) have been interpreted as due to an exciton-mediated bipolaronic mechanism. Recent experimental data on the reflectivity and its temperature dependence in La_2CuO_4 were interpreted in terms of polaronic excitons by Falck *et al.* (1992).

3.3 Localized Bipolarons

Localized bipolarons tend to be *small* bipolarons, characterized by a radius of the order of the lattice constant. For a small bipolaron, both constituting polarons can be localized either at the same lattice site (intrasite, or Anderson, bipolaron; see Anderson, 1975) or at two different lattice sites, e.g., at two neighboring lattice sites (intersite bipolaron).

For two electrons (holes) on the same site, the direct Coulomb repulsion is governed by the potential energy U. Whether or not two electrons remain at the same site is determined by U_{eff}, the effective potential that arises if both the Coulomb repulsion and the electron (hole)–phonon or polaron interaction are taken into account. If the polaron interaction dominates ($U_{\text{eff}} < 0$) the Coulomb repulsion, one speaks of *negative-U behavior* and two carriers can be occupying the same site ("double occupancy of a lattice site"); a bipolaron localized on one site arises. Negative-U bipolarons have been suggested to occur, e.g., in *chalcogenide glasses* (Anderson, 1975; Mott, 1990).

The intersite bipolaron can form *singlet* (bonding as well as antibonding) or *triplet* states (Fig. 13). Intersite singlet bipolarons are usually referred to as Heitler–London bipolarons (see, for example, de Jongh, 1988).

Localization of bipolarons was investigated by many authors including Lannoo *et al.* (1959), Hubbard (1964), Stoneham and Bullough (1971), Anderson (1975), and Mott (1990), to name a few. An extensive literature exists, and a more complete discussion is given by Fisher *et al.* (1989). Localized bipolarons have been studied in numerous oxides including Ti oxides, vanadium bronzes, $LiNbO_3$, and WO_3.

Hybridization between coexisting localized bipolarons and itinerant electrons leads to a possibility of the coherent motion of bipolarons and their phase transition into a superfluid state upon lowering the temperature to a certain critical temperature (Ranninger, 1994), which seems to be relevant to the phenomena occurring in cuprates, bismuthates, and fullerenes.

4. SPIN POLARONS

Just as electrons or holes interact with the Bose field of phonons, they can, because of the exchange interaction, interact with the magnon field. This gives rise to the formation of the *spin polaron* (or magnetic polaron). One of the first studies related to spin polarons is due to de Gennes (1960).

Intuitively one can imagine that, e.g., in an antiferromagnet, the carrier spin polarizes the spins of the surrounding magnetic ions. The spin polaron consists of the carrier with its spin together with the lattice magnetization created by the carrier spin. Similar to the case of the dielectric polaron, the physical properties of the carrier are influenced by the self-induced magnetization cloud. The mechanism for the formation of spin polarons can be easily understood for

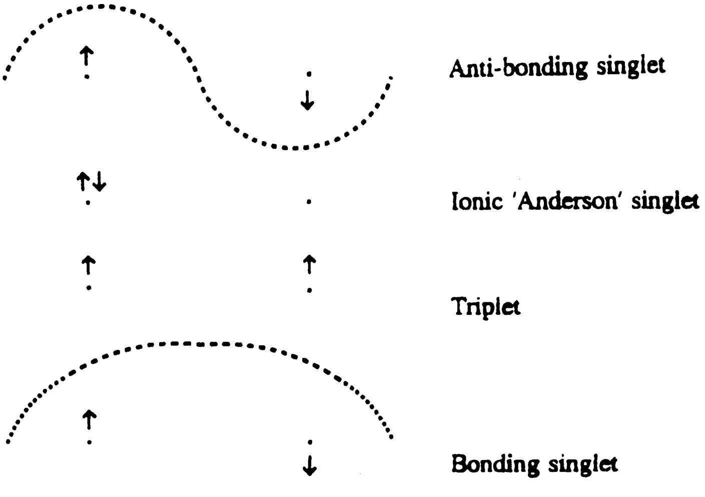

FIG. 13. States of a model two-site, two-electron system where the electrons are coupled to the interatomic coordinate; from Fisher *et al.*, (1989), reproduced by courtesy of Institute of Physics Publishing.

materials with preference for antiferromagnetic alignment; the kinetic energy of a carrier can be reduced if reversing the spins of the surrounding ions allows the carrier to move more easily around these ions (Wood, 1991).

In principle, spin polarons can arise in ferromagnets as well. Also the induced magnetic polarization can be characterized by ferromagnetic as well as antiferromagnetic alignment.

In general, it is accepted that the evidence for the occurrence of *bound magnetic polarons* is more convincing than that for *free magnetic polarons* (Benoit á la Guillaume, 1993). Bound magnetic polarons occur when the carriers are localized (e.g., because they are bound to a donor or to an acceptor).

The bound magnetic polaron has been studied primarily in Mn-based semimagnetic semiconductors (SMSC). The bound magnetic polaron then originates from the exchange interaction between a rather localized unpaired carrier and the $\frac{5}{2}$ spin of the surrounding Mn^{2+} ions. In the bound magnetic polaron, ferromagnetic order occurs within the orbit of the bound carrier and the magnetic exchange leads to an increase of its binding energy. For spin polarons, e.g., in the case of EuS doped with GdSe (Mott, 1990), the basic interaction energy between a conduction-electron spin and the surrounding magnetic moments is given by

$$J_{sf} sS |\Psi_s(0)|^2 V; \tag{31}$$

here J_{sf} is the energy of the ferromagnetic coupling between the spin s of the conduction electron and the europium ion spin S, $\Psi_s(0)$ is the wave function of the conduction electron at the origin, and V is the atomic volume.

A transparent intuitive consideration due to Mott (1990) leads to a simple expression for the spin-polaron radius:

$$R = \left(\frac{\hbar^2 \pi a^3}{4m^* J_N}\right)^{1/5}, \tag{32}$$

where J_N is the energy, per moment, needed for a transition from the antiferromagnetic alignment to the ferromagnetic alignment, m^* is the effective mass of the carrier, and a the lattice parameter of the host solid.

For the effective mass of the spin polaron, the following expression has been derived (Mott, 1990):

$$m^*_{sp} = m^* e^{\gamma R/a}, \tag{33}$$

with $\gamma \approx 1$.

5. BIPOLARONS AND HIGH-TEMPERATURE SUPERCONDUCTIVITY

Dielectric or Fröhlich bipolarons as well as spin bipolarons have been considered as possibly playing a role in superconductivity. Also the "small-polaron" interaction was studied in the framework of superconductivity.

In fact, already the charged bosons of Schafroth (1955) could be thought of as Fröhlich bipolarons. However, although the London equation indeed results from a bipolaron-gas model, a continuous behavior of the specific heat C_V was found at the superconducting transition temperature in contradistinction from experiment. After the success of the BCS theory with its Cooper pairs consisting of electrons coupling in **k** space, $|\uparrow, \mathbf{k}; \downarrow, -\mathbf{k}\rangle$, the bipolaron gas and the pairing in real space became a less central issue.

Nevertheless, theoretical models incorporating small bipolarons [localized Cooper pairs with electrons (holes) of opposite spins on nearest-neighbor sites] and small bipolarons in narrow bands and characterized by hopping conductivity were still studied (Alexandrov and Ranninger, 1981a,b; Chakraverty *et al.*, 1987). These models had not predicted high-T_c superconductivity, however.

After the discovery of high-T_c superconductivity, alternative models (with respect to BCS) for superconductivity received renewed attention. It is experimentally clear that also in the high-T_c materials, pairing of carriers in the superconducting phase takes place. However, different coupling mechanisms might be at the basis of the pairing; e.g., coupling through acoustic plasmons has been considered.

In the present context, a preliminary question is whether Fröhlich or spin bipolarons do form in high-T_c materials. This question has been investigated by Verbist *et al.* (1991), and the polaron–bipolaron phase diagram was applied, e.g., to La_2CuO_4 (see Fig.

12). In this figure the existence line for bipolarons in $YBa_2Cu_3O_6$ corresponds to $U \approx 1.49\alpha$. On the basis of the existing experimental data, it can be stated that the characteristic point in the polaron–bipolaron phase diagram for La_2CuO_4 and $YBa_2Cu_3O_6$ lies on the straight line through the origin and is shown in the figure. Presumably the Fröhlich coupling is large enough in these materials that their α overlaps with the bipolaron existence region with reference to α. It is remarkable that the existence line for bipolarons lies very close to (La_2CuO_4) or penetrates ($YBa_2Cu_3O_6$) the very narrow existence domain for large bipolarons. It seems therefore *safe to assume* that Fröhlich bipolarons do occur in some of the high-T_c materials in the normal phase; see Verbist *et al.* (1990) and Devreese *et al.* (1994).

It should be reiterated that the precise α value characterizing the high-T_c materials is not known as long as the band mass has not been determined. It may also be observed that most other "polaron materials" (alkali halides, etc.) find their characteristic points in the polaron–bipolaron phase diagram far away from the bipolaron-stability area.

Cataudella *et al.* (1992) and Iadonisi *et al.* (1995) have undertaken the study of the many-body physics of the bipolaron gas in an approximation comparable to that by Verbist *et al.* (1990), but based on a Hamiltonian formulation. Although the mathematics of this problem is complex and the studies are far from finalized, some promising features arise; especially the variation of T_c for the high-T_c materials as a function of the doping is well reproduced, be it with one fitting parameter.

On the basis of results by Verbist *et al.* (1991), also the optical spectra of bipolarons were analyzed (Devreese *et al.*, 1994); it was suggested that the mid-IR peak characterizing the high-T_c solids is related to relaxed excited states of the bipolaron.

Emin and Hillery (1989) studied the interplay between short- and long-range polaron interaction in the adiabatic (strong-coupling) approximation. This work joins the ideas of Schafroth and examines in more detail the character of the mobile charged bosons as bipolarons. (It should be noted that—in fact—the adiabatic theory does not lead to the existence of strong-coupling large Fröhlich bipolarons.)

In discussing bipolaronic superconductivity as a candidate theory for high-T_c superconductivity, one should mention the analysis by Mott in terms of *spin bipolarons.* Mott (1991) has emphasized how a metal–insulator transition occurs in several of the high-temperature superconductors, the metal being superconducting. In analyzing this transition, he suggests the formation of spin polarons and the possibility that they combine to form bipolarons. Those bipolarons could replace the Cooper pairs. Alexandrov *et al.* (1994a) have also analyzed the nondegenerate gas of bosons above T_c. In particular, the linear T behavior of the resistivity above T_c was accounted for. Mott envisaged the possibility of two different mechanisms for superconductivity depending on the hole concentration (for one of the mechanisms, the bosons do not overlap; for the other they do) and stressed the analogy of high-T_c superconductors to the behavior of superfluid ^{4}He. The quantitative formulation of this fact goes back to Alexandrov and Ranninger (1992), who used a mapping of the electronic specific heat of high-T_c superconductors onto that of ^{4}He to show that its λ-like temperature dependence may be described in terms of charged bosons.

A spin-bipolaron model has been invoked by Alexandrov *et al.* (1994b) to explain the experimental data on the Hall effect and resistivity in underdoped $YBa_2Cu_3O_{7-\delta}$ (Carrington *et al.*, 1993), based on the assumption that at zero temperature a part of the spin bipolarons are in Anderson localized states due to disorder. The agreement between theory and experiment exists at least in the region of temperatures higher than the temperature T^* at which the slope of resistivity changes. This is illustrated in Figs. 14 and 15, representing the temperature dependence of the Hall coefficient and of the resistivity for different values of the degree of reduction in $YBa_2Cu_3O_{7-\delta}$.

Also, the Cooper pairing of small polarons has been regarded as a possible mechanism of high-T_c superconductivity; see, e.g., a comprehensive review by Alexandrov and Mott (1994) and an extensive list of references therein. It was shown that there exists an intermediate region of the coupling constant, where the energy of interaction between small polarons (resulting from the interplay between the Coulomb repulsion and the at-

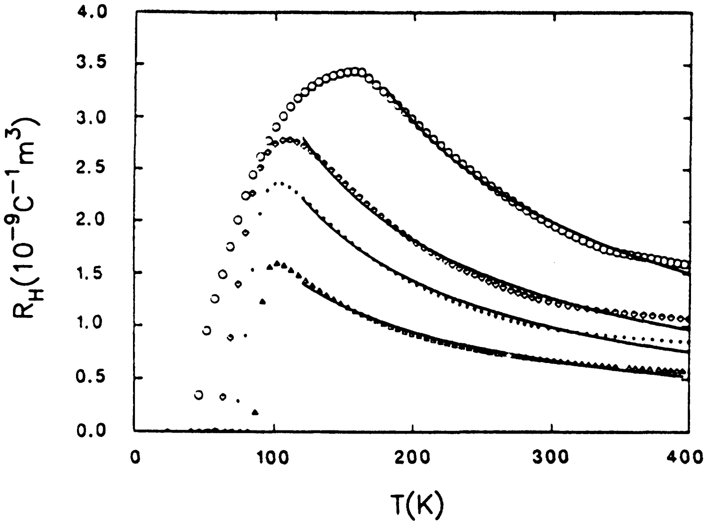

FIG. 14. Hall coefficient of YBa$_2$Cu$_3$O$_{7-\delta}$ for various degrees of reduction: $\delta = 0.05$ (△), $\delta = 0.19$ (●), $\delta = 0.23$ (◇), $\delta = 0.39$ (○) (Carrington *et al.*, 1993) in comparison with theory; from Alexandrov *et al.*, (1994b), reproduced by courtesy of Elsevier Science Publishers.

traction due to their interaction with the lattice) is smaller than, or of the same order as, the small-polaron bandwidth, both being less than or comparable to the phonon frequency. In this region, the BCS theory predicts superconductivity even for on-site repulsion, provided that it is weaker than the total intersite attraction of small polarons. It is the *polaronic narrowing* of the band that increases the estimated maximum value for the critical temperature when the interaction with high-frequency phonons is dominant. This maximum occurs for an intermediate value of the coupling constant which is situated at the boundary between the above-described polaronic superconductivity region, on the one hand, and the bipolaronic superconductivity region, on the other hand. The latter mechanism is considered to be relevant for values of the coupling constant high enough to make the energy of attraction between small polarons much larger than the small-polaron bandwidth. In this case, the ground state of the system of small polarons can be—in first approximation—regarded as consisting of a phonon field and a set of immobile small *bipolarons,* the hopping term treated as a perturbation producing the bipolaronic motion.

A serious alternative to the picture devel-

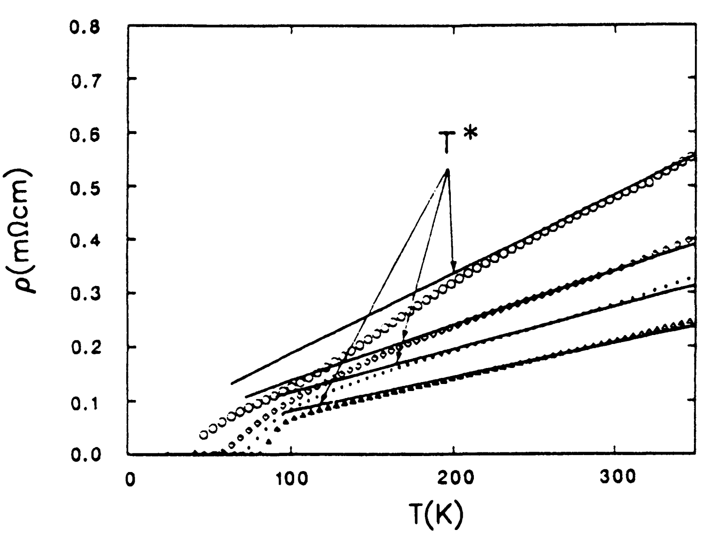

FIG. 15. Resistivity of YBa$_2$Cu$_3$O$_{7-\delta}$ for various degrees of reduction in the same denotations as in Fig. 14 (Carrington *et al.*, 1993) in comparison with theory. T^* stands for the temperature at which the change in the slope of resistivity occurs; from Alexandrov *et al.* (1994b), reproduced by courtesy of Elsevier Science Publishers.

oped by Alexandrov and Mott (1994) is the fermion–boson model considering a possibility of exchange between localized bipolarons and itinerant electrons, which goes back to Ranninger and Robaszkiewicz (1985). Depending on the filling, this model gives either a BCS-like state or a superfluid state of bipolarons with a dramatic increase of T_c beyond a critical concentration.

In summary of this section, it is realized that no definite theory for high-T_c superconductivity (whether or not it involves bipolarons) is available as yet; nevertheless dielectric bipolarons, spin-bipolaron gases, and polaron–bipolaron mixtures seem to be systems of significant promise in attempting to construct such a theory.

6. FURTHER DEVELOPMENTS OF THE POLARON CONCEPT

6.1 Polarons and Bipolarons in Polymers. Solitons

Many conjugated polymers (e.g., *trans*-polyacetylene) behave as quasi-1D semiconductors (where the energy gaps are between 1.5 and 3.0 eV). The carrier transport in those systems seems to take place through *charged defects* originating as a result of doping. These charged defects are influenced by the polaronic interaction with the surroundings. At low temperature, the transport is dominated by the motion of the defects; at higher doping levels hopping conductivity occurs. As is well known, the possible application of those polymers (including, e.g., those in aerospace technology) has attracted wide attention.

Su *et al.* (1980) have proposed a model Hamiltonian to study the physics of conjugated polymers. In analyzing the polaronic localized states related to the charged defects, nonlinear behavior, characteristic of solitons, is revealed. Polaron-type excitations were experimentally observed in polyacetylene by Su and Schrieffer (1980).

Polarons in conjugated polymers constitute a challenging and vast subject; in the context of this article only its significance and main ingredients are pointed out; the reader is referred to comprehensive reviews by Yu (1988) and Fisher *et al.* (1989). Bipolarons might occur in polymers at higher dopant levels, and the question of polaron–bipolaron conversion rates is a central one.

6.2 Modeling Systems Using the Polaron Concept

Originally the polaron was mainly studied to describe the electron–LO-phonon interaction. The concept has been generalized to several systems where one or many fermions interact with a bath of bosons or to instances where such an approximation is meaningful. Table 4 lists several such cases. The fundamental concepts concerning a piezopolaron—this is a name for a quasiparticle arising from the piezoelectric interaction between an electron and acoustic phonons—were displayed by Mahan (1972) and Lax (1972). More recently, convincing experimental evidence on the significant role of piezopolarons in the electronic transport properties of CdS has been presented by Mahan (1990).

Other examples include the motion of a ^{3}He atom through superfluid helium considered by Bardeen *et al.* (1969) and by Brosens *et al.* (1991), or the dynamics of electrons (2DEG) at the surface of liquid He—"ripplonic polaron" (Jackson and Platzman, 1981; Peeters, 1987).

Charges moving through liquids (many of which are characterized by polarizable atoms) have been modeled as "hydrated polarons" (Laria *et al.*, 1991). Even an electron in-

Table 4. Extensions of the polaron concept.

Concept	"Candidates"
Acoustic polaron	Hole in AgCl
Piezoelectric polaron	ZnO, ZnS, CdS
Electronic polaron	Band structure of all semiconductors
Spin polaron	Magnetic semiconductors
Bipolarons:	
One center	Amorphous chalcogenides
Two center	Ti_4O_7, vanadium-bronzes
Small polarons and hopping	Transition-metal oxides, polyacetylenes
Superconductivity and (bi)-polarons ("Localized electron pairs")	$BaPb_{1-x}$ (1981) Bi_xO_3 $LiTi_2O_4$ (1985) "High-T_c superconductivity"

teracting with plasmons ("plasmaron") was invoked to study the electron gas. A new subject concerns the occurrence of polarons in fullerenes (Matus *et al.*, 1992).

These extensions of the polaron concept are not treated in any detail here, but they illustrate the richness of the polaron concept.

A valuable extension of the polaron concept arises by considering the interaction between a carrier and the exciton field. One of the early formulations of this model was by Toyozawa (1963). The resulting quasiparticle is called the *electronic polaron.* The self-energy of the electronic polaron (which is almost independent of wave number) must be taken into account when the band gap of an insulator or semiconductor is calculated using pseudopotentials. For example if one calculates, with Hartree–Fock theory, the band gap of an alkali halide, one is typically off by a factor of 2. This was the original problem that was solved conceptually with the introduction of the electronic polaron (Toyozawa, 1963; Kunz, 1974). Also in the soft x-ray spectra of alkali halides exciton sidebands have been observed that seem to be due to the electronic polaron coupling (Devreese *et al.*, 1972). The standard "local density approximation" (LDA) does not provide the correct band gap for semiconductors. The solution to this problem (Hybertsen and Louie, 1987; Godby *et al.*, 1988) via the so-called GW approximation of Hedin (1965) is analogous in its basic interaction to Toyozawa's solution of the gap problem for strongly ionic crystals.

ACKNOWLEDGMENTS

The author wishes to express his gratitude to several colleagues for many stimulating discussions and contributions on polarons over the years: R. Evrard, F. Peeters, F. Brosens, L. Lemmens. He also wishes to acknowledge recent interactions and fruitful discussions in particular with V. Fomin, and also with M. Smondyrev.

This work has been partially supported by the E.E.C. Human Capital and Mobility Program under Contract No. CHRX-CT93-0124.

GLOSSARY

Anderson Bipolaron: See **Intrasite Bipolaron.**

Bipolaron: A composite quasiparticle consisting of a bound state of two conduction electrons (or holes) and the surrounding common cloud of virtual phonons.

Electronic Polaron: A quasiparticle that arises as a result of the interaction between a charge carrier and the exciton field.

Ferron: See **Magnetic Polaron.**

Free Bipolaron: A **bipolaron** characterized by translational invariance, which can move as a free particle of enhanced mass (bipolaron mass).

Free Polaron: A **polaron** characterized by translational invariance, which can move as a free particle of enhanced mass (polaron mass).

Fröhlich Bipolaron: A **large bipolaron** formed through the electron–LO-phonon interaction in a polarizable continuum.

Fröhlich Polaron: A **large (radius) polaron** formed through the electron–LO-phonon interaction in a polarizable continuum.

Heitler–London Bipolaron: A singlet **intersite bipolaron.**

"Hydrated Polaron": A quasiparticle modeling charges that move through liquids (many of which are characterized by polarizable atoms).

Intersite Bipolaron: A **bipolaron** in which the two constituent small polarons are localized at two different lattice sites.

Intrasite Bipolaron: A **small bipolaron** in which the two constituent polarons are localized at the same lattice site.

Large Bipolaron: A **bipolaron** with a **Radius** much larger than the lattice constant.

Large (Radius) Polaron: A **polaron** with a radius considerably larger than the lattice constant of the solid.

Landau Polaron: See **Fröhlich Polaron.**

Magnetic Polaron: A quasiparticle consisting of an electron together with a group of magnetic moments oriented antiparallel to the electron spin as a result of the exchange interaction of the electron with the magnon field.

Pekar Polaron: See **Fröhlich Polaron.**

Piezopolaron: A quasiparticle arising from the piezoelectric interaction between an electron and acoustic phonons.

"Plasmaron": A quasiparticle sometimes considered as consisting of an electron (or hole) interacting with plasmons in the electron gas.

Polaron: A quasiparticle consisting of a conduction electron (or hole) together with

- its self-induced polarization in a polar semiconductor or an ionic crystal;
- the self-induced lattice deformation field in a nonionic material;
- the self-induced perturbation of a Bose field (in the modern extensions of the polaron concept).

Polaronic Exciton: A composite quasiparticle, which is formed by two charge carriers of opposite sign and a surrounding cloud of virtual phonons.

Radius of a bipolaron: A characteristic size of the common cloud of virtual phonons, surrounding two electrons (or holes) in a **bipolaron.**

Radius of a polaron: A characteristic size of the fluctuations in space of an electron in a **polaron** due to emission and absorption of virtual phonons.

"Ripplonic Polaron": A quasiparticle describing the dynamics of electrons at the surface of liquid ^{4}He.

Small Bipolaron: A **bipolaron** with a **radius** of the order of the lattice constant. A small bipolaron can move by hopping from one localization position to another.

Small (Radius) Polaron: A **polaron** with a radius of the order of the lattice constant. A small polaron can move by hopping from one lattice site to another.

Spin Polaron: See **Magnetic Polaron.**

Works Cited

Adamowski, J. (1985), *Phys. Rev. B* **32,** 2588–2595.

Adamowski, J. (1989), *Phys. Rev. B* **39,** 3649–3652.

Adamowski, J., Gerlach, B., Leschke, H. (1981), *Phys. Rev. B* **23,** 2943–2950.

Aldrich, C., Bajaj, K. K. (1977), *Solid State Commun.* **22,** 157–160.

Alexandrov, A. S., Krebs, A. B. (1992), *Usp. Fiz. Nauk* **162,** 1–85 [English translation: *Sov. Phys. Usp.* **35,** 345–383].

Alexandrov, A. S., Mott, N. F. (1994), *Rep. Prog. Phys.* **57,** 1197–1288.

Alexandrov, A., Ranninger, J. (1981a), *Phys. Rev. B* **23,** 1796–1801.

Alexandrov, A., Ranninger, J. (1981b), *Phys. Rev. B* **24,** 1164–1169.

Alexandrov, A., Ranninger, J. (1982), *Solid State Commun.* **81,** 403–406.

Alexandrov, A. S., Bratkovsky, A. M., Mott, N. F. (1994a), *Phys. Rev. Lett.* **72,** 1734–1737.

Alexandrov, A. S., Bratkovsky, A. M., Mott, N. F. (1994b), *Physica C* **235–240,** 2345–2346.

Anderson, P. W. (1975), *Phys. Rev. Lett.* **34,** 953–955.

Appel, J. (1968), in: F. Seitz, D. Turnbull, H. Ehrenreich (Eds.), *Solid State Physics,* Vol. 21, pp. 193–391.

Austin, I. G., Mott, N. F. (1969), *Adv. Phys.* **18,** 41–102.

Bajaj, K. K. (1972), in: J. T. Devreese (Ed.), *Polarons in Ionic Crystals and Polar Semiconductors,* Amsterdam: North-Holland, pp. 193–225.

Bardeen, J., Baym, G., Pines, D. (1969), *Phys. Rev.* **156,** 207–221.

Bassani, F., Baldereschi, A. (1973), *Surf. Sci.* **37,** 304–327.

Bassani, F., Pastori Parravicini, G. (1975), *Electronic States and Optical Transitions in Solids,* Oxford: Pergamon.

Bassani, F., Geddo, M., Iadonisi, G., Ninno, D. (1991), *Phys. Rev. B* **43,** 5296–5306.

Benoit á la Guillaume, C. (1993), *Phys. Status Solidi B* **175,** 369–380.

Bogolubov, N. N. (1950), *Ukr. Matem. Zh.* **2,** 3–24.

Bogolubov, N. N., Tyablikov, S. V. (1949), *Zh. Eksp. Teor. Fiz.* **19,** 256–268.

Brandt, R. C., Brown, F. C. (1969), *Phys. Rev.* **181,** 1241–1250.

Brosens, F., Devreese, J. T. (1988), *Phys. Status Solidi B* **145,** 517–523.

Brosens, F., Lemmens, L., Devreese, J. T. (1991), *Phys. Rev. B* **44,** 10296–10299.

Brown, F. C. (1963), in: C. G. Kuper, G. D. Whitfield (Eds.) (1963), *Polarons and Excitons,* Edinburgh: Oliver and Boyd, pp. 323–355.

Brown, F. C. (1981), in: J. T. Devreese (Ed.) (1981), *Recent Developments in Condensed Matter Physics,* Vol. 1, New York: Plenum, pp. 575–591.

Carrington, A., Walker, D. J. C., Mackenzie, A. P., Cooper, J. R. (1993); *Phys. Rev. B* **48,** 13051–13059.

Casado, J. M., Harding, J. H., Hyland, G. J. (1994), *J. Phys.: Cond. Matter* **6,** 4685–4698.

Cataudella, V., Iadonisi, G., Ninno, D. (1992), *Europhys. Lett.* **17,** 709–714.

Chakraverty, B. K., Feinberg, D., Hang, Z., Avignon, M. (1987), *Solid State Commun.* **64,** 1147–1151.

Cheng, J.-P., McCombe, B. D., Shi, J. M., Peeters, F. M., Devreese, J. T. (1993), *Phys. Rev. B* **48,** 7910–7914.

Cohen, M. H., Economou, E. N., Soukoulis, C. M. (1984), *Phys. Rev. B* **29,** 4496–4499, 4500–4504.

Cohn, D. R., Larsen, D. M., Lax, B. (1972), *Phys. Rev. B* **6,** 1367–1375.

Das Sarma, S. (1984), *Phys. Rev. Lett.* **52,** 859–862.

Das Sarma, S., Mason, B. A. (1985), *Ann. Phys. (USA)* **163,** 78–119.

de Gennes, P. G. (1960), *Phys. Rev.* **118,** 141–154.

de Jongh, L. J. (1988), *Physica C* **152,** 171–216.

Devreese, J. T., (1963), *Bull. Belgian Phys. Soc.* **III,** 259–263.

Devreese, J. T. (Ed.) (1972), *Polarons in Ionic Crystals and Polar Semiconductors,* Amsterdam: North-Holland.

Devreese, J. T., Brosens, F. (1991), *Solid State Commun.* **77,** 583–584.

Devreese, J. T., Brosens, F. (1992), *Phys. Rev. B* **45,** 6459–6478.

Devreese, J. T., Evrard, R. (1966), *Phys. Lett.* **11,** 278–279.

Devreese, J. T., Peeters, F. (Eds.) (1984), *Polarons and Excitons in Polar Semiconductors and Ionic Crystals,* New York: Plenum.

Devreese, J. T., Peeters, F. M. (1986), *Solid State Commun.* **58,** 861–863.

Devreese, J. T., Peeters, F. M. (Eds.) (1987), *The Physics of the Two-Dimensional Electron Gas,* New York: Plenum.

Devreese, J. T., Kunz, A. B., Collins, T. C. (1972), *Solid State Commun.* **11,** 673–678.

Devreese, J., Evrard, R., Kartheuser, E., Brosens, F. (1982), *Solid State Commun.* **44,** 1435–1438.

Devreese, J. T., Verbist, G., Peeters, F. M. (1994), in: *Proceedings of Workshop on Polarons and Bipolarons in High-T_c Superconductors and Related Materials,* Cambridge, 1994 (in press).

Emin, D., Hillery, M. S. (1989), *Phys. Rev. B* **39,** 6575–6593.

Emin, D., Holstein, T., (1976), *Phys. Rev. Lett.* **36,** 323–326.

Engineer, M. H., Tzoar, N. (1972), in: J. T. Devreese (Ed.), *Polarons in Ionic Crystals and Polar Semiconductors,* Amsterdam: North-Holland, pp. 747–754.

Evrard, R. (1965), *Phys. Lett.* **14,** 295–296.

Evrard, R., Kartheuser, E., Devreese, J. T. (1970), *Phys. Status Solidi B* **41,** 431–438.

Falck, J. P., Levy, A., Kastner, M. A., Birgeneau, R. J. (1992), *Phys. Rev. Lett.* **69,** 1109–1112.

Feynman, R. P. (1955), *Phys. Rev.* **97,** 660–665.

Feynman, R. P. (March 1973), private communication.

Feynman, R. P., Hibbs, A. R. (1965), *Quantum Mechanics and Path Integrals,* New York: McGraw-Hill, p. 308.

Feynman, R. P., Hellwarth, R. W., Iddings, C. K., Platzman, P. M. (1962), *Phys. Rev.* **127,** 1004–1017.

Finkenrath, H., Uhle, N., Waidelich, W. (1969), *Solid State Commun.* **7,** 11–14.

Firsov, Yu. A. (Ed.) (1975), *Polarons,* Moscow: Nauka.

Fisher, A. O., Hayes, W., Wallace, D. S. (1989), *J. Phys.: Condensed Matter* **1,** 5567–5593.

Fomin, V. M., Devreese, J. T. (1995), *Solid State Commun.* **96,** 79–84.

Fomin, V. M., Pokatilov, E. P. (1988), *Phys. Rep.,* **158,** 205–336.

Fomin, V. M., Smondyrev, M. A. (1994), *Phys. Rev. B* **49,** 12748–12753.

Friedman, L., Holstein, T. (1963), *Ann. Phys.* **21,** 474–549.

Fröhlich, H. (1937), *Proc. Roy. Soc. London, Sect. A* **160,** 230–241.

Fröhlich, H. (1954), *Adv. Phys.* **3,** 325–361.

Fröhlich, H. (1957), *Arch. Sci. Genève* **10,** 5–6.

Fröhlich, H., Sewell, G. L. (1959), *Proc. Phys. Soc.* **74,** 643–647.

Gehlig, R., Salje, E. (1983), *Philos. Mag.* **47,** 229–245.

Godby, R. W., Schlüter, M., Sham, L. J. (1988), *Phys. Rev. B* **37,** 10159–10175.

Gurevich, V. L., Lang, I. G., Firsov, Yu. A. (1962), *Fiz. Tverd. Tela* **4,** 1252–1262 [English translation: (1962), *Sov. Phys.–Solid State* **4,** 918–925.

Haken, H. (1956), *Nuovo Cimento* **3,** 1230–1253.

Hashimoto, T., Hirasawa, R., Yoshida, T., Yonemura, Y., Mizusaki, J., Tagawa, H. (1995), *Phys. Rev. B* **51,** 576–580.

Hedin, L. (1965), *Phys. Rev. A* **139,** 796–823.

Hellwarth, R. W., Platzman, P. M. (1962), *Phys. Rev.* **128,** 1599–1604.

Hiramoto, H., Toyozawa, Y. (1985), *J. Phys. Soc. Jpn.* **54,** 245–259.

Hodby, J. W., Russell, G., Peeters, F., Devreese, J. T., Larsen, D. M. (1987), *Phys. Rev. Lett.* **58,** 1471–1474.

Höhler, G., Müllensiefen, A. (1959), *Z. Phys.* **157,** 159–165.

Holstein, T. (1959), *Ann. Phys. (USA)* **8,** 343–389.

Hopkins, M. A., Nicholas, R. J., Barnes, D. J., Brummell, M. A., Harris, J. J., Foxon, C. T. (1989), *Phys. Rev. B* **39,** 13302–13309.

Howarth, D. J., Sondheimer, E. H. (1953), *Proc. Roy. Soc. London, Sect. A* **219,** 53–74.

Hubbard, J. (1964), *Proc. Roy. Soc. London, Sect. A* **281,** 401–419.

Hybertsen, M. S., Louie, S. G. (1987), *Phys. Rev. B* **35,** 5585–5601; 5602–5610.

Iadonisi, G., Cataudella, V., Ninno, D., Chiofalo, M. L. (1995), *Phys. Lett. A* **196,** 359–364.

Jackson, S. A., Platzman, P. M. (1981), *Phys. Rev. B* **24,** 499–502.

Johnson, E., Larsen, D. (1966), *Phys. Rev. Lett.* **16,** 655–659.

Jung, W. H., Iguchi, E. (1995), *J. Phys.: Condensed Matter* **7,** 1215–1227.

Kadanoff, L. P. (1963), *Phys. Rev.* **130,** 1364–1369.

Kartheuser, E., Devreese, J., Evrard, R. (1979), *Phys. Rev. B* **19,** 546–551.

Kartheuser, E., Evrard, R., Devreese, J. (1969), *Phys. Rev. Lett.* **22,** 94–97.

Klinger, M. I. (1979), *Problems of Linear Electron (Polaron) Transport Theory in Semiconductors,* Oxford: Pergamon.

Knox, R. (1963), *Theory of Excitons,* New York: Academic.

Kunz, A. B. (1974), in: J. T. Devreese, A. B. Kunz, T. C. Collins (Eds.), *Elementary Excitations in Solids, Molecules and Atoms,* Vol. A, New York: Plenum, pp. 159–187.

Kuper, C. G., Whitfield, G. D. (Eds.) (1963), *Polarons and Excitons,* Edinburgh: Oliver and Boyd.

Landau, L. D. (1933), *Phys. Z. Sowjetunion* **3,** 664 [English translation (1965), *Collected Papers,* New York: Gordon and Breach, pp. 67–68].

Lang, I., Firsov, Yu. (1962), *Zh. Eksp. Teor. Fiz.* **43,** 1843–1860 [English translation (1963), *Sov. Phys.—JETP* **16,** 1301–1312].

Lannoo, M., Baraff, G. A., Schlüter, M. (1959), *Phys. Rev. B* **24,** 955–963.

Laria, D., Wu, D., Chandler, D. (1991), *J. Chem. Phys.* **95,** 4444–4453.

Larsen, D. M. (1969), *Phys. Rev.* **187,** 1147–1152.

Larsen, D. (1972), in: J. T. Devreese (Ed.), *Polarons in Ionic Crystals and Polar Semiconductors,* Amsterdam: North-Holland, pp. 237–287.

Larsen, D. M. (1981), *Phys. Rev. B* **23,** 628–631.

Larsen, D. (1984a), *Phys. Rev. B* **30,** 4595–4608.

Larsen, D. (1984b), *Phys. Rev. B* **30,** 4807–4808.

Larsen, D. M. (1985), *Phys. Rev. B* **32,** 2657–2658.

Larsen, D. (1991), in: G. Landwehr, E. Rashba (Eds.), *Landau Level Spectroscopy,* Vol. 1, Amsterdam: North-Holland, pp. 109–130.

Lax, B. (1972), in: J. T. Devreese (Ed.), *Polarons in Ionic Crystals and Polar Semiconductors,* Amsterdam: North-Holland, pp. 755–782.

Lee, T. D., Low, F. E., Pines, D. (1953), *Phys. Rev.* **90,** 297–302.

Lépine, Y., Matz, D. (1976), *Can. J. Phys.* **54,** 1979–1989.

Lerner, R. G., Trigg, G. L. (Eds.) (1991), *Encyclopedia of Physics,* New York: VCH, p. 941.

Lieb, E. (1977), *Studies Appl. Math.* **57,** 93–105.

Litton, C. W., Button, K. J., Waldman, J., Cohn, D. R., Lax, B. (1976), *Phys. Rev. B* **13,** 5392–5396.

Mahan, G. D. (1972), in: J. T. Devreese (Ed.), *Polarons in Ionic Crystals and Polar Semiconductors,* Amsterdam: North-Holland, pp. 553–657.

Mahan, G. D. (1990), *Many-Particle Physics,* New York: Plenum, pp. 39–40, 635–636.

Marshall, J. T., Chawla, M. (1970), *Phys. Rev. B* **2,** 4283–4287.

Masumi, T., Minami, H., Shimada, H. (1988a), *J. Phys. Soc. Jpn.* **57,** 2674–2677.

Masumi, T., Shimada, H., Minami, H. (1988b), *J. Phys. Soc. Jpn.* **57,** 2670–2673.

Matus, M., Kuzmany, H., Sohmen, E. (1992), *Phys. Rev. Lett.* **68,** 2822–2825.

Merkt, U. (1985), *Phys. Rev. B* **32,** 6699–6712.

Micnas, R., Ranninger, J., Robaszkiewicz, S. (1990), *Rev. Mod. Phys.* **62,** 113–171.

Miura, N., Nojiri, H., Imanaka, Y. (1994), in: D. J. Lockwood (Ed.), *22nd International Conference on the Physics of Semiconductors,* Vol. 2, Singapore: World Scientific, pp. 1111–1118.

Miyake, S. J. (1975), *J. Phys. Soc. Jpn.* **38,** 181–182.

Mott, N. F. (1987), *Conduction in Non-Crystalline Materials,* Oxford: Clarendon.

Mott, N. F. (1990), *Metal-Insulator Transitions,* London: Taylor and Francis.

Mott, N. F. (1991), in: D. P. Tunstall, W. Barford (Eds.), *High Temperature Superconductivity,* Bristol: Adam Hilger, pp. 271–294.

Mott, N. F., Davis, E. A. (1979), *Electronic Processes in Non-Crystalline Materials,* Oxford: Clarendon.

Nagels, P., Devreese, J., Denayer, M. (1964), *J. Appl. Phys.* **35,** 1175–1180.

Nicholas, R. J., Watts, M., Howell, D. F., Peeters, F. M., Xiaoguang, Wu, Devreese, J. T., van Bockstal, L., Herlach, F., Langerak, C. J. G. M., Singleton, J., Chery, A. (1992), *Phys. Rev. B* **45,** 12144–12147.

Peeters, F. M. (1987), in: J. T. Devreese, F. M. Peeters (Eds.), *The Physics of the Two-Dimensional Electron Gas,* New York: Plenum, pp. 393–420.

Peeters, F. M., Devreese, J. T. (1982), *Phys. Rev. B* **25,** 7281–7301.

Peeters, F., Devreese, J. T. (1984), in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics,* Vol. 38, pp. 81–133.

Peeters, F. M., Devreese, J. T. (1985), *Phys. Rev. B* **31,** 3689–3695.

Peeters, F. M., Devreese, J. T. (1986), *Phys. Rev. B* **34,** 7246–7259.

Peeters, F. M., Devreese, J. T. (1987), *Phys. Rev. B* **36,** 4442–4445.

Peeters, F. M., Xiaoguang, Wu, Devreese, J. T. (1986a), *Phys. Rev. B* **33,** 3926–3934.

Peeters, F. M., Xiaoguang, Wu, Devreese, J. T. (1986b), *Phys. Rev. B* **33,** 4338–4340.

Peeters, F. M., Xiaoguang, Wu, Devreese, J. T. (1986c), *Phys. Rev. B* **34,** 1160–1164.

Peeters, F. M., Xiaoguang, Wu, Devreese, J. T. (1988a), *Surf. Sci.* **196,** 437–444.

Peeters, F. M., Xiaoguang, Wu, Devreese, J. T. (1988b), *Solid State Commun.* **65,** 1505–1508.

Pekar, S. I. (1951), *Research in Electron Theory of Crystals,* Moscow: Gostekhizdat [German translation (1954), *Untersuchungen über die Elektronen-*

theorie der Kristalle, Berlin: Akademie Verlag; English translation (1963), *Research in Electron Theory of Crystals,* US AEC Report AEC-tr-5575].

Petelenz, P., Smith, V. H., Jr. (1981), *Phys. Rev. B* **23,** 3066–3070.

Platzman, P. M. (1962), *Phys. Rev.* **125,** 1961–1965.

Pollmann, J., Büttner, H. (1975), *Solid State Commun.* **17,** 1171–1174.

Pollmann, J., Büttner, H. (1977), *Phys. Rev. B* **16,** 4480–4490.

Ranninger, J. (1994), *Physica C* **235–240,** 277–280.

Ranninger, J., Robaszkiewicz, S. (1985), *Physica B* **135,** 468–472.

Reik, H. G. (1972), in: J. T. Devreese (Ed.), *Polarons in Ionic Crystals and Polar Semiconductors,* Amsterdam: North-Holland, pp. 679–714.

Reik, H. G., Heese, D. (1967), *J. Phys. Chem. Solids* **28,** 581–596.

Röseler, J. (1968), *Phys. Status Solidi* **25,** 311–316.

Rüscher, C., Salje, E., Hussain, A. (1988), *J. Phys. C* **21,** 3737–3749.

Sak, J. (1972), *Phys. Rev. B* **6,** 3981–3986.

Salje, E., Güttler, B. (1984), *Philos. Mag. B* **50,** 607–620.

Schafroth, M. R. (1955), *Phys. Rev.* **100,** 463–475.

Scholz, J., Koch, F., Ziegler, J., Maier, H. (1983), *Solid State Commun.* **46,** 665–668.

Seidenbusch, W., Lindemann, G., Lassnig, R., Edlinger, J., Gornik, E. (1984), *Surf. Sci.* **142,** 375–379.

Selyugin, O. V., Smondyrev, M. A. (1989), *Phys. Status Solidi B* **155,** 155–167.

Sewell, G. L. (1958), *Philos. Mag.* **3,** 1361–1380.

Shi, J. M., Peeters, F. M., Devreese, J. T. (1995), in: *March Meeting of the APS,* San José, 1995 (to appear in the Bulletin of the APS).

Sigg, P., Wyder, P., Perenboom, J. A. A. J. (1985), *Phys. Rev. B* **31,** 5253–5261.

Smondyrev, M. A. (1986), *Teor. Math. Fiz.* **68,** 29–44 [English translation: *Theor. Math. Phys.* **68,** 653].

Smondyrev, M. A., Devreese, J. T., Peeters, F. M. (1995), *Phys. Rev. B* **51,** 15008–15016.

Spicci, M., Salkola, M. I., Bishop, A. R. (1994), *J. Phys.: Condensed Matter* **6,** L361–L366.

Stoneham, A. M. (1979), *Adv. Phys.* **28,** 457–492.

Stoneham, A. M., Bullough, R. (1971), *J. Phys. C* **3,** L195–L197.

Su, W. P., Schrieffer, J. R. (1980), *Proc. Natl. Acad. Sci. U.S.A.* **77,** 5626–5629.

Su, W. P., Schrieffer, J. R., Heeger, A. J. (1980), *Phys. Rev. B* **22,** 2099–2111.

Toyozawa, Y. (1963), in: C. G. Kuper, G. D. Whitfield (Eds.) (1963), *Polarons and Excitons,* Edinburgh: Oliver and Boyd, pp. 211–232.

Toyozawa, Y. (1964), *J. Phys. Chem. Solids* **25,** 59–71.

Tyablikov, S. V. (1951), *Zh. Eksp. Teor. Phys.* **21,** 377–388.

Van Royen, J., De Sitter, J., Lemmens, L. F., Devreese, J. T. (1977), *Physica B + C* **81,** 101–106.

Verbist, G., Peeters, F. M., Devreese, J. T. (1990), *Solid State Commun.* **76,** 1005–1007.

Verbist, G., Peeters, F. M., Devreese, J. T. (1991), *Phys. Rev. B* **43,** 2712–2720.

Vigneron, J., Evrard, R., Kartheuser, E. (1978), *Phys. Rev. B* **18,** 6930–6943.

Vinetskii, V. L. (1961), *Zh. Eksp. Teor. Fiz.* **40,** 1459–1468 [English translation: *Sov. Phys.—JETP* **13,** 1023–1028].

Waldman, J., Larsen, D. M., Tannenwald, P. E., Bradley, C. C., Cohn, D. R., Lax, B. (1969), *Phys. Rev. Lett.* **23,** 1033–1037.

Wallis, R. F., Balkanski, M. (1986), *Many-Body Aspects of Solid State Spectroscopy,* Amsterdam: North-Holland.

Wood, R. F. (1991), *Phys. Rev. Lett.* **66,** 829–832.

Xiaoguang, Wu, Peeters, F. M., Devreese, J. T. (1985), *Phys. Rev. B* **31,** 3420–3426.

Xiaoguang, Wu, Peeters, F. M., Devreese, J. T. (1987), *Phys. Rev. B* **36,** 9760–9764.

Yamashita, J., Kurosawa, T. (1958), *J. Phys. Chem. Solids* **5,** 34–43.

Yamazaki, K. (1983), *J. Phys. A* **16,** 3675–3685.

Yu, Lu (1988), *Solitons & Polarons in Conducting Polimers,* Singapore: World Scientific, Chap. IV, Polarons and Bipolarons; Chap. V, Soliton and Polaron Dynamics.

Further Reading

Alexandrov, A. S., Mott N. F. (1994), *Rep. Prog. Phys.* **57,** 1197–1288.

Appel, J. (1968), in: F. Seitz, D. Turnbull, H. Ehrenreich (Eds.), *Solid State Physics,* Vol. 21, pp. 193–391.

Bogolubov, N. N. (1970), *Collected Papers,* Vol. 2, Kiev: Naukova Dumka.

Devreese, J. T. (Ed.) (1972), *Polarons in Ionic Crystals and Polar Semiconductors,* Amsterdam: North-Holland.

Devreese, J. T., Peeters, F. (Eds.) (1984), *Polarons and Excitons in Polar Semiconductors and Ionic Crystals,* New York: Plenum.

Devreese, J. T., Peeters, F. M. (Eds.) (1987), *The Physics of the Two-Dimensional Electron Gas,* New York: Plenum, Chap. II, Electron-Phonon Interaction; Chap. IV, Special Topics.

Feynman, R. P. (1972), *Statistical Mechanics,* Benjamin: Massachusetts.

Feynman, R. P., Hibbs, A. R. (1965), *Quantum*

Mechanics and Path Integrals, New York: McGraw-Hill.

Firsov, Yu. A. (Ed.) (1975), *Polarons,* Moscow: Nauka.

Fisher, A. O., Hayes, W., Wallace, D. S. (1989), *J. Phys.: Condensed Matter* **1,** 5567–5593.

Fomin, V. M., Pokatilov, E. P. (1988), *Phys. Rep.* **158,** 205–336.

Klinger, M. I. (1979), *Problems of Linear Electron (Polaron) Transport Theory in Semiconductors,* Oxford: Pergamon.

Kuper, C. G., Whitfield, G. D. (Eds.) (1963), *Polarons and Excitons,* Edinburgh: Oliver and Boyd.

Lakhno, V. D. (Ed.) (1994), *Polarons & Applications (Proceedings in nonlinear science),* Chichester: Wiley.

Mott, N. F. (1990), *Metal-Insulator Transitions,* London: Taylor and Francis.

Pekar, S. I. (1951), *Research in Electron Theory of Crystals,* Moscow: Gostekhizdat [German translation: (1954), *Untersuchungen über die Elektronentheorie der Kristalle,* Berlin: Akademie Verlag; English translation: (1963), *Research in Electron Theory of Crystals,* US AEC Report AEC-tr-5575].

Yu, Lu (1988), *Solitons & Polarons in Conducting Polimers,* Singapore: World Scientific, Chap. IV, Polarons and Bipolarons; Chap. V, Soliton and Polaron Dynamics.

POLYMER DYNAMICS

Jay D. Schieber, *Department of Chemical Engineering, University of Houston, Houston, Texas, U.S.A., and Department of Chemical and Environmental Engineering, Illinois Institute of Technology, Chicago, Illinois, U.S.A.*

Introduction 415
1. **Overview of Polymer Dynamics** 415
1.1 Goals of Polymer Dynamics .. 417
1.2 Universality of Experimental Results 419
2. **Experimental Techniques** ... 419
2.1 Conformation Studies 419
2.1.1 Light Scattering 419
2.1.2 Small-Angle Neutron Scattering 421
2.1.3 Flow Birefringence 422
2.2 Diffusion in Melts 423
2.3 Macroscopic Flow Studies 424
2.3.1 Optical Techniques 424
2.3.2 Nuclear Magnetic Resonance Techniques 425
2.4 Stress Measurements 425
2.4.1 Shear Flows 425
2.4.2 Shear-Free Flows 429
3. **Theoretical Modeling** 430
3.1 Dilute-Solution Polymers 430
3.1.1 Models of Kuhn and Kuhn, Rouse, and Kramers 430
3.1.2 Generalizations and Improvements 433
3.2 Concentrated Solutions and Undiluted Polymers 435
3.2.1 Reptation Models 435
3.2.2 Anisotropic Mobility Models .. 437
3.2.3 Mode–Mode Coupling Models 438
3.2.4 Network Models 438
3.2.5 Poisson-Bracket Formulations 439
3.2.6 Computer Simulations 439
3.3 Flows and Stresses in Complex Geometries 440
3.3.1 Numerical Solution of Constitutive Equations 440
3.3.2 Combined Stochastic Simulation and Finite-Element Techniques 441
Acknowledgments 441
Glossary 441
Works Cited 442
Further Reading 444

INTRODUCTION

This work is restricted to the dynamics of polymers at microscopic length scales in both dynamic equilibrium and nonequilibrium systems. Macroscopic dynamics of polymeric liquids is also considered, but only insofar as it is predicted from microscopic models. Continuum-mechanical considerations are otherwise ignored. The properties of flexible chains are considered in both dilute and concentrated regimes. An emphasis is placed on those properties that are largely independent of chemical detail. Exceptions are chain architecture, and the consideration of solvent quality for solutions. Special properties of liquid-crystalline polymers (which can form polycrystalline phases), or biological polymers, such as proteins, are not considered. A discussion of static equilibrium properties is limited to those properties that are important for dynamics. However, time-independent conformational properties in flows are discussed. Modeling is primarily restricted to stochastic, mean-field theories for intermediate to long time scales. Given the extremely large number of competing models, theories are broadly categorized, and many details are necessarily omitted.

1. OVERVIEW OF POLYMER DYNAMICS

Polymers are large chains made up of repeating monomers (chemical units) and have molecular weights exceeding roughly 10^4. Synthetic polymers may contain all identical monomer units, whereas biological poly-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

mers, such as proteins, are made up of many different subunits attached in various arrangements. The specific sequence of subunits on a protein may give each protein its individual conformations in the folded shape, such as α sheets, or $\alpha\beta$ barrels. Since we are concerned here with the large-scale, universal behavior of polymers, we do not consider the specific properties of proteins further, although some of the properties outlined here may also describe biological polymers.

Depending on the chemical structure, monomers of synthetic polymers may be attached in linear or branched fashions; examples of simple branched polymers are comb-shaped and star-shaped molecules. A batch of synthesized polymers is never *monodisperse* (all chains having identical molecular weights), but has a distribution of weights, and is hence called *polydisperse.* The details of the distribution of molecular weights may be found by chromatography, but are usually not known for a given batch. Oftentimes, the only information available about the distribution is the single dimensionless quantity called the *polydispersity index* (>1), which is the ratio of the weight-averaged molecular weight of chains M_w to the number-averaged weight M_n,

$$M_n := \frac{1}{n_c}\sum_{i=1}^{nc} M_i, \quad M_w := \frac{1}{n_c M_n}\sum_{i=1}^{nc} M_i^2, \tag{1}$$

where M_i is the molecular weight of chain i in a sample of n_c chains. A monodisperse batch of chains has polydispersity index of unity, and the broader the distribution, the greater the value.

Because of the large molecular weight and the more or less linear fashion in which the chains are constructed, polymers are orders of magnitude longer than they are wide, regardless of the chemical structure of the monomer. This feature suggests the existence of an intermediate length scale that is larger than a monomer, but shorter than the entire chain. If we probe only the dynamics of the chain on this length scale or larger, we expect to find results that are independent of the specific chain chemistry. These intermediate length scales are probably the ones that give rise to important macroscopic behavior. Therefore, we coarse grain the information on specific chemical structure into a few simple ideas:

1. the architecture of the chain (e.g., linear, star shaped, ring),
2. the *Kuhn step length* of the chain, and
3. the polymer–solvent interaction parameter.

The first two ideas are important for both undiluted chains and polymer solutions, whereas the last is applicable only to solutions.

Roughly speaking, the Kuhn step length (Flory, 1969, p. 12) is the average distance necessary to travel from one monomer along the chain to another monomer that has orientation statistically independent from that of the first monomer when the liquid is at equilibrium. Its value will depend upon the polymer chemistry: Molecules with stiff backbones will have values larger than those of flexible chains. A related quantity is the *persistence length* (Doi and Edwards, 1986). If two monomers separated by a distance s along the contour have orientations $\mathbf{u}_0$ and $\mathbf{u}_s$, then the correlation of the orientations falls with increasing s. If we assume that the decrease is exponential, then the persistence length is l_p, where

$$\langle \mathbf{u}_0 \cdot \mathbf{u}_s \rangle = \exp(-s/l_p). \tag{2}$$

To a good approximation, one can assume that the Kuhn step length is twice the persistence length.

The polymer–solvent interaction can be divided roughly into three categories: *good, theta,* and *poor solvents.* The quality of solvent can be described by a single quantity, such as the *Flory interaction parameter* $\chi := \chi_{\text{MS}} - \frac{1}{2}(\chi_{\text{MM}} + \chi_{\text{SS}})$, where χ_{MS} is the energy of interaction between a monomer and the solvent, for example (Flory, 1969; Yamakawa, 1971; de Gennes, 1979; des Cloizeaux and Jannink, 1990). Crudely put, the solvent is called good (poor) when it is energetically more (less) favorable for a monomer to be adjacent to a solvent particle than another monomer and $\chi < \frac{1}{2}$ ($\chi > \frac{1}{2}$). In addition to energetic interactions between monomers, there are also steric interactions. When steric and energetic interactions are balanced, the solvent is called a theta solvent. Since the quality of the solvent is temperature dependent, the *theta temperature* is that at which the solvent is a theta solvent for a given system. Static equilibrium properties of dilute polymer chains in theta solvents can be modeled as a random walk, whereas dilute polymer chains in good solvents are closer to self-avoiding random walks.

Because of the large number of degrees of freedom of the polymer chain, the chain may take effectively an infinite number of possible configurations. The size of the polymer chain may be described by the average

$$\langle R_G^2 \rangle := \frac{1}{N} \sum_{\mu=1}^{N} \sum_{\nu=1}^{N} \langle \mathbf{R}_\mu \cdot \mathbf{R}_\nu \rangle, \tag{3}$$

where R_G is the *radius of gyration*, the angular brackets represent an ensemble average of chains in the system, N is the number of Kuhn steps in a chain, and $\mathbf{R}_\nu$ is the vector from the center of mass of the chain to the νth Kuhn step.

Polymers in good solvents can be categorized as dilute, semidilute, or concentrated. In dilute solutions, polymers exist as independent coils that do not overlap. In a theta solvent, a chain's mean square radius of gyration scales as $\langle R_G^2 \rangle \sim N$, which is consistent with modeling the coarse-grained chain as a random walk. In a good solvent, the coil swells to a larger size and no longer behaves like a random walk, but $\langle R_G^2 \rangle$ scales roughly as $N^{2\nu}$, $\nu \approx \frac{3}{5}$. In a poor solvent, the coil shrinks in size and also no longer scales like a random walk.

A semidilute solution is characterized by the onset of overlap between different coils. Thus, if we call c^{**} the critical number concentration where chains begin to overlap, we can estimate its magnitude by

$$\tfrac{4}{3}\pi \langle R_G^2 \rangle^{3/2} \approx 1/c^{**}. \tag{4}$$

It is believed that chains in the semidilute region do not substantially interpenetrate.

As the concentration is increased further past the critical, molecular-weight- and chemistry-dependent *entanglement concentration* c^*, the chains begin to become entangled with one another, and the kinematics of the chains becomes much slower, and scaling dependence on molecular weight becomes much more pronounced (Graessley, 1974). There is also a minimum *entanglement molecular weight* M_e, below which the chains are not entangled. The presence of the entanglement is evident in a plateau region of the relaxation modulus described in Sec. 2.4.1. At about twice this molecular weight, the zero–shear-rate viscosity scales more strongly with molecular weight.

Brownian forces cause the subunits of the chain to rearrange continually, even at equilibrium, while the center of mass of the polymer chain exhibits Brownian motion. The instantaneous mobility of a chain depends upon its configuration, which may result in anisotropic diffusion for any given configuration. However, the average configuration of chains at equilibrium results in an overall isotropic diffusion for the entire system. The rearrangement of smaller length scales in the polymer results in a spectrum of relaxation times, one for each length scale. The simplest model that exhibits the relaxation of various length scales is the Rouse model described in Sec. 3.1.1.

Through interactions with the solvent or with surrounding chains, the dynamics of the polymer on the *microscopic* length scale can be influenced by *macroscopic* deformations of the fluid, and the average conformation of the polymer chain can be greatly distorted from that at equilibrium. In turn, the anisotropic orientation of the chains leads to anisotropic macroscopic stresses in the fluid. Thus, mechanical experiments that deform the fluid and simultaneously measure the resulting stresses can be used to infer information about the dynamics of the chains.

On the other hand, optical techniques can directly probe the average orientation and dynamics of polymer segments, either in dynamic equilibrium experiments or in nonequilibrium flow fields. When these two techniques are used together, complementary information is obtained.

After describing the goals of research in polymer dynamics in the next section, we explain the method of measuring some of the properties discussed above, as well as additional experimental observations.

1.1 Goals of Polymer Dynamics

The goals of polymer dynamics can be separated into the following broad questions:

1. What is the microscopic dynamics of the polymer chains *at equilibrium?*

2. What is the microscopic polymer dynamics *in flow fields?*
3. How does the microscopic dynamics affect the *macroscopic properties of the flowing continuum?*

The first two questions remain ongoing fields of research both theoretically and experimentally. Much greater progress has been made on the last question for optical and stress properties. Less progress has been made on thermodynamic and mass-transport properties of chains in flow fields.

From the macroscopic point of view, if one applies the idea of momentum balance to a continuum, the equation of motion, or Cauchy's equation, arises (Bird *et al.*, 1960):

$$\rho \frac{\partial \mathbf{v}_f}{\partial t} + \rho \mathbf{v}_f \cdot \nabla \mathbf{v}_f = -\nabla p - \nabla \cdot \boldsymbol{\tau} + \rho \mathbf{g}, \tag{5}$$

where ρ is the density of the fluid, $\mathbf{v}_f$ is the velocity of the fluid, t is time, p is the pressure, $\mathbf{g}$ is the gravity vector, and $\boldsymbol{\tau}$ is the extra stress tensor. To close the equation, we need a constitutive equation relating the stress $\boldsymbol{\tau}$ to the flow field $\mathbf{v}_f$.

For a Newtonian fluid, one assumes that the stress tensor depends linearly on the components of the instantaneous, local velocity-gradient tensor only. To be objective, if the fluid is incompressible and isotropic, the stress tensor depends only upon a single material constant called viscosity. If this Newtonian constitutive equation is inserted into Eq. (5), the Navier–Stokes equation results.

On the other hand, it is easily shown that polymeric fluids do not obey Newton's constitutive equation. For example, if a flow field is suddenly switched off after being applied for some time, the stress field of a polymer fluid decays to zero only slowly, whereas that for a Newtonian fluid decays to zero instantaneously. Thus, the Navier–Stokes equation does not apply, and we need to find the proper constitutive equation for $\boldsymbol{\tau}$. *No single constitutive equation has been found that correctly describes the stresses arising in all flow fields for polymeric fluids.* For this reason, this field of research is primarily experimental. We discuss below some of the progress made from a microscopic viewpoint to derive a proper constitutive equation.

Qualitatively, we know that the constitutive equation must depend upon the kinematic history of the deformation for any given material element. Thus, experiments that probe the continuum must be careful to insure that the history of the deformation of the material is known in order to infer the relationship between stress and strain. Here we focus on those flows that are homogeneous deformations at all times throughout the entire region of the sample being tested.

The deviation from Newtonian behavior is not surprising, given the above observations: Polymer chains can extend in space and exert stresses over distances, their conformations are distorted away from isotropic in flow fields, and the various length scales of the polymer rearrange on different time scales. Therefore, it may also not be surprising that if we know the dynamics of the polymer chain, it is relatively straightforward to know the polymer contribution to the stress field. Derivation of these relations is outside the scope of this article, but is based on either statistical-mechanical (Bird *et al.*, 1987) or thermodynamic (Beris and Edwards, 1994; Jongschaap, 1991; Grmela, 1985) arguments. The two approaches often yield complementary insight into polymer dynamics.

For example, if the polymer chain in a melt is modeled as elastic "connectors" whose orientations and lengths are given by the set of vectors $\{\mathbf{Q}_i\}$, then the polymer contribution to the stress tensor is called the *Kramers form for the stress tensor,*

$$\boldsymbol{\tau}_p = -n \sum_i \langle \mathbf{Q}_i \mathbf{F}_i^{(c)} \rangle, \tag{6}$$

where n is the number density of polymer chains and $\mathbf{F}_i^{(c)}$ is the force exerted by the ith connector.

We note here that linear-optical measurements probe the second moment of the chain segment orientations $\langle \mathbf{Q}_i \mathbf{Q}_i \rangle$. If a particular theory predicts that the connector force is linear in segment orientation $\mathbf{Q}_i$, then optical measurements provide a consistency check with the theory. On the other hand, if the theory predicts a tension nonlinear in $\mathbf{Q}_i$, then optical measurements provide additional information about average orientations.

Not considered further here is the fact that other linear flux laws do not apply for polymer melts. For example, anisotropic orientation leads to non-Fickian diffusion of the polymer, and it is expected that heat flux is not described by Fourier's law. Microscopic models are already giving insight into generalizations of these laws to polymeric liquids (Curtiss and Bird, 1995).

1.2 Universality of Experimental Results

Many properties of a given polymeric fluid show behavior similar to other polymers in its class (same chain architecture, and same solvent quality if in solution) when the results are properly rescaled. For example, polymer melts have a zero–shear-rate viscosity that scales with molecular weight to the 3.4 power, for long chains. All polymers exhibit a spectrum of relaxation times. However, when time is rescaled by the longest relaxation time, the results often look quite similar for all polymers in that class. Likewise, the viscosity as a function of shear rate gives similar results for all polymer melts. The results given below are meant to be indicative of all polymers in a given class.

2. EXPERIMENTAL TECHNIQUES

2.1 Conformation Studies

Scattered radiation can be used to measure static or dynamic properties of polymers in melt or solution. The general theory of scattering for polymers applies to light, x-ray, or neutron scattering. However, the details of each method often make one more suitable than the others for either melts or solutions. Here we discuss light scattering by solutions, and neutron scattering by melts.

2.1.1 Light Scattering As lucidly described in the classic text by Berne and Pecora (1976), radiation scattering can be conceptually divided into two steps: scattering from a medium whose index of refraction is fluctuating in time and the influence of a scattering center on the local index of refraction. The first step is made from a macroscopic viewpoint, whereas the second makes the connection between microscopic dynamics and macroscopic observation.

If we assume that the radiation field is weak, so that the medium's response is linear, the intensity of scattered light depends upon the fluctuations in the dielectric constant found using Maxwell's laws of radiation. If the medium has average (constant) dielectric constant ϵ_0, and a fluctuating (in general, anisotropic) tensorial component $\Delta\boldsymbol{\epsilon}(\mathbf{r},t)$, then the total dielectric tensorial field $\boldsymbol{\epsilon}$ is

$$\boldsymbol{\epsilon}(\mathbf{r},t) = \epsilon_0\boldsymbol{\delta} + \Delta\boldsymbol{\epsilon}(\mathbf{r},t), \tag{7}$$

where $\mathbf{r}$ is the position in space and $\boldsymbol{\delta}$ is the unit tensor. The intensity of scattered radiation a distance R away from the *scattering volume* (volume of the sample on which the light impinges) is given by

$$\begin{aligned} I_{if}(\mathbf{q},\omega_f - \omega_i,R) = {}& \left[\frac{I_0 k_f^4}{16\pi^2 R^2 \epsilon_0^2}\right] \\ & \times \frac{1}{2\pi}\int_{-\infty}^{\infty} \langle \Delta\epsilon_{if}(\mathbf{q},0)\Delta\epsilon_{if}(\mathbf{q},t)\rangle \\ & \times \exp[i(\omega_f - \omega_i)t]dt, \end{aligned} \tag{8}$$

where I_0 is the intensity of incident light, and ω_i (ω_f) is the frequency of incident (scattered) light. The vector $\mathbf{q}$ is the difference between the incident and scattered wave vectors,

$$\mathbf{q} := \mathbf{k}_i - \mathbf{k}_f, \tag{9}$$

where $\mathbf{k}_i$ ($\mathbf{k}_f$) is the wave vector of the incident (scattered) light, which has orientation in the direction of propagation, and magnitude ω_i/c (ω_f/c). The function $\Delta\epsilon_{if}(\mathbf{q},t)$ is the component of the spatial Fourier transform of the fluctuating part of the tensorial dielectric constant in the incident and scattered directions:

$$\Delta\epsilon_{if} := \mathbf{n}_i \cdot \left[\int \exp(i\mathbf{q}\cdot\mathbf{r})\Delta\boldsymbol{\epsilon}(\mathbf{r},t)d^3r\right]\cdot \mathbf{n}_f, \tag{10}$$

where $\mathbf{n}_i$ ($\mathbf{n}_f$) is the direction of travel of the incident (scattered) wave.

Because a laser is used for light scattering, or a collimator for neutron scattering, the incident radiation is monochromatic, so that a single value is sufficient to describe ω_i. Measuring the scattered intensity at all frequencies is equivalent mathematically to integrating $I_{if}(\mathbf{q},\omega,R)$ over all ω. When this integration is performed on Eq. (8), the exponential inside the integral on the right-hand side becomes a delta function, and the total intensity contains information about the second moment of the fluctuations in the dielectric field. Intensities measured at specific frequencies contain information about the autocorrelation function $\langle \Delta\epsilon_{if}(\mathbf{q},0)\, \Delta\epsilon_{if}(\mathbf{q},t)\rangle$. If the scattering process is elastic (i.e., $\omega_i \approx \omega_f$), then one can show that

$$q \equiv |\mathbf{q}| = (4\omega_i/c)\sin(\theta/2), \qquad (11)$$

where θ is the angle between $\mathbf{k}_i$ and $\mathbf{k}_f$. By holding the incident wave and scattering volume stationary, and moving the position of the detector that measures the intensity of scattered radiation, the magnitude of the vector $\mathbf{q}$ can be varied. In addition, one can vary the frequency of the incident radiation to vary q. The scalar q is sometimes mislabeled the "scattering vector," so that discussions are often seen on "the value of the scattering vector."

To extract information about the molecular dynamics from the fluctuations in the dielectric constant extracted from radiation-scattering measurements, one needs a model to predict the influence of molecules on the dielectric field. This step of the procedure is not nearly as "clean" as the first for polymeric systems—one typically must make several simplifying approximations that can be justified only *a posteriori* by quantitative comparison with data.

For light scattering, the scattering target is the electron cloud, which is polarizable. Hence, the effective scattering length increases monotonically with molecular weight of the atom, and most scattered light is coherent. Ability to scatter is directly proportional to the polarizability of the molecule, so that polymers are more effective scatterers than solvent molecules. Also, since polymers relax on time scales much longer than those for solvent particles, one can usually separate the influence of polymer dynamics from that of the solvent. Typical length scales for light scattering are in the range 0.025–0.2 μm.

It is generally assumed that a segment of a polymer chain can be treated as a point (or spherical) scatterer, if the size of the segment is much less than q^{-1}. The concentration and volume size are chosen so that multiple scatterings are negligible. Then, the mathematical problem is reduced to finding the scattering intensity of a single point particle. Utilizing the solution to this simpler problem, one can find the intensity of scattered light from n_c chains of N segments of point scatterers. If the chains are also identical and noninteracting, then there is no correlation between the scatterers on different chains. If we further assume that each chain segment has the same polarizability, then Eq. (8) can be simplified to depend upon the scattering from a single polymer:

$$\begin{aligned} I_{if}(\mathbf{q},\omega_i,\omega_f,R) = {} & \left[\frac{I_0\omega_f^4 N c' \alpha_m'^2 V}{16\pi^2 R^2 \epsilon_0^2 \bar{c}^4 \tilde{N}_{\mathrm{A}}}\right] \\ & \times (\mathbf{n}_i \cdot \mathbf{n}_f)^2 \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\mathbf{q},t) \\ & \times \exp[i(\omega_f - \omega_i)t]dt, \end{aligned} \qquad (12)$$

where $\bar{c}$ is the speed of light, V is the volume of the scattering sample, c' is the mass concentration of polymer, α_m' is the polarizability per unit mass of polymer, $\tilde{N}_{\mathrm{A}}$ is Avogadro's number, and $S(\mathbf{q},t)$ is the *dynamic structure factor* for a single polymer,

$$S(\mathbf{q},t) := \frac{1}{N}\sum_{lm} \langle \exp\{i\mathbf{q}\cdot[\mathbf{r}_l(t) - \mathbf{r}_m(0)]\}\rangle. \qquad (13)$$

If the scattered light is detected at all frequencies, then the only contribution from the structure factor is $S(\mathbf{q},0)$, the *static structure factor*. All other quantities in Eq. (12) except polymer molecular weight M are measurable physical quantities, or are estimated from calibration. For small values of q, by expanding the exponential in Eq. (13), one can show that

$$1/S(\mathbf{q},0) \approx 1 + \tfrac{1}{3}q^2\langle R_G^2\rangle. \qquad (14)$$

Therefore, by measuring the intensity of scattered light at several small angles and small concentrations, and performing a double extrapolation in both concentration and q (a so-called *Zimm plot*), one can obtain both the molecular weight M and the radius of gyration $\langle R_G^2\rangle$. Generalization of this derivation shows that for a polydisperse mixture of chains, the weight-averaged molecular weight is obtained, but then no simple relation exists for the second-order term in q.

Experimentally, Debye showed by light scattering that $\langle R_G^2\rangle \sim M$ for a linear chain in theta solvent, and $\sim M^{6/5}$ for a good solvent, in agreement with the predictions of Sec. 1.

By measuring the intensity of the scattered light at various frequencies, information can be obtained about the time dependence of the dynamic structure factor. From its definition Eq. (13), one sees that, in general, the dynamic structure factor will depend in complicated ways upon the dynamics of the polymer chain. Therefore, to extract useful information one begins with a model for polymer dynamics and predicts the resulting structure factor to compare with experiment. However, some information can be extracted from the experiments when the scattering angle is very small. For sufficiently small values of q, the entire molecule acts as a scatterer, if interference from intramolecular length scales is negligible. Then, one can show that $S(\mathbf{q},t) \sim \exp(-q^2Dt)$, where D is the binary solvent/polymer diffusion coefficient.

Caroline and Jones (1980) studied polystyrene in cyclohexane (at theta temperature) to find that the diffusion coefficient scales as $\sqrt{M_w}$ at infinite dilution. The largest time constant of relaxation by normal-mode analysis yielded a molecular-weight dependence to the 1.4 power.

In flow, the average chain orientation is deformed into an ellipsoid with a preferred direction. In shear flow, the major axis points in a direction somewhere between the direction of flow and 45° toward the direction of the shear gradient. If light scattering is performed in different directions relative to the flow, information can be obtained about the shape of the ellipsoid. For example, light-scattering experiments by Link and Springer (1993) on polystyrene dissolved in dioctyl phthalate (nearly a theta solvent) have found the sizes of the three principal axes and the orientation of the ellipsoid relative to the flow direction describing the average conformation of the chains during shear flows. By extrapolating their results to zero polymer concentration, they find that the orientation angle α of the ellipsoid varies with shear rate as

$$\alpha = \tfrac{1}{2}\tan^{-1}\left(\frac{2.1 + 0.1\beta}{\beta}\right), \qquad (15)$$

where β is the shear rate made dimensionless by $RT/\eta_{p,0}M_w$ and $\eta_{p,0}$ is the zero–shear-rate polymer contribution to viscosity defined in Sec. 2.4.1. They were also able to measure the three ellipsoid lengths. They find that the ellipsoid axis perpendicular to both flow and shear directions does not change in length from its equilibrium value. The major axis grows nearly linearly with shear rate, and the overall radius of gyration made dimensionless by its equilibrium value is (Link and Springer, 1993)

$$\frac{\langle R_G^2\rangle}{\langle R_G^2\rangle_{\mathrm{eq}}} = 1 + 0.017\beta^{1.4}. \qquad (16)$$

They point out, however, that the data of Cottrell *et al.* (1969) indicates that polyisobutylene in decalin has a deformability $\langle R_G^2\rangle/\langle R_G^2\rangle_{\mathrm{eq}}$ that grows nearly linearly with β.

2.1.2 Small-Angle Neutron Scattering

Small-angle neutron scattering differs from light scattering in two important ways: Neutrons at room temperature have a frequency such that they can probe length scales down to 2 Å, and they are scattered by nuclei rather than electrons. The first difference suggests that neutron scattering may be used as a tool complementary to light scattering, since it can probe smaller length scales. The second property makes neutron scattering appropriate for melts and concentrated solutions (as well as dilute solutions) for the following reasons.

For effective study of chain dynamics in the melts, it is necessary to make most of the chains "invisible" to the incident radiation, while making a few chains effective scatterers. Since light scattering relies on the electron density for scattering size, labeling a few chains in a melt requires substantial change in the chemical structure of the labeled chains, which risks changing the dynamics of the chains being probed. Since the neutron is scattered by the nucleus of the atom, isotopes can be used to create scattering contrast, without changing chemistry appreciably. Such a contrast exists, for example, between hydrogen and deuterium. However, one must be careful not to increase the temperature above a critical value causing separation of the undeuterated and deuterated chains into different phases (Bates *et al.*, 1986).

Neutron scattering on polymer melts has confirmed that the radii of gyration of polymethylmethacrylate, polystyrene, and polyethylene have dimensions comparable to dilute polymers in theta solvents, with ratios of $\sqrt{\langle R_G^2\rangle}/M_w$ equal to 0.31, 0.23, and 0.45, respectively (Higgins, 1979).

To date, neutron-scattering experiments on polymers in flow fields have been accomplished in two flow cases: scattering by solutions during flow, and scattering by melt samples that have been quenched during deformation. The latter experiment was performed by Muller *et al.* (1993) on a polystyrene melt that had been quenched during a creep experiment. By slicing the quenched sample in different planes, the coil orientations could be studied in all directions. By varying the magnitude of the scattering vector to probe different length scales, they find that the principal directions of orientation in the shear plane depend upon the length scale of the polymer; smaller length scales show less anisotropy in orientation. This is not surprising, since the small-angle neutron-scattering experiment is able to probe length scales smaller than the Kuhn length. They find that the direction perpendicular to the flow plane is unaffected by flow as in the polystyrene theta solution, and the deformation of the coil in the flow plane is less than that predicted by affine motion of the segments.

2.1.3 Flow Birefringence When polymer chains are oriented by electrical, magnetic, or mechanical means, the (tensorial) index of refraction of the fluid **n** can be made anisotropic. By careful control of the perturbing force, the anisotropic effect can be made homogeneous throughout the sample. If polarized light is made to impinge on the sample, and the intensity of light that exits from the sample is measured for different orientations of polarization, information can be obtained about the index-of-refraction tensor. Since the orientation of the chains causes the anisotropy of **n**, information is indirectly obtained about the instantaneous conformations of the chains.

Flow birefringence is the study of the anisotropy of the index of refraction of polymeric liquids that are undergoing flow (Fuller, 1995). Linear-optic theory can be used to interpret the results of such experiments. For example, consider light that travels in the z direction relative to a laboratory-fixed reference frame, and is polarized in the x direction prior to passing through a birefringent medium. The light will "see" three components of the index of refraction projected on the plane perpendicular to the direction of propagation of the impinging light. By a suitable rotation of the coordinate system, these components can be made diagonal. Thus, the components describe an ellipse of major axis, minor axis, and orientation. The intensity of light transmitted through an analyzer oriented along the y axis is given by (Riande and Saiz, 1992)

$$I = I_0 \sin^2 2\beta_e \sin^2(\delta_R/2), \tag{17}$$

where I_0 is the intensity of polarized light impinging on the sample, β_e is the angle of the ellipse relative to the x axis, called the *extinction angle*, and δ_R is the *retardation* of the medium defined by

$$\delta_R := 2\pi l \Delta n/\lambda, \tag{18}$$

where $\Delta n = n_I - n_{II}$ is the difference between the major and minor components of the index of refraction in the plane, l is the length of sample through which the light passes, and λ is the wavelength of light. By rotating the polarizer and analyzer in tandem, β_e can be varied. When the intensity of the light becomes zero, then the orientation of the index-of-refraction ellipse is known, and δ_R can be calculated from Eq. (17). Since the other variables appearing in Eq. (18) are easily measured, Δn can be found.

Janeschitz-Kriegl (1983) has shown that for all experiments involving shear flows, the stress tensor for polymer melts is, within a constant, given by the index-of-refraction tensor. The relation is called the *stress-optic law,*

$$\mathbf{n} = C\boldsymbol{\tau}, \tag{19}$$

where C is called the *stress-optic coefficient,* and, in general, is a function of material and temperature, but is not a function of strain or strain history. Theoretically, such an equivalence is predicted for any model that predicts that the stress tensor and the index-of-refraction tensor depend, within a constant, upon the same averages of the orientation of the polymer chain. Since simple, physical arguments suggest that $\mathbf{n} \sim \Sigma_i\langle\mathbf{Q}_i\mathbf{Q}_i\rangle$, it is believed that the Kramers form for the stress tensor, Eq. (6), is the proper one for polymeric melts with $\mathbf{F}_i^{(c)} \sim \mathbf{Q}$.

Practically, the stress-optic law means that some rheological measurements can be performed optically rather than mechanically. Since measuring normal stress differences can be rather difficult, the stress-optic law allows us to infer normal stress differences more easily. Many experimental data determined by flow birefringence may be found in Janeschitz-Kriegl (1983).

2.2 Diffusion in Melts

There are three different methods for defining a diffusion coefficient that are not necessarily equivalent. Each method involves the average distance that a chain may move when experiencing a given driving force. When the driving force is a concentration gradient, the diffusion coefficient may be "cooperative," whereas when the driving force is Brownian motion, and a single chain is followed, it is called the "self-diffusion coefficient." Chains may also diffuse from an external force, say an electric field. Cooperative diffusion may behave quite differently from self-diffusion, since entangled chains moving collectively may have a large contribution to the coefficient. In fact, the cooperative diffusion coefficient for semidilute chains increases with concentration to the power $\frac{2}{3}$ (de Gennes, 1979, p. 211).

Self-diffusion in polymer melts is difficult to measure because of the very long time scales associated with diffusion over very small length scales, and because of the difficulty in isolating the dynamics of a single chain from the melt. For example, neutron scattering cannot be used to measure the diffusion coefficient directly, because the time scales and length scales probed are too small for diffusion; light scattering is problematic, because isolating a chain from the others requires that the chemical structure of the polymer being probed be different from the bulk of chains.

The advantages and disadvantages of a number of methods are reviewed in an article by Tirrell (1984), where experimental diffusion results are compiled. We focus here on a single method, *forced Rayleigh scattering* (FRS), because the method appears to be the most useful and least problematic, and many other methods (e.g., fluorescence redistribution after pattern photobleaching, indirect neutron scattering, and isorefractive quasi-elastic light scattering) use analogous principles.

In FRS, one labels a small number of chains with a dye (e.g., fluorescein) that changes its adsorption spectrum when exposed to light of a particular wavelength. Since the dye group is much smaller than the chain, it is hoped that the dynamics of the labeled chains is not perceptibly altered. One then uses crossed beams of a split writing laser to alter the dye group on some of the chains. A sinusoidal interference pattern is created in the beams so that sandwiched layers of altered and unaltered chains in the melt result. Once the writing laser is switched off, the melt relaxes to a homogeneous mixture of altered and unaltered chains by diffusion. A second laser whose light is absorbed by the altered dye (but not the unaltered dye) is used to "read" the distribution in the melt, which now acts like a diffraction grating. By following with time the ability of the grating to diffract, one can calculate the diffusion coefficient (assuming Fickian diffusion over these time scales).

It is necessary to make the thickness of each layer sufficiently small that no appreciable amount of the dye decays back to its unaltered state before diffusion makes the sample homogeneous. By varying the crossing angle of the writing lasers, different layer thicknesses can be created, thereby allowing different times of diffusion to be studied.

Other techniques use asymmetric interfaces, such as a sample of large chains adjacent to a sample of small chains, by following the movement of the interface. Backscattering of tracer particles, infrared spectroscopy, or electron microscopy is used to follow the concentration profiles at the interface.

The bulk of experiments reveal some general trends (Tirrell, 1984). The self-diffusion coefficient of entangled chains in the melt scales with molecular weight as $D \sim 1/M^2$ for nylon 12, polystyrene, and polyethylene. The results for polyethylene, the most flexible of the three, are rather conclusive. However, some of the data for polystyrene suggest that the diffusion coefficient may scale with molecular weight as $1/M^{2.7}$ for long chains. Some data reported since the review of Tirrell have also suggested that D falls with increasing molecular weight more rapidly than $1/M^2$ for some systems (Nemoto *et al.*, 1984, 1985; Kim *et al.*, 1986).

If molecular-weight dependence is studied for chains at and below the entanglement length, it is found that the $1/M^2$ dependence holds well below the entanglement length. Very short chains show $D \sim 1/M$, which is consistent with the Rouse dynamics discussed in Sec. 3.1.1 when chain segment concentration is held constant.

The diffusion coefficient of star-shaped chains with three arms in the melt exhibits a dependence consistent with $D \sim M^{-1} \times \exp(-bM)$ for chains of medium length, but deviates from this form for higher molecular weights. When polyisobutylene, polystyrene, or polybutylene are unentangled, branching was found to have no important effects on the dependence of the diffusion coefficient on molecular weight.

When a single chain of weight M is probed as it diffuses through a matrix of longer chains M_m far above the glass transition temperature, the diffusion coefficient of the short chain is independent of M_m. However, when the matrix chains are much shorter than the probe chain, $D \sim 1/M$, and the matrix molecular weight becomes important if $M > M_m \approx M_e$. If a single chain is diffusing through a fixed network, its diffusion coefficient is approximately three times the self-diffusion coefficient. If the mixture is polydisperse, then $D \sim 1/M_n M_w$.

The diffusion coefficient of concentrated polystyrene in good and theta solvents has been studied extensively. The concentration dependence appears to be nearly the same in both solvents, and no simple scaling law was seen; the diffusion coefficient decreased strongly with concentration c. The slope of $\log D$ vs $\log c$ appears to be greater than -10 at the highest concentration.

2.3 Macroscopic Flow Studies

2.3.1 Optical Techniques Because polymeric liquids do not obey the Navier–Stokes equation, the flow patterns resulting in complex geometries can be dramatically different from those of Newtonian fluids. These flow patterns can give clues to the importance of shear thinning or elastic properties on the stability of roll cells or vortices, for example. When predictions are made by solution of Cauchy's equation together with the proposed constitutive equation, the results can be compared with the measured flow field.

Streamlines can be followed quite cheaply and easily by flow-visualization techniques, which use video images, or long-time exposure of photographs to follow the trajectories of brightly reflecting particles. Some beautiful photographs of these effects can be seen in Boger and Walters (1993). In these series of photographs, one sees, for example, that

1. when a rotating rod is inserted through a free surface into a container of polymeric fluid, the polymer climbs the rod;
2. upon exit from a die, the stream of a polymeric fluid expands so that the exit stream has a much greater radius than the die radius;
3. a liquid fiber can be continuously drawn from the container, and wound onto a drum;
4. if the driving force for a steady flow is suddenly switched off, the fluid *recoils* backward some substantial distance; and
5. the addition of small amounts of polymer to a solution can significantly affect the range of stable flow past a cylinder.

Macroscopically, many of these effects can be described qualitatively by the presence of a first normal-stress difference (**1**,**2**,**3**,**5**), by the concept of "memory" (**2**,**4**,**5**), or by shear-rate-dependent viscosity (**5**). None of these effects occurs for Newtonian fluids.

If more precise measurements of fluid velocities are desired, *laser Doppler anemometry* (LDA) can be used to measure the velocity of the fluid at any point in space. *Laser Doppler velocimetry* is an application of light scattering described in Sec. 2.1.1, in which quasi-elastic scattering of light from particles in the fluid is used to measure the local flow field.

There is a growing trend to measure the flow field and the stress field simultaneously in complex geometries by combining multiple techniques. For example, in a recent work, LDA experiments were used to track particles in a 10:1 contraction. The particle trajectories were then used to compute predicted stress fields and compared with data (Feigl and Öttinger, 1994). Rajagopalan *et al.*

(1992) compared finite-element predictions with birefringence measurements in flow between eccentric rotating cylinders using a multimode Giesekus model.

2.3.2 Nuclear Magnetic Resonance Techniques Nuclear magnetic resonance imaging (NMR or MRI) techniques provide an alternative way to obtain velocity fields in complex geometries, by tracking the NMR signals of nuclei in the fluid, typically hydrogen atoms. This technique does not require that all solid objects be optically transparent to light or other probing radiation, and provides a three-dimensional picture of fluid motion. The method was originally developed to track blood flow in living bodies.

The flowing sample is placed in an electric field to align the H nuclei. Then, an rf signal of narrow width is used to flip the spins of the protons in a very thin cross section of the flow, thus labeling those protons in the cross section. The protons begin to precess back to their original orientations in the field. After a chosen time t_0, a second rf signal is used to flip the protons in the cross section again. The protons that have entered the cross section after the first signal have orientation different from the protons that were present in the cross section during the first rf signal. By following the spatial dependence of the NMR signal in the cross section, it is possible to infer which sections of the region have been replenished by new fluid. By changing t_0 it is hence possible to infer the fluid velocity. This method has been used by Li and McCarthy (1994) to map the flow field in a straight pipe of circular cross section for aqueous polyacrylamide solutions. They were able to measure the axial velocities in laminar flow, and mean axial velocities in turbulent flow. By combining pressure-drop measurements with the measured shear rate at the walls, viscosity curves as functions of shear rate could be constructed. Since the pressure varies down the length of the tube, the method might be used to give pressure dependence of viscosity by measuring the shear rate at various axial positions.

2.4 Stress Measurements

Model-independent rheological measurements can be obtained in flow fields that are homogeneous, i.e., flow fields in which the velocity gradient is independent of position. Most important homogeneous flow fields can be divided into the two types: shear (transient or steady state) or shear-free (transient only) flows.

2.4.1 Shear Flows Shear flows are anisotropic, and therefore have vorticity. They can be described in the appropriate coordinate system by $dv_y/dx = \dot{\gamma}$, all other gradient components zero, where $\dot{\gamma}$ is the *shear rate.* Important measured quantities are the shear stress τ_{yx}, the first normal-stress difference $\tau_{xx} - \tau_{yy}$, and the second normal-stress difference $\tau_{yy} - \tau_{zz}$. The ratio of shear stress to shear rate is the viscosity $\eta := \tau_{yx}/\dot{\gamma}$. Important time-dependent flows are single-step strain $\dot{\gamma} = \gamma\delta(t)$ (also multiple-step strains), inception of steady shear flow $\dot{\gamma} = \dot{\gamma}_0\mathrm{H}(t)$, cessation of steady shear $\dot{\gamma} = \dot{\gamma}_0\mathrm{H}(-t)$, and small-amplitude oscillatory shear flow $\dot{\gamma} = \dot{\gamma}^0 \cos(\omega t)$. $\mathrm{H}(t)$ is the Heaviside step function. The other parameters are all held constant during an experiment, but can be varied arbitrarily from experiment to experiment, except for the strain $\dot{\gamma}^0$, for which we require that $\dot{\omega}\gamma^0 \lll 1$. Other flows are not considered here, such as exponential shearing flow: $\dot{\gamma} = \mathrm{H}(t)A \exp(\alpha t)$, where A and α are arbitrarily chosen constants for a given experiment.

Linear-viscoelastic theory (Ferry, 1980) relates all experiments of very small deformation to a single material property called the *relaxation modulus* $G(t)$. The function $G(t)$ is defined from the single-step strain experiment in the limit that $\gamma \to 0$:

$$G(t) := \tau_{yx}(t,\gamma)/\gamma, \quad \text{step strain: } \gamma \to 0. \qquad (20)$$

The relaxation modulus exhibits three regimes for polymer melts: an initial, rapidly decaying regime, a plateau, and a second decay. The plateau's height is independent of molecular weight, but its temporal location depends upon molecular weight to the 3.4 power, i.e., the longest relaxation time constant scales with molecular weight to the 3.4 power.

At higher (nonlinear) strains, the shear stress is found to be separable into a strain-dependent part $h(\gamma)$ called the *damping function,* and the time-dependent relaxation modulus:

$$\tau_{yx}(t,\gamma) = \gamma h(\gamma)G(t), \quad \text{single-step strain.} \tag{21}$$

The damping function is a dimensionless quantity that is unity at vanishingly small strain and is found to decrease monotonically with strain γ. The separability is often not observed at the shortest times. Empirically, we find that many data are described by a function of the form

$$h(\gamma) = 1/(1 + a_d\gamma^2), \tag{22}$$

where a_d is a material constant. For example, a highly branched low-density polyethylene was found by Khan and Larson (1987) to have $a_d \approx 0.07$. Concentrated solutions of linear polystyrene exhibit values of a_d between 0.11 and 0.22. The first normal-stress difference is observed to decay with the same time dependence as the viscosity; the observation is referred to as the Lodge–Meissner rule. Such experiments are typically not performed for dilute solutions.

Experiments of multiple-step strains that are multidirectional have also been performed (Meissner, 1984; Hürlimann, 1991; Venerus and Kohvand, 1994). For example, one imposes an instantaneous shearing step strain, waits a period of time, and then performs a second step in the opposite direction. Alternatively, the second step could be performed in a direction perpendicular to the first step. The results of these experiments are not easily given in a brief description, and hence we do not discuss them further here. However, it is believed that these experiments are very important to distinguish between competing molecular theories, since many theories predict results quite different from each other, and no theory has successfully described all observed multistep-strain results to date.

Small-amplitude oscillatory shear flow is thought to probe the relaxation dynamics of the chain at equilibrium, and is effectively studied using linear viscoelastic theory. From linear viscoelastic theory, and experimentally, it is found that the results of the small-amplitude oscillatory shear flow experiment can be related to $G(t)$. The first normal stress is second order in strain, and does not play a role in linear viscoelasticity. The shear stress responds with a part that is in phase with the deformation, and a part that is out of phase. If we write the shear stress in the form

$$\tau_{yx} = -\dot{\gamma}^0\mathcal{R}\{\eta^* e^{i\omega t}\}, \tag{23}$$

then the in-phase part of the shear stress is described by the real part η' of η^*, and the out-of-phase part is described by negative of the imaginary part η''. Experimentally and theoretically, it is found that the complex viscosity η^* is the Fourier transform of the relaxation modulus $G(t)$. Thus, the small-amplitude oscillatory shear flow experiment mechanically probes the various relaxations on different time scales by varying the frequency of the experiment.

The relaxation modulus is often fitted by a finite sum of exponentials: $G(t) = \Sigma_n a_n \times \exp(t/\lambda_n)$, where the $\{a_n\}$ represent the weighting factors for each of the time constants $\{\lambda_n\}$. Baumgärtel *et al.* (1990) have found a general trend for the weighting factors $\{a_i\}$ with time constants $\{\lambda_i\}$ for linear, nearly monodisperse polymer melts. They studied entangled polystyrene and polybutadiene with polydispersity indices between 1.02 and 1.13 just above the glass transition temperature. Plots of $\{\log a_n\}$ vs $\{\log \lambda_n\}$ show a general V shape. The initial linear, downward-sloping portion of the curve was found to be universal for all chains studied, and had slope of -0.84. This portion of the curve at short time constants was attributed to the glassy modes of the polymers. The final linear, upward-sloping portion of the plot depended on molecular weight and chemical structure of the chains. Polystyrene showed a slope of 0.22 (all M_w), whereas polybutadiene revealed a slope of 0.42. On the basis of these results, the authors recommended representing the relaxation modulus as a sum of $N_1 + N_2$ exponentials,

$$\begin{aligned} G(t) = &\sum_{i=1}^{N_1} g_1 a_1^{n_1(i-1)} \exp\left(\frac{-t}{a_1^{i-1}\lambda_1}\right) \\ &+ \sum_{i=1}^{N_2} g_2 a_2^{-n_2(i-1)} \exp\left(\frac{-t}{a_2^{i-1}\lambda_2}\right), \end{aligned} \tag{24}$$

where g_1, g_2, a_1, a_2, λ_1, and λ_2 are material parameters. The parameters n_1 and n_2 are the slopes described above.

Polymer melts and polymer solutions exhibit qualitatively similar behavior for η^*. That is, η' is independent of frequency ω at low frequencies and drops off with increasing ω; at high frequencies dilute solutions may show a second leveling off of η'. The out-of-phase part η'' initially rises linearly with ω, goes through a maximum, and eventually falls with increasing ω. Polydispersity tends to broaden the shapes of these curves.

The steady-state viscosity $\eta := -\tau_{yx}/\dot{\gamma}$ has many universal features for polymer melts. The initial plateau value η_0 increases with molecular weight to the first power for unentangled chains, and to the 3.4 power for entangled chains. The shear-thinning regime obeys a power law $\eta \sim \dot{\gamma}^b$, where b is between 0 and -1. This *power-law region* is independent of molecular weight or polydispersity. The dependence of viscosity η on shear rate $\dot{\gamma}$ is found to be nearly identical to the dependence of the magnitude of the complex viscosity $|\eta^*|$ on frequency ω. This observation is called the *Cox–Merz rule.*

Figure 1 shows steady-state viscosity as a function of shear rate for high-density polyethylene (HDPE, linear chains) and low-density polyethylene (LDPE, branched chains) melts of several molecular weights. The data are taken from Wales (1976). The HDPE samples have polydispersity indices of 4 and 14.5 for the samples with molecular weight 4.4×10^4 and 1.6×10^5, respectively, whereas the LDPE samples have polydispersity indices of 5, 16.7, and 4.62 for the lowest to highest molecular weights, respectively.

The first normal-stress coefficient $\Psi_1 := -(\tau_{xx} - \tau_{yy})/\dot{\gamma}^2$ exhibits a shape similar to the viscosity, but has a much stronger shear-thinning dependence. The second normal-stress coefficient, which is difficult to measure, has sign opposite to that of the first normal-stress coefficient, and approximately one-tenth the magnitude.

Figure 2 shows steady-state first normal-stress coefficient as a function of shear rate for concentrated polystyrene (linear chains) solutions of two molecular weights at a temperature of 25 °C. The data are taken from Kulicke and Wallbaum (1985). The samples have molecular weights of 9.15×10^6 and 3.2×10^6 with polydispersity indices of 1.3 and 1.1, respectively.

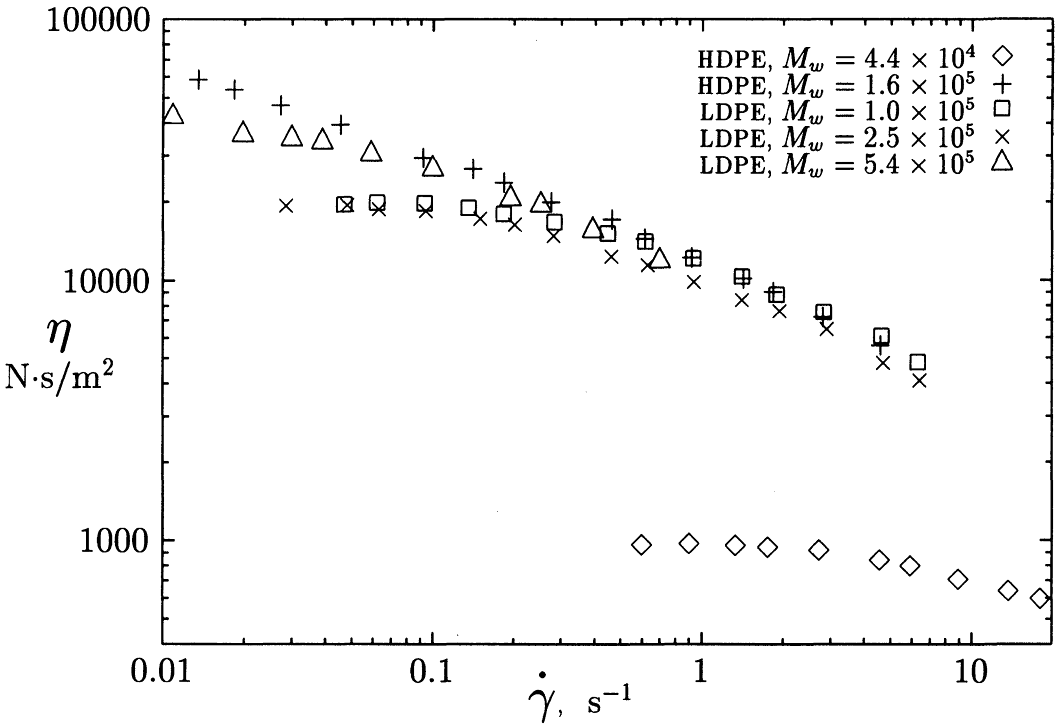

FIG. 1. Viscosity data of Wales (1976) for samples of high-density polyethylene (HDPE) and low-density polyethylene (LDPE) melts of various molecular weights; the polydispersity indices are given in the text. The data were taken using a cone-and-plate rheometer at 150 °C.

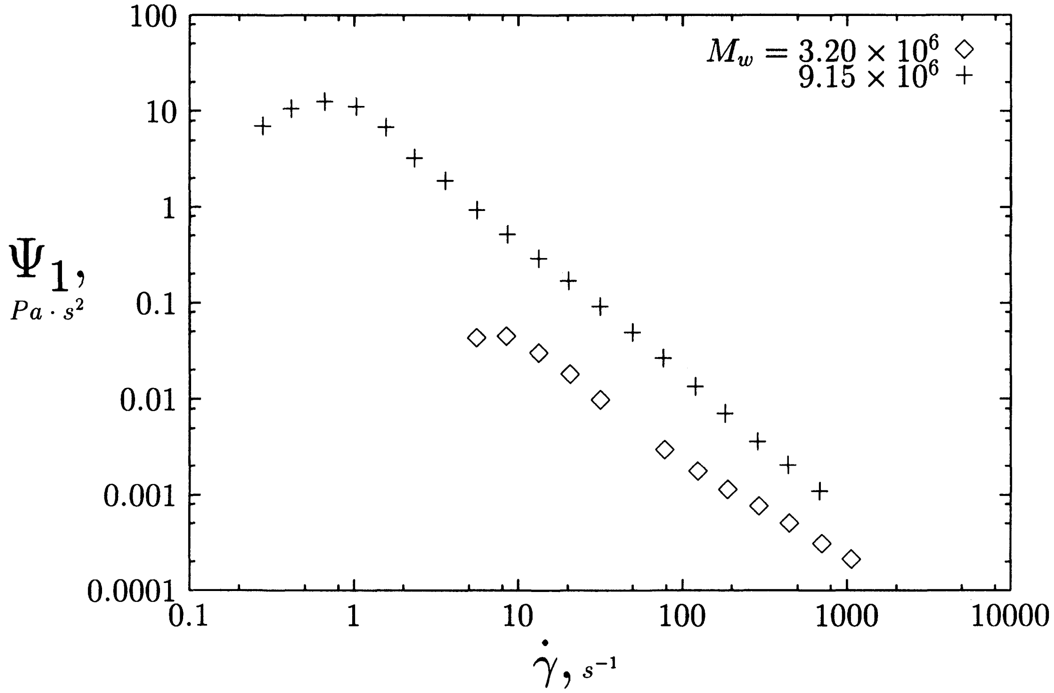

FIG. 2. First normal-stress coefficient data of Kulicke and Wallbaum (1985) for concentrated polystyrene solutions at 25 °C. The samples have molecular weights of 9.15×10^6 and 3.2×10^6 with polydispersity indices of 1.3 and 1.1, respectively.

Dilute polymer solutions again are qualitatively similar, but the steady-state viscosity reaches a second plateau at high shear rates, and at zero shear rate scales as $M^{3\nu}$ ($\nu = \frac{1}{2} \Rightarrow$ theta solvent; $\nu = \frac{3}{5} \Rightarrow$ good solvent).

Before reaching steady state, these material functions can also be studied as transient quantities. Transient material properties can be defined in a manner similar to the ones defined above. We indicate the transient property either by the addition of a superscript "+" to indicate the stress following inception of steady shear, or by a superscript "−" to indicate the stress following the cessation of steady shear flow. For example, $\eta^+(\dot{\gamma},t)$ represents the shear stress divided by the shear rate following the inception of shear rate $\dot{\gamma}$ at $t = 0$.

Both shear-stress and first normal-stress differences can be studied. It is found that both η^+ and Ψ_1^+ go through three different regimes as the shear rate is increased. At lowest shear rates, the stress growth is strictly monotonic, and the curve for the lowest shear rate forms a maximum upper curve for all experiments. At somewhat higher shear rates, an overshoot can be observed. At yet higher shear rates, an overshoot followed by at least one oscillation is observed. The first normal-stress coefficient experiences a relatively higher maximum value. The second normal-stress difference follows the behavior of the first normal-stress difference quite closely for all times.

Similar to the Cox–Merz rule, the *Gleissle mirror relation* relates the steady-state viscosity curve to the transient measurement: $\eta(\dot{\gamma}) = \eta_0^+(t = 1/\dot{\gamma})$.

After steady state has been attained, two different experiments can be performed. The first holds the sample stationary while the resulting stress is measured, whereas the second applies no stress and measures the resulting flow. The first experiment is referred to as *cessation of steady shear flow,* and the second is called *constrained recoil after steady shear flow.* The former experiment on polymer melts reveals that all stresses relax monotonically, but that the transient shear-stress coefficient η^- decays more rapidly than the first normal-stress coefficient Ψ_1^- (Gleissle and Ohl, 1990). Also, the shear and normal stresses decay more rapidly the higher the shear rate $\dot{\gamma}_0$.

In the constant-recoil experiment, after the stresses have reached steady state in the shear flow, the shear stress at the boundaries is turned off, and the sample reverses the direction of flow for some time. Results can be described by the *ultimate recoil,* which is defined as the final total strain achieved by the sample after the shear stress is turned off. The time for total recoil is comparable to the longest relaxation time of the fluid, found either by zero–shear-rate viscosity, or time required to attain steady state. In the linear viscoelastic regime, the ultimate recoil is equal in magnitude to the zero–shear-rate first normal stress divided by the zero–shear-rate shear stress, $\Psi_{1,0}\dot{\gamma}/\eta_0$. At higher shear rates, the value of ultimate recoil for linear low-density polyethylene increases linearly with shear rate, before leveling off at a strain near 3 to become independent of shear rate.

Many shear-flow data can be found to fit master curves that are temperature independent, when the time is rescaled appropriately. The principle is referred to as *time-temperature superposition,* and the temperature- and material-dependent factor necessary for rescaling is called the *time-temperature superposition shift factor,* defined as

$$a_T := \frac{\eta_0(T)T_0\rho(T_0)}{\eta_0(T_0)T\rho(T)}, \tag{25}$$

where T_0 is a reference temperature, η_0 is the zero–shear-rate viscosity, and ρ is the density. The shift factor appears to be rather universal in shifting many different curves. The factor's functional form is often fitted by either an Arrhenius expression or the so-called *WLF equation* (Williams–Landel–Ferry; see p. 274 of Ferry, 1980).

2.4.2 Shear-Free Flows Incompressible shear-free flows are described by the velocity gradient in the appropriate coordinate system as

$$\nabla \mathbf{v}_f = \begin{bmatrix} -\frac{1}{2}(1+b)\dot{\epsilon}_0 & 0 & 0 \\ 0 & -\frac{1}{2}(1-b)\dot{\epsilon}_0 & 0 \\ 0 & 0 & \dot{\epsilon}_0 \end{bmatrix} \mathrm{H}(t), \tag{26}$$

where b is a constant that specifies the type of flow field and $\dot{\epsilon}_0$ is the *elongation rate,* which can be varied from experiment to experiment. For $b = 0$ and $\dot{\epsilon}_0 > 0$, the flow is called *elongational,* and is something like stretching a filament with a rate that increases the length of the filament exponentially with time. For $b = 0$ and $\dot{\epsilon}_0 < 0$, the flow is called *biaxial stretching.* Such a flow is the opposite of elongational, and can be viewed as stretching a sheet equally in both directions, such that the width and length of the sheet increase exponentially with time. When $b = 1$, the flow is called *planar elongational,* and can be achieved by stretching the length of a rectanguloid exponentially with time, while constraining the width to stay constant, and allowing the thickness to decrease.

Important material functions in these experiments are coefficients of the normal-stress differences:

$$\begin{aligned} \bar{\eta}_1 &:= (\tau_{xx} - \tau_{zz})/\dot{\epsilon}_0, \\ \bar{\eta}_2 &:= (\tau_{xx} - \tau_{yy})/\dot{\epsilon}_0. \end{aligned} \tag{27}$$

At least one of these stresses is conveniently found by measuring the forces necessary to pull the sample in the directions described above.

These measurements are notoriously difficult, and have been successfully accomplished in very few laboratories with one very notable exception (Meissner *et al.,* 1982). In fact, a homogeneous flow field described by Eq. (26) has not been successfully achieved for polymer solutions, although some very clever experiments have been designed to impose flow fields with predominantly elongational components. We do not consider them here, although these experiments are very important in understanding the stresses arising in many complex geometries.

Achieving steady state in elongation of polymer melts has also been rather difficult, and there is evidence that steady state is rarely achieved in nonlinear strains. However, at very small strain rates linear viscoelastic theory is useful, which predicts that the steady-state value of $\bar{\eta}_1$ is three times the zero–shear-rate viscosity as $\dot{\epsilon}_0 \to 0$. This prediction is consistent with experiment.

The approach to steady state of the *elongational stress growth function* $\bar{\eta}_1^+$ (transient analog of $\bar{\eta}_1$) at small elongational rates is monotonic, and the curve forms the lower bound for results at higher elongational rates. At higher rates, the elongational viscosity initially follows the linear-viscoelastic curve at short times, but then deviates sharply upward, exhibiting an extreme *strain-hardening* effect. The strain hardening occurs at strains of approximately unity, independent of elongation rate. For sufficiently long measurement times, the coefficient goes through a maximum.

Figure 3 shows $\bar{\eta}_1^+$ data in simple elongation flows as a function of time for a branched low-density polyethylene melt at 150 °C from Zülle *et al.* (1987). The polydispersity index is not reported. However, the sample is industrial, and so presumably has a broad distribution. The different curves are various elongational rates $\dot{\epsilon}_0$.

Transient measurements during other shear-free flows have been accomplished, but are not discussed here.

3. THEORETICAL MODELING

3.1 Dilute-Solution Polymers

3.1.1 Models of Kuhn and Kuhn, Rouse, and Kramers Kinetic-theory models for dilute polymer solutions are mean-field models that treat the solvent as a continuum. However, the discrete segments of the chain are sufficiently small that Brownian forces are important, so that the equations of motion for the polymer segments are stochastic. The approach to making predictions involves two steps: solving for the microscopic dynamics, and making the connection to macroscopic observables. We begin with the former.

The chain is subdivided into discrete segments or "beads" that interact with one another and the solvent. One begins by writing the equation of motion for bead i on the chain ($i = 1,2,\ldots$):

$$m_i \frac{d}{dt}\left[\frac{d}{dt}\mathbf{r}_i - \mathbf{v}_f(\mathbf{r}_i)\right] = \mathbf{F}_i^{(c)} + \mathbf{F}_i^{(h)} + \mathbf{F}_i^{(ev)} + \mathbf{F}_i^{(b)}, \qquad (28)$$

where m_i is the mass and $\mathbf{r}_i$ is the position of bead i, $\mathbf{v}_f(\mathbf{r})$ is the solution velocity at $\mathbf{r}$, and the terms on the right-hand side of Eq. (28) are the various forces that act on the bead. The term $\mathbf{F}_i^{(c)}$ represents the "connector forces" between beads that are near one another along the contour of the chain. $\mathbf{F}_i^{(h)}$ is

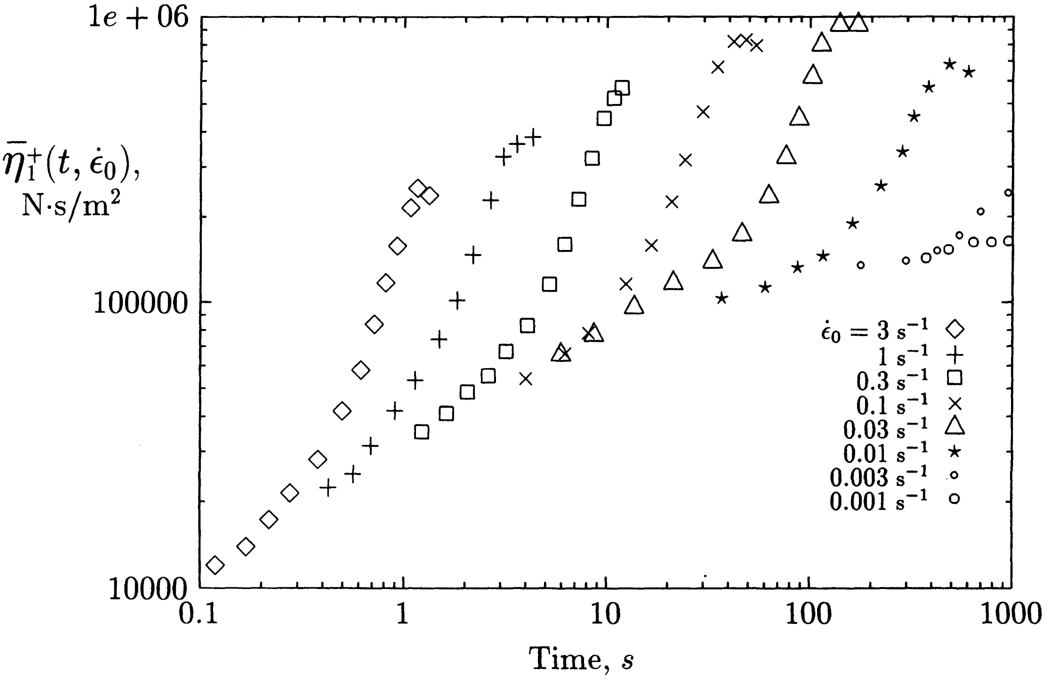

FIG. 3. Simple elongational stress data $\bar{\eta}_1^+$ as a function of time for branched polyethylene melts at 150 °C from Zülle *et al.* (1987). The different curves are various elongational rates $\dot{\epsilon}_0$.

the "hydrodynamic force," which represents the drag force exerted by the solvent on the bead. $\mathbf{F}_i^{(ev)}$ is the "excluded-volume" force that accounts for interactions between beads that may be far apart along the contour of the chain, but are near in space, and hence prevents two beads from occupying the same space. Such an effect can also be used to prevent different segments of the chain from passing through each other, or to account for the effects of a good solvent. $\mathbf{F}_i^{(b)}$ is the random, fluctuating Brownian force on the bead, whose form is fixed by statistical-mechanical considerations.

The simplest model that incorporates some of these ideas is the model of Kuhn and Kuhn (1945). Kuhn and Kuhn began with a thread model and considered only the end-to-end vector of the thread. The model is mathematically equivalent to a dumbbell model that neglects the inertia of the beads, so that the left-hand side of Eq. (28) may be set equal to zero. Only two beads are used, so that $i = 1, 2$. Excluded volume is neglected, and so we assume $\mathbf{F}_i^{(ev)} = \mathbf{0}$. The hydrodynamic force that accounts for bead–solvent interaction is assumed to be a simple Stokes's law drag with coefficient ζ,

$$\mathbf{F}_i^{(h)} = -\zeta\left[\frac{d}{dt}\mathbf{r}_i - \mathbf{v}_f(\mathbf{r}_i)\right]. \tag{29}$$

The connector force is assumed to be linear in displacement, with constant H, which is found from statistical-mechanical arguments in Sec. 3.1.2:

$$\mathbf{F}_1^{(c)} = -\mathbf{F}_2^{(c)} = H(\mathbf{r}_2 - \mathbf{r}_1). \tag{30}$$

In general, the Brownian force must be chosen to satisfy the fluctuation–dissipation theorem (Berne, 1977), which provides a relation between the Brownian force and the dissipative polymer–solvent forces. When the dissipative forces depend only upon the instantaneous velocities of the beads (i.e., when they have no "memory"), satisfaction of the equipartition of energy is usually sufficient to satisfy the fluctuation–dissipation theorem. For the Kuhn–Kuhn model, the Brownian forces are required to have an autocorrelation function that is delta-correlated in time, has no cross correlations among the different components, and has magnitude $2kT\zeta$:

$$\langle \mathbf{F}_i^{(b)}(t)\mathbf{F}_j^{(b)}(t')\rangle = 2kT\zeta\delta_{ij}\delta(t - t')\boldsymbol{\delta}, \tag{31}$$

where δ_{ij} is the Kronecker delta and $\delta(t)$ is the Dirac delta function. It is convenient to introduce the "connector vectors" to describe the orientation of the chain, $\mathbf{Q}_i := \mathbf{r}_{i+1} - \mathbf{r}_i$. For the dumbbell, a single vector is sufficient, $\mathbf{Q} \equiv \mathbf{Q}_1$. The parameters introduced give rise to a single time constant $\lambda_H := \zeta/4H$, and a single length scale $\sqrt{kT/H}$. We assume that the local velocity gradient is not large on the length scale of the size of the polymer, so that the difference in the solution velocities at the two beads can be determined from the first spatial derivative of $\mathbf{v}_f$. Using the above definitions and simplifications, and solving explicitly for $\dot{\mathbf{Q}}$, the equation of motion for a Hookean dumbbell model becomes a Langevin equation. Just as the dynamics of a Brownian particle can be explained by both a Langevin equation and Einstein's diffusion equation, the dynamics of the Hookean dumbbell can equivalently be described by a Fokker–Planck equation. Using standard techniques (Gardiner, 1985), the Fokker–Planck equation equivalent to the Langevin equation is found to be

$$\frac{\partial\psi(\mathbf{Q},t)}{\partial t} = -\frac{\partial}{\partial\mathbf{Q}} \cdot\left\{\mathbf{Q}\cdot\nabla\mathbf{v}_f - \frac{1}{2\lambda_H}\mathbf{Q} - \frac{kT}{2H\lambda_H}\frac{\partial}{\partial\mathbf{Q}}\right\}\psi, \tag{32}$$

where $\psi(\mathbf{Q},t)$ is the polymer end-to-end vector probability distribution. Note that Eq. (32) implies that the stress tensor is a functional of the history of the velocity gradient $\nabla\mathbf{v}_f$. Since the model is frame invariant (Schieber and Öttinger, 1994), it is a specific case of a viscoelastic fluid as proposed by VanArsdale (1995). In fact, nearly all kinetic-theory models to date, except liquid-crystalline polymers that can have preferred direction, are examples of such fluids.

When one measures macroscopic properties of polymer solutions, one is typically measuring certain moments of $\mathbf{Q}$. For example, light scattering measures the radius of gyration, which is proportional to $\langle Q^2\rangle = \mathrm{tr}\langle\mathbf{QQ}\rangle$. To find moments of $\mathbf{Q}$, one can take several different approaches for a given Langevin or Fokker–Planck equation. The most difficult to perform, but conceptually easiest to see, is first to solve the Fokker–Planck equation for ψ, and take the necessary integration:

$$\langle f(\mathbf{Q})\rangle = \int f(\mathbf{Q})\psi(\mathbf{Q},t)d^3Q, \tag{33}$$

where f is an arbitrary function of **Q**. The solution to Eq. (32) is a multivariate Gaussian distribution of zero moments. The second moment is more easily found by the method described next.

Alternatively, in the case where the Fokker–Planck equation is linear in **Q**, one can use Eq. (32) to generate a series of ordinary differential equations in the averages. For example, multiplying Eq. (32) by the dyadic product **QQ**, integrating both sides over all possible configurations of **Q**, and using integration by parts for the integrals on the right-hand side, one obtains

$$\begin{aligned}\frac{d}{dt}\langle\mathbf{QQ}\rangle + (\nabla\mathbf{v}_f)^{\dagger}\cdot\langle\mathbf{QQ}\rangle + \langle\mathbf{QQ}\rangle\cdot\nabla\mathbf{v}_f \\ = -\frac{1}{\lambda_{\mathrm{H}}}\langle\mathbf{QQ}\rangle + \frac{kT}{H\lambda_{\mathrm{H}}}\boldsymbol{\delta}.\end{aligned} \tag{34}$$

For any given flow field, Eq. (34) has a formal solution that contains an exponential matrix involving the integral of the velocity gradient.

If the Fokker–Planck equation derived for some model is nonlinear in **Q**, the above methods cannot be used to find an ordinary differential equation for $\langle\mathbf{QQ}\rangle$. If one attempted to create an ordinary differential equation involving the second moment of **Q**, higher moments of **Q** would arise in the equation. Attempts to write equations of the higher moments result in yet higher moments, and a closed set of differential equations would never result. When this impasse occurs, either one resorts to approximations, or the "exact" solution can be found numerically. Using a pseudorandom-number generator, one can generate stochastic trajectories of a discretized version of the Langevin equation for an ensemble of dumbbells. Then, at any point in the simulation, any desired average can be estimated by taking an ensemble average of this finite-sized ensemble of dumbbells. This *Brownian dynamics* method is used to find "exact" numerical results of many of the models discussed in the next section. A detailed discussion of stochastic simulations of polymer dynamics is given by Öttinger (1995).

To make a connection between the proposed mechanical model and measurements of stresses, one needs an expression for the extra stress tensor $\boldsymbol{\tau}$ and averages of the configuration of the chain. Statistical-mechanical arguments lead to the Kramers–Kirkwood expression for the polymer contribution to the extra stress,

$$\boldsymbol{\tau}_p = -n\sum_i \langle\mathbf{F}_i^{(\mathrm{h})}\mathbf{r}_i\rangle, \tag{35}$$

where n is the number density of polymer chains. The Kramers–Kirkwood expression has been found to be "thermodynamically consistent" for a large number of models and appears to be very generally true (Jongschaap, 1991). In this context, "thermodynamically consistent" means that the expression leads to a positive dissipation of mechanical energy in all flow fields. The Kramers–Kirkwood expression, Eq. (35), simplifies to the less general Kramers expression, Eq. (6), for the dumbbell considered here.

In contrast to experiment, the Hookean dumbbell leads to shear-rate–independent viscosity, no overshoot in the transient viscosity, and a length that can stretch without bounds in strong flows. On the other hand, the model does yield a first normal-stress coefficient, albeit shear-rate independent, and a qualitatively reasonable-looking complex viscosity in small-amplitude oscillatory shear flow.

The Hookean dumbbell predicts that the average orientation angle of the polymer coil in shear flow is $\alpha = \frac{1}{2}\tan^{-1}(m_0/\beta)$ with $m_0 = 1$. The radius of gyration is predicted to go as $1 + 2\beta^2/3m_0$. These predictions can be compared with the empirical results in Eqs. (15) and (16). The orientation predictions are qualitatively reasonable at low shear rates, but the predictions of coil stretching at high shear rates are overestimates.

An early attempt to improve this simple model is to give the chain additional degrees of freedom to allow more complex orientations than a single vector. The Rouse model allows the number of beads to be arbitrarily large, but maintains the other approximations. Such a generalization can be shown to be mathematically equivalent to using an equimolar mixture of dumbbells with different time constants. Instead of a single time constant, there is a discrete spectrum of $N_b - 1$ time constants for N_b beads,

$$\lambda_i = \zeta/8H \sin^2(i\pi/2N_b), \quad i = 1, 2, \cdots, N_b - 1. \tag{36}$$

Rouse's generalization leads to an improved complex viscosity prediction, and improves the quantitative fit for the coil stretching and orientation in shear ($m_0 = \frac{5}{2}$), but does not correct the other drawbacks of the Hookean dumbbell. Also, the Rouse model predicts that the zero–shear-rate viscosity scales as $\sim M^2$, and the diffusion coefficient scales as $\sim 1/M^2$. Both of these results are inconsistent with the experimental work, for which the scaling exponents were found to be $\frac{3}{2}$ and $\frac{1}{2}$, respectively. The model is intended for theta solvents, since no excluded-volume effects are included. Interestingly, the Rouse model describes many dynamic-equilibrium results for melts below the entanglement, molecular weight. It is believed that short-chained, nonentangled polymer melts behave like a Rouse model because of screening effects in the matrix.

In the next section, other improvements to the Hookean dumbbell are discussed. These other models yield nonlinear Fokker–Planck equations and require either mathematical approximations or numerical techniques. Generalizations that are not discussed below, but have been considered and found unimportant are inclusion of bead inertia (Schieber and Öttinger, 1988), and accounting for solvent inertia in the hydrodynamic force (Szu and Hermanns, 1974; Schieber, 1991).

Instead of using springs, Kramers (1944) and Kirkwood and Riseman (Kirkwood, 1967) have used models involving rigid rods as connectors. Kramers's connectors are freely jointed bead-rod chains, whereas Kirkwood and Riseman considered a chain in which each connector is constrained to a valence cone about the previous rod orientation. Brownian-dynamics simulation results for Kramers chain were found by Liu (1989), to show that the model does predict shear thinning for both the viscosity and the first normal-stress coefficient. The power-law region has slope of approximately $\frac{2}{3}$ for two beads down to about $\frac{1}{2}$ for 20 beads in the viscosity, and a second plateau region is predicted, all in reasonable agreement with experiment. However, the zero–shear-rate viscosity is predicted to scale with molecular weight squared.

3.1.2 Generalizations and Improvements It is generally believed that the above models are insufficient to describe data for several reasons:

1. neglect of hydrodynamic interaction;
2. infinite extensibility of springs; and
3. neglect of excluded volume.

It is also speculated that internal friction (or internal viscosity) may play an important role.

The motion of one bead on the chain will perturb the flow field felt by all other beads in the chain, giving rise to *hydrodynamic interaction* between the segments. The interaction can be accounted for by either the Rotne–Prager–Yamakawa or the Oseen–Burgers tensors (see p. 133 of Bird *et al.*, 1987), which yield the perturbation to the flow field at one point in the continuum given a hydrodynamic force at a second point. Since the effect is nonlinear in configuration, most of the work has involved mathematical approximations of the tensor.

The earliest work was by Zimm (see p. 94 of Bird *et al.*, 1987), who replaced the configuration-dependent tensor by an average taken over the equilibrium configurations of the chains. This approximation was sufficient to correct the prediction of the zero–shear-rate viscosity scaling dependence on molecular weight. Also, Zimm's approximation predicts that the longest relaxation time goes as $M\eta_{p,0}/4.04RT$, which agreed well with the subsequent experiments of Caroline and Jones (1980) given in Sec. 2.1.1. However, no other improvements are offered by Zimm's approximation.

Subsequent approximations include replacing the equilibrium-averaged hydrodynamic-interaction tensor with its consistently averaged (or iteratively averaged) values. These approximations led to improved viscosity and first normal-stress coefficient predictions. These latter approximations were then replaced by making a Gaussian closure approximation on the second-moment equation arising from the Fokker–Planck equation. The earlier approximations cannot account for fluctuations in the configurations of the chains, whereas the Gaussian closure can. This last approximation has subsequently been verified by exact Brownian-dynamics simulations by Zylka (1991) to be valid. The simulations predict that the viscosity and first normal-stress coefficient show shear thinning at moderate shear rates followed by shear thickening at high shear rates. The second normal stress has opposite sign to that of the first normal-stress coefficient at small and moderate shear rates, and then changes sign at high shear rates.

Finally, both hydrodynamic interaction and excluded volume were considered by Zylka and Öttinger (1991). Both the hydrodynamic-interaction and excluded-volume effects are accounted for as perturbations to the Rouse model. Since these effects scale with chain length to the one-half power, renormalization-group techniques are needed to find asymptotic, or universal, behavior for long chains. The excluded-volume effect is used to find predictions for dilute solutions in good solvents, and is switched off to find predictions for theta solvents. Zylka and Öttinger find the diffusion tensor for the chain in shear flow, and steady-state viscometric functions. The predictions are qualitatively unchanged from the Brownian-dynamics results for the theta-solvent conditions. However, the results do show that the viscosity grows without bounds with increasing shear rate.

The classic textbook by Yamakawa (1971) considers many dynamic-equilibrium properties of polymer solutions using a two-parameter model: the Kuhn step length, and a binary cluster parameter for excluded volume. The book also describes Fixman's use of boson operators to approximate excluded-volume interactions, and makes many comparisons with experimental measurements from dynamic light scattering.

Three results indicate that the models discussed so far (using Hookean springs) predict an overstretching of the polymer coil in flows: The coil size is found by light scattering to be much smaller than predicted, the coil stretches to infinite length for elongational rates above a critical value, and the hydrodynamic-interaction model of Zylka predicts unbounded shear thickening. These results suggest that a realistic model must include finitely extensible connectors.

The spring force arises from a coarse-graining of the random-walk chain. Derivation of the precise form of the potential is as follows. One begins by modeling the chain segment as a random walk of $N - 1$ Kuhn steps of length a. The probability of the end-to-end vector of the segment is found from a straightforward probability transformation. From the end-to-end vector distribution, the partition function for the segment is found from Gibbs's postulate. Once the partition function is available, all thermodynamic quantities can be found, including the Helmholtz free energy. Now, if one isothermally increases the length of the segment, the change in free energy is found from the thermodynamic relation just derived. This change in free energy must be the physical work done on the chain by pulling. From the change in length and the change in free energy, the force can be inferred. In this way, the force is found as a function of segment length:

$$\coth\left(\frac{aF^{(c)}(\mathbf{Q})}{kT}\right) - \frac{kT}{aF^{(c)}(\mathbf{Q})} = \frac{Q}{(N-1)a}. \qquad (37)$$

Such an expression is unwieldy, and so two approximations are generally used. The simplest is the linearization in small Q, which leads to the Hookean spring law $\mathbf{F}^{(c)}(\mathbf{Q}) = -H\mathbf{Q}$, for which the spring constant can now be identified as $3kT/a^2(N-1)$. A better approximation that is nonlinear in Q and prevents the spring from stretching farther than $(N-1)a$ is called the FENE (finitely extensible nonlinear elastic) spring:

$$\mathbf{F}^{(c)}(\mathbf{Q}) = -3kT(N-1)\mathbf{Q}/[(N-1)^2a^2 - Q^2]. \qquad (38)$$

The FENE law approximates Eq. (37) rather well and captures the important features. Most of the calculations for this model have involved a consistent-averaging approximation, whereby one replaces the Q^2 in the FENE expression with its flow-dependent average value. In other words, the average length of each segment is constrained to be less than a. Exact calculations (Rudisill and Cummings, 1992; Hua and Schieber, 1994) have confirmed the qualitative features of the approximation. Namely, shear thinning in viscosity and first normal-stress coefficient is predicted, and the stresses remain bounded in strong flows, such as steady elongation.

Also, finite extensibility reduces the amount of stretching of the coil, bringing the predictions in closer agreement with observation.

The above derivation for the connector law is derived from thermodynamic arguments, which assume a quasi steady-state stretching of the segment. Physically, one might expect that rapid deformations of the chain encounter an additional, rate-dependent, dissipative force. This additional *internal viscosity* was first proposed by Kuhn and Kuhn (1945). In addition to the spring force, one postulates an internal-viscosity force

$$\mathbf{F}^{(iv)}(\mathbf{Q}) = -\frac{\epsilon\zeta}{2}\frac{\mathbf{QQ}}{Q^2}\cdot\frac{d}{dt}\mathbf{Q}, \tag{39}$$

where ϵ is called the *internal-viscosity parameter,* whose value determines the strength of the internal viscous drag relative to the solvent drag.

Again, most calculations have been made using an approximation, such as perturbation for small ϵ, small strain rates, linearization of Eq. (39) in $\mathbf{Q}$, or Gaussian closure of the second-moment equation. However, Brownian-dynamics simulation (Hua and Schieber, 1994; Wedgewood, 1994) shows that these approximations break down for all but small values of ϵ, and often are valid only for small strain rates. The Gaussian closure is a good approximation for small ϵ at all flows, small strain rates for all values of ϵ, or short-time transients for large ϵ at large strain rates. For $\epsilon > 3$, the Brownian-dynamics simulations appear to be best approximated by rigid connectors, with a Gaussian distribution of lengths. The internal-viscosity effect is able to describe positive values for the second plateau region in the in-phase part of the complex viscosity, and predicts shear thinning at moderate shear rates. However, at large shear rates, shear thickening of viscosity is predicted for Hookean connectors, which is similar to hydrodynamic-interaction effects. The internal-viscosity effect also decreases the amount of predicted coil stretching in shear flow.

Addition of internal viscosity does predict one effect that is distinct from others. Namely, it predicts both an elastic and a viscous contribution to the stress during flow. The viscous stress contributes nothing when the flow field is switched off, whereas the elastic stress persists for some time, before decaying to zero. The other effects (e.g., FENE connectors, hydrodynamic interaction) lead to a purely elastic polymer contribution to the stress. Recent experiments of Liang and Mackay (1993) have shown that semidilute solutions of xanthan do have apparent elastic and viscous polymer contributions to the stress. However, these results could conceivably be explained by a change in solvent characteristics from the presence of the polymer, or polymer–polymer interactions.

Many different combinations of the above effects have been tried: multiple connectors, nonlinear springs, excluded volume, hydrodynamic interaction, and internal viscosity. Each of the effects has been able to explain some aspect of the data, and points to the importance of the effect. To incorporate all of these effects, one must specify two parameters that are not directly connected to the chemical structure of the polymer: bead size and ϵ. In principle, the other parameters follow from these and the Kuhn step length a, which can be estimated from the chemical structure of the chain. However, some workers choose to leave some of these as independent adjustable parameters.

3.2 Concentrated Solutions and Undiluted Polymers

3.2.1 Reptation Models The term *reptation* was coined by de Gennes (1979), who was the first to calculate quantitative predictions for the dynamics of polymer chains trapped in a gel. The ideas were later generalized to polymer melts by Doi and Edwards (1986), and then expanded by many other workers. One begins with a mean-field model, in which the surrounding matrix of chains in the gel (or melt) is replaced by a "tube" that constrains the motions of the probe chain. The chain is assumed to move rather easily along its contour, but cannot move very far laterally before "feeling" the constraints of the surrounding matrix. One then considers a *primitive chain,* which is meant to be a sort of coarse-graining of the real (probe) chain. The primitive chain is allowed to move in a coordinated fashion such that it can slink along its one-dimensional contour inside the tube. When the end of the chain exits either end of the tube, it takes a random orientation. The tube is defined by the presence of the chain, and so the tube segment at the receding end of the chain is destroyed. It is assumed that the reptative motion is driven by one-dimensional Brownian motion, and that the friction between the tube and the chain is proportional to the length of the chain.

These straightforward ideas were sufficient to allow de Gennes to predict scaling dependence of the diffusion coefficient and the longest relaxation time constant on chain length, or molecular weight for a chain in a gel. De Gennes found

$$D \sim M^{-2}, \quad \lambda_{\mathrm{R}} \sim M^{3}, \tag{40}$$

where λ_{R} is the longest relaxation time constant. Reptation was the first model to predict the correct diffusion-scaling exponent for many systems. The exponent describing the time constant does not agree with the value of 3.4 found experimentally, and the fact that the diffusion scaling predicted by reptation holds well below M_e also raises questions. However, these predictions were sufficient to generate a great deal of interest in reptation.

Predictions of the reptation model in flowing polymer melts require additional assumptions. First, the chain is assumed to relax by reptation and the additional mechanism of *chain retraction.* For example, in single-step strain, the following dynamics is assumed for the chain. First, during the imposition of the strain, the segments are deformed affinely. One can show that most equilibrium configurations of the chain are stretched after the strain. Then, on a rapid time scale, the chains are assumed to retract back to their equilibrium lengths; this motion is the first relaxation mechanism. The second relaxation mechanism is reptation of the chain to a new configuration. In other flow fields, it is generally assumed that the retraction mechanism can be treated as essentially instantaneous.

For tractability, additional mathematical approximations, such as *independent alignment,* or consistent averaging, are usually employed to make predictions of the dynamics of the chain. Independent alignment is the most severe of the approximations, but allows solution to a very large number of problems. This approximation assumes that the retraction of the entire chain can be replaced by a retraction of each of the segments of the chain independently. Then, the dynamics of the segments are decoupled, and the following Fokker–Planck equation is obtained:

$$\frac{\partial f(\mathbf{u},s,t)}{\partial t} = \frac{kT}{N\zeta}\frac{\partial^2 f}{\partial s^2} - \frac{\partial}{\partial \mathbf{u}}\cdot[\mathbf{u}\cdot\nabla\mathbf{v}_f - \nabla\mathbf{v}_f : \mathbf{uuu}]f, \tag{41}$$

where ζ is the effective friction factor felt by a segment in the tube, $\mathbf{u}$ is the unit orientation vector of a tube segment, s is the position of the tube segment relative to the chain end, and f is the probability density of finding a tube segment with orientation $\mathbf{u}$, and containing chain segment s along the chain contour. Note that since the chain diffuses randomly within the tube, s is treated as a stochastic variable that can take values between $\pm Na/2$. The boundary conditions associated with Eq. (41) couple the orientation $\mathbf{u}$ with position along the chain s:

$$f(\mathbf{u},s,t) = 1/4\pi, \quad \text{for } s = \pm Na/2. \tag{42}$$

Equation (41) has a general integral solution for f (Bird *et al.*, 1987, p. 330).

To be consistent with the rapid retraction mechanism, the tension in the chain is assumed to be the equilibrium value, independent of flow field. Thus, Doi and Edwards (1986) assume that the stress tensor is just

$$\boldsymbol{\tau} = \frac{G_e}{3}\boldsymbol{\delta} - \frac{G_e}{Na}\int_{-Na/2}^{Na/2}\int \mathbf{uu}f(\mathbf{u},s,t)\,d^2u\,ds, \tag{43}$$

where G_e is the plateau modulus and is a function of segment length a, number of chain segments N, chain concentration, and temperature.

Since the reptation model requires mathematical approximations to find quantitative predictions, the approximations have sometimes obscured the conclusions that can be drawn from quantitative comparisons with data. However, some results are conclusive, and many of the conclusions are compiled by Lodge *et al.* (1990) and by Skolnick and Kolinski (1990).

The model is able to capture qualitatively the features of the relaxation modulus, although the spectrum is too narrow. These discrepancies can be improved by accounting for polydispersity, but the general trend of relative weight for the time constants in the spectrum is inconsistent with experiments by Baumgärtel *et al.* (1990) discussed in Sec. 2.4.1.

The damping function, which arises from the retraction mechanism, is consistently underpredicted as a function of strain for linear chains. Reptation with a consistent-averaging approximation for the retraction mechanism predicts approximately that $a_d = 39/175$ in Eq. (22). Nonetheless, since the damping function contains no adjustable parameters, the comparison is impressive. After the relaxation modulus has been "corrected" (fitted to experiment), double-step strain predictions of the Doi–Edwards model are improved. However, if the second step strain is imposed before chain retraction is complete, current mathematical models are unable to describe the data (Venerus and Kohvand, 1994).

When independent alignment is assumed, the theory predicts shear thinning for both the viscosity and the first normal-stress coefficient, but the power-law region is predicted to have slopes that are too steep. Part of the discrepancy can be attributed to polydispersity in the sample. To date, there are few quantitative predictions for the reptation model that account for polydispersity (Schieber, 1987). Stress overshoot is predicted for the viscosity, but not for the first normal-stress coefficient. Elongational viscosity is also not well described by the model.

Many comparisons with data have indicated that relaxation processes in addition to reptation are needed to describe the data (Saab *et al.*, 1982). Subsequently, additional chain motions have been considered in a phenomenological manner, such as "tube renewal," "constraint release," or "tube dilation." Most of these additional mechanisms are incorporated to account for an important problem in the original formulation. Namely, the surrounding matrix of chains in a polymer melt is relaxing on the same time scale as the probe chain. Hence, it is not consistent to treat the tube as a stationary object when considering the dynamics of the chain. Some experimental results indicate that the diffusion coefficient of a shorter chain is independent of the molecular weight of the matrix for matrix chains sufficiently longer than the probe chain, as discussed in Sec. 2.2.

One of the more successful generalizations to the reptation idea is the so-called *double reptation* of des Cloizeaux (1990), to account for relaxation coupling between different chains in an *ad hoc* manner. He postulated that, since a chain segment can relax by two processes—namely its own reptation, or the relaxation of a surrounding chain—the relaxation modulus of the mixture should be the root mean square of the individual components. Apparently, the resulting mixing rules for the relaxation modulus of polydisperse systems are able to describe experimental results quantitatively, and are currently being used to estimate polydispersity distributions from complex viscosity data (Mead, 1994).

Additional models have been proposed that incorporate some form of reptation in the dynamics. For example, the work of Hess (1986, 1988) attempts to derive reptation dynamics from the Liouville equation using projection operators.

3.2.2 Anisotropic Mobility Models Additional mean-field models have been proposed for polymer melts that do not assume reptation. A few of these have incorporated anisotropic mobility for the chain segments in order to account for the restriction to mobility of chains in melts. The model of Curtiss and Bird (Bird *et al.*, 1987) assumes that it is easier for a segment of chain to move in the direction of the orientation of segments adjacent along the chain. They also assume an anisotropic Brownian motion of the chain, and that only chains that are mildly curved contribute to the stress tensor. With these assumptions, they arrive at the same Fokker–Planck equation (within a constant) as the Doi–Edwards model, Eq. (41), but an expression for the stress tensor that involves an additional adjustable parameter called the *link tension coefficient*. Within a constant, all of the results of the Doi–Edwards model are recovered when the link tension coefficient is set equal to zero.

Nonzero values of the link tension coefficient worsen the predictions of the relaxation modulus, but improve predictions of the viscosity, first normal-stress coefficient, and elastic recovery (Saab *et al.*, 1982). Also, many calculations have been made that account for polydispersity in the model, which generally compare quite well with experimental data (Schieber, 1987). However, the model has been criticized for failure to predict the stress-optic law. Also, the mild-curvature approximation does not appear to be consistent with other assumptions in the model.

An additional model containing the concept of anisotropic mobility is the constitutive equation of Giesekus (Bird *et al.*, 1987). This approach models the polymer coil as a dumbbell that can stretch more easily than it can rotate. To achieve mathematical closure in the second-moment equation of the orientation vector, a consistent-averaging approximation is made on the mobility tensor. For a given range of the adjustable parameters, reasonable qualitative results are obtained for many nonlinear experiments.

3.2.3 Mode–Mode Coupling Models The stochastic equations used for dilute-solution polymer modeling can be derived from Hamilton's equations of motion using Mori–Zwanzig projection operators (Berne, 1977). However, validity of the derivation relies on a separation of time scales between the motion of the polymers and the motion of the solvent particles. For polymer melts, such a separation is clearly not valid, as all polymer chains relax on approximately the same time scale. To account approximately for the coupling between the different modes relaxing on various time scales, it is possible to use projection operators to project the dynamics onto higher moments of the quantities of interest. An additional closure must be made on these higher moments, but it is hoped that capturing the couplings between different modes even roughly is sufficient to make good predictions. Such an approach is called mode–mode coupling (Berne, 1977).

Mode–mode coupling results in generalized Langevin equations to describe the stochastic equation of motion of the chains. The memory kernel in the history-dependent drag force on the chain depends, in principle, on the underlying interparticle potential. However, for mathematical tractability, approximations must usually be made. By considering Rouse chains that are not allowed to cross, Schweizer (1989) has considered perturbative approximations of the memory kernel to estimate some dynamic-equilibrium properties. With these approximations, the diffusion coefficients for linear chains and ideal rings are predicted to scale inversely with the molecular weight squared, whereas the plateau modulus and zero-shear viscosity are predicted to scale with molecular weight as M^3. Alternative approximations for the memory kernels lead to a scaling for viscosity more in line with experiment. Using the mode–mode coupling framework, Schweizer is able to predict many of the scaling results of reptation, without introducing any dominant form of chain motion. In an appendix, Schweizer also critically examines and compares the work of Hess.

3.2.4 Network Models Thermodynamic and kinetic-theory models of elastic rubbers have been generalized to describe polymer melts. Instead of chemical bonds that create permanent junctions between polymer strands, mechanical entanglements are assumed to provide temporary cross-link junctions. The lifetime of a given junction is modeled as a stochastic process, so that the existence of entangled strands can be treated as a birth–death process. In real systems, the destruction of an entanglement should result in the motion of nearby segments, as stress is relaxed. However, most network theories decouple the motion of segments, so that the time evolution of the probability function of a single segment may be written. If the network strands between entanglement points are assumed to deform affinely (as can be shown for a permanently cross-linked network), the creation rate and orientation of strands are assumed to be the same as at equilibrium, and a constant destruction rate is assumed, the Lodge rubberlike liquid model arises, with "convection equation" (Lodge, 1964)

$$\frac{\partial \Psi_{iN}}{\partial t} = -\left(\frac{\partial}{\partial \mathbf{Q}} \cdot [\mathbf{Q} \cdot \nabla \mathbf{v}_f] \Psi_{iN}\right) + L_{iN} \Psi_{iN,\mathrm{eq}}(\mathbf{Q}) - \frac{1}{\lambda_{iN}} \Psi_{iN}, \qquad (44)$$

where L_{iN} is the (constant) creation rate of strands and λ_{iN} is the (constant) destruction rate of strands. The distribution Ψ_{iN} ($\Psi_{iN,\mathrm{eq}}$) is the probability-density function (at equilibrium) for the configuration of entangled polymer strands. The subscript iN allows for the existence of different types and lengths of strands. The equilibrium distribution is assumed to be Maxwellian, since each strand is assumed to have a spring force associated with it. These distributions are normalized to the strand number density, rather than 1.

One also generally assumes that the proper form of the stress tensor is the Kramers form, Eq. (6), except that a sum over each type of strand must be made, and the number density of strands is already implicit in the distribution function.

The Lodge model assumes that the appropriate connector force is Hookean. The resulting constitutive equation is very similar to the Rouse model. However, the Lodge network model allows for the relaxation modulus to be a sum of exponentials with any desired spectrum of relaxation times by selecting the proper values for the L_{iN} and λ_{iN}. Thus, many of the shortcomings of the Rouse model in predicting nonlinear properties also apply for the Lodge model.

Many generalizations have been made to this straightforward model. Most of these involve nonlinear deformation of the strands, or time-dependent or conformation-dependent destruction or creation of strands. Since these generalizations make the time-evolution equation nonlinear in $\mathbf{Q}$, consistent-averaging approximations are often made. Stochastic simulations have been used by Biller and Petruccione (1990) to find exact numerical solutions to the above types of generalizations. Van den Brule and Hoogerbrugge (1994) have simulated networks of many connected strands using a stochastic dynamic simulation of birth–death processes. The simulations predict average affine motion of the strands, but shear thickening of viscosity and first normal-stress coefficient, in contradiction with experiment.

3.2.5 Poisson-Bracket Formulations Poisson brackets were developed to write deterministic equations of motion on a vector space utilizing a generator function, such as the Hamiltonian or energy, rather than from a mechanistic approach. In the mechanistic approach above, many degrees of freedom are not accounted for explicitly in the equations of motion, but rather their influence is felt through the addition of stochastic (Brownian force) and dissipative (Stokes drag) terms. When one works from Poisson brackets, the influence of the missing degrees of freedom is described through an additional "dissipative" bracket. Working with such generalized Poisson brackets has the advantage that consistency relations, such as fluctuation–dissipation relations, or thermodynamics, are straightforward to satisfy. However, a mechanistic approach often has an intuitive advantage in modeling.

In theory, the Poisson-bracket approach can be written on many levels of description, from Hamilton's equations of motion for all degrees of freedom to the phase-space configuration distribution function of a single chain, to the second moment of the connector vector of a chain segment (*conformation tensor*). In practice, however, nearly all predictions have been accomplished on the conformation-tensor level. At this level, much information is gleaned from the mechanistic approach, which makes the Poisson-bracket formulation complementary to kinetic theory. A detailed discussion of the method is described in the monograph by Beris and Edwards (1994), where many applications are made to liquid-crystalline systems (among others).

3.2.6 Computer Simulations A novel approach to model hydrodynamics in complex geometries and flows has recently been proposed (Hoogerbrugge and Koelman, 1992; Koelman and Hoogerbrugge, 1993). The method is based on modeling the fluid as a collection of fluid "particles," and, in principle, is capable of handling an extremely broad array of problems, including porous walls. These systems can be dynamic equilibrium or nonequilibrium. Dissipative particle dynamics (DPD) simulates a large number of interacting particles in a flow field by using crude, short-range interparticle potentials, in addition to some fluctuating random forces. The time steps are kept large, and only particles near in space interact. A particle can represent a "packet of fluid," a segment of polymer chain, a portion of solid particle, or a portion of solid wall. The algorithm is inspired by cellular automata programs, and shares some of the same features.

This novel approach has several advantages over existing methods (Kremer, 1983; Kremer *et al.*, 1988; Kremer and Grest, 1990). The time steps are much larger than the time steps of molecular dynamics, and, since far fewer particles interact, the required CPU time for each time step is much less. Since random forces are added, no thermostatting of temperature is required. Thus, macroscopic time scales are attainable with DPD.

Dissipative particle dynamics also exploits the universal characteristic of the hydrodynamics more directly. DPD utilizes simplified potentials for many interacting degrees of freedom. Since chemistry does not appear to be important for the observed universal macroscopic properties of supramolecular systems, DPD chooses the simplest possible potentials, but maintains the important geometric constraints. In this way, it is hoped that universal properties are captured by DPD.

Despite these advantages, dissipative particle dynamics has some drawbacks. For example, the method is an algorithm that guarantees the satisfaction of some general physical principles, such as conservation of momentum and Galilean invariance, but energy is not conserved (although the temperature is fixed). Also, there exists no equation describing the governing dynamics of the system—only individual dynamics are described.

Dynamic Monte Carlo simulations are reviewed by Baumgärtner (1992). For example, Kolinsky, Skolnick, and Yaris (Skolnick and Kolinsky, 1990) simulated multiple nonphantom chains on a lattice. Dynamic-equilibrium properties of the chains are followed with time. The simulations do predict proper scaling coefficients for the diffusion constant and the longest relaxation time constant, as found by autocorrelation functions. However, the systems studied need to be far from highly concentrated to prevent freezing of the motion, or else the dynamics appears to be glassy. Generalization to flow fields does not appear to be straightforward, except perhaps through dissipative-dynamics simulations.

Ideally, one would like to perform molecular dynamics on concentrated systems to find equilibrium-dynamics properties. However, the long-time-scale dynamics of real chains is far out of reach of current computational resources. Instead, Kremer and coworkers (1983, 1988, 1990) simulated a concentrated system of Rouse chains with excluded volume. Their results appear to give proper scaling laws, and the innermost segments exhibit reptationlike autocorrelation functions for their motion. Hence, they concluded that reptation provided a good description of their simulations. However, all chains do not exhibit these laws, and some controversy remains about whether or not the simulations confirm reptation. Also, although the chains have excluded volume, it is still possible that chains can pass through one another for some configurations. An additional drawback of the computations is that additional Brownian forces and frictional forces were added to the beads to keep a constant temperature of the system. Such stochastic forces change the fundamental nature of the equations integrated.

3.3 Flows and Stresses in Complex Geometries

3.3.1 Numerical Solution of Constitutive Equations Predictions of stresses in flow fields are useful from a practical point of view. Many industrial processes involve the flow of non-Newtonian fluids in complex geometries—injection-molding dies, for example. Also, prediction of viscoelastic-flow instabilities is desirable. For a given constitutive equation and geometry, one must simultaneously solve several equations for all points inside the geometry while satisfying the boundary conditions. In the simplest case, these equations would be the equation of motion, Eq. (5), the equation of continuity (conservation of mass), and the constitutive equation. In addition, for nonisothermal problems, one may need to solve the energy equation and an equation of state simultaneously as well. The constitutive equation may be of the differential type (depending upon local velocity gradients only), or of the integral type (depending upon the deformation history of a fluid element).

Addition of the constitutive equation leads to complications not found in solving the Navier–Stokes equation, and current efforts have centered around four chosen problems: a four-to-one contraction flow, extrudate swell, viscoelastic stability in flows around cylinders, and flow through tubes of sinusoidal cross section.

The outstanding problems posed by viscoelastic fluids in complex flows can be understood in terms of the Weissenberg and Deborah numbers. The Deborah number is defined as the ratio of the fluid's largest relaxation time, which may be estimated from the relaxation modulus, Eq. (20), to the residence time of a fluid particle in the flow regime of interest. When the Deborah number is very small, then elastic effects are not as important as viscous effects, and the fluid behaves effectively as Newtonian. On the other hand, flow fields for which the Deborah number is very large may appear more like an elastic solid. Most interesting flow fields lie between these two extremes. Hence, special problems arise in attempting to predict flow fields for viscoelastic fluids that are not seen for either purely viscous or purely elastic deformations.

The Weissenberg number is defined (by Tanner, 1985, for example) as

$$We := \lambda\sqrt{\mathrm{tr}[\nabla\mathbf{v}_f + (\nabla\mathbf{v}_f)\dagger]}. \tag{45}$$

In other words, the Weissenberg number is the ratio of the relaxation time of the polymer to a characteristic time of the local rate of deformation of the fluid. Many numerical methods work well at small *We* and *De*, but fail as these groups are made larger than 1.

Special variations of mixed finite-element methods have been used to tackle special numerical problems presented by the particular forms of the constitutive equations being used. Current methods of attack on the prediction of viscoelastic-flow fields (Crochet *et al.*, 1984; Crochet, 1986) are finite-element methods, finite-difference methods, spectral methods, and boundary elements. Several recent papers on current methods have recently been published in an issue of the *Journal of Rheology* (Vol. 38, No. 2, 1994), including a comparison of higher- and lower-order techniques (Khomami *et al.*, 1994). Other techniques were recently studied in a comparison of numerical stability by Brown *et al.* (1993).

Many constitutive equations being used are from continuum mechanics, but some molecular-based models have been used as well. However, the molecular-based models require some mathematical approximation to find the closed-form expression for the stress tensor, necessary for the numerical method.

3.3.2 Combined Stochastic Simulation and Finite-Element Techniques In order to avoid the shortcomings of using the approximate molecular models, Laso and Öttinger (1993) have proposed combining finite-element techniques with Brownian-dynamics simulations. To estimate the stress field, Brownian-dynamics simulations are performed at each element, using the linearized local velocity field. Although the method is computationally (and memory) intensive, spatial and temporal problems are attainable. As an example, they calculated the stress and velocity fields for the inception of steady shear flow of a dilute solution of FENE dumbbells. The results are qualitatively quite different from the analytic solution using the approximate FENE model with Peterlin's consistent-averaging approximation; for example, when the exact model shows a single overshoot, the approximate model shows multiple oscillations, and the steady state is reached at a much later time.

ACKNOWLEDGMENTS

The author would like to thank the following people for aiding in the preparation of this manuscript: Prof. Hans Christian Öttinger, Prof. R. Byron Bird, Prof. T. C. B. McLeish, Prof. David C. Venerus, Prof. Wesley R. Burghardt, Prof. Bamin Khomami, Prof. Raj Rajagopalan, Prof. John M. Dealy, and Mr. Chi-Chung Hua.

GLOSSARY

Constitutive Equation: An equation relating the stress in a continuum to the imposed flow field.

Cox–Merz Rule: The observation that the steady-state viscosity curve as a function of shear rate has the same form as the magnitude of the complex viscosity as a function of frequency in small-amplitude oscillatory shearing flow.

Damping Function: A function that describes the strain-dependent part of the shear stress following a shear step strain.

Dynamic Structure Factor: A time-dependent averaged quantity that describes the relative positions among different segments of the same chain.

Entanglement Concentration: A critical concentration of polymer chains in a solvent, above which the chains exhibit properties consistent with the existence of chain–chain entanglements.

Entanglement Molecular Weight: A critical molecular weight below which it is believed that concentrated chains are not substantially entangled. The onset of entanglement is indicated by the appearance of a plateau in the relaxation modulus.

Flory Interaction Parameter: A parameter that describes the relative energetic interaction between chain segments and solvent molecules.

Kinetic Theory of Polymers: The description of macroscopic behavior of polymers by the use of equilibrium or nonequilibrium statistical mechanics.

Linear Viscoelasticity Theory: The theory that relates the stress tensor in any given small-strain deformation to the stress in any other small-strain deformation.

Polydispersity: The distribution of molecular weights of a batch of polymer chains.

Polymer Fluid Dynamics: The motion of polymeric liquids as described by the equations of motion and continuity.

Polymer Rheology: The study of flows of polymeric liquids.

Power-Law Region: The intermediate region of the steady-state viscosity curve that is linear on a log–log plot as a function of shear rate.

Radius of Gyration: An averaged quantity that indicates the size of a polymer coil in solution.

Reptation: The postulation that the dominant motion of linear polymer chains in concentrated systems is a one-dimensional diffusion of the chain along its backbone.

Stress-Optic Law: The observation that, within a constant factor (called the stress-optic coefficient), the index-of-refraction tensor and the stress tensor are equal during flow for a polymer melt.

Theta Solvent: A solvent for which the energetic interaction between segments on a chain relative to solvent–segment interactions is balanced by segment–segment steric interactions, and the Flory interaction parameter is $\frac{1}{2}$.

Theta Temperature: The temperature at which a solvent is a theta solvent.

Time-Temperature Superposition: The observation that changing the temperature of an experiment does not change the functional forms of measured properties, but appears to rescale the frequency for most experiments.

Zimm Plot: A plot of light-scattering data that allows double extrapolation of chain concentration and magnitude of the scattering wave vector to obtain the molecular weight and radius of gyration of a polymer chain in solution.

Works Cited

Baumgärtel, M., Schausberger, A., Winter, H. H. (1990), "The Relaxation of Polymers with Linear Flexible Chains of Uniform Length," *Rheol. Acta* **29,** 400–408.

Baumgärtner, A. (1992), "Simulations of Macromolecules," *Top. Appl. Phys.* **71,** 285–316.

Bates, F. S., Dierker, S. B., Wignall, G. D. (1986), "Phase Behavior of Amorphous Binary Mixtures of Perdeuterated and Normal 1,4-Polybutadienes," *Macromolecules* **19,** 1938–1945.

Beris, A., Edwards, B. (1994), *Thermodynamics of Flowing Systems with Internal Microstructure,* New York: Oxford Univ. Press.

Berne, B. J., Pecora, R. (1976), *Dynamic Light Scattering with Applications to Chemistry, Biology, and Physics,* New York: Wiley.

Berne, B. J. (1977), *Statistical Mechanics, Part B: Time-Dependent Processes,* New York: Plenum.

Biller, P., Petruccione, F. (1990), *J. Chem. Phys.* **92,** 6322–6326.

Bird, R. B., Stewart, W. E., Lightfoot, E. N. (1960), *Transport Phenomena,* New York: Wiley.

Bird, R. B., Curtiss, C. F., Armstrong, R. C., Hassager, O. (1987), *Dynamics of Polymeric Liquids,* Vol. 2, *Kinetic Theory,* 2nd ed., New York: Wiley-Interscience.

Boger, D. V., Walters, K. (1993), *Rheological Phenomena in Focus,* Amsterdam: Elsevier.

Brown, R. B., Szady, M. J., Northey, P. J., Armstrong, R. C. (1993), "On the Numerical Stability of Mixed Finite Element Methods for Viscoelastic Flows Governed by Differential Constitutive Equations," *Theor. Comput. Fluid Dyn.* **5,** 77–106.

Caroline, D., Jones, G. (1980), in: V. Degiorgio, M. Corti, M. Giglio (Eds.), *Light Scattering in Liquids and Macromolecular Solutions,* New York: Plenum.

Cottrell, F. R., Merrill, E. W., Smith, K. A. (1969), "Conformation of Polyisobutylene in Dilute Solution Subjected to a Hydrodynamic Shear Field," *J. Polym. Sci. A2* **7,** 1415–1434.

Crochet, M. J., Davies, A. R., Walters, K. (1984), *Numerical Simulation of Non-Newtonian Flow,* Amsterdam: Elsevier.

Crochet, M. J. (1986), *Rubber Chem. Technol.* **62,** 426–455.

Curtiss, C. F., Bird, R. B. (1995), "Statistical Mechanics of Transport Phenomena: Polymeric

Liquid Mixtures," *Adv. Polym. Sci.* (to be published).

de Gennes, P.-G. (1979), *Scaling Concepts in Polymer Physics*, Ithaca, NY: Cornell Univ. Press.

des Cloizeaux, J. (1990), *Macromolecules* **23**, 3992–4006.

des Cloizeaux, J., Jannink, G. (1990), *Polymers in Solution*, Oxford, UK: Clarendon.

Doi, M., Edwards, S. F. (1986), *The Theory of Polymer Dynamics*, Oxford, UK: Clarendon.

Feigl, K., Öttinger, H. C. (1994), "The Flow of a LDPE Melt through an Axisymmetric Contraction: A Numerical Study and Comparison to Experimental Results," *J. Rheol.* **38**, 847–874.

Ferry, J. D. (1988), *Viscoelastic Properties of Polymers*, 3rd ed., New York: Wiley.

Flory, P. J. (1969), *Statistical Mechanics of Chain Molecules*, New York: Wiley. Has been reprinted by Hanser (Munich, 1989) and distributed in the U.S. and Canada by Oxford Univ. Press.

Fuller, G. G. (1995), *Optical Rheometry of Complex Liquids*, New York: Oxford Univ. Press.

Gardiner, C. W. (1985), *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, 2nd ed., Berlin: Springer-Verlag, pp. 93–101.

Gleissle, W., Ohl, N. (1990), "On the Relaxation of Shear and Normal Stresses of Viscoelastic Fluids following Shear Rate Experiments," *Rheol. Acta* **29**, 261–280.

Graessley, W. W. (1974), *Adv. Polym. Sci.* **16**, 1.

Grmela, M. (1985), "Stress Tensor in Generalized Hydrodynamics," *Phys. Lett. A* **111**, 41–44.

Hess, W. (1986), *Macromolecules* **19**, 1395–1404.

Hess, W. (1988), *Macromolecules* **21**, 2620–2632.

Higgins, J. S. (1979), "Polymer Conformations and Dynamics," in: G. Kostorz (Ed.), *Treatise on Materials Science and Technology: Neutron Scattering*, vol. 25, New York: Academic.

Hoogerbrugge, P. J., Koelman, J. M. V. A. (1992), "Simulating Microscopic Hydrodynamic Phenomena with Dissipative Particle Dynamics," *Europhys. Lett.* **19**, 155–160.

Hua, C. C., Schieber, J. D. (1994), "Nonequilibrium Brownian Dynamics Simulations of Hookean and Finitely Extensible Dumbbell Models with Internal Viscosity in Shear Flows," *J. Non-Newtonian Fluid Mech.* **56**, 307–332.

Hürlimann, H. P. (1991), "A multidirectional shear rheometer for polymer melts," *Rheol. Acta* **30**, 480–490.

Janeschitz-Kriegl, H. (1983), *Polymer Melt Rheology and Flow Birefringence*, Berlin: Springer-Verlag.

Jongschaap, R. J. J. (1991), "Towards a Unified Formulation of Microrheological Models," in: J. Casas-Vázquez, D. Jou (Eds.), *Rheological Modelling: Thermodynamical and Statistical Approaches, Proceedings, Sant Feliu de Guíxols, Spain 1990,* Lecture Notes in Physics No. 381, Berlin: Springer-Verlag, pp. 215–247.

Khan, S. A., Larson, R. G. (1987), *J. Rheol.* **31**, 207–234.

Khomami, B., Talwar, K. K., Ganpule, H. K. (1994), "A Comparative Study of Higher- and Lower-order Finite Element Techniques for Computation of Viscoelastic Flows," *J. Rheol.* **38**, 255–289.

Kim, H., Chang, T., Yohanan, J. M., Wang, L., Yu, H. (1986), *Macromolecules* **19**, 2737–2744.

Kirkwood, J. G. (1967), *Macromolecules*, New York: Gordon and Breach.

Koelman, J. M. V. A., Hoogerbrugge, P. J. (1993), *Europhys. Lett.* **21**, 363–368.

Kramers, H. (1944), *Physica* **11**, 1–11.

Kremer, K. (1983), *Untersuchungen zur statistischen Mechanik von linearen Polymeren unter verschiedenen physikalischen Bedingungen*, Doctoral Dissertation, Institut für Festkörperforschung, Köln, Federal Republic of Germany.

Kremer, K., Grest, G. S. (1990), *J. Chem. Phys.* **92**, 5057–5086.

Kremer, K., Grest, G. S., Carmesin, I. (1988), *Phys. Rev. Lett.* **61**, 566–569.

Kuhn, W., Kuhn, H. (1945), "Bedeutung beschränkt freier Drehbarkeit für die Viskosität und Strömungsdoppelbrechung von Fadenmolekellösungen I," *Helv. Chim. Acta* **28**, 1533–1579.

Kulicke, W. M., Wallbaum, U. (1985), "Determination of First and Second Normal Stress Differences in Polymer Solutions in Steady Shear Flow and Limitations Caused by Flow Irregularities," *Chem. Eng. Sci.* **40**, 961–972.

Laso, M., Öttinger, H. C. (1993), *J. Non-Newtonian Fluid Mech.* **47**, 1–20.

Li, T.-Q., McCarthy, K. L. (1995), "Pipe flow of Aqueous Polyacrylamide Solutions Studied by Means of Nuclear Magnetic Resonance Imaging," *J. Non-Newtonian Fluid Mech.* **57**, 155–175.

Liang, C.-H., Mackay, M. E. (1993), *J. Rheol.* **37**, 149–174.

Link, A., Springer, J. (1993), *Macromolecules* **26**, 464–471.

Liu, T. W. (1989), *J. Chem. Phys.* **90**, 5826–5842.

Lodge, A. S. (1964), *Elastic Liquids* London: Academic Press.

Lodge, T. P., Rotstein, N. A., Prager, S. (1990), *Adv. Chem. Phys.* **79**, 1–132.

Mead, D. (1994), "Determination of Molecular Weight Distributions of Linear Flexible Polymers from Linear Viscoelastic Material Functions," *J. Rheol.* **38**, 1797–1827.

Meissner, J. (1984), "Polymer Melt Rheology—A Challenge for Polymer Scientist and Engineer," *Pure Appl. Chem.* **56**, 369–384.

Meissner, J., Stephenson, S. E., Demarmels, A., Portmann, P. (1982), "Multiaxial Elongation Flows of Polymer Melts—Classification and Experimental

Realization," *J. Non-Newtonian Fluid Mech.* **11,** 221–237.

Muller, R., Pesce, J. J., Picot, C. (1993), *Macromolecules* **26,** 4356–4362.

Nemoto, N., Landry, M. R., Noh, I., Yu, H. (1984), *Polym. Comm.* **25,** 141–143.

Nemoto, N., Landry, M. R., Noh, I., Kitano, T., Wesson, J. A., Yu, H. (1985), *Macromolecules* **18,** 308–310.

Öttinger, H. C. (1995), *Stochastic Processes in Polymeric Fluids,* Berlin: Springer-Verlag.

Rajagopalan, D., Byars, J. A., Armstrong, R. C., Brown, R. A., Lee, J. S., Fuller, G. G. (1992), "Comparison of Numerical Simulations and Birefringence Measurements in Viscoelastic Flow between Eccentric Rotating Cylinders," *J. Rheol.* **36,** 1349–1375.

Riande, E., Saiz, E. (1992), *Dipole Moments and Birefringence of Polymers,* Englewood Cliffs, NJ: Prentice-Hall.

Rudisill, J. W., Cummings, P. T. (1992), "Brownian Dynamics of Model Polymer Fluids in Shear Flow. I. Dumbbell Models", *J. Non-Newtonian Fluid Mech.* **41,** 275–288.

Saab, H. H., Bird, R. B., Curtiss, C. F. (1982), *J. Chem. Phys.* **77,** 4758–4766.

Schieber, J. D. (1987), "Kinetic Theory of Polymer Melts. IX. Comparisons with Experimental Data," *J. Chem. Phys.* **87,** 4928–4936.

Schieber, J. D., Öttinger, H. C. (1988), "The Effects of Bead Inertia on the Rouse Model," *J. Chem. Phys.* **89,** 6972–6981.

Schieber, J. D. (1991), "A Hookean Dumbbell Model with Basset Forces for Dilute Polymer Solutions," *J. Chem. Phys.* **94,** 7526–7533.

Schieber, J. D., Öttinger, H. C. (1994), "On Consistency Criteria for Stress Tensors in Kinetic Theory Models," *J. Rheol.* **38,** 1909–1924.

Schweizer, K. S. (1989), *J. Chem. Phys.* **91,** 5802–5821, 5822–5839.

Skolnick, J., Kolinski, A. (1990), *Adv. Chem. Phys.* **78,** 223–278.

Szu, C. S., Hermans, J. J. (1974), *J. Polym. Sci. Polym. Phys. Ed.* **12,** 1743–1751.

Tanner, R. I. (1985), *Engineering Rheology,* Oxford, UK: Clarendon.

Tirrell, M. (1984), "Polymer Self-Diffusion in Entangled Systems," *Rubber Chem. Technol.* **57,** 523–556.

VanArsdale, W. E. (1995), "Constitutive Equations for Viscoelastic Fluids," preprint.

van den Brule, B. H. A. A., Hoogerbrugge, P. J. (1994), "Brownian Dynamics Simulation of Polymeric Networks," in: C., Gallegos, A. Guerrero, J., Muñoz, M. Berjano (Eds.), *Proceedings of the Fourth European Rheology Conference, Progress and Trends in Rheology IV,* Hemsbach, Germany: Steinkopff Verlag.

Venerus, D. C., Kohvand, H. (1994), *J. Polym. Sci. B: Polym. Phys.* **32,** 1531–1542.

Wales, J. L. (1976), *The Application of Birefringence to Rheological Studies of Polymer Melts,* Amsterdam: Delft Univ. Press.

Wedgewood, L. E. (1994), "Internal Viscosity in Polymer Kinetic Theory: Shear Flows," *Rheol. Acta* **32,** 405–417.

Yamakawa, H. (1971), *Modern Theory of Polymer Solutions,* New York: Harper and Row.

Zülle, B., Linster, J. J., Meissner, J., Hürlimann, H. P. (1987), "Deformation Hardening and Thinning in Both Elongation and Shear of a Low Density Polyethylene Melt," *J. Rheol.* **31,** 583–598.

Zylka, W. (1991), "Gaussian Approximation and Brownian Dynamics Simulations for Rouse Chains with Hydrodynamic Interaction Undergoing Simple Shear Flow," *J. Chem. Phys.* **94,** 4628–4636.

Zylka, W., Öttinger, H. C. (1991), "Calculation of Various Universal Properties for Dilute Polymer Solutions Undergoing Shear Flow," *Macromolecules* **24,** 484–494.

Further Reading

Bird, R. B., Armstrong, R. C., Hassager, O. (1987), *Dynamics of Polymeric Liquids,* Vol. 1, *Fluid Mechanics,* 2nd ed., New York: Wiley-Interscience.

Bird, R. B., Öttinger, H. C. (1992), *Annu. Rev. Phys. Chem.* **43,** 371–406.

Bird, R. B., Wiest, J. W. (1995), "Constitutive Equations for Polymeric Liquids," *Annu. Rev. Fluid Mech.* **27,** 169–193.

Curtiss, C. F., Bird, R. B. (1995), "Statistical Mechanics of Transport Phenomena: Polymeric Liquid Mixtures," *Adv. Polym. Sci.* (to be published).

Freed, K. F. (1987), *Renormalization Group Theory of Macromolecules,* New York: Wiley.

Hermans, J. J. (1978), *Polymer Solution Properties,* Part 1, *Statistics and Thermodynamics,* Part 2, *Hydrodynamics and Light Scattering,* New York: Academic.

Larson, R. G. (1988), *Constitutive Equations for Polymer Melts and Solutions,* Boston: Butterworths.

McCammon, J. A., Harvey, S. C. (1987), *Dynamics of Proteins and Nucleic Acids,* Cambridge, U.K.: Cambridge Univ. Press.

Tanner, R. I. (1985), *Engineering Rheology,* Oxford, U.K.: Clarendon.

POLYMER PROCESSING TECHNOLOGY

See TECHNOLOGY OF POLYMER PROCESSING

POLYMERIZATION AND POLYMER REACTIONS

CHARLES L. MCCORMICK, *Department of Polymer Science, The University of Southern Mississippi, Hattiesburg, Mississippi, U.S.A.*

	Introduction	445
1.	**Structural Features of Polymers**	446
2.	**Mechanisms of Polymerization**	446
2.1	Step-Growth Polymerization	446
2.1.1	The Carothers Equation	446
2.1.2	Kinetic Expressions	448
2.1.3	Examples of Linear Step-Growth Polymers	449
2.1.4	Polyesters	449
2.1.5	Polycarbonates	451
2.1.6	Polyamides	451
2.1.7	Polyimides and Polybenzimidazoles	452
2.1.8	Polyurethanes and Polyureas	453
2.1.9	Network Polymers	453
2.2	Chain-Growth Polymerization	455
2.2.1	Structural Features	456
2.2.2	Kinetics of Radical Chain Polymerization	458
2.2.3	Kinetic Chain Length and Degree of Polymerization	460
2.2.4	Free-Radical Polymerization Techniques	461
2.2.5	Free-Radical Copolymerization	462
2.2.6	Anionic Polymerization	464
2.2.7	Star and Telechelic Polymers by Anionic Polymerization	467
2.2.8	Cationic Polymerization	468
2.2.9	Cationic Copolymerization	469
2.2.10	Coordination Polymerization	469
3.	**Reactions of Functional Polymers**	470
3.1	Functional-Group Reactivity	470
3.2	Cross-Linking Reactions	471
3.3	Reactions of Saturated Polyolefins	471
3.4	Reactions of Unsaturated Polyolefins	471
3.5	Reactions of Halogenated Polymers	472
3.6	Reactions of Hydroxy- or Amine-Containing Polymers	473
3.7	Reaction of Acyl Groups	474
3.8	Controlled Cyclization Reactions	475
3.9	Functionalization for Specific Activity	475
	Glossary	475
	Works Cited	476
	Further Reading	476

INTRODUCTION

Polymers are macromolecules synthesized from small-molecule precursors or *monomers* by a process termed *polymerization.* The linkages along the polymer chain are formed by mechanisms that are generally divided into two classifications—*step growth* and *chain growth.* Reaction conditions and constituents determine the final chain structure and the number of repeating units of the macromolecule. Polymers can have relatively low molecular weights (often termed *oligomers*) or may have ultrahigh molecular weights reaching $>10^7$ g/mol.

Synthetic polymers have traditionally been divided into categories based on their respective physical properties. These include plastics, elastomers, coatings, adhesives, fibers, water-soluble resins, etc. Equally diverse in properties are *biopolymers* synthesized in biochemical processes. These include polynucleotides (deoxyribonucleic acid, DNA, and ribonucleic acid, RNA), polysaccharides, proteins, natural rubber, and naturally occurring polyesters.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

In subsequent sections for organizational purposes, polymerization or polymer-forming reactions are presented in terms of step-growth and chain-growth mechanisms for major synthetic polymer types. Postpolymerization modifications or reactions are then addressed.

1. STRUCTURAL FEATURES OF POLYMERS

General structural features of polymers are illustrated in Fig. 1. Architecture of the polymer may be linear, branched, or cross linked. Additional diversity in architecture and ultimate behavior can be built in by introducing multiple monomers into the macromolecules during or subsequent to polymerization. For example, linear chains composed of two constitutive repeating units A and B can be arranged in random, alternating, block, graft, or cross-linked fashion as illustrated in Fig. 2. The behavioral characteristics are determined by the structural features of each and would be expected to differ markedly even if the compositions of A and B units were the same. In addition to the above considerations, bond type, valence-bond and conformational angles, steric, polar, and electrostatic interactions, segmental solvation, long-range interactions, molecular weight, concentration, packing, and exclusion effects are among the many variables that dictate ultimate properties. Thus, there are nearly limitless opportunities for tailoring macromolecules for specific applications by rational monomer selection and effective polymerization procedures.

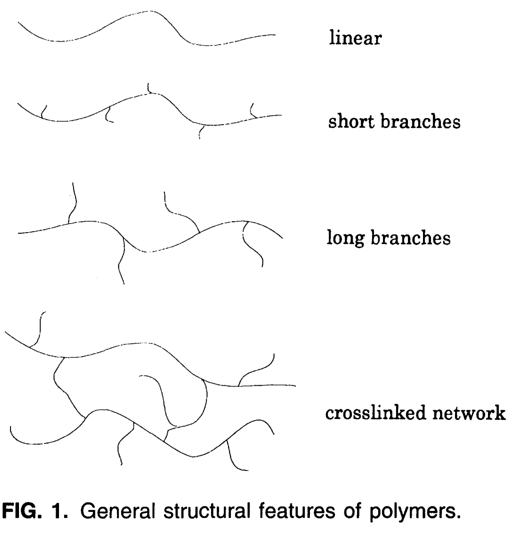

FIG. 1. General structural features of polymers.

2. MECHANISMS OF POLYMERIZATION

In early years, as the polymerization process was being studied, pioneering work by Wallace Carothers (1929) and by Paul Flory (1953) revealed that not only were structural features important, but also the mechanisms (and kinetics) of polymer-forming reactions could be separated. Carothers's classification of the two major modes, termed "condensation" reactions and "addition" reactions, was later updated to use the more precise designations of "step-growth" and "chain-growth" processes.

In step-growth polymerization reactions, functional groups on respective monomer units react in a manner to build molecular weight slowly. The linking mechanism results in formation of dimers, tetramers, pentamers, etc. In chain-growth polymerization, chain extension or propagation occurs by addition of monomer to an active chain end, which generally occurs at a rapid rate. *Initiation* and *termination* reactions (and often *chain-transfer* reactions) determine the final degree of polymerization and the distribution of molecular sizes.

Table 1 (Lenz, 1967) compares step-growth and chain-growth polymerization processes in terms of component reactions, growth, and monomer reactivity.

2.1 Step-Growth Polymerization

Step-growth polymerizations require reactive functional groups that combine to form a new linking unit between reactant pairs. In order to obtain linear polymers with high molecular weights, the reactants must be difunctional. Monofunctional reactants serve as chain stoppers, and multifunctional monomers lead to branching and/or cross linking.

2.1.1 The Carothers Equation Wallace Carothers, often called the father of step-growth polymerization, developed an equation that relates the degree of polymerization ($\bar{D}p$) to the extent of reactions, ρ, for linear

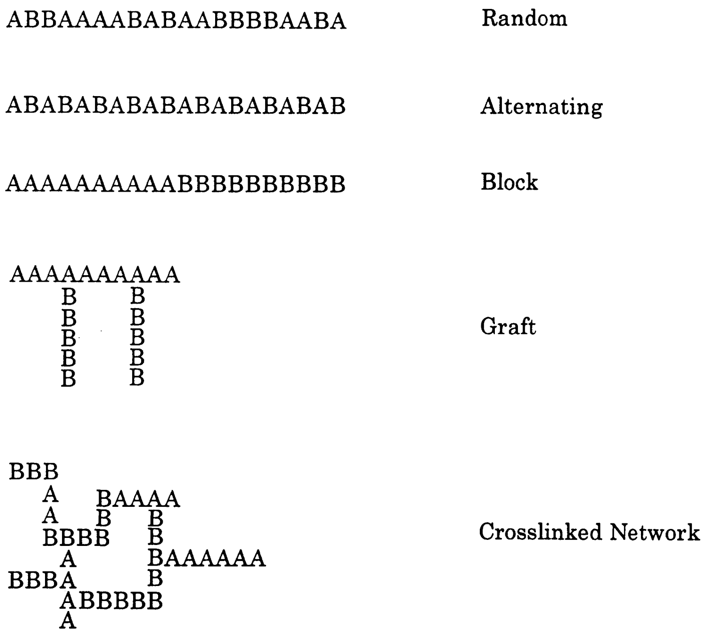

FIG. 2. Possible copolymer structures formed from respective monomers *A* and *B*.

polymers prepared by self-condensation of a monomer (*A–B*),

$$N_0A\text{–}B \rightarrow A\text{–}B\left[A\text{–}B\right]_{N_0-2}A\text{–}B, \tag{1}$$

or by reaction of monomer pairs (*A–A* plus *B–B*).

$$N_0A\text{–}A + N_0B\text{–}B \rightarrow A\text{–}A\left[B\text{–}B\text{–}A\text{–}A\right]_{N_0-2}B\text{–}B. \tag{2}$$

In these cases *A* and *B* possess reactive functional groups. At a given time, the conversion ρ is defined by

$$\rho = (N_0 - N)/N_0, \tag{3}$$

in which N represents the total number of molecules present. The term N_0 is the number of molecules originally present. Since the degree of polymerization is given by

$$\bar{D}p = N_0/N, \tag{4}$$

Table 1. Characteristics of step-growth and chain-polymerization growth.

Characteristic	Step-growth polymerization	Chain-growth polymerization
Component	Initiation, propagation, and termination reactions essentially identical in rate and mechanism	Initiation, propagation, and termination reactions significantly different in rate and mechanism
Polymer growth	Slow, random growth of polymer with high reaction conversion required for a high degree of polymerization	Rapid, preferential growth of each polymer chain, once initiated, with high degrees of polymerization attained at the start of the reaction
Reactivity of monomer	Monomer capable of combining with itself or with any other species present with equal facility, resulting in rapid disappearance of monomer at early stages of the polymerization reaction	Monomer capable of reaction only with an active end group on a polymer chain, resulting in steadily decreasing but significant concentrations of monomer throughout the polymerization reaction

then substitution yields the Carothers equation,

$$\bar{D}p = 1/(1 - \rho). \tag{5}$$

Here $\bar{D}p$ represents the number-average degree of polymerization or the average number of structural repeating units per polymer chain. The number-average molecular weight $\bar{M}_n$ is then obtained by multiplying $\bar{D}p$ by M_0, the repeating-unit molecular weight, and adding the molecular weight of the end groups.

In order for step-growth polymerization to yield a high–molecular-weight, linear polymer, high conversion must be attained. This requires nearly exact equivalency of reacting functional groups of the monomers and the absence of side reactions.

The Carothers equation can be modified for the presence of molecules N_1 with only single functionality (A or B):

$$\bar{D}p = \frac{1 + N_1/N_0}{1 + \rho + N_1/N_0}. \tag{6}$$

In A–A plus B–B polymerization, for molar ratios r of the reactant monomers less than 1, the degree of polymerization is given by

$$\bar{D}p = \frac{1 + \rho}{1 + \rho - 2r\rho}. \tag{7}$$

The extent of reaction or conversion is represented by ρ.

Equations (6) and (7) may be used to show the importance of stoichiometric balance and monomer purity in attaining high molecular weight. Organic reactions that can be utilized for step-growth polymers must proceed to virtually 100% yield of the desired linkages. It should be noted that controlled addition of monofunctional monomers or excess difunctional monomers is an effective method of controlling molecular weight. Often prepolymers for subsequent polymerization or cross linking are prepared in this manner.

The Carothers equation is also useful for predicting conversion at which gelation occurs for multifunctional monomers containing equivalent numbers of A and B groups:

$$\rho = 2/f_{\text{avg}} - 2/\bar{D}pf_{\text{avg}}. \tag{8}$$

The value of f_{avg} is calculated from

$$f_{\text{avg}} = \Sigma N_i f_i (\Sigma N_i)^{-1}, \tag{9}$$

where N_i is the number of molecules with reactive functionality. At the gel point, $\bar{D}p \rightarrow \infty$; Eq. (8) can be rewritten as

$$\rho_c = 2/f_{\text{avg}} \tag{10}$$

at the critical conversion ρ_c. Similar expressions have been derived for nonstoichiometric mixtures containing more than two monomers.

2.1.2 Kinetic Expressions Kinetic expressions for step-growth polymerization have been detailed in several reviews (Odian, 1991, Chap. 2; Manaresi and Munari, 1989; Elias, 1977, vol. 2, Chap. 17). Rates of step-growth polymerization can be easily followed by monitoring the concentrations of reacting functional groups. An example of the polycondensation of the A–A + B–B type involving a dicarboxylic acid and a diol is given below. For polymerization catalyzed by an added protic acid, HX, the reaction scheme involves first the equilibrium formation of complex I between one carboxylic acid of an A–A monomer with protic acid:

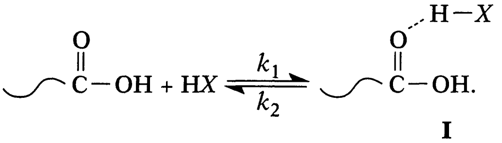

I

(11)

Complex I subsequently reacts with one of the alcohol functional groups on a B–B monomer via addition to the carboxyl carbon yielding II:

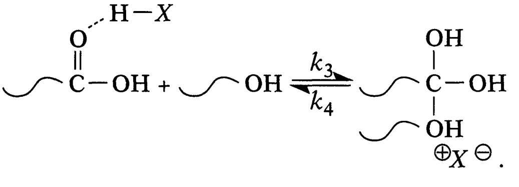

I II

(12)

In the final step, elimination of water with recovery of HX catalyst yields the ester:

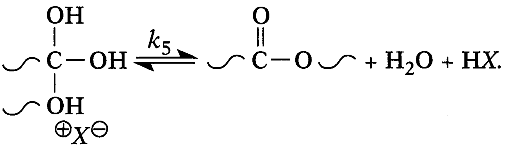

$$\sim\overset{\displaystyle OH}{\underset{\underset{\displaystyle \oplus X\ominus}{\sim OH}}{C}}-OH \underset{}{\overset{k_5}{\rightleftharpoons}} \sim\overset{\displaystyle O}{\overset{\|}{C}}-O\sim + H_2O + HX.$$

II

(13)

The successive esterifications are equilibrium reactions. In order to obtain conversions necessary for high molecular weight, the equilibrium is shifted to the right by removal of water. Under these conditions, reactions (12) and (13) are considered to be irreversible. Generally formation of II is the rate-limiting step. Since k_1, k_2, and k_5 are large compared with k_3, an expression for the rate of polymerization, R_{po}, in terms of the rate of disappearance of the carboxylic acid groups is given by

$$R_{po} = \frac{-d[\sim COOH]}{dt} = k_3[I][\sim OH]. \tag{14}$$

Because the determination of the concentration of I experimentally is not easy, the usual approach is to express I in terms of the equilibrium expression

$$K = \frac{k_1}{k_2} = \frac{[I]}{[\sim COOH][HX]}; \tag{15}$$

rearrangement and substitution into Eq. (14) yields

$$\begin{aligned} R_{po} &= \frac{-d[\sim COOH]}{dt} \\ &= k_3K[\sim COOH][\sim OH][HX]. \end{aligned} \tag{16}$$

The value of K is dependent on the strength of the external acid under the solvation and temperature conditions of the polymerization. In polymerization starting with equimolar concentrations of [~COOH] and [~OH], the molar concentration of carboxylic acid groups at time t can be related to the original concentration, $[COOH]_0$, by the conversion; Eq. (16) after integration yields an expression for the degree of polymerization,

$$\bar{D}p = \frac{1}{1+\rho} = 1 + k_3[K][\sim COOH]_0 t. \tag{17}$$

In polymerization self-catalyzed by the carboxylic acid groups on the monomers, the rate of polymerization can be shown to yield

$$\begin{aligned} R_{po} &= \frac{-d[\sim COOH]}{dt} \\ &= k_3[\sim COOH]^2[\sim OH]. \end{aligned} \tag{18}$$

For equimolar concentrations of monomers, the degree of polymerization at time t is given by

$$\bar{D}p^2 = \frac{1}{(1+\rho)^2} = 1 + 2k_3[\sim COOH]_0^2 t. \tag{19}$$

Experimental observations of the kinetics of step-growth polymerization often deviate from those predicted by Eqs. (17) and (19). The nonlinearity has been attributed to changes in the polarity of the reaction medium. Early in the reaction, small molecular species account for the solvent conditions. As the reaction proceeds the number of functional (usually polar) groups decreases, resulting in a medium of lower dielectric constant. The solvent conditions would be expected to affect the potential energy of both the ground state and transition state, thus changing the activation energy. Other kinetic factors affecting the rate constant include collision frequency and orientation. These may change with conversion for particular step-polymerization reaction types.

2.1.3 Examples of Linear Step-Growth Polymers In order for step-growth polymerizations to yield molecular weights >20 000, conversion must be high and stoichiometry must be maintained (Sec. 2.1.1). Despite an abundance of potential polymer-forming organic reactions, relatively few meet the criteria demanded by the Carothers relationship. Selected successful commercial routes are discussed below.

2.1.4 Polyesters Linear *polyesters* may be prepared by reactions of *A–B* monomers of the type HO–*R*–*X* in self-condensation reactions

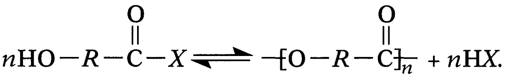

$$n HO-R-\overset{\displaystyle O}{\overset{\|}{C}}-X \rightleftharpoons -\!\!\!\!\left[O-R-\overset{\displaystyle O}{\overset{\|}{C}}\right]_n\!\!\!\!- + nHX. \tag{20}$$

Likewise difunctional alcohols (*A–A* type monomers) may be reacted with difunctional diacyl monomers (*B–B* type) as shown in

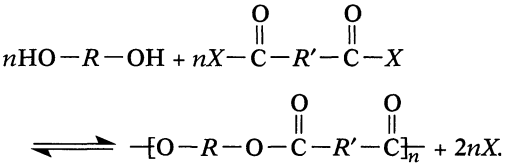

$$n\mathrm{HO}-R-\mathrm{OH}+nX-\overset{\overset{\mathrm{O}}{\|}}{\mathrm{C}}-R'-\overset{\overset{\mathrm{O}}{\|}}{\mathrm{C}}-X \rightleftharpoons -\!\!\!\left[\mathrm{O}-R-\mathrm{O}-\overset{\overset{\mathrm{O}}{\|}}{\mathrm{C}}-R'-\overset{\overset{\mathrm{O}}{\|}}{\mathrm{C}}\right]_n + 2nX. \qquad (21)$$

In both of the above examples, $-\overset{\overset{\mathrm{O}}{\|}}{\mathrm{C}}-X$ may be a carboxylic acid, anhydride, acid chloride, amide, ester, or other derivative in which *X* is a good leaving group. The terms *R* and *R'* are general representations of alkyl or aromatic groups.

A widely used method for preparation of polyesters is an *interchange reaction. Ester interchange* is shown below for the preparation of poly(ethyleneterephthalate). The preformed diethylester of terephthalic acid is reacted with diethylene glycol to yield the ester:

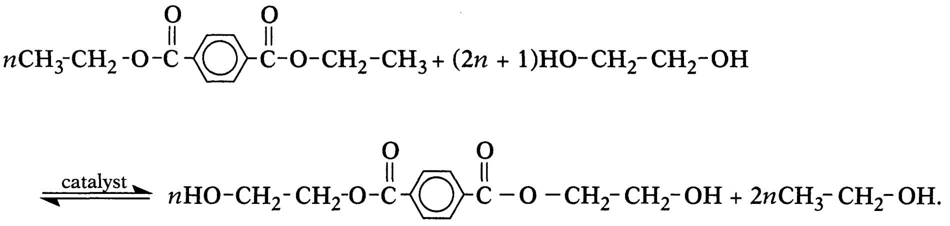

(22)

A transesterification catalyst (usually a proton donor, metallic ion, or weak base) is utilized and the equilibrium is shifted to the right by removal of ethanol. The ester-interchange intermediate is then polymerized to poly(ethyleneterephthalate) utilizing a catalyst (often a transition-metal complex) under reduced pressure:

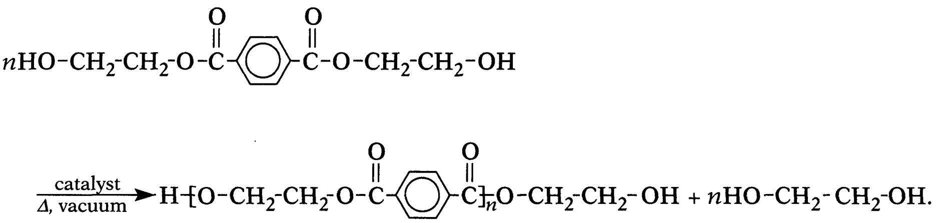

(23)

Alternatively, *direct esterification* of terephthalic acid with ethylene glycol can be accomplished to yield the polyester:

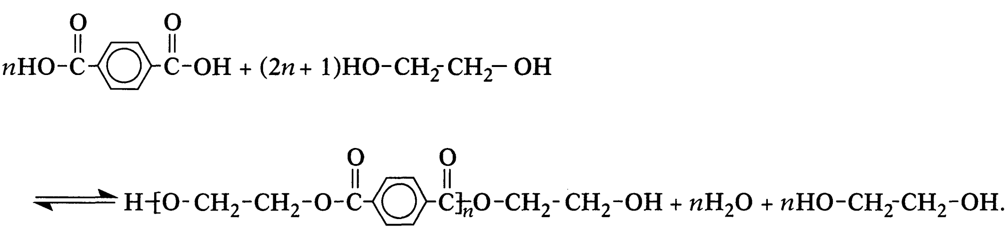

(24)

Often an excess of ethylene glycol is used to increase the rate of the initial esterification of the diacid. The excess ethylene glycol also allows conversion of any carboxyl end groups to esters, which are less likely to catalyze unwanted side reactions during processing. Unreacted ethylene glycol is removed during the final stages of polymerization under reduced pressure.

Phthalic anhydride may be reacted with ethylene glycol to produce poly(ethylene-orthophthalate):

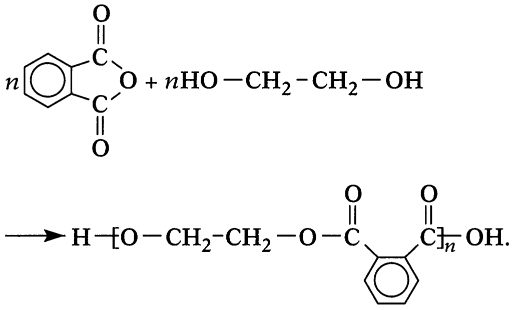

(25)

A final example of esterification is that of a diol reacting with a diacid chloride via the Schotten–Baumann reaction,

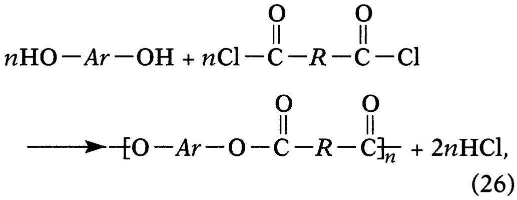

$$n\mathrm{HO}-Ar-\mathrm{OH}+n\mathrm{Cl}-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}-R-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}-\mathrm{Cl} \longrightarrow -[\mathrm{O}-Ar-\mathrm{O}-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}-R-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}]_n- + 2n\mathrm{HCl},$$

(26)

where *Ar* denotes an aromatic group. This interfacial polymerization procedure is often used with aromatic diols and is particularly facile at an interface between an aqueous alkaline solution of the aromatic diol and an immiscible organic solution of the diacid chloride. Alternatively, an amine is added as an acid receptor.

2.1.5 Polycarbonates Interfacial polymerization is also used to produce *polycarbonates* such as bisphenol A polycarbonate:

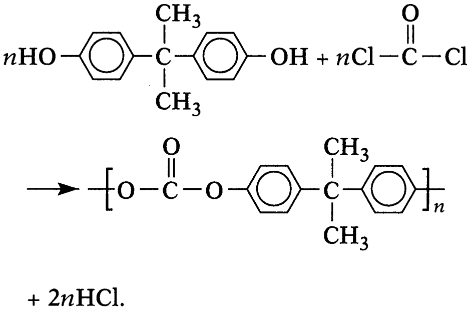

(27)

An alternative method to the above polycarbonate preparation involves the interchange reaction of the aromatic diol with diphenylcarbonate:

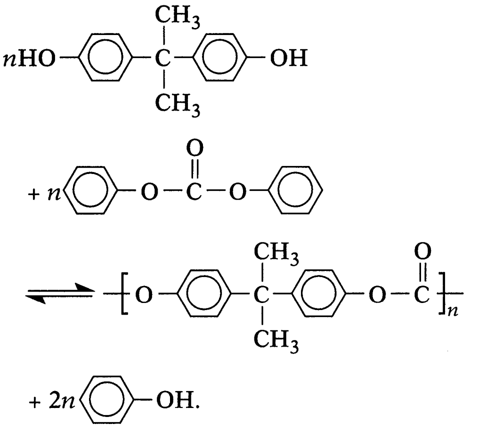

(28)

2.1.6 Polyamides *Polyamides* can be prepared by reactions of *A–B* monomers,

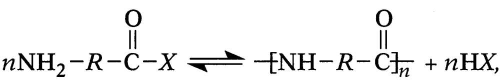

$$n\mathrm{NH}_2-R-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}-X \rightleftharpoons -[\mathrm{NH}-R-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}]_n- + n\mathrm{H}X,$$

(29)

or *A–A* + *B–B* monomer pairs,

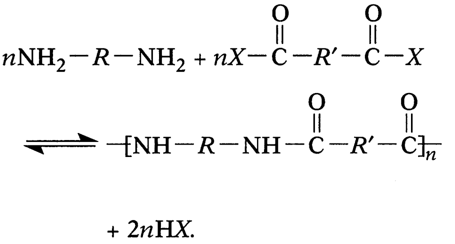

$$n\mathrm{NH}_2-R-\mathrm{NH}_2+nX-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}-R'-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}-X \rightleftharpoons -[\mathrm{NH}-R-\mathrm{NH}-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}-R'-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}]_n- + 2n\mathrm{H}X.$$

(30)

The nature of *R* and *R′* can be varied and, as with polyesterification reactions, *X* should be a good leaving group.

Direct amidation is often useful in preparing polyamides. An advantage of this method is that reaction of a diabasic acid and a diamine allows isolation of a diammonium salt prior to polymerization. The exact stoichiometry and high yields on subsequent heating under reduced pressure meet the requirement of the Carothers equations for obtaining reasonable molecular weights. Nylon 6,6 is prepared in this manner by reacting hexamethylene diamine with adipic acid:

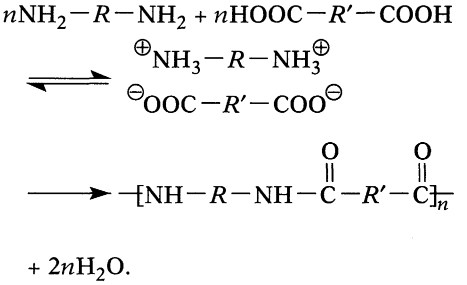

$$nNH_2-R-NH_2 + nHOOC-R'-COOH \rightleftharpoons \begin{matrix} {}^{\oplus}NH_3-R-NH_3^{\oplus} \\ {}^{\ominus}OOC-R'-COO^{\ominus} \end{matrix} \longrightarrow -\!\!\left[NH-R-NH-\overset{\overset{O}{\|}}{C}-R'-\overset{\overset{O}{\|}}{C}\right]_n + 2nH_2O.$$

(31)

Nylon-6, polyamide, is commercially prepared by the combination of *ring opening/transamidation* reaction of the cyclic lactam monomer, caprolactam:

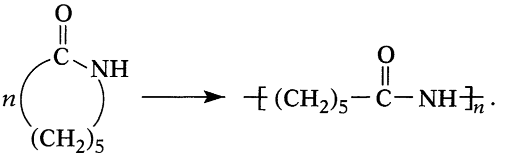

(32)

Aramids or aromatic polyamides are prepared by reacting aromatic diamines with aromatic diacids or diacid chlorides (*A–A* + *B–B* type). Alternatively an *A–B* type monomer can be polymerized. The respective reactions are

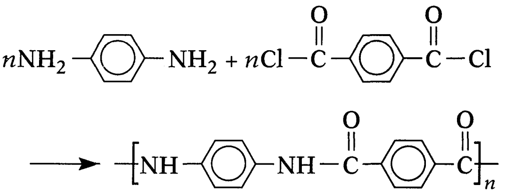

(33)

and

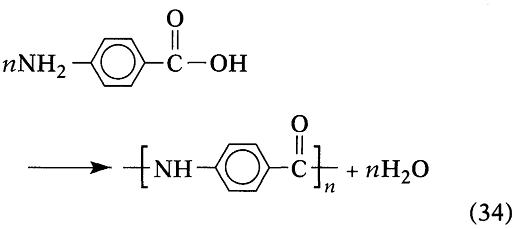

(34)

for synthesis of poly(imino-1,4-phenyleneiminoterephthaloyl) and poly(iminocarbonyl-1,4-phenylene)—both sold under the trade name Kevlar® (Odian, 1991, p. 105).

2.1.7 Polyimides and Polybenzimidazoles *Polyimides* are formed by the reaction of aliphatic or aromatic diamines with a dianhydride,

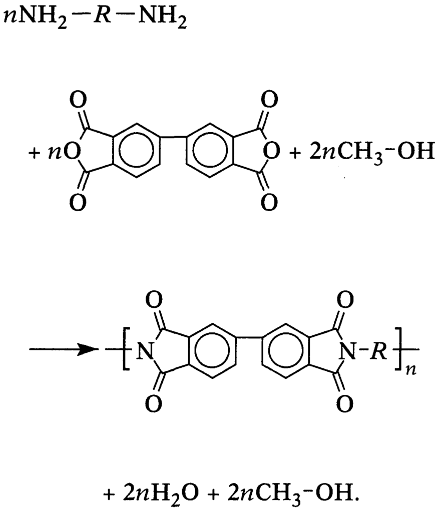

(35)

The reaction is usually carried out in a stepwise fashion to produce a polyamic acid intermediate prior to imidization.

Polybenzimidazoles are also heterocyclic polymers with high temperature stability. These are found by reactions of tetramines with diacyl monomers:

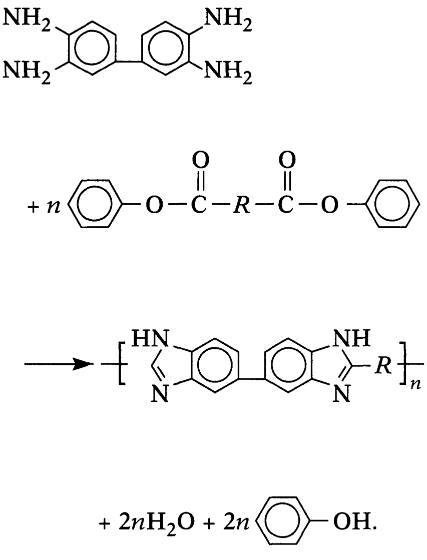

(36)

2.1.8 Polyurethanes and Polyurea *Polyurethanes* and *polyureas* are prepared from *A–A* and *B–B* monomer pairs in which a diisocyanate is added to a diol,

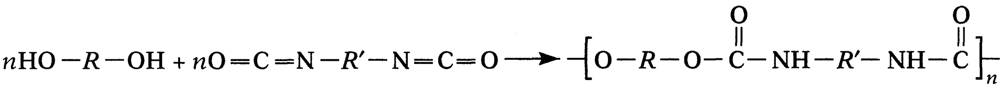

$$n\mathrm{HO{-}}R\mathrm{{-}OH} + n\mathrm{O{=}C{=}N{-}}R'\mathrm{{-}N{=}C{=}O} \longrightarrow \left[\mathrm{O{-}}R\mathrm{{-}O{-}}\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}\mathrm{{-}NH{-}}R'\mathrm{{-}NH{-}}\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}\right]_n, \tag{37}$$

or diamine

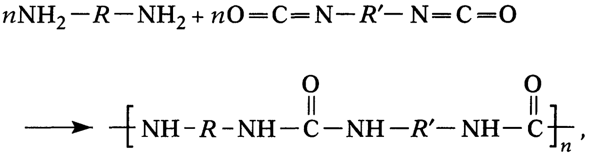

$$n\mathrm{NH_2{-}}R\mathrm{{-}NH_2} + n\mathrm{O{=}C{=}N{-}}R'\mathrm{{-}N{=}C{=}O} \longrightarrow \left[\mathrm{NH{-}}R\mathrm{{-}NH{-}}\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}\mathrm{{-}NH{-}}R'\mathrm{{-}NH{-}}\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}\right]_n, \tag{38}$$

respectively. Linear high-molecular-weight polymers are formed readily by maintaining a stoichiometric balance with the reacting monomers. The reactivity of the diisocyanate groups allows the reactions to proceed at moderate temperatures.

Polyurethane and polyurea prepolymers are often prepared for subsequent chain extensions in which *multiblock* copolymers are desired. The Carothers equation can be used to predict stoichiometry necessary to form a prepolymer, with end groups,

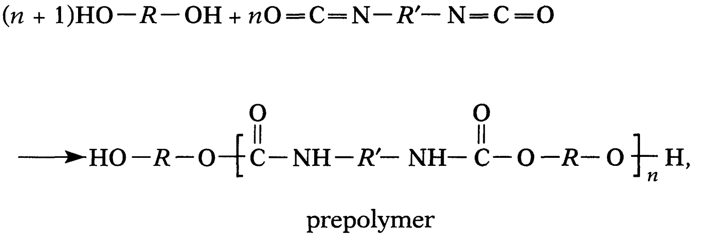

$$(n+1)\mathrm{HO{-}}R\mathrm{{-}OH} + n\mathrm{O{=}C{=}N{-}}R'\mathrm{{-}N{=}C{=}O} \longrightarrow \underset{\text{prepolymer}}{\mathrm{HO{-}}R\mathrm{{-}O}\left[\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}\mathrm{{-}NH{-}}R'\mathrm{{-}NH{-}}\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}\mathrm{{-}O{-}}R\mathrm{{-}O}\right]_n\mathrm{H}}, \tag{39}$$

which can subsequently be reacted with a second diisocyanate to form the copolymer,

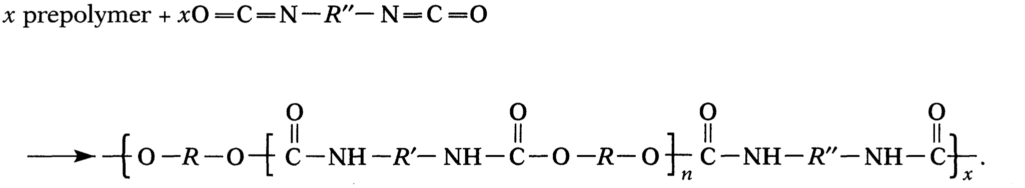

$$x\ \text{prepolymer} + x\mathrm{O{=}C{=}N{-}}R''\mathrm{{-}N{=}C{=}O} \longrightarrow \left[\mathrm{O{-}}R\mathrm{{-}O}\left[\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}\mathrm{{-}NH{-}}R'\mathrm{{-}NH{-}}\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}\mathrm{{-}O{-}}R\mathrm{{-}O}\right]_n\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}\mathrm{{-}NH{-}}R''\mathrm{{-}NH{-}}\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}\right]_x. \tag{40}$$

2.1.9 Network Polymers As discussed previously, introduction of multiple functional groups (>2) onto reactive monomers can result in network formation. The Carothers equation can be utilized to estimate the conversion necessary for this to occur; thus somewhat controlled network formation is possible with many of the step-growth reactions discussed previously. For example, polyurethane and polyurea networks can be easily formed by reacting multifunctional diols or diamines with diisocyanates. Alternatively, excess diisocyanate may be added to the linear polymers, which can react at elevated temperature with the active hydrogen atoms of the urethane or urea groups to form cross links.

Another convenient way to control network formation is to introduce functional groups that are less reactive than those participating in the initial polymer-forming reaction. Subsequent reaction (for example, heating with additional comonomer to a higher temperature) results in network formation. Reactions of this type, illustrated by

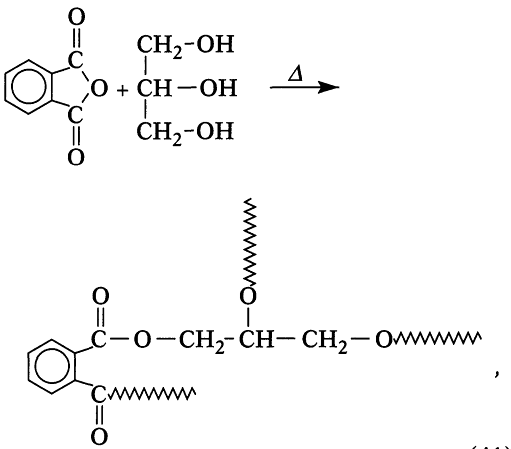

(41)

are important in technologies (plastics, coatings, elastomers) in which liquid prepolymers are converted to cross-linked materials with outstanding physical properties.

Unsaturated polyesters, polyamides, polyurethanes, etc., can be prepared by incorporating monomers containing double bonds:

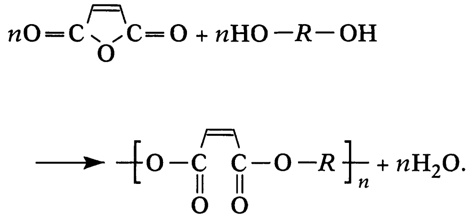

(42)

The double bonds can then serve as reactive sites for ionic or free-radical addition resulting in network formation.

Some of the earliest network polymers were prepared in the early 1900s by Leo Baekeland. The Bakelite® polymers were condensation products of the reactions of phenol with formaldehyde:

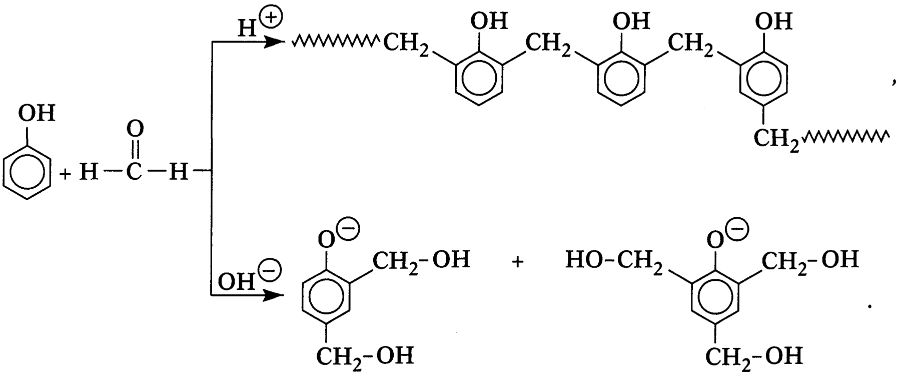

(43a)

(43b)

Under acidic conditions, (43a), "novolak" resins are formed by electrophilic aromatic substitution of methyloyl groups onto phenol; under basic conditions (43b), "resoles" are formed by nucleophilic attack of the ortho phenolate carboanion onto formaldehyde. The novolak resins have no residual $-CH_2-OH$ groups and cannot cross link without additional reagents. However, the resoles can undergo facile cross linking on heating. Cross-linked resins are also obtained by reactions of formaldehyde with urea or melamine through a series of methyloyl condensation reactions.

A final example of a cross-linkable resin formed by step-growth polymerization is that of an *epoxy* resin. An excess of the epoxide, epichlorohydrin, is reacted with bisphenol A,

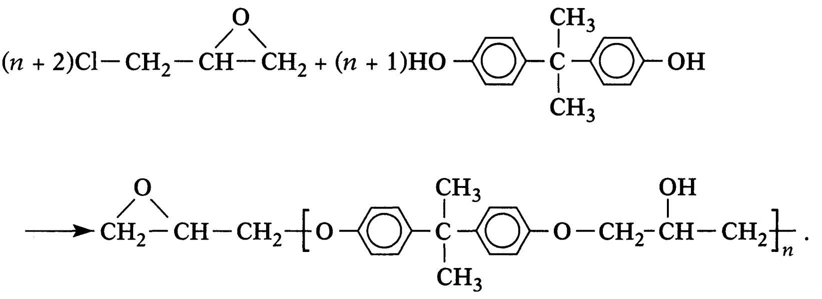

(44)

Multifunctional acyl compounds may be used to cross link the residual -OH groups in the repeating sequence. Alternatively, the epoxide end groups can be reacted with difunctional amines or alcohols. The properties of epoxy cross-linked materials vary over a wide range depending on copolymer composition and cross-link density. Current uses include adhesives, coatings, and matrix resins for high-performance composites.

2.2 Chain-Growth Polymerization

In contrast to step-growth polymerization, chain-growth polymerization proceeds in a rapid manner in which the *initiated* species continues to add monomer or *propagate* until *termination* occurs. Thus control of three essential steps—initiation, termination, and propagation—as well as possible competing reactions is vital for chain growth to proceed to high molecular weight. In the sections to follow, major types of chain-growth polymerization initiated by free-radical, anionic, cationic, and coordinated species will be examined. The general scheme for an alkene with a functional group R is given by initiation,

$$I \xrightarrow{k_d} I*, \tag{45}$$

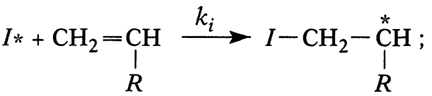

$$I* + CH_2{=}\underset{R}{CH} \xrightarrow{k_i} I{-}CH_2{-}\overset{*}{\underset{R}{CH}}\,; \tag{46}$$

propagation,

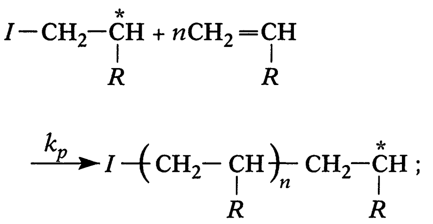

$$I{-}CH_2{-}\overset{*}{\underset{R}{CH}} + nCH_2{=}\underset{R}{CH} \xrightarrow{k_p} I\left(CH_2{-}\underset{R}{CH}\right)_n CH_2{-}\overset{*}{\underset{R}{CH}}\,; \tag{47}$$

and termination,

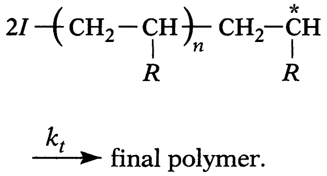

$$2I\left(CH_2{-}\underset{R}{CH}\right)_n CH_2{-}\overset{*}{\underset{R}{CH}} \xrightarrow{k_t} \text{final polymer.} \tag{48}$$

An initiator species (anionic, cationic, free radical, or coordinated) is formed in the first step with rate constant k_d. The second step of the initiation involves addition (in this case to a double bond) of the monomer. Subsequent monomer additions to the propagating chain continue with addition of n molecules of monomer until termination occurs. The respective rate constants for the second and third steps are k_p and k_t.

Successful polymerization by the pathway shown by Eqs. (45)–(48) requires careful consideration of initiator, monomer, and reaction variables.

The nature of the R group is important in selecting the proper type of initiator for the monomer. For the propagation step, the rate is given by

$$R_p = k_p[M*][M], \tag{49}$$

in which $M*$ represents the propagating species and M the monomer. The rate is therefore dependent upon the concentrations of the propagating species and the monomers as well as the magnitude of k_p. From the Arrhenius equation

$$k_p = A_p \exp(-E_p/R_gT), \qquad (50)$$

it is evident that k_p will be dependent upon the activation energy E_p for propagation, the reaction temperature T, and the pre-exponential term A_p, which includes collision frequency and orientation effects. Activation-energy values are largely determined by the nature of the R group on the unsaturated monomer, $\mathrm{CH_2{=}}\underset{R}{\mathrm{CH}}$, and reactivity of the propagating species, $\sim\mathrm{CH_2{-}}\underset{R}{\overset{*}{\mathrm{CH}}}$. Electronic and steric effects on the relative potential energy of the ground state and activated complex in transition state dictate the magnitude of the activation energy. Those factors increasing reactivity of the ground-state reagents and lowering the potential energy of the transition state lower E_p and increase the rate of the reaction.

For the above reasons, monomers with electron-donating R groups generally polymerize by cationic or coordinated cationic mechanisms. Monomers with electron-withdrawing substituents polymerize by anionic or coordinated anionic mechanisms. In the above cases, the transition states develop late in the reaction and predictions are easy. Those groups that stabilize the developing carbocation or carbanion lower ΔE_p. However, in free-radical reactions, radical character can be developed early or late along the reaction pathway. Both the potential energy of the ground-state reactants and that of the activated complex in the transition state must be assessed in order to predict reactivity.

In anionic or cationic polymerization, the concentration of the propagating species is on the order of 10^{-3} to 10^{-5} molar as compared with 10^{-7} to 10^{-9} for free-radical polymerization. The reactions are exothermic and are often carried out at low temperatures. The advantages of enhanced rates and stereocontrol inherent with ionic polymerizations are often negated by the strict requirements of avoiding termination by contaminants such as water or other acids and bases.

2.2.1 Structural Features A virtually limitless number of monomers are available for chain-growth polymerization. A number of important commercial polymers and some typical properties are given in Table 2.

In addition to repeating structures that normally arrange in a head-to-tail fashion, mono-substituted alkenes may add in a head-to-head fashion:

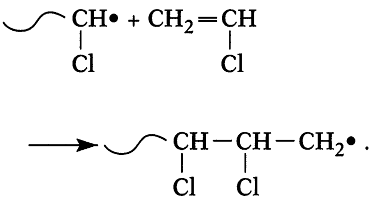

$$\sim\underset{\mathrm{Cl}}{\mathrm{CH}\bullet} + \mathrm{CH_2{=}}\underset{\mathrm{Cl}}{\mathrm{CH}} \longrightarrow \sim\underset{\mathrm{Cl}}{\mathrm{CH}}{-}\underset{\mathrm{Cl}}{\mathrm{CH}}{-}\mathrm{CH_2\bullet}. \qquad (51)$$

Steric factors in the transition state or complexation of chain ends with monomer and counterions in ionic polymerizations can yield stereoregular isotactic, (52a), or syndiotactic, (52b), segments as differentiated by the configurations along the polymer backbone:

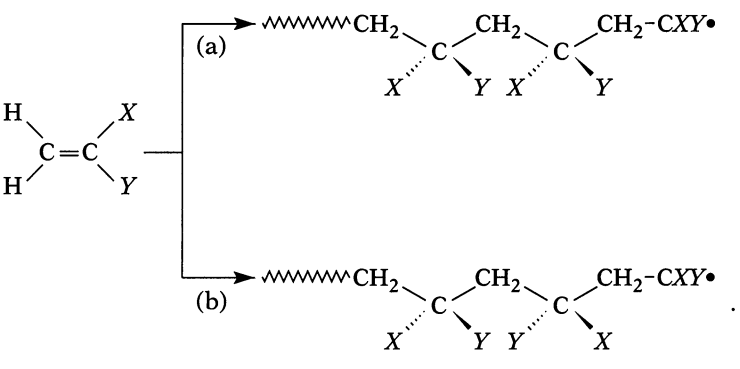

(52a)

(52b)

Table 2. Representative chain-growth polymers.

Name	Repeating unit	Properties
Poly(acrylamide)	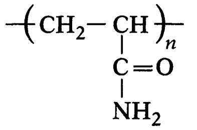 $-\!\left(CH_2-\underset{\underset{NH_2}{\overset{\vert}{C=O}}}{\overset{}{\underset{\vert}{CH}}}\right)_{n}\!-$	Water soluble, viscosifier
Poly(acrylonitrile)	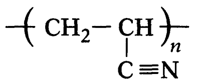 $-\!\left(CH_2-\underset{C\equiv N}{\underset{\vert}{CH}}\right)_{n}\!-$	High strength, tough, chemical resistant
Poly(vinyl acetate)	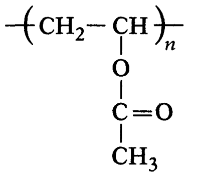 $-\!\left(CH_2-\underset{\underset{\underset{CH_3}{\overset{\vert}{C=O}}}{\overset{\vert}{O}}}{\underset{\vert}{CH}}\right)_{n}\!-$	Film former, adhesive
Poly(vinyl chloride)	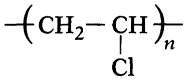 $-\!\left(CH_2-\underset{Cl}{\underset{\vert}{CH}}\right)_{n}\!-$	Chemical, moisture resistant, UV unstable
Poly(tetrafluoroethylene)	$-\!\left(CF_2-CF_2\right)_{n}\!-$	Chemically inert, low friction properties, low surface energy
Poly(ethylene)	$-\!\left(CH_2-CH_2\right)_{n}\!-$	Transparent, flexible, film former, good electrical properties
Poly(propylene)	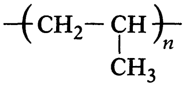 $-\!\left(CH_2-\underset{CH_3}{\underset{\vert}{CH}}\right)_{n}\!-$	High strength, hard, high softening point, chemical, moisture resistant
Poly(isoprene)	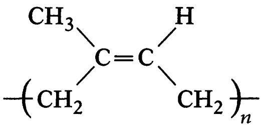	Elastomeric, properties of natural rubber
Poly(styrene)	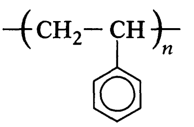	Clear, moderate glass transition temperature
Poly(methyl methacrylate)	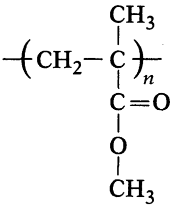 $-\!\left(CH_2-\overset{CH_3}{\overset{\vert}{\underset{\underset{\underset{CH_3}{\overset{\vert}{O}}}{\overset{\vert}{C=O}}}{\underset{\vert}{C}}}}\right)_{n}\!-$	Clear, good impact strength, mechanical properties

Atactic polymers are those with randomly placed units of both along the backbone. Dienes can polymerize through the 1,2 or 1,4 positions to yield polymers with isomeric repeating units,

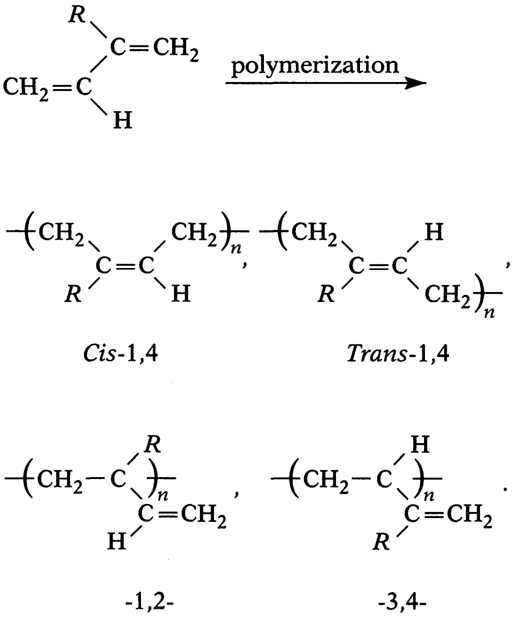

(53)

Not only can *cis* and *trans* isomers be prepared; tactic structures can result by changing appropriate polymerization conditions.

In the late 1950s, Butler first reported cyclopolymerization of nonconjugated dienes to give non–cross-linked polymer structures. The mechanism has been shown subsequently (Butler, 1989) to occur by an alternating intra-intermolecular propagation,

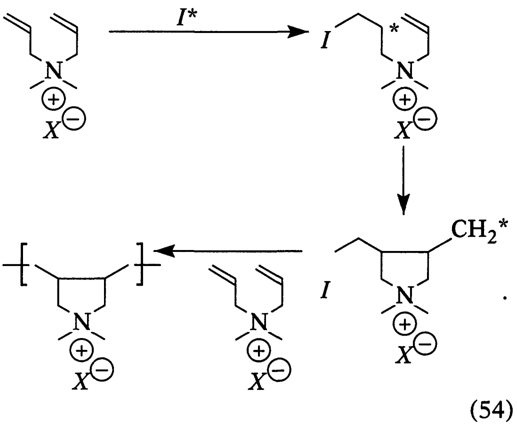

(54)

Cross-linked or network polymers can be formed by chain-growth polymerization of nonconjugated monomers provided the reactivity of the double bonds is sufficient. For example, hydrogel polymers are prepared by action of acrylamide with varying amounts of bisacrylamide,

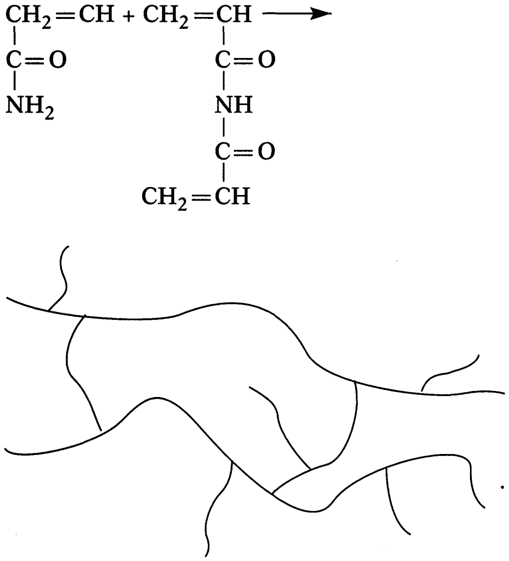

(55)

2.2.2 Kinetics of Radical Chain Polymerization Initiation in free-radical polymerization takes place in two steps. In the first, an initiating radical $I\bullet$ is formed (usually from thermal or photochemical decomposition). The initiator choice (Table 3) is extremely important for rate of reaction and molecular-weight distribution.

The rate constants for propagation for free-radical chain-growth polymerizations can be easily determined knowing the rate of disappearance of monomer, $-d[M]/dt$, the original concentration of monomer, M, and the concentration of radicals $[M\bullet]$, utilizing

$$\frac{-d[M]}{dt} = k_p[M\bullet][M]. \tag{56}$$

Unfortunately the concentration of radicals, $[M\bullet]$, is not easily obtained since it is usually 10^{-7} to 10^{-9} mol/L. An alternative approach utilizes the *steady-state* assumption discussed subsequently in this section.

In most polymerizations, the second step of the initiation is much faster than the first,

Table 3. Initiators and reaction parameters for free-radical polymerization (data from Brandrup and Immergut, 1975).

Initiator	Solvent	Temp. (°C)	K_d (s^{-1})	E ($kcal \cdot mol^{-1}$)
2,2′-azo-bis-isobutyronitrile (AIBN)	Benzene	40	5.4×10^{-7}	30.7
Phenyl-azo-triphenylmethane	Benzene	80	4.3×10^{-6}	26.8
tert-butyl-peroxide (TBP)	Benzene	80	7.8×10^{-8}	34.0
Cumyl peroxide	Benzene	115	1.6×10^{-5}	40.7
Acetyl peroxide	Benzene	35	9.5×10^{-5}	32.3
Benzoyl peroxide	Benzene	30	2.6×10^{-8}	30.4
Lauroyl peroxide	Benzene	30	2.6×10^{-7}	30.4
tert-butyl hydroperoxide	Benzene	154	4.3×10^{-6}	40.8
tert-butyl perbenzoate	Benzene	100	1.1×10^{-5}	34.7

and since two initiator fragments are formed, the rate of initiation is given by

$$R_i = 2fk_d[I_2], \tag{57}$$

in which $[I_2]$ represents the initiator concentration. The factor f is included to indicate the efficiency of initiation. It should be noted that the larger the number of chains initiated, the lower the molecular weight; the closer in time all chains start, the narrower the molecular-weight distribution.

In the propagation step, the free radical generated in the second step of the initiation reacts with monomer. During propagation, chain length grows as monomer is added. The reactivity of the growing radical chain is related to those factors previously discussed. For example, if monomer M is represented by $CH_2{=}CHR$, the nature of the substituent in regard to steric, polar, and resonance effects in both the ground state and the transition state is important. In the simplest case, the length of the growing chain is usually assumed to be unimportant for each successive step,

$$M\bullet + nM \xrightarrow{k_p} M(M)_{n-1}M\bullet \tag{58}$$

Under such conditions, the rate expression is given by

$$R_p = \frac{-d[M]}{dt} = k_p[M\bullet][M]. \tag{59}$$

Two common ways that free radicals terminate are by *combination,*

$$M_n\bullet + M_m\bullet \xrightarrow{k_{tc}} M_n - M_m, \tag{60}$$

and *disproportionation,*

$$M_n\bullet + M_m \xrightarrow{k_{td}} M_n(\mathrm{H}) + M_m(-\mathrm{H}). \tag{61}$$

The rate constants for those processes are k_{tc} and k_{td} respectively. Those monomers possessing abstractable hydrogens on carbons located in the β position relative to the propagating radical are prone to disproportionation during polymerization. This process is illustrated for termination of methyl methacrylate, which possesses five abstractable hydrogens:

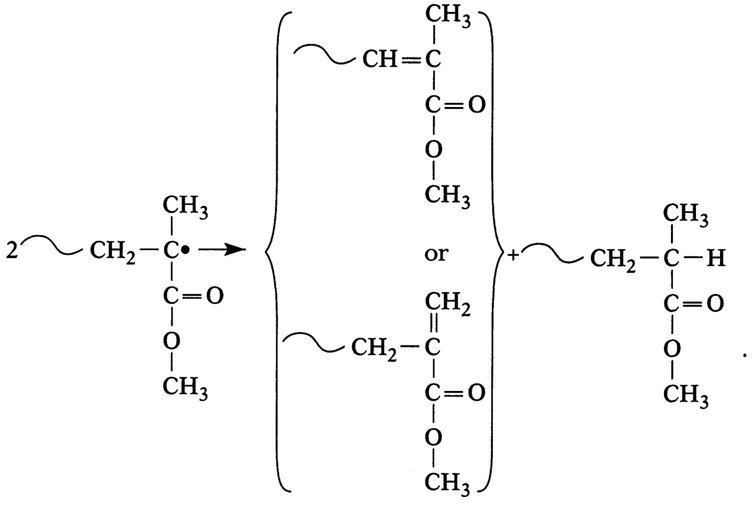

(62)

The mechanism for termination for many polymers may be a combination of both reactions (60) and (61); therefore a rate expression may be written as

$$R_t = k_{tc}[M_n\bullet][M_m\bullet] + k_{td}[M_n\bullet][M_m\bullet]. \tag{63}$$

Since $[M\bullet]$ is often difficult to measure, a simple assumption is made based on the production, after an initial induction period, of a constant concentration or "steady state" of radicals. In such cases, the rate of change of radical with time is given by

$$\frac{-d[M\bullet]}{dt} = 0. \tag{64}$$

The rate of termination should be equal to the rate of initiation,

$$R_i = R_t. \tag{65}$$

Therefore, rearrangement of

$$R_i = 2k_t[M\bullet]^2 \tag{66}$$

and solving for $[M\bullet]$ yields

$$[M\bullet] = (R_i/2k_t)^{1/2}. \tag{67}$$

Substituting into the propagation expression

$$R_p = k_p[M\bullet][M] \tag{68}$$

yields

$$R_p = k_p(R_i/2k_t)^{1/2}[M]. \tag{69}$$

Substitution of Eq. (57) for R_i yields

$$R_p = k_p[M](k_d f[I_2]/k_t)^{1/2}, \tag{70}$$

in which the constants can be combined,

$$k' = (k_p^2 k_d f/k_t)^{1/2}, \tag{71}$$

to yield the final expression for the rate of polymerization,

$$R_p = k'[M][I_2]^{1/2}. \tag{72}$$

2.2.3 Kinetic Chain Length and Degree of Polymerization The kinetic chain length, ν, in a free-radical polymerization is the number of monomers consumed per effective initiator radical and is equal to R_p/R_i as given by

$$\nu = \frac{k_p[M]}{2(k_d k_t f[I_2])^{1/2}}. \tag{73}$$

Combining constants,

$$k'' = \frac{k_p}{2(k_d k_t f[I_2])^{1/2}}, \tag{74}$$

and substituting into Eq. (73) yields the simple expression

$$\nu = k''[M]/[I_2]^{1/2}. \tag{75}$$

The degree of polymerization is then equal to 2ν for combination and ν for disproportionation.

Chain-transfer reactions involve abstraction of a labile hydrogen or halogen atom by the propagating radical:

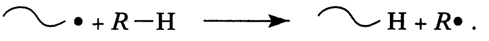

$$\sim\!\bullet + R{-}H \longrightarrow \sim\!H + R\bullet. \tag{76}$$

The abstraction may occur on initiator, solvent, monomer, or polymer. Hydrogens or halogens with low bond-dissociation energies (i.e., those on benzylic or allylic carbons) are most likely candidates for abstraction.

Since chain transfer competes with propagation, the considerations discussed previously are also important here. For example if the activation energy for addition is similar to that of chain transfer to monomer, the degree of polymerization should be low and the molecular-weight distribution wide. The expression for the degree of polymerization is given, in terms of transfer to monomer, solvent, initiator, and polymer, by

$$\bar{D}p = R_p\{(R_t/2) + k_{\mathrm{tr},M}[M\bullet][M] + k_{\mathrm{tr},S}[M\bullet][S] + k_{\mathrm{tr},I}[M\bullet][I_2] + k_{\mathrm{tr},P}[M\bullet][P]\}^{-1}. \tag{77}$$

After chain transfer, the new radical formed may then add monomer. When the rate of propagation does not change, this is termed "effective" chain transfer. If the rate is slowed, the term "degradative" chain transfer is used. In some cases, the reaction is stopped. In fact, retarders or inhibitors are often added to control chain length or to stabilize monomers during shipment and storage. Inhibitors include quinones, nitro aromatics, phenols, and aromatic amines. Note that these compounds provide stable radicals upon addition or abstraction that are of too low potential energy to add to monomer.

Intramolecular chain transfer along the polymer chain may occur especially at latter stages of the reaction. Such transfer results in branches:

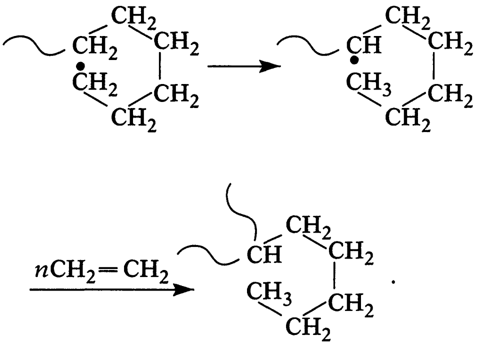

(78)

Equation (77) can be simplified by substituting expressions for R_p and R_t and combining constants. The reciprocal expression

$$\frac{1}{\bar{D}p} = \frac{1}{\bar{D}p_0} + C_m + C_s\frac{[S]}{[M]} + C_I\frac{[I]}{[M]} + C_p\frac{[P]}{[M]} \tag{79}$$

is convenient for expressing the extent of transfer in terms of chain-transfer constants to solvent, monomer, initiator, and polymer. The expression $1/\bar{D}p_0$ represents polymerization in the absence of transfer reactions and is the same as R_p/R_i. In theory, the above expression can be used to determine the magnitudes of the transfer constants

$$C_m = k_{\mathrm{tr},m}, \quad C_s = k_{\mathrm{tr},S}/k_p, \quad C_I = k_{\mathrm{tr},I}/k_p, \quad C_p = k_{\mathrm{tr},p}/k_p. \tag{80}$$

Values of chain-transfer constants for a number of polymerization reactions are compiled in Brandrup and Immergut (1975).

2.2.4 Free-Radical Polymerization Techniques Polymerization techniques for free radicals are varied and can range from those conducted under homogeneous conditions to specialized microheterogeneous polymerization. Major techniques are briefly discussed below.

In *bulk* polymerization, the monomer serves as the solvent and initiator is added. While simpler than solution polymerization in that regard, heat-transfer problems may occur or polymer solubility may diminish at a high conversion. Gel-like structures often give processing problems. In *solution* polymerization, the monomer and initiator are dissolved in a common solvent. The solvent serves as a heat sink, but isolation of polymer requires solvent removal. The techniques for *suspension* polymerization require dispersion of the monomer in an organic or water medium. Polymerization occurs in suspended droplets. An advantage of this method is that heat is dissipated and polymer is easily isolated. Suspension stability during reaction is sometimes difficult for certain monomers.

In a classical *emulsion* system, a hydrophobic monomer, a water-soluble initiator, an emulsifier, water, and stabilizers are added. Polymerization occurs in micelles. Under optimum conditions, high molecular weights may be obtained at high rates with relative ease of handling. The kinetics of the polymerization are quite different from bulk or solution polymerization. *Inverse-emulsion* techniques can be utilized with water-soluble monomers, a hydrocarbon or nonpolar solvent, and appropriate initiators. Although some systems behave as emulsion polymerizations, often the mechanism is actually that of an inverse-suspension polymerization. Both techniques have the advantage of producing water-soluble polymers, which can be easily handled.

Special cases of polymerization not discussed here are gas-phase polymerization and polymerization of monomers trapped in a continuous phase to yield compositelike structures.

2.2.5 Free-Radical Copolymerization

At the beginning of Sec. 2.2, the factors affecting the rate of homopolymerization were discussed. In copolymerization, these factors are important in determining the rates as well as microstructural compositions of the resulting copolymers. A useful approach to copolymerization is to consider two monomers M_1 and M_2 whose propagating chain ends are represented by M_1* and M_2* respectively. If reactivity of the propagating species is assumed to be dependent only on the chain end, four propagation reactions with their respective rates can be represented by

$$M_1* + M_1 \xrightarrow{k_{11}} M_1*, \tag{81}$$

$$M_1* + M_2 \xrightarrow{k_{12}} M_2*, \tag{82}$$

$$M_2* + M_1 \xrightarrow{k_{21}} M_1*, \tag{83}$$

$$M_2* + M_2 \xrightarrow{k_{22}} M_2*. \tag{84}$$

The rates of disappearance of the two monomers are given by

$$\frac{-d[M_1]}{dt} = k_{11}[M_1*][M_1] + k_{21}[M_2*][M_1] \tag{85}$$

and

$$\frac{-d[M_2]}{dt} = k_{12}[M_1*][M_2] + k_{22}[M_2*][M_2]. \tag{86}$$

The ratio of the rates of incorporation of the two monomers in the chain is obtained by dividing Eq. (85) by Eq. (86) to yield

$$\frac{d[M_1]}{d[M_2]} = \frac{k_{11}[M_1*][M_1] + k_{21}[M_2*][M_1]}{k_{12}[M_1*][M_2] + k_{22}[M_2*][M_2]}. \tag{87}$$

By assuming a steady-state concentration for each reactive species,

$$k_{21}[M_2*][M_1] = k_{12}[M_1*][M_2], \tag{88}$$

appropriate rearrangement and substitution into Eq. (87), and defining *reactivity ratios* $r_1 = k_{11}/k_{12}$ and $r_2 = k_{22}/k_{21}$, one can obtain the *copolymerization equation*

$$\frac{d[M_1]}{d[M_2]} = \frac{1 + r_1([M_1]/[M_2])}{1 + r_2([M_2]/[M_1])}. \tag{89}$$

The relative values of r_1 and r_2 are exceptionally important since they represent the propensity of the respective monomers for self addition vs cross addition.

Although there are many cases where the assumptions used to derive Eq. (89) are inappropriate, the copolymerization equation for chain-growth processes is a useful tool for predicting *instantaneous* copolymer compositions. Reactivity ratios for common comonomer pairs are given in Table 4 as compiled in Brandrup and Immergut (1975). Detailed kinetic expressions for co- and terpolymerization are reviewed elsewhere (Odian, 1991, Chap. 6).

Most copolymerizations proceed in a manner in which the relative concentrations of the reacting monomers change with the reaction time. Only at specified azeotropic compositions is composition constant with time.

Table 4. Reactivity ratios for common monomer pairs (adapted by Seymour and Carraher, 1981, from Brandrup and Immergut, 1975).

M_1	M_2	r_1	r_2	r_1r_2
Acrylamide	Acrylic acid	1.38	0.36	0.5
	Methyl acrylate	1.30	0.05	0.07
	Vinylidene chloride	4.9	0.15	0.74
Acrylic acid	Acrylonitrile (50 °C)	1.15	0.35	0.40
	Styrene	0.25	0.15	0.04
	Vinyl acetate (70 °C)	2.0	0.1	0.2
Acrylonitrile	Butadiene	0.25	0.33	0.08
	Ethyl acrylate (50 °C)	1.17	0.67	0.78
	Maleic anhydride	6.0	0	0
	Methyl methacrylate	0.13	1.16	0.15
	Styrene	0.04	0.41	0.16
	Vinyl acetate	4.05	0.06	0.24
	Vinyl chloride	3.28	0.02	0.07
Butadiene	Methyl methacrylate	0.70	0.32	0.22
	Styrene	1.39	0.78	1.08
Chlorotrifluoroethylene	Tetrafluoroethylene	1.0	1.0	1.0
Isoprene	Styrene	1.98	0.44	0.87
Maleic anhydride	Methyl acrylate	0	2.5	0
	Methyl methacrylate	0.03	3.5	0.11
	Styrene	0	0.02	0
	Vinyl acetate (70 °C)	0.003	0.055	0.0002
Methyl acrylate	Acrylonitrile	0.67	1.26	0.84
	Styrene	0.18	0.75	0.14
	Vinyl acetate	9.0	0.1	0.90
	Vinyl chloride	5	0	0
Methyl isopropenyl ketone	Styrene (80 °C)	0.66	0.32	0.21
Methyl methacrylate	Styrene	0.50	0.50	0.25
	Vinyl acetate	20	0.015	0.30
	Vinyl chloride	12.5	0	0
α-Methylatyrene	Maleic anhydride	0.038	0.08	0.003
	Styrene	0.38	2.3	0.87
Styrene	*p*-Chlorostyrene	0.74	1.025	0.76
	Fumaronitrile	0.23	0.01	0.002
	p-Methoxystyrene	1.16	0.82	0.95
	Vinyl acetate	55	0.01	0.55
	Vinyl chloride	17	0.02	0.34
	2-Vinylpyridine	0.56	0.9	0.50
Vinyl acetate	Vinyl chloride	0.23	1.68	0.39
	Vinyl laurate	1.4	0.7	0.98
Vinyl chloride	Diethyl maleate	0.77	0.009	0.007
	Vinylidene chloride	0.3	3.2	0.96
N-Vinylpyrrolidone	Styrene (50 °C)	0.045	15.7	0.71

Values of r_1 and r_2 yield insight into the nature of the copolymer formed. When r_1 and r_2 are both equal to 1, the propagating radical has no preference for either monomer. A totally random copolymer is formed. Little compositional drift occurs with conversion. When the product r_1r_2 equals 1 but r_1 and r_2 differ, a random copolymer is formed. Drift worsens with divergence of the values of r_1 and r_2. When r_1r_2 is equal to zero, an alternating copolymer results. True alternation occurs when monomers copolymerize in pairs due to ground-state or transition-state complexes. When $r_1 > 1$ and $r_2 \ll 1$, monomer one is consumed almost exclusively, and thus a homopolymer is formed. When both r_1 and r_2 are $\gg 1$, two homopolymers would form.

Alfrey and Price (see Alfrey *et al.*, 1952) proposed and later refined what is known as a *Q*,*e* scheme to predict reactivity ratios from electronic considerations. In the scheme, reactivities (polar and resonance effects) of the monomers and propagating radicals are evaluated empirically and given values *Q* and *P* respectively. The polar properties of both monomers and radicals are given *e* values based on ground-state interactions. The terms Q_1 and Q_2 are related to reactivity, and e_1 and e_2 are related to polarity of M_1 and M_2, respectively. Styrene was given standard values of $Q = 1.0$ and $e = -0.80$. Experimental reactivity ratios were then used to determine Q and e values for the other monomers. The reactivity ratios are related to Q and e according to

$$k_{11} = P_1 Q_1 \exp(-e_1^2), \tag{90}$$

$$k_{12} = P_1 Q_2 \exp(-e_1 e_2), \tag{91}$$

$$k_{22} = P_2 Q_2 \exp(-e_2^2), \tag{92}$$

$$k_{21} = P_2 Q_1 \exp(-e_1 e_2), \tag{93}$$

$$r_1 = \frac{k_{11}}{k_{12}} = \frac{Q_1}{Q_2} \exp[-e_1(e_1 - e_2)], \tag{94}$$

$$r_2 = \frac{k_{22}}{k_{21}} = \frac{Q_2}{Q_1} \exp[-e_2(e_2 - e_1)]. \tag{95}$$

Tabulated values of Q and e (Table 5) can be used to predict copolymer composition at given feed ratios. Values of r_1 and r_2 are first calculated according to Eqs. (94) and (95). These values are then substituted into the copolymerization equation [Eq. (89)].

Table 5. *Q,e* values for selected monomers (adapted from Seymour and Carraher, 1981).

Monomers	*Q*	*e*
Benzyl methacrylate	3.64	0.36
Methacrylic acid	2.34	0.65
2-Fluoro-1,3-butadiene	2.08	−0.43
p-Cyanostyrene	1.86	−0.21
p-Nitrostyrene	1.63	0.39
2,5-Dichlorostyrene	1.60	0.09
Methacrylamide	1.46	2.24
p-Methoxylstyrene	1.36	−1.11
2-Vinylpyridine	1.30	−0.50
p-Methylstyrene	1.27	−0.98
2-Vinylnaphthalene	1.25	−0.38
Isopropyl methacrylate	1.20	−0.15
Methacrylonitrile	1.12	0.81
p-Bromostyrene	1.04	−0.32
Styrene	1.00	0.80
m-Methylstyrene	0.91	−0.72
n-Amyl methacrylate	0.82	−0.32
Methyl methacrylate	0.74	0.40
Acrylonitrile	0.60	1.20
Methyl acrylate	0.42	0.60
Vinylidene chloride	0.23	0.36
Vinyl chloride	0.044	0.20
Vinyl acetate	0.026	−0.22

The *Q,e* scheme sometimes fails despite its overwhelming success for predicting reactivity of several monomer pairs. The assignment of *e* values for both the monomer and its radical is not adequate and in fact not consistent with transition-state theory. Additionally, values of *Q* and *e* are actually determined from reactivity ratios that vary greatly with experimental conditions (solvent, *p*H, concentration, and temperature). The assumptions of a steady-state concentration of each type of radical and reactivity considerations for only the end unit on a propagating radical are often violated. Finally, solvation and chain effects have been ignored in the derivation.

A number of techniques have been used to predict *copolymer microstructure,* in other words, the statistical distribution of *A* and *B* monomers along the backbone from reactivity ratios. Although beyond the scope of this monograph, statistical methods have been developed to predict sequence-length distribution of dyads, triads, pentads, and longer sequences.

Most random and alternating copolymers are prepared in free-radical reactions. Controlled graft copolymers can be formed by including sites along or pendent to a polymer backbone that can be converted to free-radical species for further monomer addition.

Commercial examples of random copolymers are poly(vinylchloride–co-vinylacetate), poly(vinylchloride–co-acrylonitrile), and poly(styrene–co-butadiene). Block copolymers are generally prepared by ionic techniques. However, block copolymers with segments or domains of random length can be produced by thermal, mechanical, or ultrasonic degradation and recombination of mixtures of the respective homopolymers. Examples are block copolymers prepared from polyethylene and polypropylene.

2.2.6 Anionic Polymerization Anionic polymerization follows the general scheme shown in reactions (45)–(48) in Sec. 2.2. However, several features distinguish the method from free-radical polymerization. First of all, a nucleophilic species is involved in all steps rather than a free radical. Since anions have little tendency to couple, concentrations of the initiating and propagating ends are much higher, 10^{-3} as compared with 10^{-8} for free radicals. Kinetics of the reaction are easily followed by monitoring the propagation step [reaction (47)]. In cases in which addition of the nucleophile to monomer is fast compared with propagation, the propagating chains start at approximately the same time. Under conditions where termination and chain-transfer reactions are slow or negligible, narrow molecular-weight distributions are obtained.

In anionic polymerization, the propagating center has an unshared pair of electrons with a charge of −1. Associated with the anion is a cation called the *counterion* or *gegenion.* The reactivity of this ion pair is determined by the nature of the functional groups to which it is attached and solvation. These factors along with temperature determine propagation rates and stereochemical control of addition.

Monomers that propagate anionically have transition states stabilized by electron-withdrawing or delocalizing groups,

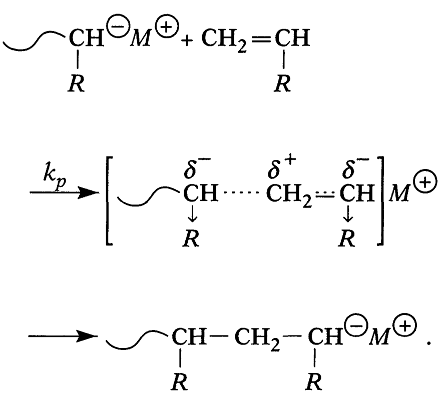

$$\sim\underset{R}{\underset{|}{CH}}^{\ominus}M^{\oplus} + CH_2{=}\underset{R}{\underset{|}{CH}} \xrightarrow{k_p} \left[\sim\underset{R}{\underset{\downarrow}{\overset{\delta^-}{CH}}}\cdots\overset{\delta^+}{CH_2}{=}\underset{R}{\underset{\downarrow}{\overset{\delta^-}{CH}}}\right]M^{\oplus} \longrightarrow \sim\underset{R}{\underset{|}{CH}}-CH_2-\underset{R}{\underset{|}{CH}}^{\ominus}M^{\oplus}. \quad (96)$$

R groups with appropriate electronic characteristics include

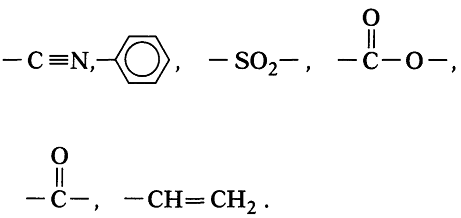

A number of anionic polymerizations involve additions to carbonyl carbons,

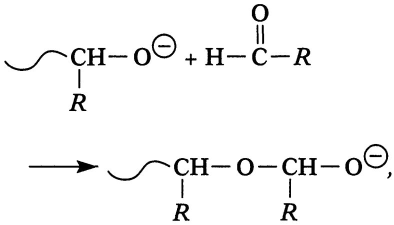

$$\sim\underset{R}{\underset{|}{CH}}-O^{\ominus} + H-\overset{O}{\overset{\|}{C}}-R \longrightarrow \sim\underset{R}{\underset{|}{CH}}-O-\underset{R}{\underset{|}{CH}}-O^{\ominus}, \quad (97)$$

or ring opening of epoxides,

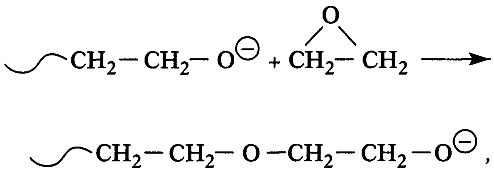

(98)

sulfides,

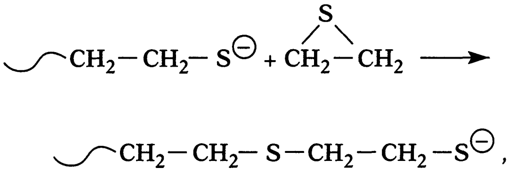

(99)

lactams,

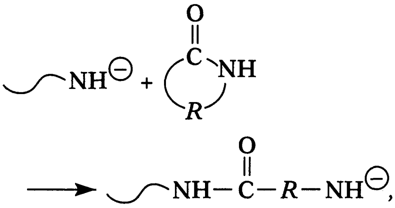

(100)

and lactones,

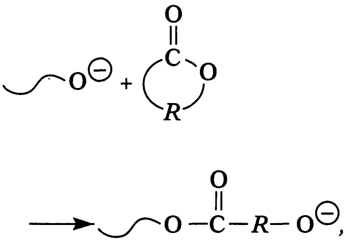

(101)

in the propagation step.

Two general mechanisms for initiation of unsaturated monomers are illustrated for substituted alkenes. The first involves nucleophilic addition to the double bond to yield a carboanion and its associated cation:

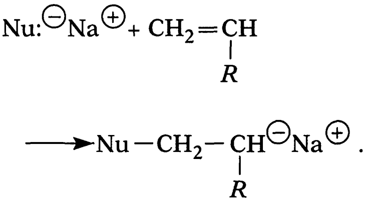

$$\mathrm{Nu{:}^{\ominus}Na^{\oplus} + CH_2{=}\underset{R}{CH}} \longrightarrow \mathrm{Nu{-}CH_2{-}\underset{R}{CH}{}^{\ominus}Na^{\oplus}}\,. \tag{102}$$

The second general mechanism involves an electron-transfer reaction to create a radical anion,

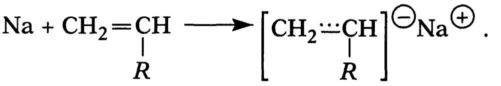

$$\mathrm{Na + CH_2{=}\underset{R}{CH}} \longrightarrow \left[\mathrm{CH_2{\overset{\cdot}{-}}\underset{R}{CH}}\right]^{\ominus}\mathrm{Na^{\oplus}}\,. \tag{103}$$

Dimerization through radical coupling yields the dianion for subsequent addition of monomer to both ends:

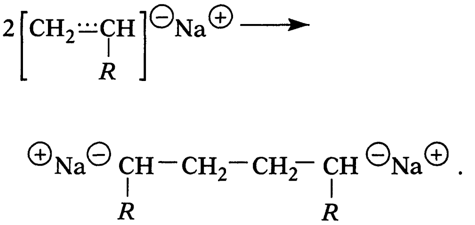

$$2\left[\mathrm{CH_2{\overset{\cdot}{-}}\underset{R}{CH}}\right]^{\ominus}\mathrm{Na^{\oplus}} \longrightarrow {}^{\oplus}\mathrm{Na}^{\ominus}\,\mathrm{\underset{R}{CH}{-}CH_2{-}CH_2{-}\underset{R}{CH}}\,{}^{\ominus}\mathrm{Na^{\oplus}}\,. \tag{104}$$

Initiators commonly used include organometallics, metal amides, Grignard reagents, metal alkoxides or cyanides, and alkali metals.

The dielectric constant of the solvent, counterion nature, and temperature are often manipulated to control reaction rate and stereochemistry (tacticity) during propagation of a given monomer:

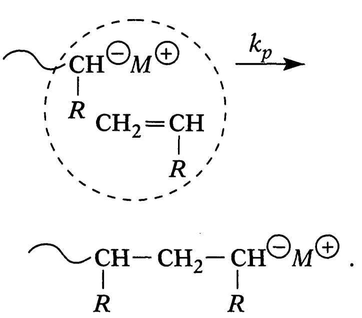

$$\sim\!\mathrm{\underset{R}{CH}{}^{\ominus}}M^{\oplus}\ \ \mathrm{CH_2{=}\underset{R}{CH}} \xrightarrow{k_p} \sim\!\mathrm{\underset{R}{CH}{-}CH_2{-}\underset{R}{CH}{}^{\ominus}}M^{\oplus}\,. \tag{105}$$

The degree of dissociation of the propagating ion pair will determine the rate of the reaction as well as the tendency for complexation of the adding monomer. For solvent-dissociated ion pairs, the counterion is not intimately involved and the approach of monomer is usually dictated by the bulkiness of the R groups on the propagating chains. Syndiotactic placements (Sec. 2.2.1) are favored by their lower activation-energy requirements. Complexation mechanisms occur at low solvent polarity, often resulting in isotactic placements. The counterion is usually involved in direct chelation with the approaching monomer. Solvents commonly utilized for anionic polymerization range from nonpolar alkenes and aromatics to the more polar ethers, dioxane, and tetrahydrofuran. Perhaps one of the most attractive attributes of anionic polymerization in nonterminating conditions is the ability to control the chain length and reaction rate by considering only the propagation step

$$R_p = k_p[M^-][M] \tag{106}$$

where M^- represents the growing anion. This condition, called *living polymerization,* is usually met in polar solvents in which the counterion is dissociated at 100% conversion; the degree of polymerization is obtained by simply dividing the total monomer concentration by the initiator concentration. More complex expressions for ion-pair equilibria in moderately polar solvents have been derived (Szwarc, 1974). Since the propagating site is not destroyed during living polymerization by transfer or termination reactions, sequential addition of monomers is possible and controlled values of M_n and narrow molecular-weight distributions are attainable.

For an extensive review of various aspects of anionic polymerization and copolymerization, the reader is referred to Eastmond *et al.* (1989a), Chaps. 25–38.

2.2.7 Star and Telechelic Polymers by Anionic Polymerization Living-polymerization techniques are most useful for sequential addition to yield block copolymers,

$$R-M \xrightarrow{nA} R-A-A-A-A-A-A-A^{\ominus}M^{\oplus}$$

$$\xrightarrow{nB} R-A-A-A-A-A-A-A-B-B-B-B-B-B-B, \qquad (107)$$

or star polymers,

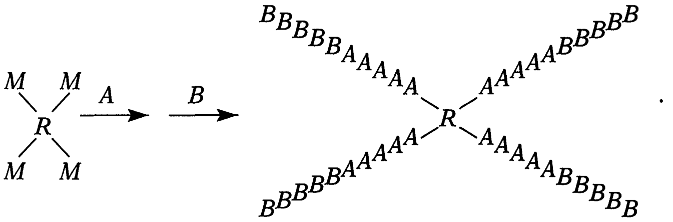
M
M
M
M
R
A
B
R
A
B

(108)

Living anionic polymerization can be used to form *telechelic* derivatives. For example, carbon dioxide may be added to polystyrene end groups to yield a difunctional carboxylated polymer,

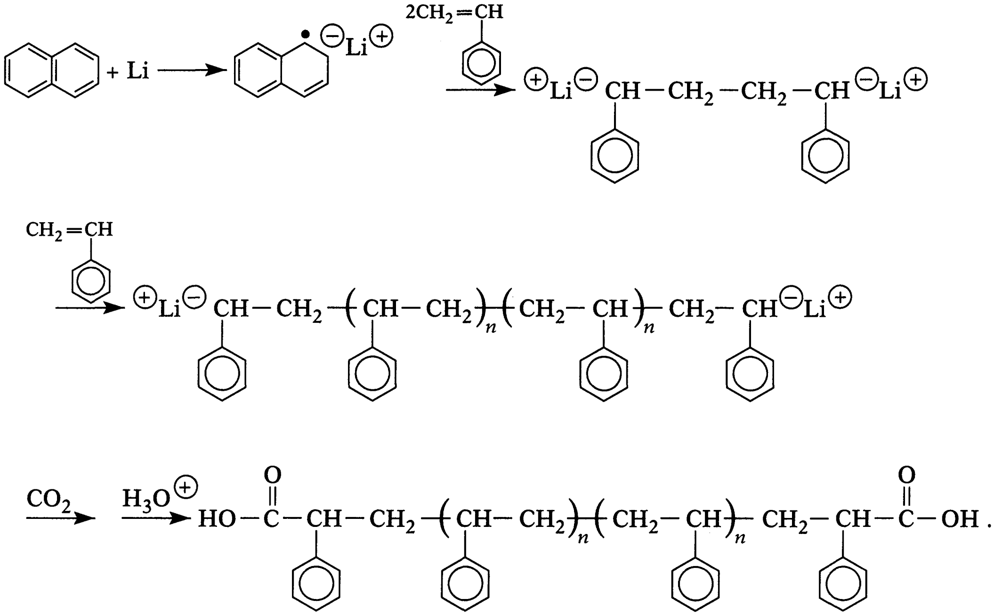
+ Li
Li
2CH₂=CH
Li
CH–CH₂–CH₂–CH
Li
CH₂=CH
Li
CH–CH₂
CH–CH₂
n
CH₂–CH
n
CH₂–CH
Li
CO₂
H₃O
HO–C–CH–CH₂
O
CH–CH₂
n
CH₂–CH
n
CH₂–CH–C–OH.
O

(109)

Likewise hydroxylated polymer can be formed by adding ethylene oxide. Terminal double bonds can be obtained by adding allylhaides.

Star polymers can be obtained by coupling living end groups with multifunctional reagents such as $SiCl_4$:

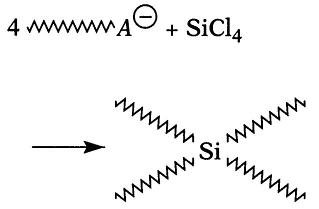

(110)

Specialty polymers with interesting morphological structures have been made by anionic techniques. For example, block copolymers of styrene with butadiene or isoprene form thermoplastic elastomers. Telechelic polymers have found application as membranes or as potential chelating sites in ionomer resins.

2.2.8 Cationic Polymerization Cationic polymerization follows the general scheme outlined in Sec. 2.2. The propagating species is generally a carbocation with its associated gegenion. The reaction is initiated by addition of a cationic species to monomer. Table 6 gives examples of representative initiators used.

The propagation step is favored by monomers that can stabilize the developing carbocation in the transition state:

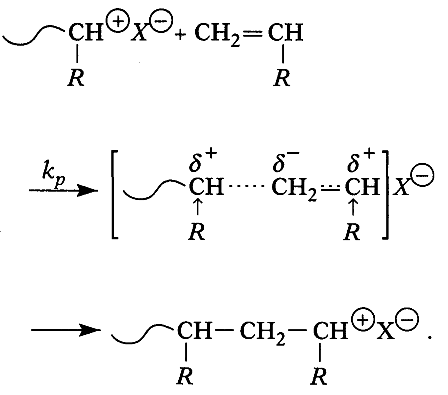

(111)

Therefore, electron-donating *R* groups, those with unshared pairs of electrons, or aromatic rings, which can delocalize the positive center in the transition state, increase the rate of addition. The rate of addition depends also upon the counterion type, solvent, and reaction temperature. Narrow molecular-weight distributions can only be obtained if the rate of initiation is faster than that of propagation, allowing the chains to start at the same time. Also chain-transfer reactions must be avoided.

Catonic polymerization can be "living" if termination reactions are not facile. As with anionic polymerization, polar solvents favor propagation by the "free" ion while nonpolar solvents would favor ion-pair propagation. Stereoregular polymers are possible under both conditions. In the former, steric factors along the chain would dictate tacticity, while complexation of incoming monomer can control direction of attack and tacticity in the latter.

Rearrangement during polymerization is a

Table 6. Cationic initiators.

Type	Examples	Initiating species	Counterion
Protic acids	H_2SO_4	$H^{\oplus}$	$HSO_4^{\ominus}$
	$HOCOCF_3$	$H^{\oplus}$	$OCOF_3^{\ominus}$
Lewis-acid conjugates	$H_2O/ALCl_3$	$H^{\oplus}$	$ALCl_3OH^{\ominus}$
	H_2O/BF_3	$H^{\oplus}$	$BF_3OH^{\ominus}$
	$(CH_3)_3CCl/Et_2AlCl$	$(CH_3)_3C^{\oplus}$	$Et_2AlCl_2^{\ominus}$1018
Other	CH_3COClO_4	$CH_3CO^{\oplus}$	$ClO_4^{\ominus}$
	$(C_6H_5)_3C^{\oplus}\ SbCl_6^{\ominus}$	$(C_6H_5)_3C^{\oplus}$	$SbCl_6^{\ominus}$

common occurrence in cationic polymerization. The carbocation first formed upon addition during propagation can rearrange by a hydride shift to the more stable carbocation that is still capable of further addition to monomer. If the rearrangement is faster than addition, pathway (112b) rather than (112a) is followed:

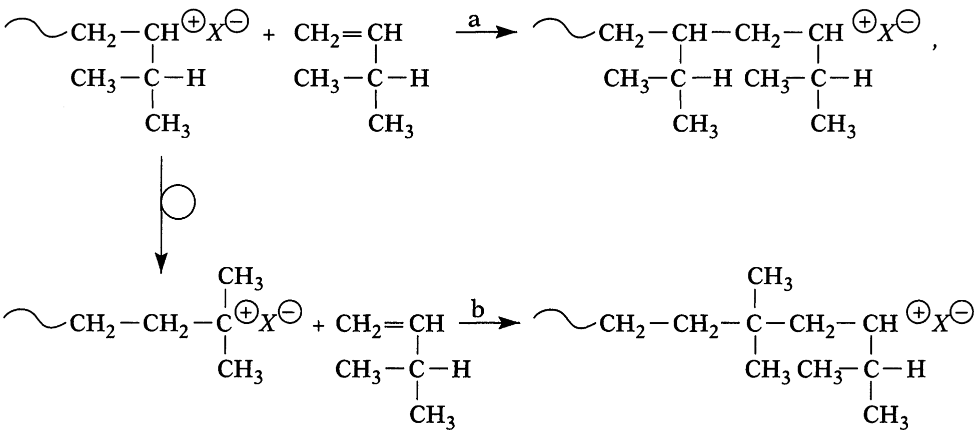

(112a)

(112b)

The major cationic polymerization with present technological importance is that of low-temperature polymerization of isobutylene to give butyl rubber. For detailed reviews of carbocationic polymerization and copolymerization, the reader is referred to Eastmond *et al.* (1989a), Vol. 3, Chaps. 39–53.

2.2.9 Cationic Copolymerization The number of monomer pairs that undergo random copolymerization by cationic initiation is much smaller than the number that undergo random radical copolymerization. There is also little tendency for alternation; usually block polymers or simply two homopolymers are formed in mixed-monomer systems.

The most useful cationic copolymers are formed by block polymerization, usually by sequential addition of monomers starting with the most reactive. Although the procedures can yield well-defined blocks with interesting morphologies, and can, like anionic polymerization, be used to prepare telechelic copolymers, rigorous purity and reaction conditions must be maintained. This currently limits commercial utility to specialty materials.

2.2.10 Coordination Polymerization *Coordination polymerization* or *polyinsertion* reactions involve a coordination of the monomer with the active sites on transition-metal complexes. The propagation proceeds via monomer insertion into the kinetic chain, leaving a vacant coordination site for complexation of the next monomer and repetition of this cycle.

Coordination polymerization is the commercial method by which most polyolefins including cyclic polyolefins are prepared. This represents almost half of the total industrial production of polymers. Typical of the initiation systems (often termed polymerization "catalysts") are the Ziegler catalysts derived from transition metals of groups IV–VIII with alkyl, aryl, or hydride metal compounds of groups I–III. A typical catalyst of this type is that prepared from $Al(C_2H_5)_3$ and $TiCl_3$. Equilibria reactions occur in which the chloride and ethyl groups can exchange on the metals. The metal complex active in polymerization may involve either the titanium only (monometallic complex) or both the titanium and aluminum (bimetallic complex). For simplicity, the monometallic complex mechanism with a single ethyl group is shown in

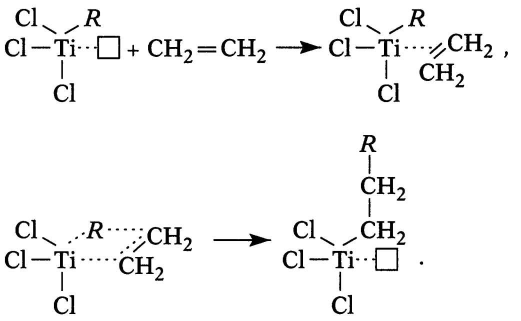

(113)

An empty *d* orbital on the transition metal (represented by □) is involved in overlap with the π system of the monomer.

When α olefins such as propylene are used, the insertion reaction can be stereoselective. With the titanium catalyst, the product is highly isotactic and the mechanism appears to be one of *catalyst site control* rather than *chain control* discussed earlier for radical and ionic polymerization. In the bimetallic mechanism, a similar pathway is followed; however, the Ti and Al are believed to "share" the alkyl chain at a coordination boundary.

The specificity of the addition is brought about by ligand coordination—in this case by the overlap of the empty *d* orbitals of the transition metal with the π orbitals of the alkene. The polyalkenes prepared in this manner have more linear structures and, therefore, can crystallize in more dense packing arrangements than the branched, low-density polyolefins prepared by free-radical polymerization. In the latter, chain transfer via a "back-biting" mechanism competes with propagation and leads to significant branching.

Both radical and ionic coordination mechanisms have been proposed for Ziegler–Natta-type polymerization (Odian, 1991, Chap. 8). Most experimental evidence gathered for polymerization of ethylene, propylene, styrene, and 1,3-dienes seems to indicate a coordinated anionic insertion (Boor, 1979). The degree of stereoregularity and the microstructure (syndiotactic vs isotactic) is dependent upon the type of the metal complex, the nature of the α olefin, the reaction temperature, and whether the catalyst is soluble or dispersed. For example, the catalyst $TiCl_4/Et_3Al$ yields largely isotactic polypropylene whereas VCl_4/Et_2AlCl yields highly syndiotactic polypropylene.

Ziegler–Natta polymerizations have some of the characteristics of "living" polymerization in that the active sites may continue to produce polymer for long periods of time. However, several individual chains may be produced during that time. The propagating chains may terminate by a β-hydride transfer to the transition metal or chain transfer to monomer. The chain may also be terminated by impurities or intentionally added reagents to "poison" the catalyst sites.

Coordination complexes formed by transition metals and alkylmetal or Lewis acids, termed metathesis catalysts, have been used for ring-opening polymerizations of cycloalkenes and bicycloalkenes and for polymerization of alkynes. The most widely used transition-metal components of the catalysts are tungsten, molybdenum, and ruthenium. Metathesis polymerization chemistry and applications are detailed in reports by Amass (1989) and Feast (1989).

Coordinated anionic polymerization termed group-transfer polymerization has been utilized successfully to prepare acrylic polymers such as methyl methacrylate (Sogah *et al.*, 1987).

3. REACTIONS OF FUNCTIONAL POLYMERS

For extensive treatments, the reader is referred to Eastmond *et al.* (1989), Seymour and Carraher (1981), Stevens (1975), and Elias (1977).

3.1 Functional-Group Reactivity

In theory, polymers with functional groups along or pendent to the macromolecular backbone should undergo reactions analogous to those of small molecules. In practice, the reaction efficiency may be much lower because of kinetic effects or thermodynamic problems related to conformational entropy changes on transformation. Major factors affecting orientation and collision frequency in polymer transformation involving isolated functional groups are solubility, chain mobility, steric and electronic effects, and neighboring-group effects. In many cases, "spacers" between the polymer backbone and the reactive functionality are required for reasonable reaction rates and conversions. On occasion, however, the polymer and its associated solvent can provide "domains" for enhanced reactions. These generally involve favorable environments in which neighboring-group assistance or local dielectric-constant changes aid in reaction kinetics. Often transformation reactions on polymers can effectively create catalytic sites in which phase-transfer reactions may be conducted on molecules held temporarily at the reaction sites.

Transformation reactions on polymers can result in systems with commercial utility. For example, cellulose, the most abundant polymer from renewable resources, can be transformed into a wide range of materials varying from membranes to water-soluble thickeners.

As with small organic molecules, chemical transformation reactions often yield by-products. These competing reactions often result in substituents permanently attached to the macromolecular backbone that can be detrimental to desired properties. Some reactions can only proceed to low conversion before precipitation occurs, and functional groups are no longer accessible. Finally, undesirable reactions may occur on polymers over a period of time such as degradation by light, oxygen, ozone, heat, or moisture.

In the sections to follow, classes of reactions representative of the myriad of polymer transformation will be presented.

3.2 Cross-Linking Reactions

Chemical cross linking through covalent, ionic, polar, or hydrophobic interactions is technologically important in developing networks that significantly alter macroscopic properties of polymers. Examples of networks are photo- or thermally cured coatings, thermoset plastics, elastomers, ionomers, and hydrogels. Since covalent network formation limits chain mobility, and thus processability, polymers of this type are formed either in one step or by preparation of linear prepolymers and subsequent cross linking. (See Sec. 2.1.8.)

Polymers capable of noncovalent network formation are usually processed from the melt or solution. Network formation occurs in response to the cooling or solvent removal. Elastomers, such as polystyrene–polyisoprene block copolymers (Sec. 2.2.7), multiblock polyurethane elastomers (Sec. 2.1.7), and telechelic polymers (Sec. 2.2.7), are example systems forming thermally reversible domains. A classical example of the cross linking of elastomers or *vulcanization* is the formation of networks from residual double bonds in polyisoprene utilizing sulfur.

3.3 Reactions of Saturated Polyolefins

In the strictest sense, polymers such as polyethylene are functional polymers in that the pendent hydrogen atoms can be substituted for. Polyethylene and other saturated polyolefins can be reacted with the halides fluorine, chlorine, or bromine in the presence of light and heat to afford partial halogenation. Competing reactions of dehydrohalogenation and chain scission must be minimized.

3.4 Reactions of Unsaturated Polyolefins

Reactions of the double bonds in alkenes along or pendent to the polymer backbone can be useful in covalent cross linking (Sec. 2.1.8), in forming graft copolymers (Sec. 1), or for changing macroscopic properties of the polymer. The simplest transformation reaction of polyisoprene is hydrogenation,

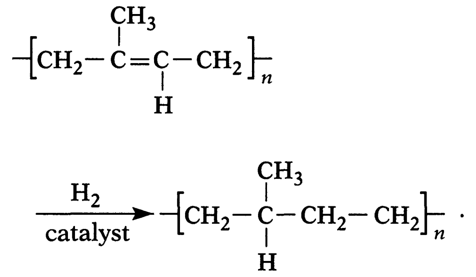

$$-\!\left[CH_2-\overset{CH_3}{\overset{|}{C}}=\underset{H}{\underset{|}{C}}-CH_2\right]_n\!- \xrightarrow[\text{catalyst}]{H_2} -\!\left[CH_2-\overset{CH_3}{\underset{H}{\overset{|}{\underset{|}{C}}}}-CH_2-CH_2\right]_n\!- \ .$$

(114)

The commercially important block copolymer Kraton® is produced by hydrogenation of the ABA triblock styrene–butadene–styrene. The double bonds of polyisoprene in linear or block copolymers can also be chlorinated.

Epoxides are easily formed by the reaction of peracetic acid or hydrogen peroxide with residual double bonds:

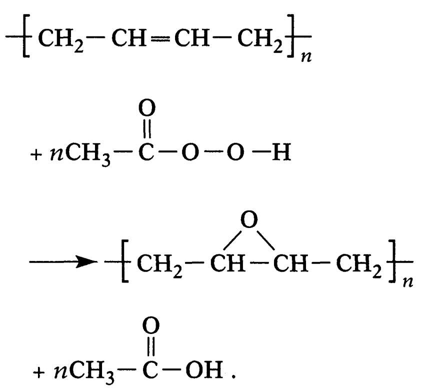

(115)

The epoxidized intermediates are easily transformed by the action of nucleophiles such as water, amines, or alcohols to functional derivatives.

Some polymers are susceptible to a number of adverse reactions, especially those with easily abstractable allylic hydrogen atoms and double bonds. Degradation reactions at moderate temperatures such as oxidation and ozonolysis rapidly break down the polymer backbone.

3.5 Reactions of Halogenated Polymers

Substitution reactions of halogens directly on the polymer backbone are generally not successful because of steric requirements and/or competing dehydrohalogenation and degradation reactions. However, halides positioned farther from the backbone, and especially those attached to benzylic or allylic carbon atoms, are easily substituted for. An example is

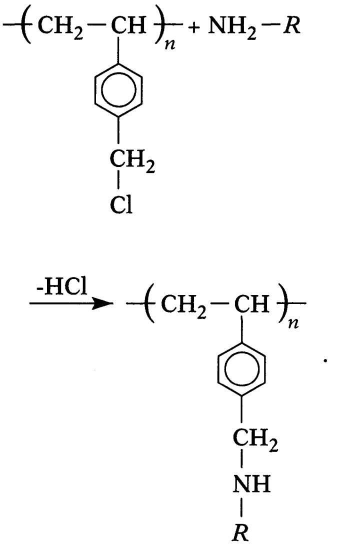

(116)

for amination of the polymerized para-substituted benzylchloride derivative of styrene.

Halogenated benzylic styrene derivatives are utilized to prepare supports upon which stepwise synthesis of peptides, nucleic acids, and polysaccharides can be conducted. In the Merrifield synthesis of polypeptides, for example, a carboxyl-protected amino acid is introduced onto a cross-linked chloromethylated polystyrene resin. The protective group is removed and another carboxy-protected amino acid is added. In this way, amino acids can be sequentially added:

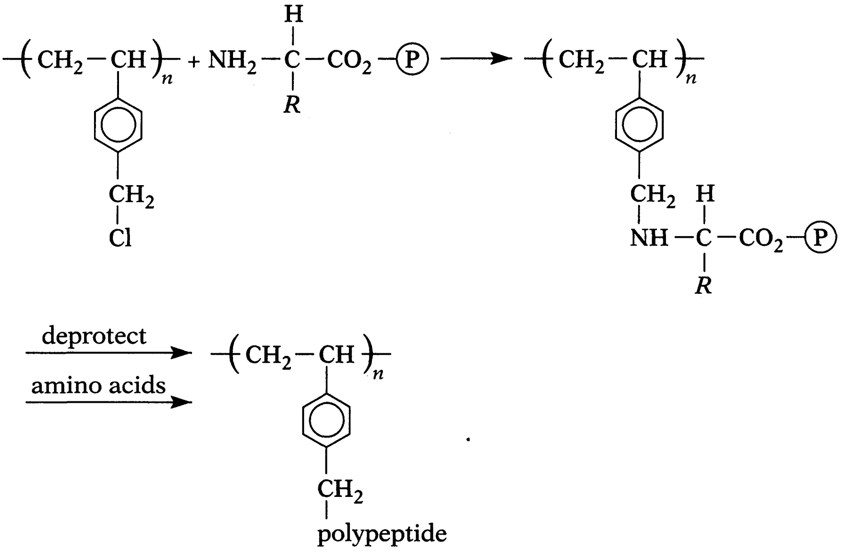

(117)

The polymer support allows facile synthesis, separation, and purification. The peptide is isolated in the last step by release from the solid support. Synthesis on solid supports has recently become important for production of polynucleotides, RNA, and DNA, utilized in recombinant biotechnology.

Halogenated aromatic polymers with aryl or benzyl substitution are also useful in electrophilic aromatic substitution:

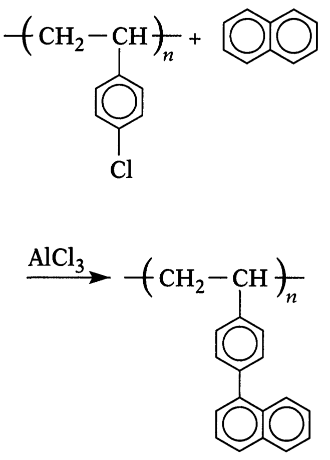

(118)

Friedel–Craft-type reactions occur readily with Lewis-acid catalysts. Competitive ring formation and intermolecular self-substitution reactions may occur along or between polymer chains.

3.6 Reactions of Hydroxy- or Amine-Containing Polymers

Polyesters or polyamides can be obtained by reacting hydroxy- or amine-functional polymers through addition-elimination reactions involving acyl acids, halides, or anhydrides as shown in

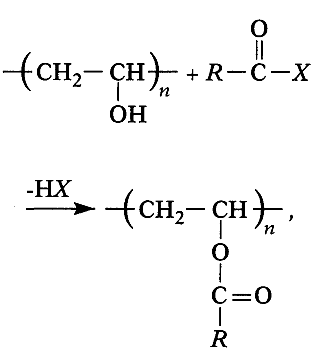

$$\left(\mathrm{CH_2}-\underset{\mathrm{OH}}{\mathrm{CH}}\right)_n + R-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}-X \xrightarrow{\mathrm{-H}X} \left(\mathrm{CH_2}-\underset{\substack{|\\ \mathrm{O}\\ |\\ \mathrm{C=O}\\ |\\ R}}{\mathrm{CH}}\right)_n ,$$

(119)

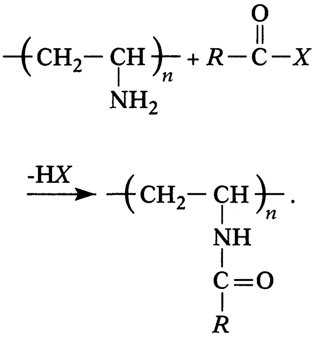

$$\left(\mathrm{CH_2}-\underset{\mathrm{NH_2}}{\mathrm{CH}}\right)_n + R-\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}-X \xrightarrow{\mathrm{-H}X} \left(\mathrm{CH_2}-\underset{\substack{|\\ \mathrm{NH}\\ |\\ \mathrm{C=O}\\ |\\ R}}{\mathrm{CH}}\right)_n .$$

(120)

Carbonate and urea derivatives can be prepared by reactions with isocyanates:

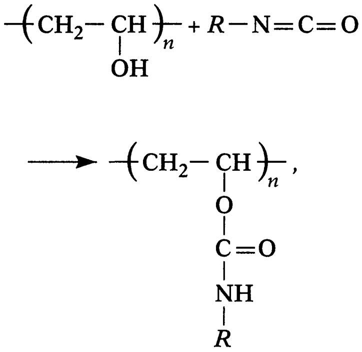

$$\left(\mathrm{CH_2}-\underset{\mathrm{OH}}{\mathrm{CH}}\right)_n + R-\mathrm{N=C=O} \longrightarrow \left(\mathrm{CH_2}-\underset{\substack{|\\ \mathrm{O}\\ |\\ \mathrm{C=O}\\ |\\ \mathrm{NH}\\ |\\ R}}{\mathrm{CH}}\right)_n ,$$

(121)

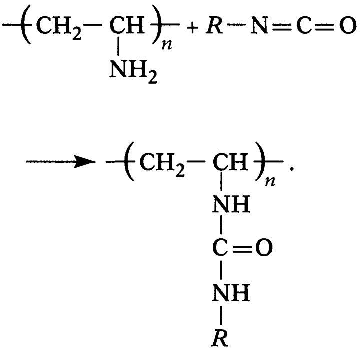

$$\left(\mathrm{CH_2}-\underset{\mathrm{NH_2}}{\mathrm{CH}}\right)_n + R-\mathrm{N=C=O} \longrightarrow \left(\mathrm{CH_2}-\underset{\substack{|\\ \mathrm{NH}\\ |\\ \mathrm{C=O}\\ |\\ \mathrm{NH}\\ |\\ R}}{\mathrm{CH}}\right)_n .$$

(122)

Polymers with side-chain ethers are prepared by reacting alkyl halides,

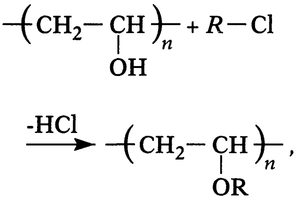

$$\left(\mathrm{CH_2}-\underset{\mathrm{OH}}{\mathrm{CH}}\right)_n + R-\mathrm{Cl} \xrightarrow{\mathrm{-HCl}} \left(\mathrm{CH_2}-\underset{\mathrm{O}R}{\mathrm{CH}}\right)_n ,$$

(123)

or epoxides with hydroxy functionality,

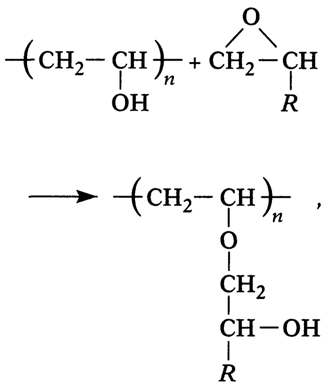

(124)

and by nucleophilic substitution reactions.

Hydroxy-containing polymers can undergo addition reactions with aldehydes to form methylol groups. Cyclization can occur, as shown in

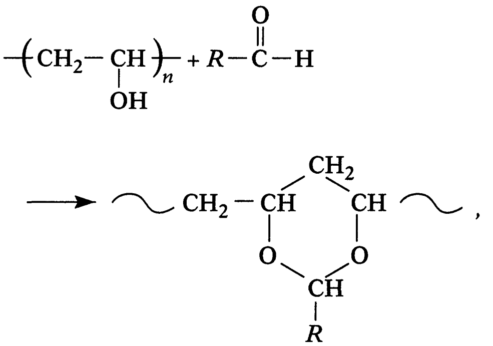

(125)

for the reaction of poly(vinyl alcohol) with aldehydes, especially when six-membered rings are possible. The product formed by the reaction of butyraldehyde, $R = C_3H_7$, which contains over 80% cyclized repeating units, is used commercially as lamination in safety glass.

Other reactions of hydroxy functionality on polymers are illustrated by alkene addition (Michael addition) of acrylonitrile,

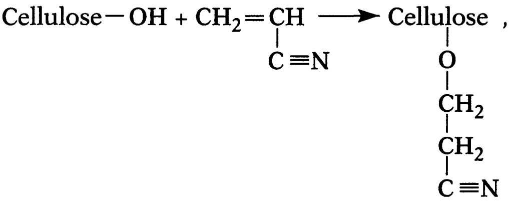

$$\text{Cellulose}-OH + CH_2=CH(C{\equiv}N) \longrightarrow \text{Cellulose}-O-CH_2-CH_2-C{\equiv}N \quad (126)$$

and nitration of cellulose,

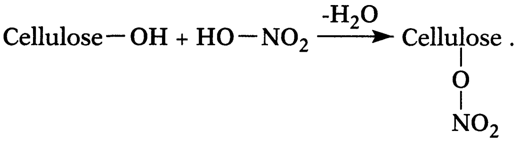

$$\text{Cellulose}-OH + HO-NO_2 \xrightarrow{-H_2O} \text{Cellulose}-O-NO_2 \quad (127)$$

The nitrile derivative can be hydrolyzed to yield water-absorbent cellulose, while nitrocellulose is used in coatings.

Carboxylated celluloses such as carboxymethyl cellulose may be prepared by reacting cellulose in caustic solution with sodium chloroactetate followed by neutralization:

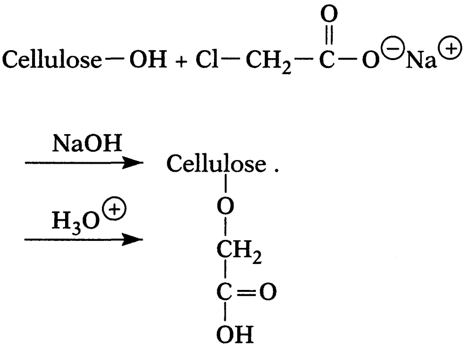

$$\text{Cellulose}-OH + Cl-CH_2-\overset{O}{\overset{\|}{C}}-O^{\ominus}Na^{\oplus} \xrightarrow{NaOH} \xrightarrow{H_3O^{\oplus}} \text{Cellulose}-O-CH_2-C(=O)-OH \quad (128)$$

3.7 Reaction of Acyl Groups

Polymers with pendent esters, acids, anhydrides, or acyl halides can undergo facile addition-elimination reactions with added nucleophiles:

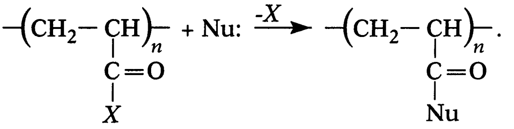

$$-\!\!\left(CH_2-\underset{\underset{X}{|}}{\underset{C=O}{|}}{CH}\right)_{\!n}\!\!- + \text{Nu:} \xrightarrow{-X} -\!\!\left(CH_2-\underset{\underset{\text{Nu}}{|}}{\underset{C=O}{|}}{CH}\right)_{\!n}\!\!- . \quad (129)$$

Common nucleophiles include water, alcohols, amines, or acids to yield polymeric acids, esters, amides, or anhydrides, respectively.

Representative of a common reaction is that of hydrolysis of the alternating copolymer prepared by chain-growth polymerization of maleic anhydride and styrene with water rendering the copolymer soluble in water:

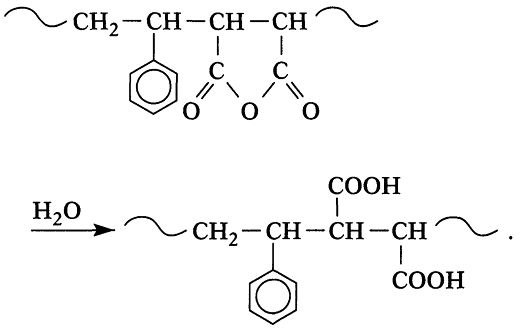

(130)

3.8 Controlled Cyclization Reactions

In addition to cyclopolymerization (Sec. 2.2.1) and polyacetal cyclization reactions [Eq. (125)], ring formation has been attempted for polymers with pendent functional groups to yield polymers with "ladder"-type structures. True "ladders" are rarely obtained, however, since isolated pendent groups are statistically predicted. Examples of ladder structures are given for the intramolecular alcohol condensation of poly(methylvinylketone),

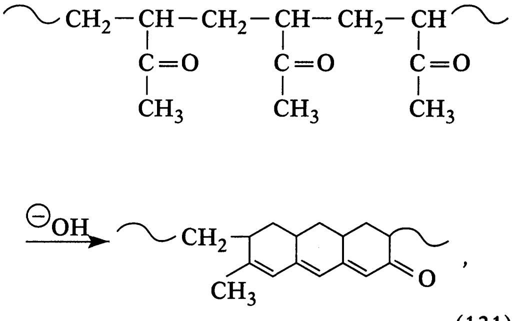

(131)

and for the controlled heating of polyacrylonitrile,

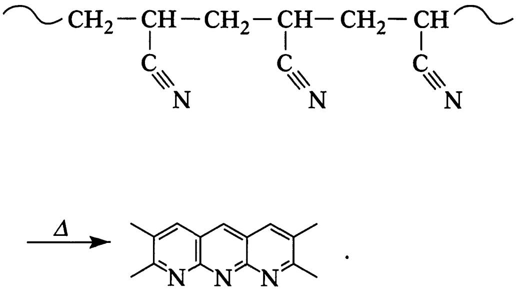

(132)

3.9 Functionalization for Specific Activity

The wide variety of functionalized polymers formed from step-growth or chain-growth processes allows synthesis of a virtually unlimited array of materials capable of serving as reagents, catalysts, sequestrants, phase-transfer sites, stabilizers, or substrates for delivery of bioactive agents. For example, polymeric reagents have been used for hydrogenation of alkenes, epoxidation reactions, stereoselective reduction of aldehydes, and oxidation/reduction reactions. Polymeric catalysts have been used for hydrogenation, hydrolysis, esterification, amidation, and photoconversion reactions. Transition metals have been chelated to functional polymers through phosphenation reactions to serve as catalysts in hydrosilation, hydroformylation, and hydrogenation reactions. Photosensitizers have been bound to polymeric centers to assist in photochemical reactions. Bioactive agents including drugs and pesticides have been bound by covalent and ionic bonds to form controlled-activity polymers. Phase-transfer catalysts have been utilized to enhance the reaction between water-soluble and water-insoluble reactants. Finally, enzymes have been immobilized onto polymer supports for carrying out biochemical and organic syntheses.

GLOSSARY

***A–A* Monomers:** Difunctional monomers in which *A* represents the reactive functional groups.

Biopolymers: Polymers synthesized in biological systems.

***B–B* Monomers:** Difunctional monomers in which *B* represents the reactive functional groups.

Cage Recombination: Coupling of radical following formation in a confined environment.

Chain Growth: Mechanism of polymerization based on initiation, propagation, and termination.

Chain Transfer: The process by which a propagating chain is terminated by transfer of an active group from another molecular species.

Coupling: Radical chain termination by combination.

Direct Amidation: Method of formation of a polyamide by reaction of a diamine with a diacid.

Direct Esterification: Method of formation of polyesters—for example, the reaction of a difunctional alcohol with a difunctional acid.

Disproportionation: Chain termination by an abstraction elimination mechanism.

Ester Interchange: Condensation reaction in which exchange of functional groups results in ester formation.

Gegenion: Associated counterion.

Initiation: The start of a chain reaction.

Interchange Reaction: Condensation reaction characterized by exchange of functional groups.

Oligomers: Polymers with low degrees of polymerization.

Polymerization: The process of preparing polymers.

Polymers: Macromolecules composed of multiple repeating units derived from reactions of small-molecule precursors.

Propagation: Chain-extension step in chain-growth polymerization.

Ring-Opening Polymerization: Mechanism of polymerization of a cyclic monomer.

Steady State: Constant concentration of reactive species.

Step Growth: Mechanism of polymerization involving stepwise coupling reactions.

Termination: End of a chain process.

Transamidation: Amide interchange reaction common in polyamide polymerization.

Vulcanization: Controlled cross linking of unsaturated elastomers utilizing sulfur.

Works Cited

Alfrey, T., Jr., Bohrer, J. J., Mark, H. (1952), *Copolymerization,* New York: Interscience.

Amass, A. J. (1989), in: G. C. Eastmond, A. Ledwith, S. Russo, P. Sigwalt (Eds.), *Comprehensive Polymer Science,* Vol. 4, Oxford: Pergamon, Chap. 6.

Boor, J. (1979), *Ziegler-Natta Catalysts and Polymerization,* New York: Academic.

Brandrup, J. Immergut, E. H. (1975), *Polymer Handbook,* New York: Wiley.

Butler, G. B. (1989), "Cyclopolymerization," in: G. C. Eastmond, A. Ledwith, S. Russo, P. Sigwalt (Eds.) *Comprehensive Polymer Science,* Vol. 4, Oxford: Pergamon, pp. 423–451.

Carothers, W. H. (1929), *J. Am. Chem. Soc.* **51,** 2548.

Eastmond, G. C., Ledwith, A., Russo, S., Sigwalt, P. (Eds.) (1989a), *Chain Polymerization, Part I, Comprehensive Polymer Science,* Vol. 3, Oxford: Pergamon Press.

Eastmond, G., Ledwith, A., Russo, S., Sigwalt, P. (Eds.) (1989b), *Polymer Reactions, Comprehensive Polymer Science,* Vol. 6, Oxford: Pergamon.

Elias, H. G. (1977), *Macromolecules,* Vol. 1 and 2, New York: Plenum.

Feast, J. W. (1989), in: G. C. Eastmond, A. Ledwith, S. Russo, P. Wigwalt (Eds.), *Comprehensive Polymer Science,* Vol. 4, Oxford: Pergamon, Chap. 7.

Flory, P. J. (1953), *Principles of Polymer Chemistry,* Ithaca, NY: Cornell Univ. Press.

Lenz, R. W. (1967), *Organic Chemistry of High Polymers,* New York: Wiley.

Manaresi, P., Munari, A. (1989), in: G. Allen, J. C. Bevington (Eds.), *Comprehensive Polymer Science,* Vol. 5, Oxford: Pergamon, Chap. 2.

Odian, George (1991), *Principles of Polymerization,* 3rd ed., New York: Wiley.

Sogah, D. Y., Hertler, W. R., Webster, O. W., Cohen, G. M. (1987), *Macromolecules* **20,** 1473.

Seymour, R. B., Carraher, C. E., Jr. (1981), *Polymer Chemistry,* New York: Marcel Dekker.

Stevens, M. P. (1975), *Polymer Chemistry. An Introduction,* MA: Addison-Wesley.

Szwarc, M. (1974), *Ions and Ion Pairs in Polymerization,* Vol. 2, New York: Wiley, Chap. 2.

Further Reading

Allcock, H. A., Lampe, F. W. (1981), *Contemporary Polymer Chemistry,* Englewood Cliffs, NJ: Prentice-Hall.

Allen, J. C., Bevington, J. C. (Eds.) (1989), *Comprehensive Polymer Science,* Collective Vol. 1–7, Oxford: Pergamon.

Billmeyer, F. W., Jr. (1962), *Textbook of Polymer Science,* 2nd ed., New York: Wiley.

Elias, Hans-Georg (1977), *Macromolecules,* Vols. 1 and 2, New York: Plenum.

Flory, P. J. (1953), *Principles of Polymer Chemistry,* Ithaca, NY: Cornell Univ. Press.

Jenkins, J. D., Ledwith, A. (1974), *Reactivity, Mechanism and Structure in Polymer Chemistry,* New York: Wiley.

Lenz, R. W. (1967), *Polymer Chemistry of Synthetic High Polymers,* New York: Wiley.

Mark, H. F., Bikales, N. M., Overberger, G. C., Menges, G., Kroschwitz, J. I. (1985–1989), *Encyclopedia of Polymer Science and Engineering,* 2nd ed., Vols. 1–7 and supplemental volumes, New York: Wiley.

Odian, G. (1991), *Principles of Polymerization,* 3rd ed., New York: Wiley.

Vollmert, Bruno (1973), *Polymer Chemistry,* New York: Springer-Verlag.

POLYMERS: CRYSTAL STRUCTURE AND MORPHOLOGY

Arthur E. Woodward, *Department of Chemistry, City College, City University of New York, New York, New York, U.S.A.*

Introduction 477
1. **Polymer Crystal Structure** 478
1.1 Chain Regularity 478
1.2 Chain Conformation, Orientation, and Packing 480
1.3 Chain Folding 480
1.4 Crystallization Methods 481
1.4.1 Crystallization from Solution 481
1.4.2 Crystallization from the Melt 481
1.4.3 Crystallization under Mechanical Stress 481
1.5 Crystal Structure Determination 481
2. **Polymer Morphology** 482
2.1 Methods of Investigation 482
2.1.1 Optical Microscopy 482
2.1.2 Electron Microscopy 482
2.1.3 Scanning Probe Microscopy 482
2.2 Morphologies of Crystallized Polymers 483
2.2.1 Single Lamellas 483
2.2.2 Dendrites, Hedrites, and Spherulites 483
2.2.3 Fibers 485
2.2.4 Epitaxial Growths 485
2.2.5 Dislocations and Defects 485
2.3 Liquid-Crystalline Mesophases 486
2.3.1 Types of Mesogens 486
2.3.2 Mesophase Textures 487
2.3.3 Disclination Imaging 487
2.4 Phase-Separated Block Copolymers 488
2.4.1 *AB*, and *ABA*, and Star-Block Types 488
2.4.2 *ABC* Type 489
2.4.3 Segmented Type 490
2.5 Phase-Separated Blends 490
2.6 Composites and Laminates 490
3. **Effects of Processing on Polymer Morphology** 491
3.1 Uniaxial Orientation during Crystallization 491
3.2 Skin/Core Morphology 491
3.3 Blend Morphologies 491
3.4 Biaxial Orientation during Molding 491
4. **Morphological Transformations of Polymers** 491
4.1 Thermal Effects 491
4.2 Mechanical Stress Effects 492
4.3 Effects of Solvents and Other Chemicals 493
4.4 Environmental Effects 493
4.5 Irradiation Effects 493
Glossary 493
Works Cited 494
Further Reading 495

INTRODUCTION

Polymer molecules are chains containing a large number of one or more repeat units linked together by covalent bonds. These molecules make up important materials that have been put to use in a variety of ways that take advantage of their unique physical, mechanical, electrical, and optical properties. Crystal structure determination, using x-ray diffraction; morphological investigation, employing optical and electron microscopy; and the viewing of individual molecules, using scanning probe microscopy, are characterization techniques currently in use by polymer scientists. These and other methods have allowed the investigation and characterization of a large number of new polymeric substances and compositions.

The form and structure, the morphology, of a polymeric material depends on the molecular structure and mass, the chain confor-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

mation, the thermal history, and the mechanical stress history. A polymer sample can be

1. an isotropic glass or rubber,
2. partially crystalline with glassy or rubbery interlayers,
3. completely crystalline,
4. in the form of a liquid-crystalline mesophase, or
5. part of a phase-separated or composite system.

Before use, polymeric materials undergo fabrication processes such as molding, extrusion, casting, and uniaxial or biaxial extension; these processes can lead to new or mixed morphologies and have a direct effect on the properties and performance. Therefore, a detailed understanding of the morphology and structure of polymeric substances and mixtures thereof is an important goal.

In the sections to follow, the requirements for polymer crystallization, chain regularity, conformation, orientation, packing, and folding are discussed, the various polymer morphologies known to date are explained and illustrated, the methods of crystallization, crystal structure determination, and morphological investigation are reviewed, and the effects of thermal changes, mechanical stress, solvents, and irradiation on the morphology are presented.

1. POLYMER CRYSTAL STRUCTURE

1.1 Chain Regularity

The ability of a polymer to form a three-dimensional repeating structure is largely, but not totally, dependent on the regularity of the atomic placement in the chain. Homopolymers that crystallize can be obtained by polymerization of unsaturated monomers, by condensation-type reactions of bifunctional monomers, and by polymerization of certain ring monomers.

Some of the repeat units obtained by polymerization of unsaturated monomers to form singly bonded carbon-atom chains are symmetrical, having the same substituents attached to each carbon atom in the chain, e.g., the repeat unit in polyethylene, —CH_2CH_2—, and some have different types of substituents on alternate carbon atoms but have a head-to-tail, isoregic, linking of the repeat units, as shown in Fig. 1. The substituents on a given carbon atom can be the same, as for poly(vinylidene chloride), —CH_2CCl_2— or they can be different, e.g., polypropylene, —$CH_2CH(CH_3)$—, with the latter referred to as asymmetric. The three arrangements, *configurations*, of three successive asymmetric repeat units, shown in Fig. 1 are possible; in this figure X, Y, and Z denote substituent groups attached covalently to carbon atoms in the chain. In some cases X and Y (or Z) are the same. Since the substituents on singly bonded carbon approximate a tetrahedral arrangement, the Y and Z groups project out in front or behind the zigzag carbon chain drawn in the plane of the paper. An isotactic placement has all Y groups projected out on one side and all Z groups projecting out the other side.

The polymers with asymmetric repeat units that crystallize usually have a regular configuration, isotactic or syndiotactic; some examples of isotactic polymers that have been prepared are polypropylene, Y = H, Z

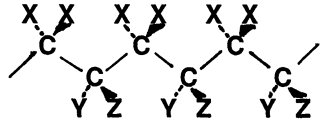

ISOTACTIC

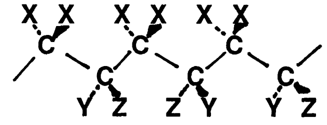

SYNDIOTACTIC

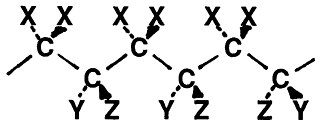

ATACTIC

FIG. 1. Types of isoregic configurations in chains containing asymmetric repeat units. X, Y, and Z are different chemical substituents. Three repeat units are shown.

= CH_3; poly(butene-1), Y = H, Z = CH_2CH_3; polystyrene, Y = H, Z = phenyl; poly(methyl ether), Y = H, Z = OCH_3; and poly(methyl methacrylate), Y = CH_3, Z = $C(O)OCH_3$. Syndiotactic polymers include polypropylene and poly(methyl methacrylate). Polymers with asymmetric repeat units in an atactic configuration that crystallize are those with small highly polar substituents such as the CN group in polyacrylonitrile, Y = H, Z = CN, and the OH group in poly(vinyl alcohol), Y = H, Z = OH. Polymers with large Y and/or Z substituents, such as poly(methyl methacrylate) and polystyrene, do not crystallize, if the configuration is atactic. An example of a polymer with an asymmetric repeat unit, shown to contain nonisoregic as well as atactic linkages, is free-radical–initiated poly(vinyl fluoride); despite these irregularities, this polymer crystallizes. Polymers synthesized from diene monomers are crystallizable if the linkage is 1,4, and if the substitution is predominately *cis* or *trans* or if an isotactic or syndiotactic arrangement occurs in the 1,2 or 3,4 substituted isomers. Representative polydiene repeat units are given in Fig. 2. In these structures R = H for polybutadiene and R = CH_3 for polyisoprene.

Polymers containing a symmetrical repeat unit and *small* amounts of a second unit, statistically placed, such as polyethylene with some propylene repeat units, are crystallizable; the repeat unit present in small amount may be excluded from the crystal stems and appear at the crystallite surface. Polymers containing two different repeat units are called copolymers. Statistical copolymers with both repeat units asymmetric usually do not crystallize.

The chains formed from unsaturated monomers are generally flexible because of the presence of many singly bonded carbon atoms. The effects of this flexibility on the morphology will be discussed below.

Many polymers are synthesized from monomers with reactive end or side groups, such as alcohols, acids, or amines. Homopolymers formed from monomers containing two reactive groups are usually crystallizable since only one isomeric choice exists and ordered arrangements can be formed. Condensation-type repeat units can be incorporated in flexible chains, e.g.: poly(6-amino hexanoic acid) and poly(ethylene terephthalate); semiflexible chains, e.g: polycarbonate; semirigid chains, e.g.: polybenzamide; and rigid chains, e.g.: polybenzimidazole. The repeat units for these are given in Fig. 3. These dif-

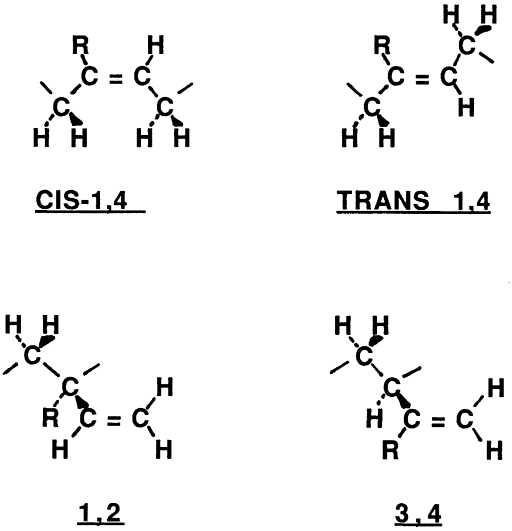
CIS-1,4

TRANS 1,4

1,2

3,4

FIG. 2. Types of polydiene repeat units. R can be H, CH_3 or Cl.

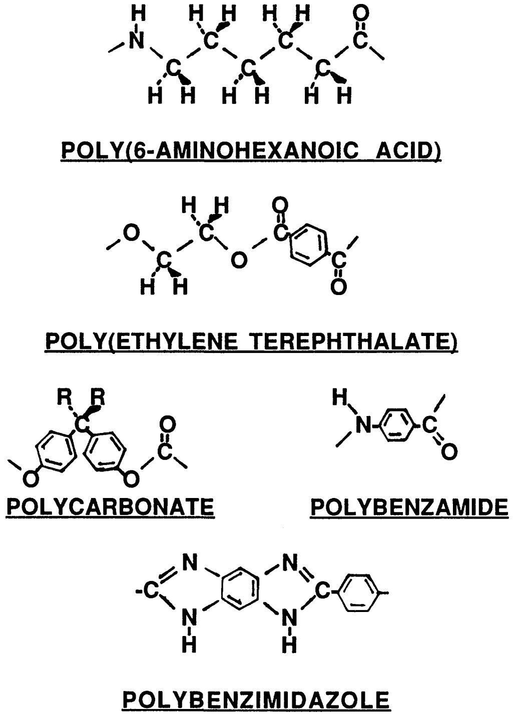
POLY(6-AMINOHEXANOIC ACID)

POLY(ETHYLENE TEREPHTHALATE)

POLYCARBONATE

POLYBENZAMIDE

POLYBENZIMIDAZOLE

FIG. 3. Polymer repeat units containing functional groups in chain (R is CH_3).

ferent types of polymers show different characteristics, particularly in their ease of crystallization and, to some extent, in their morphologies. Many of these polymers contain unsubstituted aromatic rings with para linkages—linkages on diametrically opposite atoms. The introduction of ortho (on adjacent atoms) and meta (on next-nearest neighbor atoms) linkages will cause disorder in the crystal lattice and thereby lower the crystallization ability. Substituents on the rings will have a similar effect on the crystal lattice.

The presence of trifunctional or tetrafunctional monomers in a condensation-type reaction mixture leads to highly branched or to network polymers, which usually do not crystallize. Epoxy resins are an example of this type of system. Also, the presence of a three-dimensional covalent network precludes any further molding, extrusion, or drawing.

Various small molecules with ring configurations can be polymerized to form crystalline linear-chain products. Included in this group are polymers containing silicon and phosphorus atoms in the chain. Polymers formed from ring monomers generally have flexible chains.

1.2 Chain Conformation, Orientation, and Packing

In the tacticity structures given above, when the carbon atoms are in the same plane, one of the two atoms or groups of atoms on each projects out and the other projects in. The spatial placement of each CX_2 and CYZ group and the remainder of the chain attached to it can be changed by rotation about the C—C bond between the two carbon atoms preceding the group in question. Each allowed change of the rotation angles yields another conformation for the chain. The chemical makeup and configuration of the polymer repeat units present dictate which conformations are energetically feasible as well as the relative energies of each. Rotation angles of 0, ±60, ±120, and 180 are labeled as *trans* (t), *skew*± ($s^{\pm}$), *gauche*± ($g^{\pm}$); and *cis* (c). For a polymer molecule any combination of bond conformations that results in close approach or overlap of nonbonded atoms will be of high energy and therefore will not be preferred. For polyethylene a succession of *trans* conformations (ttt), which gives a chain with the lowest energy and highest extension, is the arrangement present in the crystal. However, many other combinations of *trans, gauche*+, and *gauche*− occur for this polymer in the liquid state and in solution. For isotactic polypropylene and polystyrene a tg^+tg^+ conformation, resulting in a helix with three monomer units per turn, has the lowest energy for an isolated chain. For the polydienes, with repeat units —CH_2—CR=CH—CH_2—, *skew* and *cis* states are energetically preferred to the *trans* and *gauche* states for the single bond following the double bond in the repeat unit. Two chain conformations have low energy for *trans*-1,4-polyisoprene: s^+ts^- and cts^+cts^-. [The first (and fourth) symbol in these groups refers to rotation about the CH—CH_2 bond, the second (and fifth) symbol to rotation about the adjoining CH_2—CH_2 bond, and the third (and sixth) symbol to rotation about the subsequent CH_2—CR bond.] A polymer chain can contain relatively inflexible groups such as —C=C— units, aromatic rings, and amide groups. The all-aromatic polyamides, e.g., polybenzamide (see Fig. 3 for repeat unit structure), and polyesters are semirigid with limited conformational choices per repeat unit available. Rotation about the C(O)—NH bond takes considerable energy, resulting in a fixed conformation for that group. A rodlike polymer, such as polybenzimidazole, has only one conformation available to it and is fully extended.

During crystallization the chains adopt a particular low-energy conformation and pack together in a three-dimensional repeating array, usually with the conformed chains parallel. The unit cell usually contains only a small part of any particular chain, depending on the conformation adopted (see CRYSTALLOGRAPHY). The orthorhombic unit cell in polyethylene contains two —CH_2—CH_2— units side by side with one chain turned with respect to the other; the unit cell for the triclinic form of polytetrafluoroethylene is a helix containing 13 —CF_2—CF_2— units on the same chain having six turns. These and other unit cells are described by Wunderlich (1973).

1.3 Chain Folding

Crystallites formed from flexible-chain polymers can vary in size depending on the crystallization conditions; the dimension of the crystallite in the chain direction, referred to as the crystal stem length, can be considerably less than the total length of the chain having the particular bond conformations found in the unit cell. One very important parameter determining the crystal stem length is the degree of supercooling, given by $T_m^0 - T_c$, where T_m^0 is the equilibrium melting point and T_c is the crystallization temperature. Crystallization conditions leading to large stem lengths include high pressure, small degrees of supercooling, and long crystallization time. The remainder of the chain not incorporated in a particular crystal stem can be partially incorporated as one or more crystal stems in the same or a different crystallite or it can project from the crystallite surface. If the remainder of the chain is to be incorporated into the same crystallite, it must turn back, or fold, resulting in the appearance of bond conformations other than those in the unit cell. Chain folding, protrusion, or incorporation in two different crystallites would lead to a noncrystalline component. Folded-chain crystallites melt at temperatures below T_m^0 with the melting temperature decreasing with lamellar thickness and crystallization temperature. If the heating rate is low, reorganization with longer stem lengths can occur.

1.4 Crystallization Methods

1.4.1 Crystallization from Solution Most crystallizable flexible-chain polymers form solutions with one or more solvents at elevated temperatures. With prolonged heating crystal nuclei present due to previous crystallization are lessened in number or destroyed completely. Cooling this solution to a temperature sufficiently below that for dissolution can lead to crystallization, once nucleation occurs. Crystallization can be carried out isothermally or upon slow cooling. The crystal growth rate increases with the degree of supercooling. A seeding technique can be used to regulate the nucleation process. An essentially nucleus-free solution is subjected to a high degree of supercooling, causing rapid nucleation and crystallization; this suspension is then heated to a temperature below the minimum dissolution temperature and crystallization carried out isothermally. Crystallization can also be brought about by solvent evaporation at a temperature above that for dissolution.

1.4.2 Crystallization from the Melt Most crystallizable flexible-chain polymers form isotropic liquids at sufficiently elevated temperatures. Crystallization is usually brought about by sufficient cooling of this liquid below the equilibrium melting temperature under quiescent conditions. This process can be carried out at atmospheric or elevated pressures. Crystallizable polymers containing bulky groups in the main chain or as substituents, when cooled from the melt, can be easily quenched to a glass. Crystallization of these can be afforded by heating to a temperature above the glass transition temperature but below the melting temperature.

1.4.3 Crystallization under Mechanical Stress The application of a tensile stress to a crystallizable polymer in the rubbery liquid state can lead to chain extension and crystallization. Imposition of a tensile or a shear stress on a supercooled melt undergoing crystallization brings about changes in the morphology of the material. Vigorous stirring of a supercooled solution at temperatures above those at which crystallization occurs quiescently can bring about crystallization and a different morphology.

1.5 Crystal Structure Determination

The determination of the crystal structures of many synthetic polymers by x-ray diffraction has been carried out using quiescently crystallized samples subsequently subjected to a tensile stress that acts to orient the chains in a known direction. The resulting fiber is rotated about the orientation direction and the diffraction pattern obtained on film using the Laue or Weissenberg methods (see CRYSTALLOGRAPHY). The pattern resulting consists of lines, arcs, spots, or a combination thereof, depending on the amount of chain orientation. The periodicity and the unit cell distance along the chain can be obtained from the distances between lines or layers of arcs or spots appearing perpendicular to the direction of orientation. Unit cell distances in the two other directions are found from spacing of spots or arcs within the layers. Diffraction patterns obtained on unstretched quiescently crystallized films consist of rings or ellipses, due to random orientation of the crystallites, superimposed on a broadened diffraction background, due to the noncrystalline portions present.

Single-crystal diagrams have been obtained by selected-area electron diffraction on single platelets of polymers precipitated from dilute solution. In these samples the chain direction is usually parallel to the beam direction, affording a pattern of $h,k,0$ diffraction spots (see CRYSTALLOGRAPHY).

2. POLYMER MORPHOLOGY

2.1 Methods of Investigation

2.1.1 Optical Microscopy Solids and solid–liquid mixtures can be examined using optical microscopy at magnifications up to about 2000. Viewing of images formed with transmitted plane polarized light is done through an analyzer, a polaroid lens with its axis usually oriented at right angles to that of the polarizer. Light can be transmitted through the analyzer if parts of or if all of the sample is birefringent—that is, with a refractive index for light with electric vibrations parallel to the optic axis different from that for light perpendicular to this axis. Light will not be transmitted for the following reasons:

1. the thickness of the birefringent area is less than one wavelength,
2. the optic axis is parallel to the polarizer or analyzer direction,
3. the direction of the light is parallel, and therefore the electric vector is perpendicular, to the optic axis.

Phase contrast uses a 90° change of phase of the light diffracted from various regions of the sample; this shows up differences in refractive index within the specimen when this light is recombined with the transmitted beam.

In interference contrast one portion of the beam passes through the sample and the other part is changed in phase by a variable amount to give optimum contrast upon recombination of the two beams, showing up steps and surface tilts present and giving a three-dimensional view.

2.1.2 Electron Microscopy Electron microscopy is carried out in transmission, TEM, on relatively thin polymer specimens at magnifications up to about 2×10^6. Bright-field conditions, in which the transmitted electrons form the image, are usually used. It is possible to reverse the contrast by forming the image in dark field with one or more diffracted beams. If the sample is crystalline, a selected-area electron-diffraction pattern can also be obtained. Phase contrast is obtained by defocusing the objective lens, which produces a phase shift. By using a high-resolution electron microscope, HREM, at low dosages and moderate magnifications, followed by photographic enlargement and computer-controlled image enhancement, a crystalline or liquid-crystalline lattice can be imaged. Restrictions on sample thickness necessitate microtoming or surface replication of samples. Prior to replication or following microtoming, various treatments, such as reaction with OsO_4 or RuO_4 or etching with acids, bases, organic liquids, ions, or electrons, have been employed to highlight or reveal certain parts of the total solid sample. Shadowing with a heavy metal is carried out in order to provide contrast between the sample and substrate.

Scanning electron microscopy, SEM, uses a focused electron beam to scan the sample surface. Magnifications up to about 100 000 are possible although most work is carried out below 10 000×. In addition to having a higher resolution than optical microscopy, SEM has a much larger depth of field. Samples are thinly coated with a metal to provide a conductive layer.

2.1.3 Scanning Probe Microscopy This involves moving a probe with a finely ground tip, from one to a few atoms across, over the surface of the specimen (Wickramasinghe, 1989). The specimen surface can be imaged on an atomic scale with a small electrical current, as used in the scanning tunneling microscope, STM, or by application of a mechanical force, employed in the atomic force microscope, AFM. STM requires a conducting surface; AFM can be used to investigate insulating surfaces including polymers. In both the STM and the AFM the electron cloud at the tip of the probe is in contact with that at the specimen surface.

2.2 Morphologies of Crystallized Polymers

2.2.1 Single Lamellas Single platelets can be formed, by isothermal crystallization in a quiescent supercooled melt or solution, from a large variety of polymers with a wide range of chain flexibility (Geil, 1963; Woodward, 1988; Wunderlich, 1973). Single-polymer lamellas can have a highly regular geometry, appearing to be single crystals. Faceted lamellar geometries reported include diamond, hexagonal, pyramidal, square, and lathe shapes. Some atactic polymers form irregularly shaped nonfaceted lamellas. The observed geometries depend on the crystal lattice type adopted and the crystallization solvent.

A schematic representation of the center part of a single-polymer lamella, based on the chain folding model, is shown in Fig. 4. Each straight line represents a section of a polymer chain incorporated in the crystal lattice and contains a number of repeat units; the loops shown depict the folds, represented as

1. tight adjacent reentrant,
2. loose adjacent reentrant, and
3. nonadjacent reentrant.

The dimension along the chain direction in the crystallized portion of a single lamellar platelet from solution is usually of the order of 5–20 nm, while the lateral dimensions are in the range of 1–20 μm (1000–20 000 nm). Diamond- and square-shaped lamellas have four sectors and hexagonally shaped ones have six, corresponding to different crystallographic growth planes. Polymer lamellas can be nonplanar and therefore collapse features, such as creasing and splitting, can be evident in samples deposited from solution, as shown in Fig. 5 (Anandakumaran *et al.*, 1983). Twinning, growth of two lamellas from a common face, can occur. Growth-face splitting, leading to lamellar branching and to the formation of spiral growths, can take place.

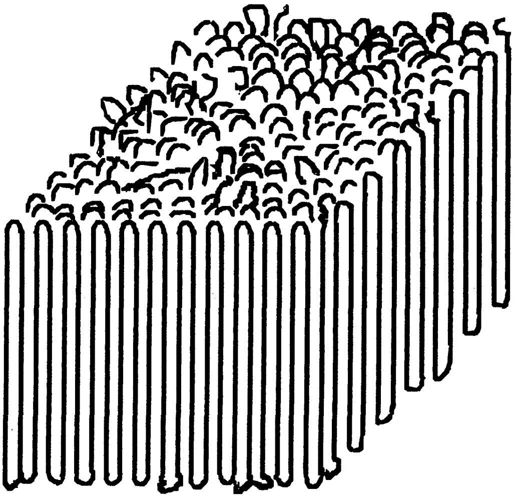

FIG. 4. Polymer single lamella with folded chains.

2.2.2 Dendrites, Hedrites, and Spherulites Under many conditions, particularly during crystallization from concentrated solution or the melt, branched and/or multilayered lamellar structures appear.

Dendrites are branched lamellas with the branches related in a definite crystallographic way to each other. Dendritic structures are found for solution-crystallized polyethylene with branches along the crystallographic *a* and *b* axes, perpendicular to the chain direction in the lamellas, and can have lengths of hundreds of micrometers (Wunderlich, 1973). At high concentrations thick multilayered dendrites are formed. Interpenetrating crosshatched dendritic structures, called quadrites, of one polymer, isotactic polypropylene, have been crystallized from solution and from the melt.

Hedrites (axialites) are stacks of centrally connected lamellas that grow from solution or the melt (Geil, 1963). A SEM micrograph of two hedrites is given in Fig. 6. During

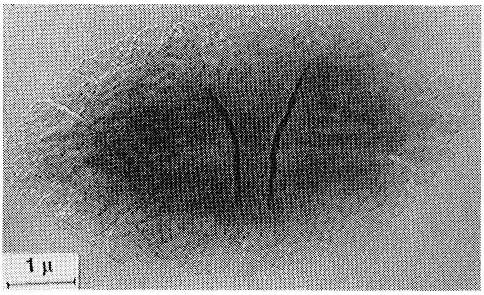

FIG. 5. Transmission electron micrograph of a *trans*-1,4-polyisoprene lamella grown from dilute amyl acetate solution at 20 °C showing wrinkling and splitting. [Reprinted from Anandakumaran *et al.* (1983). Copyright 1983, American Chemical Society.]

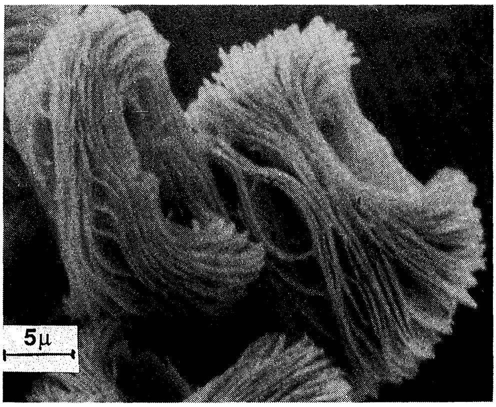

FIG. 6. Scanning electron micrograph of *trans*-1,4-polyisoprene hedrites grown from 1% amyl acetate solution at 20°C. [As in Xu and Woodward (1986). Copyright 1986, American Chemical Society.]

growth from solution, the lamellas can develop further connections as a result of the incorporation of a single chain or a group of chains in neighboring lamellas. One consequence is that two lamellas can be closely connected to each other at more than one place in the stack; repeated events of this sort give the hedrite a cellular appearance. A lamellar stack can exhibit considerable curvature, and separation or splaying of the lamellar ribbons can occur.

Spherulites are composed of lamellar ribbons that usually grow outward, radially

1. from a single (heterogeneous) nucleus,
2. from a group of lamellas similar to that in a hedrite, or
3. from a quadrite, a crosshatched lamellar array.

After sufficient supercooling of a relatively concentrated solution or of a melt, spherulites can be grown at constant or slowly decreasing temperature. The growing spherulite is filled in by lamellar branching, particularly at high degrees of supercooling. Branches can be created by nucleation of new lamellas at the surface of a parent lamella, as occurs in hedrites, or by dendritic branching, branching at the growth face, which occurs at high degrees of supercooling. When the crystallization temperature is raised, the degree of supercooling is decreased, branching diminishes, the primary lamellas increase in average thickness, and additional lamellas fill in the spherulite structures at a later time. Curvature and twisting of the lamellas in spherulites occur during growth. A common birefringence pattern for melt-crystallized spherulites is four bright quadrants separated by a dark cross, with the arms of the cross in the directions of the polarizer and analyzer. In addition, sometimes alternating bright and dark (extinction) rings are observed. Multilamellar crystallization from the melt usually yields a coherent sample made up of individual structures impinging on and linked to one another by common lamellas. An interference contrast optical micrograph of spherulitic growth from the melt is shown in Fig. 7 (Woodward, 1983). Individual spherulites can be grown from solution; two examples are shown in Fig. 8 (Xu and Woodward,

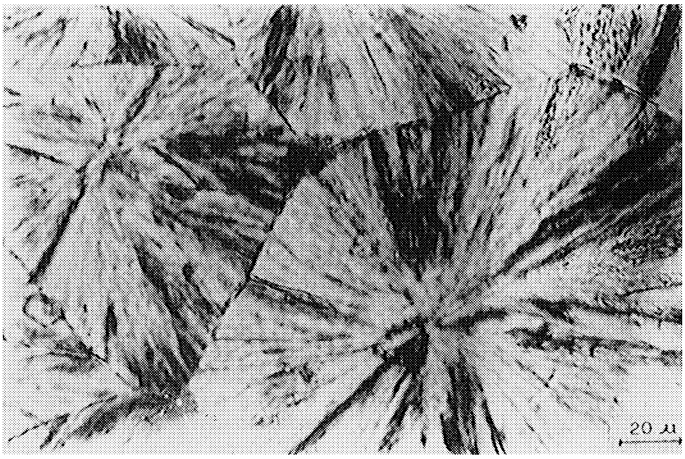

FIG. 7. Interference contrast optical micrograph of *trans*-1,4-polyisoprene spherulites grown from the melt at 55°C. [Reprinted from Woodward (1983), p. 141. Photograph by C. Kuo.]

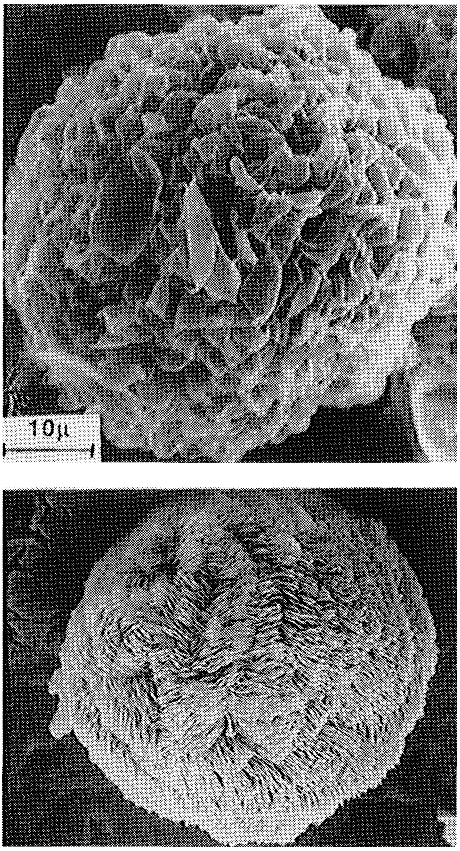

FIG. 8. Scanning electron micrographs of *trans*-1,4-polyisoprene spherulites grown from 1% amyl acetate solution at 20°C. Two samples with different molecular weights were used. [Reprinted from Xu and Woodward (1986). Copyright 1986, American Chemical Society.]

1986). Both quasicircular and scrolled lamellar spherulites have been reported (Lovinger, 1980; Vaughan, 1993).

Crystallization at high pressure (4 to 5 kbar) and increased temperature yields spherulites with highly, if not completely, extended chains. Crystallization from a liquid crystalline mesophase of semistiff chains (see Sec. 2.3) can give lamellar stacks linked together by extended chains (Hudson and Lovinger, 1993).

Hedrites and spherulities have crystallinities significantly smaller than those for solution grown lamellas, mainly because of a larger degree of supercooling that results in a smaller crystal stem length. In multilamellar structures, not all of the noncrystalline component resides in chain folds and chain ends, particularly when the nucleation rate is large. Under that condition a given chain can be incorporated in more than one lamella leading to the presence of an interlamellar traverse.

2.2.3 Fibers Fibers are long, narrow, ribbonlike structures, having a different chain orientation from that present in lamellas. In these structures the chain (*c* crystallographic axis) orientation is mainly along the ribbon direction; it is not perpendicular to it, as found in lamellas. Fibers are usually composed of smaller units, fibrils, microfibrils, and nanofibrils, oriented in the same direction. Fibrillar structures can be formed naturally, during synthesis, from solution, from the melt, from a gel, or from quiescently melt-crystallized material with the chains being placed in extended conformations during the crystallization or recrystallization process. Fibers can be prepared from crystallizable polymers with flexible chains, as well as from those with completely stiff or semistiff chains. Depending on the method of preparation, fibers from flexible-chain polymers can contain extended noncrystalline portions of chains and chain folds, or they can have perfect or nearly perfect order, as shown schematically in Figs. 9(a) and 9(b), respectively.

2.2.4 Epitaxial Growths Epitaxial or oriented overgrowth of flexible-chain polymer crystals on crystalline substrates can occur from solution or the melt.

Epitaxial crystallization of a polymer on the surface of a highly oriented crystalline film of another polymer has been carried out by pressing thin films of the two polymers together and applying a tensile stress during the melt crystallization process (Gross and Petermann, 1984).

The nucleation of folded-chain lamellar growth on a small group of extended chains of the same or a different polymer is another form of epitaxial growth. In this type of crystallization the *c* crystallographic axis, usually the chain axis, in the lamellas and in the fiber are parallel. A schematic representation of folded-chain growth on extended chains is shown in Fig. 10.

2.2.5 Dislocations and Defects Microscopic defects in crystallized polymers include

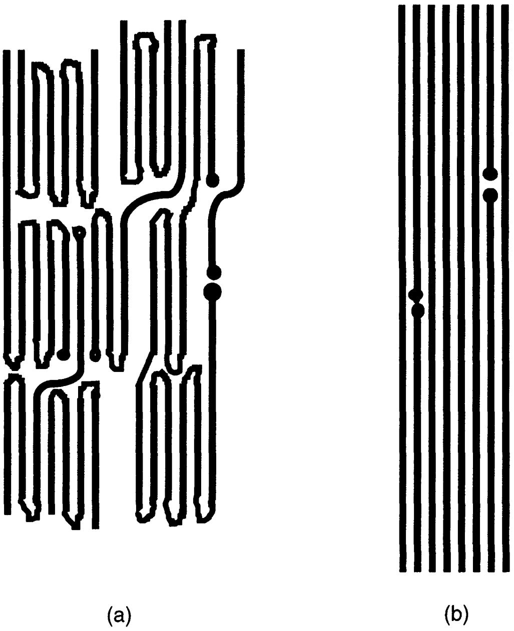

FIG. 9. Fibrillar crystallization. (a) Cold-drawn melt crystallized, (b) ultradrawn solution crystallized.

1. localized deviations from a periodic arrangement within the crystal lattice caused by screw and edge dislocations,
2. the changes in overall order occurring at lamellar surfaces and between lamellas, and
3. chain disorder within a crystal lattice.

Edge and screw dislocations in a crystal lattice are generated by line translations (see CRYSTALLINE STATE). The existence of a screw dislocation in polymer lamellas with a size of about 10 nm is readily observed by electron microscopy, because the dislocation causes the appearance of a lamellar growth spiral (Keller, 1958). Dislocations of the order of one repeat unit and dislocation networks are observed on polymer lamellas using moiré patterns generated in dark-field TEM by two superposed but misaligned lamellas.

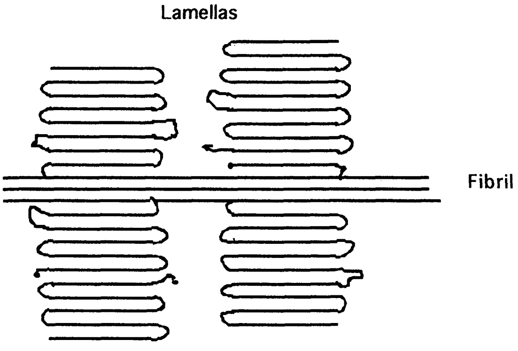

FIG. 10. Epitaxial lamellar growth on an extended chain fibril.

Chain folds and chain ends at lamellar surfaces and interlamellar chain traverses between neighboring lamellas, as discussed in Secs. 2.2.1 and 2.2.2, can be considered to be surface defects. Folds on single lamellas have been imaged using decoration techniques, employing evaporated metal atoms or evaporated and degraded polymer molecules (Wittmann and Lotz, 1985). Lamellar sectors and fold direction are delineated by those methods.

Chain disorder within the crystal lattice includes kinks, jogs, and twists in a crystal stem or stems as well as chain ends and folds. Kinks have been observed on fracture surfaces of extended chain lamellas prepared under pressure. Kinks can also be introduced into single lamellas by stress application (see Sec. 4.2).

2.3 Liquid-Crystalline Mesophases

Some polymer fluids exhibit one- or two-dimensional order involving all of the chain, parts of the main chain, parts of side chains, or parts of both the main and side chains.

2.3.1 Types of Mesogens Flexible-chain polymers can exhibit mesophases in which there is side-by-side chain alignment with disorder and movement along the chain direction. One such is the pseudohexagonal or hexatic liquid crystalline state (Kojima and Magill, 1986). Adequate cooling of this mesophase leads to crystallization and the attainment of three-dimensional order.

Polymers with chains composed of

1. alternating rigid rodlike and flexible main-chain sections,
2. rigid rodlike side-chain units and flexible main-chain sections, or
3. disk-shaped moities

can form heat-sensitive (thermotropic) liquid-crystalline mesophases. Polymers having completely inflexible chains form mesophases only in solution (lyotropic). The order in a liquid-crystalline mesophase of a polymer with rodlike mesogens can be one dimensional, nematic, with the rigid rodlike parts of the polymer molecules possessing an average molecular orientation called the director. A two-dimensional arrangement, smectic, with the rigid sections in straight or oblique rows can occur. A third arrangement of the rodlike sections is cholesteric and involves nematic layers arranged with a helical twist relative to each other. The cholesteric phase is only observed for optically active polymers or inactive polymers mixed with optically active small molecules. Molecules with rigid disk-shaped sections can form discotic liquid-crystalline mesophases in which the disk-shaped parts are stacked in columns. Some liquid-crystalline arrangements involving polymer chains are illustrated in Fig. 11.

Amphiphilic polymers in which alternate carbon atoms in the main chain have acidic and basic substituents form thermotropic liquid-crystalline mesophases, as a consequence of ion–dipole interaction between side-chain substituents on different chains. Block copolymers (see Sec. 2.4) can form lyotropic mesophases containing a close-packed array of spheres, a hexagonal array of cylinders, or a planar array of polymer layers.

2.3.2 Mesophase Textures A domain of rodlike mesogens oriented in the same direction and placed between crossed polaroids can transmit light. However, orientation of the rodlike parts of the chains along the polarizer or analyzer directions or perpendicular to the analyzer and polarizer will lead to extinction of the light and the field will appear dark. The complete disorientation of the rodlike parts in the isotropic liquid also causes extinction. Other orientations of the rods can bring about light transmition. Interaction of the light waves with the oriented mesogenic units induces a wavelength-dependent change in polarization causing the light transmitted through the analyzer to have one or more colors, when white light is used for the source.

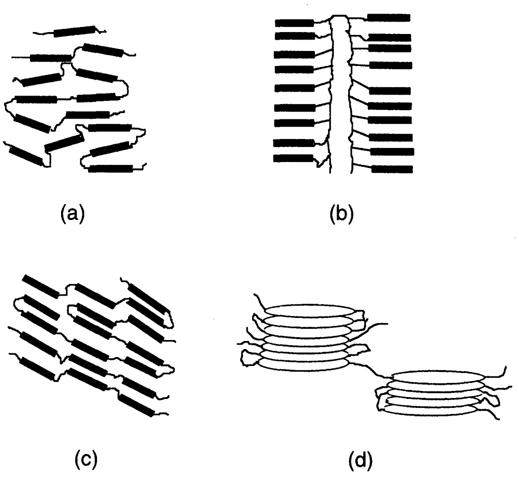

FIG. 11. Liquid crystalline mesophases: (a) main-chain nematic, (b) side-chain smectic, (c) main-chain smectic, (d) main-chain discotic.

When a nematic liquid-crystalline mesophase is first formed by melting a thin microcrystalline film sample, many domains are present and many entangled dark lines, disclinations (line rotational defects), appear in the microscopic pattern (Noel *et al.*, 1984). Usually the number and tortuosity of these lines decrease with time as a result of disclination coalescence and mesogen alignment; an example of this texture at an early time is given in Fig. 12(a). Alignment can be accelerated by shearing at low rates or by the application of a high-strength electrical or magnetic field. However, disorder will occur on cessation of shearing or field application. As the temperature is increased closed loops are formed, which eventually disappear. Thinner films of a nematic mesophase can exhibit a schlieren texture [Fig. 12(b)]. This texture contains singular points (marked by arrows in figure), connected by two or four intersecting black wavy stripes or brushes. The points are disclinations viewed end-on. The dark brushes occur where the orientation of the director is parallel to the analyzer or polarizer directions; regions transmitting light contain domains for which the director orientation is not parallel to these directions.

Smectic mesophases can display a schlieren texture with four brushes, a focal conic, or a fan-shaped texture. Cholesteric mesophases can have a fingerprint and a ribboned texture and in reflected white light show iridescent colors. Dendritic and fan-shaped textures have been reported for polymers with disk-shaped mesogens (Percec *et al.*, 1992).

2.3.3 Disclination Imaging The director field in the vicinity of a line rotational defect, or disclination, in a thin sample of a

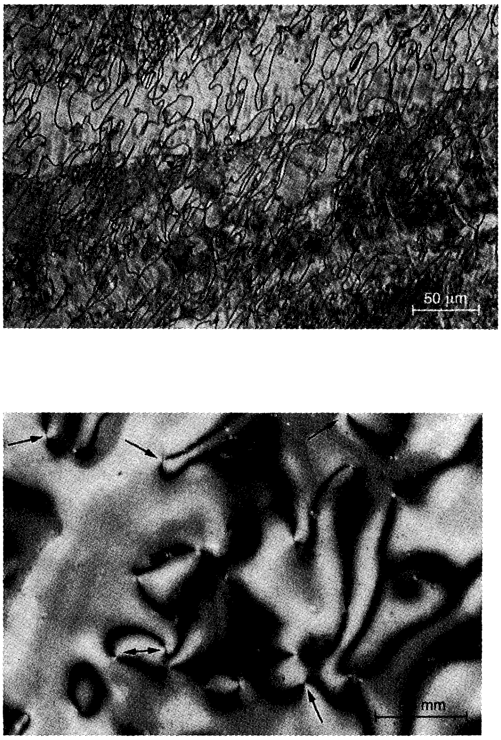

FIG. 12. Optical micrographs using polarized light for copolyesters in a nematic mesophase, showing (a) threaded and (b) schlieren textures. Courtesy of C. Noel. [(a) reprinted with permission from Noel *et al.*, (1984).]

polymeric liquid-crystalline mesophase can be imaged using TEM. The mobile mesophase is quenched to form a nematic glass; then the temperature is raised to about 20° below the crystalline melting point and the glass annealed, leading to the formation of many lamellar crystals with sizes smaller than the disclination size (Wood and Thomas, 1986). The lamellas are viewed edge-on and the director is normal to them.

Disclinations can occur in dipolar or in quadrupolar (Lehman) clusters. Nematic polymer mesophases aligned in a magnetic field of intermediate field strength display both dipolar and quadrupolar clusters, and extensional flow of oriented nematic liquid crystals brings about the appearance of many Lehman clusters (Hudson and Thomas, 1991a, 1991b). Inversion walls, where the director inverts by 180° from one side to the other, have been observed in nematic liquid-crystalline mesophases under the same conditions producing Lehman clusters.

2.4 Phase-Separated Block Copolymers

Block copolymers are polymers containing sequences of two or more repeat units. Various types of block copolymers that have been prepared are described below. Many polymers are incompatible to some degree and mixtures of them will undergo phase separation, if in solution or melt. This process also occurs in block-copolymer samples, if the blocks are long enough and if the constituent polymers are incompatible. For noncrystallizable flexible polymers the chains in phase-separated domains will be randomly coiled. If one or more of the polymers present crystallize then additional phases are formed.

2.4.1 *AB*, *ABA*, and Star-Block Types

In an *AB* block copolymer a sequence of one type of repeat unit is covalently linked to a sequence of another type; in an *ABA*-type block copolymer the *B* sequence is linked at both ends to *A* sequences; and in a star-block copolymer sequences of one type of repeat unit are attached to a multifunctional nucleus at one end and to sequences of a second type of repeat unit at the other end, as shown schematically in Fig. 13. The attachments between the *A* and *B* portions cause them to stay in close proximity to each other, greatly limiting the size of the separated phases and also affecting the morphol-

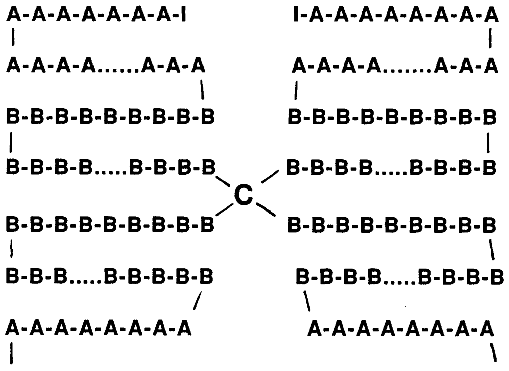

FIG. 13. Star-block copolymer structure with four arms. *A* and *B* denote different types of repeat units, *I* an end unit, and *C* a multifunctional linking agent.

ogy of the phase-separated system. The four phase-separated morphologies observed for two-component noncrystallizable block-copolymer systems (*AB*, *ABA*, and star block) are illustrated in Fig. 14 and include

1. spheres of *A* embedded in a continuous *B* phase,
2. cylinders of *A* in a continuous *B* phase,
3. the ordered bicontinuous double-diamond type, and
4. layers of *A* separated by layers of *B*.

These are listed in order of increasing volume fraction of the *A* component. The precise morphology adopted by a phase-separating *AB*-type block copolymer depends on the composition and the molecular weight. Inclusions or layers of 20–30 μm are observed. Changes in morphology can occur for *ABA* block-copolymer systems, if the components are interchanged.

Flexible-chain block copolymers containing a crystallizable component do not show the morphologies discussed above. For quiescently crystallized flexible-chain *AB* or *ABA* block copolymers, single lamellas are expected during prenucleated crystallization from dilute solution and multilamellar structures would appear from .moderately concentrated solution without seeding. If the two components are incompatible, the chains of the noncrystallizing component will separate from the folds of the crystallizing component.

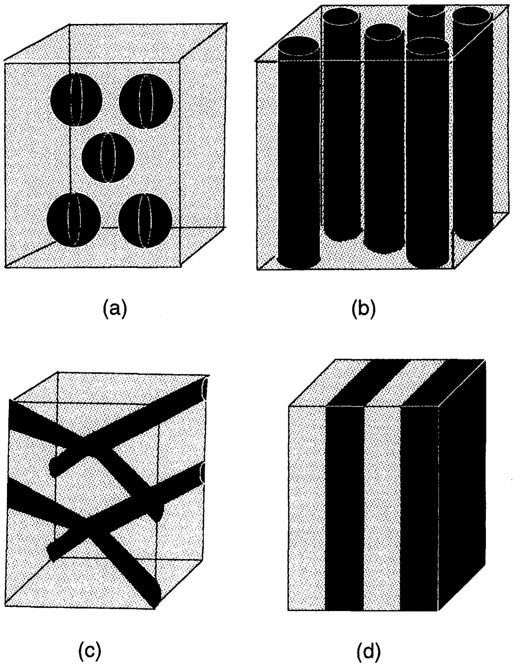

FIG. 14. Microphase separation in *AB* and star-block copolymers. (a) Spherical inclusions, (b) cylindrical inclusions, (c) ordered bicontinuous double-diamond inclusions, and (d) layers.

2.4.2 ABC Type Ten different phase-separated morphologies have been reported to date for three-component block-copolymer systems (*ABC* type with *B* the middle block). Eight of these are shown schematically in Fig. 15 and involve spheres, spheres with shells, cylinders, cylinders with shells, and layers (Woodward, 1995). Also, mutually interpenetrating double-diamond frameworks

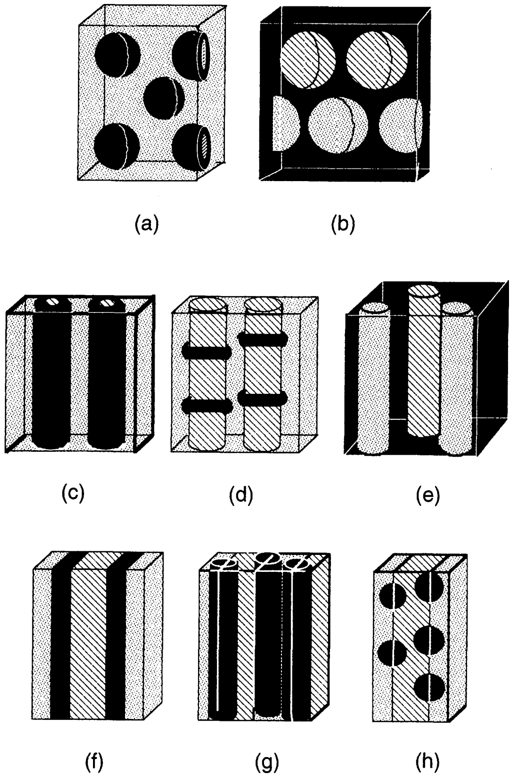

FIG. 15. Microphase separation in *ABC* block copolymers (the *B* portion is black): (a) spherical inclusions with spherical shells, (b) separate spherical inclusions of *A* and *C* in *B*, (c) cylindrical inclusions with cylindrical shells, (d) cylindrical inclusions with rings, (e) separate cylinders of *A* and *C* in *B*, (f) four layers, (g) cylindrical inclusions at layer interfaces, (h) spheres at layer interfaces.

and spheres with hexagonal shells have been found.

2.4.3 Segmented Type In this type of block, copolymer sections of one repeat unit alternate with sections of another many times per chain. These copolymers form lamellar structures on crystallization (Woodward, 1988).

2.5 Phase-Separated Blends

Mixing of two incompatible amorphous homopolymers sufficiently, prior to obtaining a rigid or semirigid mass, can lead to microphase separation in which particles of the constituent present in the smaller quantity are dispersed in the other, more prevalent, constituent. Cooling of the mixture to form a glass of one or both components gives a metastable system; at equilibrium, two separated phases would be present, one rich in one component and the other rich in the other component. Subinclusions of the dominant constituent within the dispersed particles of the lesser can form in the cooling process as a result of a decrease in miscibility with decreasing temperature. The size of the dispersed particles depends on the polymers chosen, the temperature, and the mixing speed, as well as other processing variables. The dispersed particles in an unstressed blend are expected to be approximately a spherical collection of randomly coiled chains. A spherical inclusion would minimize the surface area present and the interfacial energy. The application of mechanical stress during mixing can lead to elongated dispersed particles.

Various morphologies have been reported for binary homopolymer/block-copolymer blends, including

1. rounded particles of the microphase-separated block copolymer in a continuous homopolymer matrix,
2. small interspersed regions of each constituent,
3. a double-diamond structure, and
4. lamellar channels connected by periodic columns, which is a "mesh" or "lamellar cantenoid" structure.

When one of the constituents in a binary blend of two flexible homopolymers is crystallizable, spherulites or hedrites can appear either from the melt or from a cast film. If the two polymers are miscible, extinction rings, attributed to lamellar twisting (see Sec. 2.2.2), are evident on the spherulites where none exist for the pure polymer. If both homopolymers in a binary blend are crystallizable but do not cocrystallize, spherulites of both with large reductions in spherulite size, changes in the amount of crystallization, or elongated growths can be observed.

2.6 Composites and Laminates

Composites, heterogeneous mixtures of two or more components, can be prepared with a linear-chain polymer or a rigid network as the continuous phase and short fibers or particles of another material, not necessarily polymeric, making up the dispersed phase. Both crystallizable and noncrystallizable linear-chain polymers are used to form composites. Glass fibers, glass beads, carbon fibers, carbon black, kaolin, and mica are some of the particulate materials used. More than one type of filler can be used in the same composition.

The particle size of the dispersed material is usually many times larger than that found in blends. The shape of the dispersed particles will depend on the material used and can be spherical, rodlike, or flat. The degree of orientation of fibrous components will depend on the composition and the sample preparation method; in thick samples fiber orientation varies with sample depth. Various fiber reinforcement patterns can be used, including uniaxially oriented, biaxially oriented, and random. If a composite contains particulates, the degree of dispersion will depend on the materials and the mixing technique.

The components of a composite are incompatible and form separate phases. The dispersed component in a composite will have the same form before and after formation of the composite, although the particle size or fiber length, depending on the material, can be reduced during processing. The morphology of the continuous polymer phase in a composite containing particulates can differ from that for the polymer alone. Particulates can act as nucleating agents for crystallization and lead to a loss of spherulite texture. When certain fibers are present in a supercooled crystallizable polymer melt, transcrystallization occurs (Chatterjee *et al.*, 1975). Cylindrical columns of crystallized material, with diameters the same as those for the free spherulites present, form around the fiber. As a result of crowding on the fiber caused by the nucleation of many potential spherulitic structures, development in the long direction of the fiber is impeded, while radial growth occurs in a normal fashion.

Laminates usually contain alternating layers of two or more materials. Laminates can be formed using pieces of a woven fabric as one component and a liquid prepolymer that polymerizes to give a rigid-network polymer as the other component. Laminates can be prepared containing alternating layers of two linear-chain materials or of layers of a continuous-fiber/rigid-network composite with a change in the fiber orientation from layer to layer.

3. EFFECTS OF PROCESSING ON POLYMER MORPHOLOGY

3.1 Uniaxial Orientation during Crystallization

Application of a uniaxial tensile stress during bulk crystallization of a thin film can give lamellas with the crystallographic *c* axis along the deformation direction (Gohil and Petermann, 1982). TEM investigation of uniaxially melt-drawn polyethylene films, following annealing, shows that adjacent lamellas have many chains in common.

Extrusion of a flexible-chain polymer through a small orifice, such as present in a capillary rheometer, yields tapered interlocking lamellas epitaxially crystallized on extended chains (Bashir *et al.*, 1986). The lamellas are arranged perpendicular to the extrusion direction, while the extended chains and the crystal stems in the lamellas are oriented along the extrusion direction. The lamellar tapering taking place during crystallization is attributed to a temperature gradient existing in the extruder.

3.2 Skin/Core Morphology

Injection of a polymer melt into a mold by the application of pressure causes chains to assume elongated conformations in the direction of the applied stress. A higher degree of orientation is usually achieved with injection molding than by extrusion, because of the forces involved. If the surface of the mold is cold and the sample is thin, solidification will occur quickly, freezing in some or all of the chain orientation throughout the sample. If the surface of the mold is cold, but the sample is thick, cooling of the polymer in the interior is slower than near the surface and morphological differences will be observed due to a return to less elongated conformations of the chains in the interior prior to solidification. For some crystallizable polymers the center portion of thick samples (the core) will contain undeformed spherulites and the surface (the skin) will consist of highly oriented chains. For other polymers the spherulites in a region between the skin and the core have conical shapes due to extended lamellar growth caused by the temperature gradient present. Crystallizable polymers having high glass-transition temperatures exhibit an unoriented noncrystalline skin and a spherulitic core (Sawyer and Grubb, 1987).

3.3 Blend Morphologies

The morphology that results on the extrusion of a blend depends on the relative viscosities of the polymers present, the composition, the processing parameters, such as the flow rate and the temperature, and the sample thickness. At high flow rates the extrudate can contain fibers of the less viscous component; both irregularly shaped rods and spheres have been observed in the same sample. Injection molding of blends can lead to polymer degradation, thereby yielding unexpected morphologies.

3.4 Biaxial Orientation during Molding

If biaxial orientation occurs during molding, lamellas stacked normal to the extrusion direction and adjacent piles connected by fibrils in the extrusion direction are observed, as a result of sequential extension in two perpendicular directions.

4. MORPHOLOGICAL TRANSFORMATIONS OF POLYMERS

4.1 Thermal Effects

A temperature increase can bring about the following:

1. a change from an isotropic or liquid-crystalline glass to an isotropic or liquid-crystalline liquid, respectively,
2. transformation of a microcrystalline material with glassy interlayers to a microcrystalline material with rubbery interlayers,
3. a transition from a microcrystalline state to a liquid-crystalline or isotropic liquid state at the equilibrium melting temperature,
4. a change from one crystal form to another, or
5. a first-order transition of a mobile liquid-crystalline fluid from one type of order to another type or to a disordered isotropic liquid.

These five processes can usually be reversed. A crystal form change occurring via melting and recrystallization is not reversible and two crystal forms can exist together without spontaneous conversion to one form. Crystallizable polymers supercooled to a glass can be crystallized by heating to a temperature where the polymer is below T_m but above T_g for a time that allows adoption of the appropriate conformation of each repeat unit involved.

The morphologies before and after a cyclic thermal process may not be the same. Heating amorphous block copolymers cast from solution to a temperature above T_g for all of the components present, followed by cooling, can lead to morphological changes. When polymer crystal lamellas are annealed at temperatures above the crystallization temperature but below the equilibrium melting temperature, the lamellar thickness increases; upon cooling a change back to the original lamellar thickness does not take place.

4.2 Mechanical Stress Effects

Lyotropic liquid-crystalline mesophases subjected to shear exhibit banding that is perpendicular to the flow direction. This band formation is caused by mesogen disorientation with a periodic variation in the director, and it occurs upon removal of the field and evaporation of the solvent, if present. If the system remains in a mobile state after the bands are formed, a very slow process involving transition from banding to a polydomain situation takes place. For some fibers obtained from liquid-crystalline mesophases, banding disappears upon annealing.

Upon tensile stress application isotropic rubbers will easily deform, giving elongated chain conformations; a rubber specimen will recover its original dimensions and conformational distribution on stress release, if the system is a crosslinked one. Triblock copolymers containing a rubbery component present as spherical shells and two glassy components can be elastically deformed and returned to their original dimensions upon stress release. Cylindrical inclusions of a rubbery component in a glass matrix, subjected to a shear stress applied at a right angle to the direction of the cylinders, rotate by 90° into the shear direction.

Glassy homopolymers and blends undergo localized plastic deformation and form crazes, networks of voids, and fibrillar elements oriented in the stress direction (Kausch, 1983). Compressive stress can cause the formation of shear bands, due to the lateral movement of sections of the material without voiding. Shear banding and crazing can occur together in ductile glassy polymers under oscillatory stress.

Microcrystalline polymers will yield and undergo relatively large amounts of plastic deformation, which remains after stress release; this process is known as cold drawing. Tensile stress applied to single lamellas that have been deposited on a polymer film which is subsequently drawn causes

1. lamellar thinning in the stress direction and the appearance of a void network in the thinned regions at low stress amounts,
2. the appearance of fibrils in the deformation direction at higher deformations, and
3. shear bands at various angles and buckling more or less at right angles to the shear direction.

For lathe-shaped lamellas deformation also yields kinks and twin bands at lower deformations. For spherulites both localized yielding (inhomogeneous) and elongation and yielding of all parts (homogeneous) have been observed, with the two types usually occurring together. Both interlamellar and intralamellar yielding occur when the deformation is primarily inhomogeneous.

Ultradrawn fibers, tapes, and films, in

which there are few, if any, chain folds, have been prepared from flexible-chain polymers starting with as-polymerized and solution-cast gel samples. Biaxial drawing of dried films of ultrahigh–molecular-weight flexible-chain polymer, obtained by extrusion of a concentrated solution, can give a porous product.

Application of compressive stress to crystallized polymer samples creates shear bands at angles of ±45° with respect to the flow direction, tilting of shear bands, lamellar kinking, orientation and fragmentation, or crystal slip depending on the type and amount of compression and the sample morphology.

4.3 Effects of Solvents and Other Chemicals

Solvents, as vapor or liquid, can cause changes in morphology of a crystallized polymer by a dissolution and recrystallization process, induce crystallization in glassy crystallizable polymers, or induce crazes in glassy polymer blocks under compression. Acids and bases attack both the amorphous and crystalline portions of crystallized polymer samples, causing degradation and chemical modification.

4.4 Environmental Effects

Oxidative degradation at the surface can take place when a predominantly hydrocarbon polymer is heated in air. Environmental stress cracking can occur on polymer samples under constant load.

4.5 Irradiation Effects

Ultraviolet irradiation in air can cause surface embrittlement, raised regions on the surface, and development of spiral defect tracks along which fracture occurs.

Ion etching in air or argon gas strips away parts of polymer surfaces. The electron beam used in scanning electron microscopy can cause the appearance of grooves, cracks, and craters.

GLOSSARY

Blend: A compatible, partially compatible, or incompatible mixture of two or more polymers.

Block Copolymer: A polymer containing sequences of two or more repeat units joined covalently.

Compressive Stress: A compressive force per unit area of surface. Compression can be carried out uniaxially, where pressure is applied to the exposed end of a sample enclosed in a rigid chamber or die, or under plain strain conditions, where pressure is applied at opposite ends of an unenclosed sample.

Configuration: The way in which the atoms in a molecule are bonded to each other. Many isomers, molecules with the same chemical formula but with different atomic arrangements, are possible when the number of atoms is large as it is in polymer molecules.

Conformation: The spatial projection of a molecule due to allowed rotation about covalent bonds. For example, in the butane molecule, $CH_3CH_2CH_2CH_3$, rotation about the CH_2—CH_2 bond can occur at 25 °C, giving different spatial arrangements of the two CH_3 groups with respect to one another. The potential energy for the molecule varies with the rotation angle, and is lowest when the nonbonded atoms are farthest apart.

Crystallinity: The relative amount of crystalline regions in a sample. This is usually expressed as a fraction or a percentage of a two-phase or three-phase system. In a two-phase system the other phase is amorphous or noncrystalline; in a three-phase system a phase with intermediate order exists.

Crystallite: A collection of portions of chains in a three-dimensional repeating array. These portions are usually attached to other portions that are noncrystalline.

Degree of Supercooling: The number of degrees Celsius a crystallizable system is cooled below T_m^0 prior to the onset of crystallization. Crystallization occurs isothermally if its onset follows cooling to a constant crystallization temperature.

Equilibrium Melting Temperature, T_m^0: The melting temperature observed for a homopolymer with an infinitely large crystal stem length. The melting temperature of polymer lamellas containing folded chains is dependent on the crystal stem length and is lower than T_m^0.

Extrusion: A process in which a mobile but viscous material is forced through an orifice, giving the extruded sample a particular shape. If the sample is cooled during extrusion, crystallization or glass formation can occur, causing the sample to retain this shape although the dimensions may undergo a change.

Glass: A liquid state in which the chain segment motion is confined to oscillation about energy minima.

Glass Transition Temperature, T_g: A temperature separating the thermal region in which all or portions of the chains are confined to a particular conformation from the thermal region where the chains have conformational mobility.

Graft Copolymer: A polymer containing sequences of one repeat-unit type attached as branches on a chain of another repeat-unit type.

Homopolymer: A chain containing one type of chemical unit repeated many times with successive units held together by covalent bonds; because of valency requirements the end units are different from the ones in the body. Examples of repeat units are given in the text.

Injection Molding: A process in which a viscous fluid is pushed into a closed area having a particular shape. Following solidification, the mold is opened and the sample removed.

Liquid Crystalline Mesophase: A liquid phase with one-dimensional or two-dimensional order.

Monomer: A small molecule precursor of a synthetic polymer; these can be unsaturated, e.g., propylene, $CH_3{-}CH{=}CH_2$, or butadiene, $CH_2{=}CH{-}CH{=}CH_2$; occur as rings, e.g., ethylene oxide; or contain functional groups, e.g., caproic acid, $NH_2{-}(CH_2)_5{-}COOH$.

Nucleation: The initiation of crystallization by small ordered regions of atoms, molecules, parts of polymer chains, or foreign bodies.

Plastic Deformation: A change in dimensions caused by the application of a stress that leads to a permanent deformation in the sample after removal of the stress.

Rubber: A polymer liquid phase in which the chain segments can undergo conformational changes in the absence of an applied stress.

Shear Stress: A force per unit area applied at right angles to two opposite faces of the sample and in opposite directions.

Statistical Copolymer: A polymer with one type of repeat unit mixed statistically with another in the chain.

Strain: Change in dimensions of a body due to stress application.

Tensile Stress: A tensile force per unit area. Tension is brought about by application of forces in opposite directions at opposite surfaces of a sample.

Yielding: See **Plastic Deformation**.

Works Cited

Anandakumaran, K., Herman, W., Woodward, A. (1983), *Macromolecules* **16**, 563–567.

Bashir, Z., Hill, M. J., Keller, A. (1986), *J. Mater. Sci. Lett.* **5**, 876–878.

Chatterji, A. M., Price, F. P., Newman, S. (1975), *J. Polym. Sci. Polym. Phys. Ed.* **13**, 2369–2383.

Geil, P. H. (1963). *Polymer Single Crystals*, New York: Wiley.

Gohil, R. M., Petermann, J. (1982), *Coll. Polym. Sci.* **260**, 312–318.

Gross, B., Petermann, J. (1984), *J. Mater. Sci.* **19**, 105–112.

Hudson, S. D., Lovinger, A. J. (1993), *Polymer* **34**, 1123–1129.

Hudson, S. D., Thomas, E. L. (1991a), *Chemtracts-Macromol. Chem.* **2**, 73–93.

Hudson, S. D., Thomas, E. L. (1991b), *Phys. Rev. A* **44**, 8128–8140.

Kausch, H. H. (1983), *Adv. Polym. Sci.* **52/53**, 1–347.

Keller, A. (1958), "Growth and Perfection in Crystals," in: *Proc. Int. Conf. Crystal Growth, Cooperstown*, New York: Wiley pp. 499–528.

Kojima, M., Magill, J. H. (1986), *J. Mater. Sci.* **21**, 2651–2656.

Lovinger, A. J. (1980), *J. Polym. Sci. Polym. Phys. Ed.* **18**, 793–809.

Noel, C., Friedrich, C., Laupretre, F., Billard, J., Bosio, L., Strazielle, C. (1984) *Polymer* **25**, 263–273.

Percec, V., Cho, C. G., Pugh, C., Thomazos, D. (1992), *Macromolecules* **25**, 1164–1176.

Sawyer, L. C., Grubb, D. T. (1987), *Polymer Microscopy*, London: Chapman and Hall.

Vaughan, A. S. (1993), *J. Mater. Sci.* **28**, 1805–1813.

Wickramasinghe, H. K. (1989), *Sci. Am.* **261** (4), 98–105.

Wittmann, J. C., Lotz, B. (1985), *J. Polym. Sci., Polym. Phys. Ed.* **23**, 205–226.

Wood, B. A., Thomas, E. L. (1986), *Nature (London)* **324**, 655–657.

Woodward, A. E. (1983), in: C. E. Carraher, Jr.,

J. A. Moore (Eds.), *Modification of Polymers*, New York: Plenum.

Woodward, A. E. (1988), *Atlas of Polymer Morphology*, Munich: Hanser Publ.

Woodward, A. E. (1995), *Understanding Polymer Morphology*, Munich: Hanser Publ.

Wunderlich, B. (1973), *Macromolecular Physics*, Vol. 1, New York: Academic Press.

Xu, J. Woodward, A. E. (1986), *Macromolecules* **19**, 1114–1118.

Further Reading

Bassett, D. C., Frank, F. C., Keller, A. (1963), *Philos. Mag.* **8**, 1753–1787.

Bassett, D. C., Hodge, A. M. (1981), *Proc. R. Soc. London A* **377**, 61–71.

Blundell, D. J., Keller, A., Kovacs, A. J. (1966), *J. Polym. Sci. Part B* **4**, 481–486.

Bovey, F. A., (1982), *Chain Structure and Conformation of Macromolecules*, New York: Academic Press.

Bunn, C. W. (1954), *J. Appl. Phys.* **25**, 820–825.

Chau, C. C., Li, J. C. M. (1983), *J. Mater. Sci.* **18**, 3047–3053.

D'Orazio, L., Mancarella, C., Martuscelli, E., Casale, A., Filippi, A., Speroni, F. (1986), *J. Mater. Sci.* **21**, 989–997.

Durning, C. J., Rebenfeld, L., Russel, W. B., Weigmann, H. D. (1986), *J. Polym. Sci. Polym. Phys. Ed.* **24**, 1341–1360.

Hashimoto, T., Koizumi, S. Hasezawa, H., Izumitani, T, Hyde, S. T. (1992), *Macromolecules* **25**, 1433–1439.

Isayev, A. I., Modic, M. (1987), *Polymer Composites* **8**, 158–172.

Keith, H. D., Padden, F. J., Jr. (1966), *J. Appl. Phys.* **37**, 4027–4034.

Lee, M. P., Hiltner, A., Baer, E. (1992), *Polymer* **33**, 675–684.

Li, W., Yan, R., Jiang, B. (1992), *Polymer* **33**, 889–892.

Maganov, S. N., Shieko, S. S., Deblieck, R. A. C., Moller, M. (1993). *Macromolecules*, **26**, 1380–1386.

Mogi, Y., Hiroyuki, K., Kaneko, Y., Mori, K., Matsushita, Y, Noda, I. (1992), *Macromolecules* **25**, 5408–5411.

Paleos, C. M., Tsiourvas, D., Anaslassopoulou, J., Theophanides, T. (1992), *Polymer* **33**, 4047–4051.

Spontak, R. J., Smith, S. S., Ashraf, A. (1993), *Polymer* **34**, 2233–2236.

Takemori, M. T., Kambour, R. P. (1981), *J. Mater. Sci.* **16**, 1108–1110.

Tanaka, H., Thakur, M., Gomez, M. A., Tonelli, A. E. (1987), *Macromolecules* **20**, 3094–3097.

Thakur, M., Meyler, S. (1985), *Macromolecules* **18**, 2341–2344.

Trent, J. S., Scheinbeim, J. I., Couchman, P. R. (1983), *Macromolecules* **16**, 589–598.

Wittmann, J. C., Lotz, B., Candau, F., Kovacs, A. J. (1982), *J. Polym. Sci., Polym. Phys. Ed.* **20**, 1341–1353.

Zemel, I. S., Corrigan, J. P., Woodward, A. E. (1989), *J. Polym. Sci. Polym. Phys. Ed.* **27**, 2479–2492.

POLYMERS, ELECTRICAL AND ELECTRONIC PROPERTIES

JACEK ULAŃSKI, *Division of Polymer Physics, Institute of Polymers, Technical University of Łódź, Łódź, Poland*

MARIAN KRYSZEWSKI, *Department of Polymer Physics, Centre of Molecular and Macromolecular Studies, Polish Academy of Sciences, Łódź, Poland*

Introduction 497
1. **Basic Concepts of Electronic Structure of Insulating, Semiconducting, and Conducting Polymers** 498
2. **Dielectric Polymers** 499
2.1 Theoretical Aspects 499
2.1.1 Conductivity in dc and ac Electrical Fields 499
2.1.2 Dielectric Polarization in dc and ac Electrical Fields: Dielectric Relaxations 500
2.1.3 Dielectric Relaxations in Polymer Solutions 503
2.2 Experimental Techniques 504
2.3 Materials 505
2.3.1 General Characteristics of Polymer Dielectrics 505
2.3.2 Dielectric Thin Films 506
2.3.3 Electrets; Piezoelectric and Pyroelectric Polymers 506
2.3.4 Liquid-Crystalline Polymers 508
2.4 Electric Breakdown 508
2.4.1 Types of Breakdown and Factors Influencing the Electric Breakdown 508
2.4.2 Electrical Treeing 509
2.4.3 Measurement Techniques 509
2.5 Surface Charging 510
2.5.1 Origin of Surface Electrification 510
2.5.2 Antistatic Agents 510
2.6 Applications 510
3. **Semiconducting and Photoconducting Polymers** 511
3.1 Theoretical Aspects 512
3.1.1 Origin of Charge Carriers 512
3.1.2 Charge Transport 513
3.1.3 Charge Recombination 514
3.2 Experimental Techniques 514
3.3 Materials 515
3.3.1 Polymers with Polyconjugated Multiple Bonds 515
3.3.2 Polymers with Aromatic Amine Groups 516
3.3.3 Polymers with Large Polynuclear Aromatic Groups 516
3.3.4 Inorganic Semiconducting Polymers 516
3.3.5 Molecularly Doped Polymers . 517
3.4 Applications 518
4. **Conducting Polymers** 518
4.1 Theoretical Aspects 518
4.1.1 Basic Concepts 518
4.1.2 Doping and Charge Transport 519
4.2 Experimental Techniques 520
4.3 Materials 521
4.3.1 Synthesis 521
4.3.2 Characteristics 522
4.3.3 Processable Conducting Polymers 524
4.3.4 Heterogeneous Conducting Polymer Composites 524
4.3.5 Superconducting Polymers 526
4.4 Applications 526
Glossary 527
Works Cited 528
Further Reading 529

INTRODUCTION

Polymers in an electric field exhibit a redistribution of charges that are mobile enough to respond in the time scale of the applied field. No polymers are completely free of charges. However, in most polymeric materials, the number of free charge carriers is very small, and their mobility is very low,

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

so that polymers are essentially insulating materials. Low dc conductivity σ may often be attributed to impurities (ions) and a small concentration of electrons or holes that may be injected from electrodes, particularly at high fields. At very high fields, the injected carriers will lead to a variety of breakdown phenomena.

If the concentration of intrinsic thermally generated or photogenerated mobile carriers and their mobility is higher, polymeric materials exhibit semiconducting properties. Usually the conductivity of polymeric insulators and semiconductors varies exponentially with temperature. Because the mobility and concentration of charge carries are thermally activated, the properties of polymeric semiconductors cannot be described using classical models elaborated for covalent semiconductors.

If the diffusion of mobile charges is spatially limited, the dc conductance is absent, but ac conductivity can be observed. If the limitation of carrier mobility is important, e.g., at the phase boundary, interfacial polarization occurs. The polarization may also be induced by the redistribution or partial orientation of dipoles bound to the polymer chains in the direction of electrical field. The polarization effects are related to dielectric properties characterized by the dielectric permittivity, which is frequency dependent. Low concentrations of charges and dipoles in polymeric materials are responsible for their good dielectric properties, which can be modified by the chemical constitution and structure of the polymeric system.

Polymeric materials can exhibit high conductivity when the chemical structure of the polymer chain is such that a large number of delocalized charge carriers can be thermally generated (low ionization potential of the macromolecular chain) or enhanced by molecular doping using donor or acceptor molecules. These systems are polymeric conductors; however, the charge transport mechanism differs from that of metals.

1. BASIC CONCEPTS OF ELECTRONIC STRUCTURE OF INSULATING, SEMICONDUCTING, AND CONDUCTING POLYMERS

All materials when subjected to an electrical field show electronic and atomic polarizations. They depend on the frequency and are active at $\omega = 10^{-12}$ to 10^{-15} s^{-1}. These phenomena are responsible for optical properties. The most important for polymeric dielectrics is the polarization connected with the presence of dipoles bound to the backbone or with the pendant groups. Under an electric field, they undergo orientation that is frequency and temperature dependent. When the electric field is removed, this polarization disappears with time and with temperature increase, and it is called relaxation polarization. It defines the dielectric permittivity ϵ_r, the value of which is in consequence frequency and temperature dependent. For a good insulating material, the concentration of mobile dipoles must be low, and ϵ_r values are low. Usually, nonpolar saturated polymers are oxidized to some extent, and the presence of C=O groups is responsible for an increase of ϵ_r values. High values of ϵ_r are suitable for capacitor materials, but they exhibit dielectric losses at certain frequencies, which is an important drawback. The chemical composition and structure of the polymeric material must be designed to the needs of particular applications (McCrum *et al.*, 1967).

In the solid state, organic molecules, including macromolecular systems, retain their identity and interact weakly through van der Waals or electrostatic forces. The electron–electron interactions within a macromolecular chain are stronger then those between chains. The electrons are moving in a nonperiodic potential, which varies with time within an isolated molecule because of the existence of many possible conformations. Most polymers are amorphous or partially crystalline solids. The lack of periodic long-range order distinguishes polymers from covalent semiconductors. This fact implies that physical phenomena in an electric field are different in polymers than in covalent semiconductors and dielectrics.

When a charge carrier is injected into a polymer, it changes the electronic charge density at the molecular site that it occupies and at the neighboring molecular sites. This results in intramolecular and intermolecular relaxations, which lower the energy of the system by the relaxation energy E_r (Duke *et al.*, 1978). The intermolecular and intramolecular contributions to relaxation energies can be large. For polymers, usually E_r(intra) = E_r(inter) = 1–2 eV; therefore, molecular anions or cations are stable and contribute in polymeric insulators to conductivity and surface charging. The intramolecular contributions depend on the local molecular structure near the injected charge. Local variations in composition and chemical constitution cause spatial fluctuations in the site energies of charge in these materials. Such fluctuations are important because they create localized molecular-ion states in polymers, which are involved in contact charge exchange (i.e., contact charging of pendant group polymers, e.g., polystyrene), and extrinsic localized molecular-ion states that control the transport properties of molecularly doped polymers. The average value of intermolecular contributions to the relaxation energy depends on the dielectric response of the medium and may be determined from the measurements of electronic motion in the system (photoemission, UV absorption, electron transfer).

Polymeric solids consisting of molecular units characterized by low ionization potential or high electronic affinity exhibit semiconducting properties. To this class of polymers belong unsaturated polymers, macromolecular systems containing aromatic or heterocyclic moieties, etc. The most important is the extent of electron delocalization and the energy needed to transfer an electron from the highest occupied to the lowest unoccupied molecular orbit. The "energy gap" is usually above 1 eV, and the generation of mobile carriers as well as their mobility is thermally activated. To the large class of semiconducting polymers belong such systems as show an ability to form charge transfer complexes (CT complexes) usually with low–molecular-weight additives of low ionization potential or high electron affinity and polarizability. Conductivities of semiconducting polymers of the order of 10^{-6} to 10^{-3} S/cm are related to hopping transport on localized states (Stolka, 1988).

Electronic motion can occur along a one-dimensional molecule. Reduced dimensionality implies that the electronic states are governed by collective ground states called Fermi-surface instabilities. In a one-dimensional coupled electron-phonon system, because of the Peierls instability, distortion of the backbone lattice occurs, which creates a bond alternation and a gap at the Fermi surface. Electronic states are also localized as a result of the inherent disorder in polymers. This fact results in low conductivity and thermally activated charge mobilities, in contrast to energy-band semiconductors for which carrier mobility decreases with increasing temperature as T^{-n} ($n > 0$). Polymeric materials characterized by high extent of delocalization of electronic states in the main chain and the ability to undergo redox reactions (formations of molecular ions and/or radicals) due to molecular doping agents are conductors and show high conductivity, on the order of 10^{-1} to 10^{5} S/cm (Ku and Liepins, 1987).

2. DIELECTRIC POLYMERS

The basic theoretical concepts and theories for the dielectric properties are similar for low– and high–molecular-weight materials in crystalline or amorphous phases (more difficult interpretation in the last case). These basic concepts are presented in the article DIELECTRIC PROPERTIES OF INSULATORS, where the problems of local fields and the continuous-medium approximation are discussed.

Dielectrics are used in capacitors and as electrical insulation. The dielectric constant or dielectric permittivity of a material varies with temperature, frequency, its phase constitution, and structural defects. These factors influence the behavior of induced or permanent electrical dipoles in the polymer exposed to a steady or alternating electrical field. When the polarization lags the external field strength, it results in an electrical energy loss that appears as dissipated heat. It is proportional to the product of the relative permittivity and the tangent of the lag angle θ.

Energy loss in polymers arises also from ion and electron migration. Generally, overheating or cyclic heating results in degradation of a dielectric polymer and breakdown.

2.1 Theoretical Aspects

2.1.1 Conductivity in dc and ac Electrical Fields The relative permittivity of a polymer dielectric ϵ_r is defined as a ratio of the permittivity ϵ to the permittivity of empty space ϵ_0:

$$\epsilon_r = \epsilon/\epsilon_0. \tag{1}$$

Inserting of a dielectric into a parallel-plate capacitor whose plate area A is very large and plate spacing d is small would increase the capacity from $C_0 = (A/d)\epsilon_0$ to

$$C = (A/d)\epsilon = (A/d)(\epsilon_0\epsilon_r) = C_0\epsilon_r. \tag{2}$$

For dc, we have

$$V = \frac{Q}{C} = \int\frac{Idt}{C} \quad \text{and}\; I = C\frac{dV}{dt}. \tag{3}$$

If the dielectric contains free charges (electrons, holes, or ions), a dc field leads to a conductance of the material. Potential difference is connected with current density according to

$$\mathbf{J}_e = \mathbf{E}/\rho_e = \sigma_e\mathbf{E}, \tag{4}$$

where $\mathbf{E}$ is the electrical field [V/m] and ρ_e is called resistivity, and its reciprocal σ_e is electrical conductivity. Usually, volume resistivity ρ_e is given as a volume resistance reduced to a cubical unit volume and expressed in Ω m although the unit Ω cm is used frequently. Similarly, volume conductivity $\sigma_e = 1/\rho_e$ is used and expressed in $(\Omega\text{ cm})^{-1}$ or S/cm.

If the sinusoidal voltage is used,

$$V = V_0 \sin\omega t \tag{5}$$

where V_0 is the maximum voltage and $\omega = 2\pi f$, where f is the frequency and t is time then $I = CV_0 \cos\omega t$; i.e., the current is $\pi/2$ out of phase. This equation applies only to an ideal case, i.e., when a loss-free dielectric of $\sigma_e = 0$ is in the capacitor. For a real dielectric, the current lags voltage by $\pi/2 - \delta$, where δ is a measure of power loss:

$$W = \pi f V_0^2\epsilon_r \tan\delta, \tag{6}$$

where W is expressed in watts (W).

Small values of power loss, or loss factor, which is the product $\epsilon_r \tan\delta$, are characteristic of a good dielectric. The dissipation factor is $\tan\delta = \epsilon_1/\epsilon_2$ where ϵ_1 and ϵ_2 are given by the complex electrical permittivity function

$$\epsilon^* = \epsilon_1 + i\epsilon_2, \tag{7}$$

where ϵ_1 is the real part (the dielectric permittivity or electric inductive capacity) and ϵ_2 is the imaginary part (the dielectric absorption). Their dependence on frequency, temperature, and polymer structure will be discussed in Sec. 2.2.1.

A very important factor of major significance in judging the usefulness of an insulator is called dielectric strength. It will be discussed in Sec. 2.4.

Usually, for low-conductivity plastics—dielectrics—the values of resistivity ρ and conductivity σ are time dependent, and the apparent volume resistivity increases with increasing time. This effect is related to complex polarization phenomena that occur when an electrical field is applied. When a step voltage is applied to an insulator, its resistance and capacitance changes can be represented by known equivalent circuits: series circuit and parallel circuit. They can be used for elucidation of the above-mentioned time dependence of the current. A simple relation between current and voltage like Ohm's law is usually valid only for low fields. At higher electrical fields, super-Ohmic behavior is observed and related to the space charge (electronic or ionic charges trapped in various defects of the material under study). The theories of space-charge–limited currents (SCLC) and of the conductivity dependence on temperature and field for disordered dielectric materials are advanced and make possible the basic understanding of the dc conductance (Kao and Hwang, 1981).

2.1.2 Dielectric Polarization in dc and ac Electrical Fields: Dielectric Relaxations

When a dielectric is inserted between the charged plates of a capacitor, dielectric dipoles are formed whose electric moment is given by

$$\boldsymbol{\mu}_e = q\mathbf{r}, \tag{8}$$

where q is a positive charge separated from a negative charge of the same value by the distance $\mathbf{r}$. The origin of the vector $\boldsymbol{\mu}$ is at the negative charge. The total dipole moment per unit of volume is called the polarization $\mathbf{P}$, which is

$$\mathbf{P} = \epsilon_0\mathbf{E}(\epsilon_r - 1). \tag{9}$$

The field at a molecule in the interior of a dielectric between the plates of a charged capacitor is larger than the applied field **E**. Because of polarization that occurs within and on the surfaces of the dielectric, the dipole moment induced in the molecule depends on local field $\mathbf{E}_{\text{loc}}$, thus:

$$\boldsymbol{\mu}_{\text{loc}} = \alpha \mathbf{E}_{\text{loc}}. \tag{10}$$

The variable α is called the polarizability of the molecule. For N molecules per unit of volume, the total polarization is

$$\mathbf{P} = N\alpha\mathbf{E}_{\text{loc}}, \tag{11}$$

and

$$\epsilon_r - 1 = \mathbf{P}/\epsilon_0\mathbf{E} = N\alpha\mathbf{E}_{\text{loc}}/\epsilon_0\mathbf{E}. \tag{12}$$

The value of $\mathbf{E}_{\text{loc}}$ depends on the model accepted for its calculation. The simplest is the Clausius–Mossotti model from which it follows that

$$\mathbf{P} = (N\alpha/3)(\epsilon_r + 2)\mathbf{E}, \tag{13}$$

or

$$N\alpha/3\epsilon_0 = (M/d)(\epsilon - 1)/(\epsilon + 2); \tag{14}$$

here, M is molecular weight and d is the density. Debye (1945) derived the relation between the permanent dipole moment $\boldsymbol{\mu}$ and dipolar polarizability:

$$\alpha_0 = \mu^2/3kT. \tag{15}$$

The polarization can be divided into two parts: distortion polarization, which is independent of temperature (electronic polarization and atomic polarization), and dipole polarization. Thus we can write the following:

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} - \frac{\epsilon_\infty - 1}{\epsilon_\infty + 2}\frac{M}{d} = \frac{N}{3\epsilon_0}\frac{\mu^2}{3kT}, \tag{16}$$

where ϵ_∞ is dielectric permittivity at very high frequency, related to electronic polarization.

The reexamination of the problem of the internal field by Onsager (1936) leads to a relation that takes into consideration interactions between the molecule and the field. When ϵ_r is large compared with ϵ_∞ (which is always small), the Onsager equation can be expressed as

$$2\epsilon_r/(\epsilon_r + 2)^2 \approx N\mu^2/3kT. \tag{17}$$

It implies many important correlations:

1. The dielectric constant is directly proportional to the number of polar molecules per unit volume.
2. The dielectric constant is proportional to the square of permanent dipole moment.
3. The dielectric constant is proportional to the reciprocal of the absolute temperature.

This equation is true for not very strong polar liquids. For polar liquids with hindered rotation of dipoles and for solids, Kirkwood (1939) generalized the treatment of the internal field. The use of the Kirkwood equation for polymers is not straightforward because of the mutual influence between segments of the chain so that the effective dipole moment decreases.

Orientation polarization occurs when the applied field interacts with the permanent electric dipoles that arise from the asymmetry of the molecule. For most materials, tanδ varies with frequency; the frequency of tanδ maximum is known as the relaxation frequency. The number and amplitudes of individual loss processes depend on the number of dipolar species and on their possible separate or cooperative motions (McCrum *et al.*, 1967). The dipole polarization is connected with reorientation of molecules or molecular segments; thus, the process is interfered with by the "friction" from the surrounding molecules. The reorientation of dipoles involves motion in viscous media, and the polarization process is delayed, reaching an equilibrium only after a relatively long time. When the electric field is removed, a long time is also needed for the polarization to disappear. This phenomenon is called dielectric relaxation. The time needed for the polarization to decay by $1/e$ of the original equilibrium value is called the relaxation time, being the reciprocal of the relaxation frequency.

Relaxation processes are evidently thermally activated, but no simple Arrhenius-type formula can be used for determination of the thermal activation energy and of the shift of the tanδ maximum with temperature. It is usual to plot the dielectric permittivity

ϵ_1 or ϵ_2 as a function of temperature or frequency as is shown in Fig. 1.

The identification and localization of loss peaks in different polymers is usually carried out using the nomenclature first suggested by Deutsch *et al.* (1954). The relaxations observed at the highest temperature (constant frequency) or at the lowest frequency (constant temperature) are called the α relaxations, and the others are called β, γ, etc. in order of decreasing temperature (or increasing frequency).

The broadening of relaxation spectra is still a subject of many experimental and theoretical studies. It is known that the simple Debye relation does not fit the experimental data for polymers, particularly for α and β relaxations. Semiempirical equations describing the broadening of the relaxation spectra [$\tan\delta = f(T)$ or $\tan\delta = f(f)$] have been suggested by many authors, e.g., Havriliak and Negami (1967). More advanced theories describing the temperature dependence of $\tan\delta = f(T)$ [or $\epsilon_1 = f_1(T)$ and $\epsilon_2 = f_2(T)$], including extended exponential relations as well as mode-coupling theory, make it possible to separate individual peaks of which the dielectric relaxation spectrum consists and to correlate the observed maxima with specific motion of particular polar groups or polar segments.

Molecular interpretations of the various loss regions should also take into account the physical structure of polymeric materials. In amorphous polymers, the loss region associated with the glass–rubber transition observed above the glass transition temperature T_g is called the α relaxation. It is related to the hindered rotation of the main chain. Below T_g, this backbone segmental motion is frozen in, and only local movements of small portions of the main chains as well as motions of pendant polar groups occur. Usually, there is a mixture of these effects, which are not easy to separate.

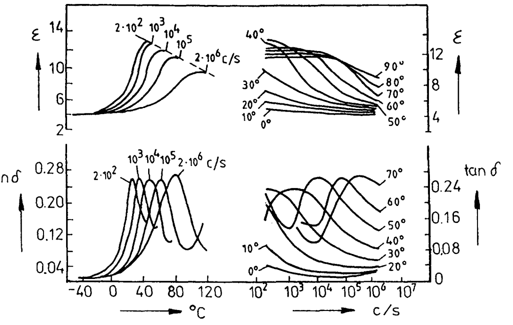

FIG. 1. Dispersion curves of ϵ and of $\tan\delta$ for poly(vinyl acetate) (Wurstlin, 1951).

In crystalline polymers, the interpretation of the loss peaks is complicated (McCrum *et al.*, 1967). At high temperature, low-frequency losses are often called α relaxations, but they are usually associated with the crystalline regions. There are further β, γ, etc. relaxations within disordered or amorphous phases. The highest amorphous peak is attributed to large segments of the main chain (associated with T_g) in amorphous phase, and the lower temperature peaks have a similar meaning as mentioned earlier.

The conformation and position of side groups play an important role in the mobility of the polymer chains, since they determine both the steric hindrance and chain packing. The chain conformation may, in a specific case, facilitate the formation of intermolecular or intramolecular hydrogen bonds leading to lowering of the mobility of main-chain segments and/or side groups. Bulky groups attached to the main chain with lower mobility increase the activation energy and relaxation time of the relaxation. It is interesting to note that large groups attached to side chains will increase partially the main-chain mobility; they still increase the hindrance, but the adjacent polymer chains are farther away. Those two effects will define the magnitude of the main-chain relaxations. The addition of low-molecular-weight material (e.g., plasticizer) to a polymer results in an increased separation of the chains, and the main-chain segment mobility increases. Addition of the plasticizer has only a small effect on side-chain relaxation because the side groups themselves act as plasticizer (internal plasticization). When the added molecules are able to form charge-transfer complexes with groups of main or side chains, an antiplasticization effect is observed to decreased molecular mobility.

Dielectric materials contain—as was mentioned before—freely migrating charges as free ions and/or electrons or holes. The migration of these charges in a dc field causes negative or positive ions to be accumulated on the positive or negative electrode, respectively. Polarization caused by charge migration is called space-charge polarization or ion polarization depending on the nature of the migrating charges. These polarization effects occur simultaneously with the polarization connected with dipole reorientation. Differentiation between the two processes, which show similar behavior, is often difficult. Much effort has been devoted to separating these two phenomena in experiments and theory, but this problem remains still actual.

Similarly to dc resistance, both dielectric permittivity and tanδ are related to the presence of moisture; ac properties are sensitive to the presence of moisture even in the absence of a continuous, low-resistance path. Thus, ac measurements respond precisely and quickly to moisture penetration (polymers with fillers, laminates, etc.). The ac properties are also sensitive to thermal aging, particularly to dielectric discontinuities such as voids, delaminations, etc.

Since ϵ_1, ϵ_2, and tanδ are determined by molecular structure, it is obvious that mechanical losses and electrical losses are closely related. One should, however, say that the mechanical and dielectric loss peaks do not always coincide because nonidentical motions are involved, e.g., nonpolar groups do not contribute to dielectric properties.

If a polymeric dielectric consists of chains built up of bulky, highly polar subunits whose orientation may not be random, the dielectric permittivity ϵ is dependent on direction and described by a third-rank tensor. The anisotropy of dielectric permittivity is important for discussion of nonlinear optical properties of polymers at high electric fields. In characterization of dielectric properties of polymers, the anisotropy of ϵ is usually neglected.

When the dielectric consists of two phases (two components) and one of them with dielectric permittivity ϵ' or conductivity σ' is dispersed in the second phase (matrix) (with ϵ'' or σ''), interfacial polarization occurs and energy is dissipated. This phenomenon was first discussed by Maxwell in 1883 and Wagner in 1914 (van Beek, 1967). The average bulk dielectric permittivity ϵ can be easily calculated, and it depends on the volume fraction of the dispersed phase. The separation of the expression that describes the dielectric permittivity ϵ of the dispersion into real and imaginary parts leads to formulas that contain the conductivity of the better conducting phase. The frequency dependence of the interfacial polarization is identical with that of simple (Debye-type) relaxation. This treatment is applicable for spherical particles without interactions. In polymeric materials, these conditions are rarely fulfilled, and particle shape is important in determining the magnitude of the effect; therefore, the effects of the shape of conducting dispersed particles (ellipsoids) and their orientation parallel or perpendicular to the field has been considered. Interfacial polarization occurs at frequencies below 1 MHz. Dielectric loss in an ac field and its maximum frequency depends on electrical properties of the components and on their distribution.

2.1.3 Dielectric Relaxations in Polymer Solutions The studies of dilute polymer solutions, where intermolecular effects are minimized, lead to much important information on the influence of chain structure on the dielectric behavior of polymers (Stockmayer, 1967). Dipole geometry defines the possible reorientations in electrical field. The dipoles can be connected with

a. flexible side chains (e.g., polyarylates),
b. rigid groups or atoms bound to the backbone [e.g., poly(vinyl chloride)], or
c. rigid groups in the backbone (e.g., polyesters).

Macromolecular dipoles arise from the vectorially summed contributions of individual chain dipoles (Volkenstein, 1963). The determination of dipole moment is a valuable pointer to polymer chain configuration and to polymer sequencing. The rotational isomeric-state model is often applied for calculation of the mean square dipole moment $\langle\mu^2\rangle$. One has to take into consideration the above-mentioned dipole contributions. If there are flexible side-chain dipoles, it is difficult to quantify how those dipoles contribute to the total polarization via backbone reorientation. The perpendicular dipoles are most common and represent the only dipoles present in poly(vinylhalides), unbranched polyoxides, polyimines, and polysulfones. The chemical structure of many monomers results in a long-chain dipole per residue; thus, they may make no contribution to $\langle\mu^2\rangle$ because of sequences of plane reflections. Perpendicular dipoles sum over all chain configurations regardless of long-range interactions, whereas parallel dipoles are subjected to some constraints such as skeletal bonds, leading to excluded volume. The knowledge of $\langle \mu^2 \rangle$ can be useful information concerning the relative equilibrium population of *gauche* and *trans* configurations in the system when reliable experimental data are available.

The rotational isomeric-state model is applicable to isolated macromolecules and to the assembly averaged total dipole. This model takes no account of internal field problems (the external field is usually modified by neighboring dipoles). The study of infinite dilute solutions and solutions in solvents of low dielectric permittivity relying on Debye's equation cannot lead to satisfactory results, particularly for macromolecules. The Onsager model with modifications (nonspherical cavities and elimination of discontinuity at the cavity edge) is more useful. It must be pointed out that the polymeric dipole moment is time dependent as a result of micro-Brownian movement of the chains and entanglements that must be correlated. The solution of the problem is the use of the correlation function approach. The theories of Kirkwood and Fröhlich provide a satisfactory basis for interpretation of the experimental results (Block and Walker, 1982).

The frequency dependences of ϵ_1 and ϵ_2 provide kinetic parameters for the relaxation process, and thermal activation parameters are accessible from variable temperature measurements. One should avoid charge migration, which is common in polyelectrolyte solutions or when macromolecular and solvent phase have different conductivities (the interfacial relaxations disturb the correct evaluation of dipole relaxations).

In semidilute solutions, detailed molecular motions or the interactions between molecules are not clarified as yet. The interpretation of the experimental data is usually done in terms of normal-mode process and segmental-mode processes.

2.2 Experimental Techniques

Electrical property measurements can be divided into low-stress phenomena, which are not particularly dependent on the electrical stress level (these measurement techniques will be described below), and high-stress phenomena, which will be analyzed in the discussion of electrical breakdown (Sec. 2.4.3).

Electrical test procedures often used for plastics materials include measurements of conductivity (resistivity) and measurements of dielectric properties. Techniques for the measurements of all important electrical properties of polymers are described in detail by Bartnikas (1987). Techniques for the measurement of resistivity (or conductivity) are described in ASTMD 257. There are three commonly used methods, which are based on the use of

1. a bridge (e.g., Wheatstone) comparison method;
2. a comparison method using a sensitive current-indicating device; or
3. the voltammeter method.

With the best available equipment, values of resistance as high as 10^{17} Ω can be measured. Insulation resistance is a combination of volume and surface resistance; that is why it is important to preserve correct measurement conditions. The leakage current from the guard to the measuring electrode is shunted in such a way that it is not measured. However, errors are introduced if the resistance between the guard and measuring electrode is less than about 100 times the internal resistance of the current-measuring device. When very high resistances have to be measured, the internal resistance, the time constant of the measurement, and the problem of guarding are increased. The surface of the dielectric measured and of the contacting electrodes must be very clean, because contamination strongly affects both the volume and the surface resistances. Electrode design is very important in the measurement of electrical resistance. It is sometimes specified in the test methods for specific materials. Round rather than square electrodes are most common. Thin coatings of conducting paints are useful, but the effect of dispersant or solvent on the plastic should be taken into consideration. Evaporated metal or thin metal foil electrodes are often applied; in the last case, a minimum of the adhesive (e.g., silicone grease) should be used, preserving good contact with the test specimen.

For resistance measurements to determine changes caused by moisture or elevated temperatures, gold electrodes are most suitable at very high temperature or under corrosive condition. Resistance measurement must be made in the desired surroundings and not after removal from the conditioning chamber. Usually, shield leads with high leakage resistance are used through the walls of conditioning chambers, and the shield is connected to the guard circuit during measurements.

For the tanδ or ϵ_1 and ϵ_2 measurements, an oscillating electrical field is applied to a specimen between two electrodes. In view of the importance of dielectric permittivity and tanδ and their dependence on frequency and temperature, there are many developments of instruments and methodology to measure these quantities. Usually, the measuring systems can be divided into dielectric spectrometer and dielectric thermal analyzer. A dielectric spectrometer scans the frequency response of permittivity and tanδ at a series of temperatures, often by Fourier transform of the "current vs time" response to a step change in the voltage. The dielectric thermal analyzer monitors tanδ at a constant frequency as the temperature is increased with time. The dc step response technique covers 10^{-4} to 10^{6} Hz.

Various methods covering the range 10^{-3} to 10^{10} Hz are mentioned here:

1. Schering-type bridge (capacitance bridge) (10^{-1} to 10^{6} Hz);
2. transformer ratio-arms bridge in which the ratio arms are closely coupled transformers; and
3. microwave bridges (10–70 GHz) or transmission lines (0.3–40 GHz).

Selection from these methods allows measurements to be made to the desired levels of accuracy and the frequency required. Resistance measurements are used to measure indirectly the influence of other factors, such as moisture, that may affect electrical strength and are useful in studying resin cure. Recently, ionic conductivity has been applied successfully to this purpose too. Dielectric thermal analysis instruments are used to investigate rheological properties, the glass transition, and other transition phenomena.

2.3 Materials

2.3.1 General Characteristics of Polymer Dielectrics Various degrees of molecular freedom are possible in polar polymers. If the crystalline lattice is highly imperfect and the shape of molecules is symmetrical so that there is considerable molecular freedom, the dipoles in a crystal may be able to make a substantial contribution to polarization effects. It is evident from Table 1 that dielectric permittivity ϵ_1, dielectric loss (tanδ), and electrical conductivity $\sigma = 1/\rho$ are interrelated and increase in about the

Table 1. Electrical properties of some polymers at room temperature (Ku and Liepins, 1987).

Polymer	Dielectric permittivity ϵ_1 (at 100 Hz)	Dielectric loss tan δ (at 100 Hz)	Volume resistivity ρ [S^{-1} cm]
Cis-polybutadiene	2.05	0.00051	3.1×10^{15}
Low-density polyethylene	2.22	0.0039	2.5×10^{15}
Cis-polyisoprene	2.26	0.0094	7.1×10^{16}
Trans-polyisoprene	2.27	0.00038	7.8×10^{16}
High-density polyethylene	2.30	0.00011	2.2×10^{16}
Polystyrene	2.31	0.00029	2.5×10^{16}
Polycarbonate	2.72	0.00115	2.2×10^{13}
Poly(dimethylsiloxane)	2.86	0.00025	5.3×10^{14}
Poly(phenylmethyl, dimethyl)siloxane	2.87	0.00010	3.0×10^{14}
Poly(diphenyl, dimethyl)siloxane	2.90	0.00041	9.8×10^{14}
Poly(methylmethacrylate)	3.03	0.057	1.2×10^{13}
Chlorosulfonyl polyethylene	5.63	0.138	2.6×10^{13}
Poly(ethylacrylate)	5.56	0.013	1.4×10^{13}
Poly(trifluoropropyl, methyl)siloxane	6.85	0.109	2.7×10^{11}
Poly(chloroprene)	7.83	0.639	8.2×10^{9}
Viton (copolymer of vinylidene) fluoride and hexafluoropropylene	8.55	0.0403	4.1×10^{11}

same fashion as the polarity of the polymer increases.

These insulation properties can be discussed from the viewpoint of dc and ac terms (McCrum *et al.*, 1967). Voltage gradients across insulating systems depend primarily on the leakage rather than on the capacitance. For this reason, insulators for high-voltage application must be formulated in such a way that they show low conductivity. In the case of high-voltage insulation, trapped charges have been taken into consideration, which can be troublesome when a sudden polarity reversal is required.

Specific requirements are to be fulfilled for low–electric-loss polymers, e.g., used in the insulation of transoceanic cables. They are met in general by polyolefins, e.g., polyethylene, which combines very low dielectric loss with moldability, toughness, and flexibility. High-performance thin polypropylene films have also been developed for various types of power capacitors. Extremely low-conductivity materials such as those used in electret microphones are required for various types of high-power capacitors. High-purity, pinhole-free, chemically inert poly(*p*-xylylene) (Szwarc, 1976) and its chlorine-substituted ultrathin film have been developed as insulation barrier coatings for semiconductor surfaces. High-temperature resistant polymers like various types of polyimides are often used as thin insulating layers in microelectronics.

2.3.2 Dielectric Thin Films Dielectric polymer thin films are rarely prepared by thermal evaporation, electron-beam evaporation, or slip casting (melting of the powder) on metallic or semiconductor surfaces; however, these techniques have created more and more interest in the fabrication of microelectronic devices. Mostly used is spin coating from dilute solutions of polymers characterized by appropriate dielectric properties. Care must be taken (rotation speed, temperature, vapor pressure of the solvent, etc.) in order to obtain pinhole-free films. Recently, the spin-coating technique has been used to obtain thin insulating films by spin coating of a dispersion of polymer spheres of low diameter on a metallic or semiconducting surface and melting the obtained monosphere-thick layer (e.g., polystyrene monodisperse lattices). Thicker, but still thin, polymer films can be obtained by extrusion and/or blowing of one-component or multicomponent layered systems.

Chemical vapor deposition and plasma polymerization have been recently applied for obtaining high-quality dielectric films, mostly organosilicon materials. Preparation conditions will decide the quality of the obtained materials, which in many cases are transparent, well adhering to the support, and insoluble.

Very thin films are prepared by polymerization of di-*p*-xylylene. The substrate is exposed to a controlled atmosphere of a pure monomer, and coating occurs as vapor-deposition polymerization (Beach, 1978). If the di-*p*-xylylene is used in plasma polymerization, it forms plasma polymers that contain a minor amount of branched and cross-linked chains, and its solubility suggests a material of low molecular weight. This process has certain similarities to vacuum metallization.

2.3.3 Electrets; Piezoelectric and Pyroelectric Polymers Electrets are dielectrics that create for a long time a strong external electric field. The phenomenon is connected with dipoles oriented during polarization treatment or with the presence of incompensated, stable charge (Sessler, 1982). Permanent polarization is usually obtained by heating or melting and solidifying a polar polymer in an external electric field (thermoelectrets) or by photoexcitation (photoelectrets). There is the possibility to induce charges in a polarized dielectric by γ irradiation (radioelectrets) or by ion implantation. Electrets may be obtained from natural dielectrics—e.g., waxes—or various polymeric dielectrics with low dielectric constant and/or of high T_g.

Most frequently, electrets are obtained by polarization of polar dielectrics in high electric fields. The charge at the surface of an electret is due to the orientation of dipoles; thus, heterocharge is formed. There are also injected charge carriers (homocharge) that disappear much more slowly with time. The resulting charge and the lifetime of an electret depend on the material used and on polarization conditions. There are electrets that can be obtained from nonpolar dielectrics as a result of charge carriers injected from the electrodes or induced by irradiation. Most important are thin-film electrets, which are obtained from good dielectrics by electron-beam irradiation whereby the charge is accumulated at a known distance from the surface of the film. Electrets of high density of charges are also prepared by corona charging and by using controlled breakdown on large surfaces. There is a method of electret preparation by the so-called "wet technique." This consists in contacting the dielectric with a liquid characterized by high conductivity in electric fields. The liquid layer is removed later on.

Electrets are subject to an aging process, but by careful treatment, one can obtain charged materials lasting a long time and creating high electric fields. There are already known techniques that are useful for determination of the charge distribution in the poled material. Important information about the relaxation parameters of oriented dipoles and trapped charges are obtained by well-elaborated, thermally stimulated depolarization techniques (van Turnhout, 1987; see also Sec. 3.2). It makes possible more detailed understanding of electret formation and its aging. Principal applications are electrostatic filters and electroacoustic transducers.

Piezoelectricity is a phenomenon related to the appearance of resultant electric moment connected with charges on the surface of a material under compression or extension (spontaneous polarization). This effect was first observed in crystals with polar axes and without a symmetry center. Pyroelectric effects are similar to piezoelectric in the appearance of opposite charges on the surfaces of a material but under homogeneous heating.

Most piezoelectric polymers are also polymer electrets. Maximum piezoelectricity is roughly related to the dielectric permittivity; thus, it is determined basically by the chemical structure of the polymer. The chain configuration and bulk morphology are equally important for piezoelectricity. Poly(vinylidene fluoride) (PVDF), because of its highly polar C–F bonds and high dipole moment ($\mu_0 = 6.4 \times 10^{-30}$ C m), has the highest dielectric constant ($\epsilon = 8.0$) and shows the highest piezoelectric constant (Wada, 1982).

Fluoropolymers and copolymers with vinylidene fluoride are good piezoelectric polymers. Other polar groups—e.g., C–Cl ($\mu_0 = 7.0 \times 10^{-30}$ C m) and C–CN ($\mu_0 = 12 \times 10^{-30}$ C m)—may also lead to useful piezoelectric macromolecular materials. Poly(vinylidene chloride) chains, because of the presence of bulky chlorine atoms, are prevented from adopting the highly polar all-*trans* conformation. The poly(vinylidene cyanide–vinyl acetate) alternating copolymer exhibits a piezoelectric constant similar to PVDF.

The very important structural feature of PVDF is high crystallinity (40–60%). It can crystallize in three major crystalline forms—α, β, and γ. The α form is obtained when the polymer crystallizes from the melt below 150 °C. The structure of the α crystals is monoclinic with the chains having *trans-gauche-trans-gauche* conformation and a dipole moment that has components normal and parallel to the chain axis. The dipole moments of two chains in the unit cell are antiparallel, and the crystal shows no spontaneous polarization. Drawing the PVDF film below 130 °C to several times its original length results in a new crystal form, the β form. This crystal has the orthorhombic unit cell in which the chains have all-*trans* conformation. These chains have a large dipole moment ($\mu_0 = 7.06 \times 10^{-30}$ C m) per monomeric unit, and they are packed in such a way that the crystal exhibits spontaneous polarization P_c (for the ideal crystal $P_c = 0.3$ C/m^2). In real β-form crystals, P_c is smaller because of defects. The stable β-form crystals should, however, have more than 3% of defects (head-to-head and tail-to-tail defects). The conversion from the α to the β form depends on draw ratio and draw temperature. The β form of PVDF crystals is characterized by the highest molecular packing. The γ form is obtained by crystallization under atmospheric pressure, just below the melting temperature, or by melt crystallization under high pressure. The γ-form crystals can be transformed into the β form by mechanical deformation of the crystal.

The PVDF electrets are prepared by annealing and poling *c*-axis oriented films containing β-form crystals. Typical poling temperature is T_p = 80–120 °C, and the electrical poling field is E_p = 50–150 MV/m. The induced surface charge is opposite to the poling field (heterocharge). There is a small part of irreversible polarization; the remaining part is stable below T_p. The mechanism of the polarization is still a subject of discussion. It is accepted that the piezoelectricity and pyroelectricity of PVDF and its large polarization depend on the number of crystallites in the β form and on its ability to align during polarization. It is correlated also with the dipole residues from amorphous regions and their incorporation into crystallites at grain boundaries. Thus, both chemical structure and physical factors play important roles.

2.3.4 Liquid-Crystalline Polymers

Liquid-crystalline polymers (LCPs) are of particular interest because of their very interesting mechanical properties in the oriented state and their nonlinear optical behavior. Liquid-crystalline side-group polymers contain dipole moments related to the pendant group (some LCPs also have a dipole moment in the backbone chain). The dipole character of LCPs makes it possible to study the molecular dynamics of those systems using broadband dielectric spectroscopy covering the frequency range 10^{-4} up to 10^{10} Hz (Kremer *et al.*, 1990). Thus, it is possible to study both fast processes (e.g., rotation of mesogenic groups around their long molecular axis) and slow relaxations (e.g., relaxation close to T_g).

For LCPs with side groups, five different dielectric relaxations have been found. For combined main-chain/side-group systems, the systematic chemical modification of dipoles revealed three relaxation processes. The α relaxation corresponds to the dynamic glass transition. The rotation of the main-chain mesogen around its long molecular axis is called β_m relaxation, whereas β_s relaxation is correlated with the rotation of the side-group mesogen around its long molecular axis. It is worth noting that for side-group LCPs, a collective relaxation process (δ process) was found and ascribed to the hindered rotation of side groups around the polymeric main chain. Similar behavior was found for combined main-chain/side-group liquid crystalline polymers as well as for some copolymers, in which it is suppressed at the smectic-crystalline transition.

The dielectric relaxation studies also give valuable information on the mobilities in various smectic phases depending on temperature. Recently, low-molecular-mass ferroelectric liquid crystals have been investigated. They show two different collective modes. Some main-chain/side-group LCPs have ferroelectric smectic phases as well. They appear when these system are oriented as a result of surface effects and/or for low magnetic fields. This allows for the combination of typical polymeric features with the properties of the ferroelectric smectic phases. Thus, in addition to their conventional use in displays, a wide field of new applications may be expected, e.g., as piezoelements made from cross-linked elastomers or as pressure sensors.

2.4 Electric Breakdown

2.4.1 Types of Breakdown and Factors Influencing the Electric Breakdown

Electric breakdown occurs when the current flowing through the sample increases, leading to its failure, and the applied voltage cannot be stable (Fisher, 1982). The applied voltage leading to material disruption is called breakdown voltage, and the corresponding voltage gradient at failure is electric strength in kV/mm. The knowledge of the breakdown mechanism is not complete, but one can distinguish the following:

1. Intrinsic breakdown, which is related to pure and defect-free material and occurs at the highest possible voltage. It is due to the increased random motion of electrons, which at the electric breakdown gain higher energy than they lose to the system.
2. Thermal breakdown, which is related to heat emission. If the rate of heating under voltage exceeds the rate of cooling by thermal transfer, thermal breakdown may occur at a limited section of the material. These restricted paths of breakdown are called Wagner breakdowns.
3. Gas-discharge–dependent breakdown is related to electric discharges in gas, which may occur on the surface or voids and on various discontinuities of electrical insulation. It can be due to ionization-induced surface flashover or arcing.
4. Physical-defect–dependent breakdown is affected by all types of cracks, voids, foreign inclusions, and factors that introduce physical deterioration of plastics.

Actual electrical breakdown may involve

two or more types simultaneously or sequentially. The true intrinsic breakdown can never be achieved experimentally, but experiments carried out at low temperature with dc voltage applied for very short times are taken as the intrinsic value.

If the thermal breakdown tests are made at high frequencies (above 1 MHz), failure occurs at relatively low voltages. Usually, $\tan\delta$ increases with increasing temperature, and the rate of heating rises rapidly. Many factors influence the thermal breakdown voltage: geometry (particularly the thickness of the sample), thermal conductivity, dielectric permittivity, absorbed moisture, etc. It depends also on the general tendency for physical degradation of the plastic under study.

Surface discharges are a source of ionization and may cause slow degradation of the polymeric material at its surface or within cracks and voids. Progressive surface degradation (tracking) is particularly important in the presence of surface contaminations. Corona and partial discharge are local effects occurring in gas where the voltage locally exceeds the gas-breakdown potential.

Breakdown connected with physical defects depends also on mechanical stress and vibrations. This type of failure depends on the nature of defects, the time under electrical stress, contaminations, and chemical constitution of the polymers. There is a certain similarity between electrical strength and mechanical strength because electrical breakdown involves physical destruction. In the same temperature regions, both modulus and dielectric strength show significant reductions in magnitude.

2.4.2 Electrical Treeing Treeing is related to sporadic discharges that are intermittent and not steady like corona discharges (Ku and Liepins, 1987). The mechanism of treeing is very similar to electric breakdown. Treeing breakdown starts in the regions of the highest and most divergent electrical stresses. The insulation system consists always of non-absolutely smooth electrodes, imperfect electrode–insulation interface, and not perfectly isotropic material, as well as there being contaminant particles or cavities. These defects influence the appearance of treeing phenomenon. Small channels of degradation develop into a treelike pattern, which may finally bridge the gap between electrodes and cause the failures. The practical importance of treeing has been emphasized in many cases, particularly for polyethylene cables. Some insulations containing stress concentration or even some kinds of trees do not fail in short periods. In fact, reports of long lifetimes of solid insulation suggest that well-designed and well-constructed insulating systems operate within the limits of electrical, mechanical, and thermal stresses for a long time.

2.4.3 Measurement Techniques Measurements of the electric breakdown strength are standardized, and the equipment and procedures for electric breakdown test are clearly described in ASTM D 149. The important features that have to be fulfilled by each equipment are an appropriate device to measure the test voltage at the specimen and not only the primary voltage of the testing transformer, a device for increasing the test voltage at uniform rate, strictly sine wave, and a device for removing the voltage at breakdown.

Usually, the corona test and the arc resistance test are made in connection with electric breakdown test. Standardized procedures are difficult to be developed, especially for volume corona, i.e., for corona discharges in voids at conducting inclusions. ASTM D 2275 is applied to measure the corona resistance. For dry arc resistance, ASTM D 495 is used. A 12.5-kV voltage (open circuit), which produces an arc between chisel-shaped tungsten electrodes, is applied.

A number of works have been carried out to elaborate various methods of treeing test (electrical and water treeing test), each of them having its advantages. ASTM D 3756 describes a test for electrical treeing between a sharply pointed needle and a rounded needle. The only quite well-standardized test for wet-treeing of cables is that of the Association of Edison Illuminating Companies [IC CS5-79].

Tracking tests have been standardized in the United States. There are four types of tracking test:

1. Dust-and-Fog Tracking and Erosion (ASTM D 2132),
2. Differential Wet Tracking (ASTM D 2302),

3. Inclined plane Tracking and Erosion ASTM D 2303, and
4. Comparative Tracking Index (ASTM D 3638).

2.5 Surface Charging

2.5.1 Origin of Surface Electrification The presence of electrostatic charge is recognized by mutual attraction or repulsion with other surfaces, possibly with small discharges (sparks), and by the dust attraction effect. When brought into intimate contact with a neutral material, e.g., by rubbing, polymers become positively or negatively charged on separation. The one that becomes positively charged acts as an electron donor, while the other, which becomes negative, acts as an electron acceptor (Ritsko, 1982; see also TRIBOLOGY).

The sequence of polymers according to their charging behavior is called the triboelectric series. It is a general rule that if two substances become charged by mutual contact, the substance with higher dielectric constant will get the positive charge. The triboelectric series can be established by contact potential differences.

The microscopic nature of electronic states involved in triboelectricity is still under study (Lowell and Rose-Innes, 1980). The molecular-ion model of bulk intrinsic states provides an acceptable microscopic description of the charge-transfer process, but further advances in the understanding of this phenomenon are needed. Very important is the rate of charge decay. It depends on the side groups that constitute the polymer chain. When side-chain substituents are involved in charging the surface, charge decay is faster. The ability of the polymer to bind ions is also a very important factor. A small amount of impurity can alter the surface charging to a high extent. Surface charging is of importance for the flow behavior of polymer powders and is a key process in all xerographic techniques.

2.5.2 Antistatic Agents The electrostatic charge can be reduced by use of antistatic agents. Even a small surface charge density can give rise to an electric field strong enough to cause arcing in air (electrical charge in carpets in dry air). Antistatic agents can be used in formulations by mixing with the basic polymer, but above all, the polymer surface must be treated to optimize the performance of the system. The tendency of polymers to store electric charge was overcome in the past by addition of carbon black or acetylene black to rubber, by addition of metallic powders to plastics, and by blending of metallic or metal-coated fibers with synthetic fibers. At present, conducting polymers, thin vacuum-deposited metal layers, or formation of charge-transfer complexes at the surface are of interest for various applications.

2.6 Applications

Polymers are used in many electrical applications as primary electrical insulation, or they may have a structural function, independent of an electrical function; thus, choice of an appropriate polymer as an insulator depends on its particular application. An interesting survey of various basic electrical properties of polymers with regard to their applications is given by Ku and Liepins (1987).

Useful dielectrics are usually divided into three principal classes: those with $\epsilon_r < 12$, those with $\epsilon_r > 12$, and ferroelectrics. The last class is represented among polymeric materials by some ferroelectric liquid-crystalline polymers. The first category may be divided into subcategories based on operating temperature; however, this division is yet not fully justified because polymeric materials depending on their chemical constitution and structure can be used as insulators and dielectrics up to 400 °C (depending on time of application of severity of temperature conditions).

Polymers fulfill generally well the typical requirements of an insulator:

1. they usually have high enough dielectric strength to withstand an electrical field between the conductors (e.g., in capacitors);
2. they show high insulation resistance to prevent the leakage of the current;
3. they exhibit a good arc resistance to prevent damage in the case of arcing;
4. they show appropriate mechanical properties to resist shock, vibrations, etc.; and
5. they usually maintain their integrity under a wide variety of environmental haz-

ards such as humidity, temperature changes, and radiation.

For domestic wiring insulation, the best electrical properties are not necessary; thus plasticized PVC is firmly established. Ester plasticizers also change the nonflammability of PVC, but some appropriately chosen compositions may compensate this insufficiency. High-quality polyethylene is used for telecommunication cables while polypropylene (biaxially oriented) and poly(ethylenetherephtalate) films are used in capacitor manufacturing.

Among some relatively new developments in thermoplastics, the polymer blends (e.g., polycarbonate/ABS blends, where ABS stands for a mixture of acrylonitrile-styrene and butadiene-styrene copolymers) are chosen for connectors and switch casings, where high impact strength, good surface quality, and low mould shrinkage are required. The low resistivity and high dielectric strength of these materials are relatively unaffected by external moisture, and so they are suitable for waterproof switches.

The permanency of storage of real and polarization charges depends on conductivity, carrier mobility, and dipolar relaxation. Insulating polymers may show a very good long-term charge-storage capability because of the low conductivity and low carrier mobility. Typical examples of polymers that can store real charges and polarization charges are polyolefins and poly(ethyleneterephthalate). Poly(tetrafluoroethylene) (Teflon™) and its fluoro-ethylene-propylene copolymers are examples of materials with good insulating properties, and poly(vinylidene fluoride) is characteristic of small dipolar relaxation.

Applications of polymers in the electronic industry are continuously increasing. Polymer piezoelectrets are used in microphones and small loudspeakers. Polymer dielectrics are rapidly gaining importance as substitutes for inorganic (ceramic-based) insulators. These applications are dominated by polyimide films, which offer particularly useful characteristics:

1. stability up to temperatures as high as 450 °C;
2. high electrical resistivity and breakdown strength, and low dielectric loss;
3. excellent chemical resistance; and
4. coating of substrates with thin films is feasible and patterning can be achieved in lithography.

All these features make it possible for this type of polymer to be applied in microelectronics as microresists and printed circuit boards (Saone and Martynenko, 1989).

Evidently, polymers are used as molded components for chip encapsulation. The encapsulant must be very pure material with an excellent barrier to moisture and contaminant ions, independently of superior electrical, physical, and manufacturing properties. Many organic polymers fulfill these conditions, both thermosetting and thermoplastic materials.

In addition to good dielectric and insulating properties, the polymers are widely used because of good processing characteristics (small and intricate shapes can be readily and economically manufactured), good mechanical properties (resistance to impact loads and stiffness, even at elevated temperatures), dimensional stability, and retention of many properties in hot and/or moist environment over long periods of time.

3. SEMICONDUCTING AND PHOTOCONDUCTING POLYMERS

The conductivity of every material can be enhanced when charge carriers are generated or injected by appropriate radiation. If UV, visible, or near-IR radiation is used, we deal with photoconductivity. In the case of x-ray, gamma-ray, or other, e.g., electron, irradiation, the term radiation-induced conductivity is usually used. High-energy irradiation inevitably results in chemical degradation of the polymer, and these problems are outside of the scope of this article. An ideal, electronically photoconductive polymer should have no free charge carriers (thus, it should show very low conductivity in the dark) and high quantum efficiency of charge-carrier photogeneration, and it should allow fast transportation of the photogenerated charges.

Three classes of polymers are known to exhibit semiconducting and photoconducting properties: polymers with π-polyconjugated multiple bonds (like polyacetylene), polymers with σ-bonded silicon or germanium atoms in the main chain (like polysilanes), and polymers with saturated backbones and pendant aromatic groups [like poly(N-vinylcarbazole)(PVK)]. Another class of polymeric semiconductors comprises molecularly doped polymers consisting of an electrically insulating and inactive polymer matrix doped with charge-transfer (CT) complexes forming additives.

In polymers with extended π-electron systems in the backbone, the molecular p orbitals broaden into filled and empty bands, and the polymer should be an intrinsic metal. However, because of bond alternation (Peierls distortion), there is a finite gap (e.g., around 1.4 eV for polyacetylene) between these bands.

In silicon-backbone polymers, significant delocalization of electrons occurs along the chain, and their electronic behavior is very different from analogous σ-bonded carbon-backbone polymers; the band gap in polysilanes is of the order of 4 eV, while in carbon analogs it is close to 8 eV.

Pendant-group polymers have electronic states similar to those of the pendant aromatic molecules, and so their properties are similar to those of the polymers molecularly doped with chromophores, where the charge transport occurs via hopping among localized sites associated with pendant groups or dopant molecules.

3.1 Theoretical Aspects

3.1.1 Origin of Charge Carriers The charge carriers necessary to transmit the electrical charge through a material can originate from four basic processes: thermal generation, photogeneration, injection, or photoinjection from electrodes. In general, polymers can be p-type or n-type transporting materials. Practically all photoconductive polymers are p type, and only holes are mobile. Although some potentially electron-transporting polymers have been synthesized [e.g., polymers with acceptor-type trinitrofluorenone (TNF) pendant groups], they exhibit poor mechanical properties because of low molecular weight, and they have insufficient electron mobility. Polymeric systems with mobile electrons or bipolar systems can be obtained by doping of the hole-transporting or inert polymers with low–molecular-weight acceptors, for example PVK-TNF.

In polymers, the thermal generation is usually negligible, even in undoped polymers with polyconjugated multiple bonds, because of the large band gap. Photogeneration is intrinsic, when it occurs by photoexcitation of the polymer molecule, or extrinsic, when the light is absorbed by pigments or dyes dispersed in a polymer matrix and the generation process occurs within the dopant molecules. Intrinsic photogeneration is confined to the UV range (i.e., the range of the polymer absorption spectrum), but its efficiency can be increased and action spectrum shifted into the visible range by doping with a suitable electron acceptor that forms a charge transfer complex with, usually donorlike, polymer.

It is generally accepted that the charge-carrier photogeneration occurring within a polymer is a multistep process that can be divided into two main phases, as shown in the scheme in Fig. 2 (Okamoto and Itaya, 1984). The first phase, the efficiency of which is field independent, is the absorption of a quantum by a molecule or a CT complex and its photoionization (most probably via unrelaxed exciplex states), followed by formation of the hole-electron pair, which is separated at some characteristic thermalization length. The efficiency of this first phase is called the primary quantum yield. The sec-

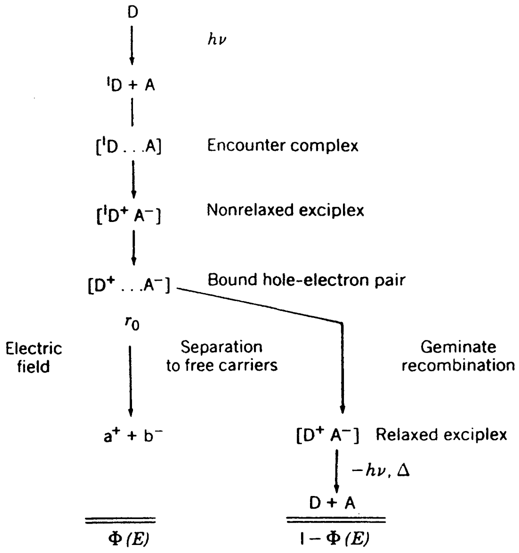

FIG. 2. A scheme of photogeneration of charge carriers; D, electron donor; A, electron acceptor (Okamoto and Itaya, 1984).

ond phase is further separation of the carriers, which contribute to the current if they avoid geminate recombination. This phase is field dependent, its mechanism is similar to the process of ion-pair dissociation in weak electrolytes (Brownian motion in the superimposed field of the gemine carrier and external field), and it has been successfully described by the Onsager theory.

3.1.2 Charge Transport In molecular, disordered solids, the overlapping of electronic orbitals is small, any excess charge is localized on the host molecule, and the existence of extended states ranging over more than several molecules is unlikely; thus, band theory cannot be applied.

In intrinsically photoconducting polymers and in molecularly doped polymers without polyconjugated multiple bonds, the charge-carrier transport occurs via thermally activated hopping (Böttger and Bryksin, 1986) among the chromophore groups or among the dopant molecules. Because all these materials are mostly amorphous and the hopping sites are randomly oriented in space and randomly distributed, there is a fluctuation of wave-function overlap between adjacent hopping sites, so-called off-diagonal disorder. The variations of the polarization energy due to positional disorder result also in the fluctuation of site energies, so-called diagonal disorder. Consequently, there is a very broad distribution of hopping probabilities. This is a reason for the relatively low mobility of charge carriers, in most cases much lower than 10^{-3} cm^2/V s.

If one considers a narrow sheet of charge carriers injected into the sample (like in the time-of-flight measurements; see Sec. 3.2), the distribution of hopping times will cause a spreading of the carrier packet as it propagates through the sample. For a Gaussian distribution of the site energies, the carriers will reach the substrate electrode as a coherent packet (although of increasing width) drifting with time-independent average velocity; the transient time is well defined, and the transport is termed nondispersive.

In many polymer photoconductors, especially at low temperatures, the motion of charge carriers is an incoherent process involving time-dependent hopping frequency, and the transport is dispersive. The current pulse in the time-of-flight experiment is a featureless trace because of the very wide, non-Gaussian spread of the carrier packet. In the pioneering work of Sher and Montroll (1975), the dispersive transport of carriers hopping among a random set of localized isoenergetic sites was described by means of the formalism of continuous-time random walk, with the assumption of off-diagonal (positional) disorder only. At present, it is commonly accepted that the dispersion is mainly due to diagonal disorder. The width of the Gaussian distribution of site energies determines the possibility that carriers attain dynamic equilibrium during a transit, the criterion for nondispersive transport (Bässler, 1984). Monte Carlo simulations for a system with a Gaussian energy distribution in addition to off-diagonal disorder predict all the characteristics of the dispersive transport in polymers: Poole–Frenkiel type field dependence of the carrier mobility, temperature dependence of the mobility, narrowing of the dispersion with increasing temperature leading to a transition from dispersive to nondispersive transport at high temperatures, and changes of the transport properties at the polymer glass transition temperature (Pautmeier *et al.*, 1990). Alternative models based on conventional multiple-trapping transport predict also non-Gaussian dispersive transport, but they are not applicable to polymers with hopping transport.

In spite of the very high concentration of impurities and defects present in polymers, trapping phenomena usually do not affect significantly the hopping transport in photoconducting polymers (Scott *et al.*, 1993). This is because the hole transport sites (chromophore groups or additives) have very low oxidation potential, and it is unlikely that other molecules present in the material (e.g., impurities) can have still lower oxidation potential, unless proper molecules are intentionally chosen and introduced. In terms of solid-state chemistry, the hole hopping transport can be described as a reversible oxidation-reduction process; the localized hole forms a cation radical and can accept an electron from the next neutral molecule, and so the hole ("cation radical") performs a random walk generally in the direction of the applied electric field (Mort and Pfister, 1982).

In polymers with σ-bonded arrays of silicon or germanium atoms in the backbone, the hole transport occurs by hopping via the

σ-derived states rather than via side-chain groups (Stolka *et al.*, 1990). As a result of considerable delocalization of charge carriers along the chain, the transport is nondispersive, and the mobility is around 10^{-4} cm²/V s.

3.1.3 Charge Recombination

Most of the photogenerated charge carriers undergo geminate recombination, and only a small part of the carriers are separated from their initial counter charges. Then, because in semiconducting polymers usually only positive charges are mobile, recombination of carriers occurs when a moving hole meets the immobile electron. The electrons are localized at the photogeneration groups or are trapped in their vicinity. The recombination rate constants in polymers are relatively high, comparable to those found in semiconductors. For materials where both holes and electrons are mobile, like in PVK heavily doped with TNF, the rate of recombination depends on the combination of carrier mobilities μ and concentration n and can be expressed by the equation

$$\frac{dn}{dt} = -\frac{4\pi e}{\epsilon}\mu n^2, \qquad (18)$$

where ϵ is the dielectric permittivity (Pearson and Stolka, 1981).

3.2 Experimental Techniques

The photoconductivity upon steady illumination is given by the equation

$$\sigma_{ph} = g_0 \tau e \mu, \qquad (19)$$

where g_0 is the number of photogenerated carriers in unit volume per second, τ is the lifetime of carriers before recombination, e is the electron charge, and μ is the drift mobility.

Often-used steady-state photoconductivity measurements with sandwich-type samples provided with semitransparent electrodes yield information about the photoconductivity of the investigated material only if the electrode–sample contacts are Ohmic, which is not easily achievable. The measured photocurrent density is equal to

$$j_{ph} = \sigma_d E + \sigma_{ph} E = \sigma_d E + (\phi I_0/d)\tau e \mu E, \qquad (20)$$

where σ_d is the dark conductivity, E is the electric field, ϕ is the quantum efficiency of photogeneration, d is the sample thickness, and I_0 is the total absorbed light (photon/cm² s). The dark current can be measured and subtracted, but the fundamental photoconductivity parameters, ϕ, τ, and μ (in the second term), cannot be separated. For this reason, the steady-state measurements can provide only a comparison of photoconductivity and dark conductivity and the relative action spectra of photoconductivity.

Quantum efficiency and mobility are determined from dynamic measurements by means of time-of-flight (TOF) technique or xerographic discharge (Mort and Pfister, 1982). The TOF method consists of generating a narrow sheet of charges by the use of a short pulse of strongly absorbed light or of an electron beam, and monitoring the current when the carriers drift across the sample in order to determine the transit time τ_T and then the mobility: $\mu = (d/\tau_T)(1/E)$. For nondispersive transport, the transit time is indicated by a sharp drop of the current; for dispersive transport, the current monotonically decreases, and only in the log-log plot of the current vs time can some characteristic point (kink) be distinguished, which is interpreted as the transit time.

In the xerographic discharge technique, the sample is provided only with a bottom electrode. The free surface of the sample is charged to a high potential V_0 by corona discharge, and then the decay of the surface potential, $V(t)$, is monitored after a short pulse of light. If the total charge generated in the sample is small, as compared with the corona-deposited charge, CV_0 (where C is the capacitance of the sample), and if the photoinjection from the deposited charge layer and the charge trapping can be neglected, the quantum efficiency is given by

$$\phi = \frac{1}{I} Ce \frac{dV}{dt}. \qquad (21)$$

When the xerographic discharge experiments are performed under space-charge–limited conditions, which means when the intensity of the light source is high enough to generate in a short time a charge comparable to CV_0, the initial potential decay rate is related to the transit time by the equation

$$\frac{dV}{dt} = -\frac{V_0}{2}\frac{1}{\tau_T}. \tag{22}$$

Consequently, the mobility can be determined from the initial potential decay rate. This method of determination of μ is recommended for materials with dispersive transport, where the TOF technique yields ambiguous results. Space-charge–limited conditions can be easily fulfilled in a two-layer configuration of the sample, when the investigated transport polymer film is provided with highly effective thin photogeneration layer, e.g., evaporated *a*-Se.

Trapping and recombination phenomena can be investigated using thermally stimulated current (TSC) and thermoluminescence (TL) techniques (Bräunlich, 1979). TSC consists in measuring a current flowing as a result of thermal release of trapped carriers in the sample heated with a constant heating rate; TL consists in measuring the light emitted through recombination of the detrapped carriers. The traps in the sample are populated at low temperature by photoexcitation, γ irradiation, or electron-beam or corona discharge. The obtained TSC and TL spectra exhibit maxima that can be attributed to the detrapping from trap levels. The analysis of the shapes, positions, and intensities of the maxima and their dependence on experimental conditions (mainly on heating rate and on electric field) yields information on the trapping parameters and mechanism. Interpretation of the TSC maxima can be ambiguous for low-mobility materials when the shape and position of the maximum can be controlled by the transport process (the so-called transport peak); for these reasons, simultaneous measurements of TSC and TL are recommended.

3.3 Materials

Photoconductivity is exhibited by many different types of polymers prepared by various methods of synthesis, like free-radical polymerization, ionic polymerization, polycondensation, ring-opening polymerization, Ziegler-Natta polymerization, electrochemical polymerization, and photopolymerization. Taking into account their structure, they can be divided into several main groups.

3.3.1 Polymers with Polyconjugated Multiple Bonds Polymers with unsaturated backbone show low-energy optical transitions. The most studied polymer with conjugated π-electron backbone is polyacetylene $(CH)_x$ (PA), which has two possible forms: *cis* and *trans* (Fig. 3; see also Secs. 4.1.1, 4.1.2, and 4.3.2). Pristine *cis*-PA is almost an insulator, with conductivity below 10^{-10} S/cm; it displays photoconductivity only in the UV range. *Trans*-PA has conductivity around 10^{-6} S/cm, is photoconducting, and shows the transport properties of a semiconductor. A large positive value of the thermopower coefficient indicates *p*-type conduction.

Perfectly conjugated polymer single crystals of macroscopic dimensions (even several centimeters) can be prepared by solid-state polymerization of diacetylene monomers. Because of the deficiency of free charge carriers, polydiacetylenes are essentially insulators, with conductivities of the order of 10^{-10} S/cm or less, but they show relatively high photoconductivity in the visible range. Polydiacetylene single crystals exhibit interesting nonlinear optical properties with high third-order susceptibility.

Many other polymers with polyconjugated multiple bonds show photoconductivity, even in the visible and in the near-IR ranges, but their dark conductivities are usually also high (Stolka, 1988).

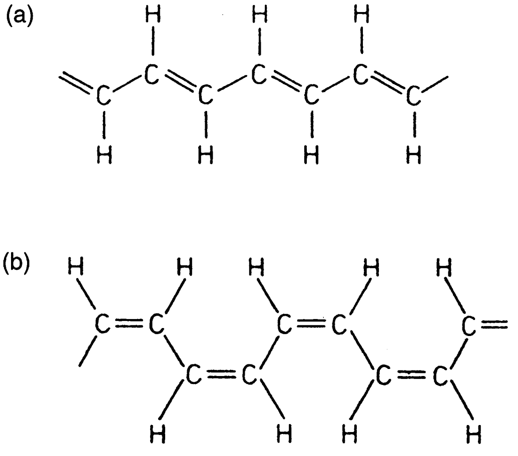

FIG. 3. Schematic structures of polyacetylene in two possible forms: (a) *trans*, (b) *cis*.

3.3.2 Polymers with Aromatic Amine Groups Many polymers containing an aromatic amine group show substantial photoconductivity, among them the most important photoconducting polymer, poly(N-vinylcarbazole) (PVK), this being the first practically applied and commercialized organic photoconductor (Mort and Pfister, 1982).

Pristine, purified PVK shows photoconductivity only in the UV range; the photogeneration spectrum has several steps in coincidence with the PVK absorption spectrum, which means with the excitation into subsequent higher singlet states. The primary quantum yield for PVK is around 0.14, and the thermalization length is in the range 22.5 Å (for excitation into the first singlet state) to 30 Å (for excitation into the third singlet state). The absolute quantum yield for photogeneration of holes is around 10^{-3} at a field of 10^5 V/cm and approaches the value 0.05 at 10^6 V/cm. The hole mobility at a field of 10^5 V/cm is around 10^{-6} cm²/V s.

PVK has been doped with numerous CT complexes forming additives like tetracyanoethylene, tetracyanoquinedimethane, chloranil, or Lewis acids; the most intensively investigated and practically used is the system PVK-TNF. It is photoconducting also in visible light because the CT band of this complex extends through the visible range. For the system with 1:1 molar ratio TNF to PVK monomer, the primary quantum yield in the visible range ($\lambda = 550$ nm) is 0.23, the thermalization length is 35 Å, and the absolute quantum yield is around 10^{-2} at 10^5 V/cm and approaches 10^{-1} at 10^6 V/cm, being the same for holes and electrons. The PVK-TNF system is bipolar; with increasing TNF concentration, the hole mobility decreases below 10^{-8} cm²/V s for the 1:1 molar ratio, while the electron mobility increases up to around 10^{-6} cm²/V s at the field 5×10^5 V/cm.

PVK has rather unfavorable mechanical properties. Its brittleness creates considerable technical problems, especially if it is to be used in the form of thin film. For this reason, several attempts have been made to obtain polymers with carbazole, or other aromatic amine pendant groups, having better application properties: various vinyl backbone polymers, polyacrylates, polymethacrylates, and copolymers of vinyl carbazole with different comonomers (including acceptor-type comonomers). Some of these polymers were sensitized with TNF and other electron acceptors. In a few cases, their mechanical and film-forming properties are improved, but usually at the expense of the photoconducting properties. Even if some of these materials exhibit higher hole mobility than PVK, their photogeneration efficiencies are lower.

3.3.3 Polymers with Large Polynuclear Aromatic Groups A variety of polymers with pendant-condensed aromatic groups have been synthesized and their photoconducting properties have been investigated (Stolka and Pai, 1978). Polymers with naphthalene, pyrene, chrysene, or acridine as pendant groups show usually low photoconductivity, constrained to the UV range only. In some cases (for example in polymers with anthracene groups, like polyvinylphenylanthracene), their photoconductivity is comparable to that of PVK, but because of oxidative instability, poor mechanical properties (brittleness), and difficult synthesis, these polymers have no practical importance.

3.3.4 Inorganic Semiconducting Polymers Polymers with silicon and germanium backbones are insulators in the pure state, but the band gap in some cases is only around 4 eV, i.e., two times smaller than in saturated polymers with a carbon backbone. In some sense, they resemble polyconjugated polymers; the absorption spectra display a red shift effect with increasing molecular weight, typical for polymers with electrons delocalized along the chain. The conductivity of polysilane derivatives can be considerably enhanced, up to around 0.1 S/cm, by suitable doping with a strong electron acceptor, e.g., polysilastyrene becomes a semiconductor when doped with arsenic pentafluoride (Miller and Michl, 1989).

It has been found recently that many polysilane derivatives are excellent photoconductors. In the pure state, they are photoconducting only in the UV range; long chain segments of *trans* conformation are the lowest-energy chromophores. These polymers can be made photoconducting in the visible range by suitable doping, for example with TNF. In poly(methylphenylsilane) (PMPS), the quantum yield of hole generation is electric-field dependent, and for high fields, it approaches 1%.

Similarly as in many amorphous organic and inorganic solids, charge transport occurs via thermally assisted hopping and is controlled by the distribution of the energies of the transport sites and their average separation. Insensitivity of the hole mobility to the nature of side groups shows that the hopping sites are associated with σ electronic states of the backbone. Because of delocalization of the charge along the chain, in polysilanes the transport sites are not localized states (like carbazole pendant groups in PVK) but domains of *trans* conformers separated by conformation irregularities.

In silicon- and germanium-backbone linear polymers, holes are the majority carriers, transport is nondispersive, and mobility around 10^{-4} cm^2/V s is usually reported at room temperature. Carrier mobility is temperature and field dependent and sensitive to the glass transition temperature T_g. Below T_g, the mobility exhibits a single activation energy over a broad temperature range; for PMPS, it is 0.24 eV at 2.5×10^5 V/cm and decreases below 0.2 eV at 6×10^5 V/cm. At high fields and temperatures, the hole mobility may approach 10^{-2} cm^2/V s. The activation energy of mobility decreases strongly above T_g, and in some cases, the mobility becomes practically temperature independent (e.g., for polysilane derivatives with aliphatic side groups). The field dependence of the mobility and of its activation energy in these polymers indicates the Poole–Frenkel effect of barrier lowering by a factor $\beta_{PF}E^{0.5}$; however, the experimentally obtained value of β_{PF} is about two times lower than the theoretical Poole–Frenkel coefficient $\beta_{PF} = (e^3/\pi\epsilon\epsilon_0)^{0.5}$ (Stolka *et al.*, 1990).

In PMPS, the trap-free, nondispersive transport results in a nearly rectangular TOF pulse, with hole mobility independent of sample thickness. The hole mobility can be reduced by doping with molecules with lower oxidation potential, creating traps for holes and resulting in trap-controlled hopping. However, when the energy level of the dopant site is close to the energy level of the transport site in the polymer, dopant molecules are transport active, and heavy doping may result in a substantial increase of the mobility. In PMPS doped with *N,N'*-bis(4-methylphenyl)-*N,N'*-bis(4-ethylphenyl)-[1,1'-(3,3'-dimethyl)biphenyl]-4,4'-diamine, the hole mobility at $E = 2.5 \times 10^5$V/cm increases by ca. three orders of magnitude up to the unusually high value 10^{-1} cm^2/V s (Stolka and Abkowitz, 1993).

3.3.5 Molecularly Doped Polymers

The molecularly doped polymers have gained a lot of interest as photoconducting materials for two reasons. First, in these materials, one can combine good mechanical properties of a chosen polymer matrix with appropriate electronic properties of dopant molecules (Ulański *et al.*, 1981). In particular, it is possible to introduce different types of dopants, e.g., dye molecules in order to extend the photoconductivity action spectrum (sensitization) and transport molecules to enhance the carrier mobility. However, to gain high mobility, up to 50 wt % of dopant (donor or acceptor type) is added to the polymer (Scott *et al.*, 1993), and so the mechanical properties of the polymer matrix are changed to a great extent.

Second, solid solutions of various dopants in an inactive polymer matrix are model systems for the investigation of charge generation and transport in photoconducting polymers. Investigations of polycarbonate (PC) doped molecularly with N-isopropylocarbazole (NIPC) have allowed for direct determination of the influence of concentration of generation and transport sites (i.e., carbazole groups) on the basic photoconduction parameters in PVK. It was shown that the vinyl chain of PVK plays no active role in the photoconduction process. Photogeneration efficiency increases with increasing NIPC concentration because of an increase of primary quantum yield. This results from an increase of the thermalization length due to higher carrier mobility at higher concentration of hopping sites (i.e., dopant molecules). The hole mobility exhibits exponential dependence on the average separation between NIPC molecules. The mobility increases with temperature, and the determined activation energy rises with decreasing dopant content; both the mobility and its activation energy are field dependent. An addition to the PC-NIPC system of a second donor-type dopant, triphenyl amine (TPA), with ionization potential lower than NIPC introduces effective traps for holes, resulting in trap-controlled hopping. The hole mobility first decreases with increasing TPA concentration. However, when the decreasing separation between the TPA molecules allows overlapping of electronic orbitals, another effective channel for hole hopping among TPA molecules is formed, and the hole mobility starts to increase again (Mort and Pfister, 1982).

3.4 Applications

Photoconducting polymers were the first electroactive polymers, and in general the first electroactive organic materials, to be practically used on a commercial scale. Polymer photoconductors, first of all the PVK-TNF system, compete successfully with inorganic materials in application in electrophotography as a result of sufficiently good electrical properties (almost trap-free transport) and the ease of production of large-area, homogeneous films at low price. In practical electrophotographic devices, the organic photoconductors are often used in the two-layer configuration, with separated photogeneration layer and transporting layer. In such a configuration, the problem of removing immobile electrons from the bulk of *p*-type photoconductors is avoided (this problem does not exist in bipolar PVK-TNF).

Large-area solar cells are another potential and much-wanted application of photoconducting polymers. Such devices have not yet been realized because of low conversion efficiency. The low costs of producing are frustrated by poor stability against solar UV light (Stolka, 1988).

4. CONDUCTING POLYMERS

Polymers with extended π-electron systems in the backbone exhibit electronic properties very different from common polymers; they have low ionization potentials and high electron affinities. As a result of these properties, they can be easily reversibly oxidized or reduced by suitable dopants, and, when doped, they can show very high conductivity, in the best cases comparable with the conductivity of metals. For this reason, they are called synthetic metals.

The conducting polymers belong to the class of synthetic metals together with other molecular conductors: charge-transfer complexes made of segregated stacks of acceptor and donor molecules, like tetrathiafulvalene-tetracyanoquinodimethane, and the so-called Krogman salts containing stacks of PtX_4 or IrX_4 (X = CN or C_2O_4), like potassium cyanoplatinate oxidized by bromine. All these systems have in common the quasi 1-D character of the charge-carrier transport, because the charge delocalization is confined to quasi 1-D structures: the polymer chains or the stacks of molecules (Ku and Liepins, 1987).

4.1 Theoretical Aspects

4.1.1 Basic Concepts Conducting polymers are in general disordered systems, and band theory elaborated for crystalline semiconductors should not be directly used; however, periodicity of the π-conjugated backbone makes possible a description of electronic structure in terms of energy bands. An idealized picture of polyacetylene shows a planar *trans*-$(CH)_x$ chain with all the bonds of equal length. Each carbon atom in the chain has in an addition to three σ electrons also one π electron. Molecular-orbital theory predicts that with increasing chain length, the energy gap between valence and conduction bands vanishes, and for high molecular weight, the *p* orbitals of the π electrons would form a half-filled energy band with the Fermi level in the middle. The polymer should be an intrinsic metal. However, the linear-combination-of-atomic-orbitals theory shows that the polyene chain is stable only if it undergoes dimerization, i.e., if alternation of bond lengths occurs. In general, a one-dimensional metal cannot be stable, because all linear chains of atoms undergo distortion and lower their symmetry leading to the creation of collective electronic ground states of reduced symmetry called Fermi-surface instabilities (charge-density waves, spin-density waves, or bond-order-alternation waves).

The self-induced metal-insulator transition, called Peierls transition, is in *trans*-$(CH)_x$ a consequence of the dimerization accompanied by a distortion of the electronic charge density creating a band gap of around 1.4 eV. Perfectly dimerized *trans*-$(CH)_x$ has two equivalent structures, which differ only in the sequence of bond alternations: single/double or double/single. If both structures are present on one chain, they are separated by a topological defect. This radical defect can move by bond flipping along the chain, not changing its energy because these two structures have the same energy (the system is degenerate in energy). At the defect, the π-wave functions change phases, and their amplitudes approach zero; the radical defect is described as a neutral soliton. A solitary wave travels without changing shape or speed on collision with another soliton. A soliton is not a point defect, but it extends over about 15 carbon-carbon bonds. The neutral soliton creates a new state in the band gap; it is a species without charge, but it has spin $\frac{1}{2}$ and can be detected by EPR. The charged soliton is an anion or cation defect, which carries charge but no spin (Lu, 1988).

Spinless transport was observed in PA but also in other polyconjugated polymers that are not degenerate in energy. In such polymers, like polypyrrole, polythiophene, poly(*p*-phenylene), or poly(*p*-phenylene sulfide), the two structures of different sequences of bond alternation have different energies: the aromatic structure has lower energy than the quinoid one. Therefore a mobile soliton cannot exist because it would migrate immediately to reduce the length of the quinoid structure. In these polymer chains, in order to restore the initial aromatic structure on the other side of the defect, the defect has to contain both a soliton and an antisoliton. Such a pair constitutes a polaron, which is a species carrying a charge and a spin $\frac{1}{2}$. In other words, the polaron is a radical ion (cation or anion) associated with distortion of the lattice around the charge. If two such coupled electron-photon states are present on the same chain, they react exothermically and form a bipolaron, a spinless double-charge carrier. The direct evidence of formation of solitons, polarons, and bipolarons comes from optical spectroscopy. Photoexcitation of an electron-hole pair, followed by charge-pair separation, results in the formation of a localized defect characteristic of the given polymer: *trans*-polyacetylene shows a photoinduced absorption band associated with the soliton level, and other polyconjugated polymers, able to form bipolaron states, show two photoinduced absorption bands corresponding to the optical transitions between the band edges and the two polaron levels in the gap (Frommer and Chance, 1988).

4.1.2 Doping and Charge Transport

The fundamental effect of doping in terms of semiconductor physics is to generate free carriers, and only in this respect can the term "doping" be used for the description of the processes of oxidation or reduction of the polyconjugated polymers by electron acceptors or donors. The mechanism of doping of conducting polymers is, however, very different from that for crystalline semiconductors. One of the main sources of this difference is a large deformability of the macromolecules. The dopant molecules are introduced between polymer chains, so that morphological changes occur on doping. Because the dopant molecules are not built into the rigid crystal lattice, they can be removed, and so the doping is reversible. The concentration of dopants in highly conducting polymers may approach 50 wt %, and so the terms "polymer salt" or "charge-transfer complex" are often preferred over the term "doped polymer" (Wegner, 1981).

Because of the charge transfer between the polymer chain and the dopant, the polymer chain is electrically charged; injected charges distort the charge-density waves on the chain and through electron-phonon coupling can alter the macromolecular geometry. In consequence, a defect is formed on the chain, and a new level is created in the band gap. The type of defect formed depends on the structure and state of the polymer. In the case of *trans*-$(CH)_x$, the interaction of the dopant with an existing neutral soliton yields a positive soliton for acceptor dopant (*p* doping) or a negative soliton for donor dopant (*n* doping). Direct oxidation (reduction) of *trans*-$(CH)_x$ by an acceptor (a donor) dopant yields a positive (negative) polaron, introducing two new levels in the band gap. At low doping and at low temperatures, the charged defect on the chain is localized by the Coulombic binding energy to the ionized dopant. With increasing doping, on some chains two polarons will be formed; on collision, they will convert into two uncorrelated charged solitons. The conductivity increases accompanied by a decrease of EPR signal intensity. In bipolaron-supporting polymers, introduction of the dopant yields also an ionized polymer chain, and polarons (radical ions) are formed. With increasing concentration, the polarons interact, eliminating radicals and producing correlated pairs of ions, i.e., bipolarons. The spinless conductivity in these polymers is attributed to bipolaron migration (Frommer and Chance, 1988).

The doping of polyconjugated polymers results in dramatic changes in absorption spectra. In *trans*-$(CH)_x$, upon oxidation by electrochemical doping, a new absorption band emerges associated with the polaron (radical cation); with increasing dopant concentration, this band evolves to the absorption band of the charged soliton (cation). In bipolaron-supporting polymers, the polaron absorption bands emerging at low dopant concentration convert to bipolaron absorption at high doping. The growth of the new absorption bands with increasing dopant concentration is accompanied by a decrease of the absorption band of the neutral polymer.

An unusual property of conducting polymers—spinless conductivity—was observed for the first time for *trans*-$(CH)_x$ with low dopant concentration, and it was related to charge transport by charged solitons. However, an isolated charged soliton cannot jump to the next, defect-free chain because it would require high energy to rearrange the bond alternation. According to the Kivelson model (1981), such a jump is possible if a neutral soliton is present on the next chain. Because the neutral soliton is more readily ionized than the defect-free chain, it is unlikely that enough neutral solitons would remain to support the increasing conductivity with increasing doping. Therefore, Kivelson's model cannot explain the spinless conductivity in *trans*-$(CH)_x$ at higher dopant concentration; evidently, it cannot explain also the spinless conductivity in other polyconjugated polymers that are not soliton supporting. A more general model of spinless transport assumes that it is executed by bipolarons [in *trans*-$(CH)_x$, the bipolaron would be formed by a pair of uncorrelated charged solitons], which can jump between chains with relatively low activation energy (Frommer and Chance, 1988). However, also this model lacks generality to the polymers with the highest doping level, where the conductivity can be described by conventional electronic mechanisms involving hopping or tunneling (Schimmel *et al.*, 1990).

The macroscopic conductivity of a conducting polymer is a function of intrachain conductivity and interchain conductivity. Each of these components is a function of many factors. The intrachain conductivity depends on conjugation length, density, and extent of defects, and interactions with dopants. The interchain conductivity in a general sense comprises not only the charge transfer between adjacent chains, but also the transport between highly conducting regions and between morphological entities, like fibers in Shirakawa-type polyacetylene (Ehinger and Roth, 1985). Therefore, the interchain conductivity depends on various morphological factors: chain packing and the orientation of the chains relative to each other, homogeneity of doping, crystallinity, if any, and size of crystallites, and morphology on the supramolecular scale (size and separation of fibers, grains, etc.) (Wegner and Ruhe, 1989). The models describing the mechanism of conductivity in doped polyconjugated polymers have to take into consideration both the characteristics of low-dimensional systems, e.g., specific charge carriers, like solitons and bipolarons, and the characteristics of polymeric materials, e.g., disorder and heterogeneity at various scales of magnitude, molecular relaxations and phase transitions, possible fractal structures, and percolation phenomena.

4.2 Experimental Techniques

The simple two-electrode method for measurements of the dc conductivity in conducting polymers may yield erroneous results because of unknown contact resistance. Also the temperature dependences of the conductivity and current–voltage characteristics can be affected by contact effects. For these reasons, a collinear four-probe method is preferably used. The method consists in measuring the voltage drop V between two inner electrodes caused by a known current I flowing between two external electrodes; the conductivity of the investigated material is given by the relation

$$\sigma = (I/V)(l/s), \tag{23}$$

where l is the distance between the inner (voltage) electrodes and s is the sample cross-sectional area.

In highly conducting materials, Joule's heat may disturb the measurements of the current–voltage characteristics at higher fields and currents. Also, the conductivity measurements at helium temperatures can be distorted by Joule's heat because of very small heat capacitance at low temperatures. Heating effects can be minimized by use of the pulse technique, consisting in measuring the voltage signals occurring at the inner electrodes caused by short, rectangular pulses of the current.

The conductivity of samples of irregular shapes (but flat and uniform in thickness) can be measured by a modified four-probe method, the van der Pauw technique, where small, pointlike electrodes are placed arbitrarily at the periphery of the samples. Two measurements are performed with commuted current and voltage electrode connections; for calculation of the conductivity value, a correction factor, numerically calculated and published by van der Pauw, is needed. Another modification of the four-probe method, the Montgomery method, is similar to the van der Pauw method but is applied to anisotropic materials; this method allows one to determine components of the resistivity tensor, i.e., the anisotropy ratio.

ac conductivity measurements have proved to be a very useful tool in investigations of the mechanism of charge transport in conducting polymers. The frequency dependences of conductivity are, together with the temperature dependences of conductivity, characteristic of various charge transport mechanisms. Most conducting polymers are inhomogeneous materials, with high conducting regions separated by barriers and low-conductivity regions; these barriers dominate the dc conductivity measurements. The ac conductivity measured at sufficiently high frequencies may give an insight into the conduction mechanism of highly conducting islands.

In the frequency range up to 10 MHz, conventional bridge techniques and automatic impedance analyzers are used. In the GHz microwave range, the method of the perturbation of a resonance-cavity Q is applied. This method consists in measurement of the frequency shift and of the broadening of the resonance line caused by introduction of the sample into the resonance cavity. It allows for calculation of the complex dielectric constant; the real part is related to dielectric properties of the material, and the imaginary part is related to ac conductivity. This method can be applied for samples smaller than the wavelength and the skin depth.

Thermopower measurements are widely used for characterization of conducting polymers, because the thermopower effect is not affected by the charge transport, and so it is less sensitive to inhomogeneities of conducting polymers than conductivity measurements. The analysis of the thermopower data, including the temperature dependence of the Seebeck coefficient, yields information concerning the sign of majority charge carriers, their density and mobility, and the conduction mechanism. In particular, it allows one to distinguish between the metallic and semiconducting types of conductivity (Pukacki, 1993).

4.3 Materials

4.3.1 Synthesis The number of conducting polymers and synthesis techniques is large and fast growing. One can distinguish four major types of synthesis yielding polymers with conjugated backbone: Ziegler-Natta polymerization, electrochemical synthesis, polycondensation, and pyrolysis (Ku and Liepins, 1987).

The most frequently applied Ziegler-Natta polymerization is the so-called Shirakawa synthesis of polyacetylene (PA) using various modifications of catalyst $[\mathrm{Ti}(\mathrm{O}\text{-}n\text{-}\mathrm{C_4H_9})_4\text{-}\mathrm{Al}(\mathrm{C_2H_5})_3]$ system. When $\mathrm{CH{\equiv}CH}$ is introduced over the catalyst solution under appropriate conditions, a $(\mathrm{CH})_x$ film is formed on the surface. The choice of the kind and concentration of the catalyst, and of the temperature (typically in the range 180–450 K), are critical. At low temperatures, the *cis* conformation dominates, yielding a reddish material; at high temperatures, the bluish *trans* conformation prevails.

Typical electrochemical polymerization may be illustrated by the synthesis of polypyrrole (PPy). A single-compartment electrochemical cell with two parallel electrodes (anode made of platinum) contains an electrolyte solution, for example, tetraethylammonium tetrafluoroborate in acetonitrile, and pyrrole. When a current is passed through the cell, under galvanostatic control, electropolymerization onto the platinum electrode occurs, accompanied by simultaneous oxidation. The resulting PPy film, "doped" *in situ* with counterion (BF_4^- in this case), is black and conducting. Other polymers, like polythiophene and its derivatives and polyaniline, can similarly be electropolymerized. All these polymers can be readily reversibly reduced and oxidized electrochemically.

Polycondensation comprises a great number of polymerization techniques, allowing the synthesis of various conducting polymers. For example, poly(*p*-phenylene) can be prepared by oxidative polymerization of benzene with the aid of the aluminum chloride–cupric chloride system, or by coupling of dihalobenzenes with the aid of copper as the catalyst (Ullman reaction), but also by condensation of aromatic nuclei in the presence of organometallic reagents or by decomposition of diazonium salts. Other polymers, like poly(*p*-phenylene sulfide), polyquinolines, polyquinoxalines, or pyrrones, have also been prepared by a variety of polycondensation techniques. The precursor polymer routes constitute a special class of polycondensation consisting in synthesis, as an intermediate, of a "prepolymer" that can be processed (e.g., solution cast as a film and oriented by stretching) and then converted by some elimination reaction to the nonprocessable polyconjugated polymer.

Pyrolysis, when applied to some polymers, like polyacrylonitrile or polyimides, leads to elimination of heteroatoms and formation of extended aromatic structures for which graphite can be a prototype. Charge carriers are delocalized along conjugated ladder-type chains.

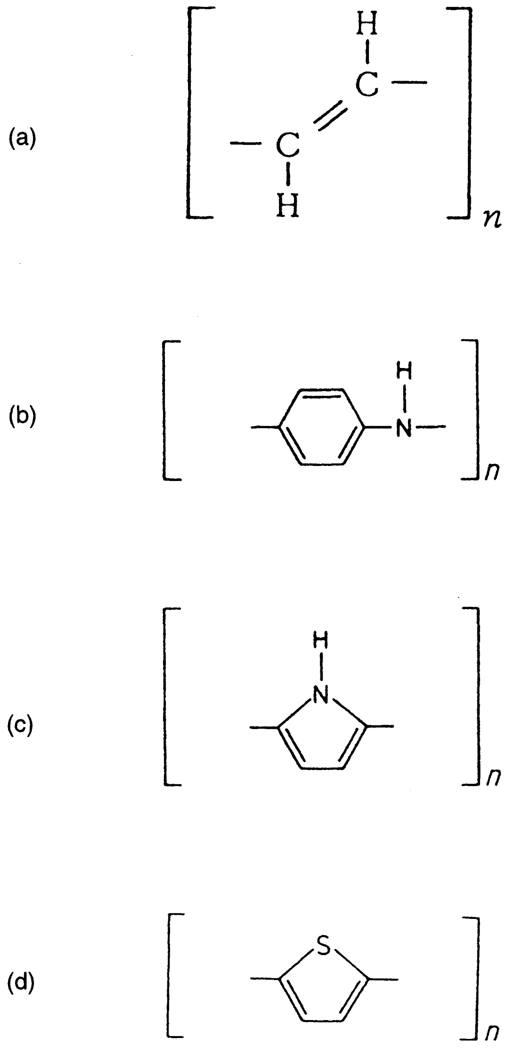

FIG. 4. Schematic structures of (a) *trans*-polyacetylene, (b) polyaniline, (c) polypyrrole, (d) polythiophene.

4.3.2 Characteristics Among numerous conducting polymers, three, polyacetylene, polypyrrole, and polyaniline (Fig. 4), seem to be the most representative.

Polyacetylene is the simplest polyconjugated polymer. It was obtained in a highly conducting state for the first time by Shirakawa *et al.* (1977), and this disclosure marks the beginning of intensive studies on conducting polymers. Its role as a model, quasi 1-D conductor with soliton-type excitations is emphasized in Sec. 4.1.

As obtained, PA is usually a mixture of *trans* and *cis* forms; isomerization to the more stable *trans* form occurs upon doping or heating. When doped with alkali metals, PA shows *n*-type conductivity and is highly environmentally unstable. Doping with oxidative agents, like iodine, bromine, SbF_6, or AsF_5 results in *p*-type conductivity. The value of conductivity of doped PA depends both on the synthesis method and on the doping and may reach 10^5 S/cm (Naarmann and Theophilou, 1987); however, typically it ranges from 10^2 to 10^3 S/cm. The differences in conductivity arise from differences in morphology, influencing the hopping conductivity between the polyacetylene chains, and from differences in conjugation length, influencing the charge transport along the PA chain.

Only the so-called "high-σ" PA, obtained by the Naarmann method, stretched, and highly doped, shows metal-like temperature dependence of conductivity at high temperatures, and it remains high also in the zero-temperature limit. The thermopower of "high-σ" PA is linearly proportional to temperature, as for the ideal metallic diffusion, indicating that it is dominated by quasi 1-D metallic conduction in highly conducting regions. The macroscopic conductivity shows largely nonmetallic behavior because it involves also charge transport through some kind of barriers, e.g., sp^3 defects.

The conductivity of unstretched PA, for example, the Shirakawa-type PA consisting of randomly oriented fibrils of diameter typically about 200 Å, is not only much smaller but also it tends to zero in the zero-temperature limit, showing the absence of continuous metallic paths. The temperature dependence can be fitted by an equation describing tunneling between small metallic islands controlled by the charging energy of the islands:

$$\sigma = \sigma_0 \exp[-(T_0/T)^{0.5}], \tag{24}$$

where σ_0 and T_0 are constants (Kaiser, 1993).

Polypyrrole was the first electrochemically synthesized conducting polymer (Diaz *et al.*, 1979), and it receives considerable interest due to the combination of relatively high conductivity with environmental stability, good mechanical properties, and simplicity of preparation. The PPy films are obtained from the monomer directly in the conducting (oxidized) state because the oxidation potential of PPY polymer is lower than that of the monomer; doping occurs simultaneously with the electropolymerization because the counterions incorporate into the polymer in order to maintain the electrical neutrality of the film. The electrolyte consists of a solvent (usually acetonitrile or water), pyrrole, and various ions, so that PPY can be easily doped with a variety of counterions. The doping (oxidation) is reversible, and PPy can be electrochemically transformed into neutral, nonconducting form.

The PPy films are amorphous and flexible, but with very low ultimate elongation. Their conductivities range in most cases from 1 to 100 S/cm, and the temperature dependence of dc conductivity (σ_{dc}) is described usually by the Mott law developed for variable-range hopping (VRH) (Mott and Davis, 1979) in disordered materials:

$$\sigma_{dc} = \sigma_0 \exp[-(T_0/T)^{0.25}], \tag{25}$$

where σ_0 and T_0 are constants (this equation holds essentially only in the low-temperature regime). However, the density of states at the Fermi level calculated from the fit to the VRH model is unacceptably higher than that obtained from independent ESR studies; also the metal-like behavior of the thermopower is in contradiction to the VRH model. The present understanding of the charge-carrier transport in PPy involves the concept of percolation on a fractal network. The measurements of spin dynamics by NMR at various temperatures indicate an anomalous diffusion in fractal structures of effective dimension $\delta = 1$ at low temperatures; the δ increases with increasing temperature. It is in agreement with the frequency dependence of the ac conductivity (σ_{ac}) following the equation

$$\sigma_{ac} \propto \omega^s \tag{26}$$

with $s = 1 - \delta/3$. At elevated temperatures, the effective dimension increases, indicating an increasing role of the interchain hopping. For three-dimensional hopping at high temperatures, the dc conductivity is given by

$$\sigma_{dc} = \sigma_0 \exp(-2\alpha R) \exp(-E_A/kT), \tag{27}$$

where R is the distance between the neighboring hopping sites, E_A is the energy barrier between the sites, and α is the inverse localization length. The investigations of 3,4-cycloalkyl-substituted PPy fully support this mechanism, showing direct correlation of R with the size of the substituent (Ruhe *et al.*, 1989).

Polyaniline (PANI), known for many years as so-called "aniline black," has attracted considerable research interest only recently, when its unusual electrochemical and optical properties were uncovered (MacDiarmid and Epstein, 1989). Polyaniline is unique among conducting polymers because its electronic properties can be altered both by redox processes involving electron transfer and by acid-base chemistry involving protonation. PANI exists in five forms very different in electrical and magnetic properties. An insulator-conductor transition, like for other conducting polymers, occurs upon charge-transfer doping, and the conducting PANI form, emeraldine salt, can be obtained directly from the leucoemeraldine, which is a large–band-gap insulator. Additionally, protonation of the semiconducting PANI form, the emeraldine base, yields the emeraldine salt by a conversion of the quinoid ring to the benzoid structure with an associated spin unpairing resulting in conductivity changes by 10 orders of magnitude. The conducting PANI form, emeraldine salt, shows good chemical stability. PANI was categorized initially as an intractable polymer; then several methods were reported to dissolve this polymer, for example, in strong acids, but an important advance was the elaboration of the counterion-induced processability of conducting PANI (see Sec. 4.3.3.).

4.3.3 Processable Conducting Polymers

Practically all initially obtained conducting polymers were not processable (insoluble and nonmelting). Insolubility causes serious problems in characterization of the molecular structure of conducting polymers since many investigation techniques can be applied only for polymer solutions. For commercial applications, it is also necessary to have materials processable with conventional plastic technologies.

The insolubility of conjugated polymers in the conducting state was related to side reactions, like cross-linking, caused by free radicals, radical ions, and ions present in the doped polymers. For these reasons, the first approaches to processing conducting polymers were limited to systems obtained via tractable precursor polymers. Nonconducting precursors can be dissolved in order to produce films or fibers; these films or fibers can be further processed, e.g., oriented by stretching, and then converted into a conducting, but no longer processable, state. Highly oriented, anisotropic PA films are obtained in this way, using the so-called Durham route (Edwards and Feast, 1980).

Further progress in processing of conducting polymers was achieved by solubilization of polyconjugated polymers by substituting side groups in them. The most known and investigated are polythiophenes (Fig. 4) substituted with alkyl chains, like poly(3-octylthiophene) (Elsenbaumer *et al.,* 1986). They are soluble in several common organic solvents and are melt processable at temperatures below 200 °C, so that they can be blended with thermoplastics and elastomers using standard techniques and then doped in order to transform the poly(3-alkylthiophene) into the conducting state.

Counterion-induced processability of conducting PANI was achieved by using a suitable functionalized acid to protonate the base form of PANI (Cao *et al.,* 1993). Polyaniline soluble in common organic solvents was synthesized by introducing counterions of large size, which reduce the interchain interactions. Using sulfonic acids, like dodecylbenzene sulfonic acid, soluble polyanilines in the conducting state were obtained. Diesters of phosphoric acid with alkyl substituents, being protonating agents for the emeraldine base, cause also plastification and solubilization of the protonated, conducting emeraldine salt; conducting films can be obtained by casting as well as by hot pressing (Laska *et al.,* 1993). Various other protonic acids acting as surfactants can be used. The conductivity obtained in PANI films cast from solutions is quite high, usually in the range 1–400 S/cm.

4.3.4 Heterogeneous Conducting Polymer Composites

Making polymer composites is one of the most effective and flexible methods in getting speciality materials. Great progress has been made in obtaining fully organic, heterogeneous, electrically conducting systems (Kryszewski, 1991). One can distinguish several main groups in terms of their composition, morphology, and conductivity.

Conducting colloid composites. Stable colloids can be made of different conducting polymers in aqueous and nonaqueous media allowing one to cast films or to press pellets of the dried dispersions. In these systems, conducting grains are surrounded by an insulating steric stabilizer. Colloidal polypyrrole (PPy) can be obtained in the form of spherical particles of diameter in the range 50–500 nm, depending on the preparation conditions, and yields films with conductivity in the range 0.1–2 S/cm. In the case of polyaniline (PANI) colloids, the grain size can be smaller; the cast films of PANI with poly(vinyl acetate) stabilizer show a conductivity of ca. 6 S/cm (Aronson *et al.*, 1991).

Blends of processable conducting polymers with inert polymers. In most of these composites, the conducting phase is made of poly(3-alkylthiophene)s (P3AT), because these polymers are soluble in common organic solvents and also can be melt-blended without decomposition (Laakso *et al.*, 1989). Further progress in such composites is expected as a result of works on the counterion-induced processability of PANI (see Sec. 4.3.3). Very high effectiveness of the conducting phase formation has been achieved by so-called gel decoration. Conducting gels of ultrahigh-molecular-weight PE and poly(3-octylthiophene) (P3OT) were prepared by gelation from solutions in decalin followed by doping with iodine. No percolation threshold was found down to 0.0005 vol fraction of P3OT, and already above 0.02 vol fraction, the conductivity saturates reaching values in the range 10^{-3} to 10^{-2} S/cm. The conducting network is formed of conjugated chains either absorbed or entangled within the PE network (Heeger *et al.*, 1991).

Mixtures of insulating polymer matrices and fillers made of conducting polymers. From the point of view of preparation, these mixtures can be most readily related to polymer–carbon-black composites. The particles of insoluble and/or infusible polymers, e.g., PPy or PANI powders, play the role of conducting filler. Utilizing the high thermal stability of polyaniline tosylate (up to 250 °C for short intervals in nitrogen) it is possible to prepare blends with a variety of conventional thermoplastics (Kulkarni *et al.*, 1991).

Composites obtained by polymerization of conjugated polymers within an insulating matrix. Some monomers, like pyrrole, can penetrate insulating polymeric matrices and, after subsequent polymerization, form conducting networks. This can be easily done with porous materials, such as paper, for which conductivity as high as 10^{-1} S/cm has been reported. Often used is the electrochemical method in which pyrrole penetrates an insulating substrate on a precoated electrode and electropolymerizes *in situ.* Conducting fibers have been obtained by chemical polymerization of pyrrole within the voids of never-dried poly(*p*-phenylene terephthalamide) fibers. At 0.57 vol fraction of PPy, the conductivity is ca. 2 S/cm (Rector *et al.*, 1991). Another route for manufacturing conducting fibers consists in polymerization of acetylene in gel fibers containing 2 wt % of ultrahigh-molecular-weight PE (Chiang *et al.*, 1988). The fibers comprise up to 80 wt % of PA, and after doping with iodine, their conductivity is 1200 S/cm; fibers drawn 2–3 times show after doping a conductivity around 6000 S/cm.

Conducting polymer salts with polymeric counterions. Very specific PPy composites can be produced in one step by electrochemical synthesis using polymeric counterions, like poly(p-styrenesulfonate) (PSS) or other anionically derivatized polymers, including elastomers. In such materials, the insulating phase, i.e., polycounterions, is an intrinsic part of the conducting phase. It has been demonstrated that the conductivity behavior of such composites is controlled by the polycounterion conformation (Glatzhofer *et al.*, 1987). Choosing proper sulfated polymers, one can get thermoplastic and elastic PPy composites (Wernet, 1991).

Reticulate doped polymers consisting of a crystalline network of "organic metals" penetrating a polymer matrix. Reticulate doped polymers (RDP) are obtained by crystallization of conducting charge-transfer (CT) complexes or salts *in situ* in a polymer matrix (Jeszka *et al.*, 1981). When conditions are optimized, a continuous network consisting of microcrystals penetrates the polymer matrix allowing one to obtain conducting materials with extremely low loading, below 0.003 vol fraction. A great number of various systems have been elaborated (Ulański, 1990). The CT complexes and salts used as the dopants include organic semiconductors, organic metals, and organic superconductors. The best systems exhibit conductivities above 1 S/cm at merely 1 wt % of the dopant; the films are transparent, flexible, and environmentally stable. A modified reticulate doping method, the so-called zone-casting technique, allows production of anisotropic composites with conducting crystals forming parallel paths, showing very high anisotropy of conductivity, 10^6 or more; another modification yields films that are only surface conducting.

While the number of polymeric conducting composites is quickly growing, the conductivity mechanism in such systems is not fully understood. Most of the models considered are oversimplified, related to polymer/conducting filler (e.g., carbon black) mixtures. In general, in heterogeneous systems, different mechanisms can contribute to the charge-carrier transport: high-mobility motion in ordered regions (e.g., within crystalline inclusions or along nondefected, oriented polymer chains); hopping among localized states in disordered regions; tunneling between highly conducting islands—all controlled by, generally speaking, the percolation process.

4.3.5 Superconducting Polymers The search for superconducting polymers was motivated by Little's hypothesis (1964) that a quasi one-dimensional polyconjugated chain with side groups in which oscillations of electronic charge can be induced may exhibit high-temperature exciton-type superconductivity. No such polymer was found; moreover, there are theoretical arguments that such a 1-D chain, isolated because of bulky side groups, cannot sustain long-range order.

The first discovered superconducting material without metallic elements was crystalline poly(sulfur nitride), $(SN)_x$. The $(SN)_x$ crystals are obtained by solid-state polymerization of S_2N_2 crystals, obtained by pyrolitic conversion of S_4N_4 in the presence of silver wool as a catalyst. The superconductivity transition temperature depends on crystal perfection and ranges from 0.26 to 0.35 K. Several experimental observations, like the specific heat anomaly and, first of all, the Meissner effect, show that $(SN)_x$ is a bulk type-II superconductor in which electron–electron Umklapp scattering between electron and hole pockets of the Fermi surface occurs rather than the usual electron–phonon–electron interactions (Gill *et al.*, 1982–83).

Recent works on organic superconductors are concentrated on other synthetic metals—molecular CT complexes. The first such system, tetramethyltetraselenafulvalene phosphorus hexafluoride, exhibits superconductivity below 0.9 K under pressure of 12 kbar. Other systems are superconducting also at ambient pressure: *bis*-ethylenedithiolotetrathiafulvalene with iodine, $(BEDT\text{-}TTF)_2I_3$, when transformed into the so-called β^* phase, shows a superconductivity transition at around 7 K (Yagubskii, 1993). Superconducting polymer composites were obtained recently by reticulate doping of polycarbonate with $(BEDT\text{-}TTF)_2I_3$ (Laukhina *et al.*, 1995). The highest superconductivity transition temperature reported so far for organic systems, approaching 33 K, was found for fullerenes, C_{60}, intercalated with alkali metals M, of general formula M_3C_{60} (Fischer and Bernier, 1993) (see FULLERENES).

4.4 Applications

Practical use of the basic property of conducting polymers, the electrical conductivity, was limited for a long time by problems with processing and stability. The first commercialized conducting, fully organic polymeric systems were the transparent, flexible, and stable films obtained by the reticulate doping technique (see Sec. 4.3.4). Conductive and transparent films are produced also by dispersing particles of an intrinsically conducting polymer (e.g., polypyrrole or polyaniline) in an inactive polymer matrix (Roth and Graupner, 1993).

Remarkable progress in the application of conducting polymers as antistatic materials, electromagnetic shielding, sheathing for electric power cables, and conducting fibers and textiles can be expected as a result of the discovery of processable (from solution or from melt) conducting polymers with improved environmental stability. Conducting films or fibers made of composites of conducting polymers with high-performance polymers yields conductivity materials of outstanding mechanical properties. The technique of decoration of a preexisting gel network by a conducting polymer should be a useful technique for producing conducting composites with very low content of conducting polymers (Heeger, 1993).

Numerous attempts have been made in order to exploit practically unusual electrical and optical phenomena induced by the injection of carriers into the π-electron system of conducting polymers. Reversibility of the electrochemical doping of PA has allowed construction of light-weight, high-power-density rechargeable batteries. Doped PA forms a Schottky-type contact to various substrates, making possible construction of photovoltaic devices, and because the band gap of PA matches perfectly the solar energy spectrum, PA-based solar cells have been built.

At present, various polymer batteries have already been commercialized. Typically, they contain a polymer anode (e.g., with PPy or PANI as the electrochemically active material), an electrolyte based on lithium salt solution, and a lithium or lithium alloy counterelectrode. The high charge-storage capacity of conducting polymers has been utilized also in the construction of capacitors with anode and cathode made of polypyrrole showing energy-storage density values two orders of magnitude bigger than conventional capacitors.

Doping/dedoping-induced changes in the absorption spectra of conducting polymers are utilized for electrochromic devices. The undoped polymer, if its energy gap is greater then 3 eV, is transparent in visible light; after doping, it has a high absorption coefficient (10^5 cm^{-1}), and thus thin films are sufficient for effective attenuation of the transmitted light. Displays made of electrochromic polymers exhibit a broad viewing angle and high stability of the image without the need of a permanently applied electric field, in contrast to the displays made of liquid crystals.

The very high third-order nonlinear optical susceptibility of conjugated polymers with subpicosecond response times and high damage thresholds are the reasons for great interest in these polymers as nonlinear optical materials for photonics.

Many other potential applications of conjugated polymers are reported: chemical sensors, making use of changes of conductivity of these polymers due to redox reactions; separation membranes, utilizing drastic changes of permeability of conducting polymers upon oxidation; electronic devices, using the ability of these materials to form rectifying junctions, Schottky barriers, etc.; LEDs built of polymers like poly(p-phenylene vinylene) showing electroluminescence; and drug release devices, where the drug molecules, incorporated as counterion into the conducting polymer, can be released during the reduction (Salaneck *et al.*, 1991).

GLOSSARY

Aromatic Amine Group: A group often present in pendant groups of photoconducting polymers, consisting of a nitrogen atom with adjacent aromatic group(s).

Bipolaron: A correlated pair of ions, being the effect of interaction of two polarons. Bipolarons carry double charge and no spin.

Charge-Transfer Complex: Product of molecular interaction between electron donor and electron acceptor.

Conductor: A material in which carriers are plentiful; either the valence band is only partially filled or the valence and conduction bands overlap.

Dispersive Transport: A hopping transport in which Gaussian statistic is not applicable because of the very broad spread of the hopping times extending into the time range of the experiment.

Electrophotography: A technique of image copying consisting in exposing of a corona-charged layered photoconductor to the image reflected from the original so that dark areas remain charged. Then the latent image is developed by electrostatic attraction of the toner particles to the charged areas and is transferred to a sheet of paper.

Ferroelectric: A material that exhibits spontaneous polarization and electric hysteresis.

Hopping Transport: A conduction mechanism in which charge carriers hop among localized sites of dimensions smaller than the average intersite distance.

Insulator: A material with few carriers of low mobility, characterized by a wide energy gap in terms of the band structure model.

Peierls Transition: A metal-insulator transition in one-dimensional systems induced by a symmetry-reducing distortion (e.g., dimerization in *trans*-polyacetylene).

Photoconductor: A semiconductor whose conductivity is sensitive to light because of the creation of electron-hole pairs by absorption of photons.

Piezoelectric: A material that becomes polarized when stressed.

Polaron: A defect formed by ionization of a polyconjugated polymer, consisting of a radical ion associated with a distortion of the lattice around the charge.

Polyconjugated Polymer: Polymer with an extended π-electron system in the backbone (with unsaturated backbone).

Relaxation Time: The time required for a disturbed system to reach $1/e$ of the final equilibrium configuration.

Relaxation Frequency: The reciprocal of the relaxation time.

Soliton: A defect in *trans*-polyacetylene, being the domain wall across which the bond alternation on the chain is reversed. The neutral soliton is the radical defect; the charged solitons are the anion or cation defects carrying charge but no spin.

Synthetic Metals: A class of conducting materials generally not containing metal atoms and usually consisting of organic molecules or macromolecules. Synthetic metals have low-dimensional structures and exhibit high electrical conductivity, in the best cases metal-like, with respect to the value of the conductivity and temperature dependences of conductivity and of thermopower. Some of these materials also exhibit superconductivity at low temperatures.

Time-of-Flight Technique: A method of determination of charge-carrier mobility consisting in measurement of the transit time of a packet of charges across the sample.

Xerographic Discharge Technique: A method of determination of quantum efficiency of photogeneration and of charge-carrier mobility, consisting in monitoring of the decay of the surface potential of the corona-charged sample after a short pulse of light.

Works Cited

Aronson, M. C., Aldissi, M., Armes, S. P., Thompson, J. D. (1991), *Synth. Met.* **41,** 837–842.

Bartnikas, R. (Ed.) (1987), *Engineering Dielectrics,* Vol. IIB, Special Technical Publication 926, Philadelphia: ASTM.

Bässler, H. (1984), *Philos. Mag.* **50,** 347–362.

Beach, W. F. (1978), *Macromolecules* **11,** 72–76.

Block, H., Walker, S. M. (1982), in: J. V. Dawkins (Ed.), *Developments in Polymer Characterization-3,* London: Applied Science Publishers, p. 75.

Böttger, H., Bryksin, V. V. (1986), *Hopping Conduction in Solids,* Weinheim: VCH Publishers.

Bräunlich, P. (Ed.) (1979), *Thermally Stimulated Relaxation in Solids,* Berlin: Springer Verlag.

Cao, Y., Smith, P., Heeger, A. J. (1993), *Synth. Met.* **57,** 3514–3519.

Chiang, J. C., Smith, P., Heeger, A. J., Wudl, F. (1988), *Polym. Commun.* **29,** 161–165.

Debye, P. (1945), *Polar Molecules,* New York: Dover Publications Inc.

Diaz, A. F., Kanazawa, K. K., Gardini, G. (1979), *J. Chem. Soc. Chem. Commun.* 635.

Deutsch, K., Hoff, E. A. W., Reddish, W. (1954), *J. Polym. Sci.* **13,** 565–582.

Duke, C. B., Salaneck, W. R., Fabish, T. J., Ritsko, J. J., Thomas, H. R., Paton, A. (1978), *Phys. Rev. B* **18,** 5717–5739.

Edwards, J. H., Feast, W. J. (1980), *Polymer* **21,** 595–596.

Ehinger, K., Roth, S. (1985), in: H. Kuzmany, M. Mehring, S. Roth (Eds.), *Electronic Properties of Polymers and Related Compounds,* Springer Series in Solid-State Science Vol. 63, Berlin: Springer, pp. 67–74.

Elsenbaumer, D. L. E., Jen, K. Y., Oboodi, R. (1986), *Synth. Met.* **15,** 169–174.

Fischer, J. E., Bernier, P. (1993), *La Recherche* **250,** 46.

Fisher, P. (1982), in: D. A. Seanor (Ed.), *Electrical Properties of Polymers,* New York: Academic Press, p. 319.

Frommer, J. E., Chance, R. R. (1988), in: J. I. Kroschwitz (Ed.), *Electrical and Electronic Properties of Polymers,* New York: J. Wiley and Sons, p. 301.

Gill, W. D., Clarke, T. C., Street, G. B. (1982–83), *Appl. Phys. Comm.* **2,** 211–282.

Glatzhofer, D. T., Ulański, J., Wegner, G. (1987), *Polymer* **28,** 449–453.

Havriliak, S., Negami, S. (1967), *Polymer* **8,** 161–172.

Heeger, A., Smith, P., Fizazi, A., Moulton, J., Pakbaz, K., Rughooputh, S. (1991), *Synth. Met.* **41,** 1027–1032.

Heeger, A. (1993), *Synth. Met.* **57,** 3471–3482.

Jeszka, J. K., Ulański, J., Kryszewski, M. (1981), *Nature* **298,** 390–391.

Kaiser, A. B. (1993), *Mater. Sci. Forum* **122,** 13–24.

Kao, K. C., Hwang, W. (1981), *Electrical Transport in Solids,* Oxford: Pergamon Press.

Kirkwood, J. G. (1939), *J. Chem. Phys.* **7,** 911–919.

Kivelson, S. (1981), *Phys. Rev. Lett.* **46,** 1344–1348.

Kremer, F., Vallerian, S. U., Zentel, R. (1990), *Adv. Mater.* **2,** 145–146.

Kryszewski, M. (1991), *Synth. Met.* **45,** 289–296.

Ku, C. C., Liepins, R. (1987), *Electrical Properties of Polymers,* Munich: Hanser Publishers.

Kulkarni, G. V., Mathew, W. R., Wessling, B., Merkle, H., Blaettner, S. (1991), *Synth. Met.* **41,** 1009–1012.

Laakso, J., Österholm, J.-E., Nyholm, P., Stubb, H., Punkka, E. (1989), *Synth. Met.* **37,** 145–150.

Laska, J., Trznadel, M., Proń, A. (1993), *Mater. Sci. Forum* **122,** 177–184.

Laukhina, E. E., Merzhanov, V. A., Pesotskii, S. I., Khomenko, A. G., Yagubskii, E. B., Ulański, J., Kryszewski, M., Jeszka, J. K. (1995), *Synth. Met.* **70,** 797–800.

Little, W. A. (1964), *Phys. Rev.* **134,** A1416–A1424.

Lowell, L., Rose-Innes, A. C. (1980), *Adv. Phys.* **29,** 947–1023.

MacDiarmid, A. G., Epstein, A. J. (1989), *Faraday Discuss. Chem. Soc.* **88,** 317–332.

McCrum, N. G., Read, B. E., Williams, G. (1967), *Anelastic and Dielectric Effects in Polymeric Solids,* New York: J. Wiley and Sons.

Miller, R. D., Michl, J. (1989), *Chem. Rev.* **89,** 1359–1410.

Mort, J., Pfister, G. (1982), in: J. Mort and G. Pfister (Eds.), *Electronic Properties of Polymers,* New York: J. Wiley and Sons, p. 215.

Mott, N. F., Davis, E. A. (1979), *Electronic Processes in Non-Crystalline Materials,* Oxford: Clarendon Press.

Naarmann, H., Theophilou, N. (1987), *Synth. Met.* **22,** 1–8.

Okamoto, K., Itaya, A. (1984), *Bull. Chem. Soc. Jpn.* **57,** 1626–1630.

Onsager, L. (1936), *J. Am. Chem. Soc.* **58,** 1486–1492.

Pautmeier, L., Richtert, R., and Bässler, H. (1990), *Synth. Met.* **37,** 271–281.

Pearson, J. M., Stolka, M. (1981), *Poly(N-vinylcarbazole),* New York: Gordon and Breach.

Pukacki, W. (1993), *Mater. Sci. Forum* **122,** 255–268.

Rector, L. P., De Groot, D. C., Schindler, J. L., Marks, T. J., Carr, S. H. (1991), *Synth. Met.* **41,** 935–939.

Ritsko, J. J. (1982), in: J. Mort, G. Pfister (Eds.), *Electronic Properties of Polymers,* New York: J. Wiley and Sons, p. 13.

Roth, S., Graupner, W. (1993), *Synth. Met.* **57,** 3623–3631.

Ruhe, J., Ezquerra, T. A., Mohammadi, M., Enkelmann, V., Kremer, F., Wegner, G. (1989), *Synth. Met.* **28,** C217–C223.

Salaneck, W. R., Clark, D. T., Sammelson, E. J. (Eds.) (1991), *Science and Application of Conducting Polymers,* Bristol: Adam Hilger.

Saone, D. S., Martynenko, Z. (1989), *Polymers in Microelectronics, Fundamentals and Applications,* Amsterdam: Elsevier.

Schimmel, Th., Schwoerer, M., Naarmann, H. (1990), *Synth. Met.* **37,** 1–6.

Scott, J. C., Pautmeier, L. Th., Moerner, W. E. (1993), *Synth. Met.* **54,** 9–19.

Sessler, G. M. (1982), in: D. A. Seanor (Ed.), *Electrical Properties of Polymers,* New York: Academic Press, p. 224.

Sher, H., Montroll, E. W. (1975), *Phys. Rev. B* **12,** 2455–2477.

Shirakawa, H., Louis, E. J., Mac Diarmid, A. G., Chiang, C. K., Heeger, A. J. (1977), *J. Chem. Soc. Chem. Commun.* 578.

Stockmayer, W. H. (1967), *Pure Appl. Chem.* **15,** 539–545.

Stolka, M., Abkowitz, M. A., Knier, F. E., Weagley, R. J., McGrane, K. M. (1990), *Synth. Met.* **37,** 295–303.

Stolka, M., Abkowitz, M. A. (1993), *Synth. Met.* **54,** 417–425.

Stolka, M. (1988), in: J. I. Kroschwitz (Ed.), *Electrical and Electronic Properties of Polymers,* New York: J. Wiley and Sons, p. 301.

Stolka, M., Pai, D. M. (1978), *Adv. Polymer Sci.* **29,** 1–45.

Szwarc, M. (1976), *Polym. Eng. Sci.* **16,** 473–479.

Ulański, J. (1990), *Synth. Met.* **39,** 13–42.

Ulański, J., Jeszka, J. K., Kryszewski, M. (1981), *Polym.-Plast. Technol. Eng.* **17,** 139–191.

van Beek, L. K. H. (1967), in: J. B. Birks (Ed.), *Progress in Dielectrics,* Vol. 7, London: Heywood Books, p. 69.

van Turnhout, J. (1987), in: G. M. Sessler (Ed.), *Electrets,* Berlin: Springer Verlag, p. 81.

Volkenstein, M. V. (1963), *Configurational Statistics of Polymers,* New York: Interscience.

Wada, Y. (1982), in: J. Mort, G. Pfister (Eds.), *Electronic Properties of Polymers,* New York: J. Wiley and Sons, p. 109.

Wegner, G., Ruhe, J. (1989), *Faraday Discuss. Chem. Soc.* **88,** 333–349.

Wegner, G. (1981), *Angew. Chem.* **20,** 361–381.

Wernet, W. (1991), *Synth. Met.* **41,** 843–848.

Wurstlin, F. (1951), *Kolloid Z.,* 84–102.

Yagubskii, E. B. (1993), *Mol. Cryst. Liq. Cryst.* **230,** 139–156.

Lu, Yu (1988), *Solitons and Polarons in Conducting Polymers,* Singapore: World Scientific.

Further Reading

Bailey, R. T., North, A. M., Pethrick, R. A. (1981), *Molecular Motions in Polymers,* Oxford: Clarendon Press, Chap. 5.

Bartnikas, R., Eichhorn, M. (Eds.) (1983), *Engineering Dielectrics,* Vol. IIA, Philadelphia: ASTM Special Technical Publication 783.

Block, H. (1979), *Adv. Polym. Sci.* **33,** 93–167.

Böttcher, C. J. F. (1973), *Theory of Dielectric Polarization,* Vol. I, *Dielectrics in Static Fields,* Amsterdam: Elsevier.

Böttcher, C. J. F., Bordewijk, P. (1978), *Theory of Electric Polarization,* Vol. II, *Dielectrics in Time Dependent Fields,* Amsterdam: Elsevier.

Bräunlich, P. (Ed.) (1979), *Thermally Stimulated Relaxation in Solids,* Berlin: Springer Verlag.

Chien, J. C. W. (1984), *Polyacetylene,* Orlando: Academic Press.

Hanack, M., Roth, S., Shier, H. (Eds.) (1991), Proceedings of the International Conference on Science and Technology of Synthetic Metals, Tubingen, West Germany, 1990, *Synth. Met.* **41–43.**

Hedvig, P. (1977), *Dielectric Spectroscopy of Polymers,* Bristol: Hilger.

Jonscher, A. K. (1983), *Dielectric Relaxation in Solids,* London: Chelsea Dielectric Press.

Kao, K. C., Hwang, W. (1981), *Electrical Transport in Solids,* Oxford: Pergamon Press.

Kremer, F., Boese, D., Meier, G., Fischer, E. W. (1983), *Progr. Coll. Polym. Sci.* **80,** 129–139.

Kroschwitz, J. I. (Ed.) (1988), *Electrical and Electronic Properties of Polymers: A State-of-the-Art Compendium,* New York: J. Wiley and Sons.

Kryszewski, M., Ulański, J., Bernier, P. (Eds.) (1990), Proceedings of the 21st Europhysics Conference on Macromolecular Physics: Electric and Optical Active Polymers, Łódź, Poland, 1989, *Synth. Met.* **37.**

Kryszewski, M. (1980), *Semiconducting Polymers,* Warszawa: PWN–Polish Scientific Publishers.

Ku, C. C., Liepins, R. (1987), *Electrical Properties of Polymers,* Munich: Hanser Publishers.

Mair, H. J., Roth, S. (Eds.) (1986), *Electrisch Leitende Kunststoffe,* Berlin: Hanser Publishers.

McCrum, N. G., Read, B. E., Williams, G. (1967), *Anelastic and Dielectric Effects in Polymer Solids,* London: J. Wiley and Sons; reprint (1991), Dover Publications Inc.

Mort, J., Pfister, G. (Eds.) (1982), *Electronic Properties of Polymers,* New York: J. Wiley and Sons.

O'Dwyer, J. J. (1964), *The Theory of Electrical Conduction and Breakdown in Solid Dielectrics,* Oxford: Clarendon Press.

Park, Y. W., Lee, H. (Eds.) (1995), Proceedings of the International Conference on Science and Technology of Synthetic Metals, Seoul, 1994, *Synth. Met.* **69–71**.

Pearson, J. M., Stolka, M. (1981), *Poly(N-vinylcarbazole),* New York: Gordon and Breach.

Prasad, S. V., Ulrich, D. R. (Eds.) (1988), *Nonlinear Optical Polymers,* New York: Plenum Press.

Seanor, D. A. (Ed.) (1982), *Electrical Properties of Polymers,* New York: Academic Press.

Sessler, G. M. (Ed.) (1987), *Electrets,* Topics in Applied Physics, Vol. 33, Berlin: Springer Verlag.

Simon, J., Andre, J.-J. (1985), *Molecular Semiconductors,* Berlin: Springer Verlag.

Skotheim, T. A. (Ed.) (1986), *Handbook of Conducting Polymers,* Vols. 1–2, New York: Marcel Dekker, Inc.

Stafström, S., Salaneck, W. R., Inganas, O., Hjertberg, T. (Eds.), (1993), Proceedings of the International Conference on Science and Technology of Synthetic Metals, Goteborg, 1992, *Synth. Met.* **55–57**.

Ulański, J., Jeszka, J. K., Kryszewski, M. (1981), *Polym.-Plast. Technol. Eng.* **17,** 139–191.

van Roggen, A. (1990), *IEEE Trans. Electr. Insul.* **25,** 95–106.

van Turnhout, J. (1975), *Thermally Stimulated Discharge of Polymer Electrets,* Amsterdam: Elsevier.

Williams, G. (1979), *Adv. Polym. Sci.* **33,** 59–92.

Wintle, H. J. (1990), *IEEE Trans. Electr. Insul.* **25,** 27–44.

POLYMERS, MECHANICAL PROPERTIES OF

DAVID T. GRUBB, *Department of Materials Science and Engineering, Cornell University, Ithaca, New York, U.S.A.*

	Introduction	531
1.	**Mechanical Properties and Testing**	532
2.	**Elastic Properties**	534
2.1	Rubber Elasticity	534
2.1.1	Statistical Principles	534
2.1.2	Phenomenological Models of Large-Strain Behavior	535
2.2	Viscoelasticity	536
2.2.1	Models of Linear Viscoelasticity	537
2.2.2	The Glass Transition and Time–Temperature Superposition	537
2.2.3	Nonlinear Viscoelasticity	538
2.3	Anisotropic Elasticity	539
3.	**Yield and Fracture**	541
3.1	Localized Shear Yield in Tension	542
3.2	Localized Shear Yield in Compression	544
3.3	Crazing and Brittle Fracture	544
3.4	Ductile–Brittle transition	545
4.	**Fatigue and Wear**	546
	Glossary	547
	Works Cited	548
	Further Reading	548

INTRODUCTION

The basic mechanical properties of a solid are its stiffness, strength, and toughness. Stiffness, the ratio of load to deflection, depends on the material and on the shape of the body. The material property itself is a modulus, a ratio of stress to strain (or a compliance, a ratio of strain to stress). The strength of a material is the stress at which the material fails by fracture or by irreversible deformation. Toughness is a measure of the work required to fracture a material. Polymeric materials come in endless chemical variety, and these materials display an enormous range of mechanical properties. Materials may be manufactured in forms that have high stiffness and strength, as fibers to reinforce tires, for example. Equally, low stiffness and low strength may be desirable for gaskets and packing materials. The properties of some typical polymers are compared with the properties of other materials in Table 1.

A complication is that polymer mechanical properties are strongly dependent on time and temperature. A polymer may appear as a stiff and rigid solid when subject to impact loading, but either as a soft elastic rubber or as a viscous liquid when the loading persists over a long time. All solids tend to flow at high temperatures, and for metals this is a serious problem for designers of jet engines. Very few users of polymers can afford to ignore the change of properties with temperature. Many natural and synthetic materials are also highly sensitive to water content and other environmental effects. Further, naming the chemical species is not enough to define the material; the molecular weight and processing conditions can have strong effects on polymer mechanical properties. Failure is strongly affected by the conditions of loading. This is a general phenomenon, as, for example, steel becomes more brittle when cold, but in polymers the range of behavior can be extreme. A single sample may be a brittle solid with very little extensibility at −10 °C, a tough solid that extends 30%–100% before breaking at room temperature, and a fluid with the consistency of soft chewing gum at 150 °C.

Polymers consist of long molecular chains. There are strong covalent bonds along the chain and much weaker bonds (hydrogen or van der Waals bonds) between chains. If the chains are frequently linked into a network, the result is a "thermosetting" polymer that does not melt. If there are only a few crosslinks, these stabilize the

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Table 1. Mechanical properties of materials.

Material	Bulk modulus (GPa)	Young's modulus (GPa)	Tensile strength (MPa)	Toughness (kJ/m^2)	Density (Mg/m^3)	Melting or softening temp. (K)
Nylons	6.5	4	100	3	1.1	360
PMMA	5	3.4	70	0.3	1.2	370
Natural rubber	3	0.05	30	—	0.9	(210)
Mild steel	200	200	420	100	7.8	1810
Window glass	45	70	3600[a]	0.01	2.5	800

[a] Compressive strength.

form of the material, even when the chains are in a mobile liquidlike state, giving an elastomer. Elastomers are soft and very extensible because the free chains between crosslinks can be extended by covalent-bond–angle rotation. Entanglements between long linear chains can act as temporary crosslinks, giving rubbery behavior at short time scales.

In a polymer glass, the chains are locked into one configuration. Above the glass-transition temperature T_g, the chains are mobile. Many properties change at T_g, including the coefficient of thermal expansion and the heat capacity, but for our purposes the changes in mechanical properties are most important. The modulus falls drastically as the temperature is increased through T_g, and the extension to failure will also change. This makes T_g a most important parameter for a polymer. Elastomers have a T_g below ambient (~ -60 °C for natural rubber) and polymers normally considered glassy have a T_g above ambient (~ 100 °C for polystyrene, PS). The chains move with a finite speed, and so a slow experiment sees them in motion, while a rapid one sees them fixed. T_g is thus frequency dependent. A true state of no motion is attained at temperatures below T_g^∞, which is some 50 K below the tabulated and practically useful value of T_g.

A polymer may be an amorphous glass, rubber, or liquid, or it may contain a degree of order. The order may be within crystalline or liquid-crystalline (LC) elements, but except in a few special cases the crystals or LC domains are very small and imperfect. In the crystalline case, the materials are described as "semicrystalline" and are modeled as containing two phases: ideal crystallites and ideal amorphous material. The fraction of crystalline phase in this model is the "degree of crystallinity," which can vary from 0.2 to 0.9. The semicrystalline material is a composite, and so its mechanical properties depend on the properties of both phases and on their arrangement. Most semicrystalline polymers are used at temperatures between T_g and the crystalline melting point T_m.

When extending the chain requires covalent-bond bending or stretching, the chain will be stiff in tension. As the bonds between chains are weak, any chain orientation will then give anisotropic mechanical properties. The ratio of perpendicular elastic moduli in tension may be as much as 100:1 in an oriented crystalline fiber, less in oriented glassy material. Failure modes that depend on chains slipping past each other and brittle failure that involves breaking bonds will also be strongly affected by chain orientation. Because chains are long, large-scale structures may be involved in deformation, and so response and relaxation times can be very long. Polymers often show a delayed elastic response, where the recovery on unloading takes a long time.

These general concepts explain the broad features of the mechanical properties of solid polymers. In detail, they are often insufficient to explain the actual properties, because of the complicated internal structure of the materials. In addition, many commercial polymers are blends or copolymers, and may contain rigid mineral filler particles and additives such as plasticizers and antioxidants that significantly affect their properties.

1. MECHANICAL PROPERTIES AND TESTING

Most mechanical testing of solid polymers uses standard tests that set out particular sample dimensions, geometry and rate of deformation, temperature, and humidity. These are excellent for quality control and very useful for comparison between materials. They are described in full detail in ASTM and DIN handbooks (ASTM, 1994; DIN, 1994). These tests can be used to measure the stiffness of a molded bar, its impact strength, or the tear strength of an elastomer. Some have been set up to test for a specific practical problem, such as stress cracking of polyethylene in detergent. The problem with this type of test is that extrapolation to different conditions of loading is difficult and uncertain. There is a continuing effort to design and validate tests for failure that have a good scientific basis and are also rapid and useful.

A picture of the polymer properties can be built up using more fundamental tests, which are

1. creep testing—measurement of displacement as a function of time at a constant load to give a time-dependent creep compliance $J(t)$;
2. stress-relaxation testing—measurement of load as a function of time at a constant displacement to give a time-dependent relaxation modulus $E(t)$;
3. dynamic mechanical testing—measurement of the amplitude and phase of displacement under a load that is a sinusoidal function of time. This gives a complex modulus $G^*(\omega) = G'(\omega) + iG''(\omega)$ or an equivalent complex viscosity $\eta^*(\nu) = G^*(\omega)/i\omega$. Thermal transitions in the material are more easily seen when the ratio G''/G' (often called the loss tangent or $\tan\delta$) is plotted as a function of temperature.

By linear-viscoelastic theory (described in Sec. 2.2.1, below) the results of these tests can used to predict behavior under more general conditions of loading (Ferry, 1970; Arridge, 1975). Unfortunately, most polymers are not linear in their mechanical behavior; that is, their stiffness depends to some degree on the value of the applied load. The conditions when this becomes important can be determined by plotting the compliance after a fixed time of creep for a range of loads. This isochronous stress–strain curve will be a horizontal line in the linear region. Good prediction of mechanical properties under general loading conditions can only be made for small deformations unless a much larger database is built up.

Commercial dynamic mechanical testing instruments are the fastest way to find the temperature dependence of mechanical properties. They apply a forced vibration to a small sample in shear, bending, or tension. The sample is held in a temperature-controlled enclosure, and the mechanical response is measured over a wide range of temperatures and at a range of frequencies. An example result is shown in Fig. 1. G', the in-phase or storage shear modulus, falls by a factor of several hundred near the glass-transition temperature, T_g = 100 °C. At the transition, there is a peak in the loss tangent $\tan\delta$. In this case, there is also a clear peak in the loss modulus G'' and a secondary tran-

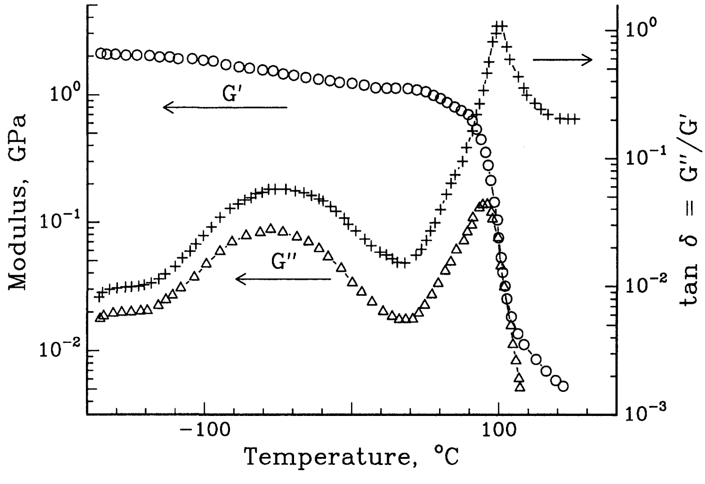

FIG. 1. Dynamic mechanical properties of unplasticized poly(vinyl chloride) (PVC). At $T_g \sim$ 100 °C the storage shear modulus G' falls dramatically, where there are peaks in both $\tan\delta$ and the loss modulus G'. A secondary transition at about −60 °C is clear in $\tan\delta$ and G'. Redrawn from Peterson and Ranby (1970).

sition below T_g at about −60 °C. These tests can be performed in tension or in shear. The conventional symbols for tensile modulus and compliance are E and D. For shear modulus and compliance they are G and J.

Elastic properties at high strain rates can be determined by ultrasonic modulus measurements; these are simple in principle and have many advantages. For unoriented materials, and for oriented materials when the acoustic waves propagate down symmetry axes of the sample, the relevant modulus component is simply the density times the square of the wave velocity. In an isotropic material longitudinal waves give E, transverse waves give G, and so any of the other elastic properties can be calculated. For example, the values for nylon 6 (and nylon 66) obtained in this way are $E = 4.0$ GPa, $G = 1.43$ GPa, bulk modulus $K = 6.5$ GPa, and Poisson's ratio $\nu = 0.40$. A pulsed laser beam focused onto the specimen has been used to find the longitudinal modulus of a fiber. The pulse creates a shock wave that moves at the speed of sound along the fiber, so that timing the shock wave gives the fiber modulus. Testing for failure at high strain rates requires special testing machines that use hydraulics or compressed air to deform the sample rapidly. They can be used for strain rates of 10–100 s^{-1}. Higher rates require ballistic testing.

2. ELASTIC PROPERTIES

2.1 Rubber Elasticity

Rubber elastic materials can be deformed reversibly to very large strains, three to ten times their original length. Their shear modulus is typically 1 MPa or less while their bulk modulus is ~1 GPa. An elastomer contains crosslinked amorphous polymer chains. The number of units between crosslinks must be reasonably large so that the chains can take up a range of configurations. The temperature must be above T_g so that the chains are not locked into one configuration.

The physical principle of rubber elasticity is that the material force restoring an elastomer to its original shape is entropy driven. An ideal rubber has no change in internal energy on deformation, only an entropy change. Simple thermodynamics shows that in this case the restoring force at constant length will be proportional to the absolute temperature. Measurements of force as a function of temperature at constant length, or measurements of length as a function of temperature at constant force (corrected for thermal expansion), are used to find out how well a given material can be modeled as an ideal rubber (Treloar, 1975; Flory, 1969). A stretched rubber shrinks on heating, and, as may be tested with an office elastic band, rapid release after stretching causes the material to cool.

Even with no crosslinks, many long-chain polymers show rubbery behavior in some range of temperature and time scale. This is because physical entanglements between the chains restrain them like crosslinks for short times. The range of the "rubbery plateau" where the modulus is constant increases with molecular weight, but the actual "plateau modulus" is independent of molecular weight. Practical elastomers are generally reinforced by the addition of rigid particles and/or by the polymer crystallizing on extension. The increase in stiffness and strength when carbon black particles are added to rubber to make tires depends strongly on the size and the surface condition of the particles. When the filler affects a large fraction of the material, it may become much less extensible. In semicrystalline polymers above T_g, the crystals act as reinforcing particles and also as cross-linking sites physically fixing the amorphous chains. However, full rubbery behavior may not be developed because the amorphous chains are short and highly constrained.

2.1.1 Statistical Principles In brief, the elastomer at equilibrium contains molecular chains in the form of random coils. When stretched these have fewer possible configurations, and it is this entropy change that drives the elasticity. Even though real molecules have restricted angles of bond rotation, a long chain can be accurately described as a set of n freely jointed links of length l, where n is less than the number of bonds in the chain. As long as the chain is not near to full extension, its three-dimensional random walk starting at the origin has a Gaussian probability distribution for the position of the end of the chain:

$$P(x,y,z)dx\,dy\,dz = \left(\frac{b}{\sqrt{\pi}}\right)^3 e^{-b^2(x^2+y^2+z^2)}dx\,dy\,dz, \tag{1}$$

where $b^2 = 3/2nl^2$. The mean square end-to-end distance for the chain, $\overline{r^2}$, is equal to nl^2. The entropy associated with the possible chain conformations is $S = C - kb^2(x^2 + y^2 + z^2)$ where k is Boltzmann's constant. Consider deforming this chain so that the end moves to x', where $x' = \lambda_1 x$. The entropy change is

$$\Delta S_x = -kb^2x^2(\lambda_1^2 - 1). \tag{2}$$

For a general deformation to $\lambda_1 x$, $\lambda_2 y$, $\lambda_3 z$, we may integrate over the distribution $P(x,y,z)$ or equivalently replace x^2, y^2, and z^2 with $\overline{r^2}/3 = nl^2/3 = 1/2b^2$. For an ideal chain with no change of internal energy on bond rotation, $\Delta A = -T\Delta S$, and so the free-energy change on deformation of an assembly of N such chains deforming at constant volume is

$$\Delta A = (NkT/2)(\lambda_1^2 + \lambda_2^2 + \lambda_3^2 - 3). \tag{3}$$

The restoring force $F = (\partial A/\partial l)_T = (\partial A/\partial \lambda)_T$ for an initial unit cube. Some simple forms of deformation give useful results. For shear, $\lambda_1 = \lambda$, $\lambda_2 = 1/\lambda$, $\lambda_3 = 1$. In pure shear, the shear strain $e = \lambda - 1/\lambda$ and so

$$\Delta A = (NkT/2)(\lambda^2 + 1/\lambda^2 - 2) = (NkT/2)e^2,$$
$$F = \sigma = NkTe. \tag{4}$$

Thus the shear modulus $G = NkT$. If the deformation is uniaxial at constant volume, $\lambda_1 = \lambda$, $\lambda_2 = \lambda_3 = \sqrt{1/\lambda}$, giving

$$\Delta A = (NkT/2)(\lambda^2 + 2/\lambda - 3),$$
$$F = NkT(\lambda - 1/\lambda^2). \tag{5}$$

This is also the nominal stress—the force divided by the original cross-sectional area A_0, here unity. The actual cross-sectional area A is $1/\lambda$, and so the true stress is $\sigma_t = G(\lambda^2 - 1/\lambda)$. For biaxial deformation, where $\lambda_1 = \lambda_2 = \lambda$, $\lambda_3 = 1/\lambda^2$,

$$\Delta A = (NkT/2)(2\lambda^2 + 1/\lambda^4 - 3),$$
$$F = 2NkT(\lambda - 1/\lambda^5). \tag{6}$$

The above results were derived for an assembly of ideal statistical chains. There are two very important facts about these relations of load and displacement. One is that for most real elastomers they are a good first approximation only. Even for unfilled material, they are not sufficient to interpret a wide range of mechanical data. There are several reasons for this: The real materials are only 80%–90% ideal; even for ideal chains, we have not considered the possibility of fluctuations in the positions of cross-links, nor the effect of a length distribution of chains. Short chains will come near to full extension and non-Gaussian behavior even at low network extensions. Theories exist to treat these complications (Treloar, 1975; Flory, 1969), but most real mechanical data, particularly at high strains, are fitted using phenomenological methods described in the next section. The other important fact is that the relations in Eqs. (4)–(6) do not depend on an entropic basis or a thermodynamic derivation. Exactly the same relations can be derived by pure mechanics, when strains are allowed to be large and a simple constitutive relation between stress and strain is assumed.

2.1.2 Phenomenological Models of Large-Strain Behavior

A general deformation at finite strains is complicated by the fact that finite shear strains change directions in the material. We can choose a set of axes such that the shear-strain components go to zero. In this case, the axes are not rotated by the deformation, and the resulting pure strain can be described by the deformation ratios along x, y, and z: λ_1, λ_2, and λ_3. The finite extensional strains are $\epsilon_{xx} = (\lambda_1^2 - 1)/2$ and so on, while the corresponding stresses are σ_{xx} = (nominal stress)$/\lambda_2\lambda_3$. The work done in deforming the body, which in an elastic material is the stored energy, must be some function of the deformation ratios. It is called the strain-energy function, W. For an isotropic material, the strain-energy function has to be independent of the choice of coordinate axes. There are three independent functions of the deformation ratios that have this property. They are called the strain invariants, I_1, I_2, and I_3, and are

$$I_1 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2,$$
$$I_2 = \left(\frac{1}{\lambda_1^2} + \frac{1}{\lambda_2^2} + \frac{1}{\lambda_3^2}\right)I_3,$$
$$I_3 = \lambda_1^2\lambda_2^2\lambda_3^2. \tag{7}$$

For an incompressible material $I_3 = 1$, so setting $W = 0$ at the original state of the material when $\lambda_1 = \lambda_2 = \lambda_3 = 1$, a general form for the unknown function W is

$$W = \sum_{i=0}^{N} \sum_{j=0}^{N} C_{i,j}(I_1 - 3)^i (I_2 - 3)^j. \qquad (8)$$

If only one term is kept in this power-law expansion, the term with $i = 1$, $j = 0$, the result is identical to the simple statistical theory of the previous section. Assuming a particular simple constitutive relation for the material gives the same result, and a material with this behavior may be called neo-Hookean. More common for real materials is to keep two terms; either the expression for W or the expression for stress in uniaxial tension derived from it may be called the Mooney–Rivlin equation:

$$W = C_1(I_1 - 3) + C_2(I_2 - 3),$$
$$\sigma = \left(\lambda - \frac{1}{\lambda^2}\right)\left(2C_1 + \frac{2C_2}{\lambda}\right). \qquad (9)$$

Practical calculations and finite-element models are likely to use the above expression for the strain-energy function. It is only an approximation and becomes very inaccurate for large tensile elongations. In physical terms, this is because the elongated material is changing its structure; the chains near full extension are no longer Gaussian and Eqs. (1) and (2) no longer apply. Figure 2 shows data for uniaxial elongation, and the fit for $\lambda < 4$ using Eq. (5). The fit with Eq. (9) is slightly better at low strains, but the difference is hardly visible on the scale of Fig. 2. A purely empirical equation for the tensile stress that only needs one parameter is the MRS equation:

$$\sigma = E\frac{\epsilon}{\lambda^2}\exp\left[A\left(\lambda - \frac{1}{\lambda}\right)\right]. \qquad (10)$$

When $A \sim 0.5$ this fits the upturn in stress in Fig. 2. The strain-energy function can also be expressed as the sum of similar functions of each of the deformation ratios, $W = \Sigma_1^3 w(\lambda_i)$. This may be useful when the data are not easily expressed in terms of the invariants. The following form of $w(\lambda)$ has been found to fit the data with three terms:

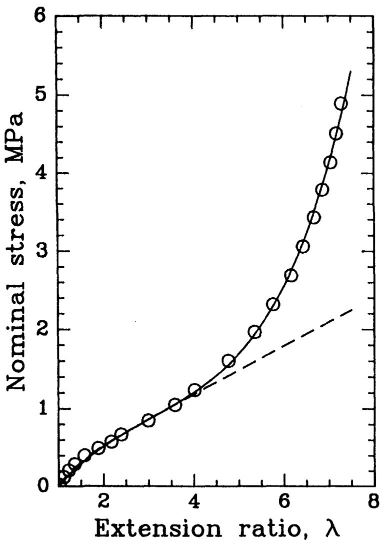

FIG. 2. Data for the uniaxial extension of a rubber, fitted by Eq. (5) for extension ratios below 4 (dashed line) and fitted by Eq. (11) for the full range of extensions (solid line). The data were taken from Treloar (1969).

$$w(\lambda_i) = \sum_n \frac{\mu_n}{\alpha_n}(\lambda_i - 1)^{\alpha_n},$$
$$\sigma_{ii} = \sum_n \mu_n \lambda_i^{\alpha_n} + P. \qquad (11)$$

Figure 2 shows how well this fits the data. A good fit is not surprising with six adjustable parameters. The point is that the *same* parameters fit the shear and biaxial extension data for the same material.

2.2 Viscoelasticity

Real polymers, whether in an apparently solid or liquid form, generally have both viscous and elastic parts to their mechanical behavior and are therefore viscoelastic. The moduli and the viscosity both decrease as the temperature is raised. The elastic properties of solid polymers are nonlinear at deformations greater than 1% or 2%, and accurate measurements at lower strains show that some have no linear region at all. The viscosity of polymers is also nonlinear, commonly decreasing with shear rate. The complex flow properties are very important for processing the polymers.

Much analysis of polymer viscoelasticity has dealt with the linear case, where material properties are independent of the magnitude of the stress, and the effects of stress or strain steps at arbitrary times are additive. (This is called the "Boltzmann superposition principle.") Linear theory is the starting point for modeling, and empirical methods can be used to predict the nonlinear behavior from linear viscoelasticity. It is important to remember that linear viscoelasticity is only a first step in describing or understanding polymer properties.

2.2.1 Models of Linear Viscoelasticity Linear viscoelasticity is a combination of linear Hookean elasticity and Newtonian linear viscosity. The simplest models of linear viscoelastic behavior have a combination of a few Hookean linear elastic springs and Newtonian linear viscous dashpots. A series combination of a spring and a dashpot is called a Maxwell element, and gives a simple model of stress relaxation. A spring and a dashpot in parallel is a Kelvin or Voigt element, and this models creep. The behavior of these networks is analogous to that of electrical circuits containing capacitors and resistors.* To get any combination of creep and stress-relaxation behavior, a combination of three or four basic units is required, such as the three-element "standard linear solid" or a series combination of Maxwell and Voigt elements.

All polymers that stress relax or creep do so over several decades of time. A Maxwell element with spring of modulus G and dashpot of viscosity η under a fixed strain ϵ has a stress-relaxation modulus $G(t) = \sigma(t)/\epsilon = \exp(-t/\tau)$ where the relaxation time $\tau = \eta/G$. The stress decays from 80% to 10% in one decade of time, from 0.2τ to 2τ. To model variation over a wider span of times, several Maxwell elements are placed in parallel, each with its own relaxation time τ_i, and a final relaxed modulus G_∞ is added:

$$G(t) = G_\infty + \sum_{I=1}^{I=N} E_i e^{-t/\tau_i}. \tag{12}$$

To create a model to fit real data, the sum is turned into an integral, and a logarithmic time scale is used, to give

$$G(t) = G_\infty + \int_{-\infty}^{+\infty} H(\tau) e^{-t/\tau} d(\ln\tau). \tag{13}$$

Here $H(\tau)$ is the relaxation-time spectrum, which gives the contribution to stress relaxation associated with relaxation times between $\ln\tau$ and $\ln\tau + d(\ln\tau)$. An equivalent model for creep, using an array of Voigt elements in series, gives an expression for the creep compliance $J(t)$ in terms of the instantaneous elastic compliance J_0 and the retardation time spectrum, $L(\tau)$:

$$J(t) = J_0 + \int_{-\infty}^{+\infty} L(\tau)[1 - e^{-t/\tau}]\, d(\ln\tau). \tag{14}$$

It seems that by this analysis, data for modulus or compliance as a function of time are merely transformed by a Laplace transform into an equivalent representation. Is there an advantage? One is that molecular theories of polymer viscoelasticity give a relaxation-time spectrum. If this is a discrete spectrum then the summing of an exponential series, as in Eq. (12), will give the modulus as a function of time, $G(t)$, for comparison with experimental data. Within the limits of linear theory $G(t)$, $J(t)$, $H(\tau)$, $L(\tau)$, and $G^*(\omega)$ are all related, and in principle any can be derived from any other (Ferry, 1970; Arridge, 1975). There are practical problems in some of these transformations, including for the time spectra the necessity for getting data over 10–15 decades of time. This normally requires the use of time–temperature superposition.

2.2.2 The Glass Transition and Time–Temperature Superposition Qualitatively, polymers soften and flow more as the temperature is raised, or as the time scale of the deformation is extended. It is found that the plots of modulus (or other elastic property) as a function of log(time) at two temperatures can be superimposed by a horizontal shift of the time scale, by $\log(a_T)$. This works at temperatures above T_g. It is made more exact by a small vertical shift of the modulus data, b_T, which takes into account the thermal expansion of the material and the increase of rubbery modulus with absolute temperature:

---

*The dissipative dashpots should be made the resistors, and the energy-storing springs the capacitors, although springs are always drawn to look like resistors do in circuit diagrams. Worse, to allow charge to be equivalent to strain and potential to stress, every parallel arrangement is analogous to the series electrical case, and vice versa.

$$G(t,T_1) = b_T G(t/a_T,T_2). \tag{15}$$

This relation allows data taken over a limited range of times and temperatures to be turned into a master curve over many decades of time. A data set taken from the work of Ferry is shown in Fig. 3, in its original form [Fig. 3(a)] and after shifting to form a master curve over 11 decades of time [Fig. 3(b)]. The shift factor a_T is not a function of time, and was found empirically to have the following form, which is called the WLF equation:

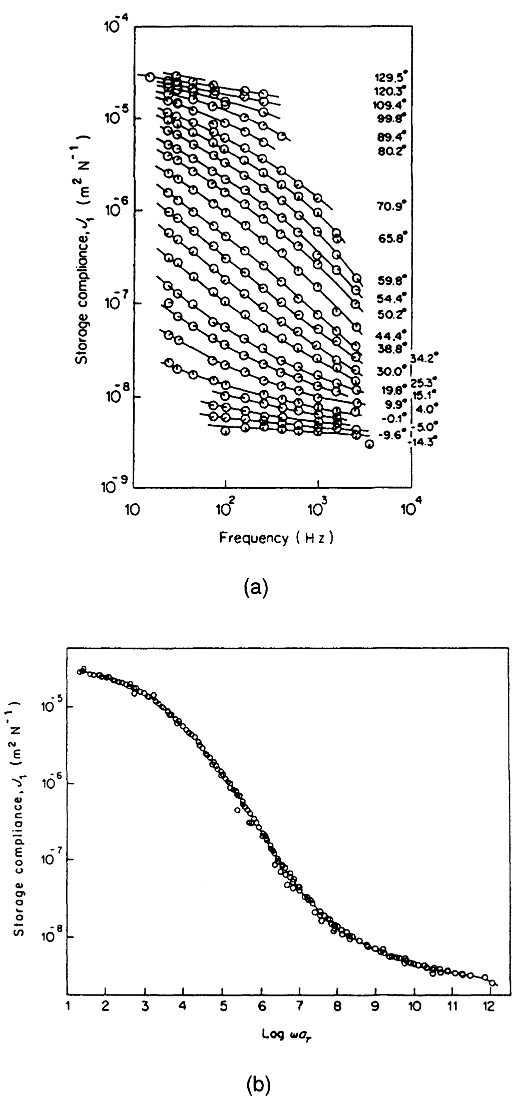

(a)

(b)

FIG. 3. (a) Storage compliance of poly(*n*-octyl methacrylate) measured in the frequency range 20–400 Hz at temperatures from −14 to 129.5 °C. (b) A master curve obtained by shifting the data shown in (a) to give the compliance at 100 °C over 11 decades of frequency. (From Ferry, 1970.)

$$\log a_T = -\frac{C_1(T - T_S)}{T - T_S + C_2}. \tag{16}$$

To a first approximation, when the reference temperature T_S is taken to be T_g as measured from volumetric thermal expansion, $C_1 \simeq 17$ and $C_2 \simeq 52$ for *all* polymers. A better fit for a wide range of materials is obtained by setting $C_1 = 8.86$ and $C_2 = 101.6$ and making T_S an adjustable parameter, generally found to be about $T_g + 50$ (Ferry, 1970). This universal behavior can be explained in terms of the free volume in the liquid and its relation to molecular mobility.

2.2.3 Nonlinear Viscoelasticity The equations describing nonlinear viscoelasticity may involve entirely empirical forms, complicated representations, or both of these together. Generally the range of materials and conditions over which any equation is applicable is quite limited. Thus polymer liquids are dealt with separately from polymer solids. An example of an empirical correlation between linear and nonlinear viscoelasticity in polymer liquids is the Cox–Mertz rule:

$$\eta(\dot{\gamma}) = |\eta^*(\omega)|_{\omega=\dot{\gamma}}. \tag{17}$$

This means that the steady-state viscosity at a shear rate $\dot{\gamma}$ is equal to the magnitude of the complex viscosity, as defined in linear theory and measured in oscillatory flow with a frequency equal to the constant shear rate. There is no theoretical justification for the Cox–Mertz rule. One of many models for polymers is called the Oldroyd fluid *B*. Its nonlinear constitutive equation for fluid flow is

$$\boldsymbol{\sigma} + \tau\frac{\delta\boldsymbol{\sigma}}{\delta t} = \eta\left(\dot{\boldsymbol{\Gamma}} + \tau'\frac{\delta\dot{\boldsymbol{\Gamma}}}{\delta t}\right). \tag{18}$$

Here $\boldsymbol{\sigma}$ is the stress tensor, $\dot{\boldsymbol{\Gamma}}$ is the strain-rate tensor, and τ and τ' are material constants. This level of complication can model shear-thinning and normal forces in polymer flow, but it is still a Maxwell-type model, and a relaxation-time spectrum needs to be introduced to model a real material.

In solids, nonlinear viscoelasticity is often concerned with creep. Figure 4 shows a schematic surface plot where the height is strain, and the two axes are stress and log(time). The surface can be defined by a few experiments, measuring creep at two different constant loads, and strain as a function of stress at a fixed time of loading. Interpolation and a modest extrapolation then give the strain under a wide range of loads and times, but only for a constant stress. If the loading sequence is more complicated, loading followed by partial unloading for example, these data are of no help. Alternative empirical approaches include power-law formulas similar to those used to describe creep in metals. These are used when the total strain is small, and again for constant-stress loading. For elastomers at large strains the MRS formula has been described in Sec. 2.1.2, Eq. (10).

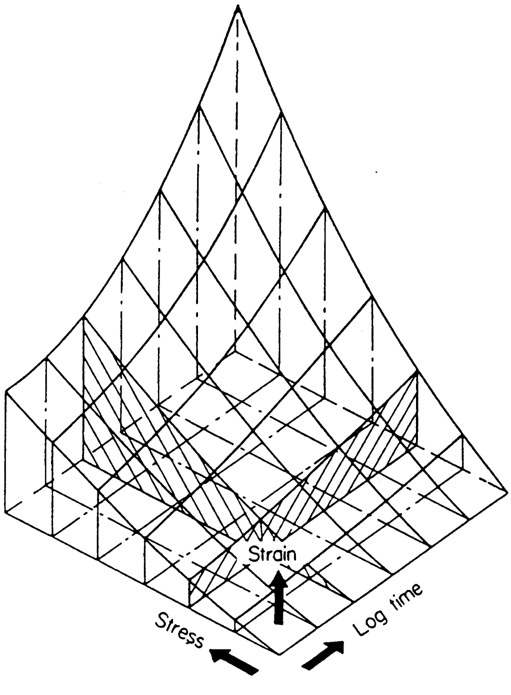

FIG. 4. Schematic plot of nonlinear elastic behavior. The height of the surface is the strain. Sections are shown at constant stress—a creep curve—and at constant time—an isochronous stress–strain curve. Two or more creep curves and one isochronous stress–strain curve may be enough to define the strain surface for design purposes. (From Ward, 1979.)

Creep and recovery experiments often show that the strain cannot be separated into a function of stress only and a function of time. In this case a multiple integral representation of the strain function has occasionally been used, although it contains a large number of parameters. A large amount of data must be collected before the consistency checks allow the representation to be anything more than curve fitting. In principle, the method is an extension of the Boltzmann superposition principle of linear viscoelasticity. In that case, the strain caused by stresses $\Delta\sigma_1$ at time $\tau_1 + \cdots + \Delta\sigma_n$ at time τ_n is given by

$$e(t) = \Delta\sigma_1 J(t - \tau_1) + \cdots + \Delta\sigma_n J(t - \tau_n). \tag{19}$$

In the multiple-integral approach, second-order terms of the form $\Delta\sigma_1\Delta\sigma_2 J_2(t - \tau_1, t - \tau_2)$ are added to the expression for $e(t)$, then third-order terms $\Delta\sigma_1\Delta\sigma_2\Delta\sigma_3 J_3(t - \tau_1, t - \tau_2, t - \tau_3)$ and so on. The J_N are functions of time that must be determined. When the stress is a continuous function of time, each set of terms becomes a multiple integral. This can be useful, but only if it is found that the data can be fitted with very few terms.

2.3 Anisotropic Elasticity

All polymer molecules are intrinsically anisotropic, and so if the molecules are oriented, in a glass or in a crystal, the material will be elastically anisotropic. A generalized Hooke's Law gives a modulus or compliance that is no longer a scalar, but a fourth-order tensor. The 3D set of equations for the linear region is conventionally given in a contracted form:

$$\sigma_p = \sum_{q=1}^{q=6} c_{pq}\epsilon_q \quad \epsilon_p = \sum_{q=1}^{q=6} s_{pq}\sigma_q. \tag{20}$$

Here **C** is the matrix of elastic constants, components c_{pq}. Each c_{pq} is the ratio σ_p/ϵ_q when all other strain components are zero. **S** is the compliance matrix, components s_{pq}. Each s_{pq} is the ratio ϵ_p/σ_q when all other stress components are zero. Indices 1, 2, and 3 refer to normal stress and extensional strain along x, y, and z respectively. Indices 4, 5, and 6 refer to shear in the yz, xz, and xy

planes (Fig. 5). Thus σ_{11} is contracted to σ_1 and σ_{23} is contracted to σ_4. External stability requires that $c_{pq} = c_{qp}$, and this reduces the number of independent components from 36 to 21. If the material has orthotropic symmetry (where there are three mutually perpendicular twofold symmetry axes), the number is reduced to 9, with a compliance matrix of the form

$$\mathbf{S} = \begin{vmatrix} s_{11} & s_{12} & s_{13} & 0 & 0 & 0 \\ s_{12} & s_{22} & s_{23} & 0 & 0 & 0 \\ s_{13} & s_{23} & s_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & s_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & s_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & s_{66} \end{vmatrix}. \qquad (21)$$

A common case of anisotropic elasticity in polymers where orthotropic symmetry exists is a sheet biaxially drawn or drawn at constant width, such as the polyethylene sheet used for plastic bags. Conventionally, in this case the z axis is taken as the draw direction of the sheet, x is transverse to the draw direction and in the plane of the sheet. The y axis is then the normal to the sheet. The polymer molecules tend to lie in the plane of the sheet, but more along z than along x. The extensional compliances in the machine and transverse directions in the plane of the sheet are s_{33} and s_{11}, respectively. They will be different and both will be less than that of an isotropic material, while the compliance perpendicular to the sheet, s_{22}, is higher. The out-of-plane compliance is difficult to measure in real thin sheets, and is also of little practical importance in many cases.

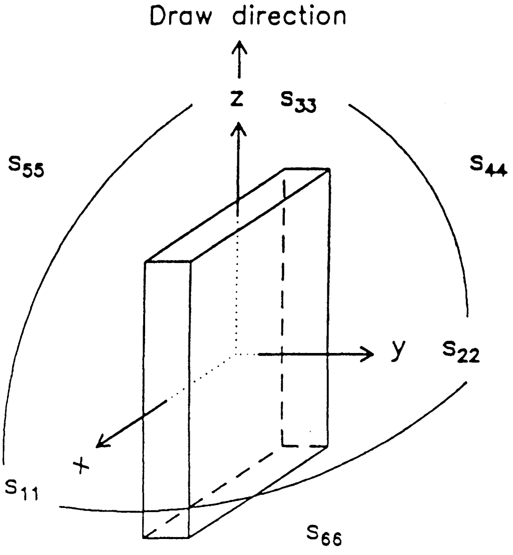

FIG. 5. Axes and compliance component directions for a polymer sheet with orthotropic symmetry (from Grubb, 1993).

If the drawing is uniaxial and the sample dimensions along x and y not too different, then the x and y axes become equivalent. This is called transverse isotropy or fiber symmetry, and such samples then require only five independent elastic constants. The x or 1 axis may be chosen as the unique (fiber draw) axis, but in polymer fibers the z or 3 axis is more common. This is because the molecular-chain axis becomes oriented along the fiber axis by drawing and the chain axis is usually labeled the c, z, or 3 axis in polymer crystals. In this case $s_{22} = s_{11}$, $s_{23} = s_{13}$, $s_{55} = s_{44}$, and $s_{66} = 2\ (s_{11} - s_{12})$.

The full compliance or stiffness matrix has been determined in only a few cases, so theoretically derived components are often used for perfect crystals, and estimates for amorphous or semicrystalline materials. Normally a few measurements are made and individual components are reported as Young's moduli or shear moduli. Consider a material with orthotropic symmetry that has a tensile stress σ applied along z. Here $\sigma_1 = \sigma$, all other stresses are zero, and there will be strains ϵ_1, ϵ_2, and ϵ_3. The tensile or Young's modulus in the 3 direction, E_3, is the ratio σ/ϵ_3, so that $s_{33} = 1/E_3$. The Poisson's ratios for this experiment are defined as $\nu_{31} = -\epsilon_1/\epsilon_3$ and $\nu_{32} = -\epsilon_2/\epsilon_3$. Thus s_{31}, which by definition is ϵ_1/σ when other stresses are zero, is equal to $-\nu_{31}/E_3$. This relation may also be expressed as $\nu_{31} = -s_{31}/s_{33}$. A shear experiment in the yz plane of the orthotropic material will give the shear modulus $G_{23} = 1/s_{44}$. Thus the compliance terms along the diagonal of the compliance matrix, Eq. (20), are the inverses of the three Young's moduli and three shear moduli, and the off-diagonal terms relate to the Poisson's ratios.

In a fiber used as a primary mechanical component such as a tire cord, the most important elastic constant is the tensile modulus in the fiber direction, $E_{\parallel}$. The tensile modulus in the transverse direction, $E_{\perp}$, may also be important, but all transverse properties are difficult to measure. The torsional modulus about the fiber axis $G_{\parallel} = 1/s_{44}$ is the only other property likely to be determined.

3. YIELD AND FRACTURE

A polymer may be brittle, failing in tension without visible plastic deformation, or may be extremely ductile, deforming to 5 or even 50 times its original length before failing. Cross linking inhibits flow, and so thermosets and elastomers may be truly brittle. Other polymers such as polystyrene are macroscopically brittle with no visible plastic deformation, but they yield and flow very locally near the crack before fracture. Tougher glassy polymers such as polycarbonate will yield, but fail in tension after 10%–20% extension. The very highly ductile materials are semicrystalline polymers tested above the T_g of their amorphous component. Below this T_g they are brittle. Multiphase polymer systems are more difficult to generalize. Adding 10% by volume of ~1 μm rubber particles to polystyrene gives "high-impact polystyrene" (HIPS), which is a tough material that fails in tension at about 50% strain. A block copolymer with glassy and rubbery domains (polystyrene and polybutadiene) can yield and draw to three times its own length, but most of the post-yield deformation is reversible.

A brittle material tested in tension fails at a low extension because of cracks initiated at flaws. In compression the cracks cannot grow and the polymer will yield and flow. The compressive strength is greater than the tensile strength not only because of a higher extension to failure but also because the modulus is larger in compression. Except for elastomers, polymers cannot be considered incompressible, and so the hydrostatic component of the stress field causes a volume change. Simply comparing this volume change with that caused by thermal expansion indicates that the change from tension to compression is equivalent to a large drop in testing temperature. As noted before, reduced mobility of the polymer chains at low temperature or high testing rates gives a higher elastic modulus. Many real situations involve bending, where part of the sample is in compression and part in tension. The effects described here combined with the nonlinear viscoelasticity of polymers can make failure in bending very difficult to analyze properly. Many reported bending (flexural) strengths have been obtained by applying linear elastic analysis.

Polymers may yield by shear or by crazing. Crazes are cracklike objects where the two faces are bridged by many fine (~20 nm diameter) fibrils of polymer. The volume fraction of fibrils and thus the relative density of material in the craze is 50% or less, and so there is a large volume increase when crazes form. This means that crazing cannot occur unless there is a net hydrostatic tensile component to the stress field. Shear yield is also affected by hydrostatic pressure. For most polymers the best criterion for shear yield in a general stress field is the von Mises criterion with a pressure-dependent yield stress:

$$(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2 = 6(\tau_y + \alpha P). \tag{22}$$

Here σ_1 to σ_3 are the principal stresses, that is, the stresses along axes chosen so that the shear stresses σ_4 to σ_6 are zero. The left-hand side is the elastic-energy density due to shear strain (change in shape). The term P is the hydrostatic pressure, $P = -(\sigma_1 + \sigma_2 + \sigma_3)/3$, and τ_y is the yield stress at zero pressure, which depends on temperature and strain rate. The coefficient α increases as temperature is increased to near T_g. At a fixed temperature the yield surface is a tapering cylinder, centered on the $\sigma_1 = \sigma_2 = \sigma_3$ axis and getting broader as P increases. There are many suggested criteria for crazing, but all involve some dilatation. In terms of stress, an expression that fits the data in plane stress ($\sigma_3 = 0$) is

$$\sigma_1 - \sigma_2 = A + B/(\sigma_1 + \sigma_2), \tag{23}$$

where A and B are constants. Figure 6(a) shows these criteria for shear and crazing at $\sigma_3 = 0$ superimposed along with the straight line of zero pressure $\sigma_1 = -\sigma_2$. Figure 6(b) shows the changes expected if $\sigma_3 < 0$ (a compressive stress). All yield is suppressed, and the line of zero pressure moves to the right. Deformation will tend to change from crazing to shear yielding when compressive stresses are applied.

Crazes are often associated with tensile failure, which occurs when a crack grows through the weaker craze material. Crazing is intrinsically inhomogeneous, although a region containing many small crazes may appear uniform on its larger scale. Shear

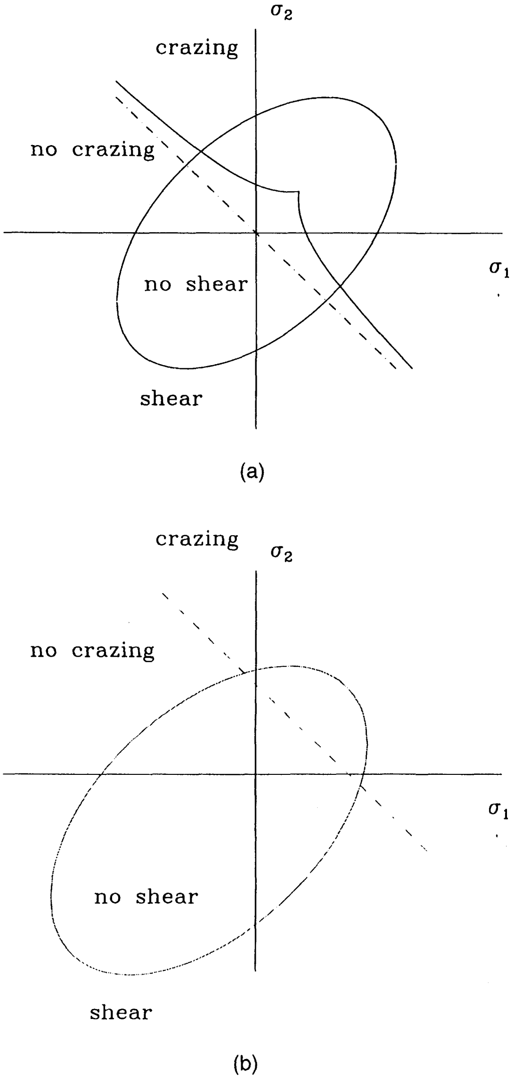

FIG. 6. Schematic plots of yield criteria for shear and for crazing. The ellipse is the von Mises criterion for shear, yielding, Eq. (20). The material will undergo shear yield when the stress is outside the ellipse. The cusped line in (a) for $\sigma_3 = 0$ is Eq. (21), and the material will craze to the right of this line. In (b) $\sigma_3 < 0$; the ellipse increases in size and shifts to the lower left. The crazing criterion moves to the upper right. In this case only biaxial tension will be sure to produce crazing.

yield may be inhomogeneous or uniform throughout the body of the material being deformed. In polymers, it is nearly always inhomogeneous. In compression or shear, localization of deformation must be due to an intrinsic strain softening of the material. This is similar to the shear-thinning behavior seen in polymer liquids. The more common experiment is tensile deformation, and this is more complicated.

3.1 Localized Shear Yield in Tension

The process of localized shear yield in tension is easy to demonstrate in a semicrystalline polymer. Cut two long strips from a plastic bag (probably made of polyethylene), cutting parallel to two different sides so that the long directions of the two strips are perpendicular. Pull on the strips. They will both deform inhomogeneously, in one case with a well-defined very narrow zone of deformation that moves slowly along the strip. After this "neck" has passed, the material has undergone a large extensional strain—it has been drawn. The draw ratio or extension ratio, which is (final length)/(initial length), may be 3 to 5. The neck is the zone of localized shear deformation. It is easy to see the process because the material strain hardens, and this stabilizes the neck. The two strips behave differently because the film used to make the bag is already oriented. The conventional stress–strain curve that would result from repeating this experiment in a tensile-testing machine is shown in Fig. 7. This

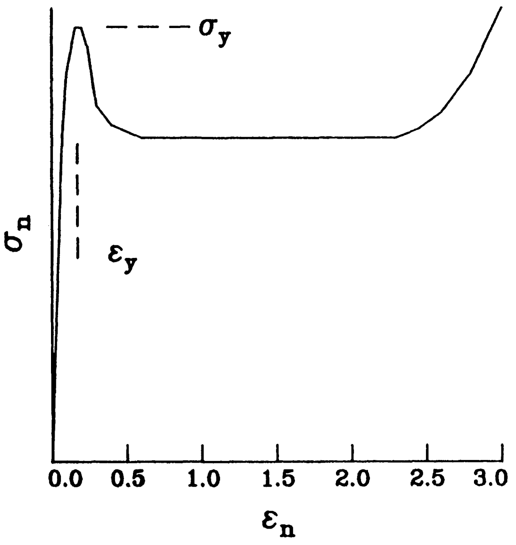

FIG. 7. Conventional stress–strain plot for a ductile semicrystalline polymer that deforms by necking. The neck forms at the yield point, which is a stress maximum, and moves through the sample at a constant drawing stress.

plots the nominal stress σ_n = force/A_0 vs the engineering or nominal strain $\epsilon_n = \Delta l/l_0$. There is a maximum in the nominal stress at the yield point σ_y, ϵ_y when the neck first forms. The long horizontal post-yield portion at the draw stress σ_d corresponds to travel of the stable neck, and the final increase in σ_n is when the neck reaches the end of the sample. This process may be called "cold drawing" or necking. In most materials, including glassy polymers, this necking down would continue until failure. Here the increasing molecular orientation in the extending region eventually increases its strength enough so that it supports the rising local stress. This defines the "natural" draw ratio.

The conventional plot of Fig. 7 cannot correspond to material behavior beyond the yield point, when some material has deformed and the rest has not. Further, the localized deformation and the load drop may be due to the local reduction in cross-sectional area. The material need not strain soften, it need only strain harden by too little to compensate for the lower area carrying the load. The effect of change in area from A_0 to A is clarified by replotting the data as true stress σ_t = force/A vs nominal strain. This plot, shown schematically in Fig. 8, is Considère's construction. At constant volume (a normal assumption for plastic deformation) there will be an extremal value of $\sigma_n = \sigma_t(1 + \epsilon_n)$ at points where a line drawn from $\sigma_t = 0$, $\epsilon_n = -1$ (i.e., $\lambda = 0$) is tangent to the curve in Considère's construction. The steeper tangent touches the curve at the yield point, and the lower one touches it at stress and strain in the region that the neck has passed through.

The lower curve (a) in Fig. 8, which has a sharp drop in σ_t, is the transformed conventional stress–strain curve, but this is not likely to be the actual behavior of the material. One possibility is the upper curve (b) where there is no decrease in true stress and therefore no strain softening. Measurement of the actual cross-sectional area in the neck is needed to give any information about the material properties between the two tangent points. Even this is not sufficient, for two reasons. One is that the strain rate is very nonuniform, being much higher in the neck region. The other is that the stress is far from simple tension in the neck region, because of the sudden change in cross section.

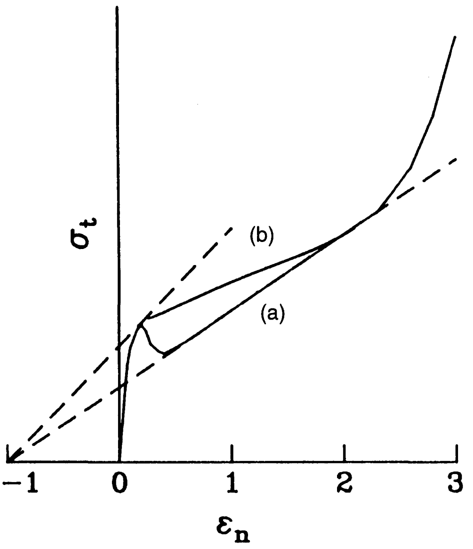

FIG. 8. Considère's construction; a plot of true stress vs nominal strain. The lower curve, *a*, is the transformed conventional stress–strain data from Fig. 7. A constant nominal or engineering stress, which was a horizontal line in Fig. 7, becomes a line passing through the point (−1,0). Material following the upper curve *b*, where there no strain softening, would give the same conventional stress–strain result.

This explains how it is that new material entering the neck can deform without reaching the yield stress. Exact modeling of the necking process requires a 3D finite-element model using nonlinear elasticity, and the anisotropy induced by deformation.

At one time it was thought that the heat generated in the deformation process played a vital role in necking, and this may be the case when the process is done at high speed. Then the local heating produces local softening, which supports local flow. It is certainly true that in commercial fiber drawing, where many filaments are deformed in an array, the heat radiated from one neck influences the position of the necks in neighboring fibers. In this way the positions of the deformation zones in all the fibers are stabilized. However, drawing at very slow rates at room temperature gives basically similar results for draw ratio and drawing stress, so that the effects of heat are only secondary in most cases.

The draw ratio of necked material is controlled by the extensibility of the molecular network. That is, the deformation after yield is like that of an elastomer, with the molecular entanglements acting as crosslinks. If the drawing is done slowly and at high temperature, the molecules have a chance to disentangle during draw. The elongation increases and the draw stress falls. Careful control allows draw ratios of 20–30 compared with the usual 3–5 in polyethylene. Even greater draw ratios can be obtained by pretreating the material to reduce the entanglement density, by dissolving and precipitating high–molecular-weight polymers.

3.2 Localized Shear Yield in Compression

In compression the deformation of glassy polymers is often localized into shear bands or kink bands, and this must be due to the intrinsic strain-softening property of the material. A single large shear band may develop that has well-defined boundaries and may be 0.1–0.5 mm across. If deformation is continued, the single shear band will pass through the entire sample. The sample will then fail by shear fracture within this shear band. The more diffuse shear zones contain many fine-scale shear bands. Continued deformation in this case increases the average shear strain so that a kink appears in the specimen. Isotropic crystalline polymers strain soften little or not at all, and so the yielding is more homogeneous.

Well-oriented high-performance polymer fibers are well known to be weak in axial compression. The compressive strength is only 10%–20% of their tensile strength. Direct measurement of the properties of a 10–30-μm-diameter fiber in compression along its axis is obviously very difficult. In compression the modulus falls with increasing strain either because of reduced chain orientation, or because of intrinsic softening by bending of the aligned chains. At strains of only ~0.1%, kink bands form. They are well-defined narrow bands of sheared material that extend through the fiber. Microscopic studies show that the molecular chains undergo sharp bends at the boundaries of kink bands. The lattice images are almost like those of twins, but molecular continuity requires that the molecules shear past each other to form the kink band. Applying tensile load to the kinked fiber pulls the kinks out of it, but in some cases the tensile strength is reduced by the compressive damage to the fiber. Modeling kink formation in fibers indicates that the process is one of fibril buckling, but in these well-ordered materials the deformation is related to the crystalline slip systems, with specific crystallographic rotations preferred for different materials.

3.3 Crazing and Brittle Fracture

Many glassy polymers are brittle, fracturing with no visible plastic deformation. The energy required to form cracks in these materials is $\sim 100\ \mathrm{J\ m^{-2}}$, while a simple model for the energy needed to break the molecular chains predicts $\sim 1\ \mathrm{J\ m^{-2}}$. In metals, a similar discrepancy is caused by plastic deformation in the region of the crack tip, which absorbs a lot of energy. The same is true in polymers, but the plastic deformation often has a special form called a craze. A craze contains voids and fine fibrils of oriented polymer (~30 nm in diameter). The voids make a continuous phase, bridged by the fibrils that pass from one side of the craze to the other. At the surface of a growing craze, the homogeneous glass flows in a very localized region to give an oriented fibril. The process is very like fiber drawing, and the extension ratio is controlled by the extensibility of the molecular network (Kramer and Berger, 1990). If the molecular weight is too low to form entanglements, there is no network; the craze is not stable, and it immediately breaks down into a crack. If the entanglement density is high, the extensibility is low, and the stress required to form a craze increases. The shear yield stress is not much affected by the entanglement density, and so crazes do not form in highly entangled material. The deformation near crack tips then occurs by shear. Crosslinking at low densities also inhibits crazing. Figure 9 is a summary of results from many polymers, blends, and crosslinked PS. The high values of network strand density ν correspond to highly entangled or crosslinked materials; these have low extensibility and deform by shear in deformation zones (DZ).

Crazes have been introduced as associated with cracks, but crazes form first, and cracks form when the craze material breaks down. In some materials such as ABS and i-PP many small crazes form on deformation, so many that the light scattering from craze

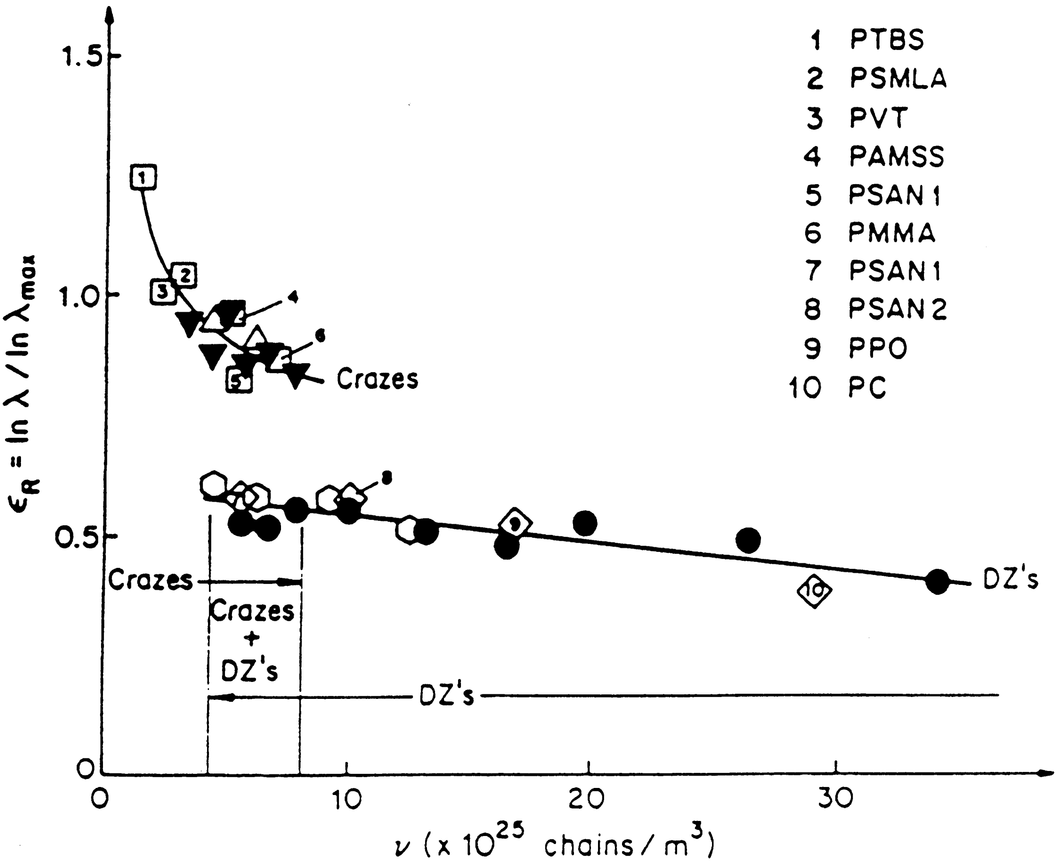

FIG. 9. True strain ratio ϵ_r in crazes and shear deformation zones (DZ) in a range of polymers, copolymers, and blends. Filled triangles and circles represent crosslinked PS. The higher crosslink density materials (higher ν) have lower extensibility and deform by shear instead of by crazing. (From Henkee and Kramer, 1984.)

surfaces turns the material white. The energy required to form these many crazes makes the material tough. In the case of i-PP, parts can be designed with a thin region that acts as a hinge. This region may be "stress whitened" at its first use, and then withstand thousands of bending cycles after this nominal "failure." The case of i-PP also illustrates that semicrystalline polymers also craze. Crazes in glassy polymers are straight and regular in simple tension, while those in crystalline materials are irregular. Figure 10 shows both types of craze in a thin film of i-PS that contains spherulites in a glassy matrix. The spherulitic structure of semicrystalline polymers gives it highly inhomogeneous and anisotropic mechanical properties on the local scale, and this makes the craze path tortuous.

3.4 Ductile–Brittle Transition

The classic ductile–brittle transition in metals is that above some transition temperature failure is ductile, and below it, failure is brittle. This is explained by the lower flow stress at higher temperatures. The same effect is seen in polymers; as already mentioned, both semicrystalline and glassy polymers tend to be brittle below T_g and ductile above it. The brittle crack propagation is usually associated with crazing and plastic flow with shear yielding. Thus the yield criteria of Fig. 6 show that brittle failure will be suppressed by the application of hydrostatic pressure. The viscoelastic properties of polymers ensure that a higher strain rate requires a higher flow stress. Thus failure under impact is more likely to be brittle. As with metals, the presence of a notch or pre-

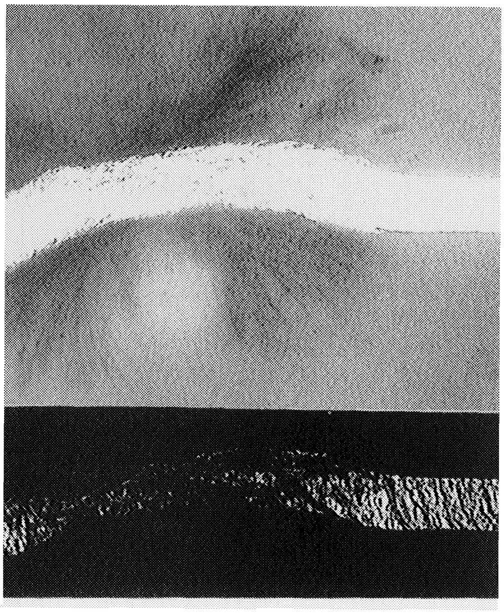

FIG. 10. Electron micrograph of a craze in partially crystallized isotactic PS. The upper image shows the spherulitic lamellar structure, made visible by staining with RuO_4. Because the drawn craze material is much thinner, it appears as a crack in this exposure. The lower image is a shorter exposure of the same area. The regular craze fibril structure is seen in the amorphous area at the right-hand side. The structure within the semicrystalline region is very confused.

existing small crack tends to make the failure brittle. This is because initiation of a crack can no longer be a barrier to fracture. To make for the worst case, and to remove the effects of small flaws in the sample, most impact testing of polymers is done on notched samples.

The environment of the polymer can have a great effect on crazing and fracture. Often the effect of a liquid environment is to weaken the material, and this is very important for practical design. The inner shell of a refrigerator must withstand the effects of butter and other oils. If the liquid can diffuse into the polymer and soften it, the effect on the craze is severe. Capillary forces draw the liquid into the craze. The region at the growing tip and the fine fibrils are quickly penetrated even if the diffusion coefficient is very low. The swollen polymer has a lower T_g and can disentangle more easily, so that the craze fibrils elongate more. The volume fraction of polymer in the craze falls, the strength of the craze falls, and cracks form. Even if there is no penetration of the liquid into the polymer, it can still affect craze formation and lower the material strength. Crazes have a very large specific surface area, and if the liquid reduces the energy of these surfaces, the energy to form the craze is reduced.

Crosslinking suppresses flow, and thus it is to be expected that an increase in crosslink density will tend to make the failure brittle. However, the opposite effect can also be obtained, if the initial material is one that flows very easily and forms weak crazes. This material appears weak and brittle. A little crosslinking reduces the extension ratio of the craze material, and increases the stress required to form a craze. The failure becomes ductile, and remains ductile until the crosslink density becomes much greater.

The thickness of the sample being tested also affects the mode of failure. A thick sample approaches the conditions of plane strain, and there is a triaxial tension within the material. A thin sample has no constraint in the thickness direction, so is in the condition of plane stress. Thus a thick sample may craze and crack, while a thin test piece may be ductile. At intermediate thickness, there can be shear at the outside edges and voiding in the center of the sample.

4. FATIGUE AND WEAR

Fatigue, failure due to repeated stresses or strains, is practically very important. The fatigue strength of a polymer may be four or five times less than its static tensile strength. Fatigue testing of a viscoelastic material requires that the frequency be carefully chosen. Frequencies greater than about 1 Hz may increase the temperature of the specimen significantly. The test frequency should be kept close to the real frequency of loading in the field in this case, although this may require long tests. Constant-stress and constant-strain tests will give very different results. Constant-stress tests show a reduction in fatigue strength for most polymers with increasing temperature; constant-strain tests will have the opposite trend. Constant strain should be chosen when appropriate to the end use of the material, as for polymer strips used as hinges, or sealants. In this case a crack may form and reduce the stress on the remaining material, and it is important to know whether the crack formation would correspond to failure in the real situation.

The fatigue life of polymers is found to be much reduced if there is complete unloading or a compressive stress in part of each fatigue cycle. This is counter to the expectation that fatigue is worse when the mean stress is raised, and is related to the destruction of the fibril structure at the craze tip when stress is reversed.

The measurement of wear and abrasion is entirely empirical. Different standard test machines use different forces and rubbing rates, and the results are difficult to interpret at best. A pair of polymeric materials may have a wear rate ratio of 2:1 on one machine, and have the opposite result, a ratio of 1:2, on another. One reason for this is that consistent results for wear of a polymer rubbing against a metal surface require control of the exact form of the surface roughness of the metal. Wear and abrasion are closely related to friction, as pieces are pulled off the sample by frictional forces during wear. Relying on this simple concept is very dangerous. Early replacement hip joints were made of PTFE on one side and surgical steel on the other. PTFE was chosen because its coefficient of friction is very low, but the wear rate of PTFE against metal in the joint was very high, and all the joints failed in a short time. Now ultrahigh–molecular-weight PE is used wherever wear and low friction are important: in replacement joints, ski bases, and industrial conveyer systems. Tires should have low wear rates combined with high friction over a wide temperature range. This is difficult to achieve, as the best adhesion comes from soft materials that do not wear well.

GLOSSARY

Craze: A cracklike object in a polymer that is bridged by many fine fibrils of oriented material.

Degree of Crystallinity: Polymers that can crystallize generally form very small crystals with many defects and a large interfacial area. Many of their properties are modeled as those of a mixture of some fraction X_C of perfect crystals and $1 - X_C$ of ideal amorphous material. X_C is the degree of crystallinity.

Dynamic Mechanical Testing: Also called dynamic mechanical analysis or DMA, this test involves applying a sinusoidally varying load to the sample, and measuring the amplitude and phase of the displacement that results. Tension, torsion, simple shear, and bending are common modes of loading.

Elastic Constant: Generally, an element c_{pq} of the compressed (6×6) elastic constant matrix **C** or an element s_{pq} of the compressed compliance matrix **S**. Either matrix describes the elastic properties of an anisotropic material.

Elastic Moduli: Ratios of stress to strain. If the stress is hydrostatic, the ratio is the *bulk modulus;* if it is a uniaxial tensile stress, the ratio is the *Young's modulus,* unique only for isotropic materials.

Elastomer or Rubbery Material: A material that contains mobile linear chains connected together to form an insoluble and infusible network. It differs from a thermoset by having a much lower density of connection points, called *crosslinks.*

Extension Ratio: The ratio (final dimension)/(initial dimension), λ; this may be called the *draw ratio* for tensile deformation (see **Strain**).

Fiber Symmetry: While technically a fiber is a long fine filament, any sample may have fiber symmetry of its mechanical properties if it has one unique axis (the fiber axis) and isotropy in the plane perpendicular to that axis.

Glass Transition: On cooling a liquid its viscosity increases dramatically over some temperature range and it becomes a glass. In the glassy state, large-scale motions do not occur and the liquid disorder is frozen in place. The glass-transition temperature T_g describes where the transition occurs.

Kink Band: A well-defined region where local shear has caused a reorientation of the polymer chains. Kink bands in well-ordered crystalline polymers become similar to twins in crystals.

Necking: Local reduction in cross section caused by localized shear during tensile deformation.

Rubber Elasticity: The property of elastomers that allows very large recoverable elastic deformations. The restoring force in an ideal rubber elastic material is due only to the entropy changes in the molecular chains when they are stretched.

Strain: The displacement per unit length; if the initial and instantaneous lengths are l_0 and l, the commonly used "nominal" strain is $(l - l_0)/l_0$. For large strains the correct or "true" extensional strain is $\ln(l/l_0) = \ln(\lambda)$.

Thermoset: A polymer where the chains are connected together to form an insoluble and infusible network. The network is often formed chemically by the application of heat to a liquid mixture, giving the name.

Time–Temperature Superposition: The principle that the mechanical response of a polymer at a different temperature is the same as its response at the same temperature after a different time of loading.

Viscoelasticity: The combination of elastic and viscous response to applied stress. In *linear viscoelasticity* the response is always linear with the applied stress.

Works Cited

Arridge, R. G. C. (1975), *Mechanics of Polymers,* Oxford: Clarendon.

ASTM (1994), *1994 Annual Book of ASTM Standards,* Philadelphia, PA: American Society for Testing and Materials.

DIN (1994), *DIN-Katalog für Technische Regeln, Zusatsband, Übersetzungen,* Beith Berlin: Deutche Institut für Normung, eV.

Ferry, J. D. (1970), *Viscoelastic Properties of Polymers,* New York: Wiley.

Flory, P. J. (1969), *Statistical Mechanics of Chain Molecules,* New York: Interscience.

Grubb, D. T. (1993), in: E. L. Thomas (Ed.), *Materials Science and Technology,* Vol 12, *Structure and Properties of Polymers,* Weinheim: VCH.

Henkee, C. S., Kramer, E. J. (1984), *J. Polym. Sci. Polym. Phys. Ed.* **22,** 721.

Kramer, E. J., Berger, L. L. (1990), *Adv. Polym. Sci.* **91/92,** 1.

Peterson, J., Ranby, B. (1970), *Makromol. Chem.* **133,** 251.

Treloar, L. R. G. (1969), *The Physics of Rubber Elasticity,* 2nd ed., (1975), 3rd ed., Oxford: Clarendon.

Treloar, L. R. G. (1976), *Proc. R. Soc. A* **351,** 301.

Ward, I. M. (Ed.) (1979), *Structure and Properties of Oriented Polymers,* New York: Wiley.

Further Reading

Arridge, R. G. C. (1975), *Mechanics of Polymers,* Oxford: Clarendon.

Ferry, J. D. (1970), *Viscoelastic Properties of Polymers,* New York: Wiley.

Flory, P. J. (1958), *Principles of Polymer Chemistry,* Ithaca, NY: Cornell Univ. Press.

Gordon, J. E. (1976), *The New Science of Strong Materials,* 2nd ed., London: Penguin.

Hearmon, R. F. S. (1961), *An Introduction to Applied Anisotropic Elasticity,* Oxford: Clarendon.

Holliday, L., White, J. W. (1971), *Pure Appl. Chem.* **26,** 545–582.

Mark, J. E., Erman, B. (1988), *Rubberlike Elasticity, A Molecular Primer,* New York: Wiley.

McCrum, N. G., Read, B. E., Williams, G. (1967), *Anelastic and Dielectric Effects in Polymeric Solids,* New York: Wiley.

McCullough, R. L. (1977), in: J. M. Shultz (Ed.), *Treatise on Materials Science and Technology,* Vol 10, Part B, New York: Academic.

Nielsen, L. E., Landel, R. F. (1994), *Mechanical Properties of Polymers and Composites,* 2nd ed, New York: Marcel Dekker.

Thomas, E. L. (Ed.) (1993), *Materials Science and Technology,* Vol 12, *Structure and Properties of Polymers,* Weinheim: VCH. This volume contains among others the following useful long articles: Crist, B., "Plastic Deformation of Polymers." Grubb, D. T., "Elastic Properties of Crystalline Polymers." Narisawa, I., Lee, A. F., "Crazing and Fracture of Polymers."

Treloar, L. R. G. (1975), *The Physics of Rubber Elasticity,* 3rd ed., Oxford: Clarendon. Earlier editions are also good.

Vincent, J. F. V., Currey, J. D. (Eds.), (1980), *Mechanical Properties of Biological Materials,* Symposia of the Society for Experimental Biology, Vol. 34, Cambridge, U.K.: Cambridge Univ. Press.

Ward, I. M. (Ed.) (1979), *Structure and Properties of Oriented Polymers,* New York: Wiley.

Ward, I. M., Hanley, D. W. (1993), *An Introduction to the Mechanical Properties of Solid Polymers,* Chichester: Wiley.

POLYMERS, MODEL MATERIALS

See SYNTHESIS, CHARACTERIZATION, AND PROPERTIES OF POLYMERS

POLYMERS, MOLECULAR PROPERTIES OF

CHARLES M. GUTTMAN AND BRUNO FANCONI, *National Institute of Standards and Technology Polymers Division, Gaithersburg, Maryland, U.S.A.*

Introduction 549
1. **Naming of Polymers** 549
2. **Methods to Determine Overall Chemical Composition** 551
2.1 Nuclear Magnetic Resonance Spectroscopy 552
2.2 Infrared Spectroscopy 556
3. **Molecular-Weight Distributions and Their Determination** 557
3.1 Naming Molecular-Weight Distributions and Their Moments 557
3.2 Methods to Determine Moments of the Molecular-Weight Distribution 557
3.2.1 Membrane Osmometry 557
3.2.2 Vapor-Phase Osmometry 558
3.2.3 Light Scattering 558
3.2.4 End-Group Analysis 559
3.2.5 Ultracentrifugation 560
3.2.6 Intrinsic Viscosity 561
3.3 Methods to Determine Molecular-Weight Distributions 561
3.3.1 Size-Exclusion Chromatography 561
3.3.2 Mass Spectrometry 562
3.3.3 Fractionation Methods 562
4. **Determination of Chemical and Conformational Order along the Chain** 563
5. **Equilibrium Dimensions and Conformations of Single Chains in Solution** 563
5.1 Models without Long-Range Interactions 563
5.2 Models with Long-Range Interaction 564
Glossary 565
Works Cited 567
Further Reading 567

INTRODUCTION

Polymers, or macromolecules, are large molecules made up of many (often many thousands) chemical groups or repeat units. Often the same group is repeated many times, or a few groups are repeated many times. In this article, we will discuss the naming of polymers and the methods to determine the chemical groups, the molecular weight, and the architecture of the polymers. Our discussion is limited primarily to macromolecules synthesized from petroleum, usually called synthetic polymers. We mention also large-volume commercial polymers derived from vegetation; cellulose is an example. For simplicity we do not include polymers dominant in biochemistry and medicine, biopolymers.

1. NAMING OF POLYMERS

Polymers may be named according to several different conventions. We name a polymer by the *kinds* of different repeat units in the macromolecule. Homopolymers have the same chemical constituent repeated many times and are represented by A–A–A–A–A–A–A–A, where A represents a repeat unit. In copolymers a fixed number of chemical constituent units are repeated. Diblock copolymers are of the form A–A–A–A–A–A–B–B–B–B–B, and a triblock copolymer refers to the form A–A–A–A–B–B–B–B–C–C–C–C, or, more commonly, A–A–A–A–B–B–B–B–B–A–A–A–A. Alternating copolymers are of the form A–B–A–B–A–B–A. These copolymers may be considered homopolymers with an A–B repeat unit. Random copolymers have at least two chemical groups repeating in random

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

order. A graft copolymer is composed of a homopolymer backbone and another repeat unit as a branch. (Grafts will also be discussed as branches.) Telechelic polymers are made with reactive groups on both ends of a linear chain to be used in subsequent reactive processes (see thermoset polymers, below). Polymers with rigid backbones form liquid-crystalline phases, and these are often called mesophase polymers. Ionomers are polymers with pendant groups that can be ionized in solutions.

Polymers may be named by the *number* of repeat units. For example, the oligomers of poly(methylene) are ethane, propane, butane, and the higher alkanes. Sometimes oligomers are given different common names than their polymers. For example, the oligomer of poly(oxyethylene) is called poly(ethylene glycol) while the polymer is called poly(ethylene oxide). Telechelic polymers are usually oligomers. A monomer is the starting material from which a polymer is made. Ethylene is the monomer for polyethylene. Some authors will refer to a repeat unit in a polymer as a monomer.

We may also name polymers by their macrotopology or *architecture.* Chains can have a linear, branched, ladder, or ring topology. Most branching occurs with random lengths with random origins along the backbone. A graft copolymer is a chain with one homopolymer for the backbone and another for the branch. Chains branched along a backbone with equally spaced branches of equal length are called combs. Stars are chains in which all the branching starts at one central point. Dendrimers are polymers in which branching occurs with a specified branch length and with a specified multiplicity of branching at each branch point. Large dendrimers approach spherical shape. Lattice animals are chains that have a finite probability of branching at each repeat unit and thus show the structure of Cayley trees. As the branching density increases, branches will themselves have branches on them. Finally a point will be reached, the *gel point,* where the entire structure acts as one molecule. These gels or *cross-linked polymers* can either be created by a chemical reaction of monomers conducive to branching or be produced by adding cross-linking agents into normally uncross-linked polymers. The formation of epoxies is an example of the former; the vulcanization of rubber, an example of the latter. Many properties of polymers are dramatically affected by the character of their branching.

Polymers are often named by their method of *chemical synthesis.* Polymers derived from natural substances are called natural polymers, e.g., cellulose or xanthate gums. Many synthetic polymers are made by addition polymerization or condensation polymerization. They are often referred to as a class by their method of synthesis as addition or condensation polymers. Often the mode of preparation indicates something about the expected molecular-weight distribution of the polymer. In addition polymerization, the repeat unit in the polymer and the reacting monomer used to create the polymer have the same chemical composition. This often produces polymers with carbon backbones and pendant substituent groups. Examples include polyethylene, polypropylene, and polystyrene. Poly(ethylene oxide) is an example of an addition polymer without a purely carbon backbone. Addition polymers can be made in reactions that involve free radicals or ionic groups. Generally polymers that involve these ionic groups in their reaction processes can be made with narrow molecular-weight distributions. In general, condensation polymers have repeat units resulting from the reaction of two different monomers. The resulting repeat unit is different from either monomer. The backbone of these polymers often contains noncarbon atoms (like oxygen or nitrogen) at regular intervals. Polyesters, polyamides, and polycarbonates are examples in this category.

Polymers are also named by the *processing conditions* used to form them into their final object. In thermoset polymers, a reaction occurs during the forming process that often causes the polymer to cross-link. Thermoset polymers cannot usually be reprocessed. Telechelic polymers are the oligomers with reactive ends often used as a starting material for thermoset polymers. Examples of thermoset polymers are polyurethanes and epoxies. Almost no chemistry occurs during the processing of a thermoplastic material. Thermoplastic polymers become solid when cooled from the melt because of a phase transition, e.g., crystallization, or by supercooling the liquid into a glass. Common thermoplastic polymers are polyethylene, polystyrene, and poly(ethylene terephthalate). Thermoplastic polymers can be reprocessed.

Polymers made from repeat units that contain an *asymmetric* carbon atom may have variations in chemical structure because of the polymerization process. For example, the repeat unit $-CH_2CHx-$, where x represents an atom or chemical group other than hydrogen, may polymerize head-to-head or head-to-tail, where "head" refers to the CH_2 group and "tail" refers to CHx. In addition, there are two possible stereoregular arrangements: one in which the arrangements of groups about each asymmetric carbon are the same, isotactic, or one in which the arrangements alternate, syndiotactic. Another possibility is a random distribution, atactic. The stereochemical regularity affects the physical properties of the polymer, mainly through the ability to crystallize.

Finally, the polymer is known by its chemical name and *chemical composition.* In Table 1, we give the common chemical name of some common polymers and the chemical structure of the repeat unit. The common chemical name is often derived from the name of the chemical species the polymer is made from, the monomer. For example, polymethylene, $CH_3-(CH_2)_n-CH_3$ is commonly called polyethylene since it is produced, mainly, from ethylene as a monomer.

2. METHODS TO DETERMINE OVERALL CHEMICAL COMPOSITION

The chemical composition of a polymer is specified by the constituent atoms and their chemical-bonding arrangement. Many analytical techniques are available for elucidating the chemical composition. Selection of appropriate methods depends on several factors, including amount and form of the specimen, surface or bulk characterization, and whether information about structure beyond the molecular level is desired. Elemental-analysis methods yield little beyond atomic

Table 1. Common chemical names of some polymers and their chemical structure.

Common chemical name	Chemical structure
Polyethylene	$-(CH_2CH_2)_n-$
Polypropylene	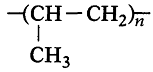 $-(\underset{\displaystyle CH_3}{\underset{\vert}{CH}}-CH_2)_n-$
Polyisobutylene	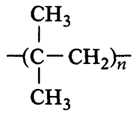 $-(\overset{\displaystyle CH_3}{\overset{\vert}{\underset{\displaystyle CH_3}{\underset{\vert}{C}}}}-CH_2)_n-$
Polyisoprene	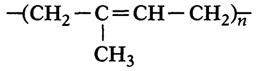 $-(CH_2-\underset{\displaystyle CH_3}{\underset{\vert}{C}}=CH-CH_2)_n-$
Polybutadiene	$-(CH_2-CH=CH-CH_2)_n-$
Polystyrene	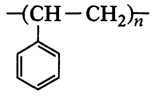
Polyacrylonitrile	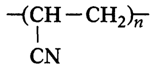 $-(\underset{\displaystyle CN}{\underset{\vert}{CH}}-CH_2)_n-$
Poly(vinyl alcohol)	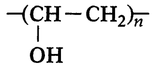 $-(\underset{\displaystyle OH}{\underset{\vert}{CH}}-CH_2)_n-$
Poly(vinyl chloride)	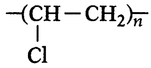 $-(\underset{\displaystyle Cl}{\underset{\vert}{CH}}-CH_2)_n-$
Poly(vinyl acetate)	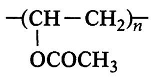 $-(\underset{\displaystyle OCOCH_3}{\underset{\vert}{CH}}-CH_2)_n-$

Table 1. Continued.

Common chemical name	Chemical structure
Poly(vinylidene fluoride)	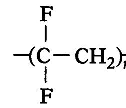 $-(\overset{F}{\underset{F}{C}}-CH_2)_n-$
Poly(methyl acrylate)	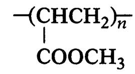 $-(\underset{COOCH_3}{CH}CH_2)_n-$
Poly(methyl methacrylate)	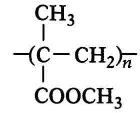 $-(\overset{CH_3}{\underset{COOCH_3}{C}}-CH_2)_n-$
Poly(formaldehyde)	$-(OCH_2)_n-$
Poly(ethylene oxide)	$-(OCH_2CH_2)_n-$
Poly(phenylene oxide)	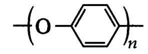
Poly(ethylene terphthalate)	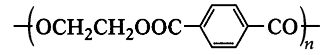
Poly(hexamethylene adipate)	$-(NH(CH_2)_6NHCO(CH_2)_4CO)_n-$
Poly(tetramethylene hexamethylene urethane)	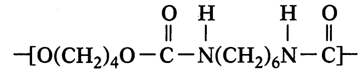 $-[O(CH_2)_4O-\overset{O}{\overset{\|}{C}}-\overset{H}{N}(CH_2)_6\overset{H}{N}-\overset{O}{\overset{\|}{C}}]-$
Poly(carbonate)	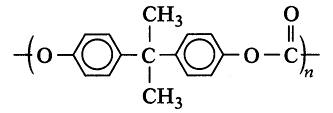
Poly(tetrafluoroethylene)	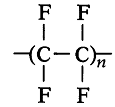 $-(\overset{F}{\underset{F}{C}}-\overset{F}{\underset{F}{C}})_n-$

composition. The constituent atoms of most commercial and naturally occurring polymers have low atomic numbers that influence the choice of techniques for elemental composition. The principal techniques for chemical composition are nuclear magnetic resonance (NMR) and infrared (IR) spectroscopies. These techniques not only are able to differentiate among the many chemical groups of polymers, but also are quantitative. In addition, these techniques have the advantages of firm theoretical foundations for predicting spectra and, as a result of their wide application, availability of spectral databases accessible by computer-aided search routines for identification of unknowns.

NMR measures transitions between nuclear-spin states of the constituent atoms in the presence of an applied magnetic field. The applied magnetic field at a particular atom is perturbed by the neighboring atoms that provide the sensitivity of the method to chemical bonding. Infrared spectroscopy determines the frequencies of the vibrations of the constituent atoms in the potential field of the chemical bonds. Raman spectroscopy is another vibrational method that complements infrared spectroscopy in several respects, but because of instrumental difficulties it is not as widely used as IR. Under favorable conditions, both NMR and vibrational spectroscopies can be used to determine not only the types of chemical groups but also their number.

2.1 Nuclear Magnetic Resonance Spectroscopy

Nuclear magnetic resonance spectroscopy (see MAGNETIC RESONANCE AND QUADRUPOLE RESONANCE, NUCLEAR) is widely used by organic chemists to deduce chemical configurations. For modest molecular weights, less than 500 g/mole, the chemical structure can be determined on samples in the liquid or solution forms. The high molecular weight of polymers prohibits detailed characterizations, but appreciable information concerning the repeat unit, or units in the case of copolymers and polymer blends, can be obtained provided the polymer is in solution, the melt phase, or a solid phase in which motional averaging within the time frame of the NMR measurement is allowed. The information derived includes stereochemistry and content of head-to-tail, head-to-head, or tail-to-tail arrangements. In many cases, it is desirable to characterize polymers in the solid state. Ordinarily, this would lead to exceedingly broad NMR lines that obscure the chemical shifts that reveal details of the structure. Fortunately, the techniques of proton decoupling have been developed to produce narrow lines and improve resolution and sensitivity in solid-state NMR.

Although many nuclei produce NMR signals, the most important in polymer characterization are the proton, fluorine, and carbon-13. Proton nuclear magnetic resonance spectroscopy is a powerful method for determining the structure and dynamics of organic molecules in solution (Bovey *et al.*, 1983). The basis of the technique is resonant transitions between energy levels of nuclear spins in the presence of applied magnetic fields. As the energy levels are sensitive to the local electronic environment of the nucleus, detailed quantitative information is obtained from the transition frequencies (chemical shifts) and intensities. The extremely narrow NMR spectral lines (approximately a few hertz) are observed for solutions because of rapid molecular motions that average interactions. As a result of the narrow spectral bands, features attributed to sequences of chemical groups are revealed. By taking advantage of the high resolution capability of NMR when applied to liquid samples, aspects of polymer structure, stereochemical configuration, branching, copolymer sequences, and molecular dynamics are elucidated.

The method may be adapted to nuclei other than protons. Fluorine NMR has been applied to fluorine-containing polymers (Bovey *et al.*, 1983); and with developments in fast Fourier transform spectroscopy and co-adding signals to improve sensitivity, NMR based on the carbon-13 (^{13}C) nucleus at natural abundance is now possible. More details concerning polymer structure are obtainable by ^{13}C NMR because of the greater chemical shifts of ^{13}C. In cases in which it is possible to dissolve the polymer, chemical microstructure, including stereochemistry, branching, and end groups, may be probed by ^{13}C NMR. Figure 1 illustrates the power

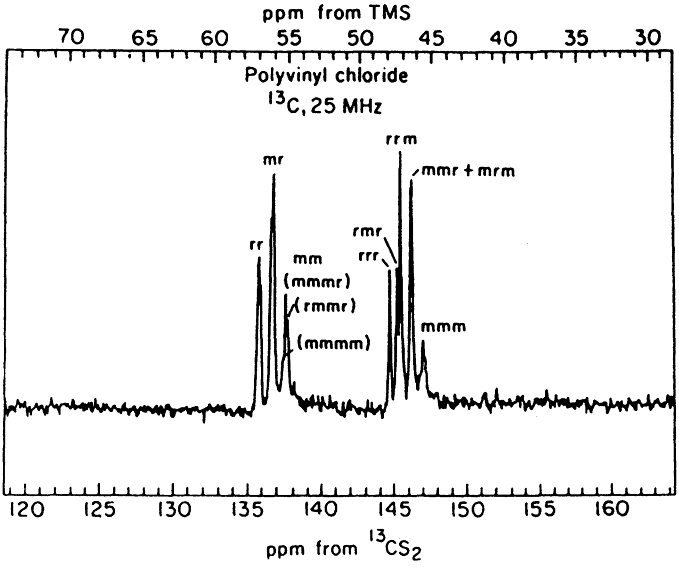

FIG. 1. 25-MHz ^{13}C NMR solution spectrum of poly(vinyl chloride). The individual resonances are labeled with the sequence of meso (m) and racemic (r) dyads (reproduced from Ivin, 1976).

of ^{13}C NMR for analyzing the stereochemistry of poly(vinyl chloride). In this figure, the cluster of NMR lines ca. 47 ppm arise from the CH_2 carbon and the lines ca. 56 ppm from the CHCl carbon. The individual NMR lines in each group are identified with specific arrangements of the chemical repeat units, $-CH_2CHCl-$. The halogen atoms of neighboring repeat units may occupy positions that are on the same side, meso (m), or opposite side, racemic (r); these two arrangements yield different ^{13}C chemical shifts that are further influenced by the next-nearest neighbors. Monomer sequences of lengths up to pentamers exhibit distinct chemical shifts.

Another application of ^{13}C NMR is determination of branch types, distributions, and concentrations. An example of quantitative information that can be obtained is illustrated in Fig. 2 for a low-density polyethylene. Resonances from methylene carbons at least four carbons removed from a branch point, branch, or chain end contribute to the intense line at 30 ppm. The remaining lines are attributable to a branch point, chain, or branch-end carbon, or carbons situated within three carbons from a branch point, chain, or branch-end carbon. In this figure, the assignments of chemical shifts to the different kinds of branches are given where Am = amyl, Bu = butyl, and Et = ethyl.

Thus far, the discussion has focused on polymers in solution; however, for various reasons it may be desirable to characterize the material in the solid state. The solid-state NMR method (Havens and Koenig, 1983) has the advantages that no special sample preparation is necessary, insoluble solids may be evaluated, and any changes in the sample due to sample preparation are avoided. Generally, NMR of solids is characterized by broad distributions of resonance frequencies that obscure the small chemical shifts from which much information is obtained on liquid or solubilized polymers. The linewidths of solid specimens arise from strong dipole–dipole interactions and the absence of motional averaging in the time scale of the NMR measurement. Linewidths of solid specimens recorded under conditions typical of liquid-state spectra may be 20 kHz or more.

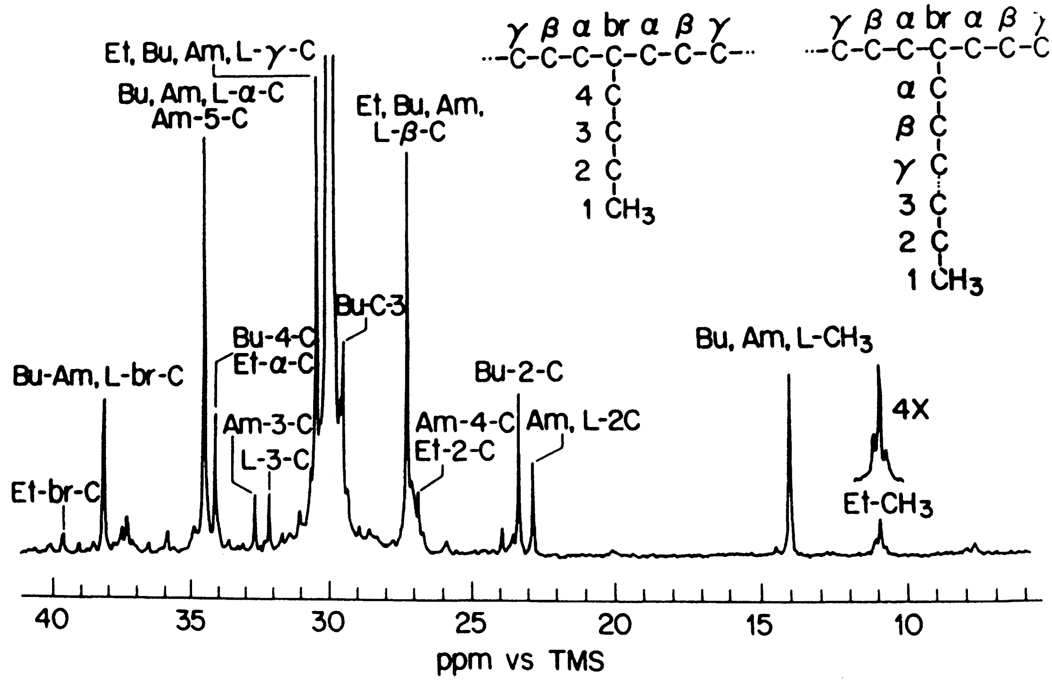

FIG. 2. ^{13}C NMR spectrum of low-density polyethylene, The individual resonances are labeled by the branch type (Am = amyl, Bu = butyl, and Et = ethyl). (Reproduced from Bovey *et al.*, 1983.)

High-resolution techniques have been implemented to eliminate or reduce solid-state line broadening, thereby achieving ^{13}C spectra of solid materials whose bandwidth approaches that typical of the liquid state. High-resolution ^{13}C NMR relies on high-power proton decoupling that practically eliminates dipolar broadening by protons and high-speed sample spinning to average out chemical-shift anisotropy of the ^{13}C nucleus. The effect of these two techniques on the NMR spectrum is to reduce the broad band typical of solids to narrow bands that reveal information typical of liquid-state samples.

One additional technique is used to improve sensitivity in solid-state NMR. The low natural abundance of ^{13}C (about 1%) means that NMR signals are inherently weak, and furthermore, because of the long spin-lattice relaxation times of the ^{13}C nucleus, substantial improvement in sensitivity by co-adding pulses requires excessively long measurement times. However, the time between pulses in the Fourier transform experiment can be shortened by a technique termed cross-polarization. In this method, the protons are used as a source of magnetization to transfer polarization from the protons to the carbons. The greater abundance of the protons allows the system to reach equilibrium more quickly.

Through incorporation of these techniques, high-resolution solid-state NMR has become an important tool for characterization of solid polymers. Figure 3 demonstrates the capabilities for analysis of epoxy-based composites. A number of distinct lines are evident in Fig. 3, including features assigned to the carboxyl carbon of the reactive species, labeled 4 in the figure, and the same carbon after reaction, labeled 4′. The degree of conversion of the reactant to polymer is determined from the intensity ratio of these two lines. The comparison of the two spectra in Fig. 3 reveals that the thermally polymerized resin achieved a higher conversion to polymer than the photo-cured resin.

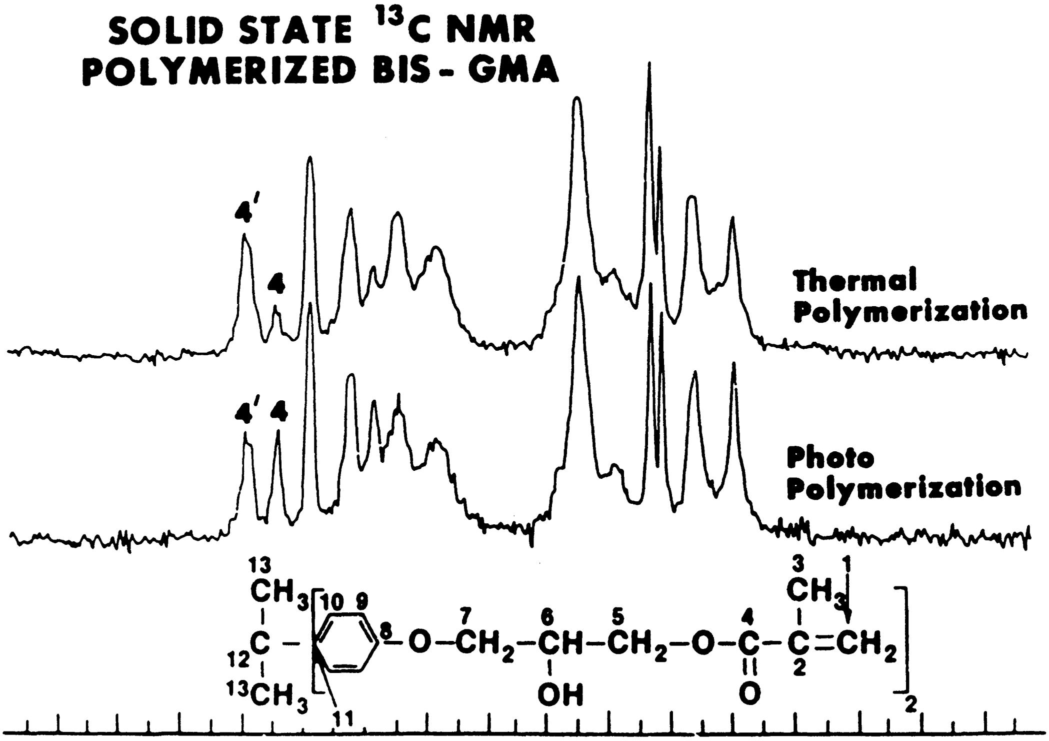

FIG. 3. ^{13}C NMR spectra of vinyl ester polymer matrix composite resin in the solid state. The two spectra compare polymerization by thermal activation and by uv light. Resonances attributed to reactive chemical species are identified.

2.2 Infrared Spectroscopy

Infrared spectroscopy is a widely used technique of polymer identification for which a large number of reference spectra of polymers, monomers, and additives are available. When combined with computer-aided search routines, this technique allows for rapid identification of unknown samples. Infrared spectroscopy is a type of vibrational spectroscopy that measures the absorption or reflection of light associated with the vibrations of the atoms in the potential field defined by the chemical bonding arrangements. The vibrational frequencies are determined by the potential field and the masses of the atoms involved in the vibrations. An isolated polymer containing N atoms will have $3N - 6$ vibrational modes, and in principle, the infrared spectrum would contain a like number of bands. However, for symmetry reasons some vibrations will be inactive in the infrared while others will have low probabilities of absorption. In addition, the vibrations tend to be localized over several atoms rather than over the entire molecule, and as a consequence, the infrared spectrum is more representative in terms of the number of bands and their frequencies to the composition of the chemical repeat unit. Nonetheless, as with infrared spectroscopy of small molecules, the infrared intensities are proportional to the number of absorbing species and the probability of absorption (Beer's law).

The IR spectra of polymers reveal details of not only chemical composition, but also of physical structure. Qualitative and quantitative information can be obtained on chemical composition of polymers, copolymers, polymer blends, composites, additives, and compounding ingredients. IR spectral frequencies and intensities are also sensitive to chemical order, i.e., stereochemistry, branching, end groups, and degradation products. The IR spectra of solid polymers provide information on the type and amount of crystalline forms, the degree of crystallinity, orientation of macromolecules, chain conformations, and the local environment of constituent chemical groups. IR spectroscopy is by far the most widely used method for characterization through vibrational properties although both inelastic neutron scattering and Raman spectroscopy have application to polymers.

Low sensitivity and inflexibility in sample type had been problems with infrared spectroscopy that have now been overcome with Fourier transform infrared spectroscopy (FTIR). Both of these problems relate to low sensitivity of conventional dispersive IR. The several hundredfold increase in signal-to-noise ratio (sensitivity) made possible with FTIR has greatly expanded the role of infrared spectroscopy for microsampling and nontransmission types of sampling. Commercially available accessories exist for specular reflection, diffuse specular reflection, diffuse reflectance sampling, and photoacoustic detection in addition to the surface-probe technique of attenuated total reflection. A wide variety of specimen types from particles to intractable solids can be examined by these techniques. All of these methods probe the surface to a large degree. Specular and diffuse specular reflection measure the index of refraction as a function of infrared wavelength. The reflection and absorption spectra are related through the Kramers–Kronig relationship. Diffuse reflection (DRF or DRIFT) is a combination of transmission and reflection of radiation by the sample. This technique is ideally suited to microsampling with spectra obtainable from a nanogram of material. Reflectance techniques are sensitive to the state of the solid specimens—different preparation methods may yield spectra in which the relative band intensities vary. This spectral sensitivity, however, may afford advantages of greater specificity in identification.

Photoacoustic spectroscopy (*q.v.*) (Carter, 1992) differs from the above-mentioned sampling techniques in that IR absorption is detected by the acoustic signals produced by the heat generated by the IR-absorbed radiation. The unusual feature of this detection method in contrast to transmission approaches is that the signal increases with absorption, rather than decreasing as is the common observation. Photoacoustic detection is suitable for studying intractable solids, specimens that cannot be prepared as thin films without destroying identifiable features.

Incorporation of a microscope into FTIR instrumentation and the use of optic-fiber waveguides for remote excitation and detection are current features that expand the type of samples and sampling arrangements that may be amenable to infrared studies.

3. MOLECULAR-WEIGHT DISTRIBUTIONS AND THEIR DETERMINATION

In contrast to small molecules that may be processed to high purity, synthetic polymers almost never exist as single–molecular-weight entities. Exceptions are oligomers of synthetic polymers and naturally occurring biopolymers. For example, normal hexatriacontane, n-$C_{36}H_{74}$, a paraffin that is an oligomer of polyethylene, may be obtained in a highly pure form from which large single crystals may be prepared. Proteins are biopolymers that occur with a single molecular weight. The molecular weights of synthetic polymers are characterized by a distribution, rather than a single value. This molecular-weight distribution arises from the chemistry of producing the polymer or, in commercial cases, is a result of the blending of the same homopolymers to produce particular final properties of the plastic piece or to obtain specific processing properties for the polymer.

3.1 Naming Molecular-Weight Distributions and Their Moments

If n_i is the number of polymer molecules with molecular weight m_i, then the total mass of molecules with molecular weight m_i is $w_i = m_i n_i$. The number molecular-weight distribution is the fraction of molecules at a given molecular weight and is given by $n_i/\Sigma n_i$. The weight molecular-weight distribution is the fraction of mass at a given molecular weight and is given by $w_i/\Sigma w_i$. Many methods to determine the molecular weight actually measure moments of the molecular-weight distribution, not the total distribution. We may define some important moments of the molecular-weight distribution as follows: the number average moment M_n,

$$M_n = \sum w_i \Big/ \sum n_i;$$

the weight average moment M_w,

$$M_w = \sum m_i w_i \Big/ \sum w_i;$$

the z-average moment M_z,

$$M_z = \sum m_i^2 w_i \Big/ \sum m_i w_i;$$

the z + 1-average moment M_{z+1},

$$M_{z+1} = \sum m_i^3 w_i \Big/ \sum m_i^2 w_i;$$

and the viscosity average molecular weight M_v,

$$M_v = \sum \left(w_i m_i^a \Big/ \sum w_i \right)^{1/a},$$

where a is the power coefficient in the Mark–Kuhn–Houwink–Sakurada equation discussed later (Sec. 3.2.6).

The ratio M_w/M_n is called the polydispersity index and is often used as a measure of the breadth of the molecular-weight distribution. Depending on the available methods of synthesis, polymers can have wide or narrow distributions. For example, polystyrene prepared by anionic polymerization can be obtained with a polydispersity index of as little as 1.05. Polymers made by condensation polymerization are normally expected to have polydispersity indexes near 2. Commercial polyethylenes often have M_w/M_n ratios as high as 30.

3.2 Methods to Determine Moments of the Molecular-Weight Distribution

3.2.1 Membrane Osmometry Osmometry is one of the many colligative methods that count the number of molecules in solution. Colligative methods like melting point depression, boiling point elevation, and osmometry determine the number average molecular weight of molecules by measuring the change in a thermodynamic property of the solvent as a result of the addition of a small number of solute to a solvent. The change in the thermodynamic property arises from the entropy of mixing resulting upon the addition of the solute molecules to the solvent. In osmometry, the osmotic pressure of a dilute solution containing polymers, π_s, is compared with that of a pure solvent, π_0. A semipermeable membrane separates the two solutions. Solvent can go through the membrane while polymers cannot. See Fig. 4

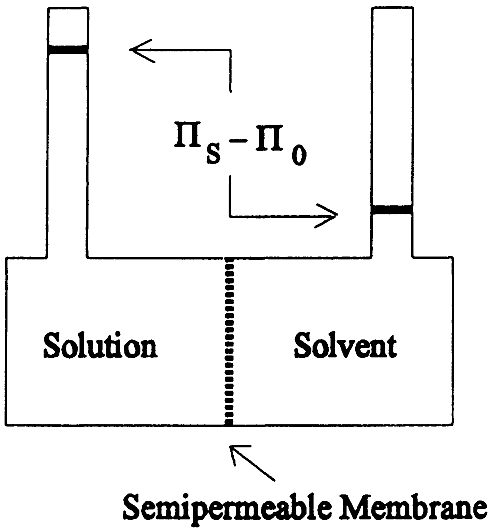

FIG. 4. Schematic of membrane osmometer. π_s is the osmotic pressure of the solution at concentration c. π_0 is the osmotic pressure of the solvent.

for a schematic of the experiment. As a result, one can measure the osmotic-pressure differential between the solution and solvent.

This measurement is sensitive enough to measure molecular weights of a polymer up to about 100 000 g/mole with good precision. It has a lower limit of approximately 8000 g/mole. This lower limit arises from the fact that the membranes available for the measurement allow polymers smaller than about 8000 g/mole through the membrane. Colligative properties can measure the molecular weight of any polymer, homopolymer, copolymers, etc., as long as the polymer can be dissolved in a dilute solution. These measurements can be made without additional measurements or assumptions about the chemical composition of the polymer or its architecture. The osmotic-pressure measurement yields the M_n moment of the molecular-weight distribution as

$$\lim_{c \to 0}(\pi_s - \pi_0)/(RTc) = 1/M_n + A_2 c + \cdots,$$

where R is the gas constant, T is the absolute temperature, c is the concentration of polymer, and A_2 is the second virial coefficient. The second virial coefficient measures the interaction between two polymers in dilute solution. To make an accurate extrapolation of the above equation to the limit of zero concentration, one often makes the molecular-weight measurement where A_2 is near zero. The temperature and solvent strength (for a mixed solvent) where A_2 is zero is called the theta point.

Membrane osmometry is one of the molecular-weight methods considered an absolute method. One needs only measure pressure, mass, and volume (for concentration) to obtain a molecular weight. No calibration against polymers of known molecular weight is required.

3.2.2 Vapor-Phase Osmometry Vapor-pressure osmometry is a molecular-weight method using the change in the vapor pressure between the polymer solution and the solvent. This could be an absolute method but to get adequate sensitivity, a droplet of solution is placed on a thermistor (a temperature-measuring device) and a droplet of solvent on another temperature-measuring device. The cooling-rate difference between them is then measured. The instrument has to be calibrated with polymers of known molecular weight. One again measures M_n. Because of the lack of sensitivity of the method, the measurement is limited to polymers of low molecular weights.

3.2.3 Light Scattering Light scattered from a solution arises from two sources, scattering from density fluctuations in the liquid or scattering from refractive-index differences arising from concentration fluctuations of solutes in the solvent. The density fluctuations can be shown to arise from the isothermal compressibility $RT(\delta\rho/\delta p)$ where p is the pressure and ρ is the solution density. The concentration fluctuations arise from the osmotic compressibility $RT(\delta c/\delta\pi_s)$ where π_s is the osmotic pressure of the polymer solution and c is the concentration. For dilute solutions away from the solvent critical point, the density fluctuations are identical to those of the pure solvent. Thus the Rayleigh ratio R_θ is related to the osmotic compressibility through

$$\lim_{\theta \to 0} Kc/R_\theta = (1/RT)(\delta\pi_s/\delta c),$$

where R_θ is the ratio of the intensity difference between solution and solvent at angle θ to the intensity of the primary beam all multiplied by the squared distance to the detector (see schematic in Fig. 5), and

$$K = (16\pi/N_A\lambda_0^4)n_0^2(dn_s/dc)^2,$$

where λ_0 is the wavelength of light, N_A is Avogadro's number, n_0 is the refractive index of the solvent, and dn_s/dc is the refractive-index increment of the polymer solution.

For molecules much smaller than $\lambda_0/20$, the angular dependence of R_θ is related to the change in scattering volume with angle. For larger molecules, light scattered from one region of the molecule interferes with light scattered from another region of the molecule causing the destruction of scattering intensity. This angular dependence is sensitive to the size and shape of the molecule. At low concentrations of the polymer, an expansion in concentration of R_θ is given by

$$\begin{aligned} Kc/R_\theta &= 1/M_w + 2A_2c \\ &\quad + 16\pi^2R_g^2/(3\lambda_0^2)\sin^2(\theta/2) + \cdots, \end{aligned}$$

where A_2 is, as before, the second virial coefficient and R_g is the radius of gyration. The term R_g is defined for homopolymers by

$$R_g^2 = \sum \langle r_{ij}^2\rangle/n^2,$$

where r_{ij} is the distance between scattering center i and scattering center j and $\langle\ \rangle$ is a configurational average. It is a measure of the size of the molecule. Light-scattering data are normally represented by a Zimm plot (Zimm, 1948) (see Fig. 6) where the molecular weight is obtained by a double extrapolation to zero concentration and zero angle.

In the measurement of molecular weight, it is often useful to use a solvent in which A_2 is near zero to facilitate extrapolation of the molecular-weight measurement to the zero-concentration limit.

Light scattering is most easily done on homopolymers. To obtain the M_w of a copolymer by light scattering requires the measurement of the light scattering in at least two solvents with considerably different refractive indexes.

Light scattering is another molecular-weight method which is absolute. No calibration against a polymer of known molecular weight is needed to obtain the molecular weight of a homopolymer.

3.2.4 End-Group Analysis The end groups of a polymer are chemically different from any interior group. The molecular weight of a homopolymer can be obtained by using any measurement technique that will count the number of end groups per mass of the polymer. Besides the spectroscopic methods to distinguish end groups, like IR and NMR described in earlier sections, chemical methods are often available. For example, polyamides (nylons) can be synthesized with amine or carboxylic acid

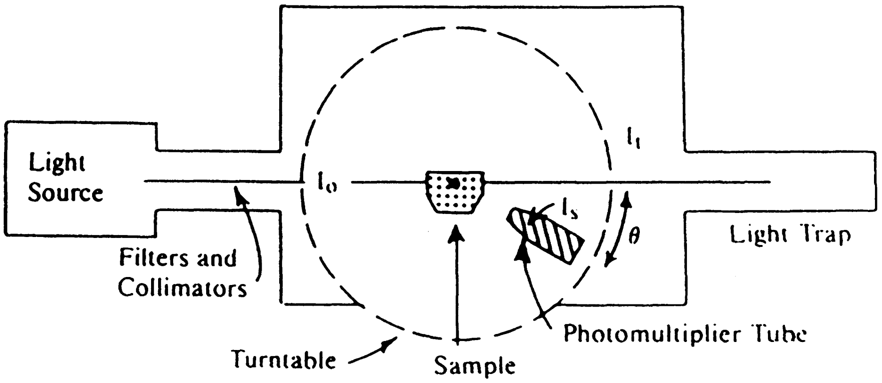

FIG. 5. Schematic of light-scattering instrument. I_0 is the intensity of the source, $I_s(\theta)$ is the scattered light from either the solvent, $I_{\text{solv}}(\theta)$, or the solution, $I_{\text{soln}}(\theta)$, at an angle θ, θ is the angle between the axis of the light source and axis of the detector, and

$$R_\theta = (I_{\text{soln}}(\theta) - I_{\text{solv}}(\theta))r^2/I_0,$$

where r is the distance from the center of the sample to the detector (from Hiemenz, 1984, p. 691).

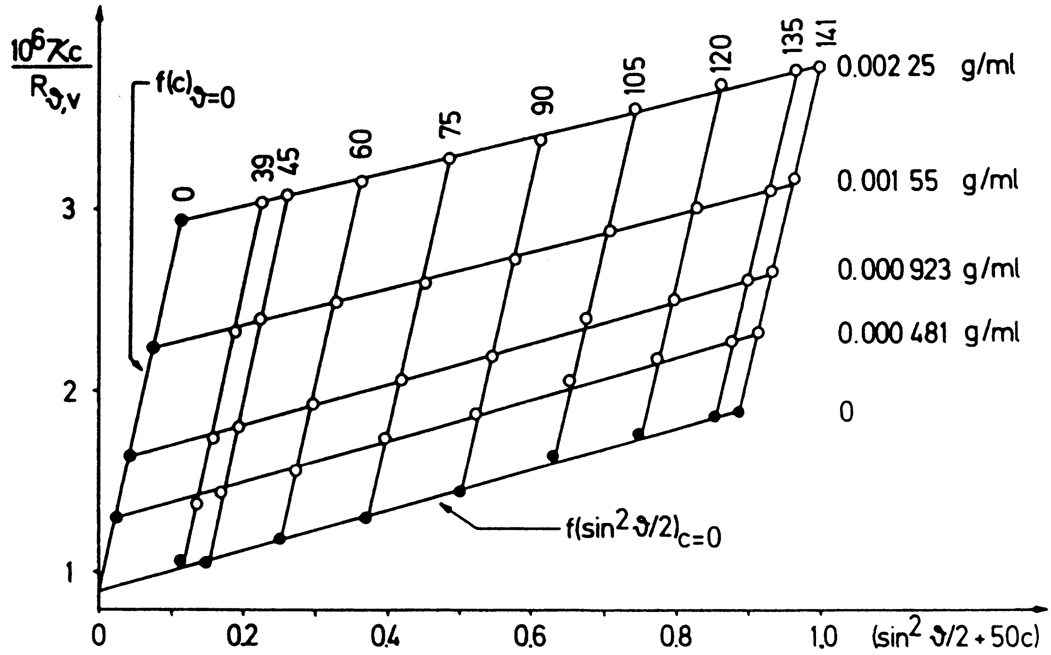

FIG. 6. Zimm plot of reciprocal intensity for poly(vinyl acetate) in butanone at 25 °C. The molecular weight is computed from the intercept (from Elias, 1977, p. 323).

end groups. These can then be counted by normal analytical methods. Assuming the method counts all end groups and the polymer is not branched, then the method yields M_n. The method is absolute under these conditions.

These conditions are seldom met unless special care is taken in the polymerization process. In many commercial processes, the initiation and termination chemical reactions that produce a polymer are often not the same, and so the same chemical groups are not on each end. Impurities often terminate the chemical reaction so that unknown groups could be on at least one end.

The sensitivity of measuring a small number of end groups in the presence of a large number of other groups is constant, independent of the polymer molecular weight. Thus, the method is limited to the determination of polymers of low molecular weight. Under the best of circumstances, the molecular weights of polymers of less than 20 000 g/mole can be measured with this method.

3.2.5 Ultracentrifugation This method is mainly used in molecular-weight studies of biological macromolecules, although it has often been used to determine the molecular weight of synthetic polymers. The equilibrium sedimentation methods of ultracentrifugation weigh the polymer molecules by obtaining an equilibrium concentration gradient within a solution sample. The gradient is a result of a balance between the force of the controlled centrifugal field on the masses of the molecules and the opposing force of the chemical potential of the molecules to redistribute uniformly in the solution (the diffusion potential). In the gradient, the lower–molecular-weight molecules move toward the centrifugal top (the meniscus) and the heavier molecules move toward the centrifugal bottom. Since this technique measures the weight of the polymer molecules, ultracentrifugation determines most accurately the first moment of the mass distribution or M_w of the polymer.

The equilibrium continuous partial fractionation of the polymer between the solution boundaries offers the opportunity to analyze the polymer molecular-weight distribution. Few successful efforts have been made to analyze molecular-weight distribution using this technique. However, the moments of the distribution curve higher than

M_w, i.e., M_z, M_{z+1}, etc., can be measured. In addition, the M_n can also be determined by an equilibrium sedimentation experiment conducted under conditions in which the equilibrium polymer concentration becomes approximately zero in the meniscus region of the solution sample. The uncertainty in M_n and M_z is much greater than the uncertainty in M_w determined by this experimental procedure.

Like light scattering and osmometry, equilibrium sedimentation is an absolute method. In addition to it being an absolute method, the other major advantage of the equilibrium sedimentation procedure is that it determines the M_w more precisely than any other method. Furthermore, practically useful values for M_n, M_z, and M_{z+1} can also be determined in the same single experiment. The singular major problem with the equilibrium sedimentation method is the time, often a week, required to achieve equilibrium.

3.2.6 Intrinsic Viscosity The viscosity of a dilute solution of a polymer is greater than that of the solvent. Usually a small concentration of a high-molecular-weight polymer will increase the viscosity of the solution greatly. The greater the size of the polymer, the more greatly the viscosity increases. Thus before other methods were available to measure the molecular weight of the polymer, viscosity–molecular-weight relations were developed.

The intrinsic viscosity is defined as the limiting slope of the viscosity concentration function. Thus

$$[\eta] = \lim_{c \to 0}(\eta_s - \eta_0)/c,$$

where η_s is the viscosity of the solute at the concentration c in the solution, and η_0 is the viscosity of the solvent. The relationship of intrinsic viscosity to molecular weight is called the Mark–Kuhn–Houwink–Sakurada relationship and is experimentally

$$[\eta] = K_\eta M^a,$$

where the value of K_η depends on the polymer and the solvent and the value of a depends on the solvent quality only. The value of a varies from 0.5 for a chain near its theta point to 0.8 or so for a chain in a good solvent. The Mark–Kuhn–Houwink–Sakurada equation is valid for higher molecular weights. For example, the equation holds above a molecular weight of 10 000 g/mole for polystyrene in tetrahydrofuran, a common polymer solvent for size-exclusion chromatography.

3.3 Methods to Determine Molecular-Weight Distributions

3.3.1 Size-Exclusion Chromatography Size-exclusion chromatography (SEC) is one form of high-performance liquid chromatography (see CHROMATOGRAPHY). SEC is often called gel permeation chromatography (GPC) or gel filtration chromatography. The column packings for nonaqueous SEC are usually porous gel beads of styrene divinyl-benzene or porous glass beads.

In any liquid chromatographic process using a porous medium as the chromatographic bed, the average elution volume in chromatography reflects the partitioning of the polymer between the flowing "mobile phase" and the "stationary phase." In the stationary phase, the polymer is localized in the bead pore. In the simplest quasiequilibrium model of chromatography, the average elution volume V_e is given by

$$V_e = V_m + K_D V_p,$$

where V_m is the mobile phase volume, V_p is the pore volume, and K_D is the partition coefficient. For size-exclusion chromatography, K_D is less than 1.

In size-exclusion chromatography, the polymer has no interaction with the pore wall except to be excluded by it. The polymers that are largest are excluded from the pore (they cannot get in) while the smaller polymers can get into the pores. Thus the largest polymers can only sample the mobile phase V_m. The larger polymers have K_D near zero and the smaller polymers have K_D closer to 1.

Since the separation process is on the basis of molecular size, not molecular weight, the method requires calibration. Furthermore, since the size of a very high–molecular-weight branched polymer can be the same as a much lower–molecular-weight linear polymer, the molecular weight can only be determined by a calibration with polymers of similar repeat units and architecture.

The method called "universal calibration" has alleviated this calibration problem for unknown polymers. In universal calibration, the volume elution from a given column is experimentally found to be a function of the hydrodynamic volume of the polymer independent of polymer architecture and chemical composition (see Fig. 7) (Grubisec *et al.*, 1967). The hydrodynamic volume is the product of the molecular weight and intrinsic viscosity, $[\eta]$. SEC detectors are available that can measure the intrinsic viscosity on line. Using universal calibration, the molecular weight and molecular-weight distribution of any unknown polymer can be measured. There are instances where universal calibration has been shown to fail.

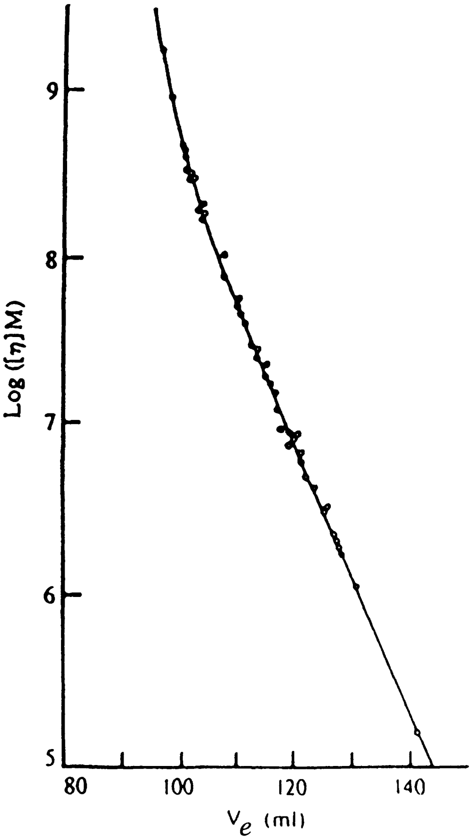

FIG. 7. Plot of $\log([\eta]M)$ vs elution volume for various polymers. The polymers in the figure include linear and branch polystyrene, poly(methyl methacrylate), poly(vinyl chloride), poly(phenyl siloxane), polybutadiene, and branched, block, and graft copolymers of styrene and methyl methacrylate (from Grubisec *et al.*, 1967).

Other detectors, ultraviolet or differential refractive-index detectors, commonly used in SEC, measure concentration of the polymer. When combined with a flow-through light-scattering detector, the molecular weight of a homopolymer at each elution volume can be measured. With this method, the molecular-weight distribution of a homopolymer can be measured.

3.3.2 Mass Spectrometry Generally mass spectrometry has been limited to measuring molecular weights for molecules of masses less than 4000 g/mole. This is due to the difficulty of obtaining the whole, not degraded, molecule in the gas phase. Recently, methods have become available to get whole polymer molecules into the gas phase.

In the most common of these, MALDI (matrix-assisted laser-desorption ionization), the polymer is dissolved in dilute solution of a solid matrix. A laser is used to oblate the matrix. The laser energy is absorbed by the matrix, not the polymer. The polymer is carried into the gas phase by the ablated matrix molecules and ionized by the transfer of charge from the matrix molecules to the polymer. The polymer once in the gas phase can then be analyzed using a variety of methods. For high-molecular-weight polymers, a time-of-flight spectrometer can measure the mass-to-charge ratio of singly or doubly charged large molecules (see, for example, Bahr *et al.*, 1992).

Mass spectrometry is the only available method to hold the possibility of determining the molecular-weight distribution directly. As pointed out earlier, SEC does not separate on the basis of molecular weight. Fractionation methods, to be discussed in the next section, often do not fractionate on the basis of molecular weight only. MALDI time-of-flight mass spectroscopy has the possibility of being an absolute method. The subject is too new to make a final judgment on this.

3.3.3 Fractionation Methods Before the appearance of SEC and mass spectrometry of polymers, most molecular-weight distributions of synthetic polymers were obtained by fractionation of the polymer followed by the measurement of some moment of the molecular weight of each fraction. If the fractionation is quantitative, the molecular-weight distribution can be expected to be quantitative.

We may fractionate a polymer from solution by using a liquid–liquid phase-separation method or a liquid–solid phase-separation method. The phase separation of a polymer solution from a single-phase system to a two-phase liquid system is little different from that which occurs with a small molecule in a solvent. As we go from a good solvent for the solute in which the solute and solvent form a single-phase liquid system to a poorer solvent by changing temperature or adding a poor or nonsolvent to the system, we obtain a two-phase system, one richer in solute than the other. The same occurs in a polymer system. However, for polymers, the temperature or solvent mixture at which the phase separation takes place is strongly molecular-weight dependent. Generally, the higher–molecular-weight species undergo phase separation at higher temperatures or better solvent strengths. Thus, we may separate polymers by molecular weight by selecting one of the two phases.

If the phase separation involves a liquid-to-solid transition, as with the formation of a crystal or precipitate, the higher–molecular-weight species usually solidifies at higher temperature or better solvent strengths.

A solid polymer can be fractionated by extracting the low–molecular-weight species at lower temperatures or poorer solvents. The polymer in its solid phase is often deposited as a film on a packed column as a substrate. Solvent strength is then varied so that each cut comes off as a function of solvent strength.

Size-exclusion chromatography has been used as a method of polymer fractionation. Generally, this technique cannot be used to obtain gram quantities of the polymer since the loading of SEC columns is very limited.

4. DETERMINATION OF CHEMICAL AND CONFORMATIONAL ORDER ALONG THE POLYMER CHAIN

Two of the methods discussed in Sec. 2, NMR and infrared spectroscopies, are utilized in determining chemical and conformational order along the polymer chain. Shielding of the applied magnetic field by neighboring atoms leads to chemical shifts in the NMR spectrum that can be related to chemical order along the polymer chain. As a result of the narrow NMR bands in liquid-state NMR, features attributed to sequences of chemical groups are revealed. Thus, aspects of polymer structure, stereochemical configuration, branching, copolymer sequences, and molecular dynamics are elucidated. An example is illustrated in Fig. 1 where ^{13}C NMR is used for analyzing the stereochemistry of poly(vinyl chloride) and poly(vinyl bromide). Besides the prevalent head-to-tail arrangement (–CH_2CHx-CH_2CHx–) (x refers to either Cl or Br), head-to-head (–$CH_2CHxCHxCH_2$–) and tail-to-tail (–$CHxCH_2CH_2CHx$–) configurations may occur as well. In addition, the halogen atoms can occupy positions that are meso (m) or racemic (r); these arrangements refer to the stereoregularity (isotactic, syndiotactic, and atactic). Sequence lengths of up to pentamers exhibit distinct chemical shifts.

Another application of ^{13}C NMR is determination of branch types, distributions, and concentrations. An example of quantitative information that can be obtained is illustrated in Fig. 2 for a low-density polyethylene. In this figure, the assignments of chemical shifts to the different kinds of branches are given.

IR spectral frequencies and intensities may also be sensitive to chemical order along the polymer chain, including stereochemistry, branching, end groups, and degradation products. The information is generally restricted to order of nearest or next-nearest neighbors of backbone atoms, however, because of the localized nature of most infrared-active vibrations. In addition, the IR spectra of solid polymers provide information on the type and amount of crystalline forms, the degree of crystallinity, orientation of macromolecules, *trans–gauche* bond ratios, and the local environment of constituent chemical groups.

5. EQUILIBRIUM DIMENSIONS AND CONFORMATION OF SINGLE CHAINS IN SOLUTION

5.1 Models without Long-Range Interactions

There are a variety of models for single-polymer chains. The simplest model, the closest model to the ideal-gas model of a gas, is the random-walk model of a polymer chain. In the simplest of these models, only the connectedness of the chain is considered. Each step in the random walk can itself be random both in length and in direction with respect to the earlier step. Thus, each step is viewed as many repeat units that may themselves have a fixed or nearly fixed bond length and perhaps have only a few rotation angles allowed. When many of these repeat units are viewed together, they appear as a random walker with random step size and random angles. This model is called the Gaussian model.

Like other forms of the random walk, only a few modifications of these models are needed to add local structure. In the freely jointed chain model, all the bonds or steps have the same length. The angle between adjacent bonds can take on any value as can the angle of rotation. In the freely rotating chain model, the bond length and the angle between successive bonds are fixed but the rotation around the bond is free.

In the most realistic of these models, the rotational isomeric state model (RIS), each step of the walk represents a repeat unit. The bond distance is fixed as is the bond angle. Only a few rotation angles are allowed. The choice of an allowed angle depends on the angle allowed in the previous few steps. Thus, the RIS describes many details of the local structure of the chain.

Equilibrium properties measured by these models include the free energy f, the average end-to-end distance squared (the average distance between the first and last segments on a linear chain), R_e^2, and the average squared radius of gyration, R_g^2. The radius of gyration, defined previously, can be measured by a light or neutron-scattering experiments.

The end-to-end distance is defined as

$$R_e^2 = \langle r_{n0}^2 \rangle,$$

where r_{n0} is the distance between the first and last beads on the chain and $\langle\ \rangle$ is the configurational average. For a long linear chain in the random walk, any squared distance measure like the end-to-end distance squared or the radius of gyration squared depends linearly on the number of steps, n. For any of the random-walk models of the polymer chain, the radius of gyration and the end-to-end distance are related as

$$R_g^2 = R_e^2/6.$$

For chains in the random-walk approximation, we find

$$R_e^2 = C_n n l^2,$$

where n is the number of monomers or repeat units in the chain, l is the bond length, often in synthetic polymers the carbon–carbon bond length, and C_n is the so-called characteristic ratio. For chains greater than 100 repeat units, C_n is independent of n.

The configurational free energy of such models, f, is for long enough chains

$$f = -kT(n - 1)\ln(z),$$

where z is the partition function per step. For the simplest lattice model of a polymer chain, $z + 1$ is the coordination number of the lattice.

5.2 Models with Long-Range Interactions

None of the above models included any interactions of chain segments separated by large numbers of other segments along the chain. These interactions are deemed long-range interactions. In three dimensions and fewer (actually in fewer than four dimensions), these interactions are important and affect the size parameters' dependence on number of segments and the functional form of the free energy.

When we include long-range interactions, the squared end-to-end distance and the radius of gyration squared in the limit of many segments have the functional form

$$R_e^2 = A_e n^{2\gamma},$$

$$R_g^2 = A_g n^{2\gamma}.$$

The value of γ depends on the nature of the long-range interactions. As with all other molecular interactions, polymer-segment interactions have two components, a repulsive core and an attractive interaction between segments. The repulsive core is usually viewed as a hard-core exclusion between polymer segments. The attractive interaction energy in solution, ϵ, is a competition between terms including polymer-segment–

polymer-segment interactions (ϵ_{pp}), polymer–solvent interactions (ϵ_{ps}), and solvent–solvent interactions (ϵ_{ss}). In lattice models, ϵ has the form

$$\epsilon = \epsilon_{ps} - (\epsilon_{ss} + \epsilon_{pp})/2.$$

In the good-solvent limit, the $\epsilon = 0$, i.e., there is no overall attractive interaction. Under these conditions, the polymer–polymer exclusion dominates and in three dimensions $\gamma = 0.588$–0.6 depending on the character of the model used. Values near $\gamma = 0.588$ are most likely correct although it is unlikely the difference can be observed experimentally from dilute-solution measurements.

(The reader should recall that in one dimension long-range interactions and short-range interactions are quite indistinguishable. In one dimension, the inclusion of local structure so that the chain cannot walk back on itself requires that the chain has $R_e^2 = A_e n^2$. That is, the chain behaves as a rod.)

In a poorer solvent where the polymer–polymer attraction is favored over polymer–solvent attraction, the polymer changes size. At the theta point, a balancing of the attraction and repulsion takes place and $\gamma = 0.5$. The competition between attraction and repulsion in the polymer makes the polymer look like a random walk. In mean field theory of Flory at this solvent strength, A_2, the second virial, vanishes. Experimental evidence supports this view in dilute solutions.

As the attraction increases, the chain collapses. The term γ takes on the value of $\frac{1}{3}$. At chain collapse of a polymer in solution, the chain is not a dense globule, but rather an open structure much larger than the condensed chain. There are a considerable number of theories and simulations that discuss chain collapse, but few experiments. The experimental nature of the transition to collapsed chain is not fully understood at this time.

Since the overall attractive energy ϵ always appears in statistical mechanics as ϵ/kT, where k is Boltzmann's constant and T is temperature, then variations in the solvent strength discussed above may be obtained by varying temperature. Another method to vary solvent strength is to use a mixed solvent. In this case, a varying solvent strength is obtained by varying concentration of good and poor solvents for the polymer. For example, cyclohexane is a poor solvent for polystyrene at room temperature with its theta point being at 34.5 °C. At higher temperature, its solvent quality improves. Benzene is a good solvent for atactic polystyrene at 34 °C. However, the mixture 74% benzene and 26% methanol, a nonsolvent for polystyrene, is a theta solvent for polystyrene at 34 °C.

The use of mixed solvents is not without problems. Often one solvent favors the polymer more than the other solvent. In dilute solution, the good solvent for the polymer may have a concentration in the vicinity of the polymer that differs from the mean solvent concentration. This situation is called preferred adsorption or solvation in polymer science. The occurrence of such preferred adsorption changes the properties of the polymer in solution.

GLOSSARY

Atactic Polymer: A polymer made from repeat units that contain an *asymmetric* carbon atom, in which the stereo arrangements of groups about the asymmetric carbons are random. Atactic polymers tend not to be able to crystallize; rather, they form glasses.

Copolymer: A polymer having a fixed number of chemical constituents or repeat units that are repeated. Diblock copolymers are of the form A–A–A–A–A–A–B–B–B–B–B, and triblock polymers are of the form A–A–A–A–B–B–B–B–C–C–C–C or, more commonly, A–A–B–B–B–B–A–A–A–A. Alternating copolymers are of the form A–B–A–B–A–B–A. These copolymers may also be considered as homopolymers of A–B repeat units. Random copolymers have at least two chemical groups repeating in random order.

End-to-End Distance, R_e: The mean square distance defined by

$$R_e^2 = \langle r_{n0}^2 \rangle,$$

where r_{n0} is the distance between the first and last beads on the chain and angular brackets denote the configurational average. For long, linear chains whose conformations are represented by random walks, the squares of the distance measures, like the end-to-end distance squared or the radius of gyration squared, depend linearly on the number of steps, n. Generally, the end-to-end distance is not directly measured; however, for linear polymers, it is related to the square of the radius of gyration according to $R_g^2 = rR_e^2$. At the theta point and for a random chain, r = $\frac{1}{6}$, independent of the number of repeat units and of the chain structure for long polymers. The constant r is about $\frac{1}{7}$ at the exclusion point. The term R_g commonly appears in theories of elasticity and of other polymer properties.

Homopolymer: A polymer having the same chemical constituent repeated many times and represented by A–A–A–A–A–A–A–A, where A denotes the chemical constituent or repeat unit.

Isotactic Polymer: A polymer made from repeat units that contain an *asymmetric* carbon atom, in which the arrangement of groups about each asymmetric carbon is identical. Isotactic polymers can crystallize.

Monomer: The chemical used in the chemical reaction to produce the polymer. For example, styrene is the monomer for homopolymer polystyrene.

Number Molecular-Weight Distribution: The fraction f_i of molecules with a mass per mole of the molecules m_i. In terms of definitions in the text,

$$f_i = n_i \Big/ \sum n_i$$

Number Average Molecular Weight, M_n: The first mass moment of the number molecular-weight distribution,

$$M_n = \sum f_i m_i,$$

where M_n is obtained from measurements of osmotic pressure, end-group analysis, or vapor-pressure osmometry on dilute solutions or from size-exclusion chromatography.

Polydispersity Index: The ratio of weight average molecular weight to number average molecular weight, M_w/M_n. It is used as a measure of the width of the molecular weight distribution and is always greater than unity.

Radius of Gyration, R_g: The mean square of the radius of gyration for a homopolymer is defined as

$$R_g^2 = \sum \langle r_{ij} \rangle / n^2,$$

where r_{ij} is the distance between repeat units i and j, the summation is over all pairs i,j, and the angular brackets denote configurational average; R_g is a measure of the size of the molecule. It is an observable directly measured by light scattering, neutron scattering, or x-ray scattering for polymers.

Second Virial Coefficient, A_2: The coefficient of the second term in the concentration expansion of the light-scattering equation or the osmotic-pressure equation. The second virial coefficient reflects the interaction between two polymers in dilute solution. The temperature or solvent strength at which $A_2 = 0$ is called the theta point.

Syndiotactic Polymer: A polymer made from repeat units that contain an *asymmetric* carbon atom, in which the arrangements of groups about successive asymmetric carbon atoms alternate. Syndiotactic polymers are known to crystallize.

Telechelic Polymer: A polymer with a chemically active end group to be used in a subsequent chemical reaction.

Thermoplastic Polymer: A polymer formed into its final product by solidification when cooled from the melt. The solidification is the result of crystallization or the supercooling from the liquid to form a glass. Common thermoplastic polymers are polystyrene, polyethylene, and poly(ethylene terephthalate). A thermoplastic polymer is susceptible to heat deformation.

Thermoset Polymer: A polymer in which chemical reactions occur during the process of forming the final product. The chemical reactions often cause the polymer to cross-link. Telechelic polymers are often used as starting materials for thermoset polymers. Thermoset polymers often cannot be reprocessed. Examples of thermoset polymers are polyurethanes and epoxies.

Theta Point: The polymer-solvent compensation point in dilute polymer solutions where the second viral coefficient A_2 is zero. At the theta point, the dimension properties have random-walk behavior.

Weight Average Molecular Weight, M_w: The ratio of the second mass moment to the first mass moment of the number molecular-weight distribution,

$$M_w = \sum m_i g_i.$$

where M_w is usually obtained from measurements of light scattering and equilibrium ultracentrifugation. The quantity g_i is defined under **Weight Molecular-Weight Distribution.**

Weight Molecular-Weight Distribution: The fraction g_i of the total mass of molecules with a mass per mole of the molecules of m_i. In terms of definitions in the text,

$$g_i = w_i \Big/ \sum w_i.$$

Works Cited

Bahr, U., Deppe, A., Karas, M., Hillenkamp, F., Glassman, U. (1992), *Anal. Chem.* **84,** 2866–2869.

Bovey, F. A., Cais, R. E., Jelinski, L. N., Schilling, F. C., Starnes, W. H., Tonelli, A. E. (1983), "Structural and Dynamic Characterization of Polymers by ^{13}C and ^{19}F NMR," in: C. D. Craver (Ed.), *Polymer Characterization: Spectroscopic, Chromatographic and Physical Instrumental Methods,* Advances in Chemistry series, Vol. 263, Washington, DC: American Chemical Society.

Carter, R. O., III (1992), *Appl. Spectrosc.* **46,** 219–224.

Elias, H. G. (1977), *Macromolecules,* Vols. 1 and 2, New York: Plenum.

Grubisec, Z., Remmp, P., Benoit, H. (1967), *Polym. Lett.* **5,** 753–759.

Havens, J. R., Koenig, J. L. (1983), "Advances of High Resolution carbon-13 Nuclear Magnetic Resonance Spectroscopy to Solid Polymers," *Appl. Spectrosc.* **37,** 226–249.

Hiemenz, P. C. (1984), *Polymer Chemistry,* New York: Marcel Dekker.

Ivin, K. J. (Ed.) (1976), *Structural Studies of Macromolecules by Spectroscopic Methods,* New York: Wiley.

Zimm, B. H. (1948), *J. Chem. Phys.* **12,** 1099–1116.

Further Reading

de Gennes, P. G. (1979), *Scaling Concepts in Polymer Science,* Ithaca, NY: Cornell Univ. Press.

Elias, H. G. (1977), *Macromolecules,* Vols. 1 and 2, New York: Plenum.

Flory, P. J. (1953), *Principles of Polymer Chemistry,* Ithaca, NY: Cornell Univ. Press.

Flory, P. J. (1969), *Statistical Mechanics of Chain Molecules,* New York: Interscience.

Hiemenz, P. C. (1984), *Polymer Chemistry,* New York: Marcel Dekker.

Koenig, J. L. (1992), *Spectroscopy of Polymers,* Washington, DC: American Chemical Society.

Schröder, E., Müller, G., Arndt, K. F. (1988) *Polymer Characterization,* Munich: Hanser.

POLYMERS, OPTICAL PROPERTIES OF

WOLFGANG KNOLL, *Max-Planck-Institut für Polymerforschung, Mainz, Germany, and Frontier Research Program, The Institute of Physical and Chemical Research (RIKEN), Wako, Saitama, Japan*

Introduction 569
1. **Polymers with Specifically Tuned Bulk Optical Properties** 570
1.1 Polymers as Transparent Substrates for Optical Disks .. 571
1.2 Solid Polyelectrolytes as Glassy Polymers with Tunable Refractive Indices and Dispersion Characteristics 571
1.3 Magneto-optical Polymers 572
1.4 Photoresists 574
1.4.1 The NOVOLAC/Diazonaphthoquinone Resist .. 576
1.4.2 Deep-UV Resists 576
1.4.3 X-ray (Synchrotron) Resists .. 576
2. **Polymers in Special Configurations** 576
2.1 Ultrathin Films 577
2.2 Integrated Optics 581
2.2.1 Planar Waveguide Structures 582
2.2.2 Waveguides from Novel Polymeric Materials 584
2.2.3 Waveguide Loss Measurements 585
2.2.4 Supramolecular Design of Waveguide Films by the Langmuir–Blodgett–Kuhn Technique 586
2.2.5 Channel Waveguides 586
2.2.6 Polymer Optical Fibers (POF) 587
3. **Nonlinear Optical Properties of Polymers** 588
3.1 General Concepts for the Functionalization of Polymers for Nonlinear Optics 588
3.2 Second-Order Nonlinear Optics with Polymers 592
3.2.1 Control of Orientational Order 592
3.2.2 Second-Harmonic Generation (SHG) 594
3.2.3 Electro-optic (EO) Modulation 597
3.3 Third-Order Nonlinear Optics . 598
3.3.1 Third-Harmonic Generation (THG) 599
3.3.2 Degenerate Four-Wave Mixing (DFWM) 599
3.3.3 All-Optical Switching 600
3.4 Photorefractive Polymers 600
4. **Outlook** 602
5. **Acknowledgments** 602
Glossary 602
Works Cited 603
Further Reading 605

INTRODUCTION

Polymers are a ubiquitous class of materials in our daily life. These macromolecules—either in solution, as melts, liquid crystals, or glasses, or in the crystalline state—play a central role also in many modern technologies. This is due to a wide range of attractive properties, e.g., mechanical, rheological, electrical, etc. As far as their optical properties are concerned, however, only polymers in the glassy state and, to a lesser extent, liquid-crystalline phases are important and optimized for various applications.

A description of these optical properties includes, but cannot be limited to, the presentation of those phenomena and properties that describe their intrinsic response to (visible) light. A rough classification of these properties could start, e.g., with transparency and birefringence, which determine the intensity decrease and change of polarization, respectively, of light being transmitted through the sample or the device. The response to external fields is also such an intrinsic property, e.g., the change of the optical properties by applying (strong) magnetic fields described by magneto-optical effects.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

In cases, however, where the lateral dimensions of the sample are in the micrometer range, i.e., approach or are comparable to the wavelength of light, the optical properties cannot be characterized by bulk features any more but need to take into account the complex sample configuration. Here the three-dimensional arrangement of the (different) polymeric materials is intimately linked to the optical properties of the system. Channel waveguides or whole opto-boards in integrated optics are examples of this.

Another important aspect of the optical properties of polymers concerns not their purely intrinsic features but the nearly unlimited possibilities for a specific functionalization. Here the polymer serves as a matrix into which the functional optical units are incorporated—either by mixing or by covalent attachment. The optical properties of these materials are given by the interaction of the whole system with light.

Some of these functionalizations are designed so as to guarantee a strong interaction with light through specific electronic excitations giving rise to a broad range of phenomena: spectral changes due to the resonance character of absorption, or secondary processes (following the excitation) like luminescence phenomena (fluorescence, phosphorescence, etc.), are examples. Many systems show changes in other properties when photoexcited—e.g., they become photoconductive. The general term photoresponsive was coined to summarize the long list of property changes that can be induced by a purposeful design of the material, typically by the proper functionalization: this list includes photoisomerization and photochemical reactions resulting in a change of optical, mechanical, structural, morphological, rheological, solubility, or (meso-)phase properties. Finally, nonlinear optical properties, relevant for the description of the response of the material to ultrastrong optical fields, i.e., to a giant laser pulse, define the most active field of research in recent years related to optical properties of polymers.

In what follows, we are able to cover only a few selected examples in trying to highlight some of the improvements that have been achieved by extensive research and development that was focused in two nearly opposite directions: One goal was to *minimize* the interaction with light, e.g., to reduce the losses in a polymer optical fiber (POF) used in (local) data communication networks or to decrease the noise in optical recording and reading for information storage on optical disks. The other strong effort was to *enhance* the interaction with light, e.g., to increase the light-intensity–dependent refractive-index change n_2 in selected polymers needed to switch "light by light" in a future photonics device application.

In Sec. 1 we concentrate on some "passive" bulk material properties related, e.g., to their transparency or their dispersion properties. Section 2 is called "polymers in special configurations" and describes polymers in ultrathin films from monomolecular layers to multilayer assemblies, integrated optics with planar- and channel-waveguide structures, and fibers. Section 3, finally, is devoted to nonlinear materials and device concepts based on second- ($\chi^{(2)}$) and third-($\chi^{(3)}$) order nonlinear optical effects.

1. POLYMERS WITH SPECIFICALLY TUNED BULK OPTICAL PROPERTIES

In this section, we concentrate on the discussion of the remarkable progress that has been made in recent years in the development of polymeric materials with specific, tailor-made property profiles. Depending on the various applications, these synthesis and processing optimizations were in some cases aimed at reducing as much as possible any interaction with light so as to not change its intensity, state of polarization, etc. Clearly, these are the applications where polymers are to replace inorganic glasses as windows, shields, eyeglasses, substrates, etc. A more recent example is the use of polymers as substrate materials for optical storage disks or highly transparent materials for polymer optical fibers.

Another important optical property of polymers is their dispersion behavior: The isotropic or anisotropic refractive index and its dependence on the wavelength of light are key parameters that need to be tunable to certain values in all areas of polymers for integrated optics. As an example, solid polyelectrolytes with a broad range of refractive-index values and dispersion characteristics are briefly discussed below.

After a short introduction to some recent developments of magneto-optical applications of polymers, we turn to probably the most important application of polymers in microelectronics: as resist materials that are optimized with respect to their absorption of photons (and, of course, with respect to other property changes like solubility following excitation by the exposure to light).

1.1 Polymers as Transparent Substrates for Optical Disks

The availability of low-cost lasers made optical information storage an attractive alternative to existing mass-storage concepts. The largest volume use of polymers for optical recording, however, is still in disk substrates. The material used for compact disks (CD), CD-ROMs (read-only memory), and video disks has to fulfill extreme requirements (Kämpf, 1985):

1. optical homogeneity (low birefringence);
2. high transparency in the visible, in some cases also in the UV (λ = 280–380 nm);
3. high shape stability (low thermal expansion, water uptake ≤0.3%);
4. isotropic expansion upon temperature or humidity increase;
5. good mechanical properties;
6. constant thickness, planarity; and
7. constant rheological properties for the molding process of the "pits".

Currently, mostly poly(methylmethacrylate) (PMMA) and poly(4-methyl-1-pentene) as well as bisdiallylpolycarbonates (PC) are used for the fabrication of substrate disks. These polymers offer the optical quality of glass. Some selected physical properties of PMMA and PC are summarized in Table 1 (after Pearson, 1988).

Table 1. Selected physical properties of PMMA and PC substrates.

	PC	PMMA
Transmittance (%) at 830 nm	90	92–93
Refractive index at 830 nm	1.58	1.49
Birefringence (nm) at 830 nm	20–30	<20
T_g (°C)	150	100
Thermal-expansion coefficient [(°C)$^{-1}$]	(6–7) × 10^{-5}	7.6 × 10^{-5}
Rockwell hardness (M scale)	75	90
Izod impact strength (kg cm/cm)	1–2	1.6
H_2O absorption (%)	0.25	0.54

1.2 Solid Polyelectrolytes as Glassy Polymers with Tunable Refractive Indices and Dispersion Characteristics

The various available polymers can have very different refractive indices and dispersion properties. However, Fig. 1 shows that these materials exhibit typically much lower refractive indices than inorganic glasses as a result of the absence of highly polarizable units like, e.g., the heavy metals with their many electrons in flint glasses. One extreme case, e.g., is poly(tetrafluorethylene) with n_D = 1.345 (data point A in Fig. 1). A material with a substantially higher refractive index like poly(N-vinylcarbazole) with n_D = 1.685 has a very high dispersion ν_D (the Abbe number ν_D is defined by $\nu_D = (n_D - 1)/(n_F - n_C)$, with n_D, n_F, and n_C being the refractive indices at three standard wavelengths: blue, λ = 486.1 nm; yellow, 589.2 nm; red, 656.3 nm, respectively).

One successful attempt to expand the available polymeric materials toward higher refractive indices was recently reported for solid polyelectrolytes, also called ionenes

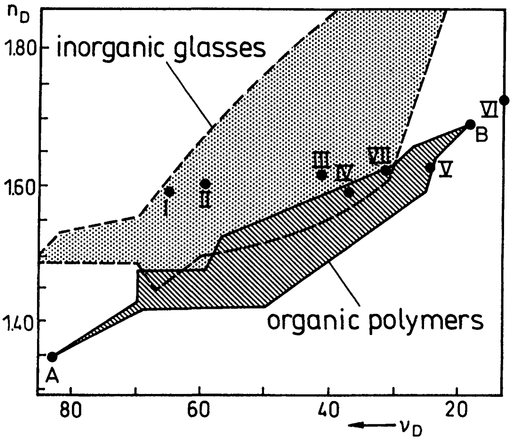

FIG. 1. Refractive indices n_D and optical dispersion ν_D of inorganic glasses (dotted area), organic polymers (dashed area), and a series of selected ionenes (numbered I–VII according to Table 2). A, poly (tetrafluorethylene); B, poly(N-vinylcarbazole).

(Simmrock *et al.*, 1989). Given their general structure

$$[R^1 - N^{\oplus}(R^3)_2 - R^2 - N^{\oplus}(R^4)_2]X^{\ominus\ominus}, \qquad (1)$$

with the main-chain segments R^1 and R^2, and R^3 and R^4 being organic moieties, e.g., aliphatic groups, there is an almost unlimited range for tuning their solid-state properties, e.g., their crystallinity or their ability to solidify as glasses, while on the other hand their optical properties can be largely varied by the introduction of suitable counterions, both monovalent and divalent. In particular, by the introduction of structural irregularity into the main chain ($R^1 \neq R^2$), glassy materials with very different polymeric properties like the glass-transition temperature could be synthesized. Table 2 indicates the progress in extending the selection of polymers toward higher refractive indices with a wide range of Abbe numbers (cf. Fig. 1).

The interrelation between high refractive index and high dispersion is demonstrated in Fig. 2 for two different ionenes (Mathy *et al.*, 1991). The first one, polyelectrolyte I (cf. Table 2), with $ZnBr_4$ as counterions, has its strongest absorption in the deep ultraviolet (DUV) range not exceeding $\lambda = 220$ nm [see Fig. 2(a)]. The Kramers–Kronig analysis with a background refractive index of $n_0 = 1.516$ shows that as a result the dispersion in the visible is very low despite a relatively high refractive index [Fig. 2(b)]. On the other hand, the introduction of HgI_4 ions into the polyelectrolyte VI extends the absorption of the material into the near-uv range [Fig. 2(c)] with a consequently much higher (resonance enhanced) refractive index but also with a very strong wavelength dependence, i.e., low Abbe number [Fig. 2(d)].

It is clear that this flexibility for tuning the linear optical properties of polymeric materials is a very promising feature for future device configurations, e.g., in integrated optics. We should also mention that the introduced counterions, of course, may also have high nonlinear polarizabilities, allowing for the preparation of highly functionalized materials, e.g., for $\chi^{(2)}$ (Weiss *et al.*, 1995) or $\chi^{(3)}$ applications (Meyer *et al.*, 1991).

1.3 Magneto-optical Polymers

Certain materials act naturally as polarization rotators, a property known as optical activity: waves with right- and left-circular polarization travel with different velocities. This property is found in media in which the molecular units have an inherently helical character (quartz, selenium, tellurium, but also many organic/polymeric materials). Materials that have different absorption coefficients for right and left circularly polarized light are said to exhibit circular dichroism.

An effect related to this, the Faraday effect, describes the polarization rotation of light passing through a material placed in a static magnetic field. The Faraday rotation angle ξ is given by

$$\xi = VBl, \qquad (2)$$

where B is the magnetic induction in the direction of the light-wave propagation, l is the length of the sample, and V is a material parameter, called the Verdet constant. This

Table 2. Chemical structures and optical properties of selected ionenes.

No.[a]	R^{1b}	R^{2b}	R^{3b}	R^{4b}	X^{b}	n_D	ν_D
I	CH_2-CH=CH-CH_2	$(CH_2)_{10}$	CH_3	CH_3	$ZnBr_4$	1.5911	65.0
II	CH_2-CH=CH-CH_2	cyclo-NC_5H_9-$(CH)_3$-cyclo-C_5H_9N	CH_3	(c)	$CdBr_4$	1.6018	59.7
III	CH_2-CH=CH-CH_2	cyclo-NC_5H_9-$(CH)_3$-cyclo-C_5H_9N	CH_3	(c)	BPh_4	1.6142	41.0
IV	$(CH_2)_6$	$(CH_2)_{10}$	CH_3	CH_3	$CdBr_4$	1.5858	37.6
V	$(CH_2)_6$	$(CH_2)_{10}$	CH_3	CH_3	CdI_2Br_2	1.6250	24.6
VI	CH_2-CH=CH-CH_2	$(CH_2)_6$	CH_2CH_3	CH_2CH_3	HgI_4	1.7245	13.2
VII	$(CH_2)_6$	$(CH_2)_{10}$	CH_3	CH_3	ZnI_2Br_2	1.6260	31.6

---

[a] Numbers corresponding to Fig. 1.

[b] Defined as in Eq. (1).

[c] Because of cyclic structure only one CH_3 group.

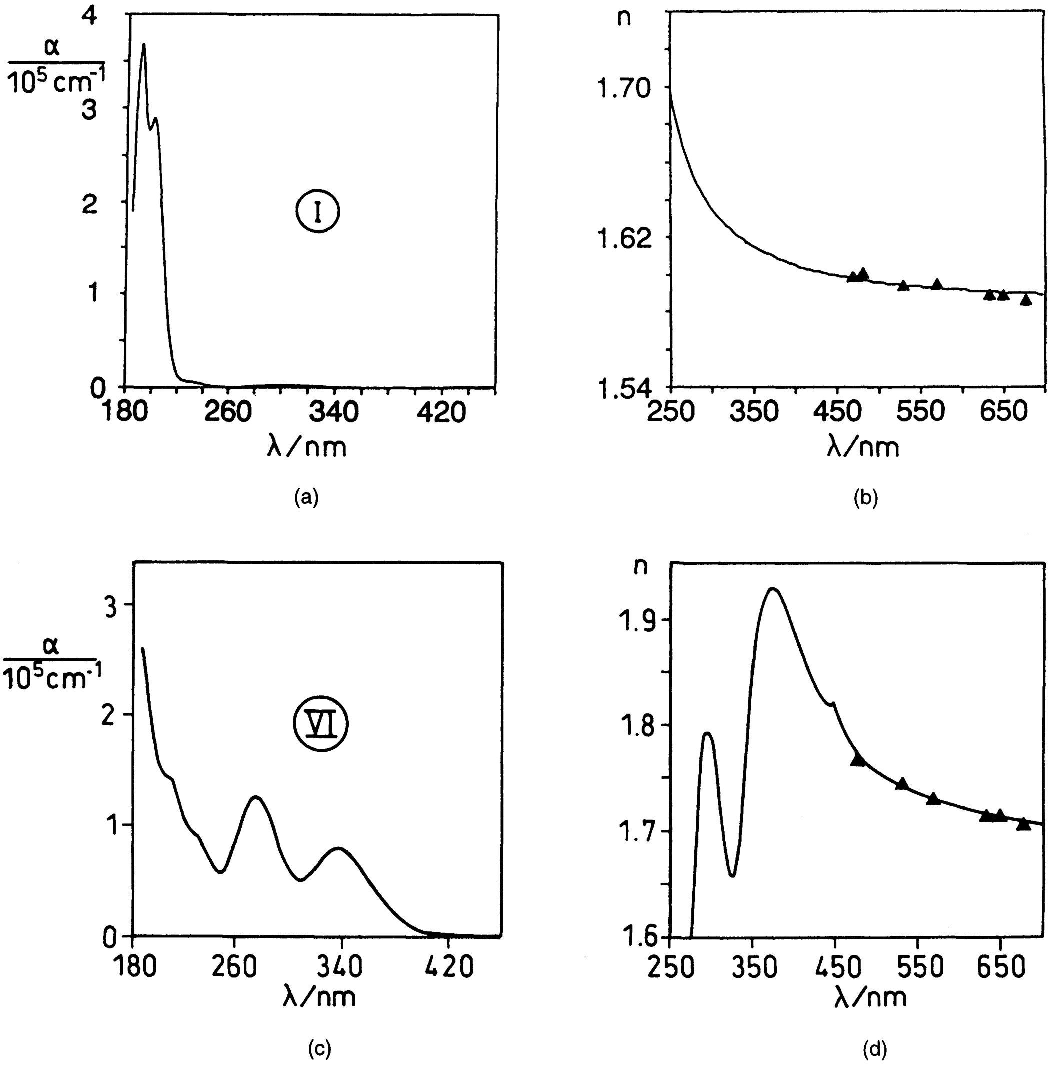

(a) (b)

(c) (d)

FIG. 2. (a) Absorption spectra of ionene I (cf. Table 2) and (c) VI; (b), (d) their respective refractive indices measured at different wavelengths together with the corresponding Kramers–Kronig analysis (full lines), with a background refractive index $n_0 = 1.515$.

constant can be relatively high in polymers containing phenyl groups [e.g., $V \approx 0.1$ min/(Oe cm) for polystyrene; Muto and Ito, 1992]. In a Faraday rotator, the sense of the rotation is governed by the direction of the magnetic field and, therefore, does not change with the reversal of the direction of propagation of the wave. This is the basic concept for optical isolators based on this phenomenon. Its principle is outlined in Fig. 3: A linearly polarized wave passing the Faraday rotator may experience a polarization rotation of, say, 45°. It passes another polarizer B tuned to this angle. After reflection it again can go through the polarizer, but is then rotated by the rotator by another 45° in the same direction. This wave is now orthogonal to the incident polarizer A and hence is blocked. Muto and co-workers (1991) have used a poly(α-methylstyrene) rod of $l = 19$ mm to construct a 45° rotator for $\lambda = 488$ nm light at 4 kOe magnetic field. Its transmission characteristic is shown in Fig. 4. Inserting the isolator into the laser beam

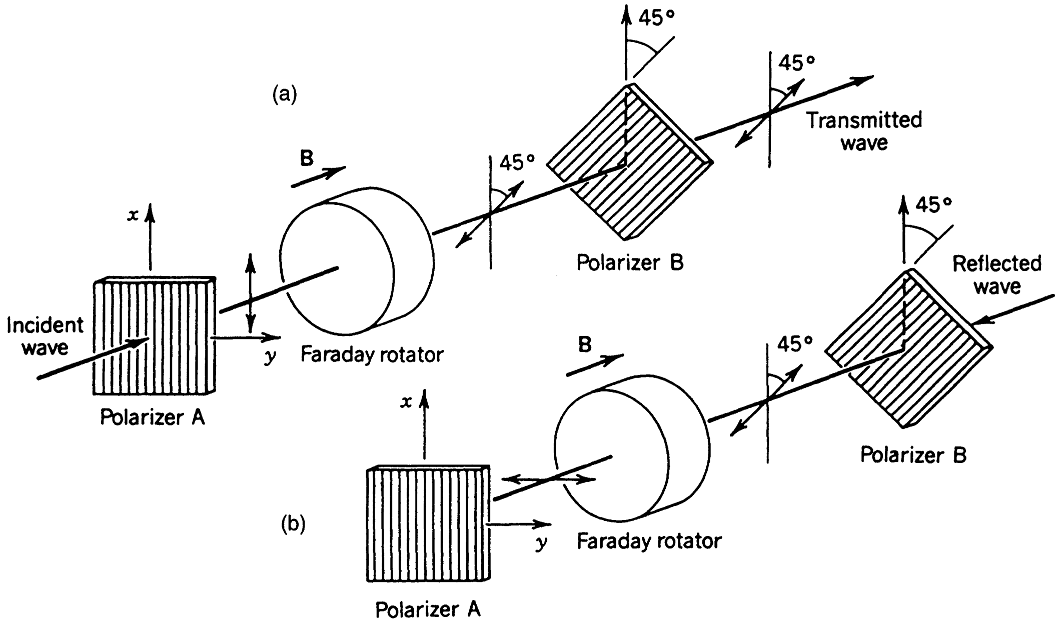
(a)
45°
45°
Transmitted wave
B
45°
x
Incident wave
y
Faraday rotator
Polarizer A
Polarizer B
45°
Reflected wave
B
45°
x
Polarizer B
(b)
y
Faraday rotator
Polarizer A

FIG. 3. An optical isolator based on the Faraday effect (a) transmits light in one direction but (b) blocks it in the opposite direction.

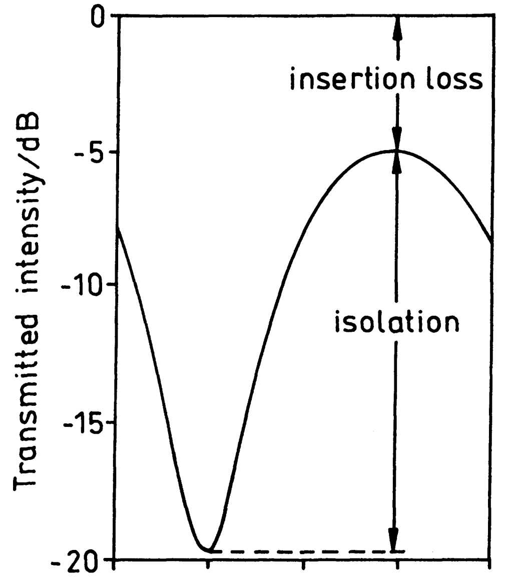

FIG. 4. Transmitted light intensity through a polymeric Faraday rotator measured as a function of the analyzer angle (polarizer B in Fig. 3) (after Muto *et al.*, 1991).

causes a loss of ca. 5 dB, but the isolation obtained is ca. 15 dB. Further improvements of the device performance are expected for materials with lower birefringence.

1.4 Photoresists

The basic lithographic processing steps are schematically depicted in Fig. 5. Onto the oxidized silicon wafer (a), a thin photoresist polymer is deposited typically by spin-coating (b). Either by use of a photomask (c1), which allows for a structured parallel exposure of the photoresponsive material, or by a setup that scans a focused beam (e.g., an electron beam) for a serial exposure across the polymer layer (c2), a latent image is stored in the resist material. By a developing process, this image is converted into a surface-relief structure (d). Depending on the chemical nature of the resist material, the exposed areas may be rendered more soluble to the developer than the unexposed areas, thereby producing a positive-tone image of the mask, or these areas are made less soluble, in which case a negative-tone image

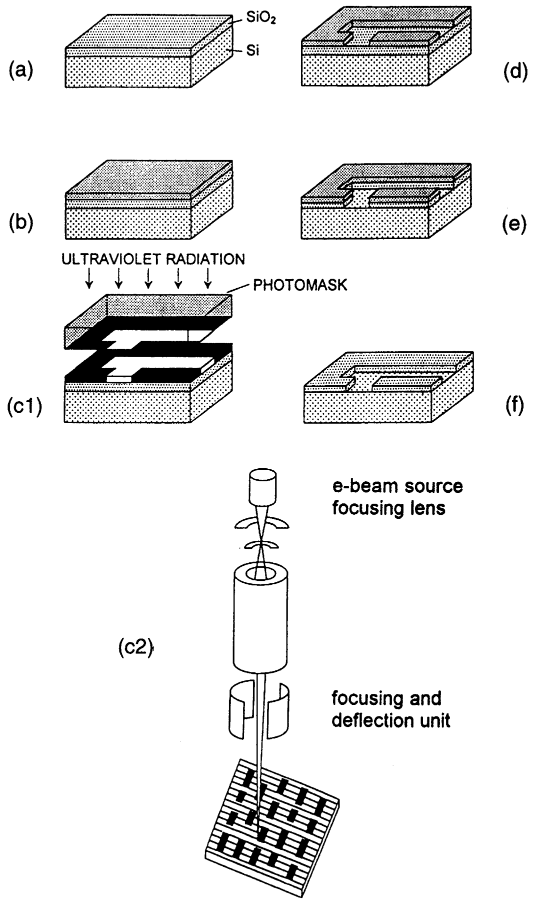

FIG. 5. Basic processing steps in lithography: (a) The oxidized wafer is (b) (spin-)coated by a thin photoresist layer. This is irradiated to form a latent image either (c1) by a photomask in a parallel exposure, or (c2) by an electron beam in a serial process. Developing leads to (d) a relief structure that allows for (e) a patterned etching of the oxide layer. (f) After stripping off the polymer the structure is ready for the next cycle. (After Bowden and Turner, 1988.)

is obtained. In either case, the development produces uncovered areas on the wafer (e) that allow for a structured etching or diffusion processing of the substrate. After the removal of the resist [the stripping (f)], the final three-dimensional replication of the two-dimensional mask pattern in the wafer surface is obtained.

This scheme of the processing steps is sufficient to appreciate some of the basic unique properties that are required of a resist material (Bowden, 1988):

1. It must be capable of being uniformly deposited as a thin (0.5–2 μm) film across the wafer of up to 8 in. (20.3 cm) in diameter.
2. It must be sensitive to different types of radiation (uv, DUV, x-ray, electron beam) and must, upon exposure, undergo a chemical change inducing a change in some physical property of the materials, e.g., a solubility change, that enables exposed and unexposed areas to be differentiated by an appropriate developing scheme.
3. It must adhere to a variety of substrates and withstand the various etching environments encountered in integrated-circuit (IC) processing, including high temperatures, corrosive etching chemicals, etc.
4. A most important requirement is the high resist resolution. The current IC mass production uses 0.8-μm lines, 0.5 μm are in the development stage, and the 64-Mb (megabit) DRAM (dynamic random-access memories), scheduled for 1995, require 0.35-μm structures to be fabricated.

The basic physical/optical limitations encountered in the image-generation process during exposure are related to the diffraction limit of the projection tools used in photolithography [or to image degradation by (secondary) scattering processes in the resist in electron-beam lithography]. This has led to the development of polymeric resist materials for the use of shorter wavelengths in DUV lithography or even synchrotron radiation in x-ray lithography. On the materials side, this has far-reaching consequences as to the strategies for converting the latent image into a surface-relief structure by the development step. In cases where the exposing radiation has not enough energy to rupture bonds directly (e.g., at the wavelengths encountered in traditional lithography: e.g., with the so-called *g* line at $\lambda = 436$ nm and the *i* line at $\lambda = 365$ nm), differential solubility schemes based on polarity changes in the matrix resulting from polarity changes of the sensitizer have been developed. This principle is used in the NOVOLAC-based photoresists. For lithography with radiation that has sufficient energy to break bonds, the mere change in molecular weight of the polymer matrix could be enough for a solubility differentiation.

1.4.1 The NOVOLAC/Diazonaphthoquinone Resist Today, all submicron imaging in photolithography is done with positive photoresist. These materials are two-component systems based on diazonaphthoquinone-sensitized NOVOLAC resins. The dissolution changes are caused by chemical modifications in the sensitizer that cause polarity changes in the irradiated matrix rendering the NOVOLAC soluble in aqueous base. This scheme is shown in Fig. 6. The system contains a cresol-formaldehyde NOVOLAC resin as the matrix material. Photosensitivity is given to the polymer by addition of diazonaphthoquinone, which undergoes photolysis to produce a ketene intermediate that rapidly reacts with the water in the resin to yield an indenecarboxylic acid. While the pristine lipophilic diazonaphthoquinone serves as a dissolution inhibitor reducing the solubility of NOVOLAC films in aqueous base, the final acidic photoproduct renders the film soluble. Hence, this system functions as a positive resist. Because the polarity of the unexposed resist is unchanged, resolution of the pattern is not limited by swelling of the unexposed areas as was the case for the older resists based on bisazide–rubber formulations. Many modifications to this basic mechanism have been developed to improve the resist performance further: These include image reversal, contrast enhancement by bilayer or multilayer resist preparations, etc. (Willson and Bowden, 1988).

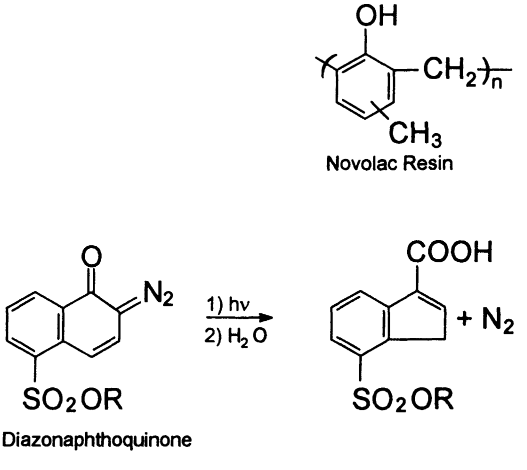

FIG. 6. The two-component photoresist based on a NOVOLAC matrix and a diazonaphtoquinone sensitizer.

1.4.2 Deep-UV Resists The usual exposure wavelength for diazonaphthoquinone systems is λ = 405–436 nm. Materials designed for an effective exposure in the mid-uv (300–350 nm)—e.g., polysilanes—have proven to increase the resolution needed for the continuous trend toward higher device integration. A selection of DUV materials is listed in Table 3. Only positive-tone resists are given. Another important feature of these systems is their organosilicon content. In an oxygen plasma used for reactive ion etching (RIE), these silicon moieties are converted to silicon dioxide at the surface of the resist, acting there as a passivation layer that protects the underlying polymer against further erosion.

Another system recently reported by Przybilla and co-workers (1992) for use in DUV lithography with KrF eximer lasers at λ = 248 nm is given in Fig. 7. Based on a partial replacement of hydroxystyrene with 2-hydroxy-hexafluoroisopropylstyrene, a resist material is obtained that is highly transparent, dry-etch resistant, and alkali soluble. By the addition of photoreactive inhibitors, a photolithographic polymer is obtained that allows for the fabrication of narrow lines of sizes far into the submicron range. An example is given in Fig. 8.

1.4.3 X-ray (Synchrotron) Resists Synchrotron radiation, i.e., the relativistic emission of light from electrons that propagate in a storage ring nearly with the speed of light, is the ultimate light source also for lithographic applications (Heuberger, 1988). Consequently, active research is conducted worldwide on materials that could be used for x-ray lithography in the wavelength range of 0.1 to 5 nm. The 1:1 projection technique with a mask–wafer separation of typically 50 μm allows for the fabrication of structures with less than 0.1-μm linewidths. Exposure times of only a few seconds have been shown to be enough to produce structures with aspect ratios >20:1 (height:width), not limited by diffraction, interference, or focal depth and independent of substrate layers or wafer topography.

2. POLYMERS IN SPECIAL CONFIGURATIONS

The foregoing discussion was focused on polymers where the bulk optical properties were of interest, specifically tailored for se-

Table 3. Organosilicon resists for DUV lithography (after Willson and Bowden, 1988).

Structure	Type	Company
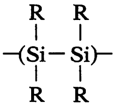 $-(\underset{R}{\overset{R}{Si}}-\underset{R}{\overset{R}{Si}})-$	Positive-tone, DUV, mid UV	IBM (1983) Sandia (1983)
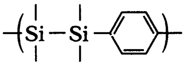	Positive-tone, DUV	Hitachi (1984)
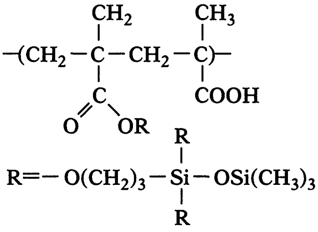	Positive-tone, electron-beam, DUV	AT&T (1984)
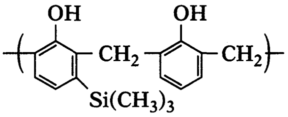 + Quinone Diazide	Positive-tone, near-UV	AT&T (1984)
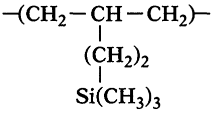 $-(CH_2-\underset{\underset{Si(CH_3)_3}{(CH_2)_2}}{CH}-CH_2)-$	Positive-tone, electron beam	Bellcore (1986)
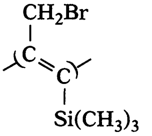	Positive-tone, DUV	Bellcore (1987)

lected applications. In this section, we concentrate on optical properties of polymers prepared in a very special configuration, where this configuration and the fabrication protocol are integral aspects of their optical properties.

2.1 Ultrathin Films

Ultrathin polymer films, in the extreme case only one molecular layer thick, are of special interest from an applied point of view for all kinds of surface modifications, be it for lubrication, isolation, adhesion, or any other possible purpose, and provide a special challenge for the experimentalist who has to develop methods that are sensitive enough to allow for an optical characterization of coatings only a few nanometers in thickness.

One example is provided by so-called Langmuir monolayers (Langmuir, 1917). Molecules like soaps, lipids, etc., can be prepared as monomolecular films at a water–air interface. Driven by the reduction of the surface free energy of water, these molecules spread across the surface when applied, e.g., from a volatile solvent. After compression these monolayers can be transferred to a solid support by dipping and withdrawing a suitable substrate through this highly organized film at the water–air interface. Thus, multilayer assemblies with a precise control of their thickness can be obtained, as is schematically sketched in Fig. 9. First reported by Blodgett (1935), this technique has

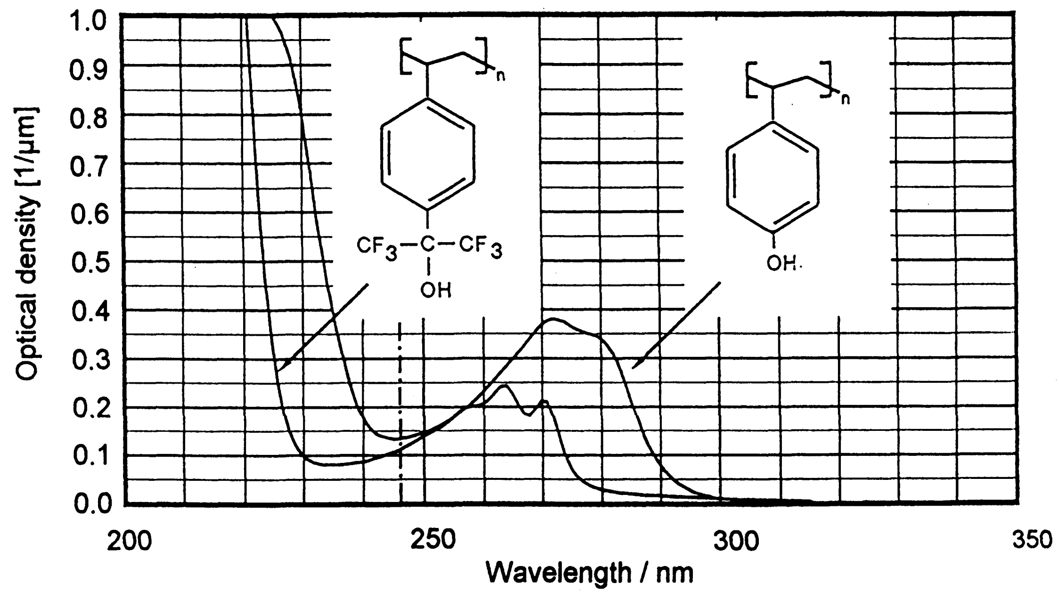

FIG. 7. Comparison of UV spectra of poly[4-(2-hydroxy-hexafluoroisopropyl)styrene] and poly(4-hydroxystyrene). Film thickness 1 mm on a quartz wafer. (After Przybilla *et al.*, 1992.)

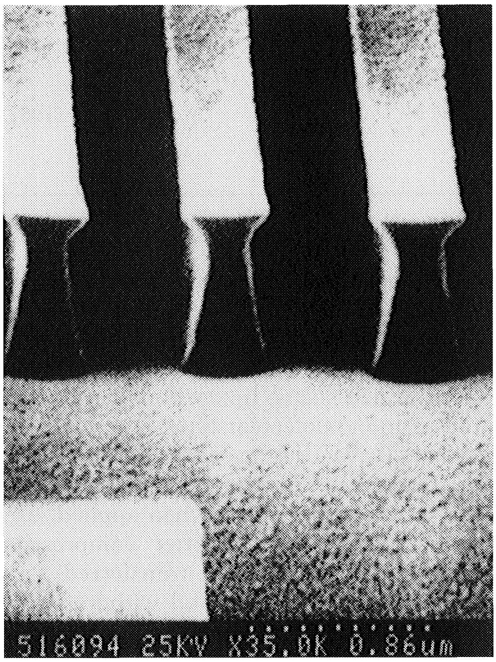

FIG. 8. 0.5-μm line and space structures obtained with dissolution-inhibition type photoresist material based on a modified polystyrene (cf. Fig. 7). Exposure at λ = 248 nm. (After Przybilla *et al.*, 1992.)

regained a worldwide interest as an excellent tool for the buildup of supramolecular architectures (Roberts, 1990) with tailor-made molecules ever since the pioneering work of Kuhn and co-workers (1972). These systems are therefore called Langmuir–Blodgett–Kuhn (LBK) layers.

Various techniques have been developed in recent years sensitive enough to allow for

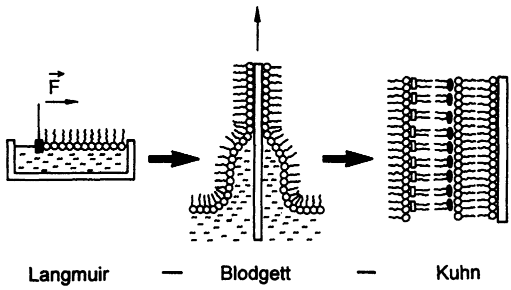

FIG. 9. Schematic representation of the Langmuir–Blodgett–Kuhn technique: The Langmuir monolayer prepared at the water–air interface at a selected surface pressure is transferred during the withdrawal of the solid substrate. Supramolecular architectures can be built up by choosing different functional units in the individual layers.

a quantitative evaluation of the optical properties of these ultrathin coatings (Swalen, 1986). One group of techniques uses evanescent waves and, in particular, surface plasmons as highly specific and sensitive interfacial light probes (Knoll, 1991a, b).

One way of introducing surface plasmons [plasmon-surface polaritons (Burstein *et al.*, 1974; Raether, 1977) or PSPs], especially in the experimental setup called the Kretschmann configuration (Kretschmann, 1972) (see Fig. 10, top), is to note that the nearly free-electron gas in the thin (≈50 nm) metal film evaporated onto the base of the prism acts as an oscillator that can be driven by the electromagnetic wave impinging upon that interface at total internal reflection conditions. Therefore, we are dealing with the resonant excitation of a coupled state between the plasma oscillations and the photons, i.e., the "plasmon–surface polaritons" (Raether, 1988). This resonance phenomenon can be clearly seen in the ATR scan (attenuated total reflection; see bottom half of Fig. 10): Below θ_c the reflectivity is very high because the metal film acts as a mirror with little transmission. Above θ_c for total internal reflection, however, a relatively narrow dip in the reflectivity curve at θ_0 indicates the resonant excitation of such a PSP wave at the metal–air interface. The coupling angle is given by the energy- and momentum-matching condition between photons and surface plasmons:

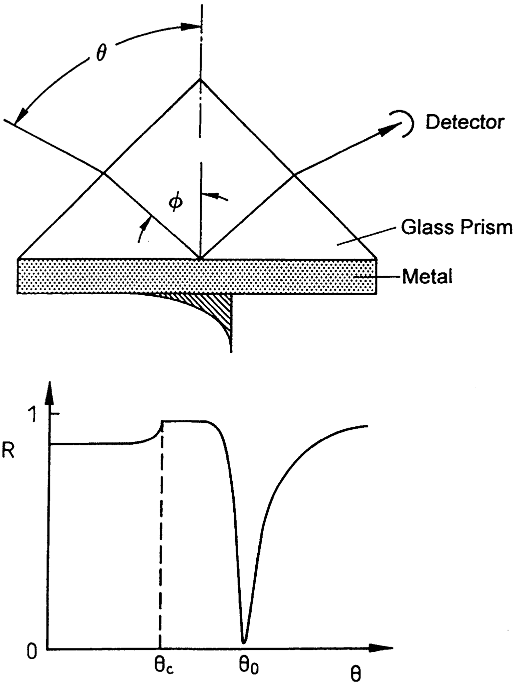

FIG. 10. ATR setup for the excitation of surface plasmons (PSPs) in Kretschmann geometry: (Top) A thin metal film ($d \sim 50$ nm) is evaporated onto the base of the prism and acts as a resonator driven by the photon field. (Bottom) The reflectivity R is recorded by a detector as a function of the angle of incidence and shows the increase to unity at θ_c, the critical angle for total reflection. The resonant excitation of the PSP wave is seen in the reflectivity curve as a sharp dip at coupling angle θ_0.

$$k_{sp}^0 = k_{ph}^0 = n_1(\omega/c)\sin\phi_0, \tag{3}$$

with k_{ph}^0 being the parallel component of the photon wave vector, ω the photon energy, c the speed of light, and ϕ_0 the (internal) coupling angle (see Fig. 10). Since we are dealing with an evanescent wave propagating along the interface, the penetration depth into air is only of the order of the wavelength. The resonance character of this excitation gives rise to an enhancement of the field E at the interface by more than a factor of 10, which is the origin of the remarkable sensitivity enhancements obtainable, e.g., in Raman spectroscopy when working with PSP light (Ushioda and Sasaki, 1983; Knobloch *et al.*, 1989). These modes obey a known dispersion relation, ω vs k_{sp}, so that each photon of energy $\hbar\omega_L$ allows for the excitation of exactly one PSP mode. This is schematically depicted in Fig. 11(a). The solid curve represents the dispersion of the surface plasmons at a metal [e.g., Ag—air interface (PSP^0)]. The horizontal line at ω_L intersects the dispersion curve at k_{sp}^0 and thus defines the coupling angle θ_0. The typical experimental setup is schematically given in Fig. 11(b). *P*-polarized photons from a laser excite PSP at the base of the prism, which is first covered with the thin metal layer by evaporation and then coated, e.g., by LBK layers of increasing thickness. The reflected intensity is monitored as a function of the angle of incidence by a photodiode. The thin dielectric coating causes a shift of the dispersion curve (PSP) to higher momentum:

$$k_{sp}^1 = k_{sp}^0 + \Delta k_{sp}, \tag{4}$$

which, according to Eq. (3), shifts the reso-

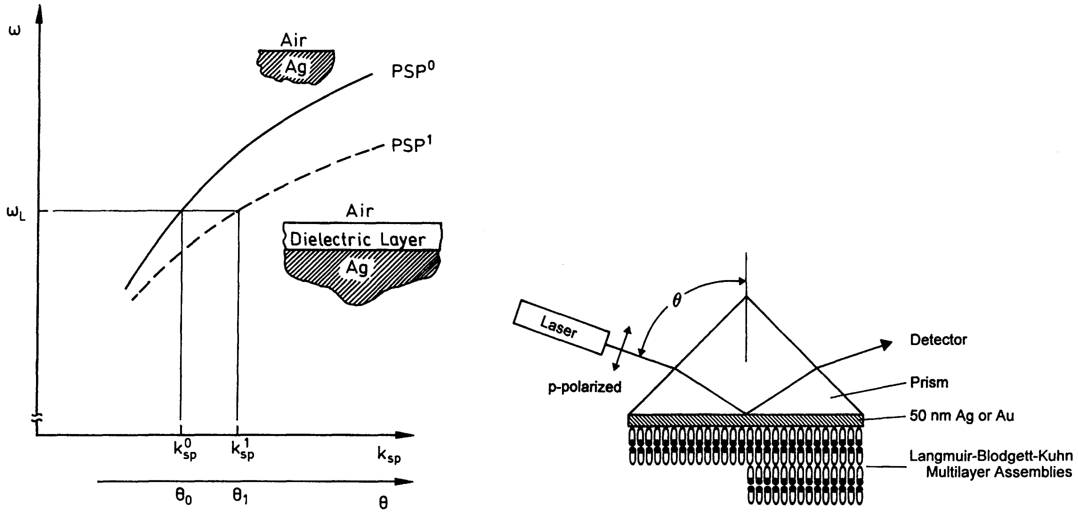

(a) (b)

FIG. 11. (a) Dispersion relation, ω vs k_{sp}, of plasmon–surface polaritons at a Ag–air (PSP^0, solid line) and at a Ag–dielectric coating–air interface (PSP^1, broken line), respectively. Laser light of energy ω_L couples at angles θ_0 and θ_1, respectively, given by the energy- and momentum-matching condition (see the intersection of the horizontal line at ω_L with the two dispersion curves). (b) Schematic of the experimental setup for surface-plasmon spectroscopy.

nance to a higher angle θ_1. From this shift and Fresnel's equations, one can calculate the optical thickness of the coating (Gordon and Swalen, 1977). This is demonstrated quantitatively for LBK multilayer assemblies prepared from polyglutamate monolayers. A schematic drawing of a transferred double layer is presented in Fig. 12(a). It shows the quasi two-dimensional nematic structure of these "hairy rods"—stiff polyglutamate helices with long, flexible alkyl side chains (Duda *et al.*, 1988).

Figure 12(b) gives the results of ATR scans obtained from the bare metal and 2, 4, 6, 10, and 20 layers of polyglutamate, respectively (Hickel *et al.*, 1990). The symbols are the experimental data points, the solid lines Fresnel calculations. All these layer preparations can be described by a constant index of refraction $n_z = 1.486$, and the multiple of a constant thickness per monolayer of $d_0 = 1.75$ nm.

The sensitivity of this technique to extremely small refractive-index changes is demonstrated in Fig. 13(a). Given is the angular scan (full circles) of the PSP resonance found for a sample consisting of Ag, 8 layers of polyglutamate, and 6 layers of an azobenzene-derivatized (and hence photoreactive) liquid-crystalline polyacrylate (structural formula is given in the inset) that could be processed by the LBK technique (Sawodny *et al.*, 1992). The spacer layers were necessary in order to decouple (electronically) the azobenzene chromophores from the metallic acceptor states in the substrate. Illuminating the sample with uv light ($\lambda = 360 \pm 30$ nm) shifts the dark-adapted all-*trans* chromophores through *trans–cis* isomerization to a new photostationary equilibrium with a high *cis*-isomer content. As a result, the optical refractive-index anisotropy is changed and the index normal to the layers is reduced. This shifts the surface-plasmon resonance to smaller angles—the LBK multilayer is now optically thinner [open triangles in Fig. 13(a)].

If this change of the optical thickness is followed on line (during illumination) by recording the reflected intensity at a fixed angle of incidence [e.g., $\theta = 47.2°$, cf. Fig. 13(a)], a kinetic analysis of the refractive-index changes can give information as to the reaction rates, the equilibrium changes, the reversibility, etc. The latter important aspect is shown in Fig. 13(b). The first rapid decrease of n_z upon uv illumination can be partly restored by switching to visible light (λ = 450

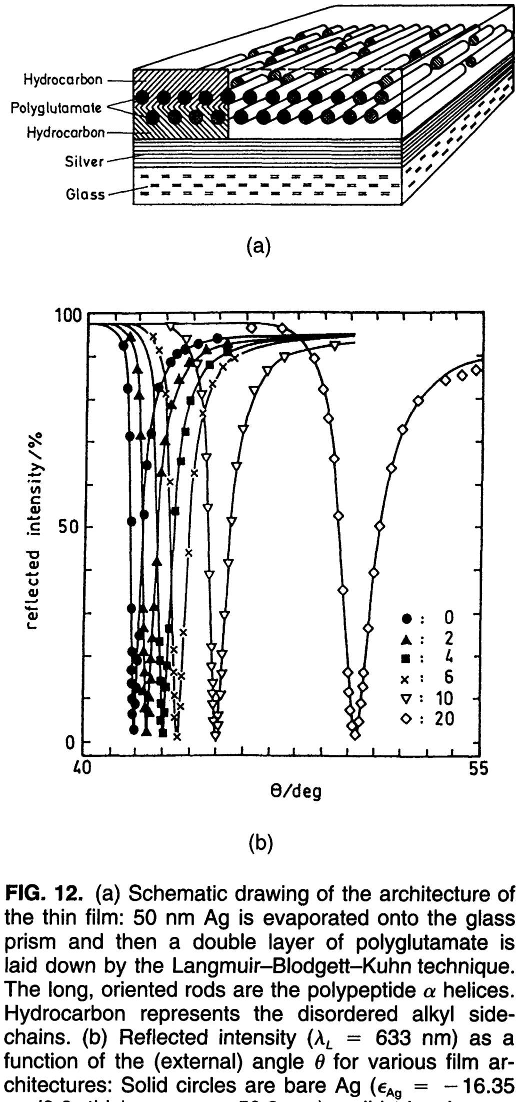

(a)

(b)

FIG. 12. (a) Schematic drawing of the architecture of the thin film: 50 nm Ag is evaporated onto the glass prism and then a double layer of polyglutamate is laid down by the Langmuir–Blodgett–Kuhn technique. The long, oriented rods are the polypeptide α helices. Hydrocarbon represents the disordered alkyl side-chains. (b) Reflected intensity (λ_L = 633 nm) as a function of the (external) angle θ for various film architectures: Solid circles are bare Ag ($\epsilon_{Ag} = -16.35 + i0.6$, thickness s = 50.8 nm); solid triangles are two layers of polyglutamate on top; solid squares = 4; crosses = 6; open triangles = 10; and diamonds = 20 layers. Solid lines are Fresnel calculations.

± 30 nm), which isomerizes the azobenzenes back into the *trans* state. These reaction cycles can be conducted many times. It is important to note that, given the signal-to-noise level of these data, index changes as small as Δn_z = 0.0001 can be monitored in reactive layers that are only 15 nm thick!

2.2 Integrated Optics

Since the concept of "integrated optics" emerged in the late 1960s, a significant growth of the areas of optical communication, signal processing, etc. has stimulated a major research and development effort in the field of organic and polymeric materials aimed at optimizing their functional properties by molecular design and synthesis (Stegeman *et al.*, 1987; Prasad and Williams, 1991; Messier *et al.*, 1989). This concerns not only their nonlinear optical properties [i.e., primarily their second- and third-order hyperpolarizabilities (see Sec. 3)] but also their linear optical response to light and, equally important, also other chemical and physical properties of the material and processing parameters like stability (against chemical attack, temperature, light, aging, delamination, etc.), ease of preparation (compatible with technologies developed for microelectronic device fabrication), cost effectiveness, and others.

One very crucial requirement of a material for the envisaged integrated-optics application is that it can be prepared as a structure capable of guiding light waves with acceptable losses. The three corresponding basic configurations (Lee, 1986; Chang, 1991) are schematically depicted in Fig. 14: 14(a), the planar-waveguide structure given in the so-called asymmetric-slab configuration with a solid substrate and air as superstrate; 14(b), the various channel waveguides (strip, embedded strip, and buried strip, from top to bottom) prepared by starting in some cases with a planar waveguide and then further structuring and etching the sample; and 14(c), the optical fiber with a high-index core and a lower-index cladding. Such waveguide structures are discussed in the literature not only with respect to their role as passive interconnects between different integrated optical-circuit (IOC) components but also in the context of various design concepts for active IOC modules like directional couplers, phase and amplitude modulators, second-harmonic generators for laser diodes, etc. (Stegeman and Stolen, 1989).

In this section, we limit the discussion to the case of a planar waveguide structure, but the general ideas allow for a straightforward extrapolation to more complicated designs. After a short presentation of various schemes for coupling plane electromagnetic waves (e.g., photons from a laser) into such structures, we give a few examples for waveguides fabricated from novel polymeric materials

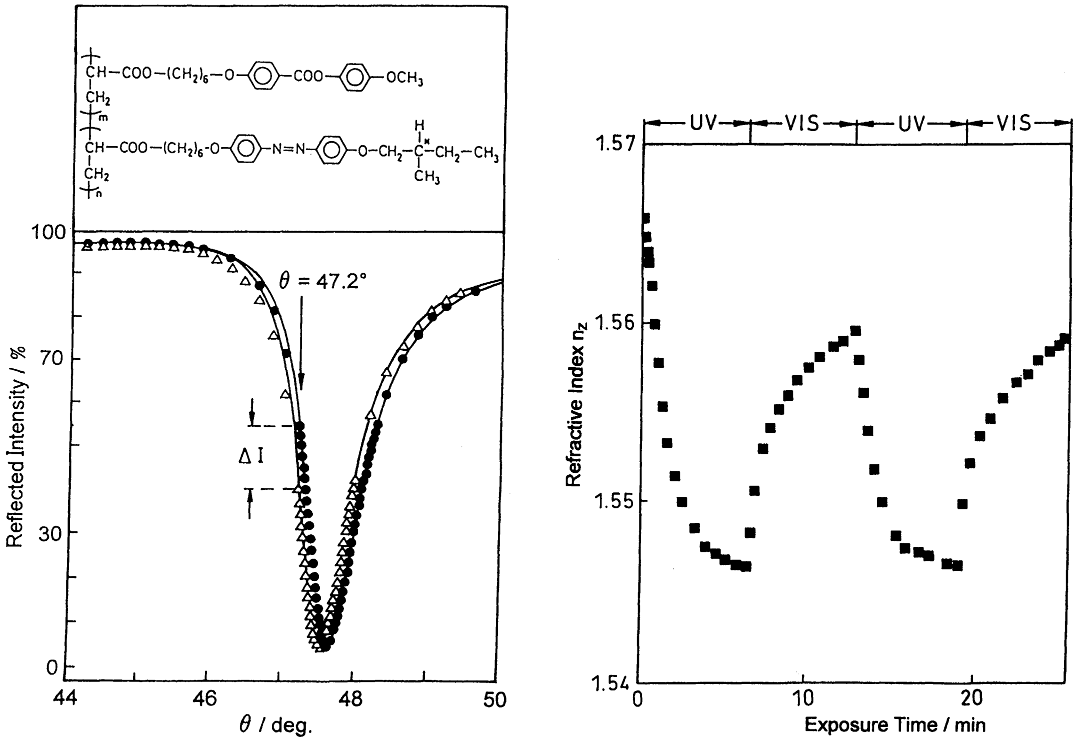

(a) (b)

FIG. 13. (a) Reflected intensity as a function of the angle of incidence for a sample consisting of Ag–8 layers polyglutamate–6 layers of the photoreactive liquid-crystalline polymer given in the inset; full circles, layers as prepared (in the dark-adapted *trans* state); open triangles, after illumination of the layers with UV light (λ = 360 nm) for 10 min. Full curves are Fresnel-fit calculations. (b) Refractive-index change as obtained by recording the intensity reflected from the same sample at a fixed angle of incidence [θ = 47.2°, cf. (a)] while first illuminating with uv light, then after 6 min switching to visible light (VIS), and so on.

with promising optical properties that can be tailored to some extent for future device applications.

2.2.1 Planar Waveguide Structures

We consider the asymmetric-slab configuration. Within Maxwell's theory it is then rather straightforward to arrive for the nonabsorbing case considered here at the well-known eigenvalue equation for waveguide modes of order m (Tien, 1969):

$$k_{2z}d = \tan^{-1}\beta_{21} + \tan^{-1}\beta_{23} + m\pi, \quad (5)$$

with k_{2z} being the wave vector of the mode of order m, d the thickness of the waveguide structure, and $2\beta_{ij} = r''_{ij}/r'_{ij}$ with $r_{ij} = r'_{ij} + ir''_{ij}$ the complex reflection coefficient at the interface between media i and j. For a given set of materials and hence combination of refractive indices (which determines $\tan^{-1}\beta_{21}$ and $\tan^{-1}\beta_{23}$ for a fixed frequency), the number of eigenmodes depends only on the thickness of the waveguide layer.

The detailed analysis of Maxwell's equations can yield the optical-field distribution for each eigenmode (Tamir, 1979). Figure 15 shows the optical-field distributions for the first three modes (with their electric field being parallel to the film along the y axis) calculated with the following indices of refraction: substrate n_3 = 1.46, waveguide film n_2 = 1.55, cladding (air) n_1 = 1. Note that the optical field extends as an evanescent wave into both the cladding and the substrate, where it decays exponentially as a function of distance from the interface, though for this asymmetric slab in an asymmetric way.

The three most important techniques for coupling light into this waveguide are sche-

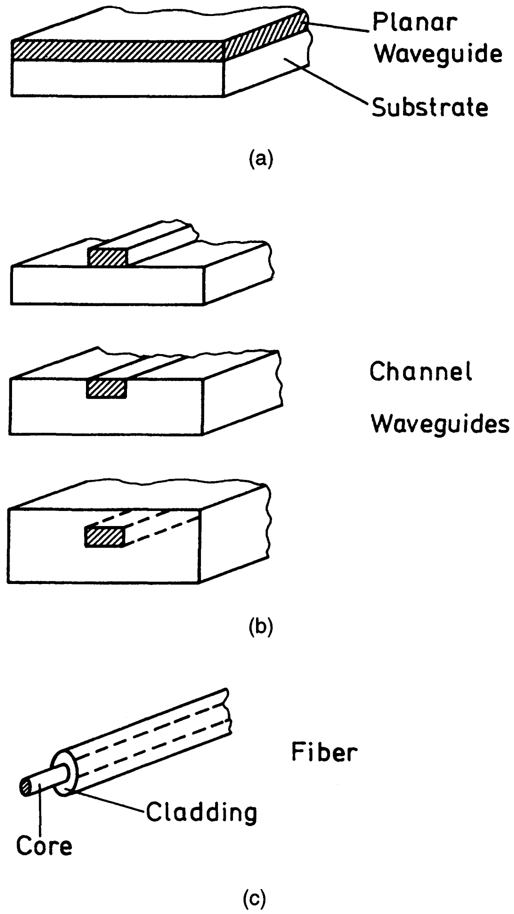

FIG. 14. Basic waveguide structures. (a) Planar waveguide (shaded) on a substrate. (b) Channel waveguides: (top to bottom) strip, embedded strip, buried strip. (c) Optical fiber with the core and the cladding of lower refractive index.

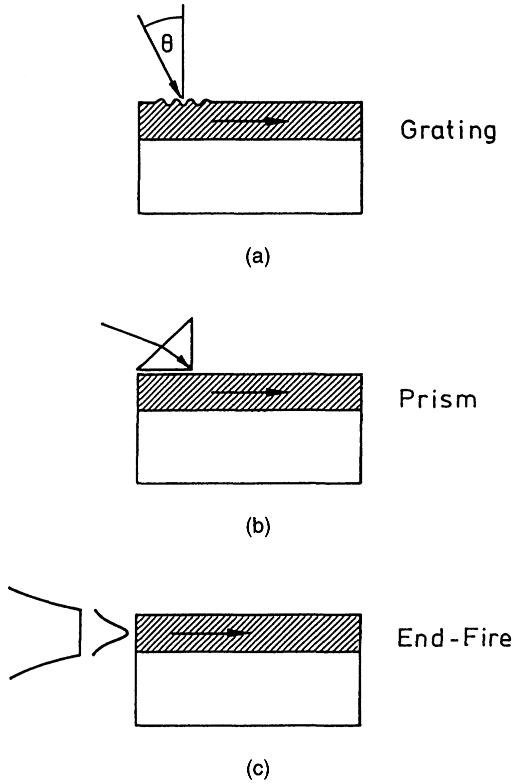

FIG. 16. The three basic coupling geometries. (a) Grating coupling, (b) prism coupling, both of which require the tuning of the angle of incidence, θ; (c) end-fire coupling, e.g., by a microscope objective.

matically sketched in Fig. 16. The grating [16(a)] and prism coupling methods [16(b)] are commonly used for planar waveguides. Efficient coupling can be achieved only under wave-vector matching conditions. For grating coupling, this requires that the photon wave-vector component parallel to the surface plus or minus a multiple of the grating vector $G = 2\pi/\Lambda$, with Λ the groove spacing of the grating, matches the waveguide propagation vector:

$$k_x^m = k_0 n_1 \sin\theta \pm mG. \tag{6}$$

In the case of prism coupling, the projection of the photon wave vector at the base of the prism must be equal to the guided wave vector:

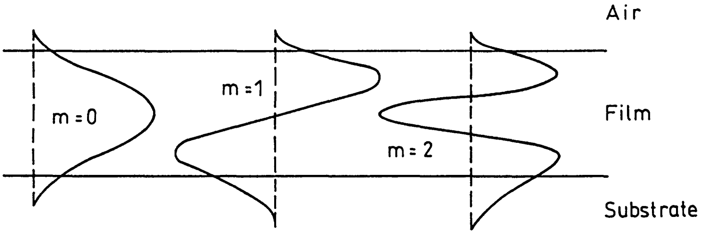

FIG. 15. Optical-field distribution in a planar waveguide structure of substrate–film–air configuration for the first three *s*-polarized modes ($m = 0,1,2$). The higher index of the substrate ($n_3 = 1.46$) compared with the superstrate (air, $n_1 = 1.0$) is the reason for the much stronger evanescent field penetrating into the substrate material.

$$k_x^m = k_0 n_p \sin\theta, \tag{7}$$

with n_p = index of refraction of the prism. This requires a prism with an index higher than the waveguide material.

The end-fire coupling (e.g., from a microscope objective to a channel waveguide) matches the incident-field spatial distribution to that of the guided wave [Fig. 16(c)].

2.2.2 Waveguides from Novel Polymeric Materials For the examples presented in the following, a modified prism-coupling technique as schematically depicted in Fig. 17(a) was used. The waveguide modes were launched with the help of a high-index prism by coupling through a metal (Au or Ag) gap of 50 nm thickness excited with p-polarized light. In this geometry, the $m = 0$ mode is the surface plasmon. Its optical intensity peaks at the metal–waveguide interface and decays (for films of typical waveguide thicknesses of ca. 1 μm or more) completely within the film. Its polarization is predominantly parallel to the film normal. Its angular position, therefore, depends only on n_z but not on the waveguide thickness. The other waveguide modes can be excited with either s- or p-polarized light and will, in general, depend on all three indices, i.e., n_x, n_y, and n_z, as well as on the film thickness [cf. Eq. (5)]. Figure 17(b) shows a series of waveguide modes obtained with p-polarized excitation of $\lambda = 633$ nm in a spin-cast film of a solid polyelectrolyte the structure formula of which is given in the inset. The various modes were measured in reflection. The broad dip in the reflected intensity at the high-angle side ($\theta \simeq 73°$) marks the surface plasmon. The crosses are the experimental data points; the full line (for better clarity somewhat shifted) is a theoretical fit based on a Fesnel calculation (Wolter, 1956). From this one obtains the thickness of the waveguide film $d = 1.85$ μm as well as its (isotropic) index of refraction $n = 1.63$. By recording the reflectivity as a function of the angle of incidence for s-polarized light (not shown), it was found that an isotropic approximation of the refractive index was sufficient in order to characterize the whole waveguide mode pattern.

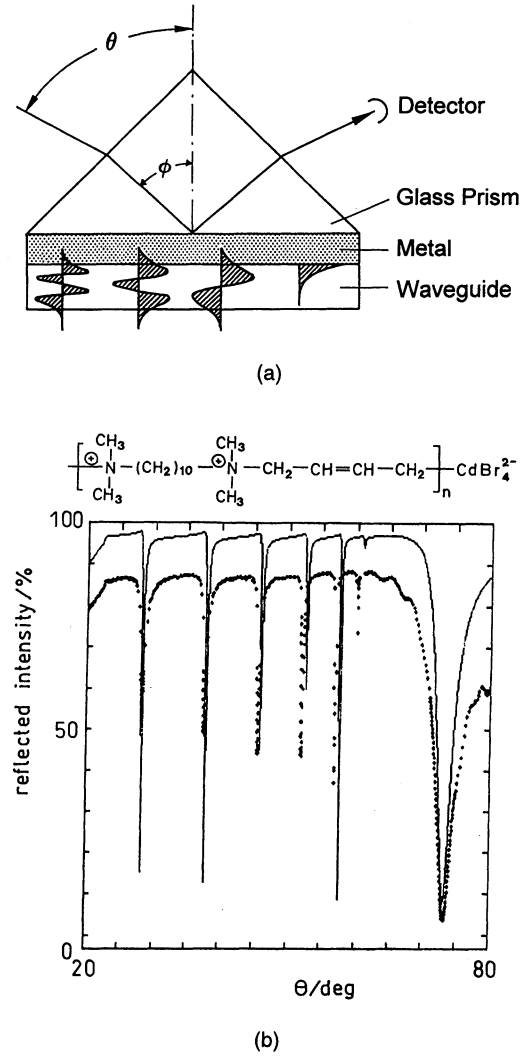

FIG. 17. (a) Attenuated total reflection setup for coupling surface-plasmon as well as waveguide modes into a planar film structure. Coupling is achieved through a thin metal gap at the base of the prism. (b) Reflected intensity recorded as a function of the angle of incidence for a thin film of a solid polyelectrolyte (structure formula given on top) spin-cast onto the Ag-coated (s = 50 nm) base of a high-index prism (SF57) as in (a). The crosses are the experimental data points, the full curve is the theoretical fit. Excitation with p-polarized light from a HeNe laser (λ = 633 nm).

Nevertheless, a closer comparison between the experimental and the theoretical curves of the angular reflectivity shows some distinct differences. First of all, the experimental modes are somewhat broader than calculated with an ideal slab geometry for the waveguiding film. This indicates some thickness and lateral refractive-index heterogeneities within the spin-cast film. In addition, not all calculated angular positions for the various modes fully coincide with the experimental ones. Given the different field-intensity distributions of the different modes (cf. Fig. 15), this discrepancy points to a certain perpendicular refractive-index variation (an index profile) within the waveguide. Both deviations from ideality within the phase and normal to the plane need to be carefully avoided in an IOC structure because such heterogeneities clearly would limit the optimum performance of the device.

2.2.3 Waveguide Loss Measurements

Typically, waveguides are designed to propagate light over long distances. This requires, first of all, the absence of any absorptive losses, either by molecular, electronic, or vibronic transitions or by dissipation, e.g., in metallic materials used as electrodes. Equally important is the control of any scattering losses, which would also result in an enhanced damping of the guided modes. This not only includes the minimization of bulk (volume) fluctuations of the optical density but requires also the ultimate control of surface-scattering processes, e.g., at interfacial roughness caused by processing steps.

One way of characterizing waveguide structures along these lines is to measure the losses of the guided waves directly. Figure 18(a) shows a schematic of an experimental setup designed for loss measurements in planar structures. Laser light is coupled to the waveguide by a grating (or prism). A coherent fiber bundle is used to transmit the light scattered off the guided streak in the planar structure, which is a measure of the local mode intensity. A photomultiplier tube with a slit diaphragm reads this scattered intensity with high spatial resolution and hence allows for the quantitative analysis of the (exponentially) attenuated optical field in the waveguide. A data-acquisition and analysis system gives the loss values typically in dB/cm (1 dB corresponds to an intensity attenuation to 79%).

Figure 18(b) gives the results obtained from a polyelectrolyte thin film prepared by spin-coating onto a low-index glass substrate, measured in air with prism coupling. Data points taken from Mathy *et al.* (1991) were measured with different laser lines and plotted correspondingly as a function of the wavelength employed. Two aspects are noteworthy:

1. For most of the wavelength range investigated, the data seem to follow a λ^{-4} power law [straight line in Fig. 18(b)], which suggests Rayleigh scattering as the predominant attenuation mechanism.
2. For wavelengths $\lambda \geq 650$ nm, losses in the range of 1 dB/cm or less seem to be feasible with these materials, which makes them also technologically interesting.

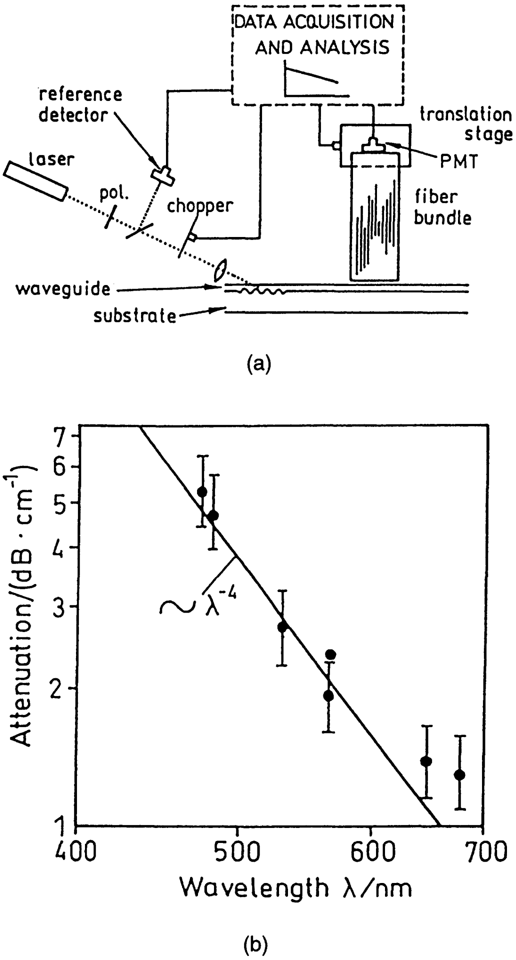

FIG. 18. (a) Schematic of the setup used for the determination of mode losses in waveguide structures. Laser light is coupled to the planar waveguide by means of a grating or prism. A fiber bundle records the local level of stray light, which is a measure of the field intensity along the guided streak. (b) Loss data for *s*-polarized (m = 0) modes (given as attenuation in dB/cm) for various wavelengths as obtained for a planar waveguide structure fabricated by spin-casting a solid polyelectrolyte. The straight line is a λ^{-4} power law. Data from Mathy *et al.* (1991).

2.2.4 Supramolecular Design of Waveguide Films by the Langmuir–Blodgett–Kuhn Technique Finally, we should briefly discuss possibilities for the ultimate control of a planar waveguide structure at the molecular level as is in principle possible by the LBK method. Here, multilayer assemblies capable of carrying guided modes are built up, layer by layer, by dipping and withdrawing a substrate many times through a monomolecular layer prepared at the water–air interface (Swalen *et al.*, 1978). The number of possible molecular structures of the individual layers is nearly infinite, and so are the functionalities that can be tailored into these systems. Even if it is unlikely that the tedious preparation process ever will make it into a large-scale production line, these LBK films nevertheless are extremely helpful as model systems to develop and test molecular structural concepts, functional units (like nonlinear chromophores, etc.), supramolecular assemblies, and device configurations.

As an example for a simple passive planar waveguide structure prepared by this LBK technique, we present in Fig. 19 results obtained with a multilayer assembly of the polyglutamates previously presented. Figure 19 demonstrates the good optical quality of a thick multilayer assembly (150 layers) capable of guiding modes with different polarizations; the upper frame was taken with *p*-polarized light, while the lower frame shows data with *s* polarization. The full circles are experimental points measured with waveguide modes propagating perpendicularly to the dipping direction, the crosses are for parallel propagation. The full lines are Fresnel fits and show the excellent theoretical description of these structures as planar slabs. The optical in-plane anisotropy can be analyzed quantitatively and amounts to only $\Delta n = n_{\parallel} - n_{\perp} = 0.01$. This demonstrates the sensitivity of the coupling conditions to subtle refractive-index variations.

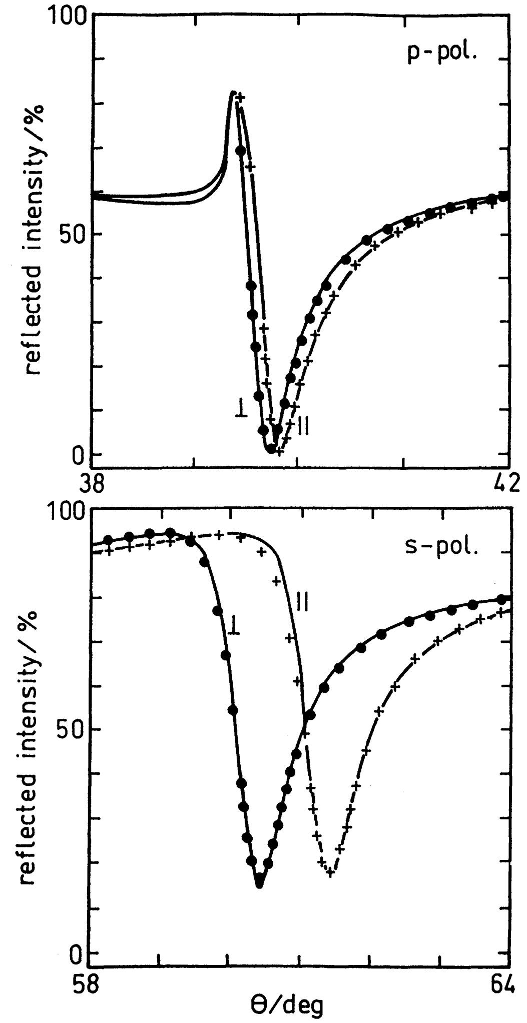

FIG. 19. Waveguide mode pattern as obtained from a polyglutamate multilayer assembly of 150 layers for (upper frame) *p*- and (lower frame) *s*-polarized light. Full circles are obtained for modes propagating perpendicular to the dipping direction, which is the preferred axis of the hairy rods. Crosses are from modes propagating parallel to the rods. The full curves are Fresnel calculations showing the corresponding in-plane anisotropy of the index of refraction.

2.2.5 Channel Waveguides The theoretical description of guiding light in a two-dimensional waveguide structure [optical confinement in two dimensions like in the channel waveguides shown in Fig. 14(b)] adds nothing new, and just requires more complex numerical techniques for the calculation of the exact dispersion relation and hence the various modal structures. These channel waveguides are eventually the ultimate structures needed for any IOC. This includes, in particular, active devices based on materials with nonlinear optical effects. Two basic structures are shown in Fig. 20. The (electro-optical) coupler, in which two channel waveguides are sufficiently close that they can interact like two weakly coupled mechanical oscillators, is shown in (a). The

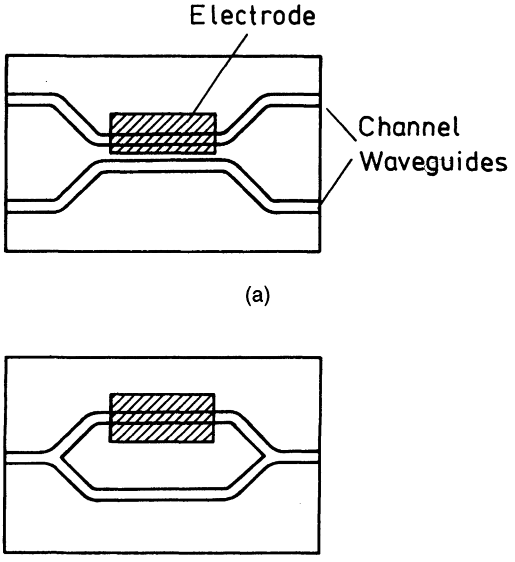

(a)

(b)

FIG. 20. Basic structures for integrated optical circuits (with special emphasis on electro-optical control): (a) channel waveguide coupler, (b) (integrated) Mach–Zehnder interferometer.

strength of the interaction that determines, e.g., the flow of light intensity from one channel to the other can be controlled by the application of a dc electric field through the Pockels effect provided the polymeric waveguide material has a correspondingly large second-order susceptibility $\chi^{(2)}$. The integrated Mach–Zehnder interferometer shown in (b) allows for a phase shift of the light in one branch relative to that propagating in the other channel. This then leads to constructive or destructive interference at the output. Again, this interaction can be controlled either by $\chi^{(2)}$-active materials through the application of an electrical field [electrodes are shown in Fig. 20(b)] or by a material with an intensity-dependent refractive index n_2 (see below).

2.2.6 Polymer Optical Fibers (POF)

Single-mode glass optical fibers are widely used as long-distance and high-bandwidth communication media because of their excellent transparency. In short-distance networks, however, where many junctions between different fibers are necessary, already slight displacements of two coupling cores each with a diameter of only about 5 μm would cause substantial coupling losses. Polymer optical fibers (POFs) with their considerably larger core diameter (ca. 1 mm) are in this respect much easier to handle. In addition, they offer other advantages like low cost, easy processing, and flexibility. However, all commercially available POFs have been of the step-index (SI) type [Fig. 21(a)]. For these multimode fibers, the different guided waves all travel with the same phase velocity but along different paths within the core, a phenomenon called modal dispersion. This is schematically depicted in Fig. 21(a). As a result of internal mode coupling, even a single input pulse [cf. Fig. 22(a)] is greatly distorted and considerably broadened. An example is given in Fig. 22(b). This means that two consecutive input pulses need to be separated in time in order to be resolved at the output as two individual pulses. The transmission bandwidth is thereby greatly reduced and currently amounts for SI-POF to ca. 5 MHz km.

A solution to this problem is the graded-index (GI) fiber. Here the refractive-index profile with the highest values in the center of the core results in a higher phase velocity of those "rays" traveling in the outer periphery of the core. Their longer pathways are therefore compensated by a faster propagation speed. As a consequence, the input pulse is far less broadened, which results in a higher bandwidth. An example that was re-

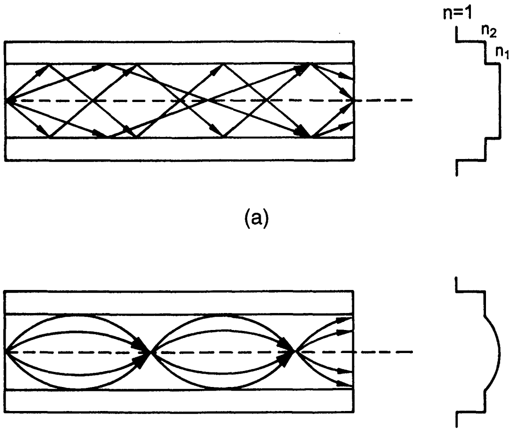

(a)

(b)

FIG. 21. Typical optical rays and refractive-index profile of (a) a multimode step-index (SI) fiber, and (b) a graded-index (GI) fiber.

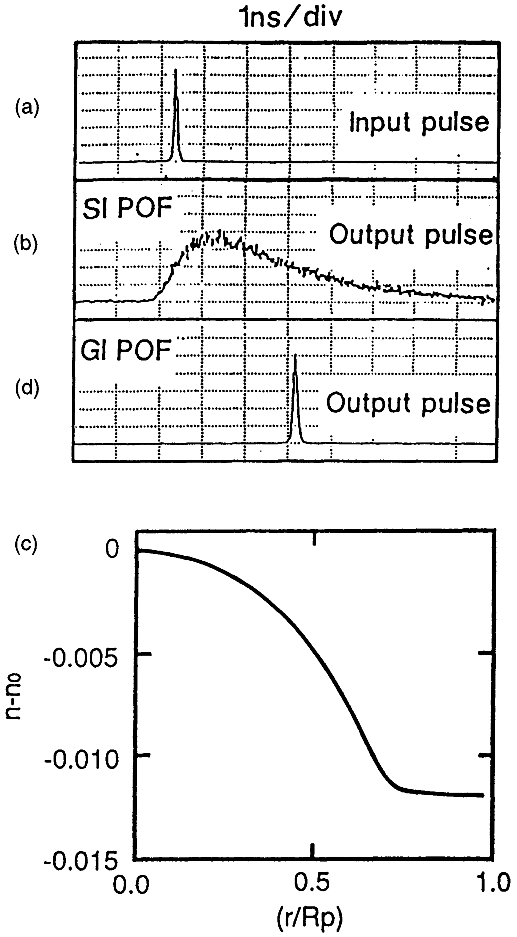

FIG. 22. (a) Short input pulse in a POF; (b) output pulse after traveling through a step-index fiber; (c) index profile of the graded-index fiber that transmits the input pulses nearly unbroadened as shown in (d). (After Ishigure *et al.*, 1992.)

cently reported for Gi-POFs prepared by a special copolymerization process (Koike, 1991) is given in Fig. 22: (c) shows the refractive index profile, (d) compares the input pulse with the output pulse. The bandwidth obtained was shown to be ca. 2 GHz km (Ishigure *et al.*, 1992).

Another improvement in the development of polymer optical fibers concerns the intrinsic losses given by absorption processes. This is indicated in Fig. 23 for a fiber made of poly(methylmethacrylate) (Groh *et al.*, 1989). The higher overtone vibrations of the C–H groups in the material are responsible for attenuation losses at $\lambda \approx 680$ nm of more than ca. 200 dB/km. Perdeuteration results in a substantial reduction of the band strengths but also in a shift of their spectral positions. Both effects reduce the loss in these fibers considerably (Fig. 23). Other attempts to reduce the absorption coefficient at relevant communication wavelengths successfully introduced fluorocarbon moieties in the material with the additional improvement of the maximum operating temperature. This is an important issue, e.g., for the implementation of POFs in automobiles (Groh *et al.*, 1989).

3. NONLINEAR OPTICAL PROPERTIES OF POLYMERS

Polymers with large nonlinear optical susceptibilities are likely to play a major role in the rapidly growing area of photonics. We concentrate in the following on some general fundamental device concepts, on materials design considerations, and on some polymer-specific issues relevant for nonlinear optics.

We first give a few remarks as to the functionalization of polymers for their use in nonlinear optical device configurations (Zyss, 1985). We then focus on some applications related to the second-order nonlinear optical susceptibility $\chi^{(2)}$ and finally summarize some third-order $\chi^{(3)}$ effects and how they could be implemented into future technologies for information processing.

3.1 General Concepts for the Functionalization of Polymers for Nonlinear Optics

The nonlinear optical response of polymers developed for future IOCs needs to be optimized through a suitable functionalization of the matrix by chromophores with correspondingly large second- or third-order susceptibilities. The two processes need, however, different strategies for an optimized material (Kajzar, 1992).

For the second-order processes, the relation between the macroscopic averaged (denoted by angular brackets) nonlinear susceptibility $\chi^{(2)}$ and the molecular nonlinear polarizability β_D is determined by the following equation (Williams, 1987; Singer *et al.*, 1987):

$$\langle \chi^{(2)} \rangle = f\beta_D N_D \langle \cos^3\vartheta \rangle, \qquad (8)$$

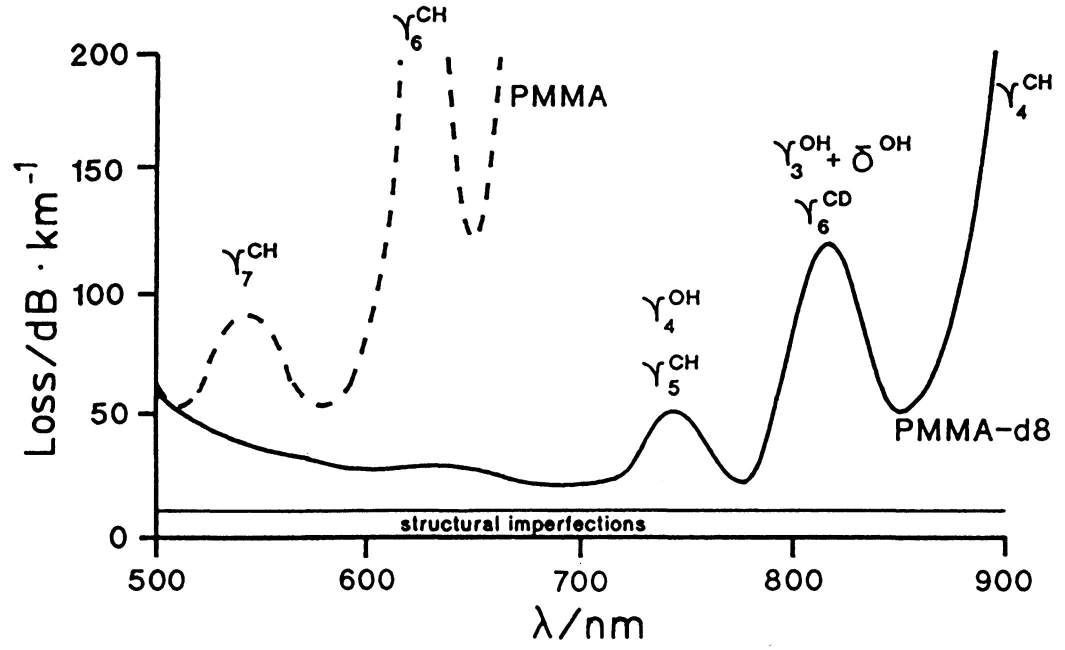

FIG. 23. Spectral-loss data for a regular PMMA POF (dashed curve) and for a fiber made from perdeuterated PMMA (full curve). Some vibrational band assignments are given. (After Groh *et al.*, 1989.)

where N_D is the density of the active molecules, f is a local-field factor of the optical field, and $\langle \cos^3 \vartheta \rangle$ is a measure of the order parameter describing the average orientational distribution of the chromphore's nonlinear optical axis relative to the applied field. Maximum $\chi^{(2)}$ is obtained if all factors in Eq. (8) are optimized.

The first step is the choice of a chromophore system with a large β_D. The most promising candidates are dyes of the charge-transfer type with a strong donor and acceptor group linked together by a conjugated π-electron system (Singer *et al.*, 1989). This leads to a strong and highly asymmetric polarizability and hence to a large nonlinearity. By now, hundreds of different chromophores have been synthesized and tested, some of which are given in Table 4 (cf. also Burland *et al.*, 1991). This compilation demonstrates that by a systematic variation of the donor and acceptor strengths and of the conjugation lengths substantial improvements in β_D can be achieved.

The next step, the maximum level of functionalization, concerns the question of how to dope the polymer matrix. Three strategies have been suggested. The first is the mere mixing of a suitable dye as the functional guest in a polymeric host material. One example of a suitable dye is presented in Fig. 24 (Levenson *et al.*, 1991). The dye's structure formula is shown above Fig. 24. The addition of 15 wt% dye to the PMMA matrix changes the linear optical properties of the polymer: The relatively flat background index of refraction of PMMA [full curve in (a)] shows for the mixture the resonant dispersion behavior (broken curve) of strong chromophores. The corresponding behavior is seen for the extinction coefficient α [Fig. 24(b)]. This material showed (after poling, see next section) decent $\chi^{(2)}$ values and could successfully be implemented into an electro-optic phase modulator.

A major drawback of the host–guest systems is the limited solubility of the dyes in the polymer: typically only 15–20 wt% can be mixed without phase segregation and crystallization. Another problem, in particular in view of device applications, arises from the typically low temporal stability of the noncentrosymmetric orientation distribution induced by the poling process (see below). Depending, of course, on the storage temperature, decay times of a few hours to

Table 4. Molecular control of β_D.

Structure	β_D at 1.9 μm ($\times 10^{-30}$ esu)
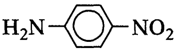	5.7
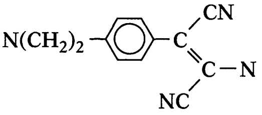	21.4
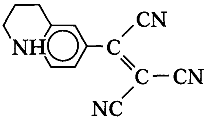	41.8
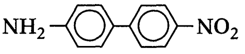	20.1
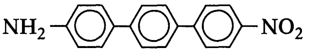	50.7
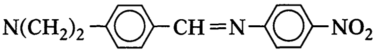	23.4
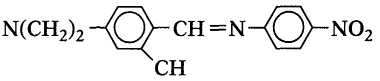	61.6
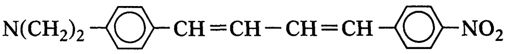	111.2

days have been reported (Page *et al.,* 1990). However, certain host–guest systems based on polyimide matrices have been demonstrated to exhibit stable $\chi^{(2)}$ values for hours even at temperatures as high as T = 300 °C (Wu *et al.,* 1991).

The aforementioned problems have led to the development of polymers where the chromophores either were an integral part of the main chain (Köhler *et al.,* 1990) or, more widely reported, were attached to the host polymer as functional side chains. A small selection of the many systems presented in the literature is given in Fig. 25; only examples based on acrylates or methacrylate polymers are added. All these attempts were aimed at fabricating a material that can easily be processed, e.g., in the form of a glassy film or channel waveguide, that shows a high β_D value of the employed chromophores, that allows for a high content of nonlinear-optically (NLO) active moieties, and that gives a stable configuration when poled (Man and Yoon, 1992). In fact, these copolymers allowed for a much higher functionalization of the material than for the host–guest systems.

Yet another approach recently reported (Eich *et al.,* 1989; Jungbauer *et al.,* 1991; Swalen *et al.,* 1991; Chen *et al.,* 1991) currently seems to be the ultimate solution for a stable chromophore orientation distribution: the incorporation of the NLO-active moieties into the polymer network as an integral part of the cross-linked matrix. One of the systems is shown in Fig. 26 (Swalen *et al.,* 1992): An epoxy monomer (a) is made to react with electric-field–oriented NLO-active amine monomers (b) to give a highly functionalized cross-linked and hence stable $\chi^{(2)}$ material (c). Very good long-term stabilities have been achieved (see below).

For the third-order nonlinear optical processes, polymeric materials with a high degree of delocalization of the conjugated electron cloud seem to be the most promising candidates. This leads to a large component γ_{xxxx} of the molecular hyperpolarizability tensor γ_D along the polymer chain (conjugation)

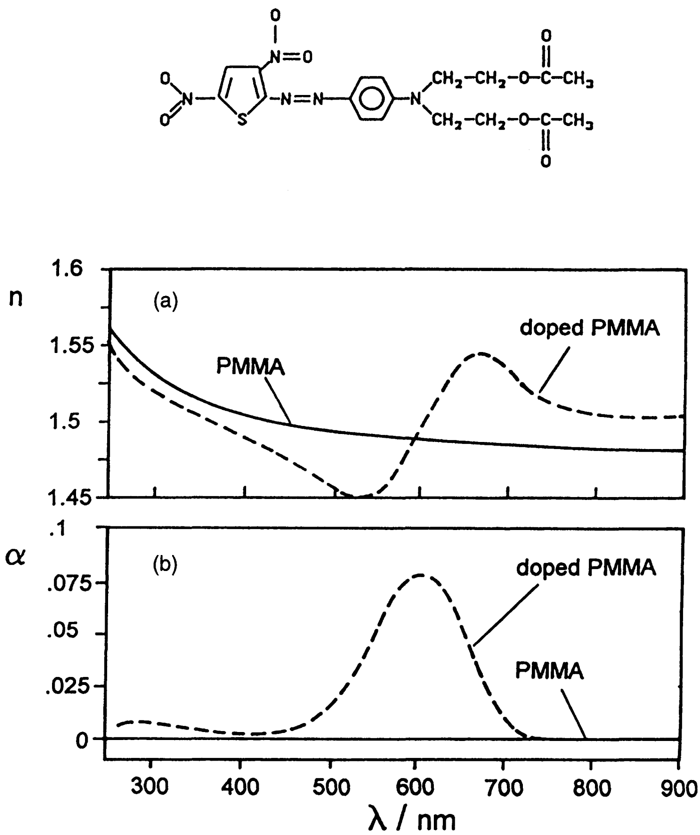

FIG. 24. Example for a host (PMMA)–guest (TP, structure formula given above figure) system. (a) The wavelength dependence of the refractive index *n* of pure PMMA (full line) and after addition of 15 wt% TP (dashed line); (b) the corresponding absorption coefficient (after Levenson *et al.*, 1991).

direction. Some examples, currently studied in many laboratories, are given in Fig. 27 (see, e.g., Jenekhe *et al.*, 1992, and references therein). According to

$$\langle\chi^{(3)}\rangle = f\gamma_{xxxx}N_D\langle\cos^4\vartheta\rangle, \tag{9}$$

this is one prerequisite for a high value of the macroscopic nonlinear susceptibility $\chi^{(3)}$. Another important issue also for third-order materials is the orientational order of the chromophores described by $\langle\cos^4\vartheta\rangle$. Although not as crucial as for the $\chi^{(2)}$ devices, which require a noncentrosymmetric arrangement of the NLO-active groups, here, too, the efficiency of the material depends on a good orientation: $\langle\cos^4\vartheta\rangle = 0.2$ for a three-dimensional disorder whereas $\langle\cos^4\vartheta\rangle = 1$ for a system where all the polymer chains are parallel to a given direction. This means that the internal order can be responsible for substantial enhancements of the device performance. The packing density of the chromophores that determines N_D [cf. Eq. (9)] is another factor that needs to be optimized for a large $\chi^{(3)}$ but has to be balanced against other materials and processing parameters. For example, most of the chromophores presented in Fig. 27 are nearly intractable, i.e., they are largely insoluble in most solvents and cannot be prepared, e.g., in thin-film form. This has led to many more or less successful routes to a final product via some soluble and hence processable precursor materials. One example is given in Fig. 27: here, a tetramethylene sulfonium chloride precursor polymer is converted as a spin-cast thin film by a thermal treatment to the final poly(*p*-phenylenevinylene) system (Lenz *et al.*, 1990), which showed decent γ_D values (Bubeck *et al.*, 1989). Another trend in current $\chi^{(3)}$-materials research concerns the incorporation of molecular structures with a two-dimensional π-electron delocalization. Some examples are given in Fig. 28 (Sasabe *et al.*, 1990; Schrader *et al.*, 1991). The ob-

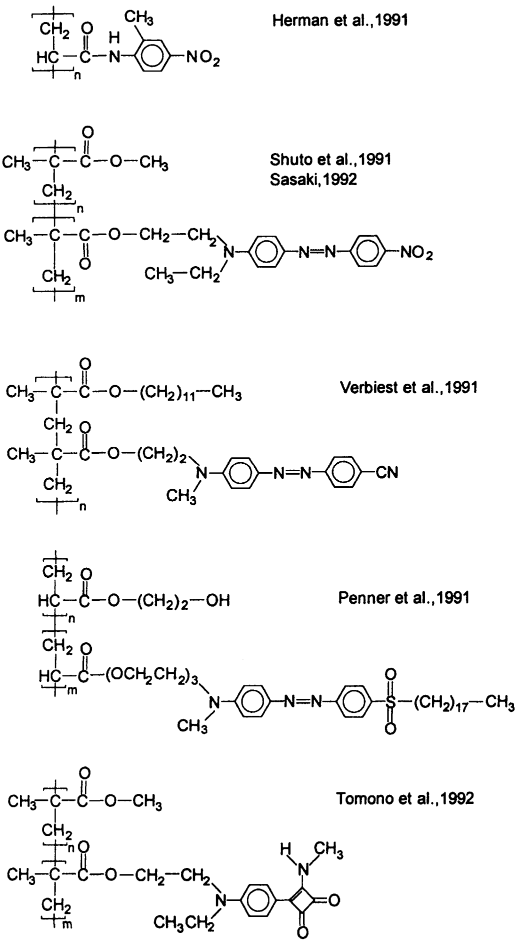

FIG. 25. Examples for different dye-substituted side-chain polymers.

tained nonresonant $\chi^{(3)}$ values, however, are not too promising as yet (Schrader *et al.*, 1991).

3.2 Second-Order Nonlinear Optics with Polymers

Among the numerous device configurations based on $\chi^{(2)}$ processes that have been realized and demonstrated in various laboratories, we focus only on two very basic concepts, i.e., a polymeric waveguide structure designed for second-harmonic generation (SHG) and an electro-optic (EO) modulator that is the key unit for many switching devices. Both modules require that the NLO-active chromophores be highly oriented in a noncentrosymmetric narrow distribution.

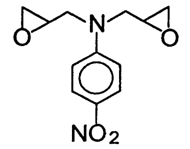

Epoxy Monomer

(a)

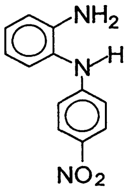

Amine Monomer

(b)

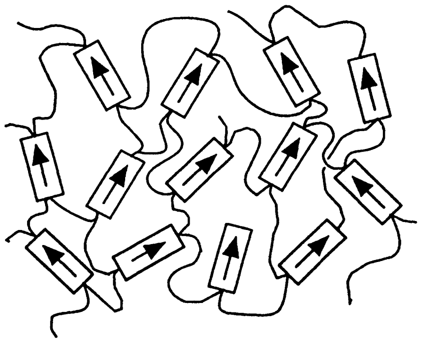

(c)

FIG. 26. Examples for a cross-linkable system composed of (a) bifunctional matrix monomer and (b) a trifunctional $\chi^{(2)}$-active chromophore. (c) A schematic of the network obtained by curing these two constituents.

3.2.1 Control of Orientational Order For polymeric materials, two strategies have been developed for obtaining oriented samples. The first is based on the LBK technique. For the usual *Y*-type deposition where one monolayer (*A*) is transferred on the downstroke and the other (*B*) on the upstroke, this requires a double-trough arrangement with the $\chi^{(2)}$-active layer being transferred, e.g., only on the upstroke whereas an optically inert material is deposited on the downstroke (Hsiung *et al.* 1991; Penner *et al.*, 1991; Aust *et al.*, 1993). A different approach

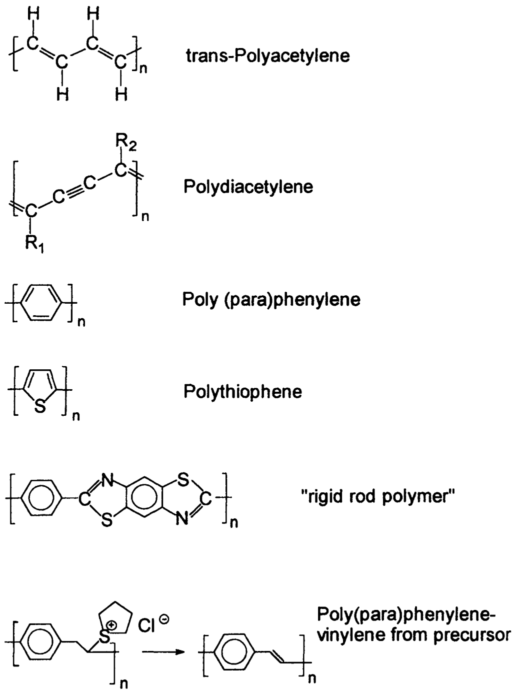

FIG. 27. Structure formulas of some selected $\chi^{(3)}$-active polymers.

to the problem of generating noncentrosymmetric structures by the LBK method has recently been reported for molecular arrangements of 2-docosylamino-5-nitropyridine, the chromophores of which are arranged in a plane parallel to both dipping directions with a net polar orientation giving rise to a large $\chi^{(2)}$ activity (Bosshard *et al.*, 1991). If two different types of chromophores with opposite β_D components are used for the *A* and for the *B* layer, respectively, then it is possible to prepare LBK multilayer assemblies in the *Y* mode with a very high chromophore density (Cresswell *et al.*, 1990). In passing, we should note that self-assembly multilayers with their intrinsically noncentrosymmetric arrangement have also been successfully introduced for $\chi^{(2)}$-active structures (Allan *et al.*, 1991).

All these systems, however, have in common that, in addition to the NLO-active chromophores, highly specialized molecular moieties need to be incorporated in order to tune the molecule's properties to the desired character (to allow for a self-assembly process, e.g.) (Bubeck *et al.*, 1991). From a technical point of view, it is also highly unlikely that such a time-consuming process as the LBK deposition of multilayers will ever make it to a commercially competitive device-fabrication level.

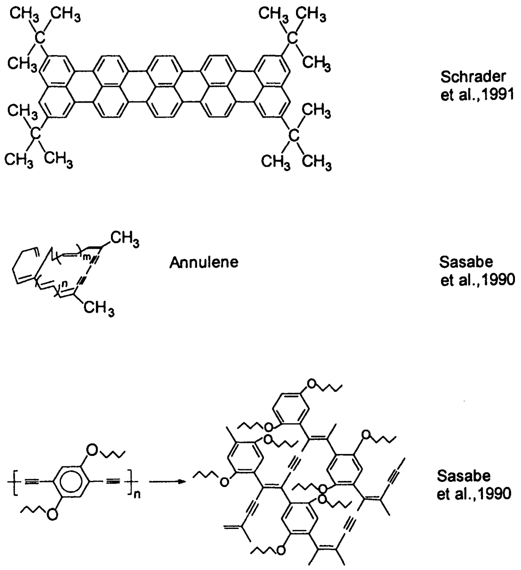

FIG. 28. Some examples for systems with a near two-dimensional π-conjugation.

Therefore, the alternative strategy, i.e., starting from a functionalized polymer with random chromophore orientation spin-cast onto a solid substrate and then subsequently oriented by a poling process, is certainly the more promising and hence more widely applied approach. In the following, we briefly review some aspects of the poling process and its stability.

Two of the basic experimental configurations used for the static electric-field–induced noncentrosymmetric orientational order of the chromophores in functionalized polymer thin films are summarized in Fig. 29. The complete EO-waveguide structure shown in (a) includes, in addition to the NLO-active core, the two buffer layers that isolate the guided mode from any absorptive losses in the (metallic) electrodes. The latter will be used in the final device as the control electrodes but can also be used to apply a dc field of the order of some 10 MV/cm during the poling process at elevated temperatures (see below). The top electrode-free poling process based on a corona discharge setup is sketched in Fig. 29(b). At present, it is not yet clear which approach—electrode or co-

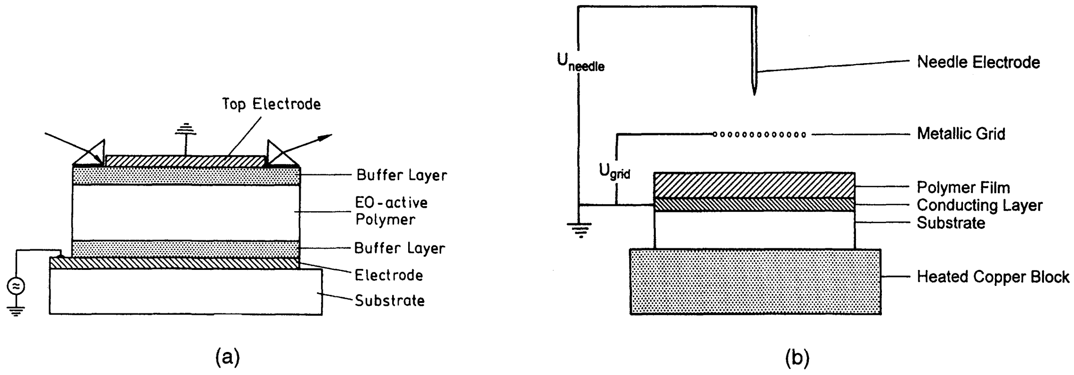

FIG. 29. (a) Schematic cross section of a polymer layer with electro-optic activity sandwiched between two buffer layers and two electrodes. This configuration can be used for (out-of-plane) poling and for (simultaneous) waveguide spectroscopy. (b) Schematic of a setup used for corona (discharge) poling.

rona poling—will yield the better results, i.e., produce a higher degree of orientation, better homogeneity, tolerance against defects, etc. (Herminghaus *et al.*, 1991).

In any case, the poling process requires an orientational (rotational) flexibility of the chromophores in the polymer matrix. This is achieved by heating the sample to a temperature close to the glass-transition temperature T_g of the material. Details of the actual temperature–electric field (applied voltage) protocol differ from laboratory to laboratory but certainly depend also on whether one is dealing with a host–guest, main-chain, or side-chain polymer or a cross-linkable system that can be cured by a thermal process. For the last case, a typical protocol is given in Fig. 30 (Swalen *et al.*, 1991). Without any voltage applied, the system is heated to a temperature where cross-linking is initiated, leading to a slight increase of T_g. This precuring period is followed by the actual poling process, which in the presence of the orienting field further cross-links the polymer–chromophore network with a concomitant increase of the glass-transition temperature.

The poling process can be followed on line by recording either the second-harmonic generation or electro-optic efficiency or, very simply, by monitoring the linear absorption data. Once the polymer sample is cooled down again to room temperature, this polar anisotropy and hence the noncentrosymmetric $\chi^{(2)}$ activity is frozen in—however, not completely, depending on the system. For device purposes this is, of course, an issue of utmost importance and hence has been the subject of many studies. With some exceptions (discussed already above), the general behavior can be classified as shown in Fig. 31: the host–guest systems show the fastest back-relaxation to a more or less isotropic distribution. Much better results are obtained with chromophores covalently attached to the polymer backbone, although clearly this reduced flexibility also may result in a reduced net orientation obtainable through poling. The best stability is found for the cross-linked system.

It should be pointed out that most of these results are obtained, at least to some extent, by a trial-and-error approach. Many of the details are not understood yet, and in many cases only very crude models have been treated. What is certainly missing is a better understanding of the correlation of these linear and nonlinear optical properties with other polymer-specific characteristics given by the structural and dynamical features of these functionalized materials.

3.2.2 Second-Harmonic Generation (SHG)

One of the key device applications of $\chi^{(2)}$-active polymers is the generation of second-harmonic light. The driving force is the search for efficient blue-light sources, in this case through frequency doubling of red laser-diode emission, e.g., for use in optical-storage concepts. The basic configuration is given in Fig. 32(a). Laser light of frequency ω, incident on the sample of thickness d at an angle θ with a fixed polarization, is refracted at the surface of the nonlinear medium and propagates through the film.

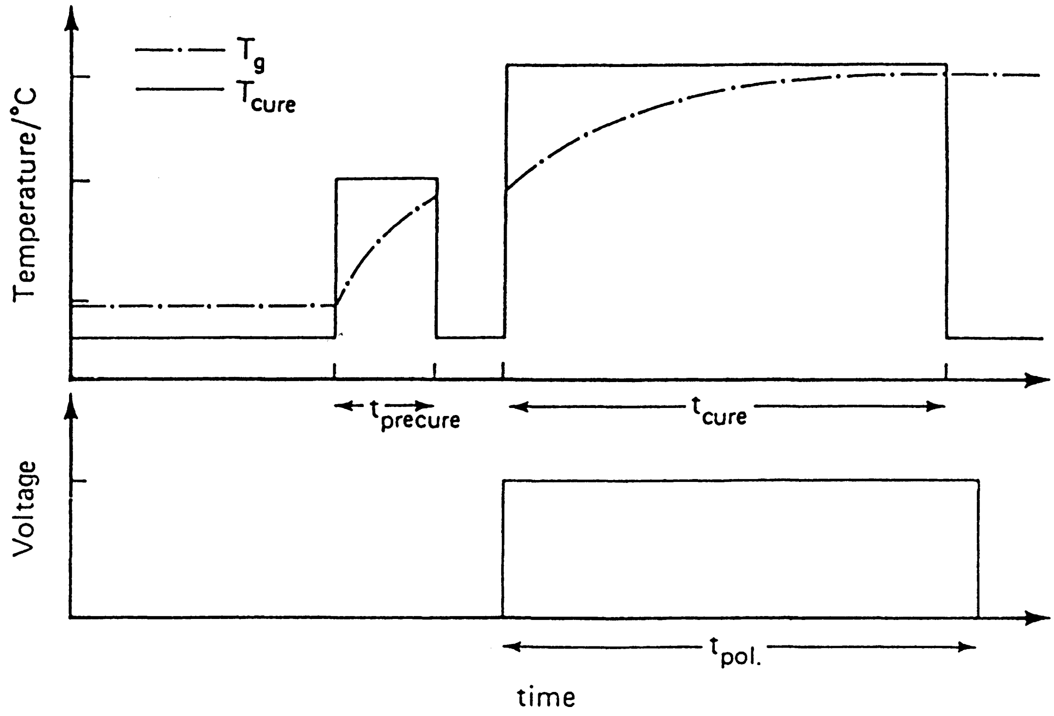

FIG. 30. The temperature–voltage protocol for a dye–polymer system that undergoes a thermally induced cross-linking reaction with a simultaneous increase of the glass-transition temperature T_g. (After Swalen *et al.*, 1991.)

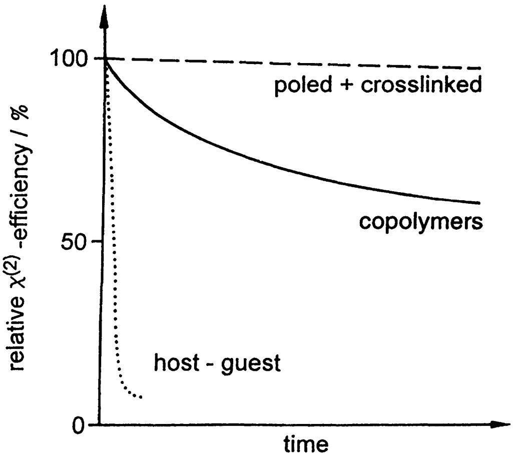

FIG. 31. Qualitative stability behavior of poled polymer films. Typically, the host–guest mixed system randomizes very fast; the copolymer system has a high residual orientation even after long times; the cross-linked system has a long-term stability.

The second-harmonic wave follows the fundamental beam and as such is called the bound wave (Jerphagnon and Kurtz, 1970). In addition, a second-harmonic wave is generated at the surface of the film; it is, however, refracted at a different angle because the refractive index $n(2\omega)$ of the NLO material is different from the value at the fundamental, $n(\omega)$. This wave is called the free wave. The two waves interfere to give the overall SH intensity, with an oscillatory pattern as one rotates the sample called Maker fringes (Ledoux *et al.,* 1987; Verbiest *et al.,* 1991). Depending on the thickness of the sample, several minima and maxima can be observed [Fig. 32(b) for a thick film, 32(c) for a thin film after Swalen *et al.,* 1992].

Now, the required high laser-light intensity for efficient nonlinear conversion is best realized in a waveguide configuration with

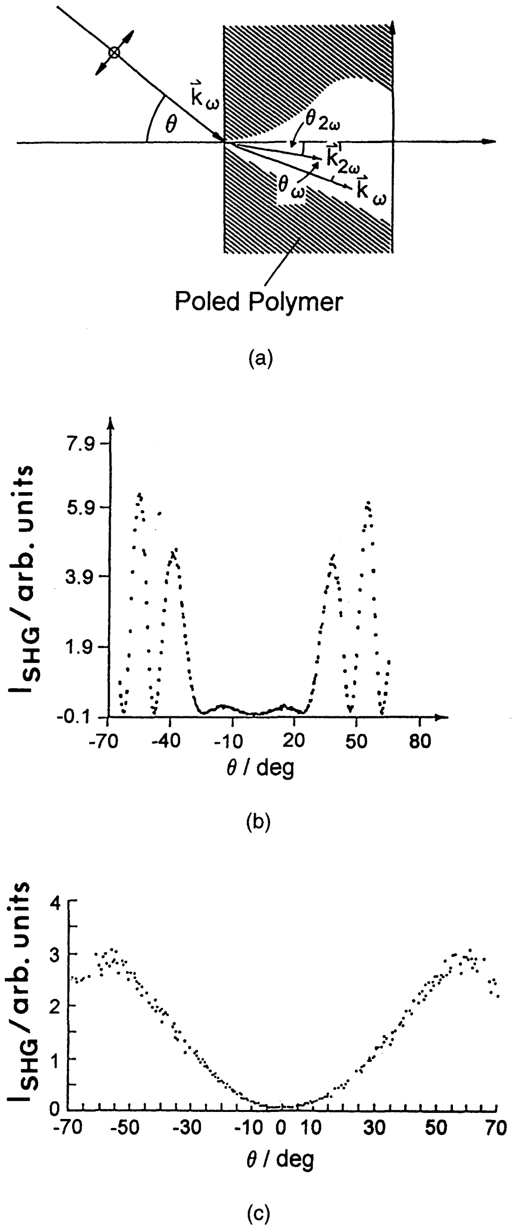

(a)

(b)

(c)

FIG. 32. (a) Schematic representation of the propagation vectors whose interference leads to the Maker fringe pattern; (b) Maker fringes for a thick polymer film; (c) one Maker fringe for a thin poled polymer film. (After Swalen *et al.*, 1992.)

strong optical-field confinement. The ultimate structure, therefore, will be the second-harmonic–generating channel waveguide (Norwood and Khanarian, 1990). However, the dispersion of the material and of the waveguide structure (modal dispersion) would result in a useless device because after a certain propagation length of the fundamental and the second-harmonic beam (called the coherence length L_c), their relative phase difference would lead to destructive interference of the second-harmonic light. Several concepts have been put forward to overcome this problem called the phase-matching problem. The first approach is based on the modal dispersion of waveguide structures: This term accounts for the fact that different modes (modes of different order) propagate with different phase velocities, and hence can be described by different effective refractive indices. Phase matching in such a concept is achieved by tuning the dimensions of the waveguide, e.g., the thickness of a planar structure, in such a way that the dispersion of the material is compensated by the modal dispersion. This then results in a configuration in which the guided fundamental mode of order m propagates with the same velocity as the second-harmonic mode of order m'. Problems arising from the need to optimize the field-intensity overlap integrals have so far hampered a major breakthrough of this approach, but can be solved.

The second concept, called quasi-phase matching, is schematically depicted in Fig. 33. After the first coherence length in the poled polymer waveguide, the $\chi^{(2)}$ activity is either destroyed (by bleaching or randomization of the chromophores) or reversed in its polar direction (by the use of correspondingly patterned and poled electrodes) [Fig. 33(a)]. The first configuration is called unidirectional poled waveguide, the second bidirectional. Both overcome the problem of destructive interference of the SHG light with the bidirectionally poled structure showing a twice as efficient buildup of the 2ω signal [Fig. 33(b)].

Finally, we should mention a third version called the Cherenkov frequency doubler. Here, the refractive indices of waveguide material and substrate are such that the fundamental light is guided in the film, but that the SH light is constantly coupled out through the substrate (because the condition for total internal reflection at the film–substrate interface is not fulfilled at 2ω). This way, also no destructive interference limits the SHG efficiency.

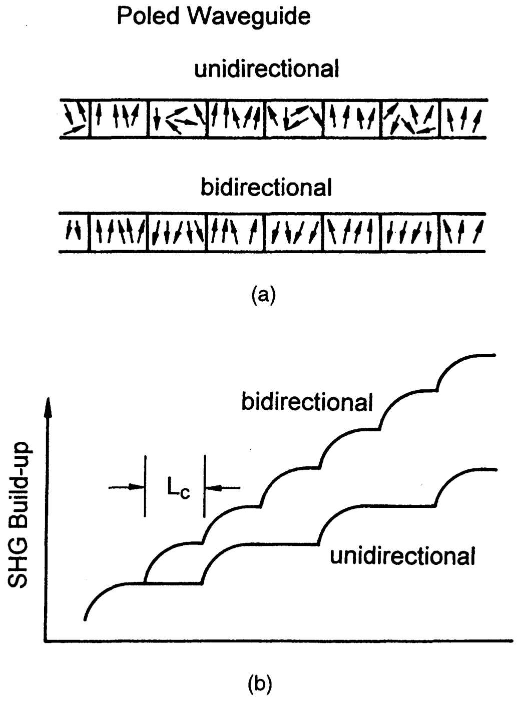

FIG. 33. (a) Schematic of unidirectional and bidirectional poling of waveguides for quasi phase-matching and (b) the resulting buildup of second-harmonic light in the waveguide.

3.2.3 Electro-optic (EO) Modulation

The second major class of $\chi^{(2)}$-based devices uses the Pockels effect, i.e., the manipulation of the refractive index of the material by an applied electric (dc) field. The ultimate structure, again, is certainly a channel waveguide (Shuto *et al.*, 1991) with an architecture schematically shown in Fig. 29(a). If only materials parameters are to be measured, various simplified architectures are sufficient. One setup operating in a reflection mode similar to ellipsometry is given in Fig. 34 (Teng and Man, 1990). The highly reflecting Au or Al electrode on the back side of the EO-active thin film is used to measure differences of the reflectivities for *s*- and *p*-polarized light propagating through the poled polymer in the absence or presence of an applied electric field, respectively. The differences of the reflected-light intensities can be quantitatively analyzed in terms of the two $\chi^{(2)}$ tensor components, $\chi^{(2)}_{13}$ and $\chi^{(2)}_{33}$, needed to describe the nonlinear (Pockels) response of a uniaxial film (Gadret *et al.*, 1991). The advantages of such a simple setup for NLO materials characterization relative to a SHG setup are obvious: No high-power laser is needed, and, therefore, wavelength-dependent measurements are much easier.

A more detailed analysis of the linear and nonlinear optical properties of planar thin-film samples is possible by surface-plasmon (Cross *et al.*, 1988) and by waveguide spectroscopy outlined in Secs. 2.1 and 2.2.1. This has been demonstrated for many samples of poled polymers. In particular, if electrostrictive effects are also present in the investigated material only the waveguide technique allows for a separation of the different processes (Dumont and Levy, 1989). It is straightforward to demonstrate the potential of the method for a waveguide structure prepared by the LBK technique. The example presented in Fig. 35 was obtained with a sample that consisted of 240 buffer layers of an inert material on each side of the EO-active *A*–*B* layer system of 20 layers each. The structure formula of the corresponding chromophore is given in the inset of Fig. 35. The three waveguide modes (m = 0,1,2 as indi-

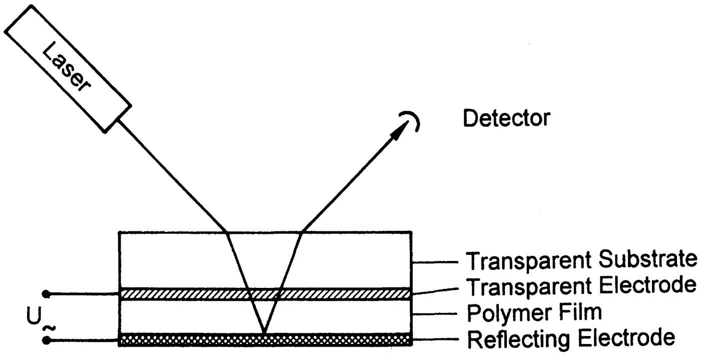

FIG. 34. Setup for EO-coefficient measurement in reflection.

cated) measured with s-polarized light (λ = 593 nm) are shown in Fig. 35(a). Figure 35(b) gives the electro-optic response of the sample under the influence of an applied voltage of 30 V peak to peak. Most remarkable is the alternating magnitude of the effect: The $m = 1$ mode shows only a very small modulation. The interpretation is obvious on the basis of the field-intensity distribution given in Fig. 15. The $m = 1$ mode has a node in the middle of the waveguide right where the EO-active material is located. Any refractive-index modulation by the applied electric field influences, therefore, the overall index configuration for this mode only very little. The even-indexed modes, however, with their intensity maximum in the middle, are influenced much more strongly. The quantitative analysis gives for all three modes the same electro-optic coefficient, $\chi^{(2)}_{13} = 14$ pm/V.

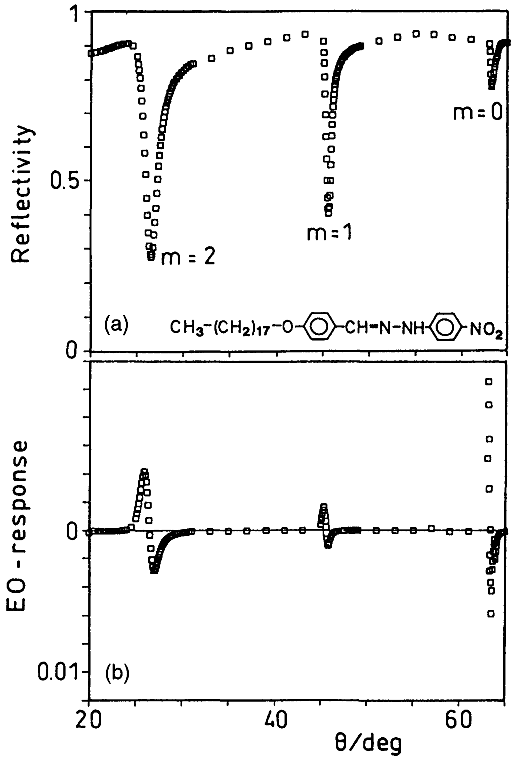

FIG. 35. (a) Waveguide mode pattern and (b) electro-optic response from applying a 30-V peak-to-peak voltage across a LBK multilayer assembly composed of 240 buffer layers, 40 A–B layers (with B being the $\chi^{(2)}$-active material), and 240 buffer layers. Structure of the NLO-active dye is given in the inset. (Aust *et al.*, 1993.)

3.3 Third-Order Nonlinear Optics

In view of possible applications of polymeric materials for all-optical information processing where (nonresonant) $\chi^{(3)}$ values in the range of 10^{-9} esu or larger are estimated to be required, it is somewhat doubtful whether linearly polyconjugated molecules allow reaching this nonresonant $\chi^{(3)}$ value. At present, it is an open question whether materials with an improved nonlinear response and/or an optimized figure of merit will be available in the future.

These figures of merit are defined such as to scale the enhancement of a desired nonlinear susceptibility, e.g., the intensity-dependent refractive-index change in the sample n_2, against other linear or nonlinear response parameters. For example, in a nonlinear directional coupler the (desired) maximum phase shift $\Delta\phi^{NL}$ is given by

$$\Delta\phi^{NL} = n_2 I(2\pi/\lambda)L, \qquad (10)$$

with L being the interaction (device) length. For a given n_2 of the material, the required phase shift of the device could be obtained simply by choosing L correspondingly (high). However, the maximum value of L is limited by the absorption in the polymer: $\max\{L\} < \alpha^{-1}$. Since $\alpha = \alpha_0 + \beta_2 I$, i.e., α contains not only the linear absorption term α_0 but also the nonlinear, e.g., two-photon, absorption term $\beta_2 I$ which is also intensity dependent, one may define (even in the absence of linear absorption, $\alpha_0 < \beta_2 I$) a figure of merit

$$T_m = 2\lambda\beta_2/n_2, \qquad (11)$$

which must be smaller than 1 for a material to be used in a nonlinear directional coupler.

On the other hand, many of these parameters are wavelength dependent. This then defines the need for a rigorous materials characterization and optimization program for those research activities committed to the development of photonic devices based on polymeric systems. First, of course, $\chi^{(3)}$ has to be determined. One experimental technique for this purpose is third-harmonic generation (THG) and is briefly presented in the next section. Of interest are not only the contributions from electronic or thermal processes or from excited states but, in particular, the spectral-dispersion behavior of $\chi^{(3)}$. This must also be known for the linear and, equally important, for the nonlinear absorption coefficients. In fact, it is currently one of the promising strategies to find spectral windows where the device performance is not limited by the two-photon (or three-photon) absorption in the materials.

Since one is interested in fast data-processing devices, with switching times $\tau < 0.1$ ps, the nonlinear recovery time is of vital importance. One experimental technique for its determination is degenerate four-wave mixing (DFWM) briefly outlined below. We should point out, once again, that in addition to these (linear and nonlinear) optical parameters, other material properties like stability, processibility into device configurations such as waveguides, etc., are equally important for a successful application.

3.3.1 Third-Harmonic Generation (THG) Third-harmonic generation describes the process in which an incident photon field of frequency ω generates, through nonlinear polarization in the medium, a coherent optical field at 3ω. It gives directly the electronic hyperpolarizability with ultrafast response time. The typical experiment performed with thin-film polymer samples prepared on suitable substrates in either transmission or reflection geometry monitors contributions from the substrate as well as from the film. Rotating the sample as outlined above for the SHG experiments here, too, leads to a coherent superposition of the free and bound waves giving rise to a fringe pattern similar to the one shown in Fig. 32 for the SHG signal. If one knows the substrate's cubic susceptibility, the experimental data can yield both the real and imaginary parts (or the modulus and the phase) of the third-order susceptibility of the polymer material. If suitable laser systems are available, these data can be obtained covering a wide range of frequencies allowing for the search for suitable spectral windows off the one-, two-, and three-photon resonances. One example is given in Fig. 36 for a polydiacetylene system (after Guo *et al.*, 1992). The experimental data show the strong spectral dependence of both the modulus and the phase of $\chi^{(3)}$. The full curve is a theoretical description of the nonlinear susceptibility of the polymer on the basis of an "essential states" picture with four levels (Guo *et al.*, 1992).

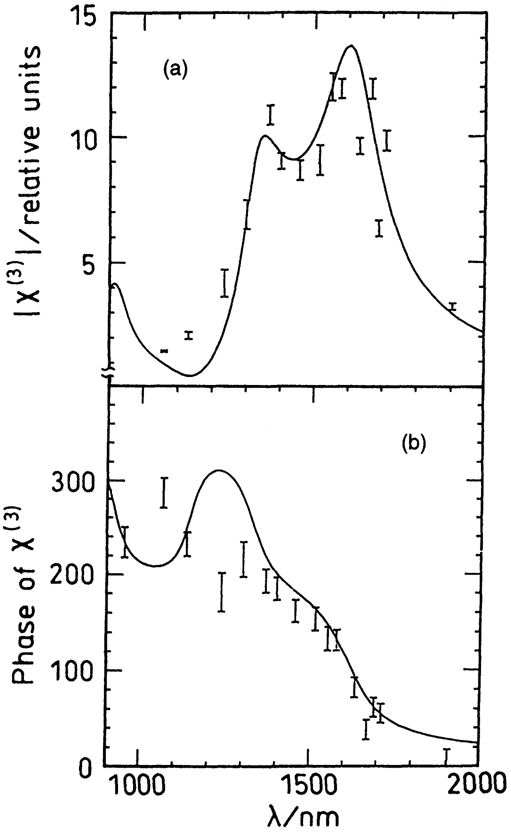

FIG. 36. (a) Modulus and (b) phase of $\chi^{(3)}$ measured for polydiacetylene (cf. Fig. 27, with $R_1 = R_2 = \{(CH_2)_4\ OCONHCH_2COO(CH_2)_3CH_3\}$) as functions of the fundamental wavelength. The full curves are theoretical calculations based on a model with four essential states (Guo *et al.*, 1992).

3.3.2 Degenerate Four-Wave Mixing (DFWM) Degenerate four-wave mixing is a holographic technique in which the output of a laser system is split into three beams. Two of them coherently interfere within the thin-film sample, thus generating a phase grating (refractive-index grating) through the intensity-dependent refractive-index change n_2 in the Kerr medium. The third beam probes this grating: Its diffraction efficiency (the fourth wave!) is monitored during the experiment. If short laser pulses are used, time-resolved experiments can be performed: If the third pulse is slightly delayed relative to the two interfering pulses, the diffracted intensity is a measure of the grating decay.

Thus, information about the important nonlinear recovery time can be obtained. (By comparison with a material of known $\chi^{(3)}$, also the magnitude of the third-order nonlinear susceptibility can be obtained.) Examples are given in the literature for DFWM experiments performed with 60-fs pulse duration showing the ultrafast response times possible with polymeric systems (Prasad, 1991).

Figure 37 summarizes some data aimed at demonstrating the influence of electronic coupling between different chromophores. The close interrelation between the detailed molecular structure of the chromophore, the supramolecular architecture of the matrix into which it is incorporated, and the linear and nonlinear optical properties of the resulting functionalized material is demonstrated for various thin-film samples of phthalocyanines. These dyes are of general interest because of their excellent light and temperature stability. The chemical structure and a few details of the different systems used in these studies are given in Fig. 37(a). In system 4, the phthalocyanine moieties were incorporated into a copolymer with styrene. Figure 37(b) shows the linear absorption spectra and (c) the decay curves as obtained from DFWM experiments at $\lambda \approx 650$ nm. Remarkable differences for the various systems are found (Kaltbeitzel *et al.*, 1989). An isolated phthalocyanine molecule like, e.g., in system 4 gives a sharp absorption spectrum. This leads to a very slow decay time of the transient grating beyond the resolution of the setup. Increasing aggregation eventually leads to an inhomogeneous broadening of the absorption band and a concomitant reduction of the transient time (by orders of magnitude!). Obviously, the electronic coupling between individual resonators affects both the linear and the nonlinear responses of a chromophore system. Various contributions to the observed phenomena may include

1. fluorescence lifetime reduction by aggregational quenching,
2. energy migration (Salcedo *et al.*, 1987), e.g., exciton diffusion into the nonilluminated area,
3. energy trapping at lower states,
4. bimolecular exciton quenching (Ho and Peyghambarian, 1988), etc.

The response times of the conjugated polymers described above are presumably also related to the relaxation processes **3** and **4.** The extended π conjugation with a tight electronic coupling via intrachain and interchain energy migration leads to the ultrafast transient times of these systems, even near resonances.

3.3.3 All-Optical Switching In this section, we just touch on a still very speculative subject, namely controlling light by light (through $\chi^{(3)}$-active materials). Various device concepts have been proposed, some of which were also experimentally tested. A very basic unit is the nonlinear directional coupler. Its realization was reported by Townsend *et al.* (1989). However, it turned out that the intensity-dependent switching of the input laser pulse from one output channel to the other was based on a thermal effect attributed to a three-photon absorption. Recently, a similar structure was presented by Kaneko *et al.* (1992), who claimed, on the basis of their experiments with Ti:Al_2O_3 laser pulses of 100-fs duration at $\lambda = 825$ nm, an electronic mechanism for the observed switching behavior. The high pulse-repetition rate still means a substantial power deposition in the sample, and it will take single-pulse experiments to show unambiguously the fast electronic nature of the all-optical switching phenomena seen in these polymer waveguide samples.

3.4 Photorefractive Polymers

The photorefractive effect was first established in $LiNbO_3$ (Chen, 1967). It arises when charge carriers, photogenerated by a spatially modulated illumination, separate and become trapped to produce a nonuniform space-charge distribution. The resulting internal field then modulates the local refractive index via the linear electro-optic effect, which results in a phase image or hologram of the illumination pattern. A photorefractive material, therefore, needs to combine some functional units that act as photoionizable charge carriers, and some that transport the photogenerated charges until they are trapped at certain sites, and it has to show a nonlinear electro-optic dependence of its refractive index upon a space-charge field. Many possible applications have been suggested including high-density optical data

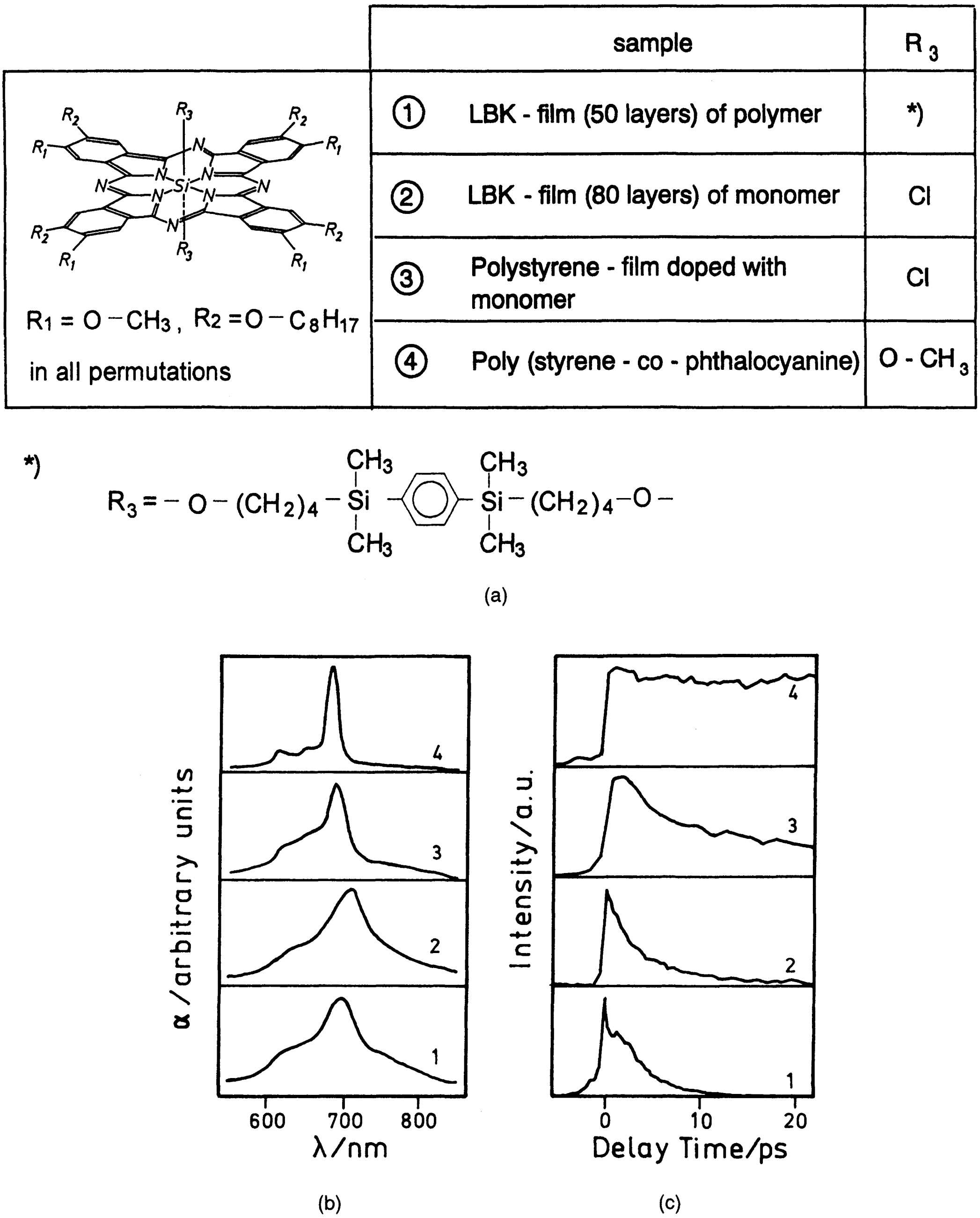

	sample	R 3
①	LBK - film (50 layers) of polymer	*)
②	LBK - film (80 layers) of monomer	Cl
③	Polystyrene - film doped with monomer	Cl
④	Poly (styrene - co - phthalocyanine)	O - CH 3

*) $R_3 = -O-(CH_2)_4-\underset{CH_3}{\overset{CH_3}{Si}}-C_6H_4-\underset{CH_3}{\overset{CH_3}{Si}}-(CH_2)_4-O-$

FIG. 37. (a) Chemical structure of phthalocyanine derivatives and sample configurations used for DFWM experiments. (b) Absorption spectra and (c) decay curves of transient gratings of some phthalocyanine thin films: 1, LBK film of 50 layers of a polymeric sample; 2, LBK film of monomeric dye; 3, polystyrene doped with monomer; 4, styrene–phthalocyanine copolymer.

storage, various image-processing techniques, phase conjugation, programmable interconnection, etc. (Günter and Huignard, 1988).

Polymers again would be ideal materials for photorefractive devices because of their excellent processability and potentially high electro-optic coefficients (see above). Also, photoconductivity in functionalized polymers is a well-established phenomenon because of its importance for the photocopy process (Burland and Schein, 1986) and for laser printing. It was, nevertheless, not until very recently that the first reports on photorefractive polymers were published (Schildkraut, 1991; Ducharme *et al.*, 1991; Tamura *et al.*, 1992) showing indeed very promising features that certainly will stimulate a very dynamic research activity in this field.

4. OUTLOOK

Polymers with their broad variety of optical properties already have their established place in modern products and technologies. The needs of our information societies and the emerging technology of the next century, photonics, mean a rapidly growing demand, chance, and challenge for ever better performing polymers with specifically tuned optical (and other) properties. Optical data storage for consumer products and in information technologies will be a growing area for the use of more and more sophisticated materials. Closely linked to this development is the requirement for a cheap yet powerful blue laser-light source. At present it is an open race whether there will first be available as a commercial mass product the blue solid-state laser diode or a device based on a red laser diode with an integrated frequency-doubling (polymeric) material. Integrated optics, in general, will have its growing share in data transmission, distribution, and processing. Here, a lot will depend on whether improved nonlinear-optical materials will be available.

These ends will be met only if by an interdisciplinary effort a better understanding of the various material parameters at the molecular level can be obtained. The close relation between macroscopic properties and functions on the one side and the microscopic aspects of structure, dynamics, and order on the other side will remain a continuous challenge for modern materials science and technology.

5. ACKNOWLEDGMENTS

It is my pleasure to thank many colleagues and friends for their support and for many helpful discussions. In particular, Y. Koike, H. Kuhn, Y. Levy, W. H. Meyer, V. Mizrahi, M. R. Philpott, H. Ringsdorf, H. Sasabe, K. Sasaki, G. I. Stegeman, J. Stumpe, J. D. Swalen, C. Urban, G. Wegner, T. Wada, and J. Yang contributed also to this article through their stimulating comments. As far as results from our group have been added, I am particularly grateful to my co-workers E. F. Aust, C. Duschl, W. Hickel, H. Motschmann, M. Sawodny, R. Reiter, and B. Rothenhäusler. Financial support came from the Deutsche Forschungsgemeinschaft, the Bundesministerium für Forschung und Technologie, The Volkswagen Stiftung, and the Leonhard-Lorenz-Stiftung.

GLOSSARY

CD-ROM: Compact disk—read-only memory.

Charge-Transfer Chromophore: Dye molecule with an electron donor and acceptor group.

Cladding: Outer part of a waveguide.

Core: Inner part of a waveguide that conducts the light.

Degenerate Four-Wave Mixing: Holographic diffraction technique with transient grating formation in materials with nonlinear polarization.

Electro-optic Modulator: Device that controls the transmission of light by the application of an electrical field to a material with a nonlinear polarization.

Host–Guest Polymer: Polymer matrix (the host) with incorporated small molecules, e.g., dyes (the guest).

Integrated Optics: Optical equivalent to electrical integrated circuits.

Langmuir Layer: Monomolecular film of amphiphilic molecules at the water–air interface.

Liquid-Crystalline Polymer: Polymer with mesophase behavior.

Optical Disk: Digital storage disk for optical readout.

Photonics: Information technology based on photons instead of electrons (electronics).

Photoresists: Polymer used in microelectronic device fabrication for lateral patterning.

(Solid) Polyelectrolyte: Polymeric material with ionic moieties.

Second-Harmonic Generation: Optical frequency doubling by a material showing nonlinear polarization.

Third-Harmonic Generation: Optical frequency tripling by a material with nonlinear polarization.

Works Cited

Allan, D. S., Kubota, F., Marks, T. J., Zhang, T. J., Liu, W. P., Wong, G. K. (1991), in: K. D. Singer, (Ed.), *Nonlinear Optical Properties of Organic Materials IV,* SPIE Proceedings No. 1560, Bellingham, WA: SPIE.

Aust, E., Hickel, W., Orendi, H., Knobloch, H., Knoll, W. (1993), *Mol. Cryst. Liq. Cryst.* **227,** 49–59.

Blodgett, K. B. (1935), *J. Am. Chem. Soc.* **57,** 1007–1013.

Bosshard, C., Küpfer, M., Flörsheimer, M., Günter, P. (1991), in: K. D. Singer, (Ed.), *Nonlinear Optical Properties of Organic Materials IV,* SPIE Proceedings No. 1560, Bellingham, WA: SPIE.

Bowden, J. J. (1988), "Polymers for Electronic and Photonic Applications," in: M. J. Bowden, S. R. Turner (Eds.), *Electronic and Photonic Applications of Polymers,* Advances in Chemistry Series No. 218, Washington, DC: American Chemical Society.

Bubeck, C., Kaltbeitzel, A., Lenz, R. W., Neher, D., Stenger-Smith, J. D., Wegner, G. (1989), in: J. Messier, F. Kajzar, P. Prasad, D. Ulrich (Eds.), *Nonlinear Optical Effects in Organic Polymers,* NATO Advanced Study Institutes Series E, Applied Sciences, Vol. 162, Dordrecht, Netherlands: Kluwer Academic Publishers, p. 143.

Bubeck, C., Laschewsky, A., Lupo, D., Neher, D., Ottenbrei, P., Paulus, W., Prass, W., Ringsdorf, H., Wegner, G. (1991), *Adv. Mater.* **3,** 54–62.

Burland, D. M., Schein, L. B. (1986), *Phys. Today* **39** (5), 46–51.

Burland, D. M., Rice, J. E., Downing, J., Michl, J. (1991), in: K. D. Singer (Ed.), *Nonlinear Optical Properties of Organic Materials IV,* SPIE Proceedings No. 1560, Bellingham, WA: SPIE.

Burstein, E., Chen, W. P., Chen, Y. J., Hartstein, A. (1974), *J. Vac. Sci. Technol.* **11,** 1004–1019.

Chang, K. (Ed.) (1991), *Handbook of Microwave and Optical Components,* Vol. 4, *Fiber and Electrooptical Components,* New York: Wiley.

Chen, F. S. (1967), *J. Appl. Phys.* **38,** 3418–3423.

Chen, Y. M., Mandal, B. K., Lee, J. Y., Miller, P., Tripathy, S. (1991), *Mater. Res. Soc. Symp. Proc.* **214,** 35–40.

Cresswell, J. P., Tsibouklis, J., Petty, M. C., Feast, W. J., Carr, N., Goodwing, M., Lvov, Y. M. (1990), in: K. D. Singer (Ed.), *Nonlinear Optical Properties of Organic Materials III,* SPIE Proceedings No. 1337, Bellingham, WA: SPIE.

Cross, G. H., Peterson, I. R., Girling, I. R., Cade, N. A., Goodwin, M. J., Carr, N., Sethi, R.-S., Marsden, R., Gray, G. W., Lacey, D., McRoberts, A. M., Scrowston, R. M., Toyne, K. J. (1988), *Thin Solid Films* **156,** 39–52.

Ducharme, S., Scatt, J. C., Twieg, R. J., Moerner, W. E. (1991), *Phys. Rev. Lett.* **66,** 1846–1848.

Duda, G., Schouten, A. J., Arndt, T., Lieser, G., Schmidt, G. F., Bubeck, C., Wegner, G. (1988), *Thin Solid Films* **159,** 221–227.

Dumont, M., Lévy, Y. (1989), in: T. Kobayashi (Ed.), *Nonlinear Optics of Organics and Semiconductors,* Springer Proceedings in Physics, Vol. 36, Berlin: Springer, p. 256.

Eich, M., Reck, B., Yoon, D. Y., Willson, C. G., Bjorklung, G. C. (1989), *J. Appl. Phys.* **66,** 3241–3246.

Gadret, G., Kajzar, F., Raimond, P. (1991), in: K. D. Singer (Ed.), *Nonlinear Optical Properties of Organic Materials IV,* SPIE Proceedings No. 1560, Bellingham, WA: SPIE.

Gordon, J. G., II, Swalen, J. D. (1977), *Opt. Commun.* **22,** 374–379.

Groh, W., Lupo, D., Sixl, H. (1989), *Adv. Mater.* **101,** 1580–1591.

Günter, P., Huignard, J.-P. (Eds.) (1988), *Photorefractive Materials and Their Applications,* Berlin: Springer-Verlag.

Guo, D., Mazumdar, S., Stegeman, G. I., Cha, M., Neher, D., Aramaki, S., Toruellas, W., Zanoni, R. (1992), *Mater. Res. Soc. Symp. Proc.* **247,** 151–161.

Herman, W. N., Rosen, W. A., Sperling, L. H., Murphy, C. J., Jain, H. (1991), in: K. D. Singer (Ed.), *Nonlinear Optical Properties of Organic Materials V,* SPIE Proceedings No. 1560, Bellingham, WA: SPIE.

Herminghaus, S., Smith, B. A., Swalen, J. D. (1991), *J. Opt. Soc. Am. B* **8,** 2311–2318.

Heuberger, A. (1988), *J. Vac. Sci. Technol.* **6,** 107–113.

Hickel, W., Duda, G., Jurich, M., Kröhl, T., Rochford, K., Stegeman, G. I., Swalen, J. D., Wegner, G., Knoll, W. (1990), *Langmuir* **6,** 1403–1407.

Ho, Z. Z., Peyghambarian, N. (1988), *Chem. Phys. Lett.* **148,** 107–109.

Hsiung, H., Rodriguez-Parada, J., Beckerbauer, R. (1991), *Chem. Phys. Lett.* **182,** 88–90.

Ishigure, T., Nihei, E., Koike, Y. (1992), *Sen-i Gakkai Symp. Preprints* No. B-4.

Jenekhe, S. A., Osaheni, J. A., Meth, J. S., Vanherzeele, H. (1992), *Chem. Mater.* **4,** 683–687.

Jerphagnon, J., Kurtz, S. K. (1970), *J. Appl. Phys.* **41,** 1667–1673.

Jungbauer, D., Reck, B., Twieg, R. J., Yoon, D. Y., Willson, C. G., Swalen, J. D. (1991), *Appl. Phys. Lett.* **56,** 2610–2612.

Kajzar, F. (1992) in: D. B. Ostrowsky, R. Reinisch (Eds.), *Guided Wave Nonlinear Optics,* Dordrecht, Netherlands: Kluwer Academic Publishers.

Kaltbeitzel, A., Neher, D., Bubeck, C., Sauer, T., Wegner, G., Caseri, W. (1989), in: H. Kuzmany, M. Mehring, S. Roth (Eds.), *Electronic Properties in Conjugated Polymers III,* Berlin: Springer-Verlag.

Kämpf, G. (1985), *Ber. Bunsenges. Phys. Chem.* **89,** 1179–1190.

Kaneko, A., Ito, A., Kuwabara, T., Sasaki, S., Sasaki, K., Sokoloff, J. P., Wada, T., Sasabe, H. (1992), *Sen-i Gakkai Symp. Preprints* No. B-118.

Knobloch, H., Duschl, C., Knoll, W. (1989), *J. Chem. Phys.* **91,** 3810–3814.

Knoll, W. (1991a), *Mater. Res. Soc. Bull.* **16,** 29–39.

Knoll, W. (1991b), *Makromol. Chem.* **192,** 2827–2856.

Köhler, W., Robello, D. R., Dao, P. T., Willand, C. S., Williams, D. J. (1990), *J. Chem. Phys.* **93,** 9157–9166.

Koike, Y. (1991), *Polymer* **32,** 1737–1745.

Kretschmann, E. (1972), *Opt. Commun.* **6,** 185–188.

Kuhn, H., Möbius, D., Bücher, H. (1972), in: A. Weissberger, B. W. Rossiter (Eds.), *Physical Methods of Chemistry, A,* New York: Wiley.

Langmuir, I. (1917), *J. Am. Chem. Soc.* **39,** 1848–1852.

Ledoux, I., Josse, D., Vidakovic, P., Zyss, J., Hann, R. A., Gordon, P. F., Bothwell, B. D., Gupta, S. K., Allen, S., Robin, P., Chustaing, E., Dubois, J. C. (1987), *Europhys. Lett.* **3,** 803–809.

Lee, D. L. (1986), *Electromagnetic Principles of Integrated Optics,* New York: Wiley.

Lenz, R. W., Han, C. C., Stenger-Smith, J. D., Karasz, F. E. (1990), *J. Polym. Sci. A* **26,** 3241–3247.

Levenson, R., Liang, J., Toussaere, E., Zyss, J. (1991), in: K. D. Singer (Ed.), *Nonlinear Optical Properties of Organic Materials IV,* SPIE Proceedings No. 1560, Bellingham, WA: SPIE.

Man, H.-T., Yoon, H. N. (1992), *Adv. Mater.* **4,** 159–168.

Mathy, A., Simmrock, H.-U., Bubeck, C. (1991), *J. Phys. D* **24,** 1003–1009.

Messier, J., Kajzar, F., Prasad, P. N., Ulrich, D. (Eds.) (1989), *Nonlinear Optical Effects in Organic Polymers,* NATO Advanced Study Institutes Series E, Applied Sciences, Vol. 162, Dordrecht, Netherlands: Kluwer Academic Publishers.

Meyer, W. H., Pecherz, J., Mathy, A., Wegner, G. (1991), *Adv. Mater.* **3,** 153–156.

Muto, S., Ito, H. (1992), *Sen-I Gakkai Symp. Preprints* No. B-17, 1–4.

Muto, S., Seki, N., Ichikawa, S., Ito, H. (1991), *Opt. Commun.* **81,** 273–275.

Norwood, R. A., Khanarian, G. (1990), *Electron. Lett.* **26,** 2105–2107.

Page, R. H., Jurich, M. C., Reck, B., Sen. A., Twieg, R. J., Swalen, J. D., Bjorklund, G. C., Willson, C. G. (1990), *J. Opt. Soc. Am. B* **7,** 1239–1246.

Pearson, J. M. (1988), "Polymers in Optical Recording," in: M. J. Bowden, S. R. Turner (Eds.), *Electronic and Photonic Applications of Polymers,* Advances in Chemistry Series No. 218, Washington, DC: American Chemical Society.

Penner, T. L., Armstrong, N. J., Willand, C. S., Schildkraut, J. S., Robello, D. R. (1991), in: K. D. Singer (Ed.), *Nonlinear Optical Properties of Organic Materials IV,* SPIE Proceedings No. 1560, Bellingham, WA: SPIE.

Prasad, P. N. (1991), *Polymer* **32,** 1746–1751.

Prasad, P. N., Williams, D. J. (1991), *Introduction to Nonlinear Optical Effects in Molecules and Polymers,* New York: Wiley.

Przybilla, K. J., Röschert, H., Pawlowski, G. (1992), *Adv. Mater.* **4,** 239–242.

Raether, H. (1977), in: G. Hass, M. H. Francome, R. W. Hoffmann (Eds.), *Physics of Thin Films,* Vol. 9, New York: Wiley.

Raether, H. (1988) *Surface Plasmons on Smooth and Rough Surfaces and on Gratings,* Springer Tracts in Modern Physics, Vol. 111, Berlin: Springer-Verlag.

Roberts, G. G. (Ed.) (1990), *Langmuir-Blodgett Films,* New York: Plenum.

Salcedo, J. R., Siegmann, A. E., Dlott, D. D., Fayer, M. D. (1987), *Phys. Rev. Lett.* **41,** 131–133.

Sasabe, H., Wada, T., Hosoda, M., Ohkawa, H., Hara, M., Yamada, A., Garito, A. F. (1990), in: K. D. Singer (Ed.), *Nonlinear Optical Properties of Organic Materials III,* SPIE Proceedings No. 1337, Bellingham, WA: SPIE.

Sasaki, K. (1992), *Sen-i Gakkai Symp. Preprints* No. B-92.

Sawodny, M., Schmidt, A., Stamm, M., Knoll, W., Urban, C., Ringsdorf, H. (1992), *Thin Solid Films* **210/211,** 500–503.

Schildkraut, J. (1991), *Appl. Phys. Lett.* **58,** 340–342.

Schrader, S., Koch, K. H., Mathy, A., Bubeck, C., Müllen, K., Wegner, G. (1991), *Prog. Colloid Polym. Sci.* **85,** 143–148.

Shuto, Y., Amano, M., Kaino, T. (1991), in: K. D. Singer (Ed.), *Nonlinear Optical Properties of Organic Materials IV,* SPIE Proceedings No. 1560, Bellingham, WA: SPIE.

Simmrock, H.-U., Mathy, A., Dominguez, L., Meyer, W. H., Wegner, G. (1989), *Angew. Chem. Adv. Mater.* **101,** 1148–1149.

Singer, K. D., Kuzyk, M. G., Sohn, J. E. (1987), *J. Opt. Soc. Am. B* **4,** 968–975.

Singer, K. D., Sohn, J. E., King, L. A., Gordon, H. M., Katz, H. E., Dirk, C. W. (1989), *J. Opt. Soc. Am. B* **6,** 1339–1348.

Stegeman, G. I., Seaton, C. T., Zanoni, R. (1987), *Thin Solid Films* **152,** 231–263.

Stegeman, G. I., Stolen, R. H. (1989), *J. Opt. Soc. Am. B* **6,** 652–681.

Swalen, J. D. (1986), *J. Mol. Electron.* **2,** 155–157.

Swalen, J. D., Tacke, M., Santo, R., Rieckhoff, K. E., Fischer, J. (1978), *Helv. Chim. Acta* **61,** 960–970.

Swalen, J. D., Bjorklund, G. C., Ducharme, S., Fleming, W., Herminghaus, S., Jungbauer, D., Jurich, M., Moerner, W. E., Reck, B., Smith, B. A., Twieg, R. J., Willson, C. G., Zentel, R. (1991), in: J. Messier *et al.* (Eds.), *Organic Molecules for Nonlinear Optics and Photonics,* Dordrecht, Netherlands: Kluwer Academic Publishers, p. 433.

Swalen, J. D., Fleming, W., Jurich, M., Moerner, W. E., Smith, B. A., Herminghaus, S., Bjorklund, G. C. (1992), *Mater. Res. Soc. Symp. Proc.* **228,** 101–112.

Tamir, T. (Ed.) (1979), *Integrated Optics,* Topics in Applied Optics, Vol. 7, Berlin: Springer-Verlag.

Tamura, K., Padias, A. B., Hall, H. K., Jr., Peyghambarian, N. (1992), *Appl. Phys. Lett.* **60,** 1408–1411.

Teng, C. C., Man, H. T. (1990), *Appl. Phys. Lett.* **56,** 1734–1736.

Tien, P. K. (1969), *Rev. Mod. Phys.* **49,** 361–420.

Tomono, T., Nishikata, Y., Pu, L. S. (1992), *Sen-i Gakkai Symp. Preprints* No. B-102.

Townsend, P. D., Jackel, J. L., Baker, G. L., Shelburne, J. A., III, Etemad, S. (1989), *Appl. Phys. Lett.* **55,** 1829–1831.

Ushioda, S., Sasaki, Y. (1983), *Phys. Rev. B* **27,** 1401–1406.

Verbiest, T., Persoons, A., Samyn, C. (1991), in: K. D. Singer (Ed.), *Nonlinear Optical Properties of Organic Materials IV,* SPIE Proceedings No. 1560, Bellingham, WA: SPIE.

Weiss, S., Aust, E., Meyer, W., Knoll, W. (1995), *J. Nonlinear Opt. Phys. Mater.*, in press.

Williams, D. J. (1987), in: D. S. Chemla, J. Zyss (Eds.), *Nonlinear Optical Properties of Organic Molecules and Crystals,* Orlando, FL: Academic.

Willson, C. G., Bowden, M. J. (1988), "Organic Resist Materials," in: M. J. Bowden, S. R. Turner (Eds.), *Electronic and Photonic Applications of Polymers,* Advances in Chemistry Series No. 218, Washington, DC: American Chemical Society.

Wolter, H. (1956), in: S. Flügge (Ed.), *Handbuch der Physik,* Berlin: Springer-Verlag.

Wu, J. W., Valley, J. F., Stiller, M., Ermer, S., Binkley, E. S., Kenney, J. T., Lipscomb, G. F., Lytel, R. (1991), in: K. D. Singer (Ed.), *Nonlinear Optical Properties of Organic Materials IV,* SPIE Proceedings No. 1560, Bellingham, WA: SPIE.

Zyss, J. (1985), *J. Mol. Electron.* **1** (2), 25–45.

Further Reading

Saleh, B. E. A., Teich, M. C. (1991), *Fundamentals of Photonics,* New York: Wiley.

Emerson, J. A., Torkelson, J. M. (Eds.) (1991), *Optical and Electrical Properties of Polymers, Mater. Res. Soc. Symp. Proc.,* Vol. 214, Pittsburgh, PA: Materials Research Society.

Bowden, M. J., Turner, S. R. (Eds.) (1988), *Electronic and Photonic Applications of Polymers,* Advances in Chemistry Series No. 218, Washington, DC: American Chemical Society.

Aust, E. F., Sawodny, M., Ito, S., Knoll, W. (1994), *Scanning* **16,** 353–361.

Aust, E. F., Ito, S., Sawodny, M., Knoll, W. (1994), *Trends Polym. Sci.* **2,** 313–323.

Kämpf, G. (1985), *Ber. Bunsenges. Phys. Chem.* **89,** 1179–1190.

Chang, K. (Ed.) (1991), *Handbook of Microwave and Optical Components,* New York: Wiley.

Prasad, P. N., Ulrich, D. R. (Eds.) (1988), *Nonlinear Optical and Electroactive Polymers,* New York: Plenum.

Prasad, P. N., Williams, D. J. (1991), *Introduction to Nonlinear Optical Effects in Molecules and Polymers,* New York: Wiley.

Zyss, J. (1994), *Molecular Nonlinear Optics,* Boston: Academic.

POLYSILICON

See SILICON, POLYCRYSTALLINE

5
Positron Spectroscopy

Paul G. Coleman

5.1 Introduction 117
5.2 Fundamentals 118
5.2.1 Annihilation of Free Positrons and Positronium 118
5.2.2 The Fate of Positrons in Condensed Matter 119
5.2.2.1 Positron Backscattering 119
5.2.2.2 Positron Implantation 120
5.2.2.3 Positron Diffusion 120
5.2.2.4 Positron Trapping 121
5.2.2.5 Positron-Surface Interactions 121
5.2.2.6 Positron Annihilation: Observables 123
5.2.2.7 Positronium Annihilation 124
5.3 Experimental Methods 125
5.3.1 Sources of Positrons 125
5.3.2 Lifetime Spectroscopy 126
5.3.2.1 Experimental Lifetime Systems 126
5.3.3 Angular Correlation of Annihilation Radiation 127
5.3.4 Doppler Broadening of Annihilation Radiation 129
5.3.4.1 Two-Detector DBAR 131
5.3.4.2 Age-Momentum Correlation (AMOC) 131
5.3.5 Positron Beams 132
5.3.5.1 Laboratory-Based Beams 132
5.3.5.2 High-Intensity Positron Beams 134
5.3.5.3 Beam Bunching 134
5.3.5.4 Positron Microbeams 134
5.3.5.5 Polarized Positron Beams 135
5.3.5.6 MeV Positron Beams 135
5.3.5.7 Trap-Based Beams 136
5.3.5.8 Positron-Beam-Based Spectroscopies 136
5.4 Examples of Research Using Positron Spectroscopies 139
5.4.1 Vacancy-Type Defects 139
5.4.2 Structural Changes 140

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

5.4.3 Nanoparticles 141
5.4.4 Interfaces 142
5.4.5 Fermi Surfaces 142
5.4.6 Nanoporous Materials and Open Volumes in Polymers 143
5.4.7 Surfaces 144
5.4.8 Positron Microscopy 146
5.4.9 Positron and Positronium Chemistry 146
Acknowledgments 148
Glossary 148
References 148
Further Reading 150

5.1
Introduction

The history of positron annihilation spectroscopy is, in many ways, an archetypal story of scientific development, from fundamental physics to applications of industrial importance. Predicted by Dirac (1930) and discovered by Anderson (1932), and found to be emitted from artificially produced isotopes in the same decade (Joliot-Curie and Joliot-Curie, 1934), the use of positrons as a subatomic probe of matter was delayed for over a decade by world events.

In 1942, the first fundamental measurement on γ-rays emitted as a result of the annihilation of positrons in solids was performed; Behringer and Montgomery (1942) found that the angle between the two γ-rays which almost always follow annihilation events did not deviate from 180° by more than 15 minutes of arc. DeBenedetti, Cowan and Konneker (1949) were the first to demonstrate that this angle deviated from 180° by an amount linked to the electron momentum and soon afterward published their vision of a new spectroscopy of electron momentum distributions – *angular correlation of annihilation radiation* (now termed *ACAR*) (DeBenedetti *et al.*, 1950). DeBenedetti was joined in his pioneering work by Page and Heinberg (1956), Stewart (1957), and Berko and Plaskett (1958).

Shearer and Deutsch (1949) pioneered the measurement of the lifetimes of positrons in gases, and it was this technique that led to the experimental discovery (Deutsch, 1951) of the positron–electron bound state, positronium (theoretically considered and named *electrum* by Mohorovičić (1934) and renamed *positronium* by Ruark (1945)). Extension of positron lifetime spectroscopy to liquids and solids required faster timing techniques, and these were developed by Bell and Graham (1953) and exploited soon thereafter by Landes, Berko and Zuchelli (1956) and others in the study of structural changes on the atomic scale.

In the 1960s, with the development of more widely available equipment, the level of activity and sophistication increased in both lifetime and angular correlation spectroscopies – for example, yielding new information on positron slowing down and annihilation in gases (Osmon, 1965) and in metals (Weisberg and Berko, 1967). In this decade, experimental problems such as the effect of the radioactive positron source on results, the analysis of data containing

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

multiple lifetime components, and struggles with electronic stability, were still being worked on. In the late 1960s, the role of positron trapping in atomic-scale open-volume point defects in solids was first realized (and explained why different laboratories often produced conflicting results for notionally the same sample). Rather than confusing the issue and burying the technique, this realization gave positron spectroscopy a new lease of life; positrons are so efficiently trapped by these defects that they soon became a sensitive probe of defect structures and open volumes in a wide variety of solids (Petersen, Thrane and Cotterill, 1974).

Further advances in instrumentation, coupled with the traditional resourcefulness and inventiveness which has characterized the positron research community over the years, led to the application of Doppler broadening spectroscopy in the 1970s – a high count-rate, lower-resolution measure of electron momentum densities – which quickly grew as its efficacy and scope for application were realized (Maier *et al.*, 1979). In the same decade, ACAR was enhanced by the development of two-dimensional measurements, first by using an array of discrete photon detectors (Mader *et al.*, 1976) and then by using position-sensitive gamma cameras (West, Mayers and Walters, 1981) or multiwire proportional chambers (Jeavons *et al.*, 1978).

One of the most significant developments in positron techniques also happened in the early 1970s, with the first usable laboratory-based beams of controllable low energy (electronvolts to kilo electronvolts) (Canter *et al.*, 1972). After a few years of application only in atomic physics (i.e., in positron scattering), these beams were applied to solids and, after a number of groundbreaking fundamental studies of positron-surface interactions (Mills, Platzman and Brown, 1978), have been used widely since the 1980s to study surface and near-surface characteristics, thin films, and interfaces (e.g., Triftshäuser and Kögel, 1982). Beam systems – both laboratory and facility based – are now available, delivering positrons with energies controllable between 0.02 eV and mega electronvolts at intensities up to almost 10^9 per second.

We focus on positron spectroscopy of condensed matter; much important fundamental research has been and continues to be performed in atomic physics, but this activity lies outside the scope of this chapter. For an excellent overview of this work, the reader is directed to the book by Charlton and Humberston (2001).

5.2
Fundamentals

5.2.1
Annihilation of Free Positrons and Positronium

The annihilation of a positron by its antiparticle, the electron, both stationary, can theoretically result in the emission of any number n of γ-rays if n is greater than 1. The γ-rays take away the rest energy $2mc^2$, where m is the mass of each particle and c is the speed of light. Zero and single γ-ray emission are forbidden for an isolated positron–electron pair at rest because both energy and momentum have to be conserved; both can theoretically occur in the presence of a third body, however. While the emission of two or more γ-photons satisfies both energy and momentum conservation, the probability of n photon emission decreases sharply as n increases (and as the number of vertices on the corresponding Feynman diagrams (Charlton and Humberston, 2001)). For example, Ore and Powell (1949) calculated that the probability of three-γ emission is about 370 times less likely than that for two-γ emission, and the contributions to experimental studies of decays with n greater than 3 are practically negligible. Consequently, almost all annihilation events involving free (or quasi-free) positrons and electrons result in the emission of two γ-rays of energy mc^2 (511 keV) in opposite directions (to conserve momentum).

A positron and electron can exist in the bound hydrogen-like system, called *positronium* (*Ps*). Because the reduced mass is half that of the electron in hydrogen, energy levels are half those in hydrogen (e.g., the ground-state binding energy is 6.8 eV) and interparticle separations are doubled. Ps annihilation events are governed by a selection rule resulting from charge parity (CP) invariance. The parity of the γ-photons is $(-1)^n$, and for ground-state Ps it is $(-1)^S$, where S is the total spin angular momentum of the Ps atom. If the positron and electron in the Ps have opposing spins (the singlet state, *para*-Ps or *p*-Ps) then the total $S = 0$ and consequently n has to be an even number. Conversely, if the positron and electron spins are parallel and $S = 1$ (the triplet state, *ortho*-Ps or *o*-Ps) then n has to be odd. Following the arguments in the previous paragraph, this means that *p*-Ps principally decays into two 511-keV antiparallel γ-rays, whereas *o*-Ps decays into three γ-rays whose total energy is $2\,mc^2$ or 1022 keV. The distribution of γ-energies from *o*-Ps decay was calculated by Ore and Powell (1949) and measured by Chang, Tang and Xi (1985), indicating that the majority of *o*-Ps decays involve the emission of two γ-rays having energies that are a large fraction of mc^2, traveling in roughly opposite directions, with a third low-energy γ-photon emitted at some angle between them. Ore and Powell also computed that the mean lifetime of *o*-Ps in vacuo is about 140 nanoseconds, over 1100 times longer than the 125 picoseconds mean life of *p*-Ps (Wheeler, 1946).

5.2.2
The Fate of Positrons in Condensed Matter

5.2.2.1 Positron Backscattering

A significant fraction of positrons incident upon a target material, whether from a radioactive source or in a monoenergetic beam, are turned around in a few collisions and leave the target with reduced energies. Measurements have shown that there is significant quasi-elastic backscattering and that the angular distribution of backscattered positrons is peaked at angles around the surface normal. Backscatter coefficients η_+ range from a few percent to almost 50%, and depend principally on the incident positron energy and the atomic number of the target material (increasing with both) (MacKenzie *et al.*, 1973; Coleman *et al.*, 1992). The consequences of backscattering have to be considered in experiments involving the detection of annihilation γ-radiation, where the detection of γ-radiation from the decay of backscattered positrons at sites other than the intended target can corrupt data.

5.2.2.2 Positron Implantation

Energetic positrons entering a sample material lose energy via electronic collisions, reaching electronvolt energies in a fraction of a picosecond and then thermalizing in a further time period from one to a few picoseconds, depending on the material. This second phase is dominated by phonon interactions, and, here, the shortest thermalization times are in metals and small band-gap semiconductors; in insulators and wide band-gap semiconductors several picoseconds may be taken for thermalization, and in extreme cases positrons may not reach thermal equilibrium before annihilation. This last scenario is also possible for monoenergetic positrons implanted at very low energies (i.e., below about 1 keV), which may retain ~0.1–1 eV when encountering the sample surface and thus are able to pass through or interact with the surface in ways different from those positrons that have fully thermalized.

Thus, we define positron implantation (or penetration) depth as that at which an incident positron reaches thermal equilibrium. Positrons emitted from the nuclei of a radioisotope decaying via positive β-decay penetrate to depths z below the surface of condensed matter targets whose distribution is traditionally given by the implantation profile $P(z) = \alpha \exp(-\alpha z)$, with $\alpha(\text{cm}) \approx 16\rho {E_m}^{-1.4}$ (ρ = density in grams per cubic centimeter, E_m = maximum β-positron energy in kilo electronvolts). This means that β-positrons can penetrate to depths of ~1 mm, and are thus used to study the bulk properties of materials. A recent experimental study (Foster *et al.*, 2007) has shown that the exponential model provides a reasonable description of β-positron implantation except at small depths, where it underestimates $P(z)$ by ~10%, and very close to the surface, where $P(z)$ should tend to zero rather than its maximum value.

If the incident positrons impinge on a target normally with an energy E, however, their implantation profile has been generally taken to be well described by a Gaussian derivative:

$$P(z, E) = 2\left(\frac{z}{{z_0}^2}\right)\exp\left(\frac{-z^2}{{z_0}^2}\right) \qquad (5.1)$$

where the parameter z_0 (in nanometers) is usually taken to be related to E and to the material density ρ (in grams per cubic centimeter) by the simple expression $(40/\rho)E^{1.6}$. z_0 is 11% larger than the mean implantation depth. While the form for $P(z, E)$ has been shown to be adequate in most cases, the prefactor and the energy exponent have been found to be material dependent (Baker *et al.*, 1991).

5.2.2.3 Positron Diffusion

The diffusive motion undergone by thermalized positrons in condensed matter may be principally considered to involve isotropic elastic scattering from acoustic phonons, and may be described by the diffusion equation because, although the positron is a quantum entity, its wavelength and mean free path (both approximately nanometers) are small compared with typical total distances traveled before annihilation. The positron diffusivity D_+ is proportional to its mobility η_+, with the constant of proportionality being (kT/e). D_+ varies with temperature as $T^{1/2}$ for scattering off acoustic phonons, as $T^{-3/2}$ for optical phonons, and $T^{3/2}$ for neutral impurities; however, the latter two dependencies are only relevant for certain materials and at very high or very low temperatures. At room temperature D_+ has values of a few square centimeters per second for most metals and semiconductors. The diffusion length of positrons $L_+ = (D_+\tau)^{1/2}$, where τ is the mean lifetime of positrons in a material (having values in the range ~100–200 picoseconds); typical values of L_+ are ~100 nm in metals to 270 nm in silicon.

5.2.2.4 **Positron Trapping**

Once considered a problem in positron measurements, the propensity for positrons to be trapped efficiently by open-volume point defects – vacancies, vacancy clusters, and voids – and at shallow trapping sites such as at a negatively charged impurity – is now the mainstay of applied positron annihilation spectroscopy. Although the classical picture of positrons being attracted to a region of lower than average positive charge is commonly used to explain positron trapping, the diffusing positron is a quantum particle and trapping should be considered to be quantum localization. Positrons provide a unique method for nondestructive characterization of such defects at concentrations as low as 10^{-7} per atom, saturating at $\sim 10^{-4}$ per atom.

There are two main models describing positron trapping. For small point defects such as vacancies, the trapping is *transition-limited* – that is, dependent upon the probability of trapping once a defect is encountered; this is described by the specific trapping coefficient ν, which has typical values in the range of 10^{14}–10^{15} per second. The total trapping rate κ is then νC, where C is the defect concentration per atom and ν for semiconductors has a temperature dependence, which can be used to identify the defect type. The second trapping model is *diffusion-limited* – that is, determined by the probability of finding the trapping sites which, once found, have a trapping probability of unity. The total trapping rate is then $\kappa = 4\pi RD_+N$, where R (centimeters) and N (cubic centimeters) are the radius and number density of the defects, respectively, and D_+ is the positron diffusion constant in square centimeters per second.

The charge state of open-volume point defects in semiconductors and insulators can play an important role in positron trapping. The specific trapping rate ν is very small for positively charged defects, whereas neutral sites are effective positron traps and negative charge can increase ν by a significant factor. While open-volume defects are deep traps for positrons, shallow trapping can occur around negatively charged vacancy or impurity defects – this is usually seen at low temperatures and identified by the temperature dependence of the trapping probability (Saarinen *et al.*, 1997).

Thermally induced detrapping from defect sites can occur, depending on the Boltzmann factor $\exp(-E_b/kT)$, where E_b is the binding energy of the positron in the trap; therefore, for example, for deep traps – that is, those for which E_b is a few electronvolts – detrapping can be considered negligible.

5.2.2.5 **Positron-Surface Interactions**

If a positron encounters the surface of a material sample after implantation it can suffer a variety of fates, all of which have found application in various experimental studies. We shall, for the moment, assume that the positron has been thermalized before reaching the surface, and not consider those backscattered positrons that penetrate the surface without any significant interaction with it.

There are three main channels open to the positron on encountering a surface: (i) trapping in the surface potential well, (ii) binding with a surface electron and leaving as Ps, and (iii) being reemitted from the surface by a negative positron work function φ_+. For those materials for which φ_+ is positive, channel (iii) is not open.

Surface Trapping The potential encountered by a positron may be considered for our purposes to be a combination of the attractive image potential, which falls off as the reciprocal of distance from the surface, and the repulsive dipole potential, which falls off significantly more quickly above the surface (i.e., within a few angstroms). The combination of these two potentials thus creates a deep well above the surface in which positrons are efficiently trapped. On an atomically clean perfect surface, the trapped positron can be used as a probe of electronic surface states. If the temperature of the material is increased then the trapped positron can be desorbed, but only via the formation of thermal-energy Ps, which supplies its 6.8-eV binding energy; this, however, is an excellent source of low-energy Ps atoms for spectroscopic measurements.

Positronium Formation For metals and semiconductors, the surface is the only site where the electron density is low enough to permit Ps formation. The formation potential for Ps at surfaces, essentially the energy possessed by the emitted Ps atom, is the balance between the sum of the electron and positron work functions – that is, the energy required to remove both particles independently from the solid – and the 6.8 eV gained by forming the bound state. In the early days of positron-surface experimentation, Ps formation was used as a signature that the positron had diffused back to the surface; however, the probability of forming Ps is critically dependent on surface conditions, which can have a positive or negative influence on measurements. Measurement of the velocity spectra of emitted Ps was also used as a probe of electronic surface states. Of great fundamental importance is the fact that small fractions of the Ps emitted can be in an excited state or as the negative ion Ps^-.

Positron Reemission The sum of the correlation and dipole potentials outlined in Section 5.2.2.5, Surface Trapping (the correlation potential is the asymptotic level of the image potential inside the material) can be positive or negative – that is, positrons are either attracted or repelled by the surface, depending on their relative magnitudes. The difference between the resultant total potential within a solid and the reference vacuum level far from the surface is the positron work function φ_+. For a number of solids the dipole potential is greater than the correlation potential and φ_+ is negative – that is, thermalized positrons are emitted from the surface with an energy of $\sim\varphi_+$. Negative φ_+ values for metals range from close to zero to ~3 eV and, like their electron equivalent, are dependent on surface composition and structure. For a perfect surface, the emission is normal to the surface, smeared only by thermal effects; for a "real" surface, the emitted positrons may undergo inelastic collisions above the surface that lower their average energy and increase their angular spread; if enough energy is lost in such a collision, then the positron may fall into the surface well and be trapped. Thus, there is a correlation between the magnitude of φ_+ and the fraction of positrons reemitted from a surface.

A small fraction of positrons from a radioactive source incident on a solid with negative φ_+ possess energies low enough (i.e., below ~10 keV) to be thermalized, return to the surface, and subsequently be reemitted with work function energies. This is the basis of the generation of monoenergetic positron beams, which have been widely used to study surface and near-surface phenomena for the past 25 years.

Surface Branching Ratios The relative probabilities of these three main channels open to the positron are described by the branching ratios ε. For a solid with a positron work function of ~–1 eV, the three branching ratios are in the approximate ratio of 1 : 1 : 1. For surfaces with a positive φ_+, approximately half the positrons form Ps and half fall into the surface trap at room temperature.

Nonthermalized Positrons If the positrons incident on a surface have energies below about 1 keV, there is a significant probability that they will return to the surface without being fully thermalized – that is, with residual energies of several hundred milli electronvolts. These epithermal positrons interact with the surface with rather different branching ratios; they can more readily overcome surface trapping and leave as positrons with energies above φ_+, and can pick up an electron and form Ps, which again leaves with higher energy than that formed by the process outlined in Section 5.2.2.5, Positronium Formation.

On encountering "real", dirty surfaces – for example, oxide-covered – the branching ratios can be hugely different from those for a clean perfect surface and, in addition, there can be significant, sometimes overwhelming, trapping at surface defects.

We next focus on the annihilation of thermal-energy positrons with electrons, and of Ps-like states, in condensed matter. Other aspects of positron and Ps interactions in solids and liquids are discussed in later sections.

5.2.2.6 Positron Annihilation: Observables

The cross section σ for annihilation is inversely proportional to positron speed v, but the annihilation rate is proportional to σv. Consequently, although annihilation during positron slowing down is possible, has been observed, and may play a role in some positron experiments (Weber *et al.*, 1999), the overwhelming fraction of annihilation events occur after thermalization, when positrons spend most of their time in diffusive motion.

Although positrons rapidly (i.e., in approximately picoseconds) thermalize in condensed matter, and so can be thought of as quasi-stationary, the electrons that annihilate them are decidedly not. Conduction electrons at the top of the Fermi sea in metals, for example, have several electronvolt energies and bound electrons can have kilo electronvolt energies. However, most positrons will perish at the hands of the lower energy conduction or valence electrons, both because of the time of interaction and because of their propensity to spend more time away from the positive ion cores.

If we initially treat the positron and electrons as independent particles (i.e., we ignore mutual interactions), then the probability $\rho(\boldsymbol{p})$ that a pair of γ-rays with total momentum $\boldsymbol{p}$ will result from the annihilation of a positron and an electron is given by

$$\rho(\boldsymbol{p}) = \left(\frac{r_0^2 c}{8\pi^2}\right) \sum_{i,j} n_i{}^{+} n_j^{-} \left| \iiint \exp (-i\boldsymbol{p}\boldsymbol{r}) \psi_i{}^{+}(r) \psi_i^{-}(r) d^3 r \right|^2 \quad (5.2)$$

where n_i^+ and n_j^- are the occupancies of the positron and electron states, the former being close to a delta function, and $\psi_i^+(\boldsymbol{r})$ and $\psi_j^-(\boldsymbol{r})$ are the positron and electron wave functions, respectively. If we assume that the positron momentum is relatively negligible, then $\boldsymbol{p}$ represents the electron momentum at the moment of annihilation.

Direct complete evaluation of the so-called momentum density $\rho(\boldsymbol{p})$ would be possible if one could perform a three-dimensional measurement of the momenta of the annihilation photons. In principal, one could achieve this by measuring the x and y components via the angles between the photons (see Section 5.3) and the z component via the Doppler shifts in the measured photon energies. In practice, we integrate over the z component and record the two-dimensional contour "map", that is, $\rho(p_x, p_y) = \int \rho(\boldsymbol{p}) dp_z$. This is the basis of the experimental technique known by the acronym 2D-ACAR (two-dimensional angular correlation of annihilation radiation).

If instead we integrate over both x and y components of $\boldsymbol{p}$, then we arrive at the momentum density in the z direction, which gives rise to the Doppler broadening of annihilation radiation (DBAR), an extremely useful spectroscopy for following changes or differences in electronic structure: $\rho(p_z) = \iint \rho(\boldsymbol{p}) dp_x dp_y$.

Finally, the total annihilation rate is given by the integral of $\rho(\boldsymbol{p})$ over all three components of $\boldsymbol{p}$: $\lambda = \iiint \rho(\boldsymbol{p}) dp_x dp_y dp_z$. Reference to Eq. (5.2) shows that λ is proportional to electron density in the vicinity of the annihilated positron. The mean positron lifetime τ is the reciprocal of the annihilation rate λ, and measurement of the various possible lifetime components for positrons in condensed matter forms the basis of positron annihilation lifetime spectroscopy (PALS). The independent particle model leading to Eq. (5.2) is inadequate for the calculation of λ; $\rho(\boldsymbol{p})$ has to be multiplied by a positron–electron enhancement factor $\boldsymbol{\varepsilon}$ to take account of the interactions between positron and electrons and between the electrons in the material being studied (Kahana, 1963).

5.2.2.7 **Positronium Annihilation**

Although the quasi-stable bound-state Ps cannot form in metals or semiconductors, as a result of electron screening (except above the surface – as discussed later), it can exist in insulating solids and liquids.

A small variety of models have been proposed to describe Ps formation in such solids and liquids, and their applicability depends to a large extent on the material under study. The Ore Model (Ore, 1949) considers Ps formation to occur principally in the range of energies from $(E_i - B)$ to E_{ex}, where E_{ex} and E_i are the threshold energies for atomic excitation and ionization, respectively, and $B = 6.8$ eV is the ground-state binding energy of Ps. Above E_{ex} other interactions are assumed to outcompete Ps formation. A second prominent model is the Spur Model, due to Mogensen (1974), in which the positron binds to an electron released in a spur during the slowing-down process, under conditions of small relative momentum. An extension of this model is to consider the end of the positron track to be a ''blob'', rather than a spur (Stepanov and Byakov, 2002). A third, particularly considered with respect to Ps in polymers, involves the formation of Ps in open volumes or holes, the electron being picked up from the surface (Brandt, Berko and Walker, 1960); if the positron is not completely thermalized, then the Ps atom may undergo thermalizing collisions with hole walls.

There have been many groundbreaking fundamental measurements on Ps over the past 50 years, from precise measurements of vacuum decay rates to measurements on excited states, observations of the negative ion and hydride, and of Ps–Ps interactions en route to a Bose–Einstein Ps condensate (Cassidy and Mills, 2007). In Section 5.4, however, we briefly summarize only some of those applications in which Ps decay has been used to characterize condensed matter on the atomic scale, mostly involving the measurement of Ps lifetimes and intensities.

5.3
Experimental Methods

5.3.1
Sources of Positrons

The two methods for obtaining positrons for use in spectroscopic measurements are (i) radioactive sources and (ii) pair production. The most common radioisotope used in the laboratory is sodium-22 (half-life 2.7 years), with cobalt-58 sometimes used in experiments requiring higher intensities over shorter periods (half-life 71 days). These and other positron-emitting sources are created in reactor cores by nuclear bombardment; a number of steps have been taken toward the in-house creation of strong positron sources using tabletop proton and deuteron accelerators to produce short-lived but high-activity radionuclides (Hirose, Nakajyo and Washio, 1997). In-house or facility-based sources are primarily used in the production of intense positron beams, as are positrons created by pair production in a linear accelerator (LINAC). In the latter case, bremsstrahlung radiation from pulsed energetic LINAC electrons created electron–positron pairs in a target of high atomic number; the energetic positrons are then moderated to form an intense, often pulsed, positron beam.

Positrons emitted via the positive β-decay of radioactive sources are longitudinally polarized (i.e., spin-polarized in the direction of their emission) as a consequence of parity nonconservation; the weak interaction that mediates β-decay leads to a nonvanishing helicity. This property has been exploited in a number of experiments to study the bulk magnetic properties of materials, principally using ACAR (Section 5.3.3). The principle underlying its application is that positrons are $\sim$1000 times more likely to annihilate an electron in the opposite spin state. Therefore, if the spin states of some of the electrons in a sample are changed (e.g., by measuring magnetic and nonmagnetic samples, or by changing the external magnetic field direction), the difference in the experimental data yields information specifically on those electrons and thus on the magnetic species.

5.3.2
Lifetime Spectroscopy

The first positron spectroscopy to be developed, and still widely used currently, involves the measurement of the mean lifetimes of positrons in condensed matter. As introduced in Section 5.2.2.5, the mean lifetime of positrons in condensed matter is related directly to the electron density in the vicinity of the positrons at their moment of annihilation. For example, if positrons are annihilated while trapped in open-volume defects, the average electron density is reduced and the mean lifetime is increased, and there is, in general, a characteristic lifetime value for every different annihilation state.

In the most general case, positrons can decay from a range of states s each with annihilation rates $\lambda(s)$ (λ being the reciprocal of the mean lifetime τ), with the probability of being in state s at annihilation being $P(s)$. The resultant lifetime spectrum is

$$I(t) = \int P(s)\exp(-\lambda(s)t)ds \tag{5.3}$$

which, if the annihilation proceed via a small, discrete states labeled i, reduces to

$$I(t) = \sum_i I_i \exp(-\lambda_i t) \tag{5.4}$$

This spectrum can be fitted with standard programs to extract the decay rates (and thus lifetimes) and corresponding intensities, deconvoluting the time resolution of the measuring system and allowing for system-related components such as that associated with positron decay in the source. In practice, the number of discrete components that can be reliably obtained from experimental spectra is crucially dependent on the range of values of λ_i and the number of states involved; a maximum i value of 3 is typical. A common strategy in many experiments is to use the average positron lifetime $\overline{\tau}$ (or decay rate $\overline{\lambda}$), defined as

$$\begin{aligned}\overline{\lambda} = \overline{\tau}^{-1} &= \int \lambda(s)P(s)ds \quad or \\ &= \sum_i I_i\lambda_i \end{aligned} \tag{5.5}$$

particularly when it is the change in λ with some external parameter that is important.

5.3.2.1 Experimental Lifetime Systems

Conventional positron lifetime spectrometers have at their heart a radioactive source, commonly sodium-22 deposited between two thin low-Z films, sandwiched between two pieces of the sample to be studied. If the sample is to be cooled or heated during the measurements, it may be mounted in a small evacuated chamber. Sodium-22 provides a "prompt" 1.28-MeV γ-ray essentially coincident with the emission of the positron, as a result of the de-excitation neon-22; this is detected by a scintillator with a fast response (e.g., barium fluoride or doped plastic) coupled to a photomultiplier. The 511-keV annihilation "death" γ-ray is recorded by a second similar detector assembly, and the fast pulses (rise time ~1 nanosecond) created are fed into a timing system, which records up to about 10^6 time intervals between birth and death γ-rays. Traditionally, the timing systems have included single channel analyzers to select pulses of appropriate

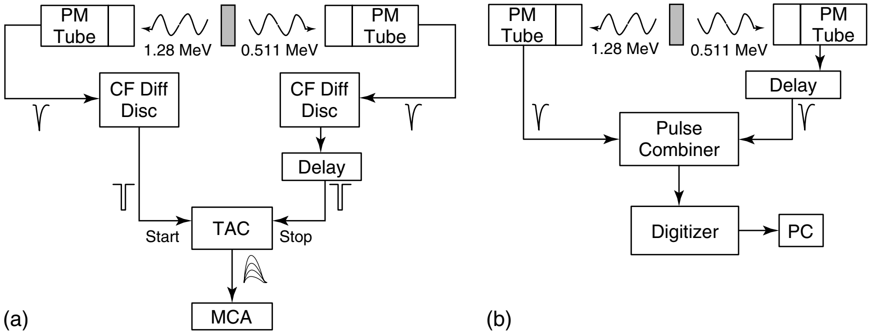

Fig. 5.1 (a) Standard fast–fast analog timing system: PM TUBE, photomultiplier tube/scintillator assembly; CF DIFF DISC, constant fraction differential discriminator; DELAY, nanosecond delay box or cable; TAC, time-to-amplitude converter; MCA, multichannel analyzer. (b) Digital equivalent of an analog timing system.

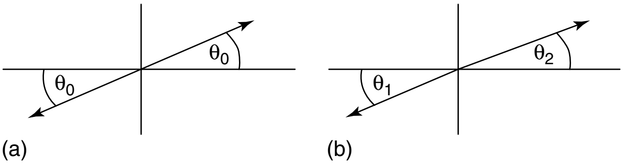

Fig. 5.2 A 2-γ annihilation in (a) center-of-mass frame and (b) laboratory frame.

size, a time-to-amplitude converter (usually with a suitable delay line before the "stop" input) and a multichannel analyzer to collect the entire range of pulse heights in parallel (i.e., the entire range of positron lifetimes). In recent years, this analog system has been replaced by a digital equivalent in which every pulse representing a time difference is digitized and stored, allowing postcollection timing optimization (Figure 5.1).

5.3.3
Angular Correlation of Annihilation Radiation

Although, as discussed in Section 5.2.1 for stationary particles, in two-γ annihilation in the center of mass frame the two photons leave the site of the annihilation with identical energies (mc^2, 0.511 MeV) in exactly opposite directions (to conserve momentum), in the laboratory frame, the momentum of the annihilating pair moving with a velocity v means that the angle between γ-rays – as shown in Figure 5.2 – is no longer 180°.

Relativistic transformation from one frame to the other yields expressions for $\tan\theta_1$ and $\tan\theta_2$. Then $\delta\theta \approx \tan(\theta_1 - \theta_2)$ which, ignoring terms in $(v/c)^2$, reduces to $(2v/c)\sin\theta_0$. Now if we set $\theta_0 \approx \theta_1 \approx \theta_2 = \theta$, then $\delta\theta \approx 2mv\sin\theta/mc = p_t/mc$, where p_t is the component of momentum of the annihilating pair in a direction transverse to the γ-emission. Measurement of $\delta\theta$ thus directly yields information on p_t. $\delta\theta$ typically ranges up to 20 mrad; in order to achieve the necessary resolution, the γ-detection regions have to be small and spaced many metres from the sample being

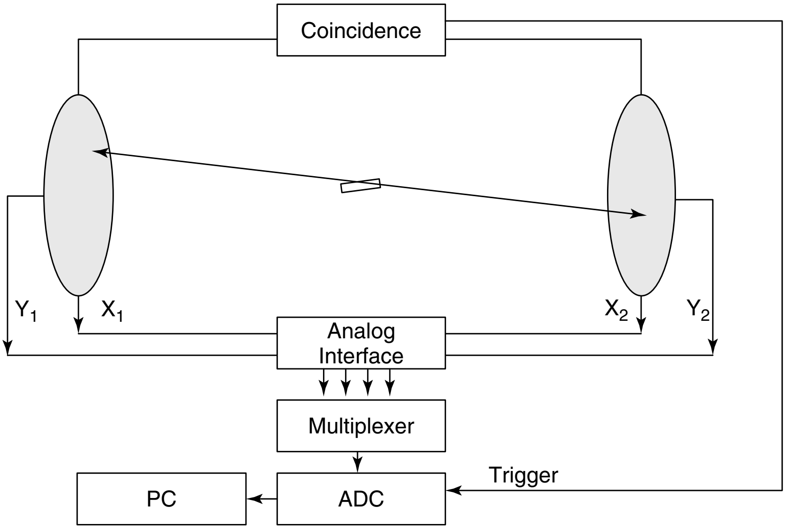

Fig. 5.3 Schematic diagram of 2D-ACAR apparatus. Positrons are guided from source to target by a strong (~1 T) magnetic field. (Courtesy Bristol positron group.)

studied. In one-dimensional studies (1D-ACAR) long thin slits – usually between lead blocks – define the γ-directions; in two-dimensional (2D-ACAR) experiments (Figure 5.3) annihilation photons are detected by position-sensitive detectors such as Anger cameras or multiwire ionization chambers. To effectively eliminate background events, positrons are guided from the approximately giga becquerel source to the samples by a strong (~1 T) magnetic field over a few centimeters from a radioisotope source, or by a guiding magnetic field in a low-energy positron beam system. About 2×10^8 coincident annihilation events are typically recorded. Samples are commonly mounted in a small vacuum chamber on a cold finger to enable measurement at temperatures down to $\sim 10^1$ K, and can be fixed at different crystalline orientations. Anger cameras have spatial resolution of 2–3 mm, so that if they are mounted at ~15 m on either side of the sample they provide angular resolution of ~0.2 mrad. A somewhat better resolution is possible with multiwire proportional counters. These high-resolution capabilities have found their place in the measurement of occupied electron states, and thus the Fermi surfaces of a number of simple and complex solids (see Section 5.4.5 for examples). The Lock–Crisp–West (LCW) procedure (Lock, Crisp and West, 1973) folds the momentum density back into the first Brillouin zone, which then gives a distribution that directly indicates the occupancy of states across the projected zone. (The Fermi surface is defined by whether a particular state is occupied or not for each band, and these occupancy breaks are revealed directly by 2D-ACAR.) The full three-dimensional Fermi surface can be reconstructed tomographically from an appropriate number of two-dimensional projections along different crystallographic orientations.

5.3.4
Doppler Broadening of Annihilation Radiation

The component of the momentum p_p of an annihilating positron–electron pair in the direction of emission of the annihilation γ-photon gives rise to a Doppler shift in the photon energy of $\pm cp_p/2$. For example, if p_p arises almost wholly from an electron moving close to the Fermi level with a kinetic energy of 5 eV, $cp/2 = 1130$ eV. This is of the same order as the energy resolution of a high-purity germanium detector. Because the electrons can have a component of momentum toward or away from the detector at the moment of annihilation, Doppler broadening around mc^2 (511 keV) is measured. Changes in the annihilation linewidth from a fraction of 1% to a few percent are typically measured and, although the technique provides a measure of the average electron momentum, which is of considerably lower resolution than ACAR, its high signal rates and relative simplicity has found wide application in the study of defects and phase changes in materials. This is referred to as *Doppler broadening spectroscopy* (*DBS*) or *DBAR*.

A typical system for DBAR is sketched in Figure 5.4. The Ge crystal is cooled by liquid nitrogen or by an electrical cooler to effectively eliminate thermal noise. A preamplifier is routinely mounted inside the cooled detector head. The optional biased amplifier allows expansion of the photopeak into a larger number of channels on the multichannel analyzer (MCA). Since small drifts can broaden the spectrum, stabilization is necessary, and this is achieved by temperature control and by using a digital stabilizer set to a nearby reference line – for example, the ^{7}Be 478-keV γ-line. Digital systems have been developed, which house all the elements shown in Figure 5.4.

Samples for DBAR are typically mounted in an evacuated chamber to avoid sample contamination on heating or cooling, and data contamination by annihilation signals in air; such mounting also enables temperature control. For traditional bulk measurements using a radioactive source, two pieces of the sample material sandwich a source held in low-Z films, and correction has to be

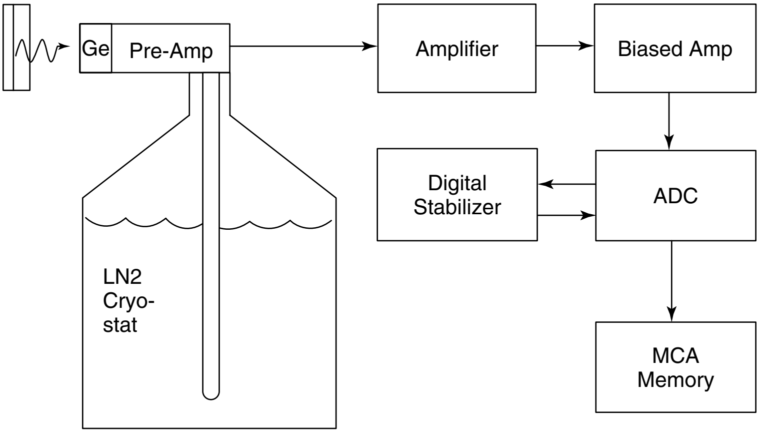

Fig. 5.4 Basic setup for Doppler broadening spectroscopy.

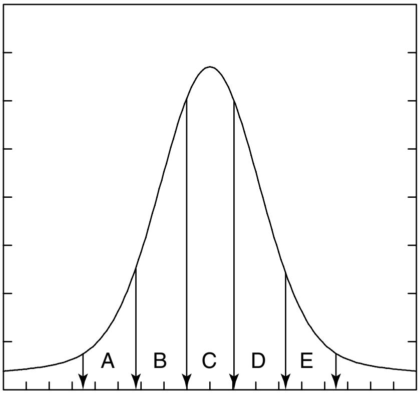

Fig. 5.5 Regions of interest defined for DBAR: sharpness parameter $S = C/T$, wing parameter $W = (A + E)/T$, where $T = A + B + C + D + E$.

made for annihilations in the source itself; if the sample is at the target end of a positron beam system, then no source correction is necessary.

Resultant "photopeak" spectra corresponding to the annihilation γ-line are quasi-Gaussian in shape; entirely free (i.e., conduction) electrons would lead to a parabolic peak, and core electrons to a number of increasingly small "side" peaks; the sum of these contributions has a Gaussian, or multi-Gaussian, appearance. The central part of the annihilation line thus preferentially contains contributions from the lower momentum conduction and/or valence electrons, while the wings contain contributions from annihilation with core electrons. Analysis routines have been developed which take account of the detailed shape of the peak, with a view to maximizing the amount of information obtained in Doppler broadening experiment, although it is more common to use the raw data to track *changes* in lineshape parameters under changing experimental conditions.

As discussed later, by decreasing background significantly, detailed analysis of the peak shape can exploit the technique's sensitivity to the elemental environment. However, we first consider the standard method for describing the Doppler-broadened linewidth – that is, the use of simple lineshape parameters. The most common parameters used are called S and W – the sharpness and wing parameters, respectively – which are defined in Figure 5.5. Three regions of the peak – here A (central) and C and E (the two wings) – are chosen, symmetrically about its centroid. The limits of these regions are chosen such that the central and total wing fractions of the peak area – which are called the S and W parameters – are $\sim$0.5 and $\sim$0.1–0.2, respectively. Background is first subtracted if deemed necessary. Following our argument above, S and W should principally reflect changes in the momentum density of lower and higher momentum electrons, respectively. For example, positron annihilation in open-volume defects typically leads to an increase in S and a decrease in W.

Neither S nor W has an absolute value, as it is their dependence on external parameters that is of common interest. However, it is common to express the parameters normalized to bulk values – that is, S/S_{bulk} and W/W_{bulk}, where the denominators are the parameters associated with the defect-free bulk material being studied, measured with the same apparatus. The choice of which parameter to use (S and W are not the only choices) and the limits of regions defining the parameters are system-dependent.

If S_f and W_f are the S and W parameters characteristic of free positrons, then we can define a new parameter $R = |(S - S_f)/(W - W_f)|$, which depends only on the nature of the defect, and not on its concentration. If R is found to be constant, then this points to the existence of only one kind of defect (Mantl and Trifthäuser, 1978).

5.3.4.1 Two-Detector DBAR

This technique was proposed many years ago by Lynn *et al.* (1977) but has been reintroduced and developed in the past decade. If a sample is viewed by a second γ-photon detector, and pulses from the two detectors (in coincidence) are fed into the inputs of a two-dimensional multichannel analyzer array, a photopeak results that not only has vastly reduced background (i.e., by two or more orders of magnitude) but the resolution is also decreased by a factor of $\sqrt{2}$. Example data for such a system are shown in Figure 5.6. This method enables one to study annihilations with core electrons and, in the case of recent studies with positron beams, identify the chemical environment in which the positron decays. The measurement typically involves the accumulation of single annihilation line spectra of high statistical precision whose shape is analyzed carefully and compared with calculations.

5.3.4.2 Age-Momentum Correlation (AMOC)

Simultaneous measurement of positron lifetime and the momentum of the annihilating pair (i.e., PALS + DBAR) can give information on thermalization and transitions between positron states (and hence on chemical reactions of positrons or Ps). The most recent version uses a mega electronvolt positron beam (Siegle *et al.*, 1997).

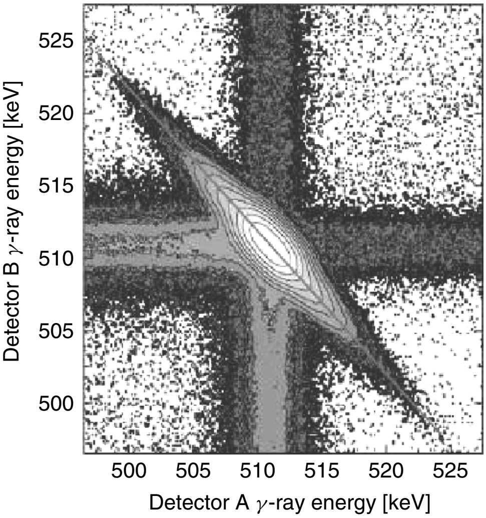

Fig. 5.6 A 2D raw data from coincidence DBAR measurement. The peak, indicated by the diagonal from top left to bottom right, is $\sqrt{2}$ narrower than a single-detector spectrum, and has essentially no background. (Courtesy Halle positron group.)

5.3.5
Positron Beams

Positron beams offer two major advantages over traditional positron systems: (i) the control of incident positron energy, and thus average implantation depths from the surface to depths of a few micrometers, and (ii) the separation of the thermalization of positrons implanted into a material from their eventual annihilation in another. This has led to new applications in surface, near-surface, interface, and thin-film studies.

The reemission of work–function–energy positrons from a solid surface was outlined in Section 5.2.2.5, Positron Reemission. In the case of simple metals, the fraction of implanted positrons reaching an exit surface is optimized if the moderator metal contains very few nonequilibrium defects. This means annealing, preferably *in situ*, to as high a temperature as possible (say $\sim$0.8 T_m) in as low an ambient pressure as possible (e.g., $\sim 10^{-6}$ Pa or less). Because of its high positron work function ($\varphi_+ \sim 2.7$ eV), however, tungsten operates well as a positron moderator after greatly varying preparation procedures. The same is not true, for example, with nickel or copper – whose $\varphi_+ \leq 1$ eV – for which careful surface cleaning and maintenance is required.

Many moderator geometries have been used; the simplest and most common is the mesh, which has quasi-transmission geometry. Moderator efficiencies are commonly in the $10^{-4} - 10^{-3}$ range. Choice of moderator material and geometry is governed by the application of the positron beam – for example, for many applications, the priority is simply to maximize moderator efficiency, whereas for others a well-collimated parallel beam requires a planar low-ϕ_+ surface cooled to minimize thermal smearing.

The highest moderation efficiencies (above 10^{-3}) recorded are for solid rare gases, in which positrons do not fully thermalize and have a long effective diffusion length, which can be condensed directly onto the radioactive source capsule.

It has long been recognized that moderation efficiency could be greatly enhanced by drifting a larger fraction of thermalized positrons to the exit surface by an internal electric field, but there has been little practical progress to date on the realization of such field-assisted moderators.

5.3.5.1 Laboratory-Based Beams

A standard slow positron beam system (Figure 5.7) comprises a flight tube pumped to high or ultrahigh vacuum; positrons ($\sim 10^5$ s^{-1} from a $\sim$4 GBq primary source) are transported in an axial magnetic field, sometimes with focusing elements. Unmoderated (β) positrons may be filtered from the beam by a curved section or an ExB velocity filter and the positrons are accelerated to the final desired energy either by raising the source of the system to a positive potential or by holding the sample at a negative potential. Beam positioning may be fine-tuned by using a pair of trim coils. Annihilation radiation from the sample target passes through a thin foil window to the Ge detector for DBAR measurements; detection of radiation from the source is minimized by shielding and by making the system a few metres in length. These considerations are relatively unimportant if the

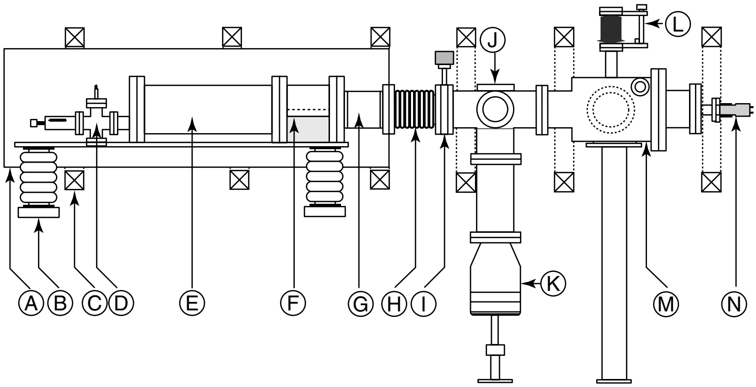

Fig. 5.7 Magnetic-transport positron beam system. A, grounded shield; B, standoff insulators; C, coils for magnetic field; D, source/moderator; E, *ExB* plates; F, lead shielding; G, accelerator; H, bellows; I, aperture; J, guiding coils; K, turbopump; L, sample manipulator; M, sample chamber; N, CEMA/CCD camera.

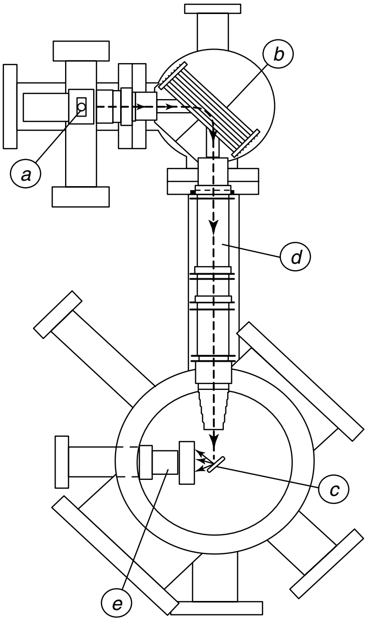

Fig. 5.8 Example of an electrostatic positron beam system. (a) Source, (b) electrostatic reflector, (c) sample, (d) electrostatic lenses, and (e) microchannel plate detector.

system is to be used for particle spectroscopies – see Figure 5.8 – especially if electrostatic transport and focusing is in use (Roach, Bakshi and Canter, 1995).

Brightness enhancement is achieved in positron beams by repeated focusing and remoderation (see Figure 5.9). Although remoderation losses may be 70%, the

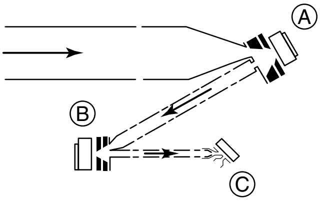

Fig. 5.9 Reflection-geometry brightness enhancement stage. A and B, remoderator surfaces; C, target.

beam area can decrease by a factor of ~50 at each stage; hence beam intensity per unit area can increase after *n* stages by a factor 15^n.

5.3.5.2 High-Intensity Positron Beams

Standard laboratory beam intensities up to $\sim10^6$ controllable-energy positrons per second are currently available, and a great range of experiments have been performed with such beams. However, to extend the capabilities and realize the full potential of positrons as a spectroscopic tool, more intense beams are essential for many studies in which the probability of an outcome is very small, for example, the combination of controlled implantation with 2D-ACAR, many surface spectroscopies where rapid measurements are required, and the study of many-positron systems.

Facility-based systems – both reactor and LINAC based – are still a main focus for intense positron beam generation, but in recent years there has been a movement toward the use of small-scale laboratory-based accelerator systems to produce short-lived but intense positron sources. In the case of LINACs (Section 5.3.1), the beam is pulsed (Section 5.3.5.3); this property can be exploited in applications where timing is an advantage. In laboratories in reactor complexes, an intense positron beam is guided far from the core and possibly subjected to remoderation, focusing, and bunching for timing applications. The production of positrons in reactor cores is not solely via the creation of intense radioactive sources, but can also proceed via pair production by energetic photons; steps have to be taken to overcome radiation damage problems in the reactor core. Several intense beam systems are in operation or are currently being developed across the world.

5.3.5.3 Beam Bunching

LINAC beam pulses are of widths from several nanoseconds to microseconds; if subnanosecond timing resolution is required, the positrons must be put through a buncher. Several buncher designs have been put into practice; the first, by Mills (1980), used magnetic mirrors. Later bunchers used radiofrequency chopping and bunching techniques – for example, the system at Munich (Bauer *et al.*, 1987), which can now achieve a timing resolution below 200 picoseconds.

5.3.5.4 Positron Microbeams

Beams of approximately micrometer dimensions have been created in the laboratory (Figure 5.10) for (i) optimum areal brightness and (ii) position-sensitive annihilation spectroscopy. Optimum areal brightness is important in positron reemission microscopy systems in which the microscope optics requires a small but very bright spot – the contrast mechanism in positron reemission microscopy (PRM) is based on the position dependence of the probability of positron

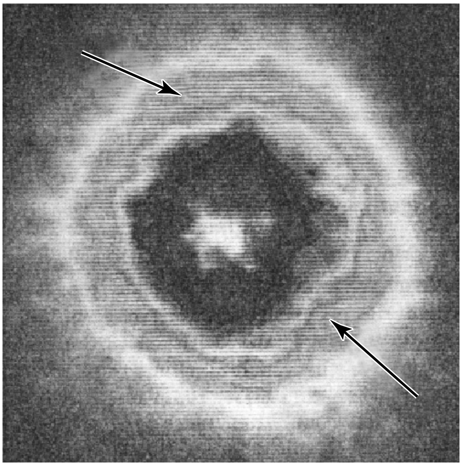

Fig. 5.10 A 10-μm square intensity cross-sectional image of a 5-keV positron microbeam at Brandeis (Canter *et al.*, 1994). The central white peak area contains approximately 465 counts per pixel, and the ring marked with arrows contains about 180 counts per pixel. Reproduced by permission, AIP.

reemission from a surface. Position-sensitive annihilation spectroscopy is required for the building of micrometer-resolution annihilation maps of near-surface structural defects by beam rastering, as has been achieved in Munich (David *et al.*, 2001).

5.3.5.5 **Polarized Positron Beams**

It was discovered in 1979 that the depolarization resulting from moderation of the β-positrons by a moderator, the basis of the creation of controllable-energy positron beams, is negligible (Zitzewitz *et al.*, 1979). Highly polarized positron beams can, in principle, probe, nondestructively, and with a degree of depth sensitivity, spin-polarized electrons in thin films or in modified near-surface regions (to depths of a few micrometers) of materials. Standard laboratory-based beams have to be adapted in two main ways: (i) the high-Z backing behind the radioactive source has to be replaced by a low-Z material to minimize backscattering, and (ii) the source has to be positioned further away from the moderator so as to select primary positrons emitted into a small forward cone (both (i) and (ii) result in loss of beam intensity). The spin polarization of a positron beam is given by

$$P = \left(\frac{v}{2c}\right)(1 + \cos\,\alpha) \qquad (5.6)$$

where v is the emission velocity, c the speed of light, and α the half-angle of the cone of acceptance of β-positrons at the moderator. (v/c is called the *helicity*.) For positrons emitted in a cone of half-angle 30° from the commonly used source ^{22}Na, $P \sim 70\%$.

There has been little exploitation of this property of positron beams, apart from an early experiment that demonstrated their capacity to probe surface magnetism (Gidley, Köymen and Capehart, 1982). The principle of the measurements is based on the comparison of DBAR or annihilation line profile data for differently polarized samples.

5.3.5.6 **MeV Positron Beams**

High-energy (approximately mega electronvolts) monoenergetic positron beams have been built and used in recent years. Megavolt accelerators used in this work have been of Pelletron or Van de Graaff type. Source-free lifetime measurements in a wide range of materials, beam-based age-momentum correlation (AMOC), and novel annihilation spectroscopies are possible with such systems – for example, positron channeling and in-flight annihilation studies, and bulk annihilation measurements, which are free of problems associated with the presence of a positron source.

5.3.5.7 Trap-Based Beams

Surko and co-workers have developed a source of very low energy positrons with extremely low energy (approximately milli electronvolts) energy spread (Gilbert *et al.*, 1997), which has found wide and important applications, particularly, but not solely, in atomic and molecular physics. The principle of their system is that positrons are extracted from a plasma of thermalized positrons stored in a Penning-type trap (Figure 5.11). The resulting beam can be DC or pulsed. Similar trap-based positron beams have been successfully employed in the antihydrogen work at CERN and other laboratories prosecuting new fundamental research.

5.3.5.8 Positron-Beam-Based Spectroscopies

The traditional spectroscopies (PALS, ACAR, and DBAR) have all been performed with positron beams, giving the extra information on depth dependence and the ability to probe depths up to approximately micrometers from a sample surface. The acronyms for such spectroscopies may be constructed by inserting VE (for variable energy) before those for the bulk spectroscopies. In all these cases, codes such as VEPFIT (Van Veen *et al.*, 1990) and POSTRAP (Aers *et al.*, 1995) are commonly used to separate and identify contributions to the measured parameters from annihilation events over a range of depths from different states such as vacancy-type defects, bulk material, interface regions, and the sample surface. The diffusion equation is solved – for example, with VEPFIT – and the positron parameter (like the lineshape parameter S) and associated effective positron diffusion length, characteristic of each chosen layer below the surface, are evaluated to achieve best fit to the experimental data. In some cases – for example, where there is one type of trapping center of concentration C_D per atom – the fitted parameter and diffusion length are linked, and self-consistency can be

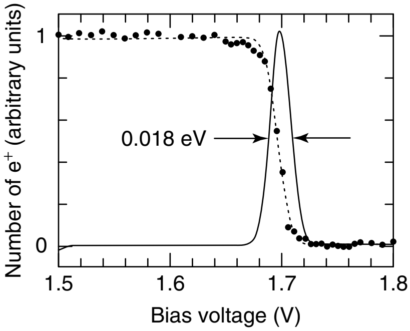

Fig. 5.11 Energy distribution of a pulse of positrons extracted from a thermalized room temperature positron plasma stored in a Penning trap (Gilbert *et al.*, 1997). Reproduced by permission, AIP.

checked. As an example, let S be the fitted lineshape parameter for a particular layer beneath a surface (i.e., between two given or fitted depths). Then $C_D = (\lambda/\nu)(S - 1)/(S_D - S)$, where λ and ν were defined in Section 5.2.2.4. Additionally C_D can also be written in terms of the effective diffusion length L as $(\lambda/\nu)[(L_+/L)^2 - 1]$, where L_+ is the positron diffusion length in the undefected "perfect" material. Equating these two expressions one arrives at $L^2 = L_+{}^2[(S - S_D)/(1 - S_D)]$.

The depth resolution of position beam techniques degrades as the mean positron implantation depth increases; the width of $P(z, E)$ (Section 5.2.2.2) is comparable to the mean depth. This means that, although resolutions of $\sim$50 nm are possible for near-surface characterization, one relies on fitting codes or known sample parameters (like film or layer thicknesses) to model depth profiles at deeper depths. The only method for maintaining depth resolutions similar to that near the surface is to employ an etch-and-measure technique, removing sequentially thin layers of material and probing the progressively revealed new surface regions with low-energy positrons. The problem with very near surface layers or thin films is the propensity for positrons to diffuse out of them, either to the surface or to nearby material, and careful fitting and/or data interpretation is required; this problem is less severe for thin layers with a very short positron diffusion length (e.g., oxides). The effects of internal electric fields must also be considered. Finally, the influence of nonthermalized positron interactions at low incident positron energies E has to be recognized, and for this reason data for E below $\sim$1 keV is often ignored.

An alternative to the use of fitting codes is the identification of different annihilation sites (e.g., layers) using the graphical method of plotting S versus W parameters for each incident energy E. An example is shown in Section 5.4.4. If there are two possible annihilation states with their characteristic S and W values – such as the sample surface (subscript S) and bulk (subscript B) – then an $S{-}W$ plot will be a straight line joining the two points (S_S, W_S) and (S_B, W_B) on the graph. Consider a third state, corresponding, say, to a defected layer with characteristic $S = S_D$ and $W = W_D$ (note that these are average *layer* values, not those specific to a particular defect); if there are energies E for which effectively all the positrons are annihilated in this layer, then the $S{-}W$ plot will have two straight lines joining (S_S, W_S), (S_B, W_B), and (S_D, W_D). Importantly, however, even if only a *fraction* of the positrons decay in the layer at any energy E, extrapolation of lines on the resulting $S{-}W$ plot will still identify the point (S_D, W_D) – just as a fitting code such as VEPFIT does. $S{-}W$ plots are thus a powerful visual tool for identifying different annihilation sites.

Positron Surface Spectroscopies Notwithstanding the fact that full realization of positron beams as a surface probe via positron diffraction, reemission, or annihilation will not come until intense positron beams are widely available, much progress has been made over the past 25 years. Positron microscopy is not mentioned in this section as it was discussed earlier in Section 5.3.5.4.

Low-energy positron diffraction (LEPD) was pioneered by Canter and co-workers at Brandeis University in 1980 (Rosenberg, Weiss and Canter, 1980) and developed by the same group over the next two decades to a level at which they could demonstrate that LEPD can be used to achieve, qualitatively and quantitatively, better agreement between experimental and theoretical $I-V$ profiles (diffracted beam intensity vs. energy), leading to significantly more reliable determinations of the surface structure than is possible using the traditional electron equivalent, low-energy electron diffraction (LEED). This improved performance is for the following reasons: (i) the phase shifts for positron scattering are less sensitive to atomic number than those for electrons, so that LEPD is more sensitive to structural parameters in multicomponent systems; (ii) the inelastic mean free path, which plays an important role in the determination of probe depth, is smaller for positrons than for electrons, and so LEPD has a greater surface sensitivity than LEED; (iii) uncertainties in the positron–electron correlation term used in LEPD calculations are less important than the equivalent electron–electron uncertainties in LEED; (iv) positrons are decelerated as they approach ion cores and so relativistic effects such as spin-orbit coupling are reduced for positron scattering from surfaces containing high-Z atoms, and LEPD $I-V$ profiles are only weakly spin-dependent when compared to LEED; and (v) as a result of the absence of spin-exchange repulsion, positrons interact more weakly with interstitial valence electrons and so the muffin-tin model works better for LEPD than for LEED for covalently bonded semiconductors.

Reflection high-energy positron diffraction (RHEPD) has been demonstrated by Kawasuso and Okada (1998). A 20-keV highly parallel positron beam is incident upon the surface being studied at glancing ($<5^\circ$) angles. The most important difference between RHEPD and its electron equivalent, RHEED, is that the high-energy positrons can undergo total reflection from the surface because of the positive crystal potential, and are thus very sensitive to the presence of adsorbate atoms, topological irregularities, and lattice vibrations in the topmost surface layer.

Positronium reflection from LiF was observed by Weber *et al.* 1988; to date, it is the only report of experimental work in this area.

Positron-annihilation-induced Auger electron spectroscopy (PAES) was developed by Weiss *et al.* (1988) in the late 1980s. Auger electron emission in PAES results from annihilation of a surface core electron by a positron implanted with very low ($\sim$10 eV) energy, rather than from impact ionization as in its electron equivalent, electron-induced Auger electron spectroscopy (EAES). The two major advantages of PAES over EAES are the elimination of secondary electron background and the extremely high surface sensitivity. Ohdaira *et al.* (1997) have developed a time-of-flight PAES system at ETL, Japan.

Positronium emission spectroscopies can provide information on electronic surface states; for example, assuming that Ps formation is a sudden process, a measurement of the Ps velocity distribution should yield information on the electronic density of states, although there remains a discrepancy between theory and experiment at low Ps energies.

2D-ACAR was first used in the mid-1980s to attempt to directly observe the positronic surface state; the observed symmetrical momentum distribution may have instead been a signature of localized surface trapping (Lynn *et al.*, 1985). 2D-ACAR was, however, successfully used to study Ps momentum distributions, and showed sensitivity to electronic structure (Chen *et al.*, 1987).

Reemitted positron spectroscopy measures the energy spectra of positrons reemitted from a film-covered surface as a function of overlayer thickness (Gidley, 1989; Ociepa *et al.*, 1990), which has been shown to be an excellent probe of any processes that affect the sum of the positron and electron bulk chemical potential, including alloying of the overlayer film.

Positronium formation spectroscopy, that is, monitoring variations in Ps yield (both in magnitude and dependence on incident positron energy) or in the Ps contribution to 2D-ACAR spectra, is an extremely sensitive tool for monitoring oxide growth on metals and semiconductors.

Future developments in positron surface studies may include *positron-induced ion desorption, surface barrier potential measurements* via very low energy (approximately electronvolts) positron reflection, glancing-angle RHEPD, and the dependence of the Ps formation probability on incident positron energy (coupled with LEED, for the surface atomic configuration, and angle-resolved ultraviolet photoemission spectroscopy, for information on electronic states). Similar measurements can give information on the atomic positions of the adatoms, even if these are hydrogen (in contrast to the conventional angle-resolved photoemission). In *inverse Ps formation spectroscopy*, a beam of Ps atoms bombards a surface, the electron is given up to an unfilled state, and the positron takes away information on that state; this spectroscopy should be more sensitive than existing probes.

The advance of positron surface science, however, awaits more widespread availability of intense positron beams. Examples of developments that may flow from such availability include *positron holography*, proposed as intrinsically more suitable than electron holography because of the positron's weak scattering and large damping in solids; *polarized PAES*, in which a highly polarized incident beam of positrons creates polarized core holes to enable novel studies of magnetic surfaces; *reemitted positron energy loss spectroscopy; Ps diffraction*; and *inverse Ps formation.*

5.4
Examples of Research Using Positron Spectroscopies

The following examples are designed to provide an overview rather than a detailed summary of every important result obtained using positron annihilation spectroscopy (PAS) and VEPAS techniques. A relatively small number of examples will thus be selected to demonstrate the applicability of the various techniques.

5.4.1
Vacancy-Type Defects

The propensity for diffusing positrons to trap in open-volume point defects has been the basis of one of the major applications of positron spectroscopy.

The measurement of formation enthalpies H_V^F for thermally generated equilibrium defects has long been a staple of PAS in bulk solids. Both DBAR and

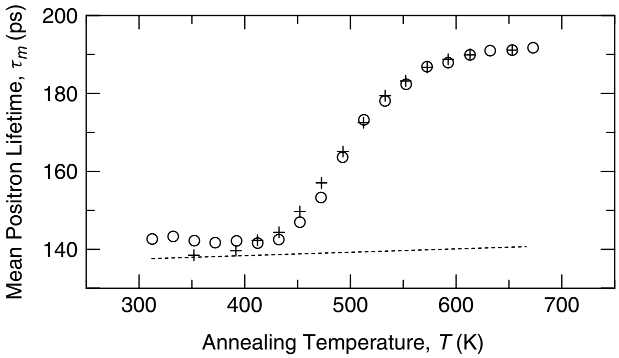

Fig. 5.12 Temperature dependence of the mean positron lifetime in Cu_3Sn. Error bars are within the points. Circles, ramping up; crosses, ramping down in temperature (Shishido *et al.*, 2007). Reproduced by permission, Wiley-VCH.

PALS have been used for such studies. Figure 5.12 shows a recent measurement of mean positron lifetime τ_m in the alloy Cu_3Sn; the rise in τ_m above 433 K indicates the formation of vacancy traps. The positron trapping rate κ is deduced from the measured positron parameters (either lineshape or mean lifetime – called P here) normalized to the bulk defect-free value P_b (so that $P_b = 1$) as

$$\kappa = \frac{(P-1)}{(P_d - P)} \tag{5.7}$$

where P_d is the parameter associated with the defect (in Figure 5.12, the high-temperature asymptotic value). The slope of the Arrhenius plot of κ versus $1/T$ then yields the vacancy formation enthalpy H_V^F. In this example, the value of H_V^F was similar to that measured for Cu, suggesting that the vacancies are being formed on Cu sites.

Nonequilibrium vacancy defects, such as those induced by strain or implantation, have also been widely studied by PAS. Figure 5.13 shows raw VEDBAR data with a typical response to defects (principally divacancies) produced by 2 MeV Si^+ ion implantation into Si for a range of ion doses from 10^{11}–10^{15} cm^{-2} (Coleman, Burrows and Knights, 2002). Observations from this and other measurements for mega electronvolt ions of widely different masses led to the universal expression for divacancy concentrations at approximately half ion range

$$C_D = (2.79 \times 10^{10})\,\phi_A{}^{0.63}\ \text{cm}^{-3} \tag{5.8}$$

where ϕ_A is the ion dose corrected by a multiplying factor equal to the vacancy/ion/Å value provided by the widely used simulation code SRIM (www.srim.org).

PAS can give information about dislocations – both in bulk materials and in thin films (e.g., relaxed SiGe) – primarily because it is thought that vacancies exist in kinks along the dislocation line and diffusing positrons are thus trapped by them.

5.4.2
Structural Changes

PAS has been, and continues to be, exploited in the study of structural

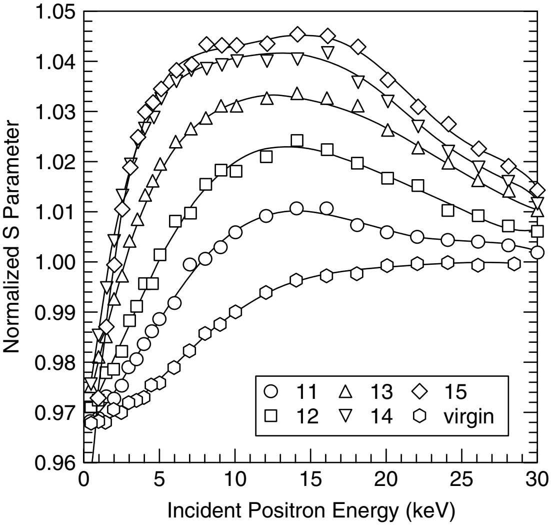

Fig. 5.13 Normalized S(E) for *FZ* Si unimplanted and implanted with 2 MeV Si^+ ions at doses of 10^n cm^{-2} ($n = 11, 12, 13, 14$, and 15) (Coleman, Burrows and Knights, 2002).

changes associated with phase transitions, precipitation, deformation, and so on, induced thermally (including aging at room temperature) or mechanically. For example, monitoring the S parameter as the steel Fe–Mn–Si–Cr–Ni is deformed (Mostafa *et al.*, 2007) shows no response below a few percent strain, as the macroscopic deformation associated with reorientation of martensite plates does not create microdefects that trap positrons. Above ~5% strain S increases as vacancy sites are created, some perhaps along dislocations, until at ~16% strain and above the response starts to saturate as all positrons become trapped; this behavior is also seen in CuZnAl shape memory alloys. The nonlinearity of the corresponding S–W plot (see Section 5.3.5.8) suggests that there is a variety of defect types created, and/or that the alloy undergoes a γ–ε phase transition during deformation (as seen in microscopic images).

5.4.3
Nanoparticles

An important recent advance in PAS has resulted from the observation that positrons can act as "magic bullets" in the study of embedded nanoparticles, either residing in the open volume around the particles or being preferentially attracted to them via their greater positron affinity. In both cases, the positrons are much more sensitive to the nanoparticles than would be expected from geometrical arguments alone. An example of this, which, in addition, illustrates the use of coincidence DBAR to gain chemical information on the atomic environment of the annihilated positrons, is shown in Figure 5.14 (Nagai *et al.*, 2000). The spectra shown are ratios of the outer

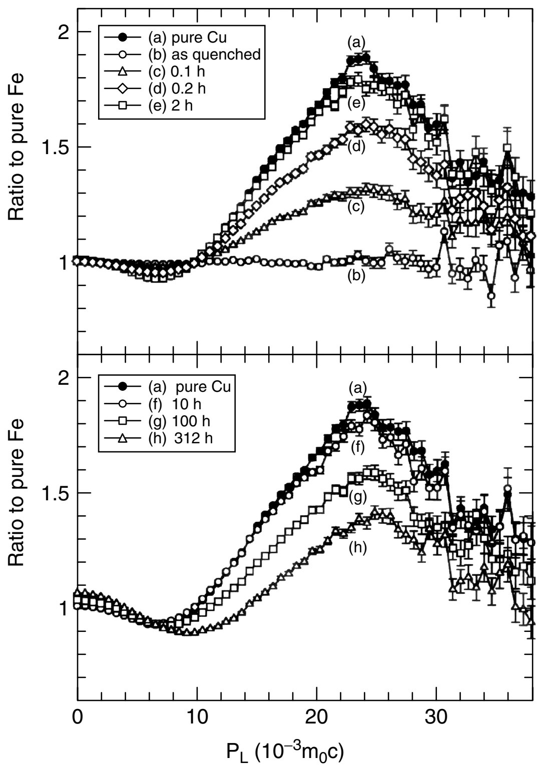

Fig. 5.14 CDBAR ratios of $Fe_{0.99}Cu_{0.01}$ to pure Fe: (a) pure Cu, (b) alloy as quenched, the after aging at 550 °C for (c) 0.1 hour, (d) 0.2 hour, (e) 2 hours, (f) 10 hours, (g) 100 hours, and (h) 312 hours (Nagai *et al.*, 2000). Reproduced by permission, APS: http://link.aps.org/abstract/PRB/v61/p6574.

wing parts of the measured annihilation line to that of Fe. Positrons are shown to be annihilated preferentially by Cu electrons as Cu nanoparticles form in a dilute (1%) alloy of Cu in Fe, by virtue of the 1 eV-deep well created by the relative positron affinities of Cu and Fe.

5.4.4
Interfaces

Generally, positrons are significantly more sensitive to interface states than geometric models would suggest, as they are commonly trapped there during diffusion. An example of VEPAS response to interfaces is shown in Figure 5.15, which also demonstrates the usefulness of the graphical parameter–parameter method; here, two *W* parameters, rather than the more usual *S* and *W*, have been plotted (Coleman *et al.*, 2007).

5.4.5
Fermi Surfaces

Knowledge of the details of Fermi surfaces can aid understanding of material properties, and 2D-ACAR represents a major tool for probing such surfaces. An example is a study of a rare-earth nickel borocarbide, which exhibits competing or coexisting antiferromagnetism and superconductivity. Dugdale *et al.* (1999) found evidence supporting this from the nesting behavior found in a particular sheet of the Fermi surface for $LuNi_2B_2C$ (Figure 5.16).

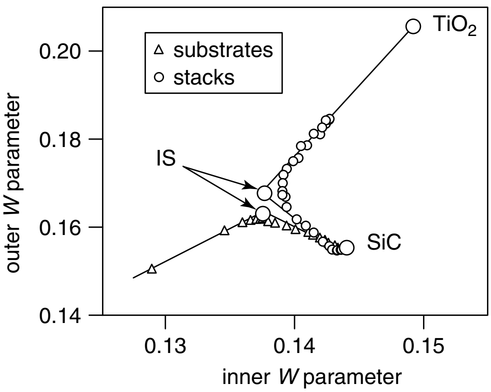

Fig. 5.15 $W_o - W_i$ map for TiO2 stacks on SiC stacks and for SiC substrates. The "pure state" points for SiC and TiO$_2$ are shown as circles, as are the two closely related interface state (IS) (Coleman *et al.*, 2007).

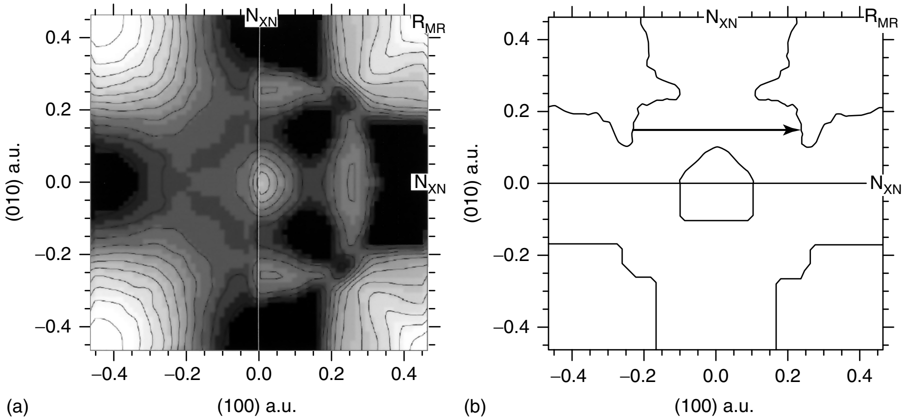

Fig. 5.16 (a): Experimental (left) and calculated electron density of LuNi$_2$B$_2$C projected along the [001] direction. Black signifies holes, and white represents electrons. (b): Fermi surface topology of LuNi$_2$B$_2$C – experimental (top) and theory for third band in the (001) plane through the Γ point. The arrow indicates the nesting feature. From Dugdale *et al.* (1999). Reproduced by permission, APS: http://link.aps.org/abstract/PRL/v83/p4824.

5.4.6
Nanoporous Materials and Open Volumes in Polymers

The application of Ps and positron lifetime measurement have long been used as a probes of open volume in polymers, often used in conjunction with other experimental techniques such as differential scanning calorimetry and ionic conductivity measurements. An example of a recent application was the study of polyethylene glycol dimethacrylate (Figure 5.17), where PALS was performed for temperatures in the range 100–370 K (Bamford *et al.*, 2001). These measurements allow the evaluation of the glass transition temperature, coefficients of expansion of the hole volume, fractional free volume, and hole number density.

There has been much activity in recent years in the area of pore size measurement and distribution in thin

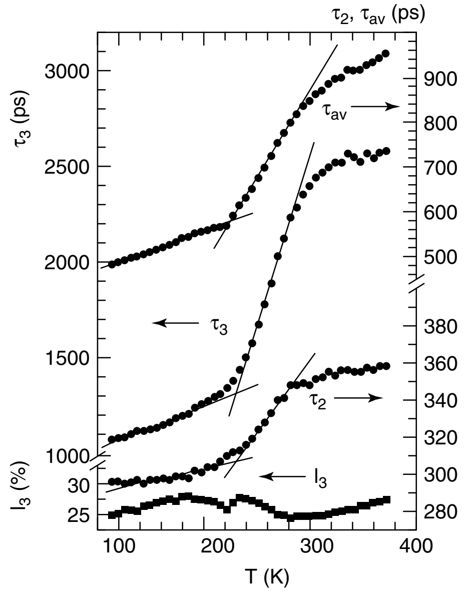

Fig. 5.17 The o-Ps lifetime τ_3, its intensity I_3, the e^+ lifetime τ_2, and the average positron lifetime τ_{av} in poly[(EG)$_{23}$DMA]; from Bamford *et al.* (2001). Reproduced by permission, AIP.

low-*k* dielectric layers grown for nanoelectronics applications. The basis of these measurements is that Ps can form in nanometer-sized pores, and its lifetime is linked to pore size. It is also possible to measure lineshape parameters and the ratio of three-γ (*o*-Ps) to two-γ decay events, both sensitive to the decay of Ps, to obtain some measure of the size, number density, and interconnectivity of the pores. These methods can be applied to films of thickness of approximately micrometers, and can provide some depth sensitivity, if linked to a positron beam system. Because Ps lifetimes can be approximately nanoseconds, timed beam systems can be constructed with relatively long timing resolutions, such as that of Gidley *et al.* (2000), Figure 5.18, who use the detection of secondary electrons to tag incident positrons. Interconnectivity can result in Ps escape as a long-lived naturally decaying entity in the vacuum space above the sample.

5.4.7
Surfaces

Although there have been several new initiatives in recent years in positron-surface studies, and activity in the ultrathin film and polymer coating field, which could be considered as surface science, we focus here on the two spectroscopies with the longest history and the most impressive results to date – that is, PAES and LEPD.

A recent example of the power of PAES, illustrating the advantages outlined in Section 5.3.5.8, Positron Surface Spectroscopies, involved the study of Se passivation layers on the Si(001) surface by time-of-flight PAES (Zhu *et al.*, 2005). The Se monolayer was found to be stable after days of air exposure, the physisorbed oxygen on the passivated surface being desorbed below 400 °C. The Se passivation layer desorbs from the Si(001) surface above 800 °C in ultrahigh vacuum (Figure 5.19).

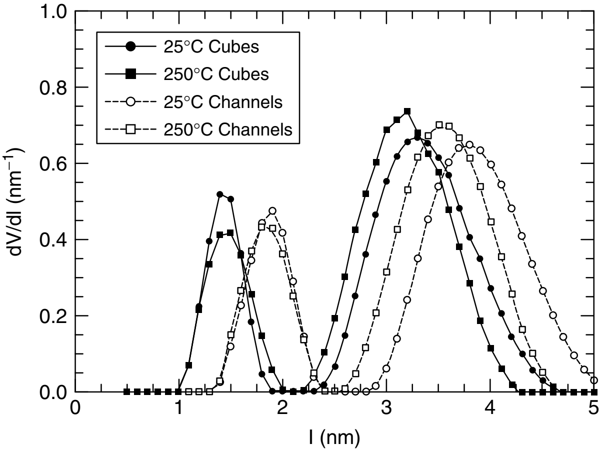

Fig. 5.18 Void size distributions derived from Ps lifetime measurements in low-dielectric thin films of methylsilsequioxane. The solid and broken lines refer to cube- or channel-shaped pores (Gidley *et al.*, 2000). Reproduced by permission, AIP.

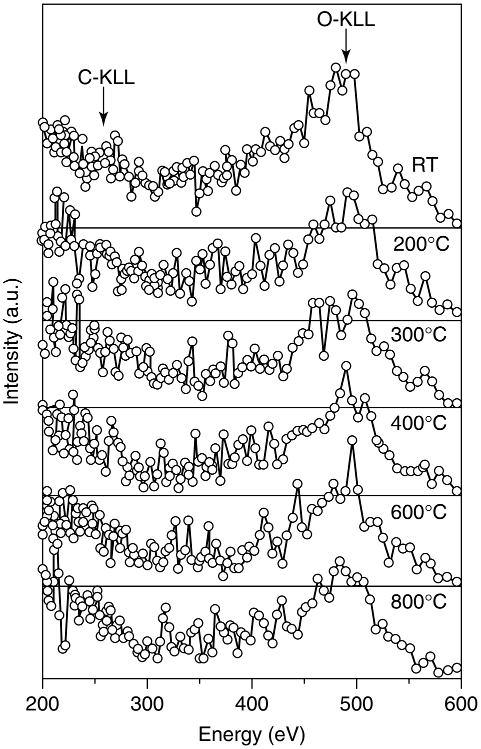

Fig. 5.19 PAES spectra for an Se-passivated Si(001) after isochronal annealing at increasing temperatures (Zhu *et al.*, 2005). Reproduced by permission, AIP.

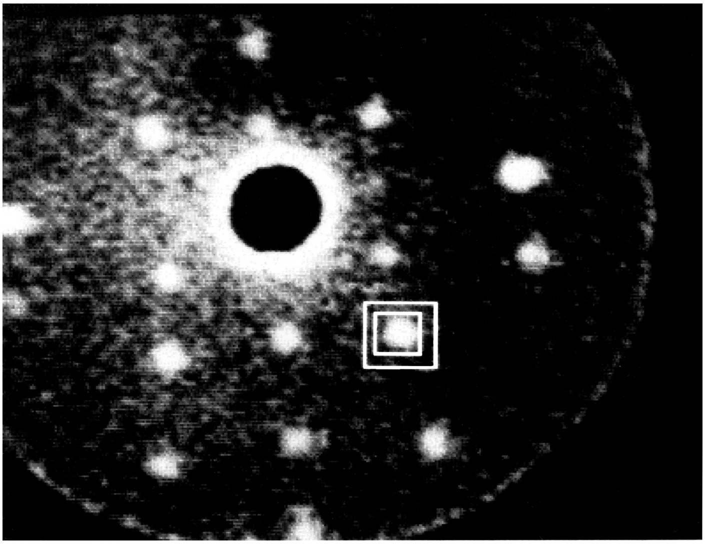

Fig. 5.20 Digital LEPD spot pattern from GaAs(110). The boxes indicate the regions used to evaluate signal intensity (Chen *et al.*, 1993). Reproduced by permission, APS: http://link.aps.org/abstract/PRB/v48/p2400.

In the clearest demonstration of the advantages of LEPD over LEED for the evaluation of the surface structure of some materials – specifically compound semiconductors – Chen *et al.* (1993) showed that the multiple-scattering theory fits LEPD data better than LEED and hence yields structural parameters with smaller uncertainties, and suggested that the differences between LEPD and LEED results were real (see Figure 5.20).

5.4.8
Positron Microscopy

The Brandeis positron reemission microscope used brightness enhancement to obtain a bright microbeam of ~5-μm diameter full width half maximum and $\sim 10^6$ positrons per second. The microscope was able to image defect structures on the surface of a thin Ni(100) film with a spatial resolution of 300 ± 10 nm (Brandes, Canter and Mills, 1988) (Figure 5.21). Improved resolution should be possible with the application of intense positron beams. More recently, there has been work aimed at further exploitation of positron reemission microscopy in Japan, and in Munich an approximately micrometer-pulsed beam (David *et al.*, 2001) was rastered across a surface exhibiting a fatigue crack on the micron scale, creating the two-dimensional positron lifetime image shown in Figure 5.22, which showed the presence of vacancy defects around the crack invisible to standard microscopic methods (Egger *et al.*, 2002).

5.4.9
Positron and Positronium Chemistry

Some current areas of activity in positron and Ps chemistry were mentioned earlier

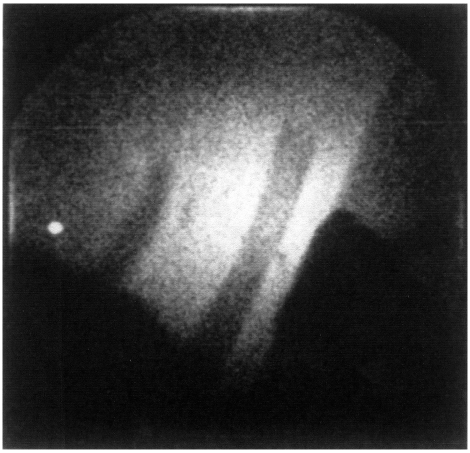

Fig. 5.21 Positron reemission microscope image of Ni foil with contrast due to positron trapping at defects. Magnification 1150. Data collection time 14 hours, white areas have ~40 counts/pixel. From Brandes, Canter and Mills, 1988. Reproduced by permission, APS: http://link.aps.org/abstract/PRL/v61/p492.

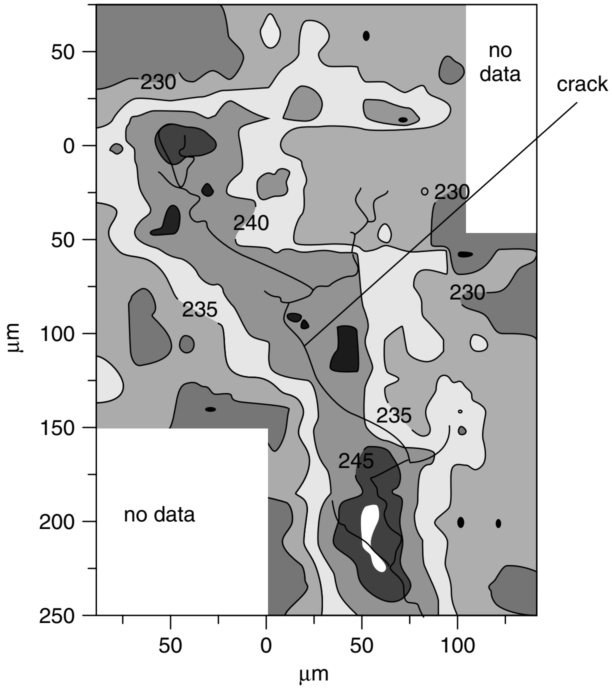

Fig. 5.22 Positron lifetime image of a fatigue crack in copper taken by the Munich scanning positron microscope; lifetimes range from 200 to 250 picoseconds. Incident positron energy = 5 keV (David *et al.*, 2001). Reproduced by permission, Elsevier.

in Section 5.4.5. The formation of Ps in liquids, its enhancement and inhibition, electron and positron scavenging reactions, the formation of positron and Ps bound states, Ps trapping in "bubbles" in liquids, Ps oxidation and spin conversion, have been studied experimentally using PALS, DBS, and ACAR spectroscopies as well as AMOC (Section 5.3.4.2). Rate constants for Ps reactions have been deduced; because Ps is a light particle its diffusion time has to be taken into account. Further details on the chemical reactions of positrons and Ps can be found in the book edited by Jean *et al.* (see the section Further Reading).

Acknowledgments

The author is grateful for the support of the Engineering and Physical Sciences Research Council, UK, over many years, and currently under grant EP/F029829/1. He also acknowledges the guidance he obtained from the earlier contribution to this Encyclopedia of Csaba Szeles and Kelvin Lynn.

Glossary

Angular correlation: The angular distribution of two annihilation γ-photons about 180°, directly related to the momentum distribution of the annihilating pair (and hence, in condensed matter, essentially the electron momentum density).

Annihilation: The decay of a positron–electron pair, with the emission of energy in the form of γ-radiation; for free particles, two photons are most commonly emitted.

Doppler broadening: The broadening of the annihilation γ-line due to the nonzero momentum of an annihilating positron–electron pair.

Lifetime: The mean life of a positron in a material.

Positron: The antiparticle of the electron.

Positronium (Ps): The quasi-stable positron–electron bound state, existing as *ortho*-Ps (spin 1) or *para*-Ps (spin 0), decaying naturally with the emission principally of three or two photons, respectively.

Specific trapping coefficient: The probability of positron trapping by a particular type of defect per second per defect site. The product of this coefficient and the defect concentration (per atom) gives the trapping rate.

Trapping: The localization of positrons (and sometimes Ps) in defects sites; vacancy-type defects are deep traps and negatively charged impurities are shallow traps.

References

Aers, G.C., Marshall, P.A., Leung, T.C., and Goldberg, R.D. (1995) *Appl. Surf. Sci.*, **85**, 196–209.

Anderson, C.D. (1932) *Science*, **76**, 238–239.

Baker, J.A., Chilton, N.B., Jensen, K.O., Walker, A.B., and Coleman, P.G. (1991) *Appl. Phys. Lett.*, **59**, 2962–2964.

Bamford, D., Dlubek, G., Reiche, A., Alam, M.A., Meyer, W., Galvosas, P., and Rittig, F. (2001) *J. Chem. Phys.*, **115**, 7260–7270.

Bauer, W., Maier, K., Major, J., Schaefer, H.-E., Seeger, A., Carstanjen, H.-D., Decker, W., Diehl, J., and Stoll, H. (1987) *Appl. Phys. A*, **43**, 261–267.

Behringer, R. and Montgomery, C.G. (1942) *Phys. Rev.*, **61**, 222–224.

Bell, R.E. and Graham, R.L. (1953) *Phys. Rev.*, **90**, 644–654.

Berko, S. and Plaskett, J.S. (1958) *Phys. Rev.*, **112**, 1877–1887.

Brandes, G.R., Canter, K.F., and Mills, A.P. (1988) *Phys. Rev. Lett.*, **61**, 492–495.

Brandt, W., Berko, S., and Walker, W.W. (1960) *Phys. Rev.*, **112**, 1289–1295.

Canter, K.F., Coleman, P.G., Griffith, T.C., and Heyland, G.R. (1972) *J. Phys. B*, **5**, L167–L169.

Canter, K.F., Dharmavaram, V., Smirnov, A.G., Wesley, S.A., Wong, K.A., Xie, R., Brandes, G.R., and Mills, A.P. Jr (1994) *Slow Positron Techniques for Solids and Surfaces* (eds E. Ottewitte and A.H. Weiss), *Conference Series 303*, AIP, New York, p. 385.

Cassidy, D.B. and Mills, A.P. Jr (2007) *Nature*, **449**, 195–197.

Chang, T.-B., Tang, X.-W., and Li, Y.-Q. (1985) *Phys. Lett.*, **B157**, 357–360.

Charlton, M and Humberston, J.W. (2001) *Positron Physics*, Cambridge University Press, Cambridge.

Chen, D.M., Berko, S., Canter, K.F., Lynn, K.G., Mills, A.P. Jr, Roellig, L.O., and Sferlazzo, P. (1987) *Phys. Rev. Lett.*, **58**, 921–924.

Chen, X.M., Canter, K.F., Duke, C.B., Paton, A., Lessor, D.L., and Ford, W.K. (1993) *Phys. Rev. B*, **48**, 2400–2411.

Coleman, P.G., Albrecht, L., Walker, A.B., and Jensen, K.O. (1992) *J. Phys. Condens. Matter*, **4**, 10311–10322.

Coleman, P.G., Burrows, C.P., and Knights, A.P. (2002) *Appl. Phys. Lett.*, **80**, 947–949.

Coleman, P.G., Burrows, C.P., Mahapatra, R., and Wright, N.G. (2007) *J. Appl. Phys.*, **102**, 014106–0141-4.

David, A., Kögel, G., Sperr, P., and Triftshäuser, W. (2001) *Phys. Rev. Lett.*, **87**, 067402–067411.

DeBenedetti, S., Cowan, C.E., and Konneker, W.R. (1949) *Phys. Rev.*, **76**, 440.

DeBenedetti, S., Cowan, C.E., Konneker, W.R., and Primakoff, H. (1950) *Phys. Rev.*, **77**, 205–212.

Deutsch, M. (1951) *Phys. Rev.*, **82**, 455–456.

Dirac, P.A.M. (1930) *Proc. Cambridge Philos. Soc.*, **26**, 361–375.

Dugdale, S.B., Alam, M.A., Wilkinson, I., Hughes, R.J., Fisher, I.R., Canfield, P.C., Jarlborg, T., and Santi, G. (1999) *Phys. Rev. Lett.*, **83**, 4824–4827.

Egger, W., Kögel, G., Sperr, P., Triftshäuser, W., Rödling, S., Bär, J., and Gudladt, H.-J. (2002) *Appl. Surf. Sci.*, **194**, 214–217.

Foster, P.J., Mascher, P., Knights, A.P., and Coleman, P.G. (2007) *J. Appl. Phys.*, **101**, 043702–043703.

Gidley, D.W., Köymen, A.R., and Capehart, T.W. (1982) *Phys. Rev. Lett.*, **49**, 1779–1782.

Gidley, D.W. (1989) *Phys. Rev. Lett.*, **62**, 811–814.

Gidley, D.W., Frieze, W.E., Dull, T.L., Sun, J., Yee, A.F., Nguyen, C.V., and Yoon, D.Y. (2000) *Appl. Phys. Lett.*, **76**, 1282–1284.

Gilbert, S.J., Kurz, C., Greaves, R.G., and Surko, C.M. (1997) *Appl. Phys. Lett.*, **70**, 1944–1946.

Jeavons, A.P., Townsend, D.W., Ford, N.L., Kull, K., Manuel A.A., Fischer O., and Peter, M. (1978) *IEEE Trans. Nucl. Sci.*, **25**, 164–169.

Joliot-Curie, I. and Joliot-Curie, F. (1934) *Nature*, **133**, 201–202.

Kahana, S. (1963) *Phys. Rev.*, **129**, 1622–1628.

Kawasuso, A. and Okada, S. (1998) *Phys. Rev. Lett.*, **81**, 2695–2698.

Landes, H.S., Berko, S., and Zuchelli, A.J. (1956) *Phys. Rev.*, **103**, 828–829.

Lock, D.G., Crisp, V.H.C., and West, R.N. (1973) *J. Physics F*, **3**, 561–570.

Lynn, K.G., Mills, A.P. Jr, West, R.N., Berko, S., Canter, K.F., and Roellig, L.O. (1985) *Phys. Rev. Lett.*, **54**, 1702–1705.

Lynn, K.G., MacDonald, J.R., Boie, R.A., Feldman, L.C., Gabbe, J.D., Robbins, M.F., Bonderup, E., and Golovchenko, J. (1977) *Phys. Rev. Lett.*, **38**, 241–244.

Mader, J., Berko, S., Krakauer, H., and Bansil, A. (1976) *Phys. Rev. Lett.*, **37**, 1232–1236.

Maier, K., Peo, M., Saile, B., Scaefer, H.E., and Seeger, A. (1979) *Phil. Mag. A*, **40**, 701–728.

Mantl, S. and Trifthäuser, W. (1978) *Phys. Rev. B*, **17**, 1645–1652.

Mills, A.P. Jr (1980) *Appl. Phys.*, **22**, 273–276.

Mills, A.P., Platzman, P.M., and Brown, B.L. Jr (1978) *Phys. Rev. Lett.*, **41**, 1076–1079.

Mogensen, O.E. (1974) *J. Chem. Phys.*, **60**, 998–1004.

Mohorovičić, S. (1934) *Astron. Nachr.*, **14**, 93–108.

Mostafa, K.M., Caenegem, N.V., De Baerdemaeker, J., Segers, D., and Houbaert, Y. (2007) *Phys. Stat. Sol.*, **4**, 3554–3558.

Nagai, Y., Hasegawa, H., Tang, Z., Hempel, A., Yubuta, K., Shimamura, T., Kawazoe, Y, Kawai, A., and Kano, F. (2000) *Phys. Rev. B.*, **61**, 6574–6578.

Ociepa, J.G., Schultz, P.J., Griffiths, K., and Norton, P.R. (1990) *Surf. Sci.*, **225**, 281–291.

Ohdaira, T., Suzuki, R., Mikado, T., Ohgaki, H., Chiwaki, M., and Yamazaki, T. (1997) *Appl. Surf. Sci.*, **116**, 177–180.

Ore, A. (1949) *Univ. Bergen Aarb. Naturvit. Rekke*, **9**.

Ore, A. and Powell, J.L. (1949) *Phys. Rev.*, **75**, 1696–1699.

Osmon, P.E. (1965) *Phys. Rev.*, **138**, B216–B218.

Page, L.A. and Heinberg, M. (1956) *Phys. Rev.*, **102**, 1545–1553.

Petersen, K., Thrane, N., and Cotterill, R.M.J. (1974) *Phil. Mag.*, **29**, 9–23.

Roach, T., Bakshi, A., and Canter, K.F. (1995) *Meas. Sci. Technol.*, **6**, 496–501.

Rosenberg, I.J., Weiss, A.H., and Canter, K.F. (1980) *Phys. Rev. Lett.*, **44**, 1139–1142.

Ruark, A.E. (1945) *Phys. Rev.*, **68**, 278.

Saarinen, K., Laine, T., Kuisma, S., Nissila, J., Hautojarvi, P., Dobrzynski, L., Baranowski, J., Pakula, K., Stepniewski, R., Wojdak, M., Wysmolek, A., Suski, T., Leszczynski, M., Grzegory, I., and Porowski, S. (1997) *Phys. Rev. Lett.*, **79**, 3030–3033.

Stewart, A.T. (1957) *Can. J. Phys.*, **35**, 168–183.

Shearer, J.W. and Deutsch, M. (1949) Minutes of the semi-centennial meeting at Cambridge, *Phys. Rev.*, **76**, 462.

Shishido, I., Yasueda, H., Mizuno, M., Araki, H., and Shirai, Y. (2007) *Phys. Stat. Sol. C*, **4**, 3563–3566.

Siegle, A., Stoll, H., Castellaz, P., Major, J., Schneider, H., and Seeger, A. (1997) *Appl. Surf. Sci.*, **116**, 140–144.

Stepanov, S.V. and Byakov, V.M. (2002) *J. Chem. Phys.*, **116**, 6178–6195.

Triftshäuser, W. and Kögel, G. (1982) *Phys. Rev. Lett.*, **48**, 1741–1744.

Van Veen, A., Schut, H., de Vries, J., Hakvoort, R.A., and IJpma, M.R. (1990) *AIP Conference Proceedings*, AIP, New York, Vol. 218, 171–196.

Weber, M.H., Tang, S., Berko, S., Brown, B.L., Canter, K.F., Lynn, K.G., Mills, A.P., Roellig, L.O., and Viescas, A.J. (1988) *Phys. Rev. Lett.*, **61**, 2542–2545.

Weber, M.H., Hunt, A.W., Golovchenko, J.A., and Lynn, K.G. (1999) *Phys. Rev. Lett.*, **83**, 4658–4661.

Weisberg, H. and Berko, S. (1967) *Phys. Rev.*, **154**, 249–257.

Weiss, A.H., Mayer, R., Jibaly, M., Lei, C., Mehl, D., and Lynn, K.G. (1988) *Phys. Rev. Lett.*, **61**, 2245–2248.

West, R.N., Mayers, J., and Walters, P.A. (1981) *J. Phys. E*, **14**, 478–488.

Wheeler, J.A. (1946) *Ann. New York Acad. Sci.*, **48**, 219.

Zhu, J.G., Nadesalingam, M.P., Weiss, A.H., and Tao, M. (2005) *J. Appl. Phys.*, **97**, 103510.

Zitzewitz, P.W., Van House, J.C., Rich, A., and Gidley, D.W. (1979) *Phys. Rev. Lett.*, **43**, 1281–1284.

Further Reading

Coleman, P.G. (ed) (2000) *Positron Beams and their Applications*, World Scientific, Singapore.

Dupasquier, A. and Mills, A.P. Jr (eds) (1994) *Positron Spectroscopy of Solids, Proceedings of the International School of Physics "Enrico Fermi", Course LXXV*, North Holland, Amsterdam.

Gidley, D.W., Peng, H.-G., and Vallery, R.S. (2006) *Ann. Rev. Mater. Res.* **36**, 49–79.

Hirose, M., Nakajyo, T., and Washio, M. (1997) *Appl. Surf. Sci.*, **116**, 63–67.

Jean, Y.C., Mallon, P.E., and Schrader, D.M. (eds) (2003) *Positron and Positronium Chemistry*, World Scientific, Singapore.

Knights, A.P., Mascher, P., and Simpson, P.J. (2007) Proceedings of the 14th International Conference on Positron Annihilation. *Phys. Stat. Sol.*, **4**, 3413–4040.

Krause-Rehberg, R. and Leipner, H.S. (1999) *Positron Annihilation in Semiconductors: Defect Studies*, Springer, Berlin.

Mackenzie, I.K., Schulte, C.W., Jackman, T., and Campbell, J.L. (1973) *Phys. Rev. A*, **7**, 135–145.

Mills, A.P. Jr, Crane, W.S., and Canter, K.F. (eds) (1986) *Positron Studies of Solids, Surfaces and Atoms*, World Scientific, Singapore.

Pethrick, R.A. (1997) *Prog. Polymer Sci.* **22**, 1–47.

Surko, C.M. and Greaves, R.G. (2004) *Phys. Plasmas* **11**, 2333.nhi

Weiss, A (1995) in *The Handbook of Surface Imaging and Visualisation* (ed. A.T. Hubbard), CRC Press.

POWER ELECTRONICS

Bimal K. Bose, *Department of Electrical Engineering, The University of Tennessee, Knoxville, Tennessee, U.S.A.*

	Introduction	1
1.	**Power Semiconductor Devices**	2
1.1	Diode	2
1.2	Thyristor	2
1.3	Triac	4
1.4	Gate Turnoff Thyristor (GTO)	4
1.5	Power Transistor	5
1.6	Power MOSFET	5
1.7	Insulated-Gate Bipolar Transistor (IGBT)	7
1.8	Static Induction Transistor (SIT)	7
1.9	MOS-Controlled Thyristor (MCT)	8
2.	**Power Electronic Converters**	8
2.1	dc-to-dc Converter	9
2.2	Phase-Controlled Rectifier	10
2.3	Cycloconverter	12
2.4	ac Controller	12
2.5	Voltage-Fed Inverter	13
2.6	Current-Fed Inverter	15
3.	**Selected Applications of Power Electronics**	16
3.1	Incandescent-Lamp Dimming	16
3.2	High-Frequency Ballast for Fluorescent Lamps	17
3.3	Switching-Mode Power Supply	17
3.4	Induction Heating	18
3.5	Uninterruptible Power Supply (UPS)	19
3.6	dc-Motor Drive	20
3.7	ac-Motor Drive	21
3.7.1	Open-Loop Volts/Hz Control	21
3.7.2	Vector Control	21
	Glossary	22
	Works Cited	23
	Further Reading	23

INTRODUCTION

Power electronics, as the name indicates, is electronics that deals with high power, and the power level may vary from a few watts to a number of gigawatts. Basically, the subject deals with conversion and control of electrical power and its applications. The core of a power electronics apparatus consists of a converter built with a matrix of power semiconductor switching devices that work under the guidance of control electronics. The converters can be classified as rectifier (ac-to-dc converter), inverter (dc-to-ac converter), chopper (dc-to-dc converter), ac-power controller (at same frequency), and cycloconverter (direct frequency converter), which will be discussed later. The motivation for using switching devices in a converter is to increase converter efficiency because an ideal switching device has no loss. A converter can be looked upon as a power amplifier. However, compared to a class A or class B linear amplifier where the efficiency is low because of large device dissipation, the switching-mode amplifier efficiency is high. One drawback for switching-mode operation of a converter is that it introduces harmonics in the source and load sides.

Recently, the application of power electronics is growing extensively in industrial, commercial, residential, aerospace, and military environments for applications such as heating and lighting control, electrochemical processes, ac- and dc-power supplies, electronic welding, and dc- and ac-motor drives. Some of these applications will be reviewed later. Power electronics is helping industrial productivity and quality improvement of products. Another aspect of power electronics application is saving of electrical energy, which indirectly helps to solve the environmental pollution problem. In this article, the power semiconductor devices are reviewed

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

first. Then, topologies of converters for different types of conversion are reviewed. Finally, a few selected applications are discussed. The references cited at the end will help the reader to broaden his horizon in power electronics and its applications.

1. POWER SEMICONDUCTOR DEVICES

The power semiconductor device (Bose, 1992) is the heart and muscle of modern power electronics. In fact, the evolution of power electronics in the present state has been possible because of the advent of so many types of advanced power semiconductor devices. It may be of interest to mention here that prior to the modern power electronics age, gas tubes, such as thyratrons, phanotrons, and pool-cathode mercury-arc devices, were used in power electronics. Although basically the power semiconductor device is an on-off switch, its practical behavior is far more complex in switching and conduction states. A power electronics engineer needs to understand the devices thoroughly for designing efficient, reliable, and cost-effective equipment. In this section, the operating principles and characteristics of important devices are reviewed.

1.1 Diode

The power diode, as shown in Fig. 1, is the simplest of all power semiconductor devices. Basically, it has a *p-n* junction, which, if forward biased, conducts freely, but blocks conduction with reverse polarity of voltage. A power diode has normally a near-intrinsic layer between the *p* and *n* layers to increase the voltage-blocking capability. High-power line-frequency *p-i-n* diode rectifiers are somewhat sluggish in recovery characteristic. They are available in ratings such as 5000 V, 2000 A. The applications of these diodes include battery charging, power supply, electroplating, traction, welding, and uninterruptible power supply systems (UPS). Fast-recovery diodes are used in high-frequency rectification and switching applications where the reverse recovery current and recovery time are required to be small. Gold or platinum diffusion is used to reduce the carrier lifetime but gives the adverse effect of larger conduction drop. Schottky diodes are formed by metal–semiconductor junctions and are defined as majority-carrier devices because the semiconductor does not store any minority carrier. These diodes are characterized by low conduction drop, very fast switching time, large reverse leakage current, and low voltage-blocking capability. The application of Schottky diodes includes high-frequency switching-mode power supplies and instrumentation. Zener diodes are specially made to operate with reverse breakdown so as to provide a reference voltage. The electrical and thermal characteristics of power diodes are somewhat similar to those of thyristors, and these are reviewed in the next section.

1.2 Thyristor

The thyristor, or silicon-controlled rectifier (SCR), is the main workhorse in today's power converters. The modern era of power

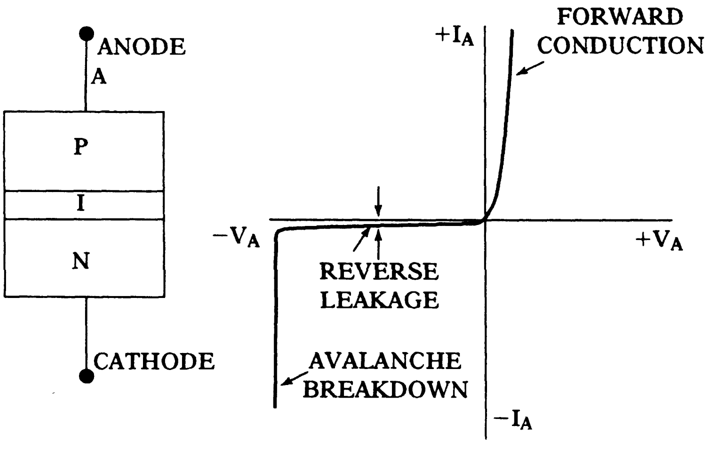

FIG. 1. Diode structure and characteristics.

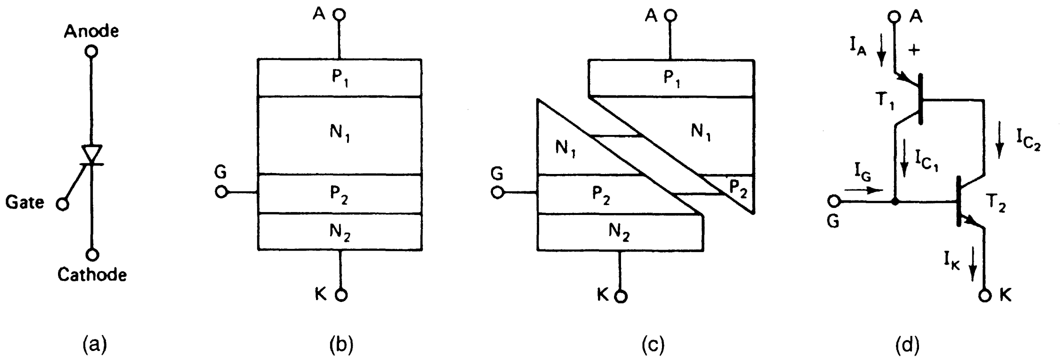

FIG. 2. Thyristor structure and two-transistor analogy.

electronics was born with the invention of the thyristor in the late 1950s. Figure 2 shows the thyristor structure, and Fig. 3 gives its volt-ampere characteristics. Basically, it is a trigger-into-conduction device that can be turned on by a positive gate-current pulse. However, once the device is on, a negative gate pulse cannot turn the device off. Commutation or turnoff is possible by applying temporarily a reverse voltage across the device. A thyristor has basically a *p-n-p-n* structure, and its triggering action can be explained by the regenerative feedback configuration of its constituent *pnp* and *npn* transistors, which give increasing current gain at higher emitter currents. A thyristor can also turn on by excessive anode voltage or its rate of rise (*dv/dt*), by a rise in temperature, or by light shining on the junctions. The device turns on rapidly, and the time is limited by the rate of rise in anode current (*di/dt*). The turnoff process is slow because the minority carriers are to be cleared from the inner junctions by "recovery" and "recombination" processes. Like other power semiconductor devices, a thyristor is connected with snubber elements for protection. A turnon snubber (inductance in series with the device) limits anode *di/dt* of the device to limit the turnon switching loss. A turnoff snubber can be unpolarized (resistance–capacitance series circuit across the device) or polarized (resistance–capacitance series circuit across the device with a bypass diode across the resistor), and it protects the device from transient voltage, reduces device *dv/dt* to prevent switch-on, and limits switching loss within the device. However, the total switching loss (including snubber loss) can be increased. A thyristor can tolerate large fault-current transiently and can be protected by a fast fuse. Within a specified junction-temperature limit (125 °C), the device can carry large current with a small duty cycle, and the current can be increased further by improved cooling. The thermal capacitance of the device and the heatsink improves the current loading within the constraint of junction temperature.

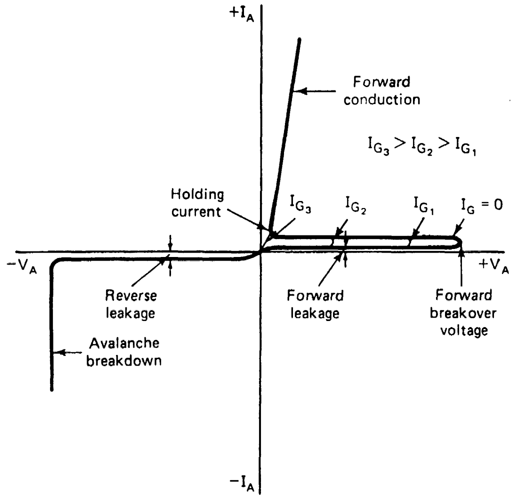

FIG. 3. Volt-ampere characteristics of thyristor.

Commercial thyristors can be classified into phase-control and inverter types. Phase-control thyristors operate on an ac line and use the line commutation (commutation by a segment of inverse line voltage) principle. Modern phase-control light-triggered thyristors are available with ratings up to 6000 V, 3500 A. While a phase-control thyristor can block voltage in either polarity, an inverter thyristor blocks voltage in the forward direction only and has faster turnoff time. Thyristors are widely used in dc- and ac-motor drives, power supplies in electrochemical processes, lighting and heating control, welding control, high-voltage dc (HVDC) conversion, and solid-state circuit breakers.

1.3 Triac

A triac is basically an integration of a pair of phase-controlled thyristors connected in inverse parallel on the same chip. Figure 4 shows the triac circuit symbol and its volt-ampere characteristics. The three-terminal device can be triggered into conduction in both positive and negative polarity of supply voltage. When the supply is positive, i.e., T_2 is positive, a positive gate pulse turns on the device. On the other hand, if T_1 is positive, a negative gate-current pulse switches on the device. A triac is less expensive than a pair of inverse-parallel thyristors, and the gate-control circuit is somewhat simpler. However, there are a few disadvantages because of complex integration of two devices on a chip. The gate-current sensitivity of a triac is poor, and turnoff time is longer because of storage-charge effect. The reapplied *dv/dt* of a triac is low on account of storage charge, and it becomes difficult to supply an inductive load by a triac converter. Triacs are normally used in 50/60-Hz power lines for incandescent lamp dimming, heating control, appliance-type low-power motor drives, and solid-state ac relays. The devices are commercially available with ratings up to 800 V, 40 A.

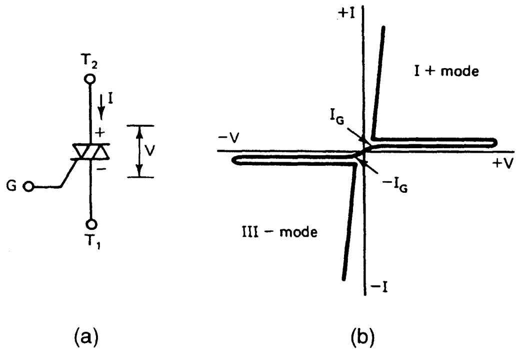

(a) (b)

FIG. 4. Triac symbol and volt-ampere characteristics.

1.4 Gate-Turnoff Thyristor (GTO)

A gate-turnoff thyristor (GTO), as the name indicates, is basically a thyristorlike device that can be turned on by a positive gate-current pulse but, in addition, has the capability of being turned off by a negative current pulse. It has thyristorlike *p-n-p-n* structure, but permits diversion of *pnp* collector current by the gate and thus breaking of the *pnp-npn* regenerative feedback effect. The GTO is characterized by poor turnoff current gain (typically 4 or 5), i.e., a 4500-V, 3000-A state-of-the-art device needs a gate drive of −750 A for turnoff. However, because of the low duty cycle of the gate-drive current, the power rating of the gate drive is not large. GTOs are available with symmetric (both directions) and asymmetric (one direction) voltage-blocking capabilities. Voltage-fed converters use asymmetric devices whereas current-fed converters use symmetric devices. These will be discussed in Sec. 2. The turnon behavior of the GTO is similar to that of a thyristor, but its turnoff characteristics are somewhat complex and may cause second-breakdown failure. The GTO is a slow-turnoff switching device. This, in addition to the long tail current, aggravates the switching loss. Therefore, a large-size well-designed snubber is essential for GTO operation. As indicated before, a snubber diverts the switching loss from the device to the snubber circuit. Because of large switching loss, the switching frequency is limited to 1–2 kHz. For a multimegawatt GTO converter, the switching frequency may be only 500 Hz. Recently, regenerative snubbers are being suggested to improve the converter efficiency. In a regenerative snubber, the stored energy in the snubber capacitor, instead of dissipating in the snubber resistor, is pumped back to the source with the help of additional circuit elements. In spite of several disadvantages, GTOs have practically replaced the force-commutated inverter thyristors in voltage-fed inverter applications because of overall advantages of reduced size and cost and improved efficiency. Commercial GTOs are available in high power ratings, and are used in dc- and ac-machine drives, UPS systems, static reactive power compensators, induction heating, and photovoltaic and fuel-cell converters.

1.5 Power Transistor

A power transistor, also known as power bipolar junction transistor (BJT) or bipolar power transistor (BPT), is a two-junction self-controlled device where the collector current is under the control of the base-drive current. Basically, it is a linear device but is operated in switching mode to improve the converter efficiency. The fault overcurrent in the device can be suppressed by inhibiting the base drive. It has asymmetric voltage-blocking capability, and application is therefore limited to voltage-fed inverters and choppers.

The current gain of a power transistor is low but can be increased to a high value in a Darlington transistor as shown in Fig. 5. Basically, it is a cascaded connection of two *npn* transistors. The bypass diode fabricated on the monolithic chip permits reverse current to flow freely. The shunt resistances R_1 and R_2 reduce the collector leakage current and help to establish bias voltages across base–emitter junctions. The intermediate base lead B_2 can be used for fast turnoff of the T_2 transistor. The base-drive power requirement is lower than that of a single transistor, but it is at the cost of slightly reduced switching frequency and higher conduction drop.

An important property of the transistor is that its current gain varies with collector current and junction temperature. The normal dc-current rating of the device can be increased with a low–duty-cycle basis within the limits of junction temperature and wire-bond melting. Double or triple Darlington power transistors are available in module form with matched parallel devices for ratings up to 1200 V and 800 A, and converters up to several hundred kVA can be built with transistors. The device turnon and turnoff times are somewhat low, and the switching frequency can be several kHz. A polarized snubber is usually used with a transistor. The switching frequency of a device is essentially dictated by the switching loss, and at higher frequency, the losses will be higher, demanding more efficient cooling.

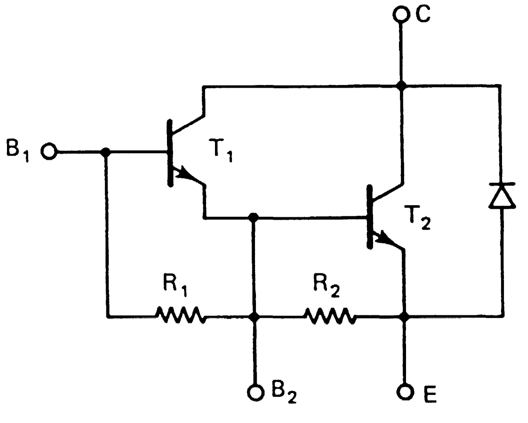

FIG. 5. Darlington transistor with bypass diode.

Power transistors are susceptible to second-breakdown failures. When the collector current is switched on, it tends to crowd on the base-emitter junction periphery (emitter crowding effect), thus constricting the collector current in a narrow area of the reverse-biased collector junction. This tends to create a hot spot, and the junction fails by thermal runaway, which is known as second breakdown. A similar problem arises when an inductive load is switched off. As the base-emitter junction becomes reverse-biased, the collector current tends to concentrate in a narrow area of the collector junction. The manufacturers provide specifications in the form of safe operating areas (SOAs) during turnon (FBSOA) and turnoff (RBSOA), as shown in Fig. 6. The load lines of the device on the volt-ampere plane should be controlled by turnon and turnoff snubbers so that FBSOA and RBSOA are not exceeded. Power transistors are popularly used in dc- and ac-motor drives, power supplies, UPS systems, static reactive compensators, induction heating, etc. Very recently, they have been challenged by IGBTs, which will be discussed later.

1.6 Power MOSFET

Unlike the bipolar transistor, a power MOSFET (metal-oxide-semiconductor field-effect transistor) is a unipolar, majority-carrier, "zero junction" voltage-controlled device. Figure 7(a) shows the circuit symbol of an *n*-channel power MOSFET. If the gate voltage is positive and beyond a threshold value, an *n*-type conducting channel will be induced that will permit current flow between the drain and the source if the drain has positive supply voltage. The gate impe-

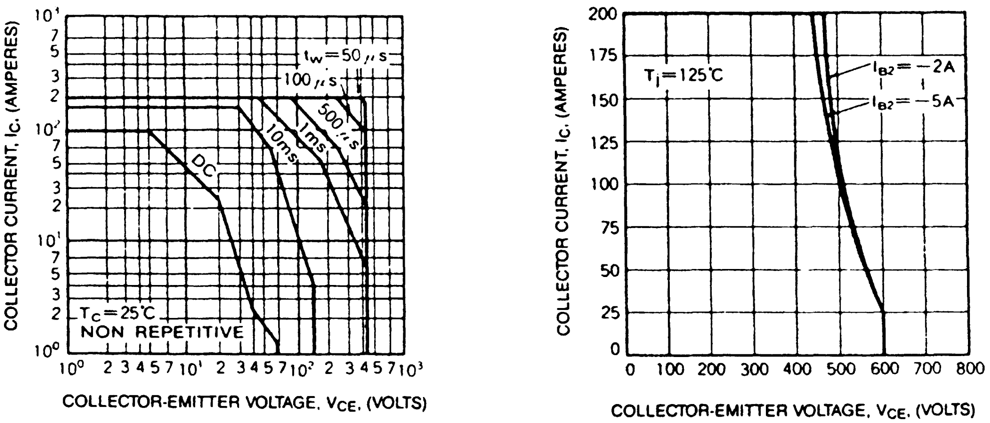

FIG. 6. Safe operating areas (FBSOA and RBSOA) of the Darlington transistor.

dance is extremely high at steady state, but the small gate-source capacitance will demand a pulse current during fast turnon and turnoff. The device is basically linear with asymmetrical blocking capability, and has an integral body diode that can carry full current in reverse direction. The diode is characterized by slow recovery and is often bypassed by external fast-recovery diodes in high-frequency applications. Being a majority-carrier device, the switching speed (turnon and turnoff times) is extremely fast, and therefore, it needs a very small-size snubber. Power MOSFETs have been used in converters with hundreds of kilohertz switching frequency. However, the device has the disadvantage of higher conduction drop, particularly with higher voltage rating. The "on" resistance of the device has a positive temperature coefficient, and therefore permits

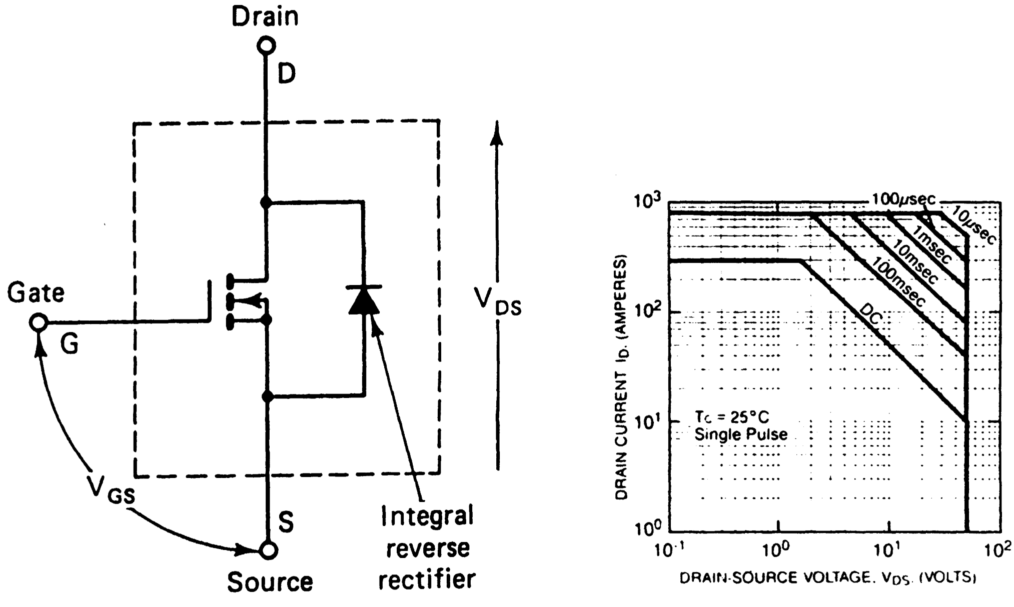
Drain
D
Gate
G
VGS
S
Source
VDS
Integral reverse rectifier
DRAIN CURRENT ID, (AMPERES)
100μsec
10μsec
1msec
10msec
100msec
DC
Tc = 25°C
Single Pulse
DRAIN-SOURCE VOLTAGE, VDS (VOLTS)

(a) (b)

FIG. 7. Power MOSFET symbol and safe operating areas.

easy paralleling of large number of devices. The second-breakdown effect of the device is negligible. If localized heating occurs for any reason, an increase of resistance forces the current distribution to be uniform. The peak current of the device can, therefore, be increased to high value on a short–duty-cycle basis, as shown in Fig. 7(b). Power MOSFETs are very popular in low-voltage, low-power, and high-frequency switching applications, such as switching-mode power supplies, brushless dc-motor drives, and solid-state relays. State-of-the-art modules are available with ratings up to 500 V, 140 A.

1.7 Insulated-Gate Bipolar Transistor (IGBT)

Recently, the IGBT is emerging as a very important device in power electronics for converter ratings up to several hundred kilowatts. Basically, it is a hybrid MOS-gated turn on/off bipolar transistor that combines the attributes of MOSFET, BJT, and thyristor. Figure 8 shows the IGBT structure and its circuit symbol. The device architecture is similar to that of the MOSFET except that the n^+ layer at the drain is replaced by a p^+ layer at the collector, as shown. The device has the high input impedance of a power MOSFET but BJT-like conduction characteristics. If the gate is positive with respect to the emitter, an *n* channel is induced in the *p* region. This forward biases the base-emitter junction of the *pnp* transistor, turning it on and causing conductivity modulation of the *n* region, causing low conduction drop compared to that of a MOSFET. The device is turned off by zero gate voltage, which removes the conduction channel in the *p* region. The device has a higher current density than the BJT and the power MOSFET, and needs only a fraction of the die size of a MOSFET. The input capacitance is very small compared to that of a MOSFET. Also, it has improved Miller feedback effect (feedback from collector to the gate due to junction capacitances) during turnon and turnoff *dv/dt*. The device does not show any second-breakdown effect, and the RBSOA and FBSOA are thermally limited by junction temperature. IGBT converters can be designed with or without snubbers.

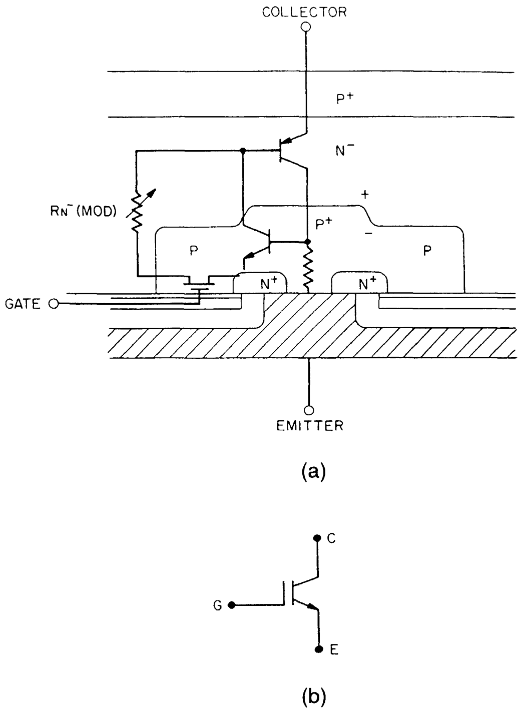

(a)

(b)

FIG. 8. IGBT structure and symbol.

Recently, IGBTs are being popularly used in applications such as dc- and ac-motor drives, UPS systems, power supplies, and drivers for solenoids, relays, and contactors. Although IGBTs are slightly more expensive than BJTs, the lower gate-drive requirement, availability of integrated gate drivers with protection capability, small snubber or snubberless operation, low switching loss, and higher switching frequency make the IGBT apparatus more efficient with less size, cost, and reliability. State-of-the art devices are available with ratings up to 2000 V, 500 A.

1.8 Static Induction Transistor (SIT)

A SIT is a high-power, high-frequency normally on device that has the characteristics of a triode vacuum tube that has been known for a long time. It is a majority-carrier *n*-channel device with gate electrodes buried in the channel. It is nearly identical to the junction FET except that the vertical and buried gate construction gives low channel resistance, causing lower conduction drop. The high-voltage, high-power device has been used in linear mode in audio, VHF/

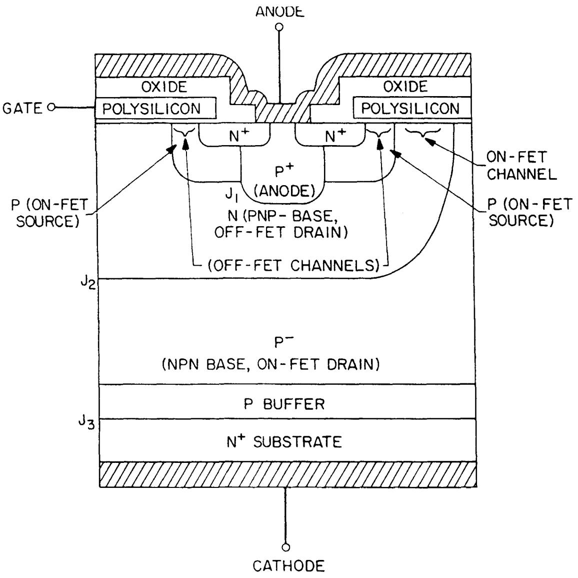

FIG. 9. MCT structure.

UHF, and microwave amplifiers and high-frequency induction heating, and ultrasonic generators in switching mode. The highest available power capability of the device is 1200 V, 300 A.

1.9 MOS-Controlled Thyristor (MCT)

A MCT, as the name indicates, is a thyristorlike trigger-into-conduction device that can be turned on and off by a voltage pulse on the MOS gate. It is more of a GTO-like switching device except that the turnoff current gain is very high. A MCT is a high-power, high-frequency, low-conduction-drop device. In switching frequency, it is somewhat comparable to the IGBT, but it has a lower conduction drop (slightly above a volt). The device has been available commercially (600 V, 75 A or 1000 V, 65 A) since 1992.

A MCT is basically a parallel connection of thousands of identical microcells on the same chip. The individual microcell structure is quite complex and is shown in Fig. 9, and Fig. 10 shows its equivalent circuit and symbol. It is turned on by a negative gate-voltage pulse (with respect to anode) and is turned off by a positive voltage pulse. The device has thyristorlike *p-n-p-n* layers, and the *pnp-npn* regenerative effect turns on the device. The device has asymmetric voltage-blocking capability, and its limited SOA demands a polarized snubber. Unlike the MOSFET, its input capacitance is fixed because of the absence of Miller effect. With the *n*-FET normally on, the device is extremely insensitive to *dv/dt* and junction-temperature triggering effect. The MCTs can be connected in parallel for higher-power equipment. In next-generation power electronics, MCTs will be serious contenders against IGBT's and GTO's.

2. POWER ELECTRONIC CONVERTERS

As mentioned before, the function of a converter is to convert and control electrical power for various applications. The conversion process may be dc to dc, ac to dc, dc to ac, or ac to ac at the same or different frequency. The type of power semiconductor used in a converter depends on the converter topology, the power rating, and the desired performance. This section will review the basic converter types for various power conversion processes (Bose, 1992).

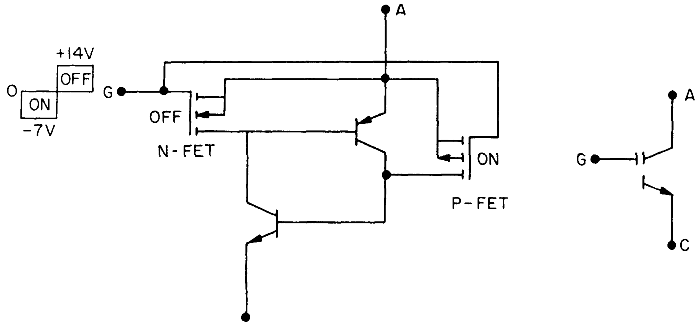

FIG. 10. MCT equivalent circuit and symbol.

2.1 dc-to-dc Converter

This type of converter changes unregulated dc voltage to regulated or variable dc voltage at the output for applications such as dc-motor speed control and switching-mode power supply. The input dc is usually a battery, or can be derived from single-phase or three-phase ac through a rectifier (as will be discussed later) and a capacitance filter. It can also be obtained from photovoltaic or fuel-cell–type dc generators. Figure 11(a) shows a simple pulse-width–modulated (PWM) dc-dc converter (also known as a chopper) for speed control of a dc motor. The power semiconductor device used is an IGBT that receives a duty-cycle–modulated wave from the control circuit. A polarized snubber is shown across the device. When the gate is positive, the IGBT switches on and the supply voltage is impressed across the machine armature. At zero gate drive, the device switches off and the freewheeling diode provides a closed path for the inductive armature current and clamps zero voltage across the armature. Therefore, by varying the duty cycle, the average dc voltage across the armature can be varied to control speed of the motor. This "one-quadrant" buck converter permits power flow from source to load only. To obtain a larger range of duty cycle, the control wave frequency can be decreased, maintaining constant on time or off time. At high speed, the motor counter emf is large, and therefore, machine current can be discontinuous. The same principle can be used in switching-mode power supply where the output dc voltage is regulated to be constant irrespective of supply and load variation. The circuit can use power MOSFETs for low-power, high-frequency application, whereas GTOs can be used for large power.

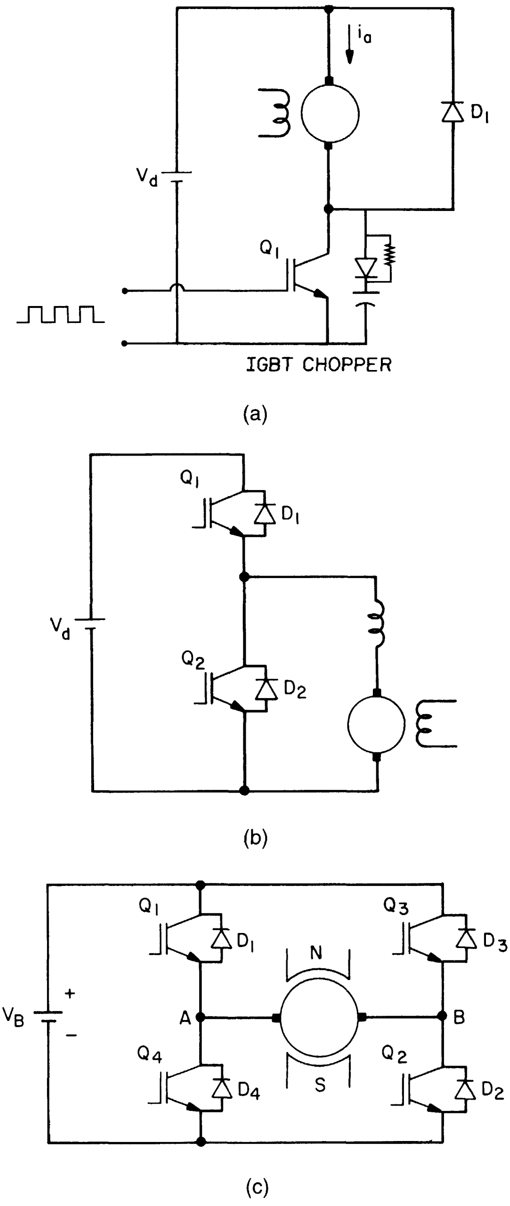

FIG. 11. dc-dc converters for speed control of dc motor: (a) buck chopper, (b) buck-boost chopper, and (c) four-quadrant chopper.

Figure 11(b) shows a "two-quadrant" buck-boost chopper for speed control with regenerative braking effect. The IGBT Q_1 with diode D_2 supplies variable voltage to the machine for motoring mode operation with or without acceleration. The devices Q_2 and D_1 operate to provide regenerative braking torque on the machine. The machine acts as a generator in this mode, and the energy is pumped back to the source. When Q_2 is on, the short-circuited armature builds armature current in the reverse direction. When Q_2 turns off, the inductor stored energy is fed to the source at higher voltage (boost operation).

A "four-quadrant" buck-boost converter for reversible and regenerative speed control using an H bridge is shown in Fig. 11(c). The four quadrants of operation can be summarized as follows:

1. Buck-converter forward motoring operation when Q_1 is on, Q_2 is chopping, and D_3–Q_1 are freewheeling.
2. Boost-converter forward regenerating operation when Q_4 is chopping and D_2–D_1 are freewheeling.
3. Buck-converter reverse motoring operation when Q_3 is on, Q_4 is chopping, and D_1–Q_3 are freewheeling.
4. Boost-converter reverse regenerating operation when Q_2 is chopping and Q_3–D_4 are freewheeling.

2.2 Phase-Controlled Rectifier

Phase-controlled rectifiers convert ac to dc by using devices such as thyristors and triacs, and are extensively used for industrial applications. As the name indicates, the converter uses the phase-angle control principle to control the dc output voltage, and a segment of ac voltage in reverse polarity is used for commutation (line commutation) of the devices. Historically, this class of converter has been available from the beginning of this century, although primitive converters used gas-tube devices, such as the mercury-arc tube (glass bulb and metal tank), ignitron, thyratron, phanotron, etc. The applications of present phase-controlled rectifiers include electrochemical processes (such as electroplating, anodizing, chemical gas production, aluminum reduction, metal refining, etc.), dc-motor speed control, dc-power supply for inverter, high-voltage dc (HVDC) converters for transmission systems, magnet power supplies, battery chargers, etc.

Figure 12 shows the popular three-phase, six-pulse bridge rectifier, and Fig. 13 gives the wave forms for a highly inductive resistance–inductance load. The delta-wye transformer on the ac-line side helps to establish the appropriate secondary voltage on the bridge and the neutral point, if needed. Each of the positive- and negative-side thyristors conducts for 120° ($\frac{1}{3}$ of a cycle), and the conduction of the negative-side device is phase shifted by 60°, as shown. The firing angle α, as indicated in (a), is controlled to control the dc output voltage. If $\alpha = 0$, the thyristor acts as a diode and the output voltage becomes maximum. Therefore, thyristors can be replaced by diodes for simple uncontrolled rectifier operation. If $\alpha = 90°$, the dc voltage or the delivered output power is zero in the ideal case. It can be shown that by varying the angle in the range $90° < \alpha < 180°$ the circuit can function as an inverter, and power from a battery (inversely connected) or energy trapped in the load inductance can be pumped to the ac line with the dc-current direction remaining the same. It is defined as a two-quadrant converter and can be used for dc-motor speed control where motoring and regenerative braking operations are possible. Because of limited load inductance, the load and line currents can be wavy, unlike that shown in Fig. 13. Again, the conduction may be continuous, as shown in the figure, or discontinuous at light load, particularly if the load contains counter emf.

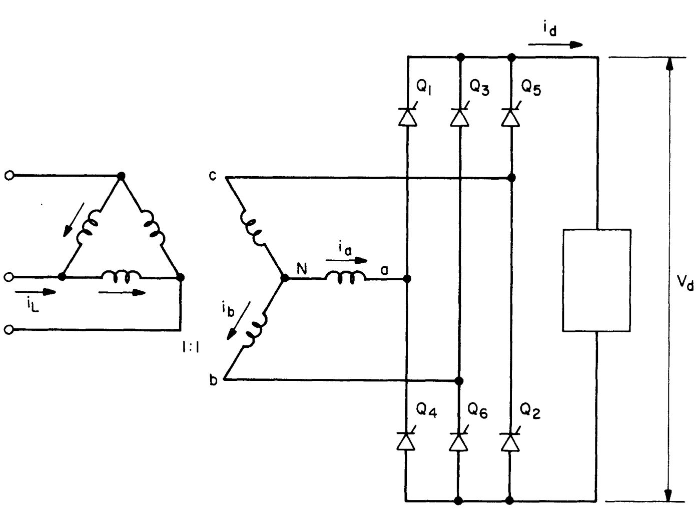

FIG. 12. Three-phase thyristor bridge rectifier.

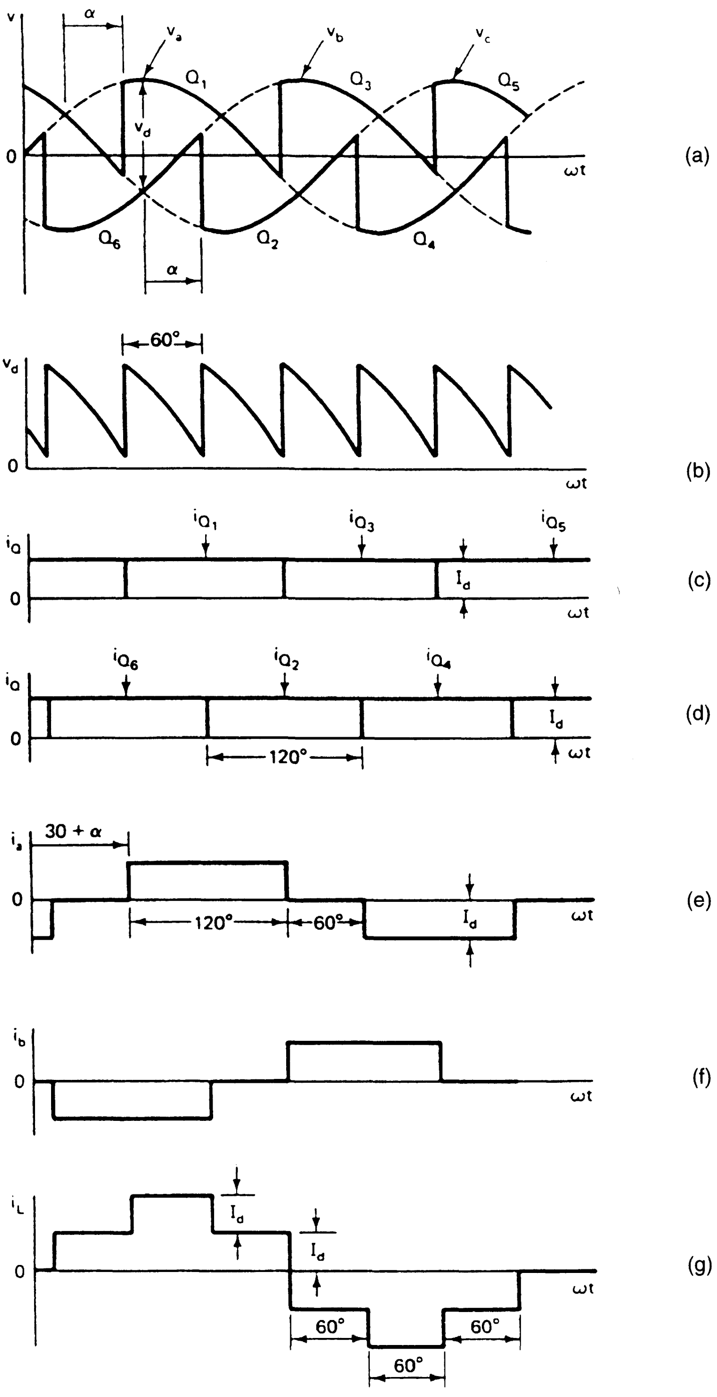

FIG. 13. Rectifier voltage and current waves: (a) envelopes for dc voltage, (b) dc-voltage wave, (c) thyristor currents for Q_1, Q_3, and Q_5, (d) thyristor currents for Q_2, Q_4, and Q_6, (e) input current i_a, (f) input current i_b, and (g) input line current.

For low power, a single-phase diode or thyristor rectifier can be used, which can have half-wave, transformer-coupled center-tapped, or full-bridge topology. A circuit can also be semibridge ($\frac{1}{2}$ of the devices are diodes) or full bridge in configuration. For high-power application, the number of phases or pulses can be multiplied (i.e., 12, 24, etc.) with a large number of devices. Such multiphase converters produce waves with low harmonic content in both dc and ac sides that can be easily filtered.

The rectifier can be considered as a nonlinear load, and the line-current and load-voltage wave forms contain rich harmonics. Besides, the fundamental component of line current lags the voltage wave form. The distorted line current tends to distort the line voltage as a result of source impedance. The large increase of phase-control–type converters on utility systems is causing power quality problems. Recently, harmonic standards, such as IEEE (Institute of Electrical and Electronic Engineers)-519 and IEC (International Electrotechnical Commission)-555 have been proposed to limit severely the harmonic injection into the utility system. Obviously, some type of filter must attenuate the harmonics, or the converter should be designed such that it does not generate any line harmonics.

2.3 Cycloconverter

A phase-controlled cycloconverter is basically a frequency changer that converts ac input power at one frequency to output at different but lower frequency with one-stage power conversion. The cycloconverters are used in ac-motor drives, slip power recovery wound-rotor induction motor drives (Scherbius drives), and varable-speed, constant-frequency (VSCF) systems, such as aircraft engine power generation at fixed frequency.

Figure 14 shows an 18-thyristor, three-phase, half-wave–type cycloconverter for an ac-machine load. For higher power, a 36-thyristor circuit can be used. In Fig. 14, each phase group consists of two half-wave converters connected in inverse parallel, which provides four-quadrant power control capability. The firing angle of each phase group is modulated sinusoidally to fabricate a mean sinusoidal voltage, as shown in Fig. 15. The other phase groups are identical except that the modulating waves are phase shifted by 120°. The output frequency (including zero frequency) and the depth of modulation can be varied for variable-frequency, variable-voltage output. The positive half-cycle of load current is supplied by the positive converter whereas the negative half-cycle is supplied by the negative converter. The nonconducting converter is blocked to prevent a short circuit.

A cycloconverter is basically a phase-controlled converter; therefore the basic characteristics, such as distortion of load voltage and line current and lagging input displacement factor, are retained. Because of modulation of firing angle, the displacement factor and harmonic content are worse. The typical output-frequency range is from zero (dc) to $\frac{1}{3}$ of the input frequency. A cycloconverter can also be operated in circulating-current mode where an intergroup reactor prevents the short circuit and limits the circulating current caused by harmonics. Although the reactor adds weight and the cycloconverter needs current derating, the advantages are that control is simple and the harmonic distortion is less, permitting higher output-frequency range (typically dc to $\frac{2}{3}$ of input frequency). Cycloconverters are used in the megawatt power range for ac-drive applications.

2.4 ac Controller

The ac controller, as the name indicates, controls the output voltage, but the output frequency remains the same as the supply

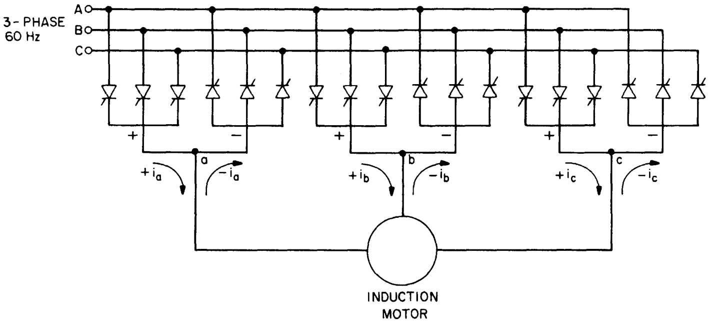

FIG. 14. 18-thyristor phase-controlled cycloconverter.

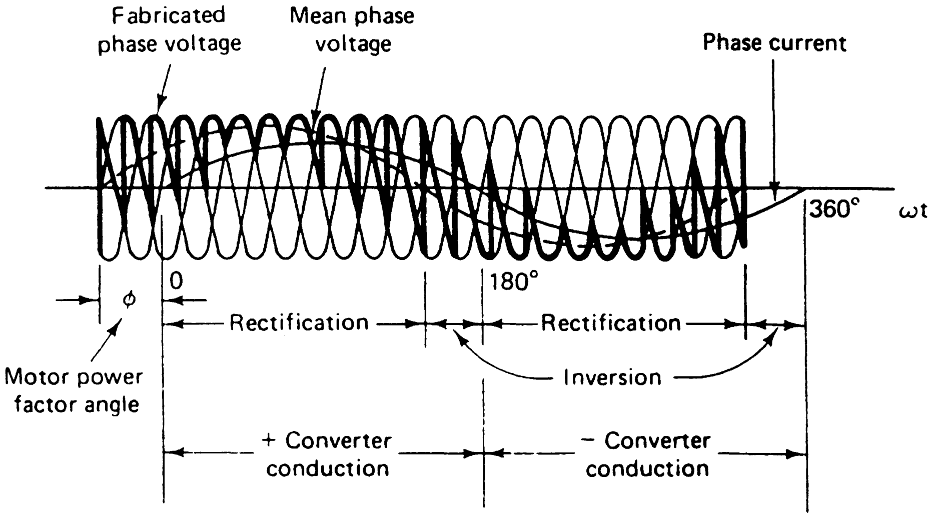

FIG. 15. Cycloconverter voltage and current waves.

frequency. An ac controller can be considered as a solid-state autotransformer where the output voltage is always less than the input voltage and the voltage control is done with the help of power semiconductor switches. Figure 16 shows a popularly used three-phase ac controller. The circuit can be used as a solid-state circuit breaker where the switches can be turned on at any instant, but can be turned off with a time delay when the device current falls to zero within a fraction of a cycle. In the "on" condition, there is power loss caused by conduction drop of the devices, which can be eliminated by bypass mechanical contactors. The control circuit can be designed to vary the thyristor firing angle gradually to get smoothly variable voltage at the output. The circuit has been popularly used as a solid-state starter for induction motors, where the voltage is gradually varied to limit the large starting current. The three-phase or single-phase version of the circuit has also been used in induction-motor speed control, transformer tap changing, induction heating, and static reactive power generation. If the load is nearly resistive, the antiparallel thyristors can be replaced by triacs. Triac-controlled circuits are used for incandescent lamp dimming control and heating control applications. Because of phase control, both line and load sides contain heavy distortion, and the line-side power factor (lagging) tends to be low. The harmonic currents flowing through the machine cause large copper loss, and the machine efficiency becomes low if the circuit is used for speed-control applications.

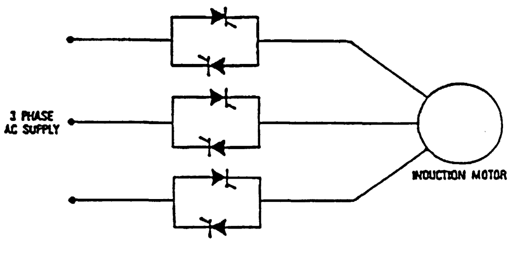

FIG. 16. Three-phase ac controller.

2.5 Voltage-Fed Inverter

The function of an inverter is to convert dc to ac, as mentioned before. In a voltage-fed inverter, the input supply is voltage, whereas in a current-fed inverter (which will be discussed later) the input supply is current. Often, a large electrolytic capacitor is connected at the input of a voltage-fed inverter. The inverter can be single-phase or polyphase type. The single-phase type can be of half bridge, full, or H bridge, or push-pull center-tap configurations. The single-phase half-bridge units can be combined to form three-phase or other multiple-phase to-

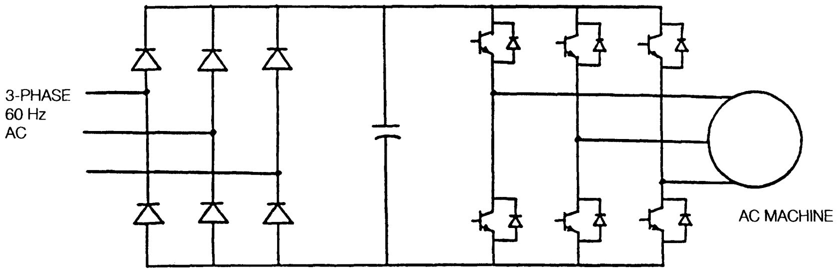

FIG. 17. Three-phase PWM inverter with diode rectifier.

pologies. Voltage-fed inverters are used for ac-machine drives, regulated voltage and frequency power supplies, uninterruptible power supplies, induction heating, ultrasonic wave generation, static reactive power generators, and active harmonic filters.

Figure 17 shows a popularly used three-phase bridge voltage-fed inverter with a diode bridge rectifier in the front end. The diode rectifier converts 60-Hz line ac into dc without any control, and the large dc-link filter capacitor filters harmonics from the voltage and provides a smooth voltage source for the inverter. The rectifier can be single-phase or three-phase type and generates discontinuous line-current wave. The inverter can generate the output voltage at a desirable frequency, as demanded by the load. The inverter uses the pulse-width modulation (PWM) principle to generate the output. Among several PWM techniques, the sinusoidal PWM (SPWM) is very popular and is explained in Fig. 18. The sine-wave signal voltage at the desirable frequency is compared with a triangular carrier wave and at intersection the upper and lower devices of an inverter leg are switched on, as indicated. There are three such controllers for the three phases where each signal wave is 120° phase shifted from the others. The output phase voltages contain harmonics of the carrier frequency, which are easy to filter. The carrier frequency is the same as the device switching frequency of the inverter. The output voltage can be regulated irrespective of supply and load variation, and the range can be from zero to a maximum value determined by the dc voltage. The fundamental frequency may vary from zero (dc) to an upper value at square wave dictated by the load. In square-wave mode, the inverter generates maximum fundamental frequency voltage. The power flow in an inverter can be controlled to be in either direction. For regenerative braking of an ac machine, the machine mechanical energy can be converted to electrical energy and pumped to the dc link through the inverter. In this mode, the in-

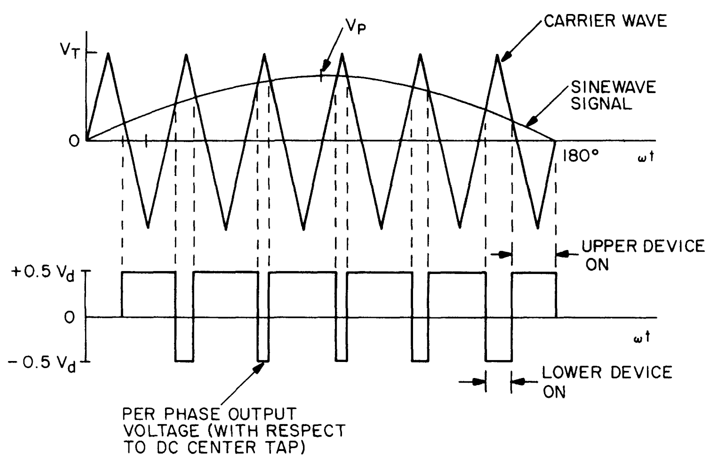

FIG. 18. Principle of sinusoidal pulse-width modulation.

verter operates as a rectifier. However, the excess dc power will raise the dc-link voltage to a dangerous level. Normally, a dynamic brake that consists of a buck chopper with a resistive load is connected to the dc link to absorb this power. In fact, the diode rectifier can be replaced by the inverter circuit to recover the energy into the ac line. Such an addition of a PWM "rectifier" on the line side not only permits full four-quadrant operation of the drive but shapes the input line current to be sinusoidal at unity power factor. Modern multimegawatt voltage-fed converter systems use GTOs, whereas IGBTs and power MOSFETs are preferred for medium- and low-power converters, respectively.

2.6 Current-Fed Inverter

A current-fed inverter requires a stiff dc-current source, as indicated before. A voltage source can be converted to a current source through an inductor and then controlling the voltage through the input converter. The phase-controlled, line-commutated, current-fed inverter (inverting mode of a rectifier) was discussed in Sec. 2.2. The inverter has various single-phase and polyphase configurations, and their applications include ac-motor drives, induction heating, ac-power supplies, static reactive power generators, and ac-line active harmonic filters.

A current-fed inverter can be classified into load-commutated, force-commutated, and self-commutated types. The former two types operate in six-pulse mode with a bridge configuration (similar to a phase-controlled converter), whereas the self-commutated type operates in PWM mode. A load-commutated inverter, which has identical topology with a phase-controlled rectifier, requires the load to be of leading power-factor type, such as an overexcited synchronous machine, so that the converter thyristors can be commutated by a segment of inverse load voltage. A practical load-commutated inverter has a phase-controlled rectifier in the front end and therefore provides easy four-quadrant speed control capability of the drive. Large multimegawatt wound-field synchronous machine drives use load-commutated converters. The converter system is very economical, but a minimum machine speed should be maintained for satisfactory load commutation by machine counter emf.

A current-fed inverter that supplies a lagging power-factor load, such as induction motor, needs forced commutation. Figure 19 shows the autosequential–type current-fed inverter (ASCI) where the capacitors and the diodes constitute the commutation elements. The diodes tend to isolate the commutating capacitors from the machine terminal. When an incoming thyristor is fired, capacitor voltage with inverse polarity is impressed across the outgoing thyristor to turn it off. This type of inverter is used in medium- to high-power applications, but recently, its importance is diminishing. The harmonic current in both line and load sides and machine torque pulsation at low frequency are problems.

The ASCI in Fig. 19 can be replaced by a self-commutated PWM GTO inverter, as shown in Fig. 20. The advantages are that

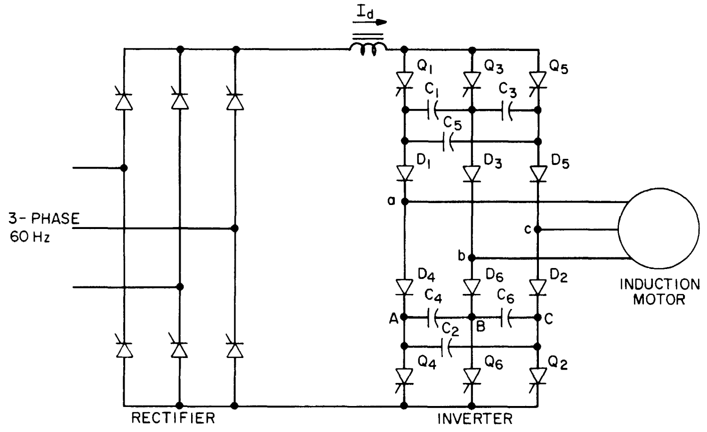

FIG. 19. Force-commutated autosequential inverter with thyristor rectifier.

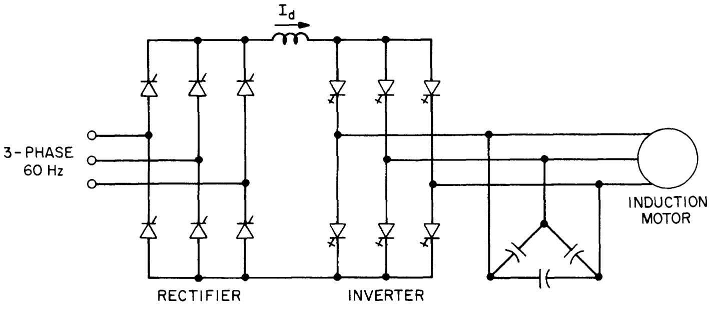

FIG. 20. PWM current-fed inverter with thyristor rectifier.

the PWM machine-current wave can be easily filtered, there is no torque pulsation problem, and the dc-link inductor size is lower. The reverse-blocking GTOs can be replaced by IGBTs with series diodes for medium-power drives. The delta shunt capacitor bank, as shown in the figure, has the functions of providing a bypass path for the load inductive current so that the GTOs can be commutated without undue voltage overshoot, and low-pass filtering the PWM harmonics so that the machine currents are nearly sinusoidal. Therefore, machine harmonic copper loss, torque pulsation, and the characteristic acoustic noise disappear. Since the inverter can carry power in either direction, the line-side phase-controlled rectifier can be replaced by the inverter unit, giving the advantage of sinusoidal line current at unity power factor.

3. SELECTED APPLICATIONS OF POWER ELECTRONICS

Many of the previous sections have touched on various applications of power electronics. In this section, a few more selected applications will be systematically reviewed.

3.1 Incandescent-Lamp Dimming

A simple application of power electronics is incandescent lamp dimming control with the help of a diac-triac phase control circuit (GE SCR Manual, 1979), shown in Fig. 21. The single-phase ac supply is connected in series with a lamp through a triac. The firing angle of the triac is controlled by the *R-C* circuit and diac to control the rms voltage across the lamp. A diac is basically an *npn* negative-resistance threshold device. The potentiometer RI is adjusted to control the dimming, where the lowest RI value gives maximum brightness of the lamp. A typical capacitor voltage (V_c) wave is shown in the figure. When the phase-lagged V_c wave crosses the diac threshold voltage V_s, a pulse of gate current flows, firing the triac. With a lower RI value, the threshold will be reached earlier in the cycle, giving a control angle in the range from α_1 to α_2. The critical resistance that will not fire the triac is shown by

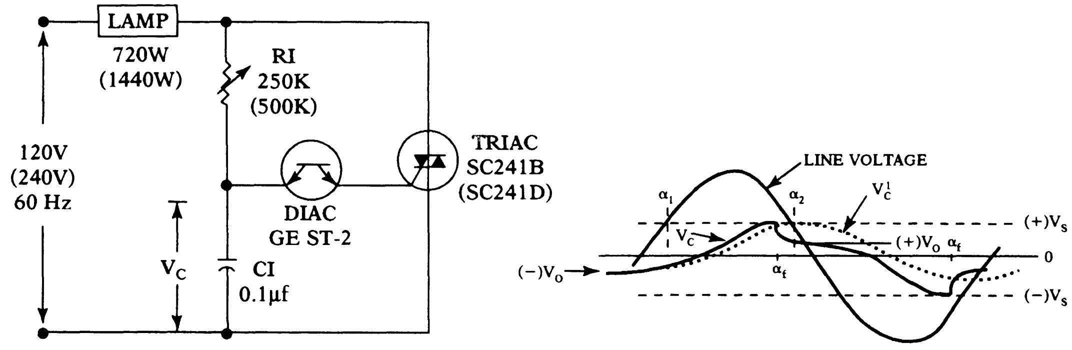

NOTE:
VALUES IN PARENTHESES ARE FOR 240VAC SUPPLY.

FIG. 21. Incandescent lamp dimming control.

the dotted wave. As the resistance is decreased from this value, V_c will rise and cross the V_s threshold, and the triac will switch on with a "snap-on" effect.

3.2 High-Frequency Ballast for Fluorescent Lamps

The efficiency of a fluorescent lamp is much higher than that of an incandescent lamp although the lamp is somewhat more expensive. The efficiency can be improved further by using high-frequency lamp ballast. A typical high-frequency ballast system (Bose, 1992) is shown in Fig. 22. The 60-Hz, single-phase supply is rectified to dc with the help of a diode-bridge rectifier and then inverted to 30-kHz ac with a half-bridge voltage-fed PWM inverter. The lamp cathodes of the two lamps are continuously heated by the three secondary windings, as shown. The inverter output is boosted by an autotransformer and applied across the two lamps in series. The transformer winding leakage inductances provide the series ballast needed for stable operation at lamp conduction when its resistance is low. A filter is normally used at the input of the rectifier to suppress conducted electromagnetic interference. A high-frequency ballast not only has smaller size, but also gives better efficiency, reduces the low-frequency stroboscopic effect, and provides lamp dimming capability by control of the tube current.

3.3 Switching-Mode Power Supply

A switching-mode power supply (SMPS) (Bose, 1992) generally converts an unregulated dc voltage to a regulated dc voltage; the unregulated dc can be obtained from a battery or ac line through a diode rectifier and a filter. The SMPS generally supplies power to an electronic circuit of a television, personal computer, VCR, tape recorder, etc. Primitive power supplies used linear regulators where efficiency was poor. The switching-mode power semiconductor device improves efficiency at the cost of harmonic generation. However, with high-frequency devices, such as power MOSFETs, only a nominal filter size is needed.

The SMPS can generally be classified into two types: PWM buck, boost, or buck-boost dc-dc converter, as indicated before, and resonant-link dc-dc converter. Figure 23 shows a typical voltage-fed parallel resonant-link converter. The primary dc voltage (from a rectifier or a battery) is split with the help of capacitors and then inverted to high-frequency (several megahertz) ac by a MOSFET half-bridge inverter. The inverter supplies a series resonance (L_r-C_r) load, and the transformer connected across the capacitor feeds the high-frequency Schottky-diode rectifier, which then supplies the dc output. The inherent body diode of the device is used, and the snubber consists of a simple capacitor, as shown. The high-frequency link, al-

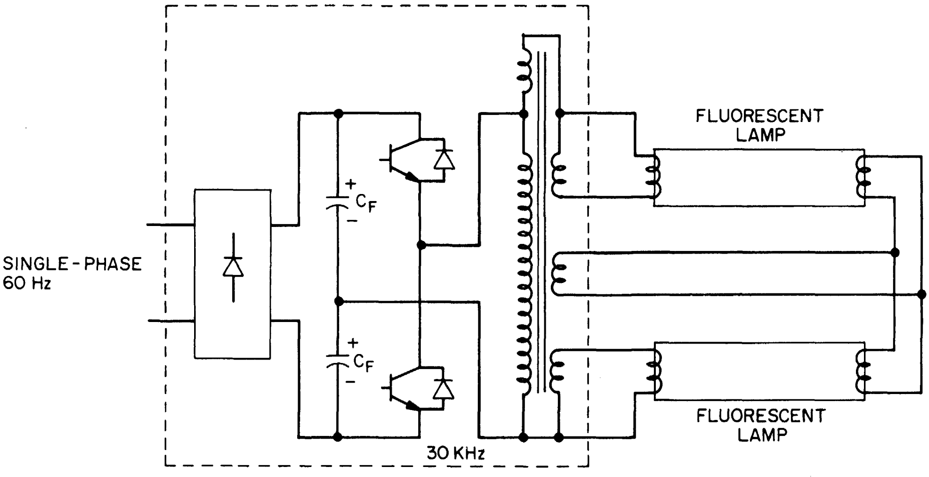

FIG. 22. High-frequency fluorescent lamp ballast.

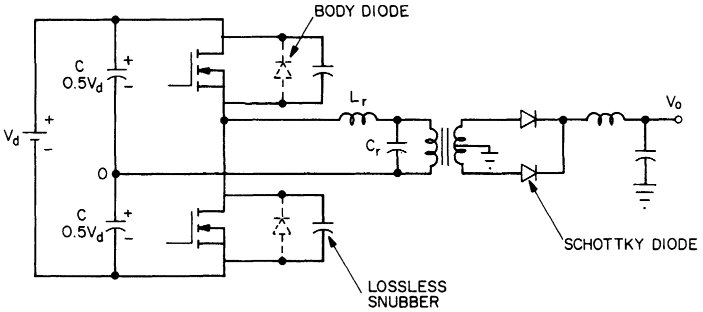
BODY DIODE
C
0.5Vd
+
-
Vd
0
C
0.5Vd
Lr
Cr
Vo
SCHOTTKY DIODE
LOSSLESS SNUBBER

FIG. 23. Resonant-link switching-mode power supply (dc-dc converter).

though it requires an additional conversion stage, provides several advantages. The operation at higher than resonant frequency permits zero-voltage switching (ZVS) of the devices, eliminating switching loss. The result is that efficiency is improved, and switching frequency much higher than that of PWM reduces the size of the isolation transformer and output passive filter. The transformer helps to change the primary voltage level. The output voltage can be regulated by varying the inverter frequency. The snubber capacitor acts as a lossless snubber, i.e., it stores energy that is returned back to the supply. For higher power, the half bridge can be replaced by a full bridge.

3.4 Induction Heating

For induction heating of metal, ac current is passed through a coil surrounding the metal workpiece. The iron loss (hysteresis and eddy-current losses) raises temperature of the metal. The induction heating for melting uses low frequency so that the heat is developed throughout the body, whereas the frequency for heat treatment may be very high where the skin effect confines the heat near the surface.

The induction heating system can use either a voltage-fed series resonance inverter or a current-fed parallel resonance inverter where the resonance-circuit inductance constitutes the work coil. Figure 24 shows a practical voltage-fed inverter induction heat-

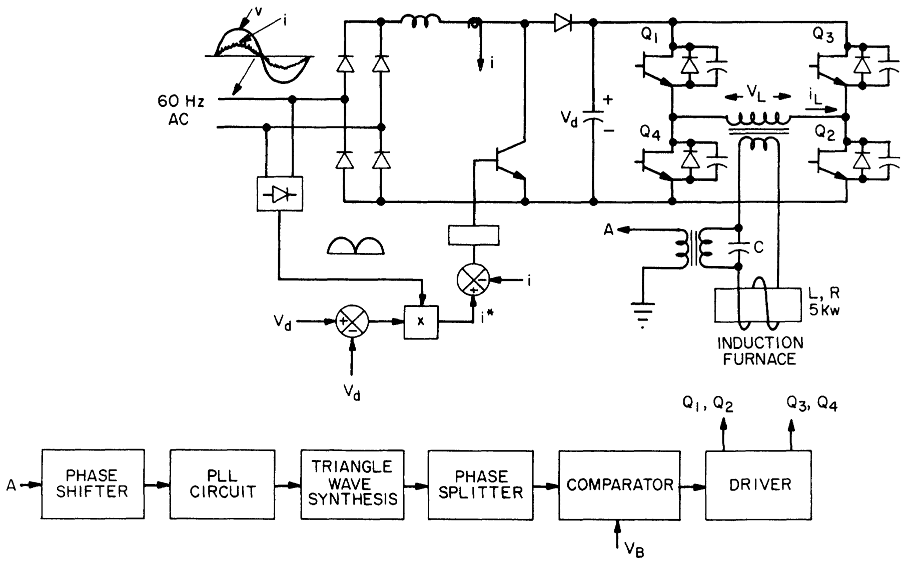
V
i
60 Hz
AC
i
Vd
+
-
Q1
Q3
Q4
Q2
VL
iL
A
C
L, R
5Kw
INDUCTION FURNACE
i
i*
x
Vd
Vd
Q1, Q2
Q3, Q4
A
PHASE SHIFTER
PLL CIRCUIT
TRIANGLE WAVE SYNTHESIS
PHASE SPLITTER
COMPARATOR
DRIVER
VB

FIG. 24. Induction heating with voltage-fed resonant inverter.

ing system using a static induction transistor (Ogiwara *et al.*, 1987). The dc supply for the inverter is obtained from a 60-Hz line through a diode bridge rectifier and a transistor boost chopper. The boost chopper not only regulates the capacitor voltage but shapes the line current to be sinusoidal at unity power factor. The feedback voltage loop error is multiplied by the full-wave rectified line-voltage wave to generate the command-current wave of the rectifier. This current is then controlled by the boost chopper using the hysteresis-band PWM control principle. The inverter is a full-bridge type where the output is stepped down through a transformer and applied to the work coil through a series tuning capacitor. The equivalent-circuit R and L parameters of the coil vary with temperature, and therefore the resonant frequency also varies. The inverter frequency tracks the resonant frequency so that a slightly higher inverter frequency always permits lossless zero-voltage switching, as indicated before. The load power flow is regulated by control of the dc-link voltage. The basic control elements of the self-controlled converter system are indicated. The capacitor voltage, which lags the coil current by 90° and the load voltage V_L by an additional lag angle, is phase shifted for control of the inverter. The signal is converted to high frequency by a phase-locked loop (PLL) circuit and then used to synthesize a triangular wave by an UP-DOWN counter. The triangular wave is phase split and compared with a bias voltage V_B to generate the logic drive signals for the devices. The bias signal generates the appropriate lockout time to prevent shoot-through fault. Since the devices are normally on, they are biased off outside the desired conduction intervals.

3.5 Uninterruptible Power Supply (UPS)

Reliable and improved-quality power supply is important for critical loads, such as computers, communication equipment, and medical systems. The utility power supply, however, has the problems of outage, brownout, and wave form distortion. In a UPS system, a battery-backed supply is available to supplement the utility supply. The UPS system may be ac or dc, and the system topology varies widely, depending on the application. Figure 25 shows a typical three-phase UPS system (Griffith, 1989) where the load can be supplied either from the utility line or from the battery-backed inverter. An electronic circuit breaker consisting of antiparallel thyristor switches helps in the fast transfer of the source. In the figure, the utility system normally supplies the load and the inverter is the standby. The inverter system consists of a diode rectifier, battery charger, storage battery, LC filter in the dc link, PWM voltage-fed inverter, and output LC filter. The inverter generates sinusoidal regulated voltage at 60 Hz at the output from the unregulated dc input. Normally, the rectifier directly supplies the inverter input when the thyristor Q remains off. At utility

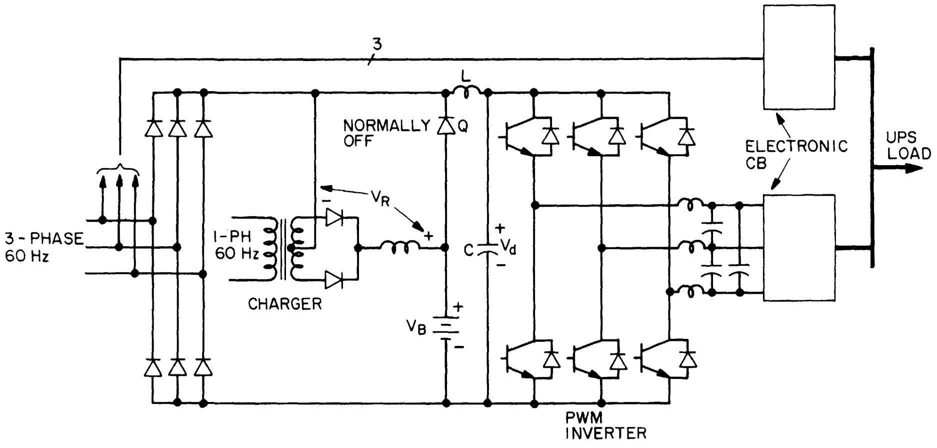

FIG. 25. Three-phase UPS system with standby inverter.

power interruption, Q is switched on and the output circuit breaker switches over the supply from the inverter. When the inverter is remaining as standby, the battery is charged by the single-phase thyristor rectifier. The rectifier boost voltage $+V_R$ acts in series with the voltage V_d to charge the battery. The transfer of load from line to inverter and back causes power interruption for fraction of a cycle, and this should be tolerated by the load.

3.6 dc-Motor Drive

One of the earliest applications of power electronics is the speed control of dc motors. dc-motor drives (Sen, 1981) in the power range of a few watts to several megawatts are widely used in industry, and the applications include instrumentation drives, robotics and machine tool drives, paper and textile mills, rolling mills, etc. In a dc drive, normally the field current is maintained constant to establish the desired flux and the armature current is varied to control the torque, which varies the speed. Generally, there are two classes of dc drives: dc-dc converter or chopper-fed drives and phase-controlled converter-fed drives. Chopper-fed one-, two-, or four-quadrant drives were already discussed in Sec. 2.1. The machines in these drives are of permanent magnet type and are of small power rating. Larger-power machines with separate field supply are generally controlled by a thyristor phase-controlled rectifier, shown in Fig. 12. The converter can provide two-quadrant speed control of the machine. This can be explained as follows: If the firing angle of the converter is reduced, the output dc voltage will increase, giving higher speed. The regenerative braking mode of operation in the same direction of rotation can be obtained by reversing the field supply and controlling the firing angle above 90°. The rectifier will then operate as an inverter with the negative counter emf load and the machine speed will fall.

Figure 26 shows a thyristor dual converter for four-quadrant speed control that is popularly used in industrial process-control applications. The positive converter (left side) can generate positive (rectifier) or negative (inverter) voltage, but the dc current is always positive. This converter will function for forward motoring and reverse regeneration. The negative converter (right side) can

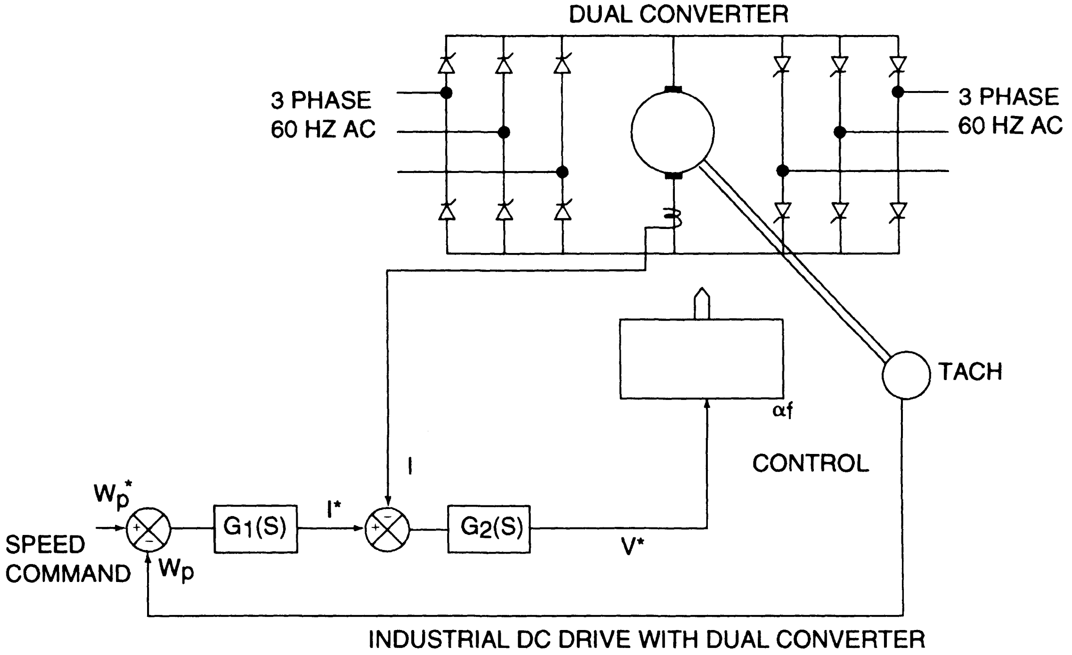

FIG. 26. Thyristor dual-converter speed control of dc motor.

generate negative (rectifier) or positive (inverter) voltage, but the dc current is always negative. This unit will permit the other two quadrants, i.e., reverse motoring and and forward regeneration. When one converter unit is operating, the other unit should be blocked to prevent short circuit. The simplified speed-control block diagram is indicated in the figure. The command speed is compared with the motor shaft speed, and the error through a controller establishes the current command for the motor armature. The armature current is then controlled by controlling the firing angle.

3.7 ac-Motor Drive

ac machines of induction and synchronous types are widely used in industry, and the application of variable-speed ac drives is increasing extensively. In this section, the popularly used induction-motor drives using squirrel-cage machines (Bose, 1986) will be discussed.

3.7.1 Open-Loop Volts/Hz Control

The speed of an induction motor can be efficiently controlled by a variable-frequency, variable-voltage power supply, as shown in Fig. 27. The converter system (same as Fig. 17) generates such a power supply for the machine. It can be shown that the airgap flux of the machine remains constant like that in a dc machine if the ratio of voltage to frequency remains constant. In Fig. 27, the speed of the machine is commanded by the frequency because the machine slip frequency is very small. The voltage signal is proportional to frequency, except that at lower speed range the signal is boosted to compensate the stator drop. The machine speed can be varied by variation of the frequency signal. However, a sudden variation of the frequency command will cause loss of machine stability. The acceleration capability is limited by the machine and inverter power ratings, and the load torque. Since there is no closed-loop control of speed, the actual speed will be slightly lower (by the slip frequency) than the desired speed. The slip frequency can be calculated from the stator voltage and current signals and added with the command signal to compensate the speed droop. The drive is simple and economical, but the performance (including the transient response) is not very good.

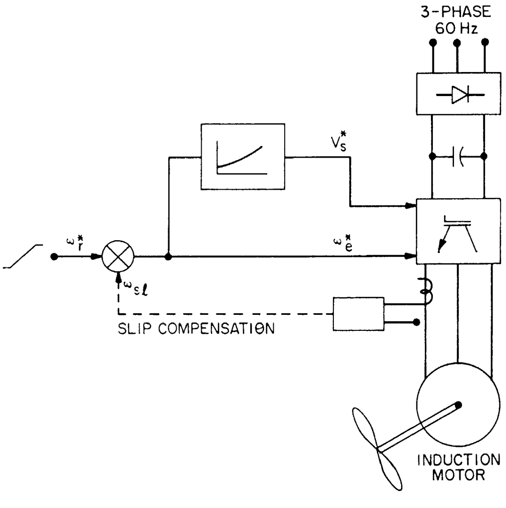

FIG. 27. Open-loop volts/Hz speed control of induction motor.

3.7.2 Vector Control

An induction motor with vector or field-oriented control gives performance that is almost identical to that of a separately excited dc motor. This means that the transient response is dc machinelike, the conventional instability problem does not arise, and the machine torque and flux can be controlled independently. The vector control can be classified into two types, i.e., direct or feedback and indirect or feedforward methods. The popularly used indirect method will be reviewed in this section. Figure 28 shows a position servo drive using the indirect vector-control method. The power circuit (same as Fig. 17) of the drive includes a dynamic brake to give four-quadrant control capability of the drive. In Fig. 28, the position-control loop generates the speed command and the speed-control loop generates the torque command (essentially the torque component of current i_{qs}), as shown. The machine is operated at a rated flux that is determined by the field component of current (i_{ds}). The slip-frequency command (ω_{sl}^*) is generated proportional to i_{qs}, and the proportionality factor is defined as the slip gain K_s. The slip frequency is added with the speed signal to

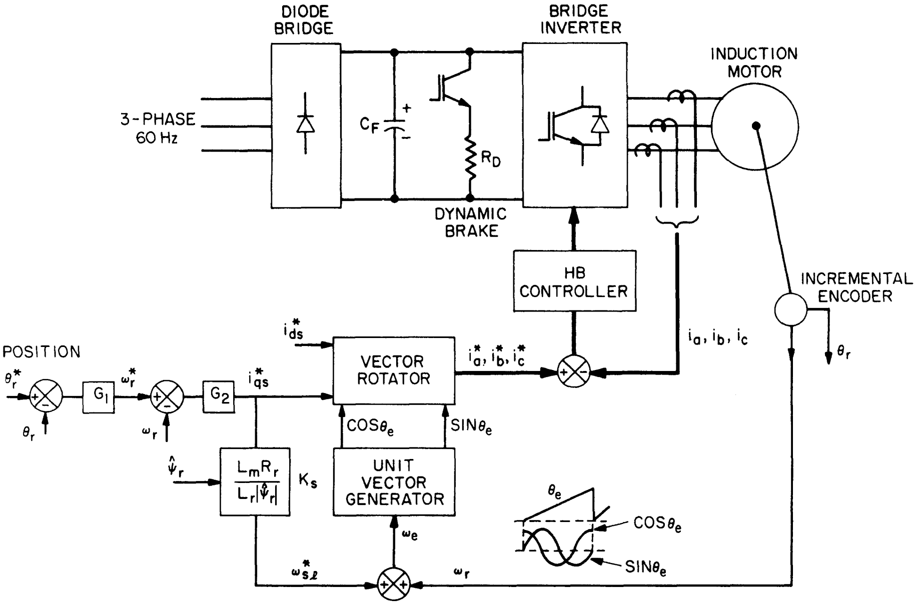
DIODE BRIDGE
BRIDGE INVERTER
INDUCTION MOTOR
3-PHASE 60 Hz
C_F
R_D
DYNAMIC BRAKE
HB CONTROLLER
INCREMENTAL ENCODER
θ_r
i_a, i_b, i_c
POSITION
θ_r^*
G_1
ω_r^*
G_2
i_{ds}^*
i_{qs}^*
VECTOR ROTATOR
i_a^*, i_b^*, i_c^*
θ_r
ω_r
$\hat{\psi}_r$
$\frac{L_m R_r}{L_r |\hat{\psi}_r|}$
K_s
$\cos\theta_e$
$\sin\theta_e$
UNIT VECTOR GENERATOR
ω_e
θ_e
$\cos\theta_e$
$\sin\theta_e$
ω_{sl}^*
ω_r

FIG. 28. Vector control of induction motor.

generate the frequency command, which is then converted to unit vector signals. The vector rotator converts the input dc-current commands into the ac stator-current commands of the machine. The command-phase currents are then compared with the actual currents, and the IGBT gate-drive signals are generated through the hysteresis-band current controller. Note that in vector control the current (instead of voltage) is controlled, and the machine frequency is essentially determined by the speed signal (self control). In braking mode, the polarity of i_{qs} becomes negative. For speed reversal, the phase sequence of the inverter output automatically reverses at zero speed.

GLOSSARY

Asymmetric Blocking: Blocking of device voltage in forward direction only.

Boost Chopper: A chopper in which the output voltage is higher than the input voltage.

Buck Chopper: A chopper in which the output voltage is lower than the input voltage.

Carrier Wave: A control wave in a PWM inverter that determines the switching frequency.

Commutation: Transfer of current from one device or circuit to another.

Firing Angle: Control phase angle of thyristor or triac.

Forced Commutation: A commutation that is performed forcibly by an auxiliary circuit.

Harmonic Frequency: A frequency that is a multiple of the fundamental frequency.

Line Commutation: Commutation that is performed by a segment of inverse ac-line voltage.

Load Commutation: Commutation that is performed by a segment of inverse ac-load voltage.

Power Factor: Ratio of effective power to apparent power.

Regenerative Braking: Braking of a machine by operating it as a generator and pumping back its stored mechanical energy to the source in electrical form.

Slip Frequency: Frequency of the induced voltage in the rotor of an induction motor.

Square Wave: Mode of operation of an inverter where each phase generates a square voltage wave.

Symmetric Blocking: Blocking of device voltage in both directions.

Vector Control: Control method of an ac machine that gives it dc-machine–like performance.

Works Cited

Bose, B. K. (1986), *Power Electronics and AC Drives,* Englewood Cliffs, NJ: Prentice-Hall.

Bose, B. K. (1992), *Proc. IEEE* **80,** 1301–1334.

General Electric SCR Manual (1979), 6th ed., Auburn, NY: General Electric.

Griffith, D. C. (1989), *Uninterruptible Power Supplies,* New York: Dekker.

Ogiwara, H., Kimura, O., Dezaki, T., Nakaoka, M., Maruhashi, T. (1987), in: M. A. Miller (Ed.), *Official Proceedings of the Fourteenth International PCI 87 Conference (SATECH 87),* Ventura, CA: Intertec Communications Inc., pp. 146–155.

Sen, P. C. (1981), *Thyristor DC Drives,* New York: Wiley.

Further Reading

Bose, B. K. (Ed.) (1981), *Adjustable Speed AC Drive Systems,* New York: IEEE.

Bose, B. K. (1989a), *IEEE Industrial Applications Society Annual Meeting Conference Record,* Piscataway, NJ: IEEE, pp. 790–797.

Bose, B. K. (1989b), *IEEE Ind. Electron. Trans.* **36,** 403–412.

Bose, B. K. (Ed.) (1992), *Modern Power Electronics,* New York: IEEE.

Dubey, G. K. (1989), *Power Semiconductor Controlled Drives,* Englewood Cliffs, NJ: Prentice-Hall.

Mohan, N., Undeland, T. M., Robbins, P. (1989), *Power Electronics,* New York: Wiley.

Murphy, J. M. D., Turnbull, F. G. (1988), *Power Electronic Control of AC Motors,* New York: Pergamon.

Rashid, M. H. (1988), *Power Electronics,* Englewood Cliffs, NJ: Prentice-Hall.

POWER ENGINEERING

See POWER TRANSMISSIONS

POWER TRANSMISSION

James R. Stewart and Laurie J. Oppel, *Power Technologies, Inc., Schenectady, New York, U.S.A.*

	Introduction	25
1.	**Power Transmission System**	26
1.1	Transmission-System Link between Generation and Customer	26
1.2	Substations	27
1.3	Transmission Lines	27
1.4	Typical Transmission Voltages and Capable Power Transfer	28
2.	**Transmission-Line Theory**	28
2.1	Low-Frequency Approximation to Electromagnetic Equations	28
2.2	Traveling-Wave Considerations	29
2.3	Pi-Equivalent Circuit	29
2.3.1	Short- and Long-Line Models	29
2.3.2	Parameters that Affect Resistance, Inductance, and Capacitance	29
2.4	Surge Impedance and Surge-Impedance Loading	30
3.	**Power-Transmission System Considerations**	31
3.1	Power Control	31
3.2	Voltage and Reactive Power Control	32
3.3	Protection	33
3.4	Stability	34
3.5	Reliability	35
3.6	Economics	35
4.	**Overhead ac Transmission Lines**	35
4.1	Line Parameters	35
4.2	Conductor Types and Selection	36
4.3	Insulation Designs and Considerations	37
4.4	Ampacity	38
4.5	Electrical Environmental Considerations	39
4.6	Support Structures: Mechanical and Civil Engineering Considerations	41
5.	**Underground ac Transmission Lines**	41
5.1	Line Parameters	41
5.2	Types of Underground Installations	42
5.3	Insulation Considerations	42
5.4	Cable Ampacity	43
5.5	Environmental Considerations	44
5.6	Specific System Considerations	44
6.	**High-Voltage Direct-Current Transmission**	45
6.1	Line Parameters	45
6.2	Types of dc Installations	46
6.3	Converter Stations	46
6.4	Electrical Environmental Considerations	47
6.5	Specific System Considerations	47
	Glossary	48
	Works Cited	48

INTRODUCTION

Electricity, in a form useful for commercial application, does not exist naturally in the same way as a source such as natural gas. Electric power must be generated by large central power plants from some other energy source, such as water at a hydroelectric plant. For a variety of reasons, the power-generating stations are normally located far from population centers, the location of the major electrical usage. The electricity then needs to be delivered to the customers through transmission and distribution lines, which form a required portion of the metallic connection between the gen-

digital Encyclopedia of Applied Physics

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

erating plant and the customer end-use appliance or device.

1. POWER TRANSMISSION SYSTEM

Figure 1 shows a simplified electric power system (Bonneville Power Administration, 1989). The transmission system efficiently transports bulk electric power long distances from the generator to the main substation that services a given area. The voltage of the transmission lines is usually above 69 kV and, in some parts of North America, may be as high as 765 kV. At the main substation, the voltage is reduced, or stepped down, to distribution-level voltages, normally 2.4–35 kV. The distribution lines carry the electricity into the neighborhoods, offices, and factories for individual customers.

Through the years, transmission lines have been built to operate at progressively higher voltages, as shown in Fig. 2 (EPRI, 1982). With each advance in voltage level, greater amounts of power could be transported more efficiently, economically, and with fewer circuits on less right of way. Unfortunately, the increase in voltage levels also brought about public conflicts. Before 345-kV transmission lines were introduced, the most noticeable effects were conflicts over land use and the aesthetic appearance of the line. With the introduction of extra-high-voltage (EHV) lines (i.e., 345 kV and above), new public conflicts occurred dealing with radio and television reception, audible noise, and electric and magnetic field effects. Section 4.5 discusses these electrical environmental considerations in more detail.

1.1 Transmission-System Link between Generation and Customer

The primary function of the transmission line is to transport electricity from the generating plant to the customer in a reliable,

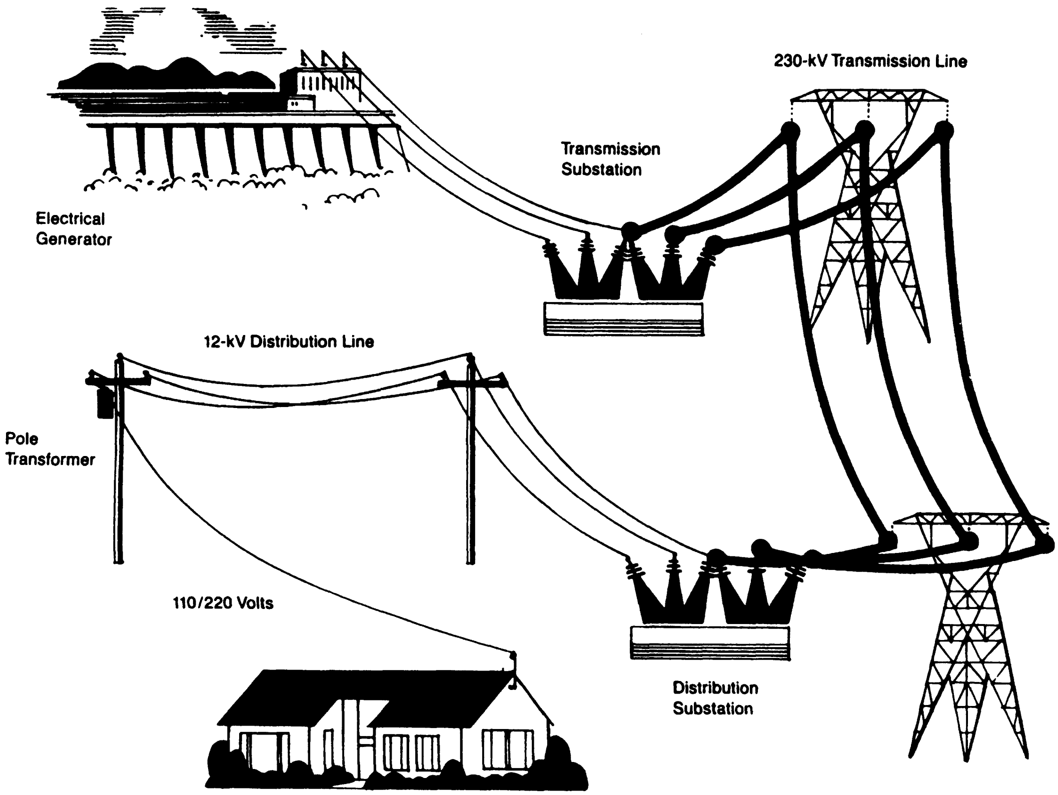

FIG. 1. Simplified electric power system.

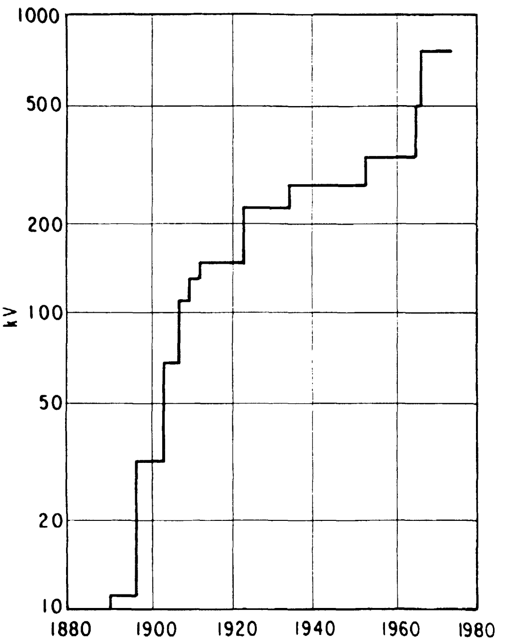

FIG. 2. Maximum ac transmission voltages in North America.

efficient, and economical manner. To meet these objectives, transmission-line design has evolved in the last century as a result of continual research and refinement. For technical and economical reasons, power transmission lines are typically constructed overhead. On the basis of aesthetic and other land use considerations, transmission lines may also be constructed underground.

1.2 Substations

In large, modern power systems, the transmission and distribution system functions to deliver bulk power from generating sources to the end users. Transmission facilities generally include step-up transformers, interconnecting transmission lines, switching stations, and step-down transformers. As an integral part of the transmission system, the substation or switching station functions as a connection and switching point for power to transmission lines, subtransmission feeders, generating circuits, and step-up/step-down transformers. The electrical equipment in a typical substation may include circuit breakers, disconnecting switches, grounding switches, current transformers, voltage (potential) transformers, power transformers, coupling capacitors, line traps, lightning arresters, shunt reactors, shunt capacitors, series capacitors, station buses, and insulators.

The substation design objective is to provide maximum reliability, flexibility, and continuity of service and to meet these objectives with the lowest investment costs that satisfy system requirements. Many factors influence the proper selection of the type of substation for a given application. The most appropriate type of station depends on voltage levels, load capacity, environmental considerations, and site space limitations.

There are two basic types of insulation for substations: air and gas. Air-insulated substations use the surrounding air as electrical insulation between energized buses, much in the same way that air is used as the primary insulation between bare conductors on transmission lines. The basic principle of gas-insulated equipment is that current-carrying parts are within a metal enclosure and are held in a concentric configuration by spacer insulators. The space between the conductor and the enclosure is filled with sulfur hexafluoride gas (SF_6) under pressure. High-voltage gas-insulated substations have been in service since the early 1960s. Operation of 800-kV gas-insulated equipment has proved successful since 1979.

1.3 Transmission Lines

Electric power worldwide is primarily transmitted by means of alternating current (ac). Direct-current (dc) transmission allows certain economic and performance advantages over ac transmission in specific situations. Specific applications where dc transmission may be superior include long-distance overhead transmission, underwater or underground transmission, and asynchronous ties between distinct power systems.

Although direct-current transmission has seen some use in the United States, the concept has been around for a long time. In the 1880s and 1890s, there was widespread controversy over whether direct-current or alternating-current systems would be best for meeting the nation's electricity needs. High-voltage direct-current (HVDC) transmission was first seriously considered in the United States in the 1930s but became a commercial reality in 1954 in Sweden.

High-voltage dc is inherently more stable than ac and requires smaller lines for transmission, less insulation, and narrower rights of way. For long-distance lines, dc lines are cheaper to build than ac lines. A "bipolar" dc line requires two phase positions as opposed to an ac transmission line, which requires three sets of conductors. dc transmission lines may also aid the overall power-system stability—the ability to maintain an electrical connection following a major disturbance. Unfortunately, even with modern technological advances, the termination facilities for converting the ac electricity into dc electricity and back again are very expensive. Therefore, dc transmission remains practical only for transmitting large amounts of power over long distances. A dc transmission line is not easily tapped or sectionalized to create a network of dc lines, as exist for the ac power system. Furthermore, dc power still cannot be directly stepped down to a level appropriate for home and business use.

1.4 Typical Transmission Voltages and Capable Power Transfer

Standard transmission voltages are established in the United States by the American National Standards Institute (ANSI). There is no clear delineation between distribution, subtransmission, and transmission level voltages. Table 1 shows the standard ac voltages listed in ANSI Standards C84 and C92.2, all of which are presently in use, and approximate loadability in megawatts (MW).

The nominal system voltages of 345, 400, 500, and 765 kV in Table 1 are classed as extra-high voltages (EHV). They are used extensively in the United States and in certain other parts of the world. The 400-kV transmission class voltage is principally used in Europe. EHV is used for the transmission of large blocks of power and for longer distances than would be economically feasible at the lower voltages. EHV may be used for interconnections between systems or superimposed upon large power-system networks to transfer large blocks of power from one area to another.

Table 1. Standard ac system voltages

Rating (KV)		Approximate loadability (MW)
Nominal	Maximum	
34.5	36.5	3
46	48.3	7
69	72.5	12
115	121	36
138	145	55
161	169	75
230	242	135
345	362	350
400	420	525
500	550	900
765	800	2250
1100	1200	5200

One voltage level above 800 kV, namely, 1100 kV nominal, is presently standardized and is referred to as ultrahigh voltage (UHV). This level has not as yet been installed commercially, although sufficient research and development have been completed to prove it technically feasible (EPRI, 1982; Barnes and Thoren, 1973; Bonneville Power Administration, 1974).

The maximum loading capability of a given transmission line is based on the line's surge impedance, which is determined by the line configuration (number of conductors per phase, phase geometry, and spacing). The actual power-transfer capacity is limited by thermal, voltage-drop, and stability considerations.

2. TRANSMISSION-LINE THEORY

2.1 Low-Frequency Approximation to Electromagnetic Equations

One wavelength at 60 Hz is 5000 km. Hence, at power frequencies of 60 Hz and below and the dimensions typically associated with electric power transmission lines, the electromagnetic fields of transmission and distribution facilities can be assumed to be quasistatic (Adler *et al.*, 1960). The quasistatic approximation is valid not only for calculation of line parameters but also for calculation of electric and magnetic field profiles and coupling to nearby objects (Olsen and Wong, 1992). If the line conductors had zero resistance, the waves would be transverse electromagnetic (TEM), and the electric and magnetic field behavior could be represented by static capacitances and inductances (Harrington, 1961). The resistance is sufficiently small for practical power transmission lines that the TEM approximation yields accurate values for the line parameters.

2.2 Traveling-Wave Considerations

Because electric power-facility dimensions are small with respect to a wavelength, it is customary to ignore traveling-wave effects for most engineering studies. The simple model shown in Figure 3 is adequate to represent power lines on a lumped-parameter basis. There are, however, several cases when traveling-wave behavior is important:

- power-frequency studies of very long lines (50 or 60 Hz),
- switching surge and lightning transients (millisecond to microsecond time frame), and
- power-line carrier communication (30–300 kHz).

Even for long lines and higher frequencies, it is still adequate to use the TEM approximation and lumped-parameter equivalent circuit, if the effects of wavelength are properly incorporated. The long-line correction is given in Sec. 2.3.1.

2.3 Pi-Equivalent Circuit

Power transmission and distribution lines can be represented for many purposes by the pi-equivalent circuit shown in Figure 3 (Bergen, 1986). The term Z represents the series impedance of the line, and $Y/2$ represents half the shunt admittance divided equally between the two line ends.

2.3.1 Short- and Long-Line Models. For "short" transmission lines, the parameters in Fig. 3 become

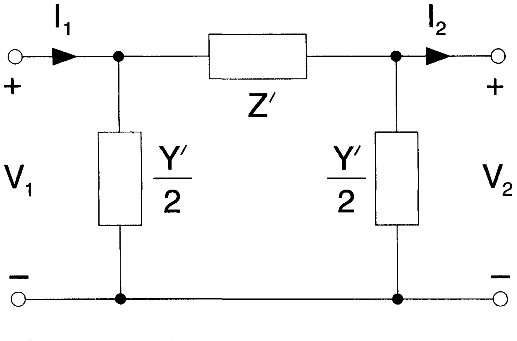

FIG. 3. Pi-equivalent circuit.

$$Z = R + j\omega L, \tag{1}$$

$$Y = G + j\omega C, \tag{2}$$

where R = series resistance, L = series inductance, G = shunt conductance, C = shunt capacitance, and $\bar{\omega}$ = angular frequency, $2\pi f$.

Most transmission and distribution lines are three phase, with three conductors per circuit. Each conductor is represented by a pi-equivalent circuit, with mutual capacitance and inductance terms between each pair of conductors. For most studies of balanced three-phase circuits, it is sufficient to use a single pi-equivalent, as if it were single phase. The terms R, L, and C are determined by a matrix transformation, called the symmetrical component transformation, from the self and mutual parameters of the three phases (Bergen, 1986). The "positive-sequence" values from the transformation are used for the single-phase pi-circuit. Fault and other studies that involve unbalanced operation of the transmission line also require the zero-sequence equivalent, which has the same form except for zero-sequence values for the parameters taken from the matrix transformation.

The short-line approximation is valid for lines shorter than about 0.05–0.1 wavelength, or about 250–500 km at 60 Hz. For longer lines, the series and shunt branches use the characteristic impedance Z_c and propagation constant γ, defined as

$$Z_c = \sqrt{Z/Y}, \tag{3}$$

$$\gamma = \sqrt{ZY}. \tag{4}$$

For the long-line model, Z and Y in the pi-equivalent circuit of Fig. 3 are replaced by

$$Z' = Z_c \sinh \gamma l, \tag{5}$$

$$Y' = (2/Z_c) \tanh (\gamma l/2). \tag{6}$$

2.3.2 Parameters that Affect Resistance, Inductance, and Capacitance. The shunt conductance G represents power dissipation in the line insulation. For an overhead line, G only represents corona loss and losses in insulators and is totally negligible compared with shunt resistance in the loads. The shunt conductance is somewhat greater for the dielectric in underground cables but is still negligible on a system basis. The only time G is considered is when one is making calculations of the current-carrying capacity (ampacity) of underground cables, where heat dissipation in the dielectric can be significant.

The series resistance R is primarily a function of the conductor size, material, and stranding. The inductance L and capacitance C are functions of the conductor diameter and geometry. The calculation procedure for positive- and zero-sequence parameters involves construction of a 3×3 series matrix and a 3×3 shunt matrix of self and mutual phase impedances for the three-phase conductors. The 3×3 matrices are converted to positive- and zero-sequence impedances and admittances by the symmetrical component matrix transformation (Bergen, 1986). Practical transmission lines normally have shield wires and may have more than one conductor per phase, or a total of from 4 to 14 conductors in a single circuit. Phase matrices of order 8×8 are required for two shield wires and two conductor bundles per phase. The 8×8 matrix is reduced to an equivalent 3×3 matrix for transformation to symmetrical components. The matrix reduction preserves the effect of shield wires and conductor bundles on the system properties of the line but eliminates detailed knowledge of shield-wire and phase-subconductor currents.

Shield wires and conductor bundles affect the line parameters. For example, conductor bundling results in a significant reduction in series reactance. Shield wires primarily affect the zero-sequence parameters, decreasing the series reactance but increasing the resistance. Shield wires do slightly change the positive-sequence values, even though hand calculation formulas ignore their effect.

The resistivity of the earth has an effect on the line resistance and inductance, which is customarily represented by Carson's series (Olsen and Pankaskie, 1983). Because of the relaxation time of the earth, the earth resistivity has no effect on line capacitance. Numerical values of line parameters are generally computed today by line-parameter computer programs, eliminating the approximations that formerly went into hand calculations.

2.4 Surge Impedance and Surge-Impedance Loading

Several factors enter into the loading capability of a given transmission line, among which are conductor thermal capacity, voltage drop, and system stability. Conductor thermal capacity is a function of the conductor resistance, the maximum design temperature of the line, and weather conditions. Voltage drop and system stability are functions not only of line reactance but also of the system to which the line is connected.

A useful measure of loading limitation due to line reactance is given by the surge-impedance loading. Surge impedance Z_s of a three-phase transmission line is the characteristic impedance Z_c assuming a lossless line and defined as

$$Z_S = \sqrt{L/C}, \tag{7}$$

where L and C are given in positive-sequence henries and farads per unit length. The units of Z_S are ohms. When a three-phase line is loaded by a Y-connected resistance of Z_S ohms per phase, it is said to be surge-impedance loaded. This is equivalent to terminating the line in its characteristic impedance, resulting in a pure forward-traveling wave. Surge-impedance loading power P_{SIL} is given by

$$P_{\text{SIL}} = \frac{N_{\text{phases}} V^2_{\text{phase-ground}}}{Z_S}. \tag{8}$$

Surge-impedance loading (SIL) is that loading where the reactive power generated in the line capacitance equals the reactive power absorbed in the line inductance. SIL is a function of the line inductance per kilometer and capacitance per kilometer and is independent of line length. SIL is a handy tool for estimating the relative loading capabilities of overhead lines of different lengths, as shown in Fig. 4 (Dunlop *et al.*, 1979). While a transmission system would not be designed on the basis of Fig. 4, the limitation on practical loading of a line can easily be seen.

For an overhead transmission line, a typical surge impedance is on the order of 300 Ω, while for an underground cable, it may be 30 Ω or less. At 345 kV, SIL of an overhead

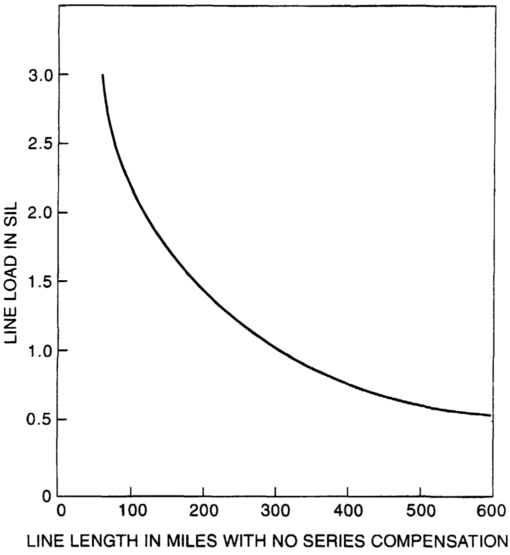

FIG. 4. Transmission-line loading in terms of surge-impedance loading.

line is approximately 400 MVA. Short overhead lines are often loaded at 2 or 2.5 times SIL and need reactive power support to maintain the voltage. Long overhead lines may be limited to 1.0 times SIL or less. SIL for a 345-kV underground cable may be on the order of 4000 MVA or an order of magnitude above the ampacity of an underground cable. Underground cables always operate well below SIL and appear as a large capacitance to the system. Differences in system behavior of overhead and underground transmission lines become apparent from an SIL analysis.

Surge-impedance loading is proportional to the square of the voltage for any transmission line configuration. If a specific line is limited by system considerations such as voltage drop or stability, there may be nothing to be gained by increasing the current rating of the line. In such a case, increasing the voltage may increase the power-handling capability of the line by the square of the factor by which the voltage is raised, assuming adequate ampacity. This variation of line loadability proportional to the square of the voltage is the driving force behind increasing transmission-system voltages.

3. POWER-TRANSMISSION SYSTEM CONSIDERATIONS

3.1 Power Control

The classical power flow equation (Stevenson, 1982) is

$$P = V_1 V_2 \sin \delta / X, \qquad (9)$$

where P = the power transfer, V_1 = the sending end voltage, V_2 = the receiving end voltage, X = the total reactance between V_1 and V_2, and δ = the power angle between V_1 and V_2.

Increasing the power flow, while keeping the voltage and reactance constant, causes the power angle to increase toward its theoretical maximum of 90°. If the power flow tries to increase further, the angle becomes greater than 90°, whereby the actual power transfer decreases and the system becomes unstable. If the voltages V_1 and V_2 are increased, the power transmitted increases by the square of the voltage, for the same reactance and power angle. For the same power flow, the power angle decreases with the higher voltage. The maximum power flow with a 90° power angle is proportional to the square of the voltage.

An operating electric utility system has multiple parallel paths, involving different circuits of different lengths and reactances and different voltage levels. For lines connecting any two buses, the bus voltages and the angle between the two voltages are the same. From Eq. (9), the power flow on each individual line is inversely proportional to the line reactance. This has two consequences:

- Power flow can be controlled only in a general overall manner by rescheduling generation. Generation changes do not change the relative flow along parallel paths.
- Power flow does not necessarily divide proportionally to the line loading capability. Particularly when an underground circuit parallels an overhead circuit, the underground circuit with the lower impedance takes a disproportionately large share of the power flow, especially when it is considered that the thermal rating (ampacity) is usually significantly less for an

underground cable than for an overhead circuit.

When it becomes necessary to control the flow on a specific circuit or group of circuits, several methods are possible: series capacitors, series inductors, or phase-shifting transformers. Series capacitors reduce the series reactance and increase power flow but have several significant drawbacks. Adding a series capacitance creates the possibility of subsynchronous resonances (Larsen *et al.*, 1990), which can cause considerable damage to generating plants and cause system stability problems. Protection of series capacitors from fault currents is difficult and costly. Series capacitors are generally only used where the reactance of very long overhead lines limits system performance (Allustiarti *et al.*, 1988).

Series inductors increase the series reactance and reduce power flow, but again, they have disadvantages of size, cost, and losses. Series inductors are generally only used when fault-current limitation is required.

The most commonly applied power control method is phase-shifting transformers, which change the power angle δ in Eq. (9). Phase-shifting transformers have the advantage of adjustment of phase angle, at the expense of cost and reliability (Bean *et al.*, 1959).

A method of power control under investigation is flexible ac transmission systems (FACTS) (Mountford *et al.*, 1993), in which thyristor technology is proposed as a means of making a solid-state phase regulator. The rapidity of phase-angle change possible with thyristors, combined with computer control, make this a promising approach for obtaining maximum possible power flow on limited transmission facilities.

3.2 Voltage and Reactive Power Control

Power in an ac circuit is given by

$$S = VI^* = P + jQ, \tag{10}$$

where S = apparent power in volt amperes, V = phasor (complex number) voltage in volts, I = phasor current in amps, P = real power in watts, Q = reactive power in vars, and the asterisk denotes complex conjugate. The difference between real and apparent power is caused by the phase-angle difference between voltage and current. The out-of-phase component, reactive power, represents flow of energy stored in inductive and capacitive elements in the system. Control of power flow on a power system involves the flow of both real and reactive power.

Changing the relative phase angle between two buses primarily changes power flow but has a secondary effect on the flow of reactive power. Small changes in bus-voltage magnitude have a significant effect on reactive power flow. Control of voltage within tight tolerances is necessary to meet equipment and customer requirements, while excessive var flow uses valuable circuit thermal capacity. Four methods are used to control voltage and secondarily reactive power flow: shunt capacitors, shunt reactors, synchronous condensers, and static var compensators.

The system effects of overhead transmission and distribution lines are dominated by their series inductance. Most power-system loads are inductive. The combination of an inductive load with an inductive source impedance leads to possible low-voltage conditions. Shunt capacitors are commonly applied to correct power factor, reduce var flow, and improve voltage. Capacitors may be permanently installed, switched by time, or switched by voltage control.

Shunt reactors are used on the highest voltage (500 or 765 kV) transmission lines to control overvoltages resulting from shunt capacitance. Shunt reactors are also common on underground transmission systems to control var flow problems resulting from the large cable capacitance.

Synchronous condensers are synchronous motors that drive no mechanical load but act as an adjustable shunt impedance that can be either capacitive (over-excited) or inductive (under-excited). The continuously variable voltage control together with the rotating inertia of the machine makes synchronous condensers attractive, but cost, durability, and maintenance problems are making them less common.

Static var compensators (SVCs) use thyristors and static impedances to mimic synchronous condensers as an adjustable shunt impedance for voltage and var control. Rapid, precise control make SVCs very attractive voltage control devices.

3.3 Protection

Power-system protection (also called power-system relaying) is essential for both lines and equipment. Protective relaying has three fundamental requirements: sensitivity, selectivity, and speed (El-Hawary, 1983). The protection must be sufficiently sensitive to detect any fault within its zone of protection, sufficiently selective to ignore faults outside its zone of protection, and sufficiently fast to remove the faulted component before damage is done to equipment or the system becomes unstable (see Sec. 3.4). Because of the importance of immediate fault removal, backup protection is provided in the event that primary relays fail to detect the fault or the circuit breaker fails to clear. A major concern in relaying is proper coordination between the relays so that only the proper relays operate in the pròper sequence. Types of relays used for power transmission include the following:

- Inverse time overcurrent relays. These respond to excessive current as a function of time. The relay responds in less time to higher currents, as shown in Fig. 5, hence the inverse time name. The time-versus-current response allows coordination between different relays.
- Distance relays. These respond to apparent impedance looking down a line. The typical characteristic of a distance relay is a circle, as shown in Fig. 6. The straight line represents impedance of the transmission line as a function of distance along the line. The relay circle is set to cross the impedance of the transmission line at about 90% of the length of the line. When the apparent impedance (voltage divided by current) seen by the relay enters the circle, the relay senses a fault and initiates a trip signal. Distance relays are more selective in their response than overcurrent relays and frequently are applied with three zones of protection with different circle sizes and time delays for backup purposes.
- Phase comparison relays. These and similar types compare current at two ends of the line to detect a fault. They are sensitive and selective, but allowance must be made in application for capacitive current on the line.

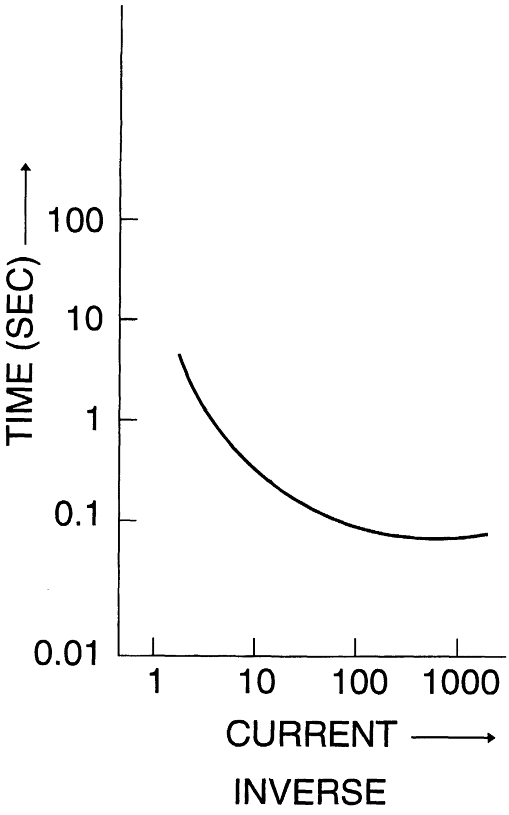

FIG. 5. Inverse time characteristic.

While it is possible to design a protection system without communication between the line ends, some form of communication is desirable for maximum protection. Phase comparison schemes require communication to compare the currents at both ends of the line. A distance relay transfer trip scheme may be used to initiate an immediate trip when the relays at both ends simultaneously see an indication. This communication may be power-line carrier, microwave, fiber optics, or leased telephone wire.

Protection must detect phase-to-phase and phase-to-ground faults. Sometimes a combination of relay types is used (e.g., distance relays for phase-to-phase and overcurrent for phase-to-ground). Other types of relays are often added—for example, directional relays to prevent tripping for faults on adjacent circuits. Where double circuit lines are involved, compensation is required to prevent ground faults on one circuit from tripping the other. System protection studies are based on load flow and fault computer pro-

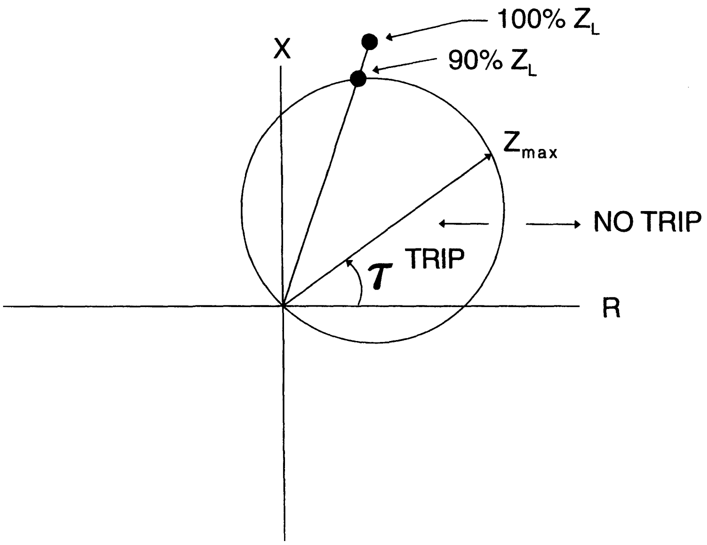

FIG. 6. Distance relay characteristic.

grams, which can model the system for normal and contingency conditions.

3.4 Stability

The classical power-flow Eq. (9) shows that any transmission circuit at its rated voltage has a maximum power flow, which cannot be exceeded without instability. This is an illustration of steady-state stability of a power system. A second stability problem is transient stability, which is a consideration whenever there is a fault or other switching of transmission lines or substation buses.

During a transmission-line fault, low system voltages instantaneously reduce power flow according to Eq. (9). The mechanical power input to the generators remains constant, but flows on the transmission lines are reduced. By conservation of energy, the mechanical power goes into stored energy in the system by speeding up generators and motors. At the same time, reduced system flows causes motors on the load side to slow down. The effect is to increase the angle δ in the milliseconds following the fault. The system is said to "swing." The detailed behavior is a function of the prefault power flow, the reactances of the transmission lines, inertia in rotating machinery, response of generation excitation and governors, and response of other equipment, such as static var compensators. If the fault is not cleared within 0.05 to 0.1 seconds, the system may swing so that δ exceeds 90°. When this happens, the system continues to swing until δ passes 180°, at which point the system conditions are mathematically identical to transmission-line faults. The system protection then properly operates, transmission lines are switched out, and the system separates into smaller isolated areas, referred to as islands. If the system separates, emergency measures must be taken by automated equipment to prevent a blackout.

Figure 7 shows an example of swing curves for a transmission-system fault where the system becomes unstable; the relative angle increases over time without bound. The two curves represent the rotor angles of two

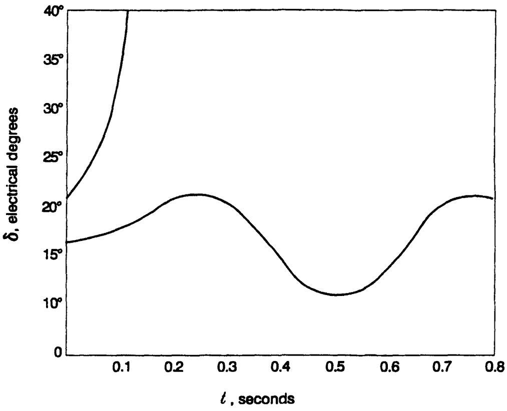

FIG. 7. Swing curves showing transient instability.

different generators relative to a common reference. One generator is speeding up with respect to the other, and the relative angle between them increases until the system separates. Because of the complexity of the transient stability problem, analysis is performed with transient stability computer programs, which represent the transmission system over a wide area and accurately model the dynamic performance of rotating machinery and response of loads to changing voltage.

3.5 Reliability

The discussion on stability points to the overall problem of power transmission reliability. Reliability involves minimization of faults and limitation of the effect of faults that do occur. Reliability considerations permeate the design of overhead and underground transmission lines, as described in Secs. 4 and 5.

Electric utilities have formed reliability councils to set reliability criteria for the bulk power system; for example, stability of the bulk power system shall be maintained for the following cases, including the effect of high-speed reclosing:

- permanent three-phase fault on any transmission circuit (as opposed to a transient fault such as a lightning trip, which causes no permanent damage);
- transformer or bus section cleared in normal time;
- simultaneous permanent phase-to-ground faults on different phases of each of two adjacent transmission circuits on a multiple circuit tower cleared in normal time; and
- permanent phase-to-ground fault on a transmission circuit with delayed clearing.

None of these events would cause a cascading failure and system separation. Reliability of individual lines is expressed in terms of outage rates, frequently in events per year. The dominant cause of transmission-line outages in the United States is lightning. Other common causes of outages are contamination on the insulator surface, trees or other objects contacting the line, and wind or ice storm damage.

3.6 Economics

Economics is a major factor in the design of transmission facilities. Selection of conductor size and type, structure type and height, foundations, span length, and other variables is a complex optimization process. The present worth of cost of losses and maintenance of the line (including all appropriate interest rates and other financial factors) must be balanced against the initial cost of construction for the overall least-cost design. Details of overhead and underground line components are given in Secs. 4 and 5. Some of the factors that enter into overhead transmission line economics are these:

- Larger conductor size results in lower losses but costs more initially and requires more massive costly structures to support.
- Longer span length requires fewer structures per kilometer, but individual structures are taller and more costly.
- Higher conductor tension results in less sag and structure height but increases mechanical loading on dead-end and angle structures.

Cost of right of way is an increasingly important economic consideration and is driving the quest for more compact lines with increased power-transfer capacity.

4. OVERHEAD ac TRANSMISSION LINES

4.1 Line Parameters

The transmission-line parameters, as shown in Sec. 2.3, may be divided into two parts: series impedance and shunt admittance. The series impedance consists of a resistance and an inductive reactance, while the shunt admittance consists of a conductance (inverse of resistance) and a capacitive reactance.

Skin effect is inherent in the conductor itself; hence when the ac resistance of a conductor is stated, what is meant is the dc resistance plus an increment that reflects the increased apparent resistance of the conductor caused solely by the skin-effect inequality of current density. Skin effect results in a decrease of current density toward the center of the cylindrical conductor (the current tends to crowd to the surface).

The per kilometer resistance, reactance, and capacitance can be determined from conductor characteristic data found in handbooks and manufacturers' literature (Aluminum Association, 1982; EPRI, 1982) and the spacing factors X_d and X'_d. The spacing factors X_d and X'_d represent the inductive reactance and capacitive reactance per conductor at a 0.3-m (1-ft) spacing, respectively. The positive-sequence resistance is listed as the 60-Hz value at 50 °C. The expression for inductive reactance per kilometer is

$$X_L = 0.001419 f \log (D/\mathrm{GMR}), \tag{11}$$

where D is the geometric mean of the three distances between phase conductors in meters, GMR is the geometric mean radius of the phase conductor in meters (as given in the conductor tables), and f is the frequency in hertz. GMR for ACSR conductor is given at 60 Hz; however, 60-Hz values of GMR can be used at other commercial power-system frequencies with small error. The expression for capacitive shunt reactance per kilometer is

$$X_c = \frac{13.45 \times 10^6}{f} \log \frac{D}{r_c}, \tag{12}$$

where r_c is the conductor radius in meters.

4.2 Conductor Types and Selection

A conductor of electricity is any substance or material that will afford continuous passage to an electric current when subjected to a difference of electric potential. The greater the density of current for a given potential difference, the more efficient the conductor.

Electrical conductors are manufactured in a variety of forms and shapes for various purposes. These may be wires, cables, flat straps, square or rectangular bars, angles, channels, or special designs for particular requirements. The most extensive use of conductors, however, is in the form of round solid wires and stranded conductors. Materials commonly used in conductors are aluminum, copper, and steel. Galvanized steel wires are combined with aluminum in the most common type of overhead conductor—aluminum conductor, steel reinforced (ACSR). The use of copper is uncommon on modern transmission lines since it usually costs and weighs considerably more than aluminum conductor of the same resistance.

Standard bare overhead conductors consist of round strands helically laid about a core in one or more layers. In a homogeneous conductor—all-aluminum conductor (AAC), hard-drawn copper conductor (CU), or all-aluminum alloy conductors (AAAC5005 or AAAC6201)—the core consists of a single strand identical to the outer strands. In a nonhomogeneous conductor—aluminum conductor, steel reinforced (ACSR), aluminum conductor, alumoweld steel reinforced (ACSR/AW), hard-drawn copper conductor, copperweld steel reinforced (CU/CW), or aluminum conductor, aluminum alloy reinforced (ACAR)—the strands in the core may or may not be of the same diameter.

In addition, there are several nonstandard overhead conductors for special-purpose applications. Self-damping ACSR conductors (ACSR/SD) are used to limit aeolian vibration to safe levels regardless of conductor tension or span length. ACSR/TW is similar to ACSR in its use of trapezoidal-shaped strands but does not have annular gaps between layers. ACSR/TW has a smaller diameter and smoother surface than conventional round-wire ACSR of the same area and thus may have reduced wind loading. Steel-supported aluminum conductors (SSAC) are similar to conventional ACSR but employ annealed aluminum alloy strands on the outside layers. The annealed aluminum has increased electrical conductivity, and the conductor has improved sag–tension characteristics for high-temperature service. T2 conductors are fabricated by twisting two conventional conductors together. Severity of wind-induced galloping when the conductor is coated with ice is reduced with T2 conductors. In conjunction with advanced line design techniques, one can consider the value of using the novel conductors for lower electrical resistance, lower thermal elongation, higher annealing temperatures, and reduced aeolian vibration.

The choice of a conductor for a transmission line, as with structure type, depends on the specific application. Once the mechanical strength requirement of the conductor is satisfied, the conductor choice considers the total costs associated with the conductor, and corona-related electrical environmental effects of radio and audible noise. Costs associated with the conductor include the cost of the conductor, support structures, and present worth of the future power losses and maintenance. Corona is a result of the electrical stress on the surface of the conductor and is a function of the voltage on the conductor, diameter and surface condition of the conductor, and the spacing between conductors and/or grounded objects. Corona also causes power loss, particularly during wet weather.

The equivalent size of a conductor can be increased either by using a larger single conductor or several smaller conductors electrically and physically connected together (bundled conductors). Typical spacing between parallel conductors for a two-conductor bundle is 47 cm. Compared with a single very large conductor, a bundle of several smaller conductors results in lower series reactance and significantly less corona and associated noise. Bundled conductors are necessary at the higher transmission voltages to limit corona to acceptable levels. They are sometimes used at lower transmission voltages to reduce reactance and provide for greater current capability.

At voltages of 345 kV and above, the minimum conductor size or the minimum number of conductors in a bundle is normally determined by the corona-related electrical environmental effects. At voltages below 345 kV, the minimum size is normally based only on conductor economics.

Since a limiting design criterion is minimum conductor height above ground (for safety reasons), the maximum sags during operation determine structure heights in terms of span lengths. Similarly in certain areas, ice can form on the conductors of sufficient weight to limit the structure heights and span lengths to maintain ground clearance. The conductor sag in the span between structures will depend on conductor materials, conductor weight, conductor strength, conductor tension, conductor temperature, and ice accumulation on the conductor. As the current in the conductor increases, the losses increase with a resultant increase in conductor temperature, causing the sag to increase. If the conductor is carrying heavy electrical load on a hot day, very significant increases in sag can occur.

4.3 Insulation Designs and Considerations

The electrical operating performance and reliability of a transmission line depend primarily on the ability of the insulation to withstand the voltage under all operating conditions. An insulator not only must have sufficient mechanical strength to support the greatest loads of ice and wind that may be reasonably expected, with an ample margin, but must be so designed as to withstand severe mechanical abuse, lightning, and power arcs without mechanically failing. It must prevent a flashover for practically any power-frequency operating condition and many transient voltage conditions, under any conditions of humidity, temperature, rain, or snow and with such accumulations of dirt, salt, and other contaminants as are not periodically washed off by rain (Kaminski, 1963).

The majority of present insulators are made of glazed porcelain. Porcelain is a ceramic product formed by the high-temperature vitrification of clay, finely ground feldspar, and silica. Insulators of high-grade electrical porcelain of the proper chemical composition free from laminations, holes, and cooling stresses have been available for many years. The insulator glaze seals the porcelain surface and is usually dark brown, but other colors such as gray and blue are used. Porcelain insulators for transmission may be disks, posts, or long rod types. Porcelain insulators have been used at all transmission-line voltages and, if correctly manufactured and applied, have high reliability.

Glass insulators have been used on a significant number of transmission lines, particularly in Europe. These are made from toughened glass and are usually clear and colorless or light green. For transmission voltages, they are available only as disk types. Most glass disk insulators will shatter when damaged but without mechanically releasing the conductor. This provides a simple method of inspection.

Synthetic insulators, originally pioneered by the General Electric Company in 1963 for high-voltage transmission lines (Nichols and Vose, 1963) and subsequently introduced by several manufacturers, are finding increasing acceptance. Most consist of a fiberglass rod covered by weather sheds of skirts of polymer (silicon rubber, polytetrafluoroethylene, cycloaliphatic resin) (Abilgaard *et al.*, 1976). Improvements in design and manufacture have made synthetic insulators increasingly attractive since their strength-to-weight ratio is significantly higher than that of porcelain and can result in reduced tower costs, especially on EHV and UHV transmission lines. Polymer insulators are usually manufactured as long rod or post types. The light weight of most designs and resistance to damage aids construction. In addition, their performance under contaminated conditions may be significantly better than that of porcelain (Karady and Lamontagne, 1976). It has been found necessary to use grading rings on some types at higher voltages to prevent damage to the sheds, and a very small number of insulators have experienced "brittle fractures," in which the fiberglass core breaks close to the end fitting. Despite these problems, it appears that reliable synthetic insulators are presently available.

The criteria for power-frequency insulation design is usually that flashover shall not occur for normal operating conditions, including reduced clearance to the structure from high wind. A typical wind design limit is the 50- or 100-year return period wind, i.e., wind velocity that occurs only once in 50 or 100 years. This velocity is obtained from local wind records and may be typically 130 to 160 km/h. In clean conditions, power-frequency voltage is not a controlling parameter for insulator design (as distinct from air-gap clearance). Design for contamination is usually expressed as centimeters of creepage per kilovolt, where the creepage distance is the length of the shortest path for a current over the insulator surface, and ranges up to 5 cm per phase-to-ground kilovolt or more for heavy contamination. Standard insulator disks (25.4 × 13 cm) have a typical creepage length of 29.2 cm per disk. To avoid very long insulator strings for contamination, disks with additional creepage distance are made. The creepage can be extended by use of lengthened skirts and deeper grooves in the underside. Fog type disks have up to 54.6 cm of creepage per 34.3 × 20.3-cm disk.

Operation of a circuit breaker on a transmission line can cause transient overvoltages, although flashovers due to such switching surges are rare in conventionally spaced lines below 500 kV. If the breaker is opening, this may be due to restrikes across the breaker contacts as they separate, although restriking has been nearly eliminated with present breaker technology. If the breaker is closing, the cause may be unequal voltage on the two sides of the breaker, including the effect of residual charge on the line from recent de-energization. The crest magnitudes of switching surges are normally defined in per unit of nominal power-frequency crest phase-to-ground voltage. The criterion for switching surge insulation design is usually that flashover shall not occur for most or all switching events.

Impulse surges on a line are caused by lightning strokes to or near the line. At transmission insulation levels, only strokes that directly intercept the line are capable of causing flashovers. It is unusual for line insulation to be determined by lightning performance alone. More typically, insulation is determined by other requirements, and the lightning performance is then verified. If this performance is unsatisfactory, it is often more efficient to change other design parameters such as shield wires or grounding than to add additional insulation.

The controlling paths for flashovers on most presently installed transmission lines are phase to ground, since there are usually grounded structure components between phases. However, compact transmission lines that do not utilize grounded support structures between the phases can employ phase-to-phase insulators. Design methods for phase-to-phase insulation are essentially the same as for phase-to-ground insulation (EPRI, 1978, 1980; Souchereau *et al.*, 1978).

4.4 Ampacity

Power lines are limited in their ability to deliver electricity. The two most important limits can be understood by considering thermal effects and stability. When current flows in the conductors, there are I^2R losses and heat is generated. This loss of power reduces transmission efficiency. The heat generated by the line losses causes a temperature rise. Since line temperature of overhead lines must be kept within safe limits to prevent excessive line sag between transmission towers (minimum ground clearances must be maintained) and to prevent irreversible stretching (at about 100 °C), this imposes limitations on the maximum safe current a line can carry. Clearly, the current-carrying capability of the line depends on the line design (conductor size, phase geometry, and spacing) and operating conditions (ambient temperature, wind velocity).

Thermal considerations impose a current-handling limitation for a given line, and, hence, lines are given current ratings that provide for safe operation. Operating the line above the rating for an extended period of time is not recommended. The thermal limitations of the terminating equipment (i.e., transformers) may be more restrictive than the ratings pertaining to the line itself.

4.5 Electrical Environmental Considerations

There are two categories of electrical environmental effects of power transmission lines. Corona effects are those caused by electrical stresses at the conductor surface that result in air ionization ("corona") and include radio, television, and audible noise. Field effects are those caused by induction to objects in proximity to the line. While the generic term is electromagnetic effects, within the electric power industry the fields are divided into two types: electric field effects and magnetic field effects. Electric fields, related to the voltage of the line, are the primary cause of induction to vehicles, buildings, and objects of comparable size. Magnetic fields, related to the current in the line, are the primary cause of induction to long objects, such as fences and pipelines.

In an electrical environmental analysis, it is important to determine the proper criteria for assessment of the impact. For example, the audible noise criterion in a commercial or industrial area would be inappropriate in a quiet residential neighborhood (Keast, 1980). Likewise, ground-level electric field criteria on a parking lot would be much more strict than these in terrain inaccessible by motor vehicles. For audible noise, the only concern is annoyance, but for electric fields, safety, annoyance, and perception levels all may need to be considered.

Corona-produced audible noise during foul weather, particularly during or following rain, can be an important design parameter for high-voltage ac transmission lines. Audible noise has two components, a random noise component and a low-frequency hum, each produced by different physical mechanisms. While the hum component is closely correlated with corona loss on the line, the random noise is not. Of these two, the most frequent cause of annoyance is the random noise, and it is this that is calculated and compared with acceptance criteria. Analyses to predict levels of audible nose consider A-weighted sound level [dB(A)] during rain including

- L_{50}, which is the level exceeded 50% of the time during rain,
- L_5, which is the level exceeded 5% of the time during rain,
- average, which is the average level of noise expected during rain, and
- heavy rain, which is the level expected during heavy rain.

One formula for calculating L_5 is given by (IEEE Task Force, 1982)

$$\begin{aligned} AN_5 = {} & -665/g + 20 \log n + 44 \log d \\ & - 10 \log D - 0.02D + AN_0 \\ & + K_1 + K_2, \end{aligned} \tag{13}$$

where g = average maximum surface gradient of conductor, kV/cm; n = number of subconductors in a phase bundle; d = diameter of subconductors, cm; D = distance from line to point at which noise level is to be calculated, m; $AN_0 = 75.2$ for $n < 3$, 67.9 for $n \geq 3$; $K_1 = 7.5$ for $n = 1$, 2.6 for $n = 2$, 0 for $n \geq 3$; $K_2 = 0$ for $n < 3$, $22.9(n - 1)d/B$ for $n \geq 3$ (B is bundle diameter in cm). The L_{50} level for each phase is obtained from

$$AN_{50} = AN_5 - \Delta A, \tag{14}$$

where $\Delta A = 14.2g_c/g - 8.2$ for $n < 3$, $14.2g_c/g - 10.4 - 8[(n - 1)d/B]$ for $n \geq 3$; $g_c = 24.4d^{-0.24}$ for $n \leq 8$, $24.4d^{-0.24} - 0.25(n - 8)$ for $n > 8$.

Radio and television interference from overhead power lines is caused by two phenomena: complete electrical discharges across small gaps (microsparks) and partial electrical discharges (corona). Gap-type sources occur at insulators, line hardware, and defective equipment and are a construction and maintenance problem rather than a design consideration. They are responsible for about 90% of noise complaints and can be located and eliminated as they occur (Loftness, 1992). Conductor and hardware corona is considered during the design phase of the line. On a properly designed line, conductor corona rarely results in complaints except perhaps in weak-signal fringe areas.

The specification of corona-free hardware is important to eliminate that source of electromagnetic interference and is especially important as lines are constructed with closer spacings and resulting higher electric fields on the hardware. Conductor clamps and other fittings that were formerly acceptable at traditional phase spacing may not be adequate for compact lines.

Two families of computation methods are available for radio noise: those based on conductor laboratory tests and analytical propagation theory (semianalytical methods) and those based on empirical formulas using data from long-term tests on operating lines (comparative methods). The comparative method is useful for conventional geometries and designs (Chartier, 1981):

$$RI = -150.4 + 120\log g + 40\log d + 20\log\left(\frac{h}{D^2}\right) + 10[1 - (\log 10f)^2], \tag{15}$$

where g = average maximum surface gradient of conductor or conductor bundle, $\mathrm{kV_{peak}/cm}$; d = subconductor diameter, mm; h = height of phase, m; D = radial distance to observer, m; f = frequency, MHz; and RI = fair-weather radio noise, dB. The term RI is calculated for each phase, and the maximum value is used as the RI of the line. There are no generally accepted RI limits in the United States, because of the impossibility of setting universal criteria for all land-use and local conditions. A Canadian standard exists for RI limits and is a useful guide (CSA, 1975).

Ground-level electric field effects of overhead power transmission lines relate to the possibility of exposure to electric discharges from objects in the field of the line. These may be steady currents or spark discharges. Other areas that have received attention are the possibility of fuel ignition and interference with individuals wearing pacemakers. The resultant electric fields in proximity to a transmission line are the superposition of electric fields due to the three phase conductors. The conducting earth must be represented by image charges located below the conductors at a depth equal to the conductor height. For a single conductor, the total ground-level electric field is given by

$$E = \frac{Q_1}{\pi\epsilon}\frac{h}{h^2 + L^2} \tag{16}$$

where Q_1 = charge on each conductor as a function of the voltage on all conductors, ϵ = permittivity, h = height of conductor above ground, and L = lateral distance to the measuring point. At ground level, the horizontal components of the electric fields of the conductor and its image cancel, and the resultant field is purely vertical. For a three-phase line, the fields of the three conductors and their images are computed separately and added.

Magnetic field coupling affects objects that parallel the line for a distance, such as fences and pipelines, and is generally negligible for vehicle- or building-sized objects. As opposed to electric field coupling, magnetic field coupling is a low-voltage, low-impedance source with relatively high short-circuit currents. Single grounds are ineffective in preventing magnetically coupled voltages, and multiple low-resistance grounds are needed. A consideration in the calculation of magnetic fields that is different from electric field calculation concerns the images. A perfectly conducting earth can be assumed for the electric field problem, even for realistic values of earth resistivity. The assumption of a transmission line in free space (no earth at all) gives a closer approximation to the ground-level magnetic fields than does the assumption of a perfectly conducting earth for measurements near the line. For a single conductor, without its image, the magnetic field may be represented by

$$B = \frac{\mu I}{2\pi r} = \frac{\mu I}{2\pi\sqrt{h^2 + L^2}}, \tag{17}$$

where I is the current in the single conductor and μ is the permeability. The magnetic field is then separated into vertical and horizontal components by multiplying by $\sin\theta$ and $\cos\theta$, where θ is the angle between the observer and the phase conductor. In general, both components must be retained. For a three-phase line, all conductors must be computed. Horizontal and vertical components of the magnetic field from the three conductors must then be combined individually as phasors, considering the angles of the different currents. The combined horizontal and vertical components in general have different time angles, causing their resultant to trace an ellipse in time. Single-axis magnetic field meters with the sensing coil oriented for a maximum reading give the magnitude of the semimajor axis of the field ellipse. A three-axis meter computes the square root of the sum of the squares of the three orthogonal field components (the "resultant" field). The resultant field can be as much as 41% greater than the major axis of the field ellipse for a circularly polarized field that results from symmetrical line configurations.

4.6 Support Structures: Mechanical and Civil Engineering Considerations

Numerous types of structures are used for supporting transmission-line conductors—for example, self-supporting steel towers, guyed steel towers, self-supporting aluminum towers, guyed aluminum towers, self-supporting steel poles, flexible and semiflexible steel towers and poles, rope suspension, wood poles, wood H frames, and concrete poles. The type of supporting structure to use depends upon such factors as the location of the line, importance of the line, desired life of the line, money available for initial investment, cost of maintenance, and availability of material. A line built with self-supporting steel towers is very satisfactory in all respects, as it requires less inspection and has a maximum life with minimum maintenance costs. Self-supporting steel poles are frequently used in congested districts where right of way is limited and short spans are necessary.

In determining the maximum forces acting on towers or poles, it is necessary to combine the transverse forces, longitudinal forces (including torsion), and vertical forces so that they act simultaneously. Other factors that may enter into the determination of the maximum forces acting on supporting structures include the horizontal and vertical components of tensions in guy wires and the addition of pole-top transformers, switches, and working platforms.

Stresses in towers can be computed analytically by the historic graphical methods or by use of one of the available computer programs for tower analysis and design. Most of the design procedures assume that the foundations are rigid. In actual practice, an uneven settlement of the foundations may produce excess stresses that must be considered and taken care of in the overload factors provided against failure. Some flexible lattice-type towers and most pole-type structures will undergo sufficiently large horizontal displacements that the nonlinear stresses due to the vertical loads acting in conjunction with these displacements must be considered.

5. UNDERGROUND ac TRANSMISSION LINES

5.1 Line Parameters

The electrical model of an underground cable is the pi-equivalent circuit described in Sec. 2. Several differences mark underground cable as compared to overhead lines:

- R is lower for an underground cable, because of the necessity to limit losses to keep the insulation temperature within limits. Conductor diameter is generally greater for an underground cable than for an overhead line, increasing the cost.
- L is lower for an underground cable, because of the smaller phase spacing.
- G is higher for an underground cable, because of dielectric losses in the solid and liquid insulation. G is not significant for the overall electrical system, but is important for considering the losses and temperature rise of the cable.
- C is higher for an underground cable, because of the higher dielectric constant of the insulation and the small spacing between the conductor and sheath.

The lower inductance and higher capacitance of an underground cable have significant system implications, as subsequently discussed in Sec. 5.6.

5.2 Types of Underground Installations

Four cable systems can be considered for application at transmission voltage levels above 69 kV (EPRI, 1992):

- High-pressure pipe-type cable; paper insulated, fluid or nitrogen filled, and pressurized to 1400 kPa (fluid filled above 135 kV). Available to 550 kV, but commercial experience to date is limited to 345 kV. This type of cable is dominant in the United States. The complete cable system includes cables, pipes, splices, terminals, oil, pumping stations, cooling stations, numerous valves, etc. The pipe is first laid in the trench, then the cables pulled in and spliced, and the pipe is filled with oil and pressurized. Distances between splices range from 460 to 1220 m.
- Self-contained fluid-filled cable, also called *low-pressure fluid-filled cable;* available in voltages through 550 kV. Instead of the pipe, the individual phase cables are sheathed with lead or aluminum. This system has evolved from the original ionization-free design developed in Italy in the 1920s and dominates underground usage in Europe, Asia, and South America.
- Extruded (solid) dielectric cable systems with polyethylene (PE), cross-linked polyethylene (XLPE), or ethylene propylene rubber (EPR) insulation.
- Compressed SF_6 gas-insulated rigid cable; available up to 500 kV for very high power levels and short (300 m) circuits. This type is an adaption of isolated phase bus developed for substations and the rigid coaxial line used for high-power UHF antenna feeds.

5.3 Insulation Considerations

At voltages above about 1.5 kV (the normal threshold of ionization), cable insulation systems fall into two categories: ionization-proof insulating materials (porcelain and other inorganic insulation such as silicone rubber and some epoxy resins) or ionization-tolerant insulations (paper-oil, varnished cloth, etc.), and ionization-free designs (XLPE cables) or ionization-suppression dielectric strength enhancement techniques (oil or gas pressurization). A power transmission cable is more complex than simple insulation over the conducting material. It is necessary to protect the insulation from water and mechanical damage, and it is especially important to grade the electric field in the insulation properly to prevent localized high fields and corona. Components of a power transmission cable (see Fig. 8) are the following:

- Conductor, aluminum or copper, stranded in segments carefully designed to minimize electrical losses to carry the load current.
- Conductor shield, a conducting or semiconducting material depending on the cable type to round out the strand conductor's irregular surface by filling or bridging the interstices between strands or segments. The purpose of the conductor shield is to make as nearly as cylindrical an electrode as possible to eliminate localized high electric field spots. A second purpose is to eliminate ionizable voids between conductor and insulation and reduce the intensity of any discharge that may occur.
- Insulation, the main insulating material to withstand the cable's voltage.

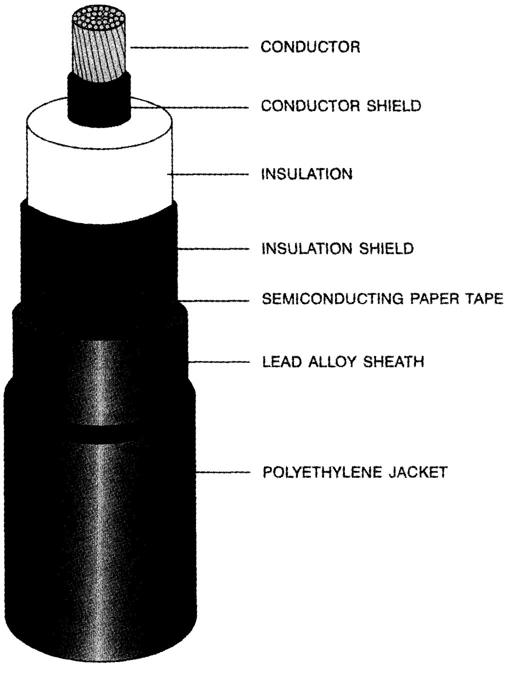

FIG. 8. Cutaway view of underground power cable.

- Insulation shield, a conducting layer over the cable with three functions: to ensure a radial electric field in the cable insulation, to carry the cable's capacitive charging current continuously, and to carry fault current.
- Sheath, in self-contained paper-insulated cables, an impermeable lead or aluminum cover to keep out water.
- Jacket, a coating layer to provide mechanical protection to the cable, additional barrier against entrance of chemicals or moisture, and sometimes a measure of electrical insulation.
- Skid wire, on cables to be installed in a pipe, a spiral-wrapped stainless steel wire on top of the sheath to protect the cable as it is pulled into the pipe and to reduce the friction during the pull.

In addition to the cable insulation, careful attention must be paid to attachments, splices, and terminations. A cable is terminated in an insulating bushing called a pothead whose function is to grade the electrical stress to prevent flashover across the outside of the insulation. If the insulating covering were merely stripped off without attention to the electric field distribution, surface flashovers could occur. Terminations have skirts to extend the leakage distance, much like the insulators on an overhead transmission line.

5.4 Cable Ampacity

An overhead electrical conductor dissipates heat directly to the adjacent air. Heat generation in an underground cable must pass through the thermal resistance of the cable insulation, pipe (if any), air space and duct (if any), and earth before being released to the atmosphere. The thermal circuit of an underground cable is significantly more complicated than for an overhead line, and the ability to dissipate heat is correspondingly reduced for an underground cable. Conversely, the thermal time constant for an underground cable is much longer than for an overhead line, allowing thermal storage to be a consideration in setting emergency loading levels. Cable losses and sources of heat fall in three categories:

- Conductor, proportional to I^2R. Shield/sheath losses are proportional to I^2R for current and resistance of the shield. Conductor losses are reduced by increasing conductor diameter, and properly stranding and segmenting the conductor. Shield/sheath losses are sometimes reduced in a three-phase cable by cross-bonding, that is, insulating the three shields from each other and transposing the connections to them along the cable length to cancel inductively coupled circulating currents.
- Dielectric, proportional to $V^2C \tan \delta$, where $C \tan \delta$ is the insulation loss factor.
- Magnetic, a function of the permeability of the pipe (if any).

Thermal resistances are the cable insulation itself, the insulating fluid (if any), the pipe and its coating (if any), and the earth (this is the dominant factor). An electrical analog of the thermal circuit of a high-pressure fluid-filled cable is given in Fig. 9. Forced cooling of the circulated insulating fluid in pipe-type cables has been employed to increase heat dissipation by controlling the ambient temperature of the cable and by lowering the thermal resistance. Additional losses are incurred by the circulating pumps and mechanical cooling equipment.

The classic reference that is still the basis of cable ampacity calculations is Neher and

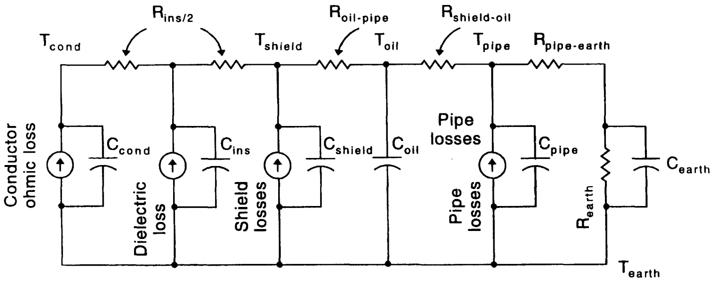

FIG. 9. Electrical analog of high-pressure fluid-filled cable thermal circuit.

McGrath (1957). This work has been extended for larger cable sizes (Katz *et al.*, 1978) and for self-contained cables (IEC, 1969).

5.5 Environmental Considerations

Underground power transmission involves several environmental considerations: opening a trench for installation and maintenance, and the location, repair, and consequences of leaks of insulating fluid. Because of the cable shield, there is no external electric field from an underground cable, and so electric fields are not an environmental factor.

Earth is not a shield for magnetic fields from underground cables. Figure 10 shows a magnetic field lateral profile for a solid dielectric cable compared with an overhead line of the same voltage and current. (Magnetic field at 60 Hz is a function of line current, not voltage. Voltage is included in this comparison to ensure comparable transmission capability.) The magnetic field of the underground cable falls off with distance more rapidly than the magnetic field of the overhead line, but the maximum values closest to the lines are comparable. The underground circuit's phase conductors have closer spacing and, hence, greater field cancellation. However, the closest distance from the observer to the underground cable is closer than that to the overhead line. These relative distances account for the difference in profiles.

The magnetic field profile of an underground circuit depends on several variables: cable geometry, sheath connections, and effects of the cable pipe. The geometry depends on the installation, whether direct buried side by side, in ducts, constrained by spacers, or in a pipe. Generally, the closer and more symmetrical the arrangement, the lower the magnetic field. When sheaths are cross-bonded to reduce losses and improve ampacity, they have no circulating current and do not affect the magnetic field profile. When the sheaths are multigrounded, allowing inductive circulating current to flow, the net magnetic field is reduced by the amount of the magnetic field from the sheath currents. Cables installed in a pipe are constrained in their geometry, have magnetic shielding from the ferromagnetic behavior of the pipe, and have shielding from the circulating currents induced in the pipe. The overall effect is difficult to predict exactly, but generally, the magnetic field of a pipe-type cable is a factor of 25 below that of a comparable cable of the same geometry without the pipe.

5.6 Specific System Considerations

System considerations of underground cables mainly result from the large capacitance compared to an overhead line. The overhead system mainly appears as a series inductance, while an underground cable system is dominated by the shunt capacitance. A capacitive current fed from an inductive source results in a voltage rise, as shown in Fig. 11. This voltage rise is significant, especially un-

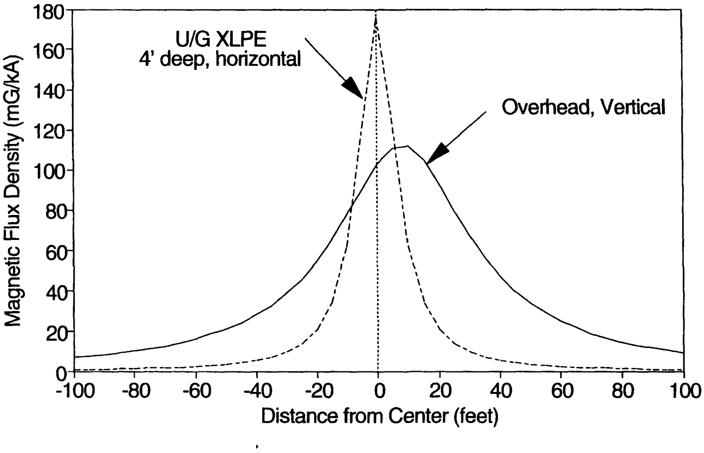

FIG. 10. Comparison of magnetic fields produced by overhead (solid line) and underground (dashed line) 230-kV transmission lines.

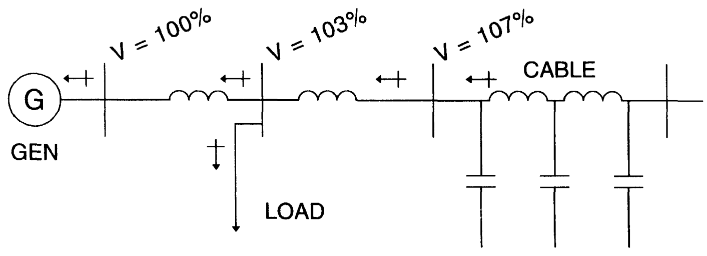

FIG. 11. Voltage rise due to cable capacitance.

der two operating conditions: light load hours when system load current is small and excessive voltage is possible and abnormal situations such as restoration of power following a blackout. During restoration, the supply system may be "weak," that is, high reactance. Energizing cable circuits with a weak source can result in overvoltages. Both problems are serious for utilities that have extensive underground transmission.

There is a fundamental limit to the length of a cable circuit. Beyond a certain length, the entire ampacity of the cable is taken by the capacitive charging current. At shorter distances, capacitive current may impair the ability of the cable to carry load current.

Overvoltages and capacitive current limitations are controlled by addition of shunt reactors to absorb the capacitive current. Placement and sizing of shunt reactors must be done in conjunction with a digital-computer power-flow study to ensure proper voltage and var control under all operating conditions. Switching and system restoration considerations may dictate more smaller reactors, sometimes switched with the cables. The comparatively low series reactance of underground cables presents a problem of distribution of power in parallel paths with overhead circuits. Section 3 addresses practical matters of real and reactive power control.

Overvoltage and surge protection is especially important for underground cable applications. Where a flashover on an overhead line may disrupt the line operation but causes no permanent damage, a fault on an underground cable damages the insulation and may require a considerable time (days or more) to repair. Thus, it is necessary to provide surge arrester and lightning protection sufficient to prevent surge damage to the cables.

Reliability of underground transmission is a misunderstood topic. Frequently it is assumed that reliability is improved by placing transmission lines underground, out of sight. While it is true that underground cables have fewer tripouts per year than overhead lines, when a failure does occur on an underground cable, the repair time is sufficiently long that a prolonged outage may result. For this reason, underground cable systems are frequently designed with more redundancy (another cost factor) than overhead lines.

6. HIGH-VOLTAGE DIRECT-CURRENT TRANSMISSION

High-voltage direct-current (HVDC) transmission has a number of significant benefits, including lower costs for long distances, aid in system stability during disturbances, controllable power transfer, and more power transfer across a given right of way than alternating current. HVDC overhead transmission has been in commercial operation in the United States since 1969 when an 800-mile, ±400 kV line was energized. This line, the "Pacific Intertie," runs from The Dalles in Oregon to near Los Angeles in California.

6.1 Line Parameters

dc lines may be overhead or underground. The electrical model of an HVDC line is the pi-equivalent circuit described in Sec. 2. Several differences mark dc lines as compared to ac lines:

- R modeling is required for steady-state and transient performance
- L modeling is needed for the transient analysis.
- C modeling will demonstrate the very high-frequency characteristics during transient conditions. The capacitance also be-

comes important in underground dc cables, as discussed in Sec. 5.

6.2 Types of dc Installations

In HVDC transmission, ground return may be used as one conductor. Based on this fundamental principle, the following basic circuit arrangements may be considered: monopolar, bipolar, or homopolar. With monopolar transmission, only one transmission pole (or "phase") is installed. Ground return is permanently used. Monopolar transmission is used in systems of comparatively low power rating, primarily with underground or underwater transmission. A bipolar dc transmission line consists of two poles: positive and negative. A ground return is not necessary in bipolar transmission, but is normally installed to provide additional reliability in case one pole fails, and the line will temporarily operate as a monopolar line. A homopolar dc transmission lines consists of two poles with the same polarity. This is applicable for an overhead line feeding a monopolar cable.

HVDC transmission systems can be classified into three major categories: back-to-back systems; two-terminal, or point-to-point, systems; and multiterminal systems. In back-to-back HVDC systems, the rectifier (conversion of ac to dc) and the inverter (conversion of dc to ac) are located in the same substation and usually in the same building. The rectifier and inverter are normally joined through a reactor. A back-to-back dc system is used to connect two asynchronous ac systems (e.g., one system operating at 50 Hz and one operating at 60 Hz, or two ac systems operating at the same frequency but not synchronized).

Two-terminal dc systems can be either bipolar or monopolar. A bipolar configuration is commonly used with overhead lines and consists of two distinct "phase" locations, one for each polarity (positive and negative), which carry equal currents. A monopolar dc system will have only one "phase," either positive or negative polarity, utilizing the earth or another metallic conductor as the return path. The monopolar, ground-return configuration has been used for underwater dc installations where the return path is the water. In a monopolar, metallic return system, the return current flows through the metallic return conductor. The monopolar mode of operation may be used on a bipolar installation under contingency conditions, such as a partial converter failure.

Multiterminal dc installations may operate in either parallel or serial configuration. Parallel configurations can be either radial or mesh where all converters operate at the same dc voltage, similar to ac installations. In the parallel operation mode, one converter determines the operating voltage, while the other converters control the current flow. In a series operation of multiterminal dc lines, all converters operate at the same current where one converter sets the current, while the other converters operate in a voltage control mode. The power output of the noncurrent-controlling converter is varied by adjusting the voltage.

6.3 Converter Stations

Since generation and most of the transmission and distribution systems are ac, HVDC transmission requires conversion from ac to dc (rectification) and from dc to ac (inversion). Rectification and inversion are accomplished by converter stations. The converter terminal comprises the following major components: solid-state valves and controls, converter transformer, reactors, filters, and reactive power supplies.

A three-phase, full-wave bridge converter (Graetz circuit or three-phase, six-pulse converter) may be used for conversion from ac to dc and from dc to ac. The dc output voltage of a six-pulse bridge contains harmonics of orders multiples of six, i.e., 6, 12, 18, . . . , $6n$, where $n = 1, 2, 3, \ldots$. Unless the harmonic voltages are filtered, telephone interference along the dc line may result. The ac current from a six-pulse bridge contains current harmonics of orders 5, 7, 11, 13, 17, 19, . . . , $6n \pm 1$. If the current harmonics are not filtered, voltage distortion and telephone interference will result on the ac system.

A three-phase, 12-pulse converter that is composed of two three-phase, six-pulse converters may also be used for conversion from ac to dc and back again. The two six-pulse bridges are identical except that the ac supply voltages of the two are shifted in phase by 30°. As a result of the 30° phase shift, certain harmonics listed above are cancelled, thus leaving voltage harmonics on the dc side of orders 12, 24, 36, . . . , 12*n*, and current harmonics on the ac side of orders 11, 13, 23, 25, . . . , $12n \pm 1$. This reduction in harmonic filter requirements is one of the major reasons why most solid-state HVDC systems operate with twelve-pulse converters.

6.4 Electrical Environmental Considerations

HVDC systems environmentally are often more compatible and provide less operational difficulties than comparable ac systems. Similar to ac lines, the two categories of dc line electrical environmental considerations are corona effects and field effects. With an HVDC line, the corona effects may be broken into three categories: audible noise, radio interference, and air ions. The field effects of an HVDC line are electric and magnetic fields.

The positive-polarity conductor is the primary source of dc transmission-line audible noise and radio interference. The radio interference produced by the negative pole is about one-half that from the positive pole. Rain causes a slight reduction in audible noise. dc radio interference levels are decreased by rain, wet snow, and other atmospheric conditions that thoroughly wet the conductor. However, radio interference may increase slightly during the initial wetting period. Wind also affects dc radio interference levels. The radio interference levels are increased by wind, with the greatest influence being when the direction of air flow is from the negative to the positive pole (Chartier, 1981).

Air ions are charged molecules produced by the breakdown of the air molecules adjacent to the conductor (corona). Because of the nonalternating nature of the direct-current transmission, the air ions migrate away from the conductor. Because both conductors of a dc line have an electric field, both can produce corona and therefore air ions. Most air ions are attracted to the conductor opposite the one that generated them. Neutralization occurs when air ions combine with those of opposite polarity. Most air ions from HVDC lines are neutralized; approximately 10% of the ions escape and migrate away from the transmission line. The migrating air ions are carried away by wind, much like dust particles or pollen. Therefore, few air ions produced by the dc line are present on the upwind side of the line (Maruvada *et al.*, 1984; Johnson, 1983).

HVDC lines produce static (or constant) electric and magnetic fields. A typical electric field for a 500-kV dc transmission line is about 15 kV/m, and the magnetic field is about 100 mG at the edge of the right of way. The earth also has static electric and magnetic fields on the order of 0.09–0.13 kV/m and 480–600 mG, respectively.

6.5 Specific System Considerations

The decision to select a dc transmission system instead of an ac system may be summarized with one question: "Will taps or new substations be needed along the route in the future to supply additional loads or incorporate additional power sources?" dc will compete with ac systems where long distances are involved, charging current limits the ampacity, asynchronous interconnection is feasible, right-of-way availability demands a higher power density, controllability of dc enhances system stability, or lower short-circuit currents are desired.

Long cable circuits are a natural application for dc. With ac, the charging current utilizes part of the cable ampacity, whereas with dc, the full ampacity is available for power transfer. Underground (or underwater) cable lengths greater than 15–25 km require shunt reactors for ac installation. Higher dc voltages than ac voltages can be tolerated by a given cable design. Hence, a dc project will require fewer cables or less costly cables.

During steady-state operation, a rectifier can be considered as a load and the inverter as a generator. Power across a dc line can be set anywhere within line rating by simply requesting the desired power level. Loss of a converter is similar to load rejection or generator tripping or, in the case of a dc line between two points in the same synchronous system, the loss of an ac line carrying the same power. A dc line can be built to move a specified amount of power.

Dynamic operation of a dc line or back-to-back converter is very different from the operation of a comparable ac system. There are three capabilities that dc can provide that make it unique:

- The transfer over the line can be reduced almost instantaneously to a lower power-transfer level or all the way to zero, if necessary.
- If equipped with temporary overload capability, the dc line can be ramped very quickly to a power level that exceeds the normal operating level.
- The real and/or reactive power level can be smoothly varied or "modulated."

GLOSSARY

A-Weighted Sound Level [dB(A)]: Measured sound pressure level adjusted to correspond approximately to the sensitivity of human hearing according to a specified frequency response curve.

Ampacity: Current-carrying capacity of a wire or cable under specified thermal conditions.

Bundled Conductor: An assembly of two or more conductors used as a single conductor and employing spacers to maintain a predetermined configuration.

Bus: A heavy conductor in a substation used to connect substation components at a specific voltage level.

Coupling Capacitor: A capacitive voltage divider used to connect communications signals to power transmission conductors.

Current Transformer: A transformer designed and applied to maintain a specific ratio of current from primary to secondary for metering and protective relay applications.

Distribution: The part of an electric power system that transports power from high-voltage substations to the end users.

Geometric Mean Radius (GMR): The equivalent radius of a single conductor that gives the same impedance as a bundled conductor.

Grading Ring: A metallic ring (e.g., mounted at the end of an insulator) to lower the electric field on some component.

Line Trap: A coil mounted in a substation to isolate communications signals to a specific power transmission line.

Positive Sequence: The sequence resulting from the symmetrical component transformation that represents normal balanced operation of a three-phase system.

Quasistatic: Low-frequency approximation where static (dc) electric and magnetic field equations adequately represent time-varying fields.

Reactive Power: The product of the voltage, current, and sine of the phase angle between them in a sinusoidal circuit. Unit of measure is volt-ampere reactive (var).

Relaxation Time: The time required for the deviation from equilibrium of some system parameter to diminish to $1/e$ of its initial value.

Shield Wire: A wire employed for the purpose of reducing the effects on electric supply or communication circuits from extraneous sources; in overhead power transmission, a grounded wire situated above the phase conductors for lightning protection.

Subconductor: One of the individual conductors constituting a bundled conductor.

Substation: A power station used to interconnect transmission and/or distribution lines, often at different voltages.

Subtransmission: A term sometimes used for that part of the power transmission system connecting bulk high-voltage substations and distribution substations.

Symmetrical Components: A matrix transformation for a three-phase system that decomposes the phase voltages, currents, and impedances into three sequences called positive, negative, and zero sequence.

Transmission: The part of an electric power system devoted to bulk transport of power.

Transverse Electromagnetic (TEM): An electromagnetic field mode whose electric and magnetic field vectors are both normal to the direction of propagation.

Voltage (Potential) Transformer: A transformer designed and applied to maintain a specific ratio of voltage from primary to secondary for metering and protective relay applications.

Zero Sequence: The sequence resulting from the symmetrical component transformation that represents the average current of the three phases that returns somewhere other than through the phase conductors themselves.

Works Cited

Abilgaard, E. H., Bauer, E. A., de Lussanet de la Sableniere, K. L. (1976), "Composite Longrod Insulators and Their Influence on the Design of

Overhead Lines," *Conf. Int. Grands Reseaux Electr. Haute Tension, Bull.* No. 26, paper 22–03.

Adler, R. B., Chu, L. J., Fano, R. M. (1960), *Electromagnetic Energy Transmission and Radiation,* New York: Wiley.

Allustiarti, R., Hoexter, H., Lai, P., Samuelsson, J., Lindberg, P., Rocamora, R. G. (1988), "Design and Operating Performance of 500-kV Metal-Oxide-Protected Series-Capacitor Banks on the Table Mountain-Tesla Line," *IEEE Trans. Power Del.* **3,** 1951–1959.

Aluminum Association (1982), *Aluminum Electrical Conductor Handbook,* 2nd ed., Washington, DC: The Aluminum Association.

Barnes, H. C., Thoren, B. (1973), "The AEP-ASEA UHV Test Station and Line," paper C73-319-1, presented at the IEEE Summer Power Meeting, Vancouver, B.C., 15–20 July (unpublished).

Bean, R. L., Chackan, N., Jr., Moore, H. R., Wentz, E. C. (1959), *Transformers for the Electric Power Industry,* Pittsburgh: Westinghouse Electric Corp.

Bergen, A. R. (1986), *Power Systems Analysis,* Englewood Cliffs, NJ: Prentice-Hall.

Bonneville Power Administration (1974), *Development of Ultra-High Voltage Transmission,* Portland, R: Bonneville Power Administration.

Bonneville Power Administration (1989), *Electrical and Biological Effects of Transmission Lines: A Review,* Washington, DC: U.S. Department of Energy.

Canadian Standards Association (1975), "Tolerable Limits and Methods of Measurement of Electromagnetic Interference from Alternating Current High Voltage Power Systems," CSA Standard C108.3.1-1975, Ottawa: Canadian Standards Association.

Chartier, V. L. (1981), "Investigations of Corona and Field Effects of AC/DC Hybrid Transmission Lines," *IEEE Trans. Power Apparat. Syst.* **PAS-100,** 72–80.

Dunlop, R. D., Gutman, R., Marchenko, P. P. (1979), "Analytical Development of Loadability Characteristics for EHV and UHV Transmission Lines," *IEEE Trans. Power Apparat. Syst.* **PAS-98,** 606–617.

El-Hawary, M. E. (1983), *Electrical Power Systems Design and Analysis,* Reston, VA: Reston Publishing Co.

EPRI (1978), *Transmission Line Reference Book, 115–138 kV Compact Line Design,* Palo Alto, CA: Electric Power Research Institute.

EPRI (1980), *Bundled Circuit Design for 115–138 kV Compact Transmission Lines,* Report No. EL 1314, Palo Alto, CA: Electric Power Research Institute.

EPRI (1982), *Transmission Line Reference Book, 345 kV and Above,* Palo Alto, CA: Electric Power Research Institute.

EPRI (1992), *Underground Transmission Systems Reference Book,* Palo Alto, CA: Electric Power Research Institute.

IEC (1969), *Calculation of the Continuous Rating of Cables (100% Load Factor),* Publication No. 287, Geneva: International Electrotechnical Commission.

IEEE Task Force (1982), "A Comparison of Methods for Calculating Audible Noise of High Voltage Transmission Lines," *IEEE Trans. Power Apparat. Syst.* **PAS-101,** 4090–4099.

Harrington, R. F. (1961), *Time-Harmonic Electromagnetic Fields,* New York: McGraw-Hill.

Johnson, G. B. (1983), "Electric Fields and Ion Currents of a ±400 kV HVDC Test Line," *IEEE Trans. Power Apparat. Syst.* **PAS-102,** 2559–2568.

Kaminski, J., Jr. (1963), "Long Time Mechanical and Electrical Strength in Suspended Insulators," *Trans. AIEE* **67,** 446–451.

Karady, G., Lamontagne, G. (1976), "Electrical and Contamination Performance of Synthetic Insulators for 735 kV Transmission Lines," Paper A76-502-5, presented at the IEEE Power Engineering Society Summer Meeting, Portland, Oregon, 18–23 July.

Katz, C., Eager, P. S., Jr., Seman, G. W., Garcia, F. G., Smith, W. G., McCourt, J. W. (1978), "Progress in the Determination of AC/DC Resistance Ratios of Pipe-Type Cable Systems," *IEEE Trans. Power Apparat. Syst.* **PAS-97,** 2262–2271.

Keast, D. N. (1980), "Assessing the Impact of Audible Noise from AC Transmission Lines: A Proposed Method," *IEEE Trans. Power Apparat. Syst.* **PAS-99,** 1021–1031.

Larsen, E. V., Baker, D. H., Imece, A. F., Gerin-Lajoie, L., Scott, G. (1990), "Basic Aspects of Applying SVC's to Series-Compensated AC Transmission Lines," *IEEE Trans. Power Del.* **5,** 1466–1473.

Loftness, M. O. (1992), *Power Line Interference, A Practical Handbook,* NRECA Research Project 90-30, Washington, DC: Rural Electric Research.

Maruvada, P. S., Dallaire, R. D., Heroux, P., Rivest, N. (1984), "Long Term Statistical Study of the Corona Electric Field and Ion Corona Performance of a ±900 kV Bipolar HVDC Transmission Line Configuration," *IEEE Trans. Power Apparat. Syst.* **PAS-103,** 76–83.

Mountford, J. D., Koessler, R. J., Zelingher, S. (1993), "Practical Application of FACTS Technology in System Control," NTUA-IEEE/PES Joint International Power Conference, September.

Neher, J. H., McGrath, M. H. (1957), "The Calculation of Temperature Rise and Load Capability of Cable Systems," *AIEE Trans. Power Apparat. Syst.* **32,** 752–772.

Nichols, F. S., Vose, F. C. (1963), "A Polymer Insulator for High Voltage Transmission Lines," *Elec. Eng.* **82,** 684–688.

Olsen, R. G., Wong, P. S. (1992), "Characteristics of Low Frequency Electric and Magnetic

Fields in the Vicinity of Electric Power Lines," *IEEE Trans. Power Del.* **7,** 2046–2055.

Olsen, R. G., Pankaskie, T. A. (1983), "On the Exact, Carson and Image Theories for Wires at or Above the Earth's Interface," *IEEE Trans. Power Apparat. Syst.* **PAS-102,** 769–774.

Stevenson, W. D., Jr. (1982), *Elements of Power System Analysis,* 4th ed., New York: McGraw-Hill.

Souchereau, N., Sabourin, G., Cayer, P., Tsui, Y. T. (1978), "Validation of a Chainette Tower for a 735 kV Line," *Conf. Int. Grands Reseaux Electr. Haute Tension, Bull.* No. 27, paper No. 22-04.

PRECIOUS METALS

See SILVER, GOLD AND OTHER PRECIOUS METALS

PROBABILITY THEORY

Lajos Takács, *Case Western Reserve University, Cleveland, Ohio, U.S.A.*

1. Introduction 51
2. Random Trials 52
3. General Theorems 54
4. Combinatorial Methods 55
5. Geometric Methods 56
6. Conditional Probabilities 57
6.1 The Theorem of Total Probability 58
6.2 Bayes's Theorem 58
7. Independent Events and Trials 59
8. Random Variables 60
9. Expectation of Random Variables 62
10. Moments, Variance, and Covariance 63
10.1 The Correlation Coefficient 64
11. Generating Function, Characteristic Function, and Laplace–Stieltjes Transform 64
12. Discrete Distributions 65
12.1 The Bernoulli Distribution 65
12.2 The Negative Binomial Distribution 65
12.3 De Moivre's Limit Theorem 65
12.4 The Poisson Distribution 66
12.5 The Hypergeometric Distribution 66
12.6 The Multinomial Distribution 66
13. Continuous Distributions 67
13.1 The Exponential Distribution 67
13.2 The Gamma Distribution 67
13.3 The Beta Distribution 67
13.4 The Normal Distribution 68
13.5 The χ^2 Distribution 68
13.6 The Multivariate Normal Distribution 68
14. Laws of Large Numbers 68
14.1 The Weak Law of Large Numbers 69
14.2 The Strong Law of Large Numbers 69
15. Limit Distributions 69
15.1 The Central Limit Theorem 69
15.2 The Arcsine Law 70
Further Reading 70

1. INTRODUCTION

The mathematical theory of probability provides a systematic approach to the study of random mass phenomena, that is, experiments that can be performed any number of times under identical conditions. This theory was initiated by B. Pascal (1623–1662) and P. Fermat (1601–1665) in 1654. The pioneering results were achieved by Ch. Huygens (1629–1695), Jakob Bernoulli (1654–1705), P. R. Montmort (1678–1719), and A. De Moivre (1667–1754). In the early nineteenth century, great progress was made by P. S. Laplace (1749–1827) and J. Lagrange (1736–1813). At the end of the nineteenth century and at the beginning of the twentieth century, P. L. Chebyshev (1821–1894) and A. A. Markov (1856–1922) contributed significantly to the theory. The foundations of modern probability theory were laid down by P. Lévy (1886–1971), A. Ya. Khintchine (1894–1959), and A. N. Kolmogorov (1903–1987).

An experiment is called a *random trial* if it has several possible outcomes and the actual outcome is not determined by the given conditions. For example, the experiment may be the tossing of a balanced coin, and the result may be either heads or tails, but we cannot predict which one. The various outcomes of a random trial are called *events* (simple or compound) and are frequently denoted by Latin capital letters. We speak about a simple event if it can occur only in one way, and a compound event if it can occur in more than one way. In the example of the coin, let

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

A denote the event that the coin shows heads and B, the event that it shows tails. Both A and B are simple events.

Let us suppose that we toss a coin repeatedly n times and we obtain $k_n(A)$ heads and $k_n(B)$ tails. The ratio $k_n(A)/n$ is called the *frequency* of A in the n trials, and $k_n(B)/n$, the frequency of B in the n trials. It is an empirical observation that in repeated trials the frequency of any event shows stability if the number of trials increases. If n increases, then, in a random way, $k_n(A)/n$ approaches a real number p, and $k_n(B)/n$ approaches a real number q. By an approach in a random way, we mean that as n increases, in an overwhelming majority of the trials the deviations

$$\left|\frac{k_n(A)}{n} - p\right| \quad \text{and} \quad \left|\frac{k_n(B)}{n} - q\right|$$

are very small. The numbers $\mathbf{P}\{A\} = p$ and $\mathbf{P}\{B\} = q$ are called the *probabilities* of the events A and B respectively.

Since

$$0 \le \frac{k_n(A)}{n} \le 1, \quad 0 \le \frac{k_n(B)}{n} \le 1,$$

and

$$k_n(A)/n + k_n(B)/n = 1,$$

for every $n = 1, 2, \ldots$, we must have $0 \le p \le 1$, $0 \le q \le 1$, and $p + q = 1$.

In general, if in a random trial an event A has probability p where $0 \le p \le 1$, and if we repeat the trial n times, we expect that the observed frequency of A will be close to p, provided that n is sufficiently large.

We cannot determine the probability of an event by theoretical reasoning. The probability of any event is a physical constant, like the mass of an atom. Relying on our experience with random trials, we accept the existence of such a constant, and our aim is to build up a mathematical theory that reflects this fact and makes it possible to give predictions about the nature of the outcomes of repeated random trials. Thus we can define a mathematical model for the trial of tossing a coin. If the probability of heads is p where $0 \le p \le 1$, then the probability of tails is necessarily $q = 1 - p$. By the theory of probability, we can derive several general results for the coin-tossing trials. Let us suppose that the coin is tossed n times under identical conditions, and denote by ν_n the number of heads observed, that is, $\nu_n = k_n(A)$ in our previous notation. As usual, ν_n is called a *random variable* (random quantity), which may take on the values $0, 1, \ldots, n$. The probability of the event $\nu_n = k$, that is, the probability that we obtain exactly k heads, is

$$\mathbf{P}\{\nu_n = k\} = \binom{n}{k} p^k q^{n-k}$$

for $k = 0, 1, \ldots, n$. Denote by

$$\delta_n = \nu_n/n - p$$

the deviation between the frequency and the probability of heads in the n trials. For any positive real number a, we have

$$\mathbf{P}\{|\delta_n| \le a\sqrt{pq/n}\} = \sum_{|k-np| \le a\sqrt{npq}} \mathbf{P}\{\nu_n = k\}.$$

This probability is arbitrarily close to

$$P(a) = \frac{1}{\sqrt{2\pi}} \int_{-a}^{a} e^{-x^2/2}\, dx$$

if n is sufficiently large. If $a = 3$, then $P(3) = 0.9973 \cdots$. If in our trial we use a balanced coin, then by symmetry we expect that both heads and tails have the same probability, that is, $p = q = \frac{1}{2}$. In this case, it is reasonable to choose the model for which $p = \frac{1}{2}$. Then by the above general results we can draw the conclusion that if n is sufficiently large, in n trials, the deviation between the frequency and the probability of heads is $\le 3/2\sqrt{n}$ with probability $0.9973 \cdots$. If the coin is loaded, then we should select another model for which $p \ne \frac{1}{2}$. If p is unknown, then we can still use the theory of probability to design tests that make it possible to estimate p within a prescribed error with high accuracy.

2. RANDOM TRIALS

In order to study random trials mathematically, we associate a *probability space* $(\Omega,\mathcal{B},\mathbf{P})$ with every random trial considered. Here Ω is the *sample space*, the set of all the possible simple outcomes of the random trial, and $\mathcal{B}$ is the class of events with which we desire to associate probabilities. If A is an event belonging to $\mathcal{B}$, then $\mathbf{P}\{A\}$ is the probability of A.

Let us consider the example of throwing a six-sided die. We observe the number of points showing on the top face of the die. This number may be 1, 2, 3, 4, 5, or 6. Let us define the sample space as $\Omega = \{1,2,3,4,5,6\}$. Each sample point represents a simple event. We also want to consider compound events, such as A = the result is an even number, B = the result is an odd number, and C = the result is a prime number. We also include Ω, the sure event that always occurs, and $\varnothing$, the impossible event, which never occurs. There are $2^6 = 64$ admissible events associated with the trial considered. We can describe every event as a subset of Ω. Each simple event is represented by a single sample point. A compound event contains more than one sample point. The events $A = \{2,4,6\}$, $B = \{1,3,5\}$, $C = \{2,3,5\}$, and $\Omega = \{1,2,3,4,5,6\}$, the sure event, are compound events. The impossible event, $\varnothing$, is represented by the empty set.

Now let us consider various possible relations among events. If the occurrence of an event A implies the occurrence of an event B, then we write $A \subset B$ (A is a subset of B). The complementary event of A will be denoted by $\bar{A}$ (A does not occur). If we have two events A and B, then we can define a new event by the condition "both A and B occur." This event, the simultaneous occurrence of A and B, will be denoted by AB or by $A \cap B$ (the intersection of the two sets A and B). The event "at least one of the two events A and B occurs" will be denoted by $A + B$ or by $A \cup B$ (the union of the two sets A and B). More generally, $ABC \cdots$ or $A \cap B \cap C \cap \cdots$ means the simultaneous occurrence of $A, B, C, \ldots$ (the intersection of $A, B, C, \ldots$), and $A + B + C + \ldots$ or $A \cup B \cup C \cup \cdots$ means that at least one event occurs among $A, B, C, \ldots$ (the union of the sets $A, B, C, \ldots$). Two events are mutually exclusive if they cannot occur simultaneously, that is, if $AB = \varnothing$. More generally, any number of events $A, B, C, \ldots$ are mutually exclusive if among them no two events can occur simultaneously.

A class of events (some subsets of Ω) $\mathcal{B}$ is called an *algebra* if it is closed under any finite number of set operations (forming intersection, union, and complementary sets). Equivalently, $\mathcal{B}$ is an algebra of some subsets of Ω if and only if (a) $A \in \mathcal{B}$ implies that $\bar{A} \in B$, and (b) $A \in \mathcal{B}$ and $B \in \mathcal{B}$ imply that $A + B \in \mathcal{B}$.

If Ω, the sample space, is finite, we always assume that $\mathcal{B}$, the class of random events, is an algebra. If $\mathcal{B}$ contains all the possible subsets of Ω, then $\mathcal{B}$ is obviously an algebra.

First, let us assume that Ω is finite, say, $\Omega = \{\omega_1,\omega_2,\ldots,\omega_m\}$, consisting of m sample points, and $\mathcal{B}$ is an algebra of some subsets of Ω. With every event $A \in \mathcal{B}$, we want to associate a nonnegative real number, $\mathbf{P}\{A\}$, the probability of A. Since there are various relations among events, we expect that the probabilities of the different events cannot be chosen freely; they are subject to some restraints. These restraints are expressed by the following system of axioms: (i) $\mathbf{P}\{A\} \geq 0$ for $A \in \mathcal{B}$, (ii) $\mathbf{P}\{\Omega\} = 1$, and (iii) if $A \in \mathcal{B}$, $B \in \mathcal{B}$, and $AB = \varnothing$, then

$$\mathbf{P}\{A + B\} = \mathbf{P}\{A\} + \mathbf{P}\{B\}.$$

These are obvious requirements. They are based on our intuitive interpretation of the probability of an event.

The system of axioms implies that the probability is finitely additive; that is, if A_1, $A_2, \ldots, A_n$ are mutually exclusive events, then we have

$$\mathbf{P}\{A_1 + A_2 + \cdots + A_n\} = \mathbf{P}\{A_1\} + \mathbf{P}\{A_2\} + \cdots + \mathbf{P}\{A_n\}$$

for $n = 2, 3, \ldots$.

If Ω, the sample space, is finite, then every probability space $(\Omega,\mathcal{B},\mathbf{P})$ that satisfies the above requirements provides a mathematical model for a random trial.

If Ω contains m sample points, that is, $\Omega = \{\omega_1,\omega_2,\ldots,\omega_m\}$, if $\mathcal{B}$ contains all the 2^m subsets of Ω, including the empty set and the whole set, and if the event $A_i = \{\omega_i\}$ ($i = 1,2,\ldots,m\}$ has probability $\mathbf{P}\{A_i\} = p_i$ ($i = 1,2,\ldots,m$), then it follows from the system of axioms that $p_i \geq 0$ for all $i = 1, 2, \ldots, m$ and

$$p_1 + p_2 + \cdots + p_m = 1.$$

If A is a subset of Ω that contains k sample points, namely, $A = \{\omega_{i_1},\omega_{i_2},\ldots,\omega_{i_k}\}$, then by the system of axioms,

$$\mathbf{P}\{A\} = p_{i_1} + p_{i_2} + \cdots + p_{i_k}.$$

In this case to every choice of the nonnegative real numbers $p_1, p_2, \ldots, p_m$ whose sum is $p_1 + p_2 + \cdots + p_m = 1$, there corresponds a probability space $(\Omega,\mathcal{B},\mathbf{P})$ that describes a random trial. The case where p_i does not depend on i has special importance. Then necessarily $p_i = 1/m$, and if A is defined as above, then

$$\mathbf{P}\{A\} = k/m.$$

This is a simple rule for calculating the probabilities of various events. If there are m possible cases, if they are equally probable, and if an event A can occur in k different ways (i.e., the number of cases favorable for A is k), then the probability of A is simply the ratio of the number of favorable cases to the total number of cases. In the age of classical probability theory, this model was almost exclusively the only one considered, and the above rule served as the definition of probability.

If Ω, the sample space, contains infinitely many sample points, then, in defining the associated probability space, $(\Omega,\mathcal{B},\mathbf{P})$, we impose some further restrictions on $\mathcal{B}$ and $\mathbf{P}$. We require that $\mathcal{B}$, the class of random events, be closed not only for finite, but also for countably infinite numbers of set operations. If this is satisfied, then we say that $\mathcal{B}$ is a σ-algebra. We can see that $\mathcal{B}$ is a σ-algebra of some subsets of Ω if and only if (a) $A \in \mathcal{B}$ implies that $\bar{A} \in B$, and (b) $A_n \in \mathcal{B}$ for $n = 1, 2, \ldots$ implies that

$$\sum_{n=1}^{\infty} A_n \in \mathcal{B}.$$

We require that the probability be not only finitely additive, but also countably additive. In this case, the system of axioms is as follows: (i) $\mathbf{P}\{A\} \geq 0$ for $A \in \mathcal{B}$, (ii) $\mathbf{P}\{\Omega\} = 1$, and (iii) if $A_i \in \mathcal{B}$ for $i = 1, 2, \ldots$ and $A_iA_j = \emptyset$ for $i \neq j$, then

$$\mathbf{P}\left\{\sum_{i=1}^{\infty} A_i\right\} = \sum_{i=1}^{\infty} \mathbf{P}\{A_i\}.$$

Briefly, the probability $\mathbf{P}\{A\}$ is a finite, nonnegative, countably additive (σ-additive) set function on a σ-algebra $\mathcal{B}$ of subsets of Ω.

3. GENERAL THEOREMS

Let $(\Omega,\mathcal{B},\mathbf{P})$ be a probability space describing a random trial. Let $A, B, A_1, A_2, \ldots$ be random events belonging to $\mathcal{B}$. It is always assumed that $\mathcal{B}$ is a σ-algebra of subsets of Ω and $\mathbf{P}\{A\}$ is a probability defined for $A \in \mathcal{B}$, which satisfies the system of axioms.

From the system of axioms, we can deduce several general theorems for $\mathbf{P}\{A\}$, $A \in \mathcal{B}$. Here are a few theorems.

If $A \subset B$, then $\mathbf{P}\{A\} \leq \mathbf{P}\{B\}$. The probability of the impossible event is $\mathbf{P}\{\emptyset\} = 0$. The probability of the complementary event of A is $\mathbf{P}\{\bar{A}\} = 1 - \mathbf{P}\{A\}$.

Boole's inequality: For an arbitrary sequence of events A_i ($i = 1,2,\ldots$) we have

$$\mathbf{P}\{A_1 + A_2 + \cdots\} \leq \mathbf{P}\{A_1\} + \mathbf{P}\{A_2\} + \cdots.$$

Let us consider n events $A_1, A_2, \ldots, A_n$. Define $S_0 = 1$ and

$$S_j = \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} \mathbf{P}\{A_{i_1}A_{i_2} \cdots A_{i_j}\}$$

for $j = 1, 2, \ldots, n$. Here S_j is the sum of $\binom{n}{j}$ probabilities. Then we have *Poincaré's theorem:* The probability that at least one event occurs among $A_1, A_2, \ldots, A_n$ is

$$\mathbf{P}\{A_1 + A_2 + \cdots + A_n\} = \sum_{j=1}^{n} (-1)^{j-1} S_j.$$

We also have *Jordan's theorem:* The probability that exactly k events occur among $A_1, A_2, \ldots, A_n$ is

$$P_k(A_1,A_2,\ldots,A_n) = \sum_{j=k}^{n} (-1)^{j-k}\binom{j}{k}S_j$$

for $k = 0, 1, \ldots, n$.

If $A_1, A_2, \ldots, A_n, \ldots$ is an infinite sequence of events, then

$$A^* = \limsup_{n\to\infty} A_n := \prod_{n=1}^{\infty}\sum_{i=n}^{\infty} A_i$$

is the event that infinitely many events occur in the sequence $A_1, A_2, \ldots, A_n, \ldots$, and

$$A_* = \liminf_{n\to\infty} A_n := \sum_{n=1}^{\infty}\prod_{i=n}^{\infty} A_i$$

is the event that all but a finite number of events occur in the sequence $A_1, A_2, \ldots, A_n, \ldots$. If $A_n \in \mathcal{B}$ for $n = 1, 2, \ldots$, then $A^* \in \mathcal{B}$ and $A_* \in \mathcal{B}$, and

$$\mathbf{P}\{A^*\} = \lim_{n\to\infty} \mathbf{P}\left\{\sum_{i=n}^{\infty} A_i\right\}$$

and

$$\mathbf{P}\{A_*\} = \lim_{n\to\infty} \mathbf{P}\left\{\prod_{i=n}^{\infty} A_i\right\}.$$

Obviously, $A_* \subset A^*$ and $\mathbf{P}\{A_*\} \le \mathbf{P}\{A^*\}$.

A remark about notations: We can use interchangeably the notations $A + B$ and $A \cup B$, and likewise AB and $A \cap B$. In this article, the notations $A + B$ and AB are used for a finite or a countably infinite number of events, and the notations $\cup$ and $\cap$ are used for nondenumerable events. It is ambiguous to write, for example, $\mathbf{P}\{A \cup B \cap C\}$. The notation $\mathbf{P}\{A + BC\}$ indicates immediately the hierarchy of operations. Multiplication (intersection) has preference over addition (union) just like in arithmetic.

4. COMBINATORIAL METHODS

In Sec. 2, we have already mentioned an important model for random trials in which the sample space Ω contains m sample points, $\mathcal{B}$ contains all the possible 2^m subsets of Ω, and if an event $A \in \mathcal{B}$ contains exactly k sample points, then

$$\mathbf{P}\{A\} = k/m,$$

that is, the probability of A is the number of favorable cases divided by the number of possible cases. We can write also that $\mathbf{P}\{A\} = |A|/|\Omega|$ where $|A|$ is the number of elements of the set A, and similarly $|\Omega|$ is the number of elements of the set Ω. It is usual to apply combinatorial methods for the determination of $|A|$ and $|\Omega|$.

Let us consider a few examples.

Example 1. We have n boxes, and each box contains a white balls and b black balls. We draw a ball from each box. What is the probability that among the n balls drawn there will be k white balls and $n - k$ black balls? If all the possible $(a + b)^n$ cases are equally probable, then the probability in question is

$$P_k = \binom{n}{k}\frac{a^k b^{n-k}}{(a + b)^n} = \binom{n}{k} p^k q^{n-k},$$

where $p = a/(a + b)$, $q = b/(a + b)$, and $k = 0, 1, \ldots, n$. The answer is the same if we have only one box that contains a white balls and b black balls, and we draw n times with replacement, that is, after each drawing we put back into the box the ball drawn.

Example 2. We have a box that contains a white balls and b black balls, and we draw n times without replacement, that is, we do not put the balls drawn back into the box. What is the probability that among the n balls drawn there will be k white balls and $n - k$ black balls? If $n \le a + b$ and if all the possible $(a + b)(a + b - 1) \cdots (a + b - n + 1)$ cases are equally probable, then the probability in question is

$$Q_k = \binom{n}{k}\frac{a(a-1)\cdots(a-k+1)b(b-1)\cdots(b-n+k+1)}{(a+b)(a+b-1)\cdots(a+b-n+1)} = \frac{\binom{a}{k}\binom{b}{n-k}}{\binom{a+b}{n}}$$

for $k = 0, 1, \ldots, n$. We obtain the same probability if we take all the n balls simultaneously out of the box and assume that all the possible $\binom{a+b}{n}$ cases are equally probable.

Example 3. A box contains m cards marked $1, 2, \ldots, m$. We draw n cards with replacement. There are m^n possible results. If all these m^n cases are equally probable, then the probability that the sum of the n numbers drawn is k is given by

$$P_k = \frac{1}{m^n}\sum_{j=0}^{[(k-n)/m]} (-1)^j \binom{n}{j}\binom{k - 1 - jm}{n - 1}$$

for $n \le k \le mn$. This formula was discovered by P. R. Montmort in 1710 and A. De Moivre in 1711.

Example 4. A box contains n cards marked $k_1, k_2, \ldots, k_n$ where $k_1, k_2, \ldots, k_n$ are nonnegative integers with sum $k_1 + k_2 + \cdots + k_n = k$ where $k \le n$. We draw all the n cards without replacement. Let us assume that all the $n!$ outcomes are equally probable. Then the probability that the sum of the first r numbers drawn is less than r for every $r = 1, 2, \ldots, n$ is given by

$$P(n,k) = (n - k)/n.$$

This problem is a generalization of the classical ballot problem of J. Bertrand from 1887.

Example 5. We distribute n balls in m boxes. All the m^n arrangements are assumed to be equally probable. The probability that the ith box contains k_i balls for $i = 1, 2, \ldots, m$ is

$$P(k_1,k_2,\ldots,k_m) = \frac{n!}{k_1!k_2!\cdots k_m!}\left(\frac{1}{m}\right)^n$$

if $k_1 + k_2 + \cdots + k_m = n$. In statistical mechanics the Maxwell–Boltzmann statistics is based on the above model.

5. GEOMETRIC METHODS

First, let us say a few words about Borel sets and the Lebesgue measure in finite-dimensional Euclidean spaces. In one dimension, an elementary set is defined as the union of a finite number of disjoint intervals. Let us consider the smallest class of sets that contains all the elementary sets and that is closed under any finite or countably infinite number of intersections and unions. If a set belongs to this class, then the set is called a Borel set. With every Borel set S, we can associate a nonnegative real number $\mu(S)$, the measure of S. If S is an interval, then $\mu(S)$ is simply the length of S. If an interval S is the union of a finite or countably infinite number of disjoint intervals $S_1, S_2, \ldots$, then the length of S satisfies the following property:

$$\mu(S) = \mu(S_1) + \mu(S_2) + \cdots.$$

The Lebesgue measure of a Borel set S is simply an extension of the notion of length to more general sets than intervals in such a way that above-mentioned additive property is preserved. It can be proved that there is a unique extension. Accordingly, if S is a Borel set and if S is the union of a finite or countably infinite number of disjoint Borel sets S_i, $i = 1, 2, \ldots$, then

$$\mu(S) = \mu(S_1) + \mu(S_2) + \cdots.$$

In higher dimensions, we define Borel sets and the Lebesgue measure in a similar way. In two dimensions, we start with the area of a rectangle; in three dimensions, with the volume of a parallelepiped, and so on.

Let Ω be a Borel set with a finite, positive Lebesgue measure $\mu(\Omega)$. We choose a point at random in Ω. We can describe this random trial by the probability space $(\Omega,\mathcal{B},\mathbf{P})$ where $\mathcal{B}$ is the σ-algebra of Borel subsets of Ω, and if $A \in \mathcal{B}$, then $\mathbf{P}\{A\} = \mu(A)/\mu(\Omega)$ where $\mu(A)$ is the Lebesgue measure of A. In this case, we say simply that the random point has a uniform distribution over the set Ω. The problem of finding the probabilities of various events can be reduced to the problem of finding the measures of various sets in one, two, three, or higher dimensions.

Here are a few examples.

Example 1. We choose a point at random in the interval $(0,t)$. Suppose that the point has a uniform distribution over $(0,t)$. Let $0 \le x \le t$. What is the probability that the random point belongs to the interval $(0,x)$? The answer is evidently x/t.

Example 2. Let us choose two points at random in the interval $(0,t)$. The outcome of this random trial can be described by an ordered point pair (u,v) where $0 \le u \le t$ and $0 \le v \le t$. It is obvious to define the sample space as $\Omega = \{(u,v):0 \le u \le t, 0 \le v \le t\}$, that is, Ω is a square in the plane. Let us assume that the point pair (u,v) has a uniform distribution over Ω. In this case, we define $\mathcal{B}$ as the σ-algebra of Borel subsets of Ω, and the probability of an event $A \in \mathcal{B}$ is $\mathbf{P}\{A\} = \mu(A)/\mu(\Omega)$ where now $\mu(A)$ is the two-dimensional Lebesgue measure of A and $\mu(\Omega) = t^2$, the area of Ω. If we want to find the probability that both points belong to the interval $(0,x)$ where $0 \le x \le t$, then the answer is $\mathbf{P}\{A\}$ where now $A = \{(u,v):0 \le u \le x, 0 \le v \le x\}$. In this case, A is a square and its area is x^2. Thus $\mathbf{P}\{A\} = x^2/t^2$.

Example 3. What is the probability that in the preceding example the distance between the two points is $\leq x$ where $0 \leq x \leq t$? The answer is simply $\mathbf{P}\{B\}$ where now $B = \{(u,v):|u - v| \leq x\}$. We can easily see that the area of B is $t^2 - (t - x)^2 = 2tx - x^2$ and so the probability in question is $(2tx - x^2)/t^2$.

Example 4. In 1733, Comte de Buffon (G. L. Leclerc) proposed the following problem in geometric probability. A plane is ruled with a series of parallel lines a distance of $2a$ apart. A needle of length $2l (l < a)$ is thrown at random on the plane. Determine the probability that the needle will touch or intersect one of the lines. We can describe the outcome of this random trial by a pair of numbers (x,α), where x is the distance from the center of the needle to the closest line and α is the angle that the needle forms with this line. We have $0 \leq x \leq a$ and $0 \leq \alpha \leq \pi$. In this case, we can define the sample space as $\Omega = \{(x,\alpha):0 \leq x \leq a, 0 \leq \alpha \leq \pi\}$, that is, Ω is a rectangle in the plane and its area is $\mu(\Omega) = a\pi$. If we assume that the point pair (x,α) has a uniform distribution over Ω, then the probability in question is $\mathbf{P}\{C\} = \mu(C)/\mu(\Omega)$ where now C is the subset of Ω that contains the point pairs (x,α) for which the needle touches or intersects one of the lines. This happens if and only if $0 \leq x \leq l \sin\alpha$, that is, $C = \{(x,\alpha):0 \leq x \leq l \sin\alpha, 0 \leq \alpha \leq \pi\}$. The area of C is

$$\mu(C) = l \int_0^\pi \sin\alpha \, d\alpha = 2l,$$

and consequently, $\mathbf{P}\{C\} = 2l/(a\pi)$. In 1850, a Swiss astronomer, J. R. Wolf, performed 5000 trials in the case where $l/a = 0.8$ and estimated π empirically. He found that $\pi \sim 3.159 \cdots$.

6. CONDITIONAL PROBABILITIES

The following consideration leads in a natural way to the notion of conditional probability. Let us consider a random trial with probability space $(\Omega,\mathcal{B},\mathbf{P})$. Let $A \in \mathcal{B}$ and $B \in \mathcal{B}$ be two given events. Let us perform the random trial n times under identical conditions and observe the number of occurrences of A, B, and AB. Suppose that in n trials A occurs $k_n(A)$ times, B occurs $k_n(B)$ times, and AB occurs $k_n(AB)$ times. If n increases, $k_n(A)/n$, $k_n(B)/n$, and $k_n(AB)/n$ approach in a random way the probabilities $\mathbf{P}\{A\}$, $\mathbf{P}\{B\}$, and $\mathbf{P}\{AB\}$ respectively.

The frequency of A in the $k_n(B)$ trials in which B occurs is

$$\frac{k_n(AB)}{k_n(B)} = \frac{k_n(AB)/n}{k_n(B)/n}.$$

If $\mathbf{P}\{B\} \neq 0$, then $k_n(AB)/k_n(B)$ will approach $\mathbf{P}\{AB\}/\mathbf{P}\{B\}$ in a random way. We can interpret $\mathbf{P}\{AB\}/\mathbf{P}\{B\}$ as the probability of A given that B occurs. We shall denote it by $\mathbf{P}\{A|B\}$ and call it the *conditional probability* of A for given B (or if it is known that B occurred, or on the hypothesis B.) If $\mathbf{P}\{B\} > 0$, then for an arbitrary event A we define

$$\mathbf{P}\{A|B\} = \mathbf{P}\{AB\}/\mathbf{P}\{B\}.$$

The conditional probability $\mathbf{P}\{A|B\}$ remains undefined when $\mathbf{P}\{B\} = 0$. Then necessarily $\mathbf{P}\{AB\} = 0$.

Let us suppose that the random trial consists in throwing a symmetric die and we assume that each side has the same probability of being the top face. Denote by A the event that the result is an even number, that is, $A = \{2,4,6\}$, and B, the event that the result is ≤ 3, that is, $B = \{1,2,3\}$. Then $\mathbf{P}\{A\} = \mathbf{P}\{B\} = \frac{3}{6}$ and $\mathbf{P}\{AB\} = \frac{1}{6}$. The conditional probability

$$\mathbf{P}\{A|B\} = \tfrac{1}{3}.$$

By the definition of the conditional probability, it follows that

$$\mathbf{P}\{AB\} = \mathbf{P}\{A|B\}\mathbf{P}\{B\}$$

if $\mathbf{P}\{B\} \neq 0$, and by symmetry

$$\mathbf{P}\{AB\} = \mathbf{P}\{A\}\mathbf{P}\{B|A\}$$

if $\mathbf{P}\{A\} \neq 0$. More generally, for any n events $A_1, A_2, \ldots, A_n$ we have the so-called chain rule,

$$\mathbf{P}\{A_1A_2 \cdots A_n\} = \mathbf{P}\{A_1\}\mathbf{P}\{A_2|A_1\}\mathbf{P}\{A_3|A_1A_2\} \cdots \mathbf{P}\{A_n|A_1 \cdots A_{n-1}\},$$

provided that the conditional probabilities on the right-hand side are meaningful.

6.1 The Theorem of Total Probability

Let $\{A_i, i \in I\}$ be a complete system of events, that is, a sequence of mutually exclusive and exhaustive events. By this, we mean that $A_iA_j = \varnothing$ for $i \neq j$ and $\Sigma_{i \in I} A_i = \Omega$. Then for any event B we have

$$\mathbf{P}\{B\} = \sum_{i \in I} \mathbf{P}\{A_i\}\mathbf{P}\{B|A_i\}.$$

If it happens that for some i the probability $\mathbf{P}\{A_i\} = 0$, then in the sum the ith term should be interpreted as 0. This formula is useful because the evaluation of the conditional probabilities $\mathbf{P}\{B|A_i\}$ is sometimes easier than a direct evaluation of $\mathbf{P}\{B\}$.

Example 1. Let us consider $m + 1$ boxes numbered $i = 0, 1, \ldots, m$. In box i, there are i white balls and $m - i$ black balls. We choose a box at random. We may choose any of the $m + 1$ boxes with the same probability. We draw n balls with replacement from the box chosen. What is the probability that we get k white balls and $n - k$ black balls?

Let A_i denote the event of choosing box i. Then $\mathbf{P}\{A_i\} = 1/(m + 1)$ for $i = 0, 1, \ldots, m$. Denote by B the event of drawing k white balls and $n - k$ black balls. Then

$$\mathbf{P}\{B|A_i\} = \binom{n}{k} \frac{i^k(m - i)^{n-k}}{m^n}$$

for $i = 0, 1, \ldots, m$, and

$$\mathbf{P}\{B\} = \sum_{i=0}^{m} \mathbf{P}\{A_i\}\mathbf{P}\{B|A_i\}$$
$$= \binom{n}{k} \frac{1}{(m + 1)m^n} \sum_{i=0}^{m} i^k(m - i)^{n-k}.$$

6.2 Bayes's Theorem

In 1763, an important theorem was discovered in the manuscripts of the Rev. Thomas Bayes, who had died two years earlier. By this theorem if $\mathbf{P}\{B\} > 0$ and $\{A_i, i \in I\}$ is a complete system of events, then

$$\mathbf{P}\{A_i|B\} = \frac{\mathbf{P}\{A_iB\}}{\mathbf{P}\{B\}} = \frac{\mathbf{P}\{A_i\}\mathbf{P}\{B|A_i\}}{\Sigma_{j \in I}\mathbf{P}\{A_j\}\mathbf{P}\{B|A_j\}}.$$

This theorem can be applied in the following situation. We perform a random trial and observe the occurrence of an event B. If $\{A_i\}$ is a complete system of events, then one and only one event necessarily occurs. Here $\mathbf{P}\{A_i\}$ is the *a priori* (before observation) probability of A_i. The conditional probability $\mathbf{P}\{A_i|B\}$ is the *a posteriori* (after observation) probability of A_i, that is, the probability of A_i calculated under the hypothesis that we observe the occurrence of B.

Example 2. Let us suppose that we have $m + 1$ boxes. For each $i = 0, 1, \ldots, m$ there is one box that contains i white balls and $m - i$ black balls. We choose a box at random. We may choose any of the $m + 1$ boxes with equal probability. From it we draw n balls with replacement. We get k white balls and $n - k$ black balls. What is the probability that we have chosen the box that contains i white balls and $m - i$ black balls? If we use the same notations as in example 1, then the answer is

$$\mathbf{P}\{A_i|B\} = \frac{\mathbf{P}\{A_i\}\mathbf{P}\{B|A_i\}}{\Sigma_{j=0}^{m} \mathbf{P}\{A_j\}\mathbf{P}\{B|A_j\}} = \frac{i^k(m - i)^{n-k}}{\Sigma_{j=0}^{m} j^k(m - j)^{n-k}}$$

for $i = 0, 1, \ldots, m$.

Example 3. A box contains m balls, white balls and black balls in unknown proportion. We draw n balls with replacement. Among the n balls, we get k white balls and $n - k$ black balls. What is the probability that in the box there are i white balls and $m - i$ black balls?

Let A_i $(i = 0,1,\ldots,m)$ be the event that there are i white balls and $m - i$ black balls in the box, and denote by B the event that in the n drawings with replacement we get k white balls and $n - k$ black balls. Then

$$\mathbf{P}\{B|A_i\} = \binom{n}{k} \frac{i^k(m - i)^{n-k}}{m^n}$$

and write $\mathbf{P}\{A_i\} = \omega_i$ $(i = 0,1,\ldots,m)$. By Bayes's theorem

$$\mathbf{P}\{A_i|B\} = \frac{\mathbf{P}\{A_i\}\mathbf{P}\{B|A_i\}}{\Sigma_{j=0}^{m}\mathbf{P}\{A_j\}\mathbf{P}\{B|A_j\}}$$
$$= \frac{\omega_i i^k(m - i)^{n-k}}{\Sigma_{j=0}^{m}\omega_j j^k(m - j)^{n-k}}$$

for $i = 0, 1, \ldots, m$. Now $\omega_i = \mathbf{P}\{A_i\}$ denotes the *a priori* probability that there are i white and $m - i$ black balls in the box, and $P_i(n) = \mathbf{P}\{A_i|B\}$ $(i = 0,1,\ldots,m)$ is the *a posteriori* (conditional) probability of the event that there are i white balls and $m - i$ black balls in the box. To find this probability, we must accept some hypothesis as to how the box was filled with balls. For example, we may suppose that all possible proportions of the white balls to the black balls are equally probable. Then $\omega_i = 1/(m + 1)$ for $i = 0, 1, \ldots, m$ and

$$P_i(n) = \frac{i^k(m - i)^{n-k}}{\sum_{j=0}^{m} j^k(m - j)^{n-k}}$$

for $i = 0, 1, \ldots, m$. This hypothesis is justified, for example, if we know that the box was chosen among the $m + 1$ boxes considered in example 2 and all the $m + 1$ possible choices were equally probable. There are many other possible hypotheses. For example, the box may contain as many white balls as the number of heads obtained in tossing m balanced coins, and as many black balls as the number of tails obtained. If we know that the box was filled with balls in that way, then we should choose

$$\omega_i = \binom{m}{i}\frac{1}{2^m}$$

for $i = 0, 1, \ldots, m$.

7. INDEPENDENT EVENTS AND TRIALS

The conditional probability $\mathbf{P}\{A|B\}$ generally does not equal the probability $\mathbf{P}\{A\}$. The information as to whether B has occurred changes the probability of the occurrence of the event A. Only when $\mathbf{P}\{A|B\} = \mathbf{P}\{A\}$ does this information not permit any inference as to the occurrence of A. In this case we say that A is *independent* of B. The notion of independence in probability theory is interpreted in a stochastic or statistical sense. Since $\mathbf{P}\{AB\} = \mathbf{P}\{A|B\}\mathbf{P}\{B\}$, the condition $\mathbf{P}\{A|B\} = \mathbf{P}\{A\}$ can be expressed in the following form: $\mathbf{P}\{AB\} = \mathbf{P}\{A\}\mathbf{P}\{B\}$. This equation is symmetric in A and B, and shows that whenever A is independent of B so is B of A.

Definition: Two events A and B are said to be independent if

$$\mathbf{P}\{AB\} = \mathbf{P}\{A\}\mathbf{P}\{B\}.$$

This definition is accepted also if $\mathbf{P}\{A\} = 0$ or $\mathbf{P}\{B\} = 0$, although in these cases $\mathbf{P}\{B|A\}$ and $\mathbf{P}\{A|B\}$ are not defined.

If A and B are independent, then $\bar{A}$ and B, A and $\bar{B}$, and $\bar{A}$ and $\bar{B}$ are similarly independent.

More generally, we say that n events $A_1, A_2, \ldots, A_n$ are mutually independent if any event and the simultaneous occurrence of any other events are independent. This definition is equivalent to the following: The events $A_1, A_2, \ldots, A_n$ are mutually independent if for any $i_1, i_2, \ldots, i_r$ satisfying the inequalities $1 \le i_1 < i_2 < \cdots < i_r \le n$ we have

$$\mathbf{P}\{A_{i_1}A_{i_2}\cdots A_{i_r}\} = \mathbf{P}\{A_{i_1}\}\mathbf{P}\{A_{i_2}\}\cdots\mathbf{P}\{A_{i_r}\}.$$

The above requirements mean $\binom{n}{r}$ conditions for any $r = 2, 3, \ldots, n$. The total number of conditions is

$$\binom{n}{2} + \binom{n}{3} + \cdots + \binom{n}{n} = 2^n - (n + 1).$$

If $\mathbf{P}\{A_{i_1}A_{i_2}\} = \mathbf{P}\{A_{i_1}\}\mathbf{P}\{A_{i_2}\}$ for $1 \le i_1 < i_2 \le n$, then we say that $A_1, A_2, \ldots, A_n$ are pairwise independent. Pairwise independence does not imply mutual independence. This fact will be illustrated by the following example. Two dice are thrown simultaneously. Let us suppose that all the 36 possible outcomes are equally probable. Three events A, B, and C are defined as follows: A means odd number with the first die, B means odd number with the second die, and C means odd sum. In this case, $\mathbf{P}\{A\} = \mathbf{P}\{B\} = \mathbf{P}\{C\} = \frac{18}{36} = \frac{1}{2}$ and $\mathbf{P}\{AB\} = \mathbf{P}\{AC\} = \mathbf{P}\{BC\} = \frac{9}{36} = \frac{1}{4}$. Consequently, A, B, and C are pairwise independent: $\mathbf{P}\{AB\} = \mathbf{P}\{A\}\mathbf{P}\{B\}$, $\mathbf{P}\{AC\} = \mathbf{P}\{A\}\mathbf{P}\{C\}$, and $\mathbf{P}\{BC\} = \mathbf{P}\{B\}\mathbf{P}\{C\}$. But A, B, C are not mutually independent: $\mathbf{P}\{ABC\} \neq \mathbf{P}\{A\}\mathbf{P}\{B\}\mathbf{P}\{C\}$. For $\mathbf{P}\{ABC\} = 0$. The three events cannot occur simultaneously; $ABC = \varnothing$, the impossible event that has probability 0.

An infinite sequence of events $A_1, A_2, \ldots, A_n, \ldots$ is said to be mutually independent if for any $n = 2, 3, \ldots$, the events $A_1, A_2, \ldots, A_n$ are mutually independent.

Our next aim is to introduce the notion of *independent trials*. Previously, we defined the notion of independent events concerning a random trial. That is, we considered a probability space $(\Omega, \mathcal{B}, \mathbf{P})$ and defined the notion of independence for events in $\mathcal{B}$.

Now let us consider a *compound random*

trial (e.g., random trials performed simultaneously, a succession of random trials, or, in particular, repeated random trials). If any outcome of a constituent random trial is independent of the simultaneous occurrence of any outcomes of any other constituent trials in the compound random trial, we say that the constituent random trials are independent. Conversely, if we consider random trials whose outcomes do not interact on each other, then we suppose that they are independent, that is, the compound random trial has the property that the outcomes of the constituent random trials are independent. Clearly, ordinary independence, defined in terms of occurrences, implies stochastic independence defined in terms of probability alone. The only problem is to show that there always exists a compound trial with the described property. In 1933, A. N. Kolmogorov and in 1939, more generally, B. Jessen proved that there always exists a so-called product probability space that possesses the indicated properties.

For example, let us suppose that we perform two random trials and in each trial we choose a point at random in the interval $(0,t)$. Assume that each point has a uniform distribution over $(0,t)$. Let us associate with the compound random trial the probability space defined in example 2 of Sec. 5. In this case, the two constituent trials are independent in the compound trial. Conversely, if the two trials do not interact on each other, we define the compound trial in the same way as in example 2 of Sec. 5.

8. RANDOM VARIABLES

So far, we have dealt with random events occurring in random trials, and our aim has been to determine the probabilities of various events. However, in many cases it is more convenient to work with random variables. Intuitively, a random variable is a random quantity whose value depends on the outcome of a random trial. The outcome of a random trial is completely determined by one, and only one, sample point. Consequently, the value of a random variable is determined if it is known which simple event occurs. We shall consider mostly real random variables and denote them by lowercase Greek letters such as ξ, η, ζ, χ, ν, τ. For example, let us suppose that the random trial consists in throwing a die. The number of points on the top face of the die is a random variable ξ that may assume the values 1, 2, 3, 4, 5, 6. If the sample space is $\Omega = \{1,2,3,4,5,6\}$, then $\xi = \xi(\omega) = \omega$ for $\omega \in \Omega$. Accordingly, a random variable is a function defined on the sample space. Many random variables can be defined on the same sample space. For instance, let us suppose that there are two players P_1 and P_2 and one of them throws a die. If the score is even, then P_2 pays one dollar to P_1, and if the score is odd, then P_1 pays one dollar to P_2. Denote by η_1 the gain of P_1 and η_2 the gain of P_2. Then $\eta_1(\omega)$, $\omega \in \Omega$, and $\eta_2(\omega)$, $\omega \in \Omega$, are random variables. Another example: Denote by A the event that in throwing the die the result is an even number. Define a random variable χ such that $\chi = 1$ if A occurs and $\chi = 0$ if A does not occur. Such a random variable is called an *indicator random variable.* It has only two possible values, 0 and 1. Table 1 illustrates the above-mentioned random variables.

Let $(\Omega,\mathcal{B},\mathbf{P})$ be a probability space and $\xi(\omega)$, a function defined on the sample space. We shall not consider arbitrary functions $\xi(\omega)$, but only those for which probability questions can be answered. An example of such questions is "what is the probability that $\xi \in S$ where S is a linear Borel set?" In order for all such questions to be answerable, it is necessary and sufficient that the set $\{\omega: \xi(\omega) \in S\}$ belong to $\mathcal{B}$. This is assured by the following definition:

A real, single-valued function $\xi(\omega)$, defined on the sample space Ω, is called a real random variable, if for every choice of the real number x $(-\infty < x < \infty)$ the set $\{\omega: \xi(\omega) \le x\}$ [the set of all $\omega \in \Omega$ for which the inequality $\xi(\omega) \le x$ is satisfied] belongs to $\mathcal{B}$. Briefly, a real random variable is a measurable function on the sample space (measurable with respect to $\mathcal{B}$).

The function

Table 1. Random variables.

ω	1	2	3	4	5	6
ξ	1	2	3	4	5	6
η_1	-1	1	-1	1	-1	1
η_2	1	-1	1	-1	1	-1
χ	0	1	0	1	0	1

$$\mathbf{P}\{\xi \le x\} = F(x)$$

defined for every real x is called the *distribution function* of the random variable ξ. The event $\xi \le x$ is described by the set of sample points $\omega \in \Omega$ for which $\xi(\omega) \le x$. We define $F(\infty) = \lim_{x\to\infty} F(x)$ and $F(-\infty) = \lim_{x\to-\infty} F(x)$. If $\xi(\omega)$ is finite for every $\omega \in \Omega$, then $F(\infty) = 1$ and $F(-\infty) = 0$. If we know the distribution function $F(x)$ of a random variable ξ, and if S is any linear Borel set, then the probability $\mathbf{P}\{\xi \in S\}$ is uniquely determined by $F(x)$.

A distribution function $F(x)$ is a nondecreasing function of x. If $F(x)$ is an absolutely continuous function of x, then there exists a function $f(x)$ such that $f(x) \ge 0$, $f(x)$ is integrable over $(-\infty,\infty)$ and

$$F(x) = \int_{-\infty}^{x} f(u)du$$

for all x. Such an $f(x)$ is called a *density function* of a real random variable. The density function is not determined uniquely by $F(x)$, but any two versions are equal almost everywhere. They can differ only on a set of measure 0. If $f(x)$ is a density function of ξ, then

$$\frac{dF(x)}{dx} = f(x)$$

for almost every $x \in (-\infty,\infty)$. The derivative $dF(x)/dx$ exists for every $x \in (-\infty,\infty)$ except perhaps for a set of measure 0.

If $\mathbf{P}\,\{\xi = a\} = F(a) - F(a - 0) > 0$, then $\xi = a$ is a possible value of the random variable ξ and $F(x)$ has a jump of magnitude $F(a) - F(a - 0)$ at $x = a$. The set of possible values of a real random variable is always countable (empty, finite, or countably infinite). If $F(x)$ increases only in jumps, then we say that $F(x)$ is a step function and ξ is a discrete random variable.

If in a random trial $(\Omega,\mathcal{B},\mathbf{P})$ we consider n random variables $\xi_1, \xi_2, \ldots, \xi_n$, that is, n functions $\xi_1(\omega), \xi_2(\omega), \ldots, \xi_n(\omega)$ defined on Ω and measurable with respect to $\mathcal{B}$, then for every choice of the real numbers $x_1, x_2, \ldots, x_n$ the events $\{\omega\colon \xi_i(\omega) \le x_i\} \in \mathcal{B}$. The simultaneous occurrence of these n events also belongs to $\mathcal{B}$ and thus the following function is well defined:

$$F(x_1,x_2,\ldots,x_n) = \mathbf{P}\{\xi_1 \le x_1, \xi_2 \le x_2, \ldots, \xi_n \le x_n\}.$$

It is called the *joint distribution function* of the random variables $\xi_1, \xi_2, \ldots, \xi_n$. The function $F(x_1,x_2,\ldots,x_n)$ is an n-dimensional distribution function.

If there exists a nonnegative, integrable function $f(x_1,x_2,\ldots,x_n)$ such that

$$F(x_1,x_2,\ldots,x_n) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_n} f(u_1,u_2,\ldots,u_n) \times du_1 du_2 \cdots du_n,$$

then we say that $f(x_1,x_2,\ldots,x_n)$ is a *joint density function* of the random variables $\xi_1, \xi_2, \ldots, \xi_n$. If such a function exists, then

$$\frac{\partial^n F(x_1,x_2,\ldots,x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n} = f(x_1,x_2,\ldots,x_n)$$

for almost every $(x_1,x_2,\ldots,x_n)$. For any n-dimensional Borel set S we have

$$\mathbf{P}\{(\xi_1,\xi_2,\ldots,\xi_n) \in S\} = \int\int\cdots\int_S f(x_1,x_2,\ldots,x_n) \times dx_1 dx_2 \cdots dx_n.$$

If

$$\mathbf{P}\{\xi_1 \le x_1,\ldots,\xi_n \le x_n\} = \mathbf{P}\{\xi_1 \le x_1\} \cdots \mathbf{P}\{\xi_n \le x_n\}$$

for every choice of $x_1, \ldots, x_n$, then we say that the random variables $\xi_1, \xi_2, \ldots, \xi_n$ are *mutually independent*. If $\xi_1, \xi_2, \ldots, \xi_n$ are mutually independent, then for any linear Borel sets $S_1, S_2, \ldots, S_n$ we have

$$\mathbf{P}\{\xi_1 \in S_1, \xi_2 \in S_2, \ldots, \xi_n \in S_n\} = \mathbf{P}\{\xi_1 \in S_1\}\mathbf{P}\{\xi_2 \in S_2\} \cdots \mathbf{P}\{\xi_n \in S_n\}.$$

If each variable ξ_i $(i = 1,2,\ldots,n)$ has a density function $f_i(x)$, then the variables $\xi_1, \xi_2, \ldots, \xi_n$ have a joint density function

$$f(x_1,x_2,\ldots,x_n) = f_1(x_1)f_2(x_2)\cdots f_n(x_n).$$

If ξ_1 and ξ_2 are independent random variables with distribution functions $\mathbf{P}\{\xi_1 \le x\} = F_1(x)$ and $\mathbf{P}\{\xi_2 \le x\} = F_2(x)$ respectively, then the distribution function of the sum, $\mathbf{P}\{\xi_1 + \xi_2 \le x\} = H(x)$, can be calculated by the following Riemann–Stieltjes integral:

$$H(x) = \int_{-\infty}^{\infty} F_1(x - u)dF_2(u).$$

We can write briefly that $H(x) = F_1(x) * F_2(x)$ and say that $H(x)$ is the *convolution* of $F_1(x)$ and $F_2(x)$. If ξ_1 has a density function $f_1(x)$ and ξ_2 has a density function $f_2(x)$, then $\xi_1 + \xi_2$ also has a density function $h(x)$ given by

$$h(x) = \int_{-\infty}^{\infty} f_1(x - u)f_2(u)du.$$

We can determine the distribution function of the sum of mutually independent random variables by forming repeated convolutions.

9. EXPECTATION OF RANDOM VARIABLES

Let us consider a random trial with which a probability space $(\Omega,\mathcal{B},\mathbf{P})$ is associated. Let $\xi = \xi(\omega)$ be a real random variable defined on Ω.

First, let us suppose that ξ is a discrete random variable with possible values a_j ($j \in I$) where I is a finite or countably infinite set and the corresponding probabilities are $\mathbf{P}\{\xi = a_j\} = p_j$ for $j \in I$. Repeat the random trial n times under identical conditions and observe the value of ξ in each trial: $X_1, X_2, \ldots, X_n$. Suppose that the value a_j is observed n_j times. Then $\Sigma_{j\in I} n_j = n$. The arithmetic mean of the observations is

$$\frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\Sigma_{j\in I} n_j a_j}{\Sigma_{j\in I} n_j} = \sum_{j\in I} a_j \frac{n_j}{n}.$$

In many cases if $n \to \infty$, the observed mean shows stability, namely, it approaches a certain constant in a random way. The frequency of the event $A_j = \{\xi = a_j\}$, that is, n_j/n, approaches p_j, the probability of A_j. Consequently, the mean value of the observations should approach

$$\sum_{j\in I} a_j p_j.$$

If

$$\sum_{j\in I} |a_j| p_j < \infty,$$

then

$$\sum_{j\in I} a_j p_j$$

is a definite number and it is called the expectation of the random variable ξ.

Definition: If ξ is a discrete random variable that assumes the values a_j ($j \in I$) with corresponding probabilities $\mathbf{P}\{\xi = a_j\} = p_j$, then the *expectation* (or mean, or average) of ξ is defined by

$$\mathbf{E}\{\xi\} = \sum_{j\in I} a_j p_j = \sum_{j\in I} a_j \mathbf{P}\{\xi = a_j\}$$

provided that the series converges absolutely. In this case, we say that ξ has a finite expectation; the expectation exists. If

$$\sum_{j\in I} |a_j| p_j = \infty,$$

then we say that ξ has no finite expectation. (The expectation does not exist: infinite or indefinite.)

If we throw a symmetric die and ξ is the number of points on the top face of the die, then $\mathbf{P}\{\xi = j\} = \frac{1}{6}$ for $j = 1,2,\ldots,6$ and the expectation of ξ is

$$\mathbf{E}\{\xi\} = \frac{1 + 2 + \cdots + 6}{6} = \frac{7}{2} = 3.5.$$

If we repeat the trial n times and in each trial we observe the number of points obtained, then we expect that the arithmetic mean of the observations will be approximately 3.5 if n is large enough.

If $\mathbf{P}\{\xi \le x\} = F(x)$ and ξ has a density function $f(x)$, then the expectation of ξ is defined by

$$\mathbf{E}\{\xi\} = \int_{-\infty}^{\infty} xf(x)dx,$$

provided that

$$\int_{-\infty}^{\infty} |x| f(x)dx < \infty.$$

In the general case when $\mathbf{P}\{\xi \le x\} = F(x)$, the expectation of ξ is defined by

$$\mathbf{E}\{\xi\} = \int_{-\infty}^{\infty} xdF(x),$$

provided that

$$\int_{-\infty}^{\infty} |x| dF(x) < \infty.$$

Here the integrals are Riemann–Stieltjes integrals.

If ξ is a nonnegative real random variable with finite expectation $\mathbf{E}\{\xi\}$, then *Chebyshev's inequality* states that

$$\mathbf{P}\{\xi \geq a\} \leq \frac{\mathbf{E}\{\xi\}}{a}$$

for $a > 0$.

One of the most frequently used theorems is the following: If $\xi_1, \xi_2, \ldots, \xi_n$ are real random variables and

$$\eta_n = \xi_1 + \xi_2 + \cdots + \xi_n,$$

then the expectation of η_n is

$$\mathbf{E}\{\eta_n\} = \mathbf{E}\{\xi_1\} + \mathbf{E}\{\xi_2\} + \cdots + \mathbf{E}\{\xi_n\},$$

provided that each ξ_i ($i = 1,2,\ldots,n$) has a finite expectation.

If $\xi_1, \xi_2, \ldots, \xi_n$ are mutually independent random variables having finite expectations, then

$$\mathbf{E}\{\xi_1\xi_2 \cdots \xi_n\} = \mathbf{E}\{\xi_1\}\mathbf{E}\{\xi_2\} \cdots \mathbf{E}\{\xi_n\}.$$

10. MOMENTS, VARIANCE, AND COVARIANCE

If ξ is a real random variable and $g(x)$ is a continuous function, or more generally a Borel-measurable function, then $\eta = g(\xi)$ is also a random variable. If $\mathbf{P}\{\xi \leq x\} = F(x)$, then the expectation of η is

$$\mathbf{E}\{\eta\} = \int_{-\infty}^{\infty} g(x) dF(x)$$

provided that

$$\int_{-\infty}^{\infty} |g(x)| dF(x) < \infty.$$

If ξ is a discrete random variable with possible values a_j ($j \in I$) and corresponding probabilities $\mathbf{P}\{\xi = a_j\} = p_j$, then

$$\mathbf{E}\{\eta\} = \sum_{j \in I} g(a_j) p_j,$$

provided that the series is absolutely convergent. If ξ has a density function $f(x)$, then

$$\mathbf{E}\{\eta\} = \int_{-\infty}^{\infty} g(x) f(x) dx,$$

provided that

$$\int_{-\infty}^{\infty} |g(x)| f(x) dx < \infty.$$

If, in particular, $g(x) = x^m$ where $m = 0, 1, 2, \ldots$, then $\mathbf{E}\{\xi^m\}$ is called the mth *moment* of the random variable ξ. If $\mathbf{E}\{\xi\} = a$ exists, and $g(x) = (x - a)^m$ where $m = 0, 1, 2, \ldots$, then $\mathbf{E}\{(\xi - a)^m\}$ is called the mth *central moment* of ξ.

The second central moment plays an important role. It is called the *variance* of ξ and is denoted by

$$\mathrm{Var}\{\xi\} = \mathbf{E}\{(\xi - a)^2\} = \mathbf{E}\{\xi^2\} - [\mathbf{E}\{\xi\}]^2.$$

The positive square root of the variance is called the *standard deviation.*

The variance measures the spread of the values of ξ around the expectation. If $\mathbf{E}\{\xi\} = a$ and $\mathrm{Var}\{\xi\} = \sigma^2$ where σ is the standard deviation, and if we apply Chebyshev's inequality to the random variable $\eta = (\xi - a)^2$, then we obtain that

$$\mathbf{P}\{|\xi - a| \geqq \lambda\sigma\} = \mathbf{P}\{(\xi - a)^2 \geqq \lambda^2\sigma^2\} \leq 1/\lambda^2$$

for any $\lambda > 0$.

If $\mathbf{E}\{\xi\} = a$ and $\mathrm{Var}\{\xi\} = \sigma^2$, where $\sigma > 0$, then the random variable

$$\xi^* = (\xi - a)/\sigma$$

is called the *standardized random variable* corresponding to ξ. It has expectation 0 and variance 1.

If $\xi_1, \xi_2, \ldots, \xi_n$ are real random variables with finite expectations $\mathbf{E}\{\xi_i\} = a_i$ ($i = 1,2,\ldots,n$) and finite variances $\mathrm{Var}\{\xi_i\} = \sigma_i^2$ ($i = 1,2,\ldots,n$), then

$$\mathrm{Var}\{\xi_1 + \xi_2 + \cdots + \xi_n\} = \sum_{i=1}^{n} \mathrm{Var}\{\xi_i\} + 2 \sum\sum_{1 \leq i < j \leq n} \mathrm{Cov}\{\xi_i, \xi_j\},$$

where the *covariance* of ξ_i and ξ_j is defined by

$$\begin{aligned}\operatorname{Cov}\{\xi_i,\xi_j\} &= \mathbf{E}\{(\xi_i - a_i)(\xi_j - a_j)\}\\ &= \mathbf{E}\{\xi_i\xi_j\} - \mathbf{E}\{\xi_i\}\mathbf{E}\{\xi_j\}.\end{aligned}$$

If the random variables $\xi_1, \xi_2, \ldots, \xi_n$ are pairwise independent, then $\operatorname{Cov}\{\xi_i,\xi_j\} = 0$ for $1 \le i < j \le n$, and

$$\operatorname{Var}\{\xi_1 + \xi_2 + \cdots + \xi_n\} = \sum_{i=1}^{n} \operatorname{Var}\{\xi_i\}.$$

10.1 The Correlation Coefficient

Let ξ_1 and ξ_2 be two real random variables with finite expectations a_1 and a_2 and finite positive variances σ_1^2 and σ_2^2, respectively. The covariance of the standardized random variables $\xi_1^* = (\xi_1 - a_1)/\sigma_1$ and $\xi_2^* = (\xi_2 - a_2)/\sigma_2$ is called the *correlation coefficient* of ξ_1 and ξ_2 and is denoted by

$$\rho(\xi_1,\xi_2) = \operatorname{Cov}\{\xi_1^*,\xi_2^*\} = \operatorname{Cov}\{\xi_1,\xi_2\}/\sigma_1\sigma_2.$$

The correlation coefficient measures the linear dependence of the two random variables ξ_1 and ξ_2. We have always $|\rho(\xi_1,\xi_2)| \le 1$, and $\rho(\xi_1,\xi_2) = \pm 1$ if and only if $\xi_2 = a\xi_1 + b$ where a and b are constants. If ξ_1 and ξ_2 are independent, then $\rho(\xi_1,\xi_2) = 0$.

11. GENERATING FUNCTION, CHARACTERISTIC FUNCTION, AND LAPLACE–STIELTJES TRANSFORM

If ξ and η are real random variables, then $\zeta = \xi + i\eta$ is a complex random variable. If both ξ and η have finite expectations, then we say that the expectation of ζ exists and is defined by

$$\mathbf{E}\{\zeta\} = \mathbf{E}\{\xi\} + i\mathbf{E}\{\eta\}.$$

If $\zeta_1, \zeta_2, \ldots, \zeta_n$ are complex random variables and $\mathbf{E}\{\zeta_k\}$ exists for $k = 1, 2, \ldots, n$, then

$$\begin{aligned}\mathbf{E}\{\zeta_1 + \zeta_2 + \cdots + \zeta_n\} = \mathbf{E}\{\zeta_1\} &+ \mathbf{E}\{\zeta_2\}\\ &+ \cdots + \mathbf{E}\{\zeta_n\}.\end{aligned}$$

We say that $\zeta_k = \xi_k + i\eta_k$ $(k = 1,2,\ldots,n)$ are mutually independent random variables if the n events $(\xi_k,\eta_k) \in S_k$ $(k = 1,2,\ldots,n)$ are mutually independent for any choice of the two-dimensional Borel sets S_k $(k = 1,2,\ldots,n)$. If $\zeta_k = \xi_k + i\eta_k$ $(k = 1,2,\ldots,n)$ are mutually independent random variables, then

$$\mathbf{E}\{\zeta_1\zeta_2\cdots\zeta_n\} = \mathbf{E}\{\zeta_1\}\mathbf{E}\{\zeta_2\}\cdots\mathbf{E}\{\zeta_n\},$$

provided that the expectations on the right-hand side exist.

If ξ is a real random variable and z is a complex number, then z^ξ is a complex random variable. If we know the distribution function $\mathbf{P}\{\xi \le x\} = F(x)$, then the expectation of z^ξ is a function of z and is given by

$$G(z) = \mathbf{E}\{z^\xi\} = \int_{-\infty}^{\infty} x^x\, dF(x),$$

provided that the expectation exists. It is called the *generating function* of ξ.

If $z = e^{it}$ where t is a real number, then the expectation of $e^{it\xi}$, that is,

$$\psi(t) = G(e^{it}) = \int_{-\infty}^{\infty} e^{itx}\, dF(x),$$

always exists and is called the *characteristic function* of ξ.

If $z = e^{-s}$ where s is a complex number and if the expectation of $e^{-s\xi}$, that is,

$$\varphi(s) = G(e^{-s}) = \int_{-\infty}^{\infty} e^{-sx}\, dF(x),$$

exists, then it is called the *Laplace–Stieltjes transform* of ξ.

The distribution function of ξ is uniquely determined by any of the aforementioned three transforms, and if the moments of ξ exist, then they can be determined by forming the derivatives of these transforms. In particular,

$$\mathbf{E}\{\xi\} = G'(1) = -i\psi'(0) = -\varphi'(0)$$

and

$$\mathbf{E}\{\xi^2\} = G''(1) + G'(1) = -\psi''(0) = \varphi''(0).$$

If $\xi_1, \xi_2, \ldots, \xi_n$ are mutually independent real random variables, then $z^{\xi_1}, z^{\xi_2}, \ldots, z^{\xi_n}$ are also mutually independent and

$$\mathbf{E}\{z^{\xi_1+\xi_2+\cdots+\xi_n}\} = \mathbf{E}\{z^{\xi_1}\}\mathbf{E}\{z^{\xi_2}\}\cdots\mathbf{E}\{z^{\xi_n}\},$$

provided that the expectations on the right-hand side exist. By this equation, we can conclude that the generating function, or the characteristic function, or the Laplace–Stieltjes transform of a sum of n mutually independent real random variables is the product of the appropriate transforms of the n random variables.

It is convenient to use generating functions for discrete random variables that take on nonnegative integer values only, and Laplace–Stieltjes transforms for nonnegative real random variables.

If $G(z)$ is the generating function of a discrete random variable ξ that takes on nonnegative integer values only, then

$$\mathbf{P}\{\xi = k\} = \frac{1}{k!}\left(\frac{d^kG(z)}{dz^k}\right)_{z=0}.$$

If a real random variable ξ has a density function $f(x)$, then its characteristic function is

$$\psi(t) = \int_{-\infty}^{\infty} e^{itx}f(x)dx.$$

Conversely, if $\psi(t)$ is the characteristic function of a real random variable ξ, and if

$$\int_{-\infty}^{\infty} |\psi(t)|dt < \infty,$$

then ξ has a density function $f(x)$ and

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\psi(t)dt.$$

12. DISCRETE DISTRIBUTIONS

12.1 The Bernoulli Distribution

Let us assume that in a sequence of repeated independent trials there are only two possible outcomes: "success", which has probability p, and "failure", which has probability q, where $p > 0$, $q > 0$, and $p + q = 1$. In this case, we speak about a sequence of *Bernoulli trials*. Denote by ν_n the number of successes in the first n trials and τ_n the number of trials performed up to the occurrence of the nth success. Then

$$\mathbf{P}\{\nu_n = k\} = \binom{n}{k} p^kq^{n-k}$$

for $k = 0, 1, \ldots, n$. We say that the random variable ν_n has a *Bernoulli distribution* (or binomial distribution) with parameters n and p. The expectation of ν_n is $\mathbf{E}\{\nu_n\} = np$ and its variance is $\mathrm{Var}\{\nu_n\} = npq$. Furthermore,

$$\mathbf{E}\{|\nu_n - np|\} = 2sq\mathbf{P}\{\nu_n = s\},$$

where $s = [np] + 1$. The generating function of ν_n is

$$\mathbf{E}\{z^{\nu_n}\} = (q + pz)^n$$

for any z.

12.2 The Negative Binomial Distribution

We have $\{\tau_n \leq n + k\}$ if and only if $\{\nu_{n+k} \geq n\}$, that is

$$\mathbf{P}\{\tau_n \leq n + k\} = \mathbf{P}\{\nu_{n+k} \geq n\}$$

for $n \geq 1$ and $k \geq 0$. Hence

$$\mathbf{P}\{\tau_n = n + k\} = \binom{n + k - 1}{n - 1}p^nq^k$$

for $n \geq 1$ and $k \geq 0$. We can write also that

$$\mathbf{P}\{\tau_n = n + k\} = (-1)^k\binom{-n}{k}p^nq^k$$

for $n \geq 1$ and $k \geq 0$, and say that τ_n has a *negative binomial distribution.* The expectation of τ_n is $\mathbf{E}\{\tau_n\} = n/p$ and its variance is $\mathbf{Var}\{\tau_n\} = nq/p^2$. The generating function of τ_n is

$$\mathbf{E}\{z^{\tau_n}\} = \left(\frac{p}{1 - qz}\right)^n$$

for $|z| < 1/q$.

12.3 De Moivre's Limit Theorem

In 1730, A. De Moivre proved that if

$$a \leq \frac{k - np}{\sqrt{npq}} \leq b,$$

where $[a,b]$ is an arbitrary finite interval, then

$$\mathbf{P}\{\nu_n = k\} \sim \varphi\left(\frac{k - np}{\sqrt{npq}}\right)\frac{1}{\sqrt{npq}},$$

where

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

is the *normal density function.* The symbol ~ means that the left-hand side is asymptotically equal to the right-hand side, that is, the ratio of the two sides tends to 1 as $n \to \infty$. Hence it follows that

$$\lim_{n\to\infty}\mathbf{P}\left\{a \le \frac{\nu_n - np}{\sqrt{npq}} \le b\right\} = \Phi(b) - \Phi(a)$$

for $a \le b$ where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-x^2/2}\,du$$

is the *normal distribution function.*

12.4 The Poisson Distribution

Another limit theorem for the Bernoulli distribution is the Poisson limit theorem. Let us perform n Bernoulli trials with probability $p = p_n$ for success, and denote again by ν_n the number of successes. Then $\mathbf{E}\{\nu_n\} = np_n$. If $\lim_{n\to\infty} np_n = a$ where a is a positive real number, then

$$\lim_{n\to\infty}\mathbf{P}\{\nu_n = k\} = e^{-a}a^k/k!$$

for $k = 0, 1, \ldots$.

If ν is a discrete random variable with possible values $0,1,2,\ldots$, and

$$\mathbf{P}\{\nu = k\} = e^{-a}a^k/k!$$

for $k = 0,1,2,\ldots$ where a is a positive real number, then we say that the random variable ν has a *Poisson distribution* with parameter a. Then $\mathbf{E}\{\nu\} = \mathrm{Var}\{\nu\} = a$ and

$$\mathbf{E}\{|\nu - a|\} = 2s\mathbf{P}\{\nu = s\}$$

where $s = [a] + 1$. The generating function of ν is

$$\mathbf{E}\{z^\nu\} = e^{-a(1-z)}$$

for every z.

12.5 The Hypergeometric Distribution

If ξ denotes the number of white balls in n drawings in example 2 of Sec. 4, then we have

$$\mathbf{P}\{\xi = k\} = \binom{a}{k}\binom{b}{n-k}\bigg/\binom{a+b}{n}$$

for $k = 0, 1, \ldots, n$. In this case, we say that ξ has a *hypergeometric distribution.* Now $\mathbf{E}\{\xi\} = na/(a + b)$,

$$\mathrm{Var}\{\xi\} = \frac{nab(a + b - n)}{(a + b)^2(a + b - 1)}$$

and

$$\mathbf{E}\left\{\left|\xi - \frac{na}{a+b}\right|\right\} = \frac{2s(b + s - n)}{a + b}\mathbf{P}\{\xi = s\},$$

where $s = [na/(a + b)] + 1$.

12.6 The Multinomial Distribution

Finally, we mention the *multinomial distribution.* Consider a sequence of repeated independent trials. Let $A_1, A_2, \ldots, A_m$ be a complete system of events, that is, mutually exclusive and exhaustive events. Let $\mathbf{P}\{A_i\} = p_i$ for $1 \le i \le m$. Then $p_i \ge 0$ and $p_1 + p_2 + \cdots + p_m = 1$. Let us denote by $\nu_i(n)$ the number of occurrences of A_i in the first n trials. Then

$$\mathbf{P}\{\nu_1(n) = k_1, \nu_2(n) = k_2, \ldots, \nu_m(n) = k_m\} = \frac{n!}{k_1!k_2!\cdots k_m!}p_1^{k_1}p_2^{k_2}\cdots p_m^{k_m}$$

if $k_1 + k_2 + \cdots + k_m = n$. In this case, we say that the random variables $\nu_1(n), \nu_2(n), \ldots, \nu_m(n)$ have a multinomial distribution. The binomial distribution is the special case $m = 2$. Now $\mathbf{E}\{\nu_i(n)\} = np_i$, $\mathrm{Var}\{\nu_i(n)\} = np_i(1 - p_i)$, and

$$\mathrm{Cov}\{\nu_i(n),\nu_j(n)\} = -np_ip_j$$

for $i \ne j$.

13. CONTINUOUS DISTRIBUTIONS

13.1 The Exponential Distribution

Let us consider the disintegration of radioactive atoms. Denote by ξ the lifetime of an atom. Suppose that the event that an existing atom disintegrates during a time interval of length t does not depend on the age of the atom. If we write $\mathbf{P}\{\xi > t\} = q(t)$ for $t \geq 0$, then $q(t)$ is a nonincreasing function of t and $q(t + u) = q(t)q(u)$ for $t \geq 0$ and $u \geq 0$. The only solution of this functional equation is $q(t) = e^{-\mu t}$ for $t \geq 0$ where μ is a positive real constant. Accordingly, $\mathbf{P}\{\xi \leq x\} = F(x)$ where

$$F(x) = 1 - e^{-\mu x}$$

for $x > 0$ and $F(x) = 0$ for $x \leq 0$. We say that the random variable ξ has an *exponential distribution function* with parameter μ. Now ξ has a density function $f(x)$ and

$$f(x) = \mu e^{-\mu x}$$

for $x > 0$ and $f(x) = 0$ for $x \leq 0$. We have $\mathbf{E}\{\xi\} = 1/\mu$ and $\mathrm{Var}\{\xi\} = 1/\mu^2$. The characteristic function of ξ is

$$\mathbf{E}\{e^{it\xi}\} = \mu/(\mu - it)$$

for any real t.

13.2 The Gamma Distribution

The gamma function $\Gamma(a)$ is defined for $a > 0$ by the integral

$$\Gamma(a) = \int_0^\infty e^{-x}x^{a-1}dx.$$

We have $\Gamma(a) = (a - 1)\Gamma(a - 1)$ for $a > 1$. In particular, $\Gamma(1) = 1$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

If a random variable η_a has the distribution function

$$G_a(x) = \frac{1}{\Gamma(a)}\int_0^x e^{-u}u^{a-1}du$$

for $x > 0$ and $G_a(x) = 0$ for $x \leq 0$, then we say that η_a has a *gamma distribution function* with parameter a. Then η_a has a density function

$$g_a(x) = \frac{1}{\Gamma(a)}e^{-x}x^{a-1}$$

for $x > 0$ and $g_a(x) = 0$ for $x \leq 0$. We have $\mathbf{E}\{\eta_a\} = \mathrm{Var}\{\eta_a\} = a$. The characteristic function of η_a is

$$\mathbf{E}\{e^{it\eta_a}\} = (1 - it)^{-a}$$

for any real t.

The gamma distribution has the following important property:

$$G_a(x) * G_b(x) = G_{a+b}(x).$$

Accordingly, if we add mutually independent random variables each having a gamma distribution, then the sum also has a gamma distribution. In particular, if we add n independent random variables each having the same exponential distribution function $G_1(x) = 1 - e^{-x}$ for $x > 0$, then the sum has a gamma distribution function with parameter n, that is,

$$G_n(x) = 1 - \sum_{j=0}^{n-1} e^{-x}\frac{x^j}{j!}$$

for $x > 0$, and $G_n(x) = 0$ for $x \leq 0$.

13.3 The Beta Distribution

The beta function $B(a,b)$ is defined for $a > 0$ and $b > 0$ by the integral

$$B(a,b) = \int_0^1 x^{a-1}(1 - x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}.$$

If a random variable $\eta_{a,b}$ has the distribution function

$$F_{a,b}(x) = \frac{1}{B(a,b)}\int_0^x u^{a-1}(1 - u)^{b-1}du$$

for $0 \leq x \leq 1$, $F_{a,b}(x) = 1$ for $x > 1$ and $F_{a,b}(x) = 0$ for $x < 0$, then we say that the random variable $\eta_{a,b}$ has a *beta distribution function* with parameters a and b. We have $\mathbf{E}\{\eta_{a,b}\} = a/(a + b)$ and $\mathbf{E}\{\eta_{a,b}^2\} = a(a + 1)/(a + b)(a + b + 1)$.

13.4 The Normal Distribution

If $\mathbf{P}\{\xi \le x\} = \Phi(x)$ where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du,$$

then we say that ξ has a *standard normal distribution function.* We have $\mathbf{E}\{\xi\} = 0$ and $\operatorname{Var}\{\xi\} = 1$. If $\eta = \sigma\xi + a$ where $\sigma > 0$, then we say that η has a normal distribution of type $N(a,\sigma^2)$. The first parameter is $\mathbf{E}\{\eta\} = a$ and the second parameter is $\operatorname{Var}\{\eta\} = \sigma^2$. The distribution function of η is

$$\mathbf{P}\{\eta \le x\} = \Phi\left(\frac{x-a}{\sigma}\right),$$

and the characteristic function of η is

$$\mathbf{E}\{e^{it\eta}\} = \exp(ita - \sigma^2 t^2/2).$$

If we add n mutually independent random variables having normal distributions of types $N(a_i,\sigma_i^2)$ $(i = 1,2,\ldots,n)$ respectively, then the sum has a normal distribution of type $N(a_1 + a_2 + \cdots + a_n, \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2)$.

13.5 The χ^2 Distribution

[F. R. Helmert (1876) and K. Pearson (1900)]. Let $\xi_1, \xi_2, \ldots, \xi_n$ be mutually independent random variables each having the same standard normal distribution $N(0,1)$. Define

$$\chi_n^2 = \xi_1^2 + \xi_2^2 + \cdots + \xi_n^2.$$

Then $\mathbf{P}\{\chi_n \le x\} = H_n(x)$ is given by

$$H_n(x) = \frac{1}{2^{(n-2)/2}\Gamma(n/2)} \int_0^x e^{-u^2/2}\, u^{n-1}\, du$$

for $x > 0$ and $H_n(x) = 0$ for $x \le 0$. Since $\mathbf{P}\{\xi_i^2/2 \le x\} = G_{1/2}(x)$ for $i = 1, 2, \ldots, n$, we have

$$\mathbf{P}\{\xi_1^2 + \xi_2^2 + \cdots + \xi_n^2 \le x\} = G_{n/2}(x/2),$$

and consequently $H_n(x) = G_{n/2}(x^2/2)$.

13.6 The Multivariate Normal Distribution

Let

$$Q(x_1,x_2,\ldots,x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j$$

be a positive definite quadratic form with coefficient matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix},$$

where $a_{ij} = a_{ji}$. We say that $Q(x_1,x_2,\ldots,x_n)$ is positive definite if $Q(x_1,x_2,\ldots,x_n) \ge 0$ and $Q(x_1,x_2,\ldots,x_n) = 0$ if and only if $x_1 = x_2 = \cdots = x_n = 0$. Then $A = \operatorname{Det}[\mathbf{A}] > 0$. The n-variate function

$$f(x_1,x_2,\ldots,x_n) = \frac{\sqrt{A}}{(2\pi)^{n/2}} e^{-Q(x_1-m_1,x_2-m_2,\ldots,x_n-m_n)/2}$$

is an n-dimensional density function for any choice of $m_1, m_2, \ldots, m_n$. We say that n random variables $\xi_1, \xi_2, \ldots, \xi_n$ have an n-dimensional normal distribution if their joint density function has the above form. In this case, we have $\mathbf{E}\{\xi_i\} = m_i$ for $i = 1, 2, \ldots, n$. If we define $\sigma_{ij} = \mathbf{E}\{(\xi_i - m_i)(\xi_j - m_j)\}$ for $i, j = 1, 2, \ldots, n$, then the covariance matrix

$$\Delta = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix} = \mathbf{A}^{-1};$$

that is, Δ, the covariance matrix, is the inverse of $\mathbf{A}$, the coefficient matrix.

14. LAWS OF LARGE NUMBERS

Let $(\Omega,\mathcal{B},\mathbf{P})$ be a probability space and $\{\eta_n(\omega), n \ge 1\}$ be a sequence of real random variables defined on Ω. Let also $\eta(\omega)$ be a random variable defined on Ω.

We say that the sequence $\{\eta_n\}$ *converges in probability* to the random variable η if for every $\epsilon > 0$ we have

$$\lim_{n\to\infty} \mathbf{P}\{|\eta_n - \eta| < \epsilon\} = 1.$$

Define

$$C = \{\omega: \lim_{n\to\infty} \eta_n(\omega) \text{ exists}\}.$$

The set C represents a random event, that is, $C \in \mathcal{B}$ and consequently, the probability $\mathbf{P}\{C\}$ is determined. If $\mathbf{P}\{C\} = 1$, then we say that the sequence $\{\eta_n(\omega)\}$ converges with probability 1. In this case, there exists a random variable $\eta(\omega)$ such that the event $\{\lim_{n\to\infty}\eta_n(\omega) = \eta(\omega)\}$ has probability 1, and we say that the sequence of random variables $\{\eta_n\}$ *converges with probability* 1 to the random variable η. We can prove that $\eta_n \to \eta$ with probability 1 if and only if for every $\epsilon > 0$ we have

$$\lim_{n\to\infty} \mathbf{P}\{\sup_{k\geq n} |\eta_k - \eta| \geq \epsilon\} = 0.$$

14.1 The Weak Law of Large Numbers

Let $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ be pairwise independent and identically distributed random variables for which $\mathbf{E}\{\xi_k\} = a$ exists. Then $(\xi_1 + \xi_2 + \cdots + \xi_n)/n$ converges in probability to a, that is

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n} - a\right| < \epsilon\right\} = 1$$

for every $\epsilon > 0$.

This theorem was proved by J. Bernoulli around 1680 in the case where $\mathbf{P}\{\xi_n = 1\} = p$, $\mathbf{P}\{\xi_n = 0\} = q$, and $p + q = 1$, and by P. L. Chebyshev in 1867 in the case where $\mathbf{E}\{\xi_k^2\} < \infty$. The above general theorem was proved by A. Ya. Khintchine in 1929.

14.2 The Strong Law of Large Numbers

Let $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ be mutually independent and identically distributed random variables for which $\mathbf{E}\{\xi_k\} = a$ exists. Then $(\xi_1 + \xi_2 + \cdots + \xi_n)/n$ converges with probability 1 to a, that is,

$$\mathbf{P}\left\{\lim_{n\to\infty} \frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n} = a\right\} = 1.$$

In the particular case where $\mathbf{P}\{\xi_n = 1\} = \mathbf{P}\{\xi_n = 0\} = \frac{1}{2}$, this theorem was proved by É. Borel in 1909. In the case where $\mathbf{P}\{\xi_n = 1\} = p$, $\mathbf{P}\{\xi_n = 0\} = q$, and $p + q = 1$, the theorem was proved by F. P. Cantelli in 1917. The above general theorem was proved by A. N. Kolmogorov in 1930.

15. LIMIT DISTRIBUTIONS

The following limit theorem, which G. Pólya (1920) named the *central limit theorem*, is one of the most frequently used theorems in probability theory.

15.1 The Central Limit Theorem

Let $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ be mutually independent and identically distributed random variables such that $\mathbf{E}\{\xi_n\} = a$ and $\mathrm{Var}\{\xi_n\} = \sigma^2$ where σ is a finite positive number. Then

$$\lim_{n\to\infty} \mathbf{P}\{(\xi_1 + \xi_2 + \cdots + \xi_n - na)/\sigma\sqrt{n} \leq x\} = \Phi(x),$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du$$

is the normal distribution function.

In the particular case where $\mathbf{P}\{\xi_n = 1\} = p$, $\mathbf{P}\{\xi_n = 0\} = q$, and $p + q = 1$, the central limit theorem was proved by A. De Moivre (1730) and P. S. Laplace (1812). The above general case was proved by P. L. Chebyshev (1890) and A. A. Markov (1898).

Now let us suppose that $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ are mutually independent random variables and define $\zeta_n = \xi_1 + \xi_2 + \cdots + \xi_n$ for $n \geq 1$. The problem arises under what conditions there exist constants a_n and $b_n > 0$ such that

$$\lim_{n\to\infty} \mathbf{P}\{(\zeta_n - a_n)/b_n \leq x\} = H(x),$$

where $H(x)$ is a nondegenerate distribution function. This problem was solved by B. V. Gnedenko (1939) and W. Doeblin (1941) for identically distributed random variables, and A. Ya. Khintchine (1937), P. Lévy (1937), and B. V. Gnedenko (1939) for the general case.

15.2 The Arcsine Law

In the central limit theorem, the limit distribution is independent of the distribution of the random variables $\{\xi_n\}$. This independence is sometimes described as an invariance principle. There are many other examples where the invariance principle is valid. One such example is the *arcsine law*. Let $\zeta_k = \xi_1 + \xi_2 + \cdots + \xi_k$ for $k \geq 1$ where $\xi_1, \xi_2, \ldots, \xi_k, \ldots$ are mutually independent and identically distributed random variables for which $\mathbf{E}\{\xi_k\} = 0$ and $\mathrm{Var}\{\xi_k\} = 1$. Denote by Δ_n the number of subscripts $k = 1, 2, \ldots, n$ for which $\zeta_k > 0$. Then

$$\lim_{n\to\infty}\mathbf{P}\{\Delta_n/n \leq x\} = (2/\pi)\arcsin\sqrt{x}$$

for $0 \leq x \leq 1$. In the particular case where $\mathbf{P}\{\xi_k = 1\} = \mathbf{P}\{\xi_k = -1\} = \frac{1}{2}$, the theorem was proved by P. Lévy in 1939 and in the general case by P. Erdös and M. Kac in 1947.

Further Reading

Chung, K. L. (1974), *A Course in Probability Theory*, 2nd ed., New York: Academic.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton, NJ: Princeton Univ. Press.

Feller, W. (1966), *An Introduction to Probability Theory and its Applications*, Vol. 2, New York: Wiley.

Feller, W. (1968), *An Introduction to Probability Theory and its Applications*, Vol. 1, 3rd ed., New York: Wiley.

Gnedenko, B. V. (1967), *The Theory of Probability*, 4th ed., New York: Chelsea.

Gnedenko, B. V., Kolmogorov, A. N. (1954), *Limit Distributions for Sums of Independent Random Variables*, Cambridge, MA: Addison-Wesley.

Kolmogorov, A. N. (1950), *Foundations of the Theory of Probability*, New York: Chelsea.

PROPERTIES AND APPLICATIONS OF AMORPHOUS MATERIALS

Toshikazu Shimada, *Central Research Laboratory, Hitachi Ltd., Tokyo, Japan* (Sec. 1)

Ken Takahashi, *Hitachi Research Laboratory, Hitachi Ltd., Ibaraki, Japan* (Sec. 2)

Takashi Sato, *Department of Metallurgical Engineering, Tokyo Institute of Technology, Tokyo, Japan* (Sec. 3)

Yasuo Okazaki, *Steel Research Laboratories, Nippon Steel Corporation, Chiba, Japan* (Sec. 3)

Introduction 72
1. **Amorphous Semiconductors** 72
1.1 Advantages and Disadvantages of Amorphous Materials 73
1.2 Structural, Electrical, and Optical Properties of Amorphous Semiconductors .. 74
1.2.1 Amorphous Si and Tetrahedrally Coordinated Materials 74
1.2.2 Amorphous Se and Chalcogenide Materials 76
1.3 Applications of Amorphous Semiconductors 77
1.3.1 Thin-Film Transistors 77
1.3.2 Solar Cells 79
1.3.2.1 Single-Junction Cells 79
1.3.2.2 Stacked-Junction Cells 80
1.3.3 Two-Dimensional Image Pickup Tubes 81
1.3.3.1 Vidicon-Type *a*-Se Image Pickup Tubes 81
1.3.3.2 Avalanche-Multiplication–Type *a*-Se Image Pickup Tubes 81
1.3.4 Photoreceptors 82
1.3.5 Linear Photosensors and Solid-State Imagers 83
2. **Functional Glasses** 84
2.1 Fundamental Properties of Glasses 84
2.1.1 What Is Glass? 84
2.1.2 Structure and Features of Glasses 85
2.1.3 Synthesis of Glasses 85
2.1.3.1 Unconventional Forming Method 85
2.1.3.2 Deposition Methods 86
2.1.3.3 Sol-Gel Method 86
2.1.3.4 Solid-State Transformation Method 86
2.1.4 Molecular-Dynamics Simulation 86
2.2 Applications of Glasses 87
2.2.1 Function of Glasses 87
2.2.2 Optical Fibers 87
2.2.3 Active Optical Devices 90
2.2.4 Nonlinear Optical Glasses 91
2.2.5 Electronic Substrates 91
2.2.6 Chemically Durable Glasses ... 92
2.2.7 Biomedical Glasses 92
2.3 Polymeric Materials 93
2.3.1 Features of Polymers 93
2.3.2 Applications of Polymers 93
3. **Amorphous Metals** 94
3.1 Magnetic Properties of Rapidly Quenched Metals 94
3.2 Applications of Amorphous Metals 97
3.2.1 Distribution Transformers 97
3.2.2 High-Frequency Power Devices 99
3.2.2.1 Switching-Mode Power Supply 99
3.2.2.2 High-Power Pulse Compressor 100
3.2.3 Sensors 101
3.2.3.1 High-Permeability 102
3.2.3.2 Magnetomechanical Effect 103
3.2.3.3 Large Barkhausen Effect and Matteucci Effect 104
3.2.4 Magnetic Shielding 105
3.2.4.1 Cable and Gauge Shieldings .. 106
3.2.4.2 Composite Shielding Panel 106
3.2.5 Magnetic Separation 106
3.2.6 Medical Uses 107
3.2.6.1 Intratissue Hyperthermia 107
3.2.6.2 Energy-Transferring Coils for Totally Implantable Devices .. 107

digital Encyclopedia of Applied Physics

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

3.3 Final Comments 107
Glossary 108
Works Cited 109
Further Reading 113

INTRODUCTION

In the last thirty years, amorphous materials have been developed that allow new practical applications never before realized by crystalline materials. This rapid expansion of the applicational field originates not only from epoch-making events in basic physics, but also from breakthroughs in technical developments. This article reviews three special areas of new applications: electronic and optoelectronic devices using amorphous semiconductors, functional usages with optical, mechanical, and chemical properties of glasses, and ferromagnetic functions of amorphous metals. In each section, the physical properties of each material are first briefly explained. Next, important applications and their operational principles are described. Section 1 discusses thin-film transistors for liquid-crystal display, solar cells, and imaging devices. Section 2 looks at optical fibers, laser glasses, nonlinear optical glasses, chemical and biomedical glasses, and functional polymers. Finally, in Sec. 3, magnetic cores for distribution transformers, high-frequency power device components, magnetic nonlinear uses, and high-permeability applications are reviewed. Despite recent developments, the progress in fundamental understanding of amorphous materials has been slow because of the lack of the symmetry and long-range order that is found in crystals. Many problems are still to be solved. A continuous effort in fundamental studies and in technical developments is necessary to expand the useful application areas.

1. AMORPHOUS SEMICONDUCTORS

Recently amorphous materials have been receiving a great deal of attention for possible application in various devices as well as for their interesting physical properties. Of particular significance are amorphous thin-film semiconductors such as amorphous Se (*a*-Se) and hydrogenated amorphous Si (*a*-Si:H), which are now opening useful new fields of application in electronics. Liquid-crystal displays (LCD), solar cells, and image-sensing devices are the most important applications for electronic devices at this time.

Thirty years ago, amorphous semiconductors were mainly studied as interesting materials from the purely scientific viewpoint. Great applicative efforts on amorphous semiconductors have taken place since the decade of the 1950s. Pioneer work on an image-sensing device using amorphous Se was done by Weimer and Cope (1951), and the commercial use of image pickup tubes using amorphous Se-based material was successfully reported by Maruyama *et al.* (1974).

In 1975, successful conductivity control over ten orders of magnitude and conduction-type control (of p and n types) using boron and phosphorus doping into amorphous Si prepared by glow-discharge decomposition of a monosilane (SiH_4) source gas were reported by Spear and LeComber (1975, 1976). After the publication of this key report, various applications were attempted that dramatically increased in parallel with studies on basic material properties.

Spear and LeComber also demonstrated that this material could be used in thin-film transistors (TFT) (LeComber *et al.*, 1979). Currently, liquid-crystal displays driven by *a*-Si TFTs are widely used, for example, in flat-panel televisions, laptop computers, car navigation systems, and so on.

The application of hydrogenated amorphous Si to solar calls was first demonstrated by Carlson and Wronski (1976). Over 50% of electronic pocket calculators selling throughout the world are powered by amorphous Si solar cells. Applications are expanding mainly into various wireless and limited-power appliances such as electric power sources for wrist watches, car ventilation fans, and so on.

The high dark resistivity and high photoconductivity of amorphous Si and Se are useful for the photoreceptors in copy machines or laser-beam printers and also for photosensors.

Amorphous semiconductors can be classified into two material groups: tetrahedrally coordinated materials and chalcogenide materials. Almost all the applications for amorphous semiconductors are listed in Table 1. The device applications italicized in the table are major commercial products that are discussed later.

1.1 Advantages and Disadvantages of Amorphous Materials

Amorphous materials have great advantages compared with crystalline semiconductor materials. These are as follows:

1. Readily controlled electronic properties of films over a wide range. In the chalcogenide amorphous group, Se is the principal material and S, Te, Si, Ge, Sb, and As are used mainly as alloying elements to control electronic properties. In tetrahedrally coordinated materials such as hydrogenated amorphous *a*-Si:H, optical and electrical properties can be varied widely by alloying with Ge, C, Sn, N, or O. B or P doping into *a*-Si:H and related alloys allows *p*- or *n*-type control similar to crystalline Si (*c*-Si).
2. Production of the films at low temperatures. In contrast to the growth temperatures over 1000 °C required for *c*-Si, only deposition temperatures up to about 500 °C are required for amorphous Si. Typically, deposition temperature is lower than 100 °C for *a*-Se and 300 °C for *a*-Si:H. Low-temperature deposition permits low interdiffusion of compositional elements and impurities between multilayers, allowing precise control of compositional and doping profiles; the use of various kinds of substrate materials such as glass (Ohnishi *et al.*, 1984), flexible polymer (Okaniwa *et al.*, 1982), and stainless steel (Okamoto *et al.*, 1979; Haruki and Uchida, 1983); and the conservation

Table 1. Electronic-device application of amorphous semiconductors.

Phenomena	How to use the phenomenon	Materials: Chalcogenide	Materials: Tetrahedral	Device applications
Photoconducting effect	Conductivity change by optical absorption	Se, Se–Te, Se–As, Se–As–Te	Si:H, SiC:H, SiN:H, SiO:H, Si:H(B)	*Image pickup tube,*[a] *photoreceptor* for copy machine or laser-beam printer, *linear photosensor,* color sensor, *solid-state imager*
Nonlinear effect	Nonlinear response by light or electric field	Se, Te–Ge–Sb–S, Te–As–Si–Te		*Avalanche-multiplication device,* varistor
Electric-field effect	Conductivity change by varying electric field		Si:H	*Thin-film transistor,* charge-coupled device, static-induction device
Carrier-injection effect	Interaction and/or recombination of electrons and holes		Si:H	Bipolar transistor, light-emitting diode, injection-type photosensor
Photovoltaic effect	Photoinduced carrier separation by *p-i-n,* Schottky, MIS junctions		Si:H, SiGe:H, Si:H(B), μc-Si:H	*Solar cell,* photosensor, pattern sensor
Quantization effect	Tunneling, quantum confinement, Coulomb blockade, carrier separation, low-dimensional effect, etc.		Si:H, SiC:H, SiN:H, SiO:H, μc-Si:H	Solar cells, thin-film transistor, light-emitting diode, Coulomb-blockade device

---

[a] The italicized devices, which are commercially used in specific application fields, are explained in this section.

of heating-up energy required for substrates during deposition.

3. Easily produced thin and multilayered films over a large area. Nonepitaxial deposition of amorphous films can be achieved on various foreign materials. Deposition is typically undertaken by using the thermal or plasma chemical vapor-deposition method [see CVD (CHEMICAL VAPOR DEPOSITION)] from gas-phase materials, or conventional resistive or electron-beam heating evaporation method from solid source materials. Large-area deposition to the order of 0.5 × 0.5 m^2 is now standard technology on commercial production lines.
4. Materials conservation. Amorphous thin-film semiconductors are directly prepared on substrates by using various deposition methods that conserve materials.

Amorphous materials have some disadvantages compared with crystalline materials. These are as follows:

1. Instability of structural and electronic properties. The unstable physical properties of amorphous materials, which have nonequilibrium metastable structures, affect the reliability of devices. With chalcogenide amorphous materials, thermal structural change such as crystallization readily occurs even at temperatures of less than 100 °C. With tetrahedrally coordinated materials, light-induced structural changes brought about by the Staebler–Wronski effect (Staebler and Wronski, 1977, 1980) still remain a problem that needs to be solved. The present reliability of devices is greatly increased through the effective choice of device structures and preparation conditions for materials. Materials having these instabilities are primarily used in new functional devices such as nonvolatile memories.
2. Low carrier mobility. Carrier mobilities of amorphous semiconductors are several orders of magnitude lower than those for crystalline materials, as shown for *a*-Si:H and *c*-Si in Table 2. Although it is difficult to expect these to achieve realizing high speed in electronic devices, amorphous semiconductors are suitable for high-impedance and low–power-dissipation devices such as the switching elements for liquid-crystal displays. With optoelectronic applications such as solar cells and photosensors, the high optical-absorption coefficient compensates for disadvantages of low mobility and short minority-carrier diffusion length.

Table 2. Electronic properties of *c*-Si and *a*-Si:H.

Electronic properties	*c*-Si	*a*-Si:H
Character of optical absorption	Indirect	Quasidirect
Band/optical gap (eV)	1.1	1.65–1.8
Refractive index	3.42	3.6
Electron mobility ($cm^2/V \cdot s$)	~1500	~1
Minority-carrier diffusion length (μm)	~1000	~1
Minimum gap state density (cm^{-3} eV^{-1})	~0	$>10^{15}$

1.2 Structural, Electrical, and Optical Properties of Amorphous Semiconductors

From the viewpoint of electronic applications, the most important materials are *a*-Si:H and its related alloy group. In addition to this, chalcogenide materials such as *a*-Se and its alloys play an important role for those purposes. We concentrate on these materials in this section. The main physical properties of *a*-Se and *a*-Si:H are listed in Table 3.

1.2.1 Amorphous Si and Tetrahedrally Coordinated Materials The local atomic-bonding structure or the so-called short-range order (SRO) of *a*-Si is based on the covalent bonding of an sp^3-type configuration. The hybridized eightfold sp^3 valence-electron orbitals of a Si atom are split into four occupied bonding and four unoccupied antibonding orbitals in a solid, having strong directionality and making valence and conduction bands, respectively. The lack of a long-range order (LRO) in the atomic-bonding structure erases the sharp Van Hove singularities and eliminates the wave vector as a good quantum number in the material. The strong directionality of four sp^3 covalent bonds and the lack of LRO induce bond-length and -angle deviations, and dangling bonds. In other words, medium-range order

Table 3. Physical properties of *a*-Se and *a*-Si:H.

	a-Se	*a*-Si:H
Chemical bonding structure	Twofold coordinated chainlike structure	Fourfold coordinated three-dimensional structure
Hardness (Vickers, kg/mm^2)	~30	~2000
Optical gap (eV)	~2.0 (nonbonding–antibonding)	~1.7 (bonding–antibonding)
Dark conductivity (S/cm)	~10^{-15}	~10^{-10} ~10^{-13} (B compensated)
Mobility (cm^2/V·s) Electron	~5 × 10^{-3}	~1
Mobility (cm^2/V·s) Hole	~0.2	~10^{-2}
Maximum storage temperature (°C)	60	200–300

(MRO) characterizes the various properties of amorphous materials. The bond-length and -angle deviations induce band-edge tails, in which carrier mobilities are very low. The dangling bonds introduce midgap states, which act as traps and recombination centers. This prevents Fermi-level control by impurity doping.

In contrast to *a*-Si, *a*-Si:H contains hydrogen from 5 to 20 at.% bonded to Si. The Si–H bonds primarily relax strain in the material and passivate the dangling bonds. The reduction of strain reduces band-edge tailing. The midgap nonbonding states vanish on account of the passivation of dangling bonds. Namely, passivation pushes the antibonding states up above the conduction-band edge and the bonding states down below the valence-band edge, because of the strong bonding energy of Si–H compared with that of Si–Si. As a result, sharper band edges, wider band gap, and lower gap-state density are achieved in an *a*-Si:H network.

The electronic structure of *a*-Si:H is schematically shown in Fig. 1. Since *a*-Si:H has a lower gap-state density (down to 10^{15} cm^{-3} eV^{-1}), Spear and LeComber (1975) were able to achieve *p*-type or *n*-type valence control by PH_3 and B_2H_6 gas-phase doping during deposition, as shown by Fig. 2. Figure 2(a) shows room-temperature conductivity as a function of doping gas ratio, and Fig. 2(b) plots the activation energy of conductivity. Undoped *a*-Si:H shows weak *n*-type characteristics. Consequently, a small amount of B doping creates an electronically intrinsic *a*-Si:H.

In highly doped *p*- and *n*-*a*-Si:H, the state density around the Fermi level and the gap-state density increase greatly. Good *p*-*n* junction characteristics have not been obtained yet. The *p*-*i*-*n* junction shows fine rectifying characteristics as illustrated in Fig. 3.

Even with *a*-Si:H, the effect of wave vector **k**, which is not a good quantum number, appears during the interband optical-absorption and emission processes. The optical-absorption coefficient in the indirect transition region between 1.8 and 3.4 eV of *c*-Si is in the order of ~10^4/cm. The *a*-Si:H, however, has direct transition characteristics in which the optical-absorption coefficient is about one order higher than that of the *c*-Si shown in Fig. 4.

In order for *c*-Si to be used in optoelectronic devices such as solar cells and photosensors in the visible region, a thickness of 10 μm is necessary to absorb incident photons. However, an *a*-Si:H thickness of about 1 μm is sufficient to absorb incident photons. Figure 5 shows the Si-content dependence of optical gaps for *a*-Si(Ge or C):H al-

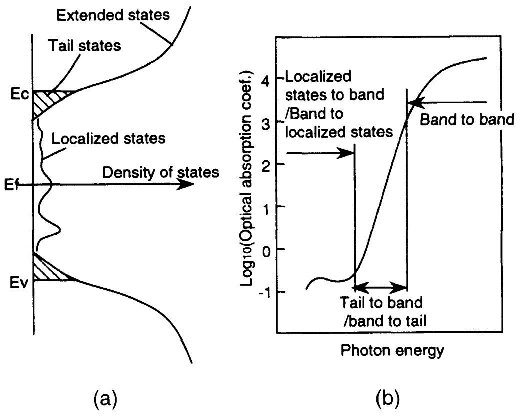
Extended states
Tail states
Ec
Localized states
Density of states
Ef
Ev
(a)
Log10(Optical absorption coef.)
4
3
2
1
0
-1
Localized states to band /Band to localized states
Band to band
Tail to band /band to tail
Photon energy
(b)

FIG. 1. (a) Electronic density of states and (b) optical-absorption spectrum of amorphous semiconductors.

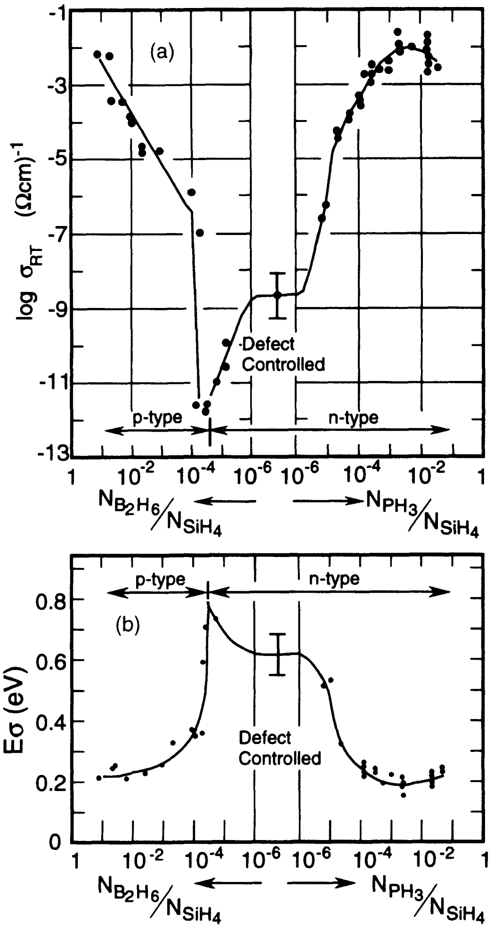

FIG. 2. *p*- or *n*-type doping characteristics of *a*-Si:H (Spear and LeComber, 1975).

loys (Anderson and Spear, 1977; Shimada *et al.*, 1979; Morimoto *et al.*, 1982), which exhibit quasidirect optical-transition characteristics. Alloying with Ge or C allows us to achieve systematic control of optical gaps.

1.2.2 Amorphous Se and Chalcogenide Materials The local atomic-bonding structure of amorphous Se is based on covalent bonding in the p^4 valence-electron configuration. This has a chainlike structure. The hybridized six *p* orbitals are split into two occupied bonding and two occupied nonbonding orbitals and two unoccupied antibonding orbitals in the solid. The occupied nonbonding and the unoccupied antibonding orbitals make valence and conduction bands, respectively. The splitting between nonbonding and antibonding states forms an energy gap.

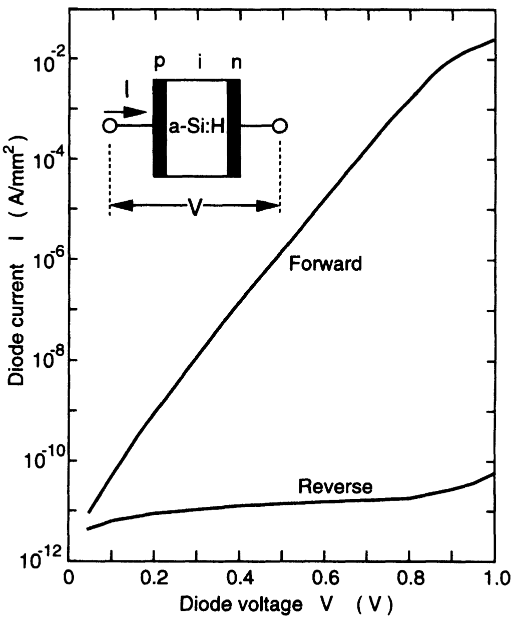

FIG. 3. Voltage vs current of an *a*-Si:H *p*-*i*-*n* diode.

The chainlike structure originating from the two occupied bonding orbitals is flexible compared with the tetrahedrally bonded structure of *a*-Si:H. As a result, very low gap-state density can be achieved in *a*-Se without the use of any strain-relaxation procedures such as the hydrogenation required for *a*-Si.

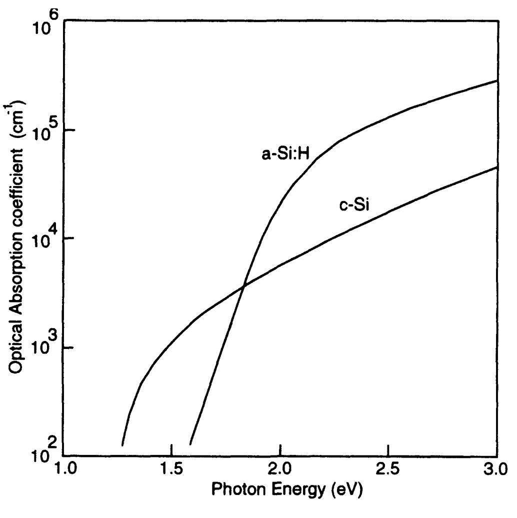

FIG. 4. Optical-absorption spectra of *a*-Si:H and *c*-Si.

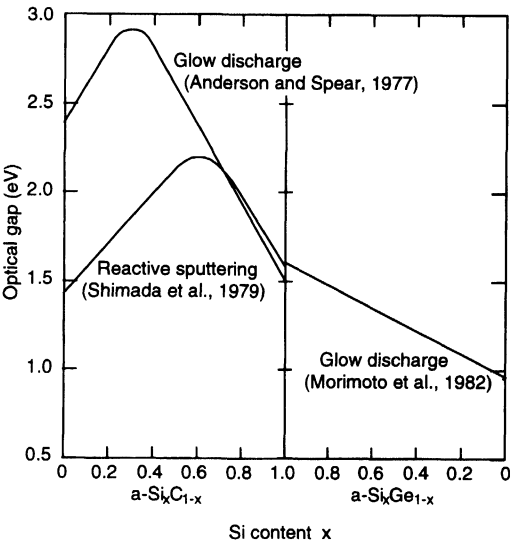

FIG. 5. Si-content dependence of optical gaps of *a*-Si(Ge or C):H alloys (Anderson and Spear, 1977; Shimada *et al.*, 1979; Morimoto *et al.*, 1982).

In contrast to *a*-Si:H, *p*, *n* doping cannot be performed in *a*-Se. This limits the various electronic applications—for example, use in solar cells, *p*-*n* or *p*-*i*-*n* photodiodes, and so on. Amorphous Se has high photoconductive yield, up to 1, and this is a solitary instance in which clear avalanche-multiplication effects have been observed in amorphous materials (Juska and Arlauskas, 1980; Tanioka *et al.*, 1987; Tsuji *et al.*, 1994). Image pickup device application by use of the avalanche-multiplication effect is one application.

1.3 Applications of Amorphous Semiconductors

1.3.1 Thin-Film Transistors In 1962, thin-film transistors (TFTs) using CdSe were reported for the first time (Weimer, 1962). Stable fabrication processes and operation of TFTs, however, could not be realized until the discovery of *a*-Si:H and the means of *p*-, *n*-doping control (Spear and LeComber, 1975, 1976). An *a*-Si:H TFT is useful to cover large areas and for use in high-impedance circuits such as the switching transistors in liquid-crystal displays (LeComber *et al.*, 1979; Powell, 1989; Tsumura *et al.*, 1991). Figure 6 shows a cross-sectional view of a typical inverted stagger-type *a*-Si:H TFT. The carrier channel is situated in the interface between the *a*-Si:H and gate insulator. Source and drain electrodes have been formed on the opposite side of the channel. This is the origin of the name "inverted stagger-type," in contrast to the *c*-Si MOSFET (metal-oxide-semiconductor field-effect transistor) structure shown in the figure for comparison.

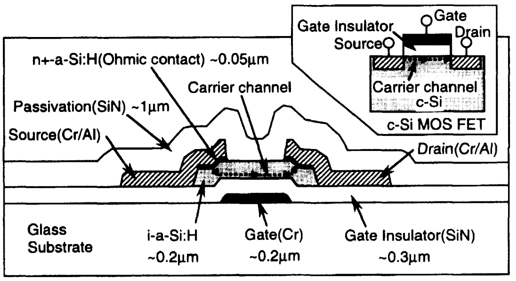

FIG. 6. Cross-sectional structure of an inverted stagger-type *a*-Si:H TFT. Inset: A *c*-Si MOSFET shown for comparison.

Figure 7 shows gate-voltage V_G vs drain-current I_D characteristics of an *a*-Si:H TFT.

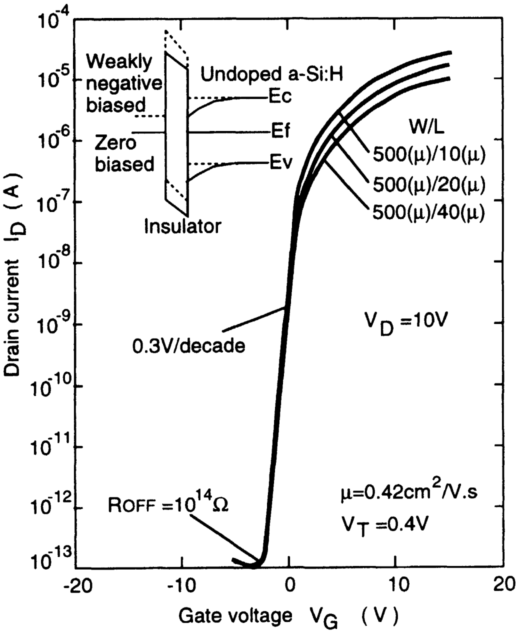

FIG. 7. Gate-voltage–drain-current characteristics of *a*-Si:H TFT. *W*, gate width; *L*, gate length; V_D, drain voltage; *m*, carrier mobility; V_T, threshold voltage; R_{OFF}, "off" resistance; E_C/E_V, conduction-/valence-band edges; E_F, Fermi energy.

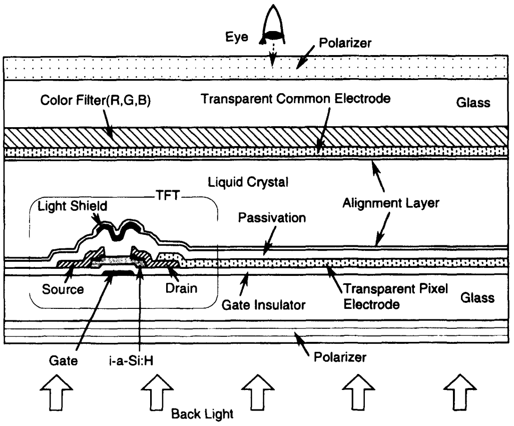
Eye
Polarizer
Color Filter(R,G,B)
Transparent Common Electrode
Glass
Liquid Crystal
TFT
Light Shield
Alignment Layer
Passivation
Source
Drain
Transparent Pixel Electrode
Gate Insulator
Glass
Gate
i-a-Si:H
Polarizer
Back Light

FIG. 8. Cross-sectional view of a color TFT/LCD element.

On application of a weak negative bias voltage to the gate, I_D falls to on the order of 10^{-13} A. This corresponds to the estimated value of 10^{-13} A obtained from the intrinsic resistivity of *a*-Si:H in the flat-band scheme (shown by the dotted line in the inset).

The *a*-Si TFT and liquid-crystal optical switch are combined in the single pixel shown in Fig. 8. They are arranged in two dimensions as shown by Fig. 9 using an equivalent circuit. The operation of this device is as follows. White and unpolarized light arrives from the bottom and is polarized by the bottom polarizer. When the TFT is in the "off" state, molecules are aligned parallel to the alignment layers and perpendicular to one another as shown in Fig. 10(a). The direction of polarization of the incident light is rotated 90°. As a result, the light is blocked by the top polarizer [Fig. 10(a)]. When the TFT is in the "on" state, molecules are aligned vertically to the alignment layers. The polarized light can go through the top polarizer [Fig. 10(b)]. Sets of three color pixels display color dots. Each pixel is connected as shown by Fig. 9.

Initially, the first series of TFTs connected to the line of the first gate are in an "on" state. Data are sent in parallel to the series of TFTs from the data-input bus line. Next, the second series of TFTs is in an "on" state, and data are written in the same manner. The storage capacitors maintain the written data until the next scan.

A *c*-Si LSI is used for the scanning transistors at this time. To reduce the cost of TFT/LCDs, a high-mobility TFT using poly-Si is expected to be used commercially instead of the *c*-Si LSI presently used for the scanner and the *a*-Si:H TFT presently used for matrix switches.

The electron channel at the insulator–*a*-Si:H interface is opened by applying positive bias voltage to the gate. The current I_D increases exponentially and tends toward the relation $I_D = (W/2L)\mu C_{\text{INS}}(V_G - V_T)^2$. It is finally limited by the series resistance of the device. In the above, W and L are the channel width and length, respectively; μ is the field-effect electron mobility in the channel, V_T is the threshold voltage, and C_{INS} is the gate-insulator capacitance per unit area. The

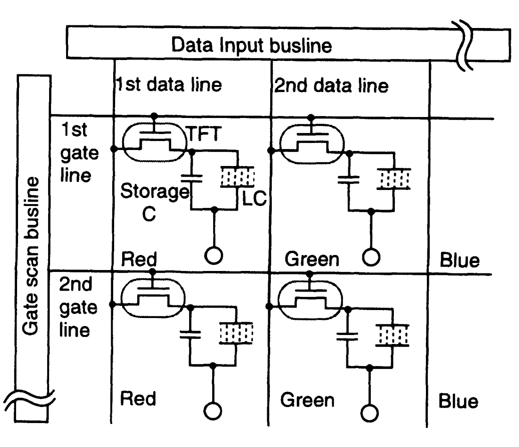
Data Input busline
1st data line
2nd data line
Gate scan busline
1st gate line
TFT
Storage C
LC
Red
Green
Blue
2nd gate line
Red
Green
Blue

FIG. 9. Matrix circuit of color TFT/LCD.

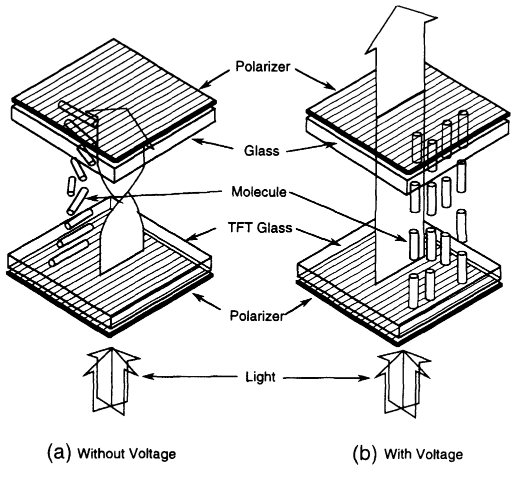
Polarizer
Glass
Molecule
TFT Glass
Polarizer
Light
(a) Without Voltage
(b) With Voltage

FIG. 10. Schematic explanation of LCD operation.

estimated values of μ and V_T in this case are 0.42 cm^2/V·s and 0.4 V, respectively. The on-off ratio of I_D at V_G = −3 and 10 V exceeds 10^8, which is a value large enough for a high-definition LCD.

1.3.2 Solar Cells

1.3.2.1 Single-Junction Cells. Carlson and Wronski (1976) first reported on a Pt/*a*-Si:H Schottky-type solar cell with a conversion efficiency of 5%. A *p-i-n*–type *a*-Si:H solar cell, which is the standard structure for *a*-Si:H solar cells, was developed by the members of Prof. Hamakawa's group at Osaka University (Okamoto *et al.*, 1979). Their group also developed an *a*-SiC/*a*-Si heterojunction solar cell having a conversion efficiency of more than 8% (Tawada *et al.*, 1982). The Taiyoyuden and Electrotechnical Laboratory group showed that the employment of a textured transparent conductive oxide (TCO) electrode is effective in increasing conversion efficiency by means of optical confinement. The rough surface scattering of incident light by the so-called "textured surface" reduces optical-reflection loss in a wide range of light wavelength compared with an optically flat surface. Recently, conversion efficiencies of about 13% have been reported for the *a*-SiC/*a*-Si heterojunction solar cell (Miyachi *et al.*, 1992; Sawada *et al.*, 1993).

A typical *p-i-n* heterojunction solar cell with a glass/textured SnO$_2$/*p*-*a*-SiC/*i*-*a*-Si/μc-Si/Ag structure is schematically shown in Fig. 11(a), and its potential distribution is shown in Fig. 11(b). Because of the short minority-carrier diffusion length in *a*-Si:H compared with that of crystalline Si and the high gap-state density of heavily doped *a*-Si:H, a *p-i-n* junction has been used instead of a *p-n* junction.

Conversion efficiency η is defined as shown in Fig. 12, where V_{OC}, J_{SC}, V_m, and J_m are open-circuit voltage, short-circuit current density, and voltage and current at maximum power output, respectively; *FF* is fill factor, and P_{in} is incident-light power density.

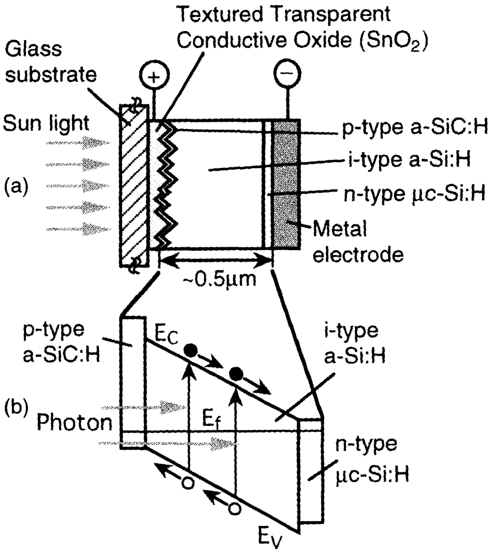

FIG. 11. (a) Cross-sectional structure of single-junction *a*-SiC:H/*a*-Si:H heterostructure solar cells and (b) potential distribution.

In a crystalline-material solar cell such as a *c*-Si cell, the diffusion-type photovoltaic effect is the essential operation mechanism. In contrast to this, the mechanism converting sunlight to electric power in *a*-Si:H solar cells is based on the drift-type photovoltaic effect, in which the generated photocarriers are transported by the internal field. The internal electric field in an *a*-Si:H solar cell is very sensitive to change as a function of gap-state density and photoinduced space-charge distributions.

The drift length of photogenerated carriers is strongly dependent on electric field and comparable to an *a*-Si thickness of 0.1–1 μm that absorbs sunlight effectively. The *p* and *n* layers only act as pinning for the Fermi level to establish an internal electric field, but do not contribute to photocarrier

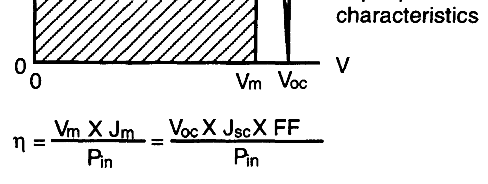

$$\eta = \frac{V_m \times J_m}{P_{in}} = \frac{V_{oc} \times J_{sc} \times FF}{P_{in}}$$

FIG. 12. Definition of output characteristics of solar cells.

generation directly. It is very important to achieve large internal electric-potential difference and optimized internal electric-field distribution in the cell.

In this cell structure, B-doped *a*-SiC:H is used for the *p* layer, which increases the internal electric field and acts as an optical window layer for the *i* layer. The textured TCO and metal back layer are used for electrodes as well as for the optical-confinement layers of incident light.

1.3.2.2 Stacked-Junction Cells. A stacked-junction cell structure has been introduced to increase the conversion efficiency. Each cell can separately convert a certain range in the spectrum of incident sunlight. The cell is fabricated by using various optical-gap materials—for example, *a*-SiGe:H and narrow-gap *a*-Si with a low H content of less than 5 at.%. A primary concern, to obtain higher efficiency in the stacked cell, is to develop high-quality narrow–optical-gap materials.

An *a*-SiGe:H alloy is a promising candidate in the manufacture of narrow-gap materials. At this stage, however, the quality of this alloy is poor compared with that of *a*-Si:H. The preferential attachment of H to Si and Ge and the microscopic structural fluctuations of Si and Ge compositions seem to degrade the quality of the materials. In other words, structural fluctuation in the medium-range order (MRO) apparently prevents the preparation of high-quality *a*-SiGe:H.

MRO in *a*-Si:H and *a*-SiGe:H has been investigated using small-angle x-ray scattering (SAXS) measurement with the aid of the thermal-evolution measurement of hydrogen and IR spectroscopy (Muramatsu, 1989, 1991). Results obtained by SAXS suggest that hydrogenated amorphous silicon contains compositional fluctuations of Si and/or voids. In addition to this, the MRO in *a*-SiGe:H alloys is strongly affected by the clustering of Ge atoms in the alloys.

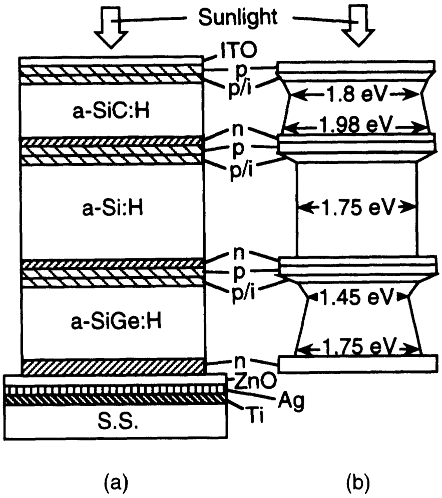

FIG. 13. (a) *a*-SiC/*a*-Si/*a*-SiGe stacked solar cell and (b) its band-gap profile (Nakata *et al.*, 1990b).

A microwave CVD method was introduced to reduce the preferential attachment of H and structural fluctuation in the medium range (Watanabe, 1987). This method has high decomposition and ionization yields for SiH_4 and GeH_4 source gases. By optimizing deposition conditions, minimum gap-state density N_d can be less than 10^{15} cm^{-3} eV^{-1} in *a*-SiGe:H with an optical gap of 1.5 eV. Such an optical gap is comparable to that realized in *a*-Si:H.

The stacked three-junction structure cell is a most promising structure with both high stability and high conversion efficiency (Nakata *et al.*, 1990a). Inside each subcell, precise band-gap profiling is useful in obtaining more efficient and more stable solar cells. By application of the band-gap profiling technique, *a*-SiC/*a*-Si/*a*-SiGe stacked cells were fabricated by Nakata *et al.* (1990b) as shown in Fig. 13(a). The band-gap profile is in Fig. 13(b). The solar-cell characteristics are shown in Fig. 14. By use of this cell structure, light-induced degradation is greatly suppressed. An expected degradation ratio of 14% in the first year was estimated from an

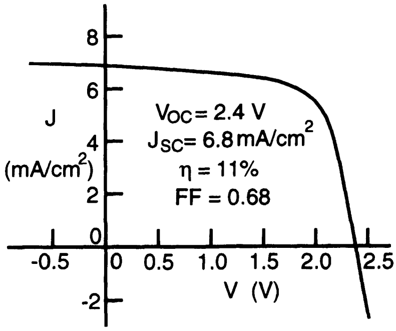

FIG. 14. Output characteristics of the stacked solar cell (Nakata *et al.*, 1990b).

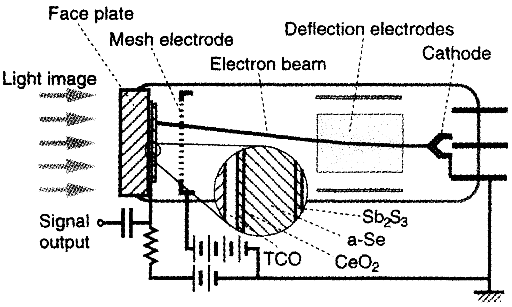

FIG. 15. Schematic illustration of Vidicon-type *a*-Se image pickup tube.

acceleration test. Ichikawa *et al.* (1992) also reported a total-area efficiency of 10.5% in a 30 × 40-cm² tandem cell with a large surface area.

1.3.3 Two-Dimensional Image Pickup Tubes

1.3.3.1 Vidicon-Type* a*-Se Image Pickup Tubes. Application of *a*-Se to image pickup tubes was first attempted by Weimer and Cope (1951). Commercial use was achieved by Maruyama *et al.* (1974). The cross-sectional structure of an *a*-Se image pickup tube is schematically shown in Fig. 15. Two-dimensional light images are converted into time-sequential electric signals by electron-beam scanning. The electric charge on a photoconductive *a*-Se film deposited by electron-beam scanning is discharged by incident light. Recharging occurs at the next electron-beam scan. Current for recharging is from electric-signal output.

Figure 16 shows the structure of a photoconductive *a*-Se film in detail (Shidara *et al.*, 1981). Elimination of Te in the first layer maintains hole-blocking characteristics, which prevents hole injection from the transparent conductive oxide (TCO) electrode. Te in the second layer enhances red-light sensitivity. Graded doping of As in the third layer provides uniform field distribution, contributing to the effective collection of photocarriers. The small amount of As in the *a*-Se prevents crystallization of the film. The fourth layer achieves small-device capacitance. The fifth layer, of porous Sb_2S_3, acts as the beam-landing layer to suppress secondary electron emission as well as acting as the electron-blocking layer to prevent injection of electrons into the *a*-Se. The thin CeO_2 layer, 20 nm, inserted between the TCO (made of indium tin oxide, ITO) and the first layer also suppresses injection of holes from the TCO and crystallization of the *a*-Se layer. Instead of *a*-Se, the use of *a*-Si:H has also been attempted for image pickup tubes (Imamura, 1979).

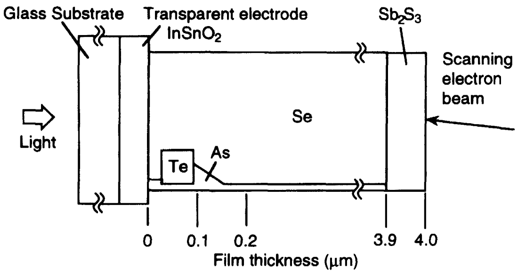

FIG. 16. Photoconductive-layer composition of an *a*-Se image pickup tube (Shidara *et al.*, 1981).

1.3.3.2 Avalanche-Multiplication–Type* a*-Se Image Pickup Tubes. In 1987, Tanioka *et al.* achieved steady and uniform avalanche multiplication of photocurrent in an *a*-Se thin film (Tanioka *et al.*, 1987). An ultrahigh-sensitive image pickup tube was successfully fabricated using this *a*-Se film as the photoconductive target. This was the first imaging device to use the avalanche multiplication of photocarriers in amorphous materials.

The avalanche-multiplication effect is very widespread in crystalline materials—for example, Si, Ge, and GaAs. An avalanche photodiode is already commercially available for high-speed photodetection in optical communication systems. However, only a few studies have been conducted on the effect in amorphous materials. Only one pioneer work suggesting this effect was reported by Juska and Arlauskas (1980), and this was studied in detail by Tsuji *et al.* (1987, 1991).

The structure of an avalanche-multiplication–type *a*-Se image pickup tube is essentially the same as that in Fig. 15. The amorphous Se-layer thickness, up to 25 μm, exceeds the 4 μm in a conventional *a*-Se vidicon tube. The applied target voltage, up to 2500 V, is extremely high compared with the approximately 50 V used in conventional ones.

The operation mechanism of this tube is shown in Fig. 17. The electron-blocking layer is charged by landed electrons when the

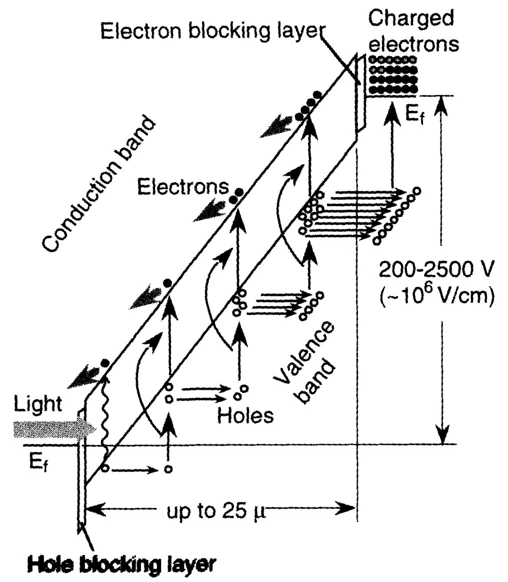

FIG. 17. Operation mechanism of avalanche-multiplication–type image pickup tube.

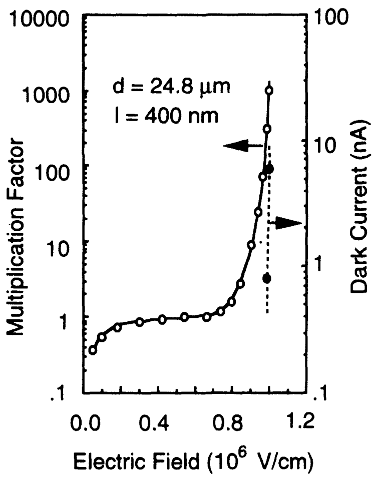

FIG. 18. Electric-field dependence of avalanche-multiplication factor and dark current in the tube (Tsuji *et al.*, 1991).

scanning electron beam just passes through this point in the target area. Incident light creates electron–hole pairs. Electrons and holes are accelerated by the electric field in opposite directions, and these create new electron–hole pairs by collision. The creation of new electron–hole pairs is called the avalanche-multiplication effect, and it happens easily with holes as compared with electrons in *a*-Se. Only avalanche-multiplication events initialized by holes are shown in the figure. The holes reaching the electron-blocking layer recombine with the charged electrons (recombined electrons are shown by gray circles), reducing the surface potential of the target. When the electron beam comes through this point during the next scanning cycle, the number of lost electrons is restored. This restoring current comes out as the signal.

Figure 18 shows the electric-field dependence of the avalanche-multiplication factor for 400-nm light in a tube with an *a*-Se thickness of 25 μm and a dark current as the reference. Ultrahigh sensitivity with an equivalent quantum efficiency larger than 1000 was obtained by Tsuji *et al.* (1991).

In addition to this, excess noise induced by avalanche multiplication is greatly reduced because of the large ratio of the creation rate of electron–hole pairs created by one hole collision to that by one electron collision. The effects of negative feedback due to the operation mode of charge storage for this device also contribute to reductions in excess noise.

This tube can have a high sensitivity to the ultraviolet or x-ray wavelength regions by changing the window material. In principle, it is also applicable to solid-state imaging devices.

1.3.4 Photoreceptors Amorphous Se and *a*-Si:H were successfully used in photoreceptors for an electrophotographic copy machine and a laser-beam printer. The photoreceptor consisted of an *a*-Se(Te) or *a*-Si:H photoconductive film deposited onto an aluminum cylinder.

Figure 19 explains the operational steps (so-called Carlson process: Carlson, 1938) for the laser-beam printer using the amorphous photoreceptor (positively charged case). Step 1: The surface of the drum is charged up to about 1000 V by the corona-discharging method. Step 2: The drum surface is scanned using laser light modulated by a pattern signal. A charge pattern is formed on the drum surface by the recombination of photocarriers with surface charges. Step 3: Fine carbon particles with thermally adhesive materials ("toner") are deposited on the surface by electrostatic force. Step 4: The patterned

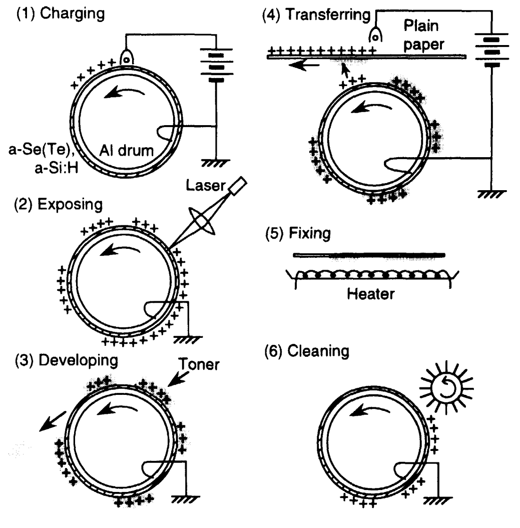
(1) Charging
a-Se(Te),
a-Si:H
Al drum
(2) Exposing
Laser
(3) Developing
Toner
(4) Transferring
Plain paper
(5) Fixing
Heater
(6) Cleaning

FIG. 19. Operation steps of laser-beam printer using amorphous photoreceptor (Carlson process).

toner is transferred onto plain paper electrostatically. Step 5: The patterned toner on the paper is fixed by heating. Step 6: The drum surface is cleaned up. The operation step returns to step 1.

An *a*-Se photoreceptor is used in a plain-paper copy machine. Various types of *a*-Si:H–based photoconductive layers have been reported (Oda *et al.*, 1981; Inoue and Shimizu, 1982; Yamamoto *et al.*, 1981; Kawamura *et al.*, 1983). The structure of a multilayered *a*-Si:H photoreceptor is shown in Fig. 20. Incident light is mainly absorbed in the carrier-generation layer and forms electron–hole pairs. In the case of positive charge, photoelectrons compensate for surface positive charge and holes running through the carrier-transport layer to the Al conductive substrate.

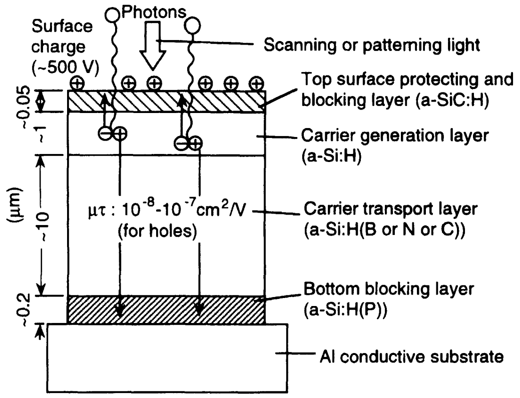
Surface charge (~500 V)
Photons
Scanning or patterning light
Top surface protecting and blocking layer (a-SiC:H)
Carrier generation layer (a-Si:H)
μτ : 10⁻⁸-10⁻⁷cm²/V (for holes)
Carrier transport layer (a-Si:H(B or N or C))
Bottom blocking layer (a-Si:H(P))
Al conductive substrate
~0.05
~1
~10
~0.2
(μm)

FIG. 20. Sample structure of a multilayered *a*-Si:H photoreceptor.

Photoconductivity gains up to 0.8 are obtained in a wavelength range of 400–800 nm. The top and bottom blocking layers prevent the injection of surface holes and electrons into the respective carrier-transport layers. The high resistivity of the top blocking layer maintains the electrostatic latent image and the toner image on the surface.

Recently, however, organic photoconductive materials have mainly been used for photoreceptors. Consequently, performance and durability have advanced remarkably and production costs have been greatly reduced. Development of a low-cost production process is desirable for amorphous photoreceptors.

1.3.5 Linear Photosensors and Solid-State Imagers Amorphous Si:H contact-type linear photosensors are widely used in facsimile machines to read documents. Cassanhiol *et al.* (1974) first reported the non-storage mode of operation of a contact-type linear photosensor using chalcogenide material. Kagawa *et al.* (1982) fabricated *a*-Si:H photosensors. The sensor consisted of 1000–6000 photoconductive elements. The elements were spaced 200–250 mm in a line for standard A4-size facsimile machines. Several thousand terminals for photoconductive elements have to be connected to addressing LSI chips.

Addressing *a*-Si:H TFTs are integrated on glass substrates using *a*-Si:H photodiodes to reduce the number of terminals. Figure 21 shows the equivalent circuit of a matrix-scanning–type integrated linear sensor with 2560 elements with 104 terminals. The 2560 elements are divided into 40 blocks. Each block operating in the parallel storage mode is read out and the readout converted into serial signals by an analog-multiplexer LSI. An analog multiplexer has been used because of the low switching speed of TFTs.

Various types of matrix-scanning integrated linear sensors have been reported (Credelle, 1988; Kobayashi *et al.*, 1991; Miyake *et al.*, 1991). These linear sensors are used mostly with lens arrays or lensless contact schemes using illumination light

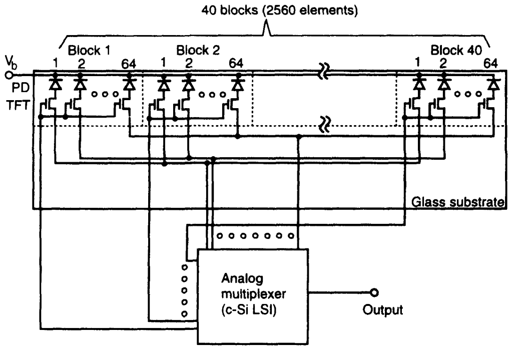

FIG. 21. Equivalent circuit of matrix-scanning line sensor (Miyake *et al.*, 1991).

sources. Current commercially available facsimile machines mainly employ this type of sensing unit.

An *a*-Se or an *a*-Si:H photoconductive film is integrated onto a *c*-Si MOS- or CCD-type two-dimensional scanner to obtain a highly sensitive and compact solid-state imager (Tsukada *et al.*, 1979; Baji *et al.*, 1982; Sasaki *et al.*, 1990; Manabe *et al.*, 1991). Recently, two-dimensional solid-state color imagers were made using *a*-Si:H photosensors and *a*-Si:H TFTs (Weisfield, 1993), and using *a*-Si:H phototransistors (Yamaguchi *et al.*, 1993). More intensive study and development, however, are expected to allow for actual commercial use.

2. FUNCTIONAL GLASSES

2.1 Fundamental Properties of Glasses

2.1.1 What Is Glass? Glass has long been defined as an inorganic product of fusion that has cooled to a rigid condition without crystallizing. This is true for the glasses used for windows, bottles, lenses, lamp bulbs, etc. However, along with other amorphous materials, glasses can now be made by deposition from the vapor phase, by sputtering in a low-pressure system, or through chemical processes—that is, by methods other than cooling a melt. So a more accurate definition of glass is an amorphous material that exhibits the glass transition. The glass transition is a phenomenon by which a solid amorphous phase exhibits a more or less sudden change in the derivative thermodynamic properties with changing temperature; these properties may include heat capacity and expansion coefficient, which change from crystal-like to liquidlike values (Rawson, 1980).

Figure 22 shows the volume change of a liquid, a supercooled liquid, glasses, and a crystal with temperature. Crystallization of the melt on cooling is usually accompanied by a marked decrease in volume at the melting point T_m. If the crystallization rate is low enough, the melt can be cooled below the melting point without crystallization taking place. This viscous liquid below the melting point is a supercooled liquid. No sudden volume change occurs when the melt is supercooled. The volume decreases along the same line with the liquid according to the change in the structure of the melt, reflecting the equilibrium configuration at a particular temperature. Further cooling results in the viscosity rising to as high a value as 10^{12} Pa s so that no further such changes are possible. Then a decrease in slope is found in the volume–temperature curve. With a further drop of temperature, the decrease in volume is due almost entirely to the decreasing amplitude of the atomic vibrations (Rawson, 1980; Boyd and Thompson, 1985).

The change in slope with temperature is characteristic of a glass as described in its definition. The temperature at which the change in slope occurs is called the glass-

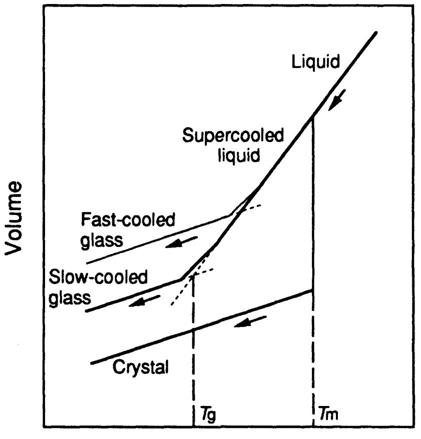

FIG. 22. Volume change of a liquid, a supercooled liquid, glasses, and a crystal with temperature. T_m, melting point; T_g, glass-transition temperature.

transition temperature T_g, and below T_g the material is described as a glass. Thus among amorphous materials, only those that exhibit the glass transition are glasses.

The change from supercooled liquid to glass is not a sudden one, and so T_g is not a well-defined temperature for any particular glass. The temperature at which the change in slope occurs decreases as the cooling rate is decreased. The structure that is frozen in during the glass transition persists at lower temperatures. Thus the structure of a glass depends on the cooling rate.

2.1.2 Structure and Features of Glasses

Most glasses are composed of metal–oxygen (or nitrogen, halogen, or chalcogen) bondings. In silicate glasses, the basic structural unit is the silicon–oxygen tetrahedron in which a silicon atom is tetrahedrally coordinated to four surrounding oxygen atoms. These tetrahedra are joined together at their corners. The structural unit is almost the same as that of crystalline silica. However, the bonding features between silicon–oxygen tetrahedra in glasses are different from those in the crystalline material, as shown in Fig. 23. Structure of glasses is treated in separate articles in the Encyclopedia in connection with experimental and theoretical techniques for structure determination (see AMORPHOUS MATERIALS, STRUCTURE OF; GLASSES).

The random-network theory of glass structure proposes that the atoms form a three-dimensional connected structure without periodic order and with energy content comparable to that of the corresponding crystalline material. The structure of silicate glasses described above is in accordance with this theory. In reality, the situation is not likely to be so simple, however, and the random-network hypothesis cannot be applied to every inorganic glass-forming material. Conceptions of chemical and structural inhomogeneities of glasses have been proposed (Porai-Koshits, 1990).

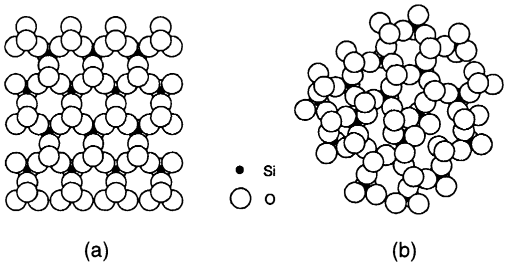

FIG. 23. Structure of SiO_2 in (a) crystalline and (b) glassy forms.

The similarity in the structural unit between the glass and crystalline forms results in similarities in optical, electrical, and magnetic properties. A glass exhibits most of the features peculiar to the crystalline form with the same composition, but less remarkably. Only in rare cases does a glass exhibit more remarkable features than the crystalline form—for example, in chemical resistance. Glasses can be made by combinations of almost any of the elements in the periodic table and with nonstoichiometric compositions. Consequently, the number of kinds of glass is infinite and many properties are controlled continuously over a wide range.

Another important feature of glass from the standpoint of practical application is that glass is more easily formed to any desired shape and to large-size components than ceramics or single crystals, resulting in lower machining costs. Moreover, glass is chemically stable compared with steel or plastics and more stable to temperature changes than plastics. The main elements constituting glass are generally among the most common elements on Earth, and suitable for industrial purposes.

2.1.3 Synthesis of Glasses

Most glass articles are manufactured by a process in which raw materials are melted at high temperatures and then formed into the articles (Boyd and Thompson, 1985). Recent demands for new high-performance glasses, however, have led to new preparation methods. Soga (1990) has reviewed many of the new glass-forming methods with an emphasis on their advantages for preparing the new types of glasses over the conventional melting method.

2.1.3.1 Unconventional Forming Method.

Rapid quenching is a new glass-forming method in which a liquid is rapidly cooled below the glass-transition temperature without the occurrence of crystallization. Rapid quenching is achieved by quenching the liquid into a bath of water, mercury, or liquid nitrogen, splat cooling onto a cold plate, smashing a molten droplet with a fast-moving cold piston, or dropping the liquid between two cold rotating rollers. In the case of two cold rotating rollers with a rotating speed of several thousand rpm, a cooling rate of 10^7 K/s can be realized. This method has allowed a variety of glasses to be gotten with new compositions, including the Bi–Sr–Ca–Cu–O system noted for its high-T_c superconducting property.

2.1.3.2 Deposition Methods. Evaporation is an old and well-established nonreactive deposition method involving only physical processes. In place of the usual electrical heating in metal boats or coils, electron-beam heating is used to vaporize a raw material. An example is deposition of antireflection coatings of Ta_2O_5 or SnO.

In the sputtering process (see SPUTTERING), atoms ejected by the starting material target when hit by energetic ions or charged particles deposit on the substrate. To obtain an amorphous film, the substrate temperature is kept low enough to avoid crystallization of the deposited film. Radio-frequency (rf) sputtering is commonly used for inorganic materials. A high-rate magnetron sputtering apparatus that combines magnetic and electric fields is effective to achieve a high sputtering rate.

Chemical vapor deposition [see CVD (CHEMICAL VAPOR DEPOSITION)] involves chemical reactions and is an effective method to produce bulk glasses. Vaporizable–decomposable organometallic or halide compounds are oxidized in a flame or a glow-discharge plasma to produce ultrafine amorphous particles, which are deposited on the solid surface, forming a high-purity glass soot. The soot is then sintered to make a solid, inclusion-free, glass body. This method has been explored extensively as a method to fabricate high-purity optical-waveguide fibers.

An advantage of deposition from a vapor phase is the possibility of making a glass that cannot be realized by any melting method. One example is aluminophosphate glass, which is difficult to make by any melting process, because Al_2O_3 and P_2O_5 have quite different melting temperatures and vapor pressures.

Finally, the thermal-oxidation process is a variation of the reactive deposition method, in which a metal surface, such as Si, is oxidized in an oxygen-containing atmosphere at elevated temperatures to form an amorphous oxide film.

2.1.3.3 Sol-Gel Method. In the sol-gel method, a glass is formed from a liquid solution. A sol of chemically polymerized organometallic compounds or a colloidal solution of metallic salts is transformed into a gel through chemical reaction at low temperatures. By heat treatment of the gel at temperatures lower than the usual melting temperature of that material, a glass with no detectable difference from the melt-driven glass is obtained.

The advantages of the sol-gel method are

1. preparation of the glass at lower temperatures than usual;
2. high homogeneity (components can be mixed at an atomic or molecular level);
3. synthesis of new kinds of multicomponent glasses with novel functions, which otherwise cannot be obtained because of crystallization and phase-separation tendencies, or the requirement of very high melting temperatures; and
4. the possibility of doping organic compounds or making organic–inorganic hybrid materials.

Long processing times are the most serious disadvantage of the sol-gel method. Photoirradiation has been proved to enhance the reaction effectively in the sol-gel process and to lower the temperature and shorten the time for the heat treatment (Ohishi *et al.*, 1992).

2.1.3.4 Solid-State Transformation Method. Extra energy given to a crystalline solid through impacting bodies or particles destroys the lattice structure at low temperatures and produces amorphous materials. This transformation occurs upon neutron bombardment in a nuclear reactor, exposure to shock waves as in a meteorite impact, and ion implantation.

2.1.4 Molecular-Dynamics Simulation

Soga (1990) has treated molecular-dynamics simulation as a virtual method for preparing glasses, because it can be a tool for determining whether some glasses can be made in a given system.

Molecular-dynamics simulation can provide a realistic model of glass structure when suitable pair potentials are used. Each atom in the system is assumed to take a random configuration corresponding to the Boltzmann distribution of velocities at a temperature high enough to form a liquid state. Then, the temperature is decreased by reducing the average kinetic energy stepwise or suddenly to room temperature. By selecting a slow cooling rate, a glass can be simulated having a real quenching rate. This requires a large number of steps and extremely long computing times. Unrealistically high quenching rates can also be simulated by this method.

2.2 Applications of Glasses

2.2.1 Function of Glasses New types of glasses with active functions completely different from those of conventional glasses, such as containers or building materials, are often called new glasses (Sakka, 1987; Tomozawa, 1987). New glasses are contributing to marked progress in advanced technologies, such as information and communications technology, electronics, energy technology, and biotechnology. Sophisticated research work is required to elucidate the mechanism involved in various functions of new glasses and to find applications for them. The development of new glasses is important also from the standpoint of glass industries. Glass containers (bottles) have been replaced by plastic, metallic, and paper containers. Their production is decreasing every year. Eyeglass lenses are also commonly plastic lenses nowadays. Glass industries have to introduce new glasses and new products in order to survive.

New glasses involve

1. glasses;
2. amorphous materials [typical elemental semiconductors (Si and Ge), metals, carbons, and high polymers are excluded];
3. crystalline materials made from glasses and amorphous materials; and
4. composite materials based on glass, amorphous materials, or crystals obtained from them.

New glasses are classified by areas of function: optical function, electronic function, thermal and mechanical function, chemical function, and biochemical function (Sakka, 1987). Examples of glasses and applications are shown in Table 4.

2.2.2 Optical Fibers Glasses have been used as a medium for transmission, reflection, refraction, absorption, and dispersion of light for centuries, and optical functions are still the most important and useful. Formability into fibers and capability to change properties such as refractive index continuously by incorporating additives into the structure are advantages of glasses over crystals for application to transmitting light. Glasses are used in the form of fibers or planar waveguides for this purpose.

Light-wave communications via glass fibers is replacing electronic message transmission on wires. The frequencies of visible or infrared radiation used in optical communications are vastly higher than those of the electronic carrier used in wire transmissions. Since the information-carrying capacity increases with the carrier frequency, the potential for light-wave communications is enormous. So optical fibers are indispensable for construction of multimedia infrastructures (see FIBER OPTICS; OPTICAL COMMUNICATIONS).

Developments have been made of low-loss glass fibers for light-wave communication systems mainly using silica glass. A glass fiber consists of a core of very high transparency and a transparent cladding layer of lower refractive index surrounding the core. The refractive-index difference between the core and cladding is about 0.01. The structures of optical fibers are classified into three types according to refractive-index profiles as shown in Fig. 24 (Pearson *et al.*, 1985; MacChesney and DiGiovanni, 1991). These fiber diameters are of the order of 100 μm. In the single-mode fiber, the core diameter is 1 to 5 μm. In the multimode fiber, the core diameter is over 30 μm, typically 60 μm. The fiber is called multimode since there are many paths the light may follow. The multimode fiber was designed for use with an incoherent light-emitting diode (LED) source, and the single-mode fiber with an injection laser emitting coherent light.

In stepped-index fibers, the core and the cladding have uniform refractive indices and the change in the refractive index at the core–cladding interface is an abrupt one. Mode dispersion broadens the light pulse injected into the core of a multimode stepped-index fiber. All the rays of different mode orders travel at the same speed, but on

Table 4. Classification of new glasses and their applications.

Glasses	Applications
Optical functions	
Silica-based optical fibers	Low-loss light-transmission fibers
Fluoride-glass optical fibers	Ultralow-loss light-transmission fibers
Chalcogenide-glass optical fibers	IR-transmission fibers
Neodymium-doped laser glass	Laser nuclear fusion
Erbium-doped glass fiber	Optical amplifiers
Nonlinear optical glasses	Optical switches, memories, modulators
Glasses for optical waveguides	Optical integrated circuits
Photochromic glasses	Sunglasses
Faraday rotator glasses	Optical isolators
Glasses with selective transmission or reflection of light	Heat-radiation–shield glasses, mirrors, nonreflecting glasses
Gradient-index glasses	GRIN lenses
Electronic functions	
Super ion-conducting glasses	Solid cells
Photoconducting chalcogenide glasses	Pickup tubes
Ultrasonic delay line glasses	TV sets, video apparatus
Substrate glasses	Microelectronics packaging
Electrochromic glasses	Displays
Thermal and mechanical functions	
Low-expansion glasses	Photomask substrate
Low-temperature–melting glasses	Bonding, sealing
High–elastic-modulus oxynitride glasses	Reinforcement
Machinable glass ceramics	Insulators, structural materials
Chemical functions	
Chemically durable glasses	Radioactive-waste solidification
Alkaline-resistant glasses	Glass-reinforced cement
Porous glasses	Supporting enzymes or catalysts Gas-molecule separation
Biochemical functions	
Bioactive glass ceramics	Artificial bone, tooth root

different paths with different lengths, and the time width of the pulse increases as the light travels down the fiber core. This broadening effect would cause the trailing edge of the pulse to overlap the leading edge of the next pulse, and the information being transmitted would be lost. To decrease mode dispersion of multimode fibers, the graded-index fiber configuration is used. In such a structure, the refractive index of the core in-

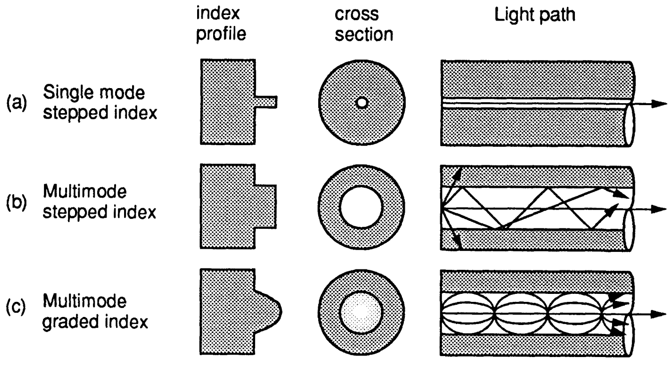
Refractive index profile
Fiber cross section
Light path
(a) Single mode stepped index
(b) Multimode stepped index
(c) Multimode graded index

FIG. 24. Structures and light paths of optical fibers with different refractive-index profiles (MacChesney and DiGiovanni, 1991).

creases gradually from its periphery to the center as shown in Fig. 24(c), so that the speed of the light that travels a longer distance is higher and the modes can arrive at the end of the fiber almost simultaneously.

The wavelength range of the light used in light-wave communications is 0.8 to 2 μm. The most important characteristic of optical fibers is the attenuation or loss in this wavelength range. Optical loss L, measured in terms of dB/km, is expressed as $L = (10/l) \times \log(I_0/I)$, where I_0 is the input intensity and I is the intensity measured after a distance l. Low loss of 0.2 dB/km at a wavelength of 1.55 μm has been achieved by silica-glass fiber. This means attenuation of 4.5% in traveling a distance of 1 km. An example is shown in Fig. 25 (Snitzer, 1989).

Optical loss at short wavelengths is caused by ultraviolet absorption due to electronic transitions of the glass cations; at longer wavelengths it is caused by infrared absorption due to molecular vibrations. Rayleigh scattering is caused by compositional fluctuations and density fluctuations on a scale smaller than the wavelength of the light being used, and causes loss increasing at shorter wavelengths.

Other losses are superimposed on these intrinsic loss mechanisms (Pearson *et al.*, 1985). Material absorption is primarily caused by transition-metal ions: iron, cobalt, nickel, copper, chromium, manganese, and vanadium. A few tens of parts per billion of these metals dissolved in a soda-lime silicate glass will cause enough absorption to preclude lightguide use. Water can also be a serious contaminant. Macroscopic inhomogeneities such as bubbles and crystallites with dimensions that are large compared with the wavelength give rise to light scattering. The structure of the lightguide affects the lightwave transmission and becomes a cause of loss. For example, irregularities at the core–cladding interface might result in some of the light striking the interface at an angle such that it is lost into the cladding.

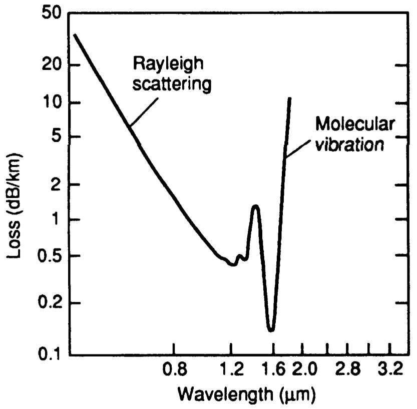

FIG. 25. Optical loss of low-loss silica fibers and its causes (Snitzer, 1989). At shorter wavelengths, loss is caused by electronic transitions.

Fiber-fabrication technologies aim at producing a fiber having the highest purity to achieve low losses and with a precise index profile to achieve low dispersion at the particular wavelength of operation (MacChesney and DiGiovanni, 1991).

Multicomponent glass fibers are fabricated by a double-crucible technique. In this process an ingenious continuous-casting system composed of concentric platinum crucibles is used. A thin stream of core glass flows from the upper crucible, passes through the reservoir of cladding glass, and is concentrically surrounded by the cladding.

Low-loss silica fibers are fabricated by vapor-deposition techniques, using oxidation of silicon tetrachloride vapor to produce submicron amorphous silica particles and of other chloride vapors such as germanium tetrachloride and phosphorus oxychloride as sources of dopants in the silica. There are inside and outside processes. For inside processes, modified chemical vapor deposition (MCVD) or plasma chemical vapor deposition (PCVD) is used to deposit doped silica layers on the inside wall of a silica tube. The tube is collapsed to a rod and drawn into fiber.

There are two versions of outside processes. One is outside vapor deposition (OVD) and the other is vertical-axial deposition (VAD). In the former, a soot of SiO_2 particles is deposited layer by layer on a horizontal, rotating mandrel. A core is deposited first, followed by a cladding. After the deposition, the mandrel is removed and the preformed tube is sintered to form vitreous silica. Then it is drawn to fiber. The central hole is collapsed during either the sintering or the fiber drawing.

In the VAD process, a porous soot cylinder without a central hole is formed by depositing the core on the end of the cylinder and cladding around the cylinder simultaneously using two torches. The core profile is controlled by precise control of the gaseous constituents in the torch flame producing the growing soot face. When the deposition is completed, the body is sintered and drawn to fiber.

Several attempts have been made to develop optical fibers with transmission characteristics superior to the silica-glass fiber. Heavy-metal fluoride glasses are expected to have lower intrinsic optical attenuation than a silica fiber (MacChesney and DiGiovanni, 1991; Lucas, 1991). Fluoride glasses based upon the ZrF_4–BaF_2 system—for example, ZBLA (ZrF_4–BaF_2–LaF_3–AlF_3) or ZBLAN (ZrF_4–BaF_2–LaF_3–AlF_3–NaF) glasses—are being developed. Fluoride glasses are unstable and not easy to fabricate. Bi_2O_3–PbO–based heavy-metal oxide glasses having high infrared transmission characteristics with better stability have been developed. Chalcogenide glasses, such as the GeSbSe system, are also expected to be low-loss glasses with high infrared transmission (Elliott, 1991). As a practical matter, however, these glasses require development of longer-wavelength light sources and detectors to take full advantage of the lowest-loss region at 2 μm, so that silica glass still holds an advantage.

2.2.3 Active Optical Devices Glass lasers are being developed and utilized for high-power applications, because it is easy to produce various shapes and sizes of laser material with high optical quality. The laser glass has been used in forms ranging from fibers to disks with diameters tens of centimeters across. Large pieces of laser glass are used for the amplifier disks in laser-induced fusion facilities. It would be enormously expensive and probably technically impossible to make those with single crystals of similar size. The function of glass lasers is based on doping with rare-earth luminescent ions. Rare-earth ions used include Nd, Yb, Ho, Er, and Tm. The Nd ion is practically useful because its light emission takes place at room temperature and the absorption bands spread over wavelengths ranging from the visible to near-IR regions.

Figure 26 shows the energy levels of the laser glass. Pumping light excites the active ions in the glass to some upper energy level. The radiative transition back to the ground state yields a photon. This radiative transition is either spontaneous or stimulated. Stimulated emission is in phase with the signal photon promoting deexcitation. A laser is made by placing semitransparent mirrors and promoting stimulated emissions all in phase continuously between them. The spectral range of glass lasers extends from about 0.5 to 3.0 μm. Wide–band-gap fluoride glasses or fused silica are used as hosts for the ultraviolet emission and heavy-metal halide or chalcogenide glasses as hosts for the infrared. Because the optical spectra of ions in glass are inhomogeneously broadened, glass lasers provide a degree of tunability ($\leq 10\%$) that is not possible in crystalline hosts (Weber, 1990, 1991; MacChesney and DiGiovanni, 1991).

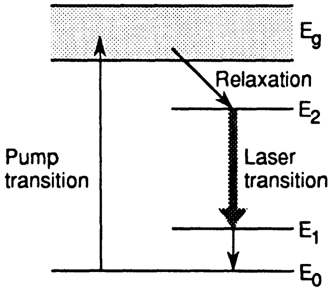

FIG. 26. Energy levels and transitions of laser glass.

New developments of rare-earth–doped glass are as fiber amplifiers and planar-waveguide optical amplifiers. A single-mode fiber core doped with ions of a rare earth and pumped with (injected by) light of an appropriate wavelength emits photons stimulated by the incident signal, and it can act as an amplifying device. Fiber amplifiers are of particular interest for applications in telecommunications, because they can amplify signal input directly without transforming it to an electrical one and reconverting it to an optical one after electrical amplifying. Optical amplification was successfully demonstrated in erbium-doped silica-fiber amplifiers, which allow amplification of signals in the range of 1.5–1.6 μm, a very important region for communication systems. Figure 27 illustrates the basic configuration of such an amplifier. See AMPLIFIERS, OPTICAL.

An Er-doped planar-waveguide amplifier was also reported (Kitagawa *et al.*, 1993). It has an Er concentration of 0.47% and exhib-

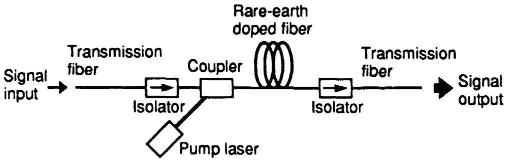

FIG. 27. Rare-earth–doped fiber amplifier configuration.

its a gain of 24 dB for signal wavelength of 1.533 μm. Planar-waveguide amplifiers should contribute to the development of planar light-wave circuits and optical integrated circuits.

The introduction of a large amount of a rare earth into the now available low-loss oxide fibers is difficult. New processes and nonoxide glasses such as fluoride fibers are being considered.

2.2.4 Nonlinear Optical Glasses Nonlinear glasses have attracted attention because of their potential application in all-optical devices.

High intensity of radiation induces a large polarization P in a medium, which is not linear in the external optical electric field E, but expressed as its power series. Nonlinear parts of this series lead to many desirable effects such as second-harmonic generation, an intensity-dependent refractive index, and optical bistability. These effects allow an optical signal to affect the medium, and thereby its own propagation or the propagation of other optical signals, so that it can be utilized for optical switches, memories, or modulators. These devices are key components of planar light-wave circuits.

Glasses with small values of the linear refractive index and dispersion generally have small nonlinearities. Simple SiO_2 has the smallest linear and nonlinear refractive indices of the oxide glasses. These values increase as modifier ions are added (Weber, 1991).

Another approach to produce highly nonlinear glasses is based upon composites with glass matrix and semiconductor dispersed particles. In semiconductor-doped glasses, the nonlinear behavior appears as intensity-dependent changes in the absorption and refractive indices. These effects originate from the generation of a short-lived free-electron plasma and band-filling effects. The size of the semiconductor microcrystals in the glass affects the linear and nonlinear optical properties through quantum confinement effects in the potential accompanying the particles. Typical semiconductor-doped glasses include I–VII [CuCl, CuBr, Ag(Cl,Br,I)] compounds, II–VI [CdS, CdSe, CdS_xSe_{1-x}, CdTe] compounds, and III–V compounds (Weber, 1991). Semiconductor-doped glasses have nonlinear refractive-index coefficient values 10^5 times higher than silicate glasses (Tomozawa, 1987).

2.2.5 Electronic Substrates Thermal, mechanical, and electronic functions of glasses allow their utilization for various kinds of substrates, which is becoming the most important usage in the electronics field. Glasses are used as substrates for flat-panel displays (liquid-crystal display, plasma display panel, electroluminescence display, and electrochromic display), for disks (magnetic disk, optical disk, and magneto-optical disk), for solar cells, for photomasks, and for microelectronics packaging.

Requirements as microelectronics packaging coming from developments in semiconductor technology are, for example, a lower dielectric constant of substrate material, low-temperature processing, a thermal-expansion rate match to silicon, improved thermal conductivity, multilayer processing, and high mechanical strength. These objectives have been achieved by the development of various low-temperature, cofirable glass ceramics and glass-plus-filler composites. Metals such as copper, gold, and silver can be used as conductors in the low-temperature cofired systems around 900 °C. These metals have approximately one-third the resistivity of the tungsten or molybdenum metals used in high-temperature alumina-based systems (Geiger, 1990).

Table 5 compares low-temperature cofired glass ceramics with high-temperature alumina ceramics. Glass ceramics are derived through glass crystallization. Glass-plus-filler composites are formed by admixing filler materials such as granular ceramics and metals into a glass. The filler acts not only to toughen the material or to increase the strength, but also to control physical properties such as thermal expansion (Nishigaki and Fukuta, 1989; Niwa *et al.*, 1989).

Figure 28 shows a schematic diagram of a

Table 5. Comparison of low-temperature cofired glass ceramics with high-temperature alumina ceramics.

Property	Glass ceramics	Alumina
Sintering temperatures (°C)	800–1100	1500–1600
Cofired conductors	Cu, Ag, Ag/Pd, Au	W, Mo, Mo–Mn
Flexural strength (MPa)	150–300	280–420
Thermal conductivity (W/m·K)	2–10	20–40
Dielectric constant (at 1 MHz)	4–8	9–10
Thermal expansion ($\times 10^{-7}$/°C)^a	30–80	65–75

^{a}Si: 35

typical low-temperature cofired multilayer process using the green-sheet technology. A three-dimensional electrical circuit is formed inside the multilayer substrate with layer-to-layer conductive connections through via holes (Geiger, 1990). Several tens of layers are laminated and as many as 100 LSI chips are mounted for multilayer substrate modules used in mainframe computer systems (Tummala, 1992).

2.2.6 Chemically Durable Glasses Glasses are relatively inactive chemically compared with plastics and metals. For applications such as physicochemical apparatus and medical appliances, the chemical durability of glasses is sufficiently high.

Long-term chemical durability is the major concern in high-level radioactive-waste management (Tomozawa, 1987). The most popular waste-solidification method is vitrification. Glass for the high-level radioactive-waste solidification is required to be stable in structure at 500–600 °C under high radiation exposure and resistant against leaching to water or, when used with cement, to alkaline aqueous solution formed by reaction between water and cement, which causes migration of nuclides from the glass to the underground water. Glass compositions including borosilicate glass, aluminosilicate glass, and phosphate glass have been investigated, and, currently, borosilicate glass is being most widely investigated. Borosilicate glasses for this application have 40–50 mass % SiO_2 and melt at 1100–1200 °C, because several toxic materials volatilize during melting above 1300 °C.

2.2.7 Biomedical Glasses Some glasses and glass ceramics can be used as implants for hard tissue (bones and teeth). Desirable properties of glass and glass ceramics as biomaterials are these:

1. bioactivity, i.e., the property to bond with hard tissue (bones) of the human body;
2. biocompatibility, i.e., the biologically compatible intergrowth of the biomaterial and the tissue of the human body without a rejection reaction or possible toxic effect;
3. higher mechanical strength than living bones and long-term stability in living bodies;
4. ease of forming and machining;
5. ease of sterilization.

The hard tissue of human bodies is constituted of apatite [$Ca_{10}(PO_4)_6(OH)_2$]. An example of bioactive glass is based on the $Na_2O–CaO–P_2O_5–SiO_2$ system. This glass forms calcium-phosphate layers and apatite

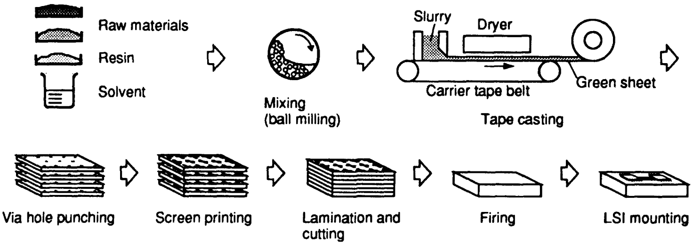

FIG. 28. Schematic diagram of the typical fabrication process for a low-temperature cofired multilayer substrate.

in the living body in reaction with the fluid and bonds to the bones.

On the other hand, bioactive glass ceramics are based on parent glasses yielding apatite on a suitable crystallization treatment, so that they contain apatite crystals already in the base material. It is possible in glass ceramics to control the apatite crystallization and the formation of other crystal phases in order to develop materials having a combination of special properties. Vogel and Höland (1990) developed glass ceramics in the SiO_2–Al_2O_3–MgO–CaO–Na_2O–K_2O–F–P_2O_5 system. This system precipitates apatite and mica/phlogopite crystals simultaneously and, consequently, bioactive and machinable glass ceramics are produced. They have long-term stability and can be machined with standard metal-working tools by a surgeon, if necessary during an operation. Living bones and apatite crystals have flexural strength of 100–200 MPa, and bioactive glass ceramics of 100–600 MPa.

Bioactive machinable glass ceramics have widely and successfully been applied in the middle ear, nose, and jaw and have been implanted as thoracic vertebra substitutes or cervical spine substitutes and as tooth root implants.

2.3 Polymeric Materials

2.3.1 Features of Polymers There are two types of polymers: thermoplastic polymers and thermosetting polymers. Thermoplastic polymers melt on heating. Cooling of the melt causes crystallization in some polymers (crystalline polymers) such as polyethylene. Some regions remain amorphous, however, and further cooling reduces these regions to glass states through the glass-transition temperature. Other polymers such as polystyrene cannot be crystallized even partially and form amorphous solids through glass transition upon cooling from the melt (noncrystalline or amorphous polymers). Properties of polymers change drastically at the glass-transition point in the same manner as with inorganic-glass materials.

Thermosetting polymers have a network structure with a large number of cross-linking points on the polymer chains. Prepolymers are the materials before the cross-linking reaction and they are liquid at room temperature. They become hard glassy plastics through cross-linking reactions upon heating and do not melt with further heating, but proceed to harden until decomposed.

2.3.2 Applications of Polymers Some polymers resemble inorganic-glass materials in properties such as hardness, transparency of light, electrical insulation, and ease of forming. Others have completely different features from glasses. For example, a wide range of ductility and elasticity allows several hundred percent of stretching and restoration. The variation in dynamic properties of polymers comes from differences in the states of aggregation and thermal agitation of molecular chains as well as in the molecular structure of the polymer chains themselves. This variety in characteristics has made polymers very useful in many fields. Only a part of electronics and photonics applications is outlined below.

The main application of electrical properties of polymers is as insulators. Polymers such as phenol resin have been used as insulators for a long time, and polymers are now indispensable to the development of new microelectronics.

In the fabrication of integrated circuits, polymers sensitive to UV light, electrons, x-rays, and ions are used as resists in the lithographic process. The change in solubility rate with exposure allows developer to remove either the exposed (positive tone) or unexposed (negative tone) areas of the resist and a developed image is produced. The developed image is transferred into the semiconductor substrate by using resists as etching masks. Solubility characteristics, thermal stability, and chemical-etch resistance of the materials are key requirements (Watts, 1987). Polyimides and polyimidelike materials are widely used for photoresists.

A continuing increase in the scale of integration in microelectronics requires line-width control of smaller than 1 μm. In place of conventional photolithography, which lacks the resolution to meet these requirements, alternative lithographic technologies, such as electron-beam lithography, short-wavelength (deep UV) photolithography, and x-ray and ion-beam lithography are under development. Each technology requires resist materials specifically designed and optimized for the particular radiation being applied.

Polyimides are also used as a protective overcoat (passivant), providing a barrier for transport of chemical species that could attack the underlying device, and as an interlevel insulator film for multilevel interconnecting systems. For these usages, the material must be an excellent electrical insulator and must adhere well to the substrate. In addition, planarity and a good etching property to form fine via holes for reliable interconnections between the metal layers above and beneath the film (the "via contact") are required of interlevel insulator films (Senturia, 1987).

Plastic packaging of microelectronics devices provides lower-cost, nonhermetic packages. Epoxy encapsulation compounds are used to protect chips during handling and subsequent assembling. Multilayer printed circuit boards are manufactured by laminating circuitized epoxy laminates containing glass fabric (Oka *et al.*, 1992; Higgins, 1992).

Conducting polymers have been developed and are expected to find application as electrodes or electromagnetic interference shielding (Baker, 1987).

In the application field of photonic process and devices, polymers are used as protective coatings for silica optical fibers. The coating affects the fiber strength and transmission loss. UV-curable epoxy acrylates or urethane acrylates are used for in-line application processes as the fiber is drawn and nylon or polyester elastomers are used for off-line application processes (Blyler *et al.*, 1987). Polymeric optical fibers such as those of deuterated polymethyl methacrylate (PMMA) are useful for local communication loops and links. In optical-disk technology, polymers are not only applied as inexpensive substrates, but also used as passive layers (e.g., antiscratch, defocusing overcoats) and as recording layers. PMMA and polycarbonate (PC) are used for disk substrates. Moreover, polymer glasses, such as PMMA and polystyrene, have potential use as nonlinear optical materials for integrated optics (Sohn *et al.*, 1987).

3. AMORPHOUS METALS

The pioneers in the development of rapidly quenched metals were Pol Duwez and his group in 1960 (Klement *et al.*, 1960). They obtained an Au–Si alloy in an amorphous state by plunging a droplet of it onto a Cu substrate. The first product was in the form of a flake. A prototype method for producing an amorphous ribbon continuously was developed by Pond and Maddin (1969). This technology, called centrifugal melt spinning, contributed greatly to accelerating progress in the study of amorphous metals. The most refined method is planar flow casting (PFC), developed by Narasimhan (1980). In PFC, the molten alloy is ejected onto the outer surface of a rotating wheel through a slotted nozzle placed close to the top of the wheel. Today, most wide ribbons are produced by this method.

The characteristic features of amorphous metals stem from the random structures in atomic arrangement, similar to that in liquids. The lack of long-range order endows amorphous metals with various unique properties, such as soft magnetic behavior, high mechanical strength, high corrosion resistance, and combinations of these, which have not been easily realized in crystalline metals.

These unique properties provide a variety of applications. This section focuses on the magnetic properties, because most applications of amorphous metals currently are connected with the magnetic properties. In respect to types of materials, only rapidly quenched amorphous-metal ribbons are dealt with. Even though amorphous metals can also be produced by other techniques, such as gaseous-phase deposition, electrodeposition and electroless deposition, and ion implantation, the rapid quenching from the liquid phase is the most mass productive and therefore the least expensive. Nanocrystalline metals are also discussed as among the rapidly quenched metals, because they are obtained by annealing of amorphous metals.

3.1 Magnetic Properties of Rapidly Quenched Metals

The amorphous metals are commonly very soft magnetically. The magnetic softness—that is, low coercivity and high permeability—results from the absence of magnetocrystalline anisotropy, grain boundaries, and secondary phases. Also, their high electrical resistivity and thinness of the ribbon contribute to reducing energy wasted when they are used in connection with an alternating field.

On the other hand, amorphous metals have a deficiency of low-saturation magnetic induction. In this respect, the addition of alloying elements and the change in electronic structure are concerned. Amorphous metals made by rapid quenching generally contain 1–25 at.% metalloid elements—B, Si, P, and C—or transition-metal elements, typically like Zr, Ti, or Hf. These elements are added to enhance glass-forming ability and to stabilize amorphous phases. However, the addition of glass formers not only dilutes the concentration of ferromagnetic atoms, but also reduces the magnetic moment per atom because of a change in electronic structure, so-called charge transfer. As a result, the saturation magnetization and Curie temperature of amorphous metals are greatly decreased compared with those of crystalline metals. The decrease in magnetization acts unfavorably in power applications, giving rise to an increase in the size of cores.

An additional advantage of amorphous magnetic metals is that one can expect to realize a desirable combination of several properties. For instance, one can attain soft magnetic and hard mechanical properties simultaneously; in traditional crystalline materials, soft magnetic materials are also mechanically soft. These features not only improve performance through use of amorphous metals, but also allow the development of unique applications.

Figure 29 shows the composition dependence of saturation induction, coercivity, and magnetostriction in an amorphous $(Fe,Co,Ni)_{78}Si_8B_{14}$ alloy system. This figure combines the experimental results of Ohnuma *et al.* (1977) and Jagielinsky *et al.* (1977). The Fe-rich region has high saturation induction and high magnetostriction, and the compositions along the line from Co-rich toward Ni-rich regions exhibit zero magnetostriction, resulting in very low coercivity. Figure 29 also indicates that the magnetic properties change continuously with chemical composition. Hence it is easy to find the best composition to meet given requirements. This is another characteristic of amorphous metals. On change in composition, crystalline materials often encounter phase transformations where many properties change drastically.

Nowadays, commercial amorphous magnetic materials are largely classified into Fe-based and Co-based alloys, both of which contain B and Si as glass formers. Fe-based alloys have core losses approximately 70% lower than that of 3% silicon steel, whereas the saturation induction is 20% less than

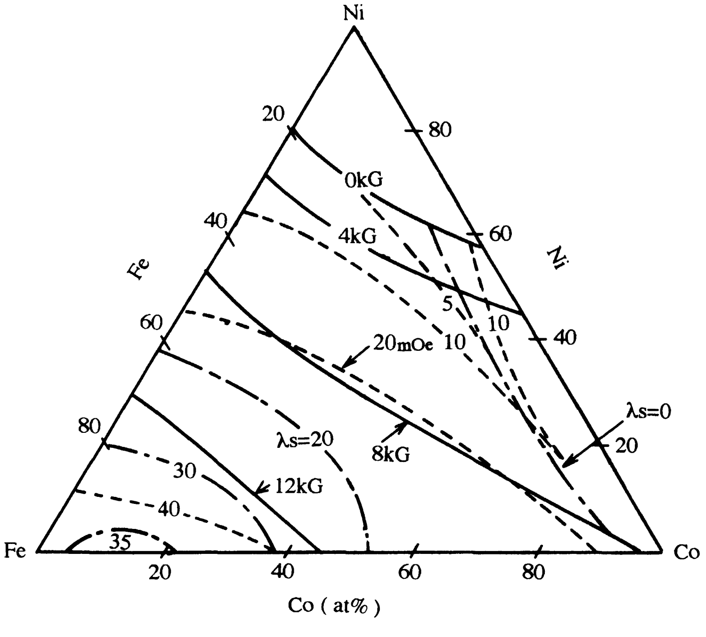

FIG. 29. Composition dependence of several magnetic properties in amorphous $(Fe–Co–Ni)_{78}Si_8B_{14}$ quasi-ternary system. Solid lines denote saturation magnetic induction in kG, dashed lines coercivity in mOe, and chained lines saturation magnetostriction in 10 times.

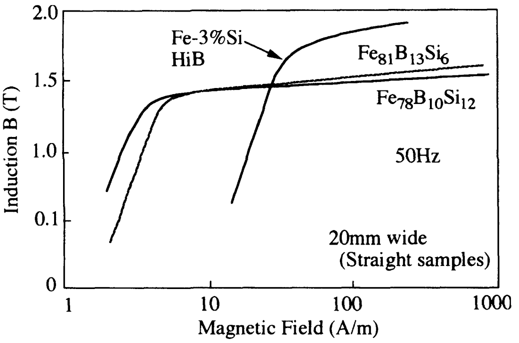

FIG. 30. Typical ac magnetic induction curves as a function of magnetic field for amorphous alloys and silicon-steel HiB.

that of silicon steel. Examples of the magnetization curves and core-loss characteristics of Fe-based alloys are given in Figs. 30 and 31 (Sato, 1983). The amorphous $Fe_{78}B_{13}Si_9$ alloy was proposed by DeCristofaro *et al.* (1982) as one of the best alloys for power transformers, and has been commercialized under the trademark name Metglas 2605S2 by Allied Signal.

Co-based alloys exhibit remarkably high permeability and low coercivity with the addition of several percent of Fe, comparable to that of Supermalloy. (Supermalloy, an alloy of composition 79 wt.% Ni, 5 wt.% Mo, and the balance Fe, is the alloy having the highest known permeability, ~1 000 000.) This superior property originates from the zero-magnetostrictive nature, involving the least stress sensitivity.

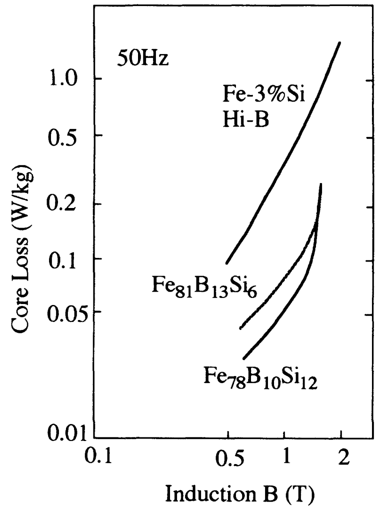

FIG. 31. Core losses as a function of magnetic induction for amorphous alloys and silicon-steel HiB.

One of the other characteristic features of amorphous alloys is that the magnetic anisotropy can be easily changed by annealing. Figure 32 shows that *B-H* loops vary greatly according to the presence or absence and the direction of a magnetic field during annealing. This indicates that one can enable the material to meet various application requirements by choosing the annealing condition.

Fe-based nanocrystalline alloys may also be regarded as a class of rapidly quenched magnetic materials. The first nanocrystalline magnetic alloy was developed by Yoshizawa *et al.*, (1988). They produced a nanocrystalline structure of an FeSiBNbCu alloy by annealing the original amorphous alloy at temperatures above the crystallization temperature. Their grain sizes of 10–20 nm are much smaller than the width of a magnetic domain wall (>100 nm). The grain boundaries therefore do not act as obstacles to domain-wall motion. Also, the nanocrystals have random orientation, and thus their apparent magnetic anisotropy is averaged out. The nanocrystalline FeSiBNbCu alloy exhibits very low coercivity and high permeability due to its very low magnetostriction. The values of permeability are comparable to those of Co-based amorphous alloys, and the saturation magnetic induction is much greater than that of Co-based amorphous alloys. The nanocrystalline FeSiBNbCu alloy is commercialized under the trademark name Finemet by Hitachi Metals.

The occurrence of low coercivity in nanocrystalline materials has been interpreted theoretically by Herzer (1990). Smaller grain

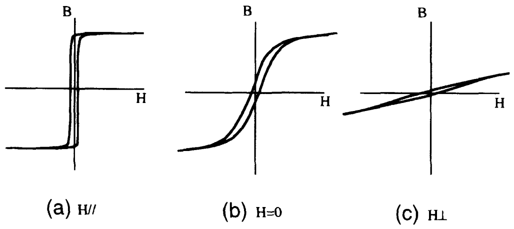

(a) H// (b) H=0 (c) H⊥

FIG. 32. Examples of hysteresis loops depending on the direction of a magnetic field applied during annealing.

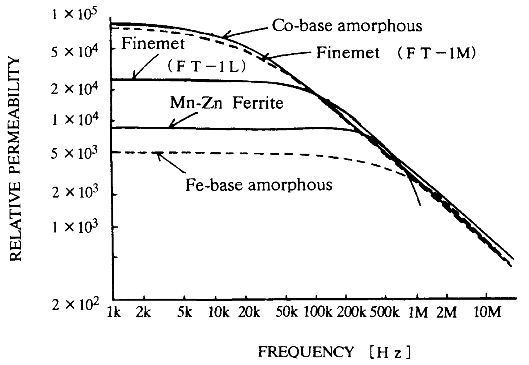

FIG. 33. Relative permeability of Finemet compared with those of Co-based and Fe-based amorphous alloys and Mn-Zn ferrite. FT-1L and FT-1M represent different anneal conditions.

sizes have so far been believed to give rise to large coercivity. The magnetic characteristics of nanocrystalline alloys are compared in Figs. 33 and 34 with those of Co-based amorphous alloys and other magnetic materials. Nanocrystalline alloys also behave in a manner similar to amorphous alloys with respect to magnetic field applied during annealing. By choice of anneal conditions, various *B-H* loops can be obtained as shown previously in Fig. 32. In addition, nanocrystalline materials have a characteristic feature superior to amorphous materials. This is a high thermal stability due to the crystalline nature, suggesting a longer service life of products made with nanocrystalline material.

3.2 Applications of Amorphous Metals

3.2.1 Distribution Transformers The use of amorphous metals in distribution transformers was first proposed by Luborsky (1978). He indicated that replacing traditional grain-oriented silicon steel by amorphous metal could save huge amounts of energy because of its low core losses. The project of realizing his idea was first carried out in the U.S.A., followed by Japan and some other countries. Test results showed a 70% reduction of core losses. Today, the largest market for amorphous metals being used in practice is for distribution transformers. The authors estimate that more than 300 000 amorphous-alloy–cored trans-

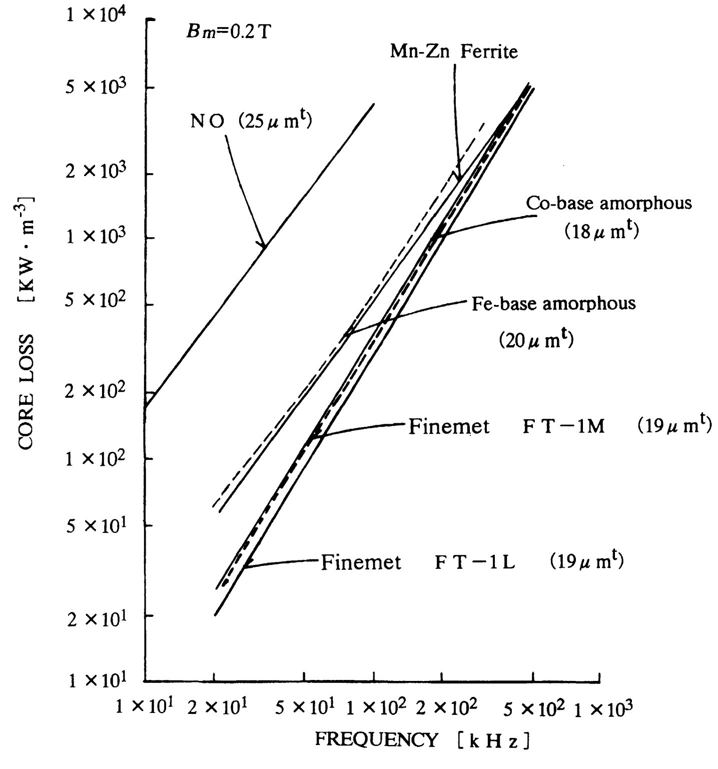

FIG. 34. Core losses of Finemet as a function of frequency compared with those of Co-based and Fe-based amorphous alloys and Mn–Zn ferrite.

formers had been built over the world by the end of 1994.

There have been two types of cores, wound cores and stacked cores, for power transformers. Wound cores, relatively small in size, are generally used for distribution transformers such as pole-mounted transformers.

Amorphous material is produced in ribbon form. Its thickness is about ten times thinner than that of silicon steel. Too thin materials are not beneficial in building stacked cores, and therefore wound cores are used in practice for amorphous-metal transformers. Wound cores are roughly classified into no-cut cores and cut cores. Figure 35 illustrates typical structures of the two types. No-cut cores are made up of continuous strips with no cut between each wrap, resulting in good magnetic performance. Hence, this type has been widely adopted for amorphous cores, which are sensitive to stress. In no-cut cores the coils are wound after fabrication, and thus a cruciform core cross section is more effective in increasing core space factor for the round coil window, but less advantageous in terms of cost because of the necessity of preparing core materials of various widths.

Cut cores have a joint for each turn, as is shown in Fig. 35(b). This makes it possible to insert core materials of identical widths into the rectangular window of the prefabricated coil, resulting in a larger coil-opening space factor and hence a smaller tank size. Almost all silicon-steel wound-core transformers are built of this type. However, there are considerable difficulties with thin and sometimes brittle materials such as amorphous metals. Cut cores are sometimes poorer in magnetic performance compared with no-cut cores for a given magnetic path length. But actual cores of rectangular shape have a shorter average length of core legs and thus result in core losses comparable to those of no-cut cores.

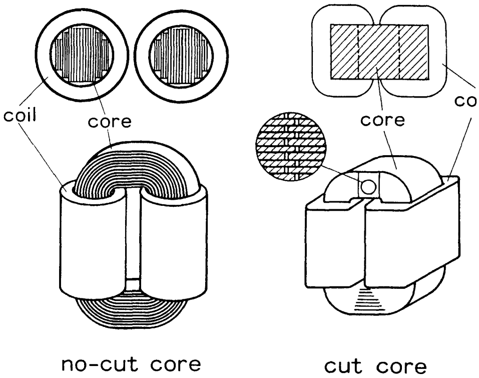

FIG. 35. Schematic illustrations of (a) no-cut core and (b) cut core.

In the first stage of commercialization, the no-cut type was adopted because of the ease of the core-winding process and good magnetic performance, but recently the proportion of cut cores is increasing. A number of evaluations of the magnetic performance have been conducted on commercial-size amorphous-metal distribution transformers. The results show that the core losses are 70% lower than for silicon-steel transformers, as expected, regardless of rated capacity. Also, the results of field tests indicate that amorphous-alloy transformers have a lifetime long enough for field service, as reported by Ng *et al.* (1991).

Attempts have also been made to apply amorphous materials to stacked-core transformers. The results showed as low core losses as for wound-core transformers. However, the thinness of amorphous material requires much more labor and time for cutting and stacking as compared with the use of silicon steel.

To improve core-processing efficiency, thick materials have been developed. One is a consolidated lamination of conventional ribbons, which has been patented by Kushnick *et al.* (1985). This material, commercialized under the trademark name Powercore, is 120–140 μm thick for 6–12 layers. Efforts have also been made to produce thick ribbon directly by rapid quenching. A new casting method proposed by Sato *et al.* (1989) uses a multiple-slotted nozzle as shown in Fig. 36. An increase in nozzle plurality enhances the

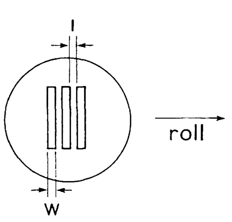

FIG. 36. Geometry of a multiple-slotted nozzle.

cooling ability substantially, and consequently ribbons thicker than 100 μm can be obtained in the amorphous state. Increasing the breadth of a slot in planar flow casting does not result in a significant increase in ribbon thickness, as reported by Fiedler *et al.* (1984).

Ribbons of 60 and 100 μm in thickness made using multislotted nozzles are as ductile as is needed to be wound without fracture around 2- and 10-mm-diam rods, respectively, as reported by Sato *et al.* (1989). The advantage of using thick ribbon is more prominent for stacked cores than for wound cores. The magnetic performance of wound cores tends to be degraded with increasing ribbon thickness, probably as a result of the incompleteness of stress relief by annealing. On the other hand, for stacked cores thick ribbon brings about good results. A 90° joint stacked core of 58-μm-thick ribbon exhibited core losses 30% less than one of 28 μm. Too thick ribbon, however, does not provide good performance, as reported by Sato *et al.* (1989). The core losses of a 92-μm ribbon core are approximately equal to those of a 28-μm ribbon core, and the exciting power of the former is considerably degraded. This is probably related to a difficulty in cutting ribbons due to the brittleness of thick ribbon. Uneven edges of cut thick ribbon would yield a large magnetic resistance at the corner joints.

3.2.2 High-Frequency Power Devices

3.2.2.1 Switching-Mode Power Supply. To miniaturize electronic devices and drive them efficiently, switched-mode power supplies are widely used today. Increasing switching frequency makes it possible to reduce the size of power supplies. In switched-mode power supplies, various kinds of magnetic components are mounted as shown in Fig. 37: common-mode chokes, normal-mode chokes, smoothing chokes, saturable reactors, and so on. These components can be further reduced in size with further increase in switching frequency. Thus the use of amorphous and nanocrystalline materials is particularly beneficial because of their excellent characteristics at high frequencies.

Most of the magnetic components used in switching power supplies can be replaced by components made with amorphous and nanocrystalline material. Co-based amorphous saturable reactors, used in magnetic-amplifier–type power supplies, contribute greatly to an increase in switching frequency because of their low core losses and the high squareness (ratio of remanence to saturation) of their *B-H* loops. Amorphous and nanocrystalline materials are commonly used today in power supplies operating at frequencies 200–500 kHz; the use of traditional Permalloy cores has been limited to below 50 kHz. Recently ribbons as thin as 5 μm have been produced by Yagi and Sawa (1990) using vacuum casting. Sawa *et al.* (1993) reported that a switching power supply using ultrathin material exhibited excellent performance for operation at 1.6 MHz.

Common-mode chokes are magnetic components to remove high-frequency noise from input current. Thus high initial permeability at high frequencies is needed for core material. Co-based amorphous and Fe-based nanocrystalline alloys are suitable for common-mode chokes because they exhibit the highest initial permeability. Common-mode chokes are often required to cut off impulse noise of high voltage. In this case, Fe-based amorphous alloys are used because of their relatively high saturation magnetization. The cores are annealed to round shape [cf. Fig. 32(b)] to obtain a low remanence. Furthermore, saturable reactors of Co-based amorphous alloys have been developed to absorb the surge voltages and spike-shaped current noise caused by recovery of diodes. The lead wire of a diode has only to be inserted into the window of a coil-free saturable core.

Amorphous alloys are also applied to

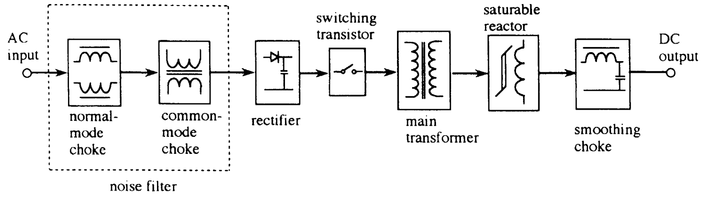

FIG. 37. Main components used in switched-mode power supply.

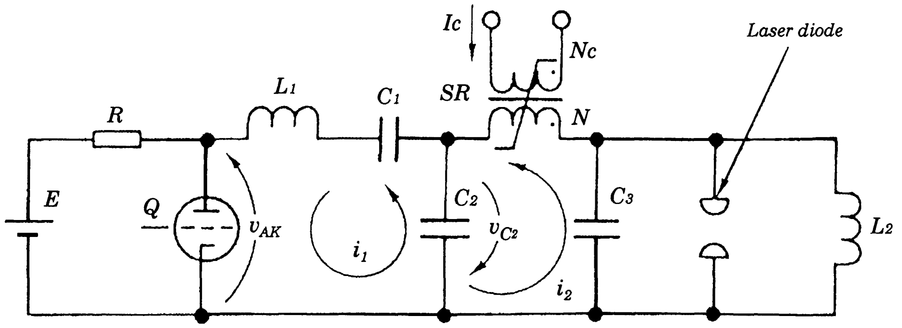

FIG. 38. Pulse generator using a magnetic pulse-compression circuit.

smoothing chokes to rectify ac current and remove ripples from dc current. To avoid the saturation of the cores due to biasing dc current, high saturation induction and a flat *B-H* loop are required for the cores. For this purpose, Fe-based amorphous-alloy cores are used in the form of cut cores, or they are annealed without the cutting process to a partially crystallized state.

3.2.2.2 High-Power Pulse Compressor. Pulsed high-power compressors are used as energy sources to drive pulsed lasers and linear accelerators. The characteristics of pulsed power required are short pulse length, large pulse peak height, high repetition frequency, and so on. For example, an excimer laser needs a pulse of several tens of kilovolts and of a repetition frequency of 500 Hz. In order to generate such sharp and high-power pulses, a pulse-compression circuit such as is shown in Fig. 38 is employed. In this circuit a saturable reactor plays a substantial role as a magnetic switch. After the main switch *Q* turns on, energy stored in the capacitor C_1 flows only into capacitor C_2 because the current flow to C_3 is blocked by the high permeability of a saturable core in unsaturated state. At the moment when the core is saturated, the current begins to flow out of C_2, and almost all of the stored energy is transferred to C_3 because the inductance L_1 is much higher than the inductance of the saturated core. The peak value of the current i_2 is enhanced to τ_1/τ_2 times the peak value of i_2, where τ_1 and τ_2 are the widths of pulse currents i_1 and i_2, respectively. A larger value of τ_1/τ_2 gives a higher pulse-compression ratio or gain, requiring a higher permeability of the saturable core in the unsaturated state. The wave forms of the current and voltage in a pulse-compression circuit are shown schematically in Fig. 39. The total

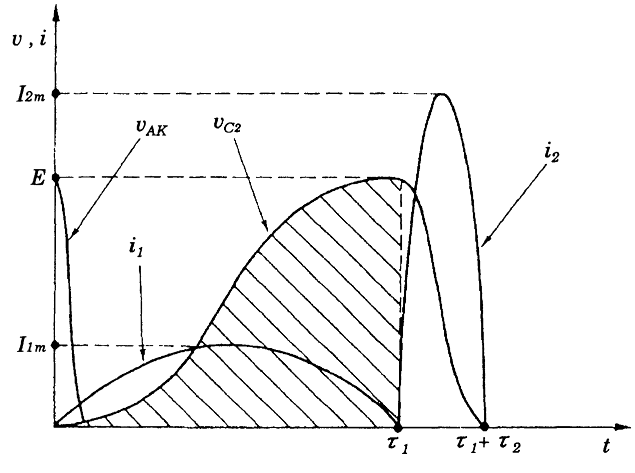

FIG. 39. Voltage and current wave forms in magnetic pulse-compression circuit.

compression ratio can be enhanced further by use of a series pulse-compression circuit.

The magnetic properties of materials required for the saturable core of a pulsed-power compressor are

1. large saturation induction,
2. low core losses,
3. large ratio of the permeability in the unsaturated state to that in the saturated state, and
4. large squareness of the *B-H* loop.

In view of these requirements, amorphous or nanocrystalline material can be favorably applied. The use of amorphous metals for magnetic pulse compression has been proposed by Birk *et al.* (1985). Nanocrystalline material is also a candidate. Nakajima *et al.* (1993) conducted a test for comparison between Fe-based and Co-based amorphous alloys and nanocrystalline material. The result showed a superior performance of nanocrystalline material (Finemet), which meets both requirements of downsizing and reducing core losses.

3.2.3 Sensors Amorphous metals have unique natures that are beneficial in sensor applications. For example, they have both high mechanical strength and soft magnetic properties. This nature is employed in magnetic heads requiring high wear resistance and high permeability. Amorphous metals also exhibit an extremely high electromechanical coupling factor. Thus transducers or sensors using amorphous material are expected to convert, effectively, mechanical quantities into electrical signals or vice versa. Actually, numerous sensor devices using amorphous metals have so far been proposed, and some of them have already been commercialized. Amorphous sensors can be classified into two categories according to the magnitude of the magnetostriction of the materials used. For low magnetostriction and high magnetostriction, Co-based and Fe-based alloys, respectively, are available. High magnetostriction generally causes a high electromechanical coupling factor; low magnetostriction yields a large inductive effect due to high permeability.

The magnetomechanical or magnetoelastic effect is a reflection of stress-sensitive magnetic properties arising from the large magnetostriction, which is favorably applied to mechanical stress or strain measurement.

Amorphous magnetic wires made by water quenching (Ohnaka *et al.,* 1982) have very unique properties as sensor materials, as has been reported by Mohri *et al.* (1992). This is attributed to their characteristic domain structure. Figure 40 illustrates a model proposed by Humphrey *et al.* (1987) for the domain structures of as-cast amorphous wires. Figure 40(a) shows the domain structure of a positive-magnetostriction wire such as Fe–Si–B alloys, and Fig. 40(b) that of a negative-magnetostriction wire such as Co–Si–B alloys. The outer shell of the positive-magnetostriction wire is composed of domains magnetized perpendicularly to the wire surface, while that of the negative-magnetostriction wire is magnetized in the circumferential direction. These domain structures are considered to result from peculiar quenched-in stresses that are compressive for the outer region and tensile for the inner region.

The inner-core domain structure is stable in the absence of domain walls. As soon as a reverse domain is nucleated at one end of a wire by application of a magnetic field in the direction opposite to that of the magnetization of the inner core, the domain wall propagates very fast in the direction of the field and is eliminated at the other end. The reason is that the state of the absence of a domain wall is the lowest in magnetostatic energy. This bistable flux reversal generates a sharp voltage pulse in a coil wound around

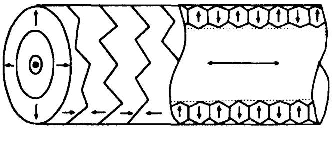

(a) $\lambda > 0$ wire

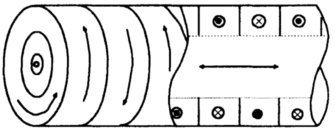

(b) $\lambda < 0$ wire

FIG. 40. Domain-structure models proposed for (a) positive-magnetostriction wire and (b) negative-magnetostriction wire.

Table 6. Amorphous sensors.

- A. High permeability (zero-magnetostrictive alloys).
  - A-1. Magnetic recording head
  - A-2. Current sensor
  - A-3. Magnetic-field sensor
  - A-4. Displacement sensor
    - data tablet
    - magnetometer
    - direction sensor
    - magnetic cartridge
- B. Magnetomechanical effect (large magnetostrictive alloys)
  - B-1. Torque sensor
  - B-2. Pressure sensor
  - B-3. Distance sensor
    - shock sensor
    - data tablet
- C. Large Barkhausen effect and Matteucci effect
  - C-1. Revolution-counter sensor
  - C-2. Field sensor
  - C-3. Security sensor
    - distance sensor
    - rotation-speed sensor
    - data tablet

the wire, which is known as a large Barkhausen effect. The large Barkhausen effect is used to detect a magnetic field with high sensitivity.

On the other hand, a negative-magnetostriction wire generates a voltage between its longitudinal ends when a magnetic field is applied perpendicular to the wire length. This voltage is induced by a flux change due to circumferential domains. This effect, called the Matteucci effect, is also applied to sensing a magnetic field at high sensitivity.

Examples of amorphous sensors are listed in Table 6. The numbered items in the table are described below in more detail. They are already used in practice, or are on the way to practical use.

3.2.3.1 High-Permeability. A-1. Magnetic recording head: Magnetic heads for audio and video devices were the first application of the high permeability of zero-magnetostrictive amorphous alloys. Cores were made by an etching process. Nowadays, the amorphous material for heads is shifting from ribbon to thin film prepared by a sputtering process (see SPUTTERING).

A-2. Current sensor: A current sensor using amorphous metal wire has been reported by Kashiwagi *et al.* (1991), whose circuit is shown in Fig. 41. Two amorphous-wire cores in a multivibrator bridge detect a field around a conductor bar and measure several hundreds of ac amperes. The sensibility of detecting a field and its range can be controlled by changing the length of the amorphous-wire cores. The detectable range is within ±300 A at 0–2 kHz, which is applicable to vector control for invertor-excited induction motors (see MACHINES, ELECTRICAL).

A-3. Magnetic-field sensor: Figure 42(a)

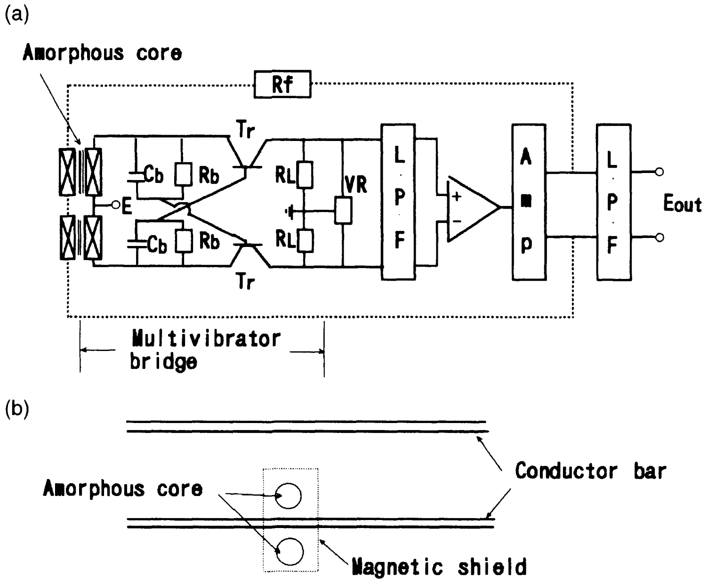

FIG. 41. (a) Current sensor composed of multivibrator bridge circuit and feedback circuit and (b) arrangement of amorphous core. RL, dummy resistance; LPF, low-pass filter; Tr, transistor; AMP, amplifier.

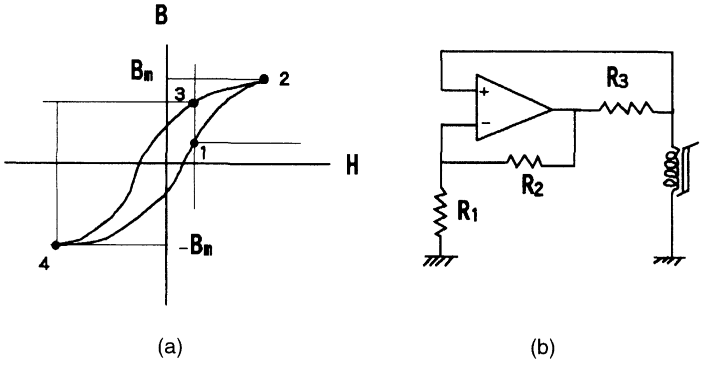

(a) (b)

FIG. 42. (a) Magnetization curve of an amorphous-ribbon core used for magnetic-field sensor, and (b) fundamental circuit for the field sensor.

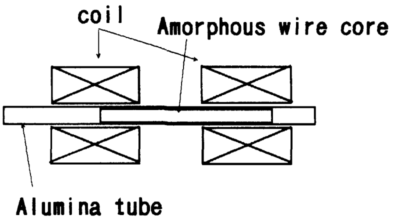

FIG. 43. Main portion of displacement-sensor head.

shows the magnetization curve of an amorphous-ribbon core. When an external field exists, the initial position for excitation corresponds to point 1 on the hysteresis loop. The difference ΔB_1 between $+B_2$ and B_1 is smaller than ΔB_2 between B_4 and B_3. A circuit to detect the field, shown in Fig. 42(b), checks the difference in exciting time between ΔB_1 and ΔB_2, obtained in practice by the integrated output voltage, and measures the strength and the direction of the field. This type of field sensor was presented by Takeuchi *et al.* (1990), providing the detection range (dc) of 0 to ±100 μT.

A-4. Displacement sensor: Mutual displacement of an amorphous core made from Co-based wire and two detection coils generates an inductance change depending on the position of the core, as shown in Fig. 43 (Yabe and Oono, 1988). The induction change of the two coils can detect the displacement. A multivibrator circuit for excitation at 0.5–30 kHz is used. The measuring range is ±(2–5) mm, and the accuracy is 0.5%.

3.2.3.2 Magnetomechanical Effect. B-1. Torque sensor: A torque sensor of high sensitivity and noncontact detection employing Fe-based amorphous ribbons has been developed by Hase and Wakamiya (1988). The sensor is used to measure the torque of a shaft in a screwdriver system. Fe-based amorphous strips are glued onto the surface of a shaft in a chevron pattern as shown in Fig. 44. The detectable torque ranges from +20 to −20 kgf-m with an accuracy of ±2.5% full scale.

When a torque is introduced on the shaft, tensile and compressive stresses of $\pm\sigma$ in the direction of ±45° relative to the axis are induced, where the stress σ is $16T/\pi d^3$ (T is torque; d is diameter), and the strain ϵ is σ(1

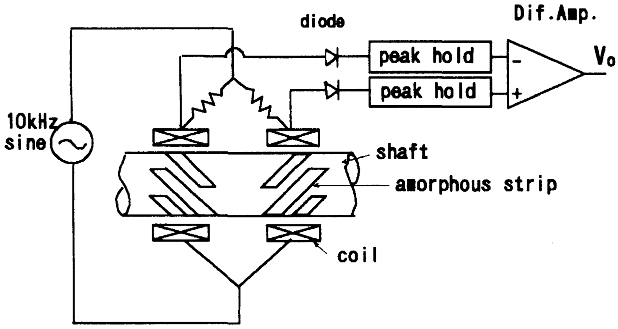

FIG. 44. Structure and electric circuit of torque sensor.

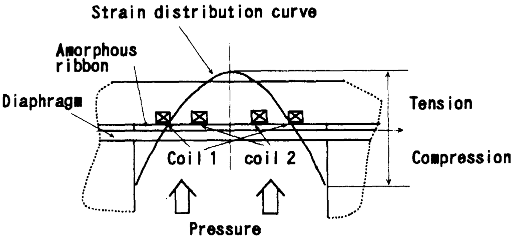

FIG. 45. Pressure sensor using amorphous ribbon fixed on a diaphragm, and schematic strain-distribution curve.

+ $\nu)/E$ (ν is Poisson's ratio; E is Young's modulus). Accordingly, the permeability of the strips changes in opposite senses and the value of an applied torque can be obtained by detecting the output voltage V_0, which corresponds to the difference in the end voltages of the elements.

To glue amorphous ribbons firmly on a shaft, glazing and laser treatment of Fe-based amorphous strips have been proposed by Hirose *et al.* (1994).

B-2. Pressure sensor: An oil-pressure sensor for automotive hydraulic control systems has been developed by Fukumaru *et al.* (1992). An amorphous strip is fixed on a diaphragm as shown in Fig. 45, where a curve of the strain distribution induced in the strip by oil pressure is also illustrated as tension and compression. The induced strain changes the permeability of the strip. In Fig. 45, coil 1 is set at the neutral position of the strain curve, and coil 2 is set at the position of tension. As shown in Fig. 46, a linear self-inductance change of coil 2 was obtained for pressures of 0 to 3 MPa. The voltage difference ΔV between coils 1 and 2 determines the magnitude of the strain, where the output of coil 1 is constant. The detectable pressure is 0–3 MPa, and the sensor output change is $\pm 3\%$ at 3 MPa at a temperature of -40 to $+120$ °C.

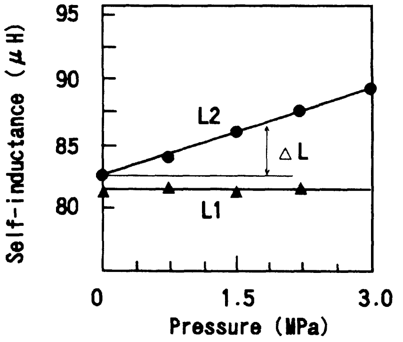

FIG. 46. Self-inductance changes of coils 1 and 2 placed at positions shown in Fig. 45 as a function of pressure.

A biomedical monitoring system as another application of a magnetostrictive sensor has been reported by Klinger *et al.* (1992).

B-3. Distance sensor: A distance sensor using an Fe-based amorphous-alloy ribbon as a magnetostrictive delay line was reported by Arai *et al.* (1976). The giant electromechanical coupling characteristics of Fe-based amorphous alloys are employed. An acoustic pulse generated by an exciting coil propagates along an amorphous ribbon as shown in Fig. 47. The magnetostrictive wave speed is 4–5 km/s, and a detection coil wound along the ribbon generates a pulse voltage in a position where a local field is applied by a magnet. The position is determined by the delay time measured from the time when the acoustic pulse is excited. The propagation speed of the pulse is too fast to control the sensitivity enough and to remove reflected waves. Instead, the use of domain-wall motion has been proposed by Uchiyama *et al.* (1989). The speed of domain-wall motion is an order of magnitude lower than that of the magnetostrictive wave, and also the reflection problem can be avoided.

3.2.3.3 Large Barkhausen Effect and Matteucci Effect. C-1. Revolution-counter sensor: A revolution-counter sensor using amorphous metals has been developed by Ogasawara (1987), who applied the large Barkhausen effect of Fe-based amorphous wires. An amorphous-wire core detects the field of a code disk with magnets embedded as shown in Fig. 48. The stable and sharp

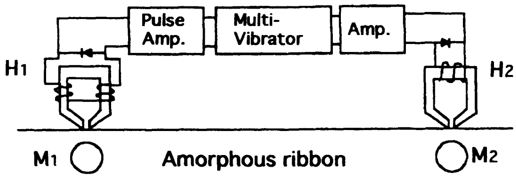

FIG. 47. Schematic setup of magnetostriction delay sensor. H_1, head for generating a supersonic wave; H_2, sensing head; M_1 and M_2, bias magnets.

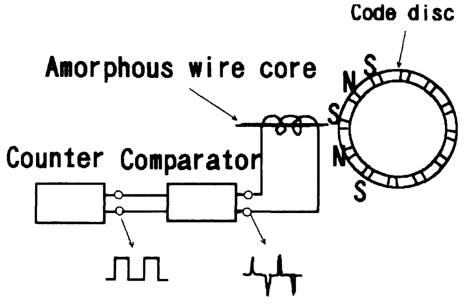

FIG. 48. Revolution-counter sensor using amorphous-wire core. Pulsed voltages are generated according to change in the sign of magnetic poles due to the revolution of code disk.

pulses induced by the field of the small magnet detect the angle of revolution. A rotary encoder with eight cores for an eight-channel code disk has been realized with a resolving power of 512 pulses (9 bits) per revolution, and maximum frequency of 10 kHz (900 rpm). The angle error is within ±1°.

C-2. Field sensor using magnetoinductive element: The magnetoinductive element (MI element) is an application of the Matteucci effect using amorphous wire. A circumferential flux change in an amorphous wire induces an ac voltage e_L between the two ends of the wire, when a high-frequency current I_w is applied to the wire with slightly negative magnetostriction, the magnetic domains of which are shown in Fig. 49 (Kawashima *et al.*, 1993). The voltage e_L is given by

$$e_L = \frac{d\phi}{dt} = \frac{1}{4\pi} l\mu_\theta(\omega,B_m)\frac{dI_w}{dt},$$

where l is the wire length and μ_θ is the permeability. If an external field H_{ex} is applied parallel to the longitudinal direction of the wire, the magnetization rotates to the axial direction and the amplitude $|e_L|$ of e_L is reduced by the reduction of the circumferential flux component as shown in Fig. 49. In Fig. 50, the normalized $|e_L|/e_L(H_{ex} = 0)$ is shown. Thus the MI element can detect a field.

The characteristics of the MI element are that

1. it is a simple device (no exciting coil),
2. it yields a large impedance change (50% for several gauss),
3. it has high sensitivity (10^{-5}–10^{-6} G with a 200-MHz multivibrator), and
4. it has high directivity.

Applications to data tablets and magnetic-head detecting fields are being developed.

C-3. Security sensor: When an amorphous wire is excited in an ac field of over 0.1 Oe, a high-harmonic voltage is generated by the Barkhausen effect. By use of this principle, goods with an amorphous wire attached can be detected when they pass through a gate installed with this sensing system. This system is put into practice in shopping centers and libraries.

3.2.4 Magnetic Shielding Amorphous ribbons that exhibit high permeability are suitable for magnetic shielding. Magnetic shields using amorphous alloys were first made by weaving narrow as-cast ribbons (Mendelsohn *et al.*, 1976). Shielding ratios obtained with these woven shields were comparable to those of Permalloy 80Ni–20Fe. The amorphous shielding has the additional advantages of being flexible and less sensitive to mechanical strain. Other applications of amorphous wires or ribbons that are woven into braids for a conductive shield are reported by Dismukes and Sellers (1978).

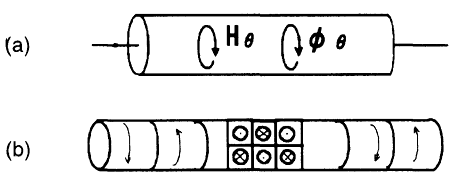

FIG. 49. (a) Circumferential magnetic field H_θ induced by the current flowing through the length of amorphous wire; (b) domain-structure model for tension-annealed amorphous wire.

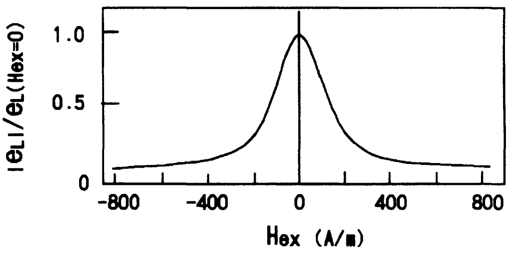

FIG. 50. Normalized inductive voltage due to circumferential flux change as a function of external field.

Shields using amorphous flakes produced by the cavitation method are reported by Ishii and Kaneko (1987). They randomly distributed and laminated as-cast amorphous fibers between two sheets of polyester films. But these two examples are inferior in performance because of the presence of numerous air gaps, which degrade the effective permeability and thus the shielding factor.

3.2.4.1 Cable and Gauge Shieldings. Co-based amorphous ribbon is wound helically around a cable to shield against magnetic-field leakage from the cable. A second layer wound in a counter sense is added to improve the shielding efficiency. It is reported by Hilzinger (1985) that a shielding factor $S = 100$ is obtained for fields up to 10 A/cm at 60 Hz.

Automobile gauge shielding using amorphous ribbon has been developed by Ichimura *et al.* (1991). As-cast Fe-based amorphous ribbon wound in a toroid is set inside the gauge case, instead of a drawn cup case made of Permalloy.

3.2.4.2 Composite Shielding Panel. A large-scale cylindrical magnetic-shielding case has been fabricated by Sasada and Oonaka (1994). This is a composite of amorphous alloy and Permalloy with a shaking system. Magnetic shaking refers to the application of an alternating magnetic field to a shielding material to keep the magnetic domain walls in vibrating motion (Bozorth, 1951). Co-based amorphous ribbons are spirally wound in 16 layers on the outer surface of a cylindrical Permalloy tube, as shown in Fig. 51. A toroidal coil wound around the amorphous layers and a meander coil arranged between the amorphous layers generate an ac magnetic-shaking field to enhance shielding effects. The typical effects of a biased shaking field are that the hysteresis curve narrows and its slope increases. Use of the system described provides an extremely low-field environment needed for SQUID (superconducting quantum interference device) measurements.

Another composite material, composed of amorphous metal and silicon steel, has been developed by Okazaki and Mogi (1994). Fe-based amorphous ribbons, annealed with no field, and silicon-steel sheets are stacked into a panel. If a shaking field is applied, this panel exhibits a remarkably large shielding factor of $S > 400$, as shown in Fig. 52. Okazaki and Ueno (1992) showed that such large shielding factors result from a combination of the circular magnetic domains of the amorphous ribbon used and ac magnetic shaking.

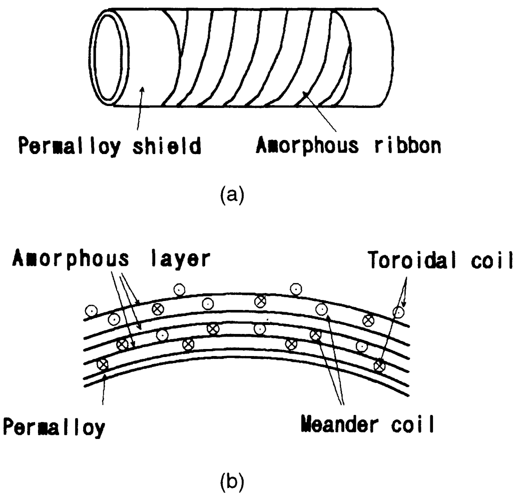

FIG. 51. (a) Schematic view of the cylindrical magnetic shield and (b) close-up view of the part of the cross section.

3.2.5 Magnetic Separation When a narrow soft-magnetic wire is placed in a magnetic field, a high magnetic-field gradient

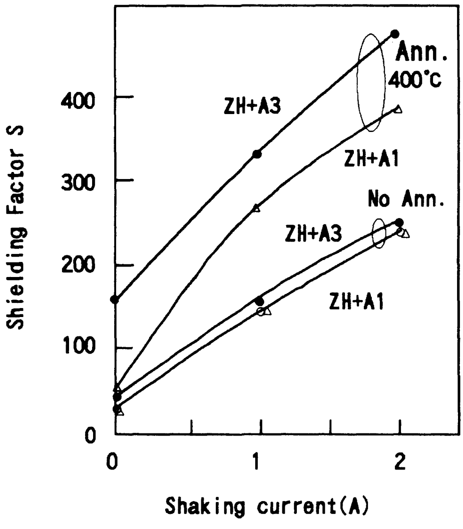

FIG. 52. Shaking effect on the shielding factor of silicon-steel sheets (ZH), with two Fe-based amorphous sheets (ZH + A1) and six amorphous sheets (ZH + A3). The effect of annealing also is shown.

is generated at the surface of the wire. Because a magnetic wire attracts magnetic particles strongly at the region of high gradient, this principle is applied to the filtration of magnetic contaminants from fluids. As a filter element, narrow amorphous-alloy ribbons are favorable for use. The reason is that amorphous ribbons have good corrosion-resistant and mechanical properties, together with soft-magnetic properties. The high corrosion resistance allows amorphous ribbons to be used in various kinds of fluids, and the high tensile strength and elasticity endure the pressure of the fluids. Additionally, a ribbon yields a much stronger attractive force at the edges than a round wire, as a result of the greater curvature and much smaller demagnetization coefficient across the fiber axis and parallel to the plane. It is reported by Kurahashi *et al.* (1990) that magnetic separation systems using amorphous-metal fibers have been put into practice to recycle hydraulic fluids and process fluids in the steel industry. They facilitate washing processing for removing magnetic particles by developing a rotary-type system. The use of amorphous magnetic filters greatly improves the efficiency of removing iron particles from skin-pass rolling oils, as shown in Fig. 53. Similar systems are also used as cleaners of pressing lubricants in the automobile industry.

3.2.6 Medical Uses

3.2.6.1 Intratissue Hyperthermia. For intratissue hyperthermia, which is expected to be one of the cancer therapies to heat tumors effectively (Dewey *et al.*, 1977), temperature-sensitive amorphous magnetic flakes coated with anticancer agents in capsules have been under development. Since the temperature rise of amorphous flakes is limited by the Curie temperature of the flakes, the maximum temperature can be controlled automatically when they are excited by an alternating magnetic field. Matsuki *et al.* (1994a) found that the maximum temperature also depends on the mean length of the flakes. Figure 54 shows their result that the maximum temperature for the $Fe_{66}P_{13}C_7Cr_{14}$ alloy with a saturation flux density of 0.44 T changes from 51 to 47.5 °C with an increase in the mean length of the flakes.

3.2.6.2 Energy-Transferring Coils for Totally Implantable Devices. A flexible transformer consisting of a pair of coils incorporating copper and amorphous magnetic fibers has been presented by Matsuki *et al.* (1994b) for such devices as an electrically driven artificial heart. Co-based amorphous fibers are wound radially on spiral copper coils as shown in Fig. 55. One of the coils is implanted in the body and receives energy at a position *g* apart from the other. Earlier work based on the same concept was reported by Smith *et al.* (1987).

3.3 Final Comments

The magnetic properties and applications of amorphous metals are outlined. Their unique natures resulting from the random

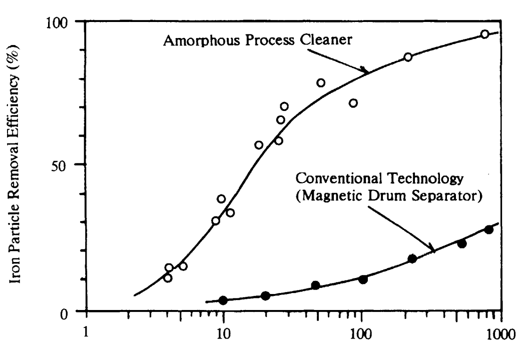

FIG. 53. The efficiency of removing iron particles from a fluid for the amorphous magnetic filter, compared with that of the conventional magnetic drum separator.

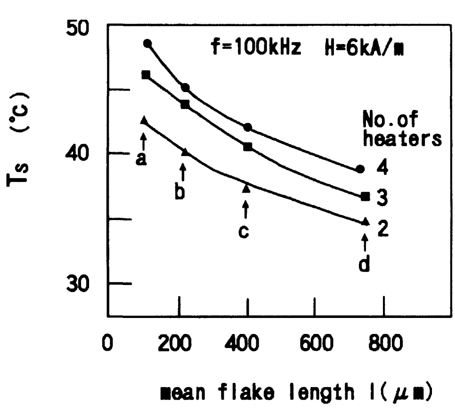

FIG. 54. Maximal temperature T_s as a function of mean flake length. T_c (°C): a = 51.0, b = 49.5, c = 48.0, d = 47.5. (Heaters filled with amorphous flakes were immersed in agar phantom.)

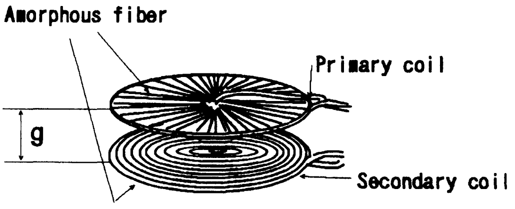

FIG. 55. A pair of transmitting coils with amorphous fibers.

atomic arrangement have been shown to be effectively used in various applications. Nevertheless, further improvements are needed for expanding the application fields drastically. One requirement is to produce thick sheets. Thick materials enable amorphous material to be applied to power transformers of stacked-core type and to motors. Though thick ribbons are made on a laboratory scale, they are not available commercially. Another requirement is to improve the workability of amorphous metals. The brittleness of amorphous metals, which increases with increase in ribbon thickness, may be the biggest problem to restrict the practice of amorphous metals. If heavy working such as deep drawing is possible at room temperature or a little higher, amorphous metals can also be used as structural materials with high corrosion resistance, bringing about a huge market. Readers who want to know about the mechanical and chemical properties of amorphous metals are referred to the book edited by Liebermann (1993) listed under Further Reading.

GLOSSARY

Amorphous Material: A material lacking the long-range order of a crystalline material. Includes noncrystalline, glassy, or disordered material.

Amorphous Si Solar Cell: Solar cell with *i*-type *a*-Si (and/or with Ge and C):H; the semiconductor provides optical absorption, electron–hole pair creation, and electron and hole separation layers.

Avalanche-Multiplication Effect: The phenomenon in which an electron or hole accelerated by an electric field creates an electron and/or hole by collision, which is accelerated and again creates an electron and/or hole, producing a cascade multiplication. The initial electron or hole can be produced by photon absorption, by injection from an electrode through a barrier, etc.

Boltzmann Distribution: A function giving the probability distribution of generalized position and momentum coordinates for a molecule of a gas in thermal equilibrium obeying classical mechanics.

Carrier Mobility: Carrier velocity under unit electric-field strength.

CCD: Charge-coupled device.

Chalcogen: Element from group VI of the periodic table (especially S, Se, Te).

Chalcogenide: Chemical compound containing an element, other than oxygen, from column VI of the periodic table (S, Se, Te).

CVD: Chemical vapor deposition.

Dangling Bond: The nonbonding state of a valence electron that usually forms a chemical bonding state in the covalently bonded crystalline material. Localized states induced by dangling bonds prevent the *p*- or *n*-type control of the material.

Dark Conductivity: Conductivity in the absence of incident light.

Electrochromic: Changeable in color when injected with positive or negative charges.

Electroluminescence: The emission of light resulting from application of an electric field to a material.

Faraday Rotation: Rotation of polarization of a beam of linearly polarized light when it passes through a transparent material in the direction of an applied magnetic field.

Gap States: States distributed in an energy gap. Gap states consist of localized states and tail states.

GRIN Lens: Graded refractive-index lens, consisting of a rod with a refractive index that has its maximum value on the axis, decreasing approximately as the square of the distance from the axis.

Halide: Compound material containing a halogen.

Halogen: Element from group VII of the periodic table (i.e., F, Cl, Br, I, At).

Harmonic: A sinusoidal component of a periodic wave, having a frequency that is an integral multiple of the fundamental frequency.

Hydrogenated Amorphous Silicon (*a*-Si:H): Amorphous silicon alloyed with hydrogen. Nominal hydrogen content ranges between 1 and 30 at.%. Typically, *a*-Si:H is prepared by glow-discharge decomposition of monosilane (SiH_4).

Image Pickup Tube: A vacuum tube in which a two-dimensional input image is converted into a one-dimensional electric signal over time by electron-beam scanning. An example is Vidicon-type operation.

IR Spectroscopy: Infrared spectroscopy. IR absorption and emission spectroscopy gives information on vibrational spectra of chemically bonded atoms and/or molecules.

ITO: Indium tin oxide.

Laser-Beam Printer: Carlson-process printer using a laser-beam–scanned image.

LCD: Liquid-crystal display.

Light-Induced Structural Change: Structural change in an amorphous material induced by light; this takes place more easily than in a crystal because of the nonequilibrium metastable nature of the amorphous material. In the case of *a*-Si:H, it is called the Staebler–Wronski effect.

Molecular Dynamics: A method for the simulation of the structure and/or motion of a collection of atoms by following the time evolution of the classical motion under a particular interaction potential.

Optical Bistability: The property of a substance or device that has two stable states of transmission, high and low, for a single input light intensity.

Organometallic Compound: A compound containing an alkyl or aryl radical bonded to a metal.

Passivation of a Dangling Bond: Termination of a dangling bond by an appropriate element or chemical fragment. Localized states induced by the dangling bond in the electronic gap are effectively eliminated.

Phase Separation: Separation of a two-component or multicomponent melt into two liquid phases differing markedly in composition from each other, resulting from immiscibility within certain ranges of composition and temperature.

Photochromism: Change in color when exposed to visible or near-visible light, and restoration in the dark.

Photoconductivity: Conductivity resulting from light irradiation. The light-irradiation condition should be specified.

Photoreceptor: A main part of the Carlson-process printer. The input image is converted into an electric-charge image on the photoreceptor surface, where it is held for a certain time (0.1–100 s).

Photovoltaic Effect: Photoinduced electromotive force in a heterogeneous structure involving a semiconductor, such as a *p-n* junction or a metal–semiconductor junction. An absorbed photon creates an electron–hole pair, and an internal electric field formed by the heterogeneous structure separates the electron and hole.

Quantum Efficiency: The ratio of the number of output quanta to the number of input quanta.

Random Network: The amorphous structure of atoms bonded into a network that shows no order beyond a few atomic dimensions.

SAXS: Small-angle x-ray scattering. Used mainly to characterize medium-range structures.

Tetrahedrally Coordinated Material: Covalently bonded material with sp^3 hybridized orbitals, such as Si, Ge, diamond, and gray tin.

Van Hove Singularity: Singular point in the energy–momentum dispersion curve. The phase velocity vanishes at the singular point.

Vitrification: Formation of a glassy or noncrystalline material.

Works Cited

Anderson, D. A., Spear, W. E. (1977), *Philos. Mag.* **35,** 1–16.

Arai, K., Tsuya, N., Yamada, M., Matsumoto, T. (1976), *IEEE Trans. Magn.* **MAG-12,** 936–938.

Baji, T., Shimomoto, Y., Matsumarau, H., Koike, N., Akiyama, T., Sasano, A., Tsukuda, T. (1982), *Jpn. J. Appl. Phys.* **21,** 269–273.

Baker, G. L. (1987), in: M. J. Bowden, S. R. Turner (Eds.), *Polymers for High Technology,* Washington, DC: American Chemical Society, p. 558.

Birk, D. L., Cook, E. G., Hawkins, S. A., Poor, S. A., Reginato, L., Schmidt, J., Smith, M. W. (1985), in: S. Steeb, H. Warlimont (Eds.), *Proceedings of the 5th International Conference on Rapidly Quenched Metals,* Vol. 2, pp. 1637–1638.

Blyler, L. L., DiMarcello, F. D., Hart, A. C., Huff, R. G. (1987), in: M. J. Bowden, S. R. Turner (Eds.), *Polymers for High Technology,* Washington, DC: American Chemical Society, Chap. 34.

Boyd, D. C., Thompson, D. A. (1985), in: M. Grayson (Ed.), *Encyclopedia of Glass, Ceramics, Clay and Cement,* New York: Wiley, p. 462.

Bozorth, R. M. (1951), *Ferromagnetism,* Princeton, NJ: Van Nostrand; reprint (1978), New York: Dover; reprint (1993), New York: IEEE Press, p. 546.

Carlson, C. F. (1938), U.S. Patent No. 2,221,776.

Carlson, D. E., Wronski, C. R. (1976), *Appl. Phys. Lett.* **28,** 671–673.

Cassanhiol, B., Cornet, J., Rossier, D. (1974), in: J. Stuke, W. Brenig (Eds.), *Proceedings of the 5th International Conference on Amorphous and Liquid Semiconductors,* London: Taylor and Francis, pp. 571–579.

Credelle, T. J. (1988), in: L. F. Weber, W. E. Howard (Eds.), *Proceedings of the International Display Research Conference,* Piscataway, NJ: IEEE Inc., pp. 208–214.

DeCristofaro, N. J., Datta, A., Davis, L. A., Hasegawa, R. (1982), in: T. Masumoto, K. Suzuki (Eds.), *Proceedings of the 4th International Conference on Rapidly Quenched Materials,* Vol. 2, Sendai: Japan Institute of Metals, pp. 1031–1033.

Dewey, W. C., Hopwood, L. E., Sapareto, S. A., Gerweck, L. E. (1977), *Radiology* **123,** 463–474.

Dismukes, J. P., Sellers, G. J. (1978), in: B. Cantor (Ed.), *Proceedings of the 3rd International Conference on Rapidly Quenched Metals, Brighton,* Vol. 2, pp. 205–212.

Elliott, S. R. (1991), in: J. Zarzycki (Ed.), *Materials Science and Technology,* Vol. 9, Weinheim: VCH, Chap. 7.

Fiedler, H., Muhlbach, H., Stephani, G. (1984), *J. Mater. Sci.* **19,** 3229–3235.

Fukumaru, T., Machino, H., Takagi, M., Nakajima, T. (1992), *Proc. Soc. Automotive Eng. Jpn.* **3** (924138), 97–100.

Geiger, G. (1990), *Am. Ceram. Soc. Bull.* **69,** 1131–1136.

Haruki, H., Uchida, Y. (1983), in: Y. Hamakawa (Ed.), *Amorphous Semiconductor—Technologies and Devices,* Tokyo: Ohmsha; Amsterdam: North-Holland, pp. 216–227.

Hase, H., Wakamiya, M. (1988), *Nat. Tech. Rep.* **34,** (4), 10–15.

Herzer, G. (1990), *IEEE Trans. Magn.* **MAG-26,** 1397–1402.

Higgins, L. M, III (1992), in: *1992 Proceedings International Symposium on Microelectronics,* Reston, VA: ISHM—The Microelectronics Society, pp. 216–224.

Hilzinger, H. R. (1985), in: S. Steeb, H. Warlimont (Eds.), *Proceedings of the 5th International Conference on Rapidly Quenched Metals,* Würzburg: North-Holland, Vol. 2, pp. 1695–1698.

Hirose, F., Takagi, M., Mori, H., Imura, T. (1994), *Mater. Sci. Eng.* **A181/A182,** 1359–1362.

Humphrey, F. B., Mohri, K., Yamasaki, J., Kawamura, H., Malmhall, R., Ogasawara, I. (1987), in: *Proceedings of the International Symposium on Magnetic Properties of Amorphous Metals,* pp. 110–116.

Ichikawa, Y., Ihara, T., Saito, S., Ota, H., Fujikake, S., Sakai, H. (1992), in: L. Guimares, W. Palz, D. DeReyff, H. Kiess, P. Helm (Eds.), *Proceedings of the 11th EC-Photovoltaic Solar Energy Conference,* Chur, U.K.: Harwood Academic, pp. 203–206.

Ichimura, E., Kurihara, K., Horikawa, M. (1991), *Tech. Rep. Kantoseiki* **2,** 95–99.

Imamura, Y., Ataka, S., Takasaki, Y., Kusano, C., Hirai, Y., Maruyama, E. (1979), *Appl. Phys. Lett.* **35,** 349–351.

Inoue, E., Shimizu, I. (1982), *Photogr. Sci. Eng.* **26,** 148–152.

Ishii, H., Kaneko, M. (1987), *J. Magn. Soc. Jpn.* **11,** 413–416.

Jagielinsky, T., Arai, K. I., Tsuya, N., Ohnuma, S., Matsumoto, T. (1977), *IEEE Trans. Magn.* **MAG-13,** 1553–1555.

Juska, G., Arlauskas, K. (1980), *Phys. Status Solidi A* **59,** 389–393.

Kagawa, T., Tatsumoto, N., Kawabe, K. (1982), *Jpn. J. Appl. Phys.* **21,** 251–256.

Kashiwagi, Y., Kondo, T., Uematsu, T. (1991), *Nat. Conv. Rec. IEE Jpn.* **14,** S.19-3, pp. s19-9–s19-12.

Kawamura, T., Yamamoto, N., Nakayama, Y. (1983), in: Y. Hamakawa (Ed.), *Amorphous Semiconductor—Technologies and Devices,* Tokyo: Ohmsha; Amsterdam: North-Holland, pp. 325–336.

Kawashima, K., Kohsawa, T., Yoshida, H. K., Mohri, K., Panina, L. V. (1993), *J. Magn. Soc. Jpn.* **17,** 423–428.

Kitagawa, T., Hattori, K., Shuto, K., Oguma, M., Temmyo, J., Suzuki, S., Horiguchi, M. (1993), in: *19th European Conference on Optical Communication. Proceedings,* Vol. 3, Zurich: Swiss Electrotechnical Association (SEV), pp. 41–44.

Klement, W., Willens, R. H., Duwez, P. (1960), *Nature (London)* **187,** 869–870.

Klinger, T., Pfuetzner, H., Schoenhuber, H. P., Hoffmann, K., Bachi, N. (1992), *IEEE Trans. Magn.* **MAG-28,** 2400–2402.

Kobayashi, K., Abe, T., Miyake, H., Kashimura, H., Ozawa, T., Hamano, T., Fennek, L., Turner, W., Weisfield, R. (1991), in: W. C. Chang, J. R. Melch (Eds.), *Camera and Input Scanner Systems,* SPIE Proceedings No. 1448, Bellingham, WA: SPIE, pp. 157–163.

Kurahashi, M., Takemoto, M., Oishi, N. (1990), in: *Proceedings of the Japan International Tribology Conference,* Nagoya: Japan Society of Tribologists, pp. 71–76.

Kushnick, J. H., Sawhney, D. L., Hathaway, R. E. (1985), U.S. Patent No. 4,529,458.

LeComber, P. G., Spear, W. E., Ghaith, A. (1979), *Electron. Lett.* **15,** 179–181.

Luborsky, F. E. (1978), *IEEE Trans. Magn.* **MAG-14,** 1008–1012.

Lucas, J. (1991), in: J. Zarzycki (Ed.), *Materials*

Science and Technology, Vol. 9, Weinheim: VCH, Chap. 8.

MacChesney, J. B., DiGiovanni, D. J. (1991), in: J. Zarzycki (Ed.), *Materials Science and Technology,* Vol. 9, Weinheim: VCH, Chap. 15.

Manabe, S., Matsunaga, Y., Furukawa, A., Yano, K., Endo, Y., Miyagawa, R., Iida, Y., Egawa, Y., Shibata, H., Nozaki, H., Sakuma, N., Harada, N. (1991), *IEEE Trans. Electron Devices* **ED-38,** 1765–1771.

Maruyama, E., Hirai, T., Goto, N., Isozaki, Y., Shidara, K. (1974), in: J. Stuke, W. Brenig (Eds.), *Proceedings of the 5th International Conference on Amorphous and Liquid Semiconductors,* London: Taylor and Francis, pp. 581–586.

Matsuki, H., Yanada, T., Sato, T., Murakami, K., Minakawa, S. (1994a), *Mater. Sci. Eng.* **A181/A182,** 1366–1368.

Matsuki, H., Matsuzaki, T., Suzuki, A. (1994b), *Mater. Sci. Eng.* **A181/A182,** 1363–1365.

Mendelsohn, L. I., Nesbitt, A., Bretts, G. R. (1976), *IEEE Trans. Magn.* **MAG-12,** 924–926.

Miyachi, K., Ishigo, N., Miyashita, T., Yanagawa, N., Tanaka, H., Koyama, M., Ashida, Y., Fukuda, N. (1992), in: L. Guimares, W. Palz, C. DeReyff,H. Kiess, P. Helm (Eds.), *Proceedings of the 11th EC-Photovoltaic Solar Energy Conference,* Chur, U.K.: Harwood Academic, pp. 88–91.

Miyake, Y., Sakai, K., Abe, T., Sakai, Y., Hatta, H., Sugino, H., Ito, H., Ozawa, T. (1991), in: W. C. Chang, J. R. Melch (Eds.), *Camera and Input Scanner Systems,* SPIE Proceedings No. 1448, Bellingham, WA: SPIE, pp. 150–156.

Mohri, K., Kohzawa, T., Kawashima, K., Yoshida, Y., Panina, L. V. (1992), *IEEE Trans. Magn.* **MAG-28,** 3150–3152.

Morimoto, A., Miura, T., Kumeda, M., Shimizu, T. (1982), *J. Appl. Phys.* **53,** 7299–7305.

Muramatsu, S., Shimada, T., Kajiyama, H., Azuma, K., Watanabe, T., Kamiyama, T., Suzuki, K. (1989), *Jpn. J. Appl. Phys.* **28,** L1092–L1095.

Muramatsu, S., Matsubara, S., Watanabe, T., Shimada, T., Kamiyama, T., Suzuki, K., Matsuda, A. (1991), *Jpn. J. Appl. Phys.* **28,** L2006–L2008.

Nakajima, S., Arakawa, S., Yamashita, Y., Shiho, M. (1993), *Nucl. Instrum. Methods Phys. Res.* **A331,** 318–322.

Nakata, Y., Sannomiya, H., Moriuchi, S., Inoue, Y., Nomoto, K., Yokota, A., Itoh, M., Tsuji, T. (1990a), *Optoelectronics Device Technol.* **5,** 209–221.

Nakata, Y., Sannomiya, H., Moriuchi, S., Yokota, A., Inoue, Y., Itoh, M., Itoh, H. (1990b), in: P. C. Taylor, M. J. Thompson, P. G. LeComber, Y. Hamakawa, A. Madan (Eds.), *Material Research Society Research Symposium No. 192,* Pittsburgh: MRS, pp. 15–26.

Narasimhan, M. C. (1980), U.S. Patent No. 4,221,257.

Ng, H. W., Hasegawa, R., Lee, A., Lowdermilk, L. A. (1991), *Proc. IEEE* **79,** 1608–1623.

Nishigaki, S., Fukuta, J. (1989), in: M. F. Yan, K. Niwa, H. H. O'Bryan, Jr., W. S. Young (Eds.), *Ceramic Substrate and Packages for Electronic Applications,* Westerville, OH: American Ceramic Society, pp. 199–215.

Niwa, K., Imanaka, Y., Kamehara, N., Aoki, S. (1989), in: M. F. Yan, K. Niwa, H. H. O'Bryan, Jr., W. S. Young (Eds.), *Ceramic Substrate and Packages for Electronic Applications,* Westerville, OH: American Ceramic Society, pp. 323–337.

Oda, S., Saito, Y., Shimizu, I., Inoue, E. (1981), *Philos. Mag. B* **43,** 1079–1089.

Ogasawara, O. (1987), *Sensor Gijutsu* **8** (1), 116–120.

Ohishi, T., Maekawa, S., Katoh, A. (1992), *J. Non-Cryst. Solids* **147&148,** 493–498.

Ohnaka, I., Fukusako, T., Ohmichi, T., Masumoto, T., Inoue, A., Hagiwara, M. (1982), in: T. Masumoto, K. Suzuki (Eds.), *Proceedings of the 4th International Conference on Rapidly Quenched Metals,* Sendai: Japan Institute of Metals, Vol. 1, pp. 31–34.

Ohnishi, M., Nishiwaki, H., Tsuda, S., Nakamura, N., Nakano, S., Matsuoka, T., Kuwano, Y. (1984), in: A. F. Forestieri (Ed.), *Conference Record of the 17th IEEE Photovoltaic Specialists Conference,* Piscataway, NJ: IEEE, pp. 70–75.

Ohnuma, S., Watanabe, K., Masumoto, T. (1977), *Phys. Status Solidi A* **44,** K151–K154.

Oka, S., Kikuchi, T., Kanegae, H., Nakajima, H. (1992), in: *1992 Proceedings International Symposium on Microelectronics,* Reston, VA: ISHM—The Microelectronics Society, pp. 58–63.

Okamoto, H., Nitta, Y., Adachi, T., Hamakawa, Y. (1979), *Surf. Sci.* **86,** 486–491.

Okaniwa, H., Asano, M., Nakatani, K., Yano, M., Suzuki, K. (1982), *Jpn. J. Appl. Phys. Suppl.* **21-2,** 239–244.

Okazaki, Y., Mogi, H. (1994), *Mater. Sci. Eng.* **A181/A182,** 1374–1377.

Okazaki, Y., Ueno, K. (1992), *J. Magn. Magn. Mater.* **112,** 192–194.

Pearson, A. D., MacChesney, J. B., French, W. G. (1985), in: M. Grayson (Ed.), *Encyclopedia of Glass, Ceramics, Clay and Cement,* New York: Wiley, p. 439.

Pond, R., Jr., Maddin, R. (1969), *Trans. Metall. Soc. AIME* **245,** 2475–2476.

Porai-Koshits, E. A. (1990), *J. Non-Cryst. Solids* **123,** 1–13.

Powell, M. J. (1989), *IEEE Trans. Electron Devices* **36,** 2753–2763.

Rawson, H. (1980), *Properties and Applications of Glass,* Amsterdam: Elsevier, Chap. 1.

Sakka, S. (1987), in: *Proceedings of the First International Symposium on New Glass,* Tokyo: The Association of New Glass Industries, pp. 1–13.

Sasada, I., Oonaka, Y. (1994), *J. Magn. Soc. Jpn.* **18,** 473–476.

Sasaki, M., Miyagawa, R., Manabe, S. (1990), in: *Extended Abstracts of the 22nd Conference on SSDM,* Tokyo: Japan Society of Applied Physics, pp. 705–708.

Sato, T. (1983), in: Y. Sakurai (Ed.), *Japan Annual Reviews in Electronics, Computers and Telecommunications,* Vol. 10, *Recent Magnetics for Electronics,* Chap. 13.

Sato, T., Yamada, T., Ozawa, T. (1989), in: T. Masumoto (Ed.), *Proceedings of the MRS International Meeting on Advanced Materials,* Pittsburgh: Material Research Society, Vol. 3, pp. 303–316.

Sato, T., Yamada, T., Ozawa, T. (1990), *An. Fis. B* **86,** 148–150.

Sawa, T., Nakagawa, K., Maruyama, J., Inoue, N., Takagi, K., Yoshida, J., Kobayashi, H. (1993), *Technical Report of IEICE,* PE92-66, pp. 13–18.

Sawada, T., Tarui, H., Terada, N., Tanaka, M., Takahama, T., Tsuda, S., Nakano, S. (1993), in: R. J. Schwartz (Ed.), *Conference Record of the 23rd IEEE Photovoltaic Specialists Conference,* Louisville: Electron Devices Society of the IEEE, pp. 803–810.

Senturia, S. D. (1987), in: M. J. Bowden, S. R. Turner (Eds.), *Polymers for High Technology,* Washington, DC: American Chemical Society, Chap. 36.

Shidara, K., Goto, N., Maruyama, E., Hirai, T., Nonaka, Y. (1981), *IEEE Electron Device Lett.* **EDL-2,** 101–102.

Shimada, T., Katayama, Y., Komatsubara, K. F. (1979), *J. Appl. Phys.* **50,** 5530–5532.

Smith, B., Peckham, P. H., Keith, M. W., Roscoe, D. D. (1987), *IEEE Trans. Biomed. Eng.* **MBE-34,** 499–508.

Snitzer, E. (1989), in: H. Jain, A. R. Cooper, K. J. Rao, D. Chakravorty (Eds.), *Current Trends in the Science and Technology of Glass,* Singapore: World Scientific, p. 375.

Soga, N. (1990), *J. Non-Cryst. Solids* **123,** 363–376.

Sohn, J. E., Singer, K. D., Kuzyk, M. G. (1987), in: M. J. Bowden, S. R. Turner (Eds.), *Polymers for High Technology,* Washington, DC: American Chemical Society, Chap. 33.

Spear, W. E., LeComber, P. G. (1975), *Solid State Commun.* **17,** 1193–1196.

Spear, W. E., LeComber, P. G. (1976), *Philos. Mag.* **33,** 935–949.

Staebler, D. L., Wronski, C. R. (1977), *Appl. Phys. Lett.* **31,** 292–294.

Staebler, D. L., Wronski, C. R. (1980), *J. Appl. Phys.* **51,** 3262–3268.

Takeuchi, S., Sakai, F., Hara, K., Yoshimuta, K. (1990), Papers of Technical Meeting on Magnetics, *IEEE Trans. Magn.* **MAG-26,** 1409–1411.

Tanioka, K., Yamazaki, J., Shidara, K. S., Taketoshi, K., Kawamura, T., Ishioka, S., Takasaki, Y. (1987), *IEEE Electron Device Lett.* **EDL-8,** 392–394.

Tawada, Y., Kondo, M., Okamoto, H., Hamakawa, Y. (1982), *Solar Energy Mater.* **6,** 299–315.

Tomozawa, M. (1987), in: *Proceedings of the First International Symposium on New Glass,* Tokyo: The Association of New Glass Industries, pp. 15–26.

Tsuji, K., Takasaki, Y., Hirai, T., Yamazaki, J., Tanioka, K. (1987), in: *Extended Abstracts of the 19th Conference on SSDM,* Tokyo: Japan Society of Applied Physics, pp. 91–94.

Tsuji, K., Ohshima, T., Hirai, T., Gotoh, N., Tanioka, K., Shidara, K. (1991), in: A. Madan, Y. Hamakawa, M. J. Thompson, P. C. Taylor, P. G. LeComber (Eds.), *Materials Research Society Symposium No. 219,* Pittsburgh: MRS, pp. 507–518.

Tsuji, K., Takasaki, Y., Hirai, T., Yamazaki, J., Tanioka, K. (1994), *Optoelectron. Device Technol.* **9,** 367–387.

Tsukuda, T., Baji, T., Yamamoto, H., Takasaki, Y., Hirai, T., Maruyama, E., Ohba, S., Koike, N., Ando, H., Akiyama, H. (1979), in: *IEDM Technical Digest,* Piscataway, NJ: IEEE Inc., pp. 134–136.

Tsumura, M., Kitajima, M., Funahata, K., Wakai, Y., Saito, R., Mikami, Y., Nagae, Y., Tsukuda, T. (1991), in: J. Morreale (Ed.), *SID '91 International Symposium, Digest of Technical Papers,* Playa del Rey, CA: SID, pp. 215–218.

Tummala, R. R. (1992), in: *1992 Proceedings International Symposium on Microelectronics,* Reston, VA: ISHM—The Microelectronics Society, pp. 357–362.

Uchiyama, T., Mohri, K., Konno, Y., Sato, T. (1989), *J. Magn. Soc. Jpn.* **13,** 439–442.

Vogel, W., Höland, W. (1990), *J. Non-Cryst. Solids* **123,** 349–353.

Watanabe, T., Tanaka, M., Azuma, K., Nakatani, M., Sonobe, T., Shimada, T. (1987), *Jpn. J. Appl. Phys.* **26,** L288–L290.

Watts, M. P. C. (1987), in: M. J. Bowden, S. R. Turner (Eds.), *Polymers for High Technology,* Washington, DC: American Chemical Society, p. 160.

Weber, M. J. (1990), *J. Non-Cryst. Solids* **123,** 208–222.

Weber, M. J. (1991), in: *J. Zarzycki (Ed.), Materials Science and Technology,* Vol. 9, Weinheim: VCH, Chap. 12.

Weimer, P. K., Cope, A. D. (1951), *RCA Rev.* **12,** 314–334.

Weimer, P. K. (1962), *Proc. IRE* **50,** 1462–1469.

Weisfield, R. L. (1993), *J. Non-Cryst. Solids* **164–166,** 771–776.

Yabe, T., Oono, M. (1988), *Sensor Technol.* **8,** 183–185.

Yagi, M., Sawa, T. (1990), *IEEE Trans. Magn.* **MAG-26,** 1409–1411.

Yamaguchi, M., Kneko, Y., Tsutsui, K. (1993), *Jpn. J. Appl. Phys.* **32,** 458–461.

Yamamoto, N., Nakayama, Y., Wakita, K., Nakano, M., Kawamura, T. (1981), *Jpn. J. Appl. Phys. Suppl.* **20-1,** 305–310.

Yoshizawa, Y., Oguma, S., Yamauchi, K. (1988), *J. Appl. Phys.* **64,** 6044–6046.

Further Reading

Amorphous Semiconductors

Elliott, S. R. (1991), "Amorphous Materials, Structure of," in: G. L. Trigg (Ed.), *Encyclopedia of Applied Physics*, Vol. 1, New York: VCH, pp. 559–592.

Elliott, S. R., Davis, E. A., Robertson, J. (Eds.) (1993), *Amorphous Semiconductors—Science and Technology*, Parts 1 and 2, Amsterdam: North-Holland.

Fritzsche, H. (Ed.) (1989), *Advances in Disordered Semiconductors—Amorphous Silicon and Related Materials A and B*, Vol. 1, Singapore: World Scientific.

Hamakawa, Y. (Ed.) (1982, 1983, 1984), *Amorphous Semiconductor—Technologies and Devices*, Tokyo: Ohmsha; Amsterdam: North-Holland.

Mott, N. F., Davis, E. A. (1979), *Electronic Processes in Non-Crystalline Materials*, 2nd ed., Oxford: Clarendon.

Sakurai, Y., Hamakawa, Y., Masumoto, T., Shirae, K., Suzuki, K. (Eds.) (1993), *Current Topics in Amorphous Materials*, Amsterdam: North-Holland.

Takahashi, K., Konagai, M. (1986), *Amorphous Silicon Solar Cells*, London: North Oxford Academic.

Functional Glasses

Bowden, M. J., Turner, S. R. (Eds.) (1987), *Polymers for High Technology*, Washington, DC: American Chemical Society.

Davidson, T. (Ed.) (1984), *Polymers in Electronics*, Washington, DC: American Chemical Society.

Jain, H., Cooper, A. R., Rao, K. J., Chakravorty, D. (Eds.) (1989), *Current Trends in the Science and Technology of Glass*, Singapore: World Scientific.

Zarzycki, J. (Ed.) (1991), *Materials Science and Technology*, Vol. 9, Weinheim: VCH.

Amorphous Metals

Hernando, A. (Ed.) (1993), *Nanomagnetism*, NATO Advanced Studies Institute Series E, Applied Sciences, No. 247, Dordrecht: Kluwer Academic.

Liebermann, H. H. (Ed.) (1993), *Rapidly Solidified Alloys*, New York: Marcel Dekker.

Luborsky, F. E., Hamakawa, Y., Masumoto, T., Shirae, K., Suzuki, K. (Eds.) (1983), *Amorphous Metallic Alloys*, London: Butterworths.

Sakurai, Y. (Ed.) (1993), *Current Topics in Amorphous Materials*, Amsterdam: North-Holland.

Warlimont, H. (Ed.) (1985), *Proceedings of the 5th International Conference on Rapidly Quenched Metals*, Vol. 2, Würzburg: North-Holland.

PROSTHESES AND ARTIFICIAL LIMBS

RICHARD F. ff. WEIR AND DUDLEY S. CHILDRESS, *Northwestern University Prosthetics Research Center, Chicago, Illinois, U.S.A.*

Introduction ... 115
1. **Desirable Attributes for Prosthesis Control** ... 116
2. **Upper-Limb Prostheses** ... 116
2.1 Design Constraints (Why Human, Prosthetic, and Robotic Arms Have So Little in Common) ... 117
2.1.1 Arm Replacement or Tool? ... 119
2.1.2 Function vs Appearance ... 119
2.2 Prehension or Grasp ... 120
2.3 Prehension with Prostheses ... 121
2.3.1 Handlike Prehensors ... 122
2.3.2 Body-Powered Non-Handlike Prehensors ... 122
2.3.3 Voluntary-Opening versus Voluntary-Closing Devices ... 122
2.3.4 Externally Powered Non-Handlike Prehensors ... 123
2.3.5 Passive Prostheses ... 123
2.4 Artificial Arms ... 123
2.4.1 Body-Powered Prostheses ... 124
2.4.2 Externally Powered Systems .. 125
2.4.3 Hybrid Systems ... 126
2.5 Control ... 127
2.5.1 Myoelectric Control ... 128
2.5.2 Myoacoustic and Neuroelectric Control ... 128
2.6 The Issue of Feedback ... 129
2.7 Parallel Development of Robots, Manipulators, and Prostheses ... 129
2.7.1 Extended Physiological Proprioception ... 130
2.8 Surgery for Improved Control ... 132
2.9 Medical Manipulators or Rehabilitation Robots ... 133
3. **Lower-Limb Prostheses and Walking Robots** ... 135
4. **Conclusion** ... 137
Glossary ... 137
Works Cited ... 139

INTRODUCTION

It is the goal of artificial limb replacement to restore as much of the function and appearance lost through amputation as possible. This having been said, it should be noted that the history of the development of this field was tied to developments in amputation surgery. In turn, the development of amputation surgery was to a large extent dominated by the ebb and flow of human warfare.

Artificial limb replacements in the form of peglegs date from antiquity. In the middle ages, armorers made artificial limbs to hide amputations incurred by their knight masters in battle. However, it was not until the late nineteenth century that the enterprise of artificial limb replacement flourished. During that period mortality from amputation was greatly reduced by the advent of aseptic surgical techniques and the use of anesthesia. This was particularly true during and following the American Civil War. In the United States, the *modern era* of prosthetics can be dated from 1945—the end of World War II—with the founding of the Committee on Artificial Limbs of the National Research Council.

The field of artificial limb replacements can be broadly divided into the areas of the upper limb and lower limb. The area of upper-limb prosthetics is dominated by considerations of prosthesis control and power, while the area of lower-limb prosthetics is dominated by considerations of passive mechanisms, material properties, and limb–prosthesis fit and alignment. For further reading on the history of artificial limb replacement or prosthetics the reader is referred to American Academy of Orthopaedic Surgeons (1960).

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. DESIRABLE ATTRIBUTES FOR PROSTHESIS CONTROL

The level of amputation is an important element in the restoration of functional control over a prosthesis. *The level of amputation is inversely proportional to the number of control sources available from the person, but the number of control sources needed is directly proportional to the level of amputation.* That is, the higher the level of amputation the fewer the available control sources but the greater the amount of function that must be replaced. A control source is the means used by the amputee to control a specific function or degree of freedom of the prosthesis—e.g., opening and closing of an artificial hand. Childress (1992) presented the following attributes of prosthesis control as desirable. While some of these attributes may be difficult, if not impossible, to achieve in practice, they are seen as desirable goals.

1. Low mental loading or subconscious control. This means that the prosthesis can be used without undue mental involvement. Successful control systems enable the users to use their artificial limbs almost subconsciously, the way we use our own limbs. In other words, the prosthesis should serve the user; the user should not be the servant of the prosthesis. This kind of control may require physiologically appropriate proprioceptive and sensory feedback in order to be achieved.
2. User friendly or simple to learn to use. This implies that the use of a prosthetic device should be intuitive and natural. The amputee should be able to learn to use the prosthesis quickly and easily.
3. Independence in multifunctional control. Control of any function, or degree of freedom, should be able to be executed without activating or interfering with the other control functions of a multifunctional prosthesis. For example, a person with prostheses on both arms should be able to use each arm independently. Operation of one prosthesis should not activate the prosthesis on the other side.
4. Simultaneous, coordinated control of multiple functions (parallel control). This is the ability to coordinate multiple functions simultaneously in effective and meaningful ways without violating the first and third attributes.
5. Direct access and instantaneous response (speed of response). All functions should, if possible, be directly accessible to the user and without time delay. Prosthetic systems should respond immediately to inputs.
6. No sacrifice of human functional ability. The control system should not encumber any natural movement that an amputee can apply to useful purposes. In general it is not wise to sacrifice a useful body action for the control of a prosthesis. The prosthesis should be used to supplement, not subtract from, available function.
7. Natural appearance. If possible, the control system should be operated in ways that are aesthetically attractive. Likewise, the mechanical response should be graceful. Control techniques that allow aesthetically pleasing action (e.g., smooth, flowing, graceful movement) are important to the dynamic appearance (dynamic cosmesis) of the prosthesis just as is static appearance (static cosmesis) in the form of shape and color. Movements that appear mechanical in nature attract unwanted attention in social situations and may not be pleasing to the eye.

2. UPPER-LIMB PROSTHESES

The task of replacing a lost hand or arm is a daunting one. There are 33 muscles acting on the hand. There are 18 intrinsic muscles and 15 extrinsic muscles controlling hand function. The human hand has 27 major bones, and over 20 joint articulations with 27 or more degrees of freedom. The arm contributes another seven degrees of freedom. Even if the hand is replaced with a simple single–degree-of-freedom gripper, as is usually the case, this still remains an extremely challenging control problem. Robotic hands that mimic the operation of the human hand have been built—for example, the M.I.T./Utah dexterous hand of Animate Systems Inc., Utah (Jacobsen *et al.*, 1982). However, the actuators and computer system required to control this hand occupy a space equal to that occupied by two small filing cabinets, and power is supplied externally from electrical mains. Thus, weight, size, and power considerations preclude the use of robotic devices, like the dexterous hand, in prosthetics applications.

The primary role of the human arm is to position the hand in space. The control of the arm is directed at controlling the position of the hand. Even though humans control their arms with what appears to be great simplicity, this is a highly complex and demanding task. Consider the example of the backhoe operator. The backhoe is essentially a mechanical arm that is under the control of the operator. When working this mechanical arm the operator uses both arms to pull levers, both feet to press pedals to operate the arm, and both eyes to monitor the task being performed by the digger. All this attention is needed, just to control a single mechanical arm! Now consider the plight of a person who has lost both of his or her arms at a high level (e.g., above the elbow) and one begins to have some appreciation of the task a limbless person faces in controlling an artificial (prosthetic) arm or arms. The problem of providing persons with artificial arms that are anywhere near as functional as the arms that were lost has been called one of the most challenging problems of medical engineering.

The famous physiologist Nikolai Bernstein (1967) observed that a blacksmith can precisely hit the same point on an anvil with a hammer time after time, but that the joint angles of his arm varied somewhat for each swing. This example highlights a problem that exists in deciphering how humans control motion. The position of the joints and their angles do not really matter as long as the hammer strikes its target. There is massive redundancy, in terms of degrees of freedom, in the human arm and torso. Small changes in posture are accommodated by changes in other joints such that the hand/tool still ends up on target. A robot doing the same task would travel in exactly the same, rigidly defined, trajectory each time. Robot arms generally have six degrees of freedom. This is the minimum required to specify both orientation and position of a point in space. The human arm and trunk, on the other hand, have in excess of thirty-four degrees of freedom to specify the same point.

The problem of the lack of specification of joint angles is known as Bernstein's redundancy problem and arises as a result of how the brain controls a motion. In Bernstein's opinion "... the structure of the central complex which governs the production of a given series of movements is much more closely related to spatial form than to muscle scheme ..."; or, to rephrase: The human control of movement appears concerned with spatial position rather than with control of a particular set of muscles.

In addition to this lack of specificity, the human arm is compliant (springlike); it has "give." When the hammer of the blacksmith strikes the surface of the anvil, the arm "gives" to absorb the shock and prevent damage to the joints. This "give" or compliance also stops the arm of the blacksmith from going into oscillation or instability upon contact (contact instability) with the anvil. The problem of contact instability is frequently encountered by robotic arms, unless the robot, the environment, and the nature of the interaction between the two (e.g., hammering the anvil) are precisely mathematically and mechanically defined. Contact instabilities can occur if this mathematical definition of the problem is even slightly off. The human arm does not have this problem because of its inherent compliance. This compliance is not a fixed quantity but can be varied depending on the task requirements: a stiff arm for bracing oneself against an expected blow, a relaxed arm for playing the piano. This change in arm compliance is achieved by the cocontraction of agonist/antagonist muscle pairs (simultaneous contraction of opposing muscle groups) that in turn varies joint impedance (stiffness). Considerable effort has gone into designing this type of compliance into robot arms and manipulators. In the field of prosthetics, Hogan (1982) developed impedance control for a prosthetic elbow. Human motor function, robot function, and prosthesis function would seem to have much in common; nevertheless, their different substrates and different constraints have so far limited the degree to which exchange of knowledge between these fields has been useful.

2.1 Design Constraints (Why Human, Prosthetic, and Robotic Arms Have So Little in Common)

Robotics concepts have had little impact on commercial prosthetics because of the severe physical constraints required for a prosthetic device to be successful. In any prosthetic device, size, weight, and power are all at a premium, far more so than in the related fields of robot or manipulator design. Limb replacements, it is thought, should be anthropomorphic (of human form) in general shape, size, and outline. Anthropomorphic does not necessarily imply that the artificial limb look exactly like its human counterpart, but it does mean that any replacement should have similar kinematics and kinetics, i.e., there should be a joint that operates (moves) like an elbow joint where one would expect to see an elbow joint, and the various limb segments should be of a size consistent with a normal human being. In remote manipulation a robot arm or manipulator is under the control of an operator who may be physically separated from the robot arm by a large distance, such as in space-based systems where a robot arm in space may be controlled by an operator located on earth. In such systems, a slave arm (the remote robot arm) is controlled by a master arm that is worn by the operator. This use of master–slave arms has led to the use of an anthropomorphic design for the general-purpose manipulator (see Fig. 1). The amount of mental effort that must be exerted (mental loading) is reduced if the slave arm is kinematically similar to the master arm. The master arm in turn is usually fitted to the operator's own arm to make the control more intuitive and subconscious.

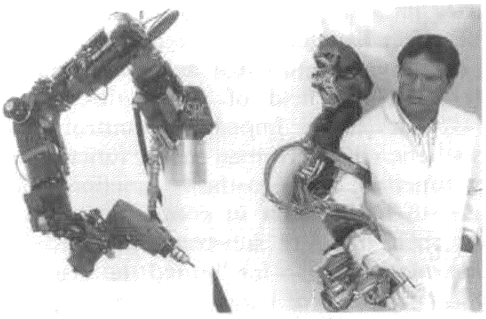

FIG. 1. Picture of an operator in the SARCOS dexterous, master–slave, force-reflecting teleoperator. The operator "wears" a master arm to control a remote slave arm. Consequently an anthropomorphic shape for both arms eases the mental control burden placed on the operator. Forces experienced by the slave are reflected back to the master and experienced by the operator. (Photograph courtesy of SARCOS Research Corporation, Utah, UT.)

It is important to realize that *people see what they expect to see,* so that if a person with an artificial limb replacement looks and moves in an expected manner the fact of the artificial limb will pass unnoticed to the casual observer. Often the terminal device, the hand replacement, will look only generally like a natural hand but because it is moved about in a natural-looking manner it will go undetected until closely scrutinized. This property is referred to as dynamic cosmesis. Static cosmesis is achieved when the device also looks like the limb it is replacing. Although a device might be statically "correct" in appearance it can look "wrong" or lifeless if it does not also have dynamic cosmesis.

The issue of weight is also very important to the success of any prosthetic limb. Contrary to what one might think, an artificial limb should usually be lighter than the limb it replaces. Arm replacements need to be as light as possible or else they may be rejected by the users. To be effective, artificial arms need to be worn by the user for periods in excess of 12 hours a day. The lack of an intimate connection between amputee and limb replacement means that the prosthesis will be perceived as an external load and therefore as something that must be carried.

Power must be conserved in prostheses. If the power source is the amputee, i.e., body power, then the prosthesis should not require excessive effort so as not to tire the user over the course of the day (frictional losses need to be minimized). If the artificial limb is externally powered (i.e., uses a power source other than the body, usually electric storage), the limb should be able to run all day from the same power source without needing to be recharged. In addition, it is desirable for the power source to be contained within the prosthesis. Electric batteries are the main source of energy for modern prosthetic arms, although pneumatic gas cylinders have been used. Both are heavy and occupy space but electric storage provides greater energy/weight and energy/volume ratios. The problem of portable prosthesis power is analogous to the power issues in the laptop computer industry where a major contributor to the weight of a portable computer is the battery.

Although some size, weight, and power constraints must be placed on robots and remote manipulators, these constraints are far less severe than in prosthetics. Robotic actuators can often be as large and as heavy as required to achieve a specific result. Power is usually not an issue since it can be obtained from the power mains. As a result of the radically different design constraints, robotics and prosthetics design have much less in common than one might expect.

2.1.1 Arm Replacement or Tool? The control of a joint's compliance or impedance (impedance control) is considered by some designers to be a physiologically correct form for prosthesis control. It would seem to follow that this kind of control is desirable and ought to be incorporated into prosthetic arm systems. This raises a philosophical question as to whether an artificial limb should be designed as a limb replacement or as a tool. A true limb replacement would be a fully functioning substitute limb that is intimately connected to and controlled by the amputee. The goal of such a limb replacement would be to match as closely as possible the properties and function of the original limb. Alternatively, the prosthesis as a tool can be regarded as an interchangeable device to be worn by the amputee, used as needed, and then ignored. The tool makes no pretense of trying to replace the lost limb physiologically but is there as an aid to help provide some of the functions that were lost. While the difference might seem to be subtle it will permeate the discussion throughout this section. The goal of some research in the prosthesis field is directed toward the prosthesis as a limb replacement, and this does seem desirable; however, in current practice we are mostly limited to the prosthesis as a tool. The major causes of this limitation are practical ones due to the severe design constraints previously mentioned and the difficulty in finding a sufficient number of appropriate control sources to control the requisite number of degrees of freedom.

Impedance control comes under the guise of the prosthesis as a limb replacement. As with the neuromuscular system, this type of control requires that a control signal be constantly applied/maintained to hold a particular position. This places a constant drain on the power source, which is technically undesirable. The presently available body-powered wrist, elbow, and shoulder joints can be considered to have a limited form of compliance control or two-state compliance control. Through the use of locking mechanisms these devices have more or less infinite impedance when a joint is locked or close to zero impedance when the joint is unlocked and free to be moved.

2.1.2 Function versus Appearance The relative importance of function vs appearance is another issue that faces the amputee, and assisting clinicians, when deciding on a particular type of prosthesis, if any. Static vs dynamic cosmesis has been previously discussed but very often there is a tradeoff to be made between function and appearance. Some amputees may be solely concerned with the visual presentation and reject a body-powered prosthesis because of the unsightly appearance of the control harness or of a hook-shaped terminal device. Others might find the function provided by these devices to outweigh poor appearance. Choice of prosthesis is ultimately based on many psychological and practical factors. Other factors affecting choice are age, sex, occupation, and degree of physical activity of the amputee, the amputee's attitude toward training, the type of amputation involved, and whether only one arm is affected (unilateral) or both arms are affected (bilateral).

While the physical constraints mentioned previously are severe, ultimately it is more the control constraints (shortage of independent control sources) and not the physical constraints that impede the development of today's prosthetic systems. It is for this reason that upper-limb prosthetics is often dominated by consideration of control. Presently available terminal devices have a single degree of freedom (opening and closing). Prosthetic arms requiring multi–degree-of-freedom control use sequential control with locking mechanisms to switch control from one component or degree of freedom to the next. Generally, since vision is the primary source of feedback, the greatest number of functions that are controlled in parallel is two. Otherwise the mental burden becomes excessive. In spite of these limitations, many amputees become skilled users of these devices and show surprising adaptability and inventiveness in overcoming functional challenges faced by them in everyday life.

2.2 Prehension or Grasp

The reduction of the hand to a single–degree-of-freedom terminal device in most artificial limbs was done to reduce the number of control sources required to control prehension. The function of the human hand may be divided into two subgroups: nonprehensile functions and prehensile functions. Nonprehensile functions of the hand are those functions where grasping is not required; for example, pushing an object, holding an object between the body and forearm, flicking, brushing, percussive motions such as playing the piano, etc. Prehensile hand functions are those cases where an object is grasped and held partly or wholly within the hand.

Numerous different classifications of hand prehension have been defined but it is the six patterns of Keller *et al.* (1947) that have endured the test of time and are the most widely accepted in the field of prosthetics (see Fig. 2). These patterns, as described in Klopsteg and Wilson (1954), are tip prehension, lateral prehension, palmar prehension, cylindrical prehension, spherical prehension, and hook prehension.

Tip prehension, lateral prehension, and palmar prehension are primarily the function of the thumb working in opposition with the index and middle fingers. Napier (1956) described these patterns as precision grips. Tip prehension, or fingernail pinch, is used mainly to grasp small objects. In lateral prehension the thumb holds an object against the side of the index finger as is the case when using a key. In palmar prehension (sometimes referred to as tri-digital pinch or three-jaw chuck) the thumb opposes either a single finger or two or more fingers. This is the grip most commonly used in daily activities. Cylindrical and spherical prehension are examples of power grasps as described by Napier. Power grasps use all the fingers of the hand to provide an encompassing grasp that firmly stabilizes the object being held. Hook prehension is achieved by flexing the fingers into a hook; the thumb is either alongside the index finger or used to lock the object held. Carrying a briefcase is a good illustration of this kind of prehension.

The grip of the hand is further improved by the ability of the hand to adapt passively to the shape of an object grasped. A grasped object depresses, or indents, the skin and underlying soft tissues of the hand, at first meeting little resistance. Consequently, the

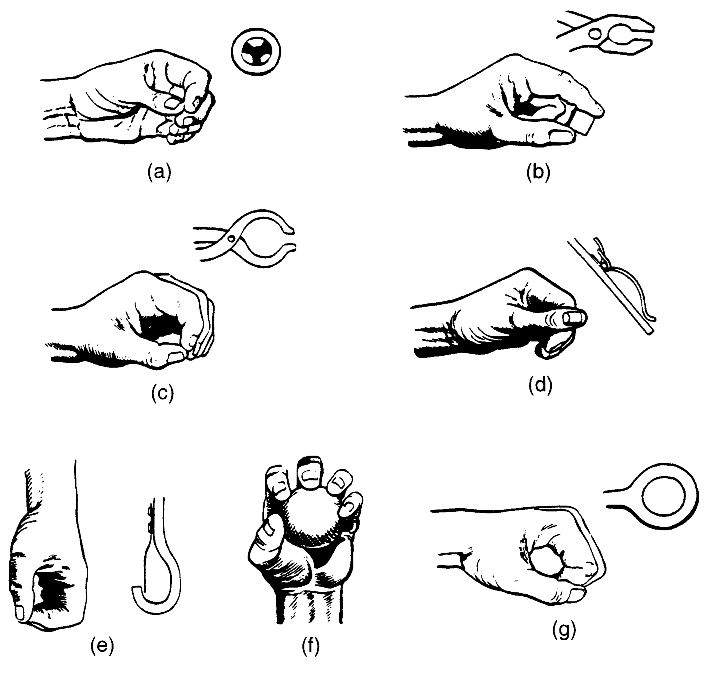

FIG. 2. Sketches of the prehension patterns defined by Keller *et al.* (1947). (a) Palmar prehension. Opposition of thumb with digits II and III in the near-closed position gives a three-jawed chuck prehension. (b) Palmar prehension. At openings of 1 to 3 in., opposition between thumb and digit II is in the manner of a pliers grip. (c) Tip prehension. Distal phalangeal segments of digits II and III and of thumb are strongly flexed to bring the finger and thumb tips into opposition. (d) Lateral prehension. Ball of thumb opposes lateral surface of digit II, usually at the second phalanx. (e) Hook prehension. Load supported by hooked terminal phalanges, thumb acting largely to prevent slipping volarly. (f) Spherical prehension. In addition to elements of grasp prehension, curving across the knuckle line permits conformity to spherical objects. (g) Cylindrical/grasp prehension. Digits curled about object, such as a handle; thumb curves and overlies finger tips to close the prehension "ring."

soft tissue adapts easily to the shape of the object grasped. However, the mechanical properties of the soft tissue are quite nonlinear and the conforming tissue becomes turgid as pressure is increased. This is demonstrated by blanching on the sides of the pads of the finger tips if sufficient force is applied. The rise in tissue stiffness after conformation to shape enables objects to be grasped with great security. This feature of the human hand would seem to be useful for robotic and prosthetic systems.

2.3 Prehension with Prostheses

The finding by Keller *et al.* (1947) that palmar prehension was the most frequently used pattern has meant that most prosthetic hands incorporate this pattern as the dominant pattern. The persistence of this configuration in prosthetic hand designs and its general acceptance over the years tend to support this finding. The passive adaptability afforded by the soft tissue of the hand is mimicked, to some extent, by lining the device with a soft plastic and covering it with a glove that has been molded to resemble the natural hand (a cosmetic glove).

Adequate reproduction of all six prehension patterns in an artificial handlike device requires three or four degrees of freedom. A four–degree-of-freedom device typically uses two for the thumb, one for the index finger, and one for the remaining three fingers. If flexion of the fingers is prevented, three degrees of freedom can be used: two for the thumb; and one to splay the fingers. Prosthetic hands with this number of degrees of freedom cannot at present be controlled in a meaningful way on account of the lack of suitable control sources. However, it is hoped that the use of surgical techniques to create additional control sites will remedy this problem. In the area of remote manipulation such hands have been developed for use with master–slave manipulators, the terminal device of the SARCOS system (Jacobsen *et al.*, 1990) shown in Fig. 1 being a case in point (see Fig. 3).

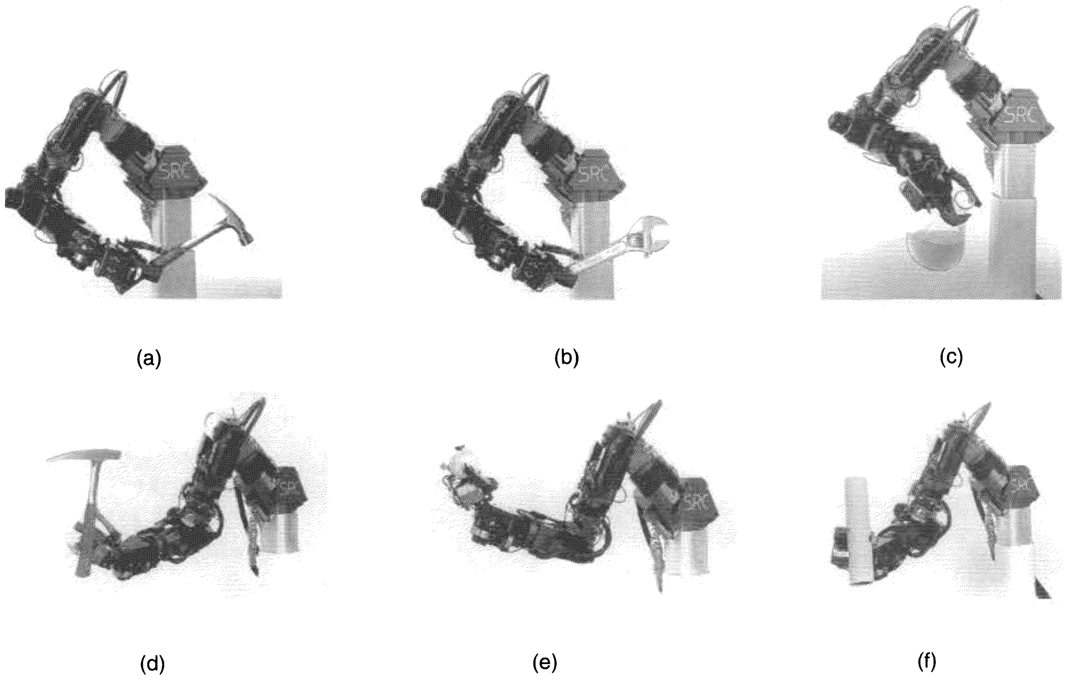

FIG. 3. Photographs showing various grasp patterns for the three–degree-of-freedom SARCOS dexterous manipulator terminal device. The index finger (digit II) is fixed. The thumb has two degrees of freedom to permit lateral and palmar prehension while digit III has the third degree of freedom to provide a stabilizing grip for spherical and cylindrical prehension patterns. (Photographs courtesy of SARCOS Research Corporation, Utah, UT.)

2.3.1 Handlike Prehensors Commercially available handlike prehensors are predominantly externally powered. The de facto standard handlike prosthesis is a mechanism manufactured by Otto Bock Orthopedic Inc. This mechanism is covered with a hand-shaped plastic liner over which is pulled a PVC or silicone rubber cosmetic glove (see Fig. 4). Hugh Steeper, Ltd. (Roehampton, England), and Centri (Sweden) also manufacture handlike devices for the adult. Child-size hands are also available, and there is a trend today to fit young children with birth or traumatic below-elbow deficiencies as early as possible. A number of child-size hands are manufactured by Variety Village Systems, Inc. (VASI), Otto Bock Orthopedic Inc., and Hugh Steeper, Ltd., among others (Michael, 1986; Heckathorne, 1992).

Palmar prehension in these handlike prehensors is achieved by using monocoque (single joint) fingers that are fixed in slight flexion at a position approximating the interphalangeal joints of the human hand. The resulting finger shape also creates a concave inner prehension surface that can be used to provide cylindrical prehension. Body-powered handlike prehensors are available but few are used effectively as active terminal devices. Frictional losses and poor mechanical advantage in the hand mechanism limit pinch force of body-powered hands. In addition, restriction of motion by the glove and/or plastic liner and the bulkiness of the hand's shape, which blocks vision, both serve to diminish the effectiveness of these devices (Fryer and Michael, 1992).

2.3.2 Body-Powered Non-Handlike Prehensors Non-handlike prehensors have the advantage of being able to incorporate many of the previously mentioned prehension patterns into a single–degree-of-freedom device. The split hook (Fig. 5) is an example of such a prehensor. It can perform tip, lateral, cylindrical, and hook prehension. This in large part accounts for the popularity of the split-hook prosthesis in spite of its less than appealing appearance. Supplied by Hosmer-Dorrance Corp., Otto Bock Orthopedic Inc., and other companies, split hooks come in many variations on the basic theme (Fryer and Michael, 1992).

2.3.3 Voluntary-Opening versus Voluntary-Closing Devices Body-powered terminal devices can be classified according to their mode of operation. They usually fall into one of two categories: voluntary-opening devices or voluntary-closing devices. Voluntary-opening components have an inherent closing force acting on the terminal device when the control cable is relaxed. Voluntary-opening components have an inherent opening force acting on the terminal device when the control cable is relaxed. The closing force of a voluntary-opening component

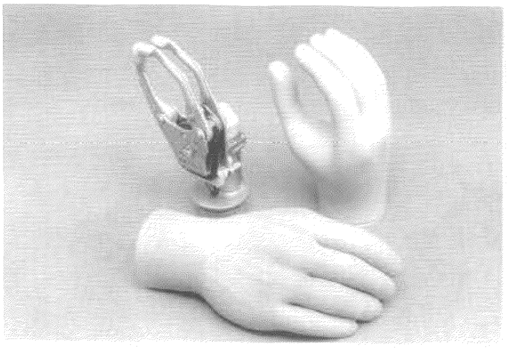

FIG. 4. Otto Bock Industries, Inc., externally powered handlike prosthesis. This device is the de facto standard for externally powered handlike prostheses. It consists of a drive mechanism (left) that is covered by a hand-shaped plastic liner (right). A cosmetic glove (bottom) is then worn over this liner. (Photograph courtesy of Northwestern University Prosthetics Research Laboratory, Chicago, IL.)

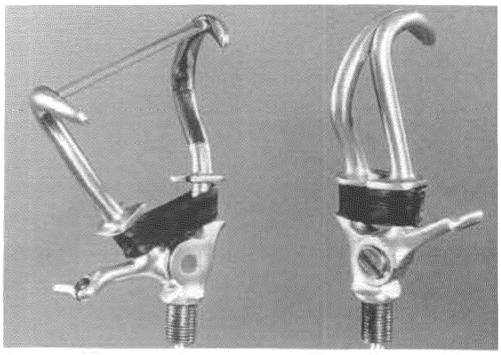

FIG. 5. The traditional voluntary-opening split-hook–type terminal device. This device exerts a closing force that must first be overcome before it can be opened. In the left-hand picture the hook is being held open by a spacer bar. (Photograph courtesy of Northwestern University Prosthetics and Orthotics Center, Chicago, IL.)

must be overcome before the terminal device will open. Prehension force with these devices is limited to the closing force (see Fig. 6, top). The opening force of a voluntary-closing component must be overcome before the terminal device will open. Prehension force in the terminal device is directly proportional to the force applied to the device (see Fig. 6, bottom).

The question of whether voluntary opening is superior to voluntary closing really depends on the design philosophy used. If a prosthesis is to be viewed primarily as a tool, to be used and ignored, then voluntary opening is appropriate. The terminal device acts simply as a vise that can be positioned anywhere in the amputee's work space. If, however, the prosthesis designer is striving for a limb replacement that is intimately joined to the amputee then the more physiologically correct control of voluntary closing (prehension force proportional to applied cable force) seems more appropriate. Nevertheless, the bulk of body-powered prostheses in use today employ voluntary-opening devices.

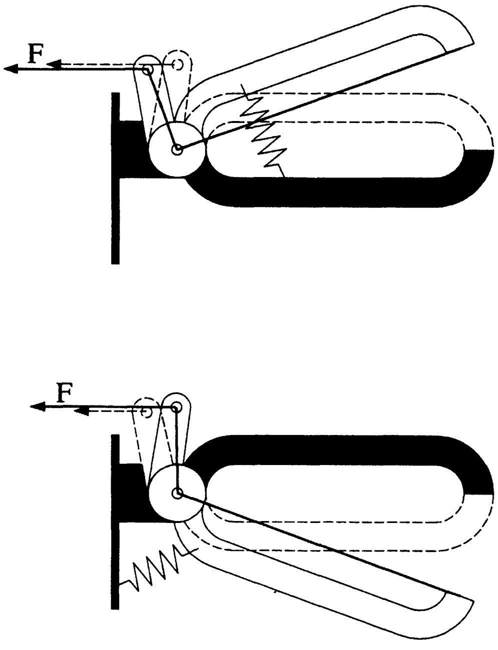

FIG. 6. Schematic showing the difference in operation between a voluntary-opening device (top schematic) and a voluntary-closing device (bottom schematic). A voluntary-opening device has a maximum grip force equal to the force that must be overcome before it will open. A voluntary-closing device has a maximum grip force equal to that applied by the user less the opening force that must be overcome.

2.3.4 Externally Powered Non-Hand-like Prehensors In an effort to capitalize on the success of the split hook, the Hosmer-Dorrance NU-VA Synergetic prehensor uses a split hook in an externally powered configuration. The term synergetic refers to the fact that this prehensor uses two motors within the device, one to provide speed and the other to provide force (Childress, 1973). The Steeper Powered Gripper, available through Liberty Technology, and Otto Bock's latest line of child-size prosthetic hand mechanisms are also devices that use the principle of synergy to boost performance. The Otto Bock "Greifer" is a non-handlike prehensor that uses only one motor. The NU-VA Synergetic Prehensor, the Otto Bock Greifer, and the Steeper Powered Gripper can perform the equivalent of tip, lateral, and cylindrical prehension patterns (see Fig. 7) (Heckathorne, 1992).

2.3.5 Passive Prostheses Devices in this class, usually hands, are regarded as passive or passively functional prostheses. They have no moving parts and require no cables or batteries for operation. They are typically lightweight and reliable. Generic (standard) passive hand prostheses may consist of a cosmetic outer glove over a soft plastic hand with wire reinforcements in the fingers. Highly realistic hands, fingers, and finger parts can be custom sculpted and colored to match an individual's remaining hand.

Another class of passive prosthesis is the work or recreational prosthesis. A custom-designed system that serves only one function may aid the wearer more than one that is supposed to be multifunctional. In such cases the prosthetic device is worn on those occasions when it is needed. These devices range from tool adapters to sports mitts.

2.4 Artificial Arms

If the philosophy of the prosthesis as limb replacement is followed, arm prostheses ought to be designed as an integrated whole rather than a series of modular components. Most of the early artificial arms for the high-level amputee were designed in this manner.

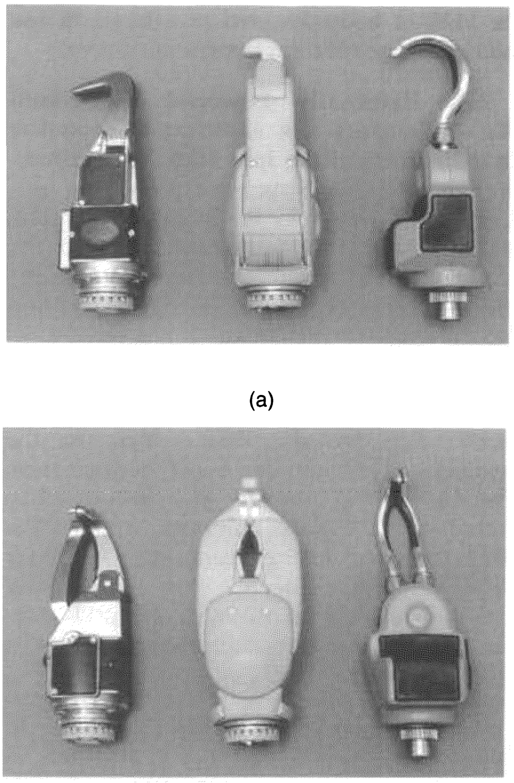

(a)

(b)

FIG. 7. (a) Lateral and (b) "palmar" views of the Steeper Powered Gripper (left), Otto Bock System Electric Greifer (center), and Hosmer NU-VA Synergetic Prehensor (right) (photograph courtesy of Northwestern University Prosthetics Research Laboratory, Chicago, IL).

But as these systems failed to provide the expected function, there was a move away from designing complete arm prostheses to the design of specific or modular components, such as externally powered or body-powered elbow joints, powered or passive wrist rotators, passive humeral rotators, and whole ranges of different hooks, hands, and prehensors (Fryer and Michael, 1992; Heckathorne, 1992; Michael, 1986).

The current trend is to use a "mix and match" approach to make good use of the function available. This modular approach has the advantage of providing great flexibility and practicality for system design. However, it will probably never be able to attain the high functional goals that may be possible from a more integrated standpoint.

Prosthetic components for arm prostheses fall in one of two categories: body-powered or externally powered. Body-powered devices are functional prostheses that use some motion of the body to provide both the excursion and force necessary to power and control the prosthetic component. Externally powered devices, as their name suggests, receive their power from a source external to the body, usually electric.

2.4.1 Body-Powered Prostheses

Modern body-powered prostheses were greatly influenced by techniques used in aircraft flight-surface control. Of particular note was the adaptation by the Northrop Aircraft Co. of the Bowden cable for use in the prosthetics field. These cables, named for their inventor and the founder of the Raleigh bicycle company, have found extensive use in prosthetics applications. Their original use was as brake cables for bicycles. A Bowden cable consists of an inner core cable that is free to move within a sleeve cable that is fixed in place at either end.

The basic principles of the prosthetic design for body-powered devices are the same today as they were in the 1950s. These devices require a harness, to be worn about the shoulders, to which one or more Bowden cables are attached. The conventional below-elbow, body-powered prosthesis has a single control cable that runs from the harness to a terminal device. Terminal-device opening and closing are then controlled by shoulder shrug (biscapular abduction) and/or flexion of the residual upper arm (glenohumeral flexion) (see Fig. 8). For an above-elbow amputee an additional control cable is used to switch control of the harness from terminal-device opening to elbow flexion by unlocking the elbow. Figure 9 shows a typical above-elbow prosthetic fitting (the locking cable passes over the shoulder to the front of the elbow mechanism).

Body-powered prostheses of many kinds are available, and they are the most common kind of prosthesis used in the United States. The reason for their success is thought to be due to the intimate connection the control cable provides between the input and output. Just as a pilot "feels" connected to a plane's control surfaces, so the user of a body-powered limb feels closely connected to the operation of the prosthesis. These prostheses

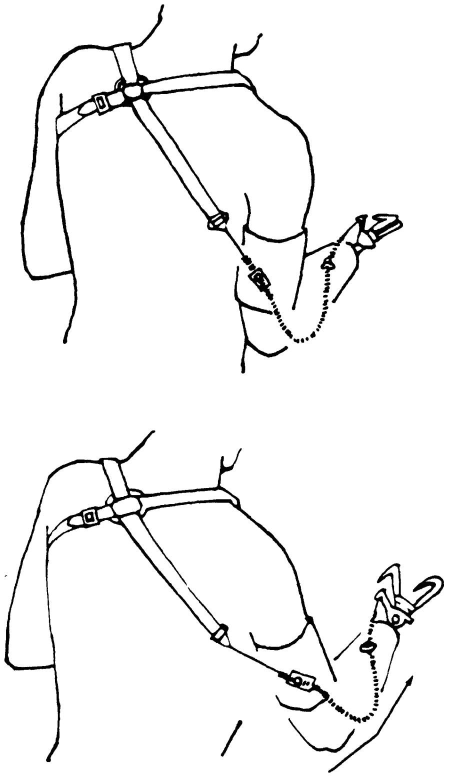

FIG. 8. Sketches of a body-powered control harness showing how glenohumeral flexion is used to control terminal-device opening (sketches courtesy of Northwestern University Prosthetics and Orthotics Center, Chicago, IL).

are also lightweight, durable, and of relatively low cost.

Although generally successful, body-powered prostheses have a number of shortcomings. The major issues are the uncomfortable harness mechanism; the somewhat ungainly control motions, particularly in the case of above-elbow prostheses; restricted range of motion; and limited load-lifting capacity (Fryer and Michael, 1992; Fryer, 1992).

2.4.2 Externally Powered Systems Externally powered prostheses are a relatively new (last 15 to 20 years) addition to the armamentarium of prosthetic devices. The electrically powered prosthesis under the control of myoelectric signals from residual muscles did not become commercially available until the late 1960s and did not gain widespread clinical acceptance until the early 1980s. A myoelectric signal is an electric signal that is generated as a by-product of a contracting muscle. Electrodes on the surface of the skin, placed near the contracting muscle, can pick up this signal and use it for the control of a prosthesis. Electrically powered hands for the below-elbow amputee are the most widely used externally powered prosthetic systems (see Fig. 10).

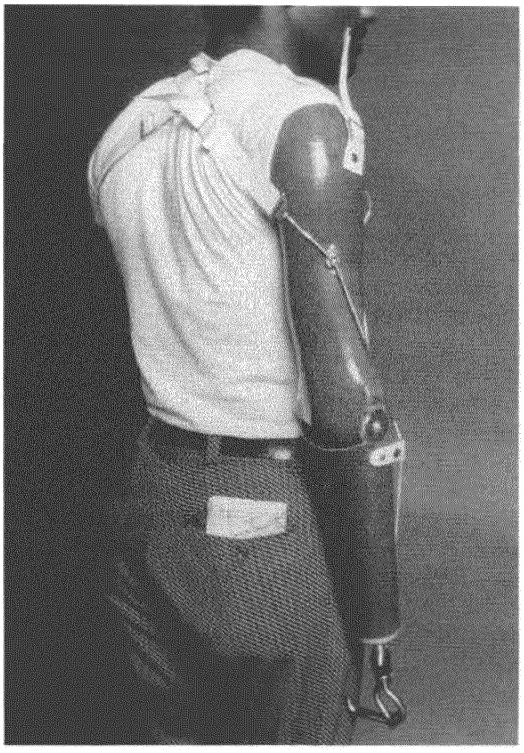

FIG. 9. Above-elbow body-powered prosthetic fitting. The same configuration of control harness as the below-elbow harness is used with the addition of a second "elbow-locking" cable. This cable switches the control from elbow flexion to terminal-device opening by locking and unlocking the elbow in place. (Photograph courtesy of Northwestern University Prosthetics and Orthotics Center, Chicago, IL.)

Currently available externally powered arms consist of an externally powered elbow and terminal device with the possible addition of a powered wrist rotator. The three most common externally powered elbows are the Liberty Technology Boston Elbow II, the Motion Control Utah Elbow, and the Hosmer-Dorrance NY Elbow. All these elbows can be used in conjunction with the previously described powered terminal devices (Heckathorne, 1992; Atkins and Meier, 1989).

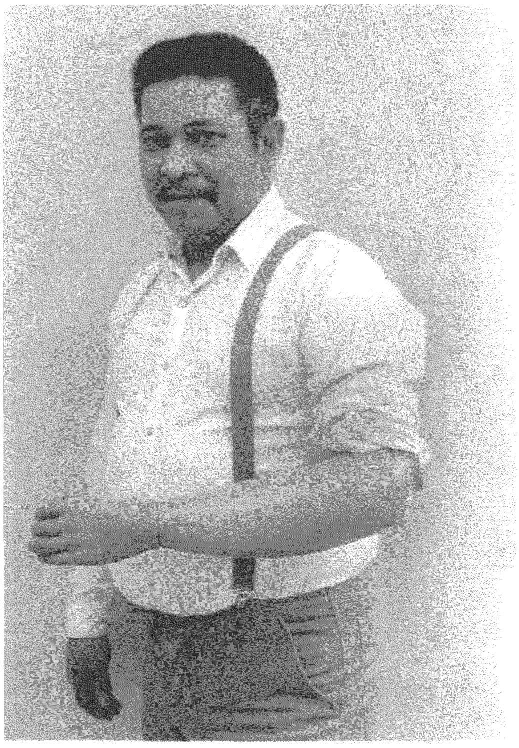

FIG. 10. A definitive below-elbow, externally powered, myoelectrically controlled, prosthetic fitting (photograph courtesy of Northwestern University Prosthetics Research Laboratory, Chicago, IL).

During the 1960s and 1970s much effort was invested in the development of externally powered, multifunctional, complete arm systems. While many of the ideas tried during this period failed to make it from the laboratory into everyday prosthetic practice, a number of multifunctional hands ended up making the transfer to the robotics field. The Belgrade hand (Razic, 1973) and Sven hand (Herberts *et al.*, 1973), while not commercial products, remain in use in robotics research applications. For example, the Belgrade hand has influenced the robotics field in the form of the Belgrade/USC robotic hand (Beattie *et al.*, 1994).

2.4.3 Hybrid Systems When body-powered and externally powered systems are linked together they are called hybrid systems. Hybrid systems are used most frequently with persons who have amputations above the elbow or who have bilateral arm amputations. Such systems can provide the user with the high gripping and/or high lifting capacities of powered systems and the fine control of body power. An example of hybrid control is illustrated in prosthetic limbs for persons with loss of both arms at the shoulder level. By providing the amputee with a body-powered limb on one side and an electric-powered limb on the other side, they enable the wearer to use the limb that is most appropriate for a specific task. This method also enables the limbs to be operated independently of each other, i.e., the body motions required to operate the body-powered side do not influence the state of the powered limb, and vice versa; they are decoupled. A hybrid system of the kind described is shown in Fig. 11.

The amputee in the photograph has a body-powered, single-cable, four-function system fitted on the right side. Chin levers

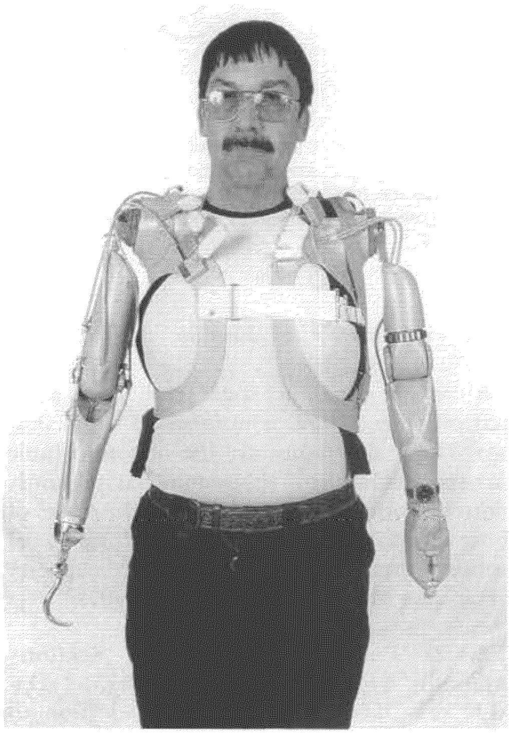

FIG. 11. Hybrid fitting of a bilateral shoulder disarticulation amputee. The body-powered prosthesis (right side in picture) is used as the dominant arm, with the electric limb (left side) assisting as the nondominant arm. See text for description of operation. (Photograph courtesy of Northwestern University Prosthetics Research Laboratory, Chicago, IL.)

are used to lock and unlock both shoulders (MICA locking shoulder joints), the elbow, and the wrist rotator. A lever on the wrist unlocks and locks the wrist flexion unit. Biscapular abduction (shoulder shrug) is then used to actuate unlocked components. The wrist rotator and wrist flexion unit are spring loaded to return these devices to their initial position when unlocked. Gravity is used to return the elbow to its initial position. The shoulder joints are positioned using gravity and bending of the trunk. By locking of all the joints the arm becomes a rigid extension of the user and the principle of extended physiological proprioception applies (see Sec. 2.7.1).

The left side is fitted with a powered elbow (Liberty Technology Boston Elbow) controlled by shoulder elevation pulling against a two-position pull switch. The powered prehensor (Otto Bock Greifer) and an Otto Bock powered wrist rotator are controlled by chin movement against rocker switches actuated by the chin.

The photograph of Fig. 11 illustrates a number of points made previously:

1. Because one side is mechanically powered, while the other is electrically powered, each side can be controlled independently of the other.
2. This fitting shows how components are modularly linked together to create a total arm prosthesis.
3. It shows how high-level amputees are fitted with body-powered and externally powered prostheses.
4. It illustrates how the limbs can be made rigid (high impedance) or free for movement (low impedance).
5. It highlights the difficulty in obtaining sufficient control sources to control a multi–degree-of-freedom arm. Many different methods must be used to control each degree of freedom.

2.5 Control

Numerous methods of control for upper-limb devices have been explored. Chief among these were harness and Bowden cable and/or mechanical switches for body-powered devices and electromechanical switches and myoelectric control for externally powered devices. The available control options can be broken into biomechanical and bioelectric/acoustic. A partial list of current and potential control options follows.

- I. Biomechanical
  - A. Movement/force from a body joint or multiple joints (position, force/pressure)
    - 1. Chin and head force/movement
    - 2. Glenohumeral flexion/extension abduction/adduction
    - 3. Biscapular and scapular abduction (shoulder shrug)
    - 4. Shoulder elevation and depression
    - 5. Chest expansion
    - 6. Elbow or wrist movement
  - B. Direct force/motion from muscle(s)
    - 1. Force/motion from a muscle tunnel cineplasty
    - 2. Force/motion from skin that is adherent to underlying muscle
    - 3. Krukenberg surgical procedure (long transradial amputation)
- II. Bioelectric/bioacoustic
  - A. Myoelectric potentials (muscle electricity)
  - B. Myoacoustic (muscle sounds)
  - C. Neuroelectric potentials (neuron and nerve signals)

Biomechanical inputs have been extensively used in the control of body-powered prosthetic components and have been mentioned in passing throughout the text (see Fig. 11). These same motions can be used to control externally powered components through the use of the appropriate electromechanical switch or force transducer (see Fig. 11). In general, either on/off or proportional velocity control is used with externally powered devices.

Proportional control is a form of control where the amount/intensity of the controlled output variable is directly related (proportional) to the amount/intensity of the input signal. For example, the output speed of a variable-speed power drill may be proportional to the amount the trigger is pressed. This is an illustration of velocity (speed) control; that is, output speed is proportional to the amount of input signal (in this case how far the drill trigger is pressed in). Proportional control is used where a graded response to a graded input is sought.

Position control of upper-extremity function relates prosthesis position directly to the input control signal. The amputee's ability to perceive and control prosthesis position is directly determined by his ability to perceive and control the input signal. With position control, unlike velocity control, the user must maintain the input signal to hold an output level other than zero. As the goal of the arm is to position the hand in space, the use of position control for positioning of the terminal device in space is considered to be most appropriate. Velocity control may be better suited for the control of prehension (Childress, 1992).

2.5.1 Myoelectric Control While most of the currently available externally powered upper-limb prosthetic systems use some form of velocity control, the merits of velocity control are being reevaluated, particularly in the context of position control, which in some cases appears preferable. Myoelectric control has received considerable attention since it first appeared in the 1940s. It was advanced as a natural approach for the control of prostheses since it made it possible for amputees to use similar mental processes to control prosthesis function as had previously been used in controlling their physiological limb. Myoelectric control uses the remaining neuromuscular systems in the amputation stump.

Myoelectric control derives its name from the electromyogram (EMG) signal, which it uses as a control input. When a muscle contracts it produces electric potentials (recorded as the electromyogram or EMG), as a by-product of that contraction. The intensity of the EMG signal produced increases as muscle tension increases. Surface electrodes placed on the skin near a muscle can detect this signal. The signal can then be electronically amplified and processed and used to control a prosthesis (see Fig. 12). See Parker and Scott (1985) for further reading.

2.5.2 Myoacoustic and Neuroelectric Control Myoacoustic signals are auditory sounds created as a by-product of muscle contraction. A myoacoustic control system is very similar in structure to a myoelectric control system but the elimination of unwanted acoustic noise may be a bigger problem than the elimination of unwanted electrical noise in myoelectric systems (Barry and Cole, 1990). It has not gained widespread use.

Neuroelectric control or the control of a prosthetic device by way of nerve impulses would appear to be a natural control approach. Although there is, and has been, research concerning prosthesis connections with nerves and neurons (Edell, 1986; Kovacs *et al.*, 1989) the practicality of human–machine interconnections of this kind is still problematic. Nervous tissue is sensitive to mechanical stresses and this form of control requires the use of implanted systems.

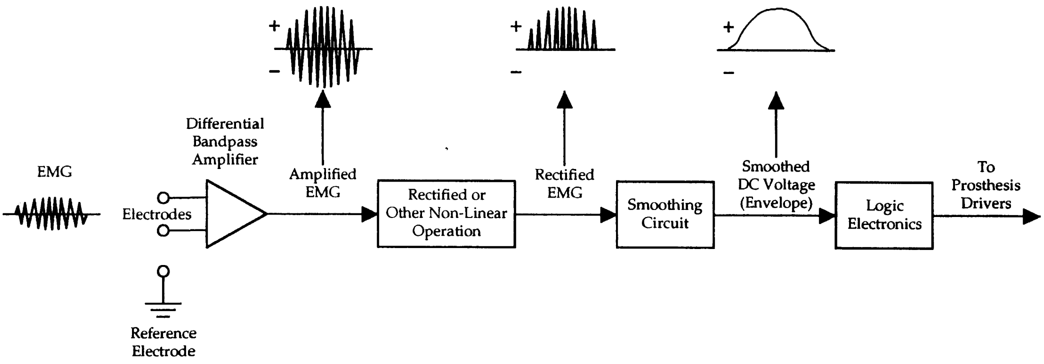

FIG. 12. Drawing showing myoelectric signal (EMG) processing in a typical myoelectric control system. Three "dry" metal electrodes are always associated with each muscle site. The small ac potential from the muscle is amplified and changed into a dc potential, often by rectification or by squaring. In a typical circuit this dc potential is commonly smoothed with a low-pass filter. The smoothed dc voltage E_s can be compared in a logic circuit with a threshold voltage E_{th}. If E_s is greater then E_{th}, then power is supplied to the prosthesis motor. If E_s is less than E_{th}, no power is supplied, and the motor is off.

2.6 The Issue of Feedback

Myoelectrically controlled systems are usually open-loop feed-forward in nature. That is, a control signal is sent from the user to the device but nothing in the way of feedback is returned except visual information. This lack of proprioceptive and sensory feedback is considered significant with either myoelectric or switch-controlled externally powered systems. Without proprioceptive feedback, subconscious multifunctional control of a prosthesis is probably not possible.

Constant visual monitoring is required for effective operation of velocity-control systems. For the control of multiple degrees of freedom this places an excessive mental burden on the user, greatly diminishing any benefits a prosthesis of complexity might offer. Visual and auditory feedback are slower, less automated, and less programmed than normal proprioceptive feedback and therefore place a greater mental burden on the operator (Soede, 1982).

Upper-limb prostheses are currently controlled primarily through *visual feedback* with some assistance from feedback that is incidental rather than by design (see Fig. 13). Examples of *incidental feedback* are motor whine, prosthesis vibration, socket forces, etc. Deliberate attempts have been made to provide feedback by means of mechanical vibration of the skin and by electrical stimulation of the skin. These forms of feedback are called *supplementary sensory feedback (s.s.f),* because they supplement the normal sensory feedback paths of the body. *Artificial reflexes,* which bypass the human operator altogether, are another method employed to reduce mental loading and improve performance. Artificial reflexes seek to remove the operator from the control loop and automatically respond to some external sensor input. They are essentially closed loops within the mechanism/prosthesis itself. A slip detector, for example, detects slippage of an object grasped by a prehensor and automatically increases prehension force until it detects that the slippage has stopped.

Feedback information can be received through the control interface, as in power steering for an automobile. *Control interface feedback* can provide feedback to the operator concerning the state of the prosthesis through the same channel through which the prosthesis is controlled. Information concerning prosthetic joint position, velocity, and the forces acting on it can be available to the operator through the proprioceptors of the controlling human joint. Feedback through the control interface can be in forms that are easily interpreted by the user and can be interpreted at a subconscious level. Consider the example of power steering for a car. When parking, the driver learns subconsciously to equate wheel position with steering-wheel position. Central to the implementation of feedback through the control interface is the design of the controller. The majority of prosthesis controllers, however, do not provide control-interface feedback, and it is for this reason, it could be argued, that externally powered limb replacements have not gained the widespread acceptance expected of them (Childress, 1992).

2.7 Parallel Development of Robots, Manipulators, and Prostheses

If one looks at the parallel development of aircraft, manipulator, and automobile control systems, the progression starts out with

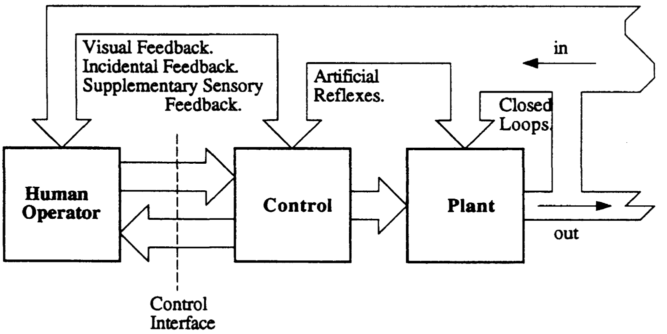

FIG. 13. Block diagram of the various feedback paths used in upper-limb prosthetics. See text for explanation.

the flight surfaces, the steering wheels, the remote manipulator, or the prosthetic joint controlled directly by mechanical linkages between the operator and the device. In early aircraft the tail flaps and rudder were physically connected to the joystick by way of control cables. In early automobiles the front wheels were physically connected to the steering wheel by a rack-and-pinion mechanism. The early master–slave manipulators were physically interconnected by cables. The early prostheses used a cable attached from the artificial joint to the shoulder harness. In all these cases the power to perform the control task was supplied by the human and feedback was available to the operator via the control interface. The pilot could "feel" what was happening to the aircraft flight surfaces through the joystick. The amputee had a sense of prosthesis state through the control cable.

As aircraft got larger and faster hydraulic power assist was added to the cable controls. As the automobile got heavier power steering was developed. The control input (the operator at a joystick or steering wheel) remained the same but instead of the operator powering the system, power was now provided by hydraulic systems that augmented the cable systems of aircraft and the steering of cars. But the same type of feedback, control-interface feedback, was maintained. Position, velocity, and force information were still fed back within these "boosted" systems. A similar pattern of development occurred with remote manipulators (i.e., human-powered cable systems were followed by electric-motor–boosted cable systems).

Unfortunately, it is at this stage that prosthetics development diverges from that of aircraft, automobiles, and remote manipulators. When prosthetic arm technology moved to externally powered systems, the control modalities shifted away from cable inputs that provided sensory and proprioceptive feedback to the operator and with one or two exceptions used open-loop velocity-control techniques (such as myoelectric and switch control) that provide little feedback to the operator beyond visual feedback.

This divergence from the path taken by aircraft, automobile, and remote manipulators can be attributed to the demanding physical design constraints of volume, weight, and power that must be met if a prosthesis is to be successful. The constraint problem resulted in essentially open-loop, velocity-controlled, externally powered devices. The major reason for this approach was simplicity. The actuator of choice, the electric dc motor, is inherently a rate-controlled device (i.e., its output speed is directly proportional to the input voltage). In addition a velocity-controlled system does not need to draw power to maintain a particular position.

While upper-limb prosthetics control development seems to have been retarded, the development of controls for the aircraft and manipulator industries has moved into the next stage and beyond. Aircraft controls, building on work from the field of teleoperation or remote manipulation, have done away with the direct physical interconnection of the pilot and flight surfaces and now use fly-by-wire systems. In these systems the pilot is connected to the output through electrical wire connections. Force feedback or force reflection is used to provide the pilot with the appropriate feedback or "feel."

For advanced fighter planes the stage beyond fly-by-wire has been reached. These planes are inherently unstable, i.e., the speed of response and accuracy required to ensure stable operation are beyond human capability. Thus, a human operator is incapable of controlling them alone, and so now "boosted control" has been added to "boosted power." A similar stage of development has been reached in the area of force-reflecting remote manipulators. Here the kinematic and dynamic properties of a microslave manipulator are scaled, using a technique called bilateral impedance shaping (Colgate, 1993) to match those of the controlling master arm so that the contact forces experienced by this manipulator "feel right" to the operator. Bilateral impedance shaping is a means of constructively altering the impedance of a task. This is particularly valuable if the task impedance or feel is very different from those normally experienced by the human limb, such as in the case of the human arm controlling a micromanipulator. See Thring (1983) for further reading.

2.7.1 Extended Physiological Proprioception The work of Simpson (1974) was an exception to the trend taken by upper-limb prosthetics into open-loop velocity control. Simpson advocated essentially the same developmental process as that taken by the aircraft and auto industries, that of augmented cable control. He demonstrated that this was a viable approach for the subconscious control of multifunctional prosthetic arms [the Edinburgh arm (Simpson, 1969)] and coined the phrase "extended physiological proprioception" (e.p.p.) to indicate that the body's own natural physiological sensors are used to relate to the operator the state of the prosthetic arm. That is, the operator extends his own proprioception into the prosthesis; the prosthesis becomes an extension of the amputee's self. Simpson's work was largely ignored at the time and is only now beginning to receive serious attention in the clinical setting. The success of the conventional, body-powered, Bowden-cable–operated prostheses, it could be argued, can be attributed to this idea of e.p.p.

The use of tools such as pencils, pens, knives, hammers, and the like illustrates the simplest form of e.p.p. We naturally use these extensions of our bodies without thinking about their control, because we extend our proprioception through these tools. Similarly, blind persons uses long canes to extend their range of feel of the outside world. A golfer or tennis player does not watch the head of the club or racket during a swing. Both are able to judge the angle and position of the head of the club or racket in space based on how it feels in the hand. In both cases the proprioceptive capabilities of the wrist and other joints have been extended to include the club or racket. The club or racket becomes a natural extension of the player's hand/arm.

The success of the below-elbow myoelectric hand prosthesis can be attributed to e.p.p. in its simplest form. The below-elbow prosthesis socket can be viewed as an extension of the elbow joint with the hand itself behaving in a mode similar to a pair of pliers. A prosthesis for the lower-limb, below-knee amputee is another example of simple e.p.p. The below-knee prosthesis becomes an extension of the shank from the knee.

Control of *another joint* by a physiological joint through e.p.p. is more complex than the e.p.p. control of a simple rigid extension. The artificial joint may be powered by the physiological joint as in body-powered prostheses, or it can be externally powered, receiving only the control input from the physiological joint. A prerequisite for the e.p.p. control of another joint is that the two joints, the physiological joint and the artificial joint, be *mechanically interconnected.* In this way, motions and forces experienced by one joint result in equivalent motions and forces being experienced by the other, and vice versa (see Fig. 14). This type of system is known as an unbeatable position servomechanism. A hand-operated bicycle brake is a rudimentary example of an e.p.p. system. These systems are cable-operated manual-control systems similar to the cable-control mechanisms of early biplanes and body-powered prostheses. The cyclist is able to achieve fine braking control because it is possible to sense how hard the brakes are being actuated through the cable connecting the brake to the brake handle.

As previously mentioned, the success of conventional body-powered devices has been attributed to the e.p.p. that is inherent in their mode of operation, particularly those devices for below-elbow amputees. The use of Bowden cables means that external forces experienced by the artificial joint/terminal device are transmitted, via the cable, to the shoulder harness and felt across the shoulders.

Control of an externally powered prosthetic component by e.p.p. is similar in concept to power steering for a car. The driver feels, through the handling of the steering wheel, the action of the car's front wheels. Furthermore, the driver cannot "beat" the response of the front wheels because the steer-

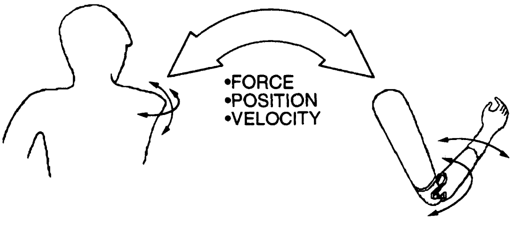

FIG. 14. Diagrammatic representation of a two–degree-of-freedom e.p.p.-controlled electric prosthesis. Shoulder elevation/depression would be linked to elbow flexion/extension. Shoulder protraction/retraction would be linked to wrist pronation/supination. Feedback is through the control interface that actuates the prosthesis so that the operator gives and receives information in a way that is similar to the interaction found in extension tools used in the human hand (e.g., hammer).

ing wheel and the front wheels move together in a one-to-one correspondence.

E.p.p.-like control can be implemented using two motors instead of using Bowden cables for the linkage between the natural and artificial joints. In such a system, one motor is used to drive the prosthesis and the second motor is used to implement the restraining linkage that the cables previously provided. The second motor provides the proprioceptive feedback. This combination of two motors is referred to as a master–slave system. One motor is slaved to the other. This use of a dual-motor master–slave servo system has been applied in the area of human amplification and remote manipulation, where reduction of weight, space, and power consumption are not primary concerns (see Thring, 1983).

2.8 Surgery for Improved Control

The role of surgery is vital to the success of a prosthetic fitting. History has shown that the greatest success in prosthetic fitting is achieved when the surgeon, amputee, prosthetist, physical therapist, physician, and engineer work together as a team. Muscle-tunnel cineplasty is a good example of where this interdisciplinary approach was needed to ensure success. In muscle-tunnel cineplasty surgical intervention is used to provide direct mechanical connection between muscle and prosthesis. The procedure involves the construction of a skin-lined tunnel, by way of skin grafts, through a muscle belly (see Fig. 15). Developed in its original form by the German surgeon Ferdinand Sauerbruch (1916), the procedure owes much of its success to the team approach used at Sauerbruch's facility. Sauerbruch's technique was modified by Lebsche in Munich, and the modified approach was used in the United States after World War II.

The use of surgical techniques to create pectoral or other tunnel cineplasties offers a means of creating badly needed control sources for persons with high levels of amputation. The use of tunnel cineplasties as control inputs to externally powered prosthetic components provides a means of physically linking a skeletal muscle to an externally powered device. If the controller embodies the concept of extended physiological proprioception (e.p.p.) then feedback information about prosthesis state can be given to the amputee through the cineplastized muscles' own proprioceptive and sensory apparatus. It is possible to envision a below-elbow amputee with independent multidigit control through the use of multiple miniature forearm cineplasties providing control inputs to multiple e.p.p. controllers, one cineplasty per controller per digit to be controlled (see Fig. 16). Alternatively it might be possible to create multiple miniature pectoral cineplasties, for high-level above-elbow amputees, to control a multi–degree-of-freedom arm prosthesis in a more subconscious manner than is now possible.

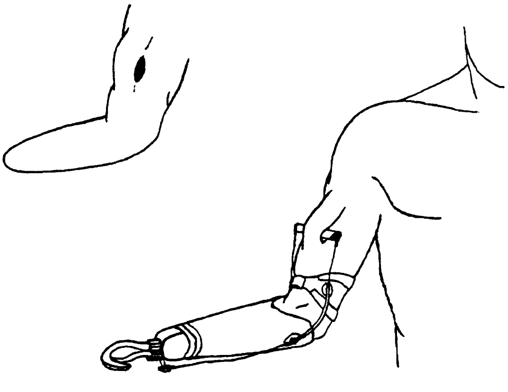

FIG. 15. Sketch of a Sauerbruch/Lebsche biceps muscle-tunnel cineplasty and prosthesis. A skin-lined tunnel is surgically constructed through a skeletal muscle that has been freed from its point of insertion (the biceps in this case). A pin is then placed in the tunnel and mechanically attached by way of cables to a prosthetic terminal device. Traditionally a voluntary-closing split hook was used in the United States. Voluntary contraction of the cineplastized muscle causes the pin to move, resulting in terminal-device closure.

A variation on the surgical procedure for muscle-tunnel cineplasty is the tendon-exteriorization cineplasty procedure (Beasley, 1966). The procedure uses tendon transfers combined with skin flaps to bring a tendon loop outside the body. This procedure might be used to create exteriorized tendons for the control of a four–degree-of-freedom prosthetic hand (see Fig. 17).

In Sweden, pioneering work in the area of direct skeletal attachment of prostheses to humans has been carried out by Dr. Per-Ingvar Brånemark and his team. These surgeons and orthopedic engineers have created

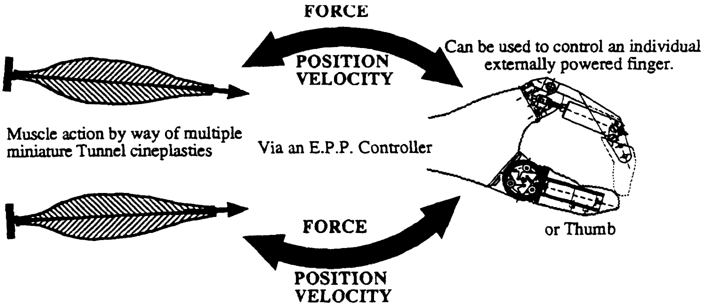
FORCE
POSITION
VELOCITY
Can be used to control an individual externally powered finger.
Muscle action by way of multiple miniature Tunnel cineplasties
Via an E.P.P. Controller
FORCE
POSITION
VELOCITY
or Thumb

FIG. 16. Schematic of multiple miniature tunnel cineplasties, used in conjunction with e.p.p. controllers, controlling a multifunctional hand prosthesis. It is envisaged that each prosthetic digit would be under the control of an individual muscle with a miniature tunnel cineplasty via an e.p.p. controller. This presents the intriguing possibility of someday achieving subconscious coordinated multidigit control for externally powered hand prostheses.

infection-free interfaces for direct skeletal attachment systems for upper- and lower-limb amputations. This work evolved from work with titanium implants in the jawbone for holding artificial teeth and progressed to the placement of titanium sockets in the skull to retain maxillofacial prostheses and for the attachment of highly efficient bone-conduction hearing aids (Brånemark, 1985). Brånemark seems to have overcome the infection problems that have curtailed previous direct skeletal attachment efforts (Hall *et al.*, 1967; Hall and Rostoker, 1980) with innovative surgical techniques and post-operative procedures. Should direct skeletal attachment prove itself viable, it could revolutionize the prosthetic fitting of amputees (upper and lower).

2.9 Medical Manipulators or Rehabilitation Robots

Medical manipulators form a special class of artificial arm replacement. These devices are mechanical arms that attempt to provide arm function for persons who have lost control of their arms, such as in high-level quadriplegia. These devices fall into one of three categories:

1. powered orthoses,
2. industrial or custom-designed manipulators for use with a dedicated work station in highly structured environments, or
3. custom-designed manipulators for use in unstructured environments.

The initial work in the field of rehabilitation

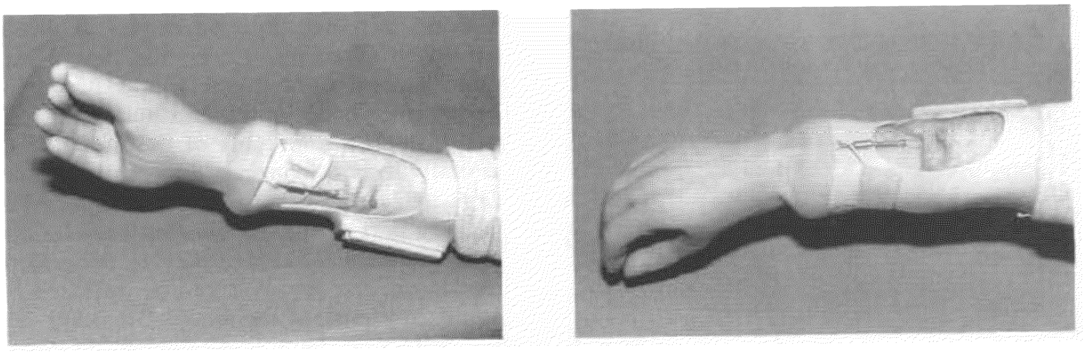

(a) (b)

FIG. 17. Subject with tendon-exteriorization cineplasties wearing a prototype full e.p.p.-controlled, externally powered hand prosthesis. Top photograph shows the connection to the flexor compartment tendon loop. Bottom photograph shows the connection to the extensor compartment tendon loop. The e.p.p. controller and mechanical linkage are arranged to provide full motion of a modified Otto Bock hand mechanism. This was a proof-of-concept fitting done to examine powered prosthesis control using direct muscle attachment. This was the first step toward the goal of controlling a four–degree-of-freedom hand using direct muscle attachment. (Photographs courtesy of Northwestern University Prosthetics Research Laboratory, Chicago, IL.)

manipulators used the idea of anthropomorphic systems and developed powered orthoses that were essentially externally powered exoskeletons that moved a paralyzed person's arm to perform a desired manipulation. These systems later evolved into mechanical arms that were mounted on a wheelchair or in a dedicated work space. The paralyzed limb is not transported in later systems. See American Academy of Orthopaedic Surgeons (1981) for further reading.

Persons with high-level quadriplegia do not have the use of their arms; consequently the anthropomorphic master–slave approach of conventional manipulator design is not necessarily the most practical design approach. Yet many medical manipulators evolved from or use upper-limb prosthetic components. The APL/JHU manipulator (Seamone and Schmeisser, 1972) (Fig. 18), developed at the Applied Physics Laboratory of Johns Hopkins University, evolved from a shoulder-level arm prosthesis, while the Stanford Robotic Aid (Leifer *et al.*, 1978) (Fig. 19), which is built around a commercially available industrial robot, the Unimation PUMA-260 arm, uses an Otto Bock "Greifer" prosthetic prehensor as its end effector. Both these systems are remote from the user and operate in dedicated work spaces. See Leifer (1981) and Preising *et al.* (1991) for review.

As with myoelectrically controlled prostheses, the control of these devices is essentially open-loop feed-forward in nature. The burden of feedback and monitoring of manipulator position is placed totally on the visual system. Subconscious control is lost, with concomitant increase in mental loading to the user.

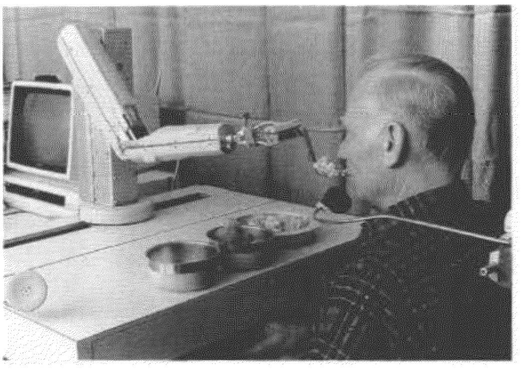

FIG. 18. The APL/JHU rehabilitation manipulator. This device evolved from an externally powered shoulder-level prosthesis system. The end effector is a modified voluntary-opening split hook. (Copyright © 1986, The Johns Hopkins University/Applied Physics Laboratory. All rights reserved. Reproduced by permission.)

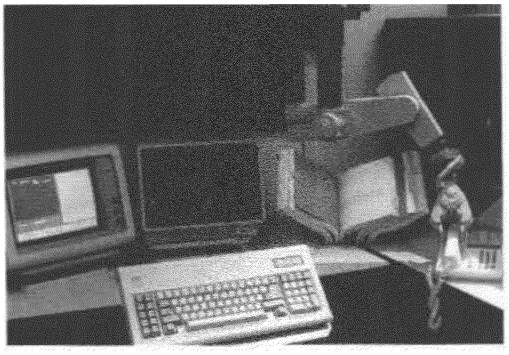

FIG. 19. DeVAR, the VA/Stanford Robotic Aid. Built around a small industrial robot, the Puma-260 arm, this device uses a prosthetic hand replacement, the Otto Bock Greifer, as its terminal device. This system is used in a highly structured environment. (Photograph courtesy of Machiel Van der Loos, Department of Veterans Affairs Medical Center, Palo Alto, CA.)

The Handy 1 Rehabilitation Robot from Rehab Robotics Ltd., Staffordshire, England (Fig. 20), has achieved the greatest field success so far in the area of personal robotic assistants. It is the first commercially available robotic aid capable of assisting the severely disabled with everyday tasks such as eating, drinking, shaving, etc. Developed at Keele University (Topping, 1995) this device has found success through the simple nature of the interface between the user and the device and through the mobility of the unit (it can be easily moved from room to room). A single switch is the sole mode of control and is the only part of the system that may be different for each individual person. Because of this feature the Handy 1 can be operated by almost any part of the body. Originally designed for persons with cerebral palsy, stroke accident, multiple sclerosis, and muscular dystrophy, the Handy 1, in the eating mode, presents the user with food. The user then uses gross trunk movements to take the food from the robot. The device can be used in a similar fashion to assist with grooming and personal hygiene. It is this limited idea of presenting objects (e.g., comb, washcloth,

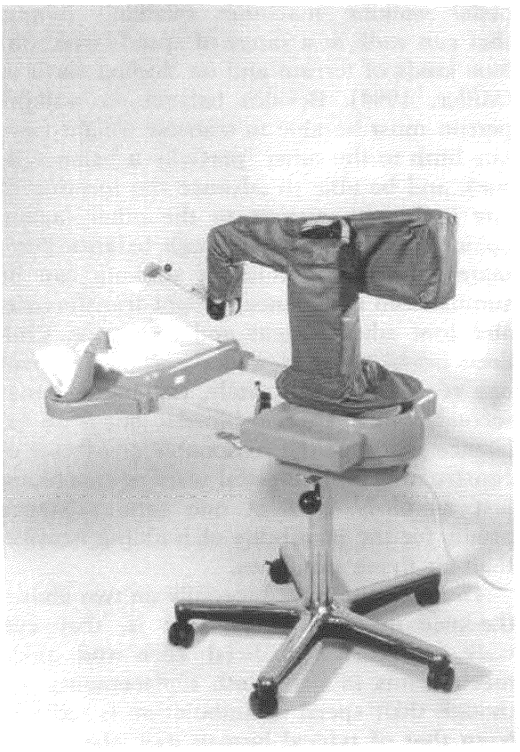

FIG. 20. The Handy 1 Rehabilitation Robot. This device owes its success to the simplicity of its user interface and to its ability to be moved easily from place to place. An object (food, toiletries, etc.) is "presented" to the user who then employs gross trunk movements to eat or to use the presented object. (Photograph courtesy of Rehab Robotics, Keele, England.)

etc.) rather than trying to be a mechanical arm replacement that is at the heart of this device's success.

3. LOWER-LIMB PROSTHESES AND WALKING ROBOTS

Humans can ambulate reasonably well on artificial legs, particularly if the amputations are below the knee. Persons with one limb missing below the knee can walk almost normally if they are reasonably healthy, if they have good gait training, and if they have a well-fitted functional prosthesis. Walking on two artificial below-knee prostheses is something akin to walking on stilts and many people can also walk quite naturally with two limbs missing below the knee. Today's below-knee prostheses use modern materials to achieve combinations of shape and compliance that are appropriate for walking and running. Figures 21 and 22 illustrate prostheses of this nature. The runner in Fig. 21 has run 100 meters in 11.65 s. These prostheses, products of Flex-Foot Inc., are made of composite carbon fiber material. The foot is shaped in the form of a flexible cantilever beam that bends during walking. In addition, the pylon to which the cantilever is attached, in Fig. 22, is pistonlike and will shorten under load through the readily visible side leaf spring. Control of the limbs comes through intimate fitting sockets that provide extended physiological proprioception (e.p.p.) from the knee. Foot function is achieved through shape and through bending (compliance),

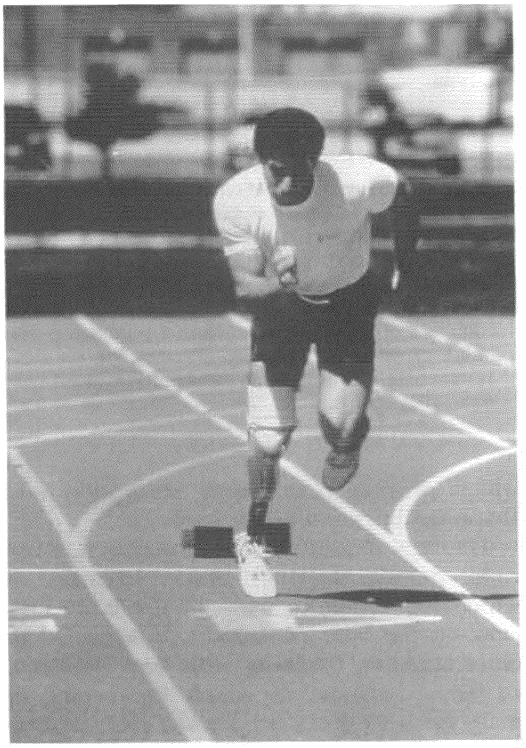

FIG. 21. Dennis Oehler, one of the fastest leg amputees in the world, running on a Flex Foot®. He has run the 100 meters in 11.65 s. The Flex Foot® has a flexible keel that stores energy during the stance phase of walking or running and releases this stored energy at toe off to provide an added boost to the user. Feet of this type have been termed dynamic response feet. (Photograph courtesy of Flex Foot® Inc., San Diego, CA.)

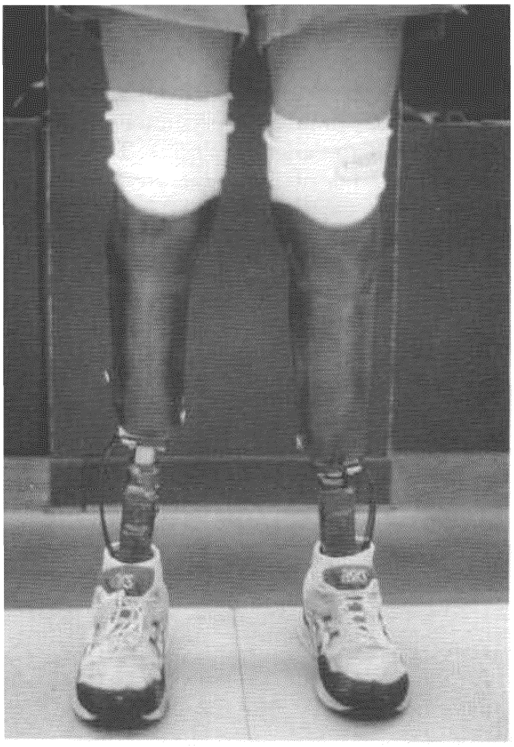

FIG. 22. The latest incarnation of the Flex Foot®, the Reflex Foot®, still has the carbon-fiber keel of the original Flex Foot® but now also includes a vertical leaf spring and piston on the shaft of the leg. These springs can be seen in the above photograph. These springs act essentially like shock absorbers, limiting the amount of shock transmitted to the amputee during stair climbing and walking. (Photograph courtesy of Northwestern University Prosthetics Research Laboratory, Chicago, IL.)

which provide both shock absorption and shape modification under load.

Humans require good balance to stay upright during walking. Static balance, in which the center of mass of the body projects on the base of support, is achieved in quiet standing (walking with zero velocity). Dynamic balance, in which the center of mass often projects outside the base of support, is required for bipedal walking. Walking is often regarded as a process of falling and catching oneself. Perry and associates (1981) have described walking as a continuous event of falling from one foot to the other to move the body forward. Balance is the key to bipedal walking, and the lack of good artificial balancing ability is one of the primary reasons it is difficult to build bipedal walking machines (walking robots) that can walk at a range of speeds over various kinds of terrain and on sloping surfaces (Miller, 1994). Besides balance, a walking person must be able to transfer weight from one limb to the other (partially a balance issue), and be able to advance the location of one limb with respect to the other (again, balance is involved). Although balance dominates, the main issues of walking can be summarized as balance, weight transference, and foot advancement and clearance. Children begin to learn walking skills at about one year of age. Ultimately, walking becomes second nature. Human bipedal gait is an existence proof for the reasonable possibility of constructing agile bipedal walking machines, just as birds were at one time existence proofs for the possibility of building heavier-than-air flying machines.

Humans can walk bipedally on two above-the-knee artificial limbs; that is, they can walk on passive artificial knee and ankle mechanisms in both limb replacements. Although their speed of ambulation is reduced from that of typical human gait, the ambulation ability they achieve is a tribute to their excellent balance system and to their control of muscles around the hip joint and pelvis (see Fig. 23). A person amputated at the hip level on one side can also walk reasonably well on a hip disarticulation prosthesis. Walking ability of this nature testifies also to the capacity of a remaining limb to compensate for one that is missing.

Legged locomotion is an interesting area of robotics, prosthetics, and animal studies. Raibert (1984, 1990) points out that legged locomotion has potential practical applications for vehicular mobility over rough terrain. He also points out that technical development of walking machines may assist with the understanding of animal movement, including human bipedal ambulation. Likewise, he notes that animal studies may assist with the design of walking machines. It might be added that study of walking on prostheses may help us understand normal human walking and/or robotic walking and that knowledge about robotic walking and human walking may influence artificial limb designs. This kind of interplay between disciplines occurs often in science and engineering. It is worth noting, for example, that the photographic studies of animal locomo-

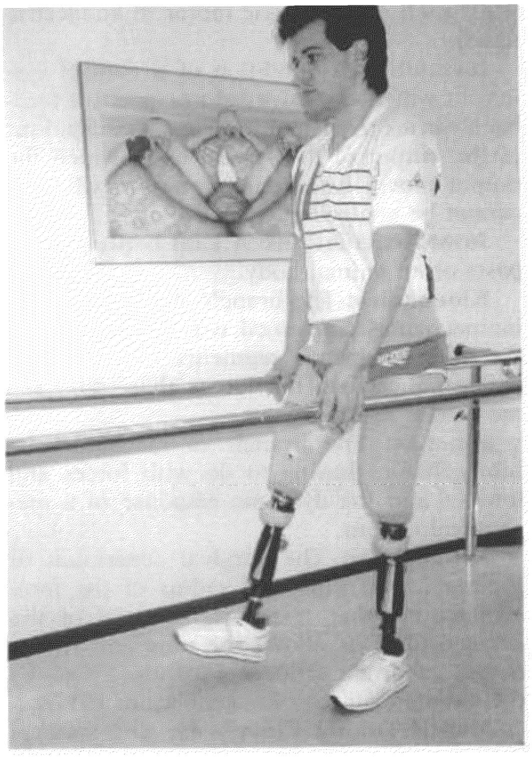

FIG. 23. A bilateral above-knee amputee walking bipedally on two above-the-knee artificial limbs. In the photograph he is shown walking between parallel bars during his early gait training. He now walks with the aid of a cane, which is primarily an aid to balance. (Photograph courtesy of Northwestern University Prosthetics Research Laboratory, Chicago, IL.)

tion (originally studies of horses) by Edward Muybridge in the United States and by E. J. Marey in France played important roles in the development of cinematography.

Comprehensive introductions to locomotion and walking machines may be gained through Todd (1985) and Raibert (1986), and through two special issues on legged locomotion published by the *International Journal of Robotics Research* (see Raibert, 1990, 1984). Suggested introductions to human walking are Perry (1992) and Rose and Gamble (1994). Perry's book examines human pathological ambulation as well as normal walking. The Rose and Gamble book contains a chapter by C. W. Radcliffe concerned with walking on prostheses. McMahon (1984) provides readers with a general overview of human and animal locomotion issues.

4. CONCLUSION

Presently, the field of limb prosthetics finds itself in a stage of development similar to that of the aircraft industry back in the 1930s, when the DC-3 airplane was introduced. In the aircraft industry this plane represented a watershed between seat-of-the-pants empirical design and scientific design using theoretical constructs of aerodynamics. The field of prosthetics seems to be at a similar point in its development. Significant advances have been made, to a great extent by empirical means. Now, near the end of the twentieth century, it is anticipated that further gains will be made through the enhancement of empiricism by theoretical and computational science.

GLOSSARY

Abduction (Adduction): The lateral movement of the limbs away from (toward) the median plane of the body.

Above-Elbow: Refers to the those amputations that occur above the elbow and below the shoulder (transhumeral).

Agonist/Antagonist: Muscles acting in groups or pairs. For any given movement some muscles in the group or one muscle in the pair must be either relaxing or contracting. The agonist muscle is the muscle that is directly engaged in contraction as distinguished from muscles that have to relax at the same time; thus in bending of the elbow, the biceps brachii is the agonist and the triceps the antagonist.

Anthropomorphic: Human shaped. Attribution of human qualities to nonhuman organisms or objects.

Artificial Reflex: A reflex that bypasses the human operator altogether for the purpose of reducing the mental loading placed on the operator and to perform an operation automatically without requiring intervention from the operator. Artificial reflexes seek to remove the operator from the control loop and automatically respond to some external sensor input. For example, a slip detector detects the slippage of an object and automatically increases prehension force until the slippage stops.

Below-Elbow: Refers to those amputations that occur below the elbow (transradial).

Bilateral: Both sides.

Bilateral Impedance Shaping: Used in the area of remote manipulation, bilateral impedance shaping is a means of constructively altering the impedance or "feel" of a task. This is particularly valuable if the task impedance or feel is very different from those normally experienced by the human limb.

Body-Powered: Driven by power provided by the amputee through some body motion.

Cocontraction: The simultaneous contraction of both muscles in an agonist/antagonist muscle pair.

Contact Instability: The instability that can occur when a robot manipulator comes into contact with a surface.

Control-Interface Feedback: Sensory feedback information that is received through the control interface—e.g., as in power steering for an automobile. Control-interface feedback provides feedback to the operator concerning the state of the prosthesis through the same channel through which the prosthesis is controlled.

Degrees of Freedom: The number of axes of rotation or joints that need to be controlled.

Dynamic Cosmesis: The appearance of a prosthesis or device when it is in motion. Good static cosmesis does not necessarily ensure good dynamic cosmesis.

Extended Physiological Proprioception (e.p.p.): A concept used to explain how a person is able to extend the body's proprioceptive capabilities into objects held in the hand or controlled by physically linking the object to a physiological joint.

Externally Powered: Driven by power from a source external to the body—e.g., batteries to drive an electric motor.

Feed Forward: Process control in which changes are detected at the process input and an anticipating correction signal is applied before process output is affected.

Flexion/Extension: The act of bending/unbending or condition of being bent/extended; decrease/increase of the angle between the bones forming a joint (e.g., in bringing the hand to the mouth the elbow of the arm flexes).

Incidental Feedback: Those forms of sensory information that are available to the amputee by default rather than by design (e.g., pitch of an electric motor in an electric hand).

Instability: A condition of a control system in which excessive gain or positive feedback causes persistent, unwanted oscillations in the output of the system or in which the output moves to a constraint from which it cannot be controlled.

Joint Articulations: A joint between rigid parts of an animal body.

Kinematics: The branch of mechanics or biomechanics concerned with description of the movements of segments of the body without regard to the forces that cause the movement to occur.

Kinetics: The branch of mechanics or biomechanics having to do with forces and torques and the dynamic response of a mechanical system.

Krukenberg: The surgical separation of the remaining ulna and radius of the forearm stump after traumatic removal of the lower arm. This allows the ulna and radius to act as crude pincers. First described by the German surgeon Krukenberg in 1917.

Muscle-Tunnel Cineplasty: The surgical alteration of the muscles in a stump after amputation so that it is possible for them to impart motion and direction to an artificial limb. Developed in its modern form by the German surgeon Sauerbruch in 1916.

Myoacoustic: The acoustic sounds created as a by-product of muscle contraction.

Myoelectric: The electric signal created as a by-product of muscle contraction.

Neuroelectric: The electric signal created by the firing of a nerve.

Open-Loop: A system is open-loop when the system outputs are controlled by system inputs only, and no account is taken of *actual* system output.

Oscillation: Any effect that varies periodically back and forth between two values.

Passive Prosthesis: A prosthesis that is neither body-powered or externally powered.

Prehensor: A gripping device that may or may not look like a hand.

Remote Manipulation: The control of a manipulator that is physically remote from the operator.

Static Cosmesis: The appearance of a prosthesis when it is stationary. Good static cosmesis implies an aesthetically pleasing device.

Supplementary Sensory Feedback (s.s.f.): The deliberate attempt to provide additional sensory feedback, usually by mechanical means (e.g., mechanical vibration of the skin). These forms of feedback supplement the normal sensory feedback paths of the body.

Terminal Device: The gripping mechanism used by a prosthesis, robot, or manipulator.

Unbeatable-Position Servomechanism: The core of an extended physiological proprioceptive system. The input and output joints are physically linked to each other so that one cannot lag or lead the other.

Work Space: The volume surrounding an amputee that can be reached with the aid of prostheses.

Works Cited

American Academy of Orthopaedic Surgeons (AAOS) (Eds.) (1960), *Orthopaedic Appliances Atlas: Artificial Limbs,* Vol. 2, Ann Arbor, MI: J. W. Edwards, pp. 1–22.

American Academy of Orthopaedic Surgeons (AAOS) (Eds.) (1981), *Atlas of Limb Prosthetics, Surgical and Prosthetic Principles,* St. Louis, MO: C. V. Mosby.

Atkins, D. J., Meier, R. H. (Eds.) (1989), *Comprehensive Management of the Upper-Limb Amputee,* New York: Springer-Verlag.

Barry, D. T., Cole, N. M. (1990), *IEEE Trans. Biomed. Eng.* **BME-37,** 525–531.

Beasley, R. W. (1966), *Inter-Clinic Information Bulletin (ICIB),* New York: Committee on Prosthetics Research and Development, **V** (8), pp. 6–8.

Beattie, D., Iberall, T., Sukhatme, G. S., Bekey, G. A. (1994), in: D. C. Bacon, T. Rahmin, W. S. Harwin (Eds.), *Proceedings of the Fourth International Conference on Rehabilitation Robotics (ICORR), Wilmington, Delaware,* Wilmington, DE: Applied Science & Engineering Laboratories, University of Delaware, A. I. duPont Institute, pp. 67–72.

Bernstein, N. (1967), *The Co-ordination and Regulation of Movements,* London: Pergamon.

Brånemark, P.-I. (1985), *Tissue-Integrated Prostheses-Osseointegration in Clinical Dentistry,* Chicago: Quintessence.

Childress, D. S. (1973), in: M. M. Gavrilovic, A. B. Wilson (Eds.), *Advances in External Control of Human Extremities, Proceedings of the Fourth International Symposium on External Control of Human Extremities, Dubrovnik, Yugoslavia,* Belgrade, Yugoslavia: Yugoslav Committee for Electronics and Automation (ETAN).

Childress, D. S. (1992), in: J. H. Bowker, J. W. Michael (Eds.), *Atlas of Limb Prosthetics, Surgical, Prosthetic, and Rehabilitation Principles,* 2nd ed., St. Louis, MO: Mosby-Year Book, Inc., pp. 175–199.

Colgate, J. E. (1989), *IEEE Trans. Robotics Automation* **9,** 374–384.

Edell, D. J. (1986), *IEEE Trans. Biomed. Eng.* **BME-33** (2), 203–214.

Fryer, C. M. (1992), in: J. H. Bowker, J. W. Michael (Eds.), *Atlas of Limb Prosthetics, Surgical, Prosthetic, and Rehabilitation Principles,* 2nd ed., St. Louis, MO: Mosby-Year Book, Inc., pp. 133–151.

Fryer, C. M., Michael, J. W. (1992), in: J. H. Bowker, J. W. Michael (Eds.), *Atlas of Limb Prosthetics, Surgical, Prosthetic, and Rehabilitation Principles,* 2nd ed., St. Louis, MO: Mosby-Year Book, Inc., pp. 107–133.

Hall, C. W., Eppright, R. H., Engen, T., Liotta, D. A. (1967), in: G. E. Schreiner, J. M. Hayes, J. P. Ryan (Eds.), *Transactions of the American Society on Artificial Internal Organs,* Vol. 13, Washington DC: American Society for Artificial Internal Organs (ASAIO), pp. 329–331.

Hall, C. W., Rostoker, W. (1980), *Bull. Prosthet. Res.* **10** (34), 98–100.

Heckathorne, C. W. (1992), in: J. H. Bowker, J. W. Michael (Eds.), *Atlas of Limb Prosthetics, Surgical, Prosthetic, and Rehabilitation Principles,* 2nd ed., St. Louis, MO: Mosby-Year Book, Inc., pp. 151–175.

Herberts, P., Almström, C., Kadefors, R., Lawrence, P. (1973), *Acta Orthopaed. Scand.* **44,** 389–409.

Hogan, N. J. (1982), in: R. M. Campbell (Ed.), *Proceedings of the International Federation of Automatic Control (IFAC) Symposium: Control Aspects of Prosthetics and Orthotics, Ohio,* Oxford, New York: Pergamon, 1983, pp. 155–162.

Jacobsen, S. C., Knutti, D. F., Johnson, R. T., Sears, H. H. (1982), *IEEE Trans. Biomed. Eng.* **BME-29** (4), 249–269.

Jacobsen, S. C., Smith, F. M., Iversen, E. K., Backman, D. K. (1990), in: American Nuclear Society, Remote Systems Technology (Ed.), *Proceedings of the 38th Conference on Remote Systems Technology,* Vol. 2, Hinsdale, IL: American Nuclear Society, pp. 180–185.

Keller, A. D., Taylor, C. L., Zahn, V. (1947), *Studies to Determine the Functional Requirements for Hand and Arm Prostheses,* Los Angeles, CA: Department of Engineering, University of California at Los Angeles.

Klopsteg, P. E., Wilson, P. D. (Eds.) (1954), *Human Limbs and Their Substitutes,* 2nd ed., New York: McGraw-Hill.

Kovacs, G. T. A., Storment, C. W., Hentz, V. R., Rosen, J. M. (1989), in: J. J. Presperin (Ed.), *Proceedings of the RESNA 12th Annual Conference, New Orleans, Louisiana,* Washington DC: Resna, pp. 292–293.

Leifer, L. (1981), *Robotics Age* **3** (May/June), 4–15.

Leifer, L., Roth, B., Kelley, D., Sachs, J., Perkash, I. (1978), in: Institut de Recherche d'Informatique et d'Automatique (IRIA) (Eds.), *Colloques IRIA, Proceedings of the First International Conference on Telemanipulators for the Physically Handicapped, Rocquencourt, France,* Paris: Institut de Recherche d'Informatique et d'Automatique (IRIA), pp. 19–34.

McMahon, T. A. (1984), *Muscles, Reflexes, and Locomotion,* Princeton, NJ: Princeton Univ. Press.

Michael, J. W. (1986), *Clin. Prosthet. Orthot.* **10** (2), 66–77.

Miller, W. T. III (1994), *IEEE Control Syst.* **14** (1), 41–48.

Napier, J. A. (1956), *J. Bone Joint Surg.* **38B,** 902–913.

Parker, P. A., Scott, R. N. (1985), *CRC Crit. Rev. Biomed. Eng.* **13,** 283–310.

Perry, J. (1992), *Gait Analysis: Normal and Pathological Function,* Thorofare, NJ: Slack Inc.

Perry, J., Associates (1981), *Normal and Pathological Gait Syllabus,* Downey, CA: Pathokinesiology Service and Physical Therapy Department, Rancho Los Amigos Hospital.

Preising, B., Hsia, T. C., Mittelstadt, B. (1991), *IEEE Eng. Med. Biol.* **10** (2), 13–22.

Raibert, M. (1986), *Legged Robots That Balance,* Cambridge, MA: MIT Press.

Raibert, M. (Ed.), (1984), Special issue on Walking Robots, *Int. J. Robotics Res.* **3** (2).

Raibert, M. (Ed.), (1990), Special issue on Walking Robots, *Int. J. Robotics Res.* **9** (2).

Razic, D. (1973), in: M. M. Gavrilovic, A. B. Wilson (Eds.), *Advances in External Control of Human Extremities, Proceedings of the Fourth International Symposium on External Control of Human Extremities, Dubrovnik, Yugoslavia,* Belgrade, Yugoslavia: Yugoslav Committee for Electronics and Automation (ETAN), pp. 177–183.

Rose, J., Gamble, J. (1994), *Human Walking,* Baltimore, MD: Williams & Wilkins.

Sauerbruch, F. (1916), *Die Willkürlich Bewegbare Künstliche Hand. Eine Anleitung für Chirurgen und Techniker,* 1st ed., Berlin: Springer-Verlag.

Seamone, W., Schmeisser, G. (1972), *Bull. Prosthet. Res.* (10-17), 33–37.

Simpson, D. C. (1969), *Biomed. Eng.* **4** (3), 106–110, 119.

Simpson, D. C. (1974), in: P. Herberts, R. Kadefors, R. I. Magnusson, I. Petersén (Eds.), *The Control of Upper-Extremity Prostheses and Orthoses, Proc. Conference on the Control of Upper-Extremity Prostheses and Orthoses, Göteborg, Sweden,* Springfield, IL: Charles C. Thomas, pp. 146–150.

Thring, M. W. (1983), *Robots and Telechirs: Manipulators with Memory; Remote Manipulators; Machine Limbs for the Handicapped,* Chichester, England: Ellis Horwood Ltd.

Todd, D. J. (1985), *Walking Machines: An Introduction to Legged Robots,* New York: Chapman & Hall.

Topping, M. (1995), *Rehabilitation Robotics Newsletter,* Wilmington, DE: The Rehabilitation Robotics Research Program, The Applied Science and Engineering Laboratories, Alfred I. duPont Institute, University of Delaware.

PROSTHETIC MATERIALS

Jonathan Black, *IMN Biomaterials, King of Prussia, Pennsylvania, U.S.A.*

	Introduction	141
1.	**Biomaterials**	142
1.1	Historical Perspective	142
1.2	Definitions and Fundamental Concepts	142
1.3	The Biological Environment	143
2.	**Host Response to Implanted Materials**	144
2.1	Inflammatory Response	144
2.2	Coagulation and Hemolysis	145
2.3	Adaptation	146
2.4	Immune Response	147
2.5	Carcinogenesis	148
2.6	Metabolic Effects	149
3.	**Material Response to Implantation**	150
3.1	Swelling and Leaching	150
3.2	Corrosion and Dissolution	151
3.3	Bulk Mechanical Behavior	152
3.4	Friction, Wear, and Lubrication	153
4.	**Biomaterial Applications: Prostheses and Artificial Organs**	154
4.1	Orthopedic	154
4.2	Cardiovascular	155
4.3	Dental and Maxillofacial	157
4.4	Plastic and Reconstructive	158
4.5	Miscellaneous Applications	158
4.6	Hybrid Devices	159
5.	**Conclusions**	159
5.1	Clinical Status	159
5.2	The Challenge of New Applications	160
	Glossary	160
	Works Cited	162
	Further Reading	162

INTRODUCTION

Prosthetic materials represent a subclass of biomaterials: those materials used in the fabrication and manufacture of medical and surgical devices and implants. A very wide range of materials has been used experimentally in such applications but few have been successful: Degradation of materials by harsh biological environments and adverse biological response both to materials and to their degradation products have limited successes to a remarkably small number of biomaterials. Applications for prosthetic materials vary from surgical instruments, superficial dressings, and fluid-administration equipment to indwelling but not implanted devices such as contraceptive barriers and dental restorations and to temporary and chronic implants, such as breast-augmentation devices, heart valves, and total hip and knee replacements (Fig. 1). Implantation is normally assumed to require professional (medical/surgical) intervention, to result in breaching of one or more epithelial layers, and to leave in place materials and/or components after the initial procedure. The overall performance requirements are similar in each application, modified by device design and functional requirements, tissue type contacted, device size relative to the patient, and duration of exposure. However, this review focuses primarily on those materials used for chronic (>30 days) implant (and some dental restorative) applications. The majority of these biomaterials have previously proven themselves in other demanding environments, such as in marine and aerospace applications, and have been selected and adapted for use as biomaterials.

Despite the limited materials choices available to designers, large numbers of implants are now being used in orthopedic, cardiovascular, and other applications, providing significant improvements in life quality and duration for perhaps as many as 18 million patients in the United States alone by 1994 (Moss, 1991). Serious problems, involving both materials deterioration and adverse biological response, have arisen in plastic and reconstructive applications, revealing the

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

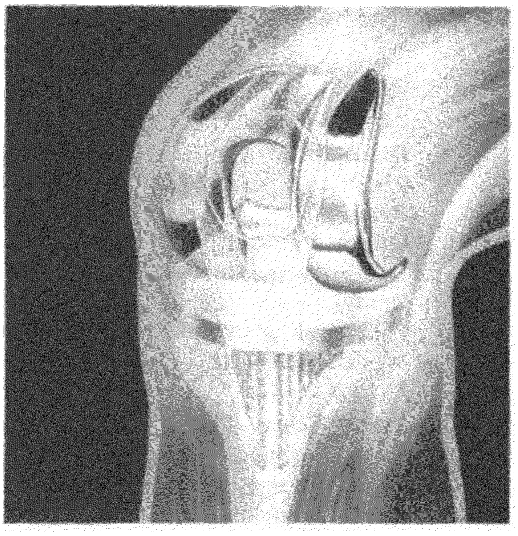

FIG. 1. Modern surgical implants, such as this posterior stabilized total knee replacement, reflect a highly sophisticated combination of biomechanical analysis and biomaterial selection. (Courtesy, Osteonics Corporation, Allendale, NJ, Series 7000 type.)

need both for better understanding of biomaterials–biological systems interactions and for the development of improved biomaterials. Increasing success in implant applications is demanding qualification of biomaterials for service lives exceeding ten years and development of new materials to meet newer applications. More recently, biomaterials are beginning to be designed and fabricated *de novo,* and composite structures with organic-derived materials or living cells are in early stages of evaluation.

1. BIOMATERIALS

1.1 Historical Perspective

The dream of replacement of diseased or lost portions of the human body is probably older than the practice of medicine itself. With the advent of the organized study of medicine and of surgery, early attempts were made using tissues and body parts from other humans and from animals. These led almost universally to failure due to infection and what is now recognized as adverse immune responses, leading to rejection.

With the development of aseptic surgical techniques by Lister in the 1860s, the search for nonreactive biomaterials began. Early applications were limited to relatively noncorrodable pure metals, such as gold, platinum, and silver. The last material was successful as a result not so much of its inertness but of the relative tissue tolerance for its corrosion products, which have also proved to be bactericidal. Catgut and other natural materials were used as sutures; here tissue reaction plays a useful part in strengthening the wound scar by promoting modest inflammation and collagen elaboration. Real progress in the successful clinical use of biomaterials began to be made in the 1940s, with the development of sulfonamide and penicillin to treat local and systemic infections and the adaptation of more corrosion-resistant alloys, especially cobalt-based superalloys, as biomaterials. World War II experience in the internal stabilization of fractures with adjunctive use of blood transfusions provided dramatic reduction in both limb loss and mortality, proving the worth of surgical implants.

The postwar era saw a dramatic expansion in the availability of commercial synthetic polymers, many of which have subsequently proven useful as biomaterials. Perhaps the real watershed events, heralding the widespread modern use of chronic implants, occurred in the late 1950s when Furman and Robinson demonstrated the first successful direct electrical "pacing" of the heart and Charnley evolved the modern polymethylmethacrylate-cemented metal-on-polymer total hip joint replacement.

1.2 Definitions and Fundamental Concepts (Williams, 1987)

A *biomaterial* is most generally a nonviable material used in a medical device, intended to interact with biological systems. This interaction may be characterized by examination of the *host response,* the reaction of a living system to a biomaterial, and of the *material response,* the reaction of a biomaterial to a living system. While biomaterials may have a pharmacological effect or contain an active agent as a component, the term "biomaterial" is reserved to those substances that remain, at least for a period after implantation, in their original form and obtain their *primary* or intended effect through physical rather than pharmacological action. In the past, it was the practice to refer to successful biomaterials as being "biocompatible," suggesting that they generally displayed low or absent host or material responses, but more recently a *biocompatible material* is defined relative to a *specific* application as having acceptable host and material response in that setting. Chemical composition, materials form, processing, cleaning, sterilization, size relative to the host, site of implantation, host disease state, and expected service life as well as nature and amount of degradation products all affect biocompatibility. Thus, a biomaterial that is highly successful in one application may well fail in a different, although related, situation. Since biocompatibility is now viewed as a relative rather than an absolute attribute of candidate biomaterials, great emphasis must be placed upon application-specific materials qualification through the use of well-defined or standard test procedures, both *in vitro* and *in vivo*, and evaluation in relation to both positive and negative control materials, when they are available. Furthermore, since biological performance is so application dependent, regulatory practice tends to focus on the resulting devices and their human clinical performance rather than upon the biomaterials from which they are fabricated.

Biomaterials may be divided into four performance classes by host response; the search for materials in these classes also mirrors the historical development of the field of biomaterials:

1. *inert* materials, which elicit little or no host response;
2. *interactive* materials, which are designed to elicit or promote a specific host response, such as hard-tissue adhesion;
3. *viable* materials, possibly incorporating host tissue and/or live cells, which are capable of being actively remodeled and/or resorbed;
4. *replant* materials, consisting of live native cells or tissue, cultured *in vitro* from cells obtained previously from specific patients or donors.

Both viable and replant materials are perhaps more properly termed *tissues* but are expanding the definition of biomaterials because of their fabrication, before clinical use, through *in vitro* (bio)engineering intervention, termed *tissue engineering*. Thus some authors distinguish between manufactured or man-made biomaterials on one hand, and natural or native biomaterials, including live cells and tissues, on the other hand.

1.3 The Biological Environment

As is the case for all materials in engineering applications, the selection and development of implantable materials are strongly affected by the intended service environment. Conditions within the human body present a high-demand, largely invariant environment. This environment is remarkable aggressive, combining high chemical activity, similar to that encountered in tropical marine applications, but enhanced by the presence of free radicals and active, energy-using enzyme systems, with a highly variable spectrum of multidirectional mechanical stresses. However, *homeostasis*, the maintenance of conditions necessary for mammalian life, results in a specific local environment of any implant that is extraordinarily constant in both physical conditions and composition. Complex, multiple control systems exist to maintain this constancy for the sake of the health of the local cells and organs; thus, deviations from established conditions caused by the presence of foreign materials, such as implants, may be expected to incite restoring responses. It is these restoring responses and their adverse effects on both the materials and the host itself that frequently limit the usability of a material as a biomaterial. For this reason, it is common to speak of prosthetic applications as being *materials limited:* That is, the ability to design and contrive implants exceeds the biological performance of materials available for their construction.

Although it is usual to speak of a single biological environment, there are regional, age, and disease-state variations in conditions experienced by implanted biomaterials. Furthermore, specific cell types are capable of altering pH and $p\text{O}_2$ significantly in their immediate or *pericellular* environment. Finally, different patients may impose different requirements upon the same implant as a result of differences in individual expectations, activity level, and life setting, and evidence is emerging for patient-specific differences in host response to biomaterials. Investigation of the utility of a biomaterial for a specific implant application requires determination of the lifetime performance requirements for that implant, termed the *implant life history.* Stainless steel has been used in temporary sutures, in fracture stabilization, and in structural elements of heart-valve replacements. Brief consideration of these three applications suggests large qualitative and quantitative differences in expected implant life history, leading to an understanding of the more generally satisfactory performance of stainless steels in the first two applications as compared with the last one.

2. HOST RESPONSE TO IMPLANTED MATERIALS

2.1 Inflammatory Response

The response to implanted biomaterials is the normal wound-healing response, specific to the tissues involved in the implant site, modulated by the degree of trauma involved but complicated by the long-term presence of a foreign body, the implant. These responses take various forms, the most general of which, always present in formed (solid) tissues, is termed the *inflammatory* response. This is a cell-mediated response directed toward stabilizing the injured area, restoring physiological status *quo ante,* removing dead or damaged tissue elements and foreign material, and correcting the structural and functional loss due to the initial insult.

The four classical signs of inflammation are redness (*rubor*), swelling (*tumor*), pain (*dolor*), and heat (*calor*). The initial biological event is a rapid dilation of local capillaries and an increased "stickiness" of their endothelial cell linings. This attracts and permits invasion of a succession of cell types, some from adjacent tissues but most from circulating blood. First to arrive are *neutrophils,* small end-state cells that are capable of engulfing (*phagocytozing*) particles of up to 5–7 μm in maximum dimension, with maximum stimulation apparently occurring for particles in the 0.1–1.0-μm range. The primary role of neutrophils is to ingest tissue fragments, foreign particles, and bacteria, which they degrade and digest by release of stored enzymes. Neutrophils are a part of the leukocyte lineage of blood cells, and they are accompanied and succeeded by *eosinophils,* another end-state phagocytic cell type, and by *monocytes,* a larger, more versatile phagocyte. Monocytes circulate in blood and are also resident in tissue, in the form of *histiocytes.* When attracted to injury sites, monocytes undergo morphological changes associated with cellular activation and are termed *macrophages.* They have metabolic capabilities and, more importantly, as a consequence of interaction with both native and foreign materials, can synthesize and release a wide range of cellular mediators, collectively termed *cytokines,* which serve as chemical messengers to a wide range of cells involved in both nonspecific and specific host responses.

If the local insult persists for prolonged periods, such as is the case in the presence of a solid, somewhat reactive implant, monocytes can fuse to form multinucleated *foreign-body giant cells,* of up to 80–100 μm in size, which are capable of engulfing much larger particles and greater volumes of foreign material than smaller mononuclear phagocytes.

Finally, in the case of healing in the absence of a solid foreign body, the injured area, cleared of dead tissue elements and foreign debris, undergoes remodeling or *resolution.* This is mediated primarily by biosynthetic cells, termed *fibroblasts,* and the results are tissue specific: Skin and muscle injuries lead to collageneous union or scarring, cartilage injuries to formation of a provisional or fibrocartilage, while cortical and cancellous bone may partly or completely regenerate. Specific tissues, such as cartilage and bone, may be reconstructed by specialized cells, in this case *chondroblasts* and *osteoblasts,* respectively, leading to regeneration rather than collageneous scarring.

The presence of an implant generally interferes with this succession of processes. Chemical reactivity of the implant may prolong the latter stages and prevent the final, or resolution, stage, producing chronic inflammation. However, if this does not occur, and in the absence of an adverse host response, such as a specific immune response, infection, or neoplastic transformation (*carcinogenesis*), resolution of an implant site will proceed to one of four end points:

1. *Extrusion:* Implants in contact with epi-

thelial tissue (skin and the lining of natural internal body cavities) will be surrounded by a downgrowing extension of such tissue, directed toward extruding the implant from the body. This is termed *marsupialization,* because of the resemblance of the tissue to a kangaroo's pouch.

2. *Resorption:* If the implant is degradable and resorbable, the tissue site may condense to a collapsed scar or, in the case of bone, completely remodel to normal (regenerated) tissue.
3. *Integration:* For a very limited number of materials, such as "bioactive" glasses of selected compositions and some metals, such as pure titanium, direct "bonding" or apparent adhesion to normal tissue may take place.
4. *Incapsulization:* In the most general case, the implant is surrounded and walled off from normal tissue by a collageneous, relatively acellular tissue termed *capsule,* which much resembles scar tissue. The thickness of this tissue, and its cellular content, predominantly histiocytes, reflect the chemical reactivity of the implant, its ability to move relative to the surrounding tissue, and features of its surface geometry. In a bony location, the capsule may be mineralized and is called a *sequestrum.*

If the implant site becomes infected, either because of previous infection, by bacteria carried in during surgery, or by bacteria seeded through the vascular system from a distant postoperative infection site, the presence of an implant interferes adversely with host defense mechanisms. The geometry of the implant–tissue interface plays a role, as may local suppression of macrophage function by corrosion and dissolution products. Implant-site infection frequently is accompanied by a local acidosis, which can increase the corrosion rate of some metallic implants. Finally, many bacteria are capable of forming a protective enveloping film (*glycocalyx*) on the surface of implants, and, as a result, implant-site infection usually requires surgical removal of the implant as an element of treatment.

Particulate debris, introduced during surgery or released by corrosion or wear processes, also has potentially adverse local and remote effects. If the particles are such that cellular degradation is limited or impossible, such as for many implanted materials, then cytokines released by phagocytes during attempted digestion produce potent stimulation of many other types of cells. This may induce local excessive fibrosis (scar/capsule elaboration) and may provoke destruction of normal tissue, such as seen in cellularly mediated bone loss (*osteolysis* or *small-particle disease*) secondary to wear-debris production by total joint replacement prostheses.

2.2 Coagulation and Hemolysis

When foreign materials contact blood, coagulation occurs. In the intact system, this process, leading to the formation of a fibrin clot, termed *thrombosis,* and subsequent tissue repair, acts to minimize blood-fluid and cellular-element loss. It is relatively nonspecific and is triggered by a wide range of physical disturbances to blood and the walls of the vascular system. Coagulation progresses through two biological cascades, with a final common pathway.

The first, or intrinsic, pathway is triggered by activation of a protein, *factor XII,* also called the *Hageman* factor, by contact with a foreign surface (other than the healthy, intact, endothelial-cell lining of vascular processes). Simple physicochemical arguments suggest that any implanted biomaterial rapidly becomes coated with proteins and other biological molecules present in tissue fluids and blood. Such films are not stable but, like adsorbed layers on surfaces in gaseous milieus, experience a progression of molecular species and concentrations. This succession, termed the *Vroman effect,* after its discoverer, in the case of coagulation involves a succession of molecules including *kininogen, prekallikrein, fibrinogen,* and *albumin.* Many of these molecules, such as prekallikrein and factor XII, on surface contact, are altered in structure and molecular weight by the forces involved in surface adsorption and thus become new or activated species.

The other, or extrinsic, pathway is triggered by release from damaged cells of *factor III,* also called *tissue thromboplastin,* and its activation by surface contact. Both pathways lead to activation of *factor V,* release of *prothrombin* (and its activation) from *platelets,* small circulating blood elements, and to the eventual activation of *factor XIII* leading to the formation of a fibrin clot.

Clots incorporate living and dead cells in a fibrous fibrin network. They may be adherent, in which case they cause local disruption of blood flow, or nonadherent, in which case they are shed and trapped in downstream capillary beds, producing damage in remote organs such as lung, kidney, and brain. There have been many approaches to minimizing thrombosis by implant materials, but most focus on presenting smooth, "slippery" surfaces with surface energies, as measured by *critical surface tension* (γ_c), near 25 dyn cm, to circulating blood. Prevention of thrombosis, particularly in blood vessels smaller than 6 mm in internal diameter, remains one of the key problems in biomaterials.

Hemolysis is the release of hemoglobin, the primary oxygen-carrying complex, from red blood cells (*erythrocytes*) primarily as a result of mechanical damage to their cell walls. It is desirable to avoid hemolysis both because of the intrinsic toxicity of hemoglobin and because of the loss of oxygen-carrying capacity due to cell loss. Both the roughness and the surface energy of biomaterial surfaces play small but significant roles in producing hemolysis. However, the dominant mechanism is mechanical, either by adverse hemodynamics produced by blood flow through implants or external blood-processing systems such as blood oxygenators, since shear stress above 1500–3000 dyne/cm^2 can cause direct noncontact cell lysis, or through physical impact, such as in heart-valve closure.

2.3 Adaptation

Mammalian tissue grows and remodels by cellular replication (*mitosis*), increase in individual cell size (*hypertrophy*), and cellular differentiation and maturation from stem-cell populations (*regeneration*). These processes, which direct normal development and restoration of the effects of physical damage and disease, are affected by the presence of implants. The abilities of tissues to adapt themselves to local requirements, including reaction to the chemical, physical, or electrical properties of implants, are collectively termed *adaptation.*

A naturally occurring example of adaptation is the progression of tissue types seen in the healing of bone fractures. This involves four stages of development. An initial *hematoma* stage rapidly progresses to the *soft callus* stage, in which a mass of mineralized collageneous tissue (*callus*) bridges the fracture plane and restores bone stiffness. This soft callus adaptively remodels to a more organized *hard callus* and then to restoration, by *remodeling,* of the original form of the bone before fracture. These last two stages are apparently driven by physical signals resulting from mechanical stress applied to the bone and associated tissues and are an expression of Wolff's law, usually understood as "The form being given, tissue adapts to best fulfill its mechanical function" (Wolff, 1892).

If a fractured bone is stabilized by implants, the adaptive response includes reduction in the volume and stiffness of the hard callus as well as, in the long term, decreased density of bone, as a consequence of load sharing between the implant and the hard tissue. This latter effect is termed *stress shielding* and also occurs in association with medullary (central) fixation elements of partial and total joint replacement devices. In the case of replacement of the femoral head with a metallic component with a intramedullary fixation element, rarefication (*osteoporosis*) and resorption occur in the proximal cortical bone, especially on the medial side, where mathematical analyses predict the greatest reduction in peak stress (from the stresses encountered in the intact bone) while overgrowth (*hypertrophy*) occurs at the distal end of the device, where elevated stress concentrations are expected. Reduction in the relative stiffness of the implant element coupled to bone, by replacing the metal alloy used by one with a lower elastic modulus or by reducing the average areal section moment, reduces the magnitude of these effects in experimental animal models.

Hard tissue also shows other adaptive responses. Healing bony tissue will grow onto implant-surface features of almost any size and into interconnected pores of at least 100 μm in diameter. Nonloaded implants or ones located in regions in which bony tissue is not normally found, such as medullary spaces, display later resorption of such ingrown tissue.

Soft tissue also shows adaptive effects, primarily related to use or disuse. Although there are general claims made for a soft-tissue adaptive law similar to Wolff's law, the evidence is much less clear than in the case of hard tissues. While excitable soft tissue (muscle) clearly adapts to use, there has been no widespread success in mediating soft-tissue response to implants. Perhaps the most pressing problem has been the desire to direct the formation of a new endothelial-cell layer on the inner wall of blood-conducting devices, such as arterial grafts. Despite many claims to the contrary, the primary response still seen is a firm fibrin clot, with occasional islands of endothelial cells (a *pseudointima*), rather than the desired mature, continuous endothelial-cell lining or *neointima*. Similarly, attempts to regenerate tendons and ligaments by providing a disintegrating or resorbable scaffold most generally still lead to elaboration of scar tissue, with some functional reorganization, rather than to formation of an appropriate vascularized, enervated tissue replacement.

Finally, there is a growing recognition that electrical signals may play various roles in the natural organization and maintenance of both hard and nonexcitable soft tissues. Research and clinical application have shown that faradic, capacitive, and more complex electromagnetic signals at field strengths that produce tissue potential gradients of 0.1–1 mV/cm are able to produce a range of stimulatory effects, although the biological mechanisms are as yet unclear. The former effect, of stimulation of growth by a negatively charged (cathodic) electrode, represents a special case of an implanted biomaterial. The primary clinical application has been in the treatment of recalcitrant *nonunions,* i.e., fractures of bone that will not unite without further treatment.

2.4 Immune Response

In addition to the nonspecific responses to foreign materials described in Secs. 2.1 and 2.2, specific composition- and/or structure-mediated responses, termed collectively the *immune* reaction, are produced by a system of agents and cells that reside in all mammalian tissues. This system is based upon the production of specific molecules (*antibodies*), which recognize and combine with foreign materials, or *antigens.* With the exception of introduced proteins, as in the case of implants fabricated from insufficiently processed human or animal tissues, the local biological reaction to biomaterials is actually mediated by reaction to adsorbed native proteins whose structure and/or composition may be altered by chemical or physical reaction.

The common recognized form of immune response is a *type IV delayed hypersensitivity.* This is produced by *antigen* reaction between antibodies bound to the surface of a leukocyte subpopulation, the *T lymphocytes,* and has local clinical hallmarks of chronic inflammation, except that both *lymphocytes* and *neutrophils* may persist in the implant site for very long periods of time.

Systemically, remote-site *dermatitis* (skin irritation), secondary to release of free antibody, is a recognized complication of biomaterial implantation. A large but controversial clinical and experimental literature exists that suggests the possibility of exacerbation or possibly primary causation of a number of immune-system disorders, including *asthma, scleroderma, myositis,* and *lupus erythematosis.*

Metals, particularly chromium, cobalt, and nickel, can produce such immune responses, with the antigen probably being a *hapten* or complex formed between corrosion products and soluble proteins. Immune sensitivity to metals, primarily due to workplace exposure, is a well-recognized clinical phenomenon with perhaps 10% of adult populations in developed countries sensitive to at least one of these three metals. Activation of a prior sensitivity and acquisition of new sensitivities have been demonstrated as consequences of chronic implantation of metallic biomaterials. Despite the existence of a number of convincing case reports, the quantitative contribution of specific immune response to the causality of more widely recognized adverse clinical symptoms, such as pain, local swelling, and loosening of devices anchored to bone, has not been clearly established. A long-term concern is the ability of these haptens to suppress the chemotactic abilities of neutrophils in sensitive individuals, thus possibly interfering with resistance to bacterial infection in implant sites.

Some monomers, especially methacrylate

monomer, released from polymers, have been shown experimentally to cause human immune sensitivity, especially after repeated contact, as experienced by dental laboratory technicians, but demonstration of clinical symptoms related to implantation of such materials is still unclear. There is current significant clinical concern about immune response to poly(dimethyl)siloxane (silicone elastomer), especially in the form of gel released from breast-augmentation implants. Although immune-system disorders are reported by some patients and several investigators have apparently detected specific antibodies to finely divided silica (silicon oxide) and to silicone elastomer, the causal relationship between the presence of such antibodies and clinical symptoms is unproven, as is the general case for all such responses.

Human immune responses to either permanent or resorbable implanted ceramics have not been reported.

2.5 Carcinogenesis

Cancer (*neoplasia*) in mammals is recognized as uncontrolled multiplication and spread within the individual of abnormal forms of host cells. The key features of such growth are seemingly uncontrolled cellular replication, autonomy of growth, and intrusion into other tissues and organs. Such neoplastic growth is the direct consequence of inheritable changes in cellular genetic code.

Cancer is a multiphase process that can be initiated and/or promoted by a variety of conditions and agents. The two primary mechanisms of concern related to the use of implanted biomaterials are *chemical carcinogenesis* and *foreign-body carcinogenesis.*

There is a widespread societal concern about chemical carcinogenesis. Alarm concerning dioxin contamination, due to industrial accidents as in Serveso, Italy, or deliberate dispersion, as in Times Beach, Missouri, in the U.S., and concerning Alar™ pesticide residues on fruit has tended to obscure some of the realities of our 20th-century experience with cancer.

Despite the common perception that cancer is increasing, overall age-adjusted cancer death rates, with the exception of death due to lung cancer and, perhaps, brain cancers, have remained constant or, in the case of cancers of the liver, the stomach, and, in women, the uterus, have declined steadily over the past 50 years. However, increasing life expectancy, especially in the older, more cancer-prone segments of the population, and reductions in age-adjusted death rates due to heart attack and stroke have increased the relative incidence of death due to cancer. This combined with the continuing increase in deaths of men due to lung cancer, believed largely related to increases in tobacco smoking since WWI, and a parallel incidence increase for women since the 1960s have further heightened the impression of overall increase.

Similarly, repeated "scares," such as previously noted, have conveyed the impression that "everything" causes cancer. Insofar as chemical agents are concerned, the opposite is the case: While there are probably more than 1 500 000 identified chemical compounds, only a few thousand of these are suspected carcinogens, based upon observations in animals, and fewer than three dozen can be classified as known—that is, presumptive—carcinogens in humans.

However, despite this relatively rosy outlook, there are continuing concerns about possible chemical carcinogenesis related to the use of chronic implants. Nickel, cobalt, and hexavalent chromium are presumptive human carcinogens, and other agents, including vanadium and some polymeric degradation products, especially those released by polyvinyl halides and some polyurethanes, are suspected human carcinogens.

Despite the long history of the use of many biomaterials, the evidence for implant-mediated chemical carcinogenesis is sparse. Chemical carcinogenesis typically requires a long latent time between exposure and tumor expression: At low doses in humans, this period can easily exceed twenty years, as is the case for asbestos-induced *mesothelioma* (a tumor of the lining of the chest cavity). Many implants, such as orthopedic or cardiovascular devices, are placed in tissues that show low intrinsic rates of neoplastic transformation. However, there are now several apparently reliable reports of implant-site tumors occurring ten years or more after total hip replacement with components fabricated of stainless steel and/or cobalt-based alloys.

Epidemiology of implant patients is sufficiently weak that there is great difficulty in relating release of corrosion or degradation products from one site with possible changes in incidence and/or progression of tumors in other sites. Several studies of total hip joint replacement patients suggest the possibility of an elevated incidence of tumors of the circulatory system, especially lymphomas, which is consistent with the concern for such possible systemic or remote-site causation. Interest in this clinical area remains active and is encouraged by an apparent trend to implant permanent orthopedic devices in younger patients.

Foreign-body carcinogenesis is well recognized in rodents. In many strains of mice and rats, subcutaneous implantation of chemically inert solid materials, of a wide range of compositions, results in neoplastic transformation of the adjacent soft tissues. This phenomenon was first observed in the 1940s and has come to be called the *Oppenheimer effect,* after its discoverers. Concern about the possibility of such an effect in human clinical populations has been widely expressed, especially in relation to implants used in plastic and reconstructive surgery, which are often placed in analogous tissue sites. Extensive retrospective clinical follow-up studies have failed to demonstrate elevated tumor incidence (vs operated but unimplanted soft-tissue sites) for periods of up to fifteen years. While these negative findings are encouraging, they should be viewed with caution, as latency times may be extremely long for such chemically passive implants, and implant-site inflammation, due to chemical or mechanical irritation, tends to suppress such tumor formation in animal models. Furthermore, the histological appearance of the tissue capsule surrounding chronic implants closely resembles the *benign fibrous histiocytoma* associated with tissue volumes in which local blood circulation has failed, such as *infarcts* in bone. Such conditions in nonimplant sites display a low but significant rate of transformation to *malignant* fibrous histiocytoma, a tumor type that is the most frequent of the tumors reported in implant sites. However, it may prove to be the case that the human immune system, which is much more sophisticated than that of rodents, may be able to provide long-term protection from foreign-body carcinogenesis.

2.6 Metabolic Effects

Mammals have a well-developed circulatory system that supplies oxygen and nutrients to cells in all tissues while conveying metabolic and catabolic end products to excretory organs. This system is characterized by relatively high flux rates in the presence of local maintenance of chemical composition and concentrations.

There are two major consequences of this situation that are of importance in the consideration of the effects of implants on metabolic processes. First, tissues are extremely clean or pure, with the elements oxygen, carbon, hydrogen, and nitrogen, in declining order of abundance, making up 96.9 wt % of all tissues. Second, other elements (apart from the so-called *physiological elements:* calcium, phosphorus, potassium, sulfur, sodium, chlorine, and iron), although present typically at only parts per billion to parts per million concentrations, have closely controlled concentrations and play "catalytic" or amplified roles in metabolic processes. For example, chromium, with a total body burden (in a 70-kg individual) of <6.6 mg and a daily dietary intake of ~0.2 mg, is an essential constituent or cofactor in all sugar metabolism, Cobalt, although a key ingredient in the essential vitamin B_{12}, is sufficiently toxic that it is absorbed only in trace amounts as a bacterial preformed compound (*cobalamine*) and thus is the rarest quantified element in the mammalian body, with a typical human body burden of <1.5 mg.

Dietary materials enter the human body through the gastrointestinal (GI) tract, primarily by way of active (energy requiring) absorption processes. Passing through this first control interface, the GI intima, they are taken up and distributed through the circulating plasma pool, involving both arterial blood supply and venous blood and lymphatic fluid return. In a 70-kg individual, this volume contains ~3.2 L of water and is coupled to the GI tract through fairly rapid diffusion and exchange coefficients, which are dependent upon molecular weight, molecular composition, and deviation from normal (control) concentrations of solutes. The plasma pool communicates directly with an interstitial pool (~15.7 L) that consists of all extracellular fluid within the tissues outside of the circulating blood and lymphatic pathways. This interstitial pool is subdivided kinetically into a "fast" compartment (~8.4 L), probably representing ions and molecules in solution, and a "slow" compartment (~7.3 L), probably representing precursor combination and physisorption to cell walls and membranes. The fast compartment, in turn, communicates with cellular function through both active transport and passive diffusion across cell walls. The intracellular space, containing genetic material and in which metabolic and catabolic processes occur, has an effective volume of ~23 L, representing 55% of body water content and 33% of body weight.

Soluble degradation products from implants, whether released by dissolution, leaching, or corrosion, effectively pass initially into the fast interstitial compartment. Thus the act of surgically implanting a biomaterial bypasses at least one and probably more regulatory barriers. Released materials may then enter local cells directly or pass into the plasma pool where transport and remote-site storage compete with excretion through renal excretion and bile synthesis pathways. Clinical studies of patients and analyses of cadavers with metallic implants have shown small but significant serum concentration elevations of metal content (50%–300%) that can be related to 10–100-fold elevations in tissues both surrounding the implant and in remote-storage tissues, especially in the liver, even in the presence of satisfactory clinical performance.

All metallic ions released are rapidly complexed with organic molecules, binding either nonspecifically to albumin or specifically to high–molecular-weight carrier proteins from the immune globulin family. These binding processes usually represent relative amplification of normal physiochemical processes, related to transport, utilization, storage, and excretion. However, differing metallic ionic radii with multiple valence states may result in adverse biological consequences because of ionic species competition. Thus, chromium[III] may compete with iron[III] for binding to *transferrin,* the normal iron carrier molecule, reducing tissue iron availability and inducing relative *anemia,* as shown in animal studies. Similarly, the higher binding avidity of aluminum[III] than of magnesium[II] to many organic ligands may play a role in aluminum's ability to induce osteoporosis and possibly promote Alzheimer's disease.

The issue of release of organic agents from implants is more complex. The presence of lipids and other nonpolar species in body tissues and fluids results in an ability to solubilize, transport, and store a wide range of both polar and nonpolar organic molecules. It is possible for released molecular species to resemble normal enzyme substrates sufficiently that they can enter normal metabolic pathways. In some cases this leads to simple catabolic degradation; however, metabolic intermediates, such as aromatic amines produced by metabolism of some low–molecular-weight polyurethanes, may have unexpected toxicity. There is a poor understanding of the metabolic/catabolic pathways for all but the most simple synthetic organic molecules, leading to a common practice of performing metabolic fate studies, utilizing radiotracer-marked materials in a small-animal model, as part of the biocompatibility testing and qualification of new biomaterials.

Finally, cell-mediated and passive transport of wear and other small debris from implant sites is often observed. Surgical exploration reveals accumulation of macrophage-borne debris in regional lymph nodes near the implant site, while animal and human cadaver studies demonstrate transport of debris to remote organs, including the lungs, liver, and spleen. Asymmetric particles, such as wire fragments, can move passively through tissues to great distances, as a result of muscular contraction and relaxation forces. The biological significance of such findings is still in dispute, as is the general issue of bioavailability of biomaterial degradation products in general: That is, the mere finding of a foreign material within a tissue may not be easily related to the effect that its constituents can have upon intracellular processes.

3. MATERIAL RESPONSE TO IMPLANTATION

3.1 Swelling and Leaching

Implanted materials are placed in an aqueous phase with both polar and nonpolar components. Thus, there is the possibility for exchange of solutes between the implant and the surrounding fluids and tissue elements. This is a consideration in the use of polymers as biomaterials.

Polymers, as fabricated for a variety of industrial applications, usually contain a wide range of additives, including plasticizers, fillers, colorants, and mold-release agents as well as residual catalysts, monomer, and, in the case of chemosets, reaction products. When these polymers are adapted for biomaterial use, there is a general policy to eliminate or at the least minimize such low–molecular-weight products that can be removed by leaching. Unfortunately, in many cases, a residual level of plasticizers and low–molecular-weight polymers is required to produce fabricatability and maintain desirable mechanical properties. Leaching of these materials may cause local inflammation, generally related in magnitude to the interfacial concentration of such species. Loss of low–molecular-weight species from polymers may result in increased tensile modulus, thus stiffening the implants, but also may reduce yield strength.

The ability of polymers to release low–molecular-weight species by diffusion is used in the fabrication of reservoir systems for release of antibiotics and other treatment agents. A common example is the use of antibiotic-impregnated PMMA during total joint replacement after removal of a device from an infected site. The combination of a high–diffusion-rate reservoir with a low–diffusion-rate "metering" membrane can provide relatively constant rates of release over periods of months. This design approach is used in the design of implantable fertility-control systems.

Polymers can also absorb water and, in some cases, lipids and other nonpolar solutes. In addition to generally decreasing tensile modulus and raising yield stress, absorption may reflect a lack of osmotic equilibrium with the local biological environment. Under this condition, absorption may generate sufficient internal pressure to cause static tensile failure, seen as cracking and crazing.

3.2 Corrosion and Dissolution

All metallic materials will release elemental metals as a consequence of implantation. Alloys intended to serve as biomaterials are selected both for their low intrinsic or uniform corrosion rate and for the biological acceptability of the released corrosion products. Thus, the principal alloy systems in use are ones that are either electrochemically noble (gold or platinum based) or are able to be rendered highly passivated by preimplantation chemical formation of surface oxides/hydroxides. The primary alloy systems used are thus low-carbon stainless steels and cobalt-based superalloys (both rendered passive by chromium–oxygen reaction layers) and titanium-based alloys, rendered passive by formation of TiO_2–Ti_2O_5 surface layers.

Two points are important to take note of with regard to uniform corrosion:

1. The traditional engineering definition of absence of corrosion (total dissolved metal concentration at equilibrium $\leq 10^{-6} M$) is an inadequate selector, since this threshold concentration corresponds to 50–60 ppb for metallic elements, such as cobalt and chromium, that have confirmed biological activities at two orders of magnitude lower concentration.
2. Protective nonconductive coatings, such as paint, or layers of sacrificial elements, such as zinc galvanize, generally cannot be incorporated into biomaterials because of the generally adverse host response to such materials and to their degradation products.

Initial *in vivo* uniform corrosion rates are generally predictable by suitable *in vitro* measurements, if performed in physiologic (0.85%) saline solution at values of pH and pO_2 appropriate to the intended application. Most useful biomaterial alloys, when evaluated at near neutral pH, after adequate cleaning and passivation, display uniform corrosion rates $\approx 10^{-8}$–10^{-7} g/cm^2 day (Steinemann, 1980), with titanium-based and noble alloys somewhat more corrosion resistant than passivated chromium-containing alloys. However, the presence of monatomic free oxygen radicals, hydrogen peroxide intermediates, and active specific and nonspecific protein ligands renders such determinations only approximate. Furthermore, in the absence of surface damage or relative movement between implant and adjacent tissue, the maturation of adsorbed protein layers and continued growth of passivating layers may reduce uniform corrosion rates with increasing time of implantation.

Unfortunately, the fabrication of biomaterials into functional implants produces a number of other corrosion effects that may produce adverse material or host response, including the following.

The high ionic conductivity of interstitial and other body fluids permits the possibility of galvanic corrosion when two metallic components of an implant have different electrochemical potentials as a result of differences in composition or processing. It is generally held that cobalt-based superalloys, titanium-based alloys, and bulk carbon (either pyrolitic or vitreous) can be coupled in this fashion without risk, but none of these materials are safely coupled to stainless steels.

Crevice corrosion is also a problem encountered when multipart devices admit body fluids into narrow, deep interpart spaces, such as the screwhead–plate interface in fracture fixation devices. Pitting, intergranular, and leaching corrosion may also take place but are usually related to inappropriate processing. Relative motion between passivated parts may lead to fretting corrosion as an apparently accelerated form of wear. In this respect, there is significant evidence that *in vitro* studies, in the absence of serum proteins, tend to overestimate the fretting rate for any particular configuration of components.

Detailed mechanistic study of *in vivo* corrosion is still at an early stage. However, evidence has been found in model systems that organic ligands can bind and remove metal without charge separation, that surface protein layers on metals turn over, producing continuing fresh binding, and that adsorption and diffusion of amino acids can alter corrosion rates, apparently secondarily to changes in the electrical conductivity of passivation layers. Finally, there is some evidence for direct biological corrosion, reflecting attempts by phagocytic cells, especially monocytes, to dissolve implants or, as previously stated, a strong acidic shift in *p*H in the presence of infection.

Both soluble (absorbable) polymers and ceramics, especially in the calcium phosphate systems, undergo passive dissolution, primarily controlled by local *p*H and concentration of dissolved species. At least two modes of *in vivo* dissolution are recognized: surface recession and internal dissolution. The latter, more common for hydrophilic materials, presents a possible hazard of fine particulate release, which may produce local adverse cellular response (see Sec. 2.1).

3.3 Bulk Mechanical Behavior

Mechanical function of biomaterials in implant applications is dominated by three fundamental intrinsic materials properties: elastic modulus, yield strength, and fracture (or ultimate) strength. These properties, when expressed in the design of an implant, affect respectively the ability to deform reversibly, the maximum load without irreversible deformation, and the minimum load to produce fragmentation. In specific applications, other mechanical properties such as hardness, toughness, ductility, creep rate, and fatigue strength may become important.

The elastic modulus, whether in tension/compression or in shear, is primarily determined by the nature of atomic bonding within materials and thus is little affected by exposure to the biological environment. The elastic moduli of metals, ceramics, and composites are very little affected by implantation. However, polymers may be affected by leaching and absorption (see Sec. 3.1) and by chemical processes that alter the pattern of covalent bonds within their bulk. In general, chain scission processes, which lower average molecular weight, tend to reduce elastic moduli while cross-linking tends to raise elastic moduli, particularly for polymers with large side groups. Implantation may also result in oxidation and a concomitant increase in moduli.

As is the case for elastic moduli, only polymers undergo real changes in yield strength upon implantation. The most usual reaction is an increase in yield strength, with a concomitant reduction in ductility and toughness. Thus many polymers become brittle after prolonged implantation.

Ultimate strength is dependent upon both bulk structure and surface properties. Metallic, polymeric, and ceramic materials, as fabricated into biomaterials, are all polycrystalline to a degree, thus possessing grain or domain boundaries. Corrosion, dissolution, and absorption phenomena all selectively affect these boundaries, reducing the strength of the material. Furthermore, most fracture processes depend upon propagation of defects to produce a fracture plane. Therefore, local surface or grain-boundary attack increases the number and size of such defects, reducing strength of the material.

Finally, corrosion processes, by promoting dissolution of passivation layers, generally act to reduce fatigue strength in low-stress, high-cycle applications. This phenomenon is especially seen in stainless steels but may also occur in cobalt-based alloy systems. Titanium-based alloy systems with their higher oxide stabilities generally do not show this effect, although high local chloride or sulfide concentrations may have a deleterious effect.

The effects of implantation on composites depend upon two factors: the effects on the phases of the composite and on the phase interfaces. If composite properties depend upon maintenance of particular interfacial property values, then *in vivo* degradation becomes extremely difficult to predict.

It is worth noting here that *in vitro* simulation of *in vivo* degradation of bulk properties is generally unsuccessful. While *in vitro* conditions designed to reproduce specific biological environments appear to produce lower average degradation than later encountered in *in vivo* exposure, the effects of specific biological processes, associated with enzymatic attack, molecular complexing, and surface calcification, are difficult to predict. Thus, long-term animal implantation studies are a prerequisite to clinical introduction of new material formulations.

3.4 Friction, Wear, and Lubrication

Tribological properties (friction, wear, and lubrication) (see TRIBOLOGY) become of extreme importance in applications in orthopedics and maxillofacial surgery that involve partial or total replacement of joints. Natural joints are highly developed structures in which the porous resilient articular (*hyaline*) cartilage on each side of the interface provides very low coefficients of friction combined with low wear rates and considerable ability to absorb energy in high-velocity impacts. Synovial tissues within the surrounding soft-tissue capsule of each joint secrete both low–molecular-weight phospholipids, which act as surface lubricants, and high–molecular-weight hyalauronates, which assemble to carrier proteins to produce bottle-brush-shaped, extremely hydrophilic complexes, rendering the resulting joint fluid viscous and maintaining joint surface separation. A final lubricating feature of cartilage is its high water content (<70%), a portion of which is expressed reversibly under load, providing a high-pressure fluid film that aids in reducing both friction coefficients and wear rates.

Joint replacement materials are nonporous and rigid in comparison with cartilage. The most common wear pair in use today is cobalt-based superalloy (either cast or forged) articulating against ultrahigh–molecular-weight polyethylene (UHMWPE). This and other combinations work best when the UHMWPE component is either concave or flat; convex components show a 3–5-fold higher wear rate. Synovial tissues left in the site after surgery reproduce a nearly normal synovial fluid as a lubricant. This combination demonstrates *in vivo* an adequate dynamic coefficient of friction of 0.05–0.10 (~15–20-fold higher than the natural joint) with a linear wear rate (in the hip) of about 0.1 mm/yr.

The wear observed in this material combination is primarily of the UHMWPE, and 70%–90% of the resulting particles are <1 μm in maximum size. Wear rates in knee replacement applications are similar, but fewer, somewhat larger particles are typically produced. The principal wear mechanism appears to be a combination of abrasion, adhesion, and fatigue. *In vitro* simulations in the absence of proteins or hyalauronates produce polymeric transfer films that are periodically shed, resulting in cigar-shaped particles as long as 300 μm; however, these are very rarely seen in recovered human tissues. UHMWPE, because of its softness, is quite prone to plowing damage by the roughness of the counterface. Thus scratching of the metallic head or inclusion of "third bodies" (metal fragments, bone chips, etc.) into the interface can produce dramatically accelerated polymeric wear. As a result, titanium and titanium alloys have proven less successful as articulating counterfaces for UHMWPE. No other polymers have been found so far that routinely display lower *in vivo* wear rates than UHMWPE.

Harder, smoother surfaces, such as alumina and zirconia, appear to produce lower polymeric wear rates when articulated against UHMWPE. So-called "hard-on-hard" articulating pairs, especially metal–metal (cobalt-based) and alumina–alumina, while having somewhat higher coefficients of friction, have wear rates reportedly as low as 2.5–5 μm/yr. However, design problems and concern about biological hazards associated with cobalt-alloy use and about brittle fracture of ceramics have limited the use of these European systems in the U.S. so far primarily to clinical trials.

4. BIOMATERIAL APPLICATIONS: PROSTHESES AND ARTIFICIAL ORGANS

4.1 Orthopedic

Orthopedic implants may be divided into two main categories: those used for treatment of trauma and those used in musculoskeletal reconstruction.

The use of implants in treatment of musculoskeletal trauma is of ancient origin but, as previously noted, gained real momentum during WWII, due to the availability of improved, more corrosion-resistant alloys, sulfonamides and antibiotics to fight infection, and blood transfusions. These advances made it possible both to preserve life of the multiply injured soldier and to perform extensive reconstructive surgery in areas, such as the skull, neck, shoulder, spine, and hip, where extensive blood loss would be expected because of the anatomic inability to apply pressure bandages or tourniquets. German surgeons, especially Küntschner, pointed the way; their advances were eagerly imitated by French, British, and U.S. surgeons with much of the postwar development then occurring in Switzerland, within a surgical-industrial alliance called the AO-Group. The result is that today there is an extensive armamentarium of hardware for treatment of almost any fracture, from massive disassociation fractures of the pelvis to isolated fractures of the small bones in the wrist, hand, and foot. The material of choice for most of this hardware is one or another variant of a low-carbon version of 316 stainless steel, alloys containing typically ~18% chromium, ~14% nickel, and typically less than 0.08% carbon. These alloys combine good corrosion resistance and low relative cost with moderate tensile and fatigue strength yet are sufficiently ductile so as to permit intraoperative adjustment of implant shape. Although modern alloys in this family are quite stable for long periods of implantation, such hardware is still viewed as a temporary implant and it is recommended that it be removed by two years after implantation, if the patient's medical condition and life expectancy warrant a second procedure, because of concerns for stress shielding, secondary hematogeneous infection, and release of corrosion products. Recently, polylactic–polyglycolic copolymers have been used as isolated resorbable pins, screws, and connectors to obviate the need for later removal, but lack of strength and concerns about premature dissolution have limited this development.

The great breakthrough in orthopedic procedures involving biomaterials has been the increasingly successful total replacement of the knee and hip joints. These joints may be damaged by trauma, be attacked by infectious or autoimmune diseases, or simply wear out through overuse or old age. Previous to the 1950s, surgical correction depended upon anatomical realignment (*osteotomy*), various procedures involving interposition of natural or artificial materials, or fusion. The general result was continued disability and reduction of life expectancy, due to loss of cardiovascular capability.

John Charnley, working at Wrightington, U.K., developed a total hip joint replacement (THR) combining the following elements: a polymeric acetabular cup, bedded in the pelvis with cure-in-place chemosetting methyl methacrylate (PMMA) "bone cement," and a metallic femoral component with a small-diameter (originally 22 mm) head, also secured with bone cement (Fig. 2). The combination of the metal–polymer [originally stainless steel–polytetrafluoroethylene (later quickly replaced by UHMWPE, molecular weight ~3–6 million)] articulating interface with the nonadhesive PMMA cure-*in-situ* grouting material to distribute stresses to the underlying bone has proven so successful that such systems now can be expected to show survival [= absence of surgical revision (re-

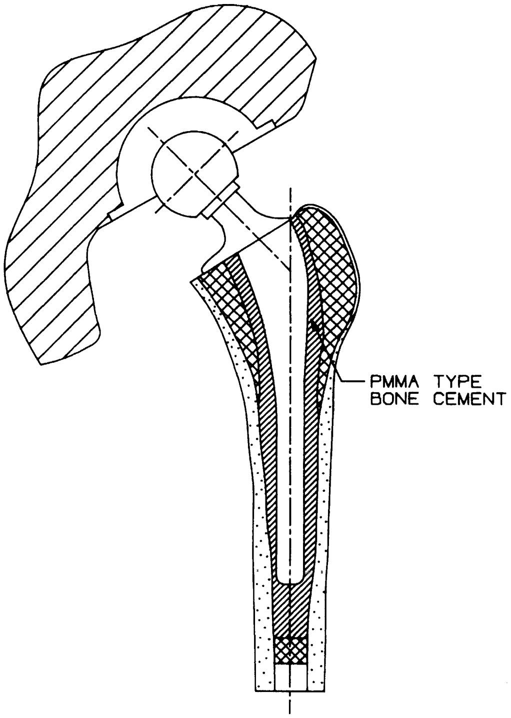

FIG. 2. Generic PMMA-cemented total hip replacement utilizing the design elements introduced by Sir John Charnley. Modern surgical technique involves meticulous preparation of the bony beds, sizing components to provide thin (2–3 mm) uniform cement mantles, and pressurization of cement after insertion. (Courtesy Howmedica Orthopaedic Div., Rutherford, NJ.)

placement) for pain related to loosening] of >95% at ten years after insertion. This has resulted in an estimated U.S. clinical population of perhaps 1.5 million THR patients, with an annual operative rate probably exceeding 120 000.

Modern developments include the substitution of high-strength cobalt-based alloys (typically 27–30 wt % chromium, 5–7 wt % molybdenum, various minor alloying agents, balance cobalt) for stainless steel in the femoral component, the use of modular or multipart components permitting better fitting of the device to the underlying bone, and use of ceramic femoral heads for lower polymer wear, and introduction of alternative fixation systems, frequently applied to lower-modulus titanium-based alloys (such as 6 wt % aluminum, 4 wt % vanadium, balance titanium), utilizing porous metallic or "bioactive" calcium hydroxyapatite surface coatings. Concerns remain for the long-term failure of these devices, used both with and without PMMA cement, secondary to osteolysis mediated by submicron UHMWPE wear debris.

The same technology has also yielded great success in replacement of the knee, with such procedures possibly now exceeding total hip replacement in frequency, and with somewhat less success in replacement of the shoulder, elbow, ankle, and small joints in the hand and wrist. The latter applications are dominated by relatively non-load-bearing alignment or spacing devices, fabricated primarily from silicone elastomer with various fabric reinforcing elements. Polymeric materials are also used for replacement of ligaments but with quite poor long-term clinical success.

Ceramics are little used in orthopedic applications because of their inherent susceptibility to brittle fracture under tensile or shear loading. Alumina and, more recently, zirconia have been used as femoral heads in THR, particularly in Europe and Japan. Various resorbable or nonresorbable ceramic materials, primarily calcium phosphates such as calcium hydroxyapatite and tricalcium phosphate, have been used to augment or replace bone grafts.

4.2 Cardiovascular

The use of implanted biomaterials in treatment of cardiovascular disease has centered on three applications: repair of the valves and body of the heart, partial or total replacement of the heart, and replacement of portions of central and peripheral arteries and veins. All three applications have been hindered by a continuing inability to provide truly thromboresistant (nonclotting) artificial surfaces to circulating blood.

There have been various attempts to replace lost valve function in the heart. Modern designs provide either a passive movable occluder, such as a ball, or flapper elements, which are constrained by a ring and/or cage enclosure. Figure 3 illustrates a typical modern design in which a single free pyrolitic carbon disk is retained by a series of struts in a titanium-alloy ring covered with a woven polyester or polyethylene terephthalate

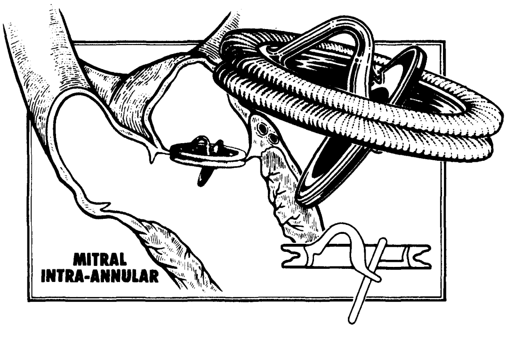

FIG. 3. Modern mitral valve replacement of the Hall type. This valve is passive, requiring blood backflow during diastole to close the orifice during a portion of each heartbeat cycle. (Courtesy Medtronic, Inc., Minneapolis, MN.)

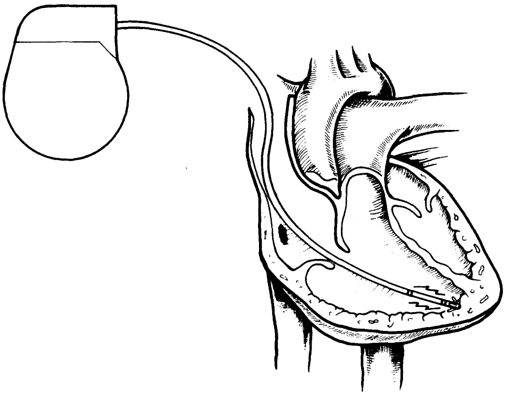

FIG. 4. Single-chamber pacemaker system. The different environments and performance requirements require different materials and combinations for each component (left to right): pacer, lead (with insulation), and dipole stimulating electrode. (Courtesy Medtronic, Inc., Minneapolis, MN.)

sewing ring for attachment to the opening left after excision of the diseased natural mitral valve. Continuing problems exist associated with wear of the occluder and fragmentation of the sewing ring. The latter component presents a significant problem if implant-site infection occurs. Such infections also promote thrombosis and are sufficiently life threatening that surgical replacement is frequently necessary.

Valves are also fabricated from porcine heart valves, with various chemical processes used to increase tissue strength and reduce antigenicity. Note that such devices are not meant to be resorbable but permanent and are limited in their utility by a tendency for the tissue leaflets to calcify with time *in vivo,* in a manner similar to naturally occurring atherosclerosis. Perhaps 30 000 heart valves of all types are implanted annually in the U.S. with probably ten times that number in use (Moss, 1988).

The implantation of electronic devices to regulate heart activity and, more recently, to permit closed-chest resynchronization (*defibrillation*) of erratic heart activity represents another major advance in biomedicine. Early devices were bulky, had battery lives as short as 1–2 y, and frequently experienced mechanical lead failure. Figure 4 illustrates a typical modern design. The electronic components, with a lithium battery that may last 5–7 y, are enclosed in an electron-beam–welded titanium-alloy case that is implanted in the chest cavity. The pacing lead is typically fabricated of a high–spring-constant cobalt- or platinum-based alloy in a polyurethane or polysiloxane insulating sheath. The electrode assembly is made of similar materials and is designed to be self-retaining after insertion. This device illustrates the point that there is no "ideal" biomaterial: Biomaterials, with appropriate biological performance of the application, must be selected for the particular processing and performance requirements of each component of the particular design.

Nearly half a million (Moss, 1988) U.S. patients have heart pacers today, with the incidence expected to rise rapidly as the population ages and materials and device technology continues to advance.

Vascular prostheses, as well as patches for heart-wall defect repair, are typically knitted or woven from polyester fiber, although some are fabricated from nonwoven polyester and/or polytetrafluoroethylene materials. The success of these devices depends upon achieving an appropriate porosity, generally measured as water permeability. Higher permeabilities favor tissue growth through the wall of the implant, leading to a more viable neointimal layer. However, higher permeabilities also permit greater blood loss intraoperatively and in the early healing period. All of these prostheses tend to shed small thrombi and to encourage excessive tissue overgrowth (*hyperplasia*) where they are connected to natural vessels, particularly at the "downstream" end.

Although vascular prostheses below 5-6 mm internal diameter are used, they tend to become blocked by thrombosis within weeks to months after implantation and require surgical clearance or replacement. Such periodic replacement is preferred to limb amputation, but it is clear that the problem of the small-bore vascular prosthesis remains unsolved.

While extracorporeal blood pumps and oxygenators are in routine use in surgical procedures, attempts to produce long-term implantable artificial hearts and, more recently, left ventricular assist devices have been largely unsuccessful. The primary clinical limitation is the combination of hemolysis and small-thrombi shedding, which may result in regional lung and brain infarction, produced by the biomaterials lining the blood conduit. More successful has been repetitive renal dialysis, in the presence of kidney malfunction or failure. Many patients are able to tolerate the required 2–3 times weekly 4–6-h contact with biomaterials in the blood conduit; however, kidney transplantation, from either donors or cadavers, is still a more definitive and long-term solution. Although chronic arterial–venous shunts, of various materials, are implanted to ease this procedure, no implantable dialysis system is in use clinically.

4.3 Dental and Maxillofacial

Dental and maxillofacial implant applications include filling defects in teeth produced by infection (*dental caries*), replacing portions or all of adult teeth lost through trauma or disease, providing bony anchorage for artificial teeth or groups of teeth (*dental bridges*), and augmenting bone for preservation of teeth or for retention of removable dental devices (*dentures*). Because of the proximity of the mouth to the GI tract, concern for biological performance of dental implant materials is usually extended to a wide variety of materials used in preparing dental implants, including cleaning agents, adhesives and cements, and impression and casting materials.

The success of modern dentistry is based upon the widespread use of silver–mercury amalgams for replacing diseased tooth tissue (*enamel* and *dentine*). Typically these consist of a solid phase, 40–70 wt % silver, 22–30 wt % tin, 2–30 wt % copper, which is mixed just before insertion 1:1 with metallic mercury. The formation of various mercury-containing intermetallic compounds produces a relatively stiff, strong, self-retaining body, which can be used within an hour and matures in 5–7 days. More recent concern about the possibility of mercury release and its conversion to biologically active compounds, such as methyl mercury, has resulted in various ceramic-particle–filled polymer-based chemoset composites being used in this application.

Tooth replacements were traditionally made of gold alloys, with strengthening alloying agents added, covered by ceramic enamels for prosthetic considerations. Recently, other precious-metal alloys, as well as nonprecious-metal alloys, such as cobalt- and nickel-based superalloys, have begun to be widely used for crown and bridge prostheses, driven by cost considerations. Cobalt-based alloys are widely used for fully implanted components, such as alveolar ridge augmentations designed to support artificial teeth. In fact, the initial use of such alloys for biomedical applications was in the related field of orthodonture where they won favor as being "tasteless," in contradistinction to the stainless steel alloys that they have displaced.

Pergingival devices, such as tooth posts, that permanently penetrate soft tissue became widely successful over the past decade after the observation of Bränemark that bone would bond directly (*osseointegrate*) to the surface of chemically pure titanium. Retention rates exceeding 95% for periods of up to ten years have been reported. More recently, more complex structures, involving surface textures and internal pores to permit bone ongrowth or ingrowth, have supplemented earlier smooth designs, and calcium hydroxyapatite coatings, believed to promote earlier bone apposition, have been added as adjuncts.

A wide variety of materials has been used to replace bone lost around teeth through periodontal disease and to reinforce and reconstruct bony support in individuals who have lost some or all of their teeth. The most widely used such material is microporous, granular calcium hydroxyapatite (100–300 μm), but other materials, including resorbable glasses and processed natural materials, such as processed bovine cancellous bone and coral, have seen some success.

The nature of dental practice, with most procedures performed in office rather than hospital settings, makes the annual volume of procedures difficult to estimate although one recent study (Langer and Vicanti, 1993) suggests an annual total of 10 000 000, with many patients receiving multiple implants.

4.4 Plastic and Reconstructive

Plastic and reconstructive applications focus on three problems: repair of birth defects, reconstruction of the residual effects of trauma or cancer surgery, and perceived need for esthetic enhancements. The first two applications represent some of the highest achievements of the surgical art. Soft materials, such as polyurethanes, polysiloxanes, and low- and mid-molecular-weight polyethylenes, are used in conjunction with soft- and hard-tissue materials, both harvested from other sites within the patient's body and obtained through an extensive tissue and organ bank system within the U.S. The use of vascularized flaps and bone segments and extensive stretching techniques to ensure skin coverage further enhances the functional and esthetic aspects of the results of such procedures. In some cases, when extensive reconstruction is not possible, as in the loss of portions of the nose or ears, fully synthetic external orthoses may be used, anchored to osseointegrated titanium percutaneous posts. Such devices have proven very successful for periods exceeding ten years. However, load-bearing applications, such as permanent anchorage of partial limb-replacement orthoses with percutaneous devices, still are not successful, because of continuing breakdown of the so-called *three-phase junction,* the point where tissue, implant, and air meet, leading to pain, infection, and marsupialization and requiring eventual removal of the device.

It is unfortunate that these truly miraculous advances have been to a considerable degree obscured by the profligate use of implants to enhance sexual function and attractiveness. The former goal, primarily in men, involves procedures to place solid or inflatable implants within the penis, for correction of impotence related to circulatory disease or as a consequence of pelvic surgery. Such devices, as well as prosthetic replacements for the testis, have proven of uneven clinical success.

Far more common, in the U.S. at least, has been the use of implants to increase breast size in women and, in some cases, to reconstruct breast volume and appearance after cancer surgery. Breast implants have been fabricated from polyurethane and polysiloxane elastomers and filled with silicone gel or saline solution. Implants with early grades of polyurethanes are no longer used, because of concerns about possible chemical carcinogenesis by aromatic amine degradation products. Implants with silicone gel fillers are no longer used because of problems of leakage, rupture, and implant-site calcification and scar contraction. Before such procedures were discontinued, hearsay estimates (in the absence of sound data) suggested that between 500 000 and 1 000 000 U.S. women had received breast implants. There is ongoing litigation in the U.S. concerning possible systemic and remote-site effects, believed due primarily to release and distribution of silicone gel. This litigation has resulted in a tentative class action damage settlement of $\$4.2 \times 10^9$, which has had a distinct chilling effect both on the supply of currently used polymeric biomaterials and on efforts to develop better replacement materials.

4.5 Miscellaneous Applications

One of the earliest successful modern implant applications in the U.S. was the development of the polysiloxane indwelling shunt for draining fluid from the cranial cavity into the spinal canal, thus preventing progressive brain damage in children born with the condition of *hydrocephalus.* The most common implant, the Holter shunt, is named for its inventor, John Holter, an aerospace engineer who pursued this problem because of the birth of a son with this defect. Related neurosurgical applications include closure of the skull after trauma or surgical procedure, frequently with tantalum plates or mesh, and attempts to encourage nerve regrowth by provision of sheaths fabricated from various metallic and polymeric materials.

Small–material-volume but high–procedure-rate devices play vital roles in modern medical management. These include polymeric vent tubes to drain the middle ear during infection (~300 000/yr) and polymethylmethacrylate intraocular lenses as a treatment for cataracts and other degenerative conditions of the eye in the elderly (~800 000/yr) (Moss, 1991).

Biomaterials play an extensive role in fertility control, especially outside the U.S. The use of intrauterine devices (IUDs) promotes an adverse local host response preventing implantation of the fertilized ovum in the uterus. The most common material used is pure copper, an ordinarily "incompatible" material, on a polymeric support. Polymers are also used as implantable reservoirs for long-term release of hormones for female fertility control. Male fertility may be controlled reversibly by injection of *in situ* curing polysiloxane into the vas deferens, preventing admission of fertile spermatozoa into the ejaculate.

4.6 Hybrid Devices

A number of hybrid devices, utilizing cells from patient or donor sources cultured *in vitro* and combined with resorbable or metabolizable supports and matrices, are in various stages of animal and preclinical evaluation. As of now, none have entered routine clinical use. There is general concern over regulatory issues related to the potential for inadvertently causing infection, especially by human immune viruses and other difficult to detect, very long-expression-time biological agents, such as Creutzfeldt-Jakob and Epstein-Barr viruses. These concerns are also present in the increasing use of "live" (unprocessed) tissue transplants. Recently, in the U.S., the FDA has indicated that it intends to regulate the quality and safety of transplanted tissues and organs. This will occur initially through oversight of the conduct and practices of tissue banks, and will certainly lead to active regulation of the safety and effectiveness of hybrid biomaterials.

5. CONCLUSIONS

5.1 Clinical Status

Biomaterials are in broad use today in the U.S. and the developed countries. Their use in the developing countries of the third world is still limited by lack of manufacturing facilities, lower surgical standards, poor hygiene, and, in some cases, cultural opposition. In the U.S., various estimates made in the early 1990s suggest that as many as twenty million permanent (>30 days duration) surgical implants may now be in use, with orthopedic and cardiovascular applications attracting the most attention. Recent poor experience with silicone-gel–filled silicone breast implants and polytetrafluoroethylene-containing interposition spaces for the temporomandibular joint (TMJ) have led to a current re-evaluation of the commercial, legal, and ethical risks associated with the use of chronic implants. One immediate effect is the present and pending withdrawal from the U.S. market by many manufacturers of materials used in implants. Since most current biomaterials are modified or selected grades of commodity materials used in vastly greater volumes for other applications, such withdrawal in the face of rising product-liability litigation is understandable but is raising questions about the future availability of biomaterials.

Implant performance is strongly device and application oriented; thus, it is not possible to provide a list of generally recognized and accepted (thus, safe and effective) implantable biomaterials. However, the list given below is of inert and interactive biomaterials currently in human clinical use in the U.S. in at least one application.

Polymers:
- Polyacetal
- Polyamide (Nylon™ 6,6)
- Polycarbonate
- Polyethylene (ultrahigh molecular weight)
- Polyethylene terephthalate (Dacron™)
- Poly(tetrafluoroethylene) (Teflon™, Gore-Tex™)
- Polylactic/polyglycolic acid copolymers (resorbable)
- Poly(methyl methacrylate)
- Polyurethanes
- Polypropylene
- Poly(dimethylsiloxane) (silicone elastomer) (Silastic™)
- Polysulfone
- Polyvinyl chloride

Metals and Alloys:
- Cobalt–chromium–molybdenum based al-

  - loys (collectively but incorrectly termed Vitallium™)
  - Gold and gold-based alloys (dental only—usually not implanted)
  - Copper–mercury–tin–silver amalgam (dental only—not implanted)
  - Platinum- and palladium-based alloys
  - Stainless steels (primarily type 316 low carbon; type 304 used in very restrictive applications)
  - Tantalum
  - Titanium
  - Titanium–aluminum–vanadium (α–β type) alloy
  - Titanium-based β-type alloys
- Ceramics:
  - Aluminum oxide (alumina) (bulk, particulate in resin base)
  - Calcium hydroxyapatite (coatings, powder, granules, porous/dense block)
  - Carbon (coatings, pyrolitic, vitreous)
  - Zirconium oxide (zirconia) (bulk)

5.2 The Challenge of New Applications

The field of biomaterials is undergoing rapid change and development. The search for inert materials has been essentially abandoned, with the exception of continuing interest in surface modification, both to reduce wear and debris formation and to render tissue reaction more benign. Major efforts center on interactive materials. Extensive efforts are being made to decipher the messages that cells receive from surfaces. It is now clear that some cell types can sense the chemical, morphological, and electrical aspects of surfaces independently and that preferred combinations of these aspects of the biomaterial–tissue interface selectively mediate cellular response. Molecular design, made possible by microbiological and tissue engineering techniques, to replicate specific amino acid sequences in polymeric proteins, has produced both novel materials and surface modifications exhibiting specific molecular sequences utilized by cells for cell–cell and cell–matrix communication and binding.

The exciting future lies, however, with the development of hybrid materials, with live cells included within a synthetic matrix and perhaps actually encapsulated so as to permit diffusion of small molecules but prevent interaction with the patient's immune system. Applications include hybrid synthetic skin and small-diameter blood vessels as well as encapsulated pancreatic islet cells for the treatment of diabetes.

These approaches are still only a stepping stone to the eventual future of biomaterials. The artificial heart and left ventricular assist devices, as temporary implants, are now seen as bridges to transplantation of whole organs from donors. Similarly, renal dialysis is viewed, for all but the elderly and those with multiple organ diseases, as a temporizing procedure until an immune-matched donor organ can be obtained. As control of tissue growth and elaboration *in vitro* becomes increasingly better understood, we can look forward to biomaterials serving as bridges until the implantation, or, better, replantation of cells, tissues, and organs grown from tissue removed from the specific patient under treatment. Such techniques are already in use in cancer therapy, involving selective growth of leukocyte subpopulations *in vitro* followed by transfusion to the donor. As these techniques become perfected and patients begin to be able to receive live tissues, containing their specific genetic code and immune factors, then biomaterials will have achieved its true promise of helping to alleviate human disability and disease.

GLOSSARY

Adaptation: The ability of tissues to adapt themselves to local requirements, including reaction to the chemical, physical, or electrical properties of implants.

Bioactive: The property of an **Implant** surface of adhering directly to soft or hard tissue without an intermediate layer of modified tissue.

Biocompatible Material: A material having acceptable host and material response in a *specific* application. (See **Response, Host; Response, Material.**)

Biomaterial: A nonviable material used in a medical device, intended to interact with biological systems. Classified into Biomaterial, Inert: A biomaterial that elicits little or no **Host Response;** Biomaterial, Interactive: A biomaterial that is designed to elicit or promote a specific **Host Response,** such as hard-tissue adhesion; Biomaterial, Manmade (or Manufactured): A biomaterial that is significantly processed from raw materials of inorganic or organic origin; Biomaterial, Native (or Natural): A biomaterial that is obtained from natural sources and is implanted essentially unprocessed; Biomaterial, Replant: A biomaterial consisting of live native cells or tissue, cultured *in vitro* from cells obtained previously from specific patients or donors; Biomaterial, Viable: A biomaterial that incorporates host tissue and/or live cells, and is capable of being actively remodeled and/or resorbed.

Carcinogenesis: Malignant, inheritable change in mammalian cells. Classified into Carcinogenesis, Chemical: (with respect to an implant) Carcinogenesis induced by the chemical composition of an **Implant** or its degradation products; Carcinogenesis, Foreign Body: Carcinogenesis induced by the physical form of an **Implant,** independent of its chemical composition.

Environment, Biological: Conditions encountered by an **Implant** within an animal or human body.

Environment, Pericellular: Conditions encountered by an **Implant** immediately adjacent to living cells within an animal or human body.

Epithelial Tissue: Skin and the lining of natural internal body cavities.

Extrusion: Resolution in which an **Implant** in contact with **Epithelial Tissue** is surrounded by a downgrowing extension of such tissue, directed toward extruding the implant from the body. Also called **Marsupialization,** from the resemblance of the tissue to a kangaroo's pouch.

Glycocalyx: A protective enveloping film formed on the surface of **Implants** by some types of bacteria.

Homeostasis: The maintenance of conditions necessary for mammalian life.

Hybrid Devices: Devices using cells from patient or donor sources cultured *in vitro* and combined with **Resorbable** or metabolizable supports and matrices.

Implant: A device placed within an animal or human body by the act of **Implantation.** Classified as Implant, Acute: A device that remains *in situ* for less than 30 days; Implant, Chronic (or Permanent): A device that remains *in situ* for 30 or more days; Implant, Intraoperative: A device removed within hours or days, at the termination of the surgical or therapeutic procedure; Implant, Percutaneous: A device that, after placement, penetrates the skin.

Implantation: Placement of a device or material within the body of an animal or human by a medical or surgical professional, in such a way as to breach one or more epithelial layers, and to leave materials and/or components in place after the initial procedure.

Implant Life History: Lifetime performance requirements for an **Implant.**

Incapsulization: Resolution in which the **Implant** is surrounded and walled off from normal tissue by a collagenous, relatively acellular tissue termed *capsule,* which much resembles scar tissue. In a bony location, the capsule may be mineralized and is called a *sequestrum.*

Inflammatory Response: The cell-mediated response directed toward stabilizing injured tissue, restoring physiological *status quo ante,* removing dead or damaged tissue elements and foreign material, and correcting the structural and functional loss due to the initial insult. The four classical signs of inflammation are redness, swelling, pain, and heat.

Integration: Resolution for a very limited number of materials, such as "bioactive" glasses of selected compositions, or some metals, such as pure titanium, for which direct "bonding" or apparent adhesion to normal tissue may take place.

Marsupialization: See **Extrusion.**

Oppenheimer Effect: Induction of tumors in rodents through a foreign-body mechanism; named after its discoverers, E. and B. S. Oppenheimer.

Osseointegration: Clinical stability of an **Implant** anchored in bone. Often taken to refer to implants with **Bioactive** coatings.

Osteolysis: Cellularly mediated bone loss (also called *small-particle disease*) secondary to wear-debris production by total joint replacement prostheses.

Phagocytosis: The process of internalizing small particles by mammalian cells.

Physiological Elements: Those nontrace elements (calcium, phosphorus, potassium, sulfur, sodium, chlorine, and iron), other than oxygen, hydrogen, nitrogen, and carbon, required for mammalian **Homeostasis.**

Pseudointima: Tissue, consisting of a firm fibrin clot with occasional islands of endothelial cells, formed by **Resolution** on interiors (blood-contacting surfaces) of cardiovascular **Implants.**

Resolution: The end state of the **Inflammatory Response** in the presence of an **Implant.**

Resorbable: Of an **Implant,** able to be dissolved or digested, and thus to disappear after **Implantation.**

Resorption: Resolution associated with **Resorbable Implants** in which the tissue site condenses to a collapsed scar or, in the case of bone, completely remodels to normal (regenerated) tissue.

Response, Host: The reaction of a living system to a **Biomaterial.**

Response, Material: The reaction of a **Biomaterial** to a living system.

Stress Shielding: A decreased density of bone, as a consequence of load sharing between an **Implant** and the hard tissue.

Three-Phase Junction: With reference to percutaneous **Implants,** that point where tissue, **Implant,** and air meet.

Tissue Engineering: Elaboration of cells and tissues outside a living organism, for replantation, by use of engineering methods and techniques.

Vroman Effect: The succession of molecular species adherent to surfaces of **Implants,** named after its discoverer, L. Vroman.

Wolff's Law: "The form being given, tissue adapts to best fulfill its mechanical function" (Wolff, 1892).

Works Cited

Langer, R., Vicanti, J. P. (1993), *Science* **260,** 920–926.

Moss, A. J. (1991), *Use of Selected Medical Device Implants in the United States, 1988.* Advance Data from Vital and Health Statistics of the National Center for Health Statistics No. 191 (February 26), 1–24.

Steinemann S. G. (1980), in: G. D. Winter *et al.* (Eds.), *Evaluation of Biomaterials,* Chichester: Wiley, pp. 1–34.

Williams, D. F. (Ed.) (1987), *Definitions in Biomaterials,* Amsterdam: Elsevier.

Wolff, J. (1892), *The Law of Bone Remodelling,* translated by P. Maquet and R. Furlong, Berlin: Springer-Verlag.

Further Reading

Agnew, W. F., McCreery, D. B. (Eds.) (1990), *Neural Prostheses Fundamental Studies,* Englewood Cliffs: Prentice-Hall.

Black, J. (1987), *Electrical Stimulation: Its Role in Growth, Repair and Remodeling in the Musculoskeletal System,* New York: Praeger.

Black, J. (1988), *Orthopaedic Biomaterials in Orthopaedic Research and Practice,* New York: Churchill-Livingstone.

Black, J. (1992), *Biological Performance of Materials: Fundamentals of Biocompatibility,* 2nd ed., New York: Marcel Dekker.

Byrom, D. (Ed.) (1991), *Biomaterials. Novel Materials from Biological Sources,* New York: Macmillan.

Cameron, H. U. (Ed.) (1993), *Bone Implant Interface,* St. Louis: C. V. Mosby.

Craig, R. G. (Ed.) (1993), *Restorative Dental Materials,* 9th ed., St. Louis: C. V. Mosby.

Greco, R. S. (Ed.) (1994), *Implantation Biology,* Boca Raton: CRC Press.

Harker, L. A., Ratner, B. D., Didisheim, P. (Eds.) (1993), "Cardiovascular Biomaterials and Biocompatibility," *Cardiovascular Pathology* **2** (3) (Suppl.), 1S–224S.

Langer, R., Vicanti, J. P. (1993), "Tissue Engineering," *Science* **260,** 920–926.

Mertz, W. (Ed.) (1986), *Trace Elements in Human and Animal Nutrition,* 5th ed., New York: Academic.

Morrey, B. F. (Ed.) (1993), *Biological, Material and Mechanical Considerations of Joint Replacement,* New York: Raven Press.

Ravaglioli, A., Krajewski, A. (1992), *Bioceramics,* London: Chapman & Hall.

Remes, A., Williams, D. F. (1992), "Immune Response in Biocompatibility," *Biomaterials* **13,** 731–743.

Shoen, F. J. (1989), *Interventional and Surgical Cardiovascular Pathology,* Philadelphia: W. B. Saunders.

Savio, J. A. III, Overcamp, L. M., Black, J. (1994), "Size and Shape of Wear Debris," *Clin. Mater.* **15,** 101–147.

Silver, F., Doillon, C. (1989), *Biocompatibility: Interactions of Biological and Implantable Materials.* Vol. 1. *Polymers,* New York: VCH Publishers.

Tencer, A. F., Johnson, K. D., Kyle, R. F., Fu, F. H. (1993), "Biomechanics of Fractures and Fracture Fixation," in: J. D. Heckman (Ed.), *Instutional Course Lectures,* Chicago: American Academy of Orthopedic Surgeons.

von Recum, A. F. (Ed.) (1986), *Handbook of Biomaterials Evaluation,* New York: Macmillan.

Williams, D. F., Roaf, R. (1973), *Implants in Surgery,* London: W. B. Saunders Co. Ltd.

PROTEIN DYNAMICS

G. Ulrich Nienhaus, *Department of Physics, University of Illinois at Urbana–Champaign, Urbana, Illinois, U.S.A.*

Robert D. Young, *Department of Physics, Illinois State University, Bloomington–Normal, Illinois, U.S.A.*

Introduction 163
1. **Protein Structure and Dynamics—Overview** 164
1.1 Structure 164
1.2 Dynamics 165
2. **Concepts of Protein Dynamics** 165
2.1 Protein States, Substates, and Motions 166
2.2 Nonergodicity, Temperature, and Time Scales 166
2.3 Reaction-Rate Theory for Proteins 166
2.4 Time and Temperature Dependence of Conformational Motions 167
2.5 Conformational Dynamics and Protein Reactions 167
2.5.1 Distributed Kinetics 168
2.5.2 Averaged Kinetics 168
2.5.3 Time-Dependent Reaction-Rate Coefficients 168
2.6 Viscosity Dependence of Reaction Rates and Protein Motions 169
2.7 Protein Dynamics and Function 169
2.8 Emerging Concepts 170
2.8.1 Hierarchy of Substates 170
2.8.2 Proteins, Glasses, and Spin Glasses 170
3. **Molecular Dynamics Simulations** 171
3.1 Essentials of Molecular Dynamics 172
3.2 Some Applications of Molecular Dynamics to Myoglobin 172
4. **Experimental Methods** 173
4.1 X-Ray Crystallography 173
4.2 Mössbauer Spectroscopy 175
4.3 Neutron Scattering 176
4.4 Nuclear Magnetic Resonance 177
4.5 Hydrogen Exchange 178
4.6 Fluorescence Spectroscopy 179
4.7 Transient Optical Spectroscopies 179
4.7.1 Ultraviolet Spectroscopy 180
4.7.2 Visible/Near-Infrared Spectroscopy 180
4.7.3 Infrared Spectroscopy 181
4.7.4 Raman Scattering 181
4.7.5 Other Experimental Techniques 181
5. **Protein Dynamics—A Reprise** 182
Glossary 182
Works Cited 183
Further Reading 184

INTRODUCTION

Proteins perform a wide variety of tasks in living organisms (Stryer, 1988, and the preceding article Proteins and Enzymes): Some proteins have structural roles, for example, as constituents of muscle and fibrous tissue. Other proteins store or transport energy, electrons, or small molecules. Moreover, many different enzymes exist in living systems that catalyze a large number of essential reactions, and almost all enzymes are proteins. Antibodies are proteins elicited by the immune system to remove foreign materials from the organism.

To perform these diverse functions, protein molecules have specific three-dimensional structures, designed and fine-tuned by evolution over billions of years. Perutz and Kendrew solved the first protein structures with atomic resolution using x-ray crystallography (Perutz *et al.*, 1960; Kendrew *et al.*,

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1960). The remarkable detail and intrinsic beauty of the molecular structures led to the misconception that proteins are rigid molecules, and most attempts to understand protein function have been based on rigid models. For example, the specificity and efficiency of an enzymatic reaction have been explained by complementary interactions between rigid enzyme and substrate that were likened to the fitting of a key within a lock.

In the past 20 years, the static view of proteins has undergone a complete revision. It is now generally recognized that proteins are fluctuating entities; they constantly change the details of their conformation, and the x-ray structure reveals only an average over the many different conformational substates that the molecule can assume within one conformation. This development toward a dynamic picture of proteins was promoted by advances in experimental techniques that allow the study of motions on all relevant time scales and by advances in theory. Furthermore, the availability of powerful computers makes it possible to study protein motions on time scales up to a few nanoseconds using molecular dynamics (MD) simulations.

At present, the detailed mechanisms by which structure and dynamics give rise to the particular function of a given protein are still only partially understood. Research focuses on a few proteins that have been characterized well enough to allow for the detailed study of the structural dynamics. Of particular importance are heme proteins, especially the oxygen-binding myoglobin and hemoglobin.

We begin this article with a brief overview of the structure and dynamics of proteins. We then explain some essential concepts of protein dynamics. The wide field of molecular dynamics simulation of protein dynamics is introduced in Sec. 3. In Sec. 4, we give an overview of some major experimental techniques with which motions in proteins are studied. Since not all techniques can be covered, the reader is referred to the books listed under Further Reading at the end of the article.

1. PROTEIN STRUCTURE AND DYNAMICS—OVERVIEW

1.1 Structure

Proteins are linear polymers of amino acids, which are covalently linked by peptide bonds, as illustrated in Figs. 1 and 2 of PROTEINS AND ENZYMES. The polypeptide chain is intrinsically flexible since many of the covalent bonds permit rotations. The monomers of the polypeptide chain are chosen from a set of 20 different amino acids, which have the same backbone structure (except for proline) but differ in structural and chemical properties of the sidechains. The amino acids can be divided in two classes, hydrophilic residues, which are more soluble in water, and the less soluble hydrophobic residues. Hydrophobic residues are neutral and relatively nonpolar, whereas hydrophilic residues may be charged or possess dipole moments.

The structure of a given protein is encoded in the DNA (deoxyribonucleic acid) in a linear sequence of nucleotide triplets. In the ribosome, this information is used to synthesize a linear polymer chain of amino acids, called the primary structure of the protein. In the proper environment, the polypeptide chain folds into its compact, characteristic three-dimensional structure. This process occurs spontaneously for simple proteins. Many proteins are subsequently modified by attachment of prosthetic groups, metal ions, or carbohydrates.

The native structure of a globular protein represents an efficient packing of the atoms, comparable to that in a dense liquid. The interactions that lead to the compact, native structure are not fully understood; thus it is currently not yet possible to predict the three-dimensional structure from the amino-acid sequence. The attractive interactions that determine the native structure include protein–protein, protein–solvent, and solvent–solvent interactions. The most important type of protein–protein interaction is hydrogen bonds. Many groups along the chain are hydrogen-bond donors and acceptors. Upon folding, hydrogen bonds to water molecules are replaced by internal hydrogen bonds. Hydrogen bonds between the amide and carbonyl groups of the amino-acid backbone are established by regular folding arrangements of the backbone, like α helices and β sheets. These folding patterns are called secondary structure. In addition to hydrogen bonds, weakly directional dispersion (van der Waals) forces lead to tight packing of internal protein groups. In some proteins, one or a few disulfide bridges between cysteine residues establish additional covalent cross links.

The three-dimensional arrangement of the polypeptide chain is termed tertiary structure. Hydrophobic sidechains are predominantly found in clusters in the interior of the protein, excluded from the polar solvent. Charged amino-acid sidegroups are located on the exterior, where they can favorably interact with water molecules. In the interior, charged groups are normally paired with groups of opposite charge. In larger proteins, the tertiary structure can be subdivided into domains. In many proteins, spontaneous aggregation of different polypeptide chains creates oligomers; this structural arrangement is called quaternary structure.

Since the interactions that stabilize the three-dimensional structure are weak and only weakly directional, a folded protein molecule does not exhibit a unique state of lowest free energy. The ground state is highly degenerate, so that in each conformation, the protein exists in a large number of almost isoenergetic structures, called conformational substates. They can be represented as minima in a multidimensional conformational energy surface. While the overall structure of the conformational substates is the same, they differ slightly in the arrangement of the atoms. However, their functional properties may be vastly different.

1.2 Dynamics

A polypeptide chain consists of fairly rigid groups of atoms, linked by bonds that permit rotations. Examples of rigid groups are the CONH unit of the backbone and the phenyl group of the phenylalanine amino acid. Within the rigid groups, only small relative displacements of the atoms occur because of the large energy expense of changing bond lengths and angles and rotating around multiple bonds. Substantial displacements arise from torsional motions around single bonds, which are easily rotated. Since proteins are densely packed, conformational rearrangements can only occur in a concerted way; for one group to move, neighboring groups have to give way. Therefore, the dynamics of proteins show strong analogies to the dynamics of viscous liquids. Motions on time scales ranging from $\sim 10^{-14}$ s up to macroscopic time scales are observed. An overview of the properties of protein motions under physiological conditions is given in Table 1 (McCammon, 1984).

On picosecond time scales, molecular groups vibrate in cages created by neighboring groups and, on the protein surface, by the solvent shell. These motions are of rather small amplitude. Large-scale displacements occur as collective motions involving several residues and therefore happen on time scales of nanoseconds and longer. Since the topological arrangement of the atoms changes in these motions, they are described as transitions between conformational substates.

2. CONCEPTS OF PROTEIN DYNAMICS

A number of concepts play a crucial role in protein dynamics. These concepts also form part of the intellectual framework of condensed matter physics and chemical dynamics. However, the special features of protein structure and biological function add a new dimension. The concepts apply to all biomolecules, including proteins and nucleic acids.

Table 1. Typical features of internal motions of proteins.

Motion	Spatial extent (nm)	Amplitude (nm)	Logarithm of characteristic time (s)
Relative vibration of bonded atoms	0.2–0.5	0.001–0.01	−14 to −13
Elastic vibration of globular region	1–2	0.005–0.05	−12 to −11
Rotation of sidechains at surface	0.5–1	0.5–1	−11 to −10
Torsional libration of buried groups	0.5–1	0.05	−11 to −9
Relative motion of different globular regions ("hinge-bending")	1–2	0.1–0.5	−11 to −7
Rotation of medium-sized sidechains in interior	0.5	0.5	−4 to 0
Allosteric transitions	0.5–4	0.1–0.5	−5 to 0
Local denaturation	0.5–1	0.5–1	−5 to +1

2.1 Protein States, Substates, and Motions

Frequently, protein molecules exhibit a few structurally well distinguishable states or conformations. For example, an enzyme exists in the substrate-free or substrate-bound state; the oxygen-storage protein myoglobin exists in the deoxy state Mb or ligand-bound state, MbCO or MbO_2. Each of the states of a working protein has an overall structure determined by the amino-acid sequence. However, because of the complexity of a protein, a given protein state consists of many conformational substates (CS). The conformational substates of a given protein state have the same overall structure and perform the same biological function, but differ in structural details and possibly in other properties, such as the rate at which the function is performed. The conformational energy surface of a protein is rugged with many energy minima and maxima—the energy minima are the conformational substates (Frauenfelder *et al.*, 1979; Frauenfelder *et al.*, 1988). Conformational transitions can occur in two different ways:

1. Equilibrium fluctuations. Since proteins are mesoscopic systems, they show considerable fluctuations of energy, entropy, and volume, due to their interaction with the thermal bath (Cooper, 1976). Consequently, their molecular structure fluctuates continuously.
2. Nonequilibrium relaxations. These occur as a response to a perturbation. For example, absorption of a light quantum or dissociation of a ligand can leave the molecule in a nonequilibrium state from which it subsequently relaxes.

In the linear response regime, both types of motions are connected by fluctuation–dissipation theorems (Landau and Lifshitz, 1980).

2.2 Nonergodicity, Temperature, and Time Scales

The rugged energy landscape of proteins leads to a strong dependence of protein motions on temperature and time. Within a given time interval, for example the characteristic time of a particular measurement, a protein molecule may not be able to visit all conformational substates. *Nonergodicity* is the term that characterizes this situation. At physiological temperature, a protein may be able to explore all substates of its rugged conformational energy surface on a time scale of microseconds, and the protein is *ergodic* on this time scale. However, as temperature is decreased, equilibrium fluctuations slow, and the protein becomes more and more confined in a subregion of conformational space within a given time interval, until finally, the protein molecule is caught in a local minimum on any reasonable experimental time scale (Palmer, 1982). For example, carbonmonoxy-myoglobin (MbCO) in glycerol/water solvent has been shown to experience a transition to a glasslike state at $T_g \approx 180$ K (Frauenfelder *et al.*, 1991).

2.3 Reaction-Rate Theory for Proteins

Chemical reactions are the basis for all aspects of biological processes, from transfer of matter, charge, and information to enzyme catalysis. *Transition-state theory*, derived from studies of two-body interactions of small molecules in gas phase, provides the starting point for understanding biological reactions (Frauenfelder and Wolynes, 1985): If we consider the two-state reaction $B \to A$, where all molecules are in state B at time $t = 0$, and neglect back transitions $A \to B$, the time dependence of the fraction $N(t,T)$ of species B at time t and temperature T is given by a simple exponential, $N(t,T) = \exp[-k_{BA}(T)t]$. If the reaction is thermally activated, the rate coefficient $k_{BA}(T)$ is expressed by an Arrhenius relation,

$$k_{BA}(T) = A_{BA} \exp[-H_{BA}/RT]. \qquad (1)$$

Here, H_{BA} is the activation enthalpy, and the pre-exponential A_{BA} includes an attempt frequency, a transmission factor, and an activation entropy S_{BA}. The gas constant $R = 8.314$ J K^{-1} mol^{-1}. H_{BA} is schematically illustrated in Fig. 1(a). The activation free energy G_{BA} can be expressed in terms of the activation enthalpy H_{BA} and entropy S_{BA}, $G_{BA} = H_{BA} - TS_{BA}$. In elementary transition-state theory, the pre-exponential $A_{BA}^{*} = \nu_0 \exp[S_{BA}/R]$, where $\nu_0 \approx 10^{12}$ s^{-1} in small-molecule, gas-phase reactions. In biological reactions, the activation entropy plays a crucial role,

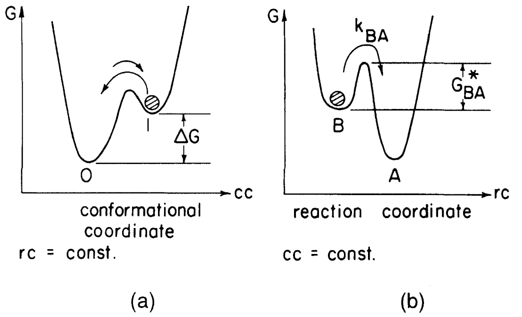

FIG. 1. Two cross sections through a highly simplified energy hypersurface of a protein. (a) At constant reaction coordinate (e.g., B in panel b) the protein possesses two conformational substates 0 and 1. (b) At constant conformational coordinate (e.g., 0 in panel a) the protein possesses two states. The bullet represents the protein–ligand system.

and ν_0 and the pre-exponential A_{BA} can differ considerably. For example, in the binding of CO and O_2 to myoglobin or of NO to azurin (Az) at low temperature, the pre-exponential $A_{BA} \approx 10^9\ s^{-1}$ (Steinbach *et al.*, 1991; Ehrenstein and Nienhaus, 1992). This implies that a typical activation entropy barrier S_{BA} for bond formation in protein reactions is given by $S_{BA} \approx -7R$.

Biological reactions can proceed by quantum-mechanical tunneling (Chance *et al.*, 1979). In this situation, the Arrhenius relation, Eq. (1), is inappropriate and must be replaced by a more general quantum-mechanical expression. Molecular tunneling has been observed in CO binding to myoglobin below 10 K; electron tunneling is involved in the charge separation in photosynthetic reaction centers at all temperatures.

2.4 Time and Temperature Dependence of Conformational Motions

Relaxation methods are frequently used in studies of protein motions. The sample is rapidly brought to a nonequilibrium state by a perturbation. The return to equilibrium is characterized by a relaxation function $\Phi(t,T)$. Simple systems show exponential (Debye) behavior. In proteins, the complexity of the system often results in a nonexponential time dependence of the relaxation function, analogous to relaxation processes in glasses (Frauenfelder *et al.*, 1991). The time dependence in both glasses and proteins can usually be well parameterized by a stretched exponential:

$$\Phi(t,T) = \exp(-[\kappa(T)t]^{\beta}), \tag{2}$$

where $0 < \beta < 1$. In proteins, $\beta \approx 0.3$–0.5 has been observed, indicating that the relaxation extends over many orders of magnitude in time. The effects of the rugged energy surface of the system are also reflected in the temperature dependence of the characteristic rate coefficient, $\kappa(T)$. Fits to relaxation data for proteins using an Arrhenius relation, $\kappa(T) = \kappa_0 \exp[-E/RT]$, are good only over a small range in temperature. The locally determined activation energy E increases strongly with decreasing temperature. Empirical relations for $\kappa(T)$ derived from glass-forming liquids can be used to describe the temperature dependence of relaxation processes in proteins (Frauenfelder *et al.*, 1991; Frauenfelder and Wolynes, 1994).

2.5 Conformational Dynamics and Protein Reactions

A protein exists in a large number of different structures or conformational substates, and the rate of a protein reaction will generally be different for different substates. The kinetics of a biomolecular reaction is then controlled by both the reaction rate coefficients within the substates and the fluctuation rate coefficients among the substates. The simplest model that allows us to discuss the interplay between the reaction and conformational transitions uses only two different protein states, A and B, along the reaction coordinate (*rc*) and two different conformational substates (0 and 1) along the conformational coordinate (*cc*). We stress that a single conformational coordinate with only two substates is an oversimplification and that there may also be additional protein states along the reaction coordinate. Figure 1 illustrates the model: Figure 1(a) shows a cross section through the energy surface at constant reaction coordinate, say $rc = B$. Figure 1(b) gives a cross section at constant conformational coordinate, say $cc = 1$. The rate coefficients for the reactions $B \leftrightarrow A$ depend on the protein substate: The rate coefficients for $B \rightarrow A$, for example, are written k^i_{BA}, where $i = 0, 1$ is the substate index. In the following, we assume an absorbing state A for simplicity and hence neglect the transition $A \rightarrow B$. The rate coefficients for transitions between conformational substates, $0 \leftrightarrow 1$, are given by κ_{ij}, $i \neq j = 0, 1$. The ratio of the relative populations of the two conformational substates 0 and 1 is $w_0/w_1 = \exp[-\Delta G(T)/RT]$, where the free-energy difference $\Delta G(T)$ between the two substates is $\Delta G(T) = \Delta H - T\Delta S$. Three different cases can be distinguished, dependent on the relative magnitudes of k^i_{BA} and κ_{ij}.

2.5.1 Distributed Kinetics When the rate coefficients κ_{ij} for conformational transitions are small compared with the rebinding rate coefficients, the survival fraction for the B states is biexponential, $N(t,T) = w_0(T_g) \times \exp[-k^0_{BA}t] + w_1(T_g) \exp[-k^1_{BA}t]$. This case is observed in proteins at low temperature, and the relative populations of the substates are frozen in at the glass transition temperature, T_g. In general, the number of conformational substates is very large so that the survival fraction can be described by a distribution (Frauenfelder *et al.*, 1988):

$$N(t,T) = \int g(H_{BA}) \exp[-k(H_{BA},T)t]dH_{BA}, \qquad (3)$$

where $g(H_{BA})dH_{BA}$ is the probability to find protein molecules with barriers between H_{BA} and $H_{BA} + dH_{BA}$. The distribution function $g(H_{BA})$ is essentially temperature independent below T_g. Substates that are described by distributions such as $g(H_{BA})$ are called *statistical substates.* Figure 2 illustrates the nonexponential time dependence of the rebinding kinetics of MbCO following photodissociation at temperatures between 60 and 160 K. The temperature-independent distributions of activation enthalpy barriers for the low-temperature recombination of MbCO and AzNO are also shown in Fig. 2.

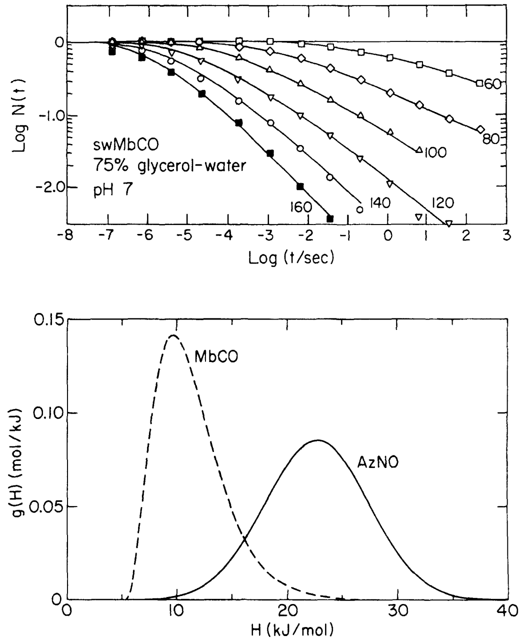

FIG. 2. Upper panel: rebinding kinetics of CO to Mb after flash photolysis between 60 and 160 K. Lower panel: activation enthalpy distributions $g(H_{BA})$ for MbCO and AzNO. Solvent: 75% glycerol/buffer.

2.5.2 Averaged Kinetics If the rate coefficients κ_{ij} are large relative to k^i_{BA}, the survival fraction $N(t,T)$ is exponential, $N(t,T) = \exp[-\langle k_{BA}\rangle t]$, and the average rate coefficient $\langle k_{BA}\rangle$ is given by $\langle k_{BA}\rangle = w_0(T)k^0_{BA} + w_1(T)k^1_{BA}$. Here, the $w_i(T)$ denote the equilibrium populations at temperature T. In the case of a large number of conformational substates, the average rate coefficient must be expressed in terms of a distribution of rate coefficients. Experimentally, averaged kinetics are observed for reactions slow compared with conformational fluctuations, e.g., bimolecular rebinding of ligands to heme proteins at sufficiently high temperature.

2.5.3 Time-Dependent Reaction-Rate Coefficients The most complicated case arises when the time scales for conformational transitions and for the reaction are similar. Reaction-rate theory including dynamic modulation of the microscopic rate coefficients by a fluctuating environment is under active development. Protein motions can result in a time-dependent rate coefficient $k_{BA}(t,T)$ so that the survival fraction becomes nonexponential in time,

$$N(t,T) = \exp\left[-\int_0^t k_{BA}(\tau,T)d\tau\right]. \qquad (4)$$

A time-dependent rate coefficient has been used, for example, to model the bond formation step between CO or NO and myoglobin in flash photolysis experiments (Petrich *et al.*, 1991; Steinbach *et al.*, 1991). Motions are also necessary for ligands or substrates to enter or exit the binding site of a protein. For example, fluctuating gate models, in which passage of a ligand through the protein depends on the opening and closing of channels by conformational fluctuations, have been used to explain these dynamic phenomena (Beece *et al.,* 1980).

2.6 Viscosity Dependence of Reaction Rates and Protein Motions

Proteins function in solvents or lipid membranes of differing viscosity. Since proteins are flexible, their motions and, hence, biochemical reactivity are influenced by the solvent composition and viscosity. In particular, the transition rates between conformational substates depend strongly on the solvent viscosity η (Beece *et al.,* 1980). The transition-state expression, Eq. (1), does not properly include the effects of solvent viscosity on transition rate coefficients. The approach developed by Kramers is more appropriate (Frauenfelder and Wolynes, 1985). For proteins, a modified Kramers relation should be used because of shielding of the protein interior from the external solvent. The rate coefficient for conformational transitions in a wide range of solvents can be expressed as

$$\kappa = \kappa_0(\eta) \exp[-\Delta G^*/RT]. \qquad (5)$$

Here ΔG^* is the free-energy barrier for the conformational transition. The pre-exponential $\kappa_0(\eta)$ depends strongly on solvent viscosity, a fact not included in the phenomenological Arrhenius relation in Eq. (1) for a simple two-state, gas-phase reaction. The viscosity dependence arises since ligands and substrates enter a binding site of a biomolecule as a result of protein fluctuations that couple to the solvent.

2.7 Protein Dynamics and Function

Heme proteins such as hemoglobin and myoglobin have played the role of model systems in developing our present understanding of protein dynamics, as described in the previous sections. Here we briefly sketch the connection between protein motions and function for a simple biological process, the binding of small ligands such as CO and O_2 to myoglobin (Mb). The basic features of the reaction have been elucidated, but many details are still not entirely understood and are being investigated at present. In the simplest model, the ligand binding reaction can be described with the reaction energy surface in Fig. 3, which involves three states A, B, and S. State A refers to the ligand-bound state, say MbCO, where the CO molecule is bound to the heme iron. State B refers to the unbound state where the CO is in a relatively open region of the protein, known as the heme pocket. In state S, the ligand has escaped from the protein into the solvent; the protein in state S is called deoxymyoglobin. Two barriers are involved in the binding process. The inner barrier between A and B is associated with the final binding step at the heme iron, and the outer one between B and S may represent the gate between the heme pocket and the solvent.

At physiological temperatures, this energy surface is by no means static but dynamic as a result of fluctuations and relaxations of the protein molecules. The x-ray structures of myoglobin and hemoglobin do not reveal an open path to the binding site at the heme iron (Perutz and Mathews, 1966). Protein motions are essential for ligand escape from and entry to the protein. For the reaction surface in Fig. 3, this fact implies that the outer barrier between B and S fluctuates. Kinetic studies show that these equilibrium fluctuations are strongly affected by the solvent viscosity and composition (Frauenfelder *et al.,* 1991).

Another, functionally important dynamic

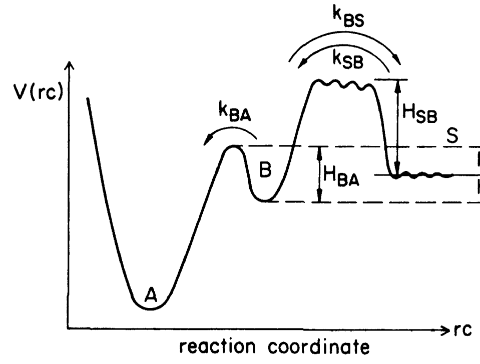

FIG. 3. Reaction energy surface of MbCO with three states, ligand bound (A), ligand in the heme pocket (B), and ligand in the solvent (S).

process affects the inner barrier between B and A. X-ray studies show significant differences in structure between the ligand-bound state A (MbCO) and the deoxy state S (Mb). For example, the iron in the ligand-bound state A is nearly in the plane of the heme group, while it is significantly out of plane by about 0.3 Å in deoxy Mb. Several kinetic and spectroscopic studies indicate that the protein relaxes to the equilibrium Mb structure after ligand dissociation. This relaxation has been described by a nonexponential time dependence of the type described in Sec. 2.4 (Steinbach *et al.*, 1991; Petrich *et al.*, 1991). More recent results suggest that the relaxation involves a sequence of discrete intermediate steps (Nienhaus *et al.*, 1994). The enthalpy barriers for binding from state B to state A increase considerably as the iron moves further out of the plane of the heme. The resultant slowing of the rebinding rate may be crucial for allowing efficient escape of the ligand from the protein. This relaxation of the inner barrier is also dependent on solvent viscosity (Ansari *et al.*, 1992).

2.8 Emerging Concepts

Proteins share essential characteristics with glasses and spin glasses. The dynamics of all these complex systems can be described theoretically with the concept of rugged energy surfaces. Glasses are simpler than proteins, and hence they can guide the formulation of concepts and theories of protein dynamics. The rugged conformational energy surface of a protein forms the scaffold for protein motions. Detailed knowledge of the organization of conformational substates (minima) within the energy surface is crucial for an understanding of protein dynamics and function. Ligand binding studies in myoglobin have yielded insights into the organization of conformational substates in the energy landscape.

2.8.1 Hierarchy of Substates A significant body of evidence implies that protein substates are organized in a hierarchical structure (Frauenfelder *et al.*, 1991). The hierarchical arrangement of substates may reflect the primary, secondary, and tertiary spatial structure described in Sec. 1. The most salient results are summarized in Fig. 4, which gives a schematic diagram of the hierarchical organization of conformational substates (CS) in myoglobin, the most extensively characterized protein from the point of view of its dynamic properties. The figure shows that MbCO assumes a large number of different structures. The top tier (CS0) corresponds to the largest differences in structures among CS. MbCO can assume at least three different conformations that are characterized by different CO stretch frequencies and different enthalpies, entropies, and volumes. Since the CS0 can be characterized individually, they are called *taxonomic substates.* Many proteins show evidence of taxonomic substates. The three CS0 in myoglobin bind CO with different rates, and their relative populations depend markedly on environmental factors such as *p*H, temperature, pressure, and protein–protein interactions. Taxonomic substates may be essential for protein function and control: External factors, for instance, can switch a protein from one taxonomic substate to another and change the reaction rate. This is also the crucial assumption of Koshland's *induced-fit model* of enzymatic function (Stryer, 1988; PROTEINS AND ENZYMES).

Within a given CS0, a protein can assume a large number of substates of tier 1, CS1. This follows from the observation that each individual CS0 of myoglobin rebinds CO, after photodissociation below about 160 K, with kinetics that are highly nonexponential in time. The nonexponential time dependence results from different proteins in the experimental sample becoming stuck in different CS, each with a different rate coefficient for rebinding at low temperature. The substates of tier 1 are *statistical substates* since their properties must be described by appropriate distributions such as $g(H_{BA})$ (see Sec. 2.5.1). The furcation continues: For example, glasslike behavior of the specific heat in proteins is observed below 1 K, implying that motions occur between two-level systems with very low barriers even at this low temperature.

2.8.2 Proteins, Glasses, and Spin Glasses The stretched exponential, Eq. (2), is an empirical relation that provides a consistent framework for analyzing experimental data in protein dynamics. This equation also emerges in experimental studies of glasses and spin glasses. In addition, both glass-

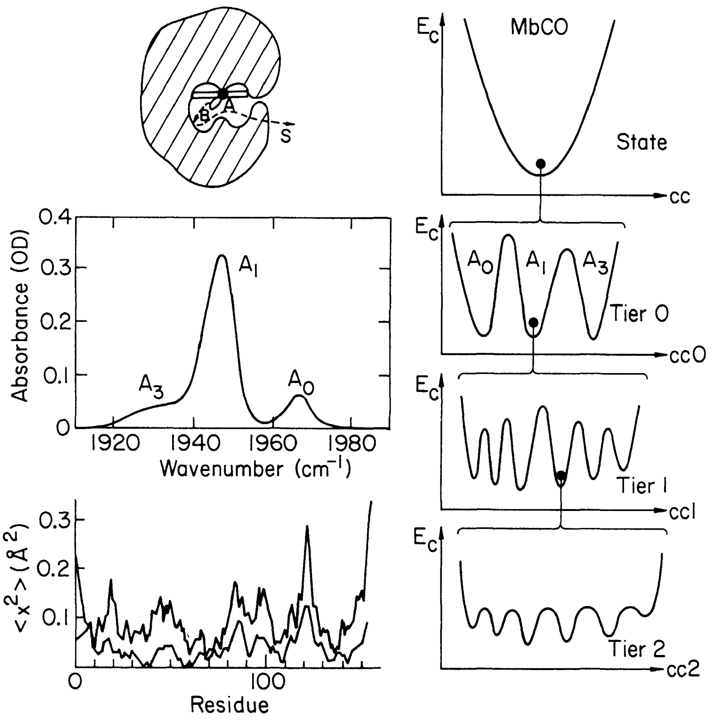

FIG. 4. Structure and conformational energy surface of MbCO. Right column: simplified scheme of the conformational energy, E_c, in three different tiers of the hierarchical energy surface, as a function of a conformational coordinate, *cci*. Note that in reality, E_c is a hypersurface in a many-dimensional conformational space. Left column: Top, coarse structure of MbCO including a possible binding path of the CO ligand from the solvent (S) to the pocket (B) and the iron binding site (A), with an expanded view of the binding geometry. Middle, infrared spectrum of the stretch bands of CO bound to the heme iron; three bands can be distinguished that are assigned to the three taxonomic substates A_0, A_1, and A_3 in the right column of the figure. Bottom, mean squared displacements of amino acids (backbone averages N, C_α, C) of met myoglobin at 300 K (upper) and 80 K (lower) as function of residue number.

forming liquids and proteins experience transitions to long-lived metastable states. This suggests connections among proteins, glasses, and spin glasses. However, an understanding requires theoretical justification of these relations and their connection to the rugged conformational energy surface. The random-energy model (REM) in which the energy minima in the rugged energy surface have uncorrelated free energies has been developed to describe spin glasses and has also been applied to neural nets and molecular evolution (Mezard *et al.*, 1987). The REM has also been successfully applied to protein folding (Frauenfelder *et al.*, 1991). This suggests that complex systems including glasses, spin glasses, neural nets, and proteins share essential properties and that a theory of dynamics on complex energy surfaces may become a reality. However, biomolecules go far beyond physical systems in one crucial aspect: the connection between dynamics and function and between dynamics and evolution. Biomolecules may provide superb systems for development of both theoretical and experimental understanding of complex systems (Frauenfelder and Wolynes, 1994).

3. MOLECULAR DYNAMICS SIMULATIONS

Molecular dynamics (MD) simulations (see SIMULATION BY MOLECULAR DYNAMICS) afford the most detailed theoretical look at the dynamic behavior of proteins (Brooks *et al.*, 1988; McCammon, 1984). Molecular dynamics studies of the atomic motions of proteins and nucleic acids are of relatively recent origin, dating from the first simulation of the atomic motions of a small protein, pancreatic trypsin inhibitor, in 1977 (McCammon *et al.*, 1977). Since then progress has been rapid as a result of the development of faster computers and improved computer algorithms and software. The most important limitation of a molecular dynamics simulation is that it is approximate because of the use of finite-difference techniques and approximate interaction potentials. A second limitation arises from the very large range of time scales involved in protein motions and reactions (Table 1)—relevant times range from $\sim 10^{-14}$ to ~ 1 s. Current simulations can extend to $\sim 10^{-8}$ s, and large-scale parallel processor arrays may extend simulation times to $\sim 10^{-6}$ s, depending on the size of the protein and the number of solvent molecules included. Current MD simulations involve numerical solution of the coupled classical equations of motion for the interacting atoms of a protein. A full treatment of chemical reactions must include a quantum-mechanical approach. However, MD simulations have been useful for determining thermodynamic properties of proteins as well as providing information concerning protein motions themselves. We discuss the essential aspects of MD and some applications to protein dynamics.

3.1 Essentials of Molecular Dynamics

In a molecular dynamics simulation, classical equations of motion for all atoms of the biomolecule are solved by various numerical techniques to obtain the time evolution of the protein structure. Trajectories of proteins have been calculated at constant energy and at constant temperature. Most MD simulations of biomolecular dynamics have been based on potential energy functions that are sums of the types of interactions used in the description of simpler molecular systems. A typical potential function has the form

$$V = \tfrac{1}{2}\sum_{\text{bonds}} K_b(b - b_0)^2 + \tfrac{1}{2}\sum_{\substack{\text{bond}\\\text{angles}}} K_\theta(\theta - \theta_0)^2 + \tfrac{1}{2}\sum_{\substack{\text{dihedral}\\\text{angles}}} K_\phi[1 + \cos(n\phi - \delta)] + \sum_{\substack{\text{nonbonded}\\\text{pairs}}}\left[\frac{A}{r^{12}} - \frac{C}{r^6} + \frac{q_1q_2}{Dr}\right]. \tag{6}$$

The first sum includes a term for every covalent bond in the biomolecule. The second term includes deformation energy for valence angles among covalent bonds for each atom; the third term accounts for intrinsic deformation energy for rotating about an axis through covalently bonded atoms. These three terms represent variations in the covalent binding energy of the protein. The first two terms are treated in the harmonic approximation, which is reasonable at temperatures where fluctuations in the bond lengths and angles are small. The three terms in the sum over nonbonded pairs of atoms separated by a distance r correspond respectively to the core repulsion, dispersion, and electrostatic interactions. The Lennard-Jones 6–12 and the electrostatic potentials in the last sum represent a potential of mean force rather than a simple potential energy if the dielectric coefficient $D \neq 1$ is used to describe any charges that are not explicitly included in the model. Various energy minimization algorithms are used to explore the rugged conformational energy surface of the protein.

3.2 Some Applications of Molecular Dynamics to Myoglobin

The number of applications of molecular dynamics to biomolecules is large and growing (Brooks *et al.*, 1988): MD is used to refine protein structures obtained by x-ray diffraction and NMR, to obtain free energies and other thermodynamic properties of both wild-type and genetically engineered mutant proteins including solvent, to explore the conformational energy surface and map the multiple minima, and to calculate trajectories of protein–ligand and enzyme–substrate reactions. Solvent molecules can be explicitly included in the MD simulations, and the effects of protein hydration can be systematically explored.

The most detailed information may well be available for the oxygen-storage protein myoglobin, which is sometimes called the "hydrogen atom of biology":

1. Molecular dynamics calculations provide the motions of non-hydrogen atoms in myoglobin extending over 300 ps. This dynamic information, together with energy minimization, was used to map the conformational energy surface of Mb (Elber and Karplus, 1987).
2. Molecular dynamics simulations were also used to calculate incoherent neutron-scattering spectra of myoglobin from 80 to 325 K (Smith *et al.*, 1990). These simulations involve time scales (0.3–100 ps) that overlap experimental time scales. The simulated mean square displacements $\langle x^2 \rangle$ for hydrogen atoms are in excellent agreement with the incoherent neutron-scattering data (see Fig. 6 and Sec. 4.3).
3. Trajectories for escape of oxygen from the binding site in Mb have been calculated using molecular dynamics (Case and Karplus, 1979). The trajectories identified a small number of possible pathways for ligand escape from the protein and rein-

forced the importance of protein motions in the escape process. Barriers of ~20 kJ/mol for ligand escape were estimated for the relaxed myoglobin structure.

4. Molecular dynamics simulations of the heme–imidazole environment in Mb subsequent to ligand dissociation have provided insights into the nonexponential time courses for rebinding observed even at room temperature for times less than 300 ps (Petrich *et al.*, 1991).

These simulations have demonstrated the importance of both conformational substates and protein relaxation among the substates in producing nonexponential time courses.

4. EXPERIMENTAL METHODS

Since proteins show an impressively rich spectrum of dynamic phenomena, a particular experimental technique is only sensitive to certain aspects of the motions. To understand the complex dynamics of proteins, different approaches have to be used yielding complementary results. Here we describe important experimental techniques that have been used to probe protein dynamics and to develop and refine the concepts of Sec. 2. Excellent introductions to the various techniques are given by Cantor and Schimmel (1980).

4.1 X-Ray Crystallography

X-ray crystallography is the most important technique to determine the molecular structure of proteins. Since x rays are scattered by electrons in the protein crystal, the method yields the electron density in the molecule, from which atomic coordinates of the non-hydrogen atoms are determined. The periodic arrangement of molecules in a protein crystal gives rise to a periodic pattern of reflections due to constructive interference of the waves at scattering angles that fulfil the Bragg condition (Fig. 5). The structural in-

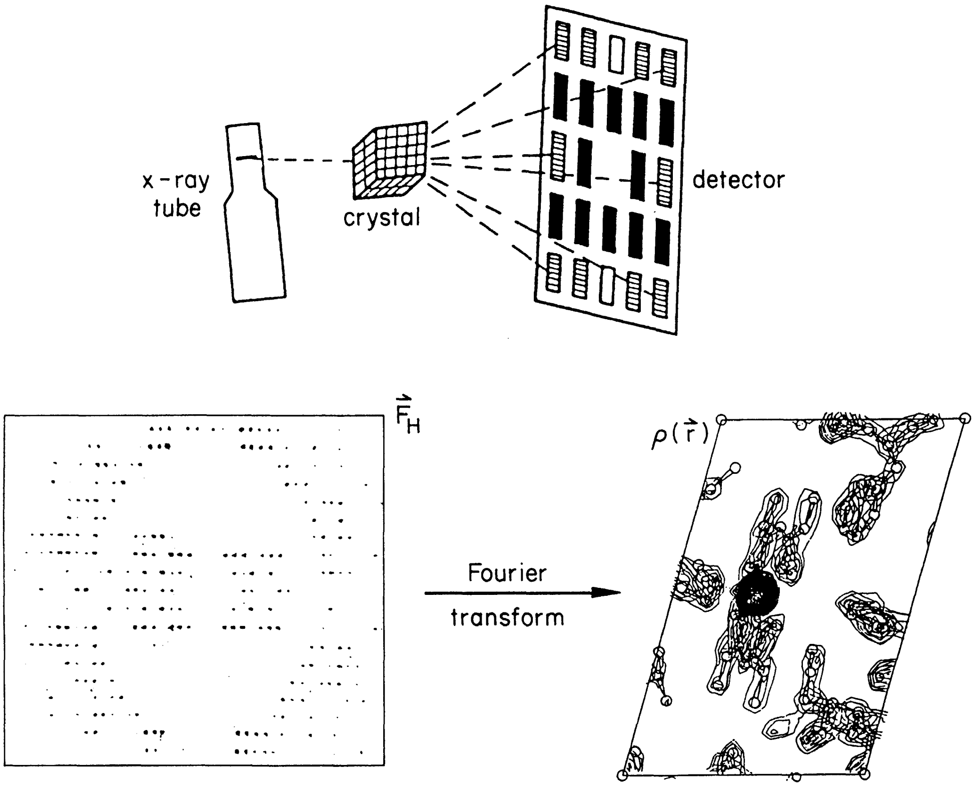

FIG. 5. X-ray structure analysis. The intensities of reflections from x rays scattered by a protein crystal are measured. The electron density $\rho(\mathbf{r})$ is obtained as the Fourier transform of the structure factors $\mathbf{F}_H$.

formation is encoded in the intensities of the reflections, which are the squares of the (complex) structure factors. For a protein structure analysis, the intensities of a large number of reflections (typically 10^4–10^5) are measured.

The electron density is the Fourier transform of the structure factors, and both amplitudes and phases of the structure factors have to be determined. Apart from scaling factors, the amplitude is simply the square root of the intensity. Phases are measured in indirect ways—for example, by additional measurements on crystals where reference scatterers are introduced at well-defined positions in the unit cells (isomorphous replacement).

The presence of conformational substates in proteins implies that the positions of the atoms vary slightly among different molecules in the protein crystal. If we assume an isotropic, Gaussian spread of the atoms around their average positions, the x-ray scattering amplitude of each atom j is decreased by the Debye–Waller factor,

$$f_{\mathrm{DW}} = \exp[-8\pi^2\langle x_j^2\rangle^x(\sin^2\theta)/\lambda^2]. \qquad (7)$$

Here, the mean squared displacement, $\langle x_j^2\rangle^x$, characterizes the disorder of the atom j. We use the superscript "x" to indicate that the mean squared displacement was determined by x-ray crystallography. The scattering angle is denoted by θ, and λ is the wavelength of the x rays. The reflections become weaker with increasing $\langle x_j^2\rangle^x$, and θ.

The preliminary model that is obtained by Fourier transformation is improved by inserting additional information about bond geometry in a least-squares fit procedure called restrained refinement, which improves the agreement between the observed structure factors and those calculated from the model. In high-resolution data sets, $\langle x_j^2\rangle^x$ of each atom is taken as an adjustable fitting parameter in the refinement (Petsko and Ringe, 1984).

Different types of disorder contribute to the mean squared displacement in protein crystals. These contributions are assumed to be uncorrelated, so that the mean squared displacement of an individual atom in the protein, $\langle x^2\rangle^x$, can be written as a sum,

$$\langle x^2\rangle^x = \langle x_s^2\rangle + \langle x_v^2\rangle + \langle x_c^2\rangle. \qquad (8)$$

In protein crystals, the molecules may occupy the unit cells in slightly different ways, which leads to static lattice disorder, represented by $\langle x_s^2\rangle$. The term $\langle x_v^2\rangle$ indicates harmonic vibrations of the lattice, a dynamic disorder contribution to the mean squared displacement. The interesting term in the context of protein dynamics is $\langle x_c^2\rangle$. This term represents "disorder" arising from the inhomogeneous ensemble of protein molecules in different conformational substates.

Energy and time are conjugate variables. Since x-ray crystallography does not measure the energy exchange between the x-ray photons and the sample, the method inherently lacks time resolution; it cannot distinguish between static and dynamic disorder. Information about motions is obtained indirectly (Frauenfelder *et al.*, 1979): Since thermal energy drives fluctuations among the CS, transitions among conformational substates can be slowed by lowering the temperature, until finally, a static distribution of protein molecules is obtained.

As an example, Fig. 6 shows the average mean squared displacement of all non-hydrogen atoms, $\overline{\langle x^2\rangle^x}$, as a function of temperature for myoglobin (Parak *et al.*, 1987). As the temperature is lowered, $\overline{\langle x^2\rangle^x}$ decreases, but a

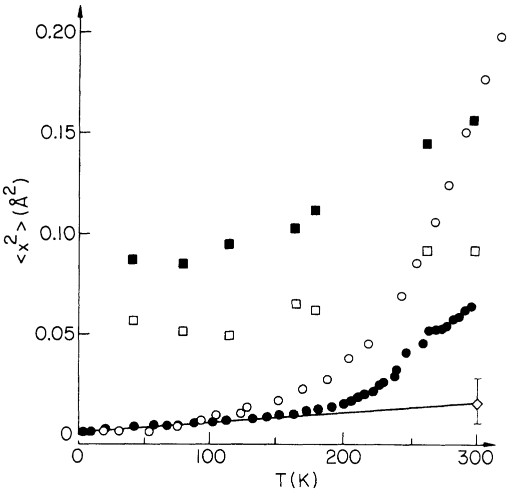

FIG. 6. Mean squared displacements, $\langle x^2\rangle$, measured as a function of temperature. Filled squares, $\overline{\langle x^2\rangle^x}$, average of non-hydrogen atoms (x-ray diffraction); open squares, $\langle x^2\rangle^x_{\mathrm{Fe}}$ of the heme iron (x-ray diffraction); filled circles, $\langle x^2\rangle^\gamma$ of the heme iron (Mössbauer spectroscopy); open circles, $\langle x^2\rangle$, average of hydrogen atoms (incoherent neutron scattering); open diamond, $\langle x^2\rangle^x$ due to lattice vibrations (RSMR).

large amount of disorder prevails at low temperatures. At room temperature, the proteins fluctuate rapidly among the conformational substates. As the temperature is lowered, protein fluctuations get more and more restricted to substates in lower regions of the conformational energy surface, and the spatial distribution narrows. Since there are many with similar conformational energies but slightly different structures, a significant amount of disorder remains even below 100 K. There, the thermal energy is insufficient to overcome the barriers, and the molecules are trapped within a static ensemble of conformational substates.

An isotropic Gaussian distribution is a rather crude model for the disorder. In general, atomic potentials are anharmonic and anisotropic. Recently, algorithms have been developed to use crystallographic data as restraints in molecular dynamics (MD) simulations, such that the MD simulation generates an ensemble of conformational substates, the average of which is restrained to be consistent with the observed crystallographic data (Gros *et al.*, 1990). Then, the disorder of an individual atom is given by the distribution within the ensemble of slightly different structures generated by the MD simulation.

The intensity lost by crystal disorder appears as a thermal diffuse scattering (TDS) background that varies smoothly with scattering angle. Analysis of the TDS background yields information about spatial correlations of the displacements. For example, correlated motions of molecular segments within a protein molecule can be revealed.

X-ray structure analysis used to be a tedious procedure involving measurement times of several weeks. With the advent of powerful x-ray sources and area-sensitive detectors, measurement times have become short enough to follow a chemical reaction—for example, ligand binding—using x-ray crystallography (Rasmussen *et al.*, 1992).

4.2 Mössbauer Spectroscopy

Mössbauer spectroscopy (see MÖSSBAUER EFFECT) involves the absorption of γ rays by atomic nuclei. Normally, a radioactive source is employed that emits γ quanta during a nuclear decay into the ground state. When the nuclei are embedded in a solid matrix, a fraction of the decaying nuclei transfers its entire energy to the γ photons, which leads to an extremely small linewidth of the radiation. These γ quanta can be resonantly absorbed by nuclei of the same kind in the sample under study.

Most often, ^{57}Fe is used, a nucleus that is produced by the decay of ^{57}Co. It emits γ radiation with an energy $E_0 = 14.4$ keV. The energy width Γ of the Mössbauer line is extremely small, 5×10^{-9} eV, reflecting the lifetime $\tau = 141$ ns of the excited nuclear state ($\Gamma\tau = h/2\pi$, where h is Planck's constant).

To measure an absorption spectrum, the γ quanta are counted while the energy of the source line is scanned across the sample resonance. The energy scan is performed using mechanical transducers that move the source relative to the sample, giving rise to a Doppler energy shift $\Delta E = (v/c)E_0$, where v is the source velocity and c is the velocity of light. Mössbauer spectra give the transmission through the sample as a function of source velocity. Since the nuclear levels are sensitive to charge density, magnetic fields, and electric-field gradients in the vicinity of the nucleus (*hyperfine interactions*), the energy levels of the absorber are usually shifted with respect to the source and split into multiple lines. From the Mössbauer spectra, information about the electronic structure at the nucleus can be extracted; it is, therefore, an important technique to characterize electronic properties of iron-containing active sites in proteins. Motions in iron-containing proteins are reflected in the spectra in three different ways:

1. Hyperfine parameters. Motions around the Mössbauer atom change the local fields and, consequently, change the hyperfine parameters, i.e., the positions and splittings of the lines. Information about the motions is usually obtained from the temperature dependence of the hyperfine parameters.
2. Quasielastic lines. Diffusive motions with characteristic times on the order of the coherence time of the radiation (141 ns for ^{57}Fe) or less lead to broad, quasielastic lines that can be understood as a Doppler effect from the motions. The widths of the lines reflect the time scales of the motions. Mössbauer spectroscopy with ^{57}Fe

is sensitive to diffusive motions with characteristic times in the range 10^{-6}–10^{-11} s. Both glass-forming liquids and proteins exhibit wide temporal distributions of dynamic processes, indicating close analogies between a protein and a glass.

3. Temperature-dependent changes of the spectral area. At temperatures near 0 K, all motions except for zero-point vibrations are arrested, and the area of the Mössbauer spectrum is maximal. With increasing temperature, the onset of fast, vibrational motions leads to a decrease of the spectral area. In the Gaussian approximation, the fractional area is related to the mean squared displacement through the Lamb–Mössbauer factor,

$$f_{\mathrm{LM}} = \exp[-k_0^2\langle x^2\rangle^{\gamma}]. \tag{9}$$

Here, k_0 denotes the magnitude of the wave vector, $2\pi/\lambda$, with $\lambda = 0.86$ Å for ^{57}Fe. The mean squared displacement, $\langle x^2\rangle^{\gamma}$, includes only displacements from motions fast compared with the coherence time of the radiation.

As an example, Fig. 7 shows spectra of deoxymyoglobin crystals at 200 and 283 K (Frauenfelder *et al.*, 1988) consisting of quadrupole doublets. The doublet spectra arise from the coupling of an electric-field gradient and the nuclear quadrupole moment. Note the drastically decreased area and the appearance of a broad, quasielastic component in the spectrum at 283 K. The line broadening indicates diffusive motions of the Mössbauer nucleus with time scales around 100 ns. Figure 6 shows the temperature dependence of $\langle x^2\rangle^{\gamma}$ for deoxymyoglobin crystals (Parak *et al.*, 1987). The dynamics below ~200 K are governed by harmonic vibrations, as evidenced by the linear temperature dependence. Above 200 K, $\langle x^2\rangle^{\gamma}$ increases much more steeply, indicating that fluctuations among conformational substates on time scales faster than ~100 ns become thermally excited. By contrast, the individual mean squared displacement of the iron obtained from the x-ray structure analysis, $\langle x^2\rangle^{x}_{\mathrm{Fe}}$, does not show such a pronounced change. This different behavior is a consequence of the fact that the x-ray method cannot distinguish between static and dynamic disorder. In protein crystals, the dynamic disorder at high temperature turns smoothly into static disorder at low temperature.

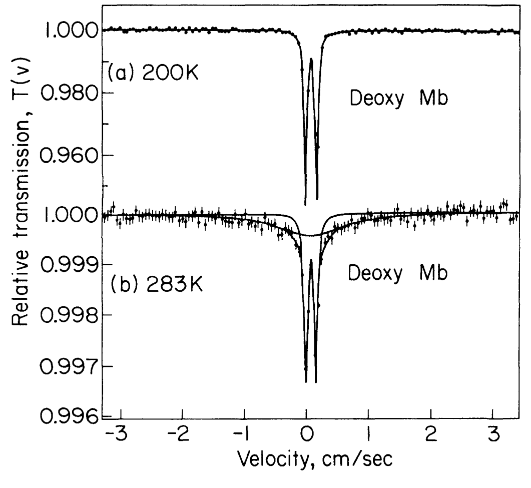

FIG. 7. Mössbauer spectra of deoxymyoglobin crystals at 200 and 283 K. Note different vertical scales for the two curves.

The Mössbauer effect is also used in scattering experiments. Rayleigh scattering of Mössbauer radiation (RSMR) combines the spatial sensitivity of x-ray diffraction experiments with the temporal sensitivity of the Mössbauer technique. It is essentially an x-ray scattering experiment using a Mössbauer γ source instead of an x-ray tube. The sample need not contain Mössbauer nuclides, as the γ quanta are scattered by the electrons in the sample. The exchange of energy in the scattering can be examined by introducing a resonant absorber in the scattered beam. Only γ quanta that have been scattered without (or with very small) energy exchange with the sample can still be absorbed resonantly in the absorber. RSMR is in principle a more powerful technique than Mössbauer absorption spectroscopy, since it allows investigation of spatial correlations of protein fluctuations. For example, RSMR experiments on myoglobin single crystals have shown that lattice vibrations ($\langle x_v^2\rangle$ in Eq. 8) in protein crystals are only a minor contribution to the total disorder at room temperature (Fig. 6) (Nienhaus *et al.*, 1989).

4.3 Neutron Scattering

Neutron scattering is used as a crystallographic technique in a way similar to x-ray diffraction. The neutron wavelength is given by the de Broglie relation, $\lambda_0 = h/mv_0$ (h = Planck's constant, m = neutron mass, v_0 = neutron velocity). Thermal and cold neutrons from a nuclear reactor are useful for crystal structure analysis, since they have wavelengths on the order of atomic distances. Neutrons interact with the atomic nuclei in the crystal, and the strength of the interaction is given by the scattering length b. X-ray diffraction cannot resolve hydrogen atoms, whereas neutron scattering can because of the substantial coherent neutron-scattering length of hydrogen. Furthermore, the large contrast between hydrogen and deuterium in neutron scattering is exploited in studies of hydrogen-exchange processes.

The high energy resolution of neutron scattering makes it possible to analyze the energy exchange between neutrons and target. Inelastic and quasielastic neutron scattering are powerful methods to study motions in proteins on time scales ranging from 10^{-13} to 10^{-7} s (Middendorf, 1984; Doster *et al.*, 1989). The intensity of incoherent scattering is dominated by the hydrogen atoms. Consequently, neutron spectroscopy yields mainly information about the (average) dynamics of the hydrogen atoms. In Fig. 6, the temperature dependences of the mean squared displacements of ^{57}Fe and the hydrogen are similar. The amplitudes are larger for the hydrogens, as expected for lighter atoms; furthermore, the iron is located in the less flexible interior of the protein.

4.4 Nuclear Magnetic Resonance

In recent years, nuclear magnetic resonance (NMR) spectroscopy (see MAGNETIC RESONANCE AND QUADRUPOLE RESONANCE, NUCLEAR) has developed into a powerful tool for investigation of the structure and dynamics of proteins in solution (Wüthrich, 1986). Since nuclei are characterized by a spin angular momentum that can only assume discrete values with respect to a chosen quantization axis, the energy of the nuclei splits into discrete levels in a magnetic field, because of the interaction between the field and the magnetic moment associated with the spin. Most important for NMR on proteins are the nuclei $^1_1$H, $^{13}_6$C, $^{15}_7$N, and $^{31}_{15}$P, which all have spin quantum number $\frac{1}{2}$, giving rise to two energy levels in a magnetic field. Transitions between the levels are induced by irradiation with radio-frequency (rf) photons with energies matching the energy difference among the nuclear levels. Since the lower energy level is more populated than the upper level in thermal equilibrium, resonant absorption of energy occurs. The local magnetic field at the nucleus differs from the external field because the surrounding electrons induce an additional field. Since each nucleus has a slightly different environment, the individual resonances occur at different frequencies; this effect is called *chemical shift*. The large number of protons in a protein molecule leads to complex, crowded spectra with many overlapping resonances. For the determination of protein structures, data are collected using two-dimensional (2D) and multidimensional NMR techniques involving special rf pulse sequences that disperse the resonances and identify neighboring nuclei through measurement of *spin–spin interactions.* Correlated spectroscopy (COSY) experiments probe the connectivity of nuclear spins by through-bond interactions. Nuclear Overhauser enhancement spectroscopy (NOESY) establishes the proximity of nuclear spins by measuring through-space dipolar interactions. The three-dimensional structure is obtained by sequential assignment of resonances along the polypeptide backbone.

Protein motions affect the spin interactions by changing the magnitude and direction of the local magnetic field with respect to the external field. In solution, rapid overall reorientation of the protein molecules averages the nuclear spin interactions to their isotropic values, leading to sharp resonances, which is essential for high-resolution NMR structure determination. Large biomolecules can be too immobile for solution NMR, or crystals and amorphous samples are to be investigated. For these applications, high-resolution solid-state NMR techniques are available, where rf irradiations and mechanical sample spinning replace molecular motions as line-narrowing mechanisms. Dynamic information is obtained from NMR in three different time ranges:

1. Slow processes ($\tau \gtrsim 10$ min) are monitored by successive acquisition of NMR spectra during the time course of the reaction, for example, in hydrogen exchange studies.

2. Processes on an intermediate time scale ($\tau \approx 1\text{–}10^{-5}$ s) are reflected in the particular shape of the spectrum, and line-shape analysis reveals the time scales of the nuclear motions. Consider a nucleus that can assume two distinct positions in the molecule. The different environments give rise to two peaks at frequencies ν_A and ν_B. Two sharp resonance lines are observed as long as the residence time τ in either of the positions is longer than the inverse of the frequency separation of the two lines, $\tau \gg \Delta\nu^{-1}$. Jump processes are usually thermally activated, so that with increasing temperature, the jumps become more frequent. Concomitantly, the lines broaden and the separation decreases, until they finally merge into a single line, as sketched in Fig. 8.

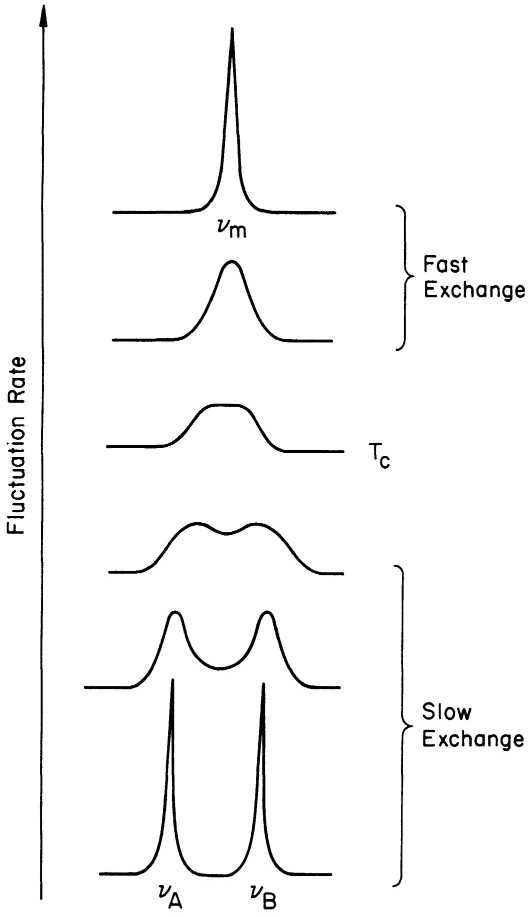

FIG. 8. NMR spectrum of a nucleus with two distinguishable positions, as a function of fluctuation rate.

3. Fast processes ($\tau \lesssim 10^{-9}$ s) are studied with relaxation-time experiments—that is, spin–lattice relaxation (T_1), spin–spin relaxation (T_2), and cross-relaxation rates in nuclear Overhauser enhancements.

4.5 Hydrogen Exchange

Polar side-chain protons and main-chain peptide-group protons exchange with protons from the solvent. Hydrogen exchange (Englander and Kallenbach, 1984) requires transient breaks of hydrogen bonds and accessibility of the sites to the solvent. Consequently, the method is sensitive to conformational fluctuations of protein molecules. The experimentally observed exchange rates span many orders of magnitude. Hydrogen-exchange experiments reveal that the protected, hydrogen-bonded protons are exchanged by way of local unfolding of small structural segments as cooperative units. Three different experimental approaches are used to monitor the exchange:

1. NMR. The protein is dissolved in heavy water (D_2O). The amplitudes of the proton resonances are proportional to the concentration of ^1H at a particular site. Exchange rates of well-separated resonances can be determined by measuring spectra at different times after transfer into D_2O, since hydrogen resonances disappear as the protons are exchanged for deuterons. Fast exchanges can be studied through line-shape analysis.
2. Neutron diffraction. Protein crystals are soaked in D_2O. The scattering lengths of ^1H and ^2H are distinctly different, and, therefore, deuterons and hydrogens are distinguishable in the density map.
3. Tritium exchange. The protein sample is incubated with tritiated water for a certain amount of time that may be varied from minutes to days to select a population of hydrogens within a certain range of exchange rates (exchange-in phase). Then, the tritiated solvent is separated from the protein using biochemical procedures. Tritium (^3H) is a radioactive isotope of hydrogen, and the number of tritium decays immediately after removal of the tritiated solvent is proportional to the fraction of sites where hydrogen has been replaced by tritium. The tritium atoms

will subsequently exchange with the solvent (exchange-out phase). A second exchange of the solvent is performed after a certain delay time, terminating the exchange-out phase. Nuclear decays after the second exchange are due to tritium atoms that have not exchanged out during the delay time, and hence the decrease in radioactivity measures the fraction of protons that have escaped from the protein into the solvent during the delay period. By varying the length of the exchange-out phase, the distribution of rates is obtained.

4.6 Fluorescence Spectroscopy

Fluorescence spectroscopy (Lakowicz, 1983) is based upon the emission of light by fluorescent molecules, which typically decay into the ground state 1–10 ns after electronic excitation. Both the intrinsic fluorescence (aromatic side chains, prosthetic groups) and the fluorescence emitted by dyes attached to protein molecules are extremely sensitive to a wide range of interactions and perturbations during the lifetime of the excited state that influence the fluorescence emission, including proton-transfer reactions, ligand binding, solvent-cage relaxation, local conformational transitions, and diffusional processes. The method is highly sensitive so that only minute amounts of material are needed.

Rotational diffusion and segmental motions are studied by measurement of fluorescence anisotropies. The sample is illuminated with linearly polarized light, and only a minor fraction of the sample, with transition dipoles for absorption preferentially parallel to the polarization vector, is excited. This principle is called *photoselection.* Since emission and absorption transition dipoles are sterically coupled, photoselection introduces an anisotropy in the fluorescent emission that is averaged out by orientational motions of the fluorophore. Measurements of the anisotropy decay yield information about the time scales of the motions.

Collision of fluorescent molecules with oxygen (and heavy atoms like iodide or cesium ions) leads to fluorescence quenching. This process requires diffusion of the quencher to the fluorescent group within the lifetime of the excited state. When the fluorophores are on the exterior of the protein, studies of collisional quenching allow determination of the rates with which small molecules diffuse to the protein. When the fluorophore is buried in the interior, permeation of the protein by the quencher can be studied.

Fluorophores show time-dependent spectral shifts, since polar groups reorient themselves after excitation in response to the changed dipole moment. Time-resolved measurements of the fluorescence emission spectrum reveal the time scales on which the protein reacts to the change of the electric field after excitation.

Time-dependent fluorescence measurements are performed most directly in the time domain with pulse response measurements. In recent years, frequency-domain experiments have become very popular. In these measurements, the sample is continuously illuminated by a modulated excitation beam, and both phases and modulation amplitudes of the resulting fluorescence emission are registered as functions of the excitation frequency (Gratton *et al.,* 1984).

4.7 Transient Optical Spectroscopies

The availability of pulsed lasers and fast optical detectors makes transient optical spectroscopy a powerful tool for the study of protein dynamics over an enormous time range (10^{-14}–10^{6} s). External perturbations are applied that force the system out of equilibrium, and the subsequent relaxation toward equilibrium can be monitored by measuring transient optical spectra (Eigen and de Mayer, 1974). A nonequilibrium situation can be created in various ways—for instance, by a temperature jump (irradiation with microwaves or infrared laser pulses, and by current pulses), pressure jump, electric-field jump, pulse radiolysis (irradiation with high-energy radiation), or photoexcitation (flash photolysis).

Photoexcitation is a frequently used relaxation technique. Intense monochromatic laser pulses are ideal for triggering a variety of different reactions in proteins. Often, the protein relaxes back to the initial state within a short time so that repetitive excitations can be performed at high frequency (signal averaging). Many systems are damage resistant and do not change their kinetic properties even after thousands of flashes. Figure 9 shows a classical example of photoexcitation, the photocycle of bacteriorhodopsin, a molecule that utilizes light energy to pump protons across the cell membrane (Lozier *et al.,* 1975; Mathies *et al.,* 1991). When chemical bonds are broken by the laser pulse, the photoexcitation technique is referred to as flash photolysis.

A number of spectroscopic techniques are used to monitor the spectral changes after photoexcitation. Different spectral regions give complementary information. Transient spectra are observed in different ways, dependent upon the time scales: Slow processes ($\tau \gtrsim 10$ s) are observed with conventional dispersive or Fourier-transform spectrometers. On intermediate time scales ($\tau \gtrsim 10^{-9}$ s), transient changes at single wavelengths can be monitored continuously as a function of time after the laser flash. To obtain experimental data on entire spectra, gated optical multichannel analyzers (OMA) are used. The time dependence is obtained by opening the detector gate for a short time at various time delays after the laser pulse. On fast time scales ($\tau \lesssim 10^{-9}$ s), pulse/probe techniques are employed: The laser pulse is split, and one part is used for photoexcitation, whereas the other part, shifted in time by an optical delay line, is used to probe the transient spectrum.

Protein spectra are generally inhomogeneously broadened as a result of the presence of conformational substates. The inhomogeneity can be shown using selective photoexcitation of subpopulations by illumination at different positions across the band (*spectral hole burning*). If the band position of a homogeneous subpopulation is coupled in a systematic way to a kinetic variable—for example, the enthalpy barrier for ligand rebinding—band shifts are observed as the reaction proceeds (*kinetic hole burning*) (Campbell *et al.,* 1987, Ormos *et al.,* 1990).

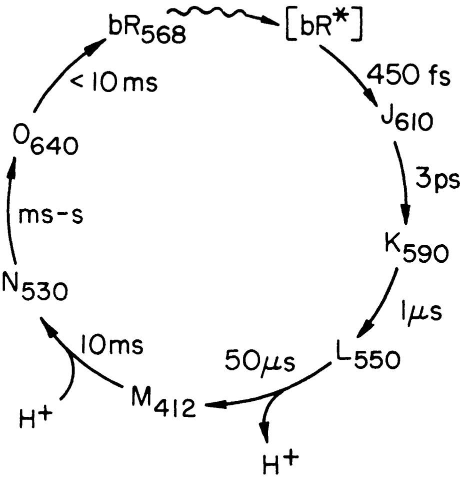

FIG. 9. Photocycle of bacteriorhodopsin, showing the predominant intermediates and the transition rates. Subscripts are absorption maxima in nanometers.

Linear dichroism (LD) experiments determine the absorption as a function of the polarization of the monitoring beam. LD experiments in combination with photoselection (see Sec. 4.6) monitor reorientation of the absorbing chromophores. *Circular dichroism* (CD) experiments are based on the fact that biological molecules show optical activity and absorb left-circularly and right-circularly polarized light differently.

4.7.1 Ultraviolet Spectroscopy Ultraviolet (UV) spectroscopy is based upon electronic transitions induced by the absorption of UV light. The peptide unit dominates the spectrum around 200 nm. The spectrum changes in a characteristic way with the type of secondary structure. From 230 to 300 nm, the aromatic sidechains of phenylalanine, tyrosine, and tryptophan and the histidine and disulfide sidechains absorb. The absorption of prosthetic groups may also fall into the UV region. UV absorption and CD spectra depend upon the folding pattern and are used for structural and kinetic studies, for instance, of helix-coil transitions.

4.7.2 Visible/Near-Infrared Spectroscopy Many biological molecules show strong absorbance in the visible (VIS) and near-infrared (NIR) regions of the spectrum, due to the presence of metal ions and prosthetic groups with extended π-electron systems, like chlorophyll, heme, flavin, and retinal. The bands are sensitive to the polypeptide environment and reflect structural changes, oxidation states, and the binding of ligands, making transient visible spectroscopy an excellent technique for physical studies of enzyme action. As an example, the visible spectra of heme change markedly upon photodissociation of a ligand like oxygen or carbon monoxide. The absorption bands in the visible are superb markers for the binding of ligands to heme proteins (Fig. 10).

4.7.3 Infrared Spectroscopy In protein molecules, absorption in the infrared (IR) region of the electromagnetic spectrum arises from transitions between vibrational levels due to bond stretching and bending. IR vibrations occur between 5 000 cm^{-1} (2 μm) and 200 cm^{-1} (50 μm). The contribution from a particular bond is usually weak, so that concentrated samples are needed. Water absorbs strongly in the region where protein bands are found, but the problem can be alleviated by the use of D_2O instead of water. IR spectra are superpositions of a large number of overlapping bands. Specific information is obtained by IR difference spectroscopy, where spectra are taken before and after photoexcitation. The absorbance difference spectrum contains only lines from bond changes resulting from the photoprocess. Often, only a few lines are observed, which can be assigned to particular bonds by site-directed mutagenesis or isotope labeling techniques.

Fourier-transform infrared (FTIR) spectrometers are routinely used for infrared studies of protein dynamics. Collecting FTIR spectra takes on the order of seconds to minutes, so that only slow processes, like ligand binding at cryogenic temperatures, can be studied. For samples that allow repetitive photoexcitation, step-scan and stroboscopic FTIR techniques are used with time resolution in the microsecond region. Technology for direct IR detection on the picosecond time is not yet available. To probe on fast time scales, the IR beam is mixed with visible light in a nonlinear crystal, and the intensity at the sum frequency in the visible is monitored.

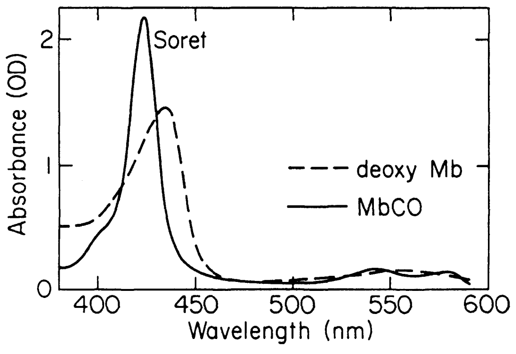

FIG. 10. Absorption spectra of carbonmonoxymyoglobin (MbCO) and deoxymyoglobin (Mb) in the visible. The Soret band on the blue side of the spectrum is the strongest electronic transition in heme proteins.

4.7.4 Raman Scattering For Raman experiments, the sample is illuminated with monochromatic light. Besides the elastic (Rayleigh) line, the scattered light contains sidebands (Stokes and anti-Stokes lines) resulting from energy exchange between photons and sample through vibrational transitions. Therefore, the method is closely related to IR absorption. Instead of electric dipole changes, Raman bands arise from polarizability changes due to nuclear motions. For small, centrosymmetric molecules, IR and Raman activities are mutually exclusive; hence IR absorption and Raman spectroscopy complement each other. In large, asymmetric molecules, the same vibrational bands are often seen with both techniques. An advantage of Raman spectroscopy for biological samples is that water scatters weakly and does not obscure much of the spectral range, as it does with IR absorption. Raman bands are usually weak and difficult to study, unless they are coupled to electronic transitions: By exciting at an electronic absorption wavelength, an enormous signal intensity enhancement is observed. This technique, called *resonance Raman scattering,* can be used for the study of biological samples in dilute solutions. Since resonance enhancement is limited to modes coupled to the chromophore, the spectrum is greatly simplified. Most often, the chromophore is located at the active site, so that resonance Raman spectroscopy focuses in on the most interesting region.

4.7.5 Other Experimental Techniques The techniques that are discussed above are the most frequently used tools for experimental studies of protein dynamics. Other popular techniques include EXAFS, electron microscopy, electron paramagnetic resonance, photoacoustic spectroscopy, differential scanning calorimetry (DSC), and ultracentrifugation (and other hydrodynamic techniques). For a general introduction to these techniques, see Cantor and Schimmel (1980) and Hoppe *et al.* (1983).

5. PROTEIN DYNAMICS—A REPRISE

Proteins are dynamic systems because of their structure and the various competing interactions among their constituent atoms. On the basis of studies of deuterium/hydrogen exchange in proteins, Linderstrøm-Lang and Schellman held the view as early as 1959 that a protein did not possess a unique structure but instead "exists mainly as a group of structures not too different from one another in free energy, but frequently differing considerably in energy and entropy" (Linderstrøm-Lang and Schellman, 1959). In 1966, Perutz and Mathews pointed out that the x-ray structures of hemoglobin and myoglobin do not allow access of the ligand to the binding site, and hence they concluded that the proteins must have flexible elements that permit a dioxygen molecule to reach the binding site at the heme iron (Perutz and Mathews, 1966). Klotz (1966) and Weber (1972) suggested that a protein with a given primary sequence exists in many conformational states and is dynamic. Flash photolysis studies of small-ligand binding to myoglobin established the functional relevance of the manifold of protein conformational substates (Frauenfelder *et al.*, 1988). In particular, it was shown that an ensemble of proteins below about 160 K is described by an inhomogeneously broadened distribution of rate coefficients for ligand rebinding after photodissociation. The spatial disorder in proteins due to the manifold of conformational substates was demonstrated with x-ray diffraction (Frauenfelder *et al.*, 1979).

In spite of significant progress in the last few decades, understanding the connections among structure, dynamics, and function continues to be a major challenge in protein science. The enormous development of computer technology has made possible the study of protein motions on subnanosecond time scales by computer simulations. However, many biomolecular processes occur on much longer time scales, and other theoretical approaches are necessary. Statistical models, borrowed from the physics of disordered materials, have led to exciting new insights. For the development of theories, experimental studies of the conformational energy landscape have played a guiding role, and they will continue to do so. The arsenal of experimental methods that are available for the study of proteins is impressive. However, because of the enormous complexity of biomolecules, every technique reveals only a projection of the many dimensions of the problem. Combining the results of many different experiments, performed on selected model systems, may finally lead to fundamental theories of the dynamics of proteins.

ACKNOWLEDGMENT

This work was supported by grant GM 18051 from the National Institutes of Health.

GLOSSARY

In this Glossary, definitions of crucial concepts in the field of protein dynamics are listed. Readers are referred to the article Proteins and Enzymes, preceding this article, for definitions of terms and concepts from biochemistry.

Complex System: A system, such as a glass or protein, with a rugged conformational energy landscape with many relative energy minima (conformational substates) separated by relative energy maxima.

Conformation, Conformational State: The three-dimensional structure of an equilibrium state of a protein as determined, for example, by x-ray diffraction.

Conformational Energy: The potential energy of a conformational state of a protein resulting from interactions among all the constituent atoms of the molecule.

Conformational Energy Landscape: A rugged conformational energy surface of a complex system, such as a protein or glass, obtained by plotting the conformational energy as a function of all relevant spatial or conformational coordinates of the molecule.

Conformational Substate: One of a manifold of states of a protein molecule having the same overall structure (i.e., conformational state) and performing the same biological function, but with different structural details and possibly other properties such as the rate at which the function is performed or absorption spectra. A conformational substate refers to a local minimum in the conformational energy landscape of the protein.

Conformational Transition: A process in which the conformational state or substate of a protein changes.

Distribution of Activation Barriers, $g(H_{BA})$: A continuous distribution of activation barriers resulting from a large manifold of conformational substates with slightly different barriers for the biochemical transition, $B \rightarrow A$.

Equilibrium Fluctuation: A conformational transition between two different conformational states or substates of a protein resulting from interactions with the thermal bath.

Ergodic: Refers to a protein experiencing equilibrium fluctuations among all conformational substates of a given conformational state on the time scale of an experimental observation.

Molecular Dynamics: A computational theory in which the coupled classical equations of motion for all atoms of a protein are solved by any one of several algorithms using high-speed digital computers. Molecular dynamics also includes algorithms for minimizing the potential energy of a protein.

Nonequilibrium Relaxation: A conformational transition between two different conformational states or substates of a protein occurring after a perturbation that leaves the protein in a nonequilibrium state.

Nonergodic: Refers to a protein molecule being trapped in a region of the conformational energy landscape on the time scale of an experimental observation.

Relaxation Function, $\Phi(t,T)$: A function that describes the relaxation of an experimental observable of a system from an initial nonequilibrium state to an equilibrium state. The relaxation function is defined to be 1 at time $t = 0$ and to be 0 at time $t = \infty$.

Statistical Substates: Substates of a protein with properties, such as reaction rates and absorption spectra, that must be described by distributions resulting from the very large number of substates.

Taxonomic Substates: A relatively small number of substates of a protein with properties that can be easily classified, including thermodynamic parameters, reaction rates, and absorption spectra. Distributions need not be used to describe taxonomic substates.

Works Cited

Ausari, A., Jones, C. M., Henry, E. R., Hofrichter, J., Eaton, W. A. (1992), *Science* **256,** 1796–1798.

Beece, D., Eisenstein, L., Frauenfelder, H., Good, D., Marden, M. C., Reinisch, L., Reynolds, A. H., Sorensen, L. B., Yue, K. T. (1980), *Biochemistry* **19,** 5147–5157.

Brooks, C. L., Karplus, M., Pettitt, B. M. (1988), *Proteins: A Theoretical Perspective of Dynamics, Structure, and Function,* New York: Wiley Interscience.

Campbell, B. F., Chance, M. R., Friedman, J. M. (1987), *Science* **238,** 373–376.

Cantor, C. R., Schimmel, P. R. (1980), *Biophysical Chemistry,* San Francisco: Freeman.

Case, D., Karplus, M. (1979), *J. Mol. Biol.* **132,** 343.

Chance B., DeVault, D. C., Frauenfelder, H., Marcus, R. A., Schrieffer, J. R., Sutin, N. (1979), *Tunneling in Biological Systems,* New York: Academic Press.

Cooper, A. (1976), *Proc. Natl. Acad. Sci. U.S.A.* **73,** 2740–2741.

Doster, W., Cusack, S., Petry, W. (1989), *Nature* **337,** 754–756.

Eigen, M., Mayer, L. de (1974), in: A. Weissberger (Ed.), *Techniques of Organic Chemistry,* Vol. 6, Pt. II, New York: Wiley.

Elber, R., Karplus, M. (1987), *Science* **235,** 318.

Ehrenstein, D., Nienhaus, G. U. (1992), *Proc. Natl. Acad. Sci. U.S.A.* **89,** 9681.

Englander, S. W., Kallenbach, N. R. (1984), *Q. Rev. Biophys.* **16,** 521–655.

Frauenfelder, H., Petsko, G. A., Tsernoglou, D. (1979), *Nature* **280,** 558–563.

Frauenfelder, H., Wolynes, P. G. (1985), *Science* **229,** 337–345.

Frauenfelder, H., Parak, F., Young, R. D. (1988), *Annu. Rev. Biophys. Biophys. Chem.* **17,** 451–479.

Frauenfelder, H., Sligar, S. G., Wolynes, P. G. (1991), *Science* **254,** 1598–1603.

Frauenfelder, H., Wolynes, P. G. (1994), *Phys. Today* **47** (2), 58–64.

Gratton, E., Jameson, D. M., Hall, R. D. (1984), *Annu. Rev. Biophys. Bioeng.* **13,** 105–124.

Gros, P., Gunsteren, W. F. van, Hol, W. G. J. (1990), *Science* **249,** 1149–1152.

Hoppe, W., Lohmann, W., Markl, H., Ziegler, H. (Eds.) (1983), *Biophysics,* Berlin: Springer.

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., Shore, V. C. (1960), *Nature* **185,** 422.

Klotz, I. M. (1966), *Arch. Biochem. Biophys.* **116,** 92.

Lakowicz, J. R. (1983), *Principles of Fluorescence Spectroscopy,* New York: Plenum.

Landau, L. D., Lifshitz, E. M. (1980), *Statistical Physics,* London-Paris: Pergamon.

Linderstrøm-Lang, K. U., Schellman, J. A. (1959), *Enzymes* **1,** 443.

Lozier, R. H., Bogomolni, R. A., Stoeckenius, W. (1975), *Biophys. J.* **15,** 955.

Mathies, R. A., Lin, S. W., Ames, J. B., Pollard, W. A. (1991), *Annu. Rev. Biophys. Biophys. Chem.* **20,** 491–518.

McCammon, J. A., Gelin, B. R., Karplus, M. (1977), *Nature* **267,** 585.

McCammon, J. A. (1984), *Rep. Prog. Phys.* **47,** 1–46.

Mezard, M., Parisi, G., Virasoro, M. A. (1987), *Spin Glass Theory and Beyond,* Teaneck, NJ: World Scientific.

Middendorf, H. D. (1984), *Annu. Rev. Biophys. Bioeng.* **13,** 425–451.

Nienhaus, G. U., Heinzl, J., Huenges, E., Parak, F. (1989), *Nature* **338,** 665–666.

Nienhaus, G. U., Chu, K., Mourant, J. R., Frauenfelder, H. (1994), *Biochemistry* **33,** 13413–13430.

Ormos, P., Ansari, A., Braunstein, D., Cowen, B. R., Frauenfelder, H., Hong, M. K., Iben, I. E. T., Sauke, T. B., Steinbach, P. J., Young, R. D. (1990), *Biophys. J.* **57,** 191–199.

Parak, F., Hartmann, H., Nienhaus, G. U. (1987), in: R. Austin, E. Buhks, B. Chance, D. De Vault, P. L. Dutton, H. Frauenfelder, V. I. Goldanskii (Eds.), *Protein Structure: Molecular and Electronic Reactivity,* New York: Springer, pp. 65–84.

Palmer, R. G. (1982), *Adv. Phys.* **31,** 669.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., North, A. C. T. (1960), *Nature* **185,** 416.

Perutz, M. F., Mathews, F. S. (1966), *J. Mol. Biol.* **21,** 199.

Petrich, J. W., Lambry, J.-C., Kuczera, K., Karplus, M., Poyart, C., Martin, J.-L. (1991), *Biochemistry* **30,** 3975–3987.

Petsko, G. A., Ringe, D. (1984), *Annu. Rev. Biophys. Bioeng.* **13,** 331–371.

Rasmussen, B. F., Stock, A. M., Ringe, D., Petsko, G. A. (1992), *Nature* **357,** 423–424.

Smith, J., Kuczera, K., Karplus, M. (1990), *Proc. Natl. Acad. Sci. U.S.A.* **87,** 1601.

Steinbach, P. J., Ansari, A., Berendzen, J., Braunstein, D., Chu, K., Cowen, B. R., Ehrenstein, D., Frauenfelder, H., Johnson, J. B., Lamb, D. C., Luck, S., Mourant, J. R., Nienhaus, G. U., Ormos, P., Philipp, R., Xie, A., Young, R. D. (1991), *Biochemistry* **30,** 3988.

Stryer, L. (1988), *Biochemistry,* New York: Freeman.

Weber, G. (1972), *Biochemistry* **11,** 864.

Wüthrich, K. (1986), *NMR of Proteins and Nucleic Acids,* New York: Wiley.

Further Reading

Austin, R., Buhks, E., Chance, B., De Vault, D., Dutton, P. L., Frauenfelder, H., Goldanskii, V. I. (Eds.) (1987), *Protein Structure: Molecular and Electronic Reactivity,* New York: Springer.

Blundell, T. L., Johnson, L. N. (1976), *Protein Crystallography,* New York: Academic Press.

Campbell, I. D., Dwek, R. A. (1984), *Biological Spectroscopy,* New York: Addison-Wesley.

Cantor, C. R., Schimmel, P. R. (1980), *Biophysical Chemistry,* San Francisco: Freeman.

Eaton, W. A., Szabo, A. (Eds.) (1991), *Chem. Phys.* **158,** 191–531.

Fleming, G. R., Hañggi, P. (1993), *Activated Barrier Crossing,* Singapore: World Scientific.

Frauenfelder, H., Marden, M. C. (1981), "Biological Physics," in: H. L. Anderson (Ed.), *Physics Vade Mecum,* New York: American Institute of Physics.

Hoppe, W., Lohmann, W., Markl, H., Ziegler, H. (Eds.) (1983), *Biophysics,* Berlin: Springer.

McCammon, J. A., Harvey, S. C. (1987), *Dynamics of Proteins and Nucleic Acids,* Cambridge: Cambridge University Press.

Rousseau, D. L. (Ed.) (1984), *Optical Techniques in Biological Research,* New York: Academic Press.

Schulz, G. E., Schirmer, R. H. (1979), *Principles of Protein Structure,* New York: Springer.

Spiro, T. G. (Ed.) (1988), *Biological Applications of Raman Spectroscopy,* New York: Wiley.

Stein, D. L. (Ed.) (1992), *Spin Glasses and Biology,* Singapore: World Scientific.

PROTEINS AND ENZYMES

Hans Bisswanger, *Physiologisch-chemisches Institut der Universität Tübingen, Tübingen, Germany*

Introduction 185
1. **Protein Structure** 186
1.1 Primary Structure 186
1.2 Secondary Structure 188
1.3 Tertiary Structure 188
1.4 Quaternary Structure 189
1.5 Cellular distribution of Proteins 189
1.6 Posttranslational Modification of Proteins 189
2. **Protein Functions** 190
2.1 Structural and Contractile Functions 190
2.2 Binding, Transport, and Storage 190
2.3 Infection Defense and Blood Clotting 190
2.4 Membrane Proteins and Membrane Transport 191
3. **Enzymes** 192
3.1 Enzyme Catalysis 192
3.2 Catalytic Center 193
3.3 Mechanisms of Enzyme Catalysis 193
3.4 *p*H Dependence of Enzyme Reactions 194
3.5 Temperature Dependence and Arrhenius Theory 194
3.6 Enzyme Regulation 195
3.7 Cellular Organization of Enzymes 196
4. **Kinetics of Enzyme Reactions** 197
4.1 Michaelis–Menten Equation and Steady-State Relationship 197
4.2 Graphical Representation of the Michaelis–Menten Equation 197
4.3 Enzyme Inhibition 199
4.4 Multisubstrate Reactions 200
4.5 Allosteric Enzymes 201
5. **Isolation and Purification of Proteins** 202
5.1 Cell Disintegration and Fractionation 202
5.2 Precipitation and Concentration 202
5.3 Chromatographic Techniques 203
6. **Techniques for Analysis and Structure Determination of Proteins and Enzymes** 204
6.1 Electrophoretic Techniques 204
6.2 Protein Determination and Enzyme Tests 204
6.3 Amino Acid Analysis and Protein Sequencing 205
6.4 Determination of the Three-Dimensional Structure 205
6.5 Molecular Mass Determination 206
6.6 Measurement of Fast Reactions 208
6.7 Peptide and Protein Synthesis 209
7. **Technical Applications of Proteins and Enzymes** 209
7.1 Immobilization of Proteins and Enzymes 209
7.2 Enzyme Reactors 210
7.3 Biosensors 210
7.4 Medical Applications 211
7.5 Protein Engineering and Artificial Enzymes 211
Glossary 211
Works Cited 212
Further Reading 213

INTRODUCTION

Proteins belong to the most important class of compounds of the living world. The first proteins, fibrin from blood and casein from milk, were isolated by the Swedish chemist J. J. Berzelius in 1813. The word *protein* (from the Greek word *proteton,* "precedence," "superiority") was introduced by G. J. Mulder in 1838. He found that proteins

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

are the main components of the living matter and also that they are similar in their composition. The catalytic activity of a protein was first demonstrated by A. Payen and J. F. Persoz with the example of the liquefication of starch by amylase (diastase) from malt extract in 1833. For catalytic proteins, W. F. Kühne created in 1876 the word *enzyme* (from the Greek phrase *en zyme*, "in yeast"). In 1897, H. E. A. Buchner discovered the first metabolic pathway in a cell-free system, alcoholic fermentation in yeast juice. J. B. Sumner crystallized in 1926 the first enzyme, urease, and after crystallization of several digestive enzymes by J. H. Northrop, the protein nature of enzymes became generally accepted. In 1984, however, S. Altman and in 1986, T. Cech found ribonucleic acids with catalytic activities (ribozymes). The first amino acid sequence of a protein, that of insulin, was unraveled in 1955 by F. Sanger, and in 1963, W. H. Moore and S. Stein elucidated the first amino acid sequence of an enzyme, ribonuclease. Six years later, R. B. Merrifield reported the total synthesis of this enzyme by the solid-phase peptide synthesis. The first three-dimensional structures of protein molecules, myoglobin and hemoglobin, were determined by J. C. Kendrew and M. F. Perutz with the x-ray diffraction technique in 1960 (Florkin and Stotz, 1972).

Up to now, an overwhelming number of different proteins have been isolated and characterized from all types of organisms from bacteria and viruses to man. Proteins are essentially involved in every process occurring in living organisms, like reproduction of the genetic material; fertilization; development of cells and of whole organisms; maintenance of structures; selection, uptake, and metabolism of nutrients; energy generation; photosynthesis; muscle contraction; sensory functions like vision; nerve conduction, and brain function. Within a short review, it is impossible to give more than a superficial summary on proteins. At first, the structure of proteins is discussed, followed by a survey on some important functions. A broader space is given to the most significant proteins, the enzymes. The ability to accelerate chemical reactions by several orders of magnitude is one of the most fascinating inventions of nature and a fundamental prerequisite for life. As chemical catalysts enzymes obey the pertinent rules (cf. CHEMICAL KINETICS), but being proteins, they possess various possibilities to adapt their catalytical capacity to the special requirements of the cell and the organism, and they follow their own rules of enzyme kinetics.

The methods for isolation, characterization, and structure determination are similar both for proteins in general and for enzymes. Enzymes, however, possess the advantage of being detectable by specific, sensitive, enzyme tests even in tissue homogenates. The recently developed techniques of genetic engineering revolutionized the investigation of proteins. These methods not only allowed the unraveling of an immense number of primary structures from different proteins within a short time period. They also make possible the modification of proteins in any imaginable mode and open a broad field for altering functional domains like catalytic centers and binding sites to create new functions or to accept alternative ligands. The modifications of the protein molecule concerned are designed on a computer screen and realized by the methods of genetic engineering. Increasing interest is won also by the application of enzymes for industrial processes, e.g., in bioreactors and biosensors and enzyme therapy in medicine.

1. PROTEIN STRUCTURE

1.1 Primary Structure

Proteins are built up from amino acids. With rare exceptions of untypical examples (e.g., peptide antibiotics) all proteins so far described share 20 different amino acids, which are summarized in Fig. 1. They all carry the amino group at the α (or C2) position, creating an asymmetric center (with the exception of glycine) uniformly assuming the L conformation. With respect to their chemical character, the amino acids are classified as neutral, acid, basic, hydrophobic, and aromatic. Integrated in the protein molecule, the amino acids preserve their character. Most proteins possess polar amino acid residues on their surface, rendering the protein molecule soluble in the aqueous medium, while hydrophobic residues are buried to the inside and mediate also noncovalent contacts between different protein subunits.

In proteins, a large number of amino ac-

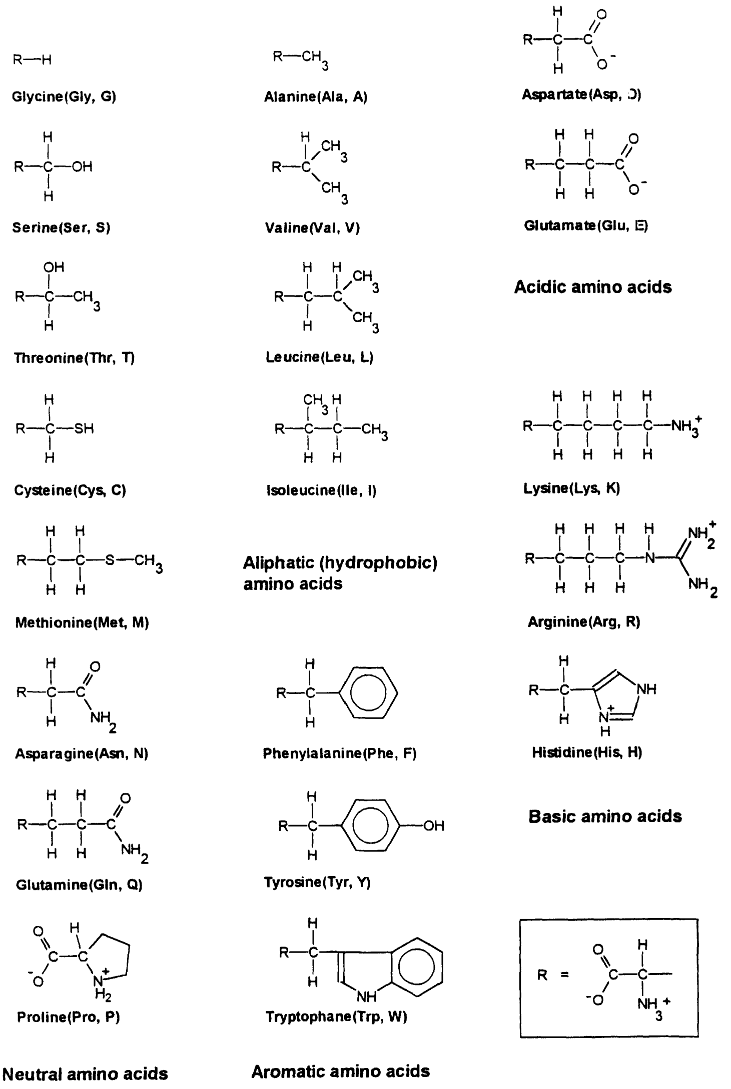
Glycine(Gly, G)
Alanine(Ala, A)
Aspartate(Asp, D)
Serine(Ser, S)
Valine(Val, V)
Glutamate(Glu, E)
Threonine(Thr, T)
Leucine(Leu, L)
Acidic amino acids
Cysteine(Cys, C)
Isoleucine(Ile, I)
Lysine(Lys, K)
Methionine(Met, M)
Aliphatic (hydrophobic) amino acids
Arginine(Arg, R)
Asparagine(Asn, N)
Phenylalanine(Phe, F)
Histidine(His, H)
Glutamine(Gln, Q)
Tyrosine(Tyr, Y)
Basic amino acids
Proline(Pro, P)
Tryptophane(Trp, W)
R =
Neutral amino acids
Aromatic amino acids

FIG. 1. Structures of proteinogenic amino acids (in brackets, abbreviations in the three- and the one-letter code).

ids (mostly 100–500, sometimes more than 1000) are linked together in a strictly defined sequence. This *primary sequence* is constant and characteristic for each special protein. The amino acids are linked together by acid amide bonds between the carboxy group of the one and the α-amino group of the second amino acid, the *peptide bond* (Fig. 2). In this way, a covalent *polypeptide* chain is formed with a uniform backbone, the amino acid side chains sticking out. There exist also smaller chains, the *peptides*, consisting of 2–10 amino acids, and the *oligopeptides*, containing 11–100 amino acids. They differ in their functions from the large protein molecules and act mainly as hormones or antibi-

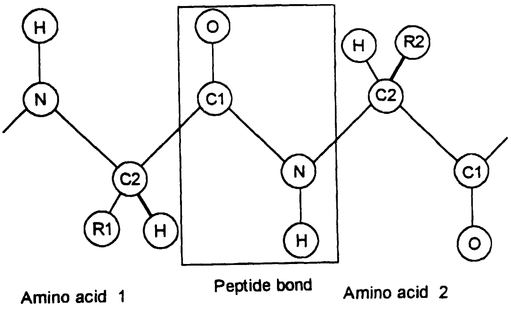
H
N
O
C1
H
R2
C2
N
C2
C1
R1
H
H
O
Amino acid 1
Peptide bond
Amino acid 2

FIG. 2. Peptide bond.

otics. Since the molecular mass of amino acids on the average is 110, most polypeptide chains have molecular masses between 10 000 and 60 000.

1.2 Secondary Structure

Native proteins cannot be regarded as flexible strings; rather, they assume defined stable three-dimensional structures. Three types of folding units, the secondary structure elements, are found in most proteins: α helix, β-pleated sheet, and random coil. In the *α helix,* the backbone of the peptide bonds is arranged in a right-handed spiral with 3.6 amino acid residues per turn and a pitch of 0.54 nm, stabilized by hydrogen bonds between the carboxy and the amino groups of the backbone (Plate 1). It has a rodlike structure with a diameter of about 0.5 nm and a maximum length of 4 nm, though α helices up to 100 nm occur in some structural proteins like myosin from muscle, keratin from hair, or fibrin from blood. The amino acid side chains point outward from the helix.

In *β-pleated sheets,* linear strands of the peptide backbone are arranged in a parallel or an antiparallel manner in a sheetlike structure, stabilized by hydrogen bonds between carboxy and amino groups from adjacent strands (Plate 1). Two to six strands form one sheet, but sometimes large structures up to 15 strands occur.

Native proteins are composed of various α-helix and β-pleated sheet elements. The regions between these elements appear more irregular and are designated as *random coil,* though they form also stable structures containing regular units like *β turns,* which often appear as a link between two antiparallel α helices. The β turns are hairpin turns where the polypeptide chain changes its direction by an angle of 180°. They are built from a sequence of four amino acids stabilized by hydrogen bonds. A special left-handed helix with 3.3 residues per turn is found in the structural protein collagen. This protein has a regular amino acid sequence with glycine at each third position, the following position frequently occupied by proline.

Which secondary structure element becomes realized at a defined region of the protein molecule is dictated by the amino acid sequence encoded in the primary structure. Both the angles of the peptide bonds and influences of the amino acid side chains determine the spatial arrangement of the polypeptide chain. Consequently, the amino acid sequence should allow the prediction of the secondary structure of a protein. However, because of an immense number of forces and restrictions governing macromolecular structures it has been impossible to deduce a convincing system for precise structure predictions. Out of numerous approaches, the most accepted one is that of Chou and Fasman (1978). A coefficient is ascribed to each amino acid for its probability to occur in one of the secondary structure elements, e.g., glutamic acid has a strong preference for α helices, while proline acts as a helix breaker. The coefficient was deduced empirically from a series of known protein structures. The Chou–Fasman rules give a high degree of correct predictions.

The secondary structure elements are arranged in proteins in *supersecondary structures,* like the $\beta\alpha\beta$ motif, where an α helix is embedded between two β strands. An extension of this motif is the eightfold $\alpha\beta$ barrel structure that appears in a series of enzymes like the triosephosphate isomerase and the pyruvate kinase. Eight alternating α helices and β strands are arranged like a barrel with a hole in its middle, which serves as a catalytic center of the enzymes. The Rossmann fold ($\beta\alpha\beta\alpha\beta$) is generated by fusion of two $\beta\alpha\beta$ units, while three antiparallel β strands connected by two β turns form a β meander (Rossmann and Argos, 1981).

1.3 Tertiary Structure

Larger proteins fold into distinct folding units, the *domains.* Each domain consists of 100–150 amino acid residues. Different domains are linked together by small peptide bridges. Domains are often functional units of proteins carrying binding sites, regulatory sites, or catalytic centers. One protein may fold into two or more domains, each contributing a special function that all together cooperate in the intrinsic physiological function of the native protein. It is assumed that in evolution larger proteins evolved by combination of domains with the proper functions.

Folding of polypeptide chains brings into close contact with one another distinct amino acid residues that are quite separate in the primary structure. In this way, stable three-dimensional centers are created, where the amino acid residues are kept in a precise spatial orientation in an optimum arrangement to allow specific interactions with ligands for specific binding, proper orientation, and activation for the catalytic conversion. These functional sites are often hydrophobic pockets, shielded from the surrounding aqueous phase and allowing the entry of the ligand molecule only through a narrow gap.

The outside surface of the protein is formed preferentially from hydrophilic amino acid residues that mediate the contacts to the hydrophilic environment and maintain the solubility of the protein molecule. Hydrophobic regions on protein surfaces serve as an anchor for integration into membranes of cells or cell organelles.

1.4 Quaternary Structure

Protein molecules with simple functions, like binding proteins (e.g., myoglobin) or degrading enzymes (e.g., proteases, ribonucleases), consist mainly of one single polypeptide chain (*monomer*). Most proteins, however, are aggregates of several polypeptide chains (*subunits*), which may be identical (homo-oligomers) or nonidentical (hetero-oligomers). Also, both identical and nonidentical subunits may occur in one protein molecule. The number of subunits of a distinct protein is strictly defined, and symmetrical structures are formed. Oligomers with a small, even number of subunits predominate: 2 (dimer), 4 (tetramer, assuming preferentially tetrahedral structures), or 6 (hexamer). Aggregation is important for stabilizing the protein structure (especially in the case of identical subunits), for regulatory mechanisms (e.g., allosteric enzymes, see Sec. 4.5), or to combine subunits with different functions. Aggregates of subunits with different catalytic functions are *multienzyme complexes,* like the pyruvate dehydrogenase complex or the membrane-associated respiratory chain. They contain several enzyme components, each catalyzing one step of a reaction sequence, so that intermediates formed by one enzyme component are channeled to the next enzyme component without diffusion into the surrounding fluid.

Protein subunits are usually attached to one another by strong, noncovalent forces, like salt bridges and hydrophobic bonds. The subunits can be separated by *p*H changes, high ionic strength, urea, or detergents. Sometimes subunits are connected by covalent dithiol bridges between cystein residues of different polypeptide chains. Such bonds are cleaved by reduction with thiol reagents like dithiothreitol.

1.5 Cellular Distribution of Proteins

Proteins are synthesized by a complicated machinery of the ribosomes in the cytosol of the cell, translating the genetic information encoded in the DNA sequence into the amino acid sequence (see NUCLEIC ACIDS). Since for many proteins the place of action is not in the cytosol, but in distinct cell compartments or membranes, they are synthesized originally as preforms carrying additional signal sequences of 15–35 amino acids at their amino terminus. This sequence directs the protein to its cell compartment (e.g., the mitochondria) and promotes its translocation through the membrane. At the final place, the signal sequence is split off, and the protein becomes activated (Pfanner and Neupert, 1990).

1.6 Posttranslational Modification of Proteins

Proteins are often modified after their synthesis (*processing*). Membrane proteins, antibodies, and other proteins are glycosylated by attachment of carbohydrates, branched oligosaccharide chains containing sugar components like glucose, mannose, or galactose, partially derivatized with amino and acetyl residues. The oligosaccharide residues cover the cell surface and mediate cell-cell recognition. The structural protein collagen is modified by hydroxylation of prolin and lysin residues. Protein modification serves also as a regulatory mechanism. Various enzymes are controlled by reversible phosphorylation of distinct serine or tyrosine residues. Several proteins, e.g., digestive enzymes (trypsin, chymotrypsin) or hormones (insulin), are modified by cleavage of peptide fragments, as described already for the signal sequences.

2. PROTEIN FUNCTIONS

Only a short survey on the major functions of proteins can be given here (enzymes will be treated separately in Sec. 3). The broad variety of protein functions is based on two intrinsic features of this class of compounds:

1. the ability of recognition and binding of a distinct ligand or substrate out of a large variety of compounds in the cell; and
2. the ability to undergo reversible conformational transitions, allowing the change of activity in answer to an external signal.

2.1 Structural and Contractile Functions

Proteins are structural determinants for cells and organisms. *Collagen,* which amounts to $\frac{1}{3}$ of all proteins in the human organism, is a major protein component of connective tissues, cartilage, tendons, and cornea. It has a special conformation, where three polypeptide chains are twisted in a helical manner, leading to stable elastic fibers. The *keratins* are the major components in hair, horn, silk, and skin and in epithelial sheets. They are fibrillar elastic proteins, which confer high flexibility to these substances. Together with neurofilaments in neurons and desmin filaments in muscle cells they belong to the family of intermediate filament (IF) proteins. They are components of the cytoskeleton, a cellular meshwork responsible for the shape and motility of the cell. Another component of the cytoskeleton is the microtubules, long cylindric structures about 30 nm in diameter and built up by the proteins α and β *tubulin.* These proteins form also the cilias of microorganisms. *Actin* and *myosin* are components of contractile microfilaments in the cytoskeleton. Together with *tropomyosin* they form the contractile elements of muscles.

2.2 Binding, Transport, and Storage

Proteins are responsible for transport and storage of substrates and metabolites. *Hemoglobin* in erythrocytes of the blood and *myoglobin* in muscle are both responsible for the binding and providing of oxygen for cellular respiration. Both these proteins have been intensely studied as models for metabolic regulation by proteins. Myoglobin, a monomeric protein, shows normal oxygen binding, while the tetrameric hemoglobin binds oxygen in a cooperative manner controlled by subunit-subunit interactions and was the first example for an allosteric protein (see Sec. 4.5). This cooperativity is influenced in a positive or negative manner by metabolic effectors, like the *p*H of the blood. Acid *p*H favors the dissociation of oxygen into the tissues while oxygen uptake in the lung releases protons from hemoglobin and liberates CO_2 from HCO_3^+. In both proteins, oxygen binds to a complex-bound Fe^{2+} ion of a porphyrin-ring system.

Iron is transported by *transferrin* in the blood and stored by *ferritin* in the liver. Ferritin binds iron up to one-quarter of its own molecular mass. Copper is transported in the blood by *ceruloplasmin.* Essential transport proteins for fatty acids and lipids in the blood are *serum albumin* and *lipoprotein.* Opsin, a membrane protein in the retina of the eye, combines with the light-absorbing chromophore retinal to form the eye pigment *rhodopsin.*

2.3 Infection Defense and Blood Clotting

Antibodies (*immunoglobulins*) protect the organisms against infections. They are of uniform structure consisting of two light and two heavy polypeptide chains linked together by disulfide bridges. Antigens, like viruses or bacteria, are inactivated by binding to highly variable regions on the immunoglobulins, which are able to adapt their amino acid sequence to the structure of the foreign substance. A large number of different proteins are involved in a complicated cascade mechanism of blood coagulation. The principle of this cascade is the activation of a series of inactive protein precursors after damage or trauma, by cleavage of small peptide fragments. This causes an amplification of the whole process. The final step is the cleavage of the fibrous protein fibrinogen to *fibrin,* which polymerizes to form the blood clot.

2.4 Membrane Proteins and Membrane Transport

Proteins integrated into membranes can be regarded as a special protein class. Though their general structure is similar to that of other proteins, they possess some remarkable features. Integration into the membrane is mediated by regions on the protein surface with a high content of hydrophobic amino acid residues. Biological membranes, which are essential components of the walls of cells and cell compartments, are lipid bilayers formed by fatty acids, cholesterol, and other lipid residues. These are amphiphilic compounds with both a hydrophobic hydrocarbon region and polar groups like carboxy residues or phosphates. They are arranged in thin layers, the polar groups oriented to the one, the hydrophobic residues to the other side. Two such layers form a membrane bilayer: the hydrophobic sites facing one another and the polar sides interacting with the polar surrounding liquid (Fig. 3). Hydrophobic regions of proteins dive into the lipid phase of the bilayer, the hydrophilic parts sticking outside. Membrane binding of proteins is very stable, and dissociation from the membrane as well as reorientation in the vertical direction through the membrane is quite improbable. In the lateral direction across the membrane, free flotation of the proteins occurs and allows different membrane proteins to come into direct contact. Because of their tight membrane integration, resolvation of the proteins from the membrane is difficult and results often in inactivation of the protein function. Membrane proteins are investigated either in membrane fragments or in the presence of lipids or detergents.

With respect to membrane binding of proteins, three types can be differentiated. Various proteins are only loosely associated with the membrane and can be resolved by applying mild conditions. The second class dip into the membrane with one part, while the other part looks outside, e.g., into the cytosolic fluid. This is the case for many membrane-bound enzymes like those from the respiratory chain. The third class is completely integrated into the membrane with a central hydrophobic region and with polar regions looking outside at both faces of the membrane (integral membrane proteins). These proteins can interact with components from both sides of the membrane and are able to transfer signals from the one to the other side. This is the principle of the receptor function, where hormones or transmitter substances bind to specific receptors outside the membrane, triggering a reaction inside of it. A characteristic example is the β-adrenergic hormone receptor. Its central part consists of seven parallel transmembrane helices, which are integrated into the mem-

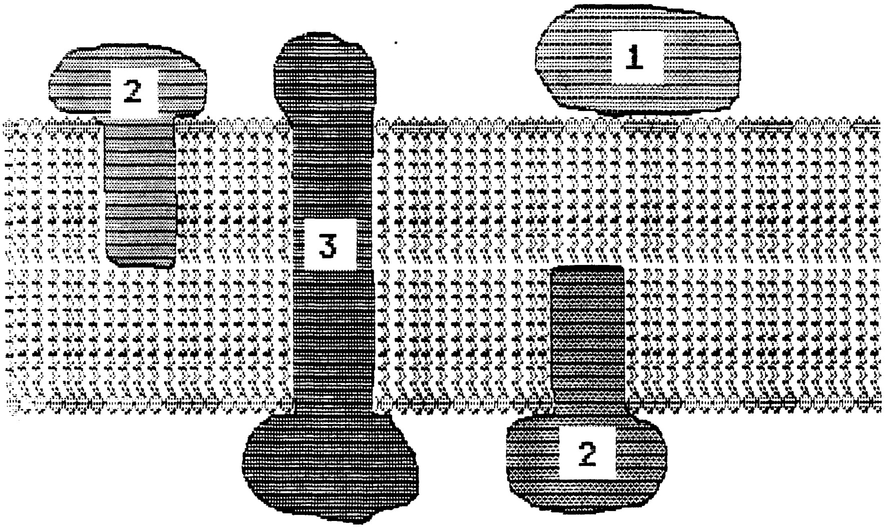

FIG. 3. Bilayer membrane with (1) loosely associated, (2) integral, and (3) transmembrane proteins.

brane. The hormone adrenaline binds to the receptor at the outside of the cell membrane. To the inside, a regulatory *G* protein is associated, which is an aggregate of three different subunits: G_α, G_β, and G_γ. Hormone binding triggers the dissociation of the G_α subunit, which binds to a membrane-bound enzyme, the adenylate cyclase. The binding activates this enzyme, and it catalyzes the synthesis of cyclic adenosine monophosphate (cAMP), a cell hormone or second messenger that regulates various processes in the cell. Also, for other peptide hormones and neurotransmitters like acetyl choline and γ aminobutyrate (GABA), there exist specific membrane-protein receptors. Neurotransmitters act by opening specific ion channels in the membrane.

Ion channels play an essential role in maintaining ion gradients between the two sides of the membrane. These gradients are important for various cellular processes like energy generation, signal transduction, cellular regulation, muscle contraction, and nerve conduction. The ion channels (ion pumps) consist of transmembrane proteins that bind the respective ions (Ca^{2+}, Na^+, Cl^-, H^+, etc.) with high specifity. The transport across the membrane against the concentration gradient occurs by a conformational change of the protein. The energy is provided by the breaking of an energy-rich phosphoric acid anhydride bond of adenosine triphosphate (ATP), the main energy mediator of the cell. Restoration of the energy-rich bond is brought about in the reverse process also by a transport system, the membrane-bound proton pump. The cellular respiration on the inner mitochondrial membrane generates a proton gradient by selective segregation of protons to the outside. To compensate the concentration gradient, protons migrate to the inside of the membrane through a specific proton channel, and the energy of this process is used to induce a conformational change of proteins of the system that mediates the ligation of energy-rich ATP bonds. Membrane-bound proteins, like the photosynthetic reaction center, are responsible for the photosynthetic fixation of carbon dioxide.

Transport of metabolites and other substances like amino acids or carbohydrates through membranes is achieved by integral membrane proteins. The metabolites are specifically recognized by receptor regions at the outside of the membrane. Translocation through the membrane occurs via a conformational change of the protein driven by breakage of energy-rich ATP bonds. Selective, energy-driven carrier systems that control the uptake of substances from the environment enable the organism to segregate only such compounds as are required for its metabolism and to concentrate them within the cell.

3. ENZYMES

3.1 Enzyme Catalysis

Enzymes, the most important group of proteins, have a twofold function for the living organism: they are the catalysts and also the regulators of all reactions of the cell metabolism. In comparison with chemical catalysts, enzymes are very efficient, enhancing the reaction rates by factors of more than 10^{10}. Thus even reactions that appear quite unfavorable become accessible to the metabolism. As true catalysts, however, enzymes can enhance the reaction rate, but cannot influence the thermodynamic equilibrium. If the state of the equilibrium of a reaction favors the substrate *A* instead of the product *B* (*endergonic reactions*—the free energy ΔG has a positive value):

$$A \overset{+\Delta G}{\rightleftharpoons} B,$$

it cannot be shifted by the enzyme in order to provide the cell with sufficient product. But even such reactions can become accessible by coupling them to one (or more) strongly *exergonic reactions* (negative ΔG values). The second reaction releases energy, which is taken to drive the first one. It is only essential that the net amount of free energy remains negative. The principle of this coupling is the fast removal of the product of the first enzyme reaction by the enzyme catalyzing the second one so that the reaction becomes *kinetically driven:*

$$A \overset{+\Delta G_1}{\rightleftharpoons} B \overset{-\Delta G_2}{\rightleftharpoons} C, \quad \Delta G_1 < -\Delta G_2.$$

In most cases, the cleavage of an energy-rich

ATP bond serves to supply unfavorable reactions with the necessary energy, and the main task of the energy metabolism of the cell is to regenerate ATP as a motor of the whole cell metabolism.

Each metabolic reaction is catalyzed by a special enzyme. This is true even for the same type of reactions if the substrates differ, e.g., dehydrogenation of compounds like ethanol and lactate. This apparently uneconomic principle makes possible independent regulation of each reaction and the realization of a complex network of hundreds of metabolic reactions occurring in parallel in the same cell compartment at the same time.

3.2 Catalytic Center

The catalytic conversion occurs at a distinct region of the enzyme molecule, the *catalytic center* or *active site.* It is a hydrophobic pocket inside of the enzyme molecule with an entry for the substrate. It is formed by a precise three-dimensional arrangement of functional amino acid residues of the protein chain. Often it is found in a cleft between different domains, and also elements from different subunits of a protein molecule can contribute to an active site. The active site is responsible for the recognition and discrimination of its substrate molecule out of a mixture of partly very similar compounds in the cell (*enzyme specificity*). For correct recognition, close interaction between substrate and catalytic site by proper positioning of charges and steric obstacles is important. This adaptation is supported by conformational changes at the active site, which are induced by the substrate upon binding (*induced fit*). Conformational changes cause further stabilization of the transition state, which initiates the catalytic process.

Besides different functional amino acid residues of the protein, frequently nonprotein components bound covalently (*prosthetic groups*) or noncovalently (coenzymes, cofactors) to the catalytic center are involved in enzyme catalysis. Many of these coenzymes or their precursors cannot be synthesized by the human organism itself and must be supplied by food (*vitamins*). Examples of such coenzymes are nicotinamide adenine dinucleotide (NAD) and flavins FMN (flavin mononucleotide) and FAD (flavin dinucleotide), both of which are involved in redox reactions; thiamine diphosphate (ThDP), taking part in decarboxylation reactions; and coenzyme A, mediating the acyl group transfer.

Many enzyme reactions use metal ions as cofactors. Magnesium occurs in all reactions with the energy mediator ATP, zinc is found in alcohol dehydrogenase, copper is found in the cytochrome oxidase of the respiratory chain, and molybdenum is found in MoFeS clusters that participate in the nitrogen fixation of the nitrogenase reaction. Iron is involved in a variety of reactions, either complex bound in a porphyrin ring system (e.g., in cytochromes of the respiratory chain and in the oxygen carrier hemoglobin) or as iron-sulfur clusters in redox reactions.

3.3 Mechanisms of Enzyme Catalysis

The principle of enzyme catalysis is the reduction of the activation energy by stabilizing the transition state. To achieve this, the reactants are brought into direct contact with one another by binding to the same catalytic center and by fixing them in a proper orientation. This effect results in a large increase of the efficient substrate concentration and is an entropic contribution to enzyme catalysis (Jencks, 1975).

In various enzyme reactions, covalent intermediates are formed by nucleophilic attack of functional groups of the enzyme to electrophilic regions at the substrate molecule. Nucleophilic groups in the catalytic site are carboxy, amino, hydroxy, imidazol, or thiol groups, and acyl, glycosyl, or phosphoryl intermediates can be formed from appropriate substrates. In a second step, the intermediate becomes transferred to a nucleophilic acceptor molecule, e.g., water in hydrolytic reactions.

Another principle for the acceleration of catalytic reactions is to exert strains or distortions on the substrate molecule upon binding to the enzyme. The substrate becomes destabilized to facilitate the switch into the transition state. This is achieved by twisting of bond angles, steric compression, or electrostatic repulsion between charged groups of the substrate molecule and the catalytic site (Walsh, 1979).

Many enzyme reactions proceed by general acid-base catalysis, e.g., proton transfer to or from the transition state. At neutral *p*H, the imidazol group of the amino acid histidine is a very efficient base, and it is often found in catalytic sites. An exceptional example of general acid-base catalysis is the chymotrypsin reaction. The catalytic site of this enzyme contains three functional residues in juxtaposed position: an imidazol group from histidine, a hydroxy group from serine, and a carboxy group from aspartate (*catalytic triad*). The reaction starts by a proton transfer from the hydroxy group to a nitrogen of the imidazol ring, which is stabilized by the carboxy group of aspartate. A nucleophilic attack of the hydroxy group on a carbonyl group of the substrate and a proton transfer from imidazol to the amino group of the substrate destabilizes the peptide bond. The amine product is released, and an acyl intermediate remains bound to the serine residue. Hydrolysis of this intermediate by a water molecule and release of product is a reversion of the acylation mechanism (Fig. 4).

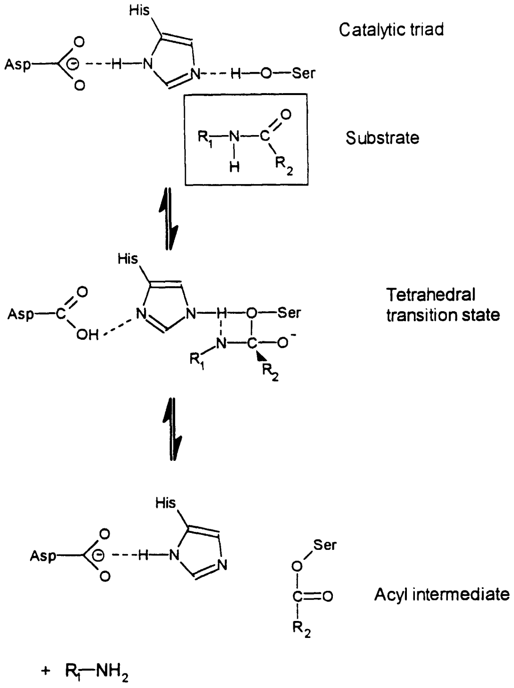

FIG. 4. Catalytic mechanism of chymotrypsin.

3.4 *p*H Dependence of Enzyme Reactions

The activity of enzymes is strongly dependent on the *p*H value of the solution. Optimum curves are obtained with a maximum mainly within the neutral range (*p*H 7.0–8.0) according to neutral *p*H of the cell, but there are examples of very extreme *p*H optima, like acid proteases, e.g., pepsin of the gastric juice, which has its highest activity at *p*H 2.0. Both at low *p*H (<3) and at high *p*H (>11), the enzyme activity declines to zero. The *p*H optimum is determined by two effects:

1. direct influences by ionized or protonated groups in the active center, and
2. indirect influences by residues essential for maintenance of the native structure.

The catalytic residues show sigmoidal *p*H titration curves in dependence of the enzyme activity. Their inflexion points correspond to the pK_a values, i.e., the *p*H value at which the residue is half protonated. While *p*H-dependent dissociation of active site residues is usually reversible, activity losses due to changes of the ionic state of structurally important residues often cause irreversible denaturation of the enzyme.

As a result of their protein nature, enzymes carry a large number of charged groups (amino, carboxy, thiol, hydroxy residues). Within the protein molecule, positive and negative charges are mutually neutralized, but usually there remains a residual amount of surplus charges at physiological *p*H. At a distinct *p*H value, the *isoelectric point* (*pI*), also this net charge becomes zero. The *pI* is a characteristic value for proteins, where they show their lowest solubility in water and tend to precipitate. At this *p*H value, the protein does not migrate in the electric field. Various preparative techniques, like isoelectric focusing and precipitation methods, make use of this fact. Generally, the surface charges of protein molecules are important for their solubility in polar solvents like water, and many analytical and preparative methods (e.g., electrophoresis, ion chromatography) are based on the polyionic nature of proteins.

3.5 Temperature Dependence and Arrhenius Theory

To convert substrate into product, the energy barrier of the transition state has to be overcome, and the reaction proceeds only if the system possesses sufficient energy to surmount the barrier. The principle of enzyme catalysis is to reduce the energy barrier by stabilizing the transition state. The activation energy E_a of the transition state is given by the Arrhenius equation:

$$k = Ae^{-E_a/RT}, \tag{1}$$

where k is the rate constant of the reaction, A is a pre-exponential factor containing the collision frequency and a steric component, R is the gas constant, and T is the absolute temperature (K). According to this equation, reaction rate is a logarithmic function of the reciprocal absolute temperature:

$$\ln k = \ln A - E_a/RT. \tag{2}$$

A plot of ln k against $1/T$ should yield a straight line with the slope of E_a/R.

Enzyme-catalyzed reactions, like chemical reactions, are accelerated by a factor of 2–3 if the temperature is raised by 10 °C. However, because of their metastable protein structure, enzymes denature irreversibly at higher temperatures, and the reaction velocity drops to zero. In this region, the straight line in an Arrhenius plot declines abruptly. In contrast to the pH optimum, the temperature maximum cannot be regarded as optimal for the enzyme activity, since denaturation occurs even at the maximum temperature. Individual enzymes differ considerably with respect to their temperature ranges of denaturation and their thermal stability. Most enzymes denature between 40 and 50 °C, but there exist extremely temperature-sensitive enzymes like alcohol dehydrogenase, which denatures already at 30 °C, while enzymes from thermophilic microorganisms living in hot volcanic springs remain stable up to 100 °C. Thermal denaturation is a complex, irreversible, multistep process. Usually, one initial step causes the loss of catalytic activity, followed by a series of processes leading to a successive unfolding of the native protein structure. Investigation of such processes is important for the understanding of the native enzyme structure.

3.6 Enzyme Regulation

In contrast to spontaneous chemical reactions, catalyzed reactions can be influenced by changing both the amount and the efficiency of the catalyst. Both these regulation principles are realized. There exist mechanisms for adaptation of the amount and the activity of enzymes to the actual metabolic requirements. The amount of the enzyme is regulated at the level of enzyme biosynthesis by control of gene expression (see NUCLEIC ACIDS). Specific metabolites that are directly or indirectly involved in the enzyme reaction, like substrate, products, or intermediates, either enhance (*induction*) or repress (*repression*) the gene expression. Usually, substrate molecules act as inducers, and their accumulation favors enzyme synthesis, while products are repressors signaling the stop of the synthesis. Many compounds and metabolites are synthesized in sequences of several reaction steps, each being catalyzed by a specific enzyme. Such *metabolic pathways* are regulated by interaction of their end product as a repressor and of their initial substrate as an inducer with a regulatory gene (operator) of the gene cluster (operon) containing the structural genes of the respective enzymes (Fig. 5). Too high amounts of enzymes already present in the cell are digested by specific degradative enzymes, the *proteases*.

The efficiency of the enzyme catalyst is controlled by binding and interaction of regulatory substances (*effectors*) to the enzyme. This regulation principle is similar to gene regulation: reaction substrates activate, products inhibit. Metabolic pathways are subject to a concerted regulation: The first enzyme is inhibited by the end product of the whole pathway (*feedback inhibition*) so that accumulation of intermediates is prevented. The inhibitors and activators react either by direct interaction, e.g., blocking of the active site of the enzyme, or by binding to a distinct regulatory site that transfers the regulatory signal to the active center by conformational changes and subunit-subunit interactions of the protein (*allosteric enzymes*, cf. Sec. 4.5). Another regulation principle is effector-induced covalent but reversible modification of the enzyme, frequently via phosphorylation of distinct serine or tyrosine residues. This effect causes either activation or inhibition depending on the special type of enzyme. Phosphorylation is further controlled by the regulatory metabolite cyclic adenosine monophosphate (cAMP), a *second messenger* of the cell. The cAMP level is itself

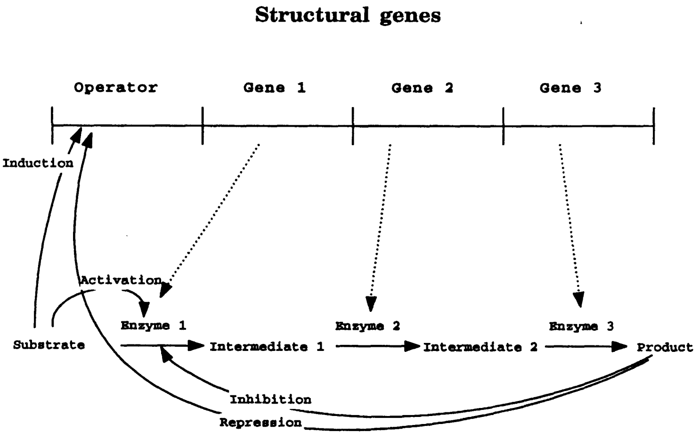

FIG. 5. Regulation of enzyme activity on the gene and the enzyme levels.

subject to hormonal control, thus connecting hormone action to enzyme activity.

3.7 Cellular Organization of Enzymes

In the cells, the enzymes are partitioned in various compartments to separate different pathways. In the cytosol, the metabolization of carbohydrates in the *glycolytic pathway* (Embdem–Meyerhof pathway) and the *pentose-phosphate pathway* occur and fatty acids are there synthesized. Degradation of fatty acids for energy generation and for disposition of precursors for metabolite synthesis occurs in the mitochondria, where it is coupled to the *citric acid cycle* (Krebs cycle) and the *respiration chain,* the main energy generating machinery of the cell. Lysosomes, small cell particles, contain enzymes for degradation of proteins, polysaccharides, and lipids, and the peroxisomes concentrate enzymes that form and destroy the highly toxic hydrogen peroxides.

Usually one reaction is catalyzed by one enzyme. The enzyme molecule either may consist of one polypeptide chain carrying one catalytic center or (more frequently) may be an oligomer consisting of some (about 2–6) identical polypeptide chains, each possessing the same catalytic site. The reasons for aggregation and for concentration of catalytic sites of the same type in one molecule may be the stabilization of the native structure and the reduction of particle volumes within the limited cell space. Also regulatory mechanisms are based on subunit-subunit interactions (cf. Sec. 4.5). Various enzymes possess more than one type of catalytic center and catalyze different, usually sequential reaction steps (*multienzyme complexes*). The reaction intermediates are directly transferred from one center to the next without dilution into the cell liquid. Different catalytic centers are usually located on different subunits so that multienzyme complexes consist of several nonidentical subunits, and each subunit type can appear in such complexes in several identical copies. The pyruvate dehydrogenase complex, the largest multienzyme complex of the cell, consists of three different enzyme components, two of them contributing 60 and the third one contributing 12 identical copies so that the total molecular mass amounts to 8×10^6 for the whole enzyme complex. Sometimes different catalytic sites are located on the same polypeptide chain (*multifunctional enzymes*), as in the case of the fatty acid synthase, where seven catalytical centers are arranged on one single subunit.

Higher organized structures are realized by membrane fixation of proteins and enzymes. An example is the respiratory chain, which consists of several functional protein complexes integrated into the inner mitochondrial membrane. It mediates the transport of reduction equivalents from various metabolites to molecular oxygen by formation of water. Segregation of this chemically simple reaction into several enzyme-controlled reaction steps allows conservation of the energy liberated by this process. Protons produced in the course of this reaction sequence are vectorially released into the intermembrane space between the inner and the outer mitochondrial membranes. A proton gradient is generated, and in order to equalize the concentration difference, protons migrate through specific proton channels, which converts the energy gained by this process into the chemical energy of acid anhydride bonds of ATP. Three such bonds are formed by one passage of reduction equivalents through the respiratory chain, corresponding to 40% efficiency.

4. KINETICS OF ENZYME REACTIONS

4.1 Michaelis–Menten Equation and Steady-State Relationship

The central law for enzyme analysis and enzyme kinetics is the equation formulated in 1913 by L. Michaelis and M. Menten in Berlin on the basis of earlier studies of A. J. Brown and V. Henry. It was originally derived for a single substrate reaction:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P,$$

where the reversible formation of the ES complex (*Michaelis complex*) between enzyme E and substrate S is followed by the catalytic conversion of substrate into product P and its dissociation from the enzyme in a quasi-irreversible process. Michaelis and Menten assumed the catalytic process to be very slow compared to the initial binding equilibrium; the rate constants should be related by $k_1 \sim k_{-1} \gg k_2$. Under these conditions, the reaction velocity v, which is proportional to the concentration of the ES complex ($v = k_2[ES]$), can be taken as a direct measure for the reversible binding equilibrium, which was thought to be not influenced by the following catalytic step. On the basis of this assumption, the *Michaelis–Menten equation* for the dependence of the enzyme velocity on the substrate concentration was derived:

$$v = V[S]/(K_m + [S]). \qquad (3)$$

The term V is the maximum velocity for the case that the total amount of enzyme in the sample $[E]_0$ is saturated with substrate and is involved in the reaction; $V = k_2[E]_0$. It can be attained only at infinite substrate concentration. According to the Michaelis–Menten assumption, the *Michaelis constant* K_m is identical to the dissociation constant of the ES complex, $K_d = k_{-1}/k_1$.

In 1923, G. E. Briggs and J. B. S. Haldane suggested a modification of the Michaelis–Menten equation that extends its general validity. They realized that the catalytic conversion may be fast enough to influence the binding equilibrium, for which v is no longer a direct measure. But even under these conditions, the concentration of the ES complex remains constant for a distinct time, where formation of ES and decomposition into E, S, and P are counterbalanced. In contrast to a real (time-independent) equilibrium, this state, which lasts only for a limited time period of several seconds to minutes, is designated as *steady state.* Equation (3) remains valid also for this modification, with the exception that the Michaelis constant is extended by the catalytic constant k_2: $K_m = (k_{-1} + k_2)/k_1$. It cannot be further regarded as a true thermodynamic dissociation constant. The steady-state range can only be observed if $[S] \gg [E]$ and is characterized by an initial linear zero-order product formation, which turns into nonlinear first-order kinetics after a considerable decrease of the initial substrate concentration. In this modified form, the Michaelis–Menten equation is valid also for more complicated enzyme reactions, including more than one substrate, enzyme inhibition or activation, and reversible reactions, if initial reaction rates are measured and all parameters besides the one under investigation are kept constant.

4.2 Graphical Representation of the Michaelis–Menten Equation

The dependence of the enzyme velocity on the substrate concentration according to the Michaelis–Menten Eq. (3) obeys a hyperbolic function [Fig. 6(a)] and approaches asymptotically the maximum velocity V since v becomes V for $[S] \rightarrow \infty$. At half saturation ($v = V/2$), the substrate concentration corresponds to the Michaelis constant K_m. Both K_m and V are characteristic enzyme constants. The Michaelis constant reflects the affinity of the substrate for the catalytic center, as long as k_2 is small. The K_m values range usually be-

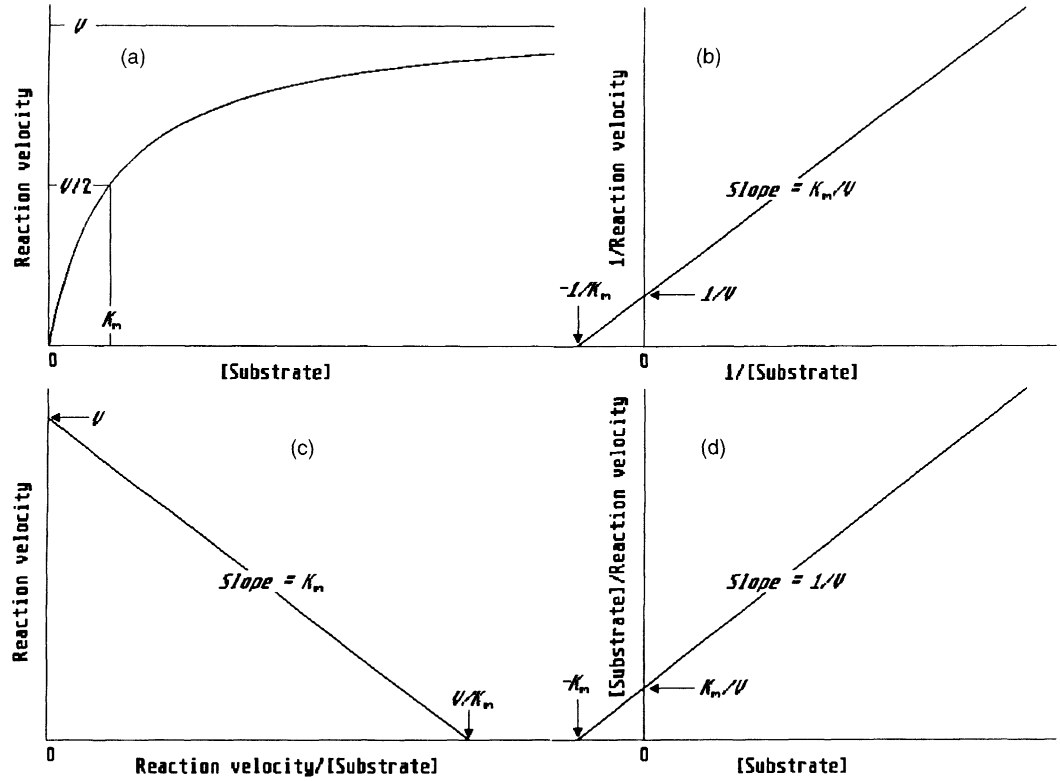

FIG. 6. Graphical representation of the Michaelis–Menten equation: (a) direct plotting; (b) Lineweaver–Burk diagram; (c) Eadie–Hofstee diagram; (d) Hanes diagram.

tween 10^{-6} and 10^{-3} mol/L. The term V is defined as the amount of substrate converted per time unit (mol/s). It is a measure of the catalytic efficiency of the enzyme and depends on the amount of enzyme and thus on the purity of the preparations. Of more general significance is the catalytical constant or turnover number k_{cat} [k_2 in Eq. (3)], which is obtained by dividing V by the enzyme concentration $[E]_0$. The k_{cat}/K_m ratio is a measure for the enzyme specificity, large values indicating high specificity.

Though both K_m and V can be obtained by direct plotting of v against $[S]$, as shown in Fig. 6(a), this procedure is often ambiguous because of the problems of fitting nonlinear curves to scattering data points. This can be overcome by transformation of the Michaelis–Menten equation into linearized forms. Three simple ways of linearization exist. In the double reciprocal *Lineweaver–Burk diagram* the variables are plotted reciprocally, $1/v$ against $1/[S]$ (Fig. 6(b)), according to the following equation:

$$\frac{1}{v} = \frac{K_m}{V}\frac{1}{[S]} + \frac{1}{V}. \tag{4}$$

The *Eadie–Hofstee diagram* plots v against $v/[S]$ [Fig. 7(c)]:

$$v = K_m \frac{v}{[S]} - V, \tag{5}$$

and $[S]/v$ against $[S]$ is plotted in the *Hanes diagram* [Fig. 7(d)]:

$$\frac{[S]}{v} = \frac{[S]}{V} + \frac{K_m}{V}. \tag{6}$$

The Lineweaver–Burk plot is used most frequently, though the reciprocal plotting distorts the error margins of the data points and prohibits application of linear regression. This problem is less important with the other plots, but there the two variables v and $[S]$ are not separated by the coordinates. The kinetic constants K_m and V are obtained from these diagrams from slopes and coordinate intercepts of the straight lines, as indicated in Fig. 6.

4.3 Enzyme Inhibition

Enzyme reactions can be specifically inhibited by distinct compounds, the inhibitors. Inhibition is of great physiological significance for the regulation of the cell metabolism. Inhibition is a valuable tool for the study of enzyme mechanisms. Selective blocking of enzyme reactions by specific inhibitors is also applied in medical therapy.

Three main inhibition types are differentiated. If the inhibitor binds directly to the substrate binding site at the catalytic center, the inhibition is *competitive*. This is the case if the inhibitor is structurally closely related to the substrate but does not participate in the catalytic reaction. Both substrate (S) and inhibitor (I) cannot bind simultaneously to the enzyme but compete for the same binding site:

$$\begin{array}{l} E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P \\ + \\ I \\ \downarrow\uparrow K_i \\ EI \end{array}$$

Here K_i, the binding constant of the inhibitor I, is the *inhibition constant*. The Michaelis-Menten Eq. (3) changes to

$$v = \frac{V[S]}{K_m(1 + [I]/K_i) + [S]}, \tag{7}$$

and the K_m value is extended by a term containing the inhibitor concentration $[I]$ and the inhibition constant. Here V remains unchanged, since very high amounts of the substrate displace the inhibitor completely from the binding site. In the Lineweaver–Burk plot, the competitive inhibition is characterized by a common ordinate intercept at $1/V$ of straight lines that are obtained by determination of the enzyme activity in the presence of different inhibitor concentrations [Fig. 7(a)]. The term K_i is obtained from the abscissa intercepts of the lines, which have the meaning $-1/K_m(1 + [I]/K_i)$.

The product of an enzyme reaction, if present already at the start of the reaction, behaves like a competitive inhibitor, since it shares the same binding site as the substrate (*product inhibition*).

If there is no direct competition between inhibitor and substrate, the inhibition is *noncompetitive*. The inhibitor may bind either to a second binding site at the catalytic center (e.g., for cofactor or cosubstrate) or to a distant regulatory site. An inactive ternary EIS complex appears:

$$\begin{array}{lll} E + S & \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} & ES \xrightarrow{k_2} E + P \\ + & & + \\ I & & I \\ \downarrow\uparrow K_i & & \downarrow\uparrow K_{ii} \\ EI + S & \rightleftharpoons & EIS \end{array} .$$

Two inhibition constants exist: K_i and K_{ii}. They are identical if both substrate and inhibitor bind independently of one another. In this case, V, but not K_m, becomes changed:

$$v = \frac{V[S]/(1 + [I]/K_{ii})}{K_m + [S]}. \tag{8}$$

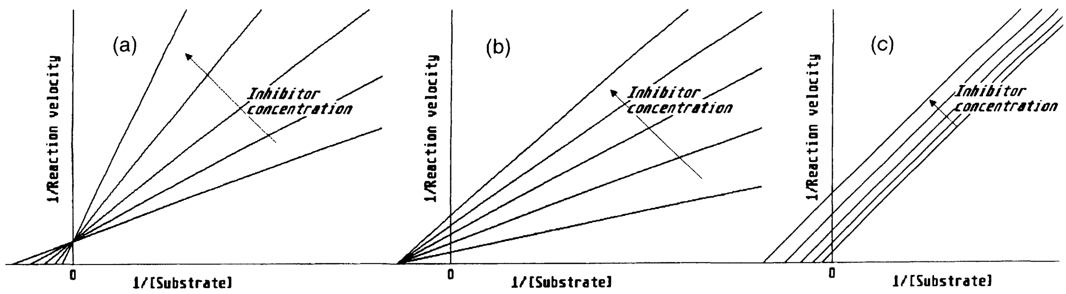

FIG. 7. Enzyme inhibition in the Lineweaver–Burk diagram: (a) competitive inhibition; (b) noncompetitive inhibition; (c) uncompetitive inhibition.

Indicative for this inhibition type is a common abscissa intercept of all lines in the Lineweaver–Burk plot [Fig. 7(b)]. The term K_{ii} is obtained from the ordinate intercept $(1 + [I]/K_{ii})/V$. If binding of inhibitor influences the binding of substrate and vice versa, then $K_i \neq K_{ii}$, and the rate equation reads

$$v = \frac{V[S]}{K_m(1 + [I]/K_i) + (1 + [I]/K_{ii})[S]}. \tag{9}$$

This case is sometimes designated as *mixed inhibition.* Now both K_m and V are changed, and the lines in the Lineweaver–Burk plot intersect at a common point to the left of the ordinate. Both K_i and K_{ii} are derived from abscissa and ordinate intercepts, as described above.

A rare type is *uncompetitive inhibition.* The inhibitor binds exclusively to the ES complex; thus, substrate binding is necessary for inhibitor action:

$$\begin{array}{l} E + S \overset{k_1}{\underset{k_{-1}}{\rightleftharpoons}} ES \overset{k_2}{\rightarrow} E + P \\ \qquad\quad\;\; + \\ \qquad\quad\;\; I \\ \qquad\quad\;\; \downarrow\!\uparrow K_{ii} \\ \qquad\quad\;\; EIS \end{array} .$$

The rate equation is

$$v = \frac{V[S]}{K_m + (1 + [I]/K_{ii})[S]}. \tag{10}$$

Parallel lines appear in the Lineweaver–Burk plot; K_{ii} is obtained from the ordinate intercepts [Fig. 7(c)]. A special case of uncompetitive inhibition is *substrate inhibition.* A second substrate molecule binds to the ES complex, forming an inactive ESS complex. At high substrate concentrations, the reaction velocity slows down and characteristic deviations from the straight lines are observed in the linearized diagrams.

4.4 Multisubstrate Reactions

In most enzyme reactions, two and sometimes even three different substrates are involved. Three main mechanisms of multisubstrate reactions are differentiated with respect to the order of substrate binding: random, ordered, and ping-pong mechanisms. The numbers of substrates involved and of products released are added to the name of the mechanism (uni, bi, and ter for one, two, and three components), e.g., ordered bibi for an ordered mechanism with two substrates converted into two products. The Michaelis–Menten Eq. (3) and the linearization procedures remain principally valid as long as all substrates but one are kept constant (Cleland, 1970).

For the *random mechanism,* independent binding of the substrates ($S1$, $S2$), the following is assumed:

$$\begin{array}{lcl} E + S1 & \overset{K_{iS1}}{\rightleftharpoons} & ES1 \\ + & & + \\ S2 & & S2 \\ \downarrow\!\uparrow K_{iS2} & {\scriptstyle K_{mS1}} & \downarrow\!\uparrow K_{mS2} \\ ES2 + S1 & \rightleftharpoons & ES1S2 \end{array} .$$

For each substrate, there exist two types of constants, an *inhibition constant* (K_{iS1},K_{iS2}) for the binding to the free enzyme and a *Michaelis constant* (K_{mS1},K_{mS2}) for the formation of the active $ES1S2$ complex. The constants are related by the relationship $K_{iS1}/K_{iS2} = K_{mS1}/K_{mS2}$. Both the K_i and K_m values for the same substrate are identical if the different substrates bind independently, and they differ if the substrates mutually influence each other in the binding process. The rate equation is

$$v = \frac{V[S1][S2]}{K_{iS1}K_{mS2} + K_{mS2}[S1] + K_{mS1}[S2] + [S1][S2]}. \tag{11}$$

Multisubstrate reactions are analyzed by varying the amount of one substrate at constant cosubstrate concentrations. A series of straight lines with a common abscissa intercept is obtained for the random mechanism in the Lineweaver–Burk plot.

The different substrates bind in a strictly sequential mode, one after the other, in the *ordered mechanism:*

$$E \overset{+S1}{\rightarrow} ES1 \overset{+S2}{\rightarrow} ES1S2.$$

Equation (11) remains valid, and the inhibition constants for binding of a substrate differ considerably from its Michaelis constant. The lines in the Lineweaver–Burk plot intersect at a common point to the left of the ordinate.

In the *ping-pong mechanism,* the first substrate forms an intermediate with the enzyme E' while the first product becomes released. The second substrate reacts with the enzyme intermediate, forming the second product and restoring the original enzyme form E:

$$E \xrightarrow{+S1} ES1 \xrightarrow[-P1]{} E' \xrightarrow{+S2} E'S2 \xrightarrow[-P2]{} E.$$

Examples for a ping-pong mechanism are aminotransferases, an enzyme class with pyridoxal phosphate as a prosthetic group. They catalyze the transfer of an amino group from an amino acid to an α-oxo acid by formation of a covalent enzyme-pyridoxamine intermediate.

The rate equation for this mechanism is

$$v = \frac{V[S1][S2]}{K_{mS2}[S1] + K_{mS1}[S2] + [S1][S2]}. \qquad (12)$$

Parallel lines are obtained in the Lineweaver–Burk plot.

4.5 Allosteric Enzymes

Allosteric enzymes represent one of the major regulatory principles of the cell. Enzymes belonging to this group show three significant features:

1. composition of identical subunits;
2. sigmoidal saturation behavior of ligands, mainly the substrate (*homotropic effects*); and
3. regulatory centers (*allosteric sites*) for binding and action of effectors (inhibitors or activators) distant from the catalytic site (*heterotropic effects*).

The sigmoidal shape of the saturation curves is generally ascribed to cooperativity due to subunit-subunit interactions. The first ligand binding to a subunit increases the affinity for the following ones. The steepness of the sigmoidal saturation curves is a measure of the strength of cooperativity and of subunit-subunit interaction. Allosteric inhibitors strengthen the cooperative effects (though the enzyme activity is reduced), while activators weaken it and convert the sigmoidal curves to hyperbolic ones (Fig. 8).

Two models were proposed that are equally suited to describe the behavior of allosteric enzymes. According to the *concerted* or *symmetry model* of Monod *et al.* (1965), the enzyme molecule exists in at least two conformational states: an inactive T state (T = tense) and an active R state (R = relaxed). The transition between the two states occurs spontaneously in a symmetrical manner, all identical subunits of the enzyme molecule assuming the same conformation at the same time. In the absence of substrate, the T state predominates. Substrates and activators bind preferentially to the active R state, which becomes stabilized and is drawn out of the equilibrium between the free states. To restore this equilibrium, a corresponding amount of enzyme in the free T state changes into the R state. Conversely, inhibitors stabilize the T state and reduce the amount of the active R state.

Also, the *sequential model* of Koshland *et al.* (1966) assumes the enzyme to exist in an inactive state in the absence of substrate. Substrate binding induces a change to the active conformation, but only of those enzyme subunits to which the substrate actually binds, while free subunits remain in their inactive state. It is further assumed that interactions between active and inactive subunits facilitates binding of substrate molecules to the inactive states and their transition into the active form. Activators act in

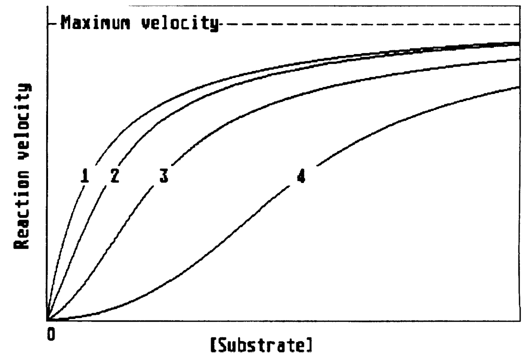

FIG. 8. Sigmoidal saturation behavior of allosteric enzymes: (1) normal hyperbolic curve; (2–4) sigmoidal curves: (2) in the presence of an activator; (3) without effector; (4) in the presence of an inhibitor.

the same way, while inhibitors impede the transition.

Allosteric enzymes control the initial steps of metabolic pathways (*key enzymes*) and are regulated by the end product of the pathway (*feedback inhibition*). One of the well-documented allosteric enzymes is aspartate transcarbamoylase, which possesses catalytic and regulatory centers on different subunits. It catalyzes the initial step of the biosynthetic pathway of pyrimidines, one of the two central building blocks for nucleic acids. The enzyme is inhibited by the end product of the pathway, cytidine triphosphate. ATP, the end product of the biosynthetic pathway of purines, the second building block for nucleic acids, acts as an activator. This regulation guarantees that both building blocks are synthesized in the same ratio as is needed for nucleic acid synthesis. Other examples for allosteric enzymes are phosphofructokinase, the main regulatory enzyme of the glycolytic pathway of carbohydrate degradation, and glycogen phosphorylase, which controls the mobilization of glycogen, the glucose (and energy) storage of the cell. Hemoglobin, a protein without catalytic activity, was the first example of an allosteric protein and is up to now the most thoroughly investigated representative of this group. Oxygen binding shows sigmoidal saturation behavior, which is allosterically influenced by 2,3-bisphosphoglycerate and by the Bohr effect, i.e., the regulation of oxygen binding by interaction with CO_2 and the proton concentration. The allosteric mechanism has been resolved by the brilliant work of Perutz (1989). Oxygen binds to the sixth coordination of an iron complex bound in a porphyrin ring system and causes a transition from the high-spin to the low-spin state of the iron. This transition is accompanied by movement of the iron into the plane of the porphyrin ring, which in turn triggers a conformational change of the whole protein molecule.

5. ISOLATION AND PURIFICATION OF PROTEINS

5.1 Cell Disintegration and Fractionation

The methods applied for the isolation of proteins or enzymes depend on the respective organism and tissue from which a particular protein should be extracted and also on its localization in the cell. Material must be available in sufficient quantities and must contain considerable amounts of the pertinent protein. Microorganisms are easy to cultivate. Their proteins can be specifically modified by mutation, which is very helpful for the investigation of protein functions. Microbial proteins are often closely related to corresponding proteins from higher organisms with respect both to structure and to function and can serve as valuable models. Genetic engineering allows the transfer of genes from higher organisms into microorganisms to isolate the respective proteins. For direct isolation of proteins from higher organisms, plants, laboratory animals like mice, rats, and guinea pigs, or tissues like heart, brain, kidney, or liver from the slaughter are taken. The preparation must be performed in *p*H-stabilized buffered solution with defined ionic strength in the cold room to prevent denaturation or degradation of the protein.

The initial isolation step is the disintegration of the tissue to obtain crude extracts. Tissues from higher organisms like liver, muscle, or kidney are minced into small particles with mincing machines, homogenizers, or fast blenders. The cell walls, especially of microorganisms, are disintegrated also by ultrasonication, digestion with degradative enzymes, detergents, or osmotic shock. Cell compartments like mitochondria, ribosomes, or nuclei are separated from cell homogenates by differential centrifugation. Membrane-bound proteins must be solubilized with detergents (Harris and Angal, 1989).

5.2 Precipitation and Concentration

The principle of protein precipitation is the disorganization of the hydration shell around the protein molecule. Ammonium sulfate or polyethylene glycol are efficient precipitants. Proteins precipitate at characteristic concentrations of the precipitant (*fractional precipitation*). Precipitation by *p*H changes makes use of the fact that proteins possess their lowest solubility at their isoelectric point (cf. Sec. 3.4). Precipitation at higher temperature is applied to separate thermostable proteins from the thermosensitive bulk proteins, especially after cloning of the gene of a thermophilic protein into a mesophilic microbial host. Ammonium sulfate and polyethylene glycol precipitation is also a method for concentrating proteins.

Dialysis of protein solutions against semipermeable membranes, which can be penetrated only by small molecules and not by macromolecules like proteins, is used to remove small impurities like metabolites, salts, or precipitants or to exchange buffer systems. The protein solution is included into a dialysis bag consisting of a cellulose membrane. The bag is transferred into a vessel containing a large volume of the desired buffer solution and gently stirred for several hours during which the outside solution is exchanged several times to remove traces of the contaminants. Special apparatus for dialysis of small volumes have been developed. Dialysis can also be used for protein concentration whereby the dialysis bag is embedded into hygroscopic substances (e.g., aquacide), which draws the water out of the bag.

Ultrafiltration is a technique for concentration of dilute protein solutions. The solutions are pressed through semipermeable membranes either mechanically or applying the pressure of compressed gasses, centrifugal force, or vacuum. Membranes of different materials (cellulose, nitrocellulose, etc.) and pore sizes are used and serve also for a rough separation of proteins. To concentrate large proteins, sedimentation in the gravitational field of an *ultracentrifuge* can be taken. Applying differential gradient centrifugation allows separation of proteins according to their size (cf. Sec. 6.5).

5.3 Chromatographic Techniques

The principle of column chromatography is the separation of protein mixtures on matrices in long columns. The column dimensions depend on the amount of protein to be separated, the binding capacity of the material, and the accuracy of separation of different protein bands (Scopes, 1987).

Ion-exchange chromatography is a powerful method for protein isolation. It makes use of the fact that the protein surface carries a large number of charged groups. DEAE (diethylaminoethyl) and QAE (diethyl-[2-hydroxypropyl]aminoethyl) groups as anion exchangers and CM (carboxymethyl) or SP (sulfopropyl) groups as cation exchangers fixed to cellulose, dextran, agarose, or synthetic polmer matrices are applied. Protein adsorbed to the exchangers is eluted by buffer or salt gradients with increasing ionic strength or *p*H change.

Adsorption chromatography is related to ion chromatography. It relies mainly on nonspecific dipole-dipole interactions between protein and the matrix. Inorganic materials like calcium phosphate or hydroxy-apatite serve as matrices.

Affinity chromatography makes use of the fact that proteins and especially enzymes possess specific binding sites for substrates and cofactors (Jakoby and Wilchek, 1974; Dean *et al.*, 1985). Such ligands are covalently attached to matrices like dextrans, agarose, cellulose, or polyacrylamide. The matrix has to be activated beforehand with cyanogen bromide or epoxide and the ligand is fixed either directly or via spacer molecules like 1,6-diaminohexane to confer high mobility onto the fixed ligand for the interaction with the protein. Besides enzyme substrates and coenzymes, also binding proteins like antibodies or lectins are used. Affinity chromatography is a very efficient purification method since only proteins that specifically interact with the fixed ligand are retarded. Elution is achieved with a solution of the unbound ligand.

Hydrophobic chromatography with alkyl (e.g., octyl), aminoalkyl, or aryl (e.g., phenyl) residues immobilized to matrices like agarose is based on interactions of these groups with hydrophobic regions on the protein surface. Polar conditions like high ionic strength favor binding while a decrease in polarity (lowering of ionic strength, nonpolar solvents) promote elution (Ochoa, 1978).

Gel filtration is a method to separate macromolecules according to their size. The dextran, agarose, or polyacrylamide matrices consist of porous particles with defined pore diameters. Smaller molecules penetrate the particles and are retarded on their way across the column while larger molecules pass the column directly. With this method, molecular masses of proteins also can be estimated.

The efficiency of column chromatography has been highly improved with respect both to the speed and to the quality of separation by the high performance liquid chromatography (HPLC) and the fast protein liquid chromatography (FPLC) techniques. The principle of both methods is the operation of column chromatography at high pressure (up to 400 bar for HPLC). HPLC has a broad spectrum of applications from small compounds to macromolecules, while the FPLC technique has been developed especially for purification of sensitive compounds like enzymes. The separation principles (ion exchange, gel filtration, hydrophobic chromatography) are equivalent to those of conventional chromatography, only the matrices are stabilized to endure the high-pressure conditions. One column run lasts some minutes compared to hours and days for conventional chromatography. Very sharp bands are obtained with high reproductivity (Rossomando, 1987; Mant and Hodges, 1991).

6. TECHNIQUES FOR ANALYSIS AND STRUCTURE DETERMINATION OF PROTEINS AND ENZYMES

6.1 Electrophoretic Techniques

Electrophoresis serves mainly as a control for the purity of protein preparations and for molecular mass determinations. Preparative electrophoresis is rarely used because of its complicated manipulation. The migration velocity of proteins in the electrical field depends on their net number of charges and, inversely, on their size (Stokes radius). Electrophoresis is performed with buffered protein solutions on matrices like agarose or polyacrylamide. Frequently, the electrophoresis is carried out in the presence of sodium dodecyl sulfate (SDS). This detergent associates to the protein molecule and unravels its secondary and tertiary structure so that it dissociates into its subunits. The negative charge of the bound SDS molecules exceeds the net charge of the protein, and migration to the anode is determined essentially by the protein size because of retardation within the matrix network (Hames and Rickwood, 1990).

Isoelectric focussing makes use of the fact that the net charge and thus the migration of a protein at its isoelectric point (*pI*) is zero. A *p*H gradient is created in an electric field by special amphoteric polymeric compounds, the ampholytes. Proteins migrate to the position of their respective *pI*'s, which can be determined by this method (Righetti, 1975; Chrambach, 1985).

Isotachophoresis is a type of electrophoresis that is performed at weakly alkaline *p*H where most proteins are negatively charged and migrate to the anode. The proteins migrate between a leading anion with fast mobility like phosphate and a terminator anion with low mobility like glycine in a high electric field. The anions in the sample arrange in the order of their mobility (Everaerts *et al.*, 1976; Righetti, 1975).

6.2 Protein Determination and Enzyme Tests

A large number of methods exist for the determination of the amount of protein in extracts and solutions. They are mainly based on the formation of colored complexes as in the biuret and the Lowry assay. The Bradford method makes use of the adsorption of the dye Coomassie brilliant blue to proteins. The absorption of aromatic amino acids, especially of tryptophan at 280 nm, is used both for quantitative protein determination and for detection of protein bands in chromatographic techniques.

Enzymes in extracts, solutions, or eluates from columns are identified by specific enzyme tests based on their catalytic reactions. The amount of product formed or of substrate converted as a function of time is measured, and the enzyme activity is expressed in katal (kat) as the amount of enzyme forming 1 mol product within 1 s (according to the SI nomenclature) or in enzyme units (U), the amount of enzyme producing 1 μmol product per minute. Enzyme activity divided by the protein concentration of the sample, the *specific enzyme activity*, is a measure for the purity of enzyme preparations. Enzyme tests must be performed at distinct standard conditions with respect to temperature, *p*H value, and ionic strength. Since enzymes differ considerably in their features, general standard conditions cannot be defined. Usually, enzymes are tested at 25 °C and at neutral *p*H (7.0–7.6), stabilized by a buffer system with moderate ionic strength (~0.1 M).

Enzyme reactions are preferentially followed spectrophotometrically or fluorimetrically if there is any detectable spectroscopic difference between substrate and product. Representative examples are the nicotinamide adenine dinucleotide (NAD) dependent dehydrogenases. The reduced form of this cofactor, NADH, possesses an additional absorption peak at 340 nm compared with the oxidized form and shows also considerable fluorescence emission. Reactions that proceed without spectral changes may be coupled with photometrically detectable reactions. In other cases, the reaction is stopped after a defined time, and the product produced or the remaining substrate is determined by chemical analysis. Also, the HPLC method is taken for the quantitative identification of the reactants.

6.3 Amino Acid Analysis and Protein Sequencing

The amino acid composition is determined after acid or basic hydrolysis of the respective protein or after its enzymatic hydrolysis with specific proteases. The amino acids from the hydrolysate are separated by chromatographic techniques, e.g., ion-exchange chromatography, HPLC, or gas-liquid chromatography. They are detected after conversion into colored derivatives (e.g., with ninhydrine or fluorescamine). Automatic amino acid analyzers have been developed (Needleman, 1970).

Some information about the occurrence of special amino acids and especially about their functional and structural roles in distinct proteins can be obtained from specific amino acid modifications. There exist different reagents reacting with distinct amino acids, like the sulfhydryl reagent 5,5′-dithio-bis-2-nitrobenzoic acid (DTNB, Ellman's reagent), the fluorescent amino group reagent 5-dimethylaminonaphthalene-1-sulfochloride (DANSYL chloride), or the tryptophan reagent 2-hydroxy-5-nitrobenzyl bromide (Koshland's reagent). In native proteins, only such amino acid residues react as are accessible by the reagent, and from this, information about their localization in the protein molecule can be gained.

Proteins from microorganisms can be modified by mutation. These are induced by special mutagens like bromouracil, diethylsulfate, or UV and x-ray radiation. With this method, amino acid exchanges, breaks, or deletions of parts of the polypeptide chain are created, and the changes in protein structure and function can be studied. Defined exchanges of special amino acids, e.g., in binding regions or catalytic centers, are obtained by the method of site-directed mutagenesis. Short oligonucleotides with the nucleotide sequence of the respective gene region including the desired mutation are synthesized with an automatic DNA synthesizer. This fragment hybridizes with a single-stranded DNA vector containing the original gene, and a DNA polymerase completes the oligonucleotide to the complementary DNA strand. The resulting mutant gene is transformed into a microbial host like *Escherichia coli* that expresses the modified protein. Site-directed mutagenesis allows introduction of any desired amino acid exchange. In combination with computer modeling, predicted changes—e.g., altering of binding specificity or functions of proteins—can be realized.

The amino acid sequences of proteins are resolved by successive splitting and determination of the terminal amino acids. There exist several methods for determination of both the amino terminal and the carboxy terminal residues of polypeptides, such as the conversion with phenylisothiocyanate (PITC, Edman's reagent) or fluorodinitrobenzene (Sanger's reagent). Specific enzymes, the aminopeptidases and the carboxypeptidases, cleave amino acid residues from the amino and the carboxy terminus, respectively (Needleman, 1970; Hugli, 1989). Protein sequences can be derived also from the DNA sequence of the respective genes. After cloning and isolation by methods of genetic engineering (e.g., the polymerase chain reaction, PCR), the gene is sequenced according to the Sanger or Maxam–Gilbert technique (Alberts *et al.*, 1989).

6.4 Determination of the Three-Dimensional Structure

The most powerful method to resolve the three-dimensional structure of a protein is x-ray crystallography. Electron-density maps are obtained from scattering of x rays or of synchrotron radiation by a protein crystal. For the interpretation of the density maps, a strongly scattering atom like a heavy-metal atom should be introduced into the protein molecule, e.g., by modification of side chains with mercurials. The density data are treated by Fourier transformation, and, in combination with the amino acid sequence, spatial arrangement of the polypeptide chain and association of protein subunits can be constructed (Cantor and Schimmel, 1980).

With other physical methods, only fragmentary information rather than a complete picture of the protein structure is obtained. However, such methods allow the study of proteins in solutions and especially of dynamic processes such as interaction with ligands and conformational changes. α helix, β-pleated sheet, and random coil, the major elements of the secondary structure of proteins, show characteristic circular dichroism (CD) spectra. CD spectroscopy serves to estimate the relative portion of these structures in proteins and allows the observation of changes of these structures. Information about secondary structure elements is revealed also from far-UV absorption spectra (200–220 nm). Especially aromatic amino acids contribute to the UV spectra of proteins in the range between 200 and 340 nm. Thus, the UV spectra of proteins give information about amino acid content, structure, and interaction with ligands. More sensitive is fluorescence spectroscopy, where the intrinsic protein fluorescence of aromatic amino acids can be taken as a specific signal from distinct positions of the protein molecule. Furthermore, amino acid residues can be labeled with fluorescent chromophores like fluorescein or dimethylaminonaphthalene as probes. With the method of fluorescence energy transfer, distances between different sites at a protein molecule can be estimated by introducing a pair of chromophores as donor and acceptor at the respective sites. The excitation energy of the donor is nonradiatively transferred to the acceptor, and the transfer efficiency is a measure of the distance. The size and shape of macromolecules can be obtained by fluorescence polarization and anisotropy measurements. Fluorescence is excited with linearly polarized light, and the polarization of the emitted light is measured as a ratio of the fluorescence intensities in and perpendicular to the plane of polarization. Depolarization occurs in dependence on molecular motion and thus the size of the molecule. Independent movements of distinct regions or groups on the protein surface can be detected (Brand, 1967).

Valuable information about the protein structure is obtained from electron paramagnetic resonance (EPR) and nuclear magnetic resonance (NMR) spectroscopy. EPR is an extremely selective technique to detect unpaired, paramagnetic electrons, whose magnetic moment orients in an external magnetic field, and the resonance absorption of a microwave radiation is measured as a function of the magnetic field. Proteins themselves are not paramagnetic and give no EPR signal. Paramagnetic electrons appear in oxidative processes, e.g., in iron-sulfur clusters in the respiratory chain. Metal ions bound as cofactors to various proteins, like manganese, copper, or cobalt, give signals in the EPR spectroscopy. A large number of stable spin labels are available, which can be covalently attached to specific groups of the protein molecule and which act as reporter groups giving information about interactions in the environment of this region. Most frequently, the nitroxide radical stabilized as tetramethylpiperidinyl-1-oxyl (TEMPO) and coupled to a site-specific reactive group is used.

The NMR spectroscopy deals with the orientation of the nuclear magnetic moment in an external magnetic field. Transitions between distinct energy levels are induced by a perturbing field of a microwave oscillator, and the absorption of the energy of this field is determined. NMR signals in proteins are observed mainly with the isotopes ^{1}H, ^{13}C, and ^{31}P. Proton NMR gives information about H–C, H–N, and H–O bonds. Because of the large number of such bonds in a protein molecule, the picture is very complex and hard to unravel. Measurements must be performed in D_2O to eliminate the strong H_2O resonance. More specific signals are obtained by ^{13}C NMR. The intensities, however, are lower than with proton NMR, and the isotope must be incorporated into the protein (Freifelder, 1976).

6.5 Molecular Mass Determination

Molecular masses of polypeptide chains and protein subunits can be calculated from their amino acid sequences or from the DNA sequence of the respective genes. This gives, however, no information about the subunit composition of the protein. Similar information is obtained, in the absence of sequence data, from electrophoresis in the presence of the detergent sodium dodecylsulfate on polyacrylamide or agarose gels. The detergent adsorbs the protein, which dissociates into its subunits, and the negative charge of the detergent determines the migration rate of the protein subunits. In the narrow network of the gel, the protein becomes retarded according to its size, and in comparison to known standard proteins, the relative molecular masses of the protein subunits can be estimated.

A simple, but not very accurate, method for the determination of molecular masses of native proteins in their aggregated state is gel filtration chromatography (see Sec. 5.3). A more accurate technique is the analytical ultracentrifuge. Large proteins migrate in the gravitational field of an ultracentrifuge. The sedimentation velocity v depends on the molecular mass M_r, the partial specific volume $\bar{v}$ of the protein, the frictional coefficient f, and the density ρ of the solution:

$$\frac{M_r(1 - \bar{v}\rho)}{Nf} = \frac{v}{\omega^2 r} = s,$$

where N is Avogadro's number, ω is the angular velocity, and r is the distance from rotation center. The sedimentation coefficient s (usually expressed in Swedberg units $S = 1 \times 10^{13}$ s) is a characteristic value for the macromolecule from which the molecular mass can be deduced. The sedimentation behavior of the protein is followed during the centrifugation run with special ultracentrifugation cells equipped with observation windows. The sedimentation of the sample is observed by UV absorption, or Schlieren or interference optics. Two different ultracentrifugation techniques are applied. The *sedimentation velocity method* registers the velocity v of the particle in the gravitational field. From v, r, and the angular velocity ω, the sedimentation coefficient s is obtained. More accurate is the *sedimentation equilibrium method*. During sedimentation, the particles form a boundary moving to the bottom of the ultracentrifugation cell. This migration is opposed by diffusion. The two effects are counterbalanced at equilibrium, and the boundary remains at a constant position. From the steepness of the boundary, the molecular mass can be calculated.

Molecular masses of proteins and enzymes even in crude preparations can be estimated by the technique of differential gradient centrifugation. In the ultracentrifuge tube, a linearly increasing concentration gradient of a dense medium like sucrose is prepared before centrifugation and the protein sample is applied onto the top of the gradient. In the gravitational field, the protein migrates according to its mass as a single band. From the migration distance of the protein band to the top of the gradient, the sedimentation coefficient and the molecular mass of the sample can be calculated in relation to known standard proteins.

During recent years, mass spectroscopy (MS) has become increasingly important also for proteins. The protein is ionized by fast-atom bombardment, and molecular masses are determined with high accuracy up to 400 000 without need for reference proteins for calibration. The structure of fragments from the protein produced by collisions with small neutral molecules can be analyzed in tandem mass spectrometers and gives information about structural aspects of the protein. Bombarding is done with argon or xenon atoms or cesium ions with kinetic energies of 8–35 keV for peptides and smaller proteins. For larger proteins, the fission products of ^{252}Cf with kinetic energies of 80 MeV are taken. The proteins are adsorbed on a nitrocellulose layer (plasma-desorption MS) and the time of flight (TOF-MS) of the ions is determined. For laser-desorption MS, the protein is embedded in an ultraviolet-absorbing matrix and ionized by a pulsed ultraviolet laser. With the electrospray method ionization is achieved by spraying the protein solution across an electrostatic field gradient. Besides molecular masses of proteins, changes due to mutations or modifications like glycosylation can be detected. Collision-induced decomposition (CID) by helium, argon, or xenon is used for sequencing of peptides and proteins with a tandem MS. The peptides are scanned in a first mass spectrometer, a selected beam passes a collision cell, and the resulting fragments are analyzed in a second mass spectrometer (Biemann, 1992).

Information about the size and the structure of proteins can be obtained also by electron microscopy. Especially for large protein molecules and aggregates like multienzyme complexes, the overall molecular shape and the subunit composition become visible. To create contrasts for the electron beam, the proteins must be pretreated with heavy metals like uranyl acetate or phosphotungstate, which, however, may cause some damage to the native conformation of the protein. A more gentle treatment is the shadowing of freeze-dried samples with platinum and carbon, though the resolution is poorer. An improvement of the resolution of electron micrographs and the elimination of artifacts are achieved by computer-aided image reconstruction, where a series of images of the molecule in the same projection is superimposed.

6.6 Measurement of Fast Reactions

For the study of extremely fast processes in catalysis, ligand binding, and conformational mobilities of proteins and enzymes, special techniques have been developed: flow or rapid mixing techniques and relaxation methods (Hiromi, 1979). The principle of the flow techniques, which have been developed by Chance (1974), is the minimization of the mixing time of the reactants, which is usually the time-limiting process for conventional measurements. Plungers of two syringes containing the reactant solutions are simultaneously compressed by hydraulic or mechanical forces, and the solutions are forced through narrow nozzles into a mixing chamber. In the *continuous-flow apparatus,* the reaction mixture flows at constant rate through an observation tube. During the flow, the reaction proceeds, and its age becomes proportionally to the distance from the mixing chamber. To follow the reaction course, the absorption or the conductivity of the reaction mixture is measured at different points of the observation tube. The disadvantage of this method is the need of large sample volumes.

For the *stopped-flow apparatus,* the observation tube is replaced by a reaction cell, and the drive of the plungers is stopped when the cell is filled with the reaction mixture (Fig. 9). Simultaneously, a trigger is released to start the detection. Most frequently, optical methods like absorption or fluorescence spectroscopy are applied. The signal is registered by a storage oscillograph or fast computer storage. The resolution time of the flow methods lies within the range of 1 ms.

The principle of the relaxation methods, developed by M. Eigen (Eigen and De Mayer, 1974), is the disturbance of an equilibrium by a rapid change of environmental parameter and the observation of the relaxation of the system to the new equilibrium conditions. Since the equilibrium constants depend on temperature, pressure, and field strength, distinct apparatus for each parameter have been constructed. For biological studies, the *temperature-jump apparatus* is most convenient. An observation cell contains the sample solution between two platinum electrodes. A high tension is applied and discharged by a spark gap. A current surge crosses the observation cell and ele-

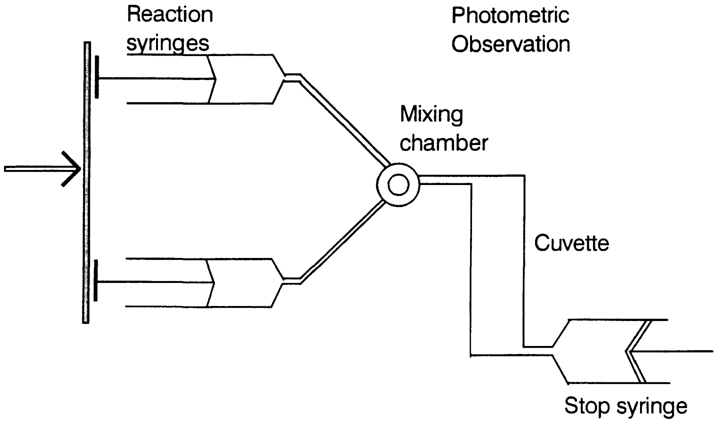

FIG. 9. Schematic drawing of a stopped-flow apparatus.

vates the temperature by some degrees by Joule heating within about 1 μs. A trigger starts the optical detection.

Instantaneous expansion of high pressure in the sample solution is the principle of the *pressure-jump technique.* Up to 100 atm are applied to the sample through compressed air. Rupture of a metal diaphragm with a steel needle allows the air to escape and releases the trigger. The time resolution is between 20 and 100 μs. Since high reactant concentrations are needed to obtain detectable signals, the method is less suited for enzymatic studies. Nearly no application for protein and enzyme studies is made of the *electric field-jump technique.* High electric fields are alternately applied and removed to the observation cell within short time intervals, and the relaxation of the sample to equilibrium is observed. Because of the considerably high field strength, the sample warms up within a short time, causing denaturation of proteins.

6.7 Peptide and Protein Synthesis

For the synthesis of a distinct protein, the amino acids must be arranged and combined in a defined sequence. The first step is the protection of amino acid side chains, which should not take part in the reaction, by selective protection groups. N-Benzyloxycarbonyl (cbo, cbz, Z), N-*tert*-butoxycarbonyl (boc), 4-toluenesulfonyl (tosyl), and triphenylmethyl (trityl) residues are protection groups for amino residues, and benzyl and *tert*-butyl residues for carboxy groups. These groups must be removed from the protein after the synthesis without impairing the peptide bonds and other amino acid side chains. A second prerequisite for protein synthesis is an extremely high degree of purity of the amino acids used, since even minute contaminations accumulate during the synthesis cycles of the growing protein molecule and cause incorrect incorporation of amino acids. For the same reason, completion of the reaction at each step must be assured. These problems limit the length of artificially synthesized polypeptide chains. Larger proteins must be constructed by combination of shorter fragments.

Two different principles for peptide synthesis exist: the conventional peptide synthesis in solution and the solid-phase technique. Synthesis in solution has the disadvantage that the reaction intermediate must be isolated after each cycle. Therefore, the development of the *solid-phase technique* by R. B. Merrifield in 1963 was a valuable improvement (Erickson and Merrifield, 1976). The growing peptide chain is fixed to solid matrices like polystyrene, and the amino acids are linked to it in standardized reaction cycles. Automatic synthesizers are constructed for this technique. A disadvantage is that purification of intermediates is impossible.

7. TECHNICAL APPLICATIONS OF PROTEINS AND ENZYMES

7.1 Immobilization of Proteins and Enzymes

Enzymes as very efficient catalysts are highly recommended for technical applications. A serious obstacle, however, is their instability and the fact that the extremely valuable enzyme material becomes lost within the reaction batch, and reisolation from there is unprofitable. This problem can be overcome at least for some special enzymes of industrial interest by immobilization to solid supports. These can be removed from the reaction batch and regenerated for the next process. Immobilization is also applied for proteins with other functions like antibodies. Different immobilization techniques are used: entrapment into microcapsules, fixation to solid supports, and cross-linking.

Entrapment is the most gentle immobilization technique since the proteins retain their native structure. Proteins have been included into microcapsules formed by polymerization of nylon or nitrocellulose in an organic phase of an emulsion with the aqueous protein solution. Also, microcapsules consisting of phospholipid or cholesterol liposomes are applied. Protein entrapment is also possible in hollow fibers. These are long narrow tubes (about 1-mm diameter) of semipermeable cellulose or synthetic polymer membranes. They provide large surfaces for the exchange of the entrapped protein with the surrounding solution. Polymer lattices like polyacrylamide or agarose are further used for protein entrapment. A disadvantage of protein entrapment, however, is the limitation of free diffusion of the reaction components from the outside solution to the entrapped protein. Steric obstacles, charges, or polarity may limit both the diffusion of the substrate to its enzyme and that of the product outside, so that inside the microcapsule, a region of substrate depletion and product accumulation will be generated. Since most enzyme reactions are inhibited by larger product concentrations, this effect may cause a severe activity decline.

Fixation to *solid supports* can be achieved by noncovalent adsorption of the polyionic proteins to anion or cation exchangers under defined conditions of ionic strength and *p*H value. The proteins keep their native structure and can be stripped off from the matrix by change of *p*H or ionic strength after the reaction process. Covalent attachment of proteins to solid matrices yields stable adducts but may influence the native protein structure and function. As matrices dextrans, cellulose, synthetic polymers, silanized glass, and ceramics are used. They must be activated to introduce reactive groups for the covalent protein fixation. Cyanogen bromide, epoxide, or transition-metal salts like $TiCl_4$ or $SnCl_4$ are applied for activation. The protein is linked to the activated matrix either directly or via spacer molecules like hexamethylene diamine or glutardialdehyde to prevent steric hindrance between support and protein.

Cross-linking of proteins with bifunctional reagents like glutardialdehyde without supports yields large, insoluble aggregates, which can easily be separated from the solution after reaction. The cross-linking can be done either with homogeneous proteins or by co-immobilization between the sample and an inert stabilizing protein, like serum albumine.

If instability of an enzyme allows no immobilization, whole cells containing considerable activities of the pertinent enzyme are immobilized. An example is the transformation of steroids by bacterial cells used for synthesis and modification of drugs. Entrapment techniques or covalent linkage are applied in this case. Mammalian cells immobilized as monolayer cultures to solid supports with glycoproteins are used for hormone detection.

7.2 Enzyme Reactors

To perform reaction processes with immobilized enzyme on industrial scales, special reactors have been constructed. In the *stirred-tank reactor,* the immobilized enzyme particles are stirred in a vessel together with the substrate. After the reaction, the enzyme particles are collected by filtration or centrifugation. The procedure is laborious, but it is applicable to viscous and particulate substrates, which may block membranes of other reactor types. The stirred-tank reactor can be run continuously by constant pumping of the substrate solution into the reaction vessel and withdrawing the reaction product through a membrane. The efficiency of such reactions can be increased by packing the immobilized enzyme into columns. The substrate solution flows up or down the column (*fixed-* or *packed-bed reactor*). The *fluidized-bed reactor* is a combination of the two principles. The flow rate of the upward stream of the substrate solution balances the sedimentation of the enzyme particles.

Enzyme reactors are applied for amino acid synthesis, isomerization of glucose to fructose, production of wort from unmalted barley in brewing, antibiotic processing, and drug synthesis and modification. Immobilized systems have been developed for the regeneration of coenzymes that are involved in the reactions, which may otherwise be lost during the process (Woodward, 1985).

7.3 Biosensors

Biosensors (enzyme electrodes) gain increasing importance in industrial processes like fermentation and bioreactors, where they control the reaction progress. In medicine, biosensors are used for distinct diseases to detect metabolites in serum and urine. Biosensors consist of a biological system, usually an enzyme, entrapped or fixed to a support and shielded from the surrounding liquid by a membrane permeable for the enzyme substrate. This compartment is connected to a transducer. The transducer is an electrode sensitive for oxygen, carbon dioxide, ions, or *p*H, or an optoelectrical detector combined with a field-effect transistor. It transduces the enzyme reaction, e.g., oxygen formation or *p*H change, into an electrical signal, which is monitored after amplification. Biosensors are used to determine glucose (e.g., in the blood), urea (in the urine), amino acids, cholesterin, antibiotics, and peroxides (Turner, 1987). Enzyme thermistors monitor the heat generated in the course of an enzyme reaction. The substrate solution passes a column containing the immobilized enzyme, and the temperature of the eluate is measured with a thermistor. Enzyme thermistors allow the detection of extremely small metabolite concentrations.

7.4 Medical Applications

Proteins and especially enzymes play an important role in medicine. In the course of different diseases and metabolic disorders, the level of distinct enzymes in the organism changes. The variation of enzyme and metabolite levels is tested in enzyme diagnostics in the blood or in tissue samples to gain information about the type and course of the disease and also for prophylaxis. Inborn errors like phenylketonuria or glycogen storage diseases can be identified in this way. Inhibition of enzymes by specific inhibitors acting as substrate analogs is widely used in therapy, e.g., the reduction of cholesterol synthesis to prevent formation of arterial plaques. Various drugs and toxic substances like heavy metals and cyanide block specifically enzyme reactions or other protein functions like hormone receptors (e.g., beta blockers). In special cases, direct administration of enzymes, especially into the blood, serves in therapy (*enzyme therapy*) to remove toxic substances. Blood clots are dissolved by the administration of the enzymes streptokinase and urokinase. In tumor therapy, the enzyme asparaginase is used, which cleaves asparagine, a metabolite essential for some tumours.

7.5 Protein Engineering and Artificial Enzymes

Site-directed mutagenesis allows the exchange of special amino acids in proteins. In combination with computer modeling, this is a powerful tool to create proteins and enzymes with modified features with respect to substrate specifity, inhibitor sensitivity, thermal stability, etc. With special computer programs, the three-dimensional structure of an enzyme (so far as known from x-ray crystallography or other methods) can be visualized on a screen, and distinct amino acid residues can be exchanged and adapted to the overall protein structure, like bulky or charged residues. For example, catalytic centers can be constructed to accept alternative substrates. These predictions are tested and verified by site-directed mutagenesis, which realizes the projected amino acid exchanges. In this way, the substrate specificity of the enzyme xylose isomerase has been changed from xylose to glucose (Meng *et al.,* 1991).

A recently developed method to create artificial enzymes is the construction of catalytic antibodies. This technique is based on the assumption that the principle of enzyme catalysis is the stabilization of the transition state of a reaction. Monoclonal antibodies are raised against stable transition-state analogs to form complementary structures at its antigen binding sites. Several examples of reactions catalyzed by catalytic antibodies have already been demonstrated (Benkovic, 1992).

Less promising are attempts to construct artificial enzymes by chemical synthesis (synzymes). As models of binding sites for substrate molecules, cyclic compounds with a hole in their middle are chosen, like cyclodextrines and crown ethers with different diameters. Functional groups to perform general acid/base catalysis and other mechanisms are introduced. Reactions like hydrolysis, proton transfer, acyl transfer, proceeding partially stereospecific, could be demonstrated. The reaction rates of artificial enzymes, however, remain far behind those of natural enzyme reactions (Komiyama and Bender, 1984; Stoddart, 1984).

GLOSSARY

Cell Compartments: Particles and organelles within the living cell representing distinct functional regions: the nucleus as carrier of the genetic information, the mitochondria responsible for energy generation and cell respiration, the chloroplasts as the place of photosynthesis, and the lysosomes, which concentrate degradative activities.

Cellular Respiration: Process of energy generation in the cell. Hydrogen from various reductive processes is transmitted to molecular oxygen in a multistep reaction by the mitochondrial respiratory chain. The energy released by this reaction is stored in the form of energy-rich phosphoric acid anhydride bonds.

Coenzymes (Cofactors): Nonprotein portion of enzymes that supports the catalytic process. Coenzymes are often vitamins or metal ions.

Cytoplasm: Total cell contents besides the cell organelles. Many essential cell functions like protein biosynthesis and various metabolic pathways occur in the cytoplasm. The soluble fraction of the cytoplasm is the cytosol.

Enzymes: Biocatalysts controlling all reactions within living organisms. Enzymes are protein molecules, partly associated with a nonprotein portion, the coenzyme or prosthetic group.

Hormones: Biologically active substances produced by distinct organs and acting specifically on other organs, tissues, or cells. Hormones are important regulators of development, growth, and metabolism of organisms.

Immobilized Enzymes: Enzymes fixed to solid supports or entrapped in microcapsules. They serve as catalysts in biotechnical processes.

Ligands: Compounds, mainly of low molecular mass, that specifically interact with macromolecules, e.g., enzyme substrates or hormones binding to receptors.

Metabolism: Sum of all (usually enzyme-catalyzed) reactions occurring in the living cell.

Metabolites: Compounds that participate in the cell metabolism as substrates and products of enzyme reactions, as intermediates of metabolic pathways, and as precursors of biological active substances. Metabolites act often as regulators of metabolic pathways or related processes like gene regulation.

Mitochondria: Large organelles of eukaryotic cells executing various essential functions of the cell. Mitochondria possess an outer and an inner membrane. The inner membrane carries the major components of the respiratory chain. The inner mitochondrial space, the matrix, contains important enzymes for energy metabolism and fatty acid degradation.

Nucleotides: Building units of the nucleic acids consisting of a heterocyclic base, a sugar (ribose or deoxyribose), and phosphoric acid.

Nucleic Acids: Large nucleotide polymers, which act as carriers of the genetic information and are responsible for the translation of this information into functional components of the cell like proteins.

Oligosaccharides: Polymeric molecules consisting of sugar units, like glucose. Oligosaccharides serve as energy storage (starch, glycogen) and have structural functions (cellulose, cell wall components).

Prosthetic Groups: Nonprotein components covalently bound to catalytic centers of enzymes and essentially involved in the catalytic mechanisms.

Proteins: Macromolecules consisting of polymeric chains from different α-amino acids arranged in distinct sequences. Proteins are involved in all essential functions of living organisms like cell metabolism, energy generation, gene reproduction, and translation, motility, sensory, and structural functions. The most important group of proteins are the enzymes.

Reaction Order: Classification of chemical reactions according to the number of reactants involved; e.g., reaction of two substances is second order.

Receptors: Regions at the surfaces and membranes of cells that interact with specific signal substances like hormones or neurotransmitters and transduce the signal into the cell. Receptors are mainly built from membrane proteins.

Ribosomes: Cellular particles that are responsible for the translation of the genetic information encoded in nucleic acids into the primary sequence of proteins (protein biosynthesis). Ribosomes consist of proteins and nucleic acids.

Second Messengers: Cellular substances accepting signals from hormones and neurotransmitters from the outside of the cell and causing the transformation of the messages into cellular functions. The most important second messenger is cyclic AMP.

Works Cited

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J. (1989), *Molecular Biology of the Cell,* New York: Garland Publ.

Benkovic, S. J. (1992), *Annu. Rev. Biochem.* **61,** 29–54.

Biemann, K. (1992), *Annu. Rev. Biochem.* **61,** 977–1010.

Brand, L., Witholt, B. (1967), in: C. H. W. Hirs (Ed.), *Enzyme Structure,* Methods in Enzymology Vol. 11, New York: Academic, pp. 776–856.

Chance, B. (1974), in: A. Weissberger (Ed.), *Techniques of Organic Chemistry,* Vol. 6, Part II, New York: Wiley, pp. 5–62.

Chou, P. Y., Fasman, G. D. (1978), *Annu. Rev. Biochem.* **47,** 251–276.

Chrambach, A. (1985), *The Practice of Quantitative Gel Electrophoresis,* Weinheim: VCH.

Cleland, W. W. (1970), *The Enzymes* **2,** 1–65.

Dean, P. D. G., Johnson, W. S., Middle, F. A. (1985), *Affinity Chromatography: A Practical Approach,* Oxford: Oxford University Press.

Eigen, M., De Mayer, L. (1974), in: A. Weissberger (Ed.), *Techniques of Organic Chemistry,* Vol. 6, Part II, New York: Wiley, pp. 63–146.

Erickson, B. W., Merrifield, R. B. (1976), in: H. Neurath (Ed.), *The Proteins,* 3rd ed., Vol. 2, New York: Plenum, pp. 255–527.

Everaerts, F. M., Beckers, J. L., Verheggen, T. P. E. M. (1976), *Isotachophoresis, Theory, Instrumentation and Applications,* Amsterdam: Elsevier.

Florkin, M., Stotz, E. H. (Eds.) (1972), *Comprehensive Biochemistry,* Vol. 30, Amsterdam: Elsevier.

Freifelder, D. (1976), *Physical Biochemistry,* San Francisco: Freeman.

Hames, B. D., Rickwood, D. (1990), *Gel Electrophoresis of Proteins. A Practical Approach,* Oxford: Oxford University Press.

Harris, E. L. V., Angal, S. (1989), *Protein Purification Methods. A Practical Approach,* Oxford: Oxford University Press.

Hiromi, K. (1979), *Kinetics of Fast Enzyme Reactions,* New York: Wiley & Sons.

Hugli, T. E. (Ed.) (1989), *Techniques in Protein Chemistry,* San Diego: Academic Press.

Jakoby, W. B., Wilchek, M. (1974), *Affinity Techniques,* Methods in Enzymology Vol. 34, New York, Academic.

Jencks, W. P. (1975), *Adv. Enzymol.* **43,** 219–410.

Komiyama, M., Bender, M. L. (1984), "Cyclodextrines as Enzyme Models," in: M. I. Page (Ed.), *New Comprehensive Biochemistry,* Amsterdam: Elsevier, pp. 505–527.

Koshland, D. E., Nemethy, G., Filmer, D. (1966), *Biochemistry* **5,** 365–385.

Mant, C. T., Hodges, R. S. (1991), *High Performance Liquid Chromatography of Peptides and Proteins: Separation, Analysis, and Conformation.* Boca Raton, FL: CRC Press.

Meng, M., Lee, C., Bagdsarian, M., Zeikus, J. G. (1991), *Proc. Natl. Acad. Sci. USA* **88,** 4015–4019.

Monod, J., Wyman, J., Changeux, J.-P. (1965), *J. Mol. Biol.* **12,** 88–118.

Needleman, S. B. (1970), *Protein Sequence Determination,* Berlin: Springer Verlag.

Ochoa, J. L. (1978), *Biochimie* **60,** 1–15.

Perutz, M. (1989), *Q. Rev. Biophys.* **22,** 139–236.

Pfanner, N., Neupert, W. (1990), *Annu. Rev. Biochem.* **59,** 331–353.

Righetti, P. G. (1975), *Isoelectric Focusing and Isotachophoresis,* Amsterdam: Elsevier Publ.

Rossmann, M. G., Argos, P. (1981), *Annu. Rev. Biochem.* **50,** 497–532.

Rossomando, E. F. (1987), *High Performance Liquid Chromatography in Enzymatic Analysis. Application to the Assay of Enzyme Activity.* New York: Wiley.

Scopes, R. K. (1987), *Protein Purification, Principles and Practice,* New York: Springer-Verlag.

Stoddart, J. F. (1984), "Crown Ethers as Enzyme Models," in: M. I. Page (Ed.), *New Comprehensive Biochemistry,* Amsterdam: Elsevier, pp. 529–561.

Turner, A. P. F. (1987), *Biosensors. A Practical Approach,* Oxford: Oxford University Press.

Woodward, J. (1985), *Immobilized Cells and Enzymes. A Practical Approach,* Oxford: Oxford University Press.

Further Reading

Bergmeyer, H. U. (Ed.) (1983–1986), *Methods of Enzymatic Analysis,* Weinheim: Verlag Chemie. Extensive encyclopedia on enzymology and enzyme tests. Detailed description of the different enzyme classes.

Cantor, C. R., Schimmel, P. R. (1980), *Biophysical Chemistry,* San Francisco: Freeman. The fundamental book of biophysical chemistry with brilliant and detailed descriptions of macromolecular and protein structure and behavior and an extensive presentation of biophysical methods.

Kornberg, H. L. (1978), *Techniques in Protein and Enzyme Biochemistry,* County Clare: Elsevier/North-Holland. A collection of concise reviews of outstanding authors on the essential physical and chemical methods for the investigation of proteins.

Palmer, T. (1985), *Understanding Enzymes,* Chichester: Ellis Horwood. An easy to read introduction into enzymology.

Price, N. C., Stevens, L. (1989), *Fundamentals of Enzymology,* Oxford: Oxford University Press. A comprehensive treatise of enzymology.

Schulz, G. E., Schirmer, R. H. (1979), *Principles of Protein Structure,* New York: Springer. Theoretical background of how the structure of proteins is determined.

Segel, I. H. (1975), *Enzyme Kinetics,* New York:

Wiley. The most detailed and fundamental treatise on enzyme kinetics.

Stryer, L. (1988), *Biochemistry,* New York: Freeman. A excitingly and engagingly written survey on the whole field of biochemistry.

Suelter, C. H. (1985), *A Practical Guide to Enzymology,* New York: J. Wiley & Sons. A short introduction to theoretical and practical enzymology.

Walsh, C. (1979), *Enzymatic Reaction Mechanisms,* San Francisco: Freeman. The different catalytic mechanisms of enzymes are extensively discussed.

PSYCHOLOGICAL ACOUSTICS

See ACOUSTICS, PSYCHOLOGICAL

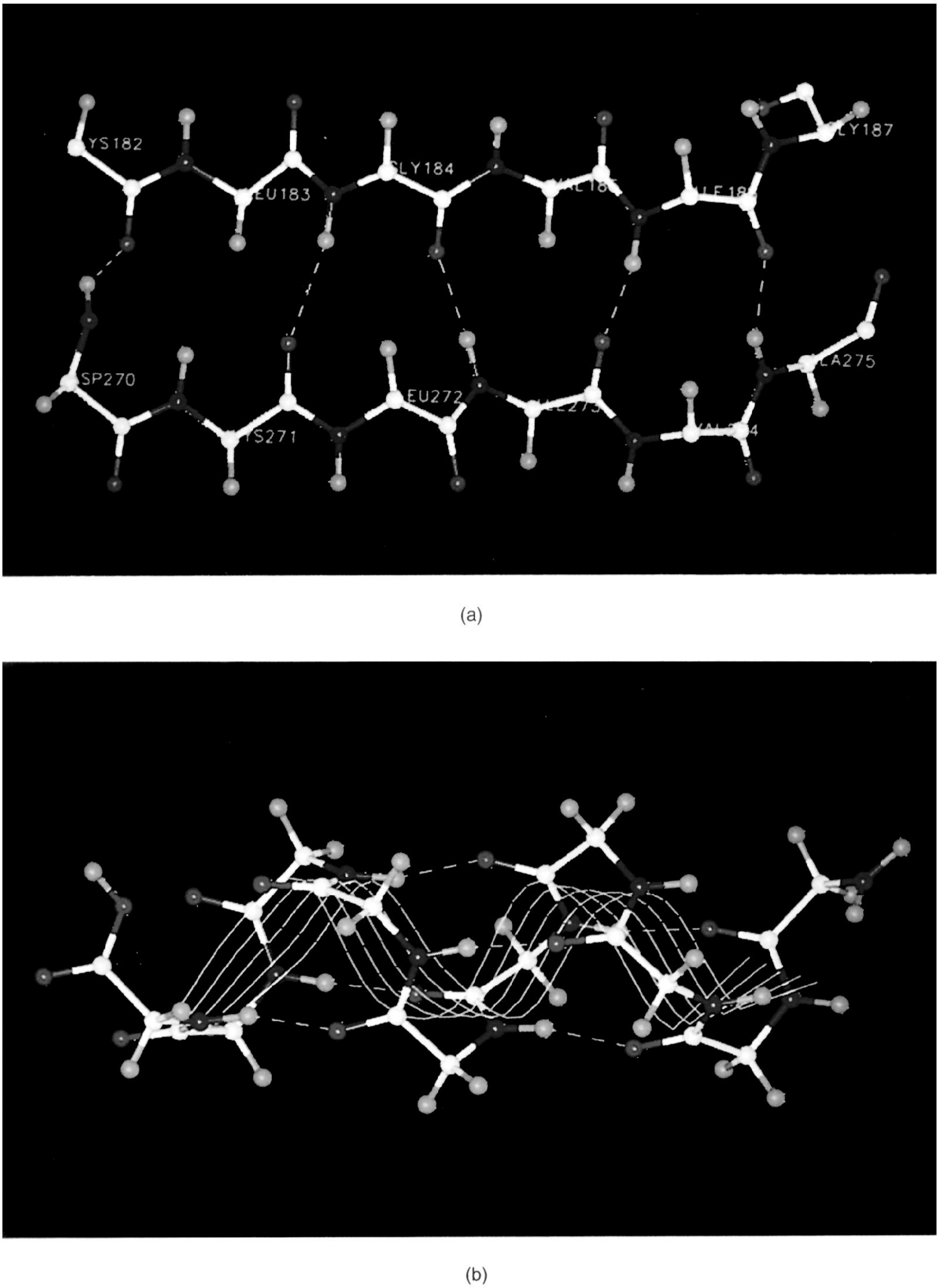
YS182
EU183
LY184
LY187
SP270
YS271
EU272
A275
(a)
(b)

Vol. 15 – PLATE 1. Secondary structures of proteins: (a) β-pleated sheet, (b) α helix. White, carbon atoms; dark blue, nitrogen atoms, light blue, protons. The dashed yellow lines are hydrogen bonds.

QUANTUM ELECTRODYNAMICS

TOICHIRO KINOSHITA, *Newman Laboratory, Cornell University, Ithaca, New York, U.S.A.*

Introduction 215
1. **Development of QED** 216
1.1 Brief Review of Quantum Mechanics 216
1.2 Quantization of a Free Electromagnetic Field 217
1.3 Quantization of a Free Electron Field 218
1.4 Interaction of the Electron Field and the Electromagnetic Field 219
1.5 Divergence Difficulty 220
1.6 Renormalized Quantum Electrodynamics 221
2. **Experimental Tests of QED** 222
2.1 High-Energy Tests 223
2.2 High-Precision Tests of QED at Low Energies 223
2.2.1 Anomalous Magnetic Moment of the Electron 223
2.2.2 Anomalous Magnetic Moment of the Muon 224
2.2.3 Hyperfine Structure of One-Electron Atoms 224
2.2.4 The Hydrogen Lamb Shift 225
3. **High-Precision Determination of α** 225
3.1 α Determined by the Quantum Hall Effect 225
3.2 α Determined by the ac Josephson Effect 225
3.3 α Determined by QED 226
3.4 High-Precision Test of Condensed-Matter Physics 226
4. **Conclusions** 226
Glossary 227
Works Cited 227
Further Reading 228

INTRODUCTION

Quantum electrodynamics is a natural generalization of quantum mechanics that satisfies the requirement of relativity. It is applicable without restriction to all phenomena involving the electromagnetic interaction. Its validity has been confirmed up to a few parts in 10^{12} in precision and down to the length scale of about 10^{-17} m.

In order to understand quantum electrodynamics, it is useful to recall what was wrong with classical electrodynamics, which after all was one of the most successful theories in the history of physics. Classical electrodynamics is applicable to most electromagnetic phenomena and describes them quite accurately. However, it is found to disagree with observations on a microscopic scale, the spectrum of the blackbody radiation and the photoelectric effect being two such phenomena discovered around the end of the nineteenth century. Furthermore, it is incapable of providing a satisfactory explanation for the structure of atomic spectra or the stability of atoms and molecules. From a more fundamental point of view, it suffers from internal inconsistencies such as the infinite energy of self-interaction of charged particles.

As is well known, efforts to resolve some of these difficulties led to the discovery of quantum mechanics. Quantum mechanics has been extremely successful. Through it, the structure of the spectra of atoms and molecules is understood for the first time. It describes qualitatively the emission and absorption of photons by atoms and molecules, although the proper quantitative treatment of the photon had to wait for the development of a quantum theory of photons. It has become the foundation of modern physics and of most modern technologies including lasers, semiconductors, and superconductors. Even in the seemingly macroscopic domains such as establishing practical standards of electricity and magnetism, high-precision re-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

alization of such standards now depends on quantum effects such as the ac Josephson effect of superconductors and the quantum Hall effect of semiconductors at very low temperature (Cohen and Taylor, 1987).

Insofar as practical applications are concerned, there seems to be no need to go beyond the conventional quantum mechanics. However, this is not completely satisfactory. First of all, quantum mechanics cannot possibly be exactly correct since it does not incorporate special relativity, which is one of the cornerstones of modern physics. It is not capable of explaining why electrons obey the Pauli exclusion principle. Furthermore, its predictions often diverge if one tries to calculate beyond lowest-order approximations. The first defect was recognized as soon as quantum mechanics was formulated around 1925, and efforts were made to make quantum mechanics compatible with relativity. However, this task turned out not to be so trivial, and it took more than 20 years before a theory that satisfies the requirement of relativity and incorporates the exclusion principle was formulated successfully. The renormalized quantum electrodynamics (hereafter abbreviated as QED) is the outcome of this effort.

QED has turned out to be much more than a mere marriage of quantum mechanics and relativity. The wave/particle duality is now fully incorporated in the theory, and electrons and photons are treated on the same footing, viz., as quantized fields. It enables us to handle not only the motion of particles dynamically but also the creation and destruction of photons and electron-positron pairs.

What makes QED particularly intriguing is that, for the first time in the history of physics, we have found a theory that has no obvious intrinsic limitations and enables us to calculate physical phenomena from first principles to any accuracy we need in terms of a few measurable parameters such as the elementary electric charge e and the electron mass m_e. Predictions of QED have been confirmed by extensive experiments without exception. In recent years, it has become the prototype of a more general theory, called the standard model, that unifies the electromagnetic force with the weak and strong forces found in nuclear and elementary particle physics.

Throughout this article, only more recent works not found in standard textbooks will be cited. Others may be found in textbooks quoted in Further Reading.

1. DEVELOPMENT OF QED

1.1 Brief Review of Quantum Mechanics

Since QED is nothing but a relativistic extension of ordinary quantum mechanics, it will be useful to begin by recalling the basic features of the latter. Quantum mechanics was discovered in the efforts to make sense out of various observations of atomic scale that contradicted predictions of classical physics. Experiments showed convincingly that classical waves behave like particles (e.g., in the photoelectric effect) and classical particles behave like waves (e.g., in the diffraction of electrons by a crystal lattice) at the atomic scale. A crucial step from classical physics to quantum physics is to accept this wave/particle duality as the fundamental experimental fact. As was noted by A. Einstein and L. de Broglie, the energy E and the momentum p carried by a particle are then related to the frequency ν and the wavelength λ of the associated wave by

$$E = h\nu, \quad p = h/\lambda, \tag{1}$$

where $h = 6.626\,075\,5\,(40) \times 10^{-34}$ J s is Planck's constant. The quantity enclosed in the parentheses represents the measurement uncertainty in the final digits of the numerical value.

One of the characteristic features of a wave is that the peak of its oscillation cannot be located much more precisely than its wavelength. This is translated in quantum theory to the statement that the position x and the corresponding momentum p_x cannot be measured precisely at the same time. Mathematically, p_x can be expressed as a differential operator

$$p_x = \frac{\hbar}{i}\frac{\partial}{\partial x}, \quad \text{where } i = \sqrt{-1}. \tag{2}$$

As is easily checked, the products xp and px, acting from the left on a function of x, are not equal but related to each other by the "commutation relation"

$$xp_x - p_x x = i\hbar \quad \text{or}\ [x,p_x] = i\hbar,$$

where $[x,p_x] \equiv xp_x - p_x x$ is called a commutator. This leads to the Heisenberg uncertainty relation

$$\Delta x \Delta p_x \geq \hbar,$$

where Δx and Δp_x are uncertainties in the position x and the momentum p_x when they are measured simultaneously. Similar commutation relations and uncertainty relations hold for the y and z axes, too. In addition, the energy E becomes an operator

$$E = -\frac{\hbar}{i}\frac{\partial}{\partial t}, \tag{3}$$

leading to another uncertainty relation,

$$\Delta t \Delta E \geq \hbar.$$

In classical physics, the energy E of a free particle of mass m and momentum p is described by the equation

$$E = p^2/2m. \tag{4}$$

If one applies the substitution rules [Eqs. (2) and (3)] to Eq. (4) and supplies a "wave function" $\psi(t,x)$ on which the differential operators operate, one obtains the wave equation

$$i\hbar \frac{\partial}{\partial t}\psi(t,x) = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi(t,x).$$

This is the Schrödinger equation, which gives a complete quantum description of the motion of a free particle in one space dimension. The square of the absolute value of the wave function $\psi(t,x)$ gives the probability for finding the particle at the location x at the time t. Generalizations to three space dimensions and to several particles, and inclusion of the potential energy, are straightforward.

1.2 Quantization of a Free Electromagnetic Field

Shortly after the birth of quantum mechanics, P. A. M. Dirac, W. Heisenberg, and W. Pauli attempted to quantize the energy contained in the electric field $\mathbf{E}(t,\mathbf{x})$ and magnetic field $\mathbf{B}(t,\mathbf{x})$—as required by quantum mechanics for the energy of any physical system. This is not simple because $\mathbf{E}$ and $\mathbf{B}$ are not only functions of time t but also of position $\mathbf{x}$ and hence have infinitely many degrees of freedom. Nevertheless, they managed to quantize it by requiring that $\mathbf{E}$ and $\mathbf{B}$ satisfy a commutation relation analogous to that of x and p_x. The actual form of commutation relations for $\mathbf{E}$ and $\mathbf{B}$ is fairly complicated. It is simpler to write them for the four-vector potential A^μ.

In dealing with relativity, it is convenient to treat the position $\mathbf{x}$ and ct, where c is the speed of light in vacuum and t is the time, together as a vector in four-dimensional space-time. This combination is called a four-vector and denoted as x^μ, $\mu = 0, 1, 2, 3$, where $x^0 = ct$, $x^1 = x$, $x^2 = y$, $x^3 = z$. If one introduces another four-vector $x_\mu \equiv \Sigma_{\nu=0}^3 g_{\mu\nu} \times x^\nu$, where $g_{00} = 1$, $g_{11} = g_{22} = g_{33} = -1$, $g_{\mu\nu} = 0$ otherwise, the principle of relativity can be reduced to the geometrical statement that the "scalar product" of four-vectors x^μ and y_μ, $x^\mu y_\mu \equiv x^0 y^0 - \mathbf{x}\cdot\mathbf{y}$, has the same value in any inertial frame; in other words, it is invariant under the Lorentz transformation. Here and elsewhere in this article, we drop the summation symbol following the convention. Some important four-vectors are the energy-momentum four-vector $p^\mu = (E/c,p_x,p_y,p_z)$ and the four-vector potential $A^\mu = (\phi,A_x,A_y,A_z)$ mentioned already, where ϕ and $\mathbf{A}$ are the scalar and vector potentials in the conventional sense. When there is no danger of confusion, p^μ is often written as p and $p_\mu p^\mu$ as $p\cdot p$ or p^2. Another convention we adopt hereafter is natural units, in which $\hbar = c = 1$ where $\hbar$ is Planck's constant h divided by 2π, instead of SI units. This is done to avoid unnecessary complication of notation. It is not difficult to reconstruct SI units from natural units.

The actual form of A^μ depends on the choice of gauge (see Glossary). In the gauge called "Feynman gauge," it can be written, if A_μ and A_ν have equal time t, as

$$[\dot{A}_\mu(t,\mathbf{x}),\dot{A}_\nu(t,\mathbf{y})] = 0,$$

$$[A_\mu(t,\mathbf{x}),A_\nu(t,\mathbf{y})] = 0,$$

$$[\dot{A}_\mu(t,\mathbf{x}),A_\nu(t,\mathbf{y})] = ig_{\mu\nu}\delta^3(\mathbf{x}-\mathbf{y}), \tag{5}$$

where $\dot{A}_\mu \equiv \partial A_\mu/\partial t$ and $\delta^3(\mathbf{x}) \equiv \delta(x^1)\delta(x^2)\delta(x^3)$,

δ being Dirac's delta function. Reflecting the infinite degrees of freedom of the electromagnetic field, the quantized energy, which can be obtained by applying the action principle to the Lagrangian

$$\mathcal{L}_M = -\tfrac{1}{4} F_{\mu\nu}F^{\mu\nu}$$

and using the commutation relation (5), can be expressed as a superposition of an arbitrary number of quanta. Here, $F_{\mu\nu} \equiv \partial A_\mu/\partial x^\nu - \partial A_\nu/\partial x^\mu$ is the covariant form of the electromagnetic field. This quantum can be identified with Planck's photon, the fundamental carrier of light and electromagnetic radiation. The quantized form of Maxwell's equation requires that the photon be a particle of zero mass, with constant velocity c ($\sim 3 \times 10^8$ m/s) and one unit of angular momentum $\hbar$—directed either along or against its direction of motion. The field amplitudes **E** and **B** [and the four-vector potential $A^\mu(x)$] are now quantum-mechanical operators that have nonzero matrix elements only between states that differ by exactly one photon. This enables us to describe correctly the emission and absorption of radiation in atomic systems.

1.3 Quantization of a Free Electron Field

During this exciting period in which developments in quantum theory came at an incredible rate, Dirac was trying to find a relativistic generalization of the (nonrelativistic) Schrödinger equation for a particle of mass m. Starting from Einstein's energy-momentum relation

$$E^2 = p^2c^2 + m^2c^4,$$

which reduces to the well-known formula "$E = mc^2$" for $p = |\mathbf{p}| = 0$, he was led to the equation

$$[i\gamma_\mu(\partial^\mu + ieA^\mu) - m]\psi(x) = 0, \tag{6}$$

which describes the motion and spin (internal angular momentum) of an electron in an (unquantized) electromagnetic potential A^μ. Here, $\partial_\mu \equiv \partial/\partial x^\mu$, ψ is a one-column four-row matrix, and γ_μ are a set of noncommuting 4×4 matrices satisfying the commutation relation

$$\gamma_\mu\gamma_\nu + \gamma_\nu\gamma_\mu = 2g_{\mu\nu}, \quad \mu, \nu = 0, 1, 2, 3.$$

The application of Eq. (6) to an electron bound in a Coulomb field led to the successful prediction of the fine structure of the energy levels of the hydrogen atom caused by the interaction between spin and orbital motion of the electron. The ratio of the fine-structure separation to the Rydberg is of the order of

$$\alpha \equiv e_0^2/4\pi\hbar c \sim 1/137.$$

This dimensionless number, called the fine-structure constant, characterizes the nature of the electromagnetic interaction and plays the central role in the theory of QED described in the following.

Unfortunately, Dirac's equation had some paradoxical features and defied simple probabilistic interpretation. P. Jordan and E. P. Wigner realized that this difficulty can be resolved if Dirac's ψ is treated not as a probability amplitude for finding a particle at **x** but as a dynamical wave similar to the electromagnetic field and is quantized accordingly. The only new feature is that a product of two ψ's must satisfy a different commutation relation, called (equal-time) anticommutation relation:

$$\{\psi_\xi(t,\mathbf{x}),\psi^\dagger_\eta(t,\mathbf{y})\} = \delta_{\xi\eta}\delta^3(\mathbf{x} - \mathbf{y}), \quad \xi, \eta = 1, 2, 3, 4, \tag{7}$$

where $\psi^\dagger$ is the Hermitian conjugate of ψ, $\{A,B\} \equiv AB + BA$, and $\delta_{\xi\eta} = 1$ if $\xi = \eta$, $= 0$ otherwise. The quanta of the ψ field can be identified with the physical electrons and positrons—particles (antiparticles) that have charge $\mp|e|$, mass m, and angular momentum $\frac{1}{2}\hbar$ directed along or against their motion and obey the Pauli exclusion principle. More specifically, the Dirac ψ field becomes a quantum-mechanical operator that destroys an electron or creates a positron. Conversely, $\bar{\psi}(\equiv \psi^\dagger\gamma^0)$ destroys a positron or creates an electron. Note that electron and positron are described symmetrically by ψ and $\bar{\psi}$ (charge-conjugation symmetry). These properties follow from the Lagrangian

$$\mathcal{L}_D = \tfrac{1}{2} i[\bar{\psi}\gamma^\mu\partial_\mu\psi - (\partial_\mu\bar{\psi})\gamma^\mu\psi] - \tfrac{1}{2} m_0[\bar{\psi},\psi] \tag{8}$$

and the anticommutation relation [Eq. (7)].

1.4 Interaction of the Electron Field and the Electromagnetic Field

Classical electrodynamics suggests (by the "correspondence principle") a quantum form of interaction energy density between the electron, positron, and photon, namely,

$$e_0 j_\mu(x) A^\mu(x), \quad \text{where } j_\mu(x) = \tfrac{1}{2}\,[\bar{\psi}(x), \gamma_\mu \psi(x)]. \tag{9}$$

(Note $j_\mu A^\mu \equiv j^0 A^0 - \mathbf{j}\cdot\mathbf{A}$.) If one accepts Eq. (9), the dynamics of the interacting photon, electron, and positron can be described by the Lagrangian density

$$\begin{aligned}\mathscr{L} = {} & \tfrac{1}{2}\, i[\bar{\psi}\gamma^\mu \partial_\mu \psi - (\partial_\mu \bar{\psi})\gamma^\mu \psi] - \tfrac{1}{2}\, m_0[\bar{\psi}, \psi] \\ & - \tfrac{1}{4}\, F_{\mu\nu}F^{\mu\nu} + \tfrac{1}{2}\, \mu_0^2 A_\mu A^\mu \\ & - \tfrac{1}{2}\, \lambda_0(\partial_\mu A^\mu)^2 - \tfrac{1}{2}\, e_0[\bar{\psi}, \gamma^\mu \psi]A_\mu.\end{aligned} \tag{10}$$

Here μ_0 and λ_0 are a (vanishingly) small photon mass and an arbitrary (gauge) parameter, respectively, introduced for technical reasons. (We must put $\mu_0 = 0$ after the calculation is finished. The final answer should not depend on the value of λ_0.)

If the electric charge e_0 is zero, $\mathscr{L}$ reduces to $\mathscr{L}_M + \mathscr{L}_D$. Upon quantization, the fields ψ, $\bar{\psi}$, and A_μ become operators, each creating or destroying an electron, a positron, or a photon at a time. The energy operator derived from $\mathscr{L}_D + \mathscr{L}_M$ has a structure such that, if a state has given numbers of photons, electrons, and positrons initially, the state remains the same at any later times. Change in the state can occur only if the interaction term is taken into account. An exact solution including this term, which makes the theory nonlinear, has not been obtained thus far.

A common method to find an approximate solution is to expand the solution into a power series in e_0 and determine the expansion coefficients. This method, called perturbation theory, works very well for QED if one restricts oneself to the lowest-order terms, the expansion parameter e_0 being rather small. The results of such calculations are in good agreement with experiment, unless the energy of the photon is too low.

In lowest-order perturbation theory, because of quantization of ψ and A_μ, the interaction term has nonzero matrix elements only between states that differ by one photon or one electron-positron pair. This may be represented graphically by the "vertex" diagrams shown in Fig. 1, where the upper and lower parts correspond to the final state and initial state, respectively. Note that the vertex conserves charge and is "local", i.e., it describes the creation or absorption of a photon at the same point $x = (t,\mathbf{x})$ in time and space where an electron (or positron) scatters or where an electron-positron pair is created or destroyed. The corresponding matrix elements are proportional to (a) $e_0\bar{u}(q)\gamma_\mu u(p)\epsilon^\mu(k)$, (b) $-e_0\bar{v}(q)\gamma_\mu v(p)\epsilon^\mu(k)$, (c) $e_0\bar{u}(q)\gamma_\mu v(p)\epsilon^\mu(k)$, (d) $e_0\bar{v}(q)\gamma_\mu u(p)\epsilon^\mu(k)$. Here $u(p)$, $v(p)$ are one-column, four-row matrices satisfying $(\gamma^\mu p_\mu - m)u(p) = 0$, $(\gamma^\mu p_\mu + m)v(p) = 0$, and $\bar{u}(p)$, $\bar{v}(p)$ are one-row, four-column matrices satisfying $\bar{u}(p)(\gamma^\mu p_\mu - m) = 0$, $\bar{v}(p)(\gamma^\mu p_\mu + m) = 0$, respectively. The term $\epsilon^\mu(k)$ represents the polarization state of a photon of four-momentum k.

The transition amplitude for more complicated physical processes can be obtained from the iteration of the basic interaction (i.e., higher-order terms of perturbation expansion in powers of e_0). In the order e_0^2, new diagrams appear in which some lines connect vertices. Such solid lines are called internal electron lines while wavy lines are called internal photon lines. Other lines are called "external." For example, the order-e_0^2 amplitude for the electron-positron collision can be computed from the "Feynman diagram" shown in Fig. 2(a). It is proportional

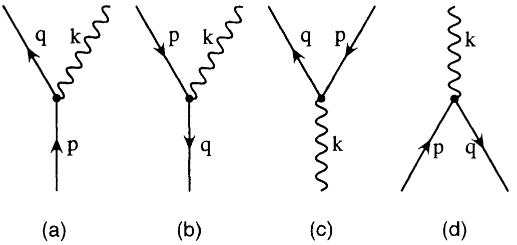

FIG. 1. These "vertex" diagrams are graphic representations of the basic interaction between electron, positron, and photon. (a) Emission of a photon (represented by a wavy line) by an electron (represented by a solid line with an upward arrow) that recoils to a different four-momentum state. (b) Emission of a photon by a positron (represented by a solid line with a downward arrow). (c) Creation of an electron-positron pair by a photon, and (d) its reverse. There are four more diagrams. Each diagram must conserve energy, momentum, and charge. The corresponding matrix elements vanish because of energy-momentum conservation if the electron, positron, and photon are free.

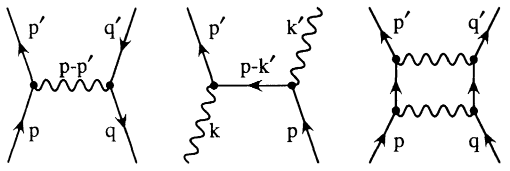

(a) (b) (c)

FIG. 2. Examples of second-order diagrams. (a) Scattering of a positron and an electron caused by an exchange of a virtual photon. (b) Scattering of a photon on an electron caused by an exchange of a virtual electron. (c) An example of higher-order diagrams in which more than one photon is exchanged.

to

$$ie_0^2\bar{u}(p')\gamma_\mu u(p)\,\frac{1}{(p-p')^2}\,\bar{v}(q)\gamma^\mu v(q'),$$

where the factor $1/(p-p')^2 \equiv [(E-E')^2 - (\mathbf{p}-\mathbf{p}')^2]^{-1}$, called a photon propagator, represents propagation of the photon carrying momentum $\mathbf{p}-\mathbf{p}'$ and energy $E-E'$ between the electron and the positron. A similar propagator $[\gamma_\nu(p^\nu - k'^\nu) - m]^{-1}$ occurs in diagrams such as Fig. 2(b) for the propagation of the internal electron or positron. The corresponding amplitude is of the form

$$ie_0^2\bar{u}(p')\epsilon_\mu(k')\gamma^\mu[\gamma_\nu(p^\nu - k'^\nu) - m]^{-1} \times \epsilon_\sigma(k)\gamma^\sigma u(p).$$

The diagram of Fig. 2(c) is an example of the e_0^4 term. These diagrams give contributions to the collision cross section that are smaller than that of Fig. 2(a) by a factor of order α. Because of the smallness of α, it is usually sufficient to calculate the lowest-order terms in order to compare with experiments.

The absolute square of the sum of all the possible amplitudes for the process gives the quantum-mechanical probability, or, in the case of a collision process, the cross section.

1.5 Divergence Difficulty

Despite the apparent correctness of the basic structure of the interaction between the photon and the electron, the good agreement with experiment is completely destroyed as soon as one tries to improve the theory by including higher-order terms of perturbation theory. The correction terms are infinitely large!

For example, the calculation of the energy shift of a free electron due to its emission and subsequent absorption of a single photon, shown in Fig. 3, gives an infinite result. This divergence difficulty can be traced to the basic interaction being pointlike (local). Attempts to solve this problem by giving a finite size to the electron have not been successful.

A breakthrough on this problem was brought forward by an extraordinary experiment performed by W. E. Lamb and R. C. Retherford in 1947. The Dirac equation for an electron bound in a Coulomb field predicted that the energy levels with the same principal quantum number n and total angular momentum j should be degenerate, i.e., they should not depend on other quantum numbers such as that of orbital angular momentum. In particular, there should be no energy separation between $2P_{1/2}$ and $2S_{1/2}$ excited levels in atomic hydrogen. Lamb's experiment, however, demonstrated conclusively that the levels were in fact separated by ~1000 MHz, a full tenth of the fine-structure separation of the $n = 2$ levels.

Then, in a historic paper of 1947, H. A. Bethe showed that the "Lamb shift" can be explained by a proper reinterpretation of the theory (inspired by H. A. Kramers's attempt to solve the self-mass divergence difficulty). He pointed out that when one calculates the

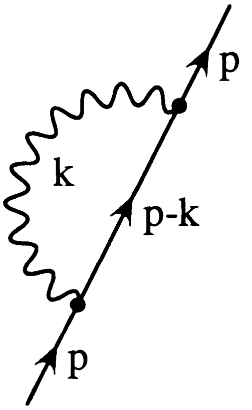

FIG. 3. Second-order electron self-energy diagram describing an electron emitting and reabsorbing a photon.

self-energy correction to a bound-state electron resulting from the emission and reabsorption of a photon, part of the energy shift—in fact, the divergent part—should be identified with the "self-energy" or "self-mass" of the electron itself. Now, the energy shifts for two bound-state levels contain the same self-mass term and hence their difference should be finite. In fact, his calculations showed that the $2S_{1/2}$ level is displaced above the $2P_{1/2}$ level by about 1080 MHz. This can be understood physically by the observation that the emission and reabsorption of the photon causes the position of the electron to fluctuate over a distance of the Compton wavelength of the electron ($\hbar/m_e c$) a fraction α of the time. The degeneracy predicted by the Dirac equation is a consequence of the strict $1/r$ behavior of the electron-nucleus Coulomb potential. This is destroyed by the orbital dependence of the effective spreading of the electron's position.

Bethe's calculation is in good agreement with the Lamb–Retherford experiment if one includes the "vacuum polarization" effect, which increases the binding of the electron and reduces the $2S_{1/2} - 2P_{1/2}$ separation by 27 MHz. This effect, found by R. Serber and E. A. Uehling in 1935, is caused by the modification of the nuclear Coulomb potential due to the presence of virtual electron-positron pairs. (The phenomenon is analogous to the polarization of dielectric matter in the presence of charge.) In fact, because two charged particles always see each other through a cloud of virtual particle-antiparticle pairs, the charge e_0 that appears as a parameter in the Lagrangian [Eq. (10)] is not precisely the physical charge e of the electron (as defined classically for an electron scattering at large distances from a test electric field). Attempts to calculate e in terms of e_0 lead to another divergent correction unrelated to that of the electron self-mass.

1.6 Renormalized Quantum Electrodynamics

Within a few years after Bethe's work, J. Schwinger, S. Tomonaga, R. P. Feynman, and F. J. Dyson demonstrated that all the divergences in quantum-electrodynamic calculations could be eliminated order by order in perturbation theory. This procedure, called the renormalization program, was shown to be covariant (independent of Lorentz frame), and despite the presence of infinite quantities in intermediate expressions, the calculations were shown always to lead to unambiguous finite results.

Briefly, one starts from the observation that the "bare charge" e_0 and "bare mass" m_0 that appear as parameters in the Lagrangian [Eq. (10)] or the equations of QED are not the physical charge and mass of the electron. To any given order in perturbation theory, one must adjust the values of m_0 and e_0 so that after calculation of all the contributing terms, electron-electron scattering at large distances will agree with classical Coulomb scattering for two particles of charge e, and that Compton scattering (elastic photon-electron scattering) at zero photon frequency will agree with the Thomson limit that defines the ratio e^2/m. After this adjustment is made, calculations not only of these but also of all *physical* quantities give finite results, although the values in perturbation theory for the quantities e_0 and m_0 are mathematically divergent.

Let me now sketch the Feynman-diagram technique for computing the transition amplitudes of QED. (For details, consult the textbooks listed in Further Reading.) For this purpose, it is useful to introduce a free-field Lagrangian $\mathscr{L}^0$ whose solution is known completely:

$$\begin{aligned}\mathscr{L}^0 = {} & \tfrac{1}{2} i[\bar{\psi}^0\gamma^\mu\partial_\mu\psi^0 - (\partial_\mu\bar{\psi}^0)\gamma^\mu\psi^0] \\ & - \tfrac{1}{2} m[\bar{\psi}^0,\psi^0] - \tfrac{1}{4} F^0_{\mu\nu}(F^0)^{\mu\nu} \\ & + \tfrac{1}{2} \mu_0^2 A^0_\mu(A^0)^\mu - \tfrac{1}{2} \lambda_0[\partial_\mu(A^0)^\mu]^2, \end{aligned} \tag{11}$$

where A^0_μ and ψ^0 are free-field operators satisfying the "initial" condition

$$\begin{aligned} A^0_\mu(0,\mathbf{x}) &= \sqrt{Z_3}A_\mu(0,\mathbf{x}), \\ \psi^0(0,\mathbf{x}) &= \sqrt{Z_2}\psi(0,\mathbf{x}), \end{aligned} \tag{12}$$

at $t = 0$, Z_2 and Z_3 being some constants to be fixed later. Note that ψ^0 has the physical mass m.

We want to express the solution of $\mathscr{L}$ in terms of the solution of $\mathscr{L}^0$, treating the difference $\mathscr{L} - \mathscr{L}^0$, which can be written

$$-\mathscr{H}_{\text{int}} \equiv -\tfrac{1}{2} e_0[\bar{\psi}^0,\gamma^\mu\psi^0]A^0_\mu + \tfrac{1}{2}\,\delta m[\bar{\psi}^0,\psi^0], \tag{13}$$

as a small perturbation, where $\delta m = m -$

m_0. The probability amplitude for the transition from a state at $t \to -\infty$ to a state at $t \to +\infty$ can then be written as a matrix element of the operator

$$S = \frac{T\left(\exp\left[-i\int \mathcal{H}_{\text{int}}(x)d^4x\right]\right)}{\left\langle 0\left|T\left(\exp\left[-i\int \mathcal{H}_{\text{int}}(x)d^4x\right]\right)\right|0\right\rangle},$$

where $|0\rangle$ means the vacuum state (state of lowest energy) and T means that all products of operators (obtained by expanding the exponential function in a power series) are ordered from left to right in the descending order of their time variable t.

It can be shown that matrix elements of S correspond one-to-one to figures called Feynman diagrams. Some examples were shown in Figs. 1 through 3. The rules for calculating the matrix element, called the Feynman–Dyson rules, have been given in a very elegant and simple form by R. P. Feynman in 1949.

Let $|i\rangle$ be the initial state containing several particles of given four-momenta and spins and $\langle f|$ be the final state similarly specified. Draw all diagrams (of given order) that connect the initial state to the final state, excluding those containing disconnected pieces and also those with self-energy insertions on external lines. Let P_i and P_f be the total four-momenta of the initial and final states. Let us introduce a matrix R by

$$\langle f|S - 1|i\rangle = -(2\pi)^4 i\delta^4(P_f - P_i)\langle f|iR|i\rangle.$$

Then the matrix element $\langle f|iR|i\rangle$ can be obtained by assigning the following factors to each diagram:

1. $-ie_0\gamma_\mu$ for each three-point vertex ($e_0 < 0$);
2. $i\delta m$ for each mass counterterm vertex;
3. $\sqrt{Z_2}u(p,s)$ or $\sqrt{Z_2}v(p,s)$ for each electron line entering the graph in the initial or final state; likewise, for each electron line leaving the graph a factor $\sqrt{Z_2}\bar{u}(p,s)$ or $\sqrt{Z_2}\bar{v}(p,s)$, where p,s are the four-momentum and spin state of the electron;
4. $\sqrt{Z_3}\epsilon^\mu(k)$ for each external photon line carrying four-momentum k;
5. $i/(\gamma_\mu p^\mu - m + i\epsilon)$ for each internal electron line with four-momentum p^μ, where ϵ is an infinitesimally small positive number;
6. $-ig_{\mu\nu}/(q^2 + i\epsilon)$ for each internal photon line with four-momentum q (in Feynman gauge only);
7. Four-dimensional integration

   $$\int \frac{d^4k}{(2\pi)^4}$$

   for each internal momentum k not fixed by momentum-conservation constraints at vertices (note that four-momentum is conserved at each vertex);
8. -1 for each closed electron loop.

Finally, renormalize the charge with $e = Z_2 Z_1^{-1}\sqrt{Z_3}e_0$. (Actually, $Z_1 = Z_2$ according to an identity proved by J. Ward.)

The quantities δm, Z_1, Z_2, and Z_3 are defined as values of $\langle f|R|i\rangle$ for some special states $\langle f|$ and $|i\rangle$. In the lowest order, $\delta m = 0$, $Z_1 = Z_2 = Z_3 = 1$. In addition, some normalization factors are needed to obtain an exact R. See Bjorken and Drell (1965) for details.

In each $\langle f|R|i\rangle$ thus constructed, infinite quantities appear during the calculation. However, as a consequence of renormalization, they all disappear in the final result expressed in terms of observed charge and mass.

In 1948, Feynman proposed an alternative formulation of quantum mechanics called the path-integral method. This method, applied to the electromagnetic interaction, provides an alternative formulation of QED. In fact, this is how Feynman was led to his diagrammatic method. The path-integral approach is useful in formulating more complicated theories such as quantum chromodynamics and the standard model.

Finally, we note that the electromagnetic interaction is invariant under the operation of space reflection P, time-reversal T, or charge conjugation C. In contrast, the weak interaction is not invariant under P or C separately, although it is invariant under the combined operation PC. All known interactions are invariant under the combined CPT operation.

2. EXPERIMENTAL TESTS OF QED

Over the history of physics, various theories have been formulated, tested, and revised by experiment. In a similar manner, QED has been tested extensively. High-energy collisions of electrons and positrons examine the validity and limitation of the theory in the region of very short distance scales, or, equivalently, very large momentum transfer. On the other hand, higher-order QED corrections are tested in various high-precision measurements of atomic energy levels and transitions.

2.1 High-Energy Tests

Very high-energy (short-distance) tests of the predictions of QED for electron-positron scattering have been performed at colliding electron-positron beam facilities at Stanford, Cornell, Orsay, Novosibirsk, Frascati, Tsukuba, Hamburg, and Geneva. The main processes studied are those in which colliding positron (e^+) and electron (e^-) produce an e^+e^- pair, two photons (γ's), or a positive-muon (μ^+), negative-muon (μ^-) pair:

$$\begin{aligned} &e^+ + e^- \rightarrow e^+ + e^-, \\ &e^+ + e^- \rightarrow \gamma + \gamma, \qquad (14) \\ &e^+ + e^- \rightarrow \mu^+ + \mu^-. \end{aligned}$$

Here the muon is an elementary particle that behaves exactly like an electron except for its heavier mass: $m_\mu/m_e = 206.768\,262\ (30)$.

QED predicts that the cross sections for the processes of Eqs. (14) fall inversely with the square of the center-of-mass energy and have a certain angular dependence. The measurements have confirmed these predictions at center-of-mass energies as high as 9×10^{10} eV (denoted 90 GeV). This means that the measurements rule out breakdown of QED down to distances of $\sim 10^{-17}$ m.

Actually, above center-of-mass energies of several hundred million eV (MeV), the e^+e^- collision is dominated by non-QED processes such as production of pions and kaons. (These particles are called "hadrons.") Nevertheless, for the processes of Eqs. (14), the prediction of QED holds precisely. At energies around 90 GeV, the contribution of the weak boson Z^0, predicted by the electroweak theory of Glashow, Weinberg, and Salam and recently confirmed at CERN experiments, dominates those of QED and production of hadrons. This does not mean, however, that QED fails at short distances but simply that the real world has forces besides electromagnetism. Indeed QED must be regarded as an integral part of a more comprehensive theory, the so-called standard model, that accommodates strong and weak forces in addition to the electromagnetic force.

2.2 High-Precision Tests of QED at Low Energies

Experiments at low energies are not very sensitive to short-distance behavior of QED. Instead, these experiments are useful in examining higher-order terms of perturbation theory and nonperturbative features of QED associated with bound states. QED has been applied to the Lamb shifts and fine and hyperfine structures of hydrogenlike atoms, as well as to more complex atoms such as helium and high-Z muonic atoms. Recent progress in the measurement of various energy levels of stripped heavier atoms has provided further confirmation of QED. The agreement of theory and experiment is excellent if all relevant higher-order corrections are included. In the following, let us review some of the high-precision results.

2.2.1 Anomalous Magnetic Moment of the Electron The classic and most basic test of QED is provided by the measurement of the anomalous magnetic moment a_e of the electron. According to the Dirac equation [Eq. (6)], the electron gyromagnetic factor g_e, the ratio of spin and rotation frequencies, must be exactly equal to 2. However, it was found by P. Kusch and H. M. Foley in a 1947 experiment that g_e is larger than 2 by about 0.1%. J. Schwinger showed that this is due to the emission and reabsorption of a photon by the electron in a magnetic field. The agreement of the experiment with his calculation of $a_e = (g_e - 2)/2$ was one of the crucial successes of the (then) newly developed theory of QED.

A more accurate measurement of the electron anomalous magnetic moment utilizes electrons that move in a helical orbit with a rotational frequency $\omega_L = eB/mc$ in a uniform magnetic field. If the electron's magnetic interaction conformed to the simple Dirac equation, the directions of its momentum and intrinsic spin would precess with exactly the same frequency. Any deviation from Dirac's prediction would then be detected by measuring the spin precession. Wilkinson and Crane (1963) have shown that the electron anomalous magnetic moment can be measured very accurately in this manner. The best result obtained by this technique is (Wesley and Rich, 1971)

$$a_e^{\text{exp}} = 0.001\,159\,6577\,(35). \tag{15}$$

More recently, based on an entirely different technique that utilizes rf resonance transitions between quantized levels of an electron in a magnetic field, Van Dyck *et al.* (1987) have improved the value of a_e by three orders of magnitude over Eq. (15). Their results for the electron and positron are

$$a_{e^-}^{\text{exp}} = 0.001\,159\,652\,1884\,(43),$$

$$a_{e^+}^{\text{exp}} = 0.001\,159\,652\,1879\,(43), \tag{16}$$

respectively. This experiment verifies the predicted invariance of QED under the combined operation of *TCP* to the level of 10^{-12}:

$$a_{e^-}^{\text{exp}} - a_{e^+}^{\text{exp}} = 0.5(6.1) \times 10^{-12}. \tag{17}$$

Thus far, the theory of a_e has been extended to the emission and reabsorption of up to four photons by the electron, which means that terms of up to eighth order in perturbation theory have been calculated (Kinoshita and Lindquist, 1990). Such a calculation requires the help of algebraic and numerical computation techniques on the fastest computers. The result can be written as

$$\begin{aligned} a_e^{th} &= 0.5\alpha/\pi - 0.328\,478\,965 \cdots (\alpha/\pi)^2 \\ &\quad + 1.176\,11(42)(\alpha/\pi)^3 \\ &\quad - 1.434(138)(\alpha/\pi)^4 + 4.4 \times 10^{-12} \\ &= 0.001\,159\,652\,1409\,(43), \end{aligned} \tag{18}$$

in good agreement with experiment. The last term (4.4×10^{-12}) consists of the effects of muon and hadrons on the photon propagator and the weak-interaction contribution. The value of α used in Eq. (18) is obtained from Eq. (22). The gyromagnetic ratio of the electron g_e is thus correctly predicted by QED to 12 significant figures!

2.2.2 Anomalous Magnetic Moment of the Muon The anomalous magnetic moment of the muon is another quantity measured very accurately. The most precise value obtained at CERN, Switzerland, based on a muon spin-precession technique similar to that of Wilkinson and Crane (1963) for the electron, is (Farley and Picasso, 1990)

$$a_\mu = 1\,165\,923(8.5) \times 10^{-9} \quad (7\text{ ppm}).$$

This is in agreement with the theoretical calculation (Kinoshita and Marciano, 1990)

$$a_\mu^{\text{theory}} = 1\,165\,919.2(1.8) \times 10^{-9}.$$

In this result, the roles of the photon-hadron interaction and the weak interaction are much more important than in Eq. (18). Thus, an accurate measurement of a_μ provides an important test of the theory of weak interaction (or the standard model). A new experiment for a more accurate measurement of a_μ is in progress.

2.2.3 Hyperfine Structure of One-Electron Atoms Other precision tests of QED involve the hyperfine splitting of the ground-state energy level of fundamental atoms such as hydrogen, muonium (atom consisting of a positive muon and an electron), and positronium (positron-electron atom). The hyperfine splitting results because of the difference in energy that occurs when the electron and nuclear magnetic moments are parallel and antiparallel. Measurements (based on the properties of the hydrogen maser) by Helwig *et al.* (1970) and Essen *et al.* (1971) have determined this separation to 13 significant figures in atomic hydrogen:

$$\Delta\nu(\text{hydrogen}) = 1420.405\,751\,7667(9)\text{ MHz}.$$

The results agree with the theoretical prediction to six significant figures. Unfortunately, the theory is sensitive to the structure of the proton at the level of a few parts per million, and cannot be improved further until the proton structure is known better.

Muonium is similar to the hydrogen atom in many ways, except that it does not share the difficulty of the hydrogen atom that arises from the finite size of the proton: The muon is structureless as far as we have been able to measure. In this sense, the muonium is a better system to test the validity of QED.

Very accurate measurement of the hyperfine splitting in muonium by Mariam *et al.*

(1982) gives

$$\Delta\nu(\text{muonium}) = 4463\,302.88(16)\ \text{kHz}. \quad (19)$$

Recent theoretical efforts (Eides *et al.*, 1991) have led to

$$\Delta\nu^{th}(\text{muonium}) = 4463\,303.63(1.34) \times (0.21)(1.0)\ \text{kHz},$$

where the first and second errors reflect the uncertainties in the measurements of the muon mass m_μ and α from Eq. (22), and the third is an order-of-magnitude estimate of the uncalculated higher-order contributions. Comparison of these results pushes the test of QED to a precision of 0.17 ppm. Further improvements in theory and measurement of $\Delta\nu$(muonium) are in progress at present.

The present experimental value for the hyperfine splitting in the ground state of positronium is (Ritter *et al.*, 1984)

$$\Delta\nu(\text{positronium}) = 203\,389.10(74)\ \text{MHz}.$$

The theory is similar to that for the muonium hyperfine structure, with the additional complications coming from the facts that the electron and positron have the same mass and that the electron and positron can annihilate each other. The latest theoretical result is (Adkins *et al.*, 1993)

$$\Delta\nu^{th}(\text{positronium}) = 203\,389.0(6)\ \text{MHz}.$$

The agreement between theory and experiment is very good, but the precision is not yet as high as that of the muonium hyperfine structure.

2.2.4 The Hydrogen Lamb Shift The most recent measurements of the hydrogen Lamb shift are

1057 862(20) kHz, (Andrews and Newton, 1976),

1057 845(9) kHz, (Lundeen and Pipkin, 1981),

1057 851(2) kHz. (Pal'chikov *et al.*, 1983). (20)

The latest theoretical values are (Sapirstein and Yennie, 1990)

1057 853(13) kHz,

1057 871(13) kHz, (21)

depending on which measured value of the proton radius is used. Although the Lamb-shift experiment played a crucial role in the development of QED, it is at present not the best system to test QED because of its dependence on the proton structure, which is hard to measure accurately. However, new experiments based on laser technology promise to give much more accurate data before long.

3. HIGH-PRECISION DETERMINATION OF α

Since the fine-structure constant α is the most important parameter of QED, let us review here how α is determined precisely. At present, the most accurate non-QED methods are those based on macroscopic quantum effects discovered in the realm of condensed-matter physics, namely the quantum Hall effect and the ac Josephson effect. Note, however, that the accuracy of these effects implicitly assumes the full validity of QED.

3.1 α Determined by the Quantum Hall Effect

Currently, the most accurate value of α outside of pure QED is one obtained using the quantum Hall effect. With the use of the exactly defined constants μ_0 (the permeability of the vacuum) and c (the speed of light), the quantum Hall resistance R_H (in SI units) is related to the fine-structure constant by

$$\alpha^{-1} = 2R_H/\mu_0 c.$$

The best result obtained thus far by the quantum Hall effect is (Cage *et al.*, 1989)

$$\alpha^{-1}(\text{QHE}) = 137.035\,9979(32)\ (0.024\ \text{ppm}). \quad (22)$$

3.2 α Determined by the ac Josephson Effect

Another way to obtain a very accurate α is to rely on a chain of four separate precision measurements, making use of the formula

$$\alpha^{-2} = \frac{c}{4R_\infty \gamma_{p'}} \frac{\mu_{p'}}{\mu_B} \frac{2e}{h}.$$

The most critical two are the ac Josephson effect, which gives a precise value for the combination $2e/h$, and the gyromagnetic ratio of the proton in water (called $\gamma_{p'}$), which is the spin-flip frequency of the proton divided by the magnetic field. This frequency can be measured very precisely. The magnetic field is obtained by constructing a precision coil with a current known from a precision current standard. The Rydberg constant, R_∞, and the proton's magnetic moment (in water) in units of the Bohr magneton, $\mu_{p'}/\mu_B$, are very accurately measured (Cohen and Taylor, 1987). The best result obtained thus far by the ac Josephson effect is (Williams *et al.*, 1989)

$$\alpha^{-1}(\text{acJ}) = 137.035\,9770(77)\ (0.056\ \text{ppm}). \tag{23}$$

3.3 α Determined by QED

If one assumes (with ample justification) that QED is valid to the precision examined, one can calculate the value of α from various QED processes. The most accurate value obtained thus far is one from the electron anomalous magnetic moment. Comparing Eqs. (16) and (18), one finds

$$\alpha^{-1}(a_e) = 137.035\,992\,22(94)\ (0.0069\ \text{ppm}). \tag{24}$$

A less accurate value is obtained from the theory and measurement of the muonium hyperfine structure (Kinoshita and Yennie, 1990):

$$\alpha^{-1}(\text{muonium}) = 137.035\,9930(224) \times (0.164\ \text{ppm}). \tag{25}$$

Very good agreement between the values of α given by Eqs. (24) and (25) shows that QED is valid to the level of 0.17 ppm. However, Eq. (25) must be improved by a factor of 20 to reach the precision of Eq. (24). Once that is accomplished, one can test the internal consistency of QED at the level of 0.01 ppm or better.

The uncertainty in Eq. (25) comes mainly from the muon mass and the theory of muonium hyperfine structure. A new experiment for a more accurate measurement of the muon mass is in progress. The same experiment will also improve $\Delta\nu$ of Eq. (19).

3.4 High-Precision Test of Condensed-Matter Physics

It is generally believed that the α determined by the quantum Hall effect and the ac Josephson effect are exact and need no theoretical correction, although this has never been proven. The difference of 21×10^{-6} between Eqs. (22) and (23), which is about three times larger than the error in Eq. (23), might be the first indication that this belief is open to question. A small discrepancy between these α's and $\alpha(a_e)$ obtained from QED also points to the possibility that small corrections might have to be applied to the predictions of condensed-matter physics.

4. CONCLUSIONS

As far as we know, QED provides a mathematically exact description of the electromagnetic properties of the electron and muon. To the extent that the electromagnetic properties of the nucleus are known, QED also provides the fundamental dynamical theory of the relativistic atom, including its external electromagnetic interactions. At present, there is no outstanding discrepancy with any of its predictions, provided that the experimental data are corrected for effects due to weak and hadronic interactions, despite our pursuing the limits of the theory to very high accuracy and very small distances. In addition, the invariance of QED under charge conjugation C, space reflection P, and time reversal T, various conservation laws, the constancy in time of α, c, etc., have all been checked to various degrees. The great success of QED is rooted deeply in the fact that it is a logical and natural extension of the classical electrodynamics and the conventional quantum mechanics, both supported strongly by experiments within their own applicabilities.

Nevertheless, it must be kept in mind that QED still has aspects that have been taken as arguments against its adequacy as a satisfactory theory, such as the following:

1. The expansion in powers of α is only asymptotic, and
2. The very presence of infinities, even though there is an internally consistent means of eliminating them, is a symptom of a deeper problem.

Besides these concerns, the adequacy of QED at shorter distances or at higher precision will have to be questioned if any of the following exist: intrinsic lepton size, nonlocal currents, heavy leptons, heavy photons, magnetic charge, or anomalous constants not given by the renormalization procedure, although these possibilities have been ruled out to some extent by the various tests discussed in this article.

More importantly, QED has served as a prototype for the development of various non-Abelian gauge theories. There is now strong evidence that one of these theories, called the standard model, can account for the properties of the strong and weak interactions observed in nuclear and elementary particle physics. Some of the particles predicted by the standard model, the W and Z bosons, which mediate the weak force, have been confirmed experimentally. Nevertheless, the standard model, of which QED is a part, is generally regarded as far from an ultimate theory. For instance, at present the mass and charge of an electron are just input parameters and not an outcome of the theory itself. To calculate them theoretically would require a deeper understanding of physics. It is not clear whether one of the (super-)string theories, which are being investigated vigorously at present, is a candidate of such a theory.

GLOSSARY

ac Josephson Effect: A phenomenon in which a pair of electrons tunnel back and forth with the frequency ν determined by the equation $h\nu = 2eV$ through a junction which separates superconductors at potential difference V.

Gauge and Gauge Invariance: The Lagrangian [Eq. (10)] is unchanged under the transformation (called gauge transformation)

$$A_\mu \rightarrow A_\mu + \frac{1}{e_0}\partial_\mu \Lambda,$$

$$\psi \rightarrow e^{i\Lambda}\psi,$$

$$\bar{\psi} \rightarrow \bar{\psi}e^{-i\Lambda}, \qquad (26)$$

where $\Lambda(x)$ is an arbitrary function of x^μ. This invariance is a consequence of masslessness of the photon and is a characteristic feature of QED. Gauge invariance of QED means that physics must be independent of the choice of the gauge function. The choice is dictated by convenience or other desirable features.

Quantum Hall Effect: When the electrons participating in a Hall-effect measurement are confined to a very thin, almost two-dimensional, layer with a strong perpendicular magnetic field and at very low temperatures, the Hall resistance tends to be quantized at certain particular values. That is, as some external parameter such as the magnetic field or a gate voltage is varied (depending on the type of device), one observes certain "plateaus" in the Hall resistance, which is defined as the ratio of the voltage (V_H) across the sample to the current (I_H) flowing along the sample.

Works Cited

Adkins, G. S., Aksu, Y. M., Bui, M. H. T. (1993), *Phys. Rev. A* **47,** 2640–2652.

Andrews, D. A., Newton, G. (1976), *Phys. Rev. Lett.* **37,** 1254–1257.

Bjorken, J. D., Drell, S. D. (1965), *Relativistic Quantum Fields,* New York: McGraw-Hill.

Cage, M. E., Dziuba, R. F., Elmquist, R. E., Field, B. F., Jones, G. R., Jr., Olsen, P. T., Phillips, W. D., Shields, J. Q., Steiner, R. L., Taylor, B. N., Williams, E. R. (1989), *IEEE Trans. Instrum. Meas.* **38,** 284–289.

Cohen, E. R., Taylor, B. N. (1987), *Rev. Mod. Phys.* **59,** 1121–1148.

Eides, M. I., Karshenboim, S. G., Shelyuto, V. A. (1991), *Phys. Lett. B* **268,** 433–436.

Essen, L., Donaldson, R. W., Bangham, M. J., Hope, E. G. (1971), *Nature* **229,** 110–111.

Farley, F. J. M., Picasso, E. (1990), in: T. Kinoshita (Ed.), *Quantum Electrodynamics,* Singapore: World Scientific, pp. 479–559.

Helwig, H., Vessot, R. F. C., Levine, M. W., Zitzewitz, P. W., Allan, D. W., Glaze, D. J. (1970), *IEEE Trans. Instrum. Meas.* **19,** 200–209.

Kinoshita, T., Lindquist, W. B. (1990), *Phys. Rev. D* **42,** 636–655.

Kinoshita, T., Marciano, W. J. (1990), in: T. Kinoshita (Ed.), *Quantum Electrodynamics,* Singapore: World Scientific, pp. 419–478.

Kinoshita, T., Yennie, D. R. (1990), in: T. Kinoshita (Ed.), *Quantum Electrodynamics,* Singapore: World Scientific, pp. 1–14.

Lundeen, S. R., Pipkin, F. M. (1981), *Phys. Rev. Lett.* **46,** 232–235.

Mariam, F. G., Beer, W., Bolton, P. R., Egan, P. O., Gardner, C. J., Hughes, V. W., Lu, D. C., Souder, P. A., Orth, H., Vetter, J., Moser, U., zu Putlitz, G. (1982), *Phys. Rev. Lett.* **49,** 993–996.

Pal'chikov, V. G., Sokolov, Yu. L., Yakovlev, V. P. (1983), *JETP Lett.* **38,** 418–420.

Ritter, M., Egan, P. O., Hughes, V. W., Woodle, K. A. (1984), *Phys. Rev. A* **30,** 1331–1338.

Sapirstein, J. R., Yennie, D. R. (1990), in: T. Kinoshita (Ed.), *Quantum Electrodynamics,* Singapore: World Scientific, pp. 560–672.

Van Dyck, R. S., Jr., Schwinberg, P. B., Dehmelt, H. G. (1987), *Phys. Rev. Lett.* **59,** 26–29.

Wesley, J. C., Rich, A. (1971), *Phys. Rev. A* **4,** 1341–1363.

Wilkinson, D. T., Crane, H. R. (1963), *Phys. Rev.* **130,** 852–863.

Williams, E. R., Jones, G. R., Jr., Ye, S., Liu, R., Sasaki, H., Olsen, P. T., Phillips, W. D., Layer, H. P. (1989), *IEEE Trans. Instrum. Meas.* **38,** 233–237.

Further Reading

Bjorken, J. D., Drell, S. D. (1964), *Relativistic Quantum Mechanics,* New York: McGraw-Hill.

Bjorken, J. D., Drell, S. D. (1965), *Relativistic Quantum Fields,* New York: McGraw-Hill.

Feynman, R. P. (1985), *QED: The Strange Theory of Light and Matter,* Princeton: Princeton University Press.

Itzykson, C., Zuber, J. -B. (1980), *Quantum Field Theory,* New York: McGraw-Hill.

Kinoshita, T. (Ed.) (1990), *Quantum Electrodynamics,* Singapore: World Scientific.

QUANTUM LOGIC

David J. Foulis, *University of Massachusetts, Amherst, Massachusetts, U.S.A. (Secs. 1–3)*

Richard J. Greechie, *Louisiana Tech University, Ruston, Louisiana, U.S.A. (Secs. 1–3)*

Maria Louisa Dalla Chiara and Roberto Giuntini, *Università di Firenze, Florence, Italy (Secs. 4–10)*

Introduction 229
1. **A Brief History of Quantum Logic** 230
1.1 The Origin of Quantum Logic 230
1.2 The Work of Birkhoff and von Neumann 230
1.3 The Orthomodular Law 231
1.4 The Interpretation of Meet and Join 232
2. **Standard Quantum Logic** 232
2.1 The Orthomodular Lattice of Projections on a Hilbert Space 232
2.2 Observables 233
2.3 States 234
2.4 Superposition of States 235
2.5 Dynamics 235
2.6 Combinations of Standard Quantum Logics 236
3. **Orthoalgebras as Models for a General Quantum Logic** ... 236
3.1 Orthoalgebras 236
3.2 Compatibility, Conjunction, and Disjunction in an Orthoalgebra 237
3.3 Probability Measures on and Supports in an Orthoalgebra . 238
3.4 Cartesian and Tensor Products of Orthoalgebras 238
3.5 The Logic of a Physical System 239
3.6 The Canonical Mapping 240
3.7 Critique of Quantum Logic ... 240
4. **The Logician's Approach** 241
5. **Algebraic and Possible-World Semantics** 242
6. **Orthodox Quantum Logic** ... 243
6.1 Semantic Characterizations of QL 243
6.2 An Axiomatization of QL 244
7. **Orthologic and Unsharp Quantum Logics** 245
8. **Hilbert-Space Models of the Brouwer–Zadeh Logics** 247
9. **Partial Quantum Logics** 248
9.1 Algebraic Semantics for WPaQL 249
9.2 An Axiomatization of Partial Quantum Logics 249
10. **Critique of Abstract Quantum Logics** 250
Glossary 250
Works Cited 254

INTRODUCTION

In formulating and studying principles of valid reasoning, logicians have been guided not only by introspection and philosophical reflection, but also by an analysis of various rational procedures commonly employed by mathematicians and scientists. Because these principles have a multitude of disparate sources, efforts to consolidate them in a single coherent system have been unsuccessful. Instead, philosophers, logicians, and mathematicians have created a panoply of competing logical formalisms, each with its own domain of applicability. Among these formalisms are Boolean-based propositional and predicate calculi, modal and multivalued logics, intuitionistic logic, and quantum logic.

Our purpose in this article is to outline the history and present some of the main ideas of quantum logic. In what follows, it will be helpful to keep in mind that there are four levels involved in any exposition of logic and its relation to the experimental sciences.

1. *Philosophical:* Addresses the epistemology

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

of the experimental sciences. Guides and motivates the activities at the remaining levels while assimilating and coordinating the insights gained from these activities.
2. *Syntactic:* Emphasizes the formal structure of a general calculus of experimental propositions.
3. *Semantic:* Focuses on the construction of classes of mathematical models for a logical calculus.
4. *Pragmatic:* Concentrates on a specific mathematical model pertinent to a particular branch of experimental science.

For instance, studies regarding the logics associated with classical physics could be categorized as follows:

1. Philosophical writings extending back at least to Aristotle.
2. Propositional and predicate calculi.
3. The class of Boolean algebras.
4. The Boolean σ algebra of all Borel subsets of the phase space of a mechanical system.

Likewise, for quantum logic, we have

1. Philosophical writings beginning with Schrödinger, von Neumann, Bohr, Einstein, *et al.*
2. Quantum-logical calculi.
3. The class of orthoalgebras.
4. The lattice of projection operators on a Hilbert space.

For expository reasons, our survey proceeds roughly in the order **1, 4, 3, 2.** Thus, we give a brief history of quantum logic in Sec. 1, outline the standard quantum logic of projections on a Hilbert space in Sec. 2, introduce orthoalgebras as models for quantum logic in Sec. 3, and discuss a general quantum-logical calculus of propositions in Sec. 4.

1. BRIEF HISTORY OF QUANTUM LOGIC

1.1 The Origin of Quantum Logic

The publication of John von Neumann's *Mathematische Grundlagen der Quantenmechanik* (1932) was the genesis of a novel system of logical principles based on propositions affiliated with quantum-mechanical entities.

According to von Neumann, a quantum-mechanical system $\mathcal{S}$ is represented mathematically by a separable (i.e., countable dimensional) complex Hilbert space $\mathcal{H}$, observables for $\mathcal{S}$ correspond to self-adjoint operators on $\mathcal{H}$, and the spectrum of a self-adjoint operator is the set of all numerical values that could be obtained by measuring the corresponding observable. Hence, a self-adjoint operator with spectrum consisting at most of the numbers 0 and 1 can be regarded as a *quantum-mechanical proposition* by identifying 0 with "false" and 1 with "true." Since a self-adjoint operator has spectrum contained in $\{0,1\}$ if and only if it is an (orthogonal) projection onto a closed linear subspace of $\mathcal{H}$, von Neumann (1955, p. 253) observed that

> . . . the relation between the properties of a physical system on the one hand, and the projections on the other, makes possible a sort of logical calculus with these.

1.2 The Work of Birkhoff and von Neumann

In 1936, von Neumann, now in collaboration with Garrett Birkhoff, reconsidered the matter of a logical calculus for physical systems and proposed an axiomatic foundation for such a calculus. They argued that the experimental propositions regarding a physical system $\mathcal{S}$ should band together to form a lattice L (Birkhoff, 1967) in which the meet and join operations are formal analogs of the *and* and *or* connectives of classical logic (although they admitted that there could be a question of the experimental meaning of these operations). They also argued that L should be equipped with a mapping carrying each proposition $a \in L$ into its negation $a' \in L$. In present-day terminology, they proposed that L forms an *orthocomplemented lattice* with $\wedge$, $\vee$, and $a \rightarrow a'$ as meet, join, and orthocomplementation, respectively (Kalmbach, 1983; Pták and Pulmannová, 1991).

Birkhoff and von Neumann observed that the experimental propositions concerning a classical mechanical system $\mathcal{S}$ can be identified with members of a field of subsets of the phase space for $\mathcal{S}$, (or, more accurately, with elements of a quotient of such a field by an ideal). In any case, for a classical mechanical system $\mathcal{S}$, they concluded that L forms a *Boolean algebra.*

An orthocomplemented lattice L is a Boolean algebra if and only if it satisfies the *distributive law:*

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z). \qquad (1)$$

An example in which $a \in L$ denotes the observation of a wave packet on one side of a plane, $a' \in L$ its observation on the other side, and $b \in L$ its observation in a state symmetric about the plane shows that

$$b = b \wedge (a \vee a') \neq (b \wedge a) \vee (b \wedge a') = 0,$$

so that the distributive law of classical logic breaks down even for the simplest of quantum-mechanical systems. As Birkhoff and von Neumann observed,

> . . . whereas logicians have usually assumed that properties of negation were the ones least able to withstand a critical analysis, the study of mechanics points to the *distributive identities* as the weakest link in the algebra of logic.

Invoking the desirability of an *"a priori* thermo-dynamic weight of states," Birkhoff and von Neumann argued that L should satisfy a weakened version of Eq. (1), called the *modular law,* and having the following form:

$$\text{If } z \leq x, \quad \text{then } x \wedge (y \vee z) = (x \wedge y) \vee z. \qquad (2)$$

If Eq. (1) holds, then so does Eq. (2) in view of the fact that $z \leq x$ implies $x \wedge z = z$. However, Eq. (2) is weaker than Eq. (1) since the projection operators for a Hilbert space of finite dimension $n \geq 2$ form a modular, but nondistributive, lattice. Thus, Birkhoff and von Neumann proposed an *orthocomplemented modular lattice* as a model for a quantum-mechanical calculus of logic, although they admitted that it would be satisfying if one could interpret the modular law in Eq. (2) "by simpler phenomenological properties of quantum physics."

Birkhoff and von Neumann also gave an example to show that the projection operators on an infinite-dimensional Hilbert space fail to satisfy the modular law. Evidently, von Neumann considered this to be a possible serious flaw of the Hilbert-space formulation of quantum mechanics as proposed in his own *Grundlagen.* Much of von Neumann's work on continuous geometries (1960) and rings of operators (Murray and von Neumann, 1936) was motivated by his desire to construct concrete complemented modular lattices carrying an *"a priori* thermo-dynamic weight of states," that is, a continuous dimension or trace function.

1.3 The Orthomodular Law

Although the projection lattice of an infinite-dimensional Hilbert space fails to satisfy the modular law in Eq. (2), it was discovered by Husimi (1937) that it does satisfy the following weaker condition, now called the *orthomodular law:*

$$\text{If } z \leq x, \quad \text{then } x = (x \wedge z') \vee z. \qquad (3)$$

If Eq. (2) holds, then so does Eq. (3) in view of the fact that $x = x \wedge (z' \vee z)$. The same condition was rediscovered independently by Loomis (1955) and Maeda (1955) in connection with their work on extension of the Murray–von Neumann dimension theory of rings of operators to orthocomplemented lattices. An orthocomplemented lattice satisfying Eq. (3) is called an *orthomodular lattice.*

In 1957, Mackey published an expository article on quantum mechanics in Hilbert space based on notes for lectures that he was then giving at Harvard. These notes were later published in the form of a monograph (Mackey, 1963) in which the basic principles of quantum mechanics were introduced in terms of a function

$$p = \text{Prob}(A,\psi,E) \qquad (4)$$

interpreted as the probability p that a measurement of the observable A in state ψ results in a value in a set of E of real numbers. The square A^2 of A is then defined by the condition

$$\text{Prob}(A^2,\psi,E) = \text{Prob}(A,\psi,F),$$

where F is the set of all real numbers x such that $x^2 \in E$. If $A = A^2$, then A is called a *question*. Under certain more or less reasonable hypotheses, it can be shown that the set of all questions forms an orthomodular lattice L.

The generality of Mackey's formulation and the natural way in which Mackey's questions give rise to an orthomodular lattice engendered the heady idea of a universal logical calculus for all of the experimental sciences. Such a calculus would be based on the class of all orthomodular lattices—including the Boolean algebras that would serve as models for the logics affiliated with classical mechanical systems. Would this be the realization of Leibniz's dream of a *calculus ratiocinator?* This captivating thought helped to motivate an ongoing study of the theory of orthomodular lattices by a relatively small but devoted group of researchers. An authoritative account of the resulting theory of orthomodular lattices as developed up to about 1983 can be found in Kalmbach (1983).

1.4 The Interpretation of Meet and Join

In spite of the appeal of a general scientific logic based on orthomodular lattices, a nagging question raised in the 1936 paper of Birkhoff and von Neumann was still unresolved. If, for a quantum-mechanical system, most pairs of observations are incompatible and cannot be made simultaneously, what experimental meaning can one attach to the meet $p \wedge q$ of two propositions? Two and a half decades after his initial paper with von Neumann, Birkhoff returned to this question (Birkhoff, 1961), calling for an autonomous quantum logic that draws its authority directly from experiments. (A similar question arises in connection with the logic of relativistic physics where the traditional notion of simultaneity is meaningless for spatially separated events.) After all, simultaneity is an indispensable constituent of classical propositional conjunction.

An obvious way to avoid the interpretation issue for $p \wedge q$ is to replace the assumption that $p \wedge q$ always exists with the weaker assumption that it exists if p and q are *compatible* in the sense that they can be *simultaneously tested* by means of a single experiment. In this connection, Birkhoff and von Neumann were careful to point out that "... one may regard a set of compatible measurements as a single composite 'measurement'."

Thus, for compatible propositions, experimental meaning can be bestowed upon the meet and join by regarding these connectives as the conjunction and disjunction in the usual sense of classical logic.

Although the mainstream effort to develop a viable quantum logic has concentrated on the use of orthomodular lattices as the basic models (Jauch, 1968; Piron, 1976; Mittelstaedt, 1978; Beltrametti and Cassinelli, 1981), alternative models have been introduced that avoid the interpretation issue for meet and join by invoking the notion of compatibility. Among these are the *orthomodular posets* introduced in the early 1960s (Foulis, 1962) and the *orthoalgebras* proposed in the late 1970s (Randall and Foulis, 1978; Hardegree and Frazer, 1981; Lock and Hardegree, 1984a,b). Thus, the evolution of quantum logic from the 1930s to the present has been the story of a slow retreat from Boolean-algebra–based logic and the concurrent development of more and more general mathematical models.

2. STANDARD QUANTUM LOGIC

2.1 The Orthomodular Lattice of Projections on a Hilbert Space

As a mathematical model for a calculus of quantum logic, the orthomodular lattice L of projection operators on a Hilbert space $\mathcal{H}$ is called a *standard quantum logic*. Wilbur (1977) has given a purely lattice-theoretic characterization of the standard quantum logics.

In the present section, we sketch the theory of standard quantum logics, considering only the special case in which $\mathcal{H}$ is a separable Hilbert space of dimension at least three over the complex number field $\mathbb{C}$. Thus, we leave aside real or quaternionic Hilbert spaces as well as the generalized Hilbert spaces of Gross and Keller (Keller, 1980). We regard $\mathcal{H}$ as the Hilbert space corresponding to a quantum-mechanical system $\mathcal{S}$. (For the time being, we do not consider superselection rules.)

If A is a bounded operator on $\mathcal{H}$, we denote by A^* the adjoint of A. Thus, $\langle A\psi|\phi\rangle = \langle\psi|A^*\phi\rangle$ for all $\psi, \phi \in \mathcal{H}$. A bounded operator

P on $\mathcal{H}$ is called a *projection* if $P = P^* = P^2$, and we define $L = L(\mathcal{H})$ to be the set of all such projection operators. If $P \in L$ and

$$\mathcal{M} = P(\mathcal{H}) = \{P(\psi) | \psi \in \mathcal{H}\} \tag{5}$$

is the range of P, then $\mathcal{M}$ is a closed linear subspace of $\mathcal{H}$; conversely, every closed linear subspace $\mathcal{M}$ of $\mathcal{H}$ is of the form given in Eq. (5) for a uniquely determined $P \in L$. If P and $\mathcal{M}$ are related as in Eq. (5), we say that P is the *projection onto* $\mathcal{M}$. The zero operator $\mathbb{0}$ is the projection onto $\{0\}$ and the identity operator $\mathbb{1}$ is the projection onto $\mathcal{H}$.

If P is the projection onto $\mathcal{M}$ and Q is the projection onto $\mathcal{N}$, we write $P \leq Q$ if and only if $\mathcal{M}$ is a linear subspace of $\mathcal{N}$. Thus, L is a partially ordered set (poset) under $\leq$. If $\mathcal{M}$ is a closed linear subspace of $\mathcal{H}$, we write $\mathcal{M}^\perp$ for the set of all vectors in $\mathcal{H}$ that are orthogonal to every vector in $\mathcal{M}$. Then $\mathcal{M}^\perp$ is again a closed linear subspace of $\mathcal{H}$. If P is the projection onto $\mathcal{M}$, we write the projection onto $\mathcal{M}^\perp$ as P'. Note that

$$P' = \mathbb{1} - P, \quad (P')' = P, \quad \mathbb{0}' = \mathbb{1}, \quad \text{and } \mathbb{1}' = \mathbb{0}.$$

Furthermore, if $P, Q \in L$ with $P \leq Q$, then $Q' \leq P'$.

If $\mathcal{M}$ and $\mathcal{N}$ are closed linear subspaces of $\mathcal{H}$, then so is the set-theoretic intersection $\mathcal{M} \cap \mathcal{N}$. If P is the projection onto $\mathcal{M}$ and Q is the projection onto $\mathcal{N}$, we define $P \wedge Q$ to be the projection onto $\mathcal{M} \cap \mathcal{N}$, noting that $P \wedge Q$ is the meet of P and Q in the poset L. If we define $P \vee Q = (P' \wedge Q')'$, we find that $P \vee Q$ is the join of P and Q in L. Also, $P \wedge P' = \mathbb{0}$ and $P \vee P' = \mathbb{1}$, and so L forms a lattice that is orthocomplemented by $P \to P'$. Furthermore, L satisfies Eq. (3), and hence it is an orthomodular lattice.

If P is the projection onto $\mathcal{M}$ and Q is the projection onto $\mathcal{N}$, then $\mathcal{M}$ is a linear subspace of $\mathcal{N}^\perp$ if and only if $P \leq Q'$. If $P \leq Q'$, we say that P and Q are *orthogonal* to each other and write $P \perp Q$. It can be shown that $P \perp Q$ if and only if $P + Q$ is again a projection operator, in which case, $P + Q = P \vee Q$.

If $(\mathcal{M}_\alpha)$ is a family of closed linear subspaces of $\mathcal{H}$, then the set-theoretic intersection $\cap_\alpha \mathcal{M}_\alpha$ is again a closed linear subspace of $\mathcal{H}$. If P_α is the projection onto $\mathcal{M}_\alpha$ for all α and P is the projection onto $\cap_\alpha \mathcal{M}_\alpha$, then P is the *greatest lower bound* in L of the family (P_α), and we write $\wedge_\alpha P_\alpha = P$. Likewise, $(\wedge_\alpha (P_\alpha)')'$ is the *least upper bound* on L of the family (P_α), and we write $\vee_\alpha P_\alpha = (\wedge_\alpha (P_\alpha)')'$. Consequently, the standard quantum logic L is actually a *complete* orthomodular lattice.

The interpretation of L as a model for a logic of quantum mechanics is based on the following premise:

> *The two-valued (true/false), experimentally testable propositions for the quantum-mechanical system* $\mathcal{S}$ *are represented by the projections in the standard quantum logic* L *for the Hilbert space* $\mathcal{H}$ *corresponding to* $\mathcal{S}$.

Furthermore, if each experimental proposition for $\mathcal{S}$ is identified with its corresponding projection $P \in L$, it is assumed that

> *if* P, Q ∈ L, *then* P ≤ Q *holds if and only if* P *and* Q *are simultaneously testable and, whenever they are both tested and* P *is found to be true, then* Q *will also be true.*

2.2 Observables

As is customary, we assume that an *observable* or *dynamical variable* for the quantum-mechanical system $\mathcal{S}$ is represented by a (not necessarily bounded) self-adjoint operator A on the Hilbert space $\mathcal{H}$. In particular, then, each projection operator $P \in L$ represents an observable that, when measured, can only produce the values 1 (true) or 0 (false). As we shall see, the connection between general observables and projection observables is effected by the celebrated *spectral theorem*.

The smallest collection of subsets of the real numbers $\mathbb{R}$ that contains all open intervals and is closed under the formation of complements and countable unions is called the σ field of *real Borel sets*. A *spectral measure* is a mapping $E \to P_E$ from real Borel sets into projections such that $P_\varnothing = \mathbb{0}$, $P_\mathbb{R} = \mathbb{1}$, and, for every pairwise disjoint sequence $E_1, E_2, E_3, \ldots$ of real Borel sets,

$$\bigvee_{k=1}^{\infty} P_{E_k} = P_{E_1 \cup E_2 \cup E_3 \cup \cdots}$$

If $E \to P_E$ is a spectral measure, $\lambda \in \mathbb{R}$, and J

$= (-\infty,\lambda]$, define $P_\lambda = P_J$. Then, by the *spectral theorem* there is a one-to-one correspondence between observables A and spectral measures $E \rightarrow P_E$ such that

$$A = \int_{-\infty}^{\infty} \lambda dP_\lambda.$$

The projections in the family (P_E) are called the *spectral projections* for the observable A.

We can now be quite explicit about the connection between observables in general and projection observables in particular. Suppose that A is an observable and that (P_E) is the corresponding family of spectral projections. Then

> $\mathrm{P_E}$ *represents the experimental proposition asserting that a measurement of the observable* A *yields a result* r *that belongs to the real Borel set* $\mathrm{E} \subseteq \mathbb{R}$.

In quantum mechanics it is understood that a family of observables (A_α) is *compatible* (that is, *jointly* or *simultaneously observable*) if and only if $A_\alpha A_\beta = A_\beta A_\alpha$ for all α,β (that is, if and only if the observables commute with each other). On the basis of this understanding, we can state the following:

> *A family of projections* $(\mathrm{P}_\alpha) \subseteq \mathrm{L}$ *is compatible (that is, simultaneously testable) if and only if* $\mathrm{P}_\alpha \mathrm{P}_\beta = \mathrm{P}_\beta \mathrm{P}_\alpha$ *for all projections in the family.*

It can be shown that two observables commute with each other if and only if their spectral projections commute with each other.

It is interesting to note that the question of whether or not two projections commute can be settled in *purely lattice-theoretic terms.* Indeed, for $P,Q \in L$,

$$PQ = QP \quad \text{if and only if} \quad P = (P \wedge Q) \vee (P \wedge Q').$$

The equation stating the condition is a special case of the distributive law in Eq. (1); hence, in a standard quantum logic, the failure of the distributive law is a direct consequence of the fact that there are incompatible pairs of quantum-mechanical ·observables.

2.3 States

A bounded self-adjoint operator W on $\mathcal{H}$ is said to be *nonnegative* if $\langle W\psi|\psi\rangle \geq 0$ for all $\psi \in \mathcal{H}$. A nonnegative operator W belongs to the *trace class* if the series

$$\mathrm{tr}(W) = \sum_{\psi \in B} \langle W\psi|\psi\rangle$$

converges for an orthonormal basis $B \subseteq \mathcal{H}$. Convergence on any one orthonormal basis implies convergence on all orthonormal bases.

A *(von Neumann) density operator* on $\mathcal{H}$ is a bounded, self-adjoint, nonnegative, trace-class operator W on $\mathcal{H}$ such that $\mathrm{tr}(W) = 1$. Denote by $\Omega = \Omega(\mathcal{H})$ the set of all density operators on $\mathcal{H}$. One of the basic assumptions of statistical quantum mechanics is the following:

> *There is a one-to-one correspondence between the possible states of the system* $\mathcal{S}$ *and the density operators* $\mathrm{W} \in \Omega$ *such that, for every experimental proposition* $\mathrm{P} \in \mathrm{L}$, *tr(*WP*) is the probability that* P *will be true when tested in the state corresponding to* W.

In accordance with this assumption, we shall identify each possible state of the system $\mathcal{S}$ with the corresponding density operator W. In particular, if A is an observable with spectral family (P_E), the probability function, Eq. (4) in Mackey's formulation, is realized as

$$\mathrm{Prob}(A,W,E) = \mathrm{tr}(WP_E). \tag{6}$$

Equation (6), one of the fundamental equations of quantum mechanics, says that

> *the probability that a measurement of the observable* A *in the state* W *yields a result* r *in the Borel set* E *is given by tr(*$\mathrm{WP_E}$*).*

By a *countably additive probability measure* on the orthomodular lattice L is meant a function $\omega{:}L \rightarrow [0,1] \subseteq \mathbb{R}$ such that, for every sequence $P_1, P_2, P_3, \ldots$ of pairwise orthogonal projections in L,

$$\omega(\vee_k P_k) = \sum_k \omega(P_k).$$

By a celebrated theorem of Gleason (1957),

ω is a countably additive probability measure on L if and only if there is a (uniquely determined) density operator $W \in \Omega$ such that

$$\omega(P) = \mathrm{tr}(WP) \quad \text{for all } P \in L.$$

2.4 Superposition of States

If $W_1, W_2, W_3, \ldots \in \Omega$ is a sequence of density operators and $t_1, t_2, t_3, \ldots$ is a corresponding sequence of nonnegative real numbers such that $\Sigma t_k = 1$, then $W = \Sigma_k t_k W_k$ is again a density operator, which is referred to as a *mixture* or an *incoherent superposition* of the states W_k. For instance, W could be regarded as the state of a statistical ensemble of systems for which t_k is the fraction of the systems that are in the state W_k.

A state W is called a *pure state* if it cannot be obtained as a mixture of other states. It is customary to assume that individual physical systems are always in a pure state and that mixed states apply only to statistical ensembles of systems each of which is in a pure state, or to physical systems that are interactively coupled with other physical systems.

It can be shown that $W \in \Omega$ is a pure state if and only if it is a projection onto a one-dimensional linear subspace $\mathcal{M}$ of $\mathcal{H}$. Thus, any normalized vector $\psi \in \mathcal{H}$ determines a unique pure state, namely the projection onto the linear subspace of complex multiples of ψ. Such a state is called a *vector state*, and two normalized vectors determine the same vector state if and only if each can be obtained from the other by multiplying by a complex number of modulus 1 (a *phase factor*). Every state W is a mixture of pure (that is, vector) states.

We define the *support* of $W \in \Omega$, in symbols supp(W), to be the set of all $P \in L$ such that $\mathrm{tr}(WP) \neq 0$. This is the same as the set of all $P \in L$ for which $WP \neq 0$. If $W = \Sigma_k t_k W_k$ is an incoherent superposition of the sequence (W_k), then it is clear that supp(W) is contained in the set-theoretic union $\cup_k \mathrm{supp}(W_k)$.

More generally, if (W_α) is a family of states, we say that the state W is a *superposition* of the states W_α if and only if

$$\mathrm{supp}(W) \subseteq \cup_\alpha \mathrm{supp}(W_\alpha)$$

(Bennett and Foulis, 1990). If W as well as every W_α is a pure state, then W is said to be a *coherent superposition* of the states W_α. For instance, if W is the vector state determined by $\psi \in \mathcal{H}$, each W_α is the vector state determined by $\psi_\alpha \in \mathcal{H}$, and ψ differs from a normalized linear combination of the ψ_α by a phase factor, then W is a coherent superposition of the W_α.

2.5 Dynamics

By *dynamics* is meant a study of the way in which the states (*Schrödinger picture*) or the observables (*Heisenberg picture*) of a system change or evolve in time. The Schrödinger and Heisenberg pictures are mathematically equivalent. For definiteness, we adopt the Schrödinger picture. Thus, if the space Ω of density operators represents the state space of the quantum-mechanical system $\mathcal{S}$, then the dynamical evolution of the system is represented by a function $f(t,W)$ of the time $t \in \mathbb{R}$ and the state $W \in \Omega$ such that

$$f(t,W) \in \Omega, \quad f(0,W) = W \quad \text{and} \quad f(t+s,W) = f(t,f(s,W)). \tag{7}$$

The understanding in Eq. (7) is that $f(t,W)$ represents the state of the system after a time interval t if it is in state W at time 0. The function f is called the *dynamical law* for the system $\mathcal{S}$.

If the dynamical law f in Eq. (7) preserves superpositions, and is continuous in a suitable sense, it can be shown (Mackey, 1963) that there is a family (U_t) of unitary operators continuously indexed by real numbers such that

$$f(t,W) = U_t W U_t^{-1}$$

holds for all $t \in \mathbb{R}$. Hence, by a celebrated representation theorem of Stone (1932), it follows that there is a self-adjoint operator H on $\mathcal{H}$ such that

$$U_t = e^{-itH} \tag{8}$$

for all $t \in \mathbb{R}$. Equation (8) is the operator form of the *Schrödinger equation* and H is the *Hamiltonian operator* for the system.

2.6 Combinations of Standard Quantum Logics

Suppose that $\mathcal{H}_1$ and $\mathcal{H}_2$ are complex separable Hilbert spaces with corresponding standard quantum logics L_1 and L_2. There are two natural ways to combine $\mathcal{H}_1$ and $\mathcal{H}_2$ to form a composite Hilbert space $\mathcal{H}$ with its own standard quantum logic L: We can form either the direct sum $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ or the tensor product $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ (Foulis, 1989). In neither case is the structure of the resulting standard quantum logic L easy to describe in terms of the structures of L_1 and L_2.

If $\mathcal{S}_1$ and $\mathcal{S}_2$ are quantum-mechanical systems represented by corresponding Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, it is customary to regard the tensor product $\mathcal{H} = \mathcal{H}_2 \otimes \mathcal{H}_2$ as the Hilbert space corresponding to the "combined system" $\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2$ (Jauch, 1968). If this is so, then in the combination $\mathcal{S}_1 + \mathcal{S}_2$ the systems can be tightly correlated, but they cannot exert instantaneous influences on each other (Kläy *et al.*, 1987).

If W is a state for the combined system $\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2$, there exist uniquely determined states W_1 for $\mathcal{S}_1$ and W_2 for $\mathcal{S}_2$ such that for all $P_1 \in L_1$ and all $P_2 \in L_2$

$$\mathrm{tr}(W_1P_1) = \mathrm{tr}(W(P_1 \otimes \mathbb{1})) \quad \text{and}$$
$$\mathrm{tr}(W_2P_2) = \mathrm{tr}(W(\mathbb{1} \otimes P_2)).$$

The states W_1 and W_2 are called *reduced states*. In general, W is not determined by W_1 and W_2, but depends on the details of the coupling between $\mathcal{S}_1$ and $\mathcal{S}_2$. However, if W is a pure state and either W_1 or W_2 is pure, then both W_1 and W_2 are pure and $W = W_1 \otimes W_2$. Therefore, if $\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2$ is in a pure state and if $\mathcal{S}_1$ and $\mathcal{S}_2$ are correlated in any way, then neither $\mathcal{S}_1$ nor $\mathcal{S}_2$ can be in a pure state.

If $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$, a *superselection rule* (Wick *et al.*, 1952) may be imposed, in which case the quantum logic L associated with $\mathcal{H}$ is understood to consist only of projections that commute with the projections P_1 and P_2 of $\mathcal{H}$ onto the subspaces $\mathcal{H}_1 \oplus \{0\}$ and $\{0\} \oplus \mathcal{H}_2$, respectively. In this case, L is isomorphic to the Cartesian product $L_1 \times L_2$ of the standard quantum logics L_1 and L_2, but L is no longer a standard quantum logic. If such a superselection rule is imposed, it is assumed that the superselected observables are those with spectral projections in L and the superselected states are represented by density operators W that commute with both P_1 and P_2.

3. ORTHOALGEBRAS AS MODELS FOR A GENERAL QUANTUM LOGIC

3.1 Orthoalgebras

In this section, we present an axiomatic mathematical structure called an *orthoalgebra* (Foulis *et al.*, 1992), which generalizes the standard quantum logics. The idea is to endow a generic orthoalgebra with an *absolute minimum of mathematical structure* so that it becomes possible to investigate the meaning and consequences of the special features that distinguish particular orthoalgebras—for instance, Boolean algebras, orthomodular lattices, or standard quantum logics—as models for a calculus of experimental propositions.

By definition, an *orthoalgebra* is a set L containing two special elements 0 and 1 and equipped with a relation $\perp$ called *orthogonality* such that, for each pair $p,q \in L$ with $p \perp q$, an *orthogonal sum* $p \oplus q$ is defined in L and subject to the following four axioms:

(*Commutativity*) If $p \perp q$, then $q \perp p$ and $p \oplus q = q \oplus p$.

(*Associativity*) If $p \perp q$ and $(p \oplus q) \perp r$, then $q \perp r$, $p \perp (q \oplus r)$, and $p \oplus (q \oplus r) = (p \oplus q) \oplus r$.

(*Orthocomplementation*) For each $p \in L$ there is a unique $p' \in L$ such that $p \perp p'$ and $p \oplus p' = 1$.

(*Consistency*) If $p \perp p$, then $p = 0$.

We note that every orthomodular lattice L becomes an orthoalgebra if we define $p \oplus q = p \vee q$ whenever $p \leq q'$. In particular, every Boolean algebra and every standard quantum logic is an orthoalgebra.

If L is an orthoalgebra and $p,q \in L$, we define $p \leq q$ to mean that there exists $r \in L$ with $p \perp r$ such that $p \oplus r = q$. It can be shown that L is partially ordered by $\leq$; $0 \leq p \leq 1$ and $p = p''$ hold for all $p \in L$; and, if $p \leq q$, then $q' \leq p'$. Also, if $p \perp q$, then with respect to $\leq$, $p \oplus q$ is a *minimal* upper bound for p and q; that is,

$$p,q \leq p \oplus q \text{ and there exists no } r \in L \text{ with } p,q \leq r < p \oplus q.$$

However, $p \oplus q$ may not be the *least* upper bound for p and q; that is, the conditions $r \in L$ and $p, q \leq r$ do not necessarily imply that $p \oplus q \leq r$.

If $x,y \in L$ have a least upper bound (respectively, a greatest lower bound), we write it as $x \vee y$ (respectively, as $x \wedge y$). By definition, an *orthomodular poset* is an orthoalgebra L satisfying the condition that $p \oplus q = p \vee q$ whenever $p \perp q$. An orthomodular lattice is the same thing as an orthoalgebra in which every pair of elements x, y has a meet $x \wedge y$ and a join $x \vee y$. A Boolean algebra is the same thing as an orthomodular lattice satisfying the condition that $x \wedge y = 0$ only if $x \perp y$.

By a *subalgebra* of the orthoalgebra L, we mean a subset $S \subseteq L$ such that $0, 1 \in S$ and, if $p, q \in S$ with $p \perp q$, then $p \oplus q \in S$ and $p' \in S$. Evidently, a subalgebra of an orthoalgebra is an orthoalgebra in its own right under the operations inherited from the parent orthoalgebra. If, as an orthoalgebra in its own right, a subalgebra B of L is a Boolean algebra, we refer to B as a *Boolean subalgebra* of L. If $p \perp q$ in L, then

$$B = \{0,1,p,q,p',q',p \oplus q,(p \oplus q)'\}$$

is a Boolean subalgebra of L, so L is a set-theoretic union of Boolean subalgebras.

3.2 Compatibility, Conjunction, and Disjunction in an Orthoalgebra

We say that a subset of an orthoalgebra L is a *compatible* set if it is contained in a Boolean subalgebra of L. A compatible set of pairwise orthogonal elements is called an *orthogonal* subset of L. If L is a standard quantum logic, then a subset of L is compatible if and only if the projections in the subset commute with one another.

Let $A = \{a_1,a_2,a_3, \ldots ,a_n\}$ be a finite orthogonal subset of L. Then, it can be shown that the least upper bound

$$\vee_B A = a_1 \vee_B a_2 \vee_B a_3 \vee_B \cdots \vee_B a_n$$

as calculated in any Boolean subalgebra B of L that contains A is independent of the choice of B. Thus, we define the *orthogonal sum*

$$\oplus A = \vee_B A$$

as calculated in any such B. If C and D are finite orthogonal subsets of L, then $\oplus C \perp \oplus D$ if and only if $C \cap D \subseteq \{0\}$ and $C \cup D$ is an orthogonal set, in which case $\oplus C \oplus \oplus D = \oplus(C \cup D)$. If $A = \{a_1,a_2\}$, then $\oplus A = a_1 \oplus a_2$. Thus, if $A = \{a_1,a_2,a_3, \ldots ,a_n\}$, is an orthogonal set, we can define

$$a_1 \oplus a_2 \oplus a_3 \oplus \cdots \oplus a_n = \oplus A$$

without notational conflict.

If $p,q \in L$ and both p and q belong to a Boolean subalgebra B of L, then the greatest lower bound $p \wedge_B q$ and the least upper bound $p \vee_B q$ of p and q as calculated in B may well depend on the choice of B. If $p \wedge_B q$ is independent of the choice of B, we define the *conjunction* p & q of p and q by

$$p \,\&\, q = p \wedge_B q.$$

Likewise, if $p \vee_B q$ is independent of the choice of B, we define the *disjunction* $p + q$ of p and q by

$$p + q = p \vee_B q.$$

It can be shown that the compatible elements p and q have a conjunction if and only if they have a disjunction. Furthermore, if $p \,\&\, q$ and $p + q$ exist, then $p \,\&\, q$ is a maximal lower bound and $p + q$ is a minimal upper bound for p and q in L. If p and q are compatible and at least one of $p \wedge q$ or $p \vee q$ exists in L, then $p \,\&\, q$ and $p + q$ exist, $p \,\&\, q = p \wedge q$, and $p + q = p \vee q$. If $p \,\&\, q$ exists, then so do $p' \,\&\, q'$ and $p' + q'$, and we have $p + q = (p' \,\&\, q')'$ and $p \,\&\, q = (p' + q')'$. If $p \perp q$, then $p + q = p \oplus q$ and $p \,\&\, q = 0$.

If L is an orthomodular poset, then any two compatible elements $p, q \in L$ have a conjunction $p \,\&\, q = p \wedge q$ and a disjunction $p + q = p \vee q$; however, there are orthoalgebras containing compatible pairs of elements that do not admit conjunctions or disjunctions. There are non-Boolean orthoalgebras in which every pair of elements forms a compatible set. There exist orthomodular posets containing three elements that are pairwise compatible, but that do not form a compatible set; however, in an orthomodular poset, every pairwise orthogonal subset is an orthogonal subset. There exist orthoalgebras containing three elements that are pairwise orthogonal, but do not form an orthogonal set.

3.3 Probability Measures on and Supports in an Orthoalgebra

By a *probability measure* on an orthoalgebra L, we mean a mapping $\omega{:}L \to [0,1] \subseteq \mathbb{R}$ such that, for $p,q \in L$, with $p \perp q$,

$$\omega(p \oplus q) = \omega(p) + \omega(q). \tag{8}$$

It is possible to define *σ-complete* orthoalgebras and *countably additive* probability measures thereon, and thus extend Eq. (8) to sequences in L, but we do not do so here. The set of all probability measures on L is denoted by $\Omega = \Omega(L)$. Evidently, Ω is a convex subset of the vector space of all real-valued functions on L.

If the elements of L are regarded as representing two-valued experimental propositions concerning a physical system $\mathcal{S}$, then a probability measure $\omega \in \Omega$ may be interpreted in any of the following ways:

(*Frequency*) ω is a *complete stochastic model* for $\mathcal{S}$ in the sense that $\omega(p)$ is the "long-run relative frequency" with which the proposition $p \in L$ will be true when repeatedly tested (D'Espagnat, 1971).

(*Subjective*) ω is a *model for coherent belief* encoding all of our current information about the system $\mathcal{S}$. Thus, if $p \in L$, then $\omega(p)$ measures our current "degree of belief," on a scale from 0 to 1, in the truth of the proposition p (Jaynes, 1989).

(*Propensity*) $\omega(p)$ is a measure on a scale from 0 to 1 of the "propensity" of the system $\mathcal{S}$ to produce the outcome 1 (= true) when the proposition p is tested (Popper, 1959).

(*Mathematical*) ω is a mathematical artifact that may be of use in making inferences about $\mathcal{S}$ using data secured by making measurements on $\mathcal{S}$ (Kolmogorov, 1933).

For $\omega \in \Omega$, we define the *support* of ω by

$$\text{supp}(\omega) = \{p \in L \mid \omega(p) > 0\}.$$

If $S = \text{supp}(\omega)$, then $1 \in S$ and, for all $p, q \in L$ with $p \perp q$,

$$p \oplus q \in S \quad \text{if and only if } p \in S \text{ or } q \in S. \tag{9}$$

A subset S of L such that Eq. (9) holds is called a *support* in L. In general, there are supports $S \subseteq L$ that are not of the form $\text{supp}(\omega)$ for $\omega \in \Omega$; those that are of this form are called *stochastic supports.* The set-theoretic union of supports is again a support, and it follows that the collection of all supports in L forms a complete lattice under set-theoretic inclusion.

3.4 Cartesian and Tensor Products of Orthoalgebras

If L_1 and L_2 are orthoalgebras, the Cartesian product $L_1 \times L_2$ becomes an orthoalgebra under the obvious componentwise operations. If L_1 is identified with $L_1 \times \{0\}$ and L_2 is identified with $\{0\} \times L_2$ in $L_1 \times L_2$, then every element in $L_1 \times L_2$ can be written uniquely in the form $p \oplus q$ with $p \in L_1$ and $q \in L_2$. This construction generalizes the superselected direct sum of standard quantum logics. Just as is the case for standard quantum logics, $\Omega(L_1 \times L_2)$ is isomorphic in a natural way to the convex hull of $\Omega(L_1)$ and $\Omega(L_2)$.

A construction for the tensor product $L_1 \otimes L_2$ of orthoalgebras based on Foulis and Randall (1981) can be found in Lock (1981). The factors L_1 and L_2 are embedded in the orthoalgebra $L_1 \otimes L_2$ by mappings $p \to p \otimes 1$ and $q \to 1 \otimes q$ for $p \in L_1$, $q \in L_2$ in such a way that $p \otimes 1$ and $1 \otimes q$ are compatible and have a conjunction $(p \otimes 1)$ & $(1 \otimes q) = p \otimes q$. Furthermore, elements of the form $p \otimes q$ generate $L_1 \otimes L_2$. If $\alpha \in \Omega(L_1)$ and $\beta \in \Omega(L_2)$, there is a unique $\gamma = \alpha\beta \in \Omega(L_1 \otimes L_2)$ such that $\gamma(p \otimes q) = \alpha(p)\beta(q)$ for all $p \in L_1$, $q \in L_2$. A probability measure on $L_1 \otimes L_2$ of the form $\alpha\beta$ is said to be *factorizable,* and a convex combination of factorizable probability measures is said to be *separable* (Kläy, 1988). The existence of probability measures on $L_1 \otimes L_2$ that are not separable seems to be a characteristic feature of the tensor product of non-Boolean orthoalgebras.

3.5 The Logic of a Physical System

We are now in a position to summarize the *quantum logic approach* to the study of *physical systems* (quantum-mechanical or not). The *basic postulate of quantum logic* for a physical system $\mathcal{S}$ is as follows:

> *(Logic postulate) The set* L *of all two-valued, experimentally testable propositions for* $\mathcal{S}$ *has the structure of an orthoalgebra such that every simultaneously testable set of propositions forms a compatible subset of* L *and every finite compatible subset of* L *is a simultaneously testable set of propositions. If* p,q $\in$ L *with* p $\perp$ q, *and if* p, q, *and* p $\oplus$ q *are tested simultaneously, then at most one of the propositions* p,q *will be true, and* p $\oplus$ q *will be true if and only if either* p *or* q *is true.*

We refer to L as the *logic* of the system $\mathcal{S}$.

It is customary to assume that there is a *state space* Ψ associated with the physical system $\mathcal{S}$. The elements $\psi \in \Psi$ are called *states,* and, at any given moment, $\mathcal{S}$ is presumed to be *in* one and only one state $\psi \in \Psi$. A state is supposed to encode all available information about the consequences of performing tests or making measurements on $\mathcal{S}$ when $\mathcal{S}$ is in that state.

Whereas the truth or falsity of an experimental proposition $p \in L$ can be determined by a suitable test, it may or may not be possible to determine the current state $\psi \in \Psi$ of $\mathcal{S}$ by a test or measurement; however, it may be possible to bring $\mathcal{S}$ into a state ψ by means of a suitable *state-preparation procedure.* The state of the system $\mathcal{S}$ can change under the action of a *dynamical law,* under a *state collapse* when an observer tests a proposition or measures an observable, because a *state-preparation procedure* is executed, or simply by virtue of a spontaneous *state transition.*

A connection between the state space Ψ for $\mathcal{S}$ and its logic L is effected as follows:

> *(Stochastic postulate) Each state* $\psi \in \Psi$ *determines a corresponding probability measure* ω_ψ *on* L *in such a way that, for* p $\in$ L, ω_ψ(p) *is the probability that the proposition* p *is true when tested with the system* $\mathcal{S}$ *in the state* ψ.

In the stochastic postulate, the probability measure ω_ψ can be interpreted in any of the four ways (frequency, subjective, propensity, mathematical) suggested in Sec. 3.3. In what follows, we denote by Σ the subset of $\Omega(L)$ consisting of all probability measures of the form ω_ψ for $\psi \in \Psi$, and we refer to each $\omega_\psi \in \Sigma$ as a *probability state* for the system $\mathcal{S}$. It is customary to identify the state ψ with the corresponding probability state ω_ψ and to speak of the elements in Σ as *states* for $\mathcal{S}$. Although this custom can lead to philosophical and mathematical difficulties (what if $\phi,\psi \in \Psi$, $\phi \neq \psi$, and yet $\omega_\phi = \omega_\psi$?), we shall follow it in the interests of simplicity.

Let $\omega \in \Sigma$ be a state and let $p \in L$ be an experimental proposition for the physical system $\mathcal{S}$. We say that p is *possible, impossible,* or *certain* in the state ω if $p \in \text{supp}(\omega)$, $p \notin \text{supp}(\omega)$, or $p' \notin \text{supp}(\omega)$, respectively. If both p and p' are possible, we say that p is *contingent* in the state ω. The state space Σ is said to be *unital* if every nonzero $p \in L$ is certain in at least one state $\omega \in \Sigma$.

If $\Lambda \subseteq \Sigma$ is a set of states, then a state $\omega \in \Sigma$ is said to be a *superposition* of the states in Λ if

$$\text{supp}(\omega) \subseteq \cup\{\text{supp}(\lambda) | \lambda \in \Lambda\}.$$

The *superposition closure* of Λ is defined to be the set Λ^{sp} of all superpositions of states in Λ. If $\Lambda = \Lambda^{sp}$, then Λ is called *superposition closed.* A state ω is *pure* if the set $\{\omega\}$ is superposition closed. If ω is a pure state, Λ is a set of pure states, and $\omega \in \Lambda^{sp}$, then ω is a *coherent superposition* of the states in Λ. In what follows, we denote by $\mathcal{L}$ the set of all superposition-closed subsets of Σ. Note that $\mathcal{L}$ is closed under set-theoretic intersection, and hence, it forms a complete lattice under set-theoretic inclusion.

It has long been a tenet of natural philosophy that affiliated with a physical system $\mathcal{S}$ is a class $\mathcal{A}$ of *attributes* or *properties.* At any given moment, some of these attributes may be *actual,* while the others are only *potential.* The attributes of $\mathcal{S}$ that are always actual are its *intrinsic* attributes; those that can be either actual or potential are its *accidental* attributes. The charge of an electron is one of its intrinsic attributes, whereas the attribute "spin up in the z direction" is accidental.

To each attribute $A \in \mathcal{A}$ there corresponds a set $\Lambda_A \subseteq \Sigma$ consisting precisely of those states ω such that A is actual whenever $\mathscr{S}$ is in the state ω. A heuristic argument, which we omit here, indicates that Λ_A should be superposition closed, so that $\Lambda_A \in \mathscr{L}$. Similar arguments suggest that every element of $\mathscr{L}$ corresponds in this way to an attribute, and thus lead us to our third postulate:

> *(Attribute postulate) Each attribute* A *determines a corresponding superposition-closed subset* Λ_A *of the state space* Σ *such that* A *is actual if and only if the system* $\mathscr{S}$ *is in a state* $\omega \in \Lambda_A$*; furthermore, every* $\Lambda \in \mathscr{L}$ *has the form* Λ_A *for some* $A \in \mathscr{A}$.

Just as we identified states with probability states, we propose to identify elements Λ of the complete lattice $\mathscr{L}$ with attributes of the system $\mathscr{S}$. (Note that, as a perhaps undesirable consequence, all of the intrinsic attributes of $\mathscr{S}$ become identified with the superposition-closed subset Σ itself.) Thus, we shall refer to the complete lattice $\mathscr{L}$ as the *attribute lattice* for the system $\mathscr{S}$.

If $\Lambda, \Gamma \in \mathscr{L}$ are attributes of $\mathscr{S}$, then $\Lambda \subseteq \Gamma$ if and only if Γ is actual whenever Λ is actual. Furthermore, the attribute $\Lambda \cap \Gamma \in \mathscr{L}$ corresponds to a bona fide *conjunction* of the attributes Λ and Γ in the sense that $\Lambda \cap \Gamma$ is actual if and only if both Λ and Γ are actual. However, the least upper bound of Λ and Γ in $\mathscr{L}$ is $(\Lambda \cup \Gamma)^{sp}$, and it can be actual in states in which neither Λ nor Γ is actual. Following Aerts (1982), we say that the attributes Λ and Γ are *separated by a superselection rule* if $\Lambda \cup \Gamma$ is superposition closed, so that the least upper bound of Λ and Γ in $\mathscr{L}$ corresponds to a bona fide *disjunction* of the attributes Λ and Γ.

3.6 The Canonical Mapping

We continue our discussion of the physical system $\mathscr{S}$ subject to the logic, stochastic, and attribute postulates of Sec. 3.5.

Von Neumann (1955, p. 249) writes,

> Apart from the physical quantities . . . , there exists another category of concepts that are important objects of physics—namely the properties of the states of the system $\mathscr{S}$.

Furthermore, he goes on to identify these properties (or attributes) with the projections in the standard quantum logic L affiliated with the quantum-mechanical system $\mathscr{S}$.

In the more general situation under discussion, it is also possible to relate propositions $p \in L$ and properties (i.e., attributes) $\Lambda \in \mathscr{L}$. For $p \in L$, define

$$[p] = \{\omega \in \Omega \mid \omega(p) = 1\}.$$

We claim that $[p]$ is superposition closed. Indeed, suppose that $\alpha \in [p]^{sp}$, but that $\alpha \notin [p]$. Then $\alpha(p) \neq 1$, and so $\alpha(p') > 0$, $p' \in \operatorname{supp}(\alpha)$, and so $p' \in \operatorname{supp}(\omega)$ for some $\omega \in [p]$. But, then, $\omega(p') > 0$, and so $\omega(p) < 1$, contradicting $\omega \in [p]$. Thus $p \to [p]$ provides a mapping from experimental propositions $p \in L$ to attributes $[p] \in \mathscr{L}$. We refer to $p \to [p]$ as the *canonical mapping* (Foulis *et al.*, 1983).

An attribute of the form $[p]$ is called a *principal attribute;* the principal attributes are those that can be identified with experimental propositions as von Neumann did. It is not difficult to show that every attribute is an intersection (i.e., a conjunction) of (possibly infinitely many) principal attributes. The state space Σ is unital if and only if $[p] = 0$ implies that $p = 0$.

Evidently, $p,q \in L$ with $p \leq q$ implies that $[p] \subseteq [q]$. If the converse holds, so that $[p] \subseteq [q]$ implies that $p \leq q$, then $\mathscr{S}$ is said to have a *full* set of states. If $\mathscr{S}$ has a full set of states and every attribute is principal, then the logic L is isomorphic to the attribute lattice $\mathscr{L}$—this is precisely what happens for a standard quantum logic and it accounts for von Neumann's identification of projections and properties.

3.7 Critique of Quantum Logic

Quantum logic is a relatively young subject, it is still under vigorous development, and many consequences of the epistemological and mathematical insights that it has already provided have yet to be exploited. Quantum-logical techniques involving the tensor product have already cast some light on the well-known Einstein–Podolsky–Rosen paradox (Kläy, 1988), and it is hoped that they will also clarify some of the other classical paradoxes (Wigner's friend, Schrödinger's cat, etc.). The problem of hidden variables can be formulated, understood, and studied rigorously in terms of quantum logics (Greechie and Gudder, 1973). Quantum-logical techniques have enhanced our understanding of group-theoretic imprimitivity methods and the role of superselection rules (Piron, 1976), and ideas related to quantum logic have been used to help unravel the measurement problem (Busch *et al.*, 1991)

There is a strong possibility that unrestricted orthoalgebras are too general to serve as viable models for quantum logic. Some orthoalgebras are extremely "pathological" and thus may be suitable only for the construction of counterexamples. It seems likely that only an appropriately specialized class of orthoalgebras, e.g., unital orthoalgebras, might prove to be adequate as models for a general logic of experimental propositions.

The main drawback of quantum logic is already evident in the standard quantum logic L of a Hilbert space $\mathcal{H}$: In the passage from the wave functions ψ in $\mathcal{H}$ to the projections $P \in L$, all phase information is lost. The lost information becomes critical when sequential measurements—e.g., iterated Stern–Gehrlach spin resolutions (Wright, 1978)—are to be performed. There are at least two ways to restore the lost information, both of which are currently being studied. One can introduce complex-valued *amplitude functions* on the logic L (Gudder, 1988), or one can introduce a general mathematical infrastructure called a *manual* or *test space* (Randall and Foulis, 1973; Foulis, 1989) that can carry phase information and that gives rise to orthoalgebras as derived structures in much the same way that Hilbert spaces give rise to the standard quantum logics.

4. THE LOGICIAN'S APPROACH

We now present an approach to quantum logic more closely aligned with that of standard logical techniques. In the preceding section, we gave an axiomatic approach to orthoalgebras, the most general mathematical structures currently used as models for quantum logic. This section deals with quantum logics by using the methods of logical tradition. In so doing, we will speak of *abstract quantum logics*. As we have seen, *standard quantum logic* is identified with the complete orthomodular lattice of the projections on a separable Hilbert space of dimension at least three over the complex number field. Thus standard quantum logic is a particular kind of semantic model for a form of abstract quantum logic. Generally, a logic **L** can be determined as a triple $\langle FL, \vdash, \models \rangle$, consisting of a formal language *FL*, a *proof-theoretic consequence* relation, and a *semantic-* (or *model-theoretic*) *consequence* relation. For the sake of simplicity, we will consider only sentential languages, generated by an alphabet containing

1. a denumerably infinite sequence of *atomic sentences* (i.e., sentences whose proper parts are not sentences),
2. a finite sequence of primitive logical connectives.

The set of the *sentences* of the language *FL* is the smallest set that contains the atomic sentences and is closed under the logical connectives.

The proof-theoretic concept of consequence $\vdash$ for **L** is defined by referring to a *calculus* (a set of *axioms* and of *rules*) that, in turn, determines a notion of *proof from a set of premises to a conclusion.* A sentence β is called a *proof-theoretic consequence* of a sentence α ($\alpha \vdash \beta$) if and only if (hereafter abbreviated as iff) there is a proof where ω is the premise and β the conclusion. The semantic-consequence relation $\models$ refers to a class of possible interpretations (*models*) of the language, which render any sentence "more or less" *true* or *false.* A sentence β is called a *semantic consequence* of α ($\alpha \models \beta$) iff in any possible model of the language, β is at least as true as α.

The two consequence relations $\vdash$ and $\models$ are *reciprocally adequate* iff they are equivalent. In other words: for any sentences α, β:

$$\alpha \vdash \beta \quad \text{iff}\ \alpha \models \beta.$$

The "if arrow" represents the *soundness* property of the logic, whereas the "only if arrow" is the semantic *completeness* property.

Naturally, a logic can be characterized by different consequence relations that turn out to be equivalent. A logic **L** is called *axiomatizable* iff it admits a proof-theoretic relation, where the notion of proof is decidable. Further, **L** is called *decidable* iff the proof-theoretic–consequence relation $\vdash$ is decidable.

5. ALGEBRAIC AND POSSIBLE-WORLD SEMANTICS

In the logical tradition, logics can be generally characterized by means of two privileged kinds of semantics: an algebraic semantics, or a possible-world semantics (called also Kripkean semantics).

These semantics give different answers to the question: What does it mean to interpret a formal language? In the algebraic semantics, the basic idea is that interpreting a language essentially means associating to any sentence an abstract *truth value* or, more generally, an *abstract meaning:* an element of an algebraic structure. Hence, generally, an algebraic model for a logic **L** will have the form

$$\mathfrak{M} = \langle \mathcal{A}, v \rangle,$$

where $\mathcal{A}$ is an algebraic structure belonging to a class $\mathfrak{K}$ of structures satisfying a given set of conditions and v transforms sentences into elements of $\mathcal{A}$, preserving the logical form (in other words, logical constants are interpreted as operations of the structure). We will consider only structures where a binary relation $\leq$ (possibly a partial order) is defined. On this basis, the semantic-consequence relation is defined as follows:

DEFINITION 5.1—β is a *semantic consequence* of α ($\alpha \models \beta$) iff for any model $\mathfrak{M} = \langle \mathcal{A}, v \rangle$, $v(\alpha) \leq v(\beta)$ (in other words, the abstract meaning of α precedes the abstract meaning of β).

In the possible-world semantics, instead, one assumes that interpreting a language essentially means associating to any sentence α the set of the *possible worlds* (or *situations*) where α holds: This set, that represents the *extensional meaning* of α, is called the *proposition associated to* α (simply, the *proposition of* α). Hence, generally, a Kripkean model for a logic **L** will have the form:

$$\mathfrak{M} = \langle I, \mathbf{R}_i, \mathbf{o}_j, \Pi, v \rangle,$$

where the meanings of the symbols are as follows.

1. I is a nonempty set of possible worlds possibly correlated by relations in the sequence $\mathbf{R}_i$ and operations in the sequence $\mathbf{o}_j$. In most cases, we have only one relation R, called the *accessibility* relation.
2. Π is a set of sets of possible worlds, representing possible propositions of sentences. Any proposition and the total set of propositions Π must satisfy convenient closure conditions that depend on the particular logic.
3. v transforms sentences into propositions preserving the logical form.

A world i is said to *verify* a sentence α ($i \models \alpha$) iff $i \in v(\alpha)$.

On this basis, the Kripkean semantic-consequence relation is defined as follows:

DEFINITION 5.2—β is a *semantic consequence* of α ($\alpha \models \beta$) iff for any model $\mathfrak{M} = \langle I, \mathbf{R}_i, \mathbf{o}_j, \Pi, v \rangle$ and for any world $i \in I$,

$$\text{if } i \models \alpha \quad \text{then } i \models \beta$$

(in other words: whenever α is verified, also β is verified).

In both semantics, a sentence α is called a *logical truth* ($\models \alpha$) iff α is the consequence of any sentence β.

An interesting variant of Kripkean semantics is represented by the *many-valued possible-world semantics,* founded on a generalization of the notion of proposition. As we have seen, in the standard possible-world semantics, the proposition of a sentence α is a set of worlds: the worlds where α holds. This automatically determines the set of the worlds where α does not hold (the "meaning" of the negation of α). Intermediate truth values are not considered. In the many-valued possible-world semantics, instead, one fixes, at the very beginning, a set of truth values $V \subseteq [0,1]$ and any proposition is represented as a function X that associates to any truth value $r \in [0,1]$ a convenient set of possible worlds (the worlds where our proposition holds with truth value r). As a consequence, the total set of propositions Π turns out to behave like a family of *fuzzy* subsets of I (see THEORY AND APPLICATIONS OF FUZZY LOGIC).

Classical logic (CL) can be characterized both in the algebraic and in the Kripkean semantics. Algebraically, it is determined by the class of all algebraic structures $\langle \mathcal{A}, v\rangle$, where $\mathcal{A}$ is a Boolean algebra and v interprets the classical connectives (negation, conjunction, disjunction) as the corresponding Boolean operations (complement, meet, join). In the framework of Kripkean semantics, instead, CL is characterized by the class of all models $\langle I,R,\Pi,v\rangle$, where

1. the accessibility relation R is the identity relation (in other words, any world is accessible only to itself);
2. the set of the possible propositions Π is the set of all subsets of I;
3. v interprets the classical connectives as the corresponding set-theoretic operations.

6. ORTHODOX QUANTUM LOGIC

In the abstract quantum-logical universe, a privileged element is represented by *orthodox quantum logic* (QL), first described "as a logic" by Birkhoff and von Neumann (Birkhoff and von Neumann, 1936). QL is a singular point in the class of all logics that are weaker than classical logic. Many logical and metalogical problems concerning QL have been solved. However, some questions seem to be stubbornly resistant to being resolved.

6.1 Semantic Characterizations of QL

Similarly to classical logic, QL can be characterized both in the algebraic and in the Kripkean semantics. The language of QL contains the two primitive connectives $\neg$ (not), $\owedge$ (and). Disjunction is supposed to be metalinguistically defined via De Morgan's law:

$$\alpha \ovee \beta := \neg(\neg\alpha \owedge \neg\beta).$$

A conditional connective can be defined as the "Sasaki hook":

$$\alpha \rightarrow \beta := \neg\alpha \ovee (\alpha \owedge \beta).$$

Definition 6.1.1—An *algebraic model* of QL is a pair $\mathfrak{M} = \langle \mathcal{A}, v\rangle$, where

1. $\mathcal{A} = \langle A,\leq,{}',\mathbf{1},\mathbf{0}\rangle$ is an orthomodular lattice;
2. v (the interpretation function) interprets the connective $\neg$ as the operation $'$, the connective $\owedge$ as the lattice-meet $\wedge$:
   - **a.** $v(\alpha) \in A$ for any atomic sentence α.
   - **b.** $v(\neg\beta) = v(\beta)'$.
   - **c.** $v(\beta \owedge \gamma) = v(\beta) \wedge v(\gamma)$.

Definition 6.1.2—A sentence α is called *true* in a model $\langle \mathcal{A}, v\rangle$ iff $v(\alpha) = \mathbf{1}$. Accordingly, we will have that β is a *consequence of* α *in the algebraic semantics* of QL ($\alpha \models^{A}_{QL} \beta$) iff $v(\alpha) \leq v(\beta)$ in any model $\langle \mathcal{A}, v\rangle$ based on an orthomodular lattice $\mathcal{A}$. Further, α is a *quantum-logical truth in the algebraic semantics* ($\models^{A}_{QL} \alpha$) iff α is true in any algebraic model of QL.

As a consequence of the orthomodular property, a semantic version of a "deduction lemma" can be proved:

Lemma 6.1.1—$\alpha \models^{A}_{QL} \beta$ iff $\models^{A}_{QL} \alpha \rightarrow \beta$. In other words, $\rightarrow$ represents a "good" conditional connective: $\alpha \rightarrow \beta$ is logically true iff β is a consequence of α.

Definition 6.1.3—A *Kripkean model* of QL has the form $\mathfrak{M} = \langle I,R,\Pi,v\rangle$, where the following conditions held:

1. The accessibility relation R is reflexive and symmetric [we will also write $i \not\perp j$ for Rij; and $i \perp j$ for not Rij. Moreover, if $X \subseteq I$, we will write $i \perp X$ for $\forall j \in X\ (i \perp j)$].

   A *possible proposition* of $\mathfrak{M}$ is a *maximal* set X of worlds, which contains all and only those worlds whose accessible worlds are accessible to at least one element of X. In other words, $i \in X$ iff $\forall j \not\perp i, \exists k \not\perp j$ with $k \in X$.

   For any $X \subseteq I$, let $X^{\obot} := \{i \in I | i \perp X\}$. One can prove that $X^{\obot}$ is a possible proposition for any $X \subseteq I$; X is a possible proposition iff $X = X^{\obot\obot}$; $\emptyset$ and I are possible propositions; if X, Y are possible propositions, then $X \cap Y$ is a possible proposition.
2. Π is a set of possible propositions closed under, I, ${}^{\obot}$, $\cap$.
3. Π is orthomodular: $X \cap (X \cap (X \cap Y)^{\obot})^{\obot} \subseteq Y$, for any $X, Y \in \Pi$.

4. **a.** $v(\alpha) \in \Pi$, for any atomic sentence α;
   b. $v(\neg\beta) = v(\beta)^{\circledperp}$;
   c. $v(\beta \owedge \gamma) = v(\beta) \cap v(\gamma)$.

DEFINITION 6.1.4—A sentence α is called *true* in a model $\mathfrak{M} = \langle I,R,\Pi,v\rangle$ iff α is verified by any world $i \in I$.

Accordingly, we will have that β is a *consequence of α in the Kripkean semantics* of QL ($\alpha \models^{K}_{\mathrm{QL}} \beta$) iff for any Kripkean model $\mathfrak{M} = \langle I,R,\Pi,v\rangle$ of QL and for any world i, if $i \models \alpha$ then $i \models \beta$. Further, α is a *quantum-logical truth in the Kripkean semantics* for QL ($\models^{K}_{\mathrm{QL}} \alpha$) iff α is true in any Kripkean model of QL.

The algebraic and the Kripkean semantics for QL turn out to characterize the same logic:

THEOREM 6.1.1—$\alpha \models^{A}_{\mathrm{QL}} \beta$ *iff* $\alpha \models^{K}_{\mathrm{QL}} \beta$. This permits us to write simply $\alpha \models_{\mathrm{QL}} \beta$ instead of $\alpha \models^{A}_{\mathrm{QL}} \beta$ and $\alpha \models^{K}_{\mathrm{QL}} \beta$.

Both the algebraic and the Kripkean models of QL admit of Hilbert-space exemplifications, which are the basis for the physical interpretations. Let $\mathcal{H}$ be the separable complex Hilbert space associated to a physical system $\mathcal{S}$. An algebraic model $\langle \mathcal{A},v\rangle$ can be constructed by taking as $\mathcal{A}$ the standard quantum logic based on $\mathcal{H}$—in other words, the orthomodular lattice of the projections on $\mathcal{H}$; whereas v will follow the intended physical meaning of the atomic sentences. At the same time, a Kripkean model $\mathfrak{M} = \langle I,R,\Pi,v\rangle$ can be constructed by putting I = the set of the pure states (represented by normalized vectors ψ of $\mathcal{H}$), R = the nonorthogonality relation between pure states, and Π = the set of the possible propositions, which are uniquely determined by the closed subspaces of $\mathcal{H}$. v will follow the physical meaning of the atomic sentences. It turns out that the propositions of the model correspond to superposition-closed subsets of the pure-state space. We will call this kind of models *Hilbertian models* of QL.

A question arises: Is QL characterized by the class of all algebraic Hilbertian models? The answer is negative as proved by Greechie (1981). For instance, there is a complicated sentence of QL (corresponding to the so-called *orthoarguesian law*) that is true in all Hilbertian models, and not true in some QL models. Let us call *Hilbertian quantum logic* (HQL) the logic that is semantically characterized by the class of all Hilbertian models. Apparently, HQL is stronger than QL. Hence, abstract quantum logic turns out to be definitely more general with respect to its physical and historical origin. The axiomatizability of HQL is still an open problem.

DEFINITION 6.1.5—A sentence α is called *semantically consistent* iff for any β, $\alpha \not\models_{\mathrm{QL}} \beta \owedge \neg\beta$ (in other words, no contradiction is a semantic consequence of α).

One can show that α is semantically consistent iff there is at least one algebraic model $\langle \mathcal{A},v\rangle$ such that $v(\alpha) \neq \mathbf{0}$ iff there exists at least one Kripkean model $\mathfrak{M} = \langle I,R,\Pi,v\rangle$ and at least one world i such that $i \models \alpha$.

6.2 An Axiomatization of QL

QL is an axiomatizable logic. Many axiomatizations are known: in the Hilbert–Bernays style (Hardegree, 1979), in the natural deduction, and in the sequent style (Gibbins, 1985; Nishimura, 1980). We will present here a calculus (Goldblatt, 1974; Dalla Chiara, 1986) that represents a kind of "logical copy" of orthomodular lattices. Our calculus (that has no axioms) is determined as a set of rules. Any rule has the form

$$\frac{\alpha_1 \vdash \beta_1, \ldots, \alpha_n \vdash \beta_n}{\alpha \vdash \beta.}$$

(If β_1 is inferred from $\alpha_1, \ldots, \beta_n$ is inferred from α_n, then β can be inferred from α.) The configurations $\alpha_1 \vdash \beta_1, \ldots, \alpha_n \vdash \beta_n$ represent the *premises* of the rule, while $\alpha \vdash \beta$ is the *conclusion*. An *improper rule* is a rule whose set of premises is empty. Instead of

$$\frac{\emptyset}{\alpha \vdash \beta},$$

we will write $\alpha \vdash \beta$.

The rules of QL are as follows:

R1 $\alpha \vdash \alpha$ (identity).

R2 $\dfrac{\alpha \vdash \beta \quad \beta \vdash \gamma}{\alpha \vdash \gamma}$ (transitivity).

R3 $\alpha \owedge \beta \vdash \alpha$.

R4 $\alpha \owedge \beta \vdash \beta$.

R5 $\dfrac{\gamma \vdash \alpha \quad \gamma \vdash \beta}{\gamma \vdash \alpha \owedge \beta}$.

R6 $\alpha \vdash \neg\neg\alpha$ (weak double negation).

R7 $\neg\neg\alpha \vdash \alpha$ (strong double negation).

R8 $\dfrac{\alpha \vdash \beta}{\neg\beta \vdash \neg\alpha}$ (contraposition).

R9 $\alpha \owedge \neg(\alpha \owedge \neg(\alpha \owedge \beta)) \vdash \beta$ (orthomodularity).

Definition 6.2.1—A *proof* is a finite sequence of configurations $\alpha \vdash \beta$ where any element of the sequence is either an improper rule or the conclusion of a proper rule whose premises are previous elements of the sequence.

Definition 6.2.2—β is a *proof-theoretic consequence* of α (or *provable* from α) ($\alpha \vdash_{\text{QL}} \beta$) iff there is a proof whose last configuration is $\alpha \vdash \beta$.

Definition 6.2.3—β is a *proof-theoretic consequence* of a set of sentences T (or *provable* from T) ($T \vdash_{\text{QL}} \beta$) iff T includes a finite subset $\{\alpha_1, \ldots, \alpha_n\}$ such that $\alpha_1 \owedge \cdots \owedge \alpha_n \vdash_{\text{QL}} \beta$.

Definition 6.2.4—A set of sentences T is called *contradictory* if $T \vdash_{\text{QL}} \beta \owedge \neg\beta$ for some sentence β; *noncontradictory*, otherwise. A sentence α is contradictory if $\{\alpha\}$ is contradictory; noncontradictory, otherwise.

The proof-theoretic and the semantic-consequence relations turn out to be equivalent. Namely, a soundness and a completeness theorem can be proved:

Theorem 6.2.1—*Soundness*

If $\alpha \vdash_{\text{QL}} \beta$ *then* $\alpha \vDash_{\text{QL}} \beta$.

Theorem 6.2.2—*Completeness*

If $\alpha \vDash_{\text{QL}} \beta$ *then* $\alpha \vdash_{\text{QL}} \beta$.

As a consequence, one obtains the result that a sentence is noncontradictory iff it is semantically consistent.

A characteristic "anomaly" of QL is the violation of a metalogical condition, which is satisfied not only by CL but also by a large class of nonclassical logics. This condition is represented by the *Lindenbaum property*, according to which any noncontradictory set of sentences T can be extended to a noncontradictory and *complete* set T^* such that for any sentence α, either $\alpha \in T^*$ or $\neg\alpha \in T^*$. The set $T := \{\neg(\alpha \rightarrow (\beta \rightarrow \alpha))\}$ (which contains the negation of the *a fortiori* principle) represents an example of a noncontradictory set that cannot be extended to a noncontradictory and complete set. The set T is noncontradictory, because in some models $\langle \mathcal{A}, v\rangle$: $v(\neg(\alpha \rightarrow (\beta \rightarrow \alpha))) \neq \mathbf{0}$. For instance, take $\langle \mathcal{A}, v\rangle$ based on the orthomodular lattice of the closed subspaces of $\mathbb{R}^2$, where $v(\alpha)$ and $v(\beta)$ are two nonorthogonal unidimensional subspaces. However, one can easily check that $v(\neg(\alpha \rightarrow (\beta \rightarrow \alpha))) = \mathbf{1}$ is impossible. Hence, $\neg(\alpha \rightarrow (\beta \rightarrow \alpha))$ cannot belong to a noncontradictory and complete set T^*, which would trivially admit a model $\langle \mathcal{A}, v\rangle$ such that $v(\beta) = \mathbf{1}$, for any $\beta \in T^*$. From an intuitive point of view, the failure of the Lindenbaum property represents a very strong incompleteness result. The *tertium non datur* principle breaks down at the very deep level: There are theories that are intrinsically incomplete, even *in mente Dei*.

Among the questions that are still unsolved, let us mention at least the following:

1. Is QL decidable?
2. Does QL admit the *finite-model property?* In other words, if a sentence is not a quantum-logical truth, is there any finite model where our sentence is not verified? A positive answer to the finite-model property would automatically provide a positive answer to the decidability question, but not vice versa.
3. Is the set of all possible propositions in the Kripkean *canonical* model of QL orthomodular? (The worlds of the canonical model are all the noncontradictory and deductively closed sets of sentences T, whereas two worlds T and T' are accessible iff whenever T contains a sentence α, T' does not contain its negation $\neg\alpha$.) This problem is correlated to the critical question whether any orthomodular lattice is embeddable into a complete orthomodular lattice. Only partial answers are known.

7. ORTHOLOGIC AND UNSHARP QUANTUM LOGICS

By dropping the orthomodular condition both in the algebraic and in the Kripkean semantics, one can characterize a weaker form of quantum logic, which is usually called *orthologic* or *minimal quantum logic* (MQL). This logic turns out to be more "tractable" from a metalogical point of view: It satisfies the finite-model property; consequently, it is decidable (Goldblatt, 1974). A calculus that represents an adequate axiomatization for MQL can be, naturally, obtained by replacing the orthomodular rule **R9** of our QL calculus with the weaker Duns Scotus rule $\alpha \owedge \neg\alpha \vdash \beta$ (*ex absurdo sequitur quodlibet:* Any sentence is a consequence of a contradiction).

A less investigated form of quantum logic is represented by *paraconsistent quantum logic* (PQL) (Dalla Chiara and Giuntini, 1989), which is a weak example of an unsharp quantum logic, possibly violating the noncontradiction and the excluded-middle principles. As we will see, unsharp quantum logics represent natural abstractions from the *unsharp approaches* to quantum theory. Algebraically, PQL is characterized by the class of all models based on an involutive lattice $\langle A,\leq,',\mathbf{1},\mathbf{0}\rangle$, with smallest element $\mathbf{0}$ and largest element $\mathbf{1}$. Equivalently, in the Kripkean semantics, PQL is characterized by the class of all models $\langle I,R,\Pi,v\rangle$, where R is a symmetric, not necessarily reflexive, relation, and Π behaves like in the MQL case. Differently from QL and MQL, a world i of a PQL model may verify a contradiction. Since R is generally not reflexive, it may happen that $i \in v(\beta)$ and $i \perp v(\beta)$. Hence: $i \models \beta \owedge \neg\beta$. An adequate axiomatization for PQL can be obtained by dropping the orthomodular rule **R9** in our QL calculus. Like MQL, also PQL satisfies the finite-model property and consequently is decidable.

Interesting unsharp extensions of PQL are the *Brouwer–Zadeh logics* first investigated by Cattaneo and Nisticò (1989). A characteristic of these logics is a splitting of the connective "not" into two forms of negation: a fuzzylike negation, which gives rise to a paraconsistent behavior, and an intuitionisticlike negation. The fuzzy "not" represents a weak negation, which inverts the truth values truth and falsity, satisfies the double-negation principle, but generally violates the noncontradiction and the excluded-middle principles. The second "not" is a stronger negation, a kind of necessitation of the fuzzy "not." As a consequence, the language of the Brouwer–Zadeh logics is an extension of the QL language, with two primitive negations: $\neg$ represents the fuzzy "not," whereas $\sim$ is the intuitionistic "not." On this basis, a necessity operator can be defined in terms of the two negations:

$$L\alpha := \sim \neg\alpha.$$

In other words: "necessarily α" means the intuitionistic negation of the fuzzy negation of α. A possibility operator is then defined in terms of L and $\neg$:

$$M\alpha := \neg L \neg\alpha.$$

We will consider two forms of Brouwer–Zadeh logics: BZL (weak Brouwer–Zadeh logic) and BZL^3, which represents a form of three-valued quantum logic. Both logics admit of Hilbert-space exemplifications. Algebraically, BZL is characterized by the class of all models $\mathcal{M} = \langle \mathcal{A},v\rangle$, where $\mathcal{A} = \langle A,\leq,',^{\sim},\mathbf{1},\mathbf{0}\rangle$ is a Brouwer–Zadeh lattice (simply a BZ lattice). In other words:

1. **a.** $\langle A,\leq,',\mathbf{1},\mathbf{0}\rangle$ is an involutive lattice with smallest element $\mathbf{0}$ and largest element $\mathbf{1}$.
   b. $^{\sim}$ behaves like an intuitionistic complement:

   $$a \wedge a^{\sim} = \mathbf{0}.$$

   $$a \leq a^{\sim\sim}.$$

   $$\text{If } a \leq b, \quad \text{then } b^{\sim} \leq a^{\sim}.$$

   c. The following relation holds between the fuzzy and the intuitionistic complement:

   $$a^{\sim\prime} = a^{\sim\sim}.$$

   d. The *regularity condition* holds:

   $$a \wedge a' \leq b \vee b'.$$

2. v interprets the fuzzy negation $\neg$ as the fuzzy complement $'$, and the intuitionistic negation $\sim$ as the intuitionistic complement $^{\sim}$.

The logic BZL, which can be equivalently characterized also by a Kripkean semantics, is axiomatizable and decidable (Giuntini, 1991). The modal operators of BZL behave similarly to the corresponding operators of the famous modal system S_5. For instance, $LL\alpha$ is equivalent to $L\alpha$; and $LM\alpha$ is equivalent to $M\alpha$.

The three-valued BZL^3 can be naturally characterized by a kind of many-valued possible-world semantics. The intuitive idea can be sketched as follows: One supposes that interpreting a language means associating to any sentence two *domains of certainty:* the domain of possible worlds where the sentence holds, and the domain of possible worlds where the sentence does not hold. All the other worlds are supposed to associate an intermediate truth value (*indetermined*) to our sentence. The models of this semantics will be called *models with positive and negative domains* (shortly, *ortho-pair models*).

Briefly, an ortho-pair model has the form $\mathcal{M} = \langle I,R,\Pi,v\rangle$, where

1. I is a nonempty set of worlds and R (the accessibility relation) is reflexive and symmetric (like in the Kripkean characterization of QL).

The possible propositions (in the sense of our definition of the Kripkean model for QL) will be here called *simple propositions.* The set Σ of all simple propositions gives rise to an ortholattice; let us indicate by $\#$, $\sqcap$, $\sqcup$ the lattice operations defined on Σ.

2. A *possible proposition* of $\mathcal{M}$ is any pair $\langle X_1,X_0\rangle$, where X_1, X_0 are simple propositions such that $X_1 \subseteq X_0^{\#}$ (in other words: X_1, X_0 are orthogonal). The following operations and relations are defined on the set of all possible propositions:
   - **a.** the fuzzy complement

     $$\langle X_1,X_0\rangle' = \langle X_0,X_1\rangle;$$

   - **b.** the intuitionistic complement

     $$\langle X_1,X_0\rangle^{\sim} = \langle X_0,X_0^{\#}\rangle;$$

   - **c.** the propositional conjunction

     $$\langle X_1,X_0\rangle \wedge \langle Y_1,Y_0\rangle = \langle X_1 \sqcap Y_1, X_0 \sqcup Y_0\rangle;$$

   - **d.** the order relation

     $$\langle X_1,X_0\rangle \le \langle Y_1,Y_0\rangle \text{ iff } X_1 \subseteq Y_1 \text{ and } Y_0 \subseteq X_0.$$

3. Π is a set of possible propositions closed under $\wedge$, $'$, $^{\sim}$, and $\mathbf{0} := \langle\emptyset,I\rangle$.
4. v (the interpretation function) maps sentences into propositions in Π and interprets the connectives $\owedge$, $\sim$, $\neg$ as the corresponding operations.

The other basic semantic definitions are like in the algebraic semantics. One can show that in any ortho-pair model the set of propositions has the structure of a BZ lattice. As a consequence, the logic BZL^3 is at least as strong as BZL. In fact, one can prove that BZL^3 is properly stronger than BZL. As a counterexample, let us consider an instance of the fuzzy excluded middle and an instance of the intuitionistic excluded middle applied to the same sentence α. One can easily check:

$$\alpha \ovee \neg\alpha \models_{\mathrm{BZL}^3} \alpha \ovee \sim \alpha \quad \text{and}$$
$$\alpha \ovee \sim \alpha \models_{\mathrm{BZL}^3} \alpha \ovee \neg\alpha.$$

However, generally

$$\alpha \ovee \neg\alpha \not\models_{\mathrm{BZL}} \alpha \ovee \sim \alpha.$$

Also BZL^3 is axiomatizable (Cattaneo *et al.*, 1993) and can be characterized by means of an algebraic semantics.

8. HILBERT-SPACE MODELS OF THE BROUWER–ZADEH LOGICS

Hilbert-space models of both BZL and BZL^3 can be obtained in the framework of the *unsharp* (or *operational*) *approach* to quantum theory that was first proposed by Ludwig (1954) and developed (among others) by Kraus (1983), Davies (1976), Busch *et al.* (1991), and Cattaneo and Laudisa (1994). One of the basic ideas of this approach is a "liberalization" of the mathematical counterpart for the intuitive notion of "experimentally testable proposition." As we have seen, in orthodox Hilbert-space quantum mechanics, experimental propositions are mathematically represented as projections P on the Hilbert space $\mathcal{H}$ corresponding to the physical system $\mathcal{S}$ under investigation. If P is a projection representing a proposition and W is a density operator representing a state of $\mathcal{S}$, the number $\mathrm{tr}(WP)$ represents the probability value that the system $\mathcal{S}$ in the state W

verifies P (*Born probability*). However, projections are not the only operators for which a Born probability can be defined. Let us consider the class $\mathcal{E}(\mathcal{H})$ of all linear bounded operators D such that for any density operator W,

$$\mathrm{tr}(WD) \in [0,1].$$

It turns out that $\mathcal{E}(\mathcal{H})$ properly includes the set $L(\mathcal{H})$ of all projections on $\mathcal{H}$. In a sense, the elements of $\mathcal{E}(\mathcal{H})$ represent a "maximal" possible notion of experimental proposition, in agreement with the probabilistic rules of quantum theory. In the framework of the unsharp approach, the elements of $\mathcal{E}(\mathcal{H})$ have been called *effects*. An important difference between projections and proper effects is the following: Projections can be associated to *sharp* propositions having the form "the value for the observable A lies in the exact Borel set F," whereas effects may represent also *fuzzy* propositions like "the value of the observable A lies in the fuzzy Borel set F." As a consequence, there are effects D that are different from the null projection $\mathbb{O}$ and that are verified with certainty by no state [for any W, $\mathrm{tr}(WD) \neq 1$]. A limit case is represented by the *semitransparent effect* $\frac{1}{2}\mathbb{1}$ (where $\mathbb{1}$ is the identity operator), to which any state W assigns probability value $\frac{1}{2}$.

The class of all effects of $\mathcal{H}$ gives rise to a structure $\langle\mathcal{E}(\mathcal{H}),\leq,{}',{}^{\sim},\mathbf{1},\mathbf{0}\rangle$ which is a BZ poset (not a BZ lattice!). In other words, $\leq$ is a partial order with largest element $\mathbf{1}$ and smallest element $\mathbf{0}$, while the fuzzy and the intuitionistic complement ($'$ and $^{\sim}$) behave like in the BZ lattices. The relation and the operations of the effect-structure are defined as follows:

1. $D_1 \leq D_2$ iff for any density operator W, $\mathrm{tr}(WD_1) \leq \mathrm{tr}(WD_2)$.
2. $\mathbf{1} = \mathbb{1}$.
3. $D' = \mathbb{1} - D$.
4. $D^{\sim}$ is the projection $P_{\mathrm{Ker}(D)}$ into the subspace $\mathrm{Ker}(D)$, consisting of all vectors that are transformed by the operator D into the null vector.
5. $\mathbf{0} = \mathbf{1}'$.

In the particular case where D is a projection, it turns out that $D' = D^{\sim}$. In other words, the fuzzy and the intuitionistic complement coincide for sharp propositions. One can show that any BZ poset can be embedded into a complete BZ lattice [for the *MacNeille completion* (Birkhoff, 1967) of a BZ poset is a complete BZ lattice (Giuntini, 1991)]. As a consequence, the MacNeille completions of the effect-BZ posets represent natural Hilbert-space models for the logic BZL.

As to BZL^3, Hilbert-space models $\langle I,R,\Pi,v\rangle$ in the ortho-pair semantics can be constructed as follows:

1. I and R are defined like in the Kripkean Hilbertian models of QL. The simple propositions turn out to be in one-to-one correspondence to the set of the projections of $\mathcal{H}$.
2. Π is the set of all possible propositions. Any effect D can be transformed into a proposition $f(D) = \langle X_1^D, X_0^D\rangle$, where

   $$X_1^D := \{\psi \in I \mid \mathrm{tr}(P_\psi D) = 1\} \quad \text{and} \quad X_0^D := \{\psi \in I \mid \mathrm{tr}(P_\psi D) = 0\}$$

   (P_ψ is the projection onto the unidimensional subspace spanned by the vector ψ). In other words, X_1^D and X_0^D represent respectively the positive and the negative domain of D (in a sense, the extensional meaning of D in the model). The map f turns out to preserve the order relation and the two complements.
3. The interpretation function v follows the intuitive physical meaning of the atomic sentences.

9. PARTIAL QUANTUM LOGICS

So far we have considered only examples of abstract quantum logics, where conjunctions and disjunctions are supposed to be always defined. However, as we have seen, the experimental and the probabilistic meaning of conjunctions of incompatible propositions in quantum theory has been often put in question. How do we construct logics where we admit that conjunctions and disjunctions are possibly meaningless? For instance, how do we give a natural semantic characterization for a logic corresponding to the class of all orthoalgebras or to the class of all orthomodular posets? Let us call these logics respectively *weak partial quantum logic* (WPaQL) and *strong partial quantum logic*

(SPaQL). Are WPaQL and SPaQL axiomatizable?

9.1 Algebraic Semantics for WPaQL

The language of WPaQL contains two primitive connectives: the negation $\neg$, and the exclusive disjunction $\boxplus$ (*aut*). A conjunction is metalinguistically defined, via De Morgan's law:

$$\alpha \square \beta := \neg(\neg\alpha \boxplus \neg\beta).$$

The intuitive idea underlying our semantics for WPaQL is the following: Disjunctions and conjunctions are considered "legitimate" from a mere linguistic point of view. However, semantically, a disjunction $\alpha \boxplus \beta$ will have the *intended meaning* only in the "well-behaved cases" (where the values of α and β are orthogonal in the corresponding orthoalgebra). Otherwise, $\alpha \boxplus \beta$ will have any meaning whatsoever (generally not connected with the meanings of α and β). A similar semantic "trick" is used in some standard treatments of the description operator ι ("the unique individual that satisfies a given property") in classical model theory.

Definition 9.1.1—An *algebraic model* of WPaQL is a pair $\mathcal{M} = \langle \mathcal{A}, v \rangle$, where

1. $\mathcal{A} = \langle A, \oplus, \mathbf{1}, \mathbf{0} \rangle$ is an orthoalgebra.
2. v (the interpretation function) satisfies the following conditions:
   $v(\alpha) \in A$, for any atomic sentence α;
   $v(\neg\beta) = v(\beta)'$, where $'$ is the orthocomplement operation that is defined in $\mathcal{A}$;
   $$v(\beta \boxplus \gamma) = \begin{cases} v(\beta) \oplus v(\gamma), & \text{if } v(\beta) \oplus v(\gamma) \text{ is defined in } \mathcal{A}, \\ \text{any element} & \text{otherwise} \end{cases}.$$

Accordingly, we will have that

$$\alpha \models_{\mathrm{WPaQL}} \beta \quad \text{iff in any WPaQL model } \mathcal{M} = \langle \mathcal{A}, v \rangle,\ v(\alpha) \leq v(\beta),$$

where $\leq$ is the partial order relation defined in $\mathcal{A}$.

9.2 An Axiomatization of Partial Quantum Logics

The logic WPaQL is axiomatizable. We present here a calculus that is obtained as a natural transformation of our QL calculus.

The rules of WPaQL are as follows:

R1 $\alpha \vdash \alpha$ (identity).

R2 $\dfrac{\alpha \vdash \beta \quad \beta \vdash \gamma}{\alpha \vdash \gamma}$ (transitivity).

R3 $\alpha \vdash \neg\neg\alpha$ (weak double negation).

R4 $\neg\neg\alpha \vdash \alpha$ (strong double negation).

R5 $\dfrac{\alpha \vdash \beta}{\neg\beta \vdash \neg\alpha}$ (contraposition).

R6 $\beta \vdash \alpha \boxplus \neg\alpha$ (excluded middle).

R7 $\dfrac{\alpha \vdash \neg\beta \quad \alpha \boxplus \neg\alpha \vdash \alpha \boxplus \beta}{\neg\alpha \vdash \beta}$ (unicity of negation).

R8 $\dfrac{\alpha \vdash \neg\beta \quad \alpha \vdash \alpha_1 \quad \alpha_1 \vdash \alpha \quad \beta \vdash \beta_1 \quad \beta_1 \vdash \beta}{\alpha \boxplus \beta \vdash \alpha_1 \boxplus \beta_1}$ (weak substitutivity).

R9 $\dfrac{\alpha \vdash \neg\beta}{\alpha \boxplus \beta \vdash \beta \boxplus \alpha}$ (weak commutativity).

R10 $\dfrac{\beta \vdash \neg\gamma \quad \alpha \vdash \neg(\beta \boxplus \gamma)}{\alpha \vdash \neg\beta}$.

R11 $\dfrac{\beta \vdash \neg\gamma \quad \alpha \vdash \neg(\beta \boxplus \gamma)}{\alpha \boxplus \beta \vdash \neg\gamma}$.

R12 $\dfrac{\beta \vdash \neg\gamma \quad \alpha \vdash \neg(\beta \boxplus \gamma)}{\alpha \boxplus (\beta \boxplus \gamma) \vdash (\alpha \boxplus \beta) \boxplus \gamma}$.

R13 $\dfrac{\beta \vdash \neg\gamma \quad \alpha \vdash \neg(\beta \boxplus \gamma)}{(\alpha \boxplus \beta) \boxplus \gamma) \vdash \alpha \boxplus (\beta \boxplus \gamma)}$.

(**R10–R13** require a weak associativity.)

The other basic proof-theoretic definitions are given like in the QL case. Some derivable rules of the calculus are the following:

D1 $\dfrac{\alpha \vdash \beta}{\beta \vdash \alpha \boxplus \neg(\alpha \boxplus \neg\beta)}$.

D2 $\dfrac{\alpha \vdash \beta}{\alpha \boxplus \neg(\alpha \boxplus \neg\beta) \vdash \beta}$.

D3
$$\frac{\alpha \vdash \neg\gamma \quad \beta \vdash \neg\gamma \, \alpha \boxplus \gamma \vdash \beta \boxplus \gamma \, \beta \boxplus \gamma \vdash \alpha \boxplus \gamma}{\alpha \vdash \beta}.$$

D4 $\dfrac{\alpha \vdash \neg\beta \quad \alpha \vdash \gamma \quad \beta \vdash \gamma \quad \gamma \vdash \alpha \boxplus \beta}{\alpha \boxplus \beta \vdash \gamma}$.

The proof-theoretic and the semantic-consequence relations for the logic WPaQL are reciprocally adequate. Namely, a soundness and a completeness theorem can be proved.

Theorem 9.2.1—*Soundness*

If $\alpha \vdash_{\mathrm{WPaQL}} \beta$ *then* $\alpha \models_{\mathrm{WPaQL}} \beta$.

Theorem 9.2.2—*Completeness*

If $\alpha \models_{\mathrm{WPaQL}} \beta$ *then* $\alpha \vdash_{\mathrm{WPaQL}} \beta$.

As to strong partial quantum logic (SPaQL), an axiomatization can be obtained by adding to our WPaQL calculus the following rule:

R14 $$\frac{\alpha \vdash \neg\beta \quad \alpha \vdash \gamma \quad \beta \vdash \gamma}{\alpha \boxplus \beta \vdash \gamma}.$$

Semantically, the models of SPaQL will be based on orthoalgebras $\mathcal{A} = \langle A,\oplus,\mathbf{1},\mathbf{0}\rangle$, satisfying the following condition: If defined, $a \oplus b$ is the sup of a and b. As we have seen in Sec. 3.1, this condition is necessary and sufficient in order to make the orthoposet induced by the orthoalgebra $\mathcal{A}$ an orthomodular poset. The soundness and the completeness theorems for SPaQL (with respect to this semantics) can be proved similarly to the case of WPaQL.

10. CRITIQUE OF ABSTRACT QUANTUM LOGICS

Do abstract quantum logics represent "real" logics or should they rather be regarded as mere extrapolations from particular algebraic structures that arise in the mathematical formalism of quantum mechanics? Different answers to this question have been given in the history of the logicoalgebraic approach to quantum theory. According to our analysis, the logical *status* of abstract quantum logics can be hardly put in question. These logics turn out to satisfy all the canonical conditions that the present community of logicians require in order to call a given abstract object a *logic:* syntactical and semantical descriptions, proofs of soundness and completeness theorems, and so on.

Has the quantum-logical research definitely shown that "logic is empirical"? At the very beginning of the history of quantum logic, the thesis according to which the choice of the "right" logic to be used in a given theoretic situation may depend also on experimental data appeared a kind of extremistic view, in contrast with the traditional description of logic as "an *a priori* and analytical science." These days, an empirical position in logic is no more regarded as a "daring heresy." At the same time, we are facing a new difficulty: As we have seen, quantum logic is not unique. Besides orthodox quantum logic, different forms of partial and unsharp quantum logics have been developed. In this situation, one can wonder whether it is still reasonable to look for the most adequate abstract logic that should faithfully represent the structures arising in the quantum world.

A question that has been often discussed concerns the compatibility between quantum logic and the mathematical formalism of quantum theory, based on classical logic. Is the quantum physicist bound to a kind of "logical schizophrenia"? At first sight, the compresence of different logics in one and the same theory may give a sense of uneasiness. However, the splitting of the basic logical operations (negation, conjunction, disjunction, . . .) into different connectives with different meanings and uses is now a well-accepted logical phenomenon that admits consistent descriptions. As we have seen, classical and quantum logic turn out to apply to different sublanguages of quantum theory that must be sharply distinguished.

GLOSSARY

Abstract Quantum Logic: A logic $\langle FL,\vdash,\models\rangle$, where the proof-theoretic and the semantic-consequence relations violate some characteristic classical principles like the distributivity of conjunction and disjunction.

Adjoint: If A is a bounded operator on a Hilbert space, then the adjoint of A is the unique bounded operator A^* that satisfies $\langle A\psi|\phi\rangle = \langle\psi|A^*\phi\rangle$ for all vectors ψ, ϕ in the space.

Algebraic Model of a Language: A pair $\langle\mathcal{A},v\rangle$ consisting of an algebraic structure $\mathcal{A}$ and of an interpretation function v that transforms the sentences of the language into elements of $\mathcal{A}$, preserving the logical form.

Algebraic Semantics: The basic idea is that interpreting a formal language means associating to any sentence an element of an algebraic structure.

Attribute: One of a class of properties affiliated with a physical system. At any given moment some of the attributes of the system may be actual, while others are only potential.

Axiomatizable Logic: A logic is axiomatizable when its concept of proof is decidable.

Boolean Algebra: An orthocomplemented lattice L that satisfies the distributive law $p \wedge (q \vee r) = (p \wedge q) \vee (p \wedge r)$ for all $p, q, r \in L$.

Borel Set: A (real) Borel set is a set that belongs to the smallest collection of subsets of the real numbers $\mathbb{R}$ that contains all open intervals and is closed under the formation of complements and countable unions.

Brouwer–Zadeh Lattice (or Poset): A lattice (or poset) L with smallest element $\mathbf{1}$ and largest element $\mathbf{0}$, equipped with a regular involution $'$ (a fuzzylike complement), and an intuitionisticlike complement $^{\sim}$, subject to the following conditions for all $p,q \in L$: (i) $p \leq q \Rightarrow q^{\sim} \leq p^{\sim}$, (ii) $p \wedge p^{\sim} = \mathbf{0}$, (iii) $p \leq p^{\sim\sim}$, (iv) $p^{\sim\prime} = p^{\sim\sim}$.

Brouwer–Zadeh Logic: A logic that is characterized by the class of all models based on Brouwer–Zadeh lattices.

Compatible: A set C of elements in an orthoalgebra L is a compatible set if there is a Boolean subalgebra B of L such that $C \subseteq$ B.

Complete Lattice: A lattice in which every subset has a least upper bound, or join, and a greatest lower bound, or meet.

Completeness: A logic $\langle FL, \vdash, \models \rangle$ is (semantically) complete when all the semantic consequences are proof-theoretic consequences (if $\alpha \models \beta$ then $\alpha \vdash \beta$).

Conjunction (in an Orthoalgebra): If L is an orthoalgebra and $p, q \in L$, then p and q have a conjunction $p \,\&\, q \in L$ if there is a Boolean subalgebra B of L with $p, q \in B$, and the meet $p \wedge_B q$ of p and q in B is independent of the choice of B, in which case $p \,\&\, q = p \wedge_B q$.

Decidable Logic: A logic is decidable when its proof-theoretic consequence relation is decidable.

Density Operator: A self-adjoint, nonnegative, trace-class operator W on a Hilbert space, such that $\mathrm{tr}(W) = 1$.

Direct Sum (or Cartesian Product) of Hilbert Spaces: The direct sum of the Hilbert spaces $\mathcal{H}$ and $\mathcal{K}$ is the Hilbert space $\mathcal{H} \oplus \mathcal{K}$ consisting of all ordered pairs (x,y) with $x \in \mathcal{H}$, $y \in \mathcal{K}$, and with coordinatewise vector operations. The inner product is defined by $\langle (x_1,y_1)|(x_2,y_2)\rangle = \langle x_1|x_2\rangle + \langle y_1|y_2\rangle$.

Disjunction (in an Orthoalgebra): If L is an orthoalgebra and $p, q \in L$, then p and q have a disjunction $p + q$ if there is a Boolean subalgebra B of L with $p, q \in B$ and the join $p \vee_B q$ of p and q in B is independent of the choice of B, in which case $p + q = p \vee_B q$.

Dynamics: The evolution in time of the state of a physical system.

Effect: A linear bounded operator A of a Hilbert space such that for any density operator W, $\mathrm{tr}(WA) \in [0,1]$. In the unsharp approach to quantum mechanics, effects represent possible experimental propositions.

Experimental Proposition: A proposition whose truth value can be determined by conducting an experiment.

Greatest vs Maximal: If L is a partially ordered set (poset) and $X \subseteq L$, then an element $b \in X$ is a greatest element of X if $x \leq b$ for all $x \in X$. An element $b \in X$ is a maximal element of X if there exists no element $x \in X$ with $b < x$.

Involution: A mapping $p \to p'$ on a poset L satisfying the following conditions for all $p, q \in L$: (i) $p \leq q \Rightarrow q' \leq p'$, (ii) $p'' = p$.

Involutive Lattice (or Poset): A lattice (or poset) with smallest element $\mathbf{0}$ and largest element $\mathbf{1}$, equipped with an involution.

Join: If L is a partially ordered set (poset) and $p, q \in L$, then the join (or least upper bound) of p and q in L, denoted by $p \vee q$ if it exists, is the unique element of L satisfying the following conditions: (i) $p, q \leq p \vee q$ and (ii) $r \in L$ with $p, q \leq r \Rightarrow p \vee q \leq r$.

Kripkean Model of a Language: A system $\langle I, R_1, \ldots, R_m, o_1, \ldots, o_n, \Pi, v\rangle$ consisting of a set I of possible worlds, a (possibly empty) sequence of world relations $R_1, \ldots, R_m$ and of world operations $o_1, \ldots, o_n$, a family Π of subsets of I (called the propositions), and an interpretation function v that transforms the sentences of the language into propositions, preserving the logical form.

Kripkean Semantics: The basic idea is that interpreting a formal language means associating to any sentence the set of the possible worlds where the sentence holds. This set is called also the proposition associated to the sentence.

Lattice: A partially ordered set (poset) in which every pair of elements p, q has a least upper bound, or join, $p \vee q$ and a greatest lower bound, or meet, $p \wedge q$.

Least vs Minimal: If L is a partially ordered set (poset) and $X \subseteq L$, then an element $a \in X$ is a least element of X if $a \leq x$ for all

$x \in X$. An element $a \in X$ is a minimal element of X if there exists no element $x \in X$ with $x < a$.

Lindenbaum Property: A logic satisfies the Lindenbaum property when any noncontradictory set of sentences T can be extended to a noncontradictory and *complete* set T^* (such that T^* contains, for any sentence of the language, either the sentence or its negation). Abstract quantum logics generally violate the Lindenbaum property.

Logic: According to the tradition of logical methods, a logic can be described as a system $\langle FL, \vdash, \models \rangle$ consisting of a *formal language*, a *proof-theoretic–consequence* relation $\vdash$ (based on a notion of *proof*), and a *semantic-consequence* relation $\models$ (based on a notion of *model* and of *truth*).

MacNeille Completion of a Brouwer–Zadeh Lattice: Let L be a Brouwer–Zadeh lattice (or poset). For $X \subseteq L$, let $X' = \{p \in L | \forall q \in X, p \leq q'\}$, $X^{\sim} = \{p \in L | \forall q \in X, p \leq q^{\sim}\}$, and $P(L) = \{X \subseteq L | X = X''\}$. The structure $\mathcal{P}(L) = \langle P(L), \subseteq, {}', {}^{\sim}, \{\ \}'', L \rangle$ is called the MacNeille completion of L. $\mathcal{P}(L)$ is a complete Brouwer–Zadeh lattice and L is embeddable into $\mathcal{P}(L)$ via the mapping $p \rightarrow \langle p]$, where $\langle p] = \{q \in L | q \leq p\}$.

Meet: If L is a partially ordered set (poset) and p, $q \in L$, then the meet (or greatest lower bound) of p and q, denoted by $p \wedge q$ if it exists, is the unique element of L satisfying the following conditions: (i) $p \wedge q \leq p, q$ and (ii) $r \in L$ with $r \leq p, q \Rightarrow r \leq p \wedge q$.

Minimal Quantum Logic: A logic that is semantically characterized by the class of all models based on orthocomplemented lattices.

Modular Lattice: A lattice L satisfying the modular law: $p \leq r \Rightarrow p \vee (q \wedge r) = (p \vee q) \wedge r$ for all $p,q,r \in L$.

Nonnegative Operator: A self-adjoint operator A on a Hilbert space $\mathcal{H}$ such that $\langle A\psi | \psi \rangle \geq 0$ for all vectors $\psi \in \mathcal{H}$.

Observable or Dynamical Variable: A numerical variable associated with a physical system the value of which can be determined by conducting a test, a measurement, or an experiment on the system.

Orthoalgebra: A mathematical system consisting of a set L with two special elements $\mathbf{0}$, $\mathbf{1}$ and equipped with a relation $\perp$ such that, for each pair $p,q \in L$ with $p \perp q$, an orthogonal sum $p \oplus q \in L$ is defined subject to the following conditions for all $p,q,r \in L$: (i) $p \perp q \Rightarrow q \perp p$ and $p \oplus q = q \oplus p$, (ii) $p \perp q$ and $(p \oplus q) \perp r \Rightarrow q \perp r$, $p \perp (q \oplus r)$ and $p \oplus (q \oplus r) = (p \oplus q) \oplus r$, (iii) $p \in L \Rightarrow$ there is a unique $p' \in L$ such that $p \perp p'$ and $p \oplus p' = \mathbf{1}$, and (iv) $p \perp p \Rightarrow p = \mathbf{0}$.

Orthocomplementation: A mapping $p \rightarrow p'$ on a poset L with smallest element $\mathbf{0}$ and largest element $\mathbf{1}$ satisfying the following conditions for all p, $q \in L$: (i) $p \vee p' = \mathbf{1}$, (ii) $p \wedge p' = \mathbf{0}$, (iii) $p \leq q \Rightarrow q' \leq p'$, and (iv) $p'' = p$.

Orthocomplemented Lattice (or Poset): A lattice (or poset) equipped with an orthocomplementation $p \rightarrow p'$.

Orthodox Quantum Logic: A logic that is semantically characterized by the class of all algebraic models based on orthomodular lattices. Standard quantum logic is a particular model of orthodox quantum logic.

Orthogonal: If L is an orthocomplemented poset and p, $q \in L$, then p is orthogonal to q, in symbols $p \perp q$, if $p \leq q'$.

Orthomodular Lattice: An orthocomplemented lattice L satisfying the orthomodular law: For all $p,q \in L$, $p \leq q \Rightarrow q = p \vee (q \wedge p')$.

Orthomodular Poset: An orthoalgebra L such that, for all $p,q \in L$, $p \perp q \Rightarrow p \oplus q = p \vee q$.

Paraconsistent Quantum Logic: A logic that is semantically characterized by the class of all models based on involutive lattices with smallest element $\mathbf{0}$ and largest element $\mathbf{1}$.

Partially Ordered Set (or Poset): A set L equipped with a relation $\leq$ satisfying the following conditions for all $p,q,r \in L$: (i) $p \leq p$, (ii) $p \leq q$ and $q \leq p \Rightarrow p = q$, (iii) $p \leq q$ and $q \leq r \Rightarrow p \leq r$.

Probability Measure: A function $\omega : L \rightarrow [0,1] \subseteq \mathbb{R}$ on an orthoalgebra L such that $\omega(\mathbf{0}) = 0$, $\omega(\mathbf{1}) = 1$, and, for all $p,q \in L$ with $p \perp q$, $\omega(p \oplus q) = \omega(p) + \omega(q)$.

Projection: An operator P on a Hilbert space that is self-adjoint ($P = P^*$) and idempotent ($P = P^2$).

Pure State: A state ψ is pure if the set $\{\psi\}$ consisting only of that state is superposition closed. In Hilbert-space quantum mechanics, the pure states are precisely the vector states.

Quantum Logic: The study of the formal structure of experimental propositions affiliated with a quantum physical system, or any mathematical model (e.g., an orthoalgebra) representing such a structure.

Regular Involution: An involution $'$ on a poset L that satisfies the regularity condition: For all $p,q \in L$: $p \leq p'$ and $q \leq q' \Rightarrow p \leq q'$. If L is a lattice, then an involution is regular iff it satisfies the Kleene condition: For all $p,q \in L$: $p \wedge p' \leq q \vee q'$.

Soundness: A logic $\langle FL, \vdash, \models \rangle$ is sound when all the proof-theoretic consequences are semantic consequences (if $\alpha \vdash \beta$ then $\alpha \models \beta$).

Spectral Measure: A mapping from real Borel sets into projection operators on a Hilbert space that maps the empty set into $\mathbb{0}$, maps $\mathbb{R}$ into $\mathbb{1}$, and maps the union of a disjoint sequence of real Borel sets into the least upper bound (join) of the corresponding projection operators.

Spectral Theorem: The theorem establishing a one-to-one correspondence between (not necessarily bounded) self-adjoint operators A on a Hilbert space $\mathcal{H}$ and spectral measures $E \rightarrow P_E$ on $\mathcal{H}$ such that, if $P_\lambda = P_{(-\infty,\lambda]}$ for all $\lambda \in \mathbb{R}$, then $A = \int_{-\infty}^{\infty} \lambda dP_\lambda$.

Spectrum: If A is a (not necessarily bounded) self-adjoint operator on a Hilbert space and $E \rightarrow P_E$ is the corresponding spectral measure, then a real number λ belongs to the spectrum of A if $P_{(\lambda-\epsilon,\lambda+\epsilon)} \neq \mathbb{0}$ for all $\epsilon > 0$.

Standard Quantum Logic: The complete orthomodular lattice L of all projection operators on a Hilbert space. For $P, Q \in L$, $P \leq Q$ is defined to mean that $P = PQ$ and the orthocomplement of P is defined by $P' = \mathbb{1} - P$.

State: The state of a physical system encodes all information concerning the results of conducting tests or measuring observables on the system. It is usually assumed that, corresponding to each state ψ of the system, there is a probability measure ω_ψ on the logic L of the system such that $\omega_\psi(p)$ is the probability that the experimental proposition $p \in L$ is true when the system is in the state ψ.

State Space: The set of all possible states of the physical system.

Strong Partial Quantum Logic: A logic that is semantically characterized by the class of all models based on orthomodular posets.

Superselection Rule: A rule that determines the possible states of a physical system. The usual quantum-mechanical superselection rules state that only vector states that commute with a certain set of pairwise orthogonal projections (i.e., projections onto superposition sectors) represent possible states of the systems.

Support: If W is a density operator on a Hilbert space $\mathcal{H}$, then the support of W is defined to be the set $\text{supp}(W)$ of all projection operators P on $\mathcal{H}$ such that $\text{tr}(WP) \neq 0$. More generally, if ω is a probability measure on an orthoalgebra L, then $\text{supp}(\omega) = \{p \in L | \omega(p) \neq 0\}$.

Superposition Closed: A set of states is superposition closed if it contains all of its own superpositions.

Superposition in Hilbert Space: For a Hilbert space $\mathcal{H}$, if (W_α) is a family of vector states determined by the corresponding family (ψ_α) of normalized vectors in $\mathcal{H}$, and if ψ is a normalized linear combination of the vectors in this family, then the vector state W determined by ψ is a (coherent) superposition of the family (W_α). If (W_α) is an arbitrary family of density operators on $\mathcal{H}$ and (t_α) is a corresponding family of nonnegative real numbers such that $\Sigma_\alpha t_\alpha = 1$, then $W = \Sigma_\alpha t_\alpha W_\alpha$ is an (incoherent) superposition (or mixtures) of the family (W_α).

Superposition in an Orthoalgebra: A probability measure ω on an orthoalgebra L is a superposition of a family (ω_α) of probability measures on L if $\text{supp}(\omega) \subseteq \cup_\alpha \text{supp}(\omega_\alpha)$.

Tensor Product of Hilbert Spaces: If $\mathcal{H}$ and $\mathcal{K}$ are Hilbert spaces, then the tensor product $\mathcal{H} \otimes \mathcal{K}$ is a Hilbert space together with a mapping $(x,y) \rightarrow x \otimes y \in \mathcal{H} \otimes \mathcal{K}$ for $x \in \mathcal{H}$, $y \in \mathcal{K}$ that is separately linear in each argument and has the property that if (ψ_i) is an orthonormal basis for $\mathcal{H}$ and (ϕ_j) is an orthonormal basis for $\mathcal{K}$, then $(\psi_i \otimes \phi_j)$ is an orthonormal basis for $\mathcal{H} \otimes \mathcal{K}$.

Trace: The trace of an operator A on a Hilbert space $\mathcal{H}$ is defined by $\text{tr}(A) = \Sigma_{\psi \in B} \langle A\psi | \psi \rangle$, where B is an orthonormal basis for $\mathcal{H}$, provided that the series converges.

Trace Class: An operator A on a Hilbert space $\mathcal{H}$ belongs to the trace class if $\text{tr}(A) = \Sigma_{\psi \in B} \langle A\psi | \psi \rangle$ converges absolutely, where B is an orthonormal basis for $\mathcal{H}$.

Unsharp Quantum Logics: Examples of paraconsistent logics where the noncontradiction principle is generally violated.

Vector State: A probability measure on the standard quantum logic L of a Hilbert space $\mathcal{H}$ determined by a normalized vector

$\psi \in \mathcal{H}$ and assigning to each projection operator $P \in L$ the probability $\langle P\psi|\psi\rangle$.

Weak Partial Quantum Logic: A logic that is semantically characterized by the class of all models based on orthoalgebras.

Works Cited

Aerts, D. (1982), "Description of Many Physical Entities without the Paradoxes Encountered in Quantum mechanics," *Found. Phys.* **12,** 1131–1170.

Beltrametti, E., Cassinelli, G. (1981), "The Logic of Quantum Mechanics," in: Gian-Carlo Rota (Ed.), *Encyclopedia of Mathematics and its Applications,* vol. 15, Reading, MA: Addison-Wesley.

Bennett, M., Foulis, D. (1990), "Superposition in Quantum and Classical Mechanics," *Found. Phys.* **20,** 733–744.

Birkhoff, G. (1967), *Lattice Theory,* 3rd ed., Providence, RI: American Mathematical Society Colloquium Publications, XXV.

Birkhoff, G. (1961), "Lattices in Applied Mathematics," in: R. P. Dilworth (Ed.), *Lattice Theory,* Proceedings of the Second Symposium in Pure Mathematics of the American Mathematical Society, Providence, RI: American Mathematical Society.

Birkhoff, G., von Neumann, J. (1936), "The Logic of Quantum Mechanics," *Ann. Math.* **37,** 823–843.

Busch, P., Lahti, P., Mittelstaedt, P. (1991), *The Quantum Theory of Measurement,* Lecture Notes in Physics, New Series m2, Berlin/Heidelberg/New York: Springer Verlag.

Cattaneo, G., Dalla Chiara, M. L., Giuntini, R. (1993), "Fuzzy-Intuitionistic Quantum Logics," *Studia Logica* **52,** 1–24.

Cattaneo, G., Laudisa, F. (1994), "Axiomatic Unsharp Quantum Mechanics (from Mackey to Ludwig and Piron)," *Found. Phys.* **24,** 631–684.

Cattaneo, G., Nisticò, G. (1989), "Brouwer-Zadeh Posets and Three-Valued Łukasiewicz Posets," *Fuzzy Sets Systems* **33,** 165–190.

Dalla Chiara, M. L. (1986), "Quantum Logic," in: D. Gabbau, F. Guenthner (Eds.), *Handbook of Philosophical Logic,* III, Dordrecht, Netherlands: Reidel.

Dalla Chiara, M. L., Giuntini, R. (1989), "Paraconsistent Quantum Logics," *Found. Phys.* **19,** 891–904.

Davies, E. B. (1976), *Quantum Theory of Open Systems,* New York: Academic.

D'Espagnat, B. (1971), *Conceptual Foundations of Quantum Mechanics,* Menlo Park, NJ: Benjamin.

Foulis, D. (1962), "A note on Orthomodular Lattices," *Port. Math.* **21,** 65–72.

Foulis, D. (1989), "Coupled Physical Systems," *Found. Phys.* **7,** 905–922.

Foulis, D., Randall, C. (1981), "Empirical Logic and Tensor Products," in: *Interpretations and Foundations of Quantum Theory,* Grundlagen der Exakt Naturwissenschaften, Vol. 5, Mannheim/Wien: Bibliographisches Institut.

Foulis, D., Greechie, R., Rüttimann, G. (1992), "Filters and Supports in Orthoalgebras," *Int. J. Theor. Phys.* **31,** 789–807.

Foulis, D., Randall, C., Piron, C. (1983), "Realism, Operationalism, and Quantum Mechanics," *Found. Phys.* **8,** 813–842.

Gibbins, P. (1985), "A User-Friendly Quantum Logic," *Log. Anal.* **112,** 353–362.

Giuntini, R. (1991), "A Semantical Investigation on Brouwer-Zadeh Logic," *J. Philos. Log.* **20,** 411–433.

Gleason, A. (1957), "Measures of Closed Subspaces of a Hilbert Space," *J. Math. Mech.* **6,** 885–893.

Goldblatt, R. (1974), "Semantic Analysis of Orthologic," *J. Philos. Log.* **3,** 19–35.

Greechie, R. (1981), "A Non-standard Quantum Logic with a Strong Set of States," in: E. Beltrametti, B. van Fraasen (Eds.), *Current Issues in Quantum Logic,* Ettore Majorana International Science Series, vol. 8, New York: Plenum.

Greechie, R., Gudder, S. (1973), "Quantum Logics," in: C. Hooker (Ed.), *Contemporary Research in the Foundations and Philosophical of Quantum Theory,* Dordrecht, Netherlands: Reidel.

Gudder, S. (1988), *Quantum Probability,* San Diego: Academic.

Hardegree, G. (1979), "The Conditional in Abstract and Concrete Quantum Logic," in: C. Hooker (Ed.), *The Logic-Algebraic Approach to Quantum Mechanics, II,* The University of Western Ontario Series in Philosophy of Science, vol. 5, Dordrecht, Netherlands: Reidel.

Hardegree, G., Frazer, P. (1981), "Charting the Labyrinth of Quantum Logics," in: E. Beltrametti, B. van Fraassen (Eds.), *Current Issues in Quantum Logic,* Ettore Majorana International Science Series, vol. 8, New York: Plenum.

Husimi, K. (1937), "Studies on the Foundation of Quantum Mechanics I," *Proc. Phys. Math. Soc. Jpn.* **19,** 766–789.

Jauch, J. (1968), *Foundations of Quantum Mechanics,* Reading, MA: Addison-Wesley.

Jaynes, E. (1989), in: R. D. Rosenkrantz (Ed.), *Papers on Probability, Statistics, and Statistical Physics,* Dordrecht/Boston/London: Kluwer.

Kalmbach, G. (1983), *Orthomodular Lattices,* New York: Academic.

Keller, H. (1980), "Ein Nicht-klassischer Hilbertscher Raum," *Math. Z.* **172,** 41–49.

Kläy, M. (1988), "Einstein-Podolsky-Rosen Experiments: The Structure of the Probability Space, I," *Found. Phys. Lett.* **1,** 205–244.

Kläy, M., Randall, C., Foulis, D. (1987), "Tensor Products and Probability Weights," *Int. J. Theor. Phys.* **26,** 199–219.

Kolmogorov, A. (1933), *Grundbegriffe der Wahrscheinlichkeitsrechnung,* Berlin: Springer; translation (1956), *Foundations of Probability,* New York: Chelsea.

Kraus, K. (1983), *States, Effects and Operations,* Lecture Notes in Physics vol. 190, Berlin/Heidelberg/New York: Springer-Verlag.

Lock, P., Hardegree, G. (1984a), "Connections among Quantum Logics, Part 1, Quantum Propositional Logics," *Int. J. Theor. Phys.* **24,** 43–53.

Lock, P., Hardegree, G. (1984b), "Connections among Quantum Logics, Part 2, Quantum Event Logics," *Int. J. Theor. Phys.* **24,** 55–61.

Lock, R. (1981), *Constructing the Tensor Product of Generalized Sample Spaces,* Ph.D. Thesis, University of Massachusetts, Amherst.

Loomis, L. (1955), *The Lattice Theoretic Background of the Dimension Theory of Operator Algebra,* Memoirs of the American Mathematical Society vol. 18.

Ludwig, G. (1954), *Die Grundlagen der Quantenmechanik,* Berlin: Springer; translation (1983), *Foundations of Quantum Mechanics, I,* Berlin: Springer.

Mackey, G. (1957), "Quantum Mechanics and Hilbert Space," *Am. Math. Monthly* **64,** 45–57.

Mackey, G. (1963), *The Mathematical Foundations of Quantum Mechanics,* New York: Benjamin.

Maeda, S. (1955), "Dimension Functions on Certain General Lattices," *J. Sci. Hiroshima Univ.* **A19,** 211–237.

Mittelstaedt, P. (1978), *Quantum Logic,* Dordrecht: Reidel.

Murray, F., von Neumann, J. (1936), "On Rings of Operators," *Ann. Math.* **37,** 116–229.

Nishimura, H. (1980), "Sequential Method in Quantum Logic," *J. Symbol. Log.* **45,** 339–352.

Piron, C. (1976), in: A. Wightman (Ed.), *Foundations of Quantum Physics,* Mathematical Physics Monograph Series, Reading, MA: Benjamin.

Popper, K. (1959), *Br. J. Phil. Sci.* **10,** 25.

Pták, P., Pulmannová, S. (1991), *Orthomodular Structures as Quantum Logics,* Dordrecht/Boston/London: Kluwer.

Randall, C. H., Foulis, D. J. (1973), "Operational Statistics II, Manuals of Operation and Their Logics," *J. Math. Phys.* **14,** 1472–1480.

Randall, C., Foulis, D. (1978), "New Definitions and Theorems," University of Massachusetts Mimeographed Notes.

Stone, M. (1932), "On One Parameter Unitary Groups in Hilbert Space," *Ann. Math.* **33,** 643–648.

von Neumann, J. (1932), *Grundlagen der Quantenmechanik,* Berlin: Springer.

von Neumann, J. (1955), *Mathematical Foundations of Quantum Mechanics,* translation of von Neumann (1932), Princeton, NJ: Princeton Univ. Press.

von Neumann, J. (1960), *Continuous Geometry,* Princeton, NJ: Princeton Univ. Press.

Wick, G., Wightman, A., Wigner, E. (1952), "Intrinsic Parity of Elementary Particles," *Phys. Rev.* **88,** 101–105.

Wilbur, W. J. (1977), "On Characterizing the Standard Quantum Logics," *Trans. Am. Math. Soc.* **233,** 265–282.

Wright, R. (1978), "Spin Manuals," in: A. R. Marlow (Ed.), *Mathematical Foundations of Quantum Theory,* New York: Academic.

QUANTUM MEASUREMENTS

Jeffrey Bub, *Department of Philosophy, University of Maryland, College Park, Maryland, U.S.A.*

Daniel M. Greenberger, *Department of Physics, City College of New York, New York, New York, U.S.A.*

Introduction 257
1. **The Interpretation of Quantum Mechanics and the Measurement Problem** 257
1.1 The Interpretation of the Quantum State 257
1.2 The Measurement Problem 258
1.3 Von Neumann's Solution 259
1.4 Schrödinger's Cat 260
1.5 Other Solutions 260
2. **The Nonintuitive Nature of Quantum Theory and Some Measurement Consequences** 261
2.1 The Superposition Principle 262
2.1.1 Wave–Particle Duality 262
2.1.2 The Wheeler "Delayed-Choice" Experiment 263
2.1.3 Schrödinger's Cat Revisited 263
2.1.4 The Renninger–Dicke Effect 264
2.1.5 The "Quantum Eraser" 264
2.1.6 The Quantum Zeno Paradox 265
2.2 Nonlocality in One-Particle Systems 266
2.2.1 The Aharonov–Bohm Effect and the Berry Phase 266
2.3 Multiparticle Superpositions 267
2.3.1 The Einstein–Podolsky–Rosen "Paradox" 267
2.3.2 The Extension to Three-Particle States 269
3. **Conclusions** 270
Glossary 270
Works Cited 271
Further Reading 272

INTRODUCTION

The following discussion is divided into two parts. In part 1, we consider some general features of quantum mechanics and show how a so-called *measurement problem* is inherent in the quantum description of events. In part 2, we deal more specifically with quantum limitations on measurement introduced by quantum interference effects, and the possibilities this allows for the appearance of new nonclassical phenomena in the behavior of measurement devices.

1. THE INTERPRETATION OF QUANTUM MECHANICS AND THE MEASUREMENT PROBLEM

1.1 The Interpretation of the Quantum State

In classical mechanics, the state of a system fixes the values of all dynamical variables of the system and is in this sense "dispersion free." We can form probability distributions over these dispersion-free states to represent a situation where we have less than perfect knowledge of a system (as in classical statistical mechanics). In quantum mechanics, states specify probabilities for ranges of values of the dynamical variables or "observables" of a system and do not fix their values. Quantum states can be either "pure" or "mixed." A mixed state can be represented (in general, nonuniquely) as a probability distribution or mixture over pure quantum states, but a pure quantum state cannot be decomposed into a mixture of other pure quantum states.

The question arises as to whether the pure quantum states can, in principle, be replaced by more complete states defined in terms of additional "hidden variables" not considered in the current theory—perhaps states that would specify definite values for the observables, instead of probabilities. If this were possible, the pure-state probabilities could be interpreted as representing a

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

measure of our ignorance as to the precise dynamical state of the system, as in classical statistical mechanics (where we can think of the hidden variables as the positions and momenta that define the dispersion-free classical states of the particles that make up the system).

The question of hidden variables underlying the quantum statistics is sometimes posed as the question of whether a statistical interpretation of the pure quantum state is appropriate—in which the pure state is taken as providing a description of the statistical properties of an ensemble of similarly prepared systems—or whether we are required to interpret the pure state as a complete description of an individual system, so that the pure state represents, in some sense, a "rock-bottom" description of the system. The statistical interpretation has been defended by Ballentine (1970) and is often attributed to Einstein, while the opposing position is associated with Bohr and the "orthodox" or Copenhagen interpretation of the theory.

There is a confusion we should dispel here between two senses of the notion of a "statistical interpretation" of the quantum state. If the quantum theory is understood "instrumentally"—as merely a tool for making predictions of measurement outcomes on systems prepared in accordance with certain operationally defined procedures—then the quantum state indeed has a purely statistical significance: Different states simply represent different ensembles of systems prepared by different operationally defined procedures that yield different distributions of outcomes for subsequent measurements. This instrumental sense of statistical interpretation is neutral with respect to the question of hidden variables underlying the quantum statistics. It is only if we interpret the theory "realistically"—as a proposed structural description of dynamical processes at the microlevel—that the question of hidden variables underlying the quantum statistics arises in any interesting sense. In this realist sense of statistical interpretation, the proposed statistical ensemble to which the quantum state refers is defined by a distribution over definite values of hypothetical hidden variables that are presumed to characterize the precise dynamical state of the system—not a distribution over measurement outcomes.

There is a long history of investigations of the hidden-variable problem. A variety of "no go" theorems have been proposed, beginning with a theorem by von Neumann (1955), and also several hidden-variable theories, notably by Bohm and collaborators (Bohm, 1952; Bohm and Bub, 1966; Bohm and Hiley, 1993). The no go theorems prove the impossibility of hidden variables satisfying various constraints. One such physically motivated constraint is locality. Following a result by Bell (1964), extended in various ways by several authors (e.g., Clauser and Horne, 1974; Garuccio and Selleri, 1980; Garg and Mermin, 1964; Greenberger *et al.*, 1989, 1990) and eventually tested experimentally (Aspect *et al.*, 1982) (see Sec. 2.3.1), it is now widely accepted that any hidden-variable theory capable of reproducing the correlations predicted by quantum mechanics for measurements on certain separated systems cannot be a local theory. Roughly, there must be a mechanism in the hidden-variable theory that allows measurement outcomes of observables of a system S_1 to depend on measurement outcomes of observables of a system S_2, where the events in question can lie outside each other's light cones. (Bohm's 1952 theory incorporates such a mechanism.)

The argument for interpreting the pure quantum state as referring to an individual system and not an ensemble rests on the rejection of the statistical interpretation in the realist sense, i.e., the rejection of hidden variables underlying the quantum statistics. If there are no hidden variables, then the probabilistic description defined by the pure state provides as complete a description as possible of the dynamical state of an individual system. But now a fundamental conceptual problem of measurement arises for this orthodox interpretation of the quantum state.

1.2 The Measurement Problem

Any measurement involves an interaction between the system measured, S, and a second system, M, that functions as a measuring instrument. In an ideal measurement, the interaction must be such as to set up a correlation between the values of the measured observable of S and the values of an indicator or "pointer" observable R of M, so that the final value of R after the interaction has ceased indicates the value of the S observable measured in the interaction. Suppose, for simplicity, that a spin-$\frac{1}{2}$ particle interacts with a measuring instrument that ideally measures the z component of spin, and suppose that S is initially in some eigenstate, $|+\rangle$ or $|-\rangle$, of z-spin. Suppose that the interaction induces a transition from an initial or ready state $|r_0\rangle$ of the instrument M (an eigenstate of the indicator observable R) to a final state $|r_+\rangle$ or $|r_-\rangle$, depending on the initial eigenstate of z-spin. Formally, the state of the composite system $S + M$ undergoes the following (unitary) transformation:

$$|+\rangle|r_0\rangle \rightarrow |+\rangle|r_+\rangle$$

or

$$|-\rangle|r_0\rangle \rightarrow |-\rangle|r_-\rangle.$$

If, now, the initial state of S is a pure state represented by a linear superposition of spin eigenstates, say

$$c_+|+\rangle + c_-|-\rangle,$$

it follows from the linearity of the transformation that the corresponding transformation of the composite state is

$$(c_+|+\rangle + c_-|-\rangle)|r_0\rangle \rightarrow c_+|+\rangle|r_+\rangle + c_-|-\rangle|r_-\rangle.$$

On the orthodox interpretation of the quantum state, the pure state $c_+|+\rangle|r_+\rangle + c_-|-\rangle|r_-\rangle$ is a complete description of the dynamical state of the composite system, $S + M$, after the measurement interaction. But after the measurement, we expect that the instrument registers either r_+ or r_-, depending on whether the spin state of the system S is found to be $+$ or $-$, i.e., we expect the final quantum state of the composite system $S + M$ to be *either* $|+\rangle|r_+\rangle$ with probability $|c_+|^2$, *or* $|-\rangle|r_-\rangle$ with probability $|c_-|^2$. In other words, the measurement interaction should result in the stochastic transition from the pure state $(c_+|+\rangle + c_-|-\rangle)|r_0\rangle$ to the mixed state representing a probability distribution over the pure states $|+\rangle|r_+\rangle$ and $|-\rangle|r_-\rangle$ with weights $|c_+|^2$ and $|c_-|^2$, respectively, rather than the deterministic transition from the pure state $(c_+|+\rangle + c_-|-\rangle)|r_0\rangle$ to the pure state $c_+|+\rangle|r_+\rangle + c_-|-\rangle|r_-\rangle$.

Of course, the pure state $c_+|+\rangle|r_+\rangle + c_-|-\rangle|r_-\rangle$ describes an ensemble in the instrumental sense: an ensemble over the possible outcomes ($+$ and $-$) of a *further measurement* by a second instrument M' on the composite system $S + M$. But this is not an ensemble over actual values of the relevant observables, z-spin of S and the pointer observable R of the instrument M. These values are indeterminate in the pure state $c_+|+\rangle|r_+\rangle + c_-|-\rangle|r_-\rangle$, just as z-spin is indeterminate, neither $+$ nor $-$, in the state $c_+|+\rangle + c_-|-\rangle$. And, of course, if the further measurement of $S + M$ by M' is treated as an interaction between M' and $S + M$, the same problem arises.

1.3 Von Neumann's Solution

Von Neumann's solution to this measurement problem—that the theory yields a pure state instead of the expected mixture for the state of the composite system $S + M$ after measurement—was to propose two modes of evolution for quantum systems (von Neumann, 1955, pp. 351, 417 ff). Systems that are not being measured undergo unitary time evolutions described by Schrödinger's time-dependent equation of motion. Measurements involve nonunitary, stochastic evolutions described by a new "projection postulate" introduced by von Neumann. [The term "projection postulate" is due to Margenau (1936), who rejected the postulate.] The projection postulate validates the transition from $c_+|+\rangle + c_-|-\rangle$ to one of the pure states $|+\rangle$ or $|-\rangle$ in a measurement of z-spin, depending on the outcome, r_+ or r_-, of the measurement, or the transition from the pure state $c_+|+\rangle + c_-|-\rangle$ to the mixed state representing a probability distribution over the pure states $|+\rangle$ and $|-\rangle$ with weights $|c_+|^2$ and $|c_-|^2$, respectively. The pure state is said to "collapse" onto one of the eigenstates of z-spin, $|+\rangle$ or $|-\rangle$, in the measurement, with probability $|c_+|^2$ or $|c_-|^2$, respectively. Similarly, a position measurement involves a collapse or "reduction" of the wave packet.

Von Neumann's justification for this duality was that a measurement divides the world into two parts: the observed system and the observer, with the boundary between the two somewhat arbitrary. On his view, we can follow all physical processes in the observed system with arbitrary precision, in principle, but not in the observer. The unitary transitions of quantum mechanics describe events in the observed part of the world, so long as they do not interact with the observing part. When such an interaction occurs, the projection postulate applies.

Von Neumann proved a consistency theorem for the two modes of evolution. Consider the measurement of S by an instrument M, and a further measurement of $S + M$ by M'. We can think of S as the system actually observed, M as the measuring instrument, and M' as the observer, or M as the instrument and M' as a recording device. The boundary between the observed system and the observer can be drawn between S and $M + M'$ or between $S + M$ and M'. In the first case, the projection postulate is applied to the interaction between S and $M + M'$. Prior to the interaction, S evolves unitarily. In the second case, the projection postulate is applied to the interaction between $S + M$ and M'. Prior to the interaction, $S + M$ evolves unitarily. Von Neumann showed that both procedures yield the same probabilities for S observables (S being the only system that belongs to the observed part of the world, however the boundary between the observer and the observed is drawn). We have consistency in the sense that the application of the projection postulate directly to the system S is consistent with its application to the system $S + M$, after a suitable unitary interaction between S and M governed by the equation of motion of the theory.

Now, showing that the postulation of a special mode of evolution for systems undergoing measurement avoids a certain kind of inconsistency with the unitary evolutions of quantum-mechanical interactions does not explain why a fundamental theory of mechanics should incorporate a special postulate applicable only to state transitions on measurement, nor why these transitions should yield mixtures from pure states. Von Neumann's explanation that unitary evolutions can be expected to apply only to the observed part of the world, and that a special process should apply to the "interaction" between the observed part and the observer, is hardly reassuring, if only because classical mechanics requires no such process.

1.4 Schrödinger's Cat

The problem is often dramatized as the problem of Schrödinger's cat (Schrödinger, 1935), where macroscopic properties of a cat become indeterminate in a measurement interaction. Suppose we take the interaction between S and M as some interaction that correlates the states $|+\rangle$ and $|-\rangle$ of a microsystem S with the states $|r_+\rangle = |\text{alive}\rangle$ and $|r_-\rangle = |\text{dead}\rangle$ of a cat (M). We might suppose that the state space of the cat is divided into two parts, representing the quantum state of the live cat and the quantum state of the dead cat. Then $|\text{alive}\rangle$ and $|\text{dead}\rangle$ represent two of these states that become correlated with microsystem states $|+\rangle$ and $|-\rangle$ in the interaction. After the interaction, the cat is alive if the final state of the composite system is $|+\rangle|\text{alive}\rangle$, the cat is dead if the final state of the composite system is $|-\rangle|\text{dead}\rangle$, but the cat is neither alive nor dead if the final state is the linear superposition $(c_+|+\rangle|\text{alive}\rangle + c_-|-\rangle|\text{dead}\rangle)$, which is the final state required by the linear dynamics. Schrödinger regarded this consequence of the theory as unacceptable.

What we expect, on the basis of the orthodox interpretation of the quantum state, is that the final state of the composite system is a mixed state representing a mixture of the pure states $|+\rangle|\text{alive}\rangle$ and $|-\rangle|\text{dead}\rangle$, with weights $|c_+|^2$ and $|c_-|^2$ respectively. But the dynamics of the theory yields a final pure state of the form $(c_+|+\rangle|r_+\rangle + c_-|-\rangle|r_-\rangle)$. The evolution to the mixed state is nonunitary, stochastic, and entropy increasing, while the evolution to the pure state is, of course, unitary and deterministic (via Schrödinger's time-dependent equation of motion), and does not increase the entropy of the composite system.

1.5 Other Solutions

Proposed solutions to the measurement problem either

1. take the dynamics of the theory seriously and modify the orthodox interpretation of the quantum state, or
2. accept the orthodox interpretation of the quantum state and modify the dynamics, or
3. accept both the dynamics and the ortho-

dox interpretation of the state and attempt to show that, for all practical purposes, the interference terms can be ignored in real measurement interactions.

Type **1** or "no collapse" solutions include Bohm's 1952 hidden variable theory, the "many-worlds" interpretation proposed by Everett (1957), and the variety of modal interpretations related to van Fraassen's "modal" interpretations. The central idea here is that a pure state can be interpreted, without inconsistency, as specifying actual values for certain observables, even when the state is not an eigenstate of these observables. Different versions of this proposal (van Fraassen, 1981; Kochen, 1985; Healey, 1989; Dieks, 1989; Bub, 1992) involve different ways of selecting these observables for a given pure state. The aim is to show that even when the state of the composite system $S + M$ is a pure state of the form $(c_+|+\rangle|r_+\rangle + c_-|-\rangle|r_-\rangle)$, both z-spin and the indicator observable R have definite values.

Type **2** or "collapse" solutions include the theory of Ghirardi *et al.* (1986) and the Bohm–Bub hidden variable theory (Bohm and Bub, 1966). Essentially, the Ghirardi–Rimini–Weber theory involves a modified dynamics for the quantum state that results in the occasional spontaneous localization of individual microsystems—so infrequently that the time evolution of a single particle remains unitary over long time intervals with high probability. For macrosystems, the modified dynamics results in the almost instantaneous transition of superpositions over states separated by macroscopic distances to the corresponding mixtures over these states. So linear superpositions over different position indicator states of a macroscopic measuring instrument are very short-lived and collapse immediately to a single pointer position state, with the appropriate probability. The theory resolves the measurement problem by modifying quantum dynamics in such a way that the behavior of microsystems is essentially unchanged, while the behavior of macroscopic systems, induced by the dynamics of their constituent microsystems, conforms to classical mechanics. The Bohm–Bub theory also introduces a modified equation of motion, with an additional nonunitary term that has a negligible effect on the quantum state, except during certain processes that include measurement processes. The nonunitary term involves hidden variables that determine the outcome of the collapse. Averaging over the hidden variables yields the quantum statistics.

Finally, type **3** or "for all practical purposes" solutions include such approaches as the Daneri–Loinger–Prosperi (1962) quantum ergodic theory of macrosystems, "decoherence" theories (Joos and Zeh, 1985; Zurek, 1993; Paz and Zurek, 1993), and various other attempts to exploit the large number of degrees of freedom of the amplifying recording device or the environment of the system as the source of a dissipative process that effectively yields a mixed state after a measurement.

While none of the proposals for resolving the measurement problem has achieved general acceptance, the detailed investigation of measurement processes in quantum mechanics from a wide variety of perspectives has considerably sharpened the debate on the conceptual foundations of quantum mechanics, and the peculiar nonclassical features of quantum interference effects have been brought out in new and unexpected ways.

We turn now to a discussion of some of these effects in quantum measurements.

2. THE NONINTUITIVE NATURE OF QUANTUM THEORY AND SOME MEASUREMENT CONSEQUENCES

The nonintuitive aspects of the theory can be ascribed primarily to three sources. First is the superposition principle, whereby quantum amplitudes add, rather than probabilities (which are determined by the square of the magnitude of the amplitude). The second is the measurement postulate, which states that a measurement yields only eigenvalues of the operator representing the observable being measured and "forces" the system into an eigenstate of this operator. In other words, the measurement acts as a filter, and forces the system to choose an eigenstate among the various possible outcomes of the experiment. This reduction or collapse of the state is also sometimes known as the "projection postulate," which we have discussed.

The main consequence of superposition is that before a measurement is made, a system may have a component of the amplitude corresponding to each possible outcome of a future measurement, and that all these possibilities may produce interference, which can be seen by varying some system parameter. Different aspects of interference are evident in one-particle and multiparticle systems. Once a measurement is made, the system will be forced into one of those states of the variable being measured, and other states cannot contribute, provided that they are experimentally distinguishable from this state.

The third nonintuitive effect is due to nonlocality, of which there are two main types. In the first, two systems that should intuitively be independent of each other can produce interference that cannot be explained by any local, realistic model, such as in the Einstein–Podolsky–Rosen "paradox." In the second, one particle confined to one of two paths nonetheless seems to be aware of what happens between the two paths, even though it has no amplitude for being there. This type of nonlocality is exemplified by the Aharonov–Bohm effect and the Berry phase, both of which seem to be topological in nature.

There are many aspects of these effects, and we shall be content with a few examples that illustrate the issues involved.

2.1 The Superposition Principle

2.1.1 Wave–Particle Duality The most obvious effect of the superposition principle is what has been called the wave–particle duality. Bohr always emphasized the principle of "complementarity," according to which one can measure either the particle or wave aspects of a quantum system, but not both together. This is a special case of the rule that if two possible paths for a system are distinguishable in principle by the experimental arrangement employed, one cannot see interference between them. This is the "particle" aspect of the system since classically, particles do not interfere. If the paths are indistinguishable, they can interfere. This interference manifests their "wave" aspect.

In the early days of the theory it was thought that this was necessarily an "all or nothing" split; however, it has been shown that in the classic two-slit interference experiment, one can characterize the system as partly wave, partly particle (Wootters and Zurek, 1979). The more particlelike the system is, the less wavelike it will be.

We illustrate some of the properties of particles and waves with a very simple model of a neutron interferometer. (The model is so simple that the particles passing through it can just as well be thought of as photons.) The neutron has shown wavelike behavior over an astonishingly large range of energies, from 10^{-7} up to 10^{9} eV. In the neutron interferometer, its energy is thermal, about 10^{-2} eV, with a de Broglie wavelength of about 10^{-8} cm, so that it undergoes coherent Bragg scattering off perfect crystals. A typical interferometer is a few centimeters long, and the neutron is inside it for about 10^{-5} s, and only about one coherent neutron per second passes through it, so that there are rarely two inside simultaneously. So each particle is clearly interfering only with itself. A simple but more detailed description of the device can be found in Greenberger (1983).

The device (Fig. 1) is called a Mach–Zehnder interferometer. It contains a beam splitter, A, such that there is a 50% chance that a particle will be either transmitted or reflected. (For probability conservation, imagine the reflected beam being retarded by 90°.) After being split, the two diverging beams are refocused by mirrors. There is also a phase shifter F, which delays the phase in one beam by a phase α. The beams converge on a second beam splitter. After the second beam splitter, there are two detectors, and a given particle will land in one of the detectors. The relative number of counts in each of the detectors depends on α, and in

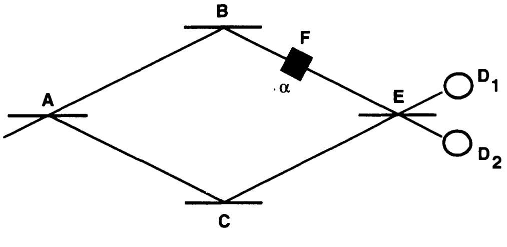

FIG. 1. Model of a neutron interferometer. This simple model shows the most important aspects of particle interferometers. Beam splitters are located at A and E to split and recombine the beams, which are also bent by mirrors at B and C. F is a phase shifter, which retards one beam by phase α. This changes the relative counting rates at the detectors D_1 and D_2.

Fig. 1, the counting rate in counter D_1 will be $\cos^2(\alpha/2)$ and in counter D_2 will be $\sin^2(\alpha/2)$. Thus the interference can be detected by varying α.

2.1.2 The Wheeler "Delayed-Choice" Experiment In the apparatus we have described, if one thinks of the incident neutron as a particle, one would expect it to take one of the two paths through the interferometer. If one thinks of the neutron as a wave, like a light wave, one might think it splits and takes both paths. An especially dramatic "proof" that both these points of view are incorrect is given by Wheeler's "delayed-choice" experiment (Wheeler, 1978), which shows nothing so much as that the results of an experiment depend critically on exactly what experiment is actually performed. For instance, imagine a neutron passing through a system such as that in Fig. 1. There is an amplitude for the neutron to take either path to the second beam splitter, where there will be interference. However, imagine that the second beam splitter at E is removed just before the neutron reaches it. Then there will be no combining of the amplitudes, and no interference. The particle in either beam will head straight toward its detector. So the particle in this case will land in one detector or the other, and the probability for landing in each detector will be 50%, independent of α, proving that it was in one beam or the other, just as one might expect for a classical particle. However, if the beam splitter at E is not removed, one will get interference, depending on α.

From a quantum-mechanical point of view this is a simple experiment, and shows that the result depends on whether the beam splitter at E is present or not. But from a classical point of view, whether the particle is to act as a particle and be completely in one beam, or act as a wave and be somehow partly in both beams, must be determined back at the beam splitter at A. And yet the decision to have the beam splitter at E present or absent need not be made until the neutron has already long since passed through the beam splitter at A, and is in fact just about to reach E. So we see that it is a mistake to think of the neutron as somehow having a "trajectory," as a classical particle would. Actual experiments equivalent to this have been performed with photons (Alley *et al.*, 1984; Hellmuth *et al.*, 1986), where they yield the expected quantum result.

2.1.3 Schrödinger's Cat Revisited The dilemma presented by the delayed-choice experiment is exactly the same as that in the famous example of Schrödinger's cat (Schrödinger, 1935), which we have already discussed as a measurement problem. There, a cat is in a closed cage with a bottle of poison gas. Whether the bottle is to be opened or not depends on the outcome of some atomic decay event. If we have not opened the cage, there is an amplitude for the decay to have happened, which translates to the same amplitude for the cat to be dead. So we have an amplitude for the cat to be either alive or dead, until we look.

There is a perfect analogy between the cat experiment and the delayed-choice experiment we have discussed. In the interferometer, after the particle has passed the first beam splitter, if the second beam splitter is absent the particle has an amplitude to be in one beam or the other, which corresponds to the cat being dead or alive. But as in that case we cannot know which beam it is in until it is detected. This is even more convincing than the cat experiment, because here we can check that a superposition really exists, by inserting the second beam splitter. But in the cat experiment, there is no way to show that a superposition of the two states really exists before we look.

So, the Schrödinger cat experiment actually works for neutrons or photons, but it is probably unlikely to do so for cats! And yet the only difference is their size, or rather their coherence properties. For a cat would transfer momentum to its cage whenever it breathed or twitched, which could be detected from the outside. The same is true for the breaking of the bottle, or even the thermodynamically irreversible escape of the gas from the bottle. So for a truly macroscopic object like a cat in its environment there is no realistic way to perform such an experiment without the information becoming available outside, to a careful observer, or frittered away inside through friction, etc. The real question is at what point does it become impossible to retain coherence in a many-body system. Here, the barriers have been pushed back, and coherence seen on a larger and larger scale (Webb and Washburn, 1988), and the question is still open. Leggett (1984) has made a careful theoretical investigation of the subject.

2.1.4 The Renninger–Dicke Effect A further complication to the notion that the reduction of a wave packet must come from the macroscopic nature of the observation apparatus comes from the possibility of reducing the wave packet by failing to measure it (Renninger, 1960; Dicke, 1981, 1986). If in our interferometer in Fig. 1 we replaced the phase shifter at F by a completely absorbing crystal, we could arrange it that if the neutron passed through beam ABE it would be absorbed by the crystal with 100% certainty. Furthermore, on absorption, the crystal would emit a γ ray, which could be detected with high efficiency. Thus if one failed to detect the γ ray, one would know that the neutron had taken the path ACE. So here the wave function has been reduced by *not* detecting the particle!

But there is a difference between detecting the particle and not detecting it. Once the counter has fired, all coherence has been lost, and so all phase knowledge has been destroyed. However if the counter has not fired, the reduction in the wave packet acts like a pure filter. The possibility that the particle has taken the path ABE has been removed, but the wave function has not been totally reduced. For example, if the neutron were in some definite spin state processing about a magnetic field, then the knowledge of the precession phase would be retained and could be made use of further down the line, in the path ACE. However, in the path ABE, any such knowledge would be destroyed by the detection of the particle.

2.1.5 The "Quantum Eraser" Another complication is revealed by Scully's "quantum-eraser" experiment (Scully *et al.*, 1978, 1991). Here, even after one has seemingly "erased" information in a system by some suitable experimental arrangement, one still has the choice of whether to confirm the information that remains available by making a measurement on the system. If one does not bother to make the measurement, the information one has "lost" is still coherent with that part of the wave function one has retained. Therefore, by further changing the experimental setup, it should be possible to regain the "erased information," whereas had one made the original measurement this would not be possible.

As an example, return to the system in Fig. 1 with a neutron beam passing through it. This time, imagine that the beam is polarized upward, out of the plane of the figure, in the z direction. As the two beams emerge from the first beam splitter at A, there is a polarization device that rotates the angle of polarization of one of the beams, say the one along the path ACE, by 90° so that this beam is now polarized in the plane of the paper, say along the x direction. The two beams are now polarized perpendicularly to each other. Assume that the detector at D_1 is sensitive to polarization. Then if it counts only z-polarized neutrons, there can be no interference, because all the neutrons it counts will have taken the path ABE, since they are polarized along the z axis. Similarly, if the detector at D_1 only counts x-polarized neutrons then there again will be no interference because all the neutrons it counts will have taken the beam ACE.

However, if the detector D_1 is now rotated so that it is in the x-z plane, at 45° from each axis, then it can count particles from both beams, and one is no longer aware of which path any neutron took, and one will have interference between the two beams, which can be seen by varying the angle α. Thus by changing the experimental setup, what appeared to be two incoherent beams are seen to be coherent. One would have had to make a measurement on the intermediate system (i.e., the separate beams) to confirm that they were incoherent. By not doing so, one can yet manipulate the system to make it coherent again. (See Kwiat *et al.*, 1992. They did a two-particle version of this experiment. The effect is much more subtle in that case, since the coherence shows up only in coincident counts between the particles. The individual particles appear to be totally incoherent.) This illustrates Wheeler's dictum "No elementary phenomenon is a phenomenon until it is a registered (observed) phenomenon" (Wheeler and Zurek, 1983, p. 192).

Scully *et al.* (1991) have suggested an even subtler version of this experiment. In one version it involves passing an atomic beam successively through two microwave cavities. In either cavity, the beam can be made to interact with the cavity such that the atom can absorb or emit a photon. If it is initially in an excited state, and after it passes the second cavity it is in its ground state, it must have given up its photon and decayed in one of the two cavities. Since it could have decayed in either cavity, the amplitudes for these two possibilities might be expected to interfere with each other.

But in fact, one can trap the photon in the cavity. Thus, by looking to see which cavity has the photon, one can tell which of the two cases occurred. Because one can distinguish these two cases experimentally, the two amplitudes cannot interfere. But if one were to couple the cavities weakly so as to mix the photon coherently between them, then one can no longer tell in which one it was created. This would then make it possible for the amplitudes to interfere again, even though by now the atom is a long way from the cavities.

Here, indistinguishability is used to recreate coherence in the system long after the two subsystems have ceased to interact. (To see interference, one would have to count coincidences between the particle beam and the cavity photons.)

2.1.6 The Quantum Zeno Paradox

The wave function of a system evolves continuously according to the Schrödinger equation. However, when a measurement is made, there is a discontinuous reduction of the wave packet. The concepts of a continuous evolution and a measurement must be rigorously distinguished. In fact, if one could make a continuous set of measurements on a system, the system would never evolve. That is the essence of the quantum Zeno Paradox (Misra and Sudershan, 1977; Peres, 1980).

We can illustrate the point simply. Consider a two-state system, $|1\rangle,|2\rangle$, that starts in state $|1\rangle$ at $t = 0$. The energy expectation values in the two states are E_1 and E_2 and there is a cross term ϵ in the Hamiltonian connecting the two states. Then the amplitude for the system to be in state $|2\rangle$ at short times t is $(i\epsilon t/\hbar)|2\rangle$, and the probability of finding the system there is $(\epsilon t/\hbar)^2$. This is quadratic in t, and so in the limit of very short times the transition rate (probability/time) for making a transition from $|1\rangle \rightarrow |2\rangle$ goes to zero.

A pretty example consists of a photon moving in the z direction passing through an optically active medium that rotates the polarization in the x-y plane. The medium is of length L, which rotates the polarization by $\pi/2$. A photon whose polarization is initially pointing along the x direction will emerge from the medium polarized along the y direction. The polarization will vary as the photon passes through the medium as $\mathbf{p} = \mathbf{i} \times \cos(\pi z/2L) + \mathbf{j} \sin(\pi z/2L)$. A polarization analyzer pointed along the x direction and placed at the far end of the medium will have probability $P = 0$ of detecting the photon.

Now imagine that the medium is split in half at $z = L/2$, and an analyzer pointed along the x axis is placed at $z = L/2$, while the one at L is retained. Then the photon will be pointed at $\pi/4$ to the x axis as it hits the first analyzer, so it has a probability $P_1 = \frac{1}{2}$ of passing through the analyzer and it will also have probability $P_2 = \frac{1}{2}$ of passing through the second analyzer. The probability that it passes through both analyzers is then $P = P_1P_2 = \frac{1}{4}$.

If there were three equally spaced analyzers, the probability for passing the first analyzer would be $P_1 = \cos^2(\pi/6) = \frac{3}{4}$, and the probability of passing all three analyzers would be $P = P_1P_2P_3 = [\cos^2(\pi/6)]^3 = 27/64 = 0.42$. For n equally spaced analyzers,

$$P = \left(\cos^2 \frac{\pi}{2n}\right)^n \underset{n \to \infty}{\longrightarrow} \left(1 - \frac{\pi^2}{4n}\right) \rightarrow 1.$$

Here, each analyzer makes a measurement of the polarization, "reducing" the wave function to the state of x polarization. In the limit of many closely spaced measurements, the transition is completely suppressed. While illustrative, of course this experiment can be interpreted classically, in terms of electric fields, rather than photons.

However a beautiful conceptually similar experiment was performed by Itano *et al.* (1990) using an atom in one of two states, 1 and 2. A resonant pulse was used to cause transitions between the states, playing the role of the medium rotating the polarization in our example. The measurement was made by using a third atomic level, which could make transitions to level 1 but not level 2. A pulse of radiation of frequency ω_{13} would cause transitions between these two levels

(1–3) and could be monitored. This could tell how many atoms were in level 1 at the time of the pulse (so that it "reduced the atomic state" to level 1, which could make a transition). The other atoms were reduced to level 2 (the Dicke–Renninger effect). By making the pulses close together, transitions between levels 1 and 2 were inhibited. While the experiment beautifully illustrates the laws of quantum theory, there is some controversy as to whether one is seeing a pure "Zeno paradox" effect. The problem is whether one has indeed destroyed all coherence by the measurement. (Our example above is subject to the same objection.)

Rather than using oscillating states, one can also see an inhibition effect with decaying states. Even exponentially decaying states go through a short period with a quadratic component, although experimentally this is almost impossible to see.

2.2 Nonlocality in One-Particle Systems

2.2.1 The Aharonov–Bohm Effect and the Berry Phase The Aharonov–Bohm effect (Aharonov and Bohm, 1959) (A–B) is a striking demonstration of the nonlocality of quantum theory, and it produces a result that is classically very puzzling, but very necessary in quantum theory. According to this effect, a beam of particles that is split and later recombined will experience a phase shift if a magnetic field is present anywhere in the area between the split beams, even though neither beam itself is exposed to a magnetic field. We discuss the effect only to the extent needed to show why it is crucial to quantum theory.

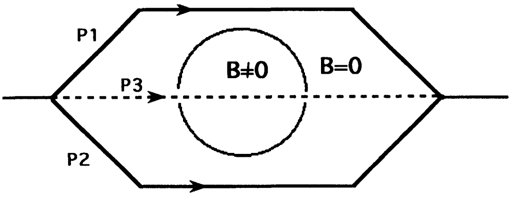

FIG. 2. The Aharonov–Bohm effect. This is a nonlocal effect that can be demonstrated by passing a beam of electrons around both sides of a solenoid (whose axis is perpendicular to the plane of the paper), enclosing a magnetic field. Even though neither beam P_1 nor P_2 ever passes through the field, there will be a phase shift between them when they recombine, proportional to the enclosed flux. One may develop an intuitive feeling for why this phase shift is necessary by considering a third beam, P_3, that does penetrate the field.

In the A–B effect, in its simplest form, an electron beam is split and the two halves go, via paths P_1 and P_2, around opposite sides of an infinite solenoid whose axis is perpendicular to the plane of the beam (Fig. 2). In both classical and quantum theory, an external magnetic field is added to the Hamiltonian by changing p to $p - eA/c$, where p is the canonical momentum and A is the vector potential. This can be accommodated in the quantum case by changing the state function ψ to $\psi\exp[(ie/\hbar c)\int_P \mathbf{A}\cdot d\mathbf{l}]$, where the line integral is along P, the classical path of the particle. When the two parts of the electron wave function recombine, the total accumulated phase difference between them is

$$\frac{ie}{\hbar c}\int_{P_1}\mathbf{A}\cdot d\mathbf{l} - \frac{ie}{\hbar c}\int_{P_2}\mathbf{A}\cdot d\mathbf{l} = \frac{ie}{\hbar c}\oint \mathbf{A}\cdot d\mathbf{l}$$
$$= \frac{ie}{\hbar c}\int \mathbf{B}\cdot d\mathbf{S} = \frac{ie}{\hbar c}\Phi,$$

where Φ is the flux enclosed by the solenoid between the two paths. So there will be a phase shift between the two beams, proportional to this flux, which is a gauge-invariant physical effect. What is so surprising about the effect is that it occurs even though neither electron beam has ever been in the magnetic field, or ever feels any force.

One can get a good insight into why such an effect must occur within quantum theory. The path integral is necessary there because it is what produces the phase changes necessary to bend the beam in a real magnetic field. But it is necessary even when the beam avoids the real fields. Consider a third beam, the dotted beam in Fig. 2, which passes along path P_3, right through the solenoid, through two tiny holes, which produce a negligible flux leak. Now consider the same experiment as above, only with one beam traveling along path P_1, and the other along path P_3. In this case, the phase difference will be only half as great, $(ie/2\hbar c)\Phi$, with path P_3 leading path P_1, but it is easy to understand. The beam inside the solenoid feels a force, bending it sideways (to the right for a + charge), away from the beam P_1, and this real deflection causes the phase shift.

Now consider yet another version of the same experiment, this time with the beam split into the paths P_3 and P_2. In this case also, the phase shift will be only half as great, with beam P_2 leading P_3, and again it is easy to understand. For the beam P_3 is again subject to a force, and again it is to the right, toward the beam P_2, and beam P_2 leads P_3, the force effects being asymmetrical in the two cases. But now we see that the phase of P_2 leads that of P_1 by the required amount (since P_2 leads P_3, and P_3 leads P_1), even though neither beam ever feels any force.

The reason for this phase shift is that the Hamiltonian contains the effects of every possible experiment that can be performed on the system. One can choose to create a wave packet that goes to the left or right of the field, but the results had better be consistent with those for a wave packet that passes through the field. Thus the integrated effects of all phase changes are contained in the solutions to the Hamiltonian, because the entire problem is nonlocal, quantum mechanically. The locality is an illusion created by choosing to perform an experiment with a local wave packet. In contradistinction to this, classically a particle is affected only by actual fields locally present along its trajectory.

So quantum mechanically the electromagnetic fields themselves do not contain enough information to determine all the physics inherent in the system, and the vector potential contains too much (since one can make gauge transformations). Actually, it is the phase factor $\exp[(ie/\hbar c)\int_P \mathbf{A}\cdot d\mathbf{l}]$ that contains exactly the information needed (Wu and Yang, 1975). This information is nonlocal, being path dependent. In practice, one always observes the difference between two paths, which is gauge invariant. The nonlocality is especially apparent in that the path enclosing the flux cannot be shrunk down to a point.

Experimentally, the A–B effect has been repeatedly confirmed, but most spectacularly with tiny, superconducting circuits (Tonomura, 1987), and the effect occurs in many forms throughout quantum physics (Olariu and Popescu, 1985; Peshkin and Tonomura, 1989; Anandan, 1990).

For example, there is also an electrostatic version (Allman *et al.*, 1993) of this experiment, where the vector potential is replaced by the scalar one, $(ie/\hbar)\int_P V\,dt$. This experiment has been done on neutrons, using the scalar potential $\boldsymbol{\mu}\cdot\mathbf{B}$, rather than the electrostatic one. (The interpretation is somewhat controversial, because although there is no force on the neutrons, there is a torque.) Finally, a symmetrical analog of the A–B effect exists, called the Aharonov–Casher effect, where the solenoid is replaced by a line electric charge, and the phase shift is produced on a split beam of particles with a magnetic moment. Again, this effect has been confirmed with neutrons (Cimmino *et al.*, 1989).

A closely related nonlocal effect is the Berry phase (Berry, 1984; see TOPOLOGICAL PHASE EFFECTS). If one imagines a particle moving in an external field—for example, a neutron in a magnetic field—then if one very slowly changes the field, the particle will slowly accommodate to the change. Classically, one can find some adiabatic invariant of the motion, which will not change as the field changes. Quantum mechanically, the particle will stay in an eigenstate corresponding to the slowly changing field (in the neutron example above, if the neutron were in the $m_z = +\frac{1}{2}$ state of the original field, it will stay in that state of the adiabatically varied field).

Finally, if the field ultimately returns to its original value, the particle will be back in its original state, and so the wave function cannot differ from the original one by more than a phase factor. Part of this phase factor has a deep topological significance in the space defined by the parameters of the external field (B_x, B_y, B_z in the case of the neutron). If one plots the changing direction of $\mathbf{B}$ in this space, $\mathbf{B}$ will return to its initial value, and thus form a cone, with the vertex at the origin. The solid angle swept out by this cone is proportional to the phase factor multiplying the neutron wave function, and is called the Berry phase. The Aharonov–Bohm effect can be considered as a special case of the Berry phase. Many experiments have been carried out to confirm the effect in many contexts, and its significance has been carried over to nonadiabatic changes (Shapere and Wilczek, 1989).

2.3 Multiparticle Superpositions

2.3.1 The Einstein–Rosen–Podolsky "Paradox" The extension of the superposition theorem to multiparticle systems brings both new phenomena into play and new insights. There is a type of nonlocality present in multiparticle systems that is different from the above for one-particle systems, and which is very troubling, from a classical point of view. This is the kind of nonlocality first discussed by Einstein, Rosen, and Podolsky (EPR) (Einstein *et al.*, 1935), and it occurs when the wave function of the two particles cannot be factored into the product of two one-particle wave functions. If the particles are independent of each other, they can be represented as such a product, $\psi(1,2) = \varphi_1(1)\varphi_2(2)$, and if the wave functions do not overlap, the particles will act independently of each other, in the sense that any forces, etc., applied to one of them will have no effect on the other.

However, when the total wave function cannot be factored (which will then be true in any representation), the particles can have long-range effects on each other. Schrödinger (1935) called such nonfactorable states "entangled states." The classic example, introduced into this discussion by Bohm (1951), is the singlet state of two spin-$\frac{1}{2}$ systems, $\psi(1,2) = 2^{-1/2}(|+\rangle_1|-\rangle_2 - |-\rangle_1|+\rangle_2)$, where $+,-$ refer to the spin states of the particles, and 1,2 refer to the particles.

EPR introduced the idea of an "element of reality," which refers to a property that objectively belongs to a particle independently of how an experimenter manipulates the system. For one-particle systems it is essentially impossible to introduce such a notion quantum mechanically, but EPR noticed that for an entangled state, such as the Bohm state above, one can do this. For example, if the particles are far apart, but one measures the spin of particle 1 to be "up" $(+)$, then one immediately knows that the spin of particle 2 must be "down" $(-)$, and vice versa. Thus one has determined the spin of particle 2, without in any way interfering with the particle, and so to EPR this seemed like an objective determination of a property already there, as the experiment could not have affected this property. The states of these two particles are said to be "perfectly correlated," because one can make a prediction of the state of the second particle with 100% accuracy by measuring the first one. Such perfect correlations are necessary for elements of reality to exist. They also introduced the idea of "locality," which they defined as an experimental situation wherein nothing done to one of the particles could have any effect on the other—for example, if they are far away from each other. Thus according to EPR, one has proven that in the Bohm state, each of the particles must have a definite spin even if we are unable to know what it is before making a measurement.

Of course quantum theory totally denies this explanation and says that the spin of particle 2 could not have been determined until one measured particle 1. The two points of view, that of EPR and that of quantum theory, are totally at odds, but it was long taken for granted that there was no experimentally detectable difference between the two viewpoints. But amazingly, Bell (1964) was able to show that by using the EPR concepts of reality, locality, and completeness (they called a theory complete if all elements of reality were accounted for by the theory), in the Bohm situation of a particle of spin 0 decaying into two particles, each of spin $\frac{1}{2}$, in an entangled state, one could actually show that the EPR point of view had detectable consequences, which violated the results of quantum theory. He proved that in an experiment that measures the correlation between the spin states of the spin-$\frac{1}{2}$ particles, it follows from the EPR assumptions that there is an upper limit to the possible correlations one can obtain between the spins. It is easy to show that a quantum-mechanical calculation of this correlation violates this limit, which has been taken to mean that by making such a measurement one could experimentally rule out the possibility of obtaining a classically realistic and local interpretation of quantum theory.

Bell's theorem set off a major experimental effort to see whether this upper limit of correlations was violated in nature. First, one had to make some plausible assumptions in order to accommodate the fact that experimental detectors are not 100% efficient (Clauser *et al.*, 1969), and then one could apply the results to actual experiments. Many such experiments have been made, and they overwhelmingly endorse the quantum-mechanical picture of the problem. The first such experiment was by Clauser and Freedman (1972), and the most famous by Aspect *et al.* (1982). (For a critical discussion of the early experiments, see Clauser and Shimony, 1978, and for a later, skeptical interpretation, see Selleri, 1990.)

Recently, a new technique, parametric down-conversion (the original experiments were Shih and Alley, 1986, and Ghosh and Mandel, 1987), has been developed, wherein one can create a definite quantum-mechanical two-photon state, which is distinct from a laser-created coherent state that may have two particles present on average. One can show that in such states one can see two-particle correlations, but one can no longer detect one-particle correlations. This means that one can measure coincidences between the two photons and see correlations in the joint counting rate, but for either photon separately one cannot detect any interference between two different paths taken by the photon (as in a Young's two-slit diffraction experiment). This is a general feature of two-particle entangled states.

With this technique one can perform clean tests of Bell's theorem (Rarity and Tapster, 1990) by measuring coincidences between the two photons, and these experiments have also come down clearly on the side of the quantum description. An interesting development here is that one may create a "Franson state" (Franson, 1989), which is an entangled state of two particles created at different times, so that one has two particles entangled in time, as well as in space (Kwiat *et al.*, 1993).

One may also produce states where one photon has been emitted by one of two different atoms, but one does not know which atom produced it (Grangier *et al.*, 1985), and these states can interfere. One can also create two-photon states that are produced by two different lasers, where one does not know in which laser each photon is produced, and the amplitudes for both cases can interfere (Pfleegor and Mandel, 1967; see also Yurke and Stoler, 1992).

All these cases discussed verify the basic consequence of superposition, which is that interference can be seen only in situations where the interfering states are not experimentally distinguishable. Here, "distinguishable" means that the experimental arrangement cannot in principle distinguish between the two states, and *not* that the experimenter does not bother to distinguish between them, even though his apparatus is capable of doing so.

For example, in our neutron interferometer (Fig. 1), one cannot experimentally determine which path the neutron takes through the interferometer, and so the two possibilities can interfere. However, if one placed some device close to one of the beams, then one could determine whether a neutron was in that beam by detecting, say, its magnetic field as it passed. But then the interaction between the neutron and the device would necessarily destroy any phase knowledge carried by the neutron wave function, and there could be no interference between the two paths through the interferometer when the paths met again. Even if one did not bother to read the output of the disturbing device, there would be no interference. The mere possibility of determining the path in that experimental arrangement is enough to destroy the interference pattern.

Dirac, in his textbook (1958), said that a photon could only interfere with itself. That was true when it was written since different photons were generally produced incoherently. But today, with correlated multiphoton systems available, the statement is ambiguous or not relevant in many experiments.

2.3.2 The Extension to Three-Particle States The Bell theorem depends on a statistical sample of spin correlations between particles, measured at arbitrary angles. It says nothing about those special cases in which a perfect correlation exists between the spins, which is the condition for EPR elements of reality to exist (in the Bohm case, when the spin component of particle 1 is measured at 0° or 180° to particle 2). In systems of three or more particles there exists a much more stringent restriction on EPR states than that given by Bell's theorem. There one can show (the GHZ theorem) (Greenberger *et al.*, 1989, 1990, 1993; Mermin, 1990, 1993) that even when perfect correlations are present, one cannot construct a realistic, local model to explain these perfect correlations. Thus in this case the EPR assumption that there exist elements of reality is directly inconsistent with their other assumptions. It leads to predictions that are the exact opposite of those of quantum theory for every individual set of test particles, rather than the statistical predictions of Bell's theorem. (Unfortunately, no such three-particle perfectly correlated state has been constructed experimentally yet.)

3. CONCLUSIONS

Thus we see that quantum theory yields many predictions that are quite disturbing to the classically trained mind. However, most of these predictions have been tested and shown to be true experimentally. Thus the problem is not whether quantum theory works or not. It is abundantly clear that it works, and to great precision. Rather, the problem has always been in interpreting the theory. No one has come up with an interpretation of the theory that is in all ways satisfying, and that explains the process of a quantum measurement in a fully acceptable way, in spite of the experimental successes of the theory.

GLOSSARY

Aharonov–Bohm Effect: An interference effect found in regions where the electromagnetic potentials are nonzero, but the fields themselves are zero. An intrinsically nonlocal quantum effect.

Bell's Theorem: Bell showed that for a two-particle system in an entangled state, the realistic, local assumptions of Einstein, Podolsky, and Rosen lead to an upper limit on the maximum correlation possible between the particles. Quantum theory predictions exceed this upper limit, and the quantum predictions have been confirmed by experiment, ruling out a large class of such alternative theories.

Berry Phase: A system carried through a series of adiabatic changes, by varying a set of parameters, will return to its initial state when the parameters have their initial values. But the state will be multiplied by a phase factor that depends on the entire history of variation of the parameters, and is thus nonlocal.

Collapse (Reduction) of Wave Packet: The process described by the projection postulate.

Complementarity: Bohr's word for the fact that in a given experimental arrangement, one can measure either the particle aspects or the wave aspects of an electron, or other "particle," but not both simultaneously.

Delayed-Choice Experiment: An experiment in which a beam is split by a beam splitter, and a system parameter is varied down-field of the split. The beams are later recombined and interfere. Such experiments are extremely troublesome for local, realistic theories that place a particle in one beam or the other.

Density Operator: A Hilbert-space operator representing a quantum state (technically: a positive, Hermitian operator with unit trace).

Eigenstate: A state of a quantum system for which the observables in some commuting (hence, simultaneously measurable) set have definite values, in the orthodox interpretation of the theory.

Einstein–Podolsky–Rosen (EPR) Paradox: If two particles are in an entangled state, then a measurement of one can affect the state of the other instantaneously, even if they are separated. It is not a true paradox, merely highly nonintuitive behavior, if one accepts the realistic and local assumptions of EPR. But the quantum predictions are fully verified by experiment.

Entangled State: A state of a two-particle or multiparticle system that cannot be factored into a product of one-particle states. Such states can have intrinsically nonlocal correlations.

GHZ Theorem: An extension of Bell's theorem to three or more particles (by Greenberger, Horne, and Zeilinger). It shows that the Einstein–Rosen–Podolsky assumptions lead to predictions that are exactly the opposite of those of quantum theory.

Hidden Variables: Hypothetical variables, in addition to the quantum state, that specify definite values for the observables of a quantum system.

Hilbert Space: The state space of quantum mechanics, a linear vector space over the complex numbers.

Instrumental Interpretation: An interpretation of a theory as purely predictive and not necessarily descriptive of the structure of reality at any level.

Many-Worlds Interpretation: An interpretation of quantum mechanics according to which the different components of a pure quantum state, expressed as a linear superposition of eigenstates of some observable, all—in some sense—represent actualities.

Mixture: A quantum state that can be decomposed (in general, nonuniquely) into a probability distribution over pure quantum states.

Modal Interpretation: An interpretation of quantum mechanics incorporating some principle for attributing actual values to observables of a quantum system, as opposed to possible values with probabilities, when the state is not an eigenstate of these observables.

Projection Operator: An operator that projects (maps) vectors in Hilbert space onto a subspace.

Projection Postulate: The postulate describing the transition of a pure quantum state to an appropriate mixture associated with the different possible outcomes of a measurement.

Pure State: A quantum state that cannot be further decomposed into other quantum states (but is still not dispersion free for all observables).

Quantum Eraser: Normally, interactions of the system with its environment produce random phases and incoherence. However, if one does not make a measurement on the system the phases may only appear random, but still maintain coherence. If one is sufficiently clever, one may recombine the parts of the system and undo the interactions, making the coherence manifest again.

Quantum Zeno Paradox: Zeno said motion was impossible because before reaching a goal, one had to get halfway there. Before that, one had to get halfway to the midpoint, etc. The quantum version is that a continuously observed system cannot make transitions. (Normal systems are not continuously observed.)

Spin-$\frac{1}{2}$ Particle: A quantum system whose spin observables have two possible values.

Superposition Principle: The principle that quantum amplitudes add, which includes both their magnitudes and phases, and which can cause interference effects. Interfering amplitudes are responsible for most of the nonintuitive aspects of quantum theory. Probabilities depend only on the square of the magnitudes.

Unitary Evolution: The dynamical evolution of a quantum state in Hilbert space, described by Schrödinger's time-dependent equation of motion (as opposed to the evolution described by the projection postulate).

Watched-Pot Effect: Another name for the quantum Zeno paradox, after the proverb, "A watched pot never boils."

Works Cited

Aharonov, Y., Bohm, D. (1959), *Phys. Rev.* **115,** 485–491.

Alley, C. O., Jakubowicz, O. G., Wickes, W. C. (1984), in: M. Namiki, Y. Ohnuki, Y. Murayama, S. Nomura (Eds.), *Proceedings of the 2nd International Symposium: Foundations of Quantum Mechanics in the Light of New Technology,* Tokyo: Physical Society of Japan.

Allman, B. E., Cimmino, A., Klein, A. G., Opat, G. I., Kaiser, H., Werner, S. A. (1993), *Phys. Rev. A* **48,** 1799–1807.

Anandan, J. S. (1990), *Quantum Coherence,* Singapore: World Scientific.

Aspect, A., Dalibard, J., Roger, G. (1982), *Phys. Rev. Lett.* **49,** 1804–1807.

Ballentine, L. E. (1970), *Rev. Mod. Phys.* **42,** 358–391.

Bell, J. S. (1964), *Physics* **1,** 195–200. Reprinted in Bell, J. S. (1987), *Speakable and Unspeakable in Quantum Mechanics,* New York: Cambridge Univ. Press.

Berry, M. V. (1984), *Proc. R. Soc. London A* **392,** 45–57.

Bohm, D. (1951), *Quantum Theory,* Englewood Cliffs, NJ: Prentice-Hall.

Bohm, D. (1952), *Phys. Rev.* **85,** 166–179, 180–193.

Bohm, D., Bub, J. (1966), *Rev. Mod. Phys.* **38,** 453–469.

Bohm, D., Hiley, B. J. (1993), *The Undivided Universe: An Ontological Interpretation of Quantum Theory,* New York: Routledge and Kegan Paul.

Bub, J. (1992), *Found. Phys.* **22,** 737–754.

Cimmino, A., Opat, G. I., Klein, A. G., Kaiser, H., Werner, S. A., Arif, M., Clothier, R. (1989), *Phys. Rev. Lett.* **63,** 380–383.

Clauser, J., Horne, M. A., Shimony, A., Holt, R. A. (1969), *Phys. Rev. Lett.* **23,** 880–884.

Clauser, J., Freedman, S. (1972), *Phys. Rev. Lett.* **28,** 938–941.

Clauser, J., Horne, M. A. (1974), *Phys. Rev. D* **10,** 526–535.

Clauser, J., Shimony, A. (1978), *Rep. Prog. Phys.* **41,** 1881–1927.

Daneri, A., Loinger, A., Prosperi, G. M. (1962), *Nucl. Phys.* **33,** 297–319.

Dicke, R. H. (1981), *Am. J. Phys.* **49,** 925–930.

Dicke, R. H. (1986), *Found. Phys.* **16,** 107–113.

Dieks, D. (1989), *Found. Phys.* **19,** 1397–1423.

Dirac, P. A. M. (1958), *Quantum Mechanics,* 4th ed., Oxford, U.K.: Clarendon.

Einstein, A., Podolsky, B., Rosen, N. (1935), *Phys. Rev.* **47,** 777–780.

Everett, H. (1957), *Rev. Mod. Phys.* **29,** 452–462.

Franson, J. D. (1989), *Phys. Rev. Lett.* **62,** 2205–2208.

Garg, A., Mermin, N. D. (1964), *Found. Phys.* **14,** 1–39.

Garuccio, A., Selleri, F. (1980), *Found. Phys.* **10,** 209–286.

Ghirardi, G. C., Rimini, A., Weber, T. (1986), *Phys. Rev. D* **34,** 470–491.

Ghosh, R., Mandel, L. (1987), *Phys. Rev. Lett.* **59,** 1903–1905.

Grangier, P., Aspect, A., Vigue, J. (1985), *Phys. Rev. Lett.* **54,** 418–421.

Greenberger, D. M. (1983), *Rev. Mod. Phys.* **55,** 875–905.

Greenberger, D. M., Horne, M. A., Zeilinger, A. (1989), in: M. Kafatos (Ed.), *Bell's Theorem, Quantum Theory, and Conceptions of the Universe,* Dordrecht: Kluwer, pp. 73–76.

Greenberger, D. M., Horne, M. A., Shimony, A., Zeilinger, A. (1990), *Am. J. Phys.* **58,** 1131–1143.

Greenberger, D. M., Horne, M. A., Zeilinger, A. (1993), *Phys. Today* **46** (8), 22–29.

Healey, R. (1989), *The Philosophy of Quantum Mechanics: An Interactive Interpretation,* Cambridge, U.K.: Cambridge Univ. Press.

Hellmuth, T., Zajonc, A. G., Walther, H. (1986), in: D. M. Greenberger (Ed.), *New Techniques and Ideas in Quantum Measurement Theory,* Annals No. 480, New York: New York Academy of Sciences.

Itano, W. M., Heinzen, D. J., Bollinger, J. J., Wineland, D. J. (1990), *Phys. Rev. A* **41,** 2295–2300.

Joos, E., Zeh, H. D. (1985), *Z. Phys.* **B59,** 223–243.

Kochen, S. (1985), in: P. Lahti, P. Mittelstaedt (Eds.), *Symposium on the Foundations of Modern Physics,* Singapore: World Scientific, pp. 151–170.

Kwiat, P. G., Steinberg, A. M., Chiao, R. Y. (1992), *Phys. Rev. A* **45,** 7729–7739.

Kwiat, P. G., Steinberg, A. M., Chiao, R. Y. (1993), *Phys. Rev. A* **47,** 2472–2475.

Leggett, A. J. (1984), in: S. Kamefuchi, H. Ezawa, Y. Murayama, M. Namiki, S. Nomura, Y. Ohnuki, T. Yajima (Eds.), *Proceedings of the 1st International Symposium: Foundations of Quantum Mechanics in the Light of New Technology,* Tokyo: Physical Society of Japan.

Margenau, H. (1936), *Phys. Rev.* **49,** 240–242.

Mermin, N. D. (1990), *Phys. Rev. Lett.* **65,** 3373–3376.

Mermin, N. D. (1993), *Rev. Mod. Phys.* **65,** 803–815.

Misra, B., Sudarshan, E. C. G. (1977), *J. Math. Phys.* **18,** 756–753.

Olariu, S., Popescu, I. I. (1985), *Rev. Mod. Phys.* **57,** 339–436.

Paz, J. P., Zurek, W. H. (1993), *Phys. Rev. D* **48,** 2728–2738.

Peres, A. (1980), *Am. J. Phys.* **48,** 931–932.

Peshkin, M., Tonomura, A. (1989), *The Aharonov-Bohm Effect,* Lecture Notes in Physics Vol. 340, Heidelberg: Springer-Verlag.

Pfleegor, R. L., Mandel, L. (1967), *Phys. Lett. A* **24,** 766–767.

Rarity, J. G., Tapster, P. R. (1990), *Phys. Rev. Lett.* **64,** 2495–2498.

Renninger, M. (1960), *Z. Phys.* **158,** 417–421.

Schrödinger, E. (1935), *Naturwissenschaften* **23,** 807–812, 823–828, 844–849. Translation by J. D. Trimmer in: Wheeler, J. A., Zurek, W. H. (Eds.), *Quantum Theory and Measurement,* Princeton: Princeton Univ. Press, 1983.

Scully, M., Shea, R., McCullen, J. D. (1978), *Phys. Rep.* **43,** 485–508.

Scully, M., Englert, B.-G., Walther, H. (1991), *Nature* **351,** 111–116.

Selleri, F. (1990), *Quantum Paradoxes and Physical Reality,* Dordrecht, the Netherlands: Kluwer.

Shapere, A., Wilczek, F. (1989), *Geometrical Phases in Physics,* Singapore: World Scientific.

Shih, Y. H., Alley, C. O. (1984), in: M. Namiki, Y. Ohnuki, Y. Murayama, S. Nomura (Eds.), *Proceedings of the 2nd International Symposium: Foundations of Quantum Mechanics in the Light of New Technology,* Tokyo: Physical Society of Japan.

Tonomura, A. (1987), *Rev. Mod. Phys.* **59,** 639–669.

van Fraassen, B. (1981), in: E. Beltrametti, B. van Fraassen, (Eds.), *Current Issues in Quantum Logic,* New York: Plenum, pp. 229–258.

von Neumann, J. (1955), *Mathematical Foundations of Quantum Mechanics,* Princeton: Princeton Univ. Press. Translated from *Grundlagen der Quantenmechanik,* Berlin: Springer, 1932.

Webb, R. A., Washburn, S. (1988), *Phys. Today,* **41** (12), 46–53.

Wheeler, J. A. (1978), in: A. R. Marlow (Ed.), *Mathematical Foundations of Quantum Theory,* New York: Academic.

Wheeler, J. A., Zurek, W. H. (Eds.) (1983), *Quantum Theory and Measurement,* Princeton: Princeton Univ. Press.

Wootters, W. K., Zurek, W. H. (1979), *Phys. Rev. D* **19,** 473–484.

Wu, T. T., Yang, C. N. (1975), *Phys. Rev. D* **12,** 3845–3857.

Yurke, B., Stoler, D. (1992), *Phys. Rev. A* **46,** 2229–2234.

Zurek, W. H. (1993), *Prog. Theor. Phys.* **89,** 281–312.

Further Reading

General

Bell, J. S. (1987), *Speakable and Unspeakable in Quantum Mechanics,* New York: Cambridge Univ. Press.

Braginsky, V. B., Khalili, F. Ya. (1992), *Quantum Measurement,* Cambridge, U.K.: Cambridge Univ. Press.

Fine, A. (1970), *Phys. Rev. D* **2,** 2783–2787.

Silverman, M. P. (1995), *More Than One Mystery: Explorations in Quantum Interference,* New York: Springer-Verlag.

Wheeler, J. A., Zurek, W. H. (Eds.) (1983), *Quantum Theory and Measurement,* Princeton: Princeton Univ. Press.

Hidden-Variable Theorems

d'Espagnat, B. (1979), *Sci. Am.* **241** (5), 128–166.

Kochen, S., Specker, E. P. (1967), *J. Math. Mech.* **17,** 59–87.

Mermin, N. D. (1993), *Rev. Mod. Phys.* **65,** 803–815.

Bohm's Hidden-Variable Theory

Bohm, D., Hiley, B. J. (1993), *The Undivided Universe: An Ontological Interpretation of Quantum Theory,* London: Routledge.

Experimental Tests of Bell's Theorem

Clauser, J. F., Shimony, A. (1978), *Rep. Prog. Phys.* **41,** 1881–1927.

Pipkin, F. M. (1978), *Adv. At. Mol. Phys.* **14,** 281–340.

Selleri, F. (1990), *Quantum Paradoxes and Physical Reality,* Dordrecht, the Netherlands: Kluwer.

Bell's Theorem without Inequalities

Greenberger, D. M., Horne, M. A., Shimony, A., Zeilinger, A. (1990), *Am. J. Phys.* **58,** 1131–1143.

Neutron Interferometer Experiments on Superposition

Greenberger, D. M. (1983), *Rev. Mod. Phys.* **55,** 875–905.

Parametric Down-Conversion Experiments on Superposition

Greenberger, D. M., Horne, M. A., Zeilinger, A. (1993), *Phys. Today* **46** (8), 22.

Topological Effects

Peshkin, M., Tonomura, A. (1989), *The Aharonov-Bohm Effect,* Lecture Notes in Physics Vol. 340, Heidelberg: Springer-Verlag.

Shapere, A., Wilczek, F. (1989), *Geometrical Phases in Physics,* Singapore: World Scientific.

1
Quantum Mechanics

Joachim Burgdörfer and Stefan Rotter

1.1 Introduction 3
1.2 Particle-Wave Duality and the Uncertainty Principle 4
1.3 Schrödinger Equation 6
1.4 Boundary Conditions and Quantization 8
1.5 Angular Momentum in Quantum Mechanics 9
1.6 Formalism of Quantum Mechanics 12
1.7 Solution of the Schrödinger Equation 16
1.7.1 Methods for Solving the Time-Dependent Schrödinger Equation 16
1.7.1.1 Time-Independent Hamiltonian 16
1.7.1.2 Time-Dependent Hamiltonian 17
1.7.2 Methods for Solving the Time-Independent Schrödinger Equation 19
1.7.2.1 Separation of Variables 19
1.7.2.2 Variational Methods 23
1.7.3 Perturbation Theory 24
1.7.3.1 Stationary Perturbation Theory 25
1.7.3.2 Time-Dependent Perturbation Theory 25
1.8 Quantum Scattering Theory 27
1.8.1 Born Approximation 28
1.8.2 Partial-Wave Method 29
1.8.3 Resonances 30
1.9 Semiclassical Mechanics 31
1.9.1 The WKB Approximation 31
1.9.2 The EBK Quantization 33
1.9.3 Gutzwiller Trace Formula 34
1.10 Conceptual Aspects of Quantum Mechanics 35
1.10.1 Quantum Mechanics and Physical Reality 36
1.10.2 Quantum Information 38
1.10.3 Decoherence and Measurement Process 39
1.11 Relativistic Wave Equations 41
1.11.1 The Klein–Gordon Equation 41
1.11.2 The Dirac Equation 42

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

Glossary 43
References 44
Further Reading 45

1.1
Introduction

Quantum mechanics (QM), also called *wave mechanics*, is the modern theory of matter, of atoms, molecules, solids, and of the interaction of electromagnetic fields with matter. It supersedes the classical mechanics embodied in Newton's laws and contains them as a limiting case (the so-called classical limit of QM). Its development was stimulated in the early part of the twentieth century by the failure of classical mechanics and classical electrodynamics to explain important discoveries, such as the photoelectric effect, the spectral density of blackbody radiation, and the electromagnetic absorption and emission spectra of atoms. QM has proven to be an extremely successful theory, tested and confirmed to an outstanding degree of accuracy for physical systems ranging in size from subatomic particles to macroscopic samples of matter, such as superconductors and superfluids. Despite its undisputed success in the description of physical phenomena, QM has continued to raise many conceptual and philosophical questions, in part due to its counterintuitive character, defying common-sense interpretation.

One distinguishes between nonrelativistic quantum mechanics and relativistic quantum mechanics depending on the characteristic speed v of the constituents of the system under study. The domain of nonrelativistic QM is characterized by speeds that are small compared with the speed of light:

$$v \ll c \tag{1}$$

where $c = 3 \times 10^8\,\mathrm{m\,s^{-1}}$ is the speed of light. For v approaching c, relativistic extensions are necessary. They are referred to as *relativistic wave equations*. Because of the possibility of production and destruction of particles by conversion of relativistic energy E into mass M (or vice versa) according to the relation (*see* chapter 2)

$$E = Mc^2 \tag{2}$$

a satisfactory formulation of relativistic quantum mechanics can only be developed within the framework of quantum field theory. This chapter focuses primarily on nonrelativistic quantum mechanics, its formalism and techniques as well as applications to atomic, molecular, optical, and condensed-matter physics. A brief discussion of extensions to relativistic wave equations is given at the end of the chapter.

The new "quantum world" will be contrasted with the old world of classical mechanics. The bridge between the latter and the former is provided by the

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

semiclassical mechanics, whose rigorous formulation and application have only recently been developed.

1.2
Particle-Wave Duality and the Uncertainty Principle

The starting point of modern quantum mechanics is the hypothesis of the particle-wave duality put forward by de Broglie (1924): the motion of each particle of mass M is associated with a wave (''matter wave'') of wavelength

$$\lambda = \frac{h}{p} \tag{3a}$$

where $h = 6.63 \times 10^{-34}$ J s is Planck's constant, discovered by M. Planck in 1900 in his investigation of blackbody radiation, and $p = Mv$ is the classical momentum of the particle. Similarly, the frequency ν of the wave is associated with the mechanical energy of the particle through the expression

$$\nu = \frac{E}{h} \tag{3b}$$

These relations, ascribing wave properties to particles, are the logical complements to postulates put forward 20 years earlier by Einstein (1905), assigning the characteristics of particles (''photons,'' particles of zero rest mass) to electromagnetic radiation. In fact, Einstein employed, in his explanation of the ''photoelectric effect,'' the expression $E = h\nu$ (Eq. 3b) to find the energy E of each photon associated with electromagnetic radiation of frequency ν. Photons can eject electrons from the atoms of a material provided that the energy exceeds the binding energy of the electron. This particle-wave duality of QM unifies the description of distinct entities of classical physics: mechanics and electromagnetism. Particles, the fundamental constituents of matter, possess, in addition to mechanical properties such as momentum and angular momentum, wave character. Conversely, electromagnetic radiation, successfully described by Maxwell's theory of electromagnetic waves, takes on, under appropriate conditions, the property of quasiparticles, the photons, carrying mechanical energy and momentum. The particle and wave character are complementary to each other. An object behaves either like a particle or like a wave.

The hypothesis of particle-wave duality has been experimentally verified in remarkable detail. Interference of matter waves due to the indistinguishability of paths was demonstrated first for electrons (Davisson and Germer, 1927), later for neutrons, and, most recently, for neutral atoms. The particle-wave duality of electromagnetic radiation and its complementarity have been explored through modern versions of Young's double-slit experiment using state-of-the-art laser technology and fast switches. Hellmuth and colleagues (1987) succeeded, for example, in switching from particle to wave character and vice versa even after the photon has passed through the slit (in the experiment, a beam splitter) by activating a switch (a Pockels cell) in one of the two optical paths of the interferometer. By use of the switch, the interference patterns are made to appear or disappear, as expected for a wavelike or particlelike behavior of the photon, respectively. This ''delayed choice'' experiment not only demonstrates the particle-wave duality for photons but also elucidates important conceptual aspects of the nonlocality of quantum mechanics. The decision as to whether the photon will behave like a particle or a wave need not be made at the double slit itself but can be spatially separated and temporarily postponed.

The consequences of the wave-particle duality for the description of physical phenomena are far reaching and multifaceted. One immediate consequence that represents the most radical departure from classical physics is the *uncertainty principle* (Heisenberg, 1927). One deeply rooted notion of classical physics is that all dynamical observables can be measured, at least in principle, with arbitrary accuracy. Uncertainties in the measurement are a matter of experimental imperfection that, in principle, can be overcome. The wave nature of particles in QM imposes, however, fundamental limitations on the simultaneous accuracy of measurements of dynamical variables that cannot be overcome, no matter how much the measurement process can be improved. This uncertainty principle is an immediate consequence of the wave behavior of matter. Consider a wave of the form

$$\psi(x) = Ae^{i(kx-\omega t)} \tag{4}$$

traveling along the x coordinate (Figure 1.1). Equation (4) is called a *plane wave* because the wave fronts form planes perpendicular to the direction of propagation. The physical interpretation of the amplitude A was provided by Born (1926) in terms of a probability density

$$P(x) = |\psi(x)|^2 = |A|^2 \tag{5}$$

for finding the particle at the coordinate x. The matter wave (Eq. 4) is further characterized by the wave number $k = 2\pi/\lambda = p/\hbar$ and the angular frequency $\omega = 2\pi\nu = E/\hbar$. The "rationalized" Planck's constant $\hbar = h/2\pi = 1.05 \times 10^{-34}$ J s is frequently used in QM.

The uncertainty principle is inherent to all wave phenomena. Piano tuners have exploited it for centuries. They sound a vibrating tuning fork of standard frequency in unison with a piano note of the same nominal frequency and listen to a beat tone between the struck tune and the tuning fork. For a fork frequency of $\nu = 440$ Hz and a string frequency of $\nu' = 441$ Hz, one beat tone will be heard per second ($\delta\nu = \nu' - \nu = 1$ Hz). The goal of the tuner is to reduce the number of beats as much as possible. To achieve an accuracy of $\delta\nu = 0.01$ Hz, the tuner has to wait for at least about $\delta t = 100$ s (in fact, only a fraction of this time since the frequency mismatch can be detected prior to completion of one beat period) to be sure no beat had

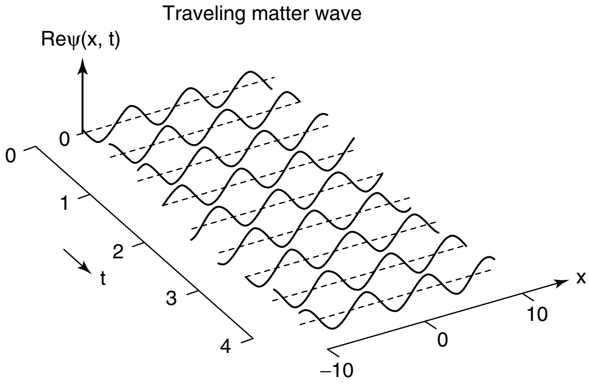

Fig. 1.1 Matter wave traveling along the x direction; the figure shows the real part of $\psi(x)$, Eq. (4) (after Brandt and Dahmen, 1989).

occurred. Piano tuning therefore relies on the frequency–time uncertainty

$$\Delta\nu\,\Delta t \gtrsim 1 \tag{6}$$

Measurement of the frequency with infinite accuracy ($\Delta\nu \to 0$) requires an infinite time period of measurement ($\Delta t \to \infty$), or equivalently, it is impossible within any finite period of time to determine the frequency of the string exactly since time and frequency are *complementary variables*. Similarly, the variables characterizing the matter wave (Eq. 4) are complementary:

$$\begin{aligned}\Delta k\,\Delta x &\geq 1\\ \Delta\omega\,\Delta t &\geq 1\end{aligned} \tag{7}$$

These relations, well known from classical theory of waves and vibrations, have unexpected and profound consequences when combined with the de Broglie hypothesis (Eq. 3) for quantum objects:

$$\begin{aligned}\Delta p\,\Delta x &\geq \hbar\\ \Delta E\,\Delta t &\geq \hbar\end{aligned} \tag{8}$$

The momentum and the position of a particle, its energy and the time, and more generally, any other pair of complementary variables, such as angular momentum and angle, are no longer simultaneously measurable with infinite precision. The challenge to the classical laws of physics becomes obvious when one contemplates the fact that Newton's equation of motion uses the simultaneous knowledge of the position and of the change of momentum.

Since $\hbar$ is extremely small, the uncertainty principle went unnoticed during the era of classical physics for hundreds of years and is of little consequence for everyday life. Its far-reaching implications become obvious only in the microscopic world of atoms, molecules, electrons, nuclei, and elementary particles. More generally, whenever the size of an object or mean distance between constituents becomes comparable to the de Broglie wavelength λ, the wave nature of matter comes into play. The size of λ also provides the key for quantum effects in macroscopic systems such as superfluids, superconductors, and solids.

1.3
Schrödinger Equation

The Schrödinger equation for matter waves ψ describes the dynamics of quantum particles. In the realm of quantum physics, it plays a role of similar importance to that which Newton's equation of motion does for classical particles. As with Newton's laws, the Schrödinger equation cannot be rigorously derived from some underlying, more fundamental principles. Its form can be made plausible, however, by combining the Hamiltonian function of classical mechanics,

$$H = T + V = E \tag{9}$$

which equals the mechanical energy, with the de Broglie hypothesis of matter waves (Eq. 4). The potential energy is denoted by V and the kinetic energy by $T = p^2/2M$. Formally multiplying the wave function $\psi(x,t)$ by Eq. (9) yields

$$H\psi(x,t) = \frac{p^2}{2M}\psi(x,t) + V\psi(x,t) \tag{10}$$

In order to connect the de Broglie relations for energy and momentum appearing in the arguments of the plane wave (Eq. 4) with the energy and momentum in the

Hamiltonian function, H and p in Eq. (10) must be taken as differential operators,

$$\begin{aligned} H &\longleftrightarrow i\hbar\frac{\partial}{\partial t} \\ p &\longleftrightarrow \frac{\hbar}{i}\frac{\partial}{\partial x} \end{aligned} \tag{11}$$

This substitution into Eq. (10) leads to the time-dependent Schrödinger equation

$$\begin{aligned} i\hbar\frac{\partial}{\partial t}\psi(x,t) &= H\psi(x,t) \\ &= \left(-\frac{\hbar^2}{2M}\frac{\partial^2}{\partial x^2} + V(x)\right) \\ &\quad\times \psi(x,t) \end{aligned} \tag{12}$$

which describes the time evolution of the matter wave. For physical systems that are not explicitly time dependent, i.e., that have time-independent potentials $V(x)$, the energy is conserved and $i\hbar\ \partial/\partial t$ can be replaced by E, giving the time-independent Schrödinger equation

$$E\psi(x) = \left(-\frac{\hbar^2}{2M}\frac{\partial^2}{\partial x^2} + V(x)\right)\psi(x) \tag{13}$$

If no interaction potential is present $(V(x) = 0)$, Eq. (12) is referred to as the *free-particle Schrödinger equation*. The plane waves of Eq. (4) are solutions of the free-particle Schrödinger equation with eigenenergies $E = (\hbar k)^2/2M$.

The *correspondence principle*, first formulated by Bohr, provides a guide to quantizing a mechanical system. It states that in the limit of large quantum numbers or small de Broglie wavelength, the quantum-mechanical result should be the same, or nearly the same, as the classical result. One consequence of this postulate is that the quantum-mechanical Hamiltonian operator in the Schrödinger equation (Eq. 12) should be derived from its corresponding classical analog, the classical Hamiltonian function (Eq. 9), through the replacement of the classical observable by the operator equivalents (Eq. 11).

The replacement of dynamical variables by differential operators leads to the *noncommutativity* of pairs of conjugate variables in QM, for example,

$$[x, p_x] \equiv x\frac{\hbar}{i}\frac{\partial}{\partial x} - \frac{\hbar}{i}\frac{\partial}{\partial x}x = i\hbar \tag{14}$$

with similar relations for other complementary pairs, such as Cartesian coordinates of position and momentum $(y,\ p_y)$, (z, p_z), or energy and time. It is the noncommutativity that provides the basis for a formal proof of the uncertainty principle, Eq. (8). The Schrödinger equation (Eqs. 12 and 13) is a linear, homogeneous, and (in general) partial differential equation. One consequence of this fact is that matter waves ψ satisfy the *superposition principle*. If ψ_1 and ψ_2 are two acceptable solutions of the Schrödinger equation, called *states*, a linear combination

$$\psi = c_1\psi_1 + c_2\psi_2 \tag{15}$$

is an acceptable solution, as well. The physical implication of Eq. (15) is that the quantum particle can be in a ''coherent superposition of states'' or ''coherent state,'' a fact that leads to many strange features of QM, such as destructive and constructive interference, the collapse of the wave function, and the revival of wave packets.

A solution of a linear, homogeneous differential equation is determined only up to an overall amplitude factor, which is arbitrary. The probability interpretation of the matter wave (Eq. 5), however, removes this arbitrariness. The probability for finding the particle somewhere is unity,

$$\int dx P(x) = \int dx |\psi(x)|^2 = 1 \tag{16}$$

Equation (16) and its generalization to three dimensions are referred to as the *normalization condition* on the wave function and eliminate the arbitrariness of the modulus of the wave function. A wave function that can be normalized according to Eq. (16) is called *square integrable*. The only arbitrariness remaining is the overall "phase factor" of the wave function, $e^{i\phi}$, i.e., a complex number of modulus 1.

1.4
Boundary Conditions and Quantization

In order to constitute a physically acceptable solution of the Schrödinger equation, i.e., to represent a state, ψ must satisfy appropriate boundary conditions. While their detailed forms depend on the coordinates and symmetry of the problem at hand, their choice is always dictated by the requirement that the wave function be normalizable and unique and that physical observables take on only real values (more precisely, real expectation values) despite the fact that the wave function itself can be complex. Take, as an example, the Schrödinger equation for a rotator with a moment of inertia I, constrained to rotate about the z axis. The classical Hamiltonian function $H = L_z^2/2I$, with L_z the z component of the angular momentum,

$$L_z = xp_y - yp_x = \frac{\hbar}{i}\frac{\partial}{\partial\phi} \tag{17}$$

gives rise, according to the correspondence principle, to the following time-independent Schrödinger equation:

$$-\frac{\hbar^2}{2I}\frac{\partial}{\partial\phi^2}\psi(\phi) = E\psi(\phi) \tag{18}$$

where ϕ is the angle of rotation about the z axis. Among all solutions of the form

$$\psi(\phi) = Ae^{i(m\phi+\phi_0)} \tag{19}$$

only those satisfying the periodic boundary conditions

$$\psi(\phi + 2\pi) = \psi(\phi) \tag{20}$$

are admissible in order to yield a uniquely defined single-valued function $\psi(\phi)$. (The implicit assumption is that the rotator is spinless, as discussed below.) The magnetic quantum number m therefore takes only integer values $m = -\infty, \dots, -1, 0, 1, \dots \infty$. Accordingly, the z component of the angular momentum

$$L_z\psi(\phi) = m\hbar\psi(\phi) \tag{21}$$

takes on only discrete values of multiples of the rationalized Planck's constant $\hbar$. The real number on the right-hand side that results from the operation of a differential operator on the wave function is called the *eigenvalue* of the operator. The wave functions satisfying equations that are of the form of Eqs (18) or (21) are called *eigenfunctions*. The eigenvalues and eigenfunctions contain all relevant information about the physical system; in particular, the eigenvalues represent the measurable quantities of the systems.

The amplitude of the eigenfunction follows from the normalization condition

$$\int_0^{2\pi} d\phi |\psi(\phi)|^2 = 1 \tag{22}$$

as $A = (2\pi)^{-1/2}$. The eigenvalues of E in Eq. (18) are $E_m = \hbar^2 m^2/2I$. Just like for a vibrating string, it is the imposition of boundary conditions (in this case, periodic boundary conditions) that leads to "quantization," that is, the selection of an infinite but discrete set of modes of matter waves, the eigenstates of matter, each of which is characterized by certain quantum numbers.

Different eigenstates may have the same energy eigenvalues. One speaks then of *degeneracy*. In the present case, eigenstates with positive and negative quantum numbers $\pm m$ of L_z are degenerate since the Hamiltonian depends only on the square of L_z. Degeneracies reflect the underlying symmetries of the Hamiltonian physical system. Reflection of coordinates at the x–z plane, ($y \rightarrow -y$, $p_y \rightarrow -p_y$ and therefore $L_z \rightarrow -L_z$) is a symmetry operation, i.e., it leaves the Hamiltonian invariant. In more general cases, the underlying symmetry may be dynamical rather than geometric in origin.

1.5
Angular Momentum in Quantum Mechanics

The quantization of the projection of the angular momentum, L_z, is called *directional quantization*. The Hamiltonian function for the interaction of an atom having a magnetic moment $\boldsymbol{\mu} = \gamma \mathbf{L}$ with an external magnetic field $\mathbf{B}$ reads

$$H = -\boldsymbol{\mu} \cdot \mathbf{B} = -\gamma \mathbf{L} \cdot \mathbf{B} \tag{23}$$

The proportionality constant γ between the magnetic moment and the angular momentum is called the *gyromagnetic ratio*. According to the correspondence principle, the Schrödinger equation for the magnetic moment in a field with field lines oriented along the z axis becomes

$$-\gamma B L_z \psi(\phi) = E_m \psi(\phi) \tag{24}$$

The eigenenergies

$$E_m = -\gamma B \hbar m \tag{25}$$

are quantized and depend linearly on the magnetic quantum number. Depending on the sign of γ, states with positive m for $\gamma > 0$, or negative m for $\gamma < 0$, correspond to states of lower energy and are preferentially occupied when the atomic magnetic moment is in thermal equilibrium. This leads to an alignment of the magnetic moments of the atoms along one particular direction and to paramagnetism of matter when its constituents carry a net magnetic moment.

The force an atom experiences as it passes through an inhomogeneous magnetic field depends on the gradient of the magnetic field, ∇B. Because of the directional quantization of the angular momentum, a beam of particles passing through an inhomogeneous magnetic field will be deflected into a set of discrete directions determined by the quantum number m. A deflection pattern consisting of a few spots rather than a continuous distribution of a beam of silver atoms was observed by Stern and Gerlach (1921, 1922) and provided the first direct evidence of directional quantization.

In addition to the projection of the angular momentum along one particular axis, for example, the z axis, the total angular momentum $\mathbf{L}$ plays an important role in QM. Here, one encounters a conceptual difficulty, the noncommutativity of different angular momentum components.

Application of Eq. (14) leads to

$$\begin{aligned}&\left[L_x, L_y\right] = i\hbar L_z, \quad \left[L_y, L_z\right] = i\hbar L_x\\&\left[L_z, L_x\right] = i\hbar L_y\end{aligned}\tag{26}$$

As noncommutativity is the formal expression of the uncertainty principle, the implication of Eq. (26) is that different components of the angular momentum cannot be simultaneously determined with arbitrary accuracy. The maximum number of independent quantities formed by components of the angular momentum that have simultaneously ''sharply'' defined eigenvalues is two. Specifically,

$$[L^2, L_z] = 0\tag{27}$$

with $L^2 = L_x^2 + L_y^2 + L_z^2$ being the square of the total angular momentum. Therefore, one projection of the angular momentum vector (conventionally, one chooses the z projection) and the L^2 operator can have simultaneously well-defined eigenvalues, unaffected by the uncertainty principle. The differential operator of L^2 in spherical coordinates gives rise to the following eigenvalue equation:

$$\begin{aligned}L^2 Y(\theta, \phi) &= -\hbar^2 \left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right] Y(\theta, \phi)\\&= \hbar^2 l(l+1) Y(\theta, \phi)\end{aligned}\tag{28}$$

The solutions of Eq. (28), $Y_l^m(\theta, \psi)$, are called *spherical harmonics*. The superscript denotes the magnetic quantum number of the z component, as above, while the quantum number $l = 0, 1, \dots \infty$ is called the *angular momentum quantum number*. Since $|L_z| \leq |L|$, we have the restriction $|m| \leq l$, that is, for given angular momentum quantum number l, the z component varies between $-l$ and l: $m = -l, \dots, -1, 0, 1, \dots l$. The total number of quantized projections of L_z is $2l+1$. The eigenvalue of the L^2 operator is, however, $\hbar^2 l(l+1)$ and not $\hbar^2 l^2$. The latter can be understood in terms of the uncertainty principle and can be illustrated with the help of a vector model (Figure 1.2). Because of the uncertainty principle, the projections L_y and L_x do not take a well-defined value and cannot be made to be exactly zero, when L_z possesses a well-defined quantum number. Therefore, the z projection $m\hbar$ will always be smaller than the magnitude of $L = \hbar\sqrt{l(l+1)}$ in order to accommodate the nonzero fluctuations in the remaining components.

The spherical harmonics are given in terms of associated Legendre functions $P_l^m(\cos\theta)$ by

$$\begin{aligned}Y_l^m(\theta, \phi) = (-1)^m \left[\frac{(2l+1)}{4\pi}\;\frac{(l-m)!}{(l+m)!}\right]^{1/2}\\ \times P_l^m(\cos\theta) e^{im\phi}\end{aligned}\tag{29}$$

The eigenfunctions of L_z (Eq. 19) can be recognized as one factor in the expression for Y_l^m. The probability densities $|Y_l^m|^2$ of the first few eigenfunctions are represented by the polar plots in Figure 1.3.

In addition to the orbital angular momentum operator **L**, which is the quantum-mechanical counterpart to the classical angular momentum, another type of angular momentum, the spin angular momentum, plays an important role for which, however, no classical analog exists. The spin is referred to as an *internal degree of freedom*. The Stern–Gerlach experiment discussed above provided the major clue for the existence of the electron spin (Goudsmit and Uhlenbeck, 1925). The orbital angular momentum quantization allows only for an odd number of discrete values of L_z ($2l+1 = 1,\ 3,\ 5, \dots$) and therefore an odd number of spots in

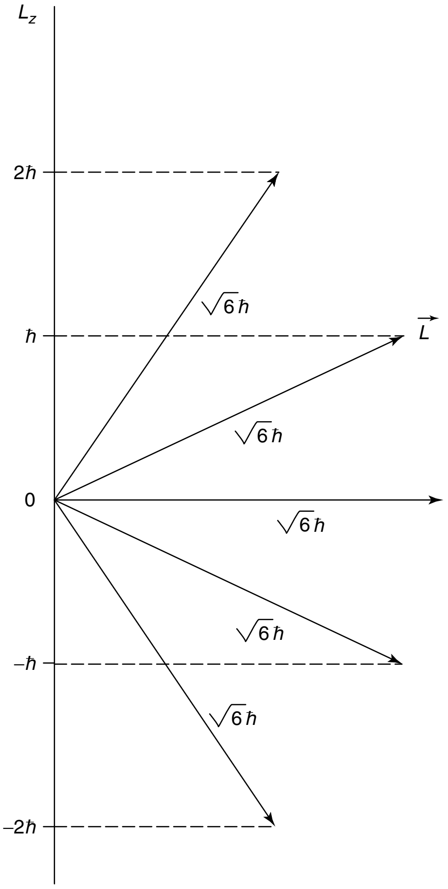

Fig. 1.2 Vector model of angular momentum **L** for quantum number $l = 2$ and $|\mathbf{L}| = \sqrt{6}\,\hbar$.

the deflection pattern. Stern and Gerlach observed, however, two discrete directions of deflection (see schematic illustration in Figure 1.14). The spin hypothesis solved this mystery. The outermost electron of a silver atom has a total orbital angular momentum of zero but has a spin angular momentum of $s = \frac{1}{2}$. The number of different projections of the spin is therefore $2s + 1 = 2$ in accordance with the number of components into which the beam was magnetically split. Within nonrelativistic quantum mechanics, the spin must be introduced as an additional degree of freedom on empirical grounds. A conceptually satisfactory description is provided by relativistic quantum mechanics (see Section 1.11). The identification of the spin as an angular momentum rests on the validity of the commutation rules (Eq. 26),

$$\left[S_x, S_y\right] = i\hbar S_z, \quad \left[S_y, S_z\right] = i\hbar S_x, \\ \left[S_z, S_x\right] = i\hbar S_y \qquad (30)$$

identical to those of the orbital angular momentum. Any set of the components of a vector operator satisfying Eq. (26) has a spectrum of eigenvalues (quantum numbers) of an angular momentum with quantum numbers $j = 0, \frac{1}{2}, 1, \ldots$ and $m = -j, -j + 1 \ldots, +j$. The quantum number j is an integer in systems with an even

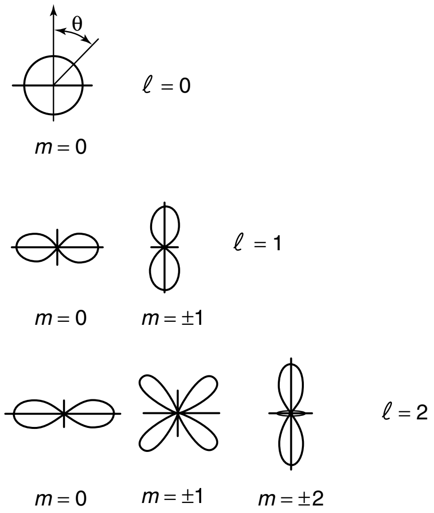

Fig. 1.3 Polar plots of the probability distributions $|Y_l^m(\theta, \phi)|^2$ of angular momentum eigenfunctions.

number of spin $\frac{1}{2}$ particles (including zero) and half-integer in systems with an odd number of spin $\frac{1}{2}$ particles. The Hamiltonian for the interaction of an electron spin with an external magnetic field is given in direct analogy to Eq. (23) by

$$H = -\gamma g \mathbf{S} \cdot \mathbf{B} \tag{31}$$

The only difference is the "anomalous" gyromagnetic ratio represented by the factor g. An approximate value of g can be deduced from the relativistic wave equation (Dirac's equation) for the electron (see Section 1.11).

1.6
Formalism of Quantum Mechanics

The quantum-mechanical description outlined in the previous section with the help of specific examples can be expanded to a general formalism of quantum mechanics. The starting point is the identification of wave functions with vectors in an infinite-dimensional vector space, the Hilbert space,

$$\psi \longrightarrow |\psi\rangle \tag{32}$$

with the vectors represented by a "ket." The ket notation was first introduced by Dirac (1930). The coherent superposition illustrated by Eq. (15) can be recognized as vector addition in Hilbert space. Physical observables such as the Hamiltonian or the angular momentum can be represented by linear and Hermitian operators. Linearity of an operator B means

$$B(c_1|\psi_1\rangle + c_2|\psi_2\rangle) = c_1 B|\psi_1\rangle + c_2 B|\psi_2\rangle \tag{33}$$

for all complex numbers c_1, c_2 and all vectors $|\psi_1\rangle, |\psi_2\rangle$. Hermitian operators are defined as those for which

$$B^\dagger = B \tag{34}$$

that is, the operator is equal to its Hermitian adjoint. The Hermiticity assures that the eigenvalue equation

$$B|\psi\rangle = \lambda|\psi\rangle \tag{35}$$

possesses real eigenvalues λ, which can be identified with measurable quantities. The time-independent Schrödinger equation, Eq. (13), is of the form in Eq. (35) with B being the Hamiltonian operator H and λ being the eigenenergy E. The result of a measurement of a physical observable yields an eigenvalue λ of the corresponding operator, B.

The noncommutativity of operators (Eq. 14) corresponds to the well-known noncommutativity of matrices in linear algebra. The overlap of wave functions can be identified with the scalar product in Hilbert space:

$$\langle\phi|\psi\rangle \equiv \int d^3r\ \phi^*(\mathbf{r})\psi(\mathbf{r}) \tag{36}$$

The adjoint of a vector in Hilbert space $\langle\psi|$ is called a *bra* and the scalar product a *bracket* (Dirac, 1930). The normalization requirement imposed on the wave function can be expressed in terms of the scalar product as $\langle\psi \mid \psi\rangle = 1$. The requirement of normalizability (Eq. 22) implies that the Hilbert space is spanned by square-integrable functions. Since eigenstates of Hermitian operators belonging to different eigenvalues are orthogonal to each other, each physical dynamical variable possesses a complete orthonormal set of eigenstates:

$$\langle\psi_i|\psi_j\rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{37}$$

The operator B is a diagonal matrix in a basis consisting of its eigenstates, the "eigenbasis" $|\psi_i\rangle$ with $i = 1, \ldots, \infty$. In its eigenbasis, the operator B possesses the following "spectral representation":

$$B = \sum_i \lambda_i|\psi_i\rangle\langle\psi_i| \tag{38}$$

containing only its eigenstates $\{\psi_i\}$ and its eigenvalues λ_i. Two operators possess the same eigenbasis when they commute with each other. Physical observables whose operators commute with each other are called *compatible observables*. A state of a physical system is therefore specified to the maximum extent possible, consistent with the uncertainty principle, by one common eigenstate of a maximum set of compatible observables. For example, the set $\{L^2, L_z\}$ is the maximum set of compatible observables for the angular momentum with the common eigenstate $|lm\rangle$.

The probability amplitude for finding the system in a particular state $|\psi_1\rangle$ is given by the projection of the state vector onto this state, i.e., by the scalar product $\langle\psi_1 \mid \psi\rangle$. The corresponding probability is

$$P(\psi_1) = |\langle\psi_1|\psi\rangle|^2 \tag{39}$$

Formally, Eq. (39) can be viewed as the expectation value of the projection operator $\overline{P} = |\psi_1\rangle\langle\psi_1|$ in the state $|\psi\rangle$. The process of a measurement is therefore associated with a projection in Hilbert space. This projection is sometimes referred to as the *collapse of the wave function*. The process of measurement thus influences the quantum system to be measured.

The state of a physical system is often not completely specified by a "pure" state described by a single Dirac ket $|\psi\rangle$ or a coherent superposition of kets $\sim c_1 \mid \psi_1\rangle + c_2 \mid \psi_2\rangle$. The limited information available may only allow us to specify a statistical mixture of states. Such a mixture is described by the statistical operator or density operator

$$\rho = \sum_i P_i|\psi_i\rangle\langle\psi_i| \tag{40}$$

with P_i denoting the probability for the system to be found in a particular state $|\psi_i\rangle$. The result of the measurement of an observable B is then given by the statistical expectation value

$$\langle B\rangle = \mathrm{Tr}(\rho B) = \sum_i \langle\psi_i|B|\psi_i\rangle P_i \tag{41}$$

i.e., by the diagonal elements of the operator B, each weighted with the probability to find the system in the corresponding state. The preferred set of states, $|\psi_i\rangle$, for which the density matrix is "diagonal," that is, of the form in Eq. (40), is sometimes referred to as the *set of pointer states.*

In light of the expression for the probability as a projection, Eq. (39), we can rewrite Eq. (5) as

$$P(\mathbf{r}) = |\psi(\mathbf{r})|^2 = |\langle\mathbf{r}|\psi\rangle|^2 \tag{42}$$

where $|\mathbf{r}\rangle$ is formally treated as a vector ("ket") in Hilbert space. This definition leads to the difficulty that $|\mathbf{r}\rangle$ cannot be defined as a square-integrable function. Nevertheless, it can be consistently incorporated into the formalism by employing the concept of generalized functions ("distributions"; see Lighthill, 1958; Schwartz, 1965). One defines a scalar product through

$$\begin{aligned}\langle\mathbf{r}|\mathbf{r}'\rangle &= \delta(\mathbf{r}-\mathbf{r}')\\ &= \delta(x-x')\delta(y-y')\delta(z-z')\end{aligned} \tag{43}$$

where the right-hand side is called *Dirac's δ function*. The δ function is zero almost everywhere:

$$\delta(x) = \begin{cases} 0, & x \neq 0 \\ \infty, & x = 0 \end{cases} \tag{44}$$

with the strength of the singularity such that the area underneath the peak is

$$\int dx\,\delta(x) = 1 \tag{45}$$

While the δ function is not an ordinary (Riemann integrable) function, it can be approximated as a limit of Riemann integrable functions, for example,

$$\lim_{\varepsilon\longrightarrow 0}\frac{1}{\pi}\frac{\varepsilon}{x^2+\varepsilon^2} = \delta(x) \tag{46}$$

Equation (46) satisfies Eq. (45) provided one evaluates the integral first and takes the limit $\varepsilon \to 0$ afterward. As the order of the operations of integration and taking the limit is not interchangeable, Eq. (46) is said to converge "weakly" toward a δ function in the limit $\varepsilon \to 0$. δ functions play an important role in the quantum theory of scattering.

The time evolution of the Hilbert-space vectors is given by Schrödinger's equation written in Dirac's ket notation:

$$i\hbar\frac{\partial}{\partial t}|\psi\rangle = H|\psi\rangle \tag{47}$$

For systems with a classical analog, the Hamiltonian operator is given, according to the correspondence principle, by the classical Hamiltonian function $H(\mathbf{r}, \mathbf{p})$ upon treating the canonical variables $(\mathbf{r}, \mathbf{p})$ as noncommuting operators (see Eq. 14). There are, however, additional dynamical variables in quantum mechanics, such as spin, for which no classical counterparts exist.

A vector in Hilbert space can be expanded in terms of any basis (a "representation"),

$$|\psi\rangle = \sum_i c_i|\phi_i\rangle = \sum_i \langle\phi_i|\psi\rangle|\phi_i\rangle \tag{48}$$

where $c_i = \langle\phi_i \mid \psi\rangle$ are the expansion coefficients of the ket $|\psi\rangle$ in the basis

$\{|\phi_i\rangle\}$. Switching from one representation to another $|\psi\rangle \to |\psi'\rangle = U|\psi\rangle$ amounts to a unitary transformation, U, since the norm of a state vector must be conserved,

$$\begin{aligned}\langle\psi'|\psi'\rangle &= \langle\psi|U^\dagger U|\psi\rangle \\ &= \langle\psi|\psi\rangle\end{aligned} \tag{49}$$

or $U^\dagger U = 1$. Transformations among representations play an important role in QM. While the maximum number of compatible observables is determined by commutation rules, the choice of observables that are taken to be diagonal or to have "good" quantum numbers depends on the specific problem under consideration. Unitary transformations provide the tool to switch from a basis within which one set of observables is diagonal to another. The transformation law for state vectors $|\psi'\rangle = U|\psi\rangle$ implies the transformation law for operators

$$A' = UAU^\dagger \tag{50}$$

The latter follows from the invariance of the expectation value $\langle\psi'|A'|\psi'\rangle = \langle\psi|U^\dagger A' U|\psi\rangle$ under unitary transformations.

Classical Hamiltonian functions as well as quantum-mechanical Hamiltonian operators possess discrete symmetries. For example, the Hamiltonian function of the harmonic oscillator,

$$H(x,p) = \frac{p^2}{2M} + \tfrac{1}{2}M\omega^2x^2 \tag{51}$$

is invariant under reflection of all spatial coordinates ($x \to -x$, $p \to -p$) since H depends only on the square of each canonical variable. In classical mechanics, this symmetry is of little consequence since the solution of Newton's equation of motion depends on the initial conditions $x(t = t_0)$, $p(t = t_0)$, which are not subject to any symmetry constraint. They may "break" the symmetry and usually do so. In quantum mechanics, however, the parity of a state vector becomes a dynamical observable.

We can define a parity operator $\hat{p}$ through

$$\hat{p}H(x,p)\hat{p}^\dagger = H(-x,-p) \tag{52}$$

Since $\hat{p}^2$ is the unit operator (two successive reflections restore the original function), we have $\hat{p} = \hat{p}^\dagger = \hat{p}^{-1}$. Eigenfunctions of $\hat{p}$, $\psi(x) = \langle x|\psi\rangle$, can now be characterized by their parity:

$$\langle x|\hat{p}|\psi\rangle = \langle -x|\psi\rangle = \pm\langle x|\psi\rangle \tag{53}$$

The ± sign in Eq. (53) is referred to as *indicating positive* or *negative parity*. A wave function cannot always be assigned a well-defined parity quantum number ±1. However, when $\hat{p}$ commutes with H,

$$\hat{p}H(x,p)\hat{p} = H(x,p) \tag{54}$$

there exists a common eigenbasis in which both the Hamiltonian operator and the parity operator are diagonal and, consequently, ψ is an eigenstate of $\hat{p}$ with a well-defined parity. A similar argument can be developed for symmetry with respect to the time-reversal operator T: $t \to -t$, $\mathbf{p} \to -\mathbf{p}$, $\mathbf{r} \to \mathbf{r}$.

An additional discrete symmetry, which is of fundamental importance for the quantum mechanics of identical particles, is the permutation symmetry. Consider a wave function describing the state of two identical particles $\psi(\mathbf{r}_1, \mathbf{r}_2)$. The permutation operator $\hat{p}_{12}$ exchanges the two particles in the wave function:

$$\hat{p}_{12}\psi(\mathbf{r}_1,\mathbf{r}_2) = \psi(\mathbf{r}_2,\mathbf{r}_1) \tag{55}$$

The spin-statistics postulate of quantum mechanics makes a general statement concerning the permutation symmetry. A many-body system consisting of identical particles of integer spin ("bosons") possesses wave functions that are even under all interchanges of two particles, whereas the wave function is odd for half-integer spin particles ("fermions"). The eigenvalue of $\hat{p}_{ij}$, the operator of permutation for any two particles of an N-body system, is $+1$ for bosons and -1 for fermions. Wave functions are totally symmetric under particle exchange for bosons and totally antisymmetric for fermions. An immediate consequence is that two fermions cannot be in identically the same state for which all quantum numbers agree (the so-called Pauli exclusion principle) since the antisymmetric wave function would vanish identically.

1.7
Solution of the Schrödinger Equation

1.7.1
Methods for Solving the Time-Dependent Schrödinger Equation

1.7.1.1 Time-Independent Hamiltonian
If the Hamilton operator is not explicitly time dependent, i.e., the total energy is conserved, the time-dependent Schrödinger equation can be reduced to the time-independent Schrödinger equation. Equation (47) can be formally solved by

$$|\psi(t)\rangle = U(t,t_0)|\psi(t_0)\rangle \tag{56}$$

where $U(t, t_0)$ is called the *time-evolution operator*, a unitary transformation describing the translation of the system in time. Substitution of Eq. (56) into Eq. (47) shows that

$$U(t,t_0) = \exp\left[\frac{-iH(t-t_0)}{\hbar}\right] \tag{57}$$

where the exponential function of the operator H is defined through its power-series expansion

$$\begin{aligned} U(t,t_0) = 1 - i\frac{t-t_0}{\hbar}H \\ -\frac{1}{2}\left(\frac{t-t_0}{\hbar}\right)^2 H\cdot H + \cdots \end{aligned} \tag{58}$$

If one is able to solve the time-independent Schrödinger equation

$$H|\psi_n\rangle = E_n|\psi_n\rangle \tag{59}$$

where n labels the eigenstate, one can calculate the time evolution of the eigenstate:

$$\begin{aligned} U(t,t_0)|\psi_n\rangle &= e^{-iH(t-t_0)/\hbar}|\psi_n\rangle \\ &= e^{-iE_n(t-t_0)/\hbar}|\psi_n\rangle \end{aligned} \tag{60}$$

using the definition in Eq. (58). The phase factor in Eq. (60) is called the *dynamical phase*. The explicit solution for an arbitrary state can then be found by expanding $|\psi(t_0)\rangle$ in terms of eigenstates of the Hamilton operator:

$$\begin{aligned} |\psi(t_0)\rangle &= \sum_n |\psi_n\rangle\langle\psi_n|\psi(t_0)\rangle \\ |\psi(t)\rangle &= U(t,t_0)|\psi(t_0)\rangle \\ &= \sum_n e^{-iE_n(t-t_0)/\hbar}|\psi_n\rangle \\ &\quad\times\langle\psi_n|\psi(t_0)\rangle \end{aligned} \tag{61}$$

Hence, for time-independent systems, the solution of the energy eigenvalue problem (Eq. 59) is sufficient to determine the time evolution for any initial state.

1.7.1.2 Time-Dependent Hamiltonian

For a time-dependent Hamiltonian $H(t)$, the solution of the time-evolution operator U in terms of the exponential function Eq. (57) must be redefined in order to take into account that the Hamiltonian operator at different instances of time may not commute, $[H(t), H(t')] \neq 0$. We have

$$\begin{aligned} U(t, t_0) &= 1 - \frac{i}{\hbar}\int_{t_0}^{t} dt' H(t') \\ &\quad -\frac{1}{\hbar^2}\int_{t_0}^{t} dt' H(t') \\ &\quad \times \int_{t_0}^{t'} dt'' H(t'') \cdots \\ &= T \exp\left(-\frac{i}{\hbar}\int_{t_0}^{t} H(t')dt'\right) \end{aligned} \tag{62}$$

Equation (62) is called the *time-ordered exponential function* and T is called the *time-ordering operator*. In two limiting cases, the evolution operator for time-dependent Hamiltonian systems can be easily calculated by reducing the problem to the time-independent Hamiltonian. One is the adiabatic limit. If the temporal changes in $H(t)$ are very slow compared with the changes in the dynamical phases of Eq. (60) for all t, the U matrix is diagonal in the slowly changing eigenbasis of $H(t)$ at each instant t,

$$|\psi_n(t)\rangle = e^{i\gamma_n(t)} \exp\left[-\frac{i}{\hbar}\int_{t_0}^{t} dt' E_n(t')\right] |\phi_{n,t}\rangle \tag{63}$$

The dynamical phase factor is determined by $E_n(t')$, the energy eigenvalue of the stationary Schrödinger equation for the Hamiltonian operator $H(t')$ at time t'. In Eq. (63), the state $|\psi_n(t)\rangle$ denotes the evolved state and the state $|\phi_{n,t}\rangle$ represents the stationary eigenstate for the Hamiltonian $H = H(t)$ at the time t, where the time plays the role of a parameter.

The additional phase $\gamma_n(t)$ is called *Berry's phase* (Berry, 1984) or the *geometric phase*, since it records the information on the excursion of the Hamiltonian $H(t')$ in parameter space in the case of adiabatic evolution. The role of $\gamma_n(t)$ is to relate the phases of the adiabatic wave function $|\psi_n(t)\rangle$ at different points in time. Insertion of Eq. (63) into Schrödinger's equation yields the equation that defines the real phase function $\gamma_n(t)$,

$$\dot{\gamma}_n(t) - i\left\langle \phi_{n,t} \middle| \frac{d}{dt}\phi_{n,t} \right\rangle = 0 \tag{64}$$

The analogy to the parallel transport of vectors on a sphere (Figure 1.4) can give a glimpse of the concept underlying Berry's phase. Consider a quantum-mechanical state to be represented by a vector at point A. By infinitely slowly (adiabatically) varying an external parameter, for example, the direction of the magnetic field, one can ''transport'' state vectors successively from point A to B, then to C, and finally back to A. After completion of the closed loop, the vector does not coincide with its original orientation but is rotated by an angle γ that is proportional to the solid angle subtended by the loop. This angle gives rise to a geometric phase, $\exp(i\gamma)$, which depends on the geometry of a loop, but unlike the dynamical phase, Eq. (60), it is independent of the time it takes to complete the transport of the state along the closed loop. If one is now able to form in a physical system a coherent superposition of two states, one accumulating the geometric phase along the adiabatic change of the external field while the other is subject to a constant field, an interference pattern will arise. The Berry phase has been observed in interference experiments involving neutron spin rotation, in molecules, in light transmission through twisted optical-fiber cables, and in the quantum Hall effect.

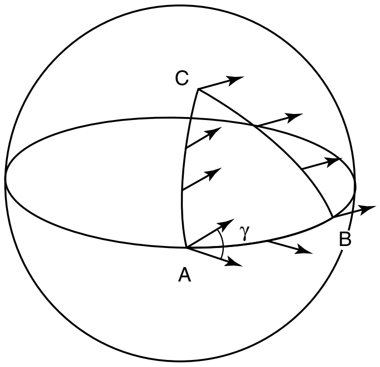

Fig. 1.4 Parallel transport of a vector on a sphere along a closed loop leads to a rotation of the vector. The vector's orientation relative to its local environment remains unchanged during parallel transport. The angle of rotation is related to Berry's phase.

In the opposite limit, where a sudden change of the Hamiltonian from $H(t) = H_0$ for all $t < 0$ to another Hamiltonian $H(t) = \overline{H}_0$ for all $t > 0$ occurs, no dynamical or geometric phase can be accumulated during the instantaneous change of the Hamiltonian at $t = 0$. The initial eigenstate of H_0, ψ_n, in which the system is prepared at $t = 0$, evolves for $t > 0$ as

$$|\psi_n(t)\rangle = \sum_{n'} \exp\left(-\frac{i\overline{\varepsilon}_{n'}t}{\hbar}\right) |\overline{\psi}_{n'}\rangle\langle\overline{\psi}_{n'}|\psi_n\rangle \tag{65}$$

In the sudden limit, the matrix U is nondiagonal. The eigenstate belonging to the "old" Hamiltonian forms a coherent superposition of eigenstates $|\overline{\psi}_n\rangle$ of the new Hamiltonian. The "sudden" formation of coherent states causes "quantum beats," observable oscillations in amplitudes and probabilities.

Consider, for example, an atom that is excited at the time $t = 0$ to a state that is a coherent superposition of a state $|\psi_1\rangle$, with energy E_1, and a state $|\psi_2\rangle$, with energy E_2, that is, $|\psi\rangle = c_1|\psi_1\rangle + c_2|\psi_2\rangle$. At a later point in time, t, the state acquires a phase factor as a function of time according to Eq. (65):

$$|\psi(t)\rangle = c_1 e^{-iE_1 t/\hbar}|\psi_1\rangle + c_2 e^{-iE_2 t/\hbar}|\psi_2\rangle \tag{66}$$

The projection onto a final state $|f\rangle$ representing, for example, the atom in the ground state and the emission of a photon,

$$\begin{aligned} P_f(t) &= |\langle f|\psi(t)\rangle|^2 \\ &= |c_1|^2|\langle f|\psi_1\rangle|^2 + |c_2|^2|\langle f|\psi_2\rangle|^2 \\ &\quad + c_1 c_2^* \exp\left[\frac{it}{\hbar}(E_2 - E_1)t\right] \\ &\quad \times \langle f|\psi_1\rangle\langle f|\psi_2\rangle^* \\ &\quad + \text{complex conjugate} \end{aligned} \tag{67}$$

will display oscillations ("quantum beats") due to the time dependence of the relative phases of the two states. These quantum beats due to dynamical phases have been observed, for example, in the Lyman-α ($n = 2 \rightarrow n = 1$) radiation of hydrogen excited in the $n = 2$ level subsequent to fast collisions with a foil (Figure 1.5).

The standard computational method of solving the time-dependent Schrödinger equation for time-dependent interactions consists of expanding a trial wave function $|\psi(t)\rangle$ in terms of a conveniently chosen truncated time-independent basis of finite size:

$$|\psi(t)\rangle = \sum_{n=1}^{N} a_n(t)|\psi_n\rangle \tag{68}$$

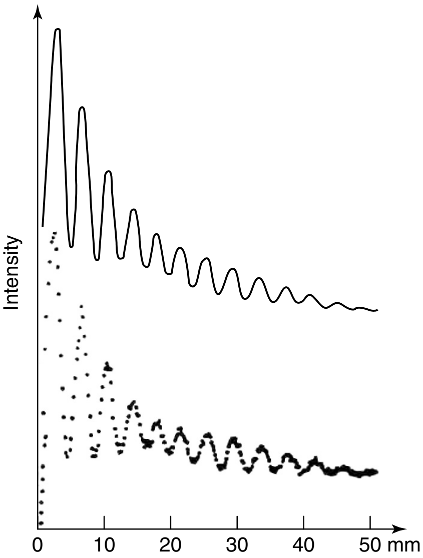

Fig. 1.5 Quantum beats in the Lyman-α photon intensity of hydrogen after collision with a carbon foil: ●, experimental data from Andrä (1974); –, theory from Burgdörfer (1981). The theory curve is shifted relative to the data for clarity.

with unknown, time-dependent coefficients $a_n(t)$. Insertion of Eq. (68) into Eq. (47) leads to a system of coupled first-order differential equations, which can be written in matrix notation as

$$i\hbar\dot{\mathbf{a}} = \mathbf{H}(t)\,\mathbf{a}(t) \tag{69}$$

using the vector notation $\mathbf{a} = (a_1, \dots a_n)$ with $\mathbf{H}$ the matrix of the Hamiltonian in the basis $\{|\psi_n\rangle\}$, $H_{n,n'} = \langle\psi_n | H | \psi_{n'}\rangle$. The solution of Eq. (69) is thereby reduced to a standard problem of numerical mathematics, the integration of a set of N coupled first-order differential equations with variable (time-dependent) coefficients.

1.7.2
Methods for Solving the Time-Independent Schrödinger Equation

The solution of the stationary Schrödinger equation, Eq. (59), proceeds mostly in the $\{|\mathbf{r}\rangle\}$ representation, i.e., using wave functions $\psi(\mathbf{r})$. Standard methods include the method of separation of variables and variational methods.

1.7.2.1 Separation of Variables

The energy eigenvalue problem, Eq. (59), requires the solution of a second-order partial differential equation. It can be reduced to an ordinary second-order differential equation if the method of separation of variables can be applied, i.e., if the wave function can be written in a suitable coordinate system (q_1, q_2, q_3) in factorized form

$$\psi(\mathbf{r}) = f(q_1)g(q_2)h(q_3) \tag{70}$$

with obvious generalizations for an arbitrary number of degrees of freedom. Equation (70) applies only to a very special but important class of problems for which the Hamiltonian possesses as many constants of motion or compatible observables as there are degrees of freedom (in our example, three). In this case, each factor function in Eq. (70) is associated with one "good" quantum number belonging to one compatible observable. This class of Hamiltonians permits an exact solution of the eigenvalue problem and is called *separable* (or *integrable*).

Consider the example of the isotropic three-dimensional harmonic oscillator. The Hamiltonian is

$$H = \frac{p_x^2}{2M} + \frac{p_y^2}{2M} + \frac{p_z^2}{2M} + \tfrac{1}{2}M\omega^2(x^2 + y^2 + z^2) \tag{71}$$

In this case, one suitable set of coordinates for the separation are the Cartesian coordinates ($q_1 = x,\ q_2 = y,\ q_3 = z$). Each of the three factor functions of Eq. (70) satisfies a one-dimensional Schrödinger equation of the form

$$\left(-\frac{\hbar^2}{2M}\frac{\partial^2}{\partial x^2} + \tfrac{1}{2}M\omega^2 x^2\right) f(x) = E_x f(x) \tag{72}$$

and $E = E_x + E_y + E_z$. The simplest solution of Eq. (72) satisfying the boundary conditions $f(x \to \pm\infty) \to 0$ for a normalizable wave function is a Gaussian wave function of the form

$$f(x) = \left(\frac{M\omega}{\pi\hbar}\right)^{1/4} e^{-(M\omega/2\hbar)x^2} \tag{73}$$

It corresponds to the lowest possible energy eigenstate, with $E_x = \frac{1}{2}\hbar\omega$. Ground-state wave functions are, in general, node free, i.e., without a zero at any finite x. The Schrödinger equation possesses an infinite number of admissible solutions with an increasing number $n = 0, 1, \ldots$ of nodes of the wave function and increasing eigenenergies

$$E_x = \hbar\omega\left(n + \tfrac{1}{2}\right) \tag{74}$$

The fact that wave functions belonging to higher states of excitation have an increasing number of zeros can be understood as a consequence of the orthogonality requirement, Eq. (37). Only through the additional change of sign of the wave function can all contributions in the orthogonality integral (Eq. 36) be made to cancel. The wave functions for excited states of a harmonic oscillator are given in terms of Hermite polynomials H_n by

$$f_n(x) = (2^n n!)^{-1/2}\left(\frac{M\omega}{\hbar\pi}\right)^{1/4} \times e^{-(M\omega/2\pi)x^2} H_n\left(\sqrt{\frac{M\omega}{\hbar}}x\right) \tag{75}$$

Examples of the first few eigenstates are shown in Figure 1.6 With increasing n, the probability density $|f_n(x)|^2$ begins to resemble the classical probability density for finding the oscillating particle near the coordinate x, $P_{\mathrm{cl}}(x)$ (Figure 1.7). The latter is proportional to the time the particle spends near x or inversely proportional to its speed:

$$P_{c1}(x) = \left(\frac{\omega}{\pi}\right)\left(2E_x/M - \omega^2 x^2\right)^{-1/2} \tag{76}$$

The increasing similarity between the classical and quantal probability densities as $n \to \infty$ is at the core of the correspondence principle. This example highlights the nonuniformity of the convergence to the classical limit: the quantum probability density oscillates increasingly rapidly around the classical value with decreasing de Broglie wavelengths λ (compared with the size of the classically allowed region). Furthermore, the wave function penetrates the classically forbidden region (''tunneling''). While the amplitude decreases exponentially, $\psi(x)$ possesses significant nonvanishing values outside the domain of classically allowed trajectories of the same energy over a distance of the order of the de Broglie wavelength λ. Only upon averaging over regions of the size of λ can the classical–quantum correspondence be

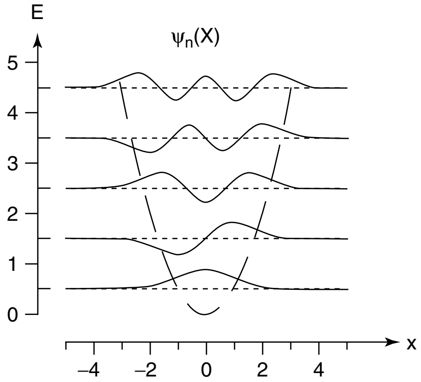

Fig. 1.6 Wave functions $f_n(x)$ for the first five eigenstates of the harmonic oscillator.

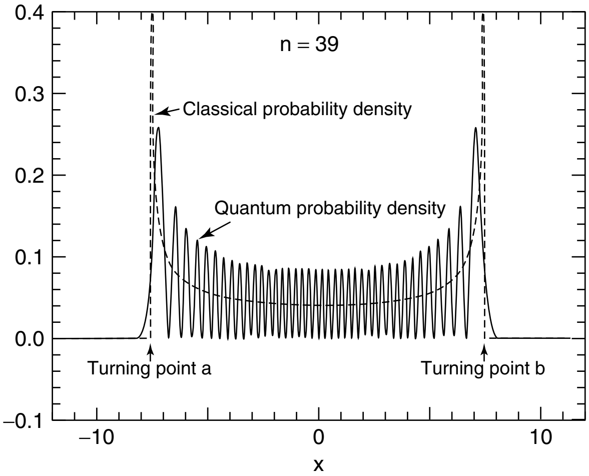

Fig. 1.7 Probability density $|f_n(x)|^2$ of a harmonic oscillator in $n = 39$ compared with the classical probability density $P_{cl}(x) = (\omega/\pi)(2E/M - \omega^2 x^2)^{-1/2}$ indicated by the dashed line. Classical turning points are denoted by a and b.

recovered. As λ tends to zero in the classical limit, classical dynamics emerges. It is the smallness of λ that renders the macroscopic world as being classical for all practical purposes.

The complete wave function for the three-dimensional isotropic oscillator (Eq. 71) consists of a product of three wave functions each of which is of the form of Eq. (75),

$$\psi_{n_x n_y n_z}(x, y, z) = f_{n_x}(x) f_{n_y}(y) f_{n_z}(z) \qquad (77)$$

with eigenenergies

$$E_{n_x n_y n_z} = \hbar\omega \left(n_x + n_y + n_z + \tfrac{3}{2}\right) \qquad (78)$$

The energy depends only on the sum of the three quantum numbers,

$$n = n_x + n_y + n_z \qquad (79)$$

but not on the individual quantum numbers n_x, n_y, and n_z, separately. Different combinations of quantum numbers yield the same energy; the spectrum is therefore degenerate. The degeneracy factor g_n, the number of different states having the same energy E_n, is

$$g_n = \frac{(n+1)(n+2)}{2} \qquad (80)$$

Only the ground state ($n = 0$) is nondegenerate, a feature generic to quantum systems and the origin of the third law (Nernst's theorem) of statistical mechanics. The degeneracy signals that alternative complete sets of commuting observables (instead of the energies in each of the three Cartesian coordinates E_x, E_y, and E_z) and corresponding alternate sets of coordinates should exist in which the problem is separable. One such set of coordinates is the spherical coordinates.

Consider the example of the hydrogen atom. One of the most profound early successes of quantum mechanics was the interpretation of the line spectrum emitted from the simplest atom, the hydrogen atom, consisting of an electron and a proton. The Schrödinger equation (in the center-of-mass frame) for the electron interacting with the proton through a Coulomb potential $V(r) = -e^2/4\pi\varepsilon_0 r$ reads as

$$\left[\frac{-\hbar^2\nabla^2}{2m_e} - \frac{e^2}{4\pi\varepsilon_0 r}\right]\psi(\mathbf{r}) = E\psi(\mathbf{r}) \qquad (81)$$

where $m_e = 9.1 \times 10^{-31}$ kg is the reduced mass of the electron, $\mathbf{r}$ is the vector that points from the nucleus to the electron, $e = 1.6 \times 10^{-19}$ C is the electric charge of the nucleus, and $-e$ is the charge of the electron. If we express the Laplace operator ∇^2 (defined as $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$) in polar spherical coordinates (r, θ, ϕ), the Schrödinger equation becomes

$$\left[\frac{-\hbar^2}{2m_e}\left(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}\right) + \frac{L^2}{2m_e r^2} - \frac{e^2}{4\pi\varepsilon_0 r}\right]\psi(r,\theta,\phi) = E\psi(r,\theta,\phi) \qquad (82)$$

Since the angular coordinates enter only through the square of the angular momentum operator (see Eq. 28), the Schrödinger equation is separable in spherical coordinates and the wave function can be written as a product $\psi(r,\theta,\phi) = R_{nl}(r)Y_e^m(\theta,\phi)$ with one factor being the spherical harmonics Y_l^m and the $R_{nl}(r)$ being the radial wave function satisfying the second-order differential equation

$$\left[\frac{-\hbar^2}{2m_e}\left(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}\right) + \frac{\hbar^2 l(l+1)}{2m_e r^2} - \frac{e^2}{4\pi\varepsilon_0 r}\right]R_{nl}(r) = ER_{nl}(r) \qquad (83)$$

The negative-energy eigenvalues of bound states can be found as

$$E_n = -\frac{m_e e^4}{2\hbar^2 n^2 (4\pi\varepsilon_0)^2} = -\frac{E_{\mathrm{H}}}{2n^2} \qquad (84)$$

where $E_H = m_e e^4/\hbar^2(4\pi\varepsilon_0)^2$ is called a *Hartree* or an *atomic unit of energy*. The energy depends only on the "principal" quantum number n, being independent of the angular momentum

$$l = 0, 1, \ldots, n-1 \qquad (85)$$

and of the magnetic quantum number m. The m independence can be found for any isotropic potential $V(|\mathbf{r}|)$ that depends only on the distance of the interacting particles. The degeneracy in l is a characteristic feature of the Coulomb potential (and of the isotropic oscillator), and it reflects the fact that the Schrödinger equation is separable not only in polar spherical coordinates but also in parabolic coordinates. The Coulomb problem possesses a dynamical symmetry that allows the choice of alternative complete sets of compatible observables. The separability in spherical coordinates is related to the set $\{H, L^2, L_z\}$ while the separability in parabolic coordinates is related to the set $\{H, L_z, A_z\}$ where $\mathbf{A}$ is the Runge–Lenz (or perihelion) vector, which, as is well known from

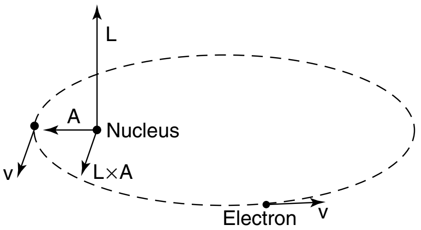

Fig. 1.8 The Runge–Lenz vector **A** and the angular momentum **L** for the classical orbit in an attractive $-1/r$ potential.

astronomy, is a constant of motion for a $1/r$ potential (Figure 1.8). The emission spectrum follows from Eq. (84) as

$$\begin{aligned}\nu &= \frac{(E_n - E_{n_0})}{h} \\ &= \frac{E_\mathrm{H}}{2h}\left(\frac{1}{n_0^2} - \frac{1}{n^2}\right)\end{aligned} \tag{86}$$

with $n_0 = 1$ and $n = 2, 3, \ldots$ giving rise to the Lyman series of spectral lines, $n_0 = 2$ and $n = 3, 4, \ldots$ giving rise to the Balmer series, and so on. The radial probability densities $r^2 R_{nl}(r)^2$ are shown in Figure 1.9

1.7.2.2 **Variational Methods**

For all but the few separable Hamiltonians, the time-independent Schrödinger equation is only approximately and numerically solvable. The accuracy of approximation can, however, be high. It may only be limited by the machine precision of modern computers. The most successful and widely used method in generating accurate numerical solutions is the variational method. The solution of the Schrödinger equation can be recognized as a solution to the variational problem of finding an extremum (in practice, a minimum) of the energy functional $\langle E\rangle = \langle\psi \mid H \mid \psi\rangle / \langle\psi \mid \psi\rangle$ by varying $|\psi\rangle$ until $\langle E\rangle$ reaches a stationary point at an approximate eigenvalue. The wave function at the minimum is an approximate solution to the Schrödinger equation. In practice, one chooses functional forms called *trial functions* $\psi(\mathbf{r})$ that depend linearly and/or nonlinearly on parameters $(c_1, \ldots, c_k)$. The trial function is optimized by minimizing the energy in the k-dimensional parameter space according to

$$\frac{d}{dc_i}\langle E\rangle = 0 \quad (1 \le i \le k) \tag{87}$$

Provided the trial function is sufficiently flexible to approximate a complete set of states in Hilbert space, ψ becomes exact in the limit as $k \to \infty$. The ground state of the helium atom consisting of an α particle as a nucleus (nuclear charge number $Z = 2$) and two electrons provides one of the most striking illustrations of the success of the variational method. The Hamiltonian (neglecting the motion of the nucleus)

$$\begin{aligned}H = &-\frac{\hbar^2}{2m_e}(\nabla_1^2 + \nabla_2^2) - \frac{Ze^2}{4\pi\varepsilon_0 r_1} \\ &-\frac{Ze^2}{4\pi\varepsilon_0 r_2} + \frac{e^2}{4\pi\varepsilon_0 |r_1 - r_2|}\end{aligned} \tag{88}$$

is nonseparable, an analytic solution is unknown, and no perturbation solution is possible since all interactions are of comparable strength. The variational method pioneered by Hylleraas (1929) and further developed by Pekeris (1958, 1959) achieves, however, a virtually exact result for the ground-state energy, $E = 2.903\,724\,377\,03\ldots E_H$, for which the accuracy is no longer limited by the method of solution but rather by the approximate nature of the Hamiltonian itself (for example, the neglect of the finite

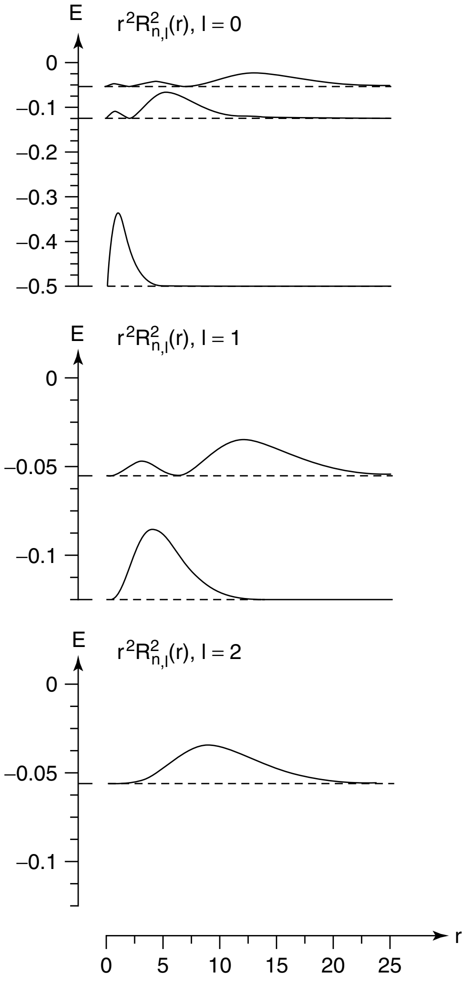

Fig. 1.9 Radial densities $r^2R_{nl}(r)$ for $n=1$, 2, 3 and $l=0$, 1, and 2 of the hydrogen atom.

size of the nucleus and of relativistic effects).

1.7.3
Perturbation Theory

A less accurate method, which is, however, easier to implement, is perturbation theory. The starting point is the assumption that the Hamiltonian H can be decomposed into a "simple," separable (or, at least, numerically solvable) part H_0 plus a perturbation gV,

$$H = H_0 + gV \tag{89}$$

which is nonseparable, but "small" such that it causes only a small modification of the spectrum and of the eigenstates of H_0. The control parameter g $(0 \le g \le 1)$ is used as a convenient formal device for controlling the strength of the perturbation and keeping track of the order of the perturbation. At the end of the calculation, g can be set equal to unity if V itself is small.

Two cases can be distinguished: stationary perturbation theory and time-dependent perturbation theory.

1.7.3.1 **Stationary Perturbation Theory**

The Rayleigh–Schrödinger perturbation theory involves the expansion of both the perturbed energy eigenvalue $E_0(g)$ and the perturbed eigenstate $|\psi_0(g)\rangle$ in powers of the order parameter g (or, equivalently, in powers of V),

$$E_0(g) = E_0 + gE^{(1)} + g^2E^{(2)} + \cdots$$
$$|\psi_0(g)\rangle = |\psi_0\rangle + g|\psi^{(1)}\rangle + g^2|\psi^{(2)}\rangle + \cdots \tag{90}$$

Inserting Eq. (90) into the Schrödinger equation

$$H(g)|\psi_0(g)\rangle = E_0(g)|\psi_0(g)\rangle \tag{91}$$

yields recursion relations ordered in ascending powers of g for the successive corrections to $E_0(g)$ and $|\psi_0(g)\rangle$. In practice, the first few terms are most important: Thus the first-order corrections for a non-degenerate eigenvalue E_0 are

$$E^{(1)} = \langle\psi_0|V|\psi_0\rangle \tag{92}$$

and

$$|\psi^{(1)}\rangle = \sum_{n\neq 0} |\psi_n\rangle \frac{\langle\psi_n|V|\psi_0\rangle}{E_0 - E_n} \tag{93}$$

where the sum over $\{|\psi_n\rangle_{n\neq 0}\}$ extends over the complete set of eigenstates of the unperturbed Hamiltonian excluding the state $|\psi_0\rangle$ for which the perturbation expansion is performed. The second-order correction to the eigenenergy is

$$E^{(2)} = \langle\psi_0|V|\psi^{(1)}\rangle = \sum_{n\neq 0} \frac{|\langle\psi_0|V|\psi_n\rangle|^2}{E_0 - E_n} \tag{94}$$

Alternative perturbation series can be devised. For example, in the Brillouin–Wigner expansion, the perturbed $E_0(g)$, rather than the unperturbed energy E_0, appears in the energy denominator of Eq. (94). The energy is therefore only implicitly given by this equation.

1.7.3.2 **Time-Dependent Perturbation Theory**

Time-dependent perturbation theory deals with the iterative solution of the system of coupled equations, Eq. (69), representing the time-dependent Schrödinger equation. One assumes again that the Hamiltonian consists of an unperturbed part H_0 and a time-dependent perturbation $V(t)$. Expansion in the basis of eigenstates of the unperturbed Hamiltonian (Eq. 68) yields a matrix $\mathbf{H}$ consisting of a diagonal matrix $\langle i\,|\,H_0\,|\,j\,\rangle = E_j\delta_{ij}$ and a time-dependent perturbation matrix $V_{ij} = \langle i\,|\,V(t)\,|\,j\rangle$.

The dynamical phases associated with the unperturbed evolution can be easily included in the perturbative solution of Eq. (69) by making the phase transformation

$$a_j(t) = c_j(t)e^{-iE_jt/\hbar} \tag{95}$$

This transformation, often referred to as the *transformation* to the *interaction representation*, removes the diagonal matrix $\langle i\,|\,H_0\,|\,j\rangle$ from Eq. (69), resulting in the new system of equations

$$i\hbar\frac{d}{dt}\mathbf{C}(t) = \mathbf{V}_I(t)\,\mathbf{C}(t) \tag{96}$$

with matrix elements

$$\langle i|V_I(t)|j\rangle = \exp\left(\frac{i}{\hbar}t(E_i - E_j)\right)\langle i|V(t)|j\rangle \tag{97}$$

and $\mathbf{C} = (c_1,c_2,\ldots,c_N)$. The perturbative solution of Eq. (96) proceeds now by assuming that initially, say at $t = 0$, the

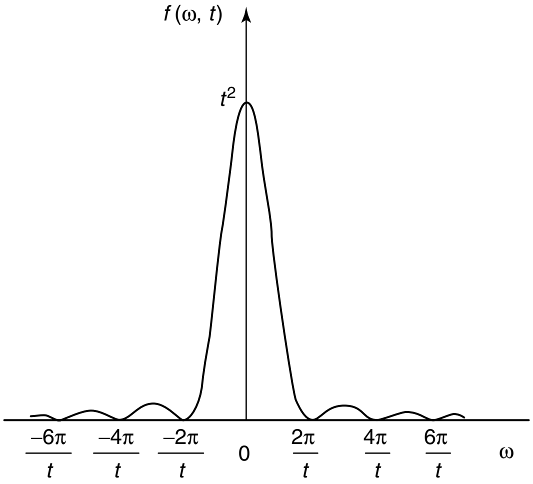

Fig. 1.10 The function $f(\omega, t)$ as a function of ω.

system is prepared in the unperturbed state $|j\rangle$ with unit amplitude, $c_j(0) = 1$. Since the perturbation is assumed to be weak, the system at a later time t will still be approximately in this state. However, small admixtures of other states will develop with amplitudes

$$\begin{aligned} c_i(t) &= \frac{1}{i\hbar}\int_0^t dt' \langle i|V_I(t')|j\rangle c_j(t') \\ &\approx \frac{1}{i\hbar}\int_0^t dt' \langle i|V_I(t')|j\rangle \end{aligned} \tag{98}$$

Therefore, to first order, the time-dependent perturbation induces during the time interval $[0, t]$ transitions $|j\rangle \rightarrow |i\rangle$ with probabilities

$$P_i(t) = |c_i(t)|^2 = \frac{1}{\hbar^2}\left|\int_0^t dt' \langle i|V_I(t')|j\rangle\right|^2 \tag{99}$$

In the special case that V is time independent, Eq. (99) gives

$$\begin{aligned} P_i(t) &= \frac{1}{\hbar^2}|\langle i|V|j\rangle|^2 \frac{|e^{i\omega_{ij}t} - 1|^2}{\omega_{ij}^2} \\ &= \frac{1}{\hbar^2}|\langle i|V|j\rangle|^2 \frac{4\sin^2(\omega_{ij}t/2)}{\omega_{ij}^2} \end{aligned} \tag{100}$$

with $\omega_{ij} = (\varepsilon_i - \varepsilon_j)/\hbar$. Plotted as a function of ω, the function $f(\omega,t) = 4\sin^2(\omega t/2)/\omega^2$ (Figure 1.10) is strongly peaked around $\omega = 0$. The peak height is t^2 and the width of the central peak is given by $\Delta\omega \simeq 2\pi/t$. In other words, f describes the energy–time uncertainty governing the transition: The longer the available time, the more narrowly peaked is the transition probability around $\omega_i - \omega_j = 0$, that is, for transition into states that are almost degenerate with the initial state. In the limit $t \rightarrow \infty$, the time derivative of Eq. (100), that is, the transition probability per unit time, or transition rate w, has the well-defined limit

$$w = \frac{d}{dt}P_i(t)_{t\longrightarrow\infty} = \frac{2\pi}{\hbar}|\langle i|V|j\rangle|^2\delta(\varepsilon_i - \varepsilon_j) \tag{101}$$

where the representation of the δ function, $\pi\delta(\omega) = lim_{t\rightarrow\infty}$ $(\sin\omega t/\omega)$, has been used. Equation (101) is often called *Fermi's golden rule* and provides a simple, versatile tool to calculate transition rates to first order. The physical meaning of the δ function is such that in the limit $t \rightarrow \infty$ only those transitions take place for which energy is conserved, as we would have expected since the interaction in Eq. (100) was assumed to be time-independent (after being switched on at time $t = 0$). For example, in scattering processes, many final states with energies very close to that of the initial states are accessible. The δ function is then evaluated within the sum (or integral) over these final states.

1.8
Quantum Scattering Theory

Scattering theory plays, within the framework of quantum mechanics, a special and important role. Its special role is derived from the fact that scattering deals with unbound physical systems: two (or more) particles approach each other from infinity, scatter each other, and finally separate to an infinite distance. Obviously, one cannot impose upon the wave function the boundary condition $\psi(r = |\mathbf{r}_2 - \mathbf{r}_2| \to \infty) \to 0$ since the probability of finding the scattered particles at large interparticle distances r is, in fact, nonvanishing. This has profound consequences: energies are not quantized since the corresponding boundary condition is missing. The energy spectrum consists of a continuum instead of a set of discrete states. The states describing the unbound motion, the "scattering states," are not square integrable and do not, strictly speaking, belong to the Hilbert space. They can only be orthonormalized in terms of a δ function,

$$\langle \psi_E | \psi_{E'} \rangle = \delta(E - E') \tag{102}$$

Mathematically, they belong to the space of distributions and can be associated with states in Hilbert space only upon forming "wave packets," that is, by convoluting the scattering state with a well localized function. Nevertheless, most of the time one can manipulate scattering states using the same rules valid for Hilbert-space vectors (see Section 1.6) except for replacing the ordinary orthonormalization condition, Eq. (37), by the δ normalization condition, Eq. (102).

The strategy for solving the Schrödinger equation, which governs both bounded and unbounded motion, is different for scattering states. Rather than determining the energy eigenvalues or the entire wave function (which is, in many cases, of no interest), one aims primarily at determining the asymptotic behavior of the wave function emerging from the interaction region, which determines the flux of scattered particles. The central quantity of both classical and quantum scattering is the cross section σ, defined as

$$\sigma = \frac{\begin{array}{c}\text{number of scattered particles}\\ \text{per unit time}\end{array}}{\begin{array}{c}\text{number of incident particles}\\ \text{per unit area and per unit time}\end{array}} \tag{103}$$

A cross section σ has the dimension of an area and can be visualized as an effective area that the target presents to the incoming projectile. Using the probabilistic description of quantum mechanics, one can replace in the definition Eq. (103) all particle currents by corresponding probability currents. Accordingly, σ can be expressed in terms of the transition rate w as

$$\sigma = \frac{w}{j} \tag{104}$$

where $j = |\psi|^2 v$ represents the incoming probability current density. For an incoming plane particle wave (Eq. 4), $j = \hbar k/M$ (with M the reduced mass and the amplitude A set equal to unity). Calculations of total cross sections σ proceed typically by first calculating differential cross sections for scattering into a particular group of final states and then summing over all possible final states. For elastic scattering of particles with incident wave vectors $\mathbf{k}_{\text{in}}$ pointing in the z direction, a group of final states is represented by a wave vector $\mathbf{k}_{\text{out}}$ of the outgoing particles pointing in a solid angle centered around the spherical angles (θ, ϕ),

$$\frac{d\sigma}{d\Omega} = \frac{M}{\hbar k_{in}} w_{\mathbf{k}_{\text{in}} \longrightarrow \mathbf{k}_{\text{out}}} \tag{105}$$

Equation (105) gives the differential cross section per unit solid angle.

1.8.1
Born Approximation

The evaluation of $d\sigma/d\Omega$ is most straightforward in the Born approximation. The first Born approximation is nothing but the application of Fermi's "golden rule," Eq. (101), to scattering:

$$\frac{d\sigma}{d\Omega} = \frac{2\pi M}{\hbar^2 k_{\text{in}}} \int_0^\infty \frac{k_{\text{out}}^2}{(2\pi)^3} dk_{\text{out}} \, |\langle \mathbf{k}_{\text{out}}|V|\mathbf{k}_{\text{in}}\rangle|^2 \times \delta(E_{\text{in}} - E_{\text{out}}) \tag{106}$$

where the integral is taken over the magnitude of k_{out} but not over the solid angle and where $k_{\text{out}}^2 dk_{\text{out}} \; d\Omega/(2\pi)^3$ represents the number of final states available for the scattered particles in the interval $dk_{\text{out}} d\Omega$. Since $k_{\text{out}}^2 dk_{\text{out}} = (M/\hbar^2) k_{\text{out}} dE_{\text{out}}$, we have

$$\frac{d\sigma}{d\Omega} = \frac{M^2}{\hbar^4 (2\pi)^2} |\langle \mathbf{k}_{\text{out}}|V|\mathbf{k}_{\text{in}}\rangle|^2 \tag{107}$$

with $|\mathbf{k}_{\text{out}}| = |\mathbf{k}_{\text{in}}|$. The first Born approximation to elastic scattering, Eq. (107), is only valid if the potential V is sufficiently weak to allow for a perturbative treatment and vanishes sufficiently rapidly at large distances for the matter wave to become a plane wave. In such cases, the matrix element in Eq. (107) can be expressed as the Fourier component of the potential,

$$\langle \mathbf{k}_{\text{out}}|V|\mathbf{k}_{\text{in}}\rangle = \int d^3 r \, e^{i\mathbf{r} \cdot (\mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}})} V(\mathbf{r}) \tag{108}$$

For a Coulomb potential $V(\mathbf{r}) = Z_1 Z_2 e^2 / 4\pi\varepsilon_0 r$ representing two charged particles with charges $Z_1 e$ or $Z_2 e$ scattering each other, Eq. (107) yields – coincidentally – the Rutherford cross section

$$\left(\frac{d\sigma}{d\Omega}\right)_{\text{R}} = \frac{Z_1^2 Z_2^2 e^4}{16 E^2 (4\pi\varepsilon_0)^2 \sin^4(\theta/2)} \tag{109}$$

This expression provided essential clues as to an internal structure and charge distribution of atoms. Two remarkable features should be noted: Prior to the advent of quantum theory, Rutherford (1911) used classical mechanics instead of quantum mechanics to derive Eq. (109). He could do so since, for Coulomb scattering, classical and quantum dynamics happen to agree. Keeping in mind that the classical limit entails the limit $\hbar \to 0$, the QM and classical cross sections should agree, since $(d\sigma/d\Omega)_{\text{R}}$ does not depend on $\hbar$. The equivalence of the classical and the QM Rutherford cross sections is due to the long range (in fact, infinite range) of the Coulomb potential. In this case, the de Broglie wavelength of the particle wave is negligibly small compared to the "size" of the target at all energies. A second coincidence lies in the fact that the first Born approximation yields the exact quantum result even though the Born approximation is, strictly speaking, not even applicable in the case of a Coulomb potential. The infinite range of the Coulomb potential distorts the matter wave at arbitrarily large distances so that it never becomes the plane wave that is assumed in Eq. (108). It turns out that all high-order corrections to the Born approximation can be summed up to give a phase factor that drops out from the expression for the cross section.

1.8.2
Partial-Wave Method

For elastic scattering by a spherically symmetric potential, that is, a potential that depends only on the distance between the particles, $V(\mathbf{r}) = V(|\mathbf{r}|)$, phase-shift analysis provides a versatile nonperturbative method for calculating differential and integral cross sections. For each angular momentum l (which is a good quantum number for a spherically symmetric potential), the phase shift for a radial wave function $u(r) = rR(r)$ can be determined from the radial Schrödinger equation

$$\left(-\frac{\hbar^2}{2M}\frac{d^2}{dr^2} + \frac{\hbar^2}{2M}\frac{l(l+1)}{r^2} + V(r)\right)u(r) = Eu(r) \tag{110}$$

The idea underlying the phase-shift analysis can be best illustrated for the phase shift δ_l for $l = 0$. In absence of the scattering potential $V(r)$, Eq. (110) represents a one-dimensional Schrödinger equation of a free particle in the radial coordinate,

$$-\frac{\hbar^2}{2M}\frac{d^2}{dr^2}u(r) = Eu(r) \tag{111}$$

whose solution is $u(r) = \sin(kr)$ with $k^2 = 2ME/\hbar^2$. Since the matter wave cannot reach the unphysical region of negative r, we have to impose the boundary condition $u(0) = 0$, which excludes the alternative solution $\sim\cos kr$. In the presence of the potential $V(r)$, the wave function will be modified within the region of the nonvanishing potential. At large distances $r \to \infty$, however, when $V(r) \to 0$, the wave function will again become $u(r) = \sin(kr+\delta_0)$ since in this region it must satisfy Eq. (111). The only effect that the potential can have on the scattered wave function at large distances is to introduce a phase shift δ_0 relative to the unperturbed wave.

Simple arguments give a clue as to the origin and the sign of the phase shift. For an attractive potential (negative V), the wave number $k(r) = \sqrt{2M(E - V(r))/\hbar^2}$ increases locally and the de Broglie wavelength $\lambda(r) = 2\pi/k(r)$ becomes shorter compared with that of a free particle of the same energy. The phase of the matter wave is therefore advanced compared with the wave of a free particle and δ_0 is positive (Figure 1.11). Similarly, if V is a repulsive potential, λ increases in the interaction region causing a phase delay and, hence, a negative phase shift δ_0. Including all angular momenta, or all "partial waves," the differential cross section can be expressed in terms of the partial-wave phase shifts δ_l as

$$\frac{d\sigma}{d\Omega} = \frac{1}{k^2}\left|\sum_{l=0}^{\infty}(2l+1)e^{i\delta_l}\sin\delta_l P_l(\cos\theta)\right|^2 \tag{112}$$

A particularly simple expression can be found for the total cross section:

$$\sigma = \frac{4\pi}{k^2}\sum_{l=0}^{\infty}(l+1)\sin^2\delta_l \tag{113}$$

For short-ranged potentials, δ_l decreases with increasing l such that only a small number of terms must be included in this sum. The scattering phase shifts $\delta_l(E)$ depend sensitively on the energy E of the incident particle.

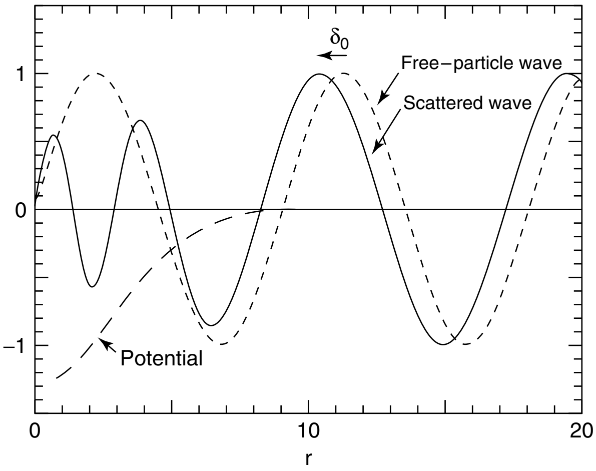

Fig. 1.11 Comparison between the scattered wave in an attractive potential and a free-particle wave (the scattering phase δ_0 is shown modulo π).

1.8.3
Resonances

Often, one phase shift $\delta_l(E)$ increases by π within a narrow energy range. This corresponds to a phase shift of one-half of a wave length, or to the addition of one node to the wave function. A sudden increase of the scattering phase by π is a signature of a "resonance" in the cross section. A resonance is characterized by a sharp maximum in the partial-wave cross section at the energy $E_{\text{res}} = \hbar^2 k_{\text{res}}^2/2M$:

$$(\sigma_l)_{\max} = \left(\frac{4\pi}{k_{\text{res}}^2}\right)(2l+1) \tag{114}$$

where the phase shift passes through $\pi/2$ (or an odd multiple thereof). Resonances appear when the scattering potential allows for the existence of "virtual levels" or quasi-bound states inside the potential well, whose probability density can leak out by tunneling through the potential well. The additional node that the scattering wave function acquires as the energy passes through E_{res} reflects the presence of a quasi-bound state. An isolated resonance can be parametrized by a formula due to Breit and Wigner. One assumes that the rapid change of the scattering phase near the resonance can be modeled by

$$\delta_l(E) \simeq \tan^{-1}\frac{\Gamma}{2(E_{\text{res}} - E)} \tag{115}$$

and that the energy dependence of the nonresonant (smooth part) of the phase shift can be neglected. The term Γ denotes the width over which the phase jump occurs. Inserting Eq. (115) into the lth term of Eq. (113) yields, for the shape of the resonant cross section, the Breit–Wigner formula

$$\sigma_l(E) = (\sigma_l)_{\text{res}}\frac{\Gamma^2/4}{(E_{\text{res}} - E)^2 + \Gamma^2/4} \tag{116}$$

The shapes of resonant cross sections become more complicated when interferences between the resonant and the nonresonant parts of the scattering amplitude come into play. The cross section can then exhibit a variety of shapes (Beutler–Fano profile; Fano, 1961), including asymmetric peaks or dips ("window resonance") depending on the relative phases of the amplitudes. In the limit that a large number of resonances with energies $E_{\text{res},i}\,(i = 1, \ldots, N)$

are situated within the width Γ ("overlapping resonances"), the cross section displays irregular fluctuations ("Ericson fluctuations") (*see* Chapter 10). These play an important role in scattering in complex systems, for example, in nuclear compound resonances (Ericson and Meyer-Kuckuk, 1966) as well as in scattering in chaotic systems Burgdörfer *et al.*, 1995).

1.9
Semiclassical Mechanics

While the correspondence principle postulates that in the limit of large quantum numbers (or, equivalently, $\hbar \to 0$), quantum mechanics converges to classical mechanics, the approach to this limit is highly nonuniform and complex. The description of dynamical systems in the limit of small but nonzero $\hbar$ is called *semiclassical mechanics*. Its importance is derived not only from conceptual but also from practical aspects since a calculation of highly excited states with large quantum numbers involved becomes difficult because of the rapid oscillations of the wave function (see Figure 1.7). Semiclassical mechanics has been developed only recently. The old Bohr–Sommerfeld quantization rules that preceded the development of quantum mechanics are now understood as an approximate form of semiclassical mechanics applicable to separable (or integrable) systems.

The wave function of the time-independent Schrödinger equation for a single particle,

$$E\psi(\mathbf{r}) = -\left(\frac{\hbar^2}{2M}\right)\nabla^2\psi(\mathbf{r}) + V(\mathbf{r})\psi(\mathbf{r}) \tag{117}$$

can be written as

$$\psi(\mathbf{r}) = A\exp\left[\frac{iS(\mathbf{r})}{\hbar}\right] \tag{118}$$

where the unknown function $S(\mathbf{r})$ replaces the exponent $\mathbf{k}\cdot\mathbf{r}$ of a plane wave (Eq. 4) for a free particle in the presence of the potential. The idea underlying the semiclassical approximation is that in the limit $\hbar \to 0$, or likewise, in the limit of small de Broglie wavelength $\lambda \to 0$ (compared with the distance over which the potential changes), $\psi(\mathbf{r})$ should behave like a plane wave with, however, a position-dependent de Broglie wavelength. Inserting Eq. (118) into Eq. (117) leads to

$$E = \frac{1}{2M}[\nabla S(\mathbf{r})]^2 + V(\mathbf{r}) - \frac{i\hbar}{2M}\nabla^2 S(\mathbf{r}) \tag{119}$$

with $\mathbf{p} = \nabla S$. Equation (119) is well suited for taking the classical limit $\hbar \to 0$ and yields the classical Hamilton–Jacobi equation. In this limit, S is the action or Hamilton's principal function. Quantum corrections enter through the term linear in $\hbar$. Approximate solutions of Eq. (119) can be found by expanding S in ascending powers of $\hbar$. The apparent simplicity of the classical limit is deceiving: the principal function S appears inside the exponent Eq. (118) with the prefactor $\hbar^{-1}$ and this causes the wave function to oscillate infinitely rapidly as $\hbar \to 0$.

1.9.1
The WKB Approximation

For problems that depend only on one coordinate, retention of the first two terms of the expansion of Eq. (119),

$$S(x) = S_0(x) + \hbar S_1(x)$$

with

$$S_0(x) = \pm \int^x p(x')dx'$$
$$S_1(x) = \tfrac{1}{2} i \ln|p(x)| + c \qquad (120)$$

leads to the Wentzel–Kramers–Brillouin (WKB) approximation where c is an integration constant to be determined by normalization of the wave function,

$$\psi^{\mathrm{WKB}}(x) = \frac{A}{\sqrt{|p(x)|}} \exp\left[\frac{\pm i}{\hbar} \int^x p(x')dx'\right] \qquad (121)$$

Several features of Eq. (121) are worth noting: the WKB wave function is an asymptotic solution of Schrödinger's equation in the limit of small λ and satisfies all rules of QM, including the superposition principle. All quantities in Eq. (121) can be calculated, however, purely from classical mechanics. In particular, the quantum probability density $|\psi^{\mathrm{WKB}}(x)|^2$ can be directly associated with the inverse classical momentum $\sim |p(x)|^{-1}$ or, equivalently, the time the particle spends near x. This result was already anticipated for the highly excited state of the harmonic oscillator (Figure 1.7). Furthermore, $S_0(x)$ is a double-valued function of x corresponding to the particle passing through a given point moving either to the right or to the left with momentum $\pm p(x)$. It is the multiple-valuedness of $p(x)$ or $S(x)$ that is generic to semiclassical mechanics and that poses a fundamental difficulty in generalizing the WKB approximation to more than one degree of freedom.

For describing bound states, a linear combination of ψ^{WKB} with both signs of the momentum in Eq. (121) is required in order to form a standing rather than a traveling wave:

$$\psi^{\mathrm{WKB}}(x) = \frac{A}{\sqrt{|p(x)|}} \sin \times \left(\frac{1}{\hbar}\int_a^x p(x')dx' + \frac{\pi}{4}\right) \qquad (122)$$

Here a is the left turning point, the boundary of the classically allowed region (Figure 1.7). The additional phase of $\pi/4$ is required to join $\psi(x)$ smoothly with the exponentially decaying solution in the classically forbidden region where $p(x)$ becomes purely imaginary ("tunneling"). With equal justification, we could have started from the classically forbidden region to the right of the turning point b and would find

$$\psi^{\mathrm{WKB}}(x) = \frac{A}{\sqrt{|p(x)|}} \sin \times \left(\frac{1}{\hbar}\int_x^b p(x')dx' + \frac{\pi}{4}\right) \qquad (123)$$

The uniqueness of $\psi^{\mathrm{WKB}}(x)$ mandates the equality of Eqs (122) and (123). The two are equal if the integral over the whole period of the trajectory (from a to b and back to a) equals

$$\left(\int_a^b + \int_b^a\right) p(x)dx = \oint p(x)dx = (n + \tfrac{1}{2})2\pi\hbar \qquad (124)$$

Equation (124) can be recognized as the Bohr–Sommerfeld quantization rule of the "old" quantum theory put forward prior to the development of quantum mechanics. Quantization rules were successful in exploring the discrete spectrum of hydrogen, one of the great puzzles that classical dynamics could not resolve. The additional term $\frac{1}{2} \times 2\pi\hbar$ in Eq. (124) was initially introduced as an empirical correction. Only within the semiclassical limit of QM does

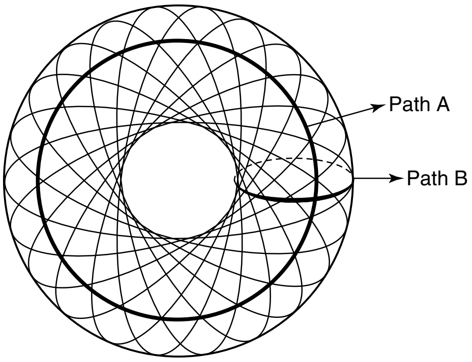

Fig. 1.12 Confinement of trajectories of an integrable Hamiltonian for constant energy E to the surface of a torus, shown here for a system with two degrees of freedom ($N = 2$) (after Noid and Marcus, 1977).

its meaning become transparent: it is the phase loss in units of $h = 2\pi\hbar$ of a matter wave upon reflection at the turning points (in multidimensional systems, at caustics). These phase corrections are of the form $\alpha(\pi/4)$ with $\alpha = 2$ in the present case. The quantity α is called the *Maslov index* and it is used to count the number of encounters with the caustics (in general, the number of conjugate points) during a full period of the classical orbit.

1.9.2
The EBK Quantization

The generalization of the WKB approximation to multidimensional systems faces fundamental difficulties: the momentum vector $\mathbf{p}(\mathbf{r})$ is, in general, not smooth but it is a highly irregular function of the coordinate vector $\mathbf{r}$ and takes on an infinite number of different values. Take, for example, a billiard ball moving in two dimensions on a stadium-shaped billiard table. A ball of a fixed kinetic energy $T = p^2/2M$ can pass through any given interior point $\mathbf{r}$ with a momentum vector $\mathbf{p}$ pointing in every possible direction. Consequently, the generalization of Eq. (121),

$$\psi(\mathbf{r}) = \sqrt{|D(\mathbf{r})|} \times \exp\left[\frac{i}{\hbar}\int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{p}(\mathbf{r}') \cdot d\mathbf{r}'\right] \tag{125}$$

is well defined only under special circumstances, that is, when $\mathbf{p}(\mathbf{r})$ is smooth and the action $S(\mathbf{r}_0, \mathbf{r}) = \int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{p} \cdot d\mathbf{r}'$ is either single valued or (finite-order) multiple valued (called a *Lagrangian manifold*; Gutzwiller, 1990). Einstein pointed out, as early as 1917, that the smoothness applies only to integrable systems, i.e., systems for which the number of constants of motion equals the number of degrees of freedom, N. For such a system, classical trajectories are confined to tori in phase space (as shown in Figure 1.12 for $N = 2$) and give rise to smooth vector fields $\mathbf{p}(\mathbf{r})$ (Figure 1.13). The corresponding quantization conditions are

$$\oint_{C_i} \mathbf{p}(\mathbf{r}) \cdot d\mathbf{r} = 2\pi\hbar\left(n + \frac{\alpha}{4}\right) \quad (i = 1, \ldots, N) \tag{126}$$

where C_i $(i = 1, \ldots, N)$ denote N topological distinct circuits on the torus, that is, circuits that cannot be smoothly deformed into each other without leaving the torus (for example, the paths A and B in Figure 1.12). The conditions of Eq. (126) are called the *Einstein–Brillouin–Keller* (*EBK*) quantization rules and are the generalization of the WKB approximation (Eq. 124). The amplitude of Eq. (125) is given by the projection of the point density of the torus in phase space onto the coordinate space. Since the point density is

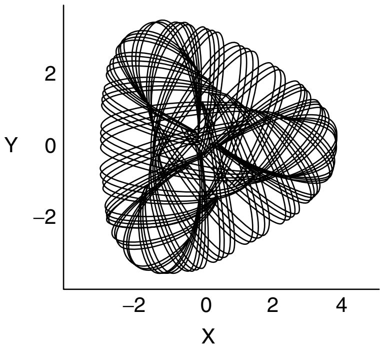

Fig. 1.13 Regular trajectory in coordinate space for an integrable system. The vector field **p**(**r**) is multiple valued but smooth (after Noid and Marcus, 1977).

uniform in the classical angle variables θ_i conjugate to the action (Goldstein, 1959), the amplitude is given by

$$|D(\mathbf{r})| \propto \det\left|\frac{\partial\theta_1\partial\theta_2\partial\theta_3}{\partial x\partial y\partial z}\right| \tag{127}$$

For a system with one degree of freedom ($N = 1$), Eq. (125) reduces to Eq. (121) since

$$\begin{aligned}\left|\frac{\partial\theta}{\partial x}\right|^{1/2} &= \left|\frac{\partial t}{\partial x}\frac{\partial\theta}{\partial t}\right|^{1/2}\\ &= \left|\frac{M}{p}\omega\right|^{1/2} \propto \frac{1}{|p|^{1/2}}\end{aligned} \tag{128}$$

1.9.3
Gutzwiller Trace Formula

As already anticipated by Einstein (1917), tori do not exist in many cases of practical interest – for example, for the motion of an electron in a stadium-shaped semiconductor heterostructure (Marcus *et al.*, 1992). WBK or EBK approximations are not applicable since the classical motion of such an electron is chaotic, destroying the smooth vector field **p**(**r**) as in the billiard discussed above. The semiclassical mechanics for this class of problems was developed only recently (Gutzwiller, 1967, 1970, 1971). The starting point is the semiclassical Van Vleck propagator (Van Vleck, 1928),

$$\begin{aligned}K^{\mathrm{SC}}&(\mathbf{r}_1,\mathbf{r}_2,t_2-t_1)\\ &= \sqrt{|D(\mathbf{r}_1,\mathbf{r}_2,t_2-t_1)|}\\ &\quad\times\exp\left[\frac{i}{\hbar}\int_{t_1}^{t_2} L(\mathbf{r}(t'))dt' - \frac{i\alpha\pi}{2}\right]\end{aligned} \tag{129}$$

where L is the classical Lagrange function for the particle traveling along the classically allowed path from $\mathbf{r}_1$ at time t_1 to $\mathbf{r}_2$ at time t_2 and α counts the number of singular points of the amplitude factor D (similar to the Maslov index). If more than one path connects $\mathbf{r}_1(t_1)$ with $\mathbf{r}_2(t_2)$, the coherent sum of terms of the form in Eq. (129) must be taken. The Van Vleck propagator was recognized only 20 years later (Feynman, 1948) as the semiclassical limit of the full Feynman quantum propagator

$$\begin{aligned}K&(\mathbf{r}_1,\mathbf{r}_2,t_2-t_1)\\ &= \int\{D\mathbf{r}\}\exp\left[\frac{i}{\hbar}\int_{t_1}^{t_2} L(\mathbf{r},t')dt'\right]\end{aligned} \tag{130}$$

where $\{D\mathbf{r}\}$ stands for the path integral, that is the continuous sum over all paths connecting $\mathbf{r}_1$ and $\mathbf{r}_2$ including all those that are classically forbidden. In the semiclassical limit $\hbar \to 0$, the rapid oscillations in the phase integral cancel all contributions except for those stemming from the classically allowed paths for which the actions $\int L dt$ possess an extremum or, equivalently, for which the phase in Eq. (130) is stationary.

Equation (129) provides a method for investigating the semiclassical time evolution of wave packets in time-dependent problems. The Fourier–Laplace transform of Eq. (129) with respect to t yields the semiclassical Green's function $G(\mathbf{r}_1, \mathbf{r}_2, E)$. The spectral density of states of a quantum system can be shown to be

$$n(E) = \sum_i \delta(E - E_i) = -\frac{1}{\pi}\mathrm{Tr}\,\mathrm{Im}\, G(\mathbf{r}_1 = \mathbf{r}_2, E) \qquad (131)$$

where the trace implies integration over the variables $\mathbf{r}_1, \mathbf{r}_2$ with constraint $\mathbf{r}_1 = \mathbf{r}_2$. Note that each of these eigenenergies causes a sharp peak (resembling the singularity of the δ function) in $n(E)$. Calculation of $n(E)$ therefore allows the determination of eigenenergies of quantum systems. It is possible to evaluate Eq. (131) in the semiclassical limit by using the semiclassical Green's function and evaluating the trace in the limit $\hbar \to 0$ by the stationary-phase method. The result is the trace formula

$$n(E) = n_{\mathrm{av}}(E) + \mathrm{Im}\frac{1}{i\hbar\pi} \times \sum_{\substack{\text{periodic}\\ \text{orbits}}} \frac{T_i}{2\sinh(T_i\lambda_i/2)} \times \exp\left[i\left(\frac{S_i}{\hbar} - \alpha_i\frac{\pi}{2}\right)\right] \qquad (132)$$

The sum extends over all unstable classical periodic orbits, each characterized by the period T_i, the instability (or ''Lyapunov'') exponent λ_i, and the action S_i, all of which are functions of the energy E. An orbit is unstable if its Lyapunov exponent is positive. The term λ_i measures the rate of exponential growth of the distance from the orbit when the initial condition is ever so slightly displaced (Gutzwiller, 1990). The smooth average density of states,

$$n_{\mathrm{av}}(E) = \frac{d}{dE}\left(\frac{V(E)}{(2\pi\hbar)^N}\right) \qquad (133)$$

is given by the rate of the change of the classical phase-space volume with energy in units of the volume of Planck's unit cell $(2\pi\hbar)^N$ associated with each quantum state. The derivation of Eq. (132) assumes two degrees of freedom and makes explicit use of the fact that the physical system is completely chaotic and integrable regions are absent. Only (hyperbolic) unstable periodic orbits are assumed to exist, each with a positive Lyapunov exponent $\lambda_i > 0$ that measures the instability of each periodic orbit. Equation (132) is therefore complementary to the EBK quantization formula (Eq. 126). This formula of ''periodic orbit quantization'' is an asymptotic series, that is, the sum over all periodic orbits is, in general, divergent. For several systems, the sum over the first few orbits has been shown to reproduce the position of energy levels quite well. The mathematical and physical properties of the Gutzwiller trace formula are still a topic of active research in quantum mechanics.

1.10
Conceptual Aspects of Quantum Mechanics

QM has proved to be an extraordinarily successful theory and, so far, has never been in contradiction to experimental observation. The Schrödinger equation has cleared every test carried out so far, including a precision test of its linearity: the upper bound for corrections due to nonlinear terms in the wave functions,

$\psi(x)^n$ ($n > 1$), in the Schrödinger equation stands currently at one part in 10^{27} (Majumder *et al.*, 1990). Even so, the conceptual and philosophical aspects of QM have remained a topic of lively discussion stimulated, to a large part, by the counterintuitive consequences of QM that apparently defy common-sense interpretation. The debate, which can be traced to the origins of QM and Einstein's objections summarized in his famous quote "Gott würfelt nicht" ("God does not play dice"), is centered around two aspects: the superposition principle, Eq. (15), and the probability interpretation of the wave function, Eq. (39).

1.10.1
Quantum Mechanics and Physical Reality

The counterintuitive aspects of the superposition principle are highlighted by the thought experiment of "Schrödinger's cat." The coherent superposition of two states, each of which is presumably an eigenstate of an extremely complex many-body Schrödinger equation, describes the state of the cat, being either dead, $|d\rangle$, or alive, $|a\rangle$:

$$|\psi_{\text{cat}}\rangle = c_d|d\rangle + c_a|a\rangle \tag{134}$$

with $|c_d|^2+|c_a|^2=1$. Only by way of measurement, that is, by projection onto the states $|a\rangle$ or $|d\rangle$,

$$P_a = |\langle a|\psi_{\text{cat}}\rangle|^2, \quad P_d = |\langle d|\psi_{\text{cat}}\rangle|^2 \tag{135}$$

can we find out about the fate of the cat. While the meaning of two sharply different outcomes of Eq. (135), "dead" or "alive," is deeply engraved in our classical intuition, the quantum state of a coherent superposition, Eq. (134), defies any classical interpretation. Yet it is an accessible state of the system. The point to be emphasized is that it is not our incomplete knowledge of an independently existing objective reality, but the fate of the cat itself, that can only be resolved by the measurement. Such a measurement is therefore, quite literally, a "demolition" measurement leading to the "collapse" of the wave function to one of the two possible states. In hypothetical thought experiments for macroscopic systems such as Schrödinger's cat problem, the average over a huge number ($\sim 10^{23}$) of uncontrolled degrees of freedom will, for all practical purposes, destroy any well-definded phase relation or coherence between the "states" dead and alive. In contrast, the axioms of QM can be put to a rigorous test for microscopic systems, for example, by performing correlation measurements on two particles emitted from the same source. Examples along these lines were, for example, put forward by Einstein, Podolsky and Rosen (1935) to challenge the purely probabilistic interpretation of QM.

Consider the emission of two electrons from, for example, a helium atom (Figure 1.14). The spin of each electron $i = 1,2$ can be in a coherent superposition state of spin up $|+\rangle$ and spin down $|-\rangle$,

$$|\psi_i\rangle = \alpha_i|+\rangle + \beta_i|-\rangle \tag{136}$$

relative to one particular Cartesian axis, for example, the z axis. One possible spin state of the electron pair (ignoring here the antisymmetrization requirement for fermions) would be an uncorrelated product state,

$$|\psi\rangle = |+\rangle_1|-\rangle_2 \tag{137}$$

Already Schrödinger pointed out that quantum theory allows to be alternatively

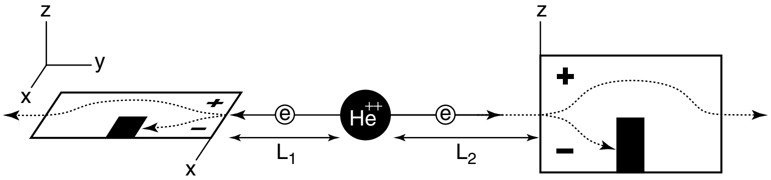

Fig. 1.14 Two-electron emission from helium with the electrons being spin-analyzed in two Stern–Gerlach-type filters with different orientations (along the x and z axis, respectively). The difference in length, $L_1 < L_2$, makes the spin selection in the left detector happen first (inducing a "wave function collapse") and in the right detector later.

in an entangled (or quantum-correlated) superposition state,

$$|\psi\rangle = \alpha|+\rangle_1|-\rangle_2 + \beta|-\rangle_1|+\rangle_2 \qquad (138)$$

In fact, the electron pair in the (singlet) ground state of helium with total spin $S = 0$ forms such an entangled pair with $\alpha = -\beta = 1/\sqrt{2}$. The hallmark of entangled states of the form Eq. (138) is that they cannot be represented as a factorized product in the form of Eq. (137).

The consequences of such quantum correlations are profound: when measuring the spin of particle 1 to be up (+), we know with certainty that particle 2 has its spin down (−). The "collapse" of the entangled wave function when measuring particle 1 forces particle 2 to switch its spin from a completely undetermined state to a well-defined value. We therefore obtain information about the orientation of spin 2 without actually performing a measurement on this particle. This correlation, often referred to as the *nonlocality of QM*, is a consequence of the probabilistic description and superposition of wave functions in QM. The emergence of a definite physical property or physical reality by "postselection" through the measurement of an entangled state goes even further. For example, the entangled state Eq. (138) with $\alpha = -\beta = 1/\sqrt{2}$,

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|+\rangle_1|-\rangle_2 - |-\rangle_1|+\rangle_2) \qquad (139)$$

is rotationally symmetric and equally applies when we inquire about the spin orientation along the x rather than along the z axis. If we postselect the spin-up direction (+) along the x axis for the first particle, the second particle will have spin down (−) along the same axis. If we, instead, measured for the second particle the polarization relative to the z axis as before (Figure 1.14), we would find both spin up and down with 50% probability, and thus a vanishing average spin. Different projection protocols for the first spin imply entirely different results for the second. The emergence of a specific "physical reality" out of a specific measurement context is often referred to as *quantum contextuality* and is unknown in classical physics. It has therefore been questioned whether the quantum mechanical description of a state such as in Eq. (138) can be considered a complete picture of physical reality (Einstein, Podolsky and Rosen, 1935). In particular, it was conjectured that particles may have additional properties that are underlying the quantum-mechanical picture but which are hidden from our view. Such "hidden variables" would ensure that physical systems have only properties that are "real," that is, independent of a measurement and that the result of any measurement is only dependent on "local" system properties rather than on a remote detection of another observable. The knowledge of "hidden variables" would allow one to remove the probabilistic aspects of QM. The role of probability in QM could then be the same as in statistical mechanics: a convenient way to quantify our incomplete knowledge of the state of a system. The hidden-variable hypothesis, or alternatively, the quantum correlation built into entangled states, Eq. (138), has been probed by testing the violation of Bell's inequality (Bell, 1969; Clauser *et al.*, 1969). Experiments with correlated pairs of particles, in particular photon pairs – for example, by Aspect and coworkers (1982) – have falsified the hidden-variable hypothesis by showing that entangled states can violate Bell's inequality while being in full agreement with the predictions of QM. More recent experiments have succeeded in measuring entangled photons at a distance of several hundred meters apart and with light polarizers (used as photon analyzers), which switch their axes so rapidly that during the photons' time of flight the polarizers have no chance to exchange information on their respective orientation. (The "collapse" of the wave function by measuring one of the two photons occurs instantaneously for both photons, but cannot be used to transmit any information and therefore does not violate the relativistic causality relations.) Even under such stringent conditions, a violation of Bell's inequality by more than 30 standard deviations was observed experimentally (Weihs *et al.*, 1998). Also so-called noncontextual hidden-variable theories and realistic models that assume a certain type of "nonlocality" are untenable (Hasegawa *et al.*, 2006; Gröblacher *et al.*, 2007).

1.10.2
Quantum Information

The most fundamental unit of information in classical information theory is the so-called bit, which can take on only one of two mutually exclusive values (typically labeled as "0" and "1"). In conventional computers, all numbers and operations with them are encoded in this binary system. The corresponding quantum analog of the bit – named *qubit* – is given by a two-state system with the two states $|+\rangle$ and $|-\rangle$ from Eq. (136), now labeled as $|0\rangle$ and $|1\rangle$. What sets the qubit apart from the conventional classical bit is precisely the ability to be in a coherent superposition of both basis states,

$$|Q\rangle = \alpha|0\rangle + \beta|1\rangle \tag{140}$$

with $|\alpha|^2 + |\beta|^2 = 1$. In quantum information theory, the intrinsic indeterminacy of the qubit's value before a measurement is exploited as a resource rather than treated as a deficiency.

One potential application of qubits is subsumed under the name "quantum computation". The basic building block of a so-called quantum computer would be a number of N entangled qubits that can be in any superposition of 2^N basis states. An example for $N = 2$ is just the maximally entangled pair in Eq. (139). While a collection of N classical bits can only be in a single one of the 2^N possible states at a given time, N qubits can be in any superposition state and thus can clearly have a much higher information content. Owing to the superposition principle, any of the (unitary) logic operations of the quantum computer performed on the initial state could process all the involved 2^N basis states simultaneously. Several algorithms have been suggested (Shor, 1994; Grover, 1996) for how this inherently quantum feature of ''parallel processing'' could be used to solve fundamental computational problems much faster, i.e., with much fewer operations than with the corresponding classical computations.

The manipulation of quantum states within the framework of quantum computation and quantum information processing involves a set of rules as, for example, the so-called no cloning theorem (Wootters and Zurek, 1982), which states the impossibility to create identical copies of an arbitrary unknown quantum state. Although only imperfect ''clones'' of a quantum state are possible, a transfer of a quantum state from one system to a second, possibly distant, system is allowed according to QM. Such a ''teleportation'' of a quantum state involves the destruction of the original state during the teleportation procedure and is typically performed through an intermediate entangled state of three particles (Bennett *et al.*, 1993; Bouwmeester *et al.*, 1997).

The laws of QM can also be used for secure information transfer. In classical cryptography, secret messages are enciphered by way of a code that the recipient of a message must know in order to decipher a message's content. To prevent the breaking of the code by a third unauthorized party, the key for the encoding algorithm is preferably defined as a random sequence of numbers, which is generated anew for each message transmitted.

Challenges to classical cryptography are the need of a truly random key and the risk of key corruption through the interception by a third party. Both of theses challenges can be met by intrinsic quantum features. A truly random sequence of numbers can be generated, for example, by repeatedly measuring the spin orientation (or photon polarization) of an unpolarized spin, a situation that may be realized by first projecting the first spin of an entangled pair along a perpendicular direction (see the example of Fig. 1.14 above).

Interception of a message shared with the recipient through an entangled pair by a third party would require a measurement, which leads to a wave function collapse and to a detectable reduction in quantum correlations. Using the principles of QM for the generation and transmission of a secure encryption key – a technique that is referred to as *quantum cryptography* – has been successfully implemented for setups involving very large distances (144 km) between the sender and recipient (Ursin *et al.*, 2007).

1.10.3
Decoherence and Measurement Process

One fundamental obstacle in practically implementing quantum computation and quantum information protocols is the challenge to keep the coherent superposition of a quantum state, in particular of entangled pairs, ''alive'' for extended periods of time. Any unintentional or intentional interaction of the system inevitably and irreversibly converts the quantum superposition into a classical probability distribution and destroys the operating principle of a quantum information device. The measurement process and the destruction of coherence, frequently called *decoherence*, are closely intertwined. In the early days of quantum theory, the measurement device that caused the ''collapse'' of the wave function was considered to be a ''classical'' apparatus decoupled from the quantum dynamics to be observed

(Bohr, 1928). The notion of open quantum systems has superseded this view and provides a widely accepted unified description of measurement and decoherence within the realm of quantum theory. The starting point here is the fundamental insight that the quantum system (S) to be observed is never completely isolated from the environment (E), which itself is quantum rather than classical. The Hamiltonian describing both system and environment is given by $H = H_S + H_E + H_{S-E}$, containing the Hamiltonian of the system (H_S) and its environment (H_E), as well as the interaction between the two (H_{S-E}). A state describing both system and environment, $|\psi\rangle$, even when at the initial time $t = 0$ prepared in a factorizable state, $|\psi(t=0)\rangle = |\psi_S^i\rangle|\psi_E^j\rangle$, will evolve in the course of time under the influence of the interaction H_{S-E} into an entangled, nonfactorizeable state

$$|\psi(t)\rangle = \sum_{i,j} a_{ij}|\psi_S^i\rangle|\psi_E^j\rangle(t) \tag{141}$$

Here, $|\psi_S^i\rangle$ (and $|\psi_E^j\rangle$) are basis states of the system (of the environment), which often, but not always, are eigenstates of the Hamilton operator of the system H_S (of the environment H_E). "Tracing out" the (unobserved) environment degrees of freedom of the entangled state by taking expectation values leads to decoherence in the (observed) system and mimics the effect of a classical measurement device.

Consider, as an illustrative example, a situation in which the system and the environment are each represented by a two-dimensional Hilbert space and described by the following entangled state:

$$|\psi(t)\rangle = \alpha|+\rangle_S|-\rangle_E + \beta|-\rangle_S|+\rangle_E \tag{142}$$

The expectation value of any physical observable represented by an operator A_S acting only on the states in the Hilbert space of the system S gives in the presence of entangled environmental degrees of freedom,

$$\begin{aligned}\langle A_S\rangle &= \langle\psi|A_S|\psi\rangle \\ &= |\alpha|_S^2\langle+|A_S|+\rangle_S \\ &\quad +|\beta|_S^2\langle-|A_S|-\rangle_S \\ &= Tr(\rho A_S)\end{aligned} \tag{143}$$

with $\rho = |\alpha|^2|+\rangle_S\,{}_S\langle+| + |\beta|^2|-\rangle_S\,{}_S\langle-|$. It is the coherent entanglement with the environment, Eq. (142), that leads to (partial) decoherence within the system (S) and to a statistical mixture described by the density operator introduced above (Eq. 40) with probabilities $P_+ = |\alpha|^2$ and $P_- = |\beta|^2$. The ubiquitous coupling of a physical system to the environment provides the natural setting for the transition from the quantum indeterminism of superpositions to the definiteness of the classical measurement. The classical limit induced by decoherence is conceptually different from the classical limit of quantum systems of large quantum numbers, which is the focus of semiclassical mechanics (see Section 1.9). In spite of recent progress, it is generally still considered an open question if decoherence concepts alone can fully explain the quantum-to-classical transition and provide definite answers to the widely discussed questions regarding the "quantum measurement problem."

The above considerations demonstrate that it is very unlikely for a macroscopic object such as Schrödinger's cat to ever be in a superposition state "dead and alive" as in Eq. (134) because it is extremely difficult to isolate the cat from its environment. For all practical purposes, an exchange of particles and heat between the cat and its environment will prevent the cat, for itself, to be in a pure quantum state. It will rather be described by a density matrix with strongly reduced quantum correlations. As both the cat and the environment are macroscopic objects with many degrees of freedom, decoherence sets in at an extremely short timescale. Recent experiments have been able to sufficiently decouple "mesoscopic" objects as superconducting rings (Friedman *et al.*, 2000) from their environment to observe these systems in superposition states.

1.11
Relativistic Wave Equations

When the characteristic speed v of particles in a physical system becomes comparable to the speed of light, $v \leq c$, the Schrödinger equation, which incorporates nonrelativistic quantum dynamics, fails. Generalizations to relativistic wave equations face, however, fundamental difficulties due to the possibility of particle creation and destruction in relativistic dynamics (*see* Chapter 2). For example, if photons have energies in excess of twice the relativistic rest energy of electrons, $E_\gamma > 2m_e c^2$, a pair consisting of an electron and its antiparticle, a positron, can be created. Therefore, the interpretation of $|\psi(\mathbf{r})|^2$ as a single-particle probability density is no longer possible. A consistent resolution of the difficulties can be accomplished within the framework of quantum field theory, which is outside the scope of this chapter. The relativistic analogs to the Schrödinger equation discussed in the following section are only applicable to relativistic systems in which particle production and destruction are not yet important.

1.11.1
The Klein–Gordon Equation

We use the same recipe that led us in Section 1.3 to the Schrödinger equation, i.e., we replace the quantities H and $\mathbf{p}$ by differential operators (Eq. 11), motivated by the de Broglie hypothesis. In this case, however, we use the relativistic relation between energy and momentum (*see* Chapter 2) instead of the nonrelativistic one, Eq. (9):

$$E - V(\mathbf{r}) = \sqrt{M^2c^4 + p^2c^2} \tag{144}$$

Following this recipe, we are led to

$$i\hbar\frac{d}{dt}\psi(\mathbf{r},t) = \left[\sqrt{-\hbar^2\nabla^2c^2 + M^2c^4} + V(\mathbf{r})\right]\psi(\mathbf{r},t) \tag{145}$$

Equation (145) has a major drawback: it treats the space and time derivatives asymmetrically. The Lorentz invariance of relativistic dynamics requires, however, that space and time coordinates be treated on equal footing. We thus start with the square of Eq. (144) and arrive at

$$\left[i\hbar\frac{\partial}{\partial t} - V(\mathbf{r})\right]^2\psi(\mathbf{r},t) = (-\hbar^2\nabla^2c^2 + M^2c^4)\psi(\mathbf{r},t) \tag{146}$$

This is the time-dependent Klein–Gordon equation, applicable to particles with integer spin. The corresponding stationary Klein–Gordon equation follows by replacing $i\hbar\partial/\partial t$ by E. Equation (146) gives an eigenvalue equation in E^2 rather than in E itself. Consequently, upon taking the square root of the eigenvalue, solutions with both signs, $\pm\sqrt{p^2c^2 + M^2c^4}$, will appear. The physical meaning of negative-energy solutions is discussed below. For a charged particle with charge q in the presence of an electromagnetic field, both the scalar potential $\phi (V = q\phi)$ and the electromagnetic vector potential $\mathbf{A}$ must be included by "minimum coupling":

$$\mathbf{p} \longrightarrow \mathbf{p} - q\mathbf{A} = \left(\frac{\hbar}{i}\right)\nabla - q\mathbf{A} \tag{147}$$

With this coupling, the Klein–Gordon equation is invariant under gauge transformations of the electromagnetic field as well as under Lorentz transformations.

1.11.2
The Dirac Equation

Instead of taking the square of Eq. (144), Dirac pursued a different approach in developing a Lorentz-invariant relativistic wave equation. He "linearized" the square root

$$\sqrt{c^2p^2 + M^2c^4} \longrightarrow \left(c\,\boldsymbol{\alpha}\cdot\mathbf{p} + \beta Mc^2\right) \tag{148}$$

with the subsidiary condition that the square of the linearized operator satisfies

$$\left(c\boldsymbol{\alpha}\cdot\mathbf{p} + \beta Mc^2\right)^2 = c^2p^2 + M^2c^4 \tag{149}$$

It turned out that solutions to this equation can be found only if all α_i $(i = x,\ y,\ z)$ and β are taken as matrices of minimum dimension 4×4. Equation (149) is satisfied if the matrices α_i, β satisfy the relations

$$\begin{aligned}\alpha_i\alpha_k + \alpha_k\alpha_i &= 2\delta_{ik}\\ \alpha_i\beta + \beta\alpha_i &= 0\end{aligned} \tag{150}$$

Accordingly, ψ is now a four-component entity (called a *spinor*), $\psi = \{\psi_i,\ i = 1, \ldots, 4\}$. The resulting Dirac equation describes spin $\frac{1}{2}$ particles. When a charged spin $\frac{1}{2}$ particle interacts with an electromagnetic field as described by minimal coupling, the stationary Dirac equation becomes

$$\begin{aligned}&[E - q\phi(\mathbf{r})]\psi(\mathbf{r})\\ &= \left[c\boldsymbol{\alpha}\left(\frac{\hbar}{i}\nabla - q\mathbf{A}\right) + \beta Mc^2\right]\psi(\mathbf{r})\end{aligned} \tag{151}$$

Dirac spinors permit the probability-density interpretation, that is, $|\psi(x)|^2 = \sum_{i=1}^{4} |\psi_i(x)|^2$. The physical meaning of the four components can be easily understood by considering a free particle. The energy eigenvalues are $E = \pm\sqrt{p^2c^2 + M^2c^4}$. In addition to the positive energies we find negative eigenenergies, which are a feature of relativistic quantum mechanics and which result from the fact that the square-root operator permits both the positive and negative sign of E. The four-spinor can be decomposed into two 2-spinors $\psi = (\phi_1,\ \phi_2)$ one of which represents the positive-energy solution and the other the negative-energy solution. Each component of the two-spinor wave function can be understood as a product of a configuration-space wave function $e^{i\mathbf{kr}}$ and a spin 1/2 eigenstate $|m_s = \pm 1/2\rangle$. A major success of the Dirac theory was that the intrinsic degree of an electron spin, which was a mere add-on within the Schrödinger theory, is incorporated into the Dirac equation from the outset. Furthermore, the g factor for the magnetic moment of the electron as well as the spectrum of the hydrogen atom can be derived from the Dirac equation to a very high degree of approximation. The pieces still missing are the quantum electrodynamic (QED) corrections that lead to small energy shifts of atomic levels ("Lamb shift") and the deviations of the g factor for the anomalous magnetic moment from

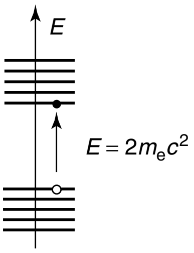

Fig. 1.15 Particle–antiparticle production visualized as the excitation of a negative-energy particle to positive energies leaving behind a hole.

2, $g = 2 \times 1.001\,16$. These corrections can be determined only within quantum field theory, the theory of the interactions of charged particles with the quantized electromagnetic field.

The negative-energy states, according to Dirac, can be interpreted in terms of antiparticle states. The vacuum state of empty space is represented by a fully occupied "sea" of negative-energy states. While this "sea" itself is not observable, "holes" created in the sea by exciting negative-energy electrons to positive energies (Figure 1.15) are observable and appear as antiparticles. For example, a missing electron with charge $-e$, energy $-E$, and momentum $\mathbf{p}$ represents a positron with charge e, energy E, and momentum $-\mathbf{p}$. The negative-energy states of electrons therefore offer a simple explanation of particle–antiparticle pair production in terms of the excitation of a negative-energy particle to a positive-energy state (the particle) leaving a hole in the negative-energy "sea" (the antiparticle) behind.

Glossary

Coherence and Decoherence: In QM, the term *coherence* stands for the ability of a particle's state to interfere. Perfectly coherent states ("pure states") are subject to the superposition principle of QM, whereas imperfections in the coherence are described by a density matrix with (partially) classical correlations. Decoherence stands for the loss of coherence as induced, for example, by coupling an isolated quantum system to its environment.

Correspondence Principle: The principle that quantum-mechanical and classical description of physical phenomena should become equivalent in the limit of vanishing de Broglie wavelength $\lambda \to 0$, or equivalently, of large quantum numbers.

de Broglie Wave: According to de Broglie, the motion of a particle of mass M and velocity v is associated with a matter wave with wavelength $\lambda = h/(Mv)$. h is Planck's quantum.

Delta (δ) Function: A particular case of a distribution, frequently used in quantum mechanics to formulate the orthogonality and normalization conditions of states with a continuous spectrum of eigenvalues.

Distributions: Also called *generalized functions*. Linear-continuous functionals operating on a metric vector space. They form the dual space of the metric vector space. In the context of quantum mechanics, the space is a Hilbert space and the operation of the functional on an element of the Hilbert space is called *formation of a wave packet*.

Eigenvalue Problem: The solution of the eigenvalue equation for the matrix A, $A\psi = \lambda\psi$ where λ is a number called an *eigenvalue* and ψ is the *eigenvector*. In quantum mechanics, ψ is a vector in Hilbert space.

Entanglement: An intrinsic feature of QM, occurring when particles or subsystems (in either case more than one), are correlated beyond a degree allowed by classical physics. Entangled systems, even when far apart, cannot be described independently of each other.

Hamilton–Jacobi Equation: A particular canonical transformation in classical dynamics that leads to the vanishing of the Hamiltonian function in the new coordinates. The transformation equation for the original Hamiltonian function is called the *Hamilton–Jacobi equation* and is equivalent to equations of motion.

Hilbert Space: Infinite-dimensional complete vector space over the complex numbers endowed with a metric induced by a scalar product that satisfies Schwartz's inequality.

Laser: Light amplication by stimulated emission of radiation. Modern source for high-intensity coherent electromagnetic radiation in the infrared, visible, and ultraviolet spectral region.

Schrödinger Equation: Fundamental equation governing the dynamics of a nonrelativistic quantum system, playing a similar role in quantum mechanics as do Newton's equations of motion in classical dynamics.

State: A mechanical system is at any given time completely characterized in quantum mechanics by the state ψ. The projection onto coordinate space, $\psi(\mathbf{r})$, is called a *wave function*. States are vectors in Hilbert space.

Superposition Principle: Any superposition in Hilbert space, $\psi = a\psi_1 + b\psi_2$, of two states ψ_1 and ψ_2 forms another physically realizable state of a physical system.

Uncertainty Principle: Two canonically conjugate variables, such as position x and momentum p, can be simultaneously measured only with an intrinsic uncertainty of at least $\Delta x \Delta p \gtrsim \hbar/2$.

Wave Function: The representation of the state $|\psi\rangle$ in coordinate space, $\psi(\mathbf{r}) = \langle \mathbf{r} \mid \psi \rangle$.

References

Andrä, H. (**1974**) *Phys. Scr.*, **9**, 257–80.

Aspect, A., Dalibard, J., and Roger, G. (**1982**) *Phys. Rev. Lett.*, **49**, 91–94; 1804–807.

(a) Bell, J.S. (**1965**) *Physics (New York)*, **1**, 195–207; (b) see also Clauser, J.F., Horner, M.A., Shimony, A., and Holt, R. (**1969**) *Phys. Rev. Lett.*, **23**, 880–83.

Bennett, C.H., Brassard, G., Crépeau, C., Jozsa, R., Peres, A., and Wootters, W.K. (**1993**) *Phys. Rev. Lett.*, **70**, 1895–99.

Berry, M. (**1984**) *Proc. R. Soc. London, Ser. A*, **392**, 45–57.

Bohr, N. (**1928**) *Nature*, **121**, 580–90.

Bouwmeester, D., Pan, J.-W., Mattle, K., Eibl, M., Weinfurter, H., and Zeilinger, A. (**1997**) *Nature*, **390**, 575–79.

Born, M. (**1926**) *Z. Phys.*, **38**, 803–27.

Brandt, S. and Dahmen, H.D. (**1989**) *Quantum Mechanics on the Personal Computer*, Springer-Verlag, Berlin.

de Broglie, L. (**1924**) Thesis; reprint (1926), *J. Phys. (Paris)*, **7**, 321–337.

Burgdörfer, J. (**1981**) *Phys. Rev. A*, **24**, 1756–67.

Davisson, C.J. and Germer, L.J. (**1927**) *Phys. Rev.*, **30**, 705–40.

Dirac, P.A.M. (**1930**) *The Principles of Quantum Mechanics*, Oxford University Press, Oxford, Reprint (1986) Oxford Scientific Publications, Oxford.

Einstein, A. (**1905**) *Ann. Phys. (Leipzig)*, **17**, 132–48.

Einstein, A. (**1917**) *Verh. Dtsch. Phys. Ges.*, **19**, 82–92.

Einstein, A., Podolsky, B., and Rosen, N. (**1935**) *Phys. Rev.*, **47**, 777–80.

Ericson, T. and Mayer-Kuckuk, T. (**1966**) *Annu. Rev. Nucl. Sci.*, **16**, 183–206.

Fano, U. (**1961**) *Phys. Rev.*, **124**, 1866–78.

Feynman, R. (**1948**) *Rev. Mod. Phys.*, **20**, 367–87.

Friedman, J.R., Patel, V., Chen, W., Tolpygo, W.K., and Lukens, J.E. (**2000**) *Nature*, **406**, 43–46.

Goldstein, H. (**1959**) *Classical Mechanics*, Addison-Wesley, Reading, PA.

Goudsmit, S. and Uhlenbeck, G. (**1925**) *Naturwissenschaften*, **13**, 953–54.

Gröblacher, S., Paterek, T., Kaltenbaek, R., Brukner, C., Zukowski, M., Aspelmeyer, M., and Zeilinger, A. (**2007**) *Nature*, **446**, 871–75.

Grover, L.K. (**1996**) Proceedings, 28th Annual ACM Symposium on the Theory of Computing (STOC) 212–19.

Gutzwiller, M. (**1967**) *J. Math. Phys.*, **8**, 1979–2000.

Gutzwiller, M. (**1970**) *J. Math. Phys.*, **11**, 1791–1806.

Gutzwiller, M. (**1971**) *J. Math. Phys.*, **12**, 343–58.

Gutzwiller, M. (**1990**) *Chaos in Classical and Quantum Mechanics*, Springer-Verlag, New York.

Hasegawa, Y., Loidl, R., Badurek, G., Baron, M., and Rauch, H. (**2006**) *Phys. Rev. Lett.*, **97**, 230401.

Heisenberg, W. (**1927**) *Z. Phys.*, **43**, 172–98.

Hellmuth, T., Walter, H., Zajonc, A. and Schleich, W. (**1987**) *Phys. Rev. A*, **35**, 2532–41.

Hylleraas, E. (**1929**) *Z. Phys.*, **54**, 347–66.

Lighthill, M.J. (**1958**) *Introduction to Fourier Analysis and Generalized Functions*, Cambridge University Press, Cambridge.

Majumder, P.K., Venema, B.J., Lamoreaux, S.K., Heckel, B.R., and Fortson, E.N. (**1990**) *Phys. Rev. Lett.*, **65**, 2931–34.

Marcus, C.M., Rimberg, A.J., Westervelt, R.M., Hopkins, P.F., and Gossard, A.C. (**1992**) *Phys. Rev. Lett.*, **69**, 506–9.

Noid, D. and Marcus, R. (**1977**) *J. Chem. Phys.*, **67**, 559–67.

(a) Pekeris, C.L. (**1958**) *Phys. Rev.*, **112**, 1649–58; (b) Pekeris C.L. (**1959**) *Phys. Rev.*, **115**, 1216–21.

Rutherford, E. (**1911**) *Philos. Mag.*, **21**, 669–83.

Schwartz, L. (**1965**) *Méthodes Mathématiques pour Les Sciences Physiques*, Hermann, Paris.

Shor, P.W. (**1994**) Proceedings of the 35th annual symposium on foundations of computer science; revised version available at: (1997). *SIAM J. Comput.*, **26**, 1484–509.

(a) Stern, O. and Gerlach, W. (**1921**) *Z. Phys.*, **7**, 249–53; (b) Stern, O. and Gerlach, W. (**1921**) **8**, 110–111; (c) Stern, O. and Gerlach, W. (**1922**) **9**, 349.

Ursin, R. *et al.* (**2007**) *Nature Physics*, **3**, 481–86.

Van Vleck, J. (**1928**) *Proc. Natl. Acad. Sci. U.S.A.*, **14**, 178–88.

Weihs, G., Jennewein, T., Simon, C., Weinfurter, H., and Zeilinger, A. (**1998**) *Phys. Rev. Lett.*, **81**, 5039–43.

Wootters, W.K. and Zurek, W.H. (**1982**) *Nature*, **299**, 802–3.

Further Reading

Bell, J.S. (**1987**) *Speakable and Unspeakable in Quantum Mechanics*, Cambridge University Press, Cambridge.

Bjorken, J.D. and Drell, S.D. (**1965**) *Relativistic Quantum Mechanics and Relativistic Field Theory*, McGraw-Hill, New York.

Bransden, B.H. and Joachain, C.J. (**1983**) *Physics of Atoms and Molecules*, Longman, New York.

Burgdörfer, J., Yang, X., and Müller, J. (**1995**) *Chaos Solitons Fractals*, **5**, 1235–73.

Cohen-Tannoudji, C., Diu, B. and Laloë, F. (**1977**) *Quantum Mechanics*, Vols. I and II, John Wiley & Sons, New York.

Courant, R. and Hilbert, D. (**1953**) *Methods of Mathematical Physics*, Wiley Interscience, New York.

Eisberg, R. and Resnick, R. (**1985**) *Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles*, 2nd edn, John Wiley & Sons, New York.

Feynman, R.P. and Hibbs, A.R. (**1965**) *Quantum Mechanics and Path Integrals*, McGraw-Hill, New York.

Feynman R.P., Leighton, R.B., Sands, M. (**1965**) *The Feynman Lectures on Physics*, Vol. III, Quantum Mechanics, Addison-Wesley, Reading, MA.

Krane, K. (**1983**) *Modern Physics*, John Wiley & Sons, New York.

Landau, L. and Lifshitz, E. (**1965**) *Quantum Mechanics, Non-Relativistic Theory*, Pergamon, Oxford.

Merzbacher, E. (**1970**) *Quantum Mechanics*, John Wiley & Sons, New York.

Messiah, A. (**1961**) *Quantum Mechanics*, North-Holland, Amsterdam.

von Neumann, J. (**1955**) *Mathematical Foundations of Quantum Mechanics*, Princeton University Press, Princeton, NJ.

Schiff, L. (**1968**) *Quantum Mechanics*, McGraw-Hill, Singapore.

Sommerfeld, A. (**1950**) *Atombau und Spektrallinien*, Vieweg, Braunschweig; Reprint (1978) Verlag Harri Deutsch, Thun, Switzerland.

Wichmann E.H. (**1971**) *Berkeley Physics Course*, Quantum Physics, Vol. **4**, McGraw-Hill, New York.

QUANTUM OPTICS

Yoshihisa Yamamoto, *ERATO Quantum Fluctuation Project, Edward L. Ginzton Laboratory, Stanford University, Stanford, California, U.S.A., and NTT Basic Research Laboratories, Atsugi-shi, Kanagawa, Japan*

Introduction 307
1. Squeezing—Control of Quantum States of Light 309
1.1 Introduction 309
1.2 Brief Review of Squeezed States 309
1.3 Principle of Squeezed-State Generation in Semiconductor Lasers 311
1.4 Squeezed-State Generation Experiments 313
1.5 Poissonian Optical-Partition Noise versus Sub-Poissonian Electrical-Partition Noise 314
1.6 Quantum Correlation between Photon Number and Junction Voltage 316
1.7 Future 316
2. Cavity Quantum Electrodynamics—Control of Spontaneous Emission 318
2.1 Introduction 318
2.2 Brief Review of Atomic Cavity Quantum Electrodynamics 318
2.3 Microcavity Laser Physics 319
2.4 Fabry–Pérot Microcavities 321
2.5 Microdisk Cavities 323
2.6 Low-Threshold Microlasers ... 324
2.7 Future 324
3. Quantum Nondemolition Measurement—Control of Measurement Back Action .. 325
3.1 Introduction 325
3.2 General Quantum Measurement and QND 325
3.3 QND Measurement for Photon Number 326
3.3.1 Heisenberg Picture 327
3.3.2 Schrödinger Picture 327
3.4 Other QND Measurements 328
4. Optical-Soliton Quantum Effect—Macroscopic Quantum Effect 328
4.1 Introduction 328
4.2 Quantization of Optical Solitons 330
4.3 Quantum Nondemolition Measurements 332
4.4 Future 334
5. Outlook 335
Glossary 335
Works Cited 335
Further Reading 337

INTRODUCTION

A coherent state of light has identical noise on two quadrature components, which are cosine and sine components of the oscillating field. The product of the two quadrature noise components satisfies a (Heisenberg) minimum-uncertainty product, $\langle \Delta \hat{a}_1^2 \rangle \langle \Delta \hat{a}_2^2 \rangle = \frac{1}{16}$. Thus, a coherent state of light is a minimum-uncertainty state. A laser pumped at far above the oscillation threshold produces a coherent state of light, if the phase diffusion (Schawlow–Townes linewidth) caused by spontaneous emission is suppressed by an appropriate phase reference beam. Preparation of the phase reference beam is possible by dividing a laser output into two arms or by using an optical phase-locked-loop (PLL) technique. These techniques are standard in various optical precision measurements and communications. Indeed, the performance limits of various optical precision measurements are all attributed to the quantum noise of a coherent state of light mentioned above. For instance, when a laser intensity or phase is measured, so-called shot noise appears and determines a signal-to-noise ratio. The origin of this shot noise is the in-phase component noise $\langle \Delta \hat{a}_1^2 \rangle = \frac{1}{4}$ for an intensity measurement and the quadrature-phase component noise $\langle \Delta \hat{a}_2^2 \rangle = \frac{1}{4}$ for a phase measurement, if only

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

cosine component $\hat{a}_1$ carries a coherent (dc) amplitude. Such performance limits are often referred to as standard quantum limits (SQL).

A ground state of light ("darkest state") is also a coherent state, whose average field is zero, $\langle \hat{a}_1 \rangle = \langle \hat{a}_2 \rangle = 0$, but quadrature noise is still $\langle \Delta \hat{a}_1^2 \rangle = \langle \Delta \hat{a}_2^2 \rangle = \frac{1}{4}$. This quadrature noise of a ground state is often referred to as a quantum-mechanical zero-point fluctuation, and the ground state itself is called a vacuum state.

Electromagnetic fields with the noise on one quadrature component reduced to below the quantum-mechanical zero-point fluctuation level and the noise on the other quadrature component enhanced to above it are currently of great interest in quantum optics because of their potential applications to various precision measurements. Such quadrature squeezed states of light are usually produced by imposing nonlinear unitary evolution on coherent (or vacuum) states.

Slusher and co-workers at AT&T Bell Laboratories produced quadrature squeezed states for the first time in 1985. A most successful generation scheme for quadrature squeezed states has been a parametric amplifier. Kimble and co-workers at California Institute of Technology produced a cw squeezed state by a resonant cavity $\mathrm{LiNbO_3}$ parametric amplifier. Kumar and co-workers at Northwestern University generated a pulse squeezed state by a traveling-wave–type $\mathrm{LiNbO_3}$ parametric amplifier. An alternative generation scheme is a four-wave mixer. Levenson and co-workers at IBM Laboratories produced quadrature squeezed states by four-wave mixing in a fiber. Haus and co-workers at Massachusetts Institute of Technology observed a pulsed squeezed state in a Sagnac fiber-loop mirror.

A coherent state of light is also a minimum-uncertainty state for photon number and phase. The photon-number noise $\langle \Delta \hat{n}^2 \rangle = \langle \hat{n} \rangle$ and the phase noise $\langle \Delta \hat{\phi}^2 \rangle = 1/4\langle \hat{n} \rangle$ of a coherent state satisfy a (Heisenberg) minimum-uncertainty product, $\langle \Delta \hat{n}^2 \rangle \langle \Delta \hat{\phi}^2 \rangle = \frac{1}{4}$. Electromagnetic fields with the photon-number noise reduced to below that of a coherent state and the phase noise enhanced to above it or vice versa are called number-phase squeezed states. Number-phase squeezed states with reduced photon-number noise and enhanced phase noise are generated directly by a constant-current–driven semiconductor laser. This is the simplest scheme for the generation of nonclassical light, and so far it has yielded the largest photon-number-noise reduction. The mutual coupling between a lasing junction and an external electrical circuit provides opportunities for exploring the macroscopic and microscopic quantum effects in open optoelectronic systems.

A quadrature squeezed state has been discussed in detail in recent monographs (Meyster and Sargent, 1991; Walls and Milburn, 1994; Vogel and Welsch, 1994), and so we do not reproduce it in this article. We concentrate here on a number-phase squeezed state.

Cavity quantum electrodynamics (QED) is also an important subject of quantum optics which has been studied intensively for the last two decades. Spontaneous emission by an atom is not an immutable property of the atom but is a consequence of atom–vacuum-field coupling. The zero-point fluctuation of a vacuum field is continuous over a relevant frequency range for an atom and also isotropic in propagation direction in free space. This is why the spontaneous emission of an atom in free space is an irreversible process with a unique decay time. If an atom is enclosed by a cavity, the distributions of the vacuum-field fluctuations in frequency and space are modulated. For instance, if the atomic transition frequency is below the cutoff frequency of the cavity, the vacuum field that is responsible for spontaneous emission is expelled from the cavity and spontaneous emission is inhibited. If the atomic transition frequency is just on resonance of the cavity, the vacuum field that is responsible for spontaneous emission is resonantly trapped in the cavity and spontaneous emission is enhanced. Moreover, when the cavity has a high enough Q value (long photon lifetime), spontaneous emission can be made even a reversible process. Such cavity QED effects have been demonstrated for a two-level atomic system.

A new class of optical microresonators is making possible the exploration of such quantum-electrodynamic phenomena in condensed-matter systems and providing microlasers with a wide range of potential applications. The coupling efficiency of spontaneous emission into a lasing mode and the spontaneous emission rate can be modified by various semiconductor microcavity structures. By increasing the coupling efficiency, semiconductor lasers with a very low threshold current and semiconductor lasers and light-emitting diodes with a high quantum efficiency, broad modulation bandwidth, and low noise can be realized. Atomic cavity QED effects have been discussed in a recent monograph (Berman, 1994), and so we concentrate here on semiconductor cavity QED effects and microlasers.

Quantum nondemolition (QND) measurements are the repeated measurements of an observable without changing the free evolution of the observable. The measurement error of the observable and the back-action noise imposed on the conjugate observable satisfy the Heisenberg uncertainty relationship. The measurement back action on the measured observable can be eliminated by choosing an observable and constructing a measuring device appropriately. The concept of QND measurements was created for improving the sensitivity of gravitational wave detection, but various QND measurement schemes have been actively studied in quantum optics.

The quantum-mechanical properties of long-lived optical pulses in a fiber known as solitons have recently been studied theoretically and observed experimentally. These experiments suggest effective ways of reducing quantum noise (squeezing) in signal transmission and of performing quantum-mechanical nondemolition measurements without destroying the signal. We discuss such soliton quantum effects with special emphasis on squeezing and QND measurements.

1. SQUEEZING—CONTROL OF QUANTUM STATES OF LIGHT

1.1 Introduction

Quantum statistical properties of laser light have been extensively studied during the last 30 years. A laser is a nonequilibrium open system. A nonequilibrium cavity internal field and population inversion, which are referred to as "systems," are coupled to thermal-equilibrium cavity external fields and pump source, etc., which are referred to as "reservoirs." Separation of system dynamics from reservoir dynamics is usually employed in the theoretical analysis of such an open system. Indeed, a quantum-mechanical reservoir theory has been successfully applied to a laser (Haken, 1970; Sargent *et al.*, 1974; Louisell, 1973). Coherent states of light, which feature quantum-mechanical zero-point fluctuations around the average field excitation, have also been extensively studied in quantum optics (Glauber, 1963). The ordinary shot noise usually observed on direct detection of the laser output light has been attributed to the quantum-mechanical zero-point fluctuations. An ideal laser pumped at well above the oscillation threshold generates a coherent state of light with the exception of random-walk phase-diffusion noise (Loudon, 1973). Various experimental facts, such as Poissonian photoelectron statistics, shot-noise–limited photocurrent spectra, and Gaussian homodyne detection output statistics, supported the above conclusion. It was also suggested that the quantum statistical properties of a semiconductor laser are essentially the same as those of other lasers (Haug and Haken, 1967). However, a semiconductor *pn*-junction laser converts electrons injected via an electrical current to coherent photons by stimulated emission.

Recent theoretical and experimental studies on quantum-noise properties of a semiconductor laser have revealed that the semiconductor *pn*-junction laser produces photon-number squeezed states with photon-number noise reduced to below the standard quantum limit (SQL) $\langle \Delta\hat{n}^2 \rangle = \langle \hat{n} \rangle$ (Yamamoto *et al.*, 1992). The conventional quantum reservoir theory cannot describe the result. We must take into account the mutual coupling between the pump electrical circuit and the lasing junction. This generation scheme of photon-number squeezed states is in sharp contrast to the conventional means of generating squeezed states of light, in which coherent states produced by a laser are unitarily transformed to squeezed states by various nonlinear optical processes (Kimble and Walls, 1987).

1.2 Brief Review of Squeezed States

The Heisenberg uncertainty principle for two conjugate observables $\hat{O}_1$ and $\hat{O}_2$ is the direct consequence of the commutation relation between them (Messiah, 1961). It can be derived by the Schwartz inequality

$$[\hat{O}_1,\hat{O}_2] = i\hat{O}_3 \Rightarrow \langle \Delta\hat{O}_1^2\rangle\langle \Delta\hat{O}_2^2\rangle \geq \tfrac{1}{4}\,|\langle \hat{O}_1,\hat{O}_2\rangle|^2 = \tfrac{1}{4}\,\langle \hat{O}_3\rangle^2, \tag{1}$$

where $\Delta\hat{O}_i = \hat{O}_i - \langle \hat{O}_i\rangle$, $i = 1,2$. A state that exactly satisfies the equality in Eq. (1) is called a minimum-uncertainty state and is mathematically defined as an eigenstate of the operator

$$\hat{O}(r) = e^r\hat{O}_1 + ie^{-r}\hat{O}_2, \tag{2}$$

where r is a squeezing parameter. That is, r determines the distribution of quantum noise between $\hat{O}_1$ and $\hat{O}_2$:

$$\begin{aligned}\langle \Delta\hat{O}_1^2\rangle &= \langle \Delta\hat{O}_1^2\rangle_{r=0}e^{-2r},\\ \langle \Delta\hat{O}_2^2\rangle &= \langle \Delta\hat{O}_2^2\rangle_{r=0}e^{2r}.\end{aligned} \tag{3}$$

A quadrature-squeezed state of the electromagnetic field is a minimum-uncertainty state, in which $\hat{O}_1$ and $\hat{O}_2$ correspond to the two quadrature phase components $\hat{a}_1 = \frac{1}{2}(\hat{a} + \hat{a}^\dagger)$ and $\hat{a}_2 = (1/2i)(\hat{a} - \hat{a}^\dagger)$. The term $\hat{O}_3$ is then a c-number ($= \frac{1}{2}$). The squeezed state features reduced quantum noise in one quadrature and enhanced quantum noise in the other quadrature. In the special case of $r = 0$, the two quadrature components share the same amount of quantum noise, namely, $\langle \Delta\hat{a}_1^2\rangle = \langle \Delta\hat{a}_2^2\rangle = \frac{1}{4}$. This is a characteristic of a coherent state (Glauber, 1965). The electromagnetic field generated by a classical oscillating current is a coherent state (Glauber, 1965). Its quadrature-component quantum noise is sometimes referred to as the standard quantum limit (SQL). In order ultimately to reduce the noise in one of the quadratures to zero, the electromagnetic-field mode must have an infinite photon number and thus an infinite energy. This is because the enhanced quadrature noise component will require part of the total photon number (Yuen, 1976):

$$\langle \hat{n}\rangle_{SS} = \langle \hat{a}_1\rangle^2 + \langle \hat{a}_2\rangle^2 + \sinh^2(r) \underset{(r\to\infty)}{\longrightarrow} \infty. \tag{4}$$

That is, there is a trade-off relation between quantum-noise reduction and the required photon number.

A number-phase squeezed state of the electromagnetic field is defined as

$$\hat{O}_1 = \hat{n} \equiv \hat{a}^\dagger\hat{a}, \tag{5}$$

$$\hat{O}_2 = \hat{S} \equiv (1/2i)[(\hat{n} + 1)^{-1/2}\hat{a} - \hat{a}^\dagger(\hat{n} + 1)^{-1/2}], \tag{6}$$

and

$$\hat{O}_3 = \hat{C} \equiv \tfrac{1}{2}[(\hat{n} + 1)^{-1/2}\hat{a} + \hat{a}^\dagger(\hat{n} + 1)^{-1/2}], \tag{7}$$

where $\hat{n}$, $\hat{S}$, and $\hat{C}$ are the photon-number operator, the sine operator, and the cosine operator, respectively (Jackiw, 1968). When r is equal to $-\frac{1}{2}\ln(2\langle \hat{n}\rangle)$ and the average photon number is much larger than one, the photon-number noise and the sine-operator noise (which corresponds to the classical phase noise) are reduced to $\langle(\Delta\hat{n})^2\rangle = \langle \hat{n}\rangle$ and $\langle(\Delta\hat{S})^2\rangle = \langle \hat{C}\rangle^2/4\langle n\rangle \simeq \frac{1}{4}\langle \hat{n}\rangle$. This photon-number and sine-operator noise is also referred to as the SQL. A number-phase squeezed state with r greater than $-\frac{1}{2}\ln(2\langle \hat{n}\rangle)$ features smaller photon-number noise and larger sine-operator (phase) noise than the SQL, while the minimum-uncertainty relationship $\langle(\Delta\hat{n})^2\rangle\langle(\Delta\hat{S})^2\rangle = \frac{1}{4}\langle \hat{C}\rangle^2$ is preserved. The photon-number noise can ultimately be reduced to zero (photon-number state) without requiring an infinite photon number, because the enhanced phase noise does not consume energy at all.

The above various photon states are described by the quasi-probability density (Q representation)

$$Q(\alpha) \equiv \langle \alpha|\hat{\rho}|\alpha\rangle = |\langle \alpha|\psi\rangle|^2, \tag{8}$$

where $|\psi\rangle$ is a state vector, $\hat{\rho}$ is a density operator, and $|\alpha\rangle$ is a coherent state with the complex eigenvalue α, and the real and imaginary parts of α correspond to $\langle \hat{a}_1\rangle$ and $\langle \hat{a}_2\rangle$, respectively. Neither a squeezed state or a number-phase squeezed state possesses a positive-definite diagonal $P(\alpha)$ function (Glauber, 1963) for the density operator, which implies that these states cannot be realized by a classical mixture of coherent states. The term "nonclassical photon states" for characterizing squeezed states has been coined because of this fact. Figure 1 compares the quasi-probability densities $Q(\alpha)$ of various quantum states of light.

An unmistakable mark of squeezed-state generation is the observation of quadrature amplitude squeezing. That is, a homodyne-

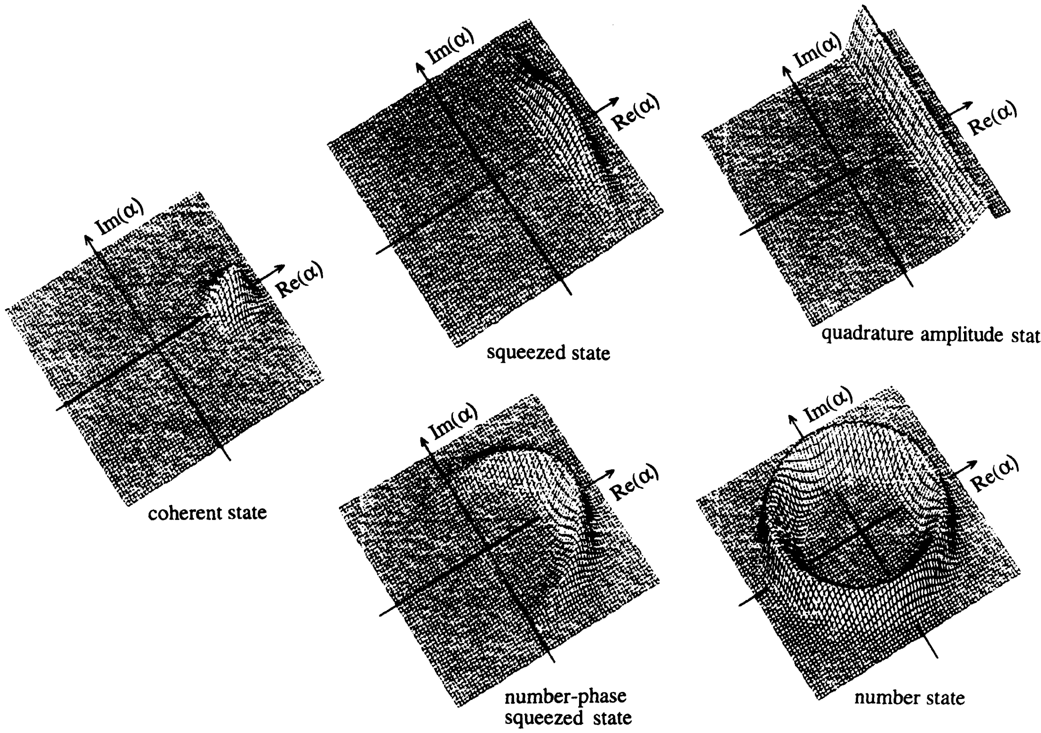

FIG. 1. Quasi-probability densities $Q(\alpha)$ of various quantum states of light.

detector output current will feature a reduced noise level below that for vacuum fields (Slusher *et al.*, 1985; Shelby *et al.*, 1986; Wu *et al.*, 1986). An unmistakable mark of number-phase squeezed-state generation, on the other hand, is the observation of amplitude squeezing. That is, a photon-counting detector output will feature a noise below the shot-noise level and thus sub-Poissonian photon statistics.

1.3 Principle of Squeezed-State Generation in Semiconductor Lasers

The macroscopic coherence of light and matter in a laser is established by the balance between the ordering force of "systems" (gain saturation in the case of a laser) and fluctuating forces from external "reservoirs." The conventional quantum-mechanical reservoir theory for a laser treats only three degrees of freedom as "systems": cavity internal field, population inversion, and dipole moment. All of the other degrees of freedom are considered as "reservoirs" and eliminated by use of the quantum-mechanical fluctuation–dissipation theorem.

A fluctuating force pertinent to our discussion is that from an external pump source. The pump fluctuating force for optically pumped lasers is usually shot-noise limited, because the pump light absorption by a gain medium is considered as a random Poisson point process. Of course, this is not a fundamental limit imposed by quantum mechanics. We may consider a thought experiment in which a gain medium is excited by photon-number states of light with a perfect absorption efficiency. The pump fluctuating force determined by the photon-number noise of pump light is reduced to zero for such a case (Golubev and Sokolov, 1984). The phase noise of pump light does not contribute to the pump noise of a laser at all, because a laser oscillator is an incoherently pumped device, in contrast to a parametric oscillator, which is a coherently pumped device.

The photon-number noise spectra for a pump-noise–suppressed laser are reduced to below SQL at frequencies below the cavity cutoff frequency at pump rates well above the threshold. The physical interpretation for the result is rather straightforward. Every absorbed pump photon is emitted sooner or later as a coherent lasing photon from the cavity, in which the stimulated emission by a strong laser field is much stronger than a random spontaneous emission at high pump rates. Therefore, if the emitted photon number is counted for a measurement time interval longer than the population lifetime and the photon lifetime, the photon statistics of the laser output should be the same as the photon statistics of the pump light. A high conversion efficiency is indeed only one criterion for such preservation of the absorbed pump-photon statistics to the emitted coherent photon statistics.

However, the phase noise spectra are increased to above SQL at frequencies below the cavity cutoff frequency because of the so-called Schawlow–Townes phase-diffusion process. The product of reduced photon-number spectrum $S_{\Delta N}(\Omega)$ and increased phase spectrum $S_{\Delta\psi}(\Omega)$ satisfies the Heisenberg uncertainty principle,

$$S_{\Delta N}(\Omega) S_{\Delta\psi}(\Omega) \geq 1. \tag{9}$$

In the limit of high pump rates, the product is higher by only a factor of 2 over the minimum-uncertainty product. A finite photon lifetime introduces the residual photon-number noise $S_{\Delta N}(\Omega)$ by a stochastic photon output process, and at the same time, it stores coherent photons inside the cavity, which leads to the stabilization of the phase noise $S_{\Delta\psi}(\Omega)$ by stimulated emission. This result suggests that the output field from a pump-noise–suppressed laser is not far from the number-phase squeezed state shown in Fig. 1. Similar states are also produced by either the quantum phase diffusion in a Kerr nonlinear medium (Kitagawa and Yamamoto, 1986) or the state reduction induced by a quantum nondemolition measurement (Kitagawa *et al.,* 1987) or by an idler photon-number measurement for a parametric amplifier output (Watanabe and Yamamoto, 1988).

Several schemes have been proposed to suppress the pump fluctuating force of a laser. One scheme utilizes sub-Poissonian space-charge–limited electron beams to obtain the Franck–Hertz effect (Teich and Saleh, 1985). Another utilizes three-level atoms as a gain medium, in which the pump process is self-regulated by a limited number of ground-state atoms (Bergou *et al.,* 1989; Haake *et al.,* 1989). The most practical and direct way of suppressing the pump fluctuating force is to use a constant-current source for a semiconductor injection laser.

The junction-current noise of a *pn* diode driven by a constant-voltage source is shot-noise limited. This fact has been erroneously identified as a "shot-noise–limited pump fluctuating force" for a semiconductor laser. The mistake in this argument is twofold. First, the junction-current noise is not a pump fluctuating force (cause) but a relaxation-current pulse (result) for the radiative recombination (photon emission) event in a junction. Second, a semiconductor laser usually operates under a constant-current mode rather than a constant-voltage mode.

The difference between constant-voltage mode and constant-current mode is schematically shown in Fig. 2. When a *pn* junction is driven by a constant-voltage source with a source resistance R_S much smaller than the diode's differential resistance R, that is, $R_S \ll R = k_B T/eI$, where k_B is the Boltzmann constant and T is temperature, the junction-voltage drop resulting from the radiative recombination of an electron–hole pair is quickly eliminated by an external circuit relaxation-current pulse. By the time the next radiative recombination event occurs, the junction voltage (number of electron–hole pairs in a junction) has recovered to the original value. In such a memoryless system, photon emission events follow a Poisson point process and the junction current carries full shot noise. On the other hand, when a *pn* junction is driven by a constant-current source with a source resistance much larger than the diode's differential resistance, that is, $R_S \gg R$, the junction voltage is not pinned by a source but is free to fluctuate. The fluctuation in the junction voltage caused by the radiative recombination has not recovered at the time of the next radiative recombination event. In such a system with memory, the photon emission events follow a sub-Poissonian process and the junction current carries noise below the shot-noise level because of self-stabilization.

The pump fluctuating force of a semiconductor laser is identified as the thermal noise

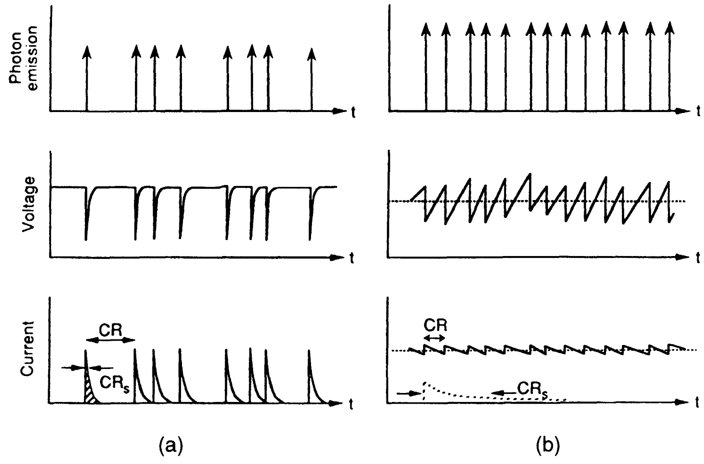

FIG. 2. The junction-voltage fluctuation and the relaxation-current pulse caused by a photon emission event for (a) $R \gg R_s$ (constant-voltage driving case) and (b) $R \ll R_s$ (constant-current driving case). C is the diffusion capacitance of a *pn* junction. The time constant CR corresponds to the electron lifetime, and CR_s is the external circuit relaxation time constant.

generated in the source resistance R_S, whose spectral density is $4k_BT/R_S$. The thermal noise is much higher than the shot-noise value $2eI$ for a constant-voltage source with $R_S \ll R$, but is much smaller than the shot-noise value for a constant-voltage source with $R_S \gg R$. The thermal-noise–limited pump fluctuating force and the dissipation rate of the junction voltage through the pump circuit satisfy the fluctuation–dissipation theorem exactly (Yamamoto *et al.*, 1991). In the case of a constant-current source, the dissipation rate $(CR_S)^{-1}$ of the junction voltage through a pump circuit is very slow and the fluctuating force from the pump circuit $4k_BT/R_S$ is also small. The two conditions necessary for the generation of photon-number squeezed states in semiconductor lasers, that is, the free voltage fluctuation and the pump-noise suppression, are satisfied simultaneously by $R_S \gg R$.

There is an analogy between the quantum states established in a *pn*-junction laser and in a superconductor tunnel junction. A Bloch oscillation at a frequency $f = I/2e$ is achieved in an ultrasmall superconductor tunnel junction driven by a constant-current source I, in which the number of tunneled electron pairs is regulated but the relative phase of the two superconductors is randomized (Likharev and Zorin, 1985). This is similar to the present scheme for the generation of photon-number squeezed states in a *pn*-junction laser driven by a constant-current source. On the other hand, an ordinary Josephson oscillation at a frequency $f = 2eV/h$ (h is Planck's constant) is observed in a macroscopic superconductor tunnel junction driven by a constant-voltage source V, in which the relative phase of two superconductors is fixed, but the number of tunneled electron pairs is random. This is similar to coherent-state generation in a conventional laser oscillator.

1.4 Squeezed-State Generation Experiments

Figure 3(a) shows a dual-balanced detector setup for measuring the photon-number noise spectrum from a semiconductor laser. A semiconductor laser output is split equally into the two arms by a half-wave plate and polarization beam splitter. One detector output current I_1 is delayed by a coaxial cable and combined with the other, undelayed, detector output current I_2 by a differential amplifier or 180° hybrid. For a fluctuation frequency Ω_{in} satisfying $\Omega_{\text{in}}\tau = 2N\pi$, where τ is the delay time and N is an integer, the differential amplifier output $I_1 - I_2$ calibrates the shot-noise value. For a fluctuation frequency Ω_{out} satisfying $\Omega_{\text{out}}\tau = (2N + 1)\pi$, the differential amplifier output $I_1 + I_2$ measures the photon-number noise of the laser output. Thus, the differential amplifier output simultaneously displays the laser noise and corresponding shot-noise values with a frequency period of $\Delta\Omega = 2\pi/\tau$ on a spectrum analyzer.

The measured noise spectra at two different pump rates for a GaAs transverse-junction stripe semiconductor laser are shown in Fig. 3(b). The noise spectra are normalized

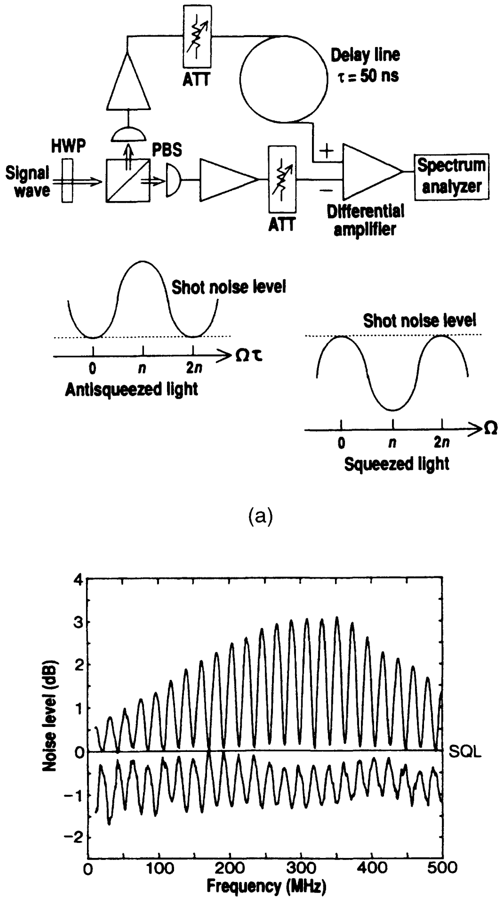

(a)

(b)

FIG. 3. (a) A dual–balanced-detector setup for measuring the intensity noise of a semiconductor laser: HWP, half-wave plate; PBS, polarization beam splitter; and ATT, attenuator and equalizer. (b) The measured intensity-noise spectra normalized by the shot-noise value for two pump rates, I/I_{th} = 1.03 (upper curve) and I/I_{th} = 13.6 (lower curve). The solid line represents the standard quantum limit (SQL).

by the shot-noise value, which is independently calibrated by a light-emitting diode with the same wavelength as the laser. The noise spectrum at the pump rate just above the threshold, $I/I_{th} = 1.03$, shows a lower noise value at Ω_{in} and a higher noise value at Ω_{out}, where I_{th} is the threshold current. This is a characteristic of antisqueezed (super-Poissonian) light. The noise spectrum at the pump rate well above the threshold, $I/I_{th} = 13.6$, shows a higher noise value at Ω_{in} and a lower noise value at Ω_{out}, which is a characteristic of squeezed (sub-Poissonian) light.

The degree of squeezing is degraded by the Poissonian partition noise associated with optical loss. To increase the light-collection efficiency, a GaAs transverse-junction stripe laser, with less than 10% front-facet reflectivity R_1 and more than 90% rear-facet reflectivity R_2, is operated at 66 K to minimize the free-carrier absorption loss and is directly coupled to a Si photodetector to minimize the coupling loss. Figure 4(a) shows theoretical and experimental photon-number noise values normalized by the shot-noise value vs the normalized pump rate $I/I_{th} - 1$. The experimental results are corrected for a detection quantum efficiency of about 89%, so that the degree of squeezing shown in Fig. 4 corresponds to that of the laser output. The experimental maximum degree of squeezing, −14 dB (≃0.04), is in reasonable agreement with the theoretical limit imposed by the output coupling efficiency,

$$\eta = \frac{\ln(1/R_1)}{\ln(1/R_1R_2)} \simeq 0.97, \tag{10}$$

of the laser.

1.5 Poissonian Optical-Partition Noise versus Sub-Poissonian Electrical-Partition Noise

If a photon-number state of light is attenuated by an optical loss, random deletion of photons introduces a so-called partition noise. In the limit of large attenuation, photon statistics approach a Poisson distribution. This is often referred to as "Poissonian partition noise."

The overall quantum efficiency [(detector output current)/(laser driving current)] in the experiment demonstrated in Fig. 4 was only 48%. If all loss processes that are responsible for this quantum efficiency could commonly introduce a Poissonian partition noise, squeezing of more than 3 dB would be impossible. However, the observed degree of squeezing was 8.5 dB. An explanation of the counterintuitive experimental result has been worked out. The GaAs laser used in the experiment has a nonlasing junction in parallel with the lasing junction [Fig. 4(b)]. Under the operating conditions of $T \simeq 66$ K and $I \simeq 30$ mA, the leakage current is estimated to be approximately equal to the current in-

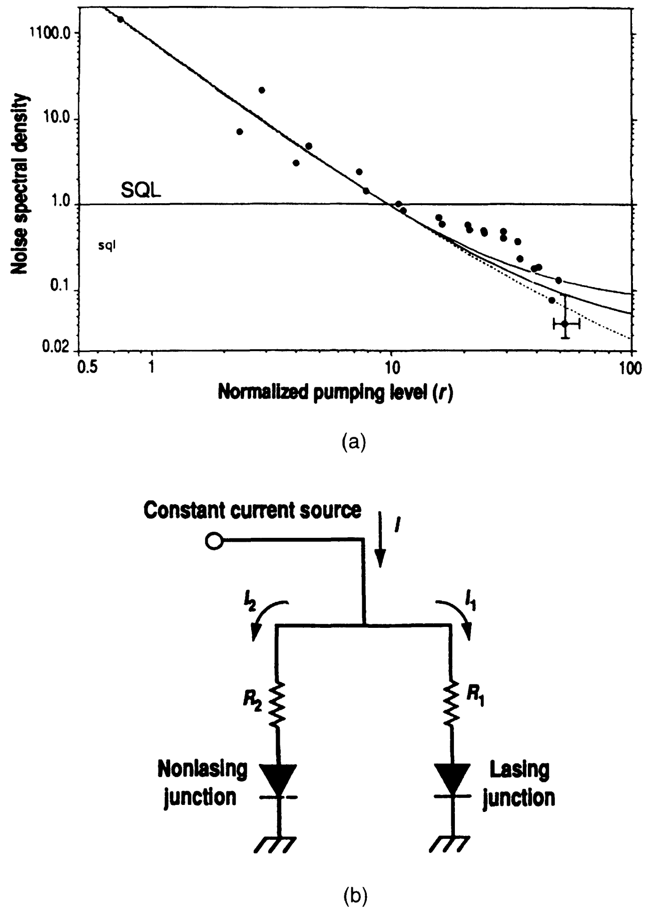

FIG. 4. (a) The theoretical (lines) and experimental (dots) intensity-noise values normalized by the shot-noise value vs the normalized pump rate $R \equiv I/I_{th} - 1$. The experimental results are corrected for a detection quantum efficiency of 89%. The three curves correspond to different internal absorption losses (the dotted line represents an ideal case of no absorption). (b) A model for current division into a lasing junction and a nonlasing junction.

jected into the lasing junction. That is, the overall quantum efficiency is degraded mainly by current division. This current division introduces a new noise source, which serves as a pump fluctuating force for the lasing junction.

A unified circuit theory has been developed for treating optical-partition noise and electrical-partition noise in the same circuit language. The current noise in two branching arms driven by a constant-current source is given by

$$i_1 = -i_2 = \frac{v_2 + v_{d2} - v_1 - v_{d1}}{2(R + R_d)}, \tag{11}$$

where it is assumed for simplicity that the two series resistances connected to the two (lasing and nonlasing) junctions are equal, $R_1 = R_2 = R$, and that the differential resistances of the two junctions are also equal, $R_{d1} = R_{d2} = R_d$. Here, v_i and v_{di} are the noise voltages associated with R and R_d. The spectra of i_1 and i_2 are given by

$$S_{i_1} = S_{i_2} = \begin{cases} 2k_BT/R & (R \gg R_d), \\ \frac{1}{2}eI & (R \ll R_d). \end{cases} \tag{12}$$

Here the quantum zero-point fluctuation $2\hbar\Omega R$ associated with R is neglected because it is much smaller than the thermal noise $4k_BTR$. Thus, if $R \ll R_d$, the current noise is only 3 dB below the shot-noise value. That is, current division introduces a Poissonian electrical-partition noise. On the other hand, if $R \gg R_d$, the current noise is much smaller than the shot-noise value. This condition is satisfied at a high current level $I \gg 2k_BT/eR$ for any small series resistance R. In this way, the sub-Poissonian character of a constant-current source is preserved against current division. The theoretical curve in Fig. 4(a) assumes a series resistance $R = 50\ \Omega$. In the case of photon flux division, the quantum zero-point fluctuation $2\hbar\omega R$ is much larger than the thermal noise $4k_BTR$ for which a Poissonian partition noise is always associated. Although such a macroscopic (circuit) theory successfully explains the experimental results, a microscopic theory, that is, a quantum theory of transport and division of charged particles in conducting wires, has yet to be worked out.

1.6 Quantum Correlation between Photon Number and Junction Voltage

The product of the reduced photon-number noise and the increased phase noise is generally higher than the minimum uncertainty product, $S_{\Delta N}(\Omega)S_{\Delta\phi}(\Omega) = 1$. The increased phase noise due to a random-walk phase diffusion is a noise inherent in any negative-conductance oscillators that do not have a restoring force for phase. On the other hand, there is a possibility that there is a hidden measurement for photon number in semiconductor lasers, which could lead to a further reduction in photon-number noise to the minimum value required by the uncertainty product.

In a constant-current–driven semiconductor laser, the junction-voltage fluctuation is related to the electron-number fluctuation through a junction capacitance. The electron-number fluctuation and the photon-number fluctuation are negatively correlated through the spontaneous emission, stimulated emission, and absorption. Consequently, it is expected that the photon-number noise can be determined by measuring the junction-voltage noise.

An experimental setup for measuring the quantum correlation between photon-number noise and junction-voltage noise is shown in Fig. 5(a). The photon-number noise of a semiconductor laser is measured by detector D2, and the shot-noise value is calibrated by a light-emitting diode and detector D1. The junction-voltage noise is measured independently and combined with the delayed photon-number noise by a wideband 180° hybrid. Figure 5(b) shows the measured junction-voltage noise (trace a), laser-intensity noise (trace b), shot noise (trace c), and combined noise of junction voltage and laser intensity (trace d). The sinusoidal variation shown by trace d manifests the correlation between the photon-number noise and the junction-voltage noise. In the absence of the correlation, trace d would be flat and the noise power numerically equal to the sum of the two noise powers given by traces a and b. The squeezed intensity noise is further reduced by the adding of the junction-voltage noise, which indicates that the correlation is negative and penetrates into a quantum region.

The junction voltage can be used as a quantum measurement probe of the photon number. Conditioned on the junction-voltage measurement result, the photon-number noise can be reduced to nearly the minimum value dictated by the uncertainty principle. This fact can be interpreted from the viewpoint of a "quantum watchdog effect" (Wheeler and Zurek, 1983). That is, the photon number is always monitored by the external pump circuit (the watchdog), even though an experimentalist does not actually read out the junction voltage. Partial (or perfect) reduction of a wave packet continuously occurs irrespective of readout actions made by the experimentalist. Such a hidden measurement introduces back-action noise on the observable conjugate to the measured observable. This inevitable back-action noise emerges as Schawlow–Townes phase-diffusion noise caused by the dephasing of dipole moments, that is, random-phase spontaneous emission.

1.7 Future

Photon-number squeezed states cannot be directly used to improve interferometric phase measurements such as a gravitational-wave–detection laser interferometer and a laser gyroscope. Such measurements require a reduced phase noise rather than a reduced photon-number noise, or more generally they require squeezed vacuum states. It is possible to change photon-number squeezed states into squeezed vacuum states (Lai *et al.,* 1991). A constant-current–driven semiconductor laser is injection-locked by an external master-laser oscillator. The phase correlation between the squeezed slave-laser output and the master-laser output is established without sacrificing appreciably the de-

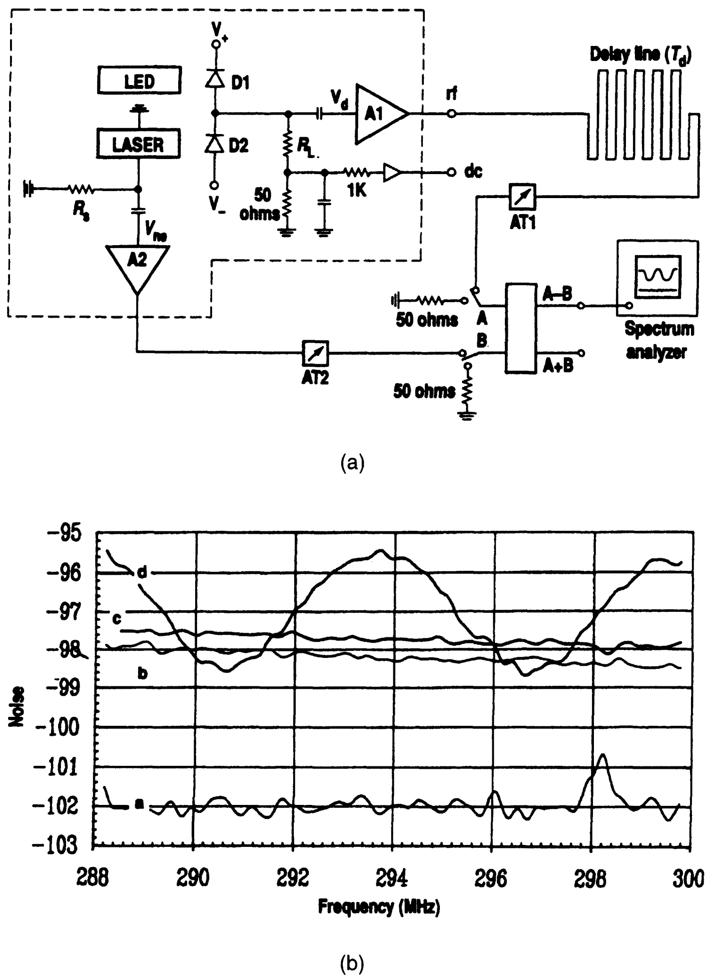

FIG. 5. (a) An experimental setup for measuring the quantum correlation between the photon-number noise and the junction-voltage noise; LED, light-emitting diode. (b) The measured junction-voltage noise (trace a), laser-intensity noise (trace b), shot noise produced by a light-emitting diode (trace c), and combined noise of junction voltage and laser intensity (trace d). The ordinate is the noise power P_n expressed as $(\mathrm{dB}_m) \equiv 10 \log_{10} P_n$ (nW). A1, A2: Amplifiers. AT1, AT2: Attenuators.

gree of photon-number squeezing. The classical signal part of the squeezed output from the slave laser can be canceled interferometrically; here, the quantum noise of the master-laser output can be suppressed with the use of a high-transmission mirror. Because a semiconductor laser emission wavelength can be chosen at any wavelength region from 0.45 to 1.7 μm, this squeezed vacuum-state generation technique can be applied to any laser interferometer system.

Photon-number squeezed states can be generated only at pump rates well above the threshold, because constantly injected electron–hole pairs are spontaneously emitted mainly into nonlasing modes at pump rates below and near the threshold. Random deletion due to such spontaneous emission introduces a Poissonian partition noise, which is even amplified at pump rates near the threshold. However, it is possible to confine all spontaneous emission into a single lasing mode by using a cavity quantum-electrodynamic technique, as described in the next section. The photon-number spectra for such a microcavity semiconductor laser with a unity coupling efficiency of spontaneous emission are reduced to below SQL at any pump rate. Only one criterion for generating photon-number squeezed states in a constant-current–driven semiconductor laser, a high conversion efficiency from electrons to photons, is satisfied by the microcavity without requiring the stimulated emission. Low-intensity squeezed light generated by such a light-emitting diode may find applications to the optical interconnection inside a large-scale integrated circuit, in which the energy dissipation for communication is already much larger than that for signal processing.

If we can decrease the current level I to less than e/τ_{sp} and satisfy the constant-current operation condition, where τ_{sp} is the spontaneous-emission lifetime, the charging time for one electron–hole pair into a junction can become longer than the radiative recombination lifetime. In such a case, coexistence of continuous uniform charging into the junction and discrete radiative recombination of an electron–hole pair will achieve the junction-voltage oscillation and single-photon emission at a frequency of $f = I/e$ (Imamoglu and Yamamoto, 1994). This is an optical analog of the Coulomb blockade, or single-electron (electron-pair) tunneling oscillation in an ultrasmall-capacitance tunnel junction. It is expected to open up a new field of combined single electronics and single photonics.

2. CAVITY QUANTUM ELECTRODYNAMICS—CONTROL OF SPONTANEOUS EMISSION

2.1 Introduction

Spontaneous emission is not a fixed property of an atom but a consequence of atom–vacuum-field coupling. It has been known for some time that the decay rate, transition energy, and radiation pattern of spontaneous emission can be altered by modifying vacuum-field fluctuations by a cavity wall. The principle, often referred to as cavity quantum electrodynamics (cavity QED), is a classic theoretical problem. Cavity-enhanced and -suppressed spontaneous decay rates were predicted by Purcell (1946). A cavity-induced radiative energy shift was pointed out by Casimir (Casimir and Polder, 1948). The nonlinear coupling of atom and vacuum fields was formulated by Jaynes and Cummings (1963), in which it was predicted that spontaneous emission becomes even reversible if an atom is put in a high-Q single-mode cavity. However, those effects have been observed only recently using either Rydberg atoms in microwave superconductor cavities or atoms in optical microcavities (Haroche and Kleppner, 1989, and references therein). The only exception is the pioneering work by Drexhage (1974), in which he demonstrated that the decay rate and radiation pattern of spontaneous emission are modified by a metal mirror.

Since spontaneous emission is a major source of energy loss, speed limitation, and noise in semiconductor optical devices, the possibility of employing cavity QED to control spontaneous emission is expected to improve the device performances fundamentally.

In this section, we review the principle of cavity QED, the performance of various microcavity structures, and the applications to semiconductor optical devices, such as light-emitting diodes and semiconductor lasers.

2.2 Brief Review of Atomic Cavity Quantum Electrodynamics

There are two distinct regimes in atomic cavity QED, i.e., a low-Q regime and a high-Q regime. A decisive parameter separating the two regimes is the vacuum Rabi frequency,

$$\Omega_R = \frac{p_{12}}{\hbar}\mathscr{E}_{\text{vac}} = \frac{p_{12}}{\hbar}\sqrt{\frac{\hbar\omega}{2\epsilon_0 V_0}}, \tag{13}$$

where p_{12} is an atomic dipole moment, $\mathscr{E}_{\text{vac}}$ is a vacuum-field amplitude, and V_0 is an optical cavity volume. When Ω_R is much smaller than the escape rate of a photon from the cavity and the decay rate of an atomic dipole moment, i.e., $\Omega_R \ll \gamma_{\text{photon}}, \gamma_{\text{dipole}}$, the atom–vacuum-field coupling is only weakly perturbed by the cavity and spontaneous emission remains an irreversible and incoherent process. Nevertheless, a cavity can modify the radiation pattern and decay rate. Spontaneous-emission rate is either inhibited or enhanced.

On the other hand, when $\Omega_R \gg \gamma_{\text{photon}}, \gamma_{\text{dipole}}$, the atom–vacuum-field coupling is strongly perturbed by the cavity and spontaneous emission becomes a reversible and coherent process. A coupled atom–field system, often referred to as a "dressed state," has two split eigenfrequencies. This vacuum Rabi splitting corresponds to a periodic energy exchange between the atom and field.

When N atoms collectively couple with the cavity field, the atom–field coupling strength is enhanced in proportion to $\sqrt{N}$. In a low-Q regime, the cavity enhances a so-called cooperative superradiant decay (Dicke, 1954). The decay rate is enhanced and the emission pattern is highly concentrated because of the formation of a gigantic dipole moment. In a high-Q regime, it corresponds to a collective Rabi splitting. The difference between the two eigenfrequencies is increased to $\Omega_R\sqrt{N}$ from Ω_R for a single-atom case.

Enhanced and inhibited spontaneous-emission rates at microwave frequencies have been demonstrated with either Rydberg atoms or cyclotron electrons. Enhanced and inhibited spontaneous emission rates at optical frequencies have also been demonstrated with either dye molecules or Rb atoms. Collective Rabi splitting has been demonstrated at both microwave and optical frequencies. Cavity QED experiments in semiconductors require a microcavity structure with an optical wavelength size (Yamamoto and Slusher, 1993, and references therein). This is because the spontaneous-emission spectral width in semiconductors is much broader than that of an atom in vacuum as a result of a very short dephasing time and a very large inhomogeneous broadening of an electron–hole pair dipole. The cavity structures used for cavity-QED experiments in semiconductors are quite similar to those utilized in vertical-cavity surface-emitting lasers. In fact, a major thrust in the study of cavity-QED effects in semiconductors is the development of small-sized, ultra-low-threshold, high-speed lasers and light-emitting diodes for applications in optical communication, parallel optical signal processing, and optical computing.

Because spontaneous emission is a major source of energy loss, speed limitation, and noise in lasers, the capability to control spontaneous emission is expected to improve laser performance. New phenomena are expected if the radiation pattern is highly directed toward a lasing mode and the fraction of spontaneous emission coupled into the lasing mode is made close to 1. The fraction of the total spontaneous emission that is coupled into the single lasing mode at low excitation rates is defined to be the spontaneous-emission coefficient β. An example of the new phenomena expected as β approaches 1 is the "thresholdless laser" (DeMartini and Jacobovitz, 1988), in which the light output increases nearly linearly with pump power instead of exhibiting a sharp turn-on at a pump threshold.

Consider a single atom centered in an optical cavity between two confocal mirrors, each with reflectivity R. This cavity radiates into a mode that subtends a solid angle $\Delta\Omega$. If the atom is located at an antinode of the vacuum field in the cavity, the spontaneous-emission rate into $\Delta\Omega$ is enhanced by a factor $4/(1 - R)$ because the vacuum field at that point is strongly enhanced relative to its free-space value. Correspondingly, when the atom is located at a node of the cavity field, the spontaneous-emission rate into $\Delta\Omega$ is inhibited by a factor of $(1 - R)/4$. If atoms are distributed over many wavelengths, strong spatial modulation of the spontaneous decay rates is not expected. Note that although the spontaneous-emission rate into the cavity mode can be strongly modified, the total spontaneous-emission rate integrated over all angles may change only a modest amount if $\Delta\Omega \ll 4\pi$.

The most important effect of modified spontaneous emission for possible applications is the altered radiation pattern and the enhanced spontaneous-emission coefficient β. The strong reflectivity dependence of the spontaneous-emission rate into a cavity mode clearly indicates that β can be increased when an atom is located at the antimode of the mode field. As the reflectivity approaches unity, the β value approaches the ideal limit of 1. Such high β values can be used to enhance the efficiency of light-emitting diodes and lower the thresholds of lasers.

2.3 Microcavity Laser Physics

The spontaneous-emission coefficient β for the relatively large-volume solid-state and gas lasers in common use today can be approximated by

$$\beta = \lambda^4/4\pi^2 V \Delta\lambda n^3. \qquad (14)$$

Here $\Delta\lambda$ is the emission linewidth, n the refractive index, and V the mode volume. This expression accounts for the number of modes in the system, which is roughly equal to the number of cubic wavelengths in the system, and the number of modes within the emission linewidth, which is proportional to

$\Delta\lambda/\lambda$. For example, even a semiconductor laser the size of a grain of salt has a β value of only 10^{-5}. One can reduce both V and $\Delta\lambda$ to obtain a larger β value. In principle, one can increase the β value to close to 1. For this purpose, one does not necessarily require a true single-mode cavity. One needs only to enhance the spontaneous-emission rate into a single mode relative to the many existing modes over the spectral-emission bandwidth, which is typically less than a few percent of the transition frequency.

One of the most important consequences of the increased spontaneous-emission coefficient is the reduction of the threshold pump rate of a laser. In a simple atomic model, where the spontaneous-emission rate and the emission spectral width are constant as functions of excitation level, the threshold pump rate as a function of β is given by (Björk and Yamamoto, 1991)

$$P_{th} = \frac{\gamma_{\text{photon}}}{2\beta}[1 + \beta + \xi(1 - \beta)]. \tag{15}$$

Here γ_{photon} is the rate of photon emission from the cavity. The quantity $\xi = N\beta V\Gamma/2\gamma_{\text{photon}}$ is the average photon number in the lasing mode when half of the atoms are in the excited state; N is the atomic density and Γ is the spontaneous-emission rate. The excitation level where the excited-atom density N_{ex} is $N/2$ is called the transparency carrier density in semiconductors, because there is no absorption and no net gain at this point. The gain G at excitation levels higher than the transparency point is assumed here to be a linear function of the excited-atom density N_{ex}: $G = \beta V\Gamma(N_{ex} - N/2)$.

The threshold pump rate in Eq. (15) includes the two distinct regions that are being studied in present semiconductor microresonator experiments, $\xi \ll 1$ and $\xi \gg 1$. When $\xi \ll 1$ and $\beta \ll 1$, P_{th} is equal to $\gamma_{\text{photon}}/2\beta$ and the threshold pump rate decreases linearly with $1/\beta$. Figure 6 shows the light output power as a function of pump rate for a set of β values in this regime. The lasing threshold, characterized by the jump in the efficiency of converting pump power to light output power, decreases with increasing β. When $\xi \gg 1$ and $\beta \ll 1$, P_{th} is independent of β and is fixed at the pump rate required to obtain transparency. One must keep the total number of atoms NV and the spontaneous decay rate Γ small to satisfy the condition $\xi \ll 1$ and obtain a decrease in the threshold pump rate with an increase in β. Of course, we are assuming that the active material can be pumped to emit a sufficient number of photons to achieve threshold at these low gain volumes. In practice, the linear gain dependence assumed in this analysis cannot be achieved if the gain volume is too small. An-

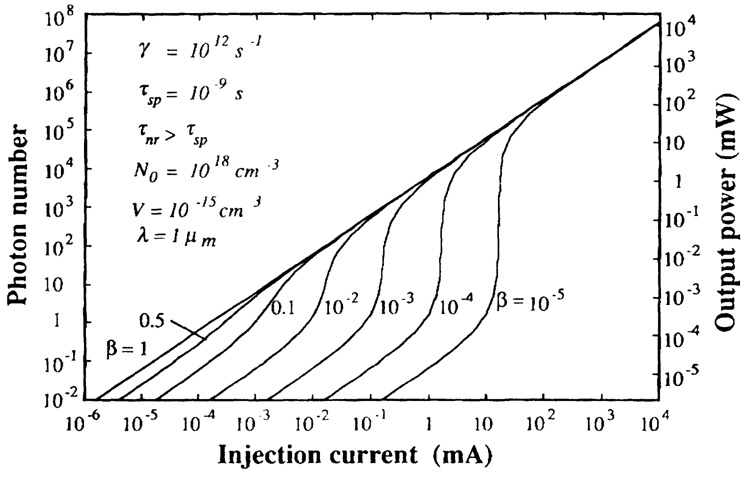

FIG. 6. Photon flux and power. The curves show the average flux in the mode (left axis) and the corresponding output power (right axis) for microcavity lasers with various spontaneous-emission coefficients β. Parameters used for the theoretical model that generated these curves are $\gamma_{\text{photon}} = 10^{12}$/s, $\Gamma = 10^{9}$/s, $N/V = 10^{18}$/cm^3, $V = 10^{-15}$ cm^3, and $\lambda = 1$ μm.

other consequence of experimental reality is that β can vary with excitation level if the spectral-emission width is not a constant.

In Fig. 6, we see that when the spontaneous-emission coefficient β is 1 there is no jump in the quantum efficiency of the conversion of pump photons to light output. This is the "thresholdless" laser, or "zero-threshold" laser. However, there is a well-defined threshold, $P_{th} = \gamma_{\text{photon}}$, even in this case, according to Eq. (15). At pump rates below threshold, the average photon number in the cavity is less than 1, so that the dominant emission process is incoherent spontaneous emission. At pump rates above threshold, the average photon number becomes greater than 1, resulting in a dominance of coherent stimulated emission. With the onset of stimulated emission, the population inversion is "pinned" at its threshold value, because each newly generated electron–hole pair very rapidly recombines in a stimulated process. The dominance of stimulated emission above threshold is also expected to result in an increased coherence of the light output and a narrowing of the output spectrum even in the $\beta = 1$ limit, as shown in Fig. 7(a).

Fluctuations of the light intensity emitted by a microlaser exhibit a resonance increase at threshold, as shown in Fig. 7(b). These light fluctuations, similar to the fluctuations found in a second-order phase transition, are a measure of threshold behavior that also persists in the limit of $\beta = 1$. Note that as β approaches 1, these fluctuations decrease and their relative width in pump power increases, thus providing another indicator of the unique characteristics in the high-β microlaser limit.

We conclude from an examination of carrier-density pinning, linewidth narrowing, and light-level fluctuations—present even when there is only approximately one photon in the cavity—that there is a well-defined threshold even for $\beta = 1$. This threshold discriminates between regions that can be described as a linear amplifier and a nonlinear laser oscillator.

2.4 Fabry–Pérot Microcavities

The simplest approach to fabricating an optical microcavity is to shrink the spacing between the mirrors of a Fabry–Pérot resonator to λ/n while reducing the lateral dimensions to the same range. This structure provides a single dominant longitudinal field mode that radiates into a narrow range of angles around the cavity axis. The first optical microcavity experiments used dye molecules between high-reflectivity dielectric mirrors in the Fabry–Pérot configuration.

Semiconductor microcavities provide high-Q Fabry–Pérot cavities for both basic studies and potential applications. The Q of a cavity is the number of radian cycles required for the optical-field energy in the cavity to decay by a factor of $1/e$. Molecular-beam epitaxy or organometallic chemical vapor deposition techniques are used to deposit high-reflectivity mirrors consisting of alternating quarter-wavelength layers of lattice-matched semiconductors. For example, 15 to 20 pairs of quarter-wave layers of $Al_{0.2}Ga_{0.8}As$ and AlAs result in a reflectivity greater than 99% and Q values greater than 500. The optically active layer in such a microcavity is typically a GaAs quantum well located at the midplane of the cavity, where the field strength is a maximum.

Planar Fabry–Pérot microcavities have been studied theoretically and experimentally in some detail. The two important parameters for laser performance, γ_{photon} and β, depend on the cavity geometry and the spectral-emission width of the quantum-well gain material. Even if the dimensions of the cavity in the plane between the mirrors are much larger than an optical wavelength, the planar cavity structure will give rise to discrete spatial modes with a finite mode radius r_m. The mirror spacing and reflectivities determine the radiating solid angle of the mode. This solid angle is inversely proportional to the coherence radius and directly proportional to the cold-cavity linewidth $\Delta\lambda_m$—the linewidth of the cavity resonance without the gain material. For example, a GaAs microcavity with a spacing of λ/n between the mirrors and with mirror reflectivities of 99% corresponds to a mode radius $r_m \simeq 7\ \mu\text{m}$, a cavity resonance linewidth $\Delta\lambda_m \simeq 0.43$ nm, and a far-field angular radiation angle of 5.7°.

The maximum β achievable with a large-area planar GaAs Fabry–Pérot microcavity is estimated to be 0.05, assuming dipole moments oriented in the plane of the cavity and negligible spectral broadening. However, the

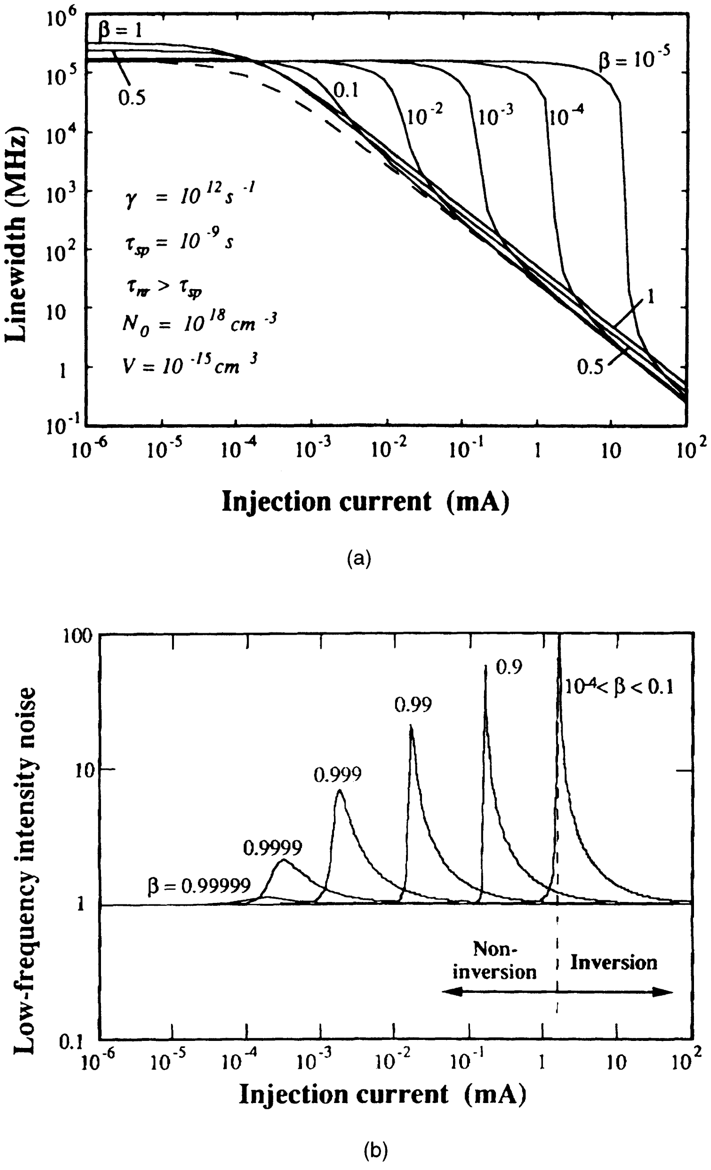

(a)

(b)

FIG. 7. Linewidth and noise. (a) The laser linewidth and (b) the intensity-noise spectral density in the light output from a microlaser as functions of the pump current are plotted here for a series of β values. The noise shown here is normalized with respect to the mean number of photons in the cavity. The parameters used in the theoretical model that generated these curves are the same as in Fig. 6.

GaAs quantum-well emission linewidth is 30 nm at room temperature and 2.5 nm at 4 K, much broader than the cold-cavity linewidth of 0.43 nm. As soon as the emission linewidth exceeds the cold-cavity bandwidth, β starts to decrease with the ratio of the two widths; β is thus estimated to be 0.05(0.43/2.5) $\simeq$ 0.01 at 4 K. The quantum efficiency for conversion of pump photons to output light is close to unity above the threshold, while it is equal to β below it, assuming negligible absorption in the cavity. Therefore β can be experimentally deduced from the step at threshold, yielding a β of 10^{-2}, as shown in Fig. 8.

The β values achievable with the large-area planar microcavities are limited by the leakage of spontaneous emission into directions centered around the plane of the cavity. A postlike structure, fabricated by wet or dry etching, can suppress this leakage of spontaneous emission and increase the β values up to 0.5 if the post diameter is on the order of the wavelength in the material and

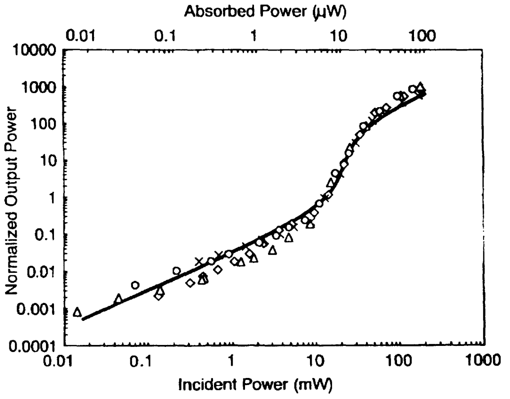

FIG. 8. Power transfer. The normalized output power from a GaAs Fabry–Pérot microcavity is plotted here as a function of incident pump power (bottom axis) and absorbed pump power (top axis). The solid line is the theoretical curve for $\beta = 0.01$.

if the quantum-well emission linewidth is much narrower than the cold-cavity linewidth. The factor-of-2 degradation in β stems from the fact that the two degenerate orthogonal polarization modes capture an equal amount of spontaneous emission below threshold, but only one of them oscillates above threshold. Such a high β has not yet been observed experimentally, in part because the emission linewidths are often broader than the cold-cavity linewidths.

An alternative way to increase β is to use a hemispherical structure with an integrated curved mirror (Matinaga *et al.,* 1993). To fabricate each of these curved mirrors, researchers formed a photoresist disk on a planar GaAs-AlGaAs optical cavity. They then melted this photoresist by heat treatment, forming it into a hemispherical disk. Finally, they transferred the curved geometry of the photoresist disk to the semiconductor by dry etching using an ion beam. They controlled the curvature and diameter of the structures by varying the diameter and the thickness of the resist disk; the resulting curvatures ranged from 5 to 20 μm. The hemispherical microcavity has a measured β value of 0.05, which compares with the theoretical prediction of 0.1.

A variety of condensed-matter Fabry–Pérot microcavities have exhibited modifications of the spontaneous-emission patterns and rates. Modified radiation patterns have been observed for the exciton emission from GaAs microcavities (Yamamoto *et al.,* 1989). Dye-molecule experiments have demonstrated both enhanced and inhibited spontaneous-emission rates (DeMartini *et al.,* 1991). One easily controls inhibition or enhancement in these experiments by adjusting the Fabry–Pérot cavity spacing to set the position of the field node in the active dye layer. Narrowing of the emission angle and enhancement of the intensity of the emitted radiation have been observed (Schubert *et al.,* 1992) from erbium-doped glass in a Fabry–Pérot cavity with mirrors formed from four alternate quarter-wave layers of silicon and SiO_2: The intensity of 1.54-μm radiation from the Er-doped glass along the cavity axis was enhanced by between one and two orders of magnitude, and the spectral width was narrowed by a factor of 10.

2.5 Microdisk Cavities

"Whispering-gallery" modes are a second successful route to achieving high-*Q* microcavities. Total internal reflection along a curved boundary between two materials with different optical indices results in high-*Q* modes propagating within a half-wavelength of the boundary. Lord Rayleigh conceived these modes over a century ago as an explanation of the efficient propagation of sound—even of whispers—along the walls of St. Paul's Cathedral in London. In fact, the ratio of the sound wavelength to the cathedral diameter is very similar to the ratio of the light wavelength to the diameter of optical microcavities being studied today.

Whispering-gallery modes in condensed-matter microstructures were pioneered using nearly spherical liquid droplets (Tzeng *et al.,* 1984). Cavities with *Q* values as high as 10^8 can be obtained because of the nearly perfect smoothness of the droplet boundaries, which in many cases is limited only by thermal fluctuations at the surface. Spectacular enhancements of the spontaneous decay rates of dye molecules—by a factor of 120—have been demonstrated in these microcavities (Campillo *et al.,* 1991). Polymer spheres have now been fabricated with the optical smoothness required for high-*Q* whispering-gallery–mode microcavities (Kuwata-Gonokami *et al.,* 1992): Optically pumping a dye-doped polystyrene sphere results in a red microlaser.

Researchers have demonstrated semiconductor whispering-gallery–mode microresonators as microlasers in the disk and cylindrical geometries (McCall *et al.*, 1992). The thickness of the disk is only 100 nm, less than a half-wavelength in the semiconductor, thus allowing only one transverse-optical mode to propagate in the plane of the disk. This microcavity geometry allows an excellent overlap of the optical gain region with the mode volume. Large overlap factors are important for achieving low microlaser thresholds. This geometry avoids the complexity of multilayer mirrors and allows the fabrication of high-Q microcavities in systems like InGaAs-InGaAsP, where the index contrast for layered mirrors is not high enough to permit fabrication of high-reflectivity distributed Bragg reflectors. For a 1-μm-radius disk and an optical wavelength of 1.5 μm, there are approximately eight optical nodes arranged around the disk edge in a whispering-gallery mode. The spectral spacing of modes with differing numbers of nodes is large enough that only one mode dominates the InGaAs quantum-well emission spectrum centered near 1.5-μm wavelength. The measured Q values for these 1-μm-radius microdisks range from 200 to 1000, and the β values are near 0.1, compared with the theoretically predicted value of 0.3.

2.6 Low-Threshold Microlasers

Optical microresonators promise ultralow-threshold laser behavior primarily because the threshold should scale down with the gain volume and with the increased coupling of spontaneous emission into the cavity mode—that is, with increased values of β. As the gain volume approaches a fraction of a cubic wavelength in the material and as β approaches 1, threshold powers below 1 μW are expected for sufficiently high Q values. This is a factor of nearly a thousand less than the thresholds near 1 mW that are typical for semiconductor lasers in common use today with volumes of several hundred cubic wavelengths and β values that range from 10^{-4} to 10^{-5}.

Estimated thresholds have been demonstrated in the 5–10-μW range for Fabry–Pérot semiconductor microlasers with one quantum well and in the 10–20-μW range for whispering-gallery–mode microlasers with six quantum wells (Yamamoto and Slusher, 1993). These experiments used optical pumping and temperatures below that of liquid nitrogen. The actual pump-light levels incident on the structures in these experiments range from 700 μW to 1 mW for the Fabry–Pérot microcavities and from 30 to 100 μW for the whispering-gallery–mode microcavities. Only the fraction of this power actually absorbed in the active region is used to obtain the estimated thresholds. Nonlasing modes reduce the β values in these microlaser experiments to the 0.01–0.1 range, because the total volumes of the cavities are still considerably larger than the ideal limit.

2.7 Future

The present trends toward smaller-scale nanostructures and the continuous development of new and improved materials should result in steady progress toward more ideal optical microcavities. Examples of new material include quantum wires and dots in semiconductor or organic systems. An interesting approach for achieving a truly isolated optical mode is the fabrication of three-dimensional photonic bandgap systems that exclude all optical modes over a limited optical spectrum (Everitt, 1992, and references therein). An "acceptor" or "donor" optical mode with a frequency in this isolated spectral region can then be formed as a dielectric defect. These defect modes can in principle form high-Q resonant modes with β values very close to 1. With advances in experimental systems, we may soon see a more substantial basis in experiment for the novel physics of QED, lasing without inversion, cavity-assisted excitonic superradiance, and squeezed-light generation (Yamamoto *et al.*, 1993, and references therein).

The limit of isolated field modes in microresonators may also have a significant impact in applied areas. For example, if visible microlasers can be developed, their many potential applications include fast holographic memories, laser projection displays, plastic fiber-optic communications, and applications where He-Ne lasers are used at present.

3. QUANTUM NONDEMOLITION MEASUREMENT—CONTROL OF MEASUREMENT BACK ACTION

3.1 Introduction

The quantum theory of measurement has been widely regarded as a problematic and dubious subject of little relevance to real physics. The reason for this irrelevance of the quantum theory of measurement is, however, only because experimentalists could not make repetitive measurements at a quantum limit on a single quantum object until very recently. In the 1980s, the experimentalists' interest has shifted from the single destructive measurement of an ensemble of microscopic systems to the repeated nondestructive measurements of a single macroscopic system. An experimentalist in quantum optics finally began to catch up with the measurement concepts such as "collapse of the wave function" of the creators of quantum mechanics.

The standard quantum theory based on unitary time evolution cannot describe the process of a measurement. A measurement is a nonunitary and indeterministic process, for which a so-called "projection postulate" is applied. This is a very reasonable postulate, but there has been no direct experimental evidence for the postulate so far. A quantum nondemolition measurement is the first experiment to confirm directly such a nonunitary collapse of the wave function by a measurement process.

3.2 General Quantum Measurement and QND

In a general quantum measurement, the observable of the signal system, $\hat{A}_s$, is measured by means of the change in the observable of the probe system, $\hat{A}_p$, after the proper interaction between the signal and probe systems, as shown in Fig. 9. The Heisenberg equations of motion for $\hat{A}_s$ and $\hat{A}_p$ are written as

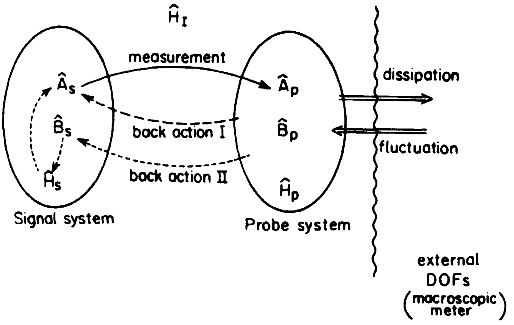

FIG. 9. QND measurement model with a probe system and a macroscopic meter. Signal observable $\hat{A}_s$ is nondemolitionally measured through demolitional measurement of probe observable $\hat{A}_p$, whose motion is affected by $\hat{A}_s$ by means of interaction Hamiltonian $\hat{H}_I$. If $\hat{H}_I$ does not commute with $\hat{A}_s$, the motion of $\hat{A}_s$ is kicked by $\hat{H}_I$ (back action I). If $\hat{H}_s$ contains $\hat{B}_s$, the motion of $\hat{A}_s$ is also kicked by the uncertainty principle between $\hat{A}_s$ and $\hat{B}_s$ (back action II).

$$-i\hbar \frac{d}{dt}\hat{A}_s = [\hat{H}_s,\hat{A}_s] + [\hat{H}_I,\hat{A}_s] \tag{16}$$

and

$$-i\hbar \frac{d}{dt}\hat{A}_p = [\hat{H}_p,\hat{A}_p] + [\hat{H}_I,\hat{A}_p], \tag{17}$$

where $\hat{H}_s$ is the unperturbed Hamiltonian of the signal system, $\hat{H}_p$ is that of the probe system, and $\hat{H}_I$ is the interaction Hamiltonian between the two. The first commutators in Eqs. (16) and (17) represent the free motion of $\hat{A}_s$ and $\hat{A}_p$. The second commutators represent the contribution of the interaction between the signal and probe systems.

In order to measure $\hat{A}_s$ using $\hat{A}_p$, $[\hat{H}_I,\hat{A}_p]$ in Eq. (17) should not be zero and H_I should be a function of $\hat{A}_s$. These are the requirements for a general quantum measurement:

1. $\hat{H}_I$ is a function of $\hat{A}_s$.
2. $[\hat{H}_I, \hat{A}_p] \neq 0$

In general, a measurement of $\hat{A}_s$ affects the motion of $\hat{A}_s$ itself in two ways (measurement back action). One way is the change of $\hat{A}_s$ due to $[\hat{H}_I,\hat{A}_s]$ in Eq. (16). Such an example is photoelectron emission in a photodetector. When a signal photon number is measured by a conventional photodetector, the photon absorption process naturally changes the signal photon number. Such a back action is shown by the dashed line in Fig. 9. If the interaction Hamiltonian H_I commutes with $\hat{A}_s$,

$$[H_I, \hat{A}_s] = 0, \tag{18}$$

H_I is called a "back-action evading" type, and this first source of back action can be eliminated.

The other way of changing $\hat{A}_s$ is through the uncertainty that is introduced on the observable conjugate to $\hat{A}_s$ by the measurement. The Heisenberg uncertainty principle dictates that when the observable $\hat{A}_s$ is measured with accuracy $\langle \Delta \hat{A}_s^2 \rangle$, the uncertainty imposed on the conjugate observable $\hat{B}_s$ should be larger than

$$\langle \Delta \hat{B}_s^2 \rangle \geq \frac{1}{4\langle \Delta \hat{A}_s^2 \rangle} |\langle [\hat{A}_s, \hat{B}_s] \rangle|^2. \tag{19}$$

If the unperturbed Hamiltonian H_s contains the conjugate observable $\hat{B}_s$, the motion of $\hat{A}_s$ is affected by this uncertainty of $\hat{B}_s$. An example is the position measurement of a free particle. When a free particle's position is measured by a γ-ray microscope, the measurement accuracy is proportional to the photon wavelength, $\Delta q \simeq \lambda$. A photon scattered by a free particle changes the particle's momentum. This back action imposed on the particle's momentum is of the order of the photon's momentum itself, $\Delta p \simeq h/\lambda$. The kicked momentum then perturbs the free motion of the particle's position according to the following relation:

$$H_s = \frac{\hat{p}^2}{2M} \Rightarrow \hat{q}(t) = \hat{q}(0) + \frac{\hat{p}(0)}{2M} t. \tag{20}$$

Here $\hat{q}(0)$ and $\hat{p}(0)$ are the position and momentum just after the first measurement. This back action is shown by the dotted line in Fig. 9.

If $\hat{A}_s(0)$ commutes with $\hat{A}_s(t)$,

$$[\hat{A}_s(0), \hat{A}_s(t)] = 0, \tag{21}$$

$\hat{A}_s$ is called a "QND observable" and the second source of back action can be eliminated. Equations (18) and (21) are the conditions for a QND measurement (Caves *et al.*, 1980, and references therein).

In a QND measurement, the observable $\hat{A}_s$ is not affected at all, even though the measurement accuracy of $\hat{A}_s$ is increased arbitrarily. In optical communication, this means that the information carried by a signal wave is fully extracted at one receiver but the signal wave still can transport exactly the same information to the other receivers.

3.3 QND Measurement for Photon Number

When the observable of the signal wave is a photon number $\hat{A}_s = \hat{n}_s$, condition (21) is satisfied, since the system free Hamiltonian is given by $H_s = \hbar\omega_s(\hat{n}_s + \frac{1}{2})$ and $[H_s, \hat{n}_s] = 0$; that is, a photon number is a "QND observable."

Let us consider the configuration shown in Fig. 10 (Imoto *et al.*, 1985). The signal wave at frequency ω_s propagates along an optical Kerr medium with a probe wave at frequency ω_p in which the phase of the probe wave is modulated by the photon number of the signal wave. The phase modulation of the probe wave is detected by a homodyne receiver, and thus the information about the signal photon number can be extracted. If the signal wave is not attenuated, the signal photon number is preserved.

The readout observable of the probe wave is a quadrature component $\hat{A}_p \equiv \hat{a}_{p2} = (\hat{a}_p - \hat{a}_p^\dagger)/2i$. The interaction Hamiltonian for the measurement is that for the interphase modulation between the signal and probe waves:

$$H_I = \hbar\chi\hat{n}_s\hat{n}_p. \tag{22}$$

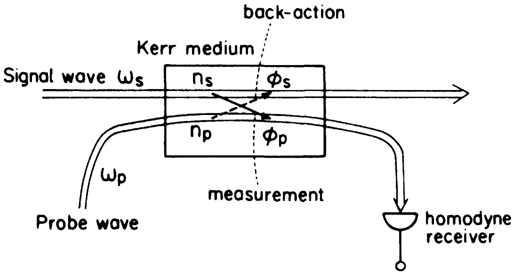

FIG. 10. Schematical view of a QND measurement scheme with an optical Kerr medium. The signal photon number is nondemolitionally measured through the homodyne detection of the probe phase, which carries the signal photon-number information by means of the optical Kerr effect.

The constant χ is equal to $(\hbar/2V\epsilon^2)\omega_p\omega_s\chi^{(3)}$, where V is the mode volume and $\chi^{(3)}$ is the third-order susceptibility; ϵ is the dielectric constant of the medium.

It is easily verified that the interaction Hamiltonian [Eq. (22)] satisfies the general quantum measurement conditions **1** and **2** in Sec. 3.2. The QND condition [Eq. (18)] is also satisfied, and thus the interaction Hamiltonian [Eq. (22)] is a "back-action evading" type.

3.3.1 Heisenberg Picture The unitary evolution for the signal- and probe-wave annihilation operators resulting from the onset of Eq. (22) is expressed by

$$\hat{a}_s(L) = \exp(i\sqrt{F}\hat{n}_p)\hat{a}_s(0), \tag{23}$$

and

$$\hat{a}_p(L) = \exp(i\sqrt{F}\hat{n}_s)\hat{a}_p(0). \tag{24}$$

Here $\sqrt{F} = \chi L/c$, L is the length of the Kerr medium, and c is the light velocity. The optical homodyne receiver with a proper local oscillator phase (90° out of phase for the coherent excitation of the probe wave) measures the quadrature phase component of $a_p(L)$,

$$\begin{aligned}\hat{a}_{p2}(L) &= (1/2i)[\hat{a}_p(L) - \hat{a}_p^\dagger(L)]\\ &= \hat{a}_{p1}(0)\sin(\sqrt{F}\hat{n}_s) + \hat{a}_{p2}(0)\cos(\sqrt{F}\hat{n}_s).\end{aligned} \tag{25}$$

Since the probe wave has the coherent excitation only along $\hat{a}_{p1}(0)$, that is, $\langle\hat{a}_{p1}(0)\rangle \neq 0$ and $\langle\hat{a}_{p2}(0)\rangle = 0$, and the phase modulation $\sqrt{F}\hat{n}_s$ is a small quantity, Eq. (25) is reduced to

$$\hat{a}_{p2}(L) \simeq \langle\hat{a}_{p1}(0)\rangle\sqrt{F}\hat{n}_s + \Delta\hat{a}_{p2}(0). \tag{26}$$

The signal photon number to be measured is expressed by

$$\hat{n}_s^{(\text{obs})} \equiv \frac{\hat{a}_{p2}(L)}{\sqrt{F}\langle\hat{a}_{p1}(0)\rangle} = \hat{n}_s + \frac{\Delta\hat{a}_{p2}(0)}{\sqrt{F}\langle\hat{a}_{p1}(0)\rangle}. \tag{27}$$

Taking the expectation value and variance of Eq. (27), we obtain

$$\langle\hat{n}_s^{(\text{obs})}\rangle = \langle\hat{n}_s\rangle, \tag{28}$$

$$\langle(\Delta\hat{n}_s^{(\text{obs})})^2\rangle = \langle\Delta\hat{n}_s^2\rangle + \frac{\langle\Delta\hat{a}_{p2}(0)^2\rangle}{F\langle\hat{n}_p\rangle}. \tag{29}$$

The second term of Eq. (29), $\langle\Delta\hat{n}_s'^2\rangle = \langle\Delta\hat{a}_{p2}(0)^2\rangle/F\langle\hat{n}_p\rangle$, represents the measurement error, which is determined by the probe-wave phase noise $\langle\Delta\hat{\psi}_p^2\rangle \simeq \langle\Delta\hat{a}_{p2}(0)^2\rangle/\langle\hat{n}_p\rangle$ and the nonlinear interaction strength $\sqrt{F}$. It can be decreased arbitrarily by increasing $\sqrt{F}$ and/or decreasing $\langle\Delta\hat{\psi}_p^2\rangle$. This QND measurement is ideal, since the expectation value and variance, that is, the statistics of measurement results, are equal to those of the signal wave itself. When the probe wave is in a coherent state, $\langle\Delta\hat{\psi}_p^2\rangle$ is given by $1/4\langle\hat{n}_p\rangle$. When the probe wave is in an optimum squeezed state, $\langle\Delta\hat{\psi}_p^2\rangle$ can be reduced to $1/4\langle\hat{n}_p\rangle(\langle\hat{n}_p\rangle + 1)$.

The phase noise of the signal wave is increased by the other interaction represented by Eq. (23). The quadrature phase component of the signal wave is similarly calculated as

$$\hat{a}_{s2}(L) \simeq \langle\hat{a}_{s1}(0)\rangle\sqrt{F}\hat{n}_p + \Delta\hat{a}_{s2}(0). \tag{30}$$

The phase noise is thus given by

$$\langle\hat{\Delta}\psi_s^2\rangle \equiv \frac{\langle\Delta\hat{a}_{s2}(L)^2\rangle}{\langle\hat{a}_{s1}(0)\rangle^2} = \langle\Delta\hat{\psi}_s^2(0)\rangle + F\langle\Delta\hat{n}_p^2\rangle. \tag{31}$$

Here $\langle\Delta\hat{\psi}_s(0)^2\rangle = \langle\Delta\hat{a}_{p2}(0)^2\rangle/\langle\hat{a}_{p1}(0)\rangle^2$ is the initial phase noise of the signal wave. The second term, $\langle\Delta\hat{\psi}_s'^2\rangle = F\langle\Delta\hat{n}_p^2\rangle$, is the back-action noise added by the measurement. With decrease of the measurement error, the back-action noise increases. When the probe wave satisfies the minimum uncertainty product, $\langle\Delta\hat{a}_{p1}(0)^2\rangle\langle\Delta\hat{a}_{p2}(0)^2\rangle = \frac{1}{16}$, the measurement error and the back-action noise satisfy the Heisenberg minimum uncertainty product:

$$\langle\Delta\hat{n}_s'^2\rangle\langle\Delta\hat{\psi}_s'^2\rangle = \tfrac{1}{4}. \tag{32}$$

The preceding argument concerning the measurement error and back-action noise suggests that the signal wave after QND measurement is "reduced" to a number-phase squeezed state. This can be easily shown by using the projection postulate.

3.3.2 Schrödinger Picture Suppose the signal and probe waves are initially in coherent states $|\alpha\rangle_s$ and $|\beta\rangle_p$. The density operator after the interphase modulation is expressed by the unitary operator $\hat U = \exp(i\sqrt{F}\hat n_s \hat n_p)$,

$$\hat\rho_{sp} = \hat U|\alpha_0\rangle_{ss}\langle\alpha_0| \otimes |\beta_0\rangle_{pp}\langle\beta_0|\hat U^\dagger. \qquad (33)$$

The density operator after the measurement of the quadrature amplitude of the probe wave is given by (for the specific readout β_2)

$$\hat\rho_s^{(\mathrm{meas},\beta_2)} = \mathrm{Tr}_p(|\beta_2\rangle_{pp}\langle\beta_2| \otimes \hat\rho_{sp}). \qquad (34)$$

The quasi-probability density $\langle\alpha|\hat\rho_s^{(\mathrm{meas},\beta_2)}|\alpha\rangle$ is shown in Fig. 11. The variances for photon-number and sine operators,

$$\begin{aligned}\langle\Delta\hat n^2\rangle &= \mathrm{Tr}_s[\hat\rho_s^{(\mathrm{meas},\beta_2)}(\hat n - \langle\hat n\rangle)^2] \\ &\simeq (1/|\alpha|^2 + 4F|\beta|^2)^{-1} \qquad (35)\end{aligned}$$

and

$$\begin{aligned}\langle\Delta\hat S^2\rangle &= \mathrm{Tr}_s[\hat\rho_s^{(\mathrm{meas},\beta_2)}(\hat S - \langle\hat S\rangle)^2] \\ &\simeq 1/4|\alpha|^2 + \tfrac{1}{2}[1 - \exp(-2F|\beta|^2)], \qquad (36)\end{aligned}$$

satisfy the number-phase minimum uncertainty product, $\langle\Delta\hat n^2\rangle\langle\Delta\hat S^2\rangle = \frac{1}{4}\langle\hat C\rangle^2$, as shown in Fig. 12.

The initial coherent state is transformed to a number-phase squeezed state by the nonunitary state reduction. QND measurement is considered as the generation process of nonclassical light. This QND measurement was experimentally demonstrated with a Kerr nonlinear medium (Friberg *et al.*, 1992, and references therein).

Two similar but slightly different configurations for QND measurement of photon number have been proposed. One utilizes a microwave photon in a high-Q superconductor cavity as a probe system (Braginsky and Vyatchanin, 1981). The other employs a nonlinear coupler based on four-wave mixing (Milburn and Walls, 1983).

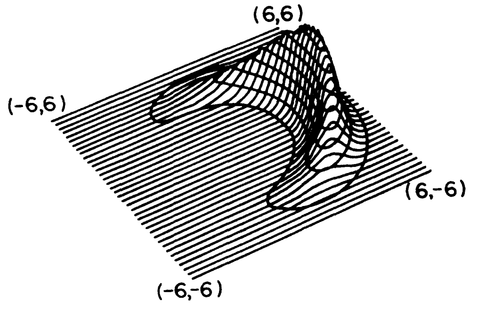

FIG. 11. Quasi-probability density $Q(\alpha)$ of the state after the measurement in the α_1-α_2 phase space, where α_1 and α_2 are the two quadrature amplitudes.

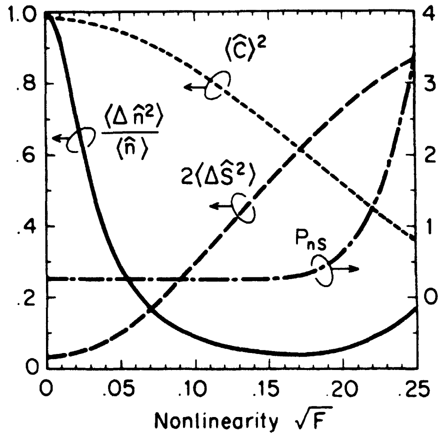

FIG. 12. Uncertainties $\langle\Delta\hat n^2\rangle$, $\langle\Delta\hat S^2\rangle$ and uncertainty product $P_{ns} = \langle\Delta\hat n^2\rangle\langle\Delta\hat S^2\rangle/\langle\hat C\rangle^2$ of the state after the measurement ($|\alpha_0| = |\beta_0| = 4$, $\beta_2 = 0$).

3.4 Other QND Measurements

The quadrature amplitude $\hat a_1$ or $\hat a_2$ is also a QND observable. A degenerate parametric amplifier with two orthogonal polarization channels can perform such a QND measurement of the quadrature amplitude (Shelby and Levenson, 1987). This QND measurement was experimentally demonstrated with a type II nonlinear crystal (Pereira *et al.*, 1994, and references therein). In this QND measurement, the back-action noise is imposed on the conjugate observable ($\hat a_2$ for the measurement of $\hat a_1$) and requires an increased energy. The repeated QND measurements leave the measured quadrature amplitude unchanged but amplify the system energy.

4. OPTICAL-SOLITON QUANTUM EFFECT—MACROSCOPIC QUANTUM EFFECT

4.1 Introduction

The first reported soliton was a water wave in a canal, observed by John Scott-Russell (Scott-Russell, 1844). A wave or "lump" of water was seen to travel over long distances without spreading out or breaking up. Since then, numerous examples of solitons have been found in various physical, chemical, and biological environments, all propagating without spreading out or breaking up. Recently, there has been great interest in solitons in optical fibers. Such solitons, predicted in 1973 (Hasegawa and Tappert, 1973) and first demonstrated in 1980 (Mollenauer *et al.*, 1980), offer advantages for high-speed, long-distance communication: propagation distances of more than 10^6 km have been demonstrated (Nakazawa *et al.*, 1991). The stability of optical solitons in fibers makes them interesting not just for optical communication systems, but also for fundamental tests of quantum mechanics.

Solitons are persistent, "solitary waves" or pulses that are particularly resistant to the disruptions that ordinarily distort a pulse as it propagates. A short pulse is made up of a range of oscillation frequencies: The shorter the pulse, the broader its spectrum. In any physical medium, the propagation velocity depends on frequency, an effect known as group-velocity dispersion. By itself, this causes different portions of the pulse spectrum to travel with different velocities, leading to pulse broadening. Temporal solitons exist in nonlinear media in which the velocity also depends on the amplitude of the pulse. If the signs of these two effects are opposite, a soliton can form that (in classical physics) would propagate without distortion. Spatial optical solitons form in a similar manner, opposing diffraction against self-focusing in a nonlinear medium (Zakharov and Shabat, 1971).

In the case of solitons in optical fibers, the balance is between the wavelength dependence and the intensity dependence of the refractive index of the fiber waveguide. Group-velocity dispersion causes shorter wavelengths to travel faster (for the case of interest here) than the longer wavelengths. Normally, the leading edge of the pulse becomes blueshifted and the trailing edge redshifted. However, the intensity dependence of the refractive index causes the phase of the pulse to be time dependent. This results in a frequency shift toward the red on the leading edge and toward the blue on the trailing edge, tending to cancel the effect of group-velocity dispersion. When these opposing effects balance, an optical-fiber soliton is formed.

Classical arguments predict that a soliton is invariant as it propagates. However, the classical picture of solitons is not correct quantum mechanically: A true quantum soliton undergoes phase diffusion and wave-packet spreading. A quantum soliton consists of a linear superposition of solitons with different photon-number and momentum eigenstates. Each component soliton has a different phase velocity and group velocity, causing the phase and the wave packet of quantum solitons to spread out as they propagate. This fact, implicit in an early work by Bethe (1931), has been explored recently in much more detail in the quantum theory of optical solitons (Haus and Lai, 1990, and references therein). We now know that solitons in optical fibers are superpositions of macroscopic quantum states that correspond to clusters of 10^8 or more photons, bound together. Nevertheless, intrinsically quantum-mechanical effects emerge and can be detected even for such macroscopically large objects. Recently, theoretical predictions of quantum phase diffusion have been confirmed experimentally in one of the few experimental tests of a simple, exactly soluble quantum-field theory. These experiments give direct evidence of the quantum nature of soliton propagation.

When two optical solitons collide, the phase and position of each soliton are correlated with the photon number and momentum, respectively, of the other soliton. This has been shown theoretically in the quantum theory of soliton collisions and has now been confirmed experimentally (Drummond *et al.*, 1993, and references therein). These experimental results not only extend our understanding of single quantum solitons to the collision of two solitons but also provide an experimental realization of a quantum nondemolition (QND) measurement of the photon number of an optical soliton.

QND measurements were first discussed by Landau and Peierls (1931). Lately, there has been renewed interest in developing various techniques for the realization of QND measurements because of their unique importance to quantum theories of measurement, and practical implications for precision measurements. Optical solitons propagate and collide with unchanging pulse shape, momentum, intensity, and energy. Consequently, the photon number of a soliton is not altered when its energy is measured by means of the phase shift of a soliton with which it collides. Accordingly, the conditions necessary for a QND measurement are met, and a QND measurement can be used to determine the number of photons in a soliton pulse with an accuracy better than the shot-noise limit.

4.2 Quantization of Optical Solitons

An optical soliton in a fiber can be described classically using the nonlinear Schrödinger equation, which is applicable to single-mode propagation in nonlinear dispersive fibers or waveguides. This is an approximate wave equation, which ignores quantum effects. In this section, we review the predictions of a quantum version of this equation, in which operators replace the classical field variables. Classically the most unexpected behavior is the existence of quantum solitons of discrete particle number and undetermined phase. More readily observable quantum effects include phase and position diffusion, in which classically stable observables have increasing quantum uncertainties with propagation. For soliton collisions, quantum theory predicts that information about phase and position is transferable from one soliton to another, allowing QND measurements.

In a reference frame moving at the group velocity v, the corresponding quantum operator equation for the photon density amplitude $\hat{\psi}$ in dimensionless form is

$$\frac{\partial}{\partial T}\,\hat{\psi}(T,X) = i\left\{\pm\tfrac{1}{2}\frac{\partial^2}{\partial X^2} + \frac{2\hat{\psi}^\dagger\hat{\psi}}{\bar{N}}\right\}\hat{\psi}(T,X), \qquad (37)$$

where

$$[\hat{\psi}(T,X),\hat{\psi}^\dagger(T,X')] = \delta(X - X'). \qquad (38)$$

The variable T represents elapsed time and is therefore proportional to the distance propagated. It is defined relative to a characteristic pulse duration t_0 and group-velocity dispersion $k_0'' = \partial^2 k/\partial\omega^2|_{\omega=\omega_0}$, so that

$$T = vt|k_0''|/t_0^2. \qquad (39)$$

The variable X represents spatial location relative to the moving reference frame:

$$X = (x/v - t)/t_0. \qquad (40)$$

The ± sign in Eq. (37) is positive for anomalous-dispersion ($k_0'' < 0$) optical fibers that support solitons, typically at wavelengths greater than ~1.4 μm. The quantity $\bar{N}$ is a scaling constant inversely proportional to n_2, the nonlinear refractive index per unit intensity:

$$\bar{N} = 2|k_0''|\mathcal{A}_c/n_2\hbar\omega_0^2 t_0. \qquad (41)$$

Here $\mathcal{A}_c$ is the effective modal cross section of the fiber. The operator $\hat{\psi}^\dagger(T,X)\hat{\psi}(T,X)$ directly represents the photon density in dimensionless units.

The operator equations show that photons propagate with a nonzero effective mass $m = \hbar/(|k''|v^3)$ in the moving reference frame. Similar changes in effective mass are observed in electrons traveling in solids. The nonlinear term in the equation is equivalent to a localized delta-function attractive force acting between the particles.

The energy eigenstates of this quantum nonlinear Schrödinger equation are known from early theoretical work on interacting Bose gases. The simplest quantum soliton is an eigenstate consisting of a cluster of photons with a well-defined energy and photon number—a photonic atom. Such a state is predicted to propagate stably, without distortion or breaking up into its component photons, just as an atom does in free space. There are forces predicted to act between two such quantum solitons, as there are between two atoms. Quantum states of this type can be pulses of macroscopic size, with 10^9 or more photons. Despite this, an isolated quantum soliton-number eigenstate has yet to be experimentally observed.

Fiber soliton experiments start with the input field in a coherent state, rather than in a state of well-defined particle number. In a coherent state, the quantum fluctuations of the electric field are identical to those of a vacuum state. As shown in Fig. 13, the variance is the same for either amplitude or phase fluctuations. Thus, the photon number N must have fluctuations such that $\langle(\Delta\hat{N})^2\rangle = \langle\hat{N}\rangle$. This means that an input field will be a superposition of quantum solitons of differ-

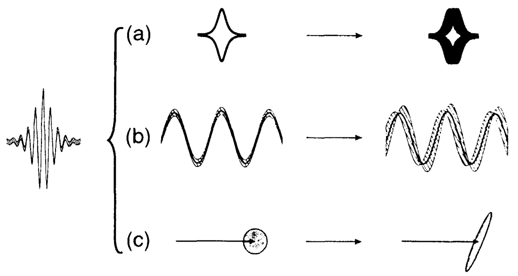

FIG. 13. Evolution of a quantum soliton. A coherent soliton, shown schematically at the left, is a superposition of a classical hyperbolic secant light pulse and quantum noise due to the zero-point fluctuations of the field. The presence of the zero-point fields imposes uncertainties in the soliton's position in time and space, its central frequency, its photon number or electric-field amplitude, and the phase of its electric field. The frequency uncertainty, coupled with the frequency dependence of the group velocity of the fiber, causes quantum spreading of the envelope of the pulse as shown in (a). The uncertainty in photon number obeys a Poisson distribution and is responsible for the "shot noise" observed when coherent light is detected. The photon-number uncertainty, along with the intensity dependence of the phase velocity, causes a spreading of the phase as shown in (b). The electric field and noise are shown for an optical period. The phase and position spreading are not due simply to an increase in uncertainty, but also to correlations with the amplitude and frequency of the pulse, respectively. This results in interference effects. The destructive interference of the noise occurs at certain phases of the optical cycle, as shown in (b). This is known as quantum-noise squeezing and can also be represented as a narrowing of the phase-space diagram of the field as shown in (c). There, the vector represents the average electric-field amplitude and the shaded area the quantum noise. The left side is a coherent state—the superposition of the field and zero-point fluctuations with equal and uncorrelated amplitude and phase uncertainties yields a circular distribution. At the right, the soliton is shown after evolving into a squeezed state. The noise is reduced in certain directions because of the correlations, yielding an elliptical distribution with a minor axis smaller than the width of the original coherent noise distribution.

ent photon number, along with some unbound "continuum" photons.

Even for superpositions of this type, there are large differences between the classical and quantum theories. We first note that the classical equations predict that coherent pulse will propagate essentially unchanged if injected with an initial amplitude of $\sqrt{\bar{N}/2} \times \operatorname{sech}(X)$. The time-dependent classical solution is given by

$$\psi_0(T,X) = \sqrt{\bar{N}/2}\,\operatorname{sech}(X)\exp[i(\theta_0 + T/2)], \tag{42}$$

where θ_0 is the initial phase. The parameter $\bar{N}$ corresponds to the mean photon number $\langle N\rangle$. Thus, the only change predicted classically is that the total phase $\theta(T) = \theta_0 + T/2$ increases linearly with distance down the fiber.

The classical result of a soliton as a phase-shifted, undistorted pulse never occurs in quantum theory. Even for a quantum soliton with a well-defined photon number, the center-of-mass position obeys the Heisenberg uncertainty principle. Position and momentum are conjugate observables and thus not simultaneously measurable, and so a soliton with some given velocity cannot have a well-defined position at all times. This holds as well for the conjugate observables of phase and photon number. Momentum and photon number are invariant during propagation. Their conjugates (position and phase) are predicted to show a quantum-mechanical time dependence—leading to increasing fluctuations as time progresses. This is the quantum diffusion of position and phase.

In order to understand quantum phase diffusion better, consider a simplified model of the input coherent pulse as a superposition of quantum solitons of different photon numbers. The classical phase shift of a soliton with N photons at elapsed time T is

$$\theta_N(T) = \frac{TN^2}{2\bar{N}^2} = \frac{x|k_0''|N^2}{2t_0^2\bar{N}^2}\,. \tag{43}$$

This phase shift can also be calculated directly from quantum theory, using the fact that the optically measured phase is the difference between the quantum phases of the N-photon and the $(N + 1)$-photon states. As many photon-number states are superimposed to give an initially coherent soliton, this leads to a corresponding diffusion of phase shifts after an elapsed time T. The enhanced phase fluctuations can be roughly estimated from the known fluctuations in photon number N, giving $\langle\Delta\theta(T)^2\rangle \approx T^2/\bar{N}$. More rigorously, the initial coherent-state phase fluctuations should also be included. This gives the final result that

$$\langle\Delta\theta(T)^2\rangle = (0.607 + T^2)/\bar{N}. \tag{44}$$

A similar result holds for the quantum diffusion of the center-of-mass position.

As a result of phase diffusion, the distribution of the amplitude fluctuations narrows, or is squeezed, for certain phases. This leads to quantum-noise reduction, as illustrated in Fig. 13. Its physical origin is the correlation of photon number and phase for each component of the superposition. Quantum noise reduction due to soliton squeezing has possible applications in low-noise measurements.

We now turn to soliton-collision experiments, which have been analyzed using quantum inverse-scattering methods. In fact, it is possible to understand soliton collisions using classical inverse-scattering results together with stochastic noise sources representing vacuum fluctuations. After two solitons collide, the position and phase of one soliton are modified by an amount that depends on the velocity and photon number of the other soliton. Quantitatively, we find

$$\Delta\theta_P^{\mathrm{COL}} = 2\left(\tan^{-1}\left[\frac{N_P + N_s}{\bar{N}|V_P - V_s|}\right] - \tan^{-1}\left[\frac{N_P - N_s}{\bar{N}|V_P - V_s|}\right]\right). \tag{45}$$

For large relative velocity differences, the result for the phase reduces to

$$\Delta\theta_P^{\mathrm{COL}} = 4N_s/\bar{N}|V_P - V_s|, \tag{46}$$

where $\Delta\theta_P^{\mathrm{COL}}$ is the phase shift experienced by a probe soliton, V_P and V_s are the velocities of the probe and signal solitons in dimensionless units, and N_P and N_s are the corresponding numbers of photons. The physical explanation of this effect is that as the solitons pass through each other, they experience a more intense field. Because of the nonlinearity of the refractive index, this increases the phase shift experienced by each soliton. The phase shift is proportional to the photon number of the other soliton—but is smaller at large relative velocities, as this reduces the interaction time.

A probe-soliton phase measurement after a collision of probe and signal solitons allows the photon number of the signal soliton to be measured, as $\Delta\theta_P^{\mathrm{COL}}$ is proportional to N_s. The significant feature of this interaction is that neither the photon number nor the velocity of either soliton is changed during the collision. In contrast, all other photon-number measurement techniques used in practice modify the photon number. Thus, soliton collisions allow a back-action–evading or quantum nondemolition measurement of the soliton photon number.

4.3 Quantum Nondemolition Measurements

For both the photon-number and the momentum eigenstates of an optical soliton, it is possible to perform two or more back-action–evading measurements that yield identical outcomes and do not change the quantum-mechanical state of the soliton. These are known as quantum nondemolition (QND) measurements. Several optical QND measurements have been proposed theoretically, and partial QND measurements have been demonstrated experimentally by several groups. However, two successive measurements on the same quantum state have yet to be achieved.

The unique characteristics of quantum solitons promise to make successive QND measurements possible. Except for solitons, all optical pulses experience broadening due to group-velocity dispersion in a dielectric medium. Thus, the photon number in any given time interval (the intensity of the pulse) is not a well-defined quantity. Much the same happens for continuous-wave optical signals, unless they are strictly monochromatic. However, quantum solitons can propagate with a well-defined photon-number state, so that two sequential measurements on the same quantum state can easily be carried out. As subsequent measurements can take place in a single optical fiber, so losses can nearly be eliminated.

A partial soliton QND measurement of photon number was recently demonstrated experimentally. Three jitter-free optical pulses—a signal, a probe, and an identical reference—were extracted from single color-center-laser pulses spectrally broadened to a 10-nm bandwidth by self-phase modulation in an optical fiber. These were launched into a 400-m, negative group-velocity dispersion, single-mode, polarization-preserving fiber, yielding a signal soliton at 1460.7 nm and identical probe and reference solitons at 1456.0 nm, all with ~1.0-nm bandwidths and 2.2-ps pulse widths. Timing was adjusted so that the signal and probe collided halfway along the fiber. The reference soliton, offset 30 ps from the probe, propagated without colliding. The three closely spaced solitons moved more quickly through the fiber than it could change as a result of thermal noise processes. This eliminated most of the detrimental effects of guided acoustic-wave Brillouin-scattering noise, which usually plagues quantum optical measurements in optical fibers.

The solitons, after colliding and leaving the fiber, were separated with a diffraction grating. The signal soliton, directed into a photodetector, gave a current proportional to its photon number. The probe and reference solitons, interfering in a soliton-collision interferometer as shown in Fig. 14, gave a current corresponding to the phase of the probe relative to that of the reference.

The results were as predicted by Eqs. (45) and (46): The phase shift of the probe was proportional to the photon number of the signal soliton. For example, when the signal soliton was switched on and off, the probe soliton's phase was modulated by 1.2 radians (70°). In the quantum regime, the fluctuations of the photon number of the signal were observed to be correlated with the phase fluctuations of the probe. Furthermore, the signal soliton's photon-number fluctuations were unchanged by the collision and shot-noise limited, showing that no excess noise was introduced back into the photon number of the signal.

The results of the experiment are shown in Fig. 15. Noise correlations were found by adding and then subtracting the two noise signals—differing results indicated a correlation. In the experiment, a delay line caused the signal-soliton shot-noise fluctuations to alternate in sign depending on the frequency. When the two noise signals were summed, correlations caused a modulation of the noise spectrum. Such a modulation, observed with a radio-frequency spectrum analyzer, is seen in the two sets of experimental data shown in Fig. 15(a). Figure 15(b) shows the modulation more clearly, after the data sets are subtracted to eliminate background variations. The results of the check for excess noise in the signal soliton, carried out immediately after the correlation measurements, are shown in Fig. 15(c). Above 12 MHz, the average difference between the noise of the signal solitons and a quantum-limited calibration signal was less than 0.1 dB, indicating no excess noise due to either the collision or external factors.

The experiment showed that soliton-collision phase measurements satisfy the criterion for a QND measurement. The QND readout (the phase shift of the probe soliton) and the QND observable (the photon number

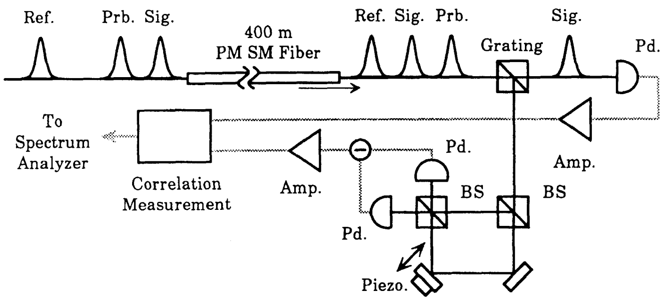

FIG. 14. A schematic diagram of the quantum nondemolition (QND) measurement. A signal soliton (Sig) at 1460.7 nm and identical probe (Prb) and reference (Ref) solitons at 1455 nm propagate in a 400-m polarization-maintaining (PM), single-mode (SM) fiber having anomalous group-velocity dispersion. The signal and probe solitons collide in the fiber, causing the phase of the probe soliton to shift in response to the photon number of the signal soliton. After leaving the fiber, the solitons are separated by a dispersing grating. A photodiode (PD) records the photon-number fluctuations of the signal soliton and an interferometer (with one arm slightly longer than the other so that the probe and reference pulses overlap) measures the phase-shift fluctuations of the probe soliton. Correlation measurements with a radio-frequency spectrum analyzer show the fluctuations to be correlated, indicating a partial QND measurement.

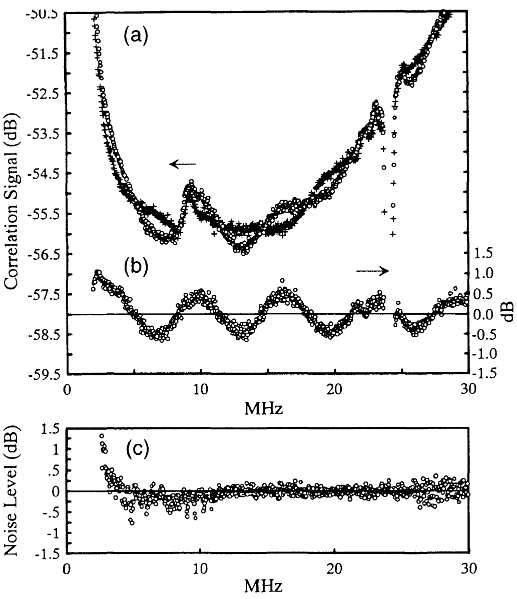

FIG. 15. Experimental results verifying the QND nature of the soliton-collision measurement. The correlation between the shot noise of the signal soliton and the phase noise of the probe soliton is indicated by the modulation of the two radio-frequency spectrum-analyzer traces in (a). Subtracting the two for clarity gives (b). The lack of excess noise due to the measurement is shown in (c), where the data points are the difference of the photon-number fluctuations of the signal soliton and a quantum-limited calibration signal. Above 12 MHz, they cluster about the standard quantum limit (solid line). Taken together, these results show the photon number of the signal soliton and the phase of the probe soliton to be quantum correlated.

of the signal soliton) were correlated; the measurement was back-action evading (no noise was introduced into the signal soliton's photon number), and the QND observable was not changed either by measurement, by propagation, or by external noise. Thus, for a soliton initially in a coherent state, a measurement as described here puts the soliton in a near photon-number state—a partial QND measurement causes photon-number squeezing. An identical second measurement on the same signal soliton, if correlated with the first measurement, could verify that both measurements gave the same result. In principle, this would be a full QND measurement that would leave the quantum-mechanical state of the soliton unaltered. It would be a classical-type measurement in that it would determine the soliton's photon number without changing it.

4.4 Future

Because solitons interact in a particlelike manner, preserving their identity when perturbed, they are also likely to be important for viable optical switching devices or logic gates. Nonsoliton implementations often suffer from pulse distortions that cause incomplete switching. From this perspective, a soliton-collision QND experiment is similar to an optical switching device—information coded on the signal soliton by means of its presence or by its photon number is transferred to the probe soliton. To a good approximation, this realizes a reversible quantum logic gate (Bennett, 1987, and references therein), a device of considerable theoretical interest. Quantum effects in the propagation and collisions of solitons may ultimately determine the limits to performance of optical switching devices. Such limits are only now beginning to be defined.

It is interesting to consider other physical systems that have optical solitons with stronger nonlinearities. Self-induced transparency in resonant two-level systems produces a slightly different type of optical soliton, mathematically described by a Maxwell–Bloch equation (McCall and Hahn, 1969). The leading edge of the pulse excites ground-state atoms to excited states, and then the trailing edge of the pulse stimulates transitions back to the ground state. In this way, the optical pulses propagate without being attenuated or distorted in the medium. As self-induced transparency solitons exist in resonant two-level systems, large nonlinearities are expected, even for solitons with low intensities. Resonant two-level systems also are expected to support spatial solitons with quantum properties similar to those described here, but with larger nonlinearities. Quantum solitons of a different variety—topological solitons—are predicted in a number of model quantum field theories used in solid-state and particle physics (Coleman, 1985). It is possible that our methods could be used to investigate directly these solitons, whose quantum properties are as yet untested by experiment.

5. OUTLOOK

Squeezing, cavity QED, and QND measurement are a nice demonstration of fundamental concepts of quantum mechanics such as Heisenberg's uncertainty relation, vacuum fluctuation, and von Neumann's projection postulate. There are also emerging subjects such as quantum cryptography and quantum computing (parallelism). Theoretical and experimental studies on various quantum optics have contributed greatly to the foundation of quantum mechanics. The field has now started to link with mesoscopic (nanostructure) physics and communication/computer technologies.

However, as Rolf Landauer has criticized repeatedly (Landauer, 1995), any information-processing scheme based on quantum coherence or superposition state is not realistic as a practical technology because it requires a perfect preparation of the system and zero dissipation. There is not such a perfect system in a real world. The computer and communication systems of today are classical ones, even though the principle of constituent devices such as a laser or a transistor has a quantum-mechanical origin. They do not utilize the quantum-mechanical nature of a linear superposition state.

The future evolution of quantum optics with nanostructure devices and atomic-scale devices is expected to give us an answer to the question "Is a real quantum effect useful for practical applications?"

GLOSSARY

Cavity Quantum Electrodynamics: Control of spontaneous emission by modifying a vacuum-field fluctuation by a cavity wall.

Heisenberg Uncertainty Principle: The statement that the quantum-mechanical uncertainties of a pair of conjugate observables such as the position and momentum of a free particle must satisfy a minimum product.

Optical Soliton: A solitary optical pulse that propagates without pulse spreading or breakup. An optical soliton in a fiber is produced by the balance between Kerr nonlinearity and negative group-velocity dispersion.

Quantum Nondemolition Measurement: A quantum measurement without the back-action noise on a measured observable. The measurement error of a measured observable and the back-action noise on the conjugate observable satisfy the Heisenberg uncertainty product.

Squeezed State: A minimum-uncertainty state of light with reduced quantum noise of one observable and enhanced quantum noise of the conjugate observable.

Vacuum Fluctuation: The ground state of an electromagnetic wave has a quantum-mechanical zero-point fluctuation that originates from the Heisenberg uncertainty principle.

Works Cited

Bennet, C. H. (1987), *Sci. Am.* **257** (5), 88–96.

Bergou, J., Davidovich, L., Orszag, M., Benkert, C., Hillery, M., Scully, M. O. (1989), *Phys. Rev. A* **40,** 5073–5085.

Berman, P. R. (1994), *Quantum Electrodynamics,* Boston: Academic Press.

Bethe, H. A. (1931), *Z. Phys.* **71,** 205–226.

Björk, G., Yamamoto, Y. (1991), *IEEE J. Quantum Electron.* **27,** 2386–2396.

Braginsky, V. B., Vyatchanin, S. P. (1981), *Dokl, Akad, Nauk SSSR* **259,** 570–572 [*Sov. Phys.—Dokl.* **26,** 686–690].

Campillo, A. J., Eversole, J. D., Lin, H.-B. (1991), *Phys. Rev. Lett.* **67,** 437–440.

Casimir, H., Polder, D. (1948), *Phys. Rev.* **73,** 360–374.

Caves, C. M., Thorne, K. S., Drever, R. W. P., Sandberg, V. D., Zimmermann, M. (1980), *Rev. Mod. Phys.* **52,** 341–392.

Coleman, S. (1985), *Aspects of Symmetry,* New York: Cambridge Univ. Press.

DeMartini, F., Jacobovitz, G. R. (1988), *Phys. Rev. Lett.* **60,** 1711–1714.

DeMartini, F., Marrocco, M., Mataloni, P., Crescentini, L., Loudon, R. (1991), *Phys. Rev. A* **43,** 2480–2483.

Dicke, R. H. (1954), *Phys. Rev.* **93,** 99–121.

Drexhage, K. H. (1974), in: E. Wolf (Ed.), *Progress in Optics,* Vol. 12, New York: North Holland, pp. 165–184.

Drummond, P. D., Shelby, R. M., Friberg, S. R., Yamamoto, Y. (1993), *Nature* **365,** 307–313.

Everitt, H. O. (1992), *Opt. Photonics News* **3** (11), 20–23.

Friberg, S. R., Machida, S., Yamamoto, Y. (1992), *Phys. Rev. Lett.* **69,** 3165–3168.

Glauber, R. J. (1963), *Phys. Rev.* **131,** 2766–2785.

Glauber, R. J. (1965), in: C. DeWitt, A. Blandin, C. Cohen-Tannoudji (Eds.), *Quantum Optics and*

Electronics, New York: Gordon & Breach, pp. 65–83.

Golubev, Y. M., Sokolov, I. V. (1984), *Zh. Eksp. Teor. Fiz.* **87,** 408–416 [*Sov. Phys. JETP* **60,** 234–245].

Haake, F., Walls, D. F., Collet, M. J. (1989), *Phys. Rev. A* **39,** 3211–3220.

Haken, H. (1970), in: S. Flügge (Ed.), *Handbuch der Physik,* Vol. 25, *Light and Matter,* Berlin: Springer-Verlag.

Haroche, S., Kleppner, D. (1989), *Phys. Today* **42** (1), 24–30.

Hasegawa, A., Tappert, F. (1973), *Appl. Phys. Lett.* **23,** 142–144.

Haug, H., Haken, H. (1967), *Z. Phys.* **204,** 262–283.

Haus, H. A., Lai, Y. (1990), *J. Opt. Soc. Am. B* **7,** 386–392.

Imamoglu, A., Yamamoto, Y. (1994), *Phys. Rev. Lett.* **72,** 210–213.

Imoto, N., Haus, H. A., Yamamoto, Y. (1985), *Phys. Rev. A* **32,** 2287–2292.

Jackiw, R. (1968), *J. Math. Phys.* **9,** 339–354.

Jaynes, E. T., Cummings, F. W. (1963), *Proc. IEEE* **51,** 89–95.

Kimble, H. J., Walls, D. F. (Eds.) (1987), Special issue on squeezed states, *J. Opt. Soc. Am. B* **4** (10).

Kitagawa, M., Yamamoto, Y. (1986), *Phys. Rev. A* **34,** 3974–3988.

Kitagawa, M., Imoto, N., Yamamoto, Y. (1987), *Phys. Rev. A* **35,** 5270–5273.

Kuwata-Gonokami, M., Takeda, K., Yasuda, H., Ema, K. (1992), *Jpn. J. Appl. Phys. Lett.* **31,** 99–102.

Lai, Y., Haus, H. A., Yamamoto, Y. (1991), *Opt. Lett.* **16,** 1517–1519.

Landau, L., Peierls, R. (1931), *Z. Phys.* **69,** 56–59.

Landauer, R. (1995), in: *Proceedings of the 4th Symposium on Quantum Nonintegrability,* Drexel University (to be published).

Likharev, K. K., Zorin, A. B. (1985), *J. Low Temp. Phys.* **59,** 347–355.

Loudon, R. (1973), *The Quantum Theory of Light* Oxford, U.K.: Clarendon.

Louisell, W. H. (1973), *Quantum Statistical Properties of Radiation,* New York: Wiley.

Matinaga, F. M., Karlsson, A., Machida, S., Yamamoto, Y., Suzuki, T., Kadota, Y., Ikeda, M. (1993), *Appl. Phys. Lett.* **62,** 443–446.

McCall, S., Hahn, E. L. (1969), *Phys. Rev.* **183,** 457–485.

McCall, S. L., Levi, A. F. J., Slusher, R. E., Pearton, S. J., Logan, R. A. (1992), *Appl. Phys. Lett.* **60,** 289–292.

Messiah, A. (1961), *Quantum Mechanics,* New York: McGraw-Hill.

Meyster, P., Sargent, M. (1991), *Elements of Quantum Optics,* Berlin: Springer-Verlag.

Milburn, J., Walls, D. F. (1983), *Phys. Rev. A* **28,** 2065–2080.

Mollenauer, L. F., Stolen, R. H., Gordon, J. P. (1980), *Phys. Rev. Lett.* **45,** 1095–1098.

Nakazawa, M., Yamada, E., Kubota, H., Suzuki, K. (1991), *Electron. Lett.* **27,** 1270–1272.

Pereira, S. F., Ou, Z. Y., Kimble, H. J. (1994), *Phys. Rev. Lett.* **72,** 214–217.

Purcell, E. M. (1946), *Phys. Rev.* **69,** 681–682.

Sargent, M., III, Scully, M. O., Lamb, W. E., Jr. (1974), *Laser Physics,* Reading, MA: Addison-Wesley.

Schubert, E. F., Vredenberg, A. M., Hunt, N. E. J., Wong, Y. H., Becker, P. C., Poate, J. M., Jacobson, D. C., Feldman, L. C., Zydzik, G. J. (1992), *Appl. Phys. Lett.* **61,** 1381–1383.

Scott-Russell, J. (1844), *Proc. R. Soc. Edinburgh* **1,** 433–434.

Shelby, R. M., Levenson, M. D., Perlmutter, S. H., DeVoe, R. G., Walls, D. F. (1986), *Phys. Rev. Lett.* **57,** 691–694.

Shelby, R. M., Levenson, M. D. (1987), *Opt. Commun.* **64,** 553–556.

Slusher, R. E., Hollberg, L. W., Yurke, B., Mertz, J. C., Valley, J. F. (1985), *Phys. Rev. Lett.* **55,** 2409–2413.

Teich, M. C., Saleh, B. E. A. (1985), *J. Opt. Soc. Am. B* **2,** 275–283.

Tzeng, H. M., Wall, K. F., Long, M. B., Chang, R. K. (1984), *Opt. Lett.* **9,** 499–502.

Vogel, W., Welsch, D. G. (1994), *Lectures on Quantum Optics,* Berlin: Akademie Verlag.

Walls, D. F., Milburn, G. J. (1994), *Quantum Optics,* Berlin: Springer-Verlag.

Watanabe, K., Yamamoto, Y. (1988), *Phys. Rev. A* **38,** 3556–3565.

Wheeler, J. A., Zurek, W. H. (Eds.) (1983), *Quantum Theory and Measurement,* Princeton, NJ: Princeton Univ. Press.

Wu, L. A., Kimble, H. J., Hall, J. L., Wu, H. (1986), *Phys. Rev. Lett.* **57,** 2520–2523.

Yamamoto, Y., Machida, S., Igeta, K., Horikoshi, Y. (1989), in: J. H. Eberly, L. Mandel, E. Wolf (Eds)., *Coherence and Quantum Optics VI,* New York: Plenum, pp. 1249–1255.

Yamamoto, Y., Machida, S., Nilsson, O. (1991), in: Y. Yamamoto (Ed.), *Coherence, Amplification and Quantum Effects in Semiconductor Lasers,* New York: Wiley, pp. 461–537.

Yamamoto, Y., Machida, S., Richardson, W. H. (1992), *Science* **255,** 1219–1224 (and references therein).

Yamamoto, Y., Slusher, R. E. (1993), *Phys. Today* **46(6),** 66–73.

Yamamoto, Y., Bjork, G., Karlsson, A., Heitmann, H., Matinaga, F. M. (1993), *Int. J. Mod. Phys. B* **7,** 1653–1695.

Yuen, H. P. (1976), *Phys. Rev. A* **13,** 2226–2241.

Zakharov, V., Shabat, A. (1971), *Zh. Eksp. Teor. Fiz.* **61,** 118–134 [(1972), *Soviet Phys. JETP* **34,** 62–69].

Further Reading

Braginsky, V. B., Khalili, F. Y. (1992), *Quantum Measurements,* New York: Cambridge Univ. Press.

Drummond, P. D., Shelby, R. M., Friberg, S. R., Yamamoto, Y. (1993), *Nature* **365,** 307–313.

Vogel, W., Welsch, D. G. (1994), *Lecture on Quantum Optics,* Berlin: Akademie Verlag.

Walls, D. F., Milburn, G. J. (1994), *Quantum Optics,* Berlin: Springer-Verlag.

Yamamoto, Y., Machida, S., Richardson, W. H. (1992), *Science* **255,** 1219–1224.

Yamamoto, Y., Slusher, R. E. (1993), *Phys. Today,* **46** (6), 66–73.

QUANTUM OPTOELECTRONIC DEVICES

Jun-ichi Nishizawa, *Tohoku University, Sendai, Japan*
Makoto Minakata, *Shizuoka University, Hamamatsu, Japan*

	Introduction	339
1.	**General Background**	339
1.1	Historical Aspects	339
1.2	Optoelectronic Devices	341
1.2.1	Light Sources	341
1.2.2	Photodetectors	341
1.2.3	Optical Fibers	342
1.2.4	Optical Modulators	342
1.2.5	Optical Logic and Bistable Devices	342
2.	**Light Sources: Laser Diodes and Light-Emitting Diodes**	343
3.	**Linear Optical Devices**	344
3.1	Optical Amplifiers	344
3.2	Optical Modulators	345
3.2.1	Optical Waveguides	345
3.2.2	Electro-optic Modulators	347
3.2.3	Electroabsorption Modulators	348
3.2.4	Other Types of Modulators	351
3.3	Optical Detectors	352
4.	**Nonlinear Optical Devices**	353
4.1	Second-Harmonic Generation	353
4.2	Raman Lasers	354
5.	**Optical Logic, Optical Bistable, and Optical Memory Devices**	356
5.1	Optical Logic Devices	356
5.2	Optical Bistable Devices	356
5.3	Optical Memory Devices	357
6.	**Photonics and Optoelectronic Integrated Circuits**	358
7.	**Summary**	361
	Glossary	362
	Works Cited	363
	Further Reading	363

INTRODUCTION

Quantum optoelectronics describes a new-generation technology that integrates quantum electronics and optics. In general, optoelectronics also encompasses their boundary fields and applications, as shown in Tables 1 and 2. These related disciplines include, among others, optical-fiber communication, optical-signal processing, space optical communication, optical spectroscopy, applied optical measurements, medical applications, optical power applications, and laser irradiation. All these technologies have rapidly developed and are now widely used in many areas such as home electronic equipment, office automation, and industrial electronics. This article summarizes the historical development, as well as the present status, of optoelectronic devices such as laser diodes, light-emitting diodes, photodetectors, optical modulators, and optical amplifiers. Finally, the latest developments in photonics and optoelectronic integrated circuits and their applications are examined.

1. GENERAL BACKGROUND

1.1 Historical Aspects

In the first days of optoelectronics, electroluminescent materials were used coupled with photoconductors. These materials were later transformed into coupled light-emitting diodes or laser diodes with photodiodes and called photocouplers. The discovery of lasers in the early 1960s laid the foundation for the establishment of optoelectronics (see Laser Physics). Indeed, the development of coherent optics and quantum electronics paved the way for numerous new technologies such as optical communication (*q.v.*), optical recording and memories (see Optical Storage), laser fusion (see Fusion Technologies), and medical applications (see Medical

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Table 1. Optoelectronic devices and related fields.

Light sources	Coherent sources	Lasers	Gas and solid lasers
			Laser diodes
		Nonlinear optical devices	Tunable and Raman lasers
			Parametric oscillator
	Incoherent sources	Signal processing	LED, LED array
		Display	Liquid-crystal display
			Photoluminescence
		Irradiation	
Optical detectors	Planar detectors (single plane)	Photoconductive type	*p-i-n* diode, APD
		Photovoltaic type	Solar battery
		Thermal-electric type	Thermistor
		Pyroelectric type	
	Two-dimensional detectors	One-dimensional (line) sensors	Diode array
		Two-dimensional (surface) sensors	Diode array
Optical devices	Active devices	Bistable devices	
		Modulator, deflector	
		Optical sensor devices	
	Passive devices	Optical components	
		Component for optical communications	

Table 2. Applications of optoelectronics.

Optical-fiber communication	Long distance, broadband and digital communication systems	
	Analog transmission systems	
	Optical interconnections	Factory Car Train Airplane Computer Others
	Fiber guides (solar guide)	
Optical-signal processing	Holography, optical disk memory, holographic memory	
	Video laser disk, compact disk (CD, DAD)	
	POS system, laser printer	
Optical spectroscopy	Spectroscopy analysis	
	Isotropic measurement and separation	
	Plasma diagnosis	
Applied optical measurement	Standard of length	
	Laser radar	Atmosphere pollution
	Laser gyro	Rotation
	Laser transit	Alignment
	Interference, beam modulation or pulse echo method	Distance Velocity
Medical applications	Eye surgery	
	Fiber scope	
Photochemistry	Photochemical vapor deposition	
	Photochemical etching	
Space optical communication	Space communication systems	
	Remote sensing	
Optical power	Heat source	Welding Ablation Cutting Milling Drilling
	Laser irradiation, laser arts	

USE OF LASERS), which make heavy use of different optoelectronic devices (see LASER TECHNOLOGY).

Since the early 1970s, optical communication systems were widely developed using single-mode fibers, laser diodes, avalanche or *p-i-n* photodiodes, and well-designed LSI optoelectronic circuits. Recently, large-capacity optical communication systems have been realized by newly developed erbium-doped optical amplifiers and photonic integrated circuits (see FIBER OPTICS; AMPLIFIERS, OPTICAL). Global communication networks are being constructed using these systems.

1.2 Optoelectronic Devices

1.2.1 Light Sources In late 1962, the first semiconductor injection lasers were reported, almost simultaneously, by four groups, headed by Hall (1962), Nathan (1962), Quist (1962), and Holonyak (1962). While earlier work had considered the possibility of light amplification in semiconductors (Watanabe and Nishizawa, 1957; Boyle and Thomas, 1960), achieving high-efficiency electroluminescence from GaAs *p-n* junctions was the event that prompted accelerated efforts to demonstrate a semiconductor laser. Pulsed radiation of 9.84 μm was obtained from a liquid-nitrogen–cooled forward-biased GaAs *p-n* junction. That same year, Pankove *et al.* (1962) demonstrated the GaAs light-emitting diode (LED). The realization of a continuous-wave (cw) laser at room temperature was later achieved by Panish *et al.* (1970), using double heterojunctions. This structure was independently first proposed by Kroemer (1963) and by Alferov and Kazarinov (1963). These light sources represented real breakthroughs in the practical use of optoelectronics, because all the devices have very small size (about 200 μm) and high efficiency and are best suited for integrated circuits. In 1971, Kogelnik and Shank (1971) proposed the distributed-feedback (DFB) laser and demonstrated the first experimental results for dye lasers. The concept was later implemented for semiconductor lasers by Nakamura *et al.* (1973). Thus, control of the oscillation frequency became possible. In 1978, Holonyak *et al.* (1980) fabricated the first quantum-well (QW) laser, which introduced quantum size effects into heterostructures and revealed excellent properties such as low threshold current, high power capability, and low noise. Tsang *et al.* (1979) achieved further reduction of the threshold current by combining quantum wells with graded-index (GRIN) distribution. The graded-index waveguide itself was previously demonstrated using organic film (Nishizawa and Otsuka, 1972), glass, and semiconductor layers (Nishizawa, 1973).

Recently, other important light sources have been developed from novel optoelectronic-device structures. Vertical-cavity surface-emitting (SE) lasers with submicron diameters have produced the lowest threshold current, the highest quantum efficiency, and the highest modulation speed (Jewell *et al.*, 1990). Furthermore, by the employment of extremely short–cavity-length and high-reflectivity distributed–Bragg-reflector (DBR) mirrors, some lasing characteristics exhibit unique features, like extremely low threshold current, single longitudinal mode, and a circular spot profile of output light (Kuchibhotla *et al.*, 1991). In addition, the ultrasmall SE lasers provide high-density two-dimensional laser arrays, which are essential for parallel optical-signal processing, optical interconnections, and optical computing. In fact, the control of spontaneous emission by microcavities is becoming of great interest for realizing the ultimate performance of SE lasers (see OPTICS, QUANTUM). Improved characteristics have also been achieved by compressively strained quantum-well (QW) lasers (Thijs *et al.*, 1991). Lasers employing tensile-strained QW show a high and a narrow gain spectrum. Such optoelectronic devices are currently of considerable interest because of their dramatically improved performance as compared with bulk and lattice-matched QW devices.

1.2.2 Photodetectors The development of efficient light detectors, like the *p-i-n* photodiode and the avalanche photodiode (APD), has prompted their wide use in optical communication. In this application, detectors receive the transmitted optical pulses and convert them into electronic pulses, with as little loss as possible. The performance requirements are high sensitivity, low noise, wide bandwidth, high reliability, and low cost. Another common application of photodetectors is the monitoring of laser transmitters. For optical communication and signal-processing systems, there is usually a need for high-speed detectors. For several other applications, high gain is necessary. Consequently, bandwidth and gain are fundamental trade-offs, and so the final application determines which type of photodetector is the most suitable. Photodetection is sometimes associated with demodulation, when a high-frequency optical signal is converted into a time-varying electrical signal for further processing and rectification. Photodetectors are used to detect optical signals over a wide range of the optical spectrum (see SIGNAL PROCESSING, OPTICAL).

Recently, coherent optical detection is rapidly gaining importance in long-distance optical communication systems. In this receiving technique, also referred to as *heterodyne detection,* a weak signal is mixed with a strong local-oscillator (LO) wave at a close enough frequency, resulting in an effective signal gain through coherent phase interference. The gain is proportional to the LO amplitude. Heterodyning allows the detection of very weak optical signals, which would not be normally converted into electrical signals through conventional direct-detection techniques.

1.2.3 Optical Fibers Depending on the application, light is transmitted through various media. In free-space communication the medium is air or vacuum. In guided-wave systems, light is transmitted by optical fibers or by dielectric waveguides, usually made of compound semiconductors. The choice of dielectric materials depends largely on the application. A typical fiber is an optical guide with an inner, high-index core and an outer, low-index cladding. The fiber properties are, to a large extent, determined by the core and cladding diameters and refractive indices. Optical fibers have their lowest loss at 1.55 μm. However, with rare-earth doping, the lowest dispersion and optical amplification have been observed at 1.3 μm. For long-distance fiber communication, the materials of choice are InP-based lattice-matched materials, or GaAs- and InP-based mismatched materials. For short-distance fiber transmission, GaAs/AlGaAs is used, since fiber losses at 0.8 μm remain within acceptable limits. Fibers are now being developed for longer wavelengths. Similarly, as semiconductor sources are developed for emission in the visible part of the spectrum, using III-V and II-VI compounds, appropriate guiding materials will also be developed (see FIBER OPTICS).

Recently, erbium-doped fiber amplifiers (EDFA) have been demonstrating their promise for advanced optical communication by allowing the transmission of soliton pulses over extremely long distances without optoelectronic repeaters (Nakazawa *et al.,* 1993).

1.2.4 Optical Modulators In the early days, one of the most promising devices for modulating a laser beam was based on the nonlinear electro-optic effect, discovered by Kerr in 1875. This effect produces a convenient and widely used way of controlling the intensity of propagating optical beams. This type of modulation is used in the *Q* switching of lasers for generating powerful optical pulses, mode locking, and optical beam deflection as well as modulation. The LD and the LED have been successfully used for optical communication since the early 1970s, because these devices can easily achieve direct modulation of the generated light corresponding to speech and video signals (see OPTICAL COMMUNICATIONS). However, direct modulation using the LD has two serious problems. One is that the modulation speed is restricted to less than 10 GHz, because of the relaxation oscillation based on the lifetime of carriers and photons. The other is that chirping produces spectral broadening of the modulated optical signal, restricting the modulation speed. Future optical transmission and processing systems will soon demand higher-speed modulation. Recently, a new type of optical modulators has been developed using multiple-quantum well (MQW), waveguide, and traveling-wave electrodes. These devices exhibit modulation speeds in excess of 20 GHz and efficient and low drive power.

1.2.5 Optical Logic and Bistable Devices For the ideal optical-signal processing required in optical computing, it is important to achieve control of an optical beam with another optical beam. Lasher (1963) and Fowler (1963) first demonstrated mutually quenched injection lasers as bistable devices. The optical logic devices, based on the competing phenomenon between modes in the LD, were built a year later by Nishizawa *et al.* (1965). The response to the change was measured at 98 GHz, which means that the time constant is very close to 10^{-12} s. These devices can be used as logic and memory devices.

More recently, Miller *et al.* (1985) produced a self–electro-optic effect device (SEED), a structure composed of MQW *p-i-n* diodes, which has a low threshold of transition light power and relatively high response. This device exhibits photonic switching, bistability, and optically induced oscillations due to the negative resistance in the photocurrent.

2. LIGHT SOURCES: LASER DIODES AND LIGHT-EMITTING DIODES

By and large, the most important optoelectronic device is the semiconductor laser or laser diode (LD). The LD, in principle, is an optical waveguide terminated by optical mirrors or reflecting faces that form a resonating cavity. The degree of confinement of the optical modes in the optical waveguide, which is usually made of dielectric or semiconductor materials, depends on the index of refraction of the surrounding media or cladding. The LD is small in size, efficient, and feasible to integrate. It has other excellent properties like high-speed switching and quenching, direct modulation of the output light, and light amplification.

The wavelength generated by the laser diode or the light-emitting diode (LED) can be varied from 0.3 to 30 μm by changing the combinations of semiconductor materials. The physical principles of the LD or LED are explained in Fig. 1, which shows the energy bands at the boundary between an *n*-type and a *p*-type region. There is a transition region, only a few angstroms thick, where population inversion exists and electrons in the conduction band can fall down and occupy holes in the valence band, emitting photons

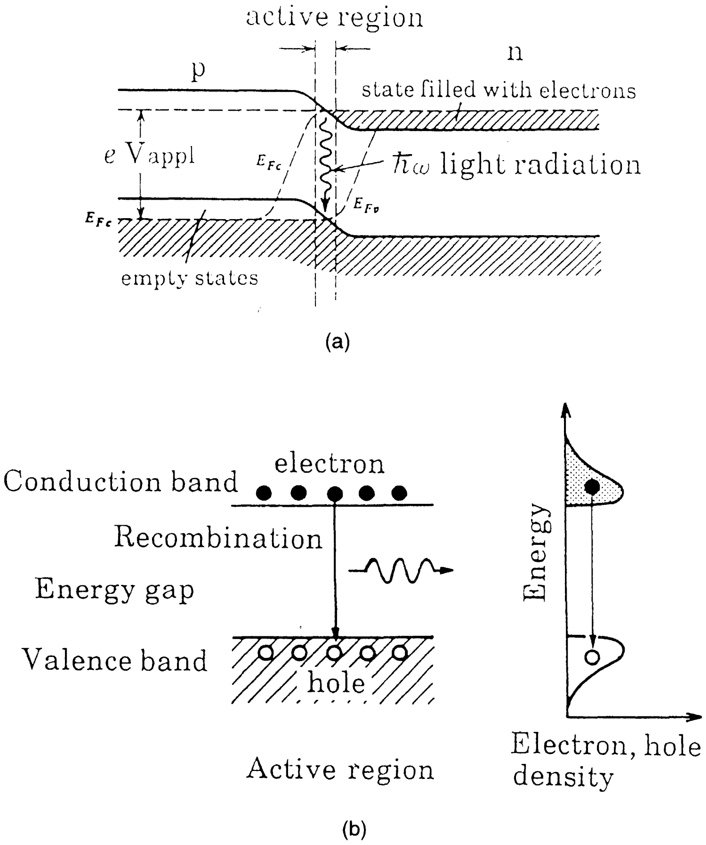
active region
p
n
state filled with electrons
eV_{appl}
E_{Fc}
$\hbar\omega$ light radiation
E_{Fv}
E_{Fc}
empty states
(a)
Conduction band
electron
Recombination
Energy gap
Valence band
hole
Active region
(b)
Energy
Electron, hole density

FIG. 1. Basic properties of a laser diode: (a) energy-band diagram and (b) recombination and population inversion.

of electromagnetic radiation. The material is in contact with a power supply that makes the electrons in the *p*-type region move toward the junction. Large numbers of electrons cross the junction when the current is flowing toward the *p*-type region, recombining with holes there. This type of laser is known as a junction diode laser, because electrons are injected into the junction region. An example of this device structure is shown in Fig. 2. Laser action takes place in the thin junction region and the emission is made into a bidirectional beam, as a result of the electromagnetic resonance by the highly flat polished end faces, which act as mirrors. The emitted light beam, multiply reflected at both end mirrors, can be changed into a polarized coherent light beam by quantum mechanically induced stimulated emission, which generates the same phase light corresponding to the phase of the standing light wave in the junction region (see Lasers, Semiconductor).

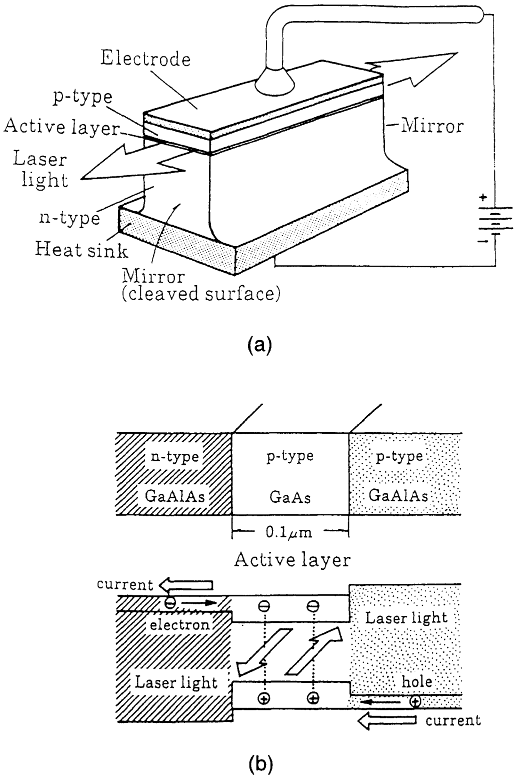

FIG. 2. Construction and operation of a laser diode: (a) system structure and (b) band diagram and carrier mechanism.

The LED is similar to the LD, but it is a structure without the stringent requirement of optical gain and optical resonator and, as a result, it has no stimulated emission. In the early days, the level of spontaneous emission from LED devices was very small. It is only recently, with the development of wider-band-gap II–VI semiconductors, that visible light emission from a biased *p-n* junction has been demonstrated. Some obvious advantages of the LED as a light source are simpler fabrication, lower cost, and simpler drive circuitry. The LED has found wide applications such as indicator lamps, display elements, and sensors (see Light-Emitting Diodes).

3. LINEAR OPTICAL DEVICES

3.1 Optical Amplifiers

The principle of the optical amplifier was already well known in the 1960s (Nishizawa, 1977). When two LDs are closely separated and one of them acts as an amplifier of the light output of the other, then the polarization direction of the incident light is maintained after amplification even if the amplifier's preferable polarization direction is different. Later, both traveling-wave–type and Fabry–Pérot–type amplifiers have been demonstrated and widely studied in the 1980s (see Amplifiers, Optical).

The LD amplifiers (LDA) are built with the same fabrication process as the LD. The traveling-wave–type LDA has nearly perfect antireflection coating on both end faces. When the input optical signal propagates in one direction along the waveguide in the LDA, the input power is gradually amplified without oscillation by injection of dc current. The typical gain for this device is 20–30 dB. The saturation output power is 8–10 dBm but the polarization dependency of the gain is relatively large. The gain × bandwidth product is in excess of 1 THz. To obtain a low noise figure, the beat noise between the optical signal and the spontaneously emitted light needs to be suppressed by using an optical narrow-band-pass filter.

For the resonant-type LDA, the optical frequency of the input light is restricted to coincide with the optical resonant frequency determined by an LD Fabry–Pérot resonator.

This LDA is similar to the LD in basic principles. The typical gain for this device is 15–20 dB. The saturation output power is −10 to −5 dBm. However, the gain × bandwidth product is as narrow as 5 GHz. The noise figure is determined by the beat noise between the optical signal and the spontaneously emitted light.

The fiber amplifier being recently developed is the revival of the concept of a glass laser and it is named for the initials of Er^{3+}-doped fiber amplifier (EDFA). The construction of the EDFA and the corresponding energy-band diagram are shown in Fig. 3. The EDFA operates at a wavelength of 1.55 μm and is composed of dielectric multiple-layer filters and an Er^{3+}-doped fiber that is pumped by an InGaAsP LD on a wavelength of 1.48 or 0.98 μm. This infrared radiation is absorbed by the Er^{3+} atoms, pumping them to a high energy level. When the excited electrons of the erbium atoms occupy a high energy level, if an optical-signal electromagnetic wave comes into the fiber, stimulated emission occurs. As a result, the power of the signal increases gradually along the fiber; that is, the optical signal is amplified. The EDFA has excellent characteristics, as follows. The typical gain is in excess of 30 dB; the gain × bandwidth product is more than 10 THz, and the noise figure less than 4 dB. Furthermore, the past few years have seen stunning developments in nonlinear optical communication with soliton propagation making full use of the advanced EDFA technology now available (Desurvire, 1994).

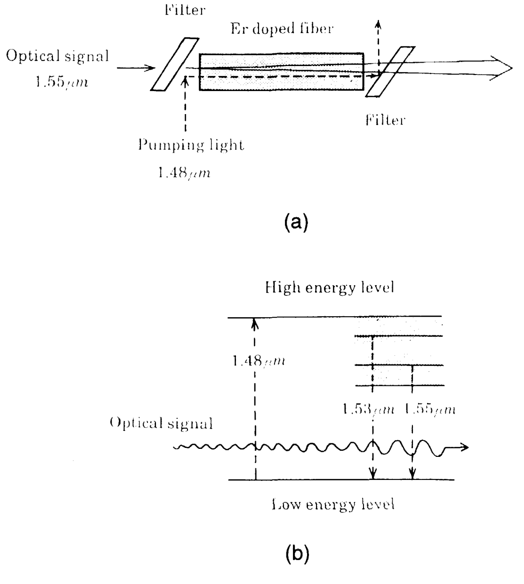

FIG. 3. (a) Basic properties of an optical amplifier and (b) band diagram of an erbium-doped fiber amplifier (EDFA).

Optical solitons were first observed in 1980 by Mollenauer *et al.* (1980). Because pulses as short as a few picoseconds can be transmitted in the soliton regime without broadening, it is evident that fiber-optic communication can potentially operate at ultrahigh bit rates such as 100 Gbit/s. Indeed, under the effect of periodic amplification, solitons can undergo significant power changes without losing their integrity. This property explains the success of EDFA-based experiments, in which the optical gain is confined to lumped fiber sections, such as the one recently reported by a Japanese team (Nakazawa *et al.*, 1993) where a 1000-km straight-line soliton transmission was achieved with four multiplexed channels modulated at a total of 40 Gbit/s.

3.2 Optical Modulators

There are many types of optical modulation, including intensity modulation, phase modulation, frequency modulation, and polarization rotating modulation. Analog or digital signals modulate the light beam. Many physical effects can be used for optical modulation and switching, such as linear (Pockels) and nonlinear (Kerr) electro-optic, magneto-optic, acousto-optic, and electroabsorption, including the quantum-confined Stark effect (QCSE). Even the Zeeman and the pressure effects can be used. These effects, phenomena, functions, and materials are summarized in Table 3. Among them, the most promising high-speed modulators are devices that make use of the Pockels effect or the QCSE.

3.2.1 Optical Waveguides An optical waveguide is composed of a core region of dielectric material through which light is propagated, surrounded by dielectric regions or air having a smaller dielectric constant, as shown in Fig. 4. This structure has been widely used for optical devices, because of the reflection-loss reduction in bulk devices, the compatibility with fibers, and the efficient modulation.

Table 3. Summary of physical effects and materials for optical modulators and optical switches.

Effect	Phenomenon	Function	Material
Electro-optic (Pockels)	Refractive-index change	Modulation: phase, amplitude, frequency, polarization Changes: reflection pass, mode	$LiNbO_3$ $LiTaO_3$ GaAs InP InGaAsP
Acousto-optic (photoelastic)	Refractive-index change	Changes: reflection, mode Modulation: frequency, polarization	$LiNbO_3$ ZnO As_2S_3
Magneto-optic (Faraday)	Polarization rotation	Isolator, circulator, modulation: polarization	YIG GaBiYIG FR-5
Opto-optic (nonlinear)	Polarization induction	SHG, parametric amplification, Raman scattering	$LiNbO_3$ $LiTaO_3$ GaP
Thermo-optic	Refractive-index change	Modulation: amplitude, phase	Glass, everything
Electroabsorption Franz–Keldysh, exciton absorptive, free-carrier absorptive	Optical power absorption, absorption-edge shift	Modulation, amplitude	GaAs InP InGaAsP
Quantum-confined Stark	Exciton absorption, refractive-index change	Modulation: amplitude, phase	GaAs InP InGaAsP

There are two types of waveguides: planar waveguides, which guide light in the vertical direction but provide no lateral confinement [Fig. 4(a)], and channel waveguides, which provide three-dimensional confinement [Figs. 4(b) and 4(c)]. For example, channel waveguides in $LiNbO_3$ and GaAs are mainly used for high-efficiency devices. An optical wave propagating will be contained in the core if this has a higher refractive index than the surrounding region. It is analogous to the silica optical fiber (see OPTICAL COMMUNICATIONS). There are two types of profile for the refractive-index distribution in the waveguides. One is a step-index profile made by epitaxial growth and sputtering. The other is a graded-index profile made by diffusion, ion exchange, and ion implantation. In the step-index waveguide, light propagates in the core region by periodic repetition of internal total reflection (ITR), as long as the incidence angle to the substrate is greater than the ITR critical angle ϕ, according to Snell's law:

$$\sin\phi = n_s/n, \quad n > n_s. \tag{1}$$

In the graded-index waveguide, the light is continuously refracted on account of the decreasing refractive index so that light propagates describing a circular arc.

In titanium-strip diffused $LiNbO_3$ devices, the core region has a refractive index higher than the surrounding region by 0.1 to 1%. The waveguide can be fabricated so that the fundamental or lowest-order mode propagates. In practice a titanium pattern is formed by using photolithographic techniques and Ti deposition on the $LiNbO_3$ substrate. Then the crystal is heated in a furnace to diffuse Ti into $LiNBO_3$ at 1000°C for 7 hours. Thus, three-dimensional waveguides are fabricated. Typical waveguide dimensions are 3–8 μm wide and 2–6 μm deep. With epitaxial layers, the necessary index difference can be achieved either by lattice-matched growth of an epitaxial layer with different composition than the substrate or

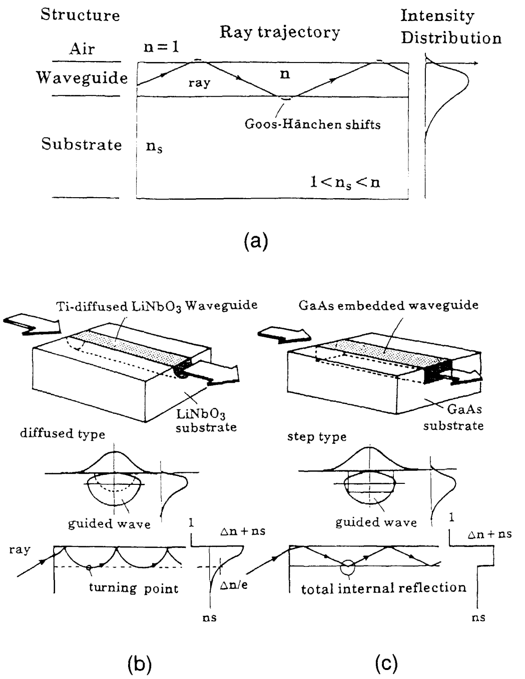
Structure
Air
Waveguide
Substrate
n = 1
Ray trajectory
ray
n
Goos-Hänchen shifts
n_s
$1 < n_s < n$
Intensity Distribution
(a)
Ti-diffused LiNbO3 Waveguide
LiNbO3 substrate
diffused type
guided wave
1
Δn + ns
ray
turning point
Δn/e
ns
(b)
GaAs embedded waveguide
GaAs substrate
step type
guided wave
1
Δn + ns
total internal reflection
ns
(c)

FIG. 4. Configuration and operation principle of optical waveguides: (a) planar, (b) Ti-diffused LiNbO$_3$, and (c) embedded GaAs.

with the same material but different doping level.

3.2.2 Electro-optic Modulators An important consequence of dielectric properties is that for a linearly polarized plane wave propagating in the *z* direction, its polarization vector is defined by the direction of the electric field. Furthermore, the phase velocity of the propagating beam depends on the direction of polarization of the electric field. This phenomenon is known as birefringence and it is only observed in anisotropic media. If a linearly polarized wave is launched in the *z* direction, with equal electric field components in the *x* and *y* directions, it can be shown that the propagation constants in the *x* and *y* directions are different as a result of birefringence. This is the basis of electro-optic modulation, and the phenomenon can be used to build efficient field-induced modulators with appropriate crystals.

The linear (Pockels) electro-optic effect provides a change in refractive index proportional to an applied electric field. The way in which this index change produces intensity modulation depends on the device configuration. The optic axis is defined as the direction in a crystal in which, when light is propagated in the same direction, the refractive index is independent of the polarization direction of light. For simplicity, the device is usually fabricated such that the incidence face is parallel to the optic axis. The polarized light upon entering the crystal can be divided into two components: the ordinary component, whose electric field is parallel to the optic axis (and which corresponds to the refractive index n_o) and the extraordinary component, perpendicular to the optic axis (and corresponding to the refractive index n_e). The propagation velocities of these two components within the crystal are different. In general, as they emerge from the crystal, the two components are out of phase. Thus, the output light is elliptically polarized, as shown in Fig. 5. In this figure, the expression for the induced phase shift is

$$\Delta\phi = \frac{2\pi}{\lambda_0} l\Delta n = \frac{\pi n_o^3 \gamma}{\lambda_0 d} V. \tag{2}$$

The electro-optical effect is described by a tensor relation. The linear change in the coefficients of the index ellipsoid due to applied fields (E_j) along the principal crystal axes is

$$(\Delta n)_i = -\frac{n^2}{2} \sum \gamma_{ij} E_j, \tag{3}$$

where $i = 1$–6 and γ_{ij} is a 6×3 electro-optic tensor. This relation can be contracted

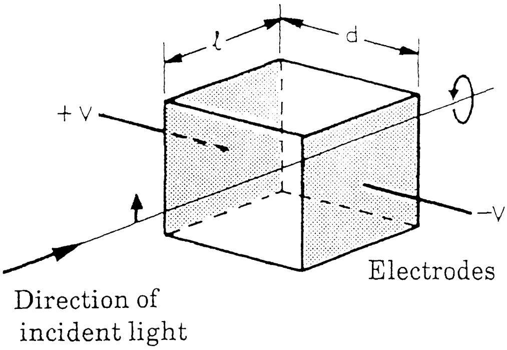
l
d
+V
−V
Electrodes
Direction of incident light

FIG. 5. A bulk-type optical phase modulator using the linear electro-optic (Pockels) effect.

with the electro-optic tensor to write the six values of Δn as the elements of a symmetric 3×3 matrix. Typical values for the electro-optic coefficients of such materials are shown in Table 4.

A typical system arrangement of an electro-optic amplitude modulator using $LiNbO_3$ is shown in Fig. 6. It consists of an electro-optic crystal placed between two crossed polarizers, which in turn are at a 45° angle. The output light from the crystal is elliptically polarized and its amplitude changes correspond to the applied voltage. The intensity of the output light can be described by

$$I = I_0 \sin^2(\pi V/2V_\pi), \tag{4}$$

where V_π is the so-called half-wave voltage. As its name implies, it is the voltage necessary to produce a retardation ($\Delta\phi$) of half a wavelength between the ordinary and extraordinary rays. From Eq. (2), we obtain that

$$V_\pi = \frac{\lambda_0 d}{(n_e^3\gamma_{33} - n_o^3\gamma_{13})l}. \tag{5}$$

If $d = 1$ mm and $l = 2$ cm, the value of V_π is about 150 V at a 633-nm wavelength.

There are numerous types of electro-optic modulator systems. These include, among others, bulk-type amplitude modulators, waveguide-type phase or intensity modulators, Mach–Zehnder interferometers, and directional-coupler–type modulators, as illustrated by Fig. 7. Typical values of the electro-optic constants for different materials are summarized in Table 4. A detailed description of such optoelectronic devices is given in a separate article (see MODULATORS AND DEMODULATORS, OPTICAL).

3.2.3 Electroabsorption Modulators

There are mainly three types of modulators based on electroabsorption phenomena, involving the Franz–Keldysh effect, exciton absorption, and free-carrier absorption.

When a reverse-bias voltage is applied to a *p-n* junction or Schottky junction, the optical absorption edge of semiconductors such as AlGaAs and InGaAs shifts to longer wavelengths, as shown in Fig. 8. Thus, the presence of a strong electric field results in the absorption of photons with energies less than the band gap of the semiconductor. This phenomenon is called the Franz–Keldysh effect, and an intensity modulator can be built making use of the difference in the absorption coefficients near the absorption edge, corresponding to the applied voltage. With the light wavelength chosen slightly smaller than the band gap, the light is fully transmitted at zero bias while it is attenuated with the application of a field. To illustrate such a system, the typical characteristics of an InGaAsP/InP substrate waveguide modulator are indicated, as follows: The modulation voltage is 12 V, the bandwidth is 8 GHz, and the insertion loss is 6 dB at a 1.58-μm wavelength.

A reverse-bias *p-i-n* diode with a multiple quantum-well (MQW) structure shows a large electroabsorption based on the quantum-confined stark effect (QCSE), as illustrated by Fig. 9. This effect is caused by

Table 4. Electro-optic constants and refractive index for different materials.

Material	Pockels constants (10^{-12} m/V)	Refractive index	$n^3\gamma$ (10^{-12} m/V)
$LiNbO_3$	$\gamma_{33} = 30.8$ $\gamma_{13} = 8.6$ $\gamma_{22} = 3.4$ $\gamma_{42} = 28$	2.286 (n_o) 2.200 (n_e) ($\lambda = 633$ nm)	328
$LiTaO_3$	$\gamma_{33} = 30.3$ $\gamma_{13} = 5.7$	2.176 (n_o) 2.180 (n_e) ($\lambda = 633$ nm)	314
GaAs	$\gamma_{41} = -1.5$	3.6 ($\gamma = 900$ nm)	−70
ZnS	$\gamma_{41} = 2$	2.368 (n_o) 2.363 (n_e) ($\lambda = 600$ nm)	27
SiO_2	$\gamma_{41} = 0.20$ $\gamma_{11} = -0.47$	1.546 (n_o) 1.555 (n_e) ($\lambda = 546$ nm)	3.4

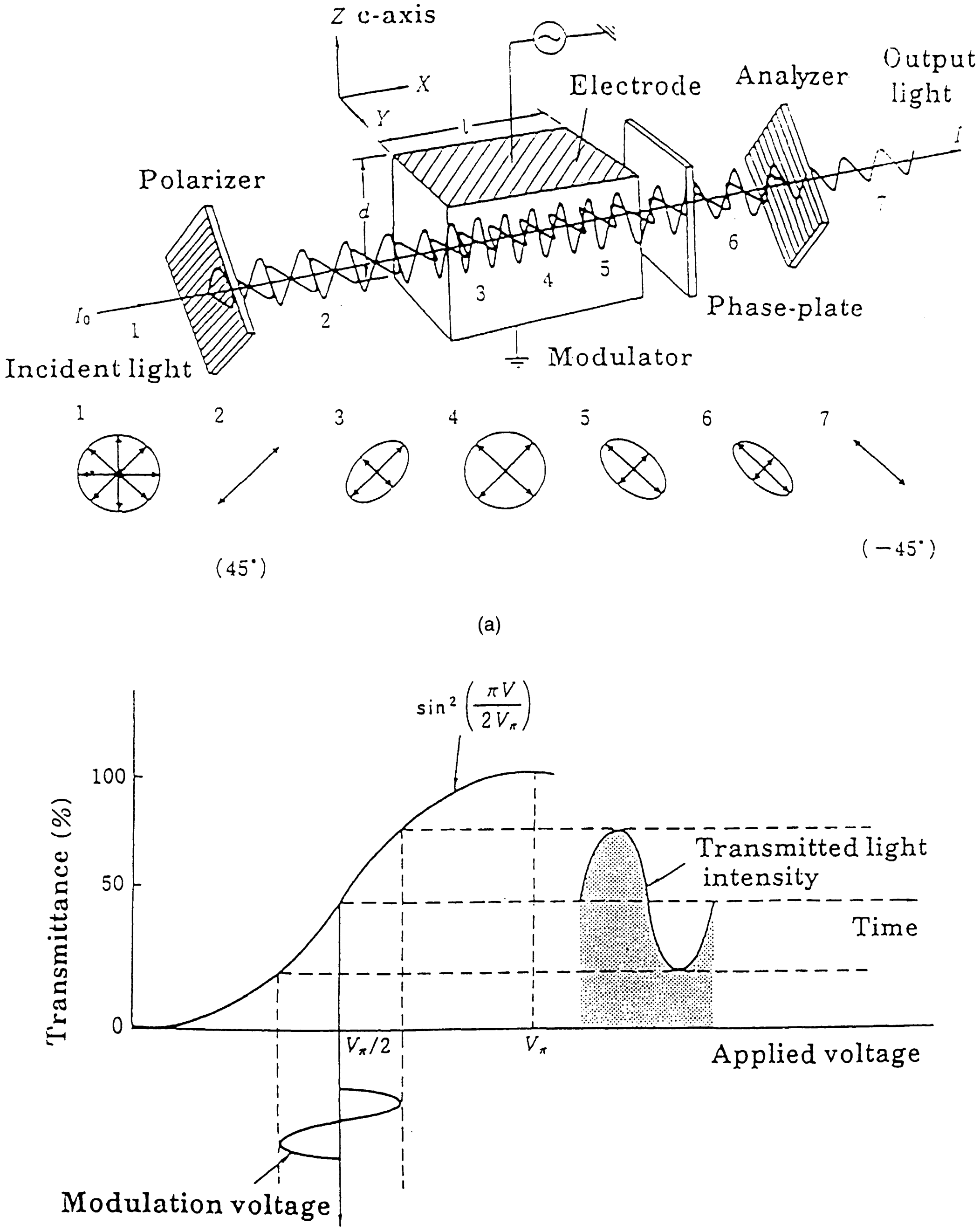
Z c-axis
X
Y
l
d
Electrode
Analyzer
Output light
I
Polarizer
I0
1
2
3
4
5
6
7
Incident light
Modulator
Phase-plate
1
2
3
4
5
6
7
(45°)
(−45°)
$\sin^2\left(\frac{\pi V}{2V_\pi}\right)$
100
50
0
Transmittance (%)
Transmitted light intensity
Time
$V_\pi/2$
V_π
Applied voltage
Modulation voltage

(a)

(b)

FIG. 6. A $LiNbO_3$ optical amplitude modulator: (a) system configuration and (b) operation principle.

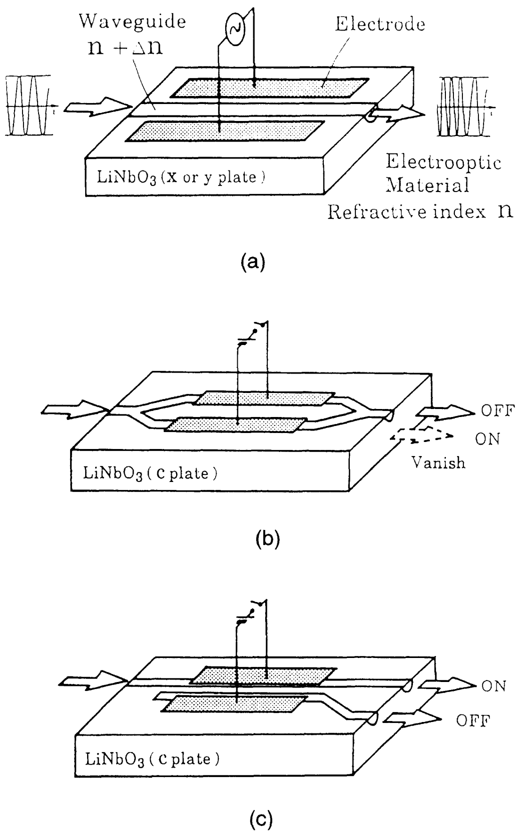

FIG. 7. Various examples of waveguide-type modulators: (a) phase modulator, (b) Mach–Zehnder amplitude modulator, and (c) directional coupler modulator.

Coulomb-coupled electron–hole pair or exciton absorption in the MQW structure, which enables the fabrication of an efficient amplitude modulator. The MQW layer, which is part of a superlattice heterostructure as shown in Fig. 9(b), consists of two periodically laminated materials with each layer usually less than 200 Å. In this two-dimensional structure, the exciton binding energy is about three times that in a three-dimensional or bulk structure. That happens because electrons and holes are confined to a region defined by the quantum-well width, so that the overlap of their wave functions is increased. This results in an increase of the oscillator strength of the interband transitions between the discrete-energy electron and hole bound states, which are produced by the size quantization. Consequently, strong resonances corresponding to the heavy-hole and light-hole transitions are seen near the band edge of the quantum-well material even at room temperature. Thus, a large exciton absorption exists at room temperature and the excitonic absorption edge shifts to longer wavelength (the so-called Burstein shift) corresponding to the reverse-bias applied voltage. The absorption-edge spectrum is abrupt compared with that of the Franz–Keldysh effect and its response speed is as high as 40 GHz. As a result, high-speed amplitude modulation (greater than 22 GHz) with low drive voltage (less than −4 V) and insertion loss less than 18 dB at a 1.54-μm wavelength have been reported for an InGaAlAs/InAlAs MQW waveguide modulator.

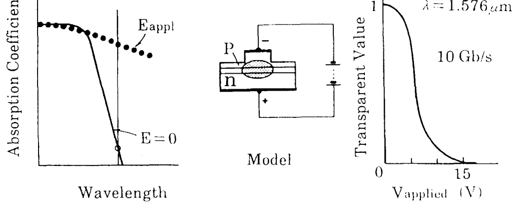

FIG. 8. System configuration and basic characteristics of an electroabsorption modulator based on the Franz–Keldysh effect.

Another mechanism used for electroabsorption modulation is free-carrier absorption. This phenomenon involves the absorption of a photon by the interaction of a free carrier within an energy band, which is raised to a higher energy level as a result. The transition of the carrier to a higher energy level within the same valley must conserve momentum. This momentum change is provided by optical or acoustic phonons or by impurity scattering. Free-carrier absorption is usually present in the long-wavelength region of the spectrum as a monotonic increase in the absorption of the form

$$\lambda^p, \quad \text{where } p = 1.5\text{–}3.5. \tag{6}$$

The value of p depends on the nature of the scattering mechanism for momentum conservation, and the absorption coefficient is proportional to the free-carrier concentration N, as follows:

$$\alpha = \frac{Nq^2\lambda^2}{4\pi m^* n_r c^3 \epsilon_0}\left\langle\frac{1}{\tau}\right\rangle, \tag{7}$$

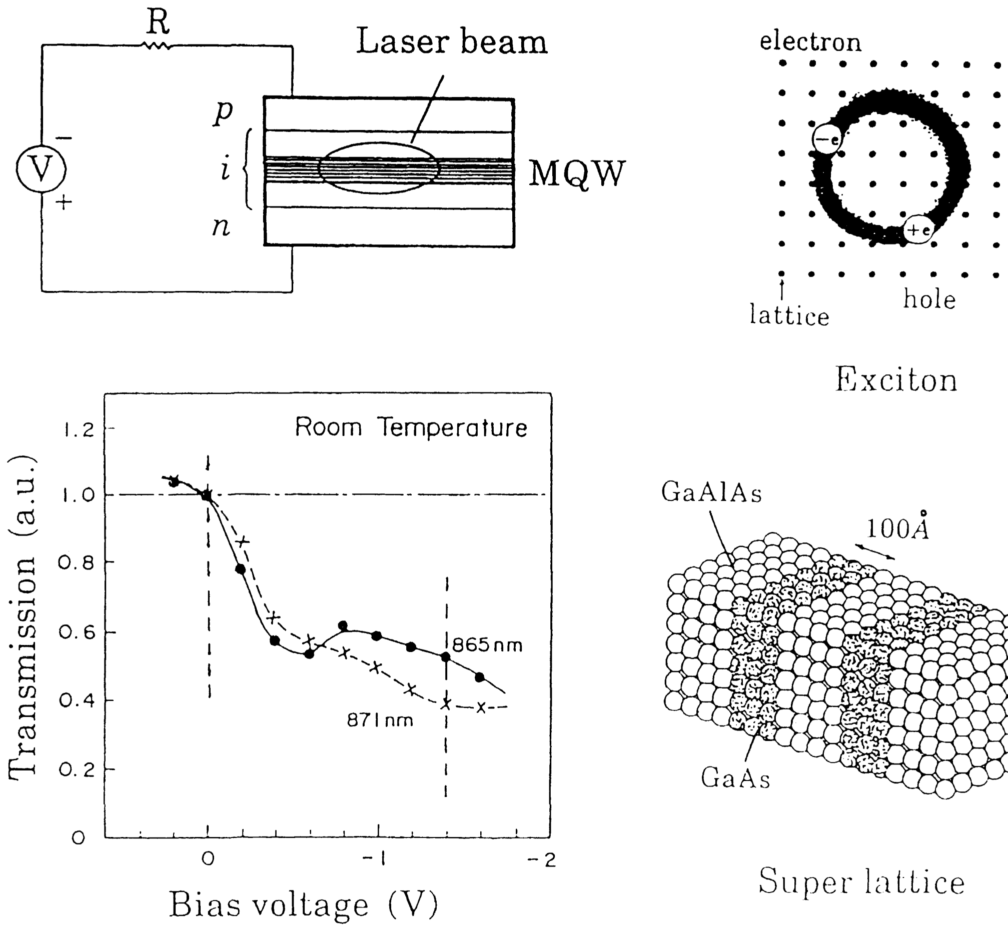

(a) (b)

FIG. 9. Configuration and characteristics of an electroabsorption modulator based on the QCSE: (a) MQW modulator system and basic transmission characteristics with wavelength as a parameter and (b) exciton (electron–hole pair) and heterostructure superlattice.

where n is the semiconductor refractive index and $\langle 1/\tau \rangle$ is the average value of the inverse of the relaxation time of the scattering mechanism.

3.2.4 Other Types of Modulators

There are many other types of modulators based on a variety of physical phenomena, such as the band-filling effect, the acousto-optic effect, the magneto-optic effect (see MAGNETO-OPTICAL DEVICES), and the thermal-optic effect (see THERMOELECTRIC DEVICES). Two of them are briefly described here.

As shown in Fig. 10, total-reflection-polarization intensity modulators have been fabricated making use of the band-filling effect. In this case, a large negative refractive-index change near the band-gap energy is induced by carrier injection. This refractive-index change is independent of the polarization direction. Typical characteristics of such a system are as follows: The switching injection is 90 mA, the switching speed is less than 1 GHz, the insertion loss is 20 dB, and the extinction ratio is about 10 dB at a 1.55-μm wavelength.

The basic principle of an acousto-optic modulator is illustrated by Fig. 11. An acoustic surface wave generated by an interdigital transducer propagates along a two-dimensional optical waveguide, inducing a distortion that acts as a grating; the pitch is equal

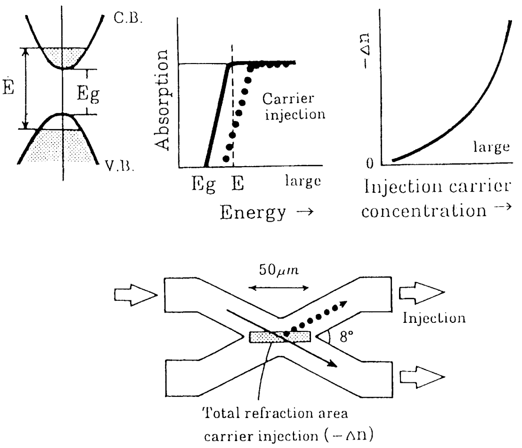
C.B.
E
Eg
V.B.
Absorption
Carrier injection
Eg E large
Energy →
−Δn
0
large
Injection carrier concentration →
50μm
Injection
8°
Total refraction area
carrier injection (−Δn)

FIG. 10. Basic properties of a total-refraction modulator.

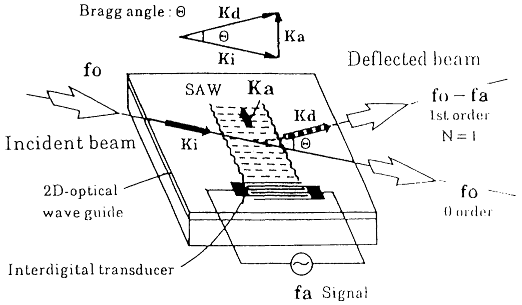
Bragg angle : Θ
Kd
Θ
Ka
Ki
fo
Deflected beam
SAW
Ka
Kd
fo − fa
1st order
N = 1
Incident beam
Ki
Θ
2D-optical wave guide
fo
0 order
Interdigital transducer
fa Signal

FIG. 11. Structure and operation of an acousto-optic modulator.

to an acoustic wavelength. An incident optical wave interacts at the grating and is reflected at certain angles that coincide with the Bragg angle between an acoustic wave vector and an optical wave vector. The spectrum of the outgoing light is changed from f_{opt} to $f_{\mathrm{opt}} \pm Nf_{\mathrm{acoust}}$, with N being an integer, so that frequency modulation can also be achieved.

3.3 Optical Detectors

When low levels of light are to be detected, it is desirable to use a photodetector with large sensitivity. For such applications, a large optical gain can be obtained with an avalanche photodiode (APD), as illustrated by Fig. 12. The APD system consists of an optical absorption region, an avalanche region, and a collector, with a reverse-bias voltage applied to the junction. The photons, incident through the window, are absorbed, generating electrons and holes. The photogenerated carriers are accelerated in the junction's depletion region, and if they acquire enough energy from the field, an ionizing collision with the binding electrons of the lattice atoms in the avalanche region can occur. Since each collision generates another electron–hole pair, the impact-ionization process continues, leading to carrier multiplication and gain. As a result, a large avalanche photocurrent is obtained. In other words, the APD has a current-magnification function corresponding to the reverse-bias voltage applied. Therefore, the APD exhibits very high sensitivity. However, its response can only be as fast as 10 GHz, since it is restricted by the carrier transit time across the avalanche region and the time constant containing the electric capacitance and the loading resistance (see DETECTORS, SEMICONDUCTOR).

The best performance of an APD is obtained when the ratio of the impact-ionization coefficients for electrons and holes is ei-

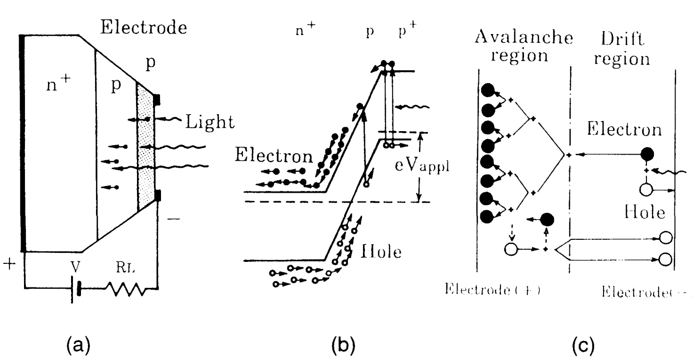
Electrode
n+
p
p
Light
+
V
RL
−
(a)
n+
p
p+
Electron
eVappl
Hole
(b)
Avalanche region
Drift region
Electron
Hole
Electrode (+)
Electrode (−)
(c)

FIG. 12. Structure and operation of an avalanche photodiode (APD): (a) configuration, (b) energy-band diagram, and (c) basic principle.

ther 0 or infinity. Unfortunately, in most compound semiconductors, this ratio is close to 1. In this case, both types of carriers continuously recycle through the depletion region, leading to long response and, consequently, small bandwidths. This is a fundamental limitation of the APD. In the last few years, however, clever ways to enhance or reduce this ratio in artificial structures have been explored (Bhattacharya, 1990). In fact, impact ionization in multilayered heterojunction structures was first proposed a decade earlier by Chin *et al.* (1980). Nowadays, superlattice or MQW APDs can be fabricated in such a way that only one type of carrier is confined in a deep quantum well (QW) and, therefore, its ionization rate is reduced (Kagawa *et al.,* 1993). The overall effect is the corresponding enhancement or reduction of the impact-ionization coefficient ratio. These tailored devices, in addition to low-noise performance and gain, can also provide tunability of the spectral response (see SPECTROMETERS, OPTICAL). In such structures, the absorption and avalanche region of the photodiode is made of multiple quantum wells of varying dimension and barrier sizes.

In the case of a *p-i-n* photodiode, the photons incident through the *p*-region window also generate electron–hole pairs. Each electron and hole flow in opposite directions, inducing a photocurrent. To obtain a high-speed response, a reverse-bias voltage is applied to the junction, but its value does not exceed the avalanche voltage. Thus, the photocurrent is very small because there is no current-magnification function. Although the *p-i-n* photodiode has no internal optical gain, by careful choice of material parameters and device design it can achieve very large bandwidths, with high-speed response of up to 60 GHz and very low noise.

4. NONLINEAR OPTICAL DEVICES

The response to small disturbances in most physical systems is linear. For instance, if the intensity of a standard laser beam is doubled, the intensities of the reflected and transmitted beams are also doubled. However, if the intensity is not increased two or three times but several orders of magnitude, it is possible to observe many interesting nonlinear effects, such as the generation of harmonics, two-photon processes, nonlinear mixing, or the Raman effect.

4.1 Second-Harmonic Generation

A most interesting physical phenomenon is second-harmonic generation (SHG). This kind of nonlinear response to the light waves occurs because the dielectric constant of materials is not really constant. In fact, any strong laser beam will be distorted when traveling through an optical material. This distortion is the result of the addition of one or more harmonics to the original laser beam. If the strength of the dipoles is defined by **P** and the applied optical field by **E**, then it may be shown that

$$\mathbf{P} = \chi \cdot \mathbf{E} + \chi_2 \cdot \mathbf{E}^2 + \chi_3 \cdot \mathbf{E}^3 + \cdots \text{etc.,} \tag{8}$$

where χ is the polarizability. From this equation, it follows that if **E** takes the form of $\mathbf{E} = \mathbf{E}_0 \sin(\omega t)$, then

$$\mathbf{P} = \chi_1 \cdot \mathbf{E}_0 \sin(\omega t) + \tfrac{1}{2}\chi_2 \cdot \mathbf{E}_0^2[1 - \cos(2\omega t)] + \cdots \text{etc.} \tag{9}$$

Equation (9) contains a term in 2ω, which corresponds to an optical wave having double the frequency of the incident wave. $\chi_2\mathbf{E}^2$ will not approach $\chi_1\mathbf{E}$ unless **E** has a value of at least 10^5 V/cm (equal to an optical power density of about 10^7 W/cm^2). Not all solids exhibit frequency doubling. Only if the material has no center of symmetry can frequency doubling be observed because χ_2 is not zero. The first researchers who observed frequency doubling focused the 6943-Å photon output from a ruby laser onto a quartz crystal and obtained blue emission at a wavelength of 3422 Å. To realize high-efficiency SHG devices, it is important that the phase (or velocity) matching between the original and the harmonic beam is satisfied. This can be achieved using appropriate optical materials, such as $LiNbO_3$, $KNbO_3$, and $KTiPO_4$ (KTP) crystals in resonant or in waveguide structure.

Figure 13 shows recently developed bulk SHG devices. For instance, a resonator composed of YAG and KTP crystals and a pair of mirrors can be excited by a LD at an 809-nm wavelength [Fig. 13(a)]. This configuration

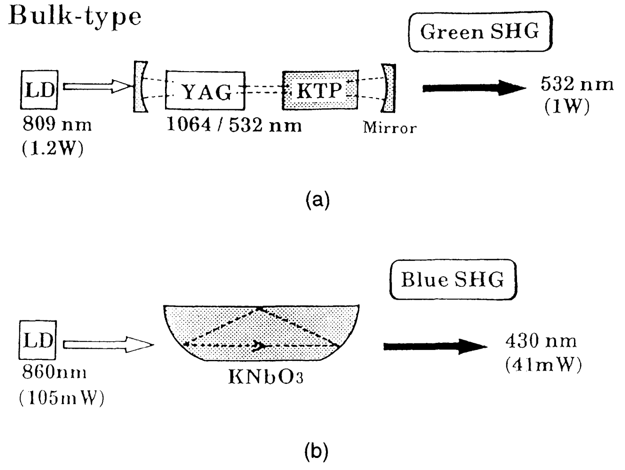

FIG. 13. Bulk-type SHG devices: (a) YAG-KTP and (b) $KNbO_3$.

allows the generation of a 532-nm green SHG beam. Likewise, a ring resonator fabricated from a $KNbO_3$ crystal, which has large nonlinear optic coefficients, allows the generation of a 428-nm blue SHG beam with 41-nW output power when an 856-nm beam from a LD with 105-mW output power is focused onto the crystal [Fig. 13(b)]. This device is highly efficient but demands both severe temperature control (within 0.3°C) and original laser wavelength control (within 0.06 nm).

The waveguide SHG devices are able to confine the original wave into waveguides so that the optical power increases about four orders of magnitude with respect to bulk SHG devices. The efficiency of waveguide SHG devices increases because it is proportional to the power density. Figure 14(a) shows a SHG device composed of a proton-exchanged $LiNbO_3$ waveguide. In this device, the phase-matching condition is automatically satisfied by use of Cherenkov radiation, involving the proton-exchanged waveguide, so that efficiency is about 1% and the direction of the output SHG beam is about 16° under the waveguide surface. In this device, a 410-nm SHG beam with 1 mW is obtained when the 820-nm original LD beam with 100 mW is focused onto the waveguide via an objective lens. The pattern of the output SHG beam has a crescent shape; therefore, a nonspherical lens is needed to convert it into a circular shape.

Recently, more efficient waveguide SHG devices have been developed by using a $LiNbO_3$ film, as shown in Fig. 14(b). To obtain the phase matching, the $LiNbO_3$ film was fabricated on a $LiTaO_3$ substrate, controlling the film thickness and orientation very precisely. The direction of the output SHG beam coincides with the original input beam. In this highly efficient device, 2.2 mW of SHG blue light is generated when only 40 mW of infrared light is focused onto the waveguides. Furthermore, to realize more efficient and practical compact devices, the quasi–phase-matching (QPM) technique can be used, which consists of the coherent domain inversion of crystals such as $LiTaO_3$, as shown in Fig. 14(c). One such recent blue-emitting SHG device is shown in Plate 1.

4.2 Raman Lasers

Another interesting nonlinear phenomenon is the Raman effect. When light of a single wavelength passes through a transparent substance, some of it is scattered. It has long been known that the scattered light contains not only the frequency of the original beam, but also other frequencies, with the frequency differences corresponding to the energy levels of the scattered molecules. These laser Raman lines are not only much more intense than the conventional Raman lines; they are also coherent, so that the laser Raman effect makes it possible to generate coherent light at new frequencies (see RAMAN SCATTERING; RAMAN SPECTROSCOPY INSTRUMENTATION). Recently developed semiconductor Raman lasers (SRL) built as a buried heterostructure with GaP Raman-active layer

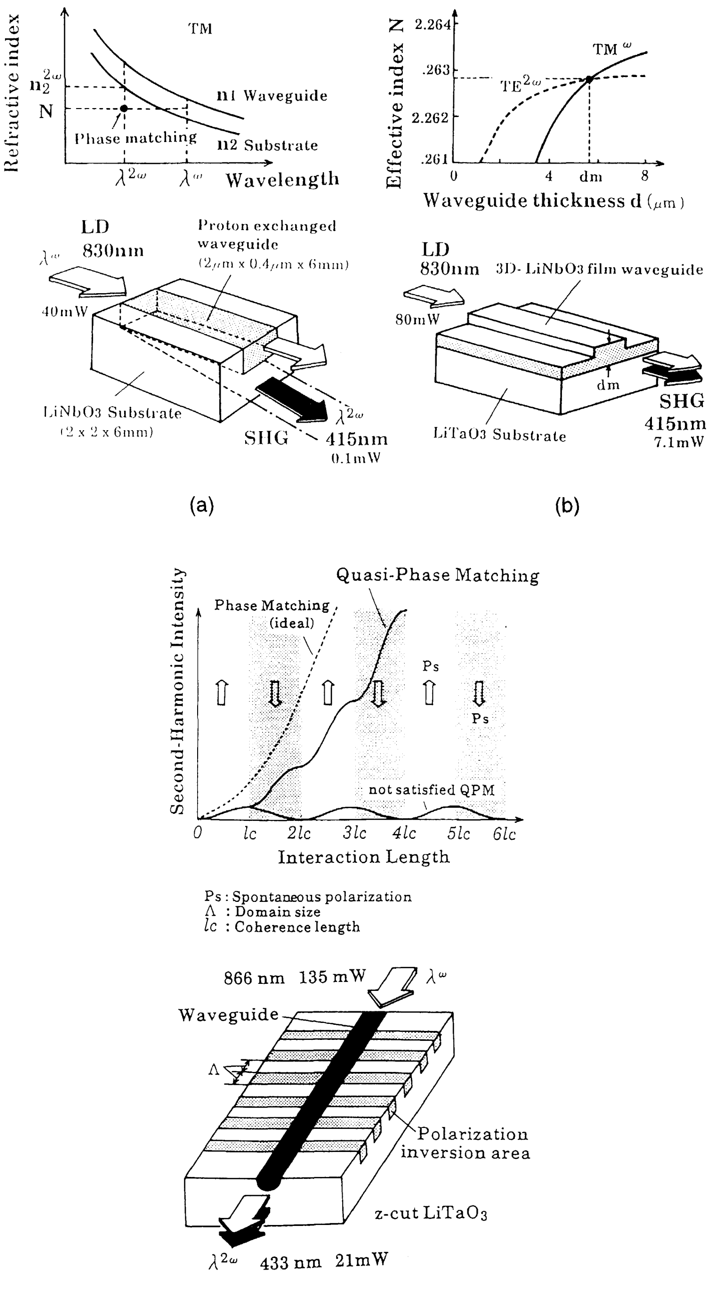
Refractive index
TM
n2 2ω
N
n1 Waveguide
Phase matching
n2 Substrate
λ2ω
λω
Wavelength
Effective index N
2.264
.263
2.262
.261
TM ω
TE2ω
0
4
dm
8
Waveguide thickness d (μm)
LD
λω 830nm
Proton exchanged waveguide
(2μm x 0.4μm x 6mm)
40mW
LiNbO3 Substrate
(2 x 2 x 6mm)
λ2ω
SHG 415nm
0.1mW
LD
830nm
3D-LiNbO3 film waveguide
80mW
dm
LiTaO3 Substrate
SHG
415nm
7.1mW
(a)
(b)
Second-Harmonic Intensity
Quasi-Phase Matching
Phase Matching (ideal)
Ps
Ps
not satisfied QPM
0
lc
2lc
3lc
4lc
5lc
6lc
Interaction Length
Ps : Spontaneous polarization
Λ : Domain size
lc : Coherence length
866 nm 135 mW
λω
Waveguide
Λ
Polarization inversion area
z-cut LiTaO3
λ2ω 433 nm 21mW
(c)

FIG. 14. Basic properties and configuration of waveguide-type SHG devices: (a) proton-exchanged $LiNbO_3$ device based on the Cherenkov radiation effect, (b) $LiNbO_3$ film SHG device based on the mode-matching method, and (c) $LiTaO_3$ quasi–phase-matching (QPM) second-harmonic generation (SHG) device.

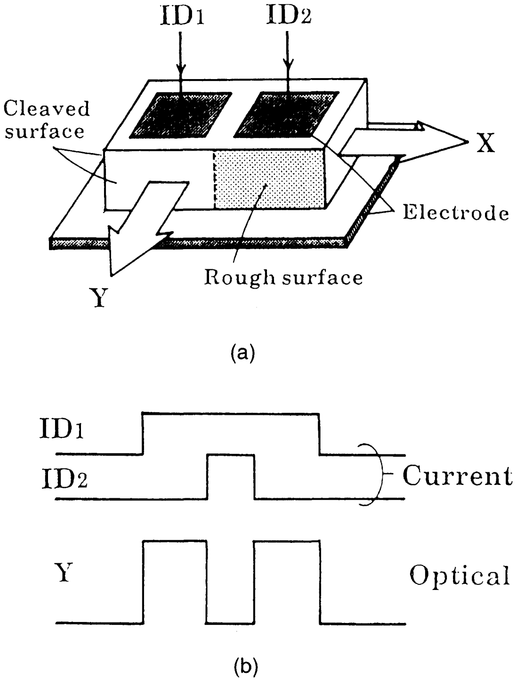

(a)

(b)

FIG. 15. Optical logic device based on a competing phenomenon between modes in a laser diode: (a) structure and (b) injection current and optical output.

and AlGaAs cladding layers could be of importance for heterodyne demodulation in wide-band optical communications, where light modulated with a frequency band exceeding 1 Thz should be detected.

5. OPTICAL LOGIC, OPTICAL BISTABLE, AND OPTICAL MEMORY DEVICES

Research in *optical computing* is already a few decades old. However, despite much progress, versatile optical computers that can replace or even compete with electronic computers remain elusive. Successful demonstrations of optical logic, utilizing the nonlinear effects in MQW structures, have been made. However, these attempts usually run into great difficulties due to the enormous amount of switching energy required and the sophisticated heat-dissipation schemes that subsequently become necessary for arrays of such single wafers. There is a real need for low-power photonic switching devices, which can be effectively used for computing and logic applications (see OPTICAL COMPUTING).

5.1 Optical Logic Devices

Optical logic devices based on the competing phenomenon between modes in the LD are shown in Fig. 15. Two diodes are fabricated in the same crystal. If an applied current is injected into surface D_1, the y direction starts lasing. However, if currents are injected simultaneously to D_1 and D_2, only the x direction is lasing because both D_2 planes are rough surfaces. That is, when the x direction is lasing, the y direction lasing is quenched as a consequence of mode competition. As a result, the optical output in the y direction, as shown in Fig. 15(b), corresponds to the injection currents ID_1 and ID_2. The response to the injection, which depends on the cavity length and the gain of the composed LD, was measured at 98 Ghz.

5.2 Optical Bistable Devices

Optical bistability allows two possible output states for the same input, as shown in Fig. 16. When the input intensity increases from zero, the output intensity gradually increases (A). The output rapidly increases (from A to B) when the input power is just over the threshold value A. It holds a constant value (C) when the input increases over

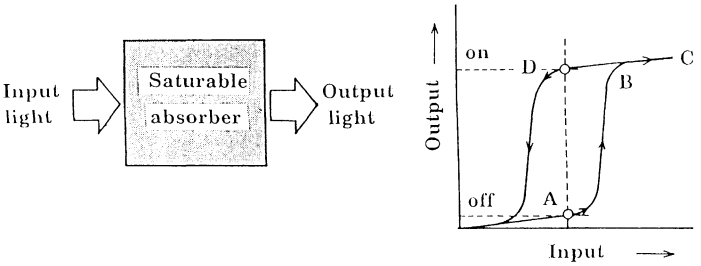

FIG. 16. Input–output characteristic of optical bistability.

the threshold. The output intensity slowly decreases with a decrease in the input power, and it rapidly decreases (from D to 0) when the input reaches another threshold value D, which is smaller than the value B. Thus, the input-versus-output characteristic has hysteresis, as shown in Fig. 16. Optical memory and logic operations can be obtained by utilizing this optical hysteresis curve. Such an operation principle is shown in Fig. 17. With the bias light intensity set at P_B, two stable states exist, according to the hysteresis curve: one "on" state, triggered by the "on" pulse, and one "off" state, triggered by the "off" pulse. If the hysteresis is free, a differential curve is obtained, which acts as a limiter for the output power corresponding to the various input values. Efficient bistability requires a strong intensity dependence to the absorption or to the index of refraction. It is desirable to obtain a large output difference and a small threshold value. There are two types of bistability; one is intrinsic, and the other is a hybrid type. The intrinsic device is fully optical; its intensity dependence arises from a direct interaction of light with matter. The hybrid device is partially optical, mixed with the electrical signal from a detector monitoring the transmitted intensity, usually applied to an intracavity phase shifter.

A reverse-bias *p-i-n* diode in a MQW structure shows a strong nonlinear effect based on the QCSE, previously described in Sec. 3.2.3. Indeed, at appropriate wavelengths, a strong negative–differential-resistance region is observed in the photocurrent-versus-bias voltage relationship of a *p-i*(MQW)-*n* structure. This arises because the photocurrent results from the change in absorption coefficient of the MQW due to the QCSE. It occurs where the heavy-hole and the light-hole peaks cross the photon energy of the input light. As long as the electron and hole densities in the quantum wells are less than 10^{11} cm^{-2}, the absorption process is essentially linear. This nonlinearity has been successfully exploited to tailor device response and demonstrate the operation of simple, optically controlled switching and logic devices. The most important of these devices is the self–electro-optic device (SEED), first demonstrated by Miller *et al.* (1985). The SEED exhibits photonic switching, bistability, and optically induced oscillations due to the negative differential resistance in the photocurrent. Recently, a symmetric self–electro-optic device (S-SEED) has been proposed, as shown in Fig. 18. Two SEED structures are electrically connected in series and inversely biased (-15 V). When input light of a constant value (2 mW) is fed into the SEED-1 and a weak control light pulse (of the order of 10 μW) is fed into the SEED-2, the output of SEED-1 shows dynamic bistability, as shown in Fig. 18(b). If the control light pulse is fed into the SEED-2, the SEED-1 shows the pulse output and has time-sequential power gain.

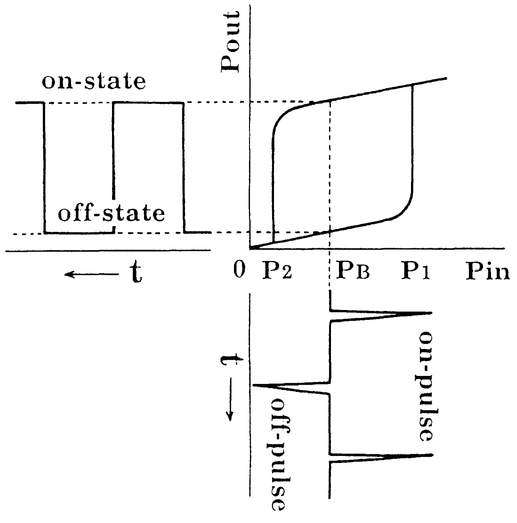

FIG. 17. Basic properties of optical memory and logic devices utilizing the optical hysteresis curve.

5.3 Optical Memory Devices

In general, optical memory can be achieved by utilizing the hysteresis curves discussed in Sec. 5.2 (see Figs. 16–18). In other words, all bistable devices can be used as optical memory devices.

Recently, a laser containing an absorption and current injection control region has been developed, as shown in Fig. 19. This device can perform as a saturable absorber. The injection-current–versus–optical characteristic has hysteresis. When the input light and the control injection-current values are small, there is no output light. However, when the input light increases beyond a threshold

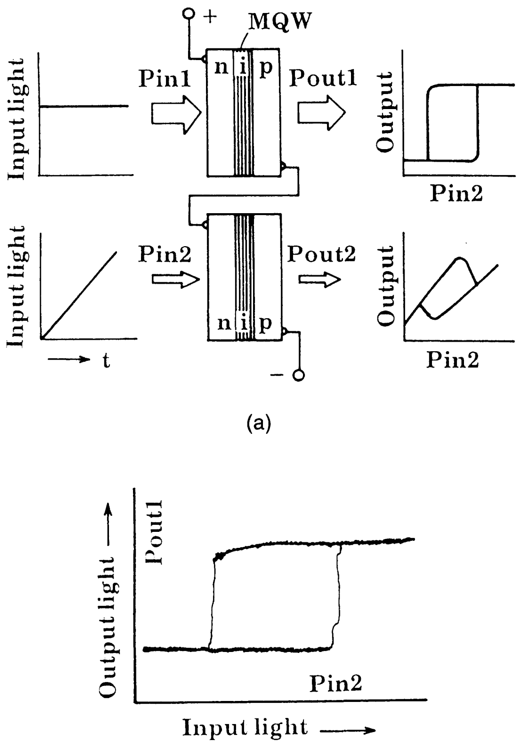

(a)

(b)

FIG. 18. (a) Structure and characteristics of a symmetric SEED (for MQW with 105-Å GaAs and 80-Å AlGaAs layers) and (b) oscilloscope trace of P_{out1} bistable characteristics.

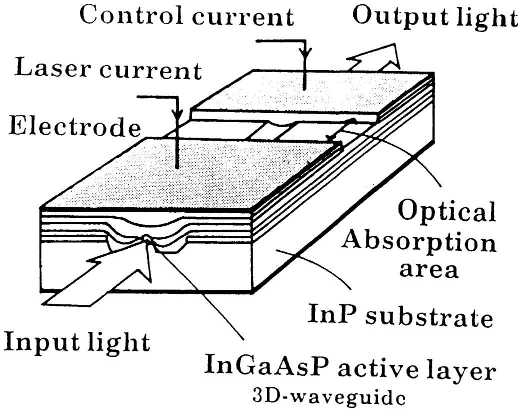

FIG. 19. Structure of an optical memory device based on a bistable LD.

value, there is a strong output light (about a hundredfold increase). Under this condition, even if the input light is eliminated, the output light continues ("on" state). That is, a memory operation can be realized. The "on" state can be switched to the "off" state by stopping the control injection current. This device acts as an optical logic and as a switching device and its response time is less than 1 ns.

Photonic devices hold the promise of massive parallelism and large connectivity. Furthermore, optical interconnections show a great potential for being programmable. Indeed, light beams could be directed from individual sources to individual detectors on the basis of a specific command, depending on the task, using two-dimensional switch arrays (see OPTICAL INTERCONNECTIONS). Thus, smart optoelectronics systems will bring an enormous degree of flexibility to future optical computing architectures.

6. PHOTONICS AND OPTOELECTRONIC INTEGRATED CIRCUITS

Following the invention of the transistor, in a few decades, electron devices dramatically evolved from electron tubes to ultra-large-scale integrated circuits, containing millions of monolithic microcomponents. Similarly, the rapid evolution of optical devices, ranging from discrete components to photonic integrated circuits, is clearly taking place. Indeed, the level of integration of optoelectronic components has increased steadily over the years, as a result of the continuous progress in semiconductor fabrication technologies. However, in contrast to the situation in electronic integrated circuits, where silicon technology completely dominates the field, a great number of different materials are used in the fabrication of optoelectronic integrated circuits (OEIC). While III-V compound semiconductors allow the integration of all active and passive components, $LiNbO_3$ is still the preferred substrate for many applications making use of the electro-optic effect. For purely passive devices, glass and silicon are commonly used. Compared with these inorganic materials, organic polymers are relative newcomers in the field. Nevertheless, the rapid advances in molecular engineering have lately allowed chemists to produce new polymer materials with tailored optoelectronic properties, optimized for applications such as waveguiding, electro-optic effects, or harmonic generation (Van Tomme, 1991).

There are two forms of optoelectronic integration, namely, hybrid and monolithic. In hybrid integration, discrete devices on separate functional blocks or chips are connected using electrical or fiber interconnections. These systems usually use a dielectric substrate, such as glass or $LiNbO_3$, which is externally coupled to light sources and detectors [see Fig. 20(a)]. The main advantage of such hybrid integration is the ability to use high-performance discrete devices available as components. The disadvantages are the lack of compactness and the enhanced parasitics in the interconnections. In monolithic integration, all optoelectronic components such as lasers, waveguide circuits, modulators, and detectors are fully integrated on a III-V compound semiconductor substrate, such as GaAs or InP [see Fig. 20(b)]. These systems, usually referred to as photonic integrated circuits (PIC), have the obvious advantages brought by size reduction, reduced parasitics, and the consequent gains in higher circuit speed and wider bandwidth. However, unlike a Si-based electronic IC in which almost all components are made with the same materials and the same processing steps, the heterostructures and the processing steps of a PIC can be significantly different. This makes the fabrication of high-performance monolithic OEICs or PICs a serious challenge.

As an alternative, some quasi-monolithic systems combine aspects of both types of integration. In these intermediate systems, for example, the light source is made of a compound semiconductor and the photodiodes are integrated on a Si substrate [see Fig. 20(c)].

Recent advances in group IV semiconductors (Si, Ge, C, and Sn) are also paving the way for a variety of new Si-based photonic and electronic devices, mostly SiGe–Si heterostructures. Device-quality SiGe–Si material has been produced by low-temperature CVD (Meyerson, 1992) and MBE techniques (Houghton *et al.,* 1991). Band-gap–engineered SiGe quantum-well and superlattice devices have opened a new realm of optoelectronic device structures that go well beyond conventional Si device systems. These structures have demonstrated new optoelectronic phenomena that are absent or weak in bulk Si.

In the future, the extension of these techniques will allow the growth of crystal alloys like SiGeC and SiGeCSn. A complete PIC can be monolithically fabricated from SiGe, SiGeC, and SiC materials, with every waveguide component connected seamlessly to its neighbors. Good progress has already been made in Si-based 1.3–1.6-μm waveguides and 1.3-μm guided photodetectors. However, efficient Si-based laser diodes, optical amplifiers, efficient LEDs, 1.55-μm photodetectors, 1.3–1.55-μm electro-optic modulators, and optical switches are the challenges that remain ahead.

During the last decade, a number of key developments have laid the foundations for the rapid advancement of photonics technology. Most important among them is the routine fabrication of DFB- and DBR-MQW laser structures, allowing high-*Q* on-chip resonators without cleaving. Another key development has been the progress achieved in high-quality computer-controlled vapor and beam crystal-growth systems for InP-based materials. The excellent large-area unifor-

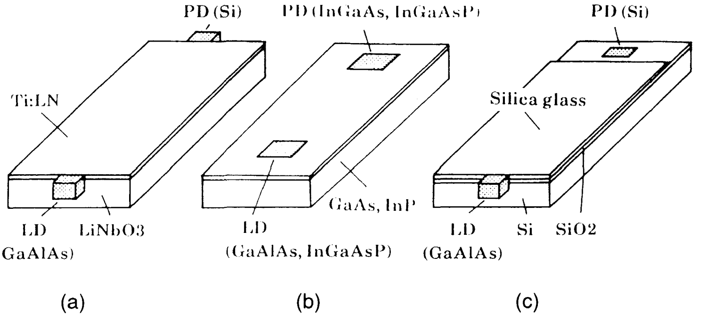

FIG. 20. Three examples of optoelectronic integrated circuits: (a) hybrid, (b) monolithic, and (c) quasi-monolithic.

mity and reproducibility achieved by these systems are crucial to any serious integration effort. In particular, the reproducible growth of highly complex vertical layer structures by use of a large number of ultrathin etch-stop layers has granted new levels of design freedom in the engineering of integrated waveguide devices. Application of these techniques, in some cases without extra growth steps, has resulted in versatile photonics technologies.

A monolithic OEIC or PIC should be constructed with single-mode waveguides in the passive structures. The structure size of its active components is of the order of one wavelength. Lasers, being nonlinear devices, are often fabricated without strictly fundamental-mode waveguides. In addition to the difficulties involved in making the buried heterostructure waveguides strictly single mode, the waveguides typical of most lasers can also be quite problematic for passive couplers or active switches because of their sensitive characteristics. An illustrative optoelectronics problem is the requirement for current blocking in lasers and current or field access to the active sections in the photonic integrated circuit, while retaining Fe-doped InP cladding for the low-loss waveguide interconnections. Nevertheless, many successful examples, such as optical transmitters and receivers constructed with lasers, photodiodes, modulators, switches, filters, integrated-optical rf spectrum analyzer, optical drive, and sensor devices, have been developed.

Figure 21 shows a quasi-monolithic OEIC consisting of a compact-disk reading head. It is constructed by integrating a focusing grating coupler (FGC), a twin-grating focusing beam splitter (TGS), and a photodiode array in a film waveguide on a Si substrate. The guided wave diverging from the laser is focused by the FGC into a point on the disk. The wave reflected by the disk is collected and coupled back into the waveguide. The TGS then divides the reflected wave into two halves, deflecting and simultaneously focusing them into two detecting points. Thus, the disk signal is read out. Whereas optical disk-drive heads are currently built with bulk micro-optics, implementation by fully integrated optics using a waveguide would lead to significant cost-effective improvements, such as better reproducibility, reduction of size and weight, and enhancement of application flexibility (see OPTICAL STORAGE).

Figure 22 shows a monolithic OEIC or PIC consisting of a transmitter of wavelength-divided multiplex. This device is constructed with three different lasers composed of DBR and injection-current control

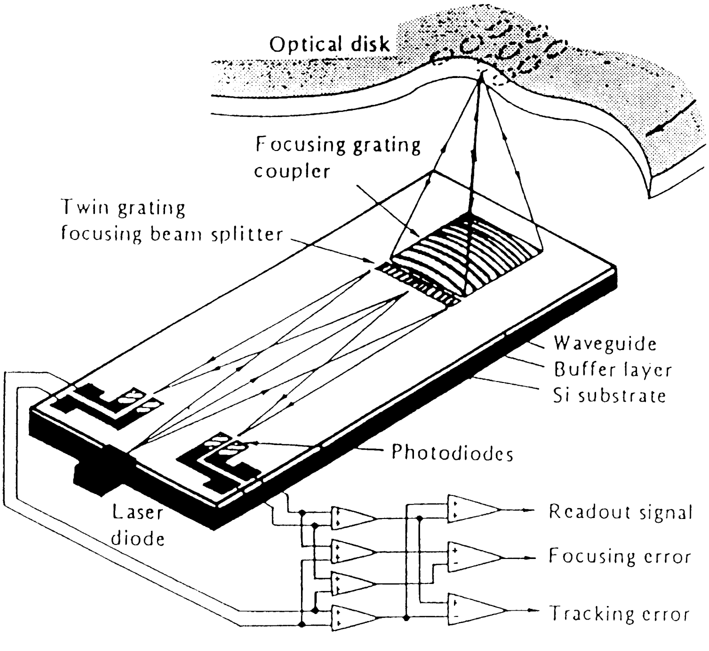

FIG. 21. A quasi-monolithic OEIC for a compact-disk reading head.

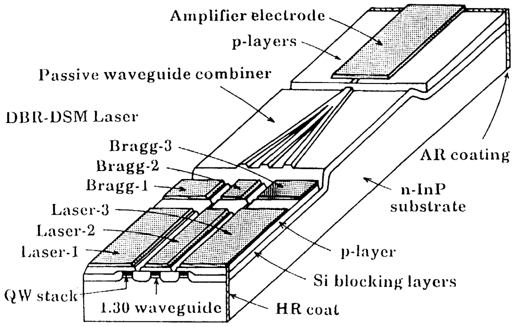
Amplifier electrode
p-layers
Passive waveguide combiner
DBR-DSM Laser
Bragg-3
Bragg-2
Bragg-1
Laser-3
Laser-2
Laser-1
QW stack
1.30 waveguide
AR coating
n-InP substrate
p-layer
Si blocking layers
HR coat

FIG. 22. A monolithic OEIC for an optical transmitter with three different-wavelength lasers.

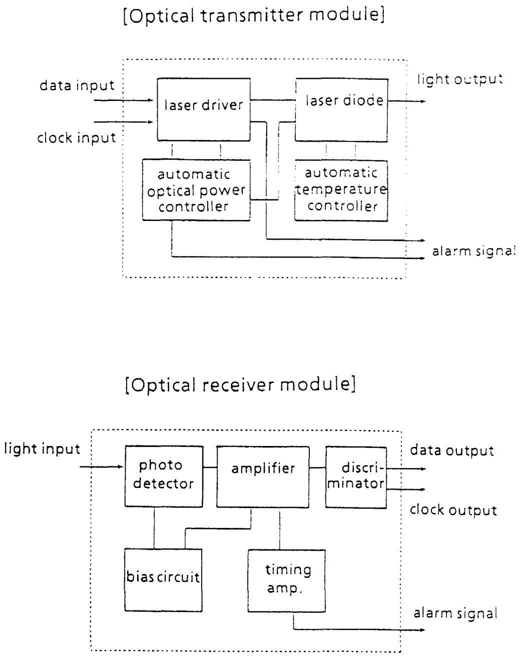
[Optical transmitter module]
data input
clock input
laser driver
laser diode
light output
automatic optical power controller
automatic temperature controller
alarm signal
[Optical receiver module]
light input
photo detector
amplifier
discri-minator
data output
clock output
bias circuit
timing amp.
alarm signal

FIG. 23. OEIC for high-speed optical transmitter/receiver modules.

regions, waveguides, and an optical amplifier. Three light waves with slightly different wavelengths around 1.55 μm emerge, and a stable high-frequency modulation (greater than 1 Gb/s) can be obtained for each laser. Such light sources, as well as similarly integrated photodetectors, will be of great importance in the development of future large-capacity optical transmission systems. Recent progress in compound-semiconductor–based electronic devices has allowed integration of complete electronic circuits with various optical devices on the same chip. Integrated receivers with highly sensitive characteristics, comparable to those of a discrete device circuit, are now being developed. Figure 23 shows the E/O (transmitter) and O/E (receiver) modules of such a system, recently built using an OEIC with high-frequency, wide-band preamplifier electronic circuits, operating at the high speed of 10 Gb/s.

Present-day optical communications systems are almost universally point-to-point in nature, where an optical fiber carries a unidirectional signal from the transmitter to the receiver. While such simple systems could benefit from some level of photonic integration, the true benefit of PIC systems is expected to become evident as optical-system architectures grow in sophistication and complexity. Indeed, their potential for high performance, compact size, high reliability, and low manufacturing cost, with much larger-scale integration and higher operating speeds, will transform them into crucial players of applications such as multichannel coherent or heterodyne (electrically demultiplexed) and incoherent (optically demultiplexed) wavelength-division multiplexed transmission, networking, and switching. In fact, this is even true in simple bidirectional links (Koch *et al.*, 1990) for broadband integrated-services digital networks (B-ISDN) and fiber-to-the-home (FTTH) architectures, which are currently being widely investigated in view of the exploding market demand for these services.

7. SUMMARY

The present article has attempted to review the development, as well as the present status, of the principal quantum optoelectronic devices, discussing their basic structure, their operation principles, and their main applications. We have emphasized the impact of new semiconductor-fabrication technologies, which have allowed us to build high-performance ultracompact optoelectronic devices with increasing levels of integration. It is expected that the powerful driving force of highly demanded market applications, most notably optical communications and computing, will accelerate the existing trends toward ever smarter and more affordable quantum optoelectronic systems.

GLOSSARY

Avalanche Photodiode (APD): Photodetector with high sensitivity due to large gains caused by impact or avalanche ionization of photogenerated carriers.

Cladding: Surrounding low-dielectric media covering the dielectric waveguide materials, which confines the lightwave propagation to the latter material.

Distributed-Feedback (DFB) or Distributed–Bragg-reflector (DBR) Laser: Semiconductor laser with high mode purity obtained by using a corrugated structure or grating that produces a periodic variation of the refractive index within the optical cavity.

Electroabsorption Effect: Absorption of photons with energies smaller than the band gap, with the application of an electric field.

Electroluminescence: The process of radiative recombination of electron–hole pairs, following carrier injection with a *p-n* junction or similar device.

Electro-optic Effect: Induced birefringence, a change in the refractive index of noncentrosymmetric crystals with electric field inversion.

Erbium-Doped Fiber Amplifier (EDFA): Optical fiber doped in its core with erbium atoms (less than 0.1%) and pumped (at 1.48 or 0.98 μm) with a LD coupled into the amplifier with a wavelength multiplexer. This infrared radiation is absorbed by the Er^{3+} atoms causing stimulated emission, which is used to amplify light signals near their lasing wavelength (1.55 μm).

Franz–Keldysh Effect: The change in semiconductor absorption that results in the absorption of photons with energies less than its band gap.

Heterodyning: Coherent phase-interference receiving technique, which allows the detection of very weak optical signals.

Laser Diode (LD): The semiconductor laser, a coherent light source that operates by stimulated emission due to population inversion.

Laser-Diode Amplifier (LDA): Two closely separated LDs, where one of them acts as an amplifier of the light output of the other.

Light-Emitting Diode (LED): A semiconductor light source that operates by spontaneous emission due to electroluminescence.

Multiple Quantum Well (MQW): Superlattice structure formed by multiple QW.

Optical Fiber: Silica-based waveguide, consisting of a high–refractive-index glass core, surrounded by a lower–refractive-index cladding; widely used in optical communications for light signal transmission.

Optical Modulator: Device for changing the intensity, phase, polarization, or frequency of a transmitting light in response to an electrical signal.

Optical Waveguide: Core region of high–refractive-index dielectric material through which light is propagated, surrounded by lower–refractive-index dielectric region or cladding.

Optoelectronic Integrated Circuits (OEIC): Systems that involve the integration of electronic and optical devices and optical interconnections.

Photoconductor: A material that exhibits an increase in electrical conductivity of a specific region with photoexcitation.

Photodetector: An optoelectronic device that absorbs optical energy and converts it to electrical energy, which gives rise to a photocurrent.

Photoluminescence: A process of radiative recombination of electron–hole pairs created by the injection of photons.

***p-i-n* Diode:** A junction photodiode in which an undoped intrinsic region is inserted between the *p* and *n* regions; it has a controlled depletion-layer width, which can be somewhat tailored to meet the requirements of photoresponse and bandwidth.

Quantum-Confined Stark Effect (QCSE): A shift of the semiconductor absorption spectrum to lower energies due to excitonic heavy-hole and light-hole resonances.

Quantum Well (QW): Synthetic structure composed of an ultrathin layer (comparable to the mean free path of carriers) of a narrow–band-gap semiconductor sandwiched between two wide–band-gap semiconductor layers, where carrier energy levels are quantized in the direction perpendicular to the layers.

Self–Electro-optic Effect Device (SEED): A structure composed of MQW *p-i-n* diodes, which exhibits photonic switching, bistability, and optically induced oscillations due to the negative resistance in the photocurrent due to the QCSE.

Stark Effect: A change in atomic energy due to electron energy-level splitting in the presence of an external electric field.

Works Cited

Alferov, Zn. I., Kazarinov, R. F. (1963), Inventors Certificate No. 181737.

Bhattacharya, P. (1990), *Int. J. Optoelectron.* **5,** 51.

Boyle, W. S., Thomas, D. G. (1960), U.S. Patent No. 3,059,117.

Chin, R., Holonyak, N., Jr., Stillman, G. E., Tang, J. Y., Hess, K. (1980), *Electronics Lett.* **16-12,** 467–469.

Desurvire, E. (1994), *Phys. Today,* January, 20–27.

Fowler, A. B. (1963), *Appl. Phys. Lett.* **3,** 1–3.

Hall, R. N., Genner, G. E., Kingsley, J. D., Soltys, T. J., Carlson, R. O. (1962), *Phys. Rev. Lett.* **9,** 366–368.

Holonyak, N., Jr., Bevacqua, S. F. (1962), *Appl. Phys. Lett.* **1,** 82–83.

Holonyak, N., Jr., Kolbas, R. M., Dupuis, R. D., Dapkus, P. D. (1980), *IEEE J. Quant. Electron.* **QE-16,** 170–186.

Houghton, D. C., Noel, J. P., Rowell, N. L. (1991), *Mater. Res. Soc. Symp. Proc.* **220,** 299–310.

Jewell, J. L., Lee, Y. H., Scherer, A., McCall, S. L., Olson, N. A., Harbison, P., Florez, L. T. (1990), *Opt. Eng.* **29,** 210–214.

Kagawa, T., Kawamura, Y., Iwamura, H. (1993), *IEEE J. Quant. Electron.* **QE-29,** 1387–1392.

Koch, T. L., Koren, U., Gnauck, A. H., Reichmann, K. C., Kogelnik, H., Raybon, G., Gnall, R. P., Oron, M., Young, M. G., de Miguel, J. L., Miller, B. I. (1990), *Technical Digest of the 12th IEEE International Semiconductor Laser Conference, Davos, Switzerland,* New York: IEEE, paper K-4.

Kogelnik, H., Shank, C. V. (1971), *Appl. Phys. Lett.* **18,** 152–154.

Kroemer, H. (1963), *Proc. IEEE* **51,** 1782–1783.

Kuchibhotla, R., *et al.* (1991), *IEEE Photonics Tech. Lett.* **3,** 354–356.

Lasher, G. L. (1963), *IBM J. Res. Dev.* **7,** 58–61.

Meyerson, B. S. (1992), *Proc. IEEE* **80,** 1592–1608.

Miller, D. A. B., Chemla, D. S., Damen, T. C., Wood, T. H., Burrus, C. A., Jr., Gossard, A. C., Wiegmann, W. (1985), *IEEE J. Quant. Electron.* **QE-21,** 1462–1476.

Mollenauer, L. F., Stollen, R. H., Gordon, J. P. (1980), *Phys. Rev. Lett.* **45,** 1095–1098.

Nakamura, M., Yariv, A., Yen, H. W., Somekh, S., Garvin, H. L. (1973), *Appl. Phys. Lett.* **22-10,** 515–516.

Nakazawa, M., Suzuki, K., Yamada, E., Kubota, H., Kimura, Y. (1993), *Electron. Lett.* **29,** 1474–1475.

Nathan, M. I., Dumke, W. P., Burns, G., Dill, F. J., Lasher, G. J. (1962), *Appl. Phys. Lett.* **1,** 62–64.

Nishizawa, J., Sasaki, I., Takahashi, K. (1965), *Appl. Phys. Lett.* **6,** 115–116.

Nishizawa, J., Otsuka, A. (1972), *Appl. Phys. Lett.* **21,** 48–50.

Nishizawa, J. (1973), *Optoelectronics* **5,** 309–321.

Nishizawa, J. (1977), *Optoelectronics,* Japan: Kyoritsu Press, chaps. 1–9.

Panish, M. B., Hayashi, I., Sumski, S. (1970), *Appl. Phys. Lett.* **16,** 326–327.

Pankove, J. I., Massoulie, M. J. (1962), *Bull. Am. Phys. Soc.* **II-7-1,** 88.

Quist, T. M., Rediker, R. H., Keyes, R. J., Krag, W. E., Lax, B., McWhorter, A. L., Zeigler, H. L. (1962), *Appl. Phys. Lett.* **1,** 91–92.

Thijs, P. J. A., Tiemeijer, L. F., Kuindersma, P. I., Binsma, J. J. M., Van Dongen, T. (1991), *IEEE J. Quant. Electron.* **27,** 1426–1439.

Tsang, W. T., Logan, R. A., Van der Ziel, J. P. (1979), *Appl. Phys. Lett.* **34,** 644–647.

Van Tomme, E. (1991), *IEEE J. Quant. Electron.* **27,** 778–787.

Watanabe, Y., Nishizawa, J. (1957), Japanese Patent No. 273,217.

Further Reading

Beesley, M. J. (1971), *Lasers and Their Applications,* London: Taylor & Francis, Ltd., chaps. 1–12.

Bhattacharya, P. (1994), *Semiconductor Optoelectronic Devices,* Englewood Cliffs, NJ: Prentice-Hall.

Brown, R. (1968), *Lasers: Tools of Modern Technology,* New York: Doubleday, chaps. 1–7.

Marcuse, D. (Ed.) (1973), *Integrated Optics,* Parts 1, 7, 8, New York: IEEE Press.

Suematsu, Y. (Ed.) (1982), *Optical Devices and Fibers,* Amsterdam: OHM, North-Holland, chaps. 1, 2.

Tamir, T. (Ed.) (1988), *Guided-Wave Optoelectronics,* New York: Springer-Verlag, chaps. 4, 6.

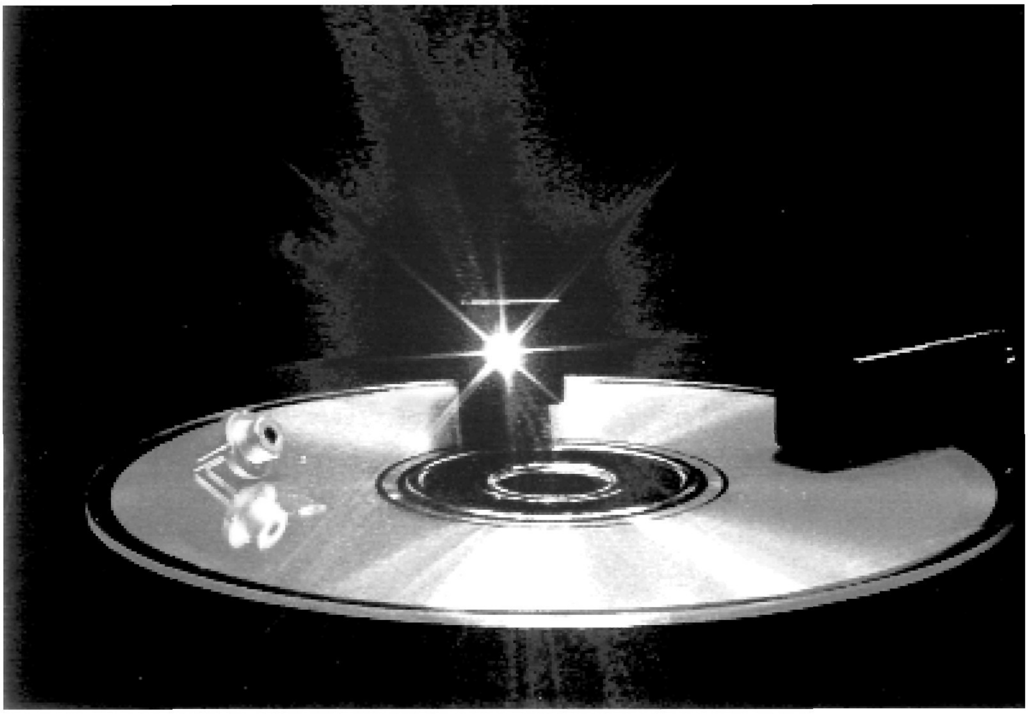

Vol. 15 – PLATE 1. Blue-emitting SHG device as used in the reading head of a CD drive

QUARTZ

Gary Johnson and Jonathan Foise, *Sawyer Research Products, Eastlake, Ohio, U.S.A.*

Introduction 365
1. **Description of Material** 365
1.1 Chemical and Crystallographic Structure 365
1.2 Physical Properties 366
1.2.1 Mechanical and Thermal 366
1.2.2 Chemical 367
1.2.3 Electrical 368
1.2.4 Optical 368
2. **Growth of Crystals** 368
2.1 Natural Occurrence and Quality 368
2.2 Cultured 368
2.2.1 History 368
2.2.2 Hydrothermal Growth 369
2.2.3 Post-Growth Processing 370
3. **Common Defects in Quartz** 370
3.1 Point Defects 370
3.1.1 Hydroxide Defects 371
3.2 Inclusions 371
3.3 Structural Defects 372
4. **Applications** 373
4.1 Electronic 373
4.1.1 Bulk Wave 373
4.1.2 Surface Acoustic-Wave Devices 374
4.2 Optical 375
4.3 Other Applications 375
Works Cited 375
Further Reading 375

INTRODUCTION

Quartz is the most common crystalline form of silicon dioxide. It is the most abundant substance in the earth's crust and exists in a variety of coarsely crystalline and microcrystalline forms. Three-quarters of the earth's crust is made up of the tectosilicates, quartz and the related minerals.

Quartz is a hard, somewhat brittle substance. It is thermodynamically and chemically very stable. It is an insulator and is piezoelectric. Its unique set of properties has made it the second most widely used electronic material. Quartz is clear, transparent from the infrared to the ultraviolet, birefringent, and optically active. This has led to many optical applications.

Most coarsely crystalline (single crystal) quartz used today is cultured, sometimes referred to as "synthetic." Hydrothermal techniques are used to produce quartz crystals. This method is one of high-temperature, aqueous solution growth. By varying such parameters as the temperature gradient, the crystallization temperature, and the solution composition, growth rate and other properties can be controlled in the finished crystal.

While an extremely pure substance, quartz can exhibit a variety of defects. These defects include point (atomic) defects, inclusions, and structural defects, such as dislocations and twins. These defects can influence the use of this material in electronic and optical applications.

1. DESCRIPTION OF MATERIAL

1.1 Chemical and Crystallographic Structure

Quartz is the principal naturally occurring form of crystalline silicon dioxide, SiO_2. The silicon is tetrahedrally bound to four oxygen atoms. Each oxygen atom is bonded to two Si atoms, and this linkage of the SiO_4 tetrahedra at their corners forms a three-dimensional network. The bonding is about 60% covalent and 40% ionic. One Si–O bond is 1.598 Å in length, the other is 1.616 Å. The bond energy is about 468 kJ mol^{-1}.

There are at least nine different ways that a SiO_2 framework can be constructed. These geometrical arrangements correspond to the nine known polymorphs of SiO_2. The naturally occurring polymorphs fall into three

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

structural categories: *low quartz,* with the lowest symmetry and most compact structure; *low tridymite,* with higher symmetry and a more open structure; and *low cristobalite,* with the highest symmetry and the most expanded structure (Klein and Hurlbut, 1977). The reconstructive transformations that convert one polymorph to another require bond breaking and thus are slow and require high energy. This allows the phases to exist metastably over a wide temperature range. Impurities or alkali-metal oxides can lower this activation energy. Tridymite is sometimes classified as a solid solution of quartz and alkali-metal oxide.

Quartz, tridymite, and cristobalite each have a low and high form [low (α) quartz and high (β) quartz are shown in Fig. 1]. These transformations are displacive, involving only minor atomic rearrangements. They occur quickly, reversibly, and at a well-defined temperature. Displacive transformations involve a much smaller symmetry change than the reconstructive transformations, with the low-temperature form showing the lower symmetry. Elevated pressures raise the transition temperatures.

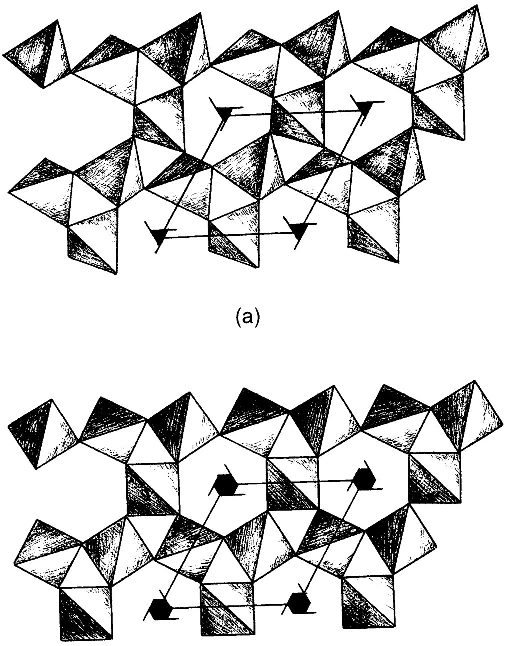
(a)

(b)

FIG. 1. The structure of the (a) low (α) quartz and (b) high (β) quartz. The unit cell, the threefold axes of low quartz, and the sixfold axes of high quartz are shown.

The common form of quartz is the α-quartz form. It exhibits trigonal symmetry (space group $P3_221$ or $P3_121$). By convention, the Z or c axis is parallel to the threefold axis and the X or a axis is parallel to one of the three twofold axes. In a set of rectangular axes, Y is at right angles to X and Z. The unit cell has 3 SiO_2 molecules. The most complete review of the lattice constants reports $a = 4.9127$ Å, $c = 5.4046$ Å for natural quartz and $a = 4.9134$ Å, $c = 5.4054$ Å for synthetic quartz (Brice, 1980). These values can vary with impurities.

At 573 °C and 1 atm pressure, α-quartz transforms to the β form, which is either $P6_222$ or $P6_422$. Upon cooling through the transition, Dauphiné twinning (see Sec. 3.3) usually occurs.

The SiO_4 tetrahedra in quartz are arranged helically on the threefold axes in both forms of quartz. This helical structure makes quartz enantiomorphic; in other words, crystals exhibit handedness, and thus are optically active. A left-handed helix results in a right-handed crystal. An observer looking along the Z axis at a plane-polarized light source through such a right-handed crystal sees the plane of polarization rotated clockwise.

Quartz occurs naturally in a variety of mineralogical forms. The naturally occurring varieties are divided into coarsely crystalline and cryptocrystalline or microcrystalline (fibrous and granular) varieties. Table 1 lists some of these varieties and their properties.

Quartz is one of the purest natural minerals. The most common impurities are Al, Fe, Ge, H, Li, Mg, and Na, with Ag, Ca, Cl, Cr, Cu, K, and Ti also reported. Total impurity concentrations are typically parts per thousand in milky grades, and in clear, colorless grades are often under 100 ppm.

1.2 Physical Properties

1.2.1 Mechanical and Thermal Quartz is a hard, somewhat brittle material. The density at room temperature is reported as 2.648 22 g cm^{-3}. The hardness on Mohr's scale is 7. The penetration hardness is 30.8

Table 1. The naturally occurring varieties of quartz, divided into coarsely crystalline and cryptocrystalline or microcrystalline (fibrous and granular) varieties, with some of their properties.

	Coarsely Crystalline
Rock crystal	Colorless quartz, commonly distinct crystals
Amethyst	Various shades of purple, color from Fe impurity and irradiation
Citrine	Various shades of yellow, Fe impurity
Rose quartz	Translucent to nearly transparent, pale pink to rose, Ti impurity
Smoky quartz	Brown to black, color from Al impurity and irradiation
Rutilated	Colorless quartz containing needles of rutile (TiO_2)
Tigereye	Yellow to brown with parallel fibers of an asbestos mineral, chatoyant
Milky quartz	Colorless, translucent, milkiness from minute fluid inclusions
	Cryptocrystalline
Chalcedony	General term for varieties composed of tiny fibrous crystals in a partly amorphous matrix
Agate	Layered chalcedony, often forming circular bands
Flint and chert	Granular variety, flint forming nodules, chert forming bedded deposits
Jasper	General term for a variety of translucent, impure chalcedony containing many inclusions

($\times$ 10^3 kg cm^{-2}) parallel to the Z axis and 22.9 perpendicular to Z. Quartz does not generally cleave along cystallographical directions. Fractures are conchoidal; that is, they form in random directions, like in glass. Quartz is much stronger in compression than in tension, and the strength varies with direction. The compression strength is 24 500 kg cm^{-2} parallel to Z and 22 400 perpendicular. Tensile strength is 1120 kg cm^{-2} parallel to Z and 850 perpendicular. The linear thermal expansion coefficient is very low; at 0 °C it is 7.10 $\times$ 10^{-6}/°C parallel to Z and 13.24 perpendicular to Z. It increases approximately linearly with temperature between -100 and 300 °C.

The third-order elastic coefficients are given in table 2, in units of GN m^{-2}. The temperature coefficients of the elastic constants range from -3300 $\times$ 10^6/°C (for c_{12}) to $+164$ $\times$ 10^6/°C (for c_{66}).

Table 2. The third-order elastic coefficients for quartz, in units of GN m^{-2}.

c_{111}	-210
c_{112}	-345
c_{113}	$+12$
c_{114}	-163
c_{123}	-294
c_{124}	-15
c_{133}	-312
c_{134}	$+2$
c_{144}	-134
c_{155}	-200
c_{222}	-332
c_{333}	-815
c_{344}	-110
c_{444}	276

Quartz is thermodynamically very stable. Its melting point is 1670 °C, with a boiling point of about 2230 °C. The vapor pressure at 1732 °C has been reported as 10 mm. The free energy of formation is 805 kJ mol^{-1} ($Si + O_2 \rightarrow SiO_2$). The heat capacity of 0 °C is 696.2 J/g. The molar specific heat has been described by the equation

$$C_p = 47.0118 + 3.4358 \times 10^{-2}\, T(\mathrm{K}) - 1.131 \times 10^6\, T^{-2}\, (\mathrm{J\ mole^{-1}\ K^{-1}}),$$

where T is the absolute temperature.

Quartz is a fair thermal insulator; at 20 °C the conductivity is 0.12 J/cm s °C parallel to the Z axis and 0.067 perpendicular to the Z.

The Curie temperature, that is the temperature of the α–β transition, is 573.3 °C (on heating), with a very small hysteresis and a latent heat of transformation of 10.5 J/g.

1.2.2 Chemical The chemical stability of quartz is also very high. Quartz is completely insoluble in water or organic solvents at room temperature and only 0.1–0.3 wt% in supercritical water. Quartz is attacked by hydrofluoric acid and solutions of fluoride salts, particularly at elevated temperature. The solubility is also greatly increased in supercritical water by the addition of mineralizing agents, such as alkali-metal salts, particularly the hydroxides and carbonates.

1.2.3 Electrical Quartz is the most widely used piezoelectric material. Broad use in the electronics industry is based on the unique combination of features quartz possesses. Among these are piezoelectricity, the existence of device orientations (cuts) with a variety of properties such as zero or high temperature coefficients, low mechanical losses (high *Q*), and ease of processing. Quartz's piezoelectricity arises from the crystallographic structure, which lacks a center of symmetry. The effect couples electrical current and voltage to elastic displacement and stress in the crystal. This coupling is described by coupling constants, which are dependent upon direction and temperature. Quartz has eight independent piezoelectric constants. The existence of the twofold symmetry axis in the *X* direction means that there is no piezoelectric effect along any axis perpendicular to *X*. The piezoelectric constants at 25 °C for right-handed quartz are given in Table 3. These constants are dependent upon temperature and pressure. The signs would be reversed for the left-handed form.

The dielectric constants for α-quartz are $\epsilon_{11}^T = \epsilon_{22}^T = 39.97$, $\epsilon_{33}^T = 41.03$, $\epsilon_{11}^S - \epsilon_{11}^T = -0.76$, and $\epsilon_{33}^S - \epsilon_{11}^T = 0$, all in units of 10^{-12} F/m. In this notation, the first number is the direction along which the voltage is applied, the second is the direction of measurement, with 1 the *x* axis, 2 the *y*, and 3 the *z*. The dielectric constants are dependent upon both temperature and field strength.

Quartz is an electrical insulator with resistivities ranging from 10^{14} Ω-cm at room temperature to 10^5 Ω-cm at 300 °C.

1.2.4 Optical Quartz's clarity, chemical and mechanical durability, and optical activity make it a widely used optical material. Clear α-quartz shows no absorption from the UV cutoff at around 160 nm to around 3 μm in the infrared. The index of refraction varies from about 1.48 at 3500 nm to 1.76 at 160 nm. It shows a slight dependence upon temperature, decreasing about 7 × 10^{-6}/°C. Quartz is birefringent, with the birefringence varying from 8.98 at 0.71 μm to 21.38 at 0.15 μm. The birefringence also shows a small decrease with increasing temperature. The rotatory power of quartz goes from 201.9 °/mm at 2265 Å to 11.589 at 7948 Å. The rotation is clockwise in right-handed quartz. The rotation power increases 1.4 × 10^{-4}/°C. The *Z* axis is the optically active axis.

Table 3. The piezoelectric constants at 25 °C for right-handed quartz.

Strain	Stress
$d_{11} = -2.31 \times 10^{-12}\ \mathrm{C\ N^{-1}}$	$e_{11} = -0.171\ \mathrm{C\ m^{-2}}$
$d_{14} = -0.727 \times 10^{-12}\ \mathrm{C\ N^{-1}}$	$e_{14} = -0.0406\ \mathrm{C\ m^{-2}}$
$g_{11} = -0.0578\ \mathrm{m^2\ C^{-1}}$	$h_{11} = -4.36\ \mathrm{N\ C^{-1}}$
$g_{14} = -0.0182\ \mathrm{m^2\ C^{-1}}$	$h_{14} = 1.04 \times 10^9\ \mathrm{N\ C^{-1}}$

2. GROWTH OF CRYSTALS

2.1 Natural Occurrence and Quality

Quartz is an abundant mineral, found particularly in igneous and metamorphic formations. Because it can be transported in water at high temperatures, quartz is also found as a secondary mineral in sedimentary rocks. Clear, colorless, single crystals (rock crystal) are widely distributed, with notable deposits in Brazil (generally recognized as the best for large, single crystals), the Alps, the Ural Mountains, and in the United States in Arkansas and New York. Important amethyst deposits have been found in Brazil, Zambia, and the Urals.

2.2 Cultured

2.2.1 History The first reported growth of quartz crystals by Schafhäult in 1845 and by de Sénarmont in 1851 produced microscopic crystals. G. Spezia, from 1898 to 1908, published reports on the growth of macroscopic crystals. He used solutions of sodium silicate, natural crystals as seeds and supply, and a silver-lined vessel. By heating the supply end of his vessel to 320–350 °C, and the other end to 165–180 °C, he obtained about 15 mm of new growth over a 200-day period. Unlike modern practice, the hotter part of vessel was at the top.

In the period before and during World War II, Richard Nacken led an extensive German effort to grow quartz crystals, especially once the supply of natural crystals from Brazil was blocked. They used an isothermal growth process that relied on the higher solubility of the vitreous silica supply to supersaturate the solution.

Following the war, researchers working with the U.S. Army Signal Corps studied the German work and developed it into a commercially viable process. C. B. Sawyer, D. R. Hale, Hans Jaffe, and others at Brush Development, and then Sawyer Research Products, Inc., developed the low-pressure sodium carbonate process (next section). R. Laudise, N. C. Lias, and others at Bell Laboratories and Western Electric developed the high-pressure sodium hydroxide process (next section). Today, these processes are used worldwide.

2.2.2 Hydrothermal Growth With a melting point of 1670 °C and high chemical resistance, quartz crystals would be very difficult to form by either melt or flux techniques. Since cooling through the Curie point at 573 °C usually produces twins, any technique for production of single crystals must be done below the Curie temperature. The process used is a solution-growth technique called hydrothermal growth. This technique is very similar to the processes by which natural quartz crystals are formed.

Hydrothermal growth is conducted in steel pressure vessels called autoclaves (Fig. 2). A supply of nutrient material, usually natural clear or milky quartz, is placed in the bottom half of the autoclave. In the top half, seed crystals, cut from either cultured (synthetic) or natural crystals are hung. The baffle, a metal plate with one or more holes cut in it to provide a constant and controlled opening (usually expressed as a percentage of the total cross section), is placed between the upper and lower halves. An aqueous solution, usually an alkali-metal hydroxide or carbonate solution, is added to the autoclave so that the solution occupies a set percentage of the total free volume of the vessel. The alkali-metal compounds acts as a mineralizer, increasing the solubility of the quartz in the solution, through the formation of soluble silicate complexes.

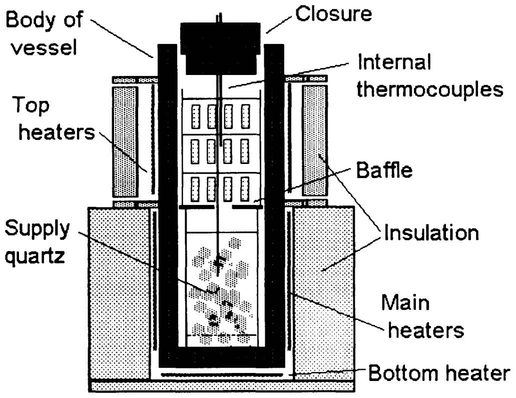

FIG. 2. Schematic drawing of an autoclave showing the various components (not to scale).

After sealing of the vessel, the autoclave is heated to operating conditions with the use of external resistance heaters. Typical operating conditions are generally between 300 and 400 °C. The internal pressure developed is controlled by the temperature and the amount of solution added to the autoclave. The critical point for pure water is well established at 374 °C. The critical points for solutions used in quartz growth have not been established but are probably altered by the addition of the mineralizer.

At operating conditions, the temperature is controlled such that a temperature difference is established between the internal temperature in the lower half of the autoclave, called the dissolving chamber, and the upper half, called the growing chamber. A temperature gradient in a closed cylinder of fluid in a gravitational field establishes a natural convection pattern called a closed thermosyphon. In the warmer dissolving chamber, quartz dissolves in the solution. The warm, less dense fluid rises. In the growing chamber, the cooler temperature leads to supersaturation of the solution for quartz and the quartz precipitates on the seed crystals. The cool, dense fluid sinks into the dissolving chamber and the cycle continues.

The temperature difference between the growing and the dissolving chambers is the chief process control of the growth rate. The temperature difference controls not only the supersaturation in the growth chamber, but together with the baffle, controls the rate of fluid and heat transfer between the two chambers. Growing temperature also has a strong effect, with growth rate increasing with increasing temperature. Other factors that influence the growth rate are pressure, mineralizer type and concentration, other chemical additives (particularly lithium), and seed orientation.

The mineralizer is a compound added to the solution to increase the solubility of the compound of interest, in this case, quartz. Typically, alkali-metal salts, particularly sodium hydroxide or sodium carbonate, are used. In natural systems, sodium chloride is believed to be the most common mineralizer. The mineralizer increases the solubility by complexing with the solute to form more soluble species. While the exact form of the dissolved species is unknown, it is believed that soluble silicates of forms such as $Si_3O_7{}^{2-}$ and $SiO_3{}^{2-}$ are formed in the presence of hydroxides and carbonates, respectively. Lithium is added at low concentration (typically 0.01 to 0.1 moles/L) to improve the relationship between proton uptake and growth rate (see Sec. 3.1.1). The mechanism of this improvement is not completely understood. Under these conditions, complex alkali-metal–iron silicates form on the steel surfaces of the autoclave. These silicates act as a passivation layer, preventing further corrosion. They are also the compounds that form inclusions (see Sec. 3.2).

Two similar processes are the most widely used in the cultured quartz industry. They are known as the low-pressure and high-pressure processes. The low-pressure process typically uses a (0.6 to 0.8)*M* sodium carbonate solution, a growing-chamber temperature of 345 °C, a temperature difference of 10 °C, and a pressure of 700–1000 bars. Typical growth rates are 0.4 mm/day in the *Z* direction. The chief advantage is the lower capital expense of autoclaves that operate at this pressure.

The high-pressure process typically uses 1.0*M* sodium hydroxide solution, a growing temperature of 360 °C, a temperature difference of 25 °C, and a pressure of 1000–1500 bars. While the capital costs for such equipment are greater, growth rates of up to 1.0 mm/day (*Z*) can be achieved at these conditions.

The anisotropic structure of quartz results in large differences in growth rate in different crystallographic directions. Growth rate in the *Z* direction can be up to three times the growth rate in the *X* direction. The orientation of the seeds used can affect not only the growth rate of the crystals but the uptake of impurities.

2.2.3 Post-Growth Processing Although most quartz is used without further treatment, sweeping is sometimes used for post-growth processing. Sweeping is a high-temperature, solid-state electrolysis of the quartz crystal. It is used to remove or replace interstitial ions in the quartz lattice.

Sweeping is conducted for one of two reasons. The first is to increase the radiation hardness of quartz. When quartz is irradiated, alkali-metal ions from defects such as $Na_{(\text{interstitial})}$–$Al_{(\text{substitutional})}$ are displaced from these defects, with the formation of Al–hole or Al–OH centers. This can cause immediate and/or permanent changes in the crystal's electrical properties. Sweeping substitutes the alkali-metal ions for protons (when swept in air) or holes (when swept in vacuum). This minimizes permanent changes in the crystal when irradiated.

The second use of sweeping is for the reduction of etch channels (see below). Though the mechanism is not completely understood, it is believed that etch channels are formed when impurity-decorated dislocations dissolve faster than the surrounding crystal, in an appropriate etchant. Sweeping reduces the formation of etch channels either by removing the impurity ions from the dislocations or by depositing an insoluble species into the dislocation.

The sweeping process involves applying an electrode material to the *Z* surface of the quartz bar. Typical electrode materials are plated metals such as gold, metal foils, or carbon films. An electric field on the order of 1 kV/cm is applied. The bar is heated to just below the Curie temperature, so as to increase the conductivity of the crystal without creating twins. Current densities on the order of μA/cm² are generated. Sweeping generally takes from days to weeks to complete.

3. COMMON DEFECTS IN QUARTZ

3.1 Point Defects

Point defects are atomic-size defects. They usually involve the replacement of one atom in the structure by a foreign atom (substitutional), or the insertion of an extra atom into the structure (interstitial). They may also involve other atomic or subatomic species such as holes and electrons (see POINT AND

EXTENDED DEFECTS IN SOLIDS). Given the openness of the quartz structure and the stability of the Si–O bond, interstitial defects are much more common in quartz. By far the most common interstitial atom is hydrogen, in the form of OH^- groups. These defects have significant effects upon the electronic properties (see below). Li, Na, K, Al, Fe, and holes are commonly found interstitially; Al, Fe, Ge, and Ti are found substitutionally, and a variety of other cations have been reported. Often a substitutional defect, such as Al^{3+} or Fe^{3+} substituted for a Si^{4+}, is paired with an interstitial species, such as H^+ or Li^+, for charge compensation. Anionic defects are much less commonly reported in quartz, with Cl particularly found in natural samples and F occasionally found in synthetic.

Quartz exhibits an anisotropic uptake of impurities; in other words, different regions of quartz have relatively higher or lower concentrations. The highest concentrations are found in the $+X$ region, followed by the $-X$ region, and then the Z regions (see Fig. 3). Concentrations in the $+X$ region can be from 3 to 10 times higher than in the Z.

The structure of the impurities sites also seems to be anisotropic in quartz. While synthetic amethyst and citrine are both formed from Fe^{3+}-containing solutions, amethyst is formed in the rhombohedral growth (particularly from seeds cut parallel to the rhombohedral faces), and citrine is formed in the Z growth (particularly from seeds parallel to the Z face).

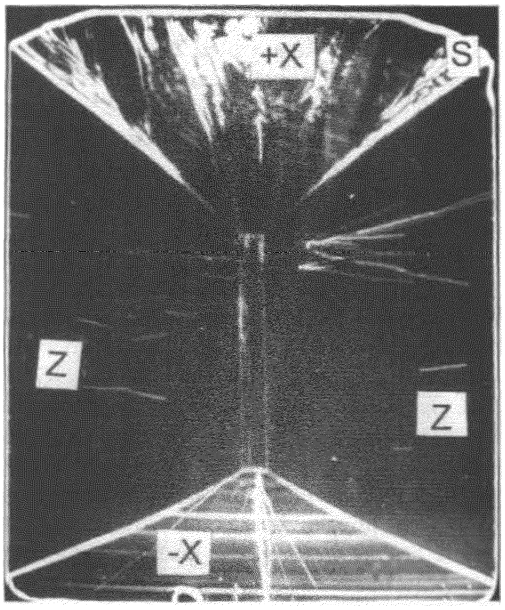

FIG. 3. X-ray topograph of a cultured quartz slice showing the growth zones.

3.1.1 Hydroxide Defects As mentioned above, protons are the most common defect in quartz. The infrared absorption at various frequencies between 3400 and 3600 cm^{-1} gives a reliable measure of the hydrogen concentration. Dodd and Fraser (1965) and Sawyer (1972) established a correlation between the extinction coefficient α and the electrical quality factor Q. This extinction coefficient is determined by the formula

$$\alpha = (1/t)\log[T_{\text{base}}/T]$$

where t is the sample thickness in centimeters, T_{base} is the percentage transmittance at a baseline wave number (nominally 3800 cm^{-1}), and T is the percentage transmittance at the wave number of interest (usually 3410, 3500, or 3585 cm^{-1}). An infrared quality factor can be determined from this.

Hydrogen concentration has also been found to be proportional to growth rate. By measuring α across the major growth direction of the crystal, the variation in the growth rate over the course of the growth cycle can be determined.

3.2 Inclusions

Inclusions are trapped solid particles or fluid bubbles, with a diameter typically between 25 and 150 μm. They are found in both natural and cultured quartz.

Fluid bubbles are fairly common in natural quartz, particularly in the milky grades. They are less common in cultured quartz, occurring mostly at the seed veil, the interface between the seed and the bulk crystal. Since bubbles are tiny samples of the solution from which the quartz formed, they have been analyzed to learn more about this solution.

Solid inclusions are common in both natural and cultured quartz. In natural quartz, the inclusions may be trapped by the growing crystal, may be formed simultaneously with the crystal, or may form after the growth of the host. Rutile is probably the most common solid inclusion in natural quartz. A wide variety of other minerals have been found, including various asbestos minerals, mica, tourmaline, hornblende, hematite, and gold.

In cultured quartz, solid inclusions are generally one of a variety of alkali-metal–iron silicates. Given the growth conditions of quartz in steel vessels, this is not surprising. Acmite, $NaFeSi_2O_6$, has traditionally been the archetypical inclusion species. However, the iron silicates {$Na_5FeSi_4O_{12}$, $Li_6Na_6Fe_2Si_{18}O_{45}$, emeleusite ($Li_2Na_4Fe_2Si_{12}O_{30}$), and riebeckite [$Na_2(Fe^{2+},Fe^{3+})_5Si_8O_{22}(OH)_2$]}, as well as lithium silicate, magnetite, and hematite, have all been identified. It appears that these inclusions are typically formed elsewhere in the autoclave, are circulated by the fluid, and are trapped by the growing crystal. Large solid inclusions can cause other defects, such as bubbles and dislocations, by disrupting the crystal structure.

Inclusion densities are estimated by immersing the crystal in an index-matching fluid and illuminating the crystal with a high-intensity light. The inclusions will appear as tiny illuminated points of light.

A more accurate method is to examine the bar microscopically. Sites in the bar are chosen and observed at moderate magnification (usually ×30). By scanning along the direction of observation, the number of inclusions in a set volume can be counted. Using a reticule with a scale, the counts can be made for a given size range of inclusions. Typically, inclusion densities are reported for inclusions 25–75, 75–100, and >100 μm in size.

3.3 Structural Defects

As in any crystalline material, various structural defects are found. Dislocations are found in single crystals of natural and cultured quartz, though they are generally more common in the latter. Screw, edge, and mixed dislocations (see POINT AND EXTENDED DEFECTS IN SOLIDS) have all been found.

Dislocations have three sources in cultured quartz. The first is continuations of dislocations in the seed. The second source is defects on the seed veil, either on the surface of the seed itself (usually damage from the cutting of the seeds) or deposited inclusions. The third source is inclusions in the bulk of the crystal.

When a quartz wafer is etched—for example, in a fluoride solution—etch pits form on the surface of the wafer and etch channels form into the bulk. Both are caused by dislocations. The increased strain energy around the dislocation causes an increase in etch rate around them. While every dislocation forms an etch pit, only about $\frac{1}{3}$ of the dislocations form etch channels. It is believed that impurity decoration of dislocations that form etch channels accounts for the difference.

A related phenomenon is that of filaments. These are line defects, generally parallel to the major growth direction, that are visible to the naked eye under intense illumination. They appear to be related to dislocations and may be dislocations decorated with fluid or other impurities.

As in other crystalline material, twins can form. There are basically two types of twins in quartz, electrical twins and optical twins. Electrical twins (also called Dauphiné twins) are regions that exhibit an X-axis orientation opposite to that of the bulk crystal. They are usually irregularly shaped. They can be caused by heating above the Curie temperature, by high pressure, or by a very intense electric field.

Optical twins (also called Brazil twins) are regions with the opposite handedness to that the bulk crystal. The boundaries of these twins are usually straight. Both types of twins can be revealed by the etching of the crystal, and then observation of the surface with an incident light source.

Cultured quartz crystals exhibit distinct growth zones or sectors around the seed. The zones, $-X$, $+X$, and Z, are named for the seed faces off which they form. In some crystals, a boundary region, called the S region, forms between the $+X$ and the Z regions (Fig. 3). The growth zones can be clearly seen in a slice taken perpendicular to the Y axis (X-Z plane) by either optical or x-ray topography, or by gamma-irradiating the sample and examining the optical darkening. The boundary between the growth zones consists of a small lattice distortion, on the order of 0.01%, and probably a tilting of the lattice on the order of 10^{-6} radians.

4. APPLICATIONS

4.1 Electronic

The Curies were the first to describe the piezoelectric effect, and determined the existence of it in quartz. During World War I, quartz plates were first used in early sonar. The first reported use of quartz as an oscillator was in the 1920s. Throughout the 1920s and 1930s, quartz-crystal units were used for radio broadcasting. The mass production of such units was brought about by the communication needs of World War II. Radio broadcasting, other commercial communications, and military applications were the only significant uses of quartz-crystal components until color television in the early 1960s.

Thereafter came the citizens' band radio, watches, and clocks, followed by digital electronics. Now, quartz is firmly established as the material of choice in frequency control. Only silicon for semiconductors is manufactured in larger quantity than single-crystal quartz.

All electronic applications of quartz involve the piezoelectric effect (see PIEZOELECTRIC DEVICES). Through this effect, electrical energy is converted into mechanical energy, and vice versa (the converse piezoelectric effect). The use of the piezoelectric effect occurs in three very different ways: exciting bulk waves, launching surface waves, or establishing static charges. Frequency-control components result from utilizing resonant elements to trap bulk or surface waves. Frequency control refers to oscillators (frequency sources) and filters (frequency selectors) (see PIEZOELECTRIC FILTERS AND RESONATORS). Resonant elements used in oscillators can be considered filter elements, selecting the frequency that will be amplified by the active part of the circuit. In an even broader sense, all resonant devices can be considered as frequency-sensitive phase-shift networks.

In principle, quartz can be used in a static mode, in which pressure is related to charge. There has been little interest in developing commercial applications based on this property.

The most common applications use the very high stability of quartz components over wide operating conditions to create components with nearly invariant characteristics over large temperature ranges, long time intervals, and so on. However, it is also possible to accentuate the effect of a particular environmental variable, temperature for example, on the resonant element and use it as a high-quality sensor. Such elements have the advantage of a digitized output (frequency), which makes it unnecessary to carry out further analog-to-digital conversion. Frequency measurement is a well-developed technology capable of higher precision than is achieved with measurement of other parameters.

The first quartz resonators used bulk-wave excitation. Resonant quartz devices are passive; they respond to stimuli but are not used as energy sources. The normal scheme is to introduce the resonant element into a broadband stimulus. The response of the element is controlled by the mechanical resonance of the device, with the mechanical excitation resulting from the conversion of the electrical stimulus to mechanical motion. Thin-film, conductive electrodes are deposited on the mechanical resonators to provide coupling between the mechanical and the electrical energy. These electrodes become part of the resonant system and must be considered in resonator design. Electrically, the presence of electrodes results in capacitance which must also be considered. Mechanically, the electrodes result in mass loading and acoustic mismatch between regions of the element with electrodes and those without. This mismatch can result in energy trapping, which is the basis of resonance.

4.1.1 Bulk Wave Bulk-wave devices are designed so that the bulk of the material is in motion during the resonant cycle. Many modes are usually excitable. For example, a cymbal is a mechanical resonator in which many modes are excited when it is excited mechanically. A tuning fork, on the other hand, responds in one mode (440 Hz, for example) very strongly, while minimizing others to the greatest extent possible. Quartz devices are designed to maximize certain vibration modes and to minimize others. The variables in the design primarily relate to device shapes and electrode characteristics. Some quartz resonators are simply rods supported at each end, analogous to a string or xylophone plate; others are shaped like tuning forks. The scale is typically a few to several millimeters.

The early bulk-wave devices were plates with strong vibrations in length-extension or face-shear modes. Later the thickness-shear modes became dominant (Fig. 4). The main reason for the increased use of thickness-shear cuts was higher stability. Experimenters had noticed that with plates of one crystallographic orientation resonant frequency increased with temperature, while with others it decreased. The conclusion was that a zero–temperature-coefficient orientation existed. This orientation was determined, and the thickness-shear mode was dominant. The resonant frequency of such a device is determined almost exclusively by the thickness of the plate.

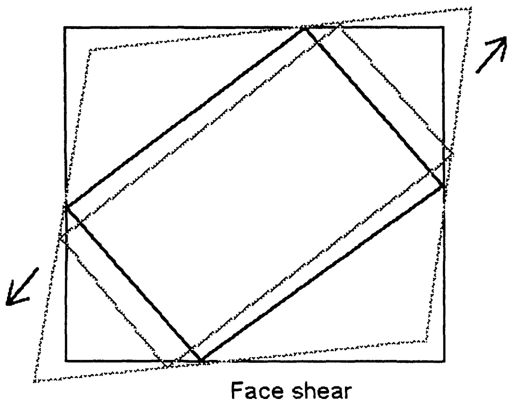

Face shear

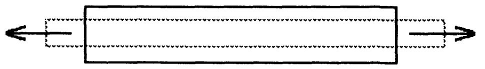

Length-extension shear

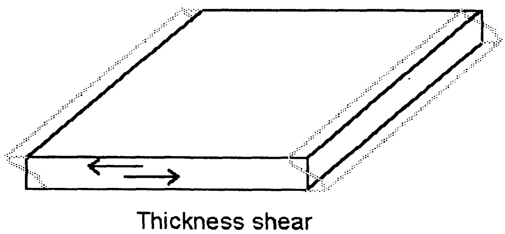

Thickness shear

FIG. 4. The various of modes of vibration for bulk-wave quartz devices.

Well-known glass fabrication methods were easily adapted to the processing of bulk-wave components, supplemented by thin-film deposition technology. Cutting, lapping, etching, and polishing are the normal procedures used for forming devices, followed by thin-film deposition. The main differences between fabricating glass and crystal quartz shapes are the final thickness control and maximum temperature of quartz. The final thickness tolerance is often the equivalent of a monolayer (such tolerances require chemical milling as the final step). The local temperature at the contact between abrasive and crystal cannot exceed 573 °C, or destructive twinning will result.

The common range of frequencies of bulk-wave devices is 30 kHz to 150 MHz based on mechanical processing. Chemical processing methods recently have been developed to increase the frequency to a few hundred MHz (at 250 MHz, the thickness of a thickness-shear device operating in the fundamental mode is about 7 μm).

The quartz crystal in wrist watches is produced in quantities of over one billion per year. It is a quartz resonant element in the shape of a tuning fork, operating at 32.768 kHz (2^{15} Hz), surrounded by an inert atmosphere in a cylindrical metal package no larger than 3 × 8 mm. The personal computer is another example of our many contacts with quartz components. The speed of the computer relates to the frequency of the internal quartz oscillator, often 10 to 66 MHz.

4.1.2 Surface Acoustic-Wave Devices

Bulk-wave devices face inherent limits to the frequency of operation related to the strict rules that relate shape and frequency. In the thickness-shear devices, for example, an upper frequency limit exists because of limits on how thin the device can be made in a controllable way. As communication technology has migrated up in frequency, related to spectrum limits and semiconductor advances, the need for GHz-range frequency-control devices has begun.

In the case of surface acoustic-wave (SAW) devices, the frequency of the launched wave is controlled by the spacing of interlaced, thin-film conductive fingers.

These fingers are alternately of positive and negative polarity, giving rise to a compression wave with $\frac{1}{4}$ of its wavelength equal to the spacing between adjacent fingers of opposite polarity. Thus, the frequency characteristic of a SAW device is controlled by the pattern of the electrodes, not the shape of the device, nor the substrate dimensions. The resulting wave is referred to as a surface wave because it is normally confined to a volume very near the surface. This is much the same as the waves on a large body of water.

The electrode pattern is established using photolithography techniques well known to the semiconductor industry. Submicron line width and spacing are required for GHz-range devices.

Another method adopted from the semiconductor industry is the concept of processing many devices on one wafer. Bulk-wave devices were typically handled one by one through manufacturing. Using the wafer philosophies and equipment of the semiconductor processors has opened many new possibilities for miniaturization, higher-frequency operation, and cost reduction.

4.2 Optical

Single-crystal quartz is also used in the optical industry, though in smaller quantities than in the electronic industry. Two properties, birefringence and optical activity, are used. Birefringence is the simultaneous existence of two indices of refraction. Optical activity relates to the rotation of the polarization vector as light passes along the optic axis. This property is used in laser optics to achieve $\frac{1}{4}$ and $\frac{1}{2}$-wave rotations in the optical path.

Birefringent elements have found their way into some consumer applications. Most home video cameras use an optical quartz device to control certain interference patterns associated with charge-coupled–device imaging elements. The compact-disc player uses a Wollaston prism in the laser path to detect the phase difference between the reference and the beam reflected from the metallized CD surface.

4.3 Other Applications

Natural quartz has for centuries served man as a gem material. Amethyst is iron-containing quartz, citrine is also iron-containing (though of different valence then amethyst) quartz, and smoky is aluminum-bearing quartz. In addition to clear, the other three important quartz-related gems have been cultured to a very high quality and uniformity. As the best grades of natural stones become less available and as the labor cost of processing small, random stones increases, the interest in cultured varieties has increased.

In addition to the cultured gems mentioned above, other colors not naturally occurring in nature have been created. Blue is the best example. Green quartz, rarely occurring naturally, has been cultured. There is very limited commercial interest in these materials.

Works Cited

Brice, J. C. (1980), *J. Mater. Sci.* **15,** 161–167.

Dodd, D. M., Fraser, D. B. (1965), *J. Phys. Chem. Solids* **26,** 673–686.

Klein, C., Hurlbut, C. S., Jr. (1977), *Manual of Mineralogy,* New York: John Wiley & Sons, p. 437.

Sawyer, B. (1972), *IEEE Trans. Sonics, Ultrason.* **SU-19,** 41–44.

Further Reading

Bottom, V. E. (1974), *The Theory and Design of Quartz Crystal Units,* Abilene, TX: Bottom.

Brice, J. C. (1980), *Rev. Mod. Phys.* **57,** 105–146.

Heising, R. A. (1946), *Quartz Crystals for Electrical Circuits: Their Design and Manufacture,* New York: D. Van Nostrand Co.

Laudise, R. A. (1987), *Chem. Eng. News* **65** (39), 30–43.

Naussau, K. (1980), *Gems Made by Man,* Radnor, PA: Chilton Book Co.

O'Donoghue, M. (1987), *Quartz,* London: Butterworths.

QUASICRYSTALS

A. R. Kortan, *AT&T Bell Laboratories, Murray Hill, New Jersey, U.S.A.*

Introduction: Quasicrystals—Impossible Crystals 377

1. **Quasicrystalline Structure** ... 379

1.1 Periodic and Quasiperiodic Structures 379

1.2 Generating Quasiperiodic Structures 379

1.2.1 Space Filling with Tile Matching 380

1.2.2 Strip-Projection Method 382

1.2.3 Generalized Dual Method 382

2. **Quasicrystal-Forming Compounds** 383

2.1 Early Metastable Compounds . 383

2.2 Stable Compounds 384

2.2.1 Al_5Li_3Cu Quasicrystal 384

2.2.2 $Al_{65}Cu_{20}(Fe, Ru, Os)_{15}$ Quasicrystals 386

2.2.3 Decagonal Quasicrystals 387

2.2.4 $Al_{70}Pd_{20}Mn_{10}$ Quasicrystal 388

2.3 Predicting New Quasicrystals; Quantum Structure Diagrams 388

3. **Quasicrystallography** 390

3.1 Icosahedral Quasicrystals 390

3.2 Decagonal Quasicrystals 391

4. **Quasicrystal Structure and Growth Models** 391

4.1 Early Models 391

4.2 Growth of Perfect Quasicrystals 393

5. **Defects in Quasicrystals** 394

5.1 Phonons and Phasons 394

5.2 Dislocations 395

6. **Properties** 395

6.1 Electronic 395

6.2 Magnetic 396

7. **Summary** 396

Glossary 396

Works Cited 397

Further Reading 397

INTRODUCTION: QUASICRYSTALS—IMPOSSIBLE CRYSTALS

Until recently, we classified solids as either crystalline, where the basic constituents, atoms or molecules, are ordered in a periodic array, or amorphous, with a disordered arrangement of constituents, as in glasses. The structure of crystals is simply described by a repetition of a single motif, called a unit cell, consisting of one or more atoms. Diffraction patterns of crystals, which are related to structure by a Fourier transform, reflect the crystal symmetries and consist of sharp diffraction peaks, indicating the presence of a long-range order. For the crystalline state, powerful theories have been developed that make use of the symmetry imposed by the periodicity. According to these theories, pentagonal (fivefold) symmetry is impossible in solids.

Discovered in 1982 by Dan Shechtman (1984), icosahedral quasicrystals exhibited an extraordinary fivefold diffraction pattern with sharp diffraction peaks, indicating the presence of a long-range order similar to that found in crystals. Additionally, peaks along a radial, straight line, were not spaced equally distanced as in crystals but were related to each other through an irrational number $\tau = 2 \cos 36°$, suggesting a new kind of ordering. Figure 1 illustrates three different state-of-the-art quasicrystalline compounds, which are quasiperiodic in one, two, and three dimensions and periodic in any remaining dimensions. Diffraction patterns are also shown along the tenfold, fivefold, and twofold directions.

The following illustration will help explain the controversy over order and fivefold symmetry. Imagine covering a planar surface with square-shaped tiles, as shown in Fig. 2(a). In this simple process, tiles are placed periodically next to each other until the plane is covered. Next, imagine covering the plane using pentagonal tiles, as shown in

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

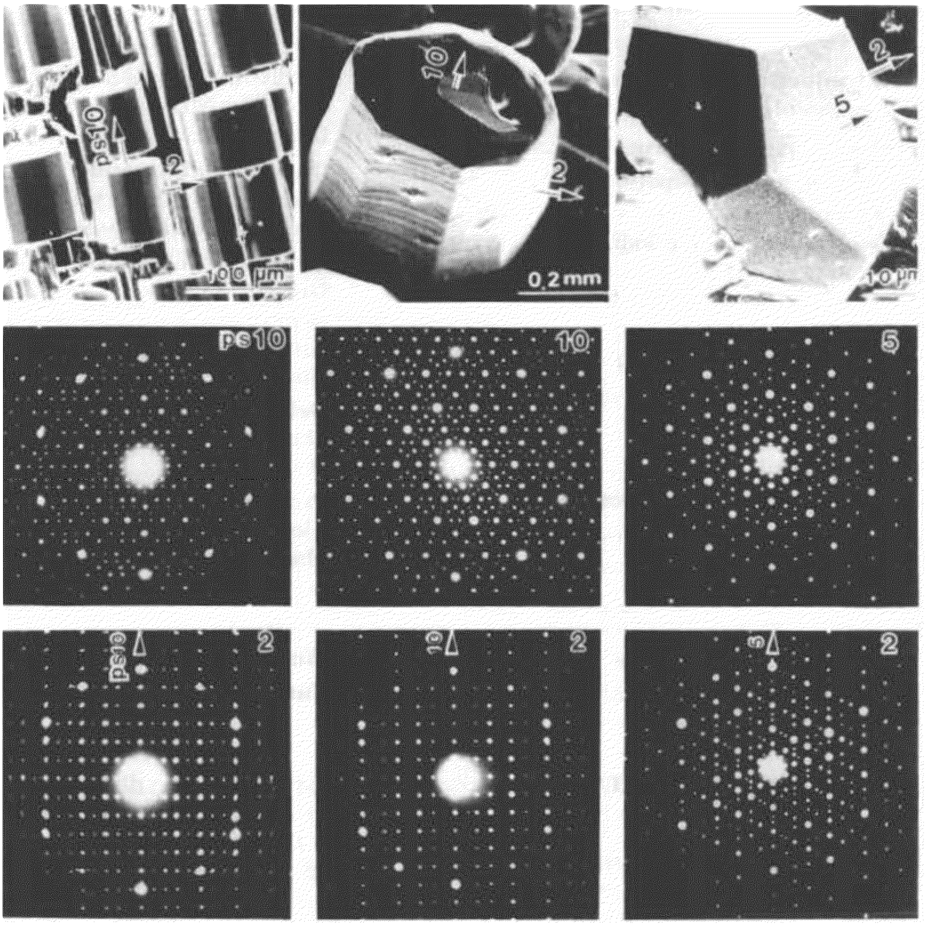

FIG. 1. Single grains of real one-, two-, and three-dimensional quasicrystalline compounds and their transmission electron diffraction patterns along fivefold, tenfold, and twofold directions. (Courtesy of A. P. Tsai.)

Fig. 2(b). These tiles leave gaps in between when the edges are matched, and the plane cannot be covered completely. Most importantly, the overall structure of the tiling appears disordered, unlike the square-tiled surface. The tiles used in these examples are direct analogs of unit cells used in describing crystalline systems, and qualitatively illustrate why fivefold order is incompatible with the high degree of periodic order found in crystals. Quasicrystals, because of their extraordinary structures, are sometimes referred to as "impossible crystals."

After a decade of intense research on quasicrystals, remarkable progress has been made in understanding this exotic form of matter. The purpose of this article is to present an introduction to quasicrystals.

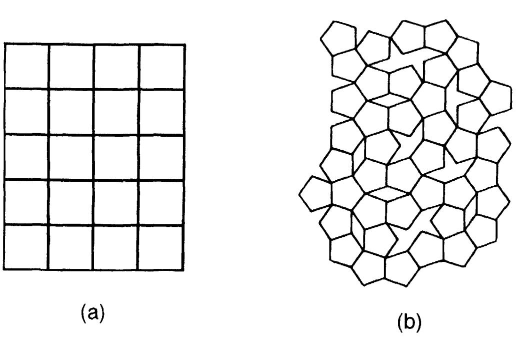
(a) (b)

FIG. 2. Tiling a planar surface demonstrates that pentagonal tiles appear disordered compared to square tiles.

1. QUASICRYSTALLINE STRUCTURE

In this section, quasiperiodic systems in one, two, and three dimensions are studied, and methods to generate them are described.

1.1 Periodic and Quasiperiodic Structures

The crystalline state of matter consists of periodically spaced atoms or molecules, whose positions are defined by the vector

$$\mathbf{R} = n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c}.$$

The primitive translation vectors **a**, **b**, and **c** describe the shortest separation between identical units, and the n's are integers. Letting the n's take all possible values, we describe a crystalline lattice. Any crystal, therefore, appears identical when viewed from points spaced by the translation vector **R** and exhibits so-called long-range periodic translational symmetry. Crystals or periodic systems also exhibit a point-group symmetry, which includes rotations and reflections of the lattice by keeping one point fixed. For crystals, rotations of only 360°/n, $n = 1, 2, 3, 4, 6$, are allowed. As demonstrated in Fig. 2, it is impossible to construct a periodic system that exhibits a fivefold point-group symmetry. (For more details, readers are referred to the article CRYSTALLINE STATE.) There is also an orientational long-range order that is common to all crystals and quasicrystals. This restricts the orientations of the unit cells to a special discrete set of directions and is demonstrated by the edges of the square and pentagonal tiles of Fig. 2, which are aligned along the same direction with respect to their neighbors, despite the disordered appearance of the pentagonal tiles.

1.2 Generating Quasiperiodic Structures

In generating quasiperiodic structures, a quasiperiodic but long-ranged translational order can be constructed. Similar to the single unit-cell concept used in crystals, for quasicrystals there exist two or more unit cells, or building blocks. These unit cells are decorated with atoms or molecules. Quasicrystalline structures are then generated by some aperiodic arrangement of these unit cells or building blocks. A one-dimensional construction that belongs to the famous mathematician Fibonacci is as follows. Imagine having two different unit length building blocks, of lengths L (for long) and S (for short). The lengths L and S are related through an irrational number, the so-called golden mean, $\tau = 2 \cos 36° = 1.618..$, by $L/S = \tau$. The Fibonacci sequence is then generated by applying the following iterative substitution. At each stage of iteration, substitute L in place of S and LS in place of L. Successive iterations generated from a starting block L then become

1. *L*
2. *LS*
3. *LSL*
4. *LSLLS*
5. *LSLLSLSL*
6. *LSLLSLSLLSLLS*
7. *LSLLSLSLLSLLSLSLLSLSL*

⋮

n . . .

This sequence has many interesting properties. First, it has a perfect aperiodic order; that is, no part of the sequence repeats itself periodically. In the limit as n approaches infinity, the ratio of the total number of L to S used in the process becomes equal to τ. It is possible to find identical local patterns in different parts of the sequence, independent of the size of these patterns. Another remarkable property is the self-similarity of the sequence upon the substitution, meaning each iteration duplicates the starting finite sequence in addition to new extra terms. This point is illustrated for a chain of atoms ordered quasiperiodically in Fig. 3. The calculated diffraction pattern of such a chain of atoms consist of sharp diffraction peaks spaced quasiperiodically.

A two-dimensional quasiperiodic packing of pentagons was first discovered by the mathematician-physicist Roger Penrose (1974). He was able to pack pentagons in two dimensions with a perfect long-range quasiperiodic order by following certain rules in matching neighboring pentagons. These matching rules were enforced by six different interlocking tiles, resembling the pieces of a jig-saw puzzle (Fig. 4). A beautiful long-range order was revealed by the perfectly aligned small line segments of the pat-

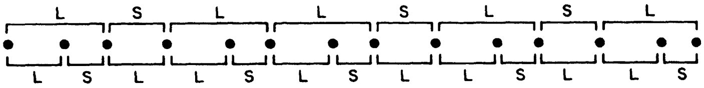

FIG. 3. Sequential spacing of atoms placed on a Fibonacci lattice remain unchanged by the iterative substitution.

tern, forming small sections of infinitely long straight lines. Here it becomes clear that matching rules are a very important part of generating good quasiperiodic order. Such a long-range quasiperiodical order is missing in the disordered pattern of Fig. 2b.

1.2.1 Space filling with Tile Matching

An ideal quasicrystal is constructed by an infinite repetition of two or more unit cells packed in a lattice that has long-range quasiperiodic order. A two-dimensional pentagonal quasicrystal can be constructed using two rhombic tiles, as shown in Fig. 5. These tiles are more convenient versions of the original Penrose tiles. Here, a long-range quasiperiodic order is now enforced through the following local matching rules. In building the two-dimensional quasiperiodic structure from individual tiles, the pointed arrows of the neighboring tiles are matched. Figure 5 also shows a two-dimensional Penrose quasicrystal where all the matching rules are obeyed. In real materials, these matching rules would be equivalent to having different atomic decorations for these two tiles. These structural models were first introduced by Levine and Steinhardt (1984). Independent of the experimental discovery, they generalized the notion of long-range order found in crystals to quasicrystals with quasiperiodic translational order. Using the pair of distinct unit cells or polyhedral tiles mentioned ear-

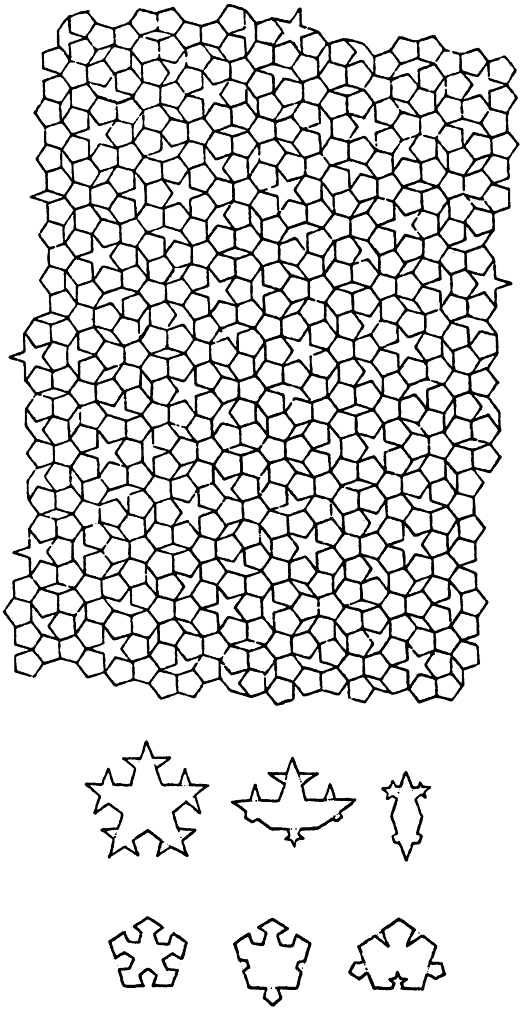

FIG. 4. Pentagons can be packed with a perfect quasiperiodic order, using six different tiles.

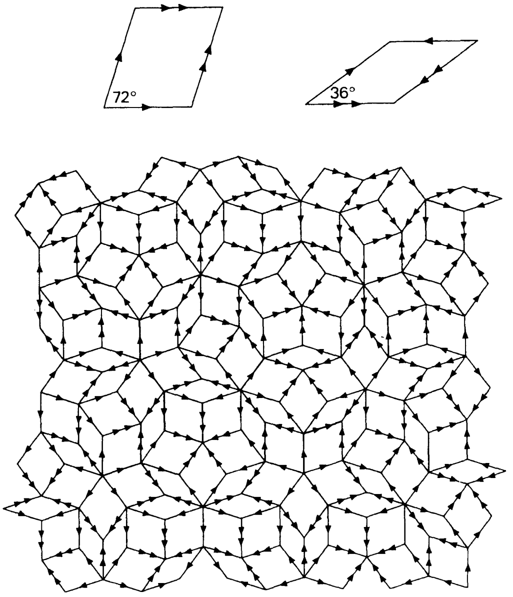

FIG. 5. Rhombic Penrose tiles with matching rules are used to generate a two-dimensional quasicrystalline structure.

lier, they showed that space could be filled quasiperiodically and calculated a diffraction pattern with an icosahedral symmetry and sharp intensities, very similar to the experimental pattern.

Following the local matching rules alone, however, is not sufficient to construct an ideal quasicrystal. Consider building a quasicrystal by adding one tile at a time, using the matching rules. Soon after adding several tiles, there will be a choice to make about which tile to use. With the wrong choice, gaps will appear in the structure that cannot be filled with any tile, as illustrated in Fig. 6. These mistakes can only be corrected by going back a few steps and starting the tiling process all over again. This example demonstrates that, in building a perfect quasicrystal from rhombic tiles, imposing matching rules is a necessary but not a sufficient condition. The self-similarity property of quasicrystals can also be used in generating quasicrystals similar to the one demonstrated for the one-dimensional Fibonacci sequence case. In two dimensions, starting from a fat or a thin rhombus, one can generate an arbitrarily large quasicrystal. This process involves iterative replacements of the fat rhombic tile with three fat and two thin tiles and the thin tile with two fat and two thin tiles, as shown in Fig. 7. These substitutions are also called inflation-deflation rules and were first applied by Penrose in conjunction with the matching rules.

Unlike crystalline structures, quasicrystalline structure patterns are not unique. In other words, in a crystalline system, given many unit cells, these cells can only generate a single pattern, irrespective of how these cells are used, simply because all the cells are identical. In quasicrystals with two unit cells, however, the pattern generated depends on the sequence of tiles that are used in building it. In ideal quasicrystals, this sequence will be only quasiperiodic. Since the pattern generated would be always finite in size, we would always be studying a small section of an infinite quasiperiodic sequence at a time. Two seemingly different patterns may therefore both belong to the same quasicrystal, as these may represent different finite sections of the same sequence. Two patterns with every finite configuration of cells in each occurring in the other are referred to as locally isomorphic, and this is called local

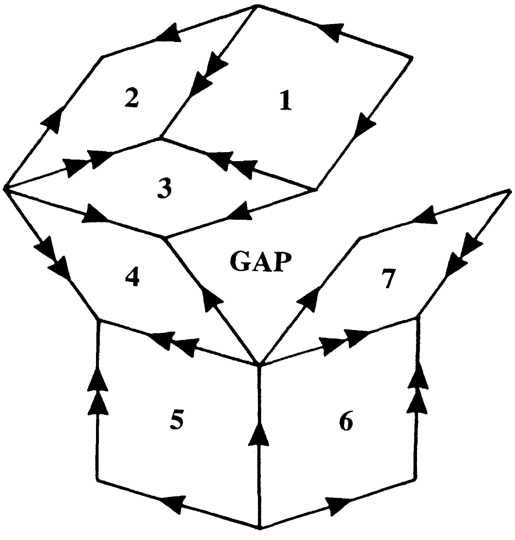

FIG. 6. Matching rules alone are not sufficient to generate perfect quasicrystalline structure.

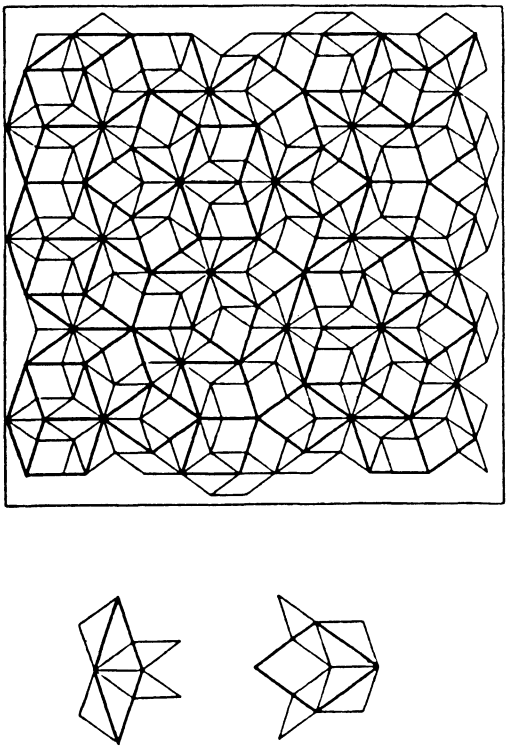

FIG. 7. A Penrose tiling generated by the substitution of each fat and thin tile (*thick lines*) by the set of smaller tiles (thin lines), as shown for individual tiles in the figure.

isomorphism (LI). Figure 8 shows three different tilings, all in different LI classes, generated by using the Penrose tiles. The pentagonal quasicrystal discussed in Fig. 5 is in the so-called Penrose local isomorphism (PLI) class. Here, we limit our discussion only to the PLI class, because of its close connection to real quasicrystals.

1.2.2 Strip-Projection Method Quasicrystals can also be generated from projections of crystalline systems from higher dimensions. Consider a two-dimensional square lattice, and its projection on a one-dimensional straight line that has a slope angle of α, as illustrated in Fig. 9. For an irrational slope, for example $\alpha = 36°$, the line does not intersect any of the lattice sites, and the projection of the lattice sites on this line will form a dense array of points. If we only project the sites that remain within a strip of unit cell dimensions, the projected point separations will then make more physical sense in terms of plausible interatomic distances. Interestingly, for the $\alpha = 36°$ choice, the sequence of projected site spacings now follow the Fibonacci sequence. Note that we generated a quasiperiodic structure in $\mathbf{R}_{\text{par}}$ or physical space by projecting a periodic structure in two-dimensional $\mathbf{R}_{\text{par}}$, $\mathbf{R}_{\text{perp}}$ space. The terms $\mathbf{R}_{\text{par}}$ and $\mathbf{R}_{\text{perp}}$ refer to parallel and perpendicular components, and $\mathbf{R}_{\text{perp}}$ is often referred to as the complementary space. This strip-projection method can be generalized to higher-dimensional quasicrystals. For instance, an icosahedral quasicrystal can be generated from projection of a six-dimensional hypercubic lattice on to the three-dimensional physical space. Just as three vectors **a**, **b**, and **c** are needed to define the vertices of a unit cell of a periodic system in three dimensions, the minimum number of vectors needed to define the vertices of a regular icosahedron is six, as shown in Fig. 10, indicating that the corresponding periodic system is in the six-dimensional space.

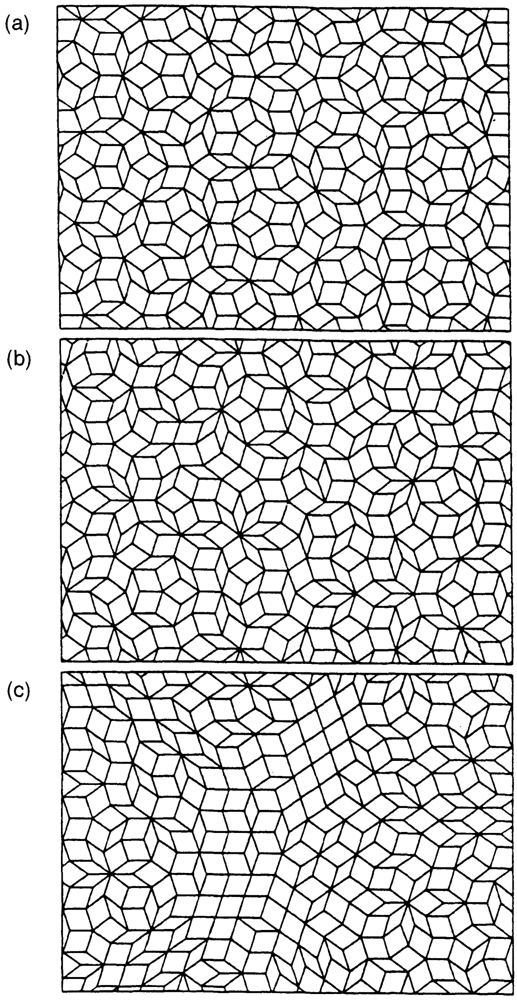

FIG. 8. Three different locally isomorphic quasiperiodic tilings. The patterns are visibly different from each other. Only (a) is generated by using the Fibonacci sequence and is in the Penrose Local Isomorphic class. (Courtesy of J. E. S. Socolar and P. J. Steinhardt.)

1.2.3 Generalized Dual Method This method starts with multiple grids of quasiperiodically spaced parallel lines in two dimensions and surfaces in three dimensions. The intersection points of these grids are

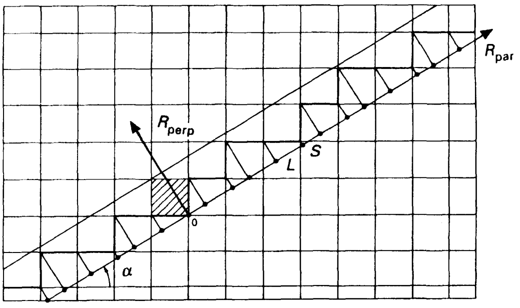

FIG. 9. A one-dimensional Fibonacci quasicrystal obtained by projection-and-cut method from a two-dimensional square lattice.

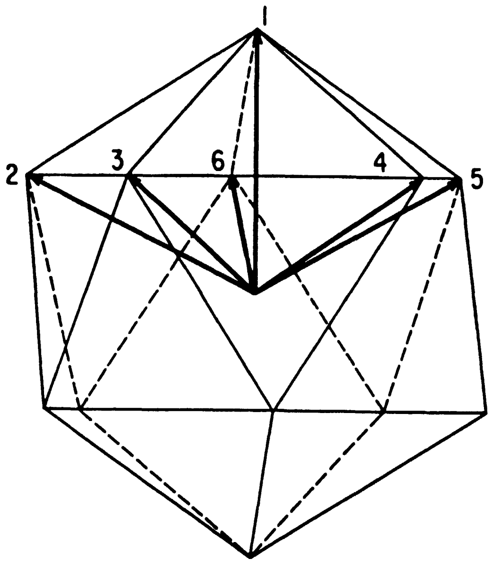

FIG. 10. A regular icosahedron and six unit-cell vectors that define the vertices.

then decorated with tiles. This construction has the following properties: a quasiperiodic translational symmetry, an orientational symmetry, and a minimal separation between atoms for a physical system. For the two-dimensional Penrose lattice, such a construction is referred to as an Ammann quasilattice, and the sequence of spacings between the parallel lines becomes the Fibonacci sequence, as illustrated in Fig. 11. Individual tiles, as shown in the figure, become decorated with sections of straight lines, and lines run continuously across the whole lattice when the tiles are properly matched. Since the model starts from the long-ranged ordered grids, the tilings always have a long-range quasiperiodic order. This method also works for generating three-dimensional quasicrystals. In three dimensions, Penrose tiles take the shape of oblate and prolate rhombohedra, as shown in Fig. 12.

2. QUASICRYSTAL-FORMING COMPOUNDS

Real materials with the exotic structure described above actually exist in nature. Almost all known quasicrystal-forming compounds have a well-known crystalline compound with a similar stoichometry. These crystalline phases usually exhibit an extensive icosahedral local order, and they are referred to as Frank–Kasper phases. Naturally, most of the search for new quasicrystals, therefore, has focussed on alloys that are known Frank–Kasper phase formers.

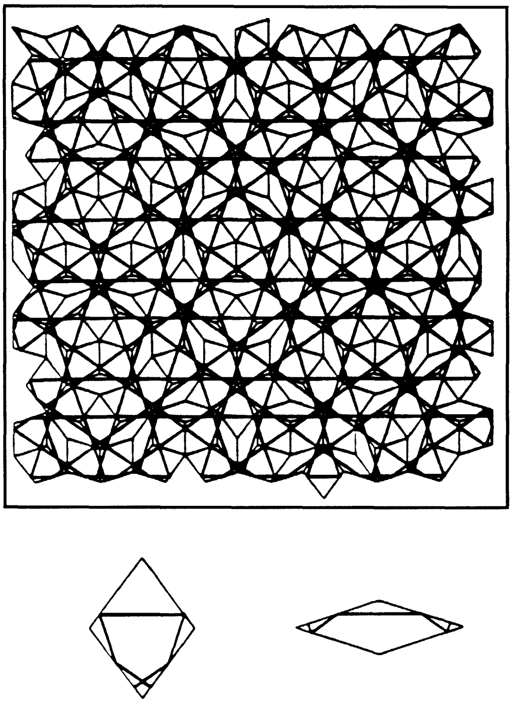

FIG. 11. A Penrose quasicrystal constructed from five sets of quasiperiodically spaced parallel Ammann lines and individual tiles with corresponding decorations.

First discovered in metastable form, quasicrystals had only few-hundred-angstrom-size defect-free regions. After several short years, the number of quasicrystal-forming compounds is now exceeding one hundred. Many new quasicrystalline compounds are discovered with one-, two-, and three-dimensional quasicrystallinity. Several compounds discovered recently by Tsai *et al.* (1989) exhibit a crystalline thermodynamic stability, and a quasicrystalline structural perfection surpassing many crystals.

2.1 Early Metastable Compounds

Quasicrystals were first discovered in aluminum-rich transition-metal (TM) alloys, in the compositional range of 10–14 at.% TM

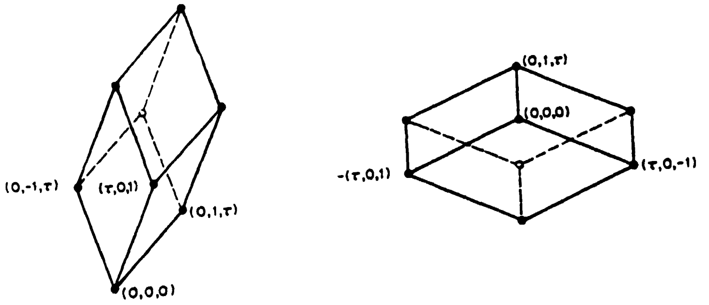

FIG. 12. Prolate and oblate rhombohedra used in three-dimensional Penrose tilings. Vertex coordinates refer to an orthogonal coordinate system.

= Mn, Fe, Cr. These alloys were prepared by rapidly quenching molten metal alloys on a rotating copper cylinder, with typical quench rates of 10^5–10^9 K/s. Ribbonlike samples were several microns thick and a few millimeters wide and contained micron-size icosahedral quasicrystalline grains as well as a few other phases. The addition of a few percent silicon to the alloy greatly improved the stability and structural perfection of the icosahedral-phase grains, which were mostly dendritic in nature. Figure 13 shows a scanning electron microscope (SEM) image of the top surface of a melt-spun ribbon of an $Al_{74}Mn_{20}Si_6$ quasicrystal. Here, individual quasicrystalline grains with sizes a fraction of a micron were resolved, which collectively defined pentagonal features on the surface, characteristic of their quasicrystal symmetry. This is explained by a dendritic growth mechanism, induced by the large undercooling.

These early icosahedral (*i*) phase alloys were all in the *i*-Al-TM class, which included *i*-$Al_{86}Mn_{20}$, *i*-$Al_{79}Cr_{17}Ru_4$, *i*-$Al_{60}Si_{20}(Mn,Cr)_{20}$, *i*-$Al_{65}Ge_{15}(Mn,Cr)_{20}$, and *i*-$Al_{75}Cu_{15}V_{10}$. The structures of these quasicrystals are closely related to the crystalline structure of $Al_9Mn_2Si_2$, and they all form peritectically, i.e., from a liquid-solid reaction. Because of kinetic limitations induced by the undercooling and slow diffusion of the liquid in solid, phase formation never reaches completion, and large compositional inhomogeneities result in a high density of structural defects.

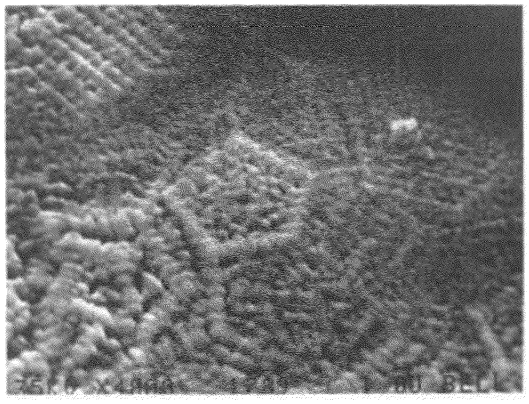

FIG. 13. Surface solidification morphology of a melt-spun $Al_{74}Mn_{20}Si_6$ alloy ribbon.

A well-known crystalline α-{AlMnSi} phase has a composition $Al_{72.5}Mn_{17.4}Si_{10.1}$, which is similar to the quasicrystalline $Al_{74}Mn_{20}Si_6$. Using the known local order found in the α phase, and a projection method, three-dimensional Penrose tiles have been constructed to describe α-{Al-(Mn,Fe)-Si} phases. Illustrated in Fig. 14, these tiles are not believed to promote the *i* phase. Instead, two slightly different packings of a 54-atom icosahedral structural unit [also called Mackay icosahedron (MI)] give rise to α or quasicrystalline phases.

2.2 Stable Compounds

2.2.1 Al_5Li_3Cu Quasicrystal This system appeared to be remarkably stable compared to earlier systems and made it possible to apply conventional techniques to grow large single grains. This ternary compound contains *s*, *p*, and *d* elements and defines a new *spd* class. Other *spd* quasicrystals include *i*-$Al_2Mg_3Zn_3$, *i*-Al_6Mg_4Cu, *i*-Al_6Li_3Au, *i*-$Al_{51}Li_{32}Zn_{17}$, *i*-$Al_{50}Mg_{35}Ag_{15}$, and *i*-$Ga_{16}Mg_{32}Zn_{52}$. Structurally, these are closely related to the $Im\bar{3}$, (bcc) Al_5Li_3Cu crystalline phase, which can be described as a body-centered close-packed structure of icosahedral-shaped

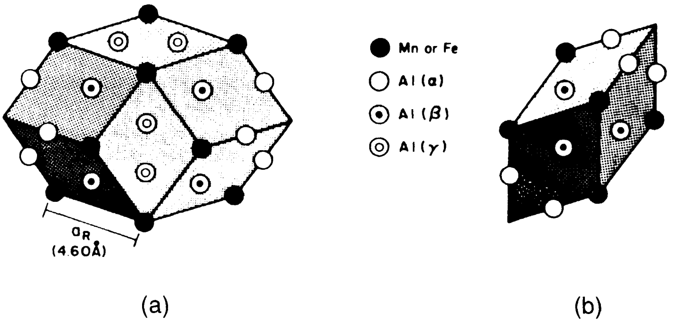

(a) (b)

FIG. 14. (a) Rhombic dodecahedron and (b) prolate rhombohedron, used in constructing α-Al(Mn,Fe)Si structures. (Courtesy of V. Elser.)

atomic clusters. Each cluster, referred to as a Samson complex, consists of shells of Al, Cu, and Li atoms distributed on five successive shells, and contains 104 atoms. These shells have the shapes of icosahedron, pentagonal dodecahedron, truncated icosahedron, and a small triacontahedron, as illustrated in Fig. 15. Based on the local order derived from the bcc phase, three-dimensional Penrose tiles have been constructed by atomic decoration of canonical Penrose tiles, as shown in Fig. 16. A reasonable agreement with the experimental diffraction data has been achieved by introducing a partial chemical disorder.

In most quasicrystal-forming systems, the quasicrystalline icosahedral phase I occurs as a noncongruently melting compound close to a congruent crystalline phase in the phase diagram. Quasicrystals then form by a peritectic reaction of the melt with the crystalline phase. For the $Al_{5.1}Li_3Cu$ quasicrystalline phase, this neighboring crystalline phase is the $Al_{4.8}Li_3Cu$ body-centered cubic phase, as shown in Fig. 17. Both compounds span a narrow range in composition and appear to be thermodynamically stable. Using a solute (Al)-rich starting liquid, large single grains are grown directly from the liquid using arbitrarily slow cooling rates. When the liquid is aluminum rich ($x > 6$, in Fig. 17), and cooled below the liquidus line, the only phase that nucleates and grows is the *i* phase. It is essential, however, in this approach to cool very slowly to minimize the coring effect and not to induce constitutional undercooling at the liquid-solid interface that will initiate a dendritic growth. Centimeter-size single quasicrystal grains have been grown using this approach. X-ray diffraction measurements find a large mosaic spread (~1°) in these samples, indicating some structural disorder. In another approach to growing quasicrystals, a stoichiometric composition can be cooled slowly through the liquidus line. This results in nucleation and growth of the cubic phase. By the time the melt is cooled to the *i*-phase boundary, the melt would contain large grains of the cubic phase, and the quasicrystal growth will be hindered by the slow diffusion of the liquid in these solid grains of the cubic phase. This approach yields kinetically limited growth and poor-quality quasicrystals. The liquidus region can also be quickly by-passed by rapid quenching of the liquid before giving the cubic phase a chance to nucleate. In principle, this results in a very small grain size because of the large nucleation density, as demonstrated by small-grain ribbons produced by melt-spinning experiments. In cast-cooled samples, single-grain quasicrystals are

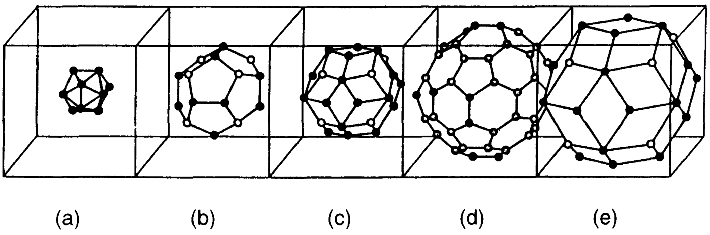

(a) (b) (c) (d) (e)

FIG. 15. Structure of the Samson complex for the **R** phase. Five successive shells are denoted by a, b, c, d, and e. (Courtesy of L. Pauling.)

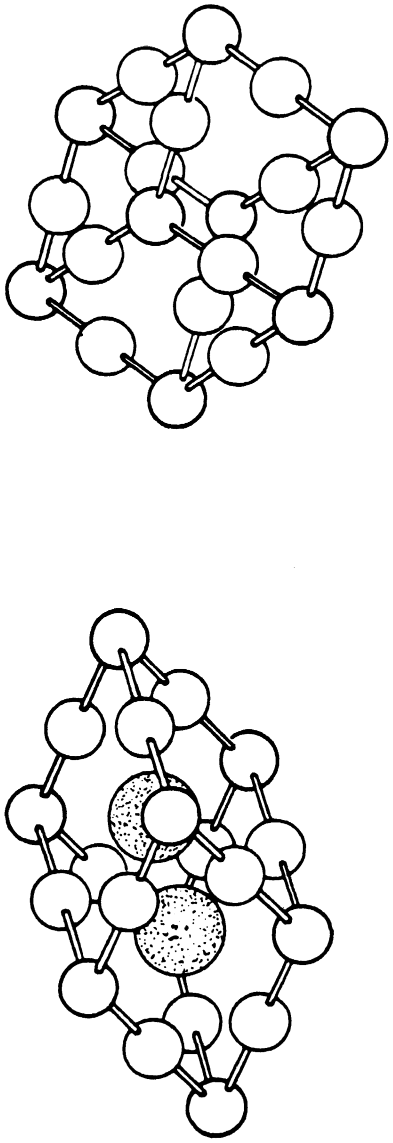

FIG. 16. Atomic decoration of prolate and oblate Penrose rhombohedra for the Al-Li-Cu quasicrystal. Small open circles refer to Al or Cu atoms, and large circles refer to Li atoms. (Courtesy of V. Elser.)

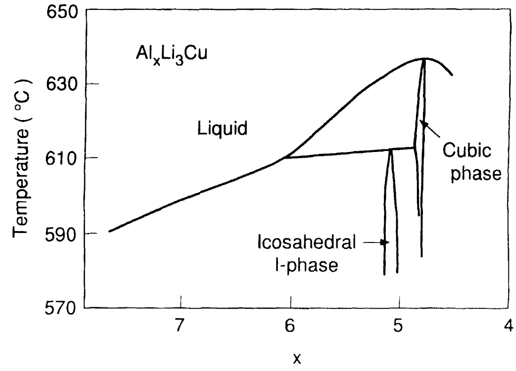

FIG. 17. A partial phase diagram of the Al-Li-Cu system. The i phase, typical of most quasicrystals, occurs next to a crystalline phase as a noncongruent compound.

observed in shrink cavities. Figure 18 shows a hundred-micron–size single grain with very well-developed facets. Solidification morphology is a triacontrahedron, with twofold planar facets, which indicates that growth is limited by a small nucleation rate on these planes.

A close structural relationship between the cubic and i phases is also demonstrated morphologically. Figure 19 shows a quasiepitaxial growth of i-phase grains on a crystalline cubic grain. Here, three distinct quasicrystalline i-phase grains nucleate and grow on the {111} vertices of a cubic grain. The threefold axes of the i-phase grains are aligned with the cubic threefold axes, and i-phase twofold planes are aligned with the cubic {111} planes. Interestingly, growth-limiting cubic {100} planes share the same alignment with the growth-limiting quasicrystalline twofold planes, suggesting a quasiepitaxial interface between the two grains.

2.2.2 $Al_{65}Cu_{20}(Fe,Ru,Os)_{15}$ Quasicrystals

Quasicrystals of this remarkable class are very well ordered, with the size of defect-free

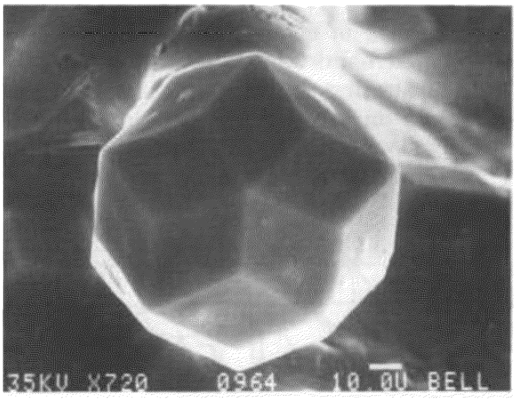

FIG. 18. A single grain of Al-Li-Cu quasicrystal. Naturally formed facets reveal the icosahedral phase symmetry.

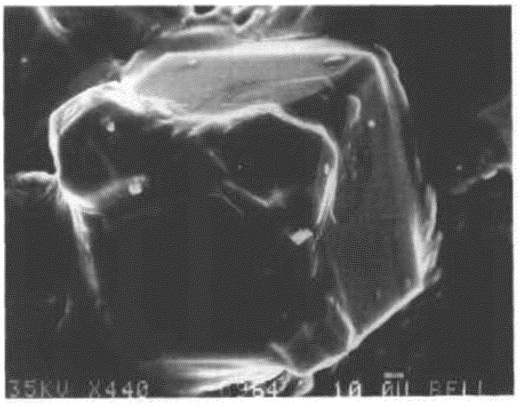

FIG. 19. *i*-phase grains grow quasiepitaxially on the cubic **R**-phase grain.

regions exceeding thousands of angstroms. The ternary phase diagram of this system, however, is very complicated. Attempts to grow large single quasicrystals were not successful because of an unusually large nucleation rate. The largest grains obtained were fractions of a millimeter in size. An additional complication has been the existence of nearby low-temperature crystalline phases, which the quasicrystalline phase sometimes transforms into. Figure 20 shows an electron microscopy image of the grain structure in a cast-cooled $Al_{65}Cu_{20}Fe_{15}$ sample. The grains have pentagonal dodecahedral morphology with very well-developed facets, indicating that the growth is limited by the layer nucleation process on the fivefold surface.

2.2.3 Decagonal Quasicrystals Discovered by Bendersky, decagonal quasicrystals are quasicrystalline in only two dimensions and crystalline in the third dimension. The atomic structure can be viewed as periodically stacked layers of two-dimensional, planar atomic structures with quasicrystalline order. Although the original discovery was made on thermodynamically unstable micron-size grains of Al-Mn alloys, stable forms occur in $Al_{65}Cu_{15}Co_{20}$ and $Al_{70}Ni_{15}Co_{15}$ ternary compounds. In these stable systems, there appears to be a vanishingly small peritectic liquidus region, which makes it possible to grow large grains under equilibrium conditions starting from stoichiometric liquids. Well-isolated single grains with ten naturally formed facets can be observed in cast-cooled samples, as shown in Fig. 21. Several millimeter-size single grains are grown under equilibrium conditions with very slow cooling rates. Single-crystal x-ray diffraction measurements of these grains indicate few-thousand-angstrom–size defect-free regions and confirm the two-dimensional quasiperiodic structure. Figure 22 shows the single-crystal scans made along the quasiperiodic and periodic directions.

The large grain sizes of decagonal quasicrystals also allow the application of conventional surface structure measurement techniques. Scanning tunneling microscopy (STM) measurements with atomic resolution also provide very valuable real-space information about the atomic structure. Figure 23 illustrates an STM image of the quasiperiod-

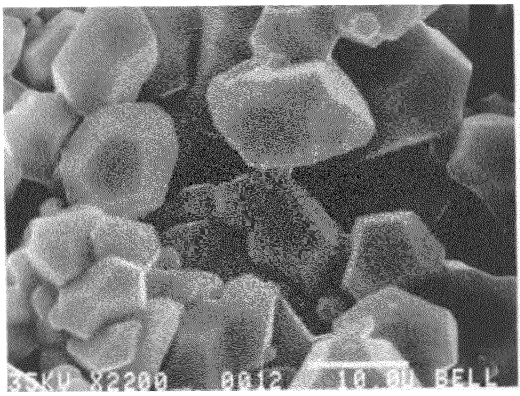

FIG. 20. Pentagonal dodecahedron-shaped single grains of $Al_{65}Cu_{20}Fe_{15}$ quasicrystal.

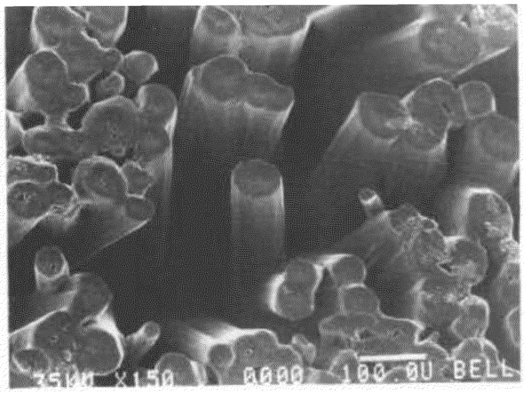

FIG. 21. Decagonal quasicrystals grow with a tenfold columnar morphology.

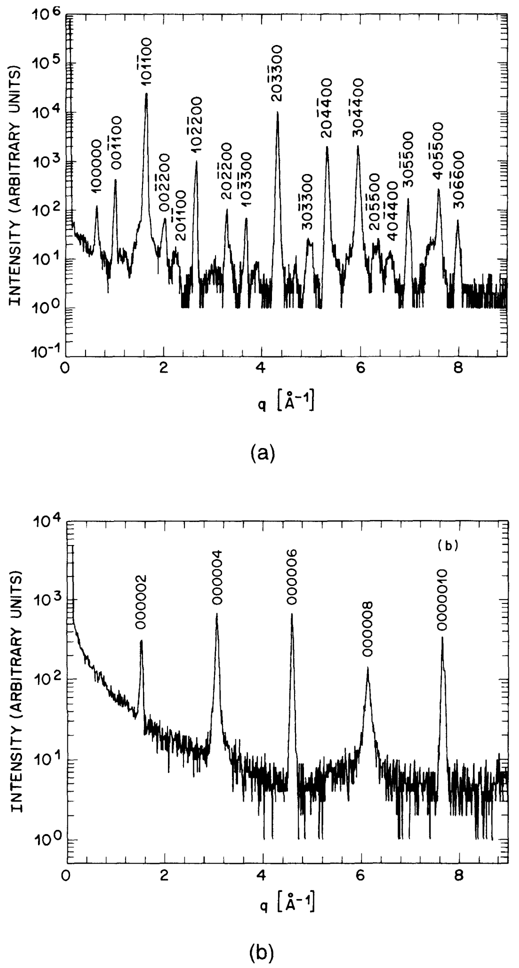

FIG. 22. Single-grain x-ray diffraction measurements of a decagonal quasicrystal, along the (a) quasiperiodic and (b) periodic directions.

ic surface layer, in both the top and tilted views. Most of the atoms and periodically spaced terraces along the *c* axis are resolved. The tilted-image view shows that individual atoms lie on straight lines, an indication that the structure is defect free on large length scales. These mass-density lines resemble the Ammann quasilattice mentioned earlier. The fact that lines run uninterrupted across terraces suggest that layers are packed on top of each other along the periodic direction. A structural model with reasonable interatomic distances can be constructed on the basis of this information. Figure 24 shows a local STM image, a Penrose lattice with atomic decoration, and their overlay. The model assumes a pentagonal network mostly made out of aluminum, and a center decoration with mostly transition-metal atoms. A similar local image illustrates two-dimensional Penrose tiling, using fat and thin tiles, as shown in Fig. 25. Individual tiles, shown in the inset, demonstrate the presence of different atomic decorations for these tiles. STM images find quasicrystalline atomic order ranging in scale from a few angstroms to several hundred angstroms.

2.2.4 $Al_{70}Pd_{20}Mn_{10}$ Quasicrystal This stable icosahedral quasicrystal is particularly important because of its vanishingly small peritectic range. A partial phase diagram of this ternary system is shown in Fig. 26. Depending on the composition, this system forms face-centered-icosahedral (FCI), simple icosahedral (SI), or decagonal-like (D) phases. The FCI phase exhibits a strong chemical ordering and a remarkable structural perfection. Figure 27(a) shows a single-grain FCI phase grown over a period of 3 weeks. The only measurable defect in this grain is a small-angle (~2°) grain boundary, as indicated by the dark line on the surface. Structural perfection is also evident in the transmission Laue x-ray pattern taken from this grain shown in Figure 27(b). Dynamical x-ray diffraction, which is extremely sensitive to structural imperfections, has also demonstrated that $Al_{70}Pd_{20}Mn_{10}$ quasicrystals are almost perfect.

2.3 Predicting New Quasicrystals; Quantum Structure Diagrams

Searching for new quasicrystalline compounds is a very difficult process, considering that there are thousands of ternary compounds. Quantum structure diagrams (QSD) have proven to be very useful in predicting new compounds, especially in *spd* systems. These diagrams are constructed from atomic size ratios, electron concentrations, electronegativity differences, and other parameters. Figure 28 shows such a diagram for all the known stable quasicrystals in a plane whose coordinates are compositionally weighted ternary size ($\overline{\Delta R}$) and electronegativity difference ($\overline{\Delta X}$). All stable quasicrystals fall within a square region, indicating that these diagrams can, in fact, distinguish small differences between different ternary compounds.

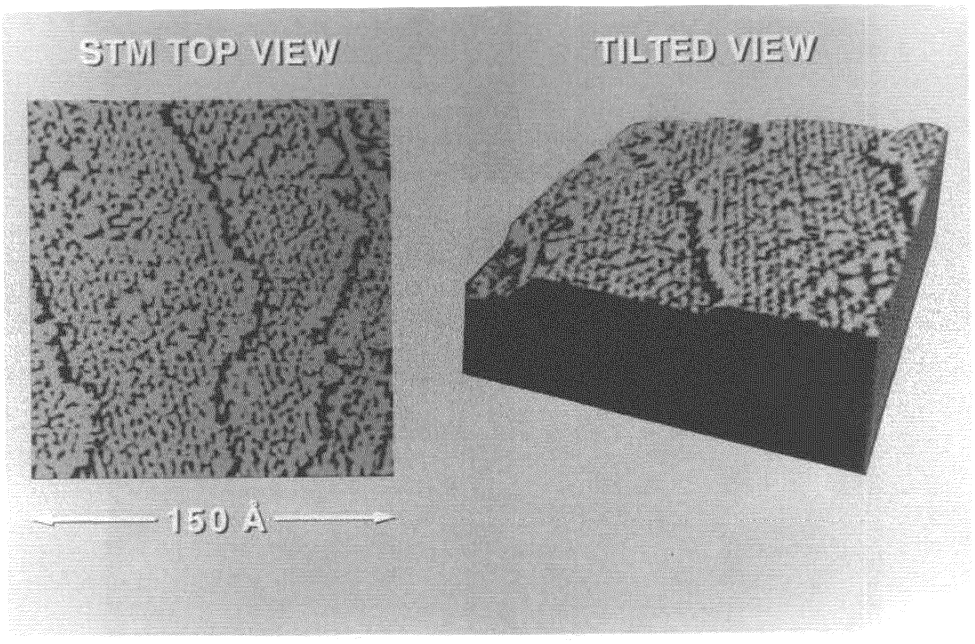
STM TOP VIEW
TILTED VIEW
150 Å

FIG. 23. A scanning tunneling microscopy image of the quasiperiodic planes of the $Al_{65}Cu_{15}Co_{20}$ decagonal quasicrystal. (Courtesy of R. S. Becker.)

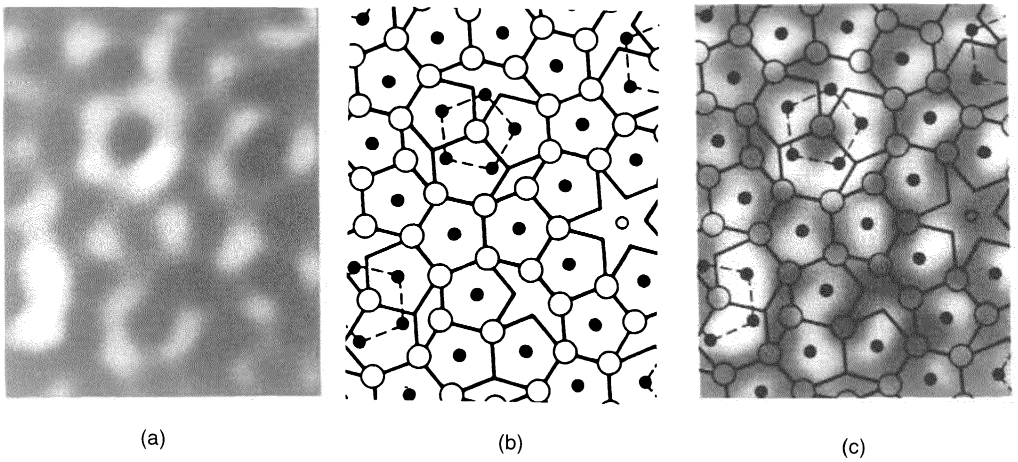
(a) (b) (c)

FIG. 24. An atomic structure model for the quasicrystalline plane.

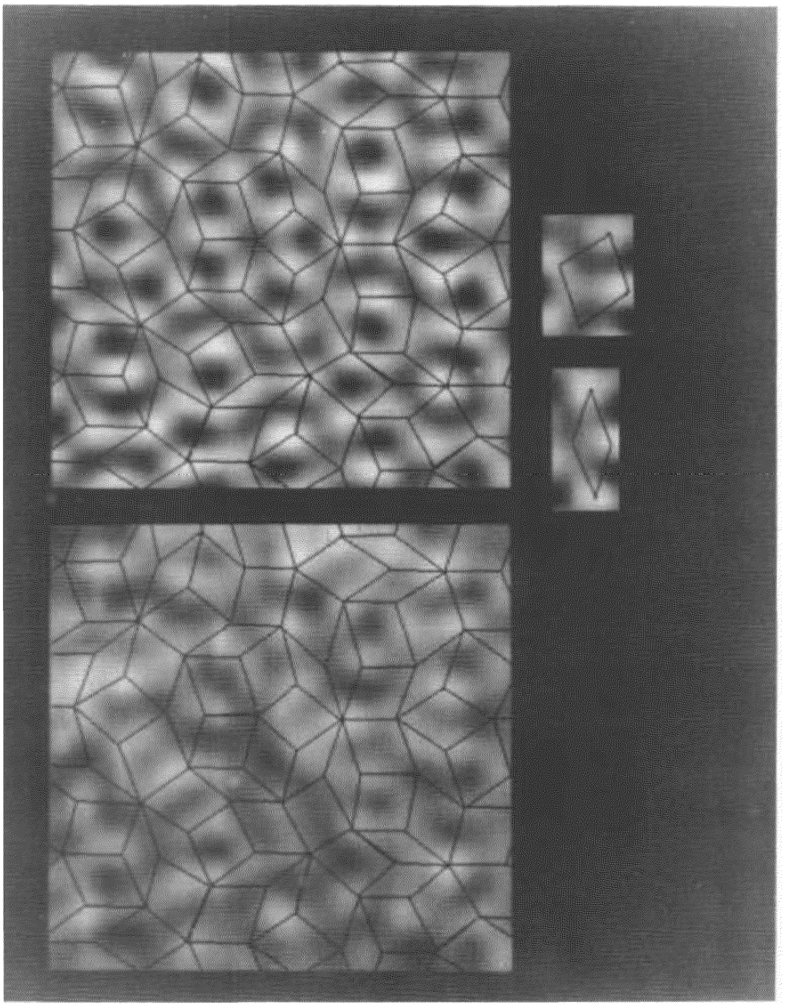

FIG. 25. Two-dimensional Penrose tiles (inset) are used to fit the raw and (upper) enhanced images.

These stable compounds are (1) $Al_{60}Li_{30}Cu_{10}$, (2) $Zn_{43}Mg_{37}Ga_{20}$, (3) $Al_{65}Cu_{20}Fe_{15}$, (4) Al_{65}-$Cu_{20}Ru_{15}$, (5) $Al_{65}Cu_{20}Os_{15}$, (6) $Al_{65}Co_{20}Cu_{15}$, (7) $Al_{70}Co_{15}Ni_{15}$, (8) $Al_{70}Pd_{20}Mn_{10}$, and (9) Al_{70}-$Pd_{20}Re_{15}$. The region enclosed by the dashed line contains *spd* quasicrystal-forming systems alone. For *spd*-type compounds, similar quantum diagrams were able to pick 55 possible ternaries out of some 2600 systems, and four new *spd* quasicrystals, Al_6Li_3Au, $Al_{51}Li_{32}Zn_{17}$, $Al_{50}Mg_{35}Ag_{15}$, and $Ga_{16}Mg_{32}Zn_{52}$, have been discovered with the help of this list.

3. QUASICRYSTALLOGRAPHY

3.1 Icosahedral Quasicrystals

There are a few different conventions in indexing an icosahedral pattern, all making use of six indices. One convenient method is choosing a cubic coordinate system alligned with the three orthogonal twofold axes of the icosahedron, as shown in Fig. 29. In this coordinate system, fivefold vertices of the icosahedron are of the form $(1\ \tau\ 0)$, and in general any of the high-symmetry axes can be

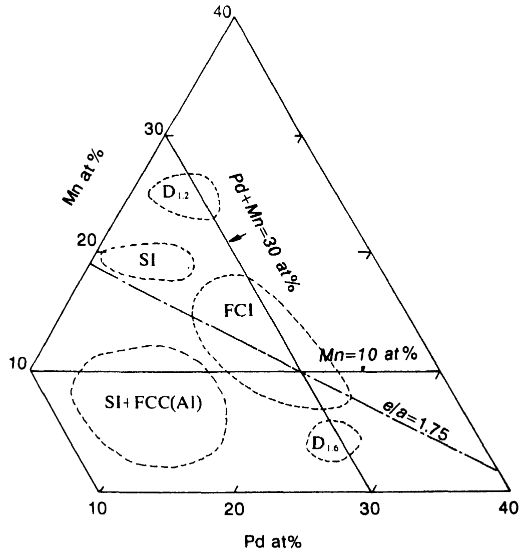

FIG. 26. Al-Pd-Cu phase diagram and regions where different quasicrystals form. (Courtesy of A. P. Tsai.)

expressed with indices of the form

$$(h + h'\tau \; k + k'\tau \; l + l'\tau)$$

or $(h/h', k/k', l/l')$ for short. Here h, h', k, k', l, l' are all integers and $\tau = 2 \cos 36°$. The need of a minimum number of six Miller indices implies the existence of a six-dimensional hypercubic system, whose projection yields three-dimensional quasicrystallinity. The six vectors along the fivefold axes then become

$$\begin{aligned} q_1 &= (1/0\ 0/1\ 0/0), \\ q_2 &= (0/1\ 0/0\ 1/0), \\ q_3 &= (0/0\ 1/0\ 0/1), \\ q_4 &= (\bar{1}/0\ 0/1\ 0/0), \\ q_5 &= (0/1\ 0/0\ \bar{1}/0), \\ q_6 &= (0/0\ \bar{1}/0\ 0/1). \end{aligned}$$

Any lattice site in reciprocal space can now be expressed as

$$Q = n_1 q_1 + n_2 q_2 + n_3 q_3 + n_4 q_4 + n_5 q_5 + n_6 q_6.$$

Here the n's are integers. In terms of three-dimensional cubic coordinates,

$$Q = \{(n_1 - n_4)/(n_2 - n_5), (n_3 - n_6)/(n_1 - n_4), (n_2 - n_5)/(n_3 - n_6)\}.$$

Combining this with the symmetry operations of the icosahedral group, certain restrictions on the indices and extinction rules are derived. Accordingly, three reciprocal lattices possible in six dimensions, primitive (P), face-centered (F), and body-centered (I), have different extinctions in the diffraction pattern. Figure 30 shows the calculated diffraction patterns of P, F, and I lattices along the twofold and fivefold zone axes. In order to make a distinction in identifying a structure, diffraction patterns along more than one zone axis must be examined.

3.2 Decagonal Quasicrystals

Decagonal quasicrystals, or two-dimensional quasicrystalline systems, require a minimum of four unit vectors to define the reciprocal lattice sites of a planar layer. A fifth index is introduced to describe the *c* axis or the periodic direction component. In terms of projection methods, this implies that the structure can be generated by projecting from a five-dimensional tetragonal lattice to three dimensions.

4. QUASICRYSTAL STRUCTURE AND GROWTH MODELS

4.1 Early Models

As mentioned earlier, a defect-free quasicrystal would have very sharp (delta function) diffraction peaks, just as a perfect crystal does, and the presence of defects would broaden these peaks. A remarkable property of quasicrystals is that even a highly defective sample, e.g., the structure in Fig. 2, would still give a quasiperiodic diffraction pattern, but with broadened peaks. This is due to the orientational order that exists even in disordered quasicrystals. An early model based on random packing of icosahedral clusters was therefore quite successful in explaining the defect-broadened diffraction patterns of earlier samples. Another model attempted to explain the forbidden symmetries by multiple twinning of very small crystallites and found reasonable agreement with these broad peaks. Stable quasicrystalline compounds discovered more recently, however, exhibit very sharp diffraction peaks, demonstrating that the observed

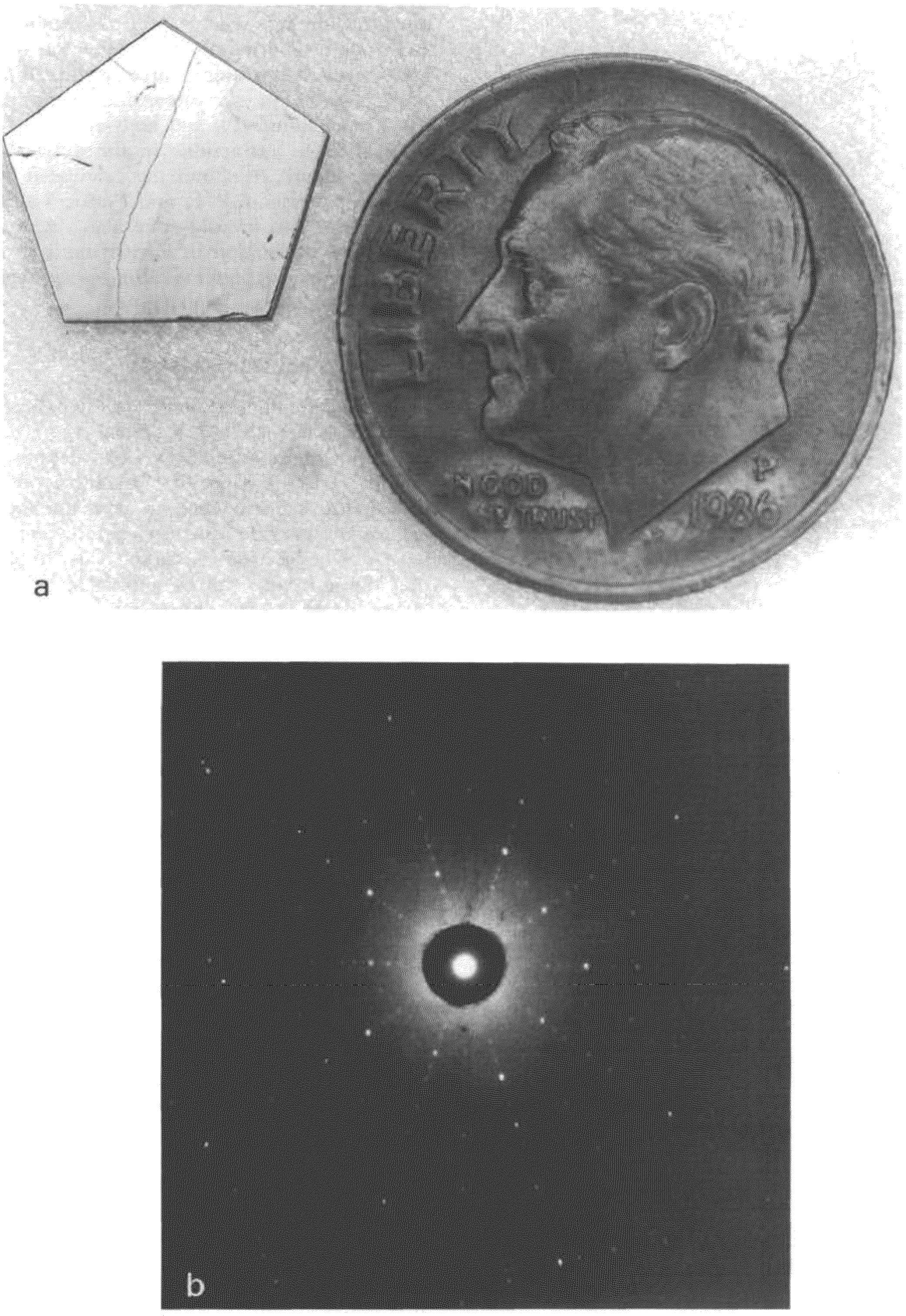

FIG. 27. (a) A single-grain Al-Pd-Mn icosahedral quasicrystal and (b) its x-ray diffraction pattern.

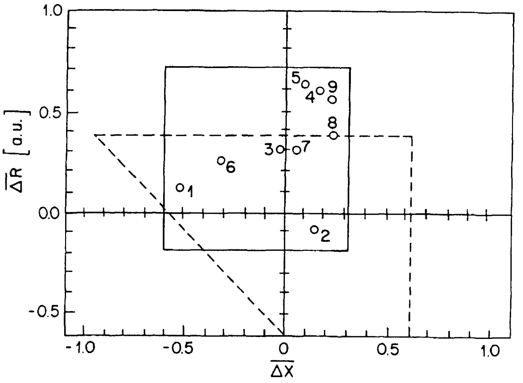

FIG. 28. A typical quantum structure diagram used in predicting new quasicrystals. Square box encloses all known stable compounds. (Courtesy of K. M. Rabe and J. C. Phillips.)

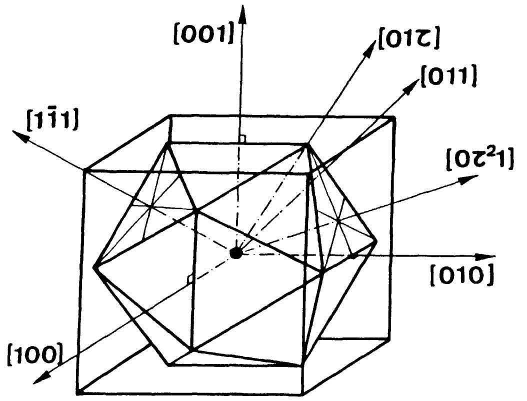

FIG. 29. An icosahedron oriented with cube axes is used in an indexing method.

disorder is not an intrinsic property of quasicrystals, and quasicrystallinity is the most reasonable structure.

4.2 Growth of Perfect Quasicrystals

An important problem in quasicrystal growth is the generation of long-range order from short-range interactions. In generating perfect tilings, long-range order was artificially introduced into the system, as in the case of Ammann lines or grid lines. If we assume that Penrose tiles with local matching rules exist in physical systems, with matching rules replaced by some interatomic forces, we face the problem explained in Fig. 7: that a defect-free quasicrystal cannot exist. As more tiles are added to the system, the long-range aperiodic translational order will be significantly reduced, and entropy will determine the degree of order that exists in the system. It is clear that the original matching rules cannot produce perfect quasicrystals.

Remarkably, however, if certain restrictions are imposed on the original matching rules by only allowing eight different types of local patterns of Fig. 31 to occur in the lattice, then the problem encountered in Fig. 7 can be avoided, and a nearly defect-free two-dimensional Penrose lattice can be constructed. This restriction on local patterns is referred to as a vertex-matching rule. This construction suggests that a long-range quasicrystalline order in real systems can be explained by a finite-range, local interaction between the atoms.

Perhaps a more physical model, the "cluster picture," offers a more refined explana-

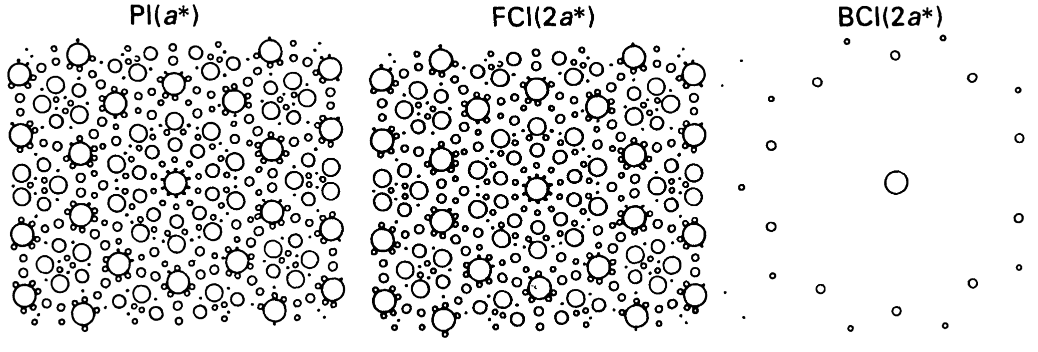
PI(*a**)
FCI(2*a**)
BCI(2*a**)

FIG. 30. Primitive, face-centered, and body-centered icosahedral phase diffraction patterns as viewed along the fivefold zone axis. (Courtesy of J. W. Cahn.)

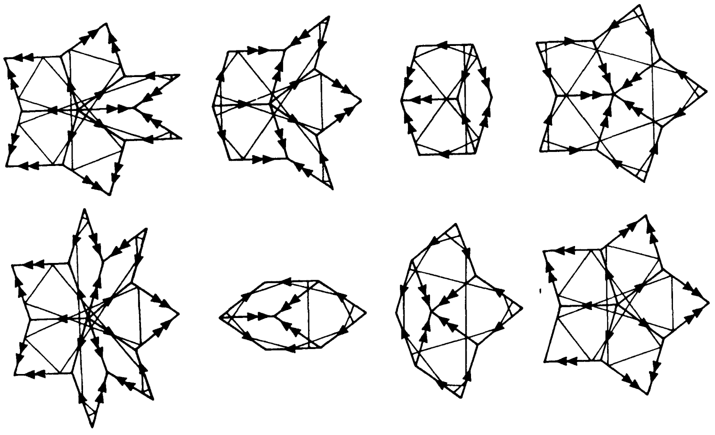

FIG. 31. Vertex-matching rules used in growing perfect quasicrystals. (Courtesy of P. J. Steinhardt.)

tion of the structure and formation of quasicrystals. Starting with a set of one or more small atomic clusters of low energy, an energetically stable quasiperiodic structure is generated by an energy-minimizing calculation that allows partial overlap of these clusters. This model generates energetically stable quasicrystals that are distinct from Penrose tilings.

5. DEFECTS IN QUASICRYSTALS

5.1 Phonons and Phasons

Quasicrystals have defects and elementary excitations that are closely related to their aperiodic translational order and rotational symmetry. These are identified in a density-wave description, where the mass density is expanded in a Fourier series:

$$\rho(\mathbf{r}_n) = \sum_{\mathbf{G}_n} \rho_{\mathrm{G}} \exp[i(\mathbf{G}_n \cdot \mathbf{r}_n)],$$

where $\mathbf{G}_n$ is a reciprocal vector in an n-dimensional space and ρ_{G} is a complex number with amplitude $|\rho_{\mathrm{G}}|$ and a phase ϕ_{G}. Consider displacing atoms of the lattice from their equilibrium positions by $\mathbf{u}$; then the phases are modified by

$$\phi_{\mathrm{G}} = \phi_{\mathrm{G}}^{0} + \mathbf{G}_n \cdot \mathbf{u}_n.$$

For the physical case of interest, i.e., for three-dimensional space, the projection method preserves the amplitudes and phases, and the density wave becomes

$$\rho(\mathbf{r}_{\mathrm{par}}) = \sum_{\mathbf{G}_{\mathrm{par}}} \rho_{\mathrm{G}} \exp[i(\mathbf{G}_{\mathrm{par}} \cdot \mathbf{r}_{\mathrm{par}})],$$

with the physical phases

$$\phi_{\mathrm{par}} = \phi_{\mathrm{G}_n} = \phi_{\mathrm{G}}^{0} + \mathbf{G}_{\mathrm{par}} \cdot \mathbf{u}_{\mathrm{par}} + \mathbf{G}_{\mathrm{perp}} \cdot \mathbf{u}_{\mathrm{perp}}.$$

Here, parallel and perpendicular components refer to the physical-space and complementary-space components, similar to the case described in Fig. 9. The $\mathbf{u}_{\mathrm{par}}$ component is called the phononlike component since it represents a real displacement in physical space. The $\mathbf{u}_{\mathrm{perp}}$ component, referred to as phasonlike modes, represents displacements in the complementary space and has the effect of locally rearranging the tiles. This tile rearrangement will only effect a local region without disturbing the rest of the lattice. In real systems, this will correspond to the local rearrangement of atoms. The $\mathbf{u}_{\mathrm{par}}$ is also referred to as the propagating mode and has an analog in crystals as phonons, whereas $\mathbf{u}_{\mathrm{perp}}$, also referred to as a diffusive mode, is unique to quasicrystals. Figure 32 illustrates phonon and phason strain fields for the two-dimensional Penrose lattice. Phononlike disturbances induce continuous and long-wavelength fluctuations in the pentagrid and will have a Debye–Waller-type effect on the diffraction pattern. Peaks will decrease in intensity and broaden with increasing q in reciprocal space. Phasonlike disturbances, meanwhile, can be detected as discontinuities or jags on the pentagrid, and cause shifts in the Bragg peak positions in proportion to the perpendicular component of the peaks. These types of defects were com-

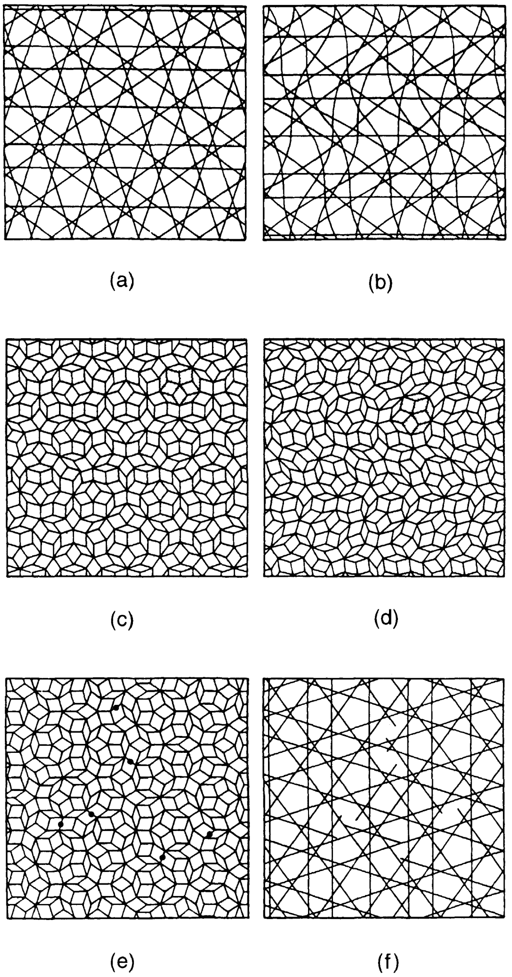

FIG. 32. Pentagrids corresponding to (a) perfect and (b) phonon-strained quasicrystals, and (c), (d) their respective Penrose tilings. Penrose-matching rules are violated (filled circles) in phason-strained tiling in (e), and the corresponding pentagrid displays jags in the pattern. (Courtesy of J. E. S. Socolar and P. J. Steinhardt.)

monly observed in early quasicrystals. $\mathbf{u}_{\text{par}}$ and $\mathbf{u}_{\text{perp}}$ are also referred to as $\mathbf{u}$ and $\mathbf{w}$, respectively.

5.2 Dislocations

Another defect that occurs in quasicrystals is the dislocation. Similar to the case of crystals, dislocations in quasicrystals are characterized by Burgers vectors. The Burgers vector $\mathbf{b}$ of a quasicrystal is, however, an n-dimensional vector (e.g., $n = 6$ for icosahedral quasicrystal). The well-known expression for crystals,

$$\mathbf{G}\cdot\mathbf{b} = 2\pi m,$$

where m is an integer, is modified for quasicrystals as

$$\mathbf{G}_{\text{par}}\cdot\{\mathbf{u}(2\pi) - \mathbf{u}(0)\} + \mathbf{G}_{\text{perp}}\cdot\{\mathbf{w}(2\pi) - \mathbf{w}(0)\} = 2\pi m.$$

Dislocations in quasicrystals have both phonon and phason contributions to the strain field. Figure 33 shows a dislocation on a two-dimensional Penrose lattice and its corresponding pentagrid.

6. PROPERTIES

6.1 Electronic

The charge-transport mechanism in crystals is closely related to the periodicity of the system. Here, the energy band concept and the Bloch theorem are all developed to treat the case of long-range periodic order. The electronic states of a crystal are extended in real space, meaning that charge carriers at the Fermi level ideally can travel freely throughout the system. For a quasicrystal, the potential energy is no longer periodic but has a dense set of Fourier components defined by peaks in reciprocal space. This translates into a dense set of energy gaps at all values of energies. Overall electrical conductivity is expected to be low. Experimentally, measured resistivities increase dramatically with the improved perfection of

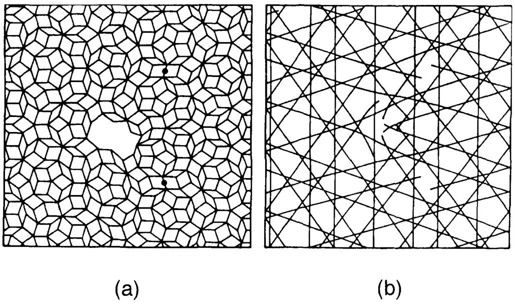

FIG. 33. (a) A dislocation in a Penrose tiling and (b) its corresponding pentagrid.

quasicrystals. Good quasicrystals are found to be marginally metallic with resistivities as high as 30 000 $\mu\Omega$ cm. Figure 34 illustrates measured resistivities of $Al_{65}Cu_{15}Co_{20}$ and $Al_{70}Ni_{15}Co_{15}$ single-grain decagonal samples. Since these samples have both periodic and quasiperiodic directions, it is possible to make valuable structure-related comparisons on the same sample. The temperature dependence of the resistivities demonstrates that along the periodic direction, conduction is metallic, whereas along the quasiperiodic direction, it is nonmetallic, and a phonon-assisted mechanism is suggested. The resistivity ratio of periodic to quasiperiodic directions is about a factor of 8.

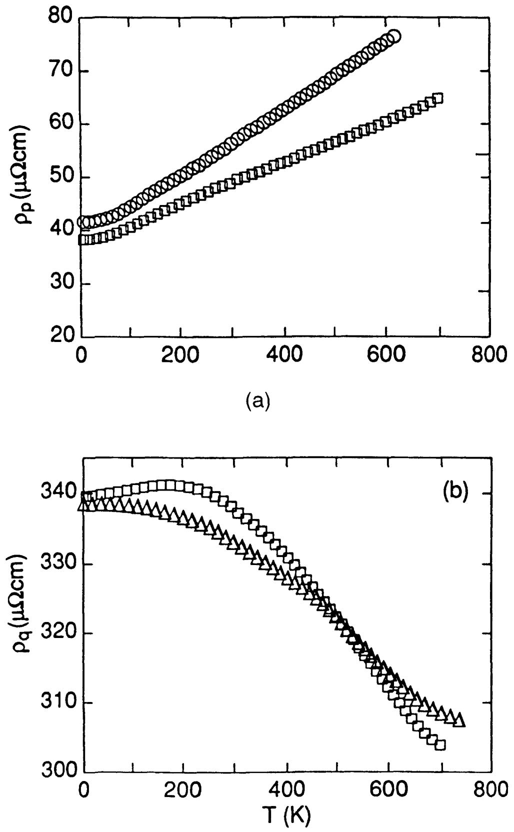

(a)

(b)

FIG. 34. Temperature dependence of the resistivities (a) along the periodic direction are different in character compared to (b) the quasiperiodic direction, for Al-Cu-Co (circles) and Al-Ni-Co (squares) decagonal quasicrystals. (Courtesy of S. Martin.)

6.2 Magnetic

Magnetism in solids is related to the splitting of the *f* and *d* energy levels under the crystal-field symmetry. Since quasicrystals are more isotropic in symmetry compared to crystals, splitting is expected to be small. Experimentally, icosahedral quasicrystals display magnetic behavior ranging from weak paramagnetism to weak diamagnetism. Good quasicrystals, however, exhibit a weakly diamagnetic behavior. This can be understood by the reduced number of conduction electrons, which give rise to paramagnetism. The observed diamagnetism may then be mainly due to core electrons and the cyclotron motion of the conduction electrons, which may become significant in quasicrystals.

7. SUMMARY

Quasicrystals are an unusual form of condensed matter with a long-range quasiperiodic order and a long-range orientational order. Discovered only a decade ago in an unstable form in Al-Mn alloys, quasicrystals are now being discovered in many different systems, some in stable form and some with structural perfections rivaled only by crystals. Remarkable progress has also been made in understanding the structure, growth, and properties of quasicrystals, but there still remain many important issues unanswered or unexplored. In this article, we introduced some of the basic features of quasicrystals including their quasicrystalline structure modeling, generation and growth models, real quasicrystals, and some properties. For a comprehensive treatment of the subject, readers are referred to the books listed below.

GLOSSARY

Ammann Quasilattice: An ordered quasilattice, with a long-range aperiodic translational order, generated by matching segments of straight lines that decorate the Penrose tiles.

Fibonacci Sequence: A one-dimensional and self-similar aperiodic sequence.

Frank–Kasper Phases: Crystalline phases with an extensive icosahedral local order.

Penrose Lattice: A special kind of quasiperiodic lattice with a perfect long-range orientational and a long-range aperiodic translational order.

Penrose Tiles: A pair of fat and thin rhombic tiles, which are the basic building blocks of a Penrose lattice.

Quasicrystal: A newly discovered form of condensed matter that has a long-range orientational order, an aperiodic translational order, and a noncrystallographic space group.

Works Cited

Levine, D., Steinhardt, P. J. (1984), *Phys. Rev. Lett.* **53,** 2477–2480.

Penrose, R. (1974), *Bull. Inst. Math. Appl.* **10,** 266–271.

Schectman, D., Blech, I., Gratias, D., Cahn, J. W. (1984), *Phys. Rev. Lett.* **53,** 1951–1953.

Tsai, A. P., Inoue, A., Masumoto, T. (1989), *Mater. Trans. JIM* **30,** 666–676.

Further Reading

Kelton, K. F., Gibbons, P. C., Goldman, A. I. (Eds.) (1993), "Proceedings of the 4th International Conference on Quasicrystals," *J. Non-Cryst. Solids,* **153 & 154,** 1–674.

Janot, C. (1992), *Quasicrystals, A Primer,* Oxford: Clarendon Press.

DiVincenzo, D., Steinhardt, P. H. (1991), *Quasicrystals; State of the Art,* Singapore: World Scientific.

Steinhardt, P. H., Ostlund, S. (1987), *The Physics of Quasicrystals,* Singapore: World Scientific.

Jaric, M. J. (Ed.) (1988), *Introduction to Quasicrystals,* New York: Wiley.

RADAR TECHNOLOGY

Josh T. Nessmith and Robert N. Trebits, *Georgia Tech Research Institute, Georgia Institute of Technology, Atlanta, Georgia, U.S.A.*

Introduction 399
1. **History** 400
2. **Physical and Technical Principles of Radar** 401
2.1 Radar Equation 401
2.2 Radar Block Diagram 402
2.3 Resolution, Detection, and Estimation 404
2.3.1 Resolution 404
2.3.2 Detection 407
2.3.3 Estimation 408
2.4 Radar Cross Section 409
2.4.1 Targets 409
2.4.2 Clutter 410
2.5 Propagation Theory 411
2.5.1 Attenuation 411
2.5.2 Multipath Interference 411
2.5.3 Refraction 412
2.6 Electromagnetic Interference 413
2.6.1 Natural Interference 413
2.6.2 Electronic Countermeasures 413
3. **Radar Design** 413
3.1 Radar Architecture 413
3.1.1 Search 414
3.1.2 Track 415
3.2 Transmitters 416
3.3 Antennas 416
3.4 Receivers and Exciters 417
3.5 Signal Processors 417
3.6 Data Processors 418
3.7 Wave Forms 418
3.7.1 The Ambiguity Function 418
3.7.2 The Ambiguity Diagram 419
3.7.3 Wave-Form Selection 420
3.8 Radar Measurements 420
4. **Applications** 422
4.1 Search and Surveillance Radars 422
4.2 Tracking Radars 423
4.3 Phased-Array Radars 423
4.4 Weather Radars 424
4.5 Ground-Probing Radars 424
4.6 Imaging Radars 424
4.6.1 Synthetic-Aperture Radars 424
4.6.2 Inverse Synthetic-Aperture Radars 425
4.7 Ground Vehicular Radar 425
4.7.1 Police Radars 425
4.7.2 Collision-Avoidance Radars 425
4.8 Missile Radars 426
Glossary 426
Works Cited 427
Further Reading 427

INTRODUCTION

Radar is an active electromagnetic sensor that operates at a wavelength within the 1-mm to 100-m range to provide detection and measurements of the location, radial velocity, size, and orientation of one or more objects. Objects or targets of interest include aircraft, missiles, satellites, space debris, ground vehicles, weather, planetary surfaces, humans, and animals. Operating nearly independently of the measurement environment, the radar is the source of energy for transmission and the receiver of energy reflected from scatterers, including the objects of interest and associated objects. Radar utilizes either pulsed wave forms or continuous wave (cw) and operates at average powers between a few milliwatts and several megawatts, with sensitivities on the order of picowatts. The radar achieves its maximum probability of detection in the presence of noise by processing the received return signal through a matched filter. In order to achieve high–measurement-accuracy estimates of target location, radial velocity, orientation, or radar cross section, the radar must isolate or resolve the desired object or target from other objects. To accomplish this resolution, the radar utilizes directive antennas to achieve angle resolution, wave-form bandwidth to achieve range resolution, and

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

observation time to achieve Doppler-shift resolution. Linear interpolations within the resolution cell boundaries provide measurement estimates of the target in position and radial velocity with respect to the resolution cell that are an order of magnitude improvement over the resolution capability, depending on the signal-to-noise ratio. Algorithms that use the measured characteristics of targets can provide highly useful radar imagery, particularly in terms of planetary surface mapping.

The original basic function of radar was to provide target detection, location, and velocity. Radar applications now include land, sea, and air vehicular control; space object and debris tracking; weapon-systems fire control; remote sensing of the Earth and other planets; and security surveillance. Research has also been directed to the phase, amplitude, and polarization measurements involved in target classification and identification. Radar has been highly successful in its roles of detection, metric measurements, and remote sensing, but less successful for object classification and identification.

The function of detection is accomplished by the controlled transmission of electromagnetic (EM) energy and reception of target-reflected energy that exceeds pre-established thresholds. These thresholds may be applied to amplitude, phase, range, angle, Doppler frequency shift, or polarization, or combinations thereof. The thresholds are set to reduce the number of false detections caused by noise, other targets, or background reflectors (clutter). False detections represent wasted expenditure of radar time and energy and overload of operator and data-processing resources. Therefore, false detections may limit the number of targets on which a radar can perform measurements.

Target angular location referenced to the radar is accomplished through the directivity of the antenna. Target radial velocity may be determined by the Doppler shift of the EM wave. Target range may be measured through the time difference between the transmission and reception of a modulated EM signal. Amplitude measurements of the EM signal give a measure of the target radar cross section (RCS). Sequential measurements of amplitude and phase, the latter with respect to the transmitted signal, can provide estimates of target size, orientation, and radial velocity. Polarization measurements, in combination with amplitude and phase measurements, can provide data for target classification and identification.

There is no single figure of merit for radar performance, because radars are used in so many different applications. The hydrometeor clutter that may limit an air-route surveillance radar in its search for aircraft targets is the target of interest for weather radar. The resolution and measurement capabilities of a radar determine its performance in its particular application. The resolution capability indicates the degree to which the radar measurements of the desired target are independent of other targets and clutter. Its metric measurement capability may be defined in terms of range, angle, and Doppler-shift accuracy for functional applications, or in terms of amplitude, phase, and polarization accuracy for phenomenology measurements.

This article begins with a brief history of the development of radar in Sec. 1. The radar equation is introduced in Sec. 2 as the vehicle through which a systematic approach to the theories used in radar can be developed, including resolution, radar cross section, detection, measurement, and electromagnetic interference. Radar system design is considered in Sec. 3, which includes discussions of each of the subsystems and the technology involved. A few of the broader applications of radar are discussed in Sec. 4.

1. HISTORY

In 1886, Hertz sought to verify Maxwell's electromagnetic theories, utilizing hardware components that are functionally identical with the radars of today. With a spark-gap transmitter operating at microwave frequencies, directive antennas, a target reflector, and a spark-gap detector, he experimentally demonstrated that electromagnetic waves would be scattered by metallic and dielectric objects in the propagation path (Hertz, 1900). Tesla recognized in 1889 that electromagnetic waves had the capability to provide location and through their use "may determine the relative position or course of a moving object, such as a vessel at sea, the distance traveled by same, or its speed."

In 1904, Christian Hülsmeyer of Germany, in his English patent No. 13 170, stated, "My invention is based upon the property of electric waves of being reflected back towards their source on meeting a metallic body, and will be readily understood by imagining a transmitting and a receiving station such as indicated placed side by side at the same point and so arranged that waves projected from the transmitter can only actuate the receiver by being reflected from some metallic body, which, at sea, would presumably be another ship" (Hülsmeyer, 1904).

Though Hülsmeyer experimentally demonstrated the concept, the maximum detection range obtained was less than one mile. Since one could see further than that, it was concluded that his system had little practical value. In 1922, however, Marconi took up the subject and stated in similar language to Hülsmeyer, "It seems to me that it should be possible to design apparatus by means of which a ship could radiate or project a divergent beam of these rays in any desired direction, which rays, if coming across a metallic object, such as another steamer or ship, would be reflected back to a receiver screened from the local transmitter on the sending ship, and thereby immediately reveal the presence and bearing of the other ship in fog or thick weather" (Marconi, 1922).

During that and the next decade, France, Germany, Great Britain, Holland, Italy, Japan, Russia, and the United States proceeded independently in the development of systems for radio location. Though the initial effort proceeded at a modest pace, the specter of war led to increased emphasis during the latter part of the 1930s. The direction of these later efforts and those during World War II was toward specific applications, including initial warning of aircraft attacks, vectoring of interceptor aircraft, and gunlaying. Airborne radar systems were developed for interceptor aircraft and for ground mapping. Ground-based radar systems were developed for landing aircraft in inclement weather (Swords, 1986).

It was in 1940 that the code name "radar" as an acronym for "Radio Detection And Ranging" was suggested by officers in the United States Navy and became widely accepted. The intense research, development, and production effort put forth in World War II produced results that otherwise might have taken a number of decades. It established a base of theory, technology, and applications that has stood up remarkably well over the succeeding half-century. Radar use in the defense of Great Britain and its use in naval battles established its importance. Radar landing systems for aircraft enabled the breaking of the Berlin blockade in the early 1950s (Guerlac, 1987).

Military development was the primary driver of radar technology and still is today. Commercial radar has been a key factor in air traffic control and in weather observations and prediction. Radar speed-detection systems have provided a degree of vehicle control on the highways. Radar astronomy was used for generating planetary imagery a number of years before space platforms carrying on-board radar provided mapping capabilities of distant planets and satellites. Maps of the Earth's surface from radars aboard aircraft and orbiting satellites provide daily information of environmental conditions on Earth.

2. PHYSICAL AND TECHNICAL PRINCIPLES OF RADAR

Radar involves a number of technologies based on detection, scattering, measurement, and propagation theories as well as the instrumentation for antennas, transmitters, signal processing, and data processing. The control of the instrumentation is another subject, as is the processing of the reflected signal and the data resulting from the metric and phenomenological measurements with the radar. The *radar equation* is used to discuss radar principles in a systematic manner, and each element of the equation must be addressed to understand radar theory.

2.1 Radar Equation

The basic analytical tool utilized to estimate the detection performance of a particular radar or radar design is the equation

$$\frac{S}{N} = \frac{P_t \tau \eta_t A_t \eta_r A_r \sigma}{4\pi R^4 \lambda^2 k T_s L_t L_r L_p L_{sp} L_o}, \qquad (1)$$

where S = signal power, N = noise power, P_t = peak power of transmitter, τ = transmitted pulse length, η_t = transmitting-aperture illumination efficiency, A_t = transmitting-aperture area (scan-angle dependent in phased-array radars), η_r = receiving-aperture illumination efficiency, A_r = receiving-aperture area (scan-angle dependent in phased-array radars), σ = radar cross section (RCS) of target (the mean or median of a statistical distribution describing target size), R = range of target, λ = wavelength of operation, k = Boltzmann's constant (1.38×10^{-23} J K^{-1}), T_s = system noise temperature (includes Ohmic receiving losses), L_t = Ohmic and non-Ohmic transmission losses between transmitter and antenna face, L_r = non-Ohmic receiving losses between antenna face and receiver including mismatches in impedances, L_p = propagation losses—multipath interference and attenuation (two-way), L_{sp} = signal-processing losses, and L_o = other losses.

The single-pulse signal-to-noise (*S*/*N*) ratio expressed in this equation is the value at the input to the detector. The major component of the noise represented by kT_s is noise internally generated by circuit elements in the receiving channel of the radar. The detrimental effects of all noise components on target detection and tracking are expressed by the inclusion of temperature components [T_a representing the external noise received by the antenna, T_r from the radio frequency (rf) plumbing, and T_e from the receiver] in the equation for system input temperature,

$$T_s = T_a + T_r + L_r T_e = T_a + T_p(L_r - 1) + L_r T_o(F_n - 1), \qquad (2)$$

where T_p is the physical temperature of the rf components, L_r is the loss in the rf components between the receiver and the antenna face, T_o is the reference temperature, 290 K, and F_n is the receiver noise figure.

There are a number of variants for the radar equation that take into account the integration of a number of single pulses that take place before detection and other losses including beam shape, i.e., the target off-center of the radar beam, or the receiving channel filters not being an optimum match to the received signal. Other variants of Eq. (1) can also be used to estimate radar performance in the presence of clutter [signal-to-clutter (*S*/*C*) ratio] and man-made interference, such as jamming [signal-to-jamming (*S*/*J*) ratio], if they rather than noise are the limiting factors (Nathanson, 1991). In general, the limitation of radar performance in clutter is a function of the noise in the radar transmitted signal, the phase stability (pulse to pulse) or coherence of the radar, the propagation path, and the clutter. The limitation in external jamming interference into the antenna main beam, as opposed to antenna sidelobes, is the energy available from the radar transmitter for target illumination.

2.2 Radar Block Diagram

A simple block diagram of a monostatic radar system, which uses a common antenna for transmission and reception, is shown in Fig. 1. Most radars built today follow this basic configuration whether they are simple or complex systems. Bistatic systems or multistatic systems that have separate transmitting and receiving systems are used where isolation between the transmitting and receiving antennas is required to reduce electromagnetic coupling or to perform interferometer measurements. Monostatic or bistatic systems may operate with either pulsed or continuous electromagnetic (EM) waves. The angular location of an object is achieved through the directivity of the antenna, whereas the object's range is determined through the time difference between transmission and reception of a modulated segment of the electromagnetic wave (pulsed or cw). Radial velocity of the object is determined by the Doppler shift of the EM wave or the differentiation of range with respect to time.

The control subsystem directs the antenna toward the desired position and then initiates the generation of a wave form of a single pulse or multiple pulses by an exciter. This wave form modulates the output of the transmitter. The duplexer directs the transmitter output energy to the antenna, isolating the high power from the sensitive circuits of the receiver. After reflection from a target, the EM wave is intercepted by the antenna and passed through the duplexer to the receiver. The duplexer may transfer megawatts of rf energy from the transmitter to the antenna, but it transfers only picowatts from the antenna to the receiver, thus

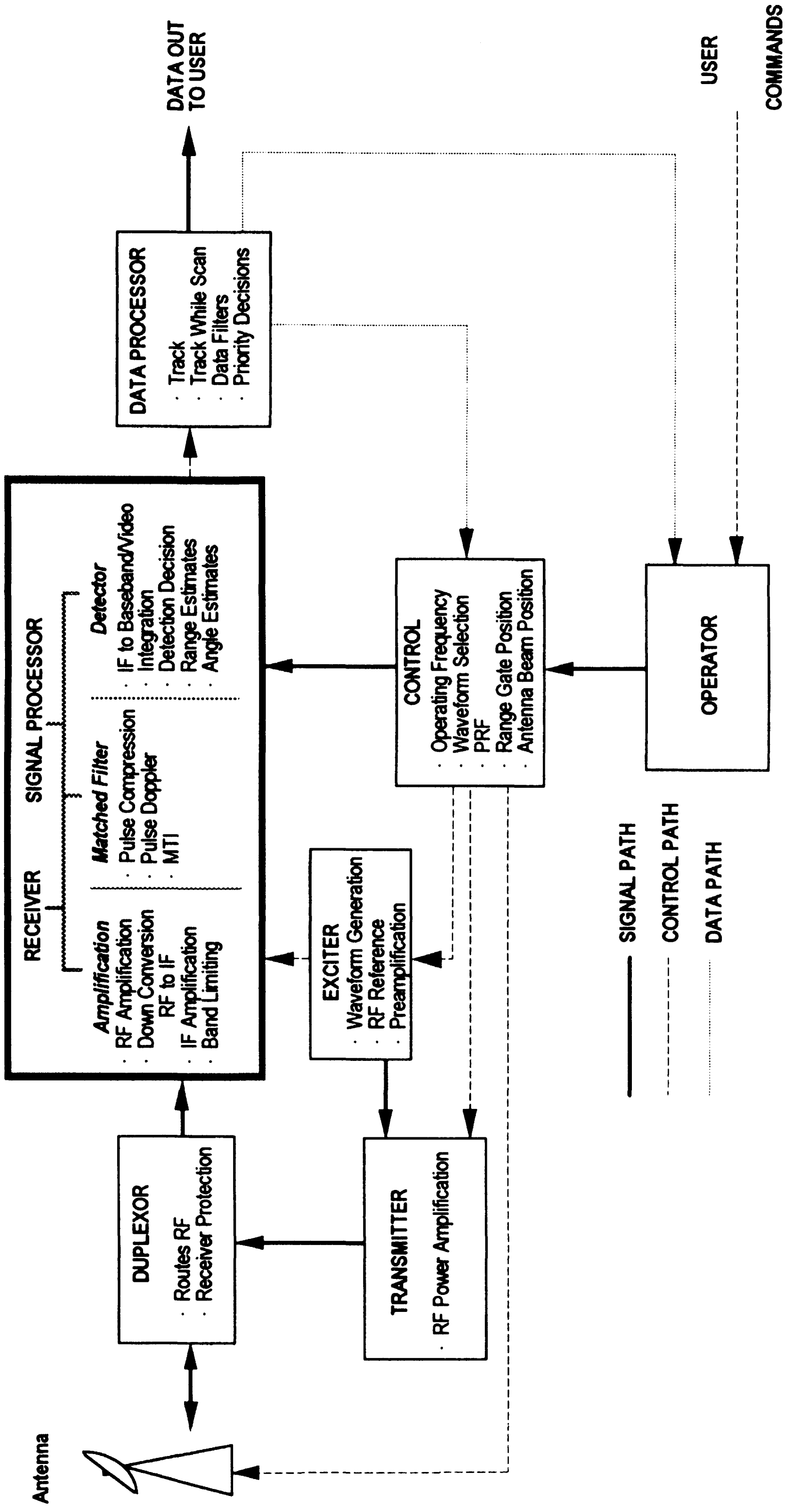
Antenna
DUPLEXOR
· Routes RF
· Receiver Protection
TRANSMITTER
· RF Power Amplification
RECEIVER
SIGNAL PROCESSOR
Amplification
· RF Amplification
· Down Conversion RF to IF
· IF Amplification
· Band Limiting
Matched Filter
· Pulse Compression
· Pulse Doppler
· MTI
Detector
· IF to Baseband/Video
· Integration
· Detection Decision
· Range Estimates
· Angle Estimates
EXCITER
· Waveform Generation
· RF Reference
· Preamplification
CONTROL
· Operating Frequency
· Waveform Selection
· PRF
· Range Gate Position
· Antenna Beam Position
OPERATOR
DATA PROCESSOR
· Track
· Track While Scan
· Data Filters
· Priority Decisions
DATA OUT TO USER
USER COMMANDS
SIGNAL PATH
CONTROL PATH
DATA PATH

FIG. 1. Radar block diagram.

protecting the receiver from burnout. The receiver provides the initial amplification of the signal return from a scatterer or target, a down-conversion from rf to a lower intermediate frequency (IF) to obtain high amplification, and bandpass filtering to isolate the received signal from rf interference. The front end (input) components of the receiver are a major source of the noise in the system, and much of the design effort goes into minimization of that noise.

The receiver or the signal processor may have a matched filter, i.e., a filter whose transfer characteristic is the Fourier transform of the complex conjugate of the transmitted wave form, to improve the S/N ratio. If there are n pulses in the waveform, then the matched filter performs a coherent integration that improves the S/N ratio over that of a single pulse by the factor n.

The signal processor, as the name implies, operates on the signal to perform one or more functions including integration, thresholding, detection decisions, and estimations of the position and Doppler shift of the target with respect to a reference. The references for estimates are the angular position of the antenna, the initial time of transmission of the signal from the antenna, and the rf carrier phase or frequency. The estimates are summed with the reference values to give absolute values of position and velocity. The functions of moving-target indication (MTI) and pulsed Doppler processing to derive Doppler shift to resolve the target from the clutter and to improve S/N ratio are performed in the receiver or signal processor. The signal processor may be either analog or digital, or a combination of both. Many modern radars perform an early transition from analog to digital format in order to better preserve phase and amplitude information.

After integration by the pulse-compression, MTI, or pulsed-Doppler filters, the IF signal is rectified by a detector to obtain the envelope (which preserves amplitude only), phase detected by a zero-crossing detector (which preserves phase only), or passed through a coherent detector (which preserves phase and amplitude). Earlier radars utilized the output of the detector to go directly in analog form to a cathode-ray tube display for the operator to make a detection decision, and a number of radars still do. But with the advent of the computer, more radars are digitizing earlier in the receiving channel. The output to both the user and the operator may be in an analog or digital format, or a combination. The data processor first converts the signal information about target presence, location, and velocity into a form suitable for display to an operator or for information to a user, such as an air traffic control or a weapon system. If the data are to be used external to the radar, they are often passed along in unfiltered format rather than filtered by further processing in the data processor. The data processor may be used to establish not only current position but also a trajectory on multiple targets (>200) and then to pass that information to the user. The data processor is also used to develop information to control the radar. This would include the data for decisions as to the next set of wave-form transmissions or the next antenna beam position.

2.3 Resolution, Detection, and Estimation

Radar functions can be divided into three successive stages: resolution, detection, and estimation. Resolution means the extent to which the radar can distinguish between closely spaced objects and can perform measurements of one or more radar parameters, i.e., azimuth and elevation angles, range, or frequency (Doppler shift), for each object. Detection is the process of determining whether the signals corresponding to each resolution cell represent the presence of a target. Estimation is the process of determining a specific target metric characteristic, typically its coordinates in range, angle, or Doppler shift, or the phenomenological characteristics of amplitude, phase, or polarization.

2.3.1 Resolution Radar systems are sometimes characterized by their resolution characteristics. Resolution is the property of radar that enables the separation of the target of interest from other targets and clutter to reduce false alarms in the detection process and to reduce the contamination of estimates in metric or phenomenological measurements. Resolution may be accomplished in terms of the antenna beamwidth for spatial resolution in two orthogonal angles, in terms of signal bandwidth for time resolution in range, and in Doppler shift for radial velocity separation. Polarization resolution occurs naturally in most radars, since most antennas have only a single polarization orientation. Circular polarization and sequential linear orthogonal polarization may be utilized in special cases where phenomenology measurement or target identification is of interest.

The resolution of the system in the two spherical coordinates, such as elevation angle and azimuth angle, is normally determined by the 3-dB beamwidths of the antenna in those planes. The angular resolution determined by the antenna beamwidth is termed the *real aperture* resolution. For an aperture of extent d in the horizontal plane, the corresponding angular resolution δ_θ is approximately λ/d (where λ is the radar wavelength). The corresponding spatial resolution δ_{CR} (often termed *cross-range resolution*) at a range R is

$$\delta_{\mathrm{CR}} \approx \lambda R/d. \tag{3}$$

Similar results are defined for the elevation plane.

For typical radar parameters and scenarios, the cross-range resolution is often much poorer than the range resolution. While acceptable in many applications, this is undesirable in radar imaging or mapping. Finer cross-range resolution can be obtained with the technique of *synthetic aperture radar* (SAR). SAR uses motion of the radar antenna relative to the area being mapped to synthesize an antenna much larger than the real aperture. As shown in Fig. 2, the physical radar antenna of aperture size d is moved from point to point by the platform motion to create a virtual linear array antenna of size D_{SAR}. The physical antenna serves successively as each element of the larger phased array. The received signals at each transmission location are collected and combined coherently in the signal processor to process the data to obtain the same spatial resolution result that would be obtained by using an array antenna whose physical extent matched the total distance traveled, D_{SAR}. The cross-range resolution achieved by a SAR system is inversely proportional to D_{SAR} and is lower bounded by half the physical aperture size (Richards, 1988), i.e.,

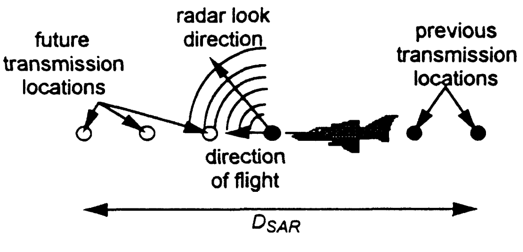
radar look direction
future transmission locations
previous transmission locations
direction of flight
D_{SAR}

FIG. 2. Virtual array-antenna generation by platform motion.

$$\delta_{\mathrm{CR,SAR}} \approx d/2. \tag{4}$$

The range resolution δ_R for a simple constant-frequency, constant-amplitude pulse of duration τ is approximately

$$\delta_{\mathrm{R}} \approx c\tau/2, \tag{5}$$

where c is the velocity of propagation. Equation (5) implies that higher resolution requires shorter pulses, but higher resolution can also be obtained by modulating the pulse in a way that increases its bandwidth and using appropriate matched filter signal processing. Phase modulation is used when it is undesirable to reduce transmitted energy by amplitude modulation of the pulse.

Matched filtering of the received phase-modulated pulses is referred to as *pulse compression* processing. The two most common types of phase modulation are *linear frequency modulation* (LFM, or *chirp*) and *biphase coding*. In the former, the instantaneous frequency is swept linearly during the pulse duration; this corresponds to a parabolic variation in phase as a function of time. In the latter, the instantaneous frequency is constant, but the absolute phase is abruptly switched by 180° a number of times within the pulse in a fixed code sequence. If the bandwidth of the modulation function is β, then δ_R will be approximately

$$\delta_{\mathrm{R}} \approx c/2\beta. \tag{6}$$

The corresponding time duration of the modulation function is approximately $1/\beta$, and this can be shorter than the actual transmission pulse length by a factor of 10 to 1000 in many systems. Bandwidths of 100 kHz up to 1 GHz are common and provide

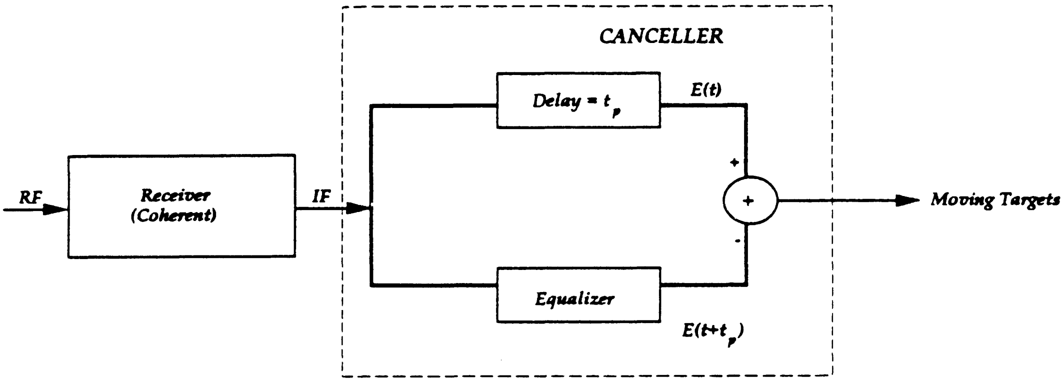
CANCELLER
Delay = t_p
E(t)
+
+
-
RF
Receiver (Coherent)
IF
Moving Targets
Equalizer
E(t+t_p)

FIG. 3. MTI block diagram.

range resolutions of 1.5 km to 0.15 m, respectively.

In nonimaging radars, Doppler-shift filtering may be applied to the received signals to resolve moving targets from stationary clutter or from one another. One form of Doppler filtering is termed *moving-target indication,* or MTI, filtering, and is accomplished by applying high-pass filters to a series of returns from the same spatial resolution cell. The simplest filter is the *two-pulse canceler* shown in Fig. 3 that subtracts successive pulses separated by the interpulse period t_{PRF}. dc signal components due to stationary clutter cancel, while Doppler-shifted components due to moving targets do not.

Pulsed Doppler processing (Fig. 4) passes the signal through a range gate set for a specific range and then sends it in parallel through a bank of bandpass filters each of which is tuned to a different value of Doppler frequency shift. Moving-target signals are passed by the bandpass filter centered nearest their actual Doppler shift. Pulsed Doppler processing is more complex than MTI filtering because it requires multiple filters, but provides more information by resolving multiple moving targets and providing estimates of their radial velocity. Attainable velocity resolution is determined by the observation period on the target (known as the observation time, t_o), which

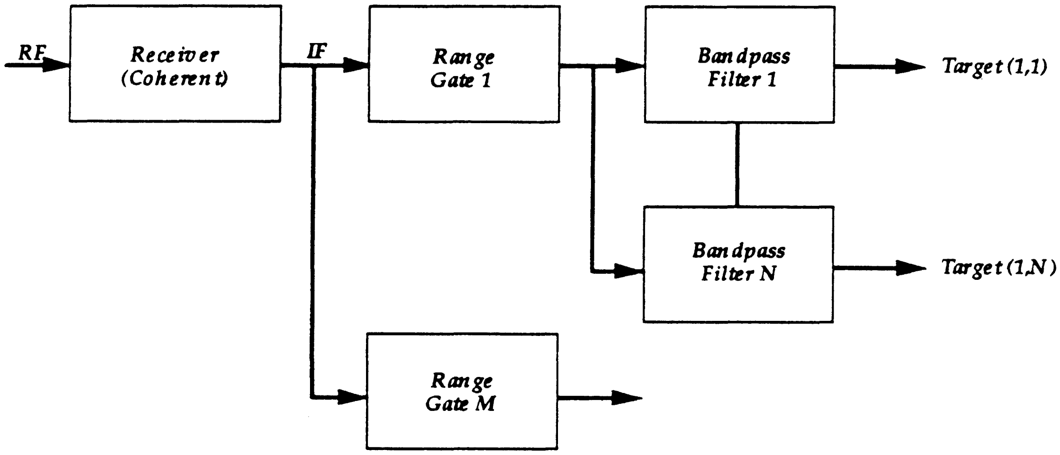
RF
Receiver (Coherent)
IF
Range Gate 1
Bandpass Filter 1
Target (1,1)
Bandpass Filter N
Target (1,N)
Range Gate M

FIG. 4. Pulse-Doppler block diagram.

sets the Doppler frequency resolution at approximately $1/t_o$. The corresponding velocity resolution is $\lambda/2t_o$.

2.3.2 Detection Detection is a term utilized in several ways in the radar literature. One connotation is the initial down-conversion of the radar return at the transmission frequency to a lower frequency (IF) for amplification in the receiver chain. The second is the conversion by a detector from IF to baseband, or video, with no carrier. The second conversion may or may not conserve phase or amplitude information. The third connotation, and the one used here, is that of a test threshold and the decision as to whether a signal from a target is present in the presence of noise, interference, or clutter.

Detection performance is optimized by maximizing the signal-to-noise (S/N) power ratio into the detector that converts from IF to baseband. This occurs when the frequency response $H(\omega)$ of the receiver constitutes a *matched filter* for the transmitted wave-form modulation, $H(\omega) = M^*(\omega)$, where $M^*(\omega)$ is the complex conjugate of the Fourier transform of the transmitted wave-form modulation $m(t)$. When a matched filter is used, the resolution and side-lobe characteristics are determined by the autocorrelation of the wave form and not the wave form itself.

If the matched filter is for a train of pulses, such as a MTI or pulsed Doppler filter, predetector (coherent) integration improves the S/N ratio. Postdetector integration prior to the threshold may also be used to improve the S/N ratio further. The latter integration may be coherent or noncoherent, based on whether both amplitude and phase are preserved by the detector. Coherent integration is generally the more efficient, with the S/N ratio improving by the number of pulses integrated, whereas noncoherent integration improvement is proportional to a factor within the range $n^{0.5}$ to $n^{0.8}$. The pulse repetition frequency (PRF) and target dynamics may also limit the number of pulses that can be integrated effectively.

After translation of the intermediate-frequency (IF) signal to video or baseband, a detection decision is made. The criterion for detection or the test to determine whether a target is present in the presence of noise or clutter can be as simple as determining whether an amplitude or phase threshold is crossed by the signal after it is processed by the detector and integrator. Since there is a probability that a noise signal alone could cross the threshold, the threshold is set high enough to limit the number of noise crossings in a specified interval. If the threshold is set too high, there will be many missed targets. If the threshold is set too low, noise crossings can overload operators or automatic detection systems. There is a considerable body of literature in the radar field on detection theory (Nathanson, 1991).

An amplitude threshold for a single pulse repetition interval can be selected to maintain a fixed probability of false alarm P_{fa} from thermal noise, i.e.,

$$P_{\text{fa}} = \exp(-V_T^2/2\varphi_0), \tag{7}$$

where V_T is the threshold setting, and φ_0 is the mean square value of the noise voltage into the detector in the receiver (Skolnik, 1980). Typical values of P_{fa} range from 10^{-8} to 10^{-2}, depending on the system type and interference statistics. The threshold is a function of the interference power and statistics, and these are rarely known *a priori* with sufficient accuracy. For this reason, the threshold used to test a given resolution cell may be estimated from the signal power values in adjoining cells under the assumption that they represent interference only. This idea is illustrated in Fig. 5. A measured value of received power is associated with each range resolution cell. The power values measured in a few range cells to either side of the test cell (representing echoes received from ranges nearer and farther than the test cell) are averaged to estimate the mean interference level. A threshold is then set based on this estimated interference level, the assumed probability density function (pdf) of the interference power, and the desired P_{fa}. If the power in the test cell exceeds the threshold, a target is declared. Though this constant false alarm rate (CFAR) threshold scheme is used in some radar systems, there are other thresholding techniques.

The probability of detection P_d for a detection threshold of V_T is

$$P_d = \int_{V_T}^{\infty} P_s(R_j)\,dR_j, \tag{8}$$

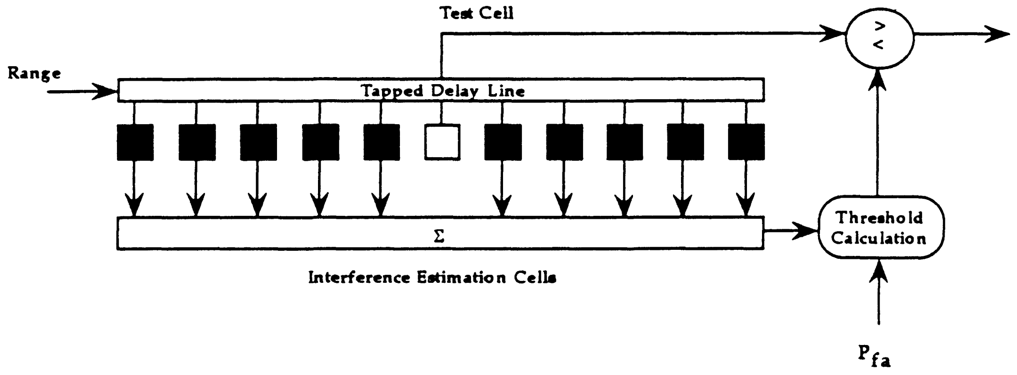

FIG. 5. Threshold for constant false-alarm rate (CFAR).

where $P_s(R_j)$ is the joint probability density function of signal and noise. The trade-off between the false alarm probability of Eq. (7) and detection probability of Eq. (8) is accomplished through the threshold setting.

2.3.3 Estimation Once a target is detected, the data in the resolution cell may be interpolated to obtain better estimates of the target position or radial velocity within the resolution cell. If a target peak is observed in resolution cell n, then a quadratic curve is fitted through the data in cells $n - 1$, n, and $n + 1$. The fractional cell location and amplitude of the peak of the quadratic can be found in terms of the three data samples, yielding an estimate of range or Doppler shift with a precision of a fraction of the resolution cell size. Though the angle can be estimated by use of such a technique, a monopulse antenna is often used to obtain high-precision angular coordinate measurements.

Monopulse is a radar technique in which the angular location of a source or target is obtained by comparison of signals received in two or more simultaneous beams that are identical but are pointed at slightly different directions and have a common volume containing the target. The simultaneity of the beams enables a single-pulse, i.e., monopulse, estimate of angle as an alternative to beam switching or conical scanning that require two or more pulses. The difference (Δ) signal is generated by taking the difference of the signals received in the two beams, and the sum (Σ) signal is generated by adding the signals received in the two beams. The plane defining the contour for equal amplitudes between the two beams is the boresight direction, or axis, where the difference signal approaches zero and the sum signal is maximized. The magnitude and phase of the ratio Δ/Σ off axis is an indicator of the angular position of the target with respect to the boresight axis. The normalization of the difference signal by the sum signal makes the estimate independent of target fluctuations in size. The detector of the signal is often a dot or cross product, with the magnitude V_e of the voltage out being

$$V_e = \mathrm{Re}(\mathbf{\Delta} \cdot \mathbf{\Sigma}/|\Sigma|^2). \qquad (9)$$

Four overlapping beams arranged in a quadrant can provide the range and the two orthogonal angle estimates simultaneously, with one transmitted pulse. Doppler shift and target size may also be obtained. The primary objective of monopulse is to minimize angle error due to target-signal amplitude fluctuations between successive pulses, which occurs when time-sequential or conical scanning beams were used.

The standard deviation σ_e of the estimated target position in a range, angle, or Doppler shift depends on the resolution δ in the same coordinate and the S/N ratio, and is given by

$$\sigma_e \approx k \frac{\delta}{\sqrt{S/N}}, \qquad (10)$$

where the proportionality constant k is typically in the range of 0.5 to 2.0. For a typical $S/N = 100$ (or 20 dB) for a detected target, the measurement error is approximately one-fifth to one-twentieth of a resolution cell.

2.4 Radar Cross Section

2.4.1 Targets The radar cross section (RCS) parameter σ in Eq. (1) is a transfer function that quantitatively relates the transmitted power incident on the target to that backscattered by the target into the radar antenna. Radar cross section is a characteristic of the target alone (at a particular radar frequency and polarization) at a particular target aspect.

The fact that radar cross section is a characteristic of a target and not its environment implies that the target is in the *far field* of the radar antenna and that the electromagnetic wave fronts incident on the target are, for all practical purposes, planar. This fact is implicit in the theoretical definition of radar cross section,

$$\sigma = \lim_{R\to\infty} 4\pi R^2 \, |E_r^2/E_i^2|, \tag{11}$$

where E_r = electric field magnitude at the radar receiver, E_i = electric field magnitude incident at the target.

The components of the received electric field vector $\mathbf{E}_r$ may be related to the components of the transmitted electric field vector $\mathbf{E}_t$ by the equation

$$\begin{bmatrix} E_h^r \\ E_v^r \end{bmatrix} = \begin{bmatrix} a_{hh} & a_{vh} \\ a_{hv} & a_{vv} \end{bmatrix} \begin{bmatrix} E_h^t \\ E_v^t \end{bmatrix}, \tag{12}$$

where the first subscript of the components of the polarization matrix denotes the transmitted polarization, and the second subscript denotes the received polarization. The symbols h and v are used, for example, for horizontal and vertical polarization in this equation. The matrix

$$\begin{bmatrix} a_{hh} & a_{vh} \\ a_{hv} & a_{vv} \end{bmatrix}$$

is called the full polarization scattering matrix and contains all of the information reflected from a target for a particular radar frequency, target orientation, and polarization vector basis set. The parameters a_{xy} are complex valued and can be related to the appropriate RCS parameters by the expression

$$a_{xy} = \sqrt{\sigma_{xy}} \exp(j\varphi_{xy}). \tag{13}$$

The polarization, i.e., the locus of the tip of the electric field vector over time, of the radar signal, as an electromagnetic wave, may be linear, circular, or in its most general form, elliptical. In the radar literature, circular polarization is defined as either right-hand or left-hand circular looking *in the direction* of energy propagation. The physics convention looks *into* the direction of energy propagation. Many radars, however, use linear polarization for transmission and reception, with circular polarization and sequential linear polarization being used for phenomenology measurements.

There are three very distinct regions for the general wavelength dependence of the RCS of an object, of which the perfectly conducting sphere may be considered representative. At low frequencies, where the radar wavelength is much greater than the sphere's circumference (the Rayleigh region, where $2\pi a/\lambda \ll 1$), the sphere's RCS varies proportionately to $(ka)^4$, where k is the wave number. Likewise, where the radar wavelength is much smaller than the sphere's circumference (the optical region, where $2\pi a/\lambda \gg 10$), the sphere's RCS is essentially wavelength independent and asymptotically approaches a value of πa^2, its geometric cross section. In between these two frequency regions (the Mie region, where $1 \ll 2\pi a/\lambda \ll 10$), the sphere's RCS oscillates about its geometric cross-section value, as a function of frequency. This oscillatory character is the result of the constructive or destructive interference of the specular radar reflection from the "front" of the sphere with the radar wave that "creeps" around the shadowed side of the sphere.

In reality, the radar cross section of an object must be described statistically, because it varies with target aspect angle. The composite radar cross section is the coherent summation of the cross sections of all the scattering centers of the target. Relative motion among the scatterers that results in differential displacement of a factor of a wavelength along the radar line of sight can result in significant radar cross-section variation. The classic statistical models that are used to define target radar cross sections have been based on

1. the reflectivity characteristics of the scattering centers on the target that is assumed to be completely illuminated by the radar, and
2. the temporal decorrelation characteristics.

Swerling (1960) defined four target models, which have since been referred to as Swerling Cases 1–4. Case 1 assumes a Rayleigh distribution function for target backscatter voltage and a slow temporal dependency, i.e., significant backscatter changes only on a scan-to-scan basis. The target radar cross section for Case 1 will be characterized by the probability density function $p(\sigma)$ defined as

$$p(\sigma) = (1/\bar{\sigma})e^{-\sigma/\bar{\sigma}}, \tag{14}$$

where $\bar{\sigma}$ is the average radar cross section. Case 2 assumes the same distribution function but a rapid signal decorrelation, i.e., significant backscatter changes on a pulse-to-pulse basis. Case 3 assumes the probability density function

$$p(\sigma) = (4\sigma/\bar{\sigma}^2)e^{-2\sigma/\bar{\sigma}} \tag{15}$$

and a slow temporal dependency. Case 4 assumes the latter distribution function but a rapid signal decorrelation. Occasionally, one also refers to a Swerling 0 target model, for the case of the nonfluctuating, target backscatter signal.

The probability density function of Eq. (14) corresponds to a target having numerous, small, statistically independent scattering centers. Many aircraft and large targets present this type of probability density function for radar cross-section characteristics. The probability density function of Eq. (15) corresponds to a target having one dominant scatterer and many smaller, independent scatterers. Small, streamlined aircraft can have distribution functions of this nature at the lower radar frequencies. The Swerling target cases corresponding to rapid signal decorrelation can be produced by radar systems that change transmitted frequency on a pulse-to-pulse basis. The four Swerling models are special cases of the chi-square distribution. The Rice and the log normal are also two density functions often used to describe the radar cross section of targets. These probability density functions must be combined with the probability density function of the type of detector used in order to determine the probability of detection.

2.4.2 Clutter Land- and sea-surface backscatter is characterized by σ^0, the radar reflectivity per unit area illuminated by the radar, or

$$\sigma^0 = \sigma/A, \tag{16}$$

where σ is the radar cross section of the illuminated area, and A is the illuminated area of the sea surface or ground. The illuminated area A should be reduced by a beam-shape factor of 1.33, due to a Gaussian rather than a rectangular beam shape. By convention, σ^0 is shown in superscript form and is usually specified as *sigma zero* or *sigma naught.* The parameter σ^0 is a complex function of

1. the radar's frequency and polarization;
2. the surface's roughness, tilt variation, and content; and
3. the geometry between the radar line of sight and the local surface normal.

For example, ground reflectivity depends on the soil makeup, water content, roughness, tilt variation, snow cover, and foliage cover, among other factors. Sea reflectivity will depend on the average wave height, wind speed and direction, and wave direction.

The shape of the instantaneously illuminated area on the surface can be a function of the incidence angle to the surface, the radar pulse length, and the radar antenna elevation and azimuthal beamwidths. The illuminated area is either elliptical or rectangular, depending on whether or not the vertical projection of the pulse length onto the surface is greater than the footprint range extent of the elevation beamwidth on the surface. At a high incidence angle that might be characteristic of an air-to-ground geometry, the reflecting surface area is an ellipse, for pulse-length projections that are larger than the footprint range extent. At a lower incidence angle that might be characteristic of a ground-to-ground geometry, where the pulse-length projection is usually less than the footprint range extent, the reflecting surface is essentially rectangular, with a width determined by the azimuthal extent of the antenna beamwidth and a length determined by the vertical projection of the pulse length.

Volumetric backscatter, such as from rain, is conventionally quantified by the parameter η that represents the radar cross section per unit volume of the scattering material. The volume of the resolution cell is defined by antenna patterns in azimuth and in elevation and by the range resolution $c\tau/2$. If the radar antenna has a pencil beam pattern, the azimuth and elevation beamwidths, defined by the 3-dB points of the beam pattern, will determine the extent of the resolution cell in the vertical and horizontal dimensions. If the azimuth and elevation beamwidths are given by θ_{az} and θ_{el}, respectively, the resolution cell cylinder will have diameters of just $R\theta_{az}$ and $R\theta_{el}$, respectively, where R is the range to the resolution cell. The range extent of the resolution cell will be simply $c\tau/2$, and the volume V of the cylinder is

$$V = [R^2\theta_{el}\theta_{az}][c\tau/2]. \tag{17}$$

The value of the normalized backscatter parameter η is then just σ/V, where σ is the actual *measured* radar cross section of the clutter within the resolution volume. The illuminated volume should be reduced by a beam-shape factor of 1.77, due to a Gaussian rather than a rectangular beam shape.

2.5 Propagation Theory

2.5.1 Attenuation In the radar equation, the propagation loss is composed of two factors: the absorption attenuation of energy (transformation to heat) and the reradiation (scattering) of energy over the entire 4π-sr volume (forward scattering and backscattering). The sum of the two effects represents the total loss of signal energy and is termed *signal extinction.* The signal extinction is taken from the optical and infrared research community and is equivalent to the radar engineering term *signal attenuation,* which does not differentiate by signal loss process.

The nomenclature of the radar engineering community will be used here, where the appropriate terminology is signal attenuation, and the conventional attenuation symbol is α (measured in dB/km). The component of the loss L_p in Eq. (1) due to both absorption and scattering processes is designated L_{att} and is quantified by the one-way signal attenuation α', expressed in nepers per meter (Np/m), assumed to be constant over the entire range extent. When the attenuation is not uniform over the entire range extent, the total, two-way attenuation loss L_{att} is found by integration over the range extent of the lossy medium:

$$L_{att} = \exp\left[-2\int_{\text{near}}^{\text{far}} \alpha'(R)dR\right]. \tag{18}$$

The propagation loss is conventionally depicted in logarithmic form, and the median loss of $10\log L_{att}$, in decibel (dB) units, is expressed as

$$\begin{aligned} 10\log L_{att} &= 10|\log e^{-2R\alpha'}| \\ &= 4.343|2R\alpha'| = 2R\alpha/1000, \end{aligned} \tag{19}$$

where the attenuation factor α has units of decibels per kilometer.

The most often encountered source of propagation loss is rain. Rain attenuation coefficient data have been empirically modeled in both the microwave and the millimeter wave regions as a two-parameter monomial

$$\alpha = AR^B, \tag{20}$$

where α is in dB/km, R is the rainfall rate in mm/h, and A and B are taken from Table 1.

2.5.2 Multipath Interference When the radar antenna beam is pointed toward a target at a low incidence angle to the surface, the radar signal may follow four different two-way paths, as shown in Fig. 6. The energy may follow a direct-direct, direct-indirect, indirect-direct, or indirect-indirect path between the radar antenna, the target, and the radar antenna again. The constructive or destructive interference among these signals at the receiving radar antenna gives rise to a lobing effect in received target backscatter signal, shown in Fig. 7. This form of the pattern is developed when the reflection coefficient is unity and the elevation beamwidth of the antenna has equal gains along the direct and indirect paths.

For a target altitude h_t much greater than

Table 1. Predicted rainfall attenuation coefficients as a function of rainfall rate (Adapted from Currie, Dyer, and Hayes, 1975).

	α(dB/km) = AR^B			
	10 GHz	35 GHz	70 GHz	95 GHz
R (mm/h)	$A = 0.009$ $B = 1.16$	$A = 0.273$ $B = 0.985$	$A = 0.634$ $B = 0.868$	$A = 1.6$ $B = 0.64$
5	0.05	1.33	2.56	4.48
10	0.13	2.64	4.68	5.98
15	0.21	3.93	6.65	9.05
20	0.30	5.22	8.54	10.88
25	0.39	6.50	10.36	12.55

radar altitude h_a, the first interference lobe will occur at an elevation angle of approximately $\lambda/4h_a$ radians. This loss L_m component of L_p is given by

$$L_m = \frac{1}{16}\csc^4\left(\frac{2\pi h_a h_t}{\lambda R}\right). \tag{21}$$

2.5.3 Refraction Index-of-refraction nonhomogenieties within the propagation medium may significantly affect the path of an electromagnetic wave. The radar beam may bend such that an incorrect (high) elevation angle and an incorrect (long) range to the target are indicated. The primary physical source for this effect is the exponential variation of the refractive index n of the atmosphere with altitude. It is convenient for radar applications to define the refractivity N as

$$N = (n - 1)10^6. \tag{22}$$

The refractivity as a function of radius r from Earth center is

$$N = N_s \exp\left(-\frac{\Delta N}{N_s}(r - r_s)\right), \tag{23}$$

where N_s is the refractivity at the Earth's surface (radius $r = r_s$).

The limiting angle $\Delta\Theta$ through which the radar beam is bent, at elevation angles $\Theta > \pi/12$ and for targets well beyond the atmosphere, may be approximated by

$$\Delta\Theta = N_s \cot\Theta \times 10^{-6}. \tag{24}$$

The corresponding limiting range error may be approximated by

$$\Delta R = 0.007\, N_s \csc\Theta. \tag{25}$$

This refractive effect will result in incorrect tracking of an airborne target, both in elevation angle Θ and in range R (Barton, 1988).

Superrefraction will occur when there exists a rapid decrease in refractive index with altitude, such as might be caused by a temperature inversion or a high humidity gradient. The radar beam becomes trapped in this *duct* in a manner not unlike an electromagnetic wave confined within a waveguide structure.

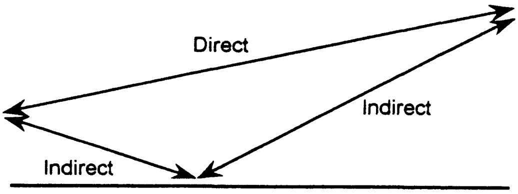

FIG. 6. Direct-direct, direct-indirect, indirect-direct, and indirect-indirect signal paths.

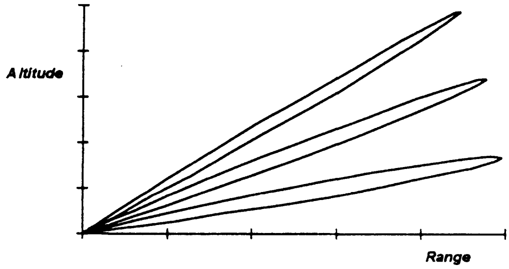

FIG. 7. Multipath interference lobing structure.

2.6 Electromagnetic Interference

2.6.1 Natural Interference Sources of naturally occurring signals that interfere with proper radar operation include

1. those external to the radar that enter through the antenna or other components and
2. noise signals generated by the radar receiver itself, as discussed under system temperature.

Examples of the former include solar and galactic noise as well as that due to Earth surface temperature. These enter through both the main lobe and sidelobes of the receiving antenna.

2.6.2 Electronic Countermeasures The term *electronic countermeasures* refers to the intentional transmission of electromagnetic signals toward a victim radar to prevent the radar from detecting or tracking a target. The term *jamming* is usually applied to such intentional transmission. The ultimate degrading effect on the victim radar's detection capability may be simply one of a decreased target signal-to-interference ratio within the target cell of resolution, or it may be the complete overload of the radar signal processor through the creation of numerous targetlike signals.

Noise jamming is the most common form of electronic countermeasures. It denies target detection by simply raising the effective noise floor of the victim radar receiver. The quantitative effect is to raise the effective radar system temperature T_s of Eq. (2) by an amount T_j given by

$$T_j = \frac{P_j G_j A_r F_j^2}{4\pi k B_j R_j^2 L_j}, \tag{26}$$

where $P_j G_j$ is the effective radiated power of the jammer, F_j is the directivity of the radar aperture in the direction of the jammer, B_j is the bandwidth of the jammer spectrum within the bandwidth of the radar, R_j is the range to the jammer, and L_j is the one-way propagation loss between the jammer and the radar. The total receiver noise temperature is then the sum of its natural amount T_s and that of the jammer, T_j (Barton *et al.*, 1991).

Common uses of noise jammers include self-protection jamming, where the jamming platform seeks to deny its own detection by the radar, and escort jamming, where the jammer is on a separate platform from the targets whose detection is being degraded. Jamming, however, does provide a signal source that can be tracked in bearing using a *home-on-jam* mode.

A separate type of electronic countermeasures is *deception jamming,* wherein false, targetlike signals are transmitted to the victim radar. These false target signals attempt to capture the radar's range or angle tracking circuits, then suddenly cease. The radar then must attempt to reacquire the true target range and bearing.

3. RADAR DESIGN

3.1 Radar Architecture

Radar design architecture is first driven by the application requirements and then bounded by the technology available. The probability of detection, range and volume to be covered, time to perform the search, and measurement accuracy are direct requirements. Target characteristics, which are statistical, and the electromagnetic and physical environment set the stage in which the radar must perform. The major variables in the radar design are wavelength, transmitter power, and antenna size. The hardware and software technology available and their costs establish another boundary.

The wavelength, antenna area, and transmitter energy are usually the key variables in the design of a radar system. Generally, the search function will require most of the energy, and high-power sources are more readily available at the longer wavelengths. Shorter wavelengths give more precise measurements in angle for a given size aperture. If search and track functions are both performed by a single radar, such as a fire-control phased-array system, there must generally be a compromise in frequency selection. Depending on the application, however, any one of the variables or combinations may be bounded by political treaty, e.g., the energy bound of radiated power for antiballistic missile radars; by the antenna size bounds established by aircraft, spacecraft, or transport vehicles; or by the losses in propagation for the shorter wavelengths. For a new design, there are generally a large number of trade-offs; consequently, an iterative selection process is required. Performed in a systematic manner, the trade-off process can converge to a realistic design.

3.1.1 Search The search function is accomplished by having the radar antenna beam scan a solid angle of Ω_s. If a surface-based search radar with a vertical fan-shaped beam of azimuth width θ_a is being rotated mechanically in azimuth at a rate ω_a and is transmitting at a PRF of f_p, the number of pulses n_p that will strike a target is approximately $n_p\theta_a/\omega_a$ on each scan. Integration of those pulses during each scan by a matched filter will improve the S/N ratio by a factor of n_p, less a beam-shape factor loss due to the target being other than at the antenna beam maximum. If the integrated-signal S/N ratio of the first scan is inadequate initially to pass the threshold required for detection, successive scans provide a cumulative probability of detection, i.e., $P_{dc} = 1 - (1 - P_{d1}) \cdots (1 - P_{ds})$, where P_{ds} is the probability of detection on a single scan. Composite search performance is usually determined by cumulative detection.

Fire-control radars, which have a pencil beam used to perform both search and track, may use a bar scan for search. A full azimuth scan is made for the first bar. At the limit of the scan a step up or down in elevation is made by approximately the antenna beamwidth, and a second azimuth scan is made in the reverse direction of the first. This is continued for several bars to cover the solid angle required. When the solid-angle scan is completed, a repeat of the scan is made, but the bars are repositioned so that the peak of the beams covers the intersection of the beams between the bars on the first solid-angle scan. When the scan is mechanical, its period is comparatively long, i.e., 1–10 s. Electronically scanned phased-array radars that sweep faster may transmit only a single pulse per beamwidth in an azimuth scan, then step down a beamwidth for the next scan, transmitting and receiving where the upper and lower beam positions intersect to form a triangular lattice pattern. A second scan may be performed with the same triangular lattice pattern, but the peaks of the pattern are positioned at the intersection of the three beams of the first scan. Detection is also accomplished through a cumulative process.

Equation (1) may be rewritten to function as a design or evaluation tool for the search and track functions. The energy E_s required of the transmitter for a single-pass search of a solid angle Ω_s, with n_b the number of beams required, i.e., $n_b = \Omega_s/\Omega_t$, is given by

$$E_s = n_b P_t \tau \geq \Omega_s \frac{1}{\eta_r A_r} \frac{(S/N) 4\pi R^4 k T_s L_t L_r L_p L_{sp} L_o}{\sigma}, \tag{27}$$

where the transmitted beam has a solid angle of $\Omega_t = \lambda^2/\eta_t A_t$. The S/N is specified to provide the requisite single-pulse detection probability for the range R and the target RCS. The other values in the equation are functions of the radar implementation. For the condition where the received beam is the same dimension and same direction as the transmitted beam, i.e., $\eta_r A_r = \eta_t A_t$, the time to perform a single-pass search is

$$t_s \geq (\Omega_s/\Omega_t)(2R/c) \tag{28}$$

for nonoverlapping coverage. Where one may be limited in detection capability by the energy available in a single pulse, overlapping coverage may be utilized and detection probability increased through the cumulative process on either a single scan or multiple scans. In such a case, the search time t_s required for a single pass is increased by the number of beams that are used to cover the search volume. The average power required of the transmitter for the search function is E_s/t_s. The average power multiplied by the effective area of the receiving antenna is referred to as the power-aperture product and is sometimes used as a figure of merit for assessing the performance of radars in performing the search function.

The formulation given does not necessarily represent an optimum search strategy for a particular radar. Target cross-section statistics and metric parameters are external constraints. The radar implementation including its average power, sensitivity, mechanically or electrically scanned antenna, and its repertoire of wave forms are internal constraints. Mechanically scanned antennas have a relatively rigid search implementation. The electrically scanned phased array usually has several more degrees of freedom, including a wider repertoire of wave forms and rapid positioning of the antenna beam. Developing an optimum search strategy in either type of radar system is one of the biggest challenges for the designer—particularly in low-*S*/*N* ratio environments (Billetter, 1989).

In the search scan, there is usually a loss in *S*/*N* ratio due to the nonuniform coverage, since the peak of the beam coinciding with target position occurs infrequently. An allowance is made in the radar equation under L_o to account for this beam-shape loss, which may be as high as 1.6 dB in mechanically scanned radars and 6 dB in phased-array radars. The latter value is primarily due to the loss in effective aperture area caused by scan angle. Losses that are incurred in the receiver and signal processor, i.e., L_{sp}, include those where the target is not centered in the filter either in range or in Doppler shift and are known as the range-gate straddle loss and Doppler-filter straddle loss with values in the ranges of 1–4 dB and 0.5–2 dB, respectively. The CFAR detection process results in typically 1.0 dB of loss.

A search radar may also include a track function in the data processor. It is referred to as track-while-scan (TWS). The search mode of a radar effectively determines the presence and current position of a target. If the estimated position of a target from a scan is stored in the data processor, the next scan position estimate enables a velocity vector to be established and a track to be established for that target. No additional energy is required to perform the track function, and the number of tracks is limited only by the data processor. A major challenge in TWS systems with a large target-number capacity is the association of position estimates on a scan-to-scan basis with existing track files and the potential of establishing false tracks. Fan-beam search radars develop tracks in two dimensions, whereas pencil-beam search radars develop three-dimensional tracks.

3.1.2 Track In search radars, the ranging and antenna pointing direction is independent of the target; in tracking radars, they are dependent. The signal from the target is used for range, Doppler-shift, and angle estimates for positioning the time and frequency filters and for pointing the antenna beam to follow the target. The single-pulse random or noise error in the range estimate for a fixed target is

$$\sigma_r = \frac{\eta_w c}{2\beta K_r \sqrt{2(S/N)}}, \tag{29}$$

where η_w is a weighting factor used in the convolution of the signal and the signal processor filter, c is the velocity of light, β is the signal bandwidth, K_r is an interpolation factor, and the *S*/*N* ratio is given by Eq. (1). For angle estimates on a fixed target, the expected error for either axis is

$$\sigma_{\alpha x,\alpha y} = \frac{\lambda / d_{x,y}}{K_{mx,my}\sqrt{2(S/N)}}, \tag{30}$$

where $d_{x,y}$ is the dimension of the antenna in either x or y, and K_m is the monopulse slope with values ranging from 1.0 to 1.8. After substituting Eq. (1) for the *S*/*N* ratio into Eq. (30) for both axes, the energy E_t required for a specified noise error in each axis for an angle measurement is

$$E_t = P_t\tau = \frac{\lambda^4}{A_r^3}\cdot\frac{2\pi R^4 k T_s L_t L_r L_p L_o}{\eta_t \eta_r^2 K_{mx} K_{my} \sigma_{ax}\sigma_{ay}\sigma}. \tag{31}$$

Since range and Doppler estimates are performed at the same time as angle estimates, there is no additional energy requirement. The Doppler-shift estimate error σ_d as limited by white noise is given by

$$\sigma_d = \frac{\eta_d}{t_o K_d \sqrt{2(S/N)}}, \tag{32}$$

where η_d is the weighting factor in the convolution of the received signal and the receiver signal processor filter, t_o is the observation time, and K_d is an interpolation factor.

The total energy required for tracking is dependent on not only the single-pulse estimate requirement but also the number and frequency of samples to provide both current and predicted positions to the required degree of precision. Losses in the tracking radar are generally considerably less than those for a search function. Since the targets are tracked continuously, the peak of the beam is always pointed toward the target, the range and Doppler-shift filters are centered on the target, and CFAR is not required. The wave-form weighting losses discussed in Sec. 3.5 have to be considered for inclusion in the signal-processing loss.

3.2 Transmitters

The variety of available tubes and solid-state rf sources permits a wide range of architecture in radar transmitter design for the more complex radar systems. A modern phased-array radar with thousands of radiating elements may utilize a tube or solid-state amplifiers in each element. A few years ago, transmitter designs were predicated upon either one or a few high-power tubes with a microwave distribution system feeding the array elements. The transmitter was a separate entity. The development of solid-state devices has led to transmit–receive (TR) modules that, when coupled to each element, form an integrated transmitter, antenna, and receiver. Current designs utilize thousands of modules in one phased-array radar face. Small tubes are also available that can fit into such modules when higher peak power levels are required.

Modern radars require transmitter spectral purity and coherence. Spectral sidebands due to power-supply–induced phase modulations as well as nonlinear phase characteristics as a function of wavelength (especially for wideband linear-FM wave forms) occur when ripple voltages from the transmitter power supply and modulator are translated into phase and amplitude modulations on the output rf signal. Extremely tight regulation and ripple reduction of the voltages required to operate both tube and solid-state devices are critical.

Linear beam tubes, such as the traveling-wave tube (TWT) and the klystron, will exhibit small-signal gain fluctuations as a function of input signal frequency due to small-signal reflections within the tube structure. These small-signal gain variations translate to insertion phase and amplitude variations at the output when the tube is operated at saturation. For pulse-compression systems that require low range sidelobes, these variations must be actively canceled out (by predistorting the input drive signal or by closed-loop cancellation around the tube, for example) or known to an accurate degree so that they can be canceled out during receiver signal processing.

3.3 Antennas

The radar antenna design starts with the function it has to perform, i.e., surveillance, tracking, fire control, imaging, communications, or combinations of functions. Single-function antennas are generally single-reflector systems that are directed mechanically. Some of the larger single-function systems such as an over-the-horizon (OTH) radar use phased arrays. Multiple-function systems, performing search and tracking, either use a single antenna per function or a phased array whose beam is directed electronically by controlling the phases among the radiating elements of the array.

The radar equation demonstrates the importance of the antenna aperture in terms of the range of the system. The larger the aperture of the antenna, the greater the energy density on the target, and the greater the amount of reflected energy received by the radar. A figure of merit for the antenna is the power gain, i.e.,

$$G_t = \frac{4\pi\eta_t A_t}{\lambda^2 L_a} = \frac{D_t}{L_a}, \qquad (33)$$

where η_t is the illumination efficiency, A is the area, λ is the operating wavelength, D_t is the directive gain, and L_a is the antenna loss. Power gain expresses the increase in power density along the axis of the antenna beam, relative to that radiated by an isotropic antenna. Directive gain considers only the impact of the aperture size and illumination and is equal to the power gain with a lossless antenna. With the advent of phased arrays and the use of TR modules and their combination of radiator, transmitter, and receiver, directivity is sometimes a more useful term as a figure of merit, since it enables one to focus on the aperture directly rather than an indeterminate antenna input.

The values of η_t and η_r are indicators of the illumination efficiency and the sidelobe level of the transmitting and receiving pattern of the antenna, respectively. Signal returns through the sidelobes of the antenna from clutter, multipath propagation, or electronic countermeasures interfere with the signal return through the main antenna beam and have adverse impacts on detection probability and measurement accuracy. In a hostile operating environment for the radar, high antenna sidelobes during radar transmission are a source of homing energy for antiradiation missiles (ARMs). In most systems $\eta_r = \eta_t$; but in more recent designs, changes in the transmission or reception illumination may be made on a pulse-to-pulse basis, depending on needs in the operating environment. Uniform illumination across the aperture gives the highest efficiency and the highest sidelobes. Illumination tapers similar to the weighting functions in Table 2 may be used to lessen sidelobes. Sidelobe levels approaching 50 dB below isotropic may be obtained, but at the expense of reducing efficiency and gain.

3.4 Receivers and Exciters

The rf reference signal and the modulation wave form are generated by the exciter for the transmitter. In an adaptive system such as a phased-array radar, the wave form and rf signal may be different on a pulse-to-pulse basis. This reference signal is also offset in frequency and furnished to the rf portion of the receiver to enable down-conversion of the received signal to the intermediate-frequency (IF) section of the receiver to ease hardware implementation requirements. In monopulse systems, separate receivers are utilized for the sum and angle channels. A factor in the monopulse angle measurement error is the match of the separate receivers in phase and amplitude response over a wide dynamic range of received signals.

Table 2. Characteristics of representative weighting functions.

Weighting function	Peak sidelobe (dB)	*S/N* loss L_{sp} (dB)	Relative 3-dB resolution
Uniform	−13	0	1
$\cos(2\pi f/\beta)$	−32	1.8	1.6
Hamming $a + b\cos(2\pi f/\beta)$ $a = 0.1, b = 0.9$	−43	1.3	1.5

3.5 Signal Processors

Signal processing in radar is the process of extracting information about the presence of radar targets and their characteristics from received echoes of the transmitted signals. Signal processing is conventionally distinguished from data processing according to whether the processing is performed on data received from single or multiple dwells. The term *dwell* refers to a single measurement period in a particular antenna beam direction with one or more transmissions. Signal processing consists of the functions carried out on a single-dwell basis, i.e., resolution, detection, and single-dwell estimation. Data-processing functions involve data from multiple dwells, and the term *observation time* is used with respect to estimation and prediction.

Radar signal processing is generally characterized by simple algorithm structures implemented on high-throughput processors. Sliding-window (nonrecursive) linear filters and correlators are used for pulse compression, Doppler filtering, image formation, and interference power estimation. Fourier analysis, implemented via the fast Fourier transform (FFT), provides an efficient means of implementing pulse compression via the frequency domain. The FFT efficiently implements a uniform bandpass filter bank for pulsed Doppler processing, and it is also used for image formation and angle-of-arrival estimation. Weighting functions for sidelobe control are used in most filtering and FFT operations. Table 2 lists characteristics of a few representative weighting functions. The loss indicated is part of L_{sp} in Eq. (1). Less common, but still important, operations are recursive (feedback) filters and rank-ordering operations.

Radar signal processors are implemented with combinations of analog and digital circuitry or, increasingly, all-digital circuitry. Current (1994) technology generally favors digital implementation for low- to moderate-bandwidth operations such as Doppler filtering and pulse compression for narrow-band (<10 MHz) wave forms. Wide-band (>10 MHz) wave forms are often candidates for analog implementation, at least through the pulse-compression stage.

The trend in signal processor technology is increasingly toward programmable digital implementations. The primary technological limitation is analog-to-digital (A/D) converter technology; however, A/D converters providing 16 bits of resolution at sampling rates up to 2 MHz, falling to 9 bits at 500 MHz, have been demonstrated (Hicks, 1992). There is also a loss associated with A/D conversion that also contributes to L_{sp} in both search and tracking radars. Commercial off-the-shelf signal processors can obtain throughputs of 100 million floating point operations per second (Mflops) on a single 6-in. by 9-in. (9-cm by 23-cm) circuit board, and this performance is rapidly improving. Several more complex radar systems use massively parallel computers offering performance in the 1 giga-floating point operations per second (Gflops) range for real-time signal processing. These performance figures, which will be obsolete in a few years, are still adequate for many radar applications.

3.6 Data Processors

The receiver and signal processor perform the detection process and then extract short-term parameter estimates from the received signal. The data processor stores and operates on these estimates to provide control for the radar and outputs to the user. With the introduction of the adaptive phased-array radar system, the data processor is an integral part of the system rather than a link between the radar and the user. In the control function, the data processor must provide direction to the radar regarding the target to be tracked, the frequency to be used, the wave form to be used, and the times for transmission and reception of the signal. It must perform short-term prediction to determine where the next range gate and antenna beam position must be. Longer-term predictions might be concerned with averting potential collisions between aircraft or directing intercepts between missiles and targets. The computer also provides a means of checking the radar hardware and the software algorithms in the computer. In the processing of data for the user, targets can be characterized and potentially identified by evaluation of their trajectories, and target priorities can be established.

In the single-target tracking radar, the data-processor role is relatively small; it performs the estimation and prediction processes required for pointing the beam and range gate and performs limited processing before sending data to the user. In a search radar used in a TWS mode, the role of the data processor and its software is more complex and is dependent on the number of targets to be tracked. In the phased-array radar, the role grows to the extent that it can be more than one-third the cost of the radar, where the cost is in the designing, programming, and testing the algorithms for controlling the radar and processing the data for the user. The software in modern phased-array radars may represent more than a third of the total development costs and may be the dominant risk area. Most data processors today are general-purpose computers, and the hardware cost is a fraction of the total cost.

Some data processing, such as beam steering and coordinate conversions in phased-array radars, is performed by special-purpose rather than general-purpose computers, since such repetitive processes can take up a large percentage of a central-processor computational time. In a multiple-target environment, the limitations may be in the speed and memory capacity of the digital interface between the radar and the data processor.

3.7 Wave Forms

3.7.1 The Ambiguity Function As already indicated, with adequate *S/N* ratio, the transmitted wave-form bandwidth determines the range resolution and range measurement capability. The dwell time determines the resolution and measurement capability in terms of Doppler shift. The capability of a radar system wave form to resolve and perform measurements on a scatterer or element of a scatterer when it is in proximity to another scatterer is indicated by the *ambiguity function,* or *ambiguity diagram,* described by Woodward (1963). Mathematically, the ambiguity function χ is

$$\chi^2(T_d,f_d) = \int_{-\infty}^{+\infty} u(t)u^*(t + T_d)\exp(-j2\pi f_d t)dt, \tag{34}$$

where $u(t)$ and $u^*(t)$ are the normalized signal returns from the first and second scatterers, respectively, T_d is the time difference in range, and f_d is the difference in frequency shift between the two. If the integral is equal to 1, no resolution of the two scatterers can occur. If the integral is equal to 0, then the targets are fully resolved. The function $\chi(T_d,f_d)$ is essentially the correlation function for a *combined* time and frequency shift, or, equivalently, range and Doppler shift.

The ambiguity function also describes the output of a radar receiver with a matched filter. Woodward pointed out that the Fourier transform of this correlation function or the signal energy spectrum as an output of the matched filter is a sufficient description of the radar system resolving power in time and frequency.

The ambiguity function does not give a figure of merit of one wave form with respect to another. It specifies a distribution in the time–frequency domain from which one may estimate the relative degree of resolution. A wave form that might be ideal for measurements of a particular class of targets may be inadequate for another class.

3.7.2 The Ambiguity Diagram The *ambiguity diagram* has the time and frequency domain as two coordinates, with the energy density plot $|\chi(T_d,f_d)|^2$ orthogonal to the time–frequency plane. An ambiguity diagram of a rectangular pulse train with constant-frequency pulses from a target signal passing through a matched filter is given in Fig. 8, with the magnitude of $|\chi(T_d,f_d)|^2$ represented by the size of the spectral components as displayed on the diagram. The width of the pulse (dwell to null) in the time domain is 2τ after passing through the matched filter. The bandwidth of a single pulse in the frequency domain is $1/\tau$. A train of four pulses would require a dwell time of t_d, and the frequency-domain bandwidth of the individual line spectra would be $2/t_d$ (null to null). This gives much better resolution in the frequency domain than from the single pulse. If one examines the frequency axis at $t = 0$, for a single rectangular pulse, a $\sin^2 x/x^2$ spectrum would be observed representing the self-convolution of the transmitted spectrum. For a train of pulses, the $\sin^2 x/x^2$ would be the envelope of the spectral lines

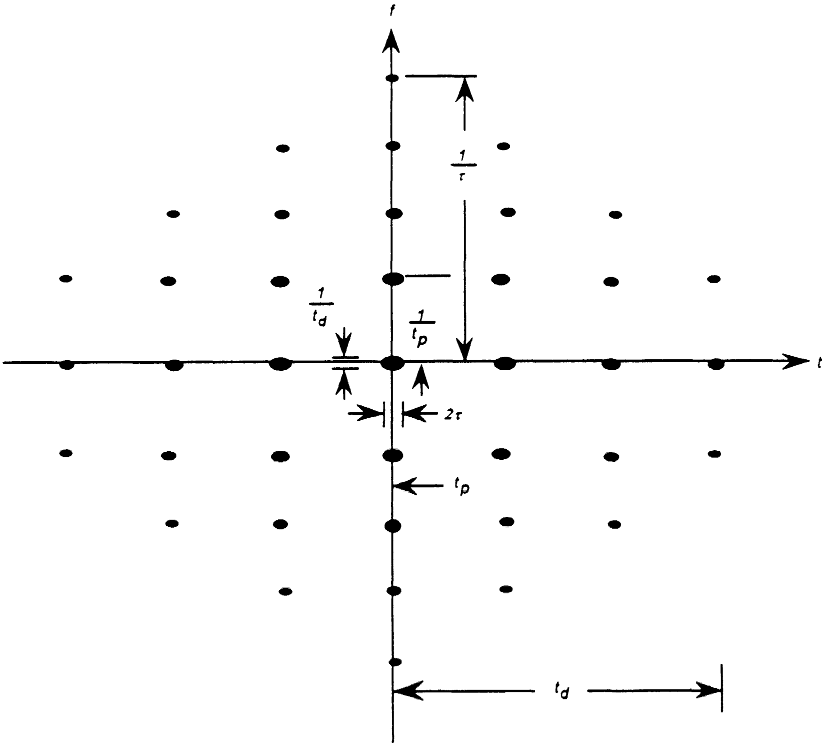

FIG. 8. Ambiguity diagram of four-pulse train.

whose spacing is approximately the PRF, or $1/t_{\mathrm{PRF}}$, but individually modified by the Doppler-shift difference at each frequency.

Many search and tracking radars work unambiguously, i.e., their instrumented range is less than the range corresponding to the PRF period (t_{PRF}) and their instrumented Doppler shift is less than the PRF. Long ranges and high radial velocities of a target may require a radar to operate ambiguously in either range or Doppler shift, or both. This process is equivalent to undersampling a signal. This is particularly true of an airborne radar in a search mode covering long extents of ground clutter, where the search is performed at a high PRF (>100 kHz) so that a target can be readily resolved from the ground clutter by virtue of its Doppler shift. The target return will be highly ambiguous in range. A fundamental problem in developing radar systems designed to detect target returns in the presence of large clutter returns via Doppler filtering is the suppression of such range/Doppler ambiguities. Ambiguities are typically resolved by changing the PRF from one coherent integration interval to the next and then using residue arithmetic techniques to estimate the true range and Doppler coordinates of the target.

3.7.3 Wave-Form Selection As with the ambiguity function, the ambiguity diagram is a useful tool in determining the resolving power of a wave form. The design of wave forms to meet requirements for resolution, estimation, and ambiguity usually starts with simple wave forms of known characteristics. Through an iterative process, using the target and its environment as one set of bounds and limitations imposed by the hardware and software mechanization as a second set, a repertoire of wave forms may be developed for a particular radar. The more complex radar systems may need a set of 20 or more wave forms to meet all requirements. Table 3 lists a variety of wave forms used in radar. The choice depends on the target and the environment. Nathanson (1991) provides excellent descriptions of the appropriate criteria and the selection process.

Table 3. Typical wave forms used in radar.

Wave form	Modulation
Continuous wave	Unmodulated
Train of cw pulses	No frequency modulation
Train of FM/cw pulses	Linear frequency modulation
	Nonlinear frequency modulation
Train of phase-coded pulses	*N*-bit biphase (Barker)
	N-bit polyphase (Frank)
Train of pulses with subpulses frequency coded	*N* subpulses noncoincidental in time and frequency

3.8 Radar Measurements

The primary function of radar is to perform measurements in order to furnish data on target position with a specified degree of accuracy at a rate consistent with user needs. Accuracy is defined as the quality of freedom from error and is expressed in terms of the error. The total error in radar measurements is the root sum square (rss) of the precision σ_n and residual systematic error σ_{rse}, i.e,

$$\sigma_e = (\sigma_n^2 + \sigma_{\mathrm{rse}}^2)^{1/2}. \tag{35}$$

Precision is defined in terms of the repeatability of measurement data and is generally equated to the random or noise error component, i.e., σ_n. The random error is a component of error whose magnitude and direction vary in a random manner in a sequence of measurements. It is primarily a measure of the error due to the thermal noise in the radar receiver, though there are other random errors in the radar and the propagation path. The systematic error is an inherent bias (offset) in the radar or propagation path. The residual systematic error σ_{rse} is that error left after correction or calibration of a systematic or bias error and may include registration error with the coordinate system of the user.

In a well-designed radar system of given resolution capability, i.e., bandwidth and beamwidth, that has been calibrated to reduce the systematic errors, the errors due to thermal noise may tend to dominate the accuracy, particularly if the radar data are being used to predict the future position of the target. The expected random or noise error on a single-pulse basis, i.e., Eqs. (29) and (30), can be reduced through integration, either predetector or postdetector, or in a data-processor tracking filter and through observation time.

Since most users of radar data are interested in predicting the future target position, the lower bound of the random error for a constant-velocity target is given by Gelb (1974) as

$$\sigma_n^2 = 4\frac{\sigma_m^2}{n} + 12\frac{\sigma_m^2}{nt_0}t_p + 12\frac{\sigma_m^2}{nt_0^2}t_p^2, \qquad (36)$$

where σ_m is the standard deviation of the measurements given by Eqs. (29) and (30) translated from the curvilinear radar coordinate system into a linear coordinate system for estimation and prediction. The terms t_p and t_o are the prediction and observation times, respectively, and n is the number of independent samples. The output samples from a matched filter that integrates more than one pulse would not be independent, and n would need to be adjusted accordingly. The first term in Eq. (36), which is an endpoint estimate for constant-velocity targets rather than Eqs. (29) and (30), is the noise-error component of the total tracking error for constant-velocity targets, though several texts refer to Eqs. (29) and (30) as the tracking errors. All three of these equations apply equally well to TWS systems and tracking radars. The primary difference between the two is the amount of predetector and postdetector integration prior to the data furnished to the tracking filter in the TWS system. Larger amounts of integration tend to reduce noise errors at the expense of increasing residual systematic or bias errors.

Equation (36) shows the importance of both the number of pulses integrated and the time interval over which a target is observed for prediction. From empirical data, however, there appears to be a limit on the integration that can be accomplished for a number of reasons, i.e., target motion, correlation between samples, measurements contaminated by radar instrumentation errors, and target and propagation noise with spectra much lower in frequency content than thermal noise. While not found in the literature, empirical data indicate that a total variance reduction of approximately 10 to 20 may be expected, though higher values are achieved in instrumentation systems.

The systematic error σ_{rse} in the accuracy equation is the error that is correlated over time or position with respect to an external reference and that can be measured and removed, i.e., calibrated. Consequently, the limit of systematic error is that due to the calibration errors and the electrical and mechanical stability of the radar. These residual errors are the antenna beam position and the range-resolution cell position.

Propagation errors due to refraction and multipath interference have systematic and random components that contribute primarily to elevation angle and range errors. Knowing the refractive index, one can correct propagation error to less than 10% of the original error. Short-term variations cannot be corrected. Often, when only a relative measurement between two targets is required, the error due to refraction is minimized. Though a number of techniques have been proposed to reduce multipath interference, there are none known to be particularly effective. The impact of multipath interference can be minimized by keeping the main antenna beam above the surface of reflection and reducing the antenna elevation sidelobes.

Clutter also has an impact on elevation and range measurement accuracy. The limit is established by the stability of the clutter and the ability of the radar to resolve the target from that clutter. The degree of resolution depends on the purity of the transmitted signal, stability of the radar, and the observation period or integration time. The errors due to clutter are proportional to the energy of the residual clutter in the time–Doppler-shift resolution cell and that of the signal.

In radars that utilize pulse compression as a means of getting more energy on target, the compression techniques produce not only the main return but also time sidelobes. Through these sidelobes, clutter and multipath interference returns can flood many of the range–Doppler-resolution cells. Their impact on measurements in a flooded cell can be reduced by 30–40 dB by weighting functions, such as those in Table 2.

Personal contact with highway speed radars may lead one to question Doppler-shift measurement accuracy. The small resolution cell available with long observation periods, as in pulsed-Doppler radars, enables a precise estimate of radial velocity well beyond the differentiation of position estimates. Cross-range velocities are still dependent on the time derivative of angle estimates; consequently, Doppler shift is most useful in resolution but has a place in situations where radial velocity estimates become of paramount importance.

4. APPLICATIONS

The applications of radar do not necessarily define the type of radar. In the process of air traffic control, one may find over-the-horizon (OTH) radars for tracking aircraft over the ocean, en-route surveillance radars for monitoring aircraft over land between terminal areas, terminal-area surveillance radars to monitor the aircraft near an airport, weather radars to determine optimum flight paths en route, landing radars as a complement to instrument landing systems, windshear radars to measure potential microbursts in the landing path, and surface detection equipment radars to monitor the ground movement of aircraft on the airport runways and taxiways. Aboard the aircraft will be the weather mapping radar and the radar altimeter.

Table 4 provides an indication of the relatively broad spectrum over which radar applications are found.

4.1 Search and Surveillance Radars

The search radar is one used "for detection of targets in a particular volume of interest," whereas a surveillance radar is used "to maintain cognizance of selected traffic within a selected area" such as an airport terminal area or air route (Radar Systems Panel of the IEEE Aerospace and Electronic Systems Society, 1990). Most radars operate with wavelengths in the 1-mm to 1-m regions. Their propagation characteristics are such that they are considered line-of-sight (LOS) radars with the propagation path in a straight line.

OTH radars operating in the 10- to 100-m wavelength region utilize either surface-wave propagation or sky-wave propagation, with the former having ranges of about 300 to 400 km, and the latter up to 4000 km on a single reflection from the ionosphere. OTH surface-wave systems depend on the low attenuation characteristics of the longer wavelengths when launched in a vertically polarized wave over a saltwater surface. Azimuth, range, and Doppler data can be obtained for aircraft and ships. Because of sea clutter, it is necessary to depend on Doppler information to separate the target from the background. Because of the long wavelength, the Doppler shift is low for aircraft targets, and careful design is required to separate the moving target from clutter from both the surface and the ionosphere. Sky-wave propagation is similar to that in communication systems, except that there is a two-way path. At these long wavelengths, the time of day and the impact of solar radiation on the ionosphere determine the frequency to be used and the ranges obtained. There are particular problems with auroras and the free-electron velocities encountered.

Table 4. Radar wavelengths and applications.

Band	Wavelength (m)	Frequency	Applications
HF	10–100	3–30 MHz	Over-the-horizon radar
VHF	1–10	30–300 MHz	Very long-range search
UHF	0.3–1	300–1000 MHz	Very long-range surveillance
L	0.15–0.3	1–2 GHz	Long-range surveillance, en-route air traffic control
S	0.075–0.15	2–4 GHz	Surveillance, long-range weather characterization, terminal air traffic control
C	0.0375–0.075	4–8 GHz	Fire control, metric instrumentation, short-range weather characterization
X	0.025–0.0375	8–12 GHz	Fire control, air-to-air missile seeker, marine radar, airborne weather characterization
K_u	0.0167–0.025	12–18 GHz	Short-range fire control, satellite altimeter, airborne surveillance radar equipment, remote sensing
K	0.011–0.0167	18–27 GHz	Water vapor attenuation band
K_a	0.0075–0.011	27–40 GHz	High-resolution remote sensing, weapon guidance
V	0.004–0.0075	40–75 GHz	Remote sensing, weapon guidance
W	0.0027–0.004	75–110 GHz	Remote sensing, weapon guidance
MM	0.001–0.0027	110–300 GHz	Experimental, weapon guidance

Both airborne and ground-based, line-of-sight, search and surveillance radar ranges are generally limited by the S/N ratio of the target return or by clutter and attenuation at the shorter wavelengths. These radars are primarily used for airborne targets, though there are some, such as the airport surface detection equipment (ASDE), that are utilized to monitor ground targets. Civilian use is primarily directed toward air traffic control, whereas military use is directed toward the detection of threat targets so that defensive action can be taken.

The limitation on the maximum number of simultaneous targets in search radars is the data processor. There are limits on the accuracy of the data, primarily because of the number of samples available from each target, the low update rate of the search scan, the bandwidth of the transmitted signal (which is small to reduce the number of range cells to be searched), and the beamwidth in elevation (which is large to provide coverage from the horizon to zenith).

4.2 Tracking Radars

The narrow pencil-beam tracking radar is used to achieve high-accuracy position data. Accuracy is usually considered the root mean square sum of the precision of the radar and the residual errors after calibration. Theoretical range- and angle-measurement precision for a fixed target are defined in Eqs. (28) and (29), and the tracking precision for a constant-velocity target is defined by Eq. (35). The radar can provide excellent range data but is generally limited in angle, as compared to optical systems, because of the λ/d ratio, where d is the antenna diameter. The overall accuracy is limited by

1. calibration errors in range gate and beam position of the radar with respect to a reference, and
2. propagation errors.

The highest-accuracy tracking radars are those whose antenna angle position is calibrated utilizing a collimated telescope and known stars and where measurements of the propagation condition allow correction of atmospheric refraction effects. Partially correlated instrumentation and propagation errors, in turn, produce phase and amplitude errors in the signal and limit multiple-pulse tracking accuracy.

Tracking radars were initially developed during World War II to direct antiaircraft artillery. Derivatives of these radars were developed for not only fire-control functions in ground and airborne air-defense operations but also instrumentation for military development and civilian space programs. They are utilized to determine accurate current and projected positions of powered and ballistic targets, both airborne and in space. Search and surveillance radars can provide two-dimensional data trajectories but of much lower accuracy. The state-of-the-art accuracy for end-point estimates of monopulse tracking radars performing satellite tracks is less than 0.1 mr in angle and 3–5 m in range. The performance of tracking radars for air-defense and fire-control systems is limited by the RCS of the target. Transponders operating at the radar frequencies are used on cooperative targets to improve the S/N ratio; these systems are called Information, Friend or Foe (IFF), in military use and have civilian counterparts.

4.3 Phased-Array Radars

Since most tracking radars are single-beam systems, the number of targets is limited to those in the narrow pencil beam, usually one. The need to track a multiplicity of threat targets, however, drove the development of phased-array radars. Though the theory of phased-array antennas has been known for many decades, development was limited because of the cost of phase shifters and time-delay units and the inability to control the flexibility of the system. The number of these units needed for a phased-array radar antenna is approximately the area divided by $0.3\lambda^2$. A second limitation, not recognized early, was lack of a control system that could take advantage of the adaptability of the antenna, transmitter, and signal processor. Development of the control system paralleled, but ran behind, the development of the digital computer. In the last two decades, multiple systems have been fielded. Though most are surface based, there are a number of airborne systems.

Its adaptability means that the design of the phased array to perform both the search and the track functions in a multiple-target environment is generally a compromise of wavelength, time, energy, and costs. The search function leads to the need for longer wavelengths to cover a large volume, and the track function leads to the need for shorter wavelengths. Algorithms for radar control must permit real-time compromise between search and track for energy and time resources while providing immediate access to these resources for the search for and track of high-priority targets. The final limitation is cost.

The initial systems were developed for surface-based air defense and ballistic-missile defense in several countries. Examples include the PATRIOT, AEGIS, PAVE PAWS, and BMEWS (Ballistic Missile Early Warning System) in the United States and the AMB-3, the Hen House, and the SA series of air- and missile-defense radars in the U.S.S.R. (Blake, 1989). It might be noted that the BMEWS radars were originally separate search and single-target tracker radars that were upgraded to phased-array radar systems. A number of other countries have built phased-array radars for air-defense and air traffic-control applications. In the latter, they are utilized for en-route surveillance and monitoring and controlling the landing of aircraft.

4.4 Weather Radars

In weather radars, in addition to the azimuth, elevation, and range measurements, there is an estimate of the zero, first, and second moments of the returned power spectral density, which in the case of precipitation are measures of the reflectivity, mean Doppler shift, and Doppler frequency dispersion. These data provide information about precipitation accumulation, storm identification and tracks, tornadoes, and atmospheric turbulence. Precipitation accumulation can be used to predict flooding, and storm identification and tracks can be used for warning.

The Terminal Doppler Weather Radar (TDWR) focuses directly on the problem of air travel and air safety. There have been a large number of fatalities and accidents due to wind shear during the landing of aircraft. The TDWR provides automatic detection of microbursts and low-level wind shear through the Doppler measurement of the radial velocity of rain or airborne microparticulates to obtain a direct indication of the presence of wind shear.

4.5 Ground-Probing Radars

Some degree of success has been achieved employing radar to characterize the depth profile of subsurface conditions. These so-called ground-probing radars generate a pulse of energy from a very broad-bandwidth antenna positioned in close proximity to the ground. Pulse lengths tend to be very short, in the 1- to 5-ns range, depending on range resolution and probe depth. The objective of such a radar is to detect and map subsurface interfaces and buried objects. The range to a subsurface layer or target is proportional to the two-way propagation time, which is influenced by such factors as the relative dielectric constant, loss tangent, and bulk attenuation characteristics of the media probed. These factors, in turn, are functions of mineral content, soil homogeneity, stratification, discontinuities, packing density, and surface conditions. Environmental factors that also affect performance include temperature, seasonal and diurnal variation, and especially the moisture content of the soil.

Applications of ground-probing radar systems have tended to use microwave frequencies below 3 GHz and have included detection and identification of nonmetallic mines, location of buried pipes, and location of air voids within highway pavements. Penetration performance is optimum under dry-medium conditions. The effectiveness of all of these applications is adversely affected when the pavement or underlying soil is wet. Salt penetration of roadbeds can create excessive signal attenuation in addition to high surface reflections.

4.6 Imaging Radars

4.6.1 Synthetic-Aperture Radars The synthetic-aperture radar (SAR) effects very fine cross-range resolution by artificially creating a very large linear-array antenna as discussed under Resolution in Sec. 2.3.1. Radar backscatter signals are recorded from an aircraft or satellite over several seconds time as if they had been captured by a real antenna array. The recorded data captured over a particular dwell time are then coherently summed to create an image of the surface, where the ideal cross-range resolution is one-half the actual antenna aperture, independent of range. Range resolution is obtained through traditional short-pulse or pulse-compression techniques.

The Fourier-transform–based SAR signal processing exploits the Doppler frequency relationship between the radar backscatter from a resolution cell on the surface with that abeam of the moving platform. Within a fairly broad angular approximation, the Doppler-shifted frequency from a resolution cell is linearly proportional to the distance from a point at the same range on the surface abeam of the platform, where the Doppler shift is zero.

SAR imaging has enabled cartographers to map regions of Central and South America, for example, that are usually shrouded from optical camera surveillance by constant cloud cover. Spatial resolutions of a meter or less are routinely achieved. Satellite imaging of the Earth and Venus has been routinely performed by satellite-borne and space-shuttle–based synthetic-aperture radar systems. Earth images have been used for such diverse purposes as remote sensing of geographical features, surveillance of military emplacements, and location of ships at sea. Magellan SAR-derived imagery of the planet Venus has been providing contour and reflectivity information about the Venerian surface, from which important geophysical information on that planet has been obtained.

4.6.2 Inverse Synthetic-Aperture Radars Synthetic-aperture radar signal processing creates imagery from the mathematical relationship between the instantaneous Doppler frequency shift of the backscattered signal from a surface resolution cell and the cross-range resolution cell position. This Doppler frequency shift can be the result of the target's rotational motion instead of that of the radar platform. If a target is rotating about an axis perpendicular to the radar line of sight, there will exist a Doppler-shifted signal from a resolution cell of the object that can be correlated with cross-range position. This signal-processing concept is called inverse synthetic-aperture radar, or ISAR.

Applications of ISAR signal processing include imaging of several planets and our own moon from a radar located on the Earth, imaging of ships from their slow rolling motion in the water, and imaging of satellites and reentry vehicles due to their apparent rotation with respect to a radar line of sight. ISAR images of military vehicles are routinely obtained by recording radar backscatter from targets placed on a turntable as it rotates within the illuminated volume.

4.7 Ground Vehicular Radar

4.7.1 Police Radars The well-known police radar used to monitor automobile speeds has a small antenna and requires operation at *X* (3 cm) and *K* bands (1.5 cm) to achieve resolution. These are usually homodyne radars in which the received signal is beat against the transmitted signal. For a 90-km/h velocity at *X* band, the Doppler frequency is about 1167 Hz. The output is amplitude limited, and the signal is passed into a digital counter that counts the zero crossings over a short period of time. The number of crossings per unit time gives the measured speed. Where the police radar is moving, the Doppler signal from nearby clutter intercepted by the radar beam is used as the measure of the police radar velocity, and the target vehicle has a Doppler frequency plus or minus that of the police vehicle. This type of radar is also used to measure baseball velocities or the speed with which a pitch is delivered.

4.7.2 Collision-Avoidance Radars Recent research efforts have been under way in Germany, France, the United States, and Japan in the development of short-range radar systems for automatic collision avoidance for automobiles and trucks. The concept usually takes the form of detecting the closing rate between the platform automobile and another vehicle, based on detecting a Doppler-shifted backscatter signal. Detection then triggers an automatic response of braking the automobile at a rate commensurate with the rate of range closing that was determined.

False-alarm reduction has remained a challenging area of development. False detections can result from antenna sidelobe detection of backscatter from roadside objects such as building, signs, and other cars. Other challenges include reducing the overall system cost and increasing the system robustness to make the radar application more commercially attractive.

4.8 Missile Radars

The first missile radars were the proximity fuzes developed during World War II. These were simple cw homodyne radars, where the received signal was mixed directly with the transmitted signal. The fuze was activated when the detected target-return amplitude achieved a preset threshold value. Today's more complex systems may utilize cw or pulsed electromagnetic waves but with modulations and signal processing that enable resolution in both the time and frequency domains. Many anti–air-missile seekers have such wide beamwidths that they require no search. Some antennas may be dual purpose, however, scanned to provide search coverage and acquisition, with a monopulse feed that enables the missile to track.

Some of these radars operate in a bistatic mode, that is, the host radar illuminates the target with a radar signal and the missile radar receives the reflection from the target. A rearward-pointing antenna on the missile may also allow the missile to determine its location with respect to the host radar and its range from the target. Modern missiles fired in a fire-and-forget mode, i.e., a monostatic mode, operate in a completely autonomous manner, both illuminating the target and receiving and processing the target's backscattered signal, without requiring the host's illumination of the target.

GLOSSARY

Accuracy: Freedom from error.

Backscatter Coefficient: The radar cross section per unit illuminated area (for surface clutter) or illuminated volume (for rain, smoke, fog, etc.).

Bandwidth: The frequency extent of a signal or the passband of an electrical component, usually measured between frequencies whose probability density function value is 3 dB below the average peak value.

Beamwidth: The angular extent of a one-way antenna power pattern, typically between points 3 dB down from angular position of maximum antenna gain.

Clutter: Any unwanted radar echoes, e.g., ground or sea surface, rain, etc.

Dwell: The period of time that the antenna beam is pointed at a specific resolution cell or cells in search radars.

ECCM: Electronic counter-countermeasures, i.e., radar hardware, signal-processing, or operational techniques that preserve radar detection and tracking performance in the presence of jamming.

ECM: Electronic countermeasures, i.e., jamming/antijamming techniques designed to deny or degrade radar detection or tracking.

Jamming: Electronic signals designed to generate interference in the receiving band of the radar.

MTI: Moving target indication: radar signal-processing techniques that detect targets having a component of velocity along the radar line of sight.

Observation Time: The period of time that the antenna beam is pointed at a target—particularly for targets being tracked. Can be equal to dwell time or multiple dwells in search radars.

Polarization: The spatial dependence of the electric field direction of an electromagnetic wave: linear, circular, or elliptical. Note that the engineering definition of circular polarization is defined looking in the direction of energy propagation and is opposite to the physics convention.

Precision: The repeatability of a measurement as limited by noise.

Pulse Compression: Any technique that uses coding on transmission and decoding on reception to transmit a long pulse length and receive a short pulse length, resulting in fine range resolution.

Pulse Length: The time duration of a transmitted radar signal.

Pulse Repetition Frequency (PRF): The number of transmitted pulses per second.

Pulsed Doppler Radar: A pulsed radar that utilizes the Doppler effect to determine the component of a target's velocity along the radar line of sight.

Radar Cross Section: A measure of the reflectivity of an object: 4π times the ratio of the power per unit solid angle scattered in a specified direction to the power per unit area in a plane wave incident on the object from a specified direction.

Range: Distance between the radar and a reflecting object (target).

Range Resolution: The ability to distinguish between two objects at the same radar bearing, separated only in range.

Swerling Model: A mathematical target model based on the echo-amplitude distribution-function characteristics and temporal dependence.

Track-While-Scan (TWS): A radar process that automatically tracks objects in range and bearing, with periodic position updates from a scanning antenna.

Works Cited

Barton, D. K. (1988), *Modern Radar System Analysis,* Norwood, MA: Artech House, pp. 304–305.

Barton, D. K., Cook, C. E., Hamilton, P. C. (1991), *Radar Evaluation Handbook,* Norwood, MA: Artech House.

Billetter, D. R. (1989), *Multifunction Array Radar,* Norwood, MA: Artech House.

Blake, B. (1989), *Jane's Weapon Systems 1988–89,* 19th ed., Surry, U.K., and Alexandria, VA: Jane's Information Group, Inc.

Currie, N. C., Dyer, F. B., Hayes, R. D. (1975), *Analysis of Radar Rain Return at Frequencies of 9.375, 35, 70, and 95 GHz,* Technical Report No. 2, Contract DAAA25-73-0256, Atlanta, GA: Georgia Institute of Technology.

Gelb, A. (1974), *Applied Optimal Estimation,* Cambridge, MA, and London, U.K.: MIT Press, Massachusetts Institute of Technology.

Guerlac, H. E. (1987), *Radar in World War II,* American Institute of Physics Series in the History of Modern Physics, Vol 8., New York: Tomash.

Hertz, H. (1900), *Electric Waves,* New York: Macmillan Company.

Hicks, R. C. (1992), "A Survey of Analog-to-Digital Converters for Radar Applications," in: *Proceedings of the IEEE International Conference on Radar (Radar 92),* New York: IEEE, pp. 534-537.

Hülsmeyer, C. (1904), *Hertzian-wave Projection and Receiving Apparatus Adapted to Indicate or Give Warning of the Presence of Metallic Body, Such as a Ship or a Train, in the Line of Projection of Such Waves,* British Patent No. 13 170.

Marconi, S. G. (1922), "Radio Telegraphy," *Proc. IRE* **10,** 215–238.

Nathanson, F. E. (1991), *Radar Design Principles,* 2nd ed., New York: McGraw Hill, Inc.

Radar Systems Panel of the IEEE Aerospace and Electronic Systems Society (1990), *IEEE Standard Radar Definitions,* IEEE Standard No. 686–1990, New York: IEEE.

Richards, M. A. (1988), "Synthetic Aperture Processing," in: G. V. Morris (Ed.), *Airborne Pulsed Doppler Radar,* Norwood, MA: Artech House, Chap. 9.

Skolnik, M. I. (1980), *Introduction to Radar Systems,* New York: McGraw-Hill, Inc.

Swerling, P. (1960), "Probability of Detection for Fluctuating Targets," *IRE Transactions* **IT-6,** 269–308.

Swords, S. S. (1986), *Technical History of the Beginnings of RADAR,* London, U.K.: Peter Peregrinus Ltd.

Woodward, P. M. (1963), *Probability and Information Theory, With Applications to Radar,* London: Pergamon Press.

Further Reading

Barton, D. K. (1988), *Modern Radar System Analysis,* 2nd ed., Norwood, MA: Artech House.

Currie, N. C. (Ed.) (1984), *Techniques of Radar Reflectivity Measurements,* Dedham, MA: Artech House.

Currie, N. C., Hayes, R. D., Trebits, R. N. (1992), *Millimeter Wave Radar Clutter,* Norwood, MA: Artech House.

Eaves, J. L., Reedy E. K. (1987), *Principles of Modern Radar,* New York: Van Nostrand Reinhold Company.

Levanon, Nadav (1988), *Radar Principles,* New York: Wiley.

Morchin, W. C. (1990), *Airborne Early Warning Radar,* Norwood, MA: Artech House.

Morris, G. V. (1988), *Airborne Pulsed Doppler Radar,* Norwood, MA: Artech House.

Oppenheim, A. V., Schafer, R. W. (1975), *Digital Signal Processing,* Englewood Cliffs, NJ: Prentice-Hall.

Scanlan, M. J. B. (1987), *Modern Radar Techniques,* New York: Macmillan.

Sherman, S. M. (1984), *Monopulse Principles and Techniques,* Dedham, MA: Artech House.

Skolnik, M. I. (1990), *Radar Handbook,* 2nd ed., New York: McGraw-Hill, Inc.

RADIATION BIOLOGY

See BIOLOGICAL EFFECTS OF ELECTROMAGNETIC RADIATION; BIOMEDICAL USES OF RADIATION

RADIATION DAMAGE IN CRYSTALS

P. Ehrhart, W. Schilling, and H. Ullmaier, *Institut für Festkörperforschung, Forschungszentrum Jülich, Jülich, Germany*

Introduction 429
1. **Production of Primary Defects** 430
1.1 Interaction between Radiation and Solids 430
1.2 Production of Individual Frenkel Pairs 432
1.3 Displacement Cascades 434
1.4 Ionization-Induced Defect Production 437
2. **Defect Migration and Reactions** 438
2.1 Jumps of Interstitials and Vacancies 438
2.2 Defect Recovery 439
2.3 Secondary Radiation-Induced Defects 441
2.4 Microstructural Evolution during Long-Term Irradiations 443
3. **Changes of Macroscopic Properties** 445
3.1 Metals and Alloys 445
3.1.1 Hardening and Embrittlement at $T \leq 0.35T_m$ 445
3.1.2 Swelling 447
3.1.3 Irradiation Creep and Growth 448
3.1.4 High-Temperature Embrittlement 449
3.2 Carbon-Based and Ceramic Materials 450
3.2.1 Dimensional Changes and Stored Energy 450
3.2.2 Thermal Conductivity and Mechanical Strength 452
3.3 Electronic and Optical Materials 452
3.3.1 Electrical Properties 453
3.3.2 Optical Properties 454
Glossary 454
Works Cited 455
Further Reading 456

INTRODUCTION

Irradiation with energetic particles such as neutrons, ions, and electrons, or with hard x rays, causes structural and/or chemical changes in practically all solids. This phenomenon is called "radiation damage" irrespective of whether the associated property change is beneficial or detrimental to the envisaged application. Modern radiation damage research started with the installation of the first nuclear reactors. On that occasion, Fermi and, especially, Wigner pointed out the danger of severe property changes of core components under the extended neutron irradiation. Subsequent intensive research on radiation effects led to a branching into major directions: Solid-state physicists recognized that bombardment with energetic particles offered a unique possibility to create well-defined defects in solids in a controlled manner. They carried out low-temperature, low-dose irradiations of pure metals, often in single-crystalline form, to establish the properties of the most basic radiation defects, vacancies and interstitial atoms, i.e., Frenkel pairs. As a result of this fundamental research, the basic production mechanisms and properties of radiation-induced defects are now understood in pure metals and simple ionic crystals. For semiconductors, oxides, and intermetallic compounds, however, the properties of even the simplest radiation-induced defects are still controversial or unknown. The other direction of research was dictated by the need to predict the rather complex property changes (e.g., embrittlement, swelling, or in-pile creep) that would occur in commercial multicomponent materials during many years of reactor irradiation at elevated temperatures. Over the last years an increasing cross linking between basic and applied radiation effects research can be recognized: One reason is that the development of radiation-resistant materials for future fusion reactors re-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

quires—because of the lack of powerful sources for the high-energy fusion neutrons—irradiation testing with reactor neutrons or fast ions. To transfer the results of such "simulation irradiations" to actual fusion conditions, the underlying defect production and accumulation processes must be understood in detail. Also affecting radiation effects research is the widespread industrial application of ion-beam techniques for microelectronics and thin-films technology. These techniques rely on a detailed understanding of the associated ion-induced radiation damage (see also ION BEAM MODIFICATION OF MATERIALS).

In the following section we discuss the production of the primary defects, the structure of Frenkel pairs, and the local arrangements of defects in displacement cascades. In Sec. 2 we review mobilities of Frenkel defects and the defect reactions that lead to recovery and to different microstructures during long-term irradiations at elevated temperatures. The final Sec. 3 summarizes technologically relevant changes of macroscopic properties in different classes of materials during irradiation and their relation to the radiation-induced microstructures.

1. PRODUCTION OF PRIMARY DEFECTS

1.1. Interaction between Radiation and Solids

It is now established that all macroscopic radiation effects are caused by one of the following elementary interactions between radiation and atoms: electronic excitations, nuclear transmutations, or elastic collisions. Electronic excitation usually leads to ionization of the atoms in the crystal. In most materials, the resulting electron-hole pairs do not produce lasting damage. Exceptions are organic materials, where the molecular bonds broken by the ionization may combine to form new chemical species, e.g., by cross linking, as well as special ionic crystals (see Sec. 1.4). The transient electrical photocurrents that are produced in semiconductors and insulators during the ionizing irradiation, however, are usually not considered radiation damage.

The effects of nuclear transmutations are particularly important for neutron irradiations. An example is the transmutation doping of silicon by thermal neutrons via the reaction $Si^{30}(n,\gamma)Si^{31}$ and the subsequent β decay of Si^{31} into P^{31} with a half-life of 2.6 h. By this technique large Si crystals used, e.g., in high-power devices can be precisely doped with good homogeneity. Another important class of nuclear reactions are those in which α particles are emitted, e.g., from (n,α) processes induced by fast neutrons. The helium atoms remaining after the α particles have come to rest are insoluble in most materials and will therefore precipitate in the form of bubbles. The associated problem of helium embrittlement is discussed in Sec. 3.1.4. Since the cross section for (n,α) reactions increases strongly with neutron energies above ~1 MeV, helium embrittlement will be a severe problem for future fusion reactors where the materials will be exposed to high fluxes of 14-MeV neutrons.

In the elastic collision process, the irradiating particle transfers a recoil energy T to a lattice atom, which is then called the primary knock-on atom (PKA). If this recoil energy exceeds a material-dependent threshold energy T_d, so-called displacement damage is created in the crystal. Details of this damaging process are summarized in Table 1 and is discussed in the subsequent sections.

The magnitude and the distribution of T depend on the nature and energy E of the irradiating particle. T varies between zero and the maximum energy transferred in a head-on collision,

$$T_{\max} = \frac{4Mm}{(M+m)^2}E \quad \text{and}$$
$$T_{\max} = \frac{2m_e}{M}\left(\frac{E}{m_e c^2} + 2\right)E, \tag{1}$$

for nonrelativistic particles and relativistic electrons, respectively. M and m are the masses of the recoil atom and the bombarding particle, respectively. The probability that an atom after irradiation with dose Φ has received a recoil energy between T and $T + dT$ is given by $\Phi d\sigma(T)$ (Φ is the number of particles per incident unit area or total path length of all particles per unit volume). The form of the differential cross section $d\sigma(T)$, which reflects the type of interaction between irradiating particle and lattice atom, determines the distribution of recoil angles.

Table 1. Steps in the production process of primary defects.

Time (ps)	Event	Result	Characteristic parameter
0	Transfer of recoil energy to a lattice atom	Primary recoil atom (PKA)	T = PKA energy T_{dam} = damage energy $d\sigma(E,T)$ = cross section
0–0.3	Slowing down of PKA	Collision cascades and subcascades	T_d = displacement threshold
0.3–3	Spike formation, interstitial ejection	Hot spike core, atomic mixing, replacement collision sequences (RCS)	T_{sp} = maximum spike temperature n_{sc} = number of subcascades
3–10	Spike core solidification and cooling to ambient	Stable Frenkel defects (FD), depleted zone, disordered zone, amorphous zone	$\nu(T)$ = number of stable FD's per PKA η_{diso} = disordering efficiency η_{amo} = amorphization efficiency

For a quantitative characterization of the damage introduced by a given dose of irradiating particles, each recoil event must be weighted by the amount of damage it introduces into the crystal. In this manner, the so-called damage or displacement cross section has been defined by a convention of the American Society for Testing Materials (ASTM, 1989):

$$\sigma_{\mathrm{dam}} \equiv \int_{T'=T_d}^{2.5T_d} d\sigma(T') + 0.4 \int_{T'=2.5T_d}^{T_{\max}} \frac{T_{\mathrm{dam}}(T')}{T_d}\, d\sigma(T'), \qquad (2)$$

where T_d is an effective displacement threshold energy and is tabulated partly by the ASTM (1989) and Table 2 or otherwise taken as $T_d \approx 25$ eV. The damage energy, $T_{\mathrm{dam}}(T')$, in Eq. (2) takes into account that during the slowing-down process of the PKA part of its energy is converted into electronic excitations. These electronic losses are most important for recoil energies above $T/\mathrm{keV} \geq 0.1M$ (where the mass M of the recoil atom is in atomic mass units). Equation (2) is based on the idea, originally proposed by Kinchin and Pease (1955), that recoils with $T < T_d$ do not produce displacement damage, that recoils with $T_d < T < 2.5T_d$ produce just one displacement (or Frenkel pair), and that recoils with energies above $2.5T_d$ produce a number of displacements which increases linearly with T_{dam}. In the semiconductor literature on radiation effects, instead of σ_{dam}, an "introduction rate" $\Sigma = n\sigma_{\mathrm{dam}}$ is generally

Table 2. Defect parameters for typical metallic and nonmetallic materials. $T_{d,\min}$ is the minimum recoil energy to create a stable Frenkel pair. T_d is the recommended displacement energy for dpa calculations. E_i^m and E_v^m are the activation energies for thermal migration of interstitials and vacancies, respectively.

Material	Structure	$T_{d,\min}$ (T_d) (eV)	E_i^m (eV)	E_v^m (eV)
Al	fcc	16 (25)	0.1	0.6
Cu	fcc	19 (30)	0.1	0.7
Ni	fcc	23 (40)	0.15	1.05
Fe	bcc	17 (40)	0.3	0.55
Mo	bcc	34 (60)	0.08	1.3
p-Si	Diamond	21	0	0.3–0.5 [V^{++}]
n-Si			0.3	0.2 [V^{--}]
s.i.-GaAs	Zincblende	9 [As_i]	0.5 or 1.5 [As_i]	1.7 [V_{As}]
		18 [Ga_i]	0.8 [Ga_i]	≈1.7 [V_{Ga}]
Graphite	Layered	24 (60)	<0.1	3.1
MgO	Rock salt	60 [O_i]	1.6 [O_i]	≥2 [V_0]
KCl	Rock salt	Radiolysis	0.1 [Cl_i]	0.85 [V_{Cl}^+]
				0.7 [V_{K}^-]

quoted, where n is the atomic density of the target material.

To compare the damaging effect of different irradiating particles, the irradiation dose, Φ, is calibrated in units of displacements per atom (dpa), where

$$\text{dpa} = \Phi \sigma_{\text{dam}}. \tag{3}$$

If there is a distribution in the energy E of the irradiating particles, σ_{dam} must be averaged accordingly.

For a quantitative characterization of the recoil spectra produced by different irradiating particles, the normalized damage distribution $W(T)$ is a convenient quantity. It is calculated by taking the integrals in Eq. (2) only up to $T' = T$ instead of $T' = T_{\max}$ and normalizing this cross section by σ_{dam}. In this manner, $W(T)$ gives the fraction of damage produced by recoils with energies smaller than a given value T. Examples for different irradiating particles are shown in Fig. 1. It is evident from the logarithmic scale for T that the different particles transfer their energy to the atoms in quite different regimes. For example, the median recoil energy $T_{1/2}$ (i.e., the recoil energy below which half of the displacements are produced), is about 60 eV ($\approx 2T_d$) for 5-MeV electrons and about 160 keV ($\approx 2000T_d$) for fast reactor neutrons. Other projectiles lie somewhat between these extremes. Note that neutrons produce recoils in a relatively narrow range whereas MeV protons produce recoils uniformly between T_d and $T_{\max}$.

1.2 Production of Individual Frenkel Pairs

Pairs of individual interstitials and vacancies, Frenkel pairs (FP), constitute the basic elements of radiation damage. They are generated by low-energy knock-ons, typically below 200 eV. The production mechanism and structure of Frenkel pairs have been determined only for simple metals for which molecular-dynamics computer simulations could be used. In these simulations, the classical equations of motion are solved for an assembly of 10^3–10^5 atoms interacting via a suitably chosen interatomic potential. The motion of all atoms can be followed as a function of time after one of the atoms is given an energy T. Unfortunately, reliable yet tractable interatomic potentials are only available for a few metals.

The atomic motion involved in a typical single displacement event is shown in Fig. 2 for fcc Cu. The most prominent feature is the creation of a replacement collision sequence (RCS) that propagates at supersonic velocities. Each atom in the sequence successively replaces its neighboring atom, along a close-packed atomic row. The replacements continue until so much energy has been lost to the surrounding lattice that the last atom in the sequence can no longer displace its neighbor. This last atom falls back toward its starting point which, however, is now occupied by the preceding atom in the collision sequence. Therefore, two atoms must share one lattice site at this position, i.e., an inter-

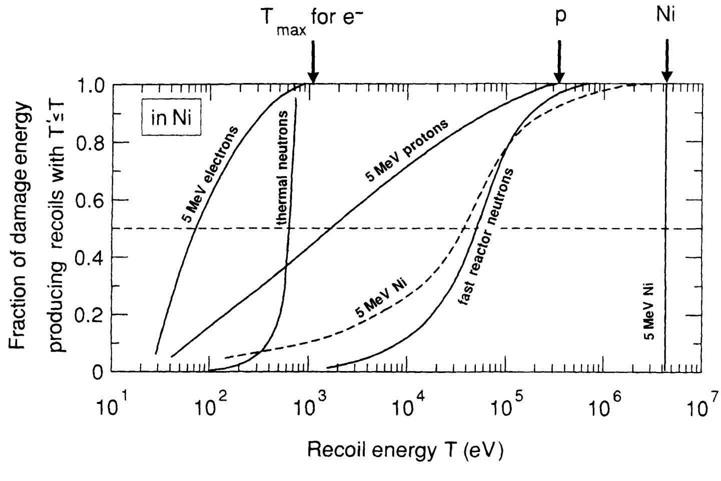

FIG. 1. Fraction $W(T)$ of the damage energy with T' smaller than a given value T produced by recoils from different irradiating particles in nickel. For the 5-MeV self ion, two curves are shown: The solid line (step at 5 MeV) applies if no distinction is made between Ni as a bombarding particle and a PKA, respectively. The dotted curve considers only energy transfers to Ni atoms and excludes the bombarding Ni. The recoils generated by thermal-neutron irradiation primarily result from the γ rays emitted after the neutron capture.

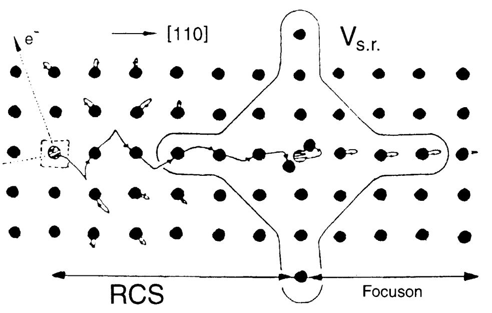

FIG. 2. Molecular-dynamics simulation of a single displacement event in fcc Cu. The figure shows the trajectories of the atoms generated by a PKA with $T = 50$ eV and recoil direction 10° off the [110] axis in the (100) plane. The PKA (shaded in its initial position) comes to rest next to the generated vacancy. The interstitial is created at the position of the other shaded atom where the replacement collision sequence (RCS) got stuck. $V_{s.r.}$ indicates the spontaneous recombination or instability volume of Frenkel pairs.

stitial is formed. Beyond the interstitial position, momentum, but no mass, continues to propagate in the collision direction as a so-called focuson. Because of the action of the RCS, the PKA usually ends up next to the vacancy.

RCS are an essential element in Frenkel pair production, at least for the close-packed metals, because they provide an efficient mechanism to separate the interstitial from its vacancy. A minimum separation is necessary to avoid subsequent athermal recombination of the interstitials and vacancies. As indicated in Fig. 2, the recombination volume V_{sr} defines the volume around the interstitial in which a vacancy is mechanically unstable to recombination. The threshold energy T_d for Frenkel-pair production can therefore be considered as the minimum energy required to separate the interstitial and vacancy beyond the range of instability. Since replacement collision sequences spread easiest along close-packed lattice directions, minimum values of T_d are expected for recoils along these directions and are, indeed, observed. The minimum threshold energies $T_{d,\min}$ listed in Table 2 thus pertain to displacements around the directions ⟨110⟩ and ⟨100⟩ in fcc and around ⟨111⟩ and ⟨100⟩ in bcc metals. For defect production calculations, a threshold energy that is averaged over all crystal directions must be employed. These energies are also listed in Table 2.

The $T_{d,\min}$ values listed in Table 2 are at least three times larger than the energy stored as formation energy of the created Frenkel pair. Apparently during the rapid formation of the Frenkel defect by the displacement process within 10^{-12} s, a large fraction of the energy $T_{d,\min}$ must be lost to neighboring atoms during the collision sequence event. Only a small fraction of $T_{d,\min}$ is stored as potential energy in the lattice.

Displacement processes in diatomic materials, with the atoms having different masses, arranged on different sublattices, and having different interaction potentials, are complicated. In addition to two kinds of vacancy and interstitial type defects on the different sublattices, antisite defects can also be produced in these materials. This complicated situation is illustrated in electron-irradiated GaAs. At the onset of damage around $T_{d,\min} \approx 9$ eV, the production of interstitials and vacancies on only the As sublattice is reported. The threshold energy for the production of Ga interstitials seems to be higher and, perhaps, associated with double displacements, generating an As interstitial simultaneously with a Ga interstitial. At recoil energies below 19 eV, it seems that recoils starting from the As sublattice, aside from Frenkel defects, can create also Ga_{As}/As_{Ga} antisite pairs. The corresponding exchange between an As and Ga lattice atom is triggered by a replacement collision with subsequent spontaneous recombination of the so-created instable interstitial with the vacancy at the starting point of the PKA. Defect production in oxides, like MgO or Al_2O_3, is similarly complicated, although antisite defects are not formed. One general trend, however, is that the threshold energies on the oxygen sublattice seem to be remarkably high, ~60 eV.

The atomic configuration of the interstitials produced by the displacement process is now established for metals and for the anion defects in alkali halides. In all cases, the interstitial consists of two atoms sharing one lattice site forming a "dumbell" whose axis is along ⟨100⟩ for fcc metals (experimentally verified for Al, Cu, Ni) along ⟨110⟩ for bcc metals (experimentally verified for Mo, W, and Fe), along ⟨110⟩ on the anion sublattice of fcc halides (experimentally verified for KCl and KBr), and along ⟨100⟩ for the bcc halides (e.g., CsBr) (see Figs. 3 and 9). The reason why dumbell configurations are preferred over the more symmetric octahedral or tetrahedral interstitial sites in metals stems from the strong repulsive interaction between the interstitial atoms and their neighbors. In the halides, on the other hand, the neutral anion interstitial seems to be stabilized by covalent bonding between the dumbell atoms and their neighbors in the ⟨110⟩ row. Whatever the detailed reason for their configuration, the strong repulsion between the atoms squeezed into the interstitial configuration creates large local displacements and leads to interstitial relaxation volumes of the order of 1–3 atomic volumes.

For other materials, especially semiconductors and oxides, little is currently known about interstitial structures and properties. This is due to the experimental difficulty that the interstitials generated during electron irradiation are mobile at low temperatures, so that their reaction with impurities cannot be avoided. Theoretical calculations, moreover, suffer from the difficulty to take properly into account the large Jahn–Teller splitting of the electronic defect levels and the accompanying large spontaneous distortion of the molecular defect center.

1.3 Displacement Cascades

Primary knock-on atoms with energies $T \approx 1\text{–}10^3$ keV produce displacement cascades. In contrast to the single-displacement event, the PKA slows down over distances much larger than a lattice distance, and it can displace many more than one atom. The secondary recoils, in turn, displace other lattice atoms and so on, creating a displacement cascade (see Table 1). Molecular-dynamics simulations, performed on modern high-speed computers, have clarified many details of the damage production in displacement cascades. According to these studies, the cascade evolution comprises three consecutive stages (see Table 1): The initial stage, called collision cascade, lasts about 0.3 ps, i.e., approximately the period of a lattice vibration. In this stage, the PKA and higher-order knock-ons slow down by two-body collisions until they reach energies below T_d and can no longer displace additional lattice atoms.

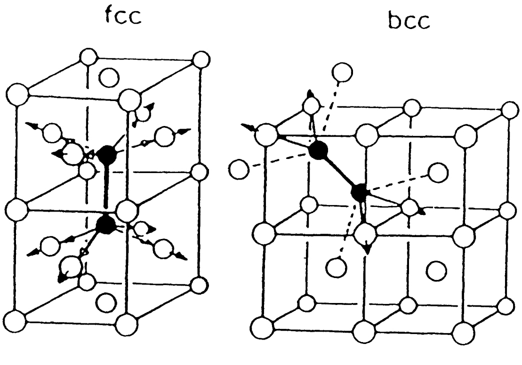

FIG. 3. The stable configurations of self-interstitials in fcc and bcc metals: Two lattice atoms (black) share one lattice site forming a so-called dumbell with the dumbell axis along ⟨100⟩ in fcc and along ⟨110⟩ in bcc metals.

In the subsequent time, from 0.3 to about 3 ps, the collision cascade develops into a heat spike during which most of the atoms localized in this region are set into violent motion. The energy density within the spike core can exceed several tenths of an eV/atom, corresponding to maximum temperatures T_{sp} of a few times the melting temperature. The atomic arrangement in the spike resembles a liquid droplet with a density lower than the surrounding lattice. This lower density is a consequence of a shock wave, thermal expansion, and the ejection of interstitial atoms. The snap shot shown in Fig. 4 illustrates these features: A highly disordered liquidlike spike core with a lower atomic density is surrounded by a shock front in which the lattice is compressed at pressures of order of the theoretical shear strength. Although the spike concept was introduced in the 1950s by Seitz and Koehler (1956), it had remained controversial whether the energy from the collision cascade dissipates slowly enough to the surrounding lattice to allow significant atomic rearrangement by diffusional processes. Today, the molecular-dynamic simulations of large cascades and much circumstantial experimental evidence strongly support the spike concept.

After a time of about 3 ps, the central molten spike zone rapidly cools and solidifies, attaining ambient temperatures after about 10 ps. In principle, both phonons and hot electrons can contribute to the heat dissipation from the hot spike core. In simple metals such as Cu or Al, however, the elec-

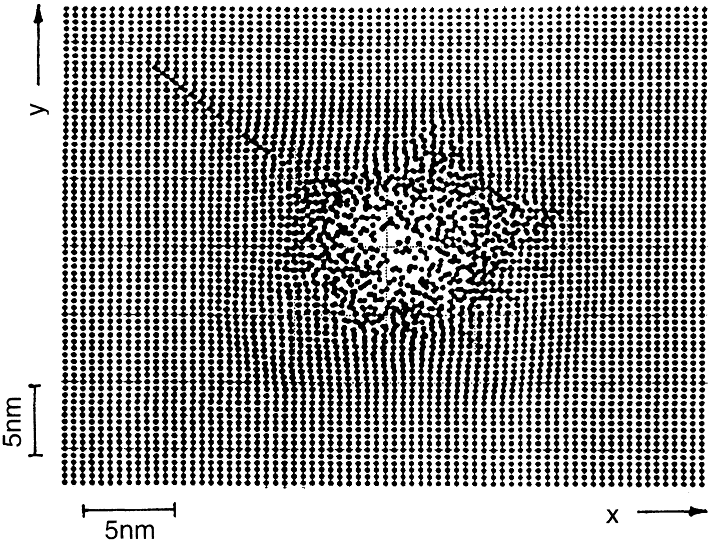

FIG. 4. Atomic configuration at 1 ps after a 10-keV PKA event in Au. The atom positions within a cross-sectional slab of three atomic planes near the spike center are projected on a {100} plane. Molecular dynamics simulations (MDS) by Ghaly and Averback (private communication). Note that the computer print shown has slightly different scales in the *x* and *y* directions of the atomic positions. By an inclined viewing of the graph the shock front and the wave along the RCSs are clearly visible.

tron-phonon coupling is too weak for the electrons to contribute to spike cooling. In other materials such as Ni, Fe, or alloys, the question whether electrons or phonons dominate the cooling rate is still open. During the rapid cooling of the spike, with rates of the order of some 10^{14} K/s, the interface between the central molten spike core and the surrounding crystal must regress with a very high velocity, of the order of the velocity of sound. This can only be achieved by a high degree of undercooling of the spike, of the order of $0.3T_m$. Only in simple lattice structures like fcc or bcc metals can atoms manage to rebuild the original crystal structure epitaxially at such high speeds. In more complicated structures, e.g., in ordered alloys in which the atoms occupy different sublattices, in compound semiconductors, or in oxides, the original crystal structure cannot regrow so quickly and a highly disordered nonequilibrium phase is frozen in. In most semiconductors (see, e.g., Fig. 5 for Si) and many oxides, e.g., SiO_2, an amorphous phase is usually produced by cascade damage. Intermetallic compounds tend to become amorphous under irradiation if their ordering energy is high but only chemically disordered if their ordering energy is low.

Because of the ejection of interstitials, the spike core will be depleted of atoms also after its solidification. During the rapid cooling the "free volume" existing in the spike is driven toward the center in a way resembling zone refining. Consequently, a compact vacancy-rich zone is formed at the end of the cooling process ("depleted zone"). Experimental investigations using field-ion microscopy, transmission electron microscopy, and

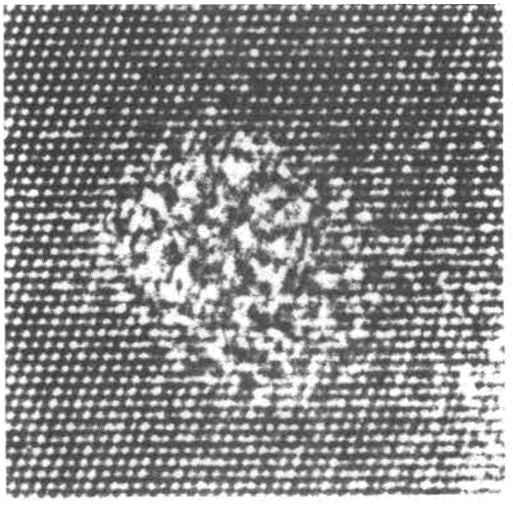

FIG. 5. High-resolution TEM micrograph of a cascade produced in Si by implantation of a 100-keV Bi^+ at 77 K. The amorphous zone in the center of the {111} slab has a rather sharp boundary. The surrounding lattice contains considerable disorder in form of disrupted ⟨110⟩ atomic chains. (From Narayan *et al.*, 1985.)

diffuse x-ray scattering have verified this structure. Since interstitials cannot survive within the initially very hot spike center, all interstitials present at the end of the spike cooling must reside at the periphery of the cascade. Figure 6 illustrates schematically the resulting defect arrangement.

Another fundamental feature of the defect arrangement produced by displacement cascades is the formation of subcascades. As illustrated in Fig. 7, for recoil energies above a critical value T_{sc}, the collision cascade branches into individual subcascades. The basic reason for the branching is that the average distance traveled by a recoil atom between energetic secondary recoil collisions increases rapidly with increasing recoil energy. When this distance is larger than the linear dimension of the secondary cascades, subcascade formation is inevitable. The formation of subcascades is, however, a statistical process and fluctuates strongly from one displacement event to the other. According to the laws for the stopping of self ions, T_{sc} depends sensitively on PKA mass; e.g., $T_{sc} \approx 25$ keV for Cu and 170 keV for Au. The average number n_{sc} of subcascades per PKA is approximately given by $n_{sc} \approx T_{\text{dam}}/T_{sc}$. This linearity suggests that at higher recoil energies subcascades can be considered as new independent damage units. For light materials, e.g., Al or Si, where $T_{sc} < 1$ keV, the spike concept is no longer applicable, because too few atoms are contained in each subcascade to establish a liquidlike core with a clear transition region to the surrounding crystalline lattice.

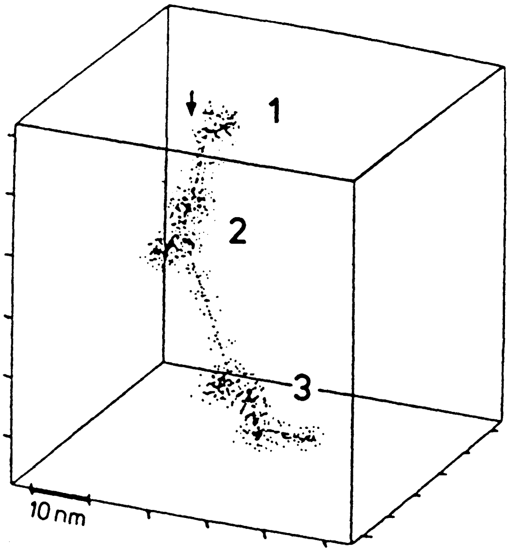

FIG. 7. Three-dimensional view of a 100-keV collision cascade in Cu ($T \approx 4T_{sc}$) generated by the MARLOWE computer code (after Heinisch, 1990). The dots represent vacancies as well as displaced atoms. All displaced atoms having a distance less than 0.8 nm to a vacancy have been eliminated from this plot (simulating spontaneous recombination). Three well separated subcascades (1 to 3) are recognized; subcascade 2 and 3 each have three lobes.

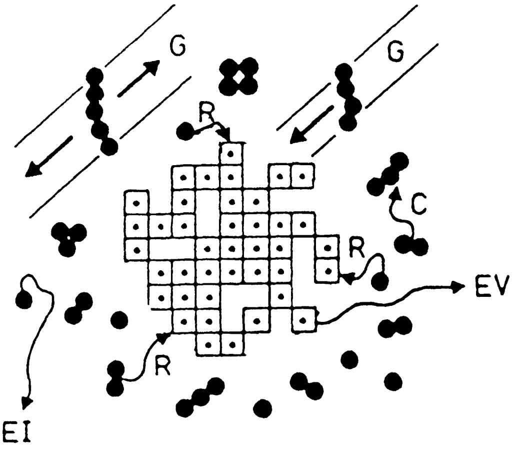

FIG. 6. Schematic representation of the defect arrangement in a displacement cascade showing the vacancy-rich core (depleted zone) and the interstitial shell containing single-, di-, tri-, ··· interstitials up to small loops. The possible interactions during defect migration are indicated as follows: R = intracascade recombination, G = glide of small interstitial loops toward, or bypassing, the depleted zone, leading to intracascade recombination or interstitial escape, respectively. EV and EI denote escaping vacancies and interstitials, respectively. C denotes intracascade clustering of interstitials.

For a quantitative characterization of the total number $\nu(T)$ of vacancies or interstitials produced per PKA of energy T, a damage efficiency is defined, i.e., $\xi(T) = \nu(T)/$dpa. As shown in Fig. 8, at $T > T_{sc}$ the damage efficiency attains a constant value ξ_{fn} of about 0.25, that is most characteristic for fast neutrons (hence the subscript) and reflects the fact that the subcascades contribute independently to the defect production. This low value shows that the energy necessary to produce a stable Frenkel defect in a cascade is about 3 to 4 times larger than in isolated low-energy recoil events and reflects

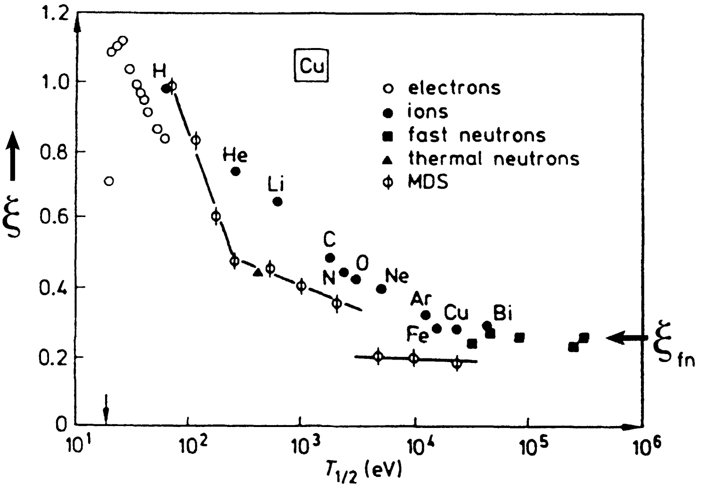

FIG. 8. Damage efficiency ξ (i.e., stable FD produced per dpa) as a function of the median recoil energy $T_{1/2}$ in copper, deduced from resistivity measurements on specimens irradiated with different particles at 4.2 K. Compilation of data by Averback *et al.* (1978) and Kinney *et al.* (1984), using T_d = 30 eV (see arrow on abscissa) and a Frenkel-pair resistivity ρ_F = 2.5 $\mu\Omega$ m. The symbols connected by the solid lines give results from MDSs by Foreman *et al.* (1992) and by Diaz de la Rubia and Guinan (1991), respectively. For the irradiations with heavy ions, e.g., 500-keV Cu, $T_{1/2}$ refers to the (secondary) recoils giving partly overlapping DCs. If the implanted ion is counted as a PKA the datum point for Cu must be plotted at $T_{1/2}$ = T_{dam} = 240 keV (from Schilling and Ullmaier, 1993).

the fact that for cascade damage only interstitials ejected far from the core of the cascade can survive. All atoms displaced within the spike core will end up at lattice sites during the resolidification process.

Similar to the efficiency for the production of interstitials and vacancies, efficiencies for disordering or amorphization by displacement cascades can be defined. These efficiencies, η_{diso} and η_{amo}, give the number of atoms contained in disordered or amorphous zones per dpa, respectively. Typical values of η_{diso} and η_{amo} are 10–30 dpa^{-1}, demonstrating a very efficient local atomic mixing within the spike, in comparison to typical efficiencies $\xi \approx 0.2$ to 0.4 dpa^{-1} for Frenkel defect production.

1.4 Ionization-Induced Defect Production

In addition to the elastic collisions, electronic excitations can also lead to permanent atom displacements in many inorganic insulators including alkali halides, alkaline-earth halides, silver halides, and several carbonates and silicates. This process, called radiolysis, results from the decay of excitons generated by the ionizing radiation. For radiolysis to occur with significant efficiency, the electronic excitations must have energies in excess of Frenkel-defect formation energies, and an efficient means for transferring the stored electronic energy into atomic motion must be available. The details of this conversion mechanism are best understood for the alkali halides (Fig. 9). A characteristic feature of this process, as shown in the figure, is that only one sublattice is involved in the formation of the Frenkel pairs. For the alkali halides, quantum efficiencies for this exciton decay mechanism as high as 0.8 have been reported; however, only a small percentage of the associated recoil processes actually lead to stable Frenkel pairs. In KCl, an average energy of $\approx 10^3$ eV is absorbed before a stable Frenkel pair is formed, and even larger amounts of energy are required for many other crystals. Nevertheless, the radiolytic process can often far exceed the defect production by atomic collisions, e.g., during irradiations in the electron microscope. There are obvious similarities between this process and the photographic process in AgBr; however, the latter process involves the presence of the surface. Other beneficial applications of radiolytic defect production include photochromic materials, crystal memories, and color centers in tunable lasers.

Completely different mechanisms of defect creation have been observed for the very high local electronic excitations achievable by GeV heavy-ion bombardment. Under these conditions the electronic energy loss of the particle is much larger than the nuclear stopping:

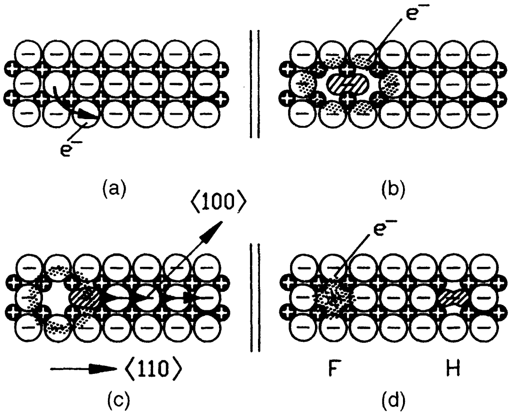

FIG. 9. Schematic view of the different steps of the radiolysis process in the (100) plane of an alkali halide with NaCl structure. (a) Ionization of an anion that yields an electron-hole pair (exciton). (b) Relaxation of the hole into the self-trapped configuration (X_2^- molecular ion on the halogen sublattice). The electron remains in a diffuse hydrogenic orbital. (c) Asymmetric decay of the exciton and initiation of a replacement collision sequence along ⟨110⟩ (or, not shown, an off-center rotation of the X_2^- molecule out of the plane). (d) Final configuration of the Frenkel pair, consisting of a so called *H* center (neutral anion interstitial) and *F* center (anion vacancy with neutral effective charge).

$$\left.\frac{dE}{dX}\right|_e \geq 2000 \left.\frac{dE}{dX}\right|_n \approx 1\ \text{keV/Å}.$$

Although the details of the damage process are still unresolved, the basic effect involves a Coulomb explosion caused by the electrons being scattered out of the track of the fast ion and subsequent rearrangement of the remaining positive ions. In many insulators, for example, amorphous tracks with diameters typically 6–10 nm along the ion paths are produced. By chemical etching, these tracks can be enlarged to diameters in the μm range, and homogeneous "nuclear track filters" can be manufactured. The ion tracks have also been shown to act as pinning centers for flux lines in ceramic high-T_c superconductors. Recently, ion tracks have even been observed in semiconductors (Fig. 10) and in intermetallic alloys.

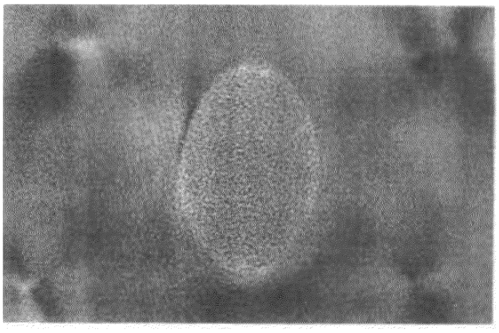

FIG. 10. High-resolution electron micrograph of an amorphous ion track in single-crystalline GeS after bombardment with 1.4-GeV U ions. The direction of the ion beam was perpendicular to the image plane and the sample has been thinned to obtain transparency for the electrons. The elliptical shape of the track (24 nm high and 16 nm wide) correlates with the anisotropy of the orthorhombic crystal (from Scholz *et al.*, 1993).

2. DEFECT MIGRATION AND REACTIONS

The primary defect arrangement considered in Sec. 1 can only be studied if the defects are frozen in during an irradiation at low temperatures. At higher temperatures, the radiation-induced interstitials and vacancies become mobile. They can then react either with each other, with impurities, or with extended sinks such as dislocations, grain boundaries, or free surfaces. In the following sections, we first review the basic information available on interstitial and vacancy mobilities and then discuss the reactions of the mobile defects in typical recovery experiments. Finally, we focus on the microstructures that evolve from defect reactions during long-term irradiations.

2.1 Jumps of Interstitials and Vacancies

Atomic motion in crystals nearly always involves jumps of interstitials or vacancies between neighboring equilibrium positions. Such jumps are activated by thermal fluctuations with a jump rate $\nu_X = \nu_0 \exp(-E_X^m/kT)$ where E_X^m is the activation energy for the jump of defect X, ν_0 is the attempt frequency, and T in this case is the absolute temperature. Using a typical value of $\nu_0 \approx 10^{13}\ \text{s}^{-1}$, one arrives at a rule of thumb for the temperature T_X^a at which the defect X can perform one jump per second of $T_X^a = (380\ \text{K/eV})E_X^m$. Typical migration energies for interstitials and vacancies are listed in Table 2. These values have usually been inferred from kinetic studies of the defect recovery after irradiation or quenching (see Sec. 2.2).

For vacancies the atomic jump mechanism is simple: a neighboring atom jumps into a vacant lattice site. In contrast, the jump mechanism for self-interstitial atoms is rather complicated, and presently it is known only for metals. In this case, the interstitials migrate by a collective process involving the interstitial dumbbell atoms and their neighbors, as shown in Figs. 11(a) and 11(b). The center of the dumbbell moves one nearest-neighbor distance while the axis of the dumbbell rotates into another crystallographic direction. Note that each atom within the interstitial configuration moves only a small distance during the jump in comparison to the translation of the interstitial center. This fact, together with a kind of local shear instability (due to the high local compression of the atoms in the interstitial configuration), accounts for the unusual low migration temperatures $T_i^a < 100$ K of interstitials in metals. In some metals, such as Au, Nb, or Cd, interstitials even migrate below 1 K, possibly under the influence of zero-point fluctuations (tunneling). Small agglomerates of interstitials can migrate at low temperatures by similar collective processes.

Interstitials and vacancies in elemental semiconductors also have low migration energies, but in III-V and II-VI semiconductors defect motion on one sublattice appears to be hindered by the presence of atoms on the other sublattice, resulting in higher migration energies—see Table 2.

For defects in nonmetals, saddle-point configurations like their equilibrium structures depend on their charge state. Therefore the same defect may have quite different migration energies in different charge states, i.e., at different positions of the Fermi level (see, e.g., the values for *n*-Si and *p*-Si in Table 2). In addition, the strong influence of the charge state on the atomic displacements at the defect center is indicative of a strong local coupling between the electrons and phonons. This coupling creates the possibility of inducing defect jumps, even at 0 K, by cycling the defect between different charge states, e.g., by successively capturing free electrons and holes in an ionizing irradiation field or during carrier injection. The electronic energy released by the electron-hole recombination is then funneled into local atomic motion which, like thermal agitation, can induce a jump of the defect into a neighboring equilibrium position. Another possibility for athermal defect migration is the Bourgoin–Corbett mechanism. In this case, the saddle-point configuration of the defect in one charge state is centered at the equilibrium site of the same defect in another charge state and vice versa. Hence, changing the charge state brings the defect from its initial equilibrium position into the saddle-point configuration of the other charge state, from where it can relax into the adjacent new equilibrium position. Cycling between two charge states can thus induce long-range defect migration, even at 0 K. A classic demonstration of athermal defect motion was performed in a series of electron paramagnetic resonance experiments (Watkins, 1975). They showed that silicon interstitials produced by electron irradiation in *p*-type silicon at 4.2 K migrated over long distances to substitutional Al impurities where they were converted and detected as Al interstitials.

2.2 Defect Recovery

Reactions among radiation-induced interstitials and vacancies can be classified as recombination, clustering, and impurity trapping. These reactions are commonly studied in recovery or annealing experiments. For

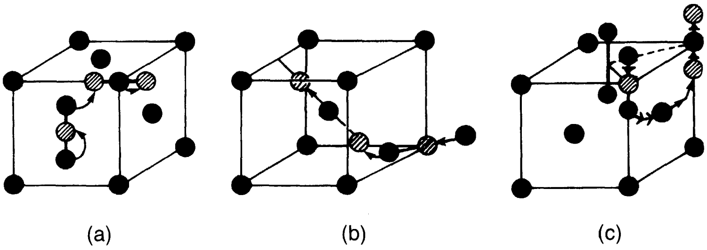

FIG. 11. Migration of (a) the ⟨100⟩- and of (b) the ⟨110⟩-split single interstitial in fcc and bcc metals, respectively. The configurations after the jump are indicated by the shaded atoms. (c) Migration of a di-interstitial (two parallel ⟨100⟩ dumbbells on nearest-neighbor sites) in the fcc lattice by a two-step jump of one of the dumbbells.

these experiments, the defects are first introduced by irradiation at low temperatures. Subsequently, the samples are warmed in a controlled manner, and the change in the defect population is monitored by measurements of suitable physical properties such as residual electrical resistivity in metals and thermal conductivity or optical absorption in semiconductors and insulators. A typical recovery curve for metals is shown in Fig. 12. It exhibits a number of annealing stages labeled stage I through stage V. After a long-lasting controversy, the following assignment of defect reactions to these stages is now widely accepted:

Stage I: In this stage, interstitials recombine either with their own vacancies or with another vacancy after free migration through the lattice. In addition to recombination, freely migrating interstitials can also form clusters or be trapped at foreign atoms. These latter two processes are responsible for the incomplete recovery in stage I (see Fig. 12). Recall that in cascades, interstitials and vacancies are separated (see Fig. 6) so that interstitial clustering reactions are strongly favored. This explains why defect survival is much larger for fast-neutron–irradiated samples than for electron-irradiated ones. After electron irradiation, stage I shows a number of substages, $I_{A,B,C}$, which are due to the collapse of close Frenkel pairs. As indicated in Fig. 13, the attractive interaction between the defects lowers the activation barrier for the jump of the interstitial toward the vacancy. Substage I_D results from the recombination of the free migrating interstitials with their own vacancies (correlated recombination) whereas substage I_E results from the free migrating interstitials recombining with distant, foreign vacancies. The latter process requires a larger number of jumps and, therefore, somewhat higher annealing temperatures. Some metals such as Au, V, Nb, or Cd do not show a well-defined annealing stage I. In these metals, it has not been possible to freeze in isolated interstitials even at the lowest irradiation temperatures, e.g., 0.3 K for Au.

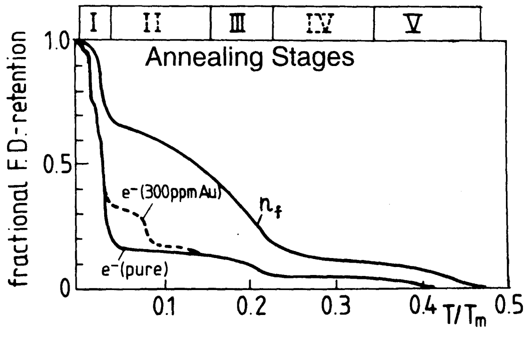

FIG. 12. Frenkel-defect retention in Cu irradiated at 4 K with fast neutrons (n_f) or 3-MeV electrons to about 10^{-5} dpa after 10-min isochronal annealings at increasing homologous temperatures (T/T_m). The dashed line indicates the influence of alloying with 300 ppm Au. The roman numerals and boxes show the common labeling and extent of the major recovery stages.

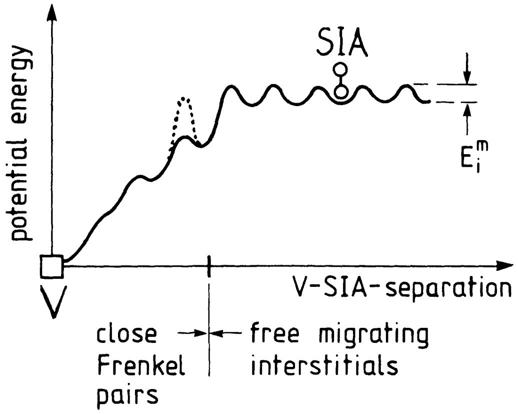

FIG. 13. Schematic representation of the potential energy of a Frenkel pair as a function of separation between the vacancy (V) and the self-interstitial (SIA) atom for metals (solid line) and semiconductors (dashed line).

Stage II: Stage II refers to the interval in temperature between the generic recovery stages I and III, but, for historical reasons, it is also called "recovery stage." In this temperature interval, interstitial clusters grow to larger sizes by mutual coalescence, a process that continues up to recovery stage V. In impure samples, interstitials that were trapped at impurities in stage I can be detrapped. This leads to distinct substages within recovery stage II (see Fig. 12).

Stage III: In this stage, vacancies begin to migrate and to annihilate at the interstitial clusters. Vacancy agglomeration also begins. If these vacancy clusters survive (together with interstitial loops) at the end of stage III, the recovery is not complete. On the other hand, some metals such as Al, Pt, and Au (only after electron irradiation) do show complete recovery. In these metals, apparently too high losses of mobile single and multiple vacancies at the interstitial clusters prevent the nucleation of stable vacancy agglomerates.

Stage IV: Similar to stage II, this is not a unique recovery stage but rather the interval between stage III and V. In this temperature interval, the vacancy clusters surviving stage III grow in size by mutual coalescence or by an Ostwald ripening process.

Stage V: In this stage, defect clusters thermally dissociate. The evaporation of single vacancies that annihilate the interstitial loops is generally the dominating annealing process and all radiation-induced Frenkel defects are removed at the end of the stage.

In contrast to metals, there is no comparable systematic understanding of the evolution of the overall defect pattern during annealing of irradiated semiconductors and oxides. The complexities stemming from athermal and charge-dependent thermal-defect mobilities and the formation of many different defect–impurity complexes (samples used for spectroscopic studies of defect recovery must unavoidably contain only very small defect concentrations) make it difficult to identify the intrinsic recombination and clustering processes of Frenkel defects. The general trend in these materials, however, is that clustering of interstitials and vacancies is more difficult than in metals, because of the long-range electrostatic repulsion of defects with equal charges (see also Sec. 2.3). In a similar vein, the recombination between interstitials and vacancies also seems often to be suppressed by the existence of deep trapping positions for interstitials near vacancies (see Fig. 13). In this case, the interstitials can overcome the final barrier for recombination with the vacancy only at higher temperatures than the temperatures at which they migrate freely.

An important aspect of interstitial trapping at impurities is the possible enhancement of impurity diffusion by the so-called kick-out mechanism. Upon trapping of a self-interstitial atom, a substitutional impurity can be "kicked out" into an interstitial site. If the so-formed interstitial impurity is able to jump from one interstitial position to another (without decaying again into a substitutional impurity plus self-interstitial), impurity diffusion can be dramatically enhanced. Examples for rapid radiation-induced diffusion via the kick-out mechanism are Ni(Si), Au(Fe), Si(Al), and Si(B).

2.3 Secondary Radiation-Induced Defects

Under prolonged irradiation at elevated temperatures, a large variety of secondary defects is observed in all kinds of materials. Examples are shown in Fig. 14. These secondary defects result from the aggregation of mobile interstitials and vacancies created continuously as primary defects by the displacement process. In contrast to the defect clusters that form in a typical recovery experiment (see Sec. 2.2), the interstitial and vacancy agglomerates considered here as secondary defects are much larger.

Interstitial-type secondary defects always have planar character (see Fig. 15). By forming a piece of a new lattice plane that is confined by an edge-dislocation loop, the interstitials are able to relieve most of their strain energy. In metals, the interstitial loops tend to precipitate on the most densely packed planes, e.g., on {111} planes in fcc and on {110} planes in bcc metals. Small loops contain an extrinsic stacking fault. In materials with high stacking-fault energy, large loops tend to unfault and are then called perfect loops [see Fig. 15(b)].

In Si and Ge, interstitial-loop morphologies that result in a minimum number of dangling bonds are observed: rodlike ⟨110⟩ dislocation dipoles and faulted loops on {311} planes.

In polyatomic nonmetals, the precipitation of interstitials requires that both stochiometry and sublattice order must be conserved on the new lattice planes. In MgO or NaCl, this requirement leads to the formation of two-layered loops on {110} planes, as illustrated in Fig. 16. The formation of such two-layered interstitial loops requires equal numbers of anions and cations to precipitate. Since most irradiation conditions—especially radiolyses—produce many more anion than cation interstitials, there must be some mechanism operating by which the missing cations are supplied for the loop formation. In the mechanism proposed by Hobbs *et al.* (1973) and outlined in Fig. 16,

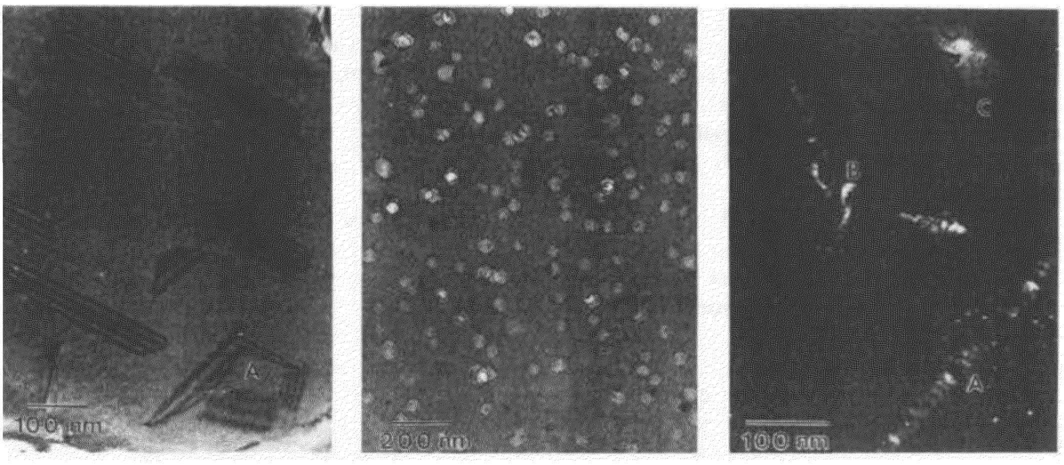

FIG. 14. TEM micrographs of typical radiation-induced secondary-defect structures in metals (W. Kesternich, private communication). Left: faulted interstitial dislocation loops produced in a NiCr-Mo alloy by irradiation with 28-MeV α-particles (dose 3.5×10^{21} α/m^2) and annealing at 750 °C. The loops are cut off by the back and front surfaces of the TEM sample. At *A*, loops on all four possible variants of {111} planes can be recognized under the ⟨110⟩ illumination of the sample. Middle: Phase-contrast picture of faceted voids in Al produced by irradiation with fast neutrons (dose 4.5×10^{25} n/m^2) at 55 °C. Right: radiation-induced Ni_3Si precipitates produced in a Ni-8%Si alloy by irradiation with 28-MeV α particles (dose 3×10^{21} α/m^2) at 475 °C. Bright-field contrast using a superlattice reflection of the ordered precipitates. At positions A, B, and C the precipitates were formed by dragging of Si atoms to grain boundaries, network dislocations, and dislocation loops, respectively.

the cations incorporated into the two-layered loop are supplied by the formation of an equal number of anion antisite defects in the surrounding lattice. To make the energetics favorable, the anion antisites form molecular bonds with neighboring anions on a lattice site.

Vacancy-type secondary defects—in contrast to the interstitial ones—may have different morphologies. One morphology is again a planar condensation of vacancies as a dislocation loop, now encircling a piece of a missing lattice plane [Figs. 15(c) and 15(d)]. In metals with low stacking-fault energy, like Au, stacking-fault tetrahedra may be energetically more favorable than planar loops. Most important, however, is the formation of three-dimensional vacancy ag-

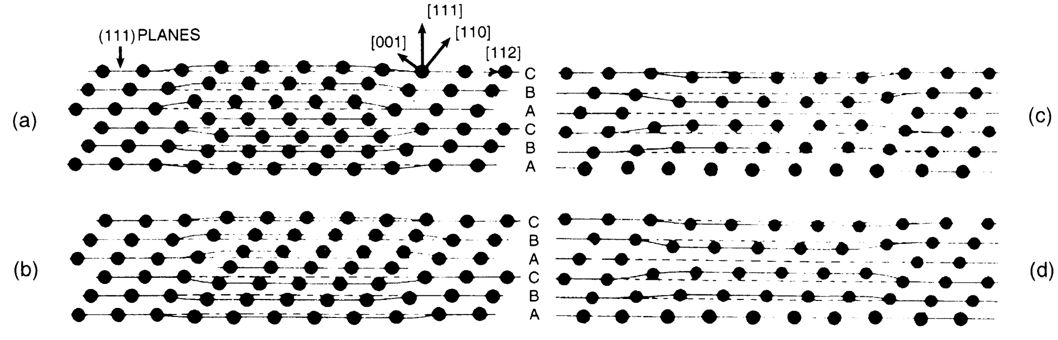

FIG. 15. Cross section along a {110} plane through interstitial- and vacancy-type dislocation loops in an fcc metal. The stacking of {111} planes in the fcc lattice is denoted by ABC. . . . (a) The additional interstitial atoms form a faulted loop with Burgers vector $b = \frac{1}{3}a$ [111]. (b) Perfect interstitial loop with $b = \frac{1}{2}a$ [110]. The stacking fault has been removed by shearing the atomic arrangement by $\frac{1}{6}a$ [112]. (c) Faulted vacancy loop created by removing atoms from a {111} plane. Note the intrinsic stacking fault. (d) Perfect vacancy loop.

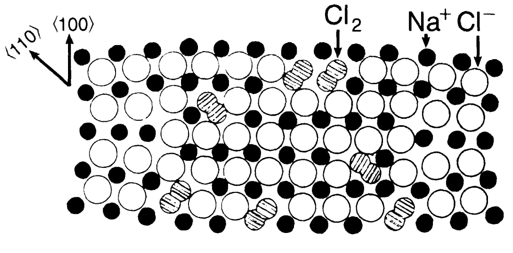

FIG. 16. (001) section through an interstitial dislocation loop with Burgers vector $\frac{1}{2}a\langle 110\rangle$ in NaCl. The two-layered loop contains equal numbers of cations and anions. To form the loop by the condensation of Cl interstitials alone, according to the Hobbs mechanism, the needed Na ions are supplied by the formation of equal numbers of Cl_{Na} antisite atoms shown as Cl_2 molecules replacing a substitutional NaCl molecule.

glomerates in the form of bubbles or voids. Such three-dimensional agglomerates are not possible for interstitials because of strain-energy considerations. In the early stage of the nucleation of voids, some gaseous impurities, such as fission gases, helium atoms from (n,α) reactions, or self-gas molecules (in oxides and halides) are necessary to prevent the collapse into a planar loop. The distinction between void and bubble derives from the gas pressure p_C inside the cavity. If $p_C < 2\gamma/r_C$ the cavity is called a void; otherwise it is a bubble. γ is the specific surface energy and r_C the cavity radius. The anisotropy of γ causes faceting of the voids along crystallographic planes with the lowest surface energy [see Fig. 14(b)].

In polyatomic nonmetals, void formation like planar defects requires the condensation of equal numbers of vacancies on all sublattices in stoichiometric proportions. In spite of this complication, three-dimensional vacancy agglomerates are frequently observed, e.g., in oxides and halides. This is true even under situations where practically only anion vacancies are produced by the irradiation. In this case, colloids are often formed. They can be considered as a local condensation of anion vacancies of neutral effective charge (i.e., containing a trapped electron) with the remaining cations precipitating out as small metal particles. Another possibility is the formation of pressurized self-gas bubbles. Such bubbles can result, e.g., from the precipitation of the molecular anion antisite defects, e.g., Cl_2 produced along with the interstitial loops.

In a broader sense, radiation-induced phase changes such as amorphization and precipitation of alloying atoms should also be included as possible secondary defect structures. For example, even if there is no direct amorphization within displacement cascades, an accumulation of disorder produced by electron irradiations of intermetallic or nonmetallic compounds may trigger a transformation of the damaged lattice into an amorphous phase after a critical irradiation dose has been achieved. In many metallic alloys, a local redistribution of their constituents is observed under irradiation. Examples are the dissolution of small, thermally stable precipitates by the atomic mixing action of displacement cascade or the precipitation of nonequilibrium phases at point-defect sinks, such as grain boundaries, dislocations, and surfaces (radiation-induced segregation; see, e.g., Fig. 14). This precipitation results from the interaction of the impurities with the migrating interstitials and/or vacancies so that the impurities are efficiently dragged to the sinks where they accumulate and finally precipitate.

2.4 Microstructural Evolution during Long-Term Irradiations

The entirety of the secondary defects, together with the network of dislocations, grain boundaries, and precipitates, comprises the microstructure of an irradiated material. Its evolution during long-term irradiations depends on the irradiation conditions (temperature, type, and flux of irradiation particles) and on the sample history (cold worked, annealed, impurity content, etc.). With such variable conditions, quite different scenarios for the buildup of the secondary defects can be envisaged, and, indeed, a large variety of different radiation induced microstructures is observed.

To predict the evolution of radiation-induced microstructures, one must model the nucleation, growth, and final interaction of the different types of secondary defects, taking into account the supersaturations of mobile interstitials and vacancies (and often gas atoms; see Sec. 1.1) that are created continuously as primary defects by the ongoing irradiation. This is a very complicated problem, because the defect supersaturation depends sensitively on the irradiation parameters and on the density and distribution of internal sinks, including the secondary defects that have accumulated from the beginning of irradiation. The usual procedure is to approximate the spatially and temporally discrete events of defect introduction and defect losses by a continuum description using chemical rate equations. This results in a large set of coupled nonlinear first-order differential equations for the time-dependent concentrations of all the different primary and secondary defects. Such equations are difficult to handle, and, quite often, they obscure the basic physics involved. In view of this complexity, we do not include elemental redistributions in the following comments.

In most technical applications, the irradiations are performed at temperatures at which the radiation-induced interstitials and vacancies are both thermally mobile. These primary defects will then either recombine locally at an antidefect generated in the same displacement event, or escape their birthplace before this region suffers another displacement event. The latter defects are called escaping interstitials and vacancies, and only these defects contribute to the microstructural evolution. Typical efficiencies for the production of escaping interstitials and vacancies are $\sim 0.5\xi_{fn}$/dpa.

After the start of an irradiation, there is always a short transient period in which the concentration of mobile defects first increases and then quickly approaches a stationary value. After this transient, the number of escaping defects that are lost per unit time by mutual recombination or annihilation at sinks equals the rate at which they are produced. (This does not mean, however, that at lower irradiation temperatures Frenkel defects can accumulate indefinitely. In this case, Frenkel defect concentrations saturate at atomic fractions $\sim 10^{-3}$ because of spontaneous recombination processes from overlapping cascades.) The time necessary to establish stationary concentrations of escaping defects is of the order of the migration time of the defects to their sinks. Since interstitials generally move much faster than vacancies, the stationary concentrations of escaping vacancies are orders of magnitude higher than those of the escaping interstitials.

After quasisteady state has been established, there are two limiting cases for the subsequent microstructural evolution: At lower and at very high (typically $>0.6\ T_m$) irradiation temperatures, recombination is the dominant loss mechanism for defects (recombination-dominated regime). Here the stationary concentrations of either the escaping or thermal-equilibrium vacancies are much larger than those of the other interstitial sinks. Therefore, the escaping interstitials will primarily recombine at vacancies, and only a few of them will arrive at fixed sinks. In this situation, the microstructural evolution is drastically slowed down.

At intermediate temperatures ($0.3 < T/T_m < 0.6$ in metals), on the other hand, the stationary vacancy concentration is much smaller than that of the other sinks. Then, practically all escaping defects will arrive at secondary defects or other sinks (sink-dominated regime), and they will contribute to the further evolution of the microstructure. Although the escaping defects are collected at the sinks in this case, the structure of these sinks changes only slowly. The reason is that escaping interstitials and vacancies are produced in equal numbers by the displacement events and, therefore, equal numbers of interstitials and vacancies must also arrive at the sinks. Thus, given only one type of sink, e.g., only dislocations or only voids, equal numbers of atoms will be added as interstitials and removed by vacancies. On the average, therefore, no climb of the dislocation or growth of the void can occur. A change in the microstructure can occur only if atoms are redistributed between different sinks under the steady supply of arriving interstitials and vacancies. The basic reason for this redistribution is that some sinks can capture interstitials more efficiently than vacancies. This bias in the sink action may have different origins. For example, dislocations, which have long-range stress fields, will elastically interact more strongly with interstitials than with vacancies, because of the larger lattice strains produced by interstitials. Cavities, on the other hand, which have only small stress fields, will capture interstitials and vacancies with similar efficiencies. If both types of sinks are present, as outlined in Fig. 17(a), there will be an imbalance in

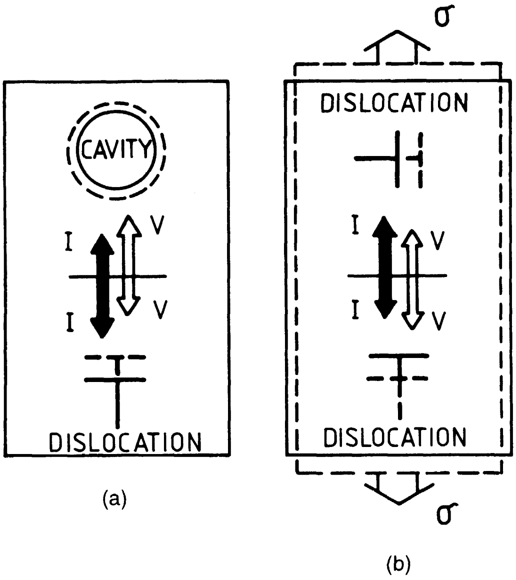

FIG. 17. Schematic representation of the partitioning of the fluxes of radiation-induced interstitials (filled arrows) and vacancies (empty arrows) to different sinks. (a) Cavities and dislocations as sinks: cavities grow by receiving more vacancies than interstitials. (b) Dislocations in different orientations with respect to a uniaxial external stress σ. At $\sigma = 0$: no climb. At $\sigma > 0$: dislocations climb producing a sample elongation along and a shrinkage perpendicular to the stress axis.

the individual fluxes of interstitials and vacancies to dislocations and cavities. As a consequence, the dislocations will climb and the voids will grow (see Sec. 3.1.2). Another reason for a "bias" in the capture of mobile defects may be a poisoning of sinks by deposited impurities or the existence of an asymmetric stress field. As outlined schematically in Fig. 17(b), such a stress field causes elastic polarization effects at the interstitials moving in different directions with respect to the stress axis. This polarization gives rise to different climb velocities of differently oriented dislocations and thus to in-pile creep (see Sec. 3.1.3). Especially large bias effects are expected when the mobile interstitials move one-dimensionally via small glissile loops, since these loops can be attracted at rather large distances from existing line dislocations or larger sessile loops. The rapid growth of these interstitial loops results in a partial incorporation of them into existing grain boundaries and/or their mutual entanglement. The associated dislocation annihilations lead to the formation of a stationary density of network dislocations, independent of the initial dislocation content of the sample. Stationary dislocation densities of the order of 10^{14}–10^{15} m^{-2} are observed in quite different materials under a large variety of irradiation conditions.

Frequently, self organization of the secondary defects is observed during prolonged irradiations. Examples are the formation of dislocation cell structures and defect walls [Fig. 18(a)], the ordering of voids [Fig. 18(b)] and bubbles in the form of a superlattice, and periodic precipitate patterns in decomposing alloys. As in other self-organization phenomena, this instability of the defect pattern arises under special circumstances because of a nonlinear feedback in the defect reactions and because the system is "thermodynamically open," i.e., because a significant amount of energy is pumped continuously into the system by the irradiation.

3. CHANGES OF MACROSCOPIC PROPERTIES

3.1 Metals and Alloys

3.1.1 Hardening and Embrittlement at $T \leq 0.35T_m$ It is often found that the yield stress σ_Y of metals and alloys increases considerably upon irradiation. The increase in yield stress σ_Y, or "radiation hardening," saturates at doses above some dpa. In contrast, the ultimate tensile stress σ_{UTS} increases much less and the work-hardening coefficient is only slightly decreased. The consequence is a reduction in the strain to fracture, ϵ_u, i.e., a loss in ductility upon irradiation (Fig. 19).

In fcc metals and alloys such as austenitic stainless steels, radiation hardening in the temperature range considered is attributed to dislocation loops at low doses and network dislocations at higher doses, acting as obstacles for dislocation movement. Employing dispersed-barrier hardening models, a reasonable agreement between the observed change in microstructure and the measured increase in yield stress can be achieved.

The often very complex behavior of bcc metals and especially alloys such as ferritic/

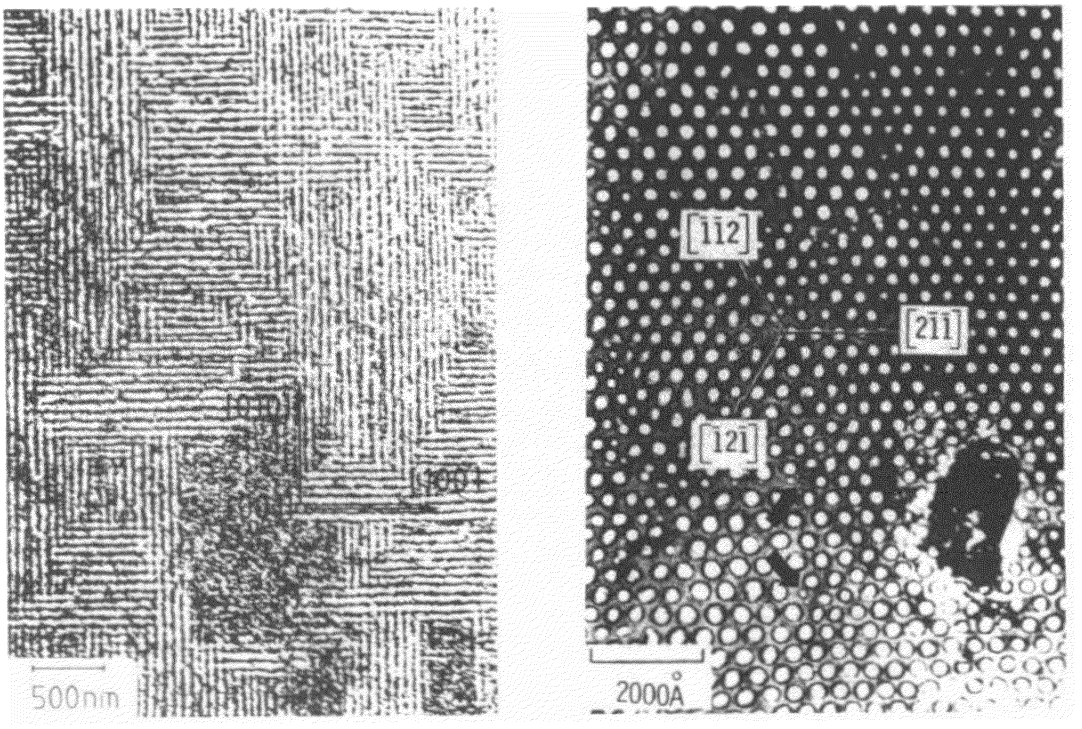

FIG. 18. TEM micrographs of ordered defect structures: (a) {100}-planar defect arrangement produced in copper under 3-MeV proton bombardment at 100 °C to 2 dpa. Regions of high cluster density appear dark (after Jäger *et al.*, 1988). (b) Void lattice in niobium irradiated with 7.5-MeV Ta^+ ions at 800 °C to 300 dpa (after Kulcinsky and Brimhall, 1973). Arrows indicate crystallographic directions within the plane of the micrograph.

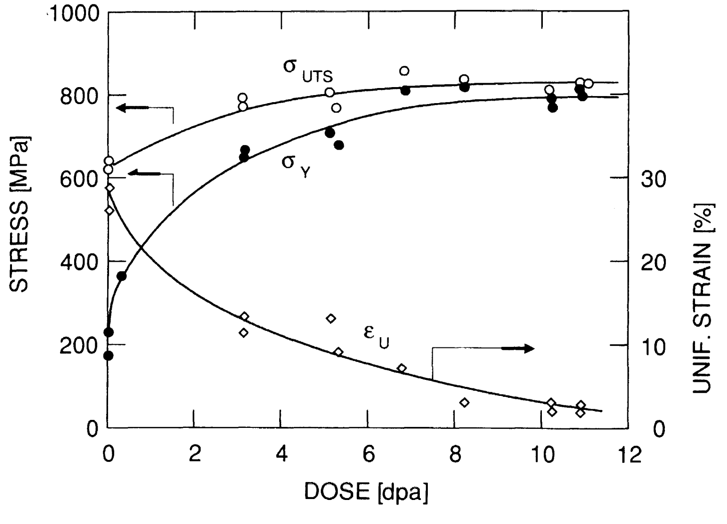

FIG. 19. Yield stress σ_Y, ultimate tensile stress σ_{UTS}, and uniform ductility ϵ_u as a function of dose of type 316 austenitic stainless steels, *neutron irradiated and* tested at 250 °C (after Elen and Fenici, 1992).

martensitic steels or Mo alloys is more difficult to analyze because of two main problems. In most bcc metals, the minor alloying or impurity elements (C, N, O, S, P, Cu, Ni, etc.) have a large influence on both the radiation behavior and the mechanical properties. The other problem is the strong increase of σ_Y with decreasing temperature which is due to the low mobility of screw dislocations in the bcc structure and often interferes with alloy or radiation hardening in a nonadditive way. The bcc-intrinsic low-temperature hardening is also the reason for the well-known ductile-to-brittle transition at a temperature DBTT, below which the yield stress exceeds the (almost temperature independent) cleavage stress. The material then fails by brittle cleavage fracture, which is accompanied by a sharp drop in fracture toughness (Fig. 20). In low-alloy ferritic steels for reactor pressure vessels, the DBTT is around 0 °C. Irradiation increases the temperature-dependent yield stress but not the cleavage stress and therefore causes a shift of the DBTT to higher temperatures. In earlier pressure-vessel steels, this shift can be so large (e.g., Fig. 20, top) that brittle behavior may be encountered in the temperature regime of normal reactor operation or of reactor shutdown. Microstructural studies using field ion microscopy, electron microscopy, and small-angle scattering have identified precipitates rich in Cu and P and microvoids, possibly stabilized by Cu and P, and metal carbides as responsible for the hardening. Modern pressure-vessel steels contain less than 0.1% Cu and 0.01% P, and they exhibit a greatly improved resistance against low-temperature irradiation embrittlement (Fig. 20, bottom). Large DBTT shifts are also of concern for Mo and its alloys, which are considered for heat-sink materials of divertors and limiters in future fusion devices.

3.1.2 Swelling Swelling by the nucleation and growth of voids was discovered by Cawthorne and Fulton (1967) by electron microscopic observations of a stainless steel irradiated in the Dounreay Fast Reactor in Scotland. In the following years, swelling was one of the dominant topics in radiation-damage research since the relative volume increase, $\Delta V/V$, can reach tens of percent (Fig. 21) at the high doses prevailing in advanced reactors. Void swelling is generally observed at irradiation temperatures between 0.3 and 0.6 of the melting temperature, T_m, with a peak around (0.45–0.5)T_m for typical

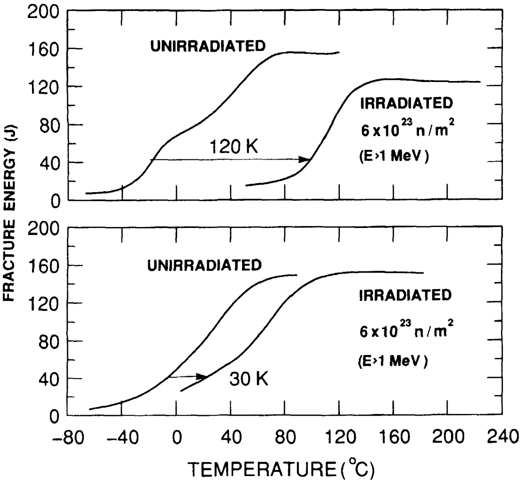

FIG. 20. Charpy V-notch ductility results showing the shift of the ductile-to-brittle transition temperature (DBTT) of A533-B ferritic pressure-vessel steel caused by irradiation to about 0.1 dpa at 290 °C, corresponding to 30 years of reactor operation. Upper plot shows material in which Cu content was not controlled; lower plot shows improved material with low Cu level (after Hawthorne, 1975).

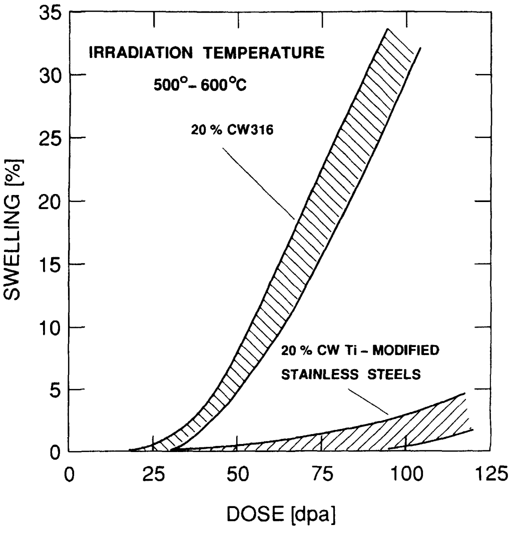

FIG. 21. Swelling as a function of dose of cold-worked 316 stainless steel and a Ti-modified variation irradiated in a fast reactor at 500–600 °C (after Maziasz, 1984).

reactor dose rates of ~10^{-6} dpa/s. Higher dose rates shift the swelling range to higher temperatures. The dose dependence of swelling is characterized by three stages (Fig. 21): an incubation period with little or no swelling, a transition stage, and a high-dose region where $\Delta V/V$ increases approximately linearly with dose, at rates on the order of 1%/dpa for many alloys.

The swelling rate in the linear dose regime is rather insensitive to materials parameters such as alloy composition, density of dislocations and precipitates, helium content, etc. However, these parameters strongly influence the duration of the incubation period. This fact is utilized in the development of swelling-resistant alloys, e.g., the addition of <1% Ti to 316 stainless steel shifts the onset of high swelling from 25 to almost 100 dpa (see Fig. 21). How the crystal structure affects void swelling is not well understood. For the technologically important steels, it is well established that bcc ferritic/martensitic steels are generally much more swelling resistant than fcc austenitic steels. On the other hand, fcc gold is highly resistant to swelling, but bcc vanadium swells readily.

Soon after its discovery, swelling was identified as being the consequence of an imbalance ("bias," see Sec. 2.4) of the fluxes of irradiation-induced vacancies and interstitials to dislocations and void embryos, respectively [Fig. 17(a)]. Rate theories based on this concept (Bullough and Perrin, 1969, and others) were able to predict qualitatively the dependences on temperature, dose rate, and dose observed in the fast-swelling regime. However, to obtain agreement with the observed absolute swelling values, unrealistically low values for the bias factors had to be assumed. More recent recognition of the importance of interstitial loops escaping from displacement cascades via unidirectional glide process has initiated a new modeling effort. This new idea promises to lead to an improved understanding of void swelling, and it seems also to have the potential for treating the incubation and transition stages for which an adequate theory is currently lacking.

Although several swelling-resistant steels are now available for fast-breeder applications, experimental and theoretical research in this field is needed, as swelling promises to be a more severe problem in fusion materials. In these machines, materials will be subjected to displacement doses of more than 100 dpa and a strongly increased helium production from nuclear transmutations induced by the 14-MeV fusion neutrons. Furthermore, the variety of elemental tailoring is restricted by the requirement to use only elements with low radioactive activation.

3.1.3 Irradiation Creep and Growth

Irradiation or in-pile creep is the plastic deformation of materials caused by the simultaneous action of irradiation and mechanical stress. It is observed at temperatures as low as $0.2T_m$. After a transient period (duration corresponding to a dose of ≈1 dpa), a constant irradiation creep rate $\dot{\epsilon}$ is observed that is essentially proportional to the displacement rate K and the applied stress σ (Fig. 22). $\dot{\epsilon}$ depends only weakly on temperature (Fig. 23), in contrast to the thermal creep rate, which increases exponentially with temperature.

The magnitude of irradiation creep depends on materials parameters. For several pure metals and solution-hardened alloys with fcc structure, an empirical correlation with the yield stress of the unirradiated material holds, showing a decrease of the normalized creep rates $\dot{\epsilon}/K\sigma$ from 3×10^{-10} to 1×10^{-11} Pa^{-1} dpa^{-1} when the yield stresses increase from 20 to 200 MPa. These values are insensitive to the initial dislocation struc-

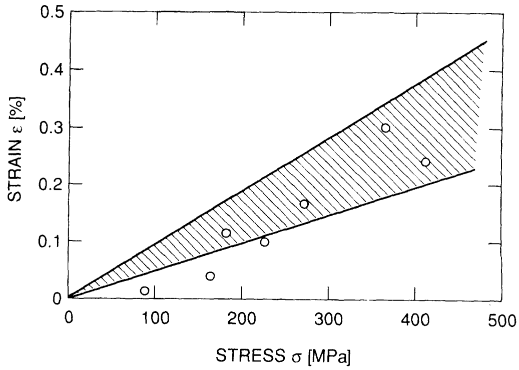

FIG. 22. Stress dependence of strain ϵ due to irradiation creep after n irradiation to 4.8 dpa in a fast reactor (FFTF, cross-hatched region) and a mixed-spectrum reactor (ORR, open circles) in cold-worked 316 stainless steel. The irradiation temperature was 330 °C (after Rowcliffe and Grossbeck, 1984).

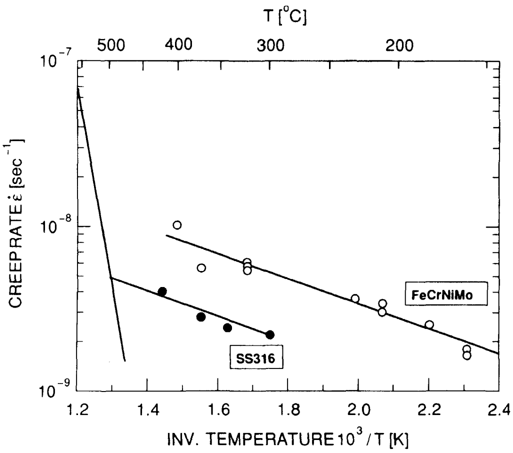

FIG. 23. Temperature dependence of irradiation creep rate $\dot{\epsilon}$ in proton-irradiated FeCrNiMo alloy and in 316 stainless steel. The applied stress was 250 and 200 MPa, respectively. The displacement rate was 10^{-6} dpa/s. The solid line on the left indicates thermal creep in 316 (after Jung and Ullmaier, 1990).

ture and to the presence of precipitates. Similar irradiation creep behavior is found in hcp Zr and its alloys, whereas no enhancement of the creep rates above the thermal values could be detected in bcc martensitic steels under irradiation.

Irradiation creep is thought to be caused by the stress-induced preferential absorption (SIPA—Heald and Speight, 1975) of interstitials and vacancies at dislocations of different orientation with respect to the applied stress [Fig. 17(b)]. The resulting dislocation climb then leads to the observed volume-conserving deformation under load. However, there are experimental results that cannot be explained by SIPA alone, and thus several other mechanisms (e.g., climb-glide, swelling driven, cascade-induced creep) have been proposed to contribute to irradiation creep.

Whereas radiation-induced swelling and embrittlement always have negative consequences on the behavior of materials in nuclear environments, irradiation creep can also have positive aspects, e.g., by relaxing high stresses generated by inhomogeneous swelling and thereby avoiding premature failure.

Irradiation growth is the shape change at constant volume that occurs during irradiation of unstressed materials of noncubic structure. Two technologically important examples are orthorhombic U metal and hcp Zr and its alloys. There is some microstructural evidence that interstitial atoms and vacancies condense unequally on the different crystallographic planes (e.g., on the prism and the basal planes in hcp metals) and most theoretical attempts to model irradiation growth are based on this anisotropy.

3.1.4 High-Temperature Embrittlement

Whereas the embrittlement effects described in Sec. 3.1.1 diminish with increasing temperature, there are other mechanisms that become operative at $T \geq 0.5\ T_m$ and that can lead to severe ductility losses in structural components in fast-breeder and particularly in future fusion reactors. It is now generally agreed that helium created by (n,α) processes is the most critical factor, and other possible mechanisms such as radiation-induced impurity segregation to grain boundaries, second-phase precipitation, voids, etc. are less important.

Under deformation with slow strain rates due to creep or fatigue loads, helium-containing materials often fail by intergranular fracture (Fig. 24). In many alloys, helium concentrations of a few appm are sufficient to induce this brittle fracture mode, which is generally accompanied by a decrease of the elongation and time to rupture (Fig. 25). The extent of this degradation increases with increasing helium concentration and temperature and with decreasing strain rate. The

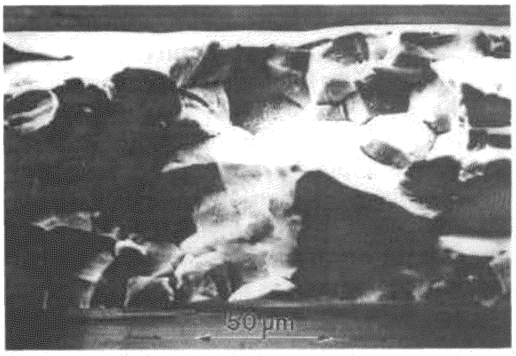

FIG. 24. Fracture surface of an austenitic stainless steel that failed after creep testing at 873 K with simultaneous He implantation. The final He concentration at rupture was 8000 appm (from Schroeder and Stamm, 1990).

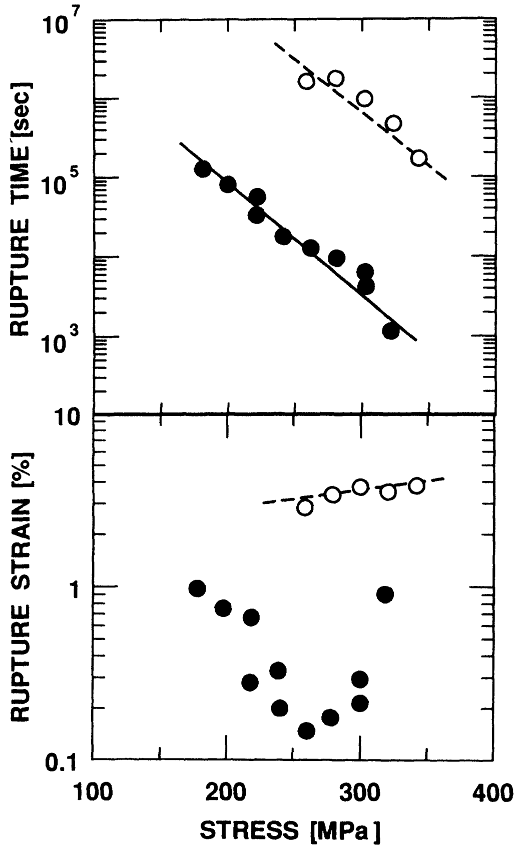

FIG. 25. Degradation of rupture time (top) and ductility (bottom) of cold-worked 316 stainless steel under creep stress at 873 K: ○ reference specimens, ● creep tests with simultaneous He implantation (after Yamamoto and Schroeder, 1988).

vulnerability of materials to "helium embrittlement" strongly depends on their initial composition and microstructure. Among the technologically important materials, Ni-based superalloys are most affected, whereas Fe-Cr ferritic/martensitic steels are resistant up to 3000 appm He and 700 °C. In austenitic steels, the addition of some minor alloying elements (e.g., Ti, which forms finely dispersed TiC precipitates) can greatly reduce the detrimental effects of helium.

There is ample experimental evidence that the microstructural cause for helium embrittlement is helium bubbles that preferentially nucleate in grain boundaries. With ongoing irradiation the bubbles grow by absorbing the continuously produced He until they initiate or accelerate grain-boundary cracks. At very high He concentrations, they perforate the grain boundaries until their contact area becomes too small to support the applied stress. The theory of helium embrittlement describing the bubble evolution in grain boundaries is rather well advanced. The available models (Trinkaus, 1986; Trinkaus and Ullmaier, 1994) are able to identify the failure criteria for the relevant ranges of temperature, stress, strain rate, and He production rate and to reproduce the observed dependencies of the creep and fatigue lifetimes on these parameters.

The main impetus for future research on high-temperature embrittlement derives from materials development for fusion power reactors. Here the availability of embrittlement-resistant alloys fulfilling several other requirements (e.g., low radioactive activation, low swelling) will be essential for a safe and economic operation of these future energy sources.

3.2 Carbon-Based and Ceramic Materials

3.2.1 Dimensional Changes and Stored Energy In spite of the complications discussed in Sec. 2.3, the basic high-dose damage mechanisms discussed for metals apply also to this class of materials and swelling is observed for many insulators as well. Here we restrict the discussion to two specific topics.

For many noncubic polycrystalline materials, internal strains are built up in addition due to anisotropic growth. The magnitudes of these dimensional changes per unit dose are most important in graphites [Fig. 26(a)]. Electron microscopy observations and a comparison of the changes in length and lattice parameter show that the growth in the *c*-axis direction is due to the growth of relatively large planar clusters of interstitial atoms, which are inserted between the existing basal planes (Kelly, 1981). The basal-plane contraction is attributed to irregular lines of the corresponding vacancies, which collapse when they reach a certain length.

Carbon-based materials used in technology (polycrystalline "graphite," pyrolytic carbon, carbon-fiber enforced carbon, etc.) are a complex and not well definable mixture of graphite crystallites of different size, density, and texture, binder materials, and pores. Ac-

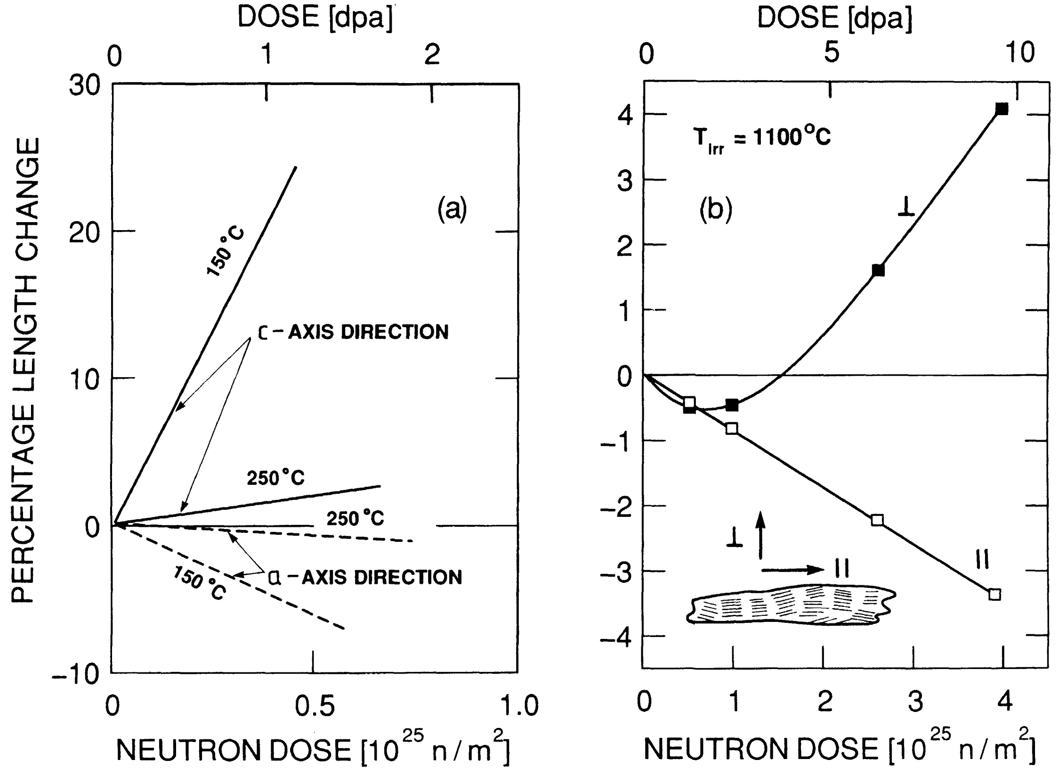

FIG. 26. Dose dependence of length changes for (a) single crystals of graphite (after Kelly, 1981) and (b) polycrystalline blocks parallel and perpendicular to the orientation of the elongated grains (after Delle *et al.*, 1974).

cordingly, the irradiation-induced dimensional changes are also very complex. Qualitatively, the dose dependence exemplified in Fig. 26(b) is typical for many carbon materials: At low doses, the material shrinks (densifies) both perpendicular and parallel to the mean orientation of the usually elongated grains. With increasing dose, the parallel dimension continues to shrink, at a decreasing rate, whereas the perpendicular dimension starts to grow. Dimensional changes under various irradiation conditions have been extensively studied for a great variety of carbon-based materials, mainly because of their importance as structural core and moderator materials in high-temperature gas-cooled reactors.

In any material, the production of defects by irradiation results in an increase of internal energy, which is stored mainly as strain energy around the lattice defects. If, upon heating, the defects annihilate or regroup, this stored energy (Wigner energy E_W) is released and can lead to a fast rise in temperature. This effect is particularly pronounced in graphites where the formation energy of Frenkel defects is high (~10 eV), and the specific heat c_p is small (~10^{-4} eV/atom·K). In graphites heavily irradiated around ambient temperatures, the energy is released upon heating in a rather narrow recovery stage, somewhat below 200 °C, and it is sufficient to raise the temperature spontaneously to 300–500 °C.

The Wigner energy rapidly decreases with increasing irradiation temperature because the stationary concentration of defects given by the balance between their production and their annihilation decreases with increasing *T*. Stored-energy effects have thus been investigated mainly in connection with the design and operation of early graphite-moderated reactors, which were operated at lower temperatures. From these studies, criteria were defined that ensure that the effective specific heat $c_p - dE_W/dT$ of the graphite will remain positive, i.e., self-sustaining releases cannot occur.

3.2.2 Thermal Conductivity and Mechanical Strength The applicability of graphite and ceramic materials for structural components in nuclear technology is often characterized by a "figure of merit," given by $\lambda\sigma_B/\alpha E$. Whereas the coefficient of thermal expansion α and Young's modulus E are rather insensitive to irradiation, considerable reductions of the breaking strength σ_B and particularly of the thermal conductivity λ are observed. λ decreases rapidly even at low doses [Fig. 27(a)] and then saturates at a level roughly between 10% and 20% of the original value for most materials investigated (graphites, Al_2O_3, SiC, AlN, MgO, etc.). The degradation of the thermal conductivity is attributed to phonon scattering at irradiation-induced point defects and their clusters. Since this contribution to the heat-flow resistance depends only weakly on temperature, the initial T^{-1} dependence of λ due to phonon-phonon interactions changes to an almost T-independent behavior. At very high irradiation and/or test temperatures, the λ values of irradiated materials merge into the curve for the unirradiated material because of defect annealing [Fig. 27(a)].

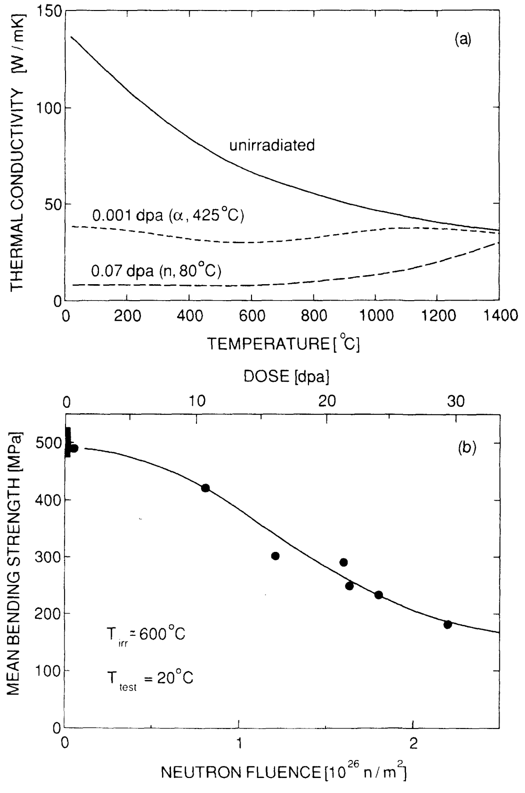

FIG. 27. (a) Decrease of thermal conductivity of isostatically hot-pressed SiC as a function of temperature due to irradiation with α particles and reactor neutrons, respectively. (b) Decrease of mechanical strength of the same material as a function of neutron fluence (after Dienst, 1992).

In contrast to the sharp decrease of λ at low doses, a threshold dose of several dpa is required before noticeable decreases in the breaking strength σ_B [Fig. 27(b)] are observed. As before, the extent of the degradation decreases with increasing irradiation and/or test temperature. Because microstructural information is lacking, no detailed explanation of the failure degradation in terms of the underlying physical processes can be offered. For a number of ceramics, however, a clear correlation between the decrease in σ_B and swelling has been found. It is also of technological importance that the irradiation-induced decrease in strength is accompanied by a considerable broadening of the strength distribution (Dienst, 1992).

3.3 Electronic and Optical Materials

Degradation of the electrical and optical performance of devices is observed at much lower defect concentrations than those discussed above for the mechanical properties. Hence, defects introduced during the processing (e.g., ion implantation) must be annealed out at high temperatures in an additional processing step. The particle fluences received during a one-year operation period in space [typically 10^{17} protons/m^2 (≥ 10 MeV) and 10^{20} electrons/m^2 (≈ 1 MeV) in a near-Earth orbit] may be considered as a typical example for the fluences that are of concern at the operation temperature of devices.

Although not specifically a topic for this section, some remarks on ionization are in order since it may influence device operation as a result of photocurrents, e.g., within the field of built-in p-n junctions. The total number of steady-state e^--h pairs produced by illumination or irradiation of doped semiconductors rarely exceeds that of the minority carriers, and therefore these currents are important only for unusual conditions, e.g., nuclear test explosions, or pulsed accelerators. Local currents of the same type may lead, however, to a discharge of built-in condensers (e.g., "single event upset") or to similar malfunctions in imaging devices (i.e., "white-out").

Device operation can also be influenced by the buildup of a net positive charge in oxide films since it can produce an inversion layer in the silicon. These inversion layers give rise to leakage currents. Similar malfunctions may arise from the disruption of chemical bonds at interfaces between semiconductors and insulators. In optical devices, the trapping of e^- or h may change the charge state of existing defects or dopants, which in turn changes the absorption band of these defects (color centers); as detrapping is efficient only above room temperature, these changes are especially important for optical waveguides and fibers.

3.3.1 Electrical Properties Irradiation-induced point defects in semiconductors generally seem to provide deep levels and act as trapping and scattering centers for the electrical carriers. With increasing fluence the carrier lifetime, the carrier concentration, and the carrier mobility are reduced appreciably. The consequences of the reduced carrier lifetime, or reduced carrier diffusion length, on the performance of silicon solar cells are shown in Fig. 28. The observed degradation shows characteristic trends that are also observed in GaAs solar cells; however, the quantitative changes depend on the detailed structural features of the devices, the materials, and the dopings. Databases have been collected on these characteristics.

Carrier removal rates (change in the number of carriers per m^3 per incident particle per m^2) may be of the order of 10^2 m^{-1} for MeV electrons and of the order of 10^3 m^{-1} for n irradiations at room temperature. After irradiation to high fluences, the defects control the position of the Fermi level, and the number of carriers, e.g., in Si (n and p type) as well as in GaAs, reaches intrinsic values, Ge becomes p type with a limiting hole concentration of $\approx 5 \times 10^{22}$ m^{-3}; InP, InSb, AlSb, InAs, CdTe, HgCdTe become n type and GaSb becomes p type.

For insulators, an increase of the ionic conductivity is often observed at higher temperatures as a result of the production of va-

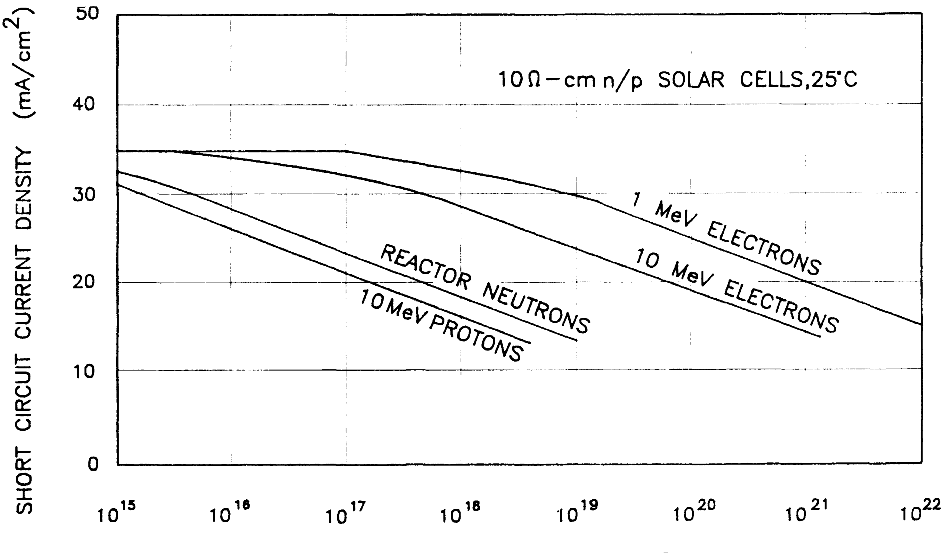

FIG. 28. Degradation of silicon solar cells with fluence for various types of irradiation. The decrease of the short-circuit current density under a solar simulator illumination (135 mW/cm^2) is shown for fluences of protons and electrons that are typically absorbed during the lifetime of a device in a near-Earth orbit (after Carter and Tuda, 1973).

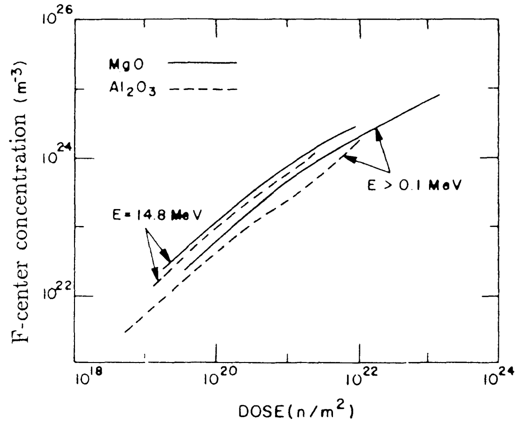

FIG. 29. Increase of the concentration of *F* centers in Al_2O_3 and MgO as a function of the fission and fusion neutron fluence. The *F*-center concentration was determined from the optical absorption within the absorption band around a photon energy of 5 eV (after Clinard and Hobbs, 1986).

cancies. An additional radiation effect, which can be important for rf windows in fusion reactors, is the increase of the dielectric loss at microwave frequencies (in addition to the loss that is proportional to the ionization-induced conductivity), as defects can act as resonance absorption centers. Displacement damage increases the dielectric loss, particularly at frequencies below 1 MHz (e.g., the loss tangent at 10^4 Hz is increased in Al_2O_3 by an order of magnitude after 10^{22} n/m^2; Fowler, 1984). In addition, defects yield changes of the dielectric breakdown fields, although the influence of surface defects and cracks is more serious for high-dose application.

3.3.2 Optical Properties Defects with electronic levels within the forbidden gap may be optically excited with bandgap light and are characterized by the corresponding absorption band or thresholds. Noteworthy examples are the vacancy-related color centers, particularly the *F* centers. Since even small changes of the absorption coefficient are unacceptable for optical fibers, materials must be carefully selected considering the optical wavelength of the signals and potential defect absorption bands.

The increase of the *F*-center concentration in MgO and Al_2O_3 for neutron fluxes that might be characteristic for window materials of future fusion reactors is shown in Fig. 29. Although there is an indication of a saturation behavior for the concentration of the centers, this behavior does not necessarily reflect the total damage, as vacancies may react and agglomerate during irradiation and no longer be optically detectable. In addition to these absorption bands, the overall absorption background increases, especially close to the gap energy.

GLOSSARY

Amorphization: Formation of amorphous phases. Special for irradiation: freezing-in of liquidlike "glassy" structures during rapid spike cooling or amorphization by defect accumulation. Amorphous zones are nonequilibrium phases that recrystallize upon annealing at elevated temperatures.

Bias: Imbalance of fluxes of interstitials and vacancies, respectively, to sinks of different capture strengths.

Bubble: Cavity filled with gas under high pressure.

Cluster: Generic term for (small) agglomerate of defects.

Damage Efficiency ξ: Number of stable Frenkel pairs per dpa. For small recoil energies $\xi \approx 1$; for very high recoil energies ξ approaches a value of $\xi_{fn} \approx 0.25$ for fast neutrons.

Damage Energy T_{dam}: Portion of the recoil energy that is available for atomic displacements (and not converted into electronic excitations).

Depleted Zone: Vacancy-rich core of displacement cascade after cooling to ambient temperature.

Dislocation Loop: Planar agglomerate of interstitial atoms or vacancies, respectively, which is equivalent to a closed dislocation line.

Disordering: Destruction of an ordered phase. Special for irradiations: freezing-in of disordered atomic structure during rapid spike cooling or disordering by defect accumulation.

Displacement Cross Section σ_{dam}: Number of displacements per atom (dpa) per unit fluence.

Displacement Energy: Energy required in order to produce a stable Frenkel pair.

Displacements Per Atom (dpa): Calculated number of displaced atoms per lattice atom [see Eq. (2)] produced by a fluence Φ of incident particles per unit area. Using dpa as a dose unit facilitates the comparison of displacement damage caused by irradiations with different particles and energies.

Ductile-to-Brittle Transition Temperature (DBTT): Temperature where the ductile failure of bcc metals and alloys (at high temperatures) changes to brittle behavior (at low temperatures).

Electronic Excitations: Energy losses of charged particles causing excitation, ionization, and charge transfer in the electronic system of irradiated matter.

Frenkel Pair (FP): Crystal lattice defect consisting of interstitial atom and vacancy.

***F* Center:** Negative-ion (anion) vacancy with trapped electron.

Ion Track: Region around the path of a GeV heavy ion with very high electronic excitation.

MDS: Molecular dynamics computer simulation. In these simulations, the classical equations of motion are solved for an assembly of atoms interacting via a suitably chosen interatomic potential. The motion of all atoms can be followed as a function of time after a recoil energy T has been given to one of the atoms.

Migration Energy: Activation energy necessary for a thermally activated jump of a defect (see Table 2).

Nuclear Transmutation: Change of a chemical element by a nuclear reaction with irradiation particles, especially neutrons.

Primary Knock-On Atom (PKA): Lattice atom that has made a direct collision with the irradiating particle and has obtained a recoil energy T.

Radiolysis: Production of stable Frenkel pairs as a consequence of the decay of electronic excitations in insulators.

Rate Equations (Chemical): Continuum description of the defect introduction rates and defect losses by recombination and capture at different secondary defects and sinks. This generally leads to a coupled set of nonlinear first-order differential equations.

Recoil Energy: Energy transferred to a lattice atom by a collision with the irradiation particle; the maximum recoil energy is given by Eq. (1).

Recombination-Dominated Regime: Regime on the temperature (and flux) coordinate where during continuous irradiation the dominating fraction of the produced interstitial atoms recombines with vacancies (or vice versa) before they can reach other defect sinks.

Recombination Volume (Spontaneous): Volume $V_{s.r.}$ around an interstitial atom (or vacancy) where the antidefect is not stable and can recombine athermally.

Recovery Stages: The annealing or recovery of a damaged crystal usually proceeds in several steps as a function of the annealing temperature. These steps or stages can be related to different defect reactions.

Replacement Collision Sequence (RCS): A linear chain of collisions along a close-packed lattice direction whereby one colliding atom replaces its neighbor.

Sink-Dominated Regime: Regime on the temperature (and flux) coordinate where during continuous irradiation the stationary vacancy concentration is much smaller than that of other sinks. Hence, the major part of the mobile interstitial atoms arrives at external sinks.

Spike (Thermal): Local region where the majority of atoms are in violent motion and where the kinetic and potential energies are randomized and converted to heat. The thermal spike is observed typically 0.3 ps after the start of a collision cascade.

Stress-induced Preferential Absorption (SIPA): Mechanism proposed for the explanation of the creep processes observed under irradiation. The absorption of mobile vacancies and interstitial atoms at dislocations depends on the orientation of the dislocations with respect to the applied stress.

Threshold Energy: Minimum energy that has to be transferred to a PKA in order to produce a stable Frenkel pair (see Table 2).

Void: Three-dimensional agglomerate of vacancies that may attain the shape of faceted spheres for large sizes.

Wigner Energy: Internal energy stored in the form of lattice defects.

Works Cited

ASTM (American Society for Testing and Materials) (1989), Standard Practice for Neutron Damage Simulation by Charged-Particle Irradiation, ASTM/E 521-89, p. 187.

Averback, R. S., Benedek, R., Merkle, K. L. (1978), *Phys. Rev.* **B18,** 4156–4171.

Bullough, R., Perrin, R. C. (1969), in: *Radiation Damage in Reactor Materials,* IAEA, Vienna, Report No. STI/PUB/230, p. 233.

Carter, J. R., Tada, H. Y. (1973), *Solar Cell Radiation Handbook,* National Aeronautics and Space Administration Report No. NASA-CR-133738.

Cawthorne, C., Fulton, E. J. (1967), *Nature* (London) **216,** 575–576.

Clinard, F. W., Hobbs, L. W. (1986), in: R. A. Johnson, A. N. Orlov (Eds.), *Physics of Radiation Effects in Crystals,* New York: Elsevier Science Publ., p. 387.

Delle, W., Haag, G., Nickel, H., Schulze, H. A., Schulz, R. E. (1974), in: *Proceedings of the 4th International Conference on Carbon and Graphite* London: Society of Chemical Industry, p. 741.

Diaz de la Rubia, T., Guinan, M. W. (1991), Lawrence Berkeley Laboratory Report No. UCRL-JC-107488.

Dienst, W. (1992), *J. Nucl. Mater.* **191–194,** 555–559.

Elen, J. D., Fenici, P. (1992), *J. Nucl. Mater.* **191–194,** 766–770.

Foreman, A. J. E., English, C. A., Phythian, W. J. (1992), *Philos. Mag. A* **66,** 655–669.

Fowler, J. D. (1984), *J. Nucl. Mater.* **122/123,** 1359–1364.

Hawthorne, J. R. (1975), ASTM-STP-570, Philadelphia: American Society for Testing and Materials, p. 83.

Heald, P. T., Speight, M. V. (1975), *Acta Metall.* **23,** 1389–1399.

Heinisch, H. L. (1990), *Radiat. Eff. Defects Solids* **113,** 53–73.

Hobbs, L. W., Hughes, A. E., Pooley, D. (1973), *Proc. R. Soc. London, Ser. A* **332,** 167–185.

Jäger, W., Ehrhart, P., Schilling, W. (1988), *Diffus. Defect Data Pt. B* **3/4,** 279–294.

Jung, P., Ullmaier, H. (1990), *J. Nucl. Mater.* **174,** 253–263.

Kelly, B. T. (1981), *Physics of Graphite,* London and New York: Applied Science Publishers.

Kinchin, G. H., Pease, R. S. (1955), *Rep. Prog. Phys.* **18,** 1–51.

Kinney, J. H., Guinan, M. W., Munir, Z. A. (1984), *J. Nucl. Mater.* **122/123,** 1028–1032.

Kulcinski, G. L., Brimhall, J. L. (1973), ASTM-STP 529, Philadelphia: American Society for Testing and Materials, p. 258.

Maziasz, J. (1984), *J. Nucl. Mater.* **122/123,** 472–486.

Narayan, J., Oen, O. S., Fathy, D., Holland, O. W. (1985), *Mater. Let.* **3,** 67–72.

Rowcliffe, A. F. Grossbeck, M. L. (1984), *J. Nucl. Mater.* **122/123,** 181–190.

Schilling, W., Ullmaier, H. (1993), in: R. W. Cahn (Ed.), *Nuclear Materials,* Materials Science and Technology, Vol. 10, Weinheim and New York: VCH-Verlag.

Scholz, R., Vetter, J., Hopfe, S. (1993), *Radiat. Eff. Defects Solids* **126,** 275–278; and private communication.

Schroeder, H., Stamm, U. (1990), ASTM-STP 1046, Philadelphia: American Society for Testing and Materials, p. 223.

Seitz, F., Koehler, J. S. (1956), in: F. Seitz, D. Turnbull (Eds.) *Solid State Physics,* Vol. 2, New York: Academic, p. 307.

Trinkaus, H. (1986), *Radiat. Eff.* **101,** 91–107.

Trinkaus, H., Ullmaier, H. (1994), *J. Nucl. Mater.* **212–215,** 303–309.

Watkins, G. D. (1975), in: F. A. Huntley (Ed.), *Lattice Defects in Semiconductors,* Institute of Physics Conference Series No. 23, London: IOP, p. 1–22.

Yamamoto, N., Schroeder, H. (1988), *J. Nucl. Mater.* **155/157,** 1043–1048.

Further Reading

Corbett, W. J. (1966), "Electron Radiation Damage in Semiconductors and Metals," in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics,* Suppl. 7, New York and London: Academic Press.

Johnson, R. A., Orlov, A. N. (Eds.) (1986), *Physics of Radiation Effects in Crystals,* Amsterdam: North Holland Phys. Publ.

Kelly, B. T. (1981), *Physics of Graphite,* London and New Jersey: Appl. Sci. Publ.

Mansur, L. K. (1988), "Mechanics and Kinetics of Radiation Effects in Metals and Alloys," in: G. R. Freeman (Ed.), *Kinetics of Nonhomogeneous Processes,* New York: Wiley-Interscience.

Nolfi, F. V. (Ed.) (1983), *Phase Transformations during Irradiation,* London and New York: Applied Science Publishers.

Peterson, N. L., Harkness, S. D. (Eds.) (1976), *Radiation Damage in Metals,* Metals Park, OH: American Society for Metals.

Schilling, W., Ullmaier, H. (1993), Physics of Radiation Damage in Metals, in: R. W. Cahn, P. Haasen (Eds.), *Nuclear Materials,* Materials Science and Technology Vol. 10, Weinheim and New York: VCH Verlag.

Spohr, R. (1990), *Ion Tracks and Microtechnology. Principles and Technological Applications,* Wiesbaden, Germany: Vieweg.

Szenes, G. (Ed.) (1992), *Physics of Irradiation Effects in Metals,* Materials Science Forum Vol. 97–99, Zürich: Trans Tech Publications (Recent Conference Volume).

Tada, H. Y., Carter, J. R., Anspaugh, B. E., Downing, R. G. (1982), *Solar Cell Radiation Handbook,* 3rd ed., JPL Publ. 82-69.

Ullmaier, H. (Ed.) (1991), *Landolt-Börnstein:*

Numerical Data and Functional Relationships in Science and Technology, Group III, *Crystal and Solid State Physics*, Vol. 25, *Atomic Defects in Metals*, Berlin: Springer Verlag.

Ullmaier, H., Schilling, W. (1980), "Radiation Damage in Metallic Reactor Materials," in: *Physics of Modern Materials*, Vol. I, IAEA, Vienna, Report No. SMR-46/105, p. 301.

Van Lint, V. A., Flanagan, T. M., Leadon, R. E., Naber, J. A., Rogers, V. C. (1980), *Mechanisms of Radiation Effects in Electronic Materials*, New York: Wiley.

RADIATION DETECTORS

See DETECTORS, SCINTILLATION; DETECTORS, SEMICONDUCTOR

RADIATION DETECTORS, INFRARED

Brenda P. Winnewisser, *Physikalisch-Chemisches Institut, Justus-Liebig-Universität, Giessen, Germany*

Introduction 459
1. **General Aspects of Infrared Radiation Detection** 460
1.1 Thermal Effects 460
1.2 Photoacoustic Effect 460
1.3 Photon Effects 460
1.4 Rectification 460
1.5 Charge-Coupled Devices (CCDs) 461
2. **Detector Performance Criteria** 461
2.1 Noise in Detectors 461
2.1.1 Thermal or Johnson Noise 461
2.1.2 Background Photon Noise 461
2.1.3 Emission Photon Noise 461
2.1.4 Current or Shot Noise 461
2.1.5 Generation–Recombination Noise 461
2.1.6 Phonon or Temperature Noise 462
2.1.7 1/*f* and Amplifier Noise 462
2.2 Responsivity 462
2.3 Detectivity 462
2.4 Quantum Efficiency 463
2.5 Frequency Response 463
2.6 Linearity and Dynamic Range 463
2.7 Geometry 463
2.8 Stability 463
3. **Types of Detectors** 463
3.1 Thermal Detectors 463
3.1.1 Bolometers 463
3.1.1.1 Semiconductor Bolometers 464
3.1.1.2 Thermistors 464
3.1.2 Pyroelectrics 464
3.1.3 Thermopile 464
3.1.4 Golay cell 464
3.2 Photoacoustic Detectors 465
3.3 Photon Detectors 465
3.3.1 Photoconductive Detectors 465
3.3.2 Hot-Carrier Photoconductors 466
3.3.3 Photovoltaic Detectors 467
3.3.4 Photoemissive Detectors 467
3.3.5 Photodiodes 468
4. **Detector Selection for Application to Broad-Banded Spectroscopy** 469
4.1 Submillimeter Region 469
4.2 Far Infrared 469
4.3 Mid-Infrared, 400–2000 cm^{-1} 469
4.4 Mid-Infrared, 2000–4000 cm^{-1} 469
4.5 Near Infrared 469
4.6 Visible 469
5. **Concluding Remarks** 470
Acknowledgments 470
Glossary 470
Works Cited 471
Further Reading 471

INTRODUCTION

The large proportional change in quantum energy and wavelength across the infrared (IR) region, defined in wavelength as 1 to 300 μm, or in terms of wave number 30 to 10 000 cm^{-1}, leads to the need for a variety of detection techniques over this range. The low power of most laboratory IR sources, and the ubiquitous and strong background in most environments, have made the development of sensitive detectors a particularly important aspect in IR applications. Solid-state semiconductor developments have revolutionized infrared detection since the 1950s.

An infrared detector is a device that transforms infrared radiation into a signal, today universally taken to be an electrical signal. This contribution reviews the mechanisms exploited for achieving this and the criteria used for evaluating a given detector element. The categories of currently available detector types are introduced. Finally, for each region within the IR, suitable detectors for one basic application are specified.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

The material covered for the far infrared (FIR), which overlaps the submillimeter-wave region, is distinguished from that covered in the entries for RADAR and RADIOASTRONOMY by the limitation here to incoherent detectors, which do not retain phase or frequency of detected radiation, whereas receivers for radar and radioastronomy rely on coherent, or heterodyne, detection. The principles and materials for the near infrared (NIR) are relevant also for the visible. Applications of IR detectors will be found in the entries CHARGE-COUPLED DEVICES and SENSORS, INFRARED.

1. GENERAL ASPECTS OF INFRARED RADIATION DETECTION

The mechanisms of detection specific to infrared radiation follow from the fact that the quantum energies involved are those of the vibrations of molecules, or, in solid-state devices, the energies are of the order of band gaps characteristic of many semiconductors. Although the detector in principle is just the fundamental element acting as a transducer, the detector generally includes a heat sink, a nonreflective coating, and electrical leads.

1.1 Thermal Effects

Infrared radiation was historically first detected by the thermal effect: Radiation is absorbed nonselectively by a black (or blackened) detecting element. This energy is then equilibrated by relaxation or collision into the translational degrees of freedom, thus appearing as heat. The heat (historically itself the signal) is transformed into an electrical signal by exploiting the temperature dependence of an electrical property of the detecting element, or of a secondary element, as a component in an electrical amplification circuit. The thermal effect may be considered to be a macroscopic effect, and is therefore broad banded. Thermal detectors include bolometers, thermistors, pyroelectric devices, thermopiles, and Golay cells.

1.2 Photoacoustic Effect

The photoacoustic or *optoacoustic effect* goes back even further than the photoelectric effect (Bell, 1881; Tyndall, 1881; Roentgen, 1881). Although with a black absorber it can be used for what is then called a photothermal device, the important application is the absorption of modulated radiation selectively, due to quantum transitions, by a sample. The absorbed energy is equilibrated, inducing a temperature change in a gas, which is exploited to generate an electrical signal in the acoustic frequency range. The effect is not limited to IR radiation, but is employed in the IR because for some applications it competes successfully with other detectors in this region of the spectrum.

1.3 Photon Effects

If the broad-banded absorption of infrared quanta by electrons in the detector material influences directly the properties of an element in an electrical circuit, we speak of a photon effect. Such specifically quantum effects are usually limited in the wavelength range of their effective use by a quantum cutoff. Photon detectors include photovoltaic devices, photoconductive and hot-electron detectors, photoemissive detectors, and photodiodes. Josephson-junction and superconductor–insulator–superconductor (SIS) devices are also photon detectors, but are only employed for coherent detection and will not be discussed here.

1.4 Rectification

The semiconductor diodes used from the radio-frequency region up into the far infrared (FIR), including Schottky-barrier diodes and metal–insulator–metal (MIM) diodes, function as mixers that preserve frequency and phase information in coherent detection systems, and will not be discussed explicitly here. However, used directly as *rectifiers*, point-contact diodes were used until the advent of the InSb hot-electron detector as incoherent detectors in the submillimeter range up to about 600 GHz. Schottky-barrier diodes are also used in the visible. The mechanism of rectification can be seen as the motion of charge carriers in the oscillating electric field of the incident radiation; current is induced in the device at the frequency of the radiation, while the nonlinearity of the voltage–current relationship results in a dc component.

1.5 Charge-Coupled Devices (CCDs)

CCDs have since 1970 come to replace many other detector systems, but, most important, they have been used for measurements not deemed realistic to attempt previously. Semiconductor photon-detector elements are deposited into a continuous array on a large semiconductor chip, creating a grid of localized potential wells and read-out conduction paths, allowing a spatial and temporal resolution not approached by any other type of system. These devices are covered in a separate article (see CHARGE-COUPLED DEVICES).

2. DETECTOR PERFORMANCE CRITERIA

No one parameter characterizing detector performance is sufficient to determine the suitability of a given detector to a given application. The definitions and some limitations on the values of the most important parameters are sketched here.

2.1 Noise in Detectors

The *noise-equivalent power* (NEP) is the rms value of the modulated incident radiant power required to produce an rms signal-to-noise (S/N) ratio of unity. The value is for a specific (postdetection) bandwidth of the modulation frequency, detector area, and field of view, and is given in watts. This parameter is determined by both the response of the detector and its noise characteristics.

Noise can be classified according to its source, so that, in theory, the square of the noise power can be written as the sum of various components:

$$
\begin{aligned}
(\text{noise})^2 = {} & (\text{thermal or Johnson noise})^2 \\
& + (\text{photon noise: background and thermal-emission photon noise, exhibited as shot noise or generation–recombination noise})^2 \\
& + (\text{shot noise due to bias current})^2 \\
& + (\text{phonon or temperature noise})^2 \\
& + (1/f \text{ and amplifier noise})^2.
\end{aligned}
$$

2.1.1 Thermal or Johnson Noise The minimum noise in a detector is that inherent in the resistive element, and is proportional to the resistance and the temperature. The temperature can be reduced by cooling the detector, but reducing the resistance generally reduces also the responsivity.

2.1.2 Background Photon Noise *Photon noise* is due to the uncertainty in the arrival time of photons at the detector, and characterizes the source noise (Rieke, 1994, p. 49). For the IR region, especially in the region of the room-temperature background-emission maximum, this is a critical quantity. It is reduced by limiting the field of view of the detector by a cooled aperture, or by limiting the background-noise bandwidth with a band-pass filter—indeed, particularly in the FIR, by using a cold band-pass filter. Since it is proportional to detector size, minimum detector size is again advantageous.

If the incident photon flux due to the signal, rather than the background, is the dominant source of noise, we speak of "photon-limited" operation of photovoltaic or photoconducting detectors. Under these conditions they exhibit the highest sensitivity.

2.1.3 Emission Photon Noise The photon noise due to thermal emission or generation of carriers in the detector itself is seldom dominant, but since it adheres to Planck's law, it is reduced by cooling the element. The thermally generated carriers are responsible for the dark current in photoconductors. This noise is also proportional to the area of the detector.

2.1.4 Current or Shot Noise Any system with current flow across a potential barrier will generate shot noise, which is proportional to the square root of the current. This contribution is the manifestation of photon noise in photovoltaic detectors, but includes the noise due to all carriers.

Systems requiring a bias current, classified as passive detectors (bolometers, photoconductors, hot-electron detectors), have a shot-noise contribution not found in active detectors (photovoltaics, pyroelectrics, photoemitters), which are run open circuit.

2.1.5 Generation–Recombination Noise Photoconductive detectors are subject to randomness in both the generation and the recombination, or relaxation, of carriers. Since photovoltaic detectors only produce generation noise, but photoconductors also have recombination noise, photoconductors have $\sqrt{2}$ times the contribution from these processes to the NEP as photovoltaics. This contribution falls off as the modulation frequency approaches the inverse of the carrier lifetime, but the responsivity also falls off in this region. Figure 1 shows the relationship between responsivity and detectivity, defined in Secs. 2.2 and 2.3, and bias current. A flat operating region in the middle of the detectivity curve is found for photoconductive detectors, dominated by generation–recombination noise and giving a linear response.

Since the dominant contributions in both cases are due to photon-generated carrier fluctuations, shot noise and generation–recombination noise are frequently treated together (Dereniak and Crowe, 1984, pp. 40–41).

2.1.6 Phonon or Temperature Noise Random fluctuations in the temperature of the detector contribute to the noise of thermal detectors.

2.1.7 1/*f* and Amplifier Noise As in most electrical circuits, semiconductors generally show 1/*f* noise. Thus this detector-noise component as well as the usual amplifier-noise contributions is reduced by operating in the frequency region where current or generation–recombination noise rather than 1/*f* noise dominates.

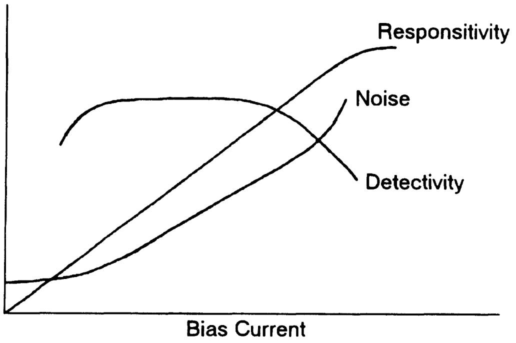

FIG. 1. Relationship between responsivity, noise, and detectivity for a photoconducting photon detector. Adapted from Morten (1971).

2.2 Responsivity

The primary measure of detector sensitivity, the signal per unit rms incident radiant power, is the *responsivity* $\mathcal{R}$ of a detector. This is usually given as a voltage signal, but may also be given as a current signal, as indicated by the units used, volts/watt or amperes/watt. The responsivity may be determined for a blackbody source or as a spectral quantity, but in either case is given for a specific modulation frequency, $\mathcal{R}(T,f)$ or $\mathcal{R}(\lambda,f)$. The noise characteristics of a detector and its responsivity together determine the NEP.

2.3 Detectivity

The parameter that summarizes the information contained in NEP and $\mathcal{R}$ and that allows a comparison of detector-element performance is the *detectivity* D^*, defined as

$$D^* = \sqrt{A_d B}/\mathrm{NEP} \tag{1}$$

in units of cm Hz$^{1/2}$/W, where A_d is the area of the detector and B is the modulation bandwidth in Hz. Some authors normalize NEP to bandwidth, in which case B can be omitted from Eq. 1 (Rieke, 1994, pp. 74, 79). Since it is normalized with respect to detector area and bandwidth, it is a suitable parameter for comparing the performance of detectors not only of different materials, but of different areas and used for different bandwidths. The value of D^* is given for a specified wavelength, modulation frequency in Hz, and postdetection bandwidth also in Hz, i.e., $D^*(10\ \mu\mathrm{m}, 900, 1)$. The bandwidth specification is not strictly necessary, since D^* is normalized with respect to this parameter. A value of D^* may also be given for a blackbody source, in which case the temperature rather than wavelength is specified.

The theoretical limiting performance of photoconductive and photovoltaic IR detectors is conventionally summarized with a plot of D^*_{BLIP}, the "background-limited infrared photoconductor" detectivity, which is defined as the spectral detectivity measured for monochromatic radiation at the wavelength of the peak detectivity for a given cooled detector, having a 2π-steradian field of view of 300-K background radiation over all wavelengths. This quantity is limited by shot or generation–recombination noise, and can be represented for a photoconducting detector as

$$D^* = \frac{\lambda}{2hc}\sqrt{\frac{\eta}{E_p}}, \tag{2}$$

where η is the quantum efficiency, defined in the next section, and E_p is the photon irradiance of the 300-K blackbody background. Photovoltaic detectors have a D^*_{BLIP} that is $\sqrt{2}$ larger than Eq. (2), since they have no recombination noise, as discussed in Sec. 2.1.6.

For many applications, the conditions under which a detector is to operate are far from BLIP conditions, so that D^*_{BLIP} is not really a useful parameter for comparing detectors, as will be shown in Sec. 3.3.3, but it is well established.

2.4 Quantum Efficiency

The fraction of incident photons that produce a free carrier in a photon detector is defined as η, the *quantum efficiency.* This is generally a function of photon energy and is therefore more precisely represented by $\eta(\tilde{\nu})$, the spectral quantum efficiency as a function of wave number $\tilde{\nu}$.

2.5 Frequency Response

The time constant τ of a detector determines its responsivity at a given modulation frequency, as seen in the relation

$$\mathcal{R} = \mathcal{R}_0/(1 + \omega^2\tau^2), \tag{3}$$

where $\mathcal{R}_0$ is the dc responsivity, $\omega = 2\pi f$, and f is the modulation frequency. The response time may be determined in a thermal detector by the thermal time constant, defined in terms of the thermal capacity and conductivity, or in a photovoltaic p-n junction device by the equivalent circuit time constant, defined by the cutoff or corner frequency $f_c = 1/2\pi RC$, where R and C are the equivalent overall resistance and capacitance of the detector and its preamplifier circuit. In most cases both time constants must be considered. The time response is a critical property in many applications.

2.6 Linearity and Dynamic Range

The output signal of most detectors is a nearly linear function of incident power up to some limit determined by intrinsic properties of the detector or by the preamplifier circuit, above which the response flattens in a saturation region. The response region in which the linearity is better than a few percent defines also the dynamic range, that is, the ratio of the largest "useful" signal to the lowest measurable signal, at $S/N = 1$. These properties are important when accurate intensity information about incident radiation is required.

2.7 Geometry

The optimum size of a detector depends on the uniformity of the material, on the relative dependence of the responsivity and noise and thus the detectivity D^* on the area, and on the spatial resolution required by the application. In general the detector element is of the order of 1 mm^2, but the exact size will depend on whether the element is part of an array, or is matched to the aperture of an optical system, and the relative priority given to detectivity vs spatial resolution.

2.8 Stability

Different detectors have different sensitivities toward mechanical stress, thermal cycling, and radiative overload. Most of the current solid-state detectors are relatively robust mechanically and thermally. Some photovoltaic detectors do not tolerate excessive radiation; cooled InSb goes into a saturated state, which, however, is reversible if the detector is warmed up to room temperature. Leads or bonding of many detectors can suffer permanent damage if badly overloaded.

3. TYPES OF DETECTORS

3.1 Thermal Detectors

3.1.1 Bolometers A *bolometer* is an element with a temperature-dependent resistance. It is usually coated with an absorbing layer, so that energy absorption by the transducer is maximized. With the temperature coefficient of the resistance defined as α, the voltage drop across a bolometer is given by

$$(\Delta V) = i_{\text{bias}}\alpha R\Delta T = i_{\text{bias}}\gamma\Delta T. \quad (4)$$

The coefficient α is defined as $\alpha = (1/R)\, dR/dT$, where R itself is a function of the temperature, and $\gamma = \alpha R = dR/dT$. The response time of a bolometer is dominated by its thermal time constant, the product of its thermal impedance and its thermal capacitance. The responsivity is also proportional to the thermal impedance, and so sensitivity and response time must be traded off. Most bolometers have response times on the order of milliseconds, limiting modulation frequencies to the kilohertz range.

3.1.1.1 Semiconductor Bolometers. The temperature dependence of the temperature coefficient and the noise behavior of semiconductor bolometers require operation at 4 K or below. Although bolometers have a flat D^* over the whole IR range, they are used primarily in the FIR, where they are not outperformed by photon detectors. Bolometer arrays are also available. Materials used for bolometers include Si, Ge, and a composite of CaF_2 and Ge (Dereniak and Crowe, 1984).

3.1.1.2 Thermistors. A bolometer consisting of a metal–oxide layer, usually on a sapphire substrate, has a coefficient γ proportional to $1/T^2$ and is called a thermistor. These devices are very robust, are broad banded, work at room temperature, and have a relatively high responsivity over the range from 0.4 to 10 μm; values of γ of up to 5%/K are obtained (Dereniak and Crowe, 1984). These detectors are Johnson-noise limited. They are frequently employed in ambient-temperature applications such as motion sensors.

3.1.2 Pyroelectrics A material that has an intrinsic electric polarization that can be changed by heating is called *pyroelectric.* A change in polarization results in a net flow of charge across the element. The signal current produced by a change in temperature in such an element is

$$i = p(T)A_d\frac{\delta T}{\delta t}, \quad (5)$$

where $p(T) = \delta P/\delta T$ is the pyroelectric coefficient, defined as the temperature derivative of the spontaneous electric polarization P, A_d is the detector area, and $\delta T/\delta t$ is the temperature rate of change. The polarization is a function of temperature, and operation at the temperature of maximum slope of this dependence is of course desired. The responsivity increases linearly with modulation frequency, but is limited by both the thermal and electrical time constants, so that there is an intermediate range of frequencies in which these detectors function optimally. Depending on the preamplifier resistors, the range can vary between 1 to 10 Hz and 10 to 100 kHz, each power of 10 in response time being bought at the price of a factor of 10 in responsivity.

Like thermistors, these materials can be used at room temperature, are relatively robust, and can have a moderately fast response. They are of course insensitive to a dc signal.

The most widely used pyroelectric material has been TGS (triglycine sulfate) and varieties thereof. Other materials used include strontium barium niobate, polyvinylidene fluoride, and lithium tantalate (Dereniak and Crowe, 1984, pp. 171–172).

3.1.3 Thermopile A sequence of thermocouples, yielding as voltage signal the sum of the corresponding work functions, provides a robust thermal detector. In the contemporary form of a thin-film stack composed of Bi and Sb in a variety of configurations, these detectors are commercially significant, especially for monitoring dc or near-dc signals.

3.1.4 Golay Cell The thermal detector devised by Golay uses a small, partially evacuated chamber, one surface of which is a blackened absorber. The residual gas in the chamber, heated by the blackened surface, exerts pressure on a diaphragm that carries a flexible mirror. The pressure distorts the mirror, which in turn determines how much of a light beam falls on a photocell. An increase in light incident on the photocell gives an electrical signal (Golay, 1949). The Golay cell has a very high responsivity but is tricky to make, is mechanically very delicate, and is slow. It was used primarily in the FIR but has now been replaced by semiconducting bolometers.

3.2 Photoacoustic Detectors

The photoacoustic effect is macroscopic in that it uses the equilibrated thermal energy of the sample, but it is a quantum effect when it responds only to radiation that can be absorbed selectively—by quantum processes—by the sample, i.e., when the sample is not black. One surface of the sample chamber, as shown in Fig. 2, consists of a microphone: a diaphragm carrying a plate or element of a capacitor. The change in capacitance produced by pressure on the diaphragm determines the electrical signal at the frequency at which the incident energy is modulated. The signal is directly proportional to the absorbed energy and thus to sample pressure in a gas cell. Like thermal detectors, photoacoustic detectors are slow. They may have a responsivity up to two orders of magnitude lower than semiconductor photon detectors (Perkowitz and Busse, 1986). However, since a photoacoustic detector does not measure total incident power or background power, but only power absorbed by the sample, it has a very low NEP. In an environment where any other detector would be saturated, but a weak signal is sought, photoacoustic detection is ideal. An example is the measurement of the weak vibrational overtone absorption of laser radiation in the near infrared (NIR) or visible by a gas in an intracavity cell (Zhan *et al.*, 1992). In the submillimeter-wave region, photoacoustic detection is used for detecting absorption of gases at low pressure, but especially with pressure-dominated line profiles (Krupnov, 1979). Recently photoacoustic detection has proved remarkably suitable together with time-resolved Fourier-transform techniques for measuring absorption-depth profiles in solid samples (Tam, 1986). Sensitivity to microphonics must be taken into consideration.

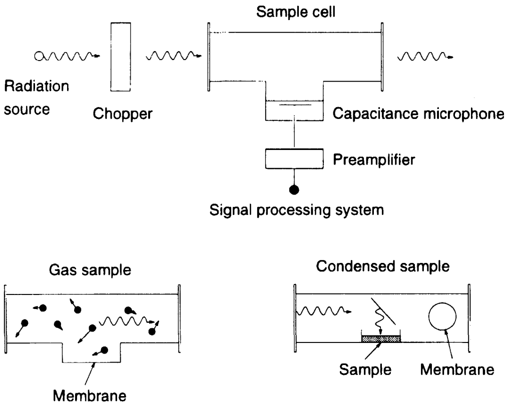

FIG. 2. Schematic structure of a photoacoustic detector, with details of cells for gas sample and solid sample.

3.3 Photon Detectors

Each of the following detector types relies on some variation of the photoelectric effect. Some semiconductor materials can be employed to exploit several of the detection mechanisms outlined here. As will be shown in Sec. 3.3.3, the parameter suitable for evaluating photon-detector efficiency, under conditions relevant to their use, is the quantum efficiency η. However, the commercial specifications are currently still given in terms of D^*.

3.3.1 Photoconductive Detectors Like bolometers, photoconductive devices are passive detectors: A bias current i_b is used to probe the conductivity, so that a change in the conductivity appears as a change in the current, δi_b. The mechanisms of various types of photoconductive devices are illustrated in Fig. 3. If an incident photon raises an electron by the photoelectric effect from the valence band to the conduction band of an intrinsic photoconductor (left-hand side of Fig. 3), the number of free charge carriers is increased, and the conductivity of the material increases correspondingly. This mechanism shows a cutoff wavelength determined by the valence–conduction energy gap, $\lambda_c = 1.23/E_g$ in microns when the gap energy is in eV. As illustrated, intrinsic photoconductors have lower cutoff wavelengths than extrinsic, or doped, photoconductors (middle of Fig. 3). The doping material serves to decrease the effective gap, and thus to increase the cutoff wavelength. These detectors have to be cooled far enough so that there are no carriers in the conduction band at equilibrium, specifically to a temperature $T_c \leq E_g/k$, where k is the Boltzmann constant; 77 K is sufficiently cold if $\lambda_c \leq 10\ \mu\text{m}$. For Ge:Zn, having the lowest E_g of less than 0.3 eV, one needs $T_c = 12$ K.

Photoconductive materials currently used include HgCdTe, in which a wide range of cutoff wavelengths, from ca. 3 to 20 μm, are

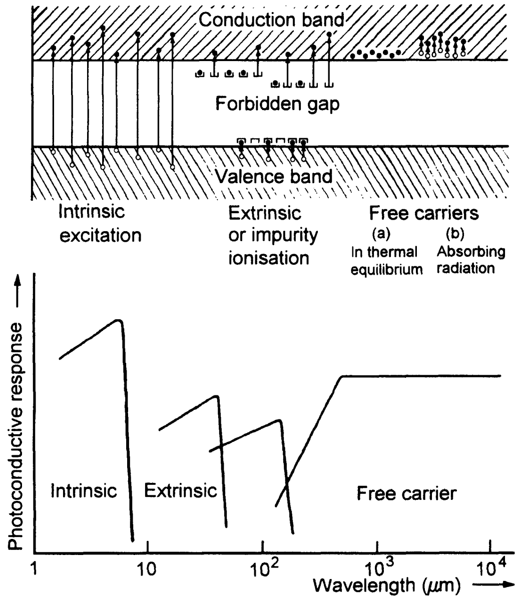
Conduction band
Forbidden gap
Valence band
Intrinsic excitation
Extrinsic or impurity ionisation
Free carriers
(a) In thermal equilibrium
(b) Absorbing radiation
Photoconductive response
Intrinsic
Extrinsic
Free carrier
1
10
10^2
10^3
10^4
Wavelength (μm)

FIG. 3. Mechanisms of semiconducting photoconductors and corresponding schematic detectivity curves as a function of wavelength. Redrawn following Putley (1964b).

achieved by varying the ratio of Hg and Cd in the material. A fixed cutoff is found for each of the doped Ge detectors Ge:Au, Ge:Hg, Ge:Cu, and Ge:Zn. Other materials also have significant detectivities, including variously doped Si. Composite devices, known as blocking–impurity-band (BIB) detectors, allow simultaneous optimization of quantum efficiency and conductivity not achieved with single-material devices (Rieke, 1994, pp. 92–99).

The detectivity as a function of wavelength of a representative set of photoconductive detectors is plotted in Fig. 4 and compared with the curve of the theoretical D^*_{BLIP} introduced in Sec. 2.3.

3.3.2 Hot-Carrier Photoconductors

InSb is a material in which the mobility of the free electrons is large, and when the material is cold (ideally 1.5 K), the relaxation between the free electrons and the lattice is slow. InSb still has some electrons in the conduction band at this temperature. If incident photons have energies below the threshold of the energy gap between the valence and conduction bands, radiation is efficiently

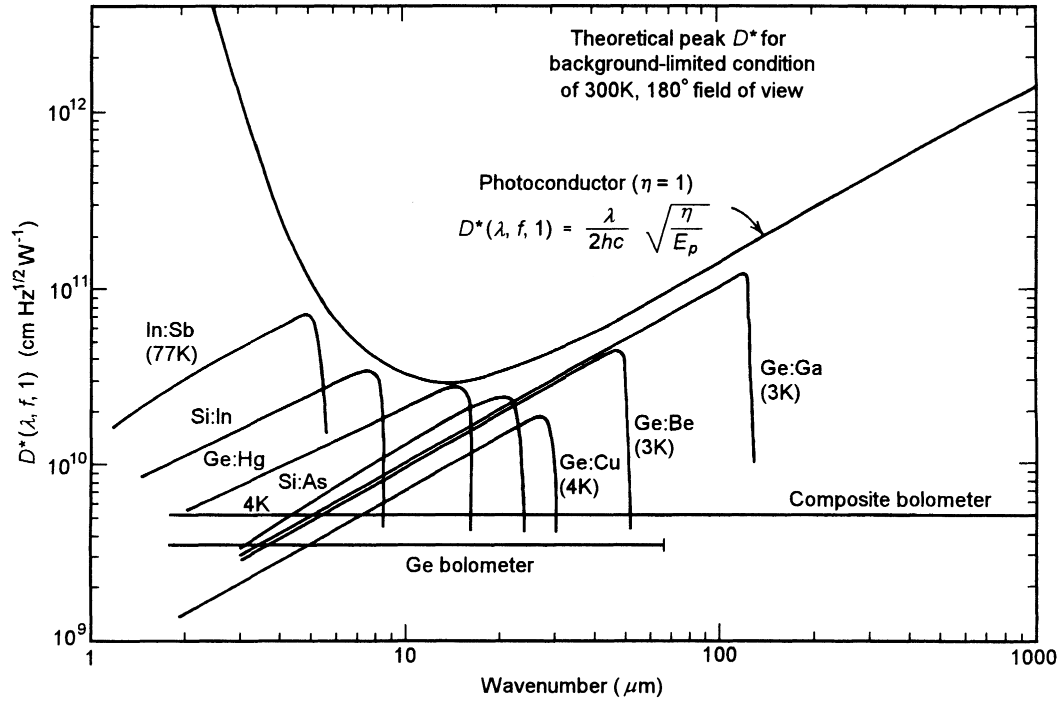

FIG. 4. Curves of D^* as a function of wavelength for a selection of photoconductive detector materials.

absorbed by the free carriers (electrons) in the conduction band at equilibrium, raising them to higher energy levels within the conduction band. This is illustrated in the right-hand side of Fig. 3. The mobility of the electrons depends strongly on their excitation, indeed as $T_{el}^{3/2}$ (QMC Instruments Ltd., 1987). (This detector type has also been classified as a bolometer.) Thus below the energy of the cutoff wavelength, InSb functions as a photoconductor (Putley, 1964a). The photoconductive efficiency is further dependent on λ^2 (Keyes, 1977, p. 23), rising to a constant value in the microwave region, so that this material is especially suited for detection of submillimeter and millimeter radiation. In addition to exhibiting the photoconductive effect, InSb and also Ge (used as a bolometer otherwise) absorb preferentially at the resonant energy of electron–cyclotron resonance, so that a peak in the responsivity can be tuned with a magnetic field. The relaxation to the lattice is fast enough that InSb can be modulated at up to about 5 MHz (Putley, 1964b). The low time constant of this detector, combined with its sensitivity, is quite unique, so that it is a competitor to bolometers in the very far IR.

3.3.3 Photovoltaic Detectors The photoelectric effect has two results in a *p-n* material with a junction potential Φ: Electrons are lifted from the valence to the conduction band, and they are accelerated by the junction potential across the junction, giving a net flow of charge across the element, which can be detected by an external circuit. This mechanism is illustrated in Fig. 5. If no bias current is used, the observed current is linearly proportional to the rate of absorption of photons. Cooling is important in most materials in order to suppress thermally excited carriers. These devices are also called photodiodes. They are characterized by a very high responsivity, rising as a function of wavelength, as for photoconductors, up to the cutoff wavelength determined by the energy gap. The cutoff wavelengths and wave numbers of a selection of photovoltaic detectors are listed in Table 1.

The roughly linear increase of D^* with wavelength seen in Fig. 4 is characteristic of all photon detectors: Whereas a thermal detector has a flat response in a plot of $D^*(\lambda)$, photon detectors have a flat response in a plot of the photon response parameter $\eta(\tilde{\nu})$. Indeed, in order to judge how well a photon detector approaches its theoretical limit or compares with other detectors in the signal-photon–limited case, the quantum efficiency $\eta(\tilde{\nu})$ defined in Sec. 2.4 should be considered rather than D^*. If $\eta(\tilde{\nu})$ is determined and plotted against $\tilde{\nu}$, and thus against photon energy, for the curves shown in Fig. 6(a), one finds the curves of Fig. 6(b), which not only are flat, but the maximum values do not follow in the same sequence of materials as in Fig. 6(a) (Birk and Brault, 1988). This is because of the different amount of background radiation to which each detector can respond under BLIP conditions (Vincent, 1989, pp. 64–67). In the signal-photon–limited case, the noise is proportional to the square root of the current, or proportional to the square root of the number of carriers, and will thus follow the curve of $\eta(\tilde{\nu})$ rather than that of D^*.

3.3.4 Photoemissive Detectors When electrons are emitted from the surface of a detector element, in the classic photoelectric effect, we speak of a photoemissive detector.

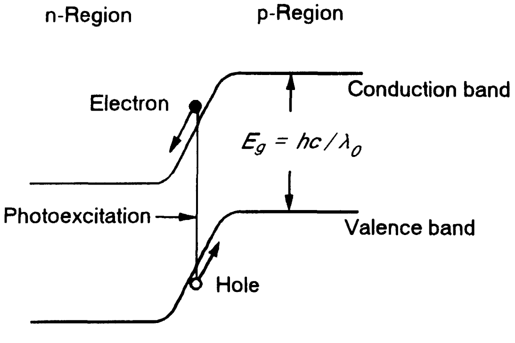

FIG. 5. Mechanism of photovoltaic detectors.

Table 1. Photovoltaic detector cutoff wavelengths λ_c and wave numbers $\tilde{\nu}_c$.

Material	λ_c/μm	$\tilde{\nu}_c$/cm^{-1}
InSb	5.4	1 850
InAs	3.5	2 900
PbS	2.9	3 500
PbSe	5.4	1 850
Ge	1.8	5 500
Si	1.1	9 100
CdSe	0.69	14 500
CdS	0.52	19 200

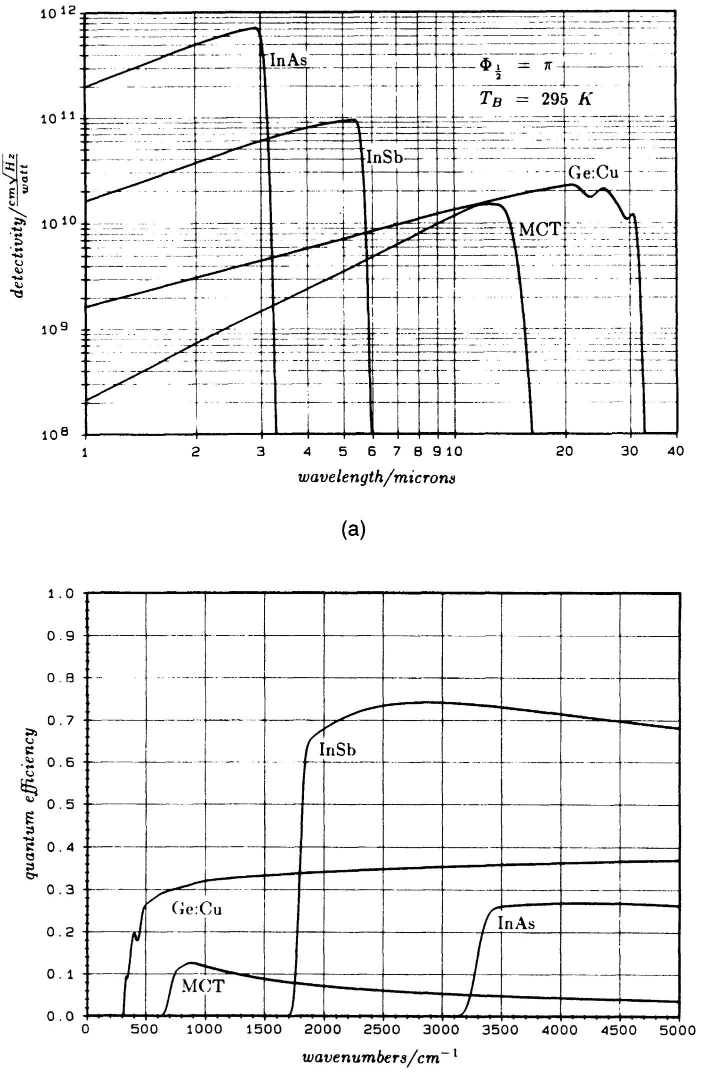

FIG. 6. (a) Spectral detectivities $D^*(\lambda)$ and (b) spectral quantum efficiencies $\eta(\tilde{\nu})$ for several photon detectors. Courtesy of Birk and Brault (1987).

These are used today for visible and shorter wavelength radiation. The electron, once emitted, is accelerated into a series of secondary electron amplifiers, generally a photomultiplier tube, giving a high responsivity. The cutoff wavelength for the use of these detectors is determined by the work function, or surface-ionization potential, of the detector materials. The photomultiplier tube is described in the article DETECTORS, SCINTILLATION.

3.3.5 Photodiodes The performance of a semiconductor as a photoconductor can be enhanced by constructing a diode junction between *n*- and *p*-doped zones within the semiconductor, creating what is called a photodiode. The doping generates a carrier-depletion zone, and thus a high impedance. Since the absorption is due to the intrinsic semiconductor, the quantum efficiency remains high. The high impedance allows good performance at room temperature. Photodiodes are made from Si, Ge, InAs, and HgCdTe and some other composite semiconductors.

Schottky diodes operate also on this principle. A barrier is created at the junction between a metal and a semiconductor, affecting the mobility of holes. Other variants of photodiodes are *pin* diodes, avalanche diodes, and multilayered quantum-well devices (Rieke, 1994, pp. 122–132).

4. DETECTOR SELECTION FOR APPLICATION TO BROAD-BANDED SPECTROSCOPY

The selection of a given detector is made not only on the basis of the performance criteria discussed in Sec. 2, but also on such factors as the basis of cost, size, need for cooling and to what temperature, incident-radiation beam geometry, thermal-resolution requirements, time-resolution requirements, mechanical stability, and required peripheral experimental equipment. In the following, for each region of the infrared spectrum, detectors with optimum properties, at the time of writing, for straight absorption spectroscopy or Fourier-transform infrared (FTIR) absorption spectroscopy against broad-banded thermal or tunable monochromatic sources are specified. For high-resolution measurements, high sensitivity is required and the bandwidth may be limited to a few hundred cm^{-1}. A narrow field of view is employed and, below 1200 cm^{-1}, a cold filter.

These choices will not be optimum for other applications. For example, for two-dimensional imaging, CCD arrays may employ a material that is not by itself optimum, but can be integrated efficiently into a CCD array. Commercially widespread yes-or-no detection systems such as motion sensors have other optimization criteria, and frequently employ thermistors or thermopiles, showing lower sensitivity but also lower cost and complexity.

4.1 Submillimeter Region

For the range 10–30 cm^{-1} there are few interferometric applications; there are no strong broadband sources and the interferometer elements have to be very large. Thus monochromatic sources (klystron harmonics, backward-wave oscillators, carcinotrons) are used for direct absorption measurements with source modulation at typically 10–40 kHz, and the detector of choice is an InSb hot-electron detector at 1.8 K. Most sources in the submillimeter region have low intensity, and a relatively high modulation frequency is needed to get well out of the $1/f$ noise. Such applications require a faster response than a bolometer can offer.

If a gas sample is at a pressure of several hundred Pa, a photoacoustic detector is competitive with an InSb.

4.2 Far Infrared

Most applications in the far infrared for straight absorption or FTIR lead to the choice of a Si or Ge bolometer. If the size of the detector and the diameter and length of the leads connecting it to the preamplifier are properly dimensioned, a bolometer can be optimized for a particular heat load, such as the weak radiation of the usual sources in the range 20–50 cm^{-1}, and achieve high detectivity if the modulation frequencies are kept low. For interferometers this means slow mirror scanning rates.

4.3 Mid-Infrared, 400–2000 cm^{-1}

The best performance is achieved in the mid-infrared with photovoltaic detectors, starting just above the far infrared with Ge:Cu or a similar doped detector; Si:B is an alternative. If the 4.2-K cooling is impractical, HgCdTe at 77 K doped for a long wavelength is suitable.

4.4 Mid-Infrared, 2000–4000 cm^{-1}

Above 1800 cm^{-1} photovoltaic InSb is the optimum detector for spectroscopy with thermal or tunable sources. In this region, signal-photon–limited performance can frequently be achieved.

4.5 Near Infrared

Here also InSb is the obvious choice. Different detector chip sizes and dimensioning for different heat loads are used for the various regions of the NIR. Above ca. 5000 cm^{-1} (below 2 μm), Ge diodes are also used.

4.6 Visible

Both Ge and, below 1000 nm, Si photodiodes and InGaAs *pin* diodes are the most efficient photovoltaic materials in the visible.

The solid-state analog of a photomultiplier tube for very narrow-banded, high-frequency (MHz) applications, such as with pulsed lasers, is provided by Ge or InGaAs avalanche diodes, which exhibit an internal amplification (Rieke, 1994, p. 123).

5. CONCLUDING REMARKS

Details of semiconductor-detector chip structure have not been discussed. The high precision, reproducibility, and reliability achieved in the technologies of semiconductor purification, doping, deposition, bonding, etching, absorbing-layer deposition, etc., have been crucial for the now predominant use of semiconductor-detector elements. Absorption, impedance matching, and heat-load matching are now well understood. As a result, many detectors now operate close to theoretical limits.

Clear-cut procedures for testing the performance of detectors and determining such parameters as the optimum load resistance, for photoconductive and bolometer detectors, have been worked out on the basis of detailed quantitative analysis of the performance of IR detectors, sketched here only qualitatively (Vincent, 1989; Rieke, 1994).

ACKNOWLEDGMENTS

The help of A. F. Krupnov, H. H. Mantsch, A. S. Pine, C. M. Walmsley, and A. Keens in locating references, J. W. Brault in discussions, and J. Preusser in preparing the figures is much appreciated.

GLOSSARY

"Background-Limited Infrared Photoconductor" Detectivity, D^*_{BLIP}: The spectral detectivity measured for monochromatic radiation at the wavelength of the peak detectivity for a given (cooled) detector, having a 2π steradian field of view of 300-K background radiation. This parameter is frequently cited, but the background conditions named are seldom relevant, and comparisons for a given application may not be useful.

Bolometer: A detector element with a temperature-dependent resistance, responding as a black absorber.

Cutoff: The wavelength or wave number at which the response of a semiconductor photon detector drops precipitously, because the energy of the photons falls below the value of the energy gap between the valence and conduction bands.

Detectivity: A parameter that normalizes information contained in the NEP and thus also in responsivity, intended to allow a comparison of detector-element performance. It is defined as $D^* = \sqrt{A_d B}/\mathrm{NEP}$ in units of cm $\mathrm{Hz}^{1/2}/\mathrm{W}$, where A_d is the area of the detector, and B is the modulation bandwidth in hertz. It may be given as a function of wavelength or wave number. The conditions under which NEP is determined must be specified.

Generation-Recombination Noise: Noise in photon detectors due to randomness in the photoelectric generation and recombination, or relaxation, of carriers. Photovoltaic detectors only produce generation noise, but photoconductors also have recombination noise.

Golay Cell: A type of thermal detector, now largely obsolete. See Sec. 3.1.4.

Hot-Carrier Photoconductor: Also called hot-electron photoconductor or bolometer. A semiconductor in which the mobility of the free electrons is large, and relaxation at low temperature between the free electrons and the lattice is slow. Incident photons with energies below the energy gap between the valence and conduction bands may be efficiently absorbed by the electrons in the conduction band, raising them to higher energy levels within the conduction band and thus increasing the conductivity.

MIM Diodes: Metal–insulator–metal diodes, used primarily for mixing radiation from the microwave up to the visible region.

Photoacoustic Effect: The generation of an electrical signal in the acoustic frequency region due to selective absorption of radiation by a sample material in contact with a microphone.

Photoconductive Device: A semiconductor element employed with a bias current in order to probe the conductivity. If an incident photon raises an electron from the valence band to the conduction band, the number of free charge carriers is increased, and the conductivity of the material increases correspondingly.

Photodiode: A photoconductor enhanced by the inclusion of a junction between *n*- and *p*-doped zones within the semiconductor. The doping generates a carrier-depletion zone, and thus a high impedance, while maintaining high quantum efficiency and allowing room-temperature application.

Photoemissive Device: A combination of a photoelectrically sensitive element, from the surface of which electrons are emitted, and a series of secondary electron amplifiers, generally a photomultiplier tube.

Photovoltaic Device: A *p-n* material in which electrons are lifted by the absorption of incident photons from the valence to the conduction band, and are accelerated by the junction potential, giving a net flow of charge.

***p-i-n* Diode:** A photodiode with an extended depletion region consisting of an intrinsic layer. This leads to a device that has a high breakdown voltage and is thus very fast.

Pyroelectric Detector: An element with an intrinsic electric polarization that can be changed by heating, yielding a signal current.

Quantum Efficiency: The fundamental performance parameter for detectors that function via the photoelectric effect. It is a function of photon energy, $\eta(\tilde{\nu})$, and varies between 0 and 1, giving the probability that an incident photon will produce a charge carrier.

Responsivity: The signal induced in a detector element per unit rms incident radiant power, given as $\mathfrak{R}$ in volts/watt or amperes/watt.

Schottky-Barrier Diode: A variety of photodiode. A barrier is created at the junction between a metal and a semiconductor, affecting the mobility of holes.

Thermistor: A bolometer consisting of a metal–oxide layer, usually on a sapphire substrate, generally used at room temperature.

Thermopile: A sequence of thermocouples, yielding as voltage signal the sum of the corresponding work functions.

Works Cited

Bell, A. G. (1881), *Philos. Mag.* **11,** 510.

Birk, M., Brault, J. W. (1988), *Microchim. Acta,* **II,** 243–247.

Dereniak, E. L., Crowe, D. G. (1984), *Optical Radiation Detectors,* New York: Wiley.

Golay, M. J. E. (1949), *Rev. Sci. Instrum.* **20,** 816–820.

Keyes, R. J. (Ed.) (1977), *Optical and Infrared Detectors,* Berlin: Springer.

Krupnov, A. F. (1979), in: G. W. Chantry (Ed.) *Modern Aspects of Microwave Spectroscopy,* New York: Academic, pp. 217–256.

Morten, G. F. (1971), in: F. A. Sowan (Ed.), *Applications of Infrared Detectors,* London: Mullard Limited.

Perkowitz, S. and Busse, G. (1986), in: K. J. Button (Ed.), *Infrared and Millimeter Waves,* Vol. 16, *Electromagnetic Waves in Matter,* Part III, New York: Academic, pp. 1–33.

Putley, E. H. (1964a), *Appl. Opt.* **4,** 649–657.

Putley, E. H. (1964b), *Phys. Status Solidi,* **6,** 571–614.

QMC Instruments Ltd. (1987), *Operating Instructions for Fast InSb Detector System, type QF I/2,* London: QMC.

Rieke, G. H. (1994), *Detection of Light: From the Ultraviolet to the Submillimeter,* Cambridge, U.K.: Cambridge Univ. Press.

Roentgen, W. A. (1881), *Ann. Phys. (Leipzig)* **12,** 153.

Tam, A. C. (1986), *Rev. Mod. Phys.* **58,** 381–431.

Tyndall, J. (1881), *Nature* **23,** 374.

Vincent, J. D. (1989), *Fundamentals of Infrared Detector Operation and Testing,* New York: Wiley.

Zhan, X., Kauppi, E., Halonen, L. (1992), *Rev. Sci. Instrum.* **63,** 5546–5551.

Further Reading

Arams, F. R. (Ed.) (1973), *Infrared-to-Millimeter Wavelength Detectors,* Dedham, MA: Artech House. A collection of original reprints of fundamental publications concerning detectors.

Dereniak, E. L., Crowe, D. G. (1984), *Optical Radiation Detectors,* New York: Wiley. A good introduction to detector theory and practice.

Kelly, W. M., Wrixon, G. T. (1980), in: K. J. Button (Ed.), *Infrared and Millimeter Waves,* Vol. 3, *Submillimeter Techniques,* New York: Academic, pp. 77–110. Basic information on Schottky-barrier diodes.

Krupnov, A. F., Burenin, A. V. (1976), in: K. Narahari Rao (Ed.), *Molecular Spectroscopy: Modern Research,* Vol. II, New York: Academic, pp. 93–126. Example of photoacoustic detection; early references.

Krupnov, A. F. (1979), in: G. W. Chantry, (Ed.), *Modern Aspects of Microwave Spectroscopy,* New York: Academic, pp. 217–256. Extended examples of types of measurements with photoacoustic detection.

Laqua, K., Schrader, B., Hoffmann, G. G., Moore, D. S., Vo-Dinh, T. (1995), *Pure Appl. Chem.*, in press. Nomenclature, Symbols, Units and their Usage in Spectrochemical Analysis—XI: Detection of Radiation. Summary of current usage and recommendations.

Mandelis, A. (Ed.) (1992), *Principles and Perspectives of Photothermal and Photoacoustic Phenomena,* Amsterdam: Elsevier. Several articles concerning absorption-depth profile techniques with photoacoustic detection.

Perkowitz, S., Busse, G. (1986), in: K. J. Button (Ed.), *Infrared and Millimeter Waves,* Vol. 16, *Electromagnetic Waves in Matter,* Part III, New York: Academic, pp. 1–33. General discussion of photoacoustic detection, history and applications in the IR.

Rieke, G. H. (1994), *Detection of Light: From the Ultraviolet to the Submillimeter,* Cambridge, U.K.: Cambridge Univ. Press. Extensive review of detector mechanisms and characteristics, with application examples.

Tam, A. C. (1986), *Rev. Mod. Phys.* **58,** 381–431. Comprehensive technical discussion of photoacoustic detection.

Vincent, J. D. (1989), *Fundamentals of Infrared Detector Operation and Testing,* New York: Wiley. Step-by-step guide to detector evaluation.

Wolfe, W. L., Zissis, G. J. (1985), *The Infrared Handbook,* Washington, D.C.: Office of Naval Research. Extensive background exposition and tables or graphs of numerical values of many parameters for every aspect of infrared applications.

RADIATION DETECTORS, PARTICLE, GAMMA, AND X-RAY

D. SCHALCH AND A. SCHARMANN, *University of Giessen, Giessen, Germany*

Introduction 474
1. **Gamma- and X-Ray Detectors** 474
1.1 Gas-Filled Detectors 475
1.2 Detection of Low-Level Gamma- and X-Ray Activities 475
1.3 Gamma- and X-Ray Detectors Used in Radiation-Protection Dosimetry 477
1.4 Gamma- and X-Ray Spectrometry 479
1.5 Detector Arrays and Imaging Systems 480
2. **Neutron Detectors** 481
2.1 Detection by (n,α) Reaction ... 481
2.1.1 Boron–Fluoride Proportional Counter 481
2.1.2 Boron Ionization Chamber 482
2.1.3 Scintillation Counter 483
2.1.4 Helium-3 Proportional Counter 483
2.1.5 Semiconductor Detector 484
2.1.6 Thermoluminescence Detector 485
2.2 Nuclear-Fission Detector 485
2.2.1 Fission Ionization Chamber .. 485
2.2.2 Fission Track Detector 486
2.3 Proton Recoil Detector 486
2.3.1 Proton Recoil Proportional Counter 486
2.3.2 Proton Recoil Scintillation Counter 487
2.3.3 Proton Recoil Semiconductor Detector 487
2.3.4 Nuclear Track Detector 487
2.4 Activation and Threshold Detectors 488
2.4.1 Activation Detector 488
2.4.2 Threshold Detector 488
2.5 Moderated Detectors 489
2.5.1 Long Counter 489
2.5.2 Rem Counter 490
2.5.3 Multisphere Detection ("Bonner" Spheres) 490
2.6 Current Developments 491
2.6.1 Bubble and Drop Detectors ... 491
2.6.2 Tissue-Equivalent Proportional Counter 492
2.7 Neutron Spectrometry 492
3. **Beta-Particle Detectors** 493
3.1 Detectors for Activity Measurements 493
3.1.1 Liquid and Solid Scintillation Counters 493
3.1.2 Proportional and Geiger-Müller Counters 495
3.1.3 Ionization Chamber 496
3.1.4 Cherenkov Counter 496
3.1.5 Beta-Particle Spectrometry 496
3.2 Detectors for Area and Environmental Control 497
3.2.1 Beta–Gamma Contamination Monitors 497
3.2.2 Room- and Waste-Air Monitors 498
3.2.3 Waste-Water Monitors 499
3.2.4 Beta Monitors for Radiation Dosimetry 500
3.3 Passive Beta Detectors 500
4. **Alpha-Particle Detectors** 501
4.1 Detectors for Activity Measurements 501
4.1.1 Liquid and Solid Scintillation Counters 501
4.1.2 Alpha-Particle Spectrometry .. 502
4.2 Detectors for Area and Environmental Control 503
4.2.1 Alpha-Contamination Monitors 503
4.2.2 Room-Air, Waste-Air, and Waste-Water Monitors 504
4.2.3 Passive Alpha-Particle Detectors 504
Glossary 505
Works Cited 506
Further Reading 507

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

X rays and radioactivity were discovered nearly simultaneously by Röntgen in 1895, and by Becquerel in 1896, respectively. Gamma rays were found by Villard in 1900. The first radiation detectors used were photographic films and fluorescent (scintillating) screens (Röntgen), and simple "ionization chambers" consisting of an open parallel-plate condenser with a gold-leaf electroscope (Pierre and Marie Curie). Soon after the discovery of the new types of radiation, their application in different fields began: x-ray radiology, e.g., in medicine (1896), radium in radiation therapy (1901), the use of radioisotopes as tracers in chemical reactions (1913). Radiation hazards to biological material and the necessity to take protective measures were perceived soon after the first applications. In 1934 the "International Commission on Radiological Protection" (ICRP) presented its first recommendations, e.g., a "tolerance dose" of 0.2 R per day, which exceeds the actual limitations appreciably. The application of radioisotopes was speeded up when the first particle accelerators were available for the production of artificial isotopes in 1932 (Fermi, Seaborg). Finally, Hahn and Strassmann discovered nuclear fission by the bombardment of uranium with neutrons (1939); now the idea of nuclear chain reactions was born. Only three years later, in 1942 the first experimental nuclear reactor was in operation in Chicago. In the following years, military and civilian utilization of nuclear energy started to a large extent. In parallel, more and more sophisticated radiation detectors were developed mainly for protective purposes.

Radiation detection is essentially based on the creation of charged particles in gases, liquids, and solids. In gas-filled detectors, ions and electrons created by the primary radiation are separated by an electrical field, thus enabling the measurement of electrical pulses or currents. In scintillator detectors, the primary radiation creates fast electrons that excite light emission along the particle or photon path. Similarly, secondary electrons are mainly responsible for the generation of visible, microscopic particle tracks in nuclear track detectors by chemical reactions. In semiconductor detectors, electron–hole pairs are generated by particles or photons and are separated by internal electric fields of suitable homojunctions or heterojunctions. A feature of neutron detectors is the necessity of converting their kinetic energy into the creation of secondary charged particles by nuclear reactions or elastic scattering, which then produce ionization events within the detector material, because neutrons are nonionizing and therefore cannot be detected directly. (For details concerning radiation interaction with matter see INTERACTION OF SOLIDS WITH PARTICLES AND RADIATION; ELECTROMAGNETIC RADIATION.)

The Cherenkov detector is based on another physical effect than the detectors described above. The so-called Cherenkov light is generated if the velocity of a charged particle is larger than the velocity of light in the detector material.

Special and highly sophisticated detectors that are used in high-energy particle physics are not considered in this article; they are described in other sections of this *Encyclopedia* (DETECTORS, PARTICLE, CALORIMETRIC; PARTICLE DETECTORS, TRACKING). The object of this article is to give a survey on the more or less classical instrumentation used in the fields of, e.g., applied research and medicine, and in industrial nuclear facilities, not the least for radiation-protection purposes. Also, it is not intended to describe basic physics of detector operation in detail; these items are covered by a number of special articles in this *Encyclopedia*. In the following, we will describe special features of the existing vast variety of detectors with respect to the kind of radiation to be detected, the wide range of particle and photon energies, and the very different demand in scientific and commercial application.

1. GAMMA- AND X-RAY DETECTORS

Gamma radiation is everywhere present in the environment, from long-living natural radioisotopes in the Earth and from cosmic radiation. Artificial radioisotopes are distributed in the environment by nuclear fallout, and in limited areas around nuclear facilities. Gamma and x rays are produced in accelerators and in all those facilities and institutes where x-ray tubes are used. So gamma- and x-ray detectors are needed in a vast variety of civilian and noncivilian establishments. The majority of natural and artificial radioisotopes decay by emission of gamma and also x radiation in a wide energy range between a few keV and about 10 MeV. So far the measurement of radiation intensities and energies is the most important task in nuclear research and application.

1.1 Gas-Filled Detectors

Gas-filled counters (proportional and Geiger-Müller counters, ionization chambers) are used in many fields for cross detection of gamma and x radiation, because they are comparatively simple in construction and in operation, and also relatively inexpensive. On the other hand, more sophisticated instruments are widely used in precision 4π counting of isotope decay, and also in spectrometry, though their importance has decreased in the last decades as a result of the development and growing application of scintillator and semiconductor detectors. Figure 1 shows a simplified cross section of a 4π gamma ionization chamber, which is applied in a large number of versions for activity measurements in the range between 100 kBq and 10 GBq, e.g., for comparative control measurements of radiopharmaceutical samples. The chambers can be easily calibrated for nearly any gamma-emitting isotope, provided the energy is not too far below 100 keV. The chambers are filled usually with noble gases like argon at normal pressure to pressures up to about 40 bar. The lower activity limit of about 100 kBq is determined by background radiation; therefore the chambers are mostly heavy-metal shielded. The sensitivity of chambers is of the order of 1 pA of ionization current per MBq, dependent on chamber size and counting-gas pressure. An advantage of this kind of chamber is the relative insensitivity to the size of samples and to the location of samples inside the well.

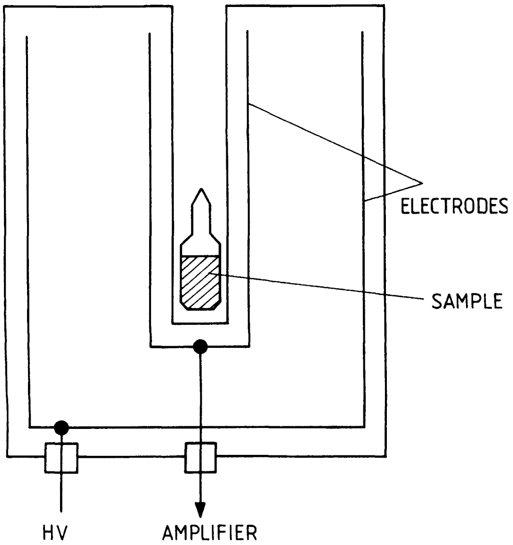

FIG. 1. Simplified cross section of a 4π well ionization chamber for precision measurement of gamma activities. The deviation from 4π geometry due to the open well is corrected by calibration procedure.

Proportional counters are suitable for energy spectrometry of low-energy gamma and x radiation. In the energy range <100 keV their energy resolution is even higher than that of scintillation counters (see Sec. 1.4). Mostly, gas-filled counters are used in radiation-protection monitoring because they are robust, handy, and comparatively inexpensive.

1.2 Detection of Low-Level Gamma- and X-Ray Activities

There is a large variety of applications in basic research and in industrial facilities, where low-level activities of gamma- and x-ray–emitting isotopes have to be measured, e.g.,

1. fallout products in the environment—in air, water, soil, and biological materials;
2. monitoring of waste water and air released from nuclear facilities;
3. nondestructive analysis of extraterrestrial material;
4. tracer techniques in nuclear medicine.

Since gamma and x rays from radioactive decay are emitted as sharp lines, and because of the good energy resolution of the available detectors, especially that of the semiconductor detectors, rapid simultaneous measurement of many isotopes in the same sample is possible, by use of multichannel analyzing techniques. Moreover, many instruments are designed to detect both gamma and beta radiation simultaneously (see also Sec. 3). Usually, NaI(Tl) scintillation detectors (for details see DETECTORS, SCINTILLATION) are used in low-level source measurements, because of their high detection efficiency (Fig. 2).

The detection efficiency is appreciably increased if specially shaped scintillator crystals are applied, which allow one to measure almost under 4π geometry. A large variety of such detectors is in use; suitable calibration procedures are necessary in each case, because the geometries of sample and scintillator, the kind of sample, and the sample–detector distance influence the detection efficiency appreciably. The most common calibration procedure is to use commercially available activity standards of the same shape as the sample, and, if necessary, to embed the standard isotope or isotope mixtures into a samplelike matrix. The efficiency curve given in Fig. 2 is the result of such a calibration, and allows prediction of the efficiency for the detection of other nuclides by interpolation of the curve.

In particular, low-level activity detectors equipped with large-sized scintillator crystals need appropriate shielding against cosmic and environmental background radiation. Large scintillators are more sensitive to the more penetrating gamma components than small ones. Electronic shielding is performed by anticoincidence techniques, physical shielding by heavy-metal plates or water tanks surrounding the detector assembly. The materials are selected with respect to low-level natural activity. Contamination due to environmental radon and its daughter products is avoided by using ventilation systems. More sophisticated systems are equipped with two NaI(Tl) scintillators, and operate in the gamma–gamma coincidence mode. Similarly, beta–gamma detector systems are built by using four detectors, two for gamma detection and two for beta detection, where plastic scintillators are used for beta detection in anticoincidence with the gamma channel (for separation of gamma and beta pulses).

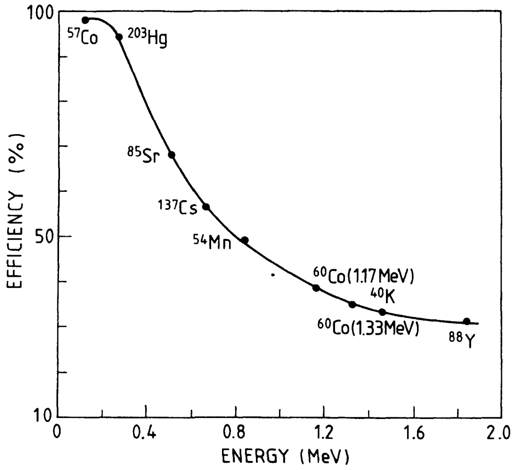

FIG. 2. Photopeak efficiency of a 4π NaI(Tl) scintillation counter system, determined by the aid of "standard" radioisotopes (Hutchinson *et al.*, 1973). Scintillator: two 8 × 8-in. well detectors, enclosing a cavity that contains the sample.

The most powerful tool, especially if low-energy gamma- and x-ray emitters in complex mixtures have to be investigated, is the semiconductor detector (for details see DETECTORS, SEMICONDUCTOR), though its detection efficiency is poorer than that of scintillation detectors, and cooling down to liquid nitrogen temperature is necessary. Today lithium-drifted and high-purity germanium crystals are in use, and advanced materials with higher *Z* are under development. Calibration, efficiency optimization, and background-suppression techniques are similar to those described above. In more sophisticated detector systems, the semiconductor detector is surrounded by a NaI(Tl) anticoincidence shield.

In the low-energy, low-level assay of gamma- and x-ray–emitting isotopes, mainly lithium-drifted silicon [Si(Li)] semiconductor detectors are applied today. A general feature of these devices is their comparatively low background sensitivity, due to the small active volumes. The detection of gammas is limited to energies of up to about 150 keV, dependent on the type of detector, because of the low *Z* and the shallow active zone of some millimeters.

Low-energy quanta are emitted by a number of isotopes in the environment, e.g., by ^{37}Ar, ^{55}Fe, ^{125}I, and ^{129}I. As an example, ^{55}Fe emits 5.9-keV x rays. For quantitative detection of this radiation, high energy resolution is necessary and care must be taken for self-absorption in the sample, absorption between sample and detector, and scattering and x-ray fluorescence in the sample and in the detector itself. In Fig. 3, two pulse-height spectra recorded with a Si(Li) detector are presented. The energy resolution is about 0.5 keV (FWHM, full width half minimum or

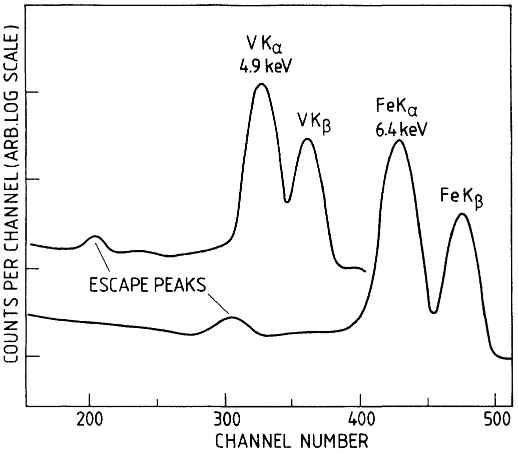

FIG. 3. *K* x-ray pulse-height spectra recorded from the electron-capture decays of ^{51}Cr → ^{51}V and ^{57}Co → ^{57}Fe (Wood *et al.*, 1971).

maximum) in this range, but can be reduced to 0.1 to 0.2 keV with state-of-the-art devices. The escape peaks are from silicon x-ray emission.

1.3 Gamma- and X-Ray Detectors Used in Radiation-Protection Dosimetry

At all sites where people are occupied with gamma- and x-ray sources, as described in the Introduction, care must be taken on radiation hazards (for health aspects of radiation see BIOLOGICAL EFFECTS OF ELECTROMAGNETIC AND PARTICLE RADIATION), and therefore monitoring of workers and working areas is decreed. The monitoring methods comprise

1. area monitoring for direct radiation from closed samples, contamination of workplaces, and airborne activities caused by open sources,
2. monitoring of waste water and air,
3. personnel monitoring with passive detectors,
4. monitoring of incorporated radioisotopes by whole- and part-body counters.

For area-monitoring purposes, mostly portable ionization chambers of various designs are used, which are calibrated in terms of Gy or Gy/h, thus giving a direct, approximate measure of the exposure of workers. The chamber walls are usually fabricated from tissue-equivalent plastics. The instruments commonly have air-filled chambers of some hundreds of cm^3 of active volume and are operated at normal pressure. The gamma-detection capability covers the range between some tens of keV to several MeV. The dose-rate range is usually 1 μGy/h to 1 Gy/h. Small "pocket chambers" are designed for personnel dosimetry. Older versions are of very simple construction and are based on capacitor discharge by ionizing events. Modern devices have sophisticated electronics for, e.g., summing up exposures over certain periods, for alarming if exposure levels are exceeded, and for self-checking of proper operation.

Special ionization chambers, so-called "flat chambers" and cylindrical "compact chambers," are constructed for dosimetry mainly in the field of gamma- and x-ray therapy (see BIOMEDICAL USE OF RADIATION). Flat chambers are used for soft radiation in the energy range between about 10 and 100 keV. They have parallel electrode plates and mostly volumes of less than 1 cm^3. The window thickness is a few mg/cm^2 in order to avoid absorption of the low-penetrating radiation. Flat chambers and compact chambers for detection of "hard" gamma and x rays have similar small volumes, but are used together with caps of about 1 cm of thickness, when the measurements are made in air, in order to ensure secondary electron equilibrium. For therapy purposes, the detection sensitivity of chambers is between 10^{-2} and 10^2 Gy, and 10^{-4} and 1 Gy/s, respectively, and the allowed error limit is some percent.

Contamination monitors, waste-air and waste-water monitors, are mostly designed to measure gamma + beta activities simultaneously. They have thin windows to avoid absorption of radiation (see Sec. 3.2). In nuclear reprocessing facilities, wastes containing numerous isotopes are handled. Detection and identification of the isotopic composition are only possible by the use of Ge semiconductor detectors. Figure 4 shows an example of the gamma spectrum of a nuclear-waste sample. The resolution of other detectors, e.g., the scintillation counter, is too poor for this purpose. A number of passive detectors are in use for personnel monitoring:

1. Film dosemeters, the blackening of which

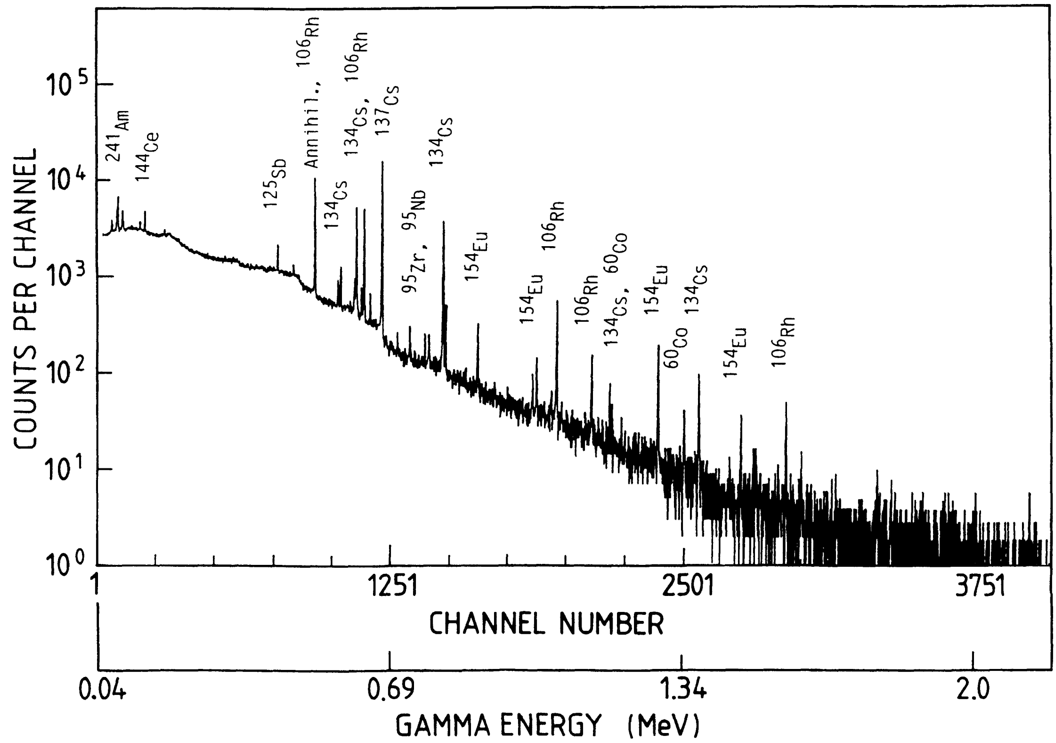

FIG. 4. Gamma spectrum recorded with a high-purity germanium spectrometer from nuclear waste produced in a reprocessing plant (Scharmann *et al.*, 1986). The record is an original one, and no corrections are made for the energy-dependent efficiency of the detector, beta-particle background, etc.

by ionization events is a measure for the exposure of workers as a result of gamma and x rays. The sensitivity of films is large at low energies and decreases with increasing energy. The lower dose-detection limit is about a tenth of a mSv at 1 MeV and some tens of μSv at 100 keV.

2. Thermoluminescence dosemeters (TLD), based on the annealing of defect centers, created by the radiation in suitable materials, mostly lithium fluoride. The annealing procedure produces a characteristic emission of light ("glow peaks"), which is related to the radiation intensity. TLDs are more sensitive to radiation than film dosemeters above a few hundred keV. Typical lower dose limits are some tens of μSv at 500 keV to some MeV.
3. Radiophotoluminescence dosemeters (RPD) are based on a similar effect: Radiation-induced defects are recovered by irradiation with light of suitable wavelength, e.g., with laser light. The induced light emission due to defect recovery is a measure for the applied dose. The sensitivity of RPD approaches that of TLD.

There are a number of other solid-state effects that are useful for passive radiation detection, e.g., exoelectron emission, lyoluminescence, electron-spin resonance, chemical reactions. Those techniques are of limited importance at the moment, and are not described in detail.

Whole- and part-body counters are constructed for detection of incorporated activities. If the activity of an isotope is concentrated in a certain organ by biological mechanisms, part-body counters are used, e.g., "lung counters" or "thyroid counters." If the activity is distributed over a number of organs or tissues, mostly whole-body counters are used. Because of absorption of radiation in tissue, especially when low-energy, low-intensity isotopes are incorporated, highly sensitive detectors have to be used, commonly scintillation detectors of the phoswich type (see Glossary). Additionally, heavy-metal and anticoincidence shielding is then necessary to suppress background radiation. Steel shields originating from old, broken-up battle ships are preferred, because these steels were fabricated in the decades before nuclear weapons were tested, and they contain low levels of activity from fallout. Nevertheless, the detection of sufficient low levels of actinides, e.g., ^{239}Pu, in a human body is extremely difficult because of low-intensity, low-energy gamma emission, and the low limits set by radiation-protection rules.

1.4 Gamma- and X-Ray Spectrometry

For gamma- and x-ray spectrometry at not too low energies, Ge(Li) or high-purity Ge semiconductor detectors and NaI(Tl) scintillation counters are nowadays the instruments of choice. When a pulse-height distribution is taken as a quantitative measure of an energy spectrum ("pulse-height analysis"), one has to be assured that all quanta are absorbed in the detector volume. Otherwise correction/calibration procedures are necessary. This is the reason why high-*Z* materials (e.g., NaI, Ge) are preferably used in solid-state gamma detectors, and the development of advanced higher-*Z* detectors is forced (e.g., HgI and CdTe).

The choice of whether to take a semiconductor detector or a scintillation detector, when the spectral analysis of a radioactive sample is to be performed, depends on the type of source: If high energy resolution is demanded, e.g., measuring a mixed isotope source like that presented in Fig. 4, a semiconductor detector should be applied. The energy resolution of scintillation detectors is about an order of magnitude poorer (Fig. 5), and there would be no chance to perceive the isotopic composition of a complex source. On the other hand, if low-level sources have to be investigated, the higher detection efficiency of the scintillation detector is of importance. In sophisticated systems, combinations of semiconductor and scintillation detectors are used to suppress background radiation. In Fig. 6, a schematic sketch of a highly sensitive spectrometer assembly is presented. The main Ge(Li) detector is surrounded by two NaI(Tl) scintillators in order to reduce the background electronically. Additional steps are taken to reduce

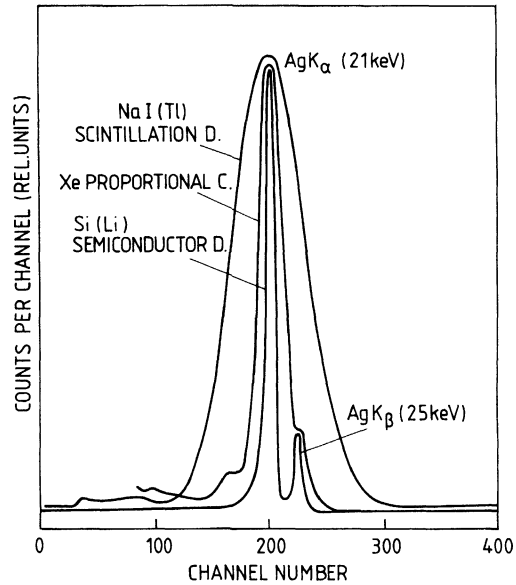

FIG. 5. Comparison of silver *K*α and *K*β x-ray spectra, recorded with different detectors (Muggleton, 1972). The two silver lines are at about 21 and 25 keV, respectively.

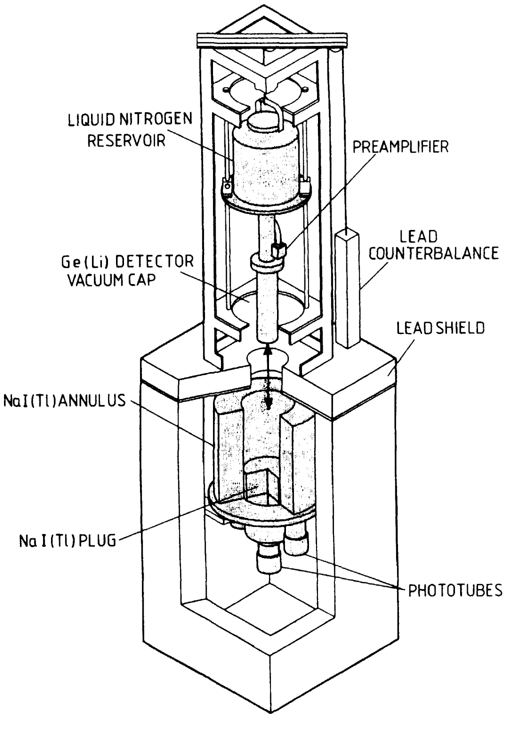

FIG. 6. Cross section of a *Ge(Li)–NaI(Tl) coincidence–anticoincidence* gamma spectrometer (Cooper and Perkins, 1971).

background pulses by means of physical shielding, as described in the preceding sections.

Spectrometry in the lowest-energy regime is nowadays mostly performed with Si(Li) detectors, which have the highest energy resolution in this range (see Fig. 3). In the intermediate-energy range, competing instrumentation is, e.g., the scintillation counter (>100 keV) or the proportional counter (<100 keV). The latter can be used down to energies below 1 keV. Under development at present is a high-resolution detector on the basis of a superconducting "Josephson junction," which exceeds even the resolution of the Si(Li) detector at low x-ray energies. It can be assumed that with advanced devices of this type, the lower-energy limit in single-photon spectrometry, which is today at about 100 eV, can be shifted to the visible region of several eV.

1.5 Detector Arrays and Imaging Systems

The oldest and most widely known imaging system is the photographic film. The films consist usually of silver bromide grains with thicknesses of up to 1 μm in a gelatin matrix. Ionizing radiation sensitizes these grains, and after subsequent processing of films, the bromide is converted to elemental silver, not only in the excited primary grain but also in a large number of grains in the vicinity of the primary one. So a macroscopic track of the quanta (or particles) is produced.

The two categories of application of films are the recording of single-particle tracks, e.g., in nuclear physics, and imaging of a large number of ionizing events in (scattered) broad beams of x or gamma radiation, the so-called radiography. The latter technique is applied, e.g., in medicine, in materials research, and also in general biological research. Figure 7 shows a principal assembly used in radiography. The different scattering and absorption of radiation in materials of different density or with different Z (e.g., tissue and bones in medical application) produce "shadows" on the processed film. By high-Z converter foils or scintillator foils, placed between the radiation source and the film emulsion, the sensitivity of the film detector can be increased by a factor of 10. When a scintillator screen is used instead of a film, an image of an object is created by the different light intensities.

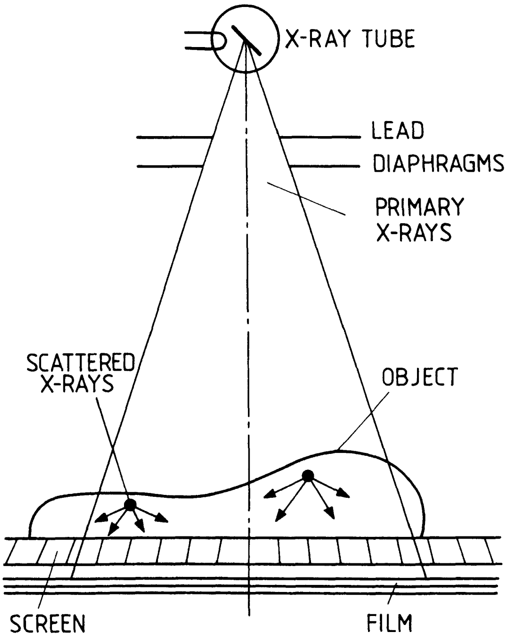

FIG. 7. Schematic sketch of an x-ray radiography assembly for diagnosis in medicine (Lang, 1979).

Another technique is the recording of coherently scattered gamma or x radiation (diffraction techniques), which allows one to observe the microscopic structure of materials.

Modern imaging systems were developed originally for use in basic research, but during the last decade there is growing interest in such systems for practical application. Whenever time-resolved ionization events are to be recorded, detectors other than film emulsion detectors are needed. In nuclear and high-energy basic research, a large number of time- and position-resolving detectors are in use, e.g., ionization chambers, proportional counters, semiconductor detectors, and scintillation counters. Figure 8 shows one of various position-sensitive semiconductor assemblies that were developed in the past (for modern devices see DETECTORS, SEMICONDUCTOR). Exemplary applications of such detectors aside from basic research are, e.g., computer tomography in medicine and the corresponding techniques in materials research.

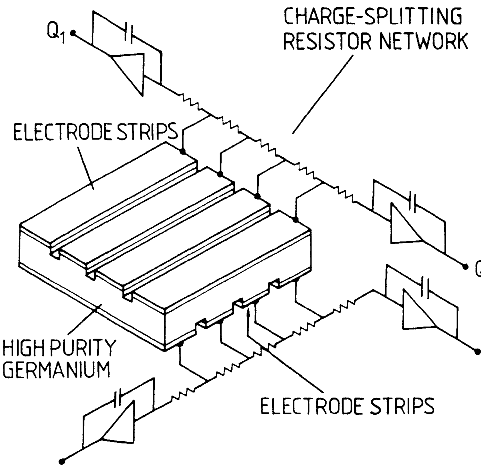

FIG. 8. Example of a position-sensitive germanium detector (Gerber *et al.*, 1977).

2. NEUTRON DETECTORS

Since neutrons are nonionizing particles, their direct detection is impossible. Therefore neutron-flux and neutron-energy measurements are always based on the detection of secondary particles, which are produced in suitable devices via nuclear reactions or elastic scattering. There are numerous techniques for the creation and detection of secondaries like alphas, protons, betas, gammas, and fission products. A need for neutron detectors aside from basic scientific interest is, e.g., in facilities of the nuclear-fuel cycle, at high-energy accelerators, and also at accelerators for medical purposes, in military nuclear facilities, and in civilian isotope-production facilities. The detectors are used, e.g., in nuclear-reaction and process control and in radiation-protection dosimetry. In nuclear and accelerator facilities and also in the cosmic-radiation environment, neutrons are produced in a large energy range (Fig. 9). Because of the complex, strongly energy-dependent interaction of neutrons with detector materials, various detection concepts are applied to cover the total range of neutron energies (e.g., Griffith *et al.*, 1990). An additional problem exists in radiation-protection dosimetry of neutron fluxes: The energy-dependent quality factor of neutrons, which is related to their biological effectiveness, has to be taken into consideration in dose-measurement techniques and instrumentation, respectively.

2.1 Detection by (n,α) Reaction

2.1.1 Boron–Fluoride Proportional Counter BF_3 gas-filled counters are the most commonly used active neutron detectors in scientific research, process control, and radiation-protection monitoring. By the reaction $^{10}B(n,\alpha)^7Li$, alpha particles are created and can be detected with high sensitivity. Because of the large specific ionization of alpha particles, it is possible to suppress a gamma background by pulse-height discrim-

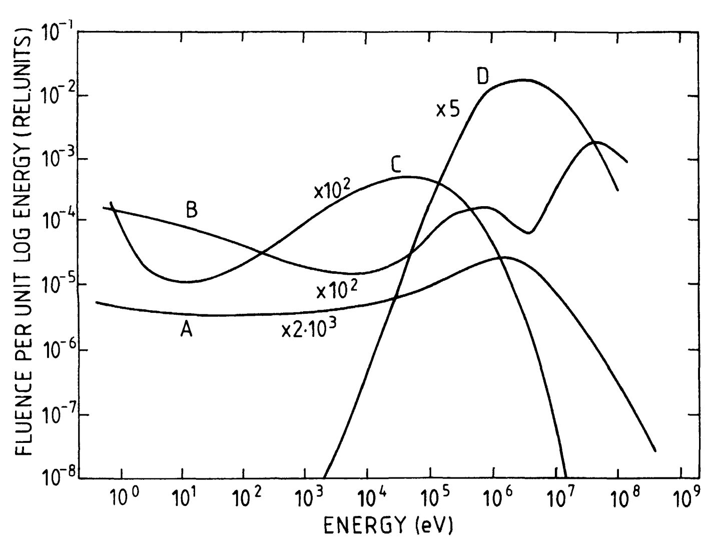

FIG. 9. Selected neutron spectra. Curve A, cosmic-ray neutron spectrum at an altitude of 12.3 km (Hewitt *et al.*, 1976). Curve B, neutron spectrum at an alternating-gradient synchrotron (location: ring top) (Distenfeld, 1976). Curve C, neutron spectrum in a commercial power water reactor (Endres *et al.*, 1981). Curve D, neutron spectrum at a Sagittaire therapy accelerator (Holeman *et al.*, 1977). Note: The given spectra are exemplary and are valid only for distinct areas at the facilities. The spectra may vary strongly dependent on the location within the facility, and in the case of cosmic radiation with geomagnetic coordinates, time, height, etc.

ination. Gamma background doses below about 1 Gy/h are then negligible. The counter dimensions are a few centimeters in diameter and 10 cm to about 1 m in length; dimensions of less than a few centimeters would result in a drastic drop of the (*n*,α) reaction yield. Usually the BF_3 pressure is around 5 × 10^4 Pa (0.5 bar). BF_3 itself shows appropriate quenching characteristics, so that the addition of a quenching-gas component is not necessary. Depending on the active counter volume and the gas pressure, the sensitivity is between 1 and 100 counts/s at a thermal neutron-flux density of 1 *n*/cm^2 s. The counter is useful up to dose rates of about 1 Gy/h (about 10^5 *n*/cm^2 s). Above this limit, gas dissociation occurs and the quenching properties diminish.

Another but more uncommon version of the boron-activated proportional counter is the "boron layer counter," where the inner counter walls are covered with boron compounds, as, e.g., BC_4 or B_2O_3. Its properties are poorer at low neutron fluxes, and therefore they are used only in the presence of high neutron fluxes.

The cross section σ of the (*n*,α) reaction is a function of the neutron energy: $\sigma \sim 1/E^{1/2}$ (Fig. 10). That is, the high sensitivity of the BF_3 counter is limited to the low-energy regime. The cross section for thermal neutrons (E = 0.025 eV) is 3.84 × 10^{-25} m^2 (3840 b). The sensitivity to neutrons of higher energies can be enlarged by the aid of surrounding moderator materials like polyethylene or paraffin. By using one or more moderator spheres (or cylinders), the detection of fast and even of energetic neutrons is possible (see Sec. 2.5).

2.1.2 Boron Ionization Chamber

Boron ionization chambers are suitable for measurement of neutron-flux densities of 10^4 up to 10^{14} *n*/cm^2 s, as they occur, e.g., in the vicinity of nuclear-reactor cores. The chambers are filled with nitrogen, argon, methane, or hydrogen; the inner walls are covered with suitable boron compounds. Gas pressures are between 10^4 and 10^5 Pa (0.1 and 1 bar), depending on chamber design. Dimensions are similar to those of boron-activated proportional counters. A further advantage of the ionization chamber is its usability at temperatures of up to 700 to 800 K, whereas the BF_3 gas-filled proportional counter or ionization chamber cannot be used at temperatures exceeding about 400 K. If one is working in the direct-current mode, more expensive compensation techniques are necessary in the presence of gamma radiation. Different methods are in use:

1. compensation by the aid of an integrated second gamma chamber that is not boron activated, and by subtracting the gamma-induced current electronically;
2. additional mechanical compensation by varying the active volume of the gamma chamber in a suitable manner, thus improving the electronic compensation mentioned before, and enlarging the measurable neutron-flux range appreciably.

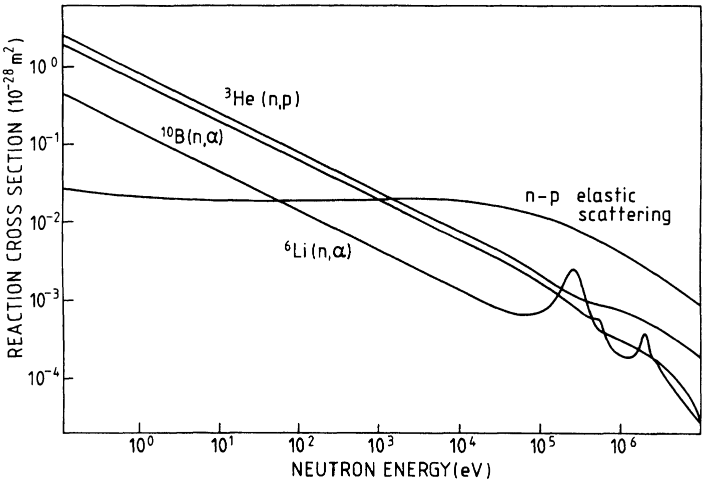

FIG. 10. Cross sections vs neutron energy for some important interactions of neutrons with detector materials.

2.1.3 Scintillation Counter

Scintillation counters (see DETECTORS, SCINTILLATION) may be used for neutron detection when charged particles like alphas or protons can be created by suitable neutron reactions within the scintillator material. Even the "normal" NaI scintillation counter is sensitive to neutrons by virtue of the reactions $^{23}Na(n,\gamma)^{24}Na$, and $^{127}I(n,\gamma)^{128}I$. Both the isotopes created by this reactions are beta–gamma emitters. Though the neutron sensitivity of the NaI counter is small, it is sometimes used for calibration purposes.

An efficient scintillator material for neutron detection is $^6LiI(Eu)$. This material is sensitive to slow neutrons because of the reaction $^6Li(n,\alpha)^3H$. The cross section σ for thermal neutrons is 9.5×10^{-26} m^2 (950 b); the energy dependence of the reaction cross section is similar to that of boron (Fig. 10). $^6LiI(Eu)$ scintillators are highly sensitive to gamma radiation, and so a gamma background must be suppressed by the aid of pulse-height and pulse-shape discrimination techniques. The usefulness of the $^6LiI(Eu)$ scintillation detector is limited to low- and medium-energy neutrons. In the MeV energy range, competing nuclear reactions occur, $^6Li(n,nd)^4He$, and $^6Li(n,p)^6He$. That is, the pulse spectrum is complex and quantification of results is hardly possible (Fig. 11). [For the response of a $^6LiI(Eu)$ counter to low- and medium-energy neutrons see Fig. 24 in Sec. 2.5.3.]

Another type of ^6Li-activated scintillator material is the "glass scintillator," containing, e.g., $^6LiO_2$, and Ce as an activator for scintillation. Those detectors can be used for thermal-neutron detection but they have poorer detection properties and are not commonly in use. Further versions of the neutron scintillation counter are based on other neutron interactions: on elastic neutron scattering (see Sec. 2.3.2), and on fission reactions. Mixtures of, e.g., ZnS(Ag) together with fissionable isotopes, as thorium, uranium, or neptunium are used as scintillators. These versions are not commonly in use.

2.1.4 Helium-3 Proportional Counter

In this type of counter, the reaction $^3He(n,p)^3H$ is applied for the detection of neutrons of energies between 10^{-2} and 10^3 eV. The reaction cross section is 5.33×10^{25} m^2 (5330 b) at thermal-neutron energies, thus larger than the corresponding one of ^{10}B. It depends on neutron energy as $1/E^{1/2}$ (Fig. 10). The ^3He proportional counter is operated at gas pressures up to 10 bar, containing pure He or He–Kr mixtures, and is more sensitive to neutrons than the BF_3 counter. The sensitivity to gamma radiation is small, and a gamma background is easy to suppress. In the presence of a broad neutron-energy distribution, some competing effects occur, e.g., He recoil nuclei from elastic n–He scattering. Also ^3H recoil nuclei, and Kr recoils in mixed-gas counters, create additional pulses. These contributions to the pulse spectrum can be well separated (Fig. 12) and suppressed by pulse-height and pulse-shape discrimination. An energy resolution of 3% between 1 eV and 1 MeV neutron energy is possible. For the response of the ^3He counter, see also Fig. 17.

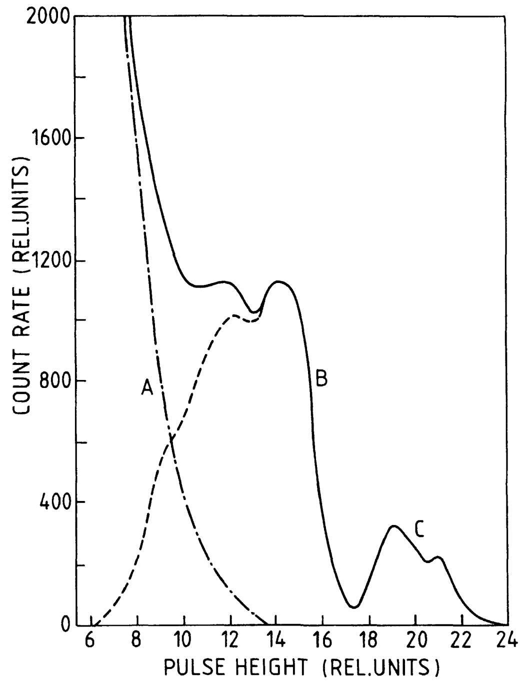

FIG. 11. Coincidence pulse spectrum due to 14-MeV neutrons measured with a LiI(Eu) scintillation detector (Neuert 1966). Curve A, gamma background; curve B, charged particles; curve C, alpha particles.

Another version of a ^3He counter consists of two semiconductor devices, between which the active gas volume is embedded. Protons and ^3H recoils are counted in coincidence, while other recoil nuclei and signals

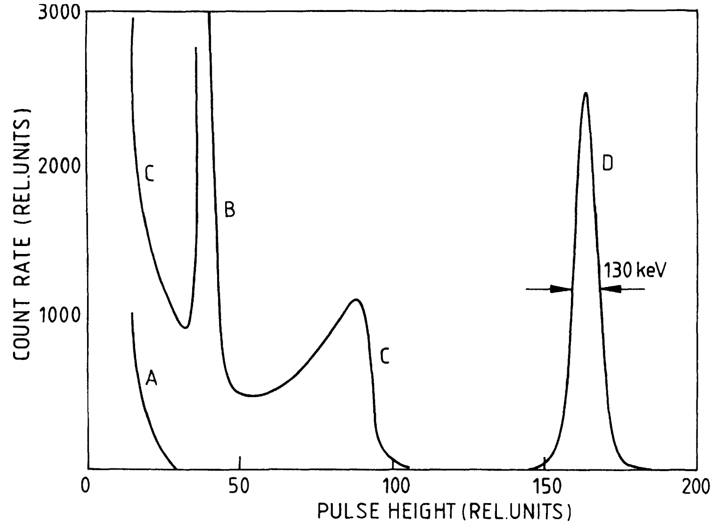

FIG. 12. Pulse spectrum due to 2.5-MeV neutrons measured with a helium-3 proportional counter; gas filling: ^{3}He + Kr + CO$_2$ (Sayres and Coppola 1964). Curve A, gamma background; curve B, epithermal neutrons; curve C, Kr and CO$_2$ recoil distribution; curve D, 2.5-MeV neutrons.

from Si(n,α) and Si(n,p) reactions are eliminated by this technique.

2.1.5 Semiconductor Detector The reaction ^{6}Li(n,α)^{3}H is also utilized by detecting neutrons with surface-barrier detectors (see DETECTORS, SEMICONDUCTOR). A pair of such detectors, one of which is covered by a thin ^{6}LiF film, is used in a sandwich structure. Because of the low range of alpha and hydrogen particles, the distance between the two detector surfaces is made as small as possible. The two particle groups can be detected separately in each device, the sum peak of both, and also a coincidence spectrum can be determined (Fig. 13). The amplitude of the sum pulses is related to the primary neutron energy. The high energy

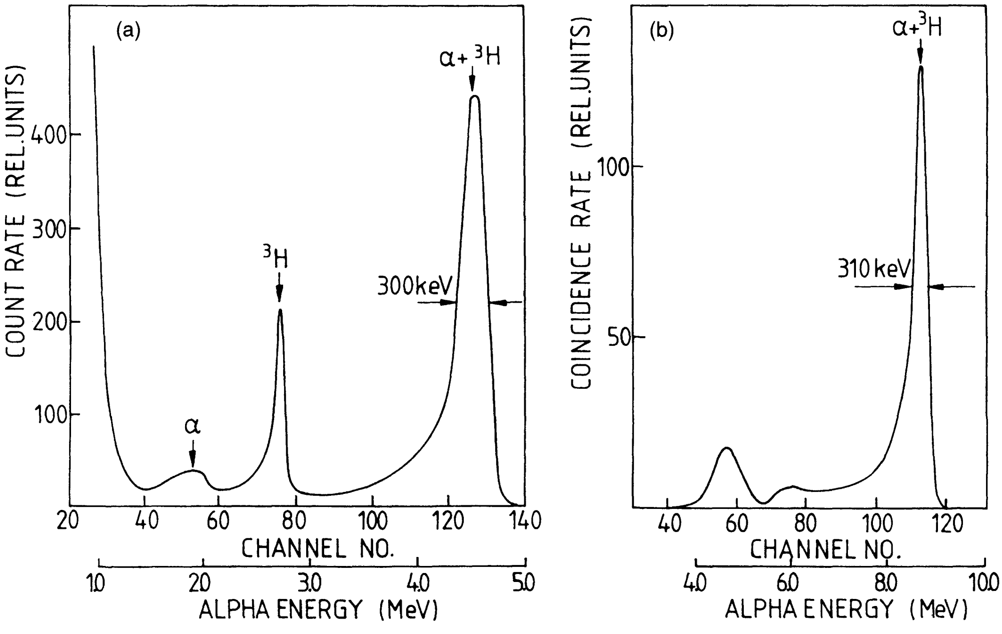

FIG. 13. Pulse spectra of secondaries from the reaction ^{6}Li(n,α)^{3}H measured with a LiF surface-barrier sandwich detector (Neuert 1966). (a) Spectrum of thermal neutron-created secondaries. (b) Coincidence spectrum of 4.15-MeV neutron-created secondaries.

resolution of the surface-barrier detectors is reduced by the LiF film thickness, which should not be too small with respect to the sensitivity of the system. Though the response of this type of detector is small (~1 count per 10^6 neutrons), they are useful for neutron spectrometry because of their good energy resolution.

2.1.6 Thermoluminescence Detector The thermoluminescence detector is a passive device that commonly contains ^{6}Li as neutron-sensitive component. If a ^{6}LiF detector is heated after neutron irradiation, thermoluminescence emission is observed, which can be separated from a gamma background by using a pair of neutron-sensitive and neutron-insensitive detectors. This type of detector is used predominantly in radiation-protection dosimetry: the so-called "albedo personnel dosemeter." The high cross section of ^{6}LiF for thermal neutrons in combination with suitable moderator materials allows the determination of "personnel doses" from the thermal- to fast-neutron range, if the neutron spectrum is well known (see Sec. 1.3).

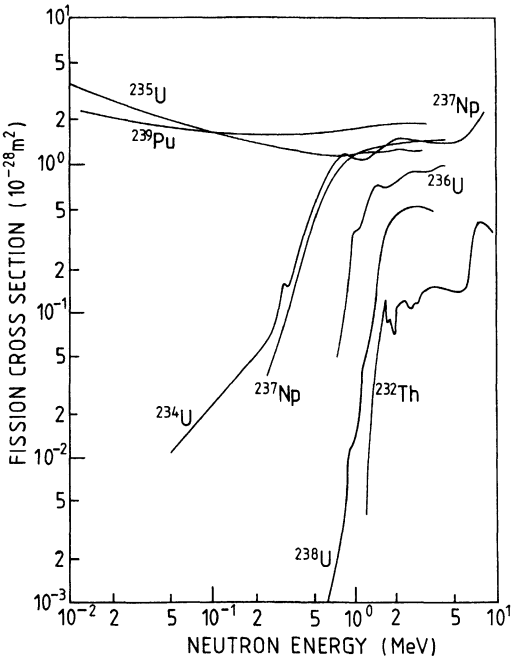

FIG. 14. Nuclear-fission cross section of isotopes used in fission detectors (Lamphere, 1960). 10^{-28} m^2 = 1 barn.

2.2 Nuclear-Fission Detector

Advantageous properties of fission detectors are the following:

1. The energies of fission particles are large (~100 MeV), and so the electronic suppression of a gamma background is easy.
2. There are a large number of suitable elements to be used in the thermal to energetic neutron range (Fig. 14).
3. A number of elements have sharp fission thresholds, so that high-energy neutrons can be detected separately from lower-energy neutrons in a comparatively easy manner.

(See NUCLEAR ENERGY, FISSION)

2.2.1 Fission Ionization Chamber The electrodes of fission chambers are covered with a thin film of a suitable fissionable element. Normally, fission chambers work in the pulse-counting mode; the current mode is used only at very high neutron fluxes. Some of the elements decay by emission of alpha particles, e.g., uranium, thorium, and plutonium. For having high sensitivity at low neutron fluxes it is necessary to suppress the alpha signals, too. This can be done electronically if the chamber is operated in the pulse-counting mode. For operation in the current mode, the chamber dimensions are made smaller than the alpha-particle ranges. The contribution of alpha-particle ionization is then small compared with the contribution of fission-fragment ionization. In Fig. 15, a miniaturized fission chamber for in-core control measurements of neutron fluxes in nuclear reactors is shown. The inner electrode is covered with ^{235}U. Those chambers can be in operation for a number of years since the fission cross section of ^{235}U is comparatively small. This is advantageous over boron ionization chambers, which would be burnt up in short times because of their large cross section for slow neutrons. The fission chamber in Fig. 15 has been constructed to withstand temperatures up to 700 K, and pressures up to 200 bar. It has a nearly linear response up to a neutron flux of 2×10^{14} *n*/cm^2 s in the presence of a gamma dose rate of 10^7 Gy/h. The outer diameter of such a chamber is only about 5 mm in order to en-

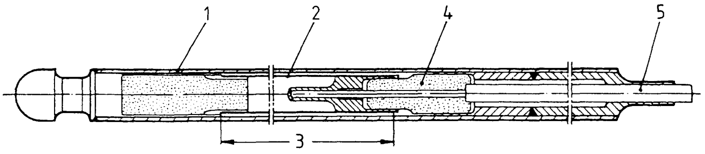

FIG. 15. Cross section of a miniaturized fission ionization chamber for use as in-core flux detector in a nuclear reactor; see text (manufacturer: AEG, Germany). 1, Chamber wall and outer electrode; 2, hollow, uranium-covered inner electrode; 3, sensitive length of the chamber; 4, metal–ceramic seal; 5, ceramic cable.

able installation within the fuel elements of a reactor.

2.2.2 Fission Track Detector Fission recoil nuclei create tracks in solid materials as a result of structural damage along their path, from elastic collisions and primary and secondary ionization events. Dependent on the detector material and the mass of the recoil fragment, the tracks are roughly of the order of nanometers in diameter and up to some tens of microns in length. Fission-track detectors are passive devices consisting mostly of a foil of fissionable material (Fig. 14) in contact with a suitable track-recording sheet. This sandwich configuration has advantages over another version of fission track detectors, where the fissionable material is incorporated in the detector material, since the detection starts with combining the two foils. Otherwise a background from time before start of measurement has to be considered. Numerous glasses and plastics are in use as detectors. Counting of tracks can be performed in different ways:

1. By electron microscopy.
2. By visual (automatic) counting, after the track dimensions have been enlarged by chemical or electrochemical etching.
3. By spark counting, applying high voltage across the (etched) detector film. The tracks across the film lead to single electrical breakthroughs.

Fission track detectors can be used over the total neutron-energy range by taking suitable fissionable materials (Fig. 14), and also for neutron spectrometry, at a large range of neutron fluxes/doses. The technique is comparatively simple, as no expensive electronic equipment is needed. The detectors are insensitive to beta and gamma radiation. The tracks are stable over long periods. As an extreme example, fission track dating is used in geochronological research, e.g., tracks from fission of natural uranium in mica allow dating of material that is up to about 10^8 years old. The detectors also enable the three-dimensional observation of fragment tracks in high-energy and cosmic-ray research.

2.3 Proton Recoil Detector

Proton recoil detectors are based on elastic collisions of fast neutrons with protons in gaseous or solid detector materials. The cross section of this non-nuclear reaction is large because of the equal masses of the two particles (Fig. 10). In hydrogen-containing gases or plastics, protons are kicked out of the molecular binding structure by fast neutrons, and act as fast ionizing particles.

2.3.1 Proton Recoil Proportional Counter Proton recoil proportional counters are filled with hydrogen–argon or methane–argon gas mixtures or the inner walls of the chamber are lined with hydrogen-containing plastics; in many devices both hydrogen gas and plastic lining are combined. Dependent on the gas filling, the gas pressure, and the dimensions of the active counter volume, neutrons with energies between about 10 keV and about 10 MeV can be detected. In Fig. 16, a cross section of a methane–argon-filled proportional counter is shown. The shaded areas represent hydrogen-containing plastic material covering the inner wall of

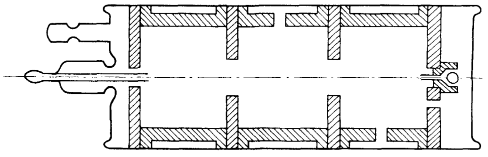

FIG. 16. Cross section of a methane–argon-filled proton recoil counter. The shaded areas represent hydrogen-containing plastic.

the detector. Because of the high specific ionization of fast secondary protons, a suppression of a gamma background by pulse-height discrimination is easy to realize. If the counter is equipped with suitable electronics, and if the counter volume is large enough to stop all secondary protons, neutron spectra can be determined by this type of counter. One disadvantage of the proton recoil proportional counter is the relatively low sensitivity (<0.1 count per *n*/cm² s) as compared with, e.g., (moderated) BF_3, He, and LiI detectors. On the other hand, neutron dose rates up to about 1 Gy/h (10^5 *n*/cm² s) can be detected.

2.3.2 Proton Recoil Scintillation Counter In this type of counter, hydrogen-containing organic scintillators are used as neutron converters, e.g., plastic scintillators, stilbene and anthracene crystals, and also liquid scintillators. Since it is possible to use large scintillator dimensions (especially with liquid scintillators), the detection probability can be made very high even at neutron energies up to about 100 MeV. With liquid scintillators, the detection efficiency at moderate energies approaches 100%. The sensitivity of organic scintillators to gamma radiation and also to competing processes within the scintillator, e.g., carbon recoil nuclei and carbon (*n*,*α*) reactions, is high (Fig. 17). Contributions from those interfering signals must be suppressed by using pulse-height and pulse-shape discrimination techniques. The lower energy limit of this detector type is about 0.5 MeV on account of nontrivial discrimination problems. For quantitative neutron-flux measurements using large scintillators, multiple scattering of neutrons within the scintillator, resulting in pulse summation, has to be considered. Using suitable electronics, neutron-energy spectra between about 0.5 and 100 MeV can be measured with the proton recoil scintillation detector.

Proton recoil scintillation counters are also built using scintillators that are covered with thin inorganic films as, e.g., ZnS(Ag) together with polyethylene, paraffin, or Plexiglas. Also, powdery mixtures of ZnS(Ag) with plastic scintillators are built, but those counters are rarely in use.

2.3.3 Proton Recoil Semiconductor Detector In this version of the proton recoil detector, the *n*-*p* converter is placed onto a silicon surface-barrier detector in a manner that 50% of the active surface is covered by the converter. Since the undesired background reactions are equal in both parts of the barrier detector, the differential pulse spectrum gives the pure proton recoil spectrum. The device is suitable for measurement at higher neutron fluxes.

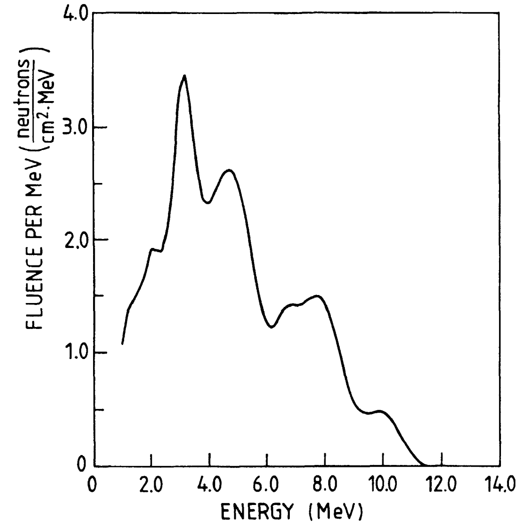

FIG. 17. Neutron spectrum from a ^{239}Pu–Be source measured with a liquid scintillation detector (Ing *et al.*, 1985).

2.3.4 Nuclear Track Detector Proton recoil track detectors are passive devices and are used predominantly in radiation-protection dosimetry. CR 39, an organic carbonate, is a very proton-sensitive material within a limited energy range. The *n*-*p* scattering cross section of CR 39 is large and varies smoothly with neutron energy. The ranges of protons in the detector are large compared with that of heavy ionizing particles, and so CR 39 is more sensitive to neutrons than other materials. Proton-induced tracks are counted after chemical or electrochemical etching of the sheets; also spark-counting techniques are applied. 1- to 10-MeV neutrons can be detected. Because of its sensitivity to *α* particles, e.g., from environmental radon, background tracks must be taken into consideration. For enlarging the response to fast and energetic neutrons, the detector can be combined with a hydrogeneous radiator sheet. The response of CR 39 devices is about 10^3 tracks/cm^2 per mSv dose of fast neutrons.

Besides the n-p scattering, (n,α) and (n,p) radiator sheets are applied with track detectors. Because of the smaller cross sections (see Fig. 10) and the lower recoil ranges within the detector, the response of these devices is poor.

2.4 Activation and Threshold Detectors

Activation detectors are passive devices, which are used to measure neutron fluxes by the aid of various nuclear reactions. Dependent on the slope of reaction cross sections (Fig. 18) with neutron energy, the detectors are classified as thermal, resonance, and threshold detectors.

2.4.1 Activation Detector Activation detectors in a narrower sense are thermal- and low-energy neutron detectors based on (n,γ) reactions (Table 1), though fast and energetic neutrons can also be detected by the use of moderators. The neutron flux can be calculated from gamma- and/or beta-emission intensities of the activated materials, and from the time of exposure to the neutron flux. It is necessary to take into consideration the self-absorption, especially of betas, within the detectors and also the decay of the reaction products. For the determination of gamma and beta activities, various monitors may be used: large-area gas counters, scintillation counters, and semiconductor counters. Because of the self-absorption effects—especially when measuring beta activities—the materials are mostly used as thin films or plates.

Table 1. Slow-neutron detectors

Reaction	Half-life of product	Energy of gamma ray (MeV)	Resonance energy (eV)
$^{23}\mathrm{Na}(n,\gamma)^{24}\mathrm{Na}$	15.0 h	1.37, 2.75	2850
$^{55}\mathrm{Mn}(n,\gamma)^{56}\mathrm{Mn}$	2.58 h	0.85, 1.81	337
$^{59}\mathrm{Co}(n,\gamma)^{60}\mathrm{Co}$	5.26 a	1.17, 1.33	132
$^{63}\mathrm{Cu}(n,\gamma)^{64}\mathrm{Cu}$	12.8 h	0.51	580
$^{115}\mathrm{In}(n,\gamma)^{116}\mathrm{In}^{m}$	54 min	1.09, 1.29	1.46
$^{181}\mathrm{Ta}(n,\gamma)^{182}\mathrm{Ta}$	115 d	1.12, 1.22	4.3
$^{186}\mathrm{W}(n,\gamma)^{187}\mathrm{W}$	23.9 h	0.48, 0.69	18.8
$^{197}\mathrm{Au}(n,\gamma)^{198}\mathrm{Au}$	2.70 d	0.41	4.9

In Fig. 18, cross sections of some activation detector materials are presented. A number of materials have cross sections $\sigma \sim 1/E^{1/2}$ over larger neutron-energy ranges (e.g., Mn), others have strong resonances at certain energies (e.g., In and Rh at about 1 eV). The spectral distribution of low-energy neutrons can be determined if a number of detectors are used simultaneously.

2.4.2 Threshold Detector Neutrons of higher energies can be detected by threshold detectors utilizing, instead of activation detectors, (n,p), $(n,2n)$, or (n,f) reactions, which are induced by neutrons above a certain threshold of the neutron energy (Fig. 19 and Table 2). Threshold detectors have low cross sections compared with activation de-

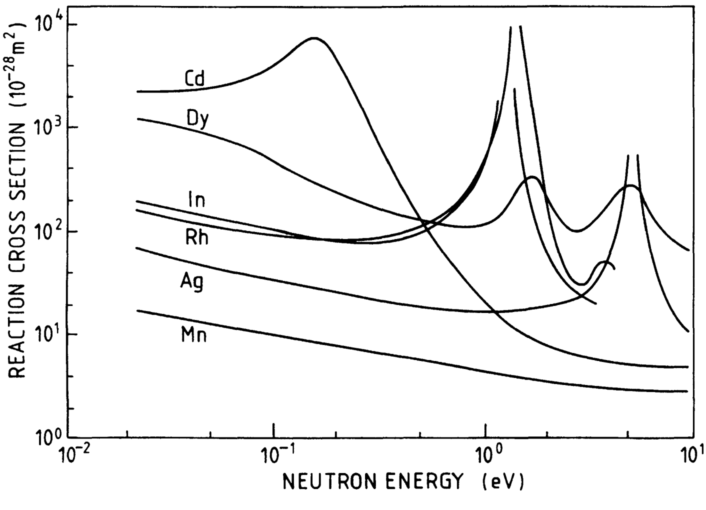

FIG. 18. Nuclear cross sections of common activation and resonance detector elements vs neutron energy (Wirtz and Beckurts, 1958); 10^{-28} m^2 = 1 barn.

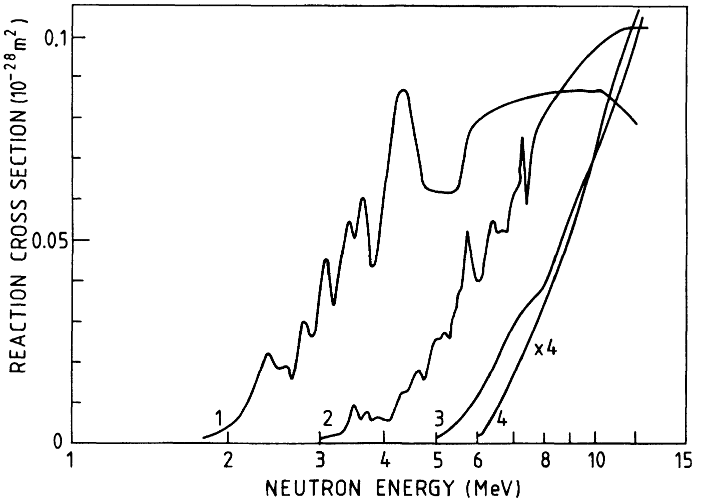

FIG. 19. Nuclear cross sections of some threshold detector elements vs neutron energy (Bresisti, 1963); 10^{-28} m^2 = 1 barn. Curve 1, ^{32}S(n,p)^{32}P; curve 2, ^{27}Al(n,p)^{27}Mg; curve 3, ^{56}Fe(n,p)^{56}Mn; curve 4, ^{27}Al(n,α)^{24}Na.

tectors. The beta and gamma emitters produced by the reactions within the detector material are normally measured with large-area gas counters, scintillation counters, or semiconductor counters. A less common method is the use of scintillator counters where the threshold material is part of the scintillator material. Together with activation detectors they are used in the presence of broad neutron spectra as, e.g., in nuclear facilities or at high-energy accelerators.

2.5 Moderated Detectors

The measurement of fast and energetic neutrons by the aid of active devices is mostly carried out with low-energy neutron detectors surrounded by suitable moderators, because of the large cross sections of low- and thermal-energy neutron detector materials, e.g., with BF_3 proportional counters and LiI scintillation counters (see above).

Table 2. Threshold detectors

Reaction	Half-life of product	Energy of gamma ray (MeV)	Approximate threshold energy (MeV)
^{12}C(n,2n)^{11}C	20.3 min	0.51	20
^{12}C(n,spall)^{7}Be	53.6 d	0.48	30
^{24}Mg(n,p)^{24}Na	15.0 h	1.37, 2.75	7.5
^{27}Al(n,p)^{27}Mg	9.5 min	0.84, 1.01	3.8
^{27}Al(n,α)^{24}Na	15.0 h	1.37, 2.75	4.9
^{27}Al(n,spall)^{22}Na	262 a	0.51, 1.28	25
^{32}S(n,p)^{32}P	14.3 d	1.71 (β)	3.3
^{56}Fe(n,p)^{56}Mn	2.58 h	0.85	5.0
^{58}Ni(n,p)^{58}Co	71.3 d	0.51, 0.81	1.3
^{58}Ni(n,2n)^{57}Ni	36.0 h	0.51, 1.37	12.6
^{65}Cu(n,p)^{65}Ni	2.56 h	1.12	3.2
^{65}Cu(n,2n)^{64}Cu	12.7 h	1.35, 0.51	10.3
^{115}In(n,n')^{115}In^{m}	4.5 h	0.34	1.5
^{197}Au(n,spall)^{149}Tb	4.1 h	0.17	600
^{209}Bi(n,f)			50
^{31}P(n,p)^{31}Si	2.6 h	1.5 (β)	2.7

2.5.1 Long Counter Long counters are used for precise measurements of neutron-flux densities in the energy range from 10^{-2} to 10^{7} eV. A sketch of a long counter is presented in Fig. 20: A BF_3 proportional counter is surrounded by a suitably designed polyethylene or paraffin moderator, which enables the detection of neutrons with nearly energy-independent sensitivity. The cadmium shield prevents the penetration of low-energy stray neutrons from outside, because cadmium is a strong low-energy neutron absorber. The detection efficiency of such devices is about 0.1 to 1%, i.e., one count per 10^2 to 10^3 neu-

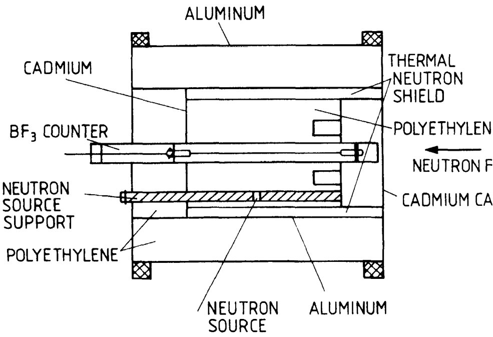

FIG. 20. Precision "long counter" for measuring fast-neutron flux density (de Pangher and Nichols, 1966).

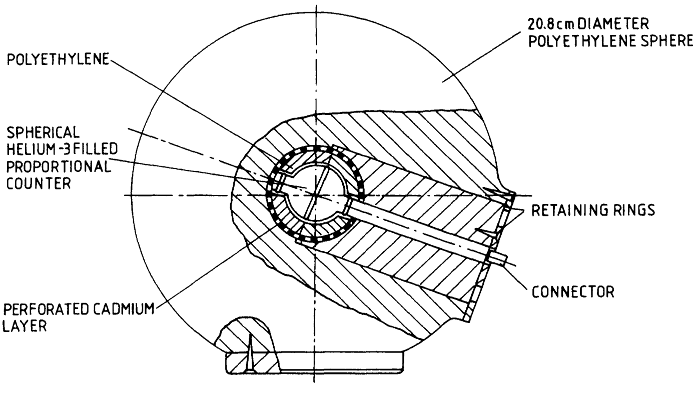
POLYETHYLENE
SPHERICAL HELIUM-3 FILLED PROPORTIONAL COUNTER
PERFORATED CADMIUM LAYER
20.8cm DIAMETER POLYETHYLENE SPHERE
RETAINING RINGS
CONNECTOR

FIG. 21. Cross section of the Leake-type "Rem counter" (Leake, 1966).

trons. It is also anisotropic with a strong maximum in the forward direction. The cadmium cap at the front side of the instrument can be used to separate thermal- and low-energy neutrons from the faster ones.

2.5.2 Rem Counter Rem counters are dose-equivalent survey meters, constructed for area monitoring in radiation protection. Their sensitivity to neutrons of different energies between 10^{-2} eV and 10^{7} eV should approach the dose equivalent per unit neutron-flux function as far as possible. The latter accounts for the biological effectiveness of neutrons as a function of their energy, and is defined by the International Commission on Radiological Protection. The Rem counter consists of a BF_3 proportional counter or more rarely of a LiF scintillation or a helium-3 counter, surrounded by a spherical or a cylindrical polyethylene moderator of suitable dimensions. In Fig. 21, one device from a variety of commercial products is presented. The proportional counter is surrounded by an inner polyethylene sphere and by a spherical perforated cadmium layer, in order to give a reduced response for lower-energy neutrons. This is because lower-energy neutrons have a smaller biological effectiveness than fast ones. Nevertheless, in the presence of a broad neutron-energy distribution the results are not exact; therefore instruments are available that have two or three replaceable moderator shells of different thickness, thus allowing more accurate dose measurements. The indicating instrument gives the result of measurement directly in equivalent dose-rate units Sv/h. The dose-rate range covered by Rem counters is commonly 0.1 μSv/h to 0.1 Sv/h. Though an isotropic response of the instruments would be very desirable for area monitoring, the response is a more or less strong function of the angle of neutron incidence. Figure 22 gives the response of Rem counters with respect to neutron energy in a unidirectional neutron field.

2.5.3 Multisphere Detection ("Bonner" Spheres) In order to overcome inaccuracy of single-sphere moderated counters in broad-energy neutron fields, precision measurements of neutron fluxes are carried out

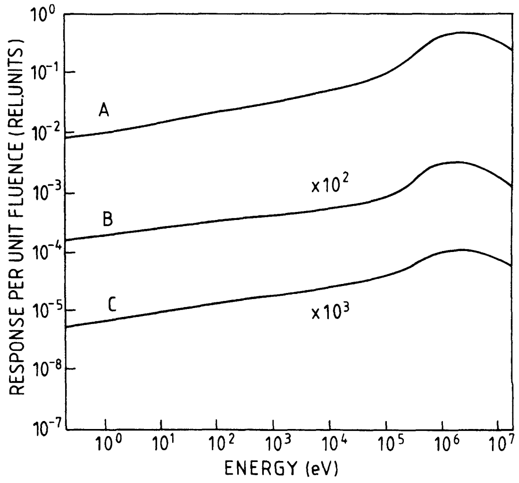

FIG. 22. Responses of Rem counters from different manufacturers. Curve A, "Snoopy" cylindrical moderator/BF_3 detector Rem counter (Hankins, 1977); curve B, Leake 95/0075 spherical moderator/helium-3 detector Rem counter (Harrison, 1979); curve C, PNR-4 spherical moderator/BF_3 detector Rem counter (Hankins, 1977).

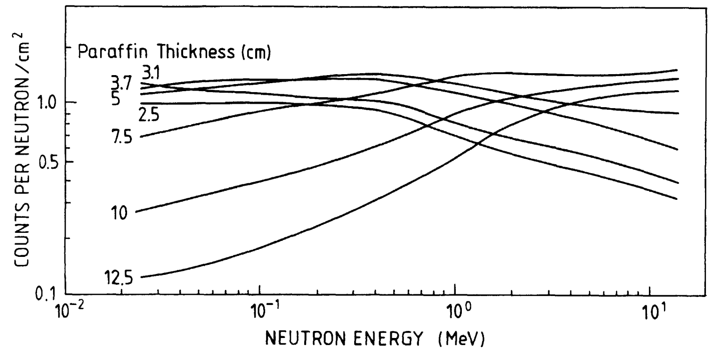

FIG. 23. Count rate of a paraffin-moderated BF_3 proportional counter vs neutron energy at different moderator thicknesses (Wallace, 1960).

with up to ten or even more replaceable moderator spheres of different diameters. The count rate C_i of each counter–moderator assembly is given by

$$C_i = \int \varphi_E(E) \times R_i(E)dE,$$

where $\varphi_E(E)$ is the spectral neutron-flux density, $R_i(E)$ the energy-dependent response, and E the neutron energy. By the aid of mathematical unfolding techniques, the spectral neutron-flux distribution $\varphi_E(E)$ can then be calculated if the functions $R_i(E)$ are well known. Responses of multisphere assemblies are given in Figs. 23 to 25. Similar measurement procedures are applied using passive low-energy detectors, e.g., activation foils. The energy resolution of multisphere detectors is poor as compared with, e.g., semiconductor or helium-3 detectors, but in many applications the covering of a wide neutron-energy spectrum is of over-riding interest.

2.6 Current Developments

2.6.1 Bubble and Drop Detectors The basic principle of this technique is well known from conventional bubble chambers. The boiling of a superheated liquid is initiated by neutrons via recoil nuclei, which deposit their kinetic energy by ionization along their track in the superheated liquid. Small devices are under development for radiation-protection purposes that are much less expensive than conventional bubble chambers. By embedding the liquid, preferably freon 12, in a gel, a large number of isolated very small drops are homogeneously dispersed

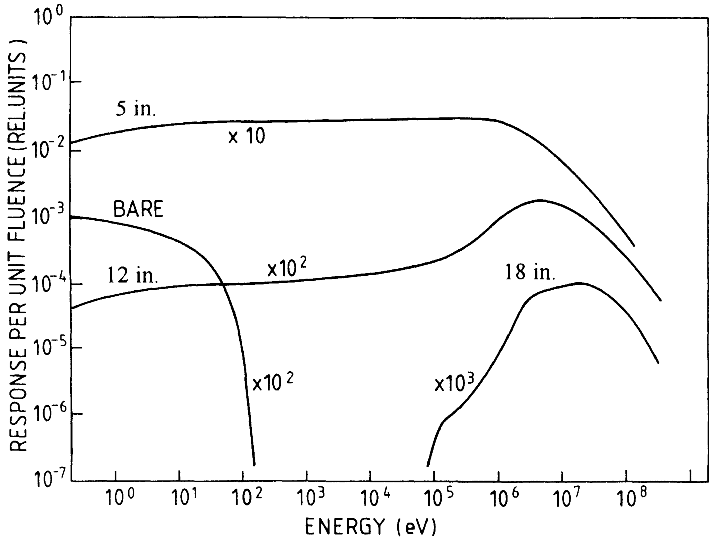

FIG. 24. Response of polyethylene-moderated LiI(Eu) scintillation detector vs neutron energy at different moderator thicknesses (Nachtigall and Burger, 1972).

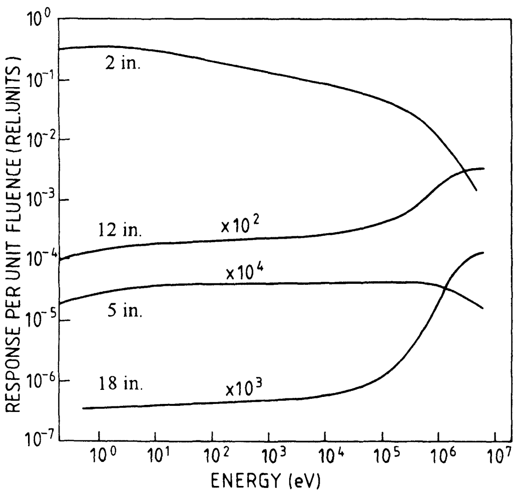

FIG. 25. Response of a polyethylene-moderated helium-3 detector vs neutron energy at different moderator thicknesses (Caizergues and Poullot, 1977).

within the gel matrix. Vaporization of drops due to neutron-induced events is detected by the aid of a piezoelectric transducer, which records the small-pressure pulses. The device enables the detection of neutrons of thermal to very high energies (Apfel and Roy, 1985).

Another approach to the use of superheated liquids for neutron detection has been utilized by Ing and Birnboim (1984): They use a solid polymer as a matrix for the small superheated drops. After vaporization by neutron-induced events, tiny immobile bubbles are formed in the polymer matrix. The bubbles can be observed visually and can be counted by simple optical instrumentation.

2.6.2 Tissue-Equivalent Proportional Counter Tissue-equivalent proportional counters (TEPC) are under development for radiation-protection monitoring in mixed-radiation fields. The TEPC chamber is constructed from tissue-equivalent wall material and is filled with tissue-equivalent gas at very low pressure. If suitably designed, the microprocessor-based TEPC gives a direct measure of absorbed radiation dose. It is possible to measure neutron, gamma, and charged-particle dose simultaneously, and to separate their contribution to dose by sophisticated electronics and mathematics. Equivalent doses and radiation quality factors can be determined numerically, too (Brackenbush *et al.*, 1985). The instruments are reasonably accurate at neutron energies between 100 keV and 14 MeV, which covers most occupational neutron exposure. TEPCs are constructed for area monitoring, and as miniaturized devices for personnel monitoring.

2.7 Neutron Spectrometry

Most of the detectors described above allow neutron-intensity measurements if the neutron spectrum is well known, but not the determination of energies in an unknown neutron field. In many applications, only rough knowledge of the neutron-energy distribution is necessary. Then, a suitable set of passive activation and threshold detectors offers a relatively simple and inexpensive chance of measurement. The application of multisphere techniques with active or passive detectors is also feasible over the total neutron-energy range, but the mathematical expenditure of unfolding procedures is large. Spectrometry in limited energy ranges is possible by the aid of BF_3, helium-3, or scintillation counters, and semiconductor detectors using discrimination and coincidence techniques. More expensive are techniques in which preselection of neutrons with respect to their energy is achieved by mechanical energy or velocity analysis, neutron diffraction, or time-of-flight analysis. With a suitable detector or if necessary with a set of detectors and if necessary with suitable moderators, the range between thermal and about 100 MeV energy can be covered spectrometrically.

In "neutron telescopes," recoil protons are produced in a thin hydrogen-containing foil (Fig. 26). Only protons that are scattered straight forward are energy-analyzed by the

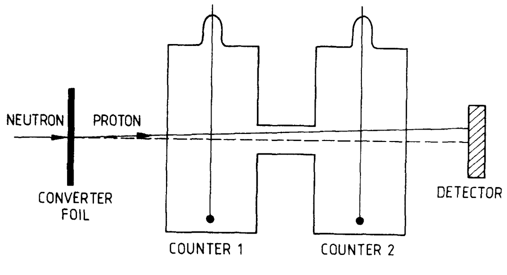

FIG. 26. Cross section of a simple neutron telescope for neutron-energy spectrometry.

detector, because they receive the total neutron energy through elastic collisions. The counting of only forward-scattered protons is enabled by operating counters 1 and 2 and also the proton detector in coincidence. Through the use of two counters, the background from stray neutrons can be appreciably reduced, compared with systems operating with only one counter + detector. A variety of counting and detecting devices is used in different types of neutron telescopes.

3. BETA-PARTICLE DETECTORS

The range of beta particles (electrons and positrons) in matter is generally much smaller than that of gamma quanta and neutrons. So self-absorption in a beta source and absorption in the detector wall must be considered. In addition, beta particles from radioisotopes have always a broad energy distribution. That is, if (self-)absorption occurs, measured beta spectra will not be identical with the original emission spectra. In Fig. 27(a) a typical beta spectrum and in Fig. 27(b) the transmission of beta particles through counter windows are shown.

Most of the artificial radioisotopes are beta emitters, and most of them emit simultaneously gamma radiation. So the activity of an isotope or an isotope mixture may in most cases be determined by the aid of a gamma detector in a simpler manner. But there are also pure beta emitters, e.g., ^{3}H, ^{14}C, ^{32}P, ^{35}S, ^{204}Tl, which are important in some applications. On the other hand, the assay of low-level beta–gamma sources is sometimes carried out by the aid of beta counters, because most of these detectors have higher efficiencies than gamma detectors. In most cases, sensitive detection of beta emission needs shielding against gamma and cosmic background radiation. Additional anticoincidence shielding leads to further reduction of background radiation.

Aside from basic scientific research, beta detectors are used in fuel-cycle facilities as well as in isotope-production facilities and in nuclear medicine departments, e.g., for process control and activity determination. In all those facilities, control of personnel exposure mainly by incorporation, and control of the release of isotopes to the environment via waste water and air is necessary.

3.1 Detectors for Activity Measurements

The activity of a source is defined as the number of nuclear disintegrations per second. This must not be confused with "beta activity" of a source, since there are isotopes that emit more than one beta particle per decay and in most cases there is simultaneous emission of one or more gamma quanta. Moreover, exact measurement is complicated if the daughter product is a beta emitter, too. Then the total activity is a function of time until equilibrium is reached (see RADIOACTIVITY). If precision is necessary, all emitted particles must be detected under 2π or 4π geometry. Otherwise the response of an instrument to betas of different energy must be well known by calibration, and also the nature of the source must be well known.

3.1.1 Liquid and Solid Scintillation Counters The liquid scintillation counter (see DETECTORS, SCINTILLATION) is in general a useful tool in activity measurement, but it is of particular importance for the assay of low-level sources, and in the presence of low-energy pure beta emitters such as ^{3}H and ^{14}C, which have maximum beta energies of only 18 and 156 keV, respectively. As a result of the homogeneous solution of the isotopic compound in the scintillator liquid, there is no primary loss of beta ionization energy. But by mixing of the two liquids quenching effects occur, and so the measurable activity is generally limited to small quantities. Background events have to be considered especially in liquid scintillation detectors, since undesired light production is possible that may even exceed that resulting from the sample decay, e.g., phosphorescence or chemiluminescence from reactions of the mixed liquids or by impurity contents. Since precision measurements are carried out by comparison with standard solutions of the isotopes of concern, the chemical identity of both samples is of importance, otherwise different background events may lead to wrong results, especially with low-energy emitters. Careful calibration is always necessary. A large number of labeled compounds are commercially available for comparative assay of nearly all isotopes of general interest. The detection probability for ^{3}H is up to 65%, and it reaches nearly 100% at higher beta energies. Though the energy res-

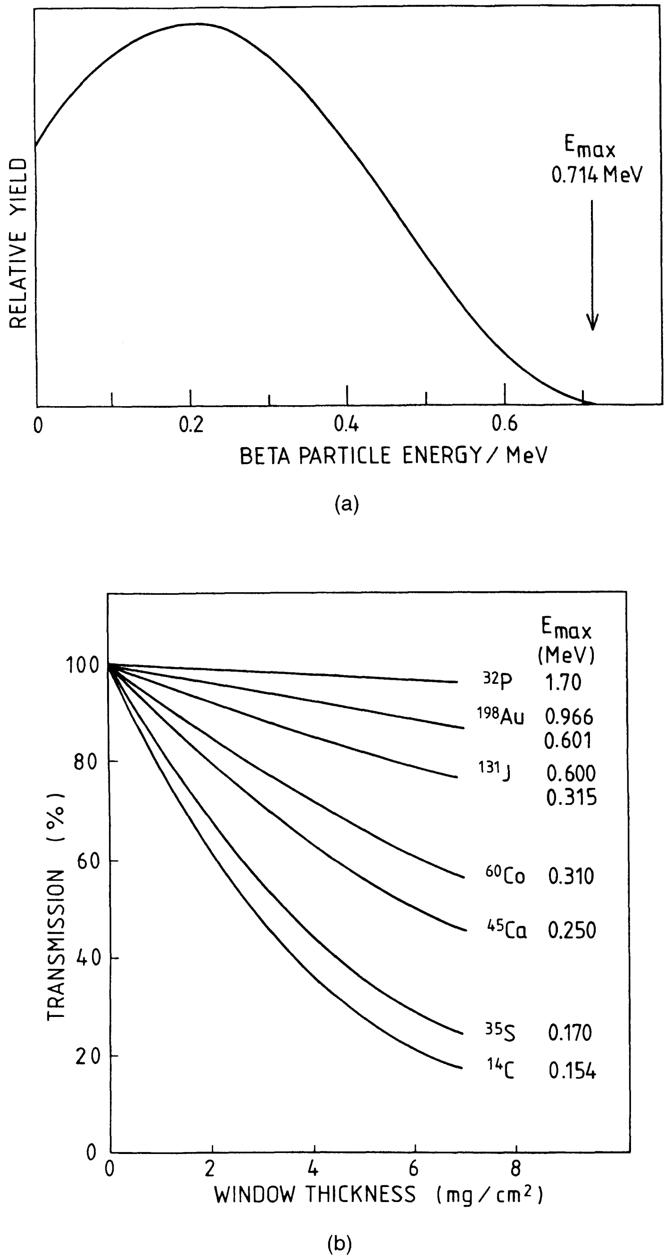

(a)

(b)

FIG. 27. Beta spectrum and transmission of beta particles through thin foils. (a) Energy distribution of beta particles emitted from ^{36}Cl. (b) Transmission of beta particles from different isotopes through window material of beta detectors (Gleason, 1951).

olution of the liquid scintillation counter is relatively poor, simultaneous measurement of isotope mixtures is possible under certain conditions (dual-label assay of two-component beta–beta or beta–alpha mixtures). By use of multichannel analyzers, suitable energy-window setting, and additional computer-supported fitting of the slope of measured spectra, the activity components of mixed isotopic compounds can be determined with reasonable success in many cases. Gaseous isotopes can also be measured with the liquid scintillation counter either by introducing the gas directly into the scintillator, or by trapping the gas in suitable liquid media, which then are mixed with the scintillator.

In solid scintillation counters for beta-particle detection, generally plastic scintillators and also organic crystals such as anthracene or stilbene are used. Those counters are well suited for beta spectrometry. By the application of large specially shaped scintillators, the sensitivity to betas but also to background gamma radiation is high. By the aid of heavy-metal and anticoincidence shielding, the background can be reduced to below 0.01 counts/s. In Fig. 28, the cross section of such an instrument is presented. By combination of two counters, a scintillator and an additional Geiger-Müller device, the gamma background is suppressed electronically.

3.1.2 Proportional and Geiger-Müller Counters Gas-multiplication detectors can be applied in solid and gaseous beta-source assay. An example of a vast variety of counter designs to be used with solid sources is shown in Fig. 29. Those counters have very thin plastic or metal foil windows with thicknesses of 0.1 to 10 mg/cm^2 in order to allow the passage of beta particles. Usually the counting gases are methane or argon–methane mixtures (proportional counter), and helium–butane mixtures (Geiger-Müller counter) at atmospheric pressure. For low-level activity assays, those devices are operated in the anticoincidence mode and are heavily shielded. On account of self-absorption in the beta source, absorption in the window, and backscattering, expensive calibration procedures are necessary. Internal counters are used for measurement of gaseous isotopes such as ^{3}H, ^{14}C, and ^{85}Kr. In most cases, the isotopes are mixed with a normal counting gas; more rarely, only the isotope acts as counting gas. Those counters exist in a large variety of designs. In low-level applications they consist usually of a

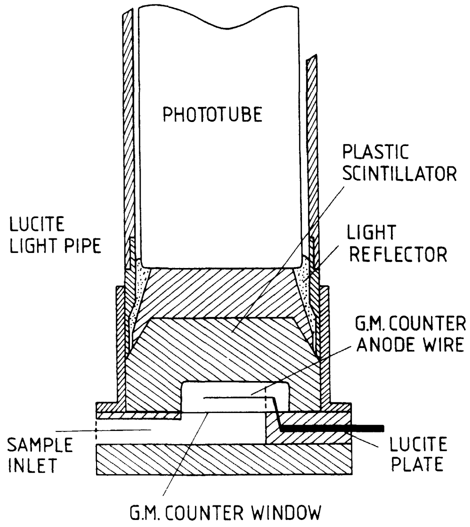
PHOTOTUBE
PLASTIC SCINTILLATOR
LUCITE LIGHT PIPE
LIGHT REFLECTOR
G.M. COUNTER ANODE WIRE
SAMPLE INLET
LUCITE PLATE
G.M. COUNTER WINDOW

FIG. 28. Cross section of a beta-particle scintillation spectrometer using a Geiger-Müller counter for coincidence detection (Tanaka, 1961).

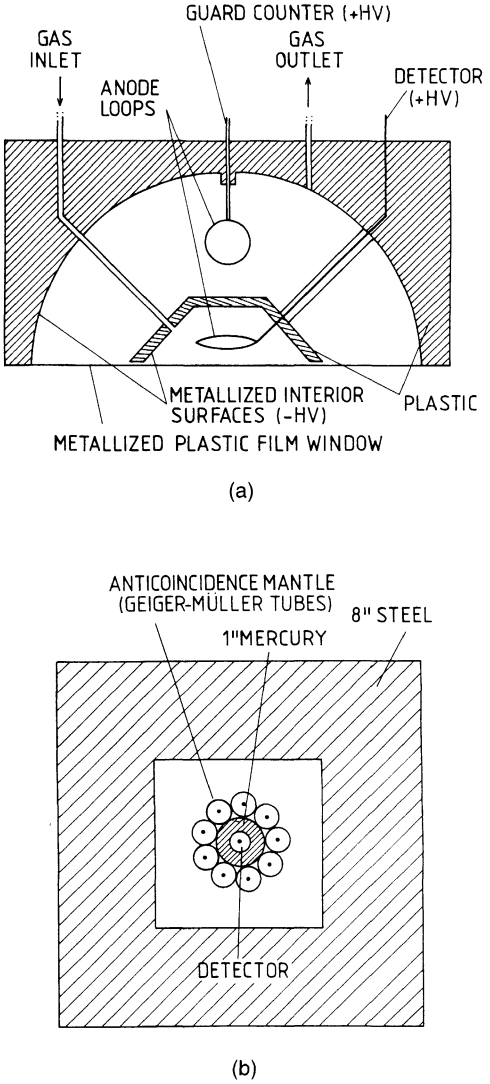
GUARD COUNTER (+HV)
GAS INLET
GAS OUTLET
ANODE LOOPS
DETECTOR (+HV)
METALLIZED INTERIOR SURFACES (-HV)
PLASTIC
METALLIZED PLASTIC FILM WINDOW
(a)
ANTICOINCIDENCE MANTLE (GEIGER-MÜLLER TUBES)
8" STEEL
1"MERCURY
DETECTOR
(b)

FIG. 29. Cross section of a gas-multiplication counter with guard detector for anticoincidence shielding (Johnson *et al.*, 1963): (a) with internal guard detector; (b) with external guard detector and heavy-metal shielding.

long metal or quartz cylinder with axial anode wire; their counting volumes range from some milliliters to several liters. Effective shielding as described above is necessary.

3.1.3 Ionization Chamber Like the proportional counter, the closed ionization chamber is disadvantageous if solid sources, especially low-energy beta emitters, are to be investigated. Nevertheless if suitable thin-film sources of not too low beta energies are used, the ionization chamber operated in the current mode is less sensitive to sample self-absorption than a pulse counter: Each high-energy beta particle creates thousands of ion pairs that contribute to the ionization current, but it creates only one pulse in the counter-type detector. The low-energy betas, which mainly undergo self-absorption, produce fewer ion pairs compared with the high-energy ones, but each represents one lost pulse in the counter. In Fig. 30, the construction of a combined beta–gamma reference ionization chamber is shown. It consists of a parallel-plate beta chamber and a cylindrical gamma chamber. Both solid and liquid sources can be introduced into the chamber below the beta chamber. Those chambers are available with fixed calibration factors for a large number of nuclides. Ionization chambers of this type are widely used as secondary standard instruments for the calibration of beta and gamma sources, because of their ease of operation with excellent reproducibility.

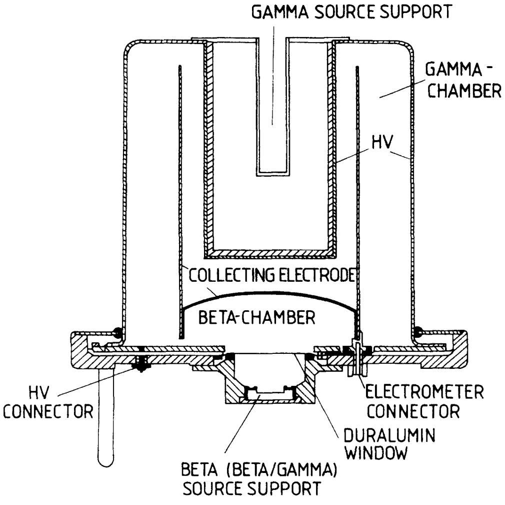

FIG. 30. Cross section of a beta–gamma ionization chamber (Robinson, 1960).

Internal ionization chambers can be used for the assay of radioactive gases such as ^{3}H, ^{14}C, and ^{85}Kr. A well-known amount of gas is filled into the chamber, and the measured ionization current is a direct measure of the gas activity. Mostly, comparative measurements are carried out by using gaseous standards.

3.1.4 Cherenkov Counter Beta particles in aqueous solutions can be detected by the aid of Cherenkov counters, if the beta energy is greater than 0.26 MeV (see ELECTROMAGNETIC RADIATION). The measurement technique is similar to liquid scintillation counting and in laboratory practice generally such equipment is applied. In practice, mostly water is used as light-producing medium. The advantage of this technique is the relative ease of sample preparation: The solid or liquid beta source is placed in a small water-filled glass or plastic vial, and the activity is counted in analogy to scintillation counting. A further advantage over liquid scintillation counting is the absence of chemical quenching effects, due to the different light-creation process. Since the Cherenkov radiation is emitted at short wavelength and anisotropically in the forward direction, a wavelength shifter is applied when using liquid scintillation coincidence counting equipment. The wavelength shifter effects a better adaption of the wavelength of light to the phototube spectral sensitivity, and it also reduces the light anisotropy due to the isotropic emission of the secondary light. The latter effect enhances the sensitivity of the equipment, because the phototubes are usually arranged diametrically opposed in liquid scintillation counters. Though the number of detectable isotopes is limited by the detection threshold, and furthermore the detection efficiency is lower than in scintillation counting, the technique is often used because of the ease of operation. Table 3 gives some examples of detection efficiencies, which strongly depend on beta energies of isotopes.

3.1.5 Beta-Particle Spectrometry Precision beta spectrometry is carried out mainly by the aid of spectrometers based on the deflection of particles in electric and

Table 3. Cherenkov-counting efficiencies for various beta-emitting nuclides (Francois, 1973). The efficiencies were measured using a liquid-scintillation counting system, with two diametrically opposed phototubes in coincidence.

Isotope	E_{mean} (MeV)	E_{max} (MeV)	Detection efficiency (%)
^{204}Tl	0.240	0.764	3
^{198}Au	0.312	0.961	6.7
^{24}Na	0.554	1.390	32.5
^{40}K	0.561	1.312	38.4
^{89}Sr	0.583	1.488	28.3
^{32}P	0.695	1.710	35.1
^{90}Y	0.935	2.284	50.6
^{76}As	1.064	1.75 (16%) 2.41 (31%) 2.97 (50%)	59
^{42}K	1.430	1.996 (17.5%) 3.521 (82.1%)	70.3

magnetic fields, having an energy resolution of about 10^{-4}. Those expensive techniques are applied only in basic scientific research. In practical application, the task is mainly to identify a beta emitter or mixtures of beta emitters by gross energy-distribution measurement.

The measurement of beta absorption in thin sheets of different thickness is a very simple method. When the counting rates are plotted vs sheet thickness on a logarithmic scale, an extrapolation of the curve to zero counting rate gives the maximum beta energy of the unknown isotope. The technique is time consuming and not very exact, especially in the presence of a gamma background and/or a mixed beta source.

The method of choice in gross beta spectrometry is scintillation counting supported by multichannel analysis. Suitable scintillators are plastics and organic crystals like anthracene or stilbene. The thickness of scintillators should be 1 to 2 cm, because total absorption of beta particles within the material is necessary. An advantage of organic scintillators over inorganic ones is the low Z, which causes low backscattering rates. A further reduction of backscatter is possible by using specially shaped scintillators (see, e.g., Fig. 28). Because of the high sensitivity of scintillation detectors to gamma radiation (anti-)coincidence techniques should be applied. In principle, also silicon surface-barrier or diffused-junction semiconductor detectors (see DETECTORS, SEMICONDUCTOR) are suitable for beta spectrometry, but their detection efficiency is lower than that of scintillation detectors, and their high energy resolution is hardly needed because of the broad beta-energy distributions [see Fig. 27(a)]. Therefore these detectors are rarely used in beta spectrometry. Thin-window proportional counters are well suited for low-energy beta spectrometry; below 100 keV their energy resolution exceeds even that of scintillation detectors. Because of the range of beta particles in gases (some meters at 1 MeV), they are not stopped totally within the counter volume at higher energies, and the measured spectra would be distorted. The upper energy limit is enhanced by using high-pressure counters.

3.2 Detectors for Area and Environmental Control

In all workplaces where beta and beta–gamma sources are handled, and the possibility exists that gases, vapors, liquids, or dusts are released from open radioactive sources or even from normally closed sources in hazardous amounts, area monitoring, personnel monitoring, and room-air and waste-water and -air monitoring are necessary. Various equipment is needed depending on the monitoring purpose and the kind of radioactive sources. Equipment of workplaces, tools, and work clothes can be controlled by portable contamination monitors. At the exit of a controlled area, personnel often are checked by large-area lock monitors. In many facilities, ventilation systems are installed and the room air is controlled by suitable monitors. Waste water and air are even analyzed spectrometrically if complex isotopic mixtures may be released under normal or accidental conditions.

3.2.1 Beta–Gamma Contamination Monitors Portable contamination monitors are usually equipped with flat, xenon-filled proportional counters. The windows are made from typically several-mg/cm^2-thick aluminum or titanium foils, and the active areas are about 100 cm^2. Mostly they give an acoustical alarm if certain isotope-related threshold pulse rates are exceeded. In Table 4, typical sensitivities for some isotopes are exemplarily given. It can be seen from the table that the monitor well meets the specifications given by concentration limits (3.7

Table 4. Isotope-related counting rates of a portable beta–gamma contamination monitor (Lab. Berthold, 1984)

Isotope	Specific counting rate ($\mathrm{Bq \cdot cm^{-2}/count \cdot s^{-1}}$)	Counting rate at the limit of 3.7 Bq $\mathrm{cm^{-2}}$ ($\mathrm{s^{-1}}$)
^{14}C	0.2	20
^{60}Co	0.07	50
^{99m}Tc	0.3	13
^{125}I	0.3	13
^{131}I	0.04	100
^{241}Am	0.3	13

Bq/cm^2) in radiation-protection regulations. ^{3}H cannot be detected by this class of monitors because of its low beta energy. If isotopes or isotope mixtures with larger gamma energies are present, these monitors detect also incorporated radioactivity. At many workplaces, the composition of isotopic mixtures is well known and constant. Then, only the control of one well-detectable "leading isotope" is necessary.

Larger counters of the type described above are used with hand or hand–foot monitors, usually with active-window areas of about 1000 cm^2. In large lock monitors, where the whole person is controlled, several counters with some square meters of active-window area are combined.

3.2.2 Room- and Waste-Air Monitors

Room-air monitors installed in ventilated areas are designed for permanent operation. Modern systems have sophisticated electronics including, e.g., permanent recording of data, self-control of important functions, automatic alarming when the concentration of isotopes exceeds certain levels, and long-term stability. Because the kind of isotope or isotope mixture differs appreciably in the various nuclear facilities, and may also differ from workplace to workplace in even one facility, a large variety of room-air monitors has been developed. There are monitors for detection of special nuclides only, and those which detect isotope mixtures, either in separate measuring channels or as sum concentrations. In many cases, it is again possible to detect only "leading isotopes" from a mixture of, e.g., beta–gamma sources. In many facilities where only beta or beta–gamma gases with not too low energies are handled, large-area, thin-window proportional counters (see Sec. 3.1) have sufficient detection sensitivity. It is important that the contaminated air pass the counter window in a continuous and well-defined flow. When radioactive aerosols are present, it is necessary to sample them by the aid of filters, and to measure the filter activity in suitable time intervals. In either case, covering of the counter window by inactive or active dust or aerosols can lead to false results. This must be measured and suppressed automatically by system electronics.

Low-energy beta emitters such as ^{3}H and ^{14}C cannot be detected by these methods, in particular not at low concentration levels as is necessary with waste-air control. The techniques of choice are real-time detection with flow-through counters, and sampling of isotopes by adsorbers and measuring in suitable time intervals with open counters. The latter technique is applied mainly in beta–alpha monitors, and is described in Sec. 4. A main problem with low-energy beta emitters is to perceive quickly a sudden, e.g., accidental, concentration increase in the room air. This needs high flow rates in the counter and also computer-supported automatic action of the system. Mostly, internal methane or methane–argon proportional counters are applied for control of low-energy beta activity in room air. In Fig. 31, the configuration of such a system, designed for independent measurement of ^{3}H and ^{14}C, is shown. The methane–air mixture flows through four flat counters—two screening counters and two measuring counters—which are not separated mechanically but by grounded grids. The two outer "screening counters" are applied for background suppression. By means of the coincidence–anticoincidence system, pulses from the two isotopes can be separated from each other and from background radiation. Additionally, the counter is heavy-metal shielded. The concentrations are not measured continuously but during successive intervals, within which the pulses are summed up. The methane–air volume ratio is 2:1. Each single detector has a volume of 1.2 L. The lower detection limit depends on the integration time. Of course, other gaseous pure beta emitters such as ^{85}Kr or ^{133}Xe can be detected by the aid of this concept. More rarely applied are other sensitive detectors, the lower detection limits of which meet the requirements described above: liq-

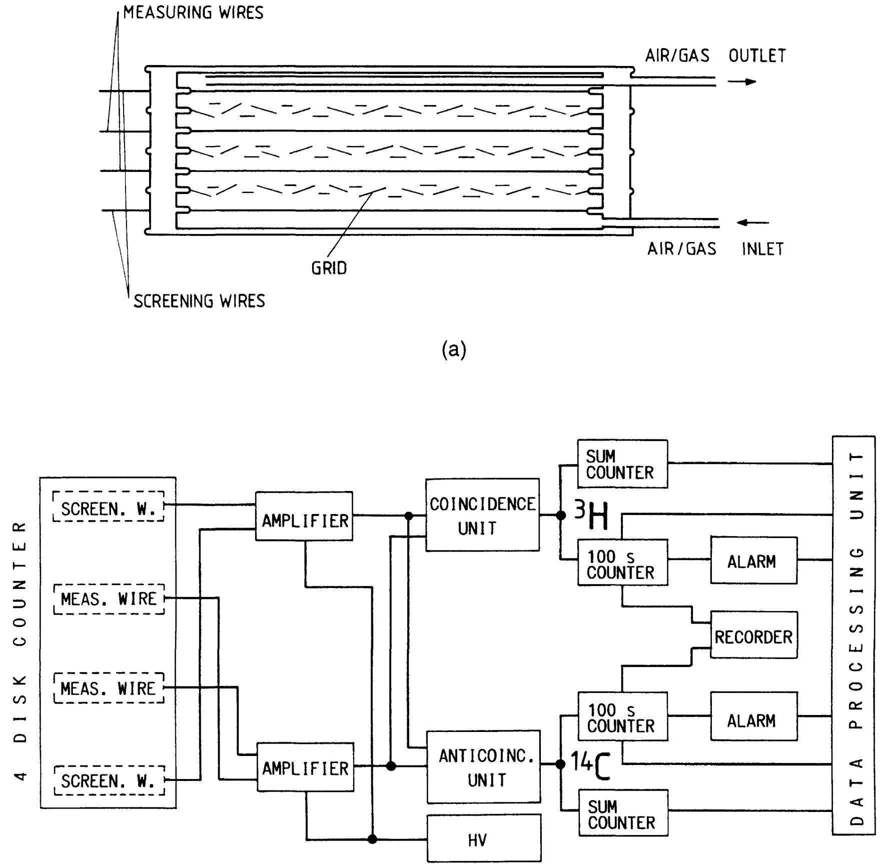
MEASURING WIRES
AIR/GAS OUTLET
GRID
AIR/GAS INLET
SCREENING WIRES
4 DISK COUNTER
SCREEN. W.
MEAS. WIRE
MEAS. WIRE
SCREEN. W.
AMPLIFIER
AMPLIFIER
COINCIDENCE UNIT
ANTICOINC. UNIT
HV
SUM COUNTER
^{3}H
100 s COUNTER
ALARM
RECORDER
100 s COUNTER
ALARM
^{14}C
SUM COUNTER
DATA PROCESSING UNIT

(a)

(b)

FIG. 31. Internal proportional counter system for detection of ^{3}H and ^{14}C gas concentrations in room and waste air (Fricke *et al.*, 1983). (a) Cross section of the "4 Disk" counter. (b) Simplified scheme of data acquisition and processing.

uid scintillation detectors, internal plastic scintillation detectors, and compensated ionization chambers.

3.2.3 Waste-Water Monitors In nuclear facilities, large amounts of waste water are produced, which have to be controlled. Generally, the same detection principles as described above are applied: thin-window proportional counters, plastic scintillation counters, and liquid scintillation counters for monitoring of very low-energy isotopes such as tritium. The simplest technique for monitoring waste water at not too low activity levels and beta energies is the use of diving counters. The detectors are preferably thin-walled, methane- or methane–argon-filled proportional counters. They can be used in streaming water and of course for activity measurements in containers. But their walls must be thicker than those of air detectors, because they must withstand larger pressures, especially dynamic pressure variations in pipes. Therefore wall thicknesses below 30 to 60 mg/cm^2 are not possible, and beta particles with energies of less than about 500 keV cannot be detected. There are also problems with active or inactive films that may be deposited after some time of use at the outer walls, and may lead to false results.

In the presence of lower-energy betas, large-area, thin-window proportional counters and more rarely plastic scintillation counters are applied, which are placed directly above water surfaces. In many cases, when the sensitivity of this method is too low, physical or chemical concentration techniques are applied before measurement. This is especially necessary if ^{3}H has to be monitored in aqueous media. Those measurements are performed using liquid scintillation counters as described in Sec. 3.1.1. A continuous, direct, and sensitive monitoring of ^{3}H in streaming water is impossible.

3.2.4 Beta Monitors for Radiation Dosimetry One task of beta measurement in radiation-protection dosimetry is to monitor the risk of man to direct beta irradiation. Because of the low penetration depth of beta radiation in tissue, this risk is small compared with the risks of incorporation from air or liquids. Therefore it is limited to those persons who deal occupationally with high-energy beta emitters such as ^{90}Sr/^{90}Y, ^{32}P, or ^{234}U or with, e.g., therapy electron accelerators. In these cases, there is a need for area and personnel monitoring at workplaces. On the other hand, high sensitivity of detectors is not necessary in this field.

For area monitoring, there is a demand for robust and handy instruments. The monitors (so-called "survey monitors") are usually air-filled ionization chambers and are operated at normal pressure. They have thin walls (several mg/cm^2) that preferably consist of tissue-equivalent materials like plastics. Chamber dimensions are between about 200 and 500 cm^3. The devices are normally not of special design for beta measurements, and the results should be taken with care, unless the response of the chamber is calibrated in a well-known reference field.

More exact beta monitors are developed in order to allow measurements in beta fields of unknown energy distribution, e.g., using a tissue-equivalent plastic scintillator combined with a multichannel analyzer for beta-spectrum determination. Beta dose rates can so be measured with an accuracy of about 25% if the gamma background is not excessive.

The "extrapolation chamber" is an instrument for precise and absolute determination of the absorbed dose rate in tissue-equivalent material. It is a special ionization chamber with parallel electrodes (Fig. 32). The volume of the filling gas can be varied by the aid of a movable piston thus making it possible to extrapolate the dose rate to zero thickness. When used in a tissue-equivalent phantom at different depths, the strongly depth-dependent absorbed dose by betas of different energy can be measured. For beta emitters with maximum energy >800 keV, the relative uncertainty of results is better than ±2%. Measurement is possible down to 50 keV but with increased uncertainty.

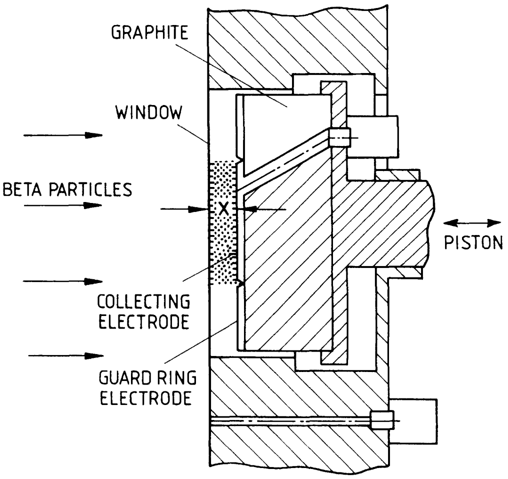

FIG. 32. Beta-particle "extrapolation chamber" for precision measurements of beta dose rates. The dotted area represents the active detector volume (Böhm, 1984).

3.3 Passive Beta Detectors

Passive beta particle detectors are mainly used in personnel monitoring for direct beta irradiation. On account of the small penetration depth of betas in tissue, the International Commission of Radiological Protection had prescribed measuring the absorbed beta dose in a corresponding tissue depth of 0.07 and eventually 3 mm. There are a number of detection techniques applied in personnel monitoring:

1. Film dosemeters, which have a lower detection limit of about 500 keV, because photographic films must be worn in a light-tight pack that absorbs lower-energy betas. But even 70-keV betas can be hazardous for human skin.
2. Thermoluminescence dosemeters (TLD), in which mostly thin lithium–fluoride films act as detector material. Their response decreases below about 200 keV, but is not zero.

3. Exoelectron dosemeters, in which the thermally stimulated emission of "exoelectrons" from beryllium-oxide thin films is applied for beta detection. These dosemeters are the most sensitive devices known up to now. Their response is appropriate even at the lowest relevant beta energies, because of their sensitive depth of about a tenth of a millimeter.

Independent of the type of dosemeter, personnel beta monitoring suffers from absorption of beta particles in air, i.e., the dose is recorded at the location where the dosemeter is worn. Therefore, recording of beta doses with personnel dosemeters is only a rough estimate of individual beta dose, and the results should be interpreted very carefully.

4. ALPHA-PARTICLE DETECTORS

Because of their mass, which is about 8000 times greater than that of beta particles, alpha particles have much lower penetration depths in matter than beta particles. The ranges of alpha particles from radioactive sources in solids and liquids are lower than about 0.01 cm, and in air about 5 cm. This means that self-absorption in sources and absorption in counter windows is a more severe problem than with beta sources, as described in the preceding section. Indeed, there is the possibility to detect alpha emitters by the aid of their gamma emission, but in most cases this gamma emission is of very low energy, and especially in mixtures of radioisotopes this method may be unsuitable. On the other hand, the ionization density of alpha particles in matter is more than two orders of magnitude higher than that of betas and gammas. In contrast to beta emission, alpha particles are emitted monoenergetically, i.e., the identification of alpha emitters in mixed sources by spectrometry is generally easier than beta spectrometry. Table 5 gives data of some selected alpha emitters. Aside from basic scientific research, alpha detectors are in use in facilities of the military and civilian nuclear-fuel cycle, including uranium mining, and in isotope-production facilities. An additional field is the environmental natural radon, especially in conventional mining.

Table 5. Data on Selected Alpha-Particle Emitters

Radioisotope	Alpha-particle energy (MeV)	Relative intensity (%)	Half-life
^{212}Po	8.7842	64.1	3×10^{-7} s
^{220}Rn	6.288	99.9	55.6 s
^{222}Rn	5.4895	99.9	3.8 d
^{232}Th	3.830	0.2	
	3.954	23	1.4×10^{10} a
	4.013	77	
^{235}U	4.368	12.3	
	4.374	6.1	7×10^{8} a
	4.400	57	
^{239}Pu	5.105	11.5	
	5.1438	15.1	2.4×10^{4} a
	5.1566	73.3	
^{241}Am	5.388	1.4	
	5.443	12.8	
	5.4486	85.2	432.6 a
	5.512	0.2	
	5.5443	0.34	
^{252}Cf	6.0757	15.5	2.64 a
	6.1181	84.5	

Concerning radiation protection, alpha emitters are the most hazardous isotopes with respect to incorporation. Therefore the allowed air concentrations of alpha emitters at workplaces are as low as, e.g., 6×10^{-1} Bq/m^3 for uranium ($^{235}UO_2$, $^{235}U_3O_8$), or 2×10^{-1} Bq/m^3 for plutonium ($^{239}PuO_2$). The detection of alpha-in-air concentrations below those limits is extremely difficult. The task is rendered more difficult by the larger concentrations of the natural alpha emitter radon and its daughters in room air, which is much less hazardous than the isotopes mentioned above.

4.1 Detectors for Activity Measurements

4.1.1 Liquid and Solid Scintillation Counters Because of the very short alpha-particle ranges, the amount of matter between detector and source should be a minimum. As with beta-particle detection described in the preceding section, scintillation counting is a suitable tool for alpha-activity measurement. But the relative large background counting rates of liquid scintillation counters (see Sec. 3.1.1) restrict the usability of this technique to low-level sources. In mixed-isotope sources, the separation of alpha energies is difficult. Using a multichannel analyzer and pulse-shape discrimination, an energy resolution of about

10% at nearly 100% detection efficiency is achievable with liquid scintillation counters if suitable alpha sources are investigated (Fig. 33). But generally this is not the method of choice in alpha spectrometry.

In solid scintillation detectors, uncoated scintillators have to be used because of the short alpha-particle ranges. Inorganic materials such as $CaF_2(Eu)$, ZnS(Ag), and CsI(Tl) are applied, which are more or less insensible to air humidity. With regard to alpha self-absorption the source should not be thicker than the range of alpha particles, when gross alpha measurements are performed. For precise spectrometry, the source thickness should be less than 1% of the particle range of about 10^{-4} cm to provide maximum energy resolution below 5%.

Gaseous alpha emitters such as the natural radon isotopes and their daughter products can be assayed with both liquid and solid scintillation detectors. A number of detectors are in use, one of which is shown in Fig. 34. The charged radon recoil particles are sampled on the ZnS(Ag)-coated anode by the aid of an electrical field.

Both liquid and solid scintillation counters should be constructed of materials that contain extremely low concentrations of natural alpha emitters, to reduce spurious background counts. Furthermore, initially clean devices may be contaminated by exposure to air, as a result of alpha-emitting daughter products of natural radon isotopes that "plate out" on surfaces.

4.1.2 Alpha-Particle Spectrometry Although scintillation detectors may be of use in this field, high-resolution spectrometry is carried out mainly with silicon surface-barrier detectors (see DETECTORS, SEMICONDUCTOR), and more rarely with pulse-ionization chambers. In Fig. 35, an alpha spectrum from ^{241}Am is presented where the different alpha-particle groups emitted by this isotope are well resolved, thus demonstrating the superior performance of this detector type (see Fig. 33). In this alpha-energy range, energy resolution is about 0.2% (FWHM), being somewhat poorer at lower alpha energies because of electronic noise.

After long-term use, the detector performance degrades as a result of radiation damage in the active detector volume, leading to increased leakage current and loss of resolution. A rough number for the occurrence of serious defects from alpha particles is 10^{11} particles/cm^2.

Pulse-ionization chambers are also used in alpha spectrometry though semiconductor detectors are in advance during the last two decades. With very large-area chambers (up

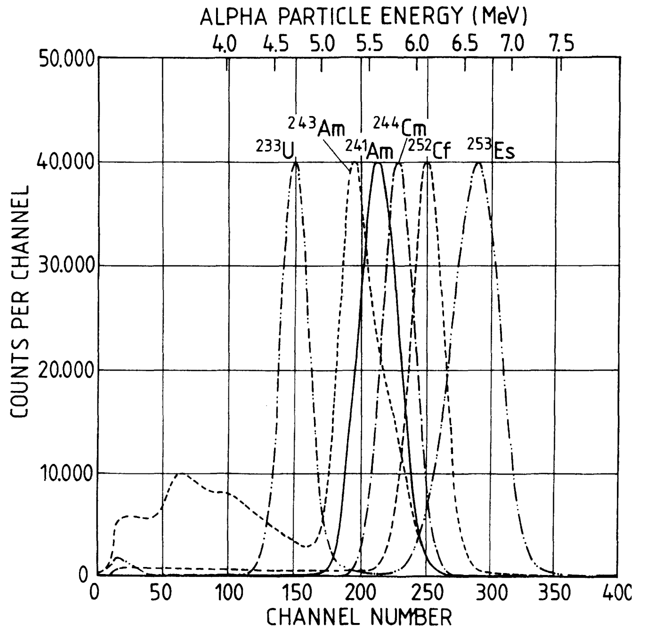

FIG. 33. Alpha spectra of selected actinides recorded with a liquid-scintillation detector (McDowell, 1971). Note: The spectra are not from a mixture of the actinides.

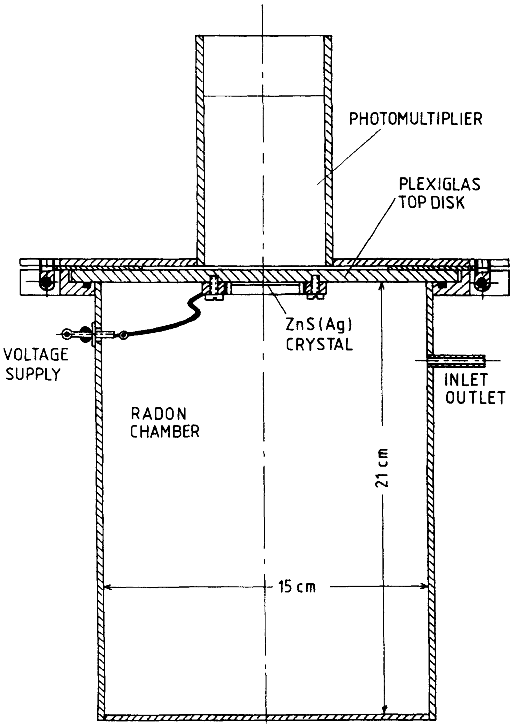

FIG. 34. Cross section of a scintillation counter for detection of radioactive gases (Mann and Seeliger, 1958).

to 1.5 m^2), a resolution of about 1% can be obtained. Further techniques applied on a small scale are mass spectrometry and calorimetry.

4.2 Detectors for Area and Environmental Control

As mentioned above, the detection of alpha-particle concentrations in room and waste air and also in waste water is very difficult because of the low concentration limits that are decreed for many alpha emitters; e.g., much less than one decaying ^{239}Pu atom/s in 1 cm^3 of air has to be detected. With the exception of radon gas, all alpha emitters are present as aerosols or dusts. A variety of sophisticated systems has been developed, the most sensitive based on aerosol and dust-sampling techniques. The control of liquid wastes is mostly carried out after physical or chemical concentration of samples. Most monitoring systems detect alpha and beta activities simultaneously.

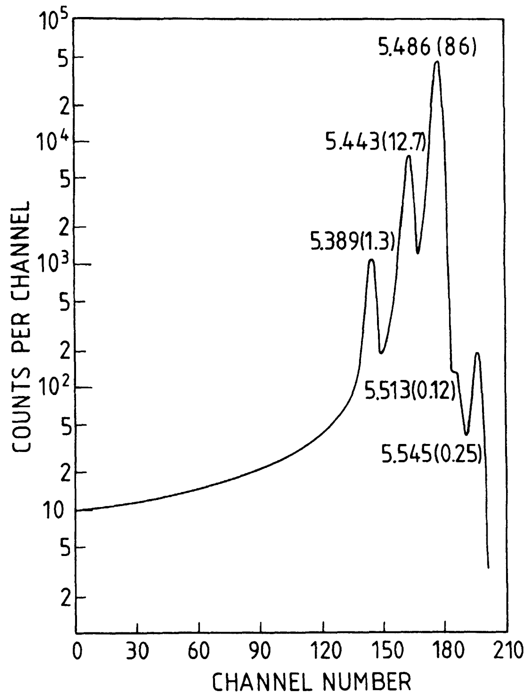

FIG. 35. Alpha spectrum of ^{241}Am recorded with a silicon surface-barrier detector (Chanda and Deal, 1970).

4.2.1 Alpha-Contamination Monitors

Alpha-contamination monitors operate mostly with very thin-window, large-area proportional counters. Windows are made from plastics, metals, and metallized plastics, 0.5 to 1 mg/cm^2 in thickness. Gas fillings are methane, butane, propane, and also noble-gas mixtures. The active counter area of portable devices is usually 100 cm^2. Most of the instruments in use detect both beta (+gamma) and alpha radiation by applying two different high-voltage levels. Because of the higher ionization density of alpha particles in the counter gas, the so-called "alpha plateau" is at lower voltages than the beta + gamma plateau. Typical alpha and beta + gamma plateau voltages are 2.3 and 3.4 kV, respectively, when working with methane, and 1.5 and 2.2 kV, respectively, when working with an argon–methane mixture. The detection efficiency of alpha monitors for, e.g., ^{241}Am is up to about 20%. The figures given vary widely depending on the type of product, i.e., the design of the instrument.
]

4.2.2 Room-Air, Waste-Air, and Waste-Water Monitors Except from some gaseous alpha emitters of comparatively low hazard potential, which can be controlled by flow-through and thin-window counters (see Sec. 3.2.2), efficient room- and waste-air control is not possible without sampling techniques. When actinide aerosols or dusts are present in room air, as is possible in nuclear reprocessing plants (limiting isotope: ^{239}Pu) or in nuclear-fuel–element production facilities (limiting isotope: ^{238}U) sampling times of up to 24 h may be necessary in order to secure that limits of isotope-in-air concentrations are not exceeded. Aerosols and dusts in room air are usually sampled by pumping 1 to 50 m^3/h of air through suitable filters. For detection of sampled activities, open scintillation counters or thin-window gas-multiplication counters are used. In many facilities, counting systems detect both alpha and beta activities simultaneously in two channels. This is realized by using two different detectors, by operating a gas counter at two high-voltage plateaus, or by using a ZnS(Ag)-coated plastic scintillator where the ZnS(Ag) coating is for alpha detection and the plastic scintillator for beta detection. One main problem of detection of low-level actinide activities is the suppression of higher-level alpha activities of natural radon and its daughter products. This can be carried out by

1. filtering the fresh air supplied from outside, so lowering the radon daughter activity concentrations;
2. applying pulse-height discrimination, since all the radon and radon daughter products have higher alpha energies than, e.g., plutonium or uranium;
3. comparing prompt and delayed count rates, and determining the decay of sampled activity. Radon and radon daughter products possess much smaller half-lives than, e.g., the relevant actinides.

Whereas room-air control is mostly performed with detection systems in which single filters are exchanged and controlled after a certain sampling period, waste-air monitors are equipped mostly with running filter belts, whereby the belts are run continuously or stepwise after certain sampling intervals. There are various commercial systems available; an example is shown in Fig. 36. The system has two alpha–beta detectors. The first one measures the prompt activity sampled on the filter, the second one measures the decayed activity after a delay of, e.g., 12 or 24 h. The highest detection sensitivity is obtained with filter belts running stepwise, since the amount of sampled activity is increased with sampling time at each belt sector. When highly resolved alpha spectrometry is necessary, e.g., in nuclear reprocessing plants, silicon surface-barrier detectors are installed.

Concerning waste-water control, especially if low-level actinide concentrations have to be detected, additional physical and chemical concentration techniques have to be applied before measurement. Often, the activity is determined by pure chemical means. This technique is also be used for independent control of activity-covered air filters.

4.2.3 Passive Alpha-Particle Detectors Nuclear track techniques are applied in many fields, e.g., in indoor and environmental radon monitoring, in uranium ore prospecting, and in geophysical research. Alpha particles produce tracks in photographic emulsions, glasses, plastics, and many natural materials. By chemical or electrochemical etching the track diameters can be enlarged and the track density can be determined by optical methods. Another technique is spark counting.

Widespread use is made of track detectors in a large number of national programs in

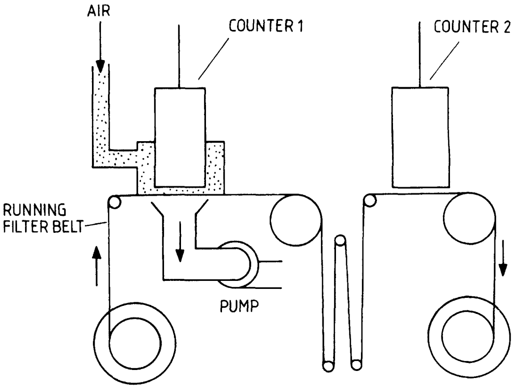

FIG. 36. Schematic drawing of a waste-air monitoring system (Schröck-Vietor, 1962).

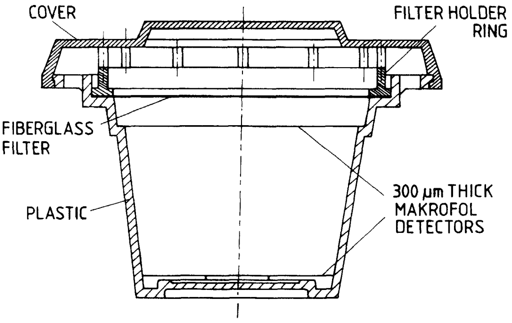

FIG. 37. Nuclear track-detector device for determination of radon gas concentration in air (Urban and Piesch, 1981).

many countries, aiming to detect natural radon and its daughter products in buildings on a large scale. Figure 37 shows a cross section of a track-detector device applied in such studies. The device is a simple diffusion chamber. A fiberglass filter holds back dusts, aerosols, and radon daughter products, so that only radon isotopes diffuse into the chamber. The upper detector is perforated. Only radon and its daughter products that are produced within the chamber and plate out on the foils are detected. The technique of filtering the daughter products present in room air allows a sufficiently accurate correlation of track density and radon concentration. This is not possible by using bare foils. The lower detection limit of the device is some Bq/m^3 of radon after an exposure period of three months.

GLOSSARY

Activity: The activity A of a radioisotope is defined as $A = -dN(t)/dt$ where dN is the number of radioactive transitions in the time interval dt. The unit is 1 Bq (1 Bequerel) = 1 transition/s irrespective of the number of particles or photons emitted per transition.

Coincidence, Anticoincidence Techniques: A coincidence unit determines if two or more detected events are coincident in time. It generates a logic signal only if they occur simultaneously, no signal if not. An anticoincidence unit rejects events that are coincident in time.

Dosimetric Units: The absorbed dose or energy dose is the energy that is absorbed per kilogram of tissue. The unit is 1 Gy (Gray) = 1 J/kg. The dose equivalent is the absorbed dose multiplied with a dimensionless quality factor Q (also called radiation-weighting factor w_R), which is related to the relative biological effectiveness of radiation. Q is 1 for photons and beta particles, 20 for alpha particles, and dependent on energy between 5 and 20 for neutrons. The unit of the dose equivalent is 1 Sv (Sievert) = 1 J/kg. Dose rates are doses per time interval; the units are 1 Sv/h and 1 Gy/h. A formerly used dose unit is the ion dose 1 R (Röntgen), which is defined in terms of the charge produced in air by photons: 1 R = 2.58 C/kg = 0.0087 Gy.

Electronvolt (eV): 1 electronvolt is the energy that a singly charged particle receives if it is accelerated by a voltage of 1 V. The unit eV is used as a measure of energy of, e.g., charged and uncharged particles or quanta in nuclear physics. 1 MeV = 1 Megaelectronvolt, 1 eV = 1.6022×10^{-19} J.

FWHM (Full Width Half Maximum or Minimum): Full width of a peak or a dip in, e.g., a measured energy distribution of particles or quanta determined at half maximum or minimum.

Moderation: The slowing down of fast neutrons by elastic collisions in matter. This process is most effective with light nuclei. Therefore, hydrogen-containing materials such as plastics are widely used as moderators.

Nuclear-Reaction Symbols: For classifying nuclear reactions generally symbols are used, e.g., (n,α), $(n,2n)$, (n,f). This means that a neutron is absorbed in the nucleus, and an alpha particle and two neutrons are emitted from the nucleus, respectively. In the third case, a neutron induces the fission of a nucleus into mostly two new nuclei.

Phoswich Detector: A phosphor sandwich detector where two scintillators with different light-decay times are coupled to one photomultiplier. The pulse-shape difference of the two scintillators can be used to discriminate low-range particles (beta and alpha particles) that are stopped in the first scintillator, and penetrating radiation (neutrons, gammas) that penetrates both scintillators.

Photopeak: The photopeak in gamma- or x-ray energy spectra is due to the photoeffect where the quanta lose their total energy to an electron of the detector material. The photopeak gives the primary energy of the quanta and enables the identification of isotopes also in mixed-radiation fields.

Pulse-Height Discrimination: Radiation of different energies or different kinds generates pulses of different height in a suitable detector. If only one radiation component should be measured, a single-channel analyzer may be applied, which has adjustable lower and upper thresholds thus defining a pulse-height acceptance window. Pulses of lower or larger heights are rejected. For pulse-height spectroscopy over a broad range of energies or radiation components, it is convenient to use a multichannel analyzer, which sorts pulses of different heights into corresponding channels.

Pulse-Shape Discrimination: The rise and decay times of detector pulses from a mixed-radiation field can deliver information on the kind of exciting particles or photons due to different ionization densities or powers in the detector. By determining the detector pulse-decay times electronically, components of the radiation can be discriminated. The technique is commonly applied in fast-neutron detection, since there is mostly a strong gamma component present.

Works Cited

Apfel, R. E., Roy, S. C. (1985), *Radiat. Protec. Dosim.* **10,** 327–330.

Böhm, J. (1984), in: *Proceedings of the International Beta Dosimetry Symposium,* Washington, DC: U.S. Nuclear Regulatory Commission, pp. 73–97.

Brackenbush, L. W., McDonald, J. C., Endres, G. W. R., Quam, W. (1985), *Radiat. Protec. Dosim.* **10,** 307–318.

Bresisti, M. (1963), *Neutron Dosimetry I,* Vienna: International Atomic Energy Agency, p. 27.

Caizergues, R., Poullot, G. (1977), *Report CEA-R-4400,* Gif-sur-Yvette, France: Commissariat à l'Énergie Atomique, Centre d'Études Nucléaires de Saclay.

Chanda, R. N., Deal, R. A. (1970), *Catalogue of Semiconductor Alpha-Particle Spectra,* Idaho Nuclear Corporation Report No. IN-1261, Springfield, VA: National Technical Information Service.

Cooper, J. A., Perkins, R. W. (1971), *Report BNWL-SA-3972,* Richland: WA: Battelle Pacific Northwest Laboratories.

De Pangher, J., Nichols, L. L. (1966), *Report BNWL-260,* Richland, WA: Battelle Northwest Laboratory.

Distenfeld, C. H. (1976), *Report BNL-21293,* Upton, NY: Brookhaven National Laboratory.

Endres, G. W. R., Aldrich, J. M., Brackenbush, L. W., Faust, L. G., Griffith, R. V., Hankins, D. E. (1981), *Final Report of Subtask A: Reactor Containment Measurements, Report PNL-3585, NUREG/CR-1769,* Richland, WA: Battelle Pacific Northwest Laboratory.

Francois, B. (1973), *Int. J. Nucl. Med. Biol.* **1,** 1.

Fricke, J. L., Haberkorn, U., Jung, D., Mohs, E. (1982), in: *Strahlenschutzmeßtechnik,* vol. FS-83-30-T, Munich: Fachverband für Strahlenschutz, p. 243.

Gerber, M. S., Miller, D. W., Schlosser, P. A., Steidley, J. W., Deutchman, A. H. (1977), *IEEE Trans. Nucl. Sci.,* **NS-24** (1), 182.

Gleason, G. I. (1951), *Nucleonics* **8,** (5), 12.

Griffith, R. V., Palfalvi, J., Madhvanath, U. (1990), *Technical Reports Series No. 318,* Vienna: International Atomic Energy Agency.

Hankins, D. E. (1977), *Proceedings of the 4th International Congress of the International Radiation Protection Association,* Vol. 2, Paris: IRPA, pp. 553–556.

Harrison, K. G. (1979), *Nucl. Instrum. Methods* **166,** 197–201.

Hewitt, J. E., Hughes, L., McCaslin, J. B., Stephens, L. D., Rindi, A., Smith, A. R., Thomas, R. H., Griffith, R. V., Welles, C. G., Baum J. W., Kuehner, A. V. (1976), *Health Phys.* **34,** 375–384.

Holeman, G. R., Price, K. W., Friedman, L. F., North, R. (1977), in: *Proceedings of the 4th International Congress of the International Radiation Protection Association,* Paris: IRPA, pp. 827–830.

Hutchinson, J. M. R., Mann, W. B., Perkins, R. W. (1973), *Nucl. Instrum. Methods* **112,** 305.

Ing, H., Birnboim, H. C. (1984), *Nucl. Tracks Radiat. Meas.* **8,** 285–288.

Ing, H., Cross, W. G., Bunge, P. J. (1985), *Radiat. Protec. Dosim.* **10,** 137–146.

Johnson, N. R., Eichler, E., O'Kelley, G. D: (1963), in: H. B. Jonassen (Ed.), *Nuclear Chemistry,* Vol. 2 of Techniques of Inorganic Chemistry, New York: Interscience.

Lab. Berthold (1984), Lab. Prof. Berthold, Wildbad, Germany, data sheet.

Lamphere, R. W. (1960), in: J. B. Marion, J. L. Fowler (Eds.), *Fast Neutron Physics I,* New York: Interscience, p. 449.

Lang, G. (1979), in: H. Reich (Ed.), *Medizinische Physik,* Heidelberg: Hüthig, pp. 39–347.

Leake, J. W. (1966), *Nucl. Instrum. Methods* **45,** 151.

Mann, W. B., Seliger, H. H. (1958), *Circular 594,* National Bureau of Standards, Washington, DC: U.S. Government Printing Office.

McDowell, W. J. (1971), in: D. L. Horrocks, C. T. Peng (Eds.), *Organic Scintillators and Liquid Scintillation Counting,* New York: Academic, p. 937.

Muggleton, A. H. F. (1972), *Nucl. Instrum. Methods* **101,** 113.

Nachtigall, D., Burger, G. (1972), in: F. H. Attix (Ed.), *Topics in Radiation Dosimetry,* Supplement 1 to *Radiation Dosimetry,* Chap. 7, New York: Academic.

Neuert, H. (1966), *Kernphysikalische Meßverfahren,* Karlsruhe (D): Braun, Germany.

Robinson, B. M. W. (1960), in: *Metrology of Radionuclides,* IAEA/STI/PUB/6, Vienna: International Atomic Energy Agency, p. 41.

Sayres, A., Coppola, M. (1964), *Rev. Sci. Instrum.* **35,** 431.

Scharmann, A., Schalch, D., Pitt, E. (1986), *Report No. BMFT KWA 3309 A7,* Düsseldorf: Bundesministerium für Forschung und Technik.

Schröck-Vietor, W. (1962), *Atomwirtschaft* **7,** 474.

Tanaka, E. (1961), *Nucl. Instrum. Methods* **13,** 43.

Urban, M., Piesch, E. (1981), *Radiat. Protec. Dosim.* **1,** 97–109.

Wallace, R. *et al.* (1960), U.S. Atomic Energy Commission Report UCRL-9214, Springfield, VA: National Technical Information Service.

Wirtz, K., Beckurts, K. H. (1958), in: *Elementare Neutronenphysik,* Berlin: Springer.

Wood, R. E., Venugopala Rao, P., Puckett, O. H., Palms, J. M. (1971), *Nucl. Instrum. Methods* **94,** 245.

Further Reading

Attix, F. H. (1986), *Introduction to Radiological Physics and Radiation Dosimetry,* New York: Wiley.

Currie, L. A. (Ed.) (1988), *Detection in Analytical Chemistry,* Washington, DC: American Chemical Society.

Debertin K., Helmer, R. G. (1988), *Gamma- and X-Ray Spectrometry with Semiconductor Detectors,* Amsterdam: North-Holland.

Durrani S. A., Bull P. K. (1987), *Solid State Nuclear Track Detection: Principles, Methods and Applications,* Oxford, UK: Pergamon.

Dyson, N. (1993), *Radiation Physics with Applications in Medicine and Biology,* 2nd ed., Englewood Cliffs, NJ: Prentice-Hall.

García-León, M., Madurga, G. (Eds.) (1991), *Low-Level Measurements of Man-Made Radionuclides in the Environment,* Singapore: World Scientific.

Jahr, R., Alberts, W. G., Menzel, H., Schraube, H. (Eds.) (1992), *Neutron Dosimetry, Radiation Protection Dosimetry,* Vol. 44, Nos. 1–4, Ashford, Kent, UK: Nuclear Technology Publishing.

Knoll, G. F. (1989), *Radiation Detection and Measurement,* 2nd ed., New York: Wiley.

Kose, V., Wagner, S. (Eds.) (1994/95), *Kohlrausch—Praktische Physik,* Vol. 2, Stuttgart: Teubner.

Leo, W. R. (1994), *Techniques for Nuclear and Particle Physics Experiments,* 2nd rev. ed., Berlin, Heidelberg: Springer.

NCRP Report No. 58 (1985), *A Handbook of Radioactivity Measurements Procedures,* 2nd ed., Bethesda, MD: National Council on Radiation Protection and Measurements.

Oberhofer, M., Scharmann, A. (Eds.) (1979), *Applied Thermoluminescence Dosimetry,* Bristol: Hilger.

Siegbahn, K. (Ed.) (1986), *Nuclear Instruments & Methods in Physics Research,* Vol. A252, (1987), Vol. A253, Amsterdam: North-Holland.

RADIATION EFFECTS IN SOLIDS

See INTERACTION OF SOLIDS WITH PARTICLES AND RADIATION

RADIATION, ELECTROMAGNETIC

See ELECTROMAGNETIC RADIATION

RADIO TELESCOPES

JOHN D. KRAUS, *The Ohio State University, Columbus, Ohio, U.S.A.*

Introduction 527

1. **General Principles and Brief History** 527

1.1 Radio Telescopes and the Radio Sky 527

1.2 History and Discoveries 527

2. **Radio-Telescope Antennas** .. 529

2.1 Antenna Patterns 529

2.2 Antenna Parameters 529

2.3 Antenna Designs and Examples 532

2.4 Aperture Distribution and Pattern 535

2.5 Brightness and Flux Density .. 535

2.6 Interferometers and Aperture Synthesis 536

2.7 Radar Telescopes 539

3. **Radio-Telescope Receivers** .. 539

3.1 Total-Power and Dicke Types 539

3.2 Low-Noise Amplifiers 541

3.3 Multichannel Receivers 541

3.4 Antenna Temperature and Sensitivity (Remote Sensing) . 542

3.5 Remote Sensing of the Earth 543

Glossary 543

Works Cited 544

Further Reading 545

INTRODUCTION

A radio telescope is a remote-sensing device that measures the equivalent temperature of a distant object by means of its radio-frequency radiation. Basically, a radio telescope has two parts: (1) an antenna for collecting the object's radiation and (2) a receiver for detecting, processing, and displaying the information, as illustrated in Fig. 1. A radio telescope is analogous to an optical telescope, except that it functions at much longer wavelengths from millimeters to hundreds of meters. The antenna is analogous to the lens or reflector of an optical telescope while the receiver-processor corresponds to a photographic plate or other optical-sensing device.

Radio telescopes are used for exploring the heavens and mapping the sky. The appearance of the radio sky is very different from the way it appears to the eye.

1. GENERAL PRINCIPLES AND BRIEF HISTORY

1.1 Radio Telescopes and the Radio Sky

Observations of the sky from the ground by either optical or radio telescopes are limited to two principal windows of the Earth's atmosphere and ionosphere:

1. in the optical region (400 to 700 nm) and
2. in the radio region (nominally 1 mm to 100 m) restricted at short wavelengths (1 mm) by molecular absorption in the atmosphere and at long wavelengths by ionospheric reflection.

Radio telescopes carried into space and directed at the earth are used to map the Earth's surface temperatures for studies of the Earth's resources and environment. In mapping the sky or the Earth, the radio telescope is a *passive* remote-sensing device. If the telescope is equipped with a radio transmitter, it can also function as an *active* remote-sensing device or *radar* telescope, providing more information. Since the echo power of a radar telescope decreases as the fourth power of the distance, astronomical radar telescopes are restricted to studies of solar-system objects.

1.2 History and Discoveries

In 1933, Karl J. Jansky, a radio engineer at the Bell Telephone Laboratories, Holmdel, New Jersey, built a rotating radio antenna to study the direction of arrival of thunderstorm static at 14.6-m wavelength. In the ab-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

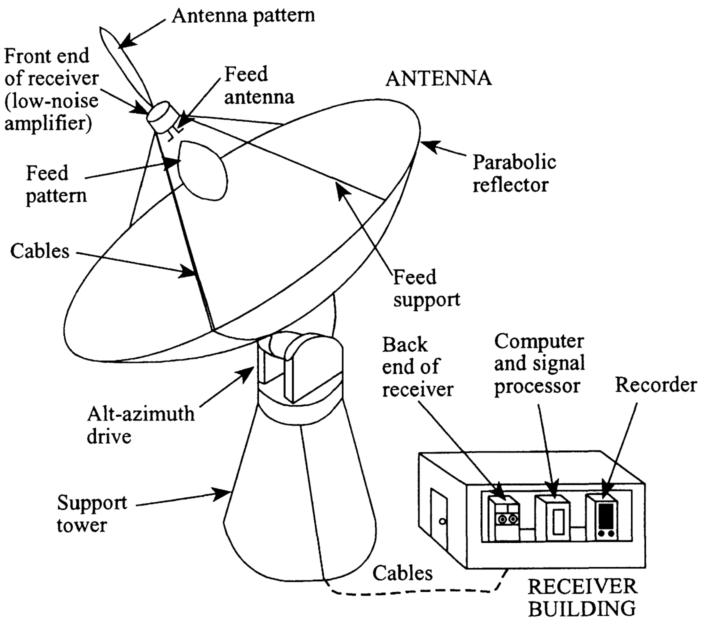

FIG. 1. Radio telescope with low-noise amplifier (LNA) at prime focus and back end of receiver, signal processor, and recorder in receiver building.

sence of any static, he noted a weak but persistent noise signal that he traced to radiation from our galaxy or Milky Way (Jansky, 1933). Prior to Jansky, all of our astronomical knowledge came from optical observations. Jansky broke the "optical barrier" and forever changed the course of astronomy.

Jansky's serendipitous discovery was followed up by another radio engineer, Grote Reber, who made the first maps of the radio sky at a wavelength of 1.9 m using a 9.5-m-diam dish antenna he constructed in the backyard of his home in Wheaton, Illinois (Reber, 1940, 1983). Reber established that the galactic radiation was nonthermal or non-Planckian in nature (stronger at longer wavelengths). It was subsequently determined (20 years later) to be synchrotron radiation from fast electrons spiraling in interstellar magnetic fields.

For almost a decade, Reber was the world's only active radio astronomer, but after WW II, with the advent of new radio devices, radio astronomy developed rapidly.

Radio telescopes have found quasars (compact radio galaxies) so distant that they are receding at over 90% of the velocity of light; radio objects traveling at apparently superluminal velocities (faster than light); neutral hydrogen clouds throughout the galaxy, giving the first picture of our galaxy's spiral structure; giant interstellar molecular clouds of hundreds of different molecules ranging from ammonia and ethyl alcohol to sulfur dioxide and water; and pulsars, rapidly rotating neutron stars, some of which spin 1000 times per second and provide long-term timekeeping comparable to the most accurate atomic clocks. Such long-term measurements of the rate of decay of the orbital period of a binary pulsar (PSR 1913 + 16) were found to agree closely with that expected from damping by gravitational radiation. Even though indirect, this constitutes observational evidence of the existence of gravitational waves, as postulated by Einstein (Taylor, 1981).

Another radio discovery is the 3-K cosmic radiation that sets a minimum background temperature for the sky. From measurements over a wide range of wavelengths, this relic radiation from the primordial fireball or "Big Bang" has been established as having the Planckian (or black body) curve of a 2.7-K object (Penzias and Wilson, 1965). This implies that all but a few hundredths of one percent of the background radiation was released within a year of the Big Bang.

2. RADIO-TELESCOPE ANTENNAS

2.1 Antenna Patterns

The antenna of a radio telescope acts as a collector of radio waves, converting radio photons to radio-frequency currents. The response of the antenna as a function of angle is given by the antenna pattern. By reciprocity, this is the same pattern measured when the antenna is transmitting. The pattern is independent of distance if measured sufficiently far from the antenna. The pattern may be expressed in field intensity E (V/m) or power flux P (W/m^2), where $P = E^2/Z$ and Z is the intrinsic impedance of space = 377 ohms/square.

A symmetrical pencil beam pattern may be described in terms of a single power flux pattern (Fig. 2) where $P(\theta)$ represents the power flux as a function of the angle θ. The half-power beam width (HPBW) and the beam width between first nulls (BWFN) are indicated. Dividing $P(\theta)$ by its maximum value, $P(\theta)_{\max}$, we obtain a normalized power pattern, which is dimensionless. In the general case, a pattern is described as a function of the angles θ and ϕ in two orthogonal planes as in Fig. 3. The normalized power pattern for this case is then

$$P_n(\theta,\phi) = P(\theta,\phi)/P(\theta,\phi)_{\max}. \tag{1}$$

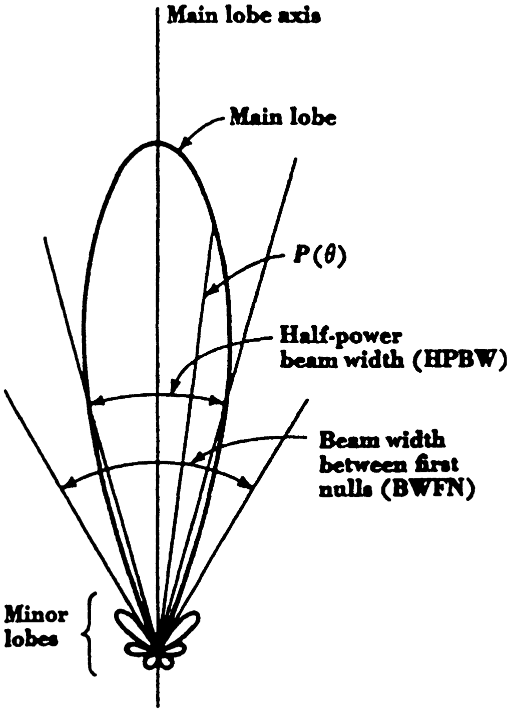
Main lobe axis
Main lobe
P(θ)
Half-power beam width (HPBW)
Beam width between first nulls (BWFN)
Minor lobes

FIG. 2. Antenna pattern in polar coordinates and linear power scale.

2.2 Antenna Parameters

Three important characteristics of a radio telescope antenna (or of any antenna) are its *directivity* D (dimensionless), *beam area* Ω_A (steradians), and *effective aperture* A_e (m^2). By reciprocity they are the same for both reception and transmission. They are related as follows:

$$D = 4\pi/\Omega_A = 4\pi A_e/\lambda^2, \tag{2}$$

where λ = wavelength (m). For example, if the beam area is 1 square degree ($=3 \times 10^{-4}$ steradians), the directivity is 41253 (or $10 \log_{10} 41253 = 46$ dB), and at a wavelength of 1 m, the effective aperture $A_e = 41253/4\pi = 3283$ m^2. Note that the solid angle in a sphere is 4π steradians or 41253 square degrees.

The beam area is given by the integral of the normalized power pattern (maximum value = unity) over a sphere (4π steradians). Thus,

$$\Omega_A = \iint_{4\pi} P(\theta,\phi) d\Omega \tag{3}$$

(dimensionless), where $P(\theta,\phi)$ = power pattern as a function of the coordinates (θ,ϕ), and $d\Omega = \sin\theta d\theta d\phi$ = element of solid angle (steradians). For a symmetrical pencil beam (Fig. 2), Eq. (3) reduces to

$$\Omega_A = 2\pi \int P(\theta) \sin\theta d\theta. \tag{4}$$

For the pencil beam, the beam area is approximately equal to the half-power beam width (HPBW) squared or $\Omega_A \cong \mathrm{HPBW}^2$. Thus, a pencil beam with HPBW = 1° has from Eq. (2) a directivity $D = 41253$. On the other hand, if the antenna is omnidirectional (same power in all directions), $P_n(\theta,\phi) = 1$, $\Omega_A = 4\pi$, and $D = 1$. This is the smallest directivity an antenna can have.

A directivity $D = 41253$ for the 1° pencil

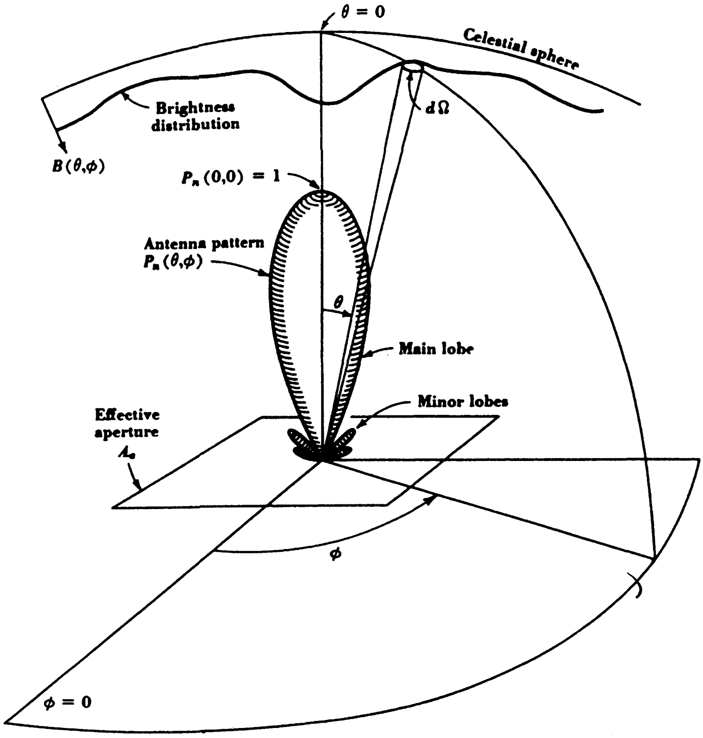

FIG. 3. Relation of antenna pattern to celestial sphere with associated coordinates.

beam means that a 1-W transmitter connected to the antenna produces a power flux in the direction of the beam equal to that from a 41253-W transmitter connected to an isotropic or omnidirectional antenna.

Directivity also has significance as a measure of the number of sources in the sky that a radio telescope antenna can resolve. The resolution of a pencil beam antenna is given *approximately* by its HPBW. A more accurate measure is the *Rayleigh resolution* (= BWFN/2) which is 12% more than the HPBW. Thus, the antenna can resolve *approximately* a number of sources $N = D$ over 4π steradians of sky where $D = 4\pi/\Omega_A = 4\pi/\text{HPBW}^2$. If, for example HPBW = 1°, the antenna can resolve of the order of 40 000 sources, assuming a uniform distribution over the sky.

The ratio k_1 of the power radiated by an antenna to its power input is unity if the antenna is lossless but less if losses are present. The gain G of an antenna is equal to its directivity multiplied by k_1. Thus $G = k_1 D$. Since k_1 is nearly unity for most antennas, the antenna gain and directivity are used almost interchangeably.

In addition to the Ohmic power loss factor k_1, other factors affecting the efficiency of large reflector antennas are the aperture utilization factor k_2, the surface accuracy factor k_3, and other factors we may combine as k_4, which include the feed efficiency, the effects of aperture blocking by the feed and feed supports, feed displacement normal to the axis (squint), feed displacement parallel to the axis (astigmatism), surface leakage, etc.

For a uniform field distribution across the aperture the aperture utilization factor k_2 is unity. However, to reduce minor lobes, the field is usually tapered to a low value at the edges. For in-phase fields $k_2 = E^2_{av}/(E^2)_{av}$ where E_{av} = average field E over aperture, $(E^2)_{av}$ = average of E^2 over aperture, and E = field at any point in the aperture. Thus, for a uniform field distribution $k_2 = 1$ while for a cosine distribution ($E = 0$ at edges), $k_2 = 0.81$.

Surface error is the deviation of the actual surface from the paraboloid of best fit. The surface accuracy or error factor k_3 may be close to unity at long wavelengths but decreases for shorter wavelengths at which the surface deviation becomes a significant fraction of the wavelength. Thus, for an rms surface deviation or error $\delta = \lambda/16$, the factor $k_3 = 0.5$ (or -3 dB) as given for small δ by the relation* (Christiansen and Högbom, 1985),

$$k_3 = \cos^2(720^\circ\ \delta/\lambda). \qquad (5)$$

The effect of surface accuracy on the gain of two large radio telescopes is shown in Fig. 4. The $k_3 = 0.5$ (or -3 dB) point occurs at a wavelength of about 8 mm on the 100-m Bonn, Germany, telescope of the Max Planck Institute (Wohlleben and Sieber, 1968) (Fig. 5), and at a wavelength of about 2.5 mm on the Nobeyama Radio Observatory 45-m telescope on Honshu Island, Japan (Fig. 6) (Akabane *et al.*, 1983).

The other factors may yield a k_4 value of 0.8 to 0.9. Combining all factors the gain G is given by

$$G = 4\pi k_1 A_e/\lambda^2, \qquad (6)$$

where k_1 = Ohmic or power loss factor, A_e = effective aperture = $k_2k_3k_4A_p = \epsilon_{ap}A_p$, m^2, λ = wavelength, m, A_p = physical aperture, m^2, ϵ_{ap} = aperture efficiency = A_e/A_p = $k_2k_3k_4$, k_2 = field distribution or utilization factor, k_3 = surface accuracy factor, and k_4 = product of other factors. All factors are dimensionless with values between 0 and 1.

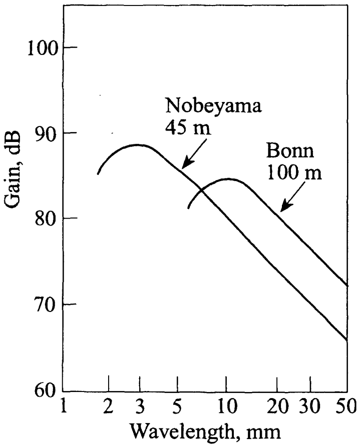

FIG. 4. Gain peaking and rolloff with decreasing wavelength for Bonn 100-m and Nobeyama 45-m radio telescopes (as given by the factor k_3).

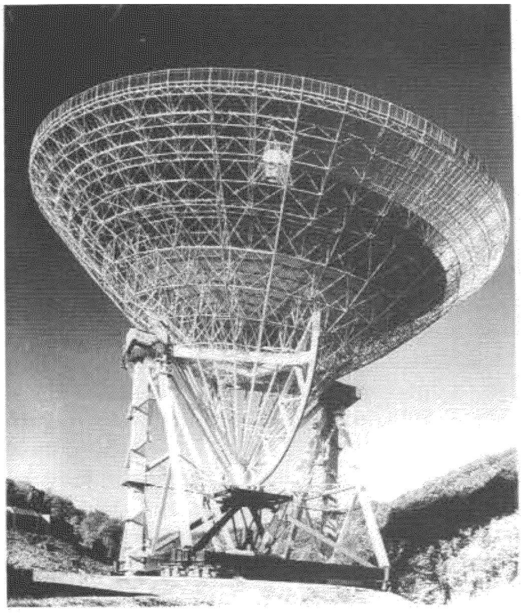

FIG. 5. The 100-m-diam fully steerable radio telescope in a mountain valley near Bonn, Germany. It stands nearly 40 stories tall. Note the massive backup structure for providing rigidity to the dish. Moving parts weigh 3200 tonnes. (Courtesy Dr. R. Wielebinski, Max-Planck-Institut für Radioastronomie.)

Useful observations with a radio telescope may be made even if the aperture efficiency is as small as 0.2. The 25-m dishes of the

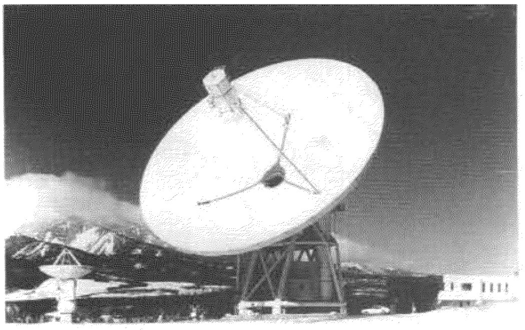

FIG. 6. Fully steerable 45-m parabolic-dish antenna for millimeter-wavelength observations at the Nobeyama Radio Observatory in the Yatsugatake Mountains, Japan. (N.R.O. photo.)

---

*Equation (5) is appropriate only for small δ ($<\lambda/12$). A more universal relation sets $k_3 = \exp-(4\pi\delta/\lambda)^2$ (Ruze, 1966).

VLA (Very Large Array) (Fig. 7) operate with aperture efficiencies of 52% at 1.5 GHz, 65% at 5 GHz, and 48% at 20 GHz (Napier *et al.*, 1983).

2.3 Antenna Designs and Examples

The main or primary reflector of a large radio telescope is constructed of many individual panels. If the panels have an rms surface accuracy of 50 μm and are set into a backup structure with an rms accuracy of 50 μm, a total error of 71 μm accrues ($71 = \sqrt{50^2 + 50^2}$) resulting in a value of $k_3 = 0.81$ at a wavelength of 2 mm. To set the panels accurately, a holographic technique may be used in which the amplitude *and* phase of a satellite beacon signal are measured over an angular extent much larger than the main beam of the antenna. The amplitude and phase of the antenna aperture field distribution are then obtained by a two-dimensional Fourier transform yielding a phase map of the reflector showing where the panels need to be adjusted to reduce the phase variation (Goldsmith and Erickson, 1986).

Radio telescopes must maintain an accurate reflector surface in spite of wind loads, temperature gradients, and gravitational deformation. For a 45-m-diam reflector, such as the one at Nobeyama, an overall rms surface deviation of 150 μm and a pointing accuracy of 1 arcsec are required. The Nobeyama telescope is on a mountain site to reduce the atmospheric attenuation at millimeter wavelengths.

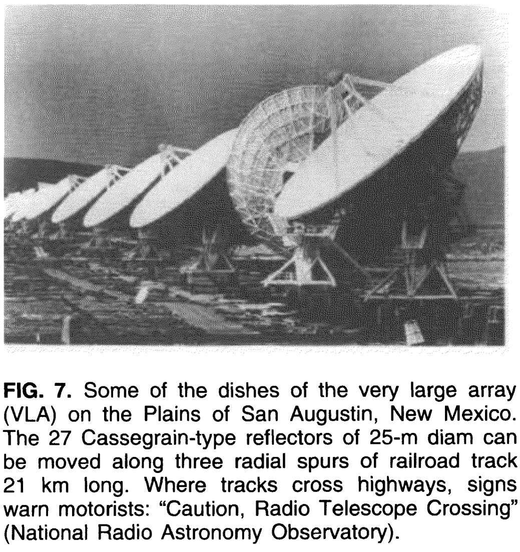

FIG. 7. Some of the dishes of the very large array (VLA) on the Plains of San Augustin, New Mexico. The 27 Cassegrain-type reflectors of 25-m diam can be moved along three radial spurs of railroad track 21 km long. Where tracks cross highways, signs warn motorists: "Caution, Radio Telescope Crossing" (National Radio Astronomy Observatory).

All ground-based steerable parabolic reflectors deform under gravity as the reflector is tilted so that for a given wavelength there is a maximum diameter limit which cannot be exceeded by adding more metal to the backup structure. However, this limit can be exceeded by a *homologous* design in which one paraboloid deforms into another as the reflector is tilted (Von Hoerner, 1977). The next limit is imposed by thermal deformation.

In the Nobeyama 45-m telescope, thermal insulation and ventilation are used to reduce thermal deformation. In the IRAM (Institut de Radio Astronomie Millimètrique) 30-m telescope at Pico Veleta, Spain (Fig. 8), insulation, ventilation, and temperature control are employed. The IRAM/SEST 15-m telescope [Swedish ESO (European Southern

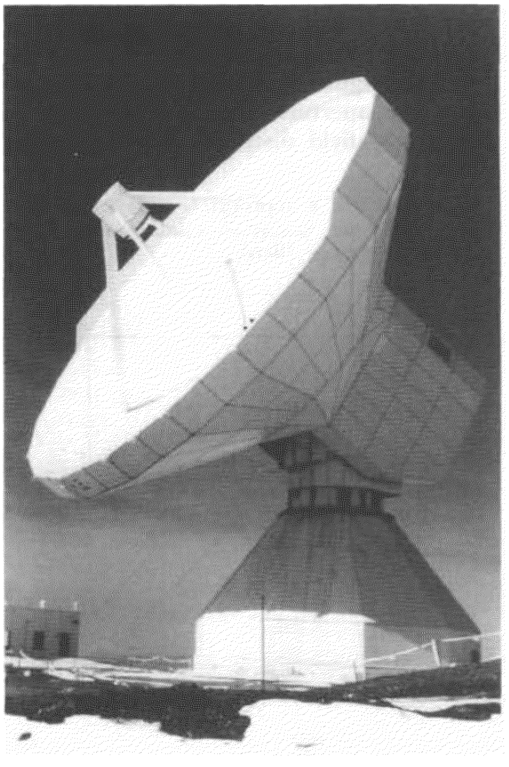

FIG. 8. 30-m telescope for millimeter wavelengths of the Institut de Radioastronomie Millimètrique (IRAM) at Pico Veleta, Spain. Note cladding to reduce thermal deformation (IRAM photo).

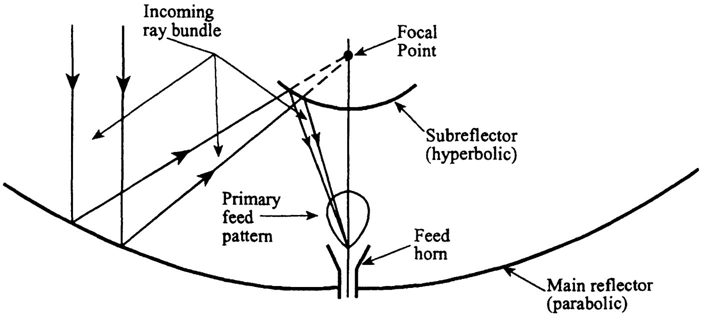

FIG. 9. Geometry for Cassegrain or shaped subreflector telescope designs.

Observatory) Submillimeter Telescope] at La Serena, Chile, has panels, feed legs, and much of the backstructure made of low-expansion carbon fiber material. Insulation, ventilation, and thermal control are also used. Panel accuracy is about 10 μm with equivalent panel positioning accuracy achieved by holographic measurements (Greve *et al.*, 1992).

Although some radio-telescope antennas have the feed antenna and receiver front end at the prime focus point as in Fig. 1, the Cassegrain design is widely used, especially in millimeter-wavelength telescopes. As shown in Fig. 9, a subreflector surrounds the prime focus of the main reflector. In the classical design, the subreflector is hyperbolic in shape. Higher aperture efficiency may be realized by shaping or modifying the hyperbolic contour of the subreflector to correct the pattern of the feed antenna and produce a more uniform field distribution across the parabolic reflector aperture.

An advantage of the Cassegrain design is its compactness, with the feed antenna (usually a horn) and associated front-end receiver components located near the vertex of the parabola (instead of far above the main reflector at the prime focus). A disadvantage is that the subreflector may increase aperture blocking and create side lobes. A constraint is that to minimize blocking the subreflector should be small compared to the main reflector but large compared to the wavelength.

The Bonn 100-m telescope can operate in either Cassegrain or prime-focus modes. The 15- and 30-m IRAM telescopes are Cassegrain types. The Nobeyama 45-m telescope has a Gregorian feed which is similar to the Cassegrain except that the subreflector is concave instead of convex.

The Bonn 100-m telescope is the world's largest fully steerable instrument. Until it was constructed in 1972, the 76-m telescope at the University of Manchester, England, was the largest of this type (Lovell, 1968).

The stationary 305-m-diam spherical reflector of the National Astronomy and Ionospheric Center (Cornell University) is suspended on cables in a mountain valley at Arecibo, Puerto Rico (Fig. 10) (Findlay, 1964). By moving the feed antenna, sup-

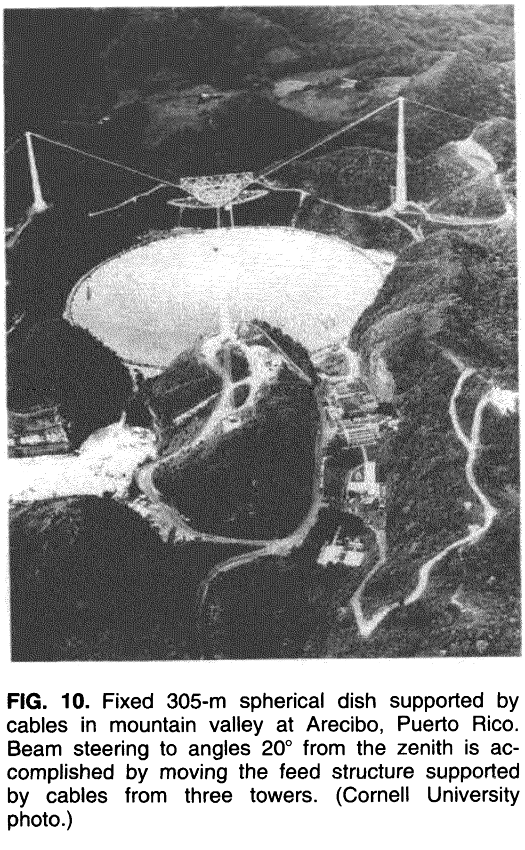

FIG. 10. Fixed 305-m spherical dish supported by cables in mountain valley at Arecibo, Puerto Rico. Beam steering to angles 20° from the zenith is accomplished by moving the feed structure supported by cables from three towers. (Cornell University photo.)

ported above the spherical reflector by cables from three towers, the antenna beam can be steered 20° in any direction from the zenith. Operation is possible to centimeter wavelengths.

A Mills-cross antenna of the University of Sydney, Australia (Mills *et al.*, 1963) is a meridian transit instrument with cylindrical parabolic reflectors 12 m wide by 1.6 km long in the form of an east–west, north–south cross. The beam is steered in declination by a combination of mechanical tilting of the east–west element on its long axis and an electrical phase shifting of the feeds of the mechanically fixed north–south element. At 408 MHz the beam width is about 3 arcmin.

The Ohio State University Radio Telescope has a 110 × 21-m^2 standing curved reflector and a 104 × 31-m^2 tiltable flat reflector with the two joined by a 15 000 m^2 aluminum-covered ground plane (Fig. 11) (Kraus, 1986, pp. 6–86, 6–90). The design provides a maximum aperture per unit cost. Also it has less than $\frac{1}{2}$% aperture blocking and reduced susceptibility to terrestrial interference because of the low profile, with feed horns at ground level and shielding by the large reflectors. It also has a spacious underground laboratory directly below the prime focus for the electronics equipment.

A similar but larger antenna of this general design is situated at Nancay, France (Fig. 12) (Blum *et al.*, 1963; Findlay, 1964). Its curved reflector measures 300 × 35 m^2.

The RATAN-600 Radio Astronomy Telescope at Zelenchukskaya, Russia, has both tiltable flat and curved reflectors (Fig. 13). The curved reflector consists of 895 automatically adjustable aluminum panels forming a ring 600 m in diameter. Operation is possible to millimeter wavelengths (Golubin, 1982; Parijskij, 1993).

A 100-m fully steerable telescope with offset feed (to avoid aperture blocking) is under construction at the National Radio Astronomy Observatory, Green Bank, West Virginia (Fig. 14). It is replacing a 91-m meridian transit telescope that collapsed there on November 15, 1988. The new telescope, being built at a cost of $75 million, is scheduled for completion in 1995. It is designed to operate efficiently at wavelengths as short as 1 cm.

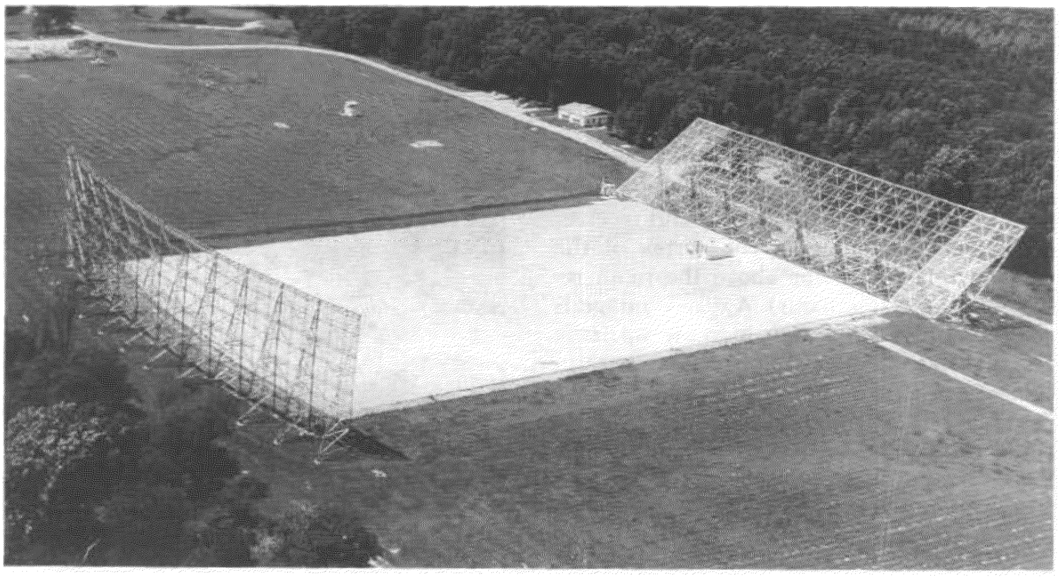

FIG. 11. 110-m standing curved reflector with tiltable flat reflector at the Ohio State University Radio Observatory. The reflectors are connected by a conducting ground plane larger than three football fields. The telescope was used to make maps of the sky at 21-cm wavelength and compile a catalog of 20 000 radio sources including the first ones found to have red shifts greater than 3.

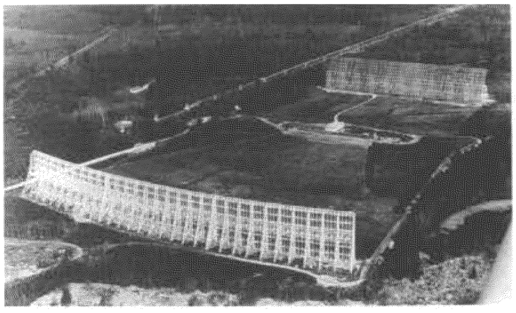

FIG. 12. Radio telescope at Nancay, France, with 300-m standing curved reflector and tiltable flat reflector. The curved reflector is spherical, which facilitates tracking sources for 2 h or more by moving the feed system. (Observatoire de Paris photo.)

2.4 Aperture Distribution and Pattern

The field distribution $E(x_\lambda)$ across the aperture of an antenna and its far-field pattern $E(\theta)$ are reciprocal Fourier transforms where $x_\lambda = x/\lambda$. The field distributions for three one-dimensional apertures of width $L_\lambda = L/\lambda$ and of uniform, Gaussian, and edge distributions are shown in Fig. 15. The latter distribution may be regarded as two spaced narrow apertures in the same configuration as in an interferometer (Sec. 2.6). The field pattern for the uniform aperture distribution has a main lobe with side lobes. Its BWFN $= 2\lambda/L$ rad $= 114.6\ \lambda/L$ deg and the HPBW $= 0.44$ BWFN. The field pattern of the Gaussian aperture distribution is also Gaussian. It has no side lobes but a larger HPBW. The field pattern of the edge distribution has many lobes of equal size but with smaller beam widths. As the separation L increases, the beam widths of the lobes decrease. This distribution is discussed further in connection with interferometers in Sec. 2.6.

2.5 Brightness and Flux Density

When the *brightness* of the sky (W m^{-2} Hz^{-1} rad^{-2}) as measured at the Earth is integrated over the area of a celestial source (Fig. 3), we obtain the *flux density* S of the source. Thus,

$$S = \iint_{\text{Source}} B(\theta,\phi)d\Omega \quad (\text{W m}^{-2}\,\text{Hz}^{-1}), \qquad (7)$$

where $B(\theta,\phi)$ = brightness of the sky as a function of position and $d\Omega$ = element of solid angle (rad^2). The unit of flux density is the jansky (Jy) = 10^{-26} W m^{-2} Hz^{-1}.

When the source is observed with a radio-telescope antenna with normalized power pattern $P(\theta,\phi)$, a flux density S_0 is obtained as given by

$$S_0 = \iint_{\text{Source}} B(\theta,\phi)P_n(\theta,\phi)d\Omega \quad (\text{Jy}). \qquad (8)$$

If the source extent is small so that $P(\theta,\phi) \cong 1$ over the source, the observed flux density S_0 will be nearly equal to the true value.

The power received from the source by an antenna of effective aperture A_e (m^2) is then

$$W = S_0 A_e \Delta f \quad (\text{W}), \qquad (9)$$

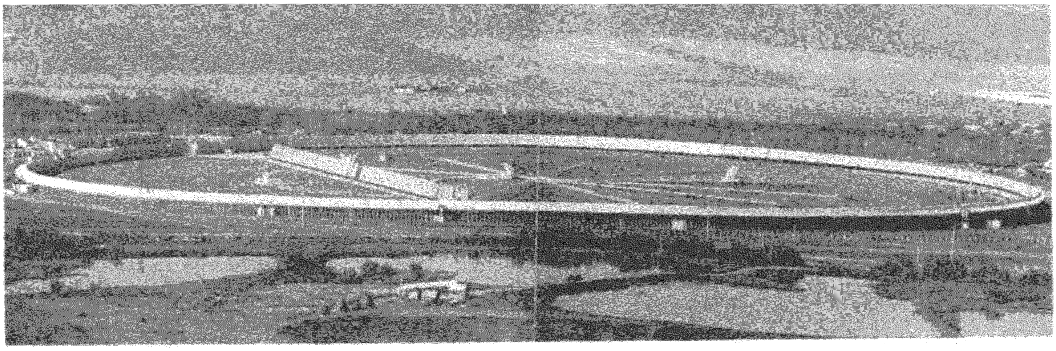

FIG. 13. The 600-m-diam RATAN-600 radio telescope near Zelenchukskaya, Russia, with tiltable flat and curved reflectors. Although the ring reflector is large enough to hold four Arecibo dishes, the collecting area is less. (Photo courtesy, Dr. Y. Parijskij.)

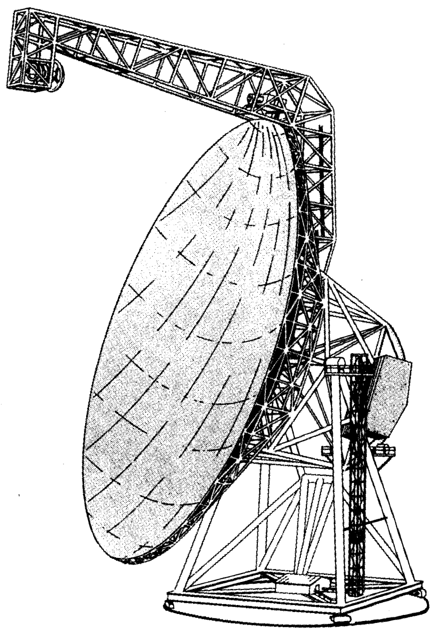

FIG. 14. Sketch of 100-m (GBT) radio telescope at the National Radio Astronomy Observatory, Green Bank, West Virginia, to be completed in 1995. It features an offset feed to provide a clear, unobstructed aperture. (N.R.A.O.-A.U.I.)

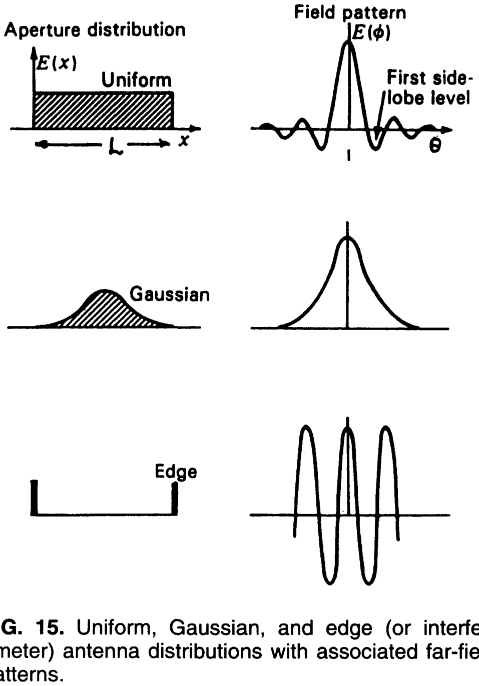

FIG. 15. Uniform, Gaussian, and edge (or interferometer) antenna distributions with associated far-field patterns.

where Δf = bandwidth (Hz). This assumes that the antenna polarization matches that of the source radiation. In general they are not matched because most celestial radio sources are unpolarized whereas an antenna is polarized (linearly, circularly, or elliptically). For this case the antenna receives one-half of the available wave power. Thus, W is one-half that of Eq. (9) for the general case.

2.6 Interferometers and Aperture Synthesis

The edge distribution of Fig. 15 in which two apertures are separated by many wavelengths constitutes a simple interferometer. Thus, if two antennas, each with a uniform distribution of 100 wavelengths in width, replace the two narrow edges and the spacing between the antennas is 1000 wavelengths, the lobes of this interferometer have a BWFN = 115°λ/L = 115°/1000 = 0.12° within the envelope of a larger lobe with BWFN = 115°/100 = 1.2° (Fig. 16). We have assumed that the angular extent of the celestial source is small compared to the width of the narrow lobes. If the interferometer antennas are separated by the Earth's diameter the resolution at 10 cm wavelength is *one millionth of a degree* (a few milliarcseconds).

A typical two-element interferometer has two steerable antennas of aperture a connected by equal-length cables to a receiver R as in Fig. 17. The antenna separation normal to the source direction is s. Whereas a single antenna has a resolution of $57.3/a_\lambda$ degrees where $a = a/\lambda$, the resolution of the interferometer is equal to $57.3/(s_\lambda + a_\lambda)$ degrees, or, if $s_\lambda \gg a_\lambda$, by $57.3/s_\lambda$ degrees. This resolution is the Rayleigh resolution (=BWFN/2). Any observed structure smaller than this angle will not be resolved.

When $d = n\lambda$ in Fig. 17, the response is maximum, and when $d = n\lambda/2$ the response is minimum (n = integer).

In Fig. 18, the interferometer response is shown as a function of source angle ϕ for three cases. In Fig. 18(a) the source extent is very small compared to the resolution angle (lobe spacing), in Fig. 18(b) it is comparable to but less than the resolution angle, and in Fig. 18(c) it is equal to the resolution angle. The amplitude $V(s_\lambda)$ of the lobe pattern as a function of the spacing s_λ is called the fringe

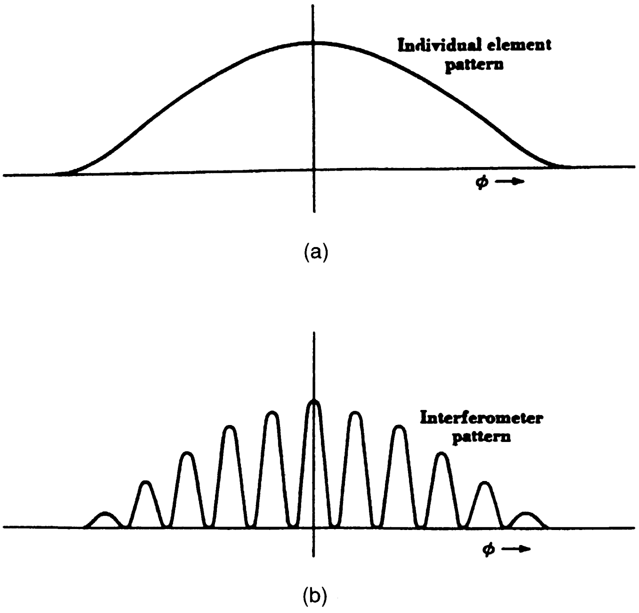

FIG. 16. Interferometer patterns (as a function of source angle ϕ) showing (a) the individual antenna pattern and (b) the actual interferometer pattern with narrow lobes within the envelope of the individual antenna pattern.

amplitude, fringe visibility, or simply the visibility as given by

$$V(s_\lambda) = \frac{S_{\max} - S_{\min}}{S_{\max} + S_{\min}}. \tag{10}$$

By moving the antennas over a full range of spacings up to L and observing $V(s_\lambda)$ at each spacing, a visibility function is obtained. Taking its Fourier transform yields the source brightness distribution. In principle this is equivalent to the distribution obtained by observations with a continuous filled aperture of length L. This process is called *aperture synthesis* (Ryle and Hewish, 1960).

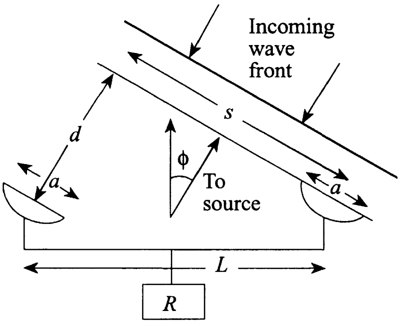

FIG. 17. Two-element interferometer with individual dish antenna diameters *a*, both antennas connected by equal-length cables to receiver *R*.

In two-dimensional aperture synthesis with many antennas (instead of just two) a celestial object is tracked by the system as the Earth rotates, with each antenna-pair combination producing a complex visibility (both amplitude and phase). A map or image of the object or region is then constructed from the visibility function. Using this technique the Very Large Array (VLA) of the National Radio Astronomy Observatory near Socorro, New Mexico (Fig. 7) (with 27 dish-reflector antennas of 25 m diam deployed at stations having a Y-shaped configuration with each arm 21 km long), is capable of resolutions at centimeter wavelengths of less than 1 arcsec providing synthesized images of 10^7 pixels (picture elements).

The spectacular radio picture of Cygnus A in Fig. 19 was obtained with this array (Perley *et al.*, 1984). The picture demonstrates the power of aperture synthesis combined, in this case, with image enhancement using "phase closure." This data-processing technique is applied following the observations and rids the image of spurious responses (Ekers, 1983; Thompson *et al.*, 1986; Kraus, 1986, pp. 6–49, 6–50).

The two bright lobes represent radio hot spots or emitting regions with relativistic jets streaming in opposite directions from a central engine midway between the bright lobes. This radio picture reveals complex filamen-

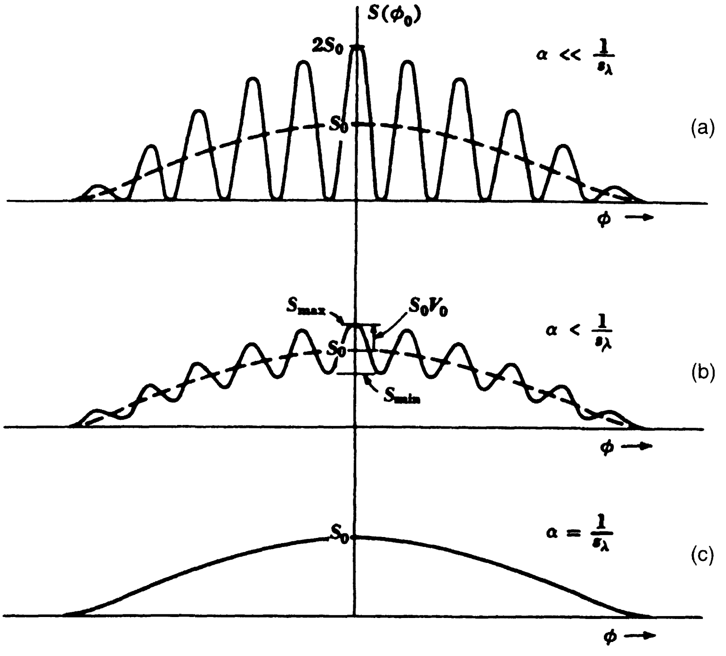

FIG. 18. Interferometer patterns (as a function of source angle ϕ) (a) for a point source, (b) for a uniform extended source of angle $\alpha < 1/s_\lambda$ and (c) for a uniform extended source of angle $\alpha = 1/s_\lambda$.

tary structure in both lobes. The overall extent of the emitting regions is about 400 000 light years or about four times our galaxy's diameter. The overall angular extent is about 115 arcsec and the resolution is 0.4 arcsec. Cygnus A is one of the strongest objects in the radio sky. Long-exposure optical photographs with the largest optical telescopes show only a small faint galaxy where the radio engine is located. There is little hint optically of the violent events that have transpired in Cygnus A.

With an actual full aperture, such as a large parabolic reflector, the incoming waves falling on each aperture area are added in phase at a focal point. In aperture synthesis, many measurements are made with different spacings of the small elemental apertures and the results summed by a computer, which may also introduce suitable weighting and phasing for controlling the synthesized antenna pattern.

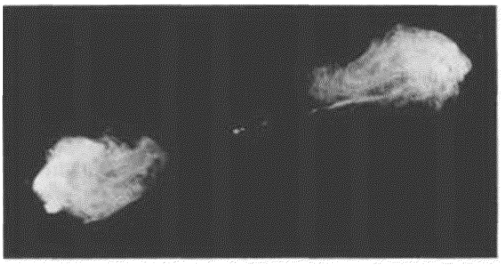

FIG. 19. The Cygnus-A double-lobe radio galaxy with filamentary structure ejected by relativistic jets from a central engine. The double structure is 400-000 light-years across. Obtained with the VLA at 5 GHz. (Photo courtesy N.R.A.O.)

The VLA is capable of antenna separations up to 36 km using physical (waveguide) radio-frequency connections. However, such connections are not feasible for separations of thousands of kilometers. Nevertheless, with antennas separated by 10 000 km it is possible to obtain milliarcsecond resolution with the following technique. In this very long-baseline interferometry (VLBI) the observed signals at widely spaced observatories are down-converted to a lower or intermediate frequency and these signals along with precise time signals are taped and sent to a central location. By synchronizing the tapes using atomic time standards, a real-time comparison can be simulated to create a radio image. Such observations require a high degree of coordination between the participating observatories and, being time-consuming, have only been performed on an intermittent basis. To avoid these problems a dedicated very long-baseline array (VLBA) has been constructed that has 10 antennas with locations coast to coast in the contiguous states of the US and in Hawaii and St. Croix (Kellermann, 1991). This array is capable of milliarcsecond resolution at centimeter wavelengths. Adding an orbiting radio telescope can increase the baselines and improve the resolution even more. The feasibility of combining ground- and space-based telescope data has already been demonstrated and plans are being made for more routine use of this technique.

To exploit the opportunities of millimeter-wave astronomy (see Sec. 3.3 and Fig. 22) a millimeter array (MMA) is planned by US astronomers. It will consist of a ring of 40 8-m dish antennas that can be deployed in a variety of configurations such as in a ring 250 m in diameter.

2.7 Radar Telescopes

By adding a transmitter to a radio telescope it can be made into a radar telescope, or active remote-sensing device, the antenna being alternately connected to the transmitter that sends pulses and then to the receiver for reception of the echoes. By reciprocity the pattern of the antenna is the same for both receiving and transmitting. According to the radar equation the echo power received, P_r, is related to the power transmitted, P_t, by

$$P_r = \frac{P_t A^2 \sigma}{4\pi r^4 \lambda^2} \quad \text{(W)}, \tag{11}$$

where A = effective aperture of antenna, m^2, σ = radar cross section of reflecting object, m^2, r = distance to object, m, and λ = wavelength, m.

The fact that the received power falls off as the fourth power of the distance places a constraint on the range of radar telescopes. Nevertheless powerful Earth-based telescopes have probed the sun, moon, Venus, Mercury, Mars, and Jupiter. Distances (or ranges) have been measured with unprecedented precision and reflectivities determined. From the Doppler spread in frequency of an echo the object's rotation can be observed. The rotation rates of Mercury (59 days) and Venus (250 days, retrograde) were first determined by radar telescopes on the Earth (Pettengill and Dyce, 1965; Goldstein and Carpenter, 1963; Goldstein, 1964). Echoes from areas on the object that are receding are downshifted in frequency and those from areas that are approaching are shifted up in frequency. From these Doppler shifts the rotation rates of Mercury and Venus were determined.

These interplanetary radar observations have resulted in improved values of planetary orbits, masses, and radii, disclosing errors in the ephemerides of the Earth and planets. The astronomical unit (Earth-sun distance) has been significantly improved. With radar, distances can be measured directly whereas optically they are inferred by indirect means.

A spectacular radar-astronomy achievement is the mapping of the surface of Venus through its opaque atmosphere by the spacecrafts Magellan and Pioneer Venus orbiting above it with data telemetered to the Earth (Fig. 20). This map covers about 85% of the surface of Venus and reveals two highlands: Ishtar Terra, as large as the continental US, and Aphrodite Terra, of comparable size. Maxwell Montes has an elevation of 8 km. The highest and lowest elevations differ by 15 km. Maat Mons is an 8-km-high volcanic peak. The radars operated at about 2.5 GHz with horizontal surface resolution of about 100 m.

3. RADIO-TELESCOPE RECEIVERS

3.1 Total-Power and Dicke Types

In the simplest, or total-power, receiver, the antenna is connected to a low-noise amplifier (LNA) and then to a mixer in which the signal is combined with that from a local oscillator to down-convert the signal to a lower or intermediate frequency (IF). This signal is then amplified, detected, integrated, and recorded [Fig. 21(a)].

In the popular Dicke receiver, the LNA is alternately switched from the antenna to a reference resistor at temperature T [Fig. 21(b)]. If T is the same as the equivalent temperature of the sky being observed, the effect of gain fluctuations in the receiver is minimized. To reduce the temperature of the resistor to sky values requires refrigeration of the resistor on ground-based telescopes. These refrigerated resistors, called "cold

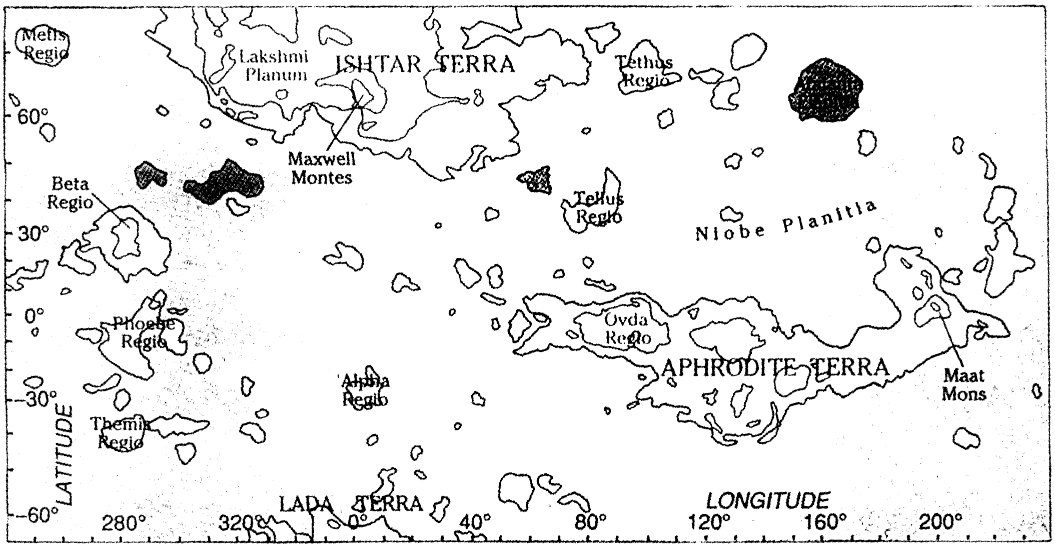
Metis Regio
Lakshmi Planum
ISHTAR TERRA
Tethus Regio
60°
Maxwell Montes
Beta Regio
Tellus Regio
Niobe Planitia
30°
0°
Phoebe Regio
Ovda Regio
APHRODITE TERRA
Maat Mons
-30°
Alpha Regio
Themis Regio
LATITUDE
-60°
280°
320°
LADA TERRA
0°
40°
80°
LONGITUDE
120°
160°
200°

FIG. 20. Surface structure of 85% of Venus obtained by orbiting radar telescopes on Pioneer Venus and Magellan (Courtesy NASA).

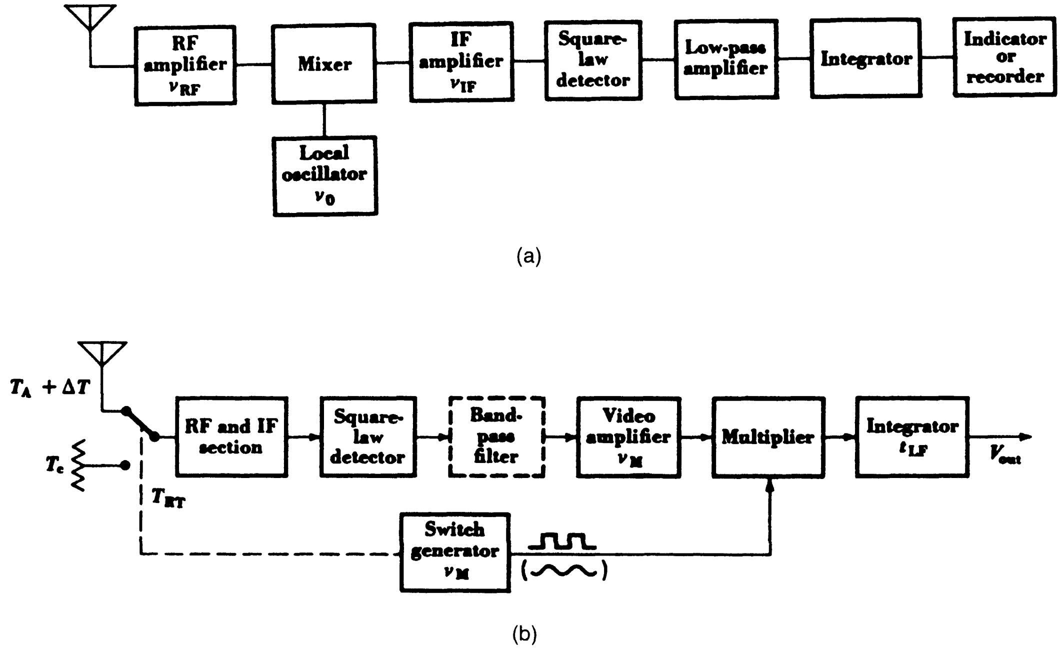
RF amplifier ν_{RF}
Mixer
IF amplifier ν_{IF}
Square-law detector
Low-pass amplifier
Integrator
Indicator or recorder
Local oscillator ν_0
(a)
$T_A + \Delta T$
T_c
RF and IF section
Square-law detector
Band-pass filter
Video amplifier ν_M
Multiplier
Integrator t_{LF}
V_{out}
T_{BT}
Switch generator ν_M
(b)

FIG. 21. (a) A superheterodyne radio-telescope receiver and (b) a switched or Dicke type.

loads," can be used to calibrate the telescope for accurate measurements of the sky temperature. It was by precise measurements of the sky temperature that the 3-K background was discovered (Penzias and Wilson, 1965).

3.2 Low-Noise Amplifiers

The front-end amplifier may be a GaAs FET (gallium arsenide field effect transistor) which can provide a gain of over 1000 (or over 30 dB gain) at 1.5 GHz with a noise temperature of only 60 K at ambient temperature (296 K) and 10 K when refrigerated to liquid nitrogen temperature (77 K). The high gain also reduces the effect of noise from the rest of the receiver. Thus, if the next section has a noise temperature of 1000 K, its contribution to the overall or system temperature is only 1 K [=(1000 K)/1000].

For millimeter-wavelength receivers the HEMT (high-electron mobility transistor) is popular because of its higher frequency of operation (over 100 GHz) and lower noise temperature. The SIS (superconductor-insulator-superconductor) quasiparticle mixer has a very low noise temperature set by the quantum limit.* Thus, when refrigerated to 4 K or less its noise temperature is about 7 K at 100 GHz (λ = 3 mm) (Tiuri and Räisänen, 1986). It is used without a radio-frequency amplifier (or front end) ahead of it, no low-noise amplifier being available at millimeter wavelengths.

3.3 Multichannel Receivers

There are over 100 interstellar and circumstellar molecular species that have been observed in the radio microwave region of the spectrum. For such observations receivers with multichannel spectrum analyzers are used with individual channel bandwidths appropriate to the observations. Molecular spectra obtained in a single observation toward the Orion-KL cloud are shown in Fig. 22 as observed between 88 and 90 GHz (3.4 mm) with the Nobeyama 45-m telescope using a receiver with 16 000 channels, each with a bandwidth of 125 kHz. Nearly 20 molecular lines are in evidence of which 15 are identified (Kaifu, 1985).

In some millimeter-wavelength telescopes an *x-y* array or cluster of feed horns is placed at the focus with each horn connected to its own receiver. This creates a radio camera giving a two-dimensional display of a molecular cloud or other region *n* times as fast as by scanning with a single feed horn, where *n* is the number of horns in the array (Erickson *et al.*, 1992).

In the search for extraterrestrial intelligence (SETI), receivers with spectral analyzers of over 1 million channels are in use with ones of a billion channels coming on

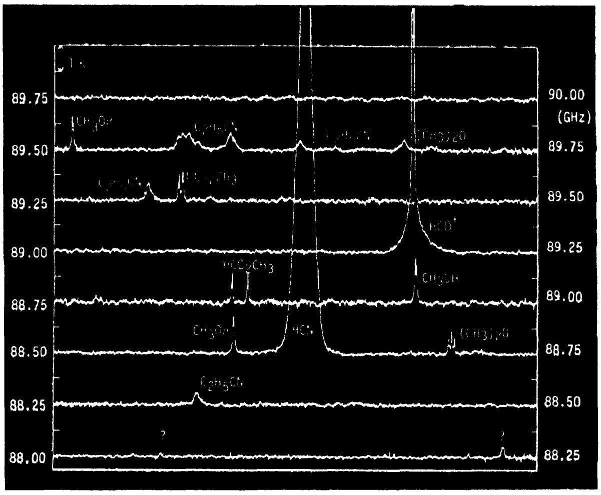

FIG. 22. Molecular spectra observed between 88 and 90 GHz in the Orion-KL cloud with the Nobeyama 45-m radio telescope. There are 15 identified molecular lines in the spectrum with HCN (hydrogen cyanide) and HCO^+ (formylium) being the strongest. (Courtesy N. H. Kaifu.)

---

*T(quantum) = $(h/k \ln 2)f$(Hz) = $0.07f$ (GHz)K and for f = 100 GHz, T(quantum) = 7 K where h = 6.63×10^{-34} J s (Planck's constant) and k = 1.38×10^{-23} J K^{-1} (Boltzmann's constant).

line. At the Harvard-Smithsonian Observatory (Horowitz and Sagan, 1993), a receiver with 8 million channels has been operating in which the incoming signal (a function of time) is fast Fourier transformed by a bank of computers into the frequency domain for spectral analysis (Fig. 23).

In the Argus concept (Dixon, 1996), an array of many simple antennas covering a large flat area is connected to a powerful computer providing multiple beams for simultaneous observation and surveillance of nearly a hemisphere of sky. Each antenna has its own receiver and digital converter. Such digital-beam-forming (DBF) arrays have great potential for the future (Kraus, 1988).

3.4 Antenna Temperature and Sensitivity (Remote Sensing)

A resistor of resistance R and temperature T matched to a receiver delivers a power $P = kT\Delta f$ watts where k = Boltzmann's constant (1.38×10^{-23} J K^{-1}), T = absolute temperature of the resistor (K), and Δf = bandwidth (Hz). If the resistor is replaced by an antenna of radiation resistance R, the radiation resistance is not at the physical temperature of the antenna but at the effective temperature of regions at which the antenna is directed (antenna assumed lossless). Thus, the radio telescope is a remote-sensing device that may be regarded as a bolometer (or heat-measuring device) for determining the

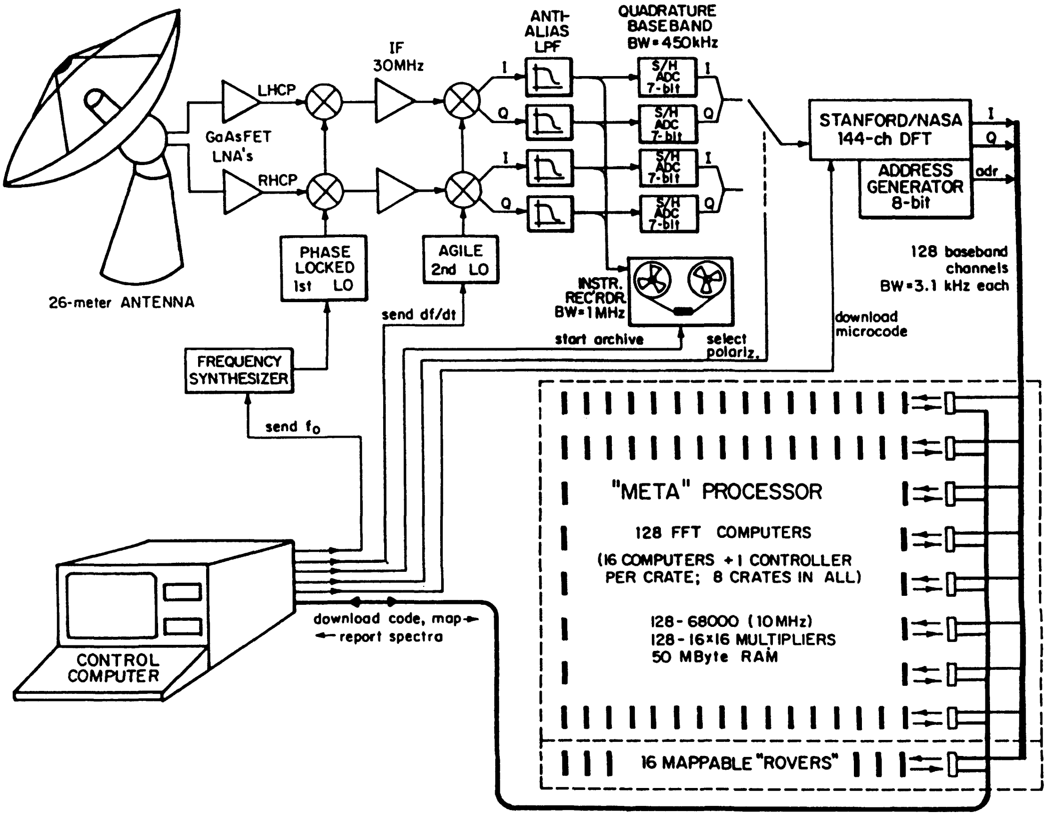
ANTI-ALIAS LPF
QUADRATURE BASEBAND BW=450kHz
IF 30MHz
LHCP
RHCP
GaAsFET LNA's
I
Q
I
Q
S/H ADC 7-bit
S/H ADC 7-bit
S/H ADC 7-bit
S/H ADC 7-bit
I
Q
I
Q
STANFORD/NASA 144-ch DFT
I
Q
ADDRESS GENERATOR 8-bit
adr
128 baseband channels BW=3.1 kHz each
download microcode
PHASE LOCKED 1st LO
AGILE 2nd LO
26-meter ANTENNA
send df/dt
INSTR. RECRDR BW=1MHz
start archive
select polarlz.
FREQUENCY SYNTHESIZER
send f₀
"META" PROCESSOR
128 FFT COMPUTERS
(16 COMPUTERS + 1 CONTROLLER PER CRATE; 8 CRATES IN ALL)
128-68000 (10MHz)
128-16×16 MULTIPLIERS
50 MByte RAM
16 MAPPABLE "ROVERS"
download code, map
report spectra
CONTROL COMPUTER

FIG. 23. Fourier spectrometer receiver of 8 × 10^6 channels with 0.05-Hz resolution and 400-kHz instantaneous bandwidth used on the Harvard-Smithsonian 26-m radio telescope in Massachusetts for the Horowitz–Sagan (1993) META (megachannel extraterrestial assay). (Courtesy Prof. P. Horowitz, Harvard University.)

effective temperature of regions within the antenna beam coupled to the system through the radiation resistance of the antenna.

The power received by a radio telescope from a celestial source is $P = SA_e\Delta f$ (W) where S = source flux density, A_e = effective aperture of telescope antenna, and Δf = bandwidth. If this power equals that of a matched (calibration) resistor at temperature T connected to the receiver in place of the antenna, the flux density of the source is given by $S = kT/A_e$. But if the source is unpolarized, as most celestial sources are, the flux density is twice this value.

Sensitivity is a function of the antenna aperture and also the *system temperature.* Thus, the sensitivity, or minimum detectable flux density, is given by

$$S_{\min} = 2kT_{\min}/A_e,$$

where $T_{\min} = K_s T_{\text{sys}}/\sqrt{\Delta f\, t}$ where K_s = receiver sensitivity constant (order of unity) (dimensionless) and t = output time constant (s). The system temperature depends principally on the effective temperature of the sky and the receiver noise temperature, especially that of the front-end amplifier.

The signals received by a radio telescope have the characteristics of thermal noise. Their magnitude is measured by comparing it with the noise signal of a resistor at a known temperature. The receiver also contributes noise in its system temperature. Thus, it is aptly said that radio astronomers use their telescopes to "measure noise, with noise, in noise."

3.5 Remote Sensing of the Earth

Radio telescopes in orbit around the Earth with antennas pointed to the ground are used in environmental studies of our planet, providing information on surface temperature and reflectivity. For example, the microwave brightness of the Antarctic continent has been measured with an orbiting radio telescope, the geostationary environmental satellite (GOES). This satellite spins on an axis parallel to the Earth's, providing an east–west scan. At the end of an east–west scan the beam is stepped to the north or south and the process repeated—resulting in a temperature map of the earth's surface. The scanning is similar to the raster scanning of the spot of a TV tube.

GLOSSARY

Aperture Efficiency: Ratio of the effective aperture area of an antenna and its physical area.

Aperture Synthesis: Method of obtaining the equivalent of a large aperture with two or more small apertures.

Beam Area: Solid angle in square radians or square degrees subtended by an antenna beam.

Brightness: A measure of the power received per unit area per unit solid angle per unit bandwidth from a radiating object such as the sky. It is expressed in watts per square meter per hertz per square radian.

BWFN: Beamwidth between the first nulls of an antenna pattern.

Dicke Receiver: Radio receiver with switched input for improving receiver stability.

Directivity: Ratio of the power radiated per unit solid angle in the direction of the main beam of an antenna to that in all directions from an isotropic or nondirectional antenna. It may also be expressed as 4π divided by the beam area, both in steradians (or square radians).

Effective Aperture: If a plane wave is incident on a lossless antenna, the effective aperture is the ratio of the power received to the incident power flux density. Thus, if the incident flux density is 1 watt per square meter and the power received is 2 watts, the effective aperture is 2 square meters.

Field Intensity: Usually the electric field intensity or strength expressed in volts per meter.

Flux Density: Power flux of an electromagnetic wave per unit area per unit bandwidth (W m^{-2} Hz^{-1}).

HEMT Amplifier: High–electron-mobility transistor low-noise amplifier.

Holographic Technique: Measurement of the amplitude *and* phase of a beacon signal for mapping the surface of a parabolic reflector.

Homologous Design: Structural design of a reflector antenna in which the parabolic reflector deforms into another as the reflector is tilted.

HPBW: Half-power beamwidth of an antenna pattern. The angular width of the pattern at the half-power level.

Interferometer: In radio astronomy, two or more spaced antennas used to obtain narrower beamwidths. See **Aperture Synthesis.**

Jansky: Unit of flux density equal to 10^{-26} W m^{-2} Hz^{-1}.

LNA: Low-noise amplifier.

Minimum Detectable Flux Density: The minimum flux density of a radio source that can be measured by a telescope in the presence of its noise.

Normalized Power Pattern: The antenna pattern divided by its maximum value so that the maximum normalized power in the direction of the main beam is unity.

Pattern: Variation of power flux (or field intensity) of an antenna as a function of angle.

Physical Aperture: Actual physical cross section of antenna. See also **Effective Aperture.**

Power Flux Density: Amount of power flow through unit area, usually expressed in watts per square meter.

Rayleigh Resolution: One half the beamwidth between first nulls of an antenna pattern. Thus, the Rayleigh resolution equals BWFN/2. It is 12% more than the half-power beamwidth or 1.12 HPBW. Two point sources separated by this angle will just be resolved.

SIS Mixer: Superconductor-insulator-superconductor quasiparticle low-noise mixer.

System Temperature: The system temperature is a factor in determining the sensitivity of a radio telescope. It depends on the noise temperature of the sky, the receiver noise temperature, and other factors. The smaller the system temperature the more sensitive the radio telescope. Front-end amplifiers of radio telescopes are often refrigerated to reduce this temperature.

Visibility: Amplitude of the lobe pattern as a function of the spacing of interferometer antennas. Also called the fringe visibility.

Works Cited

Akabane, K. M., Morimoto, M., Kaifu, N., Ishiguro, M. (1983), *Sky Telescope,* **74,** 495–498.

Blum, E. J., Boischot, A., Lequeux, J. (1963), *Proc. IRE Aust.* **24,** 208–213.

Christiansen, W. N., Högbom, J. A. (1985), *Radiotelescopes,* Cambridge, U.K.: Cambridge Univ. Press, Sec. 3.2.

Dixon, R. S. (1996), "Argus, A Future SETI Telescope," in: J. Heidmann (Ed.), Proceedings of the 44th International Astronautical Conference, Graz, Austria, *Acta Astronaut.* (to be published).

Ekers, R. D. (1983), "Radio Astronomy between Jansky and Reber" in: K. Kellermann, B. Sheets (Eds.), *Serendipitous Discoveries in Radio Astronomy, Proceedings of N.R.A.O. Workshop No. 7,* Green Bank, WV: N.R.A.O., pp. 154–159.

Erickson, N. R., Goldsmith, P. F., Novak, G., Grosslein, R. M., Viscuso, P. J., Erickson, R. B. and Predmore, C. R. (1992), *IEEE Trans. Microwave Theory Tech.* **MTT-40,** 1–11.

Findlay, J. W. (1964), *IEEE Trans. Antennas Propag.* **AP-12,** 853–864.

Goldsmith, P., Erickson, N. R. (1986), in: J. D. Kraus (Ed.), *Radio Astronomy,* 2nd ed., Powell, OH: Cygnus-Quasar, pp. 6–63, 6–67.

Goldstein, R. M. (1964), *IEEE Trans Antennas Propag.,* **AP-12,** 865–872.

Goldstein, R. M., Carpenter, R. L. (1963), *Science* **139,** 910–911.

Golubin, G. S. (1982), *Spec. Astrophys. Obs. Publ.* **36,** 5–48.

Greve, A., Dan, M., Penalver, J. (1992), *IEEE Trans. Antennas Propag.* **AP-40,** 1375–1388.

Horowitz, P., Sagan, C. (1993), *Astrophys. J.* **415,** 218–235.

Jansky, K. G. (1933), "Electrical Disturbances Apparently of Extraterrestrial Origin," *Proc. IRE* **21,** 1387–1398.

Kaifu, N. H. (1985), private communication.

Kellermann, K. I. (1991), *Sky Telescope* **82,** 247–253.

Kraus, J. D. (1986), *Radio Astronomy,* 2nd ed., Powell, OH: Cygnus-Quasar.

Kraus, J. D. (1988), *Antennas,* 2nd ed., McGraw-Hill, pp. 496–500.

Lovell, B. (1968), *The Story of Jodrell Bank,* New York: Harper and Row.

Mills, B. Y., Aitchison, R. E., Little, A. G., McAdam, W. B. (1963), *Proceedings IRE Australia,* Vol. 24, pp. 156–165.

Napier, P. J., Thompson, A. R., Ekers, R. D. (1983), *Proc. IEEE* **71,** 1295–1320.

Parijskij, Y. (1993), *IEEE Antenna Propag. Mag.* **35** (4), 7–12.

Penzias, A. A., Wilson, R. W. (1965), *Astrophys. J.* **142,** 419–421.

Perley, R. A., Dreher, J. W., Cowan, J. J. (1984), *Astrophys. J.* **285,** L35–L38.

Pettengill, G. H., Dyce, R. B. (1965), *Nature* **206,** 1240.

Reber, G. (1940), "Cosmic Static," *Astrophys. J.* **91,** 621–624.

Reber, G. (1983), "Radio Astronomy between

Jansky and Reber," in: K. Kellermann, B. Sheets (Eds.), *Serendipitous Discoveries in Radio Astronomy, Proceedings of N.R.A.O. Workshop No. 7,* Green Bank, WV: N.R.A.O., pp. 71–78.

Ruze, J. (1966), *Proc. IEEE,* **54,** 633–640.

Ryle, M., Hewish, A. (1960), *Mon. Not. R. Astron. Soc.* **120,** 220–230.

Taylor, J. H. (1981), *IAU Symposium No. 95,* Dordrecht: Reidel, pp. 361–367.

Tiuri, M. E., Räisänen, A. V. (1986), in: J. D. Kraus (Ed.), *Radio Astronomy,* 2nd ed., Powell, OH: Cygnus-Quasar, pp. 7–56, 7–57.

Thompson, A. R., Moran, J. M., Swenson, G. W., Jr., (1986), *Interferometry and Synthesis in Radio Astronomy,* New York: Wiley-Interscience, pp. 356–364.

Von Hoerner, S. (1977), *Vistas Astron.* **20** (4), 411–444.

Wohlleben, R., Sieber, W. (1968), in: *IEEE Antennas and Propagation Society, International Symposium, Syracuse, New York,* Vol. 3, New York: IEEE.

Further Reading

Christiansen, W. N., Högbom, J. A. (1969, 1985), *Radio Telescopes,* Cambridge, U.K.: Cambridge Univ. Press. Discusses all types.

Evans, J. V., Hagfors, T. (Eds.) (1968), *Radar Astronomy,* New York: McGraw-Hill.

Goldsmith, P. F. (1994), *Instrumentation and Techniques for Radio Astronomy,* Piscataway, NJ: IEEE Press. A reprint collection of 70 pertinent articles.

Hord, R. M. (1986), *Remote Sensing, Methods and Applications,* New York: Wiley.

Kellermann, K., Sheets, B. (Eds.) (1983), *Serendipitous Discoveries in Radio Astronomy,* Green Bank, WV: N.R.A.O. Historical insights.

Kellermann, K. I. (1991), "Radio Astronomy: The Next Decade," *Sky Telescope* **82,** 247–253. Describes future radio telescopes.

Kraus, J. D. (1986), *Radio Astronomy,* 2nd, ed., Powell, OH: Cygnus-Quasar Books. Basic text; complete coverage.

Kraus, J. D. (1988), *Antennas,* 2nd ed., New York: McGraw-Hill. Basic text.

Thompson, A. R., Moran, J. M., Swenson, G. W., Jr. (1986), *Interferometry and Synthesis in Radio Astronomy,* New York: Wiley-Interscience. Complete theory.

Ulaby, F. T., Moore, R. K., Fung, A. K. (1981), *Microwave Remote Sensing, Active and Passive,* Reading, MA: Addison-Wesley.

11

Radioactivity

Walter Loveland

11.1 Introduction 347
11.2 Nature of Radioactivity 347
11.2.1 Which Nuclei Decay? 347
11.2.2 Natural Radioactivity 349
11.2.3 Artificial Radioactivity 351
11.2.4 The Elements beyond Uranium 351
11.3 Radioactive Decay Kinetics 353
11.3.1 Basic Decay Law 353
11.3.2 Radioactive Decay Equilibrium 354
11.3.3 Dating 354
11.4 α-Decay 355
11.5 β-Decay 357
11.5.1 Types of β-Decay 357
11.5.2 Energetics 358
11.5.3 Decay Mechanism 358
11.5.4 Electron Capture 360
11.5.5 Double β-Decay 361
11.5.6 β-Delayed Radioactive Decay 361
11.6 Electromagnetic Decay 362
11.6.1 Types and Classification 362
11.6.2 Transition Rates 362
11.6.3 Internal Conversion 363
11.7 Other Types of Radioactive Decay 363
11.7.1 Direct Proton Emission 363
11.7.2 Heavy-Particle Radioactivity 364
11.7.3 Spontaneous Fission 364
11.8 Detection and Measurement of Radioactivity 365
11.8.1 Choice of Detector 365
11.8.2 Sample Preparation 365
11.8.3 Correction Factors and Uncertainties 365
11.9 Applications of Radioactivity 369

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

11.9.1 Tracers 369
11.9.2 Thickness Gauges 370
11.9.3 Nuclear Medicine 370
11.9.4 Nuclear Science 371
Glossary 372
References 372
Further Reading 373

11.1
Introduction

Nuclear chemistry and physics originated with the study of natural radioactivity. Many important insights into the properties of nuclear matter have been obtained through the study of radioactive decay. The study of radioactivity continues to play a significant role in nuclear science research even today.

As important as these fundamental studies of radioactivity have been and continue to be, most users of radioactivity are not nuclear physicists or chemists, but those who work in a wide variety of endeavors where radioactivity is simply a tool. This tool is no more special than any other laboratory instrument such as an infrared spectrometer, a computer, and so on. This review is meant for such users.

Any number of standard nuclear chemistry and physics textbooks contain excellent, thorough, in-depth discussions of radioactivity, its measurement, and its applications (see the section Further Reading at the end of this chapter for examples) The aim of this chapter is to provide a terse account of the most important features of radioactivity with the hope of piquing the reader's interest and serving as a source of information for further inquiries.

We begin by discussing what causes radioactive decay, and which nuclei are radioactive. In the second section, we discuss the equations that describe radioactive decay and dating. The next three sections are devoted to a more detailed discussion of the three main types of radioactive decay; α-, β-, and electromagnetic decay. In each of these sections, we seek to describe the basic decay mode, its energetics, and its occurrence and to provide a simplified summary of the decay mechanism and any appropriate selection rules. A final section on radioactive decay describes other, more exotic decay modes. Following these discussions of the physical principles behind radioactive decay, there are brief sections dealing with the detection and measurement of radioactivity and some sample applications. The section on detectors and measurements is brief because this book contains several full-length articles on individual detector types and their use.

11.2
Nature of Radioactivity

11.2.1
Which Nuclei Decay?

Radioactivity is the spontaneous disintegration of an atomic nucleus by emission

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

of some form of matter and/or energy. Nuclei that undergo radioactive decay and the decay process itself are referred to collectively as *radioactivity*. The types of radioactive decay include α-decay, β-decay, electromagnetic decay, and so on. (see Sections 11.4–11.6 for definitions and details). All nuclei will disintegrate if raised to a high enough excitation energy, but the term radioactivity is usually reserved for nuclei with little or no excitation energy. Radioactive decay is thus an exoergic process.

The release of energy in decay processes can be caused by a variety of circumstances. For example, the stable nuclei have relatively well-defined neutron-to-proton ratios (Figure 11.1). For the lighter nuclei, the n/p ratio is about 1, while this ratio approaches a value of 1.5 for the heavier nuclei. (The change in this ratio reflects the changing balance between the strong and electromagnetic forces as the nuclear composition changes.) Nuclei with n/p ratios greater than those of stable nuclei decay by β^- decay while those with n/p ratios less than those of stable nuclei decay by β^+ or electron capture (EC) decay.

Another way to represent this concept is to plot the mass of the nuclei versus the number of neutrons (N) and the number of protons (Z) in the nucleus (Figure 11.2). The stable nuclei lie at the bottom of the "valley of stability" while the radioactive nuclei lie at higher elevations and decay by dropping down into the valley of stability. For the heaviest nuclei (not shown in Figure 11.2), there is a headwall in the valley such that these nuclei decay by

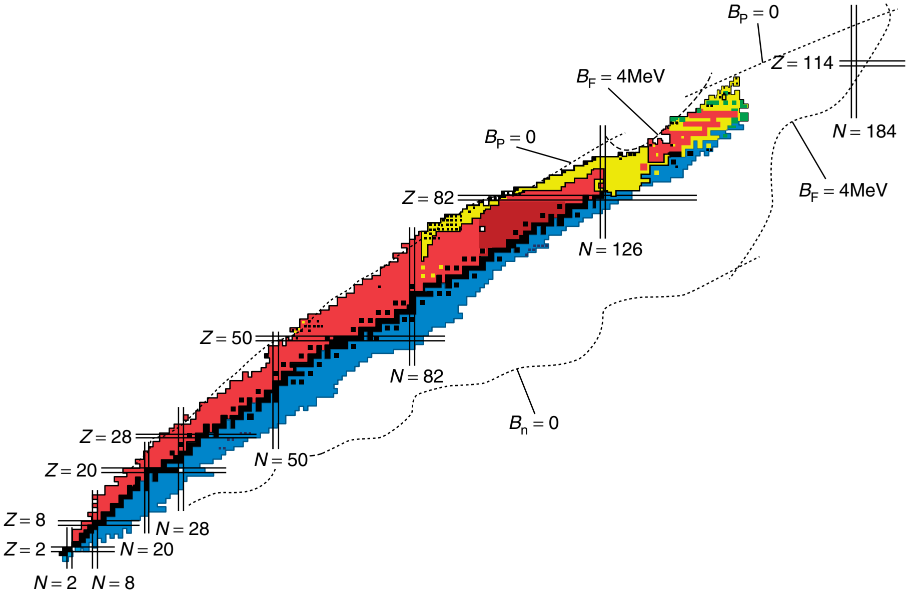

Fig. 11.1 Plot of all the known stable or radioactive nuclei, arranged by atomic number Z and neutron number N. The dashed contour shows the possible range of all bound nuclei. The black squares indicate the stable nuclei, the blue areas represent β^- emitters, the red areas represent nuclei decaying by β^+ decay or EC, the yellow regions represent the α-emitters, while the green regions indicate the nuclei decaying primarily by spontaneous fission. The "magic" numbers are shown by pairs of horizontal or vertical lines.

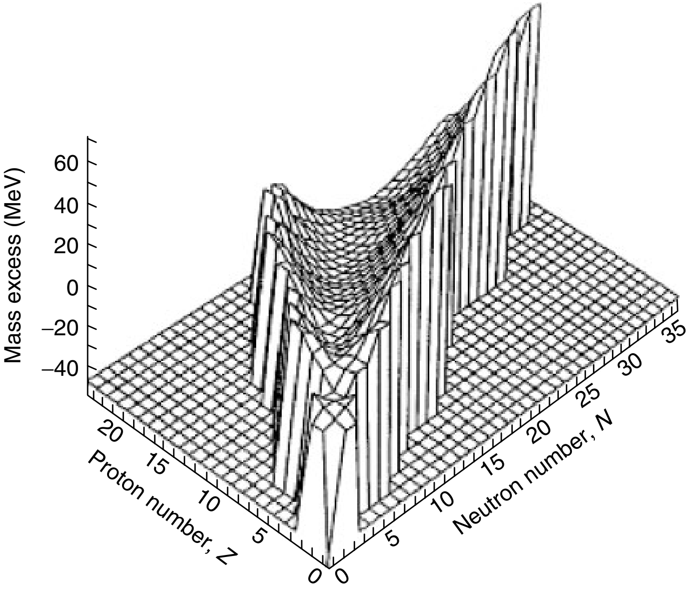

Fig. 11.2 A portion of the nuclear mass surface, showing the valley of β-stability. (From Halliday, Resnick, and Krane 2001; reproduced by permission of John Wiley & Sons, Inc.)

emitting heavy particles such as He nuclei (α-particles) or spontaneously fissioning.

11.2.2
Natural Radioactivity

One might inquire as to how a nucleus is created such that it is unstable, that is, has an unfavorable n/p ratio. All nuclides, stable and radioactive, except hydrogen, helium, and lithium-7, were made in nuclear reactions in the interior of stars or in explosive supernovas. Our solar system is composed primarily of nuclides that were formed about 4.5×10^9 years ago. If the half-life of a radioactive nuclide is greater than 4.5×10^9 years, then such original nuclei will still exist and their decay can be observed. Examples of these naturally radioactive nuclei include the heavy-element decay series, that is, the uranium, thorium, and actinium series (Figure 11.3). Each of these series consists of a long-lived parent nuclide that decays to a daughter nuclide that decays to a granddaughter, and so on, ending in a stable isotope of lead.

The thorium, uranium, and actinium series begin with long-lived primordial parents. These series are found in nature, mostly in rocks or minerals. The parent of another artificial decay series, the neptunium series, is extinct (because its half-life is less than the age of the solar system) and so this series is found only in the laboratory. The thorium, uranium, and actinium series are also known as the $4n, 4n + 2$, and $4n + 3$ series because the mass number of each member of a given series is numerically equal to $4n$ or $4n + 2$ or $4n + 3$ where n is an integer. In addition to the normal way of representing a nuclide as $^{A}_{Z}$ (Chemical

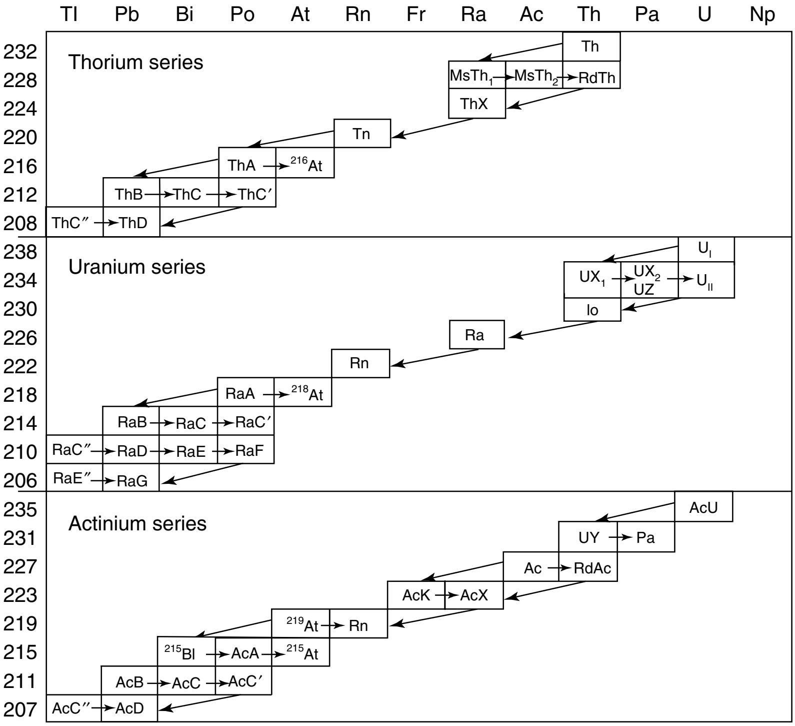
Tl
Pb
Bi
Po
At
Rn
Fr
Ra
Ac
Th
Pa
U
Np
Thorium series
232
228
224
220
216
212
208
Th
MsTh$_1$
MsTh$_2$
RdTh
ThX
Tn
ThA
^{216}At
ThB
ThC
ThC′
ThC″
ThD
Uranium series
238
234
230
226
222
218
214
210
206
U$_I$
UX$_1$
UX$_2$
UZ
U$_{II}$
Io
Ra
Rn
RaA
^{218}At
RaB
RaC
RaC′
RaC″
RaD
RaE
RaF
RaE″
RaG
Actinium series
235
231
227
223
219
215
211
207
AcU
UY
Pa
Ac
RdAc
AcK
AcX
^{219}At
Rn
^{215}Bi
AcA
^{215}At
AcB
AcC
AcC′
AcC″
AcD

Fig. 11.3 The uranium, thorium, and actinium decay series.

Symbol)$_N$ shown in Figure 11.3, one also sees an archaic representation for each member of the series such as UI, RaC, and so on. These archaic symbols date from the original work by chemists who isolated each member of the series from ores. Some of these symbols, including ThC, ThC′, RaD, RaE, and RaF, are still used. Note that the heavy-element decay series involve isotopes of radon, which is a gas. Natural decays in rocks that occur near the surface of the earth or that occur in uranium-containing materials, such as concrete, can release this gas to the environment. The hazards associated with environmental levels of radon are a subject of current research.

Other naturally occurring radionuclides exist, such as ^{40}K, ^{87}Rb, ^{113}Cd, ^{115}In, ^{138}La, ^{176}Lu, and ^{187}Re. There are also important naturally occurring cosmogenic radionuclides that are continuously produced by the interaction of cosmic rays (mostly high-energy protons) with the earth's atmosphere. Examples of these include radiocarbon (^{14}C) and tritium (^{3}H) (Lal and Suess, 1968). Another component of environmental radioactivity is those radionuclides released to the environment as a result of the activities of man, primarily through fallout from nuclear weapons testing. Prominent among these nuclides are ^{239}Pu, ^{3}H, and some long-lived fission products such as ^{90}Sr and ^{137}Cs. For example, 5.6 tons of plutonium has been injected into the environment by weapons testing. Over 99% of this material is now in the soil and sediments. Pu concentrations in water are low, being

$\sim 10^{-18}$ M. Approximately 0.1% of the plutonium released to the environment ends up in the biota (Seaborg and Loveland, 1990).

11.2.3
Artificial Radioactivity

In addition to the naturally occurring radionuclides, one can make a large number of shorter-lived radionuclides in the laboratory. Stable nuclei are transmuted into radioactive nuclei through nuclear reactions. This process was first demonstrated by Joliot and Curie (1934). Since then, the growth in the number of artificial radionuclides has been phenomenal. In Figure 11.1, one notes that the shaded areas corresponding to artificial radionuclides include over 1500 species ranging in half-life from microseconds to millions of years. One also notes the large blank areas corresponding to the 5000 or more possible radionuclides, most of which are waiting to be discovered.

11.2.4
The Elements beyond Uranium

The most important examples of artificial radioactivity are the elements beyond uranium, that is, those elements with $Z > 92$. At present, there are 25 such man-made elements, ranging in Z from 93 to 118. In Table 11.1, we list the names of these elements and the nuclear reactions used to synthesize them.

In total, there are about 260 transuranium nuclei known as of 2005. These nuclei are all radioactive, with half-lives ranging from $\sim 10^{-9}$ seconds to 10^{17} seconds. As Z increases, the half-lives of the elements become shorter, in general. All these nuclei are unstable with respect to α-decay (positive Q_α). Those with n/p ratios differing from the nuclei along the valley of β-stability decay by either β^- or EC decay. As the atomic number increases, so does the importance of spontaneous fission decay. Detailed descriptions of the nuclear and chemical properties of these elements exist (Seaborg and Loveland, 1990; Morss *et al.*, 2006). The story of the discovery of these elements is a fascinating one and the reader is referred elsewhere for details (Seaborg and Loveland, 1990; Hoffman, Ghiorso and Seaborg, 2000; Hofmann, 2002).

Up to 1970, it was thought that the practical limit of the periodic table would be reached at about element 108. By extrapolating the experimental data on heavy-element half-lives, we concluded that the half-lives of the longest lived isotopes of the heavy elements beyond about element 108 would be so short ($<10^{-6}$ seconds) due to spontaneous fission decay, that we could not produce and study them. However, in the late 1960s and early 1970s, nuclear theorists predicted that special stability against fission would be associated with proton number $Z = 114$ and neutron number $N = 184$. These ‘‘superheavy elements’’ were predicted to have half-lives of the order of the age of the universe. They were predicted to form an ‘‘island’’ of stability separated from the ‘‘peninsula’’ of known nuclei.

We now know these predictions were wrong, in part. While we believe there are a group of ‘‘superheavy’’ nuclei whose half-lives are relatively long compared to lower Z elements, we do not believe they form an ‘‘island’’ of stability. Rather, we picture them as a continuation of the peninsula of known nuclei. We also believe that their half-lives are short compared to geologic timescales. Therefore, they do not exist in nature. While most current theoretical predictions attach special stability to $N = 184$, there is a division of opinion

Tab. 11.1 Summary of transuranium element synthesis.

Atomic number	Name and symbol	Synthesis reaction
93	Neptunium (Np)	$^{238}\mathrm{U} + \mathrm{n} \longrightarrow {}^{239}\mathrm{U} + \gamma$
		$^{239}\mathrm{U} \xrightarrow{\beta^-} {}^{239}\mathrm{Np}$
		$^{238}\mathrm{U} + {}^{2}\mathrm{H} \longrightarrow {}^{238}\mathrm{Np} + 2\mathrm{n}$
94	Plutonium (Pu)	$^{238}\mathrm{Np} \xrightarrow{\beta^-} {}^{238}\mathrm{Pu}$
95	Americium (Am)	$^{239}\mathrm{Pu} + \mathrm{n} \longrightarrow {}^{240}\mathrm{Pu} + \gamma$
		$^{240}\mathrm{Pu} + \mathrm{n} \longrightarrow {}^{241}\mathrm{Pu} + \gamma$
		$^{241}\mathrm{Pu} \xrightarrow{\beta^-} {}^{241}\mathrm{Am}$
96	Curium (Cm)	$^{239}\mathrm{Pu} + {}^{4}\mathrm{He} \longrightarrow {}^{242}\mathrm{Cm} + \mathrm{n}$
97	Berkelium (Bk)	$^{241}\mathrm{Am} + {}^{4}\mathrm{He} \longrightarrow {}^{243}\mathrm{Bk} + 2\mathrm{n}$
98	Californium (Cf)	$^{242}\mathrm{Cm} + {}^{4}\mathrm{He} \longrightarrow {}^{245}\mathrm{Cf} + \mathrm{n}$
99	Einsteinium (Es)	"Mike" thermonuclear explosion (leading to $^{253}\mathrm{Es}$)
100	Fermium (Fm)	"Mike" thermonuclear explosion (leading to $^{255}\mathrm{Fm}$)
101	Mendelevium (Md)	$^{253}\mathrm{Es} + {}^{4}\mathrm{He} \longrightarrow {}^{256}\mathrm{Md} + \mathrm{n}$
102	Nobelium (No)	$^{244}\mathrm{Cm} + {}^{12}\mathrm{C} \longrightarrow {}^{252}\mathrm{No} + 4\mathrm{n}$
103	Lawrencium (Lr)	$^{250,251,252}\mathrm{Cf} + {}^{11}\mathrm{B} \longrightarrow {}^{258}\mathrm{Lr} + 3\text{–}5\mathrm{n}$
		$^{250,251,252}\mathrm{Cf} + {}^{10}\mathrm{B} \longrightarrow {}^{258}\mathrm{Lr} + 2\text{-}4\mathrm{n}$
104	Rutherfordium (Rf)	$^{249}\mathrm{Cf} + {}^{12}\mathrm{C} \longrightarrow {}^{257}\mathrm{Rf} + 4\mathrm{n} + {}^{13}\mathrm{C} \longrightarrow {}^{259}\mathrm{Rf} + 3\mathrm{n}$
105	Dubnium (Db)	$^{249}\mathrm{Cf} + {}^{15}\mathrm{N} \longrightarrow {}^{260}\mathrm{Db} + 4\mathrm{n}$
106	Seaborgium(Sg)	$^{249}\mathrm{Cf} + {}^{18}\mathrm{O} \longrightarrow {}^{263}\mathrm{Sg} + 4\mathrm{n}$
107	Bohrium (Bh)	$^{209}\mathrm{Bi} + {}^{54}\mathrm{Cr} \longrightarrow {}^{262}\mathrm{Bh} + \mathrm{n}$
108	Hassium (Hs)	$^{208}\mathrm{Pb} + {}^{58}\mathrm{Fe} \longrightarrow {}^{265}\mathrm{Hs} + \mathrm{n}$
109	Meitnerium (Mt)	$^{209}\mathrm{Bi} + {}^{58}\mathrm{Fe} \longrightarrow {}^{266}\mathrm{Mt} + \mathrm{n}$
110	Darmstadtium (Ds)	$^{209}\mathrm{Bi} + {}^{59}\mathrm{Co} \longrightarrow {}^{267}\mathrm{Ds} + \mathrm{n}$ $^{208}\mathrm{Pb} + {}^{62}\mathrm{Ni} \longrightarrow {}^{269}\mathrm{Ds} + \mathrm{n}$ $^{208}\mathrm{Pb} + {}^{64}\mathrm{Ni} \longrightarrow {}^{271}\mathrm{Ds} + \mathrm{n}$ $^{244}\mathrm{Pu} + {}^{34}\mathrm{S} \longrightarrow {}^{273}\mathrm{Ds} + 5\mathrm{n}$
111	Roentgenium (Rg)	$^{209}\mathrm{Bi} + {}^{64}\mathrm{Ni} \longrightarrow {}^{272}\mathrm{Rg} + \mathrm{n}$
112	-	$^{208}\mathrm{Pb} + {}^{70}\mathrm{Zn} \longrightarrow {}^{277}112 + \mathrm{n}$
113[a]	-	$^{209}\mathrm{Bi} + {}^{70}\mathrm{Zn} \longrightarrow {}^{278}113 + \mathrm{n}\ {}^{237}\mathrm{Np} + {}^{48}\mathrm{Ca} \longrightarrow {}^{282}113 + 3\mathrm{n}$
114	-	$^{242}\mathrm{Pu} + {}^{48}\mathrm{Ca} \longrightarrow {}^{287}114 + 3\mathrm{n}\ {}^{244}\mathrm{Pu} + {}^{48}\mathrm{Ca} \longrightarrow {}^{288}114 + 4\mathrm{n}$
115[a]	-	$^{243}\mathrm{Am} + {}^{48}\mathrm{Ca} \longrightarrow {}^{288}115 + 3\mathrm{n}$
116[a]	-	$^{245}\mathrm{Cm} + {}^{48}\mathrm{Ca} \longrightarrow {}^{290}116 + 3\mathrm{n}$
118[a]	-	$^{249}\mathrm{Cf} + {}^{48}\mathrm{Ca} \longrightarrow {}^{294}118 + 3\mathrm{n}$

[a] Unconfirmed.

as to what proton numbers have special stability, with some calculations indicating $Z = 114$ to be stable while others indicate the next proton magic number will be near $Z = 120$ or $Z = 126$. One thing is clear, however, and that is that the half-lives of he longest lived isotopes of a given Z do not continue their sharp decrease with increasing atomic number, observed for elements 93–102. Rather there is a leveling off in the observed half-lives with the reported half-lives of elements 112–118 being of the order of milliseconds to microseconds.

Synthesis reactions for heavy nuclei are divided into "cold" or "hot" fusion. Cold fusion reactions involve a heavier projectile (Ar–Kr) interacting with a Pb or Bi nucleus, where the excitation energy of the completely fused system is low (~13 MeV), giving high survival probabilities. Unfortunately, the fusion probability in such systems is low. Hot fusion reactions involve the use of lighter projectiles (^{11}B–^{48}Ca) interacting with actinide nuclei, giving a high fusion probability, but a high excitation energy (E*~30–50 MeV) with a resulting low survival probability. In Figure 11.4, we show the current cross sections for heavy-element synthesis using these two methods.

The current challenge in heavy-element science is to understand the apparent leveling off of the production cross sections using the hot fusion approach. Some attribute this trend to an increased survival probability for the nuclei as they approach the proton and neutron magic numbers.

11.3
Radioactive Decay Kinetics

11.3.1
Basic Decay Law

Radioactive decay is a first-order reaction, that is, the number of decays per unit time is directly proportional to the number of nuclei present. Thus, if $-dN/dt$ is the decay rate, we can say that

$$\frac{-dN}{dt} = \lambda N \qquad (1)$$

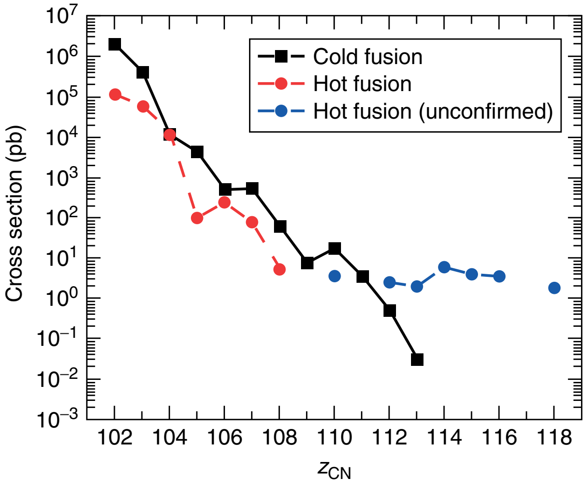

Fig. 11.4 Cross sections for heavy-element synthesis by hot and cold fusion reactions.

where λ, the decay constant, is the probability of decay per unit time and N is the number of radioactive nuclei present. The rate of disappearance of radioactive nuclei can easily be shown to be given by

$$N = N_0 e^{-\lambda t} \tag{2}$$

where N is the number of radioactive nuclei present at time t and N_0 the number of nuclei present at $t = 0$. A companion equation expresses the decrease in activity of a radioactive sample, $A(\equiv -dN/dt)$, as a function of time t, that is,

$$A = A_0 e^{-\lambda t} \tag{3}$$

where A_0 is the sample activity at time $t = 0$.

The half-life $t_{1/2}$ is defined as the time it takes the number of nuclei or the number of disintegrations per unit time to decrease by a factor of 2. A related quantity is the mean life τ, the average lifetime of a radioactive nucleus. Since the decay constant λ is the probability per unit time that a nucleus will decay, it is easy to show that τ equals $1/\lambda$. (The mean life τ is the appropriate quantity to be used in the energy–time statement of the uncertainty principle,

$$\Delta E \cdot \tau \geq \hbar \tag{4}$$

where ΔE is the uncertainty about the energy of a state while τ is its mean life.)

The natural unit of radioactivity is disintegrations per second. One *becquerel* (Bq) is defined as 1 disintegrations per second. An older unit that is still used is the curie (Ci), which corresponds to 3.7×10^{10} Bq. Most radioactive samples have activities ranging from nCi to μCi although irradiated reactor-fuel activities are frequently at the MCi level.

11.3.2
Radioactive Decay Equilibrium

When a nucleus decays, frequently a new nucleus appears that is also radioactive. This nucleus can decay to another, to another, and so on. For the general situation where species 1 decays to 2, which in turn decays to 3, and so on, one can show that

$$N_n = N_1^0 \sum_{i=1}^{n} C_i e^{-\lambda_i t} \tag{5}$$

where N_1^0 is the number of species 1 nuclei present initially and

$$C_i = \frac{\lambda_1 \lambda_2 \lambda_3 \ldots \lambda_{n-1}}{(\lambda_1 - \lambda_i)(\lambda_2 - \lambda_i)\ldots(\lambda_n - \lambda_i)} \tag{6}$$

and we omit the term in the denominator of Eq. (6) that is zero. Special cases of this general equation exist. One of the most important of these is secular equilibrium where $\lambda_1 < \lambda_2$. Here the parent is very long lived relative to the daughter. In this case, one can show that

$$\lambda_1 N_1 = \lambda_2 N_2 \tag{7}$$

or A_1 is equal to A_2. This condition applies to the heavy-element decay series discussed previously. There the half-lives of the parent nuclei are so long that the conditions for secular equilibrium are met. Consequently, the activities of all the members of a given series are equal. For further details, see Loveland, Morrissey and Seaborg (2006).

11.3.3
Dating

An important application of the basic radioactive decay law is that of dating. The basic decay law tells us that

$$P = P_0 e^{-\lambda t} \tag{8}$$

where P is the number of parent nuclei at time t and P_0 is the number at $t = 0$. Rearranging Eq. (8) we get

$$t = \frac{\ln(P_0/P)}{\lambda} \tag{9}$$

where t is the age of an object. The most famous application of Eq. (9) is that of *radiocarbon dating* (Libby, 1946; Arnold and Libby, 1949).

^{14}C ($t_{1/2} = 5730$ years) is formed continuously in the upper atmosphere by cosmic rays via the reaction $n + {}^{14}\mathrm{N} \rightarrow {}^{14}\mathrm{C} + p$, or, in a shorthand notation, $^{14}\mathrm{N}(n, p)^{14}\mathrm{C}$. This radiocarbon (^{14}C) exchanges with the stable carbon (mostly ^{12}C) in living things. If the cosmic-ray flux is constant, then a constant level of ^{14}C in all living things is found (corresponding to ∼1 atom of ^{14}C for every 10^{12} atoms of ^{12}C or about 14.9 disintegrations per minute · (g C)(P_0)). When an organism dies, it ceases to exchange its carbon atoms with the pool of radiocarbon, and its radiocarbon content decreases in accordance with Eq. (8). Measurement of the specific activity (disintegrations per minute · (g C)) of an old object allows one to calculate the age using Eq. (9). When matter has decayed for 10 or more half-lives of ^{14}C, it is no longer possible to measure directly the radioactivity content of an object. In that case, one uses *accelerator mass spectrometry* to count the number of ^{14}C atoms directly. An accelerator, such as a cyclotron or tandem Van de Graaff, is used as a mass spectrometer to separate the ^{14}C atoms from the more prevalent ^{12}C or ^{13}C. Using this technique, it has been possible to determine ages as long as 100 000 years (Kutschera and Paul, 1990).

Another variant of the dating technique is to measure the relative number of parent and daughter atoms in an object. Since every parent decay will lead to a daughter, the number of daughter atoms D at time t is

$$D = D_0 + (P_0 - P) = D_0 + P(e^{\lambda t} - 1) \tag{10}$$

where D_0 is the number of daughter nuclei present at $t = 0$. Solving for t gives

$$\begin{aligned} t &= \frac{1}{\lambda}\ln\frac{P + D - D_0}{P} \\ &= \frac{1}{\lambda}\ln\left(1 + \frac{D - D_0}{P}\right) \end{aligned} \tag{11}$$

Equations (10) and (11) assume that there is no gain or loss of either parent or daughter except by radioactive decay. If one can assume that $D_0 = 0$, then a measurement of the relative numbers of parent and daughter nuclei will give the age of the sample. (If $D_0 \neq 0$, then further assumptions can be made to deduce the age; see Wetherill, 1975). Examples of this technique involve $^{40}\mathrm{K}/^{40}\mathrm{Ar}$, $^{87}\mathrm{Rb}/^{87}\mathrm{Sr}$, and U/Pb dating.

11.4
α-Decay

α-Decay involves the spontaneous emission of a doubly charged ^{4}He nucleus (an α-particle) by another nucleus, resulting in a loss of two protons and two neutrons by the parent nucleus. The energy released in the decay, Q_α, is defined as

$$Q_\alpha = (M_A - M_{A-4} - M_\alpha)c^2 \tag{12}$$

where M_A, M_{A-4}, and M_α refer to the atomic masses of the parent nucleus, the daughter nucleus, and a ^{4}He atom and

c is the speed of light. The decay energy Q_α is shared between the α-particle ($T_\alpha = M_{A-4}Q_\alpha/M_A$) and the recoiling daughter nucleus ($T_{\text{recoil}} = M_\alpha Q_\alpha/M_{A-4}$), with a small amount being used to compensate for the change in the total electronic binding energies of the atoms ($\sim 65.3Z^{7/5}$–$80Z^{1/2}$ eV). Thus, for heavy nuclei, the α-particle carries away about 98% of the decay energy. The emitted α-particles are monoenergetic with α-particle energies ranging from 1.83 MeV (^{144}Nd) to 11.65 MeV (^{212}Po^{m}), typical energies being $\sim$6 MeV. The half-lives corresponding to these decays range from 3.0×10^{-7} seconds (^{212}Po^{m}) to 2.1×10^{15} years (^{144}Nd). This unusually strong dependence of the half-life on the α-particle kinetic energy was first noted by Geiger and Nuttall and found its explanation in the theory of α-decay developed independently by Gamow (1928) and Gurney and Condon (1928).

In this theory, we assume the α-particle exists as a preformed particle within the nucleus. In Figure 11.5, we show the potential energy $V(r)$ between the α-particle and the daughter nucleus as a function of their separation. At small distances, the two nuclei are bound by the attractive nuclear potential ($-V_0$), while at larger distances, the repulsive Coulomb force between the two nuclei defines the shape of the potential-energy function. The α-particle can be considered to be trapped in the nuclear potential well. Quantum mechanics allows the α-particle to "tunnel" through the potential-energy barriers, although it is a relatively improbable event. (e.g., for the decay of ^{238}U, $Q_\alpha = 4.27$ MeV, and $t_{1/2} \sim 4.5 \times 10^9$ years. The α-particle strikes the barrier about 10^{20} times per second for about 10^9 years before a decay occurs.) Because the probability of penetrating the barrier, P, is given by

$$P = e^{-2G} \tag{13}$$

where

$$G = \left(\frac{2m}{\hbar^2}\right)^{1/2} \int_{\text{barrier}} [V(r) - Q_\alpha]^{1/2} dr; \tag{14}$$

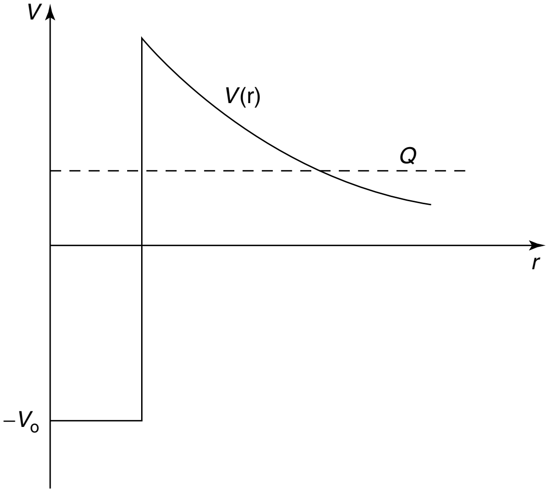

Fig. 11.5 The potential energy between an α-particle and a heavy daughter nucleus as a function of their separation. The horizontal line represents the decay energy, Q_α.

the penetrability of the barrier (i.e., the half-life for the decay) is a sensitive function of the decay energy Q_α. As a result of this, α-decay favors transitions to the ground and low-lying states of the daughter nuclei (large Q_α). From Eq. (14) one can see that $\log t_{1/2}^{\alpha}$ is a linear function of $1/Q_\alpha^{1/2}$ (the Geiger–Nuttall rule). The Geiger–Nuttall rule is used as the basis for semiempirical prescriptions relating the α-decay half-life, T_α, to Q_α, that is, the Viola–Seaborg formula (Viola and Seaborg, 1966)

$$\log_{10} T_\alpha = (aZ + b)Q_\alpha^{-1/2} + (cZ + d) + h_{\log} \tag{15}$$

where the parameters $a, b, c,$ and d are determined by fitting data for even–even nuclei and $h_{\log}$ is a hindrance factor for nuclei with unpaired nucleons. Modern versions of these semiempirical prescriptions exist (Dong and Ren, 2005; Hatsukawa, Nakahara and Hoffman, 1990) that allow one to predict T_α from Q_α to within a factor of ~1.5.

The decay of odd-A or odd–odd nuclei is hindered relative to the decay of even–even nuclei primarily by problems of the overlap (or lack of same) between the states of the parent and daughter nucleus when there are an odd number of nucleons. Typical hindrance factors, (measured $t_{1/2}$/calculated $t_{1/2}$), for even–even nuclei range from near unity to several thousands. These hindrance factors can give information about the parent and daughter states involved in the decay. Transitions with hindrance factors between 1 and 4 are called *favored* transitions; in these decays the parent and daughter states have similar wave functions.

Strict angular-momentum and parity-selection rules govern α-decay. The orbital angular momentum L_α carried away by the α-particle is given by

$$|I_i - I_f| \le L_\alpha \le I_i + I_f \tag{16}$$

where I_i and I_f are the total angular momenta of the parent and daughter states. The parity-selection rule can be written as

$$\frac{\pi_i}{\pi_f} = (-1)^{L_\alpha} \tag{17}$$

where π_i, π_f are the parities (−1 or +1) of the initial and final nuclear states. Thus, a $0 \to 3$ decay must have $L_\alpha = 3$ (Eq. (16)) and $0^+ \to 3-$ is allowed (Eq. (17)), but not $0^+ \to 3^+$. Also it should be noted that higher values of L_α are inhibited relative to lower values because of a "centrifugal potential" that raises the height of the barrier against α-decay and decreases its penetrability. A detailed discussion of α-decay is found in Rasmussen (1965).

11.5
β-Decay

11.5.1
Types of β-Decay

In nuclear β-decay, the total number of nucleons in the nucleus, A, remains constant while the number of neutrons and protons changes. The three types of β-decay are

$$\begin{aligned} n &\longrightarrow p + \beta^- + \overline{\upsilon}_e \,(\beta^-\text{decay}) \\ p &\longrightarrow n + \beta^+ + \upsilon_e \,(\beta^+\text{decay}) \\ e^- + p &\longrightarrow n + \upsilon_e \,(\text{EC decay}) \end{aligned} \tag{18}$$

In β^- decay, a nuclear neutron changes into a proton with the emission of an electron (designated β^-) and an electron antineutrino ($\overline{\upsilon_e}$). In β^+ (positron) decay, a nuclear proton changes into a neutron with the emission of a positron (β^+) and an electron neutrino (v_e). In EC decay (Alvarez, 1938), a nuclear proton changes into a neutron upon capture of an orbital atomic electron with emission of only an electron neutrino (v_e). The emitted electrons and neutrinos are created in the decay and do not exist within the nucleus. β^- decay is thus the mode of β decay for neutron-rich nuclei while EC and β^+ decay are competing decay modes for proton-rich nuclei. Examples of these decay modes are

$$\begin{aligned} {}^{14}_{6}\text{C} &\longrightarrow {}^{17}_{7}\text{N} + \beta^- + \overline{v}_e \\ &\qquad (\beta^-\text{decay}) \\ {}^{22}_{11}\text{Na} &\longrightarrow {}^{22}_{10}\text{Ne} + \beta^+ + v_e \\ &\qquad (\beta^+\text{decay}) \\ e^- + {}^{207}_{83}\text{ Bi} &\longrightarrow {}^{207}_{82}\text{Pb} + v_e \\ &\qquad (\text{electron capture}) \end{aligned} \tag{19}$$

All nuclei not lying in the valley of stability will undergo β-decay.

11.5.2
Energetics

The energy release in β-decay, Q_β, is shared among the products. In the case of EC decay, the decay energy Q_{EC} is given as

$$Q_{\text{EC}} = (M_Z^A - M_{Z-1}^A)\,c^2 - B \tag{20}$$

where M_Z^A, M_{Z-1}^A are the atomic masses of the parent and daughter species and B is the binding energy of the captured orbital electron. The recoiling daughter nucleus and the emitted neutrino share the decay energy with $T_v \approx Q_{\text{EC}}$. In β^- decay, the energy release Q_{β^-} is given as

$$Q_{\beta^-} = (M_Z^A - M_{Z+1}^A)\,c^2 \tag{21}$$

This discrete energy is shared among the recoiling daughter nucleus (a very small energy) and the two emitted particles, the electron (β^-) and the electron antineutrino ($\overline{v_e}$). Unlike EC decay, the energy of the emitted particles is variable, ranging from zero up to some maximum energy (Figure 11.6). Only the sum of the electron and antineutrino energies is a constant. On the average, about 40% of the decay energy appears as β^- energy while 60% appears as $\overline{v_e}$ energy. In β^+ decay, the decay energy Q_{β^+} is given as

$$Q_{\beta^+} = (M_Z^A - M_{Z-1}^A)c^2 - 2M_ec^2 \tag{22}$$

where M_e is the mass of the electron ($2M_ec^2 = 1.02$ MeV). The latter term in Eq. (22) reflects the need to supply additional energy to create a positron (β^+) in this decay. Thus, while EC and β^+ decay are competing modes of decay for proton-rich nuclei, EC decay is the only possible decay mode when the masses of parent and daughter nuclei differ by less than 1.02 MeV. Among the lighter elements, EC and β^+ decay are about equally probable. With increasing Z, EC decay is favored and positron emission is not observed in the heavy elements. The decay energy Q_{β^+} appears primarily as shared energy between the emitted positron and the electron neutrino (Figure 11.6).

11.5.3
Decay Mechanism

The measured half-lives for β-decay are long, varying from 10^{-3} seconds to 10^{16} years. Thus, we conclude that the processes to create simultaneously an electron and a neutrino are improbable. β-Decay must involve some sort of interaction between a nucleon, an electron, and a neutrino. This interaction is called the

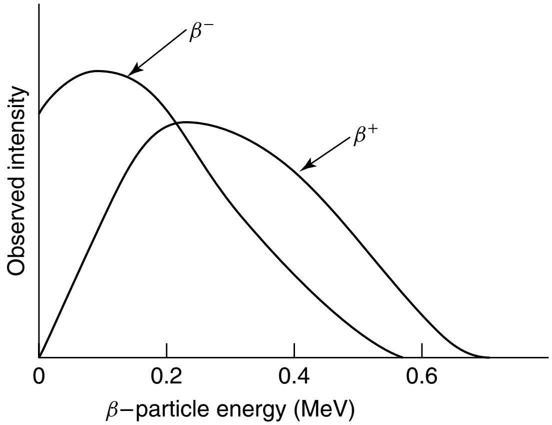

Fig. 11.6 The kinetic energy of the emitted electrons (positrons) in β^- (β^+) decay.

weak interaction and is different from the strong, electromagnetic, and gravitational forces. This weak interaction is responsible for the long half-lives associated with β-decay. The basic theory to describe β-decay was set forth by Fermi (1934). In addition to a constant g that expresses the strength of the weak interaction, analogous to the constants e and G that express the strengths of the electromagnetic and gravitational forces, the theory contains reference to the properties of the initial and final nuclear states in the decay. β-Decay will be enhanced when the initial and final states are similar. Formally, if the interaction between the initial and final states of the system is V, the decay probability will be proportional to the square of the matrix element

$$V_{\mathrm{f}i} = \int \psi_f^* V \psi_i dv \tag{23}$$

The available β-decay energy Q can be divided between the electron and the neutrino in an infinite number of ways, but these are not all equally probable. The probability of emission of a particle is proportional to the square of its momentum. Sharing the energy between two particles produces a most probable division of energy when each takes roughly half the energy. However, we also expect an attraction (repulsion) between the positively charged nucleus and the electron (positron). This effect has been reduced to a parameter $\mathrm{F}(Z, E_e)$, the Fermi function, that is tabulated. Thus, the decay constant for β-decay will be given as

$$\lambda(E_e) = (\text{constants})g^2 \times \int_0^{P_e^{\max}} |V_{\mathrm{f}i}|^2 \, \mathrm{F}(Z, E_e)(Q - E_e^2)P_e^2 dP_e \tag{24}$$

The term $((Q - E_e)^2$ is proportional to the square of the neutrino momentum. Equation (24) will describe the observed β^+, β^- spectra shown in Figure 11.6. If we assume that the matrix element does not depend on P_e (and we remove it from the integral) and we then designate the resulting integral as f, we have

$$\ln 2 = (\text{constants})g^2 |V_{fi}|^2 ft_{1/2} \tag{25}$$

or

$$ft_{1/2} = \frac{(\text{constants})}{|V_{fi}|^2} \tag{26}$$

Thus, the product $ft_{1/2}$ only depends on the square of the matrix element, that is, the similarity between initial and final nuclear states.

Transitions are classified by their change in angular momentum and parity. If the parities of the parent and daughter nuclei are the same and their angular momenta differ by one or less units, the transitions are termed *allowed*. If the spin difference is more than 1 and/or the parity changes, the transitions are "forbidden". There are different degrees of forbiddenness: first-forbidden, second-forbidden, and so on, depending on the spin change. For an increase of one degree of forbiddenness, the half-life increases by $\sim 10^3$. Values of $ft_{1/2}$ (or more conveniently, log $ft_{1/2}$) should have similar values for transitions involving similar changes in angular momentum and parity. Nomographs and semiempirical equations exist that allow one to calculate f for a given transition (Browne and Firestone, 1986). Measurement of the half-life $t_{1/2}$ along with the calculated value of f allows one to identify the angular momentum and parity change in a given β-decay. The spins of the electron and neutrino in β^- or β^+ decay can be parallel (Gamow–Teller decay) or antiparallel (Fermi decay). Individual transitions may take place by a mixture of Fermi and Gamow–Teller decays. The values of log $ft_{1/2}$ differ for Gamow–Teller and Fermi decays. A detailed discussion of β-decay is found in Wu and Moszkowski (1966).

11.5.4
Electron Capture

EC decay is the only β-decay mode for p-rich nuclei when the decay energy is <1.02 MeV and is the preferred decay mode for heavy p-rich nuclei. The captured orbital electron can come from any shell, but capture of K electrons is most probable because of their higher electron density in the nucleus. The theory of EC is similar to that of β^- or β^+ decay, except that it also involves the wave function of the captured electron, which determines its probability of being found within the nucleus. With increasing Z, the large positive charge of the nucleus causes the atomic electron shells to shrink, increasing the probability of electron capture. As mentioned previously, the only directly emitted radiation in EC decay is the electron neutrino. But the capture of an orbital electron leaves a hole in a low-lying orbital of the atomic electron configuration. The atomic electrons rearrange to fill this hole and in doing so, X-rays (especially in the heavy elements) or Auger electrons (especially in the light elements) are emitted. These extranuclear processes give rise to the useful experimental signatures of EC. The relative ratio of vacancies filled by X-ray emission or Auger-electron emission is given by the fluorescence yield or the Auger yield, which is the fraction of vacancies filled by X-ray emission or Auger-electron emission.

The rate of EC decay is sensitive to the atomic electron density at the center of the nucleus. Thus, the half-life for decay is sensitive to the chemical environment of the nucleus. Measurements of this effect in ^{7}Be, which decays only by EC, have shown a small but definite change in the half-life in different chemical environments (Johlige, Aumann and Born, 1970). In astrophysics, highly ionized nuclei occur and half-lives can be dramatically altered. In bare nuclei, EC decay can be switched off completely (^{52}Fe). In bare nuclei, internal conversion (IC) decay is also not possible.

11.5.5
Double β-Decay

There are many nuclei that are unstable with respect to the emission of two β^- particles simultaneously. Several cases are known of "two neutrino double β-decay ($\beta\beta(2\nu)$)" where two β^- particles and two electron antineutrinos are emitted, such as ^{48}Ca, ^{76}Ge, ^{82}Se, ^{96}Zr, ^{100}Mo, ^{116}Cd, ^{128,130}Te, ^{136}Xe, ^{150}Nd, and ^{238}U. This mode of decay is improbable with half-lives ranging from 10^{19}–10^{24} years (Elliott and Vogl, 2002). Of current interest is neutrinoless double β-decay, $\beta\beta(0\nu)$ where no neutrinos are emitted. This decay violates lepton conservation and is forbidden by the standard model. Searches for this mode of decay constrain the neutrino mass. Also of interest is another neutrinoless β-decay, $\beta\beta(0\nu, \chi)$ in which a light neutral boson, χ, is emitted. To date, neither of the neutrinoless decays has been observed.

11.5.6
β-Delayed Radioactive Decay

When β-decay occurs, the product nucleus can be formed in an excited state or its ground state. If the primary β-decay leads to an excited state at high enough energy, then a second decay process (such as neutron or proton emission or fission, etc.) can occur. Such processes are referred to collectively as *β-delayed radioactivities*. The most famous of these processes is β-delayed neutron emission, where the excitation energy of the daughter nucleus following β-decay is greater than the binding energy of the last neutron in that nucleus. This situation occurs frequently in *n*-rich fission products, where there are over 100 known examples of this decay mode, ranging in half-life from 0.2 to 55 seconds. Many of these nuclei are clustered above $N = 50$ and $N = 82$ because of the low values of the neutron-binding energies in the daughter nuclei. These β-delayed neutron emitters are of great technological importance in that the timescale of release of these neutrons is slow enough to allow servomechanisms controlling neutron absorbers to regulate the number of such neutrons in fissioning assemblies. This process makes possible the control of nuclear reactors. It is also possible to see β-delayed double neutron emission when the excitation energy of the daughter nucleus following β-decay is greater than the energy necessary to remove two neutrons. This phenomenon was first observed in ^{11}Be formed by the β-decay of ^{11}Li (Azuma *et al.*, 1979). In a similar manner, β-delayed triple neutron emission has also been reported for ^{11}Be (Azuma *et al.*, 1980).

When the excitation energy of the daughter nucleus following β^+ decay, E^*, is greater than the proton separation energy, β-delayed proton emission may occur. This process has been observed for a number of proton-rich nuclei. The emitted protons are monoenergetic with energies equal to $E^* - B_p$ where B_p is the proton-binding energy. Remembering that these protons have to overcome a Coulomb barrier like an α-particle to be emitted, one does not generally see delayed protons whose energy is <1 MeV. β-Delayed double proton emission has been observed in the decay (Cable *et al.*, 1983)

$$^{22}_{13}\mathrm{Al} \xrightarrow{\beta^+} {}^{22}_{12}\mathrm{Mg}^* \longrightarrow {}^{20}_{10}\mathrm{Ne} + 2p \tag{27}$$

The phenomenon of β-delayed radioactivity has a long history in nuclear science. Rutherford reported in 1916 the observation of β-delayed α-particle emission. In this case, ^{212}Bi is the parent nucleus (Figure 11.3). Numerous workers have also studied β-delayed fission where the excitation energy following a $\beta\pm$ or EC decay is large enough that the resulting nuclei undergo fission, either promptly or with some additional delay (Hall and Hoffman, 1992).

11.6
Electromagnetic Decay

11.6.1
Types and Classification

Most β- and α-decays that lead to an excited state of the daughter nucleus result in electromagnetic decay. Electromagnetic decay can take place by the nucleus emitting a photon (γ*-ray emission*) or by transferring energy to an orbital electron causing its ejection (*IC*). The result of these decay processes is to produce a lower-lying excited state or the ground state of the emitting nucleus. The emitted γ-ray photon is monoenergetic with an energy (E_γ) equal to the transition energy less a small amount of energy ($\sim 10^{-5} E_\gamma$) carried by the recoiling nucleus. The timescale for electromagnetic decay is usually 10^{-15} s. Occasionally, the emitting state will have a lifetime that is long enough to measure directly ($\sim 10^{-9}$ s). Such states are referred to as *metastable* or *isomeric* states. In such cases, the decay of the metastable state is inhibited, usually because of large spin change in its deexcitation. The γ-ray decays that depopulate isomeric states are called *isomeric transitions* (*ITs*).

The emission of γ-rays from nuclei resembles in many respects the emission of radio waves from an antenna. Both processes are due to the movement of charged particles: nucleons in the nuclear case, electrons in an antenna. In both cases, the radiation carries away energy and angular momentum. A photon can carry off any number of units of angular momentum, equal to or greater than 1 ħ. (This minimum is due to the spin of the photon, which is equal to 1 ħ.) Therefore, the initial and final nuclei must not both have zero spin, for then there would be no way to conserve angular momentum.

The greater the energy of the emitted photon, the greater the probability of γ-ray decay. The greater the angular momentum removed by the photon, the less probable γ-ray decay is. γ-Ray transitions are classified by transition type. The letters E and M are used to stand for electric and magnetic transitions; they are followed by a number equal to the number of units of angular momentum removed by the γ-ray. Thus, an E3 transition involves the loss of three units of angular momentum as a result of an electric transition. These rules are summarized in Table 11.2.

11.6.2
Transition Rates

The transition rates for γ-ray transitions are compared with those expected when a single proton makes a transition between nuclear states, as predicted by the shell model. These extreme single-particle estimates of transition rates are called the *Weisskopf estimates* after Weisskopf who first made them. These estimates are summarized in Table 11.3.

One can see that the half-lives for electric transitions are about 100 times shorter than for magnetic transitions of the same energy and multipolarity (angular momentum carried away). Each higher multipolarity for a given transition type and energy has a probability that is $\sim 10^7$ lower. Thus, γ-ray transitions usually occur by the lowest multipolarity transition possible. Quite often, γ-ray transition rates are expressed in Weisskopf units, that is, multiples of the

Tab. 11.2 γ-Ray emission rules.

Radiation type	Name	Angular momentum carried away	Does nuclear parity change?
E1	Electric dipole	1	Yes
M1	Magnetic dipole	1	No
E2	Electric quadrupole	2	No
M2	Magnetic quadrupole	2	Yes
E3	Electric octupole	3	Yes
M3	Magnetic octupole	3	No
E4	Electric 2^4 pole	4	No
M4	Magnetic 2^4 pole	4	Yes

estimates shown in Table 11.3. Most γ-ray transition rates are less than the Weisskopf estimates; the notable exception to this are E2 transitions, whose lifetimes are typically two orders of magnitude smaller than the Weisskopf estimate. (The E2 transitions are enhanced by a large-scale collective motion of several nucleons.) The result is that mixed M1–E2 transitions are common. A detailed discussion of electromagnetic decay is found in Siegbahn (1966).

11.6.3
Internal Conversion

Another form of nuclear electromagnetic decay is IC. In IC, the electric and magnetic disturbances (which cause the emission of a photon in γ-ray decay) interact directly with an orbital atomic electron, causing that electron to be ejected from the atom. The kinetic energy of the conversion electron is $E\gamma - E_B$, where E_γ is the transition energy and E_B the electron-binding energy. Another important quantity that characterizes IC is the IC coefficient α, which is defined as the (number of decays proceeding by IC)/(number of γ-ray decays). IC (large values of α) is favored by low transition energies and large spin changes in the decay. IC is favored in heavy nuclei since $\alpha \sim Z^3$. IC mostly involves ejection of K-shell electrons, since $\alpha \propto 1/n^3$ where n is the principal quantum number for the atomic electrons. Large tabulations of calculated IC coefficients exist. Measured values of α can be compared with the calculated values to determine the multipolarity of a given transition. Because no photon is emitted, E0 transitions are permitted to take place by IC.

11.7
Other Types of Radioactive Decay

11.7.1
Direct Proton Emission

Spontaneous proton emission is only energetically possible for the most proton-rich nuclei. Such nuclei can be formed in nuclear reactions with $N \approx Z$ projectile nuclei interacting with $N \approx Z$ target nuclei to create heavy nuclei with $N \approx Z$. The first example of this decay mode, observed in 1982, was ^{96}Ru(^{58}Ni,p2n)^{151}Lu.

The product ^{151}Lu was observed to decay by the emission of 1.23 MeV protons with $t_{1/2} \sim 0.085$ seconds (Hofmann *et al.*, 1982). Other cases have been observed since then.

Tab. 11.3 Gamma transition rates. The energies E are expressed in mega electronvolt. The nuclear radius has been taken as $1.3A^{1/3}$ fm and the A dependence of the transition rates has been incorporated in the formulas. It is to be noted that $t\gamma$ is the partial half-life for γ-emission only; the occurrence of internal conversion will always shorten the measured half-life.

Transition type	**Partial half-life t_γ (seconds)**	**Illustrative t_γ values for $A = 125$, $E = 0.1$ MeV (seconds)**
E1	$5.7 \times 10^{-15} E^{-3} A^{-2/3}$	2×10^{-13}
E2	$6.7 \times 10^{-9} E^{-5} A^{-4/3}$	1×10^{-6}
E3	$1.2 \times 10^{-2} E^{-7} A^{-2}$	8
E4	$3.4 \times 10^{4} E^{-9} A^{-8/3}$	9×10^{7}
E5	$1.3 \times 10^{11} E^{-11} A^{-10/3}$	1×10^{15}
M1	$2.2 \times 10^{-14} E^{-3}$	2×10^{-11}
M2	$2.6 \times 10^{-8} E^{-5} A^{-2/3}$	1×10^{-4}
M3	$4.9 \times 10^{-2} E^{-7} A^{-4/3}$	8×10^{2}
M4	$1.3 \times 10^{5} E^{-9} A^{-2}$	8×10^{9}
M5	$5.0 \times 10^{11} E^{-11} A^{-8/3}$	1×10^{17}

11.7.2
Heavy-Particle Radioactivity

In 1984, Rose and Jones reported the first example of a new mode of radioactive decay, decay by spontaneous emission of heavy particles ($Z \geq 6$), in the decay $^{223}\text{Ra} \rightarrow {}^{14}\text{C} + {}^{209}\text{Pb}$. The probability for this decay was 10^{-9} of the α-decay probability. Since that pioneering experiment, a number of other examples of heavy-particle radioactivity have been observed, such as ^{24}Ne and ^{28}Mg emission by nuclei with $Z \geq 88$ (allowing the daughter nucleus to be stabilized by the $Z = 82$ or $N = 126$ shell closures) (Price, 1989).

11.7.3
Spontaneous Fission

Spontaneous fission is a decay mode for the heaviest elements, increasing in importance with increasing atomic number. It is frequently a stability-limiting mode of decay for nuclei with $Z \geq 98$. The spontaneous fission half-lives change by a factor of 10^{29} in going from the longest-lived uranium isotopes ($t_{1/2} \sim 10^{19}$ years) to the shortest-lived isotopes of element 106 (∼milliseconds). Spontaneous fission is a barrier-penetration phenomenon with the nucleus "tunneling" from its ground state through a fission potential-energy barrier to the configuration of two touching fission fragments. The penetrability of this barrier depends on the parameter Z^2/A, which can be shown to be proportional to the ratio of the disruptive Coulomb force to the attractive nuclear force. Semiempirical relations exist that correlate the spontaneous fission half-life with Z^2/A of the fissioning system, although such predictions of spontaneous fission half-lives are only good to within a factor of 10^3–10^5.

In the late 1960s and early 1970s, there was considerable interest in another type of spontaneously fissioning nuclei, the spontaneously fissioning isomers. These were nuclei in excited states with relatively short half-lives ($\sim 10^{-9}$ s), much shorter than ordinary nuclei undergoing spontaneous fission, but that did not decay by γ-ray decay. We have since come to understand these nuclei as the first example of superdeformed nuclei, that is, ellipsoidal nuclei with a ratio of axes $\sim 2:1$. γ-Ray decay is inhibited by the large shape change involved in decaying back to the ground state, but spontaneous fission is enhanced.

11.8
Detection and Measurement of Radioactivity

11.8.1
Choice of Detector

Although the various types of nuclear radiation detectors are discussed, it seems useful to summarize the properties of these detectors in detecting radioactivity (Table 11.4).

The inverse of the resolving time (Table 11.4) is a rough measure of the maximum acceptable counting rate that a device can detect. For assay of α-emitting nuclides, the detector of choice is the semiconductor detector because of its superior energy resolution, allowing one to identify one nuclide in the presence of others. For detecting β^- emitters, a variety of devices is available, ranging from the inexpensive Geiger counter to the more sophisticated proportional counter. Occasionally, the superior energy resolution of a Si(Li) semiconductor detector is needed. The detection of the X-rays accompanying EC or IC can be done with either a proportional counter or a Si(Li) semiconductor detector. The detection of the two 0.511-MeV γ rays emitted when a β^+ annihilates is usually done with NaI(Tl) solid scintillation detectors. Most modern measurements of γ-rays from radioactive decay are made with Ge semiconductor radiation detectors. Liquid scintillation counting is the method of choice for assaying the low-energy β^- emitters, ^{3}H, and ^{14}C, as well as other nuclides frequently encountered in biological research (see (Knoll, 1989) or (Wang, Willis and Loveland, 1975), for details).

11.8.2
Sample Preparation

As important as the choice of a proper detector for radioactivity is, it is also important to prepare a sample properly for counting. Table 11.5 summarizes the sample preparation methods for α- and β-counting. No special sample preparation is needed for γ-ray counting or spectroscopy.

The preparation of samples for liquid scintillation counting is usually the most important part of the assay.

11.8.3
Correction Factors and Uncertainties

The results of a measurement of radioactivity frequently have to be corrected for a number of factors. The first of these is *background counts*, due to cosmic radiation, natural radioactivity in the vicinity of the detector, noise in the electronics, and so on. In some cases, where absolute activity measurements are being made or where a number of samples of differing size, shape, activity, and so on, are being assayed, a straightforward correction has to be made for the *solid angle* subtended by the sample with respect to the detector. The intrinsic *efficiency* of the detector for the radiation of interest must be evaluated.

Tab. 11.4 Summary of radiation detector characteristics.

Detector type	Energy discrimination	Detection medium	Reproducibility	Detector amplification factor	Relative electronic amplification required	Resolving time (microseconds)	relative background	Relative detection efficiency			Most commonly used sample form	Relative degree of difficulty in preparing sample
								α	β	γ		
Ion chamber	Yes	Gas	Excellent	1	Very high	Not applicable	Low	Good	Good	Poor	Gas	High[a]
Proportional detector	Yes	Gas	Good	10^2-10^4	High	1–2	Low	Good	Good	Poor	Solid	Medium[b]
Geiger-Müller detector	No	Gas	Fair	10^7	Low	100–300	Medium	Good	Good	Poor	Solid	Medium[b]
Solid scintillation detector	Yes	Solid	Good	10^6	Low	0.3 to >1	High	Fair	Fair	Very good	Solid or liquid	Medium[b] or low[c]
Semiconductor detector	Yes	Solid	Good	1	High	0.05	Low	Very good	Fair	Fair	Solid ($\alpha\beta\gamma$) Liquid (γ)	Low
Liquid scintillation detector	Yes	Liquid	Fair	10^6	Medium	0.05 to <1	Low	Very good	Very good	Fair	Solution	Low

[a] Usually involves individual combustion of samples to gaseous form.
[b] Usually involves preparation of planchet-mounted samples.
[c] Where well-type detector is used for liquid samples

Tab. 11.5 Sample preparation required for α- and β-assays.

Original sample form	Detector	Sample activity[a]	Sample preparation	Sample preparation time
(a) Sample preparation for α-assay				
Gas	Ionization chamber	Low–high	Count directly in gaseous form by introduction into detector in suitable mixture with counting gas	Moderate
Liquid	Semiconductor	Low–high	Convert to solid thin source and assay	Moderate–long
Liquid	Liquid scintillation counter	Low–medium	Dissolve or suspend in suitable scintillator-solvent mixture and assay	Short
Solid	Semiconductor	Low–high	If sample is very thin source, count directly. If not, dissolve sample and redeposit as thin source.	Moderate–long
Solid	Liquid scintillation counter	Low–medium	Dissolve in suitable solvent and treat as liquid, or count directly as a suspension in a gel mixture	Short–moderate
Solid	Nuclear track detectors	Low	Same as semiconductor detector	Moderate–long
(b) Sample preparation for β-assay				
Gas	Geiger–Müller	Low–medium	Sample (1) trapped in suitable solution and converted to solid form, or (2) may be counted directly in gaseous form by introduction into suitable detector chamber with counting gas	Moderate

(*continued overleaf*)

Tab. 11.5 (*continued*).

Original sample form	Detector	Sample activity[a]	Sample preparation	Sample preparation time
	Proportional counter	Low–high		
Gas	Semiconductor	Low–high	Trap sample in solution and convert to thin solid source	Long
	Solid scintillation detector	Medium–high		
Gas	Liquid scintillation counter	Low–medium	Must be trapped in suitable solution and treated as liquid sample	Moderate–long
Liquid	Geiger–Müller	Low–medium	Sample (1) preferably converted to solid form for counting, or (2) may be counted directly with suitable detector if particle energy is sufficiently high	Moderate–long
	Proportional counter	Low–high		
Liquid	Semiconductor	Low–high	Same as Geiger–Müller counter	Moderate–long
	Solid scintillation detector	Medium–high		
Liquid	Liquid scintillation counter	Low–medium	May be dissolved or suspended in a suitable scintillation solvent mixture and assayed directly without pretreatment	Short
Solid	Geiger–Müller	Low–medium	May be counted directly unless low energy of particles requires pretreatment of sample.	Short
	Proportional counter	Low–high		

(*continued overleaf*)

Tab. 11.5 (*continued*).

Original sample form	Detector	Sample activity[a]	Sample preparation	Sample preparation time
Solid	Semiconductor	Low–high	Same as Geiger–Müller counter	Short
	Solid scintillation counter	Medium–high		
Solid	Liquid scintillation counter	Low–medium	(1) Should be dissolved in suitable solvent and treated as liquid sample, or (2) may be counted directly as suspension in a suitable gel mixture, or (3), in the case of ^{3}H and ^{14}C, samples can be burned to H_2O and CO_2, which can be dissolved or trapped as in (1)	Moderate–long

[a] Low $\approx <$ 1 pCi; medium $\approx$ 1μCi; high $\approx$ >1 mCi.

Loss of counts due to *absorption* of the radiation by the sample, detector windows, and so on, must be made, as well as correction for *electronic losses* due to the finite resolving time of the detector and electronics. In β^- assay, corrections must also be made for *backscattered* radiation.

Radioactive decay is a random process and as such, consideration of the statistical nature of radioactive decay is needed to extract the maximum amount of correct information from any measurement. In Table 11.6, we summarize some pocket formulas for frequently encountered statistical quantities in radioactivity measurement (Wang, Willis and Loveland, 1975).

11.9
Applications of Radioactivity

The applications of radioactivity and radiation to medicine, industry, and science are so numerous as to defy counting. To give the reader the flavor of some of these applications, we cite the following examples.

11.9.1
Tracers

In most phenomena, the isotopes of a given element behave similarly. Thus, one can replace a stable isotope in a compound or atom by a radioactive isotope, and expect similar behavior. By using the superior sensitivity of detecting radioactivity (where one atom gives rise to one disintegration), one can measure the path, location, and so on, of a radioactive species in a system and infer the behavior of a stable species. Applications of this type have been used for decades in the study of chemical reactions. For example, consider the cyclization of 2-phenoxyacetophenone (I) to 2-phenylbenzofuran (II). Two possible mechanisms are

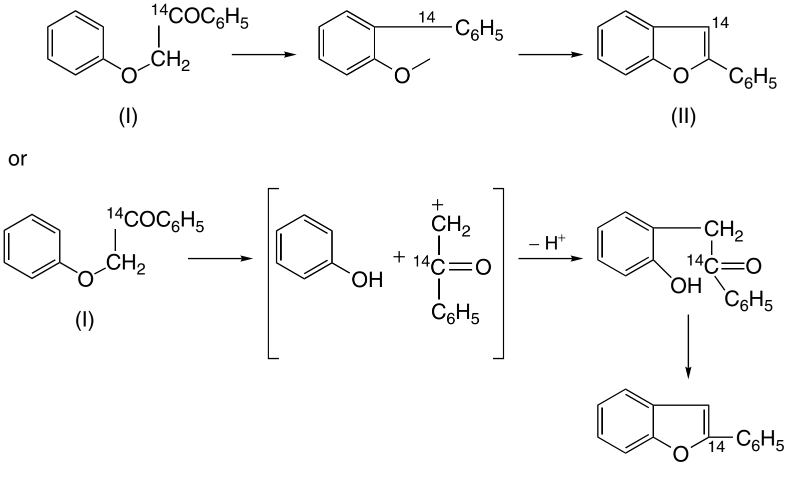

Note that the two mechanisms can be distinguished by the position of the labeled atom in 2-phenylbenzofuran (Wang, Willis and Loveland, 1975).

11.9.2
Thickness Gauges

In industry, one common use of radioactivity is in thickness gauges. The intensity of β^- radiation transmitted through a sheet of material can be used to measure its thickness. The number of β^- particles, I (from a radioactive source), that pass through a sheet of thickness x is given by

$$I = I_0 e^{-\mu x} \tag{29}$$

where I_0 is the β-intensity with no absorber and μ is a characteristic of the material whose thickness is being measured.

11.9.3
Nuclear Medicine

The application of radionuclides in both diagnosis and therapy is widespread. Radiotracers such as ^{123}I, ^{131}I, or ^{99}Tc^{m} can be caused to be taken up by an

Tab. 11.6 Frequently encountered quantities in radioactivity measurements.

Uncertainty in a single measurement of an activity, x	$\sigma_x = \sqrt{\text{no. of counts}} \equiv x^{1/2}$
Uncertainty in the average counting rate, $R = x/t$	$\sigma_R = \sqrt{R/t}$
Optimum ratio of time spent counting sample and background, t_{s+b}, to time spent counting background, t_b	$\frac{t_{s+b}}{t_b} = \sqrt{\frac{R_{s+b}}{R_b}}$
Distribution of intervals between counts (i.e., most probable interval is 0)	$I(t) = \lambda e^{-\lambda t}$
Figure of merit (FOM) for judging relative assays of sample with disintegration rate D using a detector with efficiency E and background B	$\text{FOM} = \frac{DE}{(DE+2B)^{1/2}}$

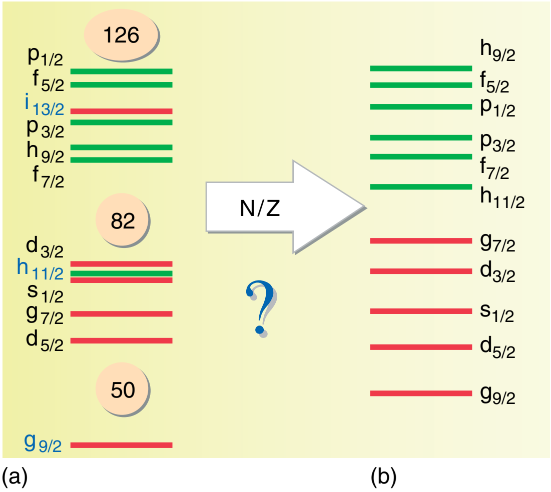

Fig. 11.7 (a) Nuclear shell structure in spherical nuclei close to the valley of stability. (b) The neutron shell structure predicted for very neutron-rich nuclei. Note the disappearance of the nuclear shell gaps due to the shallow binding potentials in the *n*-rich nuclei and the reduced spin-orbit coupling.

organ of interest. The most commonly used radionuclide is 6-hour $^{99}\mathrm{Tc}^{\mathrm{m}}$. It is used in over 10 million procedures in the U.S. each year and 20 million procedures in the world per year. $^{99}\mathrm{Tc}^{\mathrm{m}}$ is produced by the decay of $^{99}\mathrm{Mo}$. "Cows" (stock solutions) of $^{99}\mathrm{Mo}$ are "milked" to produce $^{99}\mathrm{Tc}^{\mathrm{m}}$, which will decay quickly leaving no residual radioactivity in the body. Imaging techniques can then be used to study the function of an organ or the location of a tumor. A widely used imaging technique is positron emission tomography (PET). Positron-emitting nuclides, such as $^{18}\mathrm{F}$, $^{11}\mathrm{C}$, $^{13}\mathrm{N}$, and $^{15}\mathrm{O}$, are introduced into a region to be studied. The two 0.511-MeV photons, produced when the β^+ annihilates, emerge in opposite directions and define a line passing through the point where the decay occurred. The two photons are detected by scintillation detectors placed outside the body. After observation of many decays, computer techniques are used to reconstruct a three-dimensional image of the area where the decays occurred (Krane, 1988).

11.9.4
Nuclear Science

The use of radioactivity has been a cornerstone in our studies of nuclear structure. In the past, the detection and assay of radioactivity in radiochemistry was used in studies of nuclear reactions while radioactive targets have been widely used in studies of both nuclear structure and reactions.

More recently, a great deal of attention has been paid to the use of radioactive nuclei as projectiles to induce nuclear reactions. These radioactive nuclear beams (RNBs) are also called *radioactive ion beams* (*RIBs*). The long-range plans for much of low-energy nuclear science have the use of these beams as the highest research priority.

Why the interest in RNBs? In Figure 11.1, we can see that there are about 1800 known stable and radioactive nuclei (the colored regions). But the large blank areas in Figure 11.1 correspond to some 5000 nuclei. Thus, we can say that most nuclei are unknown and unexplored at this time. As part of our ancient quest to explore the frontiers of knowledge, the study of the limits of existence of nuclei is important and RNBs are the vehicle to take us to these frontiers.

Among the nuclear structure issues that can be addressed are the location of the neutron and proton drip lines (the loci of nuclei where the Q values for neutron and proton emission become positive), and what combinations of n/p ratios are possible in bound nuclei. New neutron-rich heavy nuclei with longer half-lives are predicted to be produced with RNBs. It is anticipated that in very neutron-rich nuclei, the nuclear level spacings (shell gaps) will change leading to a washing out of the nuclear shells. (Figure 11.7)

In nuclear astrophysics, one hopes that RNBs will allow us to better understand the equation of state of neutron-rich matter such as that involved in supernovae and that we will learn more about the nuclear reactions involved in nucleosynthesis, in particular, the rapid neutron capture process (the r-process) involved in making the heavier nuclei, the rp process that makes proton-rich nuclei and the reactions that terminate the hot Carbon–Nitrogen–Oxygen (CNO) cycle.

Glossary

Atomic Number: The number of protons in the nucleus (Z).

Auger Electron: Orbital atomic electron ejected when an outer electron fills a vacancy in an inner electron shell and transfers its excess energy to another (Auger) electron, causing it to be ejected from the atom.

Even–Even Nuclide: A nucleus with an even number of neutrons and an even number of protons.

Isotopes: Nuclei with the same number of protons in the nucleus; for example, ${}^{12}_{6}\mathrm{C}$, ${}^{13}_{6}\mathrm{C}$, ${}^{14}_{6}\mathrm{C}$ are isotopes of carbon.

Mass Number: The total number of neutrons and protons in the nucleus, $A . A = N + Z$.

Neutrino/Nucleon: A neutron or a proton.

Nuclide: A specific nucleus with a given number of neutrons (N) and protons (Z).

Odd-A Nuclei: Nuclei with an odd number of particles in the nucleus; odd-A nuclei include those with even Z and odd N or odd Z and even N.

Odd–Odd Nuclei: Nuclei with both an odd number of neutrons and an odd number of protons in the nucleus.

Parity: A symmetry property of the nuclear wave function. If the signs of all the space coordinates of the wave function are changed, and the sign of the wave function does not, the system has even or $+$ parity. If the sign of the wave function does change, the system has odd or $-$ parity.

Radionuclide: A radioactive nuclide.

Secular Equilibrium: A limiting case of radioactive equilibrium in which $\lambda_1 \ll \lambda_2$ and in which the parent activity does not decrease measurably during many daughter half-lives.

Shell Closures: The complete filling of all states with a given shell (a nuclear or atomic configuration of special stability).

Strong Force (Interaction): The short-range force (interaction) responsible for nuclear binding; the attractive force between all nucleons.

References

Alvarez, L.W. (**1938**) *Phys. Rev.*, **53**, 606 .
Arnold, J.R. and Libby, W.F. (**1949**) *Science*, **110**, 678–80.

Azuma, R.E., Björnstad, T., Gustafsson, H.Å, Hansen, P.G., Jonson, B., Mattsson, S., Nyman, G., Poskanzer, A.M., and Ravn, H.L. (**1980**) *Phys. Lett.*, **96B**, 31–33.

Azuma, R.E., Carraz, L.C., Hansen, P.G., Jonson, B., Kratz, K.-L., Mattsson, S., Nyman, G., Ohm, H., Ravn, H.L., Schröder, A., and Ziegert, W. (**1979**) *Phys. Rev. Lett.*, **43**, 1652–54.

Browne, E. and Firestone, R.B. (**1986**) *Table of Radioactive Isotopes*, John Wiley & Sons, New York.

Cable, M.D., Honkanen, J., Parry, R.F., Zhou, S.H., Zhou, Z.Y., and Cerny, J. (**1983**) *Phys. Rev. Lett.*, **50**, 404–6.

Dong, T. and Ren, Z. (**2005**) *Eur. Phys. J. A*, **26**, 69–72.

Elliott, S.R. and Vogl, P. (**2002**) *Annu. Rev. Nucl. Part. Sci.*, **52**, 115–51.

Fermi, E. (**1934**) *Z. Phys.*, **88**, 161–77.

Gamow, G. (**1928**) *Z. Phys.*, **51**, 204–12.

Gurney, R.W. and Condon, E.U. (**1928**) *Nature*, **122**, 439–51.

Hall, H. and Hoffman, D.C. (**1992**) *Annu. Rev. Nucl. Part. Sci.*, **42**, 147–75.

Halliday, D., Resnick, R., and Krane, K.S. (**2001**) *Physics*, 5th edn, Vol. 2, (extended), John Wiley & Sons, New York.

Hatsukawa, Y., Nakahara, H., and Hoffman, D.C. (**1990**) *Phys. Rev. C*, **42**, 674–82.

Hofmann, S. (**2002**) *On Beyond Uranium*, CRC Press, Boca Raton.

Hofmann, S., Reisdorf, W., Münzenberg, G., Heßberger, F.P., Schneider, J.H.R., and Armbruster, P. (**1982**) *Z. Phys. A*, **305**, 111–23.

Hoffman, D.C., Ghiorso, A., and Seaborg, G.T. (**2000**) *The Transuranium People*, World Scientific Publishing Co. Ltd, Singapore.

Johlige, H.W., Aumann, D.C., and Born, H.-J. (**1970**) *Phys. Rev. C*, **2**, 1616–22.

Joliot, F. and Curie, I. (**1934**) *Nature*, **133**, 201–2.

Knoll, G.F. (**1989**) *Radiation Detection and Measurement*, John Wiley & Sons, New York.

Krane, K.S. (**1988**) *Introductory Nuclear Physics*, John Wiley & Sons, New York.

Kutschera, W. and Paul, M. (**1990**) *Annu. Rev. Nucl. Part. Sci.*, **40**, 422–38.

Lal, D. and Suess, H. (**1968**) *Annu. Rev. Nucl. Sci.*, **18**, 407–34.

Libby, W.F. (**1946**) *Phys. Rev.*, **69**, 671–72.

Loveland, W., Morrissey, D.J., and Seaborg, G.T. (**2006**) *Modern Nuclear Chemistry*, John Wiley & Sons, New York.

Morss, L.R., Edelstein, N.M., Fuger, J., and Katz, J.J. (**2006**) *The Chemistry of the Actinide and Transactinide Elements*, John Wiley & Sons, New York.

Price, P.B. (**1989**) *Annu. Rev. Nucl. Part. Sci.*, **39**, 19–42.

Rasmussen, J.O. (**1965**) in *Alpha, Beta and Gamma Ray Spectroscopy*, (ed. K. Siegbahn), Vol. I, Amsterdam, North-Holland, pp. 701–45.

Seaborg, G.T. and Loveland, W. (**1990**) *The Elements Beyond Uranium*, John Wiley & Sons, New York.

Siegbahn, K. (**1966**) *Alpha, Beta and Gamma Ray Spectroscopy*, North-Holland, Amsterdam.

Viola, V.E. and Seaborg, G.T. (**1966**) *J. Inorg. Nucl. Chem.*, **28**, 741–61.

Wang, C.H., Willis, D., and Loveland, W. (**1975**) *Radiotracer Methodology in the Biological, Environmental and Physical Sciences*, Prentice-Hall, Englewood Cliffs.

Wetherill, G.F. (**1975**) *Annu. Rev. Nucl. Sci.*, **25**, 283–328.

Wu, C.S. and Moszkowski, S.A. (**1966**) *Beta Decay*, John Wiley & Sons, New York.

Further Reading

Browne, E. and Firestone, R.B. (**1986**) *Table of Radioactive Isotopes*, John Wiley & Sons, New York. A tabulation of the detailed properties, such as half-lives, radiations, energies, and so on, of radioactive nuclei.

Evans, R.D. (**1982**) *The Atomic Nucleus*, Krieger, New York. A reprint of a 1955 classic book that remains a definitive source of information about radioactive decay kinetics.

Friedlander, G., Kennedy, J.W., Macias, E.S., and Miller, J.M. (**1981**) *Nuclear and Radiochemistry*, 3rd edn, John Wiley & Sons, New York. A "chemically oriented," somewhat out-of-date classic treatment, with unique information about the chemical manipulation of radioactivity.

Halliday, D., Resnick, R., and Krane, K.S. (**2001**) *Physics*, 5th edn, Vol. 2, (extended), John Wiley & Sons, New York. A new edition of an old classic with excellent introductory discussion of nuclear physics including radioactivity.

Harvey, B.G. (**1965**) *Nuclear Chemistry*, Prentice-Hall, Englewood Cliffs. A good,

simple discussion of the basic nature of radioactivity and its detection.

Knoll, G.F. (**2000**) *Radiation Detection and Measurement*, 3rd edn, John Wiley & Sons, New York. In-depth discussion.

Krane, K.S. (**1988**) *Introductory Nuclear Physics*, John Wiley & Sons, New York. One of the best, up-to-date, intermediate-level nuclear physics texts.

Loveland, W., Morrissey, D.J., and Seaborg, G.T. (**2006**) *Modern Nuclear Chemistry*, John Wiley & Sons, New York. Considered the successor to the Friedlander *et al*. textbook.

Seaborg, G.T. and Loveland, W. (**1990**) *The Elements Beyond Uranium*, John Wiley & Sons, New York. The radioactive properties of the heaviest elements.

Tuli, J.K. (**2005**) *Nuclear Wallet Cards*, National Nuclear Data Center, USDOE, Brookhaven. A free handy pocket-sized summary of nuclear decay properties, masses, etc. A product of the National Nuclear Data Center (*www.nndc.bnl.gov*) where up-to-date information may be obtained.

Wang, C.H., Willis, D.L., and Loveland, W.D. (**1975**) *Radiotracer Methodology in the Biological, Environmental and Physical Sciences*, Prentice-Hall, Englewood Cliffs. Details of radiotracers, liquid scintillation counting, and so on, for the biologically oriented and beginning students.

RADIOCHEMISTRY

J. Robb Grover, *Department of Chemistry, Brookhaven National Laboratory, Upton, New York, U.S.A.*

Introduction 565
1. History 566
2. Sources of Radioactive Material 566
2.1 Natural 566
2.2 Artificial 567
3. Radiochemical Technology 567
3.1 Radionuclide Production 567
3.1.1 Nuclear Reactors 567
3.1.2 Nuclear Accelerators 567
3.2 Operations with Radionuclides 568
3.2.1 Measurement 568
3.2.2 Separation and Purification 568
3.2.2.1 Wet-Chemistry Methods 568
3.2.2.2 Problems of Quantity 569
3.2.2.3 Fast and Ultrafast Methods 569
3.2.2.4 Methods without Separate Chemical Manipulation 570
3.2.3 Labeling 571
4. Applications 573
4.1 Analysis 573
4.1.1 Activation Analysis 573
4.1.1.1 Neutrons 573
4.1.1.2 Ion Beams 574
4.1.2 Isotope Dilution 574
4.1.3 Radiometric Techniques 574
4.2 Tracers 574
4.3 Isotopic Exchange and Substitution 574
4.3.1 Reaction Kinetics and Mechanisms 574
4.3.2 Electron-Exchange Reactions 575
4.3.3 Effects of Isotopic Substitution 575
4.4 Hot-Atom Chemistry 575
4.5 Artificial Elements 576
4.5.1 Observation of Chemical Properties When Only Tracer Amounts Are Available 576
4.5.2 Astatine as an Example 577
4.5.3 The Heaviest Elements 577
4.6 Joint Use of Radioactivity and Chemical Properties for Probes and Techniques 578
4.6.1 Positron Annihilation and Positronium 578
4.6.2 Muon Spin Rotation and Muonium 579
4.6.3 Perturbed Angular Correlations 579
4.6.4 Matching of Nuclear Lifetimes to Chemical Rates in Tracer Studies 580
4.7 Other Applications 582
Glossary 583
Works Cited 583
Further Reading 585

INTRODUCTION

Radiochemistry was once defined as all applications of chemistry and nuclear physics to each other. It has now come to be understood more narrowly as the methods and skills, plus the underlying basic science, concerned with the production, manipulation, and application of radioactive materials. By radioactivity is meant the spontaneous transformation of an atomic nucleus, where the energy released appears as photons and fast particles (see RADIOACTIVITY). Even with the more restricted definition, however, radiochemistry touches on a wide swath of areas in science and technology.

After a few brief remarks about the history of the development of radiochemistry, we describe the areas of technology and research that are encompassed by this field as it is presently understood. An account of where radioactivity is found or how it is produced is first presented. We then draw attention to the battery of techniques that have been developed for handling and measuring radioactive matter. Finally, a few of the many applications are described.

The following areas have much in com-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

mon with radiochemistry, but are not discussed in this article. Nuclear chemistry, once synonymous with radiochemistry, has evolved to mean principally the study and understanding of atomic nuclei, their reactions, and allied phenomena. It is now the same as nuclear physics, except that the investigators are employed in chemistry departments, and thus are more likely than physicists to utilize techniques that require chemical skills. Radiation chemistry, which began as a subfield of radiochemistry, is the study of the chemical effects of ionizing radiation on matter. The application of nuclear accelerators to chemistry has also become a well-established specialty, concerned with such matters as the study of channeling in crystals and the use of nuclear reactions for depth profiling.

1. HISTORY

The need to manipulate radioactive material chemically arose as soon as radioactivity was discovered, by Henri Becquerel, in 1896. Chemical operations were crucial, since at that time the major effort was, of necessity, concentrated on understanding the three radioactive decay chains associated with thorium and uranium. A radioactive decay chain is so called because an atom of one element makes a radioactive transformation into another element that is also radioactive, which gives rise to still another radioactive element, and so on. For thorium and uranium this phenomenon results in chains of up to 14 such transformations. Some of the scientists prominent in this earliest work were, in addition to Becquerel, William Crookes, Pierre and Marie Curie, Ernest Rutherford, and Frederick Soddy. The first synthesis of radioactive nuclides, by Irène and Frédéric Joliot-Curie, in 1933, began a tremendous expansion of the number of radioactive species demanding new chemical techniques, especially after the invention and development of nuclear accelerators in the years immediately following. Tracer applications of radioactivity, already originated by George Hevesy in 1913, were significantly extended by him when accelerator-produced activities became available. Then, the discovery of nuclear fission, by Otto Hahn and Fritz Strassmann in 1938, ushered in an entirely new era in radiochemistry, as well as in nuclear science, because it revealed a way to practical nuclear power. The subsequent wartime development of nuclear reactors and the atomic bomb vastly expanded the lexicon of techniques available for the chemical manipulation of radioactive matter and brought operations to an industrial scale. The number of practitioners burgeoned. The most noted radiochemist of that era was Glenn T. Seaborg, who, together with Edwin M. McMillan, Joseph W. Kennedy, and Arthur C. Wahl, had first synthesized and characterized plutonium. The most recent four decades have witnessed the development of powerful new instrumentation, the birth of new subdisciplines in radiochemistry, such as Mössbauer spectroscopy, and expansion into new areas such as environmental monitoring. Radiochemistry has now taken its place, along with scores of other chemical disciplines, such as x-ray crystallography or infrared spectroscopy, as a well-developed craft to be employed whenever it is the most appropriate approach to solving a problem (Rutherford *et al.*, 1930; Glasstone, 1950; Choppin and Rydberg, 1980).

2. SOURCES OF RADIOACTIVE MATERIAL

2.1 Natural

Fourteen of the 90 naturally occurring elements of the Earth's surface are radioactive because they possess isotopes having lifetimes for nuclear decay comparable to or longer than the time since the terrestrial elements were formed, about 4.6 billion years ago. Another seven, with much shorter lifetimes, belong to the decay chains of thorium or uranium, and owe their existence in nature to their membership in these chains because they possess no sufficiently long-lived isotopes to have survived since the Earth was formed. Radium is an example. In addition, sensitive measurements show that most of the elements on or near the surface of the Earth are slightly radioactive. This is caused by nuclear reactions with cosmic radiation. Also, the solar wind continuously contributes tritium, the radioactive isotope of hydrogen, to the top of the atmosphere.

2.2 Artificial

Radioactive nuclides, also called radionuclides, are usually synthesized by irradiation of the elements by neutrons, energetic ions, or photons. The most common techniques, those used in commercial production and at medical facilities, are irradiation with thermal neutrons and bombardment by hydrogen or helium ions accelerated to energies of a few million to a few tens of millions of electron volts. Thermal neutrons are available in nuclear reactors especially constructed to carry out such radioactivation, while for activation by ions small cyclotrons are customarily employed. An example of a commercial radionuclide is ^{99}Mo, which is used in medicine for diagnostic imaging. This material has a half-life, $t_{1/2}$, of 66.7 h, and is shipped packaged in kits from which its decay product ^{99m}Tc, $t_{1/2} = 6.0$ h, is readily extracted in isotonic saline solution ready for intravenous injection. Bulk radionuclides for commerce or research, such as tritium, technetium, and californium, are synthesized in nuclear power reactors or in reactors designed for such production. Radioactive nuclides for frontier research often cannot be prepared by the above methods, but must be generated in machines that feature ionic projectiles much heavier than helium, and bombarding energies much higher than simple cyclotrons can produce. Such machines include synchrocyclotrons, proton synchrotrons, and tandem Van de Graaf accelerators (see ACCELERATORS, LINEAR; PARTICLE SOURCES, ION; TARGETS FOR PARTICLE ACCELERATORS; ACCELERATORS, POTENTIAL DROP; CYCLOTRONS; SYNCHROTRONS AND STORAGE RINGS).

Radioactive isotopes are known for every element; more than 2200 radioactive nuclides with lifetimes of milliseconds or longer (including isomers) had been cataloged by 1986. For comparison, only 265 nuclides are stable (Browne *et al.*, 1986).

3. RADIOCHEMICAL TECHNOLOGY

3.1 Radionuclide Production

3.1.1 Nuclear Reactors Here, target nuclides are exposed to the reactor's neutrons, and the activated products are then recovered. The target, usually mounted in a small cylinder, is inserted into the region of high neutron flux via a long, cylindrical channel or "port." Thus, not only the target, but all of its containment structure, is subject to activation. Structural materials must therefore be chosen to minimize neutron absorption, both to avoid appreciably perturbing the neutron flux available to the target material, and to curtail unwanted radioactivity. Aluminum is often chosen, because its neutron absorption cross section is reasonably small, 0.23 b [1 b (barn) = 10^{-24} cm^2], and the ^{28}Al produced by neutron capture has a half-life of only 2.3 min. Its radioactivity therefore dies away quickly, which simplifies safe handling of the irradiated target. For short irradiations of liquids or materials incompatible with aluminum, plastic containers may be used. The plastic is made of carbon, hydrogen, nitrogen, and oxygen, which produce either nuclides that decay away rapidly, i.e., 7.2-s ^{16}N and 29-s ^{19}O, or the long-lived nuclides 12.3-y ^{3}H and 5730-y ^{14}C, radiation intensities from which build up only slowly. Irradiations must be kept short, however, to avoid destruction of the organic material. For more stringent conditions, especially for gases or liquids, sealed containers of fused silica may be used. One must then deal with 2.6-h ^{31}Si, for which waiting times of 12 h to a day for the target's radioactivity to subside (i.e., to "cool") are appropriate. Purity is important. Traces of copper in the aluminum will absorb neutrons rapidly to form 12.9-h ^{64}Cu, greatly increasing the cooling time. The boron in Pyrex glass would absorb so many neutrons, since the cross section of ^{10}B is more than 3800 b, that the flux available to the target would be seriously reduced. Aside from consideration of materials, however, targets are designed to serve a host of different applications, and so it is difficult to generalize. Some special problems are the use of extremes of low and high temperature, gas evolution and pressure buildup, self-shielding of target material when its neutron absorption cross sections are very large, degradation of fragile target materials, and investigation of activities of short lifetime.

3.1.2 Nuclear Accelerators Many radionuclides are best synthesized via the bombardment of suitable targets with energetic ions. The ion beams can either be contained within the accelerator, e.g., within the vacuum chamber of a cyclotron, or be extracted into a beam transport system, an array of evacuated tubes with focusing and guiding magnets. The latter permits much greater flexibility of target design, but higher beam intensities can be achieved with the former. As with the targets for reactors, the number of different applications defies classification, but there are some general considerations that most targets share. The target material to be exposed to the accelerator beam should be physically robust. The ion beams lose much of their energy in penetrating the targets, which must therefore survive temperature changes and stresses caused by measures taken for temperature control. In addition, the effects of severe radiation damage and of mechanical strain must be considered. Metal foils serve best. However, there are many elements for which metal foil targets are not feasible, and techniques have been developed for making stable, adherent coatings on metal foil backings. As with reactor irradiations, gaseous, and especially liquid, targets pose serious challenges. The latter become important when fast chemical processing must be carried out. (See PARTICLE SOURCES, ION; TARGETS FOR PARTICLE ACCELERATORS.)

3.2 Operations with Radionuclides

3.2.1 Measurement The most common modes of radioactive transformation are beta decay, alpha decay, deexcitation of nuclear excited states, and spontaneous fission (see RADIOACTIVITY). Beta decay may be accompanied by the creation and emission of either electrons or positrons (both are called beta particles), or by the capture of orbital electrons with concomitant emission of x rays or Auger electrons. Alpha decay is the emission of energetic ^{4}He nuclei (alpha particles). The product nucleus may be excited, and either emit photons (gamma rays) or expel orbital electrons (conversion electrons) accompanied by the resulting x rays and Auger electrons. These different kinds of emissions differ sharply in their properties, such as ease of deflection by a magnetic field, depth of penetration into matter, or density of ionization in liquids or gases, and so a variety of instruments that exploit these properties have been developed for carrying out the measurements. Some provide a relatively nondiscriminatory capability to observe the presence of radioactivity: ion chambers, proportional counters, Geiger counters. Others can resolve energy spectra of emitted particles (silicon surface barrier detectors, magnetic spectrometers), or of gamma rays and x rays [Ge(Li) detectors]. (See DETECTORS, SCINTILLATION; SPECTROMETERS, ELECTRON AND ION; SPECTROMETERS, GAMMA; PARTICLE DETECTORS, TRACKING; DETECTORS, SEMICONDUCTOR; RADIOACTIVITY.)

Amounts of radioactivity are usually measured via the intensity of emitted particles and photons. These emissions are always far more energetic than the laboratory thermal background, permitting the decays to be counted individually. Calibration of the instruments then allows the determination of absolute amounts of radioactivity, and thus of the absolute numbers of radioactive atoms via the equation

$$R = \lambda N, \quad \lambda = (\ln 2)/t_{1/2}, \qquad (1)$$

where R is the absolute radioactive decay rate of the sample in disintegrations per unit time, λ is the first-order rate constant, and N is the number of radioactive atoms. The standard units of radioactive decay rate are the *curie*, Ci, which is 3.7×10^{10} decays/s, and the *becquerel*, Bq, which is exactly one decay/s.

3.2.2 Separation and Purification

3.2.2.1 Wet-Chemistry Methods. Modern instruments obviate much of the burden once borne by chemical separations, but the latter remain indispensable. Commercial production of radionuclides and radiopharmaceuticals demands high standards of separation and purification, as do many problems in nuclear spectroscopy, assay, and tracer applications. The raw product from irradiations of targets by either neutrons or energetic ions usually contains a mixture of radioactive elements. Radiochemists process such mixtures in four steps:

1. dissolution—dissolve the target in a liquid, usually aqueous, and perhaps add some stable elements as carriers;
2. separation—carry out an appropriate

chemical separation of the desired element from the other radioactive elements;

3. purification—repeat the separation, perhaps several times, using different methods;
4. mounting—convert the isolated radioelement into the form required for the experiment.

Carrier may be added in steps **2** and **3** because when only a few atoms of an element are present, as often happens when the radioelement is different from the target material, its chemical behavior can differ from that of the bulk element (see Secs. 3.2.2.2 and 4.5.1).

Of the four steps listed above, we discuss only **2** and **3**. The main separation operations are precipitation, solvent extraction, chromatography, selective adsorption, distillation, and electrochemical deposition. Radioisotope isolation may require magnetic separation as well. The characteristics sought in these operations depend on the technical goal. For example, radioelement purity usually takes precedence over yield, or even bulk elemental purity. Sometimes carrier-free nuclides are needed, and this excludes certain manipulations. If a rapidly decaying nuclide is desired, the short half-life will dominate the separation strategy because only the fastest precipitations, solvent extractions, etc., can be used, and adaptability to automation may be demanded (see Sec. 3.2.2.3).

To exemplify the special characteristics exploited, consider precipitation. Precipitation can scrub impurities from a supernatant solution. For this, certain gelatinous or amorphous hydroxides, sulphides, or fluorides can absorb and remove many carrier-free impurity radionuclides. Conversely, the precipitate can reject radionuclide impurities. Here, one chooses a microcrystalline form for the precipitate, while holdback carriers, previously added for this purpose, are present in the supernate.

A decontamination factor $f_k(D,I)$ for purification method k can be measured for the desired element, D, and impurity element, I:

$$f_k(D,I) = (R^s_I/R^i_I)(R^i_D/R^s_D), \qquad (2)$$

where R^i_j is the initial disintegration rate of element j before the separation, and R^s_j is the corresponding disintegration rate remaining after the separation. To design a procedure to achieve a specified purity of D from I, one calculates the product of the decontamination factors for each separation step to be included.

A compilation of separation procedures, several for every element, is available as the series *Radiochemistry and Radiochemical Techniques* (see Further Reading).

3.2.2.2 Problems of Quantity. Most preparative irradiations yield very little material. For example, 1 μCi of ^{24}Na, $t_{1/2} = 15.0$ h, is equivalent to only 2.9×10^9 atoms, or 1.15×10^{-13} g of sodium. This amount of material in volumes of 1 to 100 mL of solution (10^{-11} to 10^{-14} molar) is very susceptible to surface effects. The radioactivity could be entirely adsorbed onto the vessel walls, or even onto the surface of a single dust particle. Impurities or minor constituents of a reagent, the effects of which would go unnoticed at ordinary concentrations, can become dominant, important if carrier-free nuclides are needed. Such considerations gave rise to a subfield called *tracer chemistry* (Bonner and Kahn, 1951).

At the opposite extreme, some applications require processing of kilogram amounts of radioactive nuclides, at megacurie levels of radioactivity. Mostly these involve extraction of elements, notably plutonium, from spent reactor fuel, which is fuel removed from reactors because the depletion of its fissionable component and the buildup of neutron-consuming fission products have curtailed its efficiency. Some of the special problems are self-heating of solutions, avoidance of critical assemblies of fissionable materials, radiation degradation of processing reagents, shielding of personnel from the intense radiation fields surrounding the material, and assuring that there is no escape of traces of radioactivity to contaminate the surrounding area.

The early stages of spent-fuel processing must be carried out in sealed cells behind massive radiation shielding, using mechanical manipulators. Certain processing methods, such as precipitation and filtration, are difficult to adapt to these remote operations, and so whenever possible solvent extraction and ion exchange are used instead (see NUCLEAR FUELS AND ISOTOPES; Benedict *et al.*, 1981; Choppin and Rydberg, 1980).

3.2.2.3 Fast and Ultrafast Methods. The ongoing effort to synthesize and study new nuclides makes it necessary to deal with species of ever shorter lifetimes. Although much can be learned via instrumental methods, chemical separations are often required—for example, to relieve congestion of gamma-ray lines in spectroscopic measurements, or to provide proof of elemental identity. Of the more than 2200 radioactive nuclides listed in 1986, about 800 have lifetimes of 1 min or less, and many more have been identified since then.

The problem falls naturally into three parts. First, transport the irradiated target from the reactor or accelerator. Second, carry out the chemical separation. Finally, transfer the product to the measuring instrument.

In pneumatic systems, air pressure forces a container containing the target from the site of activation into the laboratory within 0.1 to 0.2 s. The speeding target, liquid contained in a plastic capsule, smashes itself on the wall of the processing vessel, and the fast chemistry begins. In another approach, a solid target is permanently mounted at the site of irradiation, but is thin compared to the recoil range of the nuclear reaction products. Rapidly flowing gas containing aerosol particles—for example, submicroscopic clusters of sodium chloride suspended in helium—bathes the target surface. The newly produced radionuclides stop in the gas, diffuse to and are adsorbed on the aerosol particles, and then move with the flowing gas to the site of chemical separation. A heated plug of quartz wool set in the gas stream removes the activity from the aerosol and renders many of the radionuclides gaseous. Transport times in the millisecond range have been achieved. As an alternative to collection of reaction recoils, some elements will rapidly volatilize into the flowing gas from a strongly heated target.

Chemical processing may be either via a series of batches or continuous. Batch processing is appropriate for the smashed-capsule method described above. Such short times demand automated processing, and one seeks special properties of the elements under investigation that lend themselves to rapid separations in such devices. Examples are volatilization of hydrides (As, Se, Sn, Sb, Te, I, At) followed by suitable gas handling, adsorption of ions from aqueous solution onto preformed precipitates, and solvent extractions with quasisolid solvents in which the organic phase is fixed on the surfaces of a fine-grained inert carrier. These operations are combined into overall procedures, with computer control of the timing for each step. Figure 1 depicts such an apparatus. A serious drawback to batch techniques is the need to repeat the operations incessantly. Often each bombardment yields only a few events, so that hundreds to thousands of cycles must be carried out to build up adequate data.

Continuous processing offers a remedy to the disadvantages of batch processing, and also interfaces well with the gas flow transport method. Unfortunately, the necessary procedures are less well developed, because few exploitable techniques are known. Examples are thermochromatography, the use of differential sticking coefficients, and the high-speed minicentrifuge. We illustrate thermochromatography here, and an example utilizing differential sticking is presented in Sec. 4.6.4. The nuclides to be separated arrive from the accelerator adsorbed on aerosol particles and are rendered gaseous by encountering a plug of quartz wool at a high temperature in the presence of an injected reactant gas. The radioactive product gas then flows down a long tube that has a much higher temperature at the entrance than at the exit, with a smooth temperature gradient between. The radionuclides deposit on the tube walls in distinct zones, the location of the zone depending on the element and the compound in which it is bound. The tube could also be operated in vacuum or with inert carrier gases. Similarly the material of the tube's walls may be chosen to be inert or reactive. Figure 2 shows an example. For reviews, see Herrmann and Trautmann, 1982; Herrmann and Denschlag, 1969.

3.2.2.4 Methods without Separate Chemical Manipulation. Separations can sometimes be effected by ion-optic and dynamic techniques alone. One thereby gains a large advantage in speed. Most such experiments utilize an on-line isotope separator. For example, sodium isotopes can be continuously produced by GeV proton bombardment of a stack of thin thorium foils alternating with thin graphite slabs enclosed in a cylindrical tantalum foil, all maintained at 1800 °C. Recoiling sodium nuclei from fragmentation of the thorium stop in the graphite, and quickly diffuse from there into the gas phase, from

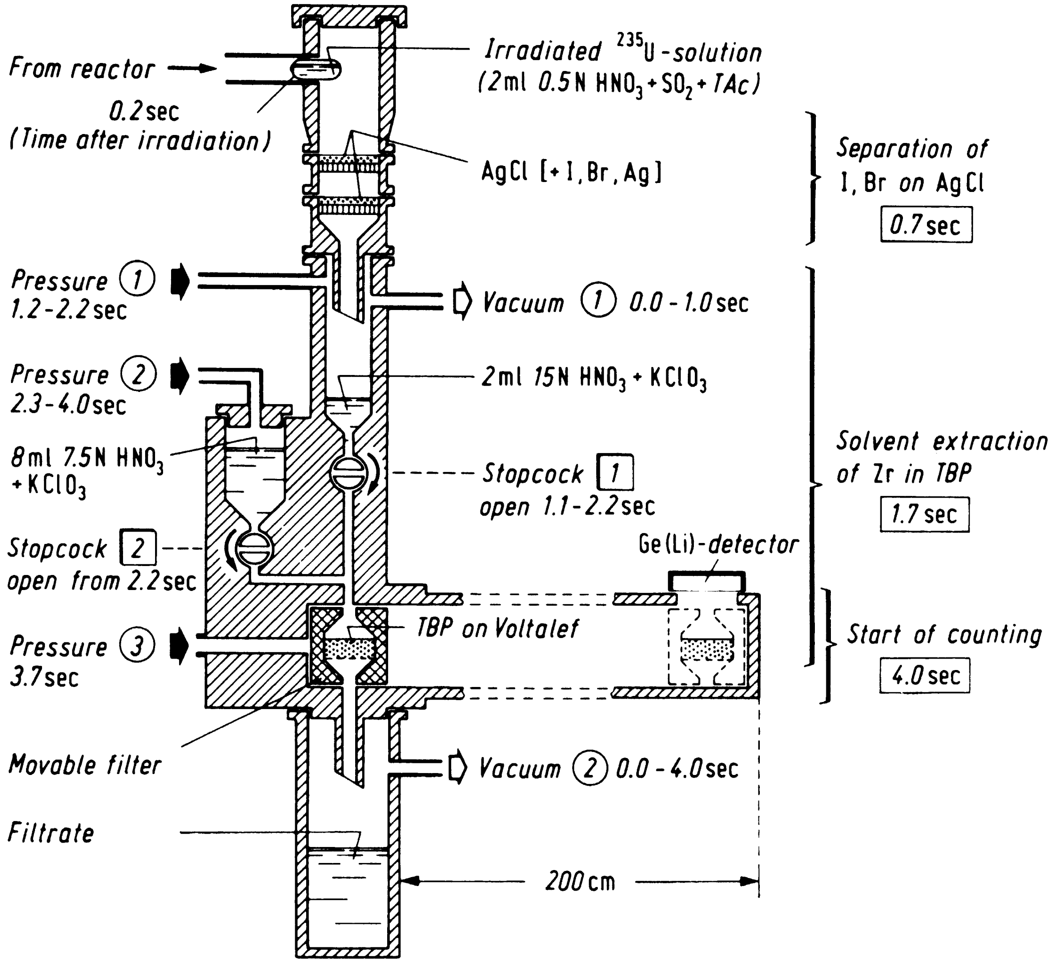
From reactor
0.2 sec
(Time after irradiation)
Irradiated ^{235}U-solution
(2 ml 0.5 N HNO_3 + SO_2 + TAc)
AgCl [+ I, Br, Ag]
Separation of
I, Br on AgCl
0.7 sec
Pressure ①
1.2-2.2 sec
Vacuum ① 0.0-1.0 sec
Pressure ②
2.3-4.0 sec
2 ml 15 N HNO_3 + $KClO_3$
8 ml 7.5 N HNO_3
+ $KClO_3$
Stopcock 1
open 1.1-2.2 sec
Solvent extraction
of Zr in TBP
1.7 sec
Stopcock 2
open from 2.2 sec
Ge(Li)-detector
Pressure ③
3.7 sec
TBP on Voltalef
Start of counting
4.0 sec
Movable filter
Vacuum ② 0.0-4.0 sec
Filtrate
200 cm

FIG. 1. Apparatus and time schedule for the automated rapid separation of zirconium from fission products. The capsule shoots in from the reactor and smashes against the opposite wall at the top. The encircled numbers on the left show the sequence in which nitrogen gas pressurizes the indicated sections. Similarly, those on the right show the pattern in which partial vacuum is applied. The numbers enclosed in squares give the sequence for manipulation of stopcocks. In the final step the zirconium is extracted into tri-*n*-butylphosphate (TBP) supported on a special powder, washed, and projected into a shielded position for gamma-ray spectroscopy (Trautmann *et al.*, 1972; reproduced by permission of Oldenbourg Verlag GmbH).

which they encounter the hot tantalum surface where they undergo surface ionization and reevaporation. Electric fields then accelerate and focus the sodium ions into a beam, which a magnetic field splits into several beams according to their mass-to-charge ratios, thus isolating the sodium isotopes. Finally, the isotopic beam of interest comes to rest in a beam stop positioned before a gamma-ray spectrometer or other instrument. In this way nuclides having lifetimes down to 0.01 s may be studied. Separation and study of a few elements other than the alkali metals can also be done with suitable target and ionizer design.

More direct techniques afford separation of nuclides, with respect to both charge and mass, as they recoil from nuclear reactions. Species with lifetimes shorter than a microsecond may thus be identified. However, the methodologies of these studies are outside our definition of radiochemistry and will not be discussed further (Hansen, 1979).

3.2.3 Labeling *Radioactive labeling* is the substitution of a radioactive isotope for a stable isotope in a molecule. Virtually all ra-

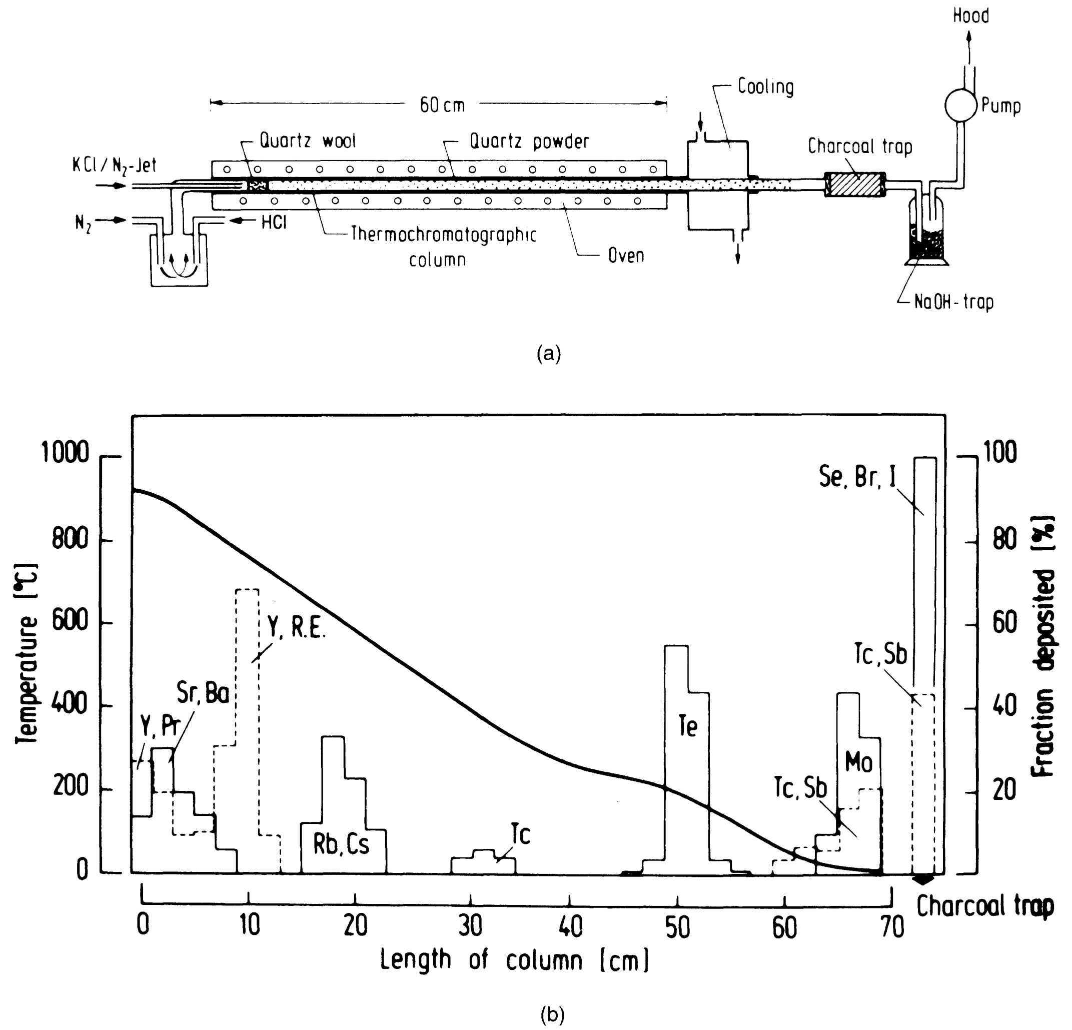
Hood
Pump
Cooling
60 cm
Quartz wool
Quartz powder
Charcoal trap
KCl / N₂-Jet
N₂
HCl
Thermochromatographic column
Oven
NaOH-trap
(a)
1000
800
600
400
200
0
Temperature [°C]
100
80
60
40
20
Fraction deposited [%]
Se, Br, I
Y, R.E.
Sr, Ba
Y, Pr
Rb, Cs
Tc
Te
Tc, Sb
Tc, Sb
Mo
0
10
20
30
40
50
60
70
Charcoal trap
Length of column [cm]
(b)

FIG. 2. (a) Diagram of a thermochromatographic system. Nitrogen gas, containing in suspension aerosol particles of KCl to which fission products are adsorbed, jets in on the left side. The analysis gas mixture indicated is nitrogen plus hydrogen chloride. (b) The distribution of separated fission product elements resulting from the use of 6% hydrogen chloride in nitrogen as the analysis gas (Hickmann *et al.*, 1980; reproduced by permission of Elsevier Science Publishers BV).

diochemical manipulations involve labeling, but for simple separation and purification of radionuclides trivial substitutions are chosen. However, labeling used in concert with tracer studies to elucidate reaction pathways can be very complex. It is an indispensable skill in research on radiopharmaceuticals. The challenges are to make the isotopic substitution cleanly in a specified position in a molecule, to achieve high specific activity, and to work with sufficient speed if the radioactivity is short lived. As a simple example, place a radiocarbon atom, C*, in either the 1- or 2- position in acetic acid. For acetic acid-2-C*,

$$\mathrm{C^{*}O_2(g)} \xrightarrow{\mathrm{H_2,catalyst}} \mathrm{C^{*}H_3OH} \xrightarrow{\mathrm{PI_3}} \mathrm{C^{*}H_3I}$$
$$\xrightarrow{\mathrm{KCN}} \mathrm{C^{*}H_3CN} \xrightarrow{\mathrm{H_2O}} \mathrm{C^{*}H_3CO_2H}.$$

For acetic acid-1-C*,

$$\mathrm{C^{*}O_2(g)} \xrightarrow{\mathrm{CH_3MgBr}} \mathrm{CH_3C^{*}O_2H}.$$

The decarbonylation of benzoylformic acid provides an example of the use of labeling:

$$C_6H_5\text{—CO—}^{14}COOH \xrightarrow{H_2SO_4} {}^{14}CO + C_6H_5\text{—COOH.}$$

Without a label, one cannot tell whether the carbon in the product carbon monoxide comes from the carbonyl group, —CO—, or from the carboxyl group, —COOH. With the labeling illustrated above, the ^{14}C tracer appears in the carbon monoxide, but not in the benzoic acid. When the carbonyl carbon is labeled instead, i.e., $C_6H_5\text{—}^{14}CO\text{—COOH}$, the ^{14}C appears in the benzoic acid but not in the carbon monoxide. This shows that the carbon monoxide arises from the carboxyl carbon (Fowler and Wolf, 1982; Murray and Williams, 1958).

High *specific activity,* i.e., a high intensity of radioactivity per unit of mass of labeled material, sometimes causes problems. The time such material can be stored can be quite limited because of radiation damage, or *radiolysis.* Some applications demand a high specific activity, however.

4. APPLICATIONS

Applications of radiochemistry appear in nearly every aspect of science and technology, but only a few examples can be mentioned here. Others are described in other articles in this Encyclopedia (see RADIOACTIVITY; MUONIC, MESONIC, AND BARYONIC ATOMS AND MOLECULES; MÖSSBAUER EFFECT; NUCLEAR FUELS AND ISOTOPES; NUCLEAR MEDICINE), or are listed in Sec. 4.7.

Most applications rely on three properties of radioactivity:

1. The disintegration rate of a sample of a given radionuclide is directly proportional to the number of its nuclei that are present.
2. The disintegration rate of a nuclide does not depend on its chemical form.
3. Chemical properties are not affected by nuclear instability.

On the other hand, violations of these properties, although slight or unusual, give rise to special applications and areas of investigation.

4.1 Analysis

The use of radioactivity is indispensable in analytical chemistry. This is now a vast area with numerous subfields, three of which are described below. The literature is voluminous. Fortunately, an exhaustive review appears biennially in a supplement to the *Journal of Analytical Chemistry* (Ehmann *et al.,* 1992; Ehmann and Vance, 1991; Tolgyessy and Kyrs, 1989).

4.1.1 Activation Analysis In this technique, radioactivity is induced in the material to be analyzed. The identity and quantities of the radionuclides produced reveal which elements are in the sample, and in what quantity. Most often the material to be analyzed is exposed to the transmuting radiation together with a sample of known composition for calibration. Usually resolved gamma and x rays, and their lifetimes, measured with Ge(Li) detectors, suffice for complete identification and determination of the elements present.

Important advantages of activation analysis are its great sensitivity, nondestructive character, applicability to nearly every element, capability for analysis of many elements simultaneously, capability for rapid results, and, with correct design, low unit cost. Disadvantages include insensitivity to chemical form, expense and awkwardness of provisions for radiation safety, and the need to choose among several techniques, since no single one is applicable to all elements (Alfassi, 1989).

Activation may be by neutrons, ion beams, or high-energy photon beams. The first two are the most useful, and are described below.

4.1.1.1 Neutrons. The usual method is to expose the sample to thermal neutrons in a nuclear reactor. This is most suitable when the induced radioactivity can be produced by neutron capture, i.e., (n,γ) reactions. However, sometimes the radioactivity needed requires energetic neutrons, via reactions such as $(n,2n)$, (n,α), or (n,p). A flux of circa 2.3-MeV neutrons is provided by the spontaneous fission of a sample of ^{252}Cf. Alternatively, a titanium tritide target can be irradiated with 100-keV deuterons from a small Cockroft–Walton accelerator to produce 14.7-MeV neutrons (see ACCELERATORS, POTENTIAL-DROP; PARTICLE SOURCES, ION; TARGETS FOR PARTICLE ACCELERATORS).

4.1.1.2 Ion Beams. Radioactivity induced by energetic charged particles allows for the analysis of elements that cannot be adequately analyzed by neutron irradiation—for example, boron, carbon, nitrogen, oxygen, and phosphorus. Small ions are most commonly used—protons, deuterons, and ^{3}He and ^{4}He nuclei—typically as beams 2 to 40 MeV in energy, produced by Van de Graaf electrostatic generators and cyclotrons (Vandecasteele, 1988; ACCELERATORS, POTENTIAL-DROP; PARTICLE SOURCES, ION; TARGETS FOR PARTICLE ACCELERATORS).

4.1.2 Isotope Dilution Let a sample that contains an unknown amount M_t of a given element be thoroughly mixed with a known, but much smaller mass, M_k, of the same element containing a radioactive isotope of specific activity S_k. Then it is only necessary to isolate a small fraction of the element from the mixture and measure its specific activity S_a to determine M_t:

$$M_t = (S_k/S_a - 1)M_k. \tag{3}$$

This technique is useful in a number of situations—for example, where quantitative isolation of the element is difficult or impossible, where other techniques would be much slower, or where the entire sample is unavailable, as in the blood of a living organism.

Recently, the most frequent application of isotope dilution analysis has been in radioimmunoassay (RIA) and related techniques. The unique specificity of antibodies and the high selectivity of RIAs has compelled widespread use of these methods to measure biologically active substances, such as drugs, enzymes, and hormones, in blood and other body fluids. (See NUCLEAR MEDICINE; Abraham, 1977.)

4.1.3 Radiometric Techniques Radioactive tracers can function as very sensitive indicators in titrations. One arranges a method where one of the reagents is radioactive and the product of the titration reaction goes into a different phase, i.e., precipitates as a solid, is extracted into a different liquid phase, is adsorbed into an ion-exchange medium, or is evolved as a gas. Depending on which reagent is radioactive, one can observe either the sharp appearance or disappearance of radioactivity in the sample to find the end point. These methods lend themselves well to automation, and also find use where end points are otherwise difficult to detect, such as when the sample is very dilute (Ehmann and Vance, 1991; Choppin and Rydberg, 1980).

4.2 Tracers

Perhaps the single most widespread and familiar use of radionuclides is as a tag to follow the behavior of some material through some process. The uses of radioactive nuclides as markers in this way are far too numerous to even classify here, and so we highlight only a few examples. The study of chemical and biological reaction paths has been one of the most productive tracer applications, and indeed the use of ^{14}C to elucidate mechanisms of organic reactions has become routine. See Sec. 3.2.3 for an example. Radioactive tracers expedite, and even make possible, the study of a great variety of diffusion and migration problems, from the diffusion of metals in metals to atmospheric and oceanic currents and movement in underground aquifers. Problems in erosion and corrosion of surfaces and studies of friction, wear, and lubrication are important in industry. Tracer applications in biochemistry are also ubiquitous, and we mention only the historic work on photosynthesis, which revolutionized our understanding of this crucial process.

4.3 Isotopic Exchange and Substitution

4.3.1 Reaction Kinetics and Mechanisms The simplest chemical equation expressing exchange is

$$AX + BX^* = AX^* + BX, \tag{4}$$

where X represents a common moiety, X^* being that moiety in which one element's isotopes serves as a radioactive label, while A and B are two other moieties. At equilibrium the specific activities of X in AX and BX are equal, and one measures the rate of approach to this equilibrium. Such measurements can be done under conditions in which equilibrium conditions already prevail among A, B, and X except for the asymmetry of the labeling. Begin with BX labeled and AX unlabeled, and at measured times thereafter recover a sample of AX for measurement. The appropriate expression for this experiment is

$$\ln(1 - f) = [-(a + b)/ab]k_r t, \tag{5}$$

where k_r is the rate for the reaction to proceed to the right, $a = [AX] + [AX^*]$, $b = [BX] + [BX^*]$, t = time elapsed since the start, and $f = a_t/a_e$, where a_t and a_e are the specific activities of X^* in AX at time t and at equilibrium. The square brackets around a chemical formula denote concentration. In this case the rate k_r is evaluated from the slope of a plot of $\ln(1 - f)$ vs t. In more complex reactions—e.g., in which species in addition to AX and BX participate, in which there is more than one reaction pathway, or in which the reaction proceeds in more than one step—the experiment and its analysis become more complicated than the simple straight-line behavior of Eq. (5).

Although the isotopic exchange technique is usually limited to reactions having half-completion times of a few minutes or longer, it has become a standard tool in chemistry. An example is the measurement, in aqueous equilibrium mixtures, of the iodide-catalyzed exchange of radioactive arsenic between arsenious and arsenic acids,

$$H_3AsO_4 + 3I^- + 2H^+ \leftrightharpoons HAsO_2 + I_3^- + 2H_2O.$$

The rate at which the reaction proceeds to the right, R, was found to be

$$R = k_r[H_3AsO_4][I^-][H^+], \tag{6}$$

with $k_r = 0.057\ L^2\ mol^{-2}\ min^{-1}$ (Friedlander *et al.*, 1981).

4.3.2 Electron-Exchange Reactions These are reactions in which a chemical species and its isotopic partner oxidize and reduce each other, in the study of which radioactive labeling has played an important role. An example is the measurement of the rates at which aqueous ferrous and ferric ions exchange labels (indicated by the asterisk),

$$Fe^{*2+} + Fe^{3+} = Fe^{*3+} + Fe^{2+},$$

as a function of the presence of other reagents, such as hydrogen or chloride ions.

The mechanisms involved in electron exchange reactions are unusually amenable to theoretical interpretation, and their study has substantially deepened our understanding of oxidation-reduction reactions. For example, such work led to the discovery that for complexed metals or their ions in aqueous solution the reactions fall into two broad classes, outer sphere and inner sphere. Outer-sphere reactions are those in which each of the two reactants retains its full complement of coordinated ligands throughout the exchange of electrons. Inner-sphere reactions are those in which the reactants share a common ligand during the transition state (Sutin, 1973).

4.3.3 Effects of Isotopic Substitution Although small, these effects are measurable under favorable circumstances. They stem from two causes, molecular symmetry and isotopic mass differences, and affect both equilibrium constants and reaction rates (*kinetic isotope effects*). Their understanding contributes insight into statistical mechanics, and into molecular potential surfaces. The changes are largest for muonium (Mu) and the isotopes of hydrogen, viz. deuterium (^2H) and tritium (^3H), and rapidly become smaller as the elements become heavier (Bigeleisen *et al.*, 1973).

4.4 Hot-Atom Chemistry

When radioactive nuclides are formed by nuclear transformations within molecules, chemical reactions may occur. The study of these processes is one branch of hot-atom chemistry. More generally, hot-atom chemistry is the study of the chemical reactions of atoms or ions that are significantly more energetic than chemical bond strengths, where the energy may be in either translation or excitation, and the energetic atoms or ions may be either free or bound in a molecule. This section focuses on the special case in which the hot atom is radioactive.

Translational energy may arise via nuclear recoil from radioactive decay or from induced nuclear reactions. Characteristic energies range from tens of MeV for fission to tenths to hundreds of eV from beta decay and gamma-ray emission. Excitation energy stems, for example, from electron capture decay and internal conversion of gamma transitions (see RADIOACTIVITY).

The chemical reaction pathways depend on whether the system is solid, liquid, or gaseous, and on other reagents that may be mixed with the target compound. For the gaseous and liquid phases, substantial efforts have been made to develop a theory of hot-atom chemistry. Since it is essentially never possible to follow the chemical events from a unique initial state to a specified final state, stochastic theories are appropriate. One example is the systematization of reactivity data as a function of translational energy and the proportion of inert material mixed with a target. The hot atom's energy is systematically reduced by a sequence of collisions until it comes into the energy range in which reaction is most likely. Much of this work is now being superseded by experiments with lasers and crossed molecular beams, although nucleogenic hot-atom studies in gases may still be the most accessible means for probing reactions at kinetic energies larger than 5 eV.

Important applications of hot-atom chemistry are provided by the Szilard–Chalmers process. For example, when liquid ethyl iodide is irradiated with thermal neutrons, the capture of a neutron by ^{127}I plus the emission of gamma rays from the resulting highly excited state of ^{128}I, at 6.83 MeV, causes the carbon–iodine bond to be broken. The bond breaks because the iodine-atom recoil energies from the gamma-ray emission are in the range of 10–200 eV while the carbon–iodine bond strength is only 2.4 eV. This phenomenon affords a convenient and efficient way to separate the radioiodine from the stable iodine compound, for when the irradiated ethyl iodide is shaken with water, most of the radioactive ^{128}I goes into the aqueous phase. Analogous systems have been developed for the production and separation of many radionuclides, often with high specific activity.

Hot-atom chemistry is also exploited for the synthesis of labeled compounds. For example, an organic reagent from which the labeled compound is to be recovered is mixed with lithium carbonate and irradiated with thermal neutrons. The reaction $^{6}\mathrm{Li}(n,{}^{3}\mathrm{H})^{4}\mathrm{He}$ produces hot tritium nuclei, which are slowed and neutralized by penetration into the surrounding reagent until the labeling reaction occurs. Sometimes a hot-atom method provides the only known route to synthesize a compound; an example is disilene, $H_2Si{=}SiH_2$, the silicon analog of ethylene.

The use or study of hot-atom effects has spread into many different fields. In solids, Mössbauer and perturbed-angular-correlation spectroscopy (see Sec. 4.6.3) have made possible studies of fast processes occurring within nanoseconds of the nuclear decay. In Earth and environmental science, hot-atom effects disrupt the equilibria of the natural radioactive series, with consequences important for geochemistry and uranium prospecting, and for the disposal of radioactive waste. In nuclear medicine, hot-atom effects on tumor cells and bulk tumors are being studied (Tominaga, 1988; Harbottle and Maddock, 1979; Maddock and Wolfgang, 1968).

4.5 Artificial Elements

The effort to discover new elements has demanded the skills of radiochemistry since early in the twentieth century. All of the recently discovered elements are radioactive, many of them short lived, for which studies made using radiochemical techniques are the principal, or even the only, source of chemical information. The transuranium elements were first studied at tracer levels, although several are now available in macroscopic or even bulk quantities. The syntheses of positronium, muonium, muonic atoms, and the like contributed a number of elements outside the traditional periodic system, and gave rise to an exotic chemistry (Ponomarev, 1973). Discussions of the chemistry of positronium and muonium are deferred to Secs. 4.6.1 and 4.6.2.

4.5.1 Observation of Chemical Properties When Only Tracer Amounts Are Available Several elements—astatine, radon, francium, and all those with atomic numbers greater than 101 (nobelium and beyond)—possess no isotopes having half-lives longer than a few days, and so studies of their chemistry must be carried out with very small samples, sometimes only a single atom at a time. Under these circumstances the experiments can be dominated in unpredictable ways by trace impurities that would not noticeably affect the same experiments done with bulk amounts. Formulation of solutions so that *p*H and oxidation-reduction couples are controlled is mandatory. Then solvent extraction, ion exchange, selective adsorption, and various forms of chromatography become reproducible. These are the same kinds of problems already mentioned in Secs. 3.2.2.2 and 3.2.2.3. A more subtle problem associated with experiments with fewer than about a hundred molecules or atoms is that chemical behavior can no longer be described straightforwardly in terms of classical thermodynamics (Guillaumont *et al.*, 1991).

4.5.2 Astatine as an Example The results of investigations on this element, the heaviest halogen, exemplify what can be learned with only tracer methods. Most work has been done with astatine's second longest-lived isotope, ^{211}At, $t_{1/2}$ = 7.2 h (for ^{210}At, $t_{1/2}$ = 8.3 h), though the short-lived ^{217}At ($t_{1/2}$ = 0.032 s) has also proved useful. Many compounds have been synthesized and characterized, both inorganic and organic, such as HAt, AtCl, CH_3At, C_6H_5At, plus aqueous species, At^-, AtO_2^-, AtO_3^-, AtO_4^-, $AtCl_2^-$. Simple molecular ions, e.g., HAt^+, CH_3At^+, $AtOH^+$, $AtCl^+$, have been identified by mass spectrometry. The compounds now known reveal that astatine, like iodine, displays formal valences from -1 to $+7$. The bond dissociation energy of HAt was measured to be 58 kcal mole^{-1} via the temperature dependence of its reaction with oxygen, and the strengths of carbon–astatine bonds in n-C_3H_7At and C_6H_5At were measured to be about 39 and 44 kcal mole^{-1}, respectively, by measurements of the rates of thermal decomposition in gaseous mixtures. In many ways astatine forms compounds and reacts as expected for a halogen heavier than iodine, but it also resembles the more metallic polonium. In aqueous solutions, oxidation potentials relating At^-, At^0, and AtO_3^- were measured by the combination of a variety of methods. By a solvent extraction method, the acid dissociation constant of aqueous astatoacetic acid, $AtCH_2COOH$, was found to be (1.5–1.8) × 10^{-4} mol L^{-1} at room temperature, ten times more acidic than acetic acid, but one-fifth as acidic as iodoacetic acid. Boiling points of many organic compounds, such as C_2H_5At (98 °C), and C_6H_5At (217 °C) have been estimated via gas chromatography and high-pressure liquid chromatography, where the apparatus was calibrated with the corresponding compounds of iodine and bromine. Organic reactions such as electrophilic substitution, diazotization, and catalyzed exchange have been successfully applied to prepare a wide range of materials. The motivation for much of this synthetic effort is that if an astatine-labeled compound can be found that concentrates in a malignant tumor, the densely ionizing and short-range alpha radiation, essentially unaccompanied by other radiation if ^{211}At is used, would be ideal for sharply localized destruction of the diseased tissue (Roessler, 1989; Kugler and Keller, 1985).

4.5.3 The Heaviest Elements The discovery and characterization of the transuranium elements demanded the development of new radiochemical techniques as novel ideas were exploited for the nuclear syntheses, and as the nuclear and chemical properties of the new elements became known. Cyclotron irradiations with deuterons and helium ions typified the earlier work, often using as targets elements that had only recently been discovered. The latter were synthesized in amounts sufficient for the experiments by prolonged irradiations in nuclear reactors, with subsequent radiochemical preparation. Further progress was achieved by bombardment of such targets with heavy ions, i.e., atoms of elements far beyond helium, such as ^{11}B, ^{12}C, and ^{22}Ne, stripped of all or most of their orbital electrons. However, two elements, einsteinium and fermium, were first identified, by radiochemical methods, among the products of thermonuclear explosions.

The elements from actinium to lawrencium were found to have properties that justify their description as actinides, i.e., the series of elements in which the 5*f* electron shell is being filled, in analogy to the lanthanides, whose properties reflect the filling of the 4*f* shell. This actinide series behavior causes a useful regularity in ion-exchange chromatography that was exploited both to help establish elemental identities and to codify some chemical properties, even when only a few atoms of a new element were present. A dramatic change of chemical properties in proceeding from lawrencium (Lr), atomic number $Z = 103$, to rutherfordium (Rf), $Z = 104$, is expected if Lr is the highest member of the actinide family while Rf belongs to the hafnium/zirconium family, Group IVb of the periodic table. That this behavior actually occurs was first established by ion-exchange chromatography, later confirmed by thermochromatography, and reconfirmed by solvent extraction experiments, even though the longest-lived isotopes of Lr and Rf then known had half-lives of only 180 and 65 s, respectively. (The elemental names used in this section are those recommended in 1995 by the Committee on Nomenclature of the American Chemical Society. However, at the time of this writing the assignments are still tentative.)

To date, hahnium, Ha ($Z = 105$) is the element of highest atomic number for which chemical studies have been carried out. For nielsbohrium, Ns ($Z = 107$), hassium, Hs ($Z = 108$), meitnerium, Mt ($Z = 109$), and elements 110 and 111, the half-lives range from 2 μs to 1 s. Identification and separation are therefore carried out by nonchemical techniques, and determination of chemical properties seems not to be possible with known methods. Recently, however, an isotope of seaborgium ($Z = 106$), ^{266}Sg, is reported to possess a half-life of 20–30 s (Seaborg, 1995; Hoffman, 1994; Seaborg and Loveland, 1990; Armbruster, 1985).

Substantial but unsuccessful efforts by several groups to synthesize and characterize still heavier elements were motivated by theoretical predictions that elements in the range $Z = 110$ to 114 might have lifetimes of many years. More recent theory predicts a region of increased stability for deformed nuclei around ^{270}Hs. The relatively long half-life of ^{266}Sg lends support to this theory (Möller and Nix, 1994).

4.6 Joint Use of Radioactivity and Chemical Properties for Probes and Techniques

4.6.1 Positron Annihilation and Positronium When a positron (antielectron) interacts with matter it undergoes a reaction with a normal electron in which both particles are annihilated, and the energy represented by their mass appears as gamma radiation. The experiments are usually performed with the energetic positrons emitted by radioactive nuclides, such as 2.6-y ^{22}Na. In condensed phases the annihilation reaction essentially does not occur, however, until the positron slows to thermal energies. Most of the gamma radiation appears as two photons of 511 keV each, emitted in opposite directions, but a small fraction appears as three photons, of total energy 1.022 MeV. The ratio of three-quantum to two-quantum annihilation depends on the chemical composition of the material in which it takes place. The explanation is that a fraction of the positrons form *positronium* atoms, analogous to hydrogen atoms with the protons replaced by positrons.

Positronium (Ps) is formed in one of two states, orthopositronium (*o*-Ps), with electron and positron spins parallel, and parapositronium (*p*-Ps), with spins opposed. Orthopositronium has an intrinsic lifetime of 10^{-7} s and decays by three-photon emission, while *p*-Ps, mean life 10^{-10} s, decays via two-photon emission. However, the interactions of *o*-Ps with matter during its relatively long lifetime usually cause its positron to decay by two-photon emission also. One such interaction is electron pickoff, where the positron annihilates with an electron other than that to which it is bound. Another is the transition from *o*-Ps to *p*-Ps induced by a magnetic field, either via an encounter with a paramagnetic species such as O_2, or by an externally imposed field. A third is chemical reactions of *o*-Ps, e.g., oxidation, reduction, or addition, which cause either an ortho-to-para conversion or the liberation of the positron from combination.

Much has been learned with relatively simple apparatus, viz., two gamma detectors tuned to 511 keV in variably delayed coincidence, arranged to view at an angle of 180° the material in which annihilation is occurring. From the lifetime of the delayed events, which at times longer than 10^{-9} s come from the converted *o*-Ps, one can measure its chemical reactivity. More complex experiments have allowed measurements of the momentum distribution of the annihilated electrons, and studies of imperfections in solids. Positron annihilation has also become an important tool for surface studies, because, for example, it is able to measure surface areas of highly convoluted materials for which other methods, such as the determination of the amount of an adsorbed gas, fail (POSITRON ANNIHILATION SPECTROSCOPY; Cheng *et al.*, 1989; Berko and Pendleton, 1980).

Positron emission tomography (PET) (see TOMOGRAPHY, POSITRON EMISSION) is an imaging technique that exploits the properties of annihilation radiation. Using PET, one has the unique capability to study ongoing metabolic processes localized within a living organism, notably the human brain. The method has spurred vigorous efforts to synthesize and test radiopharmaceuticals labeled with positron-emitting nuclides, especially 110-min ^{18}F (Kilbourn, 1990).

4.6.2 Muon Spin Rotation and Muonium Muon beams are generated by the decay of pion beams, from which process they emerge with spins aligned along the direction of motion. This polarization and the muon's long lifetime of 2.2 μs have led to the development of a unique research technique. If positively charged muons are stopped in matter, most of the polarization remains and evolves according to the properties of the sites that the muons then occupy. A muon with specified spin direction decays anisotropically into a positron, neutrino, and antineutrino, where the greatest probability is for the positron to be emitted in the direction of the spin vector, so that detection of positrons in the original forward direction of the beam allows any changes in the polarization to be monitored. Such changes are termed *muon spin rotation.* To carry out these experiments, a delayed-coincidence technique is used, where the entry of the energetic muon into the stopping region provides the start signal, and the detection of the positron is the stop. It is important to use beams of low-energy (4.1-MeV) muons, which are easily stopped. Targets can then be made small, and even gaseous, allowing a wide range of experiments. The necessary low-energy beams are obtained by tuning the beam line to collect only muons of a momentum consistent with the decay of pions at rest near the surface of the production target.

The muon possesses a substantial magnetic moment, 3.18 times that of the proton; so if a transverse magnetic field is applied to the stopping region, the residual polarization can be monitored via the precession of the muons, which appears as an oscillation superimposed on the exponential decay of the muon, as recorded by the intensity of the emitted positrons. The frequency of this oscillation depends on the imposed field plus any local field at the sites where the muons come to rest. In general, the stopping material will offer more than one kind of site and a dynamic fluctuation of the sites' field strengths, which will give rise to more than one oscillation frequency plus a gradual loss of polarization. It is the measurement of this spectrum of frequencies and their relaxations for chosen stopping materials that is the aim of most muon spin rotation experiments.

In many systems *muonium* (Mu) is formed, i.e., an atom in which an electron has become bound to the positive muon, which then serves as the nucleus. This atom behaves like a radioactive isotope of hydrogen of only 0.11 atomic mass units. Typically, in reactions such as $\mathrm{Mu} + \mathrm{Cl}_2 \rightarrow \mathrm{MuCl} + \mathrm{Cl}$ one can relate the attenuation of the muonium oscillation to the transfer of the muon from paramagnetic Mu into diamagnetic MuCl, and thereby measure the reaction rate. Rate measurements for this and similar reactions have substantially improved the theoretical understanding of isotope effects and tunneling, and thus of the reactant potential hypersurfaces. Another application is the synthesis of free radicals by muonium addition—for example, $\mathrm{C_2H_4} + \mathrm{Mu} \rightarrow \mathrm{C_2H_4Mu}\cdot$—and measurement of their muon spin rotation spectra. These data produce values of the hyperfine coupling constants, from which information about the molecular structure can be extracted.

Besides applications to chemistry (Percival, 1991; Ito, 1990), there are investigations into high-temperature superconductors (Emery, 1991; Uemura, 1991) and magnetic materials, semiconductors, metals, and other aspects of condensed matter (Cox, 1987) (see MUON SPIN ROTATION/RELAXATION/RESONANCE; MUONIC, MESONIC, AND BARYONIC ATOMS AND MOLECULES).

4.6.3 Perturbed Angular Correlations When a nucleus undergoes two successive emissions, such as two gamma-ray photons, the angle between them is not, in general, random, unless the spin of the intermediate state is less than 1. The angular distribution is described by

$$W(\theta) = \sum_n A_n P_n(\cos\theta), \quad n = 0, 2, 4, \ldots \qquad (7)$$

Here, $W(\theta)$ is the coincidence counting rate where θ is the angle between the two emissions, and $P_n(\cos\theta)$ is the Legendre polynomial of order n. Often the fit to the data is carried only as far as $n = 2$, and seldom further than $n = 4$. If the lifetime of the intermediate nuclear state is comparable to its precession rate, one can observe the effect of the environment as a time-dependent perturbation of the angular correlation:

$$W(\theta) = \sum_n A_n G_n(t) P_n(\cos\theta), \qquad (8)$$

where $G_n(t)$ is the perturbation factor. The precession can arise both from the interaction of the nuclear magnetic moment with any magnetic field present, and from the interaction of the nuclear quadrupole moment with an electric field gradient. When the transforming nuclei occupy sites in a crystal lattice, $G_n(t)$ is periodic. Thus, in delayed coincidence at a favorable angle, one sees not only the exponential decay of the intermediate state, but also a superimposed oscillation. If the transforming nuclide is the radioactive label of a molecule in liquid solution, the perturbations operate randomly as the molecule tumbles in response to thermal motion, and the factor $G_n(t)$ is not periodic, but decays exponentially at a rate characteristic of the viscosity of the solvent and the structure of the labeled molecule.

The method of perturbed angular correlations has found wide application for studies in both solids and liquids, and in the life sciences. The probe-atom concentration can be so low that the properties of host materials are not influenced. For quadrupole interactions the probes are sensitive to their local environment and thus provide information on microscopic properties. Thus, for metals, it is a powerful tool to study impurity interactions in dilute alloys, provides a spectroscopy for point-defects, and has supplied most of the experimental data on the quadrupole interaction. For semiconductors, this method is particularly sensitive to complex formation between radioactive dopant atoms and defects, and can be used to gain insight into diffusion, short-range order, segregation, ordering, and dynamic effects, where local jump motion of acceptor–donor pairs is examined via fluctuations in the strength and orientation of the electric field gradient. In liquids the method is used to study the solution chemistry of metal complexes and the rotational diffusion of macromolecules, and to measure the dissociation energies of complexes and addition compounds, and the correlation time of simultaneous equilibria for addition reactions. It is a sensitive method for measuring the electric field gradients in chemical compounds, results of which can be related to molecular structure, including protein–metal interactions. The perturbed angular correlation method complements Mössbauer spectroscopy, as it can be used for elements lacking suitable gamma-ray resonances, although it is insensitive to electric monopole interactions. It operates without solid-state and temperature restrictions, so that, for example, liquids and high-temperature phase transitions can be studied (Collins *et al.*, 1990; Bauer, 1985; Rinneberg, 1979; Adloff, 1978).

4.6.4 Matching of Nuclear Lifetimes to Chemical Rates in Tracer Studies For experiments in which a radioactive tracer is used to measure the time dependence of a chemical process, there is a regime of nuclide half-lives for which the measurement is most sensitive. We have already seen this without explicit mention in several sections above, for the reactions of Ps (Sec. 4.6.1) and muonium (Sec. 4.6.2), for perturbed angular correlations (Sec. 4.6.3), and in fast radiochemical techniques such as thermal chromatography (Sec. 3.2.2.3). In general, one has a signal rate S, and a background rate B, in counts per unit time, while there is some relationship between S and the quantity ξ that one wishes to measure. As examples ξ can be a chemical rate, a diffusion constant, or a cross section. If one wants to measure ξ to within one standard deviation $\pm\Delta\xi$, then the time T_m required to carry out the measurement is

$$T_m = \left(\frac{\xi}{\Delta\xi}\right)^2 S^{-1}\left[1 + 2\left(\frac{B}{S}\right)\left(1 + \left\{1 + \frac{S}{B}\right\}^{1/2}\right)\right] \times \left(\frac{d\,\ln\xi}{d\,\ln S}\right)^2. \qquad (9)$$

The factor $(d \ln\xi/d \ln S)^2$ quantifies the insensitivity for specifying ξ via S.

If the nuclear lifetime is too short, too much tracer will decay away before the chemical process has had time to affect it appreciably, the factor S^{-1} becomes large, and sensitivity is reduced. If the nuclear lifetime is too long, two problems may arise.

1. The chemical process is completed within a small fraction of the nuclear lifetime and therefore may have only a small effect on the nuclear counting rate, i.e., the insensitivity factor $(d \ln\xi/d \ln S)^2$ becomes large, increasing T_m.
2. The measurement may be susceptible to background contributed by the tracer itself if the experiment is of such a nature that the tracer cannot be removed after it has been used, and T_m then increases because B/S grows large.

As an example, consider the measurement of the mean time t_d that it takes radon atoms, produced by the decay of a radioactive precursor, to escape from barium stearate powder into the gas phase, where the decays of gas-phase radon atoms are counted to provide S. The value of t_d is calculated from the ratio of gas-phase to solid-phase radon atoms. Then $\xi = t_d$ and the insensitivity factor is easily shown to be $(d \ln t_d/d \ln S)^2 = ([\tau + t_d]/t_d)^2$, where $\tau = t_{1/2}/\ln 2$, $t_{1/2}$ being the radon's half-life. The resulting dependence of T_m on $t_{1/2}/t_d$, assuming that $B/S = 0$, is displayed in Fig. 3(a), and shows that for optimum sensitivity (minimum counting time) the half-life of the radon isotope chosen should be within an order of magnitude of t_d. If, for a given amount of radon radioactivity, S is too small, one needs to use an isotope of longer half-life, but if essentially all of the radioactivity is in the gas phase, an isotope of shorter half-life is needed. In this case measurement of t_d could be made over a wide range of temperature using 0.035-s ^{218}Rn, 4.0-s ^{219}Rn, and 56-s ^{220}Rn, from which it was learned that two mechanisms contribute to the emanation: diffusion, and prompt emission via nuclear recoil.

To illustrate the role of the background, consider the measurement of differential cross sections for chemically reactive scattering in crossed molecular beams, where one of the reagent molecular beams is labeled with a radioactive tracer. The experimental situation is diagrammed in Fig. 4. The beams cross each other at right angles in a vacuum chamber. A detector capable of chemical discrimination, whose position is well defined and whose aperture subtends a known small solid angle with respect to the collision zone, collects the radioactive product. The detector is movable, and product intensities, measured as counting rates of the captured product, are obtained at many positions around the collision zone until the whole scattering distribution has been mapped. Here, the background level due to the back-diffusion of small amounts of tracer limits the sensitivity. Application of Eq. (9) reveals that the most pertinent quantity is the time for the reaction product to fly from the collision zone to the detector, t_a, typically a few milliseconds. Best sensitivity, Fig. 3(b), requires that the tracer radioactive half-life $t_{1/2}$ fall in the range $0.5 \leq t_{1/2}/t_a \leq 10^2$, i.e.,

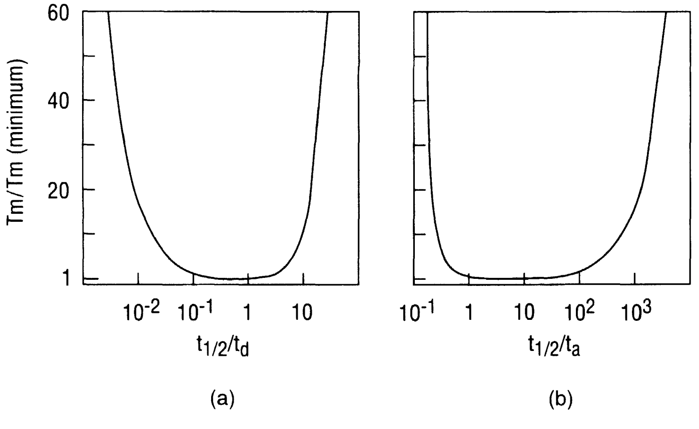

FIG. 3. The relationship between measurement sensitivity and nuclide half-life. T_m is the time required to measure to a specified precision at a given half-life, and T_m(minimum) is the minimum possible time required for that measurement. (a) Measurement of emanation rate of radon from barium stearate powder into the gas phase. (b) Measurement of differential cross sections for reactive scattering in crossed molecular beams.

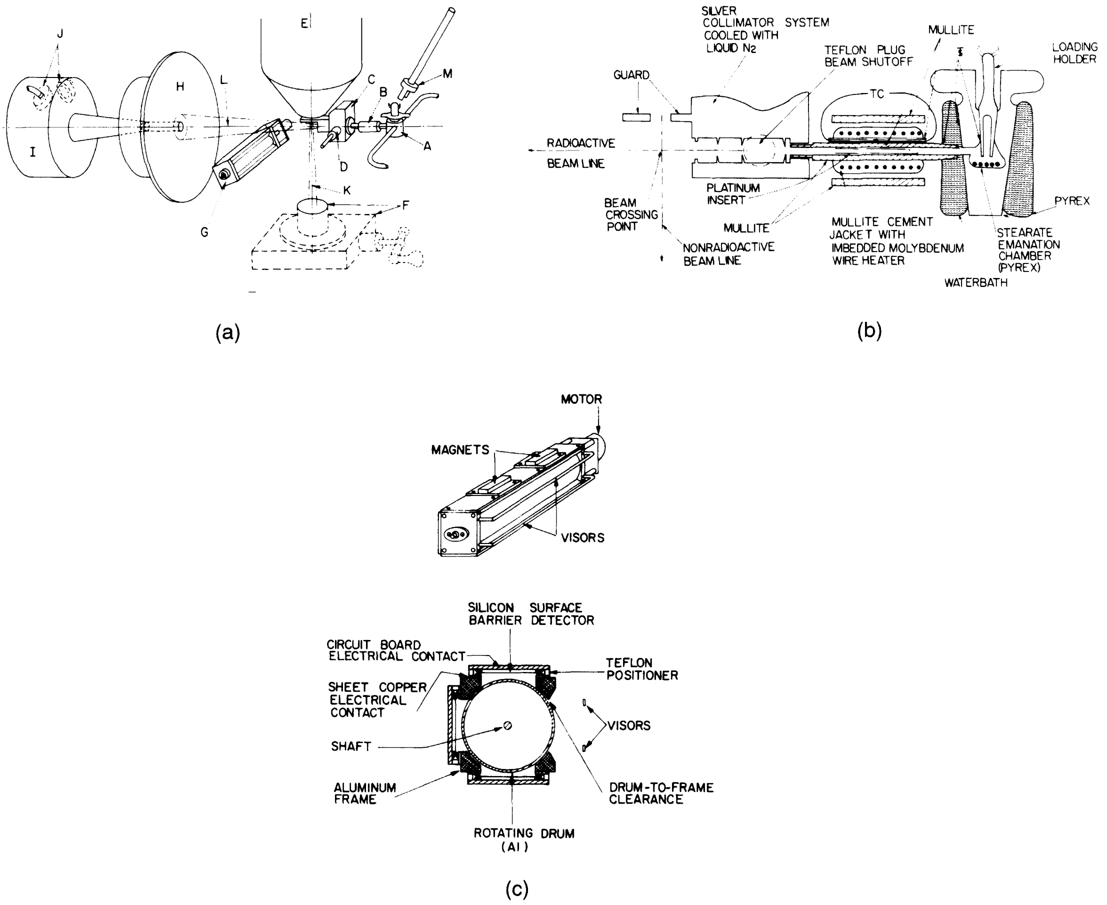
J
E
H
L
C
B
M
A
I
D
K
F
G
SILVER COLLIMATOR SYSTEM COOLED WITH LIQUID N2
MULLITE
TEFLON PLUG BEAM SHUTOFF
LOADING HOLDER
GUARD
TC
RADIOACTIVE BEAM LINE
PLATINUM INSERT
BEAM CROSSING POINT
MULLITE
NONRADIOACTIVE BEAM LINE
MULLITE CEMENT JACKET WITH IMBEDDED MOLYBDENUM WIRE HEATER
PYREX
STEARATE EMANATION CHAMBER (PYREX)
WATERBATH
MOTOR
MAGNETS
VISORS
SILICON SURFACE BARRIER DETECTOR
CIRCUIT BOARD ELECTRICAL CONTACT
TEFLON POSITIONER
SHEET COPPER ELECTRICAL CONTACT
VISORS
SHAFT
ALUMINUM FRAME
DRUM-TO-FRAME CLEARANCE
ROTATING DRUM (Al)

(a) (b)

(c)

FIG. 4. (a) Diagram of apparatus used to measure chemically reactive scattering in crossed molecular beams, where one of the beams is labeled with a short-lived nuclide. The demonstration reactions were HAt + Cl,Br → At + HCl,HBr using 0.032-s ^{217}At. A, radioactive beam source; B, scrubber-furnace; C, liquid-nitrogen–cooled collimator; D, shut-off plug; E, nozzle beam furnace and cryopump; F, gate valve to isolate cryogenically captured reagent beams for subsequent analysis; G, hodoscope detector; H, liquid-nitrogen–cooled beam dump; I, calibrated beam monitor; J, silicon surface-barrier detectors; K, halogen-atom crossing beam; L, radioactive beam; M, rotary feedthrough used to close source stopcock. (b) Detail showing the source for generation of a molecular beam of HAt purified to 99.99%. (c) Detail showing the detector selective for product At atoms. The At sticks to the rotating drum to be carried before the counters, while HAt bounces away. The apparatus shown combines five detectors side by side. (Grover *et al.* 1976; reproduced by permission of the American Institute of Physics.)

from one to a few hundred milliseconds. For suitable systems, radioactive molecular beams offer much greater sensitivity than any other known method for the measurement of differential cross sections in chemically reactive scattering (Grover, 1979; Grover *et al.*, 1976).

4.7 Other Applications

Space does not permit even a brief description of the many other areas in which radiochemistry plays an important role. We list some of those not mentioned above, with suggested reading for each: nuclear reactions (Friedlander *et al.*, 1981); Earth, solar system, cosmos studies (Das *et al.*, 1989); dating (see RADIOACTIVITY; Koppenaal, 1988, 1990; Faure, 1986; Libby, 1952); biomedical applications (see NUCLEAR MEDICINE; Iyengar, 1989; Cesareo, 1988; Abraham, 1977); Mössbauer spectroscopy (see MÖSSBAUER EFFECT; Long and Grandjean, 1989; Greenwood and Gibb, 1971); nuclear use of lasers (Otten, 1989); environment (Tolgyessy and Klehr, 1987); environmental perturbation of nuclear decay rates (Dostal *et al.*, 1977); nuclear energy fuel cycle (Benedict *et al.*, 1981; Choppin and Rydberg, 1980); radioactivation applied to forensic science (Guinn, 1974); and archaeometry (Harbottle, 1990).

GLOSSARY

Atomic Number, *Z*: The number of units of positive charge on the atomic nucleus.

Barn, b: A unit of cross section, equal to 10^{-24} cm^2.

Becquerel, Bq: A unit of nuclear disintegration rate, equal to one disintegration per second.

Carrier: A weighable quantity of an element mixed with the radioactive isotope of that element.

Carrier-Free: A radioactive isotope for which no stable isotope has been added, although one or more stable isotopes might already be present as contaminants or nuclear coproducts.

Curie, Ci: A unit of nuclear disintegration rate, equal to 3.7×10^{10} disintegrations per second, or 3.7×10^{10} Bq.

Decontamination Factor: The factor by which the ratio of a given impurity nuclide to a desired nuclide is reduced in a single purification operation.

Holdback Carrier: A weighable quantity of a stable isotope(s) of a radionuclide impurity introduced to aid in radiochemical purification.

Label: A radioactive isotope chemically bonded in a defined position in a compound.

Muonium, Mu: A radioactive analog of hydrogen, whose nucleus is a positively charged muon.

Nuclear Chemistry: The study of the atomic nucleus, nuclear reactions, and related phenomena by scientists employed in chemistry departments, who often utilize techniques that require chemical skills. Synonymous with radiochemistry in earlier times.

Nuclear Mass Number: The number of nucleons in a nucleus.

Nucleon: Generic name for a proton or neutron.

Nuclide: A species of atom for which both the atomic number and nuclear mass number are specified.

On-Line Isotope Separator: A mass spectrometer operated simultaneously with the nuclear production and ionization of radionuclides.

Positronium, Ps: An atomiclike system consisting of an electron and a positron bound together. In parapositronium the electron and positron spin vectors are opposed in direction, and it decays with a mean lifetime of about 10^{-10} s, into two photons. In orthopositronium the spin vectors are parallel in direction, and it decays with a mean lifetime of about 10^{-7} s, into three photons.

Radiation Chemistry: The study of the chemical effects of ionizing radiation on matter. Once a subfield of radiochemistry.

Radioactive Decay Chain: A sequence of more than two elements connected to one another by radioactivity.

Radioimmunoassay: An isotope dilution technique for measuring small concentrations of biologically active substances, such as hormones, drugs, enzymes, and viruses, in the body fluids of living organisms.

Radiolysis: Chemical decomposition by the action of ionizing radiation, especially self-irradiation of a radioactively labeled compound.

Radionuclide: A radioactive nuclide.

Specific Activity: Radioactive disintegration rate per gram of matter, usually expressed either as Ci/g (or mg) or Bq/g (or mg).

Spent Reactor Fuel: Nuclear fuel removed from reactors, and replaced with fresh fuel, because the depletion of its fissionable component and the buildup of neutron-consuming fission products have curtailed its efficiency.

Thermochromatography: A type of gas chromatography in which thermal gradients are exploited to effect chemical separations.

Tracer: A radioactive nuclide used as an indicator, via detection of its disintegration products.

Transuranium: Describes any element with atomic number greater than 92, i.e., greater than that of uranium.

Works Cited

Abraham, G. E. (1977), *Handbook of Radioimmunoassay,* New York: M. Dekker.

Adloff, J.-P. (1978), *Radiochim. Acta* **25,** 57–74.

Alfassi, Z. B. (Ed.) (1989), *Activation Analysis,* Boca Raton, FL: CRC Press.

Armbruster, P. (1985), *Annu. Rev. Nucl. Part. Sci.* **35,** 135–194.

Bauer, R. (1985), *Quart. Rev. Biophys.* **18,** 1–64.

Benedict, M., Pigford, T. H., Levi, H. W. (1981), *Nuclear Chemical Engineering,* 2nd ed., New York: McGraw-Hill.

Berko, S., Pendleton, H. N. (1980), *Annu. Rev. Nucl. Part. Sci.* **30,** 543–581.

Bigeleisen, J., Lee, M. W., Mandel, F. (1973), *Annu. Rev. Phys. Chem.* **24,** 407–440.

Bonner, N. A., Kahn, M. (1951), in: A. C. Wahl, N. A. Bonner (Eds.), *Radioactivity Applied to Chemistry,* New York: John Wiley & Sons, Inc., p. 102.

Browne, E., Firestone, R. B., Shirley, V. S. (1986), *Table of Radioactive Isotopes,* New York: Wiley.

Cesareo, R. (Ed.) (1988), *Techniques and Instrumentation in Analytical Chemistry,* Vol. 8, *Nuclear Analytical Techniques in Medicine,* Amsterdam: Elsevier.

Cheng, K. L., Jean, Y. C., Luo, X. H. (1989), *Crit. Rev. Anal. Chem.* **21,** 209–224.

Choppin, G. R., Rydberg, J. (1980), *Nuclear Chemistry,* New York: Pergamon Press.

Collins, G. S., Shropshire, S. L., Fan, J. W. (1990), *Hyperfine Interactions* **62,** 1–34.

Cox, S. F. J. (1987), *J. Phys. C.: Solid State Phys.* **20,** 3187–3319.

Das, H. A., Faanhof, A., van der Sloot, H. A. (1989), *Radioanalysis in Geochemistry,* Amsterdam: Elsevier.

Dostal, K.-P., Nagel, M., Pabst, D. (1977), *Z. Naturforsch.* **32a,** 345–361.

Ehmann, W. D., Vance, D. E. (1991), *Radiochemistry and Nuclear Methods of Analysis,* New York: John Wiley & Sons.

Ehmann, W. D., Robertson, J. D., Yates, S. W. (1994), *Anal. Chem.* **66,** 229R–251R. A recent example of the biennial reviews on nuclear and radiochemical analysis published by this journal.

Emery, V. J. (1991), *Hyperfine Interact.* **63,** 13–21.

Faure, G. (1986), *Principles of Isotope Geology,* 2nd ed., New York: John Wiley & Sons.

Fowler, J. S., Wolf, A. P. (1982), *The Synthesis of Carbon-11, Fluorine-18, and Nitrogen-13 Labeled Radiotracers for Biomedical Applications,* NAS-NS-3201, Technical Information Center, U.S. Department of Energy.

Friedlander, G., Kennedy, J. W., Macias, E. S., Miller, J. M. (1981), *Nuclear and Radiochemistry,* New York: John Wiley & Sons.

Glasstone, S. (1950), *Sourcebook on Atomic Energy,* New York: Van Nostrand.

Greenwood, N. N., Gibb, T. C. (1971), *Mössbauer Spectroscopy,* London: Chapman and Hall Ltd.

Grover, J. R., Iden, C. R., Lilenfeld, H. V., Kiely, F. M., Lebowitz, E. (1976), *Rev. Sci. Instrum.* **47,** 1098–1108.

Grover, J. R. (1979), *Radiochim. Acta* **26,** 65–70.

Guillaumont, R., Adloff, J.-P., Peneloux, A., Delamoye, P. (1991), *Radiochim. Acta* **54,** 1–15.

Guinn, V. P. (1974), *Annu. Rev. Nucl. Sci.* **24,** 561–591.

Hansen, P. G. (1979), *Annu. Rev. Nucl. Part. Sci.* **29,** 69–119.

Harbottle, G., Maddock, A. G. (Eds.) (1979), *Chemical Effects of Nuclear Transformations in Inorganic Systems,* Amsterdam: North-Holland.

Harbottle, G. (1990), *Top. Curr. Chem.* **157,** 59–91.

Herrmann, G., Denschlag, H. O. (1969), *Ann Rev. Nucl. Sci.* **19,** 1–32.

Herrmann, G., Trautmann, N. (1982), *Annu. Rev. Nucl. Part. Sci.* **32,** 117–147.

Hickmann, U., Greulich, N., Trautmann, N., Gäggeler, H., Gäggeler-Koch, H., Eichler, B., Herrmann, G. (1980), *Nuclear Instrum. Meth.* **174,** 507–513.

Hoffman, D. C. (1994), *Chem. Eng. News* **72** (18), 24–34.

Ito, Y. (1990), *Top. Curr. Chem.* **157,** 93–128.

Iyengar, G. V. (1989), *Elemental Analysis of Biological Systems,* Vol. I, *Sampling and Sample Preparation Techniques with Special Reference to Trace Elements,* Boca Raton, FL: CRC Press.

Kilbourn, M. R. (1990), *Fluorine-18 Labeling of Radiopharmaceuticals,* Washington, D.C.: National Academy Press.

Koppenaal, D. W. (1988), *Anal. Chem.* **60,** 113R–131R.

Koppenaal, D. W. (1990), *Anal. Chem.* **62,** 303R–324R.

Kugler, H. K., Keller, C. (Eds.) (1985), *Gmelin Handbook of Inorganic Chemistry,* 8th ed., *Astatine,* Berlin: Springer-Verlag.

Libby, W. F. (1952), *Radiocarbon Dating,* Chicago: University of Chicago Press.

Long, G. J., Grandjean, F. (1989), *Mössbauer Spectroscopy Applied to Inorganic Chemistry,* Vol. 3, New York: Plenum Publishing Corporation.

Maddock, A. G., Wolfgang, R. (1968), in: L. Yaffe (Ed.), *Nuclear Chemistry,* Vol. II, New York: Academic.

Möller, P., Nix, J. R. (1994), *J. Phys. G: Nuclear and Particle Phys.* **20,** 1681–1747.

Murray, A., Williams, D. L. (1958), *Organic Syntheses with Isotopes,* New York: Interscience Publishers, Inc.

Otten, E. W. (1989), *Investigation of Short-Lived Isotopes by Laser Spectroscopy,* New York: Harwood Academic Publishers.

Percival, P. W. (1991), *Hyperfine Interact.* **65,** 901–911.

Ponomarev, L. I. (1973), *Annu. Rev. Nucl. Sci.* **23,** 395–430.

Rinneberg, H. H. (1979), *At. Energy Rev.* **17,** 477–595.

Roessler, K. (Ed.) (1989), "Inorganic and General Astatine Chemistry," in *Radiochim. Acta* **47,** 81–168.

Rutherford, E., Chadwick, J., Ellis, C. D. (1930, 1951), *Radiations from Radioactive Substances,* London: Cambridge University Press.

Seaborg, G. T. (1995), *Accounts of Chemical Research* **28,** 257–264.

Seaborg, G. T., Loveland, W. D. (1990), *The Elements Beyond Uranium,* New York: Wiley.

Sutin, N. (1973), in: G. L. Eichorn (Ed.), *Inorganic Biochemistry,* Amsterdam: Elsevier.

Tolgyessy, J., Klehr, E. H. (1987), *Nuclear Environmental Chemical Analysis,* New York: John Wiley.

Tolgyessy, J., Kyrs, M. (1989), *Radioanalytical Chemistry,* Chichester, U.K.: Ellis Horwood Ltd.

Tominaga, T. (Ed.) (1988), "Hot Atom Chemistry," in *Radiochim. Acta* **43,** 57–128.

Trautmann, N., Kaffrell, N., Behlich, H. W., Folger, H., Herrmann, G., Hübscher, D., Ahrens, H. (1972), *Radiochim. Acta* **18,** 86–101.

Uemura, Y. J. (1991), *Physica B (Amsterdam)* **169,** 99–106.

Vandecasteele, C. (1988), *Activation Analysis with Charged Particles,* Chichester, U.K.: Ellis Horwood Ltd.

Further Reading

Adloff, J.-P., Guillaumont, R. (1993), *Fundamentals of Radiochemistry,* Boca Raton, FL: CRC Press.

Nuclear Data Sheets, New York: Academic Press, Inc. An exhaustive and critical presentation of the properties of every known nuclide, periodically updated, can be found in this journal, which is published quarterly.

Radiochemistry and Radiochemical Techniques, Washington, D.C.: U.S. Department of Energy Technical Information Center. This is a series of monographs, in which the radiochemistry of every element is described, as well as expositions of many techniques.

Wang, C. H., Willis, D. L., Loveland, W. D. (1975), *Radiotracer Methodology in the Biological, Environmental, and Physical Sciences,* Englewood Cliffs, N.J.: Prentice-Hall.

RADIOGRAPHY

See X-RAY IMAGING, RADIOGRAPHY AND TOMOGRAPHY

RAMAN SCATTERING

IRVING P. HERMAN, *Department of Applied Physics, Columbia University, New York, New York, U.S.A.*

Introduction 587
1. **Features of Raman Scattering** 588
1.1 Kinematics 588
1.2 Dynamics of Scattering 588
1.3 Nonlinear Raman Processes 590
2. **Instrumentation** 591
3. **Vibrational Raman Scattering in Solids** 593
3.1 Crystals 593
3.2 Disordered Structures 597
3.3 Two-Dimensional Structures 598
3.4 Perturbations Due to Strain, Changes in Temperature, and Electric Fields 598
3.5 Phase Transitions 600
3.6 Time-Resolved Studies 600
4. **Magnetic Raman Scattering** 600
5. **Electronic Raman Scattering** 602
5.1 Bound Electrons 602
5.2 Free Carriers 602
5.3 Two-Dimensional Structures 605
5.4 Time-Resolved Studies 606
6. **Raman Scattering in Gases and Liquids** 606
7. **Applications in Materials Characterization, Process Diagnostics, and Laser Development** 609
Glossary 610
Works Cited 610
Further Reading 612

INTRODUCTION

Raman scattering refers to most types of the inelastic scattering of electromagnetic radiation, which in quantized form consists of photons. Photons entering a medium can lose energy to the sample, which is referred to as a Stokes process, or gain energy, which is an anti-Stokes process. Raman scattering is often described in terms of the excitation or de-excitation of quantized elementary excitations in the medium, such as optical phonons in condensed matter or vibrations in molecules, magnons (magnetic excitations), and single or collective excitations of electrons. Raman scattering has been used in the fundamental spectroscopic study of excitations in a wide range of solids, liquids, and gases, and it has also been used in materials characterization and for *in situ* diagnostics during processing.

One exception to the range of inelastic scattering phenomena called Raman scattering is scattering from density changes due to pressure fluctuations, which is instead called Brillouin scattering. These fluctuations propagate as sound waves, or acoustic phonons in quantized form. A broader definition of Brillouin scattering includes the scattering of any excitation with a linear dispersion relation, such as acoustic phonons and plasmons, and any other very low-frequency excitation $<5\ \mathrm{cm}^{-1}$. Raman scattering is also to be distinguished from Rayleigh scattering, which is quasielastic scattering by nonpropagating density changes caused by entropy fluctuations. In Rayleigh scattering, the scattered light has the same frequency as the incident light but can have a broader spectral width.

Raman scattering is a coherent two-photon process in which the photon scatters only while the incident photon is present. It is formally different from the sequential absorption of a photon followed by fluorescence, though near resonance both types of processes can be important. Spontaneous Raman scattering, which is the version most commonly used in the study of materials and molecules, is very weak and only $\sim 10^{-12}$ of the incident photons are scattered and detected. Historically, it was first predicted by

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Smekal (1923) and was first demonstrated by Raman (1928) and by Landsberg and Mandelstam (1928). Several stimulated versions of Raman processes have been studied, including stimulated Raman scattering where the scattered beam is strong.

1. FEATURES OF RAMAN SCATTERING

1.1 Kinematics

The scattered frequency is $\omega_s = \omega_i \pm \omega_x$ for Stokes $(-)$ and anti-Stokes $(+)$ scattering, where ω_i and ω_x are the frequency of the incident light and the quantized excitation of the material (Fig. 1). With the conservation of momentum q, the scattered wave vector is $q_x = 2q_i \sin(\theta/2)$ when $\omega_i \gg \omega_x$, where θ is the scattering angle. In solids, when $\omega_x(q_x) = a_1 q_x$, as for acoustic phonons, ω_s is sensitive to the observed scattering angle. For excitations where $\omega_x(q_x) = \omega_x(0) + a_2 q_x^2$, as for optical phonons and plasmons, the frequency of scattered light is fairly independent of θ for near-zone-center scattering. Only near-zone-center (Γ) excitations, $q_x \sim 0$, contribute to first-order scattering in solids, where the incident photon interacts with only one excitation, because the change in photon momentum, $\sim 10^5$ cm^{-1}, is very small compared with the dimension of the Brillouin zone, $\sim 10^8$ cm^{-1}.

1.2 Dynamics of Scattering

The response of a material in the presence of an electric field E with frequency ω_i is described by the induced dipole moment $\mu = \alpha E$ (in molecules), where α is the electric polarizability, or the electric polarization $P = \epsilon_0 \chi E$ (in solids), where χ is the electric susceptibility. The total scattering cross section is $\sigma = (4\pi/3)\omega_i^4 \alpha^2/(4\pi\epsilon_0)^2 c^4$. Here χ (or α) can depend on a material excitation, such as a vibration, which is described in terms of a characteristic coordinate u that oscillates with frequency ω_x. To first order, χ can be expanded as $\chi(u) = \chi(u = 0) + (d\chi/du)u$. Classically, Raman scattering occurs because the polarization P (or μ) oscillates and therefore radiates at $\omega_i \pm \omega_x$. A vibrational transition in a molecule is Raman allowed if the polarizability varies with the normal coordinate of the vibrational mode (u), while it is allowed in absorption (IR or infrared active) if the dipole moment varies with the normal coordinate.

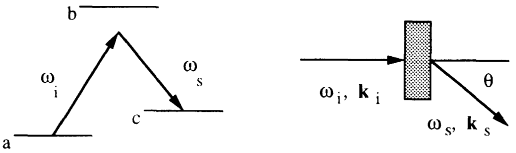

(a) (b)

FIG. 1. (a) Energy levels and (b) scattering geometry in spontaneous Raman scattering from state *a* to *c*, through intermediate state *b*. θ is the scattering angle.

In the macroscopic description of Raman scattering, the fluctuating components of the polarization lead to scattering. The spectral differential cross section for scattering photons from frequency ω_i to ω_s is (Hayes and Loudon, 1978)

$$\frac{d^2\sigma}{d\Omega\, d\omega_s} = A \frac{\omega_i \omega_s^3}{c^4} vV \frac{\eta_s}{\eta_i} |\mathbf{e}_s \cdot \chi' \cdot \mathbf{e}_i|^2 \langle u(\mathbf{q}) u^*(\mathbf{q}) \rangle_{\omega_s}, \tag{1}$$

where v is the scattering volume, V is the sample volume, η is the index of refraction, $\chi' = d\chi/du$ is the susceptibility derivative tensor, $\mathbf{e}_s$ and $\mathbf{e}_i$ are the polarization vectors of the scattered and incident beams, respectively, and $A = 1/16\pi^2$. The last term is the power spectrum of the fluctuations of the characteristic coordinate. It can be expressed in terms of quantum-mechanical second-quantization operators and then can be evaluated by use of the fluctuation-dissipation theorem. Since $\omega_s \sim \omega_i$, the frequency dependence for Raman scattering is similar to the $\sim \omega_i^4$ dependence for elastic scattering far below resonance. The rate of scattering photons is $(I_i/\hbar\omega_i) d^2\sigma/d\Omega\, d\omega_s$ where I_i is the intensity of the incident beam. A scattering efficiency S can also be defined as $S = \sigma/V$.

For phonons in solids, $\langle uu^* \rangle = (\hbar/2N\omega_p)[n(\omega_p) + 1]g(\omega_s)$ for Stokes scattering, where n is the phonon-occupation number, $1/[\exp(\hbar\omega_p/kT) - 1]$ in thermal equilibrium, ω_p is the phonon frequency, $g(\omega_s)$ is the normalized line shape whose linewidth is $\sim 1/$(phonon lifetime), and N is the number of scatterers in the volume. Using $\langle uu^* \rangle$ in Eq.

(1), the resulting V/N factor is the intensive parameter, the volume per species. The expression for anti-Stokes scattering is similar, with the corresponding anti-Stokes values used for ω_s, η, and χ in Eq. (1) and the bracket in the expression for $\langle uu^*\rangle$ replaced by $n(\omega_p)$. The final form of Eq. (1) is similar for other elementary excitations, which also have an $n + 1$ (n) factor for Stokes (anti-Stokes) scattering.

The microscopic description of Raman scattering uses quantum-mechanical time-dependent perturbation theory (Hayes and Loudon, 1978; Cardona, 1982). Photon interactions are mediated by either the $(e/mc)\mathbf{A}\cdot\mathbf{p}$ or the $(e^2/2mc^2)A^2$ term in the $(\mathbf{p} + e\mathbf{A}/c)\cdot(\mathbf{p} + e\mathbf{A}/c)/2m$ term in the Hamiltonian, where $\mathbf{p}$ is the electron momentum operator and $\mathbf{A}$ is the vector potential. The former term is important whenever interband transitions are important, as with phonons and magnons in solids and vibrations in molecules, while the latter term contributes when intraband transitions are important, as in the scattering of free electrons in vacuum. Sometimes both terms contribute, as in the scattering of free electrons in semiconductors. In molecules, $-e\mathbf{r}\cdot\mathbf{E}$, which is the same as $-\boldsymbol{\mu}\cdot\mathbf{E}$, is often used instead of $(e/mc)\mathbf{A}\cdot\mathbf{p}$, where $\mathbf{r}$ is the electron position operator.

Raman transitions from state 1, with irreducible representation Γ_1 of the space group of the crystal or molecule, to state 2 with representation Γ_2 are symmetry allowed if $\Gamma_2^* \times \Gamma_\chi \times \Gamma_1$ contains the totally symmetric representation. The term Γ_χ is the representation generated by the components of the second-rank polarization tensor χ. Consequently, Raman spectra are very sensitive to the polarization of the incident and detected scattered beams. The scattering arrangement is represented by $\mathbf{a}(\mathbf{b},\mathbf{c})\mathbf{d}$, where $\mathbf{a}$ and $\mathbf{d}$ are the directions of the incident and scattered beams, respectively, and $\mathbf{b}$ and $\mathbf{c}$ are their respective electric field polarizations. In solids, these can be related to the crystalline axes when the symmetry is at least orthorhombic. In backscattering geometry, $\mathbf{d} = \bar{\mathbf{a}}$. In a "polarized" spectrum $\mathbf{c} \parallel \mathbf{b}$, while in a "depolarized" spectrum $\mathbf{c} \perp \mathbf{b}$.

The depolarization ratio D is defined as the ratio of depolarized to polarized scattering intensities, and can indicate the symmetry and therefore the physical processes that induce the scattering. This ratio is a function of scattering angle θ and can be analyzed for incident polarization either parallel or normal to the scattering plane, which is defined by the wave vectors of the incident and scattered light. For randomly oriented species, such as the grains in polycrystalline material, molecules in most liquids, and rotating molecules in a gas, the polarization dependence must be averaged over angle. If the polarizability tensor of a given species is symmetric, then it can be diagonalized, with diagonal elements α_{xx}, α_{yy}, α_{zz}. This can be represented as a sum of two diagonal tensors, one isotropic with magnitude $a = (\alpha_{xx} + \alpha_{yy} + \alpha_{zz})/3$ and the other traceless that depends on the anisotropy factor b (or γ). When $\alpha_{xx} = \alpha_{yy} = \alpha_\perp$ and $\alpha_{zz} = \alpha_\parallel$, $a = (\alpha_\parallel + 2\alpha_\perp)/3$ and $b = |\alpha_\parallel - \alpha_\perp|$. After rotational averaging, $D(\theta = 90°) = 3b^2/(45a^2 + 4b^2) \leq \frac{3}{4}$ for incident light that is linearly polarized normal to the scattering plane, which gives an upper limit that is valid for elastic and inelastic scattering, and it is 1 for light polarized in the scattering plane. If the incident light is nonpolarized, $D(\theta = 90°) = 6b^2/(45a^2 + 7b^2) \leq \frac{6}{7}$. These upper boundaries do not apply for antisymmetric α, as in magnetic scattering. Depolarization ratios from randomly oriented species for different incident and scattered polarizations have been tabulated (Hirschfeld, 1973; Long, 1977).

The $|\mathbf{e}_s\cdot\chi'\cdot\mathbf{e}_i|^2$ term in Eq. (1) describes the polarization and angular patterns in Raman scattering for vibrational scattering and for electronic scattering that is mediated by changes in charge density. It is symmetric with respect to the exchange of the incident and scattered polarizations. In Raman scattering in magnetic materials and in electronic Raman scattering that involves spin effects or magnetic fields, the polarization term is $|(\mathbf{e}_i \times \mathbf{e}_s)\cdot\mathbf{e}_r|^2$, where $\mathbf{e}_r$ is a reference direction, such as that of the magnetic field. Raman scattering is then antisymmetric with respect to exchange of the incident and scattered polarizations.

Raman-scattering processes in solids can be divided into three areas: vibrational Raman scattering, which involves optical phonons (Sec. 3); magnetic scattering (Sec. 4), which involves collective processes specific to magnetic materials; and electronic scattering (Sec. 5), which involves transitions of free or bound electrons. In gases and liquids, vibrational Raman scattering involves excitation of molecular vibrations (Sec. 6). Additional features due to pure rotational Raman scattering appear in gases (Sec. 6).

In many materials, the inelastic light-scattering spectrum has peaks that can be assigned to excitations in either one of two or in any of the three categories. However, this division into vibrational, magnetic, and electronic scattering is not always precise. The scattering processes of these different excitations can interfere. Also, two different modes of excitation can directly couple, with either one mode or both modes having an intrinsic Raman cross section. For example, in some paramagnetic materials electron spin coupling with the lattice leads to sidebands on the optical phonon peaks (Ramdas and Rodrigues, 1988, 1991). In *p*-type Si, interference between vibrational and electronic scattering produces highly perturbed, non-Lorentzian Raman peaks (Cerdeira *et al.*, 1973). In terbium vanadate, complex Brillouin profiles are seen because the lower electronic levels of Tb^{3+} couple to acoustic phonons that propagate in specific directions (Harley *et al.*, 1980). Both excitations have scattering cross sections in each of these cases. In some solids, scattering of a mode is not allowed, but is still seen because of mode coupling. Vibrational Raman scattering is expected to be forbidden by symmetry in europium chalcogenides. It is seen only because of spin–phonon coupling (Guntherodt and Zeyher, 1984).

1.3 Nonlinear Raman Processes

Most Raman spectroscopic studies actually involve spontaneous Raman scattering, which has been discussed above. cw lasers are used, the scattered fields are weak, and in absorbing solids the intensity of the incident laser I_i is limited by laser heating. By using pulsed lasers, the incident intensity can be made much larger with minimal heating, and with these higher I_i nonlinear optical effects can become important, which can be classified as coherent and hyper Raman processes (Nibler and Knighten, 1979; Harvey, 1981; Druet and Taran, 1981; Vogt, 1982; Nibler and Yang, 1987; Levenson and Kano, 1988).

If three lasers are incident on a material, with electric fields E oscillating at frequencies ω_1, ω_2, and ω_3, the electric polarization can be written as

$$P_i(\omega) = \epsilon_0[\chi^{(1)}_{ij}E_j(\omega_1) + \chi^{(2)}_{ijk}E_j(\omega_1)E_k(\omega_2) + \chi^{(3)}_{ijkl}E_j(\omega_1)E_k(\omega_2)E_l(\omega_3) + \cdots]. \tag{2}$$

(with summation over repeated indices). The first term describes linear optical processes, while the higher-order terms detail nonlinear optical effects that are important only at strong fields. The second term describes processes such as second-harmonic generation and the electro-optic effect, and the third term describes several coherent Raman processes, which produce radiation at $\omega_s = \omega = \omega_1 \pm \omega_2 \pm \omega_3$ (Fig. 2), among other processes.

With one laser incident with frequency $\omega_1 = \omega_3$ $(= \omega_i)$, the $\chi^{(3)}$ term describes stimulated Raman scattering (SRS) at frequency $\omega = \omega_1 - \omega_3 + \omega_2 = \omega_2$ $(= \omega_s)$ [Fig. 2(a)]. There is laser gain at $\omega = \omega_1 - \omega_x$, with a gain coefficient proportional to the scattering cross section for spontaneous Raman scattering $d\sigma/d\Omega$ (Yariv, 1989). For sufficiently high gain, the scattered intensity I_s increases exponentially and there is lasing at ω. In vibrational Raman scattering, a coherent vibrational field oscillating at $\omega_x = \omega_p$ is created during SRS. This field can couple with ω_1 and ω_2 to create a series of higher-order Stokes and anti-Stokes laser lines at $\omega = \omega_1 \pm m\omega_p$, where m is an integer (Eimerl *et al.*, 1981). With the quantization of the electromagnetic fields, SRS can also be easily understood by time-dependent perturbation theory. The Raman-scattering rate is proportional to $n_s + 1$, where n_s is the number of photons at the scattered frequency ω_s per mode. The "1" term describes spontaneous Raman scattering, while the "n_s" term describes stimulated Raman scattering. Whenever two lasers are used to populate a state, say from state a to c through intermediate state b (as in Fig. 1) with lasers at frequencies ω_1 and ω_2, both sequential absorption (of ω_1 inducing $a \rightarrow b$) and stimulated emission (at ω_2 inducing $b \rightarrow c$), and the one-step, stimulated Raman process (directly $a \rightarrow c$) must be considered (Feld and Javan, 1969). Though stimulated Raman scattering is usually used to create new pulsed laser lines, it can also be used for spectroscopy by measuring the Raman amplification of a tunable

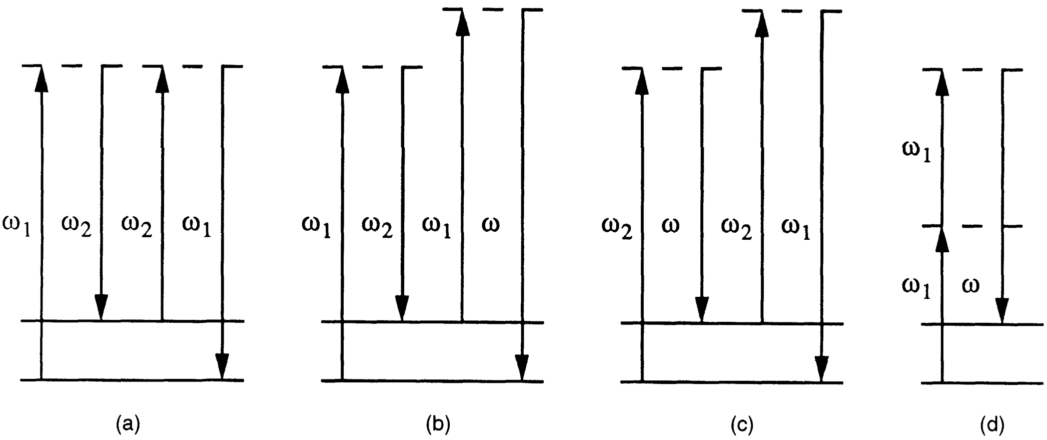

FIG. 2. Energy-level diagrams for (a) stimulated Raman scattering (SRS), $\omega_2 = \omega_1 - \omega_x$; (b) coherent anti-Stokes Raman scattering (CARS), $\omega = 2\omega_1 - \omega_2$; (c) coherent Stokes Raman scattering (CSRS), $\omega = 2\omega_2 - \omega_1$; and (d) hyper-Raman scattering (HRS), $\omega = 2\omega_1 - \omega_x$. ω_x corresponds to the excitation energy between the lowest two levels shown here.

laser in a medium that is irradiated by a strong, fixed-wavelength pump laser. The ($\omega = \omega_1 - \omega_2 + \omega_3$)-type term in Fig. 2(a) is also responsible for Raman-induced Kerr-effect spectroscopy (RIKES).

In coherent anti-Stokes Raman scattering (CARS), two pulsed laser beams at ω_1 and ω_2 ($\omega_1 > \omega_2$) intersect in a material to produce coherent light at $\omega = 2\omega_1 - \omega_2$ [Fig. 2(b)]. Usually at least one of these two lasers is tunable. CARS can be viewed as the scattering of ω_1 from the coherent oscillation produced by $\omega_1 - \omega_2 \sim \omega_x$. Spectroscopy is performed by tuning one of the incident lasers and by measuring its wavelength at resonance ($\omega_1 - \omega_2 \sim \omega_x$). This is quite different from spontaneous Raman spectroscopy where the wavelength of the weak scattered light is measured. In CARS, the interference of resonant and constant, nonresonant terms often produces an antiresonant line shape, instead of the typical Lorentzian-type resonant line shape seen in spontaneous Raman scattering. An analogous process, coherent Stokes Raman scattering (CSRS), involves the processes shown in Fig. 2(c).

Phase matching is necessary for efficient conversion in nonlinear optics. It is automatically satisfied in SRS and RIKES, but must be met by experimentally matching wave vectors of the incident laser beams in CARS and CSRS.

Just as spontaneous Raman scattering is induced by the dependence of $\chi^{(1)}$ on material excitation u, the higher-order susceptibility terms $\chi^{(2)}$ and $\chi^{(3)}$ can also depend on u, and can be expanded similarly. For example, $\chi^{(2)}(u) = \chi^{(2)}(u = 0) + (d\chi^{(2)}/du)u$. The $(d\chi^{(2)}/du)u$ term causes inelastic scattering at frequencies such as $2\omega_1 - \omega_x$ [Fig. 2(d)]. Such processes are known as spontaneous hyper-Raman scattering. It is extremely weak, even with high incident laser intensities, except when resonantly enhanced. Just as $\chi^{(1)}$ has the molecular polarizability analog α, which leads to $d\alpha/du$ terms, $\chi^{(2)}$ and $\chi^{(3)}$ have the molecular polarizability analogs β and γ, respectively, which lead to $d\beta/du$ and $d\gamma/du$ terms.

Hyper-Raman scattering has different symmetry constraints than does Raman scattering. Generally, infrared-active modes are hyper-Raman active, and so first-order hyper-Raman scattering can be allowed while first-order Raman scattering is forbidden. "Silent" modes that are both Raman and infrared forbidden can also be seen by hyper-Raman scattering. There are also stimulated versions of hyper-Raman scattering.

2. INSTRUMENTATION

Raman spectroscopy requires a light source, optics to deliver the light to the sample, optics to collect the scattered light, an instrument to provide spectral dispersion of the scattered light and to remove stray light, and a photon detector. Because the spontaneous scattering cross section is small, high-intensity light sources are essential in Raman scattering. Historically, Raman spectroscopy became relatively simple only after the development of lasers. Most studies of spontaneous Raman scattering use cw lasers such as argon-ion (4880, 5145 Å), krypton-ion (6471 Å), or dye lasers. Lasers with ~ns, ~ps, or ~multi-fs-long pulses, such as excimer lasers, Nd/YAG lasers, and pulsed and mode-locked dye lasers, are needed for nonlinear Raman scattering and time-resolved Raman spectroscopy. Time-resolved scattering is used when there is interference by high levels of background cw light and for the dynamical study of nonequilibrium distributions of phonons and carriers (Kash and Tsang, 1991). For adequate spectral resolution, the linewidth of the laser must be narrower than the linewidth of the elementary excitation, which for phonons is $\sim 2\ \mathrm{cm}^{-1}$. Tunable lasers, such as dye and Ti sapphire lasers, are useful in the study of resonance Raman scattering.

In Raman analysis of solids, the laser is often focused to a circular spot on the sample. In Raman microprobe scattering, the spot size can approach the far-field diffraction limit for visible light, ~0.5 μm. In absorbing samples, significant laser heating must be avoided. With cw laser irradiation, the temperature rise at the surface of a very strongly absorbing sample is $\Delta T = P_i/(2\sqrt{\pi}Kw)$, where P_i is the laser power transmitted into the sample, K is the thermal conductivity, and w is the laser spot size (half-width at $1/e$ intensity); this consideration puts a practical upper limit on P_i. Because the scattered light is imaged onto the slit of a spectrometer, the power of the incident laser can be increased, with no increase in heating, by focusing the laser to a line with a cylindrical lens, rather than to a circular spot.

When studying optically thick samples or thin films, experiments are usually performed in a backscattering configuration, with a scattering angle $\theta \sim 180°$. The effective scattering efficiency S^* is then

$$S^* = S\,\frac{1 - \exp[-(S + \alpha_i + \alpha_s)L]}{S + \alpha_i + \alpha_s} \times (1 - R_i)(1 - R_s), \tag{3}$$

where α_j and R_j are the absorption coefficient and (intensity) reflectance at ω_j ($j = i,s$). When $S \ll \alpha_i,\alpha_s$, the effective depth of the Raman probe is $\sim 1/(\alpha_i + \alpha_s)$, and the scattering volume v in Eq. (1) is $\sim w^2/(\alpha_i + \alpha_s)$. With transparent media, alternative geometries include side scattering ($\theta \sim 90°$) and in-line scattering ($\theta \sim 0°$), and the laser may not be focused at all. In any case, light is collected by small-f-number optics and directed, sometimes by optical fibers, to a spectrometer that must disperse the scattered light and reject any collected light from the incident laser.

It is common to collect scattered light at $\theta \sim 90°$ when studying gases and liquids. In Raman scattering of gases in cells that are external to the laser cavity, the exciting laser beam is often focused and a collecting mirror is put on the side opposite the collecting lens to collect also light scattered to $\theta \sim -90°$. To increase the incident laser intensity, either a multipass arrangement is used or the scattering cell is placed inside the laser cavity (Weber, 1979).

Triple-stage spectrometers, which have a subtractive double stage for rejecting the laser line followed by a single stage for dispersion, and double monochromators, with two additive stages, are commonly used for dispersion. An alternative approach is to use a single-stage spectrometer that is preceded by a narrow line filter to reject the laser line. This approach can have the relative advantage of greater throughput because of fewer diffraction gratings and mirrors. Improvements in laser filters, notably holographic notch filters, have lessened one earlier disadvantage of this approach, the high losses for light scattered within $\sim 200\ \mathrm{cm}^{-1}$ of the laser line; however, the lack of tunability of these filters is still a drawback. While monochromators can be used to examine line shifts as small as $\sim 5\ \mathrm{cm}^{-1}$, scanning Fabry–Pérot interferometers are needed to resolve the smaller line shifts seen in Brillouin scattering, sometimes with multipassing to reduce the level of stray light.

The classic mode of photon detection in spontaneous Raman scattering is to scan the spectrometer and to count photons with a photomultiplier. Recently, parallel detection of the dispersed radiation by multichannel detectors, such as intensity-enhanced photodiode arrays, imaging photomultipliers (which are also called mepsicrons), and, notably, charge-coupled-device (CCD) detectors (see CHARGE-COUPLED DEVICES) have greatly improved the signal-to-noise ratio in Raman spectra (Chang and Long, 1982; Tsang, 1989).

Spontaneous Raman spectra can be improved by increasing the flux of incident photons, improving photon collection, and using better detectors. Since the Raman cross section varies as ω^4, the use of shorter wavelengths improves signal strength, as does tuning ω_i to a resonance in the cross section (Cardona, 1982), the use of interference effects (Connell *et al.*, 1980), and the use of surface-enhanced Raman scattering on roughened metal surfaces (SERS) (Otto, 1984; Campion, 1985; Jha, 1985). SERS can increase Raman signals by $\sim 10^4$–10^6 and is caused by electromagnetic resonances and by other effects related to chemisorption. Because Raman processes are not surface selective, the measurement of monolayers (Campion, 1985) and very thin films (Tsang, 1989) on surfaces requires these signal-enhancing methods, and other techniques as well, such as the subtraction of the Raman signal from the substrate.

Because absolute measurement of the Raman-scattering cross sections is difficult, such measurements can be made by comparison with a standard, such as diamond or Si for solids (Cardona, 1982). The differential scattering efficiency $dS/d\Omega$ of diamond (at 2.41 eV, 1333 cm^{-1} Raman shift) and Si (at 1.9 eV, 521 cm^{-1} shift) are (6.5 ± 0.8) and $(168 \pm 50) \times (10^{-5}/\text{sr m})$, respectively. For gases, Q-band vibrational Raman scattering in gas-phase N_2 (2331 cm^{-1}) is the standard, with $d\sigma/d\Omega = (5.05 \pm 0.1) \times 10^{-48}(\tilde{\nu} - 2331\ \text{cm}^{-1})^4\ \text{cm}^6/\text{sr}$ (Schrotter and Klockner, 1979). Here $\tilde{\nu}$ is the laser wave number (in the visible and in the ultraviolet to 330 nm) expressed in cm^{-1}. For liquids, vibrational scattering in CCl_4 ($\tilde{\nu}_1$ at 459 cm^{-1}), C_6H_6 (992 cm^{-1}), and cyclohexane (C_6H_{12}) (802 cm^{-1}) are often used as standards with $(d\sigma/d\Omega)(\tilde{\nu} - \tilde{\nu}_m)^{-4}$ equal to 140, 225, and 70 respectively in units of $10^{-48}\ \text{cm}^6/\text{sr}$ (Schrotter and Klockner, 1979). The term $\tilde{\nu}_m$ is the cited vibrational quantum energy expressed in cm^{-1}.

Since Raman scattering is relatively inefficient, background photoluminescence (or fluorescence) must be kept to a minimum. Photoluminescence is not a concern when the photon energy is less than the fundamental gap E_0, and in materials that do not fluoresce at all, such as indirect semiconductors. Though scattering near the fundamental gap E_0 is inadvisable in direct semiconductors, Raman scattering can be resonantly enhanced at the associated spin–orbit-split gap $E_0 + \Delta$ or at larger gaps, such as E_1, without a strong photoluminescence background. In some materials, particularly organic compounds, the fluorescence intensity tends to increase with ω_i, thus rendering useless the ω^4 improvement in the Raman signal. Near-infrared light, such as that from Nd/YAG lasers (1.06 μm), is used to avoid fluorescence. Because optical detectors in the near infrared perform poorly, especially when employed with single-channel detection, an alternative method has developed, Fourier-transform Raman spectroscopy (Hendra *et al.*, 1991). In Fourier transform (FT) Raman, spectral dispersion and parallel detection are accomplished by using a scanning Michelson interferometer.

3. VIBRATIONAL RAMAN SCATTERING IN SOLIDS

3.1 Crystals

First-order vibrational Raman scattering is induced by the oscillation of the polarizability at vibrational-mode frequencies (Pinczuk and Burstein, 1975; Hayes and Loudon, 1978; Cardona, 1982). In nonlinear molecules with m atoms, there are $3m - 6$ vibrational modes ($3m - 5$ for linear molecules). In crystals with m atoms per primitive cell, there are $3m$ modes or branches, the three acoustic modes that can be observed in Brillouin scattering and $3m - 3$ optic modes. For relatively simple crystal structures, the vibration of the mode is parallel to the direction of propagation in $m - 1$ of these optic-mode branches [the longitudinal optical (LO) phonons], while in the other $2m - 2$ optical branches the vibration is transverse (TO). Symmetry determines whether these modes are Raman active (i.e., have a nonzero cross section for first-order scattering), infrared (IR) active (i.e., have a nonzero infrared-absorption cross section), Raman and IR active, or both Raman and IR inactive (silent). In centrosymmetric crystals and molecules, modes can be only Raman active, IR active, or silent. This is a consequence of the definite parity of vibrations in a centrosymmetric medium with respect to inversion. The matrix element of either the polarizability, which induces Raman scattering and transforms as *xx*, *xy*, *xz*, etc. (second-rank tensor), or the dipole moment, which induces IR absorption and transforms as *x*, *y*, or *z* (a vector), must vanish between vibrational states. In general, the polarization properties of Raman scattering and the number of modes that are observed can be used to determine the crystal symmetry.

The frequency of each branch $\omega_{p,i}$ is a function of wave vector $\mathbf{q}$ in the first Brillouin zone and is given by dispersion curves (see PHONONS IN CRYSTAL LATTICES). Because of momentum conservation, only phonons with $q < 10^5$ cm^{-1}, near zone center, can participate in one-phonon Raman scattering. The entire Brillouin zone can be studied by higher-order Raman processes and by neutron diffraction.

Phonon-dispersion curves can be determined by lattice-dynamics calculations with the aid of Raman and other experimental data, and at times *ab initio* electronic energy calculations (see PHONONS IN CRYSTAL LATTICES). This is relatively simple only when there are few atoms per unit cell. In more complex systems, some guidance is provided by symmetry arguments, which can determine the number and symmetry of Raman-active modes. Simplifications occur in molecular solids, where there are higher-frequency internal vibrations within the molecular group, which can be correlated with the Raman spectrum of the associated free molecule, and lower-frequency modes due to external vibration modes between these groups, librations (hindered rotation of the molecular units), and acoustic modes. Raman scattering from internal vibrations in molecular solids is similar to vibrational scattering in liquids (Sec. 6).

The symmetries and Raman-scattering tensors for Raman-active optical-phonon modes are well tabulated for $q = 0$ phonons (Hayes and Loudon, 1978; Cardona, 1982). The Raman signal varies as $|\mathbf{e}_s \cdot \mathbf{R} \cdot \mathbf{e}_i|^2$, where $\mathbf{R}$ is the Raman-scattering tensor ($\propto \chi'$), which has the symmetry of the scattered phonon, and is very sensitive to polarization. The tensor $\mathbf{R}$ is usually symmetric, though there are some cases where it is antisymmetric, such as the A_2 modes of Se, Te, and quartz.

In diamond-structure materials (O_h or $m3m$ point group), there is a single optic mode [$\Gamma_{25'}$ or Γ_5^+ (the Γ or Bethe notation), or T_{2g} (Mulliken notation) symmetry], which is threefold degenerate at zone center (LO, 2TO); it is Raman active. In backscattering, LO scattering is allowed from (100) and (111) surfaces, while TO scattering is allowed from (110) and (111) surfaces. In diamond, Si, and Ge, the LO/TO mode produces first-order Raman peaks at 1333, 521, and 301 cm^{-1}, respectively (Fig. 3). In zincblende materials, such as GaAs and ZnSe, the symmetry is T_d ($\bar{4}3m$), and the optical phonon [Γ_{15} (Γ_5) or T_2] has the same polarization selection rules as for diamond. In the closely related wurtzite structure (C_{6v}, $6mm$), as in CdS, the $A_1 + E_1 + E_2$ (or $\Gamma_1 + \Gamma_5 + \Gamma_6$) modes are Raman active. (The E modes are doubly degenerate.) In the rutile structure D_{4h} or $4/mmm$, as in TiO$_2$, MgF$_2$, and MnF$_2$, there are four Raman-active modes, $A_{1g} + B_{1g} + B_{2g} + E_g$ (or $\Gamma_1^+ + \Gamma_3^+ + \Gamma_4^+ + \Gamma_5^+$). The four Raman spectra of MnF$_2$ in Fig. 4 show how different Raman-active modes can be seen by using different polarization configurations.

High-T_c superconductors have the perovskite structure (Feile, 1989; Ferraro and Maroni, 1990; Thomsen, 1991). Tetragonal La$_{2-x}$(Ba/Sr)$_x$CuO$_4$ ($T_c \sim 40$ K) has D_{4h}^{17} or I_4/

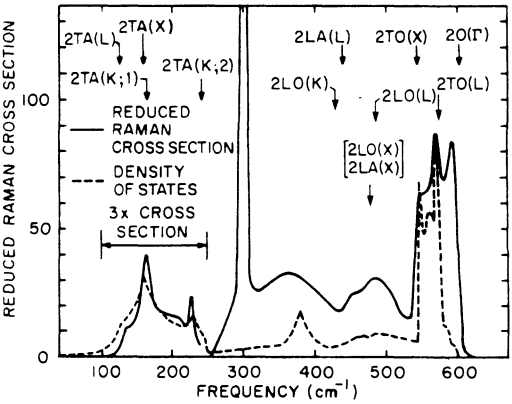

FIG. 3. Raman spectrum of Ge, showing the first-order peak near 301 cm^{-1}, which goes off scale, and the second-order peaks, which look similar to the also plotted normalized density of states (Weinstein and Cardona, 1973).

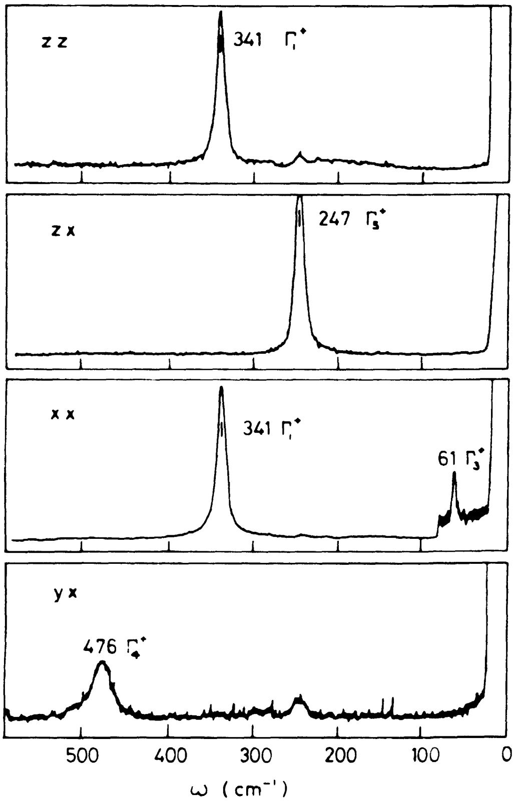

FIG. 4. Raman spectra of MnF_2 taken with different polarization configurations, in which different Raman-active modes are seen (Porto *et al.*, 1967).

mmm symmetry, and the Raman-active modes $2A_{1g} + 2E_g$. The lower-temperature, orthorhombic phase, D_{2h}^{18} or *Cmca*, has Raman-active modes $5A_g + 4B_{1g} + 3B_{2g} + 6B_{3g}$. Another high-$T_c$ material, $YBa_2Cu_3O_x$, has T_c as high as ~95 K when $x \sim 6.5$–7. With $x = 6$, it is tetragonal (D_{4h}^1 or P_4/mmm) and has Raman-active modes $4A_{1g} + B_{1g} + 5E_g$. For $x = 7$, it is orthorhombic (D_{2h}^1 or *Pmmm*) with Raman-active modes $5A_g + 5B_{2g} + 5B_{3g}$. The frequency of one Raman peak in $YBa_2Cu_3O_x$ increases linearly with oxygen content, increasing from ~475 cm^{-1} for $x = 6$ to ~501 cm^{-1} for $x = 7$.

Many metals have only one atom per primitive cell, and therefore have no optic modes to observe with Raman scattering. However, hexagonal close-packed metals have two atoms per cell and the Γ_6^+ (E_{2g}) optic mode is Raman active. In Be, Mg, and Zn, this mode leads to shifts of 455, 122, and 72 cm^{-1}, respectively. Bi has rhombohedral structure and two Raman-active modes A_{1g} and E_g near 70 and 96 cm^{-1}. The Raman spectra of these metals are weak because of the large absorption coefficient of the metal.

At zone center, LO and TO modes are degenerate for nonpolar materials. In partially ionic crystals, the zone-center frequency of the LO mode exceeds that of the doubly degenerate TO mode because it couples with the longitudinal electric field that is formed during LO vibration. Such LO vibrations are called polar modes, while those modes that do not couple with a macroscopic electric field, such as these TO vibrations and all optical modes in nonpolar crystals, are nonpolar modes. In cubic crystals, $\omega_{LO}^2 - \omega_{TO}^2 = 4\pi n_d e^{*2}/\epsilon_\infty \mu_m = \omega_{TO}^2(\epsilon_0/\epsilon_\infty - 1)$, where e^* is the ionic charge (also termed the Born transverse dynamic effective charge), n_d is the density of cation–anion pairs and μ_m is their reduced mass, and ϵ_∞ and ϵ_0 are the optical and static dielectric constants. For example, in GaAs the LO and TO modes are at 292 and 260 cm^{-1}, respectively, while in ZnSe they are at 250 and 205 cm^{-1}. In uniaxial crystals, such as the wurtzite structure, the TO further splits into two peaks at zone center when the phonon **q** is not along the *c* axis, leading to three nondegenerate optic phonons that can be observed in Raman scattering (Hayes and Loudon, 1978). The mode structure is even more complex in biaxial crystals.

At small scattering angles, nearly forward scattering, transverse (TO) phonons and the transverse electromagnetic radiation form a coupled excitation, known as phonon polaritons. This excitation looks like an unperturbed photon and a phonon at $\omega_{LO}(\Gamma)$ at $\theta \sim 0°$, a strongly coupled excitation for ~1–3° where the uncoupled photon and phonon-dispersion curves would have crossed, and an unperturbed photon and TO phonon at ω_{TO} at larger scattering angles, as is in backscattering geometry. Because of this interaction, the TO and LO modes are actually degenerate exactly at zone center. In addition to these bulk polaritons, there are surface polaritons that propagate along boundaries and exponentially decay into the bulk (Ushioda and Loudon, 1982).

Quantum mechanically, first-order vibrational Raman scattering in solids is a three-step process. First the incoming photon ω_i creates a virtual electron–hole pair through an intermediate level. Then the electron or hole interacts with the lattice, leading to the creation or annihilation of a phonon. Finally, the electron and hole recombine to form the scattered photon ω_s, leaving the material in the initial electronic state. Third-order perturbation theory leads to

$$\mathbf{e}_s \cdot \chi' \cdot \mathbf{e}_i \propto \frac{\langle v|\mathbf{p}\cdot\mathbf{e}_s|c_2\rangle\langle c_2|\mathrm{H}_{\mathrm{e-p}}|c_1\rangle\langle c_1|\mathbf{p}\cdot\mathbf{e}_i|v\rangle}{(\omega_i \pm \omega_p - \omega_0)(\omega_i - \omega_0)}, \tag{4}$$

where v is the initial valence band, c_1 and c_2 are the intermediate conduction-band states, $\mathrm{H}_{\mathrm{e-p}}$ is the electron–phonon interaction Hamiltonian, $\hbar\omega_0$ is the band gap, and $+/-$ correspond to anti-Stokes/Stokes transitions. The $\mathbf{A}\cdot\mathbf{p}$ term induces the first and third steps, which involve interband transitions. If the second step, the electron–phonon interaction, connects two different electron–hole intermediate states ($c_1 \neq c_2$ for electron scattering), which is an interband step, then three electronic states are involved and it is called a three-band process. If the same electron–phonon interaction does not change the electronic state ($c_1 = c_2$), an intraband intermediate step, it is a two-band process.

One effect that contributes to the electron–phonon coupling for both nonpolar and polar vibrations is the change in the electronic potential energy with the lateral displacement of atoms in a unit cell. This is characterized by the deformation potential and is short-range in nature. Lattice vibrations modulate the electronic energy structure, which modulates the electric susceptibility to give terms like $d\chi/du$ due to interband transitions, where u is the phonon coordinate. A second effect contributes only for polar vibrations, and it is caused by the modulation of χ by the quasistatic E_m, the macroscopic electric field, and is longer-range in nature. This interaction is described by the Fröhlich interaction, and has two components, a q-independent three-band interband term, which is allowed and can be characterized by the electro-optic tensor, and a q-dependent two-band intraband term, which is dipole forbidden away from resonance. The origin of the electro-optic term is the nonlinear polarization term $\chi^{(2)}E_iE_m$, as in Eq. (2), which can be expressed as a change in the linear susceptibility $(d\chi^{(1)}/dE_m)E_m$ with $d\chi^{(1)}/dE_m = \chi^{(2)}$.

Macroscopically, vibrational Raman scattering occurs as a result of fluctuations in u and E_m, respectively, in these two mechanisms. The ratio of the deformation potential and electro-optic contributions to this electron–phonon coupling is given by the Faust–Henry coefficient C. For most polar zinc-blende semiconductors, $C < 0$ and its magnitude is <1. Therefore, usually these two effects constructively interfere and the electro-optic effect is the stronger of the two. This means that LO scattering is usually stronger than TO scattering.

The interband terms in Eq. (4) have ingoing and outgoing resonances when the ω_i or the scattered photon frequency $\omega_s = \omega_i \pm \omega_p$ is near an energy gap. In some cases, these peaks can be resolved, though often there is one strong resonant Raman peak near $\omega_i \pm \omega_p/2$. By treating a classically bound electron, one can see the interactions in this resonance because $\chi = AF/(\omega_0^2 - \omega_i^2 - i\omega_i\gamma)$, where A is a constant, F is the oscillator strength, ω_0 is the resonance frequency, and γ is a damping (relaxation) term. Contributions to $\chi' = d\chi/du$ come from the $d\omega_0/du$ term, which is a two-band term due to diagonal matrix elements of the electron–phonon interaction, and the dF/du term, which is the three-band term due to off-diagonal elements. Integration of these highly resonant expressions [Eq. (4)] over the density of states gives line shapes that are finite at the resonances (Cardona, 1982). Occasionally, antiresonance line shapes are seen because of the interference of the resonance term with a nonresonant, constant background, as in TO phonon scattering in wurtzite ZnS near 1.9 eV.

In resonance vibrational Raman scattering, higher-order terms can also be important that can induce strong peaks not seen in the nonresonant Raman spectrum (Cardona, 1982). For example, there can be a breakdown in the symmetry properties for LO phonon scattering, so that LO phonons may be seen in a Raman spectrum for a given polarization configuration when $\hbar\omega_i$ is near a resonance, while it is forbidden off resonance. Also, in some cases multiple LO phonon lines can be seen at $\omega_s = \omega_i \pm n\omega_{\mathrm{LO}}$ with $n = 2, 3, 4, \ldots$. Since the intraband Fröhlich interaction depends on q, it does not contribute to nonresonant first-order scattering, but at resonance it can make "forbidden" first-order and higher-order scattering quite strong. Mechanistically, such multiphonon spectra can be attributed either to resonant Raman scattering or to hot luminescence, which is photoluminescence of highly excited electrons and holes that is assisted by the emission or absorption of one or more phonons.

The first step in vibrational Raman scattering is usually the formation of uncorrelated virtual electrons and holes. However, when there is strong excitonic structure, the first step may involve the production of a virtual exciton instead of an electron–hole pair. This is important in vibrational Raman scattering in Cu_2O where odd-parity forbidden Γ-point phonons are observed because of resonances with direct forbidden (quadrupole-allowed) excitons.

Second-order vibrational Raman scattering is due to the scattering of two phonons that shift ω_i by $\omega_{p1}(\mathbf{q}_1) \pm \omega_{p2}(\mathbf{q}_2)$. The phonons can be from the same branch ($p_1 = p_2$, overtones) or different branches ($p_1 \neq p_2$, combinations). Scattering cross sections for two-phonon scattering are much smaller than for first-order scattering because the higher-order susceptibility term $(d^2\chi/du_1du_2)u_1u_2$ leads to the scattering, where u_1 and u_2 are the normal-mode coordinates for the two modes. Since momentum can be conserved with contributions away from zone center, as long as $\mathbf{q}_1 + \mathbf{q}_2 \sim \mathbf{0}$, two-phonon Raman peaks are broad and resemble the joint density of states. In Stokes scattering, there is an $(n_1 + 1)(n_2 + 1)$ dependence on phonon density. The Raman peak near $2\omega_{\mathrm{LO/TO}}(q = 0)$, for example in Ge and Si, resembles the optical-phonon density of states scaled in frequency by 2, and has contributions from $\mathbf{q}$ and $-\mathbf{q}$ across the Brillouin zone. This is seen in Fig. 3 for Ge. Since $\omega(q)$ decreases from zone center for these branches, the overtone peaks below $2\omega_{\mathrm{LO/TO}}(q = 0)$. Two-phonon overtones are usually symmetry allowed for all modes, and they dominate the completely symmetric component of the spectrum Γ_1 since first-order vibrational scattering is not completely symmetric. In rocksalt crystals, such as NaCl and KBr, the optic mode is Raman inactive in first order and can only be seen in second order.

3.2 Disordered Structures

This discussion has been limited to large, pure, perfectly ordered crystals. New features in the Raman spectrum appear in crystals that are finite, impure, or disordered (see VIBRATIONAL STATES IN DISORDERED SYSTEMS). For example, when polycrystalline materials are examined with a laser focused to a spot size smaller than the grain size, the polarization selection rules can be used to determine grain orientation (Hopkins and Farrow, 1986). The polarization properties are randomly averaged over angle, as described earlier, when the grain size is smaller than the laser spot size. Further, if the grain size L is smaller than ~200 Å, phonon confinement within the grains leads to nonconservation of q, and phonons from zone center to $q \sim 1/L$ participate in Raman scattering. In Si, the Raman peak for such microcrystalline material is downshifted in frequency and asymmetrically broadened relative to that of the crystal (Campbell and Fauchet, 1986). The first-order Raman spectrum of amorphous materials resembles the optical-phonon density of states in the amorphous material, which is very similar to that of the crystal, because of the total breakdown of q conservation. This has been studied in many solids, including tetrahedrally bonded semiconductors (Si, Ge, GaAs), oxides, and chalcogenides (Brodsky, 1975). The effects of nonconservation of wave vector can also be seen in metals and highly opaque semiconductors, as a result of the large imaginary part of the photon wave vector in the material.

Point defects lead to new structure in the Raman spectrum (Srivastava, 1990). There can be local modes at frequencies above those of the host when the added impurities are lighter than the host atoms, gap modes in the frequency gaps of the host, or resonance modes within the spectrum of the host. Host modes that are forbidden in Raman scattering of the perfect crystal may appear because the defects lower crystalline symmetry.

At very high point-defect densities, the material is called a mixed crystal. Raman spectra of $A_xB_{1-x}C$ mixed crystals, such as $Ga_{1-x}Al_xAs$, often exhibit two-mode behavior with LO and TO phonons corresponding to both AC-like and BC-like vibrations, with frequencies that vary with x (Srivastava, 1990).

At $x \sim 0$ and 1, there are three peaks, the LO and TO modes of the pure crystal and that of the point defect. Some mixed crystals show one-mode behavior (a single LO and TO pair) or mixed one-mode/two-mode behavior. The polarization selection rules in these mixed crystals are relaxed from those of the *AC* and *BC* crystals because of chemical disorder, i.e., though the *A*/*B* sites have perfect crystalline order there is no *A*/*B* ordering at these sites. This can be distinguished from site disorder in microcrystalline and amorphous materials. The Raman spectrum of $Ge_{1-x}Si_x$ alloys is quite different from these other mixed crystals. It has three peaks, due to local Si–Si, Ge–Si, and Si–Si vibrations.

3.3 Two-Dimensional Structures

Raman scattering can be used to examine novel phonon features that are present in two-dimensional, layered heterostructures. This has been studied extensively for zincblende semiconductors (Klein, 1986; Jusserand and Cardona, 1989; Menendez, 1989). These heterostructures include multiple quantum wells and superlattices where very thin layers of materials *A* (e.g., GaAs) and *B* (e.g., AlAs) alternate, as in $A_{na}B_{nb}A_{na}B_{nb} \cdots$ where n_a monolayers of *A* of total width d_a (monolayer width t_a) alternate with n_b monolayers of *B* of total width d_b (monolayer width t_b). This superperiodicity lowers the symmetry, from cubic (T_d) to tetragonal (D_{2d}) for zinc-blende structures, and increases the number of atoms in the primitive cell, from 2 to $2(n_a + n_b)$.

If the frequency spans of the optic-phonon branches in bulk *A* (or *B*) do not overlap those in *B* (*A*), then optical phonons propagating along the growth direction are confined to the *A* (*B*) layer. Raman scattering will then excite standing waves in the *A* (*B*) layers corresponding to confined phonons at $\omega(m) = \omega_{\text{bulk}}(q_m)$ where $q_m = m[\pi/(n_i + 1)t_i]$ with $i = a$ (*b*), rather than at $\omega_{\text{bulk}}(q = 0)$. In $(GaAs)_{na}(AlAs)_{nb}$, optical phonons are confined in both the GaAs and the AlAs layers, as is seen in the spectra in Fig. 5, which were taken out of resonance. The $z(x,y)\bar{z}$ spectra are due to odd-*m* LO phonons, which have B_2 symmetry, while the $z(x,x)\bar{z}$ spectra are due to even-*m* LO phonons, which have A_1 symmetry. In the bulk semiconductor, phonons are seen only in $z(x,y)\bar{z}$ configuration when the laser is out of resonance.

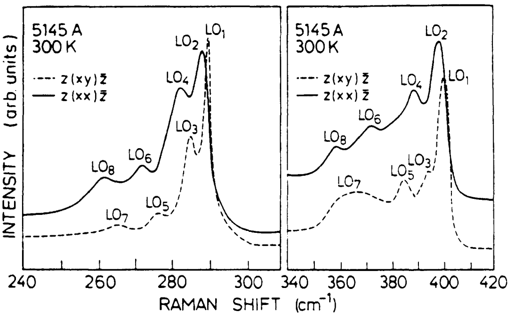

FIG. 5. Raman spectrum of $(GaAs)_8(AlAs)_8$ superlattice taken out of resonance, both for optical phonons confined to the (a) GaAs and (b) AlAs layers (Wang *et al.*, 1988).

When optical phonons propagate with a component of *q* parallel to the interface, the mode is confined near the interface. At the *AB* interface, these localized phonons, known as interface phonons, can be excited and seen in Raman scattering.

Because of the repetitive nature of the superlattice, the Brillouin-zone boundaries are at $\pm\pi/D$, where $D = d_a + d_b$, rather than at $\pm\pi/t_i$ $(i = a,b)$ for the bulk solid. This effectively leads to zone folding of the bulk dispersion curve for acoustic phonons propagating along the superlattice growth direction. In addition to the Brillouin peaks, Raman peaks now appear at $\omega_x(q)$ where $q = \pm q_{\text{SL}} + 2\pi m/D$ and $m = 1, 2, 3 \cdots$; $q_{\text{SL}} = (\omega/d)[d_a/v_a - d_b/v_b]$, where *v* is the sound speed in the respective medium. The Raman spectrum is a series of Stokes and anti-Stokes doublets with intensities that decrease with *m*; in Fig. 6 the Stokes doublets are seen for a GaAs/AlAs superlattice. These doublets are far enough away from the exciting line, ~10 cm^{-1}, to be seen with a monochromator (unlike bulk acoustic phonons), and can be seen in crystalline and amorphous superlattices.

3.4 Perturbations Due to Strain, Changes in Temperature, and Electric Fields

First-order vibrational Raman scattering is sensitive to the effects of strain on materials (Weinstein and Zallen, 1984). The appli-

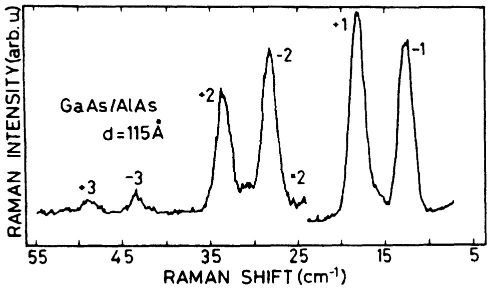

FIG. 6. Raman spectrum of folded acoustic lines in a GaAs/AlAs superlattice (Jusserand *et al.*, 1987).

cation of hydrostatic pressure (P_h) in diamond anvil cells causes uniform compressive strain and an increase in ω_p that is characterized by the mode Grüneisen parameter $\gamma = -d \ln\omega_p/d \ln V = (B/\omega_p)d\omega_p/dP_h$, where B is the bulk modulus and V is volume. The parameter γ for the TO mode at zone center increases with effective charge per cation electron for zinc-blende semiconductors, from $\gamma = 1.0$ for covalent Si to 1.9 in ZnS, while the increase of γ_{LO} with ionicity is relatively slower. By examining the pressure dependence of $\omega_{\mathrm{LO}} - \omega_{\mathrm{TO}}$ for polar modes, the change of the transverse effective charge e^* with volume can be measured. Since the two-phonon vibrational Raman spectrum has peaks from contributions near the critical points in the Brillouin zone, pressure measurements can give γ at these points.

The application of uniaxial pressure causes uniaxial/biaxial strain that splits the degenerate zone-center optical phonons (Cerdeira *et al.*, 1972; Pollak, 1991). For example, stress along the [001] and [111] directions in Si splits the triply degenerate Γ-point optical phonon into a singlet vibrating along the stress and a doublet vibrating perpendicular to it, while stress along [110] splits it into a triplet. These individual components can be seen by using different polarization configurations. Raman spectroscopy can also be used to measure biaxial strain in thin films with no applied pressure, such as in lattice-mismatched, strained-layer heterostructures that are grown commensurately by molecular beam epitaxy or MOCVD. In incommensurately grown films, film contraction in cooling from the growth temperature down to the observation temperature produces strains that can be analyzed by Raman scattering, as has been seen for Si films grown on silicon dioxide and sapphire.

Vibrational Raman scattering has shown that optical-phonon frequencies ω_p generally decrease with temperature T, while the linewidths Γ_p increase with T. This is due to anharmonicity, as can be shown by perturbation theory. The coupling of the optical phonon to two and to three phonons of lower energy leads to changes in ω_p and Γ_p that vary as $\sim T$ and $\sim T^2$, respectively, above the Debye temperature (Balkanski *et al.*, 1983). Thermal volume expansion also contributes to the decrease of ω_p with temperature, as $\sim T$.

One way to determine temperature *in situ* during processing is from the temperature dependence of ω_p and Γ_p in Raman scattering (Pazionis *et al.*, 1989). Temperature can also be determined from the the ratio of the Stokes (S) and anti-Stokes (AS) scattering intensities. Using Eq. (1), this ratio is $I_S/I_{AS} = A'(n + 1)/n = A' \exp(\hbar\omega_p/kT)$, where A' is the $\omega_s^3\eta_s\chi'^2$ ratio for the S and AS beams. Though $A' \sim 1$, it cannot be set equal to 1. To determine T without significant error, A' must be determined exactly, especially near resonances. Because $(n + 1)/n \rightarrow 1$ at high T, this method is useful only when $kT \ll \hbar\omega_p$. The ratio I_S/I_{AS} has the same temperature dependence in the scattering of other elementary excitations.

The application of a static electric field can change the symmetry of a crystal—for instance, by removing the center of inversion—and normally Raman-forbidden phonons can then be observed (Abstreiter *et al.*, 1984). When an electric field is applied along (001) of cubic perovskite paraelectrics such as $KTaO_3$ and $SrTiO_3$, the symmetry changes from O_h to C_{4v}, and the infrared-active mode F_{1u} modes split into A_1 and E components, while the silent mode F_{2u} splits into B_1 and E components. These modes are now Raman active and have a scattering intensity that varies as the square of the applied field. Electric fields can also change the polarization selection rules for modes that are Raman active with no applied field. These fields can be either externally applied or internally generated, as the internal space-charge field within the surface-depletion region of a doped semiconductor.

Changes in the Raman spectrum from changes in structure, such as those due to stresses or electric fields, are sometimes termed morphic effects.

3.5 Phase Transitions

Vibrational Raman scattering is a good probe of phase transitions (Hayes and Loudon, 1978; Cummins and Levanyuk, 1983). In a phase transition from a lower- to higher-symmetry structure, there is a loss of optical vibration modes, some of which are Raman active. This is seen in the temperature-induced transition from the orthorhombic to tetragonal phase of the perovskite high-T_c superconductor $La_{2-x}M_xCuO_4$, as was mentioned in Sec. 3.1. The frequencies of the lost modes can decrease to 0 as the transition is approached, and these are called "soft" modes. Pressure-induced phase transitions in bulk solids and superlattices have also been monitored by Raman scattering (Weinstein and Zallen, 1984; Venkateswaran *et al.*, 1992). For example, the LO/TO modes in zinc-blende GaAs disappear for $P_h > 173$ kbar because of the phase transition to a distorted NaCl structure, which has no Raman-active modes.

3.6 Time-Resolved Studies

The transient dynamics of phonons can be followed by time-resolved Raman scattering using subpicosecond lasers (Kash and Tsang, 1991). One laser beam can be split into a pumping beam, to create phonons by spontaneous Raman scattering, and a time-delayed probing beam, to analyze these phonons by spontaneous Raman scattering. Alternatively, two lasers at ω_i and $\omega_s = \omega_i - \omega_p$ can initially coherently excite lattice vibrations. The magnitude of the excitation can then be probed as a function of time by using time-resolved CARS (Bron *et al.*, 1986). Phonon lifetimes, such as ~5 ps for ZnSe, agree with the inverse of the linewidths in cw spontaneous Raman studies, and can be ascribed to anharmonic coupling to lower-energy phonons.

4. MAGNETIC RAMAN SCATTERING

In addition to the vibrational peaks, Raman spectra of magnetic solids have Stokes and anti-Stokes peaks that are shifted by collective magnetic excitations (Hayes and Loudon, 1978; Cottam and Lockwood, 1986; Ramdas and Rodriguez, 1988, 1991). Magnetic Raman scattering is observed in the low-temperature ordered states of ferromagnetic, ferrimagnetic, and antiferromagnetic solids, where the magnetic excitations are known as quantized spin waves or magnons. Scattering of electrons in the higher-temperature paramagnetic state can be classified as either magnetic or electronic Raman scattering. When collective effects are important, the magnetic scattering methods and theory in this section are most relevant. Section 5.1 on electronic Raman scattering covers paramagnetic materials where the interaction between magnetic ions is minimal. Magnetic Raman scattering is typically weaker than vibrational scattering, and the frequency shifts are typically smaller.

In the macroscopic description of scattering in ferromagnets, the electric polarization is $\mathbf{P} = \epsilon_0\chi\mathbf{E} + i\epsilon_0 G\mathbf{M} \times \mathbf{E}$ in cubic and isotropic materials, where χ is the electric susceptibility, $\mathbf{M}$ is the magnetization, and $iG = \partial\chi_{ij}(M)/\partial M_h$ (for ijh any cyclic permutation of xyz). The scattering cross section is proportional to the magnetization power spectrum $\langle M_i M_j\rangle_\omega$, and thermal fluctuations in $\mathbf{M}$ lead to the observed frequency distribution of the scattered light. Because of the $\mathbf{M} \times \mathbf{E}$ term, which arises because the magnetization and magnetic field are axial in nature (among other factors), scattering is purely antisymmetric in the incident and scattered polarizations, and linearly polarized incident light couples to circularly polarized scattered light and vice versa. There is an additional scattering contribution due to the precession of the fluctuating $\mathbf{M}$ about an applied magnetic field, which affects the magnetic susceptibility. However, its relative importance decreases with increasing ω_i, and it can usually be neglected in the scattering of visible and ultraviolet light.

The magnetization power spectrum is calculated by using the fluctuation-dissipation theorem and the Bloch equations of motion for $\mathbf{M}$ in a magnetic field $\mathbf{B}$. With $\mathbf{B}$ along the z direction, the longitudinal cross section, due to M_z, features a peak with zero frequency shift, while the transverse cross section, due to M_x and M_y, has Stokes and anti-Stokes shifts $\Delta\omega = \gamma B$, where $\gamma = g\mu_B/\hbar$ is the gyromagnetic ratio, with μ_B the Bohr magneton and g the gyromagnetic factor ($= -2$ for the electron spin). For $T \ll T_c$ (the Curie temperature), transverse scattering dominates. In these ferromagnets and paramagnets, $\Delta\omega$ is very small, $\sim 1\ \mathrm{cm}^{-1}$ for $B = 10^4$ G (0.1 T), and the spectroscopic methods used in Brillouin scattering are needed. For Stokes light, $d^2\sigma/d\Omega\, d\omega_s \propto |e_i^z e_s^+ - e_i^+ e_s^z|^2$, while for anti-Stokes e^+ is replaced by e^-, where $e^{\pm} = e_x \pm ie_y$.

In the microscopic approach, magnetic scattering in ferromagnets and paramagnets occurs by the creation or annihilation of magnons. Below the Curie temperature T_c, the magnetic moments of the transition-metal ions strongly couple to each other, through the Heisenberg exchange interaction, and to the external field B. Diagonalization of the Hamiltonian leads to elementary excitations from the ground state in units of the magnon frequency, which depends, to first order, linearly on the "exchange" field B_E and the applied field B_0. The scattering of light occurs by a two-step, electric dipole process that first couples the ground state, with the z component of spin S_z, to a virtual intermediate state, where spin and orbital motions are coupled, and then back down to the ground-state $S_z - 1$ component. Therefore the ground-state spin is decreased by 1 because of the spin–orbit coupling in the excited state. The scattering by magnons is then determined by the properly phased summations of the scattering of these individual ions. The scattering cross section has a similar form to that from macroscopic theory, but includes resonance effects when the laser frequency and the energy of the intermediate level are approximately equal.

For $T < T_c$ ($B = 0$), the ferromagnetic magnon linewidth increases with increasing T because of magnon–magnon coupling, and the peak amplitude first increases because of increasing $n(\omega)$ and then sharply decreases to 0 near T_c because of the decreasing magnetization.

The macroscopic and microscopic theories of scattering in antiferromagnets and ferrimagnets parallel those for ferromagnets. In antiferromagnets, spins line up in opposite directions on two sublattices and totally cancel the macroscopic $\mathbf{M}$. The Hamiltonian includes the exchange term (which causes ordering in each sublattice below the Néel temperature T_N), a term including the anisotropy field B_A (which acts differently on the two superlattices), and a term that couples with the external magnetic field. For each wave vector $\mathbf{q}$, there are two magnons $\uparrow\mathbf{q}$ and $\downarrow\mathbf{q}$, which correspond to a respective increase and decrease by one in the z component of spin of the crystal. The magnon frequencies are $\omega_{\downarrow\mathbf{q},\uparrow\mathbf{q}} = \omega_{\mathbf{q}} \pm \gamma B_0$, where $\omega_{\mathbf{q}}$ depends on B_A and B_E.

The differential cross section is $d\sigma/d\Omega \propto G^2 M_0\, |(\mathbf{e}_i \times \mathbf{e}_s)|^2\, (n + 1)$, where G contains the dynamical information from the macroscopic or microscopic theory, M_0 is the magnetization of a fully aligned sublattice, and n is the density of magnons. The polarization factor is seen to couple light linearly polarized in the z direction with light circularly polarized in the xy plane, with opposite senses for the two kinds of magnons. Though magnetic scattering cross sections in antiferromagnets are often smaller than for vibrational Raman scattering, the frequency shifts are on the same order, and so conventional Raman spectral dispersion methods can be used. The first-order magnetic scattering peak in $\mathrm{FeF_2}$ at zero magnetic field is at $\omega_{AF,\mathbf{q}=0} \sim 50\ \mathrm{cm}^{-1}$, as seen in Fig. 7, and symmetrically splits into two peaks when a magnetic field is applied. For $T < T_N$ ($B = 0$), as T increases ω_{AF} decreases, the magnon linewidth increases, and the peak amplitude decreases and approaches 0 near T_N.

Second-order or two-magnon scattering is observed in antiferromagnets, as is seen in $\mathrm{FeF_2}$ near $150\ \mathrm{cm}^{-1}$ in Fig. 7. As in two-phonon scattering, two-magnon scattering roughly reflects the one-magnon density of states scaled in frequency by a factor of 2. Because $\omega_{\mathbf{q}}$ increases with q, the strongest contribution in second-order magnetic scat-

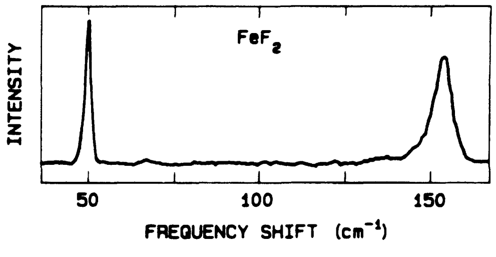

FIG. 7. Magnetic Raman scattering in antiferromagnetic $\mathrm{FeF_2}$ showing the one-magnon and two-magnon lines (Fleury and London, 1968).

tering is larger than $2\omega(q = 0)$, which is in contrast to that for optical phonons. The cross sections for one- and two-magnon scattering are of the same order, whereas two-phonon scattering is much weaker than one-phonon scattering, because second-order scattering in antiferromagnets is not caused by the first-order term taken to second order, as it is for vibrational Raman scattering, but it is due to a different mechanism. In this process, the exchange scattering mechanism, there is a double spin change of neighboring spins on the two sublattices. Because of this close spatial coupling of the magnons, the sum of the energies of the two magnons created in second-order scattering is somewhat less than twice the energy of two individual magnons. Two-magnon scattering is essentially unaffected by a magnetic field. For $B = 0$, with increasing T ($<T_N$) the peak ω decreases and Γ increases, and, unlike the one-magnon peak, the amplitude persists for $T \gg T_N$.

Coupling of spins and phonons is seen in magnetic materials (Güntherodt and Zeyher, 1984). First-order vibrational scattering in europium chalcogenides (EuX, X = O, S, Se, Te) is symmetry forbidden because they have the fcc (NaCl) structure. However, in the paramagnetic phase, broad first-order vibrational peaks are seen that are induced by spin disorder. With magnetic ordering, the spin–phonon coupling changes. In EuO and EuS, which are ferromagnets, vibrational scattering is quenched upon ordering and the shift increases from the phonon frequency ω_p to almost $\omega_p + \omega_{\text{magnon}}$. In EuTe, an antiferromagnet, scattering is enhanced upon ordering.

In some magnetic solids, in addition to the (usual) acoustic magnons, there are other excitations called optical magnons, which are caused by the anisotropy induced by the crystalline electric field. Inelastic scattering has also been observed in canted and metamagnetic antiferromagnets and in ferrimagnets, and for magnetic films and superlattices (Cottam and Lockwood, 1986).

5. ELECTRONIC RAMAN SCATTERING

5.1 Bound Electrons

Bound electrons can scatter by a two-step process, mediated by the $\mathbf{A} \cdot \mathbf{p}$ term in second order. Raman scattering between crystal-field–split electronic states with the same $4f^n$ electronic configuration has been studied in solids with rare-earth ions, such as $PrCl_3$ and $CeCl_3$. In semiconductors, electrons bound to donor atoms and holes bound to acceptor atoms can scatter efficiently to another donor/acceptor state, sometimes with the emission of an acoustic phonon, or to the conduction/valence band (Klein, 1975; Abstreiter *et al.*, 1984).

When a magnetic field is applied to a paramagnetic solid, there can be scattering between the Zeeman-split, bound electronic states of the magnetic ions with a shift $\omega_{PR} = \gamma B$ (electron paramagnetic resonance—EPR). Though it is now assumed that coupling between the magnetic ions is minimal in these materials, and individual ions, and not spin waves, are being probed, the microscopic mechanism and polarization selection rules for scattering are the same as those described for magnetic scattering in Sec. 4. One important example is scattering in diluted magnetic semiconductors (DMS), which are mixed crystals, such as $Cd_{1-x}Mn_xSe$, where there are antiferromagnetic interactions between the magnetic ions (Mn^{2+}) (Ramdas and Rodriguez, 1988, 1991). For small x, they are paramagnets at all temperatures, while for larger x, they can undergo a transition to a spin glass or antiferromagnet at low temperatures. Spin coupling with the lattice can lead to sidebands on the phonon peaks at $\omega_{LO} \pm \omega_{PR}$ and $\omega_{LO} \pm 2\omega_{PR}$. Also, spin-flip scattering between Zeeman-split levels of weakly bound electrons is seen in diluted magnetic semiconductors. The g factor for these electrons is very large, as a result of the exchange interaction between the Mn^{2+} ions and the band electrons, leading to large Zeeman splitting of the free exciton.

5.2 Free Carriers

Light scattering of free carriers in solids, as seen particularly in semiconductors, can involve scattering from single particles or collective oscillations (Klein, 1975; Abstreiter *et al.*, 1984, 1986). Single electrons (or holes) can scatter photons, and change the photon momentum by $\mathbf{q}$ and the photon frequency within the range $\hbar/m^*(q^2/2 - k_F q) < \omega_x < \hbar/m^*(q^2/2 + k_F q)$, where m^* is the effective mass of the particle and k_F is the particle wave vector at the Fermi surface. This leads to a distribution of scattered light that is centered at or near ω_i, and one from which the particle kinetic energy distribution can be determined. Light can also scatter from collective longitudinal charge-density oscillations, which are quantized as plasmons with frequency $\omega_P(\mathbf{q})$ given by $\omega_P^2(\mathbf{q}) = \omega_P^2 + (3v_F^2/5)q^2$, where the plasma frequency is $\omega_P = (4\pi ne^2/\epsilon m^*)^{1/2}$ with n the free particle density, ϵ the dielectric constant, and $v_F = \hbar k_F/m^*$. In addition to these more common optical plasmons, other collective modes can be observed in scattering in multicomponent plasmas with multivalley electrons or with electrons and holes. In these modes, there is partial or total cancellation of charge-density fluctuations, such as acoustic plasmons and intervalley density fluctuations. Optical plasmon scattering tends to dominate for $q < q_S$, where q_S is a characteristic plasma screening wave vector, while single-particle scattering dominates for $q > q_S$. The scattered wave vector is determined by $q = 2k_i \sin(\theta/2)$, where θ is the scattering angle.

Macroscopically, scattering by free electrons is attributed to fluctuations in the electron density. This can be subdivided into charge-density fluctuations, which can lead to the aforementioned single-particle and collective effects, and spin-density fluctuations, which lead only to single-particle effects in bulk nonmagnetic semiconductors. Quantum mechanically, scattering from electrons is a two-step process. In simple isotropic semiconductors with one type of carrier, the photon scattering cross section due to charge-density fluctuations is approximately (Blum, 1970)

$$\frac{d^2\sigma}{d\Omega d\omega_s} = \frac{\hbar q^2 V}{\pi e^2}\frac{m^2}{m^{*2}}\frac{\epsilon_0\epsilon_\infty^2\eta_s}{\eta_i}\frac{\omega_s}{\omega_i}\left(\frac{\omega_0^2}{\omega_0^2 - \omega_i^2}\right) \times (\mathbf{e}_i\cdot\mathbf{e}_s)^2 r_0^2[n(\omega) + 1]\mathrm{Im}\left(-\frac{1}{\epsilon(\omega,\mathbf{q})}\right) \tag{5}$$

for Stokes scattering, where $n(\omega) = 1/[\exp(\hbar\omega/kT) - 1]$, m^* is the effective mass, $r_0 = e^2/(4\pi\epsilon_0)mc^2$, and ϵ_∞ is the dielectric constant at high frequency. Equation (5) includes the contributions of the A^2 term in the Hamiltonian in first order, which describes contributions for electrons in vacuum, and the $\mathbf{A}\cdot\mathbf{p}$ term in second order, which includes the very important effects of interband scattering. This latter term adds the resonance denominator near band gap energies and the dependence on the effective mass rather than the free-particle mass. For an anisotropic band, $(\mathbf{e}_i\cdot\mathbf{e}_s)^2/m^{*2}$ is replaced by $[\mathbf{e}_i\cdot(1/\mathbf{m}_{ij}^*)\cdot\mathbf{e}_s]^2$, where m_{ij}^* is the effective mass tensor. The factor $\mathrm{Im}(-1/\epsilon(\omega,\mathbf{q}))$ describes plasmon scattering through a resonance near ω_P, and can also include the effect of a magnetic field and the coupling of plasmons with LO phonons. Since $\epsilon(\omega,\mathbf{q}) = \epsilon_\infty + \chi(\omega,\mathbf{q})$, $\mathrm{Im}(-1/\epsilon(\omega,\mathbf{q})) = (1/\epsilon_\infty) \times \mathrm{Im}(\chi(\omega,\mathbf{q})/\epsilon(\omega,\mathbf{q}))$, which is approximately $(1/\epsilon_\infty^2)\,\mathrm{Im}(\chi(\omega,\mathbf{q}))$ in dilute plasmas where $\epsilon_\infty \gg \chi(\omega,\mathbf{q})$. Therefore, the Raman spectra of dilute plasmas are determined by the factor $\mathrm{Im}(\chi)$, which is proportional to the density of states of single-particle excitations.

The scattering processes described by Eq. (5) are due to charge-density fluctuations. It is symmetric in the polarization and leads to polarized scattering in the simple case of an isotropic mass tensor. Scattering from spin-density fluctuations is mediated by the spin–orbit interaction, as are the spin-flipping terms in magnetic scattering, and has an $|\mathbf{e}_i \times \mathbf{e}_s|^2$ dependence on polarization. This is antisymmetric in the excitation and scattered polarizations, and therefore scattering from spin-density fluctuations is depolarized and $\mathbf{e}_s \perp \mathbf{e}_i$. The charge- and spin-density fluctuation terms do not interfere because their scattering amplitudes differ by i. In n-type GaAs, collective excitations are due to charge-density oscillations, while single-particle excitations are due to spin-density fluctuations.

With either scattering mechanism, the scattering cross section has an ω_s/ω_i dependence, which is different from the $\omega_i\omega_s^3 \approx \omega^4$ term in the macroscopic description of Raman scattering, as exemplified by vibrational Raman scattering below all resonances. This is analogous to behavior of the classical cross section for elastic scattering of bound electrons, which has a ω^4 dependence below resonance ($\omega \ll \omega_0$, Rayleigh scattering) and no frequency dependence for free electrons ($\omega \gg \omega_0$, Thomson scattering).

In polar semiconductors, LO phonons and plasmons interact through the longitudinal electric field that they both produce. They couple to form two modes $L_\pm$ observable in

Raman scattering at

$$\omega_\pm^2 = \tfrac{1}{2}(\omega_{\mathrm{LO}}^2 + \omega_{\mathrm{p}}^2) \pm \tfrac{1}{2}[(\omega_{\mathrm{LO}}^2 + \omega_p^2)^2 - 4\omega_p^2\omega_{\mathrm{TO}}^2)]^{1/2}. \qquad (6)$$

The TO mode is unaffected. For small electron density n, $\omega_\pm \sim \omega_{\mathrm{LO}}$ and $\omega_p \sim 0$, respectively, while for very large n, where electrons screen out the electric field that splits the LO and TO phonons, the modes are at ω_p and ω_{TO} (Fig. 8). The different scattering mechanisms (deformation potential, electro-optic, and plasmon) can be distinguished by polarization measurements. This strong LO-phonon–plasmon coupling has been seen in n-type semiconductors, but not in p-type materials.

Electronic scattering via the **A·p** terms can leave the solid in the initial electronic band (two-band term) or a different band (three-band). In the processes described so far, the free electron or hole returns to the initial state. Two-band, single-particle scattering can also affect the phonon spectrum, though it is usually less dramatic than the LO-phonon–plasmon coupling (Abstreiter *et al.*, 1984). After three-band scattering, the free carrier or bound electron, as described in the previous section, ends up in a different electronic state, as in p-type Si. Raman scattering from the light- to heavy-hole band in p-type Si, with the conduction band as the intermediate state, can lead to a spectrum that is several hundred cm^{-1} wide. Near 520 cm^{-1}, this electronic scattering interferes with the phonon peak to produce asymmetric or antiresonant profiles (Cerdeira *et al.*, 1973). This arises from the Fano-type interference between scattering with a continuum of levels (electronic) and a discrete level (vibrational). Even when electron–phonon coupling is relatively weak in doped semiconductors, phonon peaks are still affected by high carrier densities. Sometimes the peaks are softened, i.e., the peak frequency decreases.

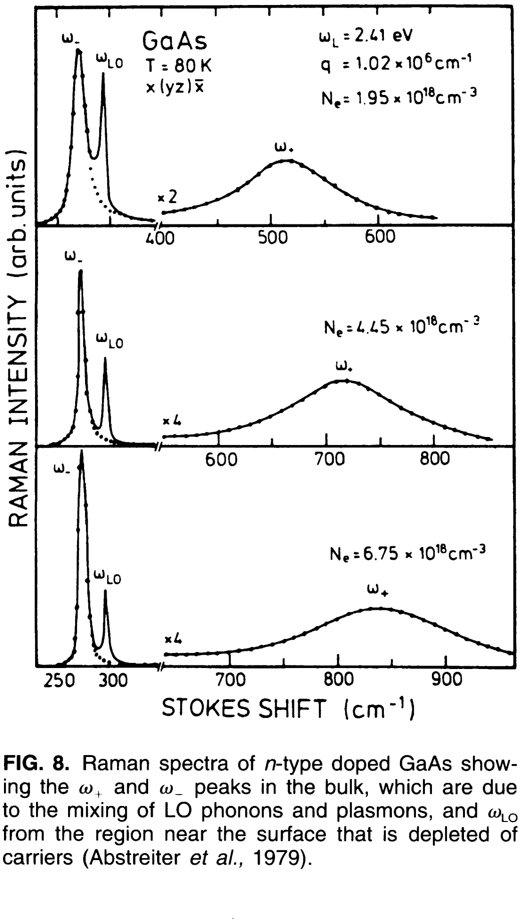

FIG. 8. Raman spectra of n-type doped GaAs showing the ω_+ and ω_- peaks in the bulk, which are due to the mixing of LO phonons and plasmons, and ω_{LO} from the region near the surface that is depleted of carriers (Abstreiter *et al.*, 1979).

Scattering from electronic excitations in high-T_c superconductors leads to several phenomena (Thomsen, 1991). In some cases, it is thought to produce the broad, constant scattering background that is seen in some polarization configurations, which rolls off at low frequencies (Reznik *et al.*, 1992). Some phonon peaks have asymmetric profiles due to a Fano-type interference between phonon and electronic scattering, as seen in semiconductors. Both the softening and hardening, and the line broadening and narrowing of some phonon peaks that occurs when $T < T_c$ (the transition temperature), have been interpreted as evidence for a sharp superconducting gap (Thomsen, 1991; McCarty *et al.*, 1991). In the conventional superconductors Nb$_3$Sn and V$_3$Si, there is a peak at the superconducting gap due to coupling to superconducting pairs (Dierker *et al.*, 1983).

Raman scattering of free electrons in a magnetic field is quite different. The motion of free electrons in a magnetic field B (z direction) is quantized in the x-y plane, leading to Landau levels $\hbar\omega_c(l + \frac{1}{2})$ for $\omega_c \gg \omega_p$, where the cyclotron frequency $\omega_c = eB/m^*c$. More generally, the Landau levels are separated by the frequency of the collective magnetoplasmon mode $\omega_P(q,B) = [\omega_P^2(q) + \omega_c^2]^{1/2}$. Because of the coupling of the electron spin to B, each of the Landau levels is split by an energy $-g\mu_B B s_z$ ($s_z = \pm\frac{1}{2}$). In spin-flip Raman scattering, $\Delta s_z = 1$, $\Delta l = 0$, and the scattered light is Stokes shifted by $\omega_x = g\mu_B B/\hbar$. Spin-flip scattering occurs via spin–orbit coupling in intermediate levels, which is similar to the scattering of magnons. The scattered phonon frequency ω_s can be tuned by changing B, and the stimulated version of this process can be used to make a tunable spin-flip Raman laser. Because of the large values of g in some semiconductors, $g \sim -50$ in InSb, large shifts are possible, ~160 cm^{-1} for $B = 10$ T. With $\mathbf{q} \perp \mathbf{B}$, transitions involving inelastic scattering between the Landau levels, $\Delta l = 1, 2 \ldots$, can occur with frequency shifts at ω_c, $2\omega_c$, ...

5.3 Two-Dimensional Structures

As in vibrational Raman scattering, novel effects appear in electronic Raman scattering in two-dimensional structures, such as quantum wells (Abstreiter *et al.*, 1984; Pinczuk and Abstreiter, 1989). In two-dimensional electron gases, as in modulation-doped GaAs–Al$_x$Ga$_{1-x}$As heterostructures, there is quantized electron motion normal to the layer planes that produces subbands. In Raman scattering, there can be vertical intersubband transitions that involve electron motion only normal to the planes ($q_\parallel = 0$), and intersubband and intrasubband transitions that involve changes in electron wave vector parallel to the planes ($q_\parallel \neq 0$). Intrasubband transitions can involve single-particle excitations and plasmons. The plasma frequency ω_P depends on the details of the two-dimensional confinement.

These different transitions can be distinguished by the scattering geometry, and can be resonantly enhanced at photon energies near gaps for the electron bands that are involved. In a general backscattering geometry, with the laser incident at an angle ϕ_1 to the normal and the scattered light collected at an angle ϕ_2, $q_\perp \gg q_\parallel$. By changing ϕ_1 and ϕ_2, $q_\parallel$ can be varied. For $\phi_1 = -\phi_2 = 45°$, $q_\parallel = 0$ and only intersubband scattering is observed.

Figure 9 shows scattering in a modulation-doped GaAs quantum well. The depolarized $z(y',x')\bar{z}$ spectrum is due to single-particle spin-density intersubband excitations, while the polarized $z(x',x')\bar{z}$ spectrum is due to collective charge-density intersubband excitations. The polarized peaks are higher in energy because of the coupling of the macroscopic electric field with collective excitations (17.7 vs 11.3 meV, etc.), and the 34.7- and 42.4-meV polarized peaks come from the coupling of collective excitations and LO phonons, which is analogous to the $L_\pm$ modes in bulk semiconductors (Pinczuk and Worlock, 1982).

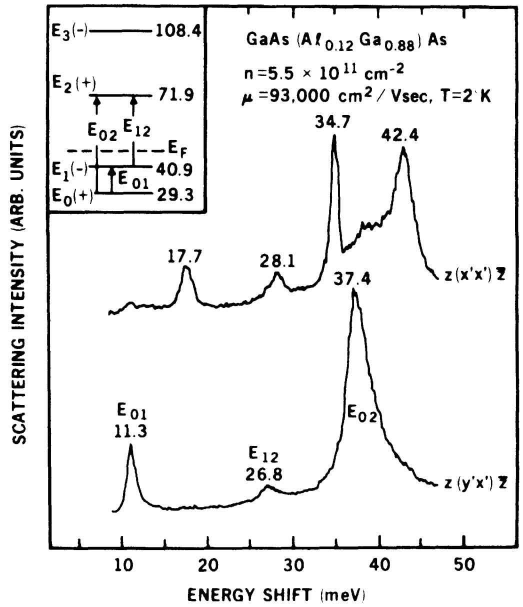

FIG. 9. Polarized (upper) and depolarized (lower) Raman spectra of modulation-doped GaAs-Ga$_{0.88}$Al$_{0.12}$As multiple quantum well, with 250-Å well thickness, due to vertical intersubband transitions (Pinczuk and Worlock, 1982).

With a magnetic field normal to the plane, motion in the plane is quantized. All transitions are discrete and are due to intersubband transitions, Landau-level transitions, and a combination of intersubband and Landau-level transitions. In-plane magnetic excitations between Landau levels that are associated with single-particle effects are called magnetic excitons, while those associated with collective modes are magnetoplasmons.

Analogous one-dimensional phenomena have been seen in quantum wires, where there is two-dimensional confinement.

5.4 Time-Resolved Studies

Pump/probe experiments with subpicosecond lasers can be used to study the dynamics of hot electrons, notably in semiconductors (Kash and Tsang, 1991). Equilibration of valence-band electrons occurs by coupling to optical phonons (~190 fs in GaAs), intervalley scattering, and carrier–carrier collisions. Hot conduction-band electrons, which are produced by a pump pulse, are relaxed by the formation of LO phonons. These phonons can be measured by monitoring the anti-Stokes Raman signal with a time-delayed pulse (von de Linde *et al.*, 1980).

6. RAMAN SCATTERING IN GASES AND LIQUIDS

Spontaneous Raman scattering in gases and liquids is a coherent two-step process, as from level a to level c in Fig. 1. In vibrational Raman scattering in gases and liquids, these two levels are different vibrational levels in the same electronic state. In gases, the rotational quantum numbers of a and c can also differ, leading to rotational–vibrational Raman scattering, or these two states can be different rotational states in the same vibrational level, which is called pure rotational Raman scattering. In vibrational and rotational scattering, the intermediate state b is in an excited electronic state. The states a and c can be different electronic states, as has been examined in atoms. Such electronic Raman scattering is relatively seldom studied.

Quantum mechanically, the term in the electric polarizability describing scattering from rotational–vibrational (rovibronic) states a to c in a molecule is

$$\alpha_{ij} = \frac{e^2}{\hbar}\sum_b\left(\frac{\langle c|r_j|b\rangle\langle b|r_i|a\rangle}{(\omega_{ab} - \omega_i - i\Gamma_b)} + \frac{\langle b|r_j|a\rangle\langle c|r_i|b\rangle}{(\omega_{bc} + \omega_i - i\Gamma_b)}\right), \qquad (7)$$

where r_i is the Cartesian component of the electron position operator, Γ_b is a damping term that describes the width of level b, and ω_{ab} is the frequency of the a–b transition (>0) (Rousseau *et al.*, 1979). The Born–Oppenheimer approximation allows factoring of each wave function, such as for state a, into three terms: a vibrational term $a_v(\mathbf{R})$, a rotational term $a_r(\mathbf{R})$, both of which are functions of the internuclear coordinate $\mathbf{R}$, and an electronic term $a_e(\mathbf{r},\mathbf{R})$, which is a function of the electron coordinate $\mathbf{r}$ and is parametrized by $\mathbf{R}$.

Integrating the rotational wave-function terms in the matrix elements over molecular orientation gives the dipole selection rules for the rotational quantum number. Integrating the electronic terms (a_e and c_e in the ground electronic state, and b_e in an excited electronic state) over the electronic coordinate gives an $\mathbf{R}$-dependent electronic dipole moment M that can be expanded about the equilibrium configuration ($\mathbf{R}_0$), as $\mathbf{R} = \mathbf{R}_0 + \mathbf{u}$, to give $M(R) = M(R_0) + [\partial M/\partial R|_{R_0}]u + \cdots$. Ignoring the nonresonant second term in Eq. (7) then gives

$$\begin{aligned}\alpha_{ij} = {} & |M(R_0)|^2\,\frac{1}{\hbar}\sum_b\frac{\langle c_v|b_v\rangle\langle b_v|a_v\rangle}{(\omega_{ab} - \omega_i - i\Gamma_b)} \\ & + M(R_0)\left.\frac{\partial M}{\partial R}\right|_{R_0}\frac{1}{\hbar} \\ & \times\sum_b\frac{\langle c_v|R|b_v\rangle\langle b_v|a_v\rangle + \langle c_v|b_v\rangle\langle b_v|R|a_v\rangle}{(\omega_{ab} - \omega_i - i\Gamma_b)} \\ & + \cdots, \qquad (8)\end{aligned}$$

where the remaining integration is over vibrational coordinates.

The first term in Eq. (8) involves the vibrational overlap Franck–Condon factors and is also known as the Albrecht A term. Its contribution to scattering is known as "Franck–Condon scattering." Far from resonance ($|\omega_{ab} - \omega_i| \gg 0$), the numerator sums to $\langle c_v|a_v\rangle$, which leads to no change in vibrational state, i.e., Rayleigh and pure rotational Raman scattering. The second term corresponds to the Albrecht B term and leads to "Herzberg–Teller scattering." Far from resonance, the numerator sums to $\langle c_v|R|a_v\rangle$, which means that the vibrational quantum number v changes by ± 1 (fundamental scattering). Overtone scattering ($|\Delta v| > 1$) is weak. Near resonance, the A term dominates vibrational Raman scattering, and both the fundamental and overtones can be strong (Rousseau *et al.*, 1979).

More generally, the separation of vibrational and rotational effects in Raman scattering can be seen by examining the electric polarizability in the space-fixed (laboratory) frame, where it depends on the vibrational and rotational coordinates, after transformation from the molecule-fixed frame, where it depends on only the vibrational coordinates.

Classically, pure rotational Raman scattering is due to the anisotropy in the static (or equilibrium) polarizability tensor, as given by $2\alpha_{zz} - \alpha_{xx} - \alpha_{yy}$. This is analogous to the requirement of a nonvanishing permanent electric dipole moment for allowed infrared transitions between rotational levels. Vibrational Raman scattering can occur if at least one of the derivatives of the polarizability tensor with respect to the normal coordinate, $\partial\alpha_{ij}/\partial u$, does not vanish when evaluated at the equilibrium position; then the dynamic polarizability, which is induced by vibration, is nonzero. Infrared vibrational transitions are allowed when at least one derivative of the dipole moment (transition dipole moment) does not vanish at the equilibrium position.

As with phonons in solids, symmetry arguments show that vibrational Raman scattering of a given mode is allowed if there is an α_{ij} that belongs to the same symmetry species, i.e., transforms the same way under symmetry operations, as does that mode (Long, 1977). Vibrations cannot be both Raman and IR active in molecules with a center of symmetry. The vibrational mode in homonuclear diatomic molecules, such as N_2, is Raman active and IR inactive. In linear symmetric triatomic molecules, like CO_2 (O–C–O), the symmetric stretch is Raman active and IR inactive, while the asymmetric stretch and the two degenerate bending modes are Raman inactive and IR active.

In molecules without a center of symmetry, modes can be both Raman and IR active. The vibrational mode in heteronuclear diatomic molecules, like HF, is both Raman and IR active, as are the three modes (asymmetric stretch, symmetric stretch, and bend) in nonlinear triatomic molecules, like H_2O. When the molecular symmetry is low, most or all modes may be both Raman and IR active, though the relative strengths of the Raman and IR modes may be very different for different modes.

In vibrational Raman scattering from a given mode of frequency ω_v with quantum number v, the Stokes spectrum corresponds to $\Delta v = 1$, as is seen for several different modes in liquid CCl_4 in the inset in Fig. 10(b), while the anti-Stokes spectrum corresponds to $\Delta v = -1$. Stokes transitions from $v = 0 \rightarrow 1$ dominate, but at higher temperatures Stokes scattering from higher v can occur. This scattering is weaker by $(v + 1) \times \exp(-v\hbar\omega_v/kT)$, and has slightly smaller shifts because of anharmonicity. Vibrational Raman cross sections have been compiled for many gases and liquids (Schrotter and Klockner, 1979). Typically, the vibrational Raman cross section per molecule is ~4× larger in a liquid than in a gas because of the effect of internal electric fields. Spontaneous and stimulated vibrational Raman scattering has been examined in numerous liquids (Fig. 10). Higher-order Stokes SRS peaks are also often seen. Thresholds for SRS are lower in liquid droplets because of morphology-dependent resonances (MDR) (Fig. 10) (Qian and Chang, 1986).

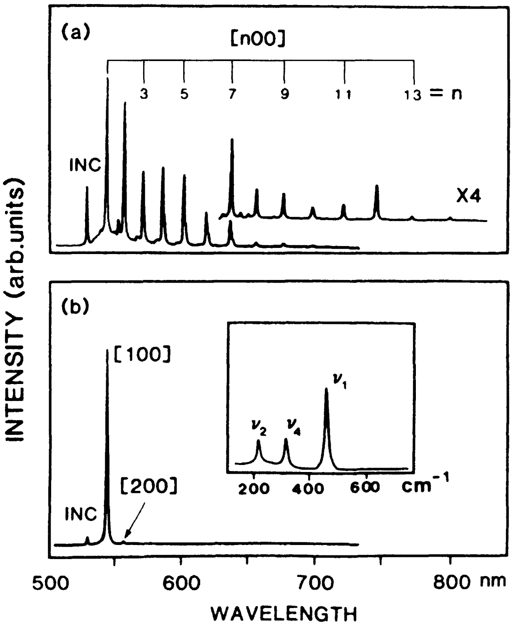

FIG. 10. Vibrational stimulated Raman scattering in (a) a CCl_4 droplet showing several higher-order Stokes lines for the ν_1 mode and in (b) liquid CCl_4 showing a strong first-order peak and a weak second-order peak. The inset shows spontaneous Raman scattering from several vibrational modes in liquid CCl_4 (Qian and Chang, 1986).

In gases, there is further structure due to changes in the rotational quantum numbers, giving rise to rotational–vibrational Raman scattering (Long, 1977). There is similar structure in pure rotational Raman scattering ($\Delta v = 0$), which has Stokes and anti-Stokes branches on either side of the Rayleigh-scattered peak as seen in Fig. 11. In linear molecules, such as N_2 and CO_2, where the rotational quantum number is J, the selection rules for rotational–vibrational scattering from totally symmetric vibrational species are $\Delta J = 0, \pm 2$; for pure rotational scattering, the selection rule is $\Delta J = 0, +2$. For degenerate vibrational species, $\Delta J = \pm 1$ is also allowed in rotational–vibrational scattering, and $\Delta J = +1$ for pure rotational scattering. The $\Delta J = -2, -1, 0, 1,$ and 2 branches are also called *O*, *P*, *Q*, *R*, and *S* branches, respectively.

In rigid-rotor approximation, the rotational energy of a linear molecule is $BJ(J + 1)$, where B is the rotational constant. In pure rotational scattering, the $\Delta J = +2$ selection rule gives rise to *S*-branch anti-Stokes lines at $\omega_i + (4J + 6)B$ [$J \geq 0$], which are upshifted in frequency by $6B$, $10B$, $14B$, ... relative to the laser line at ω_i, and *S*-band Stokes lines at $\omega_i - (4J - 2)B$ [$J \geq 2$], which are downshifted by $6B$, $10B$, $14B$, ... This is seen in Fig. 11 for N_2. These lines are equally spaced by $4B$, except for large J, where centrifugal deviations from the rigid-rotor model alter this spacing. Since ΔJ is often defined as $J' - J''$, where J' refers to the state with higher energy and J'' to that with lower energy, the selection rule for pure rotational scattering is formally $\Delta v = 0$, $\Delta J = +2$, for both Stokes and anti-Stokes scattering. The transition $\Delta v = 0$, $\Delta J = 0$ is actually Rayleigh scattering.

In rotational–vibrational Stokes scattering, the $\Delta J = \pm 2$ selection rule gives rise to a series of *S*-branch lines shifted down in frequency from the laser line ω_i by $\omega_v + 6B' + (5B' - B'')J + (B' - B'')J^2$ for $J \geq 0$ and *O*-branch lines shifted by $\omega_v + 2B' - (3B' + B'')J + (B' - B'')J^2$ for $J \geq 2$, where B' and B'' are the rotational constants for the upper and lower vibrational levels, respectively. (For Stokes scattering, these are the final and initial states, respectively.) These bands look qualitatively like those in pure rotational Raman scattering except that since usually $B' < B''$, the spacing of the *O*-branch lines increases with J, while the spacing of *S*-branch lines decreases with J. For large enough J, the shifts of *S*-branch lines may no longer increase with J but reverse sign to form a bandhead. The $\Delta J = 0$ selection rule gives *Q*-branch lines at $\omega_v + (B' - B'')J + (B' - B'')J^2$, which are closely spaced, sometimes overlapping to form a big peak, and are near but usually less than ω_v.

In symmetric-top molecules, such as H_2O

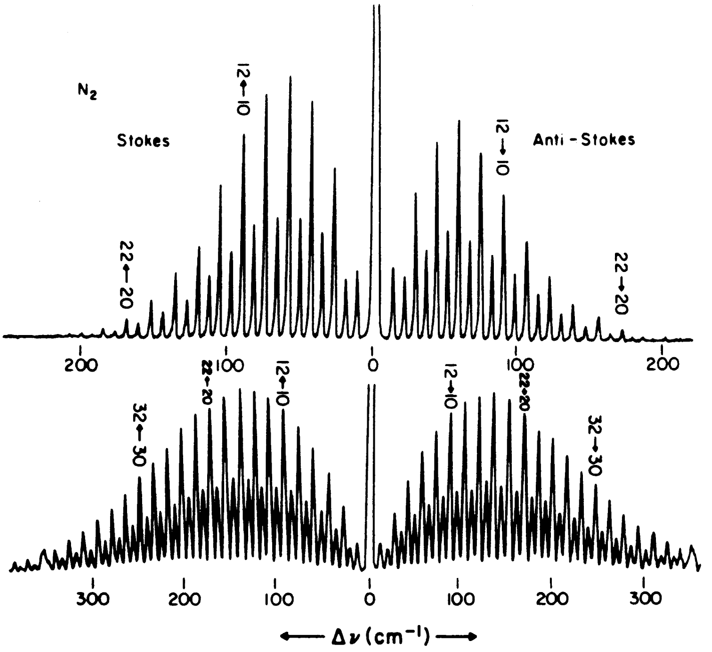

FIG. 11. Pure rotational Raman scattering in N_2 gas at 300 K (top) and at 1582 K in a H_2 diffusion flame (bottom), showing Stokes and anti-Stokes scattering (Drake, 1988).

and NH_3, there is additional rotational structure that results from the quantization of the projection of the rotational angular momentum J on the molecular axis, which is described by the quantum number K (Long, 1977). The selection rules for rotational–vibrational scattering from totally symmetric vibrational species are $\Delta J = 0, \pm 1, \pm 2$ and $\Delta K = 0$; for other vibrational species $\Delta K = \pm 1$ or ± 2 are sometimes allowed. The selection rule for rotational–vibrational scattering from totally symmetric vibrational species in spherical-top molecules, such as SF_6, is $\Delta J = 0$; for degenerate vibrational species $\Delta J = \pm 1, \pm 2$ are also allowed. In each case, the selection rules for pure rotational scattering are the same, except that negative ΔJ possibilities are excluded.

The relative strengths of the rotational lines depend on the product of the thermal population of the rotational state and the rotational matrix element (Long, 1977; Weber, 1979). The alternating line-intensity pattern sometimes seen in rotational Raman spectra is due to nuclear-spin statistics, such as the 2:1 ratio seen for N_2 in Fig. 11. The total scattering cross section of a rotational–vibrational Raman band is obtained by summing the rotational transitions, and is equal to the cross section of a space-fixed, nonrotating molecule integrated over all possible orientations (Schrotter and Klockner, 1979). High-resolution rotational–vibrational Raman spectroscopy (Broderson, 1979) and high-resolution rotational Raman spectroscopy (Weber, 1979) have been studied in many gases.

7. APPLICATIONS IN MATERIALS CHARACTERIZATION, PROCESS DIAGNOSTICS, AND LASER DEVELOPMENT

Vibrational and electronic Raman scattering are widely used in *ex situ* materials characterization, including that for semiconductors (Pollak, 1991), polymers (Bulkin, 1991), and catalysts (Stencel, 1990; Mehicic and Grasselli, 1991). Many properties of semiconductors can be probed, including damage resulting from processing such as ion implantation and reactive ion etching, strain, interface quality, grain size and orientation, and doping levels. In vibrational Raman scattering, substitutional dopant atoms can be detected by their local vibrational modes, when they are lighter than the host, and their effects on the host modes due to disorder. Disorder caused by doping or other types of damage can lead to mode softening and broadening, and the relaxation of symmetry and polarization selection rules. For example, in backscattering from a (001) surface of a zinc-blende semiconductor, TO modes may be seen in addition to the normally allowed LO modes, and disorder-activated modes, such as the DATA (disorder-activated TA) modes, may appear. Free carriers due to dopants can scatter by single-particle and collective effects, and sometimes couple to phonons, as discussed in Sec. 5.2. To see the bulk $L_\pm$ modes in n-type zinc-blende semiconductors, the Raman probe depth must exceed the width of the surface depletion layer, where only LO phonons are seen (Fig. 8). Electronic Raman scattering of intersubband transitions in semiconductor quantum wells can be used to determine band offsets (Menendez and Pinczuk, 1988).

Spontaneous vibrational Raman scattering can be used as a real-time diagnostic of the region near the surface during thin-film processing, as during molecular-beam epitaxy (MBE) (Nowak *et al.*, 1992) and laser-assisted chemical vapor deposition (CVD) (Magnotta and Herman, 1986). The gas above the substrate can be probed by vibrational and rotational Raman scattering; pure rotational Raman scattering can measure gas temperature by determining the thermal population of rotational levels as is seen in Fig. 11. Spontaneous Raman scattering is also used in the remote sensing of the atmosphere, which is called Raman lidar (light detection and ranging) (She, 1990). CARS has been used to monitor the gas phase during thin-film processing and, quite extensively, in combustion diagnostics (Nibler and Yang, 1987; Eckbreth, 1996). Stimulated Raman scattering can be used to determine the chemical concentration and size of fuel droplets in combustion (Acker *et al.*, 1990).

Stimulated Raman scattering in gases and liquids can produce coherent light at the Stokes and anti-Stokes wavelengths, as well as at higher-order Stokes and anti-Stokes wavelengths. Higher-order vibrational anti-Stokes scattering of ArF lasers (193 nm) in H_2 gas has produced vacuum ultraviolet radiation to 138 nm (Eimerl *et al.*, 1981; Dobele and Ruckle, 1984), while rotational (first Stokes) Raman scattering of CO_2 lasers (10.6 μm) in p-H_2 has efficiently produced radiation near 17 μm (Midorikawa *et al.*, 1985). Laser pulse compression can be achieved by stimulated backward Raman scattering in gases or liquids by amplifying a weak laser beam at the Stokes frequency by a counter-propagating pump beam. This can produce a pulsed laser at the Stokes wavelength that is much shorter than the pump beam (Murray *et al.*, 1979). Free-electron lasers (FEL) can be operated in the collective Raman mode (Raman FEL) where the laser wavelength is in the mm range, as well as in the single-particle, Compton regime where the laser wavelength is commonly 0.3–10 μm (Marshall, 1985).

GLOSSARY

Anti-Stokes Scattering: Inelastic scattering where the photon gains energy and deexcites the medium.

CARS: Coherent anti-Stokes Raman scattering, where two pulsed laser beams at ω_1 and ω_2 ($\omega_1 > \omega_2$) intersect in a material to produce coherent light at $\omega = 2\omega_1 - \omega_2$.

Hyper-Raman Scattering: Coherent process where a harmonic of the incident laser beam is produced, which interacts with the medium by a Stokes or anti-Stokes process.

Stimulated Raman Scattering: Coherent, nonlinear process where the Raman-scattering rate is proportional to the number of photons at the scattering frequency.

Stokes Scattering: Inelastic scattering where the photon loses energy and excites the medium.

Works Cited

Abstreiter, G., Trommer, R., Cardona, M., Pinczuk, A. (1979), *Solid State Commun.* **30,** 703-707.

Abstreiter, G., Cardona, M., Pinczuk, A. (1984), in: M. Cardona, G. Güntherodt, (Eds.), *Light Scattering in Solids IV,* Berlin: Springer, Chap. 2.

Abstreiter, G., Merlin, R., Pinczuk, A. (1986), *IEEE J. Quantum Electron.* **QE-22,** 1771-1784.

Acker, W. P., Serpenguzel, A., Chang, R. K., Hill, S. C. (1990), *Appl. Phys. B* **51,** 9–16.

Balkanski, M., Wallis, R. F., Haro, E. (1983), *Phys. Rev. B* **28,** 1928–1934.

Blum, F. A. (1970), *Phys. Rev. B* **1,** 1125–1135.

Broderson, S. (1979), in: A. Weber (Ed.), *Raman Spectroscopy of Gases and Liquids,* Berlin: Springer, Chap. 2.

Brodsky, M. H. (1975), in: M. Cardona (Ed.), *Light Scattering in Solids,* New York: Springer, Chap. 4.

Bron, W. E., Kuhl, J., Rhee, B. K. (1986), *Phys. Rev. B* **34,** 6961–6971.

Bulkin, B. J. (1991), in: J. G., Grasselli, B. J. Bulkin (Eds.), *Analytical Raman Spectroscopy,* New York: Wiley, Chap. 7.

Campbell, I. H., Fauchet, P. M. (1986), *Solid State Commun.* **58,** 739–741.

Campion, A. (1985), *J. Vac. Sci. Technol. B* **3,** 1404–1407; *Annu. Rev. Phys. Chem.* **36,** 549–572.

Cardona, M. (1982), in: M. Cardona, G. Güntherodt (Eds.), *Light Scattering in Solids II,* Berlin: Springer, Chap. 2.

Cerdeira, F., Buchenauer, C. J., Pollak, F. H., Cardona, M. (1972), *Phys. Rev. B* **5,** 580–593.

Cerdeira, F., Fjeldy, T. A., Cardona, M. (1973), *Phys. Rev. B* **8,** 4734–4745.

Chang, R. K., Long, M. B. (1982), in: M. Cardona, G. Güntherodt (Eds.), *Light Scattering in Solids II,* Berlin: Springer, Chap. 3.

Connell, G. A. N., Nemanich, R. J., Tsai, C. C. (1980), *Appl. Phys. Lett.* **36,** 31–33.

Cottam, M. G., Lockwood, D. J. (1986), *Light Scattering in Magnetic Solids,* New York: Wiley.

Cummins, H. Z., Levanyuk, A. P. (Eds.) (1983), *Light Scattering Near Phase Transitions,* Amsterdam: North-Holland.

Dierker, S. B., Klein, M. V., Webb, G. W., Fisk, Z. (1983), *Phys. Rev. Lett.* **50,** 853–856.

Dobele, H. F., Ruckle, B. (1984), *Appl. Opt.* **23,** 1040–1043.

Drake, M. C. (1988), *Mater. Res. Soc. Symp. Proc.* **117,** 203–216.

Druet, S. A. J., Taran, J.-P. E. (1981), *Prog. Quant. Electron.* **7,** 1–72.

Eckbreth, A. C. (1996), *Laser Diagnostics for Combustion Temperature and Species,* 2nd ed., Luxembourg: Gordon & Breach.

Eimerl, D., Hargrove, R. S., Paisner, J. A. (1981), *Phys. Rev. Lett.* **46,** 651–654.

Feile, R. (1989), *Physica C* **159,** 1–32.

Feld, M. S., Javan, A. (1969), *Phys. Rev.* **177,** 540–562.

Ferraro, J. R., Maroni, V. A. (1990), *Appl. Spectrosc.* **44,** 351–366.

Fleury, P. A., London, R. (1968), *Phys. Rev.* **166,** 514–530.

Güntherodt, G., Zeyher, R. (1984), in: M. Cardona, G. Güntherodt (Eds.), *Light Scattering in Solids IV,* Berlin: Springer, Chap. 4.

Harley, R. T., Lyons, K. B., Fleury, P. A. (1980), *J. Phys. C* **13,** L447–453.

Harvey, A. B. (Ed.) (1981), *Chemical Applica-*

tions of Nonlinear Raman Spectroscopy, New York: Academic.

Hayes, W., Loudon, R. (1978), *Scattering of Light by Crystals,* New York: John Wiley.

Hendra, P., Jones, C., Warnes, G. (1991), *Fourier Transform Raman Spectroscopy: Instrumentation and Chemical Applications,* New York: Ellis Horwood.

Hirschfeld, T. (1973), *Appl. Spectrosc.* **27,** 389–390.

Hopkins, J. B., Farrow, L. A. (1986), *J. Appl. Phys.* **59,** 1103–1110.

Jha, S. S. (1985), *Surf. Sci.* **158,** 190–210.

Jusserand, B., Paquet, D., Mollot, F., Alexandre, F., Le Roux, G. (1987), *Phys. Rev. B* **35,** 2808–2817.

Jusserand, B., Cardona, M. (1989), in: M. Cardona, G. Güntherodt (Eds.), *Light Scattering in Solids V,* Berlin: Springer, Chap. 3.

Kash, J. A., Tsang, J. C. (1991), in: M. Cardona, G. Güntherodt (Eds.), *Light Scattering in Solids VI,* Berlin: Springer, Chap. 8.

Klein, M. V. (1975), in: M. Cardona (Ed.), *Light Scattering in Solids,* New York: Springer, Chap. 4.

Klein, M. V. (1986), *IEEE J. Quantum Electron.* **QE-22,** 1760–1770.

Landsberg, G., Mandelstam, L. (1928), *Naturwissenschaften* **16,** 557.

Levenson, M. D., Kano, S. S. (1988), *Introduction to Nonlinear Laser Spectroscopy,* Boston: Academic.

Long, D. A. (1977), *Raman Spectroscopy,* New York: McGraw-Hill.

Magnotta, F., Herman, I. P. (1986), *Appl. Phys. Lett.* **48,** 195–197.

Marshall, T. C., (1985), *Free-Electron Lasers,* New York: Macmillan.

McCarty, K. F., Radousky, H. B., Liu, J. Z., Shelton, R. N. (1991), *Phys. Rev. B* **43,** 13751–13754.

Mehicic, M., Grasselli, J. G. (1991), in: J. G. Grasselli, B. J. Bulkin (Eds.), *Analytical Raman Spectroscopy,* New York: Wiley, Chap. 10.

Menendez, J., Pinczuk, A. (1988), *IEEE J. Quantum Electron.* **QE-24,** 1698–1711.

Menendez, J. (1989), *J. Luminescence* **44,** 285–314.

Midorikawa, K., Tashiro, H., Aoki, Y., Nagasaka, K., Toyoda, K., Namba, S. (1985), *Appl. Phys. Lett.* **47,** 1033–1035.

Murray, J. R., Goldhar, J., Eimerl, D., Szoke, A. (1979), *IEEE J. Quantum Electron.* **QE-15,** 342–368.

Nibler, J. W., Knighten, G. V. (1979), in: A. Weber (Ed.), *Raman Spectroscopy of Gases and Liquids,* Berlin: Springer, Chap. 7.

Nibler, J. W., Yang, J. J. (1987), in *Annu. Rev. Phys. Chem.* **38,** 349–381.

Nowak, C., Zahn, D. R. T., Rossow, U., Richter, W. (1992), *J. Vac. Sci. Technol. B* **10,** 2066–2071.

Otto, A. (1984), in: M. Cardona, G. Güntherodt (Eds.), *Light Scattering in Solids IV,* Berlin: Springer, Chap. 6.

Pazionis, G. D., Tang, H., Herman, I. P. (1989), *IEEE J. Quantum Electron.* **QE-25,** 976–988.

Pinczuk, A., Burstein, E. (1975), in: M. Cardona (Ed.), *Light Scattering in Solids,* New York: Springer, Chap. 2.

Pinczuk, A., Worlock, J. M. (1982), *Surf. Sci.* **113,** 69–84.

Pinczuk, A., Abstreiter, G. (1989), in: M. Cardona, G. Güntherodt (Eds.), *Light Scattering in Solids V,* Berlin: Springer, Chap. 4.

Pollak, F. H. (1991), in: J. G. Grasselli, B. J. Bulkin (Eds.), *Analytical Raman Spectroscopy,* New York: Wiley, Chap. 6.

Porto, S. P. S., Fleury, P. A., Damen, T. C. (1967), *Phys. Rev.* **154,** 522–526.

Qian, S.-X., Chang, R. K. (1986), *Phys. Rev. Lett.* **56,** 926–929.

Raman, C. V. (1928), *Indian J. Phys.* **2,** 387; Raman, C. V., Krishnan, K. S. (1928), *Nature* **121,** 501.

Ramdas, A. K., Rodriguez, S. (1988), in: J. K. Furdyna, J. Kossut (Eds.), *Semiconductors and Semimetals,* Vol. 25, *Diluted Magnetic Semiconductors,* Boston: Academic, Chap. 9.

Ramdas, A. K., Rodriguez, S. (1991), in: M. Cardona, G. Güntherodt (Eds.), *Light Scattering in Solids VI,* Berlin: Springer, Chap. 4.

Reznik, D., Klein, M. V., Lee, W. C., Ginsberg, D. M., Cheong, S.-W. (1992), *Phys. Rev. B* **46,** 11725–11729.

Rousseau, D. L., Friedman, J. M., Williams, P. F. (1979), in: A. Weber (Ed.), *Raman Spectroscopy of Gases and Liquids,* Berlin: Springer, Chap. 6.

Schrotter, H. W., Klockner, H. W. (1979), in: A. Weber (Ed.), *Raman Spectroscopy of Gases and Liquids,* Berlin: Springer, Chap. 4.

She, C. Y. (1990), *Contemp. Physics* **31,** 247–260.

Smekal, A. (1923), *Naturwissenschaften* **11,** 873.

Srivastava, G. P. (1990), *Physics of Phonons,* Bristol: Adam Hilger.

Stencel, J. M. (1990), *Raman Spectroscopy for Catalysis,* New York: Van Nostrand.

Thomsen, C. (1991), in: M. Cardona, G. Güntherodt (Eds.), *Light Scattering in Solids VI,* Berlin: Springer, Chap. 6.

Tsang, J. C. (1989), in: M. Cardona, G. Güntherodt (Eds.), *Light Scattering in Solids V,* Berlin: Springer, Chap. 6.

Ushioda, S., Loudon, R. (1982), in: V. M. Agranovich, D. L. Mills (Eds.), *Surface Polaritons,* Amsterdam: North-Holland, Chap. 12.

Venkateswaran, U. D., Cui, L. J., Weinstein, B.

A., Chambers, F. A. (1992), *Phys. Rev. B* **45,** 9237–9247, 9248–9265.

Vogt, H. (1982), in: M. Cardona, G. Güntherodt (Eds.), *Light Scattering in Solids II,* Berlin: Springer, Chap. 4.

von der Linde, D., Kuhl, J., Klingenberg, H. (1980), *Phys. Rev. Lett.* **44,** 1505–1508.

Wang, Z. P., Jiang, D. S., Ploog, K. (1988), *Solid State Commun.* **65,** 661–663.

Weber, A. (1979), in: A. Weber (Ed.), *Raman Spectroscopy of Gases and Liquids,* Berlin: Springer, Chap. 3.

Weinstein, B. A., Zallen, R. (1984), in: M. Cardona, G. Güntherodt (Eds.), *Light Scattering in Solids IV,* Berlin: Springer, Chap. 8.

Weinstein, B. A., Cardona, M. (1973), *Phys. Rev. B* **7,** 2545–2551.

Yariv, A. (1989), *Quantum Electronics,* New York: Wiley.

Further Reading

Adar, F., Griffiths, J. E., Lerner, J. M. (Eds.) (1989), *Raman Scattering, Luminescence, and Spectroscopic Instrumentation in Technology,* SPIE Proceedings Vol. 1055, Bellingham, WA: SPIE.

Cardona, M. (Ed.) (1975), *Light Scattering in Solids,* Berlin: Springer.

Cardona, M., Güntherodt, G. (Eds.) (1982–1991), *Light Scattering in Solids II–VI,* Berlin: Springer.

Ferraro, J. R., Nakamoto, K. (1994), *Introductory Raman Spectroscopy,* Boston: Academic.

Gerrard, D. L., Birnie, J. (1992), *Anal. Chem.* **64,** 502R–513R.

Griffiths, J. E., Adar, F. (Eds.) (1987), *Raman and Luminescence Spectroscopy in Technology,* SPIE Proceedings Vol. 822, Bellingham, WA: SPIE.

Lockwood, D. J., Young, J. F. (Eds.) (1990), *Light Scattering in Semiconductor Structures and Superlattices,* New York: Plenum.

Nakamoto, K. (1986), *Infrared and Raman Spectra of Inorganic and Coordination Compounds,* 4th ed., New York: Wiley.

Pemberton, J. E., Guy, A. L. (1986), in: *American Society for Metals Reference Books,* Vol. 10, *Materials Characterization,* pp. 126–138.

Valentini, J. J. (1987), in: L. J. Radziemski, R. W. Solarz, J. A. Paisner (Eds.), *Laser Spectroscopy and its Applications,* New York: Marcel Dekker, Chap. 7.

Weber, A. (Ed.) (1975), *Raman Spectroscopy of Gases and Liquids,* Berlin: Springer.

Wright, G. B. (Ed.) (1969), *Light Scattering Spectra of Solids,* New York: Springer.

RANDOM PROCESSES

MICHAEL F. SHLESINGER, *Office of Naval Research, Arlington, Virginia, U.S.A.*

	Introduction	45
1.	**A Brief History of Random Processes**	46
2.	**Gaussian Random Walks**	48
3.	**Langevin Equations**	49
3.1	Kubo Oscillator	50
4.	**Probability Equations**	51
4.1	The Chain Equation and the Master Equation	51
4.2	Fokker–Planck Equations	52
5.	**Random Walks**	53
5.1	Discrete-Time	53
5.2	Continuous-Time	55
5.3	Many Walkers	58
6.	**Reaction Kinetics**	59
6.1	Trapping	59
6.2	First- and Second-Order Kinetics	59
6.3	Reaction–Diffusion Equations	60
7.	**Fractal-Time Processes**	61
7.1	Fractal Waiting Times	61
7.2	Stretched-Exponential Relaxation	62
8.	**Fractal-Space Processes**	63
8.1	Lévy Flights	63
8.2	Lévy Walks	64
8.3	Random Walks on Fractals	65
9.	**Disordered Systems**	66
	Glossary	68
	Works Cited	69
	Further Reading	69

INTRODUCTION

Our main theme is the random motion of a particle. A random process that involves a time variable is also called a stochastic process. A physical approach to this problem, pioneered by Langevin, is to write a Newtonian equation of motion $F = Ma$ for a particle. One can include additive random forces, as well as noisy parameters. A frictional force and a force derived from a potential can be included, as well. This approach allows one to investigate random motion in a potential, where the force is written as minus the gradient of the potential. The force can be linear or nonlinear. Random motion in the linear force field of a harmonic potential is called an Ornstein–Uhlenbeck process. In the Langevin approach the particle is always in motion. In the random-walk approach, discussed below, a particle can become trapped and not move for a random length of time.

The variables in a Langevin equation, such as position and velocity, are random variables because they are determined, in part, by the noise terms in the equation. We describe how to find the time dependence of moments of these random variables. When calculating moments, the simplest case arises when the noise is uncorrelated with itself in time or space. The effect of noise correlation can, however, be of overriding importance. We will see in the Kubo oscillator example how time-correlated noise dramatically changes the high-frequency behavior of the moments of a noisy oscillator. Moments, however, only give limited information about a random process. All the possible information about a random process is contained in probability distributions. We will discuss the Fokker–Planck method for deriving an equation for the probability distribution for random variables in Langevin equations. Over all the Langevin and Fokker–Planck equations are the approach of choice when the random process is best described by an equation of motion with noisy terms.

A different approach to random processes is to work with the probabilities that govern changes in a particle's position, and not to write directly an equation for random variables. This has been called a random-walk or master-equation approach. The focus is on the elemental transition probability for the step of a particle. This is a powerful method

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

because it allows many complicated situations and features to be directly incorporated and analyzed. For example, we start with the case of a random walk where a particle makes jumps of random lengths in random directions, but then add several important generalizations. We let the time between jumps, and the jump velocity, become random. This even includes fractal probability distributions with infinite mean jump lengths and infinite mean waiting times between jumps. We can account for the random walker being restricted to a fractal space or placed in a disordered environment. Reaction processes that annihilate random walkers are also introduced. Over all the random-walk methods are the approach of choice when one wishes to model the elemental transitions in a random process.

1. A BRIEF HISTORY OF RANDOM PROCESSES

Man's conscious deliberation with random processes is as least as old as the ancient game of throwing the bones (David, 1962). The bones in question, called astrogali, are the heel bones of animals including dogs and sheep. These bones fit nicely in the hand and are the forerunners of dice. In fact, the French expression "jeu d'hasard" (games of hazard or chance) comes from the Arabic for dice, "al-zar" (Székely, 1986). In reality, gambling (from dice to actuarial tables to risk-benefit analysis) has always been in the forefront of expanding the frontiers of probability theory. The bones in question have only four faces upon which they come to rest. The four-sided spinning dreidel is the most similar, in this respect, to throwing of the bones. In ancient times a mathematics of probability did not arise. Perhaps this was because the early dice did not have equally likely outcomes and the occurrence of a rare event (landing on a less probable face) was ascribed to luck rather than sparking the development of the idea of the permanence of statistical ratios. On the other hand, a mathematics of probability also did not arise from the ancient Chinese *I Ching,* which has 64 equally likely outcomes. The *I Ching* was used as an oracle and perhaps has too many outcomes to provide the insight to inaugurate a new field of mathematics.

Despite some early writings on dice games, probability got its recognized start with a correspondence between Pascal and Fermat (Todhunter, 1865) in 1654 that cleared up a misconception between the probability to win and the expected winnings. Specifically, the expected number of sixes in four throws of a die is 4/6, which equals the expected number of pairs of sixes in 24 throws of a pair of dice. The probability to throw at least one six in four throws is $1 - (5/6)^4 \cong 0.5177$, and the probability to get at least one pair of sixes in 24 throws of a pair of dice is $1 - (35/36)^{24} \cong 0.4914$. Equal expectations do not imply equal probabilities. Having easily settled this matter, Pascal and Fermat wrestled with other problems such as how to split up the ante in games of chance that are not completed. With more than two players considered, these problems would tax many modern practitioners.

The Pascal–Fermat letters stimulated Huygens to write a treatise on probability theory. This, in turn, piqued the interest of Jacob Bernoulli, who turned his attention to questions of combinations and permutations, and games of chance. His works were published posthumously, in 1713, under the title *Ars Conjectandi* (Art of Conjecture). He introduced the Bernoulli process where a player has a probability p to win in each trial of a game, and probability $q = 1 - p$ to lose. We would today equate this to a random walk on a lattice by keeping track of a player's status along a one-dimensional axis, with success equated to a jump of one unit to the right, and failure a jump of one unit to the left. The probability $p_n(k)$ of k successes in n trials is given by $[n!/(n - k)!k!]p^k q^{n-k}$. In the limit of large n, and for $p = q$, DeMoivre showed, in 1756, that the probability to get $2h$ more successes than failures (or vice versa) in n Bernoulli trials is $(2\pi n)^{-1/2} \exp(-h^2/2n)$. This can be seen by expanding the factorials in $p_n(k)$ with $k = n/2 \pm h$. This probability function is today called a Gaussian because Gauss showed, in 1809, that DeMoivre's result holds in a more general situation. An intimate connection (Einstein, 1905) with these types of random-walk processes and Gaussian probabilities was later made to the phenomenon of diffusion called Brownian motion. For example, in Bernoulli's game (process) consider two players each starting with $R/2$ coins and with equal probability to win or lose a coin in each trial. The mean number of trials, $\langle N(R)\rangle$, before one player loses his fortune can be calculated to be $R(R + 1)/6$, and for large R, $\langle N(R)\rangle \approx R^2$. In Brownian motion, the mean-square distance a particle moves after N steps satisfies $\langle R^2(N)\rangle \approx N$.

Poisson, in the 1830s, found another possible limit for the Bernoulli process; let $p \to 0$ and $n \to \infty$ such that $np \to \lambda$, a constant. Then $p_n(k) = n(n - 1) \cdots (n - k + 1)\, p^k(1 - p)^{n-k}/k! \approx (np)^k e^{-np} e^{pk}/k! \approx \lambda^k e^{-\lambda}/k!$. This is called a Poisson probability distribution. Poisson wrote a book on probability, but devoted to social problems, such as the probability of a jury reaching the correct verdict.

Over in London, Bernoulli's contemporary De Moivre wrote, in 1718, the first edition of his famous treatise on probability, *The Doctrine of Chances,* which provided explicit odds for popular card games and, more importantly, introduced new concepts, including the use of generating functions, Sterling's formula for $\log(n!)$, the derivation of the formula $(\cos\vartheta + i\sin\vartheta)^n = \cos n\vartheta + i \sin n\vartheta$, and the proof that the Gaussian is the many-trial limit of a Bernoulli process (in the 1756 3rd edition). The Gaussian (also called the normal) is the probability distribution for a *sum* of identically distributed random variables with finite second moments. It was not until 1879 that MacAlister introduced the distribution for a *product* of random variables called the log normal distribution, as it is a Gaussian in the logarithm of the variable. De Moivre's other probability book, *Annuities upon Lives,* inaugurated actuarial science. Laplace in his 1815 treatise *Theorie Analytique des Probabilites* continued the tradition of introducing new tools in applied mathematics through probability. Before Laplace, most work in probability focused on the discrete, where all the possible outcomes of a trial could be enumerated. Card and dice games are good examples of this genre. Laplace brought the integral and continuum mathematics to the fore of probability theory with questions like: If an urn contained an infinite number of black and white tickets and one chooses A tickets of which p turn out to be white and q are black, then what is the probability that if an additional B tickets are chosen, m are white and n are black. Laplace's solution,

$$\frac{\int_0^1 x^{p+m}(1 - x)^{q+n}dx}{\int_0^1 x^p(1 - x)^q dx}, \qquad (1)$$

led the way for calculating a wide variety of integrals, many of which are well known today.

Despite opposition to probability as a real mathematics, due to a number of seeming paradoxes, and prohibitions from church and state, due to its gambling connections, the ideas set in motion by the above-mentioned books would spread and be carried on by new generations. Poisson's book found its way to Russia where Chebyshev read it and did his 1846 Master's thesis on probability, and also started a school in this area. His star students were Markov and Lyapunov whose Markov chains and Lyapunov exponents are of great importance in modern work. Kolmogorov, in Russia, placed probability on a secure mathematical foundation. The early 1900s saw the introduction of Brownian motion by Einstein and Bachelier, and reaction kinetics by Smoluchowski. Langevin, in 1908, studied equations of motion with an additive random force. Jumping over potential barriers in noisy systems (Pontryagin *et al.,* 1933; Kramers, 1940) initiated the field of noise-induced transitions. Schottky (1922) studied the phenomenon of shot noise in diodes. Specifically, if electrons arrive at a detector with a rate ν and cause a decaying time-dependent current $i(t)$, then the total measured current is

$$I(t) = \sum_{j=1}^{\infty} i(t - T_j),$$

where T_j represents the past arrival time of the jth electron. This shot noise has an average value of $\nu\int_0^\infty i(t)dt$ and variance $\nu\int_0^\infty i^2(t)dt$. This is also known as Campbell's theorem to mathematicians, and it was introduced in 1909. Chandrasekhar's review of random processes (Chandrasekhar, 1943), including random walks, Brownian motion, escape over barriers, and stellar dynamics, dominated the field for many years. Rice (Wax, 1954) reviewed advances in analyzing noise in electrical devices. The analysis of telephone switching systems, by Fry at Bell Telephone in the 1920s and by Erlang in Denmark in the 1940s, produced advances in the statistical dynamics of congestion of networks. Fisher and Tippett (1928) introduced extreme-value theory to find the largest or the smallest value from a set of random variables. Statistical physics focused on noise, for its own intrinsic properties, and introduced the master equation, entropy, and fluctuation–dissipation theorems. Lévy, in the 1920s and 1930s, introduced scale-invariant random walks and distributions for random variables with infinite moments, which were an important precursor for the field of fractals. Such a wide variety of problems depend on probability that we cannot cover all the major topics and will omit many important problems including random-matrix theory, random fields, Ising-model dynamics, combinatorics, queuing theory, reliability theory, stochastic resonance, Monte Carlo techniques, signal processing, random-growth models, multifractals, and quantum probabilities.

2. GAUSSIAN RANDOM WALKS

Bernoulli's gambling problem allowed one's fortune to increase or decrease in discrete amounts. This is equivalent to a random walk on a lattice. Consider a random walk that starts at the origin of a lattice. We begin by choosing the probability $p(l)$ for the walker in one transition to change its position by l lattice spacings. All probabilities can be written in terms of $p(l)$ and the initial condition. The probability to be at site l after $n + 1$ steps is denoted by $P_{n+1}(l)$ (see Fig. 1 for a two-dimensional example), which we can write in one dimension as

$$P_{n+1}(l) = \sum_{l'} P_n(l - l')p(l'). \tag{2}$$

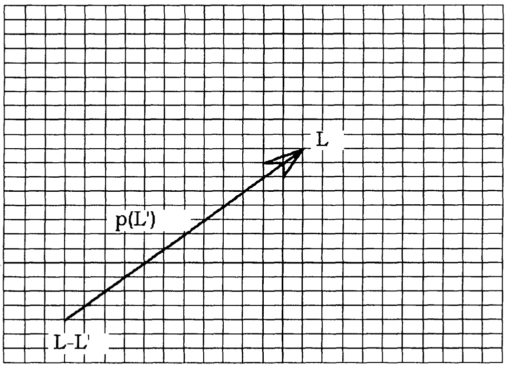

FIG. 1. A finite part of an infinite two-dimensional lattice. For a random walker on this lattice to reach site L after $n + 1$ steps, the walker first reaches some intermediate site $L - L'$ after n steps and then makes a transition with a displacement of L' at the next step with probability $p(L')$. A sum over all possible L' is necessary, as performed in Eq. (2), to obtain the probability to be at site L on the nth step.

Let us introduce the discrete Fourier transform $\tilde{p}(k)$ defined by

$$\tilde{p}(k) \equiv \sum_{l} \exp(ikl)p(l). \tag{3}$$

Fourier transforming $P_n(l)$ we have, since Eq. (2) is in the form of a convolution,

$$\tilde{P}_n(k) = [\tilde{p}(k)]^n$$

or

$$P_n(l) = \frac{1}{2\pi}\int_{-\pi}^{\pi} [\tilde{p}(k)]^n \exp(-ikl)dk. \tag{4}$$

For an unbiased random walk with no preferred direction, $\tilde{p}(k)$ is real and has a maximum value of 1 at $k = 0$. For example, a random walk with equal probabilities of $\frac{1}{2}$ to jump to the right or left one lattice site has $\tilde{p}(k) = \cos(k)$. For large n, $[\tilde{p}(k)]^n$ is small, except at and near $k = 0$. This implies that for large n (many jumps of a random walk) the integral in Eq. (4) is dominated by small values of k, and so we use a small-k expansion of $\tilde{p}(k)$ to calculate the large-n behavior of $\tilde{P}_n(k)$. We start by expanding the exponential in Eq. (4) for small k,

$$\begin{aligned}\tilde{p}(k) &= \sum_{l} [1 + ikl - \tfrac{1}{2}k^2l^2 + \cdots]p(l)\\ &= 1 + ik\langle l\rangle - \tfrac{1}{2}k^2\langle l^2\rangle + \cdots,\end{aligned} \tag{5}$$

where the moments of the jump distributions are defined as $\langle l^n\rangle \equiv \Sigma_l l^n p(l)$. For the sake of simplicity let us assume that jumps occur with equal probabilities to the right and left so that $\langle l\rangle = 0$, and that $\langle l^2\rangle$ is finite. Then, for small k values,

$$[\tilde{p}(k)]^n \approx \exp\left(-\frac{n}{2}\langle l^2\rangle k^2\right), \tag{6}$$

which is Gaussian in k. We can now extend the limits on the integral in Eq. (4) from $-\infty$ to $+\infty$ without significant error because the integrand, for large n, has a sharp maximum around $k = 0$. This yields the important result

$$P_n(l) = \frac{1}{\sqrt{2\pi\langle l^2\rangle n}} \exp\left(-\frac{l^2}{2\langle l^2\rangle n}\right) \tag{7}$$

for a many-step symmetric random walk with a finite $\langle l^2\rangle$ (see Fig. 2). This result of adding identically distributed random variables with finite second moments and arriving at a Gaussian distribution is known as the central limit theorem. This so impressed F. Galton that he wrote, "I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the Law of Frequency of Error. The law would have been personified by the Greeks and deified, if they had known of it. The larger the mob and the greater the apparent anarchy, the more perfect its sway." The Gaussian was originally applied to discuss errors in measurement, with its peak being the most probable value of the measurement and its variance related to typical errors in the measurement. The Maxwell–Boltzmann distribution is Gaussian in particle velocities, but the whole distribution has physical meaning, so that being away from the mean is not an error. In fact the whole distribution is a prediction. This was a major advance in thought.

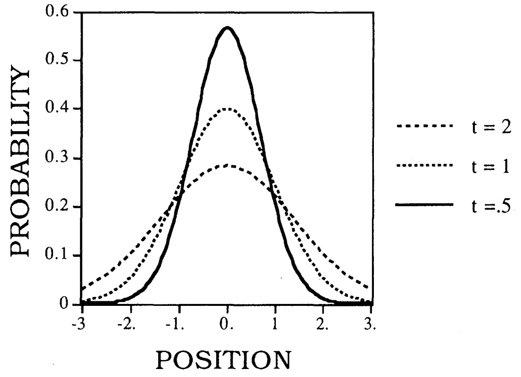

FIG. 2. Gaussian spreading. We let the number of steps become a continuous variable equivalent to the time and plot the spreading with time of a Gaussian probability. The variance of the distribution grows linearly with time.

It can be shown, via the Berry–Esséen theorem (Feller, 1971), that the value of n at which Gaussian behavior emerges is $n \approx (25/4)\langle |l^3| \rangle^2/\langle l^2\rangle^3$. The probability $P_n(l)$ satisfies the diffusion equation when we let the number of steps n and the distance l become continuous variables,

$$\frac{\partial P(l,n)}{\partial n} = D\frac{\partial^2 P(l,n)}{\partial l^2}, \tag{8}$$

with $D = \langle l^2\rangle/2\langle t\rangle$ where $1/\langle t\rangle$ is introduced as a jump rate. This is also known as the heat equation, and Bachelier, in 1900, was amazed to discover that probability could radiate in the same fashion as heat.

While the central limit theorem is very powerful, there are many situations where it does not apply. One can ask questions about the first time a random walk reaches a certain distance from its origin, or the maximum of a set of random variables, or the conditional probability for a random walker to be at a site now, given that it was at another site at an earlier time, or the probability distribution for a random walker whose mean jump distance is infinite, etc. This means that, although the Gaussian distribution plays a central role in probability theory connected to summing random variables, it is only one of many important limit theorems.

The above analysis was for a simple walk on a one-dimensional regular lattice. We began by writing down directly the transition probability for a single step. A different approach for investigating the random motion of a particle is to employ a stochastic differential equation of motion for a particle's position or velocity (the Langevin approach) and then calculate via a partial differential equation (the Fokker–Planck equation) the probability for the particle being at position x at time t. We turn to the Langevin approach now, before returning to more complex random-walk models.

3. LANGEVIN EQUATIONS

One of the major themes in probability is the mathematics of Brownian motion and its variants. The original work of Einstein involved a large Brownian particle in a fluid being buffeted continuously by smaller fluid particles. If the Brownian particle has mass M and random velocity, denoted by v, then the equation of motion for the Brownian particle's dynamics is

$$M\frac{dv}{dt} = -fv + F(t), \tag{9}$$

where mass times acceleration is on the left-hand side (LHS) and the forces are on the right-hand side (RHS). The term f is the friction coefficient and $F(t)$ is a random force representing random collisions of the Brownian particle with the smaller fluid particles. An $F = Ma$ equation that involves random forces and/or parameters is called a Langevin equation. Although, one can readily solve for $v(t)$ as

$$v(t) = v(0)e^{-\beta t} + e^{-\beta t}\int_0^t e^{\beta\tau}A(\tau)d\tau, \tag{10}$$

where $\beta = f/M$ and $A = F/M$, one cannot directly evaluate $v(t)$ because it is a random variable, which is written in terms of $A(t)$, the random force per mass. One can, however, determine the probability distribution for values of v, or of related random variables such as the position $r(t)$, which is related by $v(t) = dr(t)/dt$. To proceed further, one must consider the statistical properties of the noise. A straightforward case, called white noise, is uncorrelated noise. A white-noise random force has the properties, where the brackets indicate a time average,

$$\begin{aligned}&\langle F(t)\rangle = 0,\\ &\langle F(t=0)F(t)\rangle = 2D\delta(t).\end{aligned} \tag{11}$$

The choice of white noise yields $\langle v(t)\rangle = v(0)e^{-\beta\tau}$, and $\langle v^2(t)\rangle = \beta D + [v(0)^2 - \beta D] \times \exp(-2\beta t))$. In the long-time (equilibrium) limit $\frac{1}{2}M\langle v^2\rangle$ must be equal to $k_BT/2$, the energy of the system. Thus the choice of D is not arbitrary, but must be $D = k_BT/\beta M$. One can also solve for the moments of $r(t)$. Integrating Eq. (10) one obtains

$$\begin{aligned}r(t) = r(0) + \frac{v(0)}{\beta}(1 - e^{-\beta t})\\ - \frac{1}{\beta}e^{-\beta t}\int_0^t e^{+\beta\xi}A(\xi)d\xi + \frac{1}{\beta}\int_0^t A(\xi)d\xi.\end{aligned} \tag{12}$$

Taking averages over values of A and initial velocities $v(0)$ one obtains $\langle r(t)\rangle = r(0)$. Squaring Eq. (12) and again taking double averages over the random forces and initial velocities one obtains the classic Brownian-motion result,

$$\langle r^2(t)\rangle = \frac{2k_BT}{M\beta}t = \frac{2k_BT}{f}t = 2Dt, \tag{13}$$

plus terms that decay exponentially with time. In his theory of Brownian motion, Einstein wrote specifically $\beta = 6\pi a\eta/M$, which is Stokes's law for a spherical particle of radius a and mass M, in a fluid with viscosity η.

The next challenge is to find the probability distribution of these stochastic variables, not just their moments. Before we address this matter, we mention that Eq. (9) can be easily generalized to govern Brownian motion in a potential by adding the force term $K(x)/M$ to the RHS. For a harmonic potential (the Ornstein–Uhlenbeck problem) $K(x)/M = -\omega^2x$, and this produces an oscillatory decay to the equilibrium value of zero for the mean position (and also the mean velocity), i.e.,

$$\begin{aligned}\langle r(t)\rangle = \frac{Mv(0)}{\omega}e^{-(1/2)\beta t}\sin\omega t\\ + \frac{r(0)}{\omega}e^{-(1/2)\beta t}\left(\omega\cos\omega t + \frac{\beta}{2}\sin\omega t\right).\end{aligned} \tag{14}$$

Expressions can be similarly calculated for the second spatial and velocity moments with oscillatory decays to equilibrium values, as opposed to the exponential decays when only the frictional force is present.

3.1 Kubo Oscillator

Consider randomness entering through a multiplicative factor in a stochastic differential equation. A classic problem is an oscillator (Kubo, 1962) with amplitude $x(t)$ and fluctuating frequency $\omega(t)$. This process satisfies the equation

$$\frac{dx(t)}{dt} = i\omega(t)x(t),$$

with the first moment given by

$$\langle x(t)\rangle = \left\langle \exp\left(i \int^{t} \omega(t')dt'\right)\right\rangle. \tag{15}$$

Because $\omega(t)$ is fluctuating, $x(t)$ is a random variable. We will not calculate the probability distribution for x, but will only concentrate on its first moment here, and the effect of colored noise on this behavior. For $\omega(t) = \omega_0 + \delta\omega(t)$, with $\langle\delta\omega(t)\rangle = 0$, the frequency fluctuates around the mean value ω_0. If $\delta\omega(t)$ possesses Gaussian statistics, then

$$\begin{aligned}\langle x(t)\rangle &= e^{i\omega_0 t}\exp\left(-\tfrac{1}{2}\int_0^t dt' \int_0^t dt''\langle\delta\omega(t')\delta\omega(t'')\rangle\right)\\ &= e^{i\omega_0 t}\exp\left(-\int_0^t (t - t')\langle\delta\omega(t')\delta\omega(0)\rangle dt'\right).\end{aligned} \tag{16}$$

If $\langle\delta\omega(t)\delta\omega(0)\rangle = 2\lambda\delta(t)$, then it follows that

$$\langle x(t)\rangle = e^{i\omega_0 t}e^{-\lambda t}. \tag{17}$$

The relaxation function, defined by $\Phi(t) = \exp(-\lambda t)$, leads to the line-shape spectrum $I(\omega)$, defined as

$$\begin{aligned}I(\omega - \omega_0) &= \frac{\mathrm{Re}}{2\pi}\int_{-\infty}^{\infty} \exp[i(\omega - \omega_0)]\Phi(t)dt\\ &= \frac{1}{\pi}\frac{\lambda^2}{(\omega - \omega_0)^2 + \lambda^2}.\end{aligned} \tag{18}$$

The effect of the noisy frequency was to broaden the line shape around the peak at ω_0.

Temporally correlated "colored" noise can lead to other line shapes. If $\langle\delta\omega(t)\delta\omega(0)\rangle = (D/\tau)\exp(-|t|/\tau)$, then for $t \ll \tau$, $(\tau/D) \times \langle\delta\omega(t)\delta\omega(0)\rangle \approx 1$, and so Eq. (16) yields, for short times,

$$\langle x(t)\rangle = e^{i\omega_0 t}\exp\left(-\frac{D}{2\tau}t^2\right) \quad \text{for } t \ll \tau. \tag{19}$$

The line shape generated by this Gaussian relaxation at short times is a Gaussian at large frequencies. For long times $t \gg \tau$, $\langle\delta\omega(t)\delta\omega(0)\rangle \approx 0$, i.e., the correlation has decayed away, and so one can replace the upper limit in the last integral in Eq. (16) by infinity and recover the exponential decay of the delta-correlated noise case (effectively the $\tau = 0$ case) and Lorentzian line shape of Eq. (18) at low frequencies. This crossover from a Lorentzian to a Gaussian line shape at $\omega = 1/\tau$ illustrates the crucial role that correlated "colored" noise can play in random processes.

Equation (18) is also called a power spectrum. In general, take a random time series $h(t)$ whose mean is zero. Let $h(t)$ be zero outside a time interval T and represent $h(t)$ in terms of its Fourier transform

$$h(t) = \int_{-\infty}^{\infty} \exp(2\pi i f t)g(f)df.$$

The power spectrum $S(f)$ can be defined as

$$S(f) = \lim_{T\to\infty}\frac{2}{T}|g(f)|^2. \tag{20}$$

That also

$$4\mathrm{Re}\int_{-\infty}^{\infty}\langle h(t + \tau)h(t)\rangle \exp(2\pi i f\tau)d\tau = S(f)$$

is called the Wiener–Khintchine theorem. If the correlation function is the product of an electrical current at two different times, then its power spectrum reflects the electric power as a function of frequency; hence the appellation of power. The Lorentzian power spectrum of Eq. (18) arises from a correlation function that decays exponentially. In many complex systems one finds a noise power spectrum, for low frequencies, of the form $S(f) \approx 1/f^{\alpha}$ where α is close to unity. This is called one-over-f noise. One method to derive $1/f$ noise is to have an exponentially decaying correlation $\exp(-t/\tau)$ with a distribution of correlation times $\rho(\tau) \approx \tau^{-1-\alpha}$. This produces $S(f) \approx 1/f^{1-\alpha}$. It is still an open question to determine the physics underlying $1/f$ noise in many complex systems (Weissman, 1988).

4. PROBABILITY EQUATIONS

4.1 The Chain Equation and the Master Equation

To describe fully a stochastic process $X(t)$ one must know all the joint probabilities,

$$P_n(x_1,t_1;\cdots;x_n,t_n) \quad \text{for all } n,$$

i.e., that at time t_i, the variable $X(t)$ has the value x_i, for $i = 1, \ldots, n$. Alternatively, one can describe $X(t)$ by defining joint conditional probabilities W_n for $X(t)$ to have the value x_n at time t_n when the earlier values x_i at t_i are given for $i = 1, \ldots, n - 1$. The probability P_n can be written as

$$P_n(x_1,t_1;\cdots;x_n,t_n) = W_n(x_n,t_n|x_1,t_1;\cdots;x_{n-1},t_{n-1}) \times P_{n-1}(x_1,t_1;\cdots;x_{n-1},t_{n-1}). \tag{21}$$

Knowing P_1 and all the W_n also completely specifies $X(t)$. A Markov process is defined by

$$W_n(x_n,t_n|x_1,t_1;\cdots;x_{n-1},t_{n-1}) = W_2(x_n,t_n|x_{n-1},t_{n-1}),$$

where the value of x_n depends only on x_{n-1} and the probability of transitioning from x_{n-1} to x_n. The earlier information of the x_{n-i} does not affect the outcome of the next step in a Markov process. For a Markov process the conditional probabilities satisfy the famous Bachelier–Smoluchowski–Chapman–Kolmogorov chain equation

$$W_2(x,t|x_0,t_0) = \int W_2(x,t|x',\tau)W_2(x',\tau|x_0,t_0)dx', \tag{22a}$$

where we considered going from x_0 to x in two steps, and we have integrated over all possible intermediate states x'. One can also write the chain equation as

$$\frac{\partial P(x,t)}{\partial t} = \int [K(x,x')P(x',t) - K(x',x)P(x,t)]dx', \tag{22b}$$

which is called a master equation (Oppenheim *et al.*, 1977). To derive the master equation from the chain equation all the initial conditions have been integrated out, so that $P(x,t)$ is the probability to be at state x at time t independent of the initial condition. The quantity $K(x,x')$ is the transition rate (it has the dimension of a frequency) for the system to change from state x' to state x. The derivative measures the rate of change with time of the probability to be in state x. The first term in the integral gives the rate of the system moving into state x from state x' (integrated over all x') times the probability that the system is in state x', while the second term calculates the rate for leaving state x times the probability that the system is in state x. In our discussion of continuous-time random walks we will study semi-Markovian (temporal, but no spatial, memory) processes where a trapping of a random walker occurs between each of its jumps. For these random-walk processes, an integro-differential master equation arises with an integral over all possible waiting times in a trap [see Eq. (52)].

4.2 Fokker–Planck Equations

From the chain equation, for only small transitions in x allowed, we can derive a second-order differential equation for $P(x,t)$. Let us denote W_2 in Eq. (22a) as P and look, following Wang and Uhlenbeck (Wax, 1954), at

$$\int dyR(y)\frac{\partial P(y,t|x,0)}{\partial t},$$

where $R(y)$ is an arbitrary function that asymptotically goes to zero at infinity. We can write the above derivative as

$$\lim_{\Delta t\to 0}\frac{1}{\Delta t}[P(y,t+\Delta t|x,t) - P(y,t|x,t)] = \lim_{\Delta t\to 0}\frac{1}{\Delta t} \times \left[\int [P(y,t+\Delta t|z,t')P(z,t'|x,t)dz - P(y,t|x,y)\right], \tag{23}$$

where we have used the chain equation (22a). Expanding $R(y)$ in a Taylor series up to the second derivative and using Eq. (23) for the derivative produces the following equation for $P(x,t)$, called the Fokker–Planck equation:

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}[A(x)P(x,t)] + \tfrac{1}{2}\frac{\partial^2}{\partial x^2}[B(x)P(x,t)], \tag{24}$$

where $A(x)$ is related to the first moment of $P(x,t)$ as

$$A(x) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int P(y,\Delta t|x,t=0)(x-y)dy$$

and $B(x)$ is the same expression as above, except that the second moment of P appears through $(x - y)^2$ replacing $(x - y)$. These two terms on the RHS of Eq. (24) represent drift and diffusion in a force field. Although the expansion for $R(y)$ can be carried out to higher-order terms, they are usually small unless large jumps occur. In that case the full master equation should be used, as we do in our discussion of fractal processes. Setting the time derivative equal to zero, in Eq. (24), one has the following steady-state solution:

$$\lim_{t\to\infty} P(x,t) = \frac{\text{const.}}{B(x)} \exp\left(\int^x \frac{A(x')}{B(x')}\,dx'\right), \tag{25}$$

where the constant is chosen to insure normalization. When the first moment vanishes because there is no preferred direction, and the second moment is a constant, denoted by D, the Fokker–Planck equation reduces to the familiar diffusion equation.

The Ornstein–Uhlenbeck process for random motion in a harmonic potential has $A(x) = -\omega^2 x$, $B =$ constant. The solution of the Fokker–Planck equation for $P(x,t)$ is then

$$P(x,t) = \frac{1}{\sqrt{2\pi\sigma^2(t)}} \exp\left\{-\frac{x^2}{2\sigma^2(t)}\right\}, \tag{26a}$$

where

$$\sigma^2(t) = \frac{D}{\omega^2}[1 - \exp(-2\omega^2 t)]. \tag{26b}$$

In the long-time, steady-state limit $\omega \to 0$, $\sigma^2(t) \to 2Dt$, which brings us back to the Brownian-motion limit.

In general, a Langevin equation of the form

$$\frac{dx(t)}{dt} = F(x(t)) + G(x(t))\eta(t), \tag{27}$$

with

$$\langle \eta(t)\eta(t')\rangle = 2D\delta(t - t'),$$

has the probability $P(x,t)$ governed by the Fokker–Planck equation

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial F(x)P(x,t)}{\partial x} + D\frac{\partial}{\partial x} G(x)\frac{\partial G(x)P(x,t)}{\partial x}. \tag{28}$$

However, when a finite-correlation-time "colored" noise, e.g.,

$$\langle \eta(t)\eta(0)\rangle = (D/\tau)\exp(-|t|/\tau), \tag{29}$$

is required, a more complicated situation arises. Many techniques (Moss *et al.*, 1990) have been derived to treat stochastic differential equations with colored noise using expansions and infinite resummation in powers of D or τ.

5. RANDOM WALKS

5.1 Discrete-Time

Let us return to Eq. (2) for $P_n(l)$ and define the *generating function* $G(l;z)$ as

$$G(l;z) \equiv \sum_{n=0}^{\infty} P_n(l) z^n. \tag{30}$$

Multiplying both sides of Eq. (2) by z^{n+1} and summing over all n yields, for a walker that starts at the origin,

$$G(l;z) - z\sum_{l'} G(l - l';z)p(l') = \delta_{l,0}, \tag{31}$$

i.e., the random-walk generating function satisfies a Green's-function equation. Fourier transforming Eq. (31) gives an algebraic equation for $\tilde{G}(k;z)$, with the solution

$$\tilde{G}(k;z) \equiv \sum_{l} \exp(ikl) G(l;z) = \frac{1}{1 - z\tilde{p}(k)}. \tag{32}$$

For the case of nearest-neighbor transitions $p(l) = \frac{1}{2}[\delta_{l,-1} + \delta_{l,+1}]$, we have

$$\tilde{p}(k) = \cos(k) \tag{33}$$

and

$$P_n(l) = \frac{n!}{[\frac{1}{2}(n + l)]![\frac{1}{2}(n - l)]!} 2^{-n}. \tag{34}$$

This is the expression DeMoivre analyzed for $n/2 + l/2$ successes (jumps to the right) and $n/2 - l/2$ failures (jumps to the left) in n Bernoulli trials. He showed that $P_n(l)$ is Gaussian in the large-n limit. This is true for any symmetric random-walk process with finite mean squared jump lengths, as Gauss proved in 1809 in his central limit theorem.

The generating-function formalism is also good for calculating first-passage times. The probability $F_n(l)$ that a walker reaches site l for the first time on the nth step enters the following equation:

$$P_n(l) = \sum_{j=1}^{n} F_j(l)P_{n-j}(0) + \delta_{n,0}\delta_{l,0}, \tag{35}$$

which accounts for being at l after n steps by first reaching l after j steps and then leaving and returning to l in $n - j$ steps (for all possible j values). Multiplying Eq. (35) by z^n, and summing over all n, gives the generating-function equation

$$G(l;z) = \delta_{l,0} + F(l;z)G(l = 0;z), \tag{36}$$

where

$$F(l;z) \equiv \sum_{j=0}^{\infty} F_j(l)z^j.$$

This shows that knowledge of $G(l;z)$ allows the calculation of the first-passage-time generating function. The Green's function for return to the origin, which appears on the RHS of Eq. (36), has been calculated, for nearest-neighbor jumps, for z close to 1:

$$G(l = 0;z) = \left(\frac{1}{2\pi}\right)^D \int_0^{2\pi} \cdots \int_0^{2\pi} \frac{d^Dk}{1 - \frac{z}{D}[\cos(k_1) + \cdots + \cos(k_D)]}$$
$$= \begin{cases} (1 - z^2)^{-1/2} & \text{for } D = 1, \\ \approx -\frac{1}{\pi}\log(1 - z) & \text{for } D = 2, \\ \approx 1.516 - \frac{3}{\pi}\sqrt{\frac{3}{2}(1 - z)} & \text{for } D = 3. \end{cases} \tag{37}$$

The 2D result is for a square lattice and the 3D result is for a cubic lattice. The manner in which $G(l = 0;z)$ diverges near $z = 1$ is important for calculating first–passage-time probabilities. Consider Polya's question of what is the probability for a symmetric random walker with $\langle l \rangle = 0$, on a regular lattice, to return to its origin. This probability of return, R, is given as

$$R = \sum_{n=0}^{\infty} f_n(0) = F(l = 0;z = 1) = 1 - \frac{1}{G(0;1)}, \tag{38}$$

where we have used Eq. (36). We see from Eq. (37) that $G(0;1)$ diverges in one and two dimensions, and is finite in higher dimensions, yielding Polya's famous result that return to the origin is assured with probability 1 in one and two dimensions, but not in higher dimensions. The square-root divergence of the one-dimensional $G(0;1)$ in Eq. (37) implies that the probability that the walker has not returned by step n is

$$F_n(0) = \sum_{j=n+1}^{\infty} f_j(0) \approx \left(\frac{1}{n}\right)^{1/2}, \tag{39}$$

causing, in one dimension, that the mean number of steps

$$\langle n \rangle = \sum_{n=0}^{\infty} nf_n(0)$$

for the first return of a random walker to the origin is infinite. The walker returns with probability 1, but the mean time to do this is infinite.

What if N walkers start at the origin? Is the mean time for any one of this set to return to the origin also infinite (Lindenberg *et al.*, 1980)? The probability $f_{N,n}(0)$ that any walker from a group of N walkers returns on the nth step is, using Eq. (39),

$$f_{N,n}(0) = F^N_{n-1}(0) - F^N_n(0), \tag{40}$$

where the N on the LHS denotes the number of walkers and N is a power on the RHS.

The RHS is the difference between the probabilities that none of the walkers has returned by the $(n - 1)$st and the nth step. This difference is the probability that at least one walker returned on the nth step. The mean first time for a return to the origin from the set of N walkers is

$$\langle T_N \rangle = \sum_{n=1}^{\infty} n f_{N,n}(0) = \sum_{n=0}^{\infty} F_n^N(0). \qquad (41)$$

In one dimension $F_n(0) \approx \sqrt{1/n}$, and so the divergence of the mean return time for the set of N walkers depends on the series

$$\sum_{n=1}^{\infty} \left(\frac{1}{n}\right)^{N/2}. \qquad (42)$$

For three (or more) walkers, Eq. (42) determines that the mean time for return to the origin is finite. This further explores Polya's classic study of the effect of dimensionality on random-walk quantities.

There are several directions in which the generating-function approach to random walks can be naturally extended. One is to include finite memory of the walker's jump history. This can be accomplished through the use of internal states at each lattice site, say to keep track of whether a random walker came from the left or right. In one dimension if the probability for a walker to continue in the same direction is $p < 1$, and to change direction is $q = 1 - p$, then the generating-function method finds after many jumps that the probability density for the walker's position is Gaussian, as in Eq. (7), but with the extra factor p/q in the diffusion constant D.

Restricted reversal correlation can also be handled in terms of internal states. Suppose, in two dimensions, that a walker is not allowed to reverse its direction to go back immediately to its previous position. This is handled by having four states in a unit cell representing the four directions a walker could have come from, and not having a connecting path back to the reversed position. The mean squared displacement is doubled from the unrestricted walk, and for a periodic lattice with coordination number q, the mean squared displacement is enhanced by the factor $q/(q - 2)$ (Domb and Fisher, 1958). Correlation over a finite number of steps (Montroll, 1950) will only modify the diffusion constant, and not change the simple linear dependence on n (the number of steps) of the mean square displacement. One needs infinite correlation to go beyond simple asymptotic Gaussian probabilities for random walks with simple steps on a regular lattice. The self-avoiding random walk is one such example. This model forbids the walker to cross its own trajectory at any point. For a self-avoiding random walk on a three-dimensional cubic lattice one finds $\langle l^2(n) \rangle \approx n^{\gamma}$ with $\gamma \approx 6/5$.

5.2 Continuous-Time

So far we have treated random walks with steps taken at regular time intervals τ, 2τ, 3τ, ... We can generalize this by introducing a probability $\psi(t)dt$ for the interval between jumps to be between t and $t + dt$. The quantity $\psi(t)$ is called a waiting-time probability density, and it must be integrated over t to yield a probability. A quantity that we will find useful is $\Phi(t)$, the probability that a jump has not occurred by time t, since the previous jump. This is given by

$$\Phi(t) = 1 - \int_0^t \psi(\tau)d\tau. \qquad (43)$$

The probability density $\psi_2(t)$ that the interval between two consecutive jumps is t is given by

$$\psi_2(t) = \int_0^t \psi(t - \tau)\psi(\tau)d\tau. \qquad (44)$$

Similarly, we introduce $\psi_n(t) = \int_0^t \psi_{n-1}(t - \tau)\psi(\tau)d\tau$, and the Laplace transform of $\psi_n(t)$,

$$\psi_n^*(s) \equiv \int_0^{\infty} \psi_n(t) \exp(-st)dt = [\psi^*(s)]^n. \qquad (45)$$

If $\psi(t) = \lambda \exp(-\lambda t)$ then $\Phi(t) = \exp(-\lambda t)$. This represents a pure random process with a constant jump rate λ. By a pure process it is meant that the next jump of a walker is independent of whether you start the random walker at $t = 0$ or observe it at a random time and do not know how much time has elapsed since the last jump of the walker. This is equivalent to waiting for a bus that goes in a loop. Do you have to wait longer for the bus to return if you see that the bus just left, or if you come and no bus is present? If the time between returns to your stop is governed by an exponential $\psi(t)$ then both cases have the same result. This is only true for the exponential $\psi(t)$, and for this case the probability $N_n(t)$ that n jumps have occurred by time t is $N_n(t) = \int_0^t \psi_n(t - \tau)\Phi(\tau)d\tau$. In Laplace space,

$$N_n^*(s) = [\psi^*(s)]^n \frac{1}{1 - \psi^*(s)} = \frac{\lambda^n}{(s + \lambda)^{n+1}}. \tag{46a}$$

Taking the inverse Laplace transform we find for $N(t)$ that

$$N_n(t) = \frac{(\lambda t)^n \exp(-\lambda t)}{n!}, \tag{46b}$$

which is the Poisson distribution.

It is useful to define the moments of $\psi(t)$,

$$\langle t^n \rangle \equiv \int_0^\infty t^n \psi(t)dt. \tag{47}$$

With these preliminaries, we can now write the probability density $Q(l,t)$ to reach site l *exactly* at time t, via the relationship

$$Q(l,t) = \sum_{l'} \int_0^t Q(l - l',t - \tau)p(l')\psi(\tau)d\tau + \delta_{l,0}\delta(t). \tag{48}$$

The RHS accounts for the walker reaching the site $l - l'$ at time $t - \tau$ and then a time τ later performing a jump of displacement l' that brings the walker to site l exactly at time t. The delta functions on the RHS take into account the initial condition of the walker being at the origin. In Laplace–Fourier space, Eq. (48) has the form

$$\tilde{Q}^*(k,s) = \frac{1}{1 - \tilde{p}(k)\psi^*(s)}, \tag{49}$$

which is equivalent to the Green's function for the discrete-time random walk when in Eq. (32) we set $z = \psi^*(s)$. The Green's function acting for a time t gives the probability evolution for that time interval. It is also called a propagator. The probability $P(l,t)$ to be at site l at time t is different from $Q(l,t)$ because the walker may have reached site l at an earlier time $t - \tau$, and then no jump occurs during the interval $t - \tau$ to t. We can write $P(l,t)$ as

$$P(l,t) = \int_0^t Q(l,t - \tau)\Phi(\tau)d\tau. \tag{50}$$

In Fourier–Laplace space the probability is

$$\tilde{P}^*(k,s) = \tilde{Q}^*(k,s)\frac{1 - \psi^*(s)}{s} = \frac{1}{1 - \tilde{p}(k)\psi^*(s)} \frac{1 - \psi^*(s)}{s}. \tag{51}$$

The above analysis describes a continuous-time random walk (CTRW) (Montroll and Weiss, 1965) (see Fig. 3).

We can write the probability in Eq. (51) in terms of a nonlocal-time master equation with a transition rate $p(r)K(t)$, i.e., which generalizes Eq. (22b) for the master equation:

$$\frac{\partial P(r,t)}{\partial t} = \int_0^t \left[-P(r,t) + \sum_{r'} P(r',t)p(r - r') \right] \times K(t - \tau)d\tau, \tag{52}$$

if we set in Laplace space

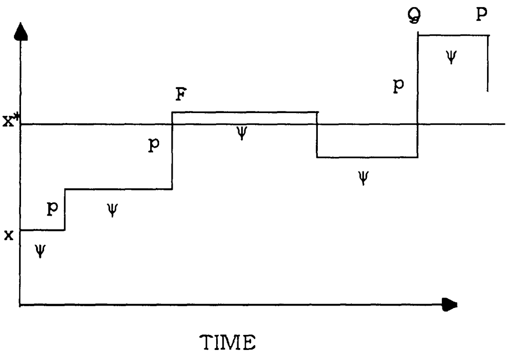

FIG. 3. A schematic plot of position x vs time t for a continuous-time random walk. The walker first waits with no movement for a time governed by $\psi(t)$, the waiting-time density. Next the walker jumps a distance r, governed by $p(r)$, then the walker again waits a random time. We choose a certain level x^*, and $F(x^*,t)$ denotes the first-passage time to this level. $Q(x,t)$ represents reaching position x exactly at time t, and $P(x,t + \tau)$ represents the probability of being at x at time $t + \tau$, which includes having arrived at an earlier time and not having left by time $t + \tau$.

$$K^*(s) = s\psi^*(s)/[1 - \psi^*(s)]. \tag{53}$$

An analysis can show that Eqs. (52) with (53) (the master equation) and (51) (the continuous-time random walk) are equivalent. The temporal memory relates how transitions (jumps) of the random walker at time t are related to the previous history of the random process. The case of $\psi(t) = \lambda \exp(-\lambda t)$ implies $K(t) = \lambda\delta(t)$. This $K(t)$, coupled with choosing $p(r)$ connecting only nearest neighbors on a one-dimensional lattice with equal probability, has the solution $P(l,t) = e^{-\lambda t}I_l(\lambda t)$, where I_l is the modified Bessel function. An analysis of the long-time behavior of the Bessel function demonstrates, in accord with the central limit theorem, that this probability asymptotically becomes Gaussian at long times.

Moments of $P(l,t)$ can be directly calculated from $\tilde{P}^*(k,s)$, e.g.,

$$\begin{aligned}\langle l(t)\rangle &= \sum_l lP(l,t) = \lim_{k\to 0}\left[-i\frac{\partial}{\partial k}\sum_l \exp(ikl)P(l,t)\right]\\ &= \lim_{k\to 0}\left[-i\frac{\partial}{\partial k}\mathscr{L}^{-1}\tilde{P}^*(k,s)\right]\\ &= \bar{l}\mathscr{L}^{-1}\left\{\frac{\psi^*(s)}{s[1-\psi^*(s)]}\right\},\end{aligned} \tag{54}$$

where the mean distance per jump is

$$\bar{l} = -i\frac{\partial}{\partial k}\tilde{p}(k=0)$$

At long times if $\bar{l}$ is not zero, and $\bar{t}$ is finite, the walker has a drift,

$$\langle l(t)\rangle \approx (\bar{l}/\bar{t})t. \tag{55}$$

The variance, $\sigma^2(t) = \langle l^2(t) - \langle l(t)\rangle^2\rangle$, can be shown to behave asymptotically as

$$\sigma^2(t) \approx \left\{\frac{\overline{l^2}}{\langle t\rangle} + \left(\frac{2\bar{l}^2}{\langle t\rangle}\right)\left[\frac{1}{2}\frac{\langle t^2\rangle}{\langle t\rangle^2} - 1\right]\right\}t. \tag{56}$$

In the later sections we will explore ways in which this result can be modified either through infinitely long jumps or waiting times, or through walks on fractal lattices.

One of the strengths of the generating-function formalism is that many quantities can be expressed in terms of the Green's function. We now show how to calculate the first-passage time within the CTRW formalism. Let $F(l,t)$ be the probability density to reach site l for the first time at time t. This is related to the first passage in n steps (with a sum over all n) times the probability density for the nth step to occur at time t,

$$F(l,t) = \sum_{n=0}^{\infty} F_n(l)\psi_n(t). \tag{57}$$

In Laplace space, we can write $F^*(l,s)$ in terms of the first–passage-time generating function $\mathscr{F}$ as

$$\begin{aligned}F^*(l,s) &= \int_0^{\infty} F(l,t)\exp(-st)dt\\ &= \sum_{n=0}^{\infty} F_n(l)[\psi^*(s)]^n = \mathscr{F}(l;\psi^*(s)).\end{aligned} \tag{58}$$

From Eq. (36) we write

$$\mathscr{F}(l,\psi^*(s)) = \frac{G(l,\psi^*(s)) - \delta_{l,0}}{G(l = 0,\psi^*(s))} \tag{59}$$

or in Fourier space,

$$\tilde{\mathscr{F}}(k,\psi^*(s)) = \frac{[1 - \tilde{p}(k)\psi^*(s)]^{-1} - 1}{G(l = 0,\psi^*(s))}. \tag{60}$$

In one dimension, using Eq. (37) for $G(l = 0;z)$, we have for s, $k \to 0$

$$\tilde{F}^*(k,s) \approx \frac{1}{\sqrt{2\langle t\rangle s}\,(\frac{1}{2}k^2\overline{l^2} + s\langle t\rangle)}. \tag{61}$$

Taking the inverse Laplace and then the Fourier transform yields, for the long-time, large-distance limit,

$$F(l,t) = \frac{|l|}{\sqrt{4\pi D}}t^{-3/2}\exp\left(-\frac{l^2}{4Dt}\right), \tag{62}$$

where $D = \overline{l^2}/2\langle t\rangle$. The function $F(l,t)$ is a one-dimensional probability density and must be integrated over time to calculate a probability. This probability may superficially look similar to a Gaussian probability (except for the $|l|$ term), but for fixed time t, a Gaussian has a maximum at $l = 0$, while $F(l,t)$ has a maximum when $l^2 = 2Dt$. In fact, $F(l,t)$ is zero at $l = 0$. An important fact to notice is that the mean first-passage time, on an infinite lattice, to any lattice site is infinite. This is because the walker can drift out to infinity in a direction away from the chosen site to which the first passage is calculated. However, the average time $\langle T(R)\rangle$ to leave a region of length scale R (e.g., a circle of radius R) is finite because we then sum over all l on the boundary of the region. Since a Gaussian probability density spreads as $\langle R^2(T)\rangle \approx T$, one might suspect that $\langle T(R)\rangle \approx R^2$, and this is correct, as we show in Sec. 5.3.

An important quantity is $S(t)$, the number of distinct sites visited on a lattice in a time t. In the field of reactions it is important to know how much territory a diffusing reactant explores in a time t. This is related to $S(t)$. We begin with the quantity S_n, the number of distinct sites visited in n steps (and we count the origin as the first new site). We can write

$$S_n = 1 + \sum_l [F_1(l) + \cdots + F_n(l)]. \qquad (63)$$

In continuous time the number of distinct sites visited in a time t is

$$S(t) = \sum_{n=0}^{\infty} S_n \int_0^t \psi_n(t - \tau)\Phi(\tau)d\tau, \qquad (64)$$

where the integral accounts for the nth jump occurring at time $t - \tau$, and then no jump occurring from time $t - \tau$ to t. Using generating-function techniques one finds for the Laplace transform of $S(t)$ the form $\psi^*(s)/sG(l = 0;\psi^*(s))$. Using the explicit form of the Green's function, we find (assuming $\langle t\rangle$ is finite) that

$$S(t) \approx \begin{cases} t/\langle t\rangle G(0;1) & \text{for } D = 3, \\ \pi t/\langle t\rangle \log(t/\langle t\rangle) & \text{for } D = 2, \\ \sqrt{8t/\pi\langle t\rangle} & \text{for } D = 1. \end{cases} \qquad (65)$$

Consider the number of distinct points visited in a set other than the entire lattice. The generating-function formalism still holds, with a sum over points in the set instead of over every lattice point. For the set being a plane in a three-dimensional cubic lattice $\langle S(t)\rangle \approx t^{1/2}$, so that after a time t the number of points in the plane visited by the walker (which can go anywhere in three dimensions) scales as $t^{1/2}$. For the special set being a line in three dimensions, $\langle S(t)\rangle \approx \log(t)$. The calculation of the probability distribution of $S(t)$, as opposed to its first moment, requires techniques beyond the generating-function approach.

5.3 Many Walkers

There are several ways to obtain noninteger exponents in random walks. We will later address the concepts of fractal space and time where fractional exponents reflect fractal properties. Here, we treat the case of more than one random walker. We consider N random walkers inside a sphere (see Fig. 4) of radius R, and ask the question for the mean first-passage time for the sphere to be empty of random walkers. This problem was posed in the context of the disentanglement time for N entangled polymer chains (Weiss *et al.*, 1988). If the walkers undergo Brownian motion we have, for the probability $p(\mathbf{r},t|\mathbf{r}_0,t = 0)$ for the walker to go between $\mathbf{r}_0$ and $\mathbf{r}$ in a time t,

$$p(\mathbf{r},t|\mathbf{r}_0,t = 0) = 4\pi\left(\frac{3}{4\pi Dt}\right)^{3/2} \exp\left[-\frac{3(\mathbf{r} - \mathbf{r}_0)^2}{4Dt}\right], \qquad (66)$$

where D is the diffusion constant. Let $E(t)$ be the probability that a particle initially in the sphere is in the sphere at time t. We can express $E(t)$ as the integral of Eq. (66) over $\mathbf{r}$ and $\mathbf{r}_0$ for all points in the sphere. For N

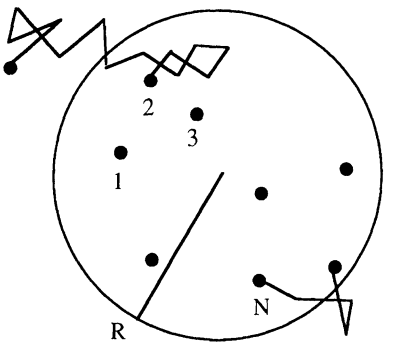

FIG. 4. *N* random walkers, denoted by ●, and labeled 1,2,3, . . ., *N*, start in random initial conditions inside a sphere of radius *R*. Each walker undergoes Brownian motion. A possible trajectory for walker 2 is shown. We are interested in the mean first-passage time for the sphere to be empty of walkers. Note that walker *N* left the circle, but reentered. This possibility complicates the calculation for the circle being empty of walkers.

noninteracting walkers the probability that at least one of them is in the sphere is given by $\Lambda(t) = \{1 - [1 - E(t)]^N\}$. The rate of change of $\Lambda(t)$ is the probability for the sphere to be empty. The first moment $\langle T_N \rangle$ for the sphere to be empty of N walkers (but not necessarily for the first time, although this is a good approximation in 3D for the first-time probability) is

$$\langle T_n \rangle = \int_0^\infty t \frac{d\Lambda(t)}{dt} dt = \int_0^\infty \Lambda(t) dt. \tag{67}$$

Now for t sufficiently large that $Dt \gg R^2$, then

$$E(t) \approx \sqrt{2\pi}\,(R^2/Dt)^{3/2}. \tag{68}$$

When N is large,

$$\begin{aligned} \langle T_N \rangle &\approx \int_0^\infty (1 - e^{-NE(t)}) dt \\ &\approx \int_0^\infty [1 - \exp[-N(T/t)^{3/2}]] dt \\ &\approx TN^{2/3} \int_0^\infty [1 - \exp(-x^{3/2})] dx, \end{aligned} \tag{69}$$

where $T = cR^2/D$, with c a constant. If N, the number of walkers, is a function of R, say $N(R) = R$, then a fractional exponent results, i.e., $\langle T_N \rangle \approx R^{8/3}$ as opposed to the single-particle result of $\langle T_1 \rangle \approx R^2$.

Another example of a fractional exponent entering a many-random-walker problem deals with interacting walkers that annihilate upon contact. In one dimension, the probability that p walkers starting near each other survive for n steps scales as $n^{-p(p-1)/4}$ (Fisher, 1984). Fractional exponents also enter naturally in trapping and reaction problems, as discussed in Sec. 6.

6. REACTION KINETICS

6.1 Trapping

Consider Brownian motion of randomly spaced particles in the presence of randomly spaced trapping regions. One might as a first-order guess consider space to be broken up into cells of length R, each with a particle and a number of traps. The mean trapping time would be proportional to R^2/D, where R is the average distance to a trap and D is the particle's diffusion constant. The probability that a particle is not trapped would then be $\Phi(t) = \exp(-2nDT/R^2)$ where n is the dimension. This expression is, in fact, a good approximation for all but asymptotic times. On a finite lattice the survival time is always exponential (den Hollander, 1985).

The problem with this approach is that it ignores spatial fluctuations in the trap positions. Let us consider a particle that starts in a region of radius R without any traps. We must now average the lifetime of a particle over the probability of being in such a region of radius R (Donsker and Varadhan, 1979). If we assume a Poisson distribution for a region of volume V to be void of traps, i.e., $P(V) = (1/V)\exp(-V/V_0)$, with volume $V \sim R^d$ then the lifetime distribution of a walker is given by

$$\begin{aligned} \Phi(t) &\approx \int_0^\infty \exp\left(-\left[\frac{R}{R_0}\right]^d\right) \exp\left(-\frac{2dDt}{R^2}\right) dR \\ &\approx \exp(-\text{const.}\, t^{d/(d+2)}). \end{aligned} \tag{70}$$

This behavior is dominated by that small fraction of walks that start very far away from any traps, and Eq. (70) is only valid at very long times.

6.2 First- and Second-Order Kinetics

We now consider various reaction–diffusion schemes. In all cases we consider diffusion to be the rate-limiting step. This means the reactions are diffusion limited and not rate limited. In bimolecular reactions, if one species is unlimited in number, the population of the other species is said to follow a pseudo-unimolecular reaction.

For a unimolecular reaction if we start with N_0 reactants, we can write an equation for the probability $P(N,t)$ of having N reactants left at time t, where $K(t)$ is the reaction rate:

$$\frac{dP(N,t)}{dt} = K(t)[(N+1)P(N+1,t) - NP(N,t)]. \tag{71}$$

The equation for the average number of reactants $\langle N(t) \rangle$ is

$$\frac{d\langle N(t) \rangle}{dt} = -K(t)\langle N(t) \rangle, \tag{72}$$

with the solution

$$\langle N(t)\rangle = N_0 \exp\left(-\int_0^t K(\tau)d\tau \right). \tag{73}$$

Usually $K(t)$ is treated as a constant.

For a bimolecular reaction with species A and B annihilating upon contact, one writes for the average populations $A(t)$ and $B(t)$ the following rate equation:

$$\frac{dA(t)}{dt} = -kA(t)B(t) = \frac{dB(t)}{dt}, \tag{74}$$

with the solution

$$\frac{1 + C/A(t)}{1 + C/A(0)} = \exp(Ckt), \tag{75}$$

where $C = B(0) - A(0)$. If $B(0) \gg A(0)$ then $C \approx B(0)$ and

$$A(t) \approx A(0) \exp[-kB(0)t]. \tag{76}$$

If, however, $A(0) = B(0)$, then expanding Eq. (75) for small C yields

$$A(t) = \frac{A(0)}{1 + A(0)kt} \approx \frac{1}{kt}. \tag{77}$$

This solution is in fact not valid for dimensions $d \leq 2$ because it assumes that the rate constant *is* a constant. Actually, k is proportional to $dS(t)/dt$, where $S(t)$ is the number of new sites visited. Since $S(t) \approx \sqrt{t}$ and $t/\ln(t)$ in 1D and 2D, k is time dependent. A more serious matter involves spatial fluctuations. In the $A + B \rightarrow 0$ case, with $A = B$, fluctuations in the density will create regions with $A > B$ and vice versa. The number of particles in a region of radius R is proportional to R^d, with fluctuations of order $R^{d/2}$. After a time $t = R^2/D$, most particles of the lower density in a region have reacted, leaving mostly the excess majority species. So an A-rich region will have about $N(A) \sim R^{d/2} \sim (Dt)^{d/4}$ particles in a volume $V \sim (Dt)^{d/2}$. The density thus decays at long times as $N/V \sim (Dt)^{-d/4}$ for $d \leq 4$, and not $1/t$ as in Eq. (77). We derive this result in the next section.

6.3 Reaction–Diffusion Equations

The above rate-equation approach for average populations and constant reaction rates is useful and holds well at early times, but it lacks spatial information about the reactants, especially the segregation into A- and B-rich regions. This segregation phenomenon will slow down the reaction rate relative to the prediction of the rate equation. It is more accurate and interesting to start with a reaction–diffusion equation, where $P_A(r,t)$ is the probability that a particle of type A is at position $\mathbf{r}$ at time t (and similarly for particles of type B), and D is the diffusion constant for type A and type B particles. Type A and type B particles annihilate upon meeting with reaction rate k:

$$\begin{aligned} \frac{\partial P_A(\mathbf{r},t)}{\partial t} &= D\nabla^2 P_A(\mathbf{r},t) - kP_A(\mathbf{r},t)P_B(\mathbf{r},t), \\ \frac{\partial P_B(\mathbf{r},t)}{\partial t} &= DP_B\nabla^2(\mathbf{r},t) - kP_A(\mathbf{r},t)P_B(\mathbf{r},t). \end{aligned} \tag{78}$$

Source and sink terms, plus boundary conditions, can also be added to the above equations. Further analysis proceeds by introducing two new variables (Ovchinnikov and Zeldovich, 1978),

$$\begin{aligned} s(\mathbf{r},t) &\equiv \tfrac{1}{2}[P_A(\mathbf{r},t) + P_B(\mathbf{r},t)], \\ \delta(\mathbf{r},t) &\equiv \tfrac{1}{2}[P_A(\mathbf{r},t) - P_B(\mathbf{r},t)]. \end{aligned} \tag{79}$$

This method immediately produces the linear diffusion equation for the difference variable,

$$\frac{d\delta(\mathbf{r},t)}{dt} = D\nabla^2\delta(\mathbf{r},t). \tag{80}$$

We will treat the case of equal numbers of A and B particles, which are randomly distributed in an uncorrelated fashion. The average of δ vanishes identically. We need to calculate the second moment,

$$\begin{aligned} \Gamma(\mathbf{r} - \mathbf{r}',t) &\equiv \langle \delta(\mathbf{r},t)\delta(\mathbf{r}',t)\rangle \\ &= \frac{\rho(t = 0)}{2(8\pi Dt)^{d/2}} \exp\left(-\frac{[\mathbf{r} - \mathbf{r}']^2}{8Dt} \right), \end{aligned} \tag{81}$$

which is the d-dimensional form of the Gaussian, and $\rho(t = 0)$ is the initial particle density. The form for Γ follows from the Gaussian form for δ when the A and B initial positions are random and uncorrelated. For $\mathbf{r} = \mathbf{r}'$ we obtain the scaling $\delta(\mathbf{r},t) \sim t^{-d/4}$ for $d < 4$ and t^{-1} for $d \geq 4$. The bimolecular reaction–diffusion equation can produce other results for other conditions. For example, Lindenberg *et al.* (1990) find for geminate recombination with A's and B's created in initially close pairs (say, e.g., in electron–hole creation) that $\delta(\mathbf{r},t) \sim t^{-(d+2)/4}$.

7. FRACTAL-TIME PROCESSES

7.1 Fractal Waiting Times

In the continuous-time random walk let us now consider the Laplace transform of the waiting-time density, as this quantity appears in calculating many properties of a random walk, such as in determining the number $S(t)$ of distinct sites visited in a time t:

$$\psi^*(s) \equiv \int_0^\infty \psi(t)\exp(-st)dt = \sum_{n=0}^{\infty} (-1)^n s^n \langle t^n \rangle. \tag{82}$$

This expansion only holds if all the moments of $\psi(t)$ are finite. Let us consider a case where this does not occur. Choose $\psi(t)$ as

$$\psi(t) = \frac{1-q}{q}\sum_{j=1}^{\infty} q^j\lambda^j \exp(-\lambda^j t) \quad (\lambda < q < 1). \tag{83}$$

This has a scaling form, with an order-of-magnitude longer waiting time (in base $1/\lambda$) occurring an order of magnitude less often (in base $1/q$). This is like intervals in the triadic Cantor set where longer lengths (in base 3) occur two times less often. Our waiting-time distribution is easily checked to be normalized to 1, but the condition that $\lambda < q < 1$ leads to $\langle t \rangle$ being infinite, i.e.,

$$\langle t \rangle = \frac{1-q}{q}\sum_{j=1}^{\infty}\left(\frac{q}{\lambda}\right)^j = \infty, \tag{84}$$

and so the straightforward use of the moment expansion in Eq. (82) is invalid. How then can one proceed with a power-series expansion of $\psi^*(s)$ to determine its dominant behavior? We can write $\psi^*(s)$ as

$$\psi^*(s) = \frac{1-q}{q}\sum_{j=1}^{\infty}\frac{(q\lambda)^j}{[s+\lambda^j]}. \tag{85}$$

The scaling relation

$$\psi^*(s) = q\psi^*\left(\frac{s}{\lambda}\right) + \frac{\lambda(1-q)}{(s+\lambda)} \tag{86}$$

allows one to determine the noninteger exponent in the expansion of $\psi^*(s)$ in terms of powers of s. The inhomogeneous term in Eq. (86) involves only integer powers of s, and so any noninteger exponent must come from the homogeneous part of the equation $\psi^*(s) = q\psi^*(s/\lambda)$, which has a solution of the form

$$\psi^*(s) \approx s^\beta, \quad \text{with } \beta = \frac{\ln(q)}{\ln(\lambda)}. \tag{87}$$

This term is missing from the series expansion in Eq. (82) because that expansion assumed that all the moments of $\psi(t)$ were finite. Adding to Eq. (82) this missing term with its noninteger exponent yields (if β is less than 1)

$$\psi^*(s) \approx 1 - s^\beta + O(s), \quad \text{as } s \to 0,$$

and

$$\psi(t) \approx t^{-1-\beta}, \quad \text{as } t \to \infty. \tag{88}$$

If β is greater than unity, then for small s, the s term will dominate the s^β term in Eq. (82), and $\psi(t)$ will decay fast enough so that $\langle t \rangle$, the average time between transitions in the random process, is finite. This implies that Poisson statistics will asymptotically govern the number of transitions in the random process, i.e., the mean number of transitions will grow linearly with time as $t/\langle t \rangle$. This is similar to the central limit theorem where the important spatial quantity was $\langle r^2 \rangle$. If this was finite, Gaussian behavior ensued. The condition $\lambda < q < 1$, which makes $\langle t \rangle$ infinite, also keeps $\beta < 1$. When β equals 1, jumps occur at a well-defined constant rate and time flows forward (as measured by the number of jumps) as a one-dimensional process. The above $\psi(t)$, with β less than 1, is called a *fractal-time* random process because the mean number of jumps in a time t grows as t^β when $\beta < 1$, as opposed to linearly with t when $\beta > 1$. Time can be described as being β-dimensional for a fractal-time random process.

7.2 Stretched-Exponential Relaxation

A random process is said to be subordinated to a second random process if an event occurring in the first process depends on the probability of an event occurring in the second random process. We consider, specifically, the first random process to be a reaction that is subordinated to a second random process of random walkers reaching the reaction site. In this scenario, when the second process involves random walks moving in fractal time, the second process will be shown to be a stretched-exponential relaxation.

Consider the problem of dielectric relaxation of glasses involving a frozen-in dipole that can be relaxed only when it is hit by a mobile defect. This problem involves the first-passage time of random walkers (defects) to reach the origin (frozen-in dipole). The mathematical analysis of the problem is in the form of letting there be V lattice sites and letting N walkers be initially randomly distributed among these sites, not including the origin. The probability $\Phi(t)$ that none of the walkers has reached the origin by time t is given by

$$\Phi(t) = \left[1 - \frac{1}{V}\sum_{r}\int_0^t F(\mathbf{r},\tau)d\tau\right]^N, \tag{89}$$

where $F(\mathbf{r},\tau)$ is the probability density that a walker starting at site $\mathbf{r}$ will reach the origin for the first time at time τ. The integral allows for a first passage of a walker to the origin in the interval $(0,t)$. The $1/V$ enters as the probability of a walker starting at a site, and a sum over all possible starting points for a walker is performed. The quantity in brackets gives the probability that a particular walker has not reached the origin, and it is raised to the Nth power for the probability that none of the walkers has yet reached the origin. The problem is easier in the limit $N \to \infty$, $V \to \infty$, but with the ratio remaining constant, $N/V = c$. In this limit,

$$\Phi(t) = \exp\left[-c\sum_{\mathbf{r}}\int_0^t F(\mathbf{r},\tau)d\tau\right] = \exp[-cS(t)]. \tag{90}$$

The term in brackets was simplified by noting that any of the sites from which a walk can reach the origin in a time t are exactly the same sites a walker starting at the origin can reach in a time t. This is exactly $S(t)$, the number of distinct sites that a walker visits in a time t, as discussed in Sec. 5.3. Equation (90) has the following form for the random-walk jumps governed by finite and infinite $\langle t\rangle$ (see Fig. 5):

$$\Phi(t) \approx \begin{cases}\exp(-\text{const.}t/\langle t\rangle) & \text{for}\,\langle t\rangle\,\text{finite},\\ \exp(-\text{const.}t^{\beta}),\ \beta<1, & \text{for}\,\langle t\rangle\,\text{infinite}.\end{cases} \tag{91}$$

In addition to introducing subordination processes, we have found that the stretched exponential is a probability limit distribution. This means that the form of $\Phi(t)$ does not depend on the details of the underlying random process, but only on whether $\langle t\rangle$ for the time between jumps is finite or infinite.

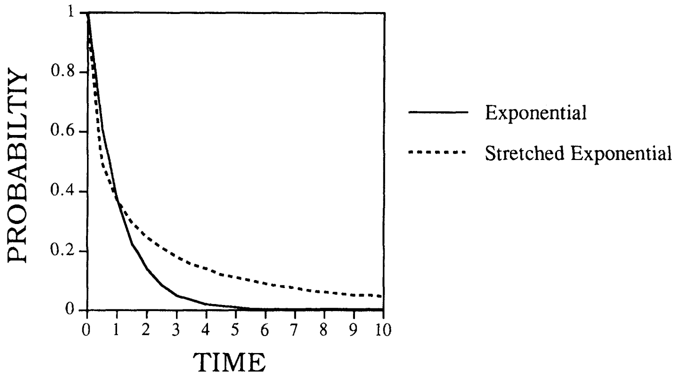

FIG. 5. The exponential decay (single time scale) compared to a stretched-exponential decay. The stretched-exponential decay is initially faster, but is slower after the $1/e$ decay of the exponential.

8. FRACTAL-SPACE PROCESSES

8.1 Lévy Flights

There are many ideas that have been introduced for generalizing Brownian motion, including fractional Brownian motion (Mandelbrot and van Ness, 1968). A fractional Brownian-motion time series would have a Gaussian distribution of values, but the variance would grow, in time t, as t^H with $0 < H < 1$. If $H > \frac{1}{2}$ the motion is called persistent and the variance grows at least as t^2, which is faster than Brownian motion, while the opposite is true for the antipersistent case of $H < \frac{1}{2}$.

A random-walk process that does not fall into the Brownian-motion class occurs when the variance of the jump lengths is infinite (Lévy, 1937). If the variance of each jump is infinite, then the variance of N jumps is also infinite. This in turn implies that the distribution of the sum of N steps should have similar properties to that of a single step. This is basically the question of fractals: When does the whole (the sum of N steps) look like its parts? Lévy found the answer that asymptotically for small k, the Fourier transform $\tilde{p}(k)$ of the probability $p(x)$ to make a jump of displacement x was of the form $\tilde{p}(k) = \exp(-|k|^\beta)$ with $\beta \leq 2$. The Gaussian is the case $\beta = 2$. The Cauchy is the case $\beta = 1$ (see Fig. 6). This is reminiscent of our analysis of fractal time where noninteger exponents entered and signaled self-similar behavior and infinite moments of a probability distribution. This time the moments are spatial, instead of temporal. Let us look at a particular pedagogical case, called the Weierstrass random walk, which illuminates the above discussion.

Begin with a random walk on a lattice with the following probabilities for jumps of length r:

$$p(r) = \frac{q-1}{2q} \sum_{j=0}^{\infty} q^{-j}[\delta_{r,b^j} + \delta_{r,-b^j}]. \tag{92}$$

In this scheme jumps an order of magnitude farther (in base q) occur an order of magnitude less often (in base b). About q jumps of size unity are made, and form a cluster of q sites, before a jump of length b occurs, and about q such clusters of visited sites are formed before a jump of length b^2 occurs, and so on until an infinite fractal hierarchy of clusters is formed. To see the mathematical condition necessary for this to occur, examine the Fourier transform of $p(r)$ to obtain

$$\tilde{p}(k) = \frac{q-1}{q} \sum_{j=0}^{\infty} q^{-j} \cos(kb^j). \tag{93}$$

This is the famous Weierstrass function, which is everywhere continuous, but nowhere differentiable, when $b > q$. Note that

$$\begin{aligned} \langle r^2 \rangle &= \sum_{r=-\infty}^{\infty} r^2 p(r) = \frac{q-1}{q} \sum_{j=0}^{\infty} \left(\frac{b^2}{q}\right)^j \\ &= -\frac{\partial^2}{\partial k^2} \tilde{p}(k=0) = \infty \end{aligned} \tag{94}$$

when $b^2 > q$. This divergence of the second moment (absence of a scale) is a sign of fractal self-similar properties. We have numerically graphed a two-dimensional version

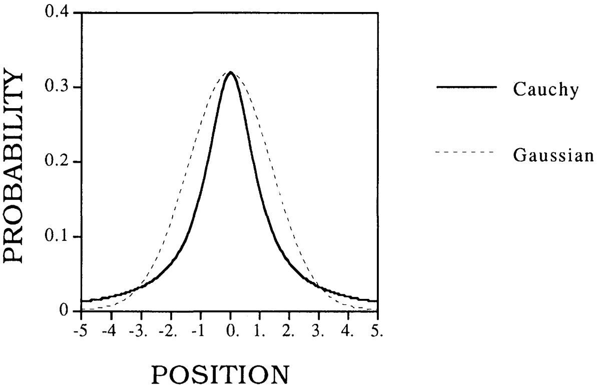

FIG. 6. A Gaussian distribution falls off rapidly enough that all of its positive moments are finite. The Cauchy distribution falls off slowly enough that its second and all higher moments are infinite. The probability in the tail is sufficiently strong to induce events far away from the typical Gaussian behavior. A random walker whose jumps are governed by a Cauchy distribution will visit in two or higher dimensions a fractal set of points with fractal dimension of 1. The traditional approach when encountering Cauchy distributions was to characterize them by their full width at half maximum, and to ignore the fractal consequences of a probability distribution with infinite moments.

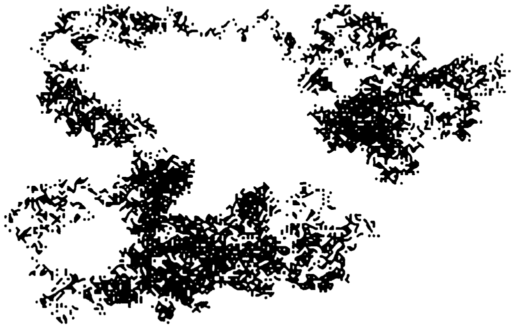

FIG. 7. A 10 000-step random walk with nearest-neighbor transitions. Eventually an infinite-step walk would fill every site in the plane and achieve a trivial self-similarity of looking the same on every scale.

of a Gaussian random walk and a Weierstrass random walk in Figs. 7 and 8 to emphasize the difference between a Gaussian and a Lévy process. We cannot write a power-series expansion of $\tilde{p}(k)$, since the coefficient of the $\langle r^2 \rangle$ term is infinite. However, as in the case of the fractal-time distribution, a scaling equation can be employed to determine the small-k behavior of $\tilde{p}(k)$,

$$\tilde{p}(k) = \frac{1}{q}\tilde{p}(bk) + \frac{q-1}{q}\cos(k). \tag{95}$$

The inhomogeneous term has no noninteger powers, and so we focus on a solution to the homogeneous part of the equation, $\tilde{p}(k) = (1/q)\tilde{p}(bk)$, which has a solution of the form k^β with $\beta = \ln(q)/\ln(b)$. When $\beta < 2$ the small-k (large distances r) behavior of $\tilde{p}(k)$ involves the exponent β, while for $\beta > 2$ the moment $\langle r^2 \rangle$ is finite and a Taylor-series expansion of $\tilde{p}(k)$ will exist, i.e.,

$$\tilde{p}(k) \approx \begin{cases} 1 - \text{const.}|k|^\beta + O(k^2) \approx \exp(-|k|^\beta) & \text{for } \beta < 2, \\ 1 - \frac{1}{2}\langle r^2 \rangle k^2 \approx \exp(-|k|^2) & \text{for } \beta > 2. \end{cases} \tag{96}$$

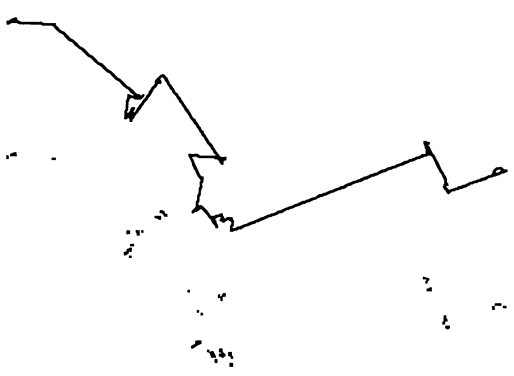

FIG. 8. A Weierstrass random walk of 1000 steps is shown with both the flight segments and the turning points (projected below) Note the self-similar nature of the cluster patterns. The more steps that are taken, the larger would be the probability that longer jumps would be made, thus starting new clusters at order-of-magnitude greater distances from the existing clusters. In this manner self-similar fractal clusters are generated.

One can calculate $P_n(r)$, following Eq. (4), in terms of the inverse Fourier transform of $[\tilde{p}(k)]^n$. For large n, $[\tilde{p}(k)]^n$ will be dominated by small k values, which are what we have calculated in Eq. (96). For the case $\beta = 1$, $P_n(r)$, in the large-n limit, has the Cauchy form $n/[\pi(n^2 + r^2)]$. Lévy's work treats a more general case allowing for a bias and additional parameters, but the above discussion captures the general spirit.

8.2 Lévy Walks

Despite the beauty of Lévy flights, they have been largely ignored in the physics literature because of their infinite moments. This infinity can be tamed by associating a velocity with each flight-trajectory segment (Shlesinger *et al.* 1987). Still in the continuous-time random-walk framework, we consider $\Psi(r,t)$ to be the probability density to make a jump of displacement r in a time t. We write

$$\Psi(r,t) = \psi(t|r)p(r)$$

or

$$\Psi(r,t) = p(r|t)\psi(t), \tag{97}$$

where $p(r)$ and $\psi(t)$ have the same meaning as before and $\psi(t|r)$ and $p(r|t)$ are conditional probabilities for a jump taking a time t given it is of distance r, and for a jump being of distance r, given it took a time t. A simple choice is

$$\psi(t|r) = \delta(t - r/V(r)). \tag{98}$$

In the study of turbulent diffusion Kolmogorov assumed a scaling law implying $V(r) \sim r^{1/3}$. If we furthermore choose

$$p(r) \approx |r|^{1+\beta}, \tag{99}$$

which for small enough values of β produces a Lévy flight with $\langle r^2 \rangle = \infty$, one finds for the mean square displacement

$$\langle r(t) \rangle^2 = \begin{cases} t^3 & \text{for } \beta \leq \frac{1}{3}, \\ t^{2+(3/2)(1-\beta)} & \text{for } \frac{1}{3} \leq \beta \leq \frac{5}{3}, \\ t & \text{for } \beta \geq \frac{1}{3}. \end{cases} \tag{100}$$

The t^3 case is called Richardson's law of turbulent diffusion. It corresponds to a Lévy walk with Kolmogorov scaling for $V(r)$ combined with a β such that the mean time spent in a segment of the trajectory is infinite. The reason we did not obtain the usual Lévy flight result of $\langle r^2(t) \rangle = \infty$ is that with the coupled space-time memory $\Psi(r,t)$ one does not calculate $\langle r^2 \rangle = -\partial^2 \tilde{p}(k = 0)/\partial k^2$, which would be infinite, but one uses $\int \exp(ikr)\psi(s|r)p(r)dr$ instead of $\tilde{p}(k)$. In effect, one is asking how far a walker has gone in a time t, instead of how long is the flight segment.

8.3 Random Walks on Fractals

Consider a random walk that is simple in the sense that it makes all of its jumps at a constant rate to a nearest neighbor. This can still lead to complex behavior if the set of sites is fractally distributed. One example is a random walk on a percolation cluster. The percolation cluster can be formed by connecting, with probability p, neighboring sites on a regular lattice. For low p values only finite-sized islands of connected sites are found. As p is increased there is a critical value when an infinite cluster of connected sites appears. Consider the two-dimensional case (see Fig. 9). If one draws a circle with radius R, then the number of sites on the connected cluster inside this circle grows as R^{d_f}, where d_f is the fractal dimension of the cluster. On a square lattice, with all sites accessible, $d_f = 2$, while on the critical percolation cluster $d_f \approx 1.89$. The cluster has many dead ends, which do not contribute to the mean square displacement, and so one may expect that $\langle R^2(t) \rangle$ grows more slowly in time than for Brownian motion in the plane.

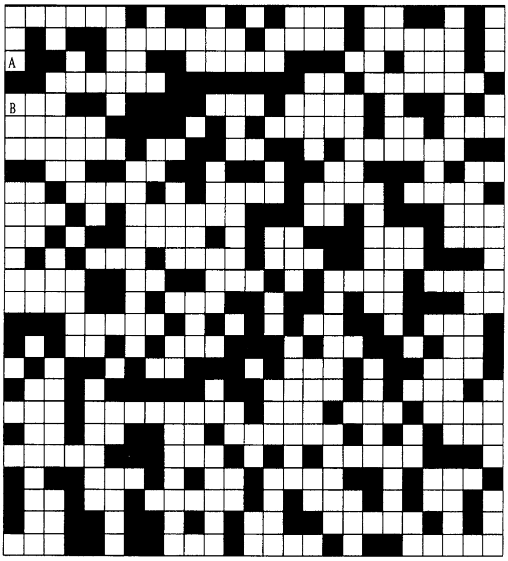

FIG. 9. The white squares represent a lattice that percolates (has continuous pathways for nearest-neighbor transitions) through the full square lattice. Many of the pathways, however, present an inefficient means for connecting two points, e.g., the nearby points A and B. A random walker on this tortuous percolation lattice only visits about $N^{2/3}$ new sites in an N-step random walk vs $N/\log(N)$ new sites on a perfect square lattice.

However, even if all the dead ends are excised and the random walk is restricted to the backbone of the percolation cluster, one still finds slower than linear growth of $\langle R^2(t) \rangle$.

For walks on fractal objects one defines the dimension of the walk, d_w, by

$$\langle R^2(t) \rangle \approx t^{2/d_w}. \tag{101}$$

Random walks on regular infinite lattices have $d_w = 2$. This is because by Polya's theorem a random walk will visit every site in two or fewer dimensions with probability 1, and so the walk trajectory dimension equals 2. Another dimension, called the spectral or harmonic dimension (Alexander and Orbach, 1982), associated with a random walker is defined via the equation for the number $S(t)$ of distinct sites visited, by

$$S(t) \approx t^{d_s/2}. \tag{102}$$

For a regular three-dimensional lattice, $d_s =$

$d_w = 2$. Using Eq. (101) for $\langle R^2(t)\rangle$ we see the scaling $R \approx t^{1/d_w}$. We can also write that after a time t the walker has explored a volume scaling as R^{d_f} that we can write as t^{d_f/d_w}. Equating the volume explored to the number of sites visited gives the following relationship between exponents:

$$d_s = 2d_f/d_w. \qquad (103)$$

In general, d_w is not necessarily equal to d_f. For a percolation lattice $d_s \approx 4/3$. One can further write d_w in terms of scaling exponents of the percolation lattice (Gefen *et al.*, 1982) For percolation the probability P_∞ that any given lattice site belongs to the infinite cluster is $(p - p_c)^\beta$, and the linear size ξ of the cluster scales as $(p - p_c)^{-\nu}$. We rewrite P_∞ as $\xi^{-\beta/\nu}$, so that the fractal dimension is $d_f = d - \beta/\nu$. By the Einstein relation, the dc conductivity σ and diffusion D are linked by $\sigma = e^2nD/kT$, where e is the charge on an electron, n is the number of electrons, k is the Boltzmann constant, and T is the temperature in kelvins. The conductivity scales as $\sigma \approx (p - p_c)^\mu \approx \xi^{-\mu/\nu}$. Using that $n \approx P_\infty$, we have $D \propto \sigma/n \propto \xi^{-\mu/\nu+\beta/\nu}$. For $p < p_c$ the largest cluster is of size ξ, so that at long times $\langle R^2(t)\rangle \approx \xi^2 \approx t^{2/d_w}$, and so $\xi \approx t^{1/d_w}$. We also write $\langle R^2(t)\rangle \approx Dt$, which implies $D \approx t^{2/d_w-1} \approx \xi^{2-d_w}$. Now using the expression for D in terms of μ, ν, and β, we have

$$d_w = 2 + (\mu - \beta)/\nu. \qquad (104)$$

The percolation lattice is a random fractal and only estimates of the various exponents governing random walks on this lattice have been achieved. For regular fractals, exact calculations of d_f, d_w, and d_s are possible. For example, on the Sierpinski gasket in d dimensions $d_w = \ln(d + 3)/\ln(2)$. For the 2D Sierpinski gasket (see Fig. 10), $2/d_w = 2 \times \ln(2)/\ln(5) = 0.862$, which is the exponent that describes the mean square displacement.

9. DISORDERED SYSTEMS

Consider a random walk of N steps on a straight line with only nearest-neighbor jumps allowed. The mean square displacement in this one-dimensional problem is the familiar $\langle r^2(N)\rangle \sim N$. Now take the straight line and deform it to follow a three-dimensional random-walk path. N units of the deformed line are separated from each other on the average only a distance of $N^{1/2}$, instead of the N units separating them on the straight line. The random walk of N steps on the deformed line has $\langle r^2(N)\rangle \sim N^{1/2}$ (Haus and Kehr, 1987). This is a simple example that provides results different from the standard random walk.

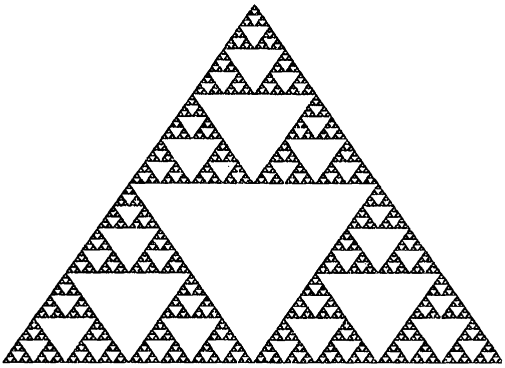

FIG. 10. A Sierpinski gasket. One performs a random walk on this structure by moving along the edges of the triangles. The jump length is the size of the side of the smallest triangle. The hierarchy of vacant triangular regions, in this fractal object, makes diffusion between widely separated sites more complicated and difficult than for a random walk on a regular square lattice. The mean square displacement of a random walker on the Sierpinski lattice grows with time as t^β with $\beta = 2 \ln(2)/\ln(5) = 0.862$. On a regular lattice $\beta = 1$.

Sinai (1982) investigated the properties of a random walk on a one-dimensional lattice with fixed probabilities to move to the right and left at each lattice site. Although these probabilities are fixed, their values are chosen from a probability distribution such that there is a strong bias at each site for the walker to go in a preferred direction. Sinai found that the walker spends many jumps going between two sites: a leftmost site that when approached from the right sends the walker back toward the right with high probability, and similarly a rightmost site that sends the walker back to the left. Many such pairs of sites can exist, especially in a nested configuration. Sinai finds $\langle r^2(t)\rangle \sim [\ln(t)]^4$ for this random-walk process. A more complicated example involves a random walker in one dimension with a random distribution of symmetric transition rates to the right and left. [See the traps in Fig. 11(b)]. For this problem, one can find (Alexander *et al.*, 1982) the anomalous statistics $\langle r^2(t)\rangle \sim t^{2(1-\beta)/(2-\beta)}$ for a distribution $g(w) \approx w^{-\beta}$ of transition rates, between neighboring sites, for small w when $\beta \leq 1$. The difficulty in generalizing the above calculations beyond one dimension has led to other approaches for random walks on disordered lattices.

Consider a random walker faced with a distribution of transition rates $g(W)$. Effective-medium theory postulates an *unknown* time-dependent rate, $W(t)$, which represents the average behavior of all the possible transitions in the random system. One determines $W(t)$ through a self-consistency argument. Change one transition between two neighboring sites in the system from $W(t)$ to a random value chosen from the distribution $g(W)$. Next calculate the Green's function $G = \langle G\rangle + \delta G$, where $\langle G\rangle$ is the random-lattice–walk Green's function for the perfect lattice where all transitions are governed by $W(t)$, and δG involves transitions T that encounter the altered bond. The Green's function can also be written symbolically as $G = \langle G\rangle + \langle G\rangle T\langle G\rangle + \langle G\rangle T\langle G\rangle T\langle G\rangle + \cdots = \langle G\rangle + \langle G\rangle TG$ (Webman and Klafter, 1982), where the second term represents movement in the effective lattice governed by $W(t)$, then an interaction T with the defective bond, followed by movement under $\langle G\rangle$ again. Higher-order terms in T have more interactions with the defective bond. Taking the average of both sides, over the distribution $g(W)$, implies that $\langle TG\rangle = 0$. The assumption for the effective-medium theory is that $\langle TG\rangle = \langle G\rangle\langle T\rangle = 0$, i.e., $\langle T\rangle = 0$. This condition determines $W(t)$, usually through a numerical evaluation.

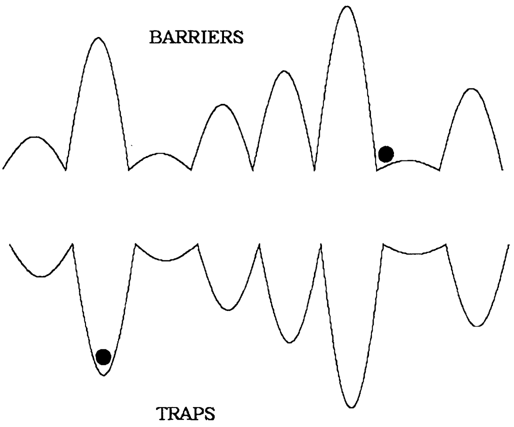

FIG. 11. The top figure represents random barriers, and the bottom one random traps. The dot, in each case, represents a random walker. In the barrier case (top), the walker has unequal probabilities at each site to go to the left or right. In this figure it is easier, from its present position, for the walker to go to the right. The walker tends to get trapped between high barriers. The walker's mean square displacement can grow logarithmically for this Sinai-type random walk. In the random-trap case (bottom), the walker always has equal probabilities to move to the right or left. The walker encounters a distribution of transition rates related to the depth of each trap. For a wide enough distribution of rates a sublinear growth of the mean square displacement with time occurs.

In the effective-medium approximation, one can let the rate to some sites be zero so that these sites are inaccessible. In this case a percolation critical probability p_c enters. If the probability is p that a site is inaccessible, and if $p < p_c$, then a walker is isolated on a finite island of connected sites, and so the diffusion constant would be zero. For a percolation model (Lax and Odagaki, 1982) the effective-medium calculation finds that the real and imaginary parts of the diffusion constant $D(\omega)$ vanish linearly and quadratically with frequency when $p < p_c$ and that $D(0)$ is proportional to $(p - p_c)/(1 - p_c)$ when $p > p_c$. Effective-medium theory was introduced in the 1950s to study scattering of waves in a random medium (Lax, 1952).

The continuous-time random walk discussed in Sec. 6.2 is an effective-medium theory. It was originally applied to electron and hole transport in disordered media (Scher and Lax, 1972). They derived the formula for the frequency-dependent diffusion constant $D(\omega)$ for a random-walk process,

$$\begin{aligned} D(\omega) &= -\frac{\omega^2}{6}\int_0^\infty \exp(-i\omega t)\langle[r(t) - r(0)]^2\rangle dt \\ &= \frac{\overline{r^2}}{6}\,\frac{i\omega\psi^*(i\omega)}{1 - \psi^*(i\omega)}. \end{aligned} \tag{105}$$

The function $\psi(t)$ represents the probability density for a walker (the electron or hole) to remain trapped between movements. The quantity $\psi(t)$ represents an effective medium and it is the same at each lattice site. An infinite first moment $\langle t\rangle$ of $\psi(t)$ was introduced (Montroll and Scher, 1973) to explain scaling behavior in a hole current set up by biased "fractal-time [$\psi(t) \approx t^{-1-\beta}$ for $\beta < 1$]" random walks. For fractal-time random walkers biased in one direction one finds that the first moment and its variance both grow as $\langle r(t)\rangle \approx \sqrt{\langle r^2(t)\rangle} \approx t^{\beta}$, and so the fluctuations (second moment) per distance traveled (first moment) are time independent. An approach to choosing $\psi(t)$ is to focus on the physics of the trapping process. If the hopping time t out of a potential well follows the Arrhenius law, $t^{-1} = \nu \exp(-E/kT)$, and the probability $\rho(E)$ of finding an energy barrier E falls off exponentially as $\exp(-E/E_0)$, then the randomness in energy induces a randomness of hopping times $\psi(t)dt = \rho(E)dE$, which produces the fractal time (see Sec. 8) $\psi(t) \sim t^{-1-\beta}$, where $\beta = kT/E_0$ when $kT < E_0$ and $\beta = 1$ if $kT > E_0$.

Random processes on random geometries is still an active area of research.

GLOSSARY

Brownian Motion: A random walk whose probability distribution is governed by the diffusion equation. The trajectory is fractal with the derivative (velocity) not well defined.

Central Limit Theorem: A theorem showing that a sum of identically distributed random variables with finite second moments converges to a Gaussian distribution as the number of variables becomes infinite.

Chapman–Kolmogorov–Smoluchowski–Bachelier Equation: An equation relating the probability for being at a site at a particular time, in terms of being at a different site at an earlier time and summing over all possible intermediate sites.

Characteristic Function: The Fourier transform of a probability function.

Colored Noise: Noise whose second-order correlation function is broader than a delta function.

First–Passage-Time Density: A probability density to reach a certain criterion (e.g., level, site) for the first time.

Fokker–Planck Equation: A second-order differential equation for a probability distribution based on an expansion relying on the existence of the first two moments of the distribution.

Fractal Brownian Motion: A fractal random process whose next step is built upon correlating its entire previous history. The uncorrelated case is Brownian motion.

Fractal Time: A random process where the mean interval between events is infinite. The set of event times looks like a random fractal set of dimension less than 1.

Gaussian Random Noise: Delta-correlated noise in time with an amplitude following Gaussian statistics.

Kubo Oscillator: An oscillator with a random frequency.

Langevin Equation: An equation of motion with an added noise term.

Lévy Process: A random walk in space with infinite second moment of the jump-size probability, or a random process with the first moment of the waiting-time distribution between events being infinite. A Lévy flight is a Lévy random walk jumping between points and a Lévy walk includes the whole trajectory as well as the end points of each jump.

Markov Process: A random process where the next change depends only on the present state of the random variable, and not on a more detailed prior history.

Master Equation: An equation relating the change with time of a probability for the value of a variable in terms of probabilities for the basic allowed transitions.

One-Over-*f* (1/*f*) noise: A power spectrum whose low-frequency behavior is described by a function that scales with frequency as $1/f$.

Power Spectrum: The Fourier transform of a second-order correlation function. The real part of the transform of a fluctuating current would correspond to the power as a function of frequency.

Random Walk: A sum of random variables. The walk can have correlated or uncorrelated steps.

Self-Avoiding Random Walk: A random walk whose steps may not overlap.

Shot Noise: The noise that arises from the arrival of randomly timed (usually Poisson arrival) distinct pulses each of which has an effect that decays away. Slow scale-invariant decays produce fractal shot noise.

Stochastic Differential Equation: A differential equation for the time evolution of random variables.

Stochastic Process: A random process that involves a time variable.

Subordinated Random Process: A process whose random event times depend on the outcome of a second random process.

Weierstrass Random Walk: A random walk whose generating function is the scaling Weierstrass function, and whose trajectory is a fractal set of hierarchical clusters.

White Noise: A noise whose second-order correlation function is a delta function, i.e., noise that is not correlated in space or time.

Wiener–Khintchine Theorem: A theorem that relates the second-order correlation function of a random process and its power spectrum as Fourier cosine transform pairs.

Works Cited

Alexander, S, Bernasconi, J., Schneider, W. R., Orbach, R. (1981), *Rev. Mod. Phys.* **53,** 175–198.

Alexander, S., Orbach, R. (1982), *J. Phys. Lett. (Paris)* **43,** L625–L631.

Chandrasekhar, S. (1943), *Rev. Mod. Phys.* **15,** 1–89.

David, F. N. (1962), *Games, Gods, and Gambling,* New York: Hafner.

den Hollander, W. Th. F. (1985), *J. Stat. Phys.* **40,** 201–204.

Domb, C., Fisher, M. E. (1958), *Proc. Cambridge Philos. Soc.* **54,** 48–59.

Donsker, M., Varadhan, S. (1979), *Commun. Pure Appl. Math.* **32,** 721–748.

Einstein, A. (1905), *Ann. Physik (Leipzig)* **17,** 549–560.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications,* 2nd ed., Vol. 2, New York: Wiley.

Fisher, R. A., Tippett, L. H. C. (1928), *Proc. Cambridge Philos. Soc.* **24,** 180–190.

Fisher, M. E. (1984), *J. Stat. Phys.* **34,** 667–729.

Gefen, Y., Aharony, A., Alexander, S. (1982), *Phys. Rev. Lett.* **50,** 77–80.

Haus, J. W., Kehr, K. (1987), *Phys. Rep.* **150,** 263–406.

Kramers, H. A. (1940), *Physica* **7,** 284–304.

Kubo, R. (1962), in: D. Ter Haar (Ed.), *Fluctuation, Relaxation, and Resonance in Magnetic Systems,* Edinburgh, U.K.: Oliver and Boyd.

Lax, M. (1952), *Phys. Rev.* **85,** 621–629.

Lax, M., Odagaki, T. (1982), in: M. F. Shlesinger, B. J. West (Eds.), *Random Walks and Their Applications in the Physical and Biological Sciences,* AIP Conference Proceedings No. 109, New York: American Institute of Physics.

Lévy, P. (1937), *Theorie de l'Addition des Variables Aleatoires,* Paris: Gauthiers-Villars.

Lindenberg, K., Seshadri, V., Shuler, K. E., Weiss, G. H. (1980), *J. Stat. Phys.* **23,** 11–25.

Lindenberg, K., West, B. J., Kopelman, R. (1990), in: F. Moss, L. A. Lugiato, W. Schleich (Eds.), *Noise and Chaos in Nonlinear Dynamical Systems,* Cambridge, U.K.: Cambridge Univ. Press.

Mandelbrot, B. B., van Ness, J. W. (1968), *SIAM Rev.* **10,** 422–437.

Montroll, E. W. (1950), *J. Chem. Phys.* **18,** 734–743.

Montroll, E. W., Scher, H. (1973), *J. Stat. Phys.* **9,** 101–135.

Montroll, E. W., Weiss, G. H. (1965), *J. Math. Phys.* **6,** 167–181.

Moss, F., Lugiato, L. A., Schleich, W. (Eds.) (1990), *Noise and Chaos in Nonlinear Dynamical Systems,* Cambridge, U.K.: Cambridge Univ. Press.

Oppenheim, I., Shuler, K. E., Weiss, G. H. (1977), *The Master Equation,* Cambridge, MA: MIT Press.

Ovchinnikov, A. A., Zeldovich, Ya. B. (1978), *Chem. Phys.* **28,** 215–218.

Pontryagin, L., Andronov, A. A., Vitt, A. (1933), *Zh. Eksp. Teor. Fiz.* **3,** 165–180.

Scher, H., Lax, M. (1972), *Phys. Rev. B* **7,** 4491–4501.

Schottky, W. (1922), *Ann. Phys. (Leipzig)* **68,** 157–176.

Shlesinger, M. F., West, B. J., Klafter, J. (1987), *Phys. Rev. Lett.* **58,** 1100–1103.

Sinai, Ya. G. (1982), *Theory Prob. Appl.* **27,** 256–268.

Székely, Gábor J. (1986), *Paradoxes in Probability Theory and Mathematical Statistics,* Dordrecht: Reidel.

Todhunter, I. (1865), *History of the Theory of Probability,* Cambridge, MA: Chelsea Pub.

Wax, N. (1954), *Selected Papers on Noise and Stochastic Processes,* New York: Dover.

Webman, I., Klafter, J. (1982), *Phys. Rev. B* **26,** 5950–5952.

Weiss, G. H., Bendler, J. T., Shlesinger, M. F. (1988), *Macromolecules* **21,** 521–523.

Weissman, M. (1988), *Rev. Mod. Phys.* **60,** 537–571.

Further Reading

Barber, M. N., Ninham, B. W. (1970), *Random and Restricted Walks: Theory and Applications,* New York: Gordon and Breach.

Blumen, A., Klafter, J., Zumofen, G. (1986), in: I. Zschokke (Ed.), *Optical Spectroscopy of Glasses,* Dordrecht: Reidel.

Bouchard, J.-P., Georges, A. (1990), *Phys. Rep.* **195,** 127–293.

Feder, J. (1988), *Fractals,* New York: Plenum.

Feller, W. (1973), *An Introduction to Probability Theory and Its Applications,* 3rd ed., Vol. 1, New York: Wiley.

Fox, R. F. (1978), *Phys. Rep.* **48,** 179–283.

Gardner, C. W. (1985), *Handbook of Statistical*

Methods for Physics, Chemistry and the Natural Sciences, Berlin: Springer-Verlag.

Gumbel, E. J. (1958), *Statistics of Extremes,* New York: Columbia Univ. Press.

Havlin, S., Ben Avraham, D. (1987), *Adv. Phys.* **36,** 695–798.

Havlin, S., Nossal, R., Shlesinger, M. (Eds.) (1991), special issue on Non-Classical Reactions, *J. Stat. Phys.* **65,** 837–1306.

Hughes, B. D. (1995), *Random Walks and Random Environments,* Oxford: Clarendon Press.

Kac, M. (1959), *Probability and Related Topics in Physical Sciences,* London: Interscience.

Lowen, S. B., Teich, M. C. (1989), *Phys. Rev. Lett.* **63,** 1755–1759.

Mandelbrot, B. B. (1982), *The Fractal Geometry of Nature,* San Francisco: Freeman.

Montroll, E. W., Shlesinger, M. F. (1984), in: E. W. Montroll, J. L. Lebowitz (Eds.), *The Wonderful World of Random Walks,* Studies in Statistical Mechanics Vol. VII, Amsterdam: North-Holland.

Montroll, E. W., West, B. J. (1979), in: J. L. Lebowitz, E. W. Montroll (Eds.), *On an Enriched Collection of Stochastic Processes,* Studies in Statistical Mechanics Vol. XI, Amsterdam: North-Holland.

Moss, F., McClintock, P.V.E. (1989), *Noise in Nonlinear Dynamical Systems,* Vols. 1–3, New York: Cambridge Univ. Press.

Moss, F., Bulsara, A., Shlesinger, M. (Eds.) (1993), special issue on Stochastic Resonance, *J. Stat. Phys.* **70,** 1–514.

Risken, H. (1984), *The Fokker-Planck Equation,* Berlin: Springer-Verlag.

Shlesinger, M. F., West, B. J. (1984), New York, AIP Conf. Proc. No. 109.

Shlesinger, M., Mandelbrot, B. B., Rubin, R. J. (Eds.) (1984), special issue on Fractals, *J. Stat. Phys.* **36,** 519–924.

Stratonovich, R. L. (1963), *Introduction to the Theory of Random Noise,* Vol. 1, New York: Gordon and Breach.

Stratonovich, R. L. (1967), *Introduction to the Theory of Random Noise,* Vol. 2, New York: Gordon and Breach.

Toussaint, D., Wilczek, F. (1983), *J. Chem. Phys.* **78,** 2642–2647.

van Kampen, N. G. (1981), *Stochastic Processes in Physics and Chemistry,* Amsterdam: North-Holland.

Weiss, G. (Chmn.) (1983), special issue on Random Walks, *J. Stat. Phys.* **30,** 249–561.

Weiss, G. H. (1994), *Aspects and Applications of the Random Walk,* Amsterdam: North Holland.

Weiss, G. H., Rubin, R. J. (1983), *Adv. Chem. Phys.* **52,** 363–505.

RARE GASES

R. A. AZIZ AND A. R. JANZEN, *Department of Physics, University of Waterloo, Waterloo, Ontario, Canada*

R. O. SIMMONS, *Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois, U.S.A.*

Introduction 72
1. **Gaseous State** 73
1.1 Interatomic Interactions 73
1.1.1 General Considerations 73
1.1.2 Like Pairs 73
1.1.3 Unlike Pairs 74
1.2 Compounds 74
1.2.1 Chemical Compounds 74
1.2.2 Clathrates 75
1.2.3 Van der Waals Complexes 76
1.3 Adsorption 76
1.4 Thermodynamics 76
1.4.1 Equations of State 76
1.4.2 Heat Capacity and Vapor Pressure 77
1.4.3 The Universal Gas Constant 77
1.5 Transport Properties 77
1.5.1 Kinetic Theory 77
1.5.2 Measurement of Viscosity and Thermal Conductivity 78
1.5.3 Measurement of Ordinary and Thermal Diffusion 78
1.6 Law of Corresponding States (LCS) 78
1.7 Optical and Dielectric Properties 79
1.7.1 Refractivity and Dielectric Virial Coefficients 79
1.7.2 Spectroscopy of the Dimers 80
2. **Liquid State and Dense Gases** 80
2.1 Interatomic Interactions in the Condensed State 80
2.2 The Structure of Liquids 81
2.3 Thermodynamic Properties 81
2.3.1 Quantum Corresponding States 82
2.3.2 Sound Velocity 82
2.4 Transport Properties 83
2.4.1 Transport Properties, Theory 83
2.4.2 Neutral and Charge Transport, Experiment 83
2.5 Microscopic Dynamics from Neutron-Scattering Evidence 83
2.6 Special Properties of Liquid Heliums and Their Isotopic Mixtures 84
3. **Crystalline State** 84
3.1 Melting, Corresponding States 84
3.2 Thermoelastic Properties 85
3.3 Microscopic Properties 86
3.3.1 Phonons, Experiment and Theory 86
3.3.2 Momentum Distribution, Kinetic Energy 87
3.4 Dielectric and Electronic Properties 87
3.5 Crystal Defects 88
3.5.1 Point Defects 88
3.5.2 Line and Surface Defects, Thin Films, Clusters 89
3.6 Solid Heliums, Unique Properties 90
3.7 Defects and Excitations in Solid Heliums 90
4. **Scientific and Technological Applications** 91
4.1 Arc Welding and Cutting 91
4.2 Passive Atmospheres and Inert Condensed Environments 91
4.3 Illumination 91
4.4 Lasers 92
4.5 Thermometry 92
4.6 Cryogenics and Refrigeration 92
4.7 Other Uses 92
Glossary 93
Works Cited 94
Further Reading 96

digital Encyclopedia of Applied Physics

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

In 1904, the Nobel prize in Chemistry was awarded to Sir William Ramsay and the prize in Physics was awarded to Lord Rayleigh, in part for their work in the discovery of rare gases as a new family of elements, a whole new column in the periodic table. This family was unknown when they began precise work on the density of nitrogen in 1892, although there was earlier evidence for two of its members: Henry Cavendish isolated a small bubble of what is now known to be argon while studying the composition of air in 1785, and P. J. C. Janssen and J. N. Lockyer had observed a brilliant new yellow line in the solar spectrum in 1868, a line that later turned out to be that of helium.

The presence of argon occupying 1% of the Earth's atmosphere produced the discrepancy in the densities of "nitrogen" obtained from air and from chemical reactions. This provided the critical clue leading to the discovery of argon, the first of the family to be isolated. Forming successively smaller proportions of the Earth's atmosphere and crust are helium, neon, krypton, xenon, and radon. It has been estimated that helium forms one-quarter of the mass of the universe, although on Earth, gravity is too weak to retain helium from escape from the atmosphere, to which it and, for example, radon are released from the Earth's crust by natural processes.

The rare gases have important places both in the development of many branches of science and in numerous technological applications. Neon provided the first clear examples of stable nuclear isotopes when J. J. Thomson, studying liquid-air residues, discovered ^{22}Ne and ^{20}Ne in 1913. The nucleus of ^{4}He (the alpha particle of radioactivity) is the most stable known, whereas all radon isotopes are radioactive. In metrology, from 1960 to 1983 the SI standard of length was defined by an orange line in the ^{86}Kr spectrum; one meter is exactly 1 650 763.73 wavelengths of this line. In thermal science, the properties of helium and neon are used in establishing the temperature scale. General properties of rare gases are given in Table 1. Selected technical applications appear in Sec. 4.

Lennard-Jones (LJ) models for the denser phases of rare-gas atoms have considerable historical significance. Such a model represents the interatomic potential as a polynomial of only two terms in the interatomic separation, one each for the repulsive and the attractive parts, respectively. The computational simplicity of the LJ interaction provided the first qualitative confirmations of the usefulness of molecular dynamics and of Monte Carlo computational methods in representing the properties of fluids, typically the properties of fluid argon. Further, a multitude of scientific articles have been written about the static and dynamical properties of fcc solid models using an LJ 12-6 potential and the mass of a rare-gas atom, again typi-

Table 1. General properties of the rare gases. (The subscripts tp, bp, and c refer to triple-, boiling-, and critical-point coordinates, respectively. For the heliums, which have complicated phase diagrams, only the bcc–fcc–liquid triple point is given.)

	^{3}He	^{4}He	Ne	Ar	Kr	Xe
Molecular mass	3.016 05	4.002 60	20.183	39.948	83.804	131.30
Atomic number	2	2	10	18	36	54
T_{tp} (K)	17.8	14.9	24.56	83.806	115.78	161.40
P_{tp} (kPa)	13 700	3050	43	69	73	82
V_{tp} (cm^3 mol^{-1})	18.83	20.70	16.18	28.24	34.32	44.31
T_{bp} (K)	3.19	4.215	27.07	87.27	119.80	165.05
T_c (K)	3.32	5.20	44.45	150.73	209.29	289.74
P_c (Mpa)	0.1167	0.2274	2.72	4.865	5.4931	5.8400
V_c (cm^3 mol^{-1})	72.8	57.51	41.7	74.7	92.3	118.3
Reduced de Boer parameters[a]						
(A)	2.88	2.50	0.543	0.172	0.094	0.058
(B)	3.08	2.68	0.594	0.186	0.102	0.063

---

[a] Reduced de Broglie wavelength. See Sec. 2.3.1 for definition. (A) Based on HFD true pair-potential parameters ϵ/k_B and σ (see Sec. 1). (B) Based on Lennard-Jones "effective" pair-potential parameters ϵ/k_B and σ.

cally argon. Among these theoretical investigations of lattice dynamics and of the statistical mechanics of the condensed state are the introduction of many interesting concepts and ingenious computational methods. Because the models can be well defined, with common dependence upon a stated pair interaction with only two parameters, comparisons between models are thus facilitated. For most properties of actual rare-gas solids, however, such models frequently have only qualitative validity. As precision in the knowledge of the pair potential between rare-gas atoms has increased and computational power is enhanced, dependence upon such a poor approximation to the pair potential is no longer necessary, although it continues to have limited usefulness in some qualitative considerations.

1. GASEOUS STATE

1.1 Interatomic Interactions

1.1.1 General Considerations A rare-gas atom consists of a positively charged nucleus surrounded by a negatively charged spherically symmetric electron cloud and hence possesses no permanent dipole. When the atoms are separated by a finite distance, they interact, and the energy resulting from the interaction is the interatomic potential energy $U(r)$, which depends only on the separation between them. Qualitatively, for small separations, the electron clouds overlap. A repulsive force then arises between the atoms, as a consequence of the Pauli exclusion principle, from the increase of energy of the electrons. For large separations, these nonpolar atoms attract one another by means of the London dispersion energy. The electrons of each atom are moving, and at any instant, one atom may possess an instantaneous dipole, which induces an instantaneous dipole in the second atom. The dipoles interact to produce an attractive "dispersion" energy.

An *ab initio* determination of the forces from quantum mechanics has proved to be very difficult except in recent years in the case of helium, the simplest rare gas. An empirical approach has been consistently more successful. In this approach, a flexible potential form is assumed that is realistic at short and at long range and that lends itself to the inclusion of theoretical values in these regions if available. In order to ascertain the potential in the intermediate range, a multiproperty fit to a judiciously chosen set of properties is performed. The set is carefully chosen so that each property probes and constrains the potential in specific regions. The various dilute-gas properties that have been used are

1. macroscopic properties such as second virial coefficients and transport properties, and
2. microscopic properties such as spectroscopic data on the dimer and beam data (differential scattering cross sections and total, or integral, cross sections in the glory region, that relative velocity region where undulations in the cross sections occur as a consequence of the presence of repulsion and attraction in the interatomic interaction).

1.1.2 Like Pairs Researchers at IBM, Oxford University, and California Institute of Technology as well as others determined fairly accurate potentials in the early 1970s. A summary of the situation to 1982 was given by Aziz (1984), who described the fine tuning of potentials for the pure gases. As more accurate data became available and more sophisticated models were developed, state-of-the-art potentials, notably the multidamped Hartree–Fock dispersion (HFD) and exchange Coulomb (XC), have been subsequently constructed. The problem of determining an accurate potential for helium from *ab initio* and semiempirical points of view has been especially difficult, and only recently has agreement been reached between the two approaches.

A particularly realistic but simple potential is the overall damped HFD in which the interaction energy is partitioned into uncorrelated (Hartree–Fock) and correlated parts and a suitable damping function to account for charge-overlap effects. The form is

$$U(r) = \epsilon U^*(x), \qquad (1)$$

where

$$U^*(x) = A^* \exp(-\alpha^* x + \beta^* x^2) - F^*(x)[c_6/x^6 + c_8/x^8 + c_{10}/x^{10}] \qquad (2)$$

with

$$F^*(x) = \begin{cases} \exp[-(D/x - 1)^2], & x < D, \\ 1, & x \geq D, \end{cases} \tag{3}$$

where $x = r/r_m$ and $c_n = C_n/\epsilon r_m^n$.

Parameters for the pure rare gases appear in Table 2. Figures 1 and 2 show a comparison between HFD and LJ potentials for argon.

Properties for which nonadditive many-body interaction energies are significant in pure rare gases are third virial coefficients, binding energies and structure of solids, pressure–volume relations, etc. The relative importance of many-body interactions is greater in the condensed phases.

1.1.3 Unlike Pairs Interactions for unlike pairs are not as well known as they are for the pure gases, largely because few spectroscopic data are available for these systems. Interatomic potentials determined from gas-phase properties for ten mixed systems are reviewed and critically assessed by Aziz (1984). Again, refinements to the knowledge of the interaction of these systems have subsequently been made as more accurate differential-scattering and diffusion data have become available.

1.2 Compounds

1.2.1 Chemical Compounds The term "chemical compound" is understood to mean a substance that is chemically stable at room temperature and pressure, held together by a chemical bond. This excludes hydrates and other clathrates as well as van der Waals complexes, in which the rare-gas atom is not held by a true chemical bond. Other substances that are excluded but that have been studied extensively include the short-lived rare-gas–cation complexes, some of which are important as lasing media.

The chemical reactivity of the rare gases increases with atomic number, as the outer electron shells get increasingly screened by the inner electrons. Nevertheless, the reactivity is always small compared with those of the other elements, because the outermost shell is filled. One would expect radon to be the most reactive rare gas, and this may indeed be the case. However, its rarity and radioactivity make it difficult to study, and its

Table 2. Parameters for HFD-B potentials (Sec. 1.1.2); σ is the value of r at which $U(r) = 0$. (Enough decimal places are displayed to avoid round-off errors.)

	He	Ne	Ar	Kr	Xe
	HFD-B3-FCI1[a]	HFD-B[b]	HFD-B3[c]	HFD-B2[d]	HFD-B1[d]
A^*	186 924.4	895 718.0	113 211.8	69 735.39	54 408.73
α^*	10.571 75	13.864 35	9.000 534	8.388 022	7.529 583
c_6	1.351 866	1.213 175	1.099 711	1.061 360	1.005 552
c_8	0.414 951	0.532 227	0.545 116	0.568 456	0.583 599
c_{10}	0.171 511	0.245 707	0.392 787	0.426 055	0.473 783
$10^{25}\ C_6(\text{J} \times \text{nm}^6)$	1.398 681	6.172 002	61.557 27	122.8311	271.7901
$10^{27}\ C_8(\text{J} \times \text{nm}^8)$	3.782 664	25.870 09	431.6150	1058.395	3006.292
$10^{28}\ C_{10}(\text{J} \times \text{nm}^{10})$	1.377 556	11.410 81	439.9169	1276.209	4651.408
β^*	−2.077 588	−0.129 938	−2.602 702	−2.796 115	−3.339 043
β^*/r_m^2 (nm^{-2})	−23.580 00	−1.360 000	−18.400 00	−17.380 00	−17.520 00
D	1.438 000	1.360 00	1.040 000	1.208 000	1.114 000
ϵ/k_B (K)	10.956	42.25	143.25	201.3	282.8
r_m (nm)	0.296 830	0.3091	0.3761	0.4011	0.436 56
σ (nm)	0.264 138	0.2759	0.3356	0.3571	0.389 10
Lennard-Jones 6–12 Parameters					
ϵ/k_B (K)	10.22	35.60	119.5	165.49	228.78
σ (nm)	0.2556	0.2749	0.3403	0.366	0.401

---

[a] Aziz *et al.*, 1995.

[b] Aziz and Slaman, 1989.

[c] Aziz and Slaman, 1990.

[d] Dham *et al.*, 1990.

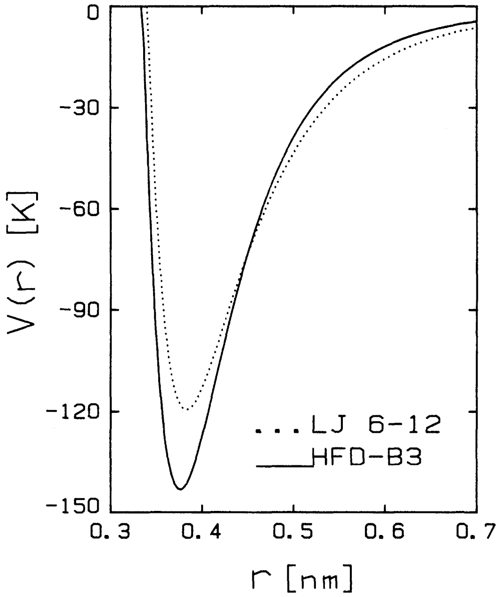

FIG. 1. The attractive region of the argon pair potential. LJ-6-12: the Lennard-Jones 6-12 argon potential of Table 2. HFD-B3: The Hartree–Fock–dispersion argon potential of Table 1.

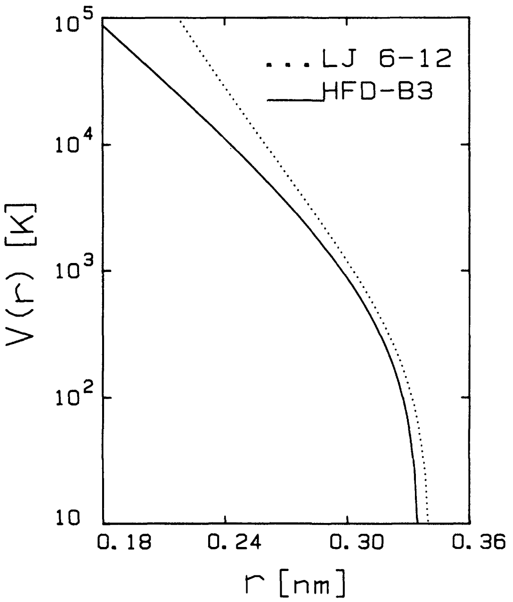

FIG. 2. The repulsive region of the argon pair potential. The same potentials as in Fig. 1.

only known compound is RnF_x, with x unknown.

Xenon is the most reactive of the nonradioactive rare gases. Xenon compounds were predicted by Linus Pauling as early as 1933, although the first one, $XePtF_6$, was not produced until 1962. Subsequently various binary compounds of xenon have been produced. The fluorides XeF_n, with $n = 2$, 4, and 6, are all stable, colorless crystals at room temperature. Evidence of XeF_8 has been reported. Xenon oxides have also been produced. XeO_3 is a white crystal and is dangerously explosive. XeO_4 has been detected but not produced in quantity and is the most volatile xenon compound known. The only other known binary compound of xenon is $XeCl_2$.

There is now a large catalog of more complex xenon compounds (Claassen, 1966). Besides the original $XePtF_6$, a yellow solid, they include $XeOF_4$, a clear, colorless liquid; metallic perxenates such as Na_4XeO_6 that, like the oxides, are very volatile; and mixed fluorides such as XeF_2SbF_5.

Krypton, with atomic number 36, is less reactive than xenon. Its only known binary compound with any stability is KrF_2, which decomposes slowly at room temperature. Some metallic krypton salts such as $KrFAuF_6$ have been made.

The three lighter gases argon, neon, and helium do not form any known stable compounds. However, calculations (Frenking *et al.*, 1990) suggest that all three may form stable compounds with beryllium and oxygen. Spectroscopic evidence for ArBeO has been reported (Thompson and Andrews, 1994), but the compound has not yet been produced in quantity.

1.2.2 Clathrates The weak van der Waals force that acts between two rare-gas atoms can also act between a rare-gas atom and an atom or molecule of another kind, resulting in a weak bond. The attachment of a rare-gas atom to other molecules is further enhanced if the molecules form a cage around the atom. Such a complex is called a clathrate. A familiar example of a clathrate is the hydrate.

Hydrates of argon, krypton, and xenon were prepared as early as 1896. All have a decomposition temperature below 0 °C, although for xenon hydrate it is just −3.4 °C. Many other rare-gas clathrates are now known including RG-hydroquinones (RG =

rare gas), RG-phenols, and the recent RG-fullerenes in which the rare-gas atom is trapped inside a C_{60} "buckyball" (Saunders *et al.*, 1993).

1.2.3 Van der Waals Complexes A physically bound two-atom or molecule-atom system, e.g., Ar_2, ArNe, H_2–Ar (X_2–Y), etc., may form under certain conditions. These dimers can arise when the temperature of the gas is reduced. Such dimers can also be formed in a pulsed supersonic jet. The concentration of dimers is small; for example, only about 0.5% of argon atoms are bound in the vapor phase near NPT (normal pressure and temperature) and are dissociated quickly by colliding atoms. A debate over whether diatomic helium could exist has persisted for many years. Dozens of potentials, *ab initio* and empirical, proposed since the 1939 Lennard-Jones (12-6) potential, produced contradictory conclusions about the existence of He_2. Generally, for potentials with reasonable long-range behavior, those with deeper wells predict the dimer while the shallower ones do not. To make an unequivocal prediction, one requires an accurate potential with, among other things, two key features: an accurate long range and an accurate well depth. The latter was determined from a knowledge of accurate low-temperature second virials provided by the standards laboratories for the ITS-90 temperature scale. Aziz *et al.* (1987) presented such a potential [HFD-B (HE); $\epsilon/k_B = 10.948$ K] from such data, which could predict a large number of other properties. This potential predicted the existence of a weakly bound dimer with a binding energy of 1.6 mK. While earlier *ab initio* potentials had shallower wells that would not predict the virials, recent *ab initio* methods have been refined to the point that *ab initio* and semiempirical potentials are in close agreement (Aziz and Slaman, 1991; Aziz *et al.*, 1995).

There have been a number of groups actively engaged in the pursuit of observing the very weakly bound He_2 dimer that would, by theoretical estimates, have the enormously long bond length of about 5.4 nm. The dimer was first detected in a supercooled beam at the University of Minnesota (Luo *et al.*, 1993) but perhaps the most convincing experimental confirmation lies in the diffraction experiments carried out in Göttingen (Schöllkopf and Toennies, 1994).

1.3 Adsorption

A great deal of experimental and theoretical work has been done on the adsorption of rare-gas atoms on surfaces of various kinds. Much of this work could be categorized as pure rather than applied, dealing, for example, with the general behavior of two-dimensional systems, but there are important applications as well.

Fundamental studies have focused upon using physisorbed layers on structured substrates such as graphite. The pattern of adsorbed rare gas depends upon its atomic size relative to the surface lattice parameter of the substrate, and the resulting layers are commensurate or not depending upon this as well as upon how many layers there are. The resulting systems are excellent prototypes for the investigation of the thermodynamic properties of two-dimensional systems (Nagler *et al.*, 1985). Both the static structures and the dynamics of motion of the adsorbed gas atoms have been studied, the latter using scattering probes such as slow neutrons and electrons.

Rare gases can be separated from gaseous mixtures by adsorption onto certain materials. For example, krypton and xenon can be adsorbed out of impure oxygen onto a certain kind of zeolite, a microporous material sometimes called a "molecular sieve." Conversely oxygen can be adsorbed out of impure argon by another type of zeolite. Neon has also been removed from helium–neon mixtures by adsorption onto activated carbon and silica gel.

1.4 Thermodynamics

1.4.1 Equations of State It has proved difficult to express the behavior of the rare gases over the whole range of pressure *P*, volume *V*, and temperature *T* in a single equation. The first success was the van der Waals equation

$$(P + a/V^2)(V - b) = RT, \qquad (4)$$

which mimics, when used with the Maxwell equal-area construction, all the qualitative features of real equilibrium fluids. The quantity a/V^2 accounts for the attraction between atoms and b accounts for the finite atomic size. The parameters a and b differ from species to species. Many empirical relations have been proposed. Song and Mason (1989) found a compact, analytical equation of state that is simple, general, and accurate (see Sec. 2.3).

A useful equation of state from both a practical and a theoretical point of view is the so-called *virial equation.* In this form, the product PV of n moles of gas is expressed as a power series in $1/v$ where $v = V/n$:

$$Pv = RT[1 + B/v + C/v^2 + D/v^3 + \cdots]. \quad (5)$$

Here B, C, D, etc. are the second, third, and fourth virial coefficients, which depend only on T and the species of gas. The coefficients $B, C, D, \ldots$ represent deviations from ideal behavior when collisions involving two, three, four, ... atoms, respectively, become significant in the gas. Hence B can be related to and provide information on the pair potential, C on the three-body potential, etc.

1.4.2 Heat Capacity and Vapor Pressure The heat capacities of the monatomic rare gases are nearly constant over a wide temperature range. The heat capacity at constant volume is nearly equal to the ideal-gas value of $\frac{3}{2}R$ while the heat capacity at constant pressure is very nearly equal to $\frac{5}{2}R$.

Very accurate values of the saturated vapor pressures P_σ for argon (Gilgen *et al.*, 1994) and other gases along the entire coexistence curve are available in the form

$$\ln(P_\sigma/P_c) = (T_c/T)(A_1\tau + A_2\tau^{1.5} + A_3\tau^{2.5} + A_4\tau^5), \quad (6)$$

where $\tau = (1 - T/T_c)$ where T_c and P_c are the critical temperature and pressure.

Accurate values of the condensed-phase enthalpy, entropy, and internal energy have been determined from accurate vapor-pressure data (Schwalbe *et al.*, 1977; Chen *et al.* 1975, 1978). In addition, accurate values of the cohesive energy at 0 K have been derived from triple-point solid enthalpies and integration of experimental heat capacity of the solid at saturated vapor pressure, C_{sat}, from T_{tp} to 0 K. Cohesive-energy data have been useful in studying many-body forces in solids.

1.4.3 The Universal Gas Constant A new value of the universal gas constant ($R = 8.314\,471 \pm 0.000\,014$ J mol^{-1} K^{-1}) was determined from measurements of the speed of sound in argon as a function of pressure at the temperature of the triple point of water in a spherical acoustic resonator at the National Institute of Standards and Technology (NIST) (Moldover *et al.*, 1988). The error assigned to the measurement is some five times smaller than that assigned to the previous measurements. This reduces the uncertainty of values derived from R, such as the Boltzmann constant k_B and the Stefan–Boltzmann constant.

1.5 Transport Properties

1.5.1 Kinetic Theory Kinetic theory was developed mainly by Maxwell and Boltzmann in the last century. From the laws of mechanics and the theory of probability, it determines the macroscopic properties of gases in terms of the motion of molecules, their collisions with one another, and the forces of interaction. Central to kinetic theory is the time-dependent distribution function. If no gradients of velocity, temperature, or concentration exist, the distribution function reduces to the time-independent Maxwellian distribution. When nonequilibrium conditions prevail, then the distribution function satisfies the Boltzmann integrodifferential equation. When a dilute rare-gas system is not too far from equilibrium, a perturbation method due to Chapman and Enskog is used to solve the equation and derive expressions for the transport properties in terms of the properties of the atoms themselves and their interactions.

Transport phenomena in gases include viscosity, thermal conductivity, and ordinary and thermal diffusion. In terms of kinetic theory, the first three processes are equivalent to the transport through the gas of momentum, kinetic energy, and mass, respectively, when gradients in the gas are present. The viscosity of a gas results from a transport of momentum along a velocity gradient. Thermal conduction is the transport of kinetic energy along a temperature gradient and (ordinary) diffusion is the transport of mass along a concentration gradient. Thermal diffusion is a secondary effect in which an initially uniform gas mixture subjected to a temperature gradient results in a tiny separation of the components.

1.5.2 Measurement of Viscosity and Thermal Conductivity The viscosities of the rare gases and their mixtures are measured (Wakeham *et al.*, 1991) mainly in one of two techniques: capillary flow and oscillating-body viscometry. With the capillary-flow technique, the estimated error of the measurement ranges from ±1% to ±2% while accuracies up to ±0.1% have been claimed with the oscillating-disc method.

Data on the viscosity were measured using the oscillating-disc method at Rostock and Brown Universities (298 to 973 K) and using the capillary method at Oxford (114 to 1600 K) and at Los Alamos National Laboratory (1110 to 2200 K).

The thermal conductivity of the gases can be measured by either a steady-state or a transient method. In the former, the thermal gradient is measured when a heat flux is passed through gas between concentric cylinders or between parallel plates. In the latter, a wire is mounted along the axis of a vertical cylindrical cell and a constant heat flux is electrically generated in it. The thermal conductivity can be determined from temperature rise in the wire as a function of time. The transient hot-wire method seems to be the current method of choice. For very high temperatures, the shock-tube method is used, in which the thermal conductivity is deduced from the measurement of heat transfer rates from the shock-heated gas to the end wall of the shock tube.

1.5.3 Measurement of Ordinary and Thermal Diffusion The most common method of measurement of the diffusion coefficient is with the use of a two-bulb apparatus (Wakeham *et al.*, 1991). In this technique, gases of unequal composition residing in each bulb are allowed to mix through a connecting capillary. The diffusion coefficient is related to the geometry of the two-bulb system and the time required for the gas to approach uniform composition.

There are two basic means of determining the associated transport coefficient, viz., the thermal-diffusion factor. These are a two-bulb technique in which each bulb is maintained at a different temperature, and a thermal-diffusion column. In order to amplify the separation in the case of the former method, the so-called Trennschaukel (swing separator) was devised, in which many two-bulb systems are connected in series. In the column method, a temperature gradient is maintained between a concentric wire and the walls of a vertical tube. The natural convection of the gas in the column amplifies the separation by sweeping the lighter component adjacent to the hot wire up and the heavier component at the cold wall down the column. The column is the prototype apparatus employed for the separation of isotopes. It was the early method for separating uranium isotopes.

Diffusion data on mixtures of the rare gases have been measured at low temperatures at Leiden. More precise and extensive measurements have been performed at Adelaide University including some measurements on thermal diffusion. Self-diffusion and extensive thermal-diffusion measurements were also carried at the Mound Facility of the U.S. Department of Energy in Ohio.

1.6 Law of Corresponding States (LCS)

When expressed in reduced form, a property will have nearly the same value for all gases in a set. This law is based on the hypothesis that all gases in a set have the same reduced intermolecular potential. The reduced potential is defined as $U^* = U(r^*,\beta_i)/\epsilon$, where $r^* = r/\sigma$. The scaling factors ϵ (the energy parameter) and σ (the length parameter) differ from species to species in a set, and β_i are potential-shape parameters, which are identical for all members of the set.

An extended law of corresponding states for the dilute rare gases (helium excepted) and their binary mixtures was developed at Brown University (Kestin *et al.*, 1972). This LCS provided thermodynamically consistent correlations for second virial coefficients and transport properties. On the basis of limited experimental measurements on some of the species, the formulation provided universal functionals for these quantities that, with a set of scaling factors for each species and its binary mixtures, allow one to predict the properties of all the species in a set over an unusually large temperature range.

Realizing the deviation from congruence of reduced potentials for the various species at small separations, the group at Brown (Najafi *et al.*, 1983) improved their LCS principle by allowing some of the shape parameters β_i to be different for different rare-gas pairs. In addition, they appealed to quantum mechanics to determine the asymptotic behavior of the properties. As a result, they were then able to extend the temperature range of the correlation to high and low temperatures and to include helium and its mixtures. The formulation effectively synthesized the equilibrium and transport properties of the five rare gases and the 26 mixtures that can be formed with them.

A further improvement was achieved by the Rostock group (Bich *et al.*, 1987), who published accurate correlations for viscosity and thermal conductivity for the pure gases. It must be noted, however, that data based on accurate interatomic potentials that are more physically based cannot only correlate the data but can extend the range of temperature inaccessible to experiment.

1.7 Optical and Dielectric Properties

1.7.1 Refractivity and Dielectric Virial Coefficients If a rare-gas atom is placed in a uniform electric field, it acquires a dipole moment and is said to be polarized. If the moment **m** is proportional to the field **E** then $\mathbf{m} = \alpha_0 \mathbf{E}$ where α_0 is the polarizability of the isolated atom. In the presence of other atoms, e.g., in a gas, the polarizability α of an atom can be defined in terms of the index of refraction n and the density ρ by the Lorenz–Lorentz equation

$$(n^2 - 1)/(n^2 + 2)\rho^{-1} = A_R + B_R\rho + C_R\rho^2 + D_R\rho^3 \cdots. \tag{7}$$

α will differ from α_0 because of the effects of atomic interactions. If ρ is in moles per unit volume, the left-hand side of Eq. (7) is the molar refraction R_m. As shown, it can be expanded as a power series in the density ρ, where A_R, B_R, C_R, D_R, . . ., respectively, are the first, second, third, fourth, . . . refractivity virial coefficients. A_R is the limiting value of R_m for a very dilute gas. That is, in the limit $\rho \to 0$, $A_R = N\alpha_0/3\epsilon_0$.

Determinations of A_R hence yield values for atomic polarizabilities α_0 for the isolated atom, while B_R, C_R, . . . are related to the excess contribution to R_m as a result of interacting pairs, triplets, . . . of atoms, respectively. Accurate values of A_R, B_R, C_R, . . . have been determined by a differential interferometric technique, and some of these are presented in Table 3 (Achtermann *et al.*, 1993).

The Clausius–Mosotti function for the dielectric constant ϵ may also be expanded as follows:

$$(\epsilon - 1)/(\epsilon + 2)\rho^{-1} = A_\epsilon + B_\epsilon r + C_\epsilon r^2 + D_\epsilon r^3 \cdots. \tag{8}$$

The left-hand side is the molar polarization P_m and A_ϵ, B_ϵ, C_ϵ, D_ϵ, . . . are respectively the first, second, third, fourth, . . . dielectric virial coefficients. Again, these coefficients are related to the contributions from individual atoms, pairs, triplets, quadruplets, . . ., respectively. The limiting value of P_m at low density is $A_\epsilon = N\alpha_0/3\epsilon_0$. For the rare gases, accurate values of the dielectric virial coefficients have been determined by a differen-

Table 3. Measured refractivity and dielectric coefficients for the rare gases (Sec. 1.7.1).

Rare gas	T (K)	A_R (cm^3 mol^{-1})	B_R (cm^6 mol^{-2})	C_R (cm^9 mol^{-3})
He	303	0.5213	−0.068	···
Ne	303	1.0012	−0.11	···
Ar	303	4.1955	1.75	−85.2
Kr	303	6.414	5.96	−275.0
Xe	303	10.345	···	···
	348	10.344	28.5	−1802.0

Rare gas	T (K)	A_e (cm^3 mol^{-1})	B_e (cm^6 mol^{-2})	C_e (cm^9 mol^{-3})
He	323	0.5196	−0.07	···
	3–18	0.006	−5.0	···
Ne	323	0.9969	−0.12	···
Ar	303	···	1.22	···
	323	4.1397	0.72	···
Kr	323	6.273	4.3	···
Xe	323	10.122	32	···

tial-capacitance technique (Huot and Bose, 1991; Gugan, 1984).

1.7.2 Spectroscopy of the Dimers Through the use of clever experimental techniques, absorption and emission bands in the vacuum ultraviolet have been observed that have provided information on the vibrational spacings (Ewing, 1975) for Ne_2, Ar_2, Kr_2, and Xe_2 dimers in the ground state and subsequently, with a spectrograph of higher resolution, on the rotational fine structure of Ar_2 and Kr_2 (Colbourn and Douglas, 1976; LaRocque *et al.*, 1986; Herman *et al.*, 1988). Such information has been invaluable for the accurate determination of ground-state constants and interatomic potentials. Results based on the potentials of Table 2 are given in Table 4. Energy levels for the argon dimer are shown in Fig. 3. Observation of vibrationally and isotopically resolved fluorescence excitation spectra has led additionally to information on some of the excited states involved in excimer lasing transitions. Infrared, microwave, radiofrequency, and Raman spectroscopy have been used to reveal the structure and properties of some of the dimers (Xu *et al.*, 1994).

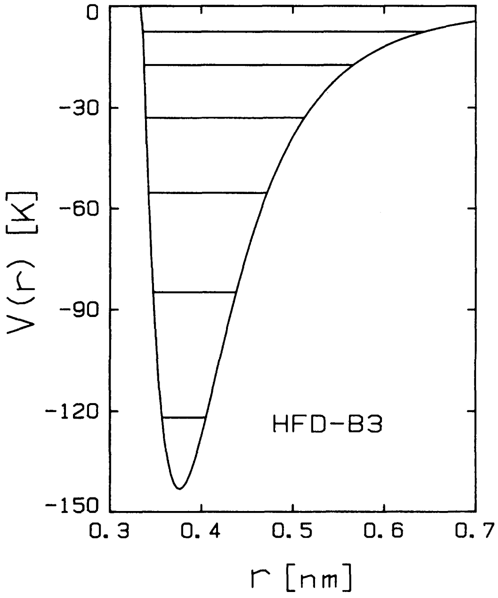

FIG. 3. Energy levels of the argon pair potential. HFD–B3: the Hartree–Fock–dispersion potential of Table 1. The energy levels are those of Table 4.

2. LIQUID STATE AND DENSE GASES

2.1 Interatomic Interactions in the Condensed State

Above the critical temperature, gases and liquids form one continuous state of matter. Coordinates of the critical points for the rare gases are given in Table 1. Experiments on rare-gas fluids can be carried out over very wide ranges of the state variables: pressure *P*, temperature *T*, and density (or molar volume *V*). Because the interatomic pair potentials are well understood, the rare gases form a prototype family of fluids upon which critical experiments can be compared with modern computations of the properties concerned. Such studies form a useful foundation upon which the complications of molecular fluids may be considered, including orientations of the constituents as well as their positions.

It cannot be said that many-body interactions in condensed rare gases are well understood. Quite often, for pressures that are not

Table 4. Vibrational levels of the rare-gas dimers calculated on the basis of potentials of Table 2 (Sec. 1.7.2). (Units in cm^{-1}.)

Level	He	Ne	Ar	Kr	Xe
$-\frac{1}{2}$	0	0	0	0	0
0 (ZPE)	7.613_6	12.3	14.85	11.35	10.23
1	···	26.0	40.57	32.52	29.84
2	···	29.1	61.09	52.20	48.29
3	···	···	76.62	69.40	65.57
4	···	···	87.50	84.51	81.66
5	···	···	94.31	97.52	96.54
6	···	···	···	108.49	110.25
Dissociation energy D_0	1.1×10^{-3}	17.1	84.71	128.56	186.32
Well depth D	7.6147_5	29.365	99.563	139.910	196.555

too large, a model of an accurate pair potential together with the Axilrod–Teller triple-dipole dispersion energy (DDD) is sufficient to explain properties of the fluids and binding energies of the solid phases. For crystals, the reason for this is that it has been assumed that contributions of the many-body dipole-interaction terms of fourth and higher order $(MBD)_{4\to\infty}$ cancel those of higher multipole third-order terms $(DDQ)_3$, $(QQD)_3$, $(DDO)_3$, and $(QQQ)_3$. D, Q, and O refer to dipole, quadrupole, and octupole, respectively, and subscripts 3 and $4 \to \infty$ refer to third and fourth to higher orders of quantum perturbation theory. In spite of the apparent success of the approach, however, it is worthwhile to point out that there are other nonadditive energies (e.g., nonadditive exchange energies) that could be of the same magnitude as (DDD) and that corrections must be incorporated when the simple approaches fail to yield results that agree with experiment, especially at high density (Barker, 1989; McLean *et al.*, 1988; Meath and Koulis, 1991).

Shock-wave experiments have been carried to high densities in dense-fluid rare gases, with compressions a factor of 3 or more (Yakub, 1994).

2.2 The Structure of Liquids

The structure of matter changes continually with time because its constituents are in motion. In the crystalline state, their displacements are small, and (except in solid helium) they are localized near their lattice sites. In the gas phase, the constituents move over such large distances that the term "structure" is rarely used. Since densities in liquids are similar to those of solids, the packings of the two states are somewhat similar. In fact, liquids possess considerable order or "structure." The static structure of a monatomic liquid may be expressed by a pair-distribution function $g(r)$, which is an equilibrium distribution of interatomic separations averaged over time and direction. Precise measurements of $g(r)$ in rare-gas liquids have been made by x-ray or neutron scattering where the intensity of scattering is obtained as a function of angle (Fig. 4). The measured distribution is then compared with one calculated using assumed potentials. Such calculations have advanced the development of theoretical methods and the understanding of the liquid state.

2.3 Thermodynamic Properties

The function $g(r)$ also plays a central role in determining the thermodynamic properties of the system. For example, the internal energy and the pressure can be expressed in terms of the pair-distribution function $g(r)$ and the potential $U(r)$ together with an ideal-gas term, if the interactions are assumed to be additive. For greater accuracy, a three-body distribution function and three-body

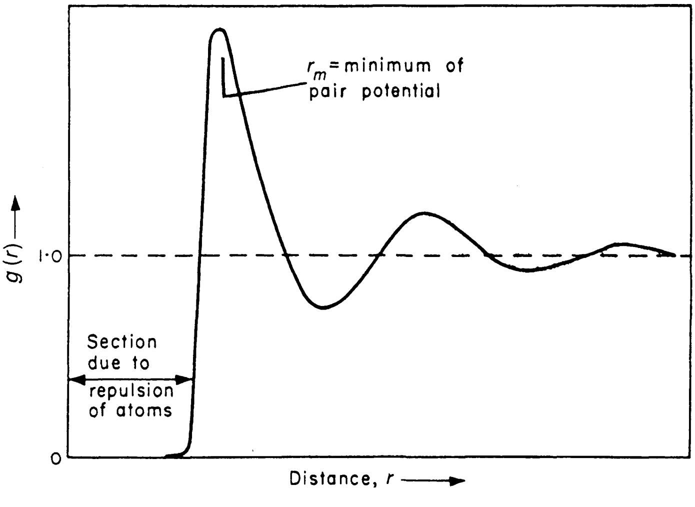

FIG. 4. A sketch of the typical form of the pair-distribution function $g(r)$ for a dense rare-gas liquid. The principal peak of $g(r)$ occurs at a position close to the minimum position r_m of the pair potential. For very low densities these two positions are coincident, and as the density is increased the maximum in $g(r)$ moves to slightly lower values of r than r_m (After Egelstaff, 1992).

potential must be introduced to reflect nonadditive contributions.

Computer-simulation methods can provide, in principle, exact values for most properties on the basis of a given potential model. There are two methods: the Monte Carlo (MC) and the molecular dynamics (MD). The MC method uses ensemble averages as in statistical mechanics and is used to calculate equilibrium properties. The MD method solves the Newtonian equations of motion of the particles and uses time averages. It is used to calculate both equilibrium and transport properties. MC can be applied to both quantum-mechanical and classical systems while MD is applicable only to classical systems. In each method, a small number of particles (less than a few hundred) are used with periodic boundary conditions to simulate the behavior of a large system. Usually an accuracy of 1% or better is attainable for many properties depending on computational speed. Such methods are useful in determining or testing interatomic potentials by comparing the results on the basis of an assumed potential with experiment. The simulated results can also provide a standard by which the results based on approximate methods may be compared.

There are a number of approximate methods to calculate thermodynamic properties that have been or are currently being employed. One of these is the so-called integral equation method. In this procedure, one introduces approximations into $g(r)$ in order to simplify the problem mathematically. One then determines whether such approximations are valid by comparing the results with experiment or computer simulation. Another approximate method meeting with success is the perturbation method (Barker and Henderson, 1976; Weeks *et al.*, 1971). In this approach, the pair potential is written as $U(r) = u_0(r) + \lambda u_i(r)$, where λ is a parameter that, when set equal to zero, gives a reference potential $u_0(r)$ (on the basis of which the properties of the reference system are well known from, say, computer-simulation studies) and, when set equal to unity, gives $U(r)$. The potential is introduced into the partition function and the thermodynamic properties are evaluated. Since the properties of the reference system are known, the properties of the real system can be determined.

Song and Mason (1989) developed a simple analytical equation of state based on perturbation theory for a hard-sphere fluid. By three separate integrations, they determine from a knowledge of the interatomic potential the second virial coefficient, an effective hard-sphere diameter, and a scaling factor as functions of temperature. Since the latter two are nearly universal functions when expressed in suitably chosen reduced units, it results that a knowledge of the second virial coefficient enables one to determine the entire equation of state for the rare gases (except helium; Song and Mason, 1993) up to 3 times the critical density. It allows one to predict such properties as the compressibility of the liquid, the vapor-pressure curve, and the critical constants.

2.3.1 Quantum Corresponding States

The classical restriction of the law of corresponding states for translational states can be removed by introduction of a reduced Planck constant, a measure of quantum effects in the condensed state. For a Lennard-Jones interaction, the de Boer parameter $\Lambda^* = h/(m\epsilon\sigma^2)^{1/2}$ may be regarded as the reduced de Broglie wavelength of a particle of mass m. A property W in reduced form is then expressed by a relation including Λ^* as $W^* = f(P^*,T^*,\Lambda^*)$, where P^* is the reduced pressure. If Λ^* is less than about 0.2, then classical theories of the condensed state are often adequate.

2.3.2 Sound Velocity

Measurements of sound velocity yield the most precise values for adiabatic compressibilities of rare-gas liquids. With knowledge that the heat-capacity ratio is a universal function of the reduced PVT coordinates, it is possible to summarize the sound velocity at saturated vapor v_σ for the heavier rare-gas liquids from their triple points to close to their critical points by a universal relation for the reduced velocity versus reduced temperature, $v_\sigma^* = f(T^*,\Lambda^*) \approx f(T^*)$. The reduction parameters for both velocity and T depend only on the energy parameter ϵ/k_B of some "effective" pair potential. By forcing corresponding states, accurate ratios of the energy parameters of krypton and xenon in terms of that of argon can be obtained (Aziz *et al.*, 1967). In fact, it is interesting to note that velocity appears to be almost perfectly linear with density.

2.4 Transport Properties

2.4.1 Transport Properties, Theory At low densities, transport properties can be accurately calculated at any temperature using exact kinetic-theory expressions. At high densities, however, there is at present no formal theory that provides an exact evaluation of transport properties in terms of a realistic characterization of the atomic interaction.

The MC method has provided significant progress toward a successful molecular theory of transport properties in dense fluids. It requires extensive numerical computation in order to determine the dependence of the transport coefficients on the experimental variables.

Enskog developed a kinetic theory of dense gases consisting of rigid spherical molecules of finite diameter. By considering only two-body collisions and the finite size of the molecule, Enskog modified the kinetic theory of the dilute gas. In this model there is an instantaneous transport of energy and momentum on collision from the center of one molecule to the center of the other; and because of the finite diameter, the frequency of collisions is altered. The density dependence of self-diffusion is given directly by the Enskog theory while viscosity and thermal conductivity require a knowledge of experimental data in order to bring theory and experiment into agreement. There is only one adjustable parameter, the molecular core size, which is determined from the data fit. In this modification as applied to the rare gases (smooth hard-sphere theory), the exact smooth hard-sphere transport coefficients P_{shs} are given by the product of the values from the Enskog theory P_{E} and computed corrections to it: $P_{\mathrm{shs}} = P_{\mathrm{E}}(P/P_{\mathrm{E}})_{\mathrm{MD}}$.

MD methods have been used to characterize the detailed time response in a model Lennard-Jones fluid (de Schepper *et al.*, 1988).

2.4.2 Neutral and Charge Transport, Experiment At the U.S. National Institute of Standards and Technology, extensive measurements have been obtained of the equation of state and the viscosity of liquid argon as a function of density, the latter using a torsional oscillating crystal. Thermal conductivity data for argon are also available. Since acceptable equations of state for krypton and xenon are unavailable, LCS can be used (Hanley *et al.*, 1974) to determine values of viscosity and thermal conductivity as follows. A reduced viscosity is defined as

$$\eta^* = \eta[r_{\mathrm{m}}^2(N/M)^{1/2}(k_{\mathrm{B}})^{-1/2}(\epsilon/k_{\mathrm{B}})^{-1/2}] \qquad (9)$$

and a reduced thermal conductivity is defined as

$$\lambda^* = \lambda[r_{\mathrm{m}}^2(M/N)^{1/2}(k_{\mathrm{B}})^{-3/2}(\epsilon/k_{\mathrm{B}})^{-1/2}] \qquad (10)$$

where ϵ/k_{B} and r_{m} are potential parameters for an "effective" pair potential. A reduced pressure is given by

$$P^* = Pr_{\mathrm{m}}^3[(\epsilon/k_{\mathrm{B}})k_{\mathrm{B}}]^{-1} \qquad (11)$$

and a reduced temperature is given by

$$T^* = T/(\epsilon/k_{\mathrm{B}}). \qquad (12)$$

To obtain η and λ of krypton or xenon at a given temperature and pressure, a corresponding temperature and pressure are first found for argon using Eqs. (11) and (12). The equation of state for argon is used to find the density, and then the transport coefficients can be determined from Eqs. (9) and (10) and the appropriate potential parameters.

Charge-transport studies in rare-gas liquids show that injected electrons are in a quasifree state of high mobility. Carrier transport has typically been determined by time-of-flight methods using electrical shutters (Shikin, 1977). More unusual have been studies of the annihilation of positrons injected into rare-gas liquids (Stewart *et al.*, 1990). A localized bubble state is formed in liquid krypton, argon, neon, and helium (and in solid helium). The bubble radius is determined by a balance between shrinking forces of bubble surface tension and external pressure and the zero-point pressure exerted by the confined positronium. The lifetime for decay of each positronium mode is a measure of the electron charge density at the positron-annihilation site.

2.5 Microscopic Dynamics from Neutron-Scattering Evidence

Slow-neutron scattering has provided a powerful tool for extending studies to include the dynamics of liquid structure. That is, it can be used to measure various correlation functions appearing in dynamical models for liquids—for example, the self-correlation (the average motion of an atom among others) and the density–density correlation (the relative motion of atoms). In practice, argon provides a rare example of a liquid system in which both the coherent scattering (from a ^{36}Ar sample) and also the incoherent scattering (from a suitable mixed ^{36}Ar–^{40}Ar sample) can be determined (Skold *et al.*, 1972). For argon, analysis of the experimental results in view of molecular-dynamics simulations shows good agreement with a picture in which the atoms exhibit free-particle behavior at very short times with a smooth transition to simple diffusion already at times about 2 ps.

At very small scattering vectors Q, i.e., long wavelengths, observed fluid modes correspond to fluctuations in the density, at frequencies $\omega = cQ$, where c is the velocity of sound, and to fluctuations in the entropy, centered at zero frequency. These three peaks broaden and merge at larger Q, about 1 nm^{-1} for the rare gases. At very large Q the modes correspond to single-particle motion. At intermediate values of Q, interpretation of the response of rare-gas fluids is somewhat controversial, although results on argon and neon show interesting differences. Helium seems to be a special case (Montfrooij *et al.*, 1992).

2.6 Special Properties of Liquid Heliums and Their Isotopic Mixtures

Superfluidity in liquid-helium systems has been treated in many thousands of research papers (see SUPERFLUIDITY: LIQUID HELIUM SYSTEMS). The behavior of liquid ^{4}He (a boson) is qualitatively different from that of liquid ^{3}He (a fermion), although both exhibit superfluid phases at low temperatures. The only atom that forms an *anisotropic* liquid is ^{3}He.

Landau gave a phenomenological theory of normal Fermi liquids in 1956, only a few years after sufficient quantities of ^{3}He became available for its thermodynamic properties to be measured. The Landau theory uses the concept of elementary excitations to clarify the properties of strongly interacting quantum systems at low temperatures. It provides a useful way to correlate measured properties directly, such as the transport coefficients, with a limited number of parameters, without the need for a microscopic theory. Subsequent microscopic derivations of the Landau theories have been produced. In ^{3}He these excitations are quasiparticles, a concept that has proven useful in other forms of condensed matter.

Thermodynamic and physical properties of isotopic helium mixtures have an extraordinary range. For example, varying the ^{4}He content in ^{3}He produces an adjustment of the Fermi temperature with little change in number density of the liquid. Measurements of thermal conduction, viscosity, diffusion, etc. have been used to check theories of Fermi liquids over a very wide range of parameters. Work on nuclear matter and on nuclei is parallel, but only a much more limited range of parameters can be studied.

We note that the properties of these extraordinary liquids can be linked to those of the heavier rare gases when the heliums are brought to high density, at which internuclear distances are nearer the characteristic length of the pair potential, and at which the statistics obeyed by the atoms is insignificant (Boninsegni *et al.*, 1994).

3. CRYSTALLINE STATE

3.1 Melting, Corresponding States

The rare-gas solids form an interesting family. Among all interatomic interactions, van der Waals cohesion is the weakest in forming condensed systems, and the pure rare gases have low condensation temperatures (Table 1) and energies (Table 5). One kind of corresponding-states analysis simply uses the triple points as scaling parameters. The search for a critical point at the end of the solid–fluid line in a P-T diagram has been carried furthest in helium, for which one can access experimentally the largest range of V of any substance. No such point has been found.

In the rare-gas family, effects of different kinds occur in a smooth progression as one goes from one row of the periodic table to another. For example, Table 5 shows that with increasing mass comes a decrease in the importance of macroscopic quantum effects (compare the ratios E_z/E_s). The solid

Table 5. Binding energies $E(0)$ of rare-gas solids at 0 K based on the potentials of Table 2. [$E(0) = E_S + E_Z$ where E_S is the static energy of face-centered-cubic lattice and E_Z is the zero-point energy including anharmonic effects. Units of J mol^{-1}. "Complete" energies include $(DDD)_3$, $(DDQ)_3$, $(QQD)_3$, $(DDO)_3$, $(QQQ)_3$, and $(MBD)_{4\to\infty}$. (See Sec. 3 for definitions of these multibody interactions.)]

Energy	Interaction	Ne	Ar	Kr	Xe
E_S	Two-body	−2609.3	−9097.1	−12 776.8	−17 855.8
E_S	Triple dipole	63.0	571.6	1 000.1	1 686.7
E_S	Two-body + triple dipole	−2546.4	−8525.5	−11 776.7	−16 169.2
E_S	Complete many-body	−2537.0	−8444.8	−11 651.9	−15 883.6
E_Z	Two-body	612.7	781.7	597.7	534.0
E_Z	Triple dipole	3.1	7.2	6.9	7.0
E_Z	Two-body + triple dipole	615.8	789.0	604.6	541.0
E_Z	Complete many-body	617.0	793.5	609.4	547.2
$E(0)$	Two-body + triple dipole	−1930.6	−7736.5	−11 172.1	−15 628.2
$E(0)$	Complete many-body	−1920.0	−7651.3	−11 042.5	−15 336.4
$E(0)$	Experiment	−1933 ± 8[a]	−7722 ± 11[b]	−11 148 ± 13[b]	−15 784 ± 35[b]

[a] McConville (1974).
[b] Schwalbe *et al.* (1977).

heliums are a special case, because they require external pressure for solidification; for them, the ratio E_z/E_s is about unity. On the other hand, with increasing mass comes an increase in the relative importance of many-body interactions.

3.2 Thermoelastic Properties

For the heavier rare-gas solids under saturated vapor pressure, the heat capacity and thermal expansivity have been measured with high precision over the entire crystalline ranges above about 1 K. Heat capacity has been well determined by adiabatic calorimetry (Korpiun and Lüscher, 1977). Thermal expansion has been precisely measured by capacitance dilatometry at low temperatures and by x-ray diffraction at temperatures up to melting. X-ray measurements of lattice parameters of single crystals have approached the best accuracy available for usual solids (Losee and Simmons, 1968). On account of very large thermal expansion, the heat capacity at constant pressure for argon, krypton, and xenon rises to about 35 J mol^{-1} deg^{-1}, which is 40% above the Dulong–Petit value $3R$. This fact, together with the relatively large vibrational energies of these solids, has attracted the attention of theorists interested in the dynamics of crystals that are highly anharmonic. For example, because the isobaric coefficient of thermal expansion corresponds to a third derivative of the free energy, this is a most demanding test of a statistical-mechanical model.

Precise values have been obtained of the thermal properties of the solid neon isotopes ^{20}Ne and ^{22}Ne over the entire temperature range. Direct measurements of kinetic energy of crystalline neon, using eV neutron scattering, show that the vibrational energy is about 35% of the cohesive energy. The vibrational energy differs for the two isotopic masses. The isotopes therefore have different internal pressures and exhibit different equilibrium lattice parameters, vapor pressures, specific heats, etc. These can be correlated by a single-particle model in which the dominant agent is the vibrational energy (Batchelder *et al.*, 1968).

Elastic constants of the rare-gas solids are not as well determined. For these solids the equilibrium vapor pressures and thermal expansions are large and the melting temperatures low, which defeats ultrasonic pulse-echo techniques as usually applied to oriented and prepared single crystals. The solid heliums are an exception; for these, ultrasonic transducers can be kept in contact with the crystals confined at constant volume. For other crystals near melting, where single crystals could be kept in good optical contact with transparent cell walls, values have been obtained by Brillouin scattering, which is inelastic scattering of a photon by an acoustic phonon.

Anisotropy in the deduced elastic constants has been attributed to two effects: for the lighter rare gases, large zero-point energies, and for the heavier ones, many-body interactions. These attributions are qualitatively correct, but detailed interpretations remain somewhat inconclusive as long as the temperature dependence of the elastic constants is not well known over the entire temperature range.

Rare-gas solids are extraordinarily compressible compared with other solids. Unlike other materials, for them it is practical to study many properties under essentially isochoric conditions or over large changes in density. Brillouin scattering has been used to measure the velocity of sound (and hence the density) in argon compressed up to 30 GPa (Polian *et al.*, 1986). Isochoric studies of heat capacity have been carried out for the heliums, for neon, and for argon. For heliums, these have been correlated with precise measurements of pressure, using the relation $(\partial S/\partial V)_T = (\partial P/\partial T)_V$. For some phonon branches, dispersion has been studied in isochoric neon and argon samples using slow-neutron scattering.

Measurement of pressure–volume–temperature (*PVT*) relations by piston-displacement techniques have given the density dependence of the bulk modulus B of the heavier rare gases (Swenson, 1977). It has been found that a Birch type of relation often represents the results well at low temperature: $B = B_0 + bP$, where P is the pressure and b is a parameter having a value 5 to 6. It is remarkable that all members of this family that have a solid–liquid–vapor triple point fall on a common curve ($\pm 15\%$). At higher temperatures for the heavier rare-gas solids, the temperature dependence of the solid equation of state is approximately given by $(dP/dT)_V$ = constant, with the constant having a value about 2.7 to 2.9 MPa K^{-1}.

3.3 Microscopic Properties

3.3.1. Phonons, Experiment and Theory Measurements of phonon energies and dispersion, by coherent inelastic scattering of neutrons, provide information about the dynamics of interatomic interactions in rare-gas solids. All rare-gas solids except radon and ^{3}He have been studied in this way, radon because it is radioactive and ^{3}He because of its enormous neutron absorption.

It is conventional to parametrize results by fitting a Born–von Kármán model. This is an atomistic formulation of lattice dynamics that sets up the interaction forces in general fashion. Calculations of properties using the parameters of the model can then be compared with corresponding properties calculated from lattice-dynamical theories. Properties of interest within the harmonic approximation include the one-phonon density of states and its frequency moments, and the lattice specific heat. By examining the energy versus wave-vector relations at very small Q, the "zero-sound" elastic constants are obtained, i.e., those at frequencies large compared with a typical phonon–phonon collision frequency. Once again, the large compressibility of rare-gas solids has permitted measurement of dispersion in some phonon branches in isochoric helium, neon, and argon crystals, which exhibit considerable anharmonicity. This has the great advantage of displaying the intrinsic temperature dependence of anharmonic effects, which are often masked by changes in phonon frequencies due to changes in crystal density from thermal expansion.

Results for the fcc phases have provided a continual object of theoretical attention, as both experiment and theory were refined. For the heavier solids such as argon, very satisfactory agreement has been reached using computer simulations. Theories that use modern interatomic potentials are now able to account for many details of the phonon-excitation line shape as well as for the dispersion structure and hence the bulk thermodynamic properties. In favorable cases, observed agreement between theory and experiment tests the theory, the potential, and the experimental technique. On the other hand, results for solid ^{4}He are interesting, showing multiple-phonon interactions, but they are incomplete for higher phonon energies.

Thermal resistivity in perfect insulators arises from umklapp processes, those phonon-scattering processes for which the rule for conservation of wave vector is $\mathbf{Q}_1 + \mathbf{Q}_2 = \mathbf{Q}_3 + \boldsymbol{\tau}$, where $\boldsymbol{\tau}$ is a vector in the reciprocal lattice. Again, the large compressibility of rare-gas solids makes it possible to investigate thermal conduction in samples at constant volume, which removes the effects of thermal expansion from the analysis. Similarly, studies of the thermal conduction at different densities are possible; such studies are extremely rare in usual solids. At the highest temperatures, accessible for argon, krypton, and xenon, thermal conductivity at constant volume goes as T^{-1}, as given by a phonon-gas model, and it follows a law of corresponding states (Batchelder, 1977). At lowest temperatures, defects and boundary scattering limit thermal conduction. In highly pure and perfect helium crystals, an enhancement of thermal conduction has been observed, which can be attributed to a regime in which Poiseuille flow of the phonons is significant. Helium-isotope crystals have provided important information on mass-defect scattering processes, because the mass ratio is very large.

3.3.2 Momentum Distribution, Kinetic Energy The momentum distribution of individual atoms can be measured by analyzing the neutron recoil-scattering process in terms of neutron energy transfer at a given momentum transfer (Simmons, 1993). Neutrons of energy about 1 eV are required. From the momentum distribution, the kinetic energy can be deduced and compared with the results of simulations in order to test both the simulations and the experiments in a very direct fashion. For the heavier rare-gas solids the results have been according to expectation; for example, the equipartition law holds for translational states regardless of density of the system.

On the other hand, measurements on helium have shown that the actual kinetic energy is significantly larger than what would be expected from the Born–von Kármán (harmonic) analysis of neutron-measured phonon dispersion. Actually, for solid helium, the high-energy excitations corresponding to single phonons are generally so broad and weak that a Born–von Kármán analysis including these states is unjustified. For helium, path-integral Monte Carlo methods have proven most satisfactory (Ceperley, 1995). An illustration is provided by considering kinetic-energy results for samples of ^{4}He at fixed number density but in different phases, bcc or hcp or liquid (depending upon P and T); such a study is only possible with helium. The result is that the kinetic energies of these three phases are essentially the same, that is, helium is a quantum system dominated by the density, not by details of the local order (Blasdell *et al.*, 1993).

3.4 Dielectric and Electronic Properties

Experimental results on many dielectric and electronic properties of the rare-gas solids can be fruitfully interpreted using atomic characteristics. These solids have very large forbidden-energy gaps for electrons (Rössler, 1977). The gaps at low pressures range from about 9.3 eV in xenon to about 30 eV in helium. For these prototype insulators, optical properties at energies below the valence-band transitions can be succinctly expressed by the Lorentz–Lorenz function that relates the dielectric constant or index of refraction to the atomic polarizability [Eq. (7)]. Quantitative (phenomenological) descriptions of the dielectric properties can be constructed on the assumption that principal contributions arise from excitons (bound electron–hole pairs, the hole being a positive charge carrier) originating from the atomic doublet $^1S_0$-$^{3,1}P_1$.

Measurements of UV and XUV (extreme ultraviolet) reflectivity and absorption data on the heavier rare-gas solids have been correlated and extended by Kramers–Kronig analysis to give the complex optical constants (Sonntag, 1977). These have been further compared with electron energy-loss spectra. Again, information on atomic properties, and theoretical calculations of electron structure, are useful for interpretation of spectral features, although explanation of the observed intensities continues to provide a challenge to theorists. Data near the x-ray absorption edges are of special interest and have been collected for all rare gases xenon through helium (Schell *et al.*, 1995). Exciton effects are prominent, and the type of excitons observed depends upon the particular rare gas (Schwentner *et al.*, 1985; Varding *et al.*, 1993).

Drastic modification of the electron structure is possible at high pressures. For example, usual xenon (fcc structure at 34.7 cm^3 mol^{-1}) has a gap between the 5p valence band and the (empty) 6s conduction band. When compressed in diamond-anvil pressure cells solid xenon is found to transform at 17 GPa to a complex hcp structure, then at about 75 GPa to hcp, then finally to a metallic state, presumably through approach of the 5p and 5d electron bands. The observed pressure of transformation is 130–150 GPa, at a density of about 11 $\mathrm{cm}^3\ \mathrm{mole}^{-1}$. In xenon, the metallization has no accompanying structural transition; x-ray diffraction shows a hcp structure from 75 to 175 GPa (Jephcoat *et al.*, 1987). For lighter rare-gas solids, expected pressures of transformation are higher, for example above 600 GPa in solid argon. The results of attempts to reach such pressures in shock-wave experiments are difficult to interpret because the accompanying large temperature rise also causes electronic excitations, apparently beginning at about 40 GPa in argon. Some insight into these processes has been gained through observations at high pressures on ionic compounds having an ion with the rare-gas–like electron configuration $(ns)^2(np)^6$ and on structural transformations in metals that involve promotion of electrons to the d band (Ross and Radousky, 1989).

Finite–sample-size effects were first seen in observations of the cutoff of excitonic series in thin films. More recently, as cluster physics and chemistry develop, measurements have been made of the soft x-ray spectra of rare-gas clusters of sizes from about 10^2 to about 10^4 atoms, produced in jets. Finally, electron energy-loss spectra have been collected both from thin films and from clusters. In films, evidence of plasma excitations has been seen.

Charge transport has been investigated using excess carriers injected by pulsed-electron excitation of the heavy rare-gas solids (Spear and le Comber, 1977) and by ionizing decay products of radioisotopes embedded in solid heliums. Charged-particle excitation has proven more convenient than photon excitation, because condensed rare gases exhibit photoemission more readily than photoconduction. For the solids formed from neon, argon, krypton, and xenon, low-field electron mobilities typically display a $T^{-3/2}$ temperature dependence, which is characteristic of quasifree-electron transport limited by phonon scattering. On the other hand, the drift mobility of excess holes (positive charge carriers) in these solids is typically several powers of ten smaller in magnitude than for electrons. In this case, it is more appropriate to describe the process as the formation and transport of localized entities, such as a polaron, a charge carrier trapped in its own polarization field. In this case the temperature dependence of the process is dominated by a term $\exp(-E_b/2k_BT)$, where E_b is a characteristic binding energy for the polaron. At high electric fields, electron velocity depends on essentially the square root of the applied field, an indication that the electron distribution becomes "hot." It saturates at a velocity near that of sound.

Altogether, rare-gas solids are providing rich sources of information about how electronic properties of condensed matter in various degrees of aggregation can be quantitatively related to the properties of the constituent atoms.

3.5 Crystal Defects

3.5.1 Point Defects The formation of equilibrium atomic vacancies has been measured in all rare-gas solids. The concentrations (n/N) are determined directly and absolutely by precise comparisons of changes in x-ray lattice parameter $(\Delta a/a)$ and of macroscopic expansion $(\Delta L/L)$ at a succession of temperatures. The applicable formula is

$$(n/N) = 3(\Delta L/L) - 3(\Delta a/a). \tag{13}$$

From the concentration the free energy of formation of the vacancies, f, follows from the relation

$$(n/N) = \exp(-f/k_BT). \tag{14}$$

The free energy has two parts: an energy e, and a nonconfigurational entropy s. From the magnitude of the entropy some idea of the local structure of the vacancy can be obtained. Table 6 displays some directly measured values for thermal-vacancy energies.

Other physical properties are affected by the presence of equilibrium defects—for example, the heat capacity and the compressibility. Inferences of vacancy content using these other properties depend, however, upon the assumption that one knows the behavior of a hypothetical perfect crystal, so that the vacancy contribution can be obtained by difference. Such assumptions have frequently proven unreliable, and so a direct method is preferred.

Rare-gas solids are composed of neutral atoms, so that comparison of their cohesive

Table 6. Some thermal-defect properties of rare-gas solids (Sec. 3.5.1). (The formation energies e correspond to the values shown for the entropies, s/k_B. The last column is a test for the influence of many-body interactions, present if the ratio is less than unity. For Ar an average value for e is used.)

Rare gas	$f/k_B T_M$	s/k_B	e (meV)	$e/(E_P - E_Z)$
Neon[a]	9.7	1.6	21	1.0
Argon[b]	8.5	0–4	62–93	0.9
Krypton[c]	7.8	2.8	86	0.8
Xenon[d]	7.0	~2	~100	0.6

[a] Schoknecht and Simmons, R.O. (1972).
[b] Schwalbe (1976).
[c] Losee and Simmons (1968).
[d] Granfors *et al.* (1981).

energies per atom with the energy of vacancy formation gives direct information about the presence of many-body forces and of local lattice relaxation around the vacancy site. Such a comparison is given in the last column of Table 6 for entropies near $2k_B$. The table shows that many-body effects become more important as one goes from solids composed of the less polarizable (neon) to the more polarizable atoms.

Self-diffusion in all rare-gas solids except ^{4}He has been measured by NMR methods or by exchange with radioisotopes in the vapor, because the usual sectioning of samples containing radiotracers is not practicable, on account of the low melting points and high vapor pressures of these solids. A consistent picture has emerged (Chadwick and Glyde, 1977). The dominant mechanism of self-diffusion in all solids except the heliums is the motion of single vacancies. In solid heliums, diffusion may occur both through direct interchange of atoms and through vacancy migration.

In principle, point defects analogous to some of the "color centers" in ionic crystals can be created in rare-gas solids by ionizing radiation. A likely example is a dimer–ion analogous to the V_K center found in alkali halides.

Extrinsic point defects such as foreign atoms, molecules, or radicals can be studied over a vast range of photon energies, microwave through x ray, because of the closed-shell character and large band gap of the rare-gas host. Low temperatures slow reaction kinetics so that ephemeral processes can be studied. Neutron scattering is also a useful tool to examine molecular excitations when the nearly (neutron) transparent ^{40}Ar is used as host (Langel, 1992).

3.5.2 Line and Surface Defects, Thin Films, Clusters Dislocations and surface defects in rare-gas solids have been studied extensively by transmission electron microscopy (Venables and Smith, 1977). Contrast methods for stacking faults (a planar defect in the crystal resulting from misstacking of successive planes of atoms) are applicable as in other thin films. Arrays of dislocations (a one-dimensional defect) are seen, and twinning as in other fcc crystals. Because of high equilibrium vapor pressures, grain boundaries that intersect a bulk surface can be thermally etched. In thin films of neon, both fcc and hcp structures are observed. The classic rare-gas structural question was "Why are most bulk structures fcc and not hcp" (as would be expected in a static lattice containing only pair interactions)? The answer is that experiments show that the energy differences are very small, and that relatively small amounts of zero-point motion, of many-body interactions, etc. are required to produce the observed fcc structures.

Clusters of rare-gas atoms can be formed in supersonic jets. Cluster size can be varied by varying the conditions of the jet. For argon, clusters of fewer than about 800 atoms exhibit a quasicrystalline (isocahedral) structure as analyzed by electron diffraction, whereas larger clusters exhibit the fcc structure of bulk argon (Kovalenko *et al.*, 1994; Schöllkopf and Toennies, 1994). Simulations suggest that in small neutral clusters there may be a dynamical coexistence between crystal and fluid structure that permits the investigation of melting mechanisms (Matsuoka *et al.*, 1992). For argon clusters, excitation by x rays or electron bombardment can produce effects such as x-ray fluorescence and fragmentation of the clusters. The phenomena can be consistently linked to atomic and bulk properties (Wörmer *et al.*, 1991). For helium clusters, however, considerable inconsistencies in such an interpretation remain (Joppien *et al.*, 1993). As usual, heliums are a special case in this rare-gas family.

3.6 Solid Heliums, Unique Properties

The two helium isotopes, ^{4}He and ^{3}He, are the only substances that do not solidify at atmospheric pressure even when their temperatures approach 0 K. External pressure is required, larger for ^{3}He (about 3.1 MPa) than for ^{4}He (about 2.5 MPa). The quantum zero-point motion of the lighter isotope is larger, producing a larger internal pressure to be overcome in order to bring the substance to a density at which the interatomic spacing is close to the distance at which the interatomic potential energy is a minimum. Table 1 indicates the remarkable differences in the properties of the helium solids from the other members of the family. These properties present a special challenge to lattice-dynamical theory. Max Born first addressed the possibility of a self-consistent phonon theory, and such theories, numerically computed, have been the most successful so far for phonons in solid heliums.

The isotope ^{3}He has a pronounced minimum in its melting curve; that is, the thermodynamic derivative along the solid–fluid coexistence line, $(\partial P/\partial T)_V$, is negative over a significant range of temperature. Qualitatively, the reason for this may be seen from the Maxwell relation, which can be used to equate the slope of the melting curve to the change in entropy with change in volume across the phase boundary: $(\partial S/\partial V)_T = (\partial P/\partial T)_V$. The liquid is less dense than the solid phase. However, in the temperature range concerned, the liquid phase is a Fermi liquid having a high degree of order in the spin system (although it has the usual positional disorder of a liquid) whereas the solid phase has much magnetic disorder (it is well above its magnetic transition temperature below 10 mK and hence has entropy about $R \ln 2$). This has led to very unusual studies such as the production of spin-polarized *solid* ^{3}He by sudden compression of the liquid, by cooling through sudden pressurization at temperatures below the minimum, and so on.

The magnetic structure of solid ^{3}He depends upon its density and upon the temperature. Magnetism in ordinary materials is accounted for by quantum-mechanical "exchange" of electron spins, which may appear to be a theoretical construct. In ^{3}He, which has only paired electrons, the spin concerned is nuclear, which, although linked to a very weak magnetic moment, becomes a dominant interaction at low enough temperatures and high enough densities. Because of the extraordinary mobility of the atoms, even in the solid state, there can be literal atomic exchanges (actually, in rings of three or of four atoms) to produce several observed magnetic phases (Roger *et al.*, 1983).

The properties of isotopic mixtures have been systematized by Edwards and Balibar (1989).

3.7 Defects and Excitations in Solid Heliums

By x-ray lattice-parameter measurements, thermal-vacancy concentrations have been directly measured in both hcp and bcc phases of both ^{3}He and ^{4}He (Simmons, 1994). The deduced activation energies have the expected variation with volume; that is, they are larger at smaller molar volumes. Nevertheless, the results are puzzling, because the concentrations are larger than those inferred indirectly from measurements (in isochoric samples) of pressure change and of specific heat. Because of the extremely high diffusivities measured by NMR in ^{3}He, explanations have been sought in vacancy excitations that are nonlocalized as a result of rapid tunneling processes. It has even been proposed, but not yet observed, that the ground states of these crystals contain vacancies. This would be one of the more bizarre consequences of the well-known nonlocalization of the atoms in solid heliums.

Negative-ion transport in the solid heliums can be attributed to motion of a hollow bubble surrounding a trapped electron (Dahm, 1986; Golov and Mezhov-Deglin, 1994). Such "bubble" states have been seen in both liquid neon and liquid helium, and in the latter have been used to probe excitations in superfluid helium samples. Bubble radii, determined by minimization of the total energy (electron confinement versus surface energy of the bubble), are thought to be about 0.7 nm in neon and from 1 to 2 nm in helium, depending upon the applied pressure.

The isotope ^{4}He also has a minimum in its melting curve, near 0.7 K, and the fact that at this temperature the adjacent liquid phase is superfluid has remarkable consequences. From the relation $(\partial S/\partial V)_T = (\partial P/\partial T)_V$ one sees that at and near this temperature the latent heat of freezing (melting) is zero or small. At the same time, the adjacent superfluid has extremely large thermal conduction, so that heat can be carried away from (or toward) the interface very rapidly. There is an interfacial energy, that is, a surface tension to be associated with the interface, so that "melting-freezing waves" can exist at this interface (Parshin, 1982; Chevalier *et al.*, 1994). Indeed, the observed frequencies of oscillation can be used to deduce information about the interfacial energy and structure. Observed velocities of propagation for these waves show the fastest freezing exhibited by any macroscopic condensed matter.

Fundamental optical studies of crystal growth have been made on ^{4}He and ^{3}He crystals growing in superfluid (the latter at 1 mK; Wagner *et al.*, 1994). Roughening transitions are seen, that is, temperatures at which the crystal–fluid surface changes from atomically smooth to rough.

4. SCIENTIFIC AND TECHNOLOGICAL APPLICATIONS

Scientific and technological applications for rare gases cover an extremely broad range. Many depend upon the inert property of the rare gases. The descriptions that follow are indicative, not comprehensive.

4.1 Arc Welding and Cutting

In volume of material, the largest use for argon and helium lies with gas-shielded welding and cutting of metals in an electric arc. They are used to exclude surrounding air from the molten weld metal. In this way, welds can be made in metals such as aluminum, magnesium, molybdenum, nickel, titanium, zirconium, and some stainless-steel alloys that are otherwise difficult to join. For some applications, one or more chemically active gases, such as hydrogen, may be added to control arc characteristics.

4.2 Passive Atmospheres and Inert Condensed Environments

The second-largest use for argon and helium, in terms of total quantity of gas, is as a passive atmosphere when processing or sparging reactive metals or other materials such as semiconductors. They are also used to regulate the rate of chemical reactions that depend on the concentration of the active gas. The excellent thermal conductivity of helium provides an efficient heat-transfer medium during the annealing of metals.

Condensed rare gases have often been used as inert matrices for spectroscopic studies of molecules and radicals that cannot otherwise be stabilized for extended study. Their transparency to electromagnetic radiation is extraordinary, and the low temperatures serve to retard kinetic processes. Spectroscopic properties of the guest molecules and radicals can therefore be studied from the microwave region through the far ultraviolet. This trapping can be carried to extreme selectivity: for examples, studies of the electron spin resonance of cesium atoms in solid helium and the influence on polarized light by rubidium implanted in solid xenon.

4.3 Illumination

A minor use of rare gases in terms of quantity, but all-pervasive in devices, is in sources of illumination. Electrical light sources fall into two groups: incandescent and gaseous discharge. In the case of incandescent lamps, a metal filament is heated to incandescence to produce light, usually in an atmosphere of argon gas together with a small amount of nitrogen. High-intensity incandescent bulbs are filled with krypton. Gaseous-discharge tubes produce light when an electric current is passed through an ionized gas between two electrodes. The color of the light produced depends on the specific gas present in the tube. Neon-discharge tubes, for example, emit light mainly in the red or orange region of the spectrum, while helium or argon or mixtures of these gases are used to produce other colors. Fluorescent lamps contain some argon, neon, or krypton to facilitate initiation of the discharge in the mercury vapor. Similarly, argon helps to start the electric discharge in mercury-vapor or sodium-vapor lamps. Xenon-filled flashlamps provide the photographer with economic, reproducible, brilliant, and brief illumination with a spectral distribution similar to that of daylight. A continuous source of extremely bright illumination for use in sports stadiums and in motion picture production is produced by an arc in high-pressure xenon placed in a quartz envelope.

4.4 Lasers

When highly collimated and monochromatic light is desired, rare-gas laser sources may be used. In the familiar helium–neon laser, laser emission results from excited levels of neutral neon atoms that have been populated by collision with excited helium atoms. Commercial versions are normally of a sealed-tube configuration. They typically have an output power of 1 mW and a lifetime about 10 000 h. Lasers based upon Ne–He mixtures are among the most widely used, including such technical applications as precise land-based surveying. Argon- and krypton-ion lasers provide high cw laser output power in the visible and near-ultraviolet regions of the spectrum but only at the expenditure of a considerable amount of electrical power. Other lasers involving rare gases are the He–Cd and pulsed rare-gas halide lasers. The latter class of "excimer" lasers produce stimulated emission as a result of transitions from a strongly bound excited state to a weakly bound ground state. For example, the laser transition occurs when $Kr^+ + F^-$ in a bound $^2\Sigma$ state undergoes a transition to the thermally unstable $X^2\Sigma$ state, which then dissociates into neutral ground-state atoms.

4.5 Thermometry

The International Temperature Scale of 1990 (ITS-90), adopted by the International Committee of Weights and Measures, is defined, in part, in terms of the properties of rare gases for its realization below 25 K. This scale (Preston-Thomas, 1990) supersedes previous scales, which have serious differences with the thermodynamic temperature. In the range from 0.65 to 5.0 K, ITS-90 is defined in terms of the *vapor pressure* of ^{3}He and ^{4}He. In the range from 3.0 K to the triple point of neon (24.5561 K), ITS-90 is defined in terms of a ^{3}He or ^{4}He gas thermometer of the constant-volume type calibrated at three experimentally realizable temperatures (the triple point of neon, the triple point of equilibrium hydrogen, and a temperature between 3.0 and 5.0 K determined in a prescribed way using a ^{3}He or ^{4}He vapor-pressure thermometer). The nonideality of the gas must be accounted for explicitly using the appropriate and accurate virial coefficient. To this end, many standards laboratories proceeded to measure these coefficients with the highest possible degree of accuracy.

4.6 Cryogenics and Refrigeration

The liquid heliums serve as cryogenic fluid baths at temperatures that can be precisely controlled through control of the vapor pressure. With readily available pumping the limiting temperatures are about 1.2 and 0.3 K for ^{4}He and ^{3}He, respectively. The isotopic mixture of ^{4}He and ^{3}He can be used to produce a continuously operating refrigerator at temperatures down to about 3 mK. ^{3}He displays about 6% solubility in the liquid mixture, even at the lowest temperatures. Concentrated ^{3}He floats on a mixture at higher temperatures, providing a source of ^{3}He, which dissolves into the dilute ^{3}He phase. This is analogous to ordinary evaporation, with accompanying absorption of heat. On the other hand, ^{3}He has a higher vapor pressure, which permits it to be distilled out of the resulting mixture and recirculated continuously.

Large-scale use of liquid ^{4}He occurs in the cooling of superconducting magnetic structures in accelerators and detectors used by high-energy physicists. This use will be important until high-T_c superconductors can be used in these structures, which experience extremely large magnetic forces and must operate with changing currents.

4.7 Other Uses

Rare gases have found an extremely wide variety of other applications. Among these for helium are as buoyant filling for nonrigid airships and balloons, and as a gas for wind tunnels at extremely high Mach numbers. A "wind" tunnel for unusual hydrodynamic regimes could use liquid helium. Perhaps most familiar to scientists are uses for highly sensitive vacuum leak detection with mass spectrometers and as a carrier gas in chromatography. Rare gases have been a common filling for gas detectors of radiation. High-energy particle physicists continue to develop the use of liquid Xe, Kr, and Ar as vacuum-ultraviolet scintillation materials that have high quantum efficiency, uniformity, and resistance to radiation damage.

An important application of rare-gas adsorption is the measurement of the total surface area of a microporous material. The area can be determined if the pores are large enough to allow the gas to enter and if its adsorption behavior on the material is well known. Conversely the nonadsorption of helium makes it useful in volume measurements of microporous materials. Helium penetrates the finest pores of a substance without being adsorbed appreciably. The volume of displaced helium then provides an accurate measurement of the volume of the substance.

All the rare gases have been investigated for biological effects and possible medical use. Helium is useful in breathing atmospheres in deep diving or in sealed environments, and it and xenon have been investigated for use as nonexplosive anesthetics. The radioactive isotopes ^{85}Kr and ^{87}Kr have found some use, for example as tracers in studies of circulation and in radiotherapy.

It is remotely possible that xenon oxides and perxenates could find application as specialized explosives or oxidizing agents (Asimov, 1966). Another possible application of solid or liquid rare-gas compounds would be as a storage method for the gas. Such compounds, if sufficiently stable, might eliminate the need for the high-pressure and cryogenic facilities currently used to store the fluids.

GLOSSARY

Chemical Compound: A substance that is chemically stable at room temperature and pressure, held together by a chemical bond.

Clathrate: A **van der Waals force** can act between a rare-gas atom and an atom or molecule of another kind, resulting in a weak bond. This attachment is further enhanced if the molecules form a cage around the rare-gas atom. Such a complex is called a clathrate.

Correlation Function: In dynamical models for liquids, for example the self-correlation (the average motion of an atom among others) and the density–density correlation (the relative motion of atoms).

Equation of State: For a gas, a relation over a range of pressure P, volume V, and temperature T in an equation, such as $PV = RT[1 + B/V + C/V^2 + D/V^3 + \cdots]$. Here B, C, D, etc. are the second, third, and fourth **virial coefficients,** which depend only on T and the species of gas. For condensed systems, the relation $P(V,T)$ generally has no such simple analytical form.

Extrinsic Point Defects: Examples are foreign atoms, molecules, or radicals.

HFD Potential: In the overall damped HFD potential the interaction energy is partitioned into uncorrelated (Hartree–Fock) and correlated parts and a suitable damping function to account for charge-overlap effects.

Kinetic Theory: From the laws of mechanics and the theory of probability, it determines the macroscopic properties of gases in terms of the motion of molecules, their collisions with one another, and the forces of interaction.

Landau Theory: A phenomenological theory of normal Fermi liquids, which are formed of spin-$\frac{1}{2}$ particles. It uses the concept of elementary excitations to clarify the properties of strongly interacting quantum systems at low temperatures. It provides a useful way to correlate measured properties directly, such as the transport coefficients, with a limited number of parameters, without the need for a microscopic theory. Subsequent microscopic derivations of the Landau theories have been produced. In ^{3}He these excitations are quasiparticles, a concept that has proven useful in other forms of condensed matter.

Law of Corresponding States: A physical relation based on the hypothesis that all gases in a set have the same reduced intermolecular potential. The reduced potential is defined as $U^* = U(r^*,b_i)/e$, where $r^* = r/s$, where e and s are scaling parameters for the energy and length, respectively. One kind of corresponding-states analysis simply uses the triple points as scaling parameters.

Many-Body Interactions: Many-body multipole interaction terms of third order are $(DDQ)_3$, $(QQD)_3$, $(DDO)_3$, and $(QQQ)_3$ and those of fourth and higher order $(MBD)_{4\to\infty}$. D, Q, and O refer to dipole, quadrupole, and octupole, respectively, and subscripts 3 and 4→∞ refer to third and to fourth to higher orders of quantum perturbation theory. Short-range many-body interactions may result from charge overlap.

Monte Carlo (MC) and Molecular Dynamics (MD) Methods: The MC method uses ensemble averages as in statistical mechanics and is used to calculate equilibrium properties. The MD method solves the Newtonian equations of motion of the particles and uses time averages. It is used to calculate both equilibrium and transport properties.

Pair-Distribution Function $g(r)$: An equilibrium distribution of interatomic pair separations averaged over time and direction, used to describe the static structure of a liquid or amorphous material.

Path-Integral Monte Carlo (PIMC) Method: A simulation technique employing Feynmann's path-integral approach to extend the applicability of the **Monte Carlo** method to quantum systems.

Phonon: In the vibrations of crystal lattices, a collective mode with well-defined energy and wave vector. Generally, there is phonon dispersion: The energy depends upon the wave vector.

Polaron: A charge carrier trapped in its own polarization field.

Quantum Corresponding States: The classical restriction of the law of corresponding states for translational states can be removed by introduction of a reduced Planck constant, a measure of quantum effects in the condensed state. For a Lennard-Jones interaction, the de Boer parameter $L^* = h/(m\epsilon s^2)^{1/2}$ may be regarded as the reduced de Broglie wavelength of a particle of mass m.

Umklapp Processes: Those **phonon**-scattering processes for which the rule for conservation of wave vector is $\mathbf{Q}_1 + \mathbf{Q}_2 = \mathbf{Q}_3 + \boldsymbol{\tau}$, where $\boldsymbol{\tau}$ is a vector in the reciprocal lattice.

Vacancy: A vacant atomic site in a crystal.

Van der Waals–London Interaction: An attraction that varies as the inverse sixth power of the separation of two electric dipole oscillators.

Van der Waals Complex: A two-atom system physically bound by the van der Waals force, e.g., Ar_2, ArNe, H_2–Ar(X_2–Y), etc.

Virial Coefficients, Refractivity and Dielectric: Those pertaining to optical refraction and to dielectric properties.

Works Cited

Achtermann, H. J., Hong, J. G., Magnus, G., Aziz, R. A., Slaman, M. J. (1993), *J. Chem. Phys.* **98,** 2308–2318.

Asimov, I. (1966), *The Noble Gases,* New York: Basic Books.

Aziz, R. A., Bowman, D. H., Lim, C. C. (1967), *Can. J. Chem.* **45,** 2079–2086.

Aziz, R. A. (1984), in: M. L. Klein (Ed.), *Inert Gases: Potentials, Dynamics, and Energy Transfer in Doped Crystals,* Berlin: Springer-Verlag.

Aziz, R. A., McCourt, F. R., Wong, C. W. (1987), *Mol. Phys.* **61,** 1487–1511.

Aziz, R. A., Slaman, M. J. (1989), *Chem. Phys.* **130,** 187–194.

Aziz, R. A., Slaman, M. J. (1990), *J. Chem. Phys.* **92,** 1030–1035.

Aziz, R. A., Slaman, M. J. (1991), *J. Chem. Phys.* **94,** 8047–8053.

Aziz, R. A., Janzen, A. R., Moldover, M. R. (1995), *Phys. Rev. Lett.* **74,** 1586–1589.

Barker, J. A. (1989), in: A. Polian, P. Loubeyre, N. Boccara (Eds.), *Simple Molecular Systems at Very High Density,* New York: Plenum, pp. 331–351.

Barker, J. A., Henderson, D. (1976), *Rev. Mod. Phys.* **48,** 587–672.

Batchelder, D. N., Losee, D. L., Simmons, R. O. (1968), *Phys. Rev.* **173,** 873–880.

Batchelder, D. N. (1977), in: M. L. Klein, J. A. Venables (Eds.), *Rare Gas Solids,* Vol. 2, London: Academic, Chap. 14.

Bich, E., Millat, J., Vogel, E. (1987), *Wiss. Z. WPU Rostock* **36** (8), 5–11.

Blasdell, R. C., Ceperley, D. M., Simmons, R. O. (1993), *Z. Naturforsch. A* **48,** 433–437.

Boninsegni, M., Pierleoni, C., Ceperley, D. M. (1994), *Phys. Rev. Lett.* **72,** 1854–1857.

Ceperley, D. M. (1995), *Rev. Mod. Phys.* **67,** 279–355.

Chadwick, A. V., Glyde, H. R. (1977), in: M. L. Klein, J. A. Venables (Eds.), *Rare Gas Solids,* Vol. 2, London: Academic, Chap. 19.

Chen, H. H., Aziz, R. A., Lim, C. C. (1975), *J. Chem. Thermodynam.* **7,** 191–199; *ibid.* (1978), **10,** 649–659.

Chevalier, E., Guthmann, C., Rolley, E., Balibar, S. (1994), *Physica B* **194–196,** 919–920.

Claassen, H. (1966), *The Noble Gases,* Boston: Heath.

Colbourn, E. A., Douglas, A. E. (1976), *J. Chem. Phys.* **65,** 1741–1745.

Dahm, A. J. (1986), in: D. F. Brewer (Ed.), *Pro-*

gress in Low-Temperature Physics, Vol. 10, Amsterdam: North-Holland, pp. 73–137.

de Schepper, I. M., Cohen, E. G. D., Bruin, C., van Rihs, J. C., Montfrooij, W., de Graaf, L. A. (1988), *Phys. Rev. A* **38,** 271–287.

Dham, A. K., Allnatt, A. R., Meath, W. J., Slaman, M. J., Aziz, R. A. (1990), *Chem. Phys.* **142,** 173–189.

Dobbs, E. R. (1994), *Solid Helium Three,* Oxford: Clarendon.

Edwards, D. E., Balibar, S. (1989), *Phys. Rev. B* **39,** 4083–4097.

Egelstaff, P. A. (1992), *An Introduction to the Liquid State,* 2nd ed., Oxford: Clarendon.

Ewing, G. F. (1975), *"Structure and Properties of van der Waals Molecules," Accounts Chem. Res.* **8,** 185–192.

Frenking, G., Koch, W., Reichel, F., Cremer, D. (1990), *J. Am. Chem. Soc.* **112,** 4240–4256.

Gilgen, R., Kleinrahm, R., Wagner, W. (1994), *J. Chem. Thermodynam.* **26,** 399–413.

Glyde, H. R. (1995), *Excitations in Liquid and Solid Helium,* Oxford: Clarendon.

Golov, A. I., Mezhov-Deglin, L. P. (1994), *Physica B* **194–196,** 951–952.

Granfors, P. R., Macrander, A. T., Simmons, R. O. (1981), *Phys. Rev. B* **24,** 4753–4763.

Gugan, D. (1984), *Metrologia* **19,** 147–162.

Hanley, H. J. M., McCarty, R. D., Haynes, W. M. (1974), *J. Phys. Chem. Ref. Data* **3,** 979–1017.

Herman, P. R., LaRocque, P. E., Stoicheff, B. P. (1988), *J. Chem. Phys.* **89,** 4535–4549.

Huot, J., Bose, T. K. (1991), *J. Chem. Phys.* **95,** 2683–2687.

Jephcoat, A. P., Mao, H. K., Finger, L. W., Cox, D. E., Helmley, R. J., Zha, C. S. (1987), *Phys. Rev. Lett.* **59,** 2670–2673.

Joppien, M., Karnbach, R., Möller, T. (1993), *Phys. Rev. Lett.* **71,** 2654–2657.

Kestin, J., Ro, S. T., Wakeham, W. A. (1972), *Physica* **58,** 165–211.

Korpiun, P., Lüscher, E. (1977), in: M. L. Klein, J. A. Venables (Eds.), *Rare Gas Solids,* Vol. 2, London: Academic, Chap. 12.

Kovalenko, S. I., Solnyshki, D. D., Verkhovtseva, E. T., Eremenko, V. V. (1994), *Low-Temp. Phys.* **20,** 758–763.

Langel, W. (1992). *Spectrochim. Acta, Part A* **48,** 405–428.

LaRocque, P. E., Lipson, R. H., Herman, P. R., Stoicheff, B. P. (1986), *J. Chem. Phys.* **84,** 6627–6641.

Losee, D. L., Simmons, R. O. (1968), *Phys. Rev.* **172,** 944–957.

Luo, F., McBane, G. C., Kim, G., Giese, C. F., Gentry, W. R. (1993), *J. Chem. Phys.* **98,** 3564–3567.

Matsuoka, H., Hirokawa, T., Matsui, M., Doyama, M. (1992), *Phys. Rev. Lett.* **69,** 297–300.

McConville, G. T. (1974), *J. Chem. Phys.* **60,** 4093.

McLean, A. D., Liu, B., Barker, J. A. (1988), *J. Chem. Phys.* **89,** 6339–6347.

Meath, W. J., Koulis, M. (1991), *J. Mol. Struct. (Theochem.)* **226,** 1–37.

Moldover, M. R., Trusler, J. P. M., Edwards, T. J., Mehl, J. B., Davis, R. S. (1988), *Phys. Rev. Lett.* **60,** 249–252.

Montfrooij, W. de Graaf, de Schepper, I. M. (1992), *Phys. Rev. B* **45,** 3111–3114.

Nagler, S. L., Horn, P. M., Rosenbaum, T. F., Birgenau, R. J., Sutton, M., Mochrie, S. G. J., Moncton, D. E., Clarke, R. (1985), *Phys. Rev. B* **32,** 7373–7383.

Najafi, B., Mason, E. A., Kestin, J. (1983), *Physica* **119A,** 387–440.

Parshin, A. Ya. (1982), *Physica* **110B,** 1819–1829.

Polian, A., Loubeyre, P., Grimsditch, M. (1986), *Phys. Rev. B* **33,** 7192–7200.

Preston-Thomas, H. (1990), *Metrologia* **27,** 3–10 and 107 (erratum).

Roger, M., Hetherington, J. H., Delrieu, J. M. (1983), *Rev. Mod. Phys.* **55,** 1–64.

Ross, M., Radousky, H. B. (1989), in: A. Polian, P. Loubeyre, N. Boccara (Eds.), *Simple Molecular Systems at Very High Density,* New York: Plenum, pp. 47–84.

Rössler, U. (1977), in: M. L. Klein, J. A. Venables (Eds.), *Rare Gas Solids,* Vol. 1, London: Academic, Chap. 8.

Saunders, M., Jimenez-Vasquez, H., Cross, R., Poreda, R. (1993), *Science* **259,** 1428–1430.

Schell, N., Simmons, R. O., Kaprolat, A., Schülke, W., Burkel, E. (1995), *Phys. Rev. Lett.* **74,** 2535–2538.

Schoknecht, W. E., Simmons, R. O. (1972), in: N. G. Graham, H. E. Hagy (Eds.), *Symposium on Thermal Expansion,* New York: American Institute of Physics, pp. 169–182.

Schöllkopf, W., Toennies, J. P. (1994), *Science* **266,** 1345–1348.

Schwalbe, L. (1976), *Phys. Rev. B* **14,** 1722–1732.

Schwalbe, L. A., Crawford, R. K., Chen, H. H., Aziz, R. A. (1977), *J. Chem. Phys.* **66,** 4493–4502.

Schwentner, N., Koch, E.-E., Jortner, J. (Eds.) (1985), *Electronic Excitations in Condensed Rare Gases,* Berlin: Springer-Verlag.

Shikin, V. B. (1977), *Sov. Phys. Usp.* **20,** 226–248.

Simmons, R. O. (1993), *Z. Naturforsch. A* **48,** 415–424.

Simmons, R. O. (1994), *J. Phys. Chem. Solids* **55,** 895–906.

Skold, K., Rowe, J. M., Ostrowski, G., Randolph, P. D. (1972), *Phys. Rev. A* **6,** 1107–1131.

Song, Y., Mason, E. A. (1989), *J. Chem. Phys.* **91,** 7840–7853.

Song, Y., Mason, E. A. (1993), *Phys. Rev. E* **47,** 2193–2196.

Sonntag, B. (1977), in: M. L. Klein, J. A. Venables (Eds.), *Rare Gas Solids,* Vol. 2, London: Academic, Chap. 17.

Spear, W. E., le Comber, P. G. (1977), in: M. L. Klein, J. A. Venables (Eds.), *Rare Gas Solids,* Vol. 2, London: Academic, Chap. 18.

Stewart, A. T., Briscoe, C. V., Steinbacher, J. J. (1990), *Can. J. Phys.* **68,** 1362–1376.

Swenson, C. A. (1977), in: M. L. Klein, J. A. Venables (Eds.), *Rare Gas Solids,* Vol. 2, London: Academic, Chap. 13.

Thompson, C. A., Andrews, L. (1994), *J. Am. Chem. Soc.* **116,** 423–424.

Varding, D., Becker, J., Frankenstein, L., Peters, B., Runne, M., Schröder, A., Zimmerer, G. (1993), *Low-Temp. Phys.* **19,** 427–436.

Venables, J. A., Smith, B. L. (1977), in: M. L. Klein, J. A. Venables (Eds.), *Rare Gas Solids,* Vol. 2, London: Academic, Chap. 10.

Wagner, R., Andreeva, O. A., Ras, P. J., Remeijer, P., Steel, S. C., van Woerkens, C. M. C. M., Frossati, G. (1994), *Physica B* **194–196,** 961–962.

Wakeham, W. A., Sengers, J. V., Nagashima, A. (Eds.), (1991), *Experimental Thermodynamics III. Measurement of the Transport Properties of Fluids,* London: Blackwell.

Weeks, J. D., Chandler, D., Andersen, H. C. (1971), *J. Chem. Phys.* **54,** 5237–5247.

Wörmer, J., Joppien, M., Zimmerer, G., Möller, T. (1991), *Phys. Rev. Lett.* **67,** 2053–2056.

Xu, Y., Jäger, W., Gerry, M. C. L. (1994), *J. Chem. Phys.* **100,** 4174–4180.

Yakub, E. S. (1994), *Low-Temp. Phys.* **20,** 579–598.

Further Reading

Allen, M. P., Tildesley, D. J. (Eds.) (1993), *Computer Simulation in Chemical Physics,* Dordrecht: Kluwer.

Barker, J. A., Auerbach, D. J. (1985), *Surf. Sci. Rep.* **4,** 1–99.

Bennemann, K. H., Ketterson, J. B. (Eds.), (1976/8), *The Physics of Liquid and Solid Helium,* Vols. I and II, New York: Wiley.

Cook, G. A. (Ed.) (1961), *Argon, Helium, and the Rare Gases,* Vols. 1 and 2, New York: Interscience.

Dobbs, E. R. (1994), *Solid Helium Three,* Oxford: Clarendon.

Duke, C. B. (Ed.) (1994), *Surface Science: The First Thirty Years,* Amsterdam: North-Holland.

Duley, W. W. (1995), *UV Lasers: Effects and Applications in Material Science,* Cambridge, U.K.: Cambridge Univ. Press.

Dymond, J. H., Smith, E. B. (1980), *The Virial Coefficients of Pure Gases and Mixtures,* Oxford: Clarendon.

Egelstaff, P. A. (1992), *An Introduction to the Liquid State,* 2nd ed., Oxford: Clarendon.

Glyde, H. R. (1995), *Excitations in Liquid and Solid Helium,* Oxford: Clarendon.

Glyde, H. R., Svensson, E. C. (1987), "Solid and Liquid Helium," in: D. L. Price and K. Skold (Eds.), *Methods of Experimental Physics,* Vol. 23, Pt. B, London: Academic, Chap. 13.

Gray, C. G., Gubbins, K. E. (1984), *Theory of Molecular Fluids,* Vol. 1, Oxford: Clarendon.

Hirschfelder, J. O., Curtiss, C. F., Bird, R. B. (1964), *Molecular Theory of Gases and Liquids,* New York: Wiley.

Horton, G. K., Maradudin, A. A. (Eds.), *Dynamical Properties of Solids,* especially Vol. 2 (1975) and Vol. 7 (1995), Amsterdam: North-Holland.

Jorgenson, C. K., Frenking, G. (1990), *Structure Bonding* **73,** 1–15.

Kauffman, G. B. (1988), *J. College Sci. Teaching* **17,** 264–268.

Klein, M. L., Venables, J. A. (Eds.) (1976/7), *Rare Gas Solids,* Vols. 1 and 2, London: Academic.

Klein, M. L. (Ed.) (1984), *Inert Gases: Potentials, Dynamics, and Energy Transfer in Doped Crystals,* Berlin: Springer-Verlag.

Maitland, G. C., Rigby, M., Smith, E. B., Wakeham, W. A. (1981), *Intermolecular Forces,* Oxford: Clarendon.

Polian, A. Loubeyre, P., Boccara, A. (Eds.) (1989), *Simple Molecular Systems at High Density,* New York: Plenum.

Price, D. L., Skold, K. (Eds.), (1987), *Methods of Experimental Physics,* Vol. 23, Pt. B, London: Academic.

Saha, G. (1992), *Fundamentals of Nuclear Pharmacy,* 3rd ed., New York: Springer-Verlag.

Schwentner, N., Koch, E.-E., Jortner, J. (Eds.) (1985), *Electronic Excitations in Condensed Rare Gases,* Berlin: Springer-Verlag.

Smith, B. L. (1971), *The Inert Gases: Model Systems for Science,* London: Wykeham.

Varotsos, P. A., Alexopoulos, K. D. (1986), *Thermodynamics of Point Defects and Their Relation with Bulk Properties,* Amsterdam: North-Holland.

Wilks, J., Betts, D. S. (1987), *An Introduction to Liquid Helium,* 2nd ed., Oxford: Oxford Univ. Press.

RAREFIED-GAS DYNAMICS

Franklin C. Hurlbut, *Department of Mechanical Engineering, University of California at Berkeley, Berkeley, California, U.S.A.*

Introduction 97
1. **Historical Summary** 98
1.1 Fundamental Beginnings, to 1939 98
1.2 Second Phase, the Expansion Period, to 1978 98
1.3 Contemporary Phase; the Growing Sophistication of Tractable Problems 99
2. **Theoretical Descriptions** 99
2.1 Preliminary Discussion 99
2.2 The Boltzmann Equation 100
2.3 Momentum and Energy Transfers in Free-Molecule Flow 101
2.3.1 Background 101
2.3.2 Momentum- and Energy-Transfer Coefficients 102
2.3.3 Forces of Lift and Drag on a Flat Plate, General Expressions 103
2.4 Direct Molecular Flow Simulation Using Monte Carlo Methods 104
2.5 Gas–Surface Interactions 105
3. **Experimental Methods** 108
3.1 Ground-Based Experimental Systems 108
3.1.1 Wind Tunnels and Shock Tunnels 109
3.1.2 Flow Production 109
3.1.3 Molecular and Ion Beams 109
3.2 Instruments and Diagnostic Methods 110
3.3 Rocket and Satellite Experimental Systems 111
4. **Applications** 111
Glossary 112
Works Cited 112
Further Reading 114

INTRODUCTION

Rarefied-gas dynamics rests upon the theoretical disciplines of statistical mechanics and the molecular theory of gases and derives essential parameters from several areas of the physical sciences. In this subject we are concerned with those dynamic processes in which the discrete particle nature of gases plays a critical role in the modes of analysis and the prediction of behavior. The gas is said to be "rarefied" wherever the ratio of the mean distance between molecular collisions (mfp) to a characteristic length of the flow exceeds a very small value, usually of the order of 0.01. This ratio, the Knudsen number, *Kn*, expresses the departure of the flow from a near-continuum regime at low values of *Kn* to one at large values of *Kn* in which intermolecular collisions occur infrequently.

At one time it was thought that rarefied gases were found principally in laboratory vacuum systems or at extreme altitude. Rarefied gases were identified as being of "low density" as compared with the "normal density" of our familiar atmosphere. One sees, however, that the phenomena of rarefaction are a matter of scale, independent of the absolute number density of molecules or the physical size of the field, so long as binary molecular collisions predominate. Flows about microscopic filaments, through small orifices, or in particle beds may be subject to the effects of rarefaction, just as are flows about orbiting bodies. Regimes of rarefied flow are customarily defined in terms of the Knudsen number, as follows:

Near-continuum flow, $0.01 < Kn < 0.1$; (1a)

Transition flow, $0.1 < Kn < 10$; (1b)

Nearly free-molecule flow, $10 < Kn < 100$. (1c)

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Discussions of the analysis and applications of rarefied flows in these regimes and in the regime of free-molecule flow in which no intermolecular collisions occur comprise the material of this article.

The physical phenomena associated with the rarefied-gas–dynamic regimes are so singularly numerous and so varied that it is not possible to categorize them in compact form. This is true in part because of the wide range of energy transfers that lie within these regimes, and in part because of the extreme variation in scale of the phenomena that lie within its domain. It is not surprising that scientists from so many areas of the physical sciences and engineering have contributed. A possible unifying factor may be found in the attractions that lie beyond the limits of continuum fluid mechanics. In addition there is always the promise of exploration, to learn, as we have learned in the past few years, of our Earth and its relation to the universe. Each period in the history of rarefied-gas dynamics has been filled with a sense of exploration, but each has had its particular nature. It is useful to distinguish among these as we approach this subject.

The formative years, which began with the rise of classical physics, were marked by the great works of Maxwell and Boltzmann during which the mathematical consolidation of the molecular hypothesis was achieved. Pioneering studies continued during this period until its close in about 1939. There next came a period of expansion and discovery in which, aided by rapidly developing research technology and by man's incursion into space, our knowledge grew with unprecedented speed. In the third and present period there exists a new departure in inquiry; we are enabled to look more deeply into gas-dynamic processes and we now have computational tools allowing the exploration of problems that once seemed unsolvable. This article examines these perceptions and skills as they relate to present applications.

1. HISTORICAL SUMMARY

1.1 Fundamental Beginnings, to 1939

Natural scientists had agreed over many years on the atomic nature of matter, but it was not until the first of the twentieth century, following Dunoyer's (1911) demonstration using a molecular beam of sodium, that the discrete particle structure of dilute gases became fully accepted. Still, the model of the particulate gas had been used by kinetic theorists for at least ninety years prior to that experiment. Maxwell had formulated the law for the distribution of molecular velocities in such a gas in 1859, and by 1866 had made substantial progress in establishing the basis for modern theories of nonuniform gases, that is, gases in which gradients of temperature, velocity, and density are present. Boltzmann (1872) published the celebrated equation that would guide the acquisition of much of our present knowledge of gas transport properties and that provided the link between continuum fluid motion and that of dilute gases. At the same time, he also presented his H theorem, which demonstrated the inevitable relaxation of nonuniform gases toward the Maxwellian distribution of velocities. Efforts to determine general expressions for the distributions of velocity in nonuniform gases met with little success until Chapman in 1916 and Enskog in 1917 published work leading to equivalent terms in expressions for the general velocity distribution function f and, thus, provided the background for the modern theory of transport coefficients. This phase in the development of rarefied-gas–dynamic theory was essentially completed in 1935 by Burnett, who introduced Sonine polynomials as terms in the expansion for f, and extended the order of expansion to the second. A readable, more detailed account of this period may be found in the well-known treatise by Chapman and Cowling (1939).

1.2 Second Phase, the Expansion Period, to 1978

The close of World War II marked the beginning of a period of an extraordinary increase in interest in the field of rarefied-gas dynamics. Attention was now given to simplification of the Boltzmann equation, as in the work of Bhatnagar *et al.* (1954), and to alternative modes of solution for the full equation, as is seen in the work of Grad (1949). Experimental laboratories were set up in departments of physics, chemistry, engineering, and aeronautics throughout the world and also in the research arms of defense departments and aeronautics agencies such as NACA, the National Advisory Committee for Aeronautics, precursor to NASA. Hirschfelder *et al.* (1954) published a monumental resource incorporating theoretical background and experimental results over a broad range of topics relating to collision processes and transport properties, covering much of importance to that year. Sputnik electrified the world in 1957 and many nations turned more directly to the application of rarefied-gas dynamics in space. An important series of symposia was begun in 1958 by Devienne at Nice, France. These meetings, the International Symposia on Rarefied Gas Dynamics, have brought together a significant number of the world's specialists every two years. The Proceedings of the symposia serve to define the field and to provide a compact entrée to it. A list of volume titles appears elsewhere in this article.

New experimental techniques were developed or adapted to a new environment. These included optical methods involving the excitation of fluorescence radiation, making possible noninvasive determinations of temperatures and velocities in monatomic gases and rotational temperatures in diatomic gases. Cryogenic pumping was introduced to enhance the extreme low-density capability of rarefied-gas wind tunnels. Several sources of flow were developed, notably the free-jet expansion. Plasma torch and arc-jet stagnation heaters were introduced to elevate the stagnation enthalpy. Space does not permit enumeration of the large number of investigators, but the record may be found in the proceedings of the RGD symposia. Rapid growth was also seen in the number and sophistication of studies on the interaction of gas molecules and surfaces. These were undertaken in numerous laboratories throughout the world by several thermal and dynamic methods. A few of these are discussed here in Sec. 3. It is particularly worth noting that the experimental study and parametrization of free-jet expansions by Ashkenas and Sherman (1966), which led to the frequent application of these flows in work with rarefied-gas wind tunnels, also led to the revolutionary method of aerodynamic intensification for molecular beams proposed by Kantrowitz and Gray (1951), now very widely used.

This second phase was one of great experimental activity, resulting in enormous extensions of our understanding of molecular collision parameters and of the aerodynamics and thermodynamics of bodies in high-speed rarefied-gas flows, and in our knowledge concerning the processes of momentum and energy exchange in gas–surface interactions. It was also a period of important theoretical activity, much of which was directed to the further exploration of the Boltzmann equation and its derivatives. Other interesting work was undertaken on "first-collision" departures from the free-molecule regime.

By the end of the 1970s there began a general reduction in activity, due, as one immediate cause, to changes in funding patterns, and we may mark the close of the second period as corresponding with this time.

1.3 Contemporary Phase; the Growing Sophistication of Tractable Problems

Beginning in the middle years of the 1980s the content of rarefied-gas dynamics broadened substantially in many aspects but most particularly in connection with the interest in more realistic treatments of molecular collisions, inelastic processes, and chemical reactions. The method of direct simulation using Monte Carlo techniques had been introduced by Bird in the late 1960s. During subsequent years the development of this method had been continued by Bird and by others, but it was not until several years had passed that the method of DSMC reached the current level of acceptance in the scientific community. This important topic is discussed more carefully here in Sec. 2.4. The contemporary period has just begun, but an important feature is visible; it is marked by a significant growth, both experimental and theoretical, in attention to inelastic and reactive collision processes, a growth coincident with and dependent upon the growing capability of DSMC.

2. THEORETICAL DESCRIPTIONS

2.1 Preliminary Discussion

Modern descriptions of rarefied gases rest upon concepts that are well verified by physical observation; gases are composed of discrete molecular particles that are small by comparison with their mutual separation and with the mean distance between collisions. Molecules are constantly in motion; their trajectories are interrupted by collisions with other molecules or with surfaces; the collisional potentials of interaction are relatively short ranged; in general analytical representations it is assumed that only binary collisions occur. The distribution of molecular velocities for such a gas in thermal equilibrium translating at velocity U with respect to a certain coordinate system is known as the translating or drifting Maxwellian and is written as follows:

$$f(\xi) = (n/2\pi RT)^{3/2} \exp[-(\boldsymbol{\xi} - \mathbf{U})^2/2RT]. \qquad (2)$$

In equation (2), n is the number density, T is the thermodynamic temperature, R is the species gas constant, $\boldsymbol{\xi}$ is the molecular velocity, and $\mathbf{U}$ is the velocity of translation. The difference between $\boldsymbol{\xi}$, the total velocity, and $\mathbf{U}$ gives the local velocity of thermal motion, the *peculiar* velocity. With the help of equation (2) and standard concepts of the kinetic theory and thermodynamics relations, we can find most of the expressions applied in free-molecule aerodynamic calculations. This matter will be taken up in subsequent paragraphs.

2.2 The Boltzmann Equation

It should be noted that the macroscopic velocity $\mathbf{U}$ is determined by averaging the total velocities of all molecules in a certain region, whether or not the gas is rarefied. Values of other quantities such as the flux of momentum across a control surface are determined by taking the average of appropriate velocity components over the velocity distribution function. However, where values of field properties such as density alter significantly within lengths of the order of the mean distance between collisions, the gas may be slightly or perhaps greatly out of equilibrium, in which event the Maxwellian distribution cannot be used in performing the weighted average. When the departures from equilibrium are small we are enabled to find velocities, temperatures, densities, and so on, through application of the Boltzmann equation. The Boltzmann equation is a conservation equation expressed in terms of the complete molecular velocity distribution, f. The equation equates the time rate of change in the number of particles of a certain velocity class within a certain small volume due to the physical movement of those particles to the change in the number within that same velocity class in that same volume due to molecular collisions. Solutions of the Boltzmann equation are found in terms of the distribution function f.

The Boltzmann equation is an integro-differential equation of considerable complexity. The several approaches to solutions are discussed in Schaaf and Chambré (1958) and in the monograph for the *Encyclopedia of Physics* (Schaaf, 1963) in which Schaaf has provided an extensive entrée to the literature to 1960. In their monumental work, Chapman and Cowling (1939) set forth a detailed and authoritative coverage. Excellent accounts are also found in Hirschfelder *et al.* (1954) and in Vincenti and Kruger (1965).

Solutions for f are often sought as a series expansion in powers of a small number, essentially the Knudsen number. If f_0 is the Maxwellian then f may be written

$$f = f_0[1 + \phi_1(\lambda/L) + \phi_2(\lambda/L)^2 + \cdots]. \qquad (3)$$

In Eq. (3) λ is the mean distance between collisions and L is a characteristic length of the flow. Clearly λ/L must be small. At normal densities, gradients of velocity and temperature are usually sufficiently small to permit the assumption that the system is in local equilibrium within each elementary region. In this event $f = f_0$. From the zeroth order of the expansion, Eq. (3), we are led directly to the Euler equations. The Navier-Stokes equations result from the retention of the first-order term in λ/L while the Burnett equations (1935) result from the retention of the second-order term. It is sometimes useful to think of small departures from this condition in terms of the ratio of Mach number M to Reynolds number. Since the kinematic viscosity ν is proportional to $a\lambda$, where a is the sound speed, and the Reynolds number is given by $Re = VL/\nu$, in which V is a velocity and L is a characteristic length, we find that $\lambda/L \approx M/Re$. Thus M/Re is a measure of rarefaction equivalent to the Knudsen number for small departures from continuum flow.

Tests of the validity of the Navier-Stokes equations and of the Burnett equations have been made under conditions that minimize the influence of wall–gas collisions. Sherman (1955) and Talbot and Sherman (1959) examined shock-wave profiles in weak shock waves of helium and air, and later in argon. It was found that the Navier-Stokes computations produced shock-wave profiles in fair agreement with experiment, but except for the very low–Mach-number calculations, the Burnett terms did not yield stable solutions. This matter is discussed again in subsequent paragraphs.

In applications of the Navier-Stokes equations in the presence of boundaries within regimes of near-continuum flow, it is necessary to apply slip-flow and temperature-jump boundary conditions. These boundary conditions express an important effect of rarefaction. Within a distance λ from the surface, the gas cannot be considered exactly at rest with respect to the surface, nor, where a gradient of temperature exists, can the gas be exactly at the surface temperature. Thus, there will be a "slip" of velocity and a "jump" in temperature at the wall, although these will be of small magnitude where the Navier-Stokes equations remain valid.

2.3 Momentum and Energy Transfers in Free-Molecule Flow

2.3.1 Background During the years when the foundations of the kinetic theory were taking shape, little thought was given to the idea that man's own satellites would orbit the Earth. However, collisionless flows were of practical and theoretical interest to physicists and chemists, so that by the early 1900s an increasing number of persons were measuring momentum and thermal transfers and attempting to understand the relation between these effects and the nature of the bounding surface. The Maxwellian distribution of molecular velocities was experimentally verified by Stern (1920), wall–gas interaction studies were conducted by Millikan (1923), and heat transfers in rarefied gases were examined by Knudsen (1911). By the end of the first quarter of the twentieth century, a great deal had become known about processes that were of low thermal order and that involved modest molecular velocities. The context broadened greatly with the flight of Sputnik, but much of the mathematical background for the description of flows about satellites had already been well established. There still remained the problem of describing the aerodynamic and thermal effects of wall–gas collisions.

Free-molecule flows are considered to occur for values of the Knudsen number greater than about 100. The flow is collisionless only in the sense that intermolecular collisions do not alter the distribution of velocities incident upon a body; however, the Maxwellian distribution assumed for the incident flow would normally have arisen from intermolecular collisions at remote distances or from diffusely scattering surroundings. Densities and molecular velocities in the vicinity of the body, on the other hand, become greatly modified by wall collisions. This point is elaborated in paragraphs to follow.

A body in free-molecule flow is assumed to be subject to the collisions of large numbers of independently moving molecules. Our knowledge of molecular energy, momentum, and number flux incident upon an element of surface or across any arbitrary elementary plane, and our knowledge of energy density and other densities, must be reported in terms of averaged values. These averages are found by integrating appropriate quantities, usually products of velocities, over the molecular velocity distribution function. Coordinates for an element of surface are illustrated in Fig. 1. As is customary in aerodynamic calculation, the attack angle θ is defined to be taken between the surface tangent and the macroscopic velocity vector **U**, and the axes are oriented to place **U** in the (x_1,x_2) plane. The number flux to an elementary surface dA is

$$dN_i = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty} f\xi_1 d\xi_1 d\xi_2 d\xi_3, \qquad (4)$$

in which f is the Maxwellian distribution from Eq. (2). It is useful to define a molecular speed ratio S as the ratio of the macroscopic speed of flow, U, to the most probable molecular speed, $(2RT)^{1/2}$. Integration of equation (4) leads to

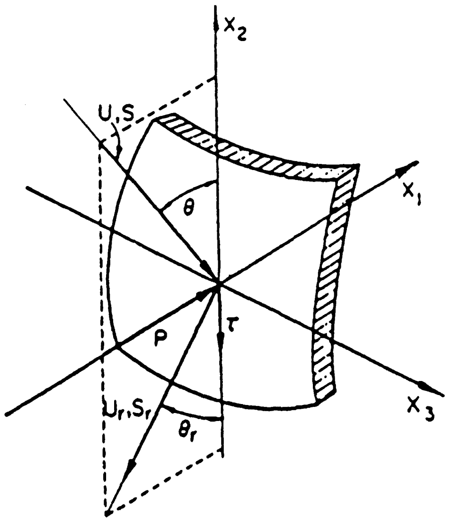

FIG. 1. Schematic diagram of coordinates for molecular flow at the surface.

$$dN_i = n\sqrt{RT/2\pi}\{\exp[-(S\sin\theta)^2] + \sqrt{\pi}(S\sin\theta)[1 + \mathrm{erf}(S\sin\theta)]\}dA. \quad (5)$$

Conservation in number density at the surface is assumed so that dN_r, the reemitted number flux, is equal to that incident, $dN_r = dN_i$. Expressions for the flux of momentum incident to the surface in the normal direction, the flux of momentum in the tangential direction, and the flux of energy are found in a similar fashion. Let p_i be the flux of normal momentum to the surface as given by the integration of $m\xi_1^2 f d\xi_1 d\xi_2 d\xi_3$, and let τ_i be the flux of tangential momentum given by the integration of $-m\xi_1\xi_2 f d\xi_1 d\xi_2 d\xi_3$. Let $S\theta = S\sin\theta$. We may then write

$$p_i = \frac{\rho U^2}{2\sqrt{\pi}S}[(S\theta)e^{-(S\theta)^2} + \sqrt{\pi}[0.5 + (S\theta)^2][1 + \mathrm{erf}(S\theta)]\}, \quad (6)$$

$$\tau_i = \frac{\rho U^2\cos\theta}{2\sqrt{\pi}S}\{e^{-(S\theta)^2} + \sqrt{\pi}(S\theta)[1 + \mathrm{erf}(S\theta)]\}. \quad (7)$$

The flux of incident energy, which may be found in an analogous manner, must include that due to the translational energy, $\frac{1}{2}m\xi^2$, and that of the internal energy, $\frac{j}{2}RT$, in which j is the number of internal degrees of freedom:

$$E_i = E_{i,\mathrm{tran}} + E_{i,\mathrm{int}}. \quad (8)$$

We find

$$E_{i,\mathrm{tran}} = \rho RT\sqrt{\frac{RT}{2\pi}}[(S^2 + 2)e^{-(S\theta)^2} + \sqrt{\pi}(S^2 + \tfrac{5}{2})(S\theta)[1 + \mathrm{erf}(S\theta)]], \quad (9a)$$

$$E_{i,\mathrm{int}} = \frac{\rho(5 - 3\gamma)}{\sqrt{\pi}(\gamma - 1)}\left(\frac{RT}{2}\right)^{3/2} \times \{e^{-(S\theta)^2} + [1 + \mathrm{erf}(S\theta)]\}. \quad (9b)$$

2.3.2 Momentum- and Energy-Transfer Coefficients Our next task is to develop expressions for the fluxes of momentum and energy that leave the surface through molecular scattering. At this point we encounter one of the least understood and most speculative areas of rarefied-gas dynamics, since our knowledge of these exit fluxes is extremely limited. Few molecular exit-velocity distributions have been obtained from laboratory experiment and none at all from observations aboard orbiting satellites. Much more is known about angular distributions of number flux, particularly in connection with laboratory studies of scattering from clean, regular surfaces at energies of thermal order. While speculations based upon reasonable conjectures can be made in light of present knowledge, there are insufficient experimental data to permit accurate performance predictions for extended shapes. A short account of our present knowledge of wall–gas interactions is given in Sec. 2.5.

The standard kinetic theory does, however, include useful formulas for drag, lift, and heat transfer for bodies in free-molecule flow, which treat the effects of gas collisions at the wall in terms of momentum- and energy-transfer coefficients. These coefficients, in essence, are measures of the degree to which the net reflected momentum or energy approaches that which would be emitted in Maxwellian equilibrium at the wall temperature. These coefficients can be connected with Maxwell's nonphysical conjecture dividing reflection into two kinds, specular and diffuse, but that is not their meaning, although some confusion persists. In the current and standard usage, thermal energy transfer is discussed in terms of α, the thermal accommodation coefficient, following von Smoluchowski (1911) and Knudsen

(1911), where

$$\alpha = (E_i - E_r)/(E_i - E_w). \tag{10a}$$

The subscript i identifies the energy carried to the surface by incident molecules, the subscript r identifies the energy carried away from the surface, and the subscript w identifies the energy that would be carried away by molecules in diffuse Maxwellian scattering.

In similar fashion we define σ, the transfer coefficient for tangential momentum, as

$$\sigma = (\tau_i - \tau_r)/(\tau_i - \tau_w), \tag{10b}$$

in which τ_w represents the tangential momentum carried from the wall by molecules scattered in Maxwellian reemission from it. However, these molecules are scattered diffusely and hence isotropically, in which event the value of τ_w is 0.

The coefficient of normal momentum transfer, σ', was introduced by Schaaf to complete the triad of surface interaction coefficients. It is seen as an essential feature in the current analysis of aerodynamic forces:

$$\sigma' = (p_i - p_r)/(p_i - p_w). \tag{10c}$$

The energy carried from the wall in Maxwellian reemission is given by

$$E_w = (4 + j)mRT_w \frac{dN_i}{dA} = \frac{\gamma + 1}{2(\gamma - 1)} mRT_w \frac{dN_i}{dA}. \tag{11}$$

We write the scattered normal momentum flux p_w in equation (10c) as

$$p_w = \tfrac{1}{2}m\sqrt{2\pi RT_w}\frac{dN_i}{dA}. \tag{12}$$

The coefficients σ, σ', and α are, in principle, independently determinable by physical measurement in particular physical systems. Through these empirical coefficients one may express the net shear stress τ, or the net pressure p, or the heat flux Q, acting on a surface, since these coefficients link the incoming with the outgoing quantities.

The quantities p_i and p_r are added to form the total normal pressure, giving

$$\begin{aligned} p = \frac{\rho U^2}{2S^2}\Bigg\{&\left(\frac{(2 - \sigma')}{\sqrt{\pi}}(S\theta) + \frac{\sigma'}{2}\sqrt{\frac{T_w}{T}}\right)e^{-(S\theta)^2} \\ &+ \Bigg[(2 - \sigma')[(S\theta)^2 + 0.5] \\ &+ \frac{\sigma'}{2}\sqrt{\frac{\pi T_w}{T}}(S\theta)\Bigg][1 + \operatorname{erf}(S\theta)]\Bigg\} \end{aligned} \tag{13}$$

and in a similar manner, letting $\chi(S\theta) = \exp[-(S\theta)^2] + \sqrt{\pi}S\theta[1 + \operatorname{erf}(S\theta)]$, we find

$$\tau = \frac{\sigma\rho U^2 \cos\theta}{2S\pi}\chi(S\theta) \tag{14}$$

and

$$\begin{aligned} Q = \alpha\rho U^2 RT\sqrt{\frac{RT}{2\pi}}\Bigg\{&\left(S^2 + \frac{\gamma}{\gamma + 1}\right. \\ &\left. - \frac{\gamma + 1}{2(\gamma - 1)}\frac{T_w}{T}\right)\chi(S\theta) - 0.5e^{-(S\theta)^2}\Bigg\}. \end{aligned} \tag{15}$$

Expressions (13), (14), and (15) provide the essential key to free-molecule analysis. It should be noted that each expression depends solely upon a single transfer coefficient, thus permitting a clear separation of effects in experimental studies. Both the pressure and the net heat flux, but not the shear stress, depend upon the wall temperature. No complication arises where T_w is known, but where it is not it may be necessary to solve for T_w with Q set equal to zero. Terms for radiation exchange will be needed.

2.3.3 Forces of Lift and Drag on a Flat Plate, General Expressions Where the transfer coefficients are known, solutions to problems of external free-molecule aerodynamics are accessible in principle but may be difficult to solve in analytic form. In general the transfer coefficients will be dependent on the angle θ, measured between the incoming gas molecules and the surface tangent. This feature may add substantially to the complication of solution. A number of authors have calculated values of aerodynamic coefficients and heat transfer for various conventional shapes, but have treated the coefficients as independent of the incident angle. For cases of high-speed flows over vehicles in orbit, where it is probable that the elementary assumpical are inadequate, it seems advisable to treat complex shapes through numerical integration over elementary areas. To this end the expressions for drag and lift per unit area for a flat plate at angle of attack θ may prove useful. These are listed as follows:

$$\begin{aligned} F_D = \frac{\rho_\infty U_\infty^2}{\sqrt{\pi} S} A \Bigg\{ & [(2 - \sigma') \sin^2\theta \\ & + \sigma \cos^2\theta] e^{-(S \sin\theta)^2} + \frac{\sigma' \pi}{2} \sqrt{\frac{T_W}{T}} \sin^2\theta \\ & + \sqrt{\pi}(S \sin\theta) \operatorname{erf}(S \sin\theta)[(2 - \sigma') \\ & \times (\sin^2\theta + \tfrac{1}{2} S^2) + \sigma \cos^2\theta] \Bigg\}, \end{aligned} \tag{16}$$

$$\begin{aligned} F_L = \frac{\rho_\infty U_\infty^2}{\sqrt{\pi} S} A \Bigg\{ & (2 - \sigma - \sigma') \cos\theta \sin\theta \, e^{-(S \sin\theta)^2} \\ & + \frac{\sigma' \pi}{2} \sqrt{\frac{T_W}{T}} \sin\theta \cos\theta \\ & + \frac{\sqrt{\pi}}{2} S \cos\theta \Bigg[2(2 - \sigma - \sigma') \sin^2\theta \\ & + \frac{(2 - \sigma')}{S^2} \Bigg] \operatorname{erf}(S \sin\theta) \Bigg\}. \end{aligned} \tag{17}$$

2.4 Direct Molecular Flow Simulation using Monte Carlo Methods

During the late 1950s scholars turned to the examination of computational models that exploited directly the dynamics of molecular motion and collisions and that left aside immediate concern with time-varying fluid behavior. Such directions were made possible by the increasing capability of high-speed computers. Alder and Wainwright (1958) used exact dynamical calculation to study the behavior of one hundred hard spheres. These were started into motion with uniform speeds and with randomized directions within a cubical enclosure. Velocities were found to approach the Maxwellian distribution following a few collisions. It was recognized, however, that the approach would quickly impose unrealistically large demands on computer capability as the number of molecules increased. Davis (1960), in what proved to be very promising directions, studied the free-molecule conductance of a tube having a right-angle bend. Randomized direction and position were given to each of a sequence of test particles started from the tube entrance. Trajectories were followed and statistical decisions were made for next directions of flight. These directions were based on diffuse scattering at the wall. The probability of success, as measured by the ratio of the number of molecules emerging to those starting, may be seen as a measure of the conductivity of the system. Haviland (1965) introduced the calculation of intermolecular collisions as an extension of the test-particle method in application to heat conduction at intermediate Knudsen numbers.

It was a natural step to move from these beginnings to the method of direct simulation of molecular dynamic processes using Monte Carlo methods (DSMC), a step that has most profoundly affected the landscape of rarefied-gas dynamics. The conceptual basis for the method and much of its operational detail are due to Professor Graeme Bird (1976). DSMC rests upon two critical assumptions. In the first of these Bird assumed that the calculation of collisional outcomes occurring within a very short period of time in a certain region of the flow may be decoupled from the positional changes of molecules occurring within the same period. It is also assumed that the outcomes of intermolecular collisions and molecule–wall collisions can be determined using statistical methods when appropriate physical parameters are given. A number of enabling assumptions are made. Molecules move within a field divided into a large number of cells, each containing 15 to 20 numbered particles; the motions and collisions of these are followed. Each particle represents a very large number of "real" gas molecules. The cell dimension is small as compared with the molecular mean free path, so that only small changes in gas properties can occur within the cell. The motion of the particle is followed through successive time intervals Δt_m, each of which is followed by calculations of collisional outcomes. The basic time interval is smaller than the mean collision time per molecule. Collision partners are selected from within the cell. Since the change in gas properties is small over cell dimensions, the exact coordinates of collision partners are immaterial, each partner being representative of molecules anywhere within the cell. A cell-by-cell record of densities and velocities is maintained for the later calculation of flow properties. The method of DSMC is developing swiftly. Numerous comparisons have been made between the results of experiment and the predictions of DSMC, leading now to general acceptance of molecular flow simulation as the mode of solution for rarefied-gas problems. Calculations incorporating chemical reactions and electron production are often made. Recent work, notably by Harvey (1989) and later by Boyd (1991) and others, has focused on the examination of phenomenological methods to guide calculation for inelastic collisions. As a result of these several advances, it would appear that some of the problems that had imposed an apparently impenetrable barrier to the advancement of understanding and application have been removed.

It is useful to consider again the determinations of temperature and density distributions within a shock wave. Experiments by Sherman (1955) and by Talbot and Sherman (1959) were conducted in the Berkeley tunnel using a free-molecule temperature probe. Argon was used in this work, in which the flow Mach numbers were low, always below 1.61. Experimental results were compared with Navier-Stokes solutions and also, as permitted, those of the Burnett equations. In addition comparisons were made with solutions to the 13-moment equations of Grad (1949) and with those from the two-temperature equations of Mott-Smith (1954). Experimental profiles of temperature compared well with density profiles from Navier-Stokes and with those from the Burnett equations (1935) at these low Mach numbers but not with those from the 13-moment solutions nor those from the two-temperature model. In subsequent work Schmidt (1969) used electron-beam fluorescence (EBF) to measure density profiles for Mach-8 flows of argon. At a later date remarkable agreement was found between Schmidt's experimental results and those from DSMC (Bird, 1976). Robben and Talbot (1965) using EBF-measured rotational temperature and density profiles for a shock wave of nitrogen at Mach 1.71. Again, results obtained by Bird (1976) using DSMC were found to be in excellent agreement with experiment. Much more recently, Lumpkin and Boyd (1993) examined flow properties for a detatched shock resulting from high–Mach-number air flow over a half-sphere. Comparisons of DSMC simulations with results of a Navier-Stokes–based computation agree well except in prediction of the flow angle at the rim, DSMC showing a much greater deflection of flow toward the central axis than given by Navier-Stokes. The Burnett equations did not yield stable solutions. Clearly, physical experiment is now required.

The foregoing material leads to the following important observations. Rarefied-gas experimentation can provide information concerning dynamic processes in high-speed flows. Where boundary conditions are simple, as, for example, for a plane shock held at the entrance to a shock holder, and the Knudsen number is small enough, Navier-Stokes solutions model the physical process fairly well. On the other hand, Monte Carlo simulations, which are also good in these cases, are not limited to low–Knudsen-number flows and accommodate more readily to complex geometries. With DSMC, as is also the case for Navier-Stokes, we are not always sure of the values of essential parameters or of the validity of physical assumptions. Physical experimentation is needed, to validate codes and to assist in the determination of essential parameters.

2.5 Gas–Surface Interactions

Collisions of gas molecules with walls occur in gas flows of any density, but at normal densities the frequency of collisions in the stream greatly exceeds that for collisions of gas molecules with solid surfaces. Under conditions of low Knudsen number, where knowledge of velocity slip and temperature jump is required, the determination of these quantities has been formulated in terms of a coefficient σ for velocity slip and α for temperature jump. In the context of the conventional formulation, σ is defined as the number of molecules diffusely reflected per unit time and surface area, $1 - \sigma$ is the number specularly reflected, while α is the standard thermal accommodation coefficient. Both α and σ are assumed equal to 1 in the absence of well-documented information.

At larger Knudsen numbers, wall–gas collisions become relatively more numerous until, in the limit of free-molecule flow, only wall–gas collisions remain. Throughout the rarefied regimes it is critically important to incorporate wall–gas collision data into computations. Where free-molecule assumptions can be made, a knowledge of the transfer coefficients, Eqs. (1a)–(1c), will permit the calculation of forces and heat transfer. In all other regimes it is essential to know the exit distributions of velocity, that is, to know the statistical description of postcollision velocities, in order to proceed in computation with DSMC.

In current conventional computions, exit distributions follow directly from the assumption that the scattering is diffuse or that it is specular. This nonphysical approximation is used in simulations of high-speed flows simply because there has been no generally accepted alternative in the aerospace context. Accordingly, it is useful to examine the understanding that has come to us from physical experiment and to look, too, at models for exit distributions that are currently being studied.

The first studies of thermal accommodation in wall–gas collisions were made using the loss of heat from a central wire through molecular transport to a cylindrical enclosure. The system could be freed of adsorbed gas molecules by baking under high vacuum. Much was learned from these experiments regarding the formation of adsorbed gas films, and values of thermal accommodation coefficients were obtained for gases of many sorts on various surfaces, both clean and with gas layers. The method was brought to its present level of development through the work over many years of Thomas (1981) and associates. Almost all gas–surface combinations exhibited incomplete thermal accommodation, i.e., $\alpha < 1$, for clean surfaces, the lowest values being found for helium in contact with clean tungsten. Values of α close to unity were found for complex molecules interacting with gas-covered surfaces. Values of α ranged from 0.0164 for helium on tungsten to 0.89–0.97 for air on etched aluminum. It seems safe to conclude that for many gas–surface combinations, some degree of incomplete thermal accommodation will be found.

Somewhat similar results were obtained in experimental studies of the tangential momentum-transfer coefficient. These studies were usually conducted in rotating-cylinder devices or by the measurement of flow through tubes. Values of σ were found to lie near to but often less than 1.0, although very infrequently, values have been found to exceed 1.0. There is no anomaly here; the result can be produced by backscattering due to roughness. Representative values for σ and for α are shown in Tables 1 and 2.

Table 1. Representative values of the tangential momentum-transfer coefficient σ, obtained from rotating-cylinder experiments.

Gas	Surface	σ	Ref.
Air	Machined brass	1.00	a
Air	Oil	0.90	a
Air	Glass	0.89	a
Air	Oil film on Al	0.90	b
N_2	Air on oil film on Al	0.8–0.93	c

Investigators: a. Millikan (1923). b. Chiang (1951). c. Hurlbut (1960).

Measurements using heat-conductivity cells or rotating cylinders, as in the foregoing, do not yield velocity or flux distributions of scattered molecules. On the other hand, from the earliest experiments using molecular-beam methods, it had been observed that distributions of scattered molecular flux were often of strikingly lobal form, although for many systems the flux patterns followed the cosine distribution to close approximation. In cosine scattering the flux of molecules scattered from the point of impingement of a collimated beam into a small element of solid angle above the surface is proportional to the cosine of θ, where θ is the angle from the surface normal to the center of the element of solid angle. Once the Kantrowitz–

Table 2. Representative values of the thermal accommodation coefficient α, obtained from measurement in thermal conductivity cells.

Gas	Surface	α	Ref.
N_2	Pt	0.50	a
N_2	W	0.35	a
Air	Machined bronze	0.89–0.93	b
Air	Machined cast iron	0.87–0.88	b
He	W 30 °C	0.0164	c
Ne	W 30 °C	0.0412	c
Ar	W 30 °C	0.271	c
Xe	W 30 °C	0.773	c
He	H_2 adsorbed on W	0.041	d
He	N_2 adsorbed on W	0.064	d
He	W, clean	0.02	d
CO_2	CO_2 adsorbed on W	0.990	d

Investigators: a. Oliver (1950). b. Wiedmann and Trumpler (1946). c. Silvernail (1954). d. Wachman (1957).

Gray sources came into common use, it was soon found that fast beams of heavier molecules could be produced by the expansion of a mixture of light and heavy species. The heavier species in beams skimmed from such mixture expansions might reach kinetic energies of one or more electron volts. The production of such "seeded" beams became standard in the art and is now used by molecular physicists and physical chemists in many research applications.

Molecular-beam methods have been used in a number of studies of the scattering of common gases from many kinds of surface including regular crystals, machined surfaces, and foils. The observations of Hinchin and Sheppard (1967) on the angular density distributions of argon scattered from platinum foil surfaces are represented in part in Fig. 2. One may see in these results certain patterns that are commonly observed and that merit discussion. It is first apparent that all distributions in this instance exhibit a strong lobality and show a remarkable regularity in the location of the lobe maxima. These positions are related to the beam reservoir temperature so that the most energetic beams scatter in patterns nearest to the surface tangent. Scattering patterns such as these, and in fact other scattering patterns that show a departure from the pure diffuse or cosine pattern, imply that the energy- and momentum-transfer coefficients also must depart from the value 1. The scattering from rough surfaces or those heavily laden with adsorbed gas layers is most often diffuse. One may summarize the results of much observation by noting that patterns of molecular scattering depend upon the nature of gas and surface, upon the surface smoothness, upon the cleanliness of the surface, and upon the energy of the gas molecule relative to the surface. Other factors may also be important if molecule internal energies are transferred or reactive events occur. Lobality, when present, tends to increase with surface cleanliness, with surface regularity, with incident velocity, and with a reduction in the angle of attack. Scattering patterns from suitably degassed and adequately crystalline surfaces are, with few exceptions, of distorted elliptical form having maxima near the exit specular angle.

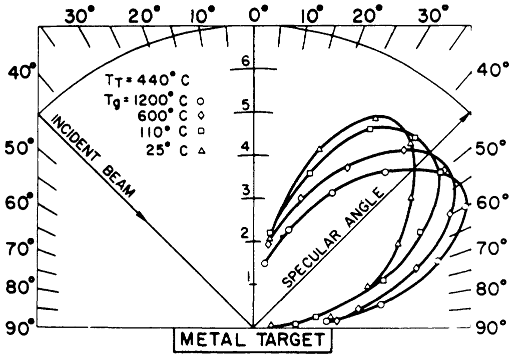

FIG. 2. Scattering of argon from degassed platinum surface (Hinchen and Shepherd, 1967).

Technologic change again intervened to make possible the measurement of distributions of velocity for the scattered molecular flux. Results from two of the relatively few studies completed are shown in Fig. 3. Both studies, Subbarao and Miller (1973) and Jih and Hurlbut (1977), employed argon beams of energy up to 1.3 eV on (111) surfaces of silver. Jih and Hurlbut observed exit velocities increasing from directions near to the surface normal to those near the surface tangent except for a pronounced dip in the region of the density peak. Subbarao and Miller, measuring only in the vicinity of the peak, observed a less pronounced dip. At higher energies, Tenner *et al.* (1986) found calculated energy distributions for 35-eV potassium ions scattered from the tungsten (110) surface to reproduce remarkably the observed experimental distributions in density and energy of exit particles. The reader is referred to surveys by Hurlbut (1986, 1989), for more detailed reviews of these and related observations.

We are also instructed by measurements of the transfer of momentum to common surfaces from high-speed beams of neutral molecules or of ions. Studies by Doughty and Schaetzle (1969) using neutral argon or nitrogen molecules of 25-eV energy on surfaces of aluminum or fresh varnish gave values for the normal momentum-transfer coefficient that decreased from values of 0.9 or smaller at normal incidence to values of 0.5 or smaller at incidence angles of 60° from the normal. Knechtel and Pitts (1969), using nitrogen ion beams incident on aluminum surfaces at 5 to 15 eV, observed coefficients for normal momentum transfer over the same range of incident angles that differed

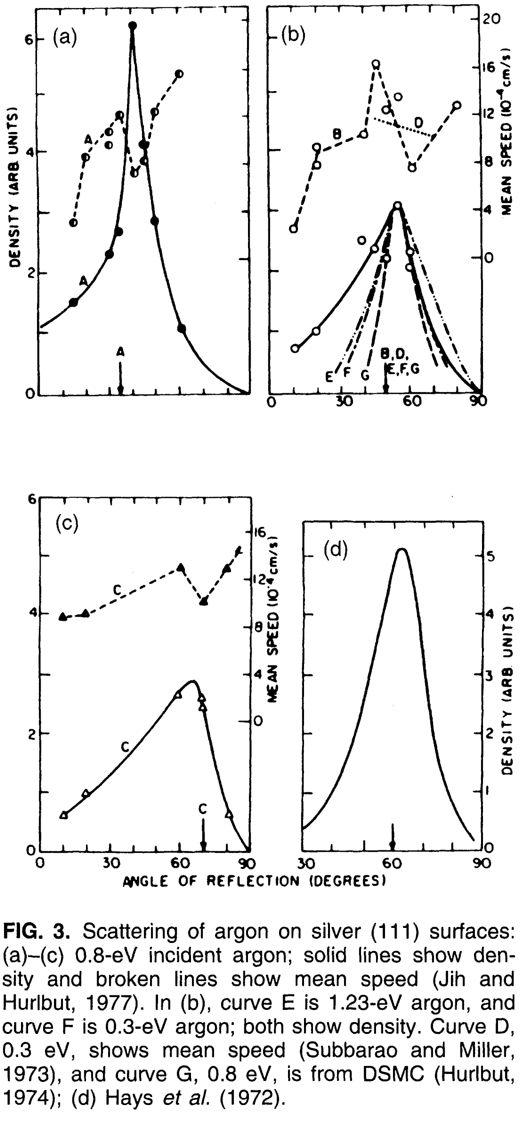

FIG. 3. Scattering of argon on silver (111) surfaces: (a)–(c) 0.8-eV incident argon; solid lines show density and broken lines show mean speed (Jih and Hurlbut, 1977). In (b), curve E is 1.23-eV argon, and curve F is 0.3-eV argon; both show density. Curve D, 0.3 eV, shows mean speed (Subbarao and Miller, 1973), and curve G, 0.8 eV, is from DSMC (Hurlbut, 1974); (d) Hays *et al.* (1972).

slightly in numerical value, as one might expect, but were qualitatively very similar to the observations of Doughty and Schaetzle. These slender fragments of data suggest very strongly that much more experimental observation is needed, some of it in the orbital environment, before one can predict with confidence in rarefied regimes.

Data from detailed measurements of velocity distributions for scattered particles must be parametrized or modeled in order to make them accessible for use in DSMC. Two general models are identified here as having merit, since each is based on its own substantive technical premise. The first derives from a conjecture due to Nocilla (1961) that lobal exit distributions could be modeled by use of suitable drifting Maxwellians consisting of a macroscopic velocity upon which a random thermal velocity is superimposed. The macroscopic velocity vector makes an angle θ_r with respect to the surface normal. Subsequently, molecular-beam experiments have shown the best representation of reflected distributions of velocity along observed directions to be the drifting Maxwellian. Hurlbut and Sherman (1968) used Nocilla's conjecture to develop a model in which parameters for the exit Maxwellian distribution are connected with those for the incident molecular flow. For a given system, experimental values of the thermal accommodation coefficient are required and a functional relation between the angle θ_i and the magnitude of the exit molecular speed ratio S_r is also required. The model is easily applied in DSMC calculations and has been used by a number of authors.

The second of these general models is based on the conjecture that the principle of reciprocity must apply between incident and exit velocity distributions in scattering processes if thermal equilibrium is to be preserved. It is further asserted that the same exit distributions found for a given set of incident conditions should correctly apply to the nonequilibrium flow. Detailed formulation and application to surface scattering appears in Cercignani and Lampis (1971). Formulas for flux distributions and for aerodynamic forces have been given in subsequent work. Both models have created substantial interest, but to the time of the present writing, neither is fully explored.

3. EXPERIMENTAL METHODS

3.1 Ground-Based Experimental Systems

Experimental systems used in the study of rarefied-gas aerodynamics and of related molecular processes are of four major types: wind tunnels, shock tunnels, molecular- and ion-beam systems, and ballistic ranges. There are also specialized systems developed for the study of momentum and energy transfers.

3.1.1 Wind Tunnels and Shock Tunnels Rarefied-gas wind tunnels are designed to produce low-density, high-speed flows of gas through a diagnostic region in which models may be placed for test. Low-density flows are frequently produced by oil-vapor booster pumps backed by lobe-type blowers that are exhausted to the atmosphere by rotary vacuum pumps. A few tunnels have been cryogenically pumped, some have been pumped by mechanical systems alone, and some have been pumped by steam-jet ejectors. All are characterized by offering an extended period of continuous testing. They are usually of large size and serve as the test bed for a wide variety of experiments. From the beginnings of orbital flight, however, it was acknowledged that limitations imposed on the magnitude of the stagnation enthalpy would make full flight simulation impossible. This notwithstanding, rarefied-gas wind tunnels are valuable research tools.

Experiments may be conducted over a wide range of flow conditions, permitting the comparison of flows in highly rarefied regimes with those in regimes of smaller Knudsen number. Long-duration flows are of particular advantage in studies involving the detection of low-level electrical or optical signals.

Many features of tunnel construction, modes of operation, instrumentation, force balance design, etc., may be found in Maslach and Schaaf (1959). Accounts of more recent design and practice may be found in the RGD Symposium Volumes 9 (1974) through 12 (1980).

Shock tunnels are, in essence, wind tunnels that derive reservoir supply gas at high temperatures from a shock tube and yield flows of substantially larger stagnation enthalpy than can be obtained in steady-state devices. While stagnation temperatures can reach several thousand kelvins, the flow duration is short, typically a few milliseconds.

3.1.2 Flow Production In continuous-flow wind tunnels and in shock tunnels, gas flows from the reservoir through a converging region and then emerges through a throat into a nozzle. Following expansion in the nozzle the flow passes into the test section. Alternatively, gas emerges from the throat directly into a test section of size sufficient to permit free-jet expansion. Such expansions contain a region of supersonic flow that is terminated by a plane shock and is bounded at its circumference by a shock barrel, as we see in the schematic, Fig. 4. Whereas the analysis of flow in nozzles is taken up in standard texts, the most satisfactory guide to free-jet expansions remains the original paper by Ashkenas and Sherman (1966).

It is useful to consider certain characteristic features of free-jet expansions. The gas, which is at sonic speed as it leaves the throat, rapidly expands to supersonic velocities along straight flow lines. These diverge from a single point of origin, as shown. Properties of the flow at any value of the nondimensional distance x/D may be fully characterized in terms of γ, the ratio of specific heats, the stagnation temperature T_0, and the ratio of the stagnation pressure p_0 to the downstream tunnel pressure p. As the Mach number increases with larger values of x/D, the local density and temperature of the flow decrease and the Knudsen number, based on the size of a diagnostic probe, can become large. Rarefied conditions are easily achieved, the flow properties can be calculated for any molecular species, and every region of the flow field is physically and optically accessible. For some purposes, these advantages are offset by the divergence of the flow.

3.1.3 Molecular and Ion Beams Molecular-beam systems are powerful tools for the study of molecular collision processes and for the study of gas collisions at the surface. A collimated stream of molecules traverses a region of high vacuum in which it can intersect a second beam or impinge on a solid test surface. Measurements of densities,

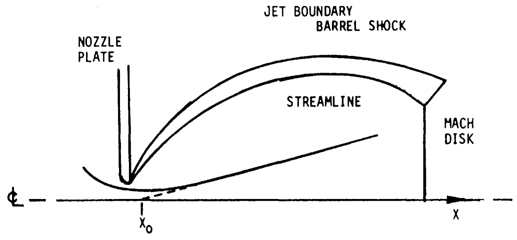

FIG. 4. Schematic diagram for free-jet expansion. (See Ashkenas and Sherman, 1966).

of composition, and of distributions of velocity and of electronic state may be obtained. In most current systems beams are formed by "skimming" from the rarefied region of a free-jet expansion, a method introduced by Kantrowitz and Gray (1951). The skimmer is simply a hollow conical shell with a small sharp-edged hole at the tip. Flow external to the skimmer is diverted to a large pump, while the skimmed beam passes into and through a high-vacuum region and thence to the target. Beam flux is higher than that from effusive sources, and because of the mechanics of expansion, the mean speed of the beam is higher and the thermal spread of velocities is much smaller. These features offer critical opportunities.

To reach beam energies above a very few electron volts it is necessary to accelerate ionized molecules. These may be used directly, or as neutral molecules following neutralization by charge exchange. Space charge limits the beam current to very low values at beam energies of the order of 10 eV. Unfortunately, energies of approximately this value are of particular interest to the experimenter interested in the effects of gases at the velocities encountered by satellites.

Beam systems are designed for operation under conditions of very high vacuum through careful attention to materials, welds, seals, and pumping equipment. A typical experiment might require scattered particles to be velocity selected, then ionized and accelerated into an electron multiplier detector. The reader is referred to the RGD Symposium proceedings for accounts of developments and applications of beams in rarefied-gas research.

3.2 Instruments and Diagnostic Methods

Wind tunnels and other experimental systems for the study of rarefied gases are, in essence, vacuum systems that must be equipped with the standard instrumentation for vacuum measurement. Balance systems are required that are suitable for the measurement of aerodynamic forces of perhaps a few hundredths of a newton. Advantage is often taken of the relatively large mean free path to position noninvasive physical probes in the field of flow. This point is illustrated by the pioneering measurement of shock-wave profiles by Sherman (1955) using a free-molecule thermocouple probe, and at an earlier time by Stalder *et al.* (1950), who measured the recovery temperature of a thermocouple probe in free-molecule flow.

Most of today's noninvasive measurement systems are based in some way upon the production and detection of light. The familiar methods of interferometry and of Schlieren flow visualization are not suitable in rarefied regimes because of insufficient optical depth. However, certain methods that employ the emission or absorption of light through the alteration of molecular electronic states have proved to be very useful. The most prominent of these methods has used an energetic electron beam to excite fluorescence radiation from well-defined regions of a flowing gas. A slender electron beam of 10- to 30-keV energy and a few milliamperes of current excites fluorescence radiation specific to the species and its state. Collection optics, focused on a portion of the electron beam, connect through a fiber optic to analyzing equipment. Detection is usually accomplished by photomultiplier and pulse counter or electrometer. Spectroscopic data from molecular gases, frequently N_2, lead to density measurement and to determinations of rotational temperatures and vibrational states. Atomic fluorescence radiation is used in density determinations, and, with the help of very high-resolution instruments such as the Fabry–Pérot interferometer, translational velocities and translational temperatures can be determined through the measurement of Doppler shifts. The method of electron-beam–induced fluorescence (EBF) in its various aspects is largely to be credited to its pioneering development by Muntz (1966) and to his continuing productive research.

During the past decade the quality and availability of tunable lasers have facilitated more general use of these systems in diagnostic applications. A valuable survey due to Lewis (1989) compares the utility of EBF, laser-induced fluorescence (LIF), Rayleigh scattering, and Raman scattering (RRS and VRS), and rates the relative difficulty of measurement over a wide range of gas densities. A bar graph from the Lewis paper is reproduced in Fig. 5. In the production of fluorescence, changes in electronic state are precisely determined by the controlled laser wavelength. Experience has shown that the laser beam is more easily focused into the

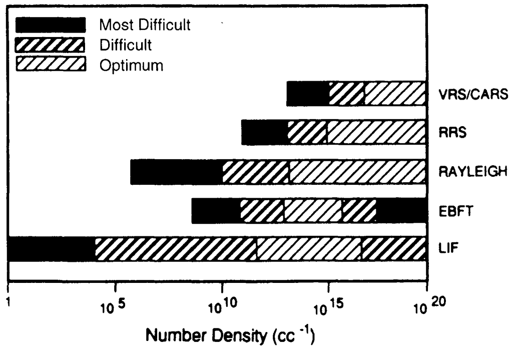

FIG. 5. Optical diagnostic comparisons (Lewis, 1989).

flow and is more stable in operation than the electron beam. Moreover, laser excitation avoids side effects due to secondary electrons, which compromise the effectiveness of EBF under some circumstances. One sees that the range of usable densities for LIF is much greater than for EBF. These advantages must be weighed against the potential for economical exploratory investigation using the electron beam as a broad-spectrum exciter.

3.3 Rocket and Satellite Experimental Systems

Once rockets were sufficiently developed to carry instruments to extreme altitudes, explorations of the thermosphere were conducted to measure atmospheric density and composition, charged-particle concentration, solar energy flux, and so on. These studies by sounding rocket have been of the greatest importance to the development of atmospheric models and to the understanding of atmospheric processes. Orbiting satellites, the next level of development, soon served as valuable experimental platforms. These vehicles may be instrumented for mass spectrometry, optical spectrometry, the detection of charged particles of various kinds, or the detection of electromagnetic fields, and with high-resolution systems for Earth viewing. Issues of astrophysical interest have largely displaced studies of the Earth's atmosphere or the gas dynamics of vehicle interactions with the flow, in part because of the altitude required for long-lived orbits. It should be noted, however, that the understanding of rarefied-gas dynamics always is important in the design of the satellite and its instrumentation.

In the late 1980s the tethered satellite became increasingly discussed as a potential research tool for atmospheric and aerodynamic studies. A research satellite, tethered to a parent satellite by a small-diameter line, may be deployed perhaps 100 km Earthward from the parent vehicle, where it will find an appropriate environment for the study of thermal and aerodynamic effects on the vehicle. The potential for aerodynamic testing using an orbiting "wind tunnel" is very attractive. By the date of this writing, two tethered satellites have been launched, the second demonstrating fully the feasibility. Dynamic and aerodynamic analysis are in progress. For further information relating to applications in space sciences and to related materials, the reader is referred to the *Journal of Geophysical Research* (see, e.g., Jacchia *et al.*, 1977, on atmospheric modeling) and to recent volumes of the *Journal of Spacecraft and Rockets.*

4. APPLICATIONS

The world's technologic life has been extraordinarily transformed through the effects of our knowledge of the behavior of rarefied gases. The influences are pervasive, touching our lives at every moment. Consider, for example, this short list of industrial processes that derive help from vacuum science and engineering: vacuum film deposition, chip manufacture, ion implantation, the production of vacuum electronic devices from x-ray sources to klystrons, diagnostic industrial control. One must also consider the essential aids to laboratory research, including electron microscopes, spectrometers, mass spectrographs, particle accelerators, and lasers.

Probably the greatest change of our times has come from the growth in our facilities for communication. This change, with its uncountable societal consequences, must be attributed to the rise of Sputnik and to the subsequent development of a vast communication umbrella based upon satellites in low Earth orbit. Every part of the technical and operational background for this enterprise rests upon aspects of rarefied-gas dynamics, just as will our continuing exploration of the Earth and its solar planetary environment.

At present much of the focus of aerospace work is on thermodynamic and thermochemical processes in very high-speed flows. This work is motivated by a need for more efficient space transportation. Specific emphasis is placed in two areas, the development of the aerospace plane and of the lifting aeropass and aerobraking shapes for reentry or orbit changes. The aerospace plane will depend upon airbreathing propulsion for much or all of its ascent and requires much new research before this concept of propulsion wherein the vehicle is an integral part of the propulsion system can be implemented. The aerobraking, lifting configurations also require much new research in order to resolve the many questions of chemistry, heat transfer, and wake flow that surround these new shapes at very high speeds and in unfamiliar atmospheres.

Both of the foregoing examples involve vehicles that will experience the full range of flow conditions from continuum to free molecule. There are in these efforts fascinating technical challenges; meeting these will result in new discoveries, and create new knowledge the impact of which cannot be predicted.

GLOSSARY

Cosine Law: Refers to molecular scattering in which the number flux of molecules scattered from a small element of surface into a small solid angle at θ with respect to the surface normal is proportional to the cosine of that angle.

Degassing: In vacuum systems, the process of removing adsorbed gas films on surfaces by heating to elevated temperatures.

Dilute Gases: A model of discrete particle gases in which only binary collisions occur.

Direct Simulation Monte Carlo Method (DSMC): A method for the simulation of the flow of rarefied gases in which the collisions and trajectories of representative particles are treated using statistical techniques.

Free-Jet Expansion: The expansion to supersonic speed of a gas issuing from an orifice into a suitably large region maintained at low pressure.

Free-Molecule Flow: Flow of very dilute gases in which the number of intermolecular collisions within or near to the system is negligibly small.

Knudsen Number: The ratio of the mean distance between molecular collisions to a characteristic length of the system. In external flows the characteristic length may be a diameter, but the choice is particular to the system.

Noninvasive Diagnostic Methods: Methods, usually optical, for determining the properties of flowing gases without altering the flow field.

Normal Momentum-Transfer Coefficient: Defined in Eq. (10c). Characterizes the effectiveness of the wall in absorbing the normal component of momentum carried to it by the incident molecular flux.

Phenomenological Methods: In the present context, statistical methods for selecting the transformation of excited states in molecular collisions from a range of alternatives.

Rarefied Gases: Dilute gases in which the ratio of the mean distance between molecular collisions to a characteristic length of the system lies between 0.01 and 100, approximately.

Specular and Diffuse Molecular Scattering: By specular scattering of molecules at surfaces it is implied that the full normal component of momentum is reversed and that the full component of tangential momentum is preserved. Diffuse scattering of molecules implies scattering according to the **Cosine Law.**

Tangential Momentum-Transfer Coefficient: Defined in Eq. (10b). Characterizes the effectiveness of the wall in absorbing the tangential component of momentum carried to it by the incident molecular flux.

Thermal Accommodation Coefficient: Defined in Eq. (10a). Characterizes the effectiveness of the wall in absorbing the thermal energy carried to it by the incident molecular flux.

Vehicle: Used in the aerospace context to refer to the entire rocket or satellite or other space system designed for flight in the upper atmosphere or beyond.

Works Cited

Alder, B. J., Wainwright, T. (1958), in: *Proceedings of the International Symposium on Transport Processes in Statistical Mechanics,* New York: Interscience, pp. 97–131.

Ashkenas, H., Sherman, F. S. (1966), "The Structure and Utilization of Supersonic Free Jets in Low Density Wind Tunnels," in: J. H. de Leeuw (Ed.), *Rarefied Gas Dynamics, Proceedings of the Fourth International Symposium,* New York: Academic Press, Vol 2, pp. 84–105.

Bhatnagar, P. L., Gross, E. P., Krook, M. (1954), "A Model for Collision Processes in Gases. I. Small Amplitude Processes in Charged and Neutral One-Component Systems." *Phys. Rev.* **94,** 511–525.

Bird, Graeme A. (1976), *Molecular Gas Dynamics,* Oxford: Clarendon Press.

Boltzmann, L. (1872), "Further Studies on the Thermal Equilibrium among Gas Molecules," *Wien. Ber.* **66,** 275.

Boyd, I. D. (1991), "Analysis of Vibrational-Translational Energy Transfer Using the Direct Simulation Monte Carlo Method," *Phys. Fluids A* **3,** 1785–1791.

Burnett, D. (1935), "The Distribution of Molecular Velocities and the Mean Motion in a Non-Uniform Gas," *Proc. London Math. Soc.* **40,** 382–400.

Cercignani, C., Lampis, M. (1971), "Kinetic Models for Gas-Surface Interactions," *Transport Theory Statist. Phys.* **1,** 101–114.

Chapman, S., Cowling, T. G. (1939), *The Mathematical Theory of Non-Uniform Gases,* 2nd ed, Cambridge, U.K.: Cambridge Univ. Press; (1952).

Chiang, S. F. (1951), Ph.D. Thesis, University of California, Berkeley.

Davis, D. H. (1960), "Monte Carlo Calculation of Molecular Flow through a Cylindrical Elbow and Pipes of Other Shapes," *J. Appl. Phys.* **31,** 1169–1176.

Devienne, F. M. (Ed.) (1960), *Rarefied Gas Dynamics; Proceedings of the First International Symposium,* Oxford: Pergamon.

Doughty, R. G., Schaetzle, W. J. (1969), "Experimental Determination of Momentum Accommodation Coefficients at Velocities up to and Exceeding Earth Escape Velocities," in: L. Trilling, H. Y. Wachman (Eds.), *Rarefied Gas Dynamics; Proceedings of the Sixth International Symposium,* New York: Academic, Vol. 2, pp. 1035–1054.

Dunoyer, L. (1911), *Compt. Rend.* **152,** 592–595.

Grad, H. (1949), "Lectures on the Kinetic Theory of Gases," *Commun. Pure Appl. Math.* **2,** 331–407.

Harvey, J. (1989), "Inelastic Collision Models for Monte Carlo Computation," in: E. P. Muntz, D. Weaver, D. Campbell (Eds.), *Rarefied Gas Dynamics; Proceedings of the 16th International Symposium,* Vol. 2, Progress in Astronautics and Aeronautics Vol. 117, Washington, DC: American Institute of Aeronautics and Astronautics, pp. 3–24.

Haviland, J. K. (1961), "Monte Carlo Application to Molecular Flows," MIT Dynamics Research Laboratory Report No. 61.5.

Hays, W. J., Rogers, W. E., Knuth, E. L. (1972), "Scattering of Argon Beams with Incident Energies up to 20 eV from a (111) Silver Surface," *J. Chem. Phys.* **56,** 1652–1657.

Hinchen, J. J., Shepherd, E. F. (1967), "Molecular Beam Scattering from Surfaces of Various Metals," in: C. L. Brundin (Ed.), *Rarefied Gas Dynamics; Proceedings of the 5th Symposium,* London: Academic, Vol. 1, pp. 239–252.

Hirschfelder, J. O., Curtis, C. F., Bird, R. B. (1954), *Molecular Theory of Gases and Liquids,* New York: Wiley.

Hurlbut, F. C. (1960), "The Influence of Pressure History on Momentum Transfer in Rarefied Gas Flows," *Phys. Fluids* **3,** 541–544.

Hurlbut, F. C. (1974), "Current Experiments and Open Questions in Gas-Surface Scattering," in: M. Becker, M. Fiebig (Eds.), *Rarefied Gas Dynamics; Proceedings of the 9th Symposium,* Porz-Wahn, Germany: DFVLR-Press, Vol. 2, pp. E-1–E-23.

Hurlbut, F. C. (1986), "Gas/Surface Scattering Models for Satellite Applications," in: J. N. Moss, C. D. Scott (Eds.), *Thermophysical Aspects of Re-Entry Flows,* Progress in Astronautics and Aeronautics Vol. 103, New York: American Institute of Aeronautics and Astronautics, pp. 97–119.

Hurlbut, F. C. (1989), "Particle Surface Interactions in the Orbital Context," in: E. P. Muntz, D. Weaver, D. Campbell (Eds.), *Rarefied Gas Dynamics; Proceedings of the 16th International Symposium,* Vol. 1, Progress in Astronautics and Aeronautics Vol. 116, Washington, DC: American Institute of Aeronautics and Astronautics, pp. 419–450.

Hurlbut, F. C., Sherman, F. S. (1968), "Applications of the Nocilla Wall Reflection Model to Free-Molecule Kinetic Theory," *Phys. Fluids* **11,** 486–496.

Jacchia, L. G., Slowey, J. W., Von Zahn, V. (1977), "Temperature, Density and Composition in the Disturbed Thermosphere from ESRO 4 Gas-Analyzer Measurements: A Global Model," *J. Geophys. Res.* **22,** 684–688.

Jih, C. T. R., Hurlbut, F. C. (1977), "Time of Flight Studies of Argon Beams Scattered from a Silver (111) Crystal Surface," in: L. Potter (Ed.), *Rarefied Gas Dynamics; Proceedings of the 10th International Symposium,* Progress in Astronautics and Aeronautics Vol. 51, New York: American Institute of Aeronautics and Astronautics, Part II, pp. 539–554.

Kantrowitz, A., Gray, J. (1951), "A High Intensity Source for the Molecular Beam, Part I," *Rev. Sci. Instrum.* **22,** 328–332.

Knechtel, E. D., Pitts, W. C. (1969), "Experimental Momentum Accommodation on Metal Surfaces of Ions near and above Earth-Satellite Speeds," in: L. Trilling, H. Y. Wachman (Eds.), *Rarefied Gas Dynamics; Proceedings of the 6th International Symposium,* Vol. 2, New York: Academic, pp. 1257–1266.

Knudsen, M. (1911), "Die Moleculäre Wärmleitungen der Gase und der Akkommodations-koeffizient," *Ann. Phys. (Leipzig)* **31,** 539–XXX.

Lewis, J. W. L. (1989), "Optical Diagnostics of Low Density Flow Fields," in: E. P. Muntz, D. Weaver, D. Campbell (Eds.), *Rarefied Gas Dynamics; Proceedings of the 16th International Symposium,* Vol. 1, Progress in Astronautics and Aeronautics Vol. 116, Washington, DC: American Institute of Aeronautics and Astronautics, pp. 107–132.

Lumpkin, F. E., Boyd, I. D., Venkatapathy, E. (1993), "Comparison of Continuum and Particle Simulations of Expanding Rarefied Flows," presented at 31st Aerospace Sciences Meeting, 11–14 January, AIAA Paper 93-0728.

Maslach, G. J., Schaaf, S. A. (1959), "Slip Flow Testing Techniques," *AGARDograph* 39, July.

Maxwell, J. C. (1867), "On the Dynamical Theory of Gases," *Philos. Trans. R. Soc.* **157,** 49. Reprinted in: W. D. Niven (Ed.), *The Scientific Papers of James Clerk Maxwell,* Vol. 2, Cambridge, U.K.: Cambridge Univ. Press, pp. 26–78.

Millikan, R. A. (1923), "Coefficients of Slip in Gases and the Law of Reflection of Molecules from Surfaces of Liquids and Solids," *Phys. Rev.* **20,** 217–238.

Mott-Smith, H. M. (1954), "A New Approach to the Kinetic Theory of Gases," MIT Lincoln Laboratory Group Report No. V-2.

Muntz, E. P. (1966), "The Direct Measurement of Velocity Distribution Functions," in: J. H. de Leeuw (Ed.), *Rarefied Gas Dynamics; Proceedings of the 4th International Symposium,* Vol. 2, New York: Academic, pp. 128–150.

Nocilla, S. (1961), "On the Interaction between Stream and Body in Free-Molecule Flow," in: L. Talbot (Ed.), *Rarefied Gas Dynamics; Proceedings of the 2nd International Symposium,* New York: Academic, pp. 169–208.

Oliver, R. N., Farber, M. (1950), Ph.D. Thesis, California Institute of Technology.

Robben, F., Talbot, L. (1965), "Measurement of Rotational Temperatures in Non-equilibrium Free-Jet Expansions of Nitrogen," *Phys. Fluids* **9,** 644–652.

Schaaf, S. A. (1963), "Mechanics of Rarefied Gases," in: S. Flügge (Ed.), *Encyclopedia of Physics,* Vol. VIII/2, Berlin: Springer, pp. 591–624.

Schaaf, S. A., Chambré, P. L. (1958), in: H. W. Emmons (Ed.), *Fundamentals of Gas Dynamics,* High Speed Aerodynamics and Jet Propulsion Series Vol. 3, Princeton, NJ: Princeton Univ. Press, pp. 687–741.

Schmidt, B. (1969), "Electron Beam Density Measurement in Shock Waves in Argon," *J. Fluid Mech.* **39,** Pt. 2, 361–373.

Sherman, F. S. (1955), "A Low Density Wind Tunnel Study of Shock Wave Structure and Relaxation Phenomena in Gases," NACA Technical Note 3298, Washington, DC: U.S. GPO.

Silvernail, W. L. (1954), Ph.D. Thesis, University of Missouri.

Stalder, J. R., Goodwin, G., Creager, M. O. (1950), "A Comparison of Theory and Experiment for High Speed Free-Molecule Flow," NACA Ames Aeronautical Laboratory Technical Note No. 2244.

Stern, O. (1920), "Eine direkte Messung der Thermischen Molekulargeschwindigkeit," *Z. Phys.* **2,** 49–56.

Subbarao, R. B., Miller, D. R. (1973), "Velocity Distribution Measurements of 0.06-1.4 eV Argon and Neon Atoms Scattered from the (111) Plane of a Silver Crystal," *J. Chem. Phys.* **58,** 5247–5257.

Talbot, L., Sherman, F. S. (1959), "Structure of Weak Shock Waves in a Monatomic Gas," NASA Memo 12-14-58W.

Tenner, A. D., Gillen, K. T., Horn, T. G. M., Los, J., Kleyn, A. W. (1986), "Energy and Angular Distribution for Scattering of K^+ from W (110) at Normal Incidence," *Surf. Sci.* **172,** 90–120.

Thomas, L. B. (1981), "Accommodation of Molecules on Controlled Surfaces—Experimental Developments at the University of Missouri, 1940–1980," in: S. Fisher (Ed.), *Rarefied Gas Dynamics; Proceedings of the 12th International Symposium,* Progress in Astronautics and Aeronautics Vol. 74, New York: American Institute of Aeronautics and Astronautics, Part I, pp. 83–108.

Vincenti, W. G., Kruger, C. H. (1965), *Introduction to Physical Gas Dynamics,* New York: Wiley.

von Smoluchowski, M. (1911), "Zur Theorie der Wärmeleitung in verfeinten Gasen und der dabei auftretenden Druckkräfte," *Ann. Phys. (Leipzig)* **35,** 983–1004.

Wachmann, Y. H. (1957), Ph.D. Thesis, University of Missouri.

Wiedmann, M. L., Trumpler, P. R. (1946), *Trans. Am. Soc. Mech. Eng.* **68,** 57–64.

Further Reading

Volumes of the International Symposia on Rarefied Gas Dynamics

Devienne, F. M. (Ed.) (1960), *Rarefied Gas Dynamics; Proceedings of the 1st International Symposium,* Paris: Pergamon.

Talbot, L. (Ed.) (1961), *Rarefied Gas Dynamics; Proceedings of the 2nd International Symposium,* New York: Academic.

Laurman, J. A. (1963), *Rarefied Gas Dynamics; Proceedings of the 3rd International Symposium,* Vols. 1 and 2, New York: Academic.

de Leeuw, J. H. (Ed.) (1965), *Rarefied Gas Dynamics; Proceedings of the 4th International Symposium,* Vols. 1 and 2, New York: Academic.

Brundin, C. L. (Ed.) (1967), *Rarefied Gas Dynamics; Proceedings of the 5th International Symposium,* Vols. 1 and 2, London: Academic.

Trilling, L., Wachman, H. Y. (1969), *Rarefied Gas Dynamics; Proceedings of the 6th International Symposium,* New York: Academic.

Dini, D., Cercignani, C., Nocilla, S. (Eds.)

(1971), *Rarefied Gas Dynamics; Proceedings of the 7th International Symposium,* Pisa: Editrice Tecnico Scientifica.

Karamcheti, K. (Ed.) (1974), *Rarefied Gas Dynamics; Proceedings of the 8th International Symposium,* New York: Academic.

Becker, M., Fiebig, M. (Eds.) (1974), *Rarefied Gas Dynamics; Proceedings of the 9th International Symposium,* Porz-Wahn, Germany: DFVLR-Press.

Potter, L. (Ed.) (1977), *Rarefied Gas Dynamics; Proceedings of the 10th International Symposium,* Progress in Astronautics and Aeronautics Vol. 51, Parts I and II, New York: American Institute of Aeronautics and Astronautics.

Campargue, R. (Ed.) (1979), *Rarefied Gas Dynamics; Proceedings of the 11th International Symposium,* Vols. 1 and 2, Paris: Commissariat à l'Energie Atomique.

Fisher, S. (Ed.) (1981), *Rarefied Gas Dynamics; Proceedings of the 12th International Symposium,* Progress in Astronautics and Aeronautics Vol. 74, Parts I and II, New York: American Institute of Aeronautics and Astronautics.

Belotserkovskii, O. M., Kogan, M. N., Kutateladze, C. S., Rebrov, A. K. (Eds.) (1985), *Rarefied Gas Dynamics; Proceedings of the 13th International Symposium,* New York: Plenum.

Ogucho, H. (Ed.) (1984), *Rarefied Gas Dynamics; Proceedings of the 14th International Symposium,* Tokyo: Univ. of Tokyo Press.

Boffi, V., Cercignani, C. (Eds.) (1986), *Rarefied Gas Dynamics; Proceedings of the 15th International Symposium,* Stuttgart: B. G. Teubner.

Muntz, E. P., Weaver, D., Campbell, D. (Eds.) (1989), *Rarefied Gas Dynamics; Proceedings of the 16th International Symposium,* Progress in Astronautics and Aeronautics Vols. 116, 117, and 118, Washington, DC: American Institute of Aeronautics and Astronautics.

Beylich, A. E. (Ed.) (1991), *Rarefied Gas Dynamics; Proceedings of the 17th International Symposium,* Weinheim: VCH.

Shizgal, B. (Ed.) (1994), *Rarefied Gas Dynamics; Proceedings of the 18th International Symposium,* Progress in Astronautics and Aeronautics Vols. 158–160, Washington, DC: American Institute of Aeronautics and Astronautics.

20
Photophysical and Photochemical Dynamics

Ottó Horváth and Kenneth L. Stevenson

20.1 **Introduction 719**
20.2 **Fundamental Principles of Photophysics and Photochemistry 719**
20.2.1 Light and Energy 719
20.2.2 Light-Absorption Principles 720
20.2.3 Quantum Yields 721
20.2.4 Quantum-Mechanical Principles of Absorption 721
20.2.5 Energy Partitioning and Dissipation Pathways 725
20.2.5.1 Photophysical Pathways: Luminescence and Nonradiative Decay 725
20.2.5.2 Photochemical Pathways 727
20.3 **Types of Photochemical and Photophysical Experiments 729**
20.3.1 Continuous (cw) Photolysis 729
20.3.2 Laser Flash Photolysis 730
20.3.3 Laser Spectroscopy 732
20.4 **Femtochemistry 733**
20.5 **Reactions of Excited Species 734**
20.5.1 Electronic Energy Transfer 734
20.5.2 Photochemistry of Organic Compounds 735
20.5.2.1 Electronic Transitions and Spectra of Organic Molecules 735
20.5.2.2 Correlation Rules and Symmetry Conservations 736
20.5.2.3 Intramolecular Processes 736
20.5.2.4 Intermolecular Processes 737
20.5.3 Photochemistry of Inorganic and Coordination Compounds 738
20.5.3.1 Ligand-Field (LF) Transitions 738
20.5.3.2 Ligand-to-Metal (LMCT) and Metal-to-Ligand (MLCT) Charge-Transfer Transitions 739
20.5.3.3 Charge-Transfer-to-Solvent (CTTS) Transitions 740
20.5.3.4 Other Charge-Transfer Transitions 740
20.6 **Conclusion 741**
Acknowledgments 741

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

Glossary 741
References 743
Further Reading 743

20.1
Introduction

The interaction of light with matter is of fundamental importance. Indeed, all of the energy we derive from fossil fuels was deposited over vast periods of time by photochemical processes, which converted energy from the sun into chemical energy. The study of chemical reactions induced by light (RIL) is usually termed *photochemistry*, whereas laser photochemistry (LP) refers to the study of reactions promoted with laser light. Photochemistry is concerned specifically with the ways that light can bring about changes in materials and has become very important in the investigation of chemical kinetics or chemical dynamics since photoinitiation is a much more selective and, in the case of the use of laser pulses, a much faster process than thermal initiation of reactions, providing greater control over the reactions. Moreover, lasers have proved invaluable not only to initiate reactions but also to analyze their products. Reaction intermediates, normally prepared thermally via precursors or collisions, can be formed directly by a selective photoprocess, permitting a vast array of reactive intermediates of known composition and energy to be available for new syntheses or experimental studies.

20.2
Fundamental Principles of Photophysics and Photochemistry

Changes in chemical structure caused by light are the subject of *photochemistry*, whereas light-induced changes in the ways electrons are distributed within a chemical entity while the atomic nuclei remain intact are the realm of *photophysics*. The initial event in any photochemical or photophysical process is the absorption of a photon by an atom, molecule, or complex ion, resulting in an electronically excited state. What happens after this event determines whether or not a photochemical reaction will occur, or whether some other photophysical process occurs such as fluorescence, phosphorescence, or the simple conversion of the light energy to heat. The following discussion describes the fundamental principles governing these processes.

20.2.1
Light and Energy

To understand the process of photolysis, that is, how light causes changes in

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

the structure and reactivity of chemical systems, it is necessary to describe some of the properties of light. On the one hand, it is convenient to think of light as an electromagnetic wave, summarized by the expression

$$\lambda = \frac{c}{\nu} \tag{20.1}$$

relating the wavelength, λ, to the frequency, ν, through the speed of propagation, c. On the other hand, light consists of quanta, or photons, each of which has energy given by

$$E = h\nu \tag{20.2}$$

where h, known as Planck's constant, can be interpreted as the proportionality constant between the energy and frequency of the wave associated with the photon.

Laser light has certain advantages over conventional, thermal light sources for use in photochemistry, such as high-power flux density, low angular spread, narrow frequency (or wavelength) distribution, and extremely short pulsewidths in pulsed lasers.

When chemical systems absorb light, individual molecules or atoms absorb only those photons with the right amount of energy to effect an excitation of the molecule from a lower to a higher energy level, if such a transition is allowed. Because light is an oscillating electric field, a transition is allowed for a normal absorption process only if the excited state has a different dipole moment than the ground state. Energy is distributed in molecules in the various levels of molecular motions, that is, translation, rotation, vibration, electronic, and so on, and the kind of electromagnetic radiation absorbed will depend on the kind of motion that is affected (see Chapter 11 by Lipson).

20.2.2
Light-Absorption Principles

The first law of photochemistry from Grotthus (1817) and Draper (1843) states that only the light that is absorbed by a molecule can produce photochemical change within the molecule. The amount of light absorbed in a homogeneous medium is defined by the Beer–Lambert law:

$$I = I_0 10^{-\varepsilon c l} \tag{20.3}$$

where I_0 and I are the light intensities at the front of the absorber and at distance l within the absorbing medium, respectively, and ε is the extinction coefficient in units consistent with the pathlength l (usually in centimeter) and the concentration c (usually in mole per liter). If one is to do quantitative continuous photolysis studies, it is important that an accurate value of the light absorbed by a photoactive sample can be measured since this determines the rate of the reaction. From Eq. (20.3) one can show that the absorbed light intensity, or radiant flux J, is given as

$$J = I_0(1 - 10^{-\varepsilon c l}) \tag{20.4}$$

Photochemists generally measure the incident light intensity, I_0, falling on a sample, and then use this equation to calculate J. The accurate measurement of light intensities, or actinometry, is very important in determining the photokinetics of the system. One convenient method for doing this is the photolysis of a chemical actinometer, which has been calibrated for the wavelength of the photoreaction under study. Almost any photoreaction system that has a well-defined quantum yield at the irradiating wavelength can be used as an actinometer, but for continuous photolysis work in the wavelength range 250–550 nm, the potassium trioxalatoferrate(III) system, developed by Hatchard and Parker (1956), is the most widely used and reliable actinometer available. For longer wavelengths to 750 nm, the photoaquation of aqueous Reinecke's salt, $KCr(NH_3)_2(NCS)_4$, works quite well (Wegner and Adamson, 1966). The common unit for J and I_0 in the aforementioned equations is einstein per second, where an einstein is 1 mol or 6.02×10^{23} photons. Occasionally it is expressed per unit volume of reactant, such as einstein per second per liter, analogous to the method for expressing rates of reactions.

20.2.3
Quantum Yields

The quantum yield or quantum efficiency of a photochemical reaction is normally defined as the fraction of photoactivated molecules that undergo some change. If there are a number of parallel pathways resulting in different products, then each pathway will have its own quantum yield, ϕ. Sometimes the primary step in a photoreaction is the initiator of subsequent thermal reactions resulting in quantum yields, which may exceed unity. The second law of photochemistry states that the absorption of light by a molecule is a one-quantum process, so that the sum of the primary process quantum yields, ϕ, must be unity, that is, $\Sigma\phi_i = 1$, where ϕ_i is the quantum yield of each of the primary processes, which may include, but are not limited to, dissociation, isomerization, charge-separation, luminescence, or radiationless deactivation.

For a reaction,

$$A \xrightarrow{h\nu} B + C + \cdots \tag{20.5}$$

the rate can be expressed as

$$\frac{-d[A]}{dt} = \phi_A J_A \tag{20.6}$$

where $[A]$ is the concentration of photoactive species. One thus determines quantum yields by measuring the ratio of the reaction rate to the absorbed light intensity, J_A.

20.2.4
Quantum-Mechanical Principles of Absorption

If A and A^* represent the ground and excited states, respectively, the process can be written as the reaction

$$A \xrightarrow{h\nu} A^* \tag{20.7}$$

For the absorption to occur, two criteria must be met by the absorbing species:

1. the difference between the energy of the excited state, E_2, and the energy of the ground state, E_1, must be exactly equal to the energy of the photon, that is,

$$h\nu = E_1 - E_2 \tag{20.8}$$

and

2. there must be a finite probability for the transition, as defined by the electronic structure of the molecule.

The Einstein transition probability for absorption, B_{12}, is given by the following expression:

$$B_{12} = \frac{8\pi^3}{3h^2c} g_2(\mu_{12})^2 \tag{20.9}$$

where g_2 is the degeneracy of the excited state, and μ_{12} is the transition moment given by the integral,

$$\mu_{12} = \int \Psi_1 \left(e \sum_i r_i\right) \Psi_2 \, dq \qquad (20.10)$$

In this equation, Ψ_1 and Ψ_2 are the wave functions of the ground and excited states, respectively; the sum is over all of the lengths of the dipole-moment vectors, r_i, of the electrons; e is the charge on the electron; and the integral is taken over the three-dimensional coordinates, q, of the molecule.

While the exact solution of Eq. (20.10) is impossible for anything but the hydrogen atom, approximate wave functions can be used to derive for various transitions qualitative selection rules that reveal when the transition moment, μ_{12}, is nonzero and thus results in an allowed transition, as opposed to one that is forbidden. Two selection rules are particularly important and relate directly to photochemical processes: the spin conservation rule and the Laporte rule. For systems containing atoms with relatively few electrons (i.e., light atoms), a reliable approximation allows for the separation of the spin wavefunctions, S_1 and S_2, from the total wavefunctions Ψ_1 and Ψ_2, such that the transition moment can be written as the product of two integrals:

$$\mu_{12} = \int \phi_1 \left(e \sum_i r_i\right) \phi_2 dq \times \int S_1 S_2 \, dq \qquad (20.11)$$

where ϕ_1 and ϕ_2 are the orbital wavefunctions of ground and excited states, respectively. The spin part of the transition moment vanishes when $\Delta S \neq 0$, that is, when the total spin quantum numbers, or spin multiplicities, of the excited and ground states are not the same, and the transition is said to be "*spin-forbidden*." The Laporte rule applies to species with a center of symmetry, such as octahedral complexes, and requires that transitions between orbitals must be either $u \rightarrow g$ or $g \rightarrow u$ (g, gerade or even and u, ungerade or uneven, that is, symmetric and antisymmetric, respectively, for inversion of the orbital about the center of symmetry), but not $g \rightarrow g$ or $u \rightarrow u$, such transitions being "parity-forbidden." The methods of group theory are often necessary in applying the Laporte rule.

In practice, not only allowed transitions but also sometimes forbidden transitions may be observed because of subtle perturbations in the system, such as spin–orbit coupling, not taken into account in the approximation procedures used to derive the selection rules. The allowed transitions, of course, are much stronger, as indicated by the bandwidth and extinction coefficient of the absorption band. The classical oscillator strength f of the band is related to the Einstein absorption probability by the following expression:

$$f = \left(\frac{m_e c^2 \omega_{12}}{\pi e^2}\right) B_{12} \qquad (20.12)$$

where m_e is the mass of the electron and ω_{12} is the frequency of the transition in cm^{-1}. Experimentally, the oscillator strength can be determined from the integral of the absorption band, that is,

$$f = 4.32 \times 10^{-9} F \int_{\omega 1}^{\omega 2} \varepsilon \, d\omega \qquad (20.13)$$

where the integration is over all frequencies from the onset to the disappearance of the band in question, and the value of F is related to the index of refraction, n, through $F = 9n/(n^2 + 2)^2$. As a very crude approximation, the oscillator strength can be estimated from

$$f \cong 4.32 \times 10^{-9} F\varepsilon_{\max}\Delta\omega_{1/2}, \quad (20.14)$$

where $\Delta\omega_{1/2}$ is the width of the band at $\varepsilon = 1/2\omega_{\max}$ (half-bandwidth).

Since the oscillator strength for a single electron in a fully allowed transition should be unity, the highest value of $\varepsilon_{\max}$ one should observe is about 5×10^4 M^{-1} cm^{-1}, for a half-bandwidth of 5000 cm^{-1}. Parity- and spin-forbidden transitions have much smaller oscillator strengths, resulting in values of ε on the order of 10^2 or less, but they are easily measured with the crudest of spectrometers. Thus, measurement of the oscillator strengths of various bands observed in a spectrum can be of help in making the correct assignments to the transitions.

The foregoing discussion has been based on the assumption that nuclear motions in ordinary molecular vibrations are so slow that they do not affect the electronic energy levels of the molecule, an approximation known as the *Born–Oppenheimer principle*. This means that the total wavefunction for the molecule can be separated into the product of an electronic part and a nuclear part:

$$\psi_{\text{total}} = \psi_{\text{elec}} \times \psi_{\text{nucl}} \quad (20.15)$$

and each part is treated independently. This permits the construction of the familiar energy diagrams of diatomic molecules, which show the electronic potential energy as a function of internuclear distance, a general example of which is shown in Figure 20.1. From a purely classical viewpoint, one would expect the molecule to reside at the exact bottom of either well, but, in fact, this is not the case because it would imply that the atoms would be stationary, that is, not vibrating.

The quantum-mechanical treatment of the simple harmonic oscillator (which approximates a diatomic molecule to the extent that the bottom parts of the potential wells are approximated by a parabola) requires that the lowest energy of the vibrating molecule must be $1/2h\nu_0$ above the minimum (a result of the uncertainty principle, which does not allow both the position and momentum to be completely defined as would be the case if the zero-point energy were 0), where ν_0 is the fundamental frequency of vibration in this ground state. Several vibrational levels for each electronic state are represented in the wells by horizontal lines that show not only the value of vibrational energy but also the turning points, that is, limits of compression and extension of the atoms during the vibrations.

Figure 20.1 shows what happens during an electronic excitation by absorption of light. Because nuclear motions are vastly slower than electronic transitions, the latter occur vertically, that is, the excited molecule has the same internuclear separation as the ground-state molecule, a rule known as the *Franck–Condon principle*. At temperatures in the vicinity of 25 °C most molecules in a thermally equilibrated system are in the ground vibrational state as a result of the Boltzmann distribution law; therefore, nearly all of the electronic transitions will arise out of this ground vibrational state, as shown in the diagram. However, because the excited electronic potential well is wider and slightly displaced to the right as a consequence of having a weaker

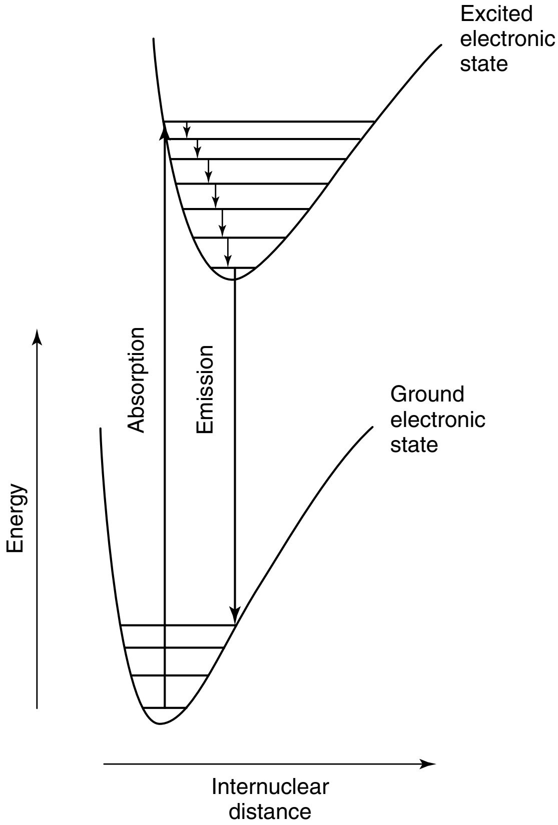

Fig. 20.1 Typical ground- and excited-state potential-energy wells for a molecule, showing vibrational relaxation and emission.

bond than in the ground electronic state, the vertical transitions end in excited vibrational levels in the upper well. Because of the necessity for optimum mixing of the initial and final vibrational wavefunctions in the transitions, those with highest probability occur between vibrational levels in which the molecule is near a turning point in its vibration (except for ground-state vibrations where this highest probability occurs at the center, or equilibrium internuclear distance). This results in a series of transitions to excited vibrational levels in the upper well, giving rise to a set of closely spaced lines in the absorption spectrum of a gas at low pressure or to a broad absorption band in a condensed medium in which the lines are broadened and overlap.

In polyatomic systems, there are numerous vibrational modes and internuclear distances, requiring a multidimensional analog of Figure 20.1 to describe the energy states. Nevertheless, one can still imagine minimum electronic potential-energy wells with superimposed vibrational levels in the hyperspace of n dimensions, where *n* is the number of atoms in the system. Thus, the simple principles of the energetics of diatomic molecules are routinely conceptually applied to more complex systems.

20.2.5
Energy Partitioning and Dissipation Pathways

What happens to the photoactive species after initial electronic excitation by absorption determines the eventual outcome of the photophysical or photochemical process. Since the excited system has a large number of its molecules in excited vibrational levels, it usually undergoes a very rapid thermal equilibration with its surroundings resulting in a Boltzmann population distribution, which means that most molecules drop back to the lowest vibrational level of the excited electronic state, a process called *vibrational relaxation* and shown by the short, downward arrows in Figure 20.1.

Certain bonds between atoms give rise to characteristic frequencies in the infrared (IR), corresponding to bond stretches and bends. Thermal sources of radiation are ineffective in breaking specific bonds, since energy cannot be supplied fast enough to stretch a bond to breaking, in competition with intramolecular and intermolecular energy relaxation.

In recent years, the topic of mode-specific photochemistry has revived in the context of a better understanding of the energy redistribution that may occur inside molecules.

20.2.5.1 Photophysical Pathways: Luminescence and Nonradiative Decay

After the vibrational relaxation of the excited electronic state, the system may return to the ground-state potential well by spontaneously emitting a photon, also shown in Figure 20.1. Note, however, that in this case, since the excited-state potential well is shifted slightly to the right, the most likely electronic transitions are from the lowest vibrational level in the excited electronic state to vibrationally excited states in the ground potential well. This gives rise to an emission band, said to be *Stokes-shifted*, which is at lower energies than, and almost an exact mirror image of, the absorption band. This form of luminescence is often referred to as *fluorescence*, a term which is generally applied to emissions of very short lifetimes between singlet energy levels. From quantum mechanics, the Einstein probability for spontaneous emission, A_{21}, is related to the oscillator strength of the absorption band, through

$$A_{21} = \left(\frac{8\pi^2 {\omega_{21}}^2}{m_e c}\right)\left(\frac{g_1}{g_2}\right) f \qquad (20.16)$$

where g_1 and g_2 are the degeneracies of the ground and excited electronic states, respectively, and the other variables have been previously defined. If emission is the only mode of decay, then

$$\tau_0 = \frac{1}{A_{21}} \qquad (20.17)$$

where τ_0 is the mean radiative lifetime for the emission, a quantity which is thus inversely proportional to the oscillator strength of the absorption.

Nonradiative decay from excited electronic states may occur if there is sufficient overlap between the excited vibrational levels of the lower electronic level and the lowest vibrational states of the excited electronic state. If this occurs between electronic levels of the

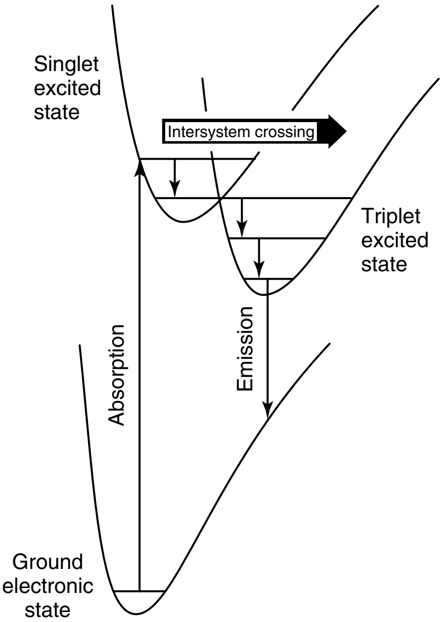

Fig. 20.2 Potential-energy diagram showing excited singlet and triplet states.

same multiplicity, it is called *internal conversion*; if electronic levels of different multiplicities are involved, it is termed *intersystem crossing*. Figure 20.2 illustrates the situation in which there are two excited electronic states of differing multiplicities, for example, singlet and triplet, which have considerable overlap in their vibrational manifolds with each other.

Intersystem crossing from the uppermost potential well to the one that crosses it should be relatively easy from the viewpoint of vibrational relaxation, but forbidden because of the spin conservation rule. However, such processes occur readily because of spin–orbit coupling. Decay of either of the two excited electronic states to the ground state is less likely from the viewpoint of vibrational relaxation because of the poor mixing of vibrational levels, but it can occur through collisional deactivation in which other molecules absorb the energy difference and behave as quenchers of the excited states without actually undergoing chemical change. Since this process is controlled by the collision rate in the system, it is easy to see why this kind of deactivation increases in liquid media at higher temperatures, causing competing luminescence processes to be decreased.

Decay from the excited state of different multiplicity back to the ground state may also occur by emission or nonradiative processes. In both cases, however, the transition has a low probability because of the different spin multiplicities of the two states and the wide separation between the two electronic levels; thus, the excited state has a rather long lifetime. Emission from such a state is often called *phosphorescence* and is characterized by lifetimes as long as seconds, minutes, or even hours after the absorption of light.

Tab. 20.1 Photophysical processes, typical rate constants, and lifetimes in liquid media.

Process	**Rate constants (s^{-1})**	**Lifetimes**
Absorption of photon	$\sim 10^{15}$	1 fs
Vibrational relaxation	$10^{11}-10^{12}$	1–10 ps
Fluorescence	$10^{8}-10^{9}$	1–10 ns
Intersystem crossing between excited states	$0-10^{10}$	>0.1 ns
Phosphorescence	$0-10^{3}$	>1 ms
Radiationless deactivation		
Internal conversion	$0-10^{9}$	>1 ns
Intersystem crossing	$0-10^{3}$	>1 ms

Table 20.1 summarizes the various photophysical processes that may occur in liquid media and the typical ranges in rate constants and lifetimes for each process. Not all emission processes fall neatly into the definitions for fluorescence and phosphorescence and therefore all types of emission can be generally referred to as *luminescence.*

In the condensed phase, vibrational relaxation typically occurs on the order of tens of femtoseconds to hundreds of picoseconds, as shown in laser pump–probe experiments (Elsaesser and Kaiser, 1991). However, it is interesting to note that in some simple liquids, such as liquid nitrogen or inert gases, vibrational relaxation times on the order of minutes have been found (Brueck and Osgood, 1976).

20.2.5.2 **Photochemical Pathways**

The extent to which photochemical products are formed will depend on how well the photochemical pathways compete with the various photophysical decay processes described above. This, in turn, will depend on the reactivity of the excited states, which are usually chemically very different from the ground states. Figure 20.3 is another way of outlining the many choices of pathways through which the excited system can dissipate the energy it gained in absorbing a photon. Nearly all types of reactions can be classified as either unimolecular or bimolecular. The excited states that are reactive may be those which were originally excited in the absorption of light, or they may be other states, labeled *reactive intermediates* in Figure 20.3, accessible by radiationless decay.

There are two special types of reactive intermediates that are formed by bimolecular processes: *excimers* that are formed by dimerization of an excited state, and *exciplexes* that are formed by a complexation reaction of an excited state with ligands. In any case, the potential-energy well of the excited reactive intermediate relative to the ground and excited states of the photoreactant can also be represented by the diagram in Figure 20.2. Photophysical processes are those which result in a return of the system to the ground state while photochemical processes result in different products.

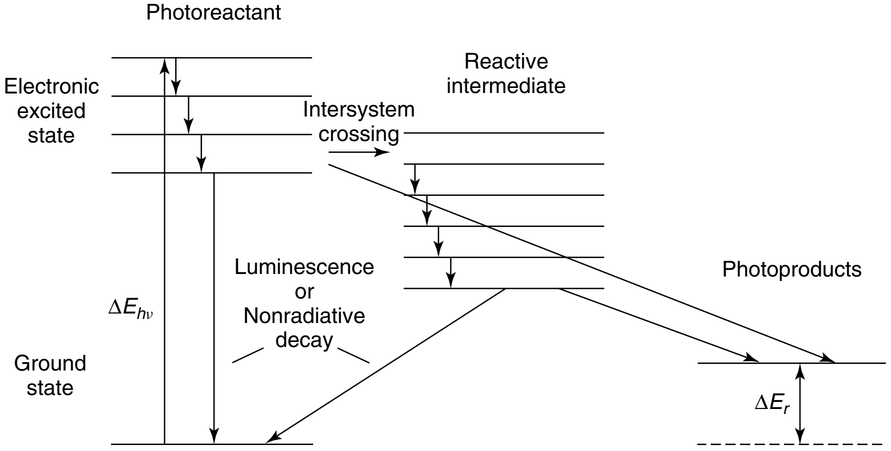

Fig. 20.3 Energy diagram (so-called Jablonski diagram) summarizing the various pathways available to an excited molecule or complex.

Many of these processes occur simultaneously and compete with each other such that one usually observes a mixture of outcomes, each with its own quantum yield. Each of the various decay pathways leading away from the excited photoreactant has its own rate, and the quantum yield for a specific pathway, ϕ_j, is the ratio of an individual rate of decay for that pathway to the sum of the rates of all of the decay pathways, such that (assuming first-order reactions);

$$\phi_j = \frac{k_j}{\sum_i k_i} \tag{20.18}$$

where k_j is the rate constant for the pathway of interest and k_i represents the rate constant for each of the other pathways.

An often-cited quantity is the lifetime, τ, of an excited state or reactive intermediate. This is defined as the time required for a sample to decay to $1/e$ of its original concentration and can be shown to be equal to the reciprocal of the sum of the first-order rate constants for all decay pathways:

$$\tau = \left(\sum_i k_i\right)^{-1} \tag{20.19}$$

It is often the case that excited states may react through bimolecular collisions with other species, thus causing the excited states to be "quenched." Such reactions offer additional pathways for the decay of the excited states, and they must be accounted for in any quantum-yield expression. Since these reactions are second-order processes, their rates will be proportional to the concentrations of both the excited state and the quencher; thus, the quantum-yield expression, Eq. (20.18), will contain a second-order rate constant times quencher concentration for each of these quenching pathways. For example, suppose that the only pathways for decay of an excited photoreactant are luminescence, nonradiative decay, or reaction with a quencher Q, to form a product, as in

$$A \underset{k_L,\, k_{\mathrm{NR}}}{\overset{h\nu}{\rightleftarrows}} A^* \xrightarrow{K_P[Q]} \text{Products} \tag{20.20}$$

where k_L and k_{NR} are the first-order luminescence and nonradiative decay constants, and k_P is the second-order constant for the bimolecular quenching, which forms the product. The quantum yield of product formation, ϕ_P, would then be

$$\phi_P = \frac{k_P[Q]}{k_L + k_{\mathrm{NR}} + k_P[Q]} \tag{20.21}$$

where $[Q]$ is the concentration of the quencher. One can establish that a second-order quenching process occurs by obtaining a linear plot of $1/\phi_P$ versus $1/[Q]$ since the reciprocal of Eq. (20.21) is

$$\frac{1}{\phi_P} = 1 + \frac{k_L + K_{\mathrm{NR}}}{k_P[Q]} \tag{20.22}$$

The slope of such a plot, known as a *Stern–Volmer plot*, can be used to calculate the second-order quenching constant if the lifetime in the absence of quencher is known. An alternative method of obtaining the second-order rate constant directly is from the slope of measured pseudo-first-order decay constants versus $[Q]$.

Many photoreaction schemes are considerably more complex than that shown in Figure 20.3. For an excellent detailed analysis of photochemical kinetics, the reader can consult the book Demas (1983).

20.3
Types of Photochemical and Photophysical Experiments

There are usually two kinds of experiments performed by photochemists: continuous (wave) cw or photolysis, and flash photolysis. Each type is useful for gaining certain kinds of information.

20.3.1
Continuous (cw) Photolysis

In cw photolysis, a continuous source of light, such as a halogen or an arc lamp, is directed or focused onto a sample in a cuvette or photochemical reactor. For the determination of quantum yield, one generally utilizes a well-collimated beam of light impinging on the sample in a cuvette. The beam may first be passed through a monochromator or filter if wavelength information is desired. The rate of reaction may be determined by monitoring at various time intervals properties that are functions of concentration, such as absorbance, pH, conductance, luminescence, or volume of gas emitted or absorbed. The light-absorption rate can be determined from actinometry (see Section 2.2).

Since photochemical processes are sensitive to specific wavelengths of light, careful attention must be paid to the spectral output of the light source selected for experiments.

For quantitative cw experiments, the usual type of lamp used is a high-pressure arc lamp containing mercury, xenon, or a mixture of mercury and xenon. These types of lamps are efficient converters of energy; they have a wide spectral response from the ultraviolet (UV) to the infrared; and because of their short-arc lengths, it is easy to focus

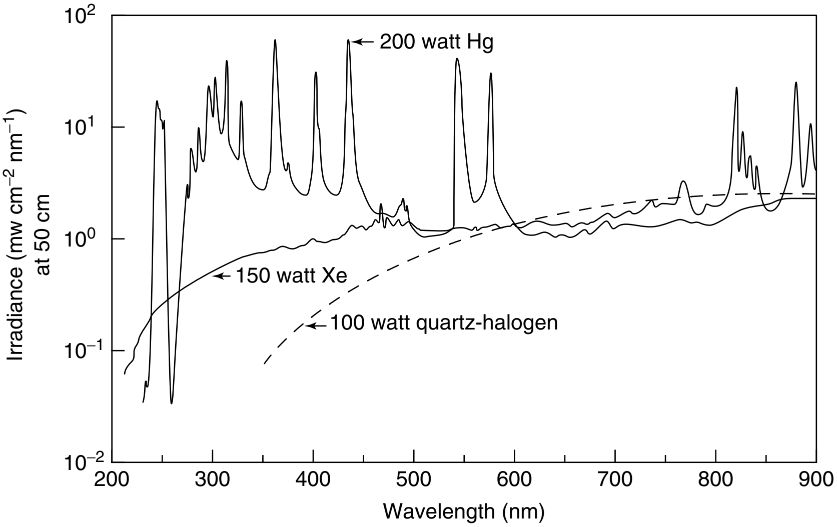

Fig. 20.4 The energy-output spectra for (solid curve) 200-W mercury, (dot-dashed curve) 150-W xenon high-pressure short-arc lamps, and (dashed curve) 100-W tungsten–halogen lamp.

their outputs through monochromators and onto a cuvette. The energy output of a high-pressure mercury arc lamp is concentrated primarily on the series of emission lines of the mercury spectrum, excluding the resonance line at 254 nm, whereas xenon arc lamps tend to give a more continuous output from the UV to the IR (Figure 20.4). The features of both types of lamp can be obtained in an arc lamp envelope filled with a mixture of xenon and mercury. Figure 20.4 also shows the visible portion of the output from a tungsten–halogen lamp, which behaves as a blackbody radiator.

20.3.2
Laser Flash Photolysis

In flash photolysis, a pulsed light source delivers a short burst of light to the sample. This generates short-lived intermediates whose properties, such as absorption or luminescence, can be monitored as a function of time and displayed on an oscilloscope screen. The kinds of information obtained for the intermediates are lifetimes and decay constants, absorption and luminescence spectra, and quantum yields of formation. Such pieces of information lead to an understanding of the intermediate steps and the mechanism of the photochemical reaction. Thus, flash photolysis can “observe” intermediates directly, even those of very fast lifetimes on the nanosecond or picosecond timescale. Both cw and flash photolysis yield important complementary information about photochemical systems.

Modern flash-photolysis experiments rely almost exclusively on pulsed lasers to produce the required high-energy and short pulsewidth flash. Lasers may operate in either the pulsed or cw mode, depending on the type. In all cases, they have extremely narrow bandwidths and virtually nonexistent beam divergences compared to noncoherent light sources. The energy output is centered only at one wavelength and depends on the material in the optical cavity. Availability of a wide selection of wavelengths and the ability to tune lasers over a limited wavelength range have emerged in the past several decades by the use of a variety of cavity materials, dye lasers, frequency modulating crystals, and optical parametric oscillators. Another type of laser, the *free electron laser*, produces coherent light at wavelengths from the far IR to the deep X-ray region, by the relativistic acceleration of a cluster of electrons in a storage ring. Such radiation is called *synchrotron radiation* (see above) and can be a very useful source of tunable laser light at almost any given pulse width, provided one has access to the rather expensive equipment required for its generation. For a complete description of the many laser configurations possible for laser flash photolysis, the reader is referred to Andrews (1992).

A simple laser flash-photolysis setup is diagramed in Figure 20.5. When the sample is excited by absorption of the laser pulse, the oscilloscope is triggered to receive the signal coming from the photomultiplier (PM) attached to the monochromator. The monochromator selects the wavelength of light of interest, and if the transient produced by the flash absorbs at this wavelength, a signal like that shown may appear as the transient decays over a selected time span. This signal can be computer-processed to yield absorbance versus time plots, and the decay constant and lifetime can be calculated. Alternatively, if the monochromator has a photodiode array detector, or if the sample is exposed to series of flashes, each one at a different wavelength, then it is possible to obtain a time-resolved spectrum, that is, an absorption spectrum of the short-lived transient at a given time after the flash. The setup also allows the user to obtain analogous luminescence decay information with the analyzing lamp turned off.

Owing to the limitations of PM rise times and oscilloscope time resolution, a rather different technique has to be used for experiments in the picosecond or femtosecond time regime. In this case, a *pump–probe* configuration is used in which time resolution is controlled by varying the pathlength of the probe pulse. In such experiments, the laser pump and probe pulses are produced in synchrony, but the probe pulse is diverted through an adjustable optical

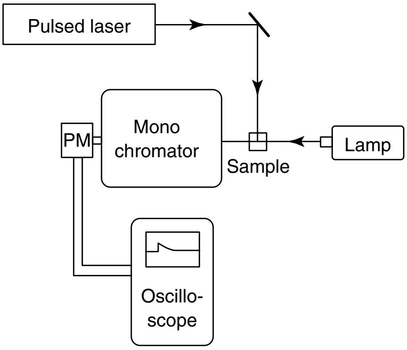

Fig. 20.5 A typical laser flash-photolysis experimental setup.

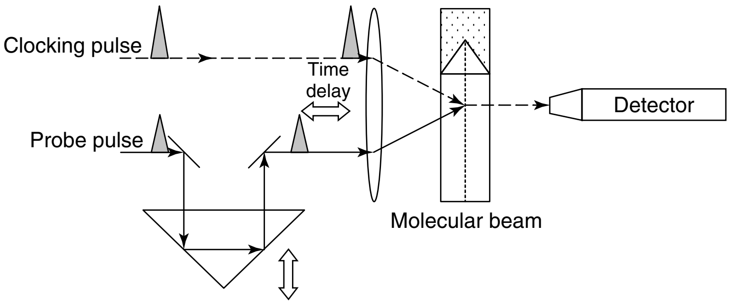

Fig. 20.6 A simplified scheme of the experimental layout for clocking in femtochemistry apparatus.

path length (Figure 20.6). Considering that light travels only about 0.3 mm in 1 picosecond and 0.3 μm in a femtosecond, accurate and precise time delays can be made by very small adjustments in the probe pathlength with respect to the pump pulse.

20.3.3
Laser Spectroscopy

The extremely high spectral purity of a laser beam, coupled with the advances in tunability of lasers, has made them indispensable tools for obtaining high-resolution spectral information of materials. Figure 20.1 shows two electronic energy levels depicted as potential-energy wells, with several vibrational levels superposed on each one. The more closely spaced molecular rotational levels that should be superposed on top of each vibrational level are not shown. In normal spectroscopy, the exciting source for the various transitions between energy levels is some kind of lamp–monochromator combination that has a fairly wide bandwidth and thus limits the degree of resolution of the energies of the transitions. On the other hand, lasers have an extremely narrow bandwidth as an excitation source, so that high-resolution spectra can be obtained. Several important laser spectroscopy techniques are briefly described. For more details on these techniques, the reader is referred to the book by Andrews (1992).

In *electronic photoabsorption spectroscopy*, a sample is excited by a visible or UV laser such that electronic excited states can be probed in the absorption mode. Determining the absorption spectrum of an excited state can yield important information about photochemical intermediates and the various reaction pathways available to a photoexcited species. Alternatively, because the bandwidth of a laser may be considerably smaller than the line width of the electronic transition, which absorbs the laser beam, replacing the conventional monochromater of a spectrometer with a tunable laser can result in highly resolved spectra of the ground state, yielding precise energies and transition probabilities of the excited electronic states. This is especially applicable to gaseous systems at low temperatures, in which the rotational and vibrational features of the spectra become minimized and the electronic transitions become "purer." One common method of significantly lowering the temperature of a gas under study is by supersonic jet-cooling, that is, ejecting the gas from a nozzle at supersonic speeds.

Laser-induced fluorescence (*LIF*) *spectroscopy* is a technique that produces a fluorescence spectrum by exciting the sample with a laser. In conventional fluorescence spectroscopy, one can obtain two types of spectra: an *excitation spectrum* in which the intensity of emission at a fixed wavelength is measured as a function of excitation wavelength and an *emission spectrum* in which intensity of emission is measured as a function of emission wavelength while exciting at a fixed wavelength. Figure 20.1 shows how the fluorescence process occurs and indicates that the excitation spectrum gives the energy changes occurring in the absorption process, whereas the emission spectrum gives the energy released when the excited species returns to the ground state. The use of a tunable laser as the excitation source significantly increases the resolution of an excitation spectrum so that vibrational and rotational fine structure can be observed in samples such as jet-cooled gases, crystals, and molecular beams. Important temporal information can also be obtained by using ultrashort pulsed lasers as the excitation source.

The absorption of laser light can sometimes occur in a multistep process, because of the very high light intensities of a laser beam. For example, a molecule can be excited to an electronic energy level that has a value of E above the ground state by the absorption of two photons of energy, $E/2$, at precisely the same time. This nonlinear effect gives rise to multiphoton absorption spectroscopy. One important advantage of this technique over conventional spectroscopy is that the electronic transitions do not obey the Laporte rule; in two-photon absorption, for example, the transitions must be $g \leftrightarrow g$ or $u \leftrightarrow u$, rather than $g \leftrightarrow u$ or $u \leftrightarrow g$. This increased access to excited states results in a greater understanding of the electronic and vibrational structure of molecules.

20.4
Femtochemistry

In femtochemistry, studies of physical, chemical, or biological changes can be made on the same timescale as that of molecular vibrations, that is, the actual nuclear motions. The femtosecond timescale is unique for the creation of coherent molecular wave packets on the atomic scale of length, a basic problem rooted in the development of quantum mechanics and duality of matter. Molecular wave functions are spatially diffuse and exhibit no motion, but superposition of a number of separate wave functions of appropriately chosen phases can produce the spatially localized and moving coherent wave packet. The powerful concept of coherence lies at the core of femtochemistry and was the key advance in observing the dynamics. Thus, the coupling of ultrafast lasers with molecular beams was essential for the initial developments in this field (Zewail, 2000a,b). The generation, amplification, and characterization of ultrashort pulses are a major part of femtochemistry experiments. Further details can be found in Zewail (1994, 2000a,b).

Such fundamental processes in chemical reactions as energy dissipation, molecular motions, and structural changes occur in the picosecond (10^{-12} seconds) to femtoseconds (10^{-15} seconds) timescale. With the advent of femtosecond lasers, it has become possible to measure events that occur on a timescale shorter than a single vibrational oscillation of a molecule. Thus, ultrafast “snapshots” can be taken of molecules, crystal lattices, unstable intermediates, or other structures in various stages of distortion, yielding valuable information about the structures of molecules and the dynamics of crystal lattices, liquids, individual molecules, and biological systems.

Many systems have recently been examined via femtosecond timescale experiments by Polanyi and Zewail (1995). For example, the dissociation of the ionic compound NaI through a covalent intermediate to the atomic products, Na + I, upon absorption of energy, occurs in less than 10 picoseconds. During this time, the activated intermediate oscillates on the average about eight times, and through an ultrafast series of probe laser pulses the amounts of activated intermediate, ionic $Na^+ I^-$ species, and product sodium atoms can be monitored, with a time resolution down to about 7 femtoseconds. The results have given direct information about the reaction time, the probability of dissociation, and the extent of ionic and covalent character in the NaI bond, and have confirmed many theoretical predictions about reaction dynamics. First, it was shown experimentally that the wave packet was highly localized in space (ca. 10 pm), thus establishing the concept of dynamics at atomic-scale resolution. Second, spreading of the wave packet was minimal up to a few picoseconds, thus establishing the concept of single-molecule trajectory. Third, vibrational (rotational) coherence was observed during the entire course of the reaction, thus establishing the concept of coherent trajectories in reactions, from reactants to products. Finally, the NaI case was the first to demonstrate the resonance behavior, in real time, of a bond converting from being covalent to being ionic along reaction coordinates (Zewail, 2000).

20.5
Reactions of Excited Species

The apparent reactivity of an excited species may be determined by the intrinsic reactivity of its electronic arrangement, by the excitation energy, and by the lifetime of the particular excited state.

The intrinsic reactivity of a species largely depends on the way in which its electrons are distributed in the available orbitals. Electronic excitation alters this arrangement and thus influences the geometry, the dipole moment, the electron-donating or -accepting ability, and the related acid–base properties.

20.5.1
Electronic Energy Transfer

An important overall mechanism of quenching in which the excitation energy is transferred to the quencher in a bimolecular process

$$M^* + Q \longrightarrow M + Q^* \qquad (20.23)$$

is also called *photosensitization*, where Q^* is some excited state of Q usually not directly accessible by optical absorption. The quencher, Q, is said to be sensitized by M, that is, although Q is by itself normally transparent at this wavelength, it is able to become excited by the excited sensitizer, M^*, whose energy should be greater than (or equal to) that of Q^*.

Energy transfer can be either radiative or nonradiative. In the previous case, the emission from M^* is reabsorbed by Q.

(Of course, the absorption spectrum of Q should overlap the emission spectrum of M^*.) Since the S_0-T_1 transition is spin-forbidden, the dominant radiative energy-transfer processes are from a singlet to a singlet or from a triplet to a singlet state.

Nonradiative energy transfers involve the mutual perturbation of the electronic structures of M^* and Q. In long-range coulombic energy transfer (Förster, 1959), the donor and acceptor molecules may be coupled through an electrostatic interaction at large (up to 10 nm) intermolecular separations. The energy-transfer processes favored by the coulombic mechanism are from a singlet to a singlet or from a singlet to a triplet state. In many sensitization reactions, triplet–triplet energy transfer plays the key role.

20.5.2
Photochemistry of Organic Compounds

If excitation affects the nature of the bonding, the shape of the molecule may also be changed. For example, while the ground state of an alkene (e.g., ethene) is planar, in the equilibrium structure of the (π, π^*) excited state the two CH_2 groups lie in perpendicular planes and only a σ bond remains between the carbon atoms.

Excited species are generally both better electron donors and better electron acceptors than the corresponding ground-state species. Therefore, excitation both decreases the ionization energy and increases the electron affinity of the molecule.

Generally, the excited singlet may be more reactive than the triplet. However, since for most organic molecules the ground state is a singlet, the triplet excited state may survive radiative and radiationless quenching so much better than the more reactive singlet. Thus, generally, the triplet is the more important excited species in terms of the overall reactivity in photoinduced organic reactions.

20.5.2.1 Electronic Transitions and Spectra of Organic Molecules

Absorption of a photon by an organic molecule causes a change in its electronic state in a variety of ways that depend on the type of the substrate and the energy of the photon. The highest energy occupied molecular orbitals (HOMOs) of an alkene are the π orbitals in the C–C bond. Absorption of a photon of wavelength around 180 nm promotes an electron from this HOMO to the lowest unoccupied molecular orbital (LUMO) π^*-orbital. The excited state is termed the (π, π^*) *excited state*, and the electronic transition leading to it is a $\pi \rightarrow \pi^*$ transition. While simple alkenes absorb at wavelengths below 200 nm, conjugation of double bonds can shift the absorption maximum into the visible range; for example, $\lambda_{max} = 476$ nm for $CH_3(CH{=}CH)_{10}CH_3$.

The HOMOs of molecules containing heteroatoms (such as N, O, or halogen) are usually nonbonding, and the lowest energy transitions for such molecules are $n \rightarrow \pi^*$ (in the presence of multiple bonds) or n $\rightarrow \sigma^*$ (for saturated molecules). Since $n \rightarrow \pi^*$ transitions, ketones and aldehydes have the lowest energy absorption maxima at 280 and 290 nm, respectively. Acids, anhydrides, and esters have absorption bands at shorter wavelengths (<250 nm), whereas aromatic substitution causes a considerable redshift (e.g.,

$\lambda_{max} \approx 340$ nm for benzophenone). The allowed $\sigma \rightarrow \sigma^*$ transitions of alkanes are less important, giving rise to bands in the vacuum UV.

Most organic molecules are singlets in their ground states; thus, their first excited triplet state is of lower energy than the corresponding singlet state, because of the repulsive nature of interactions between electrons of the same spin.

20.5.2.2 Correlation Rules and Symmetry Conservations

Electron-spin correlation rules are also valid for reactions in which chemical change occurs. Thus, in the permitted adiabatic reactions the total electron spin is conserved, that is, the sum of the spins of the reactants should be the same as that for the products. Conservation of symmetry is especially important to the understanding of pericyclic reactions, or reactions that pass through a cyclic transition state in a concerted manner, that is, with bond-breaking and -forming processes occurring essentially simultaneously. The well-known Woodward–Hoffmann rules (Woodward and Hoffmann, 1970), using the concepts of orbital symmetry, give a good compass for prediction or interpretation of these reactions.

20.5.2.3 Intramolecular Processes

Dissociation Dissociation of a diatomic molecule can give only two atomic fragments with unambiguous chemical identity, whereas that of a polyatomic molecule can sometimes yield many sets of products. For a detailed discussion of the methods used to elucidate primary dissociative mechanisms, the reader can consult the books by Calvert and Pitts (1966) and by Okabe (1978).

Photoisomerization The perpendicular configuration of ethene and its derivatives in the lowest energy (π, π^*) excited state is geometrically equivalent whether it is derived from a cis or trans ground-state molecule. Thus, the probability ratio for the formation of cis and trans isomers depends on the conversion rates of the excited state to the two (isomeric) ground-state forms.

In accordance with the Franck–Condon principle, the significant difference between the geometries of the ground and excited states forbid emission from the latter one. Therefore, in the case where the quantum yields are similar for both cis → trans and trans → cis, prolonged irradiation of either isomer results in the composition of the isomeric mixture at the photostationary state being predominantly determined by the individual absorbances of each isomer at the excitation wavelength. For simple alkenic systems, trans isomers generally possess higher extinction coefficients at longer wavelengths. Thus, irradiation in that range results in the predominance of the cis isomer in the photostationary state. Electrocyclic ring-closure reactions of dienes and trienes are well-known examples of valence and structural isomerizations. These processes leading to the formation of cycloalkenes require the s-cis conformation.

The reactions of polyenes in the excited singlet state often differ from those in the triplet state. Although direct excitation (to S_1) results in internal cyclization, triplet sensitization usually leads to (cyclic) dimerization. Photocyclization of aromatic compounds often results in nonaromatic products. Cyclic conjugated enones and dienones can also undergo structural photoisomerizations with ring rearrangement. These concerted stereospecific reactions proceed via a triplet (π, π^*) state. Another important type of photoisomerization is the di-π-rearrangement, which is formally a 1,2-shift with ring closure. This reaction, however, is probably a nonconcerted process, and promoted via the excited singlet.

Photochemical sigmatropic shift is also a significant type of photoisomerization. In these pericyclic reactions, a σ-bond migrates with respect to a system of π-electrons, resulting in a switching of double and single bonds. A more detailed discussion of these topics is given by Wagner and Hammond (1967).

Hydrogen Abstraction Intramolecular hydrogen atom abstractions, in contrast to photoisomerization processes, involve molecules almost exclusively in the (n, π^*) excited state. This type of reaction is especially significant in the photochemistry of most carbonyl compounds (Wagner, 1971).

20.5.2.4 **Intermolecular Processes**

Hydrogen abstraction can also take place from (n, π^*) excited states (especially of carbonyl compounds, see Scaiano (1973)) in intermolecular processes.

Addition Reactions Both homo- and heteroaddition reactions can be photochemically induced. The latter type of reaction can take place between unsaturated hydrocarbons and water, alcohols, and carboxylic acids or between aromatic compounds and amines. Cycloaddition of benzene and its derivatives can take place across 1,2-, 1,3-, or 1,4-positions. The 1,3-addition predominates with alkenes having only alkyl substituents on the double bonds. 1,2-addition is the main mode of reaction if the electron donor–acceptor properties of the aromatic species and the alkene are significantly different. The 1,4-cycloaddition occurs least frequently and may take place through an exciplex interaction as do the 1,2-additions as well. Cyclobutane derivatives are the main products in the photoaddition of alkenes to the unsaturated bond in α, β-unsaturated ketones. Cyclic enones undergo photocyclic addition more readily than their acyclic analogs because their reduced flexibility inhibits intersystem crossing from the excited triplet state to the ground state. The addition of the carbonyl group itself (in the excited ketones or aldehydes) to suitable alkenes can also occur, producing oxetanes (Arnold, 1968).

Oxygen is one of the most important reactants in photochemical addition reactions of unsaturated compounds. Very efficient photosensitized oxidations are observed with alkenes, dienes, dienoid heterocycles, and polycyclic aromatic compounds. The first products of these reactions are often peroxides or hydroperoxides, which may undergo subsequent oxidation steps. In most of these photosensitized oxidations, the very reactive excited singlet state of oxygen, O_2 ($^1\Delta_g$), is favored via energy transfer from the triplet state of the sensitizers (Foote, 1968). Hence, carotenoids (e.g., β-carotene) are very important photobiological reagents, because as efficient quenchers of singlet oxygen they can protect photosynthetic organisms against photooxidation sensitized by their own chlorophyll.

20.5.3
Photochemistry of Inorganic and Coordination Compounds

To illustrate the kinds of electronic transitions that occur in inorganic photochemical reactions, Figure 20.7 shows a generic molecular-orbital diagram for an octahedral transition–metal coordination complex and the various kinds of electronic transitions that may occur between the orbitals. These are categorized as follows:

1. intraligand (IL) bands, which are transitions between energy levels of the ligands (as discussed earlier about organic photochemistry) and are relatively unaffected by the metal centers;
2. ligand-field (LF) bands, which are the low-intensity transitions, usually in the visible region, caused by the splitting of the metal d orbitals (hence the other often-used term, d–d bands) by the field generated by the coordination of ligands to the metal;
3. ligand-to-metal-charge-transfer (LMCT) bands, which are transitions in which electronic charge is essentially transferred from the ligands toward the coordinating metal;
4. metal-to-ligand-charge-transfer (MLCT) bands, in which electronic charge is transferred from the coordinating metal to the ligands; and
5. charge-transfer-to-solvent (CTTS) transitions, in which electronic charge moves to the solvent. Charge-transfer bands are characterized by their tendency to produce charge separations, which may result in oxidation–reduction reactions.

Charge-transfer bands are characterized by their tendency to produce charge separations, which may result in oxidation–reduction reactions.

20.5.3.1 Ligand-Field (LF) Transitions

The transition metals are in the middle part of the periodic table where the elements have partially filled $(n-1)d$ electron subshells (where n is the period or main shell number). There are five d orbitals in a given subshell, each with its

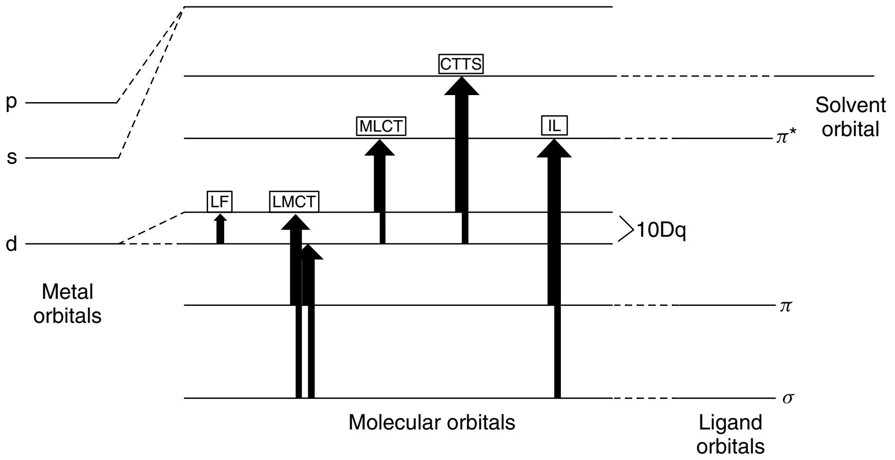

Fig. 20.7 Molecular-orbital diagram for an octahedral coordination complex and the various kinds of electronic excitation transitions that may occur between the orbitals.

own geometrical orientation in space, as shown in Figure 20.8.

In a neutral atom, the five d orbitals of Figure 20.8 are of the same energy, but in a metallic ion coordinately bound to electron-donor ligands there is a splitting of the energies, termed *crystal-field splitting*, caused by the interaction of the ligand electrons with the metal d orbitals. Which atomic orbitals undergo the strongest interaction with the ligands depends on the geometry of the complex: for an octahedral complex, the two e_g orbitals point in the directions of the incoming ligands, thus raising their energy, whereas the three t_{2g} orbitals point away from the ligands resulting in no change in energy. For a tetrahedral complex the situation is reversed, and for other geometries the splitting becomes more complicated. A molecular-orbital picture gives the same splitting patterns between metal orbitals, and these are shown for an octahedral complex in Figure 20.7 by the commonly used symbol for this difference, $10Dq$.

The split atomic orbitals of the metal ion give rise to electronic transitions resulting in LF or d–d bands. Since these transitions are Laporte-forbidden, they are rather weak such that extinction coefficients rarely exceed several hundreds, and the energies of the transitions are usually in the visible to near-IR region of the spectrum, giving a noticeable color to many transition–metal complexes.

For a transition to occur, the coordinating metal ion must have at least one d electron and there must be at least one unfilled orbital in the higher energy state to which an electron can be excited. Excitation of a metal center changes the geometrical distribution of electrons and hence may change the strength of the metal–ligand bond. Thus, LF transitions may result in ligand-exchange, isomerization, or rearrangement reactions without changing the oxidation state of the metal ion. A large number of photochemical reactions induced by LF transitions have been identified and are reviewed in the books by Balzani and Carassiti (1970), Adamson and Fleischauer (1975), Geoffroy and Wrighton (1979), and Ferraudi (1988).

20.5.3.2 Ligand-to-Metal (LMCT) and Metal-to-Ligand (MLCT) Charge-Transfer Transitions

Normally, charge-transfer bands are observed in the UV and short-wavelength visible region in complexes consisting of a donor and acceptor of electronic charge. If the donor is an electron-rich ligand, such as NH_3 or halo, and the acceptor a readily reduced metal, such as

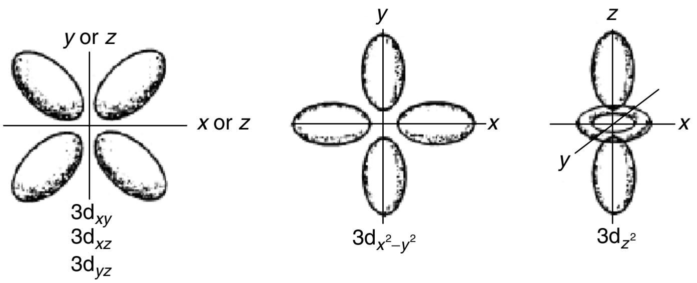

Fig. 20.8 Electron probability pictures of the three d orbitals of an atom.

Co(III) or Fe(III), the transition is LMCT; if the donor is an oxidizable metal, such as Cu(I) or Fe(II), and the acceptor an electron-poor ligand, such as 1,10-phenanthroline or bipyridine, the transition is MLCT. Such bands are Laporte- and (usually) spin-allowed yielding absorption coefficients on the order of 10^4.

One should expect that the excited species following charge-transfer absorption is at once a stronger oxidizing and reducing agent than the ground state because of the shift in electron density between metal and ligand. Thus, the photochemistry resulting from such transitions should involve redox processes of metal and ligand. This certainly seems to be the case following MLCT absorption, as exemplified by the well-known tris(bipyridyl) ruthenium (II) system (Gafney and Adamson, 1972), which has been heavily researched as a possible photocatalytic agent for water splitting (Creutz and Sutin, 1976) since the transitions occur in the visible portion of the spectrum. In addition to redox processes, MLCT transitions often result in reductive elimination.

In general, MLCT transitions usually result in processes that maintain the integrity of the metal–ligand bonds because they originate in filled or nearly filled metal d orbitals, which do not have much effect on the bonding, and they terminate in ligand-localized orbitals, which also are not usually involved in the metal–ligand bond.

By contrast, LMCT transitions often occur between ligand orbitals involved in the ligand–metal bond and a metal in a higher oxidation state with unfilled d orbitals, resulting in a coordinatively unstable, reduced metal–radical ion pair, which usually decomposes further to a complicated array of products. A useful list, prepared by Endicott (1975), tabulates the primary reduced metal and radical formed from the LMCT excitation of each of some 39 complexes of Co(III), Rh(III), Ir(III), Pt(IV), Fe(III), and Ru(III) in aqueous solution.

20.5.3.3 Charge-Transfer-to-Solvent (CTTS) Transitions

CTTS transitions are favorable when the irradiated species is easily oxidized and electron-rich, and there is some low-lying unoccupied orbital on the solvent molecule in the first or second coordination sphere of the complex. Since the first criterion for such transitions is the same as that which favors MLCT transitions, it is common for the lower oxidation states of such metals as Cu, Ru, and Fe to exhibit both MLCT and CTTS bands, sometimes with both types of transitions occurring in the same complex when an acceptor ligand is present, as has been observed, for example, in the case of the cyano complexes of copper(I) and iron(II). The most obvious indicator of a CTTS transition is the ejection by the excited species of a solvated electron into the bulk of the solvent.

20.5.3.4 Other Charge-Transfer Transitions

The number of di- and polynuclear complexes exhibiting metal–metal bonding has increased dramatically in the past few years. In such compounds, it is expected that there will be significant orbital overlap such that new metal–metal delocalized molecular orbitals will be formed. These molecular orbitals can then take part in transitions resulting in charge-transfer between the metals themselves, that is, metal–metal charge-transfer (MMCT) or between the metal–metal orbitals and ligand orbitals (perhaps labeled M_2LCT or LM_2CT, depending on the direction of the charge-transfer). For a detailed compilation of charge-transfer photochemistry of coordination compounds, see the book by Horváth and Stevenson (1993).

20.6
Conclusion

As this concise introduction demonstrates, photochemistry has great theoretical and practical significance in the natural sciences, industry, and nature. A multitude of phenomena, which have been observed and interpreted using photochemical theory and techniques, have strong connections to physics and biology, as well as to the material sciences. The development of more and more sophisticated experimental techniques in photochemistry makes possible the elucidation of the dynamics of elementary reactions. Perhaps even the energy needs of humankind will one day be met by an environmentally friendly photochemical solar energy conversion process.

Acknowledgments

This work was supported by the Hungarian Scientific Research Fund (OTKA K63494).

Glossary

Absorbance: $\log_{10}(I_0/I)$, where I_0 is the incident light intensity falling on an absorbing sample and I is the transmitted light intensity.

Actinometry: Measurement of light intensities for quantum-yield determinations.

Bond-Specific Photochemistry: Photochemistry having products determined by a specific bond or combination of bonds set into vibrational motion.

Chromophore: Light-absorbing chemical group.

Cis: A molecular structural configuration in which two like functional groups are adjacent to each other on the molecular skeleton.

Concerted Reaction: A reaction in which bond-breaking and -forming occur essentially simultaneously.

Conformation: The (possibly unstable) arrangement of atoms inside a molecule.

Conformers: Different, relatively stable, arrangements of atoms, each consisting of identical numbers and types of atoms. A dissociating or rearranging molecule may pass through different conformers. Conformers need not be capable of being physically or chemically isolated.

Conjugation: Alternating double and single bonds within a molecular chain or ring, resulting in bond stabilization via electron delocalization.

Dipole Moment: A measure of the separation between centers of positive and negative charges in polar molecules.

Excimer: A complex formed from the excited state of a molecule with an identical molecule in the ground state.

Exciplex: A complex formed from an excited molecule and a different molecule in the ground state.

Extinction Coefficient: Specific absorbance of 1-cm pathlength of solution at 1 mol l^{-1} concentration.

Flash Photolysis: Experimental study of photochemical or photophysical processes initiated by a short pulse of light.

HOMO: Highest occupied molecular orbital.

Internal Conversion: Energy loss within a given electronic excited or ground-state manifold.

Intersystem Crossing: Transition of an excited electron from a given multiplicity (such as the singlet) to a different one (such as the triplet).

Isomers: Different molecular conformations, consisting of identical atoms, that are stable, or metastable long enough to be physically or chemically separated.

Lifetime: The time required for an excited species to decay to $1/e$ of its initial concentration.

LUMO: Lowest unoccupied molecular orbital.

Mechanism: Sequence of physical and chemical steps leading to a chemical change. Sometimes used to refer to the most unique or characteristic step of an overall mechanism.

Mode-Specific Photochemistry: Photochemistry having products determined by the vibrational mode set into motion.

Monomers: Low-molecular-weight molecules or atoms that can join together to form a polymer.

Multiphoton Absorption: Direct multiphoton absorption occurs when two or more photons are absorbed at once, mediated by higher order perturbation terms of the molecular Hamiltonian interaction with the radiation field. Sequential one-photon events are sometimes called by this name, but more properly may be called *multiple-photon absorption*.

Nonradiative Energy Transfer: Loss of energy from an excited state by any process that does not produce light.

Normal Mode: Classically, molecular vibration in which (nearly) all atoms move concertedly and periodically with the same frequency, but possibly not with the same phase. Quantum mechanically, eigenfunction of the vibrational Hamiltonian.

Organometallic: A compound containing organic carbon and metal atoms.

Oscillator Strength: A measure of the area under an absorption band, or the degree to which a transition is allowed.

Pericyclic Reactions: Concerted reactions proceeding via a cyclic transition state.

Photodetachment: Ejection of an electron from a negative ion by the effect of light.

Photodissociation: Fragmentation of a molecule into two or more pieces by the effect of light.

Polarizability: The measure of charge displacement by an externally applied electric field.

Radiative Energy Transfer: Loss of energy of an excited state by emission of light.

Sensitizer: Compound that absorbs light at the irradiation wavelength, then transfers its excited-state energy to a substrate that cannot be directly excited.

Singlet: An electron configuration in which the sum of the electron-spin quantum numbers is 0, that is, all the electrons are paired with respect to their spins.

Spin–orbit Coupling: Interaction of the orbital and spin angular momenta of an electron in an atom or polyatomic species.

Trans: A molecular structural configuration in which two like functional groups are opposite to each other on the molecular skeleton.

$u \longrightarrow g$ or $g \longrightarrow u$ Transitions: Laporte-allowed transitions between wave functions that are even (gerade or *g*) or odd (ungerade or *u*) with respect to inversion about a center of symmetry.

Triplet: An electron configuration in which the sum of electron-spin quantum numbers is 1, that is, there are two unpaired electrons of parallel spin.

Vibrational Relaxation: Loss of internal vibrational energy into other modes of motion.

References

Adamson, A.W. and Fleischauer, P.D. (1975) *Concepts of Inorganic Photochemistry*, John Wiley & Sons, Ltd, New York.
Arnold, D.R. (1968) *Adv. Photochem.*, **6**, 301–423.
Andrews, D.L. (1992) *Applied Laser Spectroscopy*, Wiley-VCH Verlag GmbH, New York.
Balzani, V. and Carassiti, V. (1970) *Photochemistry of Coordination Compounds*, Academic Press, New York.
Brueck, S.R.L. and Osgood, R.M. Jr (1976) *Chem. Phys. Lett.*, **39**, 568–572.
Calvert, J.G. and Pitts, J.N. Jr (1966), *Photochemistry*, Chaps. 3–5, John Wiley & Sons, Ltd, Chichester, New York.
Creutz, C. and Sutin, N. (1976) *J. Am. Chem. Soc.*, **98**, 6384, 6385.
Demas, J.N. (1983) *Excited State Lifetime Measurements*, Academic Press, New York.
Elsaesser, T. and Kaiser, W. (1991) *Annu. Rev. Phys. Chem.*, **42**, 83–107.
Endicott, J.F. (1975) in *Concepts of Inorganic Photochemistry*, (eds A.W. Adamson and P.D. Fleischauer), Chap. 3, John Wiley & Sons, Ltd, New York.
Ferraudi, G.J. (1988) *Elements of Inorganic Photochemistry*, John Wiley & Sons, Ltd, New York.
Foote, C.S. (1968) *Acc. Chem. Res.*, **1**, 104–110.
Förster, Th. (1959) *Discuss. Faraday Soc.*, **27**, 7–17.
Gafney, H.D. and Adamson, A.W. (1972) *J. Am. Chem. Soc.*, **94**, 8238–8239.
Geoffroy, G.L. and Wrighton, M.S. (1979) *Organometallic Photochemistry*, Academic Press, New York.
Hatchard, C.G. and Parker, C.A. (1956) *Proc. R. Soc. London, Ser. A*, **235**, 518–536.
Okabe, H. (1978) *Photochemistry of Small Molecules*, Chaps. 2, 4–7, John Wiley & Sons, Ltd, Chichester, New York.
Polanyi, J.C. and Zewail, A.H. (1995) *Acc. Chem. Res.*, **28**, 119–132.
Scaiano, J.C. (1973) *J. Photochem.*, **2**, 81–118.
Wagner, P.J. (1971) *Acc. Chem. Res.*, **4**, 168–177.
Wagner, P.J. and Hammond, G.S. (1967) *Adv. Photochem.*, **5**, 21–156.
Wegner, E.E. and Adamson, A.W. (1966) *J. Am. Chem. Soc.*, **88**, 394–404.
Woodward, R.B. and Hoffmann, R. (1970) *The Conservation of Orbital Symmetry*, Wiley-VCH Verlag GmbH, Deerfield Beach, FL.
Zewail, A.H. (1994) *Femtochemistry Ultrafast Dynamics of the Chemical Bond*, Vols. I and II, World Scientific, Singapore.
Zewail, A.H. (2000a) *J. Phys. Chem.*, **104**, 5660–5694.
Zewail, A.H. (2000b) *Les Prix Nobel*, Almquist and Wiksel International, Stockholm.

Further Reading

Andrews, D.L. (1992) *Applied Laser Spectroscopy*, Wiley-VCH Verlag GmbH, New York.
Andrews, D.L. (1997) *Lasers in Chemistry*, 3rd edn, Springer-Verlag, Berlin.
Balzani, V. and Campagna, S. (2007) *Photochemistry and Photophysics of Coordination Compounds I, II (Topics in Current Chemistry)*, Springer, Berlin.
Bauerle, D. (1986) *Chemical Processing with Lasers*, Springer-Verlag, Berlin.
Boule, P. (ed.) (1999) Environmental photochemistry, in *The Handbook of*

Environmental Chemistry, (ed. O. Hutzinger), (Ed.-in-Chief), Vol. 2, Part L, Springer, Berlin.

Braun, A.M. (1991) *Photochemical Technology*, John Wiley & Sons, Ltd, New York.

Coyle, J.D. (1986) *Introduction to Organic Photochemistry*, John Wiley & Son, Ltd, Chichester, New York.

Coyle, J.D. and Hill, R.R. (eds) (1982) *Light, Chemical Change and Life: a Source Book in Photochemistry*, The Open University Press, Walton Hall, Milton Keynes, UK.

Demas, J.N. (1983) *Excited State Lifetime Measurements*, Academic Press, New York.

DeSchryver, F.C., DeFeyter, S., and Schweitzer, G. (eds) (2001) *Femtochemistry*, Wiley-VCH Verlag GmbH, Weinheim.

Ferraudi, G.J. (1988) *Elements of Inorganic Photochemistry*, John Wiley & Sons, Ltd, New York.

Fleming, G.R. (1986) *Chemical Applications of Ultrafast Spectroscopy*, Oxford University Press, New York.

Frimer, A.A. (ed) (1985) *Singlet Oxygen*, Vols. 1–4, CRC Press, Boca Raton, FL.

Gilbert, A. and Bagott, J. (1991) *Essentials of Molecular Photochemistry*, Blackwell Scientific, Oxford, UK.

Grigoropoulos, C.P. (1998) *Experimental Methods in the Physical Sciences*, Vol. 30, Academic Press, NewYork, pp. 173–223.

Griesbeck, A. and Mattay, J. (eds) (2004) *Synthetic Organic Photochemistry (Molecular and Supramolecular Photochemistry)*, CRC Press, Boca Raton, FL.

Grossweiner, L.I. (1994) *The Science of Phototherapy*, CRC Press, Boca Raton, FL.

Hollenberg, J.L. (1970) *J. Chem. Educ.*, **47**, 214.

Horspool, W.M. (ed) (1984) *Synthetic Organic Photochemistry*, Plenum, New York.

Horspool, W.M. and Lenci, F. (2003) *CRC Handbook of Organic Photochemistry and Photobiology*, Vols. 1 and 2, 2nd edn, CRC Press, Boca Raton, FL.

Horváth, O. and Stevenson, K.L. (1993) *Charge Transfer Photochemistry of Coordination Compounds*, Wiley-VCH Verlag GmbH, New York.

Ibbs, K.G. and Osgood, R.M. (eds) (1989) *Laser Chemical Processing for Microelectronics*, Cambridge University Press, Cambridge, UK.

Kimble, M. (2006) *Femtochemistry VII: Fundamental Ultrafast Processes in Chemistry, Physics, and Biology*, Elsevier, Amsterdam.

Klessinger, M. and Michl, J. (1995) *Excited States and Photochemistry of Organic Molecules*, Wiley-VCH Verlag GmbH, New York.

Kopecky, J. (1992) *Organic Photochemistry: A Visual Approach*, Wiley-VCH Verlag GmbH, New York.

Kuteladze, A.G. (2005) *Computational Methods in Photochemistry (Molecular and Supramolecular Photochemistry)*, CRC Press, Boca Raton, FL.

Letokhov, V.S. (1983) *Nonlinear Laser Chemistry*, Springer-Verlag, Berlin.

Luxon, J.T. and Parker, D.E. (1987) *Lasers in Manufacturing*, IFS, Kempston, Bedford, UK.

Manz, J. and Wöstle, L. (eds) (1995) *Femtosecond Chemistry*, Wiley-VCH Verlag GmbH, Weinheim, Germany.

Martin, M.M. and Haynes, J.T. (eds) (2004) *Femtochemistry and Femtobiology: Ultrafast Events in Molecular Science*, Elsevier, Amsterdam.

Matsumoto, J. (2005) in *Organic Photochemistry and Photophysics (Molecular and Supramolecular Photochemistry)*, (eds Ramamurthy, V. and Schanze, K.), CRC Press, Boca Raton, FL.

Monroe, B.M. and Weed, G.C. (1993) *Chem. Rev.*, **93**, 435–448.

Montalti, M., Credi, A., Prodi, L., and Gandolfi, M.T. (2006) *Handbook of Chemistry*, CRC Press, Oxford, UK.

Murov, S.L., Carmichael, I., and Hug, G.L. (1993) *Handbook of Photochemistry*, Marcel Dekker, New York.

Olivucci, M. (2005) *Computation Photochemistry*, Vol. 16 (Theoretical and Computation Chemistry), Elsevier Science, Amsterdam.

Phillips, R. (1983) *Sources and Applications of Ultraviolet Radiation*, Academic Press, New York.

Rabek, F.F. (1982) *Experimental Methods in Photochemistry and Photophysics*, John Wiley & Sons, Ltd, New York.

Roffey, C.G. (1982) *Photopolymerization of Surface Coatings*, John Wiley & Sons, Ltd, Chichester.

Shipley, T. and Crescitelli, F. (1979) *Visual Photochemistry: The Beginnings*, Pergamon Press, Oxford, UK.

Singh, J. (2006) *Photochemistry and Pericyclic Reactions*, New Age International (P) Ltd, New Delhi.

Turro, N.J. (1991) *Modern Molecular Photochemistry*, Chaps. 10–14, University Science Books, Sausalito, CA.

Wayne, R.P. (1988) *Principles and Applications of Photochemistry*, Oxford University Press, Oxford, UK.

Wayne, C.E. and Wayne, R.P. (1996) *Photochemistry (Oxford Chemistry Primers, 39)*, Oxford University Press, Oxford, UK.

RECORDING DEVICES

Hideo Fujiwara, *Center for Materials for Information Technology, The University of Alabama, Tuscaloosa, Alabama, U.S.A.*

	Introduction	149
1.	**Magnetic Recording**	150
1.1	Principles and Modes of Magnetic Recording	150
1.2	Magnetic Recording Devices	152
1.2.1	Audio Tape Recorders	152
1.2.2	Video Tape Recorders	153
1.2.3	Digital-Data Recording	154
1.3	Magnetic Heads	155
1.3.1	Inductive Heads	156
1.3.2	Magnetoresistive (MR) Heads	160
1.4	Magnetic Recording Media	163
1.4.1	Particulate Media	163
1.4.2	Thin-Film Media	166
1.4.3	Perpendicular-Recording Media	169
2.	**Optical Recording**	170
2.1	Concept of Optical Recording	170
2.2	Optical Recording Devices	170
2.3	Optical Heads	172
2.4	Optical Recording Media	173
2.4.1	Read-Only Disks	173
2.4.2	Write-Once (W-O) Disks	174
2.4.3	Rewritable (Erasable) Disks	175
2.4.3.1	MO (magneto-optic) Disks	175
2.4.3.2	Phase-Change–Type Disks	179
2.4.4	Multiplex Recording	180
2.4.4.1	Hole Burning	180
2.4.4.2	Holographic Recording	181
3.	**Conclusion**	181
	Glossary	181
	Works Cited	182
	Further Reading	183

INTRODUCTION

People since prehistoric ages have been managing to create various methods of recording to store and communicate information. Varieties of media were utilized such as rocks, bones and skins of animals, trunks and skins of trees or plants including the pith of the papyrus cut and pressed into sheets, etc. Recording was performed by putting a mark upon them by carving or painting. The inventions of paper and typography were indeed epoch-making in the history of recording technology. However, with the progress of electronics, recording technology has made a huge advance in its performances such as quality, capacity, speed, etc., and its progress is still being accelerated in accordance with the ever-increasing demand of the present information-technology era. This article focuses on magnetic and optical recording devices that will play a decisive role in the information processing, although there exist several key technologies in recording such as pen recorders, phonograph-type records, and photographic recording, about which the readers may be referred to the articles LABORATORY INSTRUMENTATION, SOUND REPRODUCTION, and RECORDING PHOTOGRAPHIC, respectively.

Magnetic recording has a history of almost one century, during which enormous development has been made in all aspects, starting from a steel wire recorder. The inventions of magnetic-powder–coated medium and the ring-type head especially have promoted surprisingly great progress in this technology, resulting in various kinds of systems such as audio and video tape recorders, disk memories of both rigid and flexible types, card systems, etc., which are all now indispensable in human life. On the other hand, it has not been much longer than fifteen years since optical recording systems were brought into market as video disk and compact disk (CD) players right after the mass-production technology for laser diodes had been accomplished. In the following, brief concepts of both magnetic and optical recordings will be given with some emphasis on the materials used recently and some idea

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

about future trends rather than historical ones.

1. MAGNETIC RECORDING

1.1 Principles and Modes of Magnetic Recording

Magnetic recording devices basically consist of two parts: a medium on which information is written in the form of magnetic patterns corresponding to the information, and a head by which the information is written and/or read out. The medium, therefore, has a magnetic layer, which is usually coated on a substrate—rigid or flexible depending on the devices in which it is used. The write head is composed of a magnetic core and a coil wound around it and is designed to generate a magnetic flux in a restricted tiny portion, the strength of which is controlled by an electric current applied to the coil. The read head can be the same write head, or a specifically designed one that is sensitive to a magnetic flux coming out of the recorded medium.

In Figs. 1(a) and 1(b) are illustrated schematically the write processes of the two typical modes (longitudinal and perpendicular; Iwasaki and Nakamura, 1977) of magnetic recording. A magnetic head having (a) a small gap or (b) a thin main pole accompanied by an auxiliary pole is put close to a recording medium, with the gap or the pole tips facing it. During writing, the head moves relative to the medium, and a current modulated according to the information to be written is applied to the coil wound around the head core (or pole), making magnetic flux flow out of the gap (or the pole tip). This flux creates a track of magnetic pattern carrying the information in the form shown in the figures, with the head width determining the track width. For readout, the track is traced by the head. This time, the leakage flux from the magnetic pattern flows into the head core or the poles passing through the coil inducing a voltage in the coil, from which the information embedded in the magnetic pattern is reproduced.

Figure 2 illustrates the principle of the magnetic recording process by use of hysteresis curves (curves of magnetization M vs field H) of the magnetic layer of a medium, taking a pulse recording as an example. Here, H is the sum of all the fields applied to the relevant spot, i.e., the head field, the leakage field from the surroundings H_{srd}, and

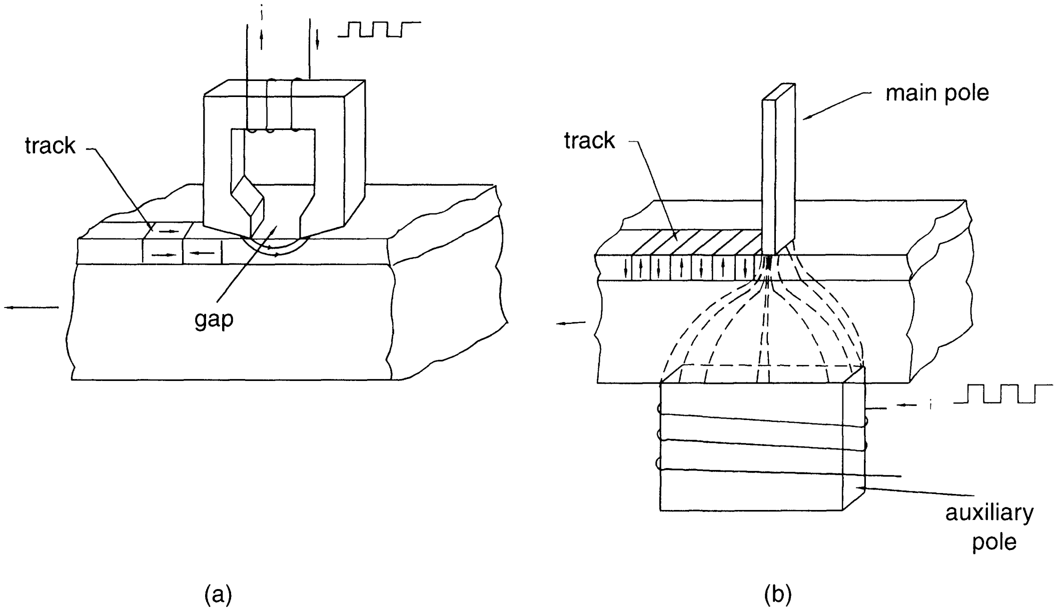

(a) (b)

FIG. 1. Write process of magnetic recording: (a) longitudinal and (b) perpendicular recording.

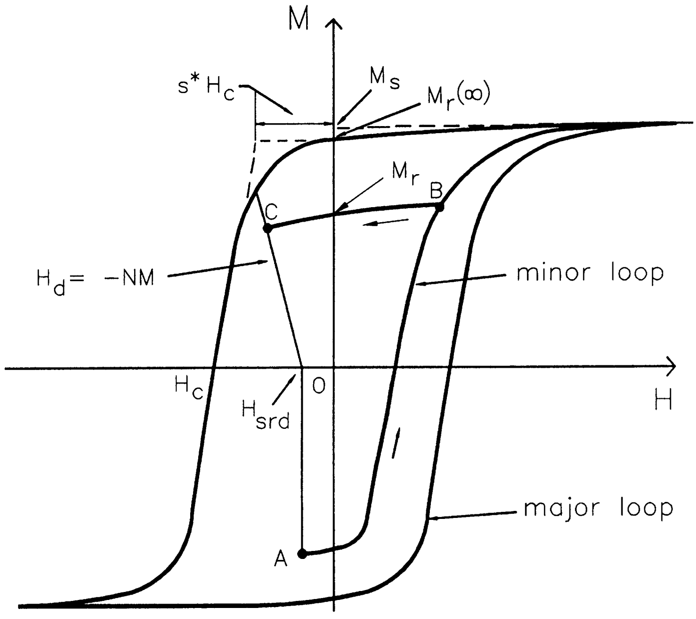

FIG. 2. Recording process on the *M–H* curves (pulse recording). *A*, initial state; *B*, maximum point (corresponding to the maximum write field), *C*, final state (recorded state); H_{srd}, field from the surroundings; H_d, demagnetization field; *N*, demagnetization factor. The saturation magnetization M_s, remanence M_r, squareness ratio M_r/M_s, coercive field H_c, and s^* (which is defined as shown in the figure) are the fundamental factors determining the recording characteristics.

the demagnetization field H_d, which is expressed as

$$H_d = -NM, \tag{1}$$

where N is a coefficient determined by the shape of the written bit and called "demagnetization factor." The recording starts from the initial point *A* and is completed at the final point *C* as is indicated by the arrows. The hysteresis curves represent the magnetic properties inherent to the medium and depend on the materials and fabrication conditions, and proper properties should be chosen for each specific device.

The modes of writing are classified basically into longitudinal and perpendicular recording and also into analog and digital recording according to the direction of the magnetization upon which signals are loaded and the nature of the signals to be recorded, respectively.

Although perpendicular recording is expected to show an excellent characteristic especially in the very high recording density, all practical systems have historically employed the longitudinal mode, with the exception of the evaporated tape case, which can be called "intermediate mode." Figures 3(a) and 3(b) show examples of the cross-sectional view of the recorded pattern of three different recording densities obtained by simulations for each mode of recording upon media having typical magnetic properties suitable for high-density recording for each case (for instance, the thickness for longitudinal recording is chosen much thinner than for perpendicular recording). The scaling ratio of the magnetization vectors for the longitudinal and perpendicular recordings is set as 3. Thus, it is seen that for the highest recording density shown here [300 kbpi (bpi: bits/inch)] the maximum recorded magnetization at the surface in perpendicular recording is much greater than that in longitudinal recording. Analog signals are recorded either by the ac bias method or by the frequency-modulation (FM) method, while digital signals are recorded by the pulse-coding method. The ac bias method was invented in order to improve the linearity of the recording response to the input head current, in which a high-frequency current (bias current) is added to the signal current. The FM method is rather similar to digital recording in the nature of the current wave form. The only difference is that in the FM method, the zero-crossing points are modulated in the analog manner, while in digital recording, they appear with intervals of integers times the predetermined minimum interval. The digital recording method

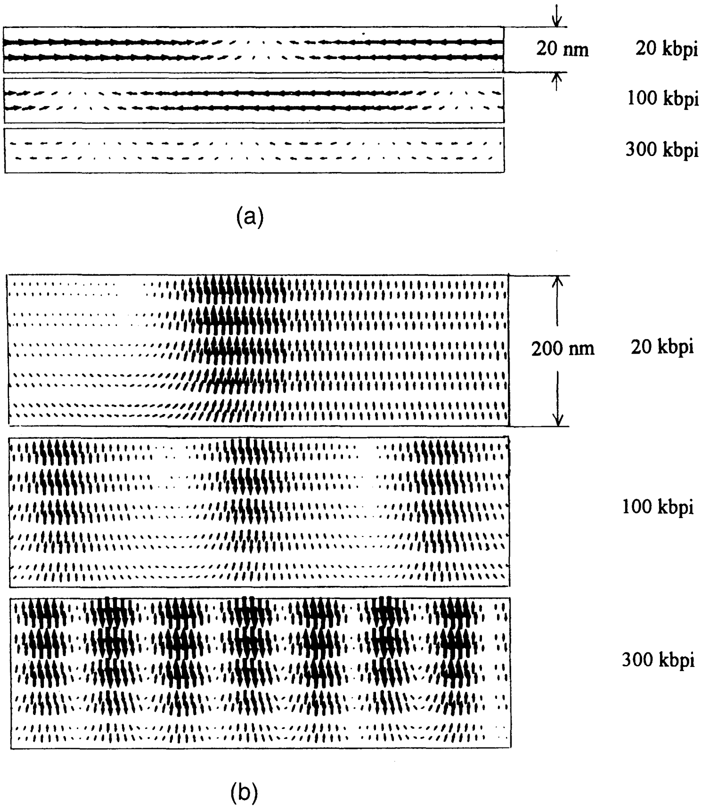

FIG. 3. Cross sections of recorded patterns for (a) longitudinal and (b) perpendicular recording at various recording densities obtained by simulation. The scaling ratio of the magnetization vectors for (a) and (b) is 3:1. (Courtesy of Dr. K. Yoshida and Mrs. F. Akagi, Hitachi.)

was first started from digital computer use. However, it is now being applied widely for all kinds of devices because of its potential perfectness in the reconstruction of the original signals by use of error-detection and -correction schemes.

1.2 Magnetic Recording Devices

A variety of devices exist that are designed to match various applications in the forms of a tape, drum, disk, card, sheet, etc. Among them, audio-video recording and digital-data storage for computers and instrumentation are the major applications. In Fig. 4, some typical magnetic recording devices are shown schematically.

1.2.1 Audio Tape Recorders Originally, audio tape recorders started from an open-reel type, in which two reels were set on a deck equipped-with a fixed read/write head and the tape was transported from one reel to the other during the read/write processes. However, they have been almost completely replaced by the two-reel "compact cassette" type, especially for consumer use, in which two stereo tracks are recorded on a 3.8-mm-wide tape.

In order to provide better sound quality and duplicability, digital tape recorders have been developed, which are divided into two categories; rotary-head type and stationary-head type. The rotary-head type recorder of the standard type is called R-DAT (3.8-mm tape), which is a modification of the VCR (video cassette-tape recorder) in which a helical-scan rotary head and a cassette tape are employed. The format called DASH (digital audio stationary head) was the first adopted for the stationary type. It is of an open-reel type carrying 2-in. tape and uses 24 or 48 heads in parallel. The plurality of heads allowed reducing the tape speed, which must have otherwise been made greater than 10 m/s. Another format of 32 channels using 1-in. tape is now also available. Recently, a stationary-head–type digital tape recorder

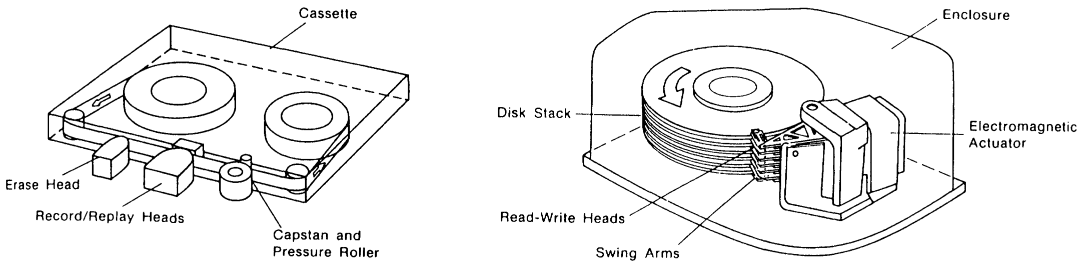

(a) Audio cassette recorder.

(c) Rigid disk drive.

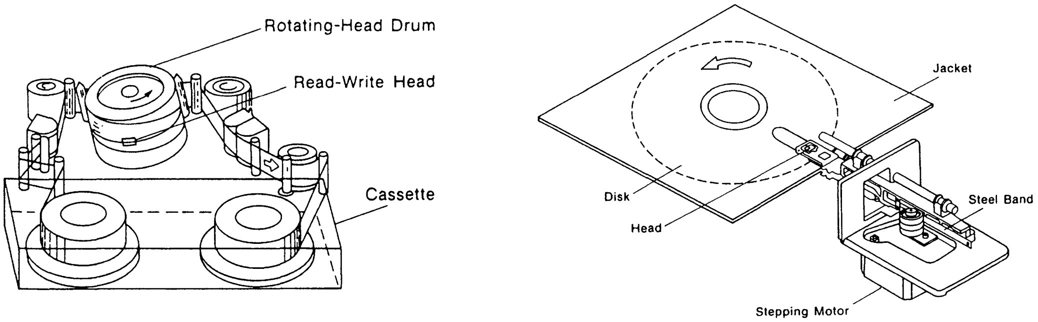

(b) Helical-scan video recorder.

(d) Floppy disk and head accessing system.

FIG. 4. Schematic illustration of typical magnetic recording devices (after Mee and Daniel, 1989).

named DCC (digital compact cassette) equipped with 18 parallel thin-film heads has been developed for consumer use. The cassette has the same basic dimensions as the conventional "compact cassette" providing read compatibility with the conventional one. The heads used for each of the above devices are designed differently from each other as will be seen in Fig. 7. In Table 1, recording characteristics and some other parameters are listed for typical audio tape recorders described above.

1.2.2 Video Tape Recorders Video signals usually carry more than two orders of magnitude wider frequency bandwidth (several megahertz) than audio ones. The frequency response of magnetic recording is not good enough to record directly such wide-band signals, and several modulation methods have been explored. Among them frequency modulation (FM) was proved to be most appropriate because of its insensitivity to the low-frequency noise inherent to electronics.

The first video recording system for broadcasting was accomplished in the form of the so-called "quadruplex head" scheme, in which a rotary drum carrying four heads that scanned a 2-in.-wide tape transversely was employed. One of its drawbacks was that a picture field was divided into 16 tracks making the electronics for reconstructing the field exceedingly complex. The "helical-scan" rotary head was thus invented to solve this problem, first using a 1-in. tape, and subsequent downsizing has continued via $\frac{3}{4}$-in. (U format) to $\frac{1}{2}$-in. (L and M formats).

On the other hand, the consumer-use video tape recorders started with the helical-scan type from the beginning. Although some of the U format were used for some time, they were soon replaced by $\frac{1}{2}$-in. video cassette-tape recorders (VCR) of β format and VHS format. The 8-mm video tape recorder was later developed for portable video cameras.

Digital video tape recorders called D-1 to D-5 have been successively developed for professional use within the past ten years. A number of experimental works have also been performed for the application of the

Table 1. Audio tape recorders.

	Open reel (Open reel)	Compact cassette (Cassette)	R-DAT (Cassette)	Digital micro (Cassette)	DCC (Cassette)
Recording scheme	Analog	Analog	Digital	Digital	Digital
Head type	Stationary	Stationary	Rotary	Rotary	Stationary (MR)
Relative speed	38, 19, 9.5 cm/s	4.76 cm/s	3 m/s		4.76 cm/s
Tape speed	Same	Same	8.15 mm/s	6.35 mm/s	Same
Track width	2.03 cm (mono) 1.05 cm (stereo)	1.6 cm (mono) 0.61 cm (stereo)	15 μm	13 μm	185 μm
Gap length (μm)	2	1.6	0.2		1.6
Tape width (mm)	12.7, 6.3	3.78	3.81	2.5	3.78
Playing time (min)	30 (7-in. reel)	120 (max.)	120 (max.)	120 (max.)	120 (max.)
Kind of tape	γ-Fe$_2$O$_3$	γ-Fe$_2$O$_3$, CrO$_2$, MP[a]	MP[a]	ME[b]	MP[a]
First delivery	1947	1966	1986	1992	1992

---

[a] Metal particle tape (acicular Fe particles).
[b] Metal evaporated tape [Co(Ni)].

digital technology to consumer use, where an 8-mm cassette is expected to be employed. However, in order to make it compatible with high-definition television (HD-TV), a bit rate as high as 130–160 Mbits/s is required, claiming the bit size as small as 1 μm^2, which is almost one order of magnitude smaller than that for D-5 (6 μm^2).

In Tables 2 and 3, the characteristics of typical analog and digital video tape recorders are listed, respectively.

1.2.3 Digital-Data Recording Tape recorders are also used for the recording of digital data. However, the demands for very quick access to the stored data gave birth to a magnetic drum, followed by the magnetic rigid disk that is now employed for the major on-line file memories for almost all computers. The head is mounted on a slider supported by a suspension that is carried by an arm. The slider is designed to float extremely close (a few thousand angstroms) to the

Table 2. Video tape recorders—professional broadcasting/home use (analog).

	2-in. type (Transverse scan)	1-in. type (Helical scan)	U (Helical)	M-II (Helical)	β-cam-sp (Helical)	UNIHI (HD-TV) (Helical)	VHS (Helical)	8-mm video (Helical)
Tape								
Width (mm)	50.8	25.35	19.0	12.65	12.65	12.65	12.56	8.0
Materials	γ-Fe$_2$O$_3$	γ-Fe$_2$O$_3$/Co-γ-Fe$_2$O$_3$	Same	MP	MP	MP	Co-γ-Fe$_2$O$_3$	MP/ME
Coercive field (Oe)	250/270	270/700	500/630	1500	1500	1500	600–700	1500/1050
Head								
Core materials	Sendust	Ferrite	Ferrite	Amorphous	Sendust (sputtered)	Sendust (multilayered)	Ferrite	Sendust (sputtered)
Number of heads	4	1 + sub.	2	2	2	4	2	2
Track width[a] (μm)	250	130	85	Y:44, C:36	Y:86, C:36	24.8	58	20.5
Relative speed (m/s)	39.5	25.59	10.26	7.09	6.9	21.4	5.8	3.8
Playing time (min)	96	180	60	95	94	63	120/540	120
Delivery	1956	1976	1970[b]	1985	1986	1989	1976	1983

---

[a] Y for brightness signals, C for color signals.
[b] Announcement of standardization.

Table 3. Video tape recorders—professional broadcasting (digital).

	D-1 (Transverse scan)	D-2 (Helical scan)	D-3 (Helical)	D-5 (Helical)	1-in. open (HD-1) (Helical)
Tape					
Width (mm)	19.01	19.01	12.65	12.65	25.4
Materials	Co–γ-Fe_2O_3	MP	MP	MP	MP
Coercive field (Oe)	850	1500	1600	1600	1500
Head					
Core materials (write/read)	Ferrite/ferrite	MIG/ferrite	Co-based multilayer/ ferrite	Fe–N thin film	Amorphous/ ferrite
Number of write heads	4	4	4	8	8
Track width (μm)	40	39.1	20	20	37
Relative speed (m/s)	35.6	27.4	21.4	21.3	10.26
Bit length (μm)	0.45	0.43	0.38	0.30	0.35
Playing time (min)	76	208	245	120	90
Delivery	1987	1988	1990	1993	1988

smoothly finished magnetic surface of the disk. The data tracks are recorded circumferentially, and high-speed access to each track is achieved by moving the arm quickly either by a linear or by a rotary actuator bestowed with a precise servoing accuracy. The disks at the early stages were as large as 24-in. in diameter, with their capacities being only 10 MB (megabytes)/disk. Downsizing has also been performed in this field through 14-, 10.8-, 8.3-, 5.25-, 3.5- to 2.5-in., and now even a 1.8-in. disk is also available. Their typical characteristics are listed in Table 4. The rigid disks are in most cases used in a stack of 2 to 10 disks set on a spindle, to each surface of which one or a plurality of heads is assigned. This allows the achievement of much higher access speed for high-capacity file memories than a single disk drive.

The flexible disk called "floppy disk" is another major digital-data storage means. The medium is a flexible plastic disk coated with magnetic paint and is put in a soft jacket or a hard cartridge removable from the disk drive. The head comes in contact with the disk surface, except for the Bernoulli-type floppy-disk drive, and is driven by a step motor. The track density is designed much lower than that for the rigid disk. These design concepts have resulted in low-cost media and drives meeting a demand in the low-end uses. The disk size started as 8-in. followed by 5.25- and 3.5-in, successively, and the capacity has been increased from 0.4 to 2 MB. Even 3.5-in. disks of 4 and 12.5 MB have already been developed. In order to attain more than 20-MB capacity on a 3.5-in. disk, an optical-servoing–type floppy disk called "floptical disk" has been developed, in which the magnetic coating is set upon a PET (polyethylene terephthalate) substrate the surface of which is embossed so as to facilitate optical tracking. Very recently, a 3.5-in. floppy-disk drive of 100 MB in which a flexible disk of ultrathin metal particle coating (several hundred nanometers thick) is employed has been announced. In Table 4, the characteristics of typical floppy disks are also listed.

1.3 Magnetic Heads

A magnetic head must be endowed with "write" and/or "read" functions. The basic structures of write heads are seen in Fig. 1. The longitudinal recording head is essentially composed of a high-permeability magnetic core that forms a ring having a narrow gap and a coil wound around the core, thus being called a "ring head." For the perpendicular recording, on the other hand, a specific head consisting of a main pole with a narrow tip at one end and an auxiliary pole wound by a coil has been proposed, which may best be called a "probe head," although a ring head is also applicable. Since both types of heads are inductive in their nature, they are called "inductive heads." Inductive heads are also used as read heads, but some

Table 4. Rigid disks, floppy disks, and data recorders.

	Rigid disks					Floppy disks				Digital data recorder QIC
	14 in.	8 in.	5.25 in.	3.5 in.	2.5 in.	8 in.	5.25 in.	3.5	Floptical (3.5 in.)	
Medium	γ-Fe_2O_3	Co-Ni-P (plated)	Co-based alloy on Cr	Same	Same	γ-Fe_2O_3	Co-γ-Fe_2O_3	Co-γ-Fe_2O_3	Ba-ferrite	Co-γ-Fe_2O_3
Head[a]	NiFe thin film (ferrite)	Ferrite	NiFe thin film	Fe-N thin film MR	Same MR	Ferrite	Ferrite	Ferrite	Ferrite	NiFe thin film MR
Rotation speed (rpm)	3600 (1500)	3000	5400	5400	3800	360	360	360	720	Tape speed 120 ips
Linear density (kbpi)	15.9 (1)	14	50	87	74.8	6.8	9.9	14.2	17.7	80
Track density (tpi)	2100 (50)	1100	2900	4100	4300	48	96	135	1245	400
Capacity (MB)	7500 (6)	400	3700	1050	344	1.6	1.6/2.0	1.6/2.0	25	10 000
Delivery	1987 (1963)	1981	1991	1993	1993	1977	1981	1985	1991	1994

[a] MR: magnetoresistive.

read-only heads, such as a Hall head and a MR (magnetoresistive) head, have been developed to obtain higher field sensitivity. The Hall* or MR element is set at the front end of a head tip or in a gap of the magnetic core (either the front gap or a gap especially made at a recessed portion of the core in order to provide a wider space for it) or to avoid direct contact of the sensor with the medium.

1.3.1 Inductive Heads In order to obtain a distribution of the head field quantitatively in the vicinity of the head gap or the pole tip, it is generally required to solve a Poisson's equation regarding a vector potential, which can only be solved numerically in most cases (see Fig. 5). However, analytical expressions obtained under some simplified but reasonable assumptions are often useful. For an infinite-permeability ring head having a sufficiently high ratio of the track width to the gap length, the field components H_x and H_y along the center of the track width are well expressed by

$$H_x = \frac{H_0}{\pi}\left(\tan^{-1}\frac{g/2 + x}{y} + \tan^{-1}\frac{g/2 - x}{y}\right), \tag{2a}$$

$$H_y = -\frac{H_0}{2\pi}\ln\frac{(g/2 + x)^2 + y^2}{(g/2 - x)^2 + y^2}, \tag{2b}$$

where the head is supposed to be placed with respect to the coordinate as shown in Fig. 5(a); g is the gap length and H_0 is the field at the center of the gap at $y = 0$. These formulas are well known as the Karlqvist expressions. For the probe head, no such simple expressions have been given, but qualitatively, H_x and H_y of the probe head have the nature of $-H_y$ and H_x, respectively, of the ring head, as is seen in Figs. 6(a) and 6(b) in which they are plotted as functions of x.

The readout voltage e to be obtained by using an inductive head is given, by applying the reciprocity law,† as follows, in terms of

---

*For increasing the recording density, narrowing the gap length is essential. However, the semiconductor film of the Hall element is difficult to make thin enough to put into the narrow gap. Therefore, Hall heads have almost been discarded from present-day magnetic recording devices.

†The reciprocity law between two coils can be applied for the system of the coil of the head and the magnetization in the medium because a magnetic moment is equivalent to a circular current.

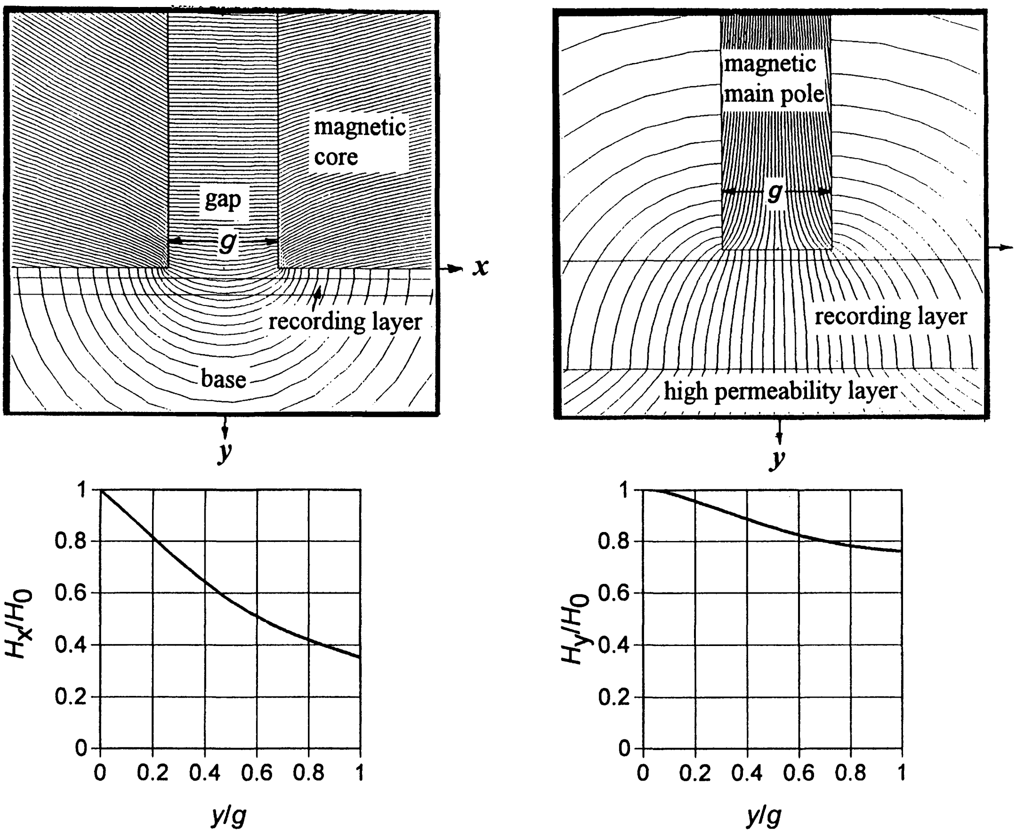

(a) (b)

FIG. 5. Magnetic flux distribution: (a) at the head gap of a ring head and (b) at the pole tip of a probe head, calculated by simulation. The head fields at the center of the gap or the pole tip are also plotted as a function of the distance along the y direction. (Courtesy of Dr. Y. Yoshida and Mrs. F. Akagi, Hitachi.)

the magnetization distribution $M(x,y,z)$ in the medium and the head field distribution $H(x,y,z)$ that would be generated in the medium if a unit current be applied to the coil:

$$e = \mu_0 v \int H(x,y,z) \frac{dM(x,y,z)}{dx} d\tau, \tag{3}$$

where μ_0 is the permeability of the space, v is the relative speed of the head and the medium in the x direction, and the integration is performed for the whole volume of the medium. Since the permeability of the core or the pole materials has a finite value, the efficiency η of the head is generally reduced from that expected for the infinite-permeability case. Thus it is highly required that the core (pole) materials possess a high permeability. The following expression regarding the efficiency explains this situation:

$$\eta = \left[1 + \frac{g/\mu_0 S_g + l_c/\mu S_c}{R_l}\right]^{-1}, \tag{4}$$

where g and S_g are the length and area of the gap, l_c, S_c, and μ are the length, area, and permeability of the core, respectively, and R_l is the reluctance of the flux leakage paths.

The attainable maximum H_0 that appears in Eq. (2a) or Eq. (2b) is proportional to the saturation induction B_s of the pole material, especially in the vicinity of the head gap or the pole tip. Therefore, high B_s is another

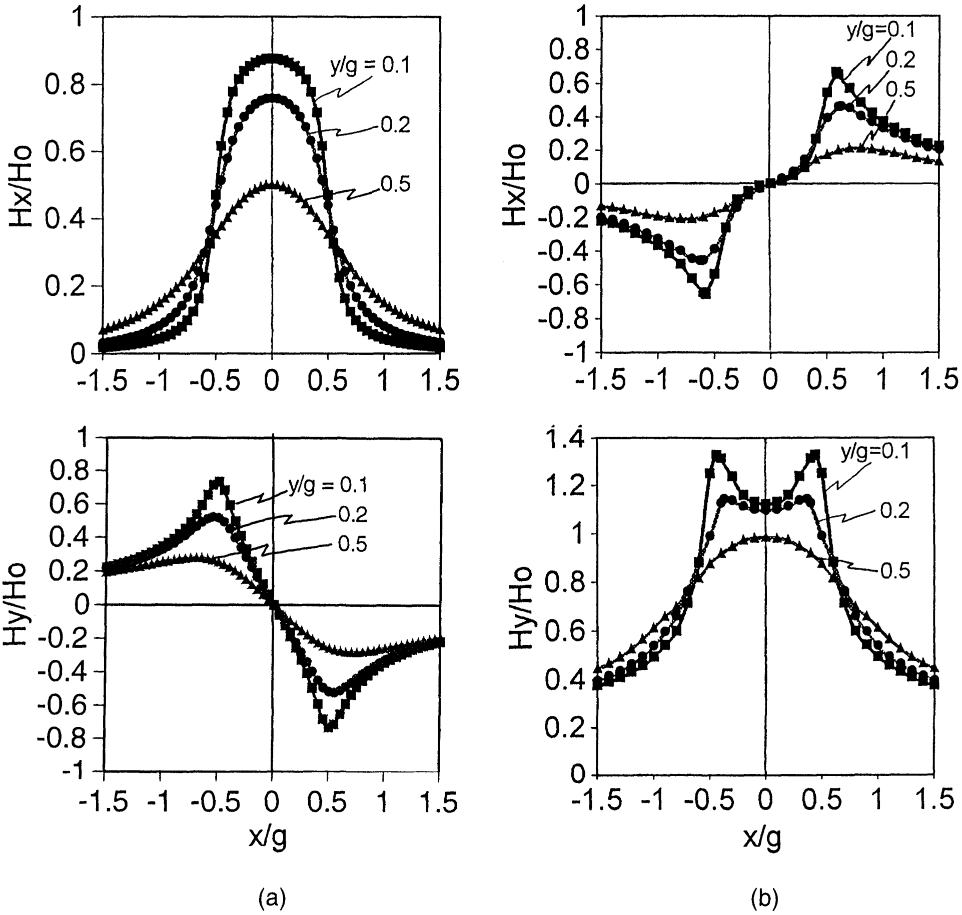

FIG. 6. Head field components H_x and H_y as functions of x (track direction) with the distance from the head surface y as a parameter: (a) of a ring head calculated by use of Karlqvist expressions, (b) of a probe head calculated assuming that the magnetizations at the main pole edge are all equal and parallel to the y direction.

important characteristic that the core/pole materials are required to assume.

Inductive heads are constructed in many ways by using bulk magnetic materials and/or thin films. In Fig. 7, the structures of typical heads are schematically illustrated. The audio heads are mostly composed of laminated Permalloy cores, the layers of which are insulated from each other to avoid eddy-current losses at high frequencies. The video heads for professional use were first composed of two parts, the back core and the front core. In order to match the video frequencies, sintered ferrite was chosen for the back core. For the front core, a mechanically hard Al–Fe–Si alloy called "Alfesil" or "Sendust"* was chosen because of its excellent precise machinability and abrasion durability. When the VCRs for consumer use were developed, single-crystalline Mn–Zn ferrites were available and were utilized as the core materials. The choice of specific crystalline orientations has been confirmed essential for obtaining both high read/write performances and a high wear resistance (Fujiwara *et al.*, 1982). A typical choice of the crystalline ori-

---

*This name came from the fact that an Al–Fe–Si alloy with a composition of Fe with Al 12%, Si 5%, which shows especially high permeability, was discovered in Tohoku University in Sendai (Japan), and was applied to making "dust" cores.

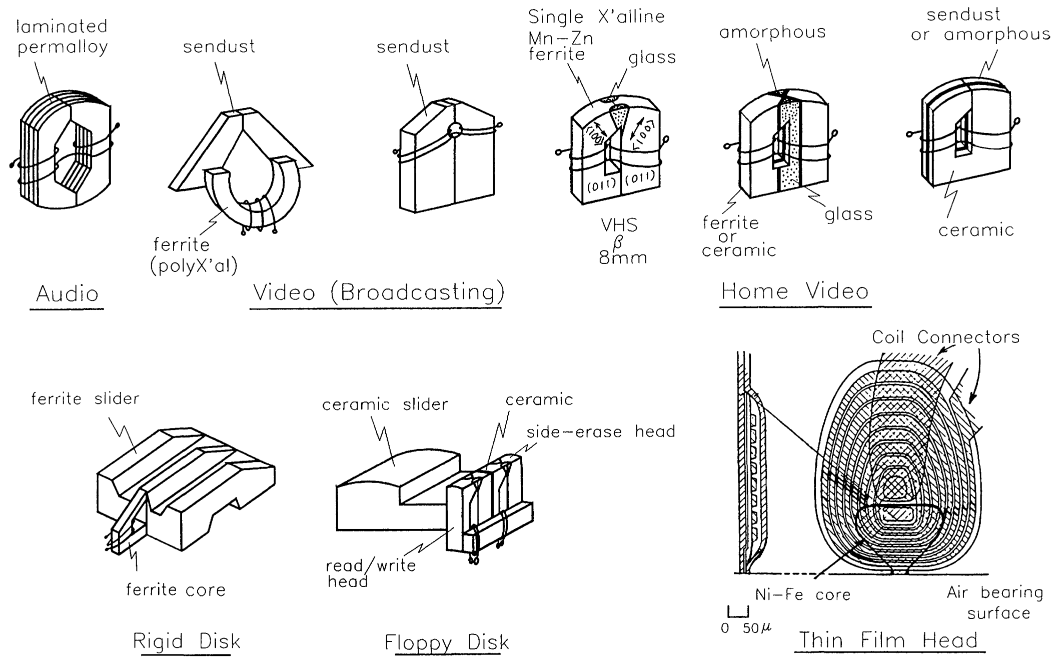
laminated permalloy
sendust
sendust
Single X'alline Mn–Zn ferrite
glass
(100)
(100)
(011)
(011)
amorphous
sendust or amorphous
ferrite (polyX'al)
VHS
β
8mm
ferrite or ceramic
glass
ceramic
Audio
Video (Broadcasting)
Home Video
Coil Connectors
ferrite slider
ceramic slider
ceramic
side-erase head
ferrite core
read/write head
Rigid Disk
Floppy Disk
Ni–Fe core
Air bearing surface
0 50μ
Thin Film Head

FIG. 7. Schematic illustration of typical magnetic heads (thin-film head: after Mee and Daniel, 1989).

entations is shown in Fig. 7. For rigid disk heads, polycrystalline Ni–Zn ferrites had long been used. This was because Ni–Zn ferrite had been believed to possess better characteristics at high frequencies than Mn–Zn ferrite on account of its very high resistivity. However, it was later confirmed that the machining effect improves the high-frequency characteristics of Mn–Zn ferrite and its higher saturation induction is more beneficial than Ni–Zn ferrite. The head cores are imbedded in a slider designed to float upon the disk surface hydraulically, the slider itself often making a part of the magnetic flux path. Floppy-disk heads are usually made of sintered polycrystalline ferrite and are imbedded in a slider.

In order to match the higher coercivity of the medium, heads having a metal gap have been developed for video tape recorders (see Fig. 7). Some of them are called MIG (metal-in-gap) or composite head as the core is mostly constructed by a ferrite and the gap portion is made of magnetic metals such as Permalloy, Sendust, and amorphous magnetic materials.

The thin-film head was introduced into disk systems first. Its chief merits are in the simplicity of its manufacturing process beneficial to mass production and in the characteristics of high permeability at high frequencies and high B_s (saturation induction). A typical structure of the thin-film head is shown in Fig. 7. Permalloy (a Ni–Fe alloy) is usually employed as a core material. The patterning of the head core and coils is performed by utilizing various fine lithography methods such as photolitography, electron lithography, and ion milling. Magnetic domain-wall stabilization or elimination is one of the principal concerns, which makes it possible to avoid the Barkhausen noise.

The demand for higher recording density, in which very high coercivity is to be employed, has stimulated the development of high-B_s head core materials. Fe-based alloys had been known as high-B_s materials, but their high crystalline anisotropy was preventing them from becoming magnetically soft enough to be employed as head core materials. The control of the grain size small enough to average out the crystalline anisotropy, which is recently referred to as nanocrystalline technology, was the breakthrough for that. This was first tried by multilayering and then by adding nitrogen and other ele-

Table 5. Head materials.

Material	Composition (wt%)	μ (dc/5 MHz)	H_c [Oe (A/m)]	B_s [kG (T)]	ρ ($\mu\Omega$ cm)	Hardness, Vickers
Mo Permalloy	4% Mo, 17% Fe, 79% Ni	11 000/–	0.025 (2.0)	8 (0.8)	50	120
Hardperm	9% Nb, 12% Fe, 79% Ni	8 000/–	0.005 (0.4)	6 (0.6)	80	350
Sendust	5.4% Al, 9.6% Si, 85% Fe	8 000/–	0.025 (2.0)	10 (1.0)	85	480
Amorphous	15% Si, 10% B, 5% Fe, 70% Co	80 000/500	0.006 (0.5)	6.6 (0.66)	134	910
	7% Mo, 8% Zr, 85% Co	–/620	0.6 (50)	10 (1.0)		1000
MnZn ferrite (single crystal)	30% MnO, 20% ZnO, 50% Fe_2O_3	4 000/500	0.05 (4)	3.8 (0.38)	10^3	640
MnZn ferrite (polycrystal)	19% MnO, 11% ZnO, 70% Fe_2O_3	3 000/800	0.1 (8)	4.8 (0.48)	10^4	650
NiZn ferrite (polycrystal)	11% NiO, 22% ZnO, 67% Fe_2O_3	850/550	0.4 (32)	3.9 (0.39)	10^7	600

ments that prevent excessive grain growth. In Table 5, properties of typical head core materials are listed.

1.3.2 Magnetoresistive (MR) Heads

The key element in MR heads is a MR element that consists of a narrow magnetoresistive thin-film stripe having a uniaxial magnetic anisotropy and current leads connected at each end of the stripe through which a current is supplied (see Fig. 8). The resistivity ρ changes as a function of the angle θ be-

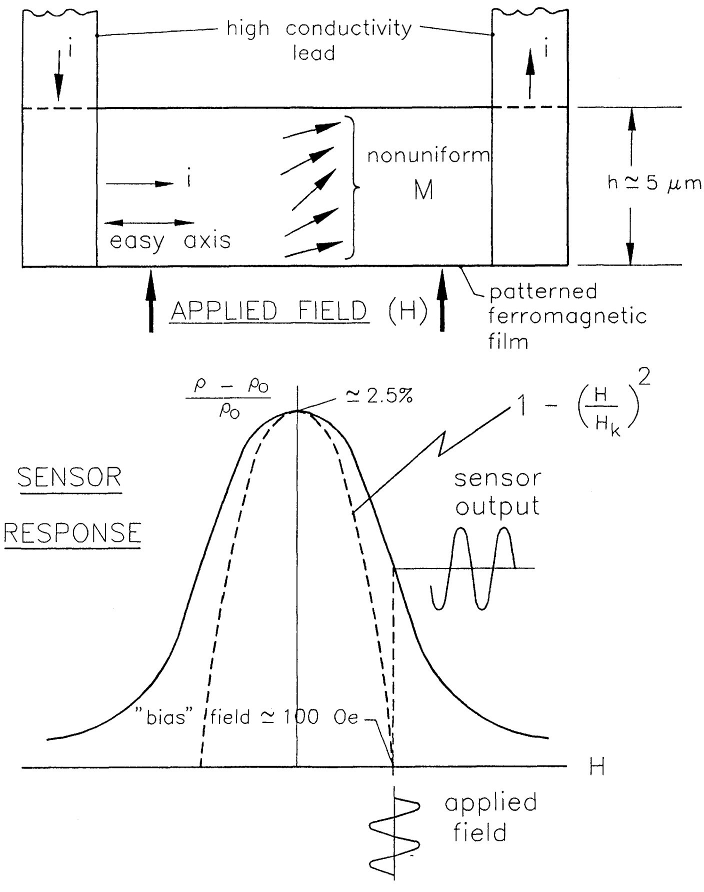
high conductivity lead
i
i
i
nonuniform M
easy axis
h ≃ 5 μm
APPLIED FIELD (H)
patterned ferromagnetic film
$\frac{\rho - \rho_0}{\rho_0}$
≃ 2.5%
$1 - \left(\frac{H}{H_k}\right)^2$
SENSOR RESPONSE
sensor output
"bias" field ≃ 100 Oe
H
applied field

FIG. 8. Response of the MR head to the applied field (after Markham and Jeffers, 1990).

tween the current and the magnetization of the element, according to the expression

$$\rho = \rho_0 + \Delta\rho \cos^2\theta, \tag{5}$$

where ρ_0 and $\Delta\rho$ depend on the material of the element. In most cases the easy axis of the element is set parallel to the stripes. Therefore ρ is expected to change quadratically with the field H as

$$\rho = \rho_0 + \Delta\rho[1 - (H/H_k)^2], \tag{6}$$

since

$$\sin\theta = H/H_k, \tag{7}$$

with H_k denoting the anisotropy field. However, because of the demagnetization field the magnetizations lying close to the edge do not deflect so much as the center, giving rise to the characteristic inflection points and the skirts in the response curve as is illustrated in Fig. 8 (Markham and Jeffers, 1990). Thus, in order to obtain a linear response, various kinds of biasing methods have been explored, typical examples of which are shown schematically in Fig. 9. The elimination of Barkhausen noise is another important issue in practical use. Therefore, some measure has to be taken to avoid the generation of reverse domains at the edges, such as putting hard magnetic or antiferromagnetic films at the ends of the element or making the element double-layered, etc.

A magnetic shield is usually used to obtain a fine resolution as shown in Fig. 10. The reciprocal relation in this case is formally expressed as

$$e \sim \phi = \mu_0 \int \mathbf{H}(x,y,z) \cdot \mathbf{M}(x,y,z) d\tau, \tag{8}$$

where ϕ is the flux entering the MR element and $\mathbf{H}(x,y,z)$ is the field distribution in the space when a unit current is applied to an imaginary coil supposed to be wound around the MR element. Analytical expressions for $\mathbf{H}(x,y,z)$ can be obtained for the infinite track width almost in the same way as the Karlqvist expressions:

$$H_x = H_{rx}(x + (b + t)/4, y) - H_{rx}(x - (b + t)/4, y), \tag{9a}$$

and

$$H_y = H_{ry}(x + (b + t)/4, y) - H_{ry}(x - (b + t)/4, y), \tag{9b}$$

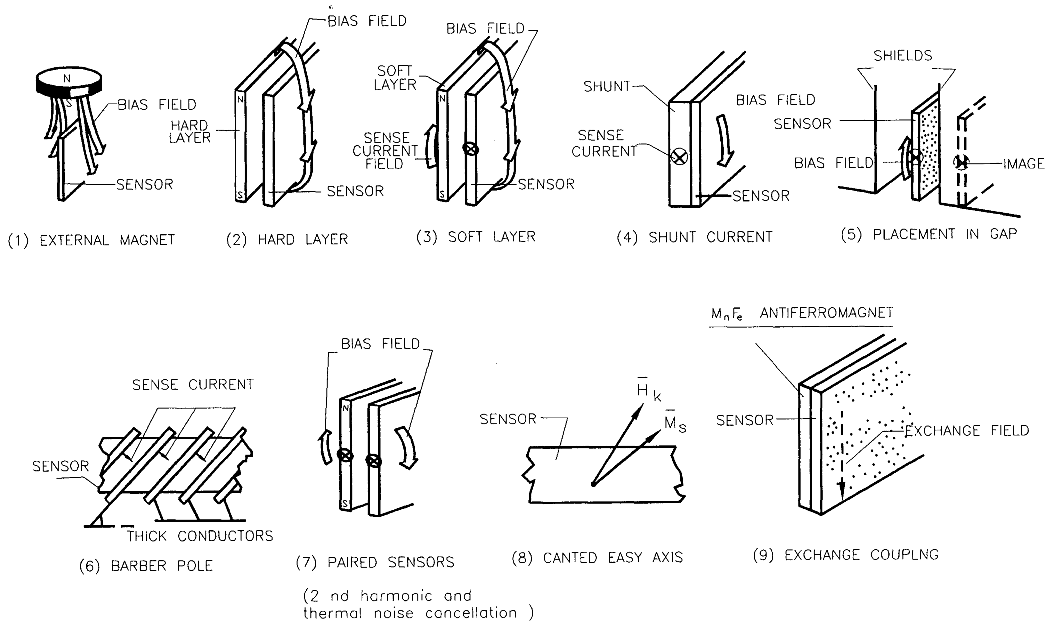
BIAS FIELD
N
S
HARD LAYER
SENSOR
BIAS FIELD
N
S
SENSOR
BIAS FIELD
SOFT LAYER
SENSE CURRENT FIELD
SENSOR
SHUNT
SENSE CURRENT
BIAS FIELD
SENSOR
SHIELDS
SENSOR
BIAS FIELD
IMAGE
SENSE CURRENT
SENSOR
THICK CONDUCTORS
BIAS FIELD
N
S
SENSOR
$\bar{H}_k$
$\bar{M}_s$
MnFe ANTIFERROMAGNET
SENSOR
EXCHANGE FIELD

(1) EXTERNAL MAGNET (2) HARD LAYER (3) SOFT LAYER (4) SHUNT CURRENT (5) PLACEMENT IN GAP

(6) BARBER POLE (7) PAIRED SENSORS (2 nd harmonic and thermal noise cancellation) (8) CANTED EASY AXIS (9) EXCHANGE COUPLNG

FIG. 9. Various biasing methods for MR heads (Jeffers, 1986; © 1986 IEEE).

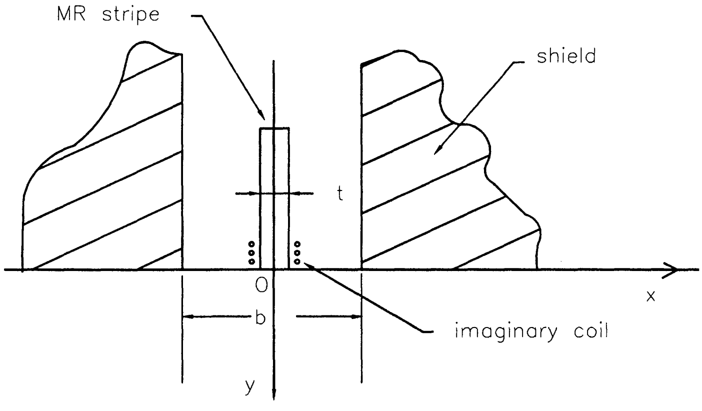

FIG. 10. Cross section of a shielded MR head, showing the coil that is imagined to exist for the purpose of computing the sensitivity function (after Potter, 1974).

where by H_{rx} and H_{ry} the Karlqvist's expressions for the ring head are denoted, with b and t denoting the gap length between the shield cores and the thickness of the MR element, respectively. The problem is, however, that the relation between e and ϕ is not usually simple. Therefore, the theoretical evaluation of the sensitivity requires some numerical calculations.

For MR heads, a high magnetoresistive ratio ($\Delta\rho/\rho_0$) is one of the basic requirements for the material. Although Permalloy has $\Delta\rho/\rho_0$ of only 2% to 3%, it has been the only material employed for practical use. The ρ_0 of Permalloy is about 20 $\mu\Omega$ cm and the applicable current density is almost 10^7 A/cm^2, leading to the maximum electric field of about 200 V/cm. This implies that for a track width of 10 μm, the signal output of 2–3 mV can be obtained taking into account the reduction of the usable amplitude to half of the maximum obtainable amplitude because of the nonlinearity in the transfer curve (Markham and Jeffers, 1990). This is almost an order of magnitude higher than the output obtainable by the inductive head at a bit density of 60 kfci (fci: flux changes per inch) and a relative speed of 1 m/s (MR heads are relatively speed insensitive by their nature). Some Ni–Co alloys show 6% to 7% $\Delta\rho/\rho_0$ ratio, but because of their high crystalline anisotropy and high magnetostriction, they have not been made applicable to magnetic heads.

In 1988, a new phenomenon called "the giant magnetoresistive (GMR) effect" was discovered in antiferromagnetically coupled multilayered films composed of a number of magnetic layers with nonmagnetic layers in between, each having the thickness of a few nanometers (Baibich *et al.*, 1988). This phenomenon comes from the spin-dependent resistivity at the interface and in the bulk of the magnetic layers and is isotropic with respect to the current direction, and is expressed as

$$\rho = \rho_0 + \Delta\rho \cos\phi, \qquad (10)$$

where ϕ is the angle between the magnetizations of the adjacent layers, in contrast to the case for the conventional MR, which is now called anisotropic MR because of its dependence on the angle of the magnetization and the direction of the current.

The GMR was first found in Fe/Cr systems at very low temperature, but soon it was observed even at room temperature in Co/Cu systems, showing MR ratios as high as 65% (Parkin *et al.*, 1991). At first, to obtain this high ratio, a magnetic field of the order of 10 000 Oe was needed, but a lot of efforts have made it possible to reduce the field drastically. Recently, a structure that gives a substantial sensitivity in fields as low as several oersteds has been proposed in which a granular nature is introduced into the magnetic layers (Hylton *et al.*, 1993). It has been made clear that the antiferromagnetic coupling between the adjacent layers is not essential to the GMR effect. Thus, any metallic magnetic bilayer structures in which the magnetization of one layer is pinned and that of the other is free exhibit the GMR effect (Shinjo and Yamamoto, 1990; Dieny *et al.*, 1991). They are now called spin valves and are expected to be applied for read heads before long.

1.4 Magnetic Recording Media

It may safely be said that the history of magnetic recording media is finding ways of increasing the recording density. This is especially true with digital recording. For longitudinal recording, assuming the magnetization function at the bit transition with a transition length a_T as

$$M_x(x) = (2/\pi)M_r \tan^{-1}(x/a_T), \tag{11}$$

where x refers to the track direction, the output voltage with use of a ring head is given by

$$e_x(x) = -2\mu_0 v w M_r t H_x(x, a_T + h), \tag{12}$$

where v is the relative velocity of the head and medium, w is the track width, M_r is the saturation remanence of the medium, and H_x is given by Eq. (2a) (h: head-to-medium spacing). Then the output pulse width p_{50} (the width at 50% of peak amplitude) is given by

$$p_{50} = 2[(g/2)^2 + (a_T + h)(a_T + h + t)]^{1/2}, \tag{13}$$

where t is the thickness of the medium. This tells that a small gap g, spacing h, thickness t, and transition width a_T are essential to obtain a high resolution. Among those factors, a_T, t, and h in part are related to the medium properties. It should be noted that the above expressions are derived by assuming that the recording is perfect from the top to the bottom, which is the case only for very thin-film media relative to g. A lot of experiments have been performed and an empirical relationship between p_{50} and M_r, H_c, and t of the medium has been obtained for the longitudinal recording as follows:

$$p_{50} \sim (M_r/H_c)^a t^b, \quad a = 0.3\text{–}0.5, \quad b = 0.3\text{–}0.8, \tag{14a}$$

while Nakamura (1991) has obtained by simulation

$$p_{50} \sim (M_r/H_c)^{0.4} t^{0.4}. \tag{14b}$$

Thus, it is concluded that in order to obtain a high resolution, H_c of the medium should be made as high as possible within the limit of the writing capability of the head, although some compromise is necessary between the resolution and output in terms of M_r and t. The surface roughness is another critical factor that should be minimized from the viewpoint of both output and resolution.

The magnetic characteristics of the medium are mostly expressed by its B–H curves, schematic illustration of which was given in Fig. 2. Other than M_r and H_c, the squareness ratio, which is defined as shown in the figure, is also critical to the resolution.

Noise reduction is another important issue. The key point is to reduce the factors that cause inhomogeneity in the medium. Among them, the reduction of the particle size (or the effective size of magnetization switching unit), their homogeneous dispersion, and the smoothing of the surface are most critical.

In accord with the guidelines stated above, varieties of recording media have been developed and brought into practical use. Historically, the first medium was a steel wire, meanwhile replaced by tapes constituted of a plastic base and iron-oxide particles coated upon it with an organic binder. A plated medium was introduced for drum devices but they were soon replaced by rigid disks of particulate coating and now thin-film media are beginning to replace them, starting with disks such as 5.25-in., 3.5-in., or smaller-sized ones. Floppy disks are again of particulate coating.

1.4.1 Particulate Media A particulate medium is basically composed of a base, with a magnetic coating and lubricant coated upon it. As base materials, biaxially oriented polyethylene terephthalate (PET), in some cases with some fillers embedded to reduce the friction coefficient, is mostly employed for flexible media, and mirror-finished Al–Mg(4%–5%) alloys for rigid disks. Other than these, polyimides, polyamides, hard glass, ceramics, etc., are now under development. For most particulate media, acicular magnetic particles are used on account of their high coercivity. For tapes and disks, the particles are aligned in the track direction by applying a magnetic field during the coating process in order to give a higher squareness ratio, while in floppy disks, they are not. This is because the floppy disks are produced by punching out of a wide plastic sheet with a magnetic coating on it. As a lubricant, fluoro resins such as perfluoropolyether are used. The typical thickness of the coatings is a few micrometers for flexible media and several hundred nanometers for rigid disks, respectively, with some exceptions such as double-coated media in which less than 300-nm thickness has been attained for flexible media (Inaba *et al.*, 1995).

The magnetic paint (or dispersion) for coating is prepared by mixing the magnetic powder, a small amount of nonmagnetic powder (α-Al_2O_3, Cr_2O_3, etc., for strengthening the durability, graphite for preventing electrostatic charging), an organic binder, and some organic solvent by use of a sand mill and/or a kneader, though some waterborne paints are under development from the environmental point of view. The binder is designed to be composed of hydrophilic and hydrophobic segments so that it may prevent the particles from aggregation. Figure 11 illustrates the function of the binder. The hydrophilic segments work as an anchor and the hydrophobic ones as a repeller. For flexible media, thermoplastic compositions such as vinyl resin and/or urethane resin are employed, while for rigid disks, thermosetting ones such as epoxy resins or phenolic resins are used. Those have been chosen to facilitate each coating process: roll coating for flexible media and spin coating for rigid disks. The volume fraction of the magnetic particles in the coatings is limited to not much greater than 35% (25% for rigid disks) in order to endow the media with sufficient durability and smooth surfaces.

Magnetic particles in practical use are listed in Table 6. Acicular γ-Fe_2O_3 particles elongated in the [110] direction are prepared according to the processes shown in Fig. 12. The magnetic anisotropy of the particles thus prepared is the sum of the crystalline and shape anisotropies, the contribution of each being almost of the same order. The dehydration and reduction involved in the process are accompanied by a substantial material transport resulting in a more or less irregular shape containing many pores, which leads to a lower coercivity than expected. In order to overcome this drawback, a new process has been developed in which acicular γ-Fe_2O_3 particles are directly precipitated without passing through α-FeOOH (goethite). Thus their coercivity ranges from 250 to 450 Oe according to their preparation methods. CrO_2 is synthesized by hydrothermal synthesis (~400 °C) from Cr_2O_5 by using seeds having a rutile structure. The particles are mostly a single crystal of well-defined shape, which is beneficent for obtaining a coercivity as high as 700 Oe. However, there was a drawback to CrO_2 that toxic hexavalent Cr ions were to be treated during the preparation process. For its replacement, Co-

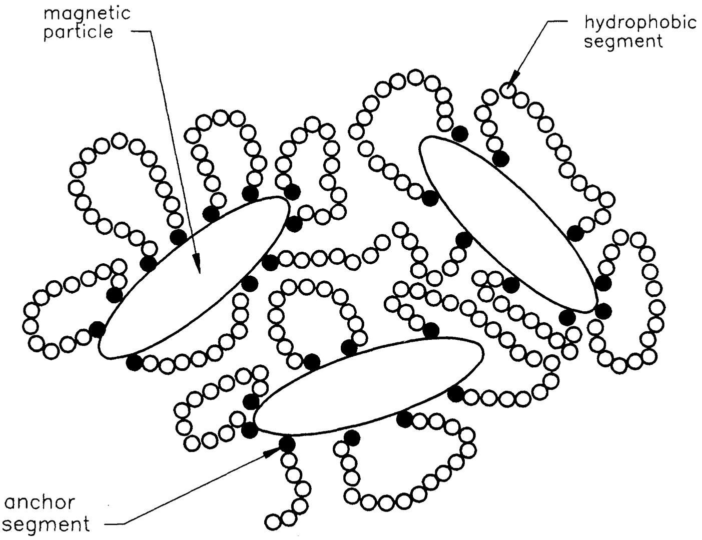

FIG. 11. Schematic model of the function of binder. The binder is composed of hydrophobic segments and hydrophilic segments located at the ends. The hydrophilic segments act as anchors to the magnetic particles.

Table 6. Magnetic particles (after Mee and Daniel, 1989).

Material	σ_s[a] (emu/g = A m^2/kg)	ρ (g/cm^3)	H$_c$ [Oe (kA/m)]	$\langle v \rangle_{av}$ (10^{-4} μm^3)	SSA[b] (m^2/g)
α-Fe	125–170	5.8	940–1600 (75–130)	<1	35–60
Fe(70 at %)Co(30 at %)	140–190	5.8	1100–2000 (90–160)	<1	35–60
CrO_2	76–84	4.8	380–730 (30–58)	0.8–10	18–40
γ-Fe_2O_3	73–75	4.8	250–450 (20–36)	1–25	12–35
Co-γ-Fe_2O_3	70–75	4.8	380–940 (30–75)	0.7–10	16–50
$BaFe_{2-2x}Co_xTi_xO_{19}$					
$x = 0.7$	58	5.3	1200 (95)	<1	20–55
$x = 0.8$	58	5.3	700 (56)	<1	20–55

---

[a] Saturation moment per gram.
[b] Specific surface area.

modified γ-Fe_2O_3 has been developed. At first, replacement of a fraction of the Fe^{2+} by Co^{2+} was tried, which was expected to increase the crystalline anisotropy substantially resulting in an increase in coercivity. However, because of the drastic temperature dependence of its anisotropy, this trial was not successful. Instead, methods in which only the surface of the particle is modified, including epitaxial growth of Co-ferrite upon the surface, have been developed. Thus, coercivity as high as 900 Oe has become available. Figure 13 demonstrates the epitaxially grown Co-ferrite upon the surface of a γ-Fe_2O_3 particle. Acicular α-Fe particles were developed to obtain higher B_s particles than oxides. They are now being produced by the reduction of acicular precursors of α-FeOOH or γ-Fe_2O_3. In order to obtain chemical stability, the surfaces are passivated by mild oxidation, and some stable oxides such as Al_2O_3 and SiO_2 are coated, although the recently developed amine-quinone polymer

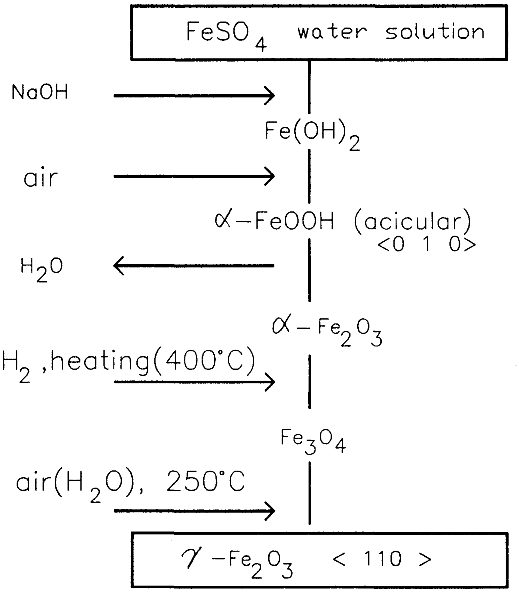

FIG. 12. Typical production process of acicular γ-Fe_2O_3 particles.

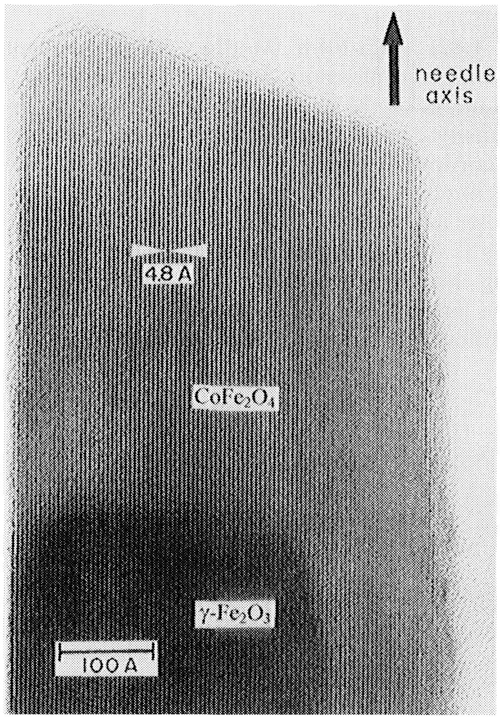

FIG. 13. Lattice image of a Co (epitaxial)–γ-Fe_2O_3 particle taken by a high-resolution electron microscope (courtesy of Hitachi Maxell).

binder may relieve the necessity of protection coating through their corrosion-protecting capability (Nikles *et al.*, 1994). Typical magnetic properties are seen in Table 6. Barium ferrite ($BaFe_{12}O_{19}$) has a hexagonal crystalline structure and has a strong uniaxial anisotropy (3.3×10^5 ergs/cm^3) with its easy axis parallel to the *c* axis. This high anisotropy results in a very high coercivity (~2000 Oe), and as a material for magnetic recording, the use was limited up to now only to master tapes or magnetic cards for specific applications. However, the idea of its application to perpendicular recording promoted its adaptability to more general use. The primary concern was the reduction of the coercivity, although it is estimated that coercivity as high as 3000 Oe will be necessary for future high-density recording. The reduction was accomplished by replacing some fraction of Ba or Fe ions by divalent and tetravalent elements such as Co–Ti or Zn–Ti. The particles are prepared either by precipitation in glass flux or by hydrothermal synthesis.

Figure 14 shows electron microscope images of typical magnetic particles mentioned above.

1.4.2 Thin-Film Media As is evident from the expressions (14a) and (14b) for p_{50}, thinness of the medium is essential to high-density recording especially for longitudinal recording. This is the reason why thin-film technology has received strong attention. Another advantage inherent to the thin-film media is the binderlessness that enables achieving higher saturation magnetizations than the particulate ones. However, obtaining a sufficiently high coercivity in thin films was the primary challenge, which was eventually achieved by utilizing crystalline anisotropy, with the aid of shape anisotropy in some cases, of the grains constituting the films. Thus, in most of the thin-film media, cobalt-based alloys having a hexagonal structure, in which a high uniaxial crystalline anisotropy is inherent, have been used, although there have been some exceptions such as gamma ferric oxide and barium ferrite. In thin magnetic films, zigzag walls as shown in Fig. 15 are apt to be formed at the recorded bit boundaries where the magnetizations point head to head (or tail to tail) against each other, reducing the magnetostatic energy at the expense of exchange and anisotropy energies. Their amplitude should be made as small as possible in recording. This can be attained by the reduction of the exchange interaction between the grains. Thus, by taking advantage of the ease of obtaining much smaller grains than in the particulate coatings, very low-noise thin-film media are now available. Protection from abrasion or mechanical damage is also important, a typical example of which is a carbon coating of the order of 10 nm thick.

Thin-film media are produced by plating, sputtering, or evaporation, the appropriate one of which is chosen according to its application: for tapes, evaporation because of its mass productivity and the ease of obtaining a high coercivity and high squareness ratio along the tape direction by making use of the oblique incident evaporation method, and for rigid disks, plating or sputtering because of its ease in producing films with isotropic high coercivity (in some cases, circumferential anisotropy is attached to the

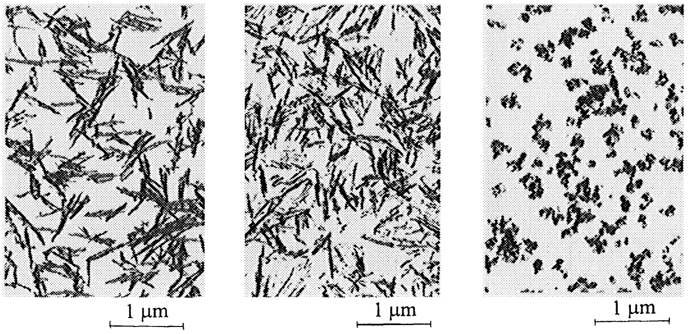

(a) Co-γ-Fe_2O_3 (b) α-Fe (c) $BaFe_2O_{19}$

FIG. 14. Electron micrographs of typical magnetic particles for recording (courtesy of Hitachi Maxell).

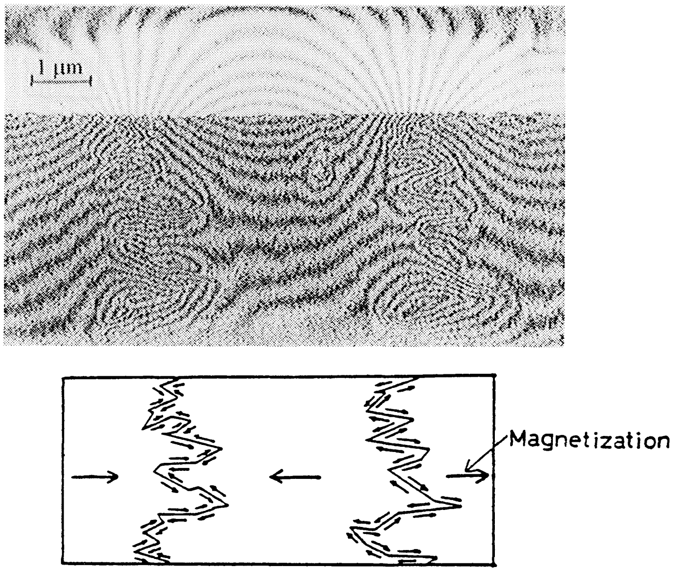

FIG. 15. Zigzag walls at the transition of the magnetization of a Co film observed by electron holography (after Yoshida *et al.*, 1983).

disk by making scratched trails on the substrate in that direction). In Table 7, the properties of some typical thin-film disks are also listed.

Plated disks are produced by electroplating or electroless plating of Co–Ni(P) upon an Al-alloy substrate covered with an electroless-plated nonmagnetic Ni(P) layer. The Ni(P) layer here facilitates obtaining a smooth finish by virtue of its hardness. The coercivity can be controlled by the composition of Co–Ni, additives, the *p*H of the plating solution, and the current density in the case of electroplating. For films having a rel-

Table 7. Properties of magnetic recording media.

Medium	H_c [Oe (kA/m)]	B_s (kG = 0.1 T)	B_r/B_s	Thickness (μm)	Areal density (Mb/in.^{2})
Tapes					
Audio cassette tape					
γ-Fe_2O_3	360–410 (28–33)	1.7–1.9	0.85–0.91	5	
CrO_2	450–650 (36–52)	1.6–2.3			
Co–γ-Fe_2O_3	610–690 (49–55)	1.7–1.8	0.89–0.91	5	
MP (α-Fe)	~1200 (95)	~3.6	0.91	4	
Video tape					
Co–γ-Fe_2O_3	650–900 (52–72)	1.7–2.0	0.82–0.88	3.5–4.5	17
MP (8 mm)	1550 (122)	3.0	0.83–0.85	2.5	90
Double-coated MP (8 mm)	1550 (122)	4.0	0.82	0.25 2.5 (nonmag.)	
ME (8 mm)	910–1400 (72–110)	4.6–6.4	0.6–0.78	0.17–0.24	128
Rigid disks					
γ-Fe_2O_3	330 (26)	1.3	0.85	0.34	
Co (Ni) (plated)	790 (63)	2.2	0.8	0.055	40
	600 (48)	8.8	0.8	0.08	15
γ-Fe_2O_3 (sputtered)	700 (56)	8	0.8	0.16	40
Co(Ta)/Cr (sputtered)	1300 (103)	9	0.8	0.05	150
Co(Pt)/Cr (sputtered)	1800 (173)	10	0.85	0.045	235
Floppy disks					
γ-Fe_2O_3	300 (24)	1.2	0.57	2.5	0.054
MP	1550 (122)	2.5	0.6	2.5	1.5
Ba-ferrite	700 (56)	1.25	0.47	0.6	2.8

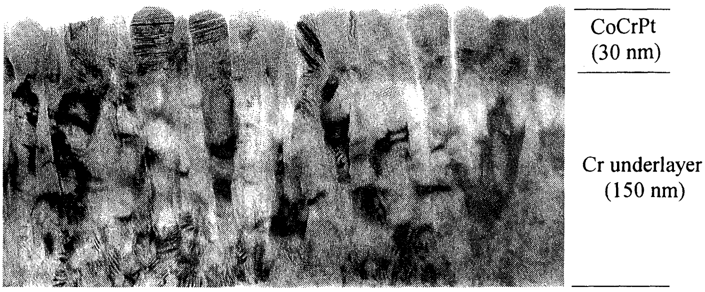

FIG. 16. Cross-sectional view of a typical Co-based–alloy sputtered disk (courtesy of Dr. M. Futamoto, Hitachi).

atively high coercivity it has been confirmed that they have an hcp structure with its *c* axis preferentially aligned in the in-plane direction forming a networklike structure or that they are composed of a mixture of hcp and fcc structures. It has been also made clear that the control of *p*H and addition of some elements such as Zn and Mn facilitate the reduction of grain sizes, leading to obtaining low-noise media (Goto and Osaka, 1993).

The γ-Fe_2O_3 sputtered disk is mostly made by reactive sputtering of Fe in an Ar–O_2–H_2 atmosphere to obtain a Fe_3O_4 film and then heat-treating it in H_2. Addition of Co increases the crystalline anisotropy. In some films of the media that show good high-density recording characteristics, it has been confirmed that the films are composed of two kinds of columns (several tens of nanometers in diameter); one is of a single crystal with the (111) plane oriented parallel to the plane of the substrate and the other is an assembly of fine grains separated by amorphous materials (Takagaki *et al.*, 1987). Co-based–alloy sputtered disks are made by depositing the alloy upon a polycrystalline Cr underlayer mostly with the (110) or (100) planes preferentially oriented parallel to the substrate surface, upon which Co alloys are epitaxially grown resulting in the *c* axis of the Co-alloy grains oriented parallel to or at a small angle with the disk surface. Circumferential trails are sometimes introduced to obtain a circumferential easy-axis orientation. As Co-based alloys, first CoNi and then CoCr, CoCrTa, and CoCrPt have been explored to obtain a higher and higher coercivity. It is now understood that Cr and Ta are likely to precipitate at the grain boundaries and give a desirable effect of the reduction of the exchange coupling between the grains. Sputtering conditions that give a low mobility to the deposited atoms upon the substrate, such as high Ar pressure, low substrate temperature, and no biasing, are also effective to reduce the noise. Figure 16 shows a cross-sectional view of a typical Co-based–alloy sputtered disk. By utilizing Co-based thin-film media, recording densities of 1–3 Gb/in.^{2} have been successfully demonstrated.

Evaporated media are prepared by such a method as is illustrated schematically in Fig. 17. A Co(Ni) alloy is electron-beam deposited obliquely onto a 6–20-μm-thick PET base film running at a speed of 10–20 m/min over a rotating drum. The incident angle of the depositing atom beams is controlled by a shutter. During the deposition, air is blown

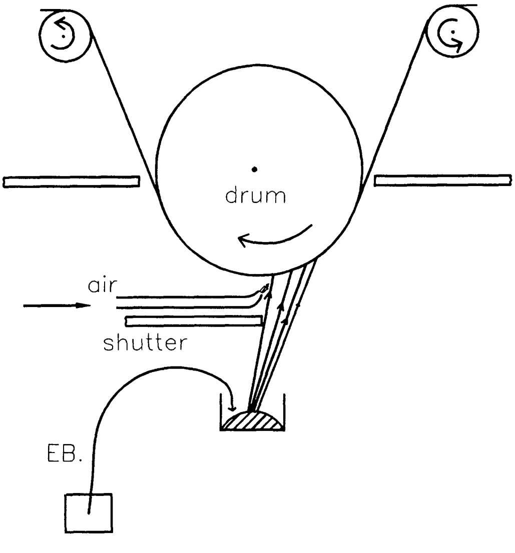

FIG. 17. Schematic illustration of the production process of evaporated metal (ME) tapes.

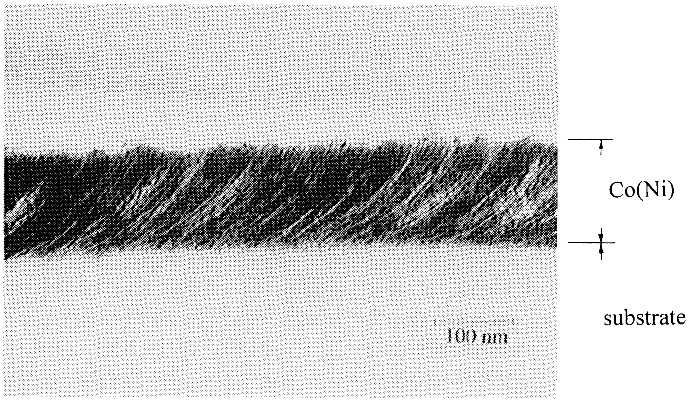

FIG. 18. Electron micrograph of a cross section of an evaporated metal (ME) tape (courtesy of Hitachi Maxell).

onto the substrate in order to form an oxide around each grain growing on the substrate. Figure 18 shows a cross section of a typical evaporated tape (ME tape). The oxide acts as a reducer or inhibitor of the exchange coupling between the grains and a protector against both corrosion and abrasion. The inclination of the grain growth with respect to the substrate surface makes the role of the perpendicular component of the magnetization greater, especially in high–recording-density regions. The optimization of the inclination angle is one of the present primary concerns.

1.4.3 Perpendicular-Recording Media

Investigations on perpendicular recording were first started by using CoCr films (Iwasaki and Nakamura, 1977). Although various other Co-based alloys (CoV, CoMo, CoW, etc.) and barium ferrite have been explored, no better recording characteristics have been obtained than with CoCr media. One of the important criteria the media should satisfy is that the quality factor Q, which is defined as the ratio $K_{\perp}/2\pi M_s^2$, be almost equal to or greater than unity, with $K_{\perp}$ denoting the intrinsic perpendicular anisotropy. For CoCr, this criterion can be satisfied in the Cr-content region greater than 15 at. %, with the limit at about 25% at which M_s tends to vanish. CoCr films with a definite perpendicular anisotropy can be formed either by sputtering or by evaporation. They are composed of very fine columnar crystallites of hcp structure with their c axes aligned almost perpendicular to the film surface. Morphological analysis has revealed that each column has a microstructure such that there exist Co-rich areas and Cr-rich areas if observed from the top of the film, an example of which is shown in Fig. 19, which facilitates high-density recording. Co-based–alloy films with a good quality factor can also be made by electroless plating in a Co–Ni–Zn–P bath with a relatively high concentration of $NiSO_4$ compared with that for the longitudinal recording media (Goto and Osaka, 1993). Barium-ferrite films have been reported successfully prepared on ceramic substrates by the sputtering/heat-treating method. A drawback of this method is that they require a high-temperature (almost 500°C or higher) heat treatment, although some breakthrough is being developed by using laser flash annealing.

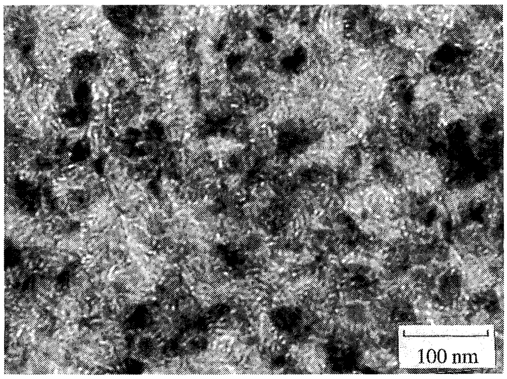

FIG. 19. Transmission electron micrograph of an etched Co–Cr film. Bright area, Co-rich; dark area, Cr-rich (courtesy of Drs. Y. Maeda and S. Hirono, NTT).

In order to match the medium to the one-sided probe-type head system, the recording layer should be formed upon a several hundred nanometers thick soft magnetic (under)layer, which acts as a flux path from the main pole to the auxiliary pole or vice versa. In most cases, Permalloy is used for that purpose. It is noteworthy that even for the ring-type head system, a very thin, rather soft magnetic underlayer (several to several tens of nanometers) improves the recording characteristics substantially (Kitakami *et al.*, 1988; Iwasaki and Ouchi, 1989)

2. OPTICAL RECORDING*

2.1 Concept of Optical Recording

The concept of optical recording is not much different from that of magnetic recording. The only primary difference between the two is that a light beam is utilized to write and read a bit in the former, a magnetic flux in the latter. Therefore, the setup for optical recording devices is quite similar to that for magnetic ones. Optical recording devices basically consist of a medium upon which signals are written and an optical head that applies an optical beam to write and/or read the signals. One of the prominent characteristics of optical recording is its extremely high-density recording capability due to the fact that the light can be focused down to the limit of the wavelength, say less than 1 μm. Other features especially to be mentioned will be the facts that the spacing between the medium and the head can be set wide enough to keep them from contact with each other and that each bit is accessible by the light beam through the transparent substrate at the surface of which the beam diameter can be made as large as about 1 mm. This provides the system with high endurance against dust, enabling the media to be made detachable from the drive and portable, with a much higher recording density than floppy disks.

Figure 20 illustrates the optical recording system schematically. An optical disk is rotated at a high speed and a head is set under the disk and is driven by an actuator so that the light beam may access each track of the disk transversely. The head system is provided with both autofocusing and autotracking functions in order to take full advantage of the high-density recording capability.

2.2 Optical Recording Devices

Optical recording is classified into three categories: read only, write once, and rewritable (erasable), according to the mechanism for recording the information. Just like magnetic recording, the medium can assume any

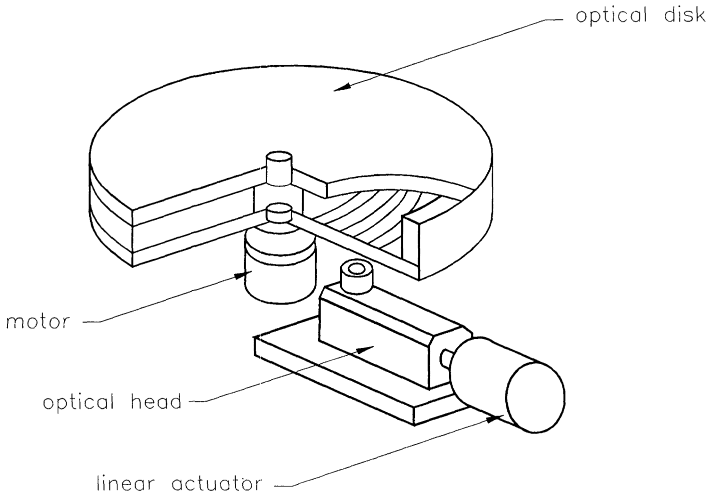

FIG. 20. Conceptual illustration of an optical disk system (after Terao *et al.*, 1990).

---

*A substantial part of this section is based on "Optical Recording" (in Japanese) (Terao *et al.*, 1990).

style—a disk, card, or tape, to meet various usages, disk being the most popular.

Among the read-only devices are video disk (VD), CD (compact disk), and its derivatives CD-ROM, CD-I (CD-interactive), CD-ROM.XA, and CD-V (CD-video). Video disks employ the FM method to record images. They are either double-sided or single-sided with diameters 200 mm (8 in.) or 300 mm (12 in.), with the playing time ranging from 20 to 120 min. The CD has a diameter of 120 mm and has been developed for recording music by the PCM (digital) method with playing time more than 60 min. An 80-mm CD has also been developed recently with the name "CD single." The CD is rotated at a constant angular velocity (CAV) whereas the VD has two types, CAV and CLV (constant linear velocity). The CD-ROM is a "read-only memory" of any data or software for digital information processors, having the same geometrical, mechanical, and optical specifications for the read procedures as those for the CD. The data structures are also made similar to each other as shown in Fig. 21. On CD-I and CD-ROM.XA are recorded both image and sound by means of a time-division multiplexing method, the latter being standardized for PC use. The CD-V has also both image and sound recordings but they are recorded separately from each other, image data on the outside and sound data inside. All the VD's and CD's have a single spiral track per surface.

Write-once [W-O or WORM (write once and read many)] and rewritable types are both intended to be used for digital information files of a huge capacity, having either a single spiral track or a large number of concentric circular tracks on each surface. Various sizes of W-O–type disks are now on the market such as 356 mm (14 in.), 300 mm (12 in.), 200 mm (8 in.), and 130 mm (5.25 in.), with various kinds of recording media. As for the rewritable ones, 130-mm (5.25-in.) and 85-mm (3.5-in.) magneto-optic

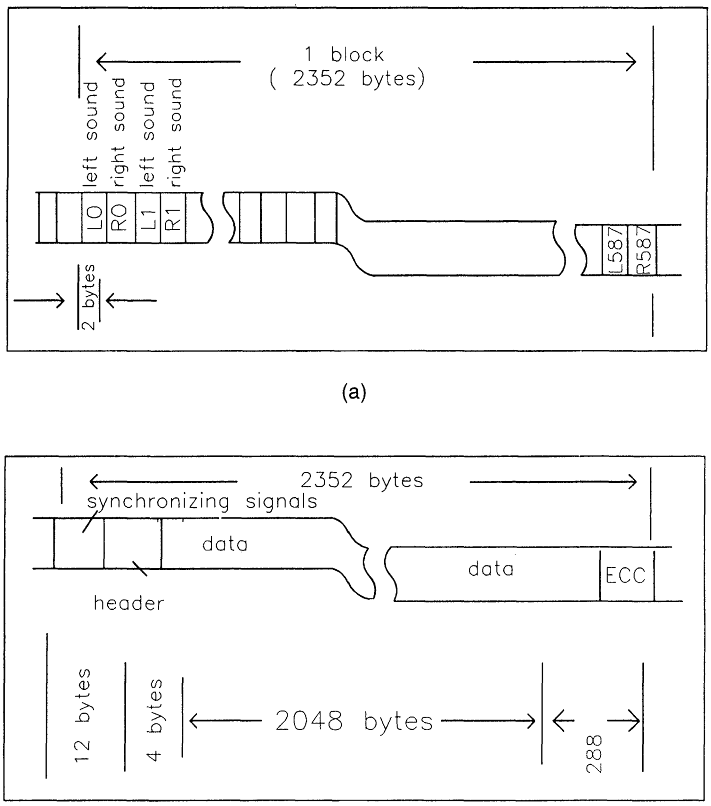

(a)

(b)

FIG. 21. Data recording schemes of (a) CD and (b) CD-ROM (after Otomo, 1988).

(MO) disks are now commonly used, although a 300-mm (12-in.) MO disk has also been developed for special applications such as video editors for broadcasting uses. The number of possible write repetitions exceeds 10^7. Phase-change–type optical disks are now also beginning to be put on the market, which will provide disks of lower cost with lower write repeatability. In Table 8 are listed the specifications and characteristics of typical optical disks and drives.

2.3 Optical Heads

In Fig. 22 is illustrated the basic system of an optical head. The beam from the laser is made parallel through the coupling lens and is shaped into a circular beam by the first prism. After passing through the polarized beam splitter, the linearly polarized light emitted from the laser is transformed into a circularly polarized beam through the $\frac{1}{4}\lambda$ (wavelength) plate, reflected by the mirror toward the medium, and focused upon the surface of the recording layer of the medium by the condenser lens mounted upon a fine positioner (not shown). The reflected beam from the medium comes back to the beam splitter after being changed into linearly polarized light by the $\frac{1}{4}\lambda$ plate and forwarded to the light detectors of the autofocusing and autotracking signals as well as the data signals. For MO recording, the $\frac{1}{4}\lambda$ plate is omitted because the linearly polarized light is utilized to read the signals.

As a laser, GaAlAs laser with 780- to 830-nm wavelength is now adopted, although, for a further increase of the recording density, light sources with shorter wavelengths are under development such as the InGaAlP laser (670 nm), the ZnSe laser (600 nm), and second-harmonic generators that generate the light of a wave length $\frac{1}{2}$ of that of the incident beam (when using a YAG laser, 532-nm light is generated).

It is well known that the diameter of the focal point (measured between points of $1/e^2$ times maximum intensity) is expressed by λ/NA and the focal depth by λ/NA^2, where λ is the wavelength of the light and NA is the numerical aperture of the condenser lens. In order to make the focal diameter small, it is necessary to employ a lens of a large NA, which makes the focal depth extremely small. Therefore, in the optical recording

Table 8. Optical disks.

	Video disk		CD	WORM disk			MO disk			MD
Disk size	200/300 mm	300 mm	120/80 mm	14 in.	12 in.	5.25 in.	12 in.	5.25 in.	3.5 in.	2.5 in.
Type of disk	Read-only	WORM	Read-only	Write-once read-many			Video	Rewritable	Rewritable	Rewritable
Capacity/side (MB)	Analog	Same	640/180	4500	1300/3500	300	11 500	340/1000	128/1230	140
Linear density (kbpi)	White peak: 9.3 MHz	Same	25		19.5/33.2	24	74.7	24/45.6	24.4/34.1	
Track density (ktpi)	16	16	16		16/17	16	21.2	16/19	16/18	16
Playing time										
CLV (min)	20/60	60	74/20	. . .	. . .	. . .	32 (both sides)	. . .	. . .	60/74
CAV (min)	–/30	30								
Rotation speed										
CLV (m/s)	10.0–11.3	Same	1.2–1.4							1.2–1.4
CAV (rpm)	1800	Same			600/1000	1800	1800	1800/3000	1800/3600	
Laser power (mW)	<1	11	<1		6	6	<10	5	5	5
Delivery year	1978	1990	1982	1991	1984/1990	1987		1987/1993	1990/1993	1993

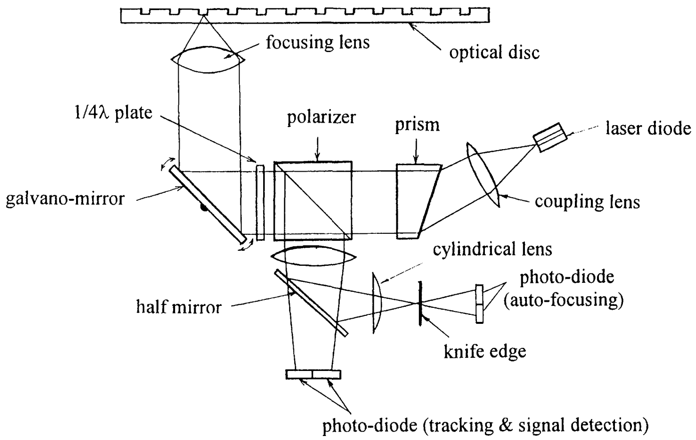

FIG. 22. Basic system of an optical head (after Terao *et al.*, 1990).

system, autofocusing is inevitable. Autofocusing is performed by detecting the change of the focus of the reflected beam condensed through a condenser system set in front of the light detectors. Several methods are proposed for it:

1. two detectors, set before and behind the nominal focal point;
2. a knife-edge shutter set at the nominal focal point and detection of the change of angular distribution of the light;
3. astigmatism method; and
4. astigmatism method with a knife edge.

In the astigmatism methods, a cylindrical lens is combined with a condenser lens in the condenser system to provide an elongated beam cross section in the straight and curved direction of the cylindrical lens right before and after the average focal point of the system, respectively. In method **3**, four detectors are set at the average focal point for the normal position, and by adding/subtracting the four outputs, the amount and the direction of the defocusing are sensed. In method **4**, a knife-edged shutter is set at the spot where the detectors are set in **3**, having its edge at 45° with the principal axis of the cylindrical lens, providing a semicircular beam cross section behind the shutter. The semicircle rotates clockwise or counterclockwise according to the defocusing direction. Therefore, by sensing the rotation direction, the current to be applied to the electromagnetic coil is determined to control the position of the condenser lens.

Autotracking methods are classified into two: continuous servoing [differential diffraction (push-pull) method and three-beam method] and sample servoing (SS). The differential diffraction method is employed when disks having guide groves are used. In this case, a two-division detector is set to receive equal zeroth- and first-order diffracted light when the beam is right on the track, and the off-tracking is sensed by the difference between the first-order diffracted beams. In most of the read-only disk drives, the three-beam method is adopted, in which the beam is split into three by a grating before entering the condenser lens and the light spot is divided into three, one main spot and two subspots aligned in a line tilted a little with respect to the data track direction. The off-track condition is sensed by the difference between the two signals detected through the two subbeams. In the sample servoing method, a number of marks are put periodically on each track (say 1000/track), each mark typically having a pair of dots aligned across the center of the track located a little before and behind in the track direction, and the data are recorded between the marks. Thus the off-track signals are obtained through the read beam.

2.4 Optical Recording Media

2.4.1 Read-Only Disks Read-only disks are produced by stamping. Figure 23 illustrates schematically the stamping process. The first three steps are those for preparing a stamper. First, an original disk in which the information is recorded as a series of

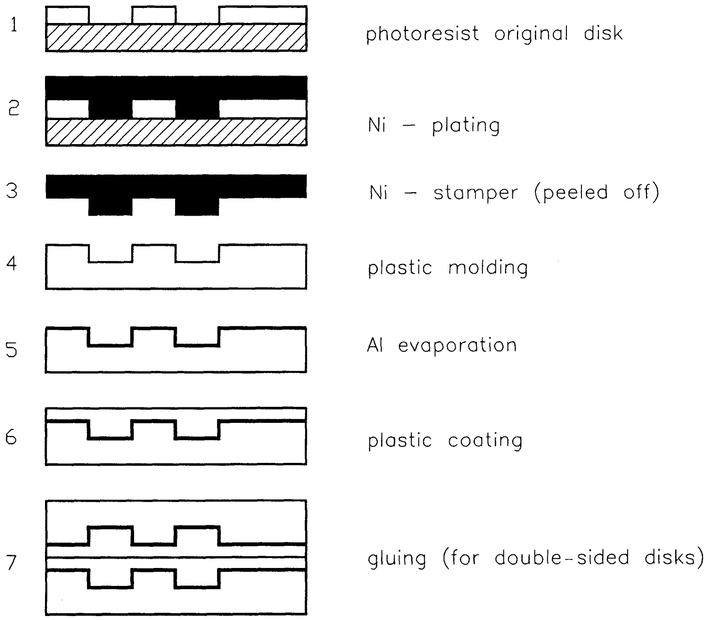
1
photoresist original disk
2
Ni – plating
3
Ni – stamper (peeled off)
4
plastic molding
5
Al evaporation
6
plastic coating
7
gluing (for double-sided disks)

FIG. 23. Production process of W-O–type optical disks (after Terao *et al.*, 1990).

pits is prepared by means of a laser-beam recording upon a photoresist-coated glass disk, then nickel is first sputtered (several micrometers) and then electrodeposited (about 0.2 mm thick), and finally the deposited nickel layer is peeled off making a mother stamper. Although not shown in the figure, actually, working stampers are made by repeating the Ni deposition and peeling similarly to the above with the aid of peeling agent, using the mother stamper. Then, plastic materials, such as polycarbonate (PC) and polymethyl methacrylate (PMMA), are compressed or injection-molded upon the stamper. After being peeled from the stamper, the mold is finished with aluminum coating for reflection and resin coating for protection (and adhered face to face when making a double-sided disk).

The signals are recorded as a series of pits having a width smaller than the read beam diameter ($\sim\lambda/\mathrm{NA}$) and a depth equal to $\lambda/4n$ (n: refractive index of the substrate). In the read process, the reflected light from a pit and its surrounding land cancel each other through interference, which gives a weaker return beam than from the land area, providing the pit signal.

2.4.2 Write-Once (W-O) Disks There are various kinds of W-O disks in terms of the mechanism of recording: deformation (hole, dip, or bubble formation), phase change, alloying, and color developing, with the deformation type, especially the hole-formation type, being the most common.

The mechanism of hole or dip formation by irradiation with a laser beam is explained as follows: When the disk is irradiated, the viscosity of the recording layer at the spot is decreased because of the elevation of the temperature, and a viscosity gradient appears with a minimum at the center, which in turn generates a gradient in the surface tension. The surface tension increases toward the perimeter, causing material to gather toward the perimeter, which makes a banklike bulge (or mound) leaving a dip inside as is schematically shown in Fig. 24; and when the viscosity becomes sufficiently low, a hole is formed at the center. There-

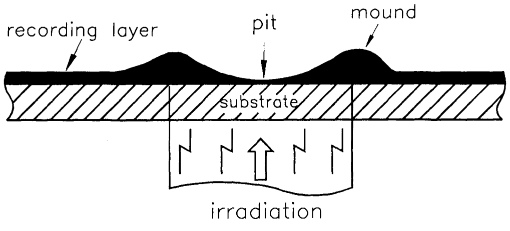

FIG. 24. Pit formation in W-O–type recording (after Oba *et al.*, 1985; Terao *et al.*, 1990).

fore, it is desirable that the materials for the recording layer possess a steep temperature dependence of the viscosity as well as a low melting point, high light-absorption coefficient, and low thermal conductivity in order to obtain high resolution and high light sensitivity. Of the many materials explored thus far, tellurium alloys such as (Pb–)Te–Se and Te–C are outstanding among inorganic materials, although there have been developed a lot of potential organic materials such as cyanine dyes, azulenium dyes, phthalocyanines, naphthoquinone dyes, etc., which have a high absorption coefficient for the light emitted from the laser diode. Generally, those dye polymers are rather apt to be deteriorated by the read beam or even by natural light, and their disk life cannot be expected to be as long as that of the inorganic ones. However, some of them show a read durability of more than 10^6 times and they are acceptable for low-cost disks.

There is another type of deformation by irradiation, in which the surface of the substrate itself is deformed through the heating up of the light-absorbing layer coated directly upon it. Au–Sn alloy on PC (polycarbonate) substrate is one of the most noticeable examples because it can provide a reflectivity contrast that is compatible with the CD's.

Te–O–Pd is known as an excellent material for phase-change–type media, in which fine Te particles of several nanometers are dispersed in the amorphous TeO_2 matrix. Without Pd, the heat/cool process under irradiation causes the recrystallization of Te in the form of long chains, which makes the spot shape irregular; Pd prevents this phenomenon. As alloying-type media, the $Sb_2Se_3/Bi_2Te_3/Sb_2Se_3$ triple-layered film coated with an Al reflection layer is well known. By irradiation, the Bi_2Te_3 layer absorbs the light and the temperature is elevated, which causes selective mutual diffusion of Bi and Se because of their strong affinity. A color-developing–type recording layer consists of a colorless dye layer and a developer layer mostly having a light-absorbing layer in between, an example being the combination of crystal violet lactone (dye), phenolphthalein (developer), and vanadyl phthalocyanine (absorber).

2.4.3 Rewritable (Erasable) Disks Although several methods have been investigated for rewritable optical recording, such as magneto-optic (MO), phase change, deformation, photochromic, liquid crystal, etc., most rewritable disks now on the market are MO disks, with some of the phase-transition types being also put into practical use.

2.4.3.1 MO (Magneto-optic) Disks. Figure 25 illustrates the principle of the write process. The recording layer is composed of a magneto-optic material of several tens of nanometers thickness, having a magnetic anisotropy with its easy axis perpendicular to the plane. Initially, the magnetization of the layer is saturated in one direction, say upward. When a spot is irradiated by a laser beam (<10 mW), the temperature of the spot is elevated and the coercive force decreases to almost zero near the Curie temperature. Thus the magnetization of the spot is reversed by the effect of the reverse field H_{srd} from the surroundings or, in most cases, by a field H_a (typically several hundred oersteds) applied through a coil set close to the condenser lens through which the beam is applied. After irradiation, the spot begins to cool accompanied by an increase in the coercive force, which sustains the small reverse domain (reversed magnetization region) even if the applied field is removed. Erasing is

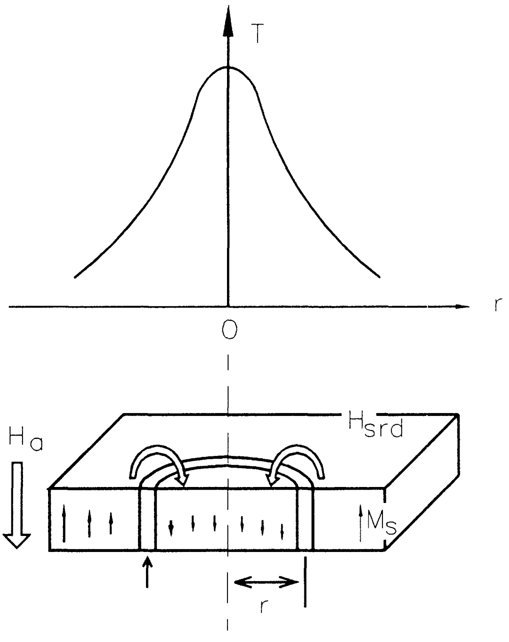

FIG. 25. Mechanism of magneto-optical (MO) recording (after Terao *et al.*, 1990).

performed before the rewrite process by applying an opposite field while irradiating. For reading, a magneto-optic effect called the polar Kerr effect is utilized. When linearly polarized light (2–3 mW) is reflected at the surface of a magnetic body, the polarization plane rotates clockwise or counterclockwise (a few tenths of a degree) according to the direction of the magnetization, which enables the sensing of the magnetization direction. Thus the magneto-optic materials for a recording layer should have

1. a high Kerr rotation angle,
2. appropriate Curie temperature (150 to 250 °C), and
3. low thermal conductivity and heat capacity, as well as sufficiently high perpendicular anisotropy $K_\perp$ to keep the magnetization perpendicular to the film plane, and high coercive force H_c, which sustains the written bit in the ambient conditions including reading.

More specific conditions that $K_\perp$ and H_c should satisfy are

$$K_\perp > 2\pi M_s^2, \tag{15}$$

$$H_c > (AK_\perp)^{1/2}/2\,M_s d, \tag{16}$$

where A is the exchange constant of the material and d is the diameter of the written spot. The readout signal-to-noise ratio is given by

$$S/N \sim R^{1/2}\theta_k \tag{17}$$

where R is the reflectivity and θ_k is the Kerr rotation angle. Therefore, the reflectivity and absorption should be appropriately compromised.

A lot of materials have been explored, and at present amorphous RE–TM (rare-earth–transition-metal) alloys are regarded as the most favorable (Imamura and Ota, 1980), although a lot of work has been devoted to manganese chalcogenides such as Mn(Cu)Bi, and to rare-earth iron garnets. This is because the RE–TM alloys have much flexibility in material designing, and their amorphousness is thought to be advantageous in making the written-spot perimeter smoother, leading to less noise than polycrystalline materials like MnBi and garnets.

Among RE–TM alloys, TbFeCo alloy is the most common, although GdTbFe with a greater Kerr effect and low-cost DyFeCo and TbDyFeCo are now also in practical use. In Fig. 26, curves of the saturation magnetization M_s and coercive force H_c vs temperature

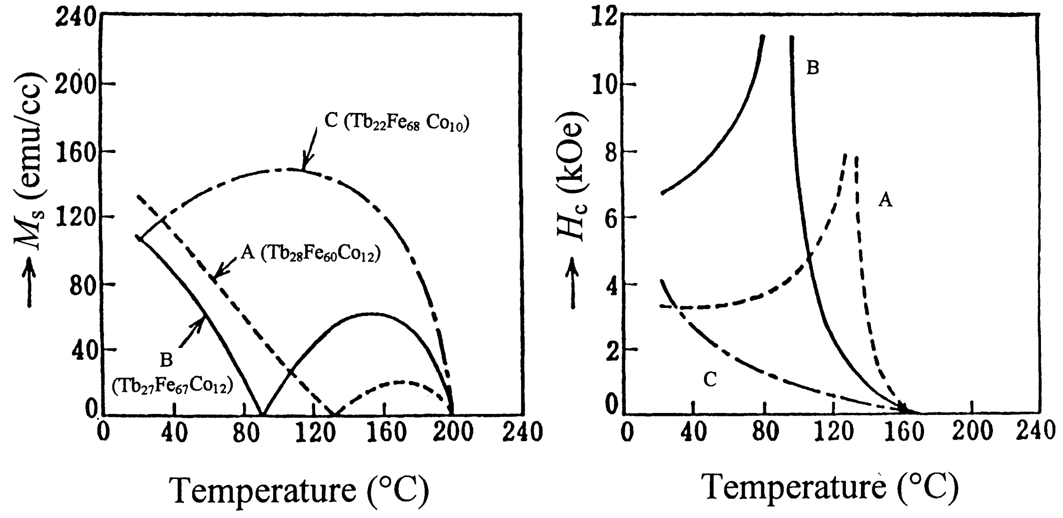

(a) (b)

FIG. 26. Temperature dependence of the saturation magnetization M_s and the coercive field H_c of typical TbFeCo magneto-optical (MO) recording films (after Takahashi *et al.*, 1988).

are shown for three typical kinds of TbFeCo alloys with different compositions. Figure 27 shows the written spots on each of the three kinds of disks for various applied fields during recording. It is seen that disk B is the most insensitive to variation of the field between 100 and 400 Oe, giving an almost constant size to the written spots. This behavior is explained by considering the balance of the forces acting on the domain wall during the write process: the forces through the applied field H_a, leakage flux of the surroundings H_{srd}, the surface tension and potential energy of the wall, and the coercive force taking into account the temperature distribution during recording:

$$H_c(r) = H_a + H_{\text{srd}} - \frac{\sigma_w}{2rM_s} - \frac{\partial M_s}{\partial r}(2M_s)^{-1}, \qquad (18)$$

where r is the radius of the written domain and σ_w is the wall energy.

For the next-generation MO disks increases in speed and recording density are now being explored. Higher speed will be attained by introducing overwritability, lighter heads, and multichannel heads. Typically, two schemes are under development to obtain overwritability: the field-modulation type and the beam-modulation type. In the former scheme, a small magnetic head is used for providing high-frequency modulation fields. To make the head small enough, the field necessary to write/erase is required to be weak. The coating of a very thin (several nanometers or less) ferromagnetic capping layer upon the recording layer is known to be extremely effective for field reduction (Ohnuki *et al.*, 1991). On the other hand, in the latter, exchange-coupled double-layered films consisting of a memory layer and an auxiliary layer or their modifications are being investigated (Saito *et al.*, 1987). Figure 28 illustrates how it works. One way to decrease the head weight to be moved is to divide the head into two parts, one consisting of the condenser-lens systems and the other the light-source and detector systems. Another is making it up with thin films. The former is almost at a practical use level, although the latter is still at a research stage. Four typical methods are being explored for increasing the recording density;

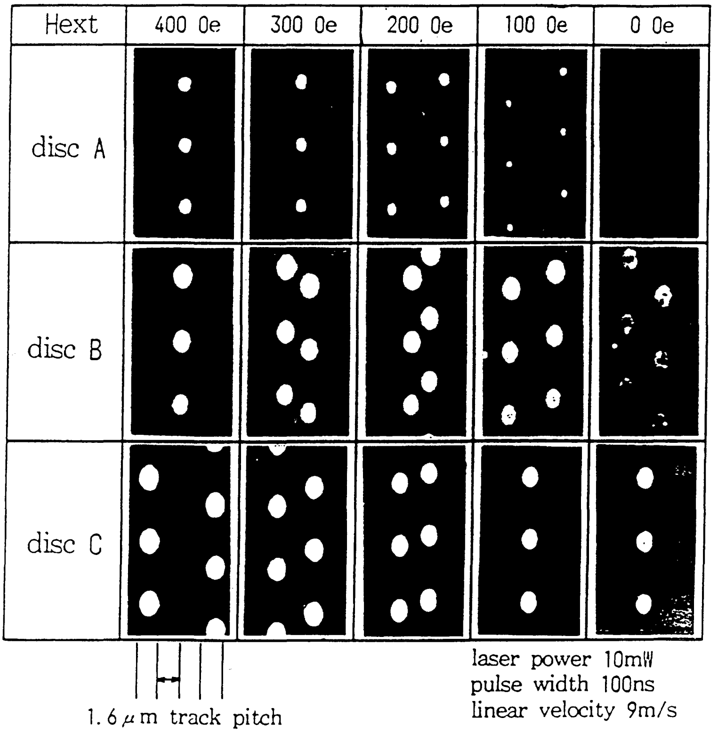

FIG. 27. Recorded patterns on the disks A, B, and C the magnetic properties of which are shown in the previous figure. The applied field was varied from 0 to 400 Oe, keeping the other conditions constant. (After Takahashi *et al.*, 1988).

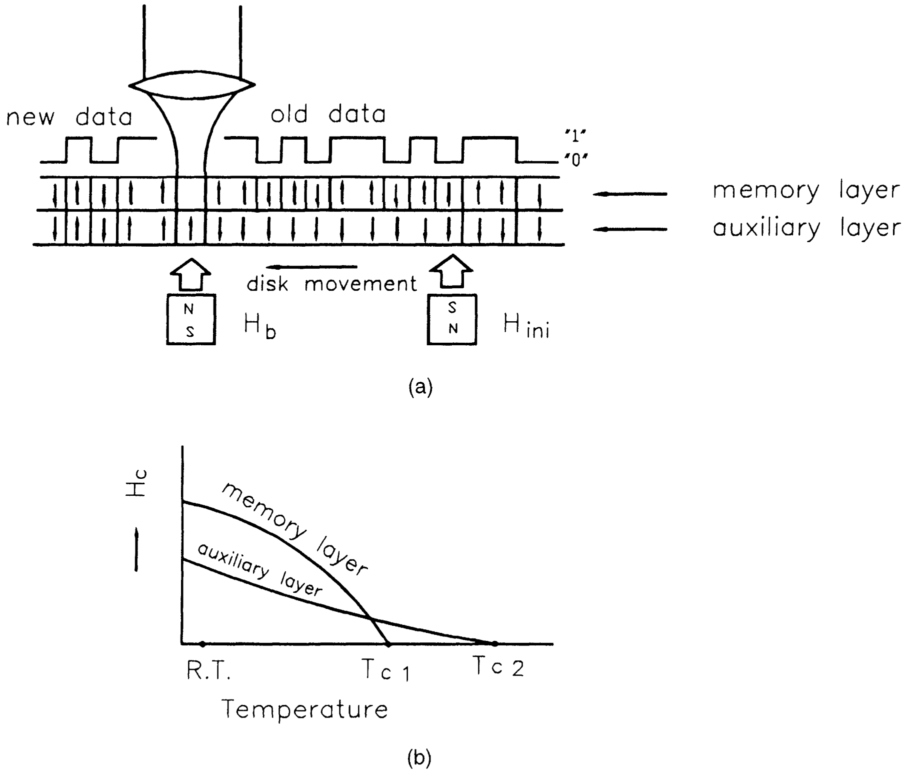

FIG. 28. Mechanism of overwritability in double-layered films for light (beam) modulation mode. The recording film consists of a memory layer and an auxiliary layer. The H_c vs temperature curve of each layer is shown in (b). Before overwriting, the auxiliary layer is initialized by a field H_{ini}. For overwriting, H_b is applied in the reverse direction. For the high-level beam ("1"), both magnetizations are aligned upward; for the low-level ("0"), the magnetization of the auxiliary layer does not change and the memory layer follows it by the exchange coupling. (After Saito *et al.*, 1987.)

1. reduction of the light wavelength λ,
2. mark-edge recording,
3. removal of track guard band, and
4. MSR (magnetically induced super resolution), described later (Ohta et al., 1991).

The reduction of λ basically depends on the development of short-wavelength laser diodes. However, materials for short wavelength are also being investigated, amorphous NdFeCo, garnets, and Co/Pt(Pd) multilayers (Zeper, *et al.*, 1989) being typical ones. Mark-edge recording, in which information is stored by the position of the edge of a recorded mark, can increase at least 1.5 times the conventional mark-position recording, and it is almost at hand. The three-beam method has been proposed for removing the guard band that enables cancellation of the crosstalk from the neighboring tracks. There are two typical types in the MSR method: front-aperture detection (FAD) and rear-aperture detection (RAD). The FAD method utilizes trilayered films exchange coupled when no beam is applied; the recording layer, a switching layer, and a readout layer (see Fig. 29). At room temperature, the recording layer has the highest coercivity and the coercivities of the readout layer and the switching layer are low enough for the magnetization of each layer to align in the same direction as that of the memory layer through exchange coupling. The Curie temperature of the switching layer is made lower than that of the other two layers. Thus, when a read beam is applied to a rotating disk while an erase field H_r is applied, the film is heated up to make a part of the

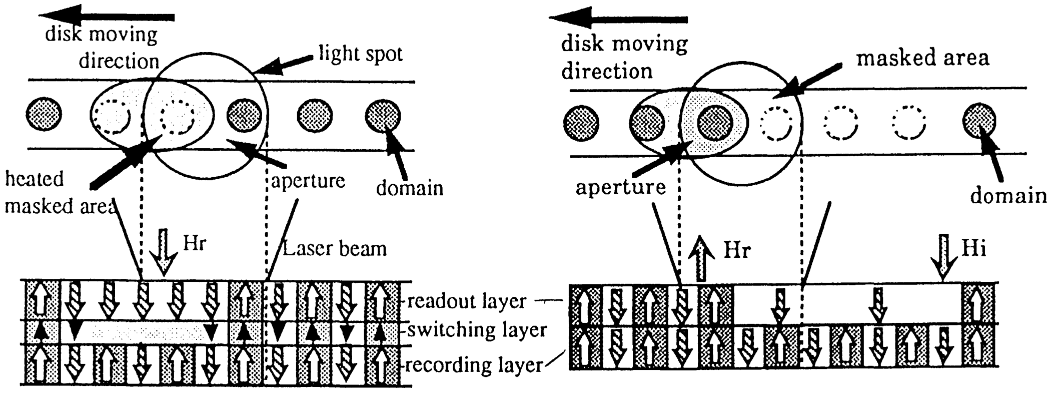
disk moving direction
light spot
heated masked area
aperture
domain
Hr
Laser beam
readout layer
switching layer
recording layer
disk moving direction
masked area
aperture
Hr
Hi
domain

(a) (b)

FIG. 29. Schematic illustration of the magnetically induced super-resolution (MSR) methods: (a) front-aperture detection (FAD) method, (b) rear-aperture detection (RAD) method (after Ohta *et al.*, 1991).

irradiated switching layer nonmagnetic, decoupling the readout layer from the recording layer and resulting in the erasure of the readout layer, which acts as a kind of mask in the spot providing an extremely high readout resolution. Regarding the mechanism of RAD, see Fig. 29(b).

Very recently, multilevel (four-level) recording has been achieved by using multilayered MO media, which will also be a candidate for a future high-density recording scheme (Shimazaki *et al.*, 1995).

2.4.3.2 Phase-Change–Type Disks. Amorphous materials have a glass point T_g above which atoms become rather easy to move around, and a melting point T_m above which they become liquids. If they are kept at a temperature between T_g and T_m for a sufficient time, they turn into crystallites, and if they are cooled rapidly from their liquid state, they become amorphous. The write/erase or overwrite in the phase-change–type disks utilizes this phenomenon. Figure 30 illustrates the recording process schematically. A series of light pulses corresponding to the signal to be written, having a sufficient height to melt the recording layer, are applied with a constant biasing power that elevates the temperature just above the crystallization temperature, thus leaving amor-

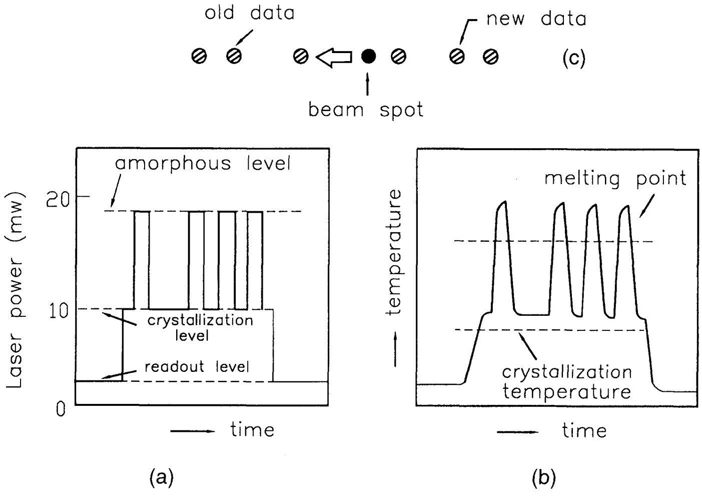
old data
new data
(c)
beam spot
Laser power (mw)
20
10
0
amorphous level
crystallization level
readout level
time
(a)
temperature
melting point
crystallization temperature
time
(b)

FIG. 30. Mechanism of phase-change–type recording (one-beam overwrite mode). (a) Form of three-level beam (amorphous, crystallization, and readout levels) applied. (b) Temperature profile generated by the irradiation. The spots where the temperature has exceeded the melting point become amorphous and the spots where the temperature has been raised a little higher than the crystallization temperature become crystallites. Thus the pattern as shown in (c) is recorded. (After Terao *et al.*, 1990.)

phous spots corresponding to the pulse peaks and crystalline spots corresponding to the valleys. Therefore, one of the crucial points to design the recording layer is to choose an appropriate rate of the phase change from amorphous to crystalline state in the range of the temperatures that the recording layers will experience.

The crystallization ratio x as a function of time t is generally given by the Avrami's formula,

$$x = 1 - \exp(-kt^n), \tag{19}$$

where n is an index called the Avrami index that takes on values from 1 to 4 according to the dimension of the crystal growth and whether the number of nuclei is kept constant or increases with time, and k is a function of the nucleation speed and the crystal-growth rate. The average crystallization rate $\langle \nu_0 \rangle$ from the beginning to the time when x becomes x_0 is given by

$$\langle \nu_0 \rangle = [-k/\ln(1 - x_0)]^{1/n}. \tag{20}$$

Reading the written bit is performed by detecting the change of the reflectivity as in the case of the write-once type.

The Ge–Sb(Co,Tl,Ag)–Te system is known as one of the most excellent material systems in terms of overwrite repeatability and erase characteristics (Akahira *et al.*, 1988; Ohta *et al.*, 1989). GeTe possesses a crystallization time as short as 30 ns and a high stability in the amorphous state because of its high crystallization temperature (Chen *et al.*, 1986). However, it is easily cracked during crystallization, as a substantial volume change takes place during the phase transition. On the other hand, Sb_2Te_3 has a long crystallization time of several microseconds but shows an excellent resistance to crack formation (Yagi *et al.*, 1987). The Ge–Sb–Te system is endowed with the compromised characteristics of those two. The addition of Co, Tl, and Ag controls the crystallization speed (Terao *et al.*, 1989). In–Se–Tl–Co, In–Sb–Te, and In–Sb have also been investigated extensively. However, none of them have surpassed the Ge–Sb–Te system in erasability and overwrite repeatability.

2.4.4 Multiplex Recording In order to obtain much higher recording density, multiplex recording methods have been investigated, such as multiwavelength recording and holographic recording, although a great deal of work is left to be done before they are put into practical use.

2.4.4.1 Hole Burning. Several media consisting of pluralities of dyes, such as multilayered LB (Langmuir–Blodgett) films of spiropirane compounds, have been shown to have the potential of being applied to multiwavelength recording. However the multiplexity number must be limited. "Hole burning," on the other hand, has demonstrated the possibility of attaining multiplexity of almost 1000 (Bjorklund *et al.*, 1988). The recording medium is composed of a host material with a guest material dispersed in it. The guest must be or must possess a light-absorption center, the spectrum of which is modulated by the inhomogeneity of the surroundings, resulting in a broad dispersion in the absorption spectrum of the medium. Recording is performed by applying a monochromatic light with a sharp spectrum making some of the absorption centers immune to the light of the same wavelength by exciting the electrons away. This generates a "hole" in the absorption spectrum of the irradiated spot, which enables reading out later. By changing the wavelength, a multiple of recordings can be done on the same spot.

A lot of materials have been investigated including both organic and inorganic materials, such as porphyrin, phthalocyanine, quinizarin, carbazole–boric acid, single-crystalline alkali halide with color centers or divalent ions, single-crystalline transition-metal oxide, glass doped with trivalent ions (Pr^{3+}, Nd^{3+}), etc. In terms of the multiplexity in recording, quinizarin/PHEMA(poly-2-hydroxyethylmethacrylate) and ABDAQ [4-amino-2,6-bis (4-butylphenoxy)-1,5-dihydroxy-anthraquinone]/PHEMA are known as outstanding. For a double-layered film of these two, 600 multiplex recording has been demonstrated successfully (Yoshimura *et al.*, 1989).

In order to obtain a reasonable signal-to-noise ratio and speed, however, the guest must simultaneously fulfill conditions with respect to both quantum efficiency and absorption cross section, which has not yet been attained. All the successful experiments have been performed at very low temperatures such as liquid He or N_2 temperatures. This is another drawback left to be overcome. At present, by reading, the record is

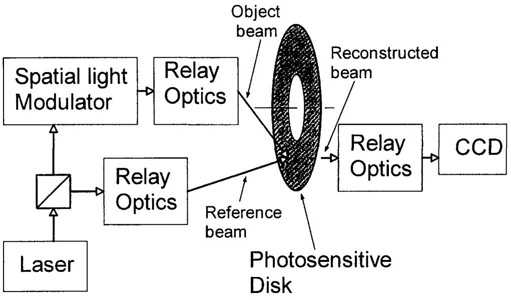

FIG. 31. Diagram of a holographic recording system (after Curtis, 1994).

substantially faded out, which is also one of the crucial issues to be improved.

2.4.4.2 Holographic Recording. Holography is a well-known image-recording technology in which both intensity and phase are recorded simultaneously in the form of interference patterns of the signal and reference beams. Figure 31 illustrates a possible system approach (Curtis, 1994). The spatial light modulator (SLM), which is either a liquid-crystal–based device or a deformable mirror, creates pages of data and multiple pages are recorded in the form of holograms (images of interference pattern) on the same area of the photosensitive recording medium by changing the wavelength or the incident angle of the reference beam. Reading is performed by applying the reference beam to generate a reconstructed beam, which is led to the light detector (CCD). Thus more than 10^5 parallel channel processing can be achieved. As recording materials, several photosensitive materials such as $LiNbO_3$, photorefractives, photopolymers, photorefractive polymers, and glasses are now under investigation. By using those materials, a recording density as high as 10^{10} pixels/cm^3 is expected to be available, 3×10^9 having been achieved experimentally by using $LiNbO_3$ (Heanue *et al.*, 1994). Although the necessary components now seem to be available, full evaluation of the total systems is still to be performed.

3. CONCLUSION

The history of recording technology has been and will continue to be, so to speak, a history of the increase in the recording density and rate. The basic trend in magnetic recording is the utilization of thin films both for heads and media. For the head, the development of high-permeability materials with a high saturation induction and the application of giant magnetoresistance will be major concerns, while for the media, thin-film media constituted of fine grains with high saturation moment and high anisotropy and the utilization of the inclined anisotropy will be. Perpendicular recording will also participate in ultrahigh-density recording in the long run. Although it was not mentioned above, it is worth pointing out that some recent studies imply that the switching speed of recording media will be a matter of serious concern when we go into the frequency range of a few hundred megahertz (He *et al.*, 1995). In optical recording, the development of shorter-wavelength laser diodes including SHG, and the media matching for them, and multilevel recording will be the main issues for increasing the recording density, although other multiplex recording technologies, such as holographic and hole-burning recording, give us an unlimited dream for the future.

GLOSSARY

Barkhausen Noise: In the magnetization process in which domain-wall motion takes place, the change in the magnetization becomes rather discontinuous because there exist a lot of wall pinning sites originating from various kinds of inhomogeneity such as holes, inclusions, crystalline boundaries, etc., generating noise. This effect was first observed by H. Barkhausen in 1919, who made an indirect verification of the presence of ferromagnetic domains in a magnetic body.

Demagnetization Field: A magnetic field that appears when a magnetic body is magnetized, due to the appearance of magnetic free poles mostly at the two ends of the body. As this field opposes the magnetization direction, resulting in some degree of demagnetization, it is called the demagnetization field. When a magnetic body is saturated in one direction with a saturation magnetization M_s, the demagnetization field H_d is expressed as $H_d = -NM_s$, where N is called a demagnetization factor.

GMR (Giant Magnetoresistance): A phenomenon found in magnetic multilayers each layer of which couples with the neighboring magnetic layers antiferromagnetically through a nonmagnetic layer in between, each layer having a thickness of a few nanometers (Baibich *et al.,* 1988). The resistance of the multilayers is changed drastically by applying a field that changes the angle of the magnetizations of the neighboring layers. This phenomenon comes from the spin-dependent resistance at the interface and in the bulk of the magnetic layers and is isotropic with respect to the applied current direction.

Inductive Head: A magnetic head consisting of a magnetic core and a coil wound around it. Because it operates through the phenomenon of inductance, it is called "inductive head." This term began to be used to distinguish the one from noninductive heads such as Hall head or MR (magnetoresistive) heads, which are used for reading only.

MR (Magnetoresistance): A phenomenon in which the resistance of a magnetic substance changes as a function of the magnetization direction with respect to the direction of the current applied. Since it is anisotropic in this sense in contrast to the giant magnetoresistive effect (GMR), it is called "anisotropic magnetoresistance."

Spin Valve: The metallic magnetic bilayer structure in which the magnetization of one layer is pinned and that of the other is free exhibiting the GMR effect. The magnetization in the free layer acts as if it were a valve for the electric current.

Works Cited

Akahira, N., Yamada, N., Kimura, K., Takao, M. (1988), in: D. B. Carlin, Y. Tsunoda, A. A. Jamberdino (Eds.), *Optical Storage Technology and Applications,* SPIE Proceedings No. 899, Bellingham, WA: SPIE.

Baibich, M. N., Broto, J. M., Fert, A., Nguyen Van Dau, F., Petroff, F. (1988), *Phys. Rev. Lett.* **61,** 2472–2475.

Bjorklund, G. C., Haarer, D., Hayes, J. M., Jankowiak, R., Lenth, W., MacFarlane, R. M., Sievers, A. J., Small, G. J. (1988), "Persistent Spectral Hole-Burning," in: W. E. Moerner (Ed.), *Science and Application,* Topics in Current Physics Series, New York: Springer-Verlag.

Chen, M., Rubin, K. A., Barton, R. W. (1986), *Appl. Phys. Lett.* **49,** 502–504.

Curtis, K., Psaltis, D. (1994), "Volume Holographic Data Storage," in *Proceedings of the ITA Sixth Annual Magnetic and Optical Media Seminar.*

Dieny, B., Speriosu, V. S., Parkin, S. S. P., Gurney, B. A., Wilhoit, R., Mauri, D. (1991), *Phys. Rev. B* **43,** 1297–1300.

Fujiwara, H., Kudo, M., Tamura, T., Sugishita, N., Shiroishi, Y., Kimura, T., Shinagawa, K., Kumasaka, N. (1982), U.S. Patent No. 4,316,228.

Goto, F., Osaka, T. (1993), "Plated Media for High Density Magnetic Recording," *Proc. Electrochem. Soc.* **93-20,** 410–422.

He, L., Doyle, W. D., Varga, L., Fujiwara, H., Flanders, P. J. (1995), presented at MRM '95 in Oxford (to be published).

Heanue, J. F., Bashaw, M. C., Hesselink, L. (1994), *Science* **265,** 749–752.

Hylton, T. L., Coffey, K. R., Parker, M. A., Howard, J. K. (1993), *Science* **261,** 1021–1024.

Imamura, N., Ota, C. (1980), *Jpn. J. Appl. Phys.* **19,** Suppl. **L,** 731.

Inaba, H., Saitoh, S., Kitahara, T., Kashiwagi, A. (1995), *IEICE Trans. Electron.* **E78-C,** 1536–1542.

Iwasaki, S., Nakamura, Y. (1977), *IEEE Trans. Magn.* **13,** 1272–1277.

Iwasaki, S., Ouchi, K. (1989), *J. Magn. Soc. Jpn.* **13,** Suppl. **S1,** 21–26.

Jeffers, F. (1986), *Proc. IEEE* **74,** 1540–1556.

Kitakami, O., Ogawa, Y., Fujiwara, H., Kugiya, F., Suzuki, M. (1988), *IEEE Trans. Magn.* **25,** 4177–4179.

Markham, D., Jeffers, F. (1990), "Magnetoresistive Head Technology," *Proc. Electrochem. Soc.* **90-8,** 185–204.

Mee, C. D., Daniel, E. D. (1989), *Magnetic Recording Handbook,* New York: McGraw-Hill.

Nakamura, Y. (1991), *J. Magn. Soc. Jpn.* **15,** S2, 497–506.

Nikles, D. E., Cain, J. L., Chacko, A. P., Webb, R. I. (1994), *IEEE Trans. Magn.* **30,** 4068–4070.

Oba, H., Abe, M., Umehara, M., Sato, T., Ueda, Y., Kunikane, M. (1985), "Organic Dye Materials for Optical Recording Medium," *Digest of Topical Meeting on Optical Data Storage,* WDD, Washington, DC: Optical Society of America, pp. 1–15.

Ohnuki, S., Shimazaki, K., Ohta, N., Fujiwara, H. (1991), *J. Magn. Soc. Jpn.* **15,** Suppl. **S1,** 399–402.

Ohta, M., Fukumoto, A., Aratani, K., Kaneko, M., Watanabe, K. (1991), *J. Magn. Soc. Jpn.* **15,** Suppl. **S1,** 319–322.

Ohta, T., Uchida, M., Yoshioka, K., Inoue, K., Akiyama, T., Furukawa, S., Kotera, K., Nakamura, S. (1989), in: G. R. Knight, C. N. Kurtz (Eds.), *Optical Data Storage Topical Meeting,* SPIE Proceedings No. 1078, Bellingham, WA: SPIE.

Otomo, Y. (1990), *Optical Disks* (in Japanese, Frontier Technology Series, No. 31), Tokyo: Maruzen.

Parkin, S. S. P., Li, Z. G., Smith, D. J. (1991), *Appl. Phys. Lett.* **58,** 2710–2712.

Potter, R. I. (1974), *IEEE Trans. Magn.* **13,** 502–508.

Saito, J., Sato, M., Matsumoto, H., Akasaka, H. (1987), *Jpn. J. Appl. Phys.* **26,** Suppl. **26-4**, 155–159.

Shimazaki, K., Yoshihiro, M., Ishizaki, O., Ohta, N. (1995), in: G. R. Knight *et al.* (Eds.), *Optical Data Storage '95,* SPIE Proceedings No. 2514, Bellingham, WA: SPIE.

Shinjo, T., Yamamoto, H. (1990), *J. Phys. Soc. Jpn.* **59,** 3061–3064.

Takagaki, T., Takaoka, H., Furusawa, K., Abe, K., Kamei, T., Sano, M. (1987), *IEEE Trans. Magn.* **23,** 3420–3422.

Takahashi, M., Niihara, T., Ohta, N. (1988), *J. Appl. Phys.* **64,** 262–269.

Terao, M., Miyauchi, Y., Ando, K., Yasuoka, H., Tamura, R. (1989), "Progress of Phase-Change Single Beam Overwrite Technology," in: G. R. Knight, C. N. Kurtz (Eds.), *Optical Data Storage Topical Meeting,* SPIE Proceedings No. 1078, Bellingham, WA: SPIE.

Terao, M., Ohta, N., Horigome, S., Ojima, M. (1990), *The Fundamentals of Optical Memories* (in Japanese), Tokyo: Corona.

Yagi, S., Fujimori, S., Yamazaki, H. (1987), *Jpn. J. Appl. Phys.* **26,** Suppl. **S26-4**, 51–54.

Yoshida, K., Okuwaki, T., Osakabe, N., Tanabe, H., Ouchi, Y., Matsuda, T., Shinagawa, K., Tonomura, A., Fujiwara, H. (1983), *IEEE Trans. Magn.* **19,** 1600–1604.

Yoshimura, M., Nishimura, T., Tsukada, N. (1989), "Ultrahigh Density Optical Memory by Photochemical Hole Burning (PHB) and Multilayered PHB System," in: G. R. Knight, C. N. Kurtz, (Eds.), *Optical Data Storage Topical Meeting,* SPIE Proceedings No. 1078, Bellingham, WA: SPIE.

Zeper, W. B., Greidanus, F., Carcia, P. F. (1989), *IEEE Trans. Magn.* **25,** 3764–3766.

Further Reading

Magnetic Recording

Bate, G. (1981), "Recent Developments in Magnetic Recording Materials," *J. Appl. Phys.* **52,** 2447–2452.

Bertram, H. N. (1994), *Theory of Magnetic Recording,* New York: Cambridge Univ. Press.

Fert, A., Bruno, P. (1994), "Interlayer Coupling and Magnetoresistance in Multilayers," in: B. Heinrich, J. A. C. Bland (Eds.), *Ultrathin Magnetic Structures II,* New York: Springer-Verlag.

Fujiwara, H. (1992), "Emerging Magnetic Technologies for Consumer Audio/Video," *J. Appl. Phys.* **73,** 5757–5762.

Guerst, J. A. (1963), "The Reciprocity Principle in the Theory of Magnetic Recording," *Proc. IEEE* **51,** 1573–1577.

Iwasaki, S. (1984), "Perpendicular Magnetic Recording—Evolution and Future," *IEEE Trans. Magn.* **20,** 657–668.

Jorgensen, F. (1980), *The Complete Handbook of Magnetic Recording,* Blue Ridge Summit, PA: TAB Books.

Karlqvist, O. (1954), "Calculation of the Magnetic Field in the Ferromagnetic Layer of a Magnetic Drum," *Trans. R. Inst. Tech.,* reprinted (1985), in: R. M. White (Ed.), *Introduction to Magnetic Recording,* New York: IEEE.

Levy, P. M. (1994), "Giant Magnetoresistance in Magnetic Layered and Granular Materials," *Solid State Phys.* **47,** 367–463.

Mallinson, J. C., Bertram, H. N. (1984), "A Theoretical and Experimental Comparison of the Longitudinal and Vertical Modes of Magnetic Recording," *IEEE Trans. Magn.* **20,** 416–467.

Mallinson, J. C. (1987), *The Foundations of Magnetic Recording,* San Diego: Academic.

Mee, C. D., Ciureanu, P., Daniel, E. D. (1989), *Magnetic Recording Handbook,* New York: McGraw-Hill.

Middelhoek, S. (1992), *Thin Film Resistive Sensors,* London: Institute of Physics Publishing.

Miura, Y. (1983), "Thin Film Head," in: Y. Sakurai (Ed.), *Recent Magnetics for Electronics,* Japan Annual Reviews in Electronics, Computers and Telecommunications Vol. 10, Tokyo: Ohmsha, Ltd., p. 77.

Potter, R. I. (1975), "Analytical Expression for the Fringe Field of Finite Pole-Tip Length Recording Head Fields," *IEEE Trans. Magn.* **11,** 80–81.

Sharrock, M. P. (1990), "Particulate Recording Media," *MRS Bull.* **15,** (3), 53–61.

White, R. M. (Ed.) (1985), *Introduction to Magnetic Recording,* New York: IEEE.

Williams, M. L., Comstock, R. L. (1971), "An Analytical Model of the Write Process in Digital Magnetic Recording," in: *Seventeenth Annual AIP Conference Proceedings,* New York: AIP, pp. 738–742.

Optical Recording

Bell, A. E. (1987), "Materials for High Density Optical Data Storage," in: M. J. Weber (Ed.), *Handbook of Laser Science and Technology,* Vol. 5, New York: CRC, Part 3.

Bouwhuis, G., Braat, J., Huijser, A., Pasman, J., van Rosmalen, G., Schouhaumer-Immink, K. (1985), *Principles of Optical Disc Systems,* Bristol, U.K.: Adam Hilger.

Brady, D., Psaltis, D. (1992), "Control of Volume Holograms," *J. Opt. Soc. Am. A* **9,** 1167–1182.

Curtis, K., Pu, A., Psaltis, D. (1994), "Method for Holographic Storage Using Peristrophic Multiplexing," *Opt. Lett.* **19,** 993–994.

Curtis, K., Psaltis, D. (1994), "Characterization of the DuPont Photopolymer for Three-dimensional Holographic Storage," *Appl. Opt.* **33,** 5396–5399.

Emmelius, M., Pawlowski, G., Vollmann, H. W. (1989), "Materials for Optical Data Storage," *Angew. Chem., Int. Ed. Engl.,* **28,** 1445–1600.

Heanue, J. F., Bashaw, M. C., Hesselink, L. (1994), "Volume Holographic Storage and Retrieval of Digital Data," *Science* **265,** 749–752.

Isailovic, J. (1985), *Videodisc and Optical Memory Systems,* Englewood Cliffs, NJ: Prentice-Hall.

Ito, F., Kitayama, K., Oguri, H. (1992), "Holographic Image Storage in $LiNbO_3$ Fibers with Compensation for Intrasignal Photorefractive Coupling," *J. Opt. Soc. Am. B* **9,** 1432–1439.

Johnson, G. H. (1989), "Information Storage Materials" in: G. E. Thomas (Ed.), *Ullman's Encyclopedia of Industrial Chemistry,* Vol. A 14, New York: VCH.

Mok, F. H., Tackitt, M. C., Stoll, H. M. (1991), "Storage of 500 High-Resolution Holograms in a $LiNbO_3$ Crystal," *Opt. Lett.* **16,** 605–607.

Nagao, Y., Sakata, H., Mimura, Y. (1993), "Improvements in Bi12SiO20 Thin Spatial Light Modulators," *Appl. Opt.* **32,** 5036–5042.

Niihara, T., Takahashi, M., Miyamoto, H., Kirino, F., Ogihara, N., Ohta, N. (1990), "Thermomagnetic Recording Mechanism on TbFeCo Disks," *J. Magn. Magn. Mater.* **88,** 177–182.

Qiao, Y., Psaltis, D., Gu, C., Hong, J., Yeh, P., Neurgaonkar, R. R. (1991), "Phase-Locked Sustainment of Photorefractive Holograms Using Phase Conjugation," *J. Appl. Phys.* **70,** 4646–4648.

Rakuljic, G. A., Leyva, V., Yariv, A. (1992), "Optical Data Storage by Using Orthogonal Wavelength-Multiplexed Volume Holograms," *Opt. Lett.* **17,** 1471–1473.

Van Heerden, P. J. (1963), "Theory of Optical Information Storage in Solids," *Appl. Opt.* **2,** 393–400.

Yu, F. T. S., Wu, S., Mayers, A. W., Rajan, S. (1991), "Wavelength Multiplexed Reflection Matched Spatial Filters Using $LiNbO_3$," *Opt. Commun.* **81,** 343–347.

RECRYSTALLIZATION, DYNAMIC

F. MONTHEILLET* AND J. J. JONAS, *Department of Metallurgical Engineering, McGill University, Montréal, Québec, Canada*

Introduction 205
1. **Discontinuous Dynamic Recrystallization** 206
1.1 Flow Curves Associated with Discontinuous DRX and the Critical-Strain Model 206
1.2 Grain-Size Evolution and the Relative–Grain-Size Model 209
1.3 Nucleus-Density Condition 211
1.4 Effect of Alloying Elements on DRX 213
2. **Dynamic Recovery and Continuous Dynamic Recrystallization** 213
2.1 Flow Curves Associated with Continuous DRX 213
2.2 Microstructural Changes Associated with Continuous DRX 215
3. **Texture Changes Associated with DRX** 218
4. **Postdynamic Softening** 221
5. **Implications for Industrial Processing** 223
Glossary 223
Works Cited 224
Further Reading 225

INTRODUCTION

The recrystallization that takes place *during* the high-temperature deformation of metals is known as dynamic recrystallization (DRX) and was initially a subject of considerable controversy. At the Sheffield conference on Deformation under Hot Working Conditions, for instance, it was still being claimed by Stüwe (1968) that the nucleation and growth of new grains (which involve grain-boundary migration) can only take place *after* straining, while the microstructural changes occurring *during* deformation are restricted to the rearrangement and removal of dislocations (dynamic recovery). It was, indeed, rather difficult to decide whether the new grains present in deformed and cooled samples were generated dynamically (i.e., during deformation) or statically, during the interval between hot working and quenching. As will be seen below, *postdynamic* restoration processes generally take place very rapidly, which makes it difficult to interpret metallographic data without ambiguity.

Because of these difficulties, the early references to DRX were based on the special shapes of the stress–strain curves sometimes observed under metal-working conditions. (Similar remarks apply to the unusual strain–time curves occasionally determined under creep conditions.) These were characterized by considerable amounts of strain softening, and more particularly by oscillations indicating that successive cycles of work hardening and flow softening were taking place. An early example of such wavy flow behavior is presented in Fig. 1, obtained from compression tests carried out on lead specimens at room temperature (Thomsen *et al.,* 1954). Since then, similar observations have been made on a large number of metals and alloys (copper, carbon steels in the austenite range, austenitic stainless steels, nickel, etc.) that have been subjected to hot-compression, tension, or torsion tests. DRX has also been reported to occur in nonmetallic materials, such as ice (Steinemann, 1958), and some minerals (Poirier, 1972).

Although the occurrence of DRX under laboratory testing conditions has now been amply assessed and analyzed, some doubt still remains about its importance during in-

---

*Permanent address: Ecole des Mines de Saint-Etienne, Centre Science des Matériaux et des Structures, URA CNRS 1884, Laboratoire Microstructures et Mise en Forme, Saint-Etienne, France.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

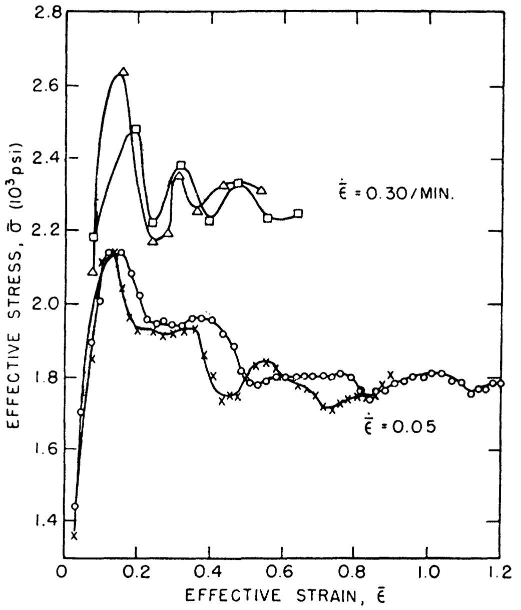

FIG. 1. Room-temperature stress–strain curves determined on lead specimens; the wavy flow behavior is an early example of the periodic softening produced by dynamic recrystallization (after Thomsen *et al.*, 1954).

dustrial hot-forming operations, as recently discussed by Jonas (1994). In processes where large strains are applied in a single operation (as in seamless tube piercing, extrusion, and planetary hot rolling), the material grain size during processing is clearly controlled by DRX. This also appears to be the case for multiple-pass processes such as hot rolling when the interpass times are short ($\leqslant 1$ s) so that interpass-restoration effects are negligible or absent, and the strain can accumulate. It will be seen later that such industrial forming conditions, which usually involve high strain rates (10 to 1000 s^{-1}), generally lead to *grain refinement* as a result of DRX and, therefore, to increases in yield stress and hardness, as well as in the resistance to cracking and fracture. No longer a "scientific curiosity," dynamic recrystallization is now becoming an industrial tool (Jonas, 1994).

Careful analysis of the stress–strain curves obtained over various ranges of strain rate and temperature, together with that of the associated microstructures, has shown that two broad classes of DRX can be distinguished. The first, usually referred to simply as "dynamic recrystallization" in the literature, involves the nucleation and growth of new grains within a strained matrix; it will be categorized here as *discontinuous dynamic recrystallization.* The second type results solely from the operation of dynamic recovery and grain-boundary migration; i.e., it does *not* involve nucleation. This mechanism can also produce "new" grains at large strains. Because of the absence of nucleation, it will be referred to here as *continuous dynamic recrystallization.*

1. DISCONTINUOUS DYNAMIC RECRYSTALLIZATION

The most striking microstructural feature of discontinuous DRX is the formation of dislocation-free nuclei, which grow during further straining. This leads to the appearance of equiaxed grains that, unlike those produced by *static* recrystallization, contain dislocations as a result of the continued straining.

1.1 Flow Curves Associated with Discontinuous DRX and the Critical-Strain Model

Some stress–strain curves that are typical of DRX are illustrated in Fig. 2(a); these were obtained from hot-compression tests of OFHC (oxygen-free high-conductivity) copper specimens at various temperatures (Blaz *et al.*, 1983). Here, the shape of the curve changes from the *multiple-peak* variety to the *single-peak* type when the temperature is decreased. A similar transition is observed when the strain rate is increased, as shown in Fig. 2(b) for the case of a 0.25% C steel (Rossard and Blain, 1959). Such equivalence between the effects of decreasing temperature and increasing strain rate is associated with the occurrence of thermally activated processes and has long been recognized in the study of hot working. It arises because an increase in strain rate decreases the time available for thermal activation (and therefore the probability of successful activation), while a decrease in temperature has an equivalent effect on successful activation by decreasing the amplitude of the thermal fluctuations. The equivalence permits use of the

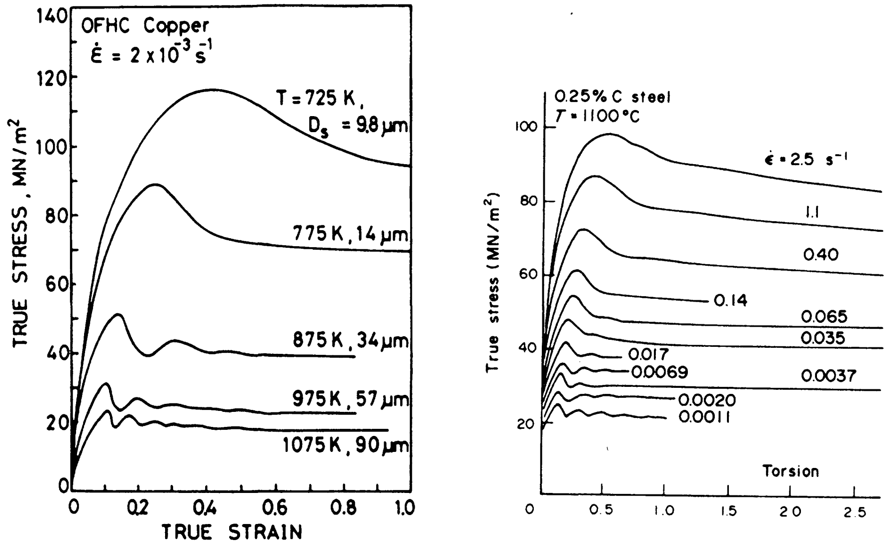

(a) (b)

FIG. 2. Stress–strain curves typical of discontinuous dynamic recrystallization obtained on (a) OFHC copper at various temperatures (after Blaz *et al.*, 1983) and (b) 0.25% C steel at various strain rates (after Rossard and Blain, 1959).

temperature-corrected strain rate, or Zener–Hollomon parameter (Zener and Hollomon, 1944):

$$Z = \dot{\epsilon}\exp(Q/RT), \qquad (1)$$

where $\dot{\epsilon}$ is the strain rate, T the absolute temperature, R the gas constant, and Q the apparent activation energy of deformation. Thus when Z is less than a critical value Z_c, multiple peaks are present, whereas when $Z > Z_c$, single-peak flow occurs.

The positive-slope portions of the oscillations (when $Z < Z_c$) can be intuitively understood as resulting from strain hardening, whereas the negative-slope regions can be associated with softening by recrystallization. Such softening can only be reflected in the flow curve when recrystallization is "synchronized" throughout the sample. Conversely, in the single-peak type of curve ($Z > Z_c$), work hardening and softening by recrystallization take place concurrently in different parts of the specimen, so that recrystallization is now "asynchronized." Starting from this very simple idea, a first model of discontinuous DRX was proposed by Luton and Sellars (1969). They introduced two characteristic strains associated with the flow curves (Fig. 3): ϵ_p, the strain to the first stress peak,* and ϵ_x, the recrystallization

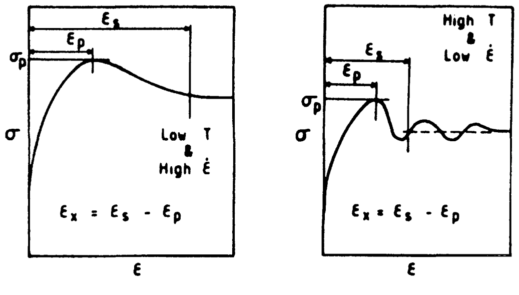
ε_s
ε_p
σ_p
σ
Low T & High $\dot{\varepsilon}$
$\varepsilon_x = \varepsilon_s - \varepsilon_p$
ε
High T & Low $\dot{\varepsilon}$
ε_s
ε_p
σ_p
σ
$\varepsilon_x = \varepsilon_s - \varepsilon_p$
ε

FIG. 3. Schematic definitions of the peak and recrystallization strains, ϵ_p and ϵ_x, respectively, in the single- (left) and multiple-peak (right) cases. ϵ_s is either the minimum strain required to attain steady-state flow or the strain at which the mean value of the flow stress is attained during the second cycle of work hardening. (After Luton and Sellars, 1969.)

---

*A more physical quantity, viz, the critical strain ϵ_c, has also been used. It refers to the strain required to initiate DRX, which is less than ϵ_p [$\epsilon_c \approx (0.6 \text{ to } 0.8)\epsilon_p$]. As ϵ_c cannot be readily determined from stress–strain curves, the "nucleation" strain is frequently specified in terms of ϵ_p rather than ϵ_c.

strain. The latter was defined as $\epsilon_s - \epsilon_p$, where ϵ_s is the minimum strain required to attain steady-state flow in single-peak curves, and the strain at which the mean value of the flow stress is attained during the second cycle of work hardening in multiple-peak curves. On the basis of their results for the torsion testing of nickel, the authors showed that when $\epsilon_p > \epsilon_x$ (low Z), DRX is synchronized, whereas when $\epsilon_p < \epsilon_x$ (large Z), asynchronized DRX takes place. By means of computer simulations, they were able to reproduce the two alternative shapes of stress–strain curves. This *critical-strain model* is consistent with the data obtained from numerous *torsion*-test investigations.

It has been shown more recently, however (Sakui *et al.*, 1977; Weiss *et al.*, 1984), that the critical-strain comparison does not apply to *tension* or *compression* experiments, at least in unmodified form. This is illustrated in Fig. 4, which shows the peak stress dependences of ϵ_p and ϵ_x for three plain carbon steels deformed in torsion, tension, and compression (Sakai and Jonas, 1984). In the case of *torsion* [Fig. 4(a)], the intersection of the two curves identifies the peak stress that corresponds to the critical value Z_c of the Zener–Hollomon parameter introduced above. The latter defines the transition from multiple-peak to single-peak flow. However, unlike the behavior displayed by the torsion results, the ϵ_p and ϵ_x curves do *not* intersect in the case of the tension [Fig. 4(b)] or compression [Fig. 4(c)] experiments, although the flow-curve–shape transition is still observed within the σ_p range investigated.

Such an unexpected disagreement between the results of torsion tests on the one hand, and those obtained from tension or compression tests on the other, has been analyzed in detail by Sakai and Jonas (1984). Their main conclusions can be summarized as follows.

1. The determination of material stress–strain curves (σ-ϵ) from torque (T) versus twist (N) data determined on solid-bar torsion specimens is subject to error, because of the radial strain and strain-rate gradients that are present. Although the problem was solved by Fields and Backofen (1957) for a wide class of materials that undergo work hardening and display rate sensitivity, their method involves calculation of the partial derivatives of T with respect to the number of twists, N, and the rate of twisting, $\dot{N}$. This requires the processing of large amounts of experimental data, and is rarely carried out. The algebraic shortcuts taken instead by most experimenters lead to the overestimation of ϵ_p.
2. Because necking develops fairly rapidly in tension after the stress peak has been attained, ϵ_x values determined by this

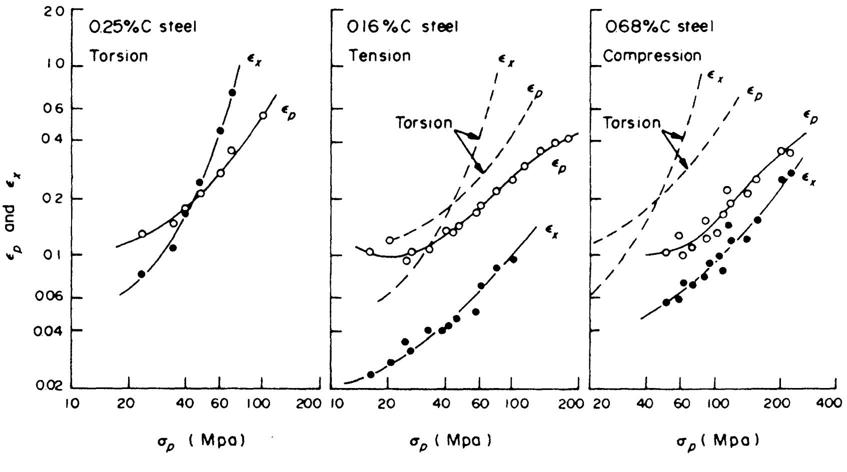

(a) (b) (c)

FIG. 4. Peak stress dependences of ϵ_p and ϵ_x for three plain carbon steels deformed in (a) torsion, (b) tension, and (c) compression (after Sakai and Jonas, 1984).

means will generally be lower than those measured using other test methods [cf. Fig. 4(b) with Figs. 4(a) and 4(c)].

3. There is also an effect of *strain path* on the rate of work hardening and therefore on the strain required to attain ϵ_p. This is because the flow stress of a material can be written in the form

$$\sigma(\epsilon) = \bar{M}(\epsilon)\bar{\tau}_c(\epsilon), \qquad (2)$$

where $\bar{M}$, the average Taylor factor, is associated with the current crystallographic texture and $\bar{\tau}_c$ is the mean value of the critical resolved shear stress, a quantity that is closely linked to the current dislocation configuration. The strain dependences of these two factors are distinctly different for specimens deformed along different strain paths (Jonas *et al.*, 1982). In particular, the flow stress of samples strained at room temperature is generally lower in torsion than in tension or compression. Such strain-path effects also contribute to the different ϵ_p and ϵ_x dependences displayed in Figs. 4(a)–4(c).

The errors and discrepancies involved in determining ϵ_p and ϵ_x from mechanical test data thus led Sakai and Jonas (1984) to conclude that the critical condition $\epsilon_p = \epsilon_x$ associated with the transition from single- to multiple-peak flow only applies, strictly speaking, to solid-bar torsion experiments and then only in an approximate and perhaps fortuitous way.

1.2 Grain-Size Evolution and the Relative-Grain-Size Model

The critical-strain approach was principally concerned with the mechanical or macroscopic aspects of DRX. Little attention was paid to the characteristics of the actual microstructural processes taking place within the material. More recently, these have been studied in some detail, principally by Sakai and co-workers (Sakui *et al.*, 1977; Sakai and Jonas, 1984). To avoid the occurrence of postdynamic restoration, they used special tensile machines in which the specimen could be quenched at a cooling rate of about 2000 K s^{-1} at the end of the deformation. By this means, these authors were able to investigate the evolution of the average grain size during DRX. In particular, they compared the dynamic "equilibrium" grain size D_s present in the steady state (at large strains) with the initial grain size D_0.

The results obtained on a 0.16% C steel tested at 940 °C and a range of strain rates are presented in Fig. 5. The strains at which stress peaks were observed are identified as P_1, P_2, etc. It can be seen that, at low strain rates, multiple-peak behavior is displayed and either grain coarsening ($\dot{\epsilon} = 2.6 \times 10^{-4}$ and $2.0 \times 10^{-3}\ s^{-1}$) or slight refinement ($\dot{\epsilon} = 2.0 \times 10^{-2}\ s^{-1}$) takes place. Conversely, at high strain rates ($\dot{\epsilon} = 1.48 \times 10^{-1}$ and 18 s^{-1}), single-peak behavior is associated with a large decrease in grain size. The critical D_0/D_s ratio above which single-peak behavior occurs can be estimated to be about 2. This can be shown more quantitatively as follows.

It is well known that the equilibrium grain size D_s is a power-law function of the steady-state flow stress σ_s for a number of metals and alloys (Derby, 1991):

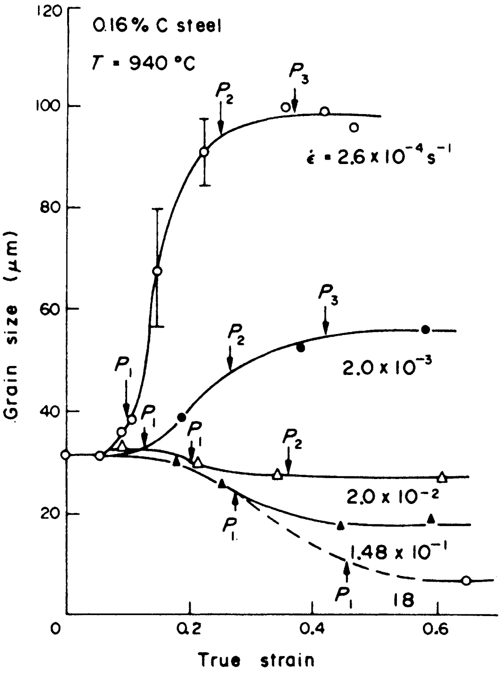

FIG. 5. Strain dependence of the average grain size during the tensile deformation of a 0.16% C steel at 940 °C. Whether grain coarsening or grain refinement takes place depends on the imposed strain rate. The strains at which stress peaks were observed are identified as P_1, P_2, etc. (After Sakui *et al.*, 1977.)

$$\sigma_s = k_1/D_s^{\alpha}. \tag{3}$$

Here, $\alpha \approx 0.66$–0.75 is an empirical constant and k_1 is a constant. Since σ_s is itself related to the Zener–Hollomon parameter Z by a power-law-relationship (provided that the strain rate and stress are not too high), viz.,

$$\sigma_s = k_2 Z^m, \tag{4}$$

Z can be expressed in the form

$$Z = k/D_s^{\alpha/m} \tag{5}$$

In the above equations, k_2 and k are further constants, and $m \approx 0.1$ to 0.2, so that $\alpha/m \approx 4$ to 8. Equation (5) essentially states that D_s depends solely on the temperature-corrected strain rate Z. This can be interpreted as signifying that the growth of new grains stops when the driving force for grain-boundary migration (the dislocation-density difference) decreases to zero, or drops below a critical level. The loss of driving force is caused by strain hardening, the rate of which must be balanced against the velocity of grain-boundary migration. It is the latter quantity that is directly related to the parameters of straining, via Z.

The relationship described by Eq. (5) is illustrated in the semilogarithmic plot of Fig. 6 for a 0.16% C steel. Here it can be seen that the dependence of the critical value of the Zener–Hollomon parameter Z_c on the initial grain size D_0 is *nearly parallel* to the Z-D_s dependence. Furthermore, the above authors showed that the Z_c vs $2D_s$ relation approximately coincides with the one pertaining to Z_c vs D_0 (Fig. 6). This means that $D_0 = 2D_s$ corresponds to $Z = Z_c$ and is thus associated with the transition from multiple- to single-peak behavior.

These experimental observations led to the *relative-grain-size model* described in Fig. 7. In this double-logarithmic plot, the Z_c-D_0 and Z-D_s relations are now represented by two parallel straight lines of slope $-\alpha/m$. The first divides the diagram into two parts, associated with single-peak (large initial grain size, high Z value) and multiple-peak (small initial grain size, low Z value) behavior, respectively. The Z-D_s line in turn distinguishes between the grain-coarsening and grain-refinement domains. From the diagram, the outcomes of two different types of tests can be predicted. "Vertical" experiments are carried out with a fixed initial grain size D_0 but at various values of Z, as in Figs. 2(a) and 2(b). According to Fig. 7, given an initial grain size D_{01}, stress oscillations will occur when a combination of relatively low testing strain rate and high tem-

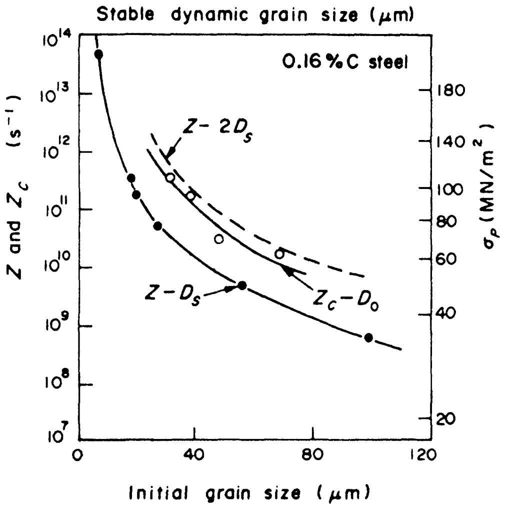

FIG. 6. Relationship between the steady-state grain size D_s and Zener–Hollomon parameter Z. The initial grain-size dependence of the critical value Z_c associated with the transition from single- to multiple-peak flow (and vice versa) is shown to be *parallel* to the latter curve. (After Sakui *et al.*, 1977.)

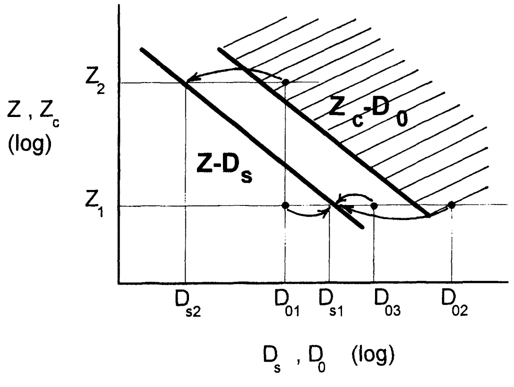

FIG. 7. Double-logarithmic plot illustrating the relative–grain-size model (see text). The Z_c-D_0 and Z-D_s relations are represented here by two parallel straight lines. The cross-hatched area corresponds to the temperature-corrected strain rate and initial grain-size conditions associated with single-peak dynamic recrystallization.

perature, e.g., $Z_1 < Z_{c1}$, is used. Each cycle of recrystallization then produces *grain coarsening* until the stable grain size D_{s1} pertaining to Z_1 is attained. Conversely, if a combination of a relatively high testing strain rate and low temperature, e.g., $Z_2 > Z_{c1}$, is used, a single peak is observed and *grain refinement* occurs until the stable grain size D_{s2} is reached.

The "horizontal" tests are carried out, in turn, at a fixed value of Z (generally with both the strain rate *and* temperature held constant) but at various values of D_0, as illustrated in Fig. 8. Thus for $Z = Z_1$, when the initial grain size D_{01} is finer than D_{s1}, multiple peaks and grain coarsening are called for, persisting until D_{s1} is attained (Fig. 7). Alternatively, when the initial grain size D_{02} or D_{03} is larger than D_{s1}, the occurrence of grain refinement is predicted; this can be associated with either single- (for $D_{02} > 2D_{s1}$) or multiple-peak (for $D_{s1} < D_{03} < 2D_{s1}$) flow. The grain size decreases in this case until it attains D_{s1} and a steady state of flow is achieved.

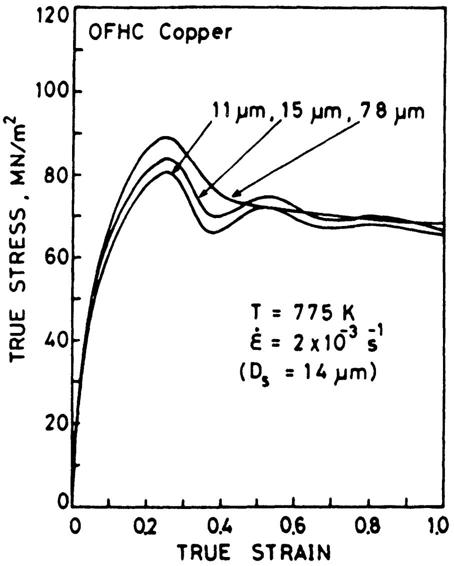

FIG. 8. Hot-compression stress–strain curves for OFHC copper specimens with various initial grain sizes, illustrating the transition from multiple- to single-peak dynamic recrystallization (after Blaz *et al.*, 1983).

1.3 Nucleus-Density Condition

A physical basis for the microstructure-mechanism map of Fig. 7 will now be considered. It is first assumed that the nucleation of new grains occurs only at or near existing grain boundaries. This is because the stored energy (dislocation density) is greater in these parts of the crystals, and also because less additional surface area need be created when a nucleus is formed. The relations between nucleus density N and grain size D are thus

$$N_0 = \beta P_0/D_0 \quad \text{and} \quad N_s = \beta P_s/D_s \qquad (6)$$

in the initial and steady states, respectively, where β is a geometrical constant. P_0 and P_s are the probabilities per unit surface area that a grain-boundary site is activated to form a nucleus. Since some recently recrystallized parts are unable to supply nuclei in the steady-state microstructure, P_s is lower than P_0; the experimental data in fact suggest that $P_0/P_s \approx 2$. Then, from Eqs. (6),

$$\frac{N_0}{N_s} = \frac{P_0}{P_s}\frac{D_s}{D_0} \approx 2\frac{D_s}{D_0}. \qquad (7)$$

Thus the relative-grain-size condition $D_0 = 2D_s$ is associated with the *nucleus-density condition*

$$N_0 = N_s. \qquad (8)$$

The transition between single- and multiple-peak behavior can now be rationalized with the help of the above criterion.

When $D_0 > 2D_s$ or $N_0 < N_s$, there is a *deficit* in the initial nucleus density with respect to its steady-state value. Because nucleation occurs at the existing grain boundaries, a "necklace" mechanism operates as illustrated in Fig. 9. Since $D_s < D_0/2$, the new grains reach their equilibrium size D_s before the whole structure has recrystallized. After the formation of the first strand of new grains, DRX continues to spread progressively toward the interiors of the deformed initial grains (Fig. 9). According to this model, at any instant in time, the different volume fractions of the sample are at different stages of recrystallization. DRX is thus a very heterogeneous process and the occurrence of softening is not at all synchronized at any

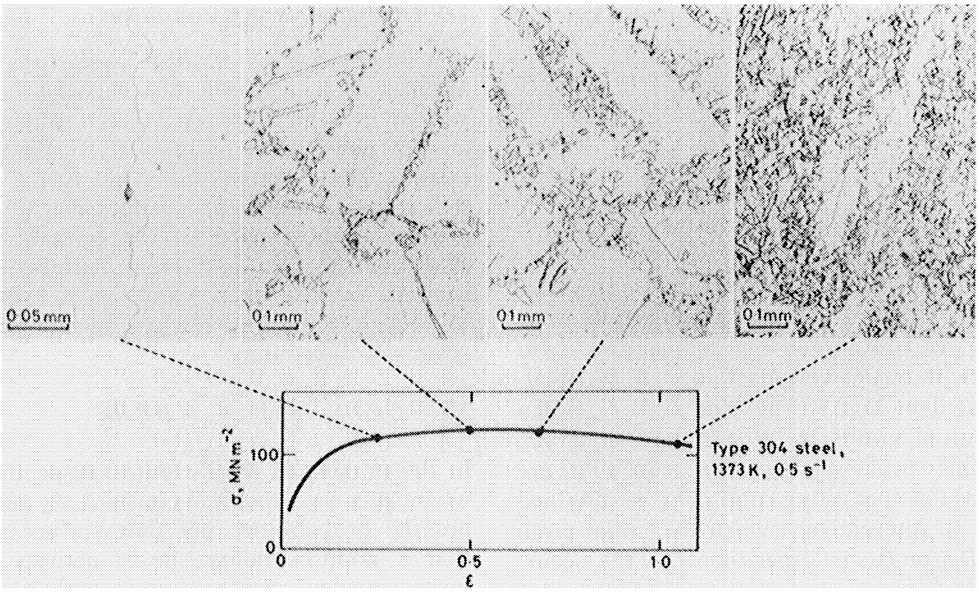

FIG. 9. Microstructures illustrating the progress of necklace dynamic recrystallization in type 304 austenitic stainless steel with a coarse initial grain size. The associated single-peak stress–strain curve is also shown. (After Roberts *et al.*, 1979.)

particular moment. Nevertheless, there is overall flow softening until the equilibrium grain size has been attained everywhere within the material, i.e., single-peak DRX occurs. The limiting case for this mechanism is a necklace containing a single strand, which corresponds to the relative–grain-size condition $D_0 = 2D_s$ or to the nucleus-density condition $N_0 = N_s$.

When $D_0 < 2D_s$ or $N_0 > N_s$, there is an *excess* in nucleus density. Since $D_s > D_0/2$, the growth of recrystallized grains is stopped by mutual impingement before they have reached their steady-state size D_s. Once the first cycle of recrystallization is complete, work hardening begins again, followed by a second cycle of recrystallization, and so on, until grain coarsening finally produces the stable grain size D_s. Such coarsening also progressively reduces the excess nucleus density until the stable level given by N_s is attained. It should be noted that, although this process generally leads to an increase in grain size (i.e., when $D_0 < D_s$), it can also be associated with a small amount of grain refinement (i.e., when $D_s < D_0 < 2D_s$), as illustrated in Fig. 5. Under these conditions, the progress of DRX is highly synchronized and the macroscopic multiple-peak flow curve readily reflects the local cycles of work hardening and recrystallization.

Unresolved Issues. Since the steady state is only attained at large strains ($\epsilon > 1$) and is rarely achieved during industrial processing, its characteristics have only been investigated experimentally in a few materials, such as Cu, Ni, and stainless steels (Ryan and McQueen, 1990). With regard to mathematical treatments, only a few tentative models have been proposed (Stüwe and Ortner, 1974; Sandström and Lagneborg, 1975). Although these can account for the stress dependence of the steady-state grain size, in their present states of development they are unable to distinguish between the grain-refinement and grain-coarsening modes of recrystallization. More recently, numerical models based on the Monte Carlo simulation technique have been used to study DRX (Rollett *et al.*, 1992). Although many of the essential features of the latter are faithfully represented, the relationship between flow-stress oscillations and grain coarsening has not yet been clearly reproduced.

During steady-state flow, both the flow stress and the microstructural parameters (e.g., the distributions of grain size and dislocation density) remain stationary (Fig. 10).

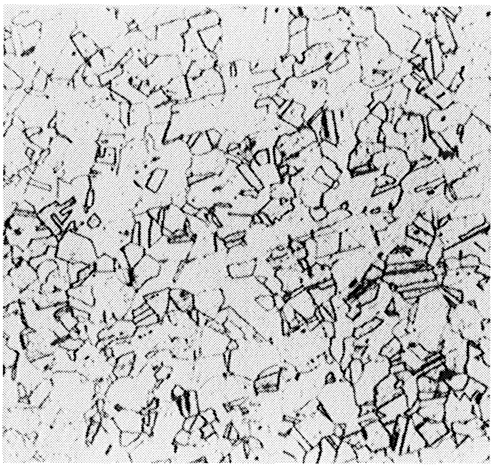

FIG. 10. Micrograph illustrating the steady-state microstructure associated with dynamic recrystallization in type 304 austenitic stainless steel deformed in torsion to $\epsilon \approx 5$ at 900 °C and 5 s^{-1}. The latter conditions correspond to single-peak flow and to grain refinement. (After Ryan and McQueen, 1990.)

Similar remarks apply to the crystallographic texture. The mechanism of nucleus formation is another topic that has been insufficiently explored. Nuclei can be created within grains from dislocation cells or subgrains (e.g., by "coalescence") or alternatively at grain boundaries, as assumed in the relative–grain-size model (e.g., by "grain-boundary bulging"). The role of oriented nucleation in the formation of nuclei remains to be clarified, as does that of selective growth. With the increasing availability and ease of use of equipment for EBSD (electron backscattering diffraction) measurements, considerable progress is expected in this area over the next few years.

1.4 Effect of Alloying Elements on DRX

The four steels compared in Fig. 11(a) all had the same initial grain size (110 μm). Their flow curves differ, however, because of differing amounts of V, Nb, or Mo in solid solution. The solutes decrease the mobility of grain boundaries in the course of DRX. This, in turn, increases the peak strain ϵ_p and decreases the steady-state grain size D_s, as shown in the figure. These effects are still more pronounced when precipitation occurs during DRX, as shown in Fig. 11(b), which pertains to the plain carbon, V, and Nb steels of Fig. 11(a) when deformed at a lower temperature. The possibility of delaying the onset of DRX by the addition of very small amounts of niobium or vanadium (e.g., 0.035% Nb or 0.115% V in the above examples) is what makes it possible to carry out the controlled rolling of HSLA (high-strength low-alloy) steels. Provided the accumulated strain in a given pass is less than the current value of ϵ_p, no DRX occurs and postdynamic recrystallization is also avoided (see below).

2. DYNAMIC RECOVERY AND CONTINUOUS DYNAMIC RECRYSTALLIZATION

The behavior described in the previous section concerned metals and alloys of medium or low stacking-fault energy (e.g., 10 to 50 mJ/m^2), such as copper, γ-iron, and austenitic steels. In these materials, the cross-slip of dislocations is relatively difficult, so that the ease of the rearrangement and removal of dislocations during deformation, i.e., *dynamic recovery*, is reduced (although it still plays an important role in the nucleation of new grains). By contrast, in materials of high stacking-fault energy (e.g., ≈100 mJ/m^2), such as aluminum, α-iron, and ferritic steels, dynamic recovery is so efficient that its operation can preclude the occurrence of discontinuous DRX. As conventional nucleation is unable to take place, the "new grains" are generated instead in a more gradual way by dynamic recovery followed by grain growth, as described in more detail below.

2.1 Flow Curves Associated with Continuous DRX

An example of such flow-stress behavior is presented in Fig. 12, where the results of hot-torsion experiments on grade 1100 aluminum carried out at 400 °C and various strain rates are reproduced (Perdrix *et al.*, 1981). The shapes of the stress–strain curves are similar at all strain rates: The strain-hardening stage leads to a smooth maximum, followed by a very slow decrease in the flow stress. Quite large strains are necessary in order to attain a true steady state, e.g., $\epsilon \approx 20$ in Fig. 12; these levels cannot be reached by other testing techniques such as

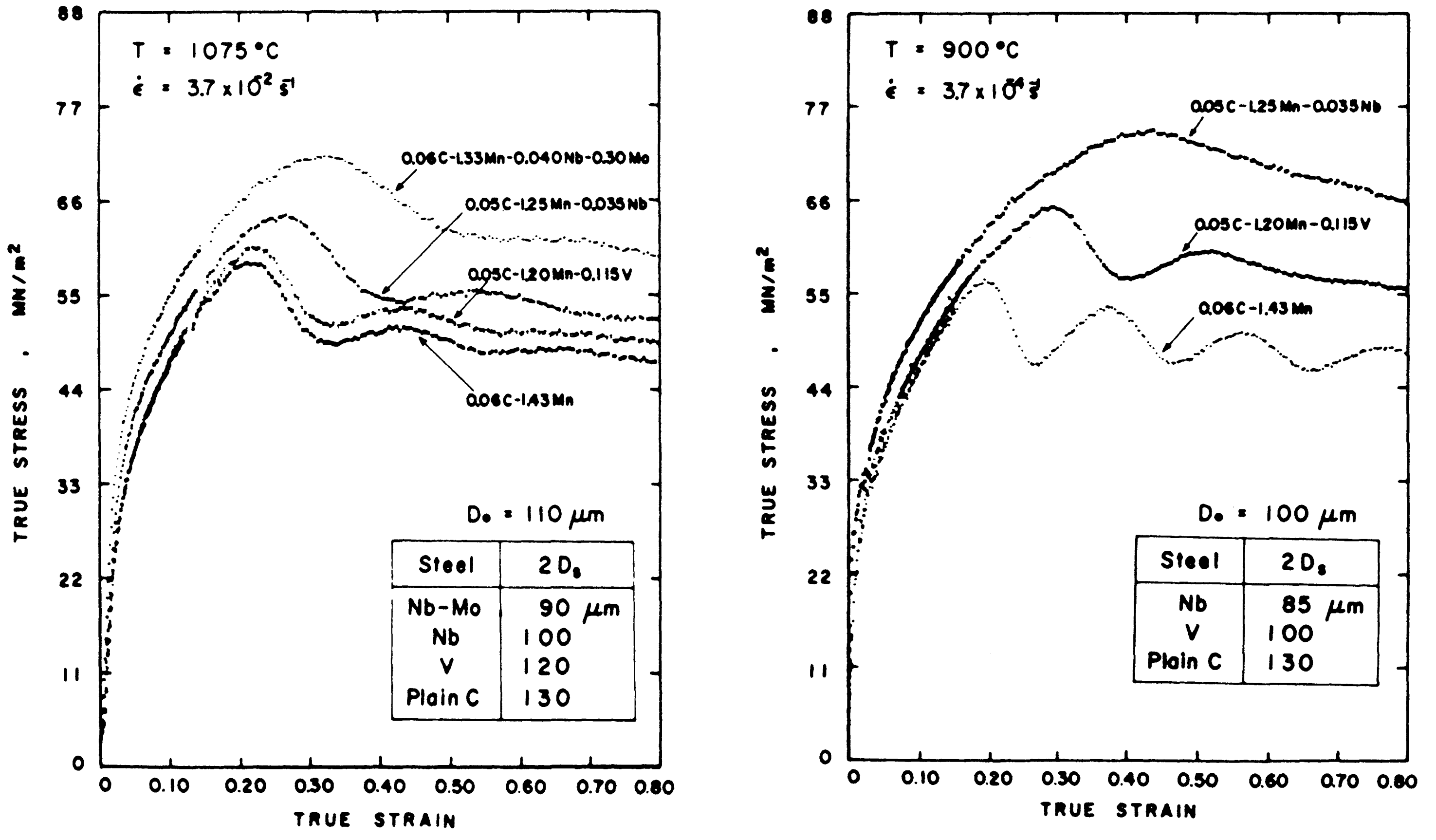

(a) (b)

FIG. 11. (a) Effect of the presence of Nb, Mo, and V in solution on the stress–strain curves associated with dynamic recrystallization of four steels with identical initial grain sizes, during deformation at the same strain rate and temperature (after Sakai *et al.*, 1982). (b) Effect of precipitation occurring during dynamic recrystallization on the stress–strain curves of the plain carbon, V, and Nb steels of (a), when deformed at a lower temperature (after Sakai and Jonas, 1986).

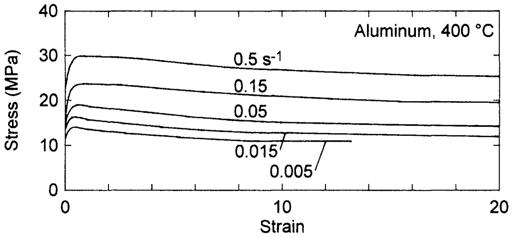

FIG. 12. Stress–strain curves typical of continuous dynamic recrystallization obtained on grade 1100 aluminum at 400 °C and various strain rates (after Perdrix *et al.*, 1981).

compression or tension. Similar behavior has been observed in α-iron and more recently during the testing of the high-temperature β phase (bcc) of titanium alloys. The smooth shape of the stress–strain curves, as well as the absence of oscillations at all temperatures and strain rates, clearly distinguishes the continuous from the discontinuous type of DRX process.

2.2 Microstructural Changes Associated with Continuous DRX

Dynamic recovery has been extensively investigated at low and moderate strains ($\epsilon < 1$), i.e., up to the maximum of the stress–strain curve. The essentially equiaxed *subgrains* formed by this process within the initial grains are illustrated in Fig. 13, obtained from an aluminum specimen tested in torsion at 400 °C to a final strain of 0.85 and rapidly quenched (Montheillet, 1981). The low misorientations that apply to these subgrains can be verified by using electron microdiffraction (or by the recently developed EBSD technique). The orientation of each subgrain with respect to the specimen axes can also be determined in this way. There are 12 different rotations θ_i (i.e., axes and angles of rotation) that can cause the orientations of two sets of rectangular coordinates to coincide. By convention, the misorientation that applies to two crystals is $\theta = \min(\theta_i)$. It can be shown that θ falls between 0 and ψ, where $\tan(\psi/2) = (\sqrt{2} - 1)(5 - 2\sqrt{2})^{1/2}$ in cubic structures, so that $\psi \approx 62.8°$ (Mackenzie, 1958). When $\theta < 15°$, the boundary is made up of one, two, or three sets of parallel dislocations and is called a *subgrain* or *low-angle boundary*. However, when $\theta > 15°$, the misorientation can no longer be accommodated by planar arrays of dislocations and is then referred to as a *grain* or *high-angle boundary*.

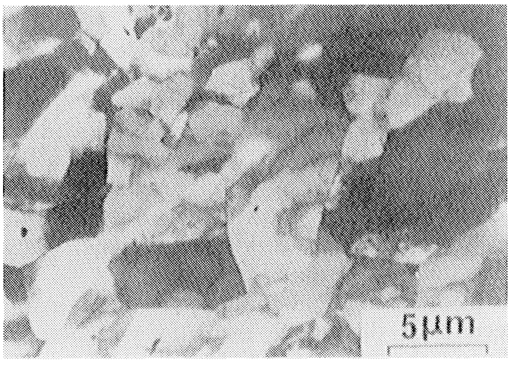

FIG. 13. Subgrain microstructure of an aluminum specimen tested in torsion at 400 °C and 4.2 × 10^{-2} s^{-1} to a strain of 0.85 (after Montheillet, 1981).

A schematic view of part of the specimen of Fig. 13 that was investigated in this way is presented in Fig. 14(a). The thick line represents a high-angle boundary, which was present in the starting material and has become somewhat serrated as a result of straining. The narrow lines in turn denote low-angle, i.e., subgrain, boundaries. The pole figure of Fig. 14(b) displays the orientations of the ⟨100⟩ (i.e., cube face) directions of the various crystals. These clearly form two families associated with the two original grains, A and B. Within each grain, the misorientations between neighboring subgrains remain small ($\theta < 15°$), a finding that is also illustrated in the misorientation-distribution diagram of Fig. 14(c).

Less work has been carried out on the dynamically recovered microstructures formed at large strains ($\epsilon > 1$) (e.g., Perdrix et al., 1981). An example of the microstructure developed at a strain of 40 in a specimen of grade 1100 aluminum twisted at 400 °C is presented in Fig. 15 (Montheillet, 1981). Comparison with the low-strain microstructure depicted in Fig. 13 shows that straight boundaries as well as 120° triple points (arrows) are much more frequent. The large-strain microstructure is further described in Fig. 16(a), from which it can be seen that numerous high-angle boundaries have formed within the material; as a result, the original 80-μm-diam grains can no longer be

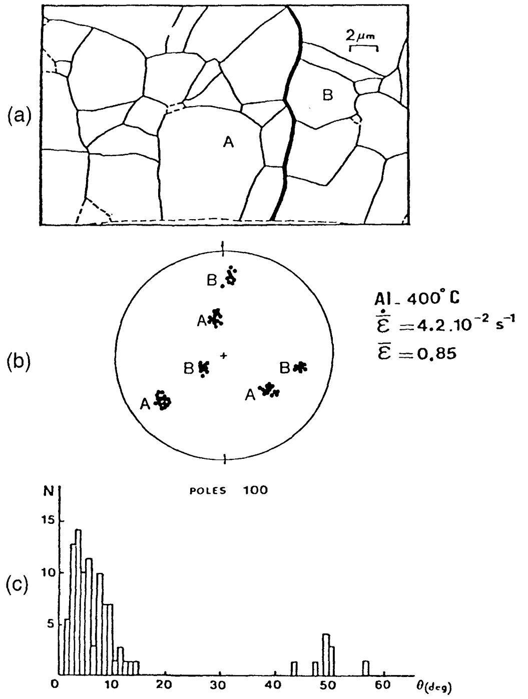

FIG. 14. (a) Schematic view of part of the specimen of Fig. 13 showing a high-misorientation boundary (thick line) and several low-angle, i.e., subgrain boundaries (narrow lines). (The dashed lines correspond to boundaries whose misorientations were not measured.) (b) (100) pole figure pertaining to the same area of the specimen. (c) Associated misorientation-distribution diagram. (After Perdrix *et al.*, 1981.)

identified. This is evident from Fig. 16(b) as well, where the ⟨100⟩ directions are seen to be located in several parts of the pole figure. Finally, Fig. 16(c) shows that the misorientations are more or less continuously distributed over the possible range.

More recently, similar results were obtained from compression tests carried out on grade 1200 aluminum specimens (Gourdet and Montheillet, 1994). In the latter case, a significant increase in the misorientation of the subgrains was observed at much lower strains ($\epsilon \approx 1$); the earlier increase in misorientation can be partly attributed to the different strain path that was employed (i.e., compression instead of torsion). It should be noted, however, that physical models remain to be developed both for the transition from low- to high-angle boundaries and for the evolution and motion of high-angle boundaries during straining.

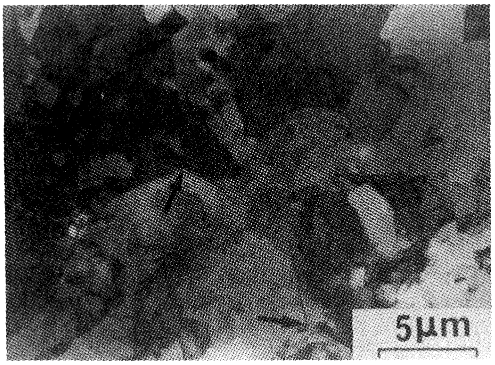

FIG. 15. Microstructure developed at a strain of 40 in a specimen of grade 1100 aluminum twisted at 400 °C and 4.2×10^{-1} s^{-1} (after Montheillet, 1981).

Two alternative models have been proposed to rationalize the above observations. Perdrix *et al.* (1981) first suggested that new high-angle boundaries form by a continuous process involving the absorption of mobile dislocations by the subgrain boundaries present at low strains. More recently, a different mechanism, referred to as *geometric dynamic recrystallization,* has been proposed by McQueen *et al.* (1989). According to these authors, the high-angle boundaries observed at large strains originate from the elongation and deformation of the boundaries present in the starting material. When the distance between two neighboring boundaries is reduced to the equilibrium size of the subgrains, pinched-off regions appear, which break up the elongated grains (Fig. 17). The repetition of this mechanism can lead eventually to the formation of an equiaxed microstructure. In contrast to the first mechanism, the second does not involve the formation of *new* high-angle boundaries. By using a simple geometrical model, it can be shown that at large torsional strains, the grain-boundary surface area is multiplied by a factor close to $\bar{\epsilon}$, the von Mises equivalent strain. Thus, starting from an initial grain size of 50 μm, an equivalent strain of about 10 is large enough to produce "geometrically recrystallized" grains of 5 μm.

In order to decide whether the first or the second of the above mechanisms operates during the hot deformation of materials of high stacking-fault energy, large-strain torsion experiments have been conducted on aluminum *single crystals* (Kassner, 1989).

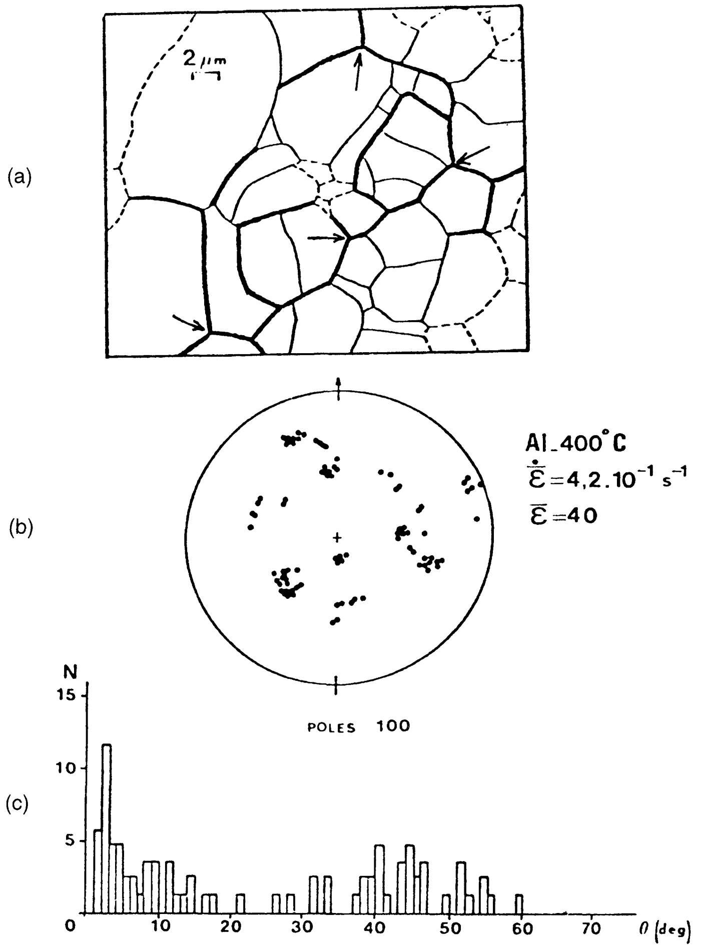

FIG. 16. (a) Schematic view of part of the specimen of Fig. 15 exhibiting numerous high-angle boundaries (thick lines) in addition to the low-angle boundaries (narrow lines). (The dashed lines correspond to boundaries whose misorientations were not measured.) (b) (100) pole figure corresponding to the same area of the specimen. (c) Corresponding misorientation distribution diagram. (After Perdrix *et al.*, 1981.)

Since only a limited number of high-misorientation interfaces was observed after straining, it was concluded that the much larger increase in the number of high-angle boundaries in deformed *polycrystals* was the result of geometric rather than continuous DRX. However, this interpretation is invalid at lower strains, in particular when the starting grain size is large, and new grains are still observed under these conditions (Gourdet and Montheillet, 1994). It therefore seems likely that solely the first mechanism operates at the strains involved in commercial processing, but that both mechanisms operate simultaneously at very large strains (e.g., $\epsilon > 10$) in polycrystalline materials.

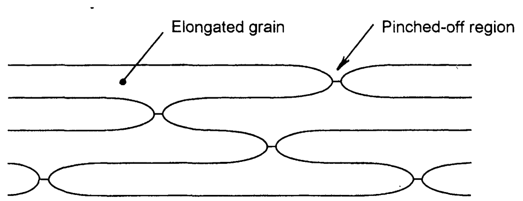

FIG. 17. Simplified schematic representation of the geometric dynamic-recrystallization mechanism.

Finally, it is of interest to note that the steady-state flow stress σ_s and stable grain or subgrain size D_s are related by a power law:

$$\sigma_s = k/D_s^{\alpha}, \tag{9}$$

where k and α are constants. Note that this relationship has the same form as in the case of discontinuous DRX [see Eq. (3)]. At low strains (i.e., when a subgrain structure is present), a number of investigations have shown that $\alpha \approx 1$. By contrast, at large strains (mixture of low- and high-angle boundaries), the available results (Montheillet, 1981) indicate that $\alpha \approx 0.75$; this suggests that a different mechanism is controlling the formation of the high-angle boundary microstructure.

By combining Eqs. (1) and (4) above, the following relationship is obtained between the steady state flow stress σ_s, strain rate $\dot{\epsilon}$, and absolute temperature T:

$$\sigma_s = A\dot{\epsilon}^m \exp(mQ/RT), \qquad (10)$$

where A is a constant. Although the strain-rate sensitivity exponent m generally increases with temperature, it is often considered to be constant to a first approximation. In such cases, the material is considered to be *power-law viscoplastic.* For both types of DRX, experimental values of m generally fall within the range 0.1 to 0.2. The temperature dependence is in turn a special form of the more general Arrhenius law, according to which Q denotes the apparent activation energy of deformation. In the case of dynamic recovery and continuous recrystallization, Q is close to the activation energy for self-diffusion (e.g., $Q \approx 150$ kJ/mol in aluminum). This can be attributed to the occurrence of dislocation climb in the subgrain boundaries, which involves self-diffusion, and which is considered to control the evolution of the microstructure. However, when discontinuous recrystallization occurs, appreciably greater values of Q are generally observed (e.g., $Q \approx 300$ kJ/mol in copper). Such a difference in temperature dependence of the flow stress has been associated with an accommodation process controlled by solute drag and by grain-boundary migration.

3. TEXTURE CHANGES ASSOCIATED WITH DRX

The changes in preferred orientation that take place during discontinuous or continuous DRX have long been of interest, in part because their control could have useful applications in deformation processing. Nevertheless, experimental data on the development of such crystallographic "textures" are scarce, particularly in the form of detailed ODFs (orientation distribution functions), in part because the rapidity of postdynamic recrystallization makes it difficult to determine the exact state of the metal at the moment when straining was completed.

Early results obtained from torsion experiments by Montheillet *et al.* (1984) indicate that, in the case of aluminum, where only continuous DRX takes place, a sharp texture develops at 400 °C [Fig. 18(a)]. The latter consists of the "twin symmetric" $\{\bar{1}12\}\langle 110\rangle/\{1\bar{1}\bar{2}\}\langle\bar{1}\bar{1}0\rangle$ ideal orientation referred to as the $B/\bar{B}$ component, where $\{hkl\}$ and $\langle uvw\rangle$ are the shear plane and shear direction, respectively [Fig. 18(d)]. In the case of copper deformed at 200 °C, although the $B/\bar{B}$ component is still evident, the $A/\bar{A}$ component, viz., $\{1\bar{1}\bar{1}\}\langle 110\rangle/\{\bar{1}11\}\langle\bar{1}\bar{1}0\rangle$, is considerably sharper [Figs. 18(b) and 18(d)]. At higher temperatures (300 to 500 °C), the $A/\bar{A}$ and $B/\bar{B}$ orientations continue to be present, but they appear to be more and more scattered [Fig. 18(c)]. This was attributed by the authors to the effect of more and more rapid discontinuous DRX, which tends to consume the deformation texture, although some amount of postdynamic recrystallization may also have occurred after straining. More recently, Tóth and Jonas (1992), Jonas and Tóth (1992), and Tóth *et al.* (1992) have characterized the textures developed during the hot torsion of copper bars in considerably more detail. For this, they employed relatively low deformation temperatures (200 and 300 °C) and fairly rapid quenches. Analysis of the ODFs confirmed that there is a considerable intensification of the $A/\bar{A}$ component with increasing temperature [Figs. 19(a) and 19(b)].

In order to account for the texture changes produced by high-temperature deformation, the above authors considered that two complementary recrystallization mechanisms take place, viz., oriented nucleation and selective growth. In this case, *oriented nucleation* concerns the preferred nucleation of particular orientations among the deformed grains; these are the ones that recover more easily than the others, and thus form many of the recrystallization nuclei. These grains (or parts of grains) have the lowest possible Taylor factor M. This is because, for a given von Mises equivalent strain increment $\Delta\epsilon$, the lowest M grains have experienced the least total microscopic slip $\Delta\Gamma$, since

$$\Delta\Gamma = M\Delta\epsilon. \qquad (11)$$

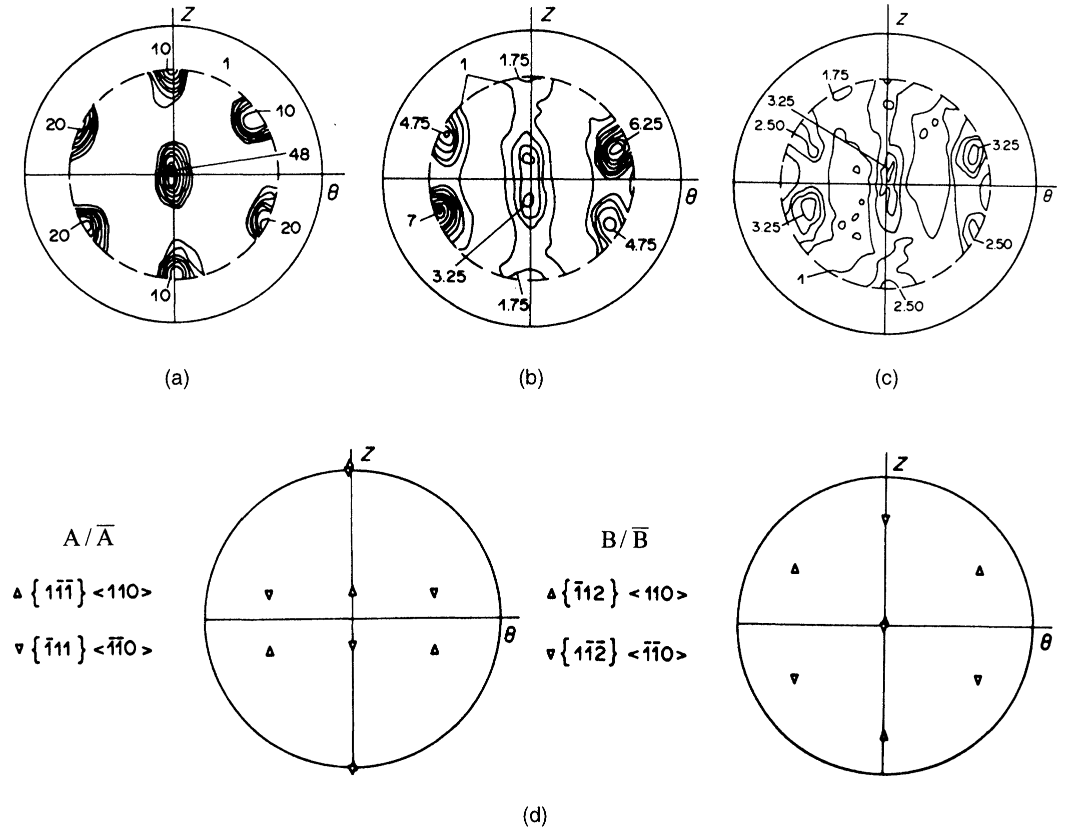

(a) (b) (c)

(d)

FIG. 18. (111) pole figures illustrating the crystallographic textures obtained from large-strain hot-torsion experiments: (a) aluminum, 400 °C; (b) copper, 200 °C; (c) copper, 400 °C; (d) *A*/*Ā* and *B*/*B̄* ideal orientations. (After Montheillet *et al.*, 1984.)

They are therefore likely to have undergone the least work hardening or increase in dislocation density, and so to be the most highly recovered at any particular strain. For the torsion (i.e., simple shear) of fcc materials, the lowest possible value of M (1.73) corresponds to the $A/\bar{A}$ component, whereas for the $B/\bar{B}$ component, $M = 2.45$.

Selective growth in turn involves the rapid increase in size of nuclei characterized by particular misorientations with respect to the matrix. In fcc metals, the orientations of these nuclei generally differ by ±40° rotations around appropriate ⟨111⟩ directions. In their simulations, Jonas and Tóth (1992) assumed that the reorientation axis is perpendicular to the most active slip plane. This choice is based on the high mobility of grain boundaries perpendicular to active slip planes, which generally contain planar arrays of dislocations and subboundaries. The high mobility is caused by impurity diffusion away from these boundaries by pipe diffusion along the planar arrays.

Comparison of the experimental and simulated textures allowed the relative importance of oriented nucleation and selective growth to be estimated. In the case of copper twisted at 300 °C, it was found that the first mechanism was the more important, in a ratio of about 2:1 [Fig. 19(c)]. Oriented nucleation can thus account for the high intensity of the minimum–Taylor-factor $A/\bar{A}$ component in copper twisted at 300 °C. By contrast, when the temperature is increased to 400 °C, as in Fig. 18(c), selective growth is likely to become more important, thus accounting for the more scattered texture evident in the latter figure. Finally, the development of the $B/\bar{B}$ (medium–Taylor-factor)

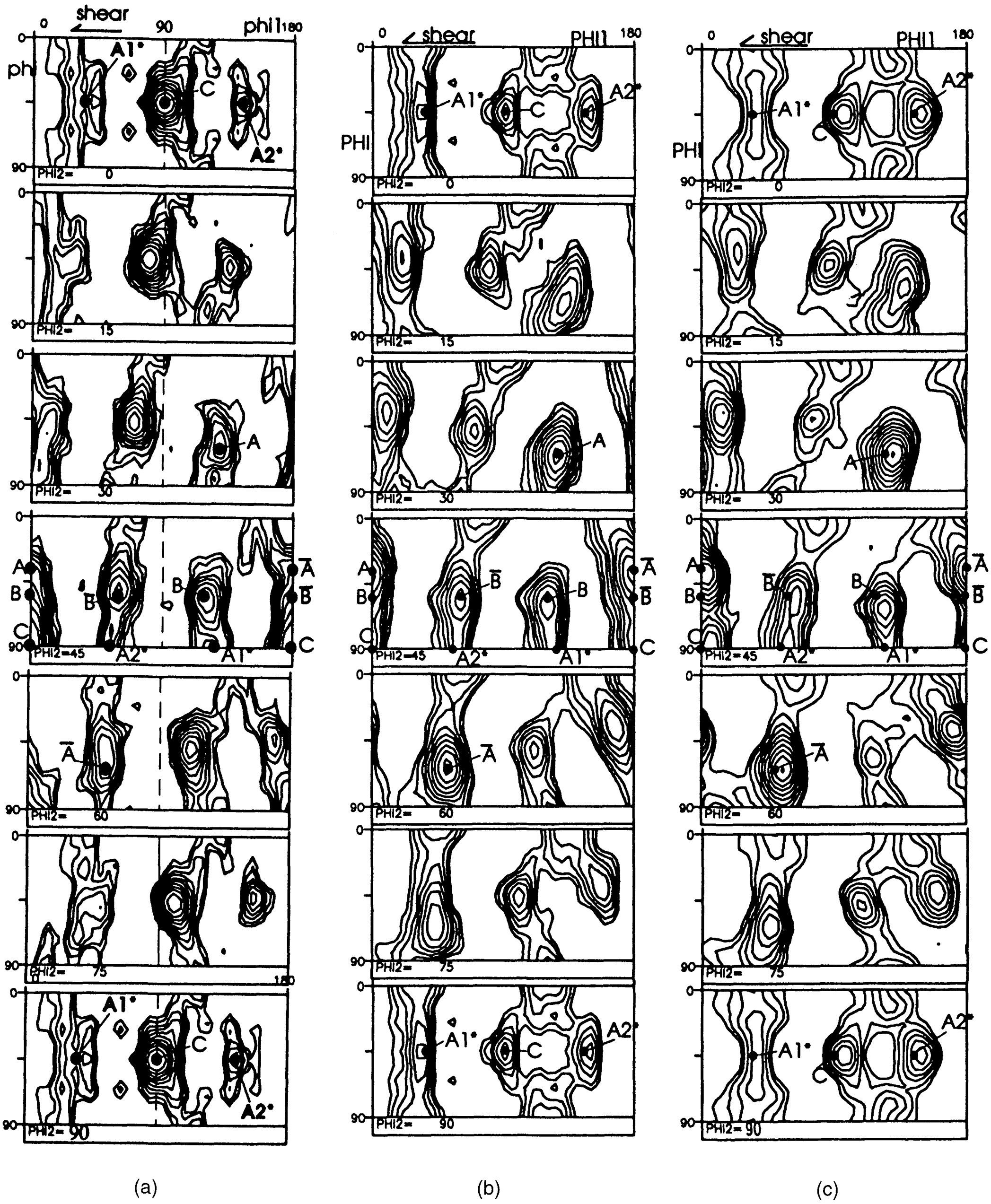

FIG. 19. Orientation-distribution functions characterizing the textures developed during the torsion testing of copper. (a) Experimental texture developed at room temperature at a strain of about 3.2. (b) Experimental texture developed at 300 °C and a strain of about 2.3. (c) Texture obtained from computer simulation under conditions corresponding to (b). (After Jonas and Tóth 1992.)

component in aluminum deformed in torsion at 400 °C still remains unexplained.

4. POSTDYNAMIC SOFTENING

As already mentioned, postdynamic restoration often takes place very rapidly in the interpass intervals of multiple-pass industrial processing, or during the final cooling stage. As a result, it can strongly influence the microstructure (e.g., grain size, texture), and therefore the properties (e.g., yield strength, fracture toughness, plastic anisotropy), of the final product. In fact, it is the characteristics of postdynamic, rather than dynamic, recrystallization that are often more relevant to industrial hot working.

The postdynamic softening mechanisms were extensively studied by Petkovic *et al.* (1975, 1979) by means of *interrupted-compression tests* carried out on several kinds of steel and copper specimens. In such experiments, a sample is first deformed at a given strain rate and temperature. It is then unloaded and held for a time t at the same temperature. Finally it is reloaded, thus allowing the softening produced by the postdynamic restoration to be measured. Samples can also be quenched for subsequent metallographic examination (although these are not easy to interpret in steel specimens because of the martensitic transformation).

Whereas only two distinct softening mechanisms can take place in a specimen held at temperature after cold deformation, *three* have been identified when hot deformation is involved:

1. static recovery;
2. conventional static recrystallization; and
3. a type of recrystallization termed *metadynamic* or *postdynamic.*

The microstructural aspects of the first two processes are exactly those of their counterparts following cold deformation. By contrast, metadynamic recrystallization takes place only if discontinuous DRX has occurred during the previous interval of hot deformation. It is not preceded by an appreciable incubation period, since it relies on further growth of the nuclei already formed during DRX. Furthermore, it proceeds much more quickly than classical recrystallization. The kinetic equations for these three mechanisms, however, are all of the Avrami type, i.e.,

$$X_i = X_i^{\infty}[1 - \exp(-k_i t^{n_i})]. \qquad (12)$$

In this equation, $X_i = X_r, X_R,$ or X_M is the fractional softening that can be associated with static recovery, static recrystallization, or metadynamic recrystallization, respectively; t denotes the time measured from the onset of the restoration process; and X_i^{∞} and k_i are parameters that depend on the detailed conditions of prior straining and holding. Experimental measurements of the exponent n_i for the hot working of copper led to the following values (Luton *et al.*, 1980): $n_r = 1.33$ (whereas $n_r = 1$ after cold working), $n_R = 2.2$, and $n_M = 1$ (to be compared with $n = 3$ to 4 in the original Avrami theory).

The progress of softening as described by Eq. (12) is illustrated in Figs. 20(a) and 20(b), which show the time dependence of the softening occurring in copper after deformation at 500 °C. In Fig. 20(a), prior straining was interrupted *before* the onset of discontinuous DRX. Accordingly, the subsequent softening is produced solely by static recovery (X_r) and static recrystallization (X_R), as shown in the figure. In Fig. 20(b), the initial compression test was stopped *after* the initiation of DRX, so that metadynamic recrystallization took place during the subsequent holding interval. In this case, static softening is described by the sum of the three components X_r, X_R, and X_M.

Metadynamic recrystallization leads to some coarsening (by a factor of about 1.5) of the dynamically recrystallized grains, as shown in Fig. 21 (Hodgson *et al.*, 1992). This diagram is based on data obtained from torsion tests carried out on 0.15% C steel specimens. It can be seen that the sizes of both the dynamically and metadynamically recrystallized grains decrease as the strain rate of the prior deformation is increased and/or the temperature is lowered (i.e., as Z is increased). It is of practical importance that, under normal hot-working conditions, the grain size produced by metadynamic recrystallization is 2 to 3 times finer than that resulting from conventional static recrystallization.

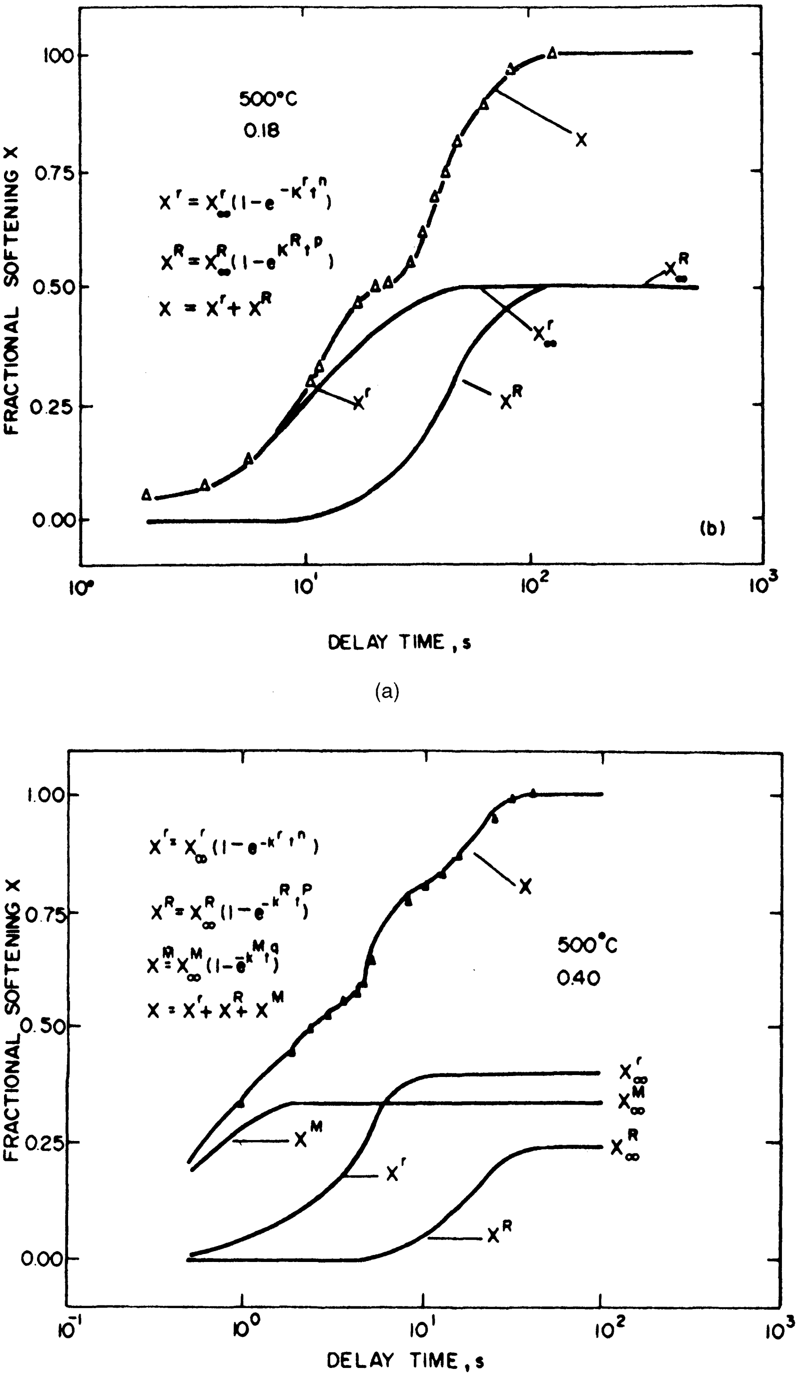

(a)

(b)

FIG. 20. Progress of softening following the hot compression of copper specimens at 500 °C to a strain of (a) 0.18, i.e., before the onset of DRX, and (b) 0.40, i.e., after DRX has started. The curves that follow the experimental data points correspond to the sum of the X_r and X_R components in the first case, and of the X_r, X_R, and X_M components in the second. These are associated with static recovery, static recrystallization, and metadynamic recrystallization, respectively. (After Luton *et al.*, 1980.)

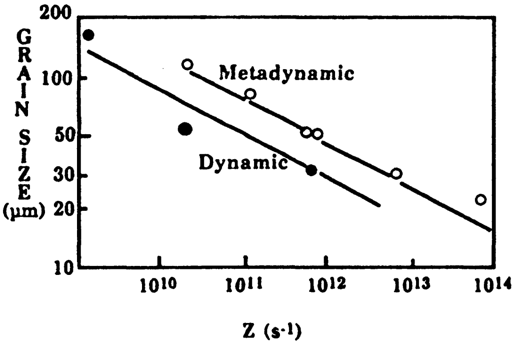

FIG. 21. Dependence of the grain sizes produced by dynamic and metadynamic recrystallization on the temperature-corrected strain rate Z (after Hodgson *et al.*, 1992).

5. IMPLICATIONS FOR INDUSTRIAL PROCESSING

In industrial practice, DRX plays a role in two types of processes. In the first, large reductions are applied in a single uninterrupted pass or operation. This description applies to planetary hot mills and extrusion presses. In the second, although multipass operations are involved, the interpass times are relatively short (i.e., <1 s). Under these conditions, there is insufficient time for interpass softening, so that strain hardening is accumulated at first, until the critical strain for DRX is attained. During subsequent straining, the flow stress remains relatively low, on account of continued dynamic recrystallization during deformation, supplemented by the occurrence of metadynamic (postdynamic) recrystallization during the interpass intervals.

The interpass softening produced by metadynamic recrystallization increases not only with the length of the interpass interval, but also with the strain rate of the previous pass. This dependence can be described by the following relationship:

$$t_{50} = A\dot{\epsilon}^{-p}\exp(Q/RT), \qquad (13)$$

where t_{50} is the hold time for 50% softening, A and p are material constants ($p \approx 0.6$), and Q is an apparent activation energy (Roucoules *et al.*, 1994). Although interpass times are progressively reduced during typical rolling operations, the strain rate simultaneously increases from pass to pass. Thus the rate of postdynamic softening also increases from pass to pass.

When little postdynamic recrystallization takes place, such low-temperature finishing has been referred to as dynamic-recrystallization–controlled rolling (DRCR). Under these conditions, the final grain size is controlled by DRX (Roucoules *et al.*, 1994; Jonas, 1994). Conversely, when there is nonnegligible postdynamic softening, it has been referred to as metadynamic-recrystallization–controlled rolling (MDRCR), in which case the final microstructure is a result of postdynamic recrystallization. The grain sizes produced by both DRX and DRX followed by metadynamic recrystallization are much finer than those attributable to the static recrystallization that takes place in conventional rolling, i.e., when longer interpass times are involved. Finally, when the rate of softening is so fast that appreciable postdynamic softening takes place, the continued occurrence of DRX is no longer possible, and the process of strain accumulation is reinitiated, as in the first few passes of DRCR and MDRCR rolling.

GLOSSARY

Apparent Activation Energy: Parameter (expressed in kJ/mol) that specifies the temperature sensitivity of the flow stress in an Arrhenius law.

Arrhenius Law: An exponential relationship between the rate of a process and the inverse absolute temperature. This is used in the present context to relate the flow stress to the inverse temperature when constant testing strain rates are used.

Avrami Equation: Relationship between the recrystallized volume fraction and time used to describe the kinetics of recrystallization.

Bulging (Grain Boundary): Local migration of a high-misorientation grain boundary, which moves into a neighboring grain containing a higher dislocation density. Such migration can lead to nucleus formation, particularly in lightly deformed materials.

Coalescence (of Subgrains): The conversion of two or more neighboring subgrains into a single subgrain of higher misorientation, by the annihilation of their common (low-misorientation) subgrain boundary. This can lead to nucleus formation and applies particularly to highly deformed materials.

Continuous Dynamic Recrystallization: DRX occurring by a progressive increase in misorientation of the subgrains formed by dynamic recovery.

Critical Resolved Shear Stress: Shear stress required to activate a slip system, i.e., to produce dislocation glide on that system.

Cross-Slip: Glide of the screw component of a dislocation when it leaves its original slip plane to travel along another plane inclined at an angle to the first.

Dislocation Climb: Movement of a dislocation in a direction perpendicular to its slip plane. Since it requires vacancy migration, rapid climb only takes place at elevated temperatures.

Dynamic Recovery: The formation of lower-energy configurations by the partial removal and rearrangement of dislocations during straining at elevated temperatures. This leads to the formation of fairly well-defined subgrain boundaries.

Dynamic Recrystallization: Nucleation and growth of initially dislocation-free new grains; this occurs *during* straining at elevated temperatures.

Geometric Dynamic Recrystallization: DRX occurring at very large strains by the elongation and subsequent pinching off and fragmentation of elongated grains.

Ideal Orientation: Low-index crystallographic orientation in the vicinity of which is located a significant volume fraction of the grains in a given specimen. Also referred to as a *texture component.*

Metadynamic Recrystallization: Type of static recrystallization that follows discontinuous DRX after the cessation of straining. Also referred to as *postdynamic recrystallization.*

Miller Index: Triplet of integer numbers used to specify the crystallographic orientation of a lattice plane or direction.

Nucleation: Generation of a nucleus, which is a small dislocation-free grain surrounded by mobile high-angle boundaries. The three most common mechanisms are grain-boundary bulging, subgrain coalescence, and particle-stimulated nucleation.

Nucleus: Small dislocation-free grain bounded by a high-misorientation boundary that is able to grow into the strain-hardened matrix.

Orientation-Distribution Function: Scalar function specifying the volume fraction of grains possessing the various possible orientations pertaining to "orientation space." Here the grain orientations of interest are expressed in terms of the Euler angles that relate a given orientation to the orientation of the specimen axes.

Self-Diffusion: Thermally activated process of bulk (or matrix) diffusion in a metal; it involves the migration of vacancies.

Stacking-Fault Energy: Parameter associated with the mobility (ease of cross-slip) of dislocations. When the stacking-fault energy is low, dislocations are widely dissociated and their mobility is low, and vice versa.

Steady-State Flow: Dynamic equilibrium that can be achieved at large strains at elevated temperatures. It is associated with both a steady flow-stress level and a statistically time-independent microstructure.

Strain Path: Type of deformation imposed on or experienced by a material, e.g., uniaxial tension or compression, plane-strain compression, simple shear (torsion).

Strain-Rate Sensitivity: Dimensionless parameter that specifies the strain-rate sensitivity of the flow stress; it involves the ratio of the logarithm of the flow-stress increase to the logarithm of the strain-rate increase.

Taylor Factor: Ratio of the macroscopic flow stress to the critical resolved shear stress for dislocation glide within a grain. It also relates the total microscopic slip taking place within a grain to the macroscopic strain.

Zener–Hollomon Parameter: Function of strain rate and temperature that accounts for the equivalent effects of these two parameters on the flow stress at high temperatures. Also referred to as the temperature-corrected strain rate.

Works Cited

Blaz, L., Sakai, T., Jonas, J. J. (1983), *Metal Sci.* **17,** 609–616.

Derby, B. (1991), *Acta Metall. Mater.* **39,** 955–962.

Fields, D. S., Jr., Backofen, W. A. (1957), *Proc. ASTM* **57,** 1259–1272.

Gourdet, S., Montheillet, F. (1994), in: "Microstructures et recristallisation," 37ème Colloque de Métallurgie de l'INSTN, Saclay, *J. Phys. (Paris)* **5,** Suppl. JPIII, No. 4, 255.

Hodgson, P. D., Jonas, J. J., Yue, S. (1992), in: A. J. DeArdo (Ed.), *Processing, Microstructure and Properties of Microalloyed and other Modern HSLA Steels*, Warrendale, PA: Iron and Steel Society of the Metallurgical Society of AIME, p. 41.

Jonas, J. J. (1994), *Mater. Sci. Eng. A* **184,** 155–165.

Jonas, J. J., Canova, G. R., Shrivastava, S. C., Christodoulou, N. (1982), in: E. H. Lee, R. L. Mallett (Eds.), *Plasticity of Metals at Finite Strain: Theory, Experiment and Computation,* Stanford University, CA, p. 206.

Jonas, J. J., Tóth, L. S. (1992), *Scripta Metall. Mater.* **27,** 1575–1580.

Kassner, M. E. (1989), *Metall. Trans. A* **20,** 2182–2185.

Luton, M. J., Petkovic, R. A., Jonas, J. J. (1980), *Acta Metall.* **28,** 729–743.

Luton, M. J., Sellars, C. M. (1969), *Acta Metall.* **17,** 1033–1043.

Mackenzie, J. K. (1958), *Biometrika* **45,** 229–240.

McQueen, H. J., Solberg, J. K., Ryum, N., Nes, E. (1989), *Philos. Mag. A* **60,** 473–485.

Montheillet, F. (1981), in: P. Costa, A. Herpin, J. C. Jousset, P. Lacombe, A. Le Bon, G. Pomey, J. L. Strudel (Eds.), *Les traitements thermomécaniques,* 24ème Colloque de Métallurgie de Saclay, Paris, France: INSTN, pp. 57–70.

Montheillet, F., Cohen, M., Jonas, J. J. (1984), *Acta Metall.* **32,** 2077–2089.

Perdrix, Ch., Perrin, M. Y., Montheillet, F. (1981), *Mém. Sci. Rev. Metall.* **78,** 309–320.

Petkovic, R. A., Luton, M. J., Jonas, J. J. (1975), *Can. Metall. Quart.* **14,** 137–145.

Petkovic, R. A., Luton, M. J., Jonas, J. J. (1979), *Acta Metall.* **27,** 1633–1648.

Poirier, J.-P. (1972), *Philos. Mag.* **26,** 713–725.

Roberts, W., Boden, H., Ahlblom, B. (1979), *Met. Sci.* **13,** 195–205.

Rollett, A. D., Luton, M. J., Srolovitz, D. J. (1992), *Acta Metall. Mater.* **40,** 43–55.

Rossard, C., Blain, P. (1959), *Mém. Sci. Rev. Metall.* **56,** 285–299.

Roucoules, Ch., Hodgson, P. D., Yue, S., Jonas, J. J. (1994), *Metall. Trans. A,* **25,** 389–400.

Ryan, N. D., McQueen, H. J. (1990), *Can. Metall. Quart.* **29,** 147–162.

Sakai, T., Akben, M. G., Jonas, J. J. (1982), in: A. J. De Ardo, G. A. Ratz, P. J. Wray (Eds.), *Thermomechanical Processing of Microalloyed Austenite,* Warrendale, PA: AIME, pp. 237–252.

Sakai, T., Jonas, J. J. (1984), *Acta Metall.* **32,** 189–209.

Sakai, T., Jonas, J. J. (1986), in: N. Hansen, D. Juul Jensen, T. Leffers, B. Ralph (Eds.), *Annealing Processes—Recovery, Recrystallization and Grain Growth,* 7th Risø International Symposium on Metallurgy and Materials Science, Roskilde, Denmark: Risø National Laboratory, p. 143.

Sakui, S., Sakai, T., Takeishi, K. (1977), *Trans. Iron Steel Inst. Jpn.* **17,** 718–725.

Sandström, R., Lagneborg, R. (1975), *Acta Metall.* **23,** 387–398.

Steinemann, S. (1958), *Beiträge Geol. Schweiz (Hydrol.)* **10,** 1–72.

Stüwe, H. P. (1968), in: W. J. McG. Tegart, C. M. Sellars (Eds.), *Deformation under Hot Working Conditions,* ISI Special Report 108, London: Iron and Steel Institute, p. 1.

Stüwe, H. P., Ortner, B. (1974), *Met. Sci.* **8,** 161–167.

Thomsen, E. G., Yang C. T., Bierbower, J. B. (1954), in: *An Experimental Investigation of the Mechanics of Plastic Deformation of Metals,* Berkeley and Los Angeles: University of California Press, p. 89, quoted by Thomsen, E. G., Yang, C. T., Kobayashi, S. (1965), *Mechanics of Plastic Deformation in Metal Processing,* New York: The MacMillan Company, p. 113.

Tóth, L. S., Jonas, J. J. (1992), *Scripta Metall. Mater.* **27,** 359–363.

Tóth, L. S., Jonas, J. J., Daniel, D., Bailey, J. A. (1992), *Textures and Microstructures,* **19,** 245–262.

Weiss, I., Sakai, T., Jonas, J. J. (1984), *Met. Sci.,* **18,** 77–84.

Zener, C., Hollomon, J. H. (1944), *Trans. Am. Soc. Met.* **33,** 163–235.

Further Reading

Jonas, J. J., Bieler, R., Bowman, K. (Eds.), (1994), *Advances in Hot Deformation Textures and Microstructures,* Warrendale, PA: The Minerals, Metals, and Materials Society.

Krauss, G. (Ed.) (1984), *Deformation, Processing and Structure* Metals Park, OH: American Society for Metals, especially Chap. 4, "Dynamic Changes That Occur during Hot Working and Their Significance Regarding Microstructural Development and Hot Workability" (W. Roberts), Chap. 5, "A New Approach to Dynamic Recrystallization" (J. J. Jonas and T. Sakai), and Chap. 6, "Static Recrystallization and Precipitation during Hot Rolling of Microalloyed Steels" (C. M. Sellars).

H. J. McQueen, Jonas, J. J. (1975), "Recovery and Recrystallization during High Temperature Deformation," in: R. J. Arsenault (Ed.), *Plastic Deformation of Materials,* New York: Academic, pp. 393–493.

Yue, S. (Ed.) (1990), *Mathematical Modelling of Hot Rolling of Steel,* Montréal: Canadian Institute of Mining and Metallurgy.

RECRYSTALLIZATION, STATIC

F. J. Humphreys, *Manchester Materials Science Centre, Manchester, United Kingdom.*

Introduction 227
1. **The Driving Force for Recrystallization** 228
2. **The Fundamental Mechanisms of Annealing** 229
2.1 Recovery 229
2.1.1 The Mechanisms of Dislocation Recovery 229
2.1.2 Subgrain Growth 232
2.2 Recrystallization of Single-Phase Alloys 232
2.2.1 Kinetics of Primary Recrystallization 232
2.2.2 The Nucleation of Recrystallization 233
2.2.2.1 Strain-Induced Grain-Boundary Migration 233
2.2.2.2 Preformed-Nucleus Model 234
2.2.2.3 Nucleation Sites 234
2.2.3 Computer Modeling of Recrystallization 234
2.3 Migration of High-Angle Grain Boundaries 235
2.3.1 Experimental Observations 235
2.3.2 The Effect of Variables on Boundary Mobility 235
3. **Recrystallization of Two-Phase Alloys** 236
3.1 Particle-Stimulated Nucleation of Recrystallization 236
3.2 Pinning Effects of Particles (Zener Drag) 238
3.3 Bimodal Alloys and the Prediction of Grain Size 238
3.4 Interaction of Precipitation and Recrystallization 239
4. **Grain Growth after Primary Recrystallization** 239
4.1 Factors Affecting Grain Growth 240
4.2 Computer Simulation of Grain Growth 240
4.3 Secondary Recrystallization 241
5. **Recrystallization Textures** 241
5.1 Orientation Effects in Nucleation 241
5.1.1 Transition Bands 241
5.1.2 Shear Bands 242
5.1.3 Prior Grain Boundaries 242
5.1.4 Second-Phase Particles 242
5.2 Orientation Effects in Growth 242
5.3 Competition between Texture Components 242
Glossary 243
Works Cited 243
Further Reading 244

INTRODUCTION

During plastic deformation of a crystalline solid, dislocations and point defects are generated in the material. Although thermodynamics would suggest that these defects should spontaneously disappear, in practice, the kinetics of the diffusion processes by which these defects migrate are often very slow at low homologous temperatures, with the result that the defects are retained after deformation. However, if the material is subsequently heated to a high temperature (or annealed), then solid-state diffusion provides a mechanism whereby the defects may be removed or alternatively arranged in lower energy configurations.

The material may be partially restored to its original state by recovery, in which annihilation and rearrangement of the dislocations occur. The changes in microstructure are shown schematically in Figs. 1(a) and 1(b). Similar softening processes (dynamic recovery) may also occur during the deformation, particularly at high temperatures.

There is a further annealing process called recrystallization in which new dislocation-free grains are formed within the de-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

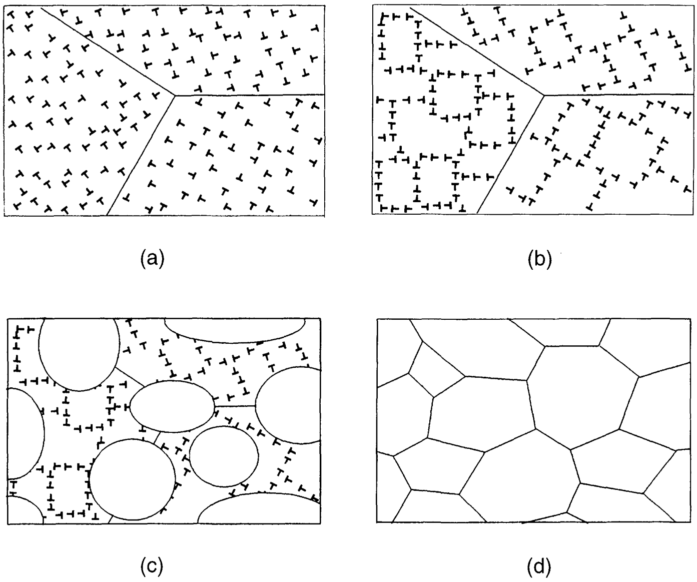

FIG. 1. Schematic sketch of recovery and recrystallization. (a) Deformed microstructure, (b) recovered structure, (c) partial recrystallization, (d) fully recrystallized specimen.

formed or recovered structure. These then grow and consume the old grains, resulting in a new grain structure with a low dislocation density [Figs. 1(c) and 1(d)].

Although recrystallization removes the dislocations, the material still contains grain boundaries, which are thermodynamically unstable. Further annealing may result in grain growth, in which smaller grains are eliminated, larger grains grow, and the grain boundaries assume a lower-energy configuration. In certain circumstances this uniform grain growth may give way to the selective growth of a few large grains, a process known as abnormal grain growth or secondary recrystallization.

The understanding and control of recrystallization is of particular importance in the mechanical processing of alloys. To a large extent the mechanical properties and behavior of a metal depend on the dislocation structure, the size of the grains, and the orientation or texture of the grains. The first of these parameters is dependent on the amount of deformation, the temperature of deformation, and the subsequent annealing treatment as well as other material and microstructural parameters. The grain size and texture are controlled by the recrystallization process.

1. THE DRIVING FORCE FOR RECRYSTALLIZATION

The nature of the deformed material is a key element in determining the annealing behavior, because the stored energy of deformation provides the driving force for recovery and recrystallization, while the mechanisms of recrystallization are determined by the microstructure of the deformed material, and in particular the microstructural heterogeneities.

Most of the work expended in deforming a metal is given out as heat, with typically only some 1% being stored in the form of dislocations and, at low deformation temperatures, point defects. In a single-phase single crystal, dislocation storage depends on the trapping of dislocations by others, and this is not easy to predict from first principles. The density of such statistically stored dislocations (ρ_S) is typically 10^{12} m^{-2} for a lightly deformed metal and 10^{16} m^{-2} for a heavily cold-rolled metal. If the metal is constrained to deform inhomogeneously by the presence of nondeformable second-phase particles, then the plastic incompatibility creates additional geometrically necessary dislocations (ρ_G) (Ashby, 1970). The density of these dislocations depends on the strain and the particle size, shape, and spacing. For example, for a volume fraction F_V of spherical particles of diameter d, Ashby shows that ρ_G is given by

$$\rho_G = 8F_V\gamma/bd, \tag{1}$$

where b is the Burgers vector and γ the shear strain. For large volume fractions or small particles, the predicted geometric dislocation density may be much larger than ρ_S. However, this may in reality be significantly lowered by dynamic recovery, even at ambient temperatures. Grain boundaries also lead to inhomogeneous deformation and hence to geometrically necessary dislocations (Ashby, 1970).

The line energy of a dislocation per unit length is given approximately by Gb^2 where G is the shear modulus. Therefore the stored energy of the deformed metal (E) is given approximately by

$$E = Gb^2(\rho_G + \rho_S). \tag{2}$$

A highly deformed metal typically has a stored energy of about 10^6 J m^{-3}. Note that this is very much smaller than the energies associated with phase transformations, e.g., only about 1% of the latent heat of fusion of the metal.

2. THE FUNDAMENTAL MECHANISMS OF ANNEALING

2.1 Recovery

During recovery, many physical and mechanical properties change, and recovery may be measured by a variety of experimental techniques. Figure 2, from the work of Schmidt and Haessner (1990), shows the annealing of high-purity aluminum that was deformed at 77 K and maintained at that temperature until the calorimetry was performed. The peak at 200 K is due to the recovery of point defects, and the peak at 260 K is due to recrystallization. A separate peak due to dislocation recovery is not detected. The high-purity aluminum recrystallizes at such a low temperature that either no dislocation recovery occurs prior to recrystallization or else the peak is too close to the recrystallization peak to be detected. The data from calorimetry are compared in Fig. 2 with those from other techniques.

One of the most important parameters in determining the amount and rate of recovery is the stacking fault energy, which, by affecting the extent to which dislocations dissociate, determines the rate of dislocation climb, which is usually the rate-controlling process during recovery. In metals of low stacking fault energy such as copper, α-brass, and austenitic stainless steel, climb is difficult, and little or no recovery of the dislocation structure occurs prior to recrystallization. However, in metals of high stacking fault energy such as aluminum and α-iron, climb is rapid, and significant recovery may occur.

2.1.1 The Mechanisms of Dislocation Recovery During recovery the stored energy of the material is lowered by dislocation movement. There are two primary processes, these being the annihilation of dislocations and the rearrangement of dislocations into low-energy configurations. Both processes are achieved by glide, climb, and cross-slip of dislocations.

A schematic drawing of a crystal containing an array of edge dislocations is shown in Fig. 3. The forces between the dislocations will depend on the Burgers vectors and relative positions of the dislocations. For example, dislocations of opposite sign on the same glide plane, e.g., A and B, may annihilate by gliding toward each other. Such processes can occur even at low temperatures, lowering the dislocation density during deformation and leading to dynamic recovery. Dislocations of opposite Burgers vector on different glide planes, e.g., C and D, can annihilate by a combination of glide and climb. As climb requires thermal activation, this can only occur at high homologous temperatures (homologous temperature is actual temperature normalized by the melting point).

If unequal numbers of dislocations of the two signs are produced during deformation then the excess cannot be removed by annihilation [Fig. 3(b)]. On annealing, these excess dislocations will arrange into lower-energy configurations in the form of regular arrays or low-angle grain boundaries. The simplest case is that shown in Fig. 3(c) in which dislocations of only one Burgers vector are involved. This process is often known

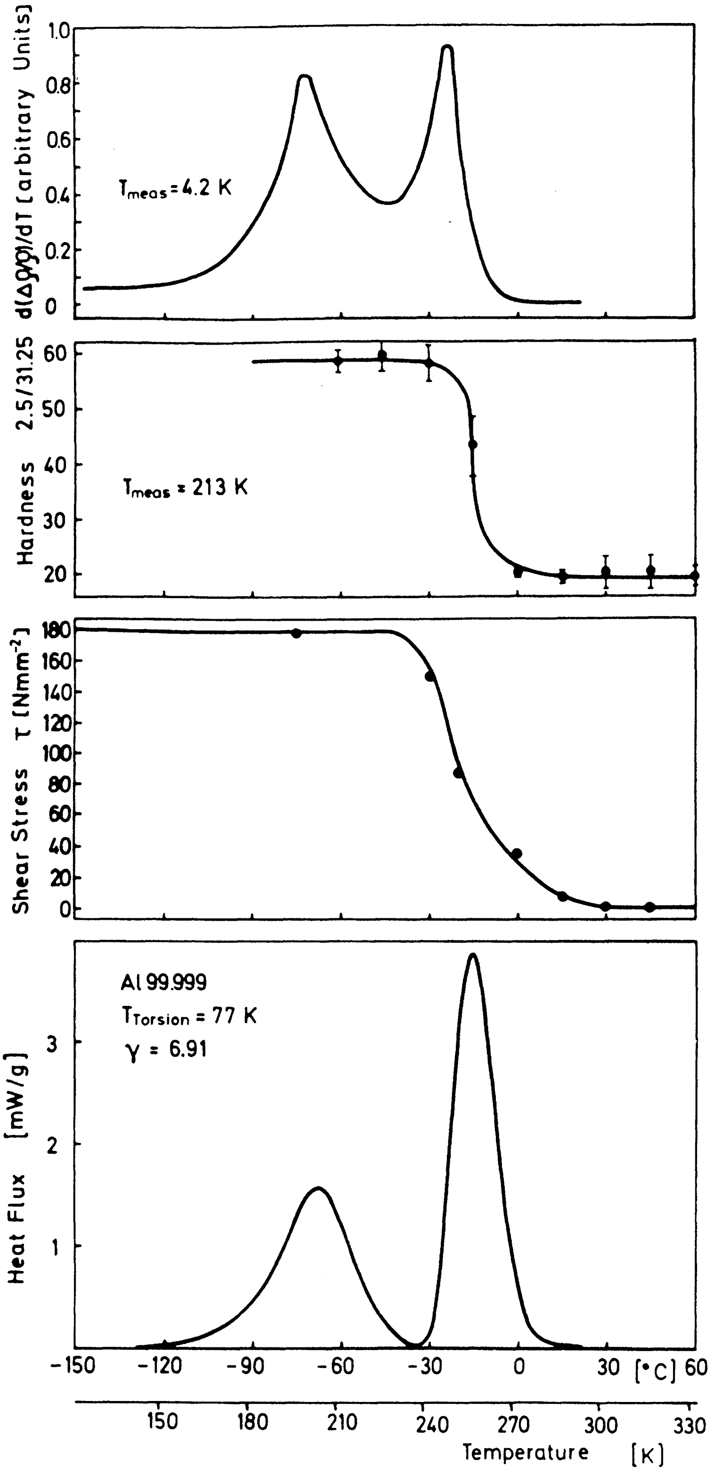

FIG. 2. Comparison of the heat flux, stress, hardness, and differentiated isochronous resistance $d(\Delta\rho/\rho)/dt$ as a function of the annealing temperature of Al 99.999, deformed to a true strain of 6.91 at 77 K (from Schmidt and Haessner, 1990).

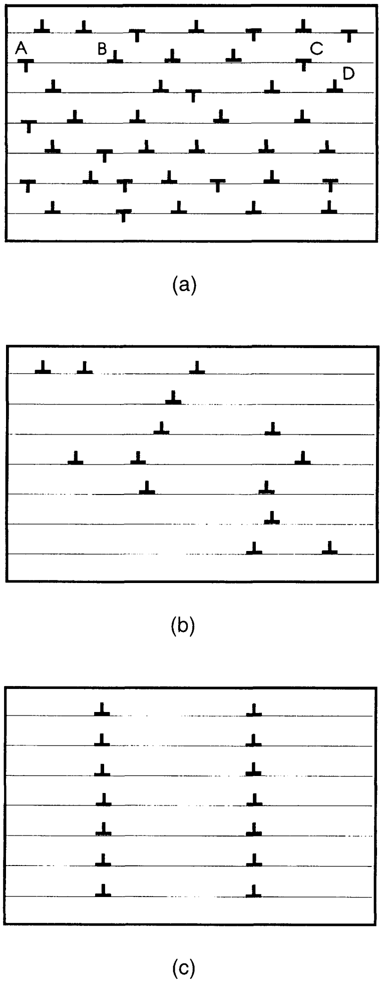

FIG. 3. (a) Crystal containing dislocations of two types. (b) The crystal after annihilation of dislocations has occurred. (c) Rearrangement of the remaining dislocations into low-angle tilt boundaries.

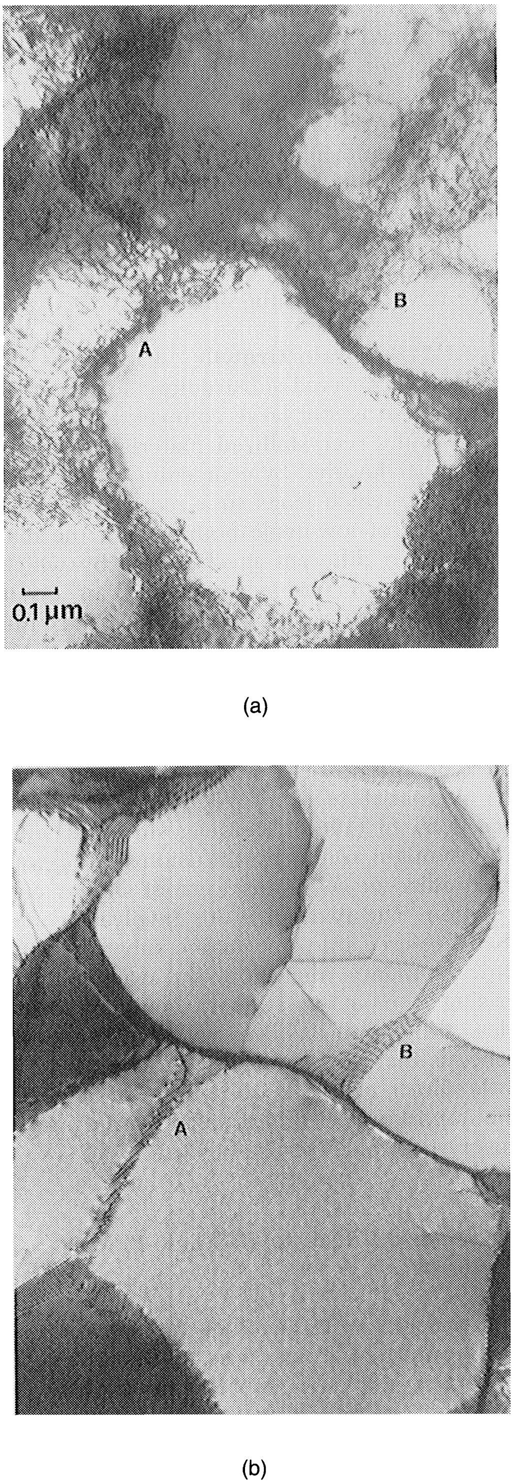

FIG. 4. Transmission high-voltage electron micrographs of aluminum deformed 10% and annealed *in situ*. (a) Deformed structure, (b) same area after 2 min anneal at 500 K.

as polygonization, and the excess dislocations form tilt boundaries. In the case of a polycrystalline material subjected to large strains the dislocation structures produced on deformation and on subsequent annealing are more complex than the simple case shown in Fig. 3 because dislocations of many Burgers vectors are involved.

In an alloy of medium or high stacking fault energy, the dislocations are typically arranged after deformation in the form of a three-dimensional cell structure, the cell walls being complex dislocation tangles. The transmission electron micrograph of Fig. 4(a) shows equiaxed cells of diameter about 1 μm in deformed aluminum. By annealing this specimen *in situ* in a high-voltage electron microscope, the recovery processes can be followed directly. Figure 4(b) shows the same area after annealing. The tangled cell walls, e.g., at A and B, have become more regular dislocation networks or low-angle grain boundaries and the number of dislocations in the cell interiors has diminished. The cells have now become subgrains.

2.1.2 Subgrain Growth The stored energy of a recovered substructure such as that of Fig. 4(b) is still large compared with that of the fully recrystallized material, and can be further lowered by coarsening of the substructure, which leads to a reduction in the total area of low-angle boundary in the material. Two different mechanisms by which coarsening of the substructure can occur have been proposed: low-angle boundary migration and subgrain rotation. Subgrain boundary migration has been frequently observed experimentally and the mobility of a low-angle boundary is known to depend on the type and misorientation of the boundary. An alternative recovery mechanism was proposed by Li (1966). He suggested that subgrains might rotate by diffusional processes until adjacent subgrains were of similar orientation. The two subgrains involved would then coalesce into one larger subgrain with little boundary migration, the driving force arising from a reduction in boundary energies. Although it has been shown that coalescence is thermodynamically feasible, there is little direct evidence of its occurrence and the importance of this mechanism during recovery has yet to be established.

2.2 Recrystallization of Single-Phase Alloys

Recrystallization involves the formation of new strain-free grains in certain parts of the specimen and the subsequent growth of these to consume the deformed or recovered microstructure [Figs. 1(c) and 1(d)]. It is convenient to divide recrystallization into two regimes, the nucleation of recrystallization, which is the formation of the new grains, and grain growth during recrystallization, which follows it.

2.2.1 Kinetics of Primary Recrystallization The progress of recrystallization with time is commonly represented by a plot of the fraction of material recrystallized (X) as a function of time. This plot usually has a characteristic sigmoidal form which is shown schematically in Fig. 5. This typically shows an incubation time before recrystallization is detected, followed by an increasing rate of recrystallization, a linear region, and finally a decreasing rate of recrystallization. Such a curve is typical of many transformation reactions (see, e.g., Shewmon, 1969), and can be described phenomenologically in terms of the constituent nucleation and growth processes. Nuclei are formed at a rate $\dot{N} = dN/dt$ and grains grow into the deformed material at a rate $\dot{G}$. The volume of spherical grains varies as the cube of their diameter, and the recrystallized fraction rises rapidly with time. However, the new grains will eventually impinge on each other and the rate of recrystallization will decrease, tending to zero as X approaches 1. Such a model leads to a kinetic equation of the form

$$X = 1 - \exp(-Bt^k), \qquad (3)$$

which is generally known as the Avrami (1939) or Johnson–Mehl (Johnson and Mehl, 1939) equation.

The constants B and k depend on the details of the model—for example on how the nucleation and growth rates vary during recrystallization—and on the shapes of the growing grains. Although this equation gives

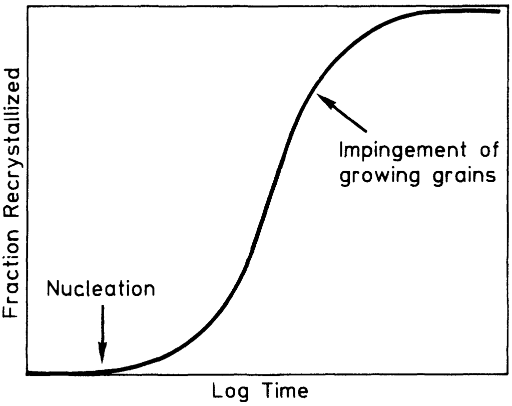

FIG. 5. The effect of time on the fraction recrystallized.

a broad description of recrystallization, careful experiments have shown that it is rarely accurately obeyed in practice. The reason for this is that the recrystallization process is far too complex to be described by an equation with two variables, and, in particular, the inhomogeneity of the distribution of stored energy and of the nucleation events has been shown to cause significant deviations from Avrami kinetics. This has led to the application of computer modeling to the problem (see Sec. 3.2.3).

Because recrystallization is a thermally activated process, the temperature of annealing has a profound influence on the rate of recrystallization as shown in Fig. 6. For isothermal annealing, the rate of recrystallization is often expressed in terms of $t_{1/2}$, the time at which the material is 50% recrystallized. For material which is isochronally annealed, it is often convenient to measure the ease of recrystallization in terms of a recrystallization temperature, the temperature at which recrystallization is complete (or 50% complete) in a fixed time, e.g., 1 h.

The amount and, to some extent, the type of deformation affect the rate of recrystallization, because the deformation alters both the amount of stored energy and the number of effective nuclei. There is a minimum amount of strain, generally a few percent, below which recrystallization will not occur. Above this strain the rate of recrystallization increases, leveling out to a maximum value at true strains of around 2–4.

The initial grain size may affect the stored energy of deformation as well as the number of nucleation sites, and it is found that a fine-grained material will recrystallize more rapidly than a coarse-grained material, as shown in Fig. 6.

The magnitude of the grain size after recrystallization can be rationalized in terms of the effects of various parameters on the nucleation and growth processes. Any parameter favoring a larger number of nuclei or a rapid nucleation rate, such as a high strain or a small initial grain size, will lead to a small final grain size.

2.2.2 The Nucleation of Recrystallization A considerable research effort has been aimed at understanding the processes by which recrystallization originates. The importance of recrystallization nucleation is that it is a critical factor in determining the size and orientation (texture) of the resulting grains. In order to control recrystallization effectively, it is necessary to understand the mechanisms and the parameters that control it. Although the term nucleation of recrystallization is universally used, it is known that recrystallization does not occur by a true nucleation event such as occurs in phase transformations. Recrystallization starts by the growth of a small crystallite that was present in the deformed microstructure.

2.2.2.1 Strain-Induced Grain-Boundary Migration. This mechanism, an example of which is seen in Fig. 7, was first reported by Beck and Sperry (1950). It involves the bulging of part of a pre-existing grain boundary leaving a dislocation-free region behind the migrating boundary. The driving force

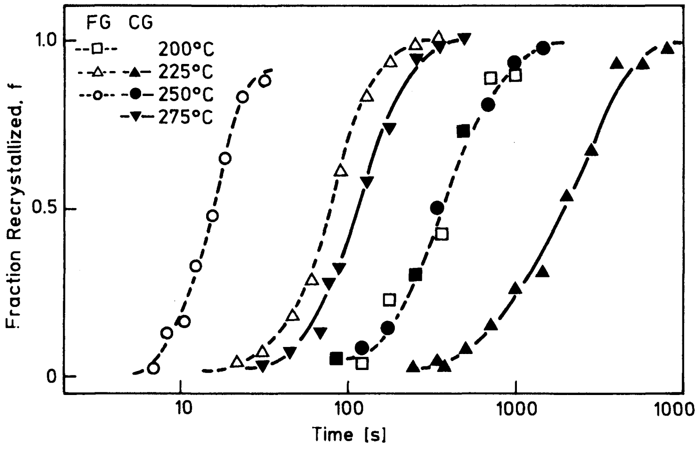

FIG. 6. The effect of temperature and initial grain size on the recrystallization kinetics of copper. The dotted curves show the behavior of material with a 15-μm grain (FG) size and the full curves show the behavior of material with a 50-μm grain (CG) size (Ryde *et al.*, 1990).

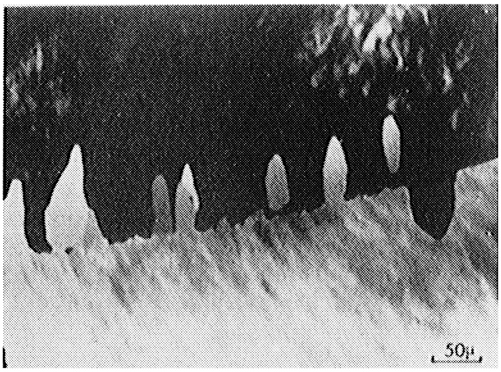

FIG. 7. Strain-induced grain-boundary migration in aluminium deformed 40% by compression (Bellier and Doherty, 1977).

for such an event arises from a difference in dislocation density on opposite sides of the grain boundary, and the process has been analyzed by Bailey and Hirsch (1964). This mechanism is particularly important after low strains. The characteristic feature is that the new grains have similar orientations to the old grains from which they have grown.

2.2.2.2 Preformed-Nucleus Model. In the process of strain-induced grain-boundary migration discussed above, the high-angle grain boundary, which is a prerequisite for recrystallization, was already in existence. However, in many cases, recrystallization originates in other regions of the material, and we need to consider how a nucleus is formed in such circumstances. It is now established beyond reasonable doubt that recrystallization originates from dislocation cells or subgrains that are present after deformation. Although there are still uncertainties about how these pre-existing crystallites become nuclei, several points are now clear:

1. The orientation of the nucleus is present in the deformed structure. There is no evidence that new orientations are formed during or after nucleation, except by twinning.
2. Nucleation occurs by the growth of subgrains by the mechanisms discussed in Sec. 2.1. Although all direct TEM observations have shown the mechanism to be one of subboundary migration, subgrain coalescence may also occur. A characteristic of nucleation is that the recovery in that area is significantly faster than in the majority of the material.
3. In order for a high-angle grain boundary to be produced by this rapid recovery, there must be an orientation gradient present.

2.2.2.3 Nucleation Sites. The site of the nucleus is very important in determining its viability, and it is clear that the general mechanism discussed above can occur at a variety of sites. The grains of a polycrystal do not deform homogeneously, but tend to split into regions of different orientations. This is a consequence of the constraining effects of neighboring grains and of the different slip systems which operate in different grains. Following the terminology used by Hatherly (1982) we note that the regions of different orientation are known as deformation bands. The boundary of a deformation band is a small region of continuous orientation change which is called a transition band. In certain cases the orientation is identical on either side of a deformation band, and in that case it is called a kink band. The transition bands at the edges of the deformation bands are ideal sites for nucleation of recrystallization, having a large dislocation density and a large orientation gradient. During rolling, another form of deformation heterogeneity called a shear band is formed. This is a planar region oriented at around 30° to the rolling plane, in which extensive shear has occurred locally. These bands are another favored site for nucleation, particularly in metals of medium and low stacking fault energy (Duggan *et al.*, 1978). These nucleation sites are shown in the sketch of Fig. 8.

2.2.3 Computer Modeling of Primary Recrystallization There have recently been several attempts to simulate primary recrystallization by computer modeling (see, e.g., Saetre *et al.*, 1986; Doherty *et al.*, 1986; Furu *et al.*, 1990). The main advantage of such models is that they can allow for realistic spatial distribution of nuclei and for complex variations of nucleation and growth rates. Such models are capable of predicting grain size distributions as well as recrystallization kinetics. However, the models are not yet sophisticated enough to give any detailed insight into the mechanisms of recrystallization.

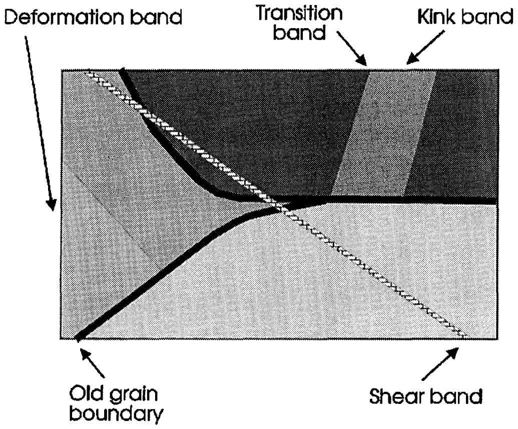

FIG. 8. Sketch of deformation inhomogeneities that may act as sites for recrystallization.

2.3 Migration of High-Angle Grain Boundaries

In the nucleation stage of recrystallization discussed above, a high-angle boundary, capable of migrating into the deformed material, is produced. In order to complete the recrystallization process, the deformed material is consumed by these migrating boundaries.

2.3.1 Experimental Observations It is generally accepted that the velocity (V) of a migrating boundary is related to the boundary mobility (M) and the driving force (P) by

$$V = MP. \tag{4}$$

If the driving force is known, then the boundary mobility can be determined. Experiments in which the velocity of a boundary during recrystallization is measured are very difficult to carry out because the driving force, arising from the stored energy, typically 10 MPa, is not easy to measure accurately, varies throughout the microstructure, and does not remain constant with time, decreasing as recovery proceeds. For this reason, many measurements of boundary mobility have been carried out on materials with a better characterized driving force, e.g., an as-cast substructure. However, it is not clear to what extent measurements of boundary mobility in materials with such low driving forces are directly applicable to migrating boundaries in recrystallizing material.

2.3.2 The Effect of Variables on Boundary Mobility As was the case for low-angle boundaries, the mobility of high-angle boundaries is very temperature dependent. The activation energy for pure metals is typically about half that for self diffusion (see Haessner and Hofmann, 1978) but the activation energy is higher for impure materials.

As boundary migration involves diffusion processes in and across the boundary, it is to be expected that the structure of the boundary should affect its mobility. Boundaries may be classified in part by reference to the number of coincidence sites, which are atom sites common to the grains on either side of the boundary. Special boundaries with a large number of coincidence sites are known to have properties different from those with few coincident sites. Aust and Rutter (1959) measured the migration rate of boundaries in zone-refined lead under a low driving force, as a function of both orientation and impurity level. Their results, shown in Fig. 9, indicate that in very pure lead, the boundary

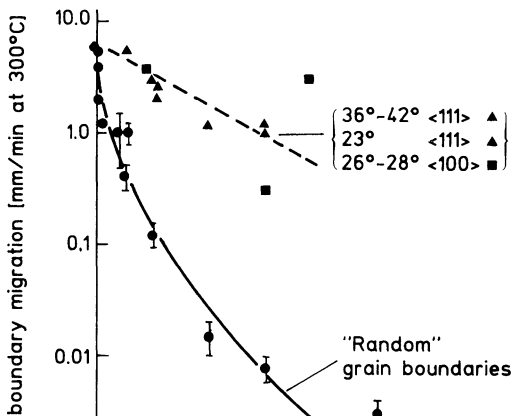

FIG. 9. The rate of grain-boundary migration at 573 K into zone-refined lead crystals doped with small amounts of tin (after Aust and Rutter, 1959).

mobility is almost independent of orientation. However, as the impurity level rises, the mobilities of the randomly oriented boundaries fall much more than those of the coincident-site-lattice boundaries, which are denoted in Fig. 9 by the angle and axis of rotation between the adjacent grains, suggesting that impurities are adsorbed more easily on the more open and disordered structure of the random boundaries.

Liebmann *et al.* (1956) found that during recrystallization of an aluminum single crystal, the boundary migration rate was highest for boundaries misoriented about a ⟨111⟩ axis by around 40° (Fig. 10). Rapid growth of grains of certain orientations will lead to the development of a recrystallization texture, and the results above have often been cited in favor of the model of oriented growth of textures (Sec. 5).

As may be seen from Fig. 9, small amounts of solute have a very large effect on the mobility of boundaries. Lücke and Detert (1957) were the first to formulate a quantitative theory of the effect of solute atoms on boundary mobility. This theory has been extended by others, and the developments are reviewed by Grant *et al.* (1984).

3. RECRYSTALLIZATION OF TWO-PHASE ALLOYS

As most alloys of commercial importance are multiphase, an understanding of the recrystallization behavior of such materials is of practical as well as of scientific interest. Second-phase particles have three important effects on recrystallization:

1. Particles may increase the stored energy and hence the driving force for recrystallization.
2. Large particles may act as nucleation sites for recrystallization.
3. Particles, particularly if closely spaced, may exert a significant pinning effect on grain boundaries.

The first two effects tend to promote recrystallization, whereas the last effect tends to prevent recrystallization. Thus the recrystallization behavior, particularly the kinetics and the resulting grain size, will depend on which of these effects dominates. The final grain size tends to be small when recrystallization is accelerated and coarse when recrystallization is retarded. The recrystallization kinetics strongly depend on both the particle size and the interparticle spacing, as was first clearly demonstrated by Doherty and Martin (1962), and it is clear that by comparison with a single-phase alloy, recrystallization is retarded or even completely inhibited by closely spaced particles, and is accelerated by widely spaced particles, as shown in Fig. 11.

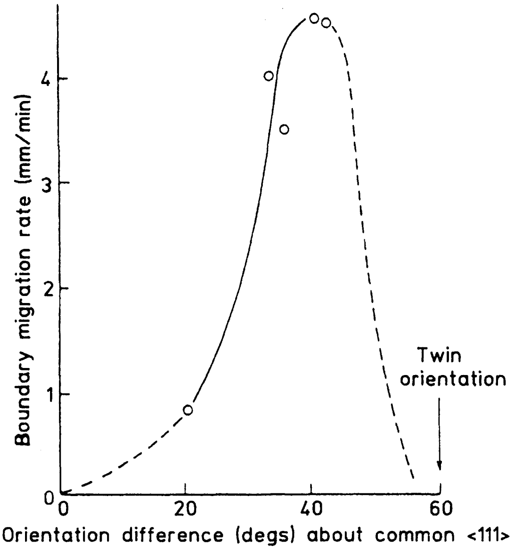

FIG. 10. Growth rates of new grains into a deformed aluminium crystal at 888 K (after Liebmann *et al.*, 1956).

3.1 Particle-Stimulated Nucleation of Recrystallization

If the particles do not deform, then geometrically necessary dislocations are generated at the particles [Eq. (1)]. The form and distribution of these dislocations are primarily functions of strain and particle size, although other factors such as shape, interface strength, and matrix are known to be important (see, e.g., Humphreys, 1985). At particles larger than around 0.2 μm, plastic relaxation results in the formation of deformation zones near the particles. These are regions of high dislocation density that are misoriented from the matrix (Humphreys, 1977, 1979). Nucleation of recrystallization may occur

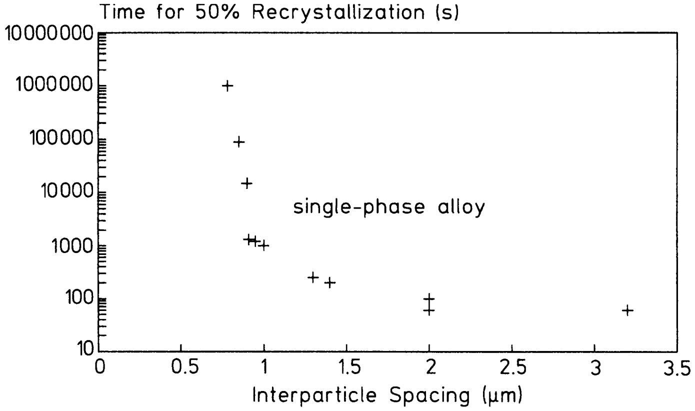

FIG. 11. The effect of interparticle spacing on the time for 50% recrystallization in Al-Cu single crystals (data from Doherty and Martin, 1964).

within these deformation zones and this is commonly referred to as particle-stimulated nucleation (PSN).

The characteristics of PSN are that nucleation occurs by rapid sub-boundary migration, and that the grain may stop growing when the deformation zone has been consumed. A sequence of electron micrographs from an *in situ* annealing sequence, Fig. 12, shows the occurrence of particle-stimulated nucleation in aluminum. There is no evidence that PSN occurs by a mechanism different from nucleation of recrystallization at heterogeneities in single-phase materials, and

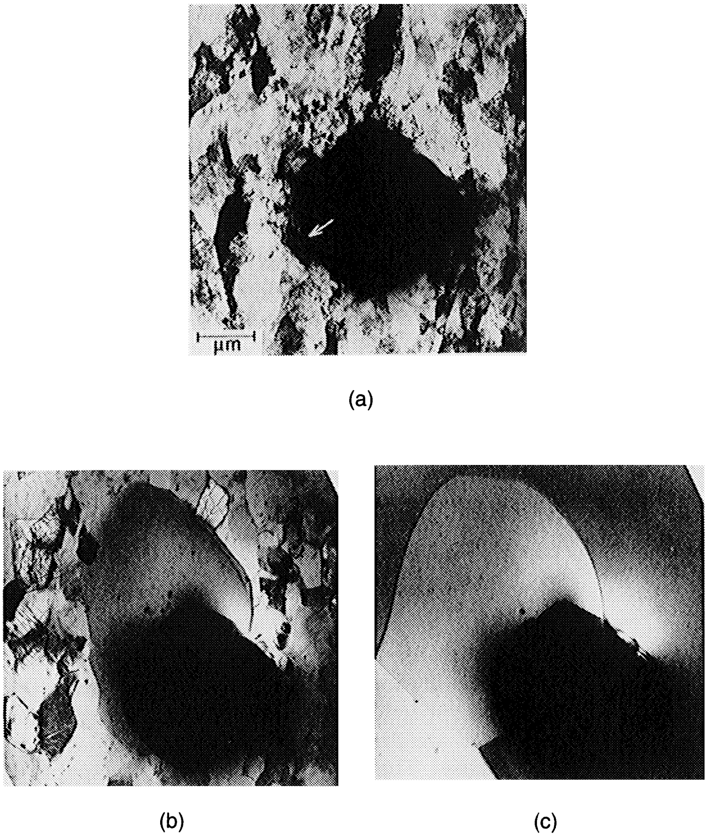

FIG. 12. *In situ* high-voltage electron microscope annealing of Al-Si. Recrystallization originated in the deformation zone (arrowed).

the models of PSN are consistent with the preformed-nucleus models discussed in Sec. 2.2.2.

The main parameters that determine whether or not PSN occurs are the strain and the particle size. This is shown in Fig. 13 for rolled aluminium containing Si particles. Two criteria must be fulfilled for growth of a nucleus beyond the deformation zone (Humphreys, 1977). First, a deformation zone with sufficient misorientation to create a high-angle boundary must be created on deformation, and second, the growing grain must also be able to grow into the surrounding matrix, which has a stored energy per unit volume of E. Humphreys showed that this condition approximated to

$$d_c = 4\sigma/E, \tag{5}$$

where d_c is the critical particle diameter and $\sigma$is the grain-boundary energy.

It is difficult to predict values of E in deformed alloys accurately. However, if reasonable values are taken (Humphreys, 1977) or if experimental subgrain data are used, then this equation gives agreement with the results of Fig. 13, showing that the growth criterion is the critical one.

3.2 Pinning Effects of Particles (Zener Drag)

A dispersion of particles will exert a retarding force on a grain boundary. The effect is known as Zener drag after the original analysis by Zener (Smith, 1948).

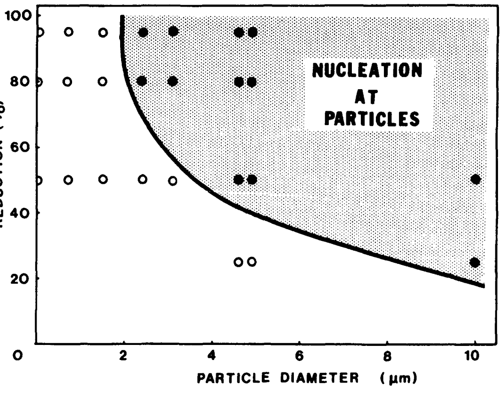

FIG. 13. The effect of rolling reduction and particle size on the occurrence of particle-stimulated nucleation (Humphreys, 1977).

When a grain boundary, of specific energy σ, migrates onto a spherical incoherent particle, when it can be assumed that the interfacial energy of the particle is unchanged by the passage of the grain boundary, then the particle, of radius r, effectively removes an area of the boundary equal to πr^2, and the energy is reduced by an amount $\pi r^2 \sigma$. To move the boundary past the particle a force F must be applied, the maximum value of which is $\pi r \sigma$. If the boundary is planar and randomly intersects particles, then the pinning force exerted on the boundary in an alloy with a volume fraction F_v of particles of radius r is given by

$$p_z = 3F_v\sigma/2r. \tag{6}$$

Coherent particles or nonspherical particles may be more effective in pinning boundaries. The Zener drag force has been examined by several authors, and the reader is referred to the reviews by Nes *et al.* (1985) and Hillert (1988) for further details. Particle pinning is undoubtedly the cause of retarded recrystallization in alloys containing closely spaced small particles, and plays a dominant role in grain growth after recrystallization (Sec. 4).

3.3 Bimodal Alloys and the Prediction of Grain Size

Many commercial alloys contain distributions of both large (>1 μm) particles, which will act as nucleation sites, and small particles, which will pin the migrating boundaries. In such a situation, the driving force is offset by the Zener pinning force and the critical particle size for nucleation [Eq. (5)] now becomes

$$d_c = \frac{4\sigma}{E - 3F_v\sigma/2r}. \tag{7}$$

Thus, as the Zener pinning force increases, the critical particle diameter for PSN increases. As there will be a distribution of particle sizes in a real alloy, this means that fewer particles are able to act as nuclei, and that the recrystallized grain size will increase. The number of particles capable of acting as nuclei (N) is the number of particles of diameter greater than d_c. The grain size will be given approximately by

$$D_N = N^{-1/3}. \tag{8}$$

There is considerable interest in being able to predict the grain size of particle-containing alloys as a function of the particle parameters and the thermomechanical processing route, Nes (1976), Nes and Hutchinson (1989), and Wert and Austin (1988) have developed models, based on the mechanisms discussed above, for predicting grain size in commercial aluminum alloys.

3.4 The Interaction of Precipitation and Recrystallization

The mutual interaction of precipitation and recrystallization has been extensively studied by Hornbogen and colleagues in Al-Cu and Al-Fe alloys, and has been comprehensively reviewed by Hornbogen and Köster (1978). If a supersaturated solid solution is deformed and annealed, then unless recrystallization is complete before precipitation begins, the two processes affect each other. Precipitation on low-angle or high-angle boundaries will hinder recovery and recrystallization, and dislocations will themselves promote the nucleation of precipitates.

4. GRAIN GROWTH AFTER PRIMARY RECRYSTALLIZATION

When primary recrystallization, which is driven by the stored energy of cold work, is complete, the structure is not yet stable, and further growth of the recrystallized grains may occur. The driving force for this is a reduction in the energy stored in the material in the form of grain boundaries. The driving pressure for grain growth is some two orders of magnitude less than that for primary recrystallization, and consequently, grain growth will be slower than during primary recrystallization and will be more affected by solutes and particles that pin grain boundaries. General reviews of grain growth have been undertaken by Higgins (1974), Doherty and Martin (1976), and Atkinson (1988).

Consider first a two-dimensional grain structure as shown in Fig. 14. If the grain boundaries are assumed to have equal energies, then the triple points are in equilibrium when the boundaries make an angle of 120° with each other. This is possible for an array of regular hexagonal grains [Fig. 14(a)] and in this case we have achieved a metastable structure that will not coarsen. However, if we take a more realistic, less regular array of grains as shown in Fig. 14(b), then the boundaries will become curved to achieve the required triple-point angles. Curved boundaries will be unstable and tend to migrate in the direction toward their centers of curvature, so as to shorten their lengths. The result is that grains of fewer than six sides will tend to shrink and eventually disappear, while those of more than six sides will tend to grow, and the average grain size therefore

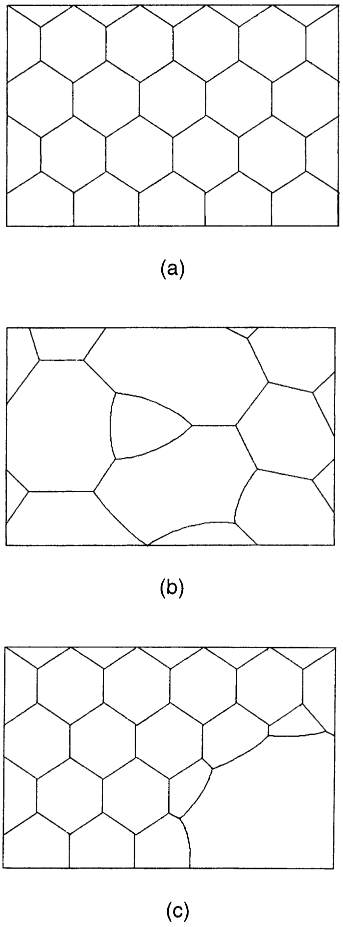

FIG. 14. Growth of a two-dimensional grain structure. (a) An array of regular hexagonal grains is stable. (b) In an irregular grain structure the boundaries are curved. (c) Secondary recrystallization.

increases with time. Extension of these ideas to a three-dimensional grain structure leads to similar results except that a truly stable three-dimensional array, equivalent to Fig. 14(b), does not appear to exist.

The kinetics of grain growth are generally found to obey the relationship

$$D^2 - D_0^2 = Kt^n, \tag{9}$$

where D is grain diameter at time t and D_0 the grain size at $t = 0$. The constant n typically lies between 0.3 and 0.5.

4.1 Factors Affecting Grain Growth

Grain growth is inhibited by a number of factors, and the effect of solutes on boundary migration has already been discussed in Sec. 2.3.

The rate of grain growth diminishes when the grain size becomes greater than the thickness of a sheet specimen (Burke, 1949). This is because the grains are now curved only in one direction rather than two, and thus the driving force is diminished. Thermal etching grooves may also be formed on the sample surface by diffusion, and these will also impede grain growth.

Grain growth may also be affected by the presence of a sharp crystallographic texture (Beck and Sperry, 1949). This arises at least in part from a large number of grains of similar orientation leading to more low-angle and hence low-energy boundaries. Thus the driving force for growth is reduced. The texture may also alter during grain growth (Abbruzzese and Lücke, 1986), thereby affecting the kinetics.

Undoubtedly the most important features affecting grain growth are second-phase particles. During grain growth of a fully recrystallized alloy, the driving force is much smaller than during primary recrystallization. Therefore the pinning effects of particles on grain boundaries discussed in Sec. 3 are relatively more important. The pinning force (p_z) exerted on a boundary by an array of particles is given by Eq. (6). During grain growth, the driving force (P_R) arises solely from the curvature of the boundaries; the magnitude of this driving force is given by $2E/R$. On average, we expect the grains to stop growing at a diameter D_z (the limiting grain size or Zener limit) when $p_Z = p_R$:

$$D_z = 4r/3F_v. \tag{10}$$

The pinning effect of even very low volume fractions of small particles is sufficient to limit the grain size significantly. For example, a volume fraction of 5×10^{-3} of particles of diameter 50 nm gives $D_Z = 67\ \mu$m. If the interparticle spacing is of a similar magnitude to the grain size then the particles will lie predominantly on the grain boundaries and the pinning force will be greater (Hellman and Hillert, 1975; Hutchinson and Duggan 1978).

4.2 Computer Simulation of Grain Growth

In a recent series of papers, Anderson and colleagues (e.g., Anderson *et al.*, 1984; Anderson, 1986) have described Monte Carlo simulation of grain growth in two and three dimensions. The material is divided into a number of discrete points each of which is given a number, corresponding to a grain orientation as shown in Fig. 15. The grain-boundary energy is then specified by the interaction between the grid sites. For example, a certain value of energy might be assigned to a 4–6 boundary and the same or a different one to a 3–9 boundary. Pairs of adjacent points are then swapped and the energy change (ΔE) measured. If the energy is lower, then there is a probability of

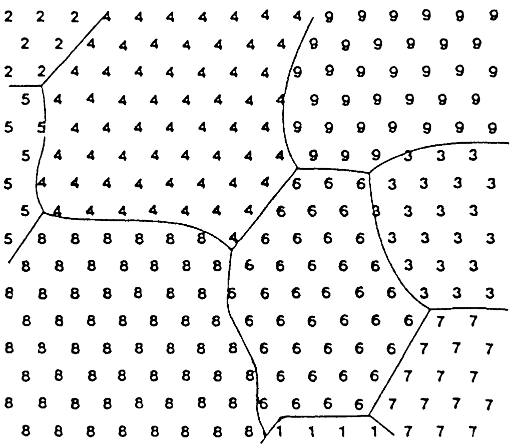

FIG. 15. The type of "microstructure" used for atomistic simulation of grain growth. The integers denote orientations and the lines are grain boundaries (Anderson, 1986).

$\exp(-\Delta E/kT)$ that the change will be accepted. Transitions occur at grain boundaries and the grains grow, showing most of the features of grain growth. If this technique is applied to three dimensions, the size of the array that can be used is limited to around 100 × 100 × 100 points.

4.3 Secondary Recrystallization

In the previous section we discussed the relatively uniform growth of grains after recrystallization. However, in certain circumstances, a few grains, such as the large grain in Fig. 14(c), may grow excessively, consuming the recrystallized grains. This may lead to a grain structure of the order of several millimeters or greater. This process is known as secondary recrystallization. Detailed discussions of this phenomenon are given by Detert (1978) and Cahn (1983). The rapidly growing grains may be larger than the average or they may have an orientation which is more favorable for growth. If the latter is the case then a pronounced texture may result on secondary recrystallization. The best-known examples of this are the iron-silicon alloys used for transformer cores (e.g., Luborsky *et al.*, 1983). In these a pronounced (110)[001] texture (a {110} plane parallel to the rolling plane and an ⟨001⟩ direction parallel to the rolling direction) is produced on secondary recrystallization. The driving force for secondary recrystallization is the energy of the grain boundaries and therefore the process is more likely to occur when pinning of boundaries by second-phase particles restricts normal grain growth (Sec. 4.1). If, during a subsequent anneal, the particles coarsen or dissolve, then secondary recrystallization commonly occurs (Detert, 1978).

5. RECRYSTALLIZATION TEXTURES

When a metal is recrystallized, the crystallographic orientations of the new grains may be quite different from those of the old ones. As textures have a significant effect on the mechanical and physical properties of a material, the understanding of the origin of textures is of great importance. The published literature on the subject of textures is extensive, and there are several contentious issues that continue to arouse much debate. For further general information the reader is referred to the reviews of Hatherly and Dillamore (1975), Grewen and Huber (1978), Nes and Hutchinson (1989), and Hatherly (1990).

The texture after annealing is primarily determined by two factors: the orientations that are produced during nucleation, and the relative growth velocity of grains of different orientations. For many years there has been heated debate as to which of these two factors controls the final texture. It has been claimed that the recrystallization texture has its origin in either the preferred nucleation of grains with the required orientation (oriented nucleation theory) or the preferred growth of such grains from a more randomly oriented array (oriented growth theory).

Recent advances in techniques for determining local textures, which enable detailed correlation of microstructure and local texture, are now providing the types of information necessary for obtaining a better understanding of the formation of annealing textures. It is clear that there are some instances in which nuclei of a limited range of orientations are produced, and this factor then will predominantly determine the final texture. Similarly, there are cases when grains of certain orientations clearly grow more rapidly. In general we must expect the final texture to be determined by both factors in proportions that will vary for different materials, deformations, and annealing conditions.

5.1 Orientation Effects in Nucleation

As discussed in Sec. 2.2, a recrystallized grain originates in a small region of the deformed material and will therefore have the orientation of that region. Although this orientation is present in the deformed material, it may not be present in sufficient quantity to be detected by bulk texture measurements and therefore is unlikely in general to be an important component of the deformation texture. In Sec. 2.2 we discussed several different sites for recrystallization, and each of these may lead to one or more texture components.

5.1.1 Transition Bands A transition band separates different parts of an old grain which has split during deformation and where the individual new parts have rotated toward different, but stable, end orientations (Hu, 1962). A transition band therefore is a region with a sharp orientation gradient bridging the two neighboring texture components. Dillamore and Katoh (1974) predicted that in fcc metals, transition bands containing the cube orientation (001)[100] would develop. This was confirmed experimentally by Ridha and Hutchinson (1982) who showed that thin pancake-shaped regions of this orientation developed into grains of the cube orientation. The success of these nuclei in developing was undoubtedly enhanced by their shape. Similar cube nuclei have also been found in aluminum by Hjelen and Nes (1986), although only in transition bands within (112)[111] grains.

5.1.2 Shear Bands Shear bands are also important sites for recrystallization. However, it is more difficult to predict the resulting textures. For example in steels, it has been shown (Haratani *et al.*, 1984) that the Goss component (110)[001] originates at shear bands. However, in copper, the nuclei that form from shear bands are of widely spread orientations (Ridha and Hutchinson, 1982).

5.1.3 Prior Grain Boundaries Strain-induced grain-boundary migration, which is an important mechanism after lower amounts of deformation, results in texture components that are present in significant amounts in the deformed material. For example, the (001)[110] and (112)[110] orientations in steel are thought to originate in this way (Inokuti and Doherty, 1977).

5.1.4 Second-Phase Particles Nucleation in the vicinity of large second-phase particles gives rise to grains whose orientations reflect those in the deformation zones. In single crystals, where a very restricted range of orientations is produced, sharp textures may be formed on recrystallization (Humphreys, 1977). However, in polycrystals, the large spread of orientations available near the particles results in a severe weakening of the deformation texture with no new major components being formed (Humphreys and Juul Jensen, 1986; Humphreys and Kalu, 1990).

5.2 Orientation Effects in Growth

The effects of orientation on grain growth were discussed in Sec. 2.3. In summary, there is clear evidence, particularly from lightly deformed fcc single crystals, that the growth rate of grains is a maximum when there is a misorientation of ~40° about a ⟨111⟩ axis. For highly deformed polycrystals there is less information, and the situation is complicated because the orientation of the deformed material changes over short distances and so no precise orientation relationship exists between the old and new grains.

Another factor which must be taken into account is that the deformation structure is heterogeneous and different regions of the microstructure contain different components of the deformation texture. Therefore a model in which oriented growth is taken into account must consider the local nuclei and their environment. Microtexture work such as that by Hjelen *et al.* (1990) is now beginning to provide the information which will enable more complete theories of texture development to be formulated.

Twinning will provide new orientations, and may also be important in producing grains of suitable orientation for rapid growth. There is still some debate about the conditions under which annealing twinning is important in determining textures, particularly in metals of higher stacking fault energy. A review by Hatherly (1990) gives a good summary of the current situation.

5.3 Competition between Texture Components

The final texture is the result of the selection of orientations during nucleation and growth. In many materials several components are formed, originating from different microstructural sites. Nucleation and growth rates for the different components may also alter with time, making it therefore very difficult to predict the final texture.

As an example, we can consider the recrystallization of commercial-purity aluminum, which contains some large second-phase particles. The growth of various texture components was followed during the annealing process and the results are shown in Fig. 16. Initially, microstructural observation showed nucleation to be predominantly at particles (random component). However, during the anneal, these grains grew slowly, whereas grains of the rolling texture (formed at the old grain boundaries) grew more rapidly. Cube grains grew very rapidly, and the

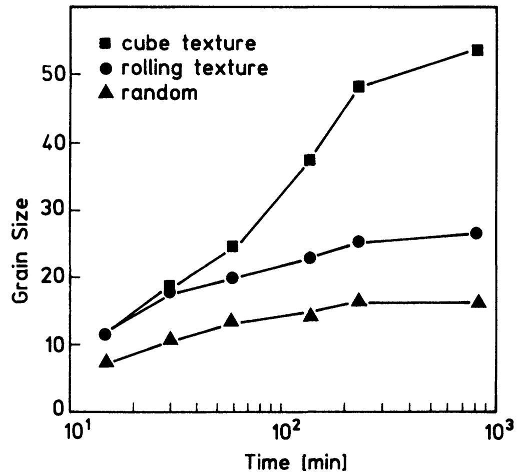

FIG. 16. The average size of grains with different orientations, measured for a series of partly recrystallized specimens of aluminium of commercial purity (Juul Jensen *et al.*, 1985).

final texture contained a high percentage of this component. This example illustrates the way in which both selection of orientations, different growth rates, and competition between the various components play a significant role in determining the final texture.

GLOSSARY

Cell: A small region, typically about one micrometer in diameter, surrounded by a tangled wall of dislocations.

Dislocation: A line defect in a crystal. Dislocations produced during plastic deformation provide the driving force for the various annealing phenomena.

Grain Growth: The growth of grains following recrystallization. The driving force for this process is a reduction in the energy of the grain boundaries.

Grains: Crystals comprising a polycrystalline aggregate.

High-Angle Grain Boundary: A grain boundary between grain misoriented by more than about 15°. In general the energy and mobility of such a boundary are independent of misorientation, although special or coincidence boundaries with low energies are found for specific misorientations.

Low-Angle Grain Boundary: A grain boundary between crystals misoriented by a small angle, typically taken to be 15°. A low-angle boundary is made up of an array of dislocations. The energy and mobility of these boundaries are dependent on the misorientation.

Particle-Stimulated Nucleation: Nucleation of recrystallization at large second-phase particles.

Recovery: The lowering of the stored energy by the annihilation and rearrangement of dislocations into low-energy configurations.

Recrystallization: The formation and growth of essentially dislocation-free grains within a deformed or recovered material.

Secondary Recrystallization: If grain growth is restricted, then a few grains may grow discontinuously thereby producing a very large grain size.

Stored Energy: The energy stored in a deformed metal, primarily in the form of dislocations.

Subgrain: Similar to a cell, but surrounded by low-angle boundaries. Recovery may convert a cell structure to a subgrain structure.

Texture: A term describing the crystallographic orientation of a polycrystalline material. Textures range from strong, if the grains have similar orientations, to weak if the grains are almost randomly oriented. A texture is produced both during deformation and during recrystallization.

Zener Drag: The pinning of boundaries by a distribution of small particles.

Works Cited

Abbruzzese, G., Lücke, K. (1976), *Acta Metall.* **34,** 905.

Anderson, M. (1986), in: N. Hansen (Ed.), *Proceedings of the 7th International Riso Symposium,* Riso, Denmark, p. 15.

Anderson, M. P., Grest, G. S., Srolowitz, D. J., Sahni, P. S. (1984), *Acta Metall.* **32,** 783.

Ashby, M. F. (1970), *Philos. Mag.* **21,** 399.

Atkinson, H. V. (1988), *Acta Metall.* **36,** 469.

Aust, K. T., Rutter, J. W. (1959), *Trans AIME* **215,** 119 and 608.

Avrami, M. J. (1939), *Chem. Phys.* **7,** 103.

Bailey, J. E., Hirsch, P. B. (1964), *Proc. R. Soc. London, Ser. A* **267,** 11.

Beck, P. A., Sperry, P. R. (1949), *Trans. AIME* **180,** 240.

Beck, P. A., Sperry, P. R. (1950), *J. Appl. Phys.* **21,** 150.

Bellier, S. P., Doherty, R. D. (1977), *Acta Metall.* **25,** 521.

Burke, J. E. (1949), *Trans AIME* **180,** 73.

Cahn, R. W. (1983), in: R. Cahn, N. Haasen (Eds.), *Physical Metallurgy,* 3rd ed., Amsterdam: North-Holland, 1595.

Detert, K. (1978), in: F. Haessner (Ed.), *Recrystallization of Metallic Materials,* Stuttgart: Dr. Riederer Verlag GmbH, 97.

Dillamore, I. L., Katoh, H. (1974), *Metal Sci.* **8,** 73.

Doherty, R. D., Martin, J. W. (1962), *J. Inst. Met.* **91,** 332.

Doherty, R. D., Martin, J. W. (1964), *Trans. ASM* **57,** 874.

Doherty, R. D., Martin, J. W. (1976), Stability of Microstructure in Metallic Systems, Cambridge, U.K.: Cambridge University Press.

Doherty, R. D., Rollett, A. R., Srolovitz, D. J. (1986), in: N. Hansen (Ed.), *Proceedings of the 7th International Riso Symposium,* Riso, Denmark, p. 53.

Duggan, B. J., Hutchinson, W. B., Hatherly, M. (1978), *Scripta Metall.* **12,** 1293.

Furu, T., Marthinsen, Nes, E. (1990), *Mater. Sci. Tech.* **6,** 1093.

Grant, E., Porter, A., Ralph, B. (1984), *J. Mater. Sci.* **19,** 3554.

Grewen, J., Huber, J. (1978), in: F. Haessner (Ed.), *Recrystallization of Metallic Materials,* Stuttgart: Dr. Riederer Verlag GmbH, p. 111.

Haessner, F., Hofmann, S. (1978), in: F. Haessner (Ed.), *Recrystallization of Metallic Materials,* Stuttgart: Dr. Riederer Verlag GmbH, p. 63.

Haratani, T., Hutchinson, W. B., Dillamore, I. L., Bate, P. (1974), *Metal Sci.* **18,** 57.

Hatherly, M. (1982) in: R. Gifkins (Ed.), *Proceedings of the 6th International Conference on Strength of Metals and Alloys,* Pergamon: Melbourne.

Hatherly, M. (1990), in: T. Chandra (Ed.), *Recrystallization '90,* Warrendale, OH: The Metals Society, p. 59.

Hatherly, M., Dillamore, I. L. (1975), *J. Australian Inst. Met.* **20,** 71.

Hellman, P., Hillert, M. (1975), *Scand. J. Metall.* **4,** 211.

Higgins, G. T. (1974), *Met Sci. J.* **8,** 143.

Hillert, M. (1988), *Acta Metall.* **36,** 3177.

Hjelen, J., Nes, E. (1986), in: N. Hansen (Ed.), *Proceedings of the 7th International Riso Symposium,* Riso, Denmark, p. 367.

Hjelen, J., Orsund, R., Nes, E. (1990), *Acta Metall.* **39,** 1377.

Hornbogen, E., Köster, U. (1978), in: F. Haessner (Ed.), *Recrystallization of Metallic Materials,* Stuttgart: Dr. Riederer Verlag GmbH, p. 159.

Hu, H. (1962), *Trans. Met. Soc. AIME* **224,** 75.

Humphreys, F. J. (1977), *Acta Metall.* **25,** 1323.

Humphreys, F. J. (1979), *Acta Metall.* **27,** 1801.

Humphreys, F. J. (1985), in: M. Loretto (Ed.), *Dislocations and Properties of Real Materials,* London: Institute of Metals, 175.

Humphreys, F. J., Juul Jensen, D. (1986), in: N. Hansen (Ed.), *Proceedings of the 7th International Riso Symposium,* Riso, Denmark, p. 93.

Humphreys, F. J., Kalu, P. N. (1990), *Acta Metall.* **38,** 917.

Hutchinson, W. B., Duggan, B. J. (1978), *Metal Sci.* **12,** 372.

Inokuti, Y., Doherty, R. D. (1977), *Texture* **2,** 143.

Johnson, W. A., Mehl, R. F. (1939), *Trans. AIME* **135,** 416.

Juul Jensen, D., Hansen, N., Humphreys, F. J. (1985), *Acta Metall.* **33,** 2155.

Li, J. C. M. (1966), in: H. Margolin (Ed.), *Recrystallization Grain Growth & Textures,* Warrendale, OH: The Metals Society, p. 45.

Liebmann, B., Lücke, K, Masing, G. (1956), *Z. Metallk.* **47,** 57.

Luborsky, F. E., Livingston, J. D., Chin, G. Y. (1983), in: R. Cahn, P. Haasen (Eds.), *Physical Metallurgy,* Amsterdam: North-Holland, 1698.

Lücke, K., Detert, K. (1957), *Acta Metall.* **5,** 628.

Nes, E. (1976), *Acta Metall.* **24,** 391.

Nes, E., Hutchinson, W. B. (1989), in: Bilde-Sorensen (Ed.), *Proceedings of the 10th International Riso Symposium,* Riso, Denmark, p. 233.

Nes, E., Ryum, N., Hunderi, O. (1985), *Acta Metall.* **33,** 11.

Ridha, A. A., Hutchinson, W. B. (1982), *Acta Metall.* **30,** 1929.

Ryde, L., Hutchinson, W. B., Jonsson, S. (1990), in: T. Chandra (Ed.), *Recrystallization '90,* Warrendale, OH: The Metals Society, p. 313.

Schmidt, J., Haessner, F. (1990), *Cond. Matter* **81,** 215.

Saetre, T. O., Hunderi, O., Nes, E. (1986), *Acta Metall.* **34,** 981.

Shewmon, P. G. (1969), *Transformations in Metals,* New York: McGraw-Hill.

Smith, C. S. (1948), *Trans. AIME* **175,** 47.

Wert, J. A., Austin, L. K. (1988), *Metall. Trans. A* **19,** 617.

Further Reading

Byrne, J. G. (1965), *Recovery, Recrystallization and Grain Growth,* New York: McMillan.

Cahn, R. W. (1983), in: R. Cahn, P. Haasen (Eds.), *Physical Metallurgy,* 3rd ed., Amsterdam: Elsevier Science Publishers, 1595.

Cotterill, P., Mould, P. R. (1976), *Recrystalliza-*

tion and Grain Growth in Metals, Surrey, U.K.: Surrey Univ. Press.

Doherty, R. D., Martin, J. W. (1976), *Stability of Microstructure in Metallic Systems,* Cambridge, U.K.: Cambridge University Press.

Haessner, F. (Ed) (1978), *Recrystallization of Metallic Materials,* Stuttgart: Dr. Riederer Verlag GmbH, p. 111.

Humphreys, F. J., Hatherly, M. (1995), *Recrystallization and Related Annealing Phenomena,* Oxford, U.K.: Pergamon.

RECYCLING

See WASTE MANAGEMENT AND RECYCLING

REFLECTOMETERS, OPTICAL

C. S. McCamy, *Wappingers Falls, New York, U.S.A.*

Introduction 247
1. **Physical and Technical Principles** 247
1.1 Basic Components of Reflectometers 247
1.2 Conditions of Measurement 247
1.3 Physical Quantities Measured 248
2. **Design, Function, and Operation** 248
2.1 Geometric Design 248
2.2 Spectral Design 249
2.3 Light Sources 250
2.4 Sensors 250
2.5 Standards 250
2.6 Quality of Measurements 250
2.7 Operation 251
3. **Manufacture** 251
4. **Applications** 251
4.1 Chemical Analysis 251
4.2 Densitometers 251
4.3 Colorimetry 252
4.4 Gloss, Distinctness of Image, and Haze 253
5. **Economic Aspects** 253
Glossary 253
Works Cited 253
Further Reading 253

INTRODUCTION

Reflectometers measure light (or other electromagnetic radiant flux) reflected from a surface, as a percentage of the flux incident on the surface or as a percentage of the flux reflected by a highly reflecting reference standard. They are used to characterize materials and to control the manufacture of colored products.

1. PHYSICAL AND TECHNICAL PRINCIPLES

1.1 Basic Components of Reflectometers

All reflectometers have the same basic components, though they differ widely in details of design, depending on the nature of the specimens to be measured and the intended use of the data. An illuminator (or irradiator in the case of radiant flux outside the visible spectrum) provides incident radiant flux, simply called the influx. The illuminator consists of a source, which radiates flux, and a director, which directs the flux to a sampling aperture, where the specimen is placed for measurement. The source may be a lamp filament, for example. The director might simply be the lamp envelope, but more often it includes various optical components called the influx optics.

Light reflected by the specimen is collected and measured by a receiver. The receiver consists of a collector, which collects flux reflected by the specimen, and a sensor, which senses the collected flux and produces a corresponding signal. The collector usually includes optical components, comprising the efflux optics. The sensor requires supporting electronics and data display. Usually, a separate or built-in computer records and processes data and interprets them in useful terms.

1.2 Conditions of Measurement

Values obtained by reflectometry depend on the angles at which the surface is illuminated and the angles at which light is collected for measurement. These angular distributions are called the geometry or geometric conditions of measurement. The values also depend on the spectral distribution of the incident light and the spectral sensitivity of the receiver, known as the spectral conditions. The geometric and spectral conditions characterize the measurement system and characterize applications, includ-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ing visual observation. To be useful, measurements must be made under conditions appropriate to the way the data will be applied. Thus the principle of simulation: The geometric and spectral conditions of measurement must simulate the geometric and spectral conditions of the application.

1.3 Physical Quantities Measured

Reflectance is the ratio of the amount of reflected flux to the amount of incident flux, under prescribed geometric and spectral conditions. Reflectance factor is the ratio of the amount of reflected flux to the amount of flux reflected by a reflection standard, under the same geometric and spectral conditions. These quantities are usually quite different. For example, white paper may reflect over 80% of incident light and absorb the rest. The eye collects a very small fraction of the light reflected in all directions, so the reflectance is very small. However, the reflectance factor (ratio of reflected light to that reflected by a white standard) might exceed 80%. If the flux reflected in all directions is collected, the adjective "total" is applied. The application dictates which quantity is appropriate. A quantity measured within a narrow wavelength band is a spectral quantity, such as spectral reflectance. Concepts, terminology, symbols, notation, and methods of describing geometry delineated by McCamy (1966) have been standardized by the American Society for Testing Materials (ASTM), the International Organization for Standardization (ISO), and the Commission Internationale de l'Éclairage (CIE).

Reflectivity is the reflectance or the reflectance factor of a generic material, such as polished gold. The reflectance factor of paper, plastics, and textiles may depend on reflections from the surface behind them. In this case, reflectivity is the reflectance factor of a specimen or stack of specimens so thick that greater thickness does not change the value.

Radiance is the radiant flux emanating from a surface, in a given direction, per unit of projected area, per unit of solid angle. Radiance factor is the ratio of the radiance from a point on a specimen to that from an ideal diffuser, under the same conditions. This quantity is the same as reflectance factor, if only reflection is involved, but if the material is fluorescent, phosphorescent, or otherwise self-luminous, the flux emanated may be partly reflected and partly emitted. Both are included in radiance factor.

2. DESIGN, FUNCTION, AND OPERATION

2.1 Geometric Design

Geometry is designed with regard for two different kinds of reflection. Incident light may be partly reflected by the first surface encountered, but some may penetrate the surface and be reflected by underlying matter. Reflection at the first surface of a glossy material is called specular reflection. The reflected light is the color of the incident light. If the incident light is white, the reflected light is white. Such highlight reflections are seen on glossy ceramics, plastics, and paints. Light that penetrates the surface is scattered and diffusely reflected by underlying pigments or other colorants and has the color imparted by them. Light reflected by a metal surface has a color characteristic of the metal.

When we examine the color of a nonmetal surface, we avoid any specular reflection of the light source. If a lamp is 45° off the normal to a surface and we observe along the normal, the specular reflection is avoided. This is one standard geometry for measurements. The colors of some materials, such as wood and some textiles, observed this way, vary if they are rotated in their own plane. This variation is eliminated by illuminating at 45°, from all azimuthal directions. Such illumination is called annular.

In another standard form of illumination, hemispherical illumination, the specimen is uniformly illuminated from nearly all directions by an integrating sphere. A hollow sphere, usually around 150 mm in diameter, is painted matte white inside. Light is introduced through an entrance port. Multiple diffuse reflections cause the whole sphere wall to be uniformly bright. A specimen placed at the sample port is diffusely illuminated. The light reflected by the specimen is measured through the exit port, a few degrees off the specimen normal. White baffles in the sphere block direct rays from the entrance port to the sample port and from the entrance port to the exit port.

Such a measurement includes light specularly reflected at the surface of a glossy specimen. To avoid including the specular component, a specular port is made in the sphere at the same angle to the normal as the exit port, but on the opposite side. A glossy specimen reflects that dark open port to the exit port, and so no light is specularly reflected to the exit port. The specular component is often excluded from measurements of glossy nonmetal specimens, because it characterizes the source, not the specimen. If desired, a white cover is placed on the specular port to include the specular component. If the specimen has a matte surface, light is reflected at the first surface in all directions and cannot be separated and excluded.

The measurement of the color of metals and the evaluation of gloss require measuring reflection in the specular direction. The illuminator and receiver are placed symmetrically about the specimen normal. Specular reflection depends on the angle of incidence, and so that angle is critical.

The sampling aperture is the area on which measurements are made. It may be limited by the influx or the efflux optics. The influx sampling aperture and efflux sampling aperture are usually not the same size. That avoids misalignment errors. Furthermore, many specimens scatter light internally, so that it is necessary to illuminate a small area and collect flux from a larger area, to include light re-emitted outside the influx sampling aperture. Alternatively, one can illuminate a large area and collect flux from a small one.

Any ray of light that passes through an optical system can pass through the same system in the opposite direction. Though this principle of reversibility pertains to rays, it is applied to the design of reflectometers and is generally valid. Illumination at 45° and collection at 0° gives the same results as illumination at 0° and collection at 45°. Likewise, the direction of rays in sphere geometry may be reversed.

2.2 Spectral Design

The spectral conditions of measurement or application are specified in terms of the spectral power distribution of the illuminator and the spectral sensitivity of the receiver, which are called the influx spectrum and efflux spectrum, respectively. These spectra are specified relative to their maximum values. The influx spectrum depends on the spectral power distribution of the source and the spectral effect of the director. The efflux spectrum depends on the spectral sensitivity of the sensor and the spectral effect of the collector. The product, at each wavelength, of the influx spectrum and the efflux spectrum is called the spectral product of the instrument or application.

A spectrophotometer measures reflected flux as a function of wavelength. The wavelength of the light used for the measurement is selected by filters, which may be dyed gelatin, glass, or interference types, or by a dispersive system. A dispersive system employs a diffraction grating or prism to disperse light into a spectrum, so that a desired wavelength can be isolated. Such an optical system is called a monochromator. Gratings are generally preferred to prisms, because they disperse a spectrum with wavelength variation proportional to distance. A monochromator may be placed in the illuminator, to illuminate the specimen with monochromatic light, or in the receiver, to sense a narrow wavelength band. In the absence of fluorescence, either can be used.

Fluorescence is the absorption of radiant energy of a given wavelength and emission of part of that energy at a longer wavelength. Characterization of the excitation spectrum and emission spectrum of a fluorescent specimen requires a monochromator in both the illuminator and receiver. Fluorescent dyes, called fluorescent whitening agents, are added to white paper, plastics, and textiles, to absorb ultraviolet energy and emit blue light, counteracting the undesirable tendency of such materials to look yellowish.

To measure the appearance of a specimen that may be fluorescent, it must be illuminated with simulated daylight or other illuminant of interest, including the ultraviolet component. Light is reflected and emitted by the specimen. The monochromator must be placed in the receiver, so that both components are measured at the designated wavelength.

To measure how light a surface appears to the eye, the reflectance factor must be spectrally weighted by the sensitivity of the human eye. Filters may be used to simulate that sensitivity, or the measured spectral reflectance factor may be mathematically integrated with respect to the standard spectral luminous efficiency function. Measured quantities of this kind are distinguished by the adjective "luminous." We differentiate radiant flux, which may have any spectral distribution, from luminous flux, which is integrated as described, and we speak of luminous reflectance. These quantities are called photometric quantities.

2.3 Light Sources

Incandescent lamps are inexpensive, simple to operate, and provide a smooth spectral power distribution over the whole visible spectrum, but are deficient in ultraviolet power. Xenon-arc lamps are powerful continuous sources with ample ultraviolet. Xenon-flash lamps, introduced in the mid-1970s, provide enough visible and ultraviolet power to simulate daylight and, having a flash duration of about 50 μs, they may be used to measure moving materials, such as paper on a web press. The momentary illumination is so high that, if measurement is made only during the flash, extraneous light, including full sunlight, on the specimen does not interfere. Reflecting properties may depend on temperature, a phenomenon known as thermochromism, and the use of a flash source avoids heating the specimen before measurement. Photodiodes are used in simple inexpensive instruments.

2.4 Sensors

Thermosensitive devices, such as thermocouples, thermopiles, and bolometers, were used as sensors before photocells were invented. Photocells were replaced by phototubes and photomultipliers and, more recently, by solid-state sensors, such as silicon diodes. An array of silicon diodes can record a number of spectral intervals simultaneously.

2.5 Standards

A documentary standard is a document recording national or international agreements on terminology, symbols, physical quantities to be measured, or methods of measurement. Standards are prepared by experts in the field and are reviewed periodically. Standard symbols and terminology facilitate concise and unambiguous communication. Standard methods of measurement assure that values obtained in different places or at different times have the same meaning. Standards on reflectometry are published by ASTM, the American National Standards Institute (ANSI), ISO, and CIE.

A physical standard is a material object, having a physical property, the quantity of which is known, used as the basis of calibrating a measuring instrument. The true value of the physical quantity is assigned by measurement in a standardizing laboratory.

Reflectance factor depends on which standard is chosen. By international agreement, unless otherwise specified, the reflection standard is taken to be an ideal diffuse reflecting surface. That condition is sometimes part of the definition of the term "reflectance factor." No ideal surface actually exists. Working standards are calibrated by standards laboratories, so that measured values can be reported as though they were obtained with an ideal standard. The first widely used white standard was calcium carbonate (chalk), then came magnesium oxide condensed from the fumes of burning magnesium, and then pure barium-sulfate powder pressed into a flat plaque. These have generally been replaced by pressed plaques of powdered polytetrafluoroethylene, sold under the registered trademark "Halon." Such standards are used to transfer primary calibrations. White ceramic tiles are generally used as working standards. For measurement of specular reflection, polished black glass or polished or deposited pure metal surfaces are used.

2.6 Quality of Measurements

The quality of measurements depends on the degree to which the method of measurement meets the standard specifications, on the precision, and on the accuracy. Precision is the degree to which repeated measurements agree, based on statistical analysis of a number of measurements. Accuracy is the closeness of the measured value to the true value, assessed by comparing measured values with those assigned by a standards laboratory. Inaccuracy is generally categorized as failure to meet geometric or spectral specifications (design inaccuracy), failure to indicate values linearly and equal to true values (photometric inaccuracy), or failure to isolate the desired spectral wavelength (wavelength inaccuracy).

2.7 Operation

Reflectometers vary so widely that only a few basic operating rules can be given. Samples must be representative and be prepared and presented for measurement in a standard way. Ideally, specimens should be plane. Options, such as inclusion or exclusion of the specular component, the spectral range, the source, and various standard computational parameters, must be chosen carefully and recorded with measured values. Instruments must be operated within the required temperature range. Some have sensors to detect temperature out of range. Some instruments require calibration with a working standard for each measurement; others require calibration a few times a day. Regularly scheduled battery checks and cleaning of optics are recommended. (Integrating spheres may be degraded by atmospheric pollutants, such as tobacco smoke.) Standard samples are measured to assure the accuracy of measurements and repeated measurements are statistically analyzed. Statistical process control is applied to this measurement process.

3. MANUFACTURE

Manufacture of reflectometers requires careful design, conformance to documentary standards, precise part manufacture (or specification and purchase), and precise assembly. Tasks are of five kinds: mechanical, optical, electronic, computational, and, where analysis is the objective, chemical. Since their introduction, personal computers have been used for computations and data storage. Portable instruments have built-in computers. Most instrument manufacturers assemble components made by firms specializing in certain components. Holographic gratings and sensor arrays have greatly reduced costs.

4. APPLICATIONS

4.1 Chemical Analysis

Infrared reflection spectra correlate with molecular structure, and so many instruments are made to perform chemical analyses. Some are general-purpose laboratory spectrophotometers; others are special-purpose field instruments to measure, for example, the moisture content of grain or the fat content of meat. Many analyses are performed by placing a specimen of blood or urine on treated paper and measuring the spectral reflectance factor of the resulting spot. Simple pocket-sized instruments are used by patients to monitor blood sugar or detect pregnancy. In chemistry, the term "colorimeter" refers to a device, sometimes a simple visual device, to compare colors of chemicals, as part of an analytical procedure.

4.2 Densitometers

Reflection densitometers are used to control photographic and color printing processes. Measurements may be made on images or on control patches printed for the purpose. A "measuring head" includes the illuminator and receiver. There are bench-type instruments, hand-held instruments, and "on-line" instruments mounted on printing presses, to measure during a run. Microdensitometers measure density on a microscopic area, to scan and quantify images and analyze image structure (Dainty and Shaw, 1974; Shaw, 1977; Swing, 1995). Measurements on very small areas involve the optical theory of partial coherence (Born and Wolf, 1965).

The standard geometry for densitometry is 45° annular. Reflectance factor is measured, but a derived quantity, reflection density, is indicated. Reflection density is the logarithm, to base ten, of the reciprocal of the reflectance factor. It increases with the amount of dye or ink present. The yellow, magenta, and cyan dyes or inks are measured in the spectral region of maximum absorption, with narrow-band blue, green, and red light, respectively. Densitometry is used to control amounts of colorants, but the colors produced depend on the colorants, and so densitometry is not colorimetry. Densitometry is standardized by ANSI and ISO.

4.3 Colorimetry

Color is a psychophysical quantity; that is, it depends not only on the spectral quality of the light source and the spectral reflectance factor of the specimen, but also on the nature of the human visual system. Data specifying the normal human observer and methods of color measurement have been standardized by the CIE and ASTM.

The spectrum of light reflected by a surface depends on the spectrum of the illumination, and so the CIE has standardized the spectra of common illuminants. An incandescent lamp, some fluorescent lamps, and daylight have been standardized. Phases of daylight are identified by correlated color temperature. (The correlated color temperature of a light is the temperature in kelvins of a thermal radiator having the same color.) Sunlight, average daylight, and north-sky daylight have correlated color temperatures of 5500, 6500, and 7500 K, respectively, and are given the standard symbols D55, D65, and D75.

CIE standardized the nature of normal human color vision, for use in colorimetry, by adopting three color-mixture functions, which are the amounts of CIE primary colored lights, in a mixture of three primaries, required to match the color of light of a given wavelength.

Usually, a spectrophotometer is used to measure the spectral reflectance factor. That quantity must be mathematically integrated with respect to a chosen illuminant and the three CIE color-mixture functions to determine tristimulus values, the amounts of the three primaries required to match the color of the given surface.

In a filter colorimeter, the specimen is illuminated by a simulated standard illuminant and reflected light is measured with sensors fitted with filters designed to simulate the three color-mixture functions, so that tristimulus values are read directly. These instruments are simpler and less expensive than a spectrophotometer, but less stable, precise, and accurate. In color technology, the term "colorimeter" usually refers to this type of instrument.

The tristimulus values are mathematically transformed in many ways to correlate with colors and color differences as they are perceived (Wyszecki and Stiles, 1982). The derived quantities are used to specify colors and color differences, compute colorant formulations to produce desired colors, correct off-color batches, and control processes that produce colors.

Color-measuring instruments may be bench type, hand-held, or on-line. Bench-type instruments are in the form of a box with an aperture on which the specimen is placed for measurement. A personal computer controls the operation of the spectrophotometer, records data, performs computations, displays results, and communicates with process-control instruments. Hand-held instruments perform the same kinds of functions, but with limited capacity. On-line instruments mounted on production machines must be resistant to vibration, moisture, heat, and chemical fumes. They are used to control the paper-making process without human intervention, a system known as closed-loop process control.

The colors of metallic automobile paints are characterized by measuring color with three different geometries. The specimen is illuminated at 45° to the normal and reflected light is measured at three widely different angles. Hand-held multiangle spectrophotometers have come into widespread use in the automotive industry. This is an example of "gonioreflectometry," the measurement of reflection as a function of angles. Standardization of such measurements is addressed by ASTM. The luster of textiles can be characterized by gonioreflectometry.

Color may be indicative of some aspect of quality, even when color itself is of little or no concern. For example, the measured color of toasted corn flakes indicates how well done they are and predicts the flavor. Sheet materials to package milk and juices for shelf storage have several layers to provide strength, retain contents, prevent influx of oxygen and other gases, block ink odors, etc. They are coextruded and to measure the thickness of each layer, the methodology of quantitative chemical analysis by infrared reflectometry is used. Very thin transparent films present interference colors, and the thickness may be gauged by colorimetry.

4.4 Gloss, Distinctness of Image, and Haze

Gloss is evaluated on a 100-point scale by measuring the specular reflectance factor, relative to highly polished black glass of a specified refractive index. The angle of incidence may be 20°, 30°, 45°, 60°, 75°, or 85°. The instrument is called a gloss meter. The spectral product must be the product of the luminous efficiency function and the spectrum of daylight. Some instruments have sensors to assure precise alignment with the specimen. Gloss meters are widely used in the paint, plastics, ceramics, and paper industries. Illuminating at 30° and receiving at 30° is called 30° gloss, receiving at 30.3° evaluates distinctness of image (DOI), and receiving at 32° or 35° evaluates 2° haze or 5° haze. Methods are standardized by ASTM.

5. ECONOMIC ASPECTS

All industries concerned with appearance depend on reflectometry to quantify and specify appearance, predict colorant formulations, control production processes, and assure quality. These industries include photography, printing, paints, plastics, and textiles, as well as all other industries, such as the automobile, clothing, housing, packaging, and advertising industries, that are supplied by these industries. Appearance is the most obvious aspect of quality perceived by the potential buyer of most products. Appearance may be a crucial factor in the decision to buy, and a variety of colors or finishes promotes sales. Reflectometry and formulation software put custom color formulation in neighborhood paint stores. Chemical analysis by reflectometry takes a small fraction of the time and effort required by gravimetric wet chemistry, and thickness measurements, easily made by reflectometry, are virtually impossible by any other nondestructive means.

GLOSSARY

Aperture: An opening permitting the flow of radiant energy.

Appearance: The way a surface looks, as characterized by color, gloss, texture, grain, etc.

Emission Spectrum: In fluorescence theory, the spectral power distribution of flux emitted by a material as a result of fluorescence.

Excitation Spectrum: The spectral distribution of energy that causes a fluorescent material to absorb energy and emit at a longer wavelength.

Filter: An optical window that transmits a selected spectrum.

Flux: The time rate of flow radiant energy.

Port: An opening in an integrating sphere.

Specimen: A sample of a class of objects, chosen to be measured.

Spectral: Pertaining to a particular wavelength, a narrow band of wavelengths, or a distribution of radiant power with respect to wavelength.

Spectrum: The spatial arrangement of components of radiant power in order of wavelength.

Specular: Reflection, as by a mirror, with the angle of reflection equal to the angle of incidence.

Wavelength: The distance in the direction of propagation between nearest maxima of an electromagnetic wave, such as a light wave. In reflectometry, the nanometer is the unit most often used.

Works Cited

Born, M., Wolf, E. (1965), *Principles of Optics,* New York: Pergamon.

Dainty, J. C., Shaw, R. (1974), *Image Science,* New York: Academic.

McCamy, C. S. (1966), *Photogr. Sci. Eng.* **10,** 314–325.

Shaw, R. (1977), *Image Analysis and Evaluation,* Washington, DC: Society of Photographic Scientists and Engineers.

Swing, R. E. (1995), *Selected Papers on Microdensitometry,* Bellingham, WA: SPIE Optical Engineering Press.

Wyszecki, G., Stiles, W. S. (1982), *Color Science: Concepts and Methods, Quantitative Data and Formulae,* 2nd ed., New York: Wiley.

Further Reading

Andrews, D. L. (1990), *Current Topics in Modern Chemical Spectroscopy,* New York: Springer-Verlag.

Burgess, C., Mielenz, K. D. (Eds.) (1987), *Ad-*

vances in Standards and Methodology in Spectrophotometry, New York: Elsevier.

Grum, F., Bartleson, C. J. (Eds.) (1980), *Optical Radiation Measurements,* Vol. 2, *Color Measurement,* New York: Academic.

Hunter, R. S., Harold, R. W. (1987), *The Measurement of Appearance,* 2nd ed., New York: Wiley.

Stearns, E. I. (1969), *The Practice of Absorption Spectrophotometry,* New York: Wiley.

REFRACTOMETERS, OPTICAL

W. NEBE, *Carl Zeiss Jena GmbH, Jena, Germany*

Introduction 255
1. **Physical Foundations** 256
1.1 Refraction of Light 256
1.2 Prism Combinations 258
1.3 Critical-Angle Refractometry .. 258
1.4 Intensity Relations 258
1.5 Interference of Light 259
1.6 Refractive-Index Gradients 260
2. **Refractometers** 261
2.1 Deviation-Measuring Refractometers 261
2.1.1 Spectrometer 261
2.1.2 Autocollimation Spectrometer 261
2.1.3 Double-Prism Refractometer .. 261
2.1.4 Vee-Prism Refractometer 263
2.2 Critical-Angle Refractometers 263
2.2.1 Abbe Refractometer 263
2.2.2 Pulfrich Refractometer 263
2.2.3 Dipping Refractometer 264
2.2.4 Hand Refractometer 265
2.3 Automatic Refractometers 265
2.3.1 Deviation-Measuring Automatic Refractometers 265
2.3.2 Intensity-Measuring Refractometers 265
2.3.3 Integrated-Optics Refractometers 266
2.4 Interference Refractometers .. 266
2.4.1 Rayleigh–Loewe Interferometer 266
2.4.2 Jamin Interferometer 267
2.4.3 Photoelectric Interferometers . 268
2.5 Refractive-Index-Gradient–Measuring Refractometers 268
2.5.1 Schlieren Apparatus 268
2.5.2 Interferometers 269
3. **Refractometry of Homogeneous Substances** ... 269
3.1 Glasses and Other Solids 269
3.2 Solutions and Binary Mixtures of Liquids 270
3.3 Ternary Mixtures of Liquids .. 271
3.4 Continuously Varying Liquids 272
3.5 Binary Mixtures of Gases 272
3.6 Ternary Mixtures of Gases 273
4. **Refractometry of Inhomogeneous Substances** 273
4.1 Transparent Solids 274
4.2 Schlieren in Liquids 274
4.3 Stratified Solutions 274
4.4 Inhomogeneous Gases 274
Glossary 275
Works Cited 275
Further Reading 276

INTRODUCTION

Refractometry means the investigation of substances by measuring their refractive index, which is a characteristic property of any homogeneous substance, and by observing variations of refractive index in inhomogeneous substances.

After the fundamental discoveries made by Snell (1591–1626) and Huygens (1629–1695)—the law of refraction and the nature of light as a wave phenomenon, respectively—the first purpose-oriented measurements of refractive index can be dated to the beginning of the 19th century. Fraunhofer (1787–1826) determined the refractive indices of glasses with the aid of the light deflection by a prism, and Arago (1786–1853) was able to ascertain the refractive index of air by means of interference methods. The decisive impulse for developing the rather sporadic measurements into a standard analytical method, known as refractometry, was Abbe's (1840–1905) idea to use the refractive index of a solution as a measure of its concentration (1869). Subsequently, the company of Carl Zeiss in Jena, Germany, developed various refractometers tailored for specific measurement of solids, liquids, or gases. Technical advancements such as the development of powerful spectral light

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

sources and the use of new glass types improved the precision of the instruments and expanded the scope of their application.

Since the 1950s, refractometry has been significantly boosted by the development of electronics. Whereas before, observations were purely visual, i.e., subjective, the use of electronic devices has created capabilities not only for objective measurement, but also for automation, which is important for the checking and control of continuously running industrial processes as well as for the monitoring of chemical reactions in the laboratory. Recent efforts are directed at miniaturizing the instruments and, above all, at minimizing the sample volumes required, by making use of integrated optical devices, which provide new possibilities of analytical applications.

This article describes instruments and methods of refractometry that are widely applied in many fields of chemical research, industrial control, and medical diagnostics.

1. PHYSICAL FOUNDATIONS

The refractive index (RI) of a transparent medium is defined as the relation of velocity of light in a vacuum (c) to that in this medium (c_M):

$$n = c/c_M. \tag{1}$$

From this there follow various relations for the behavior of light that are employed in refractometry.

1.1 Refraction of Light

If light passes the boundary area of two substances with different RI (n_A,n_B) it will be refracted according to Snell's law:

$$\sin(\alpha)/\sin(\beta) = n_B/n_A, \tag{2}$$

where α = angle of incidence, β = angle of refraction. The angles α and β and the normal to the boundary surface between the two media lie in one plane.

A light beam passing a glass prism is bent from its original direction by an angle δ that is dependent on α, the apex angle (φ), and the RI (N) of the prism and the RI of the surrounding media (n,n') (Fig. 1):

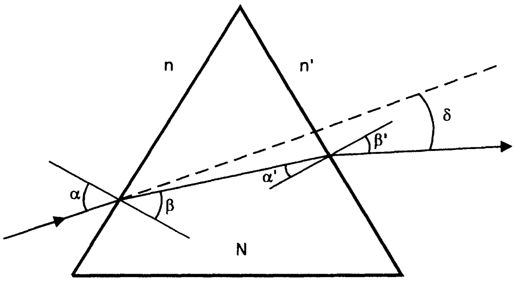

FIG. 1. Refraction by a prism.

$$\sin(\alpha)/\sin(\beta) = N/n, \quad \alpha' = \varphi - \beta;$$
$$\sin(\alpha')/\sin(\beta') = n'/N, \quad \delta = \alpha - \varphi + \beta'. \tag{3}$$

If the angle of incidence is $\alpha = 0$ then also $\beta = 0$ and

$$\delta = \beta' - \varphi, \quad \sin(\beta') = \sin(\varphi)N/n'. \tag{4}$$

Another special case will happen if the angle of deviation reaches a minimum. Then, supposing $n' = n$, the light passes the prism symmetrically, and from Eq. (3) follows

$$N = n\frac{\sin[(\delta + \varphi)/2]}{\sin(\varphi/2)}. \tag{5}$$

Generally measurements of N are made in relation to air so that $n' = n = 1$.

The RI of a medium is dependent on the wavelength of light. A beam of white light passing a prism is split up according to Eq. (3), the refracted beams forming a colored spectrum with different wavelengths, a phenomenon known as dispersion. In optical refractometry certain wavelengths in the visible region of the spectrum are preferred; they are emitted by spectral lamps, light-emitting diodes (LEDs), or lasers (Table 1). Like the RI, the dispersion of light is a characteristic parameter of a medium. It is quantified by the difference of the RI at two wavelengths:

$$D = n_F - n_C, \quad \text{or} \quad D = n_{F'} - n_{C'}. \tag{6}$$

A substance without an absorption peak in the visible region shows a normal dispersion, i.e., its RI decreases with increasing wavelength (Fig. 2). If the RI is known for three different wavelengths, the RI for any

Table 1. Light sources and preferred wavelengths for refractometry.

Light source	Color	Symbol	Wavelength (nm)
Spectral lamps			
Helium (He)	Red	···	706.5
Hydrogen (H)	Red	C	656.3
Cadmium (Cd)	Red	C′	643.8
Sodium (Na)	Yellow	D	589.3
Helium (He)	Yellow	d	587.6
Mercury (Hg)	Green	e	546.1
Cadmium (Cd)	Blue	···	508.6
Hydrogen (H)	Blue	F	486.1
Cadmium (Cd)	Blue	F′	480.0
Mercury (Hg)	Violet	g	435.8
Mercury (Hg)	Violet	h	404.7
LED	Red	···	640–660
HeNe laser	Red	···	632.8

other wavelength is represented approximately by a formula by Cauchy:

$$n = A + B/\lambda^2 + C/\lambda^4, \qquad (7a)$$

where A, B, C are constants characteristic of a substance. In the region of an absorption band, Eq. (7a) must be replaced by Sellmeier's formula

$$n^2 = 1 + A\lambda^2/(\lambda^2 - \lambda_0^2), \qquad (7b)$$

where A, λ_0 are constants, the latter corresponding to the wavelength of the absorption peak. In this region anomalous dispersion appears, i.e., RI increases with increasing wavelength.

If light passes an anisotropic medium, e.g., a crystal plate of calcite or quartz, double refraction (birefringence) may occur. The light beam is split up into two refracted beams that are polarized perpendicularly to each other and that have different angles of refraction. One beam is refracted according to Eq. (2), irrespective of the position of the crystal axis ("ordinary ray"), while the refraction of the other is dependent on this position and the angle of incidence ("extraordinary ray").

A birefringent crystal is characterized by two RI: n_o (ordinary ray) and n_e (the RI that deviates extremely from n_o, for the extraordinary ray) (Table 2). Most crystals occurring in nature possess not only one axis (uniaxial) but two crystal axes (biaxial). Therefore, they are characterized by three principal RI.

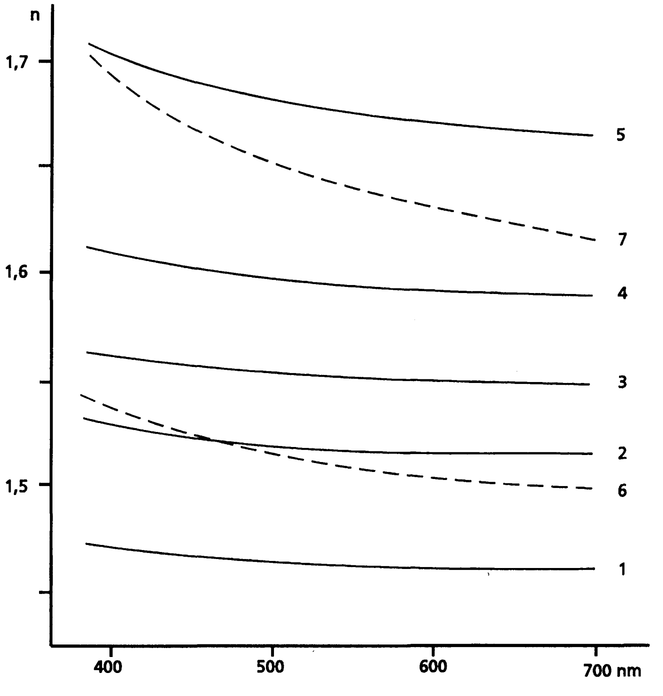

FIG. 2. Dispersion curves of some solids and liquids. (1) Fused quartz; (2) borosilicate crown; (3) crystalline quartz (*o*); (4) light flint; (5) dense flint; (6) benzene; (7) carbon disulfide.

Table 2. Refractive indices, dispersions, and Abbe numbers of some optical materials. (*o*) ordinary ray, (*e*) extraordinary ray.

Material	n_D (20°)	$n_E - n_C$	Abbe number
Quartz, fused	1.45845	0.00678	67.6
Crystalline (*o*)	1.54425	0.00778	70.0
Crystalline (*e*)	1.55336	0.00805	68.7
Calcite			
(*o*)	1.65836	0.01347	48.9
(*e*)	1.48641	0.00615	79.1
Borosilicate crown	1.5170	0.00802	64.5
Light barium crown	1.5410	0.00907	59.7
Light flint	1.5760	0.01398	41.2
Dense flint	1.6490	0.01913	33.9
Dense barium flint	1.6705	0.01409	47.5
Extra dense flint	1.7200	0.02477	29.1

1.2 Prism Combinations

Two prisms may be in contact so that the contact faces are inclined to the outer faces at a known angle φ, while the outer faces are parallel to each other. If the angle of incidence at the first prism is $\alpha = 0$ (Fig. 3) and the RI of the surrounding medium $n = 1$, the angle of deviation can be calculated from

$$\sin(\delta) = n_B \sin(\beta - \varphi), \quad \sin(\beta) = \sin(\varphi) n_A / n_B. \tag{8}$$

Let n_B be the known RI of a reference medium; then n_A is calculated from δ.

A combination of three prisms is of some importance for refractometers to avoid disturbing effects of dispersion. Prisms of different glasses (crown and flint) are arranged in contact with each other, their angles being exactly calculated so that light of the central wavelength (*D*) emerges from the triplex prism parallel to the incident light while the red (*C*) and the blue (*F*) beams are deflected in opposite directions (Fig. 4). This type is known as a direct-vision or Amici prism. The angle of deflection may reach a few degrees.

1.3 Critical-Angle Refractometry

In Eq. (2) we assume n_B to be greater than n_A. Then the largest angle of incidence, $\alpha = 90°$, results in a refractive angle γ with

$$\sin(\gamma) = n_A / n_B. \tag{9a}$$

No light will enter the second medium at an angle larger than γ [Fig. 5(a)]. γ is called the "critical angle of refraction." By means of a telescope it is possible to observe the boundary line between a bright and a dark field and to read the critical angle off a circular scale.

If, on the other hand, we assume $n_{B'}$ to be smaller than $n_{A'}$, there also exists an angle γ' such that

$$\sin(\gamma') = n_{B'} / n_{A'}. \tag{9b}$$

While a light beam with an incident angle smaller than γ' will enter the second medium, each light beam with α greater than γ' will be totally reflected back into the first medium [Fig. 5(b)]. γ' is called the "critical angle of total reflection." Here again the border line between a bright and a less bright area can be observed in the eyepiece of a telescope. The position of this border line is a measure of the critical angle and thus of the RI. If $n_A/n_B = n_{B'}/n_{A'}$, then $\gamma' = \gamma$.

1.4 Intensity Relations

With each refraction of light at the interface of two substances, a portion of light will be reflected back into the first medium. This portion is dependent on the parameters α, n_A, n_B according to Eq. (2) and to Fresnel's formulas

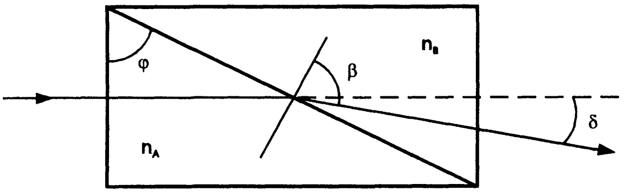

FIG. 3. Deviation of light by a double prism.

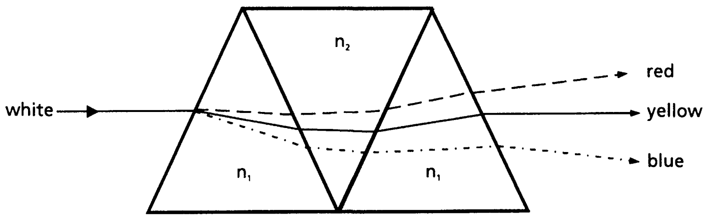

FIG. 4. Dispersion by a direct-vision prism.

$$R_s = -\frac{\sin(\alpha - \beta)}{\sin(\alpha + \beta)} I_s, \quad R_p = \frac{\tan(\alpha - \beta)}{\tan(\alpha + \beta)} I_p,$$
$$R_s' = \frac{2\cos(\alpha)\sin(\beta)}{\sin(\alpha + \beta)} I_s,$$
$$R_p' = \frac{2\cos(\alpha)\sin(\beta)}{\sin(\alpha + \beta)\cos(\alpha - \beta)} I_p, \qquad (10)$$

in which I, R, R' are amplitudes of the electric vectors of the incident, the reflected and the refracted beam; indices s and p denote the state of polarization of light perpendicular and parallel to the plane of incidence, respectively. Intensities are the squares of amplitudes. Although Eqs. (10) are not used directly in refractometry, they form the background of intensity measurements based on RI relations.

In this mode of measurement the geometrical form of the contact surface between the two substances is irrelevant: It may be plane, cylindrical, spherical, or otherwise. According to Eq. (2) n_B/n_A only depends on α and β. If the parameters α (i.e., the geometrical conditions) and I of an arrangement are kept constant, the relation n_B/n_A is found by measuring the resultant intensity of the refracted or of the reflected light.

Like the geometry of the contact face the wavelength of the incident light is not relevant either. It is necessary, however, that each refractometer is calibrated individually.

1.5 Interference of Light

Light may be understood as a propagation of waves of electromagnetic energy. The velocity of propagation in a medium is

$$c_M = \lambda\nu, \qquad (11)$$

in which λ = wavelength, ν = frequency of vibration.

If a light beam passes along a path of length s through a medium of RI n, the term sn is called "optical path length," and the difference from another optical path length is measured in terms of a multiple of the wavelength:

$$sn_2 - sn_1 = k\lambda \quad \text{or} \quad n_2 = n_1 + k\lambda/s. \qquad (12)$$

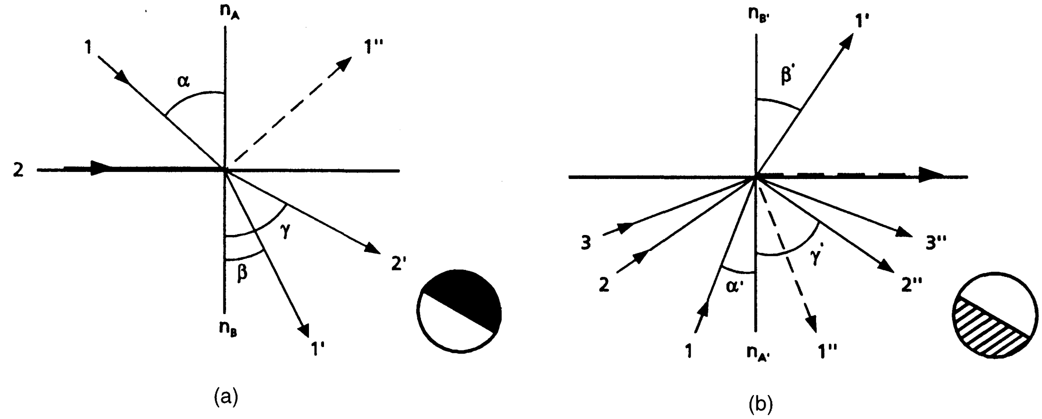

(a)

(b)

FIG. 5. Refraction (1′,2′) and reflection (1″,2″,3″) at the interface of two media; (a) n_A smaller than n_B; (b) $n_{A'}$, larger than $n_{B'}$. To the right of it, are the fields observed by a telescope.

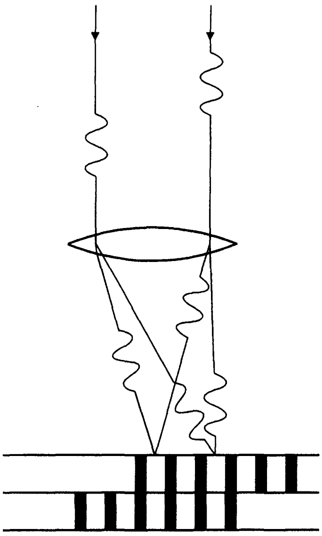

FIG. 6. Interference of two coherent light beams, with a phase difference of two wavelengths, in the focal plane of an objective. The position of the lower fringe pattern would be valid for a phase difference of zero. In the figure, only six fringes of the extended system are shown. In reality, the fringes have no sharp edges, but intensity changes continuously between minimum and maximum.

As light is emitted in quanta, it propagates in wave trains of finite length (coherence length). Two light waves are called "coherent" if they propagate in the same manner of vibration (equal wavelengths, amplitudes, and state of polarization) and if the difference of their optical paths is smaller than their coherence length. Two coherent waves superimposed on each other at one point may interfere. The superposition of the vibrations will result in an increase or decrease of the intensity at this point, so that under special phase relations complete extinction takes place at this point.

An interferometric arrangement using a *monochromatic* light source produces a system of many interference fringes. The number of fringes depends on the degree of coherence of the interfering beams and increases with the coherence length. A phase difference of the light beams causes a displacement of the fringes (Fig. 6). As all fringes look alike it is not possible to distinguish the fringe of zeroth order ($k = 0$), which marks the start position.

A *white* light source, however, because of dispersion of light, produces a superposition of fringe systems of each wavelength of the spectrum. A white zero-order fringe is seen between two black half-order fringes while the fringes of the first and higher orders show colored zones. Consequently the zero-order fringe is an excellent indication mark.

1.6 Refractive-Index Gradients

RI gradients, i.e., variations of RI within a medium, appear in the form of striae, also called "schlieren" or streaks. Striae are inhomogeneous zones within an otherwise homogeneous substance; they have no sharp borderlines, and their RI changes continuously from the margin to the center.

The most convenient method for determination of RI gradients by aid of geometrical optics is the "schlieren method" by Toepler (1864) (Fig. 7). A light beam passing striae will be deviated, the magnitude of deviation depending on the RI gradient dn/dx as well as on the path length (s) within the striae. Approximately we have

$$\delta = \frac{s}{n}\frac{dn}{dx}, \quad n = \text{RI of the surrounding.} \tag{13}$$

A knife edge ("Foucault edge") is arranged in the focal plane of the objective with its edge parallel to the image of the slit. A part of the light that is deviated by striae will fall on the knife edge and is thereby prevented from reaching the screen. Consequently,

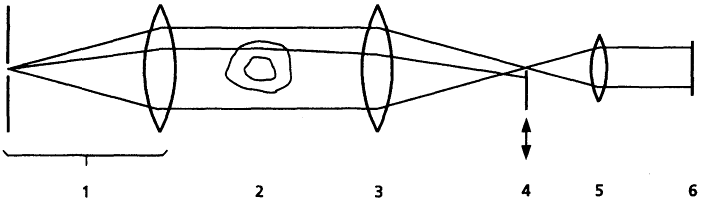

FIG. 7. Toepler arrangement for the observation of striae. (1) Collimator with entrance slit; (2) striae; (3) objective; (4) knife edge; (5) lens system; (6) screen.

striae are recognized on the screen as dark (or bright) zones in a bright (or dark) field; see Schardin (1942).

A disordered distribution of inhomogeneities generally allows only a qualitative analysis by observation of the light deviation. Special techniques allow a partially quantitative determination of the magnitude and direction of the RI gradients (Sec. 2.5.1).

For the quantitative analysis of weak inhomogeneities, interferometers of the Michelson, Twyman, or Mach–Zehnder type are applied; see INTERFEROMETERS AND INTERFEROMETRY.

In stratified liquids, there exists a constant *dn/dx* in each layer, and all gradients have the same direction. In a cell with stratified solutions, the deviation (regarding refraction at the exit window) will be

$$\delta = s\,dn/dx, \quad s = \text{length of the cell.} \qquad (14)$$

Interferometric investigations of stratified solutions are also used. They are based on the measurement of changes of the optical path length according to Eq. (12).

2. REFRACTOMETERS

2.1 Deviation-Measuring Refractometers

According to Eq. (3) these instruments measure the deflection of a defined light beam passing a homogeneous prismatic object.

2.1.1 Spectrometer By means of a collimator, a parallel bundle of light is produced and directed onto the prismatic sample, where it is deflected (Fig. 8). In the eyepiece of a telescope, the position of the image of the entrance slit is observed before and after inserting the prism in its position with predetermined angle α; e.g., for $\alpha = 0$ Eq. (4) holds. The angle of rotation of the telescope is read off at a scale. To determine the RI, the apex angle φ of the prism must be known.

The entrance slit is illuminated by a monochromatic light source. Several light sources may be fitted interchangeably with one another (Table 1).

High-precision spectrometers allow the determination of the absolute RI of glasses and other transparent solids at different wavelengths with an accuracy of some 10^{-6} RI units (RIU). The sample, however, must fulfill strong conditions: It must be transparent, homogeneous (without any striae or scattering effects), well thermostatted; its optical faces must be exactly plane, its diameter must be large, and the prism angle must be known to 1″. Moreover the RI of the surrounding medium must be controlled exactly.

The measuring range is not limited. An exact determination of dispersion is possible.

2.1.2 Autocollimation Spectrometer An "autocollimation telescope" generates a parallel bundle of light and, after reflection of the bundle back toward its original direction, allows the observation of the entrance slit (Fig. 9). This instrument is very well suited for measuring RI *and* prism angle by turning the prism around a vertical axis and reading the angles in three positions of rotation. The controlling telescope remains in a fixed position, thus providing greater convenience of measurement.

2.1.3 Double-Prism Refractometer The determination of the RI of a substance in relation to that of a reference medium is

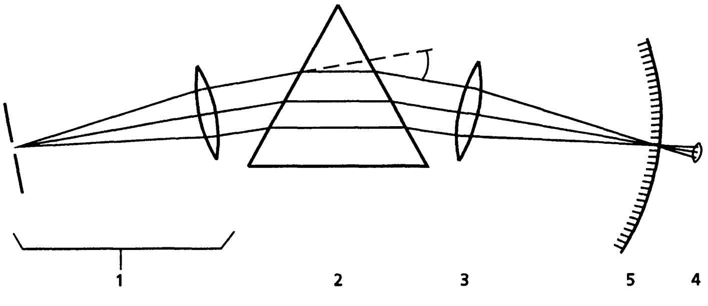

FIG. 8. Spectrometer arrangement (illumination with monochromatic light). (1) Collimator; (2) prism (sample); (3 and 4) objective and eyepiece of the telescope, respectively; (5) circular scale.

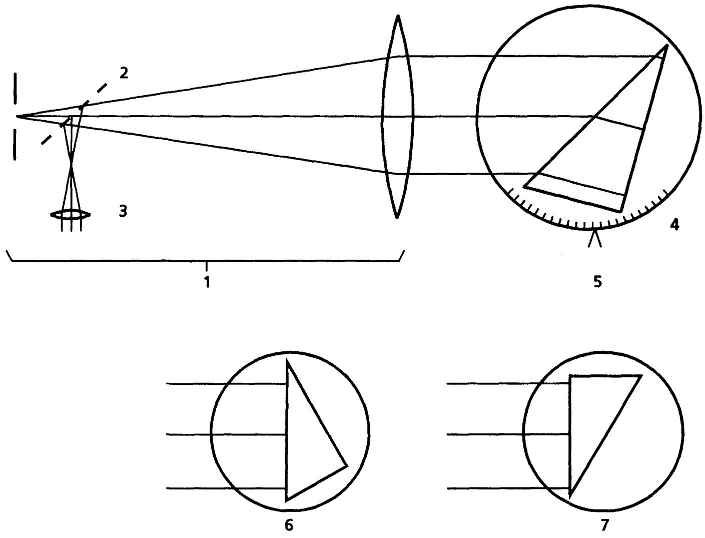

FIG. 9. Autocollimation spectrometer for the measurement of prism angle and deviation of light. (1) Autocollimation telescope with (2) semitransparent glass plate and (3) eyepiece; (4) rotating table with prism and circular scale; (5) reading device with indicative mark; (6 and 7) two further positions of the table for measuring the angle of the prism.

established if both substances take the form of prisms. For the investigation of liquids a double cell consisting of two hollow prisms is illuminated by a collimator. One prism is filled with a known reference liquid, and the other with the sample. An objective projects an image of the entrance slit onto a divided scale. The deviation of light by different substances in the double prism changes the position of the slit image; from the displacement and the focal length of the objective, the RI difference is calculated according to Eq. (8). The scale is nowadays usually replaced by a detector, such as a differential photocell or a position-sensitive detector, e.g., a charge-coupled device (CCD) array. A built-in computer calculates the RI of the sample [Fig. 10(a)].

This type of refractometer is excellently suited for the continuous measurement of

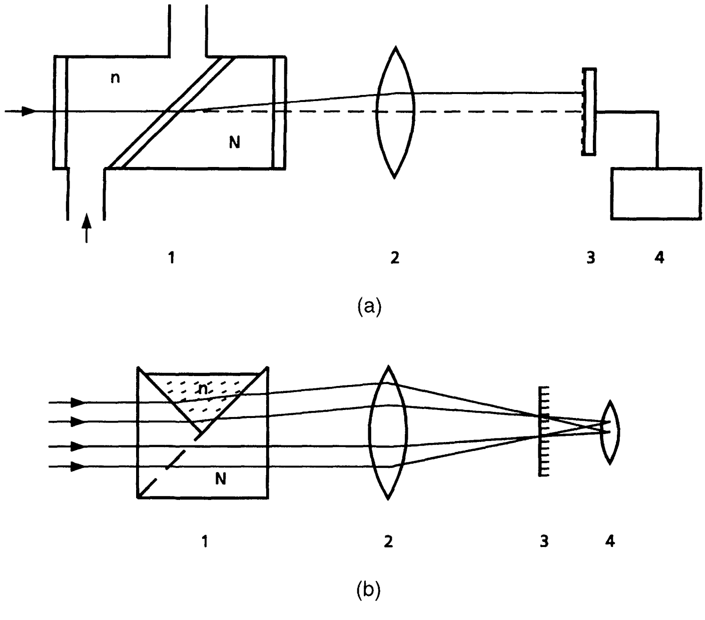

FIG. 10. Measurement of deviation with (a) flow-through refractometer with (1) double-prism cell for sample (*n*) and reference medium (*N*), (2) objective, (3) position-sensitive detector, and (4) electrical unit; (b) refractometer with (1) Vee block prism (*N*) and sample (*n*, solid or liquid), (2) objective, (3) scale, and (4) eyepiece.

liquids flowing through the sample cell. This arrangement has many advantages for continuous industrial control:

no moving parts in the instrument;
high accuracy better than 1 × 10^{-5} RIU dependent on the prism angle, the focal length of the objective, and the resolution of the detector;
the transmission of the sample may be rather low;
sedimentation of sample particles at the interior faces of the windows is largely negligible;
automatic temperature compensation by the close proximity of the liquids and, moreover, by the sample running around the reference cell before entering the flow-through cell.

2.1.4 Vee-Prism Refractometer In this refractometer a vee block consisting of two prisms of the same high-refracting glass holds the sample [Fig. 10(b)], which may be liquid or solid. The vee-block angle is exactly 90°. Light passing the lower part remains undeviated while that traversing the upper part is deviated according to the RI difference between the sample (n) and the vee block (N). From Eq. (2) there results

$$\sin(\delta) = (\sqrt{1.5\,N^2 - n^2} - \sqrt{n^2 - 0.5\,N^2})/\sqrt{2},$$
$$n^2 = N^2 - \sin(\delta)\sqrt{N^2 - \sin^2(\delta)}. \tag{15}$$

The very small entrance slit is illuminated by an interchangeable spectral lamp combined with filters to select the required monochromatic radiation. The telescope is rotated about a horizontal axis. The circular scale is divided into degrees and allows the determination of RI to an accuracy of up to 1 × 10^{-5} RIU absolute and nearly 3 × 10^{-6} RIU for the measurement of RI differences and for dispersion.

The prism block is surrounded by a hollow jacket containing a thermostatted liquid.

Solid samples must have two polished faces at right angles to each other. With two samples arranged side by side in the vee block, a very exact comparison is possible.

2.2 Critical-Angle Refractometers

The special advantage of this type of refractometer compared with deviation measurement lies in the simpler construction, as no collimating device is necessary, and in the very easy change of samples.

However, the measuring range is limited by the RI of the measuring prism. Determination of RI difference in a direct way is not possible. The sharpness of the black–dark borderline is decisive for accuracy.

2.2.1 Abbe Refractometer The original type Abbe refractometer provides a rotation of the prism relative to the telescope [Fig. 11(a)]. The measuring prism is complemented by a prism confining the sample space on the illumination side. Only a few mm^3 of sample are required. The prisms are contained in hollow jackets for thermostatting. By rotating the prism assembly, the bright–dark boundary is brought to coincidence with the measuring mark of the telescope. The angle of rotation is read off a circular scale through a microscope. The eyepieces of telescope and microscope are situated side by side for convenient observation.

Nowadays most Abbe refractometers have a fixed measuring prism with a horizontal measuring face. Rotation is effected by a rotatable mirror arranged between prism and telescope. In the eyepiece, the borderline and the circular scale are seen simultaneously. Most instruments contain two scales, one each for RI and for percentage sugar dissolved in water. The measuring range for RI extends from 1.3 to 1.7 and is limited by the RI of the measuring prism. The accuracy is about 1 × 10^{-4} RIU.

Under illumination with white light—tungsten lamp or sun light—the bright–dark borderline appears colored because of dispersion. To avoid this disturbing effect, a built-in color compensator is used, which consists of a pair of direct-vision prisms (Sec. 1.1) rotatable about the optical axis.

The Abbe-type critical-angle refractometer allows measurement with both transmitted and reflected light, so that also opaque liquids and solids as well as plastics may be observed.

2.2.2 Pulfrich Refractometer The Pulfrich refractometer in its original form provides critical-angle measurements with grazing incidence of light, $\alpha = 90°$ [Fig. 11(b)]. The measuring prism (N) has a fixed position with a horizontal measuring face and is illuminated by spectral lamps (Table 1) in com-

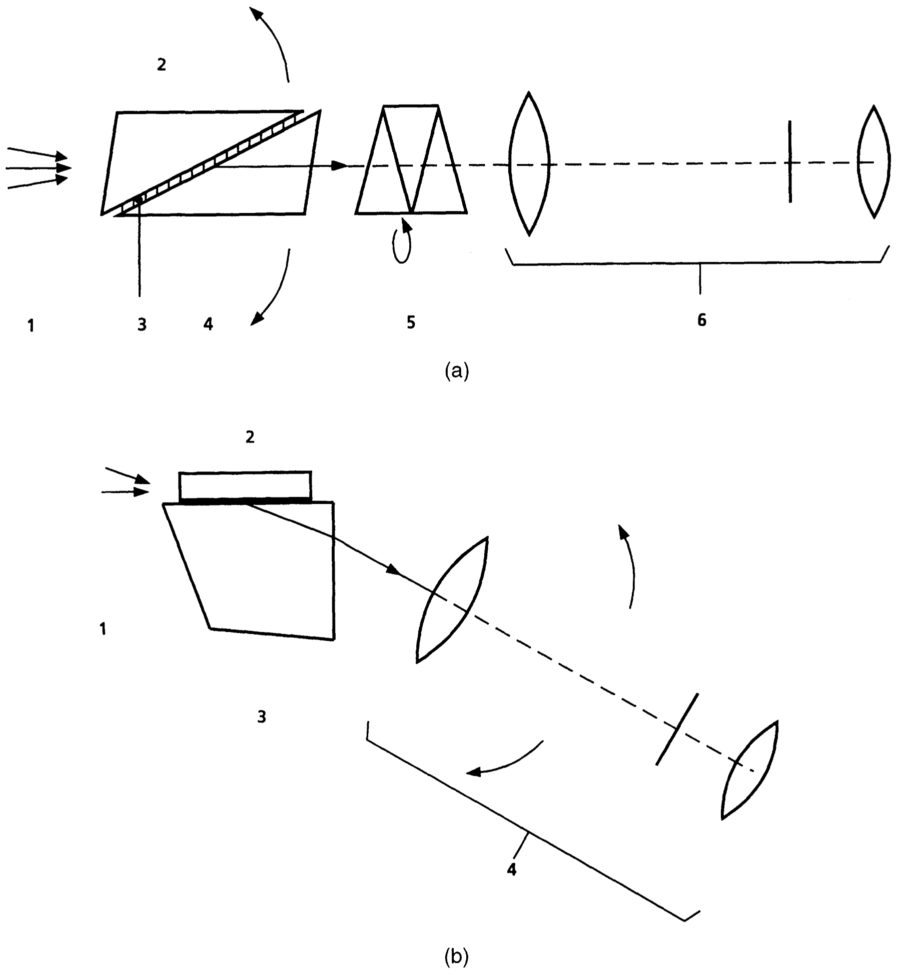

FIG. 11. Measurement of the position of the critical angle borderline. (a) Abbe refractometer: (1) diffuse illumination with white light, (2) prism for light entrance, (3) liquid sample, (4) measuring prism, (5) direct vision prism, and (6) telescope. (b) Pulfrich refractometer: (1) diffuse illumination with monochromatic light, (2) solid sample, (3) 90° measuring prism, and (4) rotatable telescope.

bination with monochromatic filters. As the prism angle is 90°, Eq. (3) results in

$$n = \sqrt{N^2 - \sin^2(\delta)}. \tag{16}$$

The borderline of the bright and dark fields is brought to the center of diagonal cross wires by rotating the telescope about a horizontal axis. The PR2 Pulfrich refractometer (Carl Zeiss Jena) has interchangeable prisms and illumination units to enable both critical-angle and deviation measurement (Sec. 2.1.4).

The measuring range of critical-angle measurement is limited by the RI of the measuring prism. The accuracy is nearly 2×10^{-5} RIU, an exactly constant temperature provided.

A special type of Pulfrich refractometer is a crystal refractometer, which has a hemisphere of highly refracting glass instead of the measuring prism with two polished plane faces.

2.2.3 Dipping Refractometer To avoid inaccuracies arising from rotatable parts, some refractometer types hold measuring prism and telescope in a fixed position. The bright–dark boundary is observed in the eyepiece and its position is directly read off a fixed scale.

To permit quick and convenient measurement and to provide an exactly grazing incidence of light at the prism surface, the prism end of the refractometer is dipped into the liquid sample contained in a small cup, which is immersed in a thermostatting water bath. More commonly, however, one uses a prism in a temperature-controlled jacket.

The measuring range is small, only about 0.03 RIU, but it may be extended by a set of interchangeable prisms with different RI.

A color-compensating prism is built in just as in the Abbe refractometer so that white-light illumination is possible. For high-precision measurements, a monochromatic light source such as a Na spectral lamp is preferred. A long focal length of the telescope objective makes a high accuracy of 2×10^{-5} RIU possible. The scale is divided into equal intervals, and RI is found from tables.

2.2.4 Hand Refractometer This type is designed for special applications in mobile use. The main characteristics of visual instruments are the rigid connection of prism and telescope; small measuring range about 0.03 RIU; scale divided into percentage of a special solution; no thermostatting. Some types of hand refractometers are temperature controlled by way of a bimetal-mounted scale so that a separate temperature measurement is superfluous.

The accuracy is up to 2×10^{-4} RIU. Light-weight instruments weigh less than 100 g.

For automatically indicating hand refractometers see Sec. 2.3.2.

2.3 Automatic Refractometers

To avoid errors caused by the subjective observation of a measuring mark, automatic refractometers have been designed on the basis of light deviation by a prism (Sec. 1.1) or of the changes in light intensity (Sec. 1.4). The major parts of these arrangements are an illuminating system with a white or monochromatic light source, a space for the interaction between the light and the sample, a detector to receive the light signal, an amplifier, and a display. Many instruments have electrical output ports for a control signal, which can be linked to recording or process-control equipment.

2.3.1 Deviation-Measuring Automatic Refractometers In these instruments, the illumination system is a collimator with an entrance slit. As far as the measurement is concerned with liquids, a prismatic differential flow-through cell is used so that the RI difference from a reference substance results in a deviation of light. The slit-image position on the detector is recognized by the signals of a differential photocell, by the modulated signal of a simple photocell, or by a semiconductor detector.

Most instruments employ a compensation method: The displacement of the slit image is reduced to zero by a rotating glass plate arranged between cell and objective or by a rotating mirror. Modern instruments use a CCD array that makes compensation superfluous, so that all movable parts can be omitted [Fig. 10(a)].

Deviation-measuring automatic refractometers are temperature compensated and yield a high accuracy of 1×10^{-5} RIU. However, there are restrictions as to the flow rate, viscosity, and absorption of the sample.

2.3.2 Intensity-Measuring Refractometers These refractometers use the variation of the intensity of light that is partially or totally reflected at the contact face between the sample and a body of glass. The form of the body is irrelevant; it may be a prism, a rod, a fiber, a spherical lens, or other shape (Fig. 12). For illumination no collimator is necessary; the reflected light is collected on the sensor by means of a simple lens. As the light does not enter the sample, opaque media can also be tested. The electrical signal is amplified and converted into RIU or percentage concentrations and displayed and/or recorded.

Hand refractometers for nonstationary applications use reflection at a prism near the critical angle and give a direct indication of RI or of percentage sugar in water.

The accuracy of intensity-measuring refractometers is less high than that of deviation-measuring instruments, reaching only close to 1×10^{-4} RIU. But they are robust in construction, of small size, and easily handled. They are applicable also for opaque and highly viscous liquids, and flow-through types allow a high flow rate. The contact face must be meticulously clean, and sample temperature must be measured separately.

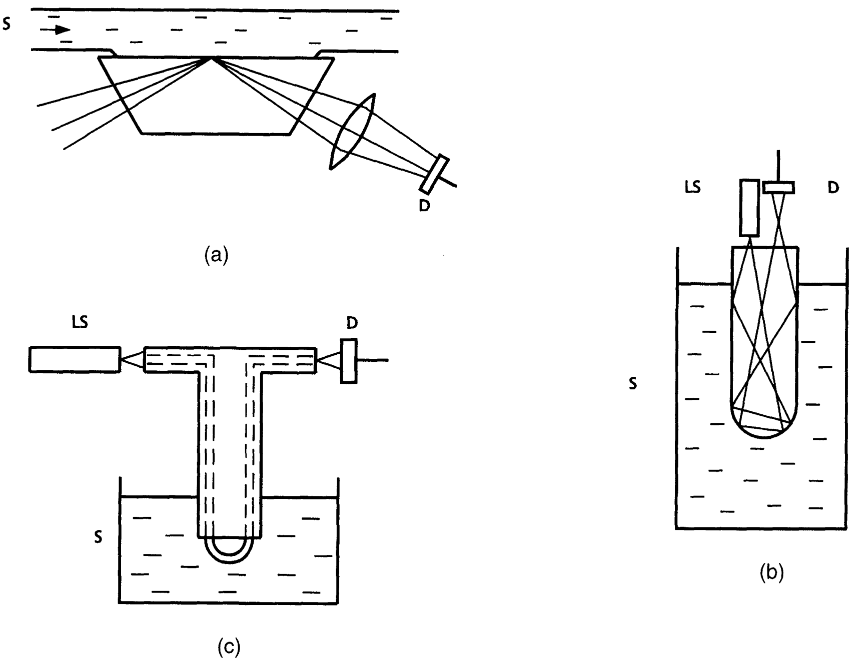

FIG. 12. Measurement of intensity of reflected light. LS: light source; S: sample; D: detector. (a) Prism in contact with a flowing sample, diffuse illumination; (b) a rod dipped in the liquid; (c) the bare zone of a glass fiber dipped in the liquid; the upper region of the fiber has a protective cover.

2.3.3 Integrated-Optics Refractometers

In recent years the development of lasers has brought forth a new branch in optics: integrated optics. Integrated optical devices make it possible to reduce the size of optical instruments and to minimize sample volume. This can be achieved by the principle of total reflection of light at the boundary of a medium with lower RI.

The main parts of an integrated-optics refractometer are a laser, an optical waveguide, and a detector. The laser is usually of the HeNe or the semiconductor type. The radiation is coupled into the waveguide via an optical fiber. The waveguide is a thin metallic layer of a thickness smaller than the wavelength of light. One surface of the layer is in contact with a substrate, e.g., a glass plate or a prism, while the other surface of about 1–2 mm^2 contacts the sample. The radiation of the laser generates waveguide modes with vibration states perpendicular to each other, which vary when the RI of the sample changes.

Different arrangements for the detection of the RI-dependent signals are described, some of them based on interferometric principles. Clerk and Lukosz (1993, 1994) inserted a diffraction grating as an output coupler and measured the outcoupling angles of the modes with position-sensitive detectors. RI differences of about 5×10^{-5} can be recognized.

2.4 Interference Refractometers

The highest accuracy in refractometric measurements is reached by refractometers based on interference of light. For analytical applications two types are commonly employed, which use geometrical or amplitude division of the wave front, respectively.

2.4.1 Rayleigh–Loewe Interferometer

Two coherent beams are generated by a double slit intercepting a parallel bundle of light [Fig. 13(a)]. Each beam traverses a chamber filled with the sample or a reference medium, respectively. The interference fringe pattern in the focal plane of an objective is observed through an eyepiece with a cylindrical lens. Two further beams passing below

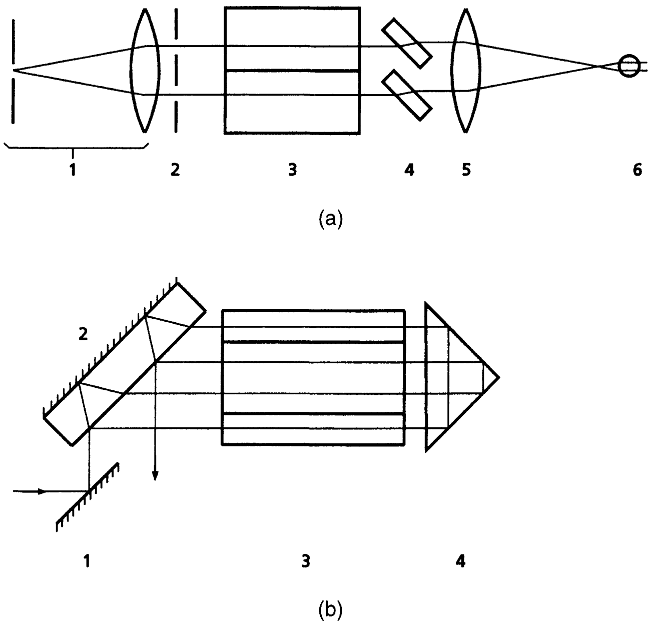

(a)

(b)

FIG. 13. (a) Rayleigh-Loewe interferometer: (1) collimator, (2) double slit, (3) double cell, (4) compensator, (5) objective, (6) eyepiece with cylindrical lens. (b) Jamin-Doi interferometer: (1) plane mirror, (2) glass plate with a semitransparent front face, (3) three-cell system with inner cell for the sample and both outer cells for a reference medium, (4) reflection prism.

the chambers generate a second pattern of interference fringes, which serves as a reference. Media with different RI (n_1,n_2) in the chambers cause a shift of the fringe pattern that, by a rotatable glass plate of a compensation device, is brought back to coincidence with the reference pattern.

White-light illumination allows an unambiguous recognition of the zero-order fringe. Because of the different wavelengths contained in white light, the spaces between the fringes differ in proportion to the wavelength. Therefore, approximate coincidence at all wavelengths exists only for fringes up to the ± first order, producing black half-order minima, while the spacings of higher-order fringes differ increasingly, so that the fringes appear colored.

According to Eq. (12), the accuracy of RI-difference determination is dependent on the length of the chambers. While in portable instruments, which use an autocollimation device, chamber length will not exceed 10 cm, in laboratory instruments it may be up to 1 m or, by internal reflections within the chamber, up to 3 m, so that an accuracy of better than 1×10^{-8} RIU will be reached in gas analyses.

2.4.2 Jamin Interferometer The coherent rays are generated by transmission and reflection of light at a semitransparent layer on a plane-parallel glass plate. After passing the chambers and a compensating device, the rays are reunited at the semitransparent layer of a second glass plate, where they interfere with one another.

Under illumination with white light, a system of interference fringes with two black fringes in the middle is observed. A RI difference of the media in the chambers causes a shift of the fringe pattern, which is either compensated (as in the Rayleigh–Loewe interferometer) or read directly off a scale.

The accuracy is nearly that of Rayleigh–Loewe interferometers.

In portable interferometers of the Jamin–Doi type, a reflecting 90° prism is arranged behind the chambers so that the light twice passes the sample chambers of, e.g., 10-cm length [Fig. 13(b)]. The recombination of the coherent rays takes place at the same layer that splits up the incoming bundle of light. The fringe pattern appears together with a scale where its position is read off directly. By a small displacement of the eyepiece, the center of the fringe pattern is shifted from the zero-order maximum to a half-order minimum, so that *one* small black fringe constitutes the zero mark.

As the chambers are short and a compensation device is omitted, these instruments are small. Moreover, by threefold reflection of light within the chambers, the size of such an instrument can be minimized. The accuracy is about 2×10^{-7} RIU.

2.4.3 Photoelectric Interferometers While the above-mentioned visual interferometers are based on the identification of the zero-order fringe in white light, photoelectric interferometers commonly use a laser source. As this monochromatic source generates an extended fringe pattern without any prominent fringe, the procedure is to vary the optical path length continuously and to determine the shift of the fringe pattern by counting the number of fringes passing a detector. This is made possible by placing the cell filled with the sample into one beam of the interferometer and rotating the cell about a defined angle.

Photoelectric interferometers are based on an amplitude division of the wave front, so that no intensity-limiting entrance slit is required. To have a large lateral separation ("shear") of the interfering beams, interferometers of the Mach–Zehnder or Michelson type are preferred (Hassan *et al.*, 1995).

2.5 Refractive-Index-Gradient–Measuring Refractometers

Inhomogeneous media show RI gradients (*dn*/*dx*), i.e., a variation of RI within the substance.

2.5.1 Schlieren Apparatus For the *qualitative* investigation of RI gradients (schlieren, striae) in a transparent medium a schlieren apparatus is employed. A Toepler arrangement according to Fig. 7 provides a high sensitivity so that very low gradients can be recognized. The visible diameter of the object is limited by the diameter of the objectives. For extended objects one prefers a Z-shaped arrangement with large mirrors instead of the objectives so that the visible field may be up to 300 mm (Fig. 14).

Instead of the knife edge a wire [Fig. 15(b)] that is as broad as the image of the slit may be inserted so that all the light deviated by striae in a direction perpendicular to the slit will produce bright zones on a dark background on the screen.

To give a higher accuracy, the wire may be replaced by a phase plate [Fig. 15(c)], i.e., a glass plate one-half of which is coated with a transparent layer of a thickness of $\lambda/2$, so that interference effects will arise that mark the presence also of very weak striae. In this way, a variation of the phase-contrast technique is achieved (Wolter, 1967).

A rough quantitative determination of striae is possible by illumination with white light and using stripes of transparent filters in different colors [Fig. 15(d)] instead of the knife edge. Thus, the object image shows strongly deflected light in a different color than weakly deflected light; light deflected in the opposite direction appears in still other colors.

For the investigation of stratified solutions, the original Toepler arrangement can easily be completed as described by Philpot

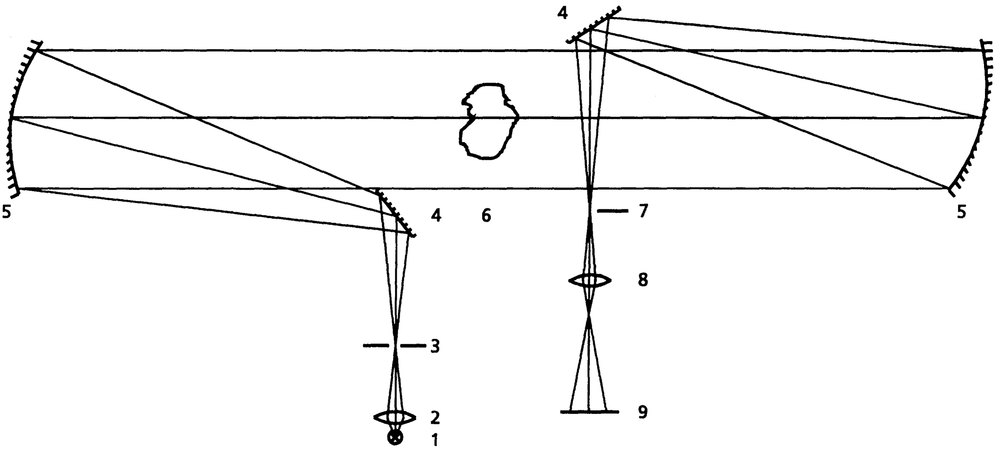

FIG. 14. Z-shaped schlieren arrangement with mirrors. (1) Light source, (2) lens, (3) entrance slit, (4) plane mirrors, (5) concave mirrors, (6) striae (deviation of light is not shown), (7) knife edge, (8) lens system, (9) screen.

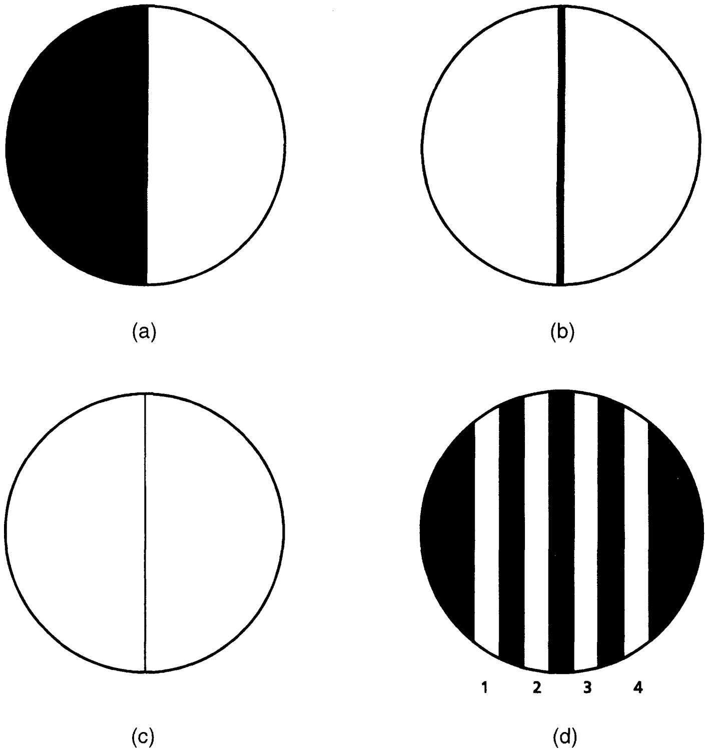

FIG. 15. Some typical sorts of insets in the focal plane of a schlieren arrangement: (a) knife edge, (b) wire, (c) phase edge, (d) colored stripes (1, 2, 3, and 4 indicate different colors).

(1938) and Svensson (1939): The knife edge is replaced by a diagonal slit, and a cylindrical lens projects an image onto the screen, which shows vertically the cell and horizontally the plane of the slit. The deviation of light caused by the gradients results in a lateral displacement of light on the screen, and a gradient curve is obtained.

2.5.2 Interferometers For a *quantitative* determination of disordered inhomogeneities, interferometers are used. Based on Eq. (12), the interference fringe pattern of an amplitude-dividing interferometer (Mach–Zehnder or Michelson type) is changed by schlieren, and the magnitude of the fringe displacement at one point allows the RI gradient to be determined. These instruments are well suited for the exact analysis of very weak inhomogeneities.

High-accuracy measurements of stratified solutions are performed by Rayleigh-type interferometers, which have a grid instead of the entrance slit, and a large vertically extended double cell (Svensson, 1939). The uniform direction of gradients allows interference fringes to be generated, which are shifted according to the magnitude of the RI difference between the sample and a homogeneous reference liquid at the respective height:

$$n_x - n_{o,x} = k_x \lambda / s; \qquad (17)$$

x marks the height in the cell.

The gradient curve is obtained by two beams of light both passing through the sample cell with a very small vertical separation (shear) between them.

The RI and the RI-gradient curve can be imaged simultaneously using a biprism in the interferometer.

3. REFRACTOMETRY OF HOMOGENEOUS SUBSTANCES

3.1 Glasses and Other Solids

Optical glasses are mainly characterized by their RI and their dispersion. The high accuracy required for close inspection is reached only by measurement of the deviation of light at a prismatic object with a spectrometer. Samples must have exactly plane, polished surfaces of several centimeters in diameter. Table 2 presents the RI and Abbe numbers of some transparent optical materials.

For routine measurements a refractometer with a vee-block prism is well suited. The sample only needs to have two unpolished plane faces approximately at right angles to each other, which are put into the vee block with a contact liquid of nearly the same RI. With this arrangement it is also possible to control the process of hardening of adhesives or other synthetics by RI measurement.

To gain comparable values, temperature measurement is necessary. While the temperature coefficients *dn/dt* of glasses are low, i.e., smaller than $\pm 2 \times 10^{-5}$ RIU/K, those of plastics may reach some 10^{-4} RIU/K.

Routine RI measurements of lower precision can be made by critical-angle methods with a Pulfrich, Abbe, or dipping refractometer. The sample must have a plane, polished face, which is brought into contact with the face of the measuring prism, and a contact liquid with a RI higher than that of the sample is needed, e.g., monobromonaphthalene.

The RI of a small piece of a transparent solid may be measured by an immersion method: A sample immersed in a liquid of the same RI is invisible. By immersing the sample into a liquid, changing the RI of the liquid (by substitution, mixing with another liquid, or change of temperature), and observing the disappearance of the contour, it is possible to identify the immersion liquid with exactly the same RI as that of the sample. The RI of that liquid can then be measured separately.

3.2 Solutions and Binary Mixtures of Liquids

Values of RI, Abbe numbers, and temperature coefficients of RI for several liquids and solutions are given in Table 3. The RI of solutions and binary mixtures depends on their concentrations. Refractometry is an excellent method of determining concentration *c* because of the simple and quick procedure. For each solution or mixture there exists a specific *n–c* relation, which must be found by calibration. Calibration data for most commonly used solutions and binary mixtures are compiled in standard tables.

Table 3. Refractive indices, Abbe numbers, and temperature coefficients of some liquids.

Liquid	n_D (20°)	Abbe number	$dn/dt \times 10^4$
Water	1.33299	55.8	−0.9
Methanol	1.3286	59.9	−4
50% in water	1.3424	···	···
Ethanol	1.3613	59.3	−4
50% in water	1.3612	···	···
Ethylene glycol	1.4319	59.8	−3
Glycerol	1.4740	59.0	−2.3
Benzene	1.5014	30.2	
Tetralin	1.5449	32.7	
Carbon disulfide	1.6277	18.4	−8.1
Monobromonaphthalene	1.6582	20.3	−4.5

The most typical example is the analysis of sugar content in water, a test widely applied in agriculture (sugar cane, sugar beet, fruit), wine growing, sugar factories, the production of jams, fruit juices, etc. Precise data of sugar concentration and its dependence on temperature have been established by the International Commission for Unified Methods of Sugar Analysis (ICUMSA, 1974).

The gradient of RI with concentration is called "refractive-index increment" (RI increment):

$$dn/dc = (n - n_o)/c, \qquad (18)$$

where c = concentration in g/mL, n, n_o = RI of solution and solvent. The RI increment has proved a characteristic constant of polymer solutions, dependent only on temperature and wavelength of light (Huglin, 1965; Lorimer, 1972). The values of *dn/dc* lie between −0.24 and 0.25. They are used for the determination of molecular weights of polymers (Baltog *et al.,* 1970).

The calibration curve of most liquid mixtures is bent. For some mixtures it has a maximum (Fig. 16), so that two different concentrations have the same RI. To get unambiguous results, the mixture is diluted 1:1 with one of its components, and the RI of the new mixture is measured.

In organic chemistry and pharmacy the terms "specific refraction" *r* and "molar refraction" *R* have become important for the identification of substances and the investigation of two- or three-component systems.

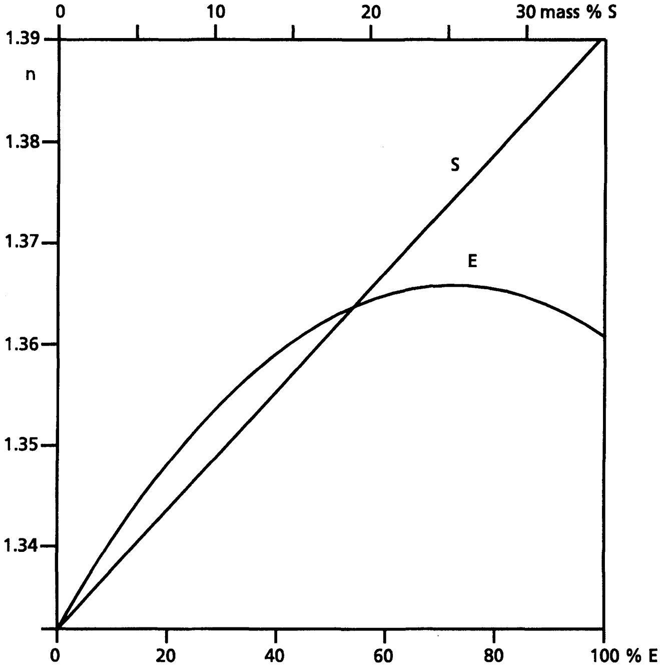

FIG. 16. Calibration curves of sugar in water (S) and of ethanol in water (E).

According to Lorenz and Lorentz the terms

$$r = \frac{n^2 - 1}{n^2 + 2}\frac{1}{d}, \quad R = rM,$$
$$M = \text{molecular weight}, \qquad (19)$$

are specific constants for each substance independent of their state of matter. Besides Eq. (19) there exist some approximations given by other authors; see Joffe (1983).

The dependence of RI of liquids on pressure is negligible, whereas that on temperature is considerable. Comparable results are gained only if the temperature t is thermostatted and noted together with the RI, the same as the wavelength.

Routine laboratory tests are commonly carried out with the Abbe refractometer. Higher precision (nearly 2×10^{-5} RIU) is reached with the dipping refractometer; differential measurements to some 10^{-6} are possible with the Pulfrich refractometer, and to about 1×10^{-7} with an interferometer. With interferometric measurements of liquids, the double cell is immersed in a water bath. Interferometers serve, e.g., for determination of the solubility of gases in water or for observation of a course of titration.

Applications outside the laboratory are widespread—e.g., for the determination of salt contents in ocean water, for testing of antifreeze agents, or for the above-mentioned measurements of sugar contents. The refractometers must be equipped with a thermometer or an automatic temperature-compensation device. They allow a direct reading of the percentage content of sugar or other substances.

3.3 Ternary Mixtures of Liquids

If in a mixture of three components the relation of concentrations c_1, c_2 of two components is a known constant, these components may be regarded as one ($c_{1,2}$) so that the problem is reduced to a binary mixture of $c_{1,2}$ and c_3. Generally, however, concentrations of each of the three components change independently so that a second parameter besides RI must be determined. Preferably this is density d, because it can be measured as easily as RI; this method is known as "refractodensitometry." The measurement of other parameters (e.g., optical activity, viscosity, velocity of sound) is also possible.

The correlations between RI, density, and concentrations c_i must be found experimentally, and calibration tables must be erected. All measurements must be made at a constant temperature.

A typical and important example is the analysis of beer. In the brewing process, part of the extract changes into alcohol, so that in the ternary system of alcohol, extract, and water the concentrations as well as the original extract are to be determined. Similar examples are the testing of wine, liqueur, and spirits.

Commonly used instruments are a dipping refractometer, a densitometer, and a thermostat. Brewery processes can be monitored with an automatic flow-through refractometer and an automatic densitometer, both equipped with temperature-controlling devices.

3.4 Continuously Varying Liquids

The development of photoelectric techniques has opened up new and highly useful applications of refractometry. This applies especially to the continuous measurement of flowing liquids or of liquids that change with time. One of the most important applications is the monitoring and control of processes in industrial production plants. Automatic process refractometers (Sec. 2.3) are well suited for the indication and control of concentration in the food and beverage industries (sugar content), of fractionating and synthesizing processes in the petrochemical industries, of spinning and cleaning processes in the textile industry, of distillation and extraction processes, etc. These instruments supply electric signals that actuate control valves to correct out-of-tolerance concentrations in the product.

Process control is the most important field of application of refractometers today.

Automatic refractometers with extremely small cells have been designed for application as detectors in liquid chromatography with columns. The necessary high accuracy of better than 1×10^{-5} RIU requires differential refractometers with a prismatic double cell. Some types of RI detectors use an autocollimation device so that the light passes the differential cell twice to gain a higher accuracy. Though in HPLC (high-performance liquid chromatography) generally a UV detector is preferred, a RI detector must be taken if the eluate and the solvent show no UV absorption.

In immunology, the amount of sample available (e.g., body fluids) is extremely small. The refractive index of a sample, which changes, e.g., in the course of reaction between antigens and antibodies, is determined with refractometers built around integrated-optics devices (Sec. 2.3.3). It is also possible with this method to trace chemical reactions in minute substance volumes.

3.5 Binary Mixtures of Gases

Refractometry is a useful tool for controlling the purity of gases or determining their concentrations by measuring RI, or the RI difference from a reference gas. The RI of gases are very low, smaller than 1.003, so that only interference refractometers are eligible. For simplification one sets $m = (n - 1) \times 10^6$ (Table 4). The RI of gases and vapors are strongly dependent on temperature and pressure. Therefore, care must be taken that these parameters are identical for sample and reference medium.

Published tables list the RI of gases for "standard conditions," i.e., temperature $t_0 = 0$ °C and pressure $p_0 = 101.325$ kPa (= 760 Torr). These are calculated from the param-

Table 4. Refractive indices of gases and vapors at standard conditions (Sec. 3.5).

Gas/Vapor	$m_e = (n_e - 1) \times 10^6$	$m = m_F - m_C$
Air, dry and CO_2 free	293.11	3.27
Helium	34.95	0.18
Neon	67.25	0.28
Hydrogen	139.37	1.81
Oxygen	272.27	3.83
Nitrogen	299.14	3.33
Carbon dioxide	450.6	5.79
Chlorine	784.0	13.9
Carbon disulfide	1477	70
Methane	443.3	6.7
Methanol	563.0	8.5
Ethanol	870	10.8
Chloroform	1412	19.9
Halothane	1582	23.8
Benzene	1759	43
Carbon tetrachloride	1781.9	30.6
Toluene	2092	...

eters of measurement, t and p, by

$$m(t_0,p_0) = Fm(t,p), \tag{20}$$

where $F = (1 + t/273)\, p_0/p$. With this, Eq. (12) yields

$$n_2(t_0,p_0) - n_1(t_0,p_0) = Fk\lambda/s. \tag{21}$$

The great dependence of the RI of air on pressure and temperature makes it possible to monitor the condition of ambient air by interferometry, which is important, e.g., in high-precision metrology.

Concentration measurements require calibration. Most mixtures show a straight calibration curve. But through the effect of dispersion (Sec. 2.4.1) the colors within the interference fringes change with increasing concentration so that the zero-order fringe becomes colored while that of a higher order becomes black. In the calibration curve this change of colors results in "jumps" by exactly one interference fringe spacing (Fig. 17).

The accuracy of a laboratory interferometer with a chamber of 1-m length is about 2 × 10^{-8} RIU for differential measurement; this corresponds, for example, to 0.01% CH_4 in air or to 0.001% benzene in air.

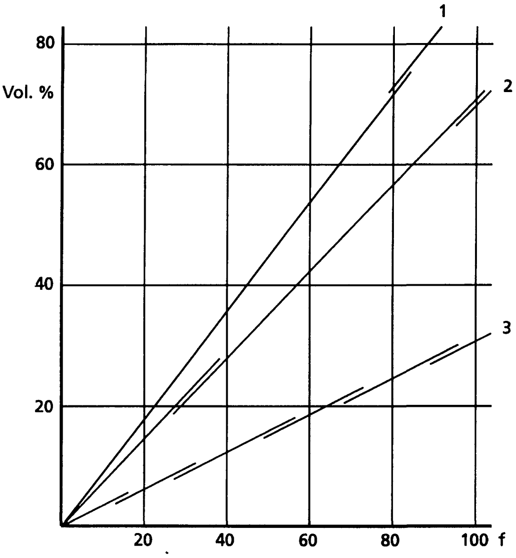

FIG. 17. Calibration curves for interferometric determination of (1) hydrogen, (2) carbon dioxide, (3) acetylene in air; Rayleigh-Loewe interferometer with cell length 1 m. f is the measured number of interference fringes.

Applications of interferometric analysis of gaseous mixtures are many and varied; examples are the calibration and testing of anesthetic vaporizers, testing of lung function, and the monitoring of air in rooms with explosion hazards. An important field of application is the monitoring of air in underground coal mines to prevent fire-damp explosions. For this purpose the small Jamin–Doi type interferometer, equipped with cells for the absorption of water vapor and CO_2 from air and with its scale divided in percentage CH_4, is used worldwide. A similar task is the testing of air in factories where organic solvents are used for production or cleaning. Most solvent vapors can generate explosive mixtures in air or are harmful to health.

The accuracy of portable interferometers is approximately 0.1% CH_4 or 0.01% benzene in air.

3.6 Ternary Mixtures of Gases

The methods employed for the investigation of ternary mixtures of gases are similar to those used for liquids (Sec. 3.3). If the relative concentration of two components is a constant these may be regarded as one gas. The problem is thus reduced to a binary system. A typical example is determination of CO_2 (or another gas) in air, as air is a mixture of essentially 79% N_2 and 21% O_2. If, however, the composition of air changes, the procedure is to measure the RI of the original sample first, and then that of the sample after absorption of CO_2, both relative to normal air. This method is applied in investigations of respiration in medical and biological tests, also in animal stables. Useful sorption media are listed by Nebe (1977).

4. REFRACTOMETRY OF INHOMOGENEOUS SUBSTANCES

Investigations of inhomogeneous substances are widespread in technical, medical, environmental and other fields. Depending on whether a more vivid visualization or an exact quantitative analysis is required, a schlieren apparatus or an interferometer is preferred. Cited below are only the most important investigations.

4.1 Transparent Solids

In glass factories and in the optical industry, a close inspection of the homogeneity of glasses is necessary. Also, optical surfaces can be tested for flatness.

In the same way, plastics are inspected. As the temperature coefficient of RI of plastics is high, heat-transfer processes can be observed. Moreover, plastics subjected to stress show effects of double refraction (photoelasticity). If such an object is brought between crossed polarization filters the stress distribution is visible. The photoelastic study of models made of transparent plastics, above all PMMA, is an important method of forecasting the behavior of products under stress in use.

4.2 Schlieren in Liquids

Within liquids striae are examined in many fields of research. Some examples are solvation processes, electrochemical processes in electrolytes, characterization of polymeric materials by measuring disperse solutions, heat current and heat transfer in liquids, and ultrasonic effects in the bulk and at the surface of liquids.

RI gradients in liquids are generally large. Therefore, interferometers are difficult to use. Instead, Toepler schlieren arrangements are preferred. Their precision may be varied by changing the focal length of the imaging objective.

4.3 Stratified Solutions

For the investigation of stratified liquids, refractometric methods have proved very useful. The optical arrangements are based on the Toepler schlieren device, with special adaptions, or on interferometric devices (Sec. 2.5). The most important measurements are those of electrophoretic, diffusion, and sedimentation processes (Weissberger, 1972).

Electrophoretic investigations are mainly aimed at the identification of colloidal substances, particularly proteins, which differ only by their rate of migration in an electric field. An important application in medical diagnostics is the differentiation of serum proteins.

Examination of diffusion and measurement of the diffusion coefficient are a useful tool for the characterization of macromolecules in chemical research. As solutions to be measured are highly diluted, very precise methods of detection, preferably interferometric methods, are applied.

Sedimentation processes are mainly observed in a high-speed ultracentrifuge because a very strong gravitational field is a prerequisite. Interferometric techniques based on Rayleigh–Svensson or Jamin-type interferometers are employed. From the measured rate of sedimentation, the sedimentation coefficient is calculated, which in combination with the diffusion coefficient allows the determination of molecular weight.

4.4 Inhomogeneous Gases

Striae in gases generally have no preferred direction of RI gradient. Very weak striae may be determined interferometrically. Strong striae, however, disturb the pattern of interference fringes, and therefore, the use of the Toepler schlieren arrangement is preferred.

The most important applications are aerodynamic investigations of all sorts of quickly moving objects by a simulation procedure in a wind tunnel; investigation of shock waves and pressure waves (Fig. 18); investigation of flames, combustions, and explosions; obser-

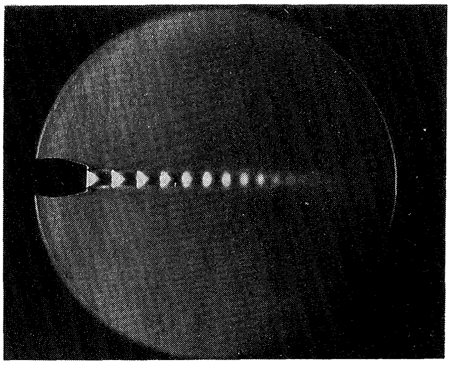

FIG. 18. Pressure waves in air escaping from a nozzle. Photo taken with a Toepler schlieren arrangement.

vation of electrical discharges and plasma research (electron concentration in plasma); heat exchange processes; and surface effects in vapor-liquid two-phase processes.

GLOSSARY

Abbe Number: Dispersive power; an expression characterizing the dispersion of a medium.

Amici Prism: See **Direct-Vision Prism.**

Beam Splitter: Glass plate with a semitransparent face that splits an incident beam of light into a refracted and a reflected one.

Birefringence: See **Double Refraction.**

Coherence: Coincidence of two electromagnetic waves in certain parameters; a condition for the interference of light beams.

Collimator: A device for the generation of a bundle of parallel light beams by illuminating a pinhole (or slit) in the focal plane of an objective.

Critical Angle of Refraction: Largest angle of refraction at the face of two substances.

Critical Angle of Total Reflection: Smallest angle of incidence for which total reflection at the face of two substances occurs.

Critical-Angle Refractometer: Refractometer based on the measurement of the critical angle of refraction/reflection.

Direct-Vision Prism: Combination of two or more prisms so that light of a selected wavelength will leave the prism undeviated.

Dispersion: $dn/d\lambda$, dependence of RI on the wavelength of light.

Double Refraction: A light beam is split into two beams; one (ordinary ray) is refracted independently of its angle to the optical axis of the medium, while the refraction of the other (extraordinary ray) depends on this angle.

Integrated-Optics Refractometer: Refractometer containing integrated-optics constituents.

Interference Fringe Pattern: A system of interference fringes generated by an interferometer.

Interference of Light: Superposition of two (or more) light waves so that they may either attenuate or amplify each other; also extinction may occur.

Refractodensitometry: Analytical method based on the measurement of RI and density.

RI = Refractive Index: Quotient of velocity of light in a vacuum and that in a medium.

RI Gradient: dn/dx; local change of RI within a medium.

RI Increment: dn/dc; change of RI of a solution with the concentration of the solute.

Schlieren: Striae, streaks; inhomogeneities (RI gradients) within a medium, caused by differences of temperature, pressure, or concentration.

Schlieren Apparatus: Instrument for recognition of striae.

Total Reflection: Reflection of light at a surface without loss of intensity.

Works Cited

Baltog, I., Ghita, C., Ghita, L. (1970), *Eur. Polym. J.* **6,** 1299–1303.

Clerk, D., Lukosz, W. (1993), *Sensors Actuators B* **11,** 461–465.

Clerk, D., Lukosz, W. (1994), *Sensors Actuators B* **18–19,** 581–586.

Hassan, G. E., El-Kashef, H., El-Baradie, B. Y., El-Labban, M. (1995), *Rev. Sci. Instrum.* **66,** 38–42.

Huglin, M. B. (1965), *J. Appl. Polym. Sci.* **9,** 3963–4024.

ICUMSA (1974), *Refractive Index,* ICUMSA Report, Subject 12, 144–156.

Joffe, B. W. (1983), *Refractometric Methods of Chemistry* (Russian), 3rd ed., Leningrad: Chimia.

Lorimer, J. W. (1972), *Polymer* **13,** 46–56 and 274–276.

Nebe, W. (1977), "Analytical Applications of Interferometry," in: G. Svehla (Ed.), *Comprehensive Analytical Chemistry,* Vol. 8, Amsterdam: Elsevier, pp. 391–546.

Philpot, J. St. L. (1938), *Nature* **141,** 283.

Schardin, H. (1942), "Die Schlierenverfahren und ihre Anwendungen," *Ergebn. Exakt. Naturwiss.* **20,** 303–439.

Svensson, H. (1939), *Kolloid Z.* **87,** 181; **90,** 141.

Toepler, A. (1864), *Beobachtungen nach einer neuen optischen Methode,* Bonn: Cohen.

Weissberger, A. (Ed.) (1972), *Techniques of Chemistry,* Vol. 1, *Physical Methods of Chemistry,* New York: Wiley Interscience.

Wolter, H. (1967), *"Schlieren-, Phasenkontrast- und Lichtschnittverfahren,"* in: S. Flügge (Ed.), *Handbuch der Physik,* Vol. 24, Berlin: Springer, pp. 555–645.

Further Reading

Born, M., Wolf, E. (1991), *Principles of Optics,* Oxford, U.K.: Pergamon.

Creath, K. (1988), "Phase-Measurement Interferometry Techniques," in: E. Wolf (Ed.), *Progress in Optics,* Vol. 26, pp. 349–393.

International Conference on Refractometry (1994), Warszawa, Abstracts; Polish Chapter of the International Society for Optical Engineering (SPIE).

Jenkins, F. A., White, H. E. (1976), *Fundamentals of Optics,* 4th ed., New York: McGraw-Hill, 746 pages.

Malacara, D. (Ed.) (1978, 1991), *Optical Shop Testing,* New York: Wiley, 523 pages.

Nebe, W. (1970), *Analytische Interferometrie,* Leipzig: Geest and Portig.

Schmitz, E. H. (1981–1993), *Handbuch zur Geschichte der Optik,* Vols. 1 to 5, Bonn and Ostende: Wayenborgh.

Weissberger, A. (Ed.) (1972), *Techniques of Chemistry,* Vol. 1, *Physical Methods of Chemistry,* New York: Wiley Interscience.

REFRACTORY MATERIALS—CERAMICS

GERALD ROUTSCHKA, *DIFK—Deutsches Institut für Feuerfest und Keramik GmbH (formerly Forschungsinstitut der Feuerfest-Industrie), Bonn, Federal Republic of Germany*

KARL-ERNST GRANITZKI, *Königswinter, Federal Republic of Germany (Sec. 5.2)*

Introduction 277
Definitions 277
1. **General Classification and Nomenclature** 278
2. **Production** 278
3. **General Composition, Properties, and Testing** 279
3.1 Porosity 280
3.2 Bulk Density, Density 280
3.3 Cold Mechanical Properties ... 281
3.4 Thermal Expansion 281
3.5 Reheat Change 281
3.6 Refractoriness under Load (R.U.L.) 281
3.7 Creep in Compression 282
3.8 Thermal Conductivity, Specific Heat, and Emission 283
3.9 Hot Mechanical Properties 283
3.10 Elasticity and Stress–Strain Behavior 283
3.11 Thermal-Shock Resistance 285
3.12 Electrical Resistance 286
3.13 Corrosion Resistance 286
4. **Types of Refractory Ceramics** 286
4.1 Dense, Shaped Refractory Products 286
4.1.1 Silica Bricks and Fused-Silica Ceramics 286
4.1.2 Fireclay and High-Alumina Bricks 288
4.1.2.1 Fireclay Bricks 288
4.1.2.2 High-Alumina Bricks 289
4.1.3 Basic Bricks 290
4.1.3.1 Magnesia Bricks 290
4.1.3.2 Doloma Bricks 291
4.1.3.3 Basic Bricks Containing Chrome 291
4.1.3.4 Spinel-Containing Bricks 292
4.1.3.5 Forsterite Bricks 292
4.1.4 Zircon- and Zirconia-Containing Bricks 292
4.1.5 Carbon and Graphite Bricks .. 293
4.1.6 Carbon- and Graphite-Containing Refractories 294
4.1.6.1 Fired Graphite-Containing Refractories 294
4.1.6.2 Pitch- and Resin-Bonded Unfired Refractories 294
4.1.7 Silicon Carbide Bricks 295
4.1.8 Fine-Grained Oxide and Nonoxide Ceramics 296
4.1.9 Fusion-Cast Bricks 297
4.1.10 Refractory Ceramics with Low Thermal Expansion 297
4.2 Unshaped Refractory Materials 297
4.2.1 Ramming Mixes 299
4.2.2 Plastic Mixes 299
4.2.3 Refractory Castables 299
4.2.4 Gunning Materials 301
4.2.5 Prefabricated Shapes 301
4.2.6 Jointing and Coating Materials 301
4.3 Heat-Insulating Ceramic Materials 301
4.3.1 Insulating Bricks 302
4.3.2 Insulating Refractory Bricks .. 302
4.3.3 Refractory Ceramic Fibers 302
5. **Selection and Use** 304
Glossary 304
Works Cited 306
Further Reading 307

INTRODUCTION

Definitions

According to ISO/R 836, refractory materials are nonmetallic materials or products (including those containing a proportion of metals) having a minimum pyrometric-cone equivalent of 1500 °C. Determination of the pyrometric-cone equivalent (refractoriness) is described in ISO/R 528. Test pieces of the refractory material in the form of small cones are raised in temperature alongside reference

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

pyrometric cones (blunt-tipped, skewed, triangular pyramids) with a defined temperature of collapse (softening).

Certain products from the refractory industry do not comply with this definition, however, but still fall under refractory classifications and are therefore conventionally termed refractories.

In the common sense, refractories are ceramic construction materials capable of withstanding temperatures above 800 °C. Numerous refractory compositions in a wide variety of shapes and forms are used in a broad range of industrial applications. The three main groups of refractory ceramics are shaped refractory products (bricks), unshaped refractory products (monolithics), and ceramic-fiber products. Their three main applications are the construction of reaction chambers (e.g., furnaces), thermal insulation, and recycling of heat in regenerators or recuperators. In some cases refractory ceramics are used as functional or mechanical parts of hot-working aggregates, especially for metal melt processing, like sliding gates and filters.

The temperature of use for "hot face" linings ranges from ca. 800 to 1800 °C, and sometimes to 2000 °C or higher.

In addition to high temperature resistance, other essential properties for refractories are resistance to temperature change (thermal shock) and to mechanical and corrosive influences.

1. GENERAL CLASSIFICATION AND NOMENCLATURE

Classifications of refractory products generally make reference only to characteristic limits for specific property values. The nomenclature and forms of designation are derived from features such as raw material base, manufacturing process, nature of bond, and typical application. The main classification criteria are the following:

1. Chemical and/or mineralogical constitution (see Table 1).
2. Total porosity: dense refractory products, <45 vol%; insulating refractory products, ≥45 vol%.
3. Form of delivery: shaped products (bricks and shapes), unshaped products (monolithics), or fibrous products (ceramic-fiber products).
4. Type of bonding: bonding at <150 °C, unfired products—ceramic bond (clay bond), inorganic chemical bond (e.g., phosphate), hydraulic bond, or organic bond (e.g., pitch, resin); bonding at 150 to 800 °C, thermally treated (tempered) products—inorganic chemical bond, hydraulic bond, organic bond (e.g., carbon bond); bonding at >800 °C—firing (ceramic bond), fusion and solidification.
5. Application field.

2. PRODUCTION

The refractories industry consumes a wide variety of raw materials, binders, and special additives (*Am. Ceram. Soc. Bull.*, 1992; Coope and Dickson, 1991).

Most of the raw materials are based on naturally occurring beneficated, upgraded, crushed minerals, usually acquired in captive production by the refractories manufacturer. Several high-purity synthetic raw materials have become important in recent years because of the increasing demand for high-performance refractories. Synthetic raw materials are produced by special suppliers.

Various methods are employed to produce refractory ceramics of different types:

1. refractory products with large grain sizes, normally ≤6 mm, in special cases ≤20 mm;
2. special fine-grained refractory products (fine ceramics);
3. products formed via a melt (i.e., fused cast shapes);
4. ceramic-fiber products.

The production process includes the following steps: preparation of mixes, shaping, drying, thermal treatment up to 800 °C, or firing, aftertreatment (if necessary), and packing.

During firing the characteristic ceramic microstructure develops as a result of reactions, recrystallization, and transformation or partial formation of a liquid phase or melt, respectively (at room temperature this phase consolidates to a glassy phase).

The main groups of refractory ceramics roughly share in total production as follows:

Table 1. Classification of refractory products (ISO 1109–1975).

Products	Limiting content of principal constituent	Criteria of subdivision and general observations
High-alumina products Group 1	$Al_2O_3 \geq 56\%$	A complete designation of these products shall include an indication of either the raw material actually used or the mineralogical constitution of the final product. In the latter case, the method of ascertaining this constitution shall be stated
High-alumina products Group 2	$45\% \leq Al_2O_3 < 56\%$	
Fireclay products	$30\% \leq Al_2O_3 < 45\%$	
Low-alumina fireclay products[a]	$10\% \leq Al_2O_3 < 30\%$ $SiO_2 < 85\%$	
Siliceous products[a]	$85\% \leq SiO_2 < 93\%$	
Silica products	$SiO_2 \geq 93\%$	Quality specifications according to use
Basic products		In view of recent and possible future developments in basic products, new subdivisions and new criteria of classification may become necessary
Magnesite (magnesia)	$MgO \geq 80\%$	Products in which the principal constituent is magnesite (magnesia)
Magnesite–chrome (magnesia chrome)	$55\% \leq MgO < 80\%$	Products in which the principal constituents are magnesite (magnesia) and chromite
Chrome–magnesite (chrome magnesia)	$25\% \leq MgO < 55\%$	Products in which the principal constituents are chromite and magnesite (magnesia)
Chromite	$Cr_2O_3 \geq 25\%$ $MgO \leq 25\%$	Products in which the principal constituent is chromite
Forsterite		Products in which the principal constituent is forsterite
Dolomite (doloma)		Products in which the principal constituent is dolomite (doloma)
Special products		Products based on carbon, graphite, zircon, zirconia, silicon carbide, carbides (other than silicon carbide), nitrides, borides, spinels (other than chromite) Products based on several oxides (other than those of basic products) Products based on pure oxides, including alumina, silica, magnesia, zirconia; products of high purity

[a] Products designated in certain countries by the name "semi-silica products" may belong either to the low-alumina fireclay class or to the siliceous products class.

silica bricks, 1%–2%; fireclay bricks, 15%; high-alumina bricks, 12%; basic bricks, 20%–30%; other bricks, 6%; and unshaped refractories, 35%–50% (mainly castables and gunning mixes).

3. GENERAL COMPOSITION, PROPERTIES, AND TESTING

The properties of refractory products are determined primarily by their chemical compositions, mineral compositions, and structure. Structure here refers to the type, quantity, size, form, and arrangement of phases in the substance involved. Refractory ceramics generally comprise three structural elements: coarse grain, binding matrix, and pores. The binding matrix is of special importance with respect to refractoriness and corrosion resistance. It is usually associated with a higher degree of porosity than the grains, which have well-defined properties.

Fired bricks possess fixed, measurable properties. By contrast, unfired and unshaped ceramic products acquire their properties through use, in the course of which a temperature gradient is usually established. This temperature gradient gives rise to considerable property gradients between the hot face (sintered at the high-temperature end) and what is often a loose mass at the low-temperature face.

Properties are ascertained by applying various tests, most of which are standardized (Annual Book of ASTM Standards, 1995; *Refractory Materials, Recommendations,* 1990). Values depend on the test method employed and, in the case of unshaped products, also on the method of preparation of test pieces. Even standardized test methods are of limited value because of problems with reproducibility and comparability, a problem especially prevalent with destructive testing.

Table 2 summarizes important property tests conducted on refractory materials.

3.1 Porosity

Porosity has a considerable effect on properties (Fig. 1). Three types of porosity can be distinguished: total, open (apparent), and closed porosity. Open porosity is normally used to characterize porosity of refractory ceramics. The amount of closed pores in refractory ceramics with total porosity exceeding 15% is small (ca. <3%.)

Relationships between open porosity and other properties can be established most accurately by determining pore-size distribution (Carswell, 1977) and gas permeability.

3.2 Bulk Density, Density

Bulk density characterizes the weight of a product per unit volume, including all pores. It is needed to calculate porosity, refractory requirement, weight of a refractory lining, stored heat, etc. Bulk density is influenced by the material density (weight per unit volume, excluding pores), which is a constant for each material, and total porosity. Densities for refractory ceramics range between 2.2 g/cm^3 (fused silica) and 5.9 g/cm^3 (ZrO_2).

Table 2. Tested characteristics of refractory ceramics.

Chemical, mineralogical	Texture	Cold mechanical	Hot mechanical	Thermophysical	Thermochemical
Chemical analysis[a]	True density	Cold compressive strength[a]	Hot compressive strength	Thermal expansion[b]	Refractoriness[b] (pyrometric-cone equivalent)
Microprobe analysis	Bulk density[a]	Cold modulus of rupture[b]	Hot modulus of rupture[b]	Thermal conductivity[b]	Corrosion resistance (slag, melts, vapors)
Microscopy	Porosity[a]	Abrasion resistance		Specific heat	CO resistance
X-ray diffraction	Permeability[b] to gas (air)			Electrical resistance	Oxidation resistance
	Pore-size distribution	Modulus of elasticity (stress–strain behavior)			Hydration resistance Acid resistance
	Nondestructive testing (mainly ultrasonic)		Change in dimensions on heating[b]		Bubble formation (only for glass-melt–contacting materials)
			Refractoriness under load[b]		
			Creep in compression		
			Thermal-shock resistance[b]		

Special and additional tests and test conditions for			
Unshaped refractories	Joining and gunning mixtures	Carbon-containing refractories	Ceramic fibers[d]
Grain-size distribution	Bonding or adhesive strength	Tests in neutral or reducing atmosphere,[c] special chemical analysis methods, coking behavior, residual carbon content[b]	Compressibility
Consistency, workability			Shot content (nonfibrous particles)
Moisture content[b]			Tensile strength

---

[a] For statistical quality control.
[b] Additional information for data sheets.
[c] Le Doussal and Prieur, 1989.
[d] Dietrichs, 1987.

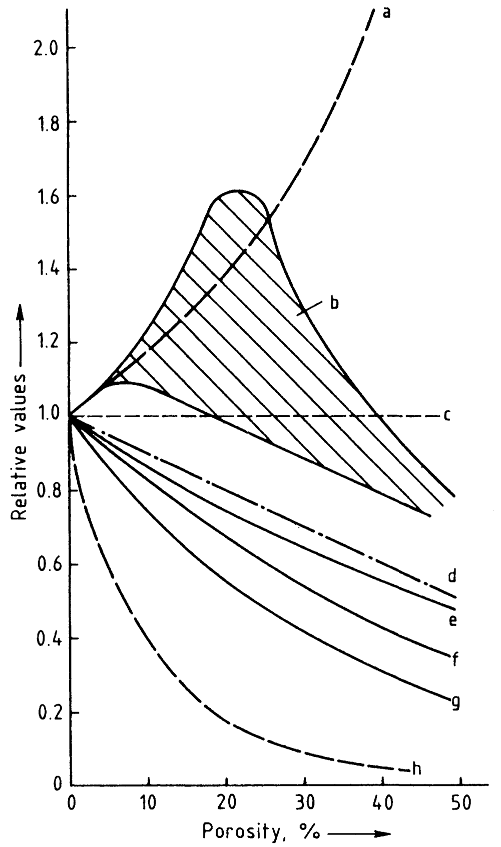

FIG. 1. Porosity effects on the properties of refractory ceramics: Curve a, corrosion; curve b, thermal-shock resistance; curve c, thermal expansion; curve d, bulk density; curve e, strength (fine ceramics); curve f, thermal conductivity; curve g, modulus of elasticity; curve h, cold compressive strength (coarse refractories).

3.3 Cold Mechanical Properties

Cold crushing and modulus of rupture (bending) tests provide information about the structure and structural uniformity of a ceramic and assist in statistical quality control.

3.4 Thermal Expansion

Thermal expansion is a material constant determined primarily by the basic component of the ceramic (e.g., SiO_2, MgO). With the exception of silica- and zirconium-oxide–containing products, fired ceramic materials have smooth thermal-expansion curves (Fig. 2). During initial heating, unfired bricks and unshaped refractories show overlap between reversible thermal expansion and irreversible expansion/shrinkage effects, which are influenced even by small loads. Reversible and permanent changes in length are of importance for construction of a refractory lining, e.g., when measuring joints. Because thermal expansion cannot be suppressed even with high loads, some fraction (usually 0.5–0.8) of the laboratory expansion value is used for dimensioning joint thickness.

3.5 Reheat Change

Heating to high temperatures and subsequent cooling often causes a permanent, time-dependent change in dimensions. If a refractory lining shows substantial shrinkage, joints or cracks may form. The opposite case, after-expansion, is also dangerous, because the resulting pressure may destroy the refractory lining. Dense bricks should not show more than 1%–1.5% linear shrinkage at maximum service temperature.

The classification-temperature criterion for insulating and unshaped refractory products is limited linear heat shrinkage at a defined temperature and holding time (5–24 h). The following shrinkage limits apply: insulating fire bricks, max. 2%; ceramic-fiber products, max. 4%; castables, max. 2%; ramming mixes, max. 2% drying and firing shrinkage; and moldable mixes, max. 3% drying and firing shrinkage.

3.6 Refractoriness under Load (R.U.L.)

Determination of the refractoriness under load (differential, with rising temperature in oxidizing atmosphere; ISO 1893) is a characteristic test for refractory products. The sample is normally a cylinder (height 50 mm, diameter 50 mm) with an internal axial bore (diameter 12.5 mm). This is heated in air under a load of 0.2 N/mm^2 at a rate of 5 °C/min until it becomes soft. Characteristic temperatures (e.g., T_{05}, the temperature at which the sample is compressed by 0.5%) are read from a height-change–temperature curve (Fig. 3). Some products display a broad softening range (e.g., fireclay bricks), while others show a sudden collapse (e.g., silica and magnesia bricks).

The precursor of ISO 1893, described for

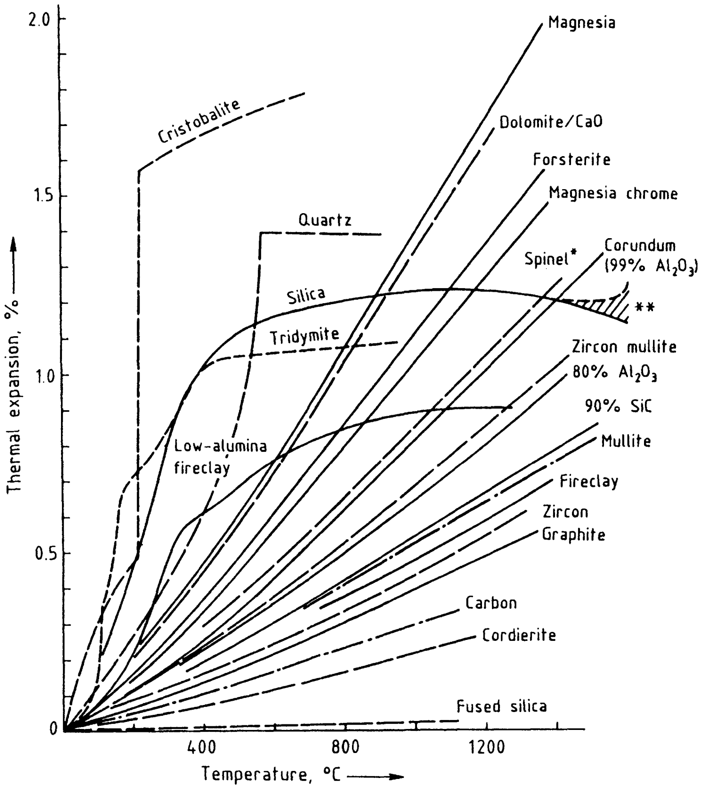

FIG. 2. Thermal-expansion curves of fired refractory materials (VDI-Richtlinie 3128, Blatt 4, 1977). *$MgO \cdot Al_2O_3$, $MgO \cdot Cr_2O_3$, chrome ore. **After expansion when residual quartz is present. For thermal expansion of ZrO_2 ceramics see Fig. 8.

instance, in DIN 51064, is still used to some extent. The test is conducted in a reducing atmosphere and the measured expansion includes the expansion of the loading columns (carbon). The characteristic temperature is t_a, the temperature at which the sample is compressed by 0.6%. Generally higher values than ISO 1893 are found, especially with materials that are sensitive to redox conditions (e.g., iron-rich magnesia).

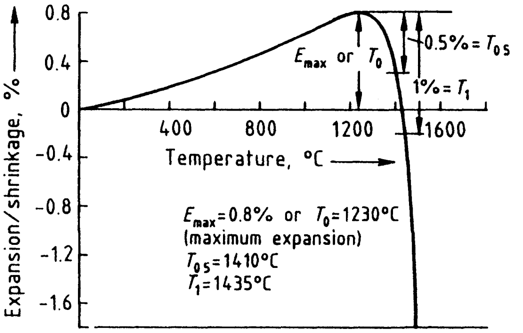

FIG. 3. Refractoriness under load curve (example according to DIN 51053, Part 1).

Softening behavior under pressure is influenced by the type of refractory material, its mineralogical and pore structure, and the quantity and the viscosity of the melts formed at testing temperature. Softening behavior improves with increasing firing temperature of the refractory material.

3.7 Creep in Compression

Refractory products are often exposed to high temperature and pressure for extended periods of time (e.g., the checkerwork of regenerators). This results in slow shrinkage and creep processes that influence refractory-lining lifetimes. Such behavior is characterized by the creep test (ISO 3187), which is performed analogous to ISO 1893 but at constant temperature and pressure over 25 h or longer (Fig. 4).

Creep over longer periods can be estimated by extrapolation. For rough calcula-

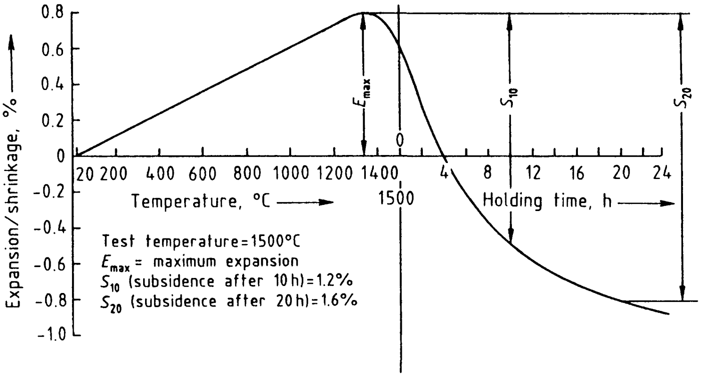

FIG. 4. Creep curve (example according to DIN 51053, Part 2).

tion, the creep increases proportional to the applied load (Konopicky *et al.*, 1971).

3.8 Thermal Conductivity, Specific Heat, and Emission

Thermal-conductivity values are used primarily to calculate the temperature distribution in a refractory wall and the heat loss of a furnace. In some cases products with high conductivity are used to keep the temperature of the hot side of the lining low, thereby reducing wear. The thermal conductivity of a refractory material depends on its chemical and mineralogical composition, and especially on its porosity and pore-size distribution. The atmosphere is also of great importance, most notably in the case of porous products (Young *et al.*, 1964).

There are two standard methods for measuring thermal conductivity: the hot-wire test (ISO 8894) and the calorimeter method (ASTM C 201). Values obtained by the two methods may differ by up to 25%. The method used should therefore be specified whenever thermal-conductivity values are given. This is particularly important for unshaped products; in such cases, special attention must also be given to pretreatment of the sample (Routschka, 1988). The thermal conductivity of refractory products ranges from ca. 0.05 (ceramic fibers) to 100 W m^{-1} K^{-1} (graphite) (Fig. 5).

Specific heat is used to calculate the heat stored by a lining, and for determining the temperature diffusivity that results from thermal conductivity in order to calculate nonstationary temperature fields in, for example, regenerators. The specific heat for conventional refractory products is ca. 1 kJ kg^{-1} K^{-1} (VDI-Richtlinie 3128, 1977).

The thermal emission of a refractory construction material depends on the type of material and the nature of the surface. Refractory products can be regarded as graybody radiators, with emissions 50%–85% as great as blackbody radiators (Glazman *et al.*, 1989).

3.9 Hot Mechanical Properties

Thermomechanical behavior is important in the application of a refractory product, and it can be characterized by high-temperature strength testing. Hot strength is a complex parameter that depends on the structure, quantity, and viscosity of the formed melts at high temperature. The temperature dependence of thermal behavior is a function of the type of loading involved (e.g., bending, compression, tension, or torsion).

One common measure of the thermomechanical character is the hot modulus of rupture test (ISO 5013), which is frequently incorporated into the quality control of refractory products. Examples of the hot modulus of rupture of different brick types are shown in Fig. 6.

3.10 Elasticity and Stress–Strain Behavior

A modulus of elasticity value is determined in order to estimate the critical stress on a refractory material brought about by mechanical and thermal loading. The static modulus of elasticity is established by measuring deformation under compression, bending, and tension. For simplicity, dy-

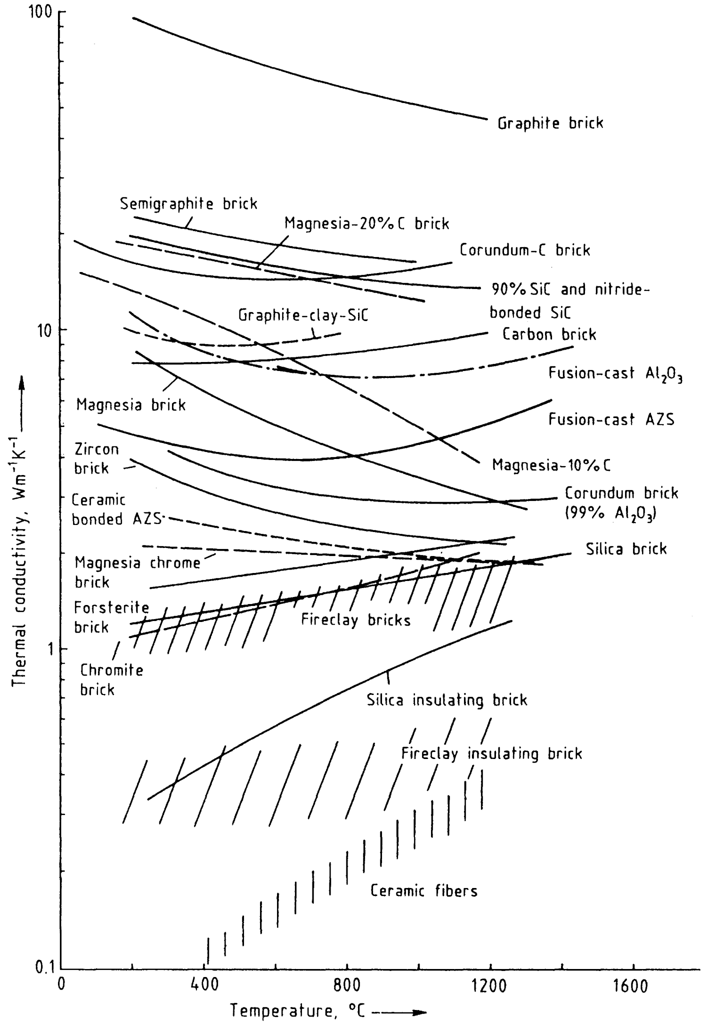

FIG. 5. Thermal conductivity of refractory products (VDI-Richtlinie 3218, Blatt 4, 1977, completed).

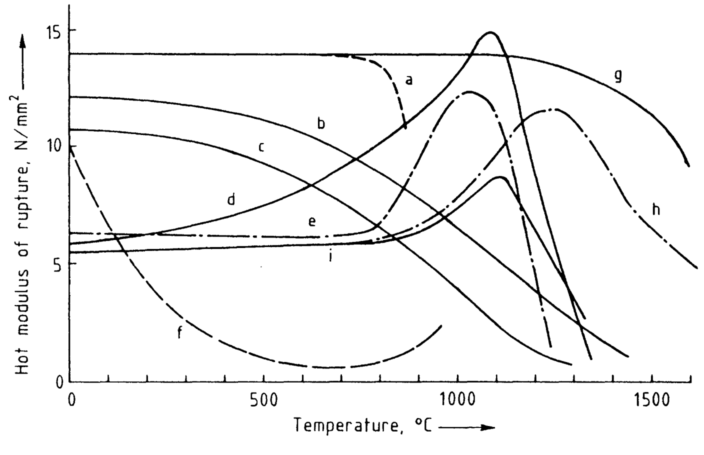

FIG. 6. Hot modulus of rupture of refractory bricks: Curve a, low-iron magnesia with 0.2% B_2O_3; curve b, corundum (99% Al_2O_3); curve c, iron-rich magnesia; curve d, low-alkali fireclay; curve e, alkali-rich fireclay; curve f, chemically bonded magnesia chrome; curve g, low-iron magnesia (C_2S bond); curve h, high-fired magnesia chrome; curve i, sillimanite.

namic elasticity is determined by means of ultrasonic velocity or resonance frequency. Values obtained by different methods can differ substantially, and they also depend on the type and structure of the refractory material under investigation (Table 3; Das and Jeschke, 1975). Refractory bricks often show strictly elastic behavior only within a narrow deformation range. Increasing stress and temperature lead to irreversible deviations from linearity. This is a consequence of deformations caused by mechanical or creep factors at higher temperatures. Irreversible changes are an indirect result of the porous, inhomogeneous textures of refractory materials.

The temperature dependence of stress–strain curves for a refractory material provides an indication of the elasticity of the microstructure (relaxation) and the onset of plastic deformation.

3.11 Thermal-Shock Resistance

Thermal-shock resistance is an important property that describes the behavior of a material during a more or less rapid change in temperature. Cracks can lead to considerable wear due to spalling, and they also increase susceptibility to corrosion. The test procedures are quite diverse (water or air quenching, dipping in metal melts, flame-impingement tests, cycling between two temperatures, etc.). Thermal-shock resistance is highly dependent on sample shape and the test conditions. Densification, uptake of foreign material, and the formation of zones of differing thermal expansion during refractory-product use can also lead to substantial changes in thermal-shock resistance. Thus, only a test procedure adapted to suit specific service conditions has any chance of success. Another way of circumventing the difficulties with thermal-shock tests is to invoke thermal-shock parameters. These represent an attempt to classify materials in terms of other properties (e.g., strength, modulus of elasticity, thermal expansion, thermal conductivity, work of fracture, and fracture toughness) (Hasselmann, 1978; Uchida *et al.*, 1990).

Two important parameters have been defined to assess the thermal-shock resistance of refractories:

Table 3. Strength and elasticity properties of refractory bricks (Das and Jeschke 1975).

		Cold strength (N/mm^2)			Static modulus of elasticity (10^4 N/mm^2)				Dynamic modulus of elasticity (10^4 N/mm^2)		
Brick type	Microstructure	Compression	Bending	Tension	Compression	Bending	Tension	Torsion	Young's modulus	Shearing modulus	Poisson's ratio
Magnesia	Crystalline, homogeneous	50	6	3.5	4.5	2.0	2.2	0.8	7.0	2.9	0.2
Magnesia-chrome	Crystalline, inhomogeneous	65	4	1.5	3.0	1.0	0.8	0.3	2.9	1.2	0.2
Fireclay	Crystalline + glassy phase	30	7	4.5	1.0	0.8	0.9	0.4	2.3	1.0	0.1
Fused silica	Glass, homogeneous	400	20	10	4.0	3.0	1.6	1.2	4.8	2.1	0.14

$R'''' = WE/S^2$

(situation involving a fracture of a catastrophic nature), and

$R_{ST} = [W/\alpha^2 E]^{0.5}$

(situation in which crack propagation is stable or quasistable), where W = work of fracture, E = modulus of elasticity; S = modulus of rupture, and α = thermal-expansion coefficient.

3.12 Electrical Resistance

The electrical resistance of a refractory material is important with respect to the insulation and operation of electrical furnaces. Conventional refractory ceramics display very high electrical resistance at room temperature ($>10^5$ Ω m), but this decreases with increasing temperature, reaching values like 1 Ω m at 1400 °C (VDI-Richtlinie 3128, 1977). Even low concentrations of foreign components can substantially reduce electrical resistance.

3.13 Corrosion Resistance

The corrosion resistance of a refractory product is influenced primarily by the chemical compositions of the corrosive agents and the refractory material. Porosity, pore-size distribution, and mineral structure of the refractory material determine the reaction kinetics. Corrosion is greatly affected by specific operating conditions (furnace atmosphere, temperature, temperature gradient in the lining, crack formation due to additional temperature changes, static or dynamic action of the corrosive media), and so conclusions drawn from laboratory slagging experiments are of limited value unless these have been carefully related to real operational conditions. The diverse character of corrosive attack on refractory materials is reflected in the numerous corrosion tests proposed (Kobayashi *et al.*, 1982).

Refractory products are subject to three main types of corrosive stress:

1. effects of the atmosphere and changes in the atmosphere,
2. attack by vapors, flue dusts, and condensates, and
3. attack by slag and other fluid media.

Crucible tests are usually performed in cavities prepared in cubes of the test material. The corrosive material (e.g., glass, slag, or a salt) is placed in the prepared cavity and the crucible is then heated. This test provides only a rough classification of materials.

The rod-immersion method has achieved importance for testing the attack of melts on refractory materials. Bars of the refractory ceramics are dipped into the melt. By rotating the bars a dynamic test is achieved.

Semi-industrial corrosion tests are somewhat extravagant in terms of time and equipment, and are a response to the demand for a test procedure that provides assessments as nearly equivalent as possible to actual operating conditions. This in turn implies that each procedure must be adapted to a specific application. Special emphasis is placed on the testing of large samples (e.g., a brickwork), which allows one also to test the joining materials.

4. TYPES OF REFRACTORY CERAMICS

4.1 Dense, Shaped Refractory Products

Characteristics of selected types of conventional refractory bricks are summarized in Tables 4 and 5.

4.1.1 Silica Bricks and Fused-Silica Ceramics Silica bricks (>93% SiO_2) (Jeschke *et al.*, 1989; Lepère *et al.*, 1992) consist mainly of silicon dioxide (SiO_2), generally described as silica. Silica occurs in a variety of crystalline modifications (e.g., quartz, cristobalite, tridymite), each with a high- and low-temperature modification, and also as an undercooled melt called fused silica. The crystal structures of the SiO_2 modifications differ widely, and so distinct density (volume) changes and steps in the thermal expansion are seen during transformation. This is of great importance during heating and cooling (i.e., during firing and use of silica refractories).

The raw materials for silica bricks are naturally occurring quartzite and other quartz materials with low Al_2O_3 and low alkali-oxide content. The grained material is

Table 4. Properties of conventional refractory bricks of the system SiO_2–Al_2O_3; for maximum service temperatures see Fig. 7.

Type	Al_2O_3 (%)	SiO_2 (%)	Fe_2O_3 (%)	$Na_2O + K_2O$ (%)	Density (g/cm^3)	Bulk density (g/cm^3)	Open porosity (%)	CCS[a] (N/mm^2)	T_{05}, °C
Fused silica, compact		99.5			2.2	2.16	0	200	1640–1690[b]
Fused silica, slip cast		>99			2.2	1.90	13–16	50–80	
Silica	<1.5	>95 CaO < 3	<0.5	<0.2	2.34–2.36	1.75–2.0	15–23	20–100	
Low-alumina fireclay	20–30		<2.5	1–3	2.5–2.65	1.9–2.3	15–25	>35	1250–1300
Fireclay	30–45		0.6–2.5	0.5–2.5	2.55–2.75	1.8–2.4	12–25	20–85	1300–1550
Andalusite, sillimanite	55–65		<1.5	<1	3.0	2.4–2.6	15–20	50–90	1470–1600
Bauxite	80–85	8–14	1.3	0.5	3.6	2.7–2.9	18–22	65–75	1400–1500
Corundum (80% Al_2O_3)	80–85		<1.3	<1	3.6	2.8	18	70	1550
Corundum (97% Al_2O_3)	>97		0.3	<0.4	3.96	3.1	<20	>50	>1650

---

[a] Cold compressive strength.

[b] Fused silica, when devitrified to cristobalite. In the glassy state T_{05} ca. 1350 °C.

Table 5. Properties of conventional ceramic-bonded basic bricks.

Type	MgO (%)	CaO (%)	SiO_2 (%)	Fe_2O_3 (%)	Al_2O_3 (%)	Cr_2O_3 (%)	Density (g/cm^3)	Bulk density (g/cm^3)	Open porosity (%)	CCS[a] (N/mm^2)	T_{05} °C
Magnesia (range)	85–95	0.8–3.9	0.2–6	0.2–6	0.1–5		3.5–3.8	2.8–3.1	14–18	40–110	1500 – >1700
Forsterite bonded	94	1.7	3.8	0.2	0.2			3.0	15	90	1700
C_2S bonded	96	2	1	0.5	0.2					60	>1700
99% MgO	99	0.1	0.1	0.1			3.58			55	>1740
Magnesia–spinel	86		0.3	0.8	13			2.95	15	55	>1600
Magnesia–chrome	55–80	<3	<5	5–12	<11	6–20	3.7–3.85	2.8–3.2	14–21	>30	>1500
Chrome–magnesia	25–25	<2.5	<5	7–15	<22	15–35	3.8–4	2.8–3.35	16–21	>40	>1500
Chromite	<25	<2	<7	9–18	<28	>25	3.8–4	3–3.3	17–21	>30	>1500
Forsterite	50–60	<1.5	<40	<8	<8	<1.5	3.35	2.6–2.8	15–21	>25	>1600
Doloma	>36	<61	<1.5	<1	<1		3.38	2.6–2.9	14–21	30–80	>1700

---

[a] Cold compressive strength.

normally bonded with about 2%–3% slaked lime and some sulfite lye solution as a temporary binder.

Fired silica bricks contain cristobalite, tridymite, perhaps some residual quartz (today usually <6%), some wollastonite ($CaSiO_3$), and about 2%–10% glassy phase. Silica bricks are characterized by very good thermal-shock resistance above 600 °C and high-temperature load-bearing capacity up to about 1600 °C.

Applications: coke ovens, hot-blast stoves, glass-melting furnaces (crowns), and furnaces for the ceramics industry.

Silica bricks in the applications cited usually show no deformation or critical wear. Acquired impurities migrate with melts formed in the bricks themselves deep into the colder part of the lining, leading to a zoned texture.

Fused-silica products are fabricated as fused blocks (manufactured with a special sink fuse method) or in special shapes starting from crushed fused grain, formed (often by slip casting), and fired at about 1050 °C. A characteristic feature of fused silica is its very low thermal expansion (i.e., high thermal-shock resistance). At about 1150 °C cristobalite crystallizes without any large volume change, but this crystallized material is sensitive to cooling below 300 °C (like silica bricks).

Applications: mainly in the glass industry, for hot repairs; submerged nozzles and shrouds for the casting of steel; foundries.

4.1.2 Fireclay and High-Alumina Bricks

These refractory products, also called alumosilicate bricks, belong to the system SiO_2–Al_2O_3 (Pask, 1990). Only three phases exist in this system: crystalline SiO_2 modifications, corundum (α-Al_2O_3), and mullite. The composition of the mullite ranges from $3Al_2O_3 \cdot 2SiO_2$ with about 72% Al_2O_3 (formed on sintering above about 1000 °C) to $2Al_2O_3 \cdot SiO_2$ with about 78% Al_2O_3 (crystallizes from melt).

The morphology of mullite crystals may vary considerably depending on the type and composition of the raw materials, the formation conditions during firing, and subsequent thermal treatment. Mullite shapes range from fine-needled crystals formed at low temperature in the presence of a melt to coarse crystals in materials with low flux content, fired at high temperatures, or in fused mullite. The shape of the mullite crystals and their arrangement in the texture of the brick (e.g., "mullite felt") have a pronounced influence on the properties of fireclay and high-alumina bricks, especially on hot strength.

An overview of refractoriness and maximum application temperatures for fireclay and high-alumina bricks is given in Fig. 7. Refractoriness in the field of fireclay refractories falls considerably below the liquidus line of the pure system SiO_2–Al_2O_3 mainly because of their high content of alkali oxides and alkaline earths (up to 4.5%). Maximum application temperatures are distinctly lower because of time-dependent shrinkage or creep, which significantly affects the life of a refractory lining. It is also necessary to distinguish between application under a thermal gradient (as in a refractory wall) and under quasi-isothermal conditions (checkers in regenerators). The broad range of maximum application temperatures for high-alumina bricks shows clearly the influence of the diverse chemical and mineralogical compositions of these bricks.

4.1.2.1 Fireclay Bricks. Fireclay bricks have an Al_2O_3 content of 10%–45%. They are manufactured from fireclay grog (see Glossary) and bonding clay mixtures. The amounts of bonding clay and added water vary, a factor that largely determines both the subsequent forming process (plastic, semidry, or dry shaping) and the properties of the resulting fired bricks. Low-alumina or siliceous fireclay bricks may contain as additives granular SiO_2 raw materials. Variation of the clays, grogs, and grain sizes permits the production of numerous brick types (see Stahl-Eisen-Werkstoffblatt 915, 917, 1984). Fireclay bricks contain up to ca. 3% Fe_2O_3 (for low-iron bricks, <1%), ≤3% TiO_2, <1% CaO + MgO, and alkali oxides, mainly K_2O (≤3.5%) and Na_2O (usually <0.25%). High-quality bricks include <0.5% Na_2O + K_2O.

Mullite, cristobalite, and/or quartz and a glassy phase (composition: ca. 80% SiO_2, 10% Al_2O_3, the rest fluxes, mainly alkalies) are the usual components of fireclay bricks. The mullite content is generally proportional to the Al_2O_3 content of the bricks. The amounts of cristobalite (0%–25%) and of the glassy phase (20%–60%) depend largely on

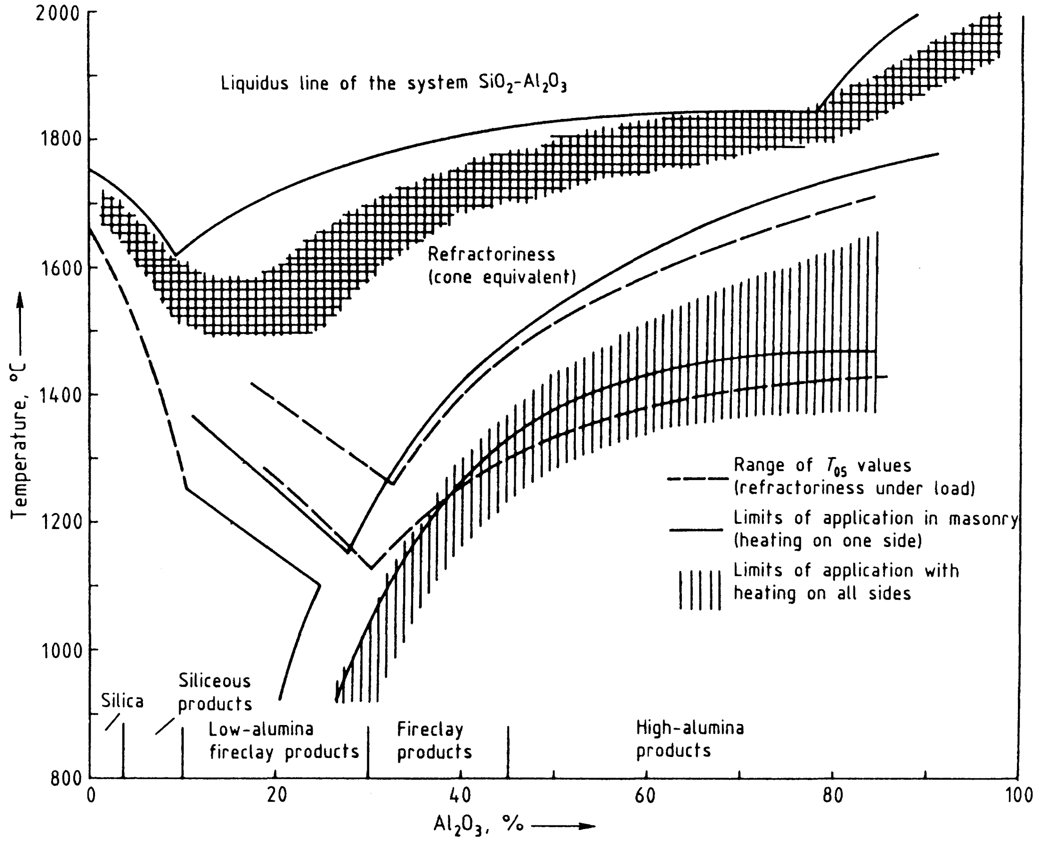

FIG. 7. Refractoriness, refractoriness under load, and maximum application temperature of fireclay and high-alumina bricks (Routschka, 1981).

the content of alkali oxides, which can be deduced from the phase diagram of Al_2O_3–SiO_2–K_2O.

Attack by slag on fireclay bricks normally occurs in relatively small reaction zones; wear progresses by erosion of the resulting highly viscous reaction zone. In contact with glass melts, alkali migrates into the brick texture and the bricks are slowly dissolved in the glass melt. Other forms of damage include alkali bursting (for bricks with >ca. 35% Al_2O_3), CO disintegration, and thermal spalling (dense bricks).

Main applications: general furnace construction, blast furnaces and hot-blast stoves, foundries, furnaces for the nonferrous-metal industry, coke ovens, glass industry, cement industry.

4.1.2.2 High-Alumina Bricks. Bricks with >45% Al_2O_3 are required for increased thermal and chemical stresses. These are produced from various high-alumina raw materials (sillimanite, andalusite, sintered and fused mullite, sintered alumina or fused alumina, often called corundum, or sintered bauxite), alone, in combination, or together with other raw-material types, mainly fireclay grog. The nomenclature applied to the products depends on the primary high-alumina raw material.

A small amount of clay (<15%) is normally used to produce ceramic-bonded fired bricks. Use of a mixture of clay and fine alumina leads to a mullite-bonded brick. A chemical bond on an aluminum–phosphate base has also proven successful regarding corrosion resistance.

Addition of Cr_2O_3 powder (5%–15%) to corundum bricks, either alone or in combination with ZrO_2-containing raw materials, leads to a highly corrosion-resistant corundum–chrome brick.

The chemical and mineralogical compositions of high-alumina bricks vary over wide ranges because firing of the bricks normally fails to produce phase equilibrium, and significant changes may occur in the properties of the bricks in service. Bauxite bricks, which constitute the majority of high-alumina bricks, show more or less after-expansion in service caused by reactions between the clay bond and corundum and resulting in mullitization.

The main mineralogical components are corundum and mullite. Sillimanite and andalusite bricks normally contain residual amounts of these minerals. Compared with fireclay refractories, the alkali-oxide content and the extent of glassy phase are low (<20%).

Properties of the different types of high-alumina bricks are compiled in Stahl-Eisen-Werkstoffblatt 912 (1984).

The corrosion resistance of such bricks to glass, slag, and iron and steel melts is good. They are prone to alkali attack: high-alumina bricks with low Al_2O_3 content react to produce low-melting alkali aluminum–silicate phases or glasses. With higher alumina content alkali bursting occurs (due to crystallization of, for example, nepheline, $Na_2O \cdot Al_2O_3 \cdot 2SiO_2$), as does the formation of β-Al_2O_3 ($Na_2O \cdot 11Al_2O_3$).

Main applications: iron and steel industry, nonferrous-metal industry, glass and cement industry, ceramic industry (also as high-temperature kiln furniture).

4.1.3 Basic Bricks (Alper, 1970)

Basic bricks belong mainly to the system MgO–CaO and consist of magnesia (periclase), doloma (dolomite), and CaO without or in combination with chrome ore, chrome oxide, spinel ($MgO \cdot Al_2O_3$), and carbon (graphite). Forsterite, chrome ore, and spinel bricks are included as well, since these materials are often used in place of or in contact with magnesia bricks. An overview is provided in Table 6.

4.1.3.1 *Magnesia Bricks.* According to ISO 1109, magnesia bricks contain at least 80% MgO. A variety of magnesia raw materials are used to manufacture magnesia and magnesia-containing products, e.g., natural magnesite sinters, seawater magnesia, or fused magnesia with various chemical and mineralogical compositions. The magnesias contain varying amounts of Fe_2O_3 (<8%), Al_2O_3 (<0.3%), CaO (<4%), SiO_2 (<4%), and B_2O_3 (<0.2%). Besides the main mineralogical component periclase, silicates and ferrites occur as secondary components. The CaO/SiO_2 ratio and the B_2O_3 content (formation of low-melting components) determine the mineral compositions and properties (hot strength, corrosion resistance) of magnesia bricks. Monticellite ($CaO \cdot MgO \cdot SiO_2$, CMS) with a CaO/SiO_2 mass ratio of 0.94 has the lowest melting point (1495 °C) of the silicates.

High-performance bricks are highly fired, have a low iron content (<1.5% Fe_2O_3), and contain dicalcium silicate $2CaO \cdot SiO_2$ (short

Table 6. Classification and types of basic bricks.

Brick type	Groups' MgO content (%)	Remarks
Magnesia	80–98	Including impregnated fired and pitch- or resin-bonded (untempered and tempered) bricks (carbon content after coking <7%)
Magnesia–doloma	50–80	
Dolomite	40	
Magnesia–carbon	30–70	Addition of natural graphite
Magnesia–doloma–carbon	C ≥7	Pitch- or resin-bonded untempered and tempered bricks
Doloma–carbon	(up to 25 C)	With or without antioxidants
Magnesia–chrome	30–80	Subgroups: direct-bonded, rebonded, and chemically bonded bricks
Chrome ore	<30	
Magnesia–spinel	60–80	Addition of sintered or fused magnesium aluminum spinel
Magnesia–zirconia	80, 90	Addition of zirconia and/or zircon
Spinel	20, 30	$MgO \cdot Al_2O_3$, $MgO \cdot Cr_2O_3$
Forsterite	40, 50	Fired and unfired bricks
CaO	CaO 90	

formula: C_2S, m.p. 2130 °C) or forsterite $2MgO \cdot SiO_2$ (short formula: M_2S, m.p. 1890 °C) between the periclase crystals. These bricks show first subsidence at temperatures ≥ 1600 °C. Magnesia bricks display the highest thermal expansion of refractory ceramics and are therefore prone to thermal shock and thermally induced stress in furnace linings. Thermal-shock resistance can be increased by a coarse-grain distribution or the addition of Al_2O_3 (spinel formation on firing, microcrack formation) or chrome ore (texture elasticity). Special products include magnesia–forsterite (olivine) bricks and zircon- or zirconia-containing magnesia bricks.

High refractoriness, good resistance to attack by basic slags, and alkali resistance characterize magnesia refractories. Slagging causes diffusion and migration of silicate or ferritic melts deep into the brick, producing a zoned structure prone to peeling (structural spalling). A pitch or resin, or carbon bond, reduces the infiltration of metal and slags under reducing atmosphere service conditions.

Magnesia bricks are sensitive to hydration (humid atmosphere between 40 and 120 °C), SO_3 (magnesium–sulfate formation), and SiO_2 attack (forsterite bursting). Excessive growth of magnesia crystals during service, promoted by acquired impurities and high temperature, can also loosen brick texture.

Applications: steel-melting and -refining furnaces and vessels, regenerators for glass-melting furnaces.

4.1.3.2 Doloma Bricks. These consist of sintered dolomite (doloma), the main components of which are evenly distributed crystals of CaO and MgO. The principal impurities are calcium compounds like dicalcium ferrite, $2CaO \cdot Fe_2O_3$ (C_2F); brownmillerite $4CaO \cdot Al_2O_3 \cdot Fe_2O_3(C_4AF)$; tricalcium aluminate, $3CaO \cdot Al_2O_3$ (C_3A); tricalcium silicate (C_3S); and dicalcium silicate (C_2S). The SiO_2 content should be $<1\%$.

Their free-lime content makes these products susceptible to hydration. For this reason, manufacture, transport, storage (packed in shrink foils), and installation require virtually moisture-free environments and early use or heating. Hydration resistance is increased by pitch or resin bonding and tempering or pitch dipping and impregnation of the fired bricks.

Doloma bricks are sensitive to attack by iron-rich and SiO_2-rich slags. The periclase component is more resistant than the CaO component. Fluid melts are formed with slags rich in FeO. MgO absorbs the FeO in a solid solution.

Applications: converters and ladles for the steel industry (pitch- or resin-bonded, tempered bricks) and the cement and lime industry (fired carbon-free bricks) in competition with magnesia and magnesia–spinel bricks.

4.1.3.3 Basic Bricks Containing Chrome. Most bricks of this type are fired mixtures of magnesia and varying amounts of chrome ore, usually in coarse to medium grain sizes, and sometimes also amounts of chrome oxide. The refractory chrome ores consist of a chromite, a spinel $(Mg,Fe^{II})O \cdot (Cr,Al,Fe^{III})_2O_3$ of variable composition depending on the ore deposit, and greater or lesser amounts of a magnesium silicate. The content of the silicate phase should be as low as possible. High-quality magnesia–chrome bricks have a SiO_2 content of $<2.5\%$ and a $CaO + SiO_2$ content of $<3.5\%$. In contrast to magnesia bricks, the CaO/SiO_2 ratio and the B_2O_3 content have little influence on hot strength.

Special brick types are high-fired bricks, with low SiO_2 content and a predominant direct bond between magnesia, residual chrome ore, and so-called secondary spinels (direct-bonded bricks); bricks containing prereacted sintered or fused magnesia–chrome grain (rebonded basic bricks), the grain in this case containing a large amount of directly bonded phases; and chromite bricks, consisting mainly of chrome ore and a small amount of magnesia, which are quite inert and very resistant to acidic and basic slags.

The advantages of adding chrome ore to magnesia mixtures are better thermal-shock resistance, increased corrosion resistance to slags with a CaO/SiO_2 ratio <1.5 (acidic slags) due to the chromite, lower thermal expansion, and lower thermal conductivity.

Applications: ladles and vessels for vacuum treatment of steel and argon–oxygen decarburization (AOD) converters; rotating cement furnaces and glass-furnace regenerators (in these applications formation of toxic hexavalent chromium occurs and chrome-free alternative bricks like magnesia–spinel or magnesia bricks are used increasingly);

open-hearth furnaces; limited applications for chromite bricks.

4.1.3.4 Spinel-Containing Bricks. The magnesium–aluminum spinel $MgO \cdot Al_2O_3$ has recently attracted greater attention in the refractory industry (Dal Maschino *et al.*, 1988). Spinel (71.7% Al_2O_3, 28.3% MgO) is the only compound in the system $MgO \cdot Al_2O_3$, and it has a melting point of 2135 °C. First formation of a melt occurs at 2020 °C, so spinel can be classed among the high-temperature materials. Sintered spinel raw materials are available with 66%–90% Al_2O_3, as are fused spinels.

The following brick types are currently in active service:

Magnesia–spinel bricks consist mainly of very low-iron and low-SiO_2 magnesia together with sintered spinel (up to 25%). This brick type was developed as a chrome-free alternative to magnesia–chrome brick in rotating cement furnaces.

A special type of spinel brick is the picrochromite brick, containing fused chrome–spinel ($MgO \cdot Cr_2O_3$) grain, which shows extraordinary resistance to blast-furnace slags and acidic coal ashes.

4.1.3.5 Forsterite Bricks. The principal component is magnesium orthosilicate, forsterite ($2MgO \cdot SiO_2$). Olivine is the main raw material, containing <8% FeO.

Applications: regenerators for glass-melting furnaces, safety linings in the iron and steel industry (restricted use).

4.1.4 Zircon- and Zirconia-Containing Bricks For characteristics of these bricks, see Table 7 (Begley and Herndon, 1971).

The primary raw material for ZrO_2-containing refractories is zirconium silicate, $ZrSiO_4$ (mineral name: zircon), with a theoretical composition of 67.2% ZrO_2 and 32.8% SiO_2. It is stable to ca. 1800 °C, but the temperature for thermal decomposition into ZrO_2 (baddeleyite) and SiO_2 may be far lower depending on the type and amount of accompanying components. Chemical processing of zircon results in powdered zirconium oxide, which is worked up to various types of refractory raw materials (sintered and fused material, stabilized or partially stabilized materials).

Pure ZrO_2 (e.g., the mineral baddeleyite) has a monoclinic lattice, which is transformed reversibly beginning at about 950 °C into a tetragonal modification. This is stable to ca. 2300 °C, at which point it is transformed into a cubic modification. The monoclinic–tetragonal transformation is accompanied by a volume change of ca. 7% together with considerable hysteresis (Fig. 8), resulting in stress and cracking in pure ZrO_2 ceramics. Consequently, such ceramics cannot be manufactured. Addition of suitable metal oxides (for refractory purposes, mainly CaO, MgO, and Y_2O_3 in solid solution with ZrO_2) permits the cubic modification to be stabilized down to room temperature. A fully stabilized ZrO_2 can be achieved by the addition

Table 7. Ceramic-bonded zircon- and zirconia-containing refractories.

Type	ZrO_2 (%)	SiO_2 (%)	Al_2O_3 (%)	Bulk density (g/cm^3)	Open porosity (%)	CCS[a] (N/mm^2)	Thermal expansion, 1000 °C (%)	Thermal conductivity, 1000 °C ($W\ m^{-1}\ K^{-1}$)
Zircon–silicate								
Pressed or slip cast	62–65	33–34	0.5–1	3.6–3.8	17–22	20–100	0.4	2
Isostatically pressed				3.6–4.2	1–17	>70		
Zircon–mullite (AZS)[b]								
Normally pressed	11–23	10–15	65–73	2.9–3.2	15–24	60	0.65	1.9–2.5
Isostatically pressed				3.2–3.5	1–14			
Zircon–zirconia (nozzle for steel casting)	85			4.25	19			
Stabilized zirconia								
MgO stab.	95			5.0	10	400	0.75	1.5
Dense (CaO stab.)				5.6	0		1.1	

---

[a] Cold compressive strength.

[b] Mineral composition: zircon 0%–30%, baddeleyite 0%–20%, mullite 25%–30%, glassy phase <10%.

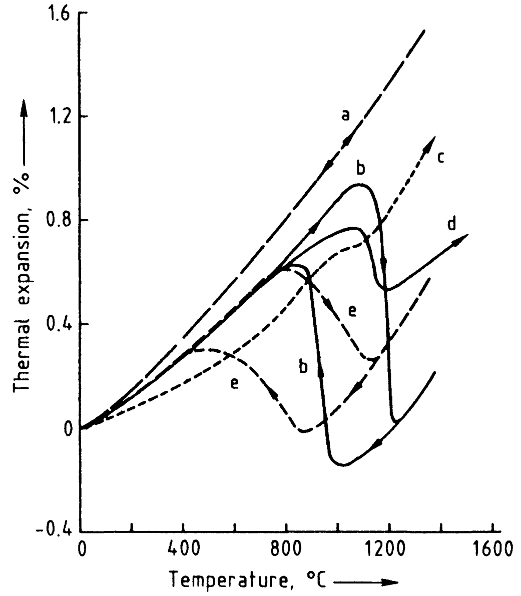

FIG. 8. Thermal expansion of ZrO_2 ceramics: Curve a, fully stabilized; curve b, baddeleyite; curve c, ceramic-bonded AZS brick; curve d, fusion-cast AZS brick; curve e, partially stabilized (Begley and Herndon, 1971, completed).

of about 5%–10% stabilizer. Partially stabilized zirconia raw materials are now used in most cases (3%–6% stabilizer added). This leads to smaller volume changes on heating and provides the additional advantage of creating microcracks, which increases the thermal-shock resistance of the brick.

The following types of ZrO_2-containing products can be distinguished:

1. ceramic-bonded zirconium–silicate bricks;
2. ceramic-bonded Al_2O_3–ZrO_2–SiO_2 (AZS) products, manufactured from mixtures of zircon and high-alumina raw material;
3. fusion-cast ZrO_2-containing bricks, such as fusion-cast AZS bricks (see Sec. 4.1.9); and
4. stabilized or partially stabilized sintered ZrO_2 products (zirconia ceramics).

Applications: zircon-containing refractories are used in glass-melting furnaces (superstructure, bottom, feeders) and in the steel industry as ladle bricks and nozzles. Zircon bricks are resistant to oil residues and especially to V_2O_5. ZrO_2 materials are used as nozzles and sliding nozzle plates for the extremely corrosive conditions associated with the teeming of special steel. Other refractory applications of ZrO_2 ceramics are in the ceramic industry as high-temperature kiln linings and kiln furniture.

4.1.5 Carbon and Graphite Bricks (Table 8) Carbon is unmeltable and not wetted by melts of metals, slag, or salts under oxygen-free conditions, but it is soluble in pig-iron melt. Carbon resists the attack of many acids. High thermal conductivity (especially of graphite) and low thermal expansion impart high thermal-shock resistance. Applications of these bricks are restricted because of susceptibility to attack by oxygen, steam, and CO_2 above 400 °C.

Raw materials with the required grain size (low-ash cokes, anthracite, natural and artificial graphite) are mixed together with tar or pitch in heated mixers and, while the mix is still hot, compacted by ramming or vibration. The cold, hardened shapes are subsequently fired at temperatures ≤1400 °C under the exclusion of air. Pure graphite shapes are produced by a second heat treatment (resistance heating) at ca. 3000 °C.

Depending on the crystallinity of the carbon or graphite and mixtures of the various raw materials, products can be fabricated with widely varying properties (thermal con-

Table 8. Carbon and graphite refractory bricks.

Type	Ash content (%)	Bulk density (g/cm^3)	Open porosity (%)	CCS[a] (N/mm^2)	Thermal expansion, 1000 °C (%)	Thermal conductivity, 1000 °C (W m^{-1} K^{-1})
Carbon	4.5–6.5	1.5–1.6	13–19	25–65	0.4	8–14
Semigraphite	1–3	1.6–1.7	13–21	25–48	0.38	14–24
Graphite	1	1.55	26	15	0.27	51

---

[a] Cold compressive strength.

ductivity and oxidation, erosion, and alkali resistance). Addition of, e.g., Al_2O_3 or SiC increases corrosion resistance (Piel *et al.*, 1989).

Applications: as large carbon blocks in the lower parts of blast furnaces; linings for electric reduction furnaces (electrolytic aluminum cells, FeSi production); acid-resistant linings for the chemical industry. Because graphite-containing bricks display high thermal conductivity they are used as a heat-abducting medium in conjunction with water-cooling systems in furnaces.

4.1.6 Carbon- and Graphite-Containing Refractories (Table 9) Combinations of different refractory materials with carbon (originating from pitch or resin bond or impregnation and/or graphite, mainly natural graphite) play a dominant role in the metallurgy of ferrous and nonferrous metals, and above all in the casting of steel.

4.1.6.1 Fired Graphite-Containing Refractories. Three types of carbon-containing fired refractory ceramics are produced.

1. Chamotte–clay–silicon carbide–graphite shapes, used mainly as crucibles for the melting and casting of nonferrous metals. The added SiC and silicon metal increase oxidation resistance and strength. To minimize oxidation during firing under reducing conditions (≤1350 °C) and during use, such shapes are glazed, or a glaze-forming additive is introduced into the mix. Silicon–carbide crucibles are manufactured with a pitch bond; i.e., a carbon bond is formed during firing.
2. Chamotte–graphite and high-alumina–graphite shapes for steel casting such as stoppers and nozzles.
3. Pitch- and resin-impregnated fired bricks (high alumina, magnesia, and doloma products).

4.1.6.2 Pitch- and Resin-Bonded Unfired Refractories. When manufacturing pitch- or resin-bonded products, preheated refractory grain and binder (usually <10%) are hot mixed (temperatures up to 160 °C depending on the viscosity of the binder) and hot

Table 9. Graphite-containing refractory products.

Type	Application	Al_2O_3 (%)	SiO_2 (%)	C (%)	ZrO_2 (%)	SiC (%)	Others[a] (%)	Bulk density (g/cm^3)	Open porosity (%)	CCS[b] (N/mm^2)
Chamotte and high-alumina graphite (fired)	Steel casting nozzles, stoppers	30–70		7–20				2–2.3	17–24	
Pitch- or resin-bonded Al_2O_3–SiO_2–SiC–C	Torpedo ladle bricks	68–90	19–30	10–17		5–8	0–6	2.7–2.8	4–10	55–80
Al_2O_3–C base[c]	Continuous steel casting (shrouds, nozzles, stopper rods)	42–56	14–24	24–33	0–3		0–5	2.2–2.45	12–18	
		60	0.9	24	0–9	4		2.7	17	
ZrO_2–C		1–2	5–7	7–10	75–80		0–2	3.55–4.2	12–16	
				20–25	67–74			3.2–3.6	15–18	
MgO–C			0–2	5–9		MgO 85–92	0–5	2.45–2.5	15–18	
			12–16	10–20		58–78		2.25–2.5	14–17	
	Converter, ladle, arc furnace bricks[c]			10–25		70–89	0–6		3–6	30–50

[a] Mainly antioxidants (metal powder, B_2O_3-containing compounds, etc.).
[b] Cold compressive strength.
[c] Thermal conductivity at 1000 °C, 6–13.5 W m^{-1} K^{-1}.

pressed. Liquid resins or solutions containing curing agents (hexamethylenetetramines) allow cold mixing. Resin hardening occurs at 200 °C (thermosetting). Impregnation of bricks with pitch or resin is carried out under vacuum and at the appropriate temperature (with resin solutions, room temperature).

Heating causes volatile matter to be released, and the "carbon bond" develops. The carbon yield depends on the type of binder and the coking conditions. The release of volatile organic compounds requires environmental precautions during the production and use of organic-bonded refractories.

The following types can be distinguished:

1. unfired pitch- or resin-bonded or tempered high-alumina bricks (after coking, ca. 5% C);
2. Al_2O_3–SiO_2–SiC–C bricks: combinations of high-alumina raw materials, SiC (5%–15%), and graphite (5%–15%);
3. pitch- or resin-bonded, unfired or tempered magnesia, magnesia–doloma, and doloma bricks and shapes (Zednicek, 1988; Zoglmeyer and Romai, 1991) (after coking, <6%, at which point additional carbon black is added up to 10%);
4. pitch- or resin-bonded or tempered oxide–graphite products from mixtures of mainly pure oxides (sintered and fused grain) together with ≤ca. 30% graphite, with or without the addition of so-called antioxidants (metal or nonoxide powders), usually <6%.

The group **4** includes corundum–carbon and zirconia–carbon products, used mainly as nozzles and shrouds for the continuous casting of steel, and magnesia–carbon bricks and shapes, used for lining steel converters, for the slag area in ladles and steel-refining vessels, and in electrical arc furnaces, and also as nozzles for continuous steel casting. Addition of fused silica or ZrO_2-containing aggregates to corundum–graphite products increases thermal-shock resistance. Special additives like BN reduce for instance the clogging of nozzles.

Especially the oxide–graphite products combine the nonwettability of graphite crystals by metal melts and slag (no infiltration into the bricks) with the high corrosion and erosion resistance of dense oxide grain. The graphite content additionally increases thermal-shock resistance. The products have low open porosity (<6%) in the delivered state, but porosity after coking increases to ca. 10%–12%.

The addition of metal powders (e.g., aluminum, magnesium, silicon) or alloys and/or nonoxide compounds like SiC, B_4C, or ZrB_2 inhibits or lowers carbon oxidation.

The effects of graphite content on essential properties of graphite-containing refractories are shown in Fig. 9.

During service, carbon-containing refractories normally show only a small oxidized working zone, where the effects of corrosion mechanisms known from comparable non–carbon-containing refractories are evident. Carbon still present behind this zone minimizes a deeper infiltration of slag or metal.

4.1.7 Silicon Carbide Bricks For characteristics of these bricks, see Table 10 (Wecht, 1977; Fickel and Völkel, 1988; Röttenbach and Pögl, 1989).

These can be divided into two classes according to their SiC content: SiC-containing refractories with ca. 40%–70% SiC, usually in combination with fireclay grog and a clay bond; and SiC products with 70%–100% SiC and different types of binding (and therefore different properties). The bonding types encountered are as follows:

1. Silicate or oxide bond, where the SiC

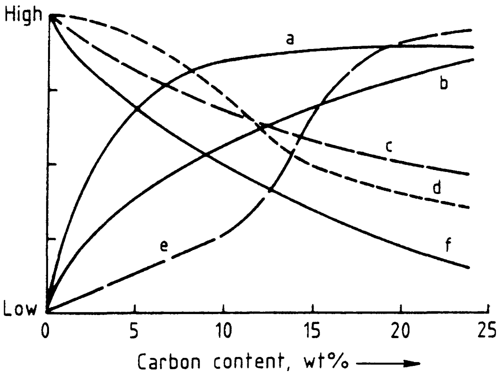

FIG. 9. Effect of graphite content on the properties of magnesia–graphite refractories: Curve a, slag-penetration resistance; curve b, thermal-shock resistance; curve c, oxidation resistance and hot strength with antioxidants; curve d, thermal expansion; curve e, thermal conductivity; curve f, oxidation resistance and hot strength without oxidants (Rathner *et al.*, 1992; Yamaguchi and Tanaka, 1992).

Table 10. Silicon–carbide refractories.

Type	Al_2O_3 (%)	SiC (%)	Si_3N_4/ SiAlON (%)	Bulk density (g/cm^3)	Open porosity (%)	CCS[a] (N/mm^2)	CMR[b] (N/mm^2)	Thermal expansion, 1000 °C (%)	Thermal conductivity, 1000 °C (W m^{-1} K^{-1})	Max. service temperature (°C)
Silicate-bonded	27	50		2.4	21	70	18	0.45	5	1450–1550
	8	90		2.5–2.7	12–20	100	30		10	
Nitride-bonded		70–85	≤25	2.6	14–20	130–150	50–70	0.45	10–15	1500–1600
Recrystallized		>99		2.6	15–18	300–700	100	0.48	22	1600 ox. 1700 red.
SiSiC		>80		3.05	0	1250	300	0.45	60	1350

[a] Cold compressive strength.
[b] Cold modulus of rupture.

grains normally are bonded with about 10%–20% bonding clay, alone or in combination with fine Al_2O_3, leading to a mullite bond; hot strength is determined largely by the composition of the formed glass phase.

2. Silicon–nitride and sialon (SiAlON) bond, produced by firing under a nitrogen-containing atmosphere a green mixture composed of SiC and Si, which may also include Al_2O_3 or other alumina-bearing compounds.
3. Self-bond, whereby a SiC bond is achieved by adding silicon and carbon and firing in an oxygen-free atmosphere.
4. Recrystallized SiC, where the bond phase (SiC) forms during high-temperature sintering (at ca. 2000 °C) by vaporization and condensation reactions of fine SiC particles in the presence of C.

A special type is the dense SiSiC ceramic, a self-bond SiC product with an infiltration of silicon metal. The exceptional properties of SiSiC ceramics include high thermal conductivity, low thermal expansion, very good resistance to thermal shock, high hot strength, up to 1350 °C (melting point of Si 1410 °C), and abrasion resistance.

SiC is sensitive to oxygen and such oxygen-containing gases as water vapor. In an oxidizing atmosphere SiC begins to decompose to SiO_2 and CO or CO_2 at ca. 900 °C. Above ca. 1200 °C the SiO_2 layer that forms on the SiC grains becomes glassy and retards further oxidation. SiC is also sensitive to alkali melts, basic slags, chlorine gas, and metal melts other than lead, zinc, and copper. Alkali resistance can be increased by using a nitride bond.

Applications: blast furnaces, kiln furniture in the ceramic industry (maximum service temperature 1500 °C), municipal incinerators and special refuse incinerators, retorts for zinc distillation, aluminum industry (submerged burner tubes), tubes for ceramic regenerators.

4.1.8 Fine-Grained Oxide and Nonoxide Ceramics (Kriegesmann, 1989) These normally very dense and highly refractory materials have not yet found appreciable application in the classical markets for refractories (exceptions include ZrO_2, Sec. 4.1.4, and SiC, Sec. 4.1.7) because of high price, a restriction to small or thin shapes, and often low thermal-shock resistance. The production of nonoxide shapes demands special fabrication methods not yet adapted to the refractory industry. Nonoxide refractory ceramics (carbides, nitrides, borides) are prone to oxidation at relatively low temperatures. They are normally not wetted by metal melts.

One of the most promising areas of potential development for fine ceramics is their use as engineering or functional refractory materials and in very special applications in industries such as the steel industry.

Examples of current uses for special fine ceramics include

1. chromium oxide bricks: used in highly corroded areas of glass tank furnaces (fiberglass production);
2. silicon nitride–based materials: skid rails in reheating furnaces for the steel industry, anchors for fixing ceramic-fiber linings, aluminum industry (not wetted by aluminum melt).

3. boron nitride–based composites with Si_3N_4 and AlN, tested also in combination with ZrB_2 and TiB_2 or mullite: break rings for horizontal continuous steel castings, reinforcement of the slag zones of submerged nozzles in continuous steel casting.

Most special ceramics are very dense, but porous materials based on Al_2O_3, ZrO_2, and SiC with defined pore-size distribution are increasingly being used for filtering steel and nonferrous-metal melts, as well as for high-temperature filtration of gases.

4.1.9 Fusion-Cast Bricks (Cichy, 1990) Fusion-cast (fused-cast) bricks are produced by melting of finely ground raw materials (alumina, zircon, zirconia, chrome ore, chrome oxide, special additives) in an electric-arc furnace and casting the melt into sand or graphite forms. The shapes produced range from small parts to blocks with lengths up to 1 m or more. During the slow cooling process (for larger blocks, 6–20 d) major cracks are prevented and the characteristic texture of a fused-cast brick develops: a dense structure (practically free of open pores) with intercrystalline bonding of the mineral phases.

The dense structure imparts high strength and both corrosion and erosion resistance. Disadvantages are low thermal-shock resistance and structural inhomogeneities caused by contraction and segregation during cooling. Location of casting voids may be partially controlled by special casting and cutting away an excess part of the block. Cutting and grinding are inherent parts of the technology of larger fusion-cast bricks. The types of fusion-cast bricks are described in Table 11. Most fusion-cast products belong to the system Al_2O_3–SiO_2–ZrO_2(–Cr_2O_3).

The different alumina types of fusion-cast bricks are produced by adding controlled amounts of soda to the raw-material alumina. The β type has a coarse friable structure with no glassy phase. The β''' type contains additionally MgO and shows high stability to alkali attack (special development as checker bricks in glass tank regenerators). The corundum–mullite type contains coarse mullite crystals in a glassy matrix.

Fusion-cast corundum–zircon blocks (also called fusion-cast AZS material) are the predominant refractory ceramics for glass-melt contact. The structure contains a corundum–zirconia eutectic phase with segregated modular and dendritic zirconia (baddeleyite) in about 20% of glassy matrix. A special feature of AZS fused-cast bricks is draining out of glassy phase during first heating up, which may lead to bubble formation in contact with glass melt. The addition of chrome oxide to corundum or AZS material considerably increases corrosion resistance caused by the formation of a solid solution of chrome–alumina.

Fused magnesia–chrome bricks have a high corrosion resistance due to its chrome–spinel phase (see Sec. 4.1.3.3.)

A newly developed fusion-cast ZrO_2 brick contains a special SiO_2-rich glassy phase that acts as a buffer against the transformation of baddeleyite. The bricks show high corrosion resistance to some special glasses (Zanoli, 1991).

Applications: primarily in glass tank furnaces and regenerators for glass furnaces (cruciform checker bricks), reheating furnaces in the steel industry (skid rails).

4.1.10 Refractory Ceramics with Low Thermal Expansion As in the case of glassy SiO_2, ceramics with low thermal expansion display extraordinarily high thermal-shock resistance, and they are especially qualified for service conditions involving frequent changes in temperature (e.g., kiln furniture). These ceramics are products based on cordierite ($2MgO \cdot 2Al_2O_3 \cdot 5SiO_2$), lithium–aluminium silicates (eucryptite, spodumene, petalite), and aluminum titanate ($Al_2O_3 \cdot TiO_2$).

Cordierite products have acquired considerable industrial significance. Fireclay grog or cordierite grog is bonded with cordierite-forming mixtures of clay, kaolin, talc, or magnesia through firing at ca. 1150–1350 °C. The cordierite content of the fired products is ca. 40%–80%. The maximum application temperature of normal cordierite shapes is 1250 °C. Addition of mullite and corundum increases both hot strength and service temperature.

4.2 Unshaped Refractory Materials (Plibrico Japan Comp., 1984; Taikabutsu Overseas, 1989)

Unshaped refractory materials, also called monolithics, are mixtures consisting of an aggregate (grains) and one or more bonds,

Table 11. Properties of fusion-cast bricks.[a]

Type	Al_2O_3 (%)	SiO_2 (%)	ZrO_2 (%)	Na_2O (%)	MgO (%)	Cr_2O_3 (%)	FeO (%)	Bulk density (g/cm^3)	Thermal expansion, 1000 °C (%)	Thermal conductivity, $W\ m^{-1}\ K^{-1}$			Mineral phases
										300 °C	800 °C	1400 °C	
Corundum (α-Al_2O_3)	99.5			0.2					0.8				Corundum
α/β-Al_2O_3	95	1.2		3.5				3.2		5	4.2	7	Corundum, $Na_2O\cdot 11A$
β-Al_2O_3	94.5			5.2				2.8					$Na_2O\cdot 11Al_2O_3$
β'''-Al_2O_3[b]	87.5	0.3		4.5	7.7			2.8		2.3	2	6.5	$15Al_2O_3\cdot 4MgO\cdot Na_2O$
Corundum–mullite	74	20	2	0.8					0.75	5	4	6.5	Corundum, mullite, 15
Corundum–zircon	50	16	33	1.1					0.75[c]	4.5	4	6.5	Corundum, baddeleyit
(AZS)	45.5	12	41	1						4.8	3.9	5.9	16%–21% glass
Zirconium oxide	0.8	4.5	94	0.4				5.3	0.7	3.6	2.5	3	Baddeleyite, 6% glass
Corundum–chrome	60	2			12	28	4	3.7					Ruby, spinel
Corundum–chrome–zircon	32	13	26			26	3	4.0	0.7	5.5	4	6	Ruby, baddeleyite, 20
Magnesia–chrome	7	2.5			57	20	12	3.2[d]	1	4	4.5	5.5	Periclase, spinel
Chromoxide–corundum	14	0.8				82		4.6	0.75				Spinel, ruby

---

[a] Open porosity <2%; cold crushing strength >200 N/mm^2.

[b] Microporosity ca. 12%; cold crushing strength 60 N/mm^2.

[c] See also Fig. 8.

[d] Porosity 15% (block).

prepared ready for use either directly or after the addition of one or more suitable liquids (usually water). These mixtures are either dense or insulating.

The ENV 1402-1 defines such products according to

1. intended application: materials for monolithic constructions, materials for repairs, materials for jointing;
2. product type and placement: castables, gunning materials, moldable materials (mainly ramming mixes, plastics, and taphole mixes), refractory cements and mortars, others (mainly dry mixes, injection mixes, and coatings, each corresponding to various types of placement), and prefabricated shapes made from unshaped materials (mainly castables and moldables).

Classification within the various categories is based on the temperature below which the material does not shrink more than a specified amount (linear shrinkage 2%–3%).

Four main types of unshaped products can be identified according to chemical composition:

1. alumina–silica refractories, including products consisting mainly of aggregates of alumina, silica, and/or alumosilicates;
2. basic refractories, mainly aggregates of magnesia, doloma, chromite, and spinel;
3. refractories containing carbon (i.e., graphite or carbon together with other nonwetting agents), including mixes for troughs of blast furnaces;
4. special products containing oxide or nonoxide aggregates different from the above (e.g., silicon carbide, zircon, zirconia).

The raw materials and compositions of monolithics correspond essentially to those of shaped products.

An overview of materials for monolithic construction is provided in Table 12.

4.2.1 Ramming Mixes These are noncoherent before use and made up of refractory aggregates, bond(s), and (if necessary) liquid(s). Processing is accomplished by hand or, more frequently, with the help of some mechanical device (pneumatic rammer, slinger, etc.) or vibration. Hardening generally occurs under the influence of heat.

Dry ramming mixes are a specialty product, often vibrated behind steel forms after mixing with a low-melting fritting agent (e.g., boric acid) or a thermosetting resin (hardening at ca. 200 °C).

4.2.2 Plastic Mixes These are ready-for-use products with a moldable, plastic consistency. The principal bonding agent is plastic clay. These mixes are supplied in a soft or preformed state (extrusion-pressed blocks). They are placed by ramming, vibration, pressing, or "extrusion."

4.2.3 Refractory Castables These are supplied dry and used after the addition of water or some other liquid followed by mixing. They are placed by casting with or without vibration (self-flowing), or by rodding, pumping, injecting, and sometimes tamping. Bond formation and hardening occur at ambient temperature. Hydraulic- or cement-bonded castables are the most common type of monolithics. Low-alumina cements with ca. 40% Al_2O_3 and 12% Fe_2O_3, and calcium–aluminate cements with 50%–80% Al_2O_3 and

Table 12. Types of unshaped refractory materials for monolithic constructions.

Type	Subdivision	Water content upon application (%)	Nature of bond
Moldable mixes	Dry ramming mixes (fritting mixes)	0–1	Ceramic
	Moist friable ramming mixes	3–6	Ceramic
	Plastic friable mixes	5–8	Chemical
	Preformed plastic mixes	6–10	Organic
Refractory castables	Dense castables	5–12	Hydraulic
	Insulating castables	10–50	Chemical
Gunning mixes	Moist friable mixes		Hydraulic
	Different types of castables	5–50	Chemical
	Special hot gunning mixes		Ceramic (special binders)

19%–38% CaO, are used as hydraulic binder (Kopanda and MacZura, 1990).

Two main families are the dense and the insulating castables, with a limit of 45 vol% porosity. The various types of dense castables are listed in Table 13. Figures 10 and 11 describe properties of castables, classifying them within the broader context of unshaped materials and comparing them with bricks.

As a result of a drastic decrease in cement content together with the presence of ultrafine powders (microsilica SiO_2, reactive Al_2O_3) and dense packing, dense castables require little mixing water and display low porosity and high cold and hot strength up to 1500 °C (due to, e.g., mullitization of the matrix). Low-cement castables without addition of microsilica show high hot strength at 1500 °C and higher (Kriechbaum *et al.*, 1994).

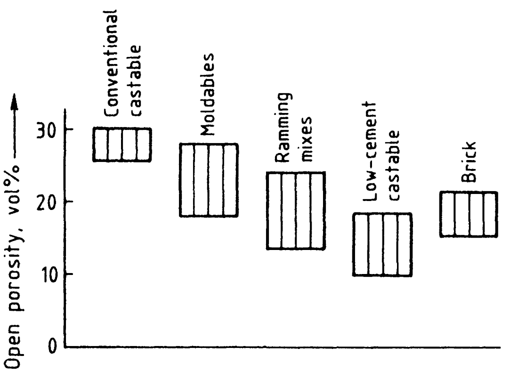

FIG. 10. Comparison of open porosities of various fired unshaped refractory materials (60%–90% Al_2O_3).

Insulating castables contain porous grains (like porous fireclay grog, thermally treated perlite, or vermiculite) and normally a high content of refractory cement (up to 40%).

Prior to heating a castable must be cured for a sufficiently long time (i.e., 24 h) to ensure full hydration and strength. Drying and

Table 13. Types of dense refractory castables.

Type	Classification ASTM C 401-89 and ENV 1402-1	Mixing water (approx.)	Essential components, bonding
Conventional castable (CC)	>2.5% CaO	9–12	Main hydraulic bonding: $CA^{a} + 10H^{b} \rightarrow CAH_{10}{}^{c}$ (<20 °C), $2CA + 11H \rightarrow C_2AH_8{}^{d} + AH_3{}^{e}$ (20–35 °C)
Low-moisture conventional castable[f]		7–9	Addition of ultrafine powders (high strength and abrasion resistance)
Low-cement castable (LCC)	1–2.5% CaO	4.5–7	Deflocculated castables containing aluminous cement, a minimum of 2% ultrafine particles (SiO_2, Al_2O_3; <1 μm), and at least one deflocculating agent
Ultralow-cement castable (ULCC)	0.2–1% CaO		Thixotropic rheological behavior (vibrational casting)
Clay-bonded castable	See LCC and ULCC	7	Contains clay (flocculated) and aluminous cement as a deflocculant
Self-flowing castable[g]		6	Consistency (rheopex rheological behavior), which allows flow and deaeration without vibration
No-cement castable	<0.2% CaO		Bonding: phosphate, alkali silicate, pitch or resin, Si or Al sol
Mixerless castable[h]			Delivered ready-mixed with water for casting (vibrating); contains 0.5% sodium silicate, thermosetting at 130–200 °C by gelation

---

[a] $CA = CaO \cdot Al_2O_3$.
[b] $H = H_2O$.
[c] $CAH_{10} = CaO \cdot Al_2O_3 \cdot 10H_2O$.
[d] $C_2AH_8 = 2CaO \cdot Al_2O_3 \cdot 8H_2O$.
[e] $AH_3 = Al_2O_3 \cdot 3H_2O$ (gel).
[f] Weavers *et al.*, 1985.
[g] Brachet *et al.*, 1992.
[h] Tabata *et al.*, 1992.

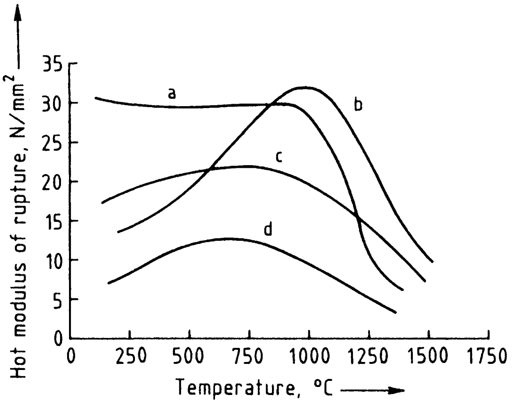

FIG. 11. Hot modulus of rupture of low-cement microsilica-containing castables, ramming mix, and dense brick (corundum-based materials) (Fletcher, 1989): Curve a, LCC; curve b, ULCC; curve c, brick; curve d, dry ramming mix. Low-cement castables without addition of microsilica achieve high hot strength at 1500 °C and higher.

first heat-up steps require careful monitoring to prevent explosive effects caused by excessive vapor pressure inside the lining.

4.2.4 Gunning Materials These are specially prepared for placement by pneumatic or mechanical projection means in cold and also hot application. Various special bonds have been developed for achieving a dense and adhesive gunned layer. Such materials are used mainly for repairs, especially in the iron and steel industry.

4.2.5 Prefabricated Shapes These are rammed or cast and may be pretreated up to 600 °C by the manufacturer as a way of providing blocks that are ready for use. Relatively large-volume parts are fabricated. The advantages are optimal manufacture, simple lining, and short installation times.

4.2.6 Jointing and Coating Materials These are essential for establishing a safe refractory brickwork or lining surface. Such products are fine-grain materials whose characteristics and composition must correspond to those of the associated bricks. Table 14 describes such materials.

Applications: principal applications of materials for monolithic construction are in the iron and steel industry, foundries, the petrochemical industry (especially insulating gunning mixes), and incinerators.

Table 14. Joints and coatings.

Material	Mixing liquid or water (%)	Type of bonding
Refractory mortars	10–25	Chemical, hydraulic, ceramic
Refractory mastics	12–30	Chemical
Adhesives	30–80	Chemical
Troweling materials	15–30	Chemical, hydraulic, ceramic
Coatings	30–60	Chemical

4.3 Heat-Insulating Ceramic Materials

Ceramic materials with a high pore volume (i.e., low thermal conductivity) are used extensively for thermal insulation of furnaces and auxiliary equipment. Apart from the effect of decreased thermal wall losses, the low weight of the insulating ceramic material allows lighter furnace construction and low heat-storage losses in discontinuous service.

Thermal conductivity depends not only on total porosity (overall gas volume) but also to a considerable extent on pore size and pore-size distribution, as well as texture ("thermal bridges") and mineralogical composition. Normally, the finer the pores, the lower the thermal conductivity. A maximum pore size of <1 mm is the target. Microporous materials have a thermal conductivity lower than that of air. Above 1000 °C thermal radiation must be taken into account as a heat-transfer mechanism.

High porosity implies low mechanical strength, high gas permeability, and low corrosion resistance, all of which must be considered in the use of insulating refractory materials. Furthermore, a porous texture leads to lower volume stability at high temperatures than in the case of dense materials (greater sensitivity to overheating) and a greater susceptibility to reducing atmospheres (reduction of, e.g., Fe_2O_3; increased shrinkage).

The main groups of insulating ceramic products are

1. insulating bricks,
2. insulating refractory bricks,
3. ceramic fibers,
4. porous or dense grains and granules of ceramic materials used as poured backup insulation,

5. insulating castables (see Sec. 4.2), and
6. special insulating products based on calcium silicate and microporous high-SiO_2 materials, with application temperatures $\leq$1100 °C.

4.3.1 Insulating Bricks Insulating bricks are based on diatomite (moler), vermiculite, and perlite, with bulk densities of ca. 0.45–0.8 g/cm^3, cold crushing strength of ca. 1–7 N/mm^2, and a thermal conductivity at 200 °C of 0.1–0.2 W m^{-1} K^{-1}. They are normally used up to 800 °C, but low-flux diatomite bricks may be applicable to 1100 °C.

4.3.2 Insulating Refractory Bricks According to official definitions, insulating refractory bricks have a total porosity of at least 45% and an application temperature of at least 800 °C. Porosity normally lies between 60% and 80%. The criteria for classification (ISO 2245) are as follows:

1. the temperature at which the material does not show a permanent linear shrinkage greater than 2%, and
2. the bulk density (which for a given product is directly related to porosity, and consequently to thermal conductivity).

Classification temperatures are not to be taken as acceptable service temperature limits, which depend on the particular service conditions.

Insulating refractory bricks are usually fabricated from fireclay, sillimanite or mullite, corundum, and silica as raw materials, together with a combustible material.

Types and properties of insulating refractory bricks are summarized in Table 15.

4.3.3 Refractory Ceramic Fibers (Cooke, 1991; Smith, 1990) A proposed definition describes ceramic fibers as synthetic mineral fibers suitable for use as heat-insulating materials and containing less than 2 wt% alkali-metal or alkaline-earth oxides. Most ceramic fibers are based on glassy or crystalline aluminosilicates and alumina.

Glassy (amorphous) refractory ceramic fibers (Al_2O_3 content ca. 45%–60%) are produced by fusing an aluminosilicate raw material (e.g., calcined kaolin) and/or sand and calcined alumina. Addition of ZrO_2, Cr_2O_3, B_2O_3, and other materials leads to changes in the properties of the fiber. The molten stream is attenuated into fiber form either by impact with a high-velocity jet of air and/or steam or by downward flow onto rotating wheels.

Polycrystalline fibers (62%–100% Al_2O_3) are sol-gel–derived fibers from, e.g., aluminum salts, colloidal silica, organosilane, and modifiers. Rotary spinning (short-staple fibers) is also practiced, as is the production of continuous filaments. After formation the fiber must be dried and calcined. Other types of oxide fibers (e.g., ZrO_2 fiber) are produced in the same way. This type of fiber is relatively expensive. To improve cost effectiveness, polycrystalline fibers are often blended with low-cost glassy fiber.

Properties and application temperatures for fibers and fiber products are shown in Table 16.

Table 15. Properties of insulating refractory bricks.

Type	Classification temperature (°C)	Al_2O_3 (%)	SiO_2 (%)	Fe_2O_3 (%)	Bulk density (g/cm^3)	Porosity (%)	CCS[a] (N/mm^2)	Thermal expansion, 1000 °C (%)	Thermal conductivity, 1000 °C (W m^{-1} K^{-1})
Silica	1500–1600	1	92 (3.5% CaO)	<0.5	0.8–1.2	50–65	3–4	1.2	0.5–0.8
Anorthite fireclay	1100–1250	36	45 (15% CaO)	0.7	0.5	85	1	0.6	0.2
Fireclay	1250–1400	35–42		<1.7	0.55–1.3	55–75	2–8	0.5–0.6	0.3–0.5
High-alumina 50–75% Al_2O_3	1400–1700	55–75		<1.5	0.8–1.5	45–75	2–12	0.6–0.7	0.35–0.55
Corundum hollow sphere	1800	98	0.6	0.1	1.2–1.4	60	8–18	0.8	0.8–1.1
Forsterite	1500	MgO 57	34	6	1.7	45	2.5	1.2	0.8

---

[a] Cold compressive strength.

Table 16. Ceramic fibers and fiber products.[a]

Type of fiber	Al_2O_3 (%)	SiO_2 (%)	Cr_2O_3 (%)	ZrO_2 (%)	Liquidus temp. (°C)	Classification temp. (°C)	Fiber diameter (μm)	Density (g/cm^3)	Maximum service temperature (°C): Oxidizing atmosphere, continuous	Oxidizing atmosphere, discontinuous	Reducing atmosphere
Glassy, amorphous	47	52			1760	1250	2–5	2.58	1100	1250	<1100
	63	36			1800	1400	3–6	2.63	1200–1400	1400 (1500[b])	Ca. 200° lower than in oxidizing atmosphere
	41	55	3		1785	1400	2–5	2.61			
	35	50		15	1810	1450	≈3	2.60			
Polycrystalline	85	15			1950	1600	≈7	3.40	1500	1600 (1700)	Ca. 100° lower than in oxidizing atmosphere
	95	5			2000	1600	≈3	3.58			

[a] For fiber products in general, the bulk density is 0.05–0.3 g/cm^3 (special products, ≤1.3 g/cm^3). Thermal-conductivity ranges are 0.08–0.12 $W\ m^{-1}\ K^{-1}$ (400 °C) and 0.16–0.40 $W\ m^{-1}\ K^{-1}$ (1000 °C). Fiber tensile strength is 400–1500 N/mm^2.

[b] Mixed fiber from high-alumina glassy fiber and alumina fiber.

The fiber is used in loose form or processed into any of a variety of flexible or essentially rigid forms. Blankets and textiles (ropes, yarns) are examples of products made by an air-laid system. Papers, felts, and boards are produced by wet-laid systems (like the traditional paper-forming process). Boards and special shapes are processed via vacuum casting. The ceramic fiber is slurried in water with organic and/or inorganic binders (e.g., SiO_2 sol) and fillers. The slurry is dewatered onto molds via a vacuum. Fiber is also used in unshaped refractory materials like gunning mixes and castables. Products include ceramic-fiber veneering elements, folded modules, and ready-for-use system blocks with anchoring.

Ceramic-fiber products have very low bulk density (0.05–0.3 g/cm^3; i.e., a porosity >90%) and unusually low thermal conductivity. In contrast to insulating bricks, thermal conductivity of ceramic-fiber products decreases with increasing bulk density. Thermal conductivity increases more with increasing temperature than in the case of comparable insulating bricks (Fig. 12). Thermal-shock resistance is extraordinarily high.

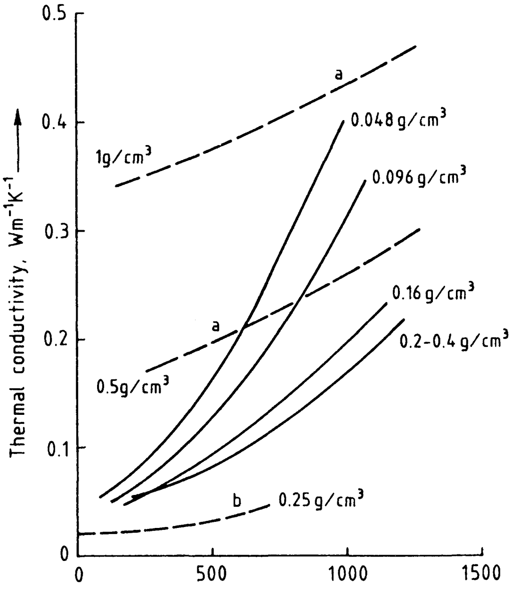

FIG. 12. Thermal conductivities of ceramic-fiber products compared with other insulating materials: Curves a, fireclay insulating bricks; curve b, microporous material.

Ceramic-fiber products are classified into groups according to the temperature (in 50 °C steps) at which the material does not show more than 4% shrinkage. Long-time application of ceramic fibers at maximum service temperature is limited significantly by shrinkage and crystallization and/or recrystallization (crystal growth). Impurities acquired from a furnace atmosphere (e.g., alkalies) may greatly increase shrinkage and crystal transformation. Ceramic-fiber products are sensitive to mechanical stress as well as, e.g., to high-velocity gas streams. Improvement is achieved by brushlike orientation of the fibers perpendicular to abrasion stresses, or by impregnating the surface with SiO_2 sol or the use of rigid ceramically bonded and prefired shapes.

Applications: Ceramic fibers have found application in all types of low- and high-temperature thermal processes, from use as a filler for joints or seals and insulation over veneering of furnace walls up to complete fiber-lined furnaces (e.g., in the ceramics industry, and heating furnaces for the steel and glass industries).

In the nonferrous-metal industry preshaped parts made from ceramic fibers have found application in direct contact with molten aluminum and copper alloys.

Proper construction and the use of appropriate joining techniques (gluing procedures, anchoring systems) are important factors in the lifetime of fiber furnace linings.

5. SELECTION AND USE

The service performance of a refractory material depends in roughly equal measure on

1. the properties of the refractory material,
2. the construction, lining and brick design, installation technique, drying, and initial heating, and
3. service conditions.

Most data sheets accompanying refractory products specify a service-temperature limit or maximum service temperature. This refers to the maximum temperature to which the refractory can be exposed during operation without corrosive influences (heating on one side only). No test yet exists for determining the maximum service temperature. Consequently, the indicated temperatures are based on practical experience supported by suitable hot tests (refractoriness under load, hot strength, reheat change). Classification temperatures for refractory products are not to be considered as service-temperature limits, which are usually lower.

In most cases, refractory ceramics are subject to a multitude of thermal, thermomechanical, mechanical, and corrosive stresses. Corrosion is the factor most likely to change the service properties of a refractory material in a significant way. A refractory material must be selected in accordance with those properties that are most important for a given service application. This requires close consultation between the refractory producer, the furnace builder, and the user. Often a refractory ceramic is "tailor made" for a specific application.

The output (tonnage) of refractory products is distributed among various industries in approximately the following way:

Iron and steel industry	55%–60%
Cement and lime industry	5%–8%
Glass industry	5%–8%
Ceramic industry	5%–8%
Nonferrous-metal industry	2%–5%
Chemical and petrochemical industry	2%–4%
Energy-producing industry	2%–4%

GLOSSARY

Aggregate: As applied to refractories, a ground mineral, consisting of particles of various sizes, used with much finer sizes (normally under 10 mm) for making formed or monolithic bodies.

Alumina: Aluminum oxide, Al_2O_3, in any of its many forms. As a general term in the refractory industry the α form (mineral name corundum), including calcined alumina, low-fired fine Al_2O_3 types, which may contain residual hydroxide or other oxides, high-fired alumina grain, called sintered (or tabular) alumina, fused alumina (often named corundum), and bubble alumina, hollow spheres of Al_2O_3. Beta alumina is not an alumina but a family of alkali-metal polyaluminate, e.g., $Na_2O \cdot 11Al_2O_3$, often a minor impurity in aluminas.

Bulk Density: The ratio of the mass to its bulk volume (including volume of total pores).

Bursting: Disintegration of refractories characterized by having the exposed face swell and grow until it breaks away from the brick mass, caused by a permanent increase in volume. Typical for chromite-containing material when exposed to iron oxide at high temperature. Alkali attack on aluminosilicate refractories leads to alkali bursting by formation of crystalline alkali aluminosilicates (e.g., nepheline, $Na_2O \cdot Al_2O_3 \cdot 2SiO_2$).

Calcine, Calcines: Refractory material, often fireclay or alumina (calcined alumina), that has been heated to eliminate volatile constituents and to produce desired physical changes.

Chamotte: See **Grog.**

CO Disintegration: A carbon deposition resulting from the decomposition of carbon monoxide gas into carbon within a critical temperature range. When deposited within the pores of refractory brick, carbon can build up sufficient pressure to destroy the bond and cause the brick to disintegrate.

Consistency: A measure of the flow, softness, or wetness of castables.

Corrosion: Destruction of refractory surfaces by the chemical action of external agencies.

Creep: Isothermal deformation of a stressed product as a function of time.

Direct-Bonded Brick: A fired refractory in which the grains are joined predominantly by a solid-state diffusion mechanism. The term "direct bond" was initially applied to fired magnesia–chrome refractories.

Erosion: Wearing away of refractories' surfaces by the washing action of moving liquids or particles in flowing gases.

Firebrick: Refractory brick of any type.

Fireclay: An earthy or stony mineral aggregate that has as the essential constituent hydrous silicates of aluminum with or without free silica, plastic when sufficiently pulverized and wetted, rigid when subsequently dried, and of suitable refractoriness for use in commercial refractory products.

Fireclay Brick: A refractory brick manufactured substantially from fireclay and fireclay grog.

Firing (Burning): The final heat treatment in a kiln to which refractory brick and shapes are subjected in the process of manufacture for the purpose of developing bond and other necessary physical and chemical properties.

Flux: A material that, when present in or added to a refractory material (even in very small amounts), appreciably lowers the temperature of formation of a liquid phase (e.g., promoting densification or sintering).

Fused Grain: Made ordinarily by arc-melting and refreezing; crushed and sized.

Fused Silica: Silica in the glassy or vitreous state. Synonyms include vitreous silica and silica glass.

Fusion Point: The temperature at which melting takes place. Most refractory materials have no definite melting points, but soften gradually over a range of temperatures.

Grog: A granular product produced by crushing and grinding calcined or burned refractory material, usually of alumina–silica composition. Grog from fireclay is sometimes named chamotte.

Gunning: An application technique that uses a pneumatic means to transport a refractory material and place it onto a cold or hot surface.

Kiln Furniture: Various supports, shapes, plates, or boxes (saggars) used to hold or contain ceramic ware during firing.

Matrix: The medium in which grains or crystal are embedded.

Modulus of Rupture: A measure of the transverse or "crossbreaking" strength of a solid body.

Monolithic Refractory: An unshaped refractory that is normally to be installed *in situ,* without joints, to form an integral structure.

Permeability: The capacity of a refractory for transmitting a fluid (gas or liquid).

Plastic Clay: A fireclay that has sufficient natural plasticity to bond together other materials that have little or no plasticity (also ball clay).

Plasticity: That property of a material that enables it to be molded into desired forms, which are retained after the pressure of molding has been released.

Pyrometric Cones: Elongated trigonal pyramids of standard size made from specified mixtures of ceramic materials, which when under stated conditions tip the supporting plaque at a specified temperature. May be used as an index of heat treatment or for testing of pyrometric-cone equivalent (refractoriness).

Pyrometric-Cone Equivalent (PCE): The number or temperature equivalent of that standard pyrometric cone whose tip would touch the supporting plaque simultaneously with a cone of the refractory material being investigated.

Rebonded Brick: Fired refractory bricks or other shapes made predominantly or entirely from fused grain.

Refractoriness: The capability of maintaining a desired degree of chemical and physical identity at high temperatures and in the environment and conditions of use. Synonym for pyrometric-cone equivalent.

Refractoriness under Load: The resistance (softening) of a refractory subjected to a constant load (normally 0.2 N/mm^2) under progressively rising temperature (5 °C/min). Temperatures corresponding to characteristic deformations are determined.

Reheat Change: The permanent changes in length (or volume) taking place in a fired refractory upon reheating.

Secondary Expansion: The property exhibited by some fireclay and high-alumina refractories of developing permanent expansion at temperatures within their useful range.

Shrinkage: The fractional reduction in dimensions or volume of a material or object when subjected to drying, calcining, or firing (sintering).

Sintering: A heat treatment that causes adjacent particles of material to cohere at a temperature below that of complete melting. Used especially when firing dense fine-grained ceramics.

Slip: A suspension or slurry (concentrated suspension) of normally finely divided ceramic materials in a liquid.

Spalling: The cracking or rupturing of a refractory material or unit caused by differential expansion due to thermal shock (thermal spalling) or crystalline inversions (structural spalling)

Tempering: Preheating at 200–600 °C of pitch- or resin-bonded refractory bricks and prefabricated blocks from castables to eliminate volatile materials or of chemically bonded bricks to gain strength.

Texture: Macroscopic relationship existing between the various shapes and sizes of grains and pores.

Thixotropy: A reversible process by which certain materials in suspension (e.g., castables) become more fluid on vibrating or mixing and less fluid or stiff on standing.

Workability: Measure of the facility with which plastic refractories can be rammed or pounded into place.

Works Cited

Alper, A. M. (Ed.) (1970), *High Temperature Oxides, Magnesia, Lime, and Chrome Refractories,* New York: Academic.

Am. Ceram. Soc. Bull. (1992), **71,** 799–824.

Annual Book of ASTM Standards (1995), Vol. 15.01, *Refractories, Carbon and Graphite Products, Activated Carbon.*

ASTM C201-93, Standard test method for thermal conductivity of refractories.

Begley, E. R., Herndon, P. O. (1971), in: A. M. Alper (Ed.), *High Temperature Oxides,* Part 4, New York: Academic, pp. 185–208.

Brachet, F. C., Avis, R., Clavaud, B. (1992), in: *Unitecr '91 Congress,* Düsseldorf: Verlag Stahleisen, pp. 264–266.

Carswell, G. P. (1977), *Refractories J.,* **52** (6), 7–19.

Cichy, P. (1990), in: *Alumina Chemicals Science and Technology,* Westerville, OH: American Ceramic Society Inc., pp. 393–426.

Cooke, T. F. (1991), *J. Am. Ceram. Soc.* **74,** 2959–2978.

Coope, B. M., Dickson, E. M. (Eds.) (1991), *Raw Materials for the Refractory Industry,* IM Refractories Survey, Metal Bulletin Ltd., London.

Dal Maschino, R., Fabbi, B., Fiori, C., (1988), *Ind. Ceram. (Faenza Italy)* **8** (3), 121–126.

Das, T. K., Jeschke, P. (1975), *Ber. Dtsch. Keram. Ges.* **52,** 126–130.

Dietrichs, P. (1987), *Silikattechnik* **38,** 349–353.

DIN 51064 (1981), Bestimmung der Druckfeuerbeständigkeit (DFB) an feuerfesten Steinen.

ENV 1402-1 (1995), (European Standard). Unshaped refractory products. Part 1: Introduction and definitions.

Fickel, A., Völkel, W. (1988), *InterCeram, Aachen Proceedings,* pp. 38–40.

Fletcher, W. (1989), *Steel Metals Mag.* **27** (5), 389–390.

Glazman, M. S., Landa, Ya. A., Litovskii, E. Ya., Puchkelevich, N. A. (1989), *Refractories (Eng. Trans.)* **30** (5), 312–319.

Hasselmann, D. P. M. (1978), *Ceramurgia Int.* **4,** 147–150.

ISO 528 (1983), Refractory products—Determination of pyrometric cone equivalent (refractoriness).

ISO/R 836 (1968), Vocabulary for the refractories industry. Trilingual edition.

ISO 1109 (1975), Refractory products—Classification of dense shaped refractory products.

ISO 1893 (1989), Refractory products—Determination of refractoriness under load (differential—with rising temperature).

ISO 2245 (1990), Shaped insulating refractory products—Classification.

ISO 3187 (1989), Refractory products—Determination of creep in compression.

ISO 5013 (1985), Refractory products—Determination of modulus of rupture at elevated temperatures.

ISO 8894-1 (1987), Refractory materials—Determination of thermal conductivity—Part 1: Hotwire method (cross-array).

Jeschke, P., Bertling, H., Breidenbech, D. (1989), in: *Unitecr '89 Proceedings*, Vol. 1, Westerville, OH: American Ceramic Society Inc., pp. 263–283.

Kobayashi, M., Nishi, M., Miyamoto, A. (1982), *Taikabutsu Overseas* **2** (2), 5–13.

Konopicky, K., Routschka, G., Hagemenn, L. (1971), *Archiv Eisenhüttenwesen* **42,** 433–437.

Kopanda, J. E., MacZura, G. (1990), in: *Alumina Chemicals Science and Technology,* Westerville, OH: American Ceramic Society Inc., pp. 171–183.

Kriechbaum, G. W., Gnauk, V., Routschka, G. (1994), *Stahl u. Eisen Speciel Okt.,* Düsseldorf: Verlag Stahleisen, pp. 150–159.

Kriegesmann, J. (Ed.) (1989), *DKG-Technische Keramische Werkstoffe,* Köln: Lose-Blatt-Sammlung, Wirtschaftsdienst.

Le Doussal, H. L., Prieur, C. (1989), *Ind. Ceram.* no. 837, 255–262.

Lepère, K. E., Overkott, E., Beckmann, R., Eschner, L. (1992), in: *Unitecr '91 Congress,* Düsseldorf: Verlag Stahleisen, pp. 450–453.

Pask, J. A. (1990), in: *Ceramic Transactions,* Vol. 6, *Mullite and Mullite Matrix,* Westerville, OH: American Ceramic Society Inc., pp. 1–13.

Piel, K. T., Wilkening, S., Lechthaler, K., Santowski, K. (1988), *InterCeram Special Issue, Aachen Proceedings,* pp. 22–25.

Plibrico Japan Comp. (1984), *Technology of Monolithic Refractories,* Tokyo.

Rathner, R., Knauder, J., Weissensteiner, H. (1992), in: *Unitecr '91 Congress,* Düsseldorf: Verlag Stahleisen, pp. 12–18.

Refractory Materials, Recommendations (1990), Brussels: Féderation Européene des Fabricants de Produits Réfractaires (PRE).

Röttenbach, R., Pögl, P. (1989), *InterCeram, Aachen Proceedings,* pp. 72–74.

Routschka, G. (1981), *Ziegelind. Intern.,* pp. 690–700.

Routschka, G. (1988), *InterCeram* **37,** 24–28, 33.

Smith, R. D. (1990), in: *Alumina Chemicals Science and Technology,* Westerville, OH: American Ceramic Society Inc., pp. 385–394.

Stahl-Eisen-Werkstoffblatt 912 (1984), *Tonerdereiche Steine,* Düsseldorf: Verlag Stahleisen.

Stahl-Eisen-Werkstoffblatt 915 (1984), *Schamottesteine für Hochöfen und Winderhitzer,* Düsseldorf: Verlag Stahleisen.

Stahl-Eisen-Werkstoffblatt 917 (1984), *Schamottesteine für allgemeine industrielle Einsatzzwecke,* Düsseldorf: Verlag Stahleisen.

Tabata, K., Yamamura, T., Katayama, M., Enoki, K. (1992), in: *Unitecr '91 Congress,* Düsseldorf: Verlag Stahleisen, pp. 46–53.

Taikabutsu Overseas (1989), *Recent Progress on Castable Refractories* **9** (1), 2–75.

Uchida, S., Tchikawa, K., Matsuo, A., Miyoshi, S. (1990), *Taikabutsu Overseas* **10** (2), 74–84.

VDI-Richtlinie 3128, Blatt 4 (1977), Induktive Erwärmung. *Physikalische Stoffeigenschaften nichtmetallischer Werkstoffe (Ofenbaustoffe),* Düsseldorf: VDI-Verlag.

Weaver, E. P., Tally, R. W., Engel, A. J. (1985), in: *Advances in Ceramics,* Vol. 13, *New Developments in Monolithic Refractories,* Westerville, OH: American Ceramic Society Inc., pp. 219–229.

Wecht, E. H. P. (1977), *Feuerfestsiliciumcarbid,* Wien: Springer-Verlag.

Yamaguchi, A., Tanaka, H. (1992), in: *Unitecr '91 Congress,* Düsseldorf: Verlag Stahleisen, pp. 32–38.

Young, R. G., Hartwig, F. J., Norton, C. L. (1964), *J. Am. Ceram. Soc.* **47,** 205–210.

Zanoli, A., DuVierre, G., Sertain, E. (1991), *Ceram. Eng. Sci. Proc.* **12,** 496–517.

Zednicek, W. (1988), *Silikattechnik* **39,** 385–390.

Zoglmeyer, G., Romai, D. (1991), *InterCeram* **40,** 81–89.

Further Reading

Brook, R. J. (1991), *Concise Encyclopedia of Advanced Ceramic Materials,* Oxford: Pergamon, p. 588.

Carniglia, S. C., Barna, G. L. (1992), *Handbook of Industrial Refractories Technology. Principles, Types, Properties and Application,* Park Ridge, NJ: Noyes, p. 627.

Ceramics and Glasses (1991), *Engineering Materials Handbook,* Vol. 4, The Materials Information Society, p. 1217.

Deutsche Gesellschaft Feuerfest- und Schornsteinbau e.V. (1994), *Feuerfestbau. Stoffe-Konstruktions-Ausführung,* Essen: Vulkan Verlag, p. 358.

Didier-Werke AG (1990), "Refractory Materials and Their Properties," company brochure, Wiesbaden, p. 104.

Fisher, R. E. (Ed.) (1985), *Advances in Ceramics,* Vol. 13, *New Developments in Monolithic Re-*

fractories, Westerville, OH: American Ceramic Society, p. 424.

Granitzki, K. E., Krönert, W., Müller, E. (1989), *Feuerfeste Stoffe im Gießereibetrieb,* Düsseldorf: Gießerei-Verlag, p. 247.

Harbison-Walker (1992), *Modern Refractory Practice,* 5th ed., Pittsburgh: Harbison Walker Refractories.

Hart, L. D. (Ed.) (1990), *Alumina Chemicals Science and Technology Handbook,* Westerville, OH: American Ceramic Society Inc., p. 617.

O'Bennon, L. S. (1984), *Dictionary of Ceramics Science and Engineering,* New York: Plenum, p. 303.

Proceedings Unitecr '95 Congress [Unified International Technical Conference on Refractories (1995), Fourth Biennial Worldwide Conference on Refractories], Kyoto, Nov. 19–22, 1995, Vols. 1–3, Düsseldorf: Verlag Stahleisen.

Routschka, G., Granitzki, K. E. (1993), *"Refractory Ceramics,"* in: *Ullmann's Encyclopedia of Industrial Chemistry,* Vol. 23, Weinheim: VCH, pp. 1–48.

Schulle, W. (1990), *Feuerfeste Werkstoffe. Feuerfestkeramik—Eigenschaften, prüftechnische Beurteilung, Werkstofftypen,* Leipzig: Deutscher Verlag für Grundstoffindustrie, p. 494.

Taikabutsu Overseas, Quarterly English Journal of The Technical Association of Refractories, Japan, Tokyo, Vol. 1, 1989, and current.

REFRACTORY MATERIALS, ELECTRONIC APPLICATIONS OF

NOBUYOSHI KOBAYASHI, *Central Research Laboratory, Hitachi, Ltd., Japan*

Introduction 309
1. **Refractory-Metal Films** 310
1.1 Formation Method 310
1.1.1 Magnetron Sputtering 310
1.1.2 Chemical Vapor Deposition (CVD) 311
1.1.3 Surface Cleaning 312
1.2 Characterization of Film Properties 312
1.2.1 Film Properties 312
1.2.2 Characterization Method 313
1.3 Patterning 313
1.3.1 Dry Etching 313
1.3.2 Etching Using Wet Chemicals 314
2. **Applications to Si ULSIs** 314
2.1 Tungsten or Molybdenum Gate Metallization for MOS LSIs 314
2.1.1 Requirements for MOS Gate Material 314
2.1.2 Tungsten or Molybdenum Gate Process 315
2.2 Refractory-Metal Barrier for Aluminum Interconnection 317
2.2.1 Barrier Properties 317
2.2.2 Layered Aluminum Interconnects with Refractory Metals 318
2.3 Contact Filling using Chemical Vapor Deposition of Tungsten 318
2.3.1 Requirements for Contact Filling 318
2.3.2 Selective Chemical Vapor Deposition of Tungsten 320
2.3.3 Blanket Chemical Vapor Deposition of Tungsten 323
2.4 Solid-Phase Reaction of Refractory Metal with Silicon 323
2.4.1 Silicide Formation 323
2.4.2 Use of Silicides in MOS Source and Drain Metallization 324
3. **Application to GaAs MESFETs** 324
4. **Acknowledgments** 325
Works Cited 325

INTRODUCTION

Industrial applications of refractory materials are very extensive, and they range from propulsive systems through electric power generation systems to nuclear systems. The maximum temperature at which a material can be used is limited by its melting point, and refractory materials that have high melting points are widely used in these applications. Besides the above applications, refractory materials have begun to be widely used in microelectronics. In particular, their use in silicon large-scale integrated circuits (Si LSIs) has become the most pervasive application. In microelectronics, refractory metals are used in the form of "thin" films, and so their technological development has been quite different from applications as bulk material. This article describes electronic applications of selected refractory metals mainly in the field of metallization for Si LSIs.

As was pointed out by Moore (1980), silicon device integration has continued to increase over the last 40 years, and it now achieves 100 megabit to gigabit integration of devices on a single silicon chip. The level of megabit to gigabit integration is usually called ULSI (ultralarge-scale integration). With this rapid technological progress in ULSI, device interconnect technology is becoming more and more significant in determining integrity and performance. Conventional interconnect technology uses aluminum (Al) for the conductive material and silicon dioxide (SiO_2) for the dielectric. The

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

SiO_2 film is still used for the dielectric, but new conductive materials and their processes have been developed in order to obtain reliable high-performance devices. Use of refractory metals has made great progress in metallization of Si ULSI. Figure 1 shows schematically the cross section of multilevel metallization in Si ULSI. The demand for closely packed devices has pushed the metallization system into such a multilevel structure, and in some cases, more than five-level metallization has been realized. This figure shows three-level metallization. Important features of refractory materials are the resistance to high-temperature processes, mechanical hardness, barrier properties to Al diffusion, and relatively low electrical resistivity. As shown in this figure, refractory metals are used for gate electrodes of MOSFETs (metal-oxide-semiconductor field-effect transistors), source and drain metallization, interconnects, contact and via filling, barrier metals to Al diffusion, and so on. Among refractory metals, tungsten (W) has low electrical resistivity and is widely used for these applications. This article discusses these electronic applications of selected refractory metals in the following aspects:

1. thin-film formation and characterization,
2. applications to Si ULSIs, and
3. applications to GaAs MESFETs.

1. REFRACTORY-METAL FILMS

In Si ULSIs, typical dimensions of devices are reduced to below 1 micron (μm), and films used for the device fabrication become very thin (e.g., less than a few hundred nm). Under the usual deposition conditions used in manufacture, thin refractory-metal films are mostly polycrystalline or amorphous rather than single-crystal. Refractory compound material is also used. It is very important for practical uses to determine how to form a refractory thin film and to characterize it.

1.1 Formation Method

Table 1 shows vapor pressures of refractory metals as a function of temperature. In practice, vapor pressures of 1–10 Pa are necessary for evaporation of refractory metals with deposition rates more than 0.5 μm/min. As shown in this table, refractory metals have high boiling points, and they are not usually deposited using the conventional resistance-heat evaporation or electron-beam evaporation for ULSI applications. Instead, sputtering or chemical vapor deposition (CVD) has been used. In particular, magnetron sputtering, where large plasma density is sustained by magnets for high deposition rate, is widely used in manufacture. A metal-film formation method that uses a gas-phase chemical reaction [CVD (Chemical Vapor Deposition), *q.v.*], has been developed to obtain the deposition of films with good step coverage (conformal deposition) onto a substrate with complicated structure.

1.1.1 Magnetron Sputtering Thornton and Penfold (1978) have proposed a magnetron sputtering method. In practice, a planar magnetron cathode is used, in which magnets are attached to the back side of the

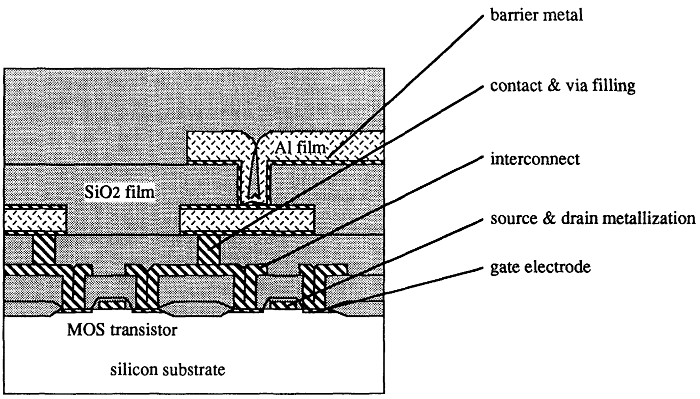

FIG. 1. Cross-sectional view of Si ULSI multilevel metallization.

Table 1. Vapor pressures of refractory metals at various temperatures.

Material	Vapor pressure (Pa)			
	1.33×10^{-4}	1.33×10^{-3}	1.33	133
	Temperature (K)			
Titanium, Ti	1508	1726	2020	2464
Chromium, Cr	1229	1403	1637	1968
Cobalt, Co	1095	1247	1454	1744
Nickel, Ni	1061	1207	1402	1681
Zirconium, Zr	1975	2265	2662	3227
Molybdenum, Mo	2090	2386	2797	···
Palladium, Pd	1336	1539	1820	2246
Tantalum, Ta	2507	2862	3329	3978
Tungsten, W	2667	3029	3502	4141
Rhenium, Re	2480	2844	3333	4036

sputtering target. It is important to obtain a high-purity refractory-metal film, and so sputtering is done with a high deposition rate at low background pressures. With advances in vacuum pumping technology, base pressures lower than 10^{-6} Pa are achieved in commercially available sputtering machines. In the sputtering, an inactive gas such as argon (Ar) is used to form the plasma, and a small number of gas atoms (around 1 at.%) can be incorporated in the metal films. Impurities have a large effect on the electrical resistivity and mechanical stress of the films.

In reactive sputtering, an active gas is intentionally added to Ar to deposit chemical compounds of refractory metals. Refractory nitrides such as titanium nitride (TiN) and zirconium nitride (ZrN) can be deposited by reactive sputtering in nitrogen (N_2) atmosphere. In addition, refractory oxides such as tantalum oxide (Ta_2O_5) and tungsten oxide (WO_3) can be obtained by adding oxygen (O_2).

1.1.2 Chemical Vapor Deposition (CVD) Chemical vapor deposition has been used in microelectronics processes to form Si and SiO_2 films. However, it has only recently been used for metal-film deposition. This is because technological problems such as the choice of suitable source gases for CVD and the design of the CVD reactor are rather difficult compared with Si or SiO_2 CVD using silane (SiH_4) or silane and oxygen (O_2). Metal precursors used for the CVD of refractory metals were summarized by Bernard *et al.* (1989). Among them, CVD of refractory metals such as tungsten (W), tungsten silicide (WSi_2), molybdenum (Mo), molybdenum silicide ($MoSi_2$), titanium nitride (TiN), and titanium silicide ($TiSi_2$) have been extensively investigated, and some are used for Si ULSI applications. An example of a CVD reactor, which is used for the CVD of these materials, is shown in Fig. 2. This figure shows a bell jar reactor that is shown in Sherman (1987). In this configuration, source gases enter at the top of the bell jar and flow between the bell jar and susceptor before exiting at the bottom. Silicon wafers are set on the susceptor. The susceptor can be heated by halogen lamps or resistance heaters located inside the bell jar. In CVD, a gas-phase chemical reaction proceeds at low pressures on a hot substrate surface the temperature of which is controlled. Typically, CVD temperature ranges from 200 to 700 °C, and total pressure from 10 to 100 Pa. The CVD reactor wall is usually cooled by water to prevent metal deposition on the wall. Moreover, some commercially available CVD

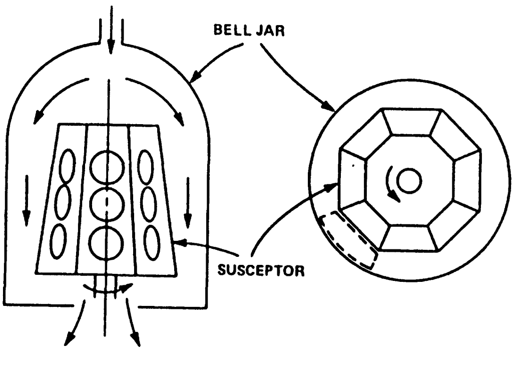

FIG. 2. Schematic figure of a bell jar CVD reactor.

machines are equipped with a plasma cleaning system for dry etching the metal deposited on the wall. With an increase of wafer size (e.g., 8-inch wafers currently in production), the single-wafer system has been widely accepted for CVD machines. A single-wafer reactor is schematically shown in Fig. 3, which is taken from Schmitz (1992). Film properties depend strongly on CVD parameters: temperature, pressure, and gas flow of source gas.

1.1.3 Surface Cleaning Film properties also depend on the physical and chemical conditions of the surface: surface roughness, surface oxidation, and contamination. Thus, it is important to clean the surface before CVD to achieve reliable metal deposition. Native oxide on a Si surface is usually removed by an HF dip. To remove native oxide on a metal surface (e.g., oxide on an aluminum surface), *in situ* Ar sputter cleaning is carried out before sputtering deposition of metal. Much work has been done to characterize refractory metals in terms of CVD parameters and surface conditions.

1.2 Characterization of Film Properties

The physical and chemical properties of thin films often differ from those of the bulk material. Characterization of thin films is important for scientific research as well as for electronic applications. General descriptions of characterization have appeared in textbooks—for example, Sze (1983)—and so this section focuses on refractory thin films.

1.2.1 Film Properties Electrical resistivity, Schottky-barrier height, mechanical stress, mass density, film composition, and microstructure (preferred orientation, grain size, subgrain structure, etc.) are important factors in determining the effectiveness of refractory metals for ULSIs. Low electrical resistivity is required for refractory thin films to be used for gates and interconnections. Low Schottky-barrier height to both n^+ and p^+ Si is preferable for Ohmic contact. Table 2 shows typical values of electrical resistivity and Schottky-barrier height of bulk refractory metals. Resistivity and barrier-height data are from the American Institute of Physics Handbook (Gray, 1972) and Rhoderick (1978). The resistivities of Mo and W are the lowest among refractory metals, and so they are widely used for metal interconnects. The electrical resistivity of refractory thin films is larger than that of the bulk material (almost twice in W and Mo) even if sputtering deposition is done at very low pressures ($<10^{-6}$ Pa). This might be because defects in the crystalline lattice of thin refractory-metal films, which are deposited at temperatures much lower than the melting point, are more numerous than those of the bulk material. Mechanical tensile stress of refractory metals is generally high, for example, greater than 1 GPa. So peeling of a refractory-metal film from the substrate is a concern in microelectronics processes. Film stress also has an influence on device characteristics such as hot-electron reliability of MOS devices and leakage current at junctions.

In general, the contact between refractory metal and Si substrate is characterized as rectifying or Ohmic, depending on dopant concentration of the Si substrate. In Ohmic contact, the most important feature is low

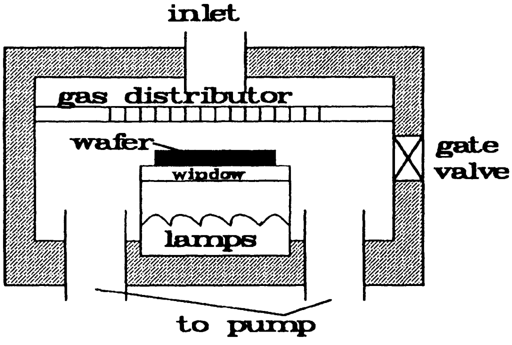

FIG. 3. Schematic figure of a single-wafer reactor.

Table 2. Resistivities and Schottky-barrier height of refractory metals.

Material	Electrical resistivity ($\mu\Omega$ cm) (bulk)	Schottky-barrier height (eV) (to n^+ Si)
Ti	47	0.5
Cr	12.9	0.6
Co	5.8	···
Ni	7.04	0.65
Zr	45	···
Mo	5.33	0.60
Hf	30.6	···
Ta	13.1	0.58
W	5.44	0.65
Re	18.6	···

resistance. For this purpose, Si substrate is heavily doped with arsenic (Ar) or phosphorus (P) for n^+ contact and with boron (B) for p^+ contact. In the case of heavy doping ($>10^{19}$ cm^{-3}), Sze (1981) discussed that tunneling current through the Schottky barrier becomes important, and the specific contact resistance (R_c) may be written as

$$R_c \sim \exp[4\pi\epsilon m^*/h(\Phi_B/\sqrt{N_D})], \qquad (1)$$

where ϵ is the dielectric permittivity of silicon, N_D is the doping concentration, h is Planck's constant, m^* is the effective mass of the charge carrier (electron or hole), and Φ_B is the Schottky-barrier height. In order to obtain low specific resistivity, refractory metal having low Φ_B is preferable.

1.2.2 Characterization Method Characterization of film properties is carried out using refractory-metal films that are formed on a substrate. In Si LSI processing, refractory-metal films are usually deposited on silicon wafers. Electrical resistivity is calculated from the sheet resistance of the film multiplied by its thickness according to the formula

$$\rho = R_s t, \qquad (2)$$

where ρ (Ω m) is electrical resistivity, R_s ($\Omega/\square$) is sheet resistance, and t (m) is metal thickness. Sheet resistance is usually measured by a four-point terminal method, and film thickness is measured from the step height that is formed by removing a metal film by means of wet chemicals and an etching mask. Other electronic properties such as electrical contact resistivity are measured by using a series-chain-contact pattern device or a Kelvin pattern device, as was reported by Loh *et al.* (1985). Mechanical stress is measured from the curvature of the substrate on which the refractory-metal film is deposited. A formula for the stress, developed assuming the substrate to be a cubic plate, is expressed as

$$\sigma = (\delta/r^2)[E/3(1 - \nu)](t_s^2/t), \qquad (3)$$

where σ is stress (N/m^2), δ is disk deflection (m), ν is substrate Poisson's ratio, E is substrate Young's modulus, t_s is substrate thickness (m), t is metal thickness, and r is disk radius (m).

In binary-metal systems such as titanium–tungsten (Ti–W) and titanium nitride (TiN), film composition is strongly related to physical and chemical film properties. Film composition is usually determined by Rutherford backscattering. The profile of the impurity distribution in films is measured by sputtering Auger electron spectroscopy. Impurities affect the microstructures and the electrical properties of films. In particular, oxygen impurity is easily absorbed in refractory-metal films, and this increases electrical resistivity. A little impurity that cannot be detected by Rutherford backscattering and Auger electron spectroscopy (less than 0.1 at.%) in films is measured by secondary-ion mass spectroscopy. Crystal orientation is determined by x-ray diffraction, and grain size is estimated from width of the x-ray diffraction peak. The microstructure of films is usually observed by SEM (scanning electron microscopy) and TEM (transmission electron microscopy). The cross section of the device structure can be observed by TEM by thinning a sample that is cut from a test wafer. Nanometer-level surface roughness is evaluated using scanning tunneling microscopy and atomic-force microscopy.

1.3 Patterning

1.3.1 Dry Etching In microelectronics applications, refractory-metal films are patterned to fine lines using photolithography and dry etching. Dry etching is a plasma-assisted pattern transfer technique using partially ionized gases consisting of ions, electrons, and neutral in a low-pressure ($\sim 10^{-2}$–10^{-3} Pa) electric discharge. In the dry-etching process, other material such as SiO_2 film and masking material exist under a refractory-metal film, and so selectivity of etching is a parameter of considerable importance. Typical etching gases used for dry etching of refractory metals are shown in Table 3. The

Table 3. Etching gases used in dry etching of refractory metals.

Material	Gases used for dry etching
Cr	$CCl_4(+O_2)$, Cl_2
Ti	CCl_4, BCl_3, $SiCl_4$
Mo	$CF_4(+O_2)$, $CBrF_3$
W	$SF_6(+O_2)$, $CF_4(+O_2)$, NF_3

vapor pressures of tungsten (W) and molybdenum (Mo) fluorides are relatively high, so that fluorine-based chemistry is used in dry etching for them. Selectivity values of 3–5 to SiO_2 can be obtained using $CF_4 + O_2$ or SF_6, although it strongly depends on the etching conditions. Dry etching of titanium (Ti) and titanium compounds uses chlorine-based chemistry such as CCl_4, BCl_3, and $SiCl_4$. On the other hand, there are no volatile compounds for nickel (Ni) and cobalt (Co), and they are physically etched using ion milling or sputtering. However, it is difficult for these metals to be used in LSI because etching using ion milling or sputtering shows little selectivity to an underlying SiO_2 film. However, Ni and Co have been investigated as metals for self-aligned silicided source and drain, where these metals are removed by wet chemicals listed in Table 4. The self-aligned silicide technology is described in Sec. 2.4.

1.3.2 Etching Using Wet Chemicals Etching of refractory metals using wet chemicals is not suitable for fine patterning. However, it is a very convenient way to remove an unnecessary metal film, or to form a relatively large pattern. As is shown in Sec. 2.4.2, wet etching of refractory metal is used for self-aligned silicide formation. In addition, several kinds of wet chemicals are used to clean Si wafers during device fabrication, and so it is important to know the resistance of refractory metals to wet chemicals when they are used in ULSI devices. Wet chemicals that are often used for patterning refractory metals in microelectronics fabrication are shown in Table 4. It should be noticed that Ti and Ta can be dissolved in HF solutions when these metals are used in the Si ULSI process.

Table 4. Wet chemicals for refractory metals.

Material	Chemical solution
Cr	$HCl/H_2O_2 = 1/1$
Co	$HCl/H_2O_2 = 1/1$
Ni	$HCl/H_2O_2 = 1/1$
Ta	$HF/HNO_3 = 1/1$
Ti	$HF/HNO_3/H_2O = 1/1/50$, $NH_4OH/H_2O_2 = 1/2$
Mo	$H_3PO_4/HNO_3/CH_3OOH/H_2O = 5/2/4/150$ $HNO_3/H_2SO_4/H_2O = 1/1/3$
W	$KH_2PO_4/KOH/K_3Fe(CN)_6/H_2O = 34/13.4/33$ (in weight) H_2O_2

2. APPLICATIONS TO Si ULSIs

2.1 Tungsten or Molybdenum Gate Metallization for MOS LSIs

2.1.1 Requirements for MOS Gate Material We use a conductive thin film for the gate electrode, which is a key component of MOS transistors. It is used to control the current under the gate by applying voltage to it. The details of MOS device operation are explained in textbooks—for example, Muller and Kamins (1986). In terms of the electrical properties of gate material, its electrical resistivity and work function are important factors for MOS device operation. Saraswat and Mohammadi (1982) calculated the *RC* time delay for gate and interconnection in ULSI circuits and pointed out the importance of low-resistivity gate and interconnection. The work function of the gate material determines the threshold voltage of the MOS device, where the current under the gate begins to flow. On the other hand, the gate material is required to have these characteristics to make it suitable for the manufacturing processes: stable interfacial electrical and mechanical properties, fine patternability, resistance to high-temperature process (800–900 °C), use as a mask for ion implantation, and so on. Figure 4 schematically shows cross sections of MOS transistors. It describes the method of fabricating MOS transistors. Gate material is deposited on the Si substrate, whose surface is covered by thin Si oxide (gate oxide), and patterned to form a gate electrode. Highly doped regions (source and drain) are formed on the substrate by ion implantation using the gate material as a mask. This is followed by annealing at 800–950 °C to activate implanted dopants. These processes constitute what is called self-aligned gate technology. In the self-aligned gate processes, the gate material is required to stop the accelerated ions and to have thermal stability during annealing.

Conventionally, polycrystalline silicon (poly Si) has been used for the gate material, because it satisfies these requirements quite well. However, a gate material with a resis-

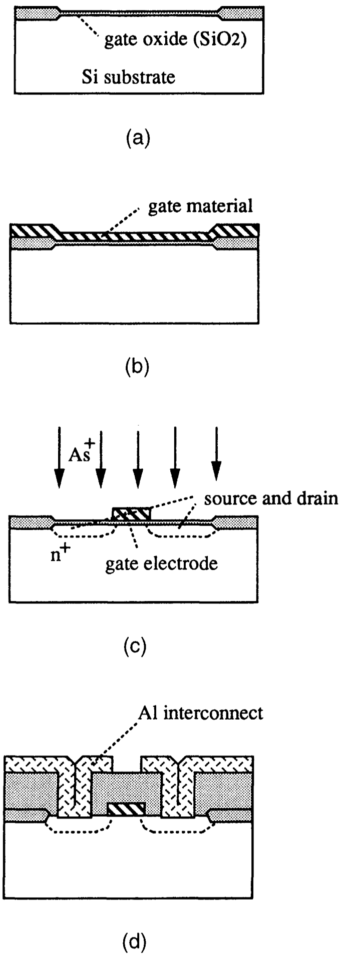

FIG. 4. Fabrication process flow of MOS transistors using the self-aligned gate technology.

tivity lower than that of poly Si is greatly needed to reduce the *RC* delay of the device as the dimensions of the gate electrode decrease. Low-resistivity refractory-metal silicide has therefore been used in manufacturing. In this application, a layered gate structure where silicide is put on poly Si is used because of its compatibility with the poly Si gate process. This gate structure is called the "polycide" gate. Wang *et al.* (1982) reported application of a polycide gate using titanium silicide ($TiSi_2$) for low-resistivity gate and interconnection. The polycide gate is used in manufacture; however, refractory metals such as W and Mo have gathered much interest for research and development as an alternative to poly Si and silicides because of their low resistivity.

2.1.2 Tungsten or Molybdenum Gate Process

Tungsten (W) and molybdenum (Mo) are the most promising candidates for low-resistivity gate materials, and they have been studied for Si LSI applications. In Table 5, the electrical resistivities of refractory-metal films are compared with those for poly Si and silicides. Brown *et al.* (1971) investigated the self-aligned Mo gate technology, although impurities were doped into the Si using doped glass diffusion sources instead of ion implantation. Iwata *et al.* (1984) compared Mo and W, and pointed out that W is more suitable for gate-electrode material because of the thermal stability of the W/SiO_2 interface. There are, however, several problems in using W (or Mo) for gate electrodes.

Generally, the adhesion between W (or Mo) and gate SiO_2 is not strong compared with that between poly Si and SiO_2. In most cases, the W (or Mo) is deposited by sputtering because it cannot be stably formed on SiO_2 by CVD. Therefore it is necessary to reduce mechanical film stress by optimizing the sputtering conditions. Figure 5 shows the internal stress change of Cr, Mo, Ta, and Pt films as a function of argon pressure during sputtering, as reported by Hoffman and Thornton (1982). Although the threshold Ar pressures, where internal stress changes from compressive to tensile, differ among these metals, the film stress changes from compressive to tensile as Ar pressure is increased. At low Ar pressures, the mean free path is a few centimeters and Ar ions easily hit the substrate surface to be incorporated in the film. The compressive stress is thought to come from interstitial argon incorporation in the W lattice. Yamamoto *et al.* (1987) reported that the film stress has an influence

Table 5. Comparison of film resistivities for MOS gate materials.

	Material	Thin-film resistivity ($\mu\Omega$ cm)
Silicon	Phosphorus-doped poly Si	700–1000
Refractory silicide	WSi_2	70–120
	$MoSi_2$	60–110
	$TiSi_2$ (after annealed at 900–1000 °C)	15–30
Refractory metal	W	8–12
	Mo	7–11

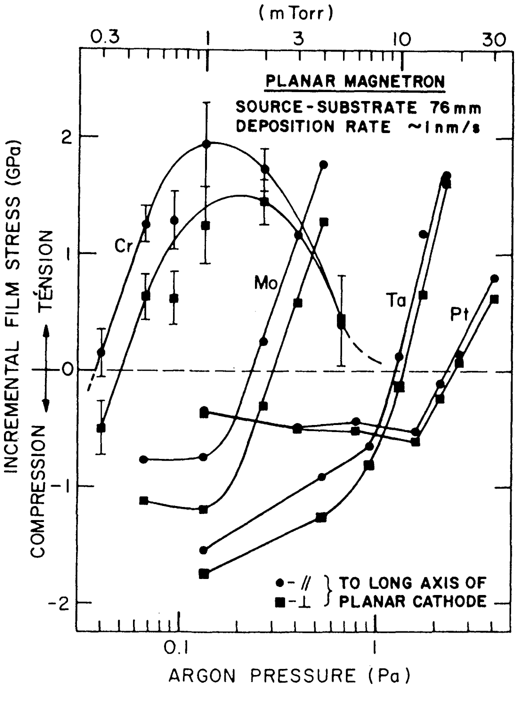

FIG. 5. Mechanical stresses as a function of Ar pressure for Cr, Mo, Ta, and Pt metal films sputter deposited from dc planar magnetron source onto glass at 1 nm/s to thickness of 50–350 nm.

on the hot-electron reliability of MOS devices. Thus, the film stress of W (or Mo) is controlled to be nearly zero by optimizing the argon pressure and other sputtering conditions.

The second problem is poor mask properties of W (or Mo) film against ion implantation to form source and drain regions. This is due to the fine columnar structure of W films, although the relatively large W atom itself has a large stopping power against the accelerated ions. In addition, ion channeling occurs through (110)-oriented W films to the Si substrate. W (or Mo) has the bcc (body-centered cubic) structure, and so its preferred orientation is (110). In practice, a 10–50-nm-thick SiO_2 film is deposited onto the W gate to prevent ion penetration through the W film.

Oxidation of W (or Mo) is also a serious problem. In the course of Si LSI fabrication, Si is oxidized in the presence of gate material. During oxidation, poly Si is covered with SiO_2, which is very stable and protective against contamination. On the other hand, W (or Mo) is easily oxidized to form unstable oxides. Figure 6 shows the temperature dependence of oxidation of sputter-deposited W films in dry O_2 ambient. SEM observation and step-height measurement of W oxides are utilized to determine the W oxide thickness. The oxidation rate is a thermally activated process, and it proceeds as $(\text{time})^{1/2}$. The oxidation rate is larger than that of poly Si by more than two orders in magnitude. Moreover, the W (or Mo) oxide is volatile at high temperatures (800–1000 °C), and its stoichiometry strongly depends on the oxidation conditions. To prevent W oxidation, an oxygen-free process is preferable. For example, measures should be taken to prevent oxygen incorporation from outside when W (or Mo) is annealed in inert gases using a quartz furnace tube. For the case that oxidation of Si is inevitable in the presence of gate material, Kobayashi *et al.* (1984) developed a wet hydrogen process, where an appropriate amount of H_2O is added to H_2, which has been developed to oxidize Si without oxidizing W.

The W/SiO_2 interface is thermally stable for obtaining reliable MOS characteristics. The interfacial reaction of refractory metal (M) with SiO_2 is expressed as

$$SiO_2 + xM = Si + M_xO_2. \tag{4}$$

This reaction favorably occurs when the

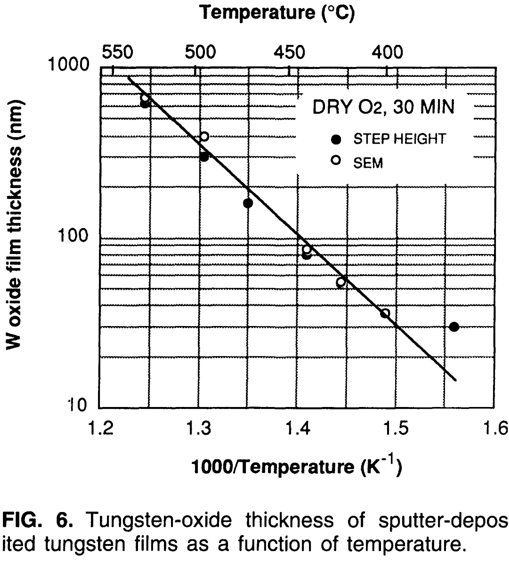

FIG. 6. Tungsten-oxide thickness of sputter-deposited tungsten films as a function of temperature.

Gibbs free energy change (ΔG) is negative. In other words, this reaction proceeds spontaneously when the absolute value of the Gibbs free energy of the metal oxide (M_xO_2) is greater than that of SiO_2. For W or Mo, ΔG is positive, and so W (or Mo) is not expected to react with SiO_2. No metallurgical reaction of the W/SiO_2 interface was observed by XPS up to 1000 °C annealing, as reported by Iwata *et al.* (1984). However, Ti or Ta reacts with SiO_2 because ΔG is negative. This reaction can degrade the electric characteristics of the gate oxide. On the other hand, it helps to reduce the contact resistance when metal can reduce the native oxide of the Si surface. Thus, Ti or Ta is used not for the gate electrode but for a low resistive contact. Although W (or Mo) is thermodynamically stable on SiO_2, its deposition method is important to obtain reliable MOS characteristics. As the gate oxide thickness decreases to less than 10 nm in the deep-submicron ULSI, damage induced by film deposition becomes serious. In sputtering deposition, plasma damage can cause degradation of the gate oxide. In CVD, the tungsten source gas (WF_6) reduces thin gate SiO_2 at high temperatures (400–500 °C). Further study is necessary in depositing W directly on thin gate SiO_2 in order to achieve good electrical performance.

Finally, avoidance of mobile-ion (Na^+, K^+, etc.) and radioactive-element (U, Th, etc.) contamination and damage induced by the sputtering deposition is critical in achieving reliable MOS characteristics. Mobile ions are easily diffused to the W/SiO_2 interface, and change the threshold voltage. May and Woods (1970) found a soft error mechanism in Si LSIs induced by alpha particles emitted in the radioactive decay of U and Th. Since then, these radioactive elements have been reduced to less than a few ppb level for interconnect and packaging materials. For this purpose, a high-purity sputtering target (99.999%–99.9999%) has been developed for use in microelectronics fabrication. Moreover, high-purity film formation by CVD has been investigated.

2.2 Refractory-Metal Barrier for Aluminum Interconnection

2.2.1 Barrier Properties Direct contact of Al to Si suffers from Al spikes caused by local dissolution of Si. The problem of Al spikes is solved by depositing Al mixed with Si. In practice, more than 1 wt.% Si is required to prevent the spikes. However, the excess Si present in the Al precipitates on the Si after a heating cycle at around 450 °C, which causes a non-Ohmic contact to Si, especially to n^+ Si. Thus, in current Al metallization, a refractory-metal barrier is used between the Al and the Si contact. Refractory-metal barriers have been used in microelectronics fabrication. Among these refractory-metal compounds, Ti–W and TiN are widely used. Ghate *et al.* (1978) reported application of Ti–W barrier metallization for Si integrated circuits. Ti–W is usually deposited by sputtering using a Ti–W compound target. On the other hand, TiN is usually deposited by reactive sputtering, and a CVD method has recently been developed.

It is known that TiN prevents Al diffusion into Si, and does not react with Al at up to 500 °C. This is why TiN is used as a barrier metal especially for high-temperature processing. Barrier properties strongly depend on the morphology and stoichiometry of the film. Pramanik and Jain (1991) reported that titanium nitride can exist as Ti_2N and TiN and nitrogen can form a solid solution with these compounds. Stoichiometric TiN has excellent barrier properties and low resistivity (50–100 $\mu\Omega$ cm). Thus, the nitrogen (N_2) gas pressure and applied voltage are key parameters in determining the film properties in reactive sputtering. However, the sputter-deposited metal barrier can be too thin to prevent the Al diffusion into Si at the bottom of a contact when the contact becomes deep. Thus, TiN CVD using $TiCl_4$ and NH_3 has been investigated for conformal deposition. Figure 7 shows the excellent step coverage of TiN film deposited in contacts with various sizes, as was reported by Yokoyama *et al.* (1991). The contact diameter is varied from 0.5 to 0.8 μm, and the substrate material is Si. In this CVD, the gas-flow ratio of $TiCl_4/NH_3$ and the CVD temperature are important parameters, and conformal TiN film can be deposited in deep contacts. High-temperature deposition (>700 °C) is necessary to reduce Cl incorporation in the films because incorporated Cl causes an increase of film resistivity. However, these deposition temperatures are too high to apply this CVD for Al metallization, and so low-temperature CVD using metal organic precursers such

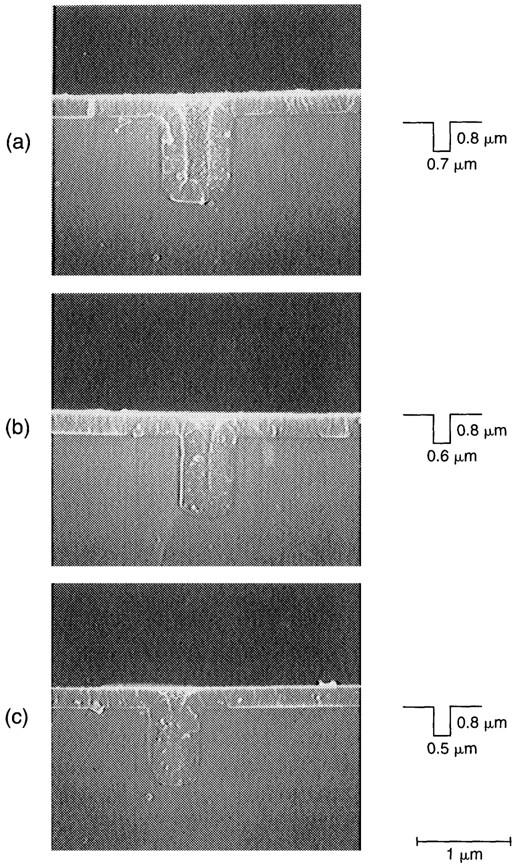

FIG. 7. SEM micrographs of 0.25-μm-thick CVD-TiN films in high-aspect-ratio contact holes with various diameters. (a) 0.7-μm diameter and 0.8-μm depth; (b) 0.6-μm diameter and 0.8-μm depth; and (c) 0.5-μm diameter and 0.8-μm depth.

as *tetrakis*-dimethylamino-titanium (TDMAT) and *tetrakis*-diethylamino-titanium (TDEAT) with NH_3 is being pursued. Moreover, a plasma-excited process using electron cyclotron resonance is being investigated.

Besides Ti–W and TiN, other refractory metals are often used as barrier metals. Holloway *et al.* (1992) reported that Ta or tantalum nitrogen alloy (Ta_2N) is a suitable barrier metal for Cu interconnect because of small reactivity with Cu and thermal stability in metal interdiffusion at the Cu/Ta structure.

2.2.2 Layered Aluminum Interconnects with Refractory Metals In device reliability, electromigration is a great concern for Al fine interconnects. Electromigration is observed as a material transport in conductive material, and it is caused by momentum transfer from electrons, moving under the influence of the electric field applied along the conductor. In addition, stress-induced migration (stress migration) has been found to cause failure of Al interconnects. In general, electromigration and stress-migration phenomena are related to Al self-diffusion along the grain boundaries, and easily occur in conductive materials with low melting points. Refractory metals have large electromigration and stress-migration resistance due to their high melting points, and this has been confirmed for W interconnects. To overcome these migration problems, layered Al interconnects with refractory metals have been widely used in microelectronics fabrication. In layered structures, refractory metal is deposited beneath the Al, and if necessary on top of the Al. Metals such as Ti–W, TiN, W, and refractory silicides are used. The use of the refractory metal beneath (and on) the Al causes an increase in resistivity because the refractory metal reacts with the Al to form high-resistivity compounds such as WAl_{12} during heat treatment at around 450 °C. The reaction of refractory metals with Al has been extensively studied.

2.3 Contact Filling Using Chemical Vapor Deposition of Tungsten

Chemical vapor deposition (CVD) of tungsten (W) achieves conformal deposition in deep contacts. Moreover, selective deposition on a conductive layer without deposition on an insulator layer can be achieved when appropriate CVD conditions are selected, and this has been widely investigated for filling submicron contacts.

2.3.1 Requirements for Contact Filling Figure 8 shows the aspect ratios of contact holes in Si LSI plotted as a function of the linewidth of the interconnects. The dotted line represents the dielectric layer thickness. Aspect ratio is defined as the dielectric layer thickness (= contact depth *b*) divided by contact diameter (*a*). The contact diameter decreases with scaling of device dimensions, whereas interdielectric layer thickness remains almost constant because a certain minimum thickness is necessary for electrical insulation between interconnects and for

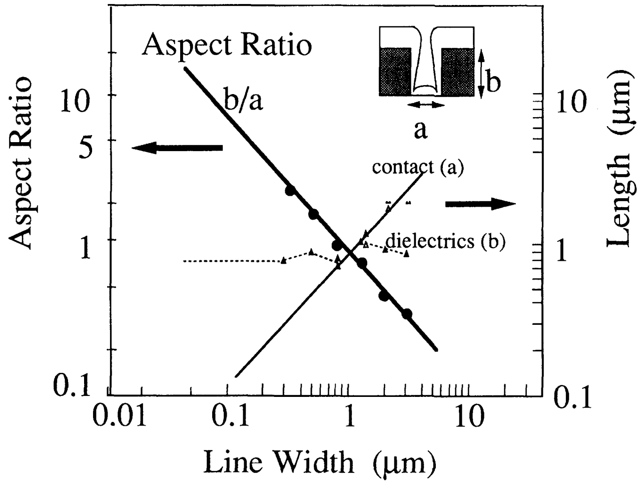

FIG. 8. Aspect ratio of contact holes as a function of linewidth of interconnect. Dotted line represents the dielectric thickness.

planarization of the multilevel metallization structure. As a result, the contact aspect ratio increases as the linewidth decreases, and it reaches around 2 in 0.5-μm interconnects. Conventional sputtering cannot achieve conformal metal deposition in high–aspect-ratio contacts. Figure 9 shows an SEM micrograph of W interconnects deposited by planar magnetron sputtering. The step coverage factor, defined as the ratio of film thickness at the bottom of the contact to that at the flat surface, is below 5% even at aspect ratio of 0.8, and it will cause electrical open-circuit failure of thin W interconnects when the aspect ratio exceeds 2. Low-pressure CVD is expected to achieve conformal deposition compared with the sputtering. For comparison, W interconnects formed by CVD are shown in Fig. 9 along with sputter-deposited W film. The step coverage factor is almost 100% for CVD-W interconnects. Because of conformal film deposition and high reliability against electromigration and stress migration, the CVD-W interconnect has begun to be used in the Si ULSI as an alternative for sputter-deposited Al interconnects. The film properties are compared between W and Al in Table 6. However, the characteristics of films deposited by CVD strongly depend on the CVD chemistry. The metals that can be deposited by CVD and that are widely used in microelectronics fabrication are restricted to W and WSi_2. CVD of Al, Cu, and TiN is now being widely investigated for LSI applications.

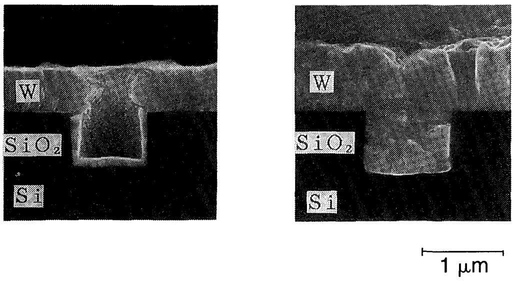

(a) (b)

FIG. 9. SEM micrograph of W interconnects deposited by planar magnetron sputtering in Si contact. (a) Sputter-deposited W interconnect and (b) CVD-W interconnect.

Figure 10 schematically shows the cross section of multilevel metallization using W-CVD. There are two approaches to fill the high–aspect-ratio contact and via using W-

Table 6. Comparison of tungsten with aluminum for interconnect application.

	W	Al (containing 1 wt.% Cu,Si)
Resistivity (μΩ cm)	8–10	3–3.5
Melting point (°C)	3380	660
Schottky barrier (eV) n^+ Si	0.5–0.6	0.65–0.75
Dry etching	SF_6	BCl_3
CVD gas	WF_6	TIBA[a], DMAH[b]

[a] TIBA: $Al(iC_4H_9)_3$
[b] DMAH: $(CH_3)_2AlH$

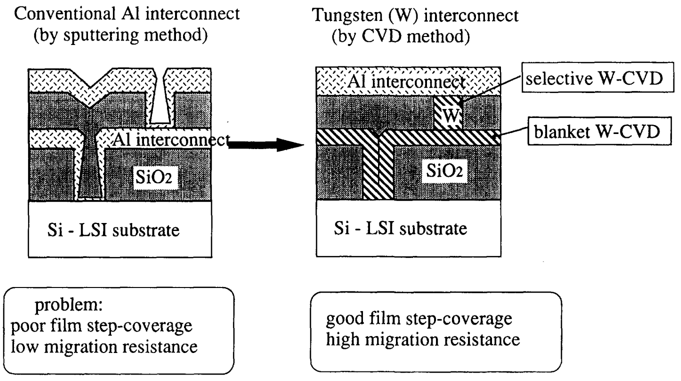
Conventional Al interconnect
(by sputtering method)
Al interconnect
SiO2
Si - LSI substrate
problem:
poor film step-coverage
low migration resistance
Tungsten (W) interconnect
(by CVD method)
Al interconnect
W
SiO2
Si - LSI substrate
selective W-CVD
blanket W-CVD
good film step-coverage
high migration resistance

FIG. 10. Cross-sectional view of multilevel metallization using tungsten CVD.

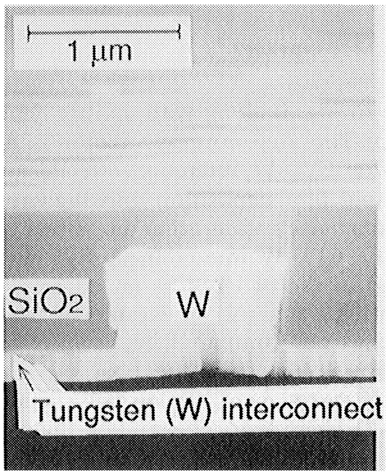
1 µm
SiO2
W
Tungsten (W) interconnect

FIG. 11. SEM micrograph of W selectively deposited on W interconnect. Tungsten is deposited using WF_6 and SiH_4.

CVD: One is selective deposition of W in contact and via holes, and the other is blanket deposition of W for W interconnect formation. In selective deposition, W is deposited only onto Si and metal without being deposited on dielectrics. Figure 11 shows SEM micrographs of W selectively deposited on W interconnects. After filling the contact and via with W, the interconnect metal (typically Al) is usually deposited by sputtering. In blanket deposition, the sputter-deposited metal film is usually deposited on the interdielectric SiO_2 film because adhesion between the W-CVD film and the SiO_2 is poor. Source gases for W-CVD that have been investigated are compared in Table 7. For Si LSI applications, tungsten hexafluoride (WF_6) has been mainly used because the vapor pressure of WF_6 is high enough for use of a conventional gas supply system and it causes relatively little damage to Si substrates. Tungsten hexachloride (WCl_6) has rarely been used in manufacture because it etches Si heavily and corrodes the stainless steel of the CVD reactor. Tungsten carbonyl [$W(CO)_6$] is a solid source with a low vapor pressure, but it is not suitable for high-quality W film deposition because of carbon incorporation in the film. When WF_6 is used for the source gas, a reducing gas for WF_6 is necessary because WF_6 itself is not decomposed at temperatures around 300–500 °C. It is necessary to point out that the characteristics of W-CVD depend strongly on the choice of reducing gas as well as the source gas. Thus, we should choose a gas to reduce WF_6 according to the application.

2.3.2 Selective Chemical Vapor Deposition of Tungsten Although the history of selective W-CVD is long, it has been exten-

Table 7. Source gases for W-CVD.

Source gas	Vapor pressure (Pa)	CVD temperature (°C)	Chemical reaction
WF_6	1.3×10^5 at 20 °C	250–500	$WF_6 + 3H_2 = W + 6HF$
WCl_6	(solid)	550–1250	$WCl_6 + 3H_2 = W + 6HCl$
$W(CO)_6$	(solid, 1330 at 60 °C)	400–500	$W(CO)_6 = W + 6CO$

sively investigated for Si LSI application for the past 10 years. Cuomo (1972) first reported selective W-CVD using WF_6 and H_2 for Si LSIs. The global chemical reaction is

$$WF_6 + 3H_2 \rightarrow W + 6HF. \tag{5}$$

McConica and Krishnamari (1986) studied growth kinetics and found that the growth rate of W film [R (nm/s)] is given by

$$R = 6.8 \times 10^4 \exp(-8000/T)P(WF_6)^0 P(H_2)^{0.5}, \tag{6}$$

where T is the deposition temperature (K) and P is the partial pressure (Pa). The origin of selectivity is thought to be the dissociative adsorption of H_2 on a conductive layer, which is expressed by

$$H_2 \rightarrow 2H. \tag{7}$$

Atomic hydrogen can react with adsorbed WF_6 to form W and HF. Broadbent (1984) explained that an activation energy in Eq. (6) corresponds to the surface migration of H on a growing W surface. When W is deposited on Si by this chemistry, the reaction of WF_6 with Si,

$$WF_6 + \tfrac{3}{2}Si \rightarrow W + \tfrac{3}{2}SiF_4, \tag{8}$$

precedes reaction (5) because reaction (8) is very rapid. Reaction (8) is called a Si reduction process, and it shows self-limiting growth after which no further W growth occurs. The self-limiting thickness depends on the CVD conditions and Si surface conditions, and it ranges from 10 to 1000 nm. In Si contacts, highly doped shallow junctions are formed to obtain an Ohmic contact, and the Si reduction process should be suppressed as much as possible because "encroachment" of W into the shallow junctions degrades leakage current characteristics. Moreover, a by-product gas of reaction (5), HF, can etch Si at relatively high temperatures and form a "worm hole" at the interface of W/Si, as was observed by Stacy *et al.* (1985).

To overcome this problem of Si damage, Kusumoto *et al.* (1987), Ohba *et al.* (1987), and Kotani *et al.* (1987) developed W-CVD using monosilane (SiH_4) as a reducing gas for WF_6. In this SiH_4-reduced W-CVD, the growth rate is high because of the high reactivity of SiH_4 with WF_6, and Si damage is greatly suppressed. The most important parameter of this chemistry is the gas-flow ratio of SiH_4/WF_6. Tungsten silicide (WSi_2) film can be deposited with ratios greater than 20. Tungsten can be selectively deposited with ratios less than 1–1.3. Film properties such as surface morphology and mechanical stress also depend on the flow ratio. Although the chemistry is not yet fully understood for SiH_4-reduced W-CVD, Kobayashi *et al.* (1991) observed two reaction pathways using *in situ* FT-IR (Fourier-transform infrared spectroscopy) of reaction gases during the selective W-CVD:

$$WF_6 + 2SiH_4 \rightarrow W + 2SiHF_3 + 3H_2, \tag{9}$$

$$WF_6 + \tfrac{3}{2}SiH_4 \rightarrow W + \tfrac{3}{2}SiF_4 + 3H_2. \tag{10}$$

Yu and Eldridge (1989) also observed reaction (10) using mass spectroscopy. A gas-phase reaction was observed even at room temperature in SiH_4/WF_6 gas mixtures when the flow ratio was larger than 1.3, and this causes loss of selectivity during deposition. Kobayashi *et al.* (1993) reported that SiH_4 is selectively dissociated at the W surface at temperatures higher than 150 °C, using the FT-IR reflection absorption spectroscopy, and this can be the cause of selective deposition on W without W deposition on SiO_2.

The growth kinetics has been studied by many researchers; however, there are some discrepancies among their data. The growth rate (R) is summarized as

$$R \propto \exp(-E/kT)P(WF_6)^a P(SiH_4)^{1+b}, \tag{11}$$

where E is the activation energy = 0.29–0.30 eV, P is the partial pressure, a is the exponent for $P(WF_6) = -0.5$ to 0, and b is the exponent for $P(SiH_4) = 0$ to 0.5. Suzuki *et al.* (1990) reported that the Si content incorporated in W films is a function of the gas-flow ratio as shown in Fig. 12. The Si content slightly increases with increasing gas-flow ratio, and decreases with increasing deposition temperature in selective deposition. On the other hand, in nonselective deposition where the gas-flow ratio exceeds 1.3, the Si content greatly increases, and it does not strongly depend on the deposition tem-

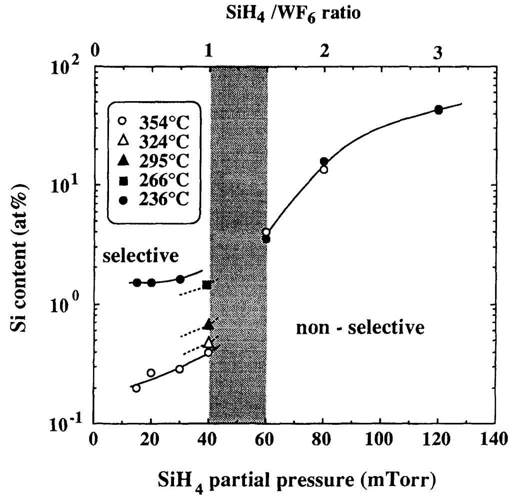

FIG. 12. The Si content in CVD-W films as a function of the gas flow of SiH_4/WF_6 at various temperatures.

perature. X-ray deflection shows that the film consists of polycrystalline α-W with gas-flow ratios less than 1, and it also involves β-W with ratios around 1.3. When the ratio exceeds 3, a broad x-ray diffraction peak of WSi_2 phase was also observed. The electrical resistivity of the film is almost proportional to the Si content and is determined by impurity scattering. This is confirmed by electrical resistivity measurement at liquid He temperatures. The superconducting transition at around 10 K was observed with W films containing 13–25 at.% Si. Kondo (1992) explained that a great increase of the superconducting transition temperature was ascribed to lattice softening in the amorphous W films. The film morphology also depends on the Si content in the film as shown in Fig. 13. The film texture changes from equiaxed to columnar surface features with increasing Si content.

In the ideal selective W-CVD, W can be deposited without any deposition of W on SiO_2. However, unnecessary W nuclei were often observed on the SiO_2 surface in the manufacturing process. These W nuclei are considered to be harmful particles produced in the process, and they should be reduced as much as possible to increase the device yield. Empirically, the cause for selectivity loss can be classified into two categories; one is the cause associated with chemical reactions during W-CVD, and the other is related to the surface condition of the SiO_2 surfaces. As for the former, besides the gas-phase reactions, reaction by-products such as tungsten subfluoride (WF_x) and silicon subfluoride (SiF_x) are reported by Creigton (1987) and Hirase *et al.* (1987) to cause selectivity loss. In the case of the latter, the SiO_2 surface is generally contaminated and damaged by conventional photolithography and dry etching, and this can cause loss of selectivity. It is known that Ar sputtering can break the Si–O chemical bond, and this provides a nucleation site for W deposition. In order to maintain the selectivity, nucleation sites on

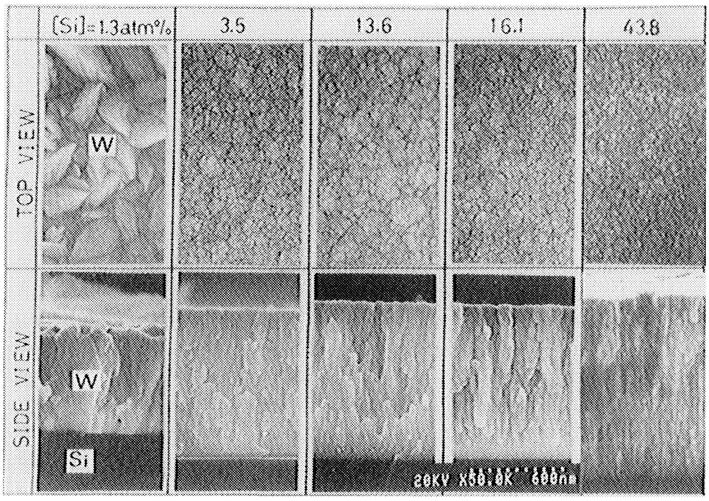

FIG. 13. Film morphology of CVD-W films as a function of the Si content.

the SiO_2 surface should be stably terminated with chemisorbed species. For this approach, *in situ* RIE (reactive ion etching) using Cl-based chemistry is being investigated and is shown to be effective to maintain the selectivity, as was reported by Tamaru *et al.* (1992).

2.3.3 Blanket Chemical Vapor Deposition of Tungsten In the formation of W interconnects, W-CVD has been widely used to fill submicron contacts. Because H_2- or SiH_4-reduced W-CVD provides selective deposition, W is not deposited on SiO_2. Thus, an adhesion metal layer is usually sputter deposited before the W-CVD when W film is formed on SiO_2. Titanium nitride (TiN), titanium–tungsten compound metal (Ti–W), and W are usually used for the adhesion layer. In the choice of adhesion layer, fine patternability, reactivity with WF_6, and contact resistivity (Schottky-barrier height) should be considered along with the adhesion to SiO_2. After deposition of an adhesion metal layer, W is deposited on the entire surface of the metal by CVD. We call this blanket W-CVD in contrast to selective W-CVD. Blanket W-CVD is used for forming W interconnection.

In blanket W-CVD, the H_2-reduced process has been used because it can provide conformal, low-resistivity W deposition. SiH_4-reduced W-CVD is not suitable for blanket deposition because of the low step coverage factor and high resistivity of the W film. The problem of H_2-reduced W-CVD is its rather high deposition temperature (450–475 °C) and low growth rate (<50 nm/min). A high-pressure process, in which H_2-reduced W-CVD is carried out at high pressures around 10^4 Pa, has been developed to increase the growth rate and obtain a smooth film surface. On the other hand, a new W-CVD chemistry has been proposed to overcome the problem. Van der Jeugd *et al.* (1991) developed W-CVD using WF_6 and germane (GeH_4), and Goto *et al.* (1991) proposed difluoro-silane (SiH_2F_2) as a reducing gas for WF_6 for low-temperature and conformal blanket deposition. The problems associated with H_2-reduced W-CVD, such as "encroachment" and "worm holes," are not crucial in blanket deposition because an adhesion layer acts as a barrier against the WF_6/Si reaction. However, the adhesion layer can be too thin to prevent the reaction at the bottom of the deep contact if the contact aspect ratio increases. So, the conformal deposition of the adhesion layer by CVD is being studied. CVD of TiN, W_2N, and WSi_2 have been reported for the formation of adhesion layers.

2.4 Solid-Phase Reaction of Refractory Metal with Silicon

As mentioned in the previous section, refractory metal is applied to interconnect material or barrier metal. In these applications, a refractory-metal film is formed directly on Si, and so solid-phase reactions of refractory metal with silicon during high-temperature processing is a great concern regarding the thermal stability of the interface between the refractory metal and Si. The refractory metal reacts with Si to form silicide, and it is associated with volume shrinkage as well as dopant redistribution. On the other hand, silicide formation is intentionally used to get a clean metal/Si interface during LSI processing.

2.4.1 Silicide Formation Typical silicidation temperatures for refractory metals, at which solid-phase reactions form stable silicide, are listed in Table 8. These data are from Murarka (1983). In case of W interconnects to n^+ Si, the silicide reaction begins at around 600 °C, and it changes the contact resistance. A sudden increase in contact resistance is observed as a result of the dopant redistribution caused by the silicide reaction of W with Si. In general, solid solubility of arsenic, phosphorus, and boron is large in silicide, and so dopant concentration at the interface will decrease after the silicide reaction. In addition, fast dopant diffusion in silicides can cause the mixing of dopants when silicides are used for interconnects connected to diffused Si areas. However, some silicides such as platinum silicide and palladium silicide are known to push the dopant out during silicidation.

A lot of mechanical stress occurs at the interface because of the volume shrinkage during silicidation, and this can degrade the electrical characteristics of the contacts. LSI processes should therefore be designed to avoid the problems associated with silicidation, when refractory metals are used in Si LSIs.

Table 8. Silicidation temperatures and resistivities for refractory-metal silicides.

Silicide	Formed by starting from	Sintering temperature (°C)	Resultant resistivity ($\mu\Omega$ cm)
$TiSi_2$	Metal on polysilicon	900	13–16
	Cosputtered	900	25
$ZrSi_2$	Metal on polysilicon	900	35–40
$HfSi_2$	Metal on polysilicon	900	45–50
	Cosputtered alloy		60–70
VSi_2	Metal on polysilicon	900	50–55
$NbSi_2$	Metal on polysilicon	900	50
$TaSi_2$	Metal on polysilicon	1000	35–45
	Cosputtered alloy	1000	50–55
$CrSi_2$	Metal on polysilicon	700	~600
$MoSi_2$	Metal on polysilicon	1100	~90
	Cosputtered alloy	1000	~100
WSi_2	Cosputtered alloy	1000	~70
$FeSi_2$	Metal on polysilicon	700	>1000
$CoSi_2$	Metal on polysilicon	900	18–20
	Cosputtered alloy	900	25
$NiSi_2$	Metal on polysilicon	900	~50
	Cosputtered alloy	900	50–60
PtSi	Metal on polysilicon	600–800	28–35
	Cosputtered alloy		
Pd_2Si	Metal on polysilicon	400	30–35

2.4.2 Use of silicides in MOS Source and Drain Metallization A silicide reaction is intentionally used to obtain the low-resistivity source and drain in MOS transistors. Shibata *et al.* (1981) showed an improvement in device performance of submicron MOS transistors by forming silicides on source and drain. We call this self-aligned silicide (SALICIDE) technology. Figure 14 shows a typical process flow for SALICIDE using titanium silicide ($TiSi_2$), as was reported by Lau (1982). After n^+ or p^+ source and drain formation, sidewall SiO_2 is formed on the gate [Fig. 14(a)]. A Ti film is deposited on the Si substrate by sputtering [Fig. 14(b)], and this is followed by annealing at silicidation temperature (650–700 °C) to form titanium silicide on the gate and source/drain. Unreacted Ti film on SiO_2 is removed by wet chemicals (H_2O_2 + NH_4OH + H_2O). After these procedures, low-resistivity $TiSi_2$ film is formed on the gate as well as the source and drain regions. This process greatly decreases the parasitic resistance of MOS transistors, and it is especially effective in obtaining high-speed performance in submicron MOS transistors. Similar processes have been investigated for reducing the base resistance of bipolar and other kinds of transistors. For SALICIDE technology, Co and Ni have also been used to form silicides on account of their low resistivity.

To obtain a reliable silicide contact, it is important to clean the Si surface before refractory-metal deposition. Native oxide on a Si surface is known to greatly suppress silicidation, and should therefore be removed to get good reproducibility of the silicidation process. Native oxide on the Si surface is usually removed using buffered HF dip. Much work has been done to investigate the silicide reaction focusing on Si surface conditions and refractory-metal deposition conditions.

3. APPLICATIONS TO GaAs MESFETs

Refractory metals are used for the self-aligned MESFET (MEtal-Semiconductor Field-Effect Transistor) gate in GaAs devices just as they are used for the MOSFET gate in Si ULSIs. The metal gate technology was reviewed by Weitzel and Doane (1986). It is important that the criteria for MESFET gate material selection are different from those for Si MOSFET. Thermal stability of the Schottky contact between the gate material and GaAs is crucial because the gate material is directly deposited on the n-type GaAs substrate in the MESFET structure. For this reason, refractory metals such as W and Mo are not suitable for the MESFET gate; however, alternatively, tungsten silicide (W_5Si_3)

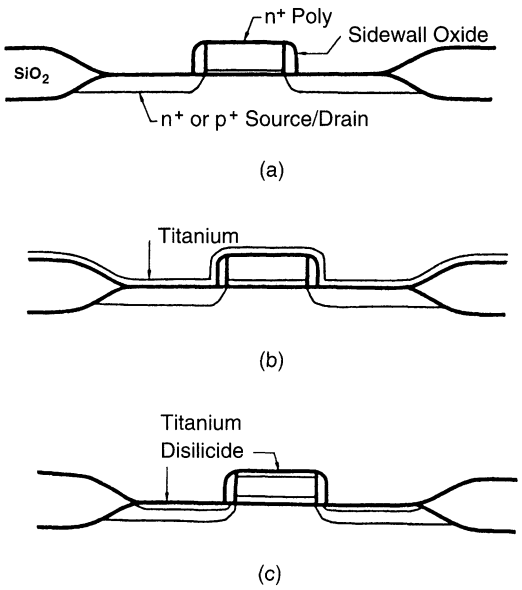

Gate oxide thickness: 250 Å
Gate polysilicon: 5000 Å
Phosphorus doped to 20 Ω/□
Source/drain n+ implant dose: 1×10^{16} arsenic
p+ implant dose: 1.5×10^{15} BF_2
Silicide thickness: 350–700 Å

FIG. 14. Fabrication process flow of the self-aligned silicide (SALICIDE) devices using titanium silicide ($TiSi_2$).

or tungsten nitride (W_2N) are adopted for the GaAs MESFET fabrication. Tungsten silicide (W_5Si_3) is the most suitable gate material in terms of thermal stability of the Schottky contact barrier height at up to 800 °C as compared with TiW or TiW silicide. Besides tungsten silicide, WN has been reported to be used for the MESFET gate by Uchitomi *et al.* (1984). These refractory metals are usually deposited by magnetron sputtering.

In the GaAs devices, Au is used for interconnects because Al can be an active dopant in GaAs. When Au interconnects make a contact to substrate, Ti or Mo films with a thickness of around 50 nm are usually sputter deposited under the Au film. They work as barrier metals against Au diffusion into GaAs. At the same time, these refractory metals play the role of an adhesion layer between the Au interconnect and the SiO_2.

Besides the applications described above, refractory compound metals such as WSi_xN_y, Ta_3N_4, and NiCr have been used as electrical resistors in GaAs devices. These metals are usually deposited by reactive sputtering. For example, WSi_xN_y films are deposited by sputtering a W_5Si_3 target in N_2/Ar mixed atmosphere.

4. ACKNOWLEDGMENTS

The author is much obliged to Dr. Seiichi Iwata, Dr. Naoki Yamamoto, and Dr. Masayoshi Saitoh for their collaboration in the study of refractory-metal applications to Si ULSIs. Above all, encouragement from Dr. Shojiro Asai in writing this article is highly appreciated.

Works Cited

Bernard, C., Madar, R., Pauleau, Y. (1989), *Solid State Technol.* **32,** (2), 79–84.

Broadbent, E. K., Ramiller C. L. (1984), *J. Electrochem. Soc.* **131,** 1427–1432.

Brown, D. M., Cady, W. R., Sprague, J. W., Salvagni D. J. (1971), *IEEE Trans. Electron Devices* **ED-18,** 931–940.

Creigton J. R. (1987), *J. Vac. Sci. Technol.* **A5,** 1739–1740.

Cuomo, J. J. (1972), in: F. A. Glaski (Ed.), *Proceedings of the 3rd International Conference on Chemical Vapor Deposition,* Hinsdale, IL: American Nuclear Society, pp. 270–291.

Ghate, P. B., Blair, J. C., Fuller, C. R., McGuire, G. E. (1978), *Thin Solid Films* **53,** 117–128.

Goto, H., Kobayashi, N., Homma, Y. (1991), in: *Extended Abstracts of International Conference on Solid State Devices and Materials,* Tokyo: The Japan Society of Applied Physics, p. 183.

Gray, D. E. (Ed.) (1972), *American Institute of Physics Handbook,* 3rd ed., New York: McGraw-Hill.

Hirase, I., Sumiya, T., Schack, M., Umishima, S., Rufin, D., Shinshikura, M., Matsuura, M., Ito, A. (1987), in: V. A. Wells (Ed.), *Tungsten and Other Refractory Metals for VLSI Applications III,* Pittsburgh: Materials Research Society, p. 133.

Hoffman, D. W., Thornton, J. A. (1982), *J. Vac. Sci. Technol.* **20,** 355–358.

Holloway, K., Fryer, P. M., Cabral, C., Jr., Harper, J. M. E., Bailey, P. J. (1992), *J. Appl. Phys.* **71,** 5433–5443.

Iwata, S., Yamamoto, N., Kobayashi, N., Terada, T., Mizutani, T. (1984), *IEEE Trans. Electron Devices* **ED-31,** 1174–1179.

van der Jeugd, C. A., Leusink, G. J., Janssen, G. C. A. M., Radelaar, S. (1991), *Appl. Phys. Lett.* **57,** 354–356.

Kobayashi, N., Yamamoto, N., Iwata, S., Mizutani, T., Yagi, K. (1984), in: *International Electron Devices Meeting Technical Digest,* New York: IEEE, p. 122.

Kobayashi, N., Goto, H., Suzuki, M. (1991), *J. Appl. Phys.* **69,** 1013–1019.

Kobayashi, N., Nakamura, Y., Goto, H., Homma, Y. (1993), *J. Appl. Phys.* **73,** 4637–4643.

Kondo, S. (1992), *J. Mater. Res.* **7,** 853–860.

Kotani, H., Tsutsumi, T., Komori, J., Nagao, S. (1987), in: *International Electron Devices Meeting Technical Digest,* New York: IEEE, p. 213.

Kusumoto, Y., Takakuwa, K., Hashinokuchi, H., Ikuta, T., Nakayama, I. (1987), in: V. A. Wells (Ed.), *Tungsten and Other Refractory Metals for VLSI Applications III,* Pittsburgh: Materials Research Society, p. 103.

Lau, C. K., See, Y. C., Scott, D. B., Bridges, J. M., Perna, S. M., Davies, R. D. (1982), in: *International Electron Devices Meeting Technical Digests,* New York: IEEE, p. 714.

Loh, W. H., Swirhun, S. E., Crabbe, E., Saraswat, K., Swanson, R. M. (1985), *IEEE Electron Devices Lett.* **EDL-6,** 441–443.

May, T. C., Woods, M. H. (1970), in: *Proceedings IEEE Reliability Physics,* New York: IEEE, p. 33.

McConica, C. M., Krishnamani, K. (1986), *J. Electrochem. Soc.* **133,** 2542–2548.

Moore, G. (1980), *Electron. Aust.* **42,** p. 14.

Muller, R. S., Kamins, T. I. (1986), in: *Device Electronics for Integrated Circuits,* 2nd ed., New York: Wiley, Chap. 9.

Murarka, S. P. (1983), *Silicides for VLSI Applications,* New York: Academic Press, p. 30.

Ohba, T., Inoue, S., Maeda, M. (1987), in: *International Electron Devices Meeting Technical Digest,* New York: IEEE, p. 213.

Pramanik, D., Jain, V. (1991), *Solid State Technol.* **34** (5), 97–101.

Rhoderick, E. H. (1978), *Metal-Semiconductor Contacts,* Monographs in Electrical and Electronic Engineering No. 19, Oxford, U.K.: Clarendon Press, p. 53.

Saraswat, K. C., Mohammadi, F. (1982), *IEEE J. Solid-State Circuits* **SC-17,** 275–280.

Schmitz, E. J. (1992), *Chemical Vapor Deposition of Tungsten and Tungsten Silicides for VLSI/ULSI Applications,* Park Ridge, NJ: Noyes Publications, p. 143.

Sherman, A. (1987), in: *Chemical Vapor Deposition for Microelectronics,* NJ: Noyes Publications, p. 35.

Shibata, T., Hieda, K., Sato, M., Konaka, M., Dang, R. L. M., Iizuka, H. (1981), in: *International Electron Devices Meeting Technical Digest,* New York: IEEE, p. 651.

Stacy, W. T., Broadbent, E. K., Norcott, M. H. (1985), *J. Electrochem. Soc.* **132,** 444–448.

Suzuki, M., Kobayashi, N., Mukai, K., Kondo, S. (1990), *J. Electrochem. Soc.* **137,** 3213–3218

Sze, S. M. (1983), in: Sze, S. M. (Ed.), *VLSI Technology,* New York: McGraw-Hill, Chap. 12.

Sze, S. M. (1981), in: *Physics of Semiconductor Devices,* 2nd ed., New York: Wiley, p. 304.

Tamaru, T., Iwata, S., Kobayashi, N., Tokunaga, T. (1992), in: *Proceedings of the Dry Process Symposium,* Tokyo: The Japan Society of Applied Physics, p. 68.

Thornton, J. A., Penfold, A. S. (1978), in: V. L. Vossen, W. Kern (Eds.), *Thin Film Processes,* New York: Academic Press, p. 75.

Uchitomi, N., Kitaura, Y., Mizoguchi, T., Ikawa, Y., Toyoda, N., Hojo, A. (1984), in: *Extended Abstracts International Solid State Devices and Materials Meeting,* Tokyo: The Japan Society of Applied Physics, p. 383.

Weitzel, C. E., Doane, D. A. (1986), *J. Electrochem. Soc.: Reviews and News* **133,** 409C.

Wang, K. L., Holloway, T. C., Pinizotto, R. F., Sobczak, Z. P., Hunter, W. R., Tasch, A. F. (1982), *IEEE Trans. Electron Devices* **ED-29,** 547–553.

Yamamoto, N., Iwata, S., Kume, H. (1987), *IEEE Trans. Electron Devices* **ED-34,** 607–614.

Yokoyama, N., Hinode, K., Homma, Y. (1991), *J. Electrochem. Soc.* **138,** 190–195.

Yu, M. L., Eldridge, B. N. (1989), *J. Vac. Sci. Technol.* **A7,** 625–629.

REFRACTORY METALS

Erwin Pink, *Erich-Schmid-Institut für Festkörperphysik der Österreichischen Akademie der Wissenschaften, Leoben, Austria*

Ralf Eck, *Metallwerk Plansee GmbH, Reutte, Austria*

Introduction 327

1. **Materials and Their Properties** 328

1.1 The Pure Metals 328

1.1.1 General Physical Properties ... 328

1.1.2 Chemical Properties 328

1.1.3 Strength and Ductility 329

1.1.4 Essential Differences between V*a* and VI*a* Metals 329

1.2 Refractory Alloys 329

1.2.1 Systematics of Alloys 329

1.2.2 Commercial Alloys 330

1.2.2.1 Substitutional Alloys 330

1.2.2.2 "Doped" Refractory Metals 330

1.2.2.3 Dispersion-Strengthened Alloys 330

1.2.2.4 Combinations 330

1.2.2.5 Dual-Phase Alloys 332

1.2.2.6 Refractory-Metal Cermets 333

1.2.2.7 Alloys Reinforced by Refractory-Metal Fibers 333

1.2.2.8 Amorphous Refractory Alloys 334

2. **Resources** 334

3. **Production Technologies** 334

3.1 Powder Compacting 334

3.2 Sintering 335

3.3 Melting 335

3.4 Forming 335

3.5 Heat Treatments and Thermomechanical Treatments 335

3.6 Recycling 336

4. **Further Processing** 337

4.1 Fastening and Joining 337

4.1.1 Mechanical Fastening 337

4.1.2 Welding 337

4.1.3 Brazing 337

4.2 Coatings 337

4.2.1 Pack and Slurry Cementation 337

4.2.2 Electrolytic Coating 338

4.2.3 Flame and Plasma Spraying .. 338

4.2.4 Chemical and Physical Vapor Deposition (CVD and PVD) ... 338

5. **Applications** 338

5.1 Main Use of Refractory Metals and Alloys 338

5.2 Special Requirements 339

5.2.1 High-Purity Materials 339

5.2.2 Porous Metals 339

5.2.3 Alloys for Thermocouples 339

5.2.4 Refractory Metals for Superconductors 339

5.2.5 Use in Thermonuclear Reactors 339

6. **Economic Aspects** 339

7. **Toxicological Aspects** 340

Glossary 340

Works Cited 341

Further Reading 342

INTRODUCTION

By convention, commercial refractory, or high-melting, metals are those that belong to groups V*a* and VI*a* of the periodic table of elements. They are vanadium, niobium, and tantalum (V*a*), and chromium, molybdenum, and tungsten (VI*a*). Their melting points exceed roughly 1900 °C, culminating at 3410 °C in the case of tungsten, which is a melting temperature comparable to those of the most heat-resisting carbides. There exist other metals with high melting points; some of them are, however, classified appropriately as precious metals (see Silver, Gold, and Other Precious Metals). The other high-melting metals, rhenium and hafnium, are used today as alloying elements for the refractory metals, and only rhenium in exceptional cases for its own metallic merits.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. MATERIALS AND THEIR PROPERTIES

1.1 The Pure Metals

1.1.1 General Physical Properties For high-temperature use, a high melting point, a low vapor pressure, low thermal expansion, and high thermal conductivity are of advantage. Other important physical properties of these metals are compiled in Table 1. All metals except for rhenium and hafnium (whose lattices are hexagonal close-packed) have a body-centered cubic crystal structure, and only hafnium exhibits a crystallographic phase change upon variation of the temperature. The microstructure of all pure refractory metals consists of one phase only.

In commercially pure metals, hydrogen is contained in amounts less than 1 μg/g. The oxygen, nitrogen, and carbon concentrations in V*a* metals are smaller than 50 μg/g, in VI*a* metals today mostly smaller than 10 μg/g. Substitutionally dissolved impurity elements (nickel, iron, or the other refractory metals) amount to 5–100 μg/g. The purity levels routinely available are better than 99.95%.

1.1.2 Chemical Properties Refractory metals have, in comparison with other metals, superior resistance to many chemical agents (Lambert, 1990). In general, they are resistant to corrosion in various media, but those of the V*a* group are outstanding in their performance. The stability of refractory metals in dilute nonoxidizing acids is good, but in oxidizing acids they corrode severely, especially molybdenum. All refractory metals are attacked by sodium and potassium hydroxides, especially when oxidizing substances are present. The resistance to solids and metal or glass melts is good to excellent. In molten vanadium pentoxide, a normally strongly corrosive medium (present in the ashes of all fossil fuels, with a melting point of 675 °C), the resistance of chromium is unsurpassed.

Refractory metals are to various extents prone to destruction through gases (Fromm and Gebhardt, 1976). This behavior, it will be shown, has consequences for the manufacture, use, and recycling methods of metals of the two groups. V*a* metals absorb large quantities of all gases (except those that are inert) and are then seriously embrittled. VI*a* metals remain unaffected. The effect of oxygen is opposite. Molybdenum and tungsten start to oxidize in air around 300–400 °C. Volatile oxides form and evaporate catastrophically above 600 °C in the case of molybdenum, above 800 °C for tungsten, and al-

Table 1. Important physical properties and mechanical room-temperature properties of pure refractory metals.

	Group V*a*			Group VI*a*			VII*a*
Property	V	Nb	Ta	Cr	Mo	W	Re
Structure	bcc	bcc	bcc	bcc	bcc	bcc	hcp
Density (g/cm^3)	6.1	8.6	16.6	7.2	10.2	19.3	21.0
Melting point (°C)	1920 ± 20	2470 ± 20	3000 ± 30	1920 ± 20	2620 ± 20	3380 ± 20	3180
Coefficient of linear thermal expansion 20 °C, (μm/m pd K)	8.3	7.6	6.6	6.2	5.3	4.6	8.4
Specific electrical resistivity at 20 °C ($\mu\Omega$ m)	0.19	0.14	0.12	0.12	0.05	0.05	0.17
Approximate interstitial solubility (μg/g), 20 °C, for C to H	10^3 to 10^4	10^2 to 10^4	70 to 4000	0.1 to 1	0.1 to 1	<0.1	<3000
Start of recrystallization (°C) (1-hour annealing)	650	750	1100	700	800	1150	900
Modulus of elasticity (GPa)	130–150	100–120	180–190	220–280	320–360	390–410	470
Poisson's ratio	0.37	0.40	0.34	0.21–0.28	0.30	0.30	0.29
Hardness HV10							
Recrystallized (min.)	70	80	80	100	150	350	250
As-worked (max.)	200	200	200	300	500	650	580
Minimum UTS (MPa) (recrystallized)	300	300	300	300	550	(480 °C) 600	1070
DBTT[a] (°C) bending (recrystallized)	−100	−200	−269	+200	−20	+400	

---

[a] Ductile–brittle transition temperature.

ready above 300 °C for rhenium. The oxidation behavior of chromium is exceptional, insofar as the progress of oxidation, once the oxide layer is formed, is much slower than even for niobium or tantalum.

Resistance to gaseous hydrocarbons and carbon oxides is defined by the temperatures of carbide formation. Nitrides form, at elevated temperatures, on surfaces of molybdenum and tungsten under high-pressure nitrogen and under ammonia at 1 bar (Martinz and Prandini, 1993–1994). In air, brittle layers grow very slowly at the usually employed temperatures below 1400 °C.

The metals niobium and tantalum have excellent biocompatibility (Zitter and Plenk, 1987).

1.1.3 Strength and Ductility The temperature dependences of strength values and fracture elongations of the recrystallized pure metals are shown in Fig. 1 in relation to each other. The strength of metals in the worked state is higher, but this gain can only be exploited in high-temperature applications as long as the recrystallization temperature is not exceeded. These limiting temperatures are also listed in Table 1, together with other mechanical characteristics.

1.1.4 Essential Differences between V*a* and VI*a* Metals The most notable feature of VI*a* metals—aside from other differences between the metals of the V*a* and VI*a* groups apparent from Table 1—is the severe embrittlement at low temperatures [below the "ductile–brittle transition temperature"; see Fig. 1(b)]. In principle, this is a common inherent feature among body-centered cubic metals (Honeycombe, 1984). However, the embrittlement temperature in the VI*a* metals is augmented and reaches, or even exceeds, room temperature when oxides and large carbides cover the grain boundaries and weaken their cohesion, especially in the recrystallized state. This is a consequence of the interplay of low interstitial solubility and actual content of impurities (see Table 1 and Sec. 1.1.1). The ductile–brittle transition temperature may be high even when low deformation rates, as in bending, are applied (see Table 1). Various means can be employed to reduce this handicap: purifying, alloying (cf. Sec. 1.2.2), and mechanical working (in order to neutralize the effect of contaminated grain boundaries).

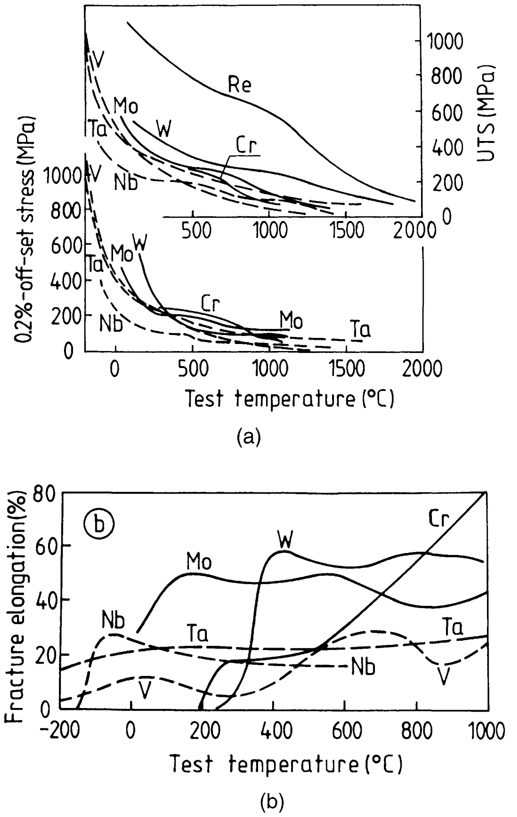

(a)

(b)

FIG. 1. The temperature dependences of (a) the ultimate tensile (UTS) and yield stresses and (b) the fracture elongations of recrystallized refractory metals, for medium strain rates (Tietz and Wilson, 1965).

1.2 Refractory Alloys

1.2.1 Systematics of Alloys In the case of refractory metals, the term "alloy" comprises—in addition to its standard meaning—also types commonly not produced from other metals. Some of these are not even "alloys" in the same sense the word is used otherwise, and so the variety of refractory alloys exceeds that of other metals. The purpose of alloying is manifold (Pink and Eck, 1996): It ranges from influencing the recrystallization behavior to raising the strength and ductility by various means, at low as well as at high temperatures, or to combining special physical properties of the alloy constituents. The common strengthening methods can equally well be employed (with the exception of transformation, i.e., martensitic, strengthening, which is not possible because of the absence of phase transformations). In addition, certain alloys are designed for special purposes to make use of mechanisms based on principles that are rarely exploited otherwise.

The alloy types are substitutional alloys, "doped" refractory metals, dispersion-strengthened alloys, combinations, dual-phase alloys, metal-ceramic composites, alloys reinforced by refractory-metal fibers, and amorphous refractory alloys.

1.2.2 Commercial Alloys Table 2 lists the most important types of alloys for the six group V*a* and VI*a* refractory metals, and the improvements that are achieved. The classes introduced in the previous section are described in the following.

1.2.2.1 Substitutional Alloys. Most solution hardening has limited importance. But rhenium additions to molybdenum and tungsten reduce or even eliminate their brittleness completely (this is not caused by a solution effect; the reasons are not yet quite explored). In certain substitutional solutions, the corrosion resistance is improved.

1.2.2.2 "Doped" Refractory Metals. Tungsten wires for high-temperature application (e.g., in incandescent lamps) need a specially modified recrystallized grain structure (Fig. 2) (Snow, 1989). This "nonsag" structure with coarse elongated and overlapping recrystallized grains with a high "grain aspect ratio" (grain length-to-width ratio) will not sag, i.e., deform along grain boundaries, as a normally recrystallized tungsten wire with a "bamboo" grain structure would. This doped or nonsag (NS) tungsten is produced through doping with aluminum chlorides and potassium silicates ("AKS" tungsten is a further denomination). The concentration of potassium that remains in the wire is smaller than 100 μg/g, but sufficient to form bubbles less than 0.1 μm in diameter, filled with metallic potassium. In the course of the severe deformation to wire diameters less than 0.5 mm, the bubbles become aligned, less than 1 μm apart, in the direction of the wire axis. They prevent the migration of the grain boundaries in the direction perpendicular to the wire axis during recrystallization, and raise the recrystallization temperature. The nonsag structure is not only creep resistant, but strengthens the matrix also at low temperatures since the bubbles serve as dispersed obstacles against dislocation motion (Pink and Eck, 1996). The overlapping grains also increase to a certain extent the bending ductility at the lower temperatures. A disadvantage is that bubbles migrate and coalesce when the wire is overheated. The effective cross section may then be reduced, and the wire becomes prone to melting when heated by direct current. This kind of doping has also been applied to produce rods, wires, sheets, and foils of molybdenum that do not embrittle as severely as the pure metals when they are recrystallized.

1.2.2.3 Dispersion-Strengthened Alloys. Commercial alloys are based on the effects of dispersed carbides that form from additions of alloying metals and carbon. Strengthening through nitrides is not technically exploited, although research has proved it equally effective. A prominent example is the molybdenum alloy TZM, containing titanium ("T") and zirconium ("Z"). "M" stands for molybdenum. At present, the strongest alloy is ZHM with hafnium ("H") carbide, which has the second-highest melting point of all carbides. In current practice, alloying is complemented by thermal treatments such as aging or strain aging only in the case of HfC-containing alloys (see Sec. 3.5).

A recent successful development is sheets and rods made from ODS (i.e., oxide-dispersion-strengthened) molybdenum. The additions are oxides of rare earths, especially of lanthanum, of high thermal stability. These materials increasingly replace the doped molybdenum because, aside from being dispersion hardened, they also develop an elongated grain structure upon recrystallization, and exhibit thus all the advantages offered by a dopant without having its drawbacks: There is no danger of bubble blistering by potassium evaporation during exposure to high temperatures, and no need for the extreme degree of deformation as is necessary for producing the bubble rows. A similar development concentrates on the production of ODS-tungsten wires.

1.2.2.4 Combinations. A further improvement is achieved by combinations of the measures described in Secs. 1.2.2.1–1.2.2.3. For instance, V*a*-group metals are strengthened by hafnium carbides, nitrides, and oxides, and their effect is enhanced by solution strengthening through tungsten, zirconium, and hafnium. Complex tungsten al-

Table 2. Properties and applications of commercial alloys (with additions in wt%).

Alloy	Oustanding properties	Applications
Vanadium alloys:		
V–(5–30)Ti–(5–15)Cr	High strength, good strength/specific-weight ratio, good corrosion resistance	Aircraft and space applications; in liquid heat exchangers, nuclear engineering
Niobium alloys:		
Nb–1Zr Nb–10W–10Hf–15C Nb–10W–2.5Zr Nb–28Ta–10W–1Zr	High strength, good strength/specific weight ratio	Space applications; nuclear engineering
Tantalum alloys:		
Ta–(2.5; 7.5; 10)W Ta–8W–2Hf Ta–10W–2.5Hf–0.01C	Good corrosion resistance; high strength	Chemical process equipment
Chromium alloys:		
Cr–(40–75)Cu	Good resistance against electrical wear; good electrical conductivity	Vacuum contacts
Molybdenum alloys:		
Doped (KS) Mo	High recrystallization temperature, elongated recrystallized grain; creep resistant, ductile at 20 °C	Wires, sheets, foils for exposure at 1600–1800 °C
Mo–(5–30)W	Resistant against liquid zinc	Zinc-producing industry
Mo–(5–41)Re	Highest ductility, high strength; electromotoric force	Ductile thermocouples for high temperatures
Dispersion alloys:		
"TZM" Mo–0.5Ti–0.08Zr–0.05C "MHC" Mo–1.2Hf–0.06C "ZHM" Mo–(0.4–0.7)Zr–(1.2–2)Hf–(0.15–0.25)C	High strength, good ductility, age hardenable (MHC, ZHM)	Construction materials for 900–1600 °C, forging tools
ODS alloy Mo–0.7La_2O_3	High recrystallization temperature, elongated recrystallized grain; creep resistant; ductile at 20 °C	Construction material for use at 1200–2200 °C
Complex alloys:		
Mo–25W–1.2Hf–0.06C ODS-ZHM-1Y_2O_3/1La_2O_3	Best strength at highest temperatures	Construction material for use at 1200–2200 °C
Tungsten alloys:		
Doped (AKS) W	High recrystallization temperature, elongated recrystallized grain; high creep resistance at high temperatures	Incandescent lamp filaments; metallizing wires
Doped W–(1–2) ThO_2	Also: low work function	Electronic emitters
W–(3–26)Re	Best low-temperature ductility, high-temperature strength; electromotoric force	X-ray targets, ductile thermocouple wires for highest temperatures
Dispersion (ODS) alloys:		
W–(1–4)ThO_2	Low work function; good machinability	Welding electrodes; plasma generators
W–(1–2)ZrO_2/CeO_2/La_2O_3	The same, as a nonradioactive alternative material for W–ThO_2	The same
Heavy alloys:		
W–(2–7)Ni–(1–3)Fe/Cu	Ductile at room temperature, excellent machinability; high specific weight	*X*- and γ-radiation shields; counter weights and high-inertia material; artillery penetrators
Electrical contact materials:		
W–(10–50)Cu/Ag W–20Cu–1Ni	Ductile at room temperature, excellent machinability; good electrical conductivity, good resistance against electrical wear	High-voltage contacts; welding electrodes

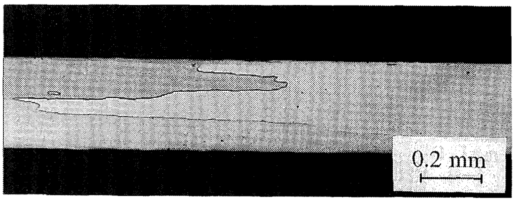

FIG. 2. The microstructure of a doped-tungsten wire with recrystallized elongated grains (courtesy of H. Yamamoto).

loys combine the advantages of rhenium and of hafnium carbides. Tungsten alloyed with rhenium can be AKS-doped to develop an elongated grain structure after recrystallization.

1.2.2.5 Dual-Phase Alloys. *Tungsten Heavy Metals.* In the early days of producing tungsten, attempts were made to create an easily machinable material by embedding the tungsten grains in a binder phase (e.g., nickel). Further development led to alloys containing 90–98 wt% tungsten and a face-centered cubic binder phase of nickel and iron. The powder mixture is heated above the melting temperature of the binder. Aside from general densification, the tungsten grains are partly or even completely dissolved during sintering (German, 1985). During cooling, the reprecipitation of the dissolved portions creates tungsten "spheroids" that are embedded in the binder phase (Fig. 3).

The Ni/Fe ratios in heavy metals are 1:1 to 4:1 (preferably 7:3, because this composition allows precipitation hardening of the matrix and is less likely to embrittle as a result of intermetallic phases at interfaces). Alloys in which copper is substituted for iron (in Ni/Cu ratios 3:2 to 4:1) have better electrical conductivity and are not ferromagnetic. Typical values of strength and fracture elongation are evident from Fig. 4. They vary with composition and with the amount of contiguity (which is a measure of the ratio of tungsten/tungsten to tungsten/matrix contact areas—a parameter of great importance for the ductility). The temperature dependence of the fracture energy obtained from Charpy tests shows a transition to brittle behavior as the temperature decreases (German, 1985). Heavy alloys can be coldworked to achieve a high ultimate tensile strength at the cost of toughness and fracture elongation.

Infiltrated (Pseudo) Alloys. Chromium, molybdenum, or tungsten is combined with silver or copper to create materials for electrical contacts (Shen *et al.,* 1990). Certain properties are combined that are incompatible in normal alloys: high electrical and thermal conductivity, high resistance to chemical reactions, a high melting point for good resistance to burn-off and spark erosion, and good mechanical strength and easy machinability. A skeleton of the refractory metal with pores of optimum size is first produced. The interconnecting pores are then infiltrated by placing the necessary amount of silver or copper in contact with the outer surface, and heating the whole above the melting point of the infiltrant. With the exception of the chromium-base alloys, the

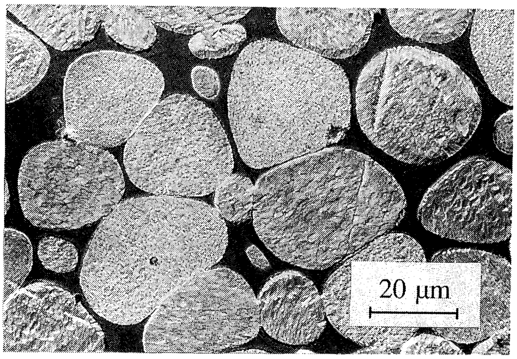

FIG. 3. The microstructure of a tungsten heavy alloy with 7 wt% nickel and 3 wt% iron (courtesy of H. Danninger). The binder phase has been etched away.

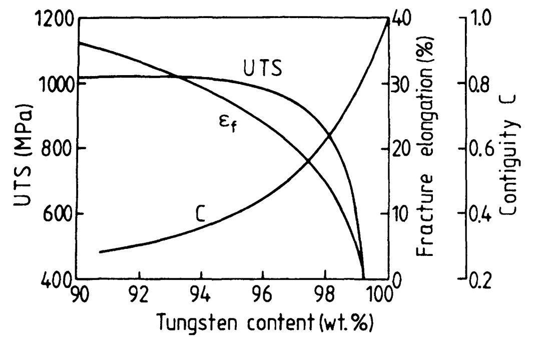

FIG. 4. Strength and fracture elongation of heavy alloys with varying tungsten contents, and their correlation with the contiguity according to Edmonds (Pink and Eck, 1996).

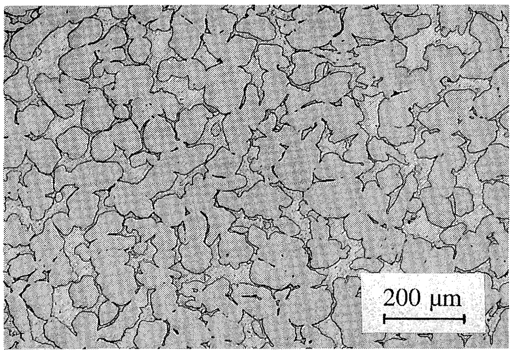

FIG. 5. The microstructure of tungsten (the larger component) infiltrated with 30 wt% copper.

components do not react with each other, and a typical microstructure of an infiltrated material (Fig. 5) therefore does not consist of the same rounded grains as the tungsten heavy alloy. Electrical contact materials with large or excess amounts of copper and silver (>50 vol%) are produced differently. A mixture of the two powder components is pressed and solid-state sintered at temperatures below the melting point of the low-melting constituent. In such structures, the refractory-metal component does not form a continuous skeleton. Upon heating above the melting point of copper or silver, the alloy would be destroyed. Such solid-state sintered tungsten-copper can be produced to densities near 90% of the theoretical. Further densification is possible by plastic forming at elevated temperatures. Some specific properties of W–Cu/Ag and Cr–Cu alloys are collected in Table 3. The strength of the composite material can be improved by cold rolling to obtain work-hardened sheet material, or—rarely done—by solution strengthening of the soft phase.

1.2.2.6 Refractory-Metal Cermets. The name designates a heterogeneous combination of a metal and a ceramic. The volume fraction of the metallic binder can be 40%–85%. Cermets with high concentrations of refractory metals are materials that are strong and wear resistant, and capable of withstanding chemical attack. In general, they can be easily machined. A combination of molybdenum with zirconia is used as sheath material for thermocouples for measuring the temperature of molten steel (Heitzinger, 1974). The strength (about 300 MPa at room temperature) varies little (or even increases slightly) with temperature up to about 1000 °C.

The following alloy variants are not standard materials.

1.2.2.7 Alloys Reinforced by Refractory-Metal Fibers. Niobium alloys reinforced with tungsten wires have been proposed as potential materials for long-term high-temperature applications in space-power systems. Wires of refractory metals (especially tungsten) strengthen other metals (primarily nickel) and superalloys intended for use at high temperatures (Lambert, 1990; see also COMPOSITE MATERIALS). Embedded within the matrix, the problem with the poor oxidation resistance of some of the refractory metals is reduced. In comparison to ceramic fibers, their thermal expansion and ductility are closer to those of the matrix materials so that they are better suited. The strength of all refractory-metal fibers is high in relation to that of most matrix metals at the applied temperatures, which are in general "low" from the point of view of tungsten fibers. Fiber–matrix interactions are possible, especially at the higher temperatures. Chemical reactions lead to the formation of brittle intermetallic phases. Traces of certain metals (e.g., nickel) that are deposited on the fiber surface or come from the matrix metal itself may diffuse into the wire and activate recrystallization of the reinforcement. This may be avoided by ductile coatings yet to be devel-

Table 3. Properties of contact alloys.

Alloy	Mean linear coefficient of thermal expansion (10^{-6}/K)	Specific electrical resistivity ($\mu\Omega$ m)	Heat conductivity (W/m K)
W–20Cu	8.5	0.05	134
W–20Ag	8.7	0.05	138
Cr–75Cu	13.5	0.04	235

oped, or by recrystallization before embedding.

1.2.2.8 Amorphous Refractory Alloys. Refractory metals are quenched into the glassy state by vapor deposition in the form of thin films onto substrates that are cooled down to temperatures below 10 K (Johnson, 1990). The pure amorphous refractory metals crystallize easily, but amorphous alloys can retain their structure upon annealing up to 1000 K. The alloys are of the metal–metal or metal–metalloid type—for instance, $(Nb,Mo)_{50}Ni_{50}$, $Mo_{30}Re_{70}$, $Mo_{52}Ru_{32}B_{16}$, or $W_{40}Re_{40}B_{20}$. Yield strength and hardness are typically high: With the Vickers hardness $HV = 2400$ of $W_{40}Re_{40}B_{20}$, for instance, and with a hardness-to-yield-point ratio of 3, common for metallic glasses, the yield stress should be 8000 MPa, which places these materials among the strongest known. Sputter deposition of amorphous layers a few microns thick delays, for instance, wear and corrosion of steels because amorphous metals have no grain boundaries.

2. RESOURCES

Specific mineral deposits of vanadium do not exist; vanadium is mainly extracted from iron, titanium, or uranium ores occurring in South Africa, Russia, China, and Finland. Increasing amounts of vanadium are obtained from petroleum. Niobium and tantalum are frequently found in close association in the mineral $(Fe,Mn)(Ta,Nb)_2O_6$, called either tantalite or columbite, depending on which metal is in excess. Cassiterite (SnO_2) deposits found in Thailand, Malaysia, Australia, etc. also contain significant amounts of tantalum and niobium. Slags from tin smelting are frequently used as tantalum feedstock. The highest-grade deposits of niobium, consisting of the mineral pyrochlore, $(Na,Ca)_2(Nb,Ta,Ti)_2O_4(OH,F)\cdot H_2O$, are found in Brazil.

Almost all chromium is obtained from deposits of chromite, $(Fe,Mg)O\cdot(Cr,Al,Fe)_2O_3$, principally in southern Africa, and in Russia, Finland, Turkey, and the Philippines. Ninety-five percent of the world production of molybdenum is from deposits of the disulfide MoS_2 that occur in the eastern Rocky Mountains of the U.S.A., in Chile, and in the Caucasus and the Ural; the most important of these are the so-called porphyry molybdenum deposits, in which molybdenum occurs with a grade of about 1% MoS_2 in conjunction with copper. Of the many tungsten minerals, only two groups are of commercial importance: the wolframite group, which consists of a solid solution of ferberite ($FeWO_4$) and huebnerite ($MnWO_4$) and occurs mainly in China, Korea, and the U.S.A.; and the scheelite group ($CaWO_4$) found primarily in Canada and Austria. Most ores contain 0.25%–2.5% WO_3.

Rhenium is obtained as a by-product from the mining of molybdenum-bearing porphyry copper ores. Major deposits occur in Chile, China, and Russia. Rhenium is particularly expensive because ores containing only 10–20 g/t have to be processed.

3. PRODUCTION TECHNOLOGIES

Since the product extracted from the ore and chemically treated is in many cases a metal powder, it is convenient to continue the further production by means of "powder metallurgy." In addition, this method is technically and economically preferable because melting at high temperatures is avoided. It needs, in the beginning, special production steps.

3.1 Powder Compacting

Powders of all commercially fabricated refractory metals (with typical particle sizes of 1–10 μm) and alloys are consolidated before sintering to reach densities averaging 60%–65% of the theoretical, mainly by two methods:

1. *Pressing of powders in rigid dies* on hydraulic presses using pressures of about 3000 bar. Dimensional precision is high. Molybdenum blanks can be machined without sintering; tungsten has to be presintered.
2. *Isostatic pressing* of powders in flexible containers. For very large parts, isostatic wet-bag pressing is the most common method, leading to homogeneous densities.

3.2 Sintering

Presintering of billets compacted from powder serves to facilitate handling and machining. The primary goal of sintering is densification to at least 90%–95% of the theoretical density, which is required to make further working operations successful. In addition to ensuring high density, purification occurs during sintering. The grain sizes after sintering are 10–30 μm.

Sintering is carried out in different ways:

1. Self-resistance heating or "direct sintering" has been used extensively for tungsten because of the extremely high temperatures required for practical sintering. Rods are heated to nearly 3000 °C in a hydrogen atmosphere. Tantalum and its alloys are sintered at 2600 °C in vacuum.
2. Ceramic-free high-temperature furnaces with vertical heating elements of tungsten sheet and heat shielding made of molybdenum alloys reach 2700 °C.
3. Induction radiant heating uses a cylindrical susceptor ring of tungsten or molybdenum that is inductively heated. The compact is heated up to 2400 °C by radiation.
4. For liquid-phase sintering and infiltration treatments, pusher-type horizontal furnaces operating up to 1600 °C are in use.

Among the many factors that influence sintering, the following are important:

1. Time and temperature are fundamental determinants of the sintering kinetics. Mass transport by diffusion is very slow for refractory metals at the sintering temperatures that are economically or technically feasible.
2. The purity of the compact greatly affects the sintering. Because molybdenum and tungsten are sintered at high temperatures, most impurities are molten and may have high vapor pressures and prevent achievement of high density.
3. Thermal gradients and heating rates are particularly important when large furnace loads are sintered.

3.3 Melting

Melting leads to a higher degree of purity in alloys, but the resulting coarse grain structure makes further processing difficult. To break up the large crystals in the ingots of molybdenum, they must be extruded further with additional cost. For vanadium, niobium, tantalum, and some alloys, arc and electron-beam melting are the most important methods.

3.4 Forming

Once a sintered ingot has been obtained, the further working follows the conventional path. Heating is mostly performed in hydrogen furnaces to avoid oxidation of molybdenum or tungsten and sublimation of their oxides.

The initial heating temperature necessary for molybdenum is 1300–1400 °C, and 1500–1600 °C for tungsten. During mechanical forming, the intermediate heating temperatures are successively lowered. After severe deformation (to produce thin wires, sheets, or foils), molybdenum and alloys can be processed at room temperature, and tungsten and alloys are rolled and drawn at temperatures of 500–200 °C without protective atmosphere. Normally, molybdenum alloys, tungsten, and tungsten alloys do not recrystallize during processing. Only pure molybdenum may recrystallize during heating at standard temperatures.

All forming processes that are common in other metal-producing industries are equally suited for refractory metals: hot and cold rolling, forging in dies, rotary swaging, extruding, flow turning, spinning, deep drawing, and wire drawing through dies. Drawing is also suited for the fabrication of long seamless tubes.

In contrast to standard metal-forming operations, the initial forming has to remove the pores that are present in the sintered condition. Figure 6 illustrates how the basic properties change during these procedures: Note how the ductile–brittle transition temperature is continuously decreased. Extremely high degrees of cold deformation can be achieved (for certain products, the diameters can be reduced by means of wire drawing to 10 μm). The production volume of certain products ranges from a few milligrams to maximum pieces weighing 5 tons.

3.5 Heat Treatments and Thermomechanical Treatments

Annealing of worked metals below their recrystallization temperature may relieve residual stresses and improve the ductility to a

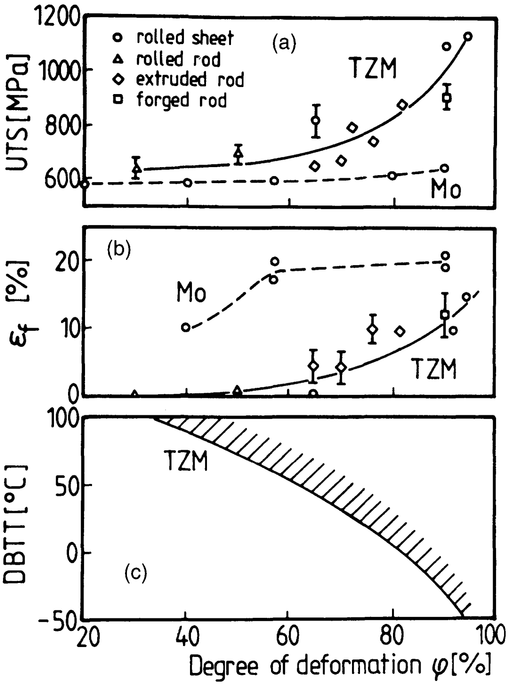

FIG. 6. The evolution of (a) the strength, (b) the fracture elongation, and (c) the ductile–brittle transition temperature (DBTT) in bending with the degree of working (Pink and Eck, 1996).

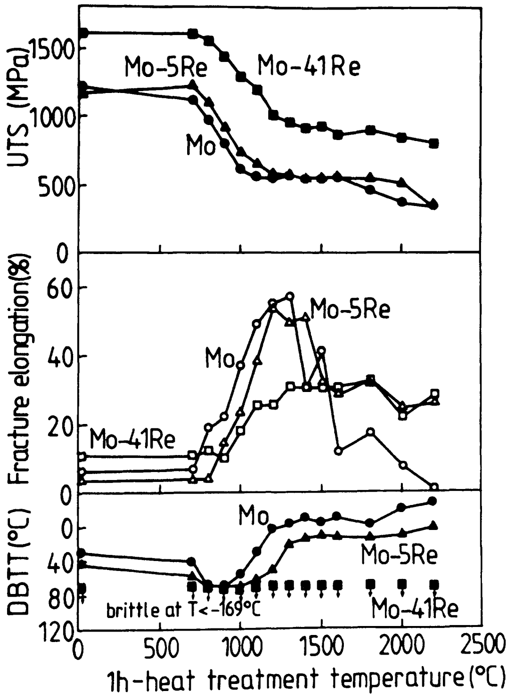

FIG. 7. The effect of annealing on room-temperature strength and elongation, and on the ductile–brittle transition temperature (DBTT) in bending, as demonstrated for the case of molybdenum and molybdenum–rhenium alloys (Pink and Eck, 1996).

certain extent (Fig. 7). Internal cracking and lamination of highly cold-worked wires and sheets can be avoided by applying intermediate partial recrystallization. In alloys, the precipitation of phases that strengthen or embrittle may be controlled by selected time-temperature cycles. Heat treatments between certain stages of forming are also performed to control final mechanical properties, e.g., in ZHM.

Heat treatments are performed under hydrogen or protective atmosphere or in vacuum. A problem of heat-treating alloys is the oxidation of alloying elements such as titanium, zirconium, and hafnium, and the depletion of carbon. In contrast to unalloyed molybdenum and tungsten, such alloys must not be annealed in nitrogen-containing atmospheres, because nitrides of the alloying elements formed on the surfaces have an embrittling effect.

Deformation in connection with the application of heat ("strain aging") may create special dislocation structures or influence the precipitation behavior, improving the strength of alloys (Trefilov and Milman, 1989).

3.6 Recycling

Most of the scrap is converted by similar chemical methods as applied for the powder production from the ore. High-purity scrap, which can be collected during the production, is recycled the easiest way, as high-purity alloying additions to superalloys and steels. Cleaned and recrystallized scrap can be recycled by comminution milling at temperatures below −40 °C using liquid nitrogen as coolant. Tungsten powder of such origin is further processed to tungsten heavy metal; molybdenum powder can be used for plasma spraying. Another process using the cooling effect of adiabatic expansion of nitrogen is called *coldstream process* and can be applied to tungsten. Pure tungsten–copper and tungsten heavy metal can be recycled by oxidation in air and successive reduction in hydrogen to the respective alloy powders, which can be reprocessed directly. Vacuum distillation could be used to recover infiltrated tungsten, but is expensive. Contaminated scrap of tantalum is pretreated mechanically and chemically, and arc-cast or electron-beam melted.

It is essential that rhenium as a "precious" metal be recovered. Chemical processing is an established procedure, as is also sublimation of the oxide.

4. FURTHER PROCESSING

4.1 Fastening and Joining

4.1.1 Mechanical Fastening For construction parts, refractory metals are preferably fastened by riveting, or with screws and nuts made of the same material. The threads of screws and nuts sinter at service temperature to the parts they are joining, and cannot be dismounted again. Rivets of molybdenum and tungsten are locally heated by torches to 800–1000 °C before pressing.

4.1.2 Welding Only "clean" welding methods are applicable (Lison *et al.*, 1989; Lambert, 1990). Welding under vacuum or pure inert gas prevents embrittling contaminations. Electron-beam (EB) and tungsten inert-gas (TIG) welding give best results. For thin parts, low-energy methods such as laser welding are also performed. Methods that do not create a liquid phase are friction welding (where temperatures of 2000 °C are reached at the contacting faces for a few seconds) and diffusion bonding. If not geometrically restricted, all joints of refractory metals and alloys can be made in this way, even when the parts to be joined are different materials.

Unalloyed tungsten and molybdenum embrittle as a result of excessive grain growth during welding. For perfect welding, the expensive filler alloys containing rhenium (Mo– or W–Re) can be used.

The welding behavior of V*a*-group metals is different from that of the group VI*a*. At high temperatures, protection from gases such as oxygen, nitrogen, hydrogen, water vapor, and carbon oxides is necessary, because the solubility for these gases is high, and even low contents have an embrittling effect (Buckman, 1988).

4.1.3 Brazing The disadvantage of brazing is the limitation imposed by the melting point of the braze, and in many cases by the temperature of the eutectic formed; another severe restriction is often the corrosion of the braze or the brazed area. Surfaces must be free of contaminations. The cleanest possible method is vacuum brazing. Brazing is important for the fabrication of refractory-metal composites where no other method is economically feasible: for brazing to ceramics, graphite, or semiconductor material (silicon).

A variety of braze materials are in use (Lambert, 1990) of which many are commercially available as foils. The temperatures where they can be used range from 700 °C in the case of silver–copper to 2200 °C for tantalum–hafnium alloys. Nickel, platinum, and rhodium are excellent brazes because of their good wetting power. Brazing materials for still higher temperatures are refractory metals or their alloys. For brazing of tungsten heavy alloys and contact materials, low-temperature brazing is sufficient: Elements contained in the binder phase such as nickel, copper, or silver are the base for brazing alloys.

4.2 Coatings

Great efforts, not too successful, have been made to find coating materials that protect molybdenum, tungsten, and rhenium against the formation of volatile oxides, and tantalum and niobium against taking atmospheric gases into embrittling solution, when these metals are exposed to elevated temperatures (Wang and Webster, 1982). The limiting temperature for creating coating under isothermal conditions is the recrystallization temperature, in order to prevent softening of the substrate material. Protective coatings develop microcracks when heated, and the smallest crack causes catastrophic failure at high temperatures.

Coatings are also deposited to protect surfaces against wear and to form layers for brazing.

4.2.1 Pack and Slurry Cementation Pack cementation based on silicides and aluminides has been developed. An improvement is to let a homogeneous slurry of complex silicides react with the refractory-metal surface (Lambert, 1990). Si–20Cr–20Fe prepared from elemental powders suspended in organic liquids is the standard coating for niobium and its alloys. Pack cementation of silicides developed for molybdenum has recently achieved commercial status. Layers may be up to 250 μm thick.

4.2.2 Electrolytic Coating

Aqueous electrolytic processes to form layers of nickel, silver, gold, platinum, rhodium, tin, and copper are more widely used for molybdenum than for tungsten. Very often a nickel flash improves adherence of subsequent layers. These coatings are necessary for brazing operations. Instead of depositing a very thin layer on wire, it turned out convenient to coat, for instance, molybdenum rods that are subsequently reduced in diameter to coated wires. Molybdenum electroplated with platinum or platinum–rhodium is widely used in the glass industry for operation in oxidizing atmospheres.

4.2.3 Flame and Plasma Spraying

One of the largest areas of consumption of molybdenum is for thermal spray coatings (Clare and Crawmer, 1982). Flame spraying (during which molybdenum wire is passed through an acetylene flame) is used for wear-resistant coatings on steel; their thickness may amount to millimeters. The wear-resistant layer contains pores that in service act as reservoirs for the lubricating oil. Plasma spraying in air (APS) and plasma spraying in vacuum (VPS, which avoids the oxidation of the layers and gives better densities up to 98%) of molybdenum starts from powder granules. Tungsten can be plasma-sprayed equally well, but is of limited importance.

Layers of iron- and/or nickel-based alloy powders containing 20 wt% chromium, 5 wt% aluminum, and up to 1% yttria protect molybdenum against oxidation. Electrodes for glass melting can be successfully protected against burner gases during the start of the furnace operation.

4.2.4 Chemical and Physical Vapor Deposition (CVD and PVD) Oxidation-resistant coatings up to 20 μm thickness can be deposited in dense layers by CVD and PVD processes. Coating compositions are based on silicides of refractory metals. Rhenium, platinum, and iridium layers are deposited by PVD processes such as magnetron sputtering and ion plating. These coatings serve as intermediate layers against carbide formation, for instance for refractory-metal–graphite composites.

CVD of layers (1 mm thick and more, with densities over 99%) of tungsten and tungsten–rhenium alloys on graphite discs serves to produce x-ray targets, large crucibles, and other shapes.

5. APPLICATIONS

5.1 Main Use of Refractory Metals and Alloys

The use of refractory metals is determined by their exceptional (both positive and negative) properties. Many applications of V*a* metals are based on their good resistance against chemical attack, which makes them a favorite material for equipment of the chemical industry. The VI*a* metals can only be applied in inert atmospheres or in vacuum. Under such conditions, their properties are exploited in the lighting industry, in the low- and high-energy electronic industry, for purposes of heat production (windings for furnaces, electrodes), and for high-temperature applications (construction material, heat shielding, jet propulsion).

Table 2 provides examples of uses of refractory alloys, based on the special properties of the materials. Up to service temperatures of 1600 °C, strain-aged ZHM and MHC are the superior refractory alloys (better than the strongest tantalum and niobium alloys or ODS nickel), especially when the data are compared in terms of a strength-to-density ratio, which is interesting for applications in aviation. From 1600 to above 2000 °C, tungsten alloys (e.g., with rhenium and hafnium carbides) are not surpassed by any other metallic materials, and are, or rather may become, important even for the aerospace industry despite their high specific weight. They are alternatives to the ceramic- and carbon-based materials.

Aside from conventional, though advanced, applications, pure refractory metals of normal quality are sometimes utilized for peculiar or exceptional purposes, e.g., chromium for sputtering targets, molybdenum for flame spraying (see Sec. 4.2.3), or rhenium as filament in carburizing atmospheres.

5.2 Special Requirements

5.2.1 High-Purity Materials Refractory metals, primarily tungsten, in use for microelectronics (microchips) need purity levels of 99.999% and 99.9999%. Lithium, sodium, and potassium contents can be reduced to below 0.1 μg/g. Sputtering targets of W–10Ti, used to deposit diffusion barriers on semiconductors, are presently produced by powder metallurgy with a purity of 99.99%, keeping the concentrations of the radioactive traces of thorium and uranium below 5 ng/g.

5.2.2 Porous Metals The most important use of tantalum is for capacitors for application in electronics (Belz, 1959). The metal surface is electrolytically oxidized to generate dielectric Ta_2O_5 layers. It is desirable to press the powders to as low a density as possible, because the capacitance is directly proportional to the surface area accessible to an electrolyte. Electron-beam melting purifies the tantalum, thereby allowing electrolytic capacitor devices to operate at higher voltages.

5.2.3 Alloys for Thermocouples For long-term use up to 2000 °C and for short-term use above, ductile Mo–41Re/Mo–5Re couples have been developed. The voltage of this type of thermocouple is equal to that of the tungsten–rhenium couples, in the range of Pt/Pt–10Rh.

5.2.4 Refractory Metals for Superconductors Pure metals of the V*a* and VI*a* groups are transition-element superconductors. Some of their alloys and compounds belong to the group of high-field type-II superconducting materials having simultaneously favorable critical transition temperatures for superconductivity as well as high upper critical magnetic fields (Warnes, 1990). NbTi alloys have the lowest superconducting characteristics (Kreilick, 1990). The A15 superconductors (Smathers, 1990) are better; they have a cubic crystal structure of type A_3B (A standing for the metals of the groups V*a* and VI*a*). In contrast to NbTi superconductors they are, as intermetallic phases, brittle. The third group of superconductors with high upper critical magnetic fields are the ternary molybdenum chalcogenides or "Chevrel" phases (Le Lay, 1990) of type $M\mathrm{Mo}_6X_8$, where M is a cation and X one of the chalcogenes sulfur, selenium, or tellurium. Amorphous modifications of refractory metals, e.g., $Mo_{30}Re_{70}$ or $Mo_{52}Ru_{32}B_{16}$, also belong to the group of high-field type-II superconductors, but their transition temperatures are below 9 K (Johnson, 1990).

It is interesting to mention that, for physical reasons, the NbTi superconductor has to be surrounded with a suitable metal like copper. The diameter of the superconducting wire must be extremely small (10 μm). Wire packages containing several hundreds or thousands of individual superconducting filaments are produced by sophisticated sequences of extruding and drawing, which can be done thanks to the good malleability of the material.

5.2.5 Use in Thermonuclear Reactors Refractory metals are favored because the operations are carried out at high temperatures, because they may retain their strength upon neutron irradiation, because of their low neutron-absorption cross section and short induced radioactivity, and because of their good corrosion resistance (e.g., to liquid alkali metals) at high temperatures (Mazey and English, 1984). Irradiation-induced void swelling and the elevation of the ductile–brittle transition are disadvantages. Refractory metals have also been considered for application in fusion reactors as "first-wall" materials, i.e., as materials situated closest to the plasma.

6. ECONOMIC ASPECTS

The production volume of refractory alloys is small compared to steel or nonferrous metals. Table 4 shows the total consumption of refractory elements in recent years. Table 5 presents a more detailed list giving the percentages of the different applications. An increase in the consumption of tungsten and molybdenum metal and alloys of 2%–3% per year is predicted. Chromium is a relatively new metal on the market. While its consumption in various forms is the highest of

Table 4. Worldwide consumption of refractory elements (in t/y).

Year	V	Nb[a]	Ta	Cr[b]	Mo	W	Re
1972	16 000	5 200					
1976		6 000	800		90 000	36 000	4
1981		20 000		2.5×10^6	100 000	40 000	5
1988			1 000		97 000	44 000	8.4
1990					105 000		15.9

[a] Nb (1989): 250–300 t metal products.

[b] Cr (1990): 75 t metal products.

all refractory elements (see Table 4), only a small fraction is produced in metallic form. The rhenium consumption will grow as new applications are found.

While prices for refractory metals fluctuate, the relations between the prices remain much the same. Molybdenum is the cheapest, followed by chromium, tungsten, vanadium, tantalum, and rhenium, in that order. Tantalum is about 10 times more expensive than molybdenum, and rhenium about 10–50 times more expensive than tantalum, depending on the market.

7. TOXICOLOGICAL ASPECTS

A certain amount of the elements of the V*a* and VI*a* groups is necessary for the human body. Of the refractory metals, only chromium, molybdenum, and vanadium appear in presently available tables of toxic materials (Goyer, 1990), yet the toxic effects appear not to be a consequence of the elementary metals, but of their chemical compounds.

Vanadium compounds as by-products of petroleum refining are present in urban atmospheres. Excess exposure can damage the respiratory tract, the kidney, liver, and bone marrow, and can cause gastrointestinal distress and affect the nervous system. Sources of chromium-compound pollution are coal-fired power plants, the metal-finishing industries, textile industries, and tanneries. Compounds of chromium are carcinogenic and provoke immune reactions. An overdose of a molybdenum compound picked up by cattle from pastures in the vicinity of molybdenum factories is responsible for anemia, diarrhea, retarded growth, and deformities of the joints. The symptoms can, in some cases, disappear when the intake is terminated.

GLOSSARY

Aging: Annealing of an alloy at a certain temperature for a certain time in order to create precipitates that strengthen the matrix (see **Dispersion-Strengthened Alloys** and **Heat Treatment**).

AKS Tungsten: Tungsten with minute additions of Al–K–Si compounds, which cause—after severe deformation—the formation of elongated grains upon recrystallization. The same effect can be also accomplished in other metals, e.g., in molybdenum.

Amorphous Metals/Alloys: Metals and alloys that could not assume their normal crystal structure during extremely rapid cooling from the melt.

Contiguity: A measure for the amount of direct contacts in a dual-phase alloy between phases (grains) of the same crystal structure.

Table 5. Percentage of refractory elements used for main products and industries.

	V	Nb	Ta	Mo	W (1990)			Re
Product	(1991)	(1972)	(1988)	(1978)	Euro.	Japan	USA	(1990)
Metal products	<2	3	83	3–4	15	17	23	10–20
In steels and (super)alloys	94	79	8	86	20	23	5	10–20
In hard metals			6		55	58	62	
For chemicals etc.		18	3	10	10	2	10	60–70

Dispersion-Strengthened Alloys: Metals that are strengthened by dispersoids (intermetallic or nonmetallic inclusions) of 5–500-nm diameter, which obstruct the movement of dislocations. The dispersoids in refractory alloys are introduced either to the metal powder or by aging treatments.

Doped Tungsten/Molybdenum: See **AKS Tungsten.**

Dual-Phase Alloys: Alloys consisting of grains of two different metals or alloys with different crystal structures, arranged in well-defined distributions (see **Contiguity**).

Ductile–Brittle Transition Temperature: Temperature below which a metal or an alloy with a body-centered cubic lattice structure loses its capability for plastic deformation, and breaks in a brittle manner.

Heat Treatment: Heating and slow or fast cooling of a metal or an alloy in order to improve its workabilities and/or its mechanical properties.

Liquid-Phase Sintering: Sintering under conditions that favor the appearance of one liquid phase which then surrounds the solid phase, thereby dissolving part of it in accordance with the solubility limits.

Nonsag (NS) Tungsten/Molybdenum: See **AKS-Tungsten.** The "sagging" is the deformation that a wire, heated to high temperatures, experiences when it is horizontally suspended.

ODS Alloys: Alloys strengthened by oxide dispersions.

Precipitates: Intermetallic or nonmetallic coherent or incoherent particles of 5–50-nm diameter that precipitate out of the crystal lattice of a supersaturated solid solution during aging.

Refractory Metals and Alloys: Metals and their alloys with high melting points (higher than 1900 °C), mainly of the V*a* and VI*a* groups of the periodic table of elements.

Sintering: Heating of a compressed metal powder at high temperatures (below the melting point), in order to achieve consolidation by diffusional material transport.

Solid Solution: An alloy in which the minor constituent is completely dissolved (and not precipitated) within the primary metal lattice.

Strain Aging: Increasing the strength by aging after a certain amount of cold work, due to a segregation of alloying atoms and/or precipitates at dislocations.

Thermomechanical Treatments: Mechanical working of an alloy in combination with a heat treatment, in order to improve the mechanical properties.

Tungsten Heavy Metal: Tungsten alloy consisting of spheroidal tungsten grains embedded in a low-melting binder phase (of iron and nickel, mostly).

Works Cited

Belz, L. H. (1959), in: W. R. Clough (Ed.), *Reactive Metals,* Metallurgical Society Conference Vol. 2, New York: Interscience, pp. 525–539.

Buckman, R. W., Jr. (1988), in: J. L. Walter, Ch. Jackson, T. Sims (Eds.), *Alloying,* Metals Park, OH: ASM International, pp. 419–445.

Clare, J. H., Crawmer, D. E. (1982), in: *Metals Handbook,* 9th ed., Vol. 7, Metals Park, OH: ASM, pp. 361–374.

Eck, R., Pink, E. (1992), *Int. J. Refract. Met. Hard Mater.* **11,** 337–341.

Fromm, E., Gebhardt, E. (1976), *Gase und Kohlenstoff in Metallen,* Berlin: Springer-Verlag.

German, R. M. (1985), *Liquid Phase Sintering,* New York: Plenum Press.

Goyer, R. A. (1990), in: *Metals Handbook,* 10th ed., Vol. 2, Metals Park, OH: ASM International, pp. 1233–1269.

Heitzinger, F. (1974), in: H. H. Hausner, W. E. Smith (Eds.), *Modern Developments in Powder Metallurgy,* Vol. 8, Princeton, NJ: American Powder Metallurgical Institute, pp. 371–390.

Honeycombe, R. W. K. (1984), *The Plastic Deformation of Metals,* London: E. Arnold.

Johnson, W. L. (1990), in: *Metals Handbook,* 10th ed., Vol. 2, Metals Park, OH: ASM International, pp. 804–821.

Kreilick, T. S. (1990), in: *Metals Handbook,* 10th ed., Vol. 2, Metals Park, OH: ASM International, pp. 1043–1059.

Lambert, J. B. (1990), in: *Metals Handbook,* 10th ed., Vol. 2, Metals Park, OH: ASM International, pp. 557–585.

Le Lay, L. (1990), in: *Metals Handbook,* 10th ed., Vol. 2, Metals Park, OH: ASM International, pp. 1077–1081.

Lison, R., Ambroziak, A., Horn, H. (1989), in: H. Bildstein, H. M. Ortner (Eds.), *Proceedings of the 12th Plansee Seminar,* Vol. 1, Reutte, Austria: Metallwerk Plansee, pp. 75–105.

Martinz, H. P., Prandini, K. (1993–1994), *Int. J. Refract. Met. Hard Mater.* **12,** 179–186.

Mazey, D. J., English, C. A. (1984), *J. Less-Common Met.* **100,** 385–427.

Pink, E., Eck, R. (1996), in: K. H. Matucha (Ed.), *Structure and Properties of Non-Ferrous Alloys,* Materials Science and Technology, Vol. 8, Weinheim: VCH-Verlag, Chap. 10, pp. 589–641.

Shen, Y. Sh., Lattari, P., Gardner, J., Wiegard, H. (1990), in: *Metals Handbook,* 10th ed., Vol. 2, Metals Park, OH: ASM International, pp. 840–868.

Smathers, D. B. (1990), in: *Metals Handbook,* 10th ed., Vol. 2, Metals Park, OH: ASM International, pp. 1060–1076.

Snow, D. B. (1989), in: E. Pink, L. Bartha (Eds.), *The Metallurgy of Doped/Non-Sag Tungsten,* London: Elsevier Science Publishers, pp. 189–202.

Tietz, T. E., Wilson, J. W. (1965), *Behavior and Properties of Refractory Metals,* London: E. Arnold.

Trefilov, V. I., Milman, Yu. V. (1989), in: H. Bildstein, H. M. Ortner (Eds.), *Proceedings of the 12th Plansee Seminar,* Vol. 1, Reutte, Austria: Metallwerk Plansee, pp. 107–131.

Wang, Ch. T., Webster, R. T. (1982), in: *Metals Handbook,* 9th ed., Vol. 5, Metals Park, OH: ASM, pp. 659–666.

Warnes, W. H. (1990), in: *Metals Handbook,* 10th ed., Vol. 2, Metals Park, OH: ASM International, pp. 1030–1042.

Zitter, H., Plenk, H., Jr. (1987), *J. Biomed. Mater. Res.* **21,** 881–896.

Further Reading

ASM-International Handbook Committee (Eds.) (1990), *Metals Handbook,* 10th ed., Vol. 2, Metals Park, OH: ASM International.

Benesovsky, F. (1979), *Wolfram, Gmelins Handbuch der anorganischen Chemie,* Ergänzungsband A1, 8 Auflage, Berlin: Springer-Verlag.

Bildstein, H., Ortner, H. M., Eck, R. (Eds.) (1985,1989,1993), *Proceedings of the 11th–13th Plansee Seminar,* Reutte, Austria: Metallwerk Plansee.

Kieffer, R., Braun, H. (1963), *Vanadin, Niob und Tantal,* Berlin: Springer-Verlag.

Kieffer, R., Jangg, G., Ettmayer, P. (1971), *Sondermetalle,* Wien: Springer-Verlag.

Pink, E., Bartha, L. (Eds.) (1989), *The Metallurgy of Doped/Non-Sag Tungsten,* London: Elsevier Sci. Publ.

Sully, A. H., Brandes, E. A. (1967), *Chromium,* London: Butterworths.

Trefilov, V. I., Mil'man, Yu. V., Firstov, S. A. (1975), *Fizicheskie Ocnovy Prochnosti Tugoplavkikh Metallov,* Kiev: Naukova Dumka.

Trefilov, V. I., Moiseev, V. F. (1978), *Dispersnie Chastitsy b Tugoplavkikh Metallakh,* Kiev: Naukova Dumka.

Yih, S. Y. H., Wang, C. T. (1979), *Tungsten: Sources, Metallurgy, Properties and Applications,* New York: Plenum Press.

REINFORCING FIBERS

A. R. Bunsell, *Centre des Matériaux P. M. Fourt, Ecole Nationale Supérieure des Mines de Paris, Evry, France*

Introduction 343
1. **Fiber Development** 344
2. **Reinforcement by Fibers** 345
3. **Fiber Processing, Structure, and Properties** 349
4. **Statistics of Fiber Reinforcement** 352
5. **Organic Fibers** 354
5.1 Molecular Structure and Mechanical Properties 354
5.2 Melt-Spun Fibers 354
5.3 Aramid Fibers 356
5.4 Theoretical Performance of Organic Fibers 357
5.5 Ultrahigh-Performance Polyethylene Fibers 358
6. **Glass Fibers** 359
7. **Boron and Silicon-Carbide Fibers on a Substrate** 360
8. **Carbon Fibers** 361
9. **Alumina Fibers** 363
10. **Ceramic Fibers from Organic Precursors** 365
11. **Whiskers** 367
12. **Conclusion** 367
Glossary 367
Works Cited 368
Further Reading 368

INTRODUCTION

A fiber is a long fine filament of matter with a diameter generally of the order of 10 μm and with an aspect ratio of length to diameter between a hundred and virtually infinity for continuous fibers. Fibers are one of the most extraordinary forms of ordinary matter. They are often much stronger and stiffer than the same material in bulk form. They have remarkable specific properties so that for the same strength, for example, a steel filament would be three or four times heavier than most fibers. The fineness of fibers allows them to be woven and draped into complex shapes. Cloth, woven or nonwoven, is a means of converting a one-dimensional filament into a two-dimensional component. Many and perhaps most fiber-reinforced composite structures are essentially two dimensional whether they are carbon-epoxy skins over a honeycomb structure, a filament-wound tube, or a car-body panel made by press molding. Three-dimensional weaving and needle punching are recent techniques of the textile industry that are also used for composites. Mechanical properties are but one aspect in the choice of a material, and others, such as thermal transfer or insulation behavior, water retention, and cost, have determined the use of fibers in the past as they do today in composite materials.

Natural fibers have long found applications in clothing, ropes, and sails; however, they were largely ignored for more structural applications up to the beginning of the twentieth century when artificial fibers were made by reprocessing the natural polymeric structure of cellulose. These viscose-rayon fibers were used to reinforce the rubber that was used for tires. The second half of the twentieth century has seen an ever widening use of synthetic fibers to reinforce plastics, metals, and ceramics. These are the subjects of this article on reinforcing fibers. The methods of production of the fibers are sketched and, most importantly, the relationships between the fibers' microstructures and their mechanical properties are treated. The mechanics of reinforcement by these long and thin filaments are explained, so revealing the importance of the rigidity of the fibers in fiber-reinforced composites but also the shear of the surrounding matrix material on the ultimate properties of the composites. The presence of often millions of fibers in

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

the cross section of any fibrous structure underlines the importance of a statistical approach to analysis of fiber properties. This is also put into context in the article.

1. FIBER DEVELOPMENT

The reinforcement of a medium by a network of fibers has evolved in the living world to be the natural solution for structures that respond to and resist mechanical stresses. The combination of two very different phases, the matrix medium and the fibers, allows a great degree of flexibility in the development of microstructural arrangements. In addition, the long fine form of the fibers, coupled with their mechanical strength and stiffness, allows natural structures to be preferentially reinforced in areas of greater stress (see COMPOSITE MATERIALS).

Although surrounded by naturally occurring fiber-reinforced materials and the wide use of wood for construction, man, with very few exceptions, only began to create fiber-reinforced structural materials at the beginning of the twentieth century. Up to that time there were no artificial fibers produced, but their development at the end of the nineteenth century coincided with the beginning of the automobile industry. Fiber-reinforced rubber forms the basis of the tire industry. Fiber-reinforced resins, which form the basis of the composite industry, have developed in the second part of the twentieth century and owe their existence to the production of organic resins and the beginning of the synthetic-fiber industry in the 1930s.

Fibers can be divided into three groups: natural, regenerated, and synthetic. Naturally occurring fibers can themselves be divided into three subgroups—vegetable, animal, and mineral fibers—as described by Moncrieff (1979). Vegetable fibers such as flax, cotton, and hemp are short or at least discontinuous and based generally on cellulose. Mineral fibers such as asbestos are also short, being rarely longer than a few centimeters, and their fully crystalline structure places them apart from the others. Animal fibers, which are based on animal proteins, can be discontinuous, as in the case of wool, or they can be virtually continuous, as with silk. The natural fibers all share a common trait of possessing molecular structures that themselves are arranged in a fibrillar manner. A cotton fiber, for example, has a microfibrillar structure arranged in a helical fashion around the fiber axis. The helix reverses in direction at points along the fiber length and is reminiscent of the structure of cotton threads. Threads consisting of discontinuous fibers owe their integrity to the intertwining of the individual fibers and the frictional forces developed between them that lock the structure together and allow a continuous thread to be produced, as indicated by Morton and Hearle (1975). The load transfer developed between the fibers within a thread is exactly analogous to the load transfer between matrix and fiber that accounts for the properties of fiber-reinforced composites; see Sec. 2.

Regenerated fibers benefit from the fibrillar nature of the molecular structure of plants, which is processed to form continuous filaments. In this way, this first class of artificial fiber to be produced avoids the difficulty of constructing its basic molecular structure.

Artificial fibers are only just one hundred years old, and the earliest produced were based on the naturally occurring cellulose molecule. Toward the end of the nineteenth century workers in Great Britain and in France had found means of dissolving natural cellulose and then extruding it through holes to produce filaments. The first commercial exploitation of these regenerated fibers was in France, where a patent to produce the earliest form of rayon was awarded to Count Hilaire de Chardonnet in 1885. This was the beginning of a revolution in fiber technology. The regenerated fibers did not exploit the total potential of the cellulose molecules, as they were broken up to shorter lengths during the fiber-manufacturing process. The cellulose was obtained from wood, usually spruce, made up of molecules containing about 1000 glucose residues, but these were broken down to about a fifth of their original length in the manufacture of rayon. The intention of the first fiber producers was to produce an artificial silk for textile purposes; however, they found a major application in the carcasses of tires for the developing automobile industry. Rayon fibers have evolved significantly over the years, and the descendants of these earliest artificial fibers are still to be found in some car tires today. Rayon fibers were used as precursors for the early production of carbon fibers, mainly in the U.S.A., and there are still some carbon fibers produced in this way; however, they have now mainly been superseded by other fibers for this purpose.

Truly synthetic organic fibers for which the long molecular structures, necessary for most fiber production, are made from relatively simple molecules have existed only since the late 1930s. The first organic truly synthetic fibers were developed independently in the U.S.A. and Germany in the late 1930s. These were polyamide or "nylon" fibers and these, together with polyester fibers developed in Great Britain in the 1940s, have become the most widely produced commodity synthetic fibers. Organic fibers have since been developed with elastic moduli rivaling that of steel but with only a fifth or less of the weight. Two approaches have been taken to achieve this remarkable behavior. First, the molecular structures of organic fibers have been made more complex and consequently stiffer. The second approach has been to align the molecular structure of simple polymers to achieve greatly improved tensile properties.

Glass fibers, first commercially produced in the U.S.A. in the early 1930s, were the first artificial fibers with a relatively high Young's modulus. The manufacture of glass fibers by drawing the molten glass through a spinneret is similar to the process used to produce fine thermoplastic fibers. However, the act of drawing aligns the molecular structure of the thermoplastic fiber producing an anisotropic structure. This is not the case with the glass fibers, which remain isotropic, but in both instances the act of drawing and plastically deforming the material during fiber production removes surface defects and accounts for the very high strength of glass fibers compared with bulk glass.

Although glass fibers possess Young's moduli considerably higher than those of conventional textile fibers, still greater tensile stiffness is required for structural composites in many applications and notably in the aerospace industry. Boron fibers produced in the early 1960s, first in the U.S.A. and then in France, allowed a leap from 70 GPa, which is the elastic modulus of glass, to over 400 GPa. These remarkable fibers were technically a success, but their use has been limited by their extremely high cost.

Carbon fibers, first produced in Great Britain in the middle 1960s and produced commercially for the first time there in 1967, were also developed in parallel in Japan, where commercial production began in 1970. These carbon fibers, which are made by the pyrolysis of polyacrylonitrile precursor fibers, have revolutionized the aerospace industry and are finding increasing applications in other industrial areas. An alternative route for producing carbon fibers is from pitch, which is the residue from oil refining or the coking of coal. Fibers with very high Young's moduli, approaching that of diamond, can be made from pitch.

Large-diameter ceramic fibers of silicon carbide on a core have existed since the early production of boron fibers; however, fine ceramic fibers started to appear at the end of the 1970s and in the 1980s. One group of these fibers, developed in Great Britain and in the U.S.A., is based on fine-grained alumina whereas others, developed in Japan, are silicon carbide or silicon nitride. This class of fibers allows light-metal alloys to be reinforced and, probably more importantly, the production of fiber-reinforced ceramics for high-temperature structural applications. They are made by the pyrolysis of precursor fibers in an analogous way to the production of carbon fibers.

Single-crystal continuous fibers can also be made. Such fibers, based on α-alumina and first commercialized in the U.S.A., usually have large diameters around 100 mm; however, the absence of grain boundaries offers the possibility of maintaining useful mechanical properties to 1500 °C and above.

The range of fibers available for reinforcement has been described in detail by Bunsell (1988).

2. REINFORCEMENT BY FIBERS

Any shape of inclusion could be put into a matrix, but a long thin form provides the most efficient reinforcement. This can be seen if a single fiber of length l and circular cross section is considered embedded in a cylindrical block of radius r_1 as illustrated in Fig. 1. Consider the case of an elastic fiber having a Young's modulus much greater than that of the matrix, which itself is elastic. In the absence of the fiber any applied

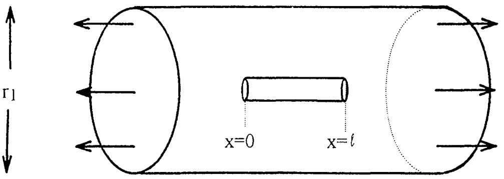

FIG. 1. Consider a short stiff elastic fiber embedded in a cylinder of elastic matrix material that is less rigid and more deformable than the fiber.

load on the specimen would produce a uniform deformation of the matrix. If the specimens were viewed from the side and divided up into identical square zones these would be transformed into identical rectangular areas. In the case in which a high-modulus fiber is embedded in the matrix and there is good adhesion between the two, a different situation occurs under load. This matrix near the fiber is limited in deformation by the lower strain produced in the fiber, whereas further away the matrix is free to deform as in the unreinforced case. Figure 2 shows that the deformations produced in the presence of a higher-modulus fiber are far from simple, as in this case the displacements produced near to and far from the fiber are different and induce shear stresses in the matrix around the fiber. The shear stresses transfer load from the matrix to the fiber, which supports a load that increases from zero at the ends, assuming no load transfer across the ends, to a maximum in the center.

This model of fiber reinforcement was first proposed by Cox (1952).

The load P generated in the fiber produces in it a deformation of U so that we can describe the load transfer due to the shear of the matrix in mathematical form by

$$\frac{dP}{dx} = H(U - V), \tag{1}$$

where V is the overall composite displacement, H is a factor of proportionality, and x is the distance along the fiber from the end.

We can also write

$$P = E_f A_f \left(\frac{dU}{dx}\right), \tag{2}$$

where E_f is the Young's modulus of the fiber, A_f is the fiber cross-sectional area, dU/dx is the strain in the fiber at x, and $P/A_F = \sigma_X$, the stress supported by the fiber at a distance x from the origin taken as one end of the fiber.

We also know that the strain e produced in the composite, considered as a continuum, would be the same all over the specimen so that

$$\frac{dV}{dx} = e = \text{const.} \tag{3}$$

From Eq. (1),

$$\frac{d^2P}{dx^2} = H\left(\frac{dU}{dx} - \frac{dV}{dx}\right).$$

So from the above three equations we have

$$\frac{d^2P}{dx^2} = H\left(\frac{P}{E_f A_f} - e\right), \tag{4}$$

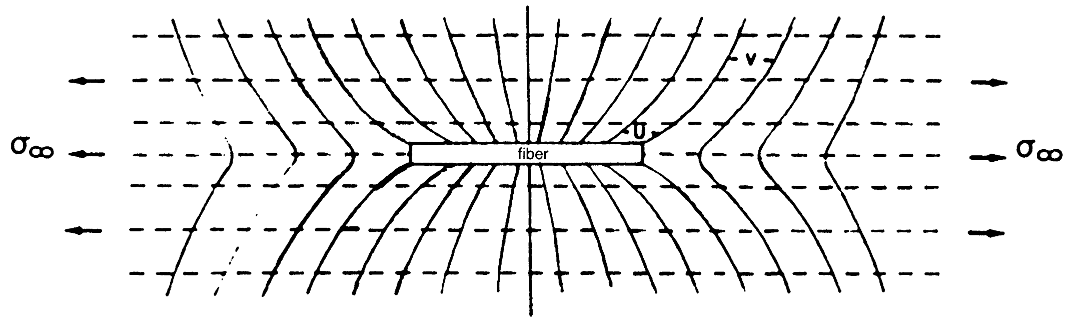

FIG. 2. A uniform tensile load parallel to a short fiber embedded in a deformable matrix produces a nonuniform strain of the matrix. The stress σ_∞ applied to the specimen produces an overall deformation of V; however, the deformation in the neighborhood of the fiber is U.

which can be solved by the substitution

$$P = E_f A_f e + R \sinh \beta x + S \cosh \beta x, \tag{5}$$

in which R, S, and β are parameters that can be determined making use of straightforward but elaborate algebra involving hyperbolic functions and by noting that $P = 0$ where $x = 0$ and l. This gives

$$\beta = \sqrt{H/E_f A_f}$$

and

$$\sigma_x = E_f e\left[1 - \frac{\cosh \beta(l/2 - x)}{\cosh \beta l/2}\right]. \tag{6}$$

Equation (6) reveals that the load supported by the fiber increases from each end to attain a maximum at the midpoint of the fiber, as is shown in Fig. 3(a). In order that the average strain in the fiber be that of the composite so that $e = \sigma_x/E_f$, the function inside the brackets in Eq. (6) must become equal to unity. This is only possible when the fiber is continuous and l tends to infinity.

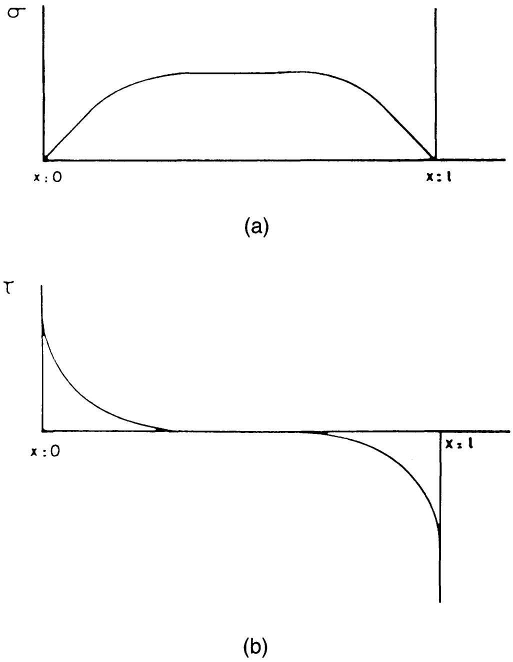

FIG. 3. (a) The stress transferred to the fiber tends to a plateau in the center of the fiber. (b) The shear of the matrix around the fiber is a maximum at the fiber ends and tends to zero at the center of the fiber. The shear in the matrix is in opposite directions at the two ends of the fiber.

The average stress in a short fiber of length l is given by

$$\sigma_m = \frac{E_f e}{l}\int_0^l\left[1 - \frac{\cosh \beta(l/2 - x)}{\cosh \beta l/2}\right]dx.$$

Simplifying gives

$$\sigma_m = \frac{E_f e}{l}\left[l - \frac{2}{\beta}\tanh\frac{\beta l}{2}\right].$$

Dividing through by l we obtain

$$\sigma_m = E_f e\left[1 - \frac{\tanh \beta l/2}{\beta l/2}\right].$$

The effective modulus of a short fiber embedded in a matrix is therefore given by

$$\frac{\sigma_m}{e} = E_{fM} = E_f\left[1 - \frac{\tanh \beta l/2}{\beta l/2}\right]. \tag{7}$$

Equation (7) shows that a short fiber does not have the same reinforcing efficiency as does a long fiber. The longer the fiber, however, the closer it gets to being as efficient as a continuous fiber, as the load-transfer lengths at each end remain the same. A practical difficulty that this analysis does not consider is that of aligning short fibers in the direction of the applied load.

The shear forces around the fiber can be calculated by considering the equilibrium state in which the shear forces in the matrix are balanced by the load in the fiber so that

$$\frac{dP}{dx} = -2\pi r_0 \tau(r_0), \tag{8}$$

in which $\tau(r_0)$ is the shear stress at the interface and r_0 is the fiber radius.

From Eq. (1) we can write

$$H = -2\pi r_0 \tau(r_0)/(U - V) \tag{9}$$

If w is the actual displacement of the matrix and r is the radial distance from the fiber axis, then at $r = r_0$, $w = U$, and at $r = r_1$, $w = V$. For equilibrium,

$$2\pi r \tau(r) = 2\pi r_0 \tau(r_0). \tag{10}$$

This shear deformation produced in the matrix is given by

$$\frac{dw}{dr} = \frac{\tau(r)}{G_m} = \frac{\tau(r_0)r_0}{G_m r}, \tag{11}$$

where G_m is the matrix shear modulus.

Integrating Eq. (11) between r_0 and r_1 given that Δw is the difference in displacements produced in the matrix at r_0 and r_1, which is $U - V$, we obtain

$$U - V = \frac{\tau(r_0)}{G_m} r_0 \ln\left(\frac{r_1}{r_0}\right), \tag{12}$$

so that from Eq. (9),

$$H = \frac{2\pi r_0 G_m}{\ln(r_1/r_0)}. \tag{13}$$

Since

$$\beta = [H/E_f A_f]^{1/2},$$

we can see that

$$\beta = \left[\left(\frac{G_m}{E_f}\right)\left(\frac{2\pi}{A_f \ln(r_1/r_0)}\right)\right]^{1/2}. \tag{14}$$

The significance of this result is that the larger G_m/E_f, the more rapidly the stress in the fiber increases with distance from the fiber ends.

Equations (6) and (8) allow us to calculate the shear stress $\tau(r_0)$ in the matrix at the interface so that, as

$$P = \pi r_0^2 \sigma_x$$

and

$$\tau(r_0) = -\frac{dP}{dx}\frac{1}{2\pi r_0},$$

it follows that

$$\tau(r_0) = e\left[\frac{G_M E_F}{2\ln(r_1/r_0)}\right]^{1/2}\left[\frac{\sinh\beta(l/2 - x)}{\cosh\beta l/2}\right]. \tag{15}$$

Figure 3(b) shows the shear-stress distribution described by Eq. (15). It can be seen that the shear stresses are a maximum at the ends of the fiber and decrease to zero at the midpoint. The shear stresses are in opposite directions at opposing ends of the fiber; hence their opposite signs.

Equations (6) and (15) reveal that for the case of a long fiber we can write

$$\frac{\tau_{\max}}{\sigma_{\max}} = \left[\frac{G_m}{2E_f \ln(r_1/r_0)}\right]^{1/2}. \tag{16}$$

As indicated the above analysis is for an elastic fiber embedded in an elastic matrix. Now let us consider an elastic fiber embedded in an ideally plastic matrix. The plastic case is simpler than the elastic situation considered above, as it can be assumed that the shear stress in the matrix is constant and equal to the yield stress τ_y, so that Eq. (8) can be directly integrated giving

$$P = -2\pi r_0 \tau_y x|_0^x,$$

which divided through by the cross-sectional area of the fiber, πr_0^2, can be written as

$$\sigma_x = 2\tau_y x/r_0. \tag{17}$$

Equation 17 shows that there is a linear buildup of stress from the fiber ends when it is embedded in a plastic matrix. In the case of a long fiber this buildup of stress is limited by the overall stress applied to the composite, as shown in Fig. 4. In this case the central section of the fiber experiences a constant stress. If the stress applied to the composite is increased the load-transfer length is increased and the central constant-stress region diminishes. Fracture will occur in the central section if the applied stress exceeds the fiber failure stress. If the fiber is long or continuous multiple fractures can occur. There clearly exists a length of fiber for which the maximum stress that can be attained at the midpoint is equal to the fiber failure stress. This is known as the critical length $2l_c c$. At the critical length Eq. (17) can be rewritten so that

$$l_c/r_0 = \sigma_f/2\tau_y, \tag{18}$$

where σ_f is the failure stress of the fiber.

It can be seen therefore that the aspect ratio l_c/r_0 is the important factor in determining the efficiency of a reinforcement. A long fine reinforcement, in other words a fiber, is clearly the ideal reinforcement.

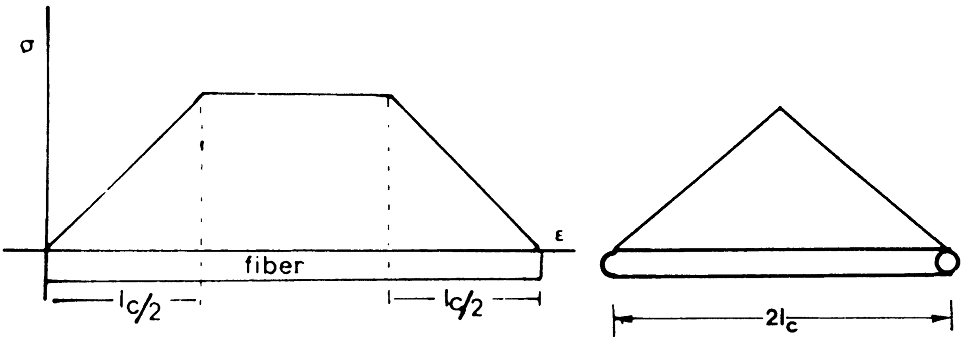

FIG. 4. An elastic fiber embedded in a perfectly plastic matrix leads to a uniform buildup of stress in the fiber. If the fiber is long this leads to a constant stress borne by the fiber in the midsection. A fiber of critical length $2l_c$ is just long enough that the maximum stress than can be attained at its midpoint is equal to its fracture stress.

3. FIBER PROCESSING, STRUCTURE, AND PROPERTIES

Many fibers, particularly organic and glass fibers, are produced by being drawn or stretched between leaving the spinneret and being wound onto a bobbin. The drawing process reduces the filament diameter from 1 or 2 mm to the order of 10 μm. This change of form has immediate consequences on the strength of the fiber even if effects of induced orientation, such as occur with organic fibers, are ignored.

Glass fibers can be several hundred times stronger than the same glass in bulk form, and this is due to the drawing of the fiber, which occurs during its manufacture. Glass is a brittle material, and its strength is determined by the size of defects it contains. Inglis (1913) first showed that irregularities such as a small crack in a brittle solid induce stress concentrations under load and that these localized high stresses σ precipitate failure. The finer the crack, the greater the stress concentration and the lower the strength of the body.

Most of the defects on glass are on the surface, as it is there that it is most easily damaged. The process of drawing the glass into a fine filament greatly deforms it and in particular modifies the surface. The movement of matter from the interior of the filament to its surface eliminates the surface defects that weaken the bulk material and so greatly increases its strength. If the filament is to retain its strength the surface must be protected from damage by abrasion and environmental effects. For this reason glass fibers receive a protective surface coating at an early stage in their manufacture.

Strength can be seen therefore not as an intrinsic quality of the material but as determined by the probability of finding defects in the body. A statistical approach to analyzing fiber failure is necessary, as is explained below.

Glass fibers are generally considered to be isotropic. The effect of drawing on polymers, however, is to orient the macromolecular structure more or less in the drawing direction, which, in the case of a fiber, is parallel to its axis. The greater degree of orientation increases both the Young's modulus and failure stress of the fiber and reduces its strain to failure. This is because the applied load is supported by an increasing number of covalent atomic bonds in the polymer chains. As a direct consequence, however, the structure of drawn organic fibers becomes anisotropic with the transverse strength being more influenced by the secondary weak atomic bonds linking the macromolecules. The alignment of a polymer structure does not always necessitate mechanical drawing, as by inducing a mesophase or liquid-crystal phase the polymer may be induced to possess an inherent alignment while still in the liquid phase, which enables very high-modulus fibers to be spun. Fibers such as the aramid fibers, as well as some pitch-based carbon fibers, are produced in such ways as to take advantage of the intrinsic anisotropy of the liquid mesophase.

The first organic fibers were based on a

Table 1. Molecular structure of several organic fibers showing the development of increasingly rigid molecules.

Fiber type	Repeat unit in the macromolecule	Maximum elastic modulus (GPa)
Polyamide 6/6 (nylon 6/6)	$—NH—CH_2—CH_2—CH_2—CH_2—CH_2—CH_2—NH—CO—CH_2—CH_2—CH_2—CH_2—CO—$	5
Polyamide 6 (nylon 6)	$—NH—CH_2—CH_2—CH_2—CH_2—CH_2—CO—$	4
Polyethylene terephthalate	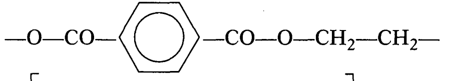	18
Poly (*m*-phenylenediamine-isophthalamide) (Nomex)	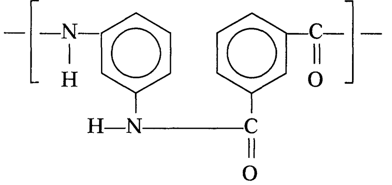	17
Poly(*p*-phenylene terephthalamide) (Kevlar)	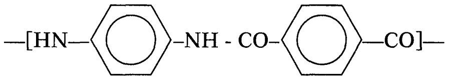	160

structure made up of linear macromolecules. Table 1 shows that polyamide fibers consist of macromolecules made up of repeat units that determine their name. The first structure shown is of polyamide 6/6, which consists of repeat units containing two series of six carbon atoms. The molecule is flexible, as it is able to bend at each bond. The structure of the fiber consists of microfibrils aligned more or less parallel to the axis and composed of regions of molecules folded into well-ordered crystallites as shown ideally in Fig. 5. The crystallites may be rectangular as indicated or rhomboidal. This type of structure is considered likely for both polyamide and polyester fibers. The folded molecules and the amorphous regions of the structure, in which the molecules are randomly organized, lead to a low Young's modulus.

The polyester fiber, however, contains an

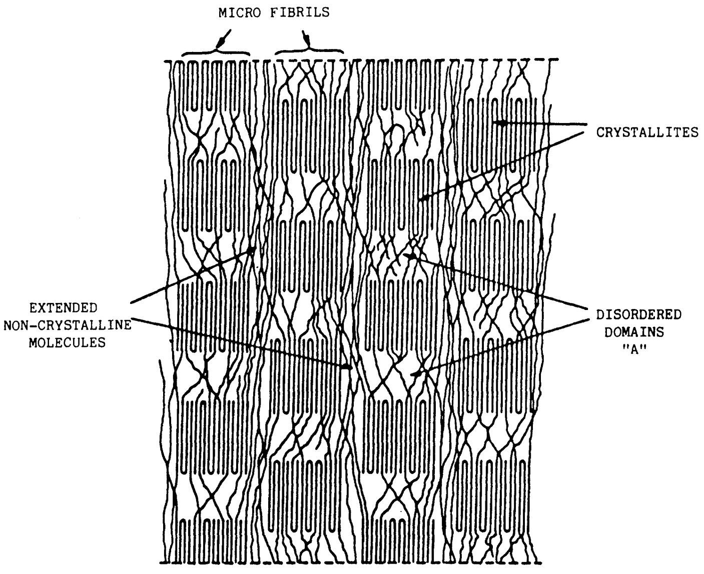

FIG. 5. Schematic and idealized representation of the microfibrillar structure of polyester and nylon fibers.

aromatic ring, which makes it less flexible than the polyamide molecule. The consequence of this increased molecular rigidity is an increase in the tensile modulus of polyester fibers compared with polyamide fibers.

Fibers with increasingly rigid molecular structures have been developed, and this has brought about better thermal stability and much higher Young's moduli. The Nomex fiber described in Table 1 has much improved thermal properties when compared with polyamide fibers, giving it an increase in useful temperature range of 100 °C, and as such is used for protective clothing where the possibility of fire may be encountered. The Nomex fiber is not produced for high mechanical properties and fails with a breaking strain of about 22% similar to that of polyester. It is used as a paper to form the core of honeycomb structures. It has an initial Young's modulus of 17 GPa.

The production of fibers from already anisotropic liquid-crystal solutions allowed a leap in initial modulus as well as final properties to be made, as described by Black and Preston (1973).

The aromatic polyamide PPT, described in Table 1 and produced by Du Pont under the name Kevlar, was the first commercial fiber of its type and launched the aramid family of fibers for composite and other applications. Table 1 shows that the high-modulus form of this fiber has a much higher tensile modulus than other organic textile fibers.

An alternative approach to producing high-modulus organic fibers is to ensure that the macromolecules are aligned parallel to the fiber axis. In order to be able to disentangle the molecular structure relatively simple polymers have so far been used such as polyethylene. Fibers of this type are produced either by spinning them from dilute solutions or by optimizing the melt-spinning process. Very high Young's moduli can be achieved, approaching that of high-performance aramid fibers. The polymer retains its low melting point, however.

The ultimate tensile properties of a solid depend on the interatomic bonds. These depend on the interaction of valence electrons associated with neighboring atoms and not on the atomic weight. This has encouraged attempts to produce high-modulus fibers from the lightest elements in the periodic table. The lightest element with which it is conceivable to make fibers is the fourth in the periodic table, beryllium. It is possible to make filaments from beryllium but the properties are not exceptional and above all the material is toxic. The fifth element, boron, however, has been made into fibers by vapor deposition onto a tungsten substrate. The properties of boron fibers, which were first produced in the early 1960s, are very good with a Young's modulus twice as high as that of steel and excellent properties in compression. The specific gravity of boron fibers is 2.45 and their diameter is normally 140 μm. The slow rate of fiber production and the fact that it is impossible to produce more than one fiber at a time with each production unit leads to an unacceptably high cost of the fiber for most applications. As a consequence almost all of the applications for which boron fibers had been considered have used carbon fibers as reinforcements.

Carbon fibers, which were developed in the middle and late 1960s, can be made with a range of properties depending on the temperature used during production. Some of the very first artificial fibers to be produced at the end of the nineteenth century were of carbon. Cotton fibers were pyrolyzed to make carbon fibers and used as elements in some of the first electric lamps produced by Edison. These fibers had little mechanical strength, and the high-performance fibers produced in the 1960s were made by the pyrolysis of polyacrylonitrile (PAN) fibers. High modulus is achieved by preventing shrinkage of the fiber from occurring during pyrolysis, which produces a 45% loss in weight. The inherent anisotropy of the precursor fiber is preserved in this way to give carbon fibers made up of polycarbon basic structural units arranged in the form of ribbons parallel to the fiber axis. This longitudinal organization is not reflected in the radial direction, in which the structure is completely random.

Most of the carbon fibers used in composites are now of the high-strength type. Remarkable progress has been made in improving the properties of these fibers through improvements in the polyacrylonitrile precursor fibers, optimization of the pyrolysis process, and a reduction of fiber diameter from 8 to 5 μm. This has produced an increase in breaking strain from 1% to more than 2%.

Pitch has attracted much attention in the U.S.A. and Japan as a source for producing carbon fibers. The high carbon yield, around 90%, together with the easy availability of pitch, seemed at first to offer the potential of a low-cost route for making carbon fibers. The difficulties of purifying the pitch and market forces seem to destine these fibers to primarily very high-modulus applications although the range of moduli is from 200 to 900 GPa. The difference in behavior of these fibers compared with that of PAN-based carbon fibers is due to a more graphitic structure, which leads both to higher moduli but also lower compressive strengths.

Alumina-based ceramic fibers are made from a variety of types of alumina. The alumina fibers possessing the highest Young's moduli are made from α-alumina, which is the most stable and crystalline form of aluminum oxide. The granular structure of α-alumina produces high-modulus but very brittle fibers. Other forms of alumina such as η and γ alumina, having much smaller grains usually associated with an amorphous silica phase, give fibers with lower Young's moduli and less brittle behavior.

Ceramic fibers, made from organo-silicide precursors that are pyrolyzed, are based on silicon associated with carbon, often oxygen, and sometimes nitrogen and titanium. Any crystalline phase is usually β-silicon carbide, which is often embedded in an amorphous matrix of silicon and oxygen, the latter element being introduced during pyrolysis. The most recent fibers of this type limit the oxygen content and hence the amorphous phase by modifying the pyrolysis conditions. The result is improved fiber properties.

4. STATISTICS OF FIBER REINFORCEMENT

Fibers are very fine filaments of matter and in any section of a composite structure there will be found thousands and most probably millions of fibers. Such large populations lend themselves to statistical analysis, and the inherent scatter in fiber properties ensures that a statistical approach to describing their properties is necessary.

It is not possible to give a single value for the strength of a fiber because they show considerable scatter and the mean strength decreases with fiber length. Figure 6 shows the scatter of strength of Toray's T300 carbon fibers. Both the scatter and the length dependence are due to the distribution of defects in the fibers, and the longer the fiber the greater chance there is of it containing a sizable defect and so being considerably weakened. Failure of the fiber as a function of applied load is therefore controlled by the random distribution of defects and requires a statistical treatment for its analysis.

The effect of a random distribution of a single type of defect on the strength of a solid has been described by Weibull, who likened the failure of the solid to the breaking of a chain in which the weakest link controls failure. If the probability of failure of a link under an applied stress σ is P_0, its probability of survival is $1 - P_0$, and the survival probability of the chain is $(1 - P_0)^n$ where n is the number of links. If P_n is the probability of the chain's failure we can write

$$1 - P_n = (1 - P_0)^n \qquad (19)$$

or

$$P_n = 1 - \exp[n \ln(1 - P_0)]. \qquad (20)$$

In the case of a body of volume V let us

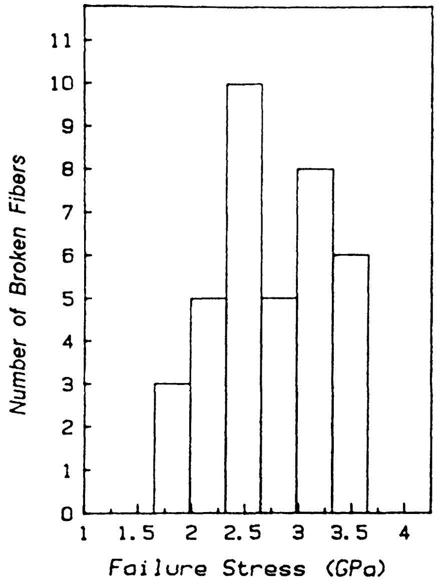

FIG. 6. The failure stress of carbon fibers shows considerable scatter.

consider it divided up into small volumes V_0 analogous to links in a chain. In this case n is analogous to V/V_0.

The quantity $\ln(1 - P_0)$ depends only on the stress if V_0 is considered constant, so that for the whole specimen the risk of failure is, from Eq. (20),

$$P_v = 1 - \exp\left[-\int_v f\left(\sigma, \frac{1}{V_0}\right)\right]dV.$$

Weibull put $f(\sigma, 1/V_0)$ equal to $[(\sigma - \sigma_u)/\sigma_0]^m$ where σ is the applied stress, σ_u is the stress below which there is a zero probability of failure, and the two material parameters σ_0 and m are measures of the density and variability of the flaws in the materials.

In a straight tensile test in which the stress σ is constant throughout the body we can write

$$P_v = 1 - \exp\{-[(\sigma - \sigma_u)/\sigma_0]^m V\}. \tag{21}$$

This is the distribution function characterizing the scatter of strengths found with repeated tests on specimens of volume V and material constants σ_u, σ_0, and m. The parameter m is known as the Weibull shape parameter and characterizes the distribution of strengths. The greater the value of m the less is the scatter of results.

Writing the probability of survival $(1 - P_v)$ as P_s we obtain from Eq. (21), if σ_u is zero,

$$\ln \ln(1/P_s) = \ln V + m \ln\sigma - m \ln\sigma_0. \tag{22}$$

If we now consider a cylindrical fiber with diameter D that can vary with different fibers taken from the same bundle but all having the same lengths, we can write from Eq. (22)

$$\ln \ln(1/P_s) = m \ln\sigma + 2 \ln D + \text{const.} \tag{23}$$

If $\ln \ln 1/P_s - 2 \ln D$ is plotted as a function of $\ln\sigma$, the Weibull shape parameter may be measured from the gradient of the curve obtained, as is shown in Fig. 7 for carbon fibers. The fiber diameter was found not to vary significantly and so was considered a constant in plotting Fig. 7. In this way the distribution of fiber properties as described by Weibull statistics can be obtained from a series of tensile tests, usually not less than twenty-five depending on the scatter of results, with specimens of equal lengths.

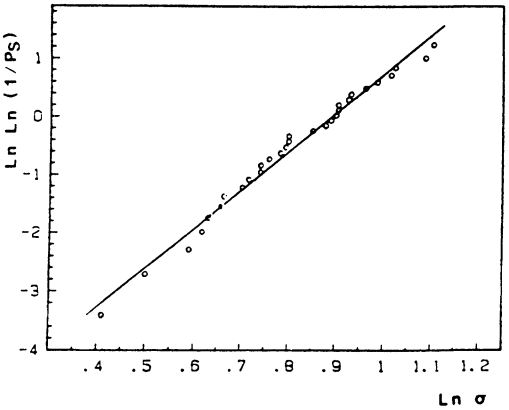

FIG. 7. A Weibull plot of the values shown in Fig. 6 reveals a straight line the gradient of which gives the Weibull modulus.

As explained above, the average strength of fibers taken from a bundle decreases as a function of length, and this offers another approach to obtaining the Weibull shape parameter. If the median failure stress (P_s = 0.5), which is not necessarily the same as the mean stress, for a number of specimens is denoted by σ_{50} Eq. (22) can be written

$$\ln \ln\left(\frac{1}{0.5}\right) = \ln l + 2 \ln D + m \ln\sigma_{50} + \text{const.} \tag{24}$$

If the mean fiber diameter from one lot of specimens to another is considered to be constant we can write

$$\ln\sigma_{50} = -(1/m) \ln l + \text{const.} \tag{25}$$

A plot of $\ln\sigma_{50}$ as a function of $\ln l$ for several gauge lengths gives a straight line curve with a gradient of $-1/m$.

Theoretically the values of m calculated by the two approaches should be identical; however, this is hardly ever the case. The difference is due to the difficulty of manipulating fine fibers during which defects may be created and, more significantly, weaker fibers are broken before they can be tested. This difficulty inevitably increases with increasing specimen length so that merely handling the fibers provides an unavoidable selection process and modifies the failure distribution curve. In addition, and although the approach based on Weibull statistics is extremely valuable and widely developed in analyzing fiber data, there are doubts as to whether the failure of fibers falls within the limits of the theory. This is because the fibers are very fine and it is not true that the same distribution of defects can be expected to exist in all small sections of the fibers.

The value of the Weibull shape parameter for most fibers is low, often around 5, which reflects the large dispersion found in their properties. This value of m is similar to that obtained with ceramics and can be compared with the values for m obtained for steel that give $m > 20$.

Because all fibers show great dispersion in their properties they do not all fail in the same manner. The mode of failure may involve slow crack propagation from surface or internal faults leading to eventual catastrophic failure. This is the case of many organic fibers, which may deform plastically, although very-high modulus organic fibers split on failure as a result of their high anisotropy. Other fibers such as carbon, glass, boron, and ceramic fibers show no plastic deformation and behave as purely brittle elastic bodies, their ultimate failure being determined by defect size as described by Griffith (1920) for elastic bodies.

5. ORGANIC FIBERS

It was not until around 1938 that the first truly synthetic organic fibers were produced from short molecules, and continuous glass fibers were also commercially produced just before this date. Both the polyamide and glass fibers were developed in the U.S.A., respectively by E. I. du Pont de Nemours and what became Owens Corning. Almost simultaneously with the work of Carothers at Du Pont, polyamide was produced in Germany by I. G. Farben.

5.1 Molecular Structure and Mechanical Properties

Since the beginning of production of polyamide, or nylon, fibers other organic fibers have been produced, and a trend toward the production of increasingly more rigid and complex molecules used in making the fibers is clearly discernible from Table 1. The aramid family of fibers made up of aromatic polyamides possesses remarkably high Young's moduli, more than twenty times that of conventional polyamide fibers. The more rigid molecule confers not only greater elastic modulus and failure stress but also better thermal resistance. This latter quality is a considerable advantage when these fibers are compared with other high-modulus organic fibers made from simpler molecules such as polyethylene made by straightening out the molecular structure and exploiting the covalent bonds in the backbone of the molecule.

5.2 Melt-Spun Fibers

Nylon (polyamide) and polyester (polyethylene terephthalate) fibers are extensively used for reinforcing rubber, and these flexible composites are used in fan belts, hoses, flexible drives, moving walkways, tires, and other applications. They account for 28% and 48%, respectively, of the total synthetic-fiber production. Both types of fibers are produced by melt spinning and are subsequently drawn to increase their Young's moduli. They are both semicrystalline with their molecular structures arranged as shown in Fig. 5 in microfibrils consisting of crystalline blocks of regularly folded molecules linked by less well-ordered molecules. Surrounding the microfibrils there are elongated but irregularly arranged molecules forming a mesomorphous phase. Crystallinity is often about 60% in polyamide fibers and somewhat less in the polyester fibers.

Neither fiber shows real elastic behavior even at low strains, and plastic deformation as well as creep occurs when the fiber, either by itself or in an elastomeric composite, is loaded. The properties required of these fibers, for example in a tire, are high strength, dimensional stability, fatigue resistance, thermal stability, and good bonding to the matrix. Tensile failure in both nylon and polyester fibers is similar, involving slow crack growth from the surface, in a radial direction. The crack opens as the material ahead of the crack deforms plastically. A point is reached at which the crack becomes unstable and rapid failure occurs. Almost identical fracture surfaces are obtained with both broken ends, one of which can be seen in Fig. 8. The composite structures for which organic

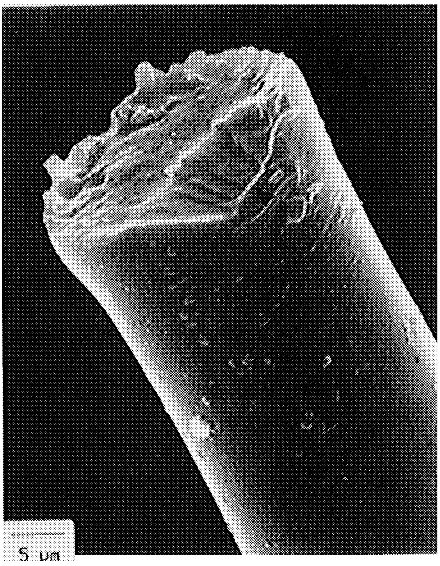

FIG. 8. The fracture morphologies of each end of a nylon or polyester fiber broken in tension or creep are very similar in appearance.

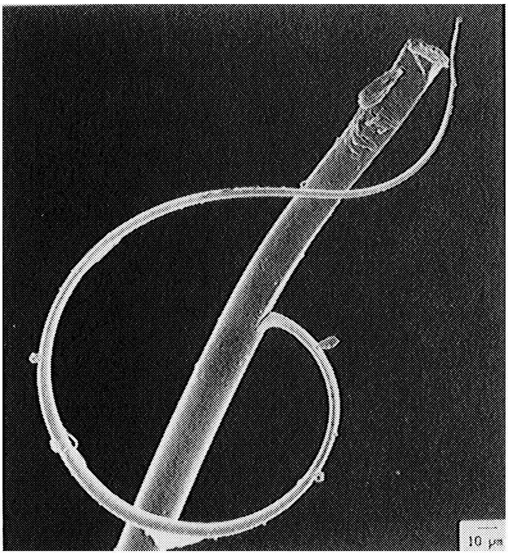

FIG. 9. Polyester and nylon fibers subjected to fatigue conditions fail with distinctive fracture morphologies.

fibers are used as reinforcements are often subjected to dynamic loading and fatigue is a major concern. Many organic fibers can fail by a distinct failure process leading to a characteristic fracture morphology involving longitudinal crack growth, which gradually reduces the load-bearing fiber cross section until failure occurs, as shown by Bunsell and Hearle (1974). This leads to a distinctive fracture morphology of which a particularly striking example is shown in Fig. 9. The loading conditions that lead to this type of failure are unusual, as a necessary criterion for fatigue failure is that the minimum cyclic load be nearly zero. Indeed, it is possible to avoid fatigue by increasing the overall loading pattern on the fiber and if this does not raise it into a regime of creep failure, the fiber will no longer break. The fatigue failure of these fibers seems to depend on the shear stresses set up between the intrafibrillar and interfibrillar amorphous zones due to differences in compliances of these regions.

Figure 10 illustrates the deformation of the molecular structures of a polyamide or polyester fiber during cyclic loading. Upon loading the tie molecules in the intrafibrillar amorphous zones are pulled slightly out of the crystalline zones so that on unloading they buckle, inducing shear stresses at the edge of the microfibril. These shear stresses lead to the breakage of molecules and a local deterioration of the fiber. Eventually the fatigue crack appears, and its development is influenced greatly by the anisotropic structure of the fiber.

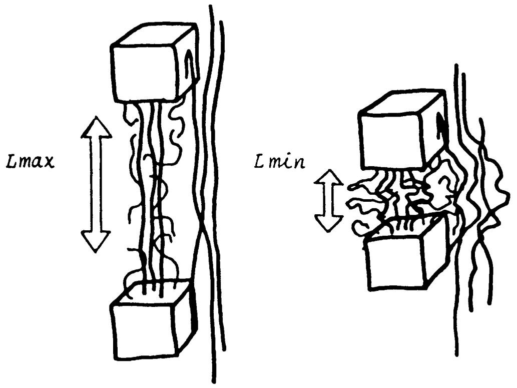

FIG. 10. During cyclic loading of nylon or polyester fibers, the tie molecules are slightly pulled out of the crystallites and buckle if the load is sufficiently reduced. This leads to the development of stresses between the microfibrils and the interfibrillar amorphous phase and the deterioration of the molecular structure.

5.3 Aramid Fibers

The aramid group of reinforced fibers, in the form of Kevlar from Du Pont de Nemours, has been available since 1972 and has been adopted for many composite structures. These fibers, which are made from a mesophase or liquid-crystal solution of the polymer in concentrated sulfuric acid, have remarkable tensile properties. Molecular poly(*p*-phenylene terephthalamide) (PPTA) alignment is produced when the mesophase solution passes through the spinneret. The shear stresses that are induced align the structure in much the same way as tree trunks are aligned during transport by river. Conventional fiber drawing is not necessary, although control of water content is important at this stage if optimum properties are to be obtained. The fibers are five times as strong as steel in air for the same weight, and this difference rises to thirty times when they are used in water, because of their low density and the buoyancy of the water. This is of great importance as these aramids find great use in cables, which are often impregnated with a resin to form a flexible composite structure, for marine and submarine applications. The highly anisotropic structure of aramid fibers results, however, in them being weak in all other directions. As a consequence failure of these fibers is nearly always highly fibrillar, as shown in Fig. 11. The structure of aramid fibers is very well ordered, and the basic molecular arrangement is shown in Fig. 12. Their highly anisotropic behavior means that the fibers are not used in primary loading structures subjected to compressive forces.

The processes that limit the compressive properties of aramid fibers confer on them, however, great tenacity, which means that they absorb much more energy than more brittle fibers when broken. This is because the fiber can deform plastically in compression. An elastic fiber can be bent to minimum radius of curvature at which the tensile stresses in the convex surface attain the breaking stress of the material. When an aramid fiber is bent slippage of the molecular structure occurs on and near the concave surface leading to the development of kink bands as can be seen from Fig. 13. The result of this plastic deformation is that the neutral stress axis is displaced from the fiber axis toward the convex surface. In this way the tensile stresses on the convex surface do not attain the failure stress of the fiber, which can be bent to a zero radius of curvature. This results in the fiber being difficult to cut as well as giving great tenacity and resistance to impact loading. At the present the aramid fibers, other than Kevlar, that are commercially available are the Arenka fiber

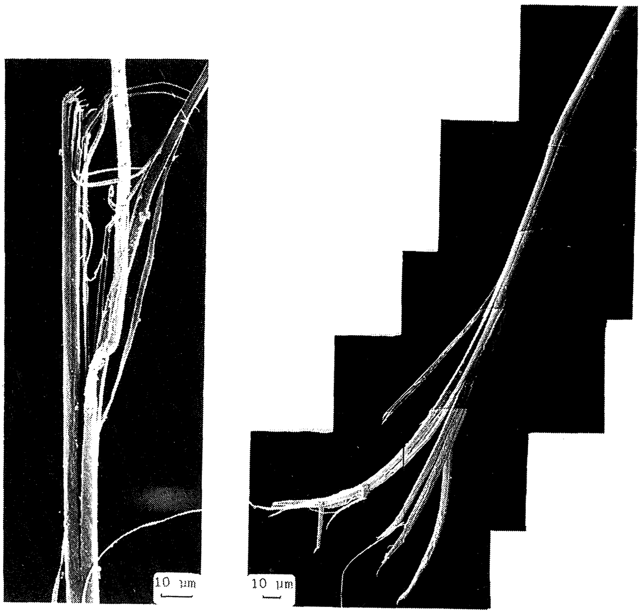

FIG. 11. Aramid "Kevlar" fibers are highly anisotropic and fibrillate on failure.

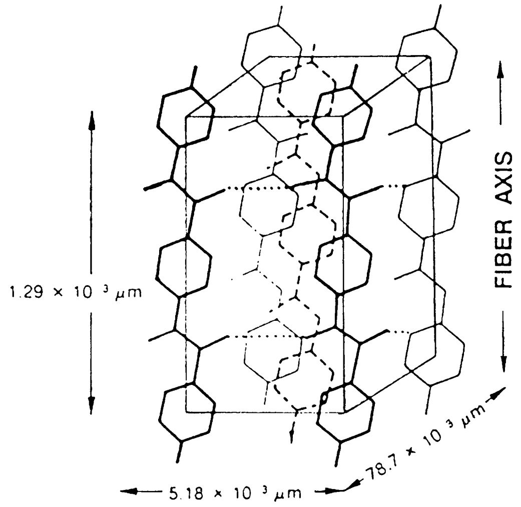

FIG. 12. The unit cell of the Kevlar aramid fiber showing the rigid PPTA molecules tied together transversely by weak secondary bonds.

from Akzo in Holland and the Technora fiber from Teijin in Japan. The Technora fiber, formally known as HM-50, from Teijin fits into the aramid family although it is not produced from a mesophase. It has an intermediate modulus.

The production of aramid fibers involves complex and expensive chemistry; however, the high-performance properties of the fibers are a result of the molecules being aligned parallel to the fiber's axis. Table 2 shows that although simpler polymers usually possess low mechanical properties they could, if all their molecules were aligned parallel to one another, give materials with the characteristics of high-performance materials.

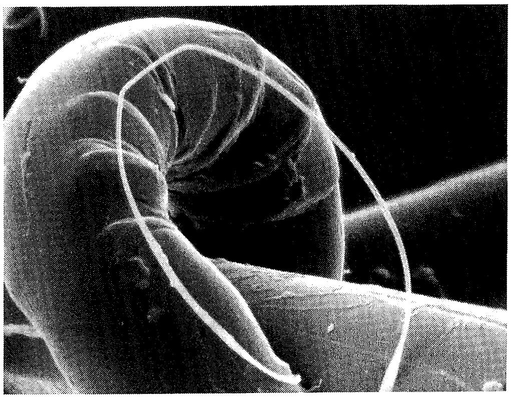

FIG. 13. Bending of aramid fibers leads to the development of kink or slip bands on the concave surface as a result of local plastic deformation.

Table 2. Ultimate strengths of various polymers.

Polymer	Density (g/cm^3)	Ultimate strength (GPa)	Strength of commercial fiber (GPa)
Polyethylene	0.96	37	0.95
Polyamide 6	1.14	32	1.00
Polyvinylalcohol	1.28	29	1.00
Poly(*p*-phenylene terephthalamide)	1.43	30	2.65
Polyethylene terephthalate	1.37	28	1.00
Poly prepylene	0.91	17	0.95
Regenerated cellulose	1.50	18	0.55

5.4 Theoretical Performance of Organic Fibers

It is possible to calculate the theoretical maximum Young's modulus for a simple organic structure for the optimum case in which all the macromolecules are aligned in the direction of the applied load. Let us consider the molecule shown in Fig. 14. The force f applied to the molecule deforms the AB and BC bonds slightly and increases the valence angle $\widehat{ABC}$. Two force parameters can be defined relating the effort and the deformation produced, such that

$$K_a = (f/\Delta a)\sin(\theta/2) \qquad (26)$$

and

$$K_\theta = fh/\theta.$$

These two parameters can be measured directly by infrared or Raman spectrometry, and for an olefin molecule we find that

$$K_a \approx 460 \text{ N m}^{-1}$$

and

$$K_\theta \approx 82 \times 20^{-20} \text{ N m}.$$

From the geometry of the molecule we see that

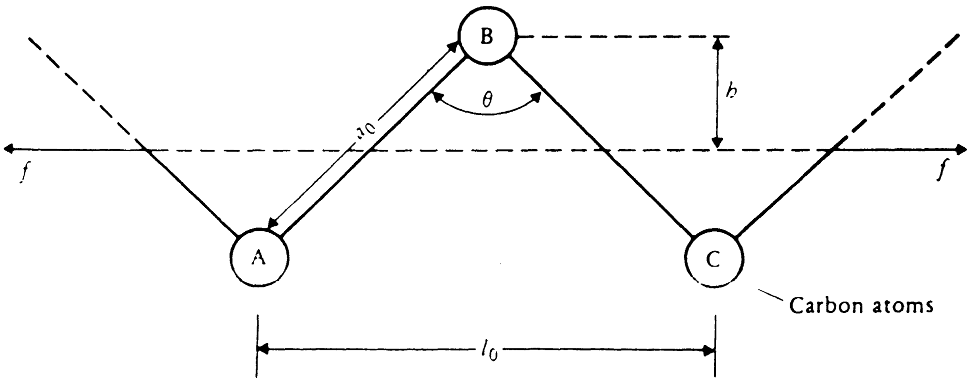

FIG. 14. Schematic representation of a part of a polymer chain being stretched by the force *f*; A, B, and C are main-chain carbon atoms.

$$l_0 = 2a\sin(\theta/2),$$

so that the deformation produced by the force f is

$$dl = 2[(a/2)\cos(\theta/2)d\theta + \sin(\theta/2)da] \tag{27}$$

and the strain

$$\epsilon = \frac{dl}{l_0} = \frac{(a/2)\cos(\theta/2)d\theta + \sin(\theta/2)da}{a\sin(\theta/2)}. \tag{28}$$

The couple fh producing a change of angle of $d\theta$ results in an angular compliance that, using Eq. (26), can be seen to be

$$\frac{d\theta}{f} = \frac{(a/2)\cos(\theta/2)}{K_\theta}. \tag{29}$$

The deformation of the AB and BC bonds is produced by the component of force, $f\cos(90° - \theta/2)$, so that the deformation compliance is

$$\frac{da}{f} = \frac{\sin(\theta/2)}{2K_a}. \tag{30}$$

The total compliance is obtained by replacing da and $d\theta$ in Eq. (28), using relations (29) and (30), and dividing by f so that

$$\frac{dl}{f} = a^2\frac{\cos^2(\theta/2)}{2K_\theta} + \frac{\sin^2(\theta/2)}{K_a}. \tag{31}$$

The Young's modulus E of the molecule can be calculated if its cross-sectional area A can be estimated, as

$$E = \frac{f/A}{dl/l_0}. \tag{32}$$

It has been estimated that A is of the order of 18×10^{-20} m^2 so that the theoretical maximum Young's modulus of a simple linear organic polymer is about 240 GPa. This is much greater than the moduli measured with most organic fibers because of their far from optimum molecular arrangements; however, ultrasonic measurements have revealed that, at a microscopic level, a simple polymer such as polyethylene possesses a modulus of 260 GPa. This observation has encouraged the optimization of the drawing processes of simple polymers with the aim of producing high-modulus fibers and films other than by producing increasingly complex and rigid molecular systems.

5.5 Ultrahigh-Performance Polyethylene Fibers

Two main manufacturing approaches exist for producing fibers by the above technique. The sol-gel approach, in which the polymer is dissolved in a solvent and then spun, is at present the only technique yielding commercially available fibers. High-modulus polyethylene fibers produced by DSM in Holland, Allied Chemicals in the U.S.A., and Toyobo in Japan have properties rivaling those of the aramid fibers with a lower specific gravity of 0.97. These fibers are limited in temperature to a maximum of 120 °C and suffer from creep. Figure 15 shows the behavior of PE fibers as a function of temperature described by Dessain *et al.* (1992). The abrupt change in slope is due to a solid-state phase change from orthorhombic to hexagonal. In the absence of an applied stress this phase change occurs around 140 °C, just below the melting point at 146 °C. However, the phase change can be induced at lower temperatures if a stress is applied. The level of stress that has to be applied to induce the phase change increases as the temperature is lowered until a critical point is reached around 5 °C below which the fiber fails before the phase change can occur.

The other manufacturing route is to draw directly from the melt as described by Bashir *et al.* (1984). This is potentially a cheaper technique but the Young's moduli of the fibers produced are not so high.

6. GLASS FIBERS

Glass has found many uses over the centuries from Roman times and earlier. A machine for producing glass fibers incorporating a type of spinneret was demonstrated at a British Association meeting in 1841. During the First World War glass fibers were produced in Germany to replace asbestos by touching a heat-softened glass rod to a spinning wheel, a technique that has been refined more recently for producing filaments of materials that are impossible to draw. In 1931 two American firms, Owen Illinois Glass Co. and Corning Glass Works, developed a method of spinning glass filaments from the melt through spinnerets. The two firms combined in 1938 to form the Owens Corning Fiberglass Corporation. Since that time extensive use of glass fibers has been made. Initially the glass fibers were destined for filters and textile uses; however, the development of heat-setting resins opened up the possibility of fiber-reinforced composites, and in the years following the Second World War the fiber took a dominant role in this type of material. Today, by far the greatest volume of composite materials is reinforced with glass fibers.

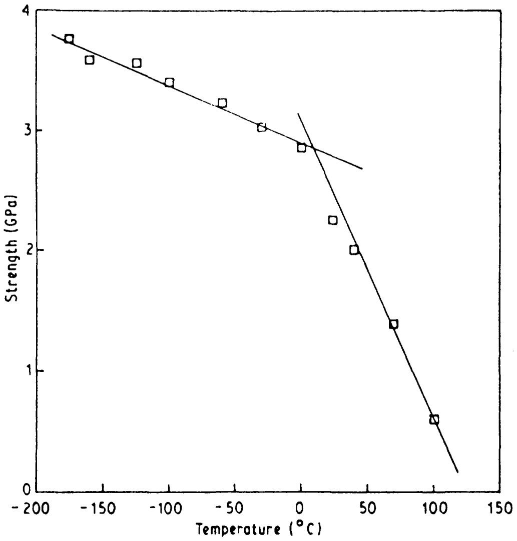

FIG. 15. The mean strength of ultrahigh-modulus polyethylene fibers as a function of temperature. The change in slope reveals a change in crystalline structure.

Several types of glass fiber exist but all are based on silica (SiO_2), which is melted and mixed with other elements to create specialty glasses (see GLASSES). Fibers of glass are produced by extruding molten glass through holes in a spinneret with diameters of 1 or 2 mm and then drawing the filaments by one of two techniques to produce fibers having diameters usually between 5 and 15 μm. The spinnerets usually contain several hundred holes so that a strand of glass fiber is produced.

The strength of glass fibers depends on the size of flaws, most usually at the surface, and as the fibers would be easily damaged by abrasion, either with other fibers or by coming into contact with machinery in the manufacturing process, they are coated with a size. The purpose of this coating is both to protect the fiber and to hold the strand together. The size may be temporary, usually a starch–oil emulsion, to aid handling of the fiber, which is then removed and replaced with a finish to help fiber-matrix adhesion in the composite. Alternatively the size may be of a type that has several additional functions, which are to act as a coupling agent and lubricant and to eliminate electrostatic charges.

Continuous glass fibers may be woven, as are textile fibers, made into a nonwoven mat in which the fibers are arranged in a random fashion, used in filament winding, or chopped into short fibers. In this last case the fibers are chopped into lengths of up to 5 cm and lightly bonded together to form a mat, or chopped into shorter lengths of a few millimeters for inclusion in molding resins.

The structure of all glass fibers is based on silica (SiO_2), which is shown in Fig. 16, mixed with other materials according to the properties desired. The structure is vitreous with no definite compounds being formed and no crystallization taking place, although at a very local level structures could be envisaged, as indicated by the dashed circles. Figure 16 shows how the silica forms a three-dimensional network, which is not regular. This open network is a result of the rapid cooling that takes place in fiber production, with the glass cooling from about 1500 °C to 200 °C in between 0.1 and 0.3 s. Despite this rapid rate of cooling there appear to be no appreciable residual stresses within the fiber and the structure is isotropic.

The most widely used glass for fiber-reinforced composites is called E glass. Glass fibers with superior mechanical properties are known as S or R glass with increased alumina content, and the compositions of a number of glasses produced in fiber form are shown in Table 3.

Glass is set to remain the most widely used reinforcement for general composites. Glass fibers are cheaper than most other relatively high-modulus fibers and, because of their flexibility, do not require very specialized machines or techniques to handle them. Their elastic modulus is, however, low when compared with other fibers that exist, and the specific gravity of glass, which for E-type glass is 2.54, is relatively high. The poor specific value of the mechanical properties of glass fibers means that they are not used for structures requiring light weight as well as high properties.

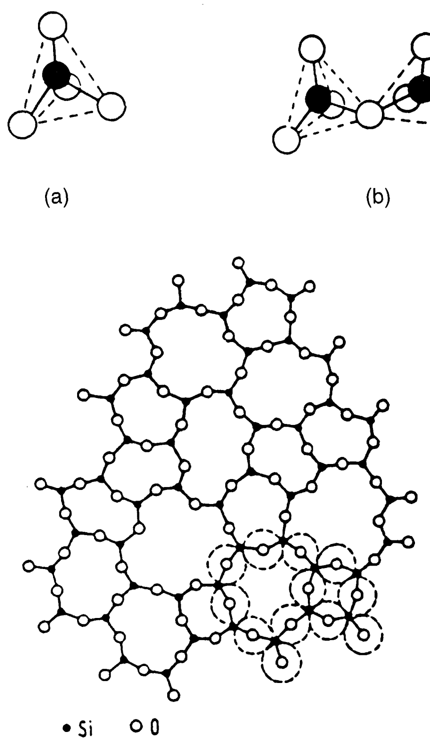

FIG. 16. The structure of glasses is a three-dimensional network based on silica but usually modified by the presence of other elements.

7. BORON AND SILICON-CARBIDE FIBERS ON A SUBSTRATE

The first fiber produced with a much increased Young's modulus compared with glass fibers was the boron fiber, made by chemical vapor deposition onto a substrate that was a tungsten wire. Boron, the fifth element in the periodic table, is the lightest element with which it was found practical to make fibers. The first boron fibers were produced in the U.S.A. and then in France and the U.S.S.R. at the beginning of the 1960s. They had remarkable properties, with a Young's modulus exceeding 400 GPa and quite extraordinary strength in compression.

Table 3. Typical compositions of various glasses used in fiber production; values given as percentages.

Glass type	E	S	R	A	C	D
SiO_2	54	65	60	72	65	74
Al_2O_3	15	25	25	0.7	4	
CaO	18		9	10	14	0.2
MgO	4	10	6	2.5	3	0.2
B_2O_3	8					
F	0.3					
Fe_2O_3	0.3					
TiO_2						0.1
NaO_2				14	8	1.2
K_2O	0.4				0.5	1.3
SO_3				0.8		
B_2O_3					5.5	23

This latter quality was enhanced by their large diameter of 140 μm. The manufacturing technique requires that each fiber be made separately and at a low speed. A mixture of boron trichloride (BCl_3) and hydrogen (H_2) heated to around 1000 °C passes over a continuous tungsten filament, having a diameter of around 13 μm, heated by electric current. Boron is deposited preferentially in the form of nodules onto the irregularities on the tungsten filament, which is produced by a wire-drawing process. The nodular structure of the fiber results in its having a surface that is not smooth, and the stress concentrations produced on the surface at the intersection of two nodules have been shown by Vega-Boggio and Vingsbo (1976) to be a factor in determining the fiber strength. Other defects associated with the core–mantle interface can also limit the strength of the fiber. Figure 17 shows a cross section of a boron fiber revealing its composite structure. The penetration of the small boron atoms into the tungsten core causes it to expand. This expansion is resisted by the boron mantle so that the core is put into compression and the mantle into tension. The cooling of the boron filament after its production induces a compressive stress into the surface.

The boron fiber shown in Fig. 17 has been additionally coated to protect it during the manufacture of a metal-matrix composite. Boron fibers proved to be a good reinforcement for light-metal alloys, particularly when the fiber was protected by a coating of SiC or B_4C. The American space shuttle is composed of a skeleton made up of more than 200 tubes of boron-fiber–reinforced aluminum, but the very high cost of the fiber has restricted wider use. Most of the applications originally foreseen for boron fibers have been taken over by carbon fibers, although they are used in some sports goods.

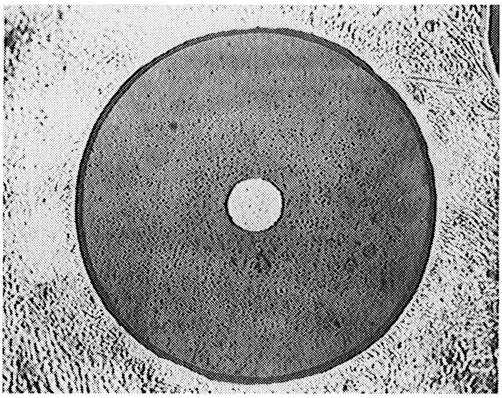

FIG. 17. Boron fibers are composed of a boron mantle deposited onto a tungsten core.

Silicon-carbide fibers, produced by a similar technique to that used to make boron fibers, remain a possible reinforcement for metal-matrix composites including titanium and intermetallic matrix composites. These fibers are based on materials that are readily available, and the substrate can either be tungsten wire or a carbon filament, which is cheaper. These SiC fibers have very good mechanical properties, with a Young's modulus of 412 GPa, and can be used at high temperature, around 1000 °C, although reactions between the fibers and the matrices they are reinforcing may limit their temperature range in a particular system. This production route for making fibers remains, however, very expensive and it is difficult to imagine a major development of this type of fiber.

8. CARBON FIBERS

Carbon fibers made by carbonizing fibers obtained from bamboo were the first filaments used in Edison's incandescent electric lamps, but they were extremely brittle and were rapidly replaced by tungsten wire. Many fibers can be converted into carbon fibers, as shown by Sitting (1980), the basic requirement being that the precursor fiber carbonizes rather than melts when heated. The route adopted in the U.S.A. in the 1950s and early 1960s to producing fibers from the next lightest element after boron was to use rayon, regenerated cellulosic fibers. This proved a slow process that did not yield spectacular results. The approach taken in Great Britain and Japan was to use polyacrylonitrile (PAN) fibers as precursors for making carbon fibers. This route, based on a modified form of acrylic textile fibers, was to prove successful and is the technique used today to make by far the greatest number of carbon fibers. The properties of carbon fibers made by this route depend on the temperature of pyrolysis used, so that it is possible to produce a family of carbon fibers with different structures. Figure 18 shows the routes adopted to make carbon fibers from both PAN and pitch precursors, as described by Toray Industries, the largest carbon-fiber manufacturer. Figures 19 and 20, taken from the results of Fitzer and Heine (1988), show how the strength of the fibers reaches a maximum at around 1500 °C whereas the elastic modulus increases continuously with increasing temperature up to around 2800 °C. The polyacrylonitrile molecule structure is linear, and conversion into carbon fiber requires its conversion into a three-dimensional structure and the elimination of all atoms except the carbon atoms. This is achieved by first heating the fiber to about 250 °C under a tensile load in air to produce a change in color from white to black and the cross linkage of the structure. Further heating in an inert atmosphere to above 1000 °C eliminates most of the elements other than carbon. Above 1500 °C the presence of elements other than carbon is negligible. The structure that results from the pyrolysis of PAN is highly anisotropic, with the basic structural units formed by the carbon-atom groups aligned parallel to the fiber axis. There is complete rotational disorder in the radial direction, and the relatively poor stacking of the carbon atoms means that a graphite structure is never achieved. For this reason it is incorrect to refer to carbon fibers made from PAN as graphite fibers. The structure contains pores, which account for the density of the fibers being less than that of densified carbon.

Table 4 shows the properties of a range of fibers used in composite materials. The values given for the carbon fibers are typical; however, there exists a considerable range of properties for these fibers.

An alternative route for making carbon fibers is from pitch obtained from either the residue of oil refining or the coking of coal. The high carbon yield of pitch, which approaches 90%, makes it an attractive and cheap source for making carbon-fiber precursors. Cost, however, is increased by the purification processes that are necessary before spinning. The pitch is converted into a mesophase or liquid-crystal solution, which is then spun, giving precursor fibers with aligned microstructures. From this point the carbon-fiber fabrication route is the same as that for the PAN-based fibers. However, the response to heating of the pitch-based precursors is not the same. There is no peak in strength at 1500 °C as this is due in the PAN-based fibers to the elimination of nitrogen, which is followed by the growth of carbon structural units resulting in a fall in strength. Pitch-based precursors heated to around

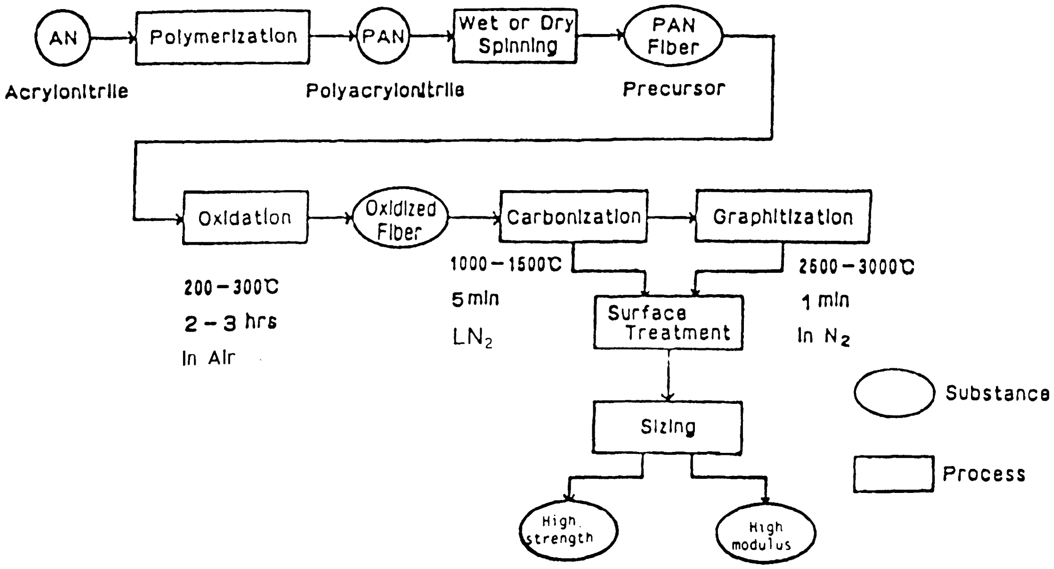

FIG. 18. Schematic flow diagram of the production of carbon fibers from PAN precursors (Toray information).

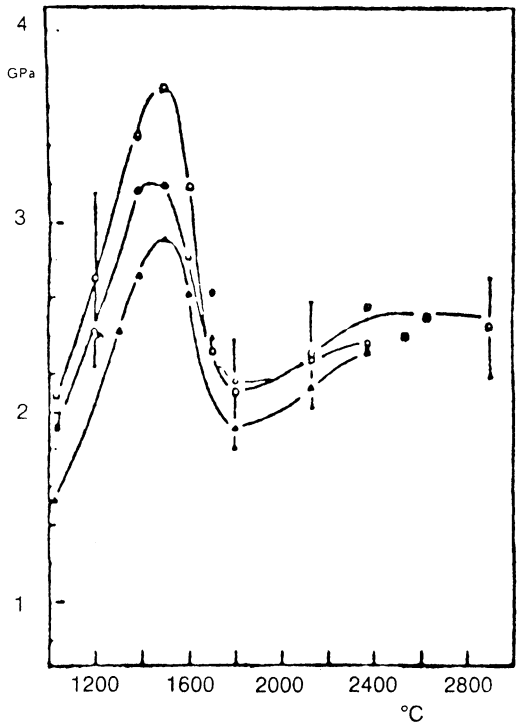

FIG. 19. The strength of carbon fibers produced from PAN depends on the heat-treatment temperature. The three curves were obtained with slightly different cross-linking temperatures but all show a maximum strength at 1500 °C (Vegga-Boggio and Vingsbo, 1977).

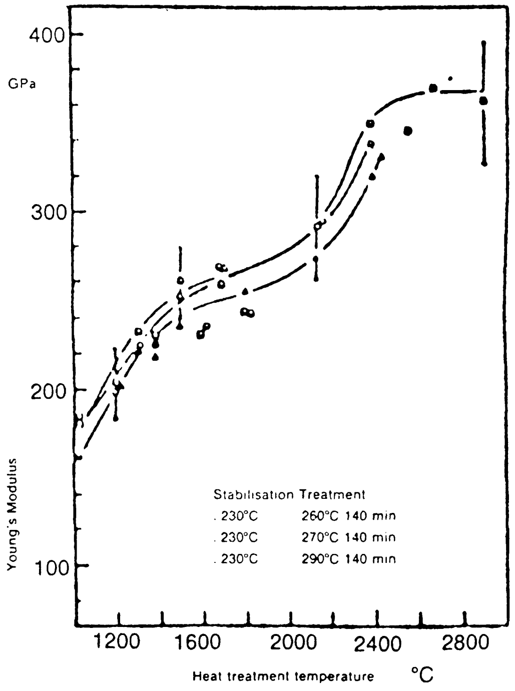

FIG. 20. The Young's modulus of carbon fibers produced from PAN increases with temperature (Fitzer and Heine, 1988).

2300 °C give fibers with Young's moduli as high as those obtained with PAN-based fibers at 2900 °C, and heating to these higher temperatures gives even greater stiffness. It is therefore more economical to produce high-modulus carbon fibers from pitch than it is from PAN. The properties are due to a less disordered, more graphitic microstructure of the pitch-based fibers, which, however, leads to lower compressive strength, as can be seen in Fig. 21.

9. ALUMINA FIBERS

Alumina is usually produced by heating aluminum hydroxide or other aluminum salts, such as ammonia alum, to above 1000 °C in air. During the heating process, the starting materials lose water and other fugitive components at 400–500 °C and thus become aluminum oxide (Al_2O_3) in chemical composition. They undergo several transformations of crystal structure, and finally become the most stable alumina called α-alumina. All other transitional aluminas are transformed into α-alumina at 1000–1100 °C. As a result of the dense packing and high regularity, α-alumina occurs as large crystal grains, generally larger than 0.1 μm; has a high degree of crystallinity, high density (3.97×10^{-3} kg m^{-3}), and a stable and inactive surface; and is strong and hard.

The first small-diameter continuous ceramic fiber, which became available in 1979 from Du Pont, consisted of almost pure α-alumina and was called Fiber FP, as described by Dhingra (1980). This fiber was made by blending the alumina in powder form with water and other compounds to form a mixture that could be spun. The filaments were then extruded through spinnerets, drawn, and fired in two steps up to 1300 °C to produce an α-alumina fiber. Properties of the fiber are shown in Table 4. It can be seen that the fiber had a low strain to fail-

Table 4. Comparison of fiber properties.

Fiber	Composition	Diameter (μm)	Specific gravity	Young's modulus E (GPa)	Strength $\bar{\sigma}$ (GPa)	Strain to failure (%)	Specific modulus E/ρ (GPa)	Specific strength σ/ρ (GPa)
Polyamide 66	PA66	20	1.2	<5	1	20	4	0.8
Polyester	PET	15	1.38	<18	0.8	15	13	0.6
Kevlar 49	PPTA	12	1.45	135	3	4.5	93	2.1
Polyethylene	PE	38	0.96	117	3	3.5	120	3.1
E glass	SiO_2	5–25	2.54	72	2	2.7	28	0.8
S glass	etc.	5–15	2.48	85	4.8	5.3	24	1.9
Boron-W	B, W	140	2.5	400	3.6	0.9	160	1.4
SiC–C	SiC, C	140	3.2	425	3.7	0.9	133	1.2
Short alumina silicate	97 Al_2O_3, 3SiO_2	3	3.3	300	2	0.7	91	0.6
Alumina	α-Al_2O_3	19	3.92	414	1.6	0.4	97	0.4
Alumina silicate	85 Al_2O_3, 15SiO_2	12–17	3.25	210	1.8	0.85	65	0.6
Saphikon	α-Al_2O_3	130	3.95	350	2.4	0.7	89	0.6
Nicalon	SiC, C, O, Si	14	2.55	220	3.0	1.4	86	1.18
Hi-Nicalon	SiC, C	14	2.74	270	2.8	1	99	1.02
Tyranno M	Si, Ti, C, O	8.5	2.37	190	3.3	1.7	80	1.39
High-strength carbon	C	5	1.82	230	4.8	2.1	126	2.64
High-modulus carbon	C	5	1.94	588	3.9	0.7	303	2.01

ure, which made it difficult to handle and limited its use. However, the FP fiber retained its strength well up to 1000 °C, showing almost no loss after 300 h when heated to this temperature in air. Above these temperatures, however, grain growth occurred making the fiber even more brittle and generally weakening it. The FP fiber is not produced commercially.

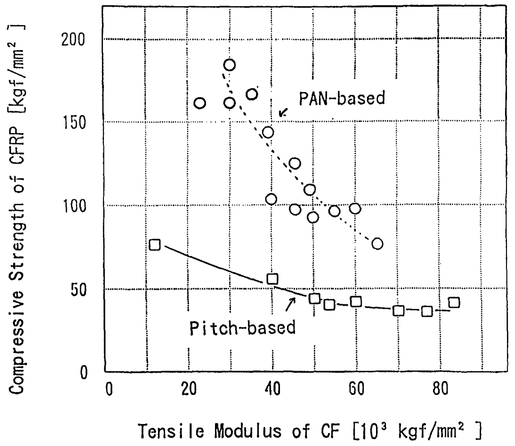

FIG. 21. The compressive strength of pitch-based carbon fibers is lower than that of PAN-based carbon fibers (Toray information).

In order to produce a fiber that could be a candidate for reinforcing high-temperature composites, Du Pont developed a fiber known as PRD-166. This fiber was described as being microcrystalline based on α-alumina but containing approximately 20 wt% of zirconia partially stabilized with yttrium oxide. The largest α-alumina grains were of the order of 0.5 μm. The zirconia showed two size populations with average sizes being 0.1 μm and 10 nm, which a study by small-angle x-ray diffraction has revealed to be in a tetragonal form.

The role of zirconia is to stabilize the structure, so improving strength and toughness. Stress concentrations associated with micropores produced a phase change in the zirconia from tetragonal to monoclinic. The associated increase in volume compressed the crack tip and impeded crack propagation. This resulted in an increase in strength although the fiber still remained brittle. The lower Young's modulus of the zirconia reduced slightly the rigidity of the PRD-166 fiber compared with the FP fiber. The PRD-166 fiber has not been made available commercially.

The properties of two other alumina fibers are shown in Table 4. Both the ICI Safimax fiber and the Sumitomo fiber contain silica, which inhibits grain growth, but in differing amounts. The greater the percentage of silica the greater the strength of the fiber but the lower the Young's modulus.

The γ-alumina fiber from Sumitomo Chemicals contains 15 wt% of silica and, as Table 4 shows, this leads to a much lower Young's modulus than that of the α-alumina fibers but a greater strain to failure and therefore easier handlability. The alumina is crystallized in the γ phase with a very fine grain size, a low density, and an active surface. The role of the silica is described as being to stabilize the alumina in the γ phase. These fibers typically show a Weibull modulus of less than 7. The fibers conserve their strength and stiffness at least up to 800 °C with little noticeable loss in tensile properties at 1000 °C. Above this temperature the tensile properties fall and at temperatures above 1150 °C the fiber deforms plastically. The fiber is observed to creep above 1000 °C. Up to 1100 °C the behavior is that of a material creeping by diffusion at grain boundaries. Above this temperature, softening of the silica seems to account for the rapid acceleration of the creep process.

The Saffil fiber is a staple fiber with a diameter of 3 μm produced by ICI since the early 1970s and consists of about 97% alumina and 3% silica. Such fibers are produced by solution spinning, which leads to little scatter in diameter. The solution consists of an aluminum salt, such as aluminum chlorohydroxide $AlC_x(OH)_y$, which contains hydrated species that eliminate water to link octahedral units by Al–O–Al bridges. As the process continues, a viscous solution is obtained, which can be spun into fibers. This is helped by the addition of organic polymers to control rheology and to lower the spinning temperature. Progressive increases in temperature burn off the organic material and convert the aluminum salt to alumina. The Saffil fiber was originally produced as a refractory felt.

The widest use of the Saffil-type fiber in composites is in the form of a mat that can be shaped into the form desired and then infiltrated with molten metal, usually aluminum alloy. It is probably the most successful fiber reinforcement for metal-matrix composites.

The Nextel series of fibers produced by 3-M are all based on mullite with varying percentages of boron oxide. The fibers are oval in cross section but a study by image analysis gives an average diameter of 11 μm. The Nextel 440 fiber has been found by x-ray diffraction to consist of two crystalline phases, one of mullite and the other of γ-alumina. The more crystalline Nextel 480 contains only mullite as a crystalline phase and much reduced boron oxide. The grain size in this latter fiber is 50 nm and the structure is found not to evolve up to 1200 °C except for the completion of the mullite crystallization. The manufacturer reports that the Nextel fibers conserve at least 75% of their properties up to 1000 °C and that they creep over 1000 °C.

The ceramic fibers based on alumina seem to be possible candidates for reinforcing light alloys with the observation, however, that their specific gravities are quite high. Such reinforcements permit metal structures to be made with enhanced properties, notably creep resistance at high temperatures. The phase change to α-alumina above about 1150 °C and subsequent grain growth would seem to exclude these fibers from most ceramic-matrix composite materials. A solution to this problem would be the use of continuous single-crystal fibers, which would contain no grain boundaries. Such fibers are produced in the form of continuous α-alumina fibers grown from the molten ceramic under the commercial name Saphikon. These fibers have diameters of around 100 μm.

10. CERAMIC FIBERS FROM ORGANIC PRECURSORS

The work of Yajima *et al.* (1987) in Japan was first published in Japan in the mid-1970s and gave rise to the first and so far the most successful fine ceramic fiber based on silicon carbide. This fiber is produced by Nippon Carbon under the name of Nicalon. The manufacture of this fiber involves the production of a polycarbosilane precursor fiber, which consists of cycles of six atoms arranged in a similar manner to the cubic structure of β-SiC.

The molecular weight of this polycarbosilane is low, around 1500, which makes drawing of the fiber extremely difficult. In addition, methyl groups (CH_3) in the polymer are not included in the Si–C–Si chain so that during pyrolysis the hydrogen is driven off, leaving a residue of free carbon. The precursor fibers are subjected to heating in air at about 300 °C to produce cross linking of the structure. This oxidation makes the fiber infusible but has the drawback of introducing oxygen into the structure, which remains after pyrolysis. The ceramic fiber is obtained by a slow increase in temperature in an inert atmosphere up to 1300 °C. The fiber that is obtained contains a majority of SiC but also significant amounts of free carbon and excess silicon combined with oxygen. The distribution of these elements is uniform across the fiber diameter, as has been shown by electron-microprobe analysis.

The strength and Young's moduli of Nicalon fibers tested in air or an inert atmosphere show little change up to 1000 °C. Above this temperature, both these properties fall with the greatest change being in the strength of the fiber.

Nicalon fibers as supplied now by the manufacturer are microcrystalline with typical grain sizes of the β-SiC crystals of around 2 nm, as is revealed by x-ray diffraction and transmission-electron microscopy. Heating above 1100 °C produces grain growth, which stabilizes with an average grain size of approximately 3 nm, at which point grain growth appears to stop. Under no load the fiber shrinks during such a heat treatment but the shrinkage stops with the stabilization of the grain size.

Simon and Bunsell (1984) have shown that when a load is applied to the fibers, it is found that a creep threshold stress exists above which creep occurs. The primary creep phase is clearly greatly influenced by the structural changes occurring. The ceramic grade 200 series Nicalon fibers, which evolved from earlier fibers, has been heat-treated in order to stabilize grain size at around 3 nm and improve the properties at high temperature.

The β-SiC crystals are embedded in an amorphous matrix consisting of Si, O, and C, which also contains aggregates of free carbon with sizes of around 2 nm. A typical composition by weight of a Nicalon fiber is 65% SiC, 15% free carbon, and an amorphous phase consisting of Si, O, and some carbon. The free carbon and the oxygen in the structure have been shown to play a determining role in the use of Nicalon fibers at high temperatures.

The existence of the creep stress thresholds is attributed to the presence of the particles of carbon in the structure, which impede its movement and rearrangement. In this case the carbon particles confer a useful property to the fiber, but their presence, together with the oxygen in the structure, finally limits the use of the fiber at high temperature. Degradation of Nicalon fibers is slower in air than in an inert atmosphere as a result of a silica coating being created at the fiber surface. This oxide coating is considered to slow the expulsion of the oxides of carbon produced in the interior of the fiber.

A later generation of these fibers known as Hi-Nicalon has been developed using the same precursor but using electron-beam radiation rather than oxidation to produce the cross linking of the precursor. The resulting fiber has a higher Young's modulus (~240 GPa) and higher strength.

Although the Nicalon fibers are the most widely available fine ceramic fiber based on silicon and carbon, the technology of this process of fiber production is being actively explored in many laboratories by a number of routes. The Tyranno fiber produced by Ube Industries is made by a process analogous to that used for the Nicalon fiber, but with significant differences. The manufacture of the Tyranno fiber begins with the production of polydimethylsilane, $(Si(CH_3)_2)_n$, as for the Nicalon fiber, which is then mixed with titanium alkoxide in a ratio of approximately 1:10, heated to 340 °C in nitrogen, and polymerized for 10 h (Gupta, 1988). The reaction product is concentrated at 310 °C for 2 h in a stream of nitrogen to produce polytitanocarbosilane. The molecular weight of this organo-metallic polymer is low, around 1500. The structure of the polytitanocarbosilane precursor is complex as a result of condensation of Si–H bonds and cross linking of titanium compounds occurring simultaneously. The precursor fiber is converted into a ceramic fiber by heating in air at about 180 °C and is then heat-treated in nitrogen at around 1000 °C. The resulting fiber contains a small amount of titanium, around 2% by weight, which is said to inhibit crystallization, and similar percentages of Si, O, C, and SiC as the Nicalon fiber. The Tyranno fiber has a diameter of around 11 μm.

The Tyranno fiber retains its strength up to 1200 °C, at least for short exposure times, although above this temperature the structure of the fiber changes to consist of β-SiC and crystallized silica.

A form of Tyranno fiber known as Tyranno LOX-E has been developed using the same electron-beam cross-linking process as mentioned for the Nicalon fiber. The Tyranno LOX-E fiber contains 4.7% oxygen.

A group of ceramic fibers that are being actively studied is based on a silicon carbide–silicon nitride system and is produced from polycarbosilazane precursors. Various routes for the manufacture of these fibers are possible, as has been shown by Lipowitz *et al.* (1987), and as yet these fibers are at the laboratory or early pilot-plant stage. The fibers can be produced by the pyrolysis of polysilazanes prepared from tri(*N*-methylamino) methylsilane polymerized by heating to 520 °C for several hours and then heated to between 1200 and 1500 °C for 2 h. During pyrolysis a weight loss of approximately 35% is reported to occur, and finally the Si_xC_2 ceramic fibers that are produced are described as being a shiny black in appearance.

This type of fiber was initially studied for the dielectric properties that it is possible to obtain with this class of material. The electrical resistivity of silicon carbide–silicon nitride is about 10^{12} times that obtained for graphite.

Other approaches to producing fibers based on Si–C–N have been studied and reported in the literature by researchers from Dow Corning. These fibers were prepared by melt spinning of amorphous thermoplastic polymers followed by a cross-linking step to gel the fibers and to make them infusible. Fibers consisting of Si–N–C were derived from hydridopolysilazane (HPZ) polymers prepared by reaction of trichlorosilane and hexamethyldisilazane. The fibers were pyrolyzed in nitrogen or argon at temperatures between 1100 and 1400 °C. The ceramic fibers produced were amorphous.

11. WHISKERS

Single ceramic crystals in the form of filaments are known as whiskers, and it has been known for many years that their mechanical properties can approach those that are theoretically possible. This is because of the lack of defects, which weaken other forms of matter. The requirement that the whiskers be free of defects means that they are extremely small, with diameters of around 1 μm and less and lengths usually not more than a few millimeters. Bigger diameters result in dramatic falls in mechanical properties.

Despite their small dimensions, whiskers have potential as reinforcements because of their aspect ratios of length to diameter, which are great enough to produce a reinforcing effect in plastic matrices. Difficulties of orienting the whiskers in a matrix and especially health hazards due to their size, which mean they can become lodged in the lungs if breathed in and so cause serious illness, have limited their use.

There are a great many possible types of whisker reinforcements. The most widely commercialized whisker type is based on silicon carbide.

12. CONCLUSION

Fibers exist in many forms and have evolved at a remarkable rate during the second half of the twentieth century. Their remarkable properties depend on their microstructure, which is often complex and anisotropic. In terms of strength fibers represent some of the strongest forms of ordinary matter, while their elastic moduli can be nearly as great as that of diamond. Such fibers can be used to reinforce all classes of solids whether they are rubber, plastics, light metals, glasses, or ceramics. The composite materials formed in this way are now and will increasingly become the engineering materials of the modern world.

GLOSSARY

Alumina: Aluminum oxide.

Aramid: A generic term for wholly aromatic polyamides.

Liquid Crystal: A solution in which the molecular structure is locally oriented.

Pitch: Residue of the oil-refining process or the coking process of oil.

Polyacrylonitrile (PAN): A polymer that can be spun to form fibers, which in turn can be converted to carbon fibers by pyrolysis.

Regenerated Fibers: Those artificial fibers that are made from naturally occurring cellulose.

Synthetic Fibers: Those of which their macromolecular structure is manmade.

Works Cited

Bashir, Z., Odell, J. A., Keller, A. (1984), *J. Mater. Sci.* **19,** 3713–3725.

Black, W. B., Preston, J. (1973), *High Modulus Wholly Aromatic Fibers,* New York: Marcel Dekker.

Bunsell, A. R. (Ed.) (1988), *Fibre Reinforcements for Composite Materials,* Amsterdam: Elsevier.

Bunsell, A. R., Hearle, J. W. S. (1974), *J. Appl. Polym. Sci.* **18,** 267–291.

Cox, H. L. (1952), *Brit. J. Appl. Phys.* **3,** 72–79.

Dessain, B., Moulaert, O., Kennings, R., Bunsell, A. R. (1992), *J. Mater. Sci.* **27,** 4515–4522.

Dhingra, A. K. (1980), *Philos. Trans. R. Soc. London, Ser. A* **294,** 411–417.

Fitzer, E., Heine, M. (1988). "Carbon Fiber Manufacture and Surface Treatment," in: A. R. Bunsell (Ed.), *Fibre Reinforcements for Composite Materials,* Amsterdam: Elsevier, Chap. 3, pp. 73–148.

Griffith, A. A. (1920), *Philos. Trans. R. Soc. London, Ser. A* **221,** 163–198.

Gupta, P. K. (1988), "Glass Fibers for Composite Materials," in: A. R. Bunsell (Ed.), *Fibre Reinforcements for Composite Materials,* Amsterdam: Elsevier.

Inglis, C. E. (1913), *Trans. Inst. Naval Architecture* **55,** 219–230.

Lipowitz, J., Freeman, H. A., Chem, R. T., Prack, E. R. (1987), *Adv. Ceram. Mater.* **2,** 121–128.

Moncrieff, R. W. (1979), *Man-made Fibres,* 6th ed., Guildford, U.K.: Butterworth.

Morton, W. E., Hearle, J. W. S. (1975), *Physics of Textile Fibres,* 2nd ed., London: Heinemann.

Simon, G., Bunsell, A. R. (1984), *J. Mater. Sci.* **19,** 3658–3670.

Sitting, M. (1980), *Carbon and Graphite Fibers: Manufacture and Applications,* Park Ridge, NJ: Noyes Dates Co.

Vega-Boggio, J., Vingsbo, O. (1976), *J. Mater. Sci.* **11,** 273–282.

Vegga-Boggio, J., Vingsbo, O. (1977), *J. Mat. Sci.* **12,** 2519–2524.

Yajima, S., Hasegawa, X., Hayashi, J., Iiuma, M. (1987), *J. Mater. Sci.* **13,** 2569–2576.

Further Reading

Bunsell, A. R. (1988), *Fibre Reinforcements for Composite Materials,* Amsterdam: Elsevier.

Moncrieff, W. E. (1979), *Man Made Fibres,* Guildford, U.K.: Butterworth.

Kelly, A., Macmillan, M. H. (1986), *Strong Solids,* Oxford, U.K.: Clarendon.

Lee, S. M. (Ed.) (1991), *International Encyclopedia of Composites,* New York: VCH.

Kelly, A. (Ed.) (1994), *Concise Encyclopedia of Composite Materials,* Oxford, U.K.: Pergamon-Elsevier.

Brook, R. J. (Ed.) (1991), *Concise Encyclopedia of Advanced Ceramic Materials,* Oxford, U.K.: Pergamon.

RELATIVITY

See GRAVITATION AND GENERAL RELATIVITY

2
Special Relativity

Tatsu Takeuchi

2.1 Introduction 49
2.2 Special Relativity Basics 49
2.2.1 The Lorentz Transformation 49
2.2.1.1 Postulates 49
2.2.1.2 Derivation of the Lorentz Transformation 50
2.2.2 The Relativity of Simultaneity, Causality, and the Light Cone 53
2.2.3 Time Dilation and Lorentz Contraction 54
2.2.3.1 Time Dilation 54
2.2.3.2 Lorentz Contraction 55
2.2.3.3 Experimental Evidence 55
2.2.3.4 Twin Paradox 56
2.2.4 Doppler Effect 58
2.2.4.1 Doppler Effect in Light 58
2.2.4.2 Application in Astronomy 59
2.2.4.3 Doppler Radar 59
2.2.5 Addition of Velocities 60
2.2.5.1 Addition of Collinear Velocities 60
2.2.5.2 Experimental Application 61
2.2.5.3 Addition of Noncollinear Velocities and Wigner Rotation 61
2.3 Relativistic Dynamics 62
2.3.1 The Covariance of $\mathbf{F} = m\mathbf{a}$ under Galilei Transformations and Rotations 62
2.3.2 Covariance under Lorentz Transformations 62
2.3.2.1 4-Vectors 62
2.3.2.2 The Relativistic Equation of Motion 63
2.3.2.3 A Simple Example 65
2.3.2.4 What is the Inertia of the Particle? 65
2.3.2.5 Action, Lagrangian, and Hamiltonian of a free particle 66
2.3.3 Parallel Transport of 4-Vectors 67
2.3.3.1 Fermi–Walker Transport 67
2.3.3.2 Thomas Precession 69

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

2.3.4 Electrodynamics 70
2.3.4.1 Covariant Formulation of Maxwell's Equations 70
2.3.4.2 Lorentz Transformation Properties of the Electromagnetic Field 71
2.3.4.3 The Action and Coupling to a Charged Particle 71
2.3.4.4 Lagrangian and the Hamiltonian in the Nonrelativistic Limit 72
2.3.4.5 Spin-Orbit Coupling 72
2.4 *Photons* 73
2.4.1 Quantum Mechanics 73
2.4.1.1 Angular Frequency and Wave Number 73
2.4.1.2 Velocities 74
2.4.1.3 Massless Particles 74
2.4.2 Doppler Effect Revisited 75
2.4.2.1 Doppler Effect on the Energy of a Photon 75
2.4.2.2 Relativistic Aberration and Relativistic Beaming 75
2.4.3 Conservation of 4-Momentum and $E = mc^2$ 76
2.4.4 Gravitational Red-Shift 77
2.4.5 Global Positioning System 78
2.5 Further Reading 79
Appendix A 79
2.A.1 Representations of the Lorentz Group 79
2.A.1.1 Conventions 79
2.A.1.2 The Lorentz Group 80
2.A.1.3 Infinitesimal Lorentz Transformations 80
2.A.1.4 The Generators of the Lorentz Group 81
2.A.1.5 Examples 82
2.A.1.5.1 Scalars 82
2.A.1.5.2 Spinors 83
2.A.1.5.3 Vectors 85
2.A.1.5.4 Antisymmetric Tensors 87
2.A.2 The Dirac Equation 89
2.A.2.1 Weyl Spinors with Majorana Masses 89
2.A.2.2 Dirac Masses, Dirac Spinors, and the Dirac Equation 90
2.A.2.3 Chiral, Dirac, and Majorana Representations 91
2.A.2.4 Coupling to the Electromagnetic Field and the Nonrelativistic Limit 92
References 93

2.1
Introduction

The special theory of relativity, or special relativity (SR) for short, is one of the foundational cornerstones of modern physics. It is particularly important in particle physics, not only because high-energy experiments routinely accelerate particles to close-to-light speeds, where their behavior can only be understood within the SR framework, but also because SR dictates the types of interactions that particles can have, leading to numerous observable consequences even at low energies. The spin-statistics theorem, which is necessary for the understanding of the periodic table of atoms, is a consequence of SR. The change in nuclear mass via fission and fusion can only be explained via the famous equation $E = mc^2$. In this chapter, we give a brief overview of the basics of SR and some of its applications.

It may be noted that SR is called the *special* theory of relativity owing to the fact that it only deals with physics in a nondynamical flat spacetime and only requires *special covariance*, that is, the laws of physics must be invariant under the *special* set of coordinate transformations, which takes the observer from one inertial frame to another. In the *general* theory of relativity, spacetime itself is a dynamical entity, the curvature of which corresponds to the gravitational field, and *general covariance* is required, that is, the laws of physics must be invariant under *general* coordinate transformations.

2.2
Special Relativity Basics

In this section, we derive the coordinate transformation, which relates observations in one inertial frame to that in another, and discuss some of its consequences.

2.2.1
The Lorentz Transformation

2.2.1.1 Postulates

The starting point of SR is the following two postulates:

- All inertial frames are equivalent, that is, no preferred inertial frame exists. Consequently, all laws of physics must be expressible in the same form in all inertial frames.
- The speed of light in vacuum, c, is the same in all inertial frames.

As with any statement regarding the physical world, the constancy of the speed of light is something that must be checked

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

experimentally. And starting from the famous Michelson–Morley experiment of 1886 (Michelson and Morley, 1887), it continues to be tested to the present day with ever-increasing experimental precision. See Chang and Zhang (1998) and Jackson (1998) for a review of developments up to 1998. Some of the more recent developments can be found in Kostelecky (1999, 2002, 2005, 2008). For the purpose of this chapter, however, we simply accept these postulates and work out the consequences.

2.2.1.2 Derivation of the Lorentz Transformation

The above postulates uniquely determine the transformation that relates the spacetime coordinates (ct, x, y, z) of an event as observed in one inertial frame to the spacetime coordinates (ct', x', y', z') of the same event as observed in another inertial frame, which is moving at constant velocity relative to the first. Here, we have multiplied the time coordinate with the universal constant c to give it the same dimension as the space coordinates. In other words, instead of measuring time intervals in *seconds*, we measure them in *meters*: the distance that a light pulse will travel in vacuum during that time interval.[1] In these units, the speed of light is dimensionless and equal to 1, so it is customary to set $c = 1$ and write ct as t with the understanding that t is measured in meters. We will, however, keep c in our expressions to facilitate the taking of the nonrelativistic limit $c \to \infty$.

To simplify our derivation of the transformation in question, we start with the $1+1$ dimensional case, and then extend the result to $3+1$ dimensions.

$1+1$ Dimensional Case Let (ct, x) and (ct', x') be the spacetime coordinates of an event as observed from two different inertial frames, which we will call frame A and frame B. We assume that the two coordinate systems share a common origin, that is, $(ct, x) = (0, 0)$ and $(ct', x') = (0, 0)$ are the same spacetime point. This can always be achieved by a simple repositioning of the spacetime origin in either frame.

Since both frames are inertial, a straight worldline in one frame must also be a straight worldline in the other. (The term *worldline* refers to the path of a particle in spacetime. The worldline of a free particle in an inertial frame will be straight.) This requires the relation between (ct, x) and (ct', x') to be linear, since a linear relation between ct and x must imply a linear relation between ct' and x', and vice versa. Therefore, (ct', x') and (ct, x) must be related through

$$\begin{bmatrix} ct' \\ x' \end{bmatrix} = B(\beta) \begin{bmatrix} ct \\ x \end{bmatrix} \tag{1}$$

where $B(\beta)$ is a 2×2 matrix, which can only depend on the velocity $\beta = v/c$ of frame B as observed in frame A. It is clear from simple physical reasoning that

$$B(0) = \mathbf{1}, \qquad B(-\beta) = B^{-1}(\beta) \tag{2}$$

Furthermore, using the parity transformation

$$P = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{3}$$

---

[1] In actuality, the current definition of a *meter* is the distance that light travels in 1/299792458 seconds. In other words, spacial distances are determined by the time it takes for light to travel that distance, and the *meter* is defined so that the speed of light in vacuum is *exactly* 299792458 ms^{-1}.

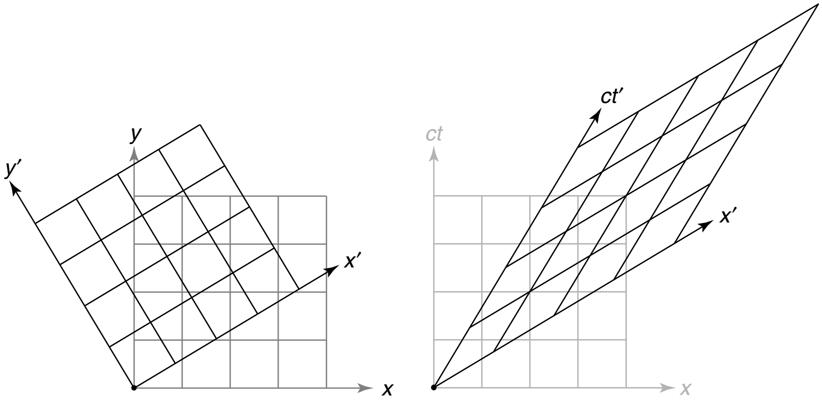

(a) (b)

Fig. 2.1 Rotations versus Lorentz Transformations: rotations (a) keep $\Delta r^2 = \Delta x^2 + \Delta y^2$ invariant while Lorentz transformations (b) keep $\Delta s^2 = c^2\Delta t^2 - \Delta x^2$ invariant.

we must have

$$P\,B(\beta)P^{-1} = B(-\beta) \tag{4}$$

since reversing the direction of the x-axis is the same thing as reversing the direction of the velocity, which together with Eq. (2) implies

$$\det B(\beta) = 1 \tag{5}$$

For Galilei transformations, the matrix $B(\beta)$ is given by

$$B_{\text{Galilei}}(\beta) = \begin{bmatrix} 1 & 0 \\ -\beta & 1 \end{bmatrix} \tag{6}$$

which clearly satisfies Eqs. (2) and (5).

Let a light pulse be emitted from point 1 with spacetime coordinates (ct_1, x_1) in frame A, and (ct'_1, x'_1) in frame B. Assume that it reaches point 2 with spacetime coordinates $(ct_2, x_2) = (c(t_1 + \Delta t), x_1 + \Delta x)$ in frame A, and $(ct'_2, x'_2) = (c(t'_1 + \Delta t'), x'_1 + \Delta x')$ in frame B. Since the speed of light is c in both frames, we must have $|\Delta x| = c|\Delta t|$ imply $|\Delta x'| = c|\Delta t'|$, and vice versa, that is,

$$\begin{aligned} &(c\Delta t)^2 - (\Delta x)^2 = 0 \\ &\quad \longleftrightarrow (c\Delta t')^2 - (\Delta x')^2 = 0 \end{aligned} \tag{7}$$

The linearity between (ct, x) and (ct', x') and the symmetry between the two frames require

$$\Delta s^2 \equiv (c\Delta t)^2 - (\Delta x)^2 \tag{8}$$

to be invariant for arbitrary spacetime intervals, not just that between the end points of a light pulse worldline for which $\Delta s^2 = 0$. This condition requires the matrix $B(\beta)$ to satisfy

$$B(\beta)^{\mathrm{T}} g\, B(\beta) = g, \quad g \equiv \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{9}$$

the solution to which can be written generically as

$$B(\beta) = \begin{bmatrix} \cosh\eta & -\sinh\eta \\ -\sinh\eta & \cosh\eta \end{bmatrix} \tag{10}$$

Note the similarity with two-dimensional rotation matrices that keep $\Delta r^2 \equiv (\Delta x)^2 + (\Delta y)^2$ invariant (see Fig. 2.1). The parameter η is known as the *rapidity* and is related to the velocity β via

$$\beta = \tanh \eta \tag{11}$$

This can be seen from the fact that the worldline of the spacial origin of frame B, $x' = 0$, corresponds to the line $-ct \sinh \eta + x \cosh \eta = 0$ in frame A. The transformation of Eq. (10) that relates the coordinates in the two frames is known as the *Lorentz transformation*. Since

$$\begin{aligned} \cosh \eta &= \frac{1}{\sqrt{1-\beta^2}} \equiv \gamma, \\ \sinh \eta &= \frac{\beta}{\sqrt{1-\beta^2}} = \gamma\beta \end{aligned} \tag{12}$$

the matrix $B(\beta)$ can also be written as

$$B(\beta) = \gamma \begin{bmatrix} 1 & -\beta \\ -\beta & 1 \end{bmatrix} \tag{13}$$

and the explicit relation between (ct', x') and (ct, x) is

$$ct' = \frac{ct - \beta x}{\sqrt{1-\beta^2}}, \qquad x' = \frac{-\beta ct + x}{\sqrt{1-\beta^2}} \tag{14}$$

(see Fig. 2.2) In the limit $c \to \infty$, this reduces to

$$t' = t, \qquad x' = -vt + x \tag{15}$$

which is the usual Galilei transformation. In the limit $\beta = v/c \to 1$, the transformation becomes singular, which reflects the fact that for the speed of light to be the same in all frames, one naturally cannot have frames moving at, or faster than, c.

Extension to 3 + 1 Dimensions Extension of this result to $3+1$ dimensions is straightforward. Following a similar argument as the $1+1$ dimensional case, the invariance of the speed of light requires

$$\Delta s^2 \equiv (c\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2 \tag{16}$$

to be invariant under transformations from one inertial frame to another. That is, the 4×4 transformation matrix Λ must satisfy the condition

$$\Lambda^{\mathrm{T}} g \Lambda = g, \quad g = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \tag{17}$$

Note that in the $3+1$ dimensional case, the usual three-dimensional spacial rotations keep Δs^2 invariant, as well as transformations such as

$$\begin{bmatrix} ct' \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cosh \eta & -\sinh \eta & 0 & 0 \\ -\sinh \eta & \cosh \eta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} ct \\ x \\ y \\ z \end{bmatrix} \tag{18}$$

which corresponds to a transformation to a frame that is moving in the positive x direction with rapidity η without any rotation of the spacial coordinate axes. Of the transformations that keep Δs^2 invariant, those that do not involve any rotations are called *Lorentz boosts*. A generic Lorentz transformation in $3+1$ dimensions is a combination of Lorentz boosts and rotations. The matrix that boosts the frame to that moving with rapidity η in the direction of unit vector $\mathbf{n}$, that is, with velocity $\boldsymbol{\beta} = \mathbf{n} \tanh \eta$, is given by

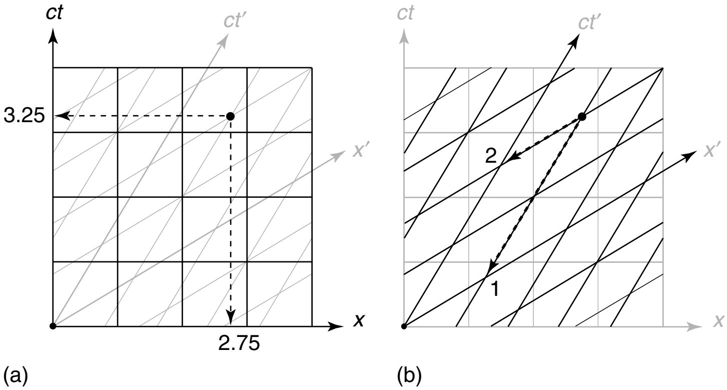

Fig. 2.2 The Lorentz transformation: the spacetime location of an event as seen in the (ct, x) frame (a) and that seen in the (ct', x') frame (b). In this example, the primed frame is moving in the positive x direction at a velocity of $0.6c$ relative to the unprimed frame.

$$B(\boldsymbol{\beta}) = \begin{bmatrix} \cosh\eta & -\mathbf{n}^{\mathrm{T}}\sinh\eta \\ -\mathbf{n}\sinh\eta & (1-\mathbf{n}\mathbf{n}^{\mathrm{T}}) + \mathbf{n}\mathbf{n}^{\mathrm{T}}\cosh\eta \end{bmatrix} \tag{19}$$

For $\mathbf{n} = (1, 0, 0)$, we recover Eq. (18).

2.2.2
The Relativity of Simultaneity, Causality, and the Light Cone

The most important aspect of the Lorentz transformation is that the familiar rule $t' = t$ of the Galilei transformation no longer holds. This implies that the concept of *simultaneity* is frame dependent. Consider two events, 1 and 2, which happen at (ct_1, x_1) and (ct_2, x_2), respectively, in frame A. If $t_1 = t_2$ they happen *at the same time* in frame A. However, if the two events are spacially separated, that is, $x_1 \neq x_2$, then Eq. (14) tells us that $t'_1 \neq t'_2$ in frame B, that is, they do NOT happen *at the same time* in frame B. Their chronological order will be $t'_2 > t'_1$ or $t'_2 < t'_1$ depending on whether $\beta(x_2 - x_1) < 0$ or $\beta(x_2 - x_1) > 0$.

If the chronological ordering of events is frame dependent, it should immediately cause concern as to whether it is compatible with the concept of *causality*, namely, the idea that the *cause* must happen before the *effect*. If event 1 is the cause and event 2 is its effect, then causality requires event 1 to happen *before* event 2, independent of the frame from which the observation is being made. In general, the spacetime separation of a pair of events can be classified into the following three categories:

$\Delta s^2 > 0$	timelike separation
$\Delta s^2 = 0$	null separation
$\Delta s^2 < 0$	spacelike separation

Since Δs^2 is invariant under Lorentz transformations, this classification is frame independent.[2] It is straightforward to prove that the chronological order of timelike or null-separated events is frame independent, while that for spacelike separated

---

[2] Some authors define Δs^2 as $\Delta s^2 = -\Delta t^2 + \Delta x^2 + \Delta y^2 + \Delta z^2$, in which case the definitions of timelike and spacelike separations are reversed.

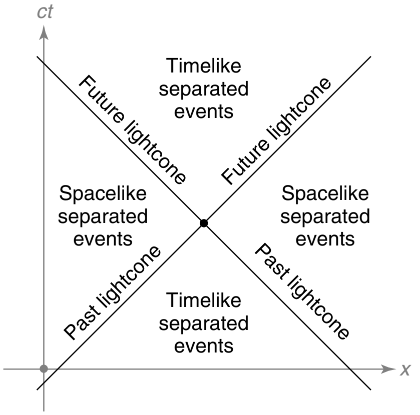
ct
x
Timelike separated events
Spacelike separated events
Spacelike separated events
Timelike separated events
Future lightcone
Future lightcone
Past lightcone
Past lightcone

Fig. 2.3 The light cone forms the boundary of the spacelike and timelike separated events. The spacelike separated events cannot have any causal connection to the event at the vertex of the light cone.

events is frame dependent. This means that spacelike separated events must not be causally connected if we are to avoid logical inconsistencies. This requirement prohibits any faster-than-light travel or communication: for two spacelike separated events to communicate with each other, the speed of the signal sent from one to the other must exceed c. Since the chronological order of the two events is frame dependent, such a signal could be received before it is sent.

For any single event, the only events that it can have any causal connection with are those that are either timelike or null separated from it. The locus of all null-separated spacetime points from each event is called the *light cone* of that event. It forms the boundary between timelike and spacelike separated regions, or regions that can and cannot have causal connections with the event (see Fig. 2.3).

2.2.3
Time Dilation and Lorentz Contraction

The relativity of simultaneity leads to some interesting consequences when considering the problem of reading clocks or measuring lengths from different inertial frames. To simplify our discussion, we will only consider one space dimension.

2.2.3.1 Time Dilation

Consider two clocks, one at rest at the spacial origin $x = 0$ of frame A, and the other at rest at the spacial origin $x' = 0$ of frame B. Frame B is moving with velocity β relative to frame A. At the spacetime origin, $(ct', x') = (ct, x) = (0, 0)$, the two clocks are at the same place at the same time, and observers in both frames will agree that both clocks read zero. At later times, each observer will compare the readings of the two clocks *at the same time in their respective frames*.

At time $t = T_A$ in frame A, clock A will be at $(ct, x) = (cT_A, 0)$ while clock B will be at $(ct, x) = (cT_A, \beta cT_A)$ whose coordinates in frame B is $(ct', x') = (cT_A/\gamma, 0)$ (see Fig. 2.4a). So when clock A reads $t = T_A$, clock B will read $t' = T_A/\gamma = T_A\sqrt{1-\beta^2}$ *at the same time in frame A*. The observer in frame A will conclude that during the time interval that clock A advanced from zero to T_A, clock B advanced from zero to $T_A/\gamma = T_A\sqrt{1-\beta^2}$. That is, clock B runs slower compared to clock A by a factor of $\sqrt{1-\beta^2}$.

At time $t' = T_B$ in frame B, clock B will be at $(ct', x') = (cT_B, 0)$ while clock A will be at

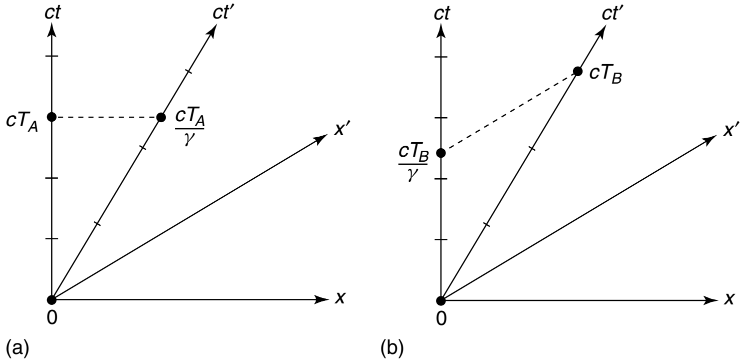

Fig. 2.4 Time dilation. (a) At time T_A in frame A (ct, x), clock at $x' = 0$ in frame B (ct', x') is observed to read T_A/γ. (b) At time T_B in frame B, clock at $x = 0$ in frame A is observed to read T_B/γ.

$(ct', x') = (cT_B, -\beta cT_B)$ whose coordinates in frame A is $(ct, x) = (cT_B/\gamma, 0)$ (see Fig. 2.4b). So when clock B reads $t' = T_B$, clock A will read $t = T_B/\gamma = T_B\sqrt{1-\beta^2}$ *at the same time in frame B*. The observer in frame B will conclude that during the time interval that clock B advanced from zero to T_B, clock A advanced from zero to $T_B/\gamma = T_B\sqrt{1-\beta^2}$. That is, clock A runs slower compared to clock B by a factor of $\sqrt{1-\beta^2}$.

Therefore, both observers conclude that the other observer's clock is running slow by a factor of $\sqrt{1-\beta^2}$. This phenomenon is known as *time dilation*.

2.2.3.2 **Lorentz Contraction**

Consider two rods, one at rest in frame A with its left and right ends, respectively, at $x = 0$ and $x = L_A$, and the other at rest in frame B with its left and right ends, respectively, at $x' = 0$ and $x' = L_B$. Frame B is moving with velocity β relative to frame A. The worldlines of the left and right ends of rod A are given by $x_{A_L}(t) = 0$ and $x_{A_R}(t) = L_A$ in frame A, which in the coordinates of frame B are given by $x'_{A_L}(t') = -\beta ct'$ and $x'_{A_R}(t') = -\beta ct' + L_A/\gamma$ (see Fig. 2.5a). The worldlines of the left and right ends of rod B are given by $x'_{B_L}(t') = 0$ and $x'_{B_R}(t') = L_B$ in frame B, which in the coordinates of frame A are given by $x_{B_L}(t) = \beta ct$ and $x_{B_R}(t) = \beta ct + L_B/\gamma$ (see Fig. 2.5b).

The *length* of an object is the separation between its left and right ends *at the same time*. So the lengths of the two rods measured by an observer in frame A are

$$\begin{aligned} x_{A_R}(t) - x_{A_L}(t) &= L_A \\ x_{B_R}(t) - x_{B_L}(t) &= L_B/\gamma = L_B\sqrt{1-\beta^2} \end{aligned} \tag{20}$$

while those measured by an observer in frame B are

$$\begin{aligned} x'_{A_R}(t') - x'_{A_L}(t') &= L_A/\gamma = L_A\sqrt{1-\beta^2} \\ x'_{B_R}(t') - x'_{B_L}(t') &= L_B \end{aligned} \tag{21}$$

Therefore, the observers will conclude that the moving rods are shorter than the ones at rest by a factor of $\sqrt{1-\beta^2}$. This phenomenon is known as *Lorentz contraction*.

2.2.3.3 **Experimental Evidence**

It is a well-known experimental fact that when an unstable particle travels at speeds

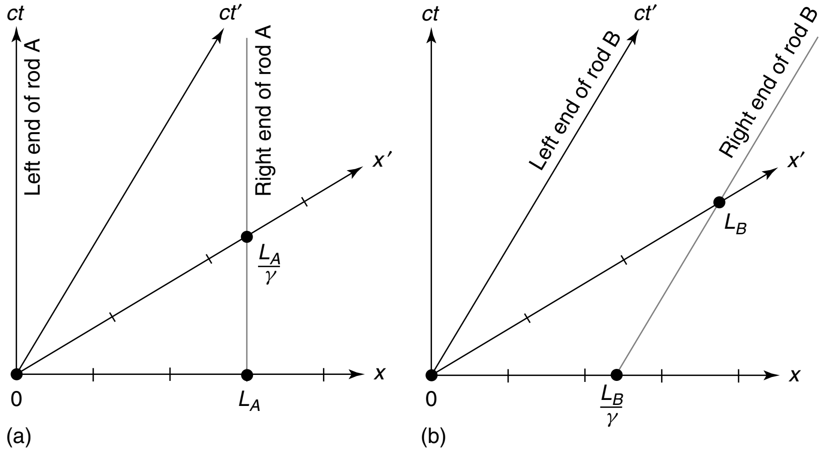

Fig. 2.5 Lorentz Contraction. (a) A rod at rest in frame A (ct, x) of length L_A is observed to have length L_A/γ in frame B (ct', x'). (b) A rod at rest in frame B of length L_B is observed to have length L_B/γ in frame A.

close to c, it lives much longer than its lifetime in its rest frame before decaying. For instance, the lifetime of a muon in its rest frame is about 2×10^{-6} seconds (Particle Data Group 2009). During that time, a particle would only be able to travel about 600 m even if it traveled at the speed of light. However, muons created in the upper regions of the atmosphere at altitudes of tens of kilometers routinely make it down to the surface of the Earth and beyond into its interior.[3]

Let the velocity of the particle in the lab frame be β. If the particle decays after time τ of its creation in its rest frame, the same time interval would be observed as $\gamma\tau$ in the lab frame owing to time dilation. The distance the particle travels before decaying in the lab frame is then

$$L = (\beta c)(\gamma\tau) = \frac{\beta c\tau}{\sqrt{1-\beta^2}} \tag{22}$$

which can be much larger than $c\tau$ when β is close to 1. In the rest frame of the particle, the same distance would be observed as $L/\gamma = \beta c\tau$ owing to Lorentz contraction. Since the lab is moving with velocity $-\beta$ in the particle's rest frame, the length of lab that travels past it during its lifetime τ is indeed $\beta c\tau$. Note that in the lab frame it is time dilation at work, while in the particle frame it is Lorentz contraction. The two effects are like two sides of the same coin, and both are necessary for the phenomenon to make frame-independent sense (see Fig. 2.6).

2.2.3.4 Twin Paradox

The fact that time dilation is symmetric between the two inertial frames A and B, with observers in each frame observing the clock in the other frame run slow, frequently leads to the following question: if both observers see the other clock run slow, how will their clocks read when they are brought back together for comparison? If we replace the clocks with twin siblings, the question becomes: if one stays at home

---

[3] The extension of the muon's lifetime due to time dilation has also been confirmed quantitatively in the lab (Bailey and Picasso, 1970).

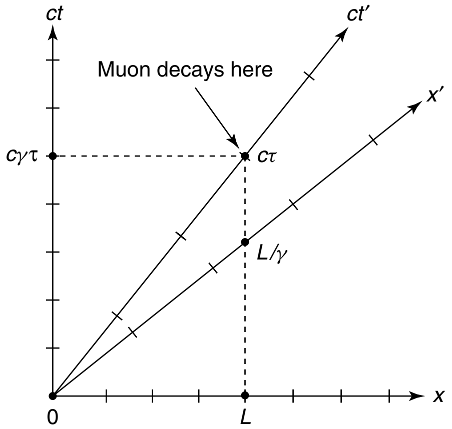

Fig. 2.6 The decay of a muon as seen from the lab frame (ct, x) and the muon's rest frame (ct', x'). The muon is created at $(ct, x) = (ct', x') = (0, 0)$. In the muon's rest frame, it decays at $(ct', x') = (c\tau, 0)$, which in the lab frame is $(ct, x) = (c\gamma\tau, L)$, where $L = c\gamma\tau\beta$. The same distance in the muon's rest frame is observed as $L/\gamma = c\tau\beta$. In this figure, the muon is traveling at $\beta = 0.8$ relative to the lab frame, and $\gamma = 5/3$.

on Earth while the other takes a trip to the stars on a fast moving spaceship, which one will be older when the two meet back on Earth? If both siblings observe the clock of the other run slow during the trip, would both not expect the other to be the younger of the two at the trip's end, leading to a paradox?

The solution to this "twin paradox" becomes apparent when one notes that as long as the two clocks (or siblings) are fixed to inertial frames A and B, they will continue to move away from each other at constant relative velocity, never to meet again at the same point in space. For the clocks to meet back again at the same place, the worldline of at least one of the clocks cannot be inertial, and this breaks the symmetry between the two clocks.

Consider two clocks A and B as in the section on time dilation, that is, clock A is fixed to the spacial origin of frame A (ct, x), while clock B is initially fixed to the spacial origin of frame B (ct', x'). Frame B is moving with velocity β relative to frame A. The two clocks are synchronized at the spacetime origin $(ct, x) = (ct', x') = (0, 0)$. At point $(ct, x) = (cT, \beta cT)$, whose B-coordinates are $(ct', x') = (cT/\gamma, 0)$, clock B turns around and starts back toward clock A with velocity $-\beta$ relative to frame A. The time elapsed until turnaround in frame A is T, while that in frame B is T/γ. Note, however, that from the point of view of frame B, the turnaround point is simultaneous to the point $(ct, x) = (cT/\gamma^2, 0)$ along the worldline of clock A.

After turnaround, the two clocks will move toward each other and meet again at point $(ct, x) = (2cT, 0)$. On the return trip, clock B is fixed to a third frame, which we will call C (ct'', x''). Since C is moving at velocity $-\beta$ relative to frame A, the C-coordinates of the turnaround point are $(ct'', x'') = (2\gamma cT - cT/\gamma, 2\gamma\beta cT)$, while those of the meeting point are $(ct'', x'') = (2\gamma cT, 2\gamma\beta cT)$. The time elapsed between these two points in frame A is T, while that in frame C is T/γ. Note, however, that from the point of view of frame C, the turnaround point is simultaneous to the point $(ct, x) = (2cT - cT/\gamma^2, 0)$ along the worldline of clock A (see Fig. 2.7).

So when the process is observed from the point of view of clock A, the outbound trip took time T, during which clock B advanced by T/γ, while the return trip again took time T, during which clock B again advanced by T/γ. The same process

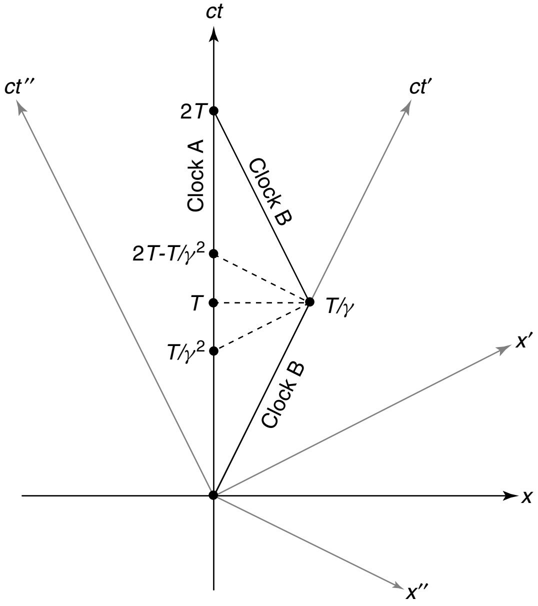

Fig. 2.7 Twin Paradox explained: Clock A stays at rest in frame A (ct, x) throughout. Clock B travels to the right at constant speed, turns around, and travels back to the left at constant speed. Before the turnaround, clock B is in frame B (ct', x'). After the turnaround, it is in frame C (ct'', x''). The point simultaneous with the turnaround point along the worldline of clock A will be different in frames B and C.

observed from the point of view of clock B is different: the outbound trip took time T/γ, during which clock A advanced by T/γ^2, while the return trip again took time T/γ, during which clock A again advanced by T/γ^2. However, at turnaround, when clock B jumps from frame B to frame C, the reading of clock A also jumps by $2T\beta^2$ from T/γ^2 to $2T - T/\gamma^2$, so the total advance of clock A is $T/\gamma^2 + 2T\beta^2 + T/\gamma^2 = 2T$. Therefore, both clocks will agree that while clock A has advanced by $2T$, clock B has only advanced by $2T/\gamma$.

Analysis of a more generic worldline for the accelerating clock, and also in $3+1$ dimensions, is possible as follows. Consider a clock whose worldline starts from $(ct, x, y, z) = (0, 0, 0, 0)$ and ends at $(ct, x, y, z) = (cT, 0, 0, 0)$, that is, it returns to its spacial starting point after time T. Between the points (ct, x, y, z) and $(c(t+dt), x+dx, y+dy, z+dz)$ along the clock's worldline, we can treat the clock as moving with constant velocity $\beta = \sqrt{dx^2+dy^2+dz^2}/cdt$, so the clock will be observed to advance by

$$d\tau = dt\sqrt{1-\beta^2(t)} \;<\; dt \tag{23}$$

Integrating from $t = 0$ to $t = T$, we find

$$\int d\tau = \int_0^T dt\sqrt{1-\beta^2(t)} \;<\; T \tag{24}$$

that is, the clock that is accelerated to return to its starting point will always be delayed compared to the one that was at rest throughout, regardless of which trajectory it took. This effect has been measured and confirmed by atomic clocks flown around the world on airplanes (Hafele and Keating, 1972).

2.2.4
Doppler Effect

2.2.4.1 Doppler Effect in Light

In SR, the speed of light relative to any source or observer is always c. Nevertheless, the observed frequency of light waves will change depending on the relative motion of the source and observer in the following manner.

Collinear case Let a light wave be propagating in the positive x-direction. Frame B (ct', x') is also traveling in the x-direction with velocity $\beta = v/c$ relative to frame A (ct, x). If the frequency of light is ν in frame A, then the worldlines of two successive wavefronts will be given by $x = ct$ and $x = ct + c/\nu$. Changing coordinates to frame B with Eq. (14), these worldlines are given by $x' = ct'$ and $x' = ct' + c/\nu'$ where

$$\nu' = \sqrt{\frac{1-\beta}{1+\beta}}\;\nu \tag{25}$$

For $\beta > 0$, the frequency is red shifted $\nu' < \nu$, while for $\beta < 0$, the frequency is blue-shifted $\nu' > \nu$.

Noncollinear case When the direction of the propagation of light and the boost direction are noncollinear, the above relation will be modified to

$$\nu' = \frac{1-\beta\cos\phi}{\sqrt{1-\beta^2}}\;\nu \tag{26}$$

where ϕ is the angle between light propagation and boost directions as observed in frame A. If we use, instead, the angle ϕ' between light propagation and boost directions as observed in frame B, which is related to ϕ via

$$\cos\phi' = \frac{\beta - \cos\phi}{1-\beta\cos\phi} \tag{27}$$

the relation becomes

$$\nu' = \frac{\sqrt{1-\beta^2}}{1-\beta\cos\phi'}\;\nu \tag{28}$$

which may be more useful for the observer. (This is just Eq. (26) with the primed and unprimed quantities reversed.) The derivations are straightforward. Note that unlike the nonrelativistic Doppler shift in sound waves, the shift is nonzero even in the perpendicular directions, $\phi = \pi/2$ or $\phi' = \pi/2$. This is due to time dilation.

2.2.4.2 Application in Astronomy

Astronomers use the Doppler effect to determine the velocities of faraway stars and galaxies. This is done by looking for atomic spectral lines in the light from stars. Since each atom emits light at multiple characteristic frequencies, it is possible to identify the pattern even when the lines are Doppler shifted. This allows for an accurate determination of the amount of Doppler shift, since the frequencies in the atom's rest frame are well known, and the velocity of the emitting atom can be inferred from Eq. (25). This technique lead to the discovery of the expansion of the universe by Hubble (1929), and the measurement of the rotation curves of spiral galaxies (Rubin, Ford and Thonnard, 1980), which is considered strong evidence for the existence of dark matter.

2.2.4.3 Doppler Radar

As we have seen above, if a light wave is emitted with frequency ν in the emitter's frame, the frequency ν' observed by an object moving away from the emitter with speed β will be given by Eq. (25). If the wave is reflected back by the object toward the emitter, the object is now the source of a light wave with frequency ν' traveling toward the observer who is again moving away with speed β. So the frequency at which the emitter receives the reflected signal will be

$$\nu'' = \sqrt{\frac{1-\beta}{1+\beta}}\;\nu' = \frac{1-\beta}{1+\beta}\;\nu \tag{29}$$

The frequency shift $\Delta\nu = \nu'' - \nu$ is then directly proportional to the speed:

$$\frac{\Delta\nu}{\nu} \approx -2\beta \tag{30}$$

This frequency shift is minute when $\beta \ll 1$, but can still be determined by letting the returning signal interfere with the original signal and measuring the beat.

The device that uses this principle is known as *Doppler radar*, and is used by law enforcement to determine the speed of motorists, by air traffic control and the military to determine the speed of airplanes, by meteorologists to determine the speed of precipitation, and at baseball games to determine the speed at which the pitcher throws the ball.

2.2.5
Addition of Velocities

The addition of velocities is also modified from the familiar Galilei–Newton case. We first consider the addition of collinear velocities, and then generalize to noncollinear ones.

2.2.5.1 Addition of Collinear Velocities

Frame B is moving with velocity $\beta_{\mathrm{BA}} = v/c$ relative to frame A. Let the velocity of an object be $\beta_{\mathrm{A}} = u/c = \Delta x/(c\Delta t)$ in frame A, and $\beta_{\mathrm{B}} = u'/c = \Delta x'/(c\Delta t')$ in frame B. Using Eq. (14) we find

$$\beta_{\mathrm{A}} = \frac{\Delta x}{c\Delta t} = \frac{\Delta x' + \beta_{\mathrm{BA}}(c\Delta t')}{(c\Delta t') + \beta_{\mathrm{BA}}\Delta x'} = \frac{\beta_{\mathrm{B}} + \beta_{\mathrm{BA}}}{1 + \beta_{\mathrm{B}}\beta_{\mathrm{BA}}} \tag{31}$$

or

$$u = \frac{u' + v}{1 + (u'v/c^2)} \tag{32}$$

If $u' = c$ then $u = c$ also, as required, regardless of the value of v. If $u', v < c$, then $u < c$, that is, adding velocities smaller than c will never lead to a sum that exceeds c. In the limit $c \to \infty$, this reduces to the usual $u = u' + v$.

The same relation can be derived by performing two Lorentz transformations in a row, that is, transform from frame A (ct, x) to frame B (ct', x'), which is moving with velocity β_{BA} relative to frame A, followed by a transformation from the frame B (ct', x') to frame C (ct'', x''), which is moving with velocity β_{CB} relative to the frame B. We find

$$\begin{aligned} B(\beta_{\mathrm{CB}})B(\beta_{\mathrm{BA}}) &= \begin{bmatrix} \cosh\eta_{\mathrm{CB}} & -\sinh\eta_{\mathrm{CB}} \\ -\sinh\eta_{\mathrm{CB}} & \cosh\eta_{\mathrm{CB}} \end{bmatrix} \times \begin{bmatrix} \cosh\eta_{\mathrm{BA}} & -\sinh\eta_{\mathrm{BA}} \\ -\sinh\eta_{\mathrm{BA}} & \cosh\eta_{\mathrm{BA}} \end{bmatrix} \\ &= \begin{bmatrix} \cosh(\eta_{\mathrm{CB}}+\eta_{\mathrm{BA}}) & -\sinh(\eta_{\mathrm{CB}}+\eta_{\mathrm{BA}}) \\ -\sinh(\eta_{\mathrm{CB}}+\eta_{\mathrm{BA}}) & \cosh(\eta_{\mathrm{CB}}+\eta_{\mathrm{BA}}) \end{bmatrix} \\ &\equiv \begin{bmatrix} \cosh(\eta_{\mathrm{CA}}) & -\sinh(\eta_{\mathrm{CA}}) \\ -\sinh(\eta_{\mathrm{CA}}) & \cosh(\eta_{\mathrm{CA}}) \end{bmatrix} \\ &= B(\beta_{\mathrm{CA}}) \end{aligned} \tag{33}$$

where we have set

$$\eta_{\mathrm{CA}} = \eta_{\mathrm{CB}} + \eta_{\mathrm{BA}} \tag{34}$$

Note that rapidities simply add, while the velocity of frame C relative to frame A is

$$\begin{aligned} \beta_{CA} &= \tanh(\eta_{\mathrm{CB}} + \eta_{\mathrm{BA}}) \\ &= \frac{\tanh\eta_{\mathrm{CB}} + \tanh\eta_{\mathrm{BA}}}{1 + \tanh\eta_{\mathrm{CB}}\tanh\eta_{\mathrm{BA}}} \\ &= \frac{\beta_{\mathrm{CB}} + \beta_{\mathrm{BA}}}{1 + \beta_{\mathrm{CB}}\beta_{\mathrm{BA}}} \end{aligned} \tag{35}$$

2.2.5.2 Experimental Application

The fact that rapidity is additive is applied in the analysis of particle collisions at hadron colliders such as the $p\bar{p}$ collider Tevatron at Fermilab and the *pp* collider LHC (Large Hadron Collider) at CERN. Since protons and antiproton are composite particles, the actual colliding particles are their constituent quarks, antiquarks, and gluons. Consequently, the center-of-mass frame of the colliding constituents is generically boosted in the beam direction relative to the lab frame, and the distribution of the particles coming out from the collision point will be Lorentz contracted in the beam direction. In order to undo this Lorentz contraction, instead of using the angles (θ, ϕ) to designate the outcoming particle's direction, where θ is the angle from the beam axis and ϕ is the angle wrapping around it, hadronic colliders replace θ with a quantity called *pseudorapidity* defined as

$$\eta \equiv \tanh^{-1}(\cos\theta) = -\ln\left(\tan\frac{\theta}{2}\right) \qquad (36)$$

and use the pair (η, ϕ) to keep track of the outcoming particles. Note that if a particle coming out of the collision point is traveling at the speed of light, $\cos\theta$ is just its velocity in the beam direction (in units of c) and η is the corresponding rapidity. A boost in the beam direction will simply shift η by a constant, which does not change the shape of the particle distribution in η. Actual particles coming out of a collision are light mesons and leptons, which will be traveling at speeds slightly less than c, but the difference is too minute to be relevant. However, since η is not exactly the rapidity for those particles, it is called *pseudo* rapidity to remind us of this fact.

2.2.5.3 Addition of Noncollinear Velocities and Wigner Rotation

Let frame B (ct', x', y', z') be moving with velocity $\boldsymbol{\beta}_{\mathrm{BA}} = (\beta_{\mathrm{BA}}, 0, 0)$ relative to frame A (ct, x, y, z). The velocity of an object is $\boldsymbol{\beta}_{\mathrm{A}} = (\beta_{\mathrm{A}x}, \beta_{\mathrm{A}y}, 0)$ in frame A, and $\boldsymbol{\beta}_{\mathrm{B}} = (\beta_{\mathrm{B}x}, \beta_{\mathrm{B}y}, 0)$ in frame B. Using Eq. (18), we find

$$\begin{aligned}
\beta_{\mathrm{A}x} &= \frac{\Delta x}{c\Delta t} = \frac{\Delta x' + \beta_{\mathrm{BA}}(c\Delta t')}{(c\Delta t') + \beta_{\mathrm{BA}}\Delta x'} \\
&= \frac{\beta_{\mathrm{B}x} + \beta_{\mathrm{BA}}}{1 + \beta_{\mathrm{B}x}\beta_{\mathrm{BA}}} \\
\beta_{\mathrm{A}y} &= \frac{\Delta y}{c\Delta t} = \frac{\sqrt{1-\beta_{\mathrm{BA}}^2}\,\Delta y'}{(c\Delta t') + \beta_{\mathrm{BA}}\Delta x'} \\
&= \frac{\sqrt{1-\beta_{\mathrm{BA}}^2}}{1 + \beta_{\mathrm{B}x}\beta_{\mathrm{BA}}}\,\beta_{\mathrm{B}y} \qquad (37)
\end{aligned}$$

Performing two Lorentz boosts in a row in noncollinear directions will yield the same result and something extra. Let the Lorentz boost from frame A to frame B be $B(\beta\mathbf{n}_{\mathrm{BA}})$, and that from frame B to frame C be $B(\beta'\mathbf{n}_{\mathrm{CB}})$, where β and β' are the speeds of B relative to A, and C relative to B, respectively, and $\mathbf{n}_{\mathrm{BA}}$ and $\mathbf{n}_{\mathrm{CB}}$ are unit vectors. Define the angle ϕ as $\mathbf{n}_{\mathrm{CB}} \cdot \mathbf{n}_{\mathrm{BA}} \equiv \cos\phi$. When the velocities are noncollinear, that is, $\phi \neq 0, \pi$, straightforward calculations using Eq. (19) show that the product of the two Lorentz boosts is not a pure boost but a product of a Lorentz boost and a rotation:

$$B(\beta'\mathbf{n}_{\mathrm{CB}})B(\beta\mathbf{n}_{\mathrm{BA}}) = R(\boldsymbol{\theta}_{\mathrm{W}})B(\beta''\mathbf{n}_{\mathrm{CA}}) \qquad (38)$$

where $\beta''\mathbf{n}_{\mathrm{CA}}$ is the velocity of frame C observed from frame A, and the rotation $R(\boldsymbol{\theta}_{\mathrm{W}})$ is known as *Wigner rotation* (Wigner, 1939). The velocity $\beta''\mathbf{n}_{\mathrm{CA}}$ is given by

$$\begin{aligned}
\beta''\mathbf{n}_{\mathrm{CA}} = {} & \frac{\beta + \beta'\cos\phi}{1 + \beta\beta'\cos\phi}\,\mathbf{n}_{\mathrm{BA}} \\
& + \frac{\beta'\sin\phi}{\gamma(1 + \beta\beta'\cos\phi)}\,\mathbf{n}_{\mathrm{BA}}^{\perp} \qquad (39)
\end{aligned}$$

where $\mathbf{n}^{\perp}_{\mathrm{BA}} \sin\phi \equiv (\mathbf{n}_{\mathrm{CB}} - \mathbf{n}_{\mathrm{BA}} \cos\phi)$, making $\mathbf{n}^{\perp}_{\mathrm{BA}}$ a unit vector perpendicular to $\mathbf{n}_{\mathrm{BA}}$, and

$$\beta'' = \frac{\sqrt{(\beta^2 + 2\beta\beta' \cos\phi + \beta'^2) - (\beta\beta' \sin\phi)^2}}{1 + \beta\beta' \cos\phi} \tag{40}$$

Equation (39) reproduces Eq. (37). The Wigner rotation is a rotation around the $\mathbf{n}_{\mathbf{CB}} \times \mathbf{n}_{\mathrm{BA}}$ axis with rotation angle θ_{W} given by

$$\tan\frac{\theta_{\mathrm{W}}}{2} = \frac{\sin\phi}{\delta + \cos\phi}, \qquad \delta = \sqrt{\left(\frac{\gamma+1}{\gamma-1}\right)\left(\frac{\gamma'+1}{\gamma'-1}\right)} \tag{41}$$

For collinear velocities $\phi = 0, \pi$, and also in the nonrelativistic limits $\gamma, \gamma' \to 1$, we obtain $\theta_{\mathrm{W}} = 0$.

2.3
Relativistic Dynamics

In this section, we discuss how Newton's second law must be modified to be covariant under Lorentz transformations. Maxwell's equations are found to be already Lorentz covariant, and the electrodynamics of a charged particle is discussed.

2.3.1
The Covariance of F = *m*a under Galilei Transformations and Rotations

The objective of dynamics is to determine the worldline of a particle (i.e., point mass) in spacetime. In Newtonian dynamics, one usually specifies the worldline by giving the position of the particle in space, $\mathbf{x}$, as a function of the time t: $\mathbf{x}(t)$. The *momentum* of the particle is defined as

$$\mathbf{p}(t) \equiv m\mathbf{u}(t) = m\frac{d\mathbf{x}(t)}{dt} \tag{42}$$

Newton's second and third laws are equivalent to the statement that for a set of interacting particles, the total momentum of the system is conserved:

$$\frac{d}{dt}\left[\sum_i \mathbf{p}_i(t)\right] = \sum_i \frac{d\mathbf{p}_i(t)}{dt} = 0 \tag{43}$$

Focusing our attention on only one particle of the system, we obtain

$$\frac{d\mathbf{p}_i(t)}{dt} = -\sum_{j\neq i} \frac{d\mathbf{p}_j(t)}{dt} \equiv \mathbf{f}_i(t) \tag{44}$$

The *force* $\mathbf{f}_i$ is the momentum transfer rate to particle i. Dropping the index i, and introducing the acceleration

$$\mathbf{a}(t) = \frac{d\mathbf{u}(t)}{dt} = \frac{d^2\mathbf{x}(t)}{dt^2} \tag{45}$$

we obtain Newton's second law:

$$\mathbf{f}(t) = m\mathbf{a}(t) \tag{46}$$

If given the force $\mathbf{f}(t)$, this equation can be solved for the worldline $\mathbf{x}(t)$.

Equation (46) is invariant under Galilei transformations since $t' = t$, $\mathbf{x}' = \mathbf{x} - \mathbf{v}t$ leaves the acceleration and force unchanged: $\mathbf{a}' = \mathbf{a}$, $\mathbf{f}' = \mathbf{f}$. It is also covariant under rotations since the acceleration $\mathbf{a}$ and force $\mathbf{f}$ inherit their transformation properties under rotations from $\mathbf{x}$ and transform as *vectors*. That is, if $\mathbf{x} \to \mathbf{x}' = R\mathbf{x}$, where R is a rotation matrix, then $\mathbf{a}' = R\mathbf{a}$, and $\mathbf{f}' = R\mathbf{f}$, and Eq. (46) remains valid after the rotation even though the individual components of the vectors themselves will change.

2.3.2
Covariance under Lorentz Transformations

2.3.2.1 4-Vectors

The covariance of $\mathbf{f} = m\mathbf{a}$ under rotations provides a template to follow in order to find its Lorentz covariant modification. Let us denote the spacetime coordinates (ct, x, y, z) of a particle collectively as a four-component object using the symbol x with a Greek superscript:

$$x^\mu = \begin{bmatrix} ct \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} ct \\ x \\ y \\ z \end{bmatrix} \tag{47}$$

The superscript μ is an index that runs from 0 to 3, with $x^0 = ct$, $x^1 = x$, $x^2 = y$, and $x^3 = z$. Denote a Lorentz transformation matrix with $\Lambda^\mu{}_\nu$:

$$x^\mu \longrightarrow x'^\mu = \Lambda^\mu{}_\nu x^\nu \tag{48}$$

Here, we have employed the Einstein convention in which repeated indices are dummies that have been summed. Δs^2 from Eq. (16) can be written as

$$\Delta s^2 = g_{\mu\nu}\, \Delta x^\mu \Delta x^\nu,$$
$$g_{\mu\nu} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \tag{49}$$

and the invariance of Δs^2 requires

$$g_{\mu\nu} \Lambda^\mu{}_\rho \Lambda^\nu{}_\sigma = g_{\rho\sigma} \tag{50}$$

which is the condition we have already encountered in Eq. (17).

We will call any four-component object which transforms in the same way as x^μ under Lorentz transformations a *4-vector* and denote them with a symbol bearing a Greek superscript, for example A^μ. To distinguish the usual spacial three-dimensional vectors (3-vectors) from 4-vectors, the 3-vectors will be denoted with a boldface symbol, For example $\mathbf{A}$, or a symbol bearing a latin index, for example A^i, with the understanding that the Latin index only runs from 1 to 3. The dual of a 4-vector A^μ is denoted with a Greek subscript:

$$A_\mu = g_{\mu\nu} A^\nu = \left[A^0, -\mathbf{A} \right] \tag{51}$$

If we denote the inverse matrix of $g_{\mu\nu}$ with $g^{\mu\nu}$, which happens to be the same matrix, we can also write

$$A^\mu = g^{\mu\nu} A_\nu \tag{52}$$

The Lorentz invariant scalar product of two 4-vectors can then be written as

$$\begin{aligned} A \cdot B \equiv g_{\mu\nu} A^\mu B^\nu &= A^\mu B_\mu = A_\nu B^\nu \\ &= A^0 B^0 - \mathbf{A} \cdot \mathbf{B} \end{aligned} \tag{53}$$

With this convention, summing a lower and upper index pair becomes a Lorentz invariant procedure, and the metric $g_{\mu\nu}$ and its inverse $g^{\mu\nu}$ can be used to lower and raise indices.

Newton's second law, being a relation between 3-vectors, $f^i = ma^i$, is manifestly covariant under rotations. To obtain a manifestly Lorentz covariant equation, we must seek a relation between 4-vectors, namely, an equation of the form $f^\mu = ma^\mu$ with f^μ and a^μ appropriately defined.

2.3.2.2 The Relativistic Equation of Motion

Proper Time As a first step in this direction, we note that describing the worldline with the function $\mathbf{x}(t)$ is inconvenient since t is a frame-dependent quantity. It makes more sense to use a parametric representation of the worldline in which the spacetime coordinates of the particle x^μ are given as functions of a parameter defined along the worldline in an observer-independent way. The simplest of such parameters would be the reading of the clock that is affixed to the moving particle. (We have already considered such a clock in our discussion on the twin paradox.) Let us call this the *proper time* τ. The proper time separating the points $x^\mu = (ct, \mathbf{x})$ and $x^\mu + dx^\mu = (c(t+dt), \mathbf{x} + d\mathbf{x})$ along the particle worldline is

$$\begin{aligned} c\,d\tau &= \sqrt{dx_\mu dx^\mu} = \sqrt{c^2 dt^2 - |d\mathbf{x}|^2} \\ &= c\,dt\sqrt{1-\beta^2} = \frac{c}{\gamma}\,dt \end{aligned} \tag{54}$$

Note that $d\tau$ is well defined as long as $\beta < 1$. Using τ, the worldline of the particle is specified by the functions $x^\mu(\tau)$.

4-Velocity: The derivative of $x^\mu(\tau)$ with respect to the proper time τ is called the particle's *4-velocity*,

$$\begin{aligned} u^\mu(\tau) &= \frac{dx^\mu(\tau)}{d\tau} = \gamma \begin{bmatrix} c \\ \mathbf{v} \end{bmatrix} = \gamma c \begin{bmatrix} 1 \\ \boldsymbol{\beta} \end{bmatrix}, \\ \boldsymbol{\beta} &= \frac{\mathbf{v}}{c} = \frac{1}{c}\frac{d\mathbf{x}}{dt} \end{aligned} \tag{55}$$

This object is a 4-vector as the symbol suggests since the proper time τ is a Lorentz scalar, and u^μ inherits its transformation properties from x^μ. Also, since the proper time was defined via $dx_\mu dx^\mu = c^2 d\tau^2$, the four components of u^μ are not independent and satisfy $u^2 = u_\mu u^\mu = c^2$. $u^\mu(\tau)/c$ is the unit tangent 4-vector to the particle worldline at τ.

4-Momentum: The particle's *4-momentum* is defined as

$$p^\mu = \begin{bmatrix} p^0 \\ \mathbf{p} \end{bmatrix} \equiv m\,u^\mu = \gamma mc \begin{bmatrix} 1 \\ \boldsymbol{\beta} \end{bmatrix} \tag{56}$$

where $m \neq 0$ is the particle's mass. Note that $p^2 = p_\mu p^\mu = m^2c^2$, and $\boldsymbol{\beta} = \mathbf{p}/p^0$. In the nonrelativistic limit $\beta \ll 1$, we find

$$\begin{aligned} cp^0 &\longrightarrow mc^2\left(1 + \tfrac{1}{2}\boldsymbol{\beta}^2 + \cdots\right) \\ &\qquad = mc^2 + \tfrac{1}{2}mv^2 + \cdots \\ \mathbf{p} &\longrightarrow mc\boldsymbol{\beta} + \cdots = m\mathbf{v} + \cdots \end{aligned} \tag{57}$$

Since $\frac{1}{2}mv^2$ is the nonrelativistic kinetic energy of the particle, we identify cp^0 as the relativistic energy E. The spacial components $\mathbf{p}$ of p^μ are identified as the relativistic 3-momentum. The energy E is then related to the 3-momentum via

$$E = \sqrt{\mathbf{p}^2c^2 + m^2c^4} \tag{58}$$

When the particle is at rest, $\mathbf{p} = \mathbf{0}$, this equation reduces to the famous $E = mc^2$, the meaning of which is discussed later.

4-Acceleration: The derivative of $u^\mu(\tau)$ with respect to τ is the particle's *4-acceleration*,

$$\begin{aligned} a^\mu &= \frac{du^\mu}{d\tau} = \frac{d^2x^\mu}{d\tau^2} \\ &= \gamma^4 \begin{bmatrix} \boldsymbol{\beta}\cdot\mathbf{a}_\beta \\ \mathbf{a}_\beta + \boldsymbol{\beta}\times(\boldsymbol{\beta}\times\mathbf{a}_\beta) \end{bmatrix}, \\ \mathbf{a}_\beta &= \frac{d^2\mathbf{x}}{dt^2} \end{aligned} \tag{59}$$

Again, this object is a 4-vector, inheriting its transformation properties from x^μ, justifying the notation a^μ. The 4-acceleration is always *perpendicular* to the 4-velocity, $u \cdot a = u_\mu a^\mu = 0$, as a consequence of $u^2 = u_\mu u^\mu = c^2$.

In the co-moving frame at τ, that is, in the inertial frame in which the particle is instantaneously at rest at τ, the 4-acceleration reduces to

$$a^\mu \longrightarrow \begin{bmatrix} 0 \\ \mathbf{a}_0 \end{bmatrix} \tag{60}$$

We can identify the spacial components of a^μ in this frame, namely $\mathbf{a_0}$, as the usual 3-acceleration of Newtonian mechanics. Therefore, $a^2 = a_\mu a^\mu$, which is a Lorentz invariant, is equal to

$$a^2 = a_\mu a^\mu = -\mathbf{a}_0^2 \tag{61}$$

4-Force: The Lorentz covariant generalization of $f^i = ma^i$ will thus be of the form

$$f^\mu = ma^\mu \tag{62}$$

where f^μ must be a 4-vector, which would be called the *4-force*. Since

$$ma^\mu = \frac{dp^\mu}{d\tau} = \frac{d}{d\tau}\begin{bmatrix} E/c \\ \mathbf{p} \end{bmatrix} \tag{63}$$

the spacial components $\mathbf{f}$ of f^μ are the momentum transfer rate $d\mathbf{p}/d\tau$ to the particle, while the time component f^0 of f^μ is the energy transfer rate $dE/d\tau$, or power, divided by c.

In the co-moving frame at τ, we have seen that the 4-acceleration reduces to Eq. (60), so the 4-force must reduce to

$$f^\mu \longrightarrow \begin{bmatrix} 0 \\ \mathbf{f}_0 \end{bmatrix} \tag{64}$$

where $\mathbf{f}_0$ is the usual Newtonian 3-force. f^μ in all other frames can be obtained by Lorentz transforming from this co-moving frame.

2.3.2.3 A Simple Example

As an example, consider motion in $1+1$ dimensions in which the particle is subject to a 2-force which is constant in the co-moving frame:

$$f^\mu_{\text{co-moving}} = \begin{bmatrix} 0 \\ ma \end{bmatrix}, \quad a = \text{constant} \tag{65}$$

This is the closest analog to motion under constant acceleration in Newtonian dynamics. Let the rapidity of the particle at proper time τ be $\eta(\tau)$ relative to the lab frame. Then, the particle's 2-velocity and 2-acceleration will be

$$u^\mu(\tau) = c\begin{bmatrix} \cosh\eta(\tau) \\ \sinh\eta(\tau) \end{bmatrix},$$
$$a^\mu(\tau) = c\begin{bmatrix} \sinh\eta(\tau) \\ \cosh\eta(\tau) \end{bmatrix}\frac{d\eta(\tau)}{d\tau} \tag{66}$$

On the other hand, the 2-force in the lab frame will be

$$f^\mu(\tau) = \begin{bmatrix} \cosh\eta(\tau) & \sinh\eta(\tau) \\ \sinh\eta(\tau) & \cosh\eta(\tau) \end{bmatrix} f^\mu_{\text{co-moving}} = ma\begin{bmatrix} \sinh\eta(\tau) \\ \cosh\eta(\tau) \end{bmatrix} \tag{67}$$

Therefore, $f^\mu = ma^\mu$ implies

$$c\,\frac{d\eta(\tau)}{d\tau} = a \tag{68}$$

which can be solved trivially to yield $c\eta(\tau) = a\tau$. Further integrating $u^\mu(\tau)$, we obtain

$$x^\mu(\tau) = \begin{bmatrix} ct(\tau) \\ x(\tau) \end{bmatrix} = \frac{c^2}{a}\begin{bmatrix} \sinh(a\tau/c) \\ \cosh(a\tau/c) \end{bmatrix} \tag{69}$$

Integration constants have been set to zero to simplify our expressions. The particle's worldline is then a hyperbola with the past and future light cones of the spacetime origin as the asymptotes (see Fig. 2.8):

$$x^2 - (ct)^2 = \frac{c^4}{a^2} \tag{70}$$

2.3.2.4 What is the Inertia of the Particle?

Now that we have a Lorentz covariant equation of motion, $f^\mu = ma^\mu = dp^\mu/d\tau$,

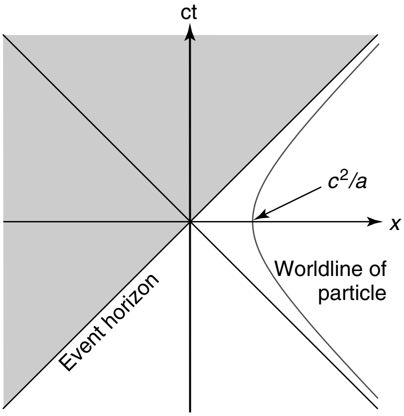

Fig. 2.8 Motion of particle under constant acceleration in the co-moving frame. No signal from the shaded region can reach the particle, effectively creating an event horizon for any observer traveling with it.

let us ask the question, what is the inertia of the particle?

Recalling that the inertial mass in Newtonian mechanics was the ratio of the rate of momentum transfer (the force) to the rate of velocity change (the acceleration), it makes sense to define the *inertia* of the particle by

$$\Delta\mathbf{p} = (\text{inertia})\Delta\mathbf{v} \tag{71}$$

even in the relativistic case. And since the relativistic 3-momentum of a particle of mass m is given by

$$\mathbf{p} = \gamma m\mathbf{v} = (E/c^2)\mathbf{v} \tag{72}$$

we can make the identification

$$(\text{inertia}) = \frac{E}{c^2} = \frac{m}{\sqrt{1-\beta^2}} \tag{73}$$

Therefore, the energy E (divided by c^2) is the inertia, and not the mass m which is Lorentz invariant.[4] In the rest frame of the particle, $E/c^2 = m$, so the inertia is equal to the mass. However, as $\beta \to 1$ the inertia blows up to infinity, making it increasingly harder to accelerate the particle as the speed of light is approached, preventing the particle from ever exceeding it. This was seen in the example problem we discussed earlier.[5]

2.3.2.5 Action, Lagrangian, and Hamiltonian of a free particle

In the absence of the 4-force, the relativistic equation of motion is simply

$$ma^\mu = \frac{dp^\mu}{d\tau} = 0 \tag{75}$$

---

[4] In some traditions, E/c^2 is called the *mass* while m is called the *rest mass*. We do not adopt this convention, however, since it is equivalent to giving two different names to E.

[5] Einstein (1905a) in his original paper proposed calling the proportionality constant relating $\mathbf{f}_0$, the Newtonian force in the co-moving frame, and $\mathbf{a}_\beta$, the Newtonian acceleration in the lab frame, the "mass". If this definition is adopted, then we find that the directions parallel and perpendicular to the velocity will have different "masses". Boosting Eq. (64) to the lab frame and comparing with Eq. (59), we find

$$\mathbf{f}_{0\parallel} = (\gamma^3 m)\mathbf{a}_{\beta\parallel}, \quad \mathbf{f}_{0\perp} = (\gamma^2 m)\mathbf{a}_{\beta\perp} \tag{74}$$

That is, $\mathbf{f}_0$ and $\mathbf{a}_\beta$ are not co-linear in general. This is not surprising considering that they are quantities measured in different inertial frames, and in the case of $\mathbf{a}_\beta$ it is not even the spacial part of a 4-vector. Einstein called $\gamma^3 m$ the *longitudinal mass* and $\gamma^2 m$ the *transverse mass*. Neither of them is equal to $E/c^2 = \gamma m$.

the solution of which is $p^\mu = mu^\mu =$ constant. This is Newton's first law, also known as the law of inertia, which is still valid in SR. The action that leads to this equation is

$$S = -mc^2 \int d\tau \sqrt{g_{\mu\nu} \frac{dx^\mu(\tau)}{c\,d\tau} \frac{dx^\nu(\tau)}{c\,d\tau}} \qquad (76)$$

The quantity in the square-root is 1 if $x^\mu(\tau)$ is the defining worldline of the proper time τ, which we take to be the stationary point of the action S. Since S is the Lorentz invariant path length of the particle worldline, this stationary point is the worldline for which its path length is at an extremum.

The requirement $\delta S = 0$ leads to the equation

$$\frac{d^2 x_\rho}{d\tau^2} + \Gamma^\rho{}_{\mu\nu} \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau} = 0 \qquad (77)$$

where we have allowed $g_{\mu\nu}$ to depend on the spacetime point and defined:

$$\Gamma^\rho{}_{\mu\nu} \equiv \tfrac{1}{2} g^{\rho\sigma} \left(\partial_\mu g_{\sigma\nu} + \partial_\nu g_{\sigma\mu} - \partial_\sigma g_{\mu\nu} \right) \qquad (78)$$

(The notation ∂_μ is introduced later in Eq. (108).) Equation (77) is the *geodesic equation*. The solution will be a *geodesic*, the "straightest" line in spacetime. For inertial frames, $g_{\mu\nu}$ is constant and $\Gamma^\rho{}_{\mu\nu} = 0$, and we recover $a^\mu = 0$. For noninertial frames, $g_{\mu\nu}$ will depend on the spacetime point x^μ, leading to a nonzero $\Gamma^\rho{}_{\mu\nu}$, and the solution to the geodesic equation will no longer be a straight line, accounting for the *inertial forces* acting on the particle.

To further justify our choice of action in Eq. (76), let us work in a particular frame and replace the proper time τ with the time t in that frame. The action reduces to

$$S = \int dt\, \mathcal{L}, \quad \mathcal{L} = -mc^2 \sqrt{1 - \frac{v^2}{c^2}} \qquad (79)$$

The momentum conjugate to x^i is then

$$p^i = \frac{\partial \mathcal{L}}{\partial \dot{x}^i} = \gamma m v^i \qquad (80)$$

and the Hamiltonian is

$$H = p^i v^i - \mathcal{L} = \gamma m c^2 \qquad (81)$$

as they should be, leading to the standard relation

$$H = \sqrt{\mathbf{p}^2 c^2 + m^2 c^4} \qquad (82)$$

2.3.3
Parallel Transport of 4-Vectors

2.3.3.1 Fermi–Walker Transport

Next, let us consider the evolution of 4-vectors that are attached to the moving particle. An example of such a vector would be the particle's intrinsic angular momentum, or *spin*. Let the vector in question be $S^\mu(\tau; \mathbf{0})$ at proper time τ when the particle's velocity is $\boldsymbol{\beta}$, and $S^\mu(\tau + d\tau; \mathbf{0})$ at proper time $\tau + d\tau$ when the particle's velocity is $\boldsymbol{\beta} + d\boldsymbol{\beta}$. The second argument on S^μ denotes the velocity of the frame from which the measurement is made, and the $\mathbf{0}$ designates the lab frame. The difference between the two vectors as observed in this frame is

$$dS^\mu(\tau; \mathbf{0}) = S^\mu(\tau + d\tau; \mathbf{0}) - S^\mu(\tau; \mathbf{0}) \qquad (83)$$

However, if $d\boldsymbol{\beta}$ is nonzero, part of this difference will be a consequence of the vector being boosted to a different frame. To subtract out this boost effect, let us first boost to the co-moving frame of the particle at τ by acting on both sides of Eq. (83) with the boost $B(\boldsymbol{\beta})$ to obtain

$$dS^{\mu}(\tau;\boldsymbol{\beta}) = S^{\mu}(\tau+d\tau;\boldsymbol{\beta}) - S^{\mu}(\tau;\boldsymbol{\beta}) \tag{84}$$

$S^{\mu}(\tau;\boldsymbol{\beta}) = B^{\mu}{}_{\nu}(\boldsymbol{\beta})\, S^{\nu}(\tau;\mathbf{0})$ is now measured in the rest frame of the particle. However, $S^{\mu}(\tau+d\tau;\boldsymbol{\beta}) = B^{\mu}{}_{\nu}(\boldsymbol{\beta})\, S^{\nu}(\tau+d\tau;\mathbf{0})$ still is not, so an additional infinitesimal Lorentz boost is necessary to bring it to its rest frame:

$$\begin{aligned} &S^{\mu}(\tau+d\tau;\boldsymbol{\beta}+d\boldsymbol{\beta}) \\ &\quad = B^{\mu}{}_{\nu}(\mathbf{a}_0\, d\tau)\, S^{\nu}(\tau+d\tau;\boldsymbol{\beta}) \end{aligned} \tag{85}$$

Note that since we are performing a pure Lorentz boost, any change in the vector's coordinates here is due to the change in the particle's velocity and not due to any rotation of the coordinate axes. Thus, the net change in the vector with the boost effect subtracted out is

$$\begin{aligned} DS^{\mu}(\tau;\boldsymbol{\beta}) = {} & S^{\mu}(\tau+d\tau;\boldsymbol{\beta}+d\boldsymbol{\beta}) \\ & - S^{\mu}(\tau;\boldsymbol{\beta}) \end{aligned} \tag{86}$$

Boosting back to the lab frame by acting on both sides of this equation by $B(-\boldsymbol{\beta})$, we obtain

$$\begin{aligned} DS^{\mu}(\tau,\mathbf{0}) = {} & \Lambda^{\mu}{}_{\nu}(\boldsymbol{\beta}+d\boldsymbol{\beta},\boldsymbol{\beta})\, S^{\nu}(\tau+d\tau;\mathbf{0}) \\ & - S^{\mu}(\tau;\mathbf{0}) \end{aligned} \tag{87}$$

where

$$\begin{aligned} &\Lambda^{\mu}{}_{\nu}(\boldsymbol{\beta}+d\boldsymbol{\beta},\boldsymbol{\beta}) \\ &\quad = B^{\mu}{}_{\lambda}(-\boldsymbol{\beta})\, B^{\lambda}{}_{\kappa}(\mathbf{a}_0\, d\tau)\, B^{\kappa}{}_{\nu}(\boldsymbol{\beta}) \end{aligned} \tag{88}$$

which can also be written in Lorentz covariant form as

$$\Lambda^{\mu}{}_{\nu} = \delta^{\mu}_{\nu} + \left(u^{\mu}a_{\nu} - a^{\mu}u_{\nu}\right)\frac{d\tau}{c^2} \tag{89}$$

Therefore, the rate of net change in S^{μ} as observed from the lab, or any other frame, is given by

$$\frac{DS^{\mu}}{d\tau} = \frac{dS^{\mu}}{d\tau} + \frac{1}{c^2}\left(u^{\mu}a_{\nu} - a^{\mu}u_{\nu}\right)S^{\nu} \tag{90}$$

For example, if we take $S^{\mu} = u^{\mu}$, we find

$$\frac{Du^{\mu}}{d\tau} = \frac{du^{\mu}}{d\tau} + \frac{1}{c^2}\left(u^{\mu}a_{\nu} - a^{\mu}u_{\nu}\right)u^{\nu} = 0 \tag{91}$$

since $u^2 = u_{\mu}u^{\mu} = c^2$ and $a \cdot u = a_{\mu}u^{\mu} = 0$, which is a consequence of the fact that the velocity of the particle is always zero in its own co-moving frame. The 2-force we encountered in Eq. (67) also satisfies the condition $Df^{\mu}/d\tau = 0$ since it was always equal to Eq. (65) in the co-moving frame.

If the vector S^{μ} is to stay the same in the particle's co-moving frame at all times, it must evolve according to the equation

$$\frac{dS^{\mu}}{d\tau} = -\frac{1}{c^2}\left(u^{\mu}a_{\nu} - a^{\mu}u_{\nu}\right)S^{\nu} \tag{92}$$

Such an evolution corresponds to *parallel transporting* the vector from one co-moving frame to the next along the particle's worldline, and is known as the *Fermi–Walker transport*. In terms of

$$\boldsymbol{\beta} = \frac{\mathbf{v}}{c} = \frac{1}{c}\frac{d\mathbf{x}}{dt}, \quad \text{and} \quad \mathbf{a}_{\beta} = \frac{d^2\mathbf{x}}{dt^2} \tag{93}$$

$\Lambda^{\mu}{}_{\nu}$ of Eq. (89) is given by

$$\begin{aligned} &\Lambda(\boldsymbol{\beta},\mathbf{a}_{\beta}) = \mathbf{1}_{4\times 4} \\ &\qquad - \gamma^{3}\begin{bmatrix} 0 & \mathbf{a}_{\beta}^{T} \\ \mathbf{a}_{\beta} & (\mathbf{a}_{\beta}\times\boldsymbol{\beta})\times \end{bmatrix}\frac{d\tau}{c} \end{aligned} \tag{94}$$

In the co-moving frame, this reduces to the pure infinitesimal Lorentz boost of Eq. (85):

$$\begin{aligned} \Lambda(\mathbf{0},\mathbf{a}_0) &= \begin{bmatrix} 1 & -\mathbf{a}_0^{T}\, d\tau/c \\ -\mathbf{a}_0\, d\tau/c & \mathbf{1}_{3\times 3} \end{bmatrix} \\ &= B(\mathbf{a}_0\, d\tau) \end{aligned} \tag{95}$$

However, in all other frames, it is clear from Eq. (94) that the same infinitesimal Lorentz transformation involves a rotation unless $\boldsymbol{\beta}$ and $\mathbf{a}_{\beta}$ are co-linear. This is an instance of Wigner rotation that we encountered earlier; note that

$$\Lambda(\boldsymbol{\beta}, \mathbf{a}_{\beta}) = B(-\boldsymbol{\beta})\Lambda(\mathbf{0}, \mathbf{a}_0)B(\boldsymbol{\beta}) \tag{96}$$

and compare with Eq. (38). What this means is that when vectors are Fermi–Walker transported along the particle's worldline, they will precess when observed from a fixed inertial frame.

2.3.3.2 **Thomas Precession**

A famous example is the precession of the particle's spin. The spin of a particle in its rest frame is a 3-vector $\mathbf{S}$. It can be treated as the spacial components of a 4-vector S^{μ}, which satisfies the condition $u^{\mu}S_{\mu} = 0$.

Consider a particle moving in a circular orbit of radius r in the xy-plane, at constant speed $v = r\omega$:

$$x = r\cos\omega t, \quad y = r\sin\omega t, \quad z = 0 \tag{97}$$

The 4-velocity and 4-acceleration of this particle are

$$u^{\mu} = \gamma\begin{bmatrix} c \\ -v\sin\omega t \\ v\cos\omega t \\ 0 \end{bmatrix}, \qquad a^{\mu} = \gamma^2\begin{bmatrix} 0 \\ -v\omega\cos\omega t \\ -v\omega\sin\omega t \\ 0 \end{bmatrix} \tag{98}$$

from which we find

$$(u^{\mu}a_{\nu} - a^{\mu}u_{\nu}) = \gamma^3 cv\omega \times \begin{bmatrix} 0 & \cos\omega t & \sin\omega t & 0 \\ \cos\omega t & 0 & -\beta & 0 \\ \sin\omega t & \beta & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{99}$$

If the force responsible for keeping the particle in this orbit does not supply a torque (when observed from the co-moving frame), then the particle's spin S^{μ} will be Fermi–Walker transported along the orbit and its evolution will be governed by Eq. (92), which in this case becomes

$$\begin{aligned} \frac{dS^0}{dt} &= -\gamma^2\beta\omega\left(S^1\cos\omega t + S^2\sin\omega t\right) \\ \frac{dS^1}{dt} &= -\gamma^2\beta\omega\left(S^0\cos\omega t - \beta S^2\right) \\ \frac{dS^2}{dt} &= -\gamma^2\beta\omega\left(S^0\sin\omega t + \beta S^1\right) \end{aligned} \tag{100}$$

Along with the condition $u^{\mu}S_{\mu} = 0$, this can be solved to yield

$$\begin{aligned} S^0 &= -\gamma\beta S\sin\gamma\omega t \\ S^1 &= S\left(\cos\omega t\cos\gamma\omega t + \gamma\sin\omega t\sin\gamma\omega t\right) \\ S^2 &= S\left(\sin\omega t\cos\gamma\omega t - \gamma\cos\omega t\sin\gamma\omega t\right) \end{aligned} \tag{101}$$

where we have chosen the initial orientation of the spin vector at $(x, y) = (r, 0)$ to be along the x-axis. (Misner, Thorne and Wheeler, 1973, p.175, Exercise 6.9).

To see what is happening when $\beta \ll 1$, we rewrite the above result as

$$\begin{bmatrix} S^1 \\ S^2 \end{bmatrix} = S\left\{\begin{bmatrix} \cos(1-\gamma)\omega t \\ \sin(1-\gamma)\omega t \end{bmatrix} + (1-\gamma) \times \sin\gamma\omega t\begin{bmatrix} -\sin\omega t \\ \cos\omega t \end{bmatrix}\right\} \tag{102}$$

Since $(1-\gamma) \approx -\beta^2/2 < 0$, the first term makes the spin precess in the retrograde direction with angular velocity $\omega_T = (\gamma - 1)\omega \approx \beta^2\omega/2$. The direction of precession of the second term depends on the sign of $\sin\gamma\omega t$, and this contribution will average to zero over time. Another way to see this is to define $\phi(t) = \tan^{-1}(S^2/S^1)$ and expand in β^2. We will find

$$\frac{d\phi}{dt} = -\frac{\beta^2\omega}{2}(1+\cos 2\omega t) + O(\beta^4) \tag{103}$$

The oscillating term comes from the second term in Eq. (102), which will average to zero. This retrograde precession of the spin with average angular velocity ω_T is known as *Thomas precession* (Thomas, 1927). For spins to precess, however, there must be a torque acting on it, and it is not necessarily obvious how there can be a torque in the lab frame when there is none in the co-moving frame. An illuminating discussion of this problem can be found in Muller (1992).

The ultrarelativistic limit $\gamma \to \infty$, on the other hand, is singular, with rapid oscillation involving the time component S^0. This is due to the absence of a frame in which the particle is at rest when $\beta = 1$. However, if we concentrate on the precession in the xy-plane, we will find

$$\phi(t) = \omega t \pm \frac{\pi}{2} \tag{104}$$

In this limit, the spin will be aligned with the particle's velocity and precess with it, which is required for massless particles (Wigner, 1939).

2.3.4
Electrodynamics

2.3.4.1 Covariant Formulation of Maxwell's Equations

Maxwell's equations in Gaussian units are given by (Jackson 1998)

$$\begin{aligned} \boldsymbol{\nabla}\cdot\mathbf{B} &= 0, & \boldsymbol{\nabla}\times\mathbf{E} + \frac{1}{c}\frac{\partial\mathbf{B}}{\partial t} &= 0 \\ \boldsymbol{\nabla}\cdot\mathbf{E} &= 4\pi\rho, & \boldsymbol{\nabla}\times\mathbf{B} - \frac{1}{c}\frac{\partial\mathbf{E}}{\partial t} &= \frac{4\pi}{c}\mathbf{J} \end{aligned} \tag{105}$$

The scalar and vector potentials, ϕ and $\mathbf{A}$, are introduced as

$$\mathbf{B} = \boldsymbol{\nabla}\times\mathbf{A}, \quad \mathbf{E} = -\boldsymbol{\nabla}\phi - \frac{1}{c}\frac{\partial\mathbf{A}}{\partial t} \tag{106}$$

so that the two equations in the first line of Eq. (105) are automatically satisfied. The two equations in the second line of Eq. (105) in terms of ϕ and $\mathbf{A}$ are

$$\begin{aligned} &\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2\right)\phi - \frac{1}{c}\frac{\partial}{\partial t}\left(\frac{1}{c}\frac{\partial\phi}{\partial t} + \boldsymbol{\nabla}\cdot\mathbf{A}\right) \\ &\quad = 4\pi\rho \\ &\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2\right)\mathbf{A} + \boldsymbol{\nabla}\left(\frac{1}{c}\frac{\partial\phi}{\partial t} + \boldsymbol{\nabla}\cdot\mathbf{A}\right) \\ &\quad = \frac{4\pi}{c}\mathbf{J} \end{aligned} \tag{107}$$

Introducing the following relativistic notations

$$\begin{aligned} \partial_\mu &\equiv \frac{\partial}{\partial x^\mu} = \begin{bmatrix} \frac{1}{c}\frac{\partial}{\partial t} & \boldsymbol{\nabla} \end{bmatrix}, \\ \partial^\mu &\equiv \frac{\partial}{\partial x_\mu} = \begin{bmatrix} \frac{1}{c}\frac{\partial}{\partial t} \\ -\boldsymbol{\nabla} \end{bmatrix} \end{aligned} \tag{108}$$

and also realizing that

$$J^\mu = \begin{bmatrix} c\rho \\ \mathbf{J} \end{bmatrix} \tag{109}$$

transforms as a 4-vector (since it is proportional to the 4-velocities of the charged particles that make up the charge and current densities), Eq. (107) can be rewritten as

$$\begin{aligned} \partial_\mu\partial^\mu\phi - \partial^0\left(\partial_0\phi + \partial_j A^j\right) &= \frac{4\pi}{c}J^0 \\ \partial_\mu\partial^\mu A^i - \partial^i\left(\partial_0\phi + \partial_j A^j\right) &= \frac{4\pi}{c}J^i \end{aligned} \tag{110}$$

The Lorentz covariance of these equations require that

$$A^\mu = \begin{bmatrix} \phi \\ \mathbf{A} \end{bmatrix} \tag{111}$$

transform as a 4-vector. Equation (110) can then be written as

$$\partial_\mu \partial^\mu A^\nu - \partial^\nu(\partial_\mu A^\mu) = \frac{4\pi}{c} J^\nu \tag{112}$$

or

$$\partial_\mu F^{\mu\nu} = \frac{4\pi}{c} J^\nu, \qquad F^{\mu\nu} \equiv \partial^\mu A^\nu - \partial^\nu A^\mu \tag{113}$$

The elements of the antisymmetric tensor $F^{\mu\nu}$ are nothing but the electric and magnetic fields:

$$F^{\mu\nu} = \begin{bmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{bmatrix} \tag{114}$$

If we act on both sides of Eq. (113) with ∂_ν, the left-hand side will vanish owing to the antisymmetry of $F^{\mu\nu}$, and consistency requires the right-hand side also to vanish:

$$\partial_\nu J^\nu = \frac{\partial \rho}{\partial t} - \nabla \cdot \mathbf{J} = 0 \tag{115}$$

Thus, the electromagnetic charge must be conserved.

2.3.4.2 Lorentz Transformation Properties of the Electromagnetic Field

Since the electric and magnetic field strengths **E** and **B** have been identified as components of the antisymmetric tensor $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$, this allows us to derive the Lorentz transformation properties of those fields from those of the 4-vectors A^μ and ∂^μ. If the fields are $(\mathbf{E}, \mathbf{B})$ in the lab frame, the fields $(\mathbf{E}', \mathbf{B}')$ in the frame moving with velocity $\boldsymbol{\beta} = \beta\mathbf{n}$ relative to it are

$$\begin{aligned} \mathbf{E}' &= \mathbf{E}_\parallel + \gamma\,(\mathbf{E}_\perp + \boldsymbol{\beta} \times \mathbf{B}) \\ \mathbf{B}' &= \mathbf{B}_\parallel + \gamma\,(\mathbf{B}_\perp - \boldsymbol{\beta} \times \mathbf{E}) \end{aligned} \tag{116}$$

where

$$\begin{aligned} \mathbf{E}_\parallel &= \mathbf{n}\,(\mathbf{n} \cdot \mathbf{E})\,, \quad \mathbf{E}_\perp = \mathbf{E} - \mathbf{E}_\parallel \\ \mathbf{B}_\parallel &= \mathbf{n}\,(\mathbf{n} \cdot \mathbf{B})\,, \quad \mathbf{B}_\perp = \mathbf{B} - \mathbf{B}_\parallel \end{aligned} \tag{117}$$

Note, however, that the following quantities are Lorentz invariant:

$$\begin{aligned} F_{\mu\nu} F^{\mu\nu} &= -2(\mathbf{E}^2 - \mathbf{B}^2) \\ \tfrac{1}{2}\,\varepsilon_{\kappa\lambda\mu\nu} F^{\kappa\lambda} F^{\mu\nu} &= -4\mathbf{E} \cdot \mathbf{B} \end{aligned} \tag{118}$$

which serve as a useful check.

The above relations show that even in the absence of the electric field in the lab frame, there will be a nonzero electric field in the moving frame as long as there is a nonzero magnetic field in the lab frame. This helps us understand the origin of the Lorentz force, which is described next.

2.3.4.3 The Action and Coupling to a Charged Particle

The action that leads to Eq. (113) is

$$S = \int d^4x \left[-\frac{1}{16\pi} F_{\mu\nu} F^{\mu\nu} - \frac{1}{c} A_\mu J^\mu \right] \tag{119}$$

For a point particle of charge q, the current density is given by

$$J^\mu(x) = q \int d\tau\, u^\mu(\tau) \delta^4(x - x(\tau)) \tag{120}$$

where x^μ is a generic spacetime point where J^μ is evaluated, and $x^\mu(\tau)$ is the worldline of the particle. (Superscripts are dropped when 4-vectors appear as arguments since the function itself does not carry the index.) The coupling term becomes

$$\begin{aligned} &\frac{1}{c} \int d^4x\, A_\mu(x) J^\mu(x) \\ &\quad \longrightarrow \frac{q}{c} \int d\tau\, A_\mu(x(\tau)) u^\mu(\tau) \end{aligned} \tag{121}$$

The total action of the electromagnetic field and the charged particle is therefore

$$S = -\frac{1}{16\pi}\int d^4x\, F_{\mu\nu}F^{\mu\nu} - \frac{q}{c}\int d\tau\, A_\mu u^\mu - mc^2\int d\tau \tag{122}$$

where the last term is shorthand for the right-hand side of Eq. (76). The equation of motion of the particle is then

$$ma^\mu = \frac{q}{c}F^{\mu\nu}u_\nu = q\gamma\begin{bmatrix} \mathbf{E}\cdot\boldsymbol{\beta} \\ \mathbf{E}+\boldsymbol{\beta}\times\mathbf{B} \end{bmatrix} \tag{123}$$

and we can identify the familiar Lorentz force law.

2.3.4.4 Lagrangian and the Hamiltonian in the Nonrelativistic Limit

Choosing to work in a particular frame allows us to write the particle-dependent part of the action as

$$S = \int dt\,\mathcal{L},$$
$$\mathcal{L} = -mc^2\sqrt{1-\frac{v^2}{c^2}} - q\left(\phi - \mathbf{A}\cdot\frac{\mathbf{v}}{c}\right) \tag{124}$$

The momentum conjugate to x^i is then

$$p^i = \frac{\partial\mathcal{L}}{\partial\dot{x}^i} = \gamma m v^i + \frac{q}{c}A^i \tag{125}$$

and the Hamiltonian is

$$H = p^i v^i - \mathcal{L} = \gamma mc^2 + q\phi \tag{126}$$

Notice that these relations can be written collectively as

$$\left(p^\mu - \frac{q}{c}A^\mu\right) = mu^\mu \tag{127}$$

This gives us the prescription that in order to couple a particle of charge q to the electromagnetic field, we should make the replacement $p^\mu \to p^\mu - (q/c)A^\mu$. The relation between H and the momentum is therefore

$$H = \sqrt{\left(\mathbf{p}-\frac{q}{c}\mathbf{A}\right)^2 c^2 + m^2c^4} + q\phi \tag{128}$$

which in the nonrelativistic limit, and neglecting irrelevant constants, becomes

$$H = \frac{1}{2m}\left(\mathbf{p}-\frac{q}{c}\mathbf{A}\right)^2 + q\phi \tag{129}$$

2.3.4.5 Spin-Orbit Coupling

When a particle with charge q also carries spin, $\mathbf{S}$, then Eq. (129) will receive a correction since the particle will then carry a magnetic dipole moment given by

$$\mathbf{m} = \frac{q}{mc}\mathbf{S} \tag{130}$$

which will interact with the magnetic field it sees in its co-moving frame. The derivation of Eq. (130) for spin $\frac{1}{2}$ particles is given in the appendix.

Consider the hydrogen atom. The electric and magnetic fields in the center-of-mass frame are

$$\mathbf{E} = \frac{e\mathbf{r}}{r^3}, \qquad \mathbf{B} = \mathbf{0} \tag{131}$$

which can be obtained from

$$\phi = \frac{e}{r}, \qquad \mathbf{A} = \mathbf{0} \tag{132}$$

and the nonrelativistic Hamiltonian of the electron is

$$H = \frac{p^2}{2m} - \frac{e^2}{r} \tag{133}$$

Then, the magnetic field that the electron sees in its co-moving frame can be obtained using Eq. (116), and we find

$$\mathbf{B}' = -\boldsymbol{\beta}\times\mathbf{E} = -\frac{\mathbf{p}}{mc}\times(-\nabla\phi) = \frac{e}{mcr^3}\mathbf{L} \tag{134}$$

where $\mathbf{L} = \mathbf{r} \times \mathbf{p}$ is the electron's orbital angular momentum. So the interaction energy that must be added to the Hamiltonian would be

$$\begin{aligned} H'_{\text{spin-orbit}} &= -\mathbf{m} \cdot \mathbf{B}' \\ &= -\left(-\frac{e}{mc}\mathbf{S}\right) \cdot \left(\frac{e}{mcr^3}\mathbf{L}\right) \\ &= \frac{e^2}{m^2c^2}\frac{\mathbf{S} \cdot \mathbf{L}}{r^3} \end{aligned} \tag{135}$$

This interaction will cause $\mathbf{S}$ to precess around the orbital angular momentum $\mathbf{L}$ with angular velocity

$$\omega_L = \frac{eB'}{mc} = \beta^2\omega \tag{136}$$

where we have used the relation $e^2/r^2 = mv^2/r$ for circular orbits, and ω is the orbital angular velocity. However, this is the result in the co-moving frame of the electron, and in the lab frame the spin also undergoes the kinematical Thomas precession, which we discussed earlier. The Thomas precession frequency was

$$\omega_T = (\gamma - 1)\omega \approx \tfrac{1}{2}\beta^2\omega \tag{137}$$

in the retrograde direction. So the net precession observed in the lab frame would be

$$\omega_L - \omega_T = \tfrac{1}{2}\omega_L \tag{138}$$

This suggests that the correction term that should be added to the Hamiltonian Eq. (133) should differ from Eq. (135) by a factor of 2:

$$H_{\text{spin-orbit}} = \frac{e^2}{2m^2c^2}\frac{\mathbf{S} \cdot \mathbf{L}}{r^3} \tag{139}$$

The argument used to derive this expression is hand-waving at best, though it is found in many textbooks, and is not very convincing. We note, however, that the correctness of Eq. (139) has been experimentally confirmed from the measurement of the fine structure of atomic spectra. A more rigorous and also elegant derivation of Eqs (130) and (139) is available via a fully relativistic treatment of spin using the Dirac equation (Darwin, 1928; Dirac 1928) (see appendix).

2.4
Photons

Photons are the quanta of the electromagnetic wave, which were introduced by Einstein (1905a) to explain the photoelectric effect in 1905. They are massless particles that travel at the speed of light c. In this section, we discuss phenomena and applications associated with the photon and the physical meaning of the famous equation $E = mc^2$.

2.4.1
Quantum Mechanics

In quantum mechanics, a particle with energy E and momentum $\mathbf{p}$ is associated with a wave with angular frequency $\omega = E/\hbar$, and wave number $\mathbf{k} = \mathbf{p}/\hbar$, where $\hbar$ is Planck's (1901) constant. We begin by checking whether this makes relativistic sense.

2.4.1.1 **Angular Frequency and Wave Number**

The phase of a plane wave of frequency ν, wavelength λ, propagating in the direction of the unit vector $\mathbf{n}$, at the spacetime point $x^\mu = (ct, \mathbf{x})$ is given by

$$2\pi\left(\nu t - \frac{\mathbf{n} \cdot \mathbf{x}}{\lambda}\right) = (\omega t - \mathbf{k} \cdot \mathbf{x}) \tag{140}$$

where $\omega = 2\pi\nu$ is the angular frequency, and $\mathbf{k} = (2\pi/\lambda)\mathbf{n}$ is the wave-number vector. Since such a phase is a physical quantity that all frames must agree upon, it must be a Lorentz invariant. This requires $(\omega/c, \mathbf{k})$ to transform as a 4-vector, which we denote as

$$k^\mu = \begin{bmatrix} k^0 \\ \mathbf{k} \end{bmatrix} = \begin{bmatrix} \omega/c \\ \mathbf{k} \end{bmatrix} \tag{141}$$

The phase can then be written in Lorentz invariant form as $k \cdot x = k_\mu x^\mu$. Recalling that the energy and momentum of a particle comprise the 4-momentum,

$$p^\mu = \begin{bmatrix} E/c \\ \mathbf{k} \end{bmatrix} \tag{142}$$

the quantum mechanical relations $E = \hbar\omega$, $\mathbf{p} = \hbar\mathbf{k}$ can be written collectively as

$$p^\mu = \hbar k^\mu \tag{143}$$

Therefore, if $E = \hbar\omega$, relativity requires $\mathbf{p} = \hbar\mathbf{k}$, and vice versa.

2.4.1.2 Velocities

Once the above identification is made, we must require $k^2 = k_\mu k^\mu = m^2c^2/\hbar^2 \equiv 1/\lambda\!\!\!^{-2}$, that is, the dispersion relation[6]

$$k^0 = \frac{\omega(\mathbf{k})}{c} = \sqrt{\mathbf{k}^2 + \frac{m^2c^2}{\hbar^2}} = \sqrt{\mathbf{k}^2 + \frac{1}{\lambda\!\!\!^{-2}}} \tag{145}$$

The phase and group velocities of the wave are then

$$\begin{aligned} v_p &= \frac{\omega}{|\mathbf{k}|} = c\frac{k^0}{|\mathbf{k}|} \\ v_g &= \frac{d\omega}{d|\mathbf{k}|} = c\frac{|\mathbf{k}|}{k^0} \end{aligned} \tag{146}$$

whereas the velocity of the particle is

$$\mathbf{v} = c\boldsymbol{\beta} = c\frac{\mathbf{p}}{p^0} = c\frac{|\mathbf{k}|}{k^0} \tag{147}$$

So the group velocity of wave packets will equal the velocity of the particle.

2.4.1.3 Massless Particles

If we set the mass of the particle, m, to zero in the above relations, we will find that all three velocities, namely the phase and group velocities of the wave and the velocity of the particle, are equal to c. So associating a massless particle, the photon, which travels at c, with electromagnetic waves whose phase velocity is c, makes perfect sense since the electromagnetic waves' group velocity is also c.

Unlike massive particles, a photon does not have a proper time associated with it since it does not have a rest frame. This means that its 4-velocity cannot be defined either. Nevertheless, as physical particles photons carry energy and momentum that is described by a 4-momentum, and the relation $p^\mu = \hbar k^\mu$ determines what they are.

Historically, the relation $E = \hbar\omega$ was suggested by Planck's theory of black-body radiation (Planck, 1901), and used by Einstein (1905a) to explain the photoelectric effect. The relation $\mathbf{p} = \hbar\mathbf{k}$ was checked via Compton scattering (Compton, 1923).

---

[6] The quantity

$$\lambda\!\!\!^{-} = \frac{\hbar}{mc} \tag{144}$$

is the only quantity with the dimension of length that can be constructed from the mass m and the fundamental constants $\hbar$ and c. It is known as the *Compton wavelength* though it is not the wavelength of any wave in particular. The quantity $\lambda = h/mc$ is also known as the Compton wavelength, and it differs from $\lambda\!\!\!^{-}$ by a factor of 2π, that is, $\lambda\!\!\!^{-} = \lambda/2\pi$.

2.4.2
Doppler Effect Revisited

2.4.2.1 Doppler Effect on the Energy of a Photon

The relation $p^\mu = \hbar k^\mu$, when applied to photons, allows us to understand the Doppler effect as shifts in the photon's energy and momentum. Indeed, if the 4-momentum of the photon were $(E/c, \mathbf{p})$ in frame A, and $(E'/c, \mathbf{p}')$ in frame B moving at velocity $\boldsymbol{\beta} = \beta\mathbf{n}$ relative to it, then Eq. (19) tells us that

$$\begin{aligned}(E'/c) &= \gamma(E/c) - \gamma\beta(\mathbf{p}\cdot\mathbf{n}) \\ &= \frac{1-\beta\cos\phi}{\sqrt{1-\beta^2}}\,(E/c)\end{aligned} \tag{148}$$

where we have made use of $E = |\mathbf{p}|c$, and ϕ is the angle between $\mathbf{p}$ and $\mathbf{n}$. Thus, Eq. (26) is recovered in a simple manner.

2.4.2.2 Relativistic Aberration and Relativistic Beaming

The same Lorentz boost tells us that the photon momenta in the two frames are related via

$$\begin{aligned}\mathbf{p}'_\parallel &= |\mathbf{p}'|\cos\phi' = |\mathbf{p}|\left(\frac{\cos\phi-\beta}{\sqrt{1-\beta^2}}\right)\mathbf{n} \\ \mathbf{p}'_\perp &= |\mathbf{p}'|\sin\phi' = \mathbf{p}_\perp\end{aligned} \tag{149}$$

from which we find

$$\begin{aligned}\cos\phi' &= \frac{\cos\phi-\beta}{1-\beta\cos\phi}, \quad \text{and} \\ \cos\phi &= \frac{\cos\phi'+\beta}{1+\beta\cos\phi'}\end{aligned} \tag{150}$$

(The first equation recovers Eq. (27) up to a minus sign since there ϕ' was defined as the angle between $\mathbf{p}'$ and $-\mathbf{n}$.) It is clear from these expressions that as β increases toward 1, the value of $\cos\phi'$ will tend toward -1 regardless of the value of $\cos\phi$, while the value of $\cos\phi$ will tend toward $+1$ regardless of the value of $\cos\phi'$. So if photons are emitted in frame B isotropically from a source at rest in frame B, then when observed from frame A, the direction of the emitted photons will be shifted toward the direction that the source is moving. This effect is known as *relativistic aberration*.[7] Combined with the fact that the photons will be blue-shifted in the direction of motion, while red-shifted in the opposite direction, the electromagnetic energy will be focused into a tight beam in the direction of motion for sources moving at ultrarelativistic speeds. This effect is known as *relativistic beaming* (see Fig. 2.9).

To be more quantitative, let the number photons emitted into solid angle $d\Omega'$ during the time interval dt' in frame B be

$$d^2N' = \frac{N}{4\pi}\,d\Omega' dt' \tag{151}$$

That is, the total number of photons emitted per unit time is N. If each photon carries energy E', then the total energy emitted into the solid angle is

$$d^2W' = \frac{N}{4\pi}E'\,d\Omega' dt' \tag{152}$$

From Eq. (150) we have

$$d\Omega' = \left(\frac{\sqrt{1-\beta^2}}{1-\beta\cos\phi}\right)^2 d\Omega \tag{153}$$

while $dt' = \sqrt{1-\beta^2}\,dt$ from time dilation. The energy of the photons in the two frames is related via Eq. (148). Therefore, the energy emitted into solid angle $d\Omega$ during the time interval dt in frame A will be

$$d^2W = \frac{NE'}{4\pi}\,\frac{(\sqrt{1-\beta^2})^4}{(1-\beta\cos\phi)^3}\,d\Omega\,dt \tag{154}$$

---

[7] The same effect is responsible for collimating the decay products of highly energetic particles into *jets* in collider experiments.

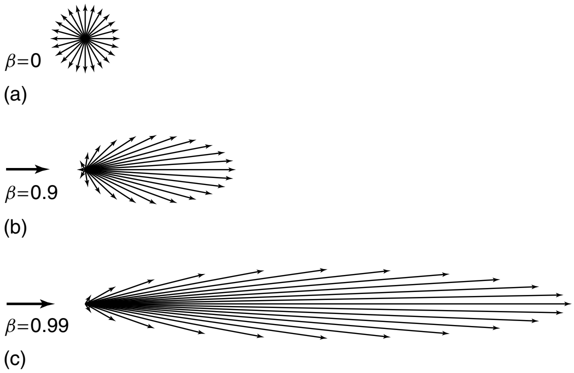

Fig. 2.9 Relativistic Beaming: A source emits monochromatic photons isotropically in its rest frame (a). The lengths of the arrows indicate the photon's energy (frequency). When observed from a frame relative to which this source is moving, the direction of the emitted photons will shift toward the direction of the source's velocity (relativistic aberration). Together with the effect of the Doppler frequency shift, the energy of the photons will be focused in the direction of motion. Shown are the $\beta = 0.9$ (b) and $\beta = 0.99$ (c) cases. The effect of time dilation is not shown.

As $\beta \to 1$, it is clear that this distribution will be highly peaked around $\phi = 0$. Note, however, that when integrated over the entire 4π solid angle, the rate of energy emission in both frames are the same:

$$\frac{dW}{dt} = \frac{dW'}{dt'} = NE' \tag{155}$$

This is due to the fact that both energy and time are the temporal components of 4-vectors and transform in the same way (Rybicki and Lightman, 1985).

2.4.3
Conservation of 4-Momentum and $E = mc^2$

For an isolated set of interacting particles, the total 4-momentum of the system,

$$P^{\mu} = \sum_i p_i^{\mu} \tag{156}$$

is conserved. The interactions can include those that change particle types and numbers. In general, while P^{μ} and the total invariant mass $P^2 = P_{\mu}P^{\mu} = M^2$ are conserved even for particle number/type changing processes, the sum of the individual particle masses $\sum_i m_i$ is not.

Following the argument of Einstein's second paper listed in Einstein (1905a), consider the process in which a particle A at rest emits two photons back to back and becomes another particle (or state) B:

$$A \longrightarrow B + \gamma + \gamma \tag{157}$$

(The γ here denotes a photon, not the γ-factor $1/\sqrt{1-\beta^2}$.) If the photon momenta are the same magnitude, B will still be at rest from momentum conservation. Then, energy conservation requires

$$m_A c^2 = m_B c^2 + 2E_{\gamma} \tag{158}$$

where E_{γ} is the energy carried away by each photon. Since the photons are massless, the total mass of the system has decreased. In this process, the photon energies must have come from the conversion of some internal energy latent in particle A, the release of which lead to a loss of mass. Thus, one can say that the mass of a particle, or any object for that matter, is a measure of the total latent energy it carries. One can also say that part of the mass of A has been converted to energy with the conversion rate given by $E = mc^2$.

2.4.4
Gravitational Red-Shift

Strictly speaking, a discussion on the *gravitational red-shift*, predicted in Einstein (1911), belongs in the article on *General Relativity*. However, since it is understandable using the tools already available, we discuss it here.

We have seen earlier that E/c^2 is the inertia of a massive particle. Invoking the equivalence of inertial and gravitational masses, $E/c^2 = m_G$, if this particle is made to climb a gravitational potential difference of $\Delta\Phi > 0$, we expect the energy of the particle to change by

$$\Delta E = -m_G\,\Delta\Phi = -\frac{E}{c^2}\,\Delta\Phi \qquad (159)$$

or

$$\frac{\Delta E}{E} = -\frac{\Delta\Phi}{c^2} \qquad (160)$$

If we assume that the same relation applies to photons as well as to massive particles, we obtain

$$\frac{\Delta\nu}{\nu} = -\frac{\Delta\Phi}{c^2} \qquad (161)$$

That is, the photon will be red-shifted if it, too, is made to climb the gravitational potential difference.

Einstein uses a slightly different (though essentially the same) argument in his article of 1911. He invokes the *equivalence principle* and replaces a system with a uniform gravitational field in the negative z direction, with gravitational acceleration g, with a system accelerating in the positive z direction with acceleration g. Let a photon be emitted from the origin upward in the positive z direction. When it reaches a height of $z = h$, the detector there will be in a frame moving faster by $\Delta v \approx gh/c$ than the frame in which the photon was emitted. Therefore, owing to the Doppler effect, the frequency at which the photon is detected is

$$\begin{aligned}\nu' &\approx \left(1 - \frac{\Delta v}{c}\right)\nu = \left(1 - \frac{gh}{c^2}\right)\nu \\ &= \left(1 - \frac{\Delta\Phi}{c^2}\right)\nu \qquad (162)\end{aligned}$$

which agrees with Eq. (161). This prediction was checked in a terrestrial experiment using γ-rays from a ^{57}Fe source (Pound and Rebka, 1959).

The implications of the gravitational red-shift are profound. (See Einstein (1911), and also Misner, Thorne and Wheeler (1973), Chapter 7.) Consider two observers: A at $z = 0$ and B at $z = h$ in the uniform gravitational field of acceleration g we just considered. Let A continuously send a light beam to B. When seen globally, the situation is completely static. The gravitational field is constant, neither A nor B is moving, and a beam of light propagates from A toward B with a fixed number of crests and troughs in between (see Fig. 2.10b). Yet, the frequency at which A emits the light beam and that at which B receives it are different. This can only mean that the rate at which time progresses is different depending on the gravitational potential. Since the frequency is inversely proportional to the rate of time progression, the rate at which time passes at A must be slower than at B. More specifically,

$$d\tau_A \approx \left(1 - \frac{\Phi_B - \Phi_A}{c^2}\right) d\tau_B \qquad (163)$$

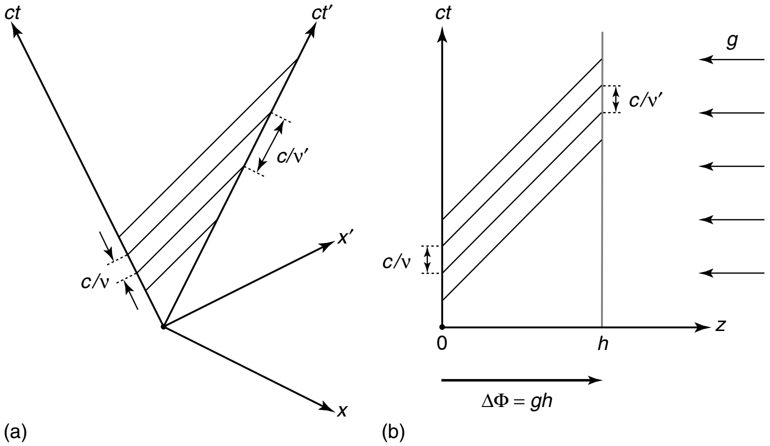

Fig. 2.10 Doppler (a) versus gravitational (b) red-shift. The proper times at $z = 0$ and $z = h$ must progress at a different rate for $\nu \neq \nu'$. The effect of gravity on the propagation of light is not shown in figure (b).

Clearly, this type of dependence cannot be encompassed within the flat spacetime formalism of SR. Therefore, gravitational red-shift actually anticipates general relativity (GR), whose result agrees with the above to leading order in $\Delta\Phi/c^2$.

2.4.5
Global Positioning System

The global positioning system (GPS), which is now quite ubiquitous and available even in cell phones, cannot function without taking time dilation and gravitational red-shift into account.

The GPS consists of a constellation of 24 satellites. Each is equipped with an atomic clock, and periodically sends out light signals indicating its exact location in space, and the exact instant in time when the signal was emitted. If a receiver, who is also carrying an atomic clock, receives such a signal, he/she can immediately calculate his/her distance from the satellite since the time interval it took for the light signal to reach him/her will be known. This will put the receiver's location somewhere on the surface of a sphere with the satellite at its center and the radius equal to the calculated distance. Signals from three satellites will then be sufficient to figure out the receiver's three-dimensional location in space. In practice, of course, no one goes around carrying an atomic clock. So signals from four satellites are used to figure out the receiver's complete four-dimensional location in spacetime. Signals from more satellites will provide redundancy in the information.

For this scheme to work, however, it is important that the clocks on all the satellites are synchronized with the clock on the Earth. And for this, both time dilation and gravitational red-shift must be taken into account. Time dilation due to the satellites' fast motion will make the satellite clocks run slower, and the shallower gravitational potential at the satellites' altitudes will make them run faster. See Ashby; Taylor, Wheeler and Bertschinger (2008) for a complete account of these, and all other relativistic corrections that must be taken into account.

2.5
Further Reading

This concludes our overview of SR. Though many of the predictions of SR often seem esoteric and beyond everyday experience, and perhaps only relevant to high-energy physicists or science fiction writers, some modern technologies, such as the GPS, have advanced to a point where SR and also GR are relevant to their function. Future technologies may further increase the relevance of SR (and GR) in our everyday lives.

Readers interested in further details of SR should consult the many excellent textbooks available on *electrodynamics* (e.g., Jackson, 1998). Textbooks on *general relativity* often (not always) have extensive discussions on SR before plunging deep into GR territory (Hartle, 2003). For applications of SR in particle physics, the readers are referred to textbooks on *particle physics* (Halzen and Martin, 1984) and *quantum field theory* (Gross, 1995, 1996, 2000). Readers interested in the foundations of SR should read the original articles of Einstein (1905a) himself, which are widely available in English translation (Hawking, 2003). Popular accounts can be found in Einstein (2006).

Appendix A

2.A.1
Representations of the Lorentz Group

In this appendix, we present a brief overview of the representations of the Lorentz Group. To simplify our notation, c is set equal to 1.

2.A.1.1
Conventions

The metric, $g_{\mu\nu}$, will be

$$g_{\mu\nu} = g^{\mu\nu} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \tag{A.1}$$

$g^{\mu\nu}$ is the inverse of $g_{\mu\nu}$, which happens to be the same matrix. The spacetime coordinates are represented by *contravariant* vectors:

$$x^{\mu} = \begin{bmatrix} t \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} t \\ \mathbf{x} \end{bmatrix} \tag{A.2}$$

A Latin letter with a lowercase Greek superscript will represent a contravariant 4-vector, the Greek index running from 0 to 3. Boldface Latin letters are 3-vectors. When we need an index running through the spacial components 1 through 3, we will use a lowercase latin letter such as $i, j, k, \cdots$. The *covariant* vectors are the *dual vectors* of the contravariant ones,

$$x_{\mu} = g_{\mu\nu} x^{\nu} = \begin{bmatrix} t \\ -x \\ -y \\ -z \end{bmatrix} = \begin{bmatrix} t \\ -\mathbf{x} \end{bmatrix} \tag{A.3}$$

and are represented by a Latin letter with a lowercase Greek subscript. (We do not necessarily represent dual vectors as row vectors.) The reason for the nomenclature, *contravariant* and *covariant*, is due to the way the components transform under a basis transformation: the *contra*variant components of a 4-vector transform under the *inverse* transformation of the basis vectors (so that the 4-vector stays the same), while the *co*variant components transform under the same transformation as the basis vectors.

Derivatives with respect to the contravariant and covariant components are denoted as

$$\partial_\mu \equiv \frac{\partial}{\partial x^\mu} = \begin{bmatrix} \dfrac{\partial}{\partial t} \\ \dfrac{\partial}{\partial x} \\ \dfrac{\partial}{\partial y} \\ \dfrac{\partial}{\partial z} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial}{\partial t} \\ \nabla \end{bmatrix},$$

$$\partial^\mu \equiv \frac{\partial}{\partial x_\mu} = \begin{bmatrix} \dfrac{\partial}{\partial t} \\ -\dfrac{\partial}{\partial x} \\ -\dfrac{\partial}{\partial y} \\ -\dfrac{\partial}{\partial z} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial}{\partial t} \\ -\nabla \end{bmatrix} \tag{A.4}$$

The sign of the totally antisymmetric tensor $\varepsilon^{\mu\nu\rho\sigma}$ is taken so that

$$\varepsilon^{0123} = +1 \tag{A.5}$$

The Kronecker delta is written δ^μ_ν with the two indices aligned vertically since the row and column indices need not be distinguished.

2.A.1.2
The Lorentz Group

Linear transformations of the 4-vector

$$x^\mu \longrightarrow x'^\mu = \Lambda^\mu{}_\nu x^\nu \tag{A.6}$$

which leave the bilinear form

$$g_{\mu\nu} x^\mu y^\nu = x_\mu y^\mu = x^\mu y_\mu \tag{A.7}$$

invariant form the Lorentz Group. Since

$$\begin{aligned} g_{\mu\nu} x'^\mu y'^\nu &= g_{\mu\nu}(\Lambda^\mu{}_\rho x^\rho)(\Lambda^\nu{}_\sigma y^\sigma) \\ &= (g_{\mu\nu}\Lambda^\mu{}_\rho \Lambda^\nu{}_\sigma) x^\rho y^\sigma \\ &= g_{\rho\sigma} x^\rho y^\sigma \end{aligned} \tag{A.8}$$

we can conclude that

$$g_{\mu\nu}\Lambda^\mu{}_\rho \Lambda^\nu{}_\sigma = g_{\rho\sigma} \tag{A.9}$$

that is, the metric is an invariant tensor. Note that this equation implies $(\det \Lambda)^2 = 1$. Therefore, $\det \Lambda = \pm 1$. Acting on both sides of this equation with $g^{\tau\rho}$ and writing $g^{\tau\rho} g_{\mu\nu} \Lambda^\mu{}_\rho = \Lambda_\nu{}^\tau$, we find

$$\Lambda_\nu{}^\tau \Lambda^\nu{}_\sigma = \delta^\tau_\sigma \tag{A.10}$$

which shows that $\Lambda_\nu{}^\tau$ is the ${}^\tau{}_\nu$ element of the inverse matrix of $\Lambda^\nu{}_\sigma$:

$$(\Lambda^{-1})^\tau{}_\nu = \Lambda_\nu{}^\tau \tag{A.11}$$

The Lorentz transformation of the covariant vector is

$$x_\mu \longrightarrow x'_\mu = \Lambda_\mu{}^\nu x_\nu = x_\nu (\Lambda^{-1})^\nu{}_\mu \tag{A.12}$$

and the invariance of Eq. (A.7) is manifest.

There are two tensors that are invariant under Lorentz transformations. The first is the metric $g_{\mu\nu}$. The other is the totally antisymmetric tensor $\varepsilon^{\mu\nu\rho\sigma}$ which transforms as

$$\Lambda^\kappa{}_\rho \Lambda^\lambda{}_\sigma \Lambda^\mu{}_\tau \Lambda^\nu{}_\upsilon \, \varepsilon^{\rho\sigma\tau\upsilon} = (\det \Lambda)\, \varepsilon^{\kappa\lambda\mu\nu} \tag{A.13}$$

so it is invariant up to a sign since $\det \Lambda = \pm 1$.

2.A.1.3
Infinitesimal Lorentz Transformations

Parametrize an infinitesimal Lorentz transformation by

$$\Lambda^\mu{}_\nu = \delta^\mu_\nu + \lambda^\mu{}_\nu, \quad |\lambda^\mu{}_\nu| \ll 1 \tag{A.14}$$

The requirement that the metric is invariant, Eq. (A.9), leads to

$$\begin{aligned} g_{\rho\sigma} &= g_{\mu\nu}(\delta^\mu_\rho + \lambda^\mu{}_\rho)(\delta^\nu_\sigma + \lambda^\nu{}_\sigma) \\ &= g_{\rho\sigma} + (\lambda_{\rho\sigma} + \lambda_{\sigma\rho}) + O(\lambda^2) \end{aligned} \tag{A.15}$$

Therefore,

$$\lambda_{\mu\nu} + \lambda_{\nu\mu} = 0 \tag{A.16}$$

that is, $\lambda_{\mu\nu}$ is an antisymmetric tensor with six independent components.[8] In matrix form, we can write

$$\lambda_{\mu\nu} = \begin{bmatrix} 0 & -\eta_1 & -\eta_2 & -\eta_3 \\ \eta_1 & 0 & -\theta_3 & \theta_2 \\ \eta_2 & \theta_3 & 0 & -\theta_1 \\ \eta_3 & -\theta_2 & \theta_1 & 0 \end{bmatrix} \tag{A.17}$$

or, raising the first index,

$$\begin{aligned} \lambda^\mu{}_\nu &= g^{\mu\rho}\,\lambda_{\rho\nu} \\ &= \begin{bmatrix} 0 & -\eta_1 & -\eta_2 & -\eta_3 \\ -\eta_1 & 0 & \theta_3 & -\theta_2 \\ -\eta_2 & -\theta_3 & 0 & \theta_1 \\ -\eta_3 & \theta_2 & -\theta_1 & 0 \end{bmatrix} \end{aligned} \tag{A.18}$$

Under an infinitesimal transformation, a contravariant vector transforms as

$$\begin{aligned} x^\mu \longrightarrow x'^\mu &= (\delta^\mu_\nu + \lambda^\mu{}_\nu)x^\nu \\ &= x^\mu + \lambda^\mu{}_\nu x^\nu \end{aligned} \tag{A.19}$$

The transformations of the time and space components written separately are

$$\begin{aligned} x^0 \longrightarrow x'^0 &= x^0 + \lambda^0{}_i x^i \\ &= x^0 - \eta_i x^i \end{aligned}$$

$$\begin{aligned} x^i \longrightarrow x'^i &= x^i + \lambda^i{}_0 x^0 + \lambda^i{}_j x^j \\ &= x^i - \eta_i x^0 + \varepsilon_{ijk} x^j \theta_k \end{aligned} \tag{A.20}$$

and we can identify $\eta_i \equiv -\lambda^0{}_i = -\lambda^i{}_0$ as the infinitesimal rapidity by which the frame is boosted in the i-direction, and $\theta_i = \frac{1}{2}\varepsilon_{ijk}\lambda^j{}_k$ as the infinitesimal rotation angle by which the frame is rotated around the i-axis.

2.A.1.4
The Generators of the Lorentz Group

Define

$$iK_1 \equiv \begin{bmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$iK_2 \equiv \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$iK_3 \equiv \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}$$

$$iL_1 \equiv \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

$$iL_2 \equiv \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

$$iL_3 \equiv \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{A.21}$$

---

[8] Recall that the Fermi–Walker transport was generated by $(u^\mu a^\nu - a^\mu u^\nu)$. This is the only antisymmetric tensor that can be constructed from the 4-velocity u^μ and the 4-acceleration a^μ.

Then, the matrix $\lambda^{\mu}{}_{\nu}$ can be expressed as

$$\begin{aligned}\lambda^{\mu}{}_{\nu} &= \begin{bmatrix} 0 & -\eta_1 & -\eta_2 & -\eta_3 \\ -\eta_1 & 0 & \theta_3 & -\theta_2 \\ -\eta_2 & -\theta_3 & 0 & \theta_1 \\ -\eta_3 & \theta_2 & -\theta_1 & 0 \end{bmatrix} \\ &= \Big[\, i\theta_j L_j + i\eta_j K_j \,\Big]^{\mu}{}_{\nu} \end{aligned} \tag{A.22}$$

The six matrices listed above are the *generators* of the Lorentz group. The K_i's generate the Lorentz boosts, while the L_i's generate the rotations. They satisfy the commutation relations

$$\begin{aligned} [\,K_i,\ K_j\,] &= -i\varepsilon_{ijk}L_k \\ [\,K_i,\ L_j\,] &= \ \ i\varepsilon_{ijk}K_k \\ [\,L_i,\ L_j\,] &= \ \ i\varepsilon_{ijk}L_k \end{aligned} \tag{A.23}$$

The 4×4 matrices we used to derive these relations provide a four-dimensional representation of this algebra.

To find the generic (finite dimensional) representation of Eq. (A.23), define

$$\begin{aligned} A_i &\equiv \tfrac{1}{2}(L_i + iK_i) \\ B_i &\equiv \tfrac{1}{2}(L_i - iK_i) \end{aligned} \tag{A.24}$$

$\lambda^{\mu}{}_{\nu}$ in terms of these linear combinations is

$$\begin{aligned} i\,(\eta_i K_i + \theta_i L_i) = i(\theta_i - i\eta_i)A_i \\ + i(\theta_i + i\eta_i)B_i \end{aligned} \tag{A.25}$$

The commutation relations among the A_i's and the B_i's are

$$\begin{aligned} \big[A_i,\ A_j\big] &= i\varepsilon_{ijk}A_k \\ \big[\,B_i,\ B_j\big] &= i\varepsilon_{ijk}B_k \\ \big[A_i,\ B_j\big] &= 0 \end{aligned} \tag{A.26}$$

which shows that the six generators of the Lorentz group $O(3,1)$ split up into two disjoint sets of generators of $SU(2)$. (Or to be more precise, $SL(2,\mathbb{C})$ since the coefficients of A_i and B_i in Eq. (A.25) are complex.) Since the representation theory of $SU(2)$ is well known, it is not reviewed here. Recalling that the representations of $SU(2)$ are labeled by a half-integer j and have dimension $(2j+1)$, we can immediately see that the representations of the Lorentz group will be labeled by a pair of half-integers (j_a, j_b) where j_a labels the representation of the A_i generators and j_b labels the representation of the B_i generators. The dimension of the representation (j_a, j_b) is $(2j_a+1)(2j_b+1)$.

Note that under the parity transformation,

$$\begin{bmatrix} t \\ x \\ y \\ z \end{bmatrix} \overset{P}{\longleftrightarrow} \begin{bmatrix} t \\ -x \\ -y \\ -z \end{bmatrix} \tag{A.27}$$

the A_i's and the B_i's are interchanged:

$$A_i \overset{P}{\longleftrightarrow} B_i \tag{A.28}$$

since the boost operators, K_i, change sign but the rotation operators, L_i, do not. This means that under parity, the (j_a, j_b) representation is mapped onto the (j_b, j_a) representation. So unless $j_a = j_b$, a theory based on the (j_a, j_b) representation cannot be invariant under parity and needs to be combined with the (j_b, j_a) representation to obtain a parity invariant theory.

2.A.1.5
Examples

Let us take a look at the first few low-dimensional representations.

2.A.1.5.1 Scalars

The representation $(0, 0)$ is the trivial representation in which all the elements of the Lorentz group are mapped onto the number 1.

2.A.1.5.2 **Spinors**

The representations $(0,\frac{1}{2})$ and $(\frac{1}{2},0)$ are two-dimensional representations. They map onto each other under parity. The generators are represented by

$$\begin{aligned} &A_i \longrightarrow 0, \quad B_i \longrightarrow \frac{\sigma_i}{2}, \\ &\qquad \text{in the } (0,\tfrac{1}{2}) \text{ representation} \\ &A_i \longrightarrow \frac{\sigma_i}{2}, \quad B_i \longrightarrow 0, \\ &\qquad \text{in the } (\tfrac{1}{2},0) \text{ representation} \end{aligned} \tag{A.29}$$

where σ_i are the Pauli matrices. In terms of rotations and boosts, we have

$$\begin{aligned} L_i &= A_i + B_i \longrightarrow \frac{\sigma_i}{2} \\ iK_i &= A_i - B_i \\ &\longrightarrow \begin{cases} -\dfrac{\sigma_i}{2} & (0,\tfrac{1}{2}) \text{ representation} \\ +\dfrac{\sigma_i}{2} & (\tfrac{1}{2},0) \text{ representation} \end{cases} \end{aligned} \tag{A.30}$$

Therefore,

$$i\left(\theta_i L_i + \eta_i K_i\right) \longrightarrow \left(i\theta_i \frac{\sigma_i}{2} \mp \eta_i \frac{\sigma_i}{2}\right) \tag{A.31}$$

We denote the two-component object (called a *spinor*) which transforms under the $(0,\frac{1}{2})$ representation as

$$\varphi_\alpha = \begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix} \tag{A.32}$$

with a lowercase Greek subscript. This is not to be confused with a covariant vector. To avoid confusion, the index of spinors will be written with $\alpha, \beta, \gamma, \cdots$ and other letters in the first half of the Greek alphabet, while the index of vectors will be written with $\kappa, \lambda, \mu, \nu, \cdots$ and other letters in the latter half of the Greek alphabet.[9] The two-component object that transforms under the $(\frac{1}{2},0)$ representation is written as

$$\chi^{\dot{\alpha}} = \begin{bmatrix} \chi^{\dot{1}} \\ \chi^{\dot{2}} \end{bmatrix} \tag{A.33}$$

with a lowercase Greek superscript. This is not to be confused with a contravariant vector. Note that dots have been placed over the indices to distinguish the $(\frac{1}{2},0)$ representation from the $(0,\frac{1}{2})$ representation. The reason for using a subscript for the $(0,\frac{1}{2})$ representation and a superscript for the $(\frac{1}{2},0)$ representation will become clear later. Under a finite Lorentz transformation, spinors transform as

$$\begin{aligned} \varphi_\alpha \longrightarrow \varphi'_\alpha &= \left[\exp\left(i\theta_i \frac{\sigma_i}{2} - \eta_i \frac{\sigma_i}{2}\right)\right]_\alpha{}^\beta \varphi_\beta \\ &\equiv b_\alpha{}^\beta \, \varphi_\beta \\ \chi^{\dot{\alpha}} \longrightarrow \chi'^{\dot{\alpha}} &= \left[\exp\left(i\theta_i \frac{\sigma_i}{2} + \eta_i \frac{\sigma_i}{2}\right)\right]^{\dot{\alpha}}{}_{\dot{\beta}} \chi^{\dot{\beta}} \\ &\equiv a^{\dot{\alpha}}{}_{\dot{\beta}} \, \chi^{\dot{\beta}} \end{aligned} \tag{A.34}$$

Some special cases:

- Rotation around the unit vector **n**:

$$\exp\left(i\theta \mathbf{n} \cdot \frac{\boldsymbol{\sigma}}{2}\right) = \cos\frac{\theta}{2} + i(\mathbf{n} \cdot \boldsymbol{\sigma}) \sin\frac{\theta}{2} \tag{A.35}$$

- Boost in the direction of the unit vector **n**:

$$\begin{aligned} \exp\left(\mp \eta \mathbf{n} \cdot \frac{\boldsymbol{\sigma}}{2}\right) &= \cosh\frac{\eta}{2} \\ &\quad \mp (\mathbf{n} \cdot \boldsymbol{\sigma}) \sinh\frac{\eta}{2} \end{aligned} \tag{A.36}$$

---

[9] The order of the letters in the Greek alphabet is as follows: $\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \eta, \theta, \iota, \kappa, \lambda, \mu, \nu, \xi, o, \pi, \rho, \sigma, \tau, \upsilon, \phi, \chi, \psi, \omega$.

Note that the rotation matrices are unitary but the boost matrices are not. However, they both have determinant 1. In general, since

$$i\theta_i \frac{\sigma_i}{2} \mp \eta_i \frac{\sigma_i}{2} \tag{A.37}$$

is a traceless complex matrix, its exponential

$$\exp\left(i\theta_i \frac{\sigma_i}{2} \mp \eta_i \frac{\sigma_i}{2}\right) \tag{A.38}$$

is always an $SL(2,\mathbb{C})$ matrix with determinant 1. Consequently, the expressions

$$\begin{aligned} \varepsilon^{\alpha\beta}\varphi_\alpha\psi_\beta &= \varphi_1\psi_2 - \varphi_2\psi_1 \\ \varepsilon_{\dot\alpha\dot\beta}\chi^{\dot\alpha}\xi^{\dot\beta} &= \chi^{\dot1}\xi^{\dot2} - \chi^{\dot2}\xi^{\dot1} \end{aligned} \tag{A.39}$$

are invariant under Lorentz transformations since

$$\begin{aligned} \varepsilon^{\alpha\beta}\varphi_\alpha\psi_\beta &\longrightarrow \varepsilon^{\alpha\beta}(b_\alpha{}^\gamma\varphi_\gamma)(b_\beta{}^\delta\psi_\delta) \\ &= (\det b)\,\varepsilon^{\gamma\delta}\varphi_\gamma\psi_\delta = \varepsilon^{\gamma\delta}\varphi_\gamma\psi_\delta \\ \varepsilon_{\dot\alpha\dot\beta}\chi^{\dot\alpha}\xi^{\dot\beta} &\longrightarrow \varepsilon_{\dot\alpha\dot\beta}(a^{\dot\alpha}{}_{\dot\gamma}\chi^{\dot\gamma})(a^{\dot\beta}{}_{\dot\delta}\xi^{\dot\delta}) \\ &= (\det a)\,\varepsilon_{\dot\gamma\dot\delta}\chi^{\dot\gamma}\xi^{\dot\delta} = \varepsilon_{\dot\gamma\dot\delta}\chi^{\dot\gamma}\xi^{\dot\delta} \end{aligned} \tag{A.40}$$

Note that we define

$$\varepsilon^{\alpha\beta} = \varepsilon_{\alpha\beta} = \varepsilon^{\dot\alpha\dot\beta} = \varepsilon_{\dot\alpha\dot\beta} = i\sigma_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \tag{A.41}$$

but many authors use a different convention, so care is necessary when comparing formulae.

Define

$$\begin{aligned} \varphi^\alpha &= \begin{bmatrix} \varphi^1 \\ \varphi^2 \end{bmatrix} \equiv \varepsilon^{\alpha\beta}\varphi_\beta = -\varphi_\beta\varepsilon^{\beta\alpha} \\ &= [i\sigma_2]^{\alpha\beta}\varphi_\beta = \begin{bmatrix} \varphi_2 \\ -\varphi_1 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \chi_{\dot\alpha} &= \begin{bmatrix} \chi_{\dot1} \\ \chi_{\dot2} \end{bmatrix} \equiv -\varepsilon_{\dot\alpha\dot\beta}\chi^{\dot\beta} = \chi^{\dot\beta}\varepsilon_{\dot\beta\dot\alpha} \\ &= [-i\sigma_2]_{\dot\alpha\dot\beta}\chi^{\dot\beta} = \begin{bmatrix} -\chi^{\dot2} \\ \chi^{\dot1} \end{bmatrix} \end{aligned} \tag{A.42}$$

Since $[i\sigma_2]^2 = -1$, we also have the relations

$$\begin{aligned} \varphi_\alpha &= -\varepsilon_{\alpha\beta}\varphi^\beta = \varphi^\beta\varepsilon_{\beta\alpha}, \\ \chi^{\dot\alpha} &= \varepsilon^{\dot\alpha\dot\beta}\chi_{\dot\beta} = -\chi_{\dot\beta}\varepsilon^{\dot\beta\dot\alpha} \end{aligned} \tag{A.43}$$

This allows us to write the invariants compactly as

$$\begin{aligned} \varepsilon^{\alpha\beta}\varphi_\alpha\psi_\beta &= \varphi_\alpha\psi^\alpha = -\varphi^\alpha\psi_\alpha \\ \varepsilon_{\dot\alpha\dot\beta}\chi^{\dot\alpha}\xi^{\dot\beta} &= \chi_{\dot\alpha}\xi^{\dot\alpha} = -\chi^{\dot\alpha}\xi_{\dot\alpha} \end{aligned} \tag{A.44}$$

Note that if the components of the spinors were just ordinary commuting numbers, then

$$\begin{aligned} \varphi_\alpha\varphi^\alpha &= \varphi_1\varphi_2 - \varphi_2\varphi_1 = 0 \\ \chi_{\dot\alpha}\chi^{\dot\alpha} &= \chi^{\dot1}\chi^{\dot2} - \chi^{\dot2}\chi^{\dot1} = 0 \end{aligned} \tag{A.45}$$

However, if they were anticommuting, which would be the case if they were Grassman numbers or Fermionic field operators, then we would have

$$\begin{aligned} \varphi_\alpha\varphi^\alpha &= \varphi_1\varphi_2 - \varphi_2\varphi_1 = 2\varphi_1\varphi_2 \neq 0 \\ \chi_{\dot\alpha}\chi^{\dot\alpha} &= \chi^{\dot1}\chi^{\dot2} - \chi^{\dot2}\chi^{\dot1} = 2\chi^{\dot1}\chi^{\dot2} \neq 0 \end{aligned} \tag{A.46}$$

It is straightforward to show that φ^α and $\chi_{\dot\alpha}$ transform as

$$\begin{aligned} \varphi^\alpha &\longrightarrow \varphi'^\alpha = (a^*)^\alpha{}_\delta\,\varphi^\delta = \varphi^\delta\,(a^\dagger)_\delta{}^\alpha \\ \chi_{\dot\alpha} &\longrightarrow \chi'_{\dot\alpha} = (b^*)_{\dot\alpha}{}^{\dot\delta}\,\chi_{\dot\delta} = \chi_{\dot\delta}\,(b^\dagger)^{\dot\delta}{}_{\dot\alpha} \end{aligned} \tag{A.47}$$

where the matrices a and b were defined in Eq. (A.34), and the asterisk denotes complex conjugation. This tells us that $\chi_{\dot\alpha}$ transforms like $(\varphi_\alpha)^*$, while φ^α transforms like $(\chi^{\dot{\alpha}})^*$. So complex conjugation either puts on or takes off the dot from the spinor index. (The positioning of the spinor indices was carefully chosen so that this would be the case.) Accordingly, we write

$$\overline{\varphi}_{\dot{\alpha}} \equiv (\varphi_\alpha)^*, \qquad \overline{\chi}^{\alpha} \equiv (\chi^{\dot{\alpha}})^* \tag{A.48}$$

The matrices a and b are related by

$$a^\dagger = b^{-1}, \qquad b^\dagger = a^{-1} \tag{A.49}$$

so that

$$\begin{aligned} \varphi^\alpha \varphi_\alpha &\longrightarrow \varphi'^\alpha \varphi'_\alpha = \varphi^\beta (a^\dagger)_\beta{}^\alpha (b)_\alpha{}^\gamma \varphi_\gamma \\ &= \varphi^\beta (b^{-1})_\beta{}^\alpha (b)_\alpha{}^\gamma \varphi_\gamma = \varphi^\beta \varphi_\beta \\ \chi_{\dot{\alpha}} \chi^{\dot{\alpha}} &\longrightarrow \chi'_{\dot{\alpha}} \chi'^{\dot{\alpha}} = \chi_{\dot{\beta}} (b^\dagger)^{\dot{\beta}}{}_{\dot{\alpha}} (a)^{\dot{\alpha}}{}_{\dot{\gamma}} \chi^{\dot{\gamma}} \\ &= \chi_{\dot{\beta}} (a^{-1})^{\dot{\beta}}{}_{\dot{\alpha}} (a)^{\dot{\alpha}}{}_{\dot{\gamma}} \chi^{\dot{\gamma}} = \chi_{\dot{\beta}} \chi^{\dot{\beta}} \end{aligned} \tag{A.50}$$

2.A.1.5.3 **Vectors**

The representation $(\frac{1}{2}, \frac{1}{2})$ is a four-dimensional representation. It can be constructed by tensoring the $(0, \frac{1}{2})$ and $(\frac{1}{2}, 0)$ representations:

$$\left(0, \tfrac{1}{2}\right) \otimes \left(\tfrac{1}{2}, 0\right) = \left(\tfrac{1}{2}, \tfrac{1}{2}\right) \tag{A.51}$$

Objects that transform under this representation have two spinor indices, one dotted and the other undotted:

$$\varphi_\alpha{}^{\dot{\beta}} = \begin{bmatrix} \varphi_1{}^{\dot{1}} & \varphi_1{}^{\dot{2}} \\ \varphi_2{}^{\dot{1}} & \varphi_2{}^{\dot{2}} \end{bmatrix} \tag{A.52}$$

It is convenient to lower the dotted index with $\varepsilon_{\dot{\alpha}\dot{\beta}}$ and deal with

$$\begin{aligned} \varphi_{\alpha\dot{\beta}} = \varphi_\alpha{}^{\dot{\gamma}} \varepsilon_{\dot{\gamma}\dot{\beta}} &= \begin{bmatrix} -\varphi_1{}^{\dot{2}} & \varphi_1{}^{\dot{1}} \\ -\varphi_2{}^{\dot{2}} & \varphi_2{}^{\dot{1}} \end{bmatrix} \\ &= \begin{bmatrix} \varphi_{1\dot{1}} & \varphi_{1\dot{2}} \\ \varphi_{2\dot{1}} & \varphi_{2\dot{2}} \end{bmatrix} \end{aligned} \tag{A.53}$$

$\varphi_{\alpha\dot{\beta}}$ transforms as

$$\begin{aligned} \varphi_{\alpha\dot{\beta}} \longrightarrow \varphi'_{\alpha\dot{\beta}} &= b_\alpha{}^\gamma (b^*)_{\dot{\beta}}{}^{\dot{\delta}} \varphi_{\gamma\dot{\delta}} \\ &= b_\alpha{}^\gamma \varphi_{\gamma\dot{\delta}} (b^\dagger)^{\dot{\delta}}{}_{\dot{\beta}} \end{aligned} \tag{A.54}$$

In matrix notation omitting the indices,

$$\varphi \longrightarrow \varphi' = b\,\varphi\,b^\dagger \tag{A.55}$$

which implies

$$\varphi^\dagger \longrightarrow \varphi'^\dagger = b\,\varphi^\dagger\,b^\dagger \tag{A.56}$$

This means that the Hermitian and anti-Hermitian parts of φ do not mix under this transformation:

$$\begin{aligned} (\varphi + \varphi^\dagger) &\longrightarrow (\varphi' + \varphi'^\dagger) = b(\varphi + \varphi^\dagger) b^\dagger \\ i(\varphi - \varphi^\dagger) &\longrightarrow i(\varphi' - \varphi'^\dagger) = b\,i(\varphi - \varphi^\dagger) b^\dagger \end{aligned} \tag{A.57}$$

Therefore, they must be treated as independently transforming Hermitian matrices. So without loss of generality, we can assume $\varphi_{\alpha\dot{\beta}}$ to be Hermitian.

Since $\det b = \det b^\dagger = 1$, the transformation $\varphi \longrightarrow b\varphi b^\dagger$ leaves the determinant of the Hermitian matrix $\varphi_{\alpha\dot{\beta}}$ invariant. Expanding $\varphi_{\alpha\dot{\beta}}$ using the Pauli matrices,

$$\begin{aligned} \varphi_{\alpha\dot{\beta}} &= \begin{bmatrix} \varphi_{1\dot{1}} & \varphi_{1\dot{2}} \\ \varphi_{2\dot{1}} & \varphi_{2\dot{2}} \end{bmatrix} \\ &= \begin{bmatrix} V^0 + V^3 & V^1 - iV^2 \\ V^1 + iV^2 & V^0 - V^3 \end{bmatrix} \\ &= V^\mu (\sigma_\mu)_{\alpha\dot{\beta}} \end{aligned} \tag{A.58}$$

where

$$(\sigma_0)_{\alpha\dot{\beta}} = \mathbf{1}_{2\times 2}, \quad (\sigma_i)_{\alpha\dot{\beta}} = \sigma_i \tag{A.59}$$

and the coefficients V^μ are all real, we find

$$\begin{aligned} \det \varphi_{\alpha\dot{\beta}} &= (V^0)^2 - (V^1)^2 - (V^2)^2 - (V^3)^2 \\ &= V_\mu V^\mu \end{aligned} \tag{A.60}$$

Thus, the invariant looks like the norm of the vector V^μ, suggesting that the transformation of $\varphi_{\alpha\dot\beta} = V^\mu(\sigma_\mu)_{\alpha\dot\beta}$ is equivalent to the transformation of a vector V^μ. Indeed, under an infinitesimal Lorentz transformation

$$\begin{aligned} b &= \exp\left(+i\theta_i\frac{\sigma_i}{2} - \eta_i\frac{\sigma_i}{2}\right) \\ &= \mathbf{1} + \left(i\theta_i\frac{\sigma_i}{2} - \eta_i\frac{\sigma_i}{2}\right) + \cdots \end{aligned} \tag{A.61}$$

we find

$$\begin{aligned} b\,\sigma_\mu\,b^\dagger &= \left[\mathbf{1} + \left(i\theta_i\frac{\sigma_i}{2} - \eta_i\frac{\sigma_i}{2}\right) + \cdots\right]\sigma_\mu \\ &\quad\times\left[\mathbf{1} + \left(-i\theta_i\frac{\sigma_i}{2} - \eta_i\frac{\sigma_i}{2}\right) + \cdots\right] \\ &= \sigma_\mu + \left(i\frac{\theta_i}{2}\left[\sigma_i, \sigma_\mu\right] - \frac{\eta_i}{2}\left\{\sigma_i, \sigma_\mu\right\}\right) \\ &\quad + \cdots \end{aligned} \tag{A.62}$$

The $\mu = 0$ and $\mu = j$ cases are

$$\begin{aligned} b\,\sigma_0\,b^\dagger &= \sigma_0 + \left(i\frac{\theta_i}{2}\left[\sigma_i, \sigma_0\right] - \frac{\eta_i}{2}\left\{\sigma_i, \sigma_0\right\}\right) \\ &\quad + \cdots = \sigma_0 - \eta_i\sigma_i + \cdots \\ b\,\sigma_j\,b^\dagger &= \sigma_j + \left(i\frac{\theta_i}{2}\left[\sigma_i, \sigma_j\right] - \frac{\eta_i}{2}\left\{\sigma_i, \sigma_j\right\}\right) \\ &\quad + \cdots = \sigma_j - \theta_i\varepsilon_{ijk}\sigma_k - \eta_j\sigma_0 + \cdots \end{aligned} \tag{A.63}$$

Therefore,

$$\begin{aligned} V^\mu\sigma_\mu &\longrightarrow V'^\mu\sigma_\mu \\ &= V^\mu(b\,\sigma_\mu\,b^\dagger) \\ &= (V^0 - \eta_j V^j)\sigma_0 \\ &\quad + (V^k - \eta_k V^0 - \theta_i\varepsilon_{ijk}V^j)\sigma_k \end{aligned} \tag{A.64}$$

which shows that

$$\begin{aligned} V^0 &\longrightarrow V'^0 = V^0 - \eta_j V^j \\ V^k &\longrightarrow V'^k = V^k - \eta_k V^0 + \theta_i\varepsilon_{kji}V^j \end{aligned} \tag{A.65}$$

This is precisely the transformation law of contravariant vectors.

Take the complex conjugate of $\varphi_{\alpha\dot\beta}$ and call it $\overline{\varphi}_{\dot\alpha\beta}$:

$$\overline{\varphi}_{\dot\alpha\beta} = (\varphi_{\alpha\dot\beta})^* = V^\mu((\sigma_\mu)_{\alpha\dot\beta})^* = V^\mu(\sigma_\mu^*)_{\dot\alpha\beta} \tag{A.66}$$

Raise the indices with the ε's:

$$\begin{aligned} \overline{\varphi}^{\dot\gamma\delta} &= (\varepsilon^{\dot\gamma\dot\alpha})\,\overline{\varphi}_{\dot\alpha\beta}\,(-\varepsilon^{\beta\delta}) = V^\mu(\sigma_2\sigma_\mu^*\sigma_2)^{\dot\gamma\delta} \\ &\equiv V^\mu(\overline{\sigma}_\mu)^{\dot\gamma\delta} \end{aligned} \tag{A.67}$$

where we have defined

$$(\overline{\sigma}_0)^{\dot\alpha\beta} = \mathbf{1}_{2\times 2}, \quad (\overline{\sigma}_i)^{\dot\alpha\beta} = -\sigma_i \tag{A.68}$$

This object transforms as

$$\begin{aligned} \overline{\varphi}^{\dot\alpha\beta} \longrightarrow \overline{\varphi}'^{\dot\alpha\beta} &= a^{\dot\alpha}{}_{\dot\gamma}\,(a^*)^\beta{}_\delta\,\overline{\varphi}^{\dot\gamma\delta} \\ &= a^{\dot\alpha}{}_{\dot\gamma}\,\overline{\varphi}^{\dot\gamma\delta}\,(a^\dagger)_\delta{}^\beta \end{aligned} \tag{A.69}$$

or in matrix notation omitting indices,

$$\overline{\varphi} \longrightarrow \overline{\varphi}' = a\,\overline{\varphi}\,a^\dagger \tag{A.70}$$

The matrices $(\sigma_\mu)_{\alpha\dot\beta}$ and $(\overline{\sigma}_\mu)^{\dot\alpha\beta}$ defined above satisfy the following relations:

$$\begin{aligned} (\sigma_\mu)_{\alpha\dot\beta}(\overline{\sigma}_\nu)^{\dot\beta\gamma} + (\sigma_\nu)_{\alpha\dot\beta}(\overline{\sigma}_\mu)^{\dot\beta\gamma} &= 2g_{\mu\nu}\delta_\alpha^\gamma \\ (\overline{\sigma}_\mu)^{\dot\alpha\beta}(\sigma_\nu)_{\beta\dot\gamma} + (\overline{\sigma}_\nu)^{\dot\alpha\beta}(\sigma_\mu)_{\beta\dot\gamma} &= 2g_{\mu\nu}\delta_{\dot\gamma}^{\dot\alpha} \\ (\sigma_\mu)_{\alpha\dot\beta}(\overline{\sigma}_\nu)^{\dot\beta\alpha} = (\overline{\sigma}_\mu)^{\dot\beta\alpha}(\sigma_\nu)_{\alpha\dot\beta} &= 2g_{\mu\nu} \\ (\sigma_\mu)_{\alpha\dot\beta}(\overline{\sigma}^\mu)^{\dot\gamma\delta} &= 2\delta_\alpha^\delta\delta_{\dot\beta}^{\dot\gamma} \end{aligned} \tag{A.71}$$

Omitting indices, we can write the first three relations as

$$\begin{aligned} \sigma_\mu\,\overline{\sigma}_\nu + \sigma_\nu\,\overline{\sigma}_\mu &= 2g_{\mu\nu} \\ \overline{\sigma}_\mu\,\sigma_\nu + \overline{\sigma}_\nu\,\sigma_\mu &= 2g_{\mu\nu} \\ \mathrm{tr}(\overline{\sigma}_\mu\sigma_\nu) = \mathrm{tr}(\sigma_\mu\overline{\sigma}_\nu) &= 2g_{\mu\nu} \end{aligned} \tag{A.72}$$

Using the trace relation on

$$\varphi_{\alpha\dot\beta} = V^\mu(\sigma_\mu)_{\alpha\dot\beta}, \qquad \overline{\varphi}^{\dot\alpha\beta} = V^\mu(\overline{\sigma}_\mu)^{\dot\alpha\beta} \tag{A.73}$$

we find

$$V^{\mu} = \tfrac{1}{2}\,\varphi_{\alpha\dot{\beta}}(\overline{\sigma}^{\mu})^{\dot{\beta}\alpha} = \tfrac{1}{2}\,\overline{\varphi}^{\dot{\alpha}\beta}(\sigma^{\mu})_{\beta\dot{\alpha}} \quad \text{(A.74)}$$

This relation implies that objects like

$$\chi_{\dot{\beta}}(\overline{\sigma}^{\mu})^{\dot{\beta}\alpha}\varphi_{\alpha}, \quad \text{or} \quad \xi^{\beta}(\sigma^{\mu})_{\beta\dot{\alpha}}\zeta^{\dot{\alpha}} \quad \text{(A.75)}$$

transform as vectors. Indeed,

$$\begin{aligned} \chi_{\dot{\beta}}(\overline{\sigma}^{\mu})^{\dot{\beta}\alpha}\varphi_{\alpha} &\longrightarrow \chi'_{\dot{\beta}}(\overline{\sigma}^{\mu})^{\dot{\beta}\alpha}\varphi'_{\alpha} \\ &= \chi_{\dot{\delta}}(b^{\dagger})^{\dot{\delta}}{}_{\dot{\beta}}(\overline{\sigma}^{\mu})^{\dot{\beta}\alpha}b_{\alpha}{}^{\gamma}\varphi_{\gamma} \\ \xi^{\beta}(\sigma^{\mu})_{\beta\dot{\alpha}}\zeta^{\dot{\alpha}} &\longrightarrow \xi'^{\beta}(\sigma^{\mu})_{\beta\dot{\alpha}}\zeta'^{\dot{\alpha}} \\ &= \xi^{\delta}(a^{\dagger})_{\delta}{}^{\beta}(\sigma^{\mu})_{\beta\dot{\alpha}}a^{\dot{\alpha}}{}_{\dot{\gamma}}\zeta^{\dot{\gamma}} \end{aligned} \quad \text{(A.76)}$$

and we can readily check that

$$\begin{aligned} b^{\dagger}\,\overline{\sigma}^{0}\,b &= \overline{\sigma}^{0} - \eta_i\overline{\sigma}^{i} \\ b^{\dagger}\,\overline{\sigma}^{j}\,b &= \overline{\sigma}^{j} - \eta_j\overline{\sigma}^{0} + \theta_i\varepsilon_{ijk}\overline{\sigma}^{k} \\ a^{\dagger}\,\sigma^{0}\,a &= \sigma^{0} - \eta_i\sigma^{i} \\ a^{\dagger}\,\sigma^{j}\,a &= \sigma^{j} - \eta_j\sigma^{0} + \theta_i\varepsilon_{ijk}\sigma^{k} \end{aligned} \quad \text{(A.77)}$$

2.A.1.5.4 **Antisymmetric Tensors**

The $(1,0)$ representation can be obtained from the symmetric tensor-product of two $(\frac{1}{2},0)$ representations:

$$\left(\tfrac{1}{2},0\right)\otimes\left(\tfrac{1}{2},0\right) = (0,0)\oplus(1,0) \quad \text{(A.78)}$$

(The antisymmetric part is just a scalar.) Similarly, the $(0,1)$ representation can be obtained from the symmetric tensor-product of two $(0,\frac{1}{2})$ representations:

$$\left(0,\tfrac{1}{2}\right)\otimes\left(0,\tfrac{1}{2}\right) = (0,0)\oplus(0,1) \quad \text{(A.79)}$$

Let us consider the $(0,1)$ representation first. Objects that transform under this representation have two undotted indices, $\varphi_{(\alpha\beta)}$, and are symmetric under the interchange of the two indices:

$$\varphi_{(\alpha\beta)} = \varphi_{(\beta\alpha)} = \begin{bmatrix} \varphi_{(11)} & \varphi_{(12)} \\ \varphi_{(12)} & \varphi_{(22)} \end{bmatrix} \quad \text{(A.80)}$$

As in the $(\frac{1}{2},\frac{1}{2})$ case, it is more convenient to raise one of the indices with $\varepsilon^{\alpha\beta}$ and work with

$$\begin{aligned} \varphi_{\alpha}{}^{\beta} &= \varphi_{\alpha\gamma}(-\varepsilon^{\gamma\beta}) \\ &= \begin{bmatrix} \varphi_{(12)} & -\varphi_{(11)} \\ \varphi_{(22)} & -\varphi_{(12)} \end{bmatrix} \\ &= \begin{bmatrix} \varphi_{1}{}^{1} & \varphi_{1}{}^{2} \\ \varphi_{2}{}^{1} & \varphi_{2}{}^{2} \end{bmatrix} \end{aligned} \quad \text{(A.81)}$$

Note that $\varphi_{\alpha}{}^{\beta}$ is traceless since $\varphi_{1}{}^{1} + \varphi_{2}{}^{2} = \varphi_{(12)} - \varphi_{(12)} = 0$. It transforms as

$$\begin{aligned} \varphi_{\alpha}{}^{\beta} \longrightarrow \varphi'_{\alpha}{}^{\beta} &= b_{\alpha}{}^{\gamma}(a^{*})^{\beta}{}_{\delta}\,\varphi_{\gamma}{}^{\delta} \\ &= b_{\alpha}{}^{\gamma}\,\varphi_{\gamma}{}^{\delta}(a^{\dagger})_{\delta}{}^{\beta} = b_{\alpha}{}^{\gamma}\,\varphi_{\gamma}{}^{\delta}(b^{-1})_{\delta}{}^{\beta} \end{aligned} \quad \text{(A.82)}$$

or

$$\varphi \longrightarrow \varphi' = b\,\varphi\,b^{-1} \quad \text{(A.83)}$$

in matrix form, which shows that

$$\mathrm{Tr}(\varphi') = \mathrm{Tr}(b\,\varphi\,b^{-1}) = \mathrm{Tr}(\varphi) = 0 \quad \text{(A.84)}$$

that is, the tracelessness is conserved. Recalling that

$$\begin{aligned} b &= \exp\left[\ \ i\theta_i\frac{\sigma_i}{2} + \eta_i\frac{\sigma_i}{2}\right] \\ b^{-1} &= \exp\left[-i\theta_i\frac{\sigma_i}{2} - \eta_i\frac{\sigma_i}{2}\right] \end{aligned} \quad \text{(A.85)}$$

notice that if we restrict our attention to rotations only ($\vec{\theta} \neq 0, \vec{\eta} = 0$), the b matrix is unitary: $b^{-1} = b^{\dagger}$. Therefore, just as in the $(\frac{1}{2},\frac{1}{2})$ case, the Hermitian part and anti-Hermitian part of $\varphi_{\alpha}{}^{\beta}$ transform among themselves under rotations. Define

$$\begin{aligned} E_i\,\sigma_i &= \begin{bmatrix} E_3 & E_1 - iE_2 \\ E_1 + iE_2 & -E_3 \end{bmatrix} \\ &\equiv \tfrac{1}{2}\left(\varphi + \varphi^{\dagger}\right) \end{aligned}$$

$$\begin{aligned} B_i\,\sigma_i &= \begin{bmatrix} B_3 & B_1 - iB_2 \\ B_1 + iB_2 & -B_3 \end{bmatrix} \\ &\equiv \frac{i}{2}\left(\varphi - \varphi^\dagger\right) \end{aligned} \tag{A.86}$$

so that

$$\varphi = (E_i - iB_i)\,\sigma_i, \quad \varphi^\dagger = (E_i + iB_i)\,\sigma_i \tag{A.87}$$

Then, just as in the $(\frac{1}{2}, \frac{1}{2})$ case, E_i and B_i transform as 3-vectors under rotations. To see how they transform under boosts, we need to figure out how the Pauli matrices transform. Under infinitesimal boosts, we find

$$\begin{aligned} b\,\sigma_i\,b^{-1} &= \left(\mathbf{1} + \eta_j \frac{\sigma_j}{2} + \cdots\right) \\ &\quad \times \sigma_i \left(\mathbf{1} - \eta_j \frac{\sigma_j}{2} + \cdots\right) \\ &= \sigma_i - i\varepsilon_{ijk}\,\eta_j\,\sigma_k + \cdots \end{aligned} \tag{A.88}$$

Therefore,

$$\begin{aligned} (E_i - iB_i)\,\sigma_i &\longrightarrow (E_i' - iB_i')\,\sigma_i \\ &= b\left[\,(E_i - iB_i)\,\sigma_i\,\right]b^{-1} \\ &= \Big[\big\{\mathbf{E} + (\eta \times \mathbf{B})\big\}_k \\ &\qquad - i\big\{\mathbf{B} - (\eta \times \mathbf{E})\big\}_k\Big]\sigma_k + \cdots \end{aligned} \tag{A.89}$$

and

$$\begin{aligned} \mathbf{E} &\longrightarrow \mathbf{E}' = \mathbf{E} + \boldsymbol{\eta} \times \mathbf{B} + \cdots \\ \mathbf{B} &\longrightarrow \mathbf{B}' = \mathbf{B} - \boldsymbol{\eta} \times \mathbf{E} + \cdots \end{aligned} \tag{A.90}$$

which is precisely how electric and magnetic fields transform under infinitesimal boosts, Eq. (116).

Next, take the complex conjugate of $\varphi_\alpha{}^\beta$ and raise the left index with $\varepsilon^{\dot\alpha\dot\beta}$:

$$\begin{aligned} \overline{\varphi}^{(\dot\alpha\dot\beta)} &\equiv \varepsilon^{\dot\alpha\dot\gamma}\,\overline{\varphi}_{\dot\gamma}{}^{\dot\beta} = \varepsilon^{\dot\alpha\dot\gamma}\,(\varphi_\alpha{}^\beta)^* \\ &= \begin{bmatrix} (\varphi_2{}^1)^* & (\varphi_2{}^2)^* \\ -(\varphi_1{}^1)^* & (\varphi_1{}^2)^* \end{bmatrix} \end{aligned} \tag{A.91}$$

Recalling that $\varphi_1{}^1 + \varphi_2{}^2 = 0$, we can see that $\overline{\varphi}^{(\dot\alpha\dot\beta)} = \overline{\varphi}^{(\dot\beta\dot\alpha)}$. Therefore, this object transforms under the $(1, 0)$ representation. As in the $(\frac{1}{2}, \frac{1}{2})$, and $(0, 1)$ cases, lowering the right index with $\varepsilon_{\dot\alpha\dot\beta}$ converts it into a traceless matrix, which makes the analysis easier:

$$\begin{aligned} \overline{\varphi}^{\dot\alpha}{}_{\dot\beta} &= \begin{bmatrix} \overline{\varphi}^{\dot 1}{}_{\dot 1} & \overline{\varphi}^{\dot 1}{}_{\dot 2} \\ \overline{\varphi}^{\dot 2}{}_{\dot 1} & \overline{\varphi}^{\dot 2}{}_{\dot 2} \end{bmatrix} = \overline{\varphi}^{\dot\alpha\dot\gamma}\,\varepsilon_{\dot\gamma\dot\beta} \\ &= \begin{bmatrix} -(\varphi_2{}^2)^* & (\varphi_2{}^1)^* \\ -(\varphi_1{}^2)^* & -(\varphi_1{}^1)^* \end{bmatrix} \end{aligned} \tag{A.92}$$

This objects transformations as

$$\begin{aligned} \overline{\varphi}^{\dot\alpha}{}_{\dot\beta} &\longrightarrow \overline{\varphi}'^{\dot\alpha}{}_{\dot\beta} = a^{\dot\alpha}{}_{\dot\gamma}\,(b^*)_{\dot\beta}{}^{\dot\delta}\,\overline{\varphi}^{\dot\gamma}{}_{\dot\delta} \\ &= a^{\dot\alpha}{}_{\dot\gamma}\,\overline{\varphi}^{\dot\gamma}{}_{\dot\delta}\,(b^\dagger)^{\dot\delta}{}_{\dot\beta} = a^{\dot\alpha}{}_{\dot\gamma}\,\overline{\varphi}^{\dot\gamma}{}_{\dot\delta}\,(a^{-1})^{\dot\delta}{}_{\dot\beta} \end{aligned} \tag{A.93}$$

In matrix form, this is

$$\overline{\varphi} \longrightarrow a\,\overline{\varphi}\,a^{-1} \tag{A.94}$$

We can continue the analysis as we did for φ, but recalling that

$$\varphi = (\mathbf{E} - i\mathbf{B}) \cdot \boldsymbol{\sigma} \tag{A.95}$$

we find

$$\overline{\varphi} = (i\sigma_2)\,\varphi^*\,(i\sigma_2) = (\mathbf{E} + i\mathbf{B}) \cdot \boldsymbol{\sigma} \tag{A.96}$$

which tells us that $\overline{\varphi}$ transforms like $(\mathbf{E} + i\mathbf{B})$. So the $(0, 1)$ and $(1, 0)$ representations combine to give objects that transform like the **E** and **B** fields, that is, like the antisymmetric tensor $F^{\mu\nu}$.

Higher dimensional representations can be constructed in a similar fashion by further tensoring the spinor representations to the ones we already have.

2.A.2
The Dirac Equation

In order to have a Lorentz invariant field theory, the Lagrangian density of the field must be a Lorentz scalar. In this section, we show how one can construct such Lagrangian densities for spinor fields, and derive the Dirac equation and some of its consequences. For the sake of simplicity, $\hbar$ is set equal to 1.

2.A.2.1
Weyl Spinors with Majorana Masses

Consider the $(0, \frac{1}{2})$ spinor field $\varphi_\alpha(x)$. The bilinears that can be constructed from φ_α and its complex (Hermitian) conjugate, $\overline{\varphi}_{\dot\alpha} = (\varphi_\alpha)^\dagger$ (we write $\dagger$ instead of $*$ here since the fields will be interpreted as operators in quantum field theory), are the scalars:

$$\begin{aligned}\varphi^\alpha\varphi_\alpha &= -\varphi_\alpha\varphi^\alpha = -\varphi_\alpha\varepsilon^{\alpha\beta}\varphi_\beta\\ \overline{\varphi}^{\dot\alpha}\overline{\varphi}_{\dot\alpha} &= -\overline{\varphi}_{\dot\alpha}\overline{\varphi}^{\dot\alpha} = -\overline{\varphi}_{\dot\alpha}\varepsilon^{\dot\alpha\dot\beta}\overline{\varphi}_{\dot\beta}\end{aligned} \tag{A.97}$$

(we assume φ_α and $\overline{\varphi}_{\dot\alpha}$ to be Grassmanian/Fermionic so these expressions do not vanish) and the vector:

$$\overline{\varphi}_{\dot\alpha}(\overline{\sigma}^\mu)^{\dot\alpha\beta}\varphi_\beta \tag{A.98}$$

from which we can construct the scalars

$$\overline{\varphi}_{\dot\alpha}(\overline{\sigma}^\mu)^{\dot\alpha\beta}(\partial_\mu\varphi_\beta), \quad \text{and} \quad (\partial_\mu\overline{\varphi}_{\dot\alpha})(\overline{\sigma}^\mu)^{\dot\alpha\beta}\varphi_\beta \tag{A.99}$$

Note that

$$\begin{aligned}\left[\varphi_\alpha\varepsilon^{\alpha\beta}\varphi_\beta\right]^\dagger &= -\overline{\varphi}_{\dot\alpha}\varepsilon^{\dot\alpha\dot\beta}\overline{\varphi}_{\dot\beta}\\ \left[\overline{\varphi}_{\dot\alpha}\varepsilon^{\dot\alpha\dot\beta}\overline{\varphi}_{\dot\beta}\right]^\dagger &= -\varphi_\alpha\varepsilon^{\alpha\beta}\varphi_\beta\\ \left[\overline{\varphi}_{\dot\alpha}(\overline{\sigma}^\mu)^{\dot\alpha\beta}(\partial_\mu\varphi_\beta)\right]^\dagger &= (\partial_\mu\overline{\varphi}_{\dot\alpha})(\overline{\sigma}^\mu)^{\dot\alpha\beta}\varphi_\beta\\ \left[(\partial_\mu\overline{\varphi}_{\dot\alpha})(\overline{\sigma}^\mu)^{\dot\alpha\beta}\varphi_\beta\right]^\dagger &= \overline{\varphi}_{\dot\alpha}(\overline{\sigma}^\mu)^{\dot\alpha\beta}(\partial_\mu\varphi_\beta)\end{aligned} \tag{A.100}$$

Therefore, the Hermitian combinations are

$$\begin{aligned}&\tfrac{1}{2}\left[\overline{\varphi}_{\dot\alpha}(\overline{\sigma}^\mu)^{\dot\alpha\beta}(\partial_\mu\varphi_\beta) + (\partial_\mu\overline{\varphi}_{\dot\alpha})(\overline{\sigma}^\mu)^{\dot\alpha\beta}\varphi_\beta\right]\\ &\quad = \tfrac{1}{2}\partial_\mu\left[\overline{\varphi}_{\dot\alpha}(\overline{\sigma}^\mu)^{\dot\alpha\beta}\varphi_\beta\right] = \text{total derivative}\\ &\tfrac{i}{2}\left[\overline{\varphi}_{\dot\alpha}(\overline{\sigma}^\mu)^{\dot\alpha\beta}(\partial_\mu\varphi_\beta) - (\partial_\mu\overline{\varphi}_{\dot\alpha})(\overline{\sigma}^\mu)^{\dot\alpha\beta}\varphi_\beta\right]\\ &\tfrac{1}{2}\left[\overline{\varphi}_{\dot\alpha}\varepsilon^{\dot\alpha\dot\beta}\overline{\varphi}_{\dot\beta} - \varphi_\alpha\varepsilon^{\alpha\beta}\varphi_\beta\right]\\ &\tfrac{i}{2}\left[\overline{\varphi}_{\dot\alpha}\varepsilon^{\dot\alpha\dot\beta}\overline{\varphi}_{\dot\beta} + \varphi_\alpha\varepsilon^{\alpha\beta}\varphi_\beta\right]\end{aligned} \tag{A.101}$$

Since total derivatives have no effect on the equation of motion, the most general free-particle Lagrangian density that can be constructed from these Hermitian operators is

$$\begin{aligned}\mathcal{L} &= \tfrac{i}{2}\left[\overline{\varphi}_{\dot\alpha}(\overline{\sigma}^\mu)^{\dot\alpha\beta}(\partial_\mu\varphi_\beta) - (\partial_\mu\overline{\varphi}_{\dot\alpha})(\overline{\sigma}^\mu)^{\dot\alpha\beta}\varphi_\beta\right]\\ &\quad -\tfrac{1}{2}m_1\left[\overline{\varphi}_{\dot\alpha}\varepsilon^{\dot\alpha\dot\beta}\overline{\varphi}_{\dot\beta} - \varphi_\alpha\varepsilon^{\alpha\beta}\varphi_\beta\right]\\ &\quad -\tfrac{i}{2}m_2\left[\overline{\varphi}_{\dot\alpha}\varepsilon^{\dot\alpha\dot\beta}\overline{\varphi}_{\dot\beta} + \varphi_\alpha\varepsilon^{\alpha\beta}\varphi_\beta\right]\\ &= \tfrac{i}{2}\left[\overline{\varphi}_{\dot\alpha}(\overline{\sigma}^\mu)^{\dot\alpha\beta}(\partial_\mu\varphi_\beta) - (\partial_\mu\overline{\varphi}_{\dot\alpha})(\overline{\sigma}^\mu)^{\dot\alpha\beta}\varphi_\beta\right]\\ &\quad -\tfrac{1}{2}\left[m\,\overline{\varphi}_{\dot\alpha}\varepsilon^{\dot\alpha\dot\beta}\overline{\varphi}_{\dot\beta} - m^*\varphi_\alpha\varepsilon^{\alpha\beta}\varphi_\beta\right]\end{aligned} \tag{A.102}$$

where m_1 and m_2 are arbitrary real parameters, and $m \equiv m_1 + im_2$.

The $(\frac{1}{2}, 0)$ spinor field $\chi^{\dot\alpha}(x)$ can be treated in an exactly analogous fashion, and the most general free particle Lagrangian density in terms of $\chi^{\dot\alpha}$ and its Hermitian conjugate $\overline{\chi}^\alpha = (\chi^{\dot\alpha})^\dagger$ is found to be

$$\begin{aligned}\mathcal{L} &= \tfrac{i}{2}\left[\overline{\chi}^\alpha(\sigma^\mu)_{\alpha\dot\beta}(\partial_\mu\chi^{\dot\beta}) - (\partial_\mu\overline{\chi}^\alpha)(\sigma^\mu)_{\alpha\dot\beta}\chi^{\dot\beta}\right]\\ &\quad -\tfrac{1}{2}\left[m\,\chi^{\dot\alpha}\varepsilon_{\dot\alpha\dot\beta}\chi^{\dot\beta} - m^*\overline{\chi}^\alpha\varepsilon_{\alpha\beta}\overline{\chi}^\beta\right]\end{aligned} \tag{A.103}$$

The terms with m and m^* that appear in Eqs. (A.102) and (A.103) are examples of *Majorana mass* terms. They are characterized by the fact that they are not invariant under the change of phase of the spinor fields. Consequently, when it is necessary to maintain the freedom to change the phases of the fields, Majorana mass terms are forbidden.

2.A.2.2
Dirac Masses, Dirac Spinors, and the Dirac Equation

When both the $(0, \frac{1}{2})$ field $\varphi_\alpha(x)$ and the $(\frac{1}{2}, 0)$ field $\chi^{\dot\alpha}(x)$ are present, the following scalars can be constructed in addition to the ones discussed above:

$$\begin{aligned} \overline{\varphi}_{\dot\alpha}\chi^{\dot\alpha} &= -\overline{\varphi}^{\dot\alpha}\chi_{\dot\alpha} = \chi_{\dot\alpha}\overline{\varphi}^{\dot\alpha} = +\overline{\varphi}^{\dot\alpha}\varepsilon_{\dot\alpha\dot\beta}\chi^{\dot\beta} \\ \overline{\chi}^{\alpha}\varphi_\alpha &= -\overline{\chi}_\alpha\varphi^\alpha = \varphi^\alpha\overline{\chi}_\alpha = -\overline{\chi}_\alpha\varepsilon^{\alpha\beta}\varphi_\beta \end{aligned} \tag{A.104}$$

Since

$$\begin{aligned} \left[\overline{\varphi}_{\dot\alpha}\chi^{\dot\alpha}\right]^\dagger &= \overline{\chi}^\alpha\varphi_\alpha \\ \left[\overline{\chi}^\alpha\varphi_\alpha\right]^\dagger &= \overline{\varphi}_{\dot\alpha}\chi^{\dot\alpha} \end{aligned} \tag{A.105}$$

the Hermitian combinations are

$$(\overline{\varphi}_{\dot\alpha}\chi^{\dot\alpha} + \overline{\chi}^\alpha\varphi_\alpha), \quad \text{and} \quad i(\overline{\varphi}_{\dot\alpha}\chi^{\dot\alpha} - \overline{\chi}^\alpha\varphi_\alpha) \tag{A.106}$$

These terms can be combined into the mass term

$$\begin{aligned} m_1(\overline{\varphi}_{\dot\alpha}\chi^{\dot\alpha} + \overline{\chi}^\alpha\varphi_\alpha) &+ im_2(\overline{\varphi}_{\dot\alpha}\chi^{\dot\alpha} - \overline{\chi}^\alpha\varphi_\alpha) \\ &= m\overline{\varphi}_{\dot\alpha}\chi^{\dot\alpha} + m^*\overline{\chi}^\alpha\varphi_\alpha \end{aligned} \tag{A.107}$$

where $m = m_1 + im_2$. In the absence of Majorana mass terms, the phase of m can be absorbed into redefinitions of the fields φ_α and/or $\chi^{\dot\alpha}$, in which case this *Dirac mass* term reduces to

$$m\left(\overline{\varphi}_{\dot\alpha}\chi^{\dot\alpha} + \overline{\chi}^\alpha\varphi_\alpha\right) \tag{A.108}$$

with m real. Note that unlike the Majorana mass, the Dirac mass term is invariant under a change of phase of the fields, as long as $\varphi_\alpha(x)$ and $\chi^{\dot\alpha}(x)$ are transformed by the same phase.

The free particle Lagrangian density is then

$$\begin{aligned} \mathcal{L} = \frac{i}{2}&\left[\overline{\varphi}_{\dot\alpha}(\overline{\sigma}^\mu)^{\dot\alpha\beta}(\partial_\mu\varphi_\beta) - (\partial_\mu\overline{\varphi}_{\dot\alpha})(\overline{\sigma}^\mu)^{\dot\alpha\beta}\varphi_\beta\right] \\ &+ \frac{i}{2}\left[\overline{\chi}^\alpha(\sigma^\mu)_{\alpha\dot\beta}(\partial_\mu\chi^{\dot\beta}) - (\partial_\mu\overline{\chi}^\alpha)(\sigma^\mu)_{\alpha\dot\beta}\chi^{\dot\beta}\right] \\ &- m\left(\overline{\varphi}_{\dot\alpha}\chi^{\dot\alpha} + \overline{\chi}^\alpha\varphi_\alpha\right) \end{aligned} \tag{A.109}$$

This expression can be rewritten in a compact form by introducing the *Dirac spinor*:

$$\begin{aligned} \psi &\equiv \begin{bmatrix} \varphi_\alpha \\ \chi^{\dot\alpha} \end{bmatrix}, \quad \psi^\dagger = \begin{bmatrix} \overline{\varphi}_{\dot\alpha} & \overline{\chi}^\alpha \end{bmatrix}, \\ \overline{\psi} &\equiv \begin{bmatrix} \overline{\chi}^\alpha & \overline{\varphi}_{\dot\alpha} \end{bmatrix} \end{aligned} \tag{A.110}$$

Under a Lorentz transformation, ψ transforms as

$$\psi \longrightarrow \psi' = \begin{bmatrix} b & 0 \\ 0 & a \end{bmatrix}\psi \tag{A.111}$$

where the matrices a and b were defined in Eq. (A.34). Define the *gamma matrices* as

$$\gamma^\mu = \begin{bmatrix} 0 & (\sigma^\mu)_{\alpha\dot\beta} \\ (\overline{\sigma}^\mu)^{\dot\alpha\beta} & 0 \end{bmatrix} \tag{A.112}$$

or more explicitly

$$\gamma^0 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \gamma^i = \begin{bmatrix} 0 & -\sigma_i \\ \sigma_i & 0 \end{bmatrix} \tag{A.113}$$

Note that $\overline{\psi} = \psi^\dagger\gamma^0$. The relations among σ^μ and $\overline{\sigma}^\mu$ given in Eq. (A.71) translate to

$$\gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu = 2g^{\mu\nu} \tag{A.114}$$

where a 4 × 4 unit matrix on the right-hand side is implied. Using this notation, Eq. (A.109) becomes

$$\begin{aligned}\mathcal{L} &= \frac{i}{2}\left\{\overline{\psi}\gamma^{\mu}(\partial_{\mu}\psi) - (\partial_{\mu}\overline{\psi})\gamma^{\mu}\psi\right\} - m\overline{\psi}\psi \\ &= \overline{\psi}\,(i\partial\!\!\!/ - m)\,\psi + (\text{total derivative})\end{aligned} \tag{A.115}$$

where $\partial\!\!\!/$ is shorthand for $\gamma_{\mu}\partial^{\mu}$. The Euler-Lagrange equation that results from this Lagrangian density is

$$(i\partial\!\!\!/ - m)\,\psi = 0 \tag{A.116}$$

which is known as the *Dirac equation* (Dirac, 1928). Note that this equation is manifestly Lorentz covariant. Acting on both sides with the operator $(i\partial\!\!\!/ + m)$, and using Eq. (A.114), we find

$$(-\partial^2 - m^2)\psi = 0 \tag{A.117}$$

which justifies our terming the parameter m a mass, given the quantum mechanical identification of $i\partial^{\mu}$ with p^{μ}.

The Dirac equation is also often seen in the following form, which treats the time component of ∂^{μ} separately from the spacial components:

$$i\partial^0\psi = \Big(-i\boldsymbol{\alpha}\cdot\nabla + m\beta\Big)\psi \tag{A.118}$$

Here, $\boldsymbol{\alpha}$ and β are matrices defined as

$$\begin{aligned}\alpha^i \equiv \gamma^0\gamma^i &= \begin{bmatrix} \sigma_i & 0 \\ 0 & -\sigma_i \end{bmatrix}, \\ \beta \equiv \gamma^0 &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\end{aligned} \tag{A.119}$$

Making the quantum mechanical identifications $i\partial^0 = H$ and $-i\boldsymbol{\nabla} = \mathbf{p}$, we can also write

$$H\psi = \big(\boldsymbol{\alpha}\cdot\mathbf{p} + m\beta\big)\,\psi \tag{A.120}$$

2.A.2.3
Chiral, Dirac, and Majorana Representations

The expressions of the Dirac spinor and the gamma matrices shown above are known as the *chiral representation*. Performing unitary transformations on ψ and γ^{μ} does not change the relation among the γ^{μ}'s, Eq. (A.114), or the form of the Lagrangian density, Eq. (A.115), or the Dirac equation, Eq. (A.116). This freedom can be exploited to change the matrix representations of the γ^{μ}'s into more convenient forms depending on the problem at hand. Starting with the chiral representation,

$$\begin{aligned}\psi_C &= \begin{bmatrix} \varphi_{\alpha} \\ \chi^{\dot{\alpha}} \end{bmatrix}, \quad \gamma_C^0 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \\ \gamma_C^i &= \begin{bmatrix} 0 & -\sigma_i \\ \sigma_i & 0 \end{bmatrix}\end{aligned} \tag{A.121}$$

where we have added a subscript to designate the representation, we define the *Dirac representation* via

$$\begin{aligned}\psi_D &= U\psi_C, \quad \gamma_D^{\mu} = U\gamma_C^{\mu}U^{\dagger}, \\ U &= \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\end{aligned} \tag{A.122}$$

Explicitly, the gamma matrices in the Dirac representation are given by

$$\gamma_D^0 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \gamma_D^i = \begin{bmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{bmatrix} \tag{A.123}$$

This representation is useful for taking the nonrelativistic limit and is used below.

Another representation that is sometimes used is the *Majorana representation*, which is defined via

$$\begin{aligned}\psi_M &= U\psi_D, \quad \gamma_M^{\mu} = U\gamma_D^{\mu}U^{\dagger}, \\ U &= \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & \sigma_2 \\ \sigma_2 & -1 \end{bmatrix}\end{aligned} \tag{A.124}$$

Explicitly, the gamma matrices in the Majorana representation are given by

$$\gamma_M^0 = \begin{bmatrix} 0 & \sigma_2 \\ \sigma_2 & 0 \end{bmatrix}$$

$$\gamma_M^1 = \begin{bmatrix} i\sigma_3 & 0 \\ 0 & i\sigma_3 \end{bmatrix}$$

$$\gamma_M^2 = \begin{bmatrix} 0 & -\sigma_2 \\ \sigma_2 & 0 \end{bmatrix}$$

$$\gamma_M^3 = \begin{bmatrix} -i\sigma_1 & 0 \\ 0 & -i\sigma_1 \end{bmatrix} \qquad \text{(A.125)}$$

Note that in this representation the gamma matrices are all purely imaginary, which has the effect of making the Dirac operator $(i\partial\!\!\!/ - m)$ real.

2.A.2.4
Coupling to the Electromagnetic Field and the Nonrelativistic Limit

As discussed in Section 2.3.4, the coupling of a charged particle to the electromagnetic field is accomplished by the substitution $p^\mu \to p^\mu - qA^\mu$ where q is the electromagnetic charge of the particle. Quantum mechanically, p^μ corresponds to the operator $i\partial^\mu$, so in order to couple the Dirac fermion to the electromagnetic field, we perform the substitution $i\partial^\mu \to i\partial^\mu - qA^\mu$ to yield

$$\left(i\partial\!\!\!/ - qA\!\!\!/ - m\right)\psi = 0 \qquad \text{(A.126)}$$

or equivalently,

$$H\psi = \left[\boldsymbol{\alpha}\cdot(\mathbf{p} - q\mathbf{A}) + m\beta + q\phi\right]\psi \qquad \text{(A.127)}$$

To find the nonrelativistic limit of this equation, we use the Dirac representation for ψ and the γ's, and write

$$\psi_D = \begin{bmatrix} \varphi \\ \chi \end{bmatrix} \qquad \text{(A.128)}$$

(Note that φ and χ are just the upper and lower components of ψ_D. They are not $(0, \frac{1}{2})$ and $(\frac{1}{2}, 0)$ fields as in the chiral representation.) The equation for the energy eigenstate with eigenvalue E in terms of φ and χ is

$$\begin{aligned}(E - m - q\phi)\,\varphi &= \left[\boldsymbol{\sigma}\cdot(\mathbf{p} - q\mathbf{A})\right]\chi \\ (E + m - q\phi)\,\chi &= \left[\boldsymbol{\sigma}\cdot(\mathbf{p} - q\mathbf{A})\right]\varphi \qquad \text{(A.129)}\end{aligned}$$

Setting $E = m + E'$ and eliminating χ, we obtain

$$\begin{aligned}(E' - q\phi)\varphi &= \left[\boldsymbol{\sigma}\cdot(\mathbf{p} - q\mathbf{A})\right]\frac{1}{2m + E' - q\phi} \\ &\quad \times \left[\boldsymbol{\sigma}\cdot(\mathbf{p} - q\mathbf{A})\right]\varphi \qquad \text{(A.130)}\end{aligned}$$

When the particle is in a uniform magnetic field $\mathbf{B}$, we can set

$$\phi = 0, \qquad \mathbf{A} = \tfrac{1}{2}(\mathbf{B}\times\mathbf{r}) \qquad \text{(A.131)}$$

Expanding Eq. (A.130) in inverse powers of m, we find

$$\begin{aligned}E'\varphi &= \left(\frac{1}{2m}\left[\boldsymbol{\sigma}\cdot(\mathbf{p} - q\mathbf{A})\right]\right. \\ &\qquad \left.\times\left[\boldsymbol{\sigma}\cdot(\mathbf{p} - q\mathbf{A})\right] + \cdots\right)\varphi \\ &= \left(\frac{1}{2m}\left[(\mathbf{p}^2 - 2q\mathbf{A}\cdot\mathbf{p} + q^2\mathbf{A}^2)\right.\right. \\ &\qquad \left.\left. - q\boldsymbol{\sigma}\cdot(\nabla\times\mathbf{A})\right] + \cdots\right)\varphi \\ &= \left(\frac{\mathbf{p}^2}{2m} - \frac{q}{2m}\mathbf{L}\cdot\mathbf{B} - \frac{q}{m}\frac{\boldsymbol{\sigma}}{2}\cdot\mathbf{B} + \cdots\right)\varphi \qquad \text{(A.132)}\end{aligned}$$

where we have used the relation

$$(\boldsymbol{\sigma}\cdot\mathbf{A})(\boldsymbol{\sigma}\cdot\mathbf{B}) = (\mathbf{A}\cdot\mathbf{B}) + i\boldsymbol{\sigma}\cdot(\mathbf{A}\times\mathbf{B}) \qquad \text{(A.133)}$$

and dropped terms of order q^2 in the last line. Equation (A.132) shows that the magnetic dipole moment associated with the spin is

$$\mathbf{m} = \frac{g}{2m}\mathbf{S} = \frac{g}{2m}\frac{\boldsymbol{\sigma}}{2} \tag{A.134}$$

with $g = 2$ as claimed in Eq. (130). Keeping terms next to leading order in $1/m$ gives the leading relativistic correction. We perform this approximation for the hydrogen atom with

$$q = -e, \quad q\phi = -\frac{e^2}{r} \equiv V(r), \quad \mathbf{A} = \mathbf{0} \tag{A.135}$$

in which case we find

$$\begin{aligned}
&(E' - V)\varphi \\
&= (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{1}{2m}\left[1 - \left(\frac{E' - V}{2m}\right) + \cdots\right](\boldsymbol{\sigma}\cdot\mathbf{p})\varphi \\
&= \left(\frac{\mathbf{p}^2}{2m} - \frac{1}{8m^2}\left\{(E' - V)\mathbf{p}^2 + \mathbf{p}^2(E' - V)\right\}\right. \\
&\quad \left. - \frac{ie}{8m^2}\left[(\boldsymbol{\sigma}\cdot\mathbf{E}), (\boldsymbol{\sigma}\cdot\mathbf{p})\right] + \cdots\right)\varphi \\
&= \left(\frac{\mathbf{p}^2}{2m} - \frac{1}{8m^2}\left\{(E' - V)\mathbf{p}^2 + \mathbf{p}^2(E' - V)\right\}\right. \\
&\quad + \frac{e}{8m^2}\left\{(\nabla\cdot\mathbf{E}) + i\boldsymbol{\sigma}\cdot(\nabla\times\mathbf{E})\right\} \\
&\quad \left. + \frac{e}{4m^2}\boldsymbol{\sigma}\cdot(\mathbf{E}\times\mathbf{p}) + \cdots\right)\varphi \\
&= \left(\frac{\mathbf{p}^2}{2m} - \frac{\mathbf{p}^4}{8m^3} + \frac{\pi e^2}{2m^2}\delta^3(\mathbf{r})\right. \\
&\quad \left. + \frac{e^2}{2m^2r^3}\frac{\boldsymbol{\sigma}}{2}\cdot\mathbf{L} + \cdots\right)\varphi
\end{aligned} \tag{A.136}$$

The second term in this expression is just the third term in the expansion

$$E = \sqrt{m^2 + \mathbf{p}^2} = m + \frac{\mathbf{p}^2}{2m} - \frac{\mathbf{p}^4}{8m^3} + \cdots \tag{A.137}$$

The third term in Eq. (A.136) is known as the *Darwin term* (Darwin 1928) and only affects *s*-wave states. The fourth term is the spin-orbit coupling that was discussed in Section 2.3.4.5.

References

Ashby, N. (**2003**) *Relativity in the Global Positioning System*, Living Rev. Relativity **6**, 1, *http://www.livingreviews.org/lrr-2003-1*.

Bailey, J. and Picasso, E. (**1970**) Anomalous magnetic moment of the muon and related topics. *Progr. Nucl. Phys.*, **12**, 43–75.

Chang, Y.-C. and Zhang, Y.-Z. (**1998**) *Special Relativity and Its Experimental Foundations*, World Scientific.

Compton, A.H. (**1923**) A quantum theory of the scattering of X-rays by light elements. *Phys. Rev.*, **21**, 483–502.

(a) Darwin, C.G. (**1928**) The wave equations of the electron. *Proc. R. Soc. Lond., A*, **118**, 654–80; (b) Foldy, L.L. and Wouthuysen, S.A. (**1950**) On the Dirac theory of spin 1/2 particles and its non-relativistic limit. *Phys. Rev.*, **78**, 29–36; (c) Kennedy, W.L. (**1988**) Partial equivalence of the correct and incorrect versions of the Darwin term. *J. Phys. A: Math. Gen.*, **21**, 3021–23.

(a) Dirac, P.A.M. (**1928**) The quantum theory of the electron. *Proc. R. Soc. London, Ser. A*, **117**, 610–24; (b) Dirac, P.A.M. (**1928**) The quantum theory of the electron, part II. *Proc. R. Soc. London, Ser. A*, **118**, 351–61.

Einstein, A. (**1905a**) Uber einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt (On a heuristic viewpoint concerning the production and transformation of light). *Ann. Phys.*, **17**, 132–48.

(a) Einstein, A. (**1905a**) Zur Electrodynamik bewegter Körper (On the electrodynamics of moving bodies). *Ann. Phys.*, **17**, 891–921; (b) Einstein, A. (**1905b**) Ist die Trägheit eines Körpers von seinen Energieinhalt abhängig? (Does the inertia of a body depend upon its energy-content?). *Ann. Phys.*, **18**, 639–41. English translations of these papers are available in Hawking 2003.

Einstein, A. (**1911**) Uber den Einfluss der Schwerkraft auf die Ausbreitung des Lichtes (On the influence of gravitation on the propagation of light). *Ann. Phys.*, **35**. English translation available in Hawking 2003.

(a) Einstein, A. (**2006**) *Relativity: The Special and the General Theory*, originally published in German in 1916, English translation by Robert W. Lawson 1920, new edition with an introduction by Nigel Calder, Penguin Classics; (b) Gamow, G. (**1993**) *Mr Tompkins in Wonderland*, first published in 1940, reprinted in *Mr. Tompkins in Paperback*, Cambridge University Press; (c) Landau, L.D. and Rumer, G.B. (**2003**) *What is Relativity?* originally published in Russian in 1959, English translation by N. Kemmer 1960, Dover; (d) Gardner, M. (**1997**) *Relativity Simply Explained*, Dover; (e) Wolfson, R. (**2003**) *Simply Einstein: Relativity Demystified*, W. W. Norton & Company; (f) Taylor, E.F. and Wheeler, J.A. (**1992**) An introductory text with some math would be (f) Taylor, E.F. and Wheeler, J.A. (1992) *Spacetime Physics*, 2nd edn., W.H. Freeman.

(a) Gross, F. (**1999**) *Relativistic Quantum Mechanics and Field Theory*, Revised edn, Wiley-Interscience; (b) Peskin, M.E. and Schroeder, D.V. (**1995**) *An Introduction to Quantum Field Theory*, Westview Press; (c) Itzykson, C. and Zuber, J.-B. (**1980**) *Quantum Field Theory*, McGraw-Hill; (d) Weinberg, S. (**1995, 1996, 2000**) *The Quantum Theory of Fields*, Vols. 1 to 3, Cambridge University Press.

(a) Hafele, J.C. and Keating, R.E. (**1972**) Around-the-world atomic clocks: predicted relativistic time gains. *Science*, **177**, 166–168; (b) Hafele, J.C. and Keating, R.E. (**1972**) Around-the-world atomic clocks: observed relativistic time gains. *Science*, **177**, 168–70.

(a) Halzen, F. and Martin, A.D. (**1984**) *Quarks and Leptons: An Introductory Course in Modern Particle Physics*, Wiley; (b) Griffiths, D. (**2008**) *Introduction to Elementary Particles*, 2nd edn, Wiley-VCH.

(a) Hartle, J.B. (**2003**) *Gravity; An Introduction to Einstein's General Relativity*, Addison Wesley; (b) Rindler, W. (**2006**) *Relativity; Special, General, and Cosmological*, 2nd edn, Oxford University Press.

Hawking, S. (**2003**) *On the Shoulders of Giants*, Running Press.

Hubble, E. (**1929**) A relation between distance and radial velocity among extra-galactic nebulae. *Proc. Natl. Acad. Sci. USA*, **15**, 168–73.

Jackson, J.D. (**1998**) *Classical Electrodynamics*, 3rd edn, Wiley.

Kostelecky, V.A. (ed.) (**1999, 2002, 2005, 2008**) *CPT and Lorentz Symmetry I, II, III, IV*, World Scientific.

Michelson, A.A. and Morley, E.W. (**1887**) On the relative motion of the earth and the luminiferous ether. *Am. J. Sci.*, **34**, 333–45.

Misner, C.W., Thorne, K.S., and Wheeler, J.A. (**1973**) *Gravitation*, W. H. Freeman.

Muller, R.A. (**1992**) Thomas precession: where is the torque? *Am. J. Phys.*, **60**, 313–17.

Particle Data Group (**2009**) *Review of Particle Physics*, *http://www.pdg.lbl.gov/*.

Planck, M. (**1901**) Uber das Gesetz der Energieverteilung im Normalspectrum (On the law of distribution of energy in the normal spectrum). *Ann. Phys.*, **4**, 533.

(a) Pound, R.V. and Rebka, G.A. Jr (**1959**) Gravitational red-shift in nuclear resonance. *Phys. Rev. Lett.*, **3**, 439–41; (b) Pound, R.V. and Snider, J.L. (**1965**) Effect of gravity on gamma radiation. *Phys. Rev.*, **140B**, 788–803.

Rubin, V.C., Ford, W.K. Jr. and Thonnard, N. (**1980**) Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 (R = 4kpc) to UGC 2885 (R = 122 kpc), *Astrophys. J.*, **238**, 471–87.

Rybicki, G.B. and Lightman, A.P. (**1985**) *Radiative Processes in Astrophysics*, Wiley-Interscience.

Taylor, E.F., Wheeler, J.A., and Bertschinger, E. (**2008**) *Exploring Black Holes: Introduction to General Relativity*, 2nd edn, Addison Wesley.

(a) Thomas, L.H. (**1926**) The motion of the spinning electron. *Nature*, **117**, 514; (b) Thomas, L.H. (**1927**) The kinematics of an electron with an axis. *Phil. Mag.*, **3**, 1–22.

(a) Wigner, E.P. (**1939**) On unitary representations of the Inhomogeneous Lorentz Group. *Ann. Math.*, **40**, 149–204; (b) Wigner, E.P. (**1957**) Relativistic invariance and quantum phenomena. *Rev. Mod. Phys.*, **29**, 255–68.

RESEARCH AND DEVELOPMENT

ROBERT A. FROSCH, *John F. Kennedy School of Government, Harvard University, Cambridge, Massachusetts, U.S.A.*

Introduction 415
1. **Varieties and Terms** 415
2. **Relationships among the Varieties of R&D** 416
3. **Interactions among Kinds of Knowledge in R&D** 417
3.1 Processes of Implementation 418
4. **Institutions in R&D** 419
4.1 Performers 419
4.2 Other Organizations 420
4.3 Industrial Policy 420
5. **Funding of R&D** 420
6. **Concluding Remarks** 421
Works Cited 421
Further Reading 421

INTRODUCTION

"Research and development" (R&D) refers to the processes by which new knowledge is discovered and new technology, products, and manufacturing processes are created. "Technology" is used here in the broad sense, to mean the capability and knowledge of "how to" make products and processes. ("A technology" or "the technology" frequently refers to both the body of knowledge and the tools and processes required to make particular kinds of products.) "Product" is also used here in the broadest possible sense, to include things and technical ideas, and abstract or nonhardware processes and products, including software.

1. VARIETIES AND TERMS

"Research" is frequently used to refer to the entire research and development process, but in its more specific sense refers to the search for new knowledge. When driven by interest in the subject and the knowledge itself, it is frequently referred to as "curiosity-driven" research, as "basic" research, or as "fundamental" research. The latter two terms imply that the new knowledge underlies knowledge obtained with other motivations, although this is not always the case.

The U.S. National Science Foundation (NSF) states (NSF, 1993, p. 94) that "the objective of basic research is to gain a more complete knowledge or understanding of the subject under study, without specific applications in mind." It adds that in industry basic research "does not have specific immediate commercial objectives, although it may be in fields of present or potential commercial interest."

"Applied research" is used for the search for new knowledge when there is a purpose in mind for the new knowledge to serve—particularly when it is wanted to help solve some specific problem, frequently a problem posed in an attempt to develop a new technology. The NSF (1993) describes it thus: "Applied Research.... to determine the means by which a specific, recognized need may be met." "Mission-oriented research" refers to research undertaken because the resulting knowledge is expected to help the work of the organization doing or sponsoring the research. "Mission-oriented research" is sometimes extended to include the idea of "mission-oriented basic research," in which the areas in which basic research is to be conducted are selected from subject areas in which new knowledge is believed to be likely to be of use in carrying out the mission purpose of the organization. It is usually used in the context of government agency research programs.

"Development" is used for the work necessary to create particular new products, processes for the manufacture of products, and software. NSF (1993) says "... systematic use of the knowledge or understanding

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

gained from research directed toward the production of useful materials, devices . . ."

"Exploratory development" has been used by the United States Department of Defense (DOD) in a sense nearly synonymous with "applied research," referring to the search for technological capabilities that may be applied to defense products.

"Advanced development" refers to work undertaken to demonstrate that a particular product or process can be created and function. Advanced development, while it demonstrates the necessary principles and frequently results in a working model, does not usually result in a product or process that has the cost, quality, reliability, and other properties required for replication and general use in the market. Advanced development is sometimes said to result in "proof of principle."

"Engineering development" refers to work that results in a product or process that has the appropriate manufacturability, quality, cost, reliability, and other properties required for routine use. Engineering development may include pilot production, the production of limited quantities of the product on a realistic, but not necessarily full-scale or full-rate, production system. Engineering development is sometimes said to result in "reduction to practice."

Engineering development may include a process of "test and evaluation" of the resulting product and/or manufacturing processes, to demonstrate that they do have the expected properties.

"Operational development" is sometimes used (e.g., in DOD) to refer to development work made necessary by product and manufacturing problems that develop in actual use.

"Innovation" refers to the whole set of processes used to bring a new product or process into the market. This generally includes business processes, as well as the technical processes of R&D.

Basic and applied research are generally considered to be in the domain of science, and more often involve scientists than engineers, while development is generally considered to be in the domain of engineering, and more often involves engineers than scientists. People educated as scientists and as engineers engage in both kinds of work.

2. RELATIONSHIPS AMONG THE VARIETIES OF R&D

The definitions of varieties of R&D described by the terms above cannot be made precise, but are approximate descriptions of activities that may be difficult to distinguish. It is frequently assumed that the terms describe a linear temporal sequence of stages in which basic research is followed by applied research, which is followed by advanced development and then engineering development, the sequence leading in an invariant orderly way to the introduction of new products or processes. This has been called the "linear model" of R&D. In actual practice, the R&D process is often far more intricate and complex.

There are cases in which discoveries in basic research lead to an understanding of possibilities for products or processes, following which there is an orderly progression through further applied research into various stages of development. Such cases, when based upon important and really new discoveries, often lead to revolutionary new products and processes. However, the development of a product may start at any stage of the presumed linear model. For example, what is called "continuous improvement" is normally a sequence of small, incremental changes made to an existing product or process to make it better in use, cheaper, more reliable, etc. These might be described as small engineering development changes, or as operational development changes, frequently undertaken without being formally described as development. It is not unusual for problems in making continuous improvements, or difficulties in engineering development, to lead to requirements for advanced development of technology, or to applied research to produce knowledge needed before the development process can successfully be completed. Sometimes problems arising in development or applied research may suggest gaps in fundamental knowledge, thus leading basic research in a new direction.

Many products and processes are based upon "craft knowledge," developed through years, even centuries or millennia, of practical experimentation, not upon systematically derived scientific knowledge. Systematic attempts are sometimes later mounted to create the scientific knowledge necessary to understand why and how these craft skills work, and what is the basis of the craft knowledge of how to do things. This is particularly the case when attempts are made to extend the use of craft knowledge beyond the domain in which it developed. Thus development may pull upon research.

Basic and applied research frequently depend upon tools that are the product of craft skills, or of much systematic development work. Even in the case when the fundamental idea for the tool arises in the basic research, it may be unusable until much development and engineering work has been done to make the tool useful.

Thus there are complex feedback and feedforward relationships among the stages of the presumed linear model.

3. INTERACTIONS AMONG KINDS OF KNOWLEDGE IN R&D

The complexity of R&D and of the R&D process, driven in part by the many possible relationships among the varieties of R&D, outlined above, is further increased by relationships among subjects and kinds of knowledge. In order to create any new product or process, it is necessary to merge knowledge and technology of various kinds. No physical product or process is exclusively electrical, electronic, chemical, or mechanical; real development always requires knowledge and technology from a variety of fields. For example, electronics must be sound from a heat-flow point of view and must be structurally supported by mechanical means; its manufacture requires chemistry. The development of new transportation vehicles may embody combustion chemistry controlled by electronics, carrying cargo by mechanical means, etc.

Thus development is likely to require a horizontal fusion of subjects and disciplines coming together into a new mix of technology and knowledge. Completely new product possibilities are sometimes created by the process of "technology fusion," as for example in the fusion of mechanical, electrical, electronic control, and computer technologies to create intelligent robotic automation. In addition to horizontal technology fusion, each of the individual subjects may require vertical extension of the underlying knowledge of the discipline to provide new solutions to new problems. All this may require support from other subjects, mathematical solutions for new chemistry, for example, or new means of doing structural or thermal computations for a regime not encountered before in previous products or processes.

The general nature of the requirements for contributions from other disciplines and technologies may be understood in advance, but the requirements are unlikely to be predictable in detail for any but the simplest developments. An ability to draw upon multiple sources of knowledge and technology is likely to be key for successful development, especially of complex products and processes. The history of applied research, and of development, is full of cases in which knowledge created by curiosity-driven basic research, not previously believed to have any connection with any application, has been crucial to the solution of applied and developmental problems. It is generally straightforward to predict what kinds of knowledge will be required for an application or a development; it has proved to be impossible to know in advance what knowledge and technology will be irrelevant. While apparently unconnected in a direct way to application, the body of basic knowledge derived from curiosity-driven research has proven to be essential to application, although sometimes with a long time lag.

The horizontal and vertical complexity of applied research and product and process development has resulted in most industrial R&D, and the R&D of mission-oriented government agencies, being strongly interdisciplinary. Academic research, not directly driven by application requirements, and thus generally in the curiosity-driven research area, has tended to be disciplinary in nature as a result of the traditional organization of scholarly subjects. However, increasing involvement of universities with industrial problems and sponsors, and the increasingly complicated relationships among the underlying disciplines, may change this.

One way of visualizing the complexity of a development process is to consider a matrix in which the rows are labeled with the various technological problems to be solved during a development project, while the columns are labeled with the kinds of knowledge expected to be required (or known from the history of a project to have been required) to complete the project successfully. Such a matrix display provides a map of the transform of knowledge and technology into product or process. Examples can deepen the understanding of the complex ways in which various kinds of knowledge and technology are the ingredients for development of products and processes.

3.1 Processes of Implementation

Since the process of developing a product or manufacturing process can be very complex, and since the nature of what is being developed may be forced by the facts of nature to change during the development, the relationship between the customer for the final development result and those who are carrying out the development is extremely important. (The immediate customer for the final result may be the ultimate market consumer or user, but it is likely that the proximate customer will be someone in a firm or government agency acting as a surrogate for the final user.) Experience suggests that it is extremely difficult to make the requirements for the development result sufficiently complete and unambiguous without also making them so constraining as to preclude the developer from finding or creating the best solution to the problem in terms of performance, cost, etc.

During applied research and development, it is often the case that the understanding of the real nature of the customer's problem may change, as well as the course of the development changing the understanding of possible solutions to the problem. The course of successful applications R&D is frequently marked by adjustments both in the understanding of the problem, and thus its redefinition, and in new understanding of the necessary knowledge and technology that must be discovered and created to solve the, possibly redefined, problem.

To deal with these changes that occur naturally in the course of applied research and development, reasonable dialog between customer and developer before, during, and after completion of a development can be an important determinant of success.

For the reasons just discussed, successful development projects are frequently the product of tightly knit interdisciplinary teams. In many industrial cases, the team must include business and marketing people if the real end requirements are to be reflected in the result of the development.

To enter the market, the product must be manufacturable. It is important that the process of development and design of the product include consideration, design, and development of the means by which it will be manufactured. The process of development and design that considers both together, with the appropriate timing, so that the product and its manufacturing means are well fitted together is referred to as "concurrent engineering." Concurrent engineering enriches the interdisciplinary development team by requiring that product development and design people and manufacturing development and design people work together as a team, creating further complexity and timing complications in the development process.

During the course of research and development, knowledge and technological understanding gained by participants in one part of the process may have to be learned and understood by those who must participate in other parts of the process. When the development has been completed, the results, whether operation of the manufacturing process, or use, maintenance, and operation of the product, will have to be understood and applied by operators and users. The process of moving the knowledge and technological "know-how" from one group of people to another is referred to as "technology transfer."

Technology transfer may occur through the reading of professional and scholarly papers or of patents and similar professional material, but most often, when successful, occurs through direct contact between people with knowledge of technology and its underlying bases and those who will use the knowledge in product and process application and use. In this sense, the continuing dialog between developers and users, referred to above, is the continuing mechanism of technology transfer—really a process in which those who need, and create the requirements for, the knowledge and technology and those who discover and create it continually educate each other.

In its broadest sense, the term technology transfer is used to describe the process by which knowledge and technology created in one organization, or part of an organization, for a particular purpose, is transferred to an entirely different organization, or part of an organization, to be used for some new and different development purpose. While this has sometimes been attempted by the transfer of documents, or simple teaching, when successful it usually involves protracted discussions and the temporary or permanent transfer of people between the organizations. The comment about interdisciplinary teams, made above with regard to development, also applies to technology transfer since it can be regarded as a logical portion of, or extension of, development.

Innovation, the total process of creating change in products, processes, or businesses, and commercialization, the process of ensuring that the results of development are a marketable product or process, may be thought of as the development and technology transfer process extended to the whole set of business or government requirements for introducing something new. The previous comments about the rich complexity and interdisciplinary team nature of the process continue to apply; the field of disciplines and knowledge must be broadened further to encompass all the business and social knowledge and technology required.

The nature of the market for a product or process is sometimes such that commercial introduction must be made by a specific time, or the opportunity for introduction will be missed. In such cases, the "window of opportunity" for introduction creates a pacing time constraint for the development, requiring that it be successfully completed by a given time. This may occur, for example, because of a race between competitors to introduce similar products or services, or because there is a limited time in which a product may be of use before there is a change to a completely new product or technology.

4. INSTITUTIONS IN R&D

4.1 Performers

There are a variety of institutions performing R&D. These include universities, industrial laboratories that are part of general businesses, private commercial laboratories performing R&D for outside customers, private nonprofit research institutions, government laboratories belonging to mission agencies, government laboratories chartered as national laboratories for particular purposes (and frequently operated by private operators under government contract), and private foundations performing R&D.

In the U.S., many universities are referred to as "research universities." In these universities, the faculty both teach and perform research. This is usually basic research, generally within the boundaries of one of the traditional academic disciplines such as physics, chemistry, or a branch of engineering, although increasing attention has been given to subjects, such as environmental issues, that require interdisciplinary work. At the time of this writing (1993), in many universities increasingly strong support ties with industrial and mission agency sponsors are changing the traditional practice in the direction of mission-oriented interdisciplinary research. Faculty members in many universities are increasingly involved in outside consulting work for private industry, or in entrepreneurial activities that involve them in the transfer of their knowledge into industrial technology, and in development work based upon their research. This trend is particularly strong in the biological and health-related sciences.

Industrial laboratories are characteristically engaged in product and process development, although many engage in applied research, and some in some basic research. Industrial R&D is characteristically multidisciplinary, since it is strongly pulled and motivated by product and process development requiring a variety of kinds of knowledge and technology.

Government laboratories belonging to mission agencies span the spectrum from a few that engage in mission-oriented basic research and applied research, to the majority, which are more like industrial laboratories dedicated to the developments required by their respective agencies. Their work is also characteristically multidisciplinary. In the United States, the Department of Defense operates the largest number of government laboratories, although many other departments and agencies have laboratories of their own.

The "national laboratories" in the United States are a special case. Created to deal with the technology required for nuclear weapons and related defense requirements, they are operated by contractors for the Department of Energy. They, too, are multidisciplinary and product and process oriented.

Private commercial laboratories are generally development oriented, performing specialized development and test work tailored to specifically defined customer needs.

4.2 Other Organizations

In addition to the R&D performing organizations, and the sponsors and funders of R&D (see Sec. 5), there are other kinds of organizations important to R&D and its participants, both performers and sponsors. Many of these serve as public markets for scientific, technological, and engineering information and knowledge, and as a means of public quality control on technical information.

These other organizations include a variety of technical societies, such as The American Physical Society, the American Chemical Society, the American Association for the Advancement of Science, and other scientific societies, and a variety of engineering societies. These serve as technical fora in which knowledge and technical information are exchanged at meetings and through technical journals published by the societies. There are also a variety of technical journals published by commercial publishers, and technical books published by university and commercial presses.

There are also trade associations of particular manufacturing sectors that serve, among other functions, as clearinghouses for technical information, and sometimes for the organization of cooperative R&D [e.g., the Electric Power Research Institute (EPRI)]. In addition, there are organizations that cut across industry and serve R&D in a functional manner, such as the Industrial Research Institute (IRI), a forum for the discussion of industrial R&D issues, especially issues of management of R&D.

In the U.S., the national academies (the National Academy of Science, the National Academy of Engineering, and the Institute of Medicine, and their operating arm, the National Research Council) play a role in providing scientific and technological advice to the federal government—frequently advice concerning R&D policies, budgets, and sponsorship.

4.3 Industrial Policy

In Europe and Japan, industrial R&D and government support of industrial R&D have been dominant. In contrast, in the period from the end of World War II until the present in the U.S., Defense Department needs and funding have tended to dominate the funding and performance of R&D. During the last decade in the U.S., there have been increasing calls for the use of the technology developed through this work to be available for industrial use. U.S. government policy has begun to move in the direction of opening government-owned technology and the capabilities and work of the national and government laboratories to use by industry. Mechanisms such as the Cooperative Research and Development Agreement (CRADA) have been developed to foster the transfer of technology from the government to industry.

At the same time, it has become clear that in many fields of technology, such as microelectronics and computer technology, industrial technology productive capability has surpassed that of the government for most purposes, and there is an increasing push to use more "civilian" technology for defense and other government purposes. Policy development in the U.S. in this area is moving very rapidly.

5. FUNDING OF R&D

An idea of the scale and performers of R&D may be obtained from Table 1.

In 1994, the split between federally and

Table 1. R&D in the U.S., European Community (EC), and Japan for 1990 (Alic *et al.*, 1992: Table 4.2, p. 92, Table 4.5, p. 100, Fig. 4.3, p. 98; OECD, 1992, Table 14, p. 22).

	Total (10^9 \$)	Government (%)	Private (%)
U.S.	145.5	44[a]	56
EC	101.9	40.8	59.2
Japan	67.0	17.9	73.1

---

[a] DOD, Dept. of Energy, NASA = 60% of the 44%.

Table 2. U.S. DOD R&D funding in 1990 by category (Alic *et al.*, 1992, Table 4.7, p. 102).

Category	Fraction (%)
Applied research (research and exploratory development)	9
Advanced development	28
Engineering and operational systems development	55
Management and support	8

Table 3. U.S. DOD R&D funding in 1994 by category (AAAS, 1994, Table 1-5, p. 49).

Category	Fraction (%)
Basic research	18.4
Applied research	18.4
Development	59.1
Facilities	4.1

nonfederally funded R&D in the U.S. was roughly 43% to 57% (AAAS, 1994, Table 1-11, p. 55). The DOD funding in 1990 was split among categories of R&D as shown in Table 2.

In 1994, a slightly different categorization gives the figures in Table 3.

The proportions are different for industry-supported R&D, somewhat favoring development. For 1993, the AAAS gives numbers as in Table 4.

Table 4. U.S. industry-supported R&D by category (AAAS, 1994, pp. 23–24).

Category	Fraction (%)
Basic, or fundamental, research	about 5
Exploratory or applied research	nearly 25
Development	73

6. CONCLUDING REMARKS

The complexity of the R&D enterprise, as an intellectual and operating endeavor, as a system of performers and sponsors, and as a social system of people engaged in it will be apparent from the preceding. The complexity continues to evolve with the changing character of modern society.

Works Cited

AAAS (1994), *AAAS Report XVIII Research and Development FY 1994,* Washington, DC: Intersociety Working Group, American Association for the Advancement of Science.

Alic, J. A., Branscomb, L. M., Brooks, H., Carter, A. B., Epstein, G. L. (1992), *Beyond Spinoff: Military and Commercial Technologies in a Changing World,* Boston: Harvard Business School Press.

NSF (1993), *National Science Board, Science and Engineering Indicators—1993,* Washington, DC: U.S. Government Printing Office (NSB 93-1).

OECD (1992), *Main Science and Technology Indicators 1992/2,* Paris: OECD.

Further Reading

AAAS (1994), *AAAS Report XVIII Research and Development FY 1994,* Washington, DC: Intersociety Working Group, American Association for the Advancement of Science.

Alic, J. A., Branscomb, L. M., Brooks, H., Carter, A. B., Epstein, G. L. (1992), *Beyond Spinoff: Military and Commercial Technologies in a Changing World,* Boston: Harvard Business School Press.

Brooks, H. (1994), *The Relationship between Science and Technology,* Vol. 23, Pt. 5, *Research Policy,* Amsterdam: Elsevier, pp. 477–486.

Landau, R. Rosenberg, N. (Eds.), (1986), *The Positive Sum Strategy: Harnessing Technology for Economic Growth,* Washington, DC: National Academy Press.

NSF (1993), *National Science Board, Science and Engineering Indicators—1993,* Washington, DC: U.S. Government Printing Office (NSB 93-1).

OECD (1992), *Main Science and Technology Indicators 1992/2,* Paris: OECD.

RESISTORS

Ronald F. Dziuba, *Electricity Division, National Institute of Standards and Technology, Gaithersburg, Maryland, U.S.A.*

Introduction 423
1. **Electrical Resistance** 424
1.1 Definition of Resistance 424
1.2 Unit of Resistance 425
1.2.1 Quantized Hall Resistance 425
1.3 Importance of Resistors 425
2. **Resistor Fabrication** 426
2.1 Base Material 426
2.2 Resistive Element 427
2.3 Resistor Terminations and Terminals 427
2.4 Protective Enclosure 428
3. **Types of Resistors** 428
3.1 Carbon Composition Resistors 428
3.2 Wirewound Resistors 428
3.3 Metal-Foil Resistors 429
3.4 Thin-Film Resistors 429
3.4.1 Metal Film 429
3.4.2 Metal-Oxide Film 430
3.4.3 Carbon Film 430
3.5 Thick-Film Resistors 430
4. **Resistor Characteristics** 430
4.1 Accuracy 430
4.2 Stability 431
4.3 Power Rating 431
4.4 Temperature Coefficient of Resistance 431
4.5 Load Coefficient 432
4.6 Voltage Coefficient 432
4.7 Humidity Effects 432
4.8 Pressure Effects 432
4.9 Frequency Effects 432
5. **Classification of Resistors** ... 433
5.1 Standard Resistors 433
5.2 Resistors for Electronic Circuits 433
5.3 Integrated-Circuit Resistors ... 433
5.4 High-Current Resistors 434
5.5 High-Voltage Resistors 434
Glossary 434
Works Cited 435
Further Reading 435

INTRODUCTION

Electrical resistance is a fundamental property of a normal electrically conductive material or conductor by which it impedes the flow of electric current in a circuit, resulting in a voltage drop across the conductor and the generation of heat energy. The electrical resistance in a circuit, measured in ohms, is equal to the ratio of the voltage drop across the resistor measured in volts divided by the current measured in amperes. *Resistors* are a class of electrical components whose primary function is to introduce resistance in an electrical circuit for purposes of operation, protection, or control. They are used for voltage measurements, current measurements, setting voltage biases, controlling amplifier gains, setting time constants, matching and loading circuits, heat generation, temperature and humidity measurements, strain gauges, and other related functions.

The first resistors used by experimenters were constructed of iron, copper, or other pure-metal wires of arbitrary lengths and sizes. It soon became evident that a "standard" of resistance was needed to permit the intercomparison of results by different experimenters. In 1861 the British Association for the Advancement of Science appointed a committee to establish resistance standards (British Association for the Advancement of Science, 1913). The first resistance standard consisted of a coil of platinum–silver alloy wire sealed in a container filled with paraffin. It was known as the British Association unit and served as a standard for nearly two decades. Its disuse was a consequence of its large temperature coefficient of resistance (TCR). By 1900 wirewound resistance standards were constructed from manganin (Pe-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

terson, 1956), a copper–manganese–nickel alloy having substantially a zero TCR near room temperature. The next major development in wire-drawn resistance alloys occurred around 1948 with the discovery of Evanohm® (Peterson, 1956) by the Wilbur B. Driver Co., an alloy of nickel, chromium, aluminum, and copper. Like manganin it also has a small TCR; however, its resistivity is 2.3 times that of manganin. Today manganin and Evanohm® are still the alloys of choice for the construction of wirewound resistors—manganin for low-value resistors (<100 Ω) and Evanohm® for high-value resistors up to 100 MΩ.

The origin of non-wirewound resistors dates back to the time of World War I with their use in communication equipment (Coursey, 1949). Frequency errors of wirewound resistors resulting from series inductance and shunt capacitance made them unsuitable for use in high-frequency amplifiers. A need also arose for multi-megohm resistors that were not available as wirewound resistors for use in radio receivers. Carbon composition resistors were developed to satisfy these needs. Mixtures of carbon and resin are altered to produce resistance materials having a wide variety of resistivities. For the construction of resistors, either the mixture is extruded as a solid rod or it can be deposited on a glass substrate. These resistors can be produced in various physical sizes, resistance values, and power ratings.

Later, film-type resistors supplanted carbon composition resistors for general-purpose use because of better accuracies, stability, lower noise, and high-frequency performance. Film-type resistors are available with thin-film or thick-film resistive elements. The materials of choice for the construction of film-type resistors are pure metals, metal alloys, conductive oxides, and semiconductors. Film-type resistors have really come into their own in integrated-circuit and surface-mount technologies.

Around 1964 the metal-foil resistor was developed (Kaufman *et al.*, 1988). This new technology resulted in the production of resistors having smaller TCRs and better long-term stability than film-type resistors. The TCR, accuracy, and stability of these metal-foil resistors are approaching those of the best precision wirewound resistors.

The first section of this article defines resistance and the unit of resistance. It also describes the use of resistors in the measurement of other physical parameters. The following sections describe resistor fabrication, types of resistors, resistor characteristics, and classification of resistors according to use.

1. ELECTRICAL RESISTANCE

1.1 Definition of Resistance

The concept of electrical resistance is based on the experimental work of Georg S. Ohm, who in 1826 formulated the quantitative relationship between voltage and current in a dc electrical circuit. This relationship or "Ohm's law" expresses the fact that the magnitude of the current flowing in an electrical circuit depends directly on the electrical potential difference (voltage) and is inversely proportional to the property of the circuit known as the resistance (Wellard, 1960). In modern terminology his experimental equation for bulk metallic conductors can be expressed as

$$R = E/I,$$

where R is the resistance and E is the potential difference across the conductor when a current I flows. Resistance is a fundamental property of any conductor and can be expressed by the following integral expression:

$$R = \int \rho dl/A,$$

where l is the length, A the cross-sectional area, and ρ the resistivity of the conductor. For thin-film resistors it is convenient to define a quantity ρ_s, called the sheet resistance, which is equal to ρ divided by the film thickness d. A thin-film resistor consisting of a simple rectangle of length l (in the direction of current) and width w has a resistance of

$$R = (\rho/d)(l/w) = \rho_s(l/w).$$

The ratio l/w is usually equated to the number of squares of side w. Then ρ_s has the units ohms per square (Ω/□), since "squares" is a pure number. Thus, the value of a thin-film resistor is equal to ρ_s multiplied by the number of squares.

The power P (measured in watts) dissipated in a resistor may be expressed by any one of the following expressions:

$$P = EI = I^2R = E^2/R,$$

where electrical energy is irreversibly converted into heat.

1.2 Unit of Resistance

In 1862 the BA Committee adopted the name "ohm" for the unit of resistance (British Association for the Advancement of Science, 1913). The following year the committee adopted the "absolute" system of units, which provided compatibility between the mechanical units and electrical units. Throughout the years many different systems of units have been in use, which culminated in 1960 with the adoption of the International System of Units (SI), which since then has been accepted internationally (Harris, 1962). In SI units the ohm is defined in terms of the mechanical units of length, mass, and time—the meter (m), kilogram (kg), and second (s), respectively—and the unit of electric current, the ampere (A). The ohm (Ω) is given by the definition

$$1\ \Omega = 1\ \mathrm{m}^2\ \mathrm{kg}\ \mathrm{s}^{-3}\ \mathrm{A}^{-2}.$$

The ohm can be realized by means of a calculable capacitor whereby its impedance, determined from length measurements and the velocity of light, is compared with a resistance through a complex chain of ac and dc bridges.

Commercial, industrial, and scientific requirements for the long-term repeatability and worldwide consistency of measurements of resistance often exceed the accuracy with which the SI unit of the ohm can be readily realized. To meet these severe demands, it has become necessary to establish a "representation" of the ohm that has superior long-term reproducibility and constancy as compared with the present direct realization of the ohm.

1.2.1 Quantized Hall Resistance By international agreement, the representation of the ohm has been based on the quantum Hall effect (QHE) since 1 January 1990 (Hartland, 1992). The QHE occurs at the interface between two semiconducting layers in suitable MOSFET or heterostructure devices when placed in a large applied magnetic field greater than 1 T and cooled to a temperature of 4.2 K or below. Under these conditions, if a constant current of less than 100 μA flows between the source and drain electrodes, plateau regions of constant Hall voltage and hence Hall resistance occur as a function of magnetic field or gate voltage depending upon the type of semiconductor device. These plateaus as shown in Fig. 1 occur at values of the Hall resistance R_H given by

$$R_H(I) = V_H/I = R_K/I,$$

where V_H is the Hall voltage, I the current through the device, I refers to the ith plateau, and R_K is the von Klitzing constant and is believed to be equal to h/e^2, h being the Planck constant and e the elementary charge. The voltage V_x along the channel is essentially zero at a plateau site, as shown in Fig. 1. The value of the von Klitzing constant adopted internationally beginning 1 January 1990 is equal to

$$R_{K-90} = 25\,812.807\ \Omega.$$

This conventional value is believed to be consistent with the SI ohm to within 0.005 Ω, corresponding to a relative uncertainty of 0.2 ppm. It is based on direct determinations of R_K using calculable capacitors and indirect measurements involving combinations of fundamental constants that yield h/e^2 independently of R_K. Still there is some small doubt concerning the equality of $R_K \equiv h/e^2$, which can only be resolved with more precise direct measurements of R_K. However, the ohm can be reproduced by means of the QHE to about 0.001 ppm.

1.3 Importance of Resistors

Resistors are important in the measurement of other electrical quantities. A large number of physical and chemical phenomena are investigated by means of measurements of voltages, and the measurement of voltage is customarily carried out by the measurement of resistance ratios. Electric

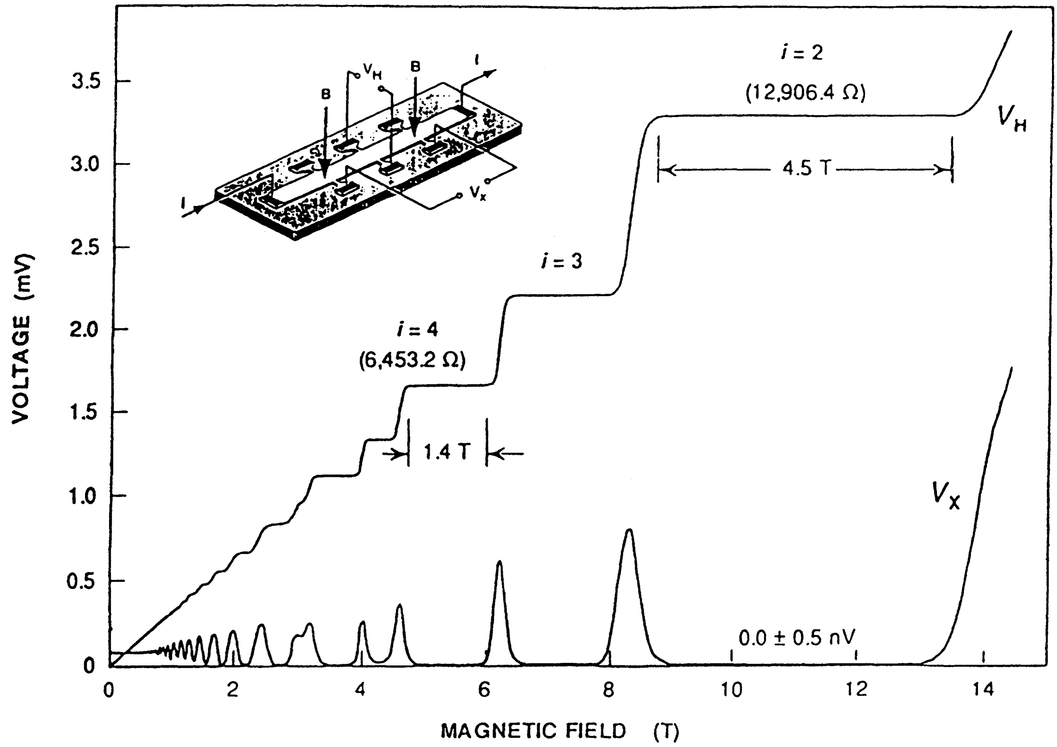

FIG. 1. The quantum Hall effect showing plateaus of constant Hall resistance with values $R_H(i)$.

current is readily most conveniently measured in terms of a voltage across a known resistor. Multimegohm resistors are used in the measurement of low currents extending downward to the picoampere range. Special low-valued resistors, called shunts, are used to measure high currents extending upward to the kiloampere range.

Resistors constructed out of pure metals such as copper or platinum have a high positive TCR and are used in the measurement of temperature. Thermistors are a special type of resistor with a very high negative or positive TCR and are also widely used to measure temperature. Changes in resistance with dimensions are utilized for measuring small displacements, and change in resistance with stress is utilized for the measurement of liquid pressures and force.

2. RESISTOR FABRICATION

The main factors involved in the selection of a resistor for a particular application are cost, resistance value, accuracy, power rating, time stability, frequency dependence, TCR, voltage coefficient of resistance (VCR), and the influence of other environmental factors such as humidity and pressure. Many of these factors can be minimized or optimized during the resistor fabrication process. The constructional features of a resistor can be separated into four general categories: base material, resistive element, terminations, and protective enclosure (Wellard, 1960).

2.1 Base Material

The base material for the resistive element must be a good insulator and have good mechanical stability over the range of ambient conditions to which the resistor is subjected. Common base materials for precision wirewound resistors are usually in the form of insulated metal cylinders, thin mica cards, or ceramic spools. For film-type resistors it is more important to consider matching the thermal-expansion coefficient of the base or "substrate" material with that of the closely adhering resistive element. Fused-sil-

Table 1. Typical electrical properties of some resistance materials.

Material	Composition (%)	ρ (20 °C) ($\mu\Omega$ cm)	TCR (20 °C) (ppm/K)
Carbon	C	3500	−500
Constantan	Cu 60, Ni 40	49	0 ± 20
Copper	Cu	1.7	+3900
Evanohm®[a]	Ni 75, Cr 20, Al 2.5, Cu 2.5	110	0 ± 5
Gold	Au	2.4	+3400
Manganin	Cu 84, Mn 12, Ni 4	48	0 ± 15
Nichrome®[b]	Ni 80, Cr 20	110	+140
Palladium–silver	Pd 60, Ag 40	42	+20
Platinum	Pt	10	+3000
Silver	Ag	1.6	+3800
Tantalum	Ta	13.5	+3100
Zeranin®[c]	Cu 92, Mn 7, Ge 1	43	0 ± 10

[a] Registered trademark, Wilbur B. Driver Company.
[b] Registered trademark, Driver Harris Company.
[c] Registered trademark, Isabellenhütte.

ica glass or alumina ceramic materials are generally used as the substrate material for thin-film, thick-film, and metal-foil resistors because of their excellent electrical and thermal properties.

2.2 Resistive Element

The main considerations in the design and construction of the resistive element are

1. the type of resistance material, and
2. the geometry of the resistance material as applied to the base material.

The material chosen for the resistive element depends on the type of resistor and its application. The main requirements are that the material have suitable resistivity and be sufficiently stable so that any changes in resistance value that occur during its operating life may be expected to fall below some prespecified value. Finally, the process used to construct the resistive element with the chosen material must be such that the final resistor can be made to meet its specifications at a reasonable cost. Tables 1 and 2 list electrical properties of some resistance materials used in the construction of wirewound and film-type resistors. Section 3 describes in more detail considerations of the geometry of the resistance material as applied to the base material.

2.3 Resistor Terminations and Terminals

Terminations are applied to the resistive element to define the exact value of resistance, while terminals provide the mechanical means by which the resistive element is connected into the electrical circuit. Termination is usually accomplished by soldering or welding lead wires to the ends of the resistive element. This is easily accomplished for wirewound elements but is more complicated for film-type elements. Termination for resistive films is commonly made by depos-

Table 2. Typical electrical properties of thin-film and thick-film resistance materials.

Type	Resistance range	TCR (ppm/K)	VCR (ppm/V)	ρ_s ($\Omega/\square$)
Carbon film	4 Ω to 20 MΩ	−250 to −400	<500	50 to 5×10^6
Cermet	10 Ω to 5 MΩ	−550 to +150	<100	10 to 10^5
Tin oxide	10 Ω to 100 kΩ	−280 to +140	<25	20 to 400
Nickel–chromium	15 Ω to 150 kΩ	−300 to +350	<1	25 to 300
Tantalum	15 Ω to 150 kΩ	−100 to +100	<1	25 to 300
Tantalum–nitride	15 Ω to 150 kΩ	−25 to +25	<1	25 to 300
Palladium, silver glaze	10 Ω to 1 MΩ	−250 to +250	<1	10 to 10^4

iting noble-metal (gold or silver) contact pads or using noble-metal conducting paint at the ends of the element. Leads then can be attached by compression, welding, or soldering. For carbon composition resistors, the leads are embedded within the resistive material.

2.4 Protective Enclosure

A resistor enclosure provides mechanical and environmental protection to the resistive element. The protective means usually takes one of three forms:

1. an enclosure usually of metal with insulating feedthroughs, ceramic, or glass,
2. a coating applied in liquid form and then cured, or
3. a molding in a suitable encapsulating material.

3. TYPES OF RESISTORS

3.1 Carbon Composition Resistors

The carbon composition resistor is perhaps the most widely used resistor in discrete circuits because of its low cost, high reliability, and small size. Basically, the resistive element is a mixture of carbon and a suitable binder. It is molded under high temperature and pressure into a cylinder with embedded solder-coated wire leads and an insulating plastic or ceramic jacket. Different resistance values are obtained by varying the carbon and filler content. Table 3 lists the standard color code (Wellard, 1960) for composition and some axial-type resistors. Different color bands are used to designate the resistance value and accuracy.

3.2 Wirewound Resistors

The base materials for precision wirewound resistors are usually in the form of insulated metal cylinders, thin mica cards, or ceramic spools. The metal cylinders are coated with an insulating enamel or varnish, and the resistance wire is wound in close thermal contact with the metal cylinder in order to dissipate readily any heat generated by the wire. These metal cylinders are ordinarily of brass, which has a coefficient of thermal expansion nearly the same as that of the wire resistance alloys. This is important to avoid small stresses in the wire introduced by temperature changes. Such stresses can change both the resistance value and its stability.

The mica-card winding form also provides a coefficient of thermal expansion compatible with the resistance wire to minimize temperature effects. The thin profile of these cards produces resistors with very thin cross section, thereby reducing inductance effects of wire loops and making the most efficient use of space. Most ceramic spools have poor heat conductivity and their temperature coefficients of thermal expansion are considerably smaller than for the resistance wires. If ceramic spools are used in the construction of precision resistors, the resistance wire is loosely wound on the spool and the structure sealed in close contact with a thermally con-

Table 3. Resistor color code.

Color of band	1st band, 1st significant figure	2nd band, 2nd significant figure	3rd band, multiplier	4th band, tolerance
Black	0	0	1	···
Brown	1	1	10	±1%
Red	2	2	10^2	±2%
Orange	3	3	10^3	···
Yellow	4	4	10^4	···
Green	5	5	10^5	···
Blue	6	6	10^6	···
Violet	7	7	10^7	···
Gray	8	8	10^8	···
White	9	9	10^9	···
Silver	···	···	10^{-2}	±10%
Gold	···	···	10^{-1}	±5%
None	···	···	···	±20%

ducting insulating fluid such as mineral oil or silicone fluid. The spools for general-purpose wirewound resistors are usually made out of ceramic, epoxy, steatite, or a vitreous material.

Almost all resistance alloys used for constructing wirewound resistors have either a copper or a nickel base (see Table 1). Manganin alloy, whose composition is approximately 84% Cu, 12% Mn, and 4% Ni, has a resistivity of about 48 $\mu\Omega$-cm and a thermal emf to copper of <3 μV/K, and is widely used to construct low-value resistors of nominal values below 100 Ω. Thermal emf refers to the Seebeck thermoelectric effect whereby a voltage is developed when the junctions of a circuit of two dissimilar metals are maintained at different temperatures (Harris, 1962). The TCR of manganin can be reduced to zero at a temperature near room temperature by proper heat treatment; however, the curvature of its resistance–temperature curve at room temperature is approximately -0.5 ppm/K^2. It is subject to surface oxidation and has a pressure coefficient of resistance (PCR) of approximately $+2.3$ ppb/hPa, where ppb refers to parts per billion (10^9).

For the construction of wirewound resistors of nominal values of 100 Ω or higher, Evanohm® is the usual alloy of choice. Its nominal composition is 75% Ni, 20% Cr, 2.5% Al, and 2.5% Cu. It has a resistivity of 110 $\mu\Omega$-cm and a thermal emf to copper of <1 μV/K. The resistance–temperature curve for Evanohm® is much flatter than that for manganin; its curvature at room temperature is approximately -0.05 ppm/K^2. Heat treatment of Evanohm® changes its temperature and time stability. A prolonged heat-treatment process of temperature cycling and soaking time is necessary in order to condition the alloy for zero TCR at a selected temperature and long-term stability. Its PCR is approximately -1.1 ppb/hPa and opposite in sign of the PCR for manganin.

The limitation of wirewound resistors at high-frequency operation is their inherent reactance that is developed as a result of the "coil" type of construction. A number of winding techniques have been developed to reduce the residual inductance and capacitance of wirewound resistors. The most common are the bifilar, thin-card, Ayrton–Perry, and reverse-pi winding methods (Wellard, 1960). The bifilar winding is accomplished by bending the wire back onto itself at its midpoint to form a long loop or "hairpin" with the two sides as close together as the wire insulation will permit. This technique reduces the inductance; however, it may result in a large capacitive effect if the loop is long. Consequently, this technique is used for short loops or low-valued resistors. Winding the resistance wire on a thin card made of mica or other insulating material produces low residual inductance due to the presence of currents in opposite direction separated only by the thickness of the card. Capacitive effects are small since the starting and finishing ends of the winding are at opposite ends of the card. The card form of construction is sometimes used with two windings in parallel, wound in opposite directions around the card, with one winding spaced between the turns of the other. This method, called the Ayrton–Perry winding, gives better cancellation of magnetic fields than the single winding. The reverse-pi method consists of windings in an even number of sections or "pi's" on a bobbin resistor form. As each section is filled, the direction of winding is reversed in the next section, so as to help cancel the inductive effects.

3.3 Metal-Foil Resistors

In the construction of this type of resistor, a layer of thin bulk metal (foil) several microns thick is cemented to a glass or ceramic substrate (Kaufman *et al.*, 1988). The foil is etched using a pattern designed for low residual reactance. The foil is compressed slightly by the substrate as a result of their unequal coefficients of expansion. The compressed foil has a negative TCR that cancels out the inherent positive TCR of the foil. As a result, the metal-foil resistor has a TCR close to zero. Since the foil has a thickness of several microns instead of the extremely thin film of a film resistor, the natural stability of the alloy is preserved, resulting in a resistor having better long-term stability.

3.4 Thin-Film Resistors

3.4.1 Metal Film Materials in thin-film form have higher resistivities than the bulk material because of the additional resistivity due to scattering of the conduction electrons at the boundary of the film (Maissel, 1970). The films are deposited on the substrate by either vacuum evaporation or cathode sputtering.

Nickel–chromium alloy is generally used in the manufacture of thin-film resistors, following logically from its use in bulk form for wirewound resistors. The properties of these films greatly depend on the film composition, which can vary over wide limits depending on deposition parameters such as substrate temperature, source temperature, vacuum, or rate of deposition. Good stable films with TCRs ranging from 5 to 100 ppm/K are obtained by selection. The films have poor stability under conditions of high humidity and therefore must be fully sealed against the entry of moisture.

Single-metal films are often used as resistive elements to avoid problems of composition control such as those occurring in the deposition of alloys such as nickel–chromium. Tantalum films are widely used because of their relatively high resistivities and low TCRs as compared with other pure-metal films and the tough, self-protective oxide that tantalum forms during heat treatment in oxygen. This oxide has excellent dielectric properties; this has led to the development of tantalum capacitors.

3.4.2 Metal-Oxide Film The most popular metal-oxide or cermet film (Maissel, 1970) used in the manufacture of resistors is chromium–silicon monoxide (Cr–SiO). This film is produced by the coevaporation of Cr and SiO and features high resistivity and stability with a TCR on the order of 100 ppm/K. The electrical resistivity can be varied over several orders of magnitude by varying the Cr content.

Tin-oxide (SnO_2) films (Maissel, 1970) used as resistive elements are formed by spraying the material in vapor form onto a heated glass or ceramic substrate. The reaction results in an extremely adherent, fully oxidized film. SnO_2 films have high resistivity and are very stable and rugged. They can withstand temperatures as high as 450 °C without deterioration. One interesting feature of SnO_2 films is that they possess a high degree of transparency. They have therefore found wide application in areas such as the manufacture of so-called “conducting glass” and as heating elements.

3.4.3 Carbon Film Carbon-film resistors (Grisdale *et al.*, 1951) have supplanted carbon composition resistors for general-purpose use because of lower cost, better tolerances, increased stability, lower noise, and better high-frequency performance. They are fabricated by pyrolytic decomposition of carbon on the surface of a cylindrical ceramic substrate. The characteristics of the resistors are sensitive to the deposition conditions, especially film thickness. During fabrication the resistance value of a carbon-film resistor is adjusted by a procedure known as spiraling, i.e., a thin grinding wheel is used to cut a groove through the film along a helical path to increase the resistance.

3.5 Thick-Film Resistors

Mixtures of metal, metal compounds, glass, and solvents are commonly used for the manufacture of resistive elements (Dummer, 1970). The usual method of applying these thick films on a flat or cylindrical alumina substrate is by dipping or rolling. The substrate with the liquid glaze (e.g., tantalum, tantalum nitride, glass, and a carrier) is fired at approximately 1000 °C to fuse the glass particles and bind the film to the substrate. This technology provides a resistive element that is impervious to environmental conditions without the need for an air-tight encapsulation. The inherent ruggedness of this glaze allows it to absorb higher voltage surges and overloads than a thin-film counterpart.

4. RESISTOR CHARACTERISTICS

A summary of key characteristics of different resistor types is listed in Table 4.

4.1 Accuracy

All resistors have a specified initial accuracy or tolerance. It expresses the maximum deviation in resistance from its nominal value in either percent or ppm. Carbon composition and some other film-type resistors use a color code to denote tolerances as listed in Table 3. Tolerances are typically 5% to 20% for carbon composition resistors, 0.5% to 10% for carbon-film resistors, 0.1%

Table 4. Typical resistor characteristics.

Characteristic	Wirewound (precision)	Metal foil	Metal film	Carbon film	Carbon composition
Resistance range	0.1 mΩ to 100 MΩ	0.5 Ω to 1 MΩ	1 Ω to 10 MΩ	1 Ω to 10 TΩ	1 Ω to 100 MΩ
Power dissipation	10 mW to 100 W	0.5 to 2 W	50 mW to 2 W	0.1 to 2 W	$\frac{1}{8}$ to 2 W
Tolerance	2 ppm to 1%	50 ppm to 1%	0.05% to 1%	0.5% to 10%	5%, 10%, 20%
Frequency limit	50 kHz	100 MHz	400 MHz	100 MHz	1 MHz
TCR (ppm/K)	±1 to ±20	0 ± 2.5	±20 to ±200	−200 to −1000	−800 to +1600
VCR (ppm/V)	0	<0.1	<1	−10	−50
Stability	Best	Good	Good	Moderate	Poorest
Advantages	Highest accuracy and stability	Low TCR, low VCR, high speed	Best frequency response	High resistance and low cost	Low cost and small size

to 1% for metal-film resistors, 0.01% to 1% for precision wirewounds, and 0.0001% (1 ppm) to 0.01% (100 ppm) for ultraprecision or standard resistors (Wellard, 1960).

4.2 Stability

Stability refers to the change in resistance value with time or environmental stress. Stresses include high- or low-temperature exposure, application of full rated power, short-time overload, moisture, soldering heat, or radiation exposure. Carbon composition resistors generally have the poorest stability; wirewounds are the most stable, followed by metal-foil, metal-film, and carbon-film resistors.

4.3 Power Rating

The power rating of a resistor is defined as the maximum specified wattage that can be continuously dissipated at a maximum specified temperature without damage to the resistor. The choice of materials in the construction of the resistor along with its physical size determines its power rating. An important consideration is the temperature rise of the resistor. If the resistor is mounted so that it is close to other heat-producing components or has restricted ventilation, the power dissipation of the resistor may have to be reduced to avoid reaching the maximum allowable temperature.

4.4 Temperature Coefficient of Resistance

All resistance materials exhibit some change in resistance with temperature, and the magnitude of this change is usually expressed in %/°C or ppm/°C. For general-purpose resistors over a limited temperature range, this change in resistance with temperature is treated as a linear function given by the relationship

$$R = R'[1 + \alpha_0(t - t')],$$

where R and R' are the resistance values at temperature t and reference temperature t', respectively, and α_0 is the mean TCR expressed in proportional parts per °C over the temperature range of operation (Harris, 1962). For most materials the resistance–temperature curve is not linear; therefore, for precision resistors, especially wirewounds, this relationship is more accurately expressed as

$$R = R'[1 + \alpha(t - t') + \beta(t - t')^2],$$

where α is the slope of the resistance–temperature curve at t' and β is the curvature at any temperature over the range of operation.

The TCRs of film-type resistors (Maissel, 1970) can span a wide range, as shown in Table 2. The main factors that cause this effect are the type of substrate, film thickness, film composition, film structure, and important parameters during deposition such as substrate temperature. Resistance in bulk materials results from electron scattering by phonons, impurities, and defects. In thin-film materials, surface boundaries cause additional scattering effects that influence the resistivities and TCRs of these films. Typically, thin films have lower TCRs compared with bulk materials primarily because of the high resistivity of thin films.

4.5 Load Coefficient

The load coefficient of a resistor is defined as the change in resistance caused by the production of joule heating (I^2R) in the resistor at the rate of one watt and is usually expressed in ppm/W. This change, however, is a function of time. When a current starts flowing in a resistor the resistance material (wire or film) very quickly takes up a temperature above that of the base material in which it is mounted. The amount of this change depends on the thermal contact between the resistance and base materials, thermal time constant of the resistance and base materials, and TCR of the resistance material, and also upon the changes in stress in the resistance material and its support. The load coefficient of a resistor is a rather indefinite quantity, varying with time of flow of current and with the environment of the resistor.

4.6 Voltage Coefficient

When voltage is applied to a carbon or film-type resistor there may be a decrease in resistance apart from changes caused by loading of the resistor (Wellard, 1960). This decrease in resistance with increase in voltage results from the breakdown of contact resistances between carbon particles in carbon composition resistors or the breakdown of boundary resistances between islands of material on substrates of film-type resistors. This change in resistance per applied volt is called the voltage coefficient of resistance. In carbon composition resistors, the change in resistance value due to applied voltage is usually -200 ppm/V or less. For carbon-film resistors, the VCR can vary from -10 to -50 ppm/V. Metal-film resistors have VCRs from -1 to -30 ppm/V depending on their wattage, whereas metal-oxide–film resistors have VCRs from -1 to -5 ppm/V. Wirewound resistors do not exhibit a VCR.

4.7 Humidity Effects

Unless the resistive element is hermetically sealed in a glass or metal container with glass-to-metal seals, moisture can permeate a resistor and change its resistance value. Moisture can produce two reversible effects:

1. On the surface of a high-value resistor or its housing it can provide a path for leakage currents and thus lower the apparent resistance.
2. If moisture is absorbed through the insulator protecting the resistive element, such as the enamel coating of resistance wire, or the jacket of a carbon composition resistor, or the overglaze film of film-type resistors, the insulator will swell, resulting in pressure being exerted on the resistive element. This effect usually causes the resistance to increase and is directly proportional to the insulator thickness and inversely proportional to the thickness of the resistive element (Starr *et al.*, 1970). The increase can vary from 0.1% to 10% for a 50% change in relative humidity depending on the type of resistor.

4.8 Pressure Effects

Resistance materials are affected by changes in pressure (Meaden, 1965). These changes of resistance as a function of pressure are not well understood, and, depending on the material, resistors may have either a positive or negative pressure coefficient of resistance. Fortunately, for well-constructed resistors these changes of resistance with pressure are only significant for those of the highest quality. For pressure changes of 500 hPa, the change in resistance would be within 1 ppm.

4.9 Frequency Effects

A resistor may be represented to a first approximation by the circuit of Fig. 2, which shows a pure dc resistance R in series with

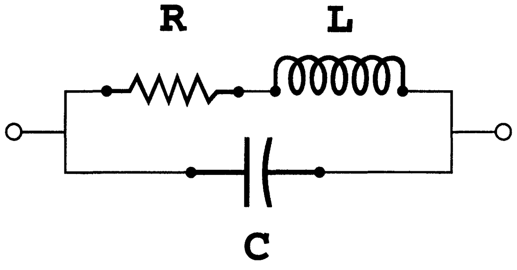

FIG. 2. Simplified model of a resistor.

an inductance L and this combination shunted by a capacitance C (Harris, 1962). The net equivalent inductance is given by

$$L' = L - CR^2.$$

The overall inductive effect of a resistor is reduced by its shunt capacitance. The time constant of a resistor is defined as

$$\tau = L'/R = L/R - CR.$$

The time constant of a resistor may be either positive or negative, depending on which of the terms, L/R or CR, predominates. For a well-designed ac resistor, the value of the time constant should be 10^{-7} or less. Low-valued resistors generally have a positive time constant, which means a net inductive effect, while high-valued resistors have a negative time constant or a net capacitive effect. The crossover point for wirewound resistors is around 100 Ω. Film-type resistors have the best high-frequency performance. The effective dc resistance of these resistors for most resistance values remains fairly constant up to 100 MHz and then decreases at higher frequencies (Hierholzer *et al.*, 1977).

For high-accuracy requirements other frequency effects need to be considered. These include dielectric losses, losses arising from eddy currents induced in nearby conductors, and losses due to skin effect. The skin effect (Harris, 1962) results from a decrease in the current density in the interior of the resistive element and an increase toward its surface as the frequency increases. The effective resistance increases and is a function of the resistivity of the material and the geometry of the resistive element.

5. CLASSIFICATION OF RESISTORS

The most common classification of resistors is according to their intended use and inherent performance.

5.1 Standard Resistors

Standard resistors (Harris, 1962) are used for calibration purposes in resistance measurements. They are usually commercially available only in decimal multiples from 0.01 mΩ to 10 TΩ. Standard resistors of nominal values to 100 MΩ are constructed of wire or strip resistance material. Manganin is the material of choice for resistors up to 10 Ω, and Evanohm® material is usually used in the construction of resistors above 10 Ω to 100 MΩ. Above 100 MΩ, the best available resistors are of the thin–metal-film construction. Standard resistors 100 Ω and below are constructed as four-terminal types to eliminate the problems of lead and contact resistances and to define the resistance more precisely. The important requirements of standard resistors are stability with time, low TCR, small thermal emfs against copper, well-defined terminations, and ruggedness. Standard resistors are adjusted to 0.01% or better from nominal value for wirewound resistors, and from 1% to 10% from nominal value for metal-film resistors.

5.2 Resistors for Electronic Circuits

The majority of resistors manufactured are intended for use in electronic circuits. These include resistors of all types of construction as listed in Sec. 3. For critical low-frequency circuits wirewounds are still in use; otherwise, metal-film resistors are used because of their excellent high-frequency performance. The other important requirements include small size, high reliability, and low cost. Resistor tolerances vary from 0.01% to 20% depending on their use. These resistors are intended for use in circuits involving potentials up to several hundred volts, but currents seldom over 10 to 100 mA.

A variety of thin- and thick-film resistor chips are available for hybrid microelectronic circuits (Hierholzer *et al.*, 1977). They measure as little as 1 mm by 1 mm, and have resistance values from a few ohms to 1 GΩ. A typical power rating is 0.25 W or less, with TCRs in the order of 20 to 200 ppm/K.

5.3 Integrated-Circuit Resistors

Integrated-circuit resistors (Fogiel, 1972) that lie in the range 20 Ω to 30 kΩ are generally formed by the diffusion of *p*-type material in an *n*-type epitaxy layer. Their tolerances, however, are poor, being in the order of ±25%. The tolerance of the ratio of two diffused resistors can be as low as ±2%. Diffused resistors are not desirable for higher values, since large chip areas are required. Above 30 kΩ, resistors can be formed from ion-implanted layers. These are quite reproducible, have a large dynamic range, and are linear.

5.4 High-Current Resistors

High-current resistors or shunts (Harris, 1962) are used for the measurement of direct current in the range from 10 A to 10 kA or higher. They are four-terminal resistors made from sheets or rods of manganin, using several elements in parallel for the lower-resistance, high-current values. They are designed for operation either in an air environment or mounted in an oil-filled container equipped with a stirring mechanism. The range of resistance is from about 10 μΩ to 10 mΩ. Because of the high power dissipation and the less well-defined thermal conditions, the accuracy obtainable is limited to 0.01% at the low-current values, and degrades to about 0.1% for high-current values.

5.5 High-Voltage Resistors

High-voltage resistors (Wellard, 1960) are designed to fulfill the special requirement for high-voltage, high-resistance units capable of dissipating moderate power. These resistors are rated from 5 to 20 kV, have a resistance range from 2 kΩ to 1 GΩ, and are rated from 5 to 20 W. These resistors are noninductive and are used primarily in high-voltage bleeder circuits, high-voltage dividers, and high-voltage networks.

GLOSSARY

Ayrton–Perry Winding: A resistor winding method where alternate winding layers are wound in opposite directions to reduce inductance by causing electrical fields to cancel.

Bifilar Winding: A resistor winding method in which a wire is folded in half into a long loop or "hairpin" with the two sides as close together as possible and then wound side by side.

Cermet: A resistance alloy consisting of a metal oxide, usually chromium–silicon monoxide (Cr–SiO).

Evanohm®: A resistance alloy of nickel, chromium, aluminum, and copper, developed by the Wilbur B. Driver Co. in 1948, having a low TCR and a resistivity 77.5 times that of copper.

Load Coefficient: The change in resistance caused by the production of joule heating in a resistor at the rate of one watt and usually expressed in ppm/W.

Manganin: A resistance alloy of copper, manganese, and nickel, developed around 1890, having a low TCR and a resistivity 27.8 times that of copper.

Ohm: The unit of resistance required to cause a voltage drop of 1 V with a flow of 1 A of current.

ppm: Parts per million (10^6), which is also 0.0001%.

Quantized Hall Resistance (QHR): A reference based on the QHE that defines a base resistance believed to be equal to the ratio of the fundamental constants h/e^2 and other quantized resistances equal to submultiples of h/e^2.

Quantum Hall Effect (QHE): An observed property at low temperatures in special microelectronic samples in which a two-dimensional electron gas exists and energy states are quantized when subjected to a strong magnetic field.

Resistance: The property of a conductor that causes it to impede the flow of electrons.

Resistance Alloy: A material, with properties of moderate to high resistivity, low temperature coefficient of resistance, and long-term stability, used in the construction of resistors.

Resistivity: The resistance per unit length of a material with uniform cross section.

Shunts: Special class of resistors of nominal values from about 10 μΩ to 10 mΩ designed to carry large currents in the range from 10 A to kA or higher.

Stability: The degree to which the resistance changes with time or environmental stress.

Standard Resistors: The highest quality of resistors used for calibration purposes in resistance measurements.

Temperature Coefficient of Resistance (TCR): The change in resistance per a degree change in temperature expressed in ppm/K or ppm/°C.

Thermal emf: The voltage generated when the junctions of a circuit of two dissimilar metals are maintained at different temperatures.

Voltage Coefficient of Resistance (VCR): The change in resistance as a function of voltage expressed as ppm/V.

Works Cited

British Association for the Advancement of Science (1913), *Reports of the Committee on Electrical Standards,* Oxford, U.K.: Cambridge Univ. Press.

Coursey, P. R. (1949), "Fixed Resistors for Use in Communication Equipment," *Proc. IEE* **96,** Pt. III, No. 41, 169–186.

Dummer, G. W. A. (1970), *Materials for Conductive and Resistive Functions,* New York: Hayden.

Fogiel, M. (1972), *Modern Microelectronics,* New York: Research and Education Association.

Grisdale, R. O., Pfister, A. C., van Roosbroeck, W. (1951), "Pyrolytic Film Resistors: Carbon and Borocarbon," *Bell System Tech. J.* 271–314.

Harris, F. K. (1962), *Electrical Measurements,* New York: Wiley.

Hartland, A., (1992), "The Quantum Hall Effect and Resistance Standards," *Metrologia* **29,** 175–190.

Hierholzer, E. L., Drexler, H. B. with Powers, J. H. (1977), "Resistors and Passive-Parts Standardization," in C. A. Harper (Ed.), *Handbook of Components for Electronics,* New York: McGraw-Hill.

Kaufman, M., Seidman, A. H., Sheneman, P. J. (1988), *Electronics Source Book for Technicians and Engineers,* New York: McGraw-Hill.

Maissel, L. I. (1970), "Thin-Film Resistors," in: L. I. Maissel, R. Glang (Eds.), *Handbook of Thin-Film Technology,* New York: McGraw-Hill, Chap. 18.

Meaden, G. T. (1965), *Electrical Resistance of Metals,* New York: Plenum.

Peterson, C. (1954), "Alloys for Precision Resistors," in: *Precision Electrical Measurements,* New York: Philosophical Library.

Starr, C. D., Schlenker, J., Graule, R. (1970), "Evaluating Moisture Resistance of Magnet Wire Enamels," *Insulation* **16** (2), 29–31.

Wellard, C. L. (1960), *Resistance and Resistors,* New York: McGraw-Hill.

Further Reading

Berry, R. W., Hall, P. M., Harris, M. T. (1968), *Thin-Film Technology,* New York: Van Nostrand Reinhold, Chap. 7.

Gore, T. S., Jr. (1989), "Resistors," in D. G. Fink, D. Christiansen (Eds.), *Electronics Engineers' Handbook,* New York: McGraw-Hill.

Marsten, J. (1962), "Resistors—A Survey of the Evolution of the Field," *Proc. IRE* **50,** 920–924.

Meeldijk, V. (1995), *Electronic Components—Selection and Application Guidelines,* New York: Wiley.

Prange, R. E., Girvin, S. M. (1990), *The Quantum Hall Effect,* New York: Springer-Verlag.

Thomas, J. L. (1948), "Precision Resistors and Their Measurement," National Bureau of Standards Circular 470, Washington, DC: U.S. Government Printing Office.

RESONANT TUNNELING

E. E. Mendez, *IBM Research Division, T. J. Watson Research Center, Yorktown Heights, New York, U.S.A.*

L. Esaki, *University of Tsukuba, Tsukuba, Ibaraki, Japan*

1. **Introduction** 437
2. **Quantum-Mechanical Tunneling** 439
2.1 Tunneling and the WKB Approximation 439
2.2 Tunneling in Semiconductor Structures 439
2.3 Tunneling Current 441
3. **Resonant Tunneling in Semiconductor Heterostructures** 441
3.1 Resonant Tunneling Probability 441
3.2 Resonant Tunneling Current 443
3.3 Charge Accumulation and Bistability 444
3.4 Resonant Tunneling in Type I Heterostructures 445
3.5 Band-Structure Effects 446
3.6 Resonant Tunneling of Holes 447
3.7 Interband Resonant Tunneling 447
3.8 Magnetic-Field Effects 447
3.9 Low-Dimensionality Resonant Tunneling 448
4. **Applications of Resonant Tunneling** 450
4.1 High-Frequency Oscillators ... 450
4.2 Other Applications 451
5. **Summary and Prospect** 452
Glossary 452
Works Cited 453
Further Reading 454

1. INTRODUCTION

Tunneling phenomena are strictly quantum mechanical, with no counterpart in classical mechanics. Their consequences, however, are easily observable even at room temperature and are at the core of new applications in electronics. In this article we describe those effects, focusing on the most spectacular of them—resonant tunneling.

When subjected to a repulsive force, a classical particle moving in one dimension at an initial speed v is slowed down and the direction of its motion can eventually be reversed if the force is sufficiently strong. In this case, the particle is turned back at a point that depends on v and on the intensity of the force. However, in the realm of quantum mechanics, in which the position of an object is described by the probability of finding it there at a given time, there is a finite probability of finding the object beyond the classical turning point.

Using a mathematical language in which a force is defined in terms of the gradient of a potential energy, we can say that quantum mechanically it is possible for an object to tunnel through a potential barrier (repulsive potential) higher than the total energy of the object, provided the width and height of that barrier are finite. In spite of the counterintuitiveness of this concept, there are numerous phenomena in nature successfully explained by tunneling, e.g., the electric-field–induced ionization of hydrogen or the alpha decay of heavy nuclei. Moreover, the concept of tunneling is the foundation of inventions like the electron-field–emission and scanning tunneling microscopes, the Esaki diode, and the Josephson junction.

The fact that in quantum mechanics a particle is described by a wave function makes it possible to draw analogies between the concept of tunneling and other undulatory phenomena. One of these analogies can be found in the exponential attenuation of an electromagnetic wave propagating from a medium with index of refraction n_1 to an-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

other of index $n_2 < n_1$. If the incident angle of the electromagnetic wave at the interface between the two media exceeds a critical value, the intensity of the wave in the second medium decays exponentially with the distance perpendicular to the interface. The wave is totally reflected if that medium is infinite, but if its thickness is comparable to the wavelength then a small fraction of the wave is transmitted, as first observed by Newton in the transmission of light between two glass plates separated by a thin air gap. The analogy with tunneling arises mathematically from the similarity between the equation for e.m. waves propagating through a thin slab and Schrödinger's wave equation for an object in a potential barrier.

Although the idea of tunneling was already used in the late 1920s to explain atomic- and nuclear-physics effects and was extensively studied in the 1950s and 1960s in relation with phenomena in solid-state physics, it has been with the advent of engineered semiconductor heterostructures and of the scanning tunneling microscope that the study of tunneling effects has reached unprecedented levels. Probably it was Iogansen (1964) who first suggested the possibility of resonant transmission of electrons in crystals (and, in particular, semiconductors) through a system of potential barriers. The development of new techniques for epitaxial growth of high-quality semiconductor films has led to the experimental demonstration of the phenomenon of *resonant tunneling,* by which under certain conditions an electron can pass unattenuated through two consecutive potential barriers confining a thin semiconductor region (see Fig. 1).

Resonant phenomena are not unique to tunneling through repulsive potentials. In fact, for certain energies an attractive potential (quantum well) can be transparent to particles normally scattered by it. A well-known example of such a phenomenon is the Ramsauer–Townsend effect, in which the cross section of electrons scattered by noble-gas atoms exhibits minima for certain resonant energies of the attractive atomic potential.

Resonant tunneling can be compared to the transmission of an electromagnetic wave through a Fabry–Pérot resonator. The equivalent of a Fabry–Pérot resonant cavity is formed by the semiconductor potential well defined by the two potential barriers, which are analogous to the Fabry–Pérot mirrors. The resonance condition is met when the energy of the electron coincides with one of the possible quantized energies of the semiconductor well (the normal modes of the Fabry–Pérot cavity).

The phenomenon of resonant tunneling could in principle be observed in multiple metal-insulator heterostructures, but it is the long mean free path of semiconductors that makes them ideal for the experimental study of the effect. It is manifested in the current-voltage characteristics of double-barrier heterostructures as regions of negative differential conductance, which are the foundation for new devices being explored for high-frequency electronics.

In this article we review the physical basis

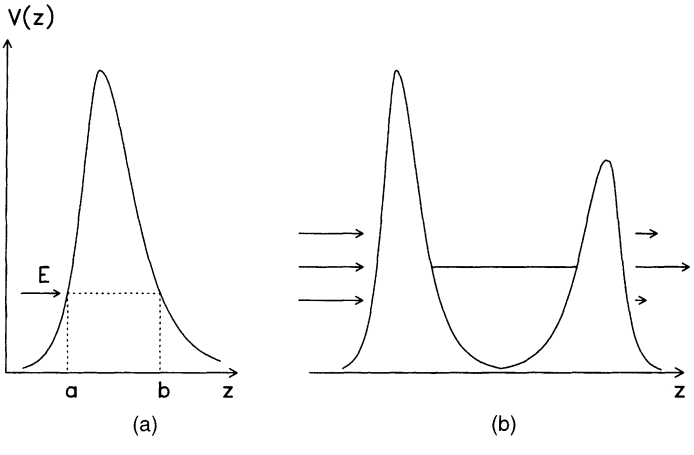

(a) (b)

FIG. 1. (a) Generic potential profile, whose gradient is equal to a force acting on an electron of energy *E*. Points *a* and *b* are the classical turn-around points. Quantum mechanically, the electron has a small (but nonzero) probability of tunneling through the potential barrier and "reappearing" at point *b*. (b) A generic double-barrier potential profile, enclosing a quantum well. When the total energy of the electron coincides with a quasibound state of the well, tunneling becomes resonant and the tunneling probability is enhanced orders of magnitude over nonresonant tunneling.

of resonant tunneling, its main consequences, and its practical implications. First we introduce in detail the concept of tunneling, before describing resonant tunneling through a double-barrier structure. We focus on its consequences for heterostructures made of semiconductors and on concepts such as charge accumulation, tunneling time, and intrinsic bistability. We then discuss manifestations of resonant tunneling in several circumstances—for instance, under a magnetic field and under lateral confinement conditions. The final section is devoted to applications based on resonant tunneling, with emphasis on the physical parameters that enter into the optimization of device performance.

2. QUANTUM-MECHANICAL TUNNELING

Tunneling is a process in which a particle passes from one continuum of states to another (classically allowed regions) by traversing a potential barrier whose height is larger (forbidden region) than the total energy of the particle. The phenomenon is characterized by the probability that a particle incident on the barrier appear on the other side of it. For a free particle of mass m such a probability is calculated by solving Schrödinger's equation, which in one dimension (the z axis) reads

$$\left[-\frac{\hbar^2}{2m}\frac{d^2}{dz^2} + V(z) - E\right]\psi(z) = 0. \qquad (1)$$

Here $V(z)$ describes the potential barrier, and $\psi(z)$ is the quantum-mechanical wave function representing the particle of energy E. The module squared, $|\psi(z)|^2$, gives the probability of finding the particle at position z.

2.1 Tunneling and the WKB Approximation

In practice, there are only a few tunneling potentials for which Eq. (1) can be solved analytically. Although in most cases it is not difficult to find numerical solutions, to get a physical insight it is frequently advantageous to use simplifying approximations for calculating the tunneling probability, the most widely used of which is the semiclassical Wentzel–Kramers–Brillouin (WKB) approximation. This general method can be applied to the probability of tunneling through a barrier, as described in detail in quantum-mechanics textbooks, provided the change of the potential within a de Broglie wavelength is small relative to the particle's kinetic energy.

In the WKB approximation, the tunneling probability T through a potential $V(z)$ is given by

$$T = C\exp(-2\beta), \qquad (2)$$

when β, defined by

$$\beta(E) = \int_a^b \left\{\frac{2m}{\hbar^2}[V(z) - E]\right\}^{1/2} dz, \qquad (3)$$

is sufficiently large. Here a and b correspond to the classical turning points [see Fig. 1(a)]. The value of the prefactor C depends on whether the quantity $V(z) - \mathrm{E}$ changes continuously or discontinuously at those turning points. If both changes are continuous, then $C = 1$; if both are discontinuous then C depends on E. The WKB approximation can be generalized to the cases where β is no longer large; the resulting expressions are then slightly more complicated than Eq. (2).

Tunneling of electrons in solids involves Bloch states and crystalline potentials, so that in principle the equations governing that process are much more complicated than those presented so far here. However, it is possible to use an effective-mass envelope-function approximation, in which the free-electron mass is replaced by an effective mass m^* and the complete potential $U(r)$ by the additional potential $V(r)$ imposed over the crystal lattice potential $U_0(r)$. The envelope wave-function solution $\psi(r)$ of Schrödinger's equation for such a reduced potential describes the envelope of the rapidly varying Bloch functions, provided $V(r)$ varies slowly on the scale of the lattice potential.

2.2 Tunneling in Semiconductor Structures

In a heterostructure of two semiconductors, say A and B, the composition varies on a distance of a few lattice constants. The potential $V(r)$ is provided by this discontinuity between the two materials. If a layer of B is imbedded in A, and if the conduction-band edge of B is higher in energy than that of A (see Fig. 2), then B acts as a potential barrier for electrons moving through the material interfaces. For a range of energies below the conduction-band offset between A and B, electrons tunnel through such a rectangular barrier.

Tunneling in semiconductors is not limited to heterostructures or to a single electronic band. In fact, under a strong electric field $\mathscr{E}$, electrons from the valence band can tunnel to empty states in the conduction band (interband tunneling) of the same material over a distance $E_g/e\mathscr{E}$, where E_g is the bandgap energy between the two bands and e is the electronic charge. (This is the mechanism for the Zener effect, which explains dielectric breakdown in solids.) If two spatial regions of a semiconductor are heavily p-type and n-type doped, respectively, so that electron and hole distributions are degenerate, both kinds of carriers have a significant probability of tunneling through to the other band. This is the basis of the Esaki or tunnel diode [Fig. 3(a)].

Another example of interband tunneling is offered by a heterostructure in which the barrier material B is such that the edge of its valence band is close in energy to that of the conduction band of material A, as indicated in Fig. 3(b). Inside the tunneling barrier, as in the p-n junction just discussed, the electron changes its character to holelike because of its proximity in energy to the valence band. In this case, it is also possible to use a WKB approximation to calculate the tunneling probability, provided the change from electronlike to holelike through the bandgap is taken into account. This can be done, e.g., by an analytical continuation into the gap of the two-band dispersion relation between energy and momentum.

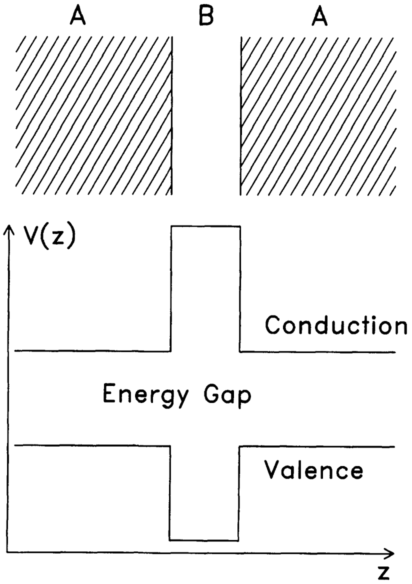

FIG. 2. Electronic potential energy along the direction z for a heterostructure consisting of a slab of material B in a material A whose conduction-band edge is higher in energy than that of B. Material B acts as a potential barrier to the motion of electrons across the interfaces between the two materials. On the other hand, when the conduction-band edge of B is lower in energy than that of A, B acts as a quantum well for electrons.

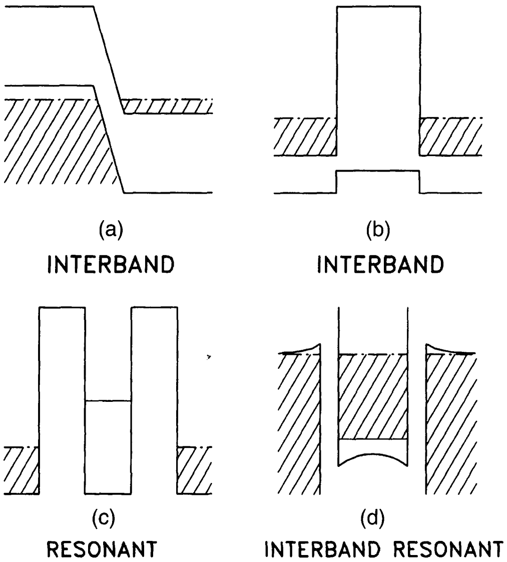

FIG. 3. Various potential profiles that can lead to negative differential resistance. (a) Interband tunneling, which can occur in a heavily doped p-n semiconductor junction. The so-called tunnel diode is based on this effect. (b) Interband tunneling through a single-barrier heterojunction. Because of the proximity in energy of the conduction and valence bands of the electrodes and the tunneling barrier, respectively, electrons from the electrodes have a hole character when tunneling through the barrier. (c) Resonant tunneling through two rectangular barriers. (d) Interband resonant tunneling, which can be seen as a particular case of (a) in which two p-n junctions are placed back to back with the n region in the center being narrow enough to exhibit quantum-size effects.

General to the various tunneling configurations outlined above are energy and momentum conservation laws. The total energy of the tunneling particle is conserved in the process, although, of course, its kinetic and potential energies are position dependent. In the simple case of a one-dimensional potential, transverse momentum $p_\perp$ ($p_\perp = \hbar k_\perp$, where $k_\perp$ is the transverse component of the wave vector) is also conserved, since there are no forces acting on the particle other than along the potential direction. As we will see later, a transverse magnetic field induces a force perpendicular to the field and the tunneling current, and under such perturbation transverse momentum is no longer conserved. Even in the absence of such a force, transverse momentum may not be conserved if a phonon is involved in the process (phonon-assisted tunneling) or if there is scattering induced by interface roughness. In practice, although weaker than the conservation of energy, the conservation of transverse momentum holds well, in a statistical sense. (That is, most electrons do not suffer scattering, which at the very least changes momentum direction.) In fact, the clearest macroscopic evidence of resonant tunneling is a direct consequence of the momentum-conservation law.

2.3 Tunneling Current

The macroscopic quantity of interest in tunneling structures is most frequently the electric current through the barrier. The tunneling current per unit area, J, is the product of the average electron velocity, the density of states, and the tunnel probability, and can be written as

$$J = \frac{e}{4\pi^3\hbar}\int dk_z d^2k_\perp f(E)T(E)\left(\frac{\partial E}{\partial k_z}\right), \tag{4}$$

where $f(E)$ is the Fermi–Dirac distribution for a total energy E, and $T(E)$ is the tunneling probability defined as the ratio between the incident and transmitted probability currents.

Under equilibrium, the net tunneling current is zero for any of the semiconductor structures mentioned above since the tunneling probability is the same for electrons moving from left to right as for those tunneling in the opposite direction. However, if a voltage is applied between the two end regions (electrodes) of A, a tunneling current will flow from the negative to the positive "electrode," expressed by

$$J = \frac{e}{4\pi^3\hbar}\int dE_z d^2k_\perp[f(E) - f(E + eV)]T(E_z). \tag{5}$$

The tunneling current can then be readily evaluated once the tunneling probability through the barrier is known.

3. RESONANT TUNNELING IN SEMICONDUCTOR HETEROSTRUCTURES

In the prototype heterostructure discussed above, the conduction-band edge of material B presented a potential barrier to electronic motion because of its higher position relative to that of material A. On the other hand, if the band edge of B is lower than that of A, layer B behaves as a quantum well, with one or more localized electronic levels at characteristic energies determined by the effective mass and the thickness of the layer (quantum-size effect). Resonant tunneling occurs in complex heterostructures comprising at least one quantum well and two barrier layers, like the one shown in Fig. 3(c).

3.1 Resonant Tunneling Probability

If an electron moves from the left to the right of such a rectangular double-barrier structure, its tunneling probability will be very small (of the order of the product of the tunneling probabilities for each individual barrier). On the other hand, when its energy coincides with one of the characteristic energies of the quantum well, its tunneling probability is enhanced drastically, to a value of the order of ratio between the transmission probabilities of the individual barriers. When such a resonance occurs we speak of resonant tunneling. If the two barriers are identical, the tunneling probability is unity, and the double-barrier system becomes totally transparent to the electron. The resonant process can be seen as a constructive interference in the well of the electron waves that tunnel through the two individual barriers.

The tunneling probability through a set of barriers and wells is frequently calculated using a transfer-matrix method (Kane, 1969; Tsu and Esaki, 1973). In such a formalism, each well/barrier is represented by a 2×2 matrix M, connecting the coefficients A and B of the general solution

$$\psi_e(z) = A \exp[ikz] + B \exp[-ikz] \qquad (6)$$

at both sides of the well/barrier. The wave vector k in each region, defined by

$$\hbar^2 k^2 / 2m = E - V, \qquad (7)$$

is real in a well region ($E - V > 0$) and imaginary in a barrier ($E - V < 0$).

If an electron is incident from the left (region 1) on a set of wells/barriers consisting of n regions, only a transmitted wave will appear in region n. The transmission probability is then (Kane, 1969)

$$T = \frac{k_1 m_n}{k_n m_1} \frac{|A_n|^2}{|A_1|^2}. \qquad (8)$$

The tunneling probability for a symmetrical double-barrier structure, calculated using the transfer-matrix method, is shown in Fig. 4(a), along with similar probabilities for triple and quintuple barriers. The sharp peaks in the transmission probability correspond to the resonant energies of the system. Near resonance, that probability is Lorentzian in shape, a result that is valid even for asymmetrical structures and that can be expressed by

$$T(E) = \frac{T_0}{1 + [(E - E_0)/\Delta E]^2}. \qquad (9)$$

Here T_0, the maximum tunneling probability, is given by (Ricco and Azbel, 1984)

$$T_0 = 4T_L T_R/(T_L + T_R)^2, \qquad (10)$$

where T_L and T_R are the individual tunneling probabilities of the left and right barriers, respectively, and ΔE is the half-width of the resonance, which is proportional to $T_L + T_R$. From Eq. (8) it follows immediately that $T_0 = 1$ when $T_L = T_R$ and that $T_0 = 4T_R/T_L$ when $T_R \ll T_L$.

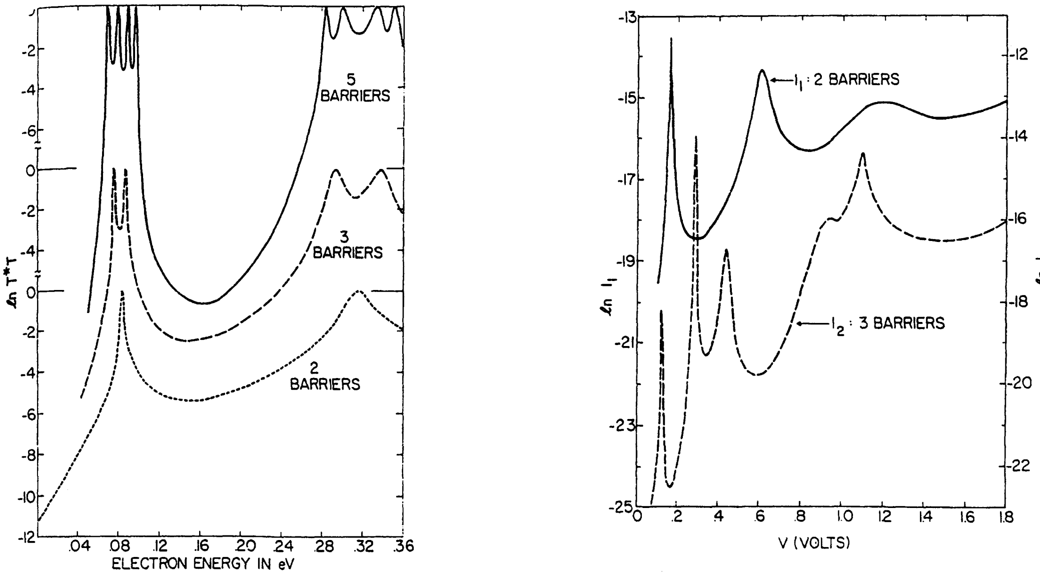

FIG. 4. (a) Probability (in logarithmic scale), calculated using the transfer-matrix method, for an electron of energy E to tunnel through two, three, or five rectangular barriers, each 0.3 eV high, 20 Å wide, and 50 Å apart from each other. (b) Calculated tunnel current density vs voltage applied to either a double- or a triple-barrier heterostructure similar to the ones described in (a). (After Tsu and Esaki, 1973.)

3.2 Resonant Tunneling Current

As for the case of a single barrier, once the tunneling probability of a double barrier is known the resonant tunneling current is calculated using Eq. (4). In semiconductor heterostructures, the two electrodes (from now on denoted emitter and collector, the former being at the lower potential) in practice have electron densities of the order of 10^{18} cm^{-3}, achieved through doping with n-type impurities. These electrons have a range of energies from 0 up to E_F, the Fermi energy. If one considers parabolic energy bands and zero temperature, the tunneling current density can be derived from Eq. (5) to be

$$J = \frac{em}{2\pi^2\hbar^3}\int_0^{E_F} dE_z(E_F - E_z)T(E_z), \qquad (11)$$

where for simplicity we have assumed that the voltage drop between the emitter electrode and the quantum well is larger than the Fermi energy divided by the electronic charge. Using a Lorentzian lineshape like Eq. (9) for the transmission probability near resonance and assuming that $\Delta E \ll E_F$ (which in practice is always the case), it follows that in this approximation the current density is

$$\begin{aligned} J &\simeq \frac{em\Delta E}{2\pi\hbar^3}T_0(E_F - E_0 + eV), \quad 0 \le E_0 \\ &\quad - eV \le E_F; \\ J &= 0 \text{ elsewhere.} \end{aligned} \qquad (12)$$

In other words, the current increases linearly with bias once the voltage aligns the localized state E_0 with the Fermi level, and it reaches a maximum at a bias such that E_0 coincides with the conduction-band edge. For higher voltages the tunneling current is ideally zero, as a consequence of conservation of transverse momentum. Although not exclusive to resonant tunneling, this abrupt drop in the current (and the corresponding negative differential resistance) is one of its defining characteristics (Fig. 5).

For nonparabolic bands and finite temperatures the expression for the current density is more complicated than Eq. (11) and its integration may have to be carried out numer-

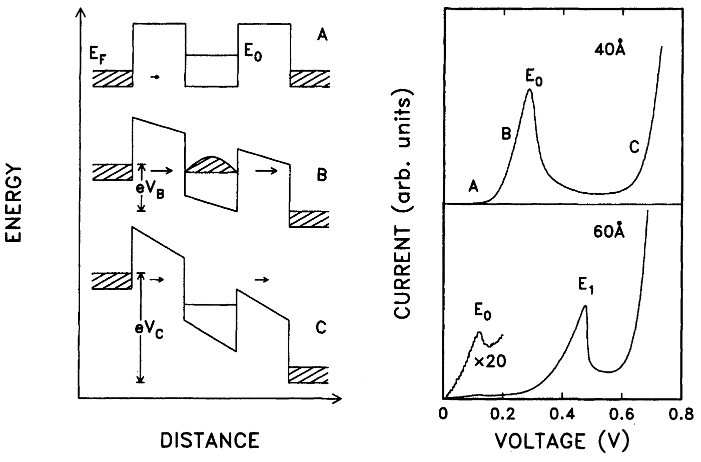

(a)

(b)

FIG. 5. (a) A double rectangular barrier under bias between two end electrodes, as exemplified by the semiconductor heterostructure n^+-GaAs/GaAlAs/GaAs/GaAlAs/n^+-GaAs. At very low voltage (A) the tunneling probability is very small and the current is negligible. When the external bias is such that the quantized state E_0 of the GaAs well lies between the Fermi level and the conduction-band edge of the emitter (B), resonant tunneling takes place and the current increases dramatically while charge accumulates in the well. When the bias pushes E_0 below the emitter's band edge, conservation of transverse momentum prevents electrons from tunneling and the current drops sharply, until at very high voltage (C) the right barrier effectively disappears and the current increases again. (b) Experimental current-voltage characteristics for two GaAs/GaAlAs double-barrier heterostructures with different quantum-well widths. A 40-Å well can accommodate only one bound state, giving rise to a single peak in its *I-V* characteristics, while a 60-Å well binds two discrete states.

ically, but that qualitative dependence of the tunneling current on voltage remains.

The negative differential resistance (NDR) characteristic of resonant tunneling provides information of two types. On one hand, the observation of NDR features implies a very high-quality heterostructure, free from materials defects and impurities and with sharp interfaces between its material components. (All these imperfections can act as scattering centers that break the $k_\perp$ conservation law.) On the other hand, the onset voltage of the resonant current and the peak voltages are directly related to the Fermi level of the electrode and the energies of the states in the quantum well, respectively.

Let us assume for a moment that when a voltage V is applied to a symmetric double-barrier heterostructure the voltage drops exclusively and uniformly in the undoped region. In this simple model, the peak voltage (times the electronic charge) corresponds to twice the quantum-state energy E_0, and the difference between the peak and onset voltages is twice the Fermi energy of the electrode. (The factor of 2 results from the fact that only half of the total voltage drops between the emitter and the center of the well.)

In practice, the peak voltages are directly related to the quantum energies, but the relation is not as simple as implied above. The reasons are multiple: part of the voltage drops at the electrodes, the drop is not uniform throughout the heterostructure, the energies of the quantum states are modified by the electric field, and there is an accumulation of electronic charge in the quantum well during the tunneling process. An exact calculation of the resonant voltage requires solving Schrödinger's equation as well as the electronic transport equations. Because of the complexity of the problem, a static approximation is frequently used, in which the accumulated charge in the well is ignored and the potential profile and energy levels at any voltage are calculated by self-consistently solving Poisson's and Schrödinger's equations.

3.3 Charge Accumulation and Bistability

Under steady-state resonant-tunneling conditions, charge builds up in the well proportionally to the tunneling current: $Q = \tau_R J$ (Goldman *et al.*, 1987). The proportionality constant, τ_R, represents the average time an electron spends in the well before tunneling out to the collector electrode. The lifetime of the electron in the well, τ_T, can be estimated semiclassically by calculating the probability per unit time for an electron of mass m and kinetic energy E to escape out of a well of width W through either the left or right barrier. The lifetime can then be written as $\tau_T^{-1} = \tau_L^{-1} + \tau_R^{-1}$, with

$$\frac{1}{\tau_T} \simeq \frac{\sqrt{2E/m}}{2W}(T_L + T_R). \qquad (13)$$

Rigorously it is shown that $\tau_T = \hbar/2\Delta E$, where ΔE is the resonance half-width. From Eqs. (10) and (12) it then follows that the resonant-tunneling current is proportional to $T_L T_R/(T_L + T_R)$ and that the accumulated charge is proportional to $T_L/(T_L + T_R)$. In an asymmetric double-barrier heterostructure, the current is then determined by the tunneling probability of the less transparent barrier. On the other hand, the charge is controlled only by the collector barrier and is the largest for an opaque collector (relative to the emitter). The maximum charge accumulated is of the order of the two-dimensional charge in a well whose energy level is below the Fermi level by an amount E_F.

This analysis assumes a picture of resonant tunneling in two sequential steps, as opposed to a coherent process (Büttiker, 1988). An electron first tunnels from the emitter into the quasibound state in the quantum well, where, by scattering, it loses its original phase. Then the electron tunnels out of the well to the collector (Luryi, 1985). As far as the current and charge accumulation are concerned, however, the two pictures are in practice equivalent. The current peak is independent of scattering whenever the emitter Fermi energy is much larger than the resonance width ΔE (Weil and Vinter, 1987). The accumulated charge is in both cases proportional to τ_R because the integrated density of states is constant, as shown by integrating the wave function in the well and adding the contributions of all tunneling electrons (Gu and Gu, 1989).

In the discussion about the relation between the resonant voltages and the alignment of the quantum states with the emitter-band edge, it was implicitly assumed that the effective mass of the electron at the electrode and in the quantum well was the same. While this is frequently the case, there are instances in which such an assumption is not valid. As demonstrated experimentally, parallel-momentum conservation requires that, depending on whether the ratio m_w/m_b is greater or less than unity, the resonant voltage occurs when the quantum state is above or below the emitter band edge.

The accumulation of charge in the quantum well during the tunneling process has an interesting consequence: the appearance of bistability in the *I-V* characteristic, produced by two stable states at voltages around the NDR region (Eaves, 1990). The accumulated electrons create a space-charge layer, which alters the voltage distribution across the heterostructure for a given total voltage *V*. It is then possible to have two potential configurations, one in which the quantum state is still slightly above the emitter-band edge and the current is large, and another in which (for the same voltage) that state is below the emitter's edge and the current is low.

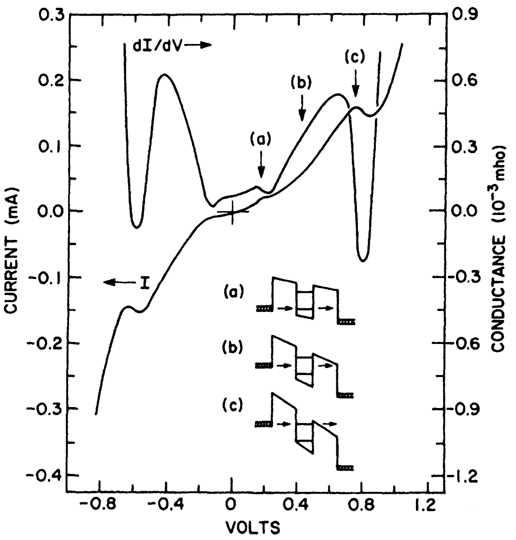

FIG. 6. First current-voltage characteristics that showed negative differential resistance due to resonant tunneling in a double-barrier semiconductor heterostructure. Although weak, the features (a) and (c) in the current (seen more clearly in the conductance) are consequences of resonant tunneling through the two quasibound states in a GaAs-GaAlAs heterostructure. (After Chang *et al.*, 1974.)

3.4 Resonant Tunneling in Type I Heterostructures

Resonant tunneling was first demonstrated experimentally in 1974, using (n^+)-GaAs-Ga$_x$Al$_{1-x}$As-GaAs-Ga$_x$Al$_{1-x}$As-(n^+)GaAs heterostructures (Chang *et al.*, 1974). The combination of two materials like GaAs and GaAlAs, whose lattice constants differ by less than 0.1% (even when $x = 1$), makes it possible the epitaxial deposition of alternate layers of them with the high degree of perfection necessary for the observation of resonant tunneling. The band gap of GaAlAs is larger than that of GaAs and its conduction-band edge is higher in energy. A GaAlAs slab acts then as a potential barrier to electrons in GaAs, and GaAlAs layers on both sides of a thin GaAs film confine its electrons in a quantum well. The barrier height increases linearly with the Al content, at an approximate rate of 0.1 eV for every 10% of Al.

The current-voltage characteristics of that first experiment showed two features, at 0.2 V and 0.8 V, associated with resonant tunneling through the two localized states in the GaAs well, as indicated in the inset of Fig. 6. The negative differential resistance (NDR) regions were observable only at liquid-nitrogen temperature (T = 77 K) or below, in a current background foreign to resonant tunneling and probably due to imperfections in the two semiconductors. Since then, dramatic advances in materials preparation by epitaxial techniques such as molecular beam epitaxy and metallorganic chemical vapor deposition have led to a drastic reduction of the leakage current, making possible the observation of resonant tunneling at room temperature and even of intrinsic bistability due to charge accumulation (Fig. 7).

Double-barrier heterostructures in other materials systems compatible for epitaxial growth (either lattice matched or strained layer) have also shown the characteristic NDR of resonant tunneling, e.g., GaAsP-GaAlAs, InGaAs-AlInAs, InGaAs-GaAs, etc. The observation of resonant tunneling has not been restricted to III-V compounds: Si-SiGe heterostructures have shown features evidencing the phenomenon, and there has even been a report of resonant tunneling in amorphous heterostructures.

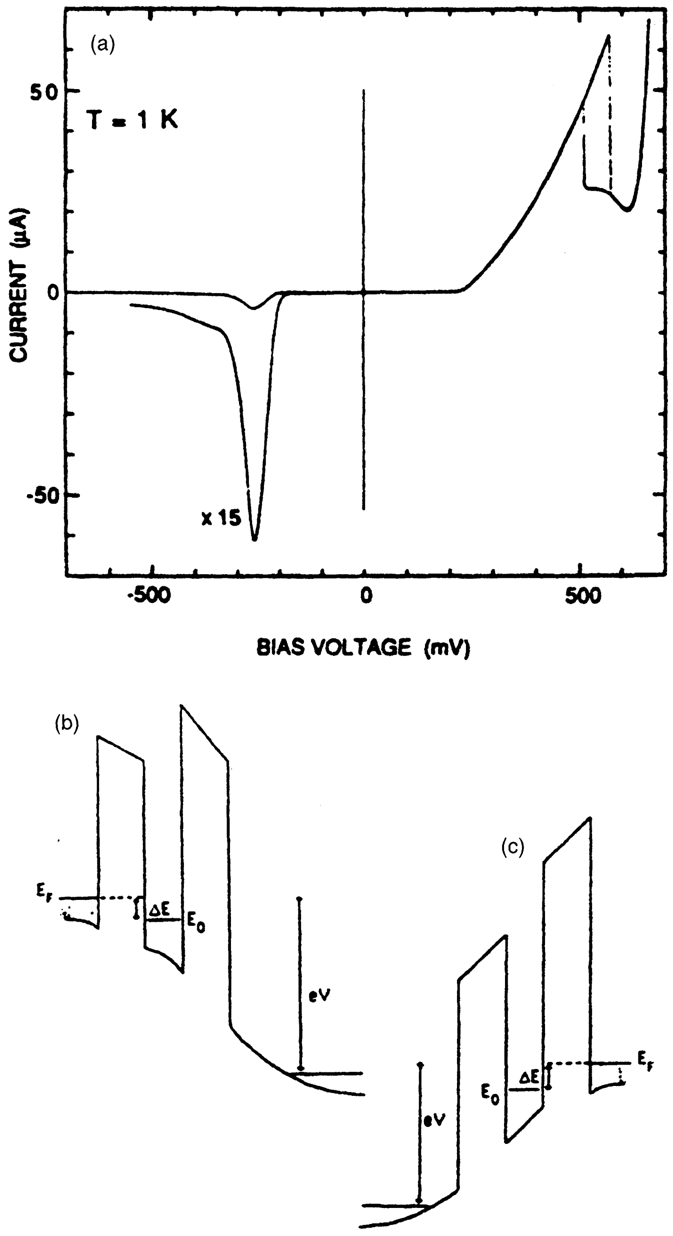

FIG. 7. (a) Current-voltage characteristics of an asymmetric double-barrier GaAs-GaAlAs heterostructure. (b) The tunneling current is much larger at forward bias because in this case, the two barriers become effectively more symmetrical and the resonant tunneling probability is larger than (c) at resonance under reverse bias. The large tunnel current and long lifetime in the well lead to a large charge accumulation in forward bias, which in turn gives rise to bistability. Under reverse bias, the accumulated charge is much smaller and no bistability is observed. (After Zaslavsky *et al.*, 1988.)

3.5 Band-Structure Effects

When studying tunneling phenomena, band-structure considerations beyond the conduction-band minima of the Brillouin zone are frequently ignored. This simplification is warranted in most cases, but there are circumstances in which close attention has to be paid to other critical points—for instance, when a high-Al composition ($x \geq 0.4$) is used in $Ga_{1-x}Al_xAs$-GaAs-$Ga_{1-x}Al_xAs$ resonant tunneling structures. Then, in the conduction band, the X point is close in energy to the Γ point and band mixing can occur. For $x > 0.45$, X is below Γ and the minimum-energy profile is a mixture of Γ (at GaAs) and X (at GaAlAs). This raises the question of which barrier determines the confined states in the quantum well—the one with the same symmetry or the lower one?

The answer has been provided by using the fact that hydrostatic pressure lowers the X point relative to Γ in GaAlAs. Starting with double-barrier $Ga_{1-x}Al_xAs$-GaAs-$Ga_{1-x}Al_xAs$ heterostructures with $x = 0.40$, it was found that pressure linearly shifted the resonant voltage to lower values, at a rate ($\simeq 1.8$ meV/kbar) that remained constant well beyond the pressure (4 kbar) at which X and Γ cross. This result proved that the Γ barrier of GaAlAs (not necessarily the lowest barrier) determines the states of a quantum well at the Γ point of GaAs (Mendez *et al.,* 1986a). The uniform voltage shift was attributed to the slightly different pressure coefficients of GaAs and GaAlAs at Γ as well as to a pressure-induced increase of the electron effective mass.

When considering quantum-size effects at points of the Brillouin zone other than Γ, it is interesting to note that the X point in AlAs is lower in energy than in GaAs. In n^+GaAs-AlAs-GaAs-AlAs-n^+GaAs structures it is then possible to draw two potential profiles: one based on the Γ point, in which GaAs acts as a well and AlAs as a barrier, and another based on the X point, where AlAs is now a well while GaAs behaves as a barrier. In addition to the Γ-like quantum states in GaAs, X-like states are formed in AlAs, as evidenced by NDR features in the low-temperature *I-V* characteristics. Although the electrons from the n^+ electrode have a Γ character, their tunneling through X-point states still preserves transverse momentum (Mendez *et al.,* 1987).

3.6 Resonant Tunneling of Holes

The double-barrier potential profile that is at the center of resonant tunneling in semiconductor heterostructures is not exclusive to the conduction band. Many materials combinations, including the most familiar, GaAs-GaAlAs, exhibit a similar profile for the valence band (although with different barrier height) and holelike discrete states are formed in the quantum well. If the end electrodes are p doped, instead of n doped as considered in the above sections, then holes can tunnel resonantly through the quantum states, which can have either light- or heavy-hole character. Experimentally, both tunneling channels appear as NDR features of the *I-V* characteristics, although normally heavy-hole resonant currents are much smaller and resolved clearly only at low temperatures (Mendez *et al.,* 1985).

3.7 Interband Resonant Tunneling

As mentioned earlier, negative differential resistance is not unique to resonant tunneling. The *I-V* characteristics of tunnel diodes show it, and so do it the characteristics of a single barrier in which interband tunneling occurs, as shown in HgCdTe-HgTe heterostructures.

An interesting combination of resonant and interband tunneling occurs in structures based in the InAs-GaSb system, which is peculiar in that the conduction-band edge of InAs is lower in energy than the top of the valence-band edge of GaSb. As a result, in a heterostructure of the two materials there is an accumulation of electrons on the InAs side of the interface and a corresponding accumulation of holes on the GaSb side. If AlSb, which is also epitaxially compatible with InAs and GaSb, is included as part of a polytype heterostructure, a large number of combinations emerge (McGill and Collins, 1993). For instance, in a (p^+)GaSb-AlSb-InAs-AlSb-(p^+)GaSb heterostructure the GaSb electrodes are p type while the InAs quantum well contains free electrons arising from the charge-transfer process [Fig. 3(d)]. This structure can then be regarded as two tunnel diodes back to back, in which the common region (n-InAs) is so thin as to give rise to quantum-size effects and therefore to resonant tunneling. A symmetrical configuration, with the role of electrons and holes reversed, is provided by an interchange of InAs and GaSb.

The band discontinuity between two semiconductors is the origin of the rectangular profiles discussed above. By compositional grading of the materials it is possible to create other profiles, e.g., triangular or parabolic, either in the well or the barrier. Resonant tunneling associated with a parabolic well (in which the energy levels are equally separated) has been observed at a temperature of 7 K (Sen *et al.,* 1987).

3.8 Magnetic-Field Effects

A magnetic field H quantizes the electronic motion perpendicular to it into a set of discrete cyclotron orbits, with energies $E_n = (n + \frac{1}{2})\hbar\omega_c$ (Landau levels), where $\omega_c =$

eH/mc is the cyclotron frequency. When the field is parallel to the direction of tunneling, the current shows new NDR features at voltages below the main resonance, which, although generally weak in the *I-V* characteristics, are clearly resolved in plots of dI/dV vs V (Mendez *et al.,* 1986b). These field-induced structures are associated with the resonant tunneling of electrons through individual Landau levels in the quantum well. Since the energy separation between these levels is inversely proportional to the effective mass of the electron, from the voltage difference between consecutive field-induced resonances it is possible to determine the electronic effective mass.

In the presence of a magnetic field, the conservation of transverse momentum requires that the Landau-level index of electrons be conserved during the tunneling process. On the other hand, when the field is perpendicular to the current no Landau levels are formed in the well and a transverse force acts on the electrons, as a result of which the NDR is shifted (Ben Amor *et al.,* 1988). Since there is a correlation between the magnetic field and the lateral momentum gained by the electron, a plot of the voltage shift with field gives direct information on dispersion relation between energy and in-plane momentum. This tool has been employed to map the nonparabolic dispersion of the valence band in GaAs (Hayden *et al.,* 1991) and even the anisotropy of holes in SiGe/Si heterostructures (Gennser *et al.,* 1991).

Much more pronounced than in conventional type I resonant tunneling (involving either electrons or holes) are the effects of a magnetic field in interband resonant tunneling, in which both electrons and holes participate. A system where the latter occurs is (p)GaSb-AlSb-InAs-AlSb-(p)GaSb, as described earlier. The small Fermi energy of holes in the GaSb electrodes, together with the large cyclotron energy of electrons in the InAs quantum well, contributes to the observation of distinct NDR structures associated with electron Landau levels even at modest magnetic fields [Fig. 8(b)]. For fields above 15 T the electronic spin splitting can be resolved, from which the Landé factor g—and its field dependence—has recently been determined (Mendez *et al.,* 1993).

The unique characteristics of that system allow us to probe by tunneling the two-dimensional electron gas in the InAs quantum well. This capability is especially interesting at high magnetic fields, under which a two-dimensional gas shows quantization of the in-plane Hall voltage and vanishing longitudinal resistance for certain fields (quantum Hall effect). In interband magnetotunneling, the zero-bias conductance exhibits a similar dependence: oscillatory behavior at low fields and zero conductance at high fields whenever the Fermi energy is between two Landau levels [Fig. 8(c)]. Because the quantum Hall effect relies on the existence of localized states that prevent transport of in-plane current, the manifestation of those states in "vertical" transport requires a better understanding of possible paths for the tunneling current, or maybe a deeper insight on the concept of localization.

3.9 Low-Dimensionality Resonant Tunneling

By quantizing the electronic energies into Landau levels, a parallel magnetic field effectively reduces the dimensionality of resonant-tunneling heterostructures. In the absence of the field, tunneling occurs between two three-dimensional regions through a two-dimensional well, whereas when the field is present tunneling is between one-dimensional electrodes via a fully quantized (zero-dimensional) well. A physical reduction of dimensionality can be achieved by lithographic means instead of by a magnetic field. In "quantum cylinders" (about 100 nm in diameter) made out of double-barrier structures, well-resolved multiple NDR have been observed at low temperatures, as a result of the quantization of the states in a GaInAs well in all three dimensions (Fig. 9). In this case, resonant tunneling takes place between two one-dimensional systems (the electrodes) via a zero-dimensional system (the quantum well).

By physically reducing the dimensions of a resonant-tunneling device, both its capacitance C and the charge accumulated in the well during the tunneling process are decreased. If the size is small enough, the energy necessary to charge the well with a single electron, $U_C = e^2/2C$, is larger than the thermal energy, kT, at temperatures around 0.1 K. The quantization of the charging en-

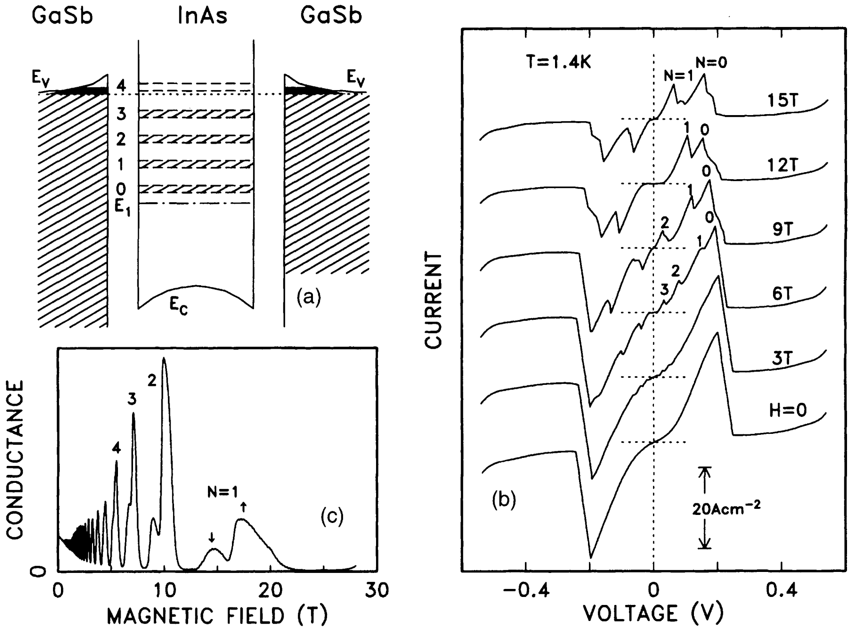

FIG. 8. (a) Band profile of a GaSb-AlSb-InAs-AlSb-GaSb heterostructure under a magnetic field *H* parallel to the materials interfaces. The InAs layer forms a quantum well for electrons while holes accumulate at the GaSb interfaces. The field creates a set of Landau levels in InAs whose separation is proportional to *H*. (b) Resonant tunneling is possible only when the energy of holes in the emitter coincides with that of an occupied Landau level in InAs, giving rise to a set of peaks in the *I-V* characteristics. The set of curves corresponds to a GaSb-AlSb-InAs-AlSb-GaSb heterostructure with 40-Å AlSb and 150-Å InAs layers. (c) Oscillations in the tunneling conductance, at $V = 0$, as a function of magnetic field, for the heterostructure of (b). In analogy with the quantum Hall effect, the conductance vanishes whenever the Fermi level lies in the gap between two Landau levels of the InAs well. (After Mendez *et al.*, 1991.)

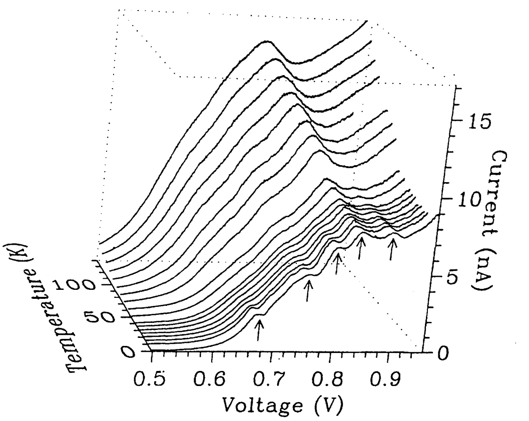

FIG. 9. Current-voltage characteristics of a 1000 Å quantum dot formed out of a GaInAs-GaAlAs double-barrier heterostructure. At high temperature, a negative differential resistance feature in the 0.7–0.8-eV range is due to resonant tunneling through the second-quantized state in the GaInAs well. At low temperatures new features are observable. They are evidence of size quantization in the plane of the interfaces. (After Reed *et al.*, 1988.)

ergy prevents tunneling for biases below a critical voltage $V_C = U_C/e$. This suppression of the tunneling current—normally denoted Coulomb blockade—has been observed below 1 K, coexisting with quantum-size effects (Su *et al.*, 1992).

The lateral dimensions of tunneling devices can be controlled by means of a Schottky-gate contact around an etched quantum cylinder. By applying a negative bias to the gate, relative to a conducting substrate, a depletion region is created around the quantum cylinder that has the effect of reducing the cylinder's cross section for the tunneling current. This geometry has allowed the observation of resonant tunneling through single impurities in quantum wells and of quantum-confinement effects controlled by a gate voltage (Dellow *et al.*, 1992; Guéret *et al.*, 1992).

Quantum wires also offer opportunities for resonant-tunneling effects in low-dimensionality systems. In an ingenious scheme

(Chang *et al.*, 1985), a thin heterostructure is grown on the edge of a conventional double-barrier structure so that the intersection of the two planes gives rise to a quantum wire. Resonant tunneling between two-dimensional electrodes through the one-dimensional well in the wire has been observed recently at $T = 0.3$ K (Dignam *et al.*, 1994).

Most experiments on resonant tunneling have utilized two different materials to create quantum well and barriers, but it is also possible to create a double-barrier configuration by in-plane modulation of the density of a two-dimensional (2D) electron gas. This modulation is achieved by two metal Schottky contacts on the surface, which, when properly biased, deplete of carriers the regions underneath them. Since the Fermi level is the same throughout the 2D plane, the conduction-band edge is raised at the depletion regions under the two contacts, creating a double-barrier configuration. An advantage of this lateral profile over the vertical one is the possibility of altering the barrier heights, and consequently the quantum states, with a change of the gate voltage. By varying such a voltage and keeping constant the bias between two Ohmic contacts (drain and source) to the 2D gas, it has been possible to observe negative differential resistance features in the drain-to-source *I-V* characteristic whenever the quantum states in the well between the two gates are aligned in energy with the conduction-band edge of the 2D gas in the electrodes (Ismail *et al.*, 1989).

4. APPLICATIONS OF RESONANT TUNNELING

The phenomenon of resonant tunneling in semiconductors not only provides experimental insight into a very fundamental concept of quantum mechanics, but also it lends itself to applications in electronics as a direct consequence of the negative differential resistance in the *I-V* characteristics of resonant-tunneling diodes discussed above. Any device that exhibits NDR can have gain, and in principle can be used as a reflection amplifier or as an oscillator. The former use, however, is not very practical because reflections at various parts of the transmission line can drive the circuit into oscillations. On the other hand, these instabilities can generate useful output power, and, in fact, they are the basis for high-frequency oscillators (Sollner *et al.*, 1983).

4.1 High-Frequency Oscillators

Compared to other devices with similar functions, resonant-tunneling diodes offer higher frequency and moderate power outputs. The cutoff frequency (above which the NDR disappears and the diode does not have gain), $f_{\max}$, and the maximum power (low frequency), w, are directly related to the electrical characteristics of the resonant-tunneling diodes through the following expressions:

$$f_{\max} = \frac{1}{2\pi RC}\sqrt{\frac{R}{R_s} - 1}. \tag{14}$$

$$w = \tfrac{3}{16} I_{\max} V_{\max}. \tag{15}$$

Here, $-R$ and C are, respectively, the negative resistance and capacitance that describe the device; R_s is the series resistance associated with the contact resistance, with the accumulation and depletion regions, and with the spreading resistance; and $I_{\max}$ and $V_{\max}$ are the maximum current and voltage excursions of the NDR region, respectively. In practice, $R \gg R_s$, so that Eq. (14) can be approximated by $f_{\max} = [2\pi\sqrt{RR_sC}]^{-1}$.

Since these parameters are related to physical and materials characteristics of the diodes, such as barrier height and width or doping level, by proper design it is possible to optimize the performance of resonant-tunneling oscillators. But because those parameters may affect performance in opposite ways, some design compromises are required. For instance, the peak current increases with decreasing barrier height and width. On the other hand, when the barrier height is reduced the thermionic current increases and so does the valley current, thus decreasing the peak-to-valley ratio and correspondingly increasing R. If the barriers become too narrow the capacitance increases and the cutoff frequency suffers as well. Low doping of the electrodes is favorable because it lowers the capacitance and increases $V_{\max}$ but has the detrimental effect of increasing the series resistance.

Band-structure parameters of the materials must be also considered for optimum design. There are cases in which a large separation in energy of critical points (e.g., the Γ and X points in InGaAs) reduces the likelihood of more than one tunneling path and enhances the peak-to-valley ratio. This (together with a reduced series resistance) explains the better performance of InGaAs-AlAs devices over those based on GaAs-AlAs. By taking advantage of all these criteria, frequencies as high as 700 GHz with an output power of 0.3 μW have been reached at room temperature in resonant-tunnel diodes based on InAs-AlSb (Fig. 10). At 100 GHz the output power was about 100 μW, which compares very favorably with the 1-μW power, at the same frequency, of the best tunnel diodes.

Traditional varistors or varactors, based on the voltage-dependent resistance or capacitance of nonlinear devices like *p-n* junctions or Schottky barriers, produce much larger power output than resonant-tunneling oscillators operating at their fundamental frequency [Eq. (14)], discussed above. However, the capability of generating higher harmonics in resonant-tunneling devices has increased their power output dramatically, up to 800 μW at 250 GHz. The variation of capacitance around zero voltage of a single, thick barrier has been exploited to generate 1 mW of power (Grönqvist *et al.*, 1991), a value comparable to that produced by Schottky varactors. By design optimization it may be possible to increase that power several times.

4.2 Other Applications

Although at the moment oscillators are the most mature devices based on resonant tunneling, by no means are they the only ones being explored. A resonant-tunneling diode can act as a switch by biasing it with a source resistance larger than its negative resistance. Their short switching time (2 ps) and low power dissipation compared with traditional semiconductor technologies could make these diodes attractive for logic circuits, but the implementation of logic operations with two-terminal devices introduces serious complications. Much more promising is the use of resonant-tunneling switches in sampling circuits, analog-to-digital converters, and multigigahertz shift registers for signal-processing applications.

Three-terminal devices based on resonant tunneling have also been proposed and demonstrated, for example, resonant-tunneling hot-electron or bipolar transistors. In the former, a double-barrier heterostructure replaces the emitter–base barrier of a conventional hot-electron transistor whereas in the latter it substitutes for the tunnel barrier at the emitter–base interface of a heterostructure bipolar transistor. In the first of these two schemes the electrons are injected into

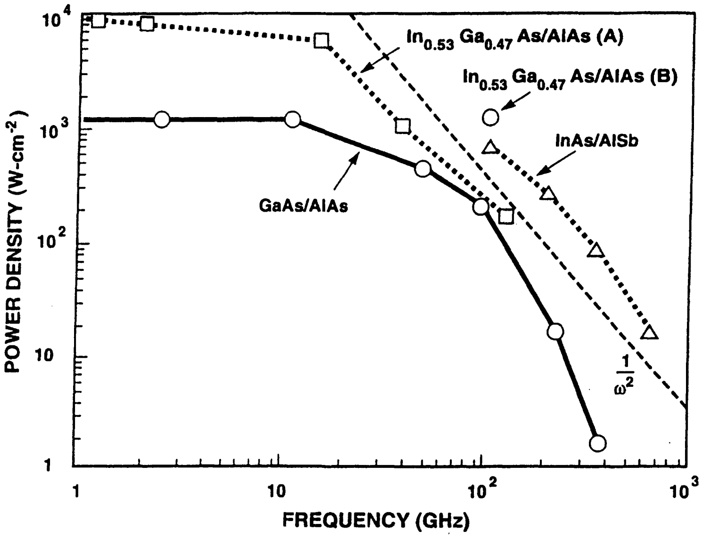

FIG. 10. A comparison of power density of resonant-tunneling oscillators based on different semiconductors. Because of the larger current densities of the InAs/AlSb pair, oscillations as high as 710 GHz with an output power of 0.3 μW have been achieved. (After Brown, 1992.)

the base by resonant tunneling, a feature that allows the device to act as a frequency multiplier or as an exclusive-NOR gate (Yokoyama *et al.*, 1985). Instead of a single transistor, MESFET technology requires eight devices for implementation of such a logic function.

The low-gain limitation of hot-electron transistors can be overcome by resonant-tunneling bipolar transistors, in which gains over 20 have been reported at 77 K, and up to 7 at room temperature, by placing the double-barrier heterostructure directly in the base (Capasso *et al.*, 1986). Recently, there has been a report of room-temperature integrated digital circuits with coexisting resonant-tunneling and double heterojunction bipolar transistors. The circuits, up to a 17-transistor full adder, required fewer components than traditional ones, thus allowing higher density and speed (Seabaugh *et al.*, 1993).

5. SUMMARY AND PROSPECT

The phenomenon of resonant tunneling reviewed here is quantum mechanical in nature, but it manifests itself in macroscopic effects even at room temperature, the most important of which is the presence of negative differential resistance in the current-voltage characteristics of semiconductor heterostructures with a double-barrier configuration. Such strong experimental evidence has made possible in the last ten years the study of resonant tunneling under a variety of circumstances and for a large group of semiconductors. By now our understanding of this phenomenon is quite complete, but its scientific interest remains unabated. An emerging area of fundamental research is the use of resonant tunneling as a spectroscopic tool to probe other quantum-mechanical systems—for example, two-dimensional gases at high magnetic fields, a topic only glimpsed here.

That the characteristic negative differential resistance of resonant tunneling is easily observed at room temperature makes some of the proposed applications good candidates for commercialization. Although it is hard to imagine that they will replace CMOS technology as the mainstream for low-power circuits, it is likely that they will find niche markets in high-frequency electronics.

The focus here has been on resonant tunneling through a double-barrier structure but the concept can be (and has been) generalized to multiple-barrier heterostructures. The ultimate case is when the barriers are so thin that there is strong coupling among the energy levels of the individual quantum wells. The electronic wave functions are no longer localized in individual wells but extended throughout the whole heterostructure. We speak then of a superlattice, by analogy with the periodic atomic lattice of a crystalline material (Esaki and Tsu, 1970; see also HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR). In contrast to the double-barrier case, electronic transport in superlattices is a much less explored subject both theoretically and experimentally. Because of the novelty of some of the phenomena that have already been predicted in the superlattice regime, the most outstanding of which is the possibility of generating electromagnetic radiation at ultrahigh frequencies (Bloch oscillations), we predict a much larger effort in this area in the next few years.

This article has dealt with semiconductors because these are the materials which have been hosts to resonant-tunneling effects so far. It is worth noting, however, that the resonant-tunneling concept for electrons in solids was first introduced in 1963 in the context of a proposed metal-insulator structure. Almost 30 years later, the first observation of room-temperature resonant tunneling in a metal-insulator ($CoSi_2$-CaF_2) double-barrier diode was reported. At the time of closing this review transistor action has just been demonstrated at 77 K in a resonant-tunneling device based on those two materials (Suemasu *et al.*, 1993). Although low materials quality still presents a challenge, these recent accomplishments are most welcome since metals have higher carrier concentrations and lower resistivities than semiconductors, both of which are favorable for high-speed operation.

GLOSSARY

Bistability: Condition of an electric circuit in which there are two stable voltage configurations for the same flow of current.

Bloch Oscillation: High-frequency electromagnetic oscillation predicted to occur when an electron moves in a periodic potential modulated by an external electric field.

Coulomb Blockade: Suppression of tunneling in a structure so small that the electrostatic energy to charge it with a single electron is larger than the thermal energy.

Cutoff Frequency: The highest frequency at which oscillations due to negative differential resistance are observed in a resonant-tunneling diode.

Heterostructure: A structure formed by two different materials. The term is normally employed in connection with semiconductor materials.

Interband Resonant Tunneling: A special case of resonant tunneling, in which the conduction and valence bands of two semiconductors are involved in the tunneling process.

Negative Differential Resistance: Electrical term that refers to a region in the current-voltage characteristic of an electronic device in which an increase in the applied voltage results in a decrease of the flow of current.

Potential Barrier: Mathematical concept used to describe a region of space where a repulsive force acts on a particle.

Quantum Well: Complementary to potential barrier, it refers to a region of space where an attractive force acts on a particle, or to a small region confined by two potential barriers.

Resonant Tunneling: A special case of tunneling, by which under certain conditions an electron has a high probability of traversing two consecutive potential barriers.

Tunneling: Quantum-mechanical phenomenon by which a moving particle can overcome a repulsive force and have a finite probability of reaching a region of space forbidden by classical mechanics.

WKB Approximation: Mathematical approximation, due to Wentzel, Kramers, and Brillouin, frequently used in quantum mechanics to calculate the transmission probability through a slowly varying potential barrier.

Works Cited

Ben Amor, S., Martin, K. P., Rascol, J. J. L., Higgins, R. J., Torabi, A., Harris, H. M., Summers, C. J. (1988), *Appl. Phys. Lett.* **53,** 2540–2542.

Brown, E. R. (1992), in: J. Shah (Ed.), *Hot Carriers in Semiconductor Nanostructures,* Boston: Academic, 469–498.

Büttiker, M. (1988), *IBM J. Res. Develop.* **32,** 63–75.

Capasso, F., Sen, S., Gossard, A. C., Hitchinson, A. L., English, J. H. (1986), *IEEE Electron Dev. Lett.* **7,** 573–576.

Chang, L. L., Esaki, L., Tsu, R. (1974), *Appl. Phys. Lett.* **24,** 593–595.

Chang, Y. C., Chang, L. L., Esaki L. (1985), *Appl. Phys. Lett.* **24,** 1324–1326.

Dellow, M. W., Beton, P. H., Langerak, C. J. G. M., Foster, T. J., Main, P. C., Eaves, L., Henini, M., Beaumont, S. P., Wilkinson, C. D. W. (1992), *Phys. Rev. Lett.* **68,** 1754–1757.

Dignam, M. M., Ashoori, R. C., Stormer, H. L., Pfeiffer, L. N., Baldwin, K. W., West, K. W. (1994), *Phys. Rev. B* **49,** 2269–2272.

Eaves, L. (1990), in: W. van Haeringen and D. Lenstra (Eds.), *Analogies in Optics and Micro Electronics,* Hingham, MA: Kluwer Academic Publishers, 227.

Esaki, L., Tsu, R. (1970), *IBM J. Res. Develop.* **14,** 61–65.

Gennser, U., Kesan, V. P., Syphers, D. A., Smith, III, T. P., Iyer, S. S., Yang, E. S. (1991), *Phys. Rev. Lett.* **67,** 3828–3831.

Goldman, V. J., Tsui, D. C., Cunningham, J. E. (1987), *Phys. Rev. B* **35,** 9387–9390.

Grönqvist, H., Kollberg, E., Rydberg, A. (1991), *Microwave Opt. Technol. Lett.* **4,** 33–35.

Gu, B.-Y., Gu, L. (1989), *Phys. Rev. B* **40,** 6124–6128.

Gueret, P., Blanc, N., Germann, R., Rothuizen, H. (1992), *Phys. Rev. Lett.* **68,** 1896–1899.

Hayden, R. K., Maude, D. K., Eaves, L., Valadares, E. C., Henini, M., Sheard, F. W., Hughes, O. H., Portal, J. C., Cury, L. (1991), *Phys. Rev. Lett.* **66,** 1749–1752.

Ioganse n, L. V. (1964), *Soviet Phys. JETP* **18,** 146–150.

Ismail, K., Antoniadis, D. A., Smith, H. I. (1989), *Appl. Phys. Lett.* **55,** 589–591.

Kane, E. O. (1969), in: E. Burstein and S. Lundqvist (Eds.), *Tunneling Phenomena in Solids,* New York: Plenum.

Luryi, S. (1985), *Appl. Phys. Lett.* **47,** 490–492.

McGill, T. C., Collins, D. A. (1993), *Semicond. Sci. Technol.* **8,** S1–S5.

Mendez, E. E., Wang, W. I., Ricco, B., Esaki, L. (1985), *Appl. Phys. Lett.* **47,** 415–417.

Mendez, E. E., Calleja, E., Wang, W. I. (1986a), *Phys. Rev. B* **34,** 6026–6029.

Mendez, E. E., Esaki, L., Wang, W. I. (1986b), *Phys. Rev. B* **33,** 2893–2896.

Mendez, E. E., Wang, W. I., Calleja, E., Goncalves da Silva, C. E. T. (1987), *Appl. Phys. Lett.* **50,** 1263–1265.

Mendez, E. E., Ohno, H., Esaki, L., Wang, W. I. (1991), *Phys. Rev. B* **43,** 5196–5199.

Mendez, E. E., Nocera, J., Wang, W. I. (1993), *Phys. Rev. B* **47,** 13937–13940.

Reed, M. A., Randall, J. N., Aggarwal, R. J., Matyi, R. J., Moore, T. M., Wetsel, A. E. (1988), *Phys. Rev. Lett.* **60,** 535–537.

Ricco, B., Azbel, M. Ya. (1984), *Phys. Rev. B* **29,** 1970–1981.

Seabaugh, A. C., Taddiken, A. H., Beam III, E. A., Randall, J. N., Kao, Y.-C., Newell, B. (1993), *IEEE IEDM Tech. Digest,* 419–422.

Sen, S., Capasso, F., Gossard, A. C., Spah, R., Hatchinson, A., Chu, S. (1987), *Appl. Phys. Lett.* **51,** 1428–1430.

Sollner, T. C. L. G., Goodhue, W. D., Tannenwald, P. E., Parker, C. D., Peck, D. D. (1983), *Appl. Phys. Lett.* **43,** 588–590.

Su, B., Goldman, V. J., Cunningham, J. E. (1992), *Science* **255,** 313.

Suemasu, T., Kohno, Y., Suzuki, N., Watanabe, M., Asada, M. (1993), *IEEE IEDM Tech. Digest,* 553–556.

Tsu, R., Esaki, L. (1973), *Appl. Phys. Lett.* **22,** 562–564.

Weil, T., Vinter, B. (1987), *Appl. Phys. Lett.* **50,** 1281–1283.

Yokoyama, N., Imamura, K., Muto, S., Hiyamizu, S., Nishi, H. (1985), *Jap. J. Appl. Phys.* **24,** L853–854.

Zaslavsky, A., Goldman, V. J., Tsui, D. C., Cunningham, J. E. (1988), *Appl. Phys. Lett.* **53,** 1408–1410.

Further Reading

Bohm, D. (1951), *Quantum Theory,* Englewood Cliffs, NJ: Prentice-Hall.

Chang, L. L., Mendez, E. E., Tejedor, C. (1991), *Resonant Tunneling in Semiconductors,* New York: Plenum.

Duke, C. B. (1969), *Tunneling in Solids,* New York: Academic Press.

Esaki, L. (1975), "Long Journey into Tunneling," in: *Les Prix Nobel en 1973,* Stockholm: Imprimerie Royale, P. A. Norstedt & Söner.

Liu, H. C., Sollner, T. C. L. G. (1994), "High-Frequency Resonant-Tunneling Devices," in: R. A. Kiehl, T. C. L. G. Sollner (Eds.), *High Speed Heterostructure Devices,* Semiconductors and Semimetals, Vol. 3, Boston: Academic.

Mendez, E. E. (1988), "Physics of Resonant Tunneling in Semiconductors," in: E. E. Mendez, K. von Klitzing (Eds.), *Physics and Applications of Quantum Wells and Superlattices,* New York: Plenum.

Mendez, E. E. (1988), "Applications of Resonant Tunneling in Semiconductor Heterostructures," in: C. R. Leavens, R. Taylor (Eds.), *Interfaces, Quantum Wells and Superlattices,* New York: Plenum.

Price, P. J. (1992), "Electron Tunneling in Semiconductors," in: T. S. Moss (Ed.) *Handbook on Semiconductors,* Vol. 1, Amsterdam: North Holland.

Wolf, E. L. (1985), *Principles of Electron Tunneling Spectroscopy,* Oxford: Oxford University Press.

RESONATORS, ACOUSTIC

See FILTERS AND RESONATORS, ACOUSTIC

REVERSIBLE AND IRREVERSIBLE MATERIALS DEFORMATION

See SOLIDS, MECHANICAL PROPERTIES OF

RHEOLOGY

Elliot Kearsley, *10413 Englishman Drive, Rockville, MD 20852*

Introduction 455
1. **Constitutive Equations** 456
1.1 Constraints 458
1.2 Continuum Mechanics 458
1.3 Linear Models 459
1.3.1 Newtonian Viscosity 460
1.3.2 Linear Elasticity 460
1.3.3 Linear Viscoelasticity 461
1.4 Nonlinear Models 463
1.4.1 Finite Elasticity 464
1.4.2 Differential Models 467
1.4.3 Models with Plastic Yield 469
1.4.4 Integral Models 470
1.4.5 Nonsimple Materials 471
2. **Rheometry** 472
2.1 Viscometry 472
2.2 Strain-Energy Functions 474
2.3 Moduli of Linear Viscoelasticity 475
2.4 Normal Stresses 476
2.4.1 Rheogoniometry—Cone and Plate 477
2.4.1.1 Steady-State Properties 477
2.4.1.2 Time-Dependent Properties 478
2.4.2 Rheometry—Other Shearing Devices 478
2.4.2.1 Torsional Flow 478
2.4.2.2 Hole Pressure and Channel Flow 479
2.5 Complex Flows 479
2.5.1 Orthogonal Rheometry 479
2.5.2 Extensional Rheometry 480
2.5.3 Plastometer 481
3. **Dimensionless Parameters** 482
Glossary 482
Works Cited 484
Further Reading 484

INTRODUCTION

Rheology as a distinct science was born as recently as 1928. In that year, the word "rheology" was coined by E. C. Bingham of Lafayette College to denote the study of the deformation and flow of materials. The following year, an international organization, the Society of Rheology, was formed to promote the science. From the beginning, it was a highly interdisciplinary field and, although rheology is considered a branch of physics and the Society is a member society of the American Institute of Physics, the subject has close connections to the fields of physical chemistry, mathematics, materials science, and engineering.

For a more detailed explanation of the meaning of rheology, recall that, while the three laws of Newton's mechanics suffice to determine the response of point masses to forces, to predict the response of extended continuous bodies it is necessary to append equations associated with the particular material of the body that describe how the material distorts or changes shape under the influence of forces. Such equations, because they are intrinsic to the material, are called *constitutive equations* or sometimes, in the simplest cases, stress–strain laws. Rheology is concerned with the problems of formulating, establishing, and applying the mechanical constitutive equations of materials. These equations are mathematical models couched in the language of continuum mechanics, but it is understood that there is a microstructure and micromechanics underlying them (e.g., macromolecules of a polymer, grain boundaries of a metal, etc.).

Prior to the Second World War, development of rheology was sporadic and generally empirical, concerned mostly with linear theories. During the war, many practical problems with rheological content such as the evaluation of synthetic rubbers, the design of flame-thrower fluid, and the design of radar domes drew the attention of scientists. In the course of solving them, rheological instruments were designed, nonlinear phenomena were examined, and the development of the necessary basis in continuum mechanics was begun. At the time, much of this work was

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

classified secret, but publication after the war showed the progress that had been made, opening up the field and stirring up considerable interest in the subject. In the decades from 1950 to 1980, mathematicians and mechanicians developed a basis for describing nonlinear rheological phenomenology, while polymer physicists and physical chemists established some of the connections between the parameters of molecular chains and the mechanical behavior of associated materials. Currently, rheological interest is broadening to encompass liquid crystals, electrorheological and magnetorheological fluids, suspensions, and granular and composite materials. The Society of Rheology now has about 1300 members, mostly in the United States, in disciplines ranging through physics, chemistry, engineering, medicine, and biology. There is an international Biorheology Society and numerous national rheology societies and groups.

One can attempt to classify the activities of rheologists under five rubrics:

1. formulation of constitutive equations,
2. relating constitutive equations to micromechanics,
3. testing the validity of constitutive equations,
4. measuring rheological properties of materials, and
5. predicting and controlling the mechanical behavior of materials.

Technological applications of rheology are ubiquitous. To indicate just a few examples: extruding and molding of plastics, rubbers and metals; handling of cement and slurries; lubrication; specification and control of materials; well drilling and oil extraction; sealants and paints; antiearthquake mounts; and food processing.

1. CONSTITUTIVE EQUATIONS

The formulation of a constitutive equation must take into account some rather subtle geometrical considerations. The internal forces between a particle of material and its neighboring particles are to be related to the motion of the material but, intuitively, it must be motion relative to an imbedded convected coordinate system, that is, as "seen" by the moving particle rather than by a fixed observer. A variety of prescriptions have been proposed for formulating constitutive equations satisfying this and other fundamental requirements. A discussion of differences between some of them can be found in Huilgol (1975).

An exposition of rheology requires a different level of mathematical notation than that necessary for classical fluid mechanics. The use of tensor fields as the variables of constitutive equations is the conventional way to simplify the geometrical problems encountered. In this article, bold-faced letters are used to denote tensors (in the general sense, including vectors, pseudotensors, tensor fields, etc.) The transpose and inverse of second-order tensors are indicated by superscripts T and -1, respectively. The trace, determinant, divergence, and gradient of a tensor, say $\mathbf{A}$, are denoted by tr$\mathbf{A}$, det$\mathbf{A}$, $\nabla \cdot \mathbf{A}$, and $\nabla\mathbf{A}$, respectively. Scalar invariants of second-order tensors will be written as $I_i(\mathbf{A})$, $i = 1, 2, 3$, except that the argument may be dropped when it is obvious from context.

One postulates a *stress tensor,* $\boldsymbol{\sigma}$, at each point of a material body, which carries the local forces of interaction between contiguous parts. The stress tensor is a real, second-order tensor in three-dimensional space, symmetric (with rare exceptions when there are body moments or couple stresses). Thus, normally, only six of its nine components are independent. The tensor may be viewed as a linear operator that acts on a vector representing area to transform it to a vector representing a force acting on the area. The physical components of a stress tensor have the dimension of force per unit area.

To describe the change of shape of a body, a *strain tensor* is defined. Strain tensors are not unique in that any invertible function of a suitable strain tensor is also suitable, so that a wide range of choices is possible. The behavior of a material may be described using any one of various strain tensors, although the form of the constitutive equation will reflect that choice.

If each point of a material body is associated with a vector giving its position in space at some instant, we call the resulting vector field the *configuration* of the body at that instant. A particular configuration, the *reference configuration,* given by position vectors $\mathbf{X}$, the *material coordinates,* is used to label the points of the body. It is often convenient to choose a reference configuration to be stress free and/or an initial configuration at some zero of time. If the points **X** of the reference configuration trace out a path in time given by the vector function $\boldsymbol{\chi}$ and **x** is the configuration at time t, a vector field, then $\mathbf{x} = \boldsymbol{\chi}(\mathbf{X},t)$. The gradient of **x** with respect to **X** is a double tensor field, $\mathbf{F}(t)$, called the *deformation gradient* (often, simply *deformation*) describing the change of shape that is experienced in the neighborhood of each point of the material in going from configuration **X** to configuration **x**. It is called a double tensor field because it depends on two configurations and acts as a tensor under coordinate transformation of either. The determinant of **F**, the relative change in volume induced by the deformation, must be greater than zero if **F** is to be physically possible. (We do not allow a finite volume of material to contract to zero or to a negative volume.)

As a direct consequence of the polar decomposition theorem, it is apparent from these conditions that **F** may be factored into a product of a proper orthogonal tensor and a positive-definite symmetric tensor. In other words, any deformation of the material is locally equivalent to three stretchings in certain mutually perpendicular directions followed by a pure rotation (or, alternatively, rotation followed by stretching). Thus,

$$\mathbf{F} = \mathbf{RU} = \mathbf{VR}, \tag{1}$$

where **R** is a proper rotation and **U** and **V** are positive-definite, symmetric tensors called the *right stretch tensor* and *left stretch tensor*, respectively. It is worth noticing that **U** represents the stretches before rotation while **V** gives the stretches in their final orientation. Thus the principal stretches, that is, the eigenvalues of **U** and **V**, are identical while their eigenvectors are not. The tensor **F** is not a suitable strain measure because its dependence on **R** (which represents a local rotation and not a true distortion of the material) is not just an expression of its orientation with respect to a reference configuration. The term **F** is said to be *observer dependent*. The stretch tensors, however, are suitable strain measures because for them **R** only specifies the change in orientation of the principal stretches and does not affect their values. But, in fact, the stretch tensors are seldom used as strain measures for the very practical reason that they are usually inconvenient to calculate.

Two tensors that are simply calculated from **F** and suitable for use as strain measures are **B** and **C**, the so-called *left and right Cauchy–Green strain tensors*, respectively, defined as follows:

$$\mathbf{B} = \mathbf{V}^2 = \mathbf{F}\mathbf{F}^{\mathrm{T}}, \tag{2}$$

$$\mathbf{C} = \mathbf{U}^2 = \mathbf{F}^{\mathrm{T}}\mathbf{F}, \tag{3}$$

where **B** and **C** are symmetric, positive-definite tensors of rank two whose physical components are dimensionless. Of course, any other invertible tensor function of **U** or **V** is also a possible strain measure.

A relative strain tensor can be formed by choosing the reference configuration to be that at some arbitrary time, say τ, and then the gradient of $\mathbf{x}(t)$ with respect to $\mathbf{x}(\tau)$ is the *relative deformation gradient* $\mathbf{F}_\tau(t)$, the subscript τ indicating the time of the reference configuration. Deformation gradients compound by multiplication, thus,

$$\mathbf{F}_\tau(t) = \mathbf{F}_\xi(t)\mathbf{F}_\tau(\xi), \tag{4}$$

where ξ refers to a time intermediate to τ and t.

It is important not to confuse the tensor **F** of Eq. (1) with this relative tensor. It is conventional among rheologists to use the base symbol from tensors with respect to a fixed reference configuration to designate the analogous relative tensor but to attach a subscript to indicate the arbitrary reference configuration. Equation (4) offers several examples of this convention.

By analogy with Eq. (3), the *right relative Cauchy–Green strain tensor* is defined by

$$\mathbf{C}_\tau(t) = \mathbf{F}_\tau^{\mathrm{T}}(t)\mathbf{F}_\tau(t) = \mathbf{U}_\tau^2(t). \tag{5}$$

This notation can be extended to various other strain tensors; thus $\mathbf{U}_\tau(t)$ is a right relative stretch tensor, $\mathbf{B}_\tau(t)$ is a left relative Cauchy–Green strain tensor, etc. The Cauchy–Green strain tensors and their inverses are commonly used in the rheological treatment of flows. The various strain tensors are related by equations of the form

$$\mathbf{B}_\tau(t) = \mathbf{C}_t^{-1}(\tau). \tag{6}$$

The tensor $\mathbf{B}_\tau(t)$ is often called the *relative Finger tensor.*

Most rheological constitutive equations that are commonly used are so-called *simple materials,* for which the stress at a material point at any time can be calculated from the history of the deformation. In this context, the term *history* means the set of values of a property at the point for that time and all prior times. In short, a simple material is one for which the stress is a *functional* of the local deformation history. There is a geometrical restriction to be satisfied by this functional, the so-called *principle of material frame-indifference* (sometimes called *material objectivity*) which says, in loose terms, that the properties of the material described by the constitutive equation must be independent of the observer. It is for this reason that we rejected **F** as a strain measure. This restriction, when worked out, leads to a representation of a general simple material in the following form:

$$\boldsymbol{\sigma}(\mathbf{x},t) = \mathbf{F}\mathcal{F}^{t}_{\tau=-\infty}(\mathbf{C}(\mathbf{X},\tau))\mathbf{F}^{\mathrm{T}}, \tag{7}$$

where $\mathcal{F}(\cdot)$ is a tensor-valued functional of the history of its argument, in this case $\mathbf{C}(\mathbf{X},\tau)$ for τ ranging from $-\infty$ to t. The pre- and postmultiplications of the functional by deformation gradients in this equation are necessary to express the current stress in laboratory coordinates.

1.1 Constraints

The simple material of Eq. (7) is a useful theoretical concept, but it is much too general for practical applications. Somewhat more useful restricted models of ideal materials are defined in terms of *constraints,* which are conditions limiting the range of possible deformations. The constraints are themselves constitutive and must satisfy material objectivity. Of particular interest are *workless constraints,* maintained by implicit, internal forces that do no work in the course of any deformation allowed by the constraints. These forces of constraint do not directly affect the energetics. As a consequence, the stresses arising from the local history of constrained deformations are only part of the complete stress since they do not include that part of the stress associated with the workless constraints.

A very important example of a workless constraint, ubiquitous in rheology, is that of *incompressibility,* for which the deformation is isochoric, i.e., volume preserving. In this case, an unspecified isotropic stress is associated with maintaining the constraint and must be added to stress determined by local deformation history. As a result, only the deviatoric part of the stress is completely determined by the local deformation. Accordingly, Eq. (7) must then be modified to read

$$\boldsymbol{\sigma}(\mathbf{X},t) - p\mathbf{1} = \mathbf{F}\mathcal{F}^{t}_{\tau=-\infty}(\mathbf{C}(\mathbf{X},\tau))\mathbf{F}^{\mathrm{T}}, \quad \det\mathbf{C} = 1, \tag{8}$$

where p is an arbitrary pressure, $\mathbf{1}$ is the unit tensor, and $\mathbf{C}$ is restricted to be unimodular to assure an isochoric deformation. Of course, one does not expect to find real materials that are strictly incompressible any more than one expects to find ideal gases in thermodynamics. The model is an excellent approximation in many cases, however, when the energy associated with volume changes is ignorable compared with the energy of shearing, as is usually the case in stretchings and flows of elastomeric materials and of flowing liquids. Incompressibility is a widely used constraint of great importance. One of its most valuable aspects is that, for models of materials with this constraint, it is possible to write useful and exact expressions for the stress produced by certain simple, highly symmetric, inhomogeneous deformations.

If models are to be of practical use, further details of the form of a particular $\mathcal{F}$ will reflect other properties of the material such as isotropy or any anisotropy, its state of homogeneity, the limits and mode of the influence of past configurations, etc.

1.2 Continuum Mechanics

Such constitutive equations are useful only within the context of continuum mechanics. The general principles of continuum mechanics are contained in field equations expressing conservation of mass, the balances of linear momentum and moment of momentum, and, if thermodynamic effects are pertinent, appropriate forms of the first and second laws of thermodynamics. When these fields vary smoothly in the interior of a body of material, the equations are conveniently expressed in the form of differential equations. The mechanical field equations may be written as follows:

$$\dot{\rho} + \rho(\nabla \cdot \mathbf{v}) = 0 \quad \text{(conservation of mass)}, \tag{9}$$

$$\nabla \cdot \boldsymbol{\sigma} + \rho \mathbf{b} = \rho \dot{\mathbf{v}} \quad \text{(balance of momentum)}, \tag{10}$$

with $\dot{\mathbf{v}} = \partial \mathbf{v}/\partial t + (\mathbf{v} \cdot \nabla)\mathbf{v}$ and $\dot{\rho} = \partial \rho/\partial t + \mathbf{v} \cdot \nabla \rho$; and

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}^{\mathrm{T}} \quad \text{(balance of moment of momentum)}, \tag{11}$$

where $\rho(\mathbf{X},t)$ is the density; a superposed dot indicates the *material derivative,* that is, the time derivative seen by a material particle; ∇ is the gradient operator; $\mathbf{v}$ is $\partial \mathbf{x}/\partial t$, the particle velocity; $\mathbf{b}$ is the body force, if any, acting on the particle; and Eq. (11) expresses the symmetry of the stress tensor in the absence of couple stresses as mentioned above. In addition, if thermal effects are pertinent, appropriate forms of the first and second laws of thermodynamics must be included with these principles. To complete the picture, to these general equations must be appended the constitutive equations that model some particular material and boundary conditions suitable to formulate a well-posed mathematical problem.

In a typical problem of continuum mechanics, the *surface tractions,* forces acting on the surface of the body, are given as well as the body forces (usually zero) or, alternatively, the surface is assigned a given shape or velocity. From the field equations and these boundary conditions, the deformation and stress in the interior are to be found. Usually, especially when there are nonlinearities present, there is no assurance that a solution exists or that it is unique even if it does. As a result, almost all known, nontrivial, exact solutions are so-called *inverse solutions* for which a tentative set of deformations is assumed, characterized by adjustable parameters and compatible with the given boundary conditions and constraints. The associated stress can then be calculated through the constitutive equation, and it is then only necessary to demonstrate that this stress can be made consistent with the field equations above by suitable adjustment of the parameters. This program can sometimes be carried out and an inverse solution found for certain constraints in highly symmetric geometries.

For incompressible materials, because Eqs. (9) and (11) are satisfied by assumption, only Eq. (10) need be considered. In that case, when there is sufficient symmetry to the problem, for a large class of constitutive equations and deformations, the arbitrary pressure term of the stress can be adjusted to satisfy Eq. (10). The adjective "controllable" is sometimes used to characterize this class of deformations.

To treat the great majority of rheological boundary-value problems for which these inverse methods are not sufficient, computational solutions must be sought through finite-element methods or other techniques for numerical solution of boundary-value problems. There is a large literature in the numerical analysis necessary for such computations because inverse solutions are rare. Furthermore, the recent amazing advances in computer simulation have made such methods often quite convenient, particularly where complex geometry and little symmetry are involved. Still, the solutions found by numerical methods usually do not have the convenient simplicity of inverse solutions.

1.3 Linear Models

The term "linear," as used here, is meaningless without an understanding of what variables depend linearly on what independent variables. For a linear stress-strain model, doubling the strain will cause a doubling of the stress. More generally, for a linear simple material, if a strain history is the sum of several simple strain histories then the resulting stress is the sum of the partial stresses associated with the simple histories. This is known as the *Boltzmann superposition principle.* This principle certainly cannot be generally valid because stresses compound additively, while strains do not. On the other hand, infinitesimal strains do compound additively (to first order in the strains) so that linear constitutive equations can be useful for treating small deformations. Furthermore, in the case of Newtonian fluids, the rate of strain (which compounds additively) rather than the strain history is relevant, so that, in this case too, there is no inherent rheological limitation to small deformations.

1.3.1 Newtonian Viscosity The *Newtonian fluid* obeys a classical constitutive equation in which the current rate of strain, rather than the whole strain history, determines the current stress. Remarkably, it is an approximation valid for many common liquids with compact molecules such as simple hydrocarbon oils or water over very large ranges of deformation rate. The concept of Newtonian fluids is so widely known and used in fluid mechanics that it is not really of much interest to the rheologist except as a model of the simplest common liquids. However, it is worthwhile here to formulate its constitutive equation in terms of deformation tensors in preparation for generalizing to models of more complex fluids.

The velocity gradient $\mathbf{L}$ is the rate of change of the deformation gradient:

$$\mathbf{L}(t) = \nabla\mathbf{v}(t) = \left.\frac{d\mathbf{F}_t(\tau)}{d\tau}\right|_{\tau=t}. \qquad (12)$$

The velocity gradient is not a materially objective quantity but it may be split into antisymmetric and symmetric parts, the rate of rotation of a particle, given by the *spin tensor* **W**, and the rate of stretching, given by the *deformation-rate tensor* **D**:

$$\mathbf{W} = \tfrac{1}{2}(\nabla\mathbf{v} - \nabla\mathbf{v}^{\mathrm{T}}), \quad \mathbf{D} = \tfrac{1}{2}(\nabla\mathbf{v} + \nabla\mathbf{v}^{\mathrm{T}}). \qquad (13)$$

The deformation rate is materially objective. The constitutive equation for an incompressible Newtonian fluid expresses the linear dependence of stress on the deformation rate, viz.,

$$\boldsymbol{\sigma} - p\mathbf{1} = 2\eta_0\mathbf{D}, \quad \mathrm{tr}(\mathbf{D}) = 0, \qquad (14)$$

where η_0 is a constant, the *Newtonian viscosity,* and p is the usual arbitrary pressure associated with incompressible media. Two characteristic properties of this model are common to all models that we would call fluids: The stress relaxes to an isotropic pressure when the strain rate goes to zero; further, any deviatoric component of stress induces an immediate nonzero rate of strain.

When Eq. (14) is put into Eq. (10), the classical Navier–Stokes equation for an incompressible Newtonian fluid results.

1.3.2 Linear Elasticity A purely elastic medium is one for which current stress is a function of current strain, that is, only the current configuration and an undeformed (that is, stress-free) configuration play a role in determining stress. *Classical infinitesimal elasticity* is based on a linear constitutive equation, an expression of *Hooke's law.* This approximation can be valid only when stretches are small and the effects of rotation unimportant. Nonetheless, it is very useful in describing brittle materials and metals in most common usages for which they are not subject to large deformations. There is a large engineering literature of various important problems in linear elasticity.

With **X** as the stress-free reference configuration, the *infinitesimal strain tensor* $\boldsymbol{\epsilon}$ is defined as

$$\boldsymbol{\epsilon} = \tfrac{1}{2}(\mathbf{F} + \mathbf{F}^{\mathrm{T}}) - \mathbf{1}. \qquad (15)$$

Then, Hooke's law for isotropic media can be written in the form

$$\boldsymbol{\sigma} = 2\mu\boldsymbol{\epsilon} + \lambda\,\mathrm{tr}(\boldsymbol{\epsilon})\mathbf{1}, \qquad (16)$$

where μ and λ are the classical *Lamé constants.* The familiar Young's modulus, E, and Poisson's ratio, ν, are related to them as follows:

$$\nu = \frac{\lambda}{2(\lambda + \mu)}, \quad E = \mu\frac{3\lambda + 2\mu}{\lambda + \mu}. \qquad (17)$$

By introducing the deviators **s** and **e** of $\boldsymbol{\sigma}$ and $\boldsymbol{\epsilon}$, respectively, Eq. (16) can be split into a scalar equation and a deviatoric equation:

$$\begin{aligned} &\mathbf{s} = 2\mu\mathbf{e}, \quad \mathrm{tr}(\boldsymbol{\sigma}) = (2\mu + 3\lambda)\,\mathrm{tr}(\boldsymbol{\epsilon}), \\ &\mathbf{s} = \boldsymbol{\sigma} - \tfrac{1}{3}\,\mathrm{tr}(\boldsymbol{\sigma})\mathbf{1}, \quad \mathbf{e} = \boldsymbol{\epsilon} - \tfrac{1}{3}\,\mathrm{tr}(\boldsymbol{\epsilon})\mathbf{1}. \end{aligned} \qquad (18)$$

An incompressible, isotropic, linear-elastic material has only one constitutive constant, an elastic modulus with dimension force/area. The constitutive equation can then be written in terms of the deviators:

$$\mathbf{s} = M\mathbf{e}, \qquad (19)$$

where M is the *shear modulus* and (in the linear approximation) **e** is equal to $\boldsymbol{\epsilon}$ since only isochoric deformations are allowed. The term modulus refers to a ratio of stress to strain, and the reciprocal of a modulus is a *compliance*.

The linear-elastic material is easily recognized as a solid because the stress deviator does not vanish no matter how long a shearing deformation is held.

1.3.3 Linear Viscoelasticity Between fluid and solid is *viscoelastic* matter which has a characteristic time and exhibits time-dependent responses. Viscoelastic materials show *stress relaxation* and *elastic recoil,* fluidlike and solidlike behavior, respectively. If a deformation is applied to a viscoelastic body and then held fixed, the stress will not remain fixed as with a purely elastic solid but will relax with time, perhaps to an isotropic pressure as with a fluid. On the other hand, if a deformation is produced by surface tractions and held for a time and then all applied tractions are suddenly released, the body will recoil as with an elastic solid, but will do so over a time span significant with respect to the characteristic time, perhaps, eventually, even returning to its original configuration. If the original configuration is not completely recovered, the residual deformation is called *permanent set.* Mechanical engineers sometimes distinguish between "viscoelastic" and "elasticoviscous" according to whether permanent set occurs, but this distinction is often unclear since relaxation is usually approximately exponential and it is generally difficult to tell whether one has waited long enough to establish permanent set. In practice, a more arbitrary definition of permanent set may be used.

If a traction is applied to a viscoelastic body and held constant, the body will *creep;* that is, the body will continue to deform for a time longer than its characteristic time—perhaps, if it is fluidlike, for as long as the traction is maintained.

A *Maxwell element* is a mechanical analog of a constitutive equation that exhibits simple exponential force relaxation. It consists of an elastic spring and a viscous dashpot connected in series fixed at one end and pulled by a force at the other so that the displacement is the sum of the responses of spring and dashpot. The characteristic time for the force at fixed displacement to relax exponentially to $1/e$ of its initial value, the quotient of the viscosity of the dashpot and the spring constant, is called the *relaxation time.*

A *Kelvin–Voigt element* connects the same components in parallel so that the force acting across the element is the sum of contributions from the spring and dashpot. Under a constant force, the displacement approaches a value determined by the spring constant. The displacement rate is a simple exponential with a characteristic time called the *retardation time.*

If a Kelvin–Voigt element is combined in series with another dashpot, a model is formed with both a retardation time and a relaxation time. When stress and strain deviators are substituted for the force and displacement in the equation for this mechanical analog, the result is the *Jeffreys model* with the following constitutive equation:

$$\boldsymbol{\sigma} + \lambda_{\mathrm{rlx}} \frac{d\boldsymbol{\sigma}}{dt} = \eta\left[\mathbf{D} + \lambda_{\mathrm{rtd}} \frac{d\mathbf{D}}{dt}\right], \tag{20}$$

where λ_{rlx} and λ_{rtd} are the relaxation time and retardation time, respectively, and η is a viscosity.

In a similar way, any number of these elements may be strung together into a passive, mechanical network analog of some particular, linear, constitutive equation. Substitution of stress and strain deviators for force and displacement in the equations for these networks leads to linear, viscoelastic constitutive equations in differential form describing isotropic, incompressible materials, valid for infinitesimal deformations. In general, such equations can be written as

$$P(\mathbf{s}) = Q(\mathbf{e}), \tag{21}$$

where P and Q are linear differential operators of the form

$$P = \sum_{r=0}^{p} p_r \frac{\partial^r}{\partial t^r}, \quad Q = \sum_{r=0}^{q} q_r \frac{\partial^r}{\partial t^r}, \tag{22}$$

in which the p_r and q_r are constitutive constants. (If the material is compressible, an analogous scalar equation on the isotropic parts of the stress and strain must be added to complete the picture.)

The differential equation, Eq. (21), may be seen to be of the same type as that which governs electrical networks of resistors and capacitors. Therefore, a similar analogy between linear viscoelasticity and passive electrical networks may be set up.

One advantage of constitutive equations formulated as linear differential equations is that an application of the Laplace transform to Eq. (21) removes the time variable and often allows the time-dependent, stress-analysis problem of a viscoelastic body to be replaced by an associated time-independent, linear-elastic problem whose boundary conditions are the transformed original boundary conditions. The solution of the viscoelastic problem is then found from the solution of the simpler elastic problem by inverse transformation.

A more common formulation of constitutive equations for linear-elastic media is in the form of integrals. Since the stress–strain relation is linear, one can superpose infinitesimal deformations to write a general stress equation for an incompressible, isotropic, viscoelastic material as follows:

$$\mathbf{s}(t) = \int_0^\infty \mathbf{e}(t - \xi)\frac{dG(\xi)}{d\xi}\,d\xi, \tag{23}$$

where $G(t)$ is the *stress-relaxation modulus*, the stress–strain ratio at time t as a result of a small constant strain applied at time zero on a body initially unstressed and at equilibrium. [The quantity $G(t)$ is a scalar since linearity implies that stress and strain are proportional.] Equation (23) expresses the stress as a summing up of contributions from configurations ranging from the present to the past.

The modulus $G(t)$ may be the stress-relaxation response of a mechanical network such as a set of Maxwell elements in parallel. Such a model would have a spectrum of relaxation times, each weighted by a modulus, so that $G(t)$ would be a sum of weighted exponential terms. On extending this idea to an infinite number of infinitesimally spaced relaxation times, a modulus with a continuous relaxation spectrum arises:

$$\begin{aligned} G(t) &= G_\infty + \int_0^\infty g(\tau)e^{-t/\tau}d\tau \\ &= G_\infty + \int_{-\infty}^\infty H(\ln\tau)e^{-t/\tau}d\ln\tau. \end{aligned} \tag{24}$$

The constant G_∞ has been added to allow a possible nonrelaxing elastic component. The function, $g(\tau)$, is called the *relaxation spectrum*. A logarithmic time scale is usually more practical and $H(\ln\tau)$, the *logarithmic relaxation spectrum*, is often encountered in the literature. The graph of any monotonically decreasing relaxation can be fitted with some appropriate relaxation spectrum. Much work has been done by polymer physicists in relating these spectra for polymeric materials to molecular structure.

If the strain history $\mathbf{e}(\tau)$ is zero from all past time up to time zero, Eq. (23) becomes

$$\mathbf{s}(t) = G(0)\mathbf{e}(t) + \int_0^t \mathbf{e}(t - \xi)\frac{dG(\xi)}{d\xi}\,d\xi, \tag{25}$$

where the first term on the right arises from the integration of the delta function associated with the jump of the stress-relaxation function at $\xi = 0$.

An equivalent constitutive relation can be formed from Eq. (25) by a change of dummy variables, $\tau = t - \xi$, and an integration by parts:

$$\mathbf{s}(t) = G(t)\mathbf{e}(0) + \int_0^t G(t - \tau)\frac{d\mathbf{e}(\tau)}{d\tau}\,d\tau. \tag{26}$$

Here, the initial term on the right is a consequence of a step discontinuity in strain history at $t = 0$.

An alternative form of the equations of an isotropic, incompressible, linear-viscoelastic medium might have been obtained by taking the current strain as determined by the history of the stress deviator, that is, by reversing the roles of stress and strain in the preceding. In that case, the equivalent of Eq. (25) would be

$$\mathbf{e}(t) = J(0)\mathbf{s}(t) + \int_{-\infty}^t \mathbf{s}(t - \xi)\frac{dJ(\xi)}{d\xi}\,d\xi, \tag{27}$$

where $J(t)$ is the *creep compliance*, the ratio of strain at time t to a constant stress that was applied to an undisturbed material at equilibrium at time zero.

From Eqs. (25) and (27), it is clear that the creep compliance and the stress-relaxation modulus are related. The relation connecting them is the following:

$$\int_0^t G(\tau)J(t - \tau)d\tau = 1. \tag{28}$$

The Laplace transform of this equation can be used to express the transform of either function in terms of the transform of the other, and, in principle, it allows one function to be calculated from the other.

It is often convenient to examine viscoelastic materials under steady-state sinusoidal excitation. If the deformation is small enough to justify linear analysis, the usual complex notation can be used to describe the response of the system. Thus, with the complex deformation history $\tilde{\mathbf{e}}(t) = \mathbf{e}_0 e^{i\omega t}$ and the corresponding complex stress $\tilde{\mathbf{s}}(t)$, the response is given by a complex modulus, $G^*(\omega)$, a function of frequency:

$$\tilde{\mathbf{s}}(t) = G^*(\omega)\tilde{\mathbf{e}}(t), \tag{29}$$

where $G^*(\omega)$ can be broken into a sum of real and imaginary parts each expressible in terms of the stress-relaxation modulus,

$$G^*(\omega) = G'(\omega) + iG''(\omega), \tag{30}$$

where $G'(\omega)$ and $G''(\omega)$, the *storage modulus* and *loss modulus*, respectively, are given by the following:

$$G'(\omega) = G(0) + \int_0^\infty \frac{dG(\tau)}{d\tau} \cos\omega\tau d\tau, \tag{31}$$

$$G''(\omega) = -\int_0^\infty \frac{dG(\tau)}{d\tau} \sin\omega\tau d\tau. \tag{32}$$

With the complex modulus, a useful equivalent to Eq. (21) can be written through use of the Fourier transform defined as follows:

$$\bar{g}(\omega) = \int_{-\infty}^{\infty} g(t)e^{-i\omega t}dt, \quad g(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \bar{g}(\omega)e^{i\omega t}d\omega, \tag{33}$$

where $\bar{g}(\omega)$ is the Fourier integral transform of $g(t)$. Then the Fourier-transformed, linear, viscoelastic, constitutive equation for an isotropic incompressible material takes the following compact form:

$$\bar{\mathbf{s}}(\omega) = G^*(\omega)\bar{\mathbf{e}}(\omega). \tag{34}$$

Of course, all these linear-viscoelastic equations can be generalized to describe compressible and/or anisotropic materials by decomposing the material properties into geometrically distinct parts and writing analogous equations for each. For example, compressible isotropic materials may be handled with the addition of another equation relating the traces of the stress and strain history.

1.4 Nonlinear Models

Linearity imposes very severe restrictions on a constitutive equation as is implied by the Boltzmann superposition principle. Consequently, the form of a linear constitutive equation is completely determined from little more than the choice of dependent and independent variables and the symmetries of the model. For materials with elasticity, linear constitutive equations can be valid only to infinitesimal deformations. On the other hand, nonlinear constitutive equations, freed from the restrictions of superposition, have great flexibility and can describe an infinite range of mechanical behaviors, often admitting phenomena quite unexpected to anyone familiar with only the linear equations of classical rheology. In fact, most materials do exhibit strongly nonlinear behavior: foods, biological materials, slurries and suspensions, soils, rocks, plastics, and rubbers, to name just a few. All elastic materials are nonlinear to some extent.

The following are some examples of the often surprising, inherently nonlinear phenomena that are observed.

1. *Poynting effect.* If a cylinder of elastic material, a length of wire or a rubber cylinder, is twisted about its axis by a force couple, the axis will lengthen. If the length is to be held fixed, a compressive axial force will be required and in certain configurations, buckling may occur.
2. *Weissenberg effect.* A phenomenon loosely related to the Poynting effect can be observed in fluids. If a rotating rod is lowered into a fixed cup of elastic fluid, the fluid will climb the rod in apparent defiance of gravity (see Fig. 1). This phenomenon is familiar to anyone who has used an eggbeater or cake mixer.
3. *Drag reduction* or *Toms effect.* The dissolution of minuscule amounts of high–molecular-weight polymer into fluids in turbulent flow will often produce amazing results. The rate of flow through a pipe

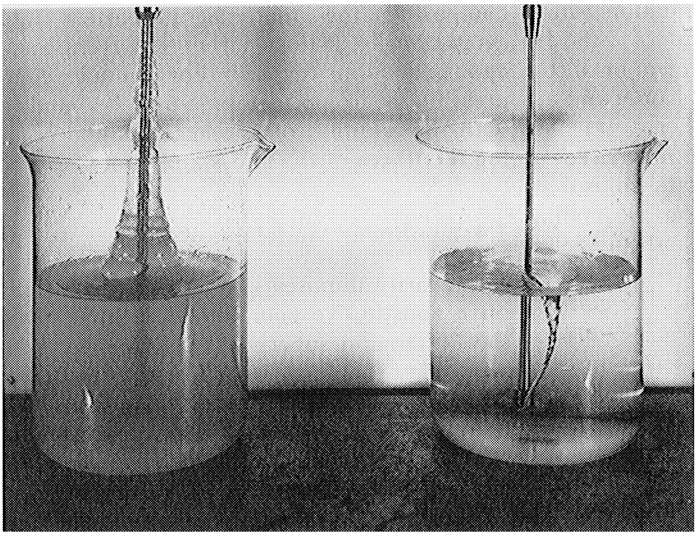

FIG. 1. The Weissenberg effect. The non-Newtonian fluid on the left (Separan AP-30) climbs the rotating rod while the Newtonian fluid (glycerol) on the right has a surface depression caused by inertia. Illustration taken from Tanner (1985), with permission from Clarendon Press.

between pressure reservoirs, for instance, can drastically increase. The details of the mechanism for this effect are still not entirely clear, but it is surely related to the nonlinearities induced by the solute.

4. *Die swell* or *Merrington effect.* An extrudate of molten plastic from a die of circular cross section will swell markedly as it leaves the die. In some cases, the diameter of the extrudate can swell to two or three times that of the die. In contrast, an extruded isothermal stream of Newtonian liquid at low Reynolds number may show a shrinking diameter; even at vanishing viscosity the swell is limited to about 15%. Die swell is very important in fiber spinning and other commercial processes.
5. *Tubeless siphon* or *Fano effect.* If a vertical tube is used to withdraw elastic liquid from a large reservoir, when the entrance of the tube is raised above the liquid level, the liquid may continue to flow up to the tube in a column supported by elastic forces (see Fig. 2).
6. *Flow through noncircular pipes.* An elastic liquid flowing through a straight tube of noncircular cross section will generally develop circulations in the plane of the cross section in addition to the flow along the axis. This would not be expected for a Newtonian fluid in isothermal laminar flow.

1.4.1 Finite Elasticity A purely elastic material is one with stress a simple function of strain, rather than a functional of strain history. If it is an isotropic material then, for purely geometrical reasons, the principal directions of the stress tensor and the strain tensor must coincide and the appropriate constitutive equation can be written as

$$\boldsymbol{\sigma} = \Lambda_0 \mathbf{1} + \Lambda_1 \mathbf{B} + \Lambda_2 \mathbf{B}^2, \tag{35}$$

where **B** is the left Cauchy–Green strain tensor with respect to a stress-free, undeformed state and the three scalars Λ_0, Λ_1, Λ_2 are scalar functions of the principal scalar invariants of **B**. Powers of **B** higher than the square need not appear in this equation because they can be reduced to polynomials of lower powers through the Hamilton–Cayley theorem.

Thermodynamic considerations imply a further condition on the constitutive equation for isothermal or adiabatic deformations of elastic media. In these cases, the work done in deforming the body from its undeformed state to any other equilibrium state of deformation is independent of the intervening deformations. Thus, a unique scalar quantity, the *strain energy,* can be associated with each state of deformation. Rigid rotations do not affect the strain energy; only the stretches contribute, and the strain energy of an isotropic, elastic material can be written as a function of the three principal scalar in-

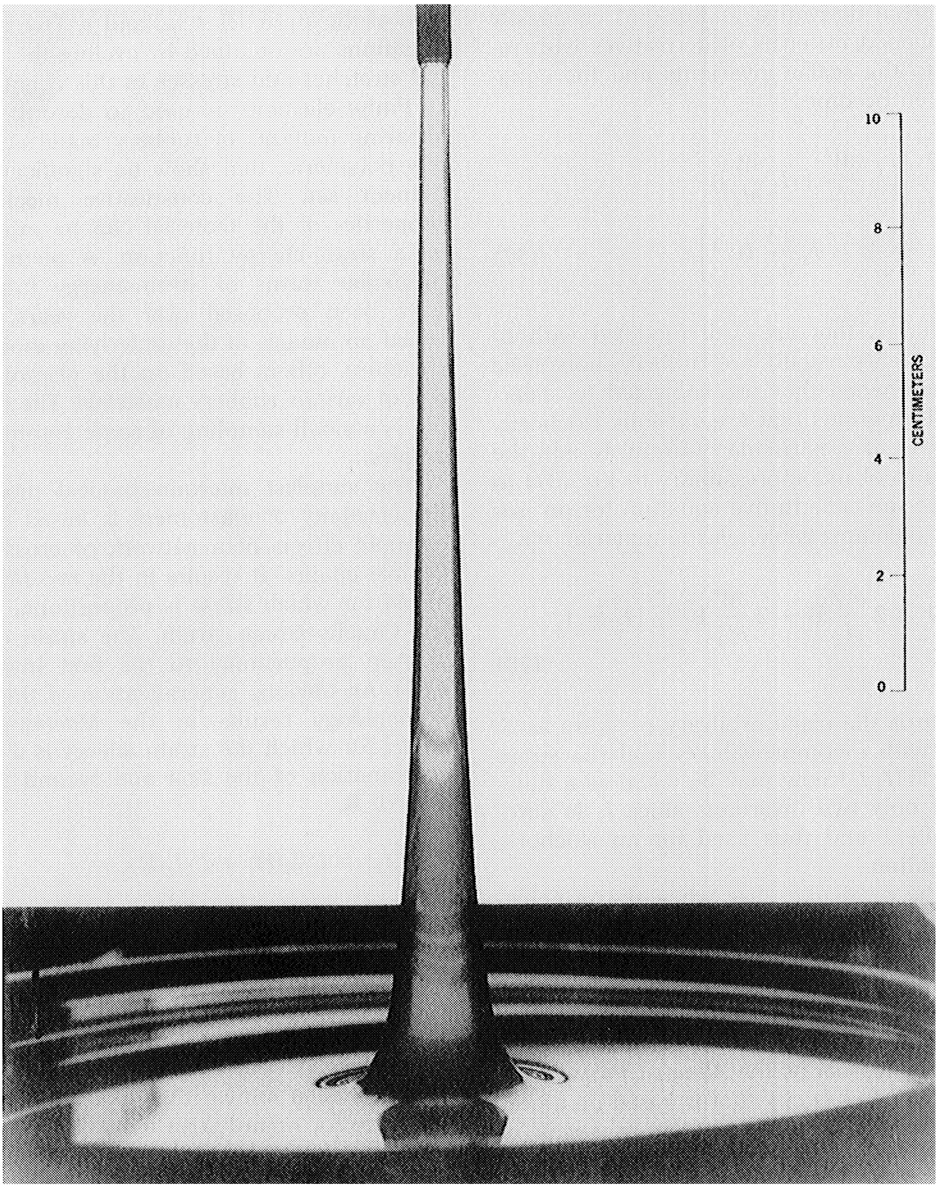

FIG. 2. The tubeless siphon or Fano effect. The siphon is begun with the tube below the surface of the elastic liquid. The liquid is sucked into the tube while the tube is raised above the surface, ultimately supporting a column of liquid. Properties of the liquid in extensional flow are related to the shape of the column. Illustration taken from Peng and Landel (1976) with permission.

variants of **B**. The result is called the *strain-energy function,* usually represented by $W(I_1,I_2,I_3)$. As a convention, the strain-energy function is written as energy per volume of material in the undeformed configuration. The mechanical properties of the material are completely contained in this function. For isothermal or adiabatic deformations, a purely elastic material has the following constitutive equation:

$$\boldsymbol{\sigma} = \frac{2}{\sqrt{I_3}} \mathbf{B} \frac{\partial W(I_1,I_2,I_3)}{\partial \mathbf{B}} . \tag{36}$$

The partial derivative in Eq. (36) can easily be expanded in terms of derivatives with respect to the scalar invariants and the equation then becomes

$$\boldsymbol{\sigma} = \frac{2}{\sqrt{I_3}}\left[\left(I_2\frac{\partial W}{\partial I_2} + I_3\frac{\partial W}{\partial I_3}\right)\mathbf{1} + \frac{\partial W}{\partial I_1}\mathbf{B} - I_3\frac{\partial W}{\partial I_2}\mathbf{B}^{-1}\right]. \tag{37}$$

Materials that are well modeled with finite elasticity usually are rubbery, polymeric materials when they are subjected to shearing deformations that are virtually isochoric. Therefore, it is particularly useful to add the constraint of incompressibility to Eq. (37) to produce the constitutive equation for an isotropic, incompressible, elastic material, viz.,

$$\boldsymbol{\sigma} - p\mathbf{1} = 2\frac{\partial W}{\partial I_1}\mathbf{B} - 2\frac{\partial W}{\partial I_2}\mathbf{B}^{-1}, \quad I_3 = 1, \tag{38}$$

where p is the usual arbitrary pressure associated with incompressibility and the strain energy $W(I_1,I_2)$ now may be taken as a function of only two invariants since I_3 is automatically 1 and thus fixed for an isochoric deformation.

In the particular case when the principal directions of strain are known, it is sometimes useful to replace Eq. (38) with relations for the principal stresses in terms of the principal stretches, the principal values of the stretch tensor, **V**. The result is a set of equations known as the *Almansi form* of the equations of elasticity that is based on an expression of the strain energy in terms of stretches. Since the principal stresses are necessarily oriented parallel to the principal strains, we will identify each stress by a subscript of the corresponding stretch. With this notation, the Almansi form of Eq. (38) can be written

$$\begin{aligned} \sigma_\lambda - p &= \lambda\frac{\partial \bar{W}(\lambda,\mu,\nu)}{d\lambda}, \\ \bar{W}(\lambda,\mu,\nu) &= W(I_1,I_2), \\ \lambda\mu\nu &= \sqrt{I_3} = 1, \end{aligned} \tag{39}$$

where σ_λ is the component of stress in the direction of principal stretch λ and the Almansi strain energy $\bar{W}$ is a symmetric function of the stretches λ, μ, and ν. Two similar equations are obtained by cycling the principal stretches and stresses in this equation.

Finite elasticity is used to describe large shearing motions of rubbery materials, usually polymeric, that show no significant permanent set. The constitutive mechanical properties of the material can be expressed as a strain-energy function. A plethora of particular forms of strain-energy functions have been proposed over the years, some based on models of the underlying molecular structure, others based on the phenomenology of various rubbery materials. The following is a small sampling of some common examples.

The simplest micromechanical model of the elasticity of elastomers is based on the entropic effects of a network of crosslinked flexible chains. It results in the *neo-Hookean model* for which stress is proportional to the left Cauchy–Green strain. The strain energy is then proportional to the first invariant, $I_1(\mathbf{B})$. An obvious generalization of this phenomenology results in the *Mooney–Rivlin model* for which the strain energy is a linear combination of the first and second invariants of **B**,

$$W(I_1,I_2) = C_1I_1(\mathbf{B}) + C_2I_2(\mathbf{B}). \tag{40}$$

Here, C_1 and C_2 are constants. The neo-Hookean model is obtained by setting C_2 equal to zero. The Mooney–Rivlin model has been widely used as the simplest model that predicts behavior at least qualitatively similar to that observed with rubbery materials, but it has long been known that it is seriously inadequate for quantitative modeling.

Rubber elasticity is known to be associated predominantly with the entropy of deformation, and this entropy is often assumed to be a sum of independent contributions from each of the principal stretches. These ideas lead to the formulation of the *Valanis–Landel model,* for which the value of the strain energy is a sum over the values of a function of the principal stretches, that is,

$$\begin{aligned} \bar{W}(\lambda,\mu,\nu) &= f(\lambda) + f(\mu) + f(\nu), \\ \bar{W}(\lambda,\mu,\nu) &= W(I_1,I_2), \end{aligned} \tag{41}$$

where $f(\cdot)$ is a constitutive function of a principal stretch that we shall call the *V-L function.* In this case, Eq. (39), the Almansi form, is often useful for calculating stresses from deformations.

These models with strain energy in terms of principal stretches were considered by Mooney (1940) as special cases of a general expression for strain energy as a power series in the invariants. Valanis and Landel (1967) developed the model further and, subsequently, a number of specific forms for the V-L function, $f(\cdot)$, have been proposed.

Examples of strain energies for compressible elastic materials are somewhat rare. One example is that of Blatz and Ko (1962), used to describe deformations of sponge rubbers. It can be written in our notation as follows:

$$\begin{aligned} W(I_1,I_2,I_3) = {} & \frac{\mu f}{2}\left[I_1 - 3 + \frac{1}{\alpha}(I_3^{-\alpha} - 1)\right] \\ & + \frac{\mu(1-f)}{2}\left[\frac{I_2}{I_3} - 1 + \frac{1}{\alpha}(I_3^{\alpha} - 1)\right], \\ & \alpha = \nu/(1 - 2\nu), \quad 1 \geq f \geq 0, \end{aligned} \tag{42}$$

where μ and ν are the shear modulus and Poisson's ratio, respectively, as observed at small deformations. The constant f is here an additional constitutive constant.

1.4.2 Differential Models Because the deformation rate $\mathbf{D}$ is a materially objective tensor, the most general nonlinear model for which stress is a function of deformation rate alone can be simply constructed by replacing the right-hand side of the model for a Newtonian fluid, Eq. (14), with a general isotropic tensorial function of $\mathbf{D}$,

$$\boldsymbol{\sigma} - p\mathbf{1} = 2\eta\mathbf{D} + 4\nu_2\mathbf{D}^2, \tag{43}$$

where η and ν_2 are material functions of the scalar invariants of $\mathbf{D}$. This is the *Reiner–Rivlin model.* It is said to be an inelastic model since it describes viscous behavior but does not describe recoil or many other phenomena associated with elasticity. The model is mostly of historical importance and is currently used only heuristically to model simple steady shearing flows. For such flows, the only nonzero scalar invariant of $\mathbf{D}$ is the second invariant, proportional to the square of the shear rate, κ^2, while $\mathbf{D}^2$ is equal to zero. The function η is a generalized, nonconstant viscosity, a monotonically increasing function of the shear rate. A common example of this form is the *power-law fluid,* in which η is proportional to a power of the shear rate, viz.,

$$\boldsymbol{\sigma} - p\mathbf{1} = C\kappa^{n-1}\mathbf{D}, \quad \kappa^2 = -I_2(\mathbf{D}), \tag{44}$$

where C is a material constant of suitable physical dimensions (depending on the dimensionless material constant n) and κ is the rate of shear with the dimension of reciprocal time. When n is set equal to 1, the equation becomes that of a Newtonian fluid, Eq. (14), with C set equal to η_0. The power-law fluid is useful because it can often be used to find exact solutions to geometrically simple problems and, for limited ranges of shear rate, it may fit data reasonably well. Usually, however, it is a poor representation of most rheologically complex fluids over wide ranges of shear rate (with the possible exception of slurries in simple, steady shearing flows).

To construct differential models of nonlinear viscoelasticity, it is necessary to define rates of change that are materially objective. The *Rivlin–Ericksen tensors* $\mathbf{A}_i(t)$, $i = 1, 2, \ldots$, are material time derivatives of the right relative strain, thus:

$$\mathbf{A}_n(t) \equiv \left.\frac{d^n\mathbf{C}_t(\tau)}{d\tau^n}\right|_{\tau=t}. \tag{45}$$

The higher derivatives can most easily be found with a recurrence relation for $\mathbf{A}_{n+1}$ in terms of $\mathbf{A}_n$, as follows:

$$\begin{aligned} \mathbf{A}_1(t) &= 2\mathbf{D}(t), \\ \mathbf{A}_{n+1}(t) &= \frac{d\mathbf{A}_n(t)}{dt} + \mathbf{v}\cdot\nabla\mathbf{A}_n(t) + \mathbf{A}_n(t)\mathbf{L}(t) \\ &\quad + \mathbf{L}^{\mathrm{T}}(t)\mathbf{A}_n(t), \quad n = 1, 2, 3, \ldots, \end{aligned} \tag{46}$$

where $\mathbf{L}$ is the velocity gradient. [See Eq. (12).] The nth-order Rivlin–Ericksen tensor has the physical dimensions of the nth power of reciprocal time.

The *nth-order Rivlin–Ericksen fluid* has the constitutive equation

$$\boldsymbol{\sigma} - p\mathbf{1} = \mathbf{f}(\mathbf{A}_1,\mathbf{A}_2,\ldots,\mathbf{A}_n), \tag{47}$$

where $\mathbf{f}(\cdot)$ is an isotropic tensor function of its arguments. There are representation theorems that may be used to develop $\mathbf{f}$ as a polynomial. For example, the stress of the second-order Rivlin–Ericksen fluid is an arbitrary isotropic tensor function of the first two Rivlin–Ericksen tensors that may be written out as the sum of terms in products of powers of the $\mathbf{A}_i$ with coefficients whose physical dimensions are those of stress times a power of time. Thus,

$$\begin{aligned}
&\boldsymbol{\sigma} = p\mathbf{1} + \boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_2 + \boldsymbol{\sigma}_3 + \boldsymbol{\sigma}_4 + \boldsymbol{\sigma}_5 + \boldsymbol{\sigma}_6;\\
&\boldsymbol{\sigma}_1 = \alpha_1\mathbf{A}_1, \quad \boldsymbol{\sigma}_2 = \alpha_2\mathbf{A}_2 + \alpha_3\mathbf{A}_1^2,\\
&\boldsymbol{\sigma}_3 = \alpha_4(\mathbf{A}_1\mathbf{A}_2 + \mathbf{A}_2\mathbf{A}_1),\\
&\sigma_4 = \alpha_5\mathbf{A}_2^2 + \alpha_6(\mathbf{A}_1^2\mathbf{A}_2 + \mathbf{A}_2\mathbf{A}_1^2),\\
&\boldsymbol{\sigma}_5 = \alpha_7(\mathbf{A}_1\mathbf{A}_2^2 + \mathbf{A}_2^2\mathbf{A}_1),\\
&\boldsymbol{\sigma}_6 = \alpha_8(\mathbf{A}_1^2\mathbf{A}_2^2 + \mathbf{A}_2^2\mathbf{A}_1^2),
\end{aligned} \tag{48}$$

where the coefficients α_i are functions of the ten scalar invariants

$$\begin{aligned}
&\mathrm{tr}(\mathbf{A}_i),\ \mathrm{tr}(\mathbf{A}_i^2),\ \mathrm{tr}(\mathbf{A}_i^3), \quad i = 1, 2,\\
&\mathrm{tr}(\mathbf{A}_1\mathbf{A}_2),\ \mathrm{tr}(\mathbf{A}_1^2\mathbf{A}_2),\ \mathrm{tr}(\mathbf{A}_1\mathbf{A}_2^2),\ \mathrm{tr}(\mathbf{A}_1^2\mathbf{A}_2^2).
\end{aligned} \tag{49}$$

The representation of isotropic functions of sets of tensors is a highly technical subject, important to rheology and well reviewed in an article by Spencer (1971). The second-order Rivlin–Ericksen fluid in Eq. (48) is based on the representation of a general isotropic tensor function of only the first two $\mathbf{A}_i$, but it is already far too complicated to be of much use in applications.

A considerable simplification is achieved by restricting the stress to a power series in the $\mathbf{A}_i$ with constant coefficients. The resulting equations containing terms with coefficients of dimensions up to nth order in time are called simply *nth order fluids.* Thus, the second-order fluid has the constitutive equation

$$\boldsymbol{\sigma} = p\mathbf{1} + \eta_0\mathbf{A}_1 + \alpha_3\mathbf{A}_1^2 + \alpha_2\mathbf{A}_2, \tag{50}$$

where η_0 is the Newtonian viscosity and the α_i are constants. This is perhaps the simplest fluid that can be used to model the Weissenberg effect. Furthermore, an unexpected and useful feature of second-order fluids is that for any plane boundary-value problem determining a flow by boundary conditions on the velocity (rather than surface tractions), the flow appropriate for a Newtonian fluid is also a possible flow for a second-order fluid of the same viscosity: Only the stresses are different. This remarkable fact is known as the *Tanner–Giesekus theorem,* and it can be used to find inverse solutions to many simple plane flow problems by direct substitution of well-known Newtonian flows. The theorem is easily extended to parallel rectilinear flows and to potential flows such as occur in boundary layers. These solutions give insight into the mechanisms for many nonlinear phenomena but, unfortunately, the second-order fluid has some serious failings. It does not explain circulations in flow through noncircular pipes, for instance, and is certainly inadequate to describe stress relaxation.

The generalization of a linear differential model, valid for small deformations, to a nonlinear model valid for large deformations is not unique. Furthermore, the material time derivatives are ambiguous for finite deformations. Roughly speaking, this problem arises because the base vectors of the coordinate system convect differently for covariant and contravariant systems. In technical terms, raising and lowering indices does not generally commute with material time differentiation. As a result, the rate of a covariant tensor is not the same as the covariant form of the rate of the contravariant form of the tensor. Yet, either rate is objective. Indeed, so is any linear combination of them. Consequently, there is an infinity of materially objective time derivatives of constitutive tensors.

Three particular rates are so commonly used that they have been given names. If $\mathbf{S}$ is any generic objective tensor, $\mathbf{L}$ is the velocity gradient defined in Eq. (12), and $\mathbf{W}$ is the spin tensor defined in Eq. (13), then the three convected derivatives or rates of change of $\mathbf{S}$ are defined and named as follows:

$$\begin{aligned}
\frac{D^{\nabla}\mathbf{S}}{Dt} &\equiv \frac{\partial\mathbf{S}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{S} - \mathbf{LS} - \mathbf{SL}^{\mathrm{T}}\ \text{(Upper)},\\
\frac{D^{\Delta}\mathbf{S}}{Dt} &\equiv \frac{\partial\mathbf{S}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{S} + \mathbf{SL} + \mathbf{L}^{\mathrm{T}}\mathbf{S}\ \text{(Lower)},\\
\frac{D^{0}\mathbf{S}}{Dt} &\equiv \frac{\partial\mathbf{S}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{S} + \mathbf{SW} - \mathbf{WS}\ \text{(Jaumann)}.
\end{aligned} \tag{51}$$

The upper-convected derivative is the rate of a contravariant tensor as seen by an observer in a body-fixed frame of covariant base vectors but expressed in a space-fixed frame. Similarly, the lower-convected derivative relates to a covariant tensor. The Jaumann or *corotational derivative* is the average of these two, and it is the rate seen by an observer whose frame rotates with the material but does not stretch.

The *Oldroyd B-fluid* is a nonlinear form of the Jeffreys model. It is obtained by replacing time derivatives in Eq. (20) with upper-convected derivatives to get

$$\mathbf{s} + \lambda_{\text{rlx}} \frac{D^{\nabla}\mathbf{s}}{Dt} = 2\eta\left[\mathbf{D} + \lambda_{\text{rtd}} \frac{D^{\nabla}\mathbf{D}}{Dt}\right]. \tag{52}$$

This is a fairly simple, properly objective equation for a model that exhibits the Weissenberg effect. If the retardation time λ_{rtd} is set equal to zero, the equation becomes that of the *upper-convected Maxwell model.* This Maxwell model is one of the simplest properly objective viscoelastic equations, and it is widely used for calculating complex flows. It is known that the model is a poor representation of materials—for instance, it cannot describe shear-dependent viscosity—but, in many situations, it often shows qualitatively correct behavior.

Differential models are often favored when relating constitutive equations to micromechanics. For example, picturing a polymer solution as a dilute solution of "dumbbells," pairs of spherical particles joined by a Hookean spring and immersed in a Newtonian fluid, leads to the *dumbbell model.* This model is phenomenologically equivalent to the upper-convected Maxwell model with the addition that the material constants η, λ_{rlx}, and λ_{rtd} are expressed in terms of parameters of the micromechanics. An elaboration of this model based on the micromechanics of a less dilute solution is the *Giesekus model* for which a concentration dependence is added to the dumbbell model through the following equation:

$$\lambda \frac{D^{\nabla}\mathbf{s}}{Dt} + \mathbf{s} + \frac{\alpha\lambda}{\eta}\mathbf{s}^2 = 2\eta\mathbf{D}, \quad 0 \le \alpha \le 1, \tag{53}$$

where α is a measure of the anisotropy induced by the flow due to the mean orientation of the dumbbells.

The *Leonov model* is based upon the idea that strain and strain rate may each be decomposed into elastic and irrecoverable parts. The irrecoverable strain rate is determined as a function of the elastic strain through a constitutive equation similar to Eq. (35) for stress. Then the elastic strain is determined through

$$\frac{D^{\nabla}\mathbf{C}_e}{Dt} + \mathbf{D}_i\mathbf{C}_e + \mathbf{C}_e\mathbf{D}_i = 0, \tag{54}$$

where $\mathbf{C}_e$ and $\mathbf{D}_i$ are the recoverable strain and the irrecoverable strain rate, respectively. The stress is found through another equation on the elastic strain of the form of Eq. (35). The Leonov model thus requires two constitutive functions similar to strain-energy functions, the elastic strain energy and the *nonequilibrium potential* for determining the irrecoverable strain rate. As a result, there is the possibility of describing very complicated behavior. The simplest version proposed, however, contains only two material constants, a modulus and a relaxation time, and, for plane flows, it is equivalent to the Giesekus equation.

1.4.3 Models with Plastic Yield

Many materials, such as metals or drilling muds, appear not to flow until the stress exceeds some minimum value. The *Bingham body* is a classical model of shearing in such materials. For this model, the constitutive equation takes different forms depending on whether or not the shear stress exceeds the *yield stress* σ_y, a constitutive constant. For simple shear, the Bingham body is defined by the following:

$$\begin{aligned} &\sigma_s - \sigma_y = \eta_p\kappa, \quad |\sigma_s| \ge \sigma_y; \\ &\kappa = 0, \quad |\sigma_s| \le \sigma_y, \end{aligned} \tag{55}$$

where σ_s is the shear component of stress, σ_y is the yield stress, η_p is the plastic viscosity, a constitutive parameter, and κ is the rate of shear. A tensor form of Eq. (55) could easily be written, but that hardly seems worth while since the model is seldom used for analyzing any deformation other than simple shearing.

The Bingham body is simply a Newtonian fluid combined through the yield condition with a rigid body. Other such constitutive equations can be easily formed by combining solidlike equations (e.g., elasticity) valid for stresses below a yield stress with a fluid differential model (e.g., a power-law fluid) above the yield. A survey of such models is to be found in Bird *et al.* (1987). *Plasticity theories* also utilize this idea of a yield criterion distinguishing loading from unloading of stresses, but further supplement this with flow rules. The classical theory of plasticity as expounded in Hill (1950) has seen many subsequent elaborations and improvements in later years. These theories are particularly useful in describing metal-forming processes.

1.4.4 Integral Models In the case of linear viscoelasticity, for smooth histories at least, each differential model corresponds to an integral model. Which of these representations to use depends on the deformation history to be described. Discontinuous histories of stress or strain are usually more easily described by integral forms while sinusoidal histories are best treated with differential forms. For nonlinear viscoelasticity, this choice is seldom available and only the simplest nonlinear models may retain this convenience.

An early model useful for unvulcanized rubbers and molten polymers was based upon the micromechanics of a certain transient network with junctions that break and reform, the *Green–Tobolsky model,* with an integral constitutive equation:

$$\boldsymbol{\sigma}(t) = p\mathbf{1} + G\int_{-\infty}^{t} e^{(\tau-t)/\lambda}\mathbf{B}_t(\tau)d\tau. \tag{56}$$

This equation is equivalent to the upper-convected Maxwell model, Eq. (52), with λ_{rtd} set to zero and η equal to $G\lambda_{\text{rlx}}$.

The *Lodge model* is based upon less restricted transient network micromechanics resulting in a somewhat more general integral form:

$$\begin{aligned} \boldsymbol{\sigma}(t) &= p\mathbf{1} + \int_{-\infty}^{t} \mu(t-\tau)\mathbf{B}_t(\tau)d\tau, \\ \mu(\xi) &= \sum_i \frac{G_i}{\lambda_i} e^{-\xi/\lambda_i}. \end{aligned} \tag{57}$$

Here, the function $\mu(\cdot)$ is a sum of relaxation processes each with a relaxation time λ_i and a modulus G_i. These quantities are to be evaluated from fitting stress-relaxation curves or, possibly, are derived from more detailed micromechanical considerations.

Because stress relaxation and other mechanical behaviors typical of viscoelastic materials are clearly not equilibrium phenomena, classical thermodynamics (which is a thermodynamics of equilibrium states) is not sufficient to describe viscoelasticity. The *perfect elastic fluid* is a thermodynamically plausible, phenomenological model of viscoelasticity encompassing entropic as well as energetic effects. The model depends upon the assumption that adding a quantity to entropy can form a variable to be used in place of entropy in a caloric equation of state that will apply to these nonequilibrium viscoelastic phenomena. The caloric equation of state of classical thermodynamics is generalized by replacing entropy with a state function for nonequilibrium states that is the sum of entropy and a quantity called *entaxy.* Entaxy is a functional of the deformation history and the thermal history of the material, in contrast to the entropy, which is simply a function of the deformation and temperature at equilibrium (Bernstein *et al.,* 1964). Thus,

$$\begin{aligned} \Sigma(t) &= \int_{-\infty}^{t} S\left(C_\tau(t), \int_t^\tau b_T(\xi)d\xi\right)b_T(\tau)d\tau, \\ \epsilon &= f(v, s+\Sigma), \quad T = \frac{\partial f(v, s+\Sigma)}{\partial s}, \end{aligned} \tag{58}$$

where $\Sigma(t)$ is the specific entaxy at time t, s is specific entropy, and v is the specific volume (the inverse of density, ρ). The function $S(\mathbf{C},\Delta t)$ is the rate of entaxy production, a function of the deformation from a past configuration and of the time interval as seen by the material. That time interval is dependent on b_T, the rate of a material clock, a positive function of temperature that is the inverse of a_T, the shift factor of time-temperature superposition (see Sec. 2.3, Moduli of Linear Viscoelasticity). The generalized caloric equation of state gives the specific internal energy ϵ as a function of specific volume v and $s + \Sigma$, the sum of entropy and entaxy. The absolute temperature T is the derivative of internal energy with respect to entropy, just as it is in equilibrium thermodynamics. The stress is given by a hydrostatic pressure, the derivative of internal energy with respect to specific volume, plus terms (not isotropic in general) that vanish at equilibrium. When a state is held until equilibrium is achieved, Σ goes to zero and classical thermodynamics is restored.

The stress equation for a perfect elastic fluid is written

$$\boldsymbol{\sigma}(t) = p(v,T)\mathbf{1} + \rho T \int_{-\infty}^{t} \mathbf{F}_t(\tau) \times \frac{\partial S\left(\mathbf{C}_\tau(t), \int_\tau^t b_T(\xi)d\xi\right)}{\partial \mathbf{C}_\tau(t)} \mathbf{F}_t^{\mathrm{T}}(\tau) b_T(\tau) d\tau. \tag{59}$$

Entaxy production is assumed to have properties such that, while a material is held fixed and isothermal after deformation until thermodynamic equilibrium is established, the integral of Eq. (59) will monotonically relax, eventually to zero—the stress will approach a purely isotropic pressure. Thus, this is a model of a fluid. On the other hand, the integrand of Eq. (59) has similarities to the right-hand side of Eq. (36) for finite elasticity so that the model shows certain elastic properties when the material is not in equilibrium. Depending on the time scale of observation [see Deborah number, Eq. (80)], this model will display either elastic or viscous behavior. The material properties of a perfect elastic fluid are contained in the function $S(\cdot,\cdot)$, the function b_T, and the equilibrium pressure function, $p(v,T)$.

The perfect elastic fluid model is generally used to describe isothermal deformations of incompressible materials for which the thermodynamic details are largely irrelevant. In that case, Eq. (59) reduces to

$$\begin{aligned}\sigma(t) &= p\mathbf{1} + \int_{-\infty}^{t} \left[\frac{\partial U(I_1, I_2, t-\tau)}{\partial I_1} \mathbf{B}_t(\tau) - \frac{\partial U(I_1, I_2, t-\tau)}{\partial I_2} \mathbf{B}_t^{-1}(\tau)\right] d\tau, \\ I_i &= I_i(\mathbf{B}_t(\tau)).\end{aligned} \tag{60}$$

This equation is called the *BKZ model* (for Bernstein, Kearsley, and Zapas; see Bernstein *et al.*, 1963, 1965) or sometimes the *Kaye-BKZ model.* The constitutive function $U(I_1, I_2, t - \tau)$ has the character of a time-dependent, weighted, strain-energy function of finite elasticity, and the integral on the right may be viewed as a weighted sum of the partial stresses arising from deformations from configurations at all past times τ to that at the present time, t. As a result, the material properties of the BKZ model are carried by the function U, which is unspecified. It must be measured or, possibly, derived from micromechanical considerations. A useful reference for applying this formulation to common deformations is found in Bernstein (1966).

The BKZ model with special forms of the function U arises often for simple micromechanically motivated models. The Green–Tobolsky and the Lodge models, Eqs. (56) and (57), are easily recognized as particular cases of Eq. (60). The *Doi-Edwards model* (1971) is based upon the idea of *reptation* proposed by de Gennes (1971) for the micromechanics of polymer melts. The chain molecules are visualized as confined by the surroundings to move within tubelike regions while the tubes themselves slowly change shape by diffusion. This rather elaborate picture turns out, at least in its simplest form, to generate a BKZ model.

Other integral models have been proposed, some of which may be viewed as generalizations of the BKZ model. For example, the derivatives of U in Eq. (60) may be replaced by arbitrary functions of the same variables, and these functions may be taken as factorable into a deformation-dependent part and time-interval–dependent part. Or, scalar invariants of the rate of deformation may be included as variables. Wagner and Stephenson (1979) propose a model in which an irreversibility is introduced by distinguishing "increasing" from "decreasing" deformation much as in plasticity theory. The form of the constitutive equation is modified according to this distinction. The resulting constitutive equation is often called the *Wagner model.*

1.4.5 Nonsimple Materials

The constitutive equations above are all simple materials in the sense that stress at a point depends only on the history of deformation at that point. Such constitutive equations may be inadequate to describe complex materials for several reasons. The stress at a point may depend on nonlocal effects as might be expected for suspensions of stiff fibers. Or, there may be microstructure introducing internal degrees of freedom at a point, say the relative rotation of particles in a granular material or dislocation density in a lattice structure. Couple stresses may be required if electromagnetic interactions are to be included. The *Cosserat continuum* (Cosserat and Cosserat, 1909) is a model material for which each point of material, in addition to its position, carries vectors called *directors* that reflect micromechanical influences. A director might express some local mean orientation of particles in a suspension, for instance. If these directors affect the energetics and rheology of the material, the balance equations, Eqs. (9)–(11), must be modified accordingly. The Cosserat continuum and other models of this type are often referred to as *micropolar models.* Expositions of theories of this type are found in Green and Rivlin (1964), Toupin (1962), and Eringen (1962).

An important example of such nonsimple constitutive equations is the *Ericksen–Leslie model* of liquid crystals (Ericksen, 1961; Leslie, 1968). In this case, the director relates to the orientation of rodlike molecules. A constitutive equation for the asymmetric stress tensor is coupled with suitably modified balance equations and boundary conditions.

2. RHEOMETRY

To confirm the validity of constitutive equations as well as to realize their practical benefits, it is necessary to devise methods of observing and measuring the mechanical properties of target materials. Measurements of rheological properties of a material can only be made in the context of some assumptions about the form of a suitable constitutive equation. Ultimately, workable methods must be devised to produce well-defined deformations and to measure the forces required. Almost always, ideal measurements can only be approached in approximation. For example, infinite geometry may be modeled with finite apparatus by ignoring various end effects. Furthermore, the practicality of a given measurement technique is usually limited to a range of the properties of the test material. As a result, a great variety of apparatus have been devised, often for measurement of the same material property, but useful for different materials and circumstances. A practicing industrial processor of materials would like one simple test device that gives a number measuring "processability." Most technological processes are so complex, however, that no single measuring device can possibly be sufficient.

In the following discussion of rheometry, we shall often assume that inertial effects, surface-tension effects, and other nonrheological effects can be ignored, and we will say little about them. Nonetheless, sometimes these effects cannot be ignored. For instance, it is clear from Eq. (10) that the inertial term $\rho(\mathbf{v}\cdot\nabla)\mathbf{v}$ could have a big influence on high-velocity flows. In some situations, this term can produce complicating nonlinearities not directly related to the material rheology and then it must be taken into account. Fortunately, the range of deformation rate of interest for many materials and processes is such that rheological measurements at low–Reynolds-number flows are adequate and inertial effects can usually be ignored.

2.1 Viscometry

The rheology of Newtonian fluids conforming to Eq. (14) is completely determined by measuring the value of η_0, the constant viscosity. A variety of instruments have been devised for achieving this. The commonest of these is the *capillary viscometer,* for which the volume rate of laminar flow through the capillary under the influence of a known pressure difference is related to the viscosity through the *Hagen–Poiseuille law:*

$$\eta_0 = \pi R^4 \Delta P/8QL, \tag{61}$$

where R is the radius of the capillary, ΔP is the difference in pressure across the capillary of length L, and Q is the volume flow per unit time. In practice, corrections to this equation are often needed, typically, a kinetic energy correction to account for the power expended in overcoming inertia and the Couette correction to account for end effects of the finite capillary through an addition to L of a fictitious corrective length. In gravity-driven capillary viscometers and in some other devices, it is actually the *kinematic viscosity,* η_0/ρ, which is measured.

A standard reference for the measurement of Newtonian viscosity by this and other classical methods is Barr (1931).

Nonlinear fluids that are modeled by a modification allowing viscosity to vary with shear rate in Eq. (14) are said to be non-Newtonian. If viscosity decreases with rate of shear, the material is said to be *pseudoplastic,* and if it increases, it is said to be *dilatant.* For some materials, such as certain suspensions and colloids, there is an internal structure when they are quiescent that is slowly broken up by shearing but, after a recovery period at rest, the fluid returns to its structured state. Such materials show a decrease in viscosity with time of shearing after rest, and they are called *thixotropic*. In rarer cases, the formation of a temporary structure is promoted by shearing and viscosity increases with time of shear. Such a material is said to be *rheopectic*.

In characterizing a non-Newtonian material, it is necessary to assign a shear rate to each measurement of viscosity. Furthermore, the Hagen–Poiseuille law is no longer valid since it is derived from the velocity profile of the flow, which differs here from the Newtonian case. Capillary viscometers can still be used, however, in the following way (based upon an equation for the gradient of the velocity). The volume flow rate Q is measured as a function of the pressure gradient, $a = \Delta P/L$. By calculation or graphical methods, the quantity κ, a rate of shear, is found from the following equation:

$$\kappa(a) \equiv \frac{1}{\pi a^2 R^3} \frac{d}{da} [a^3 Q(a)], \quad a \equiv \frac{\Delta P}{L}. \tag{62}$$

Finally, the viscosity as a function of shear rate is given by

$$\eta(\kappa) = R\Delta P/2L\kappa. \tag{63}$$

This procedure, viewed as a correction of the Hagen–Poiseuille law, is usually referred to as the *Rabinowitsch correction*. As a check on the validity of the method, measurements from capillaries of differing radii are compared. If they do not fall upon a single viscosity–rate-of-shear curve, problems with slip at the wall or other anomalies are to be suspected.

As the rate of shear is increased, even for Newtonian fluids, eventually it is found that laminar flow cannot be maintained, turbulence occurs, and the viscometer equations are no longer valid. But for many nonlinear liquids, other difficulties may interfere well before arriving at that point. The end corrections can be quite large for elastic fluids when considerable power is consumed in generating elastic energy. In that event, the Couette correction is inadequate and a procedure developed by Bagley (1957) may be used as an approximate correction for this entrance effect. As the fluid emerges from the tube, a swelling of the jet known as *die swell* (see above, Sec. 1.4) will usually occur and sometimes, at higher flow rates, an irregular unsteady extrusion known as *melt fracture*. These phenomena are presently only partially understood even though they are of great practical importance.

A *Couette viscometer* uses the laminar shearing flow between two coaxial cylinders to measure viscosity. In the most common, convenient form, the inner cylinder is rotated and the outer held fixed, but, at high Reynolds numbers, this arrangement has the drawback that Taylor vortices due to inertial forces may arise and disturb the flow. An instrument that overcomes this complication by exchanging the roles of the cylinders, thus stabilizing the flow, is sometimes called a *Searle viscometer*. In both cases, the following equation called the *Margules equation* is used to measure Newtonian viscosity:

$$\eta_0 = \frac{M}{4\pi L \Delta\Omega} \left(\frac{1}{R_i^2} - \frac{1}{R_o^2} \right), \tag{64}$$

where M is the torque acting on a cylinder, L is the length of the cylinders, $\Delta\Omega$ is the difference in angular velocity between the cylinders, and R_i and R_o are the radii of inner and outer cylinders, respectively. The Margules equation is based upon flow between infinite cylinders and, for very accurate measurements, end-effect corrections and other adjustments may be needed.

Couette viscometers may also be used to measure non-Newtonian viscosity by using a procedure analogous to the Rabinowitsch correction for capillaries, but it is somewhat more difficult to apply. Yang and Krieger (1978) consider various approximate methods of associating the measurement of viscosity with shear rate that are more practical and are often used.

To avoid the complication resulting from inhomogeneous shear rates for non-Newtonian fluids, it is convenient to measure viscosity from shearing flows between a rotating cone and a fixed plate. If the cone axis is perpendicular to the plate and the apex is just at the plate, the resulting flow is taken to be a homogeneous shearing in which fluid particles travel in circles about the cone axis. Actually, for nonlinear materials, this is an approximation true only for geometries with a small angle between the cone surface and the plane because, otherwise, normal stresses as well as inertial effects create transverse flows. For this reason, it is customary to restrict this angle to be less than 4°. With a *cone-and-plate viscometer,* for both Newtonian and non-Newtonian fluids, the viscosity and the rate of shear are found from simple equations, as follows:

$$\eta(\kappa) = \frac{3M\theta}{2\pi R^3 \Omega}, \quad \kappa = \frac{\Omega}{\theta}, \tag{65}$$

where κ is the rate of shear, M is the torque, θ is the angle between the surface of the cone and the plate, R is the radius, and Ω is the angular velocity between cone and plate.

There are many other viscometers designed to fit various practical or special technological needs. A variety of mechanisms can be used: timing the fall under gravity of a heavy ball immersed in the sample, measuring the damping of a vibrating reed or timing a disk in torsional oscillation in the test fluid, etc. A fairly complete listing of such devices as well as the classical viscometers can be found in Merrington (1949). These methods using complex flows can be used for relative viscometry of Newtonian liquids, where the linearity of the model allows comparisons of viscosity, but they are seldom practical for measuring material properties of non-Newtonian materials.

2.2 Strain-Energy Functions

The mechanical properties of a purely elastic material are contained in the strain-energy function: For an incompressible isotropic material, this is a function of two invariants of the deformation. The strain-energy function for a particular incompressible material is determined from measurements of force and displacement. [It can be seen from Eq. (38) that, to be precise, it is only the two derivatives of the strain energy with respect to the strain invariants that are experimentally accessible.] The ranges of the invariants that are physically possible are restricted by the requirement that the stretches be real positive quantities. Figure 3 illustrates the region of the $I_1(B)$–$I_2(B)$ plane containing physically meaningful sets of these two invariants. The region is enclosed by a cusp formed between two curves, loci of the values for simple extension and equibiaxial extension, respectively. The 45° line in Fig. 3 corresponds to values for pure shear. Strain energy is determined (up to an arbitrary constant of no physical significance—usually it is set to give zero strain energy in the undeformed state) by mapping the derivatives over this region. This is a laborious and demanding task at best. For some typical results, see Obata *et al.* (1970), who have measured the strain-energy derivatives of some typical rubbers as a function of degree of cross-linking and of temperature.

The most direct method of measuring the derivatives of the strain energy is through homogeneous, biaxial stretching of a sheet of material. Rivlin and Saunders (1951), in a classic description of various measurements of strain energy, describe such biaxial measurements in extensive detail. The method is useful, but it is tedious and, for practical reasons, it is not accurate for deformations close to the boundaries of the region of possible invariants shown in Fig. 3.

There are a number of inhomogeneous deformations that may be used to measure the strain-energy function of an incompressible material. Ericksen (1954) finds several classes of suitable deformations (sometimes called *controllable deformations*) for which the derivatives of the strain energy may be expressed in terms of the surface tractions needed to produce the deformations; that is, it is possible to produce them solely through the application of surface tractions no matter what the form of the strain energy. Torsion of a cylinder is an example. In this case, a distribution of normal force (because of the Poynting effect: see Sec. 1.4), in addition to a torque, must be applied to the ends of the cylinder to maintain the deformation. Of course, the actual amounts of traction required will depend upon the strain energy but, no matter what form the strain-energy function, some suitable torque and normal force will suffice. Penn and Kearsley (1976) demonstrate how measurements of torsion of an elastic cylinder can be used to evaluate the derivatives of the strain energy along the pure shear line of Fig. 3.

When the strain energy is of the Valanis–Landel type, the measurement task is much reduced. Then, the measurement of torque and normal force in torsion of a cylinder, or

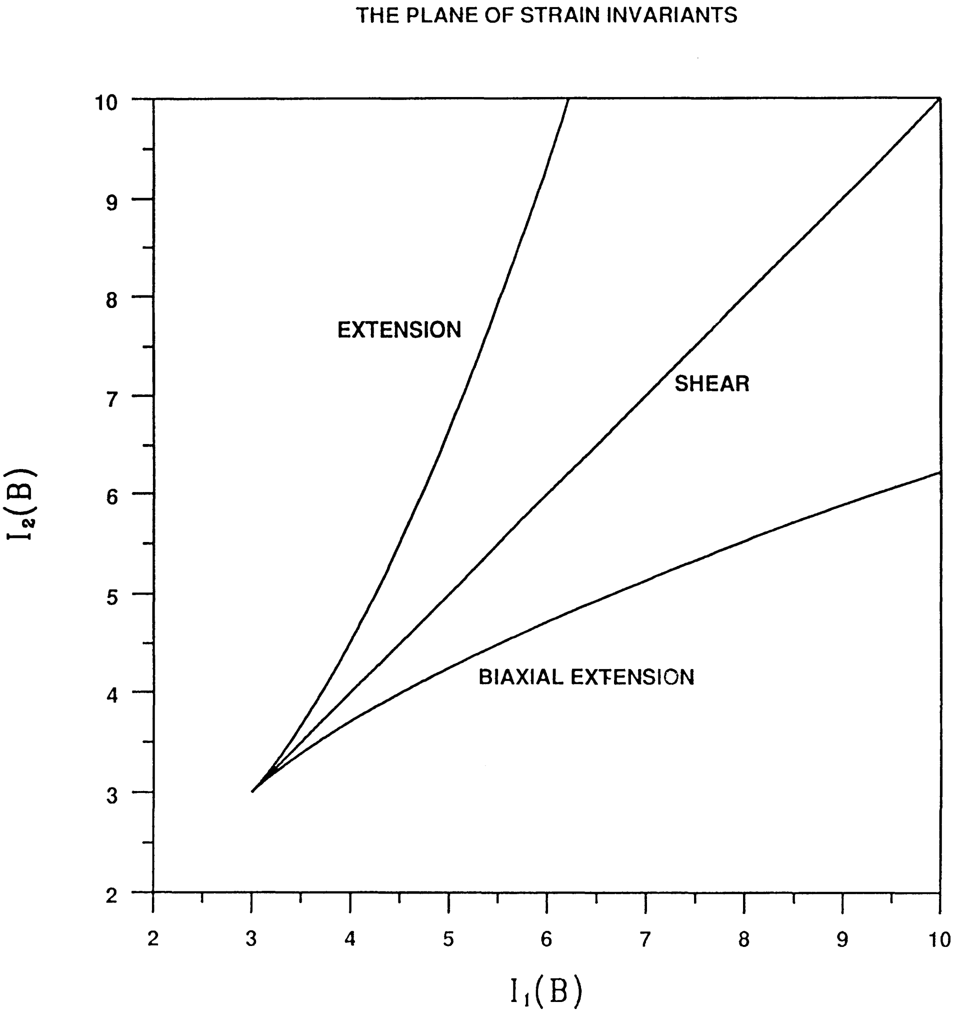

FIG. 3. Strain invariants of an incompressible material. Values of the invariants are confined to the cusped area between the lines for extension and biaxial extension because of the requirement that stretches be real and positive. The 45° line for which $I_1 = I_2$ is the locus of shear deformations.

the measurement of tractions for simple extension and equibiaxial extension, or any combination of two of these measurements is sufficient to determine the stress in any deformation (Kearsley, 1980; Kearsley and Zapas, 1980). These methods can be extremely useful because, in fact, most molecular models of rubber elasticity do predict a strain-energy function of Valanis–Landel form.

2.3 Moduli of Linear Viscoelasticity

Each of Eqs. (26), (27), and (29) formulates the mechanical properties of a linear-viscoelastic material. They indicate that the material may be characterized by measurements of stress relaxation or creep or sinusoidal deformation. For practical reasons, all three methods may be needed to cover a broad time scale of responses. For very short-time behavior, measurements of wave propagation through the material may be used. Once either $G(t)$ or $J(t)$ or $G^*(i\omega)$ is known, it is possible to calculate the stress corresponding to an arbitrary strain history, or the strain corresponding to an arbitrary stress history, by use of Boltzmann superposition principle.

For many polymeric materials, a plot of $\log J(t)$ vs $\log t$ measured at some fixed temperature T_0 can be superposed on the plot measured at another fixed temperature by translation along the log$t$ axis combined with a smaller translation along the log$J$ axis. The translation along the log$t$ axis is usually designated loga_T, where a_T is the reciprocal of the temperature-dependent rate of a material clock [see b_T of Eq. (59)]. Materials that fit this picture of a temperature-controlled, material clock are called *thermorheologically simple.* This curve-shifting procedure, known as *time-temperature superposition,* can then be used to draw a master curve of $J(t,T_0)$ at a reference temperature using superposed data from measurements at various temperatures. The shift factor a_T and the vertical shift are also plotted. In this way, extremely wide time scales can be plotted—so wide that it would be completely impractical to measure them at a single temperature. The resulting curve can be transposed to other temperatures by appropriate shifts taken from the plots.

2.4 Normal Stresses

In a certain class of flows, the strain history seen by a particle of material is independent of time in the sense that the strain history is independent of the chosen reference time, except for a possible rotation of the reference frame. The importance of this class is that for each such flow the stress for a general simple material modeled by a complicated functional as in Eq. (8) can be reduced to a set of functions of the first three Rivlin–Ericksen tensors. Such flows are called *motions with constant stretch history* (abbreviated MWCSH) because a material point experiences a deformation rate constant in time. They are often used in rheometry because of the simplicity of dealing with functions rather than functionals. A detailed analysis of the kinematics of MWCSH can be found in Huilgol (1975).

By far the commonest examples of MWCSH are steady laminar shear flows such as those occurring in viscometry. These flows constitute an important class called *viscometric flows.* A concise but readable treatment of the subject is found in Coleman *et al.* (1966). Viscometric flows are MWCSH with the added restriction that $\mathbf{L}^2 = (\nabla \mathbf{v})^2 = 0$. As a consequence, all Rivlin–Ericksen tensors of order greater than three are zero for these flows and, at each point of a flow, by rotation, a Cartesian coordinate system can be found in which the Rivlin–Ericksen tensors are as follows:

$$\mathbf{A}_1 = \begin{Vmatrix} 0 & \kappa & 0 \\ \kappa & 0 & 0 \\ 0 & 0 & 0 \end{Vmatrix}, \quad \mathbf{A}_2 = \begin{Vmatrix} 0 & 0 & 0 \\ 0 & 2\kappa^2 & 0 \\ 0 & 0 & 0 \end{Vmatrix},$$
$$\mathbf{A}_i = 0 \quad \text{when } i \geq 3. \tag{66}$$

The quantity κ and the rotation necessary to achieve this form will usually depend on position in the flow.

When Eq. (66) holds for all points of the flow in a fixed Cartesian coordinate system and κ is constant throughout the flow, the flow is a plane shearing flow with shear rate κ oriented so that the flow velocity is in direction 1, the shear gradient is in direction 2, and direction 3 is perpendicular to the other two. Locally, any viscometric flow of a simple material is equivalent to a steady plane shearing flow: Measurements of it can reveal no more rheological information than does steady plane shearing flow.

In viscometric flows, simple materials described by the functional of Eq. (8) are indistinguishable from the conceptually simpler Rivlin–Ericksen fluids described by the functions of Eq. (47) which, in light of Eq. (66), may be further simplified. Thus, for general simple materials in steady plane shearing flow a constitutive equation can be written as follows:

$$\boldsymbol{\sigma} = p\mathbf{1} + \eta(\kappa)\mathbf{A}_1 + [\psi_1(\kappa^2) + \psi_2(\kappa^2)]\mathbf{A}_1^2 - \tfrac{1}{2}\,\psi_1(\kappa^2) \tag{67}$$

where the $\mathbf{A}_\mathrm{i}$ are in the form of Eq. (66). The rheological material properties determining material response in all viscometric flows are contained in three functions of shear rate: η is the viscosity, while ψ_1 and ψ_2 are called the *first normal-stress coefficient* and the *second normal-stress coefficient,* respectively, for reasons that will now be made clear.

From Eqs. (66) and (67), the components of stress are given by

$$\begin{aligned} \sigma_{12} &= \sigma_{21} = \kappa\eta(\kappa), \\ \sigma_{11} &= \kappa^2[\psi_1(\kappa^2) + \psi_2(\kappa^2)] + p, \\ \sigma_{22} &= \kappa^2\psi_2(\kappa^2) + p, \quad \sigma_{33} = p, \end{aligned} \tag{68}$$

all other components are zero.

Here, p is the arbitrary pressure associated with incompressibility and $\eta(\kappa)$ is the viscosity generating the shear stress. The only other nonzero components of stress are the diagonal components of the stress tensor, the *normal stresses.* Because p is arbitrary, the normal stresses themselves are not uniquely determined by shear rate, and it is customary to avoid the resulting ambiguity by considering differences of normal stresses. By convention, two fundamental normal-stress differences are defined, the *first and second normal-stress functions,* N_1 and N_2, each proportional to a corresponding normal-stress coefficient, ψ_1 and ψ_2, respectively, related as follows:

$$N_1 \equiv \sigma_{11} - \sigma_{22} = \kappa^2\psi_1(\kappa^2),$$
$$N_2 \equiv \sigma_{22} - \sigma_{33} = \kappa^2\psi_2(\kappa^2). \quad (69)$$

The rheology of a simple material in a viscometric flow is completely characterized by the viscosity and the first and second normal-stress coefficients. Therefore, measuring these functions in any one viscometric flow characterizes a simple material for the whole class of flows.

For shearing flows that are not steady, direct analogs of the viscosity and normal-stress coefficients are often defined. Of course, these analogs are functions of time and they reflect the nonconstant stretch history at a particle. They apply to flows that are not MWCSH.

There are useful MWCSH other than viscometric flows. An example is steady extensional flow, that is, a pure stretching motion with constant rates of stretch. More examples are described in Huilgol (1975).

2.4.1 Rheogoniometry—Cone and Plate

As was indicated in Sec. 2.1 on viscometry, the cone-and-plate viscometer, when used with a shallow cone and at low rates of shear, is particularly convenient for measurement of nonlinear fluids because the deformation can then be taken as homogeneous and, in steady flows, viscometric. If means are supplied to measure the axial force on the cone and to vary the rate of rotation and the spacing between cone and plate, the device is called a rheogoniometer and it may be used to measure the normal-stress coefficients $\psi_1(\kappa^2)$ and $\psi_2(\kappa^2)$, as well as viscosity and, indeed, even certain mechanical properties of unsteady flows.

2.4.1.1 Steady-State Properties. In the commonest mode of operation of the rheogoniometer, the test fluid is supported in the space between cone and plate by the surface tension on a free boundary bridging from plate to cone rim. It is customary to assume that this surface is a segment of a sphere and that these surface-tension forces can be ignored with respect to the magnitudes of the normal-stress differences.

The viscosity is measured through Eq. (65) above; the first normal-stress difference, from

$$\psi_1(\kappa^2) = 2F/\pi\kappa^2R^2, \quad (70)$$

where κ is the rate of shear as given in Eq. (65), F is the axial force, and R is the radius.

Sources of error in applying Eq. (70) have been widely discussed. See Dealy (1982), for instance. Even when there is negligible transverse flow, inertial effects can influence the axial force at high shear rates. Corrections for such errors are possible. More serious are the edge effects that act on the free surface. For many materials, an indentation in the free surface appears at high shear rates, cutting deeply into the sample with increasing shear rate and completely invalidating the analysis. As a result of this and other difficulties, the method of Eq. (70) is necessarily restricted to measurements at rates of shear usually less than a few reciprocal seconds.

Measurements of the second normal-stress coefficient are even more difficult. Jackson and Kaye (1966) have assumed that a viscometric flow persists when the cone apex is displaced a distance d from contact with the plate. Upon making assumptions similar to those above, they arrive at the following equation:

$$\psi_2(\kappa^2) = -\frac{\Omega\theta}{\pi\kappa^2R}\left.\frac{\partial F}{\partial d}\right|_{d=0} - \kappa\frac{d\psi_1(\kappa^2)}{d\kappa}. \quad (71)$$

With $\psi_1(\kappa^2)$, calculated using Eq. (70), and from further data of axial force vs displacement at fixed shear rate, this equation can be used to measure the second normal-stress coefficient. However, the difficulties of accurate measurement coupled with the inaccuracies inherent in numerical or graphical evaluation of a derivative often make this method imprecise.

Another method of measuring the second normal-stress coefficient is possible with further modification of the apparatus. If means are provided to measure the pressure acting against the plate at points several distances from the apex of the cone, the logarithmic gradient of this pressure can be estimated. If the method is valid, this gradient will be a constant. Then ψ_2 can be found from this constant and ψ_1 through the following equation:

$$\psi_2(\kappa^2) = -\tfrac{1}{2}\left[\psi_1(\kappa^2) + \frac{1}{\kappa^2}\frac{dp_n(r)}{d\ln r}\right], \tag{72}$$

where $p_n(r)$ is the pressure against the plate at radial distance r.

With such measurements and with other methods, it has been found that, for polymeric materials, the second normal-stress coefficient is negative and of a magnitude only a fraction of the first normal-stress coefficient.

2.4.1.2 Time-Dependent Properties. For unsteady shearing, assuming effects due to inertia are ignorable, response functions may still be measured in principle using Eqs. (65) and (70) as in the steady shear case. Of course, in this case these responses are functions of time. For the most part, such measurements have been done on deformation histories termed *startup flow* or *cessation of steady shearing flow*, or on combinations of an oscillation or step displacement with a steady shearing. In practice, the actual measurements are usually confined to the corresponding analogs of viscosity and the first normal-stress difference since the second normal-stress difference is so difficult to measure. Usually, the cone-and-plate geometry is used.

Startup flow is achieved by suddenly shearing material that is at rest and in equilibrium, simultaneously observing the development of stress with time. The *shear-stress growth coefficient* and the *first normal-stress growth coefficient* are usually designated by $\eta^+(t,\kappa)$ and $\psi_1^+(t,\kappa)$, respectively, in an obvious generalization of the analogous steady-state response functions of Eq. (68). For many polymeric materials, in this flow a stress overshoot occurs in which the response initially grows, goes through a maximum, and then settles down to a steady-state value. The measurement is of considerable significance in relating the phenomenological rheology of the material to the micromechanics of entanglements of polymeric chains.

Cessation of steady shearing flow is achieved by suddenly stopping the shearing of a sample that had reached an equilibrium at some fixed rate of shear and recording the response as the stresses decay and the sample approaches equilibrium at rest. The corresponding response functions, the *shear-stress decay coefficient* and the *first normal-stress decay coefficient*, are defined in the obvious way and designated as $\eta^-(t,\kappa)$ and $\psi_1^-(t,\kappa)$, respectively.

Other time-dependent flows are sometimes measured with a cone-and-plate viscometer—for example, oscillatory shear or the combination of steady shear with a small sudden displacement or a small oscillation. It is also sometimes useful to do creeping shears and elastic recovery measurements.

2.4.2 Rheometry—Other Shearing Devices Because cone-and-plate rheometers produce a homogeneous rate of shear throughout the test sample, they are simple to analyze. Even in steady flow, however, the accurate measurement of ψ_2 is not easy. A number of other devices utilizing inhomogeneous shear are sometimes more or less useful. Only a few will be mentioned.

2.4.2.1 Torsional Flow. A *parallel-plate viscometer* shears the sample between a fixed plate and a parallel disk rotating with angular velocity ω. If the flow is assumed to be a slipping between plane-parallel rotating disks whose edges form a cylindrical boundary, that flow is viscometric. Because of inertia, the momentum balance of Eq. (10) cannot be exactly satisfied with such a flow, but with narrow spacing and at low shear rates it is very closely approximated. It is obvious that the shear rate is not homogeneous in this flow, and differentiation of data is needed to evaluate the viscometric coefficients. The viscosity is found from the following equation:

$$\eta(\kappa) = \frac{M}{2\pi R^3\kappa}\left[3 + \frac{d\log M(\kappa)}{d\log\kappa}\right], \quad \kappa = \frac{\omega R}{d}, \tag{73}$$

where M is the torque, R is the radius of the disk, and d is the spacing between plate and disk. Data that cannot be reduced to a unique curve when plotted as $M/(2\pi R^3)$ vs κ indicate that the flow is not as assumed.

If the thrust normal to the disk or plate can be measured as a function of ω, a combination of normal-stress coefficients can be measured. Then

$$\psi_1(\kappa^2) - \psi_2(\kappa^2) = \frac{F}{\pi R^2 \kappa^2}\left[2 + \frac{d \log F}{d \log \kappa}\right], \qquad (74)$$

where F is the total force acting normally on the disk or plate.

2.4.2.2 Hole Pressure and Channel Flow. Prior to 1968, many rheological measurements relied upon determining the fluid thrust against a solid wall by reading pressure on a gauge connected to a small hole drilled in the wall. It was known that this method worked well for a Newtonian fluid where there is only a small error caused by inertia, which can be made ignorable by reducing the diameter of the flow hole. It was assumed that this would also apply to the errors for fluids with more complex rheology since experiments showed that for such fluids also, the pressure measurements approached a constant value in the limit for decreasing diameter of the hole. However, Broadbent *et al.* (1968) found that, for a viscoelastic fluid, there is a systematic error not caused by inertia that is intrinsic to the nonlinear material properties and cannot be eliminated by reducing the hole diameter. It is important to realize that the size of this error, sometimes called *hole pressure,* is often of the same order as that of the normal-stress differences and it cannot be ignored in rheological measurements. Indeed, we now know that early measurements of the second normal-stress difference that ignored this effect did not even have the correct sign. Tanner and Pipkin (1969) illustrated the mechanism for this hole pressure by an exact calculation of a second-order fluid in a plane flow over a deep narrow slot, the simplest example of the phenomenon. Measurements of normal stresses published prior to 1969 should not be relied on without checking to see if they are free of hole-pressure error. Once the mechanism for hole pressure was understood, the phenomenon was viewed as a possible means to measure normal-stress coefficients. Calculations suggest that measurements of hole pressure for shearing flow across a slotted trench relate directly to the first normal-stress coefficient; for shearing flow along a slotted trench, to the second normal-stress difference; and for a circular hole, to some linear combination of these. Discussion of the application of these ideas will be found in Lodge and Vargas (1983).

A direct method of observing the second normal-stress difference is possible with the flow under gravity of a viscoelastic fluid down a tilted open channel. If the channel has a flat bottom with perpendicular side walls, the shearing will vary from the walls to the center of the channel. For a Newtonian fluid, this does not affect the shape of the surface, which would be flat. For a viscoelastic fluid, however, the surface will curve in response to the adjustment of the height of the fluid across the flow to counterbalance the varying normal stress. In principle, for a negative second normal-stress coefficient this will produce an upward bulge, and for positive, a concave surface. In practice, this curvature can be measured and used to estimate the second normal-stress coefficient. It is found to be an upward bulge for polymeric fluids.

2.5 Complex Flows

In general, the three viscometric coefficients, functions of shear rate, are not sufficient to characterize completely the mechanical behavior of a material, even when restricted to the special case of MWCSH. Measurements of flows more complex than simple shearing may be required. For instance, in the flow between eccentric rotating discs discussed below, five coefficients, each a function of two variables, are needed for complete characterization. Other flows, such as extensional flow, also discussed below, cannot be viewed as superpositions of shearing. They are then completely unrelated to viscometric flows but closely related to technologically important flows.

2.5.1 Orthogonal Rheometry An interesting flow, which is possible between parallel disks that rotate with the same angular velocity on axes through their slightly displaced centers, is a steady flow with circular flow lines that is a MWCSH but it is not viscometric. The deformation experienced by a particle in this flow is a shearing that rotates with respect to the particle. The particle itself revolves and rotates in laboratory coordinates. Huilgol (1969) shows that, ignoring inertia, for simple materials in this flow the resulting stress is a constant and characterized by five material functions of two parameters, the angular velocity and the eccentricity of the centers. The *eccentric disk rheometer* or *Maxwell–Chartoff orthogonal rheometer* is a device to generate and measure this flow. Actually, for practical reasons, the apparatus drive one disk and allow the other to rotate freely on a low-friction bearing. There is no control of the cylindrical surface of the sample between disks. Tangential and possibly normal forces also are measured at one disk. Kearsley (1970) has pointed out that, for the flow that is assumed to occur, since there is no net torque on the disks, all the energy dissipated by the flow must be supplied by forces on the uncontrolled cylindrical surface. Thus, in fact, the apparatus must produce a flow at variance with the one assumed. Nevertheless, the devices seem to work satisfactorily for small eccentricities and rates of rotation such that inertial effects can be ignored. In this limit, the five functions are reduced to three expressing linear properties. The storage and loss moduli of Eqs. (31) and (32) can be measured in this way. If the x and y axes are taken in the plane of a disk, parallel and perpendicular, respectively, to the projection of the line connecting centers, then the dynamic moduli may be found from the following equations:

$$G'(\omega) = \frac{F_y h}{\pi d R^2}, \quad G''(\omega) = \frac{F_x h}{\pi \Omega d R^2}, \quad \omega = \Omega, \tag{75}$$

where G' and G'' are the storage and loss moduli, respectively, F_x and F_y are components of force acting tangentially on a disk, d is the offset of the disk axes, h is the thickness of the sample filling the space between disks, R is the radius of the disks, and Ω is the angular velocity of the disks, which is equal to ω, the frequency in the argument of the moduli. An important advantage of this method is that the dynamic moduli may be measured from steady forces and no phase-angle measurement is needed.

Two comparable devices are the *balance rheometer* or *Képès rheometer* and the *eccentric cylinder rheometer,* which operate on principles very similar to those of the eccentric disk rheometer. The former uses the flow between a rotating hemispherical cup and a concentric sphere rotating at the same angular velocity but about a slightly tilted axis. The latter uses the flow between a rotating cylindrical cup and a cylindrical bob rotating at the same angular velocity about a slightly displaced parallel axis. These devices are most useful for measuring dynamic linear properties in the limiting case of small deformation. The appropriate equations and a discussion of possible sources of error are to be found in Walters (1975).

2.5.2 Extensional Rheometry. *Extensional flows* are flows in which the material is stretched in simple extension. These flows are of great technological interest because they correspond to a commercially important process, spinning of fibers, but they are also of great interest to the polymer physicist because they have a strong effect in aligning polymer chains. More general flows occurring in commercial processing involve flow in contracting or expanding pipes or channels, flows that may be viewed as combining extensional and shearing flows. Sometimes the term *elongational flows* is reserved for these mixed flows because there is reason to believe that the extensional component will play a dominant role. A survey of the literature will show that much less is known about extensional flows than simple shearing flows. The reason for this is that there are many practical difficulties in generating and measuring well-defined extensional flows. Petrie (1979) gives a thorough survey of the literature up to the time of writing.

In simple extension, a cylinder of fluid is stretched along its axis while its surface is stress free. If the length of the cylinder increases exponentially in time, the local deformation rate is constant. When any transients die out, the history is effectively a MWCSH. In that event, the stress will achieve a steady-state value and a *tensile viscosity* or *Trouton viscosity,* $\eta_E(\alpha)$, can be defined as follows:

$$\eta_E(\alpha) = \lim_{t\to\infty} \frac{f(t)L_0 \exp(\alpha t)}{A_0\alpha}, \quad L(t) = L_0 \exp(\alpha t), \tag{76}$$

where $L(t)$ is the length of the cylinder, L_0 and A_0 are the length and cross section of the cylinder at time $t = 0$, $f(t)$ is the force of stretching, and α is a constant that determines the exponential of the stretching. It is apparent from this equation that the transients may never die out for some constitutive equations, which thus predict no limiting Trouton viscosity. A Newtonian fluid has a Trouton viscosity that is independent of shear rate: It is three times the Newtonian viscosity. Most polymeric fluids have Trouton viscosities that are strong functions of the stretching rate and that, at very low rates of deformation, appear to approach the Newtonian value of three times the zero shear viscosity. On the other hand, as the deformation rate is increased for these elastic fluids, the ratio of Trouton to shear viscosity can rise very rapidly and, at very high rates of shear, a steady state may be very difficult or impossible to reach. To measure Trouton viscosity, a controlled exponential stretching must be maintained and monitored until a steady state of stress sets in. This is extremely hard to achieve, particularly for low-viscosity fluids and high rates. Only very limited and scattered data will be found in the literature. Meissner (1972) has developed an instrument in which a strand of polymer melt is stretched between a pair of gearlike rollers rotating at constant angular velocity, thus producing a constant stretch rate while rapidly thinning. With this device he has drawn samples out to almost 100 times their original length and has made measurements at constant extension rates of several reciprocal seconds. More recently, a device is described by Tirtaatmadja and Sridhar (1993) with which they have measured Trouton viscosities at stretch rates of 5 s^{-1}.

It is much easier to produce high extension rates with instruments that need not produce a MWCSH: that is, with geometry for which a particle of fluid does not experience a constant history even if the flow is steady. Because high extension rates often occur in industrial processes, constant-force extensometers and various fiber-spinning devices of this type, capable of high stretch rates, have been used for technological measurements. An inconvenience of such measurements is that the results are very much a function of the particular history. Typically, in these cases, an apparent extensional viscosity is defined as the ratio of a local tensile stress and a stretching rate at some arbitrarily designated point on the threadline. Such measurements will not yield reliable values of η_E but they can give useful rough comparisons of materials. The tubeless siphon (mentioned in Sec. 1.4, above) has been used for this purpose. Oliver and Bragg (1974) developed a triple-jet device for determining extensional flow characteristics at very high stretching rates. Two converging jets at high velocity strike a central low-velocity jet causing it to stretch. The extensional strain rate is inferred from photographs. Results are somewhat sensitive to the flow conditions in the reservoirs.

Biaxial stretching of a sheet of polymeric material occurs in the commercially important processes of film blowing and blow molding and, for this reason, measurements of *biaxial extensional viscosity* are needed. Meissner (1985) has developed rotary clamps that may be arranged to produce anisotropic biaxial stretching of sheets of polymeric materials at constant rates. With this apparatus, biaxial analogs of the Trouton viscosities may be measured.

2.5.3 Plastometer The *plastometer* or *squeeze-film viscometer* is a simple device in which a cylindrical disk of test sample is squeezed between two parallel plates by a fixed load, usually a weight. The sample disk thins and flows radially under the constant load, while the spacing between the plates is monitored. The device can be used to measure viscosity but it is most interesting as a device to measure yield stress.

There is no general agreement on the meaning of the term "yield stress." The classical, theoretical meaning is "a stress below which no flow or, at least, no unrecoverable flow occurs." For example, σ_y of the Bingham body, Eq. (55), has a yield stress. Most people would agree that real materials actually undergo some flow at any stress: it is just a matter of the time scale necessary to perceive the effect. On the other hand, a pragmatist might define yield stress in terms of a particular measurement and time scale. In practice, many materials exhibit a yield stress, for all practical purposes, in the time scale of common usage.

The plastometer is used to measure the yield stress by observing the limiting spacing between parallel plates when flow ceases. The shear yield stress is then given by

$$\sigma_y = \frac{3\pi^{1/2}h_L^{5/2}F}{2V^{3/2}}, \quad V = \pi R_L^2 h_L, \tag{77}$$

where h_L is the limiting spacing, F is the loading force, R_L is the limiting radius of the disk, and V is the volume of the sample.

The actual flow of a material with yield stress in a plastometer is not completely understood. Covey and Stanmore (1981) give a survey of several widely differing solutions for a Bingham body.

3. DIMENSIONLESS PARAMETERS

The dimensional analysis of flows of rheologically complex fluids requires dimensionless parameters beyond the Reynolds number useful for Newtonian fluids. Indeed, the concept of a Reynolds number itself must be slightly generalized to accommodate the dependence of viscosity on rate of shearing. One way of doing this is as follows:

$$R_e = \rho UL/\eta(\kappa), \tag{78}$$

where R_e is the Reynolds number, U is a characteristic velocity, L is a characteristic length, and $\eta(\kappa)$ is the viscosity at the local rate of shear, κ. In this definition, Reynolds number may vary throughout the flow.

A *Weissenberg number, W*, is defined to represent a ratio of elastic to viscous effects:

$$W = \lambda_{\mathrm{rlx}}U/L, \tag{79}$$

where λ_{rlx} is a typical relaxation time of the material, U is a typical fluid velocity, and L is a relevant length. In viscometric flows, W is a measure of the ratio of the normal force N_1 to the viscous force, $\kappa\eta(\kappa)$. More generally, it is meant to be a rough ratio of elastic forces to viscous forces.

A Deborah number, D_e, is defined as the ratio of a typical relaxation time of the material to a typical time scale of a measurement or of a residence time or flow time in a flow system. If this time scale is called T, the Deborah number is defined as

$$D_e = \lambda_{\mathrm{rlx}}/T. \tag{80}$$

The name of this dimensionless number refers, whimsically, to the biblical prophetess Deborah who affirmed, according to one reading, "The mountains flow before the Lord." To a rheologist, this might seem a metaphorical statement that, if one's time scale were long enough, the most solid of materials would appear to flow.

Accordingly, the significance of the Deborah number is that in situations with high values of D_e the material tends to appear solidlike while at low D_e the material appears more fluid. Therefore, the appropriate constitutive equation for a situation often depends upon one's time scale as measured by the Deborah number.

Clearly, the definitions of such dimensionless parameters are far from unique, and the choice of the appropriate times, lengths, and velocities to be used in calculating them is open to considerable personal taste. But sensible, consistent definitions of these concepts have proved very useful for scaling and comparing materials and flows, especially in the treatment of technological phenomena.

GLOSSARY

Blow Molding: A process in which a bubble of molten material is shaped by inflation and solidification within a mold. Ordinary polyethylene gallon milk jugs are usually made in this way.

Body Force: A force from an external influence that acts on a particle of material, e.g., gravity. Body forces are vectors with dimensions of force per unit volume.

Body Moment: A force couple from an external influence that acts on a particle of material, e.g., the effect of a magnetic field acting on a magnetic dipole distribution.

Constitutive Equation: An equation relating to a particular material giving local physical properties of the material in terms of local influences. In this article, constitutive equations give mechanical properties in terms of deformation or deformation history, e.g., a stress–strain law.

Couple Stress: An interaction by which a force couple is exerted point-wise between contiguous surfaces of a material. Couple stresses are represented by an antisymmetric component of stress tensors.

Double Tensor: A tensor calculated from points in two different spaces that transforms as a tensor under coordinate transformations of either. For example, the deformation gradient **F**(*t*) is a double tensor because it depends tensorially on the coordinate system of the configuration **x**(*t*) at time *t* as well as that of reference configuration **X**. (See Ericksen, 1960, p. 805.)

Elastomer: A polymeric substance capable of rubbery behavior in which large, essentially recoverable stretching can occur—for example, natural rubber.

Film Blowing: A process for forming polymer films in which a tube of molten plastic is extruded from an annular die, stretched axially and radially, and solidified by being drawn over a bubble of pressurized gas with rollers that flatten and pull the tube while sealing off the gas bubble.

Functional: A noun meaning a mathematical rule for calculating a quantity from a set of functions. It is a generalization of the idea of a function as a rule for calculating a quantity from a set of numbers. The functionals used in rheology are usually tensor valued, that is, they are rules for evaluating a tensor from histories of deformations.

Hamilton–Cayley Theorem: The theorem that every matrix satisfies its characteristic equation. Thus, for a second-order symmetric tensor in three-space such as stress or strain,

$$\mathbf{S}^3 - I_1(\mathbf{S})\mathbf{S}^2 + I_2(\mathbf{S})\mathbf{S} - I_3(\mathbf{S})\mathbf{1} = 0$$

where S is the tensor and the I_i are the principal scalar invariants. This theorem allows powers of these tensors to be expressed as a polynomial in lesser powers.

Isochoric: An isochoric deformation is a deformation without volume change. An incompressible material may undergo only isochoric deformations.

Polar Decomposition Theorem: A theorem of linear algebra that, as it applies to real tensors with strictly positive determinants, assures that they are equivalent to the product (in either order) of a proper orthogonal tensor and a symmetric, positive-definite tensor.

Principal Scalar Invariants: A set of scalar functions of the arguments of a matrix, invariant under orthogonal matrix transformations. For a second-order, symmetric tensor **S** in three-space, such as stress or strain, there are three scalar invariants as follows:

$$I_1(\mathbf{S}) \equiv \mathrm{tr}(\mathbf{S}), \quad I_2(\mathbf{S}) \equiv \tfrac{1}{2}[\mathrm{tr}^2(\mathbf{S}) - \mathrm{tr}(\mathbf{S}^2)],$$
$$I_3(\mathbf{S}) \equiv \det(\mathbf{S}).$$

Functions of these principal invariants—for instance, the principal values or eigenvalues—are also invariant. If the eigenvalues are available, the principal invariants may be written

$$I_1 = \lambda + \mu + \nu, \quad I_2 = \lambda\mu + \lambda\nu + \mu\nu,$$
$$I_3 = \lambda\mu\nu,$$

where λ, μ, and ν are the eigenvalues.

Shear: An isochoric deformation in which parallel surfaces of a material are displaced in their planes while remaining parallel. *Simple shear* is the isochoric deformation undergone by material sandwiched between two adherent, parallel, rigid plates that deform it by parallel displacement. *Pure shear* is a deformation without rotation composed of two orthogonally oriented stretches of reciprocal magnitudes. Since it is an isochoric deformation, the stretch in the third orthogonal direction must be unity. Simple shear may be decomposed locally into a pure shear plus a rigid rotation.

Stress Deviator: The deviatoric part of the stress tensor. The stress deviator has trace zero. The stress deviator **s** of stress $\boldsymbol{\sigma}$ is given by

$$\mathbf{s} = \boldsymbol{\sigma} - \tfrac{1}{3}\,\mathrm{tr}(\boldsymbol{\sigma})\mathbf{1}$$

Stretch: The stretch of an infinitesimal line segment is the ratio of the lengths before and after deformation. The eigenvalues of a stretch tensor give the stretches of lines oriented in the three, mutually perpendicular, principal directions of the tensor, the *principal stretches.*

Surface Traction: A vector or vector field, the force per unit area acting on a surface having, in general, both normal and tangential components. The surface may be a boundary of a body, an interface between bodies, or a notional surface in the interior of a body. A stress tensor may be thought of as a linear transformation that operates on a unit vector normal to a surface to produce a traction vector acting on the surface.

Works Cited

Bagley, E. B. (1957), *J. Appl. Phys.* **28,** 624–627.

Barr, G. (1931), *A Monograph of Viscometry,* New York: Oxford Univ. Press.

Bernstein, B. (1966), *Acta Mech.* **II/4,** 329–354.

Bernstein, B., Kearsley, E., Zapas, L. (1963), *Trans. Soc. Rheol.* **7,** 391–410.

Bernstein, B., Kearsley, E., Zapas, L. (1964), *J. Res. Natl. Bur. Stand. B* **68,** 103–113.

Bernstein, B., Kearsley, E., Zapas, L. (1965), *Trans. Soc. Rheol.* **9,** 27–39.

Bird, R. B., Dai, G. C., Yarusso, B. J. (1987), *Rev. Chem. Eng.* **1,** 1–70.

Blatz, P. J., Ko, W. L. (1962), *Trans. Soc. Rheol.* **6,** 223–251.

Broadbent, J. M., Kaye, A., Lodge, A. S., Vale, D. G. (1968), *Nature* **217,** 55–56.

Coleman, B. D., Markovitz, H., Noll, W. (1966), *Viscometric Flows of Non-Newtonian Fluids,* New York: Springer-Verlag.

Cosserat, E., Cosserat, F. (1909), *Théorie des Corps Déformables,* Paris: A. Hermann et Fils.

Covey, G. H., Stanmore, B. R. (1981), *J. Non-Newtonian Fluid Mech.* **8,** 249–260.

Dealy, J. M. (1982), *Rheometers for Molten Plastics,* New York: Van Nostrand Reinhold.

De Gennes, P. G. (1971), *J. Chem. Phys.* **55,** 572–579.

Doi, M., Edwards, S. F. (1978), *J. Chem. Soc. Faraday Trans. II* **74,** 1789–1829.

Ericksen, J. L. (1954), *Z. Angew. Math. Phys.* **5,** 466–486.

Ericksen, J. L. (1960), "Tensor Fields," in: S. Flügge (Ed.), *Handbuch der Physik,* Band III/1, Berlin: Springer-Verlag, pp. 793–858.

Ericksen, J. L. (1961), *Trans. Soc. Rheol.* **5,** 23–34.

Eringen, A. C. (1962), *Nonlinear Theory of Continuous Media,* New York: McGraw-Hill.

Green, A. E., Rivlin, R. S. (1964), *Arch. Rat. Mech. Anal.* **17,** 113–147.

Hill, R. (1950), *The Mathematical Theory of Plasticity,* New York: Oxford Univ. Press.

Huilgol, R. R. (1969), *Trans. Soc. Rheol.* **13,** 142–151.

Huilgol, R. R. (1975), *Continuum Mechanics of Viscoelastic Liquids,* New York: Wiley.

Jackson, R., Kaye, A. (1966), *Brit. J. Appl. Phys.* **17,** 1355–1360.

Kearsley, E. A. (1970), *J. Res. Natl. Bur. Stand. C* **74,** 19–20.

Kearsley, E. A. (1980), *J. Appl. Phys.* **51,** 4541–4542.

Kearsley, E. A., Zapas, L. J. (1980), *J. Rheol.* **24,** 483–500.

Leslie, F. M. (1968), *Arch. Rat. Mech. Anal.* **28,** 265–283.

Lodge, A. S., de Vargas, L. (1983), *Rheol. Acta* **22,** 151–170.

Meissner, J. (1972), *Trans. Soc. Rheol.* **16,** 405–420.

Meissner, J. (1985), *Chem. Eng. Commun.* **33,** 159–180.

Merrington, A. C. (1949), *Viscometry,* London: Edward Arnold.

Mooney, M. (1940), *J. Appl. Phys.* **11,** 582–592.

Obata, Y., Kawabata, S., Kawai, H. (1970), *J. Polym. Sci.* **8,** 903–919.

Oliver, D. R., Bragg, R. (1974), *Rheol. Acta* **13,** 830–835.

Peng, S. T. J., Landel, R. F. (1976), *J. Appl. Phys.* **47,** 4255–4260.

Penn, R. W., Kearsley, E. A. (1976), *Trans. Soc. Rheol.* **20,** 227–238.

Petrie, C. J. S. (1979), *Elongational Flow,* London: Pitman.

Rivlin, R. S., Saunders, D. W. (1951), *Philos. Trans. R. Soc. (London)* **A243,** 251–273.

Spencer, A. J. M. (1971), "Theory of Invariants," in: A. C. Eringen, (Ed.), *Continuum Physics,* Vol. I, New York: Academic Press, pp. 239–353.

Tanner, R. I., Pipkin, A. C. (1969), *Trans. Soc. Rheol.* **13,** 471–484.

Tanner, R. I. (1985), *Engineering Rheology,* Oxford, U.K.: Clarendon.

Tirtaatmadja, V., Sridhar, T. (1993), *J. Rheol.* **37,** 1081–1102.

Toupin, R. A. (1962), *Arch. Rat. Mech. Anal.* **11,** 385–414.

Valanis, K. C., Landel, R. F. (1967), *J. Appl. Phys.* **38,** 2997–3000.

Wagner, M. H., Stephenson, S. E. (1979), *J. Rheol.* **23,** 489–504.

Walters, K. (1975), *Rheometry,* London: Chapman and Hall.

Yang, T. M. T., Krieger, I. M. (1978), *J. Rheol.* **22,** 413–419.

Further Reading

Continuum Mechanics

Jaunzemis, W. (1967), *Continuum Mechanics,* New York: Macmillan.

Truesdell, C., Toupin, R. (1993), in: S. Flügge (Ed.), *Encyclopedia of Physics,* Vol. III/1, *The Clas-*

sical Field Theories, 2nd ed., New York: Springer-Verlag.

Truesdell, C., Noll, W. (1993), in: S. Flügge (Ed.), *Encyclopedia of Physics,* Vol. III/3, *The Nonlinear Field Theories of Mechanics,* 2nd ed., New York: Springer-Verlag.

General Rheology

Barnes, H. A., Hutton, J. F., Walters, K. (1989), *An Introduction to Rheology,* Amsterdam: Elsevier.

Boger, D. V., Walters, K. F. (1993), *Rheological Phenomena in Focus,* Amsterdam: Elsevier.

Cheremisinoff, N. P. (1988), *Encyclopedia of Fluid Mechanics,* Vol. 7, *Rheology and Non-Newtonian Flows,* Houston: Gulf Publishing.

Joseph, D. D. (1990), *Fluid Dynamics of Viscoelastic Liquids,* New York: Springer-Verlag.

Middleman, S. (1968), *The Flow of High Polymers,* New York: Wiley.

Constitutive Equations

Bird, R. B., Armstrong, R. C., Hassager, O. (1987), *Dynamics of Polymeric Liquids,* Vol. 1, *Fluid Mechanics,* and Vol. 2, *Kinetic Theory,* 2nd ed., New York: Wiley.

Ferry, J. D. (1961), *Viscoelastic Properties of Polymers,* New York: Wiley.

Larson, R. G. (1988), *Constitutive Equations for Polymer Melts and Solutions,* Stoneham: Butterworths.

Schowalter, W. R. (1978), *Mechanics of Non-Newtonian Fluids,* Oxford, U.K.: Pergamon.

Tanner, R. I. (1985), *Engineering Rheology,* Oxford, U.K.: Clarendon.

Tschoegl, N. W. (1989), *The Phenomenological Theory of Linear Viscoelastic Behavior: An Introduction,* New York: Springer-Verlag.

Rheometry

Coleman, B. D., Markovitz, H., Noll, W. (1966), *Viscometric Flows of Non-Newtonian Fluids,* New York: Springer-Verlag.

Dealy, J. M. (1982), *Rheometers for Molten Plastics,* New York: Van Nostrand Reinhold.

Dealy, J. M., Wissbrun, K. F. (1990), *Melt Rheology and Its Role in Plastics Processing,* New York: Van Nostrand Reinhold.

Walters, K. (1975), *Rheometry,* London: Chapman and Hall.

Computational Mechanics

Crochet, M. J., Davies, A. R., Walters, K. F. (1984), *Numerical Simulation,* Amsterdam: Elsevier.

Gordon, G. V., Shaw, M. (1994), *Computer Programs for Rheologists,* Cincinnati: Hanser/Gardner.

Washizu, K. (1968), *Variational Methods in Elasticity and Plasticity,* Oxford, U.K.: Pergamon.

RHEOLOGY CONTROLLED BY MAGNETIC FIELDS

J. M. GINDER, *Research Laboratory, Ford Motor Company, Dearborn, Michigan, U.S.A.*

Introduction 487
1. **Ferrofluids** 490
1.1 Composition 490
1.2 Field-Dependent Viscosity 490
1.3 Applications of Ferrofluids 491
2. **Magnetorheological Fluids** 491
2.1 Composition 491
2.2 Magnetic-Field–Induced Structure Formation 492
2.3 Steady-Shear Rheology 492
2.4 Oscillating-Shear Rheology 493
2.5 Magnetic Properties 494
2.6 Magnetostatic Models of Magnetorheology 494
3. **Magnetic Powders** 496
3.1 Composition 496
3.2 Magnetic-Powder Rheology 497
4. **Magnetorheological Elastomers** 497
4.1 Composition 497
4.2 Field-Dependent Viscoelastic Response 497
5. **Applications of Magnetorheological Materials** 498
5.1 Magnetic-Circuit Considerations 498
5.2 Torque-Transfer Applications 498
5.3 Damping Applications 499
5.4 Process Applications 500
5.5 Outlook 500
Glossary 501
Works Cited 501
Further Reading 502

INTRODUCTION

The formation of particle chains when magnetizable particles are exposed to a magnetic field is familiar to anyone who has experimented with iron filings and a permanent magnet. This ubiquitous phenomenon is also the basis for the sometimes dramatic field-induced changes in rheology exhibited by a variety of magnetorheological, or MR, materials (for a review of rheological phenomena, see RHEOLOGY). The application of a field to an initially random suspension of magnetizable particles results in the formation of chains, columns, or more complex solidlike structures aligned along the field direction, as shown in Fig. 1. Anisotropic interparticle forces induced by the applied field give rise to these structural and rheological changes. The formation of particle chains in an external field increases the electromagnetic energy stored in a MR suspension. The chains or columns are tilted or broken by applying a shear strain; as this deformation reduces the stored energy, a nonzero shear stress is required to impose the strain, as shown schematically in Fig. 2.

The rheology of many MR materials is commonly described by the Bingham plastic model (see, e.g., Macosko, 1994). Central to the Bingham model is the yield stress τ_y, which is a measure of the strength of solid materials. For applied stresses τ below τ_y, many materials are viscoelastic solids having nonzero shear moduli or stiffnesses G'. The shear stress required to produce a shear strain γ is then

$$\tau = G'\gamma, \quad \tau < \tau_y. \tag{1}$$

For applied stresses greater than τ_y, the material is able to flow. The Bingham equation relates the stress to the strain rate $\dot{\gamma}$ in this postyield condition:

$$\tau = \tau_y + \eta\dot{\gamma}, \quad \tau \geq \tau_y, \tag{2}$$

where η is the viscosity. These idealized pre-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

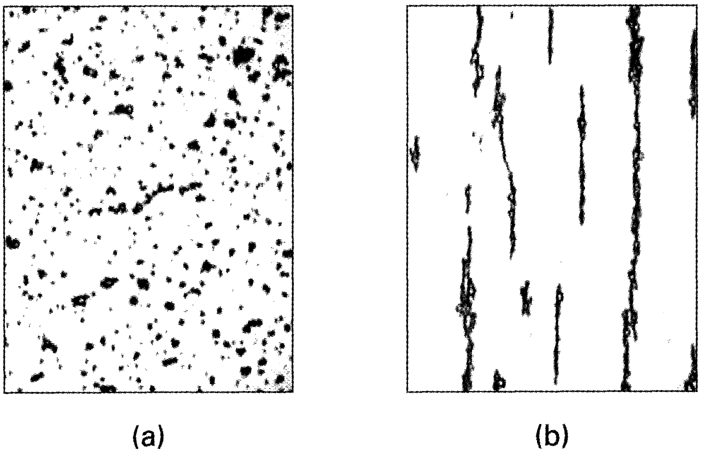

FIG. 1. Optical micrographs (300 × 400 μm) of a thin layer of a dilute suspension of a carbonyl-iron-based MR fluid: (a) zero applied field; (b) after the application of a ~20-kA/m field, showing the formation of particle chains aligned along the field direction.

and postyield behaviors are shown schematically in Fig. 3. While the scientific validity of the yield stress has been questioned, the Bingham model is undeniably quite useful in describing the flow of real materials like MR fluids. Magnetorheological phenomena involve field-dependent viscosities, yield stresses, or shear moduli.

The broad family of field-controllable magnetorheological materials includes ferrofluids, magnetorheological fluids, magnetic powders, and magnetorheological elastomers; the properties of these materials are summarized in Table 1. Ferrofluids, or colloidal magnetic fluids, typically contain magnetic particles less than 10 nm in size; consequently, the magnetorheological effect in ferrofluids is small, involving at most a tripling of the ferrofluid viscosity upon the application of a field (Rosensweig, 1985). Ferrofluids are of considerable utility in sealing, damping, and heat-transfer applications. Magnetorheological fluids are composed of micrometer-size magnetizable particles suspended at high concentration in nonmagnetic liquids. While similar in composition to ferrofluids, the larger particulates in MR fluids enable them to change reversibly from free-flowing liquids to solids having yield stresses of order 100 kPa on the application of a magnetic field. These changes enable the construction of a variety of electromechanical devices such as MR fluid clutches, brakes, and shock absorbers, several of which are presently under development for motion-control applications. Similar phenomena are already utilized in commercially available magnetic-powder clutches and brakes. Magnetorheological elastomers are obtained by dispersing micrometer-size magnetizable particles in a viscoelastic solid like a polymer gel or an elastomer. While these materials are solid under all circumstances, their modulus or stiffness can be varied by an applied field. These MR elastomers may find use in vibration-control applications.

Both the history and the physical properties of MR fluids are closely related to those of their electrical counterparts, the electrorheological (ER) fluids. ER fluids, which are

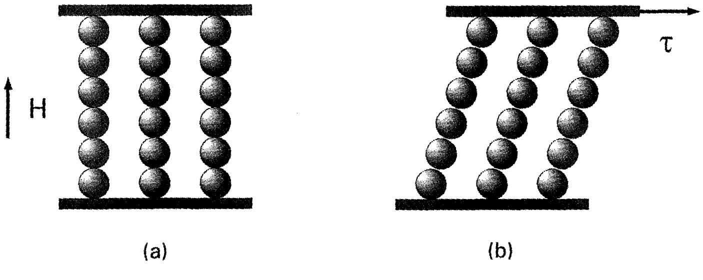

FIG. 2. Idealized structure of a MR fluid: (a) an applied field H induces the formation of chains or columns along the field; (b) a shear stress τ must be applied to deform the field-induced structure.

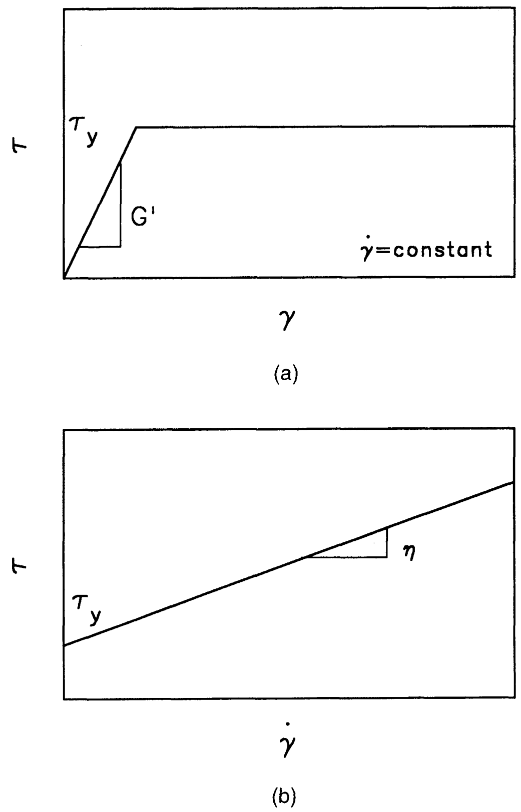

FIG. 3. Schematic rheological behavior of a Bingham solid in (a) preyield viscoelastic regime with shear modulus G' and (b) postyield plastic flow with viscosity η. The yield stress τ_y demarcates the pre- and postyield behavior.

suspensions of electrically polarizable particles in insulating liquids, undergo a transition to a solid state upon the application of a high electric field (see ELECTRORHEOLOGY). ER fluids were first developed by Willis Winslow in the 1940s (Winslow, 1947, 1949). By recognizing the analogy between the electrostatic polarization and magnetostatic magnetization of particles, Jacob Rabinow of the National Bureau of Standards was led to invent the first magnetorheological fluids and devices in the late 1940s (Rabinow, 1948, 1949, 1951). Rabinow's discovery soon inspired a flurry of activity on MR fluids and devices in the United States, England, the Soviet Union, and elsewhere (Parziale and Tilton, 1950; Blunden, 1951; Vorob'yeva, 1965). By the early 1950s it was recognized that dry magnetic powders would also function as field-controllable media for stress transfer (Grebe, 1952). Published activity in MR fluids began to decline in the late 1960s, while ER fluids began to enjoy increasing attention after the development of polymer particulate formulations in the late 1970s. The development of water-free ER fluids in the mid-1980s has led to significant improvements in their thermal and electrical performance. ER fluids having a relatively broad operating-temperature range and low electrical-power consumption are available in pilot-plant quantities from Bayer AG and Bridgestone. While these ER fluids perform well enough to permit the construction of practical devices like industrial dampers, their field-induced yield stresses are not high enough for applications in automotive torque transfer (Hartsock *et al.,* 1991). Moreover, their use requires potentially expensive high-voltage power supplies as well as the appropriate electrical wiring and connectors.

These limitations have led to a renaissance of interest in MR fluids and applications in this decade. A prototype rotary automotive shock absorber utilizing MR fluids was developed by a group at TRW (Pinkos *et al.,* 1993). MR fluids are useful in optical-surface finishing (Holton, 1993). MR-fluid–based devices are now being commercialized

Table 1. Composition and properties of magnetorheological materials.

	Ferrofluid	MR fluid	Magnetic powder	MR elastomer
Particulate material	Magnetite	Iron	Iron alloy	Iron
Particle size	2–10 nm	0.1–10 μm	10–50 μm	10–50 μm
Host material	Oils, water	Oils	Air	Elastomers
Volume fraction	0.02–0.2	0.1–0.5	0.5–0.6	0.1–0.5
Additives	Surfactants	Surfactants, thixotropes	Dispersants	···
Off viscosity (mPa-s)	2–200	100–1000	···	···
Magnetorheological changes	$\Delta\eta/\eta \sim 1$	$\tau_y \sim 100$ kPa	$\tau_y \sim 100$ kPa	$\Delta G' \sim 20$ kPa

by Lord Corporation for use in variable-resistance exercise equipment and seat vibration dampers. Because of their high yield stresses, wide operating-temperature range, and activation using electromagnets driven by low-voltage power supplies, other applications of MR fluids are likely to follow. In fact, magnetic-powder clutches and brakes have been used in motion-control and motor-testing applications for several decades; a magnetic-powder device even serves as the main clutch in an automotive power train.

Recent progress in the field of magnetorheological materials has been made in developing new MR-fluid formulations and inventing novel MR elastomers, as well as in understanding the connection between the magnetic properties of the constituents and the magnetorheological performance of the material. Questions about the long-term stability of these materials and the mass and power consumption of the devices that utilize them remain. This article is intended to review the present state of MR materials and device development as well as to provide an introductory guide for those who may wish to study or utilize these materials. Section 1 briefly reviews the composition, properties, and applications of ferrofluids. In light of the relatively weak magnetorheological effects that they possess, ferrofluids are not the main focus of this article; they nonetheless are of interest here as "proto-MR fluids." The composition, structural, rheological, and magnetic properties of magnetorheological fluids are treated in Sec. 2. Physical aspects of magnetic powders, i.e., dry MR fluids, are reviewed in Sec. 3. Section 4 treats the recently developed magnetorheological elastomers. Section 5 concludes with a discussion of the requirements for and opportunities in devices utilizing MR materials.

1. FERROFLUIDS

1.1 Composition

Ferrofluids, or colloidal magnetic fluids, are typically composed of single-domain particles of magnetite, Fe_3O_4, suspended with the aid of surfactants in polar liquids such as water or nonpolar liquids such as mineral oil. Ternary and higher ferrites, as well as metallic particulates of iron and cobalt, may also be utilized in ferrofluid formulations. The preparation of a stable ferrofluid involves a delicate balance between the attractive and repulsive interparticle forces as well as random Brownian forces on each particle (see, e.g., Rosensweig, 1985). The attractive forces are produced by magnetic dipole–dipole interactions and van der Waals forces caused by fluctuating electric dipoles. The repulsive forces are due to layers of adsorbed surfactant molecules on each particle (giving rise to so-called steric forces) as well as screened Coulomb repulsion if the particles are charged. If the particles are too small, their Brownian motion will greatly reduce the magnetic properties of the fluid. If they are too large, the attractive magnetic and van der Waals forces will cause the particles to agglomerate irreversibly. The choice of surfactant molecule necessary to stabilize the suspension depends on the nature of the suspending fluid and the surface chemistry of the particles, and is critical in preventing irreversible contact between particles. It is found that ferrofluids having particles with diameters between roughly 2 and 10 nm possess useful magnetic properties and acceptable stability (Berkovsky *et al.*, 1993).

1.2 Field-Dependent Viscosity

The relevance of ferrofluids in a discussion of magnetorheological phenomena is that, while ferrofluids remain fluid ($\tau_y = 0$) under all imposed magnetic fields, they exhibit viscosities that increase with applied field. Viscometric measurements on ferrofluids have revealed a field-dependent increase in viscosity (see, e.g., McTague, 1969). The origin of the viscosity enhancement is a single-particle effect caused by the magnetic restoring torque experienced by each suspended particle, $\mathbf{m} \times \mathbf{H}$, where $\mathbf{m}$ is the magnetic-dipole moment of the particle and $\mathbf{H}$ is the magnetic field. This torque favors the alignment of the permanent dipoles along the direction of the applied field and thus opposes the rotation of the particles in a shear flow, giving rise to an increase in the fluid viscosity. Viscosity enhancements $\Delta\eta(H)/\eta(0) \sim 2$ have been obtained (Rosensweig, 1985). Successful theories of the viscosity enhancement account for the effect of Brownian motion on the orientation of the dipoles (Shliomis, 1972).

1.3 Applications of Ferrofluids

The relatively small increases in ferrofluid viscosity with the application of a magnetic field have, to this author's knowledge, not been utilized in any applications. Ferrofluids are quite useful, however, because they are magnetic liquids: They experience magnetic-body forces in a magnetic-field gradient, and so the position of a drop of ferrofluid can be controlled by an applied field. Commercial applications of ferrofluids include rotary seals, magnetic bearings, motor dampers, and loudspeaker voice-coil dampers (Raj and Moskowitz, 1990). Monographs by Rosensweig (1985) and Berkovsky *et al.* (1993) provide excellent reviews of ferrofluid composition, properties, and applications.

2. MAGNETORHEOLOGICAL FLUIDS

2.1 Composition

The composition of MR fluids is similar to that of most ferrofluids: Magnetizable particles are suspended at high concentration in nonpolar or polar liquids with the aid of surfactant molecules. In distinct contrast to ferrofluids, the particles used in MR fluids are large enough to permit reversible particle aggregation in an applied magnetic field. Particle sizes in MR fluids may range from tens of nanometers to tens of micrometers. For magnetic particles above a critical size, which depends on particle shape and on the crystal structure of the constituent material, they consist of multiple magnetic domains. In the absence of an applied field, the particles then possess very small net dipole moments and thus experience little or no magnetic attractive forces. The application of a field causes domain-wall motion (Bean and Jacobs, 1960) that greatly increases the magnitude of the moments and thus the interparticle magnetic attraction. To obtain the largest moments and thus the largest magnetorheological effect, it is desirable to use particulates with high saturation magnetization. For example, magnetite has a saturation magnetization of 0.6 T, while iron has a saturation magnetization of 2.1 T, the largest of any element (see MAGNETIC MATERIALS). Materials for MR particulates can be produced by a number of techniques; one of the most widely used is so-called carbonyl iron, which is produced by the decomposition of iron carbonyl, $Fe(CO)_5$. Other processes to produce magnetic particulates include atomization, electrolysis, and grinding. The optimum particle-volume fraction in MR fluids is determined by a compromise between the benefits of heavy particulate loading—a reduction in sedimentation velocity and an increase in the MR effect—and a detrimental increase in zero-field viscosity. Typical volume fractions are between 0.1 and 0.5.

Since typical MR particulates are micrometers in size and are composed of metals or metal oxides having a mass density greater than that of most suspending fluids, they often suffer sedimentation under the prolonged action of gravity as well as centrifugation under the centrifugal forces encountered in rotary motion. Irreversible sedimentation or centrifugation is disastrous in practical applications of MR fluids and has been combated by several techniques. The most common is to add thixotropic agents such as stearates (Winslow, 1950) to promote the formation of a weakly associated network in the suspending fluid. The network possesses a nonzero yield stress in the absence of an applied field, locking the particles in place under quiescent conditions but permitting flow under low applied stresses. More recent approaches have involved the dispersal of highly anisotropic carbon fibers (Shtarkman, 1992) or nanometer-scale ceramic or organic particles (Shtarkman *et al.*, 1994), which also form thixotropic networks that entrain the magnetizable particles. The orientation of these rodlike fibers in shear flow may also reduce the rate of particle centrifugation. The price that is paid in these approaches is a sometimes substantial increase of the zero-field viscosity of the fluids. One promising alternative approach is to shrink the size of the magnetic particulates sufficiently to reduce the sedimentation velocity, which scales quadratically with the particle size, but not so much as to prohibit magnetorheological behavior. For example, researchers at BASF (Kormann *et al.*, 1994) have produced "nano-MRFs," ferrite-based MR fluids incorporating roughly 30-nm-diam particles suspended in polar liquids. These MR fluids appear to exhibit excellent stability against sedimentation and centrifugation and are not abrasive, but possess smaller yield stresses as a result of the reduced saturation magnetization of ferrites relative to iron.

2.2 Magnetic-Field–Induced Structure Formation

The application of a magnetic field to an isolated magnetizable particle induces the formation of a net magnetic-dipole moment **m** aligned with the field. In the low-field or linear regime, the magnitude of the moment is proportional to the magnetic field at the particle:

$$m = 4\pi a^3 \mu_0 \beta H, \tag{3}$$

where a is the particle radius, $\mu_0 = 4\pi \times 10^{-7}$ H/m is the permeability of free space, $\beta = (\mu_p - \mu_f)/(\mu_p + 2\mu_f)$, and μ_p (μ_f) is the relative permeability of the particle (fluid). In high fields in which the particle magnetization has saturated, the magnetization is independent of field and is given by

$$m = \frac{4\pi a^3}{3}\mu_0 M_s, \tag{4}$$

where M_s is the saturation magnetization of the particle. Pairs of particles experience an anisotropic force resulting from the dipole–dipole interaction energy

$$U = -\frac{m^2}{4\pi\mu_0 R^3}(3\cos^2\theta - 1), \tag{5}$$

where **R** is the vector separation between particles and θ is the angle between **m** and **R**. This anisotropic interaction, which leads to repulsive forces for particles aligned with moments head-to-head ($\theta \sim \pi/2$) and attractive forces for head-to-tail alignment ($\theta \sim 0$), promotes the formation of particle chains aligned along the field direction, as demonstrated in Fig. 1. This "pearl chaining" effect is a central feature of both MR and ER phenomena (Jones, 1995). The nature of the structures formed by this interaction has been of considerable interest; a number of optical studies of model MR-fluid systems have investigated the kinetics of chain formation and the eventual coalescence into columns or more complicated network structures (for example, Hayes, 1975; Fermigier and Gast, 1992; Liu *et al.*, 1995).

A lower bound on the response time for the magnetorheological effect is the time taken to form pairs of particles, the flocculation time τ_f. The flocculation time has been estimated from molecular-dynamics simulations and analytical models of ER fluids exposed to a stepwise increase in the field (Hass, 1993); it is predicted to be proportional to the suspending-fluid viscosity and inversely proportional to the square of the dipole moment on each particle. Utilizing the analogy between electrostatics and magnetostatics, electrorheological models can be translated to ones appropriate to magnetorheological phenomena. Following Hass (1993), an expression for the pair-formation time in a MR fluid in the linear regime can be obtained:

$$\tau_f \approx \frac{2\eta_f}{5\mu_0(\beta H)^2}\left[\left(\frac{\pi}{6\phi}\right)^{5/3} - 1\right], \tag{6}$$

where η_f is the viscosity of the suspending fluid and ϕ is the volume fraction of the suspended particles. As the field is increased, the pair-formation time decreases; at sufficiently high fields, the particle moments—and the rate of pair formation—saturate. By modifying Eq. (6) to treat saturated particles, the *minimum* pair-formation time in a MR fluid is $\tau_f(\text{sat}) \approx 18\eta_f[(\pi/6\phi)^{5/3} - 1]/5\mu_0 M_s^2$, or roughly 1 μs with $\mu_0 M_s = 2$ T, $\eta_f = 0.5$ Pa-s, and $\phi = 0.3$.

While no experimental response-time data are available on practical MR fluids, the few reports of transient MR-device performance indicate device response times of order 10 ms (Weiss *et al.*, 1993). It is likely that extrinsic factors such as inductive time constants, eddy-current effects, and hysteresis determine the structural response times of MR fluids. Moreover, a substantial magnetorheological stress cannot be obtained unless the particle network is strained. The rheological response time τ_{rheo} is thus inversely related to the shear-strain rate in steady-shear conditions: $\tau_{\text{rheo}} \sim \gamma_y/\dot{\gamma}$, where γ_y is the yield strain and $\dot{\gamma}$ is the shear-strain rate (Ginder and Ceccio, 1995).

2.3 Steady-Shear Rheology

Of central importance in both torque-transfer and damping applications is the rheological performance of MR fluids. The flows encountered in these applications typically involve shear of the working fluid. Steady-shear rheology is almost universally used to characterize the rheological performance of MR and ER fluids. Concentric-cylinder or parallel-plate rheometers are typically used: The shear stress transmitted through the fluid when it is sheared is measured. Concentric-cylinder rheometers are often preferred, since the variation in shear rate through the sample is small for small fluid gaps; the shear rate in parallel-plate rheometers varies linearly with distance from the axis of rotation (see, e.g., Macosko, 1994). Materials and measurement issues involved in designing and using rheometers for the study of MR fluids have been discussed by Janocha and Rech (1994).

Relatively few measurements of the steady-shear rheology of MR fluids are available in the literature. Parziale and Tilton (1950) measured the flux-density dependence of the shear stress transmitted through a carbonyl-iron–based MR fluid ($\phi \approx 0.5$) used in a parallel-disk MR clutch. Kordonsky and Shulman (1990) reported the rheological, magnetic, and thermal properties of several MR fluids. Lemaire and Bossis (1991) examined magnetite-based MR fluids, finding that the measured shear stresses depended on the roughness of the interface between the MR fluid and the test fixture. Weiss and co-workers (Weiss *et al.*, 1993; Weiss and Duclos, 1994) reported the shear-rate dependence of the shear stress in a proprietary MR fluid exposed to several applied magnetic fields (Fig. 4). The data are well fitted by the Bingham model [Eq. (2)] with a viscosity of order 0.3 Pa-s and a yield stress that increases subquadratically with applied field. A weak reduction of the yield stress with temperature from −50°C to +150 °C was also reported; Weiss *et al.* (1993) suggested that it was entirely consistent with the reduction in particle-volume fraction due to thermal expansion of the suspending fluid. Pinkos *et al.* (1993) inferred the apparent shear stress in a rotary damper incorporating a carbonyl-iron–based MR fluid; shear stresses approaching 50 psi, or over 300 kPa, were reported, although the nature or rate of the shear flow was not disclosed. Inverse MR fluids containing nonmagnetic particles, such as polystyrene microspheres, dispersed in ferrofluids have been studied by Kashevskii *et al.* (1989), Bossis and Lemaire (1991), and Popplewell *et al.* (1995). The latter measurements were found to be in good agreement with the predictions of a linear continuum model of MR fluids (Rosensweig, 1995a).

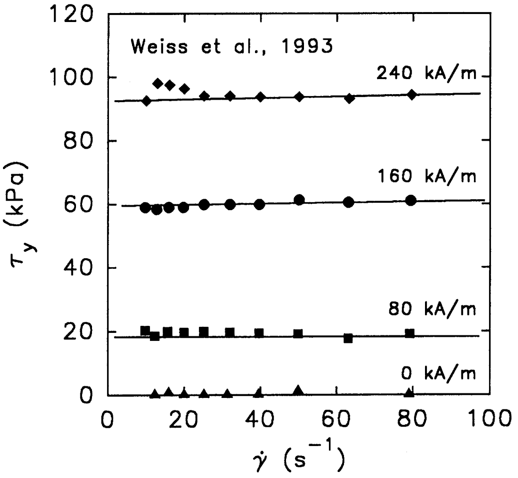

FIG. 4. Dependence of the shear stress of a MR fluid at 25 °C on shear rate and applied field (after Weiss *et al.*, 1993). The solid lines are fits by the Bingham model.

2.4 Oscillating-Shear Rheology

While many applications of MR fluids involve steady or unidirectional flows, some damping applications produce oscillatory flows. Oscillating-shear rheology is useful in defining the viscoelastic properties of complex fluids; in particulate systems, it can reveal qualitatively the nature and range of the interparticle interactions responsible for the viscoelastic behavior. These measurements may also be compared with theoretical predictions, which often are appropriate to small strain amplitudes and frequencies. They typically utilize sinuosoidal strain oscillations $\gamma^*(t) = \gamma_0 e^{i\omega t}$, where $i^2 = -1$, ω is the angular frequency of the oscillations, and t is the time, imposed on the sample; the complex stress transfer $\tau^* = \tau' + i\tau''$ is measured. The shear-storage modulus G', defined as τ'/γ_0, gives a measure of the elastic stiffness of the material. The loss modulus G'', given by τ''/γ_0, reflects viscous effects in the material. Weiss *et al.* (1994) have reported limited oscillating-shear rheological data on a proprietary MR fluid; the linear regime, in which the time dependence of the stress is linearly related to that of the applied strain, was found only for strain amplitudes below about 10^{-2}; the shear-storage moduli in this regime were as high as 2 MPa for applied fields of 160 kA/m. Magnetized MR fluids thus possess stiffnesses approaching those of common viscoelastic solids such as rubber.

2.5 Magnetic Properties

Knowledge of the magnetic properties of MR fluids can greatly assist the design of magnetic circuits for MR devices and may provide experimental tests for magnetostatic models of these materials. The variation of the flux density in a sample with applied field, $\mathbf{B} = \mathbf{B}(\mathbf{H})$ or the so-called *B-H* curve, can be measured using flux integration, vibrating-sample magnetometry, or other techniques. Carlson *et al.* (1994) reported the field dependence of the magnetization $M = B/\mu_0 - H$ for a MR fluid of unknown composition. A *B-H* curve derived from these data is shown in Fig. 5. The magnetic properties of MR fluids can be sensitive probes of the state of particle aggregation: Kordonsky and Shulman (1990) observed that the magnetization of a MR-fluid sample at a given magnetic field is larger when the particles have chained than when they are randomly distributed.

2.6 Magnetostatic Models of Magnetorheology

Models of magnetorheological phenomena are desirable both to satisfy basic scientific curiosity and to assist the development of MR fluids and devices. It was recognized early on that magnetically induced stresses were qualitatively responsible for the magnetorheological effect. More recent work has been inspired in part by considerable modeling and experimental activity on ER fluids. The connection between suspension microstructure and rheological behavior in MR fluids is likely to be very similar to that in ER fluids. Stress transfer in ER fluids occurs qualitatively through the tilting, snapping, and reformation of particle chains (Klingenberg and Zukoski, 1990). Simulations have shown how a field-dependent yield stress can result from these microstructural processes (Bonnecaze and Brady, 1992).

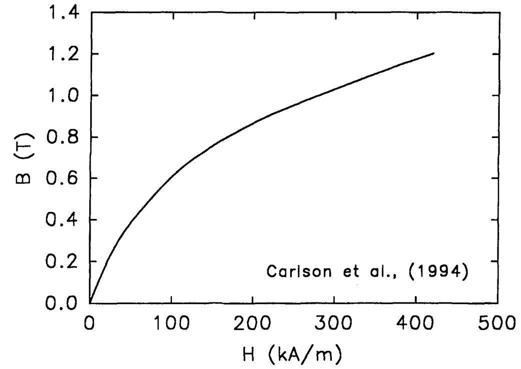

FIG. 5. *B-H* curve inferred from magnetization measurements on a MR fluid (after Carlson *et al.*, 1994).

The many approaches used to model the electrostatic interactions in ER fluids can be directly transferred to model MR fluids so long as one is confined to the linear regime, i.e., for fields low enough not to produce saturation effects in the MR particulates. These models generally involve the calculation of the fields and forces in particle chains or three-dimensionally ordered structures. The restoring force produced by shearing the structure can then be calculated as a function of applied field and the permeability and volume fraction of the magnetizable particulates. These linear approaches predict that the shear stresses, yield stresses, and shear moduli are proportional to $\phi\mu_0H^2$, where ϕ is the particle-volume fraction and H is the applied field.

The stresses measured in practical MR fluids increase with field more slowly than H^2. This serious discrepancy reflects the important roles of magnetic nonlinearity and saturation of the particles used in these fluids. These effects have been explicitly incorporated in a recent model of magnetorheology (Ginder and Davis, 1994). A nonlinear constitutive *B-H* relation was assumed, and the field distribution in an infinite linear chain of particles in a specified flux density was calculated using finite-element analysis (FEA) numerical techniques. The chain tension was then calculated from the Maxwell stress at the interface between particles and was used to infer the magnetically induced shear stress as a function of shear strain. The yield stresses predicted by this model for a saturation magnetization $\mu_0M_s = 2$ T are shown in Fig. 6(a); they increase subquadratically with flux density and saturate at as

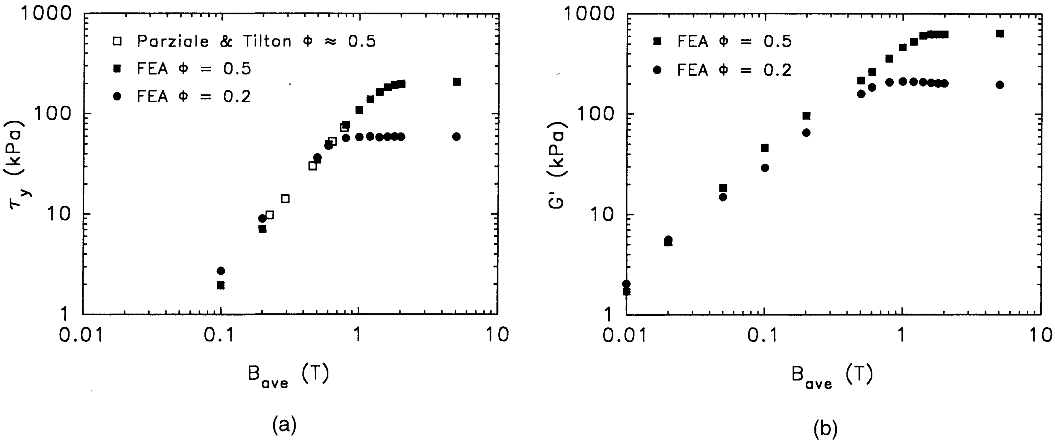

(a) (b)

FIG. 6. Predictions of a finite-element analysis model for a MR fluid with $\mu_0 M_s$ = 2 T: (a) flux-density dependence of the MR shear stress for volume fractions ϕ = 0.2 (●) and 0.5 (■). The shear stresses measured in a MR fluid with $\phi \approx$ 0.5 (□) (Parziale and Tilton, 1950) are also shown. (b) Variation of the shear modulus with flux density (after Ginder and Davis, 1994).

much as 210 kPa for $\phi = 0.5$. The predicted yield stresses compare quite favorably with experimental data from Parziale and Tilton (1950) on carbonyl-iron MR fluids with $\phi \approx 0.5$, also shown. While the shear moduli predicted by this model are as large as 600 kPa for $\phi = 0.5$ at saturation [Fig. 6(b)], they are far exceeded by the shear moduli measured on a proprietary MR fluid (Weiss *et al.*, 1994). This discrepancy is likely due to the fact that the model treats only the stresses directly resulting from magnetic interactions and not those associated with other interparticle-contact interactions—van der Waals forces, steric forces, and so on.

For applied fields too low for complete particle saturation, the consequences of saturation are still manifest in the magnetic saturation of the polar regions of each particle (von Guggenberg *et al.*, 1986; Jones and Saha, 1990; Ginder and Davis, 1994). The magnetic flux is channeled through the particles and is highest at the contact regions, causing this local saturation. A contour plot of the magnetization obtained by a FEA solution for a chain of particles (Fig. 7) reveals that the polar regions are indeed the first to saturate. Local saturation limits the increase of the field in the gaps between particles and thus reduces the quadratic dependence of the stresses on applied field or flux density. A simple analytical model (Ginder *et al.*, 1996) predicts that the yield stress in this regime is given by

$$\tau_y \approx \sqrt{6}\, \phi \mu_0 M_s^{1/2} H^{3/2}, \tag{7}$$

while the shear-storage modulus is given by

$$G' \approx 3\phi \mu_0 M_s H. \tag{8}$$

For high applied fields, the particles are completely saturated and so can be treated rigorously as dipoles. The results for the shear stress in simple dipole chains, in which the stresses scale quadratically with the dipole moment, can then be applied to calculate the yield stress at saturation (the maximum value of the yield stress). Results of FEA calculations for $\phi = 0.2$ and 0.5 (Fig. 8) also exhibit quadratic scaling with saturation magnetization. The FEA results are in agreement with the maximum (albeit not saturation-level) yield stresses reported for Lord and BASF MR fluids, as indicated in Fig. 8. The quadratic dependence of the maximum shear stress on the saturation magnetization of the particles underscores the desirability of MR particulates synthesized from elements or alloys having high M_s values. Iron–cobalt alloys having the largest known saturation magnetization of any material, $\mu_0 M_s \sim 2.4$ T, have recently been incorporated in a high-strength MR fluid (Carlson and Weiss, 1995).

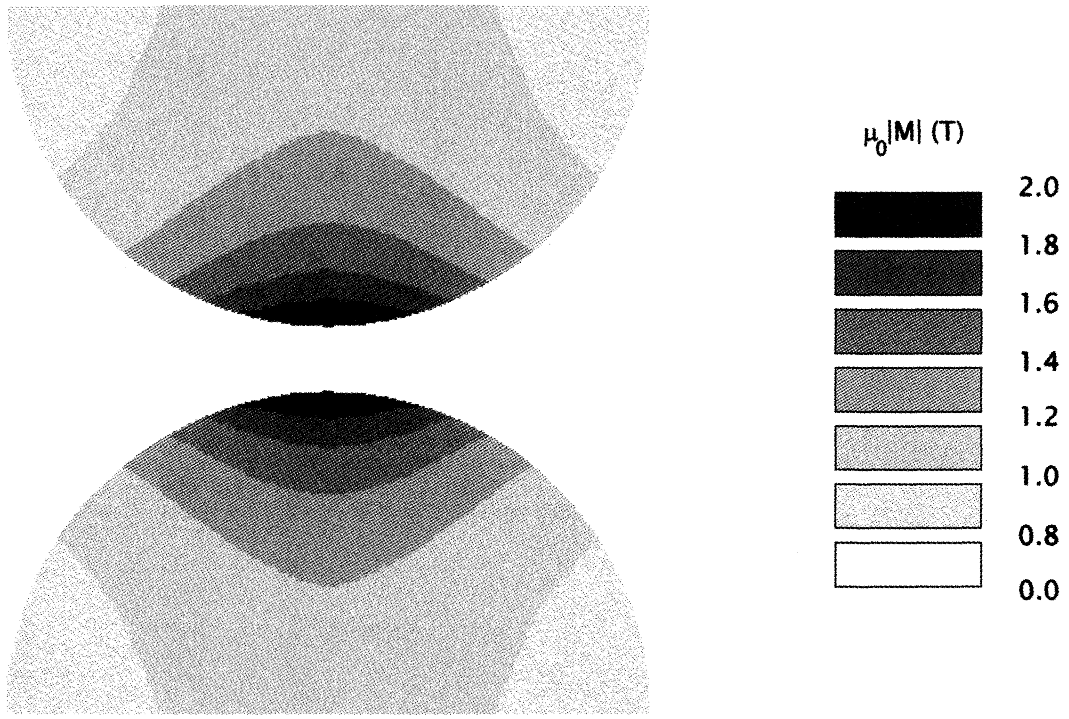

FIG. 7. Contour plot of the magnitude of the magnetization in a pair of particles embedded in an infinite chain modeled by finite-element analysis, showing the saturation of the magnetization near the contact points. The particles are assumed to have a saturation magnetization $\mu_0 M_s = 2$ T and are exposed to a flux density 0.4 T.

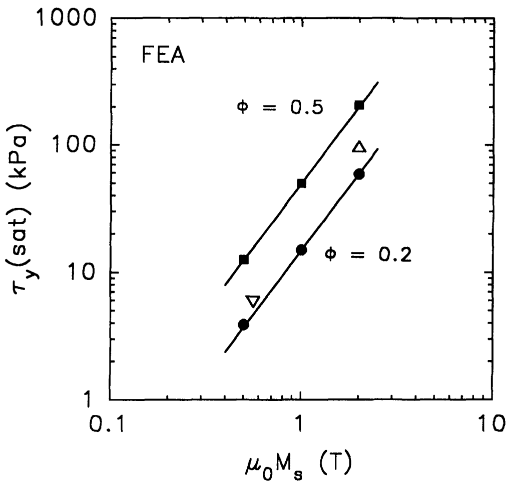

FIG. 8. Quadratic variation of the yield stress at saturation on the saturation magnetization of the particles for volume fractions $\phi = 0.2$ (●) and 0.5 (■), calculated using a FEA model (after Ginder and Davis, 1994). The largest shear stresses reported for MR fluids from Lord Corporation (△) (Weiss *et al.*, 1993) and BASF (▽) (Kormann *et al.*, 1994) are also shown.

3. MAGNETIC POWDERS

3.1 Composition

The magnetostatics of MR fluids are unaltered if the nonmagnetic suspending fluids are replaced by air to obtain dry magnetic powders. The highest-strength magnetic powders would thus involve the particulate materials with the highest saturation magnetization, iron or some of its alloys. Unfortunately, the major application of magnetic powders, in clutches and rotary brakes, is often associated with elevated temperatures that promote the eventual failure of these particles due to oxidation. Magnetic stainless steels such as 400-level steels, composed mainly of iron and chromium, are therefore used. Typical particle sizes in magnetic powders are from 10 to 100 μm; both spherically and irregularly shaped particles are used. Nonmagnetic powdered materials such as graphite, zinc oxide, or silicon dioxide may be added to act as lubricants or dispersants (Ramakrishnan and Pillai, 1980).

Magnetic powders are widely used in image development in xerography (Jones *et al.*, 1987), in which rotating drums housing arrays of permanent magnets are used to transport magnetizable particles to the photoreceptor drum on which the xerographic image is explosed. One-component xerography uses 10-μm polymeric toner particles loaded with both ferrite and pigment to form the image directly, while two-component xerography uses 100-μm ferrite or steel particles as a reusable transport medium to carry nonmagnetic 10-μm pigment particles to the photoreceptor drum.

3.2 Magnetic-Powder Rheology

The mechanics of noncohesive, isotropic powders is much more widely studied and better understood than that of magnetic powders, which are highly cohesive, anisotropic materials. The most extensive measurements of magnetic-powder rheology have been performed on magnetostabilized fluidized beds, in which a collection of magnetizable particles subjected to a stream of flowing gas or liquid is held in place by an imposed magnetic field (Rosensweig, 1995b). These beds were of interest for industrial chemical processes such as catalysis, molecular separation, and heat recovery. Several studies of the dependence of the yield stress on applied field and flow velocity have appeared (for example, Lee, 1983); as it happens, the yield stresses necessary for bed stabilization are several orders of magnitude smaller than those of interest in motion-control applications. Torque-transfer geometries have been used to characterize magnetic powders; a magnetic-powder disk clutch was used to infer the rheology of iron powders (Ramakrishnan and Pillai, 1980). The magnetically induced shear stresses were found to increase only weakly with increasing particle size and shear rate. Jones *et al.* (1987) have extended theories of powder mechanics to include magnetic interactions between particles. Such theories, first developed by Coulomb, use the conditions of static mechanical equilibrium to relate the shear and normal stresses in a powder. The application of magnetically induced stresses increases the effective normal stress and thus increases the shear stress and, in particular, its yield value.

4. MAGNETORHEOLOGICAL ELASTOMERS

4.1 Composition

Composites of magnetic particles in polymeric host media have been produced for some time, although usually for structural studies or dielectric applications. For example, the structure of a magnetostabilized bed of particles was investigated by polymerizing the suspending fluid, thereby "freezing" the particulate structure and permitting optical microscopy and magnetic measurements on the solidified specimens (Rosensweig *et al.*, 1981). Magnetic metallic particles have been aligned by magnetic fields and subsequently fixed by curing in silicone elastomers to produce transparent sheets whose conductivity along the field direction is extremely sensitive to pressure; these sheets can be used in touch-screen panels (Jin *et al.*, 1992). Highly anisotropic metallic magnetic particles have been dispersed and aligned in rigid epoxies to produce highly anisotropic dielectric materials for microwave frequencies (Behroozi *et al.*, 1990).

It has only recently been appreciated that the viscoelastic properties of magnetic-particle polymer composites can be varied by an applied field. A group at Toyota Central Research and Development has dispersed iron particles in liquid-silicone precursors, then cured the mixture to produce MR elastomers (Shiga *et al.*, 1992): viscoelastic solids whose storage and loss moduli are nonzero even in the absence of a field, but can be increased substantially, rapidly, and continuously by the application of a field. Other matrix materials may include ethylene–propylene, butadiene, and polyisoprene rubbers, as well as polyvinyl alcohol, polyacrylamide, and polystyrene gels. The preferred particle concentration ranges from 0.05 to 0.6 and can be varied spatially through the composite. After curing, these solids can be shaped or machined to the desired geometry and size.

4.2 Field-Dependent Viscoelastic Response

Shiga *et al.* (1992) measured the dynamic viscoelastic properties of several MR elastomers in an oscillating parallel-plate rheometer. One sample contained 150-μm electrolytic-iron particles dispersed randomly at $\phi \approx 0.27$ in room-temperature–curing silicone rubber. The shear-storage modulus of this material increased by 50% upon the application of a 320-Oe magnetic field. In another sample, 50-μm electrolytic iron was dispersed at $\phi \approx 0.2$ in an elevated-temperature–curable silicone-rubber precursor. A permanent magnet was used to orient the iron particles while the elastomer was cured at elevated temperature. The resulting anisotropic composite exhibited a shear-storage modulus of 7.8 kPa; the storage modulus increased to 21.5 kPa upon the application of a 740-Oe field.

5. APPLICATIONS OF MR MATERIALS

5.1 Magnetic-Circuit Considerations

Of critical importance in applications of MR materials—ferrofluids, MR fluids, magnetic powders, or MR elastomers—is the efficient delivery of the required magnetic field to the material. The production and transmission of the field are accomplished by a magnetic circuit that consists of a source of magnetomotive force, either a wire coil or a permanent magnet, and a flux path usually constructed of highly permeable materials like iron or steel (see MAGNETOSTATICS). The optimum MR-device design is that which will deliver the necessary field while consuming the least electrical power and weighing as little as possible, sometimes conflicting goals. Detailed magnetic-circuit design is greatly aided by numerical solutions of the magnetostatic field and flux density in the circuit that include the nonlinear B-H relationships of the circuit components. Several commercial numerical magnetostatic software programs based on finite-element and boundary-element techniques in two- and three-dimensional geometries are available for this purpose.

The approximate behavior of simple magnetic circuits can be understood by applying the fundamental laws of magnetostatics. Gauss's law, $\nabla \cdot \mathbf{B} = 0$, leads to the conservation of flux in the magnetic circuit. Ampére's law relates the line integral of the magnetic field along any contour (and in particular contours embedded in the magnetic circuit) to the current enclosed by it: $\nabla \times \mathbf{H} = \mathbf{J}$ or $\oint \mathbf{H} \cdot d\mathbf{l} = I_{\text{tot}}$, where J is the current density, I_{tot} is the total current contained by the contour, and $d\mathbf{l}$ is the differential line element on the contour. MR devices typically utilize electromagnets to allow control of the field in the MR material; in most cases the electromagnet is constructed around a toroidal coil of N turns such that $I_{\text{tot}} = NI$ where I is the current in the coil. If the coil has an average radius R_{coil} and a cross-sectional area A_{coil}, the electrical power consumed is $P \sim 8\rho R_{\text{coil}}(NI)^2/A_{\text{coil}}$, where ρ is the electrical resistivity of the wire. In this fashion, the electrical-power consumption can be related to the device shape and size as well as the fields in it. In typical magnetic-circuit designs, the constitutive relations between $\mathbf{B}$ and $\mathbf{H}$ are measured or provided for each material used in the magnetic circuit. The designer adjusts the magnetic-circuit geometry to obtain the desired device performance, power consumption, and packaging attributes.

5.2 Torque-transfer Applications

MR fluids and magnetic powders can be used to control the transfer of torque in clutches and rotary brakes. Magnetic-powder clutches and brakes are commercially available from a number of suppliers, including MagPower, Placid, Electroid, and T. B. Wood's Sons in the United States and Ogura and Mitsubishi Electric in Japan. These devices are used widely in manufacturing to control speed, torque, and tension, as well as in motor-testing applications. Magnetic-powder clutches are even used in automobiles: The main clutch on Subaru Justy vehicles equipped with continuously variable transmissions is a high–torque-capacity magnetic-powder clutch (Sakai, 1988). This application provides an example of the controllability afforded by the magnetorheological effect: The current to the clutch is varied continuously during engine acceleration and deceleration in order to reduce unwanted power-train transients. MR-fluid brakes are now being produced by Lord Corporation for use as variable-resistance devices on exercise machines. The smoother operating characteristics and improved heat-transfer capabilities of MR fluids are potential advantages over magnetic powders in these applications.

MR fluids and magnetic powders have

been utilized in two basic clutch or brake geometries, parallel disks and concentric cylinders, as shown schematically in Fig. 9. The torque-transfer characteristics of these geometries can be estimated if the MR fluids are assumed to obey a Bingham law with viscosity η and yield stress τ_y [Eq. (2)]. For a disk of radius R separated from each of the two opposing faces of its housing by a gap of width h, the total torque T assuming laminar flow is given by

$$T = \pi\eta R^3 \frac{R}{h} \Omega + \frac{4\pi}{3} R^3 \tau_y, \tag{9}$$

where Ω is the relative angular frequency of the disk and the housing. For concentric cylinders with a length L, inner-cylinder radius R, and gap h such that $h \ll R$, the torque is

$$T = 2\pi\eta L R^2 \frac{R}{h} \Omega + 2\pi L R^2 \tau_y. \tag{10}$$

In both of the above expressions, the first term on the right represents the contribution to the torque by Couette flow of a viscous fluid and is present at all times. The second term is due to the field-induced MR shear stress and is therefore controllable by an applied magnetic field. The rate of mechanical-power dissipation for both geometries is given by $P_{\mathrm{diss}} = \Omega T$.

5.3 Damping Applications

Another important potential application of MR materials is in vibration damping. Shock absorbers and mounts are used in many automotive, aerospace, and industrial applications to reduce the vibrations transmitted from internal or external sources to operators, passengers, or sensitive equipment. Materials like MR fluids allow the damping force to be controlled according to external inputs from sensors, enabling significant improvements in damper performance. Researchers at TRW (Pinkos *et al.*, 1993), Lord Corporation (Carlson *et al.*, 1994, 1995), and Byelocorp Scientific (Kordonsky *et al.*, 1994) have utilized MR fluids in a variety of damping applications. A field-induced increase in damping force of as much as 3 kN, comparable to that required in many passenger-car automotive applications, was obtained in a shock-absorber prototype (Weiss *et al.*, 1993). Seat mounts for on- and off-road vehicles, engine mounts, and machinery mounts are other potential applications. A schematic cross-sectional view of a MR-fluid–based damper is shown in Fig. 10.

Like their passive counterparts, most MR-fluid dampers utilize fluid flow through an orifice or channel to produce a pressure drop ΔP and thus to generate a damping force. A MR-fluid valve is an electrically controllable orifice: The pressure drop across the valve is controllable by changes in the magnetic field

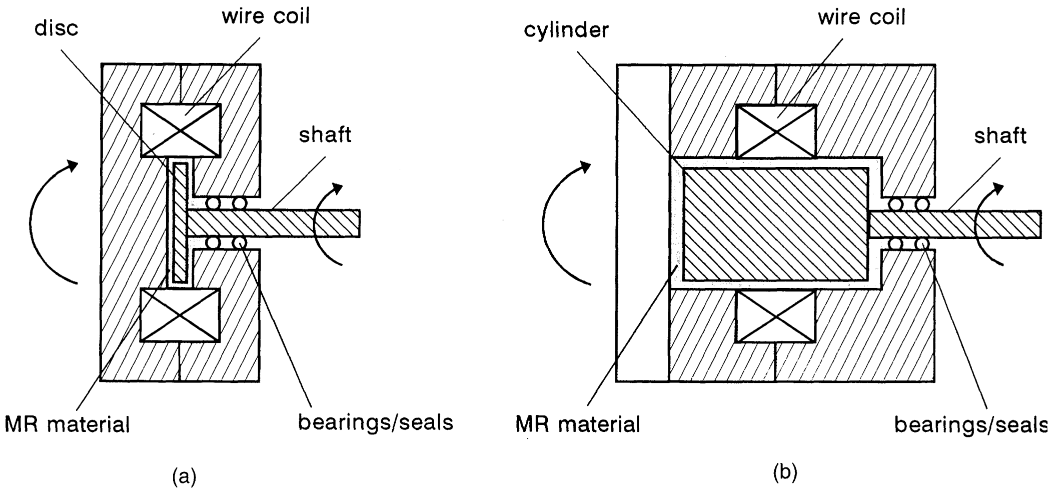

FIG. 9. Schematic cross-sectional view of (a) disk and (b) concentric-cylinder clutches.

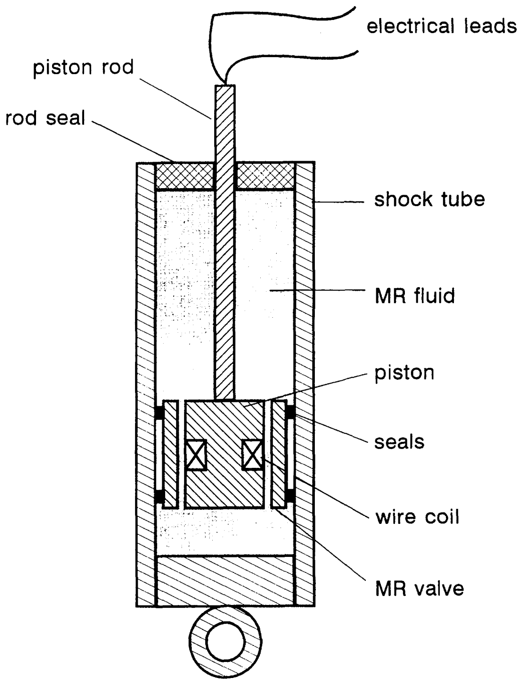

FIG. 10. Schematic cross-sectional view of a MR-fluid–based damper (after Carlson *et al.*, 1994).

in the channel. While ΔP depends on the detailed nature of the flow of solidified MR material in the channel, it is approximately given by

$$\Delta P \approx 12\eta \frac{Q}{wh^2}\frac{L}{h} + c\frac{L}{h}\tau_y, \tag{11}$$

where L, h, and w are the channel length, thickness, and width, Q is the volumetric-flow rate, and c is a numerical constant between 2 and 3. The damping force F generated in the channel is then the product of the pressure drop and the channel cross-sectional area: $F = hw\Delta P$. The first term on the right in Eq. (11) represents the pressure drop due to the (assumed) laminar Pouseille flow of the viscous MR fluid, which is proportional to the flow rate and is nominally independent of field. The second term on the right is the contribution to the pressure drop from the solidified MR fluid; it is essentially independent of flow rate but is strongly dependent on the magnetic field.

MR elastomers could potentially be used in damping and vibration isolation components such as engine mounts, although such applications have not yet appeared in the literature. Advantages of these materials include their solid-state character: No seals are required to contain fluids, and the suspended magnetizable particles do not sediment out over time.

5.4 Process Applications

Magnetorheological fluids may also find use in manufacturing and process applications. Of considerable recent interest is the use of MR-fluid–based polishing compounds in the deterministic finishing of optical components, an application being explored by Byelocorp Scientific, Inc. (Holton, 1993). Nonmagnetic abrasive particles are dispersed in an aqueous MR fluid; by applying a magnetic field, the abrasive grit is driven to the interface between the MR fluid and the optical component. The rate and spatial extent of the abrasive action can be controlled in real time by an electromagnet. Any nonferrous material may be polished by this technique.

5.5 Outlook

With the first MR-fluid devices now being commercialized, this is an exciting era in the field of magnetorheology. While the relative maturity of ferrofluid and magnetic-powder technology provides an "existence proof" for the viability of MR-fluid devices, the degree of acceptance and use of this technology in automotive and other industries will depend on a number of technical and economic factors. The not insubstantial mass and power consumption of many MR devices is an issue, although it should be noted that conventional electromechanical devices possess similar characteristics. The most efficient MR clutch designs embed wire coils inside the rotating clutch member, requiring slip rings that add cost and can fail over time. The minimum response time of MR devices has not been established. While inductive, eddy-current, and hysteretic effects conspire to slow device response, they can be combated with laminated magnetic-circuit materials and transient overvoltages. To take full advantage of the speed and controllability of the MR-device response, sensors are required as inputs to control hardware and software; while providing substantially improved performance, these sensors also add to the system cost. It might be possible to eliminate separate sensors and their associated expense if the MR material itself could be used as a strain or velocity sensor; this increase in material functionality might be achievable by using inductive or other effects. The abrasive effects of the magnetic particulates in MR fluids require much more attention than they have received; friction and wear may be particularly problematic in fluid seals. The toxicity and recyclability of MR fluids must be addressed.

Improvements in our understanding of the physics and chemistry of MR fluids should assist the development of new MR fluids. The stability of the fluids against sedimentation and centrifugation remains a central problem. The utilization of submicron magnetic particles produced by a variety of processes may improve stability. New surfactants may be identified to aid the stabilization and redispersal of the particles. Soluble polymers could be employed for fluid stabilization via the formation of a gel-like thixotropic network or by weakly flocculating the magnetizable particles. The nature of the structural deformations induced by applied stresses in MR fluids requires clarification. The assumption that isolated chains deform uniformly under an applied strain is undoubtedly a gross simplification of the situation in real, polydisperse systems. Structure deformation by fracture may be as relevant in MR fluids as it appears to be in magnetic powders. Models for the MR shear stress that account for not only the magnetostatic interactions between particles but also other colloidal-interaction forces are needed. The effects of magnetic hysteresis on the magnetic and rheological performance of MR fluids have not been investigated. The development of production devices and the emergence of new applications are powerful motivations for the continued pursuit of physical understanding of magnetorheological materials.

GLOSSARY

Magnetic Circuit: A circuit through which magnetic flux can be delivered. Usually consists of a source of magnetomotive force (either a wire coil or a permanent magnet) and a path containing iron or steel to transmit the flux.

Relative Permeability: The ratio of magnetic induction to magnetic field in a material, normalized by μ_0, the permeability of free space.

Saturation Magnetization: The maximum magnetization, or magnetic moment per unit volume, achievable in a material.

Thixotropy: A phenomenon in which a material exhibits time-dependent rheological properties.

Yield Stress: The stress at which a material deforms plastically or flows.

Works Cited

Bean, C. P., Jacobs, I. S. (1960), *J. Appl. Phys.* **31,** 1228–1230.

Behroozi, F., Orman, M., Reese, R., Stockton, W., Calvert, J., Rachford, F., Schoen, P. (1990), *J. Appl. Phys.* **68,** 3688–3693.

Berkovsky, B. M., Medvedev, V. F., Krakov, M. S. (1993), *Magnetic Fluids,* Oxford: Oxford Univ. Press.

Blunden, S. F. (1951), *Engineer* **191** (4961), 244–246.

Bonnecaze, R. T., Brady, J. F. (1992), *J. Rheol.* **36,** 73–115.

Bossis, G., Lemaire, E. (1991), *J. Rheol.* **35,** 1345–1354.

Carlson, J. D., Chrzan, M. J., James, F. O. (1994), U.S. Patent 5284330.

Carlson, J. D., Weiss, K. D. (1995), U.S. Patent 5382373.

Carlson, J. D., Chrzan, M. J., James, F. O. (1995), U.S. Patent 5398917.

Fermigier, M., Gast, A. P. (1992), *J. Colloid. Int. Sci.* **154,** 522–539.

Ginder, J. M., Davis, L. C. (1994), *Appl. Phys. Lett.* **65,** 3410–3412.

Ginder, J. M., Ceccio, S. L. (1995), *J. Rheol.* **39,** 211–234.

Ginder, J. M., Davis, L. C., Elie, L. D. (1996), in: W. Bullough (Ed.), *Proceedings of the 5th International Conference on Electrorheological Fluids and Magnetorheological Suspensions,* Singapore: World Scientific.

Grebe, O. (1952), *Elektrotech. Zeits.* **9,** 281–284.

von Guggenberg, P. A., Porter, A. J. II, Melcher, J. R. (1986), *IEEE Trans. Magn.* **MAG-22,** 614–619.

Hartsock, D. L., Novak, R. F., Chaundy, G. J. (1991), *J. Rheol.* **35,** 1305–1326.

Hass, K. C. (1993), *Phys. Rev. E* **47,** 3362–3373.

Hayes, C. F. (1975), *J. Colloid. Int. Sci.* **52,** 239–243.

Holton, W. C. (1993), *Photonics Spectra* (April), 29.

Janocha, H., Rech, B. (1994), *Rheology* (December), 199–203.

Jin, S., Tiefel, T. H., Wolfe, R., Sherwood, R. C., Mottine, J. J., Jr. (1992), *Science* **255,** 446–448.

Jones, T. B. (1995), *Electromechanics of Particles,* New York: Cambridge Univ. Press.

Jones, T. B., Saha, B. (1990), *J. Appl. Phys.* **68,** 404–410.

Jones, T. B., Whittaker, G. L., Sulenski, T. J. (1987), *Powder Technol.* **49,** 149–164.

Kashevskii, B. E., Kordonsky, V. I., Prokhorov, I. V. (1989), *Magnetohydrodynamics* **3,** 368–371.

Klingenberg, D. J., Zukoski, C. F. (1990), *Langmuir* **6,** 15–24.

Kordonsky, V. I., Shulman, Z. P. (1990), in: J. D. Carlson, A. F. Sprecher, H. Conrad (Eds.), *Proceedings of the Second International Conference on ER Fluids,* Lancaster, PA: Technomic, pp. 437–444.

Kordonsky, W. I., Gorodkin, S. R., Kolomentsev, A. V., Kuzmin, V. A., Luk'ianovich, A. V., Protasevich, N. A., Prokhorov, I. V., Shulman, Z. P. (1994), U.S. Patent 5353839.

Kormann, C., Laun, M., Klett, G. (1994), in: H. Borgmann, K. Lenz (Eds.), *Proceedings of the Fourth International Conference on New Actuators,* Bremen, Germany, 15–17 June, Bremen, Germany: AXON Technologie, pp. 271–274.

Lee, W.-K. (1983), *AIChE Symp. Ser.* **222,** 79.

Lemaire, E., Bossis, G. (1991), *J. Phys. D: Appl. Phys.* **24,** 1473.

Liu, J., Lawrence, E. M., Wu, A., Ivey, M. L., Flores, G. A., Javier, K., Bibette, J., Richard, J. (1995), *Phys. Rev. Lett.* **74,** 2828–2831.

Macosko, C. W. (1994), *Rheology,* New York: VCH.

McTague, J. P. (1969), *J. Chem. Phys.* **51,** 133–136.

Parziale, A. J., Tilton, P. D. (1950), *AIEE Trans.* **69,** 150–157.

Pinkos, A., Shtarkman, E., Fitzgerald, T. (1993), SAE International Congress and Exhibition, Detroit, MI, 1–5 March, SAE Technical Paper 930268.

Popplewell, J., Rosensweig, R. E., Siller, J. K. (1995), *J. Magn. Magn. Mater.,* in press.

Rabinow, J. (1948), *AIEE Trans.* **67,** 1308–1315.

Rabinow, J. (1949), SAE National Transportation Meeting, Cleveland, OH, 28–30 March, SAE Technical Paper 49-316.

Rabinow, J. (1951), U.S. Patent 2575360.

Raj, K., Moskowitz, R. (1990), *J. Magn. Magn. Mater.* **85,** 233–245.

Ramakrishnan, S., Pillai, K. P. P. (1980), *IEE Proc.* **127,** 81–88.

Rosensweig, R. E., Jerauld, G. R., Zahn, M. (1981), in: O. Brulin, R. K. T. Hsieh (Eds.), *Continuum Models of Discrete Systems,* Vol. 4, Amsterdam: North-Holland, pp. 137–144.

Rosensweig, R. E. (1985), *Ferrohydrodynamics,* Cambridge: Cambridge Univ. Press.

Rosensweig, R. E. (1995a), *J. Rheol.* **39,** 179–192.

Rosensweig, R. E. (1995b), *J. Electrostat.* **34,** 163–187.

Sakai, Y. (1988), SAE International Congress and Exhibition, Detroit, MI, 29 February–4 March, SAE Technical Paper 880481.

Shiga, M., Hirose, M., Okada, K. (1992), Japanese Patent laid-open No. 4-266970.

Shliomis, M. I. (1972), *Soviet Phys. JETP* **34,** 1291–1294.

Shtarkman, E. (1992), U.S. Patent 5167850.

Shtarkman, E. M., Starkovich, J. A., Davison, W. W., Peng, H.-H., Fitzgerald, T. J. (1994), U.S. Patent 5354488.

Vorob'yeva, T. M. (1965), *Electromagnetic Clutches and Couplings,* Oxford: Pergamon.

Weiss, K. D., Duclos, T. G., Carlson, J. D., Chrzan, M. J., Margida, A. J. (1993), SAE International Off-Highway and Powerplant Congress and Exhibition, Milwaukee, WI, 13–15 September, SAE Technical Paper 932451.

Weiss, K. D., Duclos, T. G. (1994), in: R. Tao, G. D. Roy (Eds.), *Electrorheological Fluids: Mechanisms, Properties, Technology, and Applications,* Proceedings of the Fourth International Conference on ER Fluids, Singapore: World Scientific, pp. 43–59.

Weiss, K. D., Carlson, J. D., Nixon, D. A. (1994), in: C. Rogers, G. Wallace (Eds.), *Proceedings of the Second International Conference on Intelligent Materials,* Lancaster, PA: Technomic, pp.' 1300–1307.

Winslow, W. M. (1947), U.S. Patent 2417850.

Winslow, W. M. (1949), *J. Appl. Phys.* **20,** 1137–1140.

Winslow, W. M. (1950), U.S. Patent 2661825.

Further Reading

Colloidal Interactions and Stability

Russel, W. B., Saville, D. A., Schowalter, W. R. (1989), *Colloidal Dispersions,* Cambridge: Cambridge Univ. Press.

Rheological Techniques

Macosko, C. W. (1994), *Rheology,* New York: VCH.

Electrorheological Fluids

Havelka, K. O., Filisko, F. E. (Eds.) (1995), *Progress in Electrorheology: Science and Technology of Electrorheological Materials,* Proceedings of the American Chemical Society Symposium on Electrorheological Materials and Fluids, New York: Plenum.

Tao, R., Roy, G. D. (Eds.) (1994), *Electrorheological Fluids: Mechanisms, Properties, Technology, and Applications,* Proceedings of the Fourth International Conference on ER Fluids, Singapore: World Scientific Press.

U.S. Department of Energy Report DOE/ER/30172 (1993), *Electrorheological (ER) Fluids: A Research Needs Assessment,* Washington, DC: U.S. Department of Energy.

Magnetism and Magnetic Materials

Bozorth, R. M. (1951), *Ferromagnetism,* Toronto: van Nostrand.

Ferrofluids

Berkovsky, B. M., Medvedev, V. F., Krakov, M. S. (1993), *Magnetic Fluids,* Oxford: Oxford Univ. Press.

Rosensweig, R. E. (1985), *Ferrohydrodynamics,* Cambridge: Cambridge Univ. Press.

ROBOTICS

John J. Craig, *Adept Technology, Inc., San Jose, California, U.S.A.*

Introduction 505
1. Description of Position and Orientation 506
2. Forward Kinematics of Manipulators 507
3. Inverse Kinematics of Manipulators 509
4. Velocities, Static Forces, Singularities 510
5. Dynamics 511
6. Trajectory Generation 511
7. Manipulator Design and Sensors 512
8. Position Control 512
9. Force Control 514
10. Computer Vision 514
11. Locomotion 515
12. Programming Robots 515
13. Off-Line Programming and Simulation 515
Glossary 516
Works Cited 517
Further Reading 517

INTRODUCTION

The study of robotics concerns itself with the desire to synthesize some aspects of human function by the use of mechanisms, sensors, actuators, and computers. Obviously, this is a huge undertaking that seems certain to require a multitude of ideas from various "classical" fields. Various definitions have been used to describe the term *robot*—for example, the following from the Robot Association of America:

> A reprogrammable, multi-functional manipulator designed to move material, parts, tools, or special devices through various programmed motions for the performance of diverse tasks.

Perhaps the broadest definition covers any computer-controlled programmable machine.

The history of industrial automation is characterized by periods of rapid change in popular methods. Use of the *industrial robot* beginning in the 1960s, along with computer-aided design (CAD) systems and computer-aided manufacturing (CAM) systems, characterizes the latest trends in the automation of the manufacturing process. These technologies are leading industrial automation through another transition, the scope of which is still unknown.

Much of the present use of industrial robots is concentrated in rather simple, repetitive tasks that tend not to require high precision. In the 1980s and early 1990s relatively simple tasks like machine tending, material transfer, painting, and welding have been economically viable for robotization. Manufacturing market analysts predict that in the 1990s and into the next century industrial robots will become increasingly viable in applications that require more precision and sensory sophistication such as assembly tasks.

The predicted increase in the capabilities of industrial robots will cause a shift in which kinds of industries employ them. The automotive industry, where robots have been economically justified since the 1970s, will continue to be the leading user in the United States. However, the major growth of the U.S. robot population will occur in nonautomotive industries. Of lesser market size, but of great technological interest, are uses of robots in outer space and under the sea.

This article focuses on the mechanics and control of the most important form of the robot, the *mechanical manipulator.* Exactly what constitutes an industrial robot is sometimes debated. The distinction lies somewhere in the sophistication of the program-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

mability of the device—if a mechanical device can be programmed to perform a wide variety of applications, it is probably an industrial robot. Machines that are for the most part relegated to one class of task are considered *fixed automation.* For the purposes of this article, the distinctions need not be debated as most topics are of a basic nature that applies to a wide variety of programmable machines.

By and large, the study of the mechanics and control of manipulators is not a new science, but merely a collection of topics taken from classical fields. Mechanical engineering contributes methodologies for the study of machines in static and dynamic situations. Mathematics supplies tools for describing spatial motions and other attributes of manipulators. Control theory provides tools for designing and evaluating algorithms to realize desired motions or force application. Electrical engineering techniques are brought to bear in the design of sensors and interfaces for industrial robots, and computer science contributes a basis for programming these devices to perform a desired task.

1. DESCRIPTION OF POSITION AND ORIENTATION

In the study of robotics we are constantly concerned with the location of objects in three-dimensional space. These objects are the links of the manipulator, the parts and tools with which it deals, and other objects in the manipulator's environment. At a crude but important level, these objects are described by just two attributes: their position and their orientation. Naturally, one topic of immediate interest is the manner in which we represent these quantities and manipulate them mathematically.

In order to describe the position and orientation of a body in space we will always attach a coordinate system, or *frame,* rigidly to the object. We then proceed to describe the position and orientation of this frame with respect to some reference coordinate system (see Fig. 1).

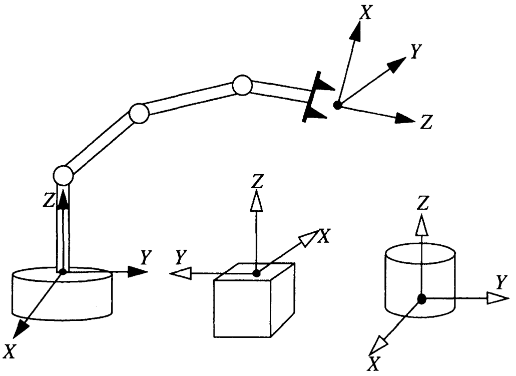

FIG. 1. Coordinate systems or "frames" are attached to the manipulator and objects in the environment.

Since any frame can serve as a reference system within which to express the position and orientation of a body, we often think of *transforming* or *changing the description of* these attributes of a body from one frame to another. Study of robotics inevitably requires conventions and methodologies for dealing with the description of position and orientation, and the mathematics of manipulating these quantities with respect to various coordinate systems.

We use the notation ^{A}P to represent a position P represented in a frame of reference A. This position is represented with a three-element vector whose individual elements are given subscripts x, y, and z:

$$^{A}P = \begin{bmatrix} p_x \\ p_y \\ p_z \end{bmatrix}. \qquad (1)$$

The leading superscript is used to indicate the frame of reference. A common problem is, given a point represented in one frame, to compute its value in a second reference frame:

$$^{A}P = {}^{A}_{B}R\,{}^{B}P + {}^{A}P_{BORG}. \qquad (2)$$

Equation (2) describes a general transformation mapping of a vector from its description in one frame to a description in a second frame. The symbol ${}^{A}_{B}R$ represents a *rotation matix* whose columns are the unit vectors of frame B written in frame A. The vector $^{A}P_{BORG}$ locates the origin of the reference system B written in the system A. Equation (2) is sometimes represented by the simpler notation

$$^{A}P = {}^{A}_{B}T\,{}^{B}P, \qquad (3)$$

where ${}^{A}_{B}T$ is called a *homogeneous transform* and has the structure

$$\left[\begin{array}{c} {}^{A}P \\ \hline 1 \end{array}\right] = \left[\begin{array}{c|c} {}^{A}_{B}R & {}^{A}P_{BORG} \\ \hline 0\ \ 0\ \ 0 & 1 \end{array}\right] \left[\begin{array}{c} {}^{B}P \\ \hline 1 \end{array}\right]. \qquad (4)$$

That is, a "1" is added as the last element of the 4 × 1 vectors, and a row "[0 0 0 1]" is added as the last row of the 4 × 4 matrix. This construction allows the manipulation of position and orientation to be cast in the form of a matrix operator.

2. FORWARD KINEMATICS OF MANIPULATORS

Kinematics is the science of motion that treats motion without regard to the forces that cause it. Within the science of kinematics one studies the position, velocity, acceleration, and all higher-order derivatives of the position variables [with respect to time or any other variable(s)]. Hence, the study of the kinematics of manipulators refers to all the geometrical and time-based properties of the motion.

Manipulators consist of nearly rigid *links* that are connected with *joints,* which allow relative motion of neighboring links. These joints are usually instrumented with position sensors that allow the relative position of neighboring links to be measured. In the case of rotary or *revolute* joints, these displacements are called *joint angles.* Some manipulators contain sliding, or *prismatic,* joints in which the relative displacement between links is a translation, sometimes called the *joint offset.*

The number of *degrees of freedom* that a manipulator possesses is the number of independent position variables that would have to be specified in order to locate all parts of the mechanism. This is a general term used for any mechanism. For example, the well-known, so-called "four-bar linkage" (see Fig. 1) has only one degree of freedom (even though there are three moving members). In the case of typical industrial robots, because a manipulator is usually an open kinematic chain, and because each joint position is usually defined with a single variable, the number of joints equals the number of degrees of freedom.

At the free end of the chain of links that make up the manipulator is the *end-effector.* Depending on the intended application of the robot, the end-effector may be a gripper, welding torch, electromagnet, or other device. We generally describe the position of the manipulator by giving a description of the *tool frame,* which is attached to the end-effector, relative to the base *frame,* which is attached to the nonmoving base of the manipulator (see Fig. 2).

A very basic problem in the study of mechanical manipulation is that of *forward kinematics.* This is the static geometrical problem of computing the position and orientation of the end-effector of the manipulator. Specifically, given a set of joint angles, the forward kinematic problem is to compute the position and orientation of the tool frame relative to the base frame. Sometimes we think of this as changing the representation of manipulator position from a *joint-space* description into a *Cartesian-space* description. By Cartesian space we mean the space in which the position of a point is given with three numbers, and in which the orientation of a body is given with three numbers. It is sometimes called *task space* or *operational space.*

The forward kinematic formulation is generally accomplished as follows. A *link frame* is attached to each link following the so-called *Denavit–Hartenberg* convention (Denavit and Hartenberg, 1955). This convention

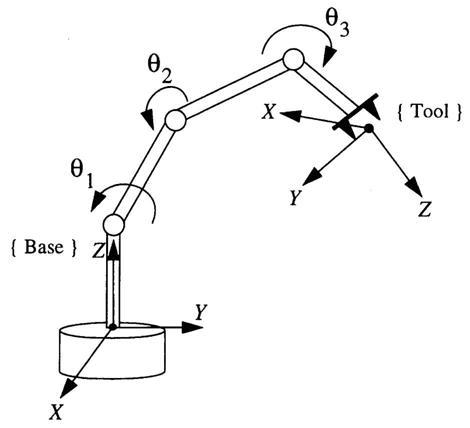

FIG. 2. Kinematic equations describe the tool frame relative to the base frame as a function of the joint variables.

specifies rules for locating link frames on links and defines four parameters that are sufficient to describe each link frame in terms of the preceding link frame. Figure 3 shows two successive link frames as well as the Denavit–Hartenberg parameters, a_i, α_i, d_i, and θ_i. The first two parameters, a_i and α_i, suffice to describe the relative spatial relationship between the lines that lie along successive joint axes. The parameter a_i, called the *link length,* is the length of the (usually unique) common perpendicular between the successive axes. The parameter α_i, called the *link twist,* is the angle between the two successive axes when projected onto a plane whose normal is the common perpendicular between the axes. The other two parameters, d_i and θ_i, have to do with the interconnection of links at joint i. The parameter d_i, called the *joint offset,* specifies the distance between the common perpendiculars describing the previous and following links at joint i. Finally, the parameter θ_i, called the *joint angle,* specifies the angle between successive common perpendiculars measured in a plane whose normal is joint axis i.

If the link frames have been attached to the links according to this convention, the following definitions of the link parameters are valid:

a_i = the distance from $\hat{Z}_i$ to $\hat{Z}_{i+1}$ measured along $\hat{X}_i$;

α_i = the angle between $\hat{Z}_i$ and $\hat{Z}_{i+1}$ measured about $\hat{X}_i$;

d_i = the distance from $\hat{X}_{i-1}$ to $\hat{X}_i$ measured along $\hat{Z}_i$; and

θ_i = the angle between $\hat{X}_{i-1}$ and $\hat{X}_i$ measured about $\hat{Z}_i$.

We usually choose $a_i > 0$ since it corresponds to a distance; however, α_i, d_i, and θ_i are signed quantities.

Given the Denavit–Hartenberg parameters for a mechanism, the transform that locates the ith frame in terms of the $(i - 1)$st frame is given by

$$
{}^{i-1}_{\ \ i}T = \begin{bmatrix} c\theta_i & -s\theta_i & 0 & a_{i-1} \\ s\theta_i c\alpha_{i-1} & c\theta_i c\alpha_{i-1} & -s\alpha_{i-1} & -s\alpha_{i-1}d_i \\ s\theta_i s\alpha_{i-1} & c\theta_i s\alpha_{i-1} & c\alpha_{i-1} & c\alpha_{i-1}d_i \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{5}
$$

where $c\theta_i \equiv \cos\theta_i$, $s\theta_i \equiv \sin\theta_i$, and so on. Note that for any one link transformation, three of the four parameters are constants, and the fourth is variable. In the case of a revolute joint, θ_i is the joint variable, and in the case of a prismatic joint, d_i is the joint variable.

Finally, the forward kinematics of a manipulator is then easily found by concatenating the link transforms to compute

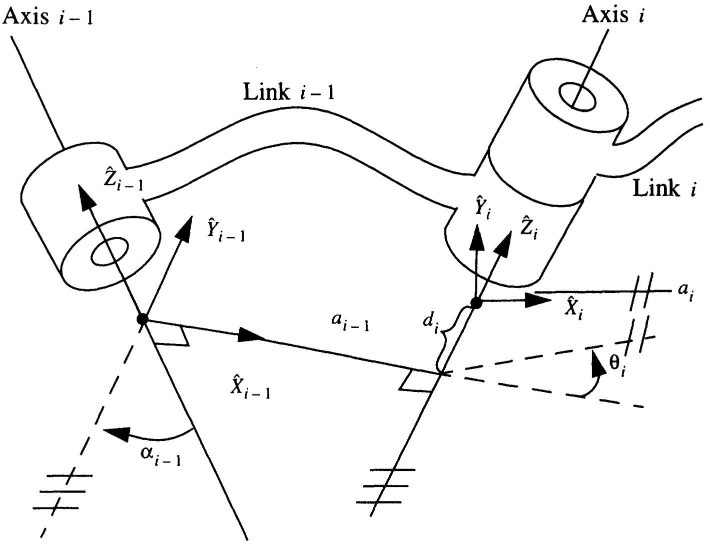

FIG. 3. Link frames are attached so that frame *i* is attached to link *i*, and has its *Z* axis aligned with the axis of rotation of joint *i*.

$$^0_NT = {}^0_1T\,{}^1_2T\,{}^2_3T \cdots {}^{N-1}_{\ \ N}T. \qquad (6)$$

This transformation, ^0_NT, will be a function of all n joint variables. If the robot's joint position sensors are queried, the Cartesian position and orientation of the last link may be computed by ^0_NT.

3. INVERSE KINEMATICS OF MANIPULATORS

The problem of *inverse kinematics* is posed as follows: Given the position and orientation of the end-effector of the manipulator, calculate all possible sets of joint angles that could be used to attain this given position and orientation (see Fig. 4). This is a fundamental problem in the practical use of manipulators.

The inverse kinematic problem is not as simple as the forward kinematics. Because the kinematic equations are nonlinear, their solution is not always easy or even possible in a closed form. Also, the questions of existence of a solution, and of multiple solutions, arise.

The existence or nonexistence of a kinematic solution defines the *workspace* of a given manipulator. The lack of a solution means that the manipulator cannot attain the desired position and orientation because it lies outside of the manipulator's workspace.

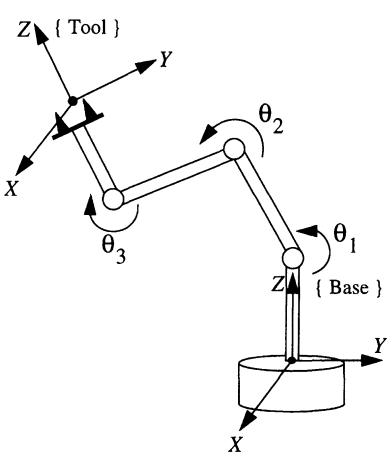

FIG. 4. For a given position and orientation of the tool frame, values for the joint variables can be calculated using the inverse kinematics.

Many industrial robots have six degrees of freedom, three of which form a *wrist* in which the last three axes of the robot are mutually orthogonal and intersect at a point. In these cases, the solution of the kinematic equations can be split into two problems: Solve for the first three joint variables in order to position the wrist point, and then solve for the final three joint angles to attain the desired orientation of the end-effector. The wrist point, or origin of the fourth link coordinate system, is given in the base system by

$$^0P_{4ORG} = {}^0_1T\,{}^1_2T\,{}^2_3T\,{}^3P_{4ORG}, \qquad (7)$$

where the left-hand side is given, and the right-hand side is a function of the joint variables of the first three degrees of freedom.

Solving a nonlinear set of equations such as Eq. (7) is difficult in general (Pieper and Roth, 1969; Tsai and Morgan, 1984). A sketch of a typical approach is as follows. Through algebraic manipulation, $n - 1$ of the n variables are eliminated, resulting in a transcendental equation in terms of the one remaining variable. Then, the trigonometric substitutions

$$u = \tan(\theta/2),$$

$$\cos\theta = \frac{1 - u^2}{1 + u^2},$$

$$\sin\theta = \frac{2u}{1 + u^2},$$

are used to convert the problem from a transcendental equation into a polynomial in powers of u. Finally, the roots of the polynomial are found to solve for one variable, and the other variables are generally then easily found through back substitution. Multiple roots of the polynomial in u correspond to multiple physical configurations of the robot that all yield the same position of the end-effector. Since quartic (and lower-order) polynomials can be solved in closed form, mechanisms whose kinematic equations can be manipulated into polynomials of order 4 or lower are called *closed-form solvable*. It turns out that what may seem to be relatively minor or subtle changes to the design of the mechanical configuration of the linkages can have a dramatic effect on the order of this "characteristic polynomial." Robots with more than six degrees of freedom are in a whole other class of device characterized by an infinity of solutions for a given end-point positioning problem (Baker and Wampler, 1988).

4. VELOCITIES, STATIC FORCES, SINGULARITIES

In addition to dealing with static positioning problems, we may wish to analyze manipulators in motion (Hunt, 1978). Often in performing velocity analysis of a mechanism it is convenient to define a matrix quantity called the *Jacobian* of the manipulator. The Jacobian specifies a *mapping* from velocities in joint space to velocities in Cartesian space (see Fig. 5). The nature of this mapping changes as the configuration of the manipulator varies. At certain points, called *singularities,* this mapping is not invertible. An understanding of the phenomenon is important to designers and users of manipulators.

In the field of robotics, we generally speak of Jacobians that relate joint velocities to Cartesian velocities of the tip of the arm. For example,

$$\mathcal{V} = J(\Theta)\dot{\Theta}, \tag{8}$$

where Θ is the vector of joint angles of the manipulator and $\mathcal{V}$ is a vector of Cartesian velocities. This 6 × 1 Cartesian velocity vector is the 3 × 1 linear velocity vector and the 3 × 1 rotational velocity vector stacked together:

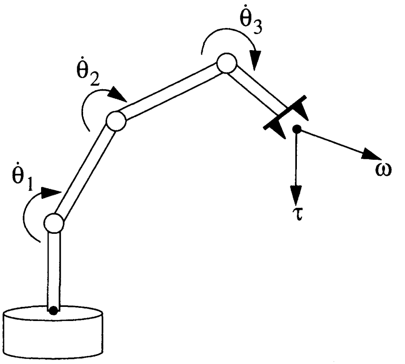

FIG. 5. The geometrical relationship between joint rates and velocity of the end-effector can be described in a matrix called the Jacobian.

$$\mathcal{V} = \begin{bmatrix} v \\ w \end{bmatrix}. \tag{9}$$

Given that we have a linear transformation relating joint velocity to Cartesian velocity, a reasonable question to ask is: Is this matrix invertible? That is, is it nonsingular? If the matrix is nonsingular then we can invert it to calculate joint rates given Cartesian velocities:

$$\dot{\Theta} = J^{-1}(\Theta)\mathcal{V}. \tag{10}$$

For example, say we wish the hand of the robot to move with a certain velocity vector in Cartesian space. Using Eq. (10) we could calculate the necessary joint rates at each instant along the path.

Most manipulators have values of Θ where the Jacobian becomes singular. Such locations are called *singularities of the mechanism* or *singularities* for short. All manipulators have singularities at the boundary of their workspaces, and most have loci of singularities inside their workspaces—for example, in special configurations of the mechanism in which two revolute joints become aligned. In such configurations, the motion of either joint causes the same motion at the end point of the kinematic chain, and so it is as if the device has lost one degree of freedom.

It turns out that the Jacobian also appears in the relationship between static forces applied at the end-effector and static actuator torques (or forces) applied at the joints. It can be shown that

$$\tau = J^T\mathcal{F}, \tag{11}$$

where $\mathcal{F}$ is a 6 × 1 Cartesian force-moment vector acting at the end-effector, and τ is an $n \times 1$ vector of torques at the joints (for an n-jointed robot).

Manipulators do not always move through space; sometimes they are also required to contact a workpiece or work surface and apply a static force. In this case the problem arises: Given a desired contact force and moment, what set of *joint torques* is required to generate them? Once again, the Jacobian matrix of the manipulator arises quite naturally in the solution of this problem.

5. DYNAMICS

Dynamics is a huge field of study devoted to studying the forces required to cause motion. In order to make a manipulator accelerate from rest, glide at a constant end-effector velocity, and finally decelerate to a stop, a complex set of torque functions must be applied by the *joint actuators.* We use joint actuators as the generic term for devices that power a manipulator; for example, electric motors, hydraulic and pneumatic actuators, muscles, etc. The exact form of the required functions of actuator torque depends on the spatial and temporal attributes of the path taken by the end-effector as well as the mass properties of the links and payload, friction in the joints, etc. One method of controlling a manipulator to follow a desired path involves calculating these actuator torque functions using the dynamic equations of motion of the manipulator.

Using any of several techniques to derive the equations of motion for a mechanical manipulator (Luh *et al.,* 1980), it is found that these equations can be written in the form

$$\tau = M(\Theta)\ddot{\Theta} + V(\Theta,\dot{\Theta}) + G(\Theta), \qquad (12)$$

where $M(\Theta)$ is the $n \times n$ *mass matrix* of the manipulator, $V(\Theta,\dot{\Theta})$ is an $n \times 1$ vector of centrifugal and Coriolis terms, $G(\Theta)$ is an $n \times 1$ vector of gravity terms, and τ is the $n \times 1$ vector of joint torques.

Each element of $M(\Theta)$ and $G(\Theta)$ is a complex function that depends on Θ, the position of all the joints of the manipulator. Each element of $V(\Theta,\dot{\Theta})$ is a complex function of both Θ and $\dot{\Theta}$. For an alternative formulation of manipulator dynamics, see Khatib (1985).

Another use of the dynamic equations of motion is in *simulation.* By reformulating the dynamic equations so that acceleration is computed as a function of actuator torque, it is possible to simulate how a manipulator would move under application of a set of actuator torques (see Fig. 6):

$$\ddot{\Theta} = M^{-1}(\Theta)[\tau - V(\Theta,\dot{\Theta}) - G(\Theta)]. \qquad (13)$$

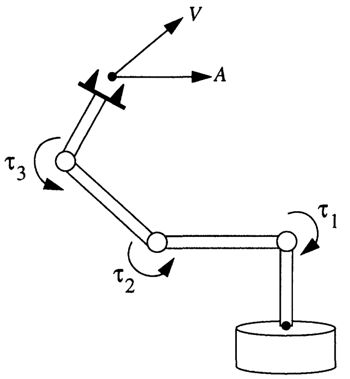

FIG. 6. The relationship between torques applied by the actuators and the resulting motion of the manipulator is embodied in the dynamic equations of motion.

We may then apply any of several known *numerical integration* techniques to integrate the acceleration to compute future positions and velocities.

6. TRAJECTORY GENERATION

A common way of causing a manipulator to move from here to there in a smooth, controlled fashion is to cause each joint to move as specified by a smooth function of time. Commonly, each joint starts and ends its motion at the same time, so that the manipulator motion appears coordinated. Exactly how to compute these motion functions is the problem of *trajectory generation* (see Fig. 7).

Often a path is described not only by a desired destination but also by some intermediate locations, or *via points,* through which the manipulator must pass en route to the destination. In such instances the term *spline* is sometimes used to refer to a smooth function that passes through a set of via points.

In order to force the end-effector to follow a straight line (or other geometric shape) through space, the desired motion must be converted to an equivalent set of joint motions. This *Cartesian trajectory generation* is a feature of most industrial robot control systems (Paul, 1981).

Whereas there is never a problem moving along smooth trajectories that are described

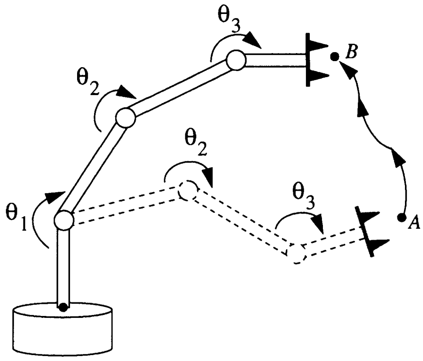

FIG. 7. In order to move the end-effector through space from point A to point B we must compute a trajectory for each joint to follow.

in terms of individual joint motions, when we impose an external path shape (such as straight-line) problems can arise.

Figure 8(a) illustrates that even though the initial and final positions of the robot are within the workspace, intermediate points may not be. A second type of problem with a straight-line path is shown in Fig. 8(b). In this case, the manipulator passes near a singularity that causes a very high joint velocity at joint 1 in order to maintain a constant tool tip velocity along the path. The closer the path comes to a singularity, the faster one or more joints are required to move. Finally, Fig. 8(c) indicates a third class of problem: Although all path points are reachable, they are not all reachable in the same kinematic solution. In this case, the path cannot be traversed. The problems illustrated in Fig. 8 are due to the realities of physical mechanisms, and are one cause of difficulty in programming robots to achieve useful tasks.

When a geometric model of the manipulator environment is available, advanced techniques can plan *collision-free paths* automatically (Lozano-Perez, 1987; Khatib, 1986). However, such functionality is presently not typically available in commercial robot systems.

7. MANIPULATOR DESIGN AND SENSORS

Although manipulators are in theory universal devices applicable to many situations, generally economics dictates that the intended task domain influence the mechanical design of the manipulator. Along with issues such as size, speed, and load capability, the designer must also consider the number of joints and their geometric arrangement. These considerations influence the manipulator's workspace size and quality, the stiffness of the manipulator structure (see Fig. 9), and other attributes.

Integral to the design of the manipulator are issues involving the choice and location of actuators, transmission systems, and internal position (and sometimes force) sensors.

8. POSITION CONTROL

Some manipulators are equipped with stepper motors or other actuators that can directly execute a desired trajectory. However, the vast majority of manipulators are driven by actuators that supply a force or a torque to cause motion of the links. In this case, an algorithm is needed to compute torques that will cause the desired motion. The problem of dynamics is central to the design of such algorithms but does not in itself constitute a solution. A primary concern of a *position control system* is to compensate

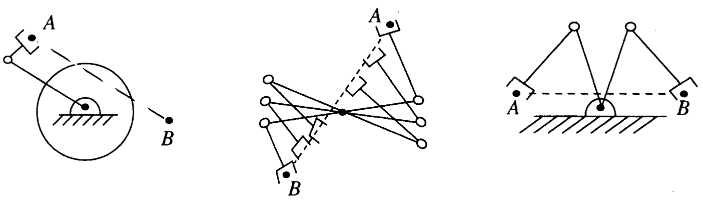

(a) (b) (c)

FIG. 8. Trajectory problem (a) intermediate points unreachable. Trajectory problem (b) high joint rates near singularity. Trajectory problem (c) no common solution for all path points.

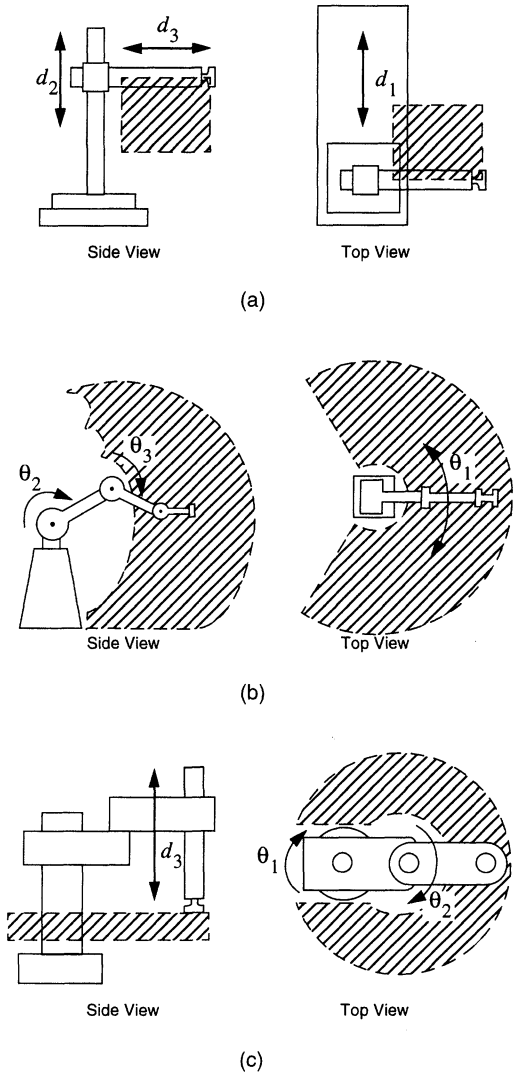

FIG. 9. Three common designs for robot mechanisms are Cartesian, Articulated, and Scara. Each type is characterized by a differently shaped workspace (as indicated) as well as other differentiating properties such as structural stiffness, positioning accuracy, etc.

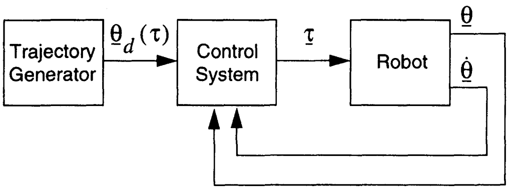

FIG. 10. In order to cause the manipulator to follow the desired trajectory, a position control system must be implemented. Such a system uses feedback from joint sensors to keep the manipulator on course.

automatically for errors in knowledge of the parameters of a system, and to suppress disturbances that tend to perturb the system from the desired trajectory (Franklin *et al.*, 1986). To accomplish this, position and velocity *sensors* are monitored by the *control algorithm,* which computes torque commands for the actuators (see Fig. 10). Many control systems are synthesized using linear approximate models of the dynamics of a manipulator. These linear methods are prevalent in current industrial practice.

Although control systems based on approximate linear models are popular in most current industrial robots, it is important to consider the complete nonlinear dynamics of the manipulator when synthesizing control algorithms. Some industrial robots now make use of *nonlinear control* algorithms in their controllers. These nonlinear techniques of controlling a manipulator promise better performance than do simpler linear schemes (Khosla, 1988; Khatib, 1987).

For a manipulator whose dynamics are given by

$$\tau = M(\Theta)\ddot{\Theta} + V(\Theta,\dot{\Theta}) + G(\Theta), \qquad (14)$$

where $M(\Theta)$ is the $n \times n$ inertia matrix of the manipulator, $V(\Theta,\dot{\Theta})$ is an $n \times 1$ vector of centrifugal and Coriolis terms, and $G(\Theta)$ is an $n \times 1$ vector of gravity terms, a nonlinear *control law* computes torques to send to the actuators as

$$\tau = \alpha\tau' + \beta, \qquad (15)$$

where τ is the $n \times 1$ vector of joint torques. We choose

$$\alpha = M(\Theta), \qquad (16a)$$

and

$$\beta = V(\Theta,\dot{\Theta}) + G(\Theta), \qquad (16b)$$

with the servo law

$$\tau' = \ddot{\Theta}_d + K_v\dot{E} + K_pE, \qquad (17)$$

where

$$E = \Theta_d - \Theta, \tag{18}$$

where Θ_d is the vector of desired joint positions, usually a function of time, and K_v and K_p are diagonal constant matrices with values selected by the control system designer for good performance. Using the above relations it is quite easy to show that the closed-loop system is characterized by the error equation

$$\ddot{E} + K_v\dot{E} + K_pE = 0. \tag{19}$$

Note that this vector equation is decoupled since the matrices K_v and K_p are diagonal so that Eq. (19) could just as well be written on a joint-by-joint basis as

$$\ddot{e}_i + k_{vi}\dot{e} + k_{pi}e = 0. \tag{20}$$

This is ideal second-order performance, as a result of the nonlinear decoupling and linearizing control above. In practice this is difficult to achieve because of imperfect knowledge of the parameters that appear in Eq. (16). The problem of imprecise knowledge of system parameters can sometimes be dealt with by means of *adaptive control* (Craig *et al.*, 1986).

9. FORCE CONTROL

The ability for a manipulator to control forces of contact when it touches parts, tools, or work surfaces seems to be of great importance in applying manipulators to many real-world tasks. *Force control* is complementary to position control in that we usually think of one or the other as applicable in a certain situation (Mason, 1978). When a manipulator is moving in free space, only position control makes sense, since there is no surface to react against. When a manipulator is touching a rigid surface, however, position control schemes can cause excessive forces to build up at the contact or may cause contact to be lost with the surface when it was desired for some application. Since manipulators are rarely constrained by reaction surfaces in all directions simultaneously, using a mixed or *hybrid* control (Raibert and Craig, 1981) is required, with some directions controlled by a *position control law* and remaining directions controlled by a *force control law.*

In Fig. 11, the task of putting a peg into a hole generally requires some form of force sensing and control. Even a seemingly simple task like this must be broken into a succession of several distinct regimes, each with a corresponding control strategy. In free space approaching the surface, position control is used. While sliding across the surface searching for the hole, force control would be used in the Z direction to maintain contact. Once started in the hole, force control would typically be used in the X and Y directions to avoid excessive side forces that might lead to jamming or wedging.

10. COMPUTER VISION

Increasingly, robots are being used with various sensors so that they can better adapt to uncertainties in their environment. *Computer vision* encompasses a suite of techniques that are used to identify, locate, or inspect objects. Many such techniques operate

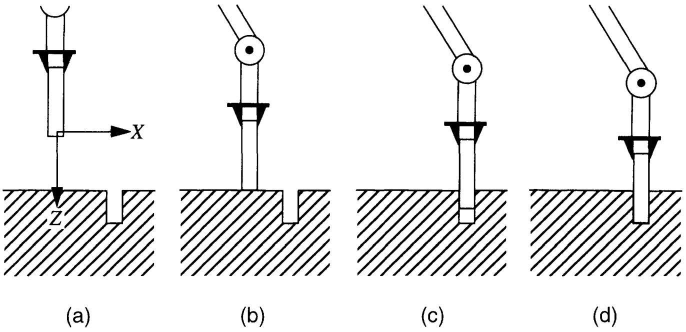

FIG. 11. In order for a manipulator to perform tasks that require contact with the environment (such as inserting a peg in a hole) the robot must be able to sense and control contact forces.

using a grayscale image acquired by a video camera. This image is then processed by various algorithms either to locate an object, to inspect an object for features, and/or to identify an object from a set of many possible objects. One simple technique uses *thresholding* to convert the grayscale image into strictly white or black regions, which are then analyzed for their size and shape. Often a *training phase* is done with the system first to train it to recognize the characteristics of the part to be identified. Another technique operates on the grayscale image to perform *edge detection* in order to convert the image into a set of lines or curves. These can then be matched to an expected set of edges in order to do object recognition, localization, or inspection. Some robot systems also use stereo vision (using two cameras) and solve a *correspondence* problem to identify features in each of the two views.

Computer vision techniques allow for more variability in the robot's workspace. For example, some arc-welding robots use a vision system to track the weld seam. Hence, if the parts have moved slightly, the operation can still proceed successfully. Robots without sensors must perform all operations by *dead reckoning,* which may lead to failures due to minor misalignments or changes in the environment. Similarly, robots use vision systems to see parts on conveyor belts, to find fiducial points on car bodies for alignment purposes, to locate a single object in a bin of parts, and so on.

11. LOCOMOTION

The vast majority of industrial robots are essentially analogous to a human arm mounted on the floor, or on a workbench. An important exception in current industrial practice is the AGV or *automatic guided vehicle.* AGVs are automatically guided carts that deliver parts in factories, or mail in some offices, or perform other operations in which mobility is required.

Although still rare, some robots have both mobility and one or more manipulators. There are also undersea and prototype outer-space robots that have both locomotion and manipulation abilities. Along with wheeled vehicles, there have been *legged* robots with one, two, four, or more legs (Raibert, 1986). These systems offer the promise of providing exploration or reconnaissance abilities in environments where it has traditionally been difficult for unmanned vehicles to go.

12. PROGRAMMING ROBOTS

A *robot programming language* serves as the interface between the human user and the industrial robot (Gruver and Soroka, 1988). Central questions arise such as: How are motions through space described easily by the programmer? How are multiple manipulators programmed so that they can work in parallel? How are sensor-based actions described in a language?

Robot manipulators differentiate themselves from fixed automation by being "flexible," which means programmable. Not only are the movements of manipulators programmable, but through the use of sensors and communications with other factory automation, manipulators can adapt to variations as the task proceeds.

The sophistication of the user interface is becoming extremely important as manipulators and other programmable automation are applied to more and more demanding industrial applications. The problem of programming manipulators encompasses all the issues of "traditional" computer programming, and so is an extensive subject in itself. Additionally, some particular attributes of the manipulator programming problem cause additional issues to arise.

13. OFF-LINE PROGRAMMING AND SIMULATION

An *off-line programming system* is a robot programming environment that has been sufficiently extended, generally by means of computer graphics, that the development of robot programs can take place without access to the robot itself (SILMA, 1995). A common argument raised in their favor is that an off-line programming system will not cause production equipment (i.e., the robot) to be tied up when it needs to be reprogrammed; hence, automated factories can stay in production mode a greater percentage of the time (see Fig. 12).

It also serves as a natural vehicle to tie

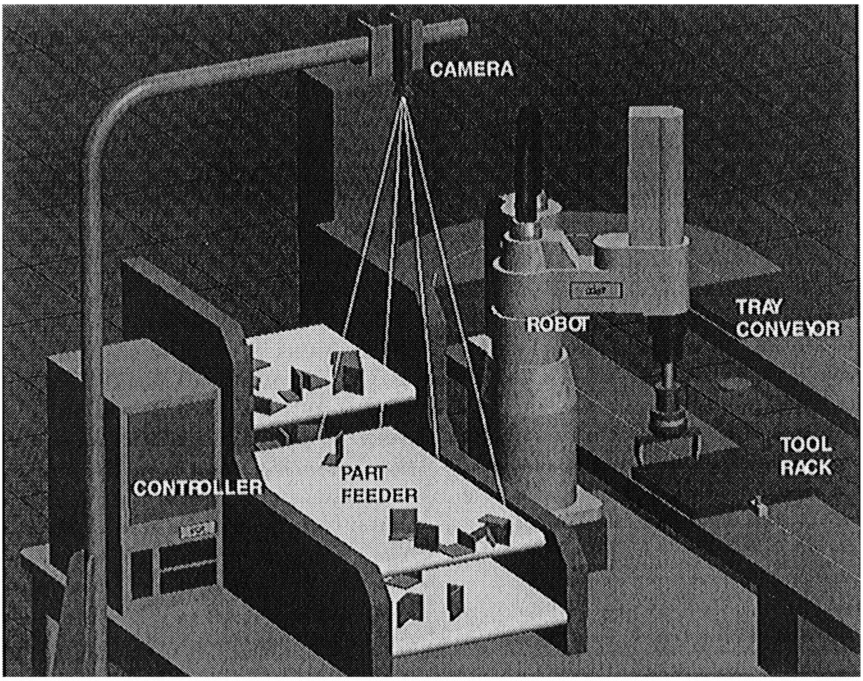

FIG. 12. Off-line programming (OLP) systems allow the programmer to interact with a 3-D simulated virtual world in order to program a robot.

computer-aided design (CAD) databases used in the design phase of a product to the actual manufacturing of the product. In some cases, this direct use of CAD data can dramatically reduce the programming time required for the manufacturing process.

GLOSSARY

Edge Detection: A technique used in computer vision in which changes in image intensity are linked together spatially to find image features that correspond to edges of objects in the image.

End-Effector: The gripper, tool, or other (generally active) object at the distal end of a manipulator.

Force Control: A style of controlling a manipulator such that the forces of contact (between the end-effector and other objects in the robot's environment) are directly controlled.

Homogeneous Transform: A 4×4 matrix containing both position and orientation information that specifies one coordinate system relative to a reference coordinate system.

Jacobian: Shortened name for the Jacobian matrix, a linear operator that maps joint velocities into linear and rotational velocities of the end point of a manipulator.

Kinematics: The study of position and motion (e.g., velocity, acceleration, etc.) without regard to the forces that cause the motion.

Mass Matrix: For an N-jointed manipulator, the $N \times N$ matrix that is a function of manipulator configuration and describes the mass distribution of all the linkages.

Off-Line Programming: The use of a computer language or system (generally augmented by the use of computer graphics) such that a robot may be programmed without access to the robot itself during programming.

Prismatic Joint: A joint between links that allows linear motion.

Revolute Joint: A joint between links that allows rotary motion.

Rotation Matrix: A matrix, usually of dimension 3 × 3, that describes the orientation of one coordinate system relative to a reference coordinate system.

Singularity: In robotics, a configuration of the mechanical linkage at which the Jacobian matrix loses full rank.

Works Cited

Baker, D., Wampler, C. (1988), "On the Inverse Kinematics of Redundant Manipulators," *Int. J. Rob. Res.* **7** (2), 3–21.

Craig, J., Hsu, P., Sastry, S. (1986), "Adaptive Control of Mechanical Manipulators," in: *Proceedings of the IEEE International Conference on Robotics and Automation, San Francisco,* Washington, DC: IEEE Computer Society Press.

Denavit, J., Hartenberg, R. S. (1955), "A Kinematic Notation for Lower-Pair Mechanisms Based on Matrices," *J. Appl. Mech.* **22,** 215–221.

Franklin, G., Powell, J. D., Emami-Naeini, A. (1986), *Feedback Control of Dynamic Systems,* Reading, MA: Addison-Wesley.

Gruver, W., Soroka, B. (1988), "Programming, High Level Languages," in: R. Dorf, S. Nof (Eds.), *The International Encyclopedia of Robotics,* New York: Wiley Interscience.

Hunt, K. (1978), *Kinematic Geometry of Mechanisms,* Cambridge, U.K.: Oxford Univ. Press.

Khatib, O. (1986), "Real-Time Obstacle Avoidance for Manipulators and Mobile Robots," *Int. J. Rob. Res.* **5** (1), 90–98.

Khatib, O. (1987), "A Unified Approach for Motion and Force Control of Robot Manipulators: The Operational Space Formulation," *IEEE J. Rob. Autom.* **RA-3** (1), 43–53.

Khosla, P. K. (1988), "Some Experimental Results on Model-Based Control Schemes," in: T. Pavlidis (Ed.), *Proceedings of the IEEE International Conference on Robotics and Automation, Philadelphia,* 3 vols., Washington, DC: IEEE Computer Society Press, p. 1380.

Lozano-Perez, T. (1987), "A Simple Motion Planning Algorithm for General Robot Manipulators," *IEEE J. Rob. Autom.* **RA-3** (3), 224–238.

Luh, J. Y. S., Walker, M. W., Paul, R. P. (1980), "On-Line Computational Scheme for Mechanical Manipulators," *J. Dyn. Syst. Meas. Control* **102** (2), 69–76.

Mason, M. (1978), "Compliance and Force Control for Computer Controlled Manipulators," M.S. Thesis, MIT AI Laboratory.

Paul, R. P. (1981), *Robot Manipulators: Mathematics, Programming, and Control,* Cambridge, MA: MIT Press.

Pieper, D., Roth, B. (1969), "The Kinematics of Manipulators under Computer Control," *Proceedings of the Second International Congress on Theory of Machines and Mechanisms, Zakopane, Poland,* Warsaw: Polytechnic Institute of Warsaw, Vol. 2.

Raibert, M., Craig, J. (1981), "Hybrid Position/Force Control of Manipulators," *J. Dyn. Syst. Meas. Control* **103** (2), 126–133.

Raibert, M. (1986), *Legged Robots That Balance,* Cambridge, MA: MIT Press.

SILMA Inc. (1995), "The CimStation User's Manual," Version 5.0, Available from SILMA Inc., 1601 Saratoga-Sunnyvale Rd., Cupertino, Calif., 95014.

Tsai, L., Morgan, A. (1984), "Solving the Kinematics of the Most General Six- and Five-Degree-of-Freedom Manipulators by Continuation Methods," Paper 84-DET-20, ASME Mechanisms Conference, Boston.

Further Reading

Craig, J. J. (1989), *Introduction to Robotics: Mechanics and Control,* 2nd ed., Reading, MA: Addison-Wesley.

Fu, K., Gonzalez, R., Lee, C. S. G. (1987), *Robotics: Control, Sensing, Vision, and Intelligence,* New York: McGraw-Hill.

Rivin, E. (1988), *Mechanical Design of Robots,* New York: McGraw-Hill.

ROCKS AND PHYSICAL GEOLOGY

See STRUCTURAL AND PHYSICAL GEOLOGY

SAFETY AND HEALTH IN THE WORKPLACE

See WORKPLACE HEALTH AND SAFETY

SATELLITES, ARTIFICIAL

Stephen J. Paddack and Bruce A. Campbell, *Advanced Missions Analysis Office, Goddard Space Flight Center, National Aeronautics and Space Administration, Greenbelt, Maryland, U.S.A.*

Enrico P. Mercanti, *Goddard Space Flight Center, National Aeronautics and Space Administration (Ret.), Annapolis, Maryland, U.S.A.*

Samuel Walter McCandless, *User Systems, Inc., Chesapeake Beach, Maryland, U.S.A.*

Introduction 519
1. **Physical and Technical Principles** 520
1.1 Orbit Principles 520
1.1.1 Orbits Used by Artificial Satellites 523
1.2 Space Environment 524
1.2.1 Orbital Debris 526
2. **Applications of Artificial Satellites** 527
2.1 Manned Spacecraft 527
2.2 Unmanned Satellites 527
2.2.1 Earth Observation 527
2.2.2 Space Environment 528
2.2.3 Planetary Exploration 529
2.2.4 Space Exploration 530
2.2.5 Commercial Satellites 531
2.2.6 Military Satellites 531
3. **Design, Function, and Operation of Artificial Satellites** 532
3.1 Satellite Design Phases and Processes 532
3.1.1 Conceptual Study Phase 533
3.1.2 Mission Analysis Phase 533
3.1.3 Definition Phase 533
3.1.4 Design and Development Phase 534
3.1.5 Operations Phase 535
3.2 Satellite Subsystems and Functions 536
3.2.1 Power Generation and Distribution 537
3.2.2 Thermal Control 537
3.2.3 Attitude Determination and Control 537
3.2.4 Orbit Maintenance and Propulsion 537
3.2.5 Structure and Mechanisms 538
3.2.6 Communications, Command, Control, and Data Handling 538
3.3 Satellite Operations 538
3.4 Space Law 539
4. **Artificial Satellite Manufacturing** 540
5. **Economic Aspects of Artificial Satellites** 542
Glossary 543
Further Reading 543

INTRODUCTION

Artificial satellites are now associated with many aspects of everyday life. Communications satellites route thousands of long-distance telephone calls around the world continuously. Most of these satellites also distribute video signals for movie channels and news stations. The pictures of the weather conditions shown in these newscasts are also provided by satellites. Some of these remote-sensing satellites also aid in monitoring the world's resources (oil, forests and crops, water) as well as evaluating the environment and determining man's effect on it. Other artificial satellites are used for navigation around the globe, exploring the solar system, examining the universe around us, and even carrying people into space.

Artificial satellites may be grouped into two major categories: manned and unmanned. While many principles of design and operation are common between these two categories, the requirements associated with providing a suitable environment for persons to survive in space are unique to manned missions and significantly affect the configuration of the spacecraft. Within these

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

two major categories, there are a number of subcategories that can be defined, mostly in terms of the objective the satellite is to achieve.

Most manned missions to date have been limited to near-Earth operations of relatively short durations. However, Soviet/Russian space stations have been the home for many cosmonauts for extended periods, some up to and beyond a year. The Apollo missions to the Moon represent the furthest that man has ventured into space (about 250 000 km away) although the missions all lasted less than two weeks each.

Unmanned space missions are many and varied, calling for different designs of the satellites that perform them. Some major categories have already been mentioned: communications, remote sensing, navigation, and scientific. Within these categories, more specific definition of the uses can be made. For instance, remote sensing satellites include weather satellites, land-resource satellites, ocean- and atmosphere-sensing satellites, and even spy satellites. The types of scientific satellites run the gamut from Earth-orbiting space-environment sensors and astronomical telescopes to interplanetary craft.

Though these missions may vary widely, and the design of the spacecraft may depend heavily on its mission, a number of principles are common to just about all systems that operate in space. After discussing these common principles, some of the particular demands on the design and/or operation of artificial satellites are explored.

1. PHYSICAL AND TECHNICAL PRINCIPLES

Satellites are constrained to operate from orbits and must exist in the environment of space for periods of time dependent on their uses.

1.1 Orbit Principles

The path through space of an artificial satellite is caused primarily by gravitational forces. In the classical two-body problem the primary body is massive with respect to the satellite and no other forces act on the bodies. Given two bodies in space, a primary and a satellite, one of two things occurs. Either the bodies collide or there will be relative motion between the bodies that follows certain predictable paths. In most cases, the primary body is used as the center of reference and the path or trajectory of the satellite is referred to as its orbit. Prior to the 1950s the motion of bodies in space was observed by astronomers and analyzed and explained by celestial mechanicians and physicists interested in central force-field problems. Since the advent of rockets and artificial satellites, however, orbital mechanics has become a popular field of study and employment.

The names of the historical observers and theorists in this field include Copernicus, Brahe, Kepler, Galileo, and Newton. Copernicus made the bold statement, based on observations, that the Earth and the other planets revolve around the Sun. Brahe made the precise observations that allowed Kepler to derive the now well-known three laws of planetary motion. These laws state that a body in orbit around another 1) describes a conic section, 2) sweeps out equal areas in space in equal times, and 3) completes one orbit in a time, the orbital period, whose square is directly proportional to the cube of the semimajor axis of the orbit. Galileo performed many experiments on the motions of bodies, and observed moons in orbit around Jupiter using the first astronomical telescope. Newton used the findings of his predecessors to derive the fundamental equations of motions based on gravitational force and energy principles.

An orbit and the position of the orbiting body in the orbit are uniquely defined by a set of six orbital elements. Figure 1 shows how the size, shape, and position of an orbiting body are defined. The three of the orbital elements associated with these properties are the semimajor axis (a), eccentricity (e), and true anomaly (v).

Orbits are classified according to shape: circular, elliptical, parabolic, and hyperbolic, as shown in Fig. 2. Newtonian mechanics show that the kinetic- and potential-energy levels of an orbiting object in space require its being in one of the conic-section classes, thereby explaining the observations of Brahe and Kepler's first law.

Circular and elliptic orbits are called "closed" orbits because a body in closed or-

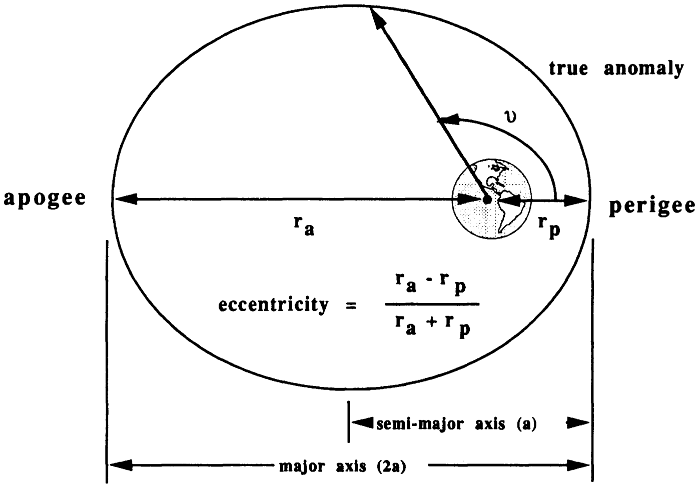
true anomaly
υ
apogee
perigee
r_a
r_p
eccentricity = $\frac{r_a - r_p}{r_a + r_p}$
semi-major axis (a)
major axis (2a)

FIG. 1. Orbital elements defining the size and shape of an orbit, and position of a body in an orbit.

bit is constrained to revolve around its primary mass. A satellite in a closed Earth orbit, for example, remains revolving around its primary, the Earth. A body in parabolic orbit possesses the minimum energy required to escape from its primary. A body in hyperbolic orbit possesses more than the minimum energy required to escape. Satellites launched from the Earth with hyperbolic velocities with respect to the Earth are used for interplanetary space travel or to escape from the solar system.

The shape of an orbit is given by its eccentricity (e). The value of eccentricity for closed orbits is bounded to be equal to or greater than zero but less than unity. A circular orbit has an eccentricity of zero. Elliptical orbits have eccentricity values greater than zero but less than one. A parabolic orbit has an eccentricity of exactly unity. Hyperbolic orbits have eccentricities greater than unity. In summary,

$e = 0 \Rightarrow$ circular orbit,

$0 < e < 1 \Rightarrow$ elliptic orbits,

$e = 1 \Rightarrow$ parabolic orbit,

$e > 1 \Rightarrow$ hyperbolic orbit.

The orientation in space of an orbit is defined by the remaining three elements. These are the argument of perigee (ω); the right ascension of the ascending node (Ω); and the inclination (i). These take on meaning only when presented in a spatial reference frame. Figure 3 shows an inertially fixed geocentric

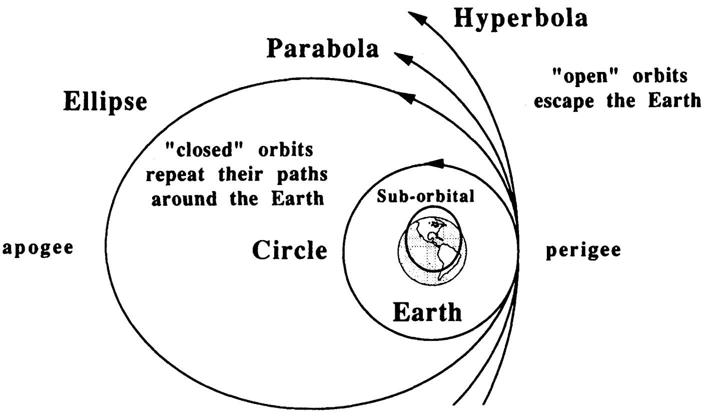
Hyperbola
Parabola
Ellipse
"open" orbits escape the Earth
"closed" orbits repeat their paths around the Earth
Sub-orbital
apogee
Circle
perigee
Earth

FIG. 2. Orbit types. Orbits around the Earth (or any other body) are conic sections.

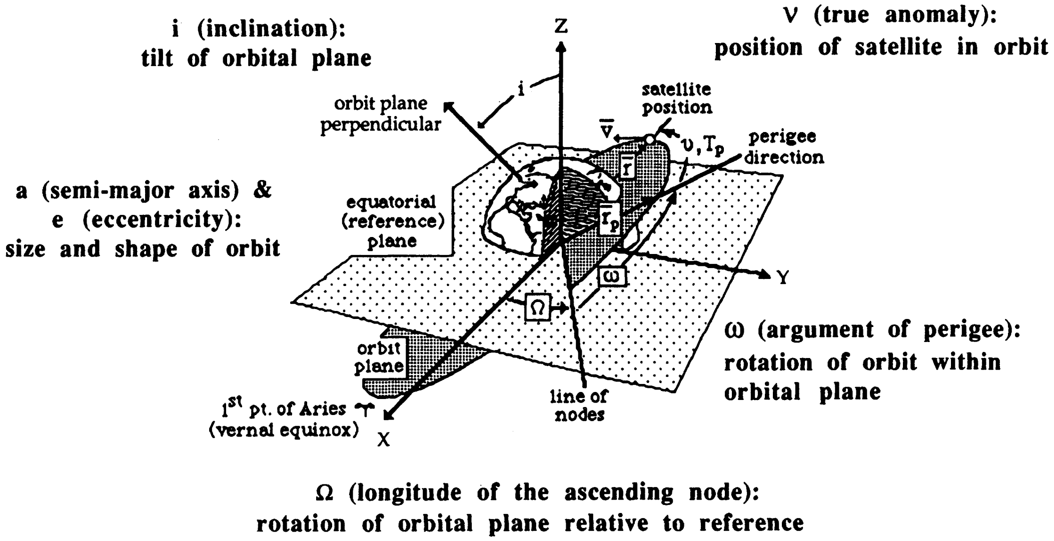
i (inclination):
tilt of orbital plane
Z
i
orbit plane
perpendicular
V (true anomaly):
position of satellite in orbit
satellite
position
$\overline{v}$
υ, T_p
perigee
direction
$\overline{r}$
$\overline{r}_p$
a (semi-major axis) &
e (eccentricity):
size and shape of orbit
equatorial
(reference)
plane
ω
Y
Ω
ω (argument of perigee):
rotation of orbit within
orbital plane
orbit
plane
1st pt. of Aries ♈
(vernal equinox)
X
line of
nodes
Ω (longitude of the ascending node):
rotation of orbital plane relative to reference

FIG. 3. Orbit description: the six "orbital elements" required to describe completely the orbit and position of a satellite in space.

Cartesian reference system where the fundamental (reference) plane is coincident with the Earth's equatorial plane and contains the *X* and *Y* axes. The positive *X* axis is directed to a specified point in space of astronomical significance. The position of the constellation Aries at a given time is often used. The *Y* axis is constructed perpendicular to the *X* axis. Finally, the *Z* axis is coincident with the Earth's spin axis with the positive direction defined as north. An orbit placed in this reference frame is also shown in the figure.

The inclination *i* is the angle between the plane of an orbit and the fundamental plane. The right ascension of the ascending node, *Ω*, is the angle in the fundamental plane measured from the positive *X* axis to the ascending nodal point (where an Earth-orbiting satellite pierces the equatorial plane when traveling from south to north). The right-hand rule is used for determining positive rotation. The argument of perigee, *ω*, is the angle measured in the plane of the orbit from the ascending node to the point of perigee.

These five parameters, *a*, *e*, *i*, *Ω*, *ω*, define the size, shape, and orientation of an orbit in space. The position of the satellite in the orbit can be given by a single angle that defines where the satellite is with respect to the point of perigee. This angle is called the true anomaly, *v*. The position of the satellite can be defined also by a time parameter, i.e., time from perigee, or the mean anomaly. The mean anomaly is based on the mean angular motion of the satellite.

In summary, the classical six orbital elements required to define completely an orbit and the position of a satellite in the orbit are

a,	semimajor axis;
e,	eccentricity;
i,	inclination;
Ω,	right ascension of the ascending node;
ω,	argument of perigee;
v,	true anomaly.

Kepler's second law is referred to as the areal law. It states simply that spatially equal areas are swept out in equal times. The most important thing this law reveals is that the speed of a satellite in an orbit will change as the orbital distance changes. Kepler's second law and its meaning to orbital velocities are shown in Fig. 4. Notice that, since its orbital distance does not change, the velocity of a satellite in a circular orbit is constant.

Kepler's third law is a mathematical statement that the time it takes for a body to complete one orbit is related to the size of the orbit. The time it takes to complete one orbit is called the orbital period, *T*. Later on,

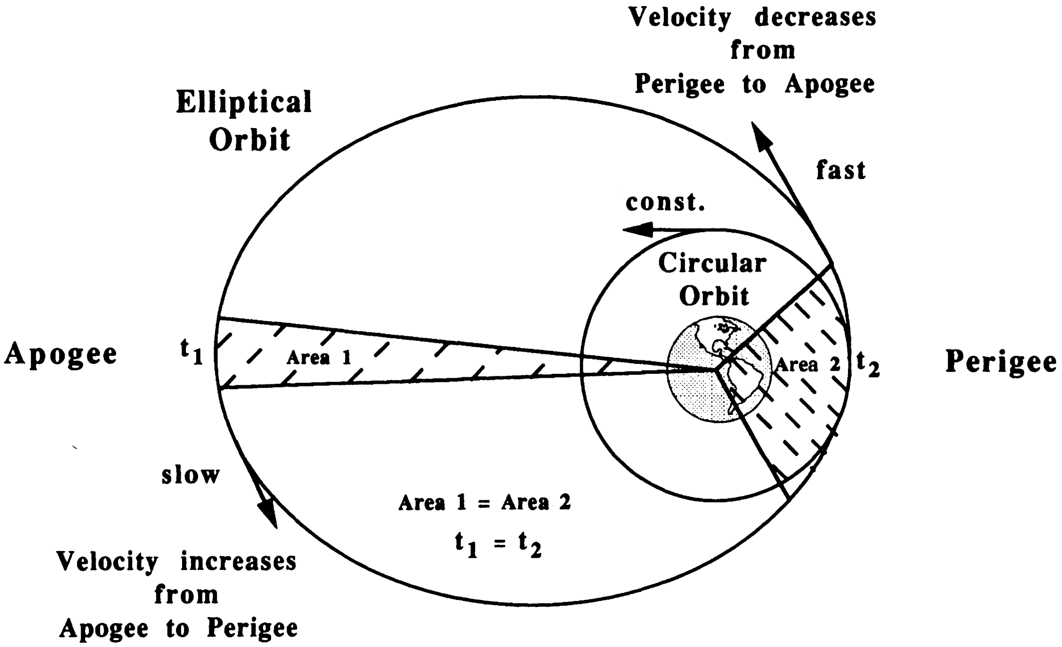

FIG. 4. Orbital velocities, illustrating Kepler's "areal" law, and its meaning for the velocity of a satellite in orbit around the Earth.

Newton was able to derive the constants associated with this relationship, which we can now describe as

$$T^2 = (4\pi^2/\mu)a^3,$$

in which μ represents the "gravitational parameter" for whatever body the satellite is orbiting. This law is constantly used by the orbit designer to satisfy the requirements of space-flight missions. One of the most obvious examples is the selection of an orbital distance giving an orbital period that precisely matches the rotation rate of the Earth, known as a "geosynchronous" orbit.

1.1.1 Orbits Used by Artificial Satellites The orbits chosen for different kinds of spacecraft depend on how the satellite is to be used. Some are low (altitude) Earth orbiting (LEO), some are geostationary (GEO), others are in high (altitude) Earth orbits (HEO), as depicted in Fig. 5. These terms, although often used in astronautics, are not uniquely defined except for the geostationary orbits. As a general rule, however, low Earth orbits are considered to be from the lowest altitude possible to allow continuing in orbit without reentering as a result of atmospheric drag to about 1000-km altitude.

Geostationary orbits are very specifically defined. As shown by Kepler's third law, close Earth-orbiting spacecraft have shorter orbital periods than spacecraft with larger orbits. It takes the Moon about 28 days to orbit the Earth once, but a spacecraft in LEO can complete one orbit in about an hour and a half. It is not difficult, then, to realize that there is a certain size orbit whose period is exactly one day, the geosynchronous orbit described earlier. If, in addition, the orbital plane of the spacecraft lies in the extended equatorial plane, and the orbit is circular (constant velocity), then the spacecraft will remain over a given longitudinal position on the Earth's equator. This special orbit is referred to as "geostationary" because a satellite in such an orbit remains fixed in the sky relative to any point on the Earth. Far from being stationary, the spacecraft is simply orbiting at an angular velocity that matches the Earth's rotational rate.

High Earth orbits, like low Earth orbits, are not uniquely defined. Rather, the term HEO is used to refer to orbits that are neither LEOs nor GEOs.

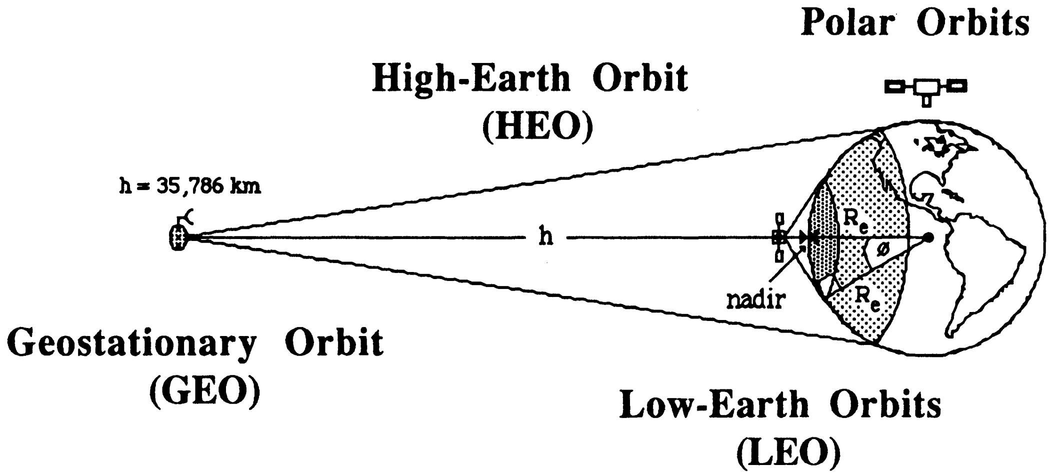

FIG. 5. Useful orbits with some common classifications of frequently used Earth orbits.

The following are some examples of space-flight missions requiring the various types of orbits. LEO orbits are used for any close Earth application such as Earth-resource monitoring and mapping or spy satellites. Because it takes more energy to achieve a higher orbit, all of the manned missions, with the exception of the Apollo flights to the Moon, have been and will continue to be exclusively in LEO orbits. The Hubble Space Telescope and other space-observing science satellites are in LEO orbits, since their fundamental requirement is to simply be out of the Earth's atmosphere to assure clarity of spacecraft sensor "vision." Some of these spacecraft are in highly inclined "polar" orbits that allow them to view the polar regions as well as the low-latitude regions.

GEO orbits are especially valuable for communications spacecraft and for global-weather spacecraft. Because the satellite remains fixed in the sky relative to any point on the Earth, it can be seen continuously by anyone in its field of vision using a fixed antenna. Figure 5 shows the large area of the Earth visible from a geostationary position, and a communications satellite can connect users thousands of miles apart. Besides providing the daily weather maps we see every night on the news, geostationary weather satellites can monitor developing storms far away from the home country and provide early warning of approaching danger.

High Earth orbits are used for scientific observations, navigation satellites, and also some communication satellites. These orbits can be highly elliptical, circular, or anything in between. An example of a science mission using high-eccentricity orbits would be those used for studying the magnetosphere, where it is important to take data close to the Earth as well as further from it. A Soviet astrodynamicist named Molniya developed a very special highly elliptical orbit that spends most of its time over the northern hemisphere. Such orbits are used by Russia for communications purposes, and a series of these satellites was named for the astrodynamicist. NASA has at least two spacecraft in very high circular orbits about half the distance to the Moon that can also be considered in HEO orbits.

Figure 5 shows that the area of the Earth that can be seen from a satellite changes with orbital distance. This is just another parameter that must be taken into account when designing an orbit for a particular application. The field of orbit design in astrodynamics combines celestial mechanics, engineering, and physics into a single discipline.

1.2 Space Environment

The environment in which satellites must operate is quite different from that of terrestrial systems. The Sun is the major environmental influence on the Earth and the space around it. The environment in the vicinity of the Earth is affected by two major products of the Sun's nuclear furnace: electromagnetic radiation, of which visible light is a part, and the outpouring of high-energy solar particles.

In the late 19th century, Stefan, Boltzmann, Wien, and Planck described the energy emitted by a blackbody as a result of its temperature. The Sun can be considered a blackbody at around 5750 K, which emits energy at essentially all wavelengths of the electromagnetic spectrum. However, the maximum energy output for such a blackbody occurs in the visible frequencies, toward which living things on Earth have developed and which the spacecraft designer uses to generate electrical power in space. The Earth also radiates electromagnetic energy, but because of its lower temperature (around 300 K) the Earth radiates its maximum energies in the infrared wavelengths. Both the Sun's and the Earth's energy must be taken into consideration in the design of satellites, especially in the area of thermal control.

In addition to the electromagnetic energy released by the Sun, a significant amount of particulate matter is emitted into space by the energy processes in and around the Sun. The Sun is composed mainly of charged particles (a plasma of loose protons and electrons), exposure to which can prove a hindrance or hazard for the operation of spacecraft and for the survival of humans in this environment. The Earth constantly encounters particulates streaming continuously from the Sun, which make up the relatively steady solar wind. However, occasional sharp increases in solar activity associated with disturbances such as sunspots and solar flares can produce a significant increase in particulate radiation that may be encountered by spacecraft in orbit around the Earth.

The environment in which spacecraft operate is affected in many ways by the electromagnetic and particulate radiations from the Sun. For example, the short-wavelength radiations from the Sun have enough energy to knock electrons out of the atmosphere, resulting in an ionized region, known as the ionosphere, that can modify and interfere with communication and remote sensing between Earth and space. The ionosphere is highly correlated with solar activity and varies dramatically geographically, seasonally, and with the day–night cycle. Radiation belts are also created, consisting of charged particles that become trapped within the Earth's geomagnetic field. These particles can affect spacecraft electronics and interfere with and threaten man's presence in space.

The Earth influences the space environment in other ways. The atmosphere extends, in continuously decreasing densities, far into space. Variations in temperature, atmospheric density, and the geomagnetic field result from complex and changing interactions between the Earth and the Sun, creating a dynamic environment in which spacecraft must operate. Figure 6 depicts the many factors involved in creating the space environment encountered by spacecraft in Earth orbit. The space-systems engineer must be familiar with this complex and fluctuating environment in order to determine how it is going to affect the spacecraft, both physically and operationally. Among other things, the effects mentioned influence thermal control, cause spacecraft charging, corrode materials, cause upsets in electronic subsystems, and create drag that modifies the spacecraft orbit. Man has also contaminated space with significant amounts of debris that create concern for the design of future missions. All of these considerations become increasingly important when humans are placed into the space environment. They are particularly susceptible to high-en-

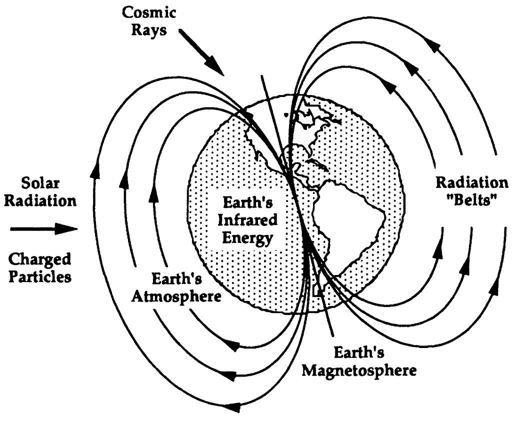

FIG. 6. Factors affecting the environment encountered by spacecraft in Earth orbit.

ergy particle radiations, which can cause cell damage and can be fatal if the radiation reaches high levels during a significant solar event.

1.2.1 Orbital Debris The world's rapidly escalating exploration of space is influencing the space environment in potentially threatening ways. The proliferation of discarded launch-vehicle parts and "dead" satellites locked in Earth orbits or traversing interplanetary space—collectively referred to as "space debris"—can cause enormous damage in a number of ways. Three of the most widely discussed dangers are the following:

1. A satellite or launch vehicle collides with a piece of debris. If the debris is not detectable and not tracked from space or the ground, the active satellite or rocket would have no knowledge it was on a collision course. Thus, if the two objects, each traveling in excess of 15 000 mph, collide there could be, at worst, the total destruction of both or, if the debris is of small mass, less but still substantial damage could ensue. The concern focuses on the geostationary orbits with their high population of unused communications and weather satellites and discarded rocket-stage hardware.
2. The orbit of the debris object decays. Gravitational forces could cause the debris to penetrate the near-Earth atmosphere and either burn up or crash on the Earth's surface at a location that may be highly populated. Because of the reentry velocity of the descending mass and its usually inoperative electric-signal–generating subsystems, it may not be detectable by ground surveillance instruments. Consequently, the incoming debris would impact the surface in much the same manner as a projectile would.
3. Contamination of other bodies. Because we do not know what life may exist on other planets, an interplanetary probe can, if it lands with a significant mass and at high speed, cause untold damage to the planet's environment, terrain, and possible life there. If the object carries contaminants from Earth, their spread on a distant celestial body could wreak havoc with humans or plant life.

These and other dangers that can result from space debris are prompting concern among international bodies such as the U.N. and countries heavily engaged in deploying satellites, shuttles, space stations, and interplanetary probes in space. This new awareness resulted in the first European Conference on Space Debris, held in Germany in April 1993. Organized by the European Space Agency (ESA), it was cosponsored by Italian, British, French, and German government space organizations. More than 100 presentations were made, 250 experts attended, and 17 countries were represented including the United States, Japan, and Russia. And certainly the U.N. had to have been involved, too. Conference organizer W. Flury stated that the purpose of the conference was to present space-debris–research results and define future research directions, identify methods of debris control, discuss needed regulations, and start considering global policy issues. He concluded, "Decision makers have to be convinced that debris mitigation measures have to be applied despite additional costs and the fact that immediate benefits to their projects may be very small."

Currently, worldwide efforts pertaining to space debris are focused on the following aspects:

1. The detection, measurement, and movements of large and particulate-sized manmade and naturally occurring objects to minimize placement of satellites in potentially dangerous domains of space. Techniques used or under consideration include radar, ground telescopes, space optics, and seismic instruments. The capture or collection of small debris particles using thin film and micropore foams is also under experimental investigation. Models are also being created and modeling techniques improved to define and predict more precisely the quantity of space debris, its location, and its motion as functions of time and position.
2. Study of the effects of impacts due to space debris on the performance and safety of rockets and satellites, with research into methods of protecting (shielding) these vehicles from damage actively in process. Manmade objects retrieved from space are now examined for space-debris effects or damage.
3. In the area of risk analysis, calculation of

the probability of collisions in the various orbits and in launch trajectories with special emphasis placed on geostationary orbits. Statistical databases, particularly of manmade abandoned or lost space segments, are under development to use in mapping safer anticollision routes for future flights.

An important risk factor pertains to nuclear-power sources in orbit that may malfunction and be abandoned in space, constituting an extremely hazardous environment should impact or even proximity occur.

4. In close conjunction with ongoing scientific and technological thrusts, study of the legal aspects involved with providers of space debris (liability, culpability, etc.), carried out internationally and by aerospace corporations, probably in much the same manner as it did when public use of aircraft soared. Policies to regulate space debris are now being formulated for later adoption by manufacturers, commercial users of space, governments, and the U.N.

2. APPLICATIONS OF ARTIFICIAL SATELLITES

2.1 Manned Spacecraft

The Soviets were first in achieving manned space flight with the launch of Air Force Major Yuri A. Gagarin in the 4706-kg (10 400-lb) Vostok 1 ("East") spacecraft on 12 April 1961 for a single orbit. Three weeks later, on 5 May 1961, the United States launched Navy Commander Alan B. Shepard, Jr., into space. The purpose behind the Mercury program was to gain basic data on the effects of space flight on human beings. Subsequent two-man Gemini flights were designed to evaluate the ability of performing the tasks in space required for a manned lunar landing. On 20 July 1969, only eight years after President Kennedy challenged the nation, astronauts Neil Armstrong and Edwin "Buzz" Aldrin descended to the surface of the Moon. After the lunar landings, using remaining Apollo equipment, the United States launched a space station, Skylab, and as an act of détente, conducted a joint space operation involving the docking of an Apollo capsule with a Soyuz spacecraft.

From 1975 to 1981 no American manned missions were launched, while design and construction of the next manned space vehicle, the Space Shuttle, proceeded. Since 12 April 1981, the shuttle fleet has flown missions carrying satellites, experiments, and crews of up to seven persons into low Earth orbit. With the exception of the *Challenger* explosion on 28 January 1986, the Space Shuttle has proven to be a capable and versatile space vehicle. Figure 7 shows a comparison of the various U.S. manned missions.

Soviet and American space programs paralleled each other in many ways. The one-man Vostok spacecraft was modified to carry a larger crew and was redesignated Voskhod ("Sunrise"). The first Soyuz ("Union") spacecraft was launched in April 1967 and has been used, with modifications, ever since. From 1971 to 1982 six successful Salyut space stations were launched and visited by a total of 29 crews. The current Russian space station, Mir ("Peace"), has been the home for cosmonauts for over a year, and several docking and laboratory modules have been added to the basic core.

The future of manned space flight is becoming more international, with Space Shuttle flights to the Russian "Mir" and with the involvement of Russia, the European Space Agency, and the Japanese in the International Space Station project.

2.2 Unmanned Satellites

In addition to manned spacecraft, the United States, the Soviet Union/Russia, and more recently many other countries have developed, have launched, and operate hundreds of unmanned spacecraft designed to perform military, national-interest, scientific, and commercial applications. The following sections highlight some of the application areas in which these spacecraft perform, and the evolution of spacecraft within each area.

2.2.1 Earth Observation Because of the unique vantage point of space, many satellites are designed to observe the Earth to collect data on the environment. TIROS 1 (Television Infrared Observation Satellite), the first weather satellite, was launched in April 1960 and returned over 22 500 pictures of the Earth. TIROS, NOAA (National Ocean-

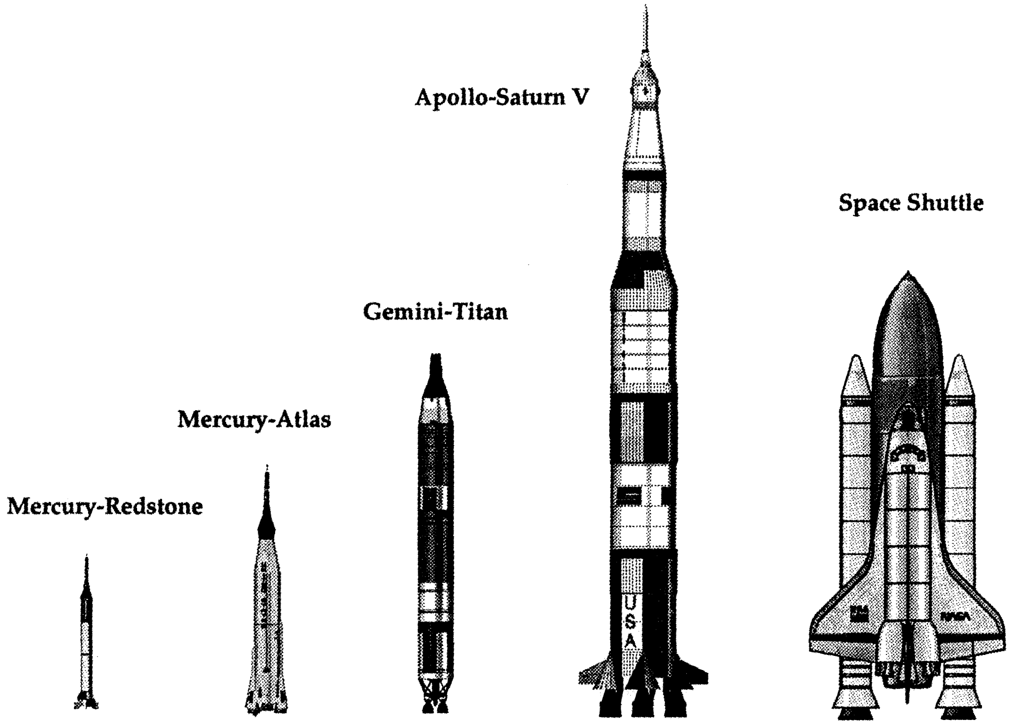

FIG. 7. U.S. manned launch vehicles. The figure, drawn approximately to scale, shows the relative sizes of the vehicles used for these missions. All of these vehicles have also been used to launch artificial satellites into orbit.

ographic and Atmospheric Administration), GOES (Geostationary Operational Environmental Satellite; Fig. 8), and the military DMSP (Defense Meteorological Satellite Program) systems monitor the Earth's weather, allowing timely prediction of climatic changes and their effects on the planet.

In September 1991, the Upper Atmosphere Research Satellite (UARS) was deployed from the Space Shuttle to conduct the first comprehensive studies of the stratosphere, mesosphere, and thermosphere regions of the atmosphere. The Total Ozone Mapping Spectrometer (TOMS), a NASA instrument that "hitched a ride" on U.S. Nimbus and Soviet/Russian Meteor weather satellites, discovered the "ozone hole" region of dangerously depleted ozone levels that appears over Antarctica each fall. Missions such as these allow researchers to map the short- and long-term changes occurring in the atmosphere and to determine what effects may be attributable to humans.

ERTS 1 (Earth Resources Technology Satellite) was launched in July 1972 into an orbit over the poles to gather information on geology, crops, population, and pollution, even gathering correlative information from ground-based instruments. ERTS evolved into the Landsat series of satellites, which continue to focus on discovering and monitoring Earth's resources in even the remotest areas of the world.

In June 1978, Seasat was launched to gather information on sea temperature, sea ice, wind speed and direction, and wave heights. Though its mission was cut short on account of an electrical failure, Seasat demonstrated the beneficial use of synthetic-aperture radar (SAR) in Earth-monitoring applications. A U.S./French satellite, TOPEX/POSEIDON, launched in August 1992, uses a radar altimeter to measure precisely the changes in global sea states to help determine the ocean's role in Earth's climate.

2.2.2 Space Environment James Van Allen used data from America's first satellite

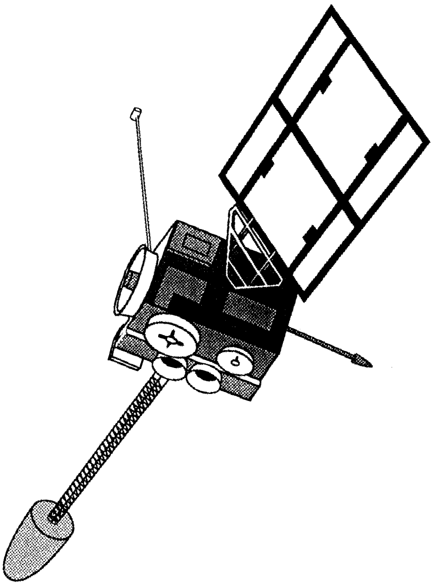

FIG. 8. GOES. The Geostationary Operational Environmental Satellites are built by NASA for the National Oceanographic and Atmospheric Administration (NOAA) of the U.S. Department of Commerce.

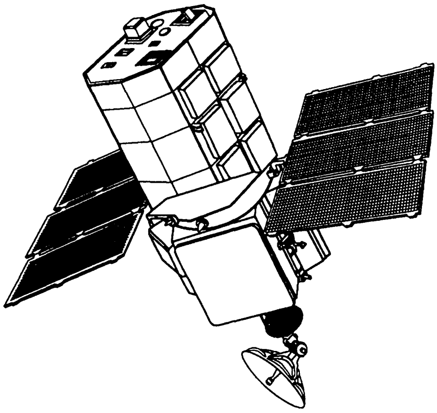

FIG. 9. SolarMax. This satellite was designed to help determine changes in the Earth's space environment due to changes in the Sun's activities.

(Explorer 1) to discover the "belts" of energetic particles encircling our world that now bear his name. Subsequent Explorer spacecraft allowed scientists to probe the solar wind and the Earth's magnetic field. The interaction between these phenomena was examined by the International Sun-Earth Explorer (ISEE) satellites, launched in 1977 and 1978, two in orbit around the Earth and one positioned in a special orbit between the Sun and the Earth. Pioneer and Voyager spacecraft, sent to visit the planets, also sampled the interplanetary environment. Now, that the Pioneer and Voyager craft are on their way out of our solar system, scientists continue to monitor their signals in hopes of gaining information on interstellar space.

The Sun has the greatest effect on the near-Earth space environment, and a series of Orbiting Solar Observatory (OSO) spacecraft launched between 1962 and 1975 were used to study solar flares and temperature differences on the Sun's surface. The Solar Maximum Mission (SMM or "SolarMax"; see Fig. 9) was launched in 1980 to observe the Sun during the maximum phase of its 11-year cycle. This spacecraft was the objective of a Space Shuttle rendezvous and EVA repair mission in 1984. Ulysses, a joint U.S.–European Space Agency spacecraft deployed from the Space Shuttle in October 1990, used a gravity assist from Jupiter to send it into a solar–polar orbit to allow mapping the solar environment in this previously unexplored region of our solar system.

Long-term effects of the space environment on spacecraft have been investigated with the Long Duration Exposure Facility (LDEF). It was released from the Space Shuttle in 1984 for what was planned as a one-year stay, but the *Challenger* accident delayed its recovery until 1990 when it was finally returned to Earth for analysis.

2.2.3 Planetary Exploration Planetary exploration began with flights to our nearest neighbor, the Moon. Before astronauts were sent there, several Ranger spacecraft were sent crashing into the lunar surface, sending back high-resolution photographs right up until impact. Lunar Orbiters photographed the Moon from orbit to locate possible landing sites, and Surveyor craft soft-landed on the Moon and sent back photographs and conducted lunar-soil sampling experiments.

Mariner 2 became the first spacecraft to fly by another planet, passing within 21 000 miles of Venus in December 1962. Subsequent Mariner and Pioneer spacecraft were sent toward Mercury, Venus, Mars, Jupiter, and Saturn—all the planets known to early astronomers! In 1976, two Viking spacecraft successfully landed on Mars, sending back pictures of the Martian landscape and conducting soil and atmospheric experiments. Spectacular pictures and a wealth of planetary information on Jupiter, Saturn, Uranus, and Neptune were sent back from the two Voyager spacecraft, now on their way out of our solar system. The Magellan spacecraft used a synthetic-aperture radar (SAR) to map more than 98% of the surface of the cloud-enshrouded planet Venus from September 1990 to September 1992. Galileo (Fig. 10), deployed in October 1989 from the Space Shuttle, used both Venus and the Earth for gravity-assist boosts to reach the planet Jupiter. Along the way, and despite a failed main-antenna deployment, Galileo flew by the asteroids Gaspra (October 1991) and Ida (August 1993), giving us our first close looks at these solar-system companions.

The Soviets sent numerous spacecraft to explore Venus, and two of their Venera craft successfully landed on the surface in March 1982. The landers were only able to transmit pictures and data for a short time until the harsh environment (temperatures that will melt lead and pressures about 100 times that on Earth) disabled the spacecraft. While the United States conducted the first flyby of a comet (Giaccobini-Zinner) in 1985, a flotilla of spacecraft from the Soviet Union, the European Space Agency, and Japan made somewhat more noteworthy flybys of Halley's comet in 1987.

2.2.4 Space Exploration The Orbiting Astronomical Observatory (OAO 2) was launched in December 1968 with a complement of ultraviolet instruments designed to examine the stars and discovered evidence of the existence of "black holes." High-Energy Astronomy Observatory (HEAO) spacecraft were launched between 1977 and 1979 to conduct x-ray, gamma-ray, and cosmic-ray surveys of the universe, and the International Ultraviolet Explorer (IUE) conducted observations in the ultraviolet (UV) wavelengths. Launched in November 1989, the Cosmic Background Explorer (COBE) carried instruments designed to search the infrared background of deep space for information about the "Big Bang" theory of the origin of our universe.

Deployed in April 1990 from the Space Shuttle, the Hubble Space Telescope (HST, Fig. 11) was the first of the "Great Observatories," facility-class satellites designed to offer scientists unparalleled opportunities to explore the universe. HST has a 2.4-m (94.5-in.) diameter mirror that can pick up objects 50 times fainter and 7 times more distant than present ground-based observatories, potentially expanding the universe visible to astronomers by 500-fold! Since it was designed for on-orbit servicing, in December 1993 astronauts were able to insert a set of corrective optics to offset the "spherical aberra-

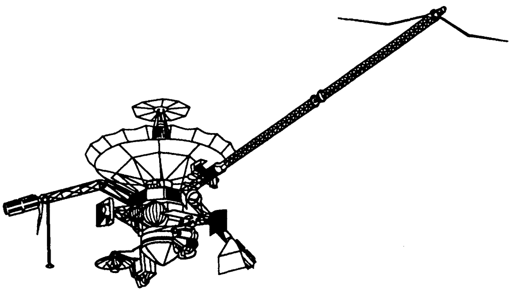

FIG. 10. Galileo. This spacecraft was designed to release a probe into the atmosphere of Jupiter and to remain in Jovian orbit to study the planet and its moons.

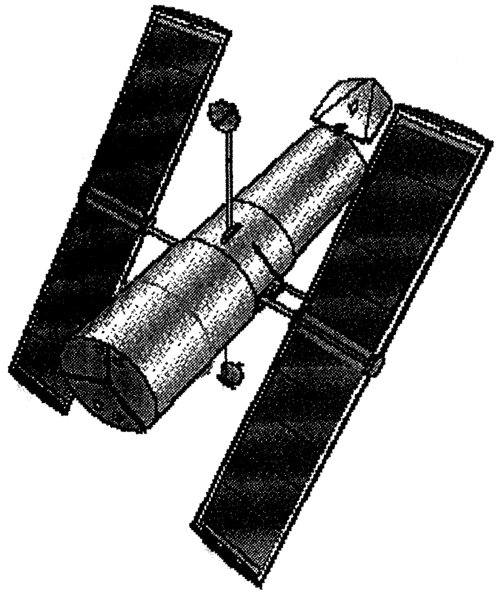

FIG. 11. Hubble Space Telescope. From its position high above the Earth's atmosphere, HST provides an unobstructed view of the universe.

tion" flaw discovered in the primary mirror that compromised HST's capabilities. The second great observatory, the Compton Gamma Ray Observatory (GRO), was deployed from the Space Shuttle in April 1992. GRO was designed to capture the relatively unimpeded gamma-ray emissions from cataclysmic cosmic events like supernovas, black holes, and even the remnants of the Big Bang. Other great observatory satellites may follow; however, the current trend is toward smaller, cheaper satellites.

2.2.5 Commercial Satellites By far, the largest commercial interest in space systems is in the relay of information or communications. Communication satellites had a humble start with Echo 1, launched in August 1960, which was simply a huge radio-reflecting balloon in low Earth orbit. By 1963, Syncom 1 demonstrated the capability and advantages of communicating via a satellite in a geostationary orbit. Today, over one hundred communication satellites, some similar to the one shown in Fig. 12, exist in this orbit, each capable of relaying thousands of telephone calls and television pictures simultaneously around the globe. In 1962, President Kennedy signed a bill creating the Communications Satellite Corporation (Comsat), a joint government–industry undertaking allowing transfer of U.S.-developed satellite communications technology for commercial use. The International Telecommunications Satellite Consortium (Intelsat), with Comsat representing the United States, promotes international use of these capabilities. Both organizations have been highly successful.

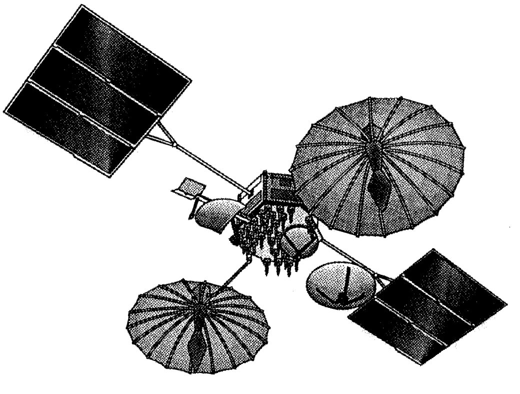

FIG. 12. Communications satellite. Though not a commercial satellite, NASA's Tracking and Data Relay Satellite (TDRS) displays the large antennas and solar arrays required to transmit and receive from geostationary orbit.

A similar arrangement exists involving dissemination of data gathered by the Landsat satellites for which demand by the commercial public has grown. The French SPOT (Satellite Probatoire de l'Observation de la Terre) satellite competes in this area, offering high-resolution photographs of points of interest on the Earth to news media, governments, and even military organizations worldwide.

2.2.6 Military Satellites The early experiments using captured V-2 rockets were conducted under the auspices of the military and led to the development of the Intercontinental Ballistic Missile (ICBM). As other uses of space became apparent, the military conducted their own experimental programs and developed systems to take advantage of the new "high ground" of space.

Today, the military routinely conducts force-enhancement missions such as communications, navigation, and remote environmental sensing using dedicated systems in space. Surveillance and monitoring missions are also routinely conducted, though the systems used are often classified. However, the worth of these systems became generally apparent during the war with Iraq, where coalition forces scored a decisive victory using these capabilities extensively.

Navigation satellites are probably the next most successful area in the commercial use of space systems. In this case, however, the private sector is taking advantage of a military satellite. The Global Positioning System (GPS, Fig. 13) and the earlier Navy Navigation System (Transit) satellites were designed for use by military forces to provide a worldwide, continuous positioning capability. As an afterthought, the U.S. government made these signals available for nonmilitary use, and the number of civilian users now far outweighs the military. Unlike the satellite communications industry, companies that provide GPS services benefit in that the cost associated with the satellites has already been borne by the U.S. government.

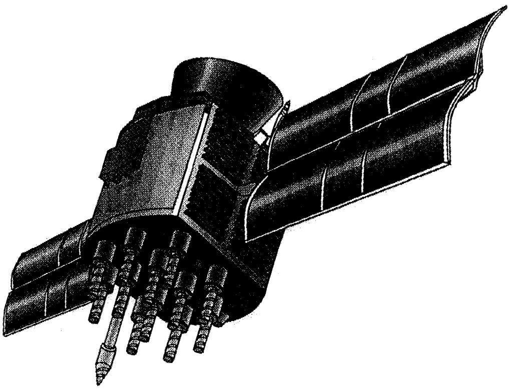

FIG. 13. Global Positioning System satellite. Though the primary mission of the constellation of GPS satellites is to provide continuous, very accurate navigation information to the U.S. military, the number of civilian and foreign users continues to grow.

3. DESIGN, FUNCTION, AND OPERATION OF ARTIFICIAL SATELLITES

3.1 Satellite Design Phases and Processes

The prerequisite for all satellite design phases and processes described is that a need exists for the mission under consideration to justify the time, effort, and sizable expense required to warrant its development. The need is often defined by individuals or groups of space scientists scattered all over the world. To further their research, they may desire the placement of instruments in space to acquire new, more, or better data on the basis of their analyses to date. Demands for artificial satellites also originate from federal military and civilian government agencies, foreign countries, and industry. Organizations such as international bodies, NASA advisory committees, the National Academy of Sciences, the National Space Council, and university research institutions also play important roles in defining needs for satellites—manned or unmanned.

Whatever the source of the initial concept for the mission, further development usually follows a structured approach. Figure 14 shows a typical mission-development approach used by NASA, which is easily compared with approaches taken by industry or other government agencies.

The approach uses defined phases associated with the level of detail expected at each stage of mission development. A number of reviews are used to ensure that the development is proceeding as expected and that all factors are such that the development can

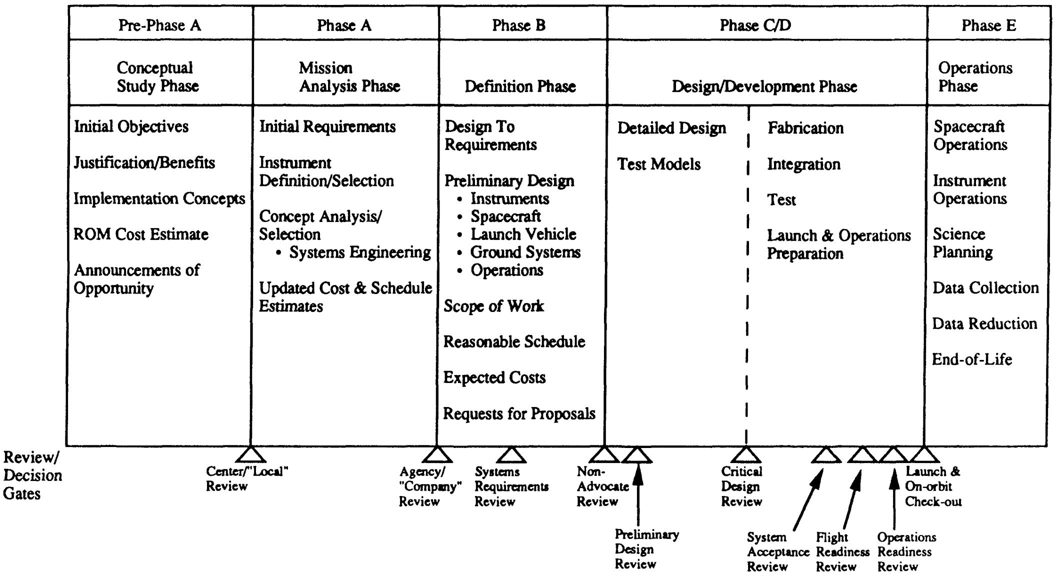

Pre-Phase A	Phase A	Phase B	Phase C/D	Phase E
Conceptual Study Phase	Mission Analysis Phase	Definition Phase	Design/Development Phase	Operations Phase
Initial Objectives Justification/Benefits Implementation Concepts ROM Cost Estimate Announcements of Opportunity	Initial Requirements Instrument Definition/Selection Concept Analysis/ Selection • Systems Engineering Updated Cost & Schedule Estimates	Design To Requirements Preliminary Design • Instruments • Spacecraft • Launch Vehicle • Ground Systems • Operations Scope of Work Reasonable Schedule Expected Costs Requests for Proposals	Detailed Design Test Models Fabrication Integration Test Launch & Operations Preparation	Spacecraft Operations Instrument Operations Science Planning Data Collection Data Reduction End-of-Life

FIG. 14. Mission development chart, showing the various design "phases" with their components and the many reviews involved.

proceed to the next level. Considering the expense associated with such missions, and the usual irretrievability of the satellite once placed into orbit, these factors include budgetary and political as well as quality-assurance and technical elements. The following sections describe these phases and the level of detail expected, usual activities, and the reviews and/or decision gates associated with continuing development of a typical satellite mission.

3.1.1 Conceptual Study Phase This initial phase (sometimes referred to as Pre-Phase A) may be accomplished by the scientific/user community, but usually involves the assistance of an organization such as a NASA center or a spacecraft-manufacturing company that has experience in spacecraft design and development. The phase begins with formulation of a clear statement of the desired objectives of the mission. Associated with these is a justification for the mission describing the benefits that will be achieved. Using descriptions of the instruments or experiments that may gather the desired information, a number of different concepts may be formulated that can achieve the mission objectives and determine the feasibility of conducting the mission. These concepts include ground system and operations as well as launch-vehicle and spacecraft descriptions, but the level of detail is usually "back of the envelope" in scope. After the concept is formulated, a rough order-of-magnitude (ROM) cost for the mission is estimated, usually also using generic cost information as a basis. In addition, preliminary overall schedules and risk assessments are generated.

The results of this phase are reviewed at a "local" level, within the NASA Center or Headquarters Division, university, laboratory, or the like, to determine if the concept warrants further consideration outside the group. Within NASA, if the concept looks promising, an Announcement of Opportunity (AO) may be issued, which is the first step in defining and selecting the instruments or experiments that will fly on the mission if it is fully developed and launched.

3.1.2 Mission Analysis Phase The purpose of the Mission Analysis Phase (Phase A) is to translate the broad mission concepts and objectives into a feasible preliminary system design. This is a refinement and expansion of the Conceptual Study Phase work made with the intention of providing a concise, clear overview of the proposed system. Implementation is done in cooperation with flight ground systems, discipline engineering, and instrument specialists. Financial analysts utilize parametric cost models to derive top-level resource estimates for the overall mission and its component elements. The typical unmanned mission study has a duration of approximately nine months to two years. Study costs are usually one to two percent of total mission costs.

Key activities during this phase are system and subsystem trade studies (comparing different systems to identify the best candidate), analyses of performance requirements, identification of advanced-technology and long lead-time items, risk assessments, end-to-end system life-cycle versus cost studies, estimates of other costs, setting schedules, and selection of system and operational concepts. More specific items addressed include justification for the selection of a specific strategy and preliminary concepts on mission management (particularly if it is not an internal NASA effort).

Specific tasks to support this phase often include feasibility assessments, cost-estimate modeling, identification of spacecraft/ground data communications interfaces and satellite tracking-system interfaces, the design of information-dissemination systems for science data, system engineering for instrument design and their interfaces, definition of alternative strategies for the mission, study of new technology needs, and risk analyses.

3.1.3 Definition Phase The purpose of the Definition Phase (Phase B) is the refinement of the mission and system architecture and the system design created during the Mission Analysis Phase. It is designed to convert the previous phase preliminary system design into a mature final design and baseline for entering the Execution Phase where fabrication begins. Functional, operational, and performance requirements are defined. Established interface requirements and specifications are allocated down to the subsystem or major-component level. Firm cost estimates and schedules are prepared for transition to the subsequent Execution Phase (C/D). Once a detailed baseline configuration that satisfies all the mission and program requirements has been established, the Definition Phase design, costs, and schedules are submitted to NASA for approval. This phase has a duration of from one to two years. The costs associated with this study are usually four to eight percent of total mission costs.

Key activities during this phase are revalidation of mission requirements and system-operation concepts, conversion of mission requirements to much more specific project requirements, risk analyses, system and subsystem studies and trades, updating schedule and life-cycle costs, and system and subsystem concept selection and validation.

This phase covers the full range of technical, management, resource, facility, and procurement assumptions, ground rules, plans, procedures, and full documentation of the entire proposed flight/ground system. Alternative designs are analyzed to allow the choice of one optimum flight and ground-system approach considering technical performance, cost, risks, and schedules. The baseline design is totally defined as well as all major interfaces and system, subsystem, and component specifications.

Specific tasks conducted during this phase include telemetry assessments, analysis of instrument hardware, Space Shuttle interface analyses (as needed), acquiring knowledge of the space environment, determining TDRSS usage, assessment of make or buy decisions of components, setting up production plans, establishing facility requirements, making risk assessments, and offset planning, estimating cost, schedule, and resources, management planning and scheduling, mission requirements assessments, payload integration plans, planning satellite operations, and devising data-transfer and -processing systems.

3.1.4 Design and Development Phase

Usually by this point a commitment has been made to develop and build the spacecraft and implement the mission. For this reason the Design Phase (Phase C) and the Development Phase (Phase D) are combined. The purpose of the Design/Development Phase (Phase C/D) is the actual implementation of the approved mission plan. Now is when the hardware is built, assembled, and tested. The successful end result of this phase is that the satellite will have achieved the desired orbit and is properly oriented in space with all space and ground systems functioning as expected to satisfy previously stipulated mission objectives.

Depending on the mission and/or the organization responsible for it, this phase is sometimes subdivided into separate Design/Development, Launch, Post-Launch checkout, and Operations phases. One example comprises the NOAA weather satellites where the spacecraft is built by either NASA or one of its contractors. (In parallel, the payload complement of meteorological and weather sensors derives from within the agency or is obtained from other contractors expert in this type of space instrumentation. Or, by virtue of a cooperative arrangement with other countries, an instrument may be supplied by a foreign government to fly on the U.S. satellite.) These sensors are then sent to the spacecraft manufacturer for integration and subsequent activities involving the entire space segment. After launch and on-orbit checkout of all systems, the control or Operations phase is transferred from NASA to NOAA.

Even if mission operations are not the responsibility of NASA, the agency may become involved if its shuttle is used to rendezvous with the satellite to perform repair, retrieval, or component- and personnel-exchange functions.

Another example of dividing this phase involves commercial or foreign satellites that are delivered to NASA or DOD sites (usually in Florida or California but sometimes on the shuttle) for Launch-phase purposes only.

The duration of this phase is extremely variable, often ranging from five to ten years depending on its complexity and whether the satellite is manned or unmanned. But subsequent copies of the same or moderately improved satellite system can be placed in orbit within much shorter time periods—sometimes in less than one year from the initial launch.

The major categories of this phase, listed below, are applicable to all mission elements including the space and ground systems:

1. design and development,
2. fabrication,
3. assembly,
4. integration,

5. environmental and other tests,
6. launch,
7. post-launch checkout, and
8. post-launch operations (may include retrieval, on-orbit change or repair, orbit changes, etc.).

Associated with these eight Design/Develop Phase categories are a host of peripheral efforts needed to assure success of the space mission or to determine what went wrong in the event of a failure. These efforts are applicable to the basic spacecraft and to its complement of life-support equipment and scientific instruments, in addition to the ground segment.

During the course of this phase, as the building, test, and assembly of the various hardware elements proceed, a concurrent effort is directed toward making the physical systems act together as a single entity. A partial list of these activities includes the following:

1. Preparing for shipping the fully assembled satellite to the launch site, installing it on the launch vehicle or shuttle, checking all systems for ground or air transit damage, and verifying correct electrical and mechanical interconnections with the launch vehicle and the shroud.
2. Because part of the Control Center is a diagnostic clinic for satellite anomalies in space, preparing and checking emergency short-term remedial control procedures that can be executed from the ground if needed. This effort also includes defining and validating fallback positions, which usually entail compromises in mission objectives.
3. Essentially the same as **2,** except that the activity is geared to anticipating and planning for unexpected changes in environmental conditions such as solar flares, micrometeorite fields, orbital debris, spacecraft outgassing that can blind sensors, and adverse weather conditions especially during the launch trajectory.
4. Establishing and agreeing upon priorities to exercise when anomalies occur in orbit. This is usually a contentious issue, with mission managers, spacecraft manufacturers, and scientists having conflicting priority requirements. For manned missions, the human element, naturally, takes precedence.
5. Planning for the uninterrupted flow of commands and data in the event one or more ground stations or relay satellites fail.

By this time, many of the difficult-to-resolve or -accomplish tasks and backup positions have become routine on prior missions and are not specifically discussed herein. Nonetheless, for every satellite flight, the issues resurface, are studied, and are satisfactorily planned for and resolved as failure modes occur.

The outputs for this phase comprise all aspects of the space and ground segments of the mission and their operation. Prior to proceeding to the Operations Phase, the launch must have been successful, the spacecraft should be deployed and in the correct orbit, the spacecraft and ground-systems communications should be in normal working order, and the spacecraft will have been operating successfully with the tracking stations or via a signal-relay satellite (e.g., TDRSS).

3.1.5 Operations Phase The purpose of the Operations Phase (Phase E) is routine operation of the satellite including data acquisition, processing, and distribution until or unless a problem develops or the mission is terminated. The conduct of this phase varies considerably depending on whether the mission is a principal investigator type, an Earth-observing–facility mission (e.g., Landsat) with many investigators and investigator organizations involved, or one transferred to another federal agency (e.g., NOAA) or the private or international sector. The duration of this phase is variable but can be five years or longer.

Key activities during this phase are command and control of the spacecraft and instruments and data-related functions such as transmission, capture, processing, archiving, and distribution. Special scientific operations may be authorized and conducted. Orbital changes may be required. Spacecraft or instrument components may be retrieved, repaired in orbit, or replaced if necessary and if possible.

During this phase routine tasks conducted include

- science-data reviews with scientists and government officials;

- flight assessments of hardware, software, science data, orbits, and mission objectives;
- flight-system monitoring;
- monitoring the acquisition, processing, dissemination, and archiving of mission data;
- compilation and production of a "lessons learned" document for the mission, which is used as a guide in the development of additional satellites in the series and for future new space missions.

3.2 Satellite Subsystems and Functions

Although satellites bear the imprint of specific designs to support particular missions, they all have certain elements in common, such as

1. a payload that captures an applications-specific objective, such as communications, remote sensing, or navigation, and
2. mission/payload-peculiar spacecraft-bus augmentations that provide special payload-related support functions possibly not offered by a standard spacecraft design. These mission/payload-peculiar support functions include such things as special data downlinks and antennas, data recorders, cryogenic coolers, data encryption, and other subsystems that are directly related to the unique needs of a particular payload.

However, no matter what the payload/mission, the satellite bus must consider the following in its design, and probably will incorporate integral subsystems that perform these functions:

1. power generation and distribution,
2. thermal control,
3. attitude determination and control,
4. orbit maintenance and propulsion,
5. structure and mechanisms,
6. communications, command, control, and data handling.

Each of these areas and the subsystems that perform them are discussed in the following sections. Figure 15 shows the dispo-

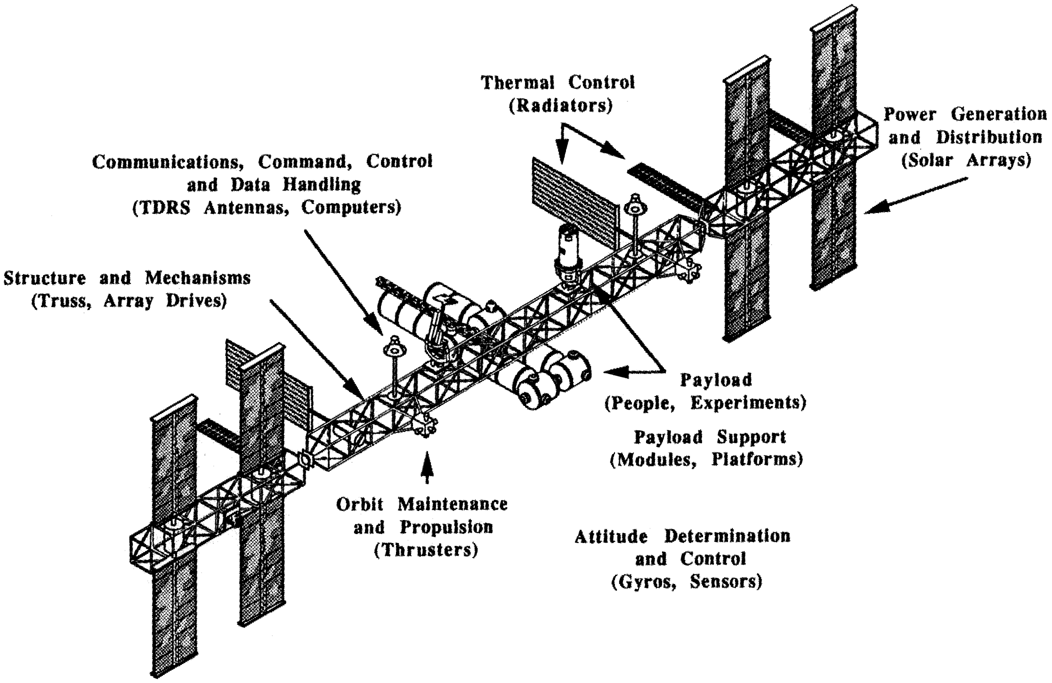
Thermal Control
(Radiators)
Power Generation
and Distribution
(Solar Arrays)
Communications, Command, Control
and Data Handling
(TDRS Antennas, Computers)
Structure and Mechanisms
(Truss, Array Drives)
Payload
(People, Experiments)
Payload Support
(Modules, Platforms)
Orbit Maintenance
and Propulsion
(Thrusters)
Attitude Determination
and Control
(Gyros, Sensors)

FIG. 15. Spacecraft components. All spacecraft have some common elements required to operate in space, as depicted in this early version of the international space station.

sition of the various components in one large satellite structure.

3.2.1 Power Generation and Distribution This subsystem provides the power used by the spacecraft and payload throughout the mission. Solar cells, which directly convert the Sun's light photons to electricity, are by far the most popular source of spacecraft power. Heat exchangers can use the Sun's radiated energy to heat a working fluid that can power a generator to create power. As a spacecraft operates further from the Sun, it may be necessary to use a different source of power such as nuclear-power generators. These systems have been used in deep space for distant planetary exploration, but have also been used in near-Earth orbits for high-power applications. Current space nuclear systems consist of radioisotope thermal generators (RTGs) that use radioactive decay to produce electrical power directly, and nuclear reactors that heat a working fluid to power a generator and produce electricity.

Power storage is achieved by using batteries to store energy. Nearly all solar-powered spacecraft use batteries to provide power during mission or orbit times when the Sun's energy is not available. Other chemical devices, such as the fuel cells used on the Space Shuttle, can be used to provide power over limited-duration time periods.

3.2.2 Thermal Control A spacecraft thermal-control subsystem keeps the temperature of the spacecraft and its mission-important elements within specified ranges. Thermal-control devices are categorized as being either passive or active. Passive devices shield, insulate, or change thermal-exposure characteristics. They include external coatings, thermal blankets, and louvers that selectively close or open spacecraft compartments to space. Active thermal-control devices include heaters, refrigerators, and heat pumps or pipes that use a working fluid to move heat energy from one spacecraft location to another.

3.2.3 Attitude Determination and Control Satellites employ a variety of techniques to control the satellite body axes to some reference orientation. Three-axis attitude-control systems use control torquing methods such as thrusters, spinning momentum wheels, and magnetic devices to maintain a satellite within an acceptable tolerance with respect to pointing references aligned with the center of the Earth, the direction of flight, and a third axis normal to these. Spin-stabilized systems rotate the entire body of the satellite to create a gyroscopically stable orientation with respect to the Earth. The spacecraft may incorporate a de-spun platform, which allows the payload to remain fixed with respect to a particular direction while the rest of the spacecraft spins for stability. Gravity-gradient control methods utilize the forces of gravity to orient the spacecraft.

All of these systems have their particular virtues and limitations. For example, spin-stabilized spacecraft produce a rotisserie effect with a beneficent thermal result. However, the spacecraft usually employ surface-mounted solar cells for power generation, and on a cylindrical spacecraft only about $\frac{1}{3}$ of these cells produce power at any time. Also, spin-stabilized satellites cannot achieve the accurate pointing control required by some remote sensors.

The control torquing methods are directed by a control system that determines the direction, amount of torque, and duration of actuation with respect to reference information provided by devices that can sense the location of the Earth, Sun, or stars. The systems designer must consider the mission/payload requirements and select the proper design from these fundamental approaches or some hybrid combination of these techniques.

3.2.4 Orbit Maintenance and Propulsion External forces, such as drag or solar-radiation pressure, cause the precise orbit desired by the mission to be perturbed. The orbit-maintenance subsystem uses thrusters periodically to restore the satellite to its desired orbit or, in many cases, to modify the orbital characteristics of a satellite (usually a change in the orbital period and associated altitude) to place the satellite into a different time- and space-sampling pattern with respect to specific Earth locales.

Most satellites utilize an integral propulsion system to transfer the satellite from an intermediary parking orbit, achieved by the launch vehicle, to the final desired orbital location. In a few cases, this system may be used to return the satellite to a maintenance orbit where it can be repaired or refueled or, at the end of life, it may be used to de-orbit the satellite from space. In many cases, an integral orbital-maintenance and propulsion system is designed having a variety of different thrusters associated with the relatively large energy requirements of orbital-plane changes and the much smaller energy requirements associated with orbital maintenance and spacecraft attitude control.

3.2.5 Structure and Mechanisms Spacecraft structural considerations must deal with the many physical configurations the craft experiences in achieving and operating in its orbit. The spacecraft must fit into the usually cramped launch-vehicle payload shroud, and survive the process of launch and orbital placement. On station, the spacecraft may need to deploy components into operational configurations, some requiring motors to keep them accurately pointed. Additionally, the structure must have sufficient strength and use the proper materials to maintain spacecraft structural integrity during mission lifetimes that often exceed 10 years of orbital stress produced by the space environment and by station-keeping and operational attitude changes.

3.2.6 Communications, Command, Control, and Data Handling Satellites send information to mission control locations about the spacecraft's state of health and operational status. Status information includes the on-or-off condition of payload and satellite subsystem elements, and health information verifies that all systems are operating within established tolerances of temperature, power, etc., or conversely that they are exceeding preset boundaries. On the basis of this information and the plan for normal or anomalous operation of the spacecraft, the ground control station will transmit commands and other information to the satellite for immediate implementation or for on-board storage for future use.

Most spacecraft are engaged in the collection or transfer of some form of information. In many cases, data rates and volumes are much greater than can be accommodated by the satellite health and status reporting system. When this happens, a separate system of transmitters, antennas, and on-board storage devices for later downlink to ground stations is needed. These special systems are extensive and expensive but are absolutely necessary for satellites that collect high-resolution remote-sensing data or act as a communications relay between widely displaced Earth locations.

Modern spacecraft use computers in nearly every aspect of the satellite system. A computer forms the heart of the attitude-control process, accepting information from the reference sensors and computing the instructions to the control torquing systems. Power-systems management is under the control and guidance of computers, and even the thermal-control system depends on computers to monitor and control component temperatures. Propulsion and orbit-maintenance systems utilize computers to determine the event time and the thrust time and direction for maneuvers, and data-handling and communication systems use computers for downlink timing control and for command authentication and execution. Finally, a master computer is required to determine the occurrence of any serious anomalous event that may affect the survival of the spacecraft, and decide the proper reaction to the event, including assuming control of spacecraft subsystems and instruments, to ensure a safe configuration until the problem can be corrected.

3.3 Satellite Operations

The mission objectives for most unmanned satellites require an operations system that provides a two-way ground-to-space communications link for the command and control of the spacecraft, its systems and subsystems, and the payload complement of instruments.

During the course of normal operations, instruments may have to be commanded into various electronic and physical configurations either routinely or in response to some anomalous situation. Typical commands include power on/partial/off; high/low-gain data-acquisition modes; sensor pointing; and activation of protective measures to prevent overheating or excessive cooling, radiation exposures, and micrometeorite damage. In some instances, the operational configuration of an instrument is varied in real time in response to computer- or scientist-based analysis of sensor data. If a computer is utilized to accomplish an operational change, it may be located on the spacecraft or at the ground control center.

The entire satellite may have to be commanded into different positional orientations during the course of normal operations to compensate for rotational limits of solar arrays (relative to the main body of the spacecraft) that may require a noon yaw turn maneuver to reposition the arrays ready for the next cycle of solar pointing. Pressurized inert gases (argon) that provide jet thrust for correcting naturally induced changes in orbital altitudes or inclinations are periodically released to accomplish these adjustments. As a function of spacecraft orientation or orbital position—both relative to the Earth—antennas and instrument probes may require routine pointing adjustments to facilitate ground communications or permit proper instrument operation.

In a limited number of cases, normal operations may require use of jet thrusters to position the spacecraft for a rendezvous with NASA's Space Shuttle for retrieval, repair, or equipment/personnel replacement purposes. Situations such as this involve complex and very precise orbital maneuvers and spacecraft orientations.

Whatever the operational mode, the main ingredient in its successful accomplishment is the communications link with the ground control center. Because of data-relay satellites, virtually round the clock communications are now possible. The NASA Tracking and Data Relay Satellite System (TDRSS) is also capable of relaying vast quantities of scientific data to NASA ground receiving stations, where it can be retransmitted via commercial communications satellites to users in any part of the world. U.S. weather satellites in geostationary Earth orbits have been retransmitting other NOAA-acquired meteorological data to anyone within the range of the antenna pattern—essentially acting simultaneously as data-relay devices. A typical ground station is shown in Figure 16.

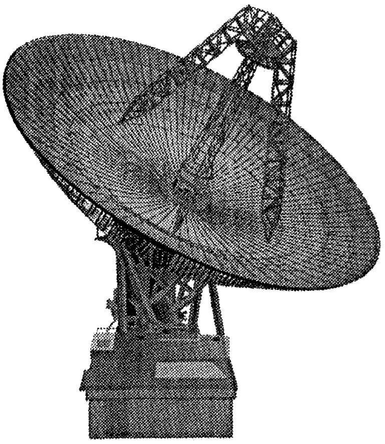

FIG. 16. A typical ground station. Many ground stations require large, movable antennas in order to ensure successful two-way communications with the satellite. The cost of these stations and their operation cannot be ignored in initial mission planning.

In addition to the NASA control centers, ground data-receiving facilities, and TDRSS, other U.S. institutions (e.g., universities; local, regional, state, and federal agencies) and foreign countries have installed their own data-receiving stations. For example, Landsat data are now routinely acquired by more than a dozen countries, various universities, U.S. government agencies, and private organizations. These facilities can process imagery data suitable for use by Earth scientists and environmentalists worldwide.

3.4 Space Law

The placement and operation of a satellite in a space environment requires satisfying seven basic "laws" related to outer space of the U.N. Committee on the Peaceful Uses of Outer Space (COPUOS). Also considered part of international law are a number of U.N. requirements mostly pertaining to notifying that international body of launch intentions; operational, debris contribution and disposal plans; and untoward occurrences. Their exact titles are as follows:

1. Treaty on principles governing the activities of states in the exploration of outer space, including the Moon and other celestial bodies (1967).
2. Agreement on the rescue of astronauts, the return of astronauts, and the return of objects launched into outer space (1968).
3. Convention on international liability for damage caused by space objects (1972).
4. Convention on registration of objects launched into outer space (1975).
5. Agreement governing the activities of states on the Moon and other celestial bodies (1979).

6. Treaty banning nuclear-weapons tests in the atmosphere, in outer space, and under water (1963).
7. Convention on the prohibition of military or any other hostile use of environmental-modification techniques (1977).

It is to be noted that not all the 154 U.N. member countries subscribe to each of these laws. For example, there is no general agreement on the definition of outer space—one commonly accepted boundary defines this domain as the region 100 km or more from the Earth's surface. But several equatorial countries claim that the country's sovereignty extends from the surface to infinity, thereby establishing their boundaries to include geostationary satellites flying overhead.

The U.N. requires notice from the launching state if a satellite is no longer in orbit, particularly if it may reenter the atmosphere and crash. If damage to life or property in another country occurs, the launching state is liable. The country where the impact occurs must search for and return the debris to the launching state, but the latter can be billed for expenses incurred. In the event astronauts are involved, the country must care for them and promptly return them to the launching country. Satellite crew members may not be detained for interrogation purposes.

If a satellite causes interference of any type (electromagnetic, radioactivity, etc.) with another satellite, its owner is liable for correcting the situation. At times orbiting optical devices have been considered suspicious if the instrument can be focused on a hostile country for espionage purposes. Situations of this nature apply to all Earth-observing satellites, but their use remains legal.

With regard to interplanetary exploration, the U.N. prohibits any activity or landing that may affect the well-being of the target planet. Contaminants, radioactive sources, thermonuclear detonations, and electronic interferences fall in this category of restrictions, as do environmental modifications. In addition, no Earth state can establish a sovereign claim on any extraterrestrial body.

In the not too distant future, it is expected that countries that launch and then abandon satellites and other debris in space must take action, still to be precisely defined, to prevent their reentry to Earth or posing a collision problem for operational satellites in orbital flight.

In summary, international laws are intended to work in the best overall interests of mankind. Compliance does add to the cost of space exploration, but only in the near term. The long-range benefits of these agreements to space exploration far outweigh the immediate inconveniences and expenses.

4. ARTIFICIAL SATELLITE MANUFACTURING

Several major institutions manufacture satellites in the United States and abroad. NASA's Goddard Space Flight Center in Maryland has all the facilities required to design, manufacture, build, test, launch, and operate scientific spacecraft. Another similar U.S. government institution is the Naval Research Laboratory in Washington, DC. U.S. commercial manufacturers include Lockheed/Martin (a merger of Lockheed with Martin Marietta, which had previously acquired the satellite facilities of GE Astrospace, formerly RCA Astrospace), Loral, TRW, and Hughes, to name a few. Foreign manufacturers include Aerospatiale, Alcatel, and Matra Marconi (France), Daimler-Benz Aerospace (Germany), Alenia Spazio (Italy), and SPAR Aerospace (Canada). There are also several small satellite manufacturers such as Orbital Sciences Corp. and CTA among others. In addition, there are several academic facilities capable of manufacturing satellites such as the Johns Hopkins University's Applied Physics Lab and the California Institute of Technology's Jet Propulsion Laboratory.

In most cases, satellite builders will obtain many of the components used in the satellite from specialized component manufacturer companies. Such companies exist for solar arrays, batteries, attitude-control sensors and systems, and most other "standard" spacecraft components.

Despite the plethora of manufacturers, many of the procedures and processes of building satellites are quite similar. In order to build and test a spacecraft properly, some very specialized equipment and facilities are required. One of the most demanding requirements is cleanliness. Contaminants can cause valves to stick, electronics to short, moving parts to wear, and sensors to be blinded. For this reason, space components are manufactured using strictly controlled procedures to minimize introduction of foreign matter. The satellites themselves are assembled in "clean rooms" (Fig. 17), which maintain a clean environment (see AEROSOLS) in a large enough area to assemble a complete spacecraft. Before being brought into a clean room, space-flight components and even tools are cleaned, usually using an evaporative solvent. Within the clean room, air is constantly circulated and filtered to keep floating particulates down to a specified level. Personnel who work in the clean room are required to wear special clothing, including booties for their feet and hats over their hair to minimize contaminants from their bodies. Gloves are always used to keep body oils off flight components.

Once a spacecraft is assembled it is tested to ensure that it was assembled correctly, that it will survive the trip into space, and that it will operate correctly when it arrives on station. Comprehensive tests exercising all of the spacecraft and instrument systems are performed both in ambient conditions within the clean room, and under conditions simulating the anticipated space environment. In order to simulate the space environment, large thermal/vacuum chambers (Fig. 18) that can contain the entire spacecraft are used. These chambers use pumps and cryogenics to reach the almost zero pressure and extremely low temperatures that will be experienced by the satellite. Pass-through connectors must be built into the walls of the chamber to allow external test connections to the spacecraft for power and data transfer.

To ensure that a satellite will survive the forces encountered during the rocket flight into space, the entire spacecraft is placed on a vibration table and "shaken" in all three axes up to and slightly beyond the levels that will be experienced. In another room, the spacecraft is subjected to low-frequency sound waves at decibel levels simulating the noises generated within the rocket's payload shroud during launch. Other pre-launch tests include electromagnetic interference/electromagnetic compatibility (EMI/EMC) tests to ensure that no spacecraft component emits any spurious signals or will not perform correctly if subjected to any expected signals. This test requires a large dedicated room covered on the inside with anechoic materials to absorb these radiations and prevent reflected signals.

Throughout this process, quality-control inspectors keep watch to ensure that proper procedures are being followed. This includes ensuring that the parts used to construct the satellite are of the proper quality and have been tested to prove adequate performance. In most cases, components and subsystems

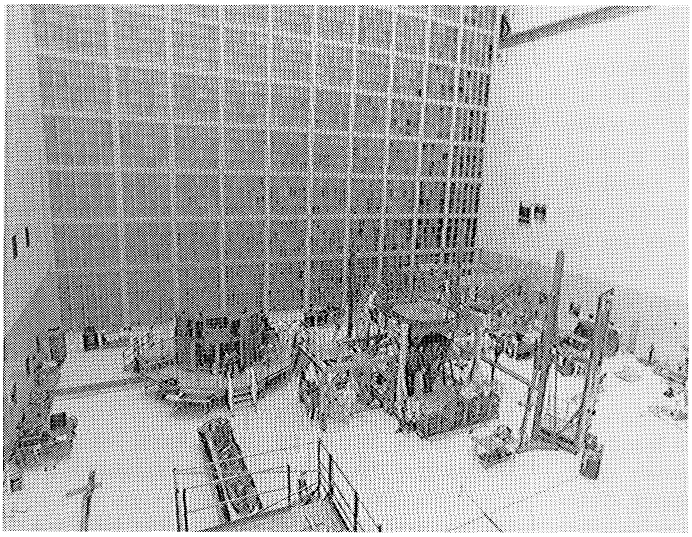

FIG. 17. Clean room. Notice the wall of filters used to keep the air in the room clean.

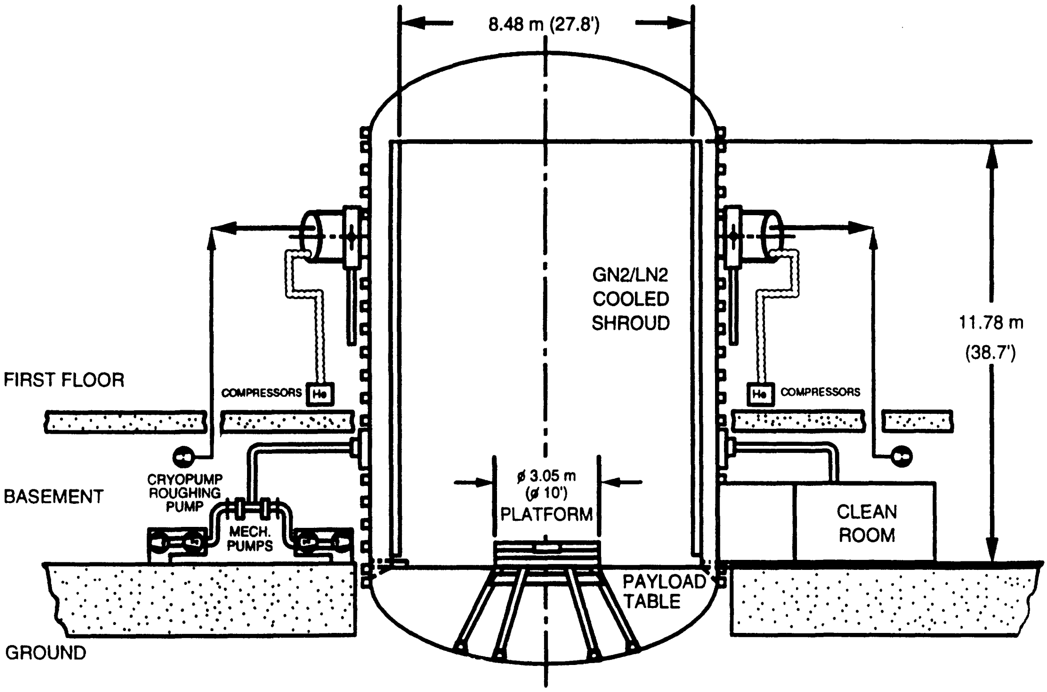

FIG. 18. Thermal vacuum chamber. Large chambers like this are required to accommodate a complete spacecraft. The chambers utilize pumps and cryogenics to expose the spacecraft to the expected environment it will experience in space.

have already been subjected to the same types of tests that the completed spacecraft will be subjected to after assembly.

5. ECONOMIC ASPECTS OF ARTIFICIAL SATELLITES

As mentioned several times previously, satellites represent an extremely large investment. A typical communications satellite may cost up to \$50 × 10^6 to buy and as much or more to place into space. Satellites also require ground stations to communicate with and control the satellite. If users are involved, remote ground equipment is also required. The GPS system mentioned earlier consists of a constellation of 24 satellites to ensure continuous positioning capability worldwide. This system represents an investment of many billions of dollars to establish, and will continue to cost taxpayers hundreds of millions to maintain and replenish satellites as they fail. The Hubble Space Telescope cost over 10^9 dollars and is scheduled for repair missions from the Space Shuttle (which itself costs around \$500 × 10^6 per flight) every two to three years. The Earth Observation System (EOS) is projected to cost over \$7 × 10^9 for the constellation of satellites that will monitor the effects of man on the Earth's environment, and the International Space Station program will cost over \$30 × 10^9.

As one can imagine, competition to supply hardware and services for these space endeavors among satellite manufacturers is intense. And while the rewards may be great, the risk is also great considering that, with the exception of communications satellites and other repeated satellite designs, each system is usually new and unproven. Since, in most cases, the satellite cannot be repaired once in space, systems must be robust and redundant or overdesigned. However, as this adds to weight and costs, the amount of robustness must be considered in design trade-offs. Usually, a particular launch vehicle is chosen at the start of design, and having to change to a larger vehicle because of increased weight may require some redesign that may cause the program to exceed funding limits. This could result in a delay of mission start, compromise of mission objectives, or even termination of the entire project, all of which have occurred in the past. Though launch vehicles are generally very reliable, insurance rates can be as high as 25% of the total cost of the satellite.

However, considering the benefits realized in communications, entertainment, navigation and national security, weather and environmental monitoring, and the general increase in knowledge of the Earth, the Universe, and our place in it, the costs associated with the use of artificial satellites are more than justified.

GLOSSARY

a: Orbit semimajor axis.

AMAO: Advanced Missions Analysis Office at NASA's Goddard Space Flight Center.

AO: Announcement of Opportunity.

COBE: Cosmic (Ray) Background Explorer.

Comsat: Communications Satellite Corporation.

COPUOS: U.N. Comm. on Peaceful Uses of Outer Space.

DMSP: Defense Meteorological Satellite Program.

DOD: Department of Defense.

e: Orbit eccentricity.

EMC: Electromagnetic compatibility.

EMI: Electromagnetic interference.

EOS: Earth Observation System.

ERTS: Earth Resources Technology Satellite.

ESA: European Space Agency.

EVA: Extravehicular activity.

GEO: Geostationary Earth orbit.

GOES: Geostationary Operational Environmental Satellite.

GPS: Global Positioning System.

GRO: Gamma Ray Observatory.

GSFC: Goddard Space Flight Center.

HEAO: High Energy Astronomy Observatory.

HEO: High-altitude Earth orbit.

HST: Hubble Space Telescope.

i: Orbit inclination.

ICBM: Intercontinental ballistic missile.

Intelsat: International Telecommunications Satellite Consortium.

ISEE: International Sun-Earth Explorer.

IUE: International Ultraviolet Explorer.

LDEF: Long Duration Exposure Facility.

LEO: Low-altitude Earth orbit.

NASA: National Aeronautics and Space Administration.

NOAA: National Oceanographic and Atmospheric Administration.

OAO: Orbiting Astronomical Observatory.

OSO: Orbiting Solar Observatory.

Phase A: Analysis phase of mission design.

Phase B: Definition phase of mission design.

Phase C/D: Execution phase of mission design.

Phase E: Operations phase of mission.

Pre-Phase A: Conceptual study phase of mission design.

ROM: Rough order of magnitude.

RTG: Radioisotope thermal generator.

SAR: Synthetic-aperture radar.

SMM: Solar Maximum Mission.

SPOT: Satellite Probatoire d' Observation de la Terre.

T: Orbit period.

TDRS: Tracking and Data Relay Satellite.

TDRSS: TDRS system.

TIROS: Television Infrared Observation Satellite.

TOMS: Total Ozone Mapping Spectrometer.

U.N.: United Nations.

UARS: Upper Atmosphere Research Satellite.

UV: Ultraviolet.

$\boldsymbol{v}$: True anomaly.

$\boldsymbol{\omega}$: Argument of perigee.

$\boldsymbol{\Omega}$: Right ascension of the ascending node.

$\boldsymbol{\mu}$: Gravitational parameter.

Further Reading

Allahdadi, F. (Ed.) (1993), *Space Debris Detection and Mitigation,* SPIE Proceedings No. 1951, Bellingham, WA: SPIE.

Bate, R., Mueller, D., White, J. (1971), *Fundamentals of Astrodynamics,* New York: Dover.

Bilstein, R. (1989), *Orders of Magnitude, A History of the NACA and NASA, 1915–1990,* Washington, DC: U.S. Government Printing Office.

Clark, P. (1988), *The Soviet Manned Space Program,* New York: Orion Books.

Dutton, L. (1990), *Military Space,* London: Brassey's.

Feher, K. (1983), *Digital Communications, Satellite/Earth Station Engineering,* Englewood Cliffs, NJ: Prentice-Hall.

Flight Projects Directorate (1994), *Project Management Handbook,* Greenbelt, MD: NASA/Goddard Space Flight Center.

Flury, W. (Ed.) (1993), Special issue on Space Debris, *Adv. Space Res.* **13** (8).

Griffin, M., French, J. (1991), *Space Vehicle Design,* Washington, DC: American Institute of Aeronautics and Astronautics.

Jasentuliyana, N. (Ed.) (1992), *Space Law: Development and Scope,* Westport, CT: Praeger.

Kerrod, R. (1988), *The Illustrated History of NASA,* Anniversary Ed., New York: Gallery Books.

Logsdon, T. (1992), *The Navstar Global Positioning System,* New York: Van Nostrand Reinhold.

McElroy, J. (1986), *Space Science and Applications,* New York: IEEE.

National Research Council Committee on Space Debris (1995), *Orbital Debris: a Technical Assessment,* Washington, DC: National Academy Press.

Pisacane, V., Moore, R. (1994), *Fundamentals of Space Systems,* New York: Oxford Univ. Press.

Schnapf, A. (1985), *Monitoring Earth's Ocean, Land, and Atmosphere from Space—Sensors, Systems and Applications,* Washington, DC: American Institute of Aeronautics and Astronautics.

Simpson, J. (Ed.) (1994), *Preservation of Near-Earth Space for Future Generations,* Cambridge, England: Cambridge Univ. Press.

Tascione, T. (1984), *Introduction to the Space Environment,* Colorado Springs: U.S. Air Force Academy.

Wertz, J., Larson, W. (1991), *Space Mission Analysis and Design,* Boston: Kluwer.

Winkler, L. (1990), "Legal Aspects of Astronomy," *Griffith Observer* **54** (8), 2–9.

SCANNING ACOUSTICAL MICROSCOPY

BUTRUS T. KHURI-YAKUB, *Edward L. Ginzton Laboratory, Stanford University, Stanford, California, U.S.A.*

GLEN WADE, *Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, California, U.S.A.*

ALBERT C. WEY, *Sonoscan, Inc., Bensenville, Illinois, U.S.A.*

Introduction 545
1. **Historical Perspectives** 546
2. **Scanning Laser Acoustic Microscope (SLAM)** 546
2.1 SLAM Configurations 547
2.1.1 Solid-Surface Systems 547
2.1.2 Holography 547
2.1.3 Tomography 548
2.2 Systems Considerations 548
2.2.1 Laser-Beam Readout 548
2.2.2 Scanning Techniques 549
2.2.3 Detection Techniques 549
2.2.4 Configuration of Elements 551
2.2.5 Numerical Aperture 551
2.2.6 Resolution 552
2.2.7 Sensitivity 552
2.3 Applications 553
2.3.1 Biomedical 553
2.3.2 Nondestructive Evaluation 553
2.3.3 Materials Technology 553
2.3.4 Electronic Components 553
3. **Scanning Acoustic Microscope (SAM)** 554
3.1 Principles of Operation 554
3.2 Transducer, Lens, and Test Media 556
3.2.1 Transducers 556
3.2.2 Buffer Rod 556
3.2.3 Lenses 556
3.2.4 Matching Layers 556
3.2.5 Operating Media 557
3.3 System Configurations 558
3.4 Scanning 559
3.5 Response of the SAM 560
3.6 Imaging Applications 562
3.6.1 Biological Applications 562
3.6.2 Materials Science 563
3.7 Metrology 565
3.7.1 Amplitude Only 565
3.7.2 Amplitude and Phase 566
4. **Near-Field Scanning Acoustic Microscope (NFSAM)** 567
4.1 Tip NFSAM 568
4.2 Pinhole NFSAM 569
5. **Conclusions** 569
6. **Acknowledgments** 569
Glossary 569
Works Cited 570
Further Reading 571

INTRODUCTION

Scanning electron microscopes, optical microscopes, scanning tunneling microscopes, and atomic-force microscopes are some of the imaging tools that have shown us, and taught us, a lot about the structure of materials in a myriad of fields and applications. Mostly, it is the surfaces of materials that have been the field of view of all these instruments, as none can show structure below the surface, especially in opaque materials. More importantly, the mechanical properties of the material and their variations are not the object of imaging of any of the above-named instruments. They measure changes in the index of refraction or topography of a material surface. Acoustic microscopes have been developed to answer a need for visualizing variations in the mechanical properties of a sample, especially below the surface. The scanning laser acoustic microscope, a transmission-imaging instrument, images the interior of samples by visualizing the shadow cast by material-property variations in the path of the acoustic beam. The scanning acoustic microscope, a reflection-imaging instrument; images the surface and near-surface variations in the mechanical properties of a sample. The near-field scanning acoustic microscope can be either a reflection or a transmission instru-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ment; however, its resolution is determined by the size of an aperture or a tip that is used to transmit and/or receive the signal. This article describes the principles of operation of these three instruments and shows some of the many applications where they have found a home.

1. HISTORICAL PERSPECTIVES

The ability to use sound for "seeing" has long been the dream of technologists throughout the world. A number of scientists and engineers have contributed ideas that have played a part in the realization of that dream. Some of the early thinkers in this regard include Leonardo da Vinci, Daniel Colladon, and Charles Sturm on passive sonar, and Lord Rayleigh, O. P. Richardson, Hiram S. Maxim, Paul Langevin, and M. C. Chilowski on active sonar (Lasky, 1977; Wade, 1987). S. J. Sokolov invented a variety of clever instruments to "see" with sound including a liquid–surface system and a bulk-diffraction system for detecting flaws in metal test pieces. In both systems the readout was accomplished by means of coherent light. He was also interested in acoustic microscopy and proposed an "ultrasonic microscope" (Sokolov, 1949) for using sound at 3 GHz where the wavelength would be short enough to make the instrument capable of resolving truly minute objects. Reimar Pohlman, a German contemporary of Sokolov, proposed the employment of a unique type of cell for instantaneous ultrasonic imaging where the image information is read out by reflected light (Pohlman, 1937).

SLAM was developed by Korpel, Kessler, and colleagues in the early 1970s (Korpel and Kessler, 1971; Korpel *et al.*, 1971). The research on SAM started with Calvin F. Quate and his colleagues in 1973 (Lemons and Quate, 1974). Since then, researchers throughout the world have joined in a sizable and effective international study of their potentialities.

Early demonstrations of near-field imaging were done using microwaves (Ash and Nichols, 1972), acoustics (Durr *et al.*, 1980), and optics (Pohl *et al.*, 1984). More recently, scanning tips such as those used in scanning tunneling and atomic-force microscopes have been used to image surfaces and variations in the mechanical properties of surfaces with atomic resolutions. These instruments are gaining in acceptance and applications not unlike the SLAM and SAM.

2. SCANNING LASER ACOUSTIC MICROSCOPE (SLAM)

SLAM works by introducing a continuous plane wave of ultrasound at frequencies up to several hundred megahertz into the bottom surface of a microscopic object as shown in Fig. 1. A fluid coupling (water in the figure) is required to bring the ultrasound to the object. As the ultrasound travels through the object, internal structural variations differentially attenuate and scatter it, resulting in a nonuniform wave pattern at the exit side of the object that corresponds to the localized acoustic properties. This pattern is duplicated onto a detector plate (labeled the coverslip in the figure), which is basically a metallized plastic block forming an optical mirror. The coverslip is placed in

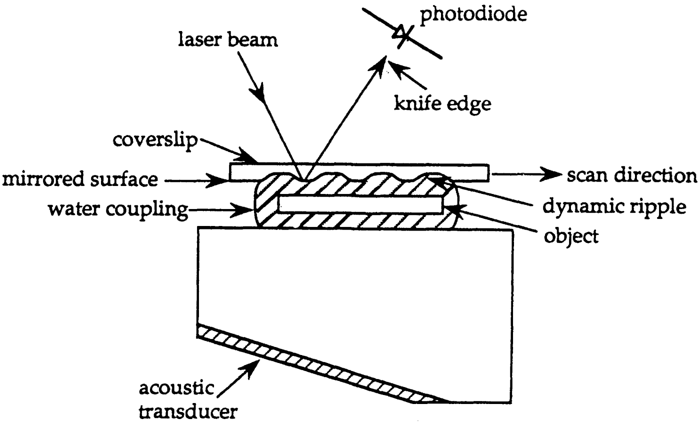

FIG. 1. Schematic diagram of the scanning laser acoustic microscope (SLAM).

close proximity to the sample and is acoustically coupled with fluid. The mirror becomes slightly distorted by the pressure of the ultrasound at each point. A laser beam is focused onto the bottom surface of the coverslip to detect the ultrasonic pattern by measuring the minute distortions of it. The reflected laser signal is converted into a video signal, and the acoustic image appears on a cathode-ray tube (CRT) screen. By rapid sweeping of the laser beam, each SLAM image is created in $\frac{1}{30}$ of a second, thus allowing true real-time inspection of the object (Kessler, 1974).

Key concepts associated with SLAM operation include point-by-point optical scanning, dynamic-ripple diffraction, real-time imaging, scanning laser-beam probe, and synchronously scanned television monitor. These concepts must be taken into account in analyzing SLAM's performance. They include classical laws of diffraction as applied to both light and sound and make use of mathematical tools worked out by such pioneer classicists as Rayleigh, Abbe, Michelson, Kirchhoff, and Fresnel. Also needed, however, in theoretical treatments of SLAM are more modern methods of systems analysis such as transfer-function theory, convolution, impulse response, sampling theory, and other methods associated with the communication-systems–theory point of view. First, we consider the different types of SLAM systems; then we address the issues relating to the performance of these systems.

2.1 SLAM Configurations

2.1.1 Solid-Surface Systems The laser beam in the SLAM impinges upon a solid surface that can support a moving ripple if struck by a continuous wave of ultrasound. The ultrasonic waves first are scattered by internal anomalies in the object. These waves then cast a nonuniform wave pattern onto the coverslip. As the waves strike the lower surface of the coverslip (see Fig. 1), they generate a moving ripple pattern on that surface. This is because the surface is dynamically distorted by particle displacements in the incident and reflected ultrasonic beams. The distortion pattern moves across the surface with velocity $V/\sin q$, where V is the sound velocity in the liquid and q the incident angle. SLAM uses a rapid-scanning laser beam to read out the information carried by the surface perturbation on the coverslip. The use of a scanning laser beam has a number of attractive features. For example, it makes it possible to discriminate between dynamically and statically scattered light because the former is Doppler shifted by the ultrasound frequency. Also, because the electrical output is at the ultrasonic frequency, electronic image processing in phase and amplitude is readily accomplished.

2.1.2 Holography The most essential characteristic of a hologram is that it records both phase and amplitude information. In contrast, SLAM, as ordinarily used, records only the amplitude. To perform holographic reconstruction, SLAM must be modified to acquire and utilize the phase information as well. When both the phase and the amplitude are employed properly in reconstructing the image, we obtain a hologram. A commercially available instrument that does this has been named HOLOSLAM.

To construct HOLOSLAM, a quadrature detector is incorporated into the circuitry of a conventional SLAM so that it can obtain the amplitude and phase information necessary for holographic reconstruction. Quadrature detection is accomplished by multiplying the detected signal with two coherent electronic reference signals, the phase of one being shifted by 90° with respect to that of the other. The two outputs thus obtained represent the real and imaginary parts of the complex amplitude of the ultrasound detected by the laser beam.

The two competing images (i.e., the real and virtual images), ordinarily present in conventional holographic reconstruction, do not both need to be reconstructed with this approach. The two bands of signal, representing respectively the positive and negative spatial-frequency components, are both processed by quadrature detectors. Because of this, HOLOSLAM has four output ports as shown in the schematic diagram of Fig. 2.

The entire operation is summarized as follows: A laser beam scans across the shiny dynamically rippled surface as in the case of SLAM. The angular displacement of the beam is detected by a knife-edge and photodiode combination, whose output is an electronic signal carrying the spatial information of the ultrasound wave field. This electronic

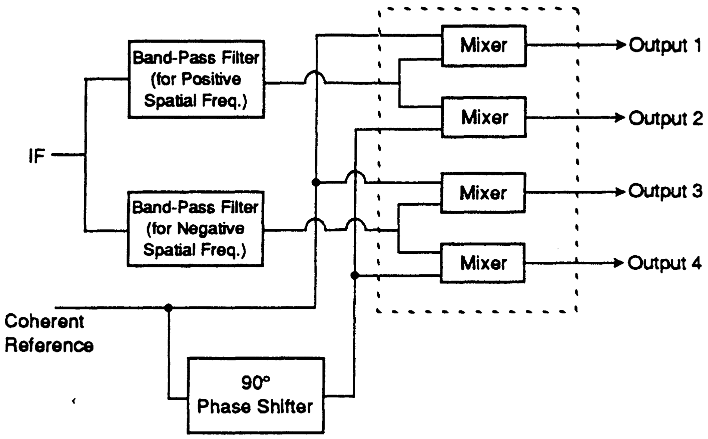
Band-Pass Filter (for Positive Spatial Freq.)
IF
Band-Pass Filter (for Negative Spatial Freq.)
Coherent Reference
90° Phase Shifter
Mixer
Mixer
Mixer
Mixer
Output 1
Output 2
Output 3
Output 4

FIG. 2. Schematic diagram of the electronic circuitry for the HOLOSLAM.

signal is down-shifted to an intermediate frequency (IF). As indicated in the figure, the IF signal is then band-pass filtered by two filters to provide two signals corresponding respectively to the positive and negative spatial-frequency variations. The two signals are then processed by two quadrature detectors to produce four output signals in the video-frequency range. These output signals constitute the holographic data that are then used in straightforward fashion for holographic image reconstruction (Lin *et al.,* 1987).

2.1.3 Tomography To enhance the performance capabilities of the SLAM, tomographic principles as well as holographic principles are used separately or simultaneously. One technique for applying the former principles is called scanning tomographic acoustic microscopy (STAM). As in the case of HOLOSLAM, STAM utilizes digital signal processing of image information obtained with a modified SLAM. It is capable of producing authentic tomograms (cross-sectional images) and thus of overcoming the ambiguities associated with SLAM images when the microscopic objects are thick and structurally complex.

STAM is most suitable for the subsurface imaging of objects with planar structure. Using plane-wave insonification, the modified SLAM performs the data acquisition, after which digital processing of these acquired data accomplishes the image reconstruction. For planar geometry, STAM image reconstruction can be efficiently accomplished through the use of the so-called back-and-forth propagation (BFP) algorithm (Chiao and Lee, 1991). If this algorithm is used only once, that is, for a single viewing angle, the reconstruction is purely holographic. That is the case in HOLOSLAM. However, by applying the BFP algorithm over and over, that is, for a number of viewing angles, the BFP algorithm will produce a tomogram. Simulation and experimental results both show that STAM images can be significantly better, that is, more highly resolved, than images from either SLAM or HOLOSLAM.

Since the resolution improvement of STAM is directly related to the detection of the phase component of the sound field, STAM image reconstruction is sensitive to phase errors in the tomographic projections. If phase errors are present, the resolving power of the microscope may be reduced and unwanted artifacts may show up in the image. To correct for phase errors we can estimate the effect of these errors in the image reconstruction and calculate the correction to obtain the final image (Chiao and Lee, 1991). Operational results confirm the validity of this correction technique.

2.2 Systems Considerations

2.2.1 Laser-Beam Readout SLAM employs rapid optical scanning for detecting the acoustic fields point by point. This provides a number of important features. For example, since the sound-field parameters are ultimately translated into electrical signals, electronic filtering may be utilized to reject unwanted background such as noise. With these signals it is easy to choose between conventional imaging and holographic imaging simply by changing between linear detection and phase detection. Both lasers and laser-beam scanning devices are readily available, and the construction of an acoustic imaging system using them is relatively straightforward. Finally, with the appropriate devices and TV-compatible circuitry, real-time imaging is possible.

2.2.2 Scanning Techniques A key element in the SLAM scanning mechanism is an all solid-state acoustic-optic (A-O) light deflector. Bragg diffraction of laser light by acoustic waves produces horizontal deflection similar to that in a television screen and permits continuous scanning at rates up to 20 kHz or greater. The constant rate of angular change that characterizes a television scan permits the use of a wide optical aperture, leading to a small spot size.

The A-O light deflector is manufactured from high-quality flint glass as the interaction medium with indium-bonded lithium-niobate piezoelectric transducers for generating the acoustic waves. These waves are frequency modulated with a sawtooth function. A light beam entering the acoustic field at nearly normal incidence is deflected at an angle proportional to the acoustic frequency.

A configuration for up-shifted Bragg diffraction is illustrated schematically in Fig. 3. When Bragg-diffraction conditions are satisfied, the beam is deflected as if it were being reflected off the moving wave fronts of the sound. The phenomenon is reminiscent of the way Bragg-diffracted x rays are deflected off atomic planes in a crystal. A sound column acts, in essence, like a continuous distribution of "reflectors," and the optical rays that are reflected off successive crests add up constructively.

A frequency deviation of 40 MHz centered at 70 MHz will deflect a laser beam of wavelength 633 nm through an angle of 6.5 mrad centered at 11.4 mrad from the undeflected beam. Since the deflector utilizes phased-array beam steering, the deflected light will have nearly constant intensity across the deflection angle.

Resolution depends on the time–bandwidth product, which is determined by the length of the optical aperture and the bandwidth of the transducer. The actual aperture of the A-O deflector is 40 mm, which represents a transit time of 10 μs. With a 40-mm aperture and a transducer bandwidth of 40 MHz, 400 spots of resolution are achieved. The transit time is the time it takes to clear the sound from the optical aperture and is thus the switch time in a random-access, 400-spot system. A tradeoff exists between resolution and access time—the smaller the beamwidth, the faster the access time, with correspondingly less resolution.

2.2.3 Detection Techniques In current commercially available SLAMs, the amplitude of the scattered acoustic waves from the object is detected by a knife-edge and photodiode combination. Figure 1 illustrates the setup. A focused laser beam is directed toward the shiny surface upon which impinge the scattered acoustic waves, having been spatially modulated by the object. The angular position of the light reflected from that surface is modulated in turn by the dynamic activity of the corrugations produced by the acoustic waves. The photodiode, which detects the power carried by the light in the beam, is not, by itself, sensitive to the angular modulation of the beam. However, an appropriately positioned obstacle, such as a knife edge, will block a fraction of the light

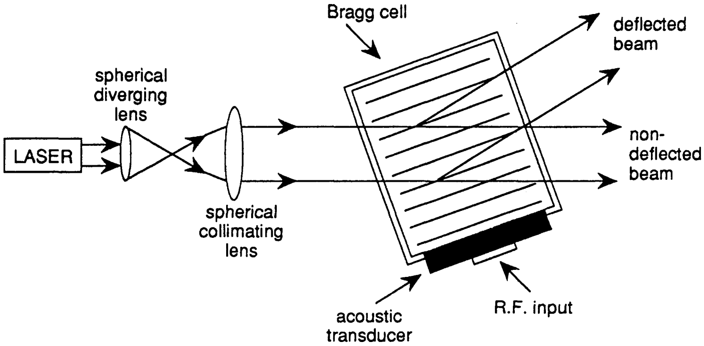

FIG. 3. Acoustic-optic light deflection using Bragg diffraction of laser light by acoustic waves.

depending upon the instantaneous position of the beam. Thus the knife edge demodulates the angular position fluctuations into corresponding intensity fluctuations. The knife-edge technique is very simple and is quite adequate in providing the detection capability needed by conventional SLAMs.

The transfer function for this approach, shown in Fig. 4, is antisymmetric in its response to spatial frequencies with motion in the scan direction. For negative spatial frequencies (corresponding to sine waves traveling on the surface in the opposite direction), the detected signal is opposite in phase to that for positive frequencies. For spatial frequencies perpendicular to the scan direction, the transfer function is symmetrical as shown (Rylander, 1982).

In ordinary SLAM operation, the insonifying wave is directed at an oblique angle of incidence against the object rather than at normal incidence. This is done because the scattering of a wave at normal incidence from an object generally produces a spatial-frequency spectrum on the shiny surface that would be centered at the origin and have its largest components there. From the transfer-function plot, we can see that such an object would not be effectively detected by the knife-edge system, at least as far as the waves traveling perpendicularly to the knife-edge direction (and hence in the scan direction) are concerned. For those waves the transfer function at the origin is zero. However, the spatial-frequency spectrum of the scattered waves may be easily biased about a finite, nonzero spatial frequency by employing obliquely (rather than normally) incident insonification. This is done in SLAM in order to avoid asymmetrical response about the spatial frequency corresponding to the bias; one sideband of the scattered spectrum is filtered out. This is quite suitable for real objects because of their generally symmetrical sidebands. Single-sideband detection with a synchronous detector can thus be used to reconstruct the images of such objects.

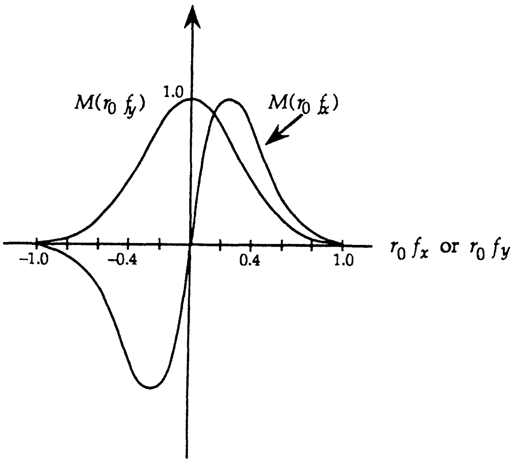

FIG. 4. Normalized transfer functions $M(r_0 f_y)$ and $M(r_0 f_x)$ of the knife-edge detector oriented perpendicular to the scan direction of the laser beam (i.e., in the x direction). The beam is assumed to have a Gaussian intensity profile. r_0 is the nominal spot radius and f_x, f_y are spatial frequencies.

The approach for image reconstruction must be somewhat different in HOLOSLAM and STAM. With these two microscopes the acquired signal constitutes the data to be digitally processed for holographic or tomographic image reconstruction. For example, the frequency spectrum corresponding to the acquired data is normalized by using inverse filtering or some other signal-processing technique to compensate for the nonuniform response of the knife-edge approach.

Finally, it should be pointed out that the peculiar shape of the knife-edge transfer function has a consequence in terms of resolution. In order to avoid the null response at the zero spatial frequency, the bandwidth of the spectrum of the detected acoustic wave field has to be limited so that it does not extend beyond the positive part of the transfer function. This, of course, brings about a limitation in the resolution of the reconstructed image.

Because of the above limitation, other, more complicated types of detection have been proposed for these microscopes. One such detection scheme uses a form of homodyne detector called the time-delay interferometric detector (TDID) (Rylander, 1982). The transfer function for the TDID is symmetric and has the maximum response for zero frequency. The detected acoustic wave field can have a greater bandwidth of the spatial-frequency spectrum with the TDID than with the knife-edge detector. Hence, higher resolution of the reconstructed image should be possible. The TDID is not currently being employed with these microscopes because the knife-edge detector is much simpler to implement.

2.2.4 Configuration of Elements Figure 5 shows all essential optical components of a typical scanning laser module in SLAM. A combination of a cylindrical lens, L1, and a spherical lens, L2, is used to shape the laser beam to make it suitable for the deflector aperture, which is 40 mm wide by 2 mm high. The laser beam is expanded horizontally by a 20× telescope (L1 and L2) to fill the 40-mm–wide aperture of the deflector. The lens L2 also vertically compresses the laser beam to 1 mm in height. After passing through the A-O deflector, the laser beam is compressed horizontally by a 0.125× telescope (L3 and L4) and vertically collimated by a lens, L3. Only the deflected laser beam is allowed to exit the scanning laser module, with a spot size of 2.5 mm in diameter. Finally, the beam is deflected vertically by a galvanometer-type mirror, driven by a current of appropriate wave form.

A knife-edge detection setup is employed to detect the periodic deflection of the probing optical beam caused by the acoustic surface perturbation of the coverslip. The knife edge lies in the Fourier-transform plane of the scanning laser module (i.e., the focal plane of L5) and therefore in the image plane of the scanner. This means that the laser beam always strikes the knife edge and the photodetector in the same place regardless of the scan angle. Conventional optical magnification may be inserted into the optical path in order to produce simultaneously a focused spot size appropriate to the resolution requirement for a given ultrasonic frequency and an optimal field of view.

2.2.5 Numerical Aperture SLAM's sensing system involves a mirror on the coverslip and a knife-edge detector. The characteristic for this system as a function of the angle of ultrasound incidence determines the numerical aperture (NA) of the system. The

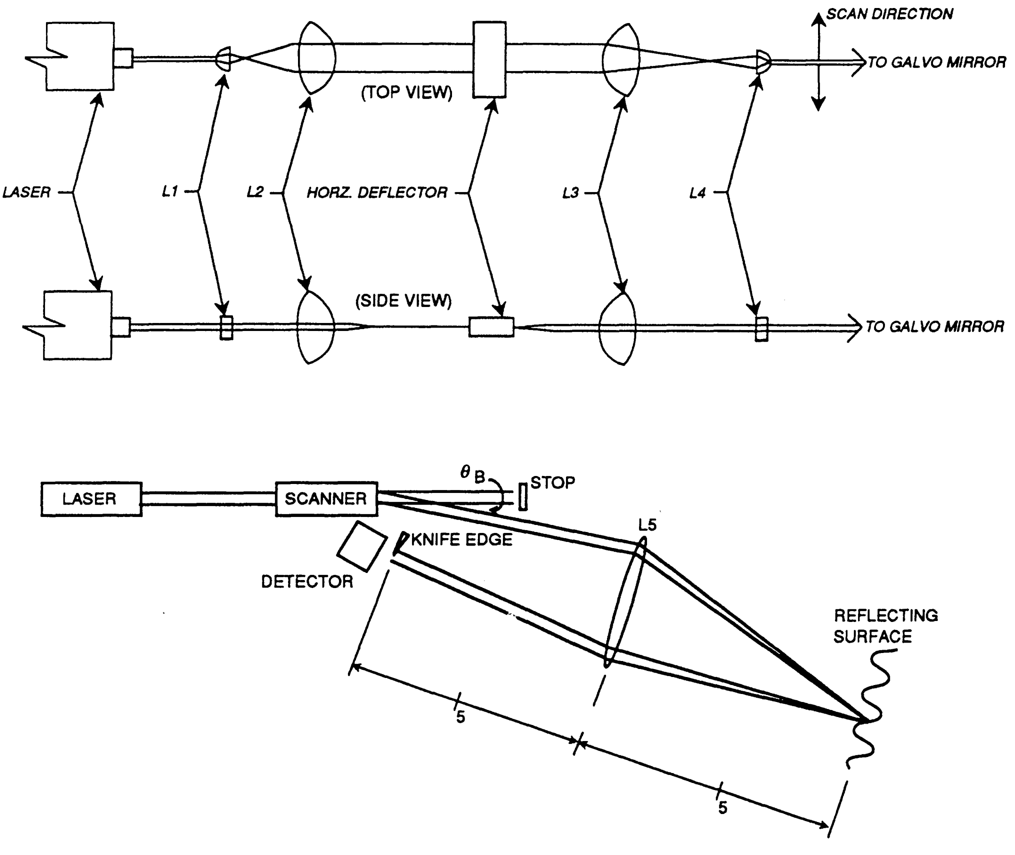

FIG. 5. The essential optical components of the scanning laser module in SLAM.

numerical aperture is defined as the sine of half the apex angle of the cone of rays admitted by the system. In SLAM, the numerical aperture is governed by the extent of the angular spectrum of sound diffracted by the specimen and detected by the coverslip. The coverslip in commercially available SLAMs is made of polycarbonate, in which the sound velocity is 2200 m/s, thus giving rise to a single critical angle at 43°. Assuming the "detectable" part of the spectrum to extend from the angle of incidence of the carrier, typically at 10°, to an angle somewhat lower than the critical angle of 43°, say 33°, we find for the numerical aperture of SLAM

$$\mathrm{NA} = \sin 33^\circ - \sin 10^\circ = 0.37.$$

2.2.6 Resolution The resolution R of SLAM is governed by the familiar relationship

$$R = \frac{\lambda/2}{\mathrm{NA}},$$

where l is the acoustic wavelength and NA, the numerical aperture. For example, at a frequency f of 100 MHz, with the sound velocity v in water at 1500 m/s and the numerical aperture at 0.37, we find for the acoustic resolution

$$R = \frac{v/2f}{\mathrm{NA}} = 20\ \mu\mathrm{m}.$$

Other ultrasonic frequencies available in a commercial SLAM are 10 and 30 MHz, which have acoustic resolutions of 200 and 67 μm, respectively.

2.2.7 Sensitivity In any imaging system the ultimate limitation on the detection of image information from weak signals is the inevitable existence of noise. Obviously no amount of amplification can make it possible to derive information from a signal whose power is substantially below the level of the noise entering the receiver. Laser-beam systems are inherently noisy because lasers suffer from high quantum noise. It is produced by the quantum limit on the possible accuracy of measurements involving an electromagnetic field. The quantum-noise power is given by $h\nu$ where h is Planck's constant and ν is the frequency of the laser beam.

For piezoelectric systems, the type of noise of greatest fundamental importance is the so-called "thermal noise." The thermal-noise power corresponding to a bandwidth B is given by the well-known expression kTB where k is Boltzmann's constant and T is the absolute temperature in kelvins. Thermal noise, in principle, can be eliminated simply by cooling all the elements of the system to absolute zero temperature. On the other hand, the expression for quantum noise shows that this noise is constant, and not amenable to reduction by physical means.

We can calculate the equivalent thermal-noise temperature of the readout process by equating the expressions for thermal noise and quantum noise and solving for T. Thus we get h/k for the noise temperature of the laser beam. If we enter this expression with the values of Planck's constant, Boltzmann's constant, and the frequency of the laser, we compute a temperature well in excess of 20 000 K. Hence the equivalent noise temperature associated with the readout process alone in the case of laser systems amounts to over 20 000 K.

From these calculations, we might guess that for ordinary circumstances the ultimate sensitivity of a laser system such as SLAM will be almost two orders of magnitude worse than that of a piezoelectric system such as SAM. Analytical study assuming ideal operating conditions for both systems has confirmed this expectation (Wang and Wade, 1974). The sensitivities achieved by these instruments in practice are, of course, not as good as those calculated in the study, the difference in the sensitivities actually achieved by the two systems being somewhat greater than the two orders of magnitude predicted.

The minimum detectable surface displacement is related to the sensitivity. It can be shown (Whitman and Korpel, 1969) that, with the knife-edge and photodiode detector, this displacement for SLAM is given by

$$\delta_{\min} = (2eB/\alpha P_0)^{1/2}(\lambda/2\pi),$$

where e is the charge on an electron, B the system bandwidth, a the responsivity of the photodiode, P_0 the laser power, and λ the optical wavelength. The conventional system uses a laser power of 10 mW and has a system bandwidth of 1 MHz, resulting in a calculated $\delta_{\min}$ of 1.3×10^{-12} m. In practice, a minimum detectable displacement of 1×10^{-11} m has been demonstrated, which renders for SLAM an approximate 70 dB of dynamic range.

2.3 Applications

SLAM instruments are used in a wide variety of ways. Four of the chief categories of use are discussed below.

2.3.1 Biomedical SLAM provides a new capability for visualizing the structural characteristics of living tissue. Viable, unfixed, and unstained tissues retain much of their intrinsic, *in vivo* properties. The unique aspect of techniques made possible with SLAM is that they can differentiate elastic, viscoelastic, and density features of structure at the microscopic level and identify subtle changes in tissue architecture.

These changes are revealed by means of corresponding changes in the acoustic index of refraction, the acoustic absorption coefficient, and the acoustic impedance. This information is available in the acoustic micrographs and in the interferograms produced by SLAM.

Some of the studies undertaken using SLAM include embryological development of mice and chickens, physiology and pharmacology of hearts in organ culture, mammalian kidney tissue, metastatic adenocarcinoma in human liver, abnormal or diseased tissue, dentine and enamel components, collagen threads, muscle mechanics and wound maturation, bone pathologies, development of the organ structure of the fruit fly larva Drosophila, plant cell structure, and amphibian limb regeneration. SLAM has also been used to measure acoustic velocity in skeletal muscle, in rat tail tendon fibers, and in rat liver as a function of fat concentration.

2.3.2 Nondestructive Evaluation The most popular industrial application of SLAM is bond evaluation. Bonding, joining, and welding processes are extremely important in the manufacturing of a wide range of products. A variety of materials may be involved such as metals, ceramics, alloys, and polymers. Diffusion bonding, heat sealing, laminating, coating, and epoxy or adhesive bonding are a few examples of the many processes. Typical applications include the nondestructive evaluation of seam and spot welds, adhesive bonds, and laminated polymers.

SLAM has difficulty in scanning complex geometries. Nevertheless, special techniques have been worked out that are suitable for accommodating many complex shapes. Inspection of aircraft-engine turbine blades, for example, differs from the customary SLAM operation and requires a special fixture to control precisely the motion of the blade in the gap between transducer and coverslip. The interior of the blade consists of a network of channels for air cooling. Ribs that are bonded to each other separate the channels in each half of the blade. SLAM can distinguish between the bonded and unbonded ribs and can measure the percentage of bond for partially bonded ribs as well.

2.3.3 Materials Technology Another popular industrial application of SLAM is the nondestructive detection of voids, inclusions, cracks, and contamination in various materials. SLAM provides the means to visualize grain structure, grain boundaries, and impurities in metal specimens such as steel, copper, aluminum, and titanium alloy.

In structural applications, ceramics are used where high temperature and light weight are important. Silicon nitride, silicon carbide, and zirconia are receiving much attention for future engine applications. However, because of the inherent brittleness of ceramic materials, small defects are very critical to the structural integrity of the materials. SLAM has been very successful in nondestructive screening of ceramic samples.

Composite materials, with the combination of materials having different properties from the components and with manufactured anisotropy, have become a challenge for acoustic microscopists. SLAM can clearly define the relevant property distribution of materials by differential attenuation of ultrasound within the field of view. SLAM is used to detect fiber–matrix disbonding, matrix cracking, and interlaminar delamination in fiber-reinforced polymers, as well as in ceramic–matrix and metal–matrix composite samples that are subjected to tensile, fatigue, and impact tests.

2.3.4 Electronic Components Much attention has been given to the use of SLAM for nondestructive evaluation of microelectronics components, and for detecting cracks, disbonds, delaminations, voids, and porosity in these devices. SLAM has been extensively employed to detect disbond at molding-compound–paddle or molding-compound–die interfaces, and at the die attachment itself in plastic integrated-circuit packages such as plastic leaded chip carriers. SLAM is also used to optimize the process for consistent bond quality in tape automatic bonding interconnections. Other applications include, but are certainly not limited to, areas such as ceramic chip capacitors, substrates, surface-mount solder joints, thick film and ceramic dual in-line packages, or other hermetically sealed packages.

3. SCANNING ACOUSTIC MICROSCOPE (SAM)

3.1 Principles of Operation

The most common configuration is the reflection-type scanning acoustic microscope (SAM), as shown in Fig. 6. A piezoelectric material is attached to the flat end of a buffer rod to excite a plane-wave ultrasonic beam into the buffer rod. The other end of that rod has a spherical or cylindrical lens that focuses the sound beam, which is transmitted into the fluid. An electrical tone burst excites the transducer, thus transmitting a packet of ultrasonic energy that is focused by the lens to a diffraction-limited spot. Resolution is determined by the size of the focal spot, as it is with optical microscopes. However, because sound waves propagate at a velocity that is about five orders of magnitude slower than the speed of light, the resulting image resolution can be comparable to that of optical microscopes when operating in the gigahertz frequency range. A portion of the ultrasonic signal that is reflected by the sample and propagated back through the lens and onto the transducer, where it generates an electrical signal that is collected by the receiving electronics.

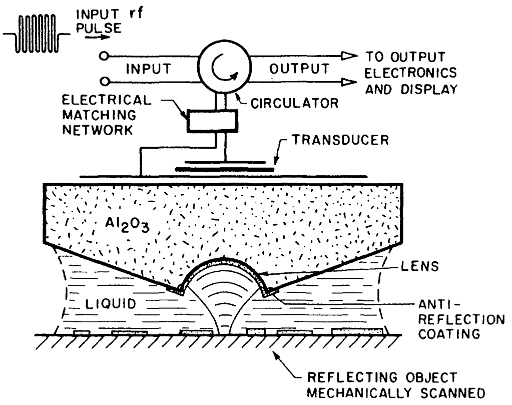
INPUT rf PULSE
INPUT
OUTPUT
TO OUTPUT ELECTRONICS AND DISPLAY
ELECTRICAL MATCHING NETWORK
CIRCULATOR
TRANSDUCER
Al_2O_3
LENS
LIQUID
ANTI-REFLECTION COATING
REFLECTING OBJECT MECHANICALLY SCANNED

FIG. 6. Schematic diagram of a transducer–lens system for a scanning acoustic microscope (Quate, 1985).

The return signal's amplitude and/or phase is collected and used to modulate the intensity of the display monitor at a location corresponding to the location of the focal spot over the sample. By scanning either the transducer or the object, an acoustic image of the sample is formed on the display monitor. Typical scanning speeds are 20–30 s/frame for operating in the gigahertz frequency range, and for a field of view of up to 0.5 × 0.5 mm^2. For lower-frequency operation, a field of 5 × 5 cm^2 can be imaged in a few minutes.

Variations in the local mechanical properties of a sample or changes in the location of the surface with respect to the focus of the lens cause changes in the amplitude and phase of the reflected signal and appear as contrasted features in the image. Local variations can be induced by changes in orientation of the grains of a material, by the presence of a different material, or by defects at or below the surface of the sample.

SAM images show new features and do not blur as a lens is brought closer to a sample. This enhanced imaging capability is due to the physical interactions that take place between the ultrasonic field and the sample. When a focused ultrasonic beam in water impinges on a solid surface placed at its focus, part of the beam is reflected back, part is transmitted as longitudinal and shear waves into the bulk of the solid, and one part is converted into a surface wave that propagates on the surface. The energy in the surface wave leaks back into the water and away from the lens. As the lens is brought closer to a sample (defocused), the energy in the surface wave leaks back into the transducer. Thus, the reflected signal collected by the transducer is now made of the interference between the specular reflection from the surface of the sample and the mode-converted leaky surface wave. If there are defects in the sample, some of the energy that is propagated into the sample as longitudinal and shear waves can be reflected back and received by the transducer. Therefore, four types of acoustic signal paths can contribute signals at the transducer, and a variation in any of these signals changes the image.

Focused C-scan is the name of one variation of an acoustic microscope whose field of application is in imaging the internal structure of samples. A schematic of a focused C-scan imaging system is shown in Fig. 7. A focused transducer is excited with a broadband pulse, typically a half cycle at the frequency of operation, and the focus of the transducer is placed below the surface of sample at a plane where the defects are to be imaged. The preferred transducer has a broad bandwidth and short impulse response to separate in time the reflections from the interface from the reflection from within the sample. The surface of the sample and the near-surface region corresponding to the pulse duration are not inspected by this technique and are known as the "dead zone." To image other planes within the sample, the location of the focus is changed, and the sample is scanned again. The peak amplitude of the return echo modulates the intensity of the display monitor where the image is shown. Alternatively, several gates are used to measure the return signal at different locations, enabling imaging at more than one plane at a time. Yet another variation monitors the amplitude and time delay of the return signal from the back side of the sample, thereby generating a velocity map of the sample together with an image of total attenuation along the path of propagation.

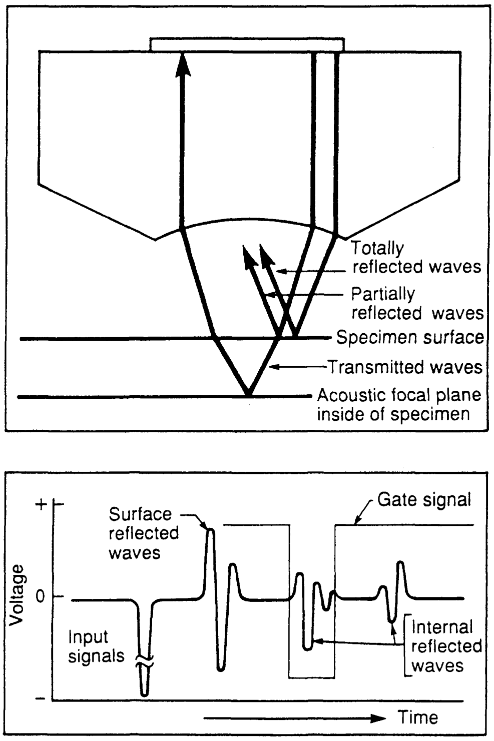

FIG. 7. Schematic diagram depicting the operation of a focused C-scan imaging system (courtesy of Olympus Corp.).

Thus differences between the C-scan and SAM are evident. The usual excitation of one is broad band (C-scan), and of the other (SAM) is narrow band. The transducer for the C-scan has a lens with a low *f*-number in order to enhance penetration into the sample, whereas for the SAM the *f*-number is low in order to excite surface waves on the sample and obtain the smallest focal spot for surface imaging. The domain of inspection of the C-scan system is the bulk of a sample, whereas the SAM does most of the work near the surface of the sample. Lastly, in a C-scan system, the longitudinal or shear wave is detected and interrogated for the presence of a defect. In a SAM, the detected signal is comprised of several components: the specular reflection, a mode-converted surface wave (or any near-surface mode that is allowed), and a possible reflection from a longitudinal or shear wave excited within the sample. It is important to note that at high frequencies the attenuation in samples reduces the penetration of the ultrasound beam, and limits the inspection to the surface of the sample.

It is also possible to defocus a transducer enough to separate the specular reflection temporally from the surface-wave reflection. This was first used in an acoustic microscope (Liang *et al.,* 1982) to make quantitative measurements of step heights and residual-stress variations. Later work (Yamanaka, 1982) used this phenomenon in a pulse-mode–excited system rather like the C-scan to measure variations of surface-wave velocity in ceramic samples. Another implementation (Gilmore *et al.,* 1986) used the surface-wave signal alone to image surfaces of samples for defects at or near the surface in a standard C-scan system; the spot size is about $0.2l_R$, where l_R is the surface-wave wavelength, at the 3-dB points because the surface wave is incident from a full circle around the focal spot. It is also possible to use the SAM with broadband excitation like a C-scan system to image solids and biological samples (Daft and Briggs, 1989).

3.2 Transducer, Lens, and Test Media

3.2.1 Transducers The transducer–lens system is at the heart of the SAM, and requires careful design to obtain good images and quantitative measurements. Various technologies and materials are available to make the transducer. Typically, a piezoelectric material such as zinc oxide or aluminum nitride is deposited in a vacuum chamber on the face of the buffer rod. Such a technology has proven capable of generating coherent ultrasonic waves in the frequency range of 30 MHz to 100 GHz. Thin metal bonding of a piezoelectric crystal, such as lithium niobate, has also been used for making ultrasonic transducers. Metal bonding has the advantage of allowing the use of materials with high piezoelectric coupling coefficients such as lithium niobate, which can be attached in different orientations to excite both longitudinal and shear waves. At frequencies below 50 MHz, it is possible to shape piezoelectric shells and use them to excite and focus the sound beam directly. The wide range of piezoelectric materials allows one to insure that the electrical impedance of the transducer is well matched to the electronics for optimum operation. Thin metal films, such as chrome–gold or aluminum, are used as the metallization defining the back contact and the top electrode of the transducer. The top electrode is used to define the active area of the transducer, and is chosen to reduce diffraction in the buffer rod and to give the desired illumination of the lens.

3.2.2 Buffer Rod Buffer rods are used as support for the transducer and to form the lens. Sapphire is the most widely used buffer-rod material at high frequencies because of its low attenuation (0.2 dB/cm at 1 GHz, with an f^2 dependence) and high ultrasonic velocity (11 100 m/s) compared with water (1500 m/s). The high velocity results in the reduction of spherical aberrations. However, sapphire has a high ultrasonic impedance ($Z_a = 44.3 \times 10^6$ kg/m^2 s) as compared with water ($Z_a = 1.5 \times 10^6$ kg/m^2 s); this impedance mismatch requires a matching layer ($Z_m = 8.15 \times 10^6$ kg/m^2 s) to improve the transmission from sapphire to water. Matching layers of borosilicate glass or carbon have been used for this purpose. These matching layers are deposited by physical sputtering or evaporation in a vacuum chamber. Fused quartz, silicon, aluminum and many other solids have also been used as buffer rods.

3.2.3 Lenses Lenses, spherical or cylindrical, may be made by grinding and polishing into the buffer-rod material. Chemical etching of materials such as silicon is another method of making lenses for acoustic microscopes. A typical 2-GHz lens in sapphire would have a radius of 50 μm and an aperture of 125 μm. At lower frequencies, both dimensions would increase proportionally. The lens surface must be free of scratches and defects, especially for high frequency operation (>1 GHz), where the wavelength of the ultrasonic beam in the water is a micron or less.

Fresnel lenses are yet a third type of lens being used in acoustic microscopes (Yamada *et al.*, 1986). Planar integrated circuit processing techniques are used to define these lenses in a process that allows making arrays of lenses, and consequently arrays of microscopes for scanning large areas.

3.2.4 Matching Layers A matching layer is usually attached to the lens to improve the transmission of the sound into the liquid. As in optics, the matching layer is a quarter wavelength thick, with the ideal impedance, which is the geometric mean between the impedances of the liquid and the buffer rod. The mechanical properties of several materials used in acoustic applications are summarized in Fig. 8. It is important to note that when the matching layer is deposited by a thin-film process, such as evaporation or sputtering, its mechanical properties will differ from those of the bulk material, requiring independent measurements to ascertain the mechanical parameters of the deposited material.

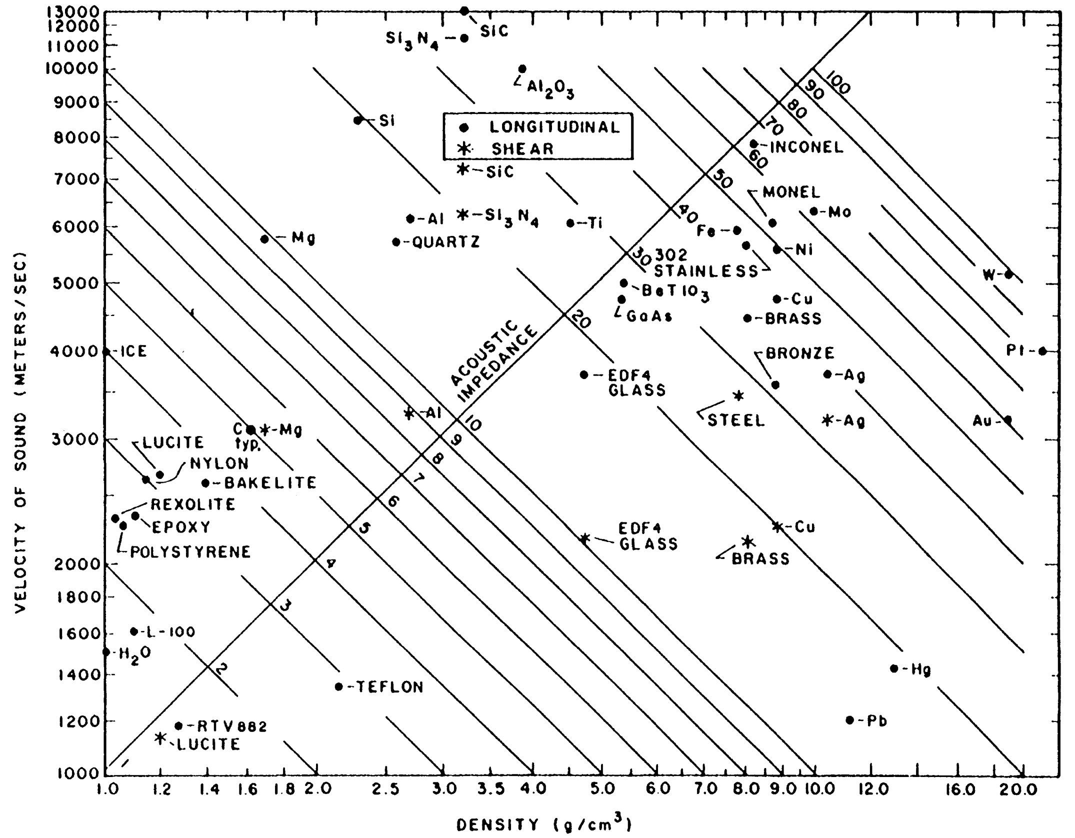

FIG. 8. Nomogram summarizing the acoustic properties of materials (after Eggelton, private communication).

3.2.5 Operating Media The properties of various liquids that are possible immersion media for acoustic microscopy are summarized in Table 1. Water is the usual medium in which samples are placed and tested, a natural and convenient choice. However, other liquids can offer advantages over water, particularly at very high frequencies (over 4 GHz) where attenuation in the water is prohibitive. Attenuation in water is

Table 1. Acoustic properties of selected liquids.

Liquid	Temperature	C (m/s)	$10^{17}a/f^2$ (dB s^2/cm)	10^{-5}Z (g/cm^2-s)
Water	25 °C	1495	191	1.5
Water	60 °C	1550	95	1.5
Methanol	30 °C	1088	262	0.866
Carbon disulfide	25 °C	1310	88 (3 GHz)	1.65
Mercury	24 °C	1449	50	16.7
Gallium	30 °C	2870	13.7	17.5
Nitrogen	77 K	850	120	0.68
Oxygen	90 K	900	86	1.0
Neon	27 K	600	201	0.72
Argon	87 K	840	132	1.2
Xenon	166 K	630	191	1.8
Helium	4.2 K	183	1966	0.023
Helium	1.95 K	227	610	0.033
Helium	0.4 K	238	15 (1 GHz)	0.035

2200 dB/cm at 1 GHz and at room temperature, and increases with the square of the frequency. For operation above 500 MHz the water is heated to about 60 °C where the attenuation is reduced to half its room-temperature value. The highest operating frequency achieved in water is 5 GHz with a resolution better than 2000 Å (Hadimioglu and Quate, 1983).

Liquid metals like gallium and indium are viable coupling fluids for acoustic imaging. As compared with water, liquid metals have high acoustic impedance and low attenuation, which are desirable features for increased signal level; however they also have larger velocities, which increase the wavelength and hence reduce the resolution. Liquid metals have been used very successfully (Attal and Cambon, 1978) to demonstrate that the improved impedance matches between the buffer-rod and liquid and between liquid and sample make it possible to penetrate deep into a sample. Features on silicon wafers have been imaged thus by looking through the back side of the wafer.

Air or gas at high pressure is another medium that has been used for imaging. Imaging in air affords the advantage of increased resolution because the speed of sound is five times smaller than in water, and it does not require wetting. One problem of working in air is the mismatch between the impedance of air and that of the sample and buffer rod. This mismatch is typically 4–6 orders of magnitude, which results in inefficient coupling of the sound into and out of the air, and which limits the instrument to imaging the topography of the sample only. Another problem is the large attenuation in the air, which is about 1.2 dB/cm at 1 MHz and which increases with the square of the frequency. These problems are somewhat reduced by operating at high pressure. An instrument operating at 500 MHz with argon gas at a pressure of 15 atm has been demonstrated (Wickramasinghe and Petts, 1982). This medium of operation has seen limited attention because of the risk of explosion in high-pressure vessels.

Cryogenic fluids such as liquid nitrogen and argon offer a medium of operation with low sound velocity and relatively low acoustic attenuation. Liquid nitrogen has been used for high-resolution imaging at frequencies up to 2.8 GHz. The best medium for high-resolution imaging by far is liquid helium at temperatures below 0.2 K, for it has very low sound velocity and negligible acoustic attenuation even at frequencies of tens of gigahertz. Imaging has been demonstrated at a frequency of 8 GHz with a resolution of 200 Å (Hadimioglu and Foster, 1984). Cryogenic liquids, like gases, allow very little coupling of the sound energy into a sample and are thus used mostly in surface-imaging instruments.

3.3 System Configurations

The first SAM was of the transmission type (Lemons and Quate, 1974). Two identical lenses were used, one to transmit the ultrasonic signal and the other to receive it after it had passed through a sample. This instrument proved difficult to align and operate, however, and yielded to the reflection-type microscope most widely used today. A typical schematic diagram for an amplitude-measuring SAM is shown in Fig. 9. The tone burst is derived by modulating a cw signal from a voltage-tuned oscillator (VTO) with a *pin* diode switch. The signal is applied to the transducer through a circulator, and the main pulse is gated out to avoid saturation of the receiver electronics. Heterodyne demodulation is used to detect the reflected pulse, which is then measured and stored in a digital computer via sample-and-hold (S/H) and analog-to-digital (A/D) conversion circuitry. The computer controls the mechanical scanner and stores the amplitude of the signal corresponding to a particular location over the sample.

Another class of systems measures both the amplitude and phase of the return signal from the sample. With the added information of phase, the systems carry out more quantitative measurements including inversion of the sample reflectance function, thin-film height, surface residual stress, image enhancement, surface roughness removal, multiple-frequency imaging, delamination thickness, and other image-processing applications (Liang *et al.*, 1985a). A schematic diagram of an amplitude- and phase-measuring instrument is shown in Fig. 10. (Reinholdtsen, 1989). The phase measurement is similar to that described in Sec. 2.1.2, the main difference being that the signal is mixed with a number of phase refer-

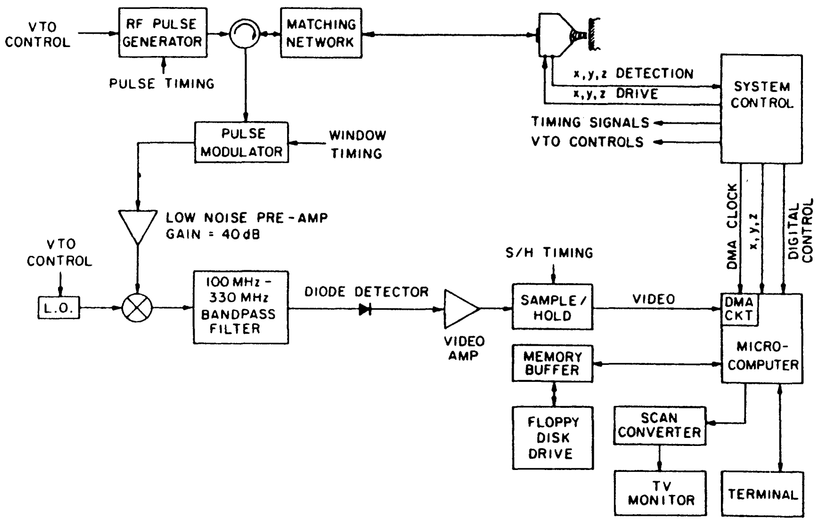
VTO CONTROL
RF PULSE GENERATOR
MATCHING NETWORK
PULSE TIMING
PULSE MODULATOR
WINDOW TIMING
LOW NOISE PRE-AMP GAIN = 40dB
VTO CONTROL
L.O.
100 MHz - 330 MHz BANDPASS FILTER
DIODE DETECTOR
VIDEO AMP
x,y,z DETECTION
x,y,z DRIVE
SYSTEM CONTROL
TIMING SIGNALS
VTO CONTROLS
DMA CLOCK
x,y,z
DIGITAL CONTROL
S/H TIMING
SAMPLE/HOLD
VIDEO
DMA CKT
MICRO-COMPUTER
MEMORY BUFFER
FLOPPY DISK DRIVE
SCAN CONVERTER
TV MONITOR
TERMINAL

FIG. 9. Schematic diagram of a scanning acoustic microscope, where VTO stands for voltage-tuned oscillator (Lam, 1983).

ences before calculation of its magnitude and phase. This type of arrangement allows the measurement of amplitude with a sensitivity of 0.1% and phase with a sensitivity of 0.1°.

3.4 Scanning

At low frequencies, scanners made with dc or stepper motors are available with relocation accuracy of the order of 1 μm, and a field of view of tens of centimeters. In these systems, the scanning time is a few minutes for a scanning field of about 5 × 5 cm^2. At high frequencies, loudspeakers, or leaf springs of beryllium copper driven by piezoelectric stacks, are used for the fast-scanning direction, and universal motors with a belt drive are used for the slow-scan direction. The resulting frame rates are typically of the order of 20–30 s for a field of view of 500 × 500 μm^2.

One exciting implementation makes use of cylindrical arrays, which allow electronic scanning in one direction. A scanning speed of a couple of seconds per image for a scanning field of 3 × 3 cm^2 while operating at 25 MHz already has been demonstrated (Kubota *et al.*, 1988). SAMs operating at higher speeds and frequencies will indeed become a reality very soon.

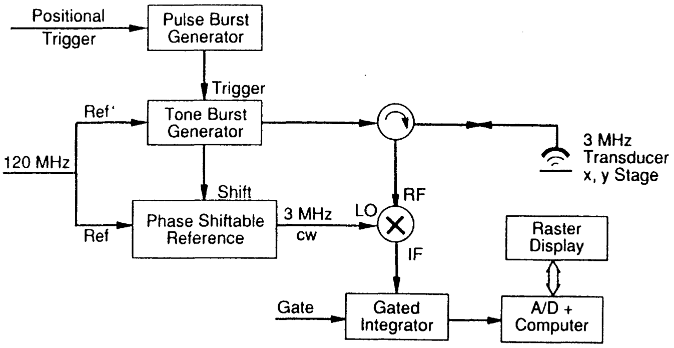
Positional Trigger
Pulse Burst Generator
Trigger
Ref'
Tone Burst Generator
120 MHz
3 MHz Transducer x, y Stage
Shift
RF
Ref
Phase Shiftable Reference
3 MHz cw
LO
IF
Raster Display
Gate
Gated Integrator
A/D + Computer

FIG. 10. Schematic diagram of an amplitude- and phase-measuring scanning acoustic microscope (Reinholdtsen, 1989).

3.5 Response of the SAM

In order to calculate the response of the SAM, it is necessary to consider the excitation, propagation, and interaction of the ultrasonic field with the sample. First, the propagation and focusing of the acoustic wave are presented. Next, the interaction of the focused beam with a sample is presented to explain the contrast mechanism of acoustic images.

The field distribution at the focus of the lens, called the point-spread function (PSF), determines the spatial resolution of the SAM. Several authors (Atalar, 1978; Wickramasinghe, 1979) have derived expressions for the PSF of an acoustic lens. These calculations are similar to those of optical lenses, and only the results of such calculations are presented in the following text.

In the system of Fig. 6, a quasi-cw excitation is applied to the transducer. The stress-field distribution inside the buffer rod undergoes large variations near the transducer, i.e., in the near field. Away from the transducer, the stress field varies slowly and decays as it propagates toward the lens. The near-field/far-field cutoff is known as the Fresnel length and is given by a^2/l, where a is the radius of the transducer and l is the wavelength of the sound in the buffer rod. Figure 11 shows the axial-field distribution at various normalized distances ($S = zl/a^2$) from the transducer. The lens can be placed at any location away from the transducer to affect its PSF.

As the ultrasonic wave reaches the lens, it is reflected in the buffer rod and refracted into the water according to Snell's law. The energy that remains within the buffer rod represents a loss in the energy available at the focus and a source of noise as it returns to the transducer through multiple reflections.

The energy that is transmitted into the water is focused at a distance equal to

$$F = R/(1 - V_W/V_S), \qquad (1)$$

where V_W and V_S are the velocity of the wave in the water and solid, respectively. The above formula is valid for small angles or for lenses with large f-numbers; however, it is usually accepted as a measure of the location of the focus.

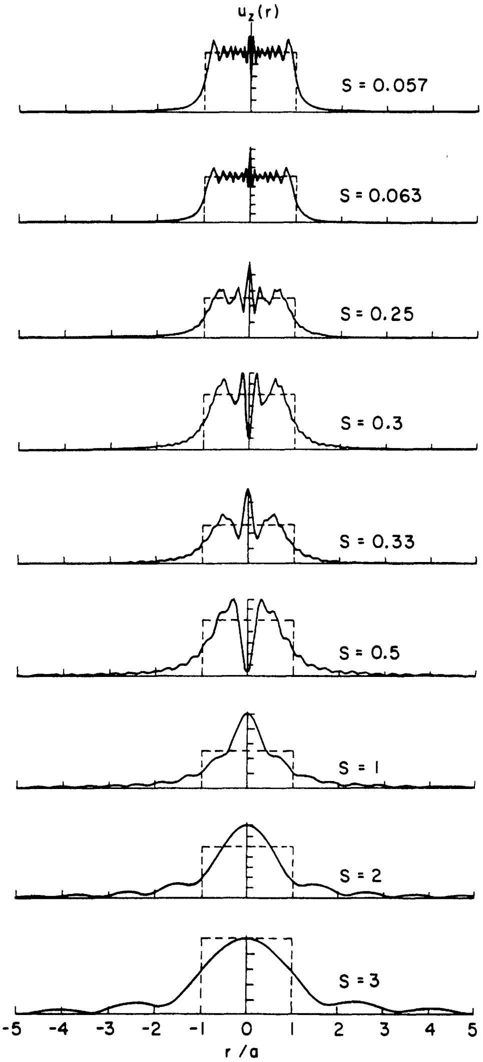

FIG. 11. Radial-field distribution due to plane piston transducer of radius a and at various distances ($S = zl/a^2$) from the transducer, for $ka = 100$ (Kino, 1987).

Figure 12 shows the PSF for a lens placed at different distances from the transducer. The specifications of the lens are f-number $= 1.65$, $f = 40$ MHz, transducer radius $=$ 3.5 mm, and normalized distances $S = 0.5$, 1.0, and 10, respectively (Chou *et al.*, 1988). The importance of varying the parameter S can be seen by noting that in Fig. 11, the field distribution on the lens for $S = 0.5$ contains a null in the center of the lens, while for $S = 1$ the field is uniformly distributed over the lens and the phase has a parabolic dependence on radial distance. For S

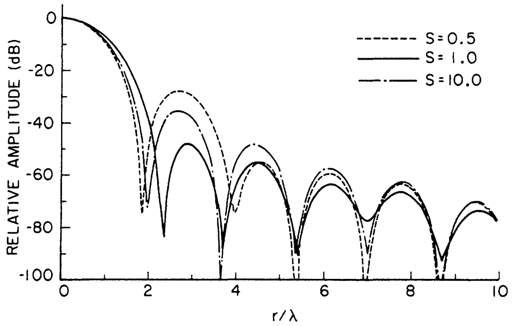

FIG. 12. Radial-field distribution at the focus of a lens with an *f*-number of 1.65, frequency of operation f = 40 MHz, where the lens is placed at different distances $S = z/a^2$ from the transducer (Chou *et al.*, 1988).

= 10, the field is almost constant in amplitude and phase over the lens. In Fig. 12, it is seen that the width of the PSF at the 3-dB points remains basically constant, while the sidelobes change by more than 20 dB. Thus, for surface imaging, it is best to use a lens located at $S = 1$ where the sidelobe level is lowest. For efficient surface-wave excitation, it is best to use a lens with the transducer located at $S = 0.5$, where the excitation off normal is most efficient.

Two important parameters of the focal spot are its lateral 3-dB width, which is equal to $f\# \times l$, and its 3-dB depth of field equal to $7.1 \times f\#^2 \times l$. Additionally, as a beam is focused within a sample its 3-dB width is the same as in water and its 3-dB depth of field is decreased by the ratio of the sound velocities of the sample to water.

Figure 13 shows three images of a sample of titanium taken at a frequency of 2 GHz and at three focus distances (Lam, 1983). At $z = 0$, the instrument is focused on the surface of the sample. Various grains in the crystal are clearly resolved because the amplitude of the reflection from each grain is different. This difference in the reflection can be due to differences in the mechanical impedance of the grains, which could be of different phases of the material or different orientations. Some of the dark spots are due to impurities or defects. The images that are taken when the sample is brought closer to the lens by 0.5 and 1.0 μm show not only improved contrast but also a contrast reversal; some of the features that were light in the image at focus are now dark and vice versa.

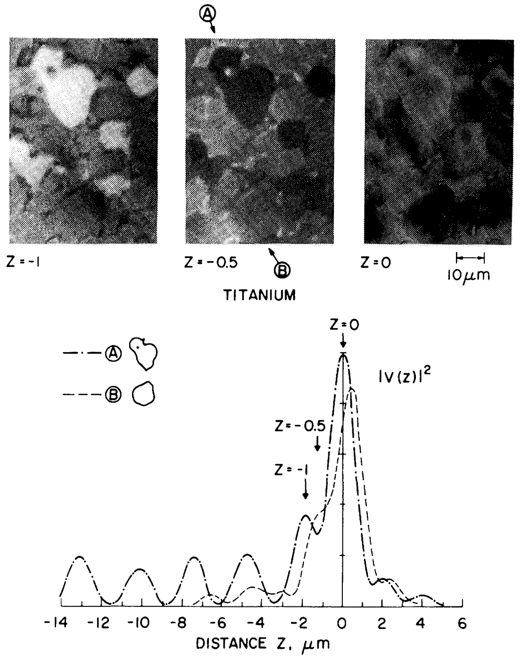

FIG. 13. Images of titanium sample along with corresponding $V(z)$ curves taken at three distances z between the lens and the object (Lam, 1983).

This contrast reversal was first noticed by several researchers, (Atalar *et al.* 1977; Weglen and Wilson, 1978) and explained theoretically and physically in an elegant fashion (Atalar, 1978; Wickramasinghe, 1979; Bertoni, 1984; Chou and Kino, 1987). Essentially, as the lens is brought closer to the sample, surface modes of propagation at the substrate–water interface are excited by the focused beam. These modes radiate their energy back into the water, where it is then collected by the lens and detected by the transducer. Often, it is the surface acoustic wave that is excited at the water–sample interface; such a "pseudosurface or leaky surface wave" is excited by rays arriving at the surface at the angle that satisfies Snell's law:

$$Q_R = \arcsin(V_W/V_R), \qquad (2)$$

where V_R is the surface-wave velocity of the sample. The surface wave leaks back into the water after propagating at the interface, and is displaced from the specular reflection by a distance known as the Schoch displacement (Schoch, 1950), a phenomenon that is also well known in optics.

The contrast-reversal mechanism is interpreted as due to the interference between the specular reflection from the water–sample interface and the radiated signal from the surface mode. Any change in the amplitude or phase of either signal, i.e., due to the change in the distance between the lens and the sample, will effect a change in the output of the transducer. The variation of the output of the transducer as a function of distance between the lens and the sample is known as the *V*(*z*) curve or the acoustic-material signature, and contains information about the mechanical properties of the sample and the illumination characteristic of the lens. Referring to any of the above-mentioned references, it is possible to write *V*(*z*) as

$$V(z/\lambda) = C \int_{2}^{2\cos\theta_0} P^2(t)R(t)e^{-j2\pi(z/\lambda)t}dt, \qquad (3)$$

where $P(t)$ is the pupil function of the lens in the spatial-frequency domain, $R(t)$ the reflectance function of the water–sample interface (reflection coefficient as a function of angle), C a constant, and q_0 the maximum opening angle of the lens. Figure 14 shows a calculation of *V*(*z*) of a silicon-nitride sample in water. The frequency of operation is 40 MHz, and the lens has an $f\# = 1.65$. Notice the typical character in the negative *z* direction (the sample is brought closer to the lens), which reflects the continuous addition of the surface wave and the specular reflections in and out of phase. The spacing between the nulls is measured to evaluate the surface-wave velocity (Kushibiki and Chubachi, 1985).

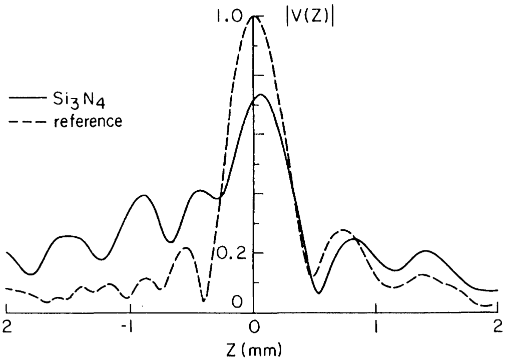

FIG. 14. *V*(*z*) of a silicon-nitride sample and reference material identical to silicon nitride but that does not allow the propagation of surface waves. The difference between the two wave forms is the surface-wave excitation, leakage, and detection (Chou *et al.*, 1988).

The above equation shows that *V*(*z*) is given by the Fourier transform of the product of the pupil function of the lens and the reflectance function of the sample. This equation expresses the power of the microscope as a quantitative tool for material characterization. Either the pupil function of a lens or the reflectance function of a sample can be measured by an inverse Fourier transform of a measured *V*(*z*). For instance, the reflectance function of a water–air, water–lead, or water–Teflon interface has a constant amplitude and phase over most angles of interest and can be factored out in the above equation. Thus, an inverse Fourier transform of a measured *V*(*z*) gives the pupil function of the lens. Once the pupil function is known, a measured *V*(*z*) off of a sample can be inverted to evaluate the reflectance function of the water–sample interface. This method has been used (Liang *et al.*, 1985b) to characterize lenses and unknown samples.

3.6 Imaging Applications

Most of the early imaging work was directed toward imaging biological samples. SAMs are well suited for this application because they have a similar resolution to optical microscopes without the need for special sample preparation. In addition, biological samples can be imaged alive in water. Today, however, the major applications of SAMs are in the field of materials science for the examination of structural materials such as plastics, rocks, semiconductors, and metals. In this section we show images taken by SAMs at various frequencies and of different materials.

3.6.1 Biological Applications Figure 15, one of the first taken with an acoustic microscope, shows an acoustic micrograph of an unstained section of human lung tissue taken at a frequency of 600 MHz and with a magnification of about 200× (Lemons, 1975). The image shows very clearly several features of the human lung tissue—the air sacs, capillaries, cell walls, and blood vessels—and is typical in the detail it shows of variations in the mechanical properties of tissue. Much better resolution can be ob-

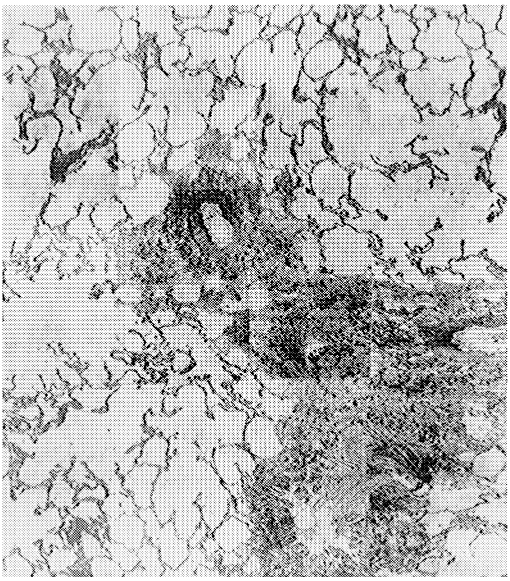

FIG. 15. Image of human lung tissue (Lemons, 1975).

tained by operating the acoustic microscope at a higher frequency, as seen in Fig. 16, which is an image of frog heart cells taken at a frequency of 1.6 GHz (Hoppe and Bereiter-Hahn, 1985). The top image is taken with the cells on a plastic substrate; the bottom image is taken on a glass substrate. The bottom image also shows a line scan of the amplitude variation over the middle of the cell. The images indicate that when a strong mismatch exists between the cells and the substrate, the image is dominated by the reflection from that interface. The image of the cell with the plastic substrate clearly shows more contrast from within the cell.

Figure 17 shows the image of a Myxobacterium taken at 8 GHz in liquid helium (Hadimioglu and Foster, 1984). This is the highest-resolution image taken to date, showing the details of the bacterium with a resolution of 200 Å. The contrast seen in the image corresponds to different levels of focus over the sample.

3.6.2 Materials Science Figure 18 shows the image of an integrated circuit taken at a frequency of 1 GHz (courtesy of Olympus Corporation). The figure is a superposition of three images taken at focus levels of $z = 0, -2$, and $-4\ \mu m$, and displays the

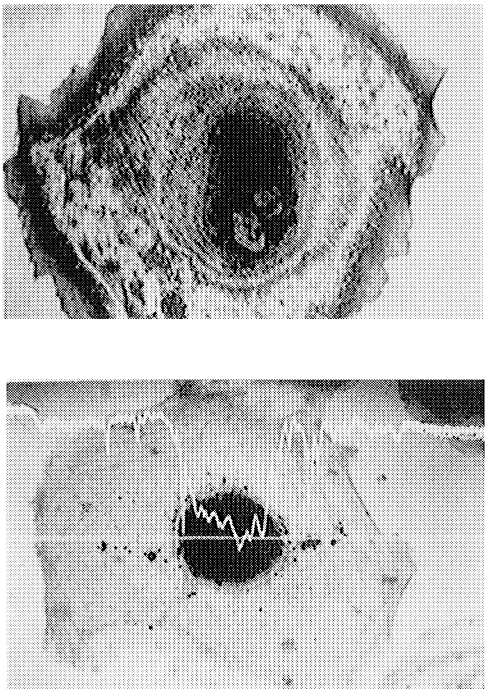

FIG. 16. Image of frog heart cells. Top image is a cell on a plastic substrate and bottom image is a cell on a glass substrate (courtesy of M. Hoppe, Leitz-Wetzlar).

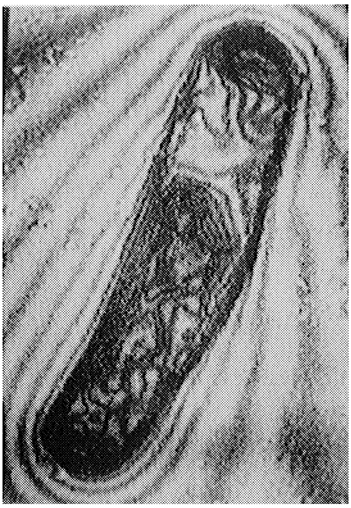

FIG. 17. Image of Myxobacterium taken with the scanning acoustic microscope operating in liquid helium (Hadimioglu and Foster, 1984).

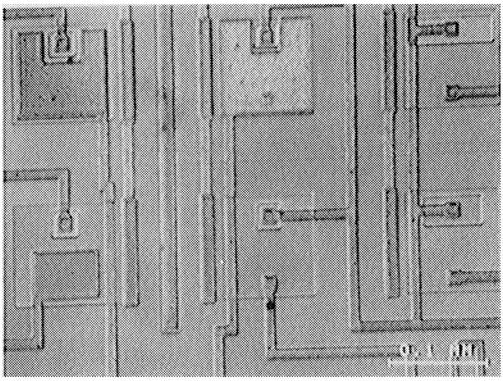

FIG. 18. Three superposed images of an integrated circuit taken at three focus distances (courtesy of Olympus Corp.).

various levels of the circuit with good contrasts; the outlines of the metallizations and diffusions are very clearly shown. Figure 19 shows a comparison between the optical and acoustic images of a chrome metallization on glass; the optical image shows the pattern with no visible defects, while the acoustic image shows clearly that there are poor-adhesion regions in the film. The two defocus images demonstrate the ability of the instrument to perform subsurface imaging.

Figure 20 shows a much lower-frequency image of impact damage in a carbon-epoxy laminate at 3 MHz, but the instrument is still a copy of its higher-frequency cousin. The image shows the layering of the many plies in the composite. It shows breaks in fibers, and it shows disbonds at various levels in depth of the sample. For these materials, acoustic microscopy is superior to dye-enhanced x-ray and focused C-scan imaging.

Figure 21 shows an image of a polished

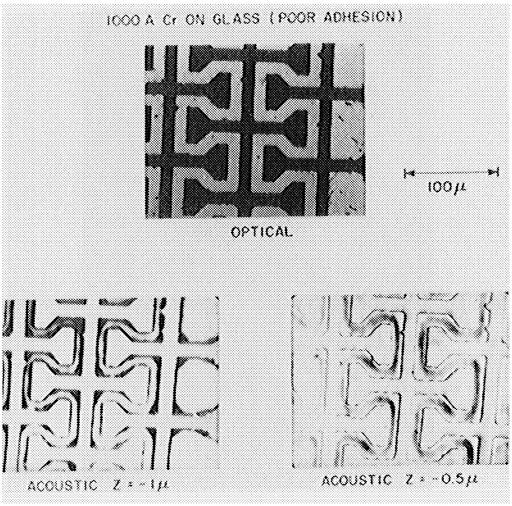

FIG. 19. A comparison between optic and acoustic images of poorly adhered chrome films on a glass substrate (courtesy of C. F. Quate).

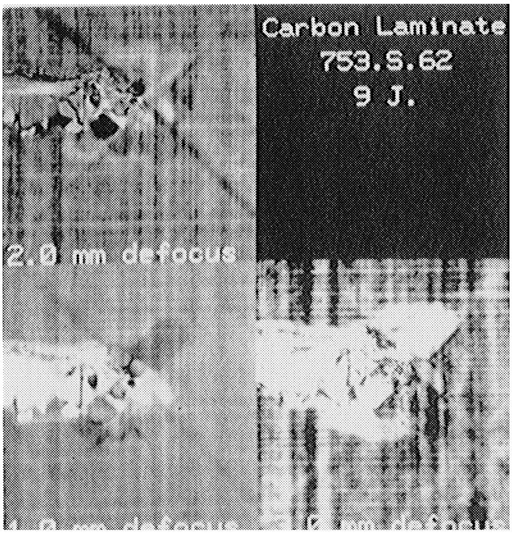

FIG. 20. Image of impact damage in a carbon-epoxy laminate. The sample was damaged with an impactor having energy 9 J, and the images were taken at three focus distances (Reinholdtsen, 1989).

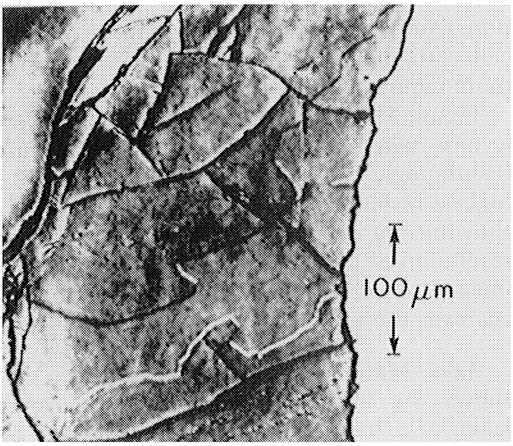

FIG. 21. Image of a granite sample showing cracks and grain boundaries (courtesy of G. A. D. Briggs).

thin section of granite. The image is taken at a frequency of 370 MHz, and shows clear details of cracking and grain boundaries in the rock (Briggs *et al.*, 1989). SEM and transmitted polarized-light images of the sample show less detail of the microcracks and the grain boundaries, illustrating yet again the power of the SAM over traditional imaging techniques.

Figure 22 (courtesy of Hitachi Construction Machinery Co.) shows a sample of a silicon-steel plate with and without residual stress. Localized differences of surface acoustic-wave velocities are observed with 6.5% tension deformation. The image was taken at a frequency of 600 MHz.

Figure 23 is an example using a focused C-scan imaging system to look inside materials. A René-95 steel sample with 0.25-, 0.375-, and 0.5-mm electric-discharge–machined holes is imaged where the plane of focus is 3.1 mm below the surface and at the end of the holes. The sample is 24 × 35 mm and the imaging plane is the bond line between two pieces of the same material. The images show the holes along with a large amount of porosity because of the large depth of field of the lens (Gilmore *et al.*, 1986).

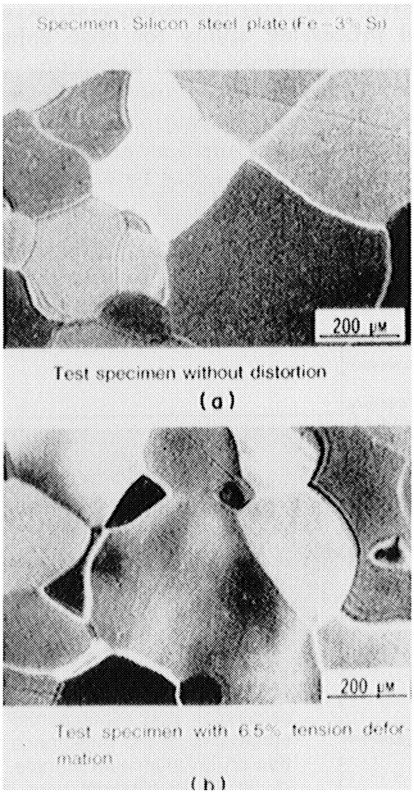
Specimen: Silicon steel plate (Fe–3% Si)
200 μ
Test specimen without distortion
(a)
200 μ
Test specimen with 6.5% tension deformation
(b)

FIG. 22. Images of a silicon-steel plate showing differences due to residual stress (courtesy of J. S. Okagima, Hitachi Construction Company).

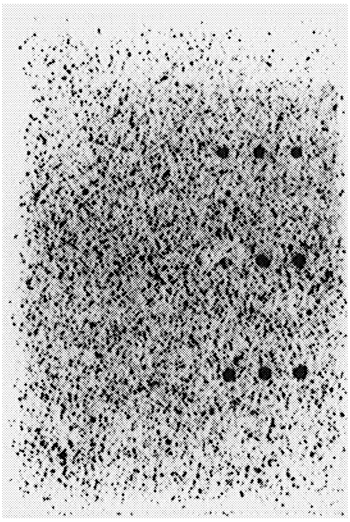

FIG. 23. C-scan image inside a René-95 specimen showing porosity and the ends of flat-bottomed holes (Gilmore, *et al.*, 1986).

3.7 Metrology

3.7.1 Amplitude Only As mentioned earlier, it is possible to measure some of the mechanical properties of a sample through Fourier inversion of $V(z)$ curve. Using a SAM that measures the amplitude of $V(z)$, it is possible to calculate the surface-wave velocity and surface-wave attenuation. The most prominent work in this area is that of the group headed by Profs. Chubachi and Kushibiki at Tohoku University in Sendai, Japan. They use cylindrical lenses that excite surface waves in only one direction and make very accurate measurements of the surface-wave velocity and attenuation (Kushibiki and Chubachi, 1985). The distance between the nulls in the negative side of $V(z)$ is given by

$$D_Z = (l_R/\sin q_R)(1 + \cos q_R)/2, \tag{4}$$

where D_z is the spacing between nulls and q_R is the critical angle for exciting the surface wave. The attenuation of the surface wave

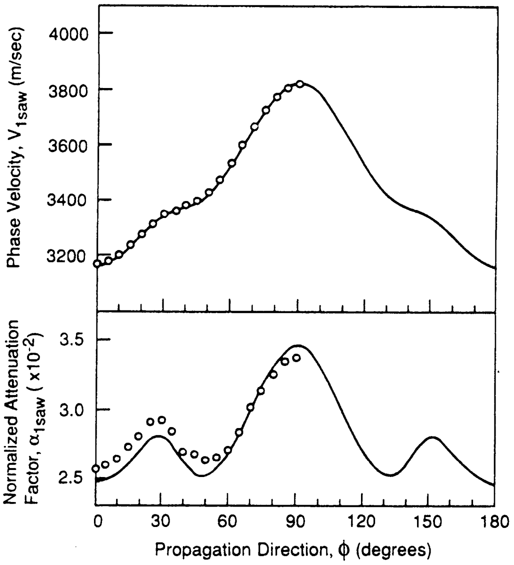

FIG. 24. Comparison of theoretical calculations and experimental measurements of the velocity and attenuation of a *Y*-cut α-quartz sample as a function of direction of propagation (Kushibiki and Chubachi, 1985).

can be deduced subsequently from the decay of the $V(z)$ curve as a function of z in the negative z direction (Weglein, 1985).

With a spherical lens, the surface wave is excited from all directions over a sample, and if the sample is anisotropic, variations in the surface-wave velocity will result in an average value for the surface-wave velocity. The cylindrical lens overcomes this problem and allows an accurate measure of anisotropy of the sample. A typical result of this work is shown in Fig. 24, where the surface-wave velocity of a sample of *Y*-cut α-quartz is shown as a function of angle (Kushibiki and Chubachi, 1985). Each point on the chart represents an independent $V(z)$ measurement and evaluation of the velocity at that angle. The agreement between theory and experiment is excellent and reflects an error of about 0.15%. The attenuation of the surface wave is also shown as a function of angle where the agreement with theory is also good.

3.7.2 Amplitude and Phase Some SAMs measure both amplitude and phase of the acoustic signal (Liang *et al.*, 1985; Reinholdtsen, 1989). The advantages offered by these instruments include enhanced height sensitivity, full $V(z)$ inversion, and increased signal processing ability for edge enhancement and subsurface-defect detection.

3.7.2.1 Reflectance-Function Inversion. Figure 25 shows the result of a full inversion of the reflectance function, using the Fourier-transform relationship between $V(z)$ and the reflectance function, of a sample of alu-

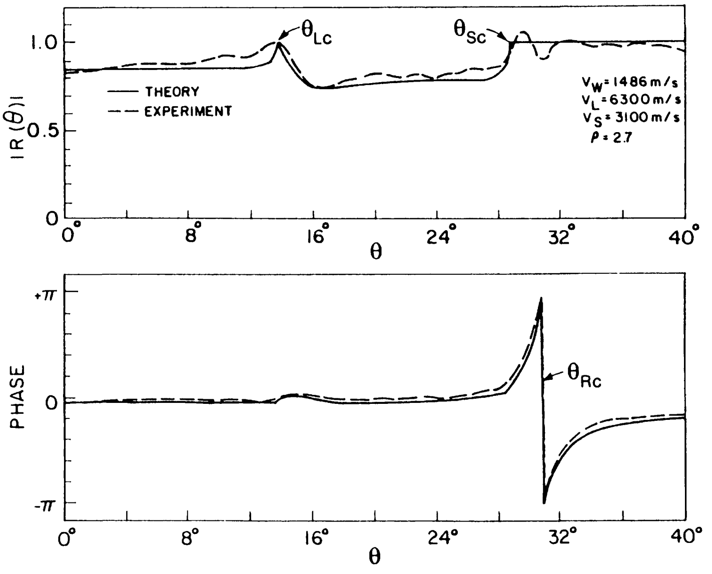

FIG. 25. Comparison of theoretical calculation and experimental measurements of the reflectance function on an aluminum sample. Here q_{Lc} is the longitudinal-wave critical angle, q_{Sc} is the shear-wave critical angle, and q_{Rc} is the surface-wave critical angle (Liang *et al.*, 1985).

minum, where q_{Lc} is the longitudinal-wave critical angle, q_{Sc} is the shear-wave critical angle, and q_{Rc} is the surface-wave critical angle. Because agreement between theory and experiment is excellent, it is possible to deduce the density and the longitudinal-wave, shear-wave, and surface-wave velocities of the sample. The ringing at the critical angles is due to the limited working distance, because the lens touches the sample for large values of defocus. The above results indicate that the best measurement system would then consist of a cylindrical lens and an amplitude- and phase-measuring instrument.

3.7.2.2 Residual-Stress Measurement. For a relatively large defocus distance, and with short-duration pulses, it is possible to separate, in the time domain, the specular reflection from the mode-converted surface wave. The phases of the two pulses can be compared to determine, very accurately, any variations of the surface-wave velocity, or the distance between the sample and lens. Such a configuration has been used to measure residual stress in a flat and homogeneous sample. A Pyrex disk 3.24 mm in thickness and 7.5 cm in diameter was treated to induce a circularly symmetric residual stress that decreased away from the center of the sample. An indentation-based measurement of the circumferential (solid circles) and radial (hollow circles) residual stresses is shown in Fig. 26. The figure also shows a measurement of the variation in the surface-wave velocity (phase of surface-wave reflection with respect to the specular reflection) in the sample at similar locations. The agreement between the two measurements is excellent. Because it is possible to measure a phase shift of 0.1°, a residual-stress change of less than 1 MPa can be measured.

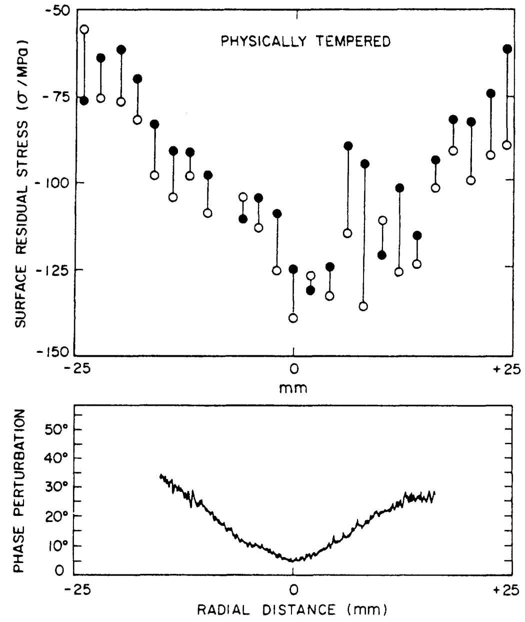

FIG. 26. Comparison of the measurement of residual stress in a Pyrex disk: top, by damage extent due to a Vickers indentor (courtesy of Marshall & Evans); bottom, by variation in surface-wave velocity (Liang *et al.*, 1985).

4. NEAR-FIELD SCANNING ACOUSTIC MICROSCOPE (NFSAM)

The transmission NFSAM is similar to the SLAM except that the sound field on the surface of a sample is detected with a tip or a pinhole that is much smaller than the wavelength. Subwavelength resolution is obtained for surface features, while subsurface features are imaged through their shadow on the surface of the sample. The reflection NFSAM is similar to the SAM; the sound can be incident through a tip or a pinhole, and variations in the reflected signal from the sample are used to image variations in the mechanical properties of a sample. In both transmission and reflection modes of operation, the signal-to-noise ratio in a NFSAM is quite small because the tips or pinholes are much smaller than the wavelength, and very small amounts of energy are collected or changed by the presence of the sample. Very simply, the sound emitted from a pinhole will propagate as a collimated beam for a distance roughly equal to the diameter of the pinhole, and thus define the near field. Beyond this distance, the beam will spread and decay rapidly in strength. Samples have to be placed in the near field in order to be imaged. The images obtained by the NFSAM represent variations in the mechanical properties of the sample with a spatial resolution equal to the size of the pinhole or tip.

The near-field scanning acoustic microscope (NFSAM) is a variation on the above two instruments where the resolution is determined by the physical dimensions of a tip or a pinhole and their distance from the

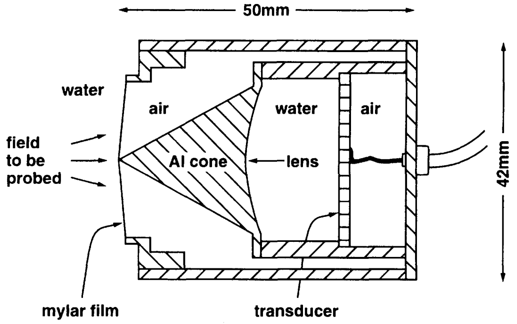

FIG. 27. Schematic diagram of a high-resolution acoustic probe. The aluminum–water interface acts as a positive lens (Durr *et al.*, 1980).

sample, rather than the dimensions of a focal spot. Two types of systems are distinguished, tip systems and pinhole systems.

4.1 Tip NFSAM

A schematic of the first tip NFSAM, reported by Durr *et al.* (1980) is shown in Fig. 27. The system is for a probe with a resolution of *l*/4. The probe consists of a metal cone whose tip receives the signal and determines the resolution. The back side of the metal cone is curved to transmit the received sound into water as a plane wave. The plane wave of sound is then detected with a flat piezoelectric transducer. This probe was used to demonstrate a resolution of about 75 μm while operating at a frequency of 4 MHz where the wavelength of sound in water is 375 μm. Zieniuk and Latuszek (1986) used a similar system with a sapphire cone to demonstrate *l*/10 resolution while operating at a frequency of 30 MHz. Guthner *et al.* (1989) used a resonant quartz-crystal tuning fork as a NFSAM. As the edge of the tuning fork was brought close to a sample, its resonant frequency and amplitude changed depending on the local properties of the sample. This system demonstrated imaging with resolution of less than 1 μm while operating at a frequency of 32 kHz.

Some impressive NFSAM images were obtained when coupling the sound in and out of atomic tips such as those used in scanning tunneling microscopy (STM) and atomic-force microscopy (AFM). Figure 28 shows a block diagram of the system of Uozumi and Yamamuro (1989). In this system a transducer is attached to the top side of a rod with a tip of atomic dimensions at the bottom. The transducer operated at a frequency of 1.44 MHz in a pulse-echo mode. The amplitude of the signal reflected from the atomic tip was measured, and variations in this signal were used to image a sample. This system was used to demonstrate imaging with a resolution of the order of 100 Å. Figure 29 shows a schematic of the system of Takata *et al.* (1989). In this system the tip was vibrated by a piezoelectric cylinder, and the transmitted acoustic signal was detected by a piezoelectric transducer on the back side of the sample. Operating at a frequency of 70 kHz, images of surface were obtained with a resolution comparable to those of the system of Uozumi and Yamamuro. In another version of a NFSAM, Kolosov and Yamanaka (1993) take advantage of the nonlinear relationship of force on an atomic tip and distance to the sample to image subsur-

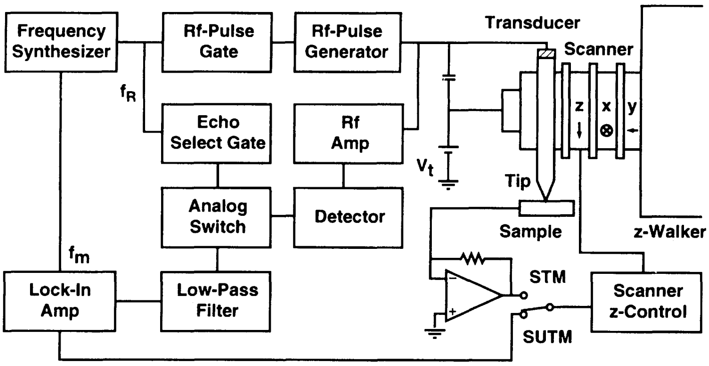

FIG. 28. Block diagram of the experimental setup of the scanning ultrasonic tip microscope (Uozumi and Yamamuro, 1989).

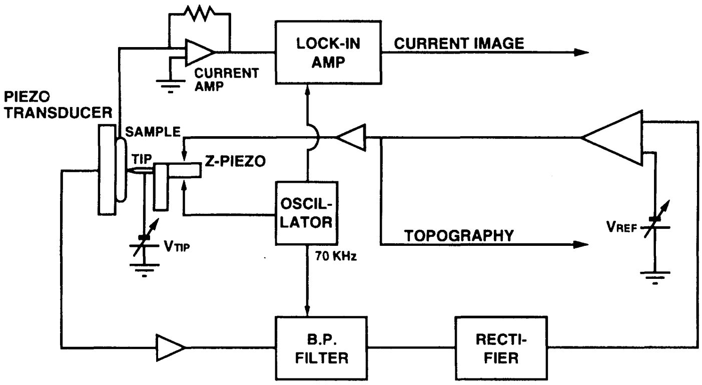

FIG. 29. Schematic diagram of the tunneling acoustic microscope (Takata *et al.*, 1989).

face features with a spatial resolution of the order of 100 Å.

4.2 Pinhole NFSAM

The first demonstration of subwavelength resolution was provided by Ash and Nicholls (1972) using electromagnetic radiation with a cavity with a pinhole. Khuri-Yakub *et al.* (1989) modified a SAM by inserting a shim with a pinhole at the focal plane of the lens to make a NFSAM. A schematic of this system is shown in Fig. 30. This instrument demonstrated imaging with an *l*/10 resolution while working at a frequency of 3 MHz. One important aspect of this system is that a simple modification to an existing SAM can be made to improve the resolution of surface images.

The NFSAM is still in its infancy compared with the other two instruments. No doubt, many applications for imaging variations of mechanical properties with atomic resolution will be available soon.

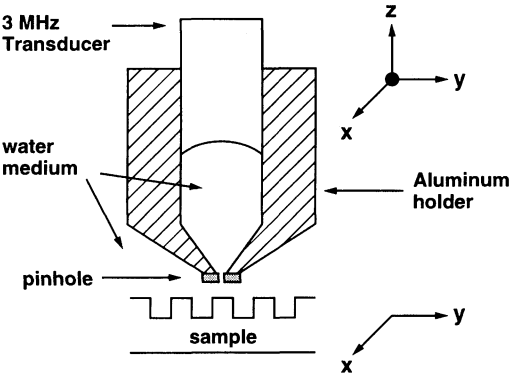

FIG. 30. Schematic diagram of a SAM lens with a pinhole (Khuri-Yakub *et al.*, 1989).

5. CONCLUSIONS

Acoustic microscopy is maturing into a very useful technology for imaging the variations of the mechanical properties of materials. Quantitative measurements are possible, and are gaining in credibility and acceptance by various communities for biological applications and material characterization. The field should continue to grow, and scanning acoustic microscopes will become traditional fixtures in materials-evaluation laboratories.

6. ACKNOWLEDGMENTS

The authors thank Andrew Briggs, Ching-Hua Chou, Babur Hadimioglu, Davis Kent, Larry W. Kessler, Gordon S. Kino, Hua Lee, Z. C. Lin, A. Meyyappan, and Calvin F. Quate for many useful discussions and comments. This work was supported by the National Science Foundation through its Center for High-Speed Image/Signal Processing at the University of California at Santa Barbara, and by the Department of Energy under Contract No. DE-FGO3-84-ER45157 at Stanford University.

GLOSSARY

Acoustic-Material Signature *V*(*z*): Output of a transducer–lens system as a function of the distance between the lens and the sample.

C-Scan Imaging System: Imaging system where the display shows defects in a plane parallel to the scanning plane.

Dead Zone: Near-surface region that is not inspected by a C-scan imaging system.

Defocus: Refers to moving a focused lens closer to or farther away from a sample.

Lamb Wave: A type of mechanical wave that is guided by a plate.

Piezoelectricity: A phenomenon where electric dipoles are generated in a solid by the application of mechanical stress, or a mechanical stress is generated as a result of the application of an electric field.

Point-Spread Function (PSF): The field distribution at the focal spot of a lens.

Rayleigh or Surface Wave: Mechanical wave that is trapped within a wavelength at the interface between a solid surface and vacuum. It is called leaky when it leaves the solid because of loading by a medium such as water.

SAM: Scanning acoustic microscope.

Works Cited

Ash, E. A., Nicholls, G. (1972), "Super-Resolution Aperture Scanning Microscope," *Nature* **237,** 510–512.

Atalar, A., Quate, C. F., Wickramasinghe, H. K. (1977), "Phase Imaging in Reflection with the Acoustic Microscope," *Appl. Phys. Lett.* **31,** 791–793.

Atalar, A. (1978), "An Angular Spectrum Approach to Contrast in Reflection Acoustic Microscopy," *J. Appl. Phys.* **49,** 5130–5139.

Attal, J., Cambon, G. (1978), "Improvements in the Resolution of the Reflective Scanning Acoustic Microscope Application to the Non destructive Evaluation of Microelectronics," in: J. deKlerk and B. R. McAvoy (Eds.), *Proceedings of IEEE Ultrasonics Symposium, 1978,* New York: IEEE, p. 212.

Bertoni, H. L. (1984), "Ray-Optical Evaluation of V(z) in the Reflection Acoustic Microscope," *IEEE Trans. Sonics Ultrason. Ind. Eng. Chem.* **31,** 105–116.

Briggs, A. (1992), *Acoustic Microscopy,* New York: Oxford Univ. Press.

Briggs, G. A. D., Daft, C. M. W., Fagan, A. F., Field, T. A., Lawrence, C. W., Montoto, M., Peck, S. D., Rodriguez, A., Scruby, C. B. (1989), in: H. Lee, G. Wade (Eds.), *Acoustical Imaging,* Vol. 17, New York: Plenum, pp. 1–16.

Chiao, R., Lee, H. (1991), *Int. J. Imaging Syst. Technol.* **3,** 334–353.

Chou, C. H., Kino, G. S. (1987), "The Evaluation of V(Z) in a Type II Reflection Microscope," *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **34,** 341–345.

Chou, C. H., Khuri-Yakub, B. T., Kino, G. S. (1988), "Lens Design for Acoustic Microscopy," *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **35,** 464–469.

Daft, C. M. W., Briggs, G. A. D. (1989), "Wideband Acoustic Microscopy of Tissue," *IEEE Trans. Sonics Ultrason. Ind. Eng. Chem.* **36,** 258–263.

Durr, W., Sinclair, D. A., Ash, E. A. (1980), "High Resolution Acoustic Probe," *Electron. Lett.* **16,** 805–806.

Gilmore, R. S., Tam, K. C., Howard, D. R. (1986), *Acoustic Microscopy from 10 to 100 MHz for Industrial Applications,* General Electric Corporate Research and Development Report No. 86CRD015.

Güthner, P., Fischer, U. Ch., Dransfeld, K. (1989), *Appl. Phys. B* **48,** 89–92.

Hadimioglu, B., Quate, C. F. (1983), "Water Acoustic Microscopy at Suboptical Wavelengths," *Appl. Phys. Lett.* **43,** 1006–1007.

Hadimioglu, B., Foster, J. S. (1984), "Advances in Superfluid Helium Acoustic Microscopy," *J. Appl. Phys.* **56,** 1976.

Hoppe, M., Bereiter-Hahn, J. (1985), "Applications of Scanning Acoustic Microscopy: Survey and New Aspects," *IEEE Trans. Sonics Ultrason. Ind. Eng. Chem.* **32,** 289–301.

Kessler, L. W. (1974), *J. Acoust. Soc. Am.* **55,** 909–918.

Kessler, L. W., Korpel, A., Palermo, P. R. (1971), *Nature* **232,** 110–111.

Khuri-Yakub, B. T., Cinbis, C., Chou, C. H., Reinholdtsen, P. A. (1989), "Near-Field Scanning Acoustic Microscope," in: B. R. McAvoy (Ed.), *Proceedings of the IEEE Ultrasonics Symposium, 1989,* pp. 805–807.

Kino, G. S. (1987), *Acoustic Waves: Devices, Imaging, and Analog Signal Processing,* Englewood Cliffs, NJ: Prentice-Hall.

Kolosov, O., Yamanaka, K. (1993), *Jpn. J. Appl. Phys.* **32,** L1095–L1098.

Korpel, A., Kessler, L. W., Palermo, P. R. (1971), "An Acoustic Microscope Operating at 100 MHz," *Nature* **323,** 110–111.

Korpel, A., Kessler, L. W. (1971), in: A. Metherell (Ed.), *Acoustical Holography,* Vol. 3, New York: Plenum, pp. 23–43.

Kubota, J., Okada, H., Musha, Y., Takishita, Y., Iwasaki, A., Sasaki, S. (1988), "Electronic Scanning of 25 MHz Ultrasound for Imaging IC Packages," in: B. R. McAvoy (Ed.), *Proceedings of the IEEE Sonics and Ultrasonics Symposium,* New York: IEEE, p. 767–770.

Kushibiki J.-I., Chubachi, N. (1985), "Material Characterization by Line-Focus-Beam Acoustic Microscope," *IEEE Trans. Sonics Ultrason. Ind. Eng. Chem.* **32,** 189–212.

Lam, L. K. (1983), *Techniques of Acoustic Microscopy,* Ph.D. Thesis, Stanford University.

Lasky, M. (1977), *J. Acoust. Soc. Am.* **61,** 283–297.

Lemons, R. A., Quate, C. F. (1973), "A Scanning Acoustic Microscope," in: J. de Klerk (Ed.), *Proceedings of the IEEE Ultrasonics Symposium,* New York: IEEE, pp. 18–20.

Lemons, R. A., Quate, C. F. (1974), "Acoustic Microscope—Scanning Version," *Appl. Phys. Lett.* **24,** 163–165.

Lemons, R. A. (1975), *Acoustic Microscopy by Mechanical Scanning,* Ph.D. Thesis, Stanford University.

Liang, K. K., Bennett, S. D., Khuri-Yakub, B. T., Kino, G. S. (1982), "Surface Wave Velocity Measurements at 50 MHz," in: B. R. McAvoy (Ed.), *Proceedings of the IEEE Ultrasonics Symposium, 1982,* Vol. 2, New York: IEEE, p. 604.

Liang, K. K., Bennett, S. D., Khuri-Yakub, B. T., Kino, G. S. (1985a), "Precise Phase Measurements with the Acoustic Microscope," *IEEE Trans. Sonics Ultrason. Ind. Eng. Chem.* **32,** 266–273.

Liang, K. K., Kino, G. S., Khuri-Yakub, B. T. (1985b), "Material Characterization by the Inversion of the V(Z)," *IEEE Trans. Sonics Ultrason. Ind. Eng. Chem.* **32,** 213–224.

Lin, Z. C., Lee, H., Wade, G., Oravecz, M. G., Kessler, L. W. (1987), *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **UFFC-34,** 293–300.

Pohl, D. W., Denk, W., Lang, M. (1984), *Appl. Phys. Lett.* **44,** 652–653.

Pohlman, R. (1937), *Z. Phys.* **107,** 497–507.

Quate, C. F. (1985), "Acoustic Microscopy," *Phys. Today* **38** (8), 34–42.

Reinholdtsen, P. (1989), *Image Processing for an Amplitude and Phase Acoustic Microscope,* Ph.D. Thesis, Stanford University.

Rylander, R. L. (1982), *A Laser-Scanned Ultrasonic Microscope Incorporating a Time-Delay Interferometric Detector,* Ph.D. Thesis, University of Minnesota.

Schoch, A. (1950), *Ergeb. Exakt. Naturwiss.* **23,** 127.

Sokolov, S. J. (1949), *D. K. Akad. Nauk SSSR* [Sov. Phys. Dokl.] **64,** 333–335.

Takata, K., Hasegawa, T., Hosaka, S., Hosoki, S. (1989), "Tunneling Acoustic Microscope," *Appl. Phys. Lett.* **55,** 1718–1720.

Uozumi, K., Yamamuro, K. (1989), "A Possible Novel Scanning Ultrasonic Tip Microscope," *J. Appl. Phys.* **28,** L1297–L1299.

Wade, G. (1987), in: H. W. Jones (Ed.), *Acoustical Imaging,* Vol. 15, New York: Plenum, pp. 1–28.

Wang, K., Wade, G. (1974), in: P. S. Green (Ed.), *Acoustical Holography,* Vol. 5, New York: Plenum, p. 239.

Weglein, R. D., Wilson, R. G. (1978), "Characteristic Material Signatures by Acoustic Microscopy," *Electron. Lett.* **14.**

Weglein, R. D. (1985), "Acoustic Micrometrology," IEEE Trans. Sonics Ultrason. *Ind. Eng. Chem.* **32,** 225–234.

Whitman, R. L., Korpel, A. (1969), *Appl. Opt.* **8,** 1567–1576.

Wickramasinghe, K. (1979), "Contrast and Imaging Performance in the Scanning Acoustic Microscope," *J. Appl. Phys.* **50,** 664–672.

Wickramasinghe, H. K., Petts, C. R. (1982), "Gas Medium Acoustic Microscopy," in: *Scanned Image Microscopy,* London: Academic.

Yamada, K., Shimizu, H., Minakata, M. (1986), "Planar Structure Focusing Lens for Operation at 200 MHz and its Application to the Reflection-Mode Acoustic Microscope," in: B. R. McAvoy (Ed.), *Proceedings of the IEEE Ultrasonics Symposium, 1986,* New York: IEEE, p. 745.

Yamanaka, K. (1982), "Broad Band Pulses of Leaky Surface Acoustic Waves in Acoustic Microscopy," in: B. R. McAvoy (Ed.), *Proceedings of the IEEE Ultrasonics Symposium, 1982,* New York: IEEE, Vol. 2, p. 609.

Zieniuk, J. K., Latuszek, A. (1986), "Ultrasonic Pin Scanning Microscope: A New Approach to Ultrasonic Microscopy," in: B. R. McAvoy (Ed.), *Proceedings of the IEEE Ultrasonic Symposium,* New York: IEEE, pp. 1037–1039.

Further Reading

Auld, B. A. (1973), *Acoustic Fields and Waves in Solids,* New York: Wiley-Interscience.

Briggs, A. (1992), *Acoustic Microscopy,* New York: Oxford Univ. Press.

Khuri-Yakub, B. T., Quate, C. F. (1992), *Selected Papers on Scanning Acoustic Microscopy,* Washington, DC: SPIE Optical Engineering Press.

Kino, G. S. (1987), *Acoustic Waves: Devices, Imaging, and Analog Signal Processing,* Englewood Cliffs, NJ: Prentice-Hall.

SCANNING TUNNELING MICROSCOPY

See TUNNELING MICROSCOPY

SCATTERING OF LIGHT

See RAMAN SCATTERING

SCHOTTKY BARRIERS

L. Magaud and F. Cyrot-Lackmann, *Laboratoire d'Etudes des Propriétés Electroniques des Solides CNRS, Grenoble, France*

Introduction 573
1. **Description and Electrical Behavior** 575
1.1 Description of the Schottky Barrier 575
1.2 Schottky Barrier and Ohmic Contact at Equilibrium 575
1.3 Schottky Barrier and Ohmic Contact under an Applied Bias 576
1.3.1 Schottky Barrier under an Applied Bias 576
1.3.2 Ohmic Contact under an Applied Bias 577
2. **Experimental Description** ... 578
2.1 General Trends 578
2.2 Phenomenological Models and Correlations 580
2.3 Schottky-Barrier Height Measurements 581
2.3.1 Capacitance Measurement, *C-V* 581
2.3.2 Current Measurement, *I-V* 581
2.3.3 Photoresponse, *I-hν* 582
2.3.4 Application of XPS to Schottky-Barrier Measurements 582
2.3.5 Ballistic-Electron Emission Microscopy (BEEM) 583
3. **Theoretical Models** 584
3.1 Schottky Model 584
3.2 Bardeen Model 584
3.3 Defect Model (UDM) 585
3.4 Cowley–Sze Model 585
3.5 Metal-Induced Gap States Model (MIGS) 586
3.6 Correlations between Schottky-Barrier Height, Valence-Band Offset, Dangling-Bond Level, and Transition-Metal Impurity Level 587
3.6.1 Charge-Neutrality Level and the Semiconductor Dangling-Bond Level 587
3.6.2 Correlation with Valence-Band Offset 587
3.6.3 Correlation with Transition-Metal Impurity Levels 587
3.7 Discussion 588
3.7.1 Variation of the Schottky-Barrier Height with Metal Coverage 588
3.7.1.1 Small Coverages, $\theta < 1$ Monolayer (ML) 588
3.7.1.2 Thick Coverages, $\theta \gg 1$ ML ... 588
3.7.2 Comparison of MIGS and Defects Models 588
4. **Microscopic Aspect: the Electronic-Structure Calculation Point of View** ... 589
5. **Conclusion** 589
Acknowledgment 590
Glossary 590
Works Cited 590
Further Reading 591

INTRODUCTION

From an electronic-structure point of view, materials can be classified into two groups: metals and insulators. If the Fermi level falls in a partially filled band, states are available for conduction, even at 0 K, and the material is metallic. On the other hand, if all bands are either empty or fully occupied, the Fermi energy is greater than or equal to the last occupied-state energy, and no neighboring state is available for conduction at 0 K, the system is an insulator. Electrons have to jump from the last filled band (called the valence band) to the lowest empty band (the conduction band) to allow conduction. The last filled states and the first empty ones are separated by a region of forbidden energy: the insulator band gap. If the gap width is not too large (typically <1.4 eV) so

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

that for $T \neq 0$ K electrons can be excited into the conduction band, the system is called a semiconductor. When different materials are brought into contact, interesting properties may be expected. This is the case for some metal–semiconductor interfaces that, depending on the sign of the applied bias, let the current flow or not. They form one type of diode, called a Schottky diode.

The rectifying properties of metal–semiconductor interfaces were reported for the first time by Braun (1874). He attributed this anomalous conductivity to the existence of a highly resistive, thin interfacial layer. In 1938, Schottky (Schottky, 1938) explained the rectifying properties by the presence of a space-charge layer on the semiconductor side of the interface that is depleted of its carriers. For this reason, these metal–semiconductor interfaces were called Schottky barriers.

A model was proposed only 60 years after Braun first reported it, by Mott (1938) and Schottky (1940); it is known as the "Schottky model." The metal and the semiconductor Fermi levels align and no interface dipole is taken into account. This results in a linear dependence of the Schottky barrier height Φ_{bn} on the metal work function Φ_m (Schottky rule). For neutrality reasons, the space charge in the depletion layer is compensated by a charge of opposite sign at the interface, on the metal side. This model is far too simple since it neglects the role of surface or interface states. In fact, it was shown (Schottky 1939) by experiments on metal–selenium contacts that new models had to take these states into account.

Bardeen, in 1947, introduced the idea of Fermi-level pinning by states in the semiconductor band gap. Charge transfer at the interface occurs between the metal and the semiconductor surface states. This idea was developed by Heine (1965) and later by Tejedor *et al.* (1977) and Tersoff (1984), who replaced surface states by interface states since Heine demonstrated that surface states do not subsist at the interface. Their model is known as MIGS, "metal-induced gap states," or IDIS, "induced density of interface states."

In 1979, Spicer (Spicer *et al.*, 1979, 1988) proposed a different origin for the states involved in Fermi-level pinning. The metal deposition liberates energy, which creates defects, and the Fermi level is then pinned at the defect level. This UDM, "unified defect model," involves donor and acceptor defects and explains experimental differences observed for low metal coverages on *n*- and *p*-type semiconductors (photoemission measurements).

Usually, Schottky-barrier heights are determined by electrical measurements: capacitance-voltage (*C-V*) or current-voltage (*I-V*). These techniques do not allow the study of microscopic mechanisms involved in the Schottky-barrier determination. Furthermore, they are sensitive to the interface quality and great care must be taken in how they are applied. X-ray photoemission spectroscopy (XPS) allows the study of the evolution of the Fermi level with the metal coverage. The influence of the deposition temperature is difficult to follow because of the different modes of growth and especially island growth, which leaves uncovered areas on the semiconductor surface.

The discrepancy between model predictions and experimental results led many experimentalists to propose correlations between the Schottky-barrier height and the chemical or physical properties of its constituents. No systematic view emerges from the enormous amount of experimental data, and these correlations only apply to a given type of material. For example, in the transition metal-silicide–silicon interface, Φ_{bn} was related to the core-level shift of the transition metal, or the silicide heat of formation (Andrews and Phillips, 1975). Effective work functions have also been proposed to replace the metal work function in the Schottky rule (Freeouf, 1980; Freeouf and Woodall, 1981). For III-V semiconductors, Φ_{bn} has also been related to the heat of formation of the most stable metal–anion compound (Brillson, 1978).

Despite years of study, no general agreement exists on the mechanisms involved in the Schottky-barrier height determination. Furthermore, the large amount of experimental and theoretical data do not provide a unique description. It appears that several mechanisms compete to determine the Schottky-barrier height but that MIGS is the main one.

Further studies of the physical properties of Schottky barriers are needed from a fundamental point of view and for practical applications. Schottky barriers play an important role in the device industry for diode and transistor fabrication. They are preferred to *p-n* junctions when high switching rates are required. The understanding of the Schottky barrier might allow the development of barriers of predetermined height and hence barrier engineering. It would be especially interesting to obtain small barrier heights on an *n*-type semiconductor for infrared detector fabrication, for example.

This article is organized as follows: Schottky barriers, Ohmic contacts, and their electrical behavior under an applied bias are described in the first section. General remarks on the Schottky-barrier values (nondependence on the metal, the doping, . . .), experimental correlations, and methods of Schottky-barrier determination are presented in the second section. Theoretical models are discussed in the third section. The fourth section deals with the microscopic aspects of the interface and electronic-structure calculations.

The reader is referred to the articles Semiconductors, Elemental, Electronic Properties and Semiconductors, Compound, Electronic Properties in this Encyclopedia for the definition of the physical quantities relevant to semiconductors.

1. DESCRIPTION AND ELECTRICAL BEHAVIOR

1.1 Description of the Schottky Barrier

The physical quantities involved in a Schottky barrier are described in Fig. 1 for the case of a metal on an *n*-type semiconductor. The Schottky-barrier height Φ_{bn} is the difference between the bottom of the conduction band and the Fermi level. At equilibrium, the Fermi levels must be the same in the metal and the semiconductor and charge transfer occurs. In the metal, the charges (Q_m) remain at the surface because of the high density of states ($\sim 10^{22}$ cm^{-3}). In the semiconductor, charges may have two contributions: a surface one (if surface or interface states exist) (Q_{ss}) and a space charge (Q_{sc}). The space charge is due to uncompensated ionized donors. Their density depends on the doping (10^{14}–10^{18} cm^{-3}) and is much lower than in the previous case so that the charge extends into the semiconductor and creates a band bending. The length scales for these contributions are different: Surface charges are on ~1 nm while band bending occurs on ~100–1000 nm. The metal work function Φ_m is the difference between the vacuum level and the Fermi-level position. The semiconductor affinity χ is the difference between the vacuum level and the bottom of the conduction band. The term E_g is the gap width. For silicon, $E_g = 1.12$ eV and $\chi = 4.07$ eV; transition-metal work functions are around 4.5–5.5 eV. In this article, any level in the gap is labeled with respect to the top of the valence band. The interface dipole Δ and the Schottky-barrier height Φ_{bn} are expressed in terms of energies. Except when it is explicitly specified, an *n*-type semiconductor is always considered.

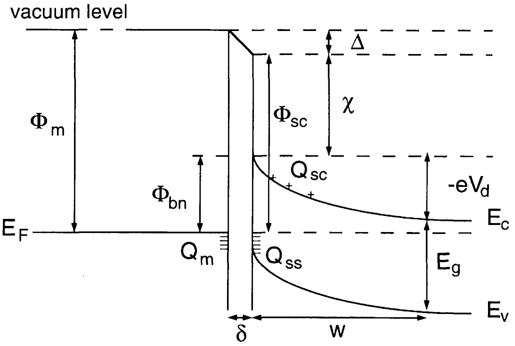

FIG. 1. Energy-band diagram for a metal–*n*-type semiconductor Schottky barrier. E_F is the Fermi level, Δ the interface dipole, E_g the gap width, E_v the top of the valence band, E_c the bottom of the conduction band, W the depletion-zone length, Φ_m the metal work function, Φ_{bn} the Schottky-barrier height, Φ_{sc} the semiconductor work function, V_d the diffusion potential, χ the semiconductor electronic affinity, Q_{sc} the space charge in the semiconductor, Q_{ss} the interface charge in the semiconductor, Q_m the interface charge in the metal, and δ the mean distance between Q_{ss} and Q_m.

1.2 Schottky Barrier and Ohmic Contact at Equilibrium

The formation of the diode is explained (Fig. 2) in a simplified case where no interface states are taken into account (Rhoderick and Williams, 1988). We first consider an *n*-type semiconductor. If the metal work function is larger than the semiconductor one [Fig. 2(a)], electrons diffuse from the semiconductor to the metal and create a band

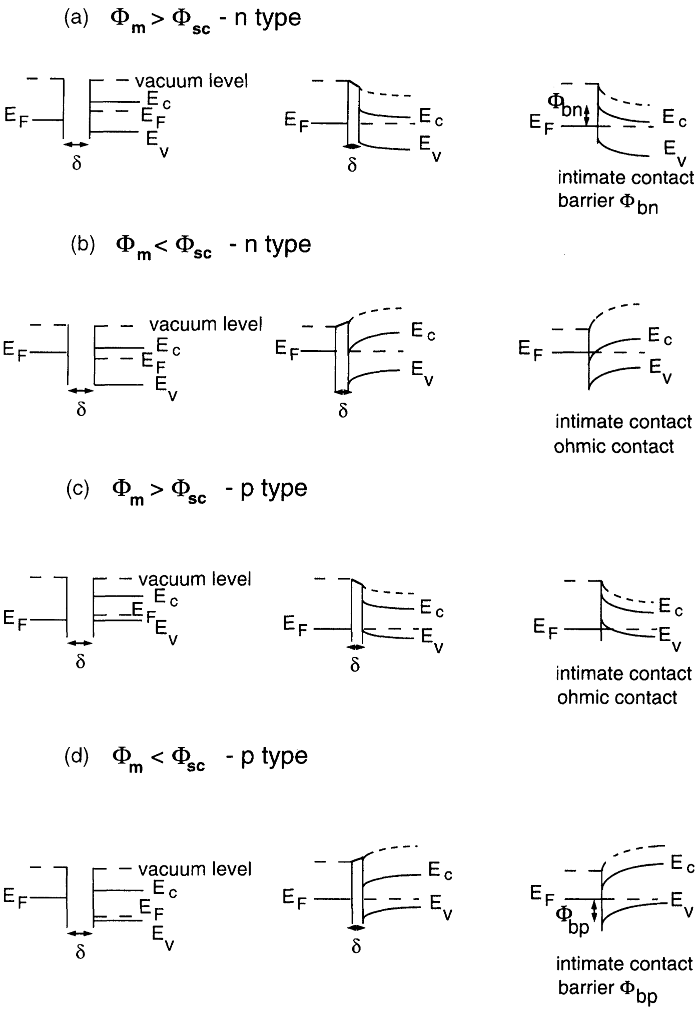
(a) Φm > Φsc - n type
vacuum level
EF
Ec
EF
Ev
δ
Φbn
intimate contact
barrier Φbn
(b) Φm < Φsc - n type
intimate contact
ohmic contact
(c) Φm > Φsc - p type
intimate contact
ohmic contact
(d) Φm < Φsc - p type
Φbp
intimate contact
barrier Φbp

FIG. 2. Metal–semiconductor contact. (a) $\Phi_m > \Phi_{sc}$: *n*-type semiconductor, Schottky barrier. (b) $\Phi_m < \Phi_{sc}$: *n*-type semiconductor, Ohmic contact. (c) $\Phi_m > \Phi_{sc}$: *p*-type semiconductor, Ohmic contact. (d) $\Phi_m < \Phi_{sc}$: *p*-type semiconductor, Schottky barrier.

bending. Electrons have to pass a barrier Φ_{bn} between the semiconductor and the metal; Φ_{bn} is the Schottky-barrier height. In the other case ($\Phi_m < \Phi_{sc}$) [Fig. 2(b)], electrons diffuse from the metal to the semiconductor. They accumulate at the bottom of the conduction band which goes closer to or even below the Fermi level and create an accumulation layer. No barrier appears and the system has the resistivity of the semiconductor; it is an Ohmic contact. For an accumulation layer, the density of states is that of the valence or the conduction band, that is 10^{19} cm^{-3}. The corresponding band bending is then smaller than for the depletion layer. The behavior is reversed for metal on a *p*-type semiconductor interface [Figs. 2(c) and 2(d)].

This description is a simplified one since it ignores the role of surface states. If surface states exist, they contain a charge Q_{ss}; band bending already exists at the free semiconductor surface (see Sec. 3.2). The type of the contact (Ohmic or rectifying) cannot so easily be deduced from the difference between the metal and the semiconductor work functions, and an interface metal–*n*-type semiconductor with $\Phi_m < \Phi_{sc}$ can exhibit rectifying properties.

1.3 Schottky Barrier and Ohmic Contact under an Applied Bias

1.3.1 Schottky Barrier under an Applied Bias If $V_{\text{sc-m}} < 0$ [Fig. 3(a)], the band bending diminishes and the conduction band CB goes up by eV. The metal–semicon-

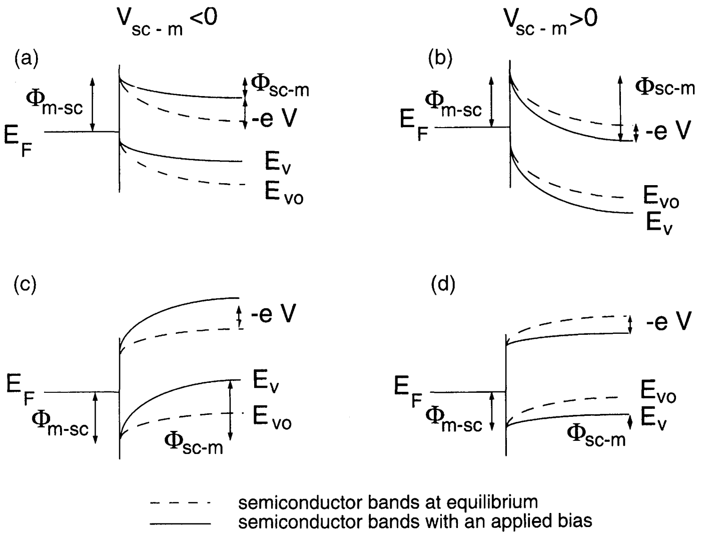
$V_{sc-m}<0$
(a)
Φ_{m-sc}
E_F
Φ_{sc-m}
-e V
E_v
E_{vo}
$V_{sc-m}>0$
(b)
Φ_{m-sc}
E_F
Φ_{sc-m}
-e V
E_{vo}
E_v
(c)
-e V
E_F
E_v
E_{vo}
Φ_{m-sc}
Φ_{sc-m}
(d)
-e V
E_F
E_{vo}
E_v
Φ_{m-sc}
Φ_{sc-m}
semiconductor bands at equilibrium
semiconductor bands with an applied bias

FIG. 3. Schottky barrier under an applied bias $V_{sc\text{-}m}$. (a) $V_{sc\text{-}m} < 0$, *n*-type semiconductor. (b) $V_{sc\text{-}m} > 0$, *n*-type semiconductor. (c) $V_{sc\text{-}m} < 0$, *p*-type semiconductor. (d) $V_{sc\text{-}m} > 0$, *p*-type semiconductor.

ductor barrier $\Phi_{m\text{-}sc}$ stays the same while the semiconductor–metal barrier $\Phi_{sc\text{-}m}$ decreases. The equilibrium is broken, and electrons diffuse from the semiconductor to the metal and create a current from the metal to the semiconductor. If the bias voltage is further increased, the flat-band regime is reached (for $V = V_d$, the diffusion potential) and the electrons can freely flow from the semiconductor to the metal.

If $V_{sc\text{-}m} > 0$ [Fig. 3(b)], the conduction band is lowered and the diffusion of the electrons from the semiconductor to the metal decreases. The metal–semiconductor barrier is unchanged. The contact is blocked. A *p*-type semiconductor exhibits opposite behavior [Figs. 3(c) and 3(d)].

1.3.2 Ohmic Contact under an Applied Bias For an Ohmic contact with an *n*-type semiconductor, electrons diffuse from the metal to the semiconductor and create an accumulation layer. The charge is an accumulation one; there is no zone without carriers and hence no resistive zone. The bias is applied to the whole semiconductor, which is more resistive than the metal. If $V_{sc\text{-}m} < 0$ [Fig. 4(a)], electrons go from the semicon-

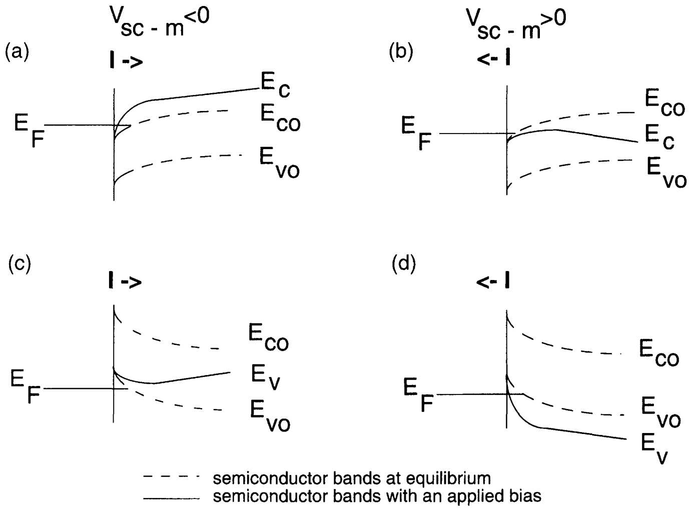
$V_{sc-m}<0$
(a)
I ->
E_F
E_c
E_{co}
E_{vo}
$V_{sc-m}>0$
(b)
<- I
E_F
E_{co}
E_c
E_{vo}
(c)
I ->
E_F
E_{co}
E_v
E_{vo}
(d)
<- I
E_F
E_{co}
E_{vo}
E_v
semiconductor bands at equilibrium
semiconductor bands with an applied bias

FIG. 4. Ohmic contact under an applied bias $V_{sc\text{-}m}$. (a) $V_{sc\text{-}m} < 0$, *n*-type semiconductor. (b) $V_{sc\text{-}m} > 0$, *n*-type semiconductor. (c) $V_{sc\text{-}m} < 0$, *p*-type semiconductor. (d) $V_{sc\text{-}m} > 0$, *p*-type semiconductor.

ductor to the metal. This flow creates a current from the metal to the semiconductor. If $V_{\text{sc-m}} > 0$ [Fig. 4(b)], the current goes from the semiconductor to the metal. Current can freely flow in the two directions. Interfaces with a *p*-type semiconductor exhibit a reverse behavior [Figs. 4(c) and 4(d)].

2. EXPERIMENTAL DESCRIPTION

Even though the study of the Schottky-barrier properties has a long history, it is difficult to draw a global view. Many experimental results are conflicting or only apply to very specific systems. We now review some general properties such as the nondependence of the barrier on the metal used or on the type of doping. We then briefly present a few correlations between the barrier height and chemical properties of its constituents. Part of the difficulty in the analysis of the results might come from the sensitivity of the measurements to experimental conditions. Different experimental methods are presented.

2.1 General Trends

Covalent or III-V semiconductor–metal Schottky barriers do not show much dependence on the nature of the metal. This is illustrated in Fig. 5 for Si and Fig. 6 for GaAs. In these figures, the plain line is the "metal-induced gap states" model prediction (see Sec. 3.5). Most of the values lie close to the mid-gap level. This is not true for barriers involving II-VI ionic semiconductors, which vary with the metal.

Table 1 shows examples of I-V measurements for different metal–semiconductor interfaces for both *n*- and *p*-type doping. In most cases, the sum of the barriers for *n* and *p* types is equal to the gap width:

$$\Phi_{bn} + \Phi_{bp} = E_g. \tag{1}$$

This means that the barrier is independent of the doping. The values tabulated correspond to thick diodes. Exceptions to this rule have been found for low metal coverages on III-V semiconductors (Al, Ga, In, Au/GaAs; Spicer *et al.*, 1988), which exhibit different Fermi-level positions for *n*- and *p*-type

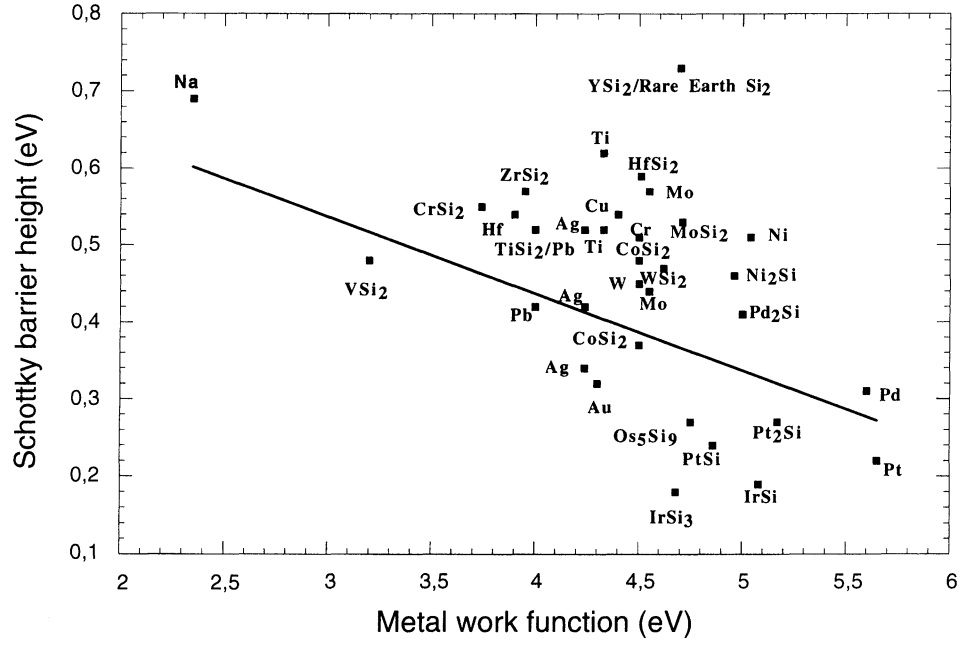

FIG. 5. $E_g - \Phi_{bn} = f(\Phi_m)$ for metal–Si interfaces. The straight line indicates the MIGS prediction.

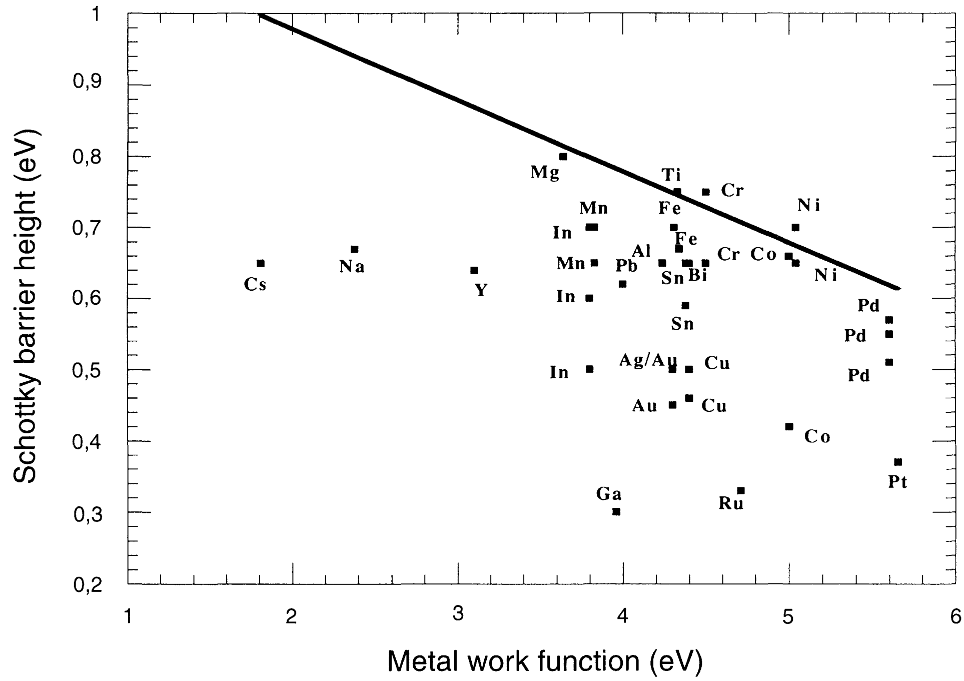

FIG. 6. $E_g - \Phi_{bn} = f(\Phi_m)$ for metal–GaAs interfaces. The straight line indicates the MIGS prediction.

doping. This is discussed in Sec. 3.3. The fact that the Schottky-barrier height does not depend on the semiconductor type of doping is the first meaning of "Fermi-level pinning." The Fermi level lies at a given position in the semiconductor band gap, whatever the doping. A second kind of pinning is usually evoked: a pinning of the Fermi level to its thick interface value for small metal coverages, generally before the deposition of one metal monolayer (1 ML). The Fermi level quickly moves from its value for isolated atoms (small coverage, $\theta \ll 1$ ML) to its saturated value for $\theta \sim 1$ ML. Calculations reproduce this behavior (Ortega *et al.*, 1988).

Table 1. I-V measurements for different metals on *n*- and *p*-type Si and GaAs.

Semiconductor	Metal	Φ_{bn}	Φ_{bp}	$\Phi_{bn} + \Phi_{bp}$
Si	Au	0.8	0.25	1.05
$E_g = 1.12$	Ag	0.67	0.47	1.14
	Cu	0.61	0.50	1.11
	Al	0.72	0.58	1.3
	Cr	0.61	0.50	1.11
	$ErSi_2$	0.39	0.70	1.09
GaAs	Ag	0.88	0.63	1.51
$E_g = 1.42$	Au	0.90	0.42	1.32
	Hg	0.72	0.68	1.4
	Cu	0.96	0.45	1.39
	Al	0.85	0.61	1.46
	Cr	0.77	0.56	1.33

Interfacial reactivity seems to have a small influence on the barrier-height determination. Ag, Au, Cu, and Pd on GaAs give nearly identical barriers (0.5, 0.55 eV) despite having completely different interface reactivities: No reaction exists at an Ag–GaAs interface while Pd on GaAs is highly reactive (Spicer *et al.*, 1988).

Other behaviors are not so clear. The interface geometry seems to have a small influence on the barrier height, but a well-known conflicting example exists: the $NiSi_2$–Si (111) interface exhibits two barriers, 0.65 and 0.75 eV, depending on whether or not $NiSi_2$ is rotated by 180° with respect to the silicon. This case will be further discussed in Sec. 4. There are also conflicting experimental results on the role of an interfacial oxygen or oxide layer.

These general trends apply to most metal–semiconductor interfaces. A huge amount of experimental data exists, and exceptions can be found in many cases.

2.2 Phenomenological Models and Correlations

Many correlations have been proposed between the Schottky-barrier height and the chemical properties of its constituents. Only a few will be discussed here; for a more detailed review, see Mönch (1990).

In the transition metal-silicide–silicon case, several relations were proposed between Φ_{bn} and the silicide bulk properties. Andrews and Phillips (1975) found a very good correlation of Φ_{bn} with the silicide heat of formation but no model was proposed. It corresponds to the straight line in Fig. 7. In this figure, more experimental points have been added and the relation is not so clear. Dependence on the transition-metal core-level shift has been proposed, too.

Propositions have been made to keep the Schottky rule and change the metal work function. In Tm–Si interfaces, Freeouf (1980) assumed the existence of a mixed phase close to the interface and replaced the transition-metal work function Φ_m by the geometrical mean, $(\Phi_m \Phi_{Si}^4)^{1/5}$, even though it does not correspond to the silicide work function. A similar approach was proposed for III-V and II-VI semiconductors. Anion microclusters at the interface were assumed to modify Φ_m so that the Schottky rule remains valid (Freeouf and Woodall, 1981), but no experimental evidence of their existence has been found.

New compounds can be formed at reactive metal–III-V or –II-VI semiconductor interfaces, even at room temperature, and Brillson (1978) related the barrier height with the heat of formation of the most stable metal–anion compound.

A common-anion rule has been proposed, too: Semiconductors with identical anions should exhibit very similar Schottky-barrier heights with the same metal, but this is contrary to what many experimental results give.

The idea of a correlation between the barrier height and bond strength (heat of formation) is very interesting but gives poor results. This might indicate that more subtle phenomena are involved in the barrier deter-

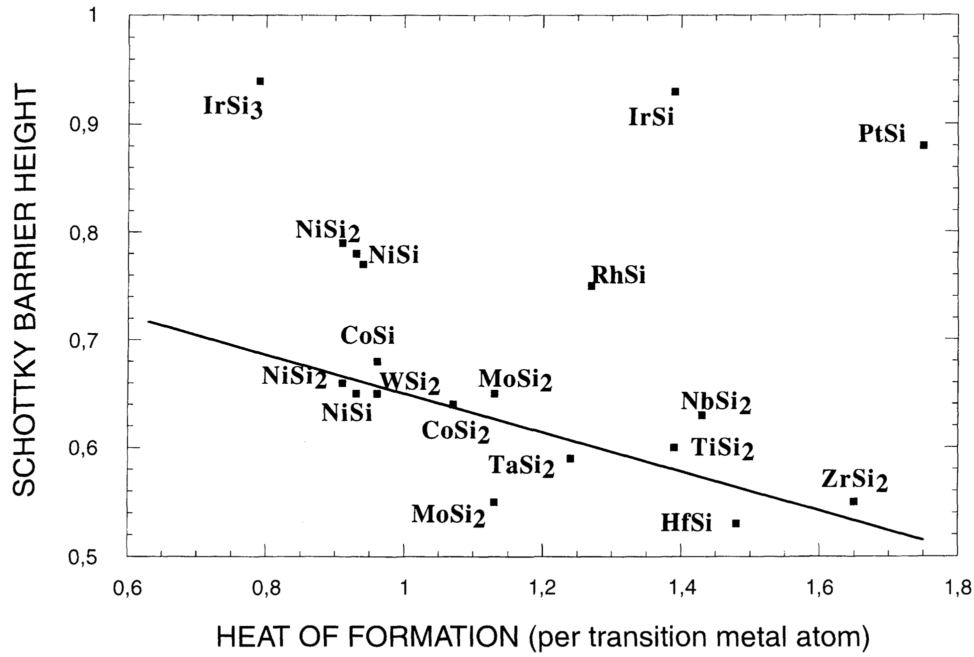

FIG. 7. $\Phi_{bn} = f(\Delta H_f)$ for transition-metal–silicon interfaces. The straight line indicates Andrews and Phillips correlation (Andrews and Phillips, 1975).

mination. It will be discussed along with the models.

2.3 Schottky-Barrier Height Measurements

Five techniques of Schottky-barrier height determination will now be presented (Rhoderick and Williams, 1988). They give the same information but face different types of problems. The combined use of at least two of them should give more confidence in the determined values.

2.3.1 Capacitance Measurement, *C-V* The relation between W, the depletion-zone width, and V_d, the diffusion potential, is given by

$$W = (2\epsilon V_d/eN_d)^{1/2}.$$

If a bias V is applied, $V = V_{\text{m-sc}} > 0$ (direct), the potential becomes $V_d - V$ and

$$W = [2\epsilon(V_d - V)/eN_d]^{1/2}. \tag{2}$$

If V is modified, W is modified too and a differential capacitance can be defined:

$$C = \delta Q/\delta V. \tag{3}$$

The charge associated with the band bending, $-e(V_d - V)$, is given by

$$Q_{sc} = [2\epsilon e N_d(V_d - V)]^{1/2}; \tag{4}$$

combining with Eqs. (2) and (3) gives

$$C = (\epsilon e N_d/2)^{1/2}(V_d - V)^{-1/2} = \epsilon/W: \tag{5}$$

C is the capacitance of a flat capacitor of width W. Equation (5) can be rewritten as

$$C^{-2}(V) = 2(V_d - V)/\epsilon e N_d; \tag{6}$$

thus $C^{-2}(V)$ is linear in V. The term N_d can be determined from the slope, and the abscissa at the origin gives V_d. If N_d is not constant, $C^{-2}(V)$ is no longer linear.

In practice, the junction is reverse polarized and small oscillations δV are superposed. If no trap exists, the variation δQ corresponds to the variation of the free-carrier concentration at the edge W of the depletion zone on a width δW. Figure 8 shows a typical C^{-2}-V spectrum.

If traps exist, they contain part of the charge δQ. To get rid of them, one has to perform measurements at high frequency (>100 kHz). For such a frequency, only the semiconductor depletion charge changes. Another problem then arises: the series resistance contribution, which disappears at low frequency. A compromise must be found to minimize these two contributions.

2.3.2 Current Measurement, I-V If thermoionic emission is assumed to be the only current transport mechanism, the current density J in the forward direction with $V > 3kT/q$ is given by

$$J = J_0 \exp(eV/kT), \tag{7}$$

where $J_0 = A^*ST^2 \exp(-\Phi_{bn}/kT)$ is the saturation current, A^* the effective Richardson constant, T the temperature, V the applied bias, and S the diode area. For simplicity, we do not introduce the Schottky-barrier lowering (due to image force), but it must be taken into account in barrier-height determination (more details can be found in Rhoderick and Williams, 1988). Very often, experimental measurements do not follow the

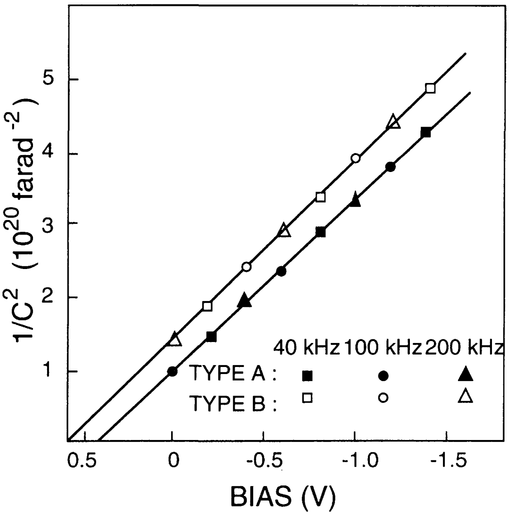

FIG. 8. Typical *C-V* characteristics for type A and B $NiSi_2$/Si (111) interfaces at three different frequencies (from Tung *et al.*, 1986). Type A: $NiSi_2$ and Si have the same crystallographic axis; type B: $NiSi_2$ is rotated by 180° with respect to Si.

linear relation between $\ln(J)$ and V. One of the reasons is that A^* and Φ_{bn} depend on the applied bias and that other current-transport mechanisms are involved. To take this effect into account, J is written as

$$J = J_0 \exp\left(\frac{eV}{nkT}\right) \quad \text{and then } n = \frac{e}{kT}\frac{\delta V}{\delta \ln(J)}, \tag{8}$$

Here n is the *ideality factor*. It must be close to 1 to deduce Φ_{bn} from the I-V characteristic ($n < 1.1$). If $\ln(J)$ is plotted against V, J_0 is the extrapolated value for $V = 0$. The value of n is deduced from the slope. Then Φ_{bn} is given by

$$\Phi_{bn} = kT \ln(A^* ST^2/J_0). \tag{9}$$

The choice of A^* is not crucial; a change by a factor 3 results in a variation of kT for Φ. A typical example is given in Fig. 9.

2.3.3 Photoresponse, *I-hν* A monochromatic light is directed upon the metal surface (front illumination). When a photon arrives at the surface, an electron of the metal can take its energy $h\nu$. If this energy is large enough, the electron can go from the metal to the semiconductor conduction band. In the semiconductor, electrons will be accelerated because of the electric field in the space-charge zone. Electrons can cross the barrier if $h\nu > \Phi_{bn}$, and a photocurrent I_{ph} is induced. The process is described in Fig. 10. An electron with an energy E_1 receives the photon energy $h\nu$. It is excited to $E_2 = E_1 + h\nu$ and creates a hole of energy E_1. If $E_2 > \Phi_{bn}$, the electron can pass the barrier and is collected in the semiconductor. The hole, even if E_1 is smaller than E_v, will be pushed back to the metal. The photocurrent I_{ph} from the semiconductor to the metal is given by Fowler's formula:

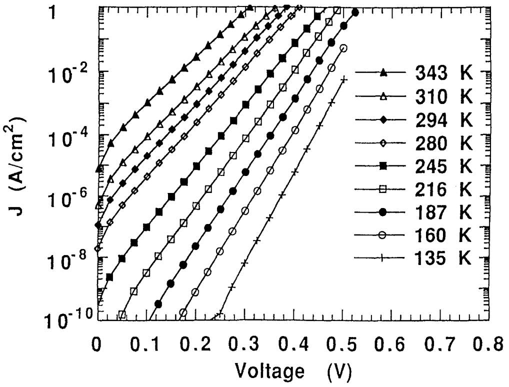

FIG. 9. Typical *I*-*V* characteristics for Al-*n*-GaAs samples for various temperatures. Ideality factors are between 1.01 and 1.03 (from Muret *et al.*, 1992).

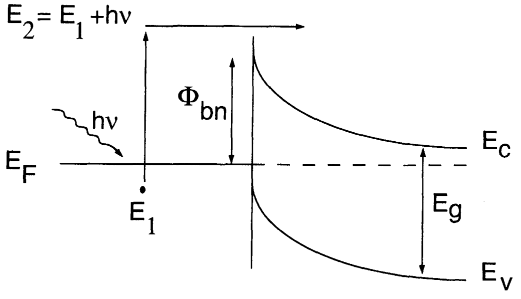

FIG. 10. Principle of photoresponse measurement.

$$I_{ph} = AT^2\{x^2/2 + \pi^2/6 - (e^{-x} - e^{-2x}/2^2 + e^{-3x}/3^2 - \cdots)\} \quad \text{for } x > 0. \tag{10}$$

Here A is a constant, T is the temperature, and $x = (h\nu - \Phi_{bn})/kT$. If $x > 6$, only the parabolic term is kept and

$$I_{ph} \sim (h\nu - \Phi_{bn})^2. \tag{11}$$

When $(I_{ph})^{1/2}$ is plotted against ν, a straight line should be obtained and Φ_{bn} is given by the extrapolated value at the energy axis. A typical spectrum is shown in Fig. 11. If the metal thickness is too large, the photocurrent will not be linear and absorption occurs in the metal. The thickness should be of the order of the electron mean free path (a few times 10 nm).

2.3.4 Application of XPS to Schottky-Barrier Measurements XPS has been extensively used for band bending and Schottky-barrier height determination for over a decade (Hollinger, 1987). It allows the examination of the first stages of Schottky-barrier formation. The principle is explained in Fig. 12.

Because of the small penetration length of XPS, bands can be considered as flat in the region of interest. The binding energy E_c of a semiconductor core level is measured for the free surface and for increasing metal cover-

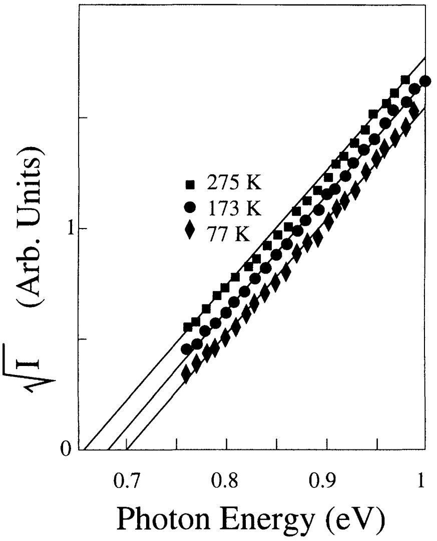

FIG. 11. Typical *I*-$h\nu$ spectra of $CoSi_2$–Si(*n*) (from Duboz *et al.*, 1989).

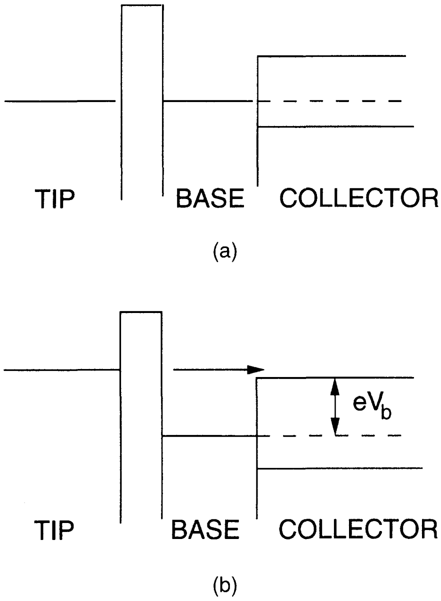

FIG. 13. Schematic energy-band diagram for the three-terminal BEEM experiment. Collector current I_c is measured between the base (metal) and the collector (semiconductor). (a) The energy-band diagram for zero tunnel bias, $V = 0$. (b) The energy-band diagram for tunnel bias greater than the barrier voltage, $eV > eV_b$. (From Kaiser and Bell, 1988.)

age. The core-level shift δ ($=E_{ci} - E_c$) reproduces the variation of the band bending. It is then possible to follow the evolution of the Schottky-barrier height $\Phi_{bn} = CB_i - E_{Fi} + \delta$ as a function of the metal coverage.

2.3.5 Ballistic-Electron Emission Microscopy (BEEM) This new technique for Schottky-barrier height determination was proposed by Kaiser and Bell (1988; see also Bell *et al.*, 1988). It is based on a scanning tunneling microscopy (STM) method and can be used for Schottky barriers, semiconductor heterojunctions, and other buried-interface studies. The principle of BEEM is described in Fig. 13.

A STM tip is positioned near the metal surface to allow electron tunneling between the tip and the metal (base). This injects ballistic electrons into the base. The tunnel current I_t is kept constant. If the metal width is smaller than the attenuation length (>10 nm), these ballistic electrons may propagate up to the interface. If the metal–tip bias V is smaller than V_b, electrons cannot enter the semiconductor. When V is increased above V_b, ballistic electrons enter the semiconductor (collector) and the collector current I_c increases drastically. Figure 14 shows a BEEM I_c-V spectrum for the Au–Si structure recorded by Kaiser and Bell. A theoretical treatment was developed by these authors, who proposed that for V just above V_b, the

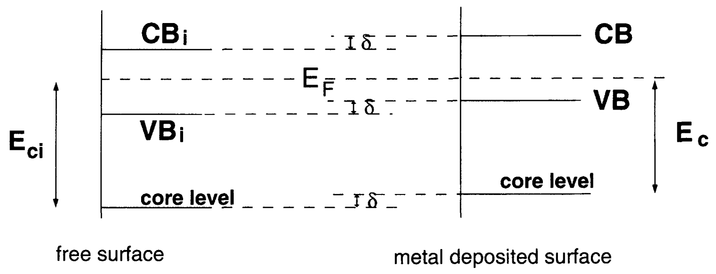

FIG. 12. Application of XPS to Schottky-barrier height measurement.

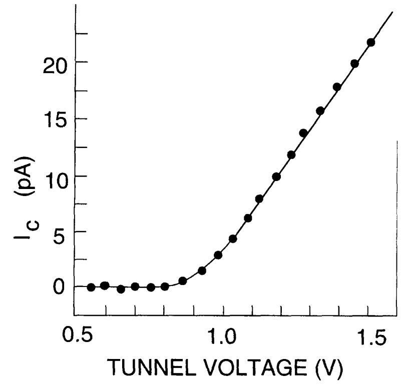

FIG. 14. BEEM I_c-V (dots) for the Au–Si structure recorded at a tunneling current I_t = 0.87 mA (from Kaiser and Bell 1988).

I_c-V spectrum behaves as $(V_b - V)^2$, but the power of this dependence is disputed.

This method allows the determination of the spatial variations of the Schottky-barrier height with nanometer-scale resolution.

3. THEORETICAL MODELS

The first model for Schottky barriers (Mott, 1938: Schottky, 1940) appeared more than 60 years after Braun's report and was soon shown incorrect by experiments. In 1947, the understanding of the semiconductor electronic structure allowed Bardeen to propose the idea of Fermi-level pinning by states in the semiconductor band gap, which is the key point of today's theories.

3.1 Schottky Model

This model (Fig. 15) (Schottky, 1940, 1942) assumes that no state exists in the band gap of the semiconductor so that no interface dipole can be induced. Only space charges exist in the semiconductor. At equilibrium, the Fermi levels must align in the metal and the semiconductor. Charge transfer occurs and creates an accumulation charge Q_m on the metal surface and a space charge Q_{sc} in the semiconductor. The whole system is neutral so that $Q_m = -Q_{sc}$.

As Fig. 15 shows,

$$\Phi_{bn} = \Phi_m - \chi. \tag{12}$$

The Schottky barrier depends linearly on the metal work function. It is a property of bulk metal and semiconductor. The same idea has been developed for band lineup in heterojunctions (the reader is referred to the article HETEROJUNCTIONS, SEMICONDUCTOR in this Encyclopedia for further information). This model gave good results for ionic semiconductors but does not agree with experiments on covalent and III-V semiconductors.

3.2 Bardeen Model

The Schottky model ignores the role of surface states at the metal–semiconductor interface. It is well known that a semiconductor free surface may show states in the gap, and Bardeen (1947) proposed a model where these states play a major role. They are filled up to some level Φ_0 and carry a charge Q_{ss} [Fig. 16(a)]. Because of the neutrality of the system, Q_{ss} must be compensated by a space charge Q_{sc} inside the semiconductor so that $Q_{ss} + Q_{sc} = 0$. This space charge is associated with band bending and depends on the density of surface states.

At the metal–semiconductor contact [Fig. 16(b)], charge transfer creates a supplementary charge Q_{ss}, compensated by the same charge of opposite sign on the metal side. If the density of surface states is high enough, the surface locally behaves like a metal

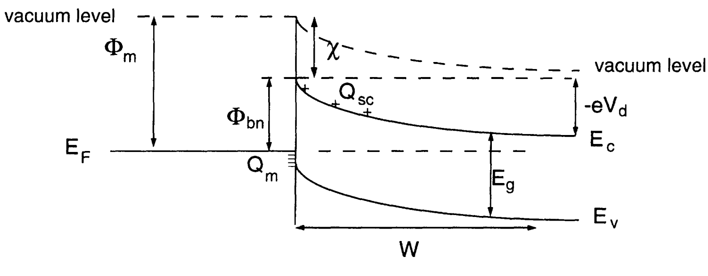

FIG. 15. Schottky-model description of the Schottky barrier for an n-type semiconductor. $\Phi_{bn} = \Phi_m - \chi$.

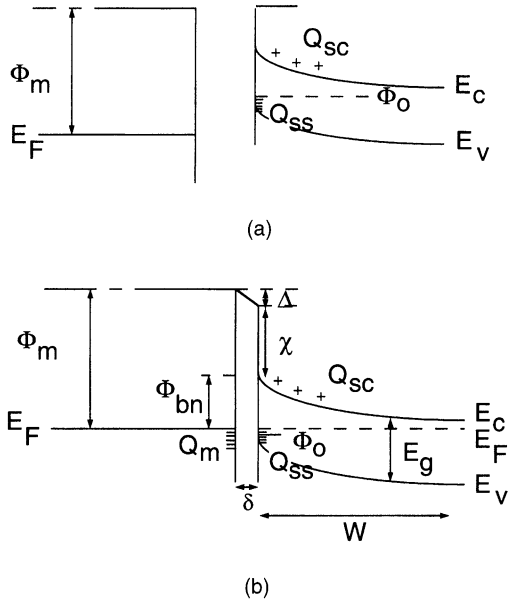

FIG. 16. Fermi-level pinning by states in the semiconductor band gap (Bardeen, Cowley–Sze, MIGS models). (a) Metal and semiconductor free surfaces. (b) Metal and semiconductor in contact.

whose Fermi level is Φ_0. The high density of surface states can accommodate all the charge transferred from the metal without noticeably moving E_F from Φ_0. The charge transfer creates a dipole Δ that tends to align E_F in the metal with Φ_0. The Schottky barrier is given by

$$\Phi_{bn} = E_g - \Phi_0. \tag{13}$$

The space-charge region in the semiconductor remains unchanged. The Schottky barrier is a property of the semiconductor surface and is completely independent of the metal. Here E_F is pinned near Φ_0 by the high density of surface states.

Bardeen is the originator of the idea of Fermi-level pinning by states in the semiconductor band gap. This is the key point of the models, but the nature of these states is still under discussion.

3.3 Defect Model (UDM)

Cleaved 110 surfaces of III-V semiconductors show no surface states in the band gap and the Bardeen model cannot be applied. Spicer *et al.* (1979) proposed a model where the Fermi level is pinned by defect levels in the band gap. The energy needed for the defect creation is provided by the metal deposition (condensation heat, etc.). The model applies to III-V–insulator and III-V–metal interfaces; this is why it is called the "unified defect model" (UDM). If Φ_1 is the defect level inside the semiconductor band gap, relative to the top of the valence band, and if the density of defects is high enough, charge transfer between the metal and the defect states creates a dipole that aligns E_F and Φ_1. E_F is pinned by these states and

$$\Phi_{bn} = E_g - \Phi_1. \tag{14}$$

When this model was first proposed, not much was known about defects, and Spicer first attributed their origin to anion or cation vacancies. Further studies of the GaAs system showed that antisites are the main defects. In GaAs, the key defect is the $\mathrm{As_{Ga}}$ antisite, which is a double donor with levels at 0.52 and 0.75 eV above the top of the valence band. These donors are partly compensated by acceptors, most likely $\mathrm{Ga_{As}}$ antisite. The Fermi-level position depends on the doping and the ratio $R = \mathrm{As_{Ga}}/\mathrm{Ga_{As}}$. In fact, $R > 1$ because GaAs crystals are grown from the As-rich side of the phase diagram. Interfaces with *n*- (*p*-) type GaAs are pinned at 0.75 (0.52) eV. (For further details, see Spicer *et al.*, 1988).

This model has been successfully applied to many III-V–metal systems. Zur *et al.* (1983) modeled the role of donor and acceptor defects on the Schottky-barrier height in two different cases: thick overlayer and thin overlayer (free semiconductor surface with defects). They found that for thick diodes Fermi levels cannot be very different (~0.1 eV) for *n*- and *p*-type semiconductors, but for submonolayer coverages, substantial differences may exist. The UDM explains why E_F can be pinned at different levels in the gap for *n*- and *p*-type semiconductors since it involves donor and acceptor defects.

3.4 Cowley–Sze Model

There is a general agreement on the pinning of the Schottky barrier by interface states, but the nature of these states is still controversial. The development proposed by

Cowley and Sze (1965) only assumes their existence and that their density D_s is constant between Φ_0, the filling level of surface states, and the final Fermi level [Fig. 16(b)]. ϕ_0 corresponds to a local charge neutrality of the surface. The interface charge due to metal–semiconductor charge transfer is

$$Q_{ss} = -eD_s(E_g - \Phi_0 - \Phi_{bn})\ \mathrm{C\,cm^{-2}}. \qquad (15)$$

The calculation of the depletion width W of a semiconductor with a band bending $-eV_d$ is a classical electrostatic problem. It is done by solving Poisson's equation. Under the assumptions that for $x < W$, $\rho = -eN_d$, and for $x > W$, $\rho = 0$, $dV/dx = 0$, we find

$$W = (2\epsilon/eN_dV_d)^{1/2}; \qquad (16)$$

here N_d is the donor density and is assumed to be constant in the whole depletion zone. At the interface, the charge per unit area associated with the band bending is given by

$$Q_{sc} = [2\epsilon_s N_d(\Phi_{bn} - \phi_n)]^{1/2}\ \mathrm{C\,cm^{-2}}. \qquad (17)$$

The value of $\phi_n = E_c - E_f$ depends on the doping. The whole system is neutral so that $Q_m = -(Q_{sc} + Q_{ss})$. The dipole Δ at the interface can be obtained by Gauss's law on the surface of the metal and the semiconductor:

$$\Delta = -e\delta Q_m/\epsilon_i. \qquad (18)$$

From Fig. 16(b), Δ can also be expressed as

$$\Delta = \Phi_m - \chi - \Phi_{bn}, \qquad (19)$$

which comes from the fact that E_F is constant in the whole system. Combining relations (17), (18), and (19), one can express Φ_{bn} as

$$\Phi_{bn} = (1 + \alpha)^{-1}(\Phi_m - \chi) + [1 - (1 + \alpha)^{-1}] \times (E_g - \Phi_0) + V, \qquad (20)$$

where $\alpha \doteq e^2D_s\delta/\epsilon_i\epsilon_0$. For $\delta \sim 0.4$–0.5 nm, $N_d < 10^{18}\ \mathrm{cm^{-3}}$, $\epsilon_i \sim 1$, $\epsilon_s \sim 12$, V can be neglected. If only the work function changes when the metal changes, the barrier varies with a slope $S = (1 + \alpha)^{-1}$. Equation (20) can be rewritten as

$$\Phi_{bn} = S(\Phi_m - \chi) + (1 - S)(E_g - \Phi_0). \qquad (21)$$

If $D_s \to 0$, $S \to 1$, $\Phi_{bn} = \Phi_m - \chi$; this is the Schottky-model prediction. If $D_s \to \infty$, $S \to 0$, $\Phi_{bn} = E_g - \Phi_0$; this is the Bardeen-model prediction. The Schottky and Bardeen models are two limiting cases of this description. For the general case, $0 < S < 1$.

3.5 Metal-Induced Gap States Model (MIGS)

In 1965, Heine argued that real surface states cannot exist at the metal–semiconductor interface but that they are replaced by metal-induced states in the semiconductor band gap. The metallic wave-function tails decrease exponentially in the semiconductor and create a charge transfer. The semiconductor surface states go through a resonance (in opposition to Bardeen's theory). The model has been further developed by Tejedor *et al.* (1977), who introduced the "charge-neutrality level" (CNL) concept: Surface or interface states in the gap equally come from the valence and the conduction bands. States closer to the valence (conduction) band have a predominantly valence (conduction) character. For some level Φ_0, the valence- and the conduction-band contributions are equal. The term Φ_0 corresponds to a local neutrality of the surface and is the so-called "charge-neutrality level."

The charge transfer between the metal and the semiconductor creates a dipole that tends to set the metal Fermi level equal to Φ_0. If the density of interface states is high enough, the difference between the metal Fermi level and the semiconductor CNL is nearly completely screened by the charge transfer. The Schottky barrier is then given by Eq. (21) for small S. Common values of S are 0.1 for covalent and III-V semiconductors and 0.3 for ionic II-VI semiconductors.

The small dependence of the Schottky-barrier height on the metal used for a given semiconductor can be explained within this model. If the density of interface states is high enough, they can absorb a charge transfer due to a change in the metal work function without noticeably moving the Fermi level of the whole system.

Tersoff (1984) calculated the CNL for different semiconductors by coupling the states of the semi-infinite semiconductor complex band structure to the metallic wave-function tails. They are shown in Table 2. For silicon, $\Phi_0 = 0.36$ eV. This gives Schottky-barrier heights in the 0.6–0.8-eV range for all metal–n-Si interfaces, in good agreement with most experimental results. These values were obtained for a continuum of metallic states. If the metal DOS were made of discrete peaks for the energies in the band gap region, E_F would be pinned at one of these peaks, maybe far away from Φ_0.

The physical meaning of the CNL is fairly vague. Its value has been related to the semiconductor dangling-bond level (see Sec. 3.6.1, and Lannoo and Friedel, 1991). The existence and the role of the interface states have been confirmed by first-principles calculations (Yndurain, 1971; Louie *et al.*, 1977). They showed that a smooth density of states appears in the semiconductor band gap, close to the interface.

MIGS and IDIS are two different names for the same model. IDIS (induced density of interface states) applies for metal–semiconductor and semiconductor–semiconductor interfaces while MIGS is only for metal–semiconductor junctions.

3.6 Correlations between Schottky-Barrier Height, Valence-Band Offset, Dangling-Bond Level, and Transition-Metal Impurity Level

The charge-neutrality level is the basic key of the MIGS model, but its definition is rather vague. Lannoo and Friedel (1991) established the equivalence of the CNL to the semiconductor dangling-bond level and gave it a more physical meaning. Within this framework, the correlations between the Schottky-barrier height and various properties such as valence-band offset and transition-metal impurity levels can easily be explained.

Table 2. Charge-neutrality levels calculated by Tersoff (1984).

Sc	CNL	Sc	CNL
Si	0.36	GaSb	0.07
Ge	0.18	GaP	0.81
AlAs	1.05	InP	0.76
GaAs	0.7		
lnAs	0.5		

3.6.1 Charge-Neutrality Level and the Semiconductor Dangling-Bond Level We first consider an ideal metal–covalent-semiconductor interface with no rearrangement of atoms (no relaxation nor reconstruction). Each dangling bond interacts with the metal states and becomes a resonant state (Lannoo and Friedel, 1991). If the resonance is symmetric, it is centered on the dangling-bond level E_{db}. There is one electron per dangling bond and so the Fermi level is equal to E_{db}. The surface is neutral, and E_{db} is the charge-neutrality level.

III-V semiconductor nonpolar surfaces exhibit two different dangling bonds, one for the anion (level E_a) and one for the cation (E_c), which go through a resonance because of the coupling with the metal wave functions. If the resonance is broad enough (the resonance width large compared with $E_c - E_a$) and the metal–anion and metal–cation couplings do not differ too much, E_F is pinned at the average dangling-bond level $(E_c + E_a)/2$. The CNL is then $(E_c + E_a)/2$, and E_F can be pinned in the gap even though the free surface does not show states in it.

3.6.2 Correlation with Valence-Band Offset The IDIS model is generally used to describe semiconductor heterojunctions. The same argument as above can be applied to band lineup, and one then finds that the two semiconductor dangling-bond levels must align. If two interfaces between the same metal M and two different semiconductors $SC1$ and $SC2$ have a difference in the Schottky-barrier heights $\Delta\Phi$ ($= \Phi_1 - \Phi_2$, Fig. 17), $\Delta\Phi$ is then the valence-band offset ΔEv of the heterojunction $SC1$–$SC2$.

3.6.3 Correlation with Transition-Metal Impurity Levels The vacuum level was first proposed as a reference level for transition-metal impurities in semiconductors. Tersoff and Harrison in 1987 argued that transition-metal impurities are pinned on the bulk cation vacancy level.

Delerue *et al.* (1988) with a defect molecule model and charge-dependent tight-binding calculations proposed that transition-metal impurity levels are pinned at the average dangling-bond level. Transition atom levels can then be used to determine Schottky-barrier height or valence-band offset.

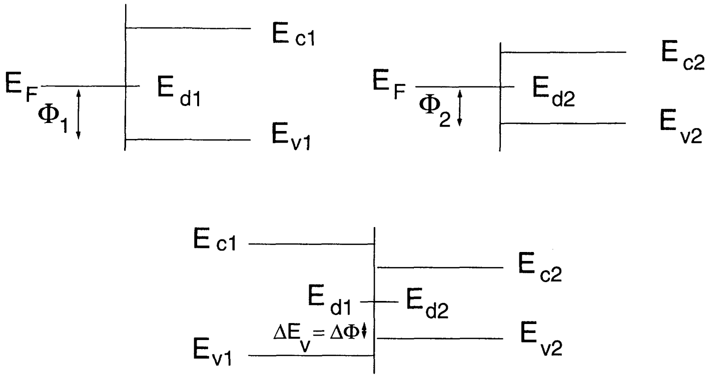

FIG. 17. Correlation between Schottky-barrier height and valence-band offset. E_{d1} (E_{d2}) is the dangling-bond level of semiconductor 1 (2).

3.7 Discussion

It is difficult to choose between the defects or the MIGS model because most of the time they give the same predictions. In the following section, we review some experimental evidence of the Schottky-barrier behavior in light of the two models.

3.7.1 Variation of the Schottky-Barrier Height with Metal Coverage

3.7.1.1 Small Coverages, $\theta < 1$ Monolayer (ML). For very small metal coverages, $\theta < 1$ ML, some systems have been found to have two different pinning positions depending on the type of doping. This is the case for Cs, In, Ga, and Al on GaAs (Spicer *et al.*, 1988). In this case, metal atoms do not have their bulk metallic properties and the MIGS model does not apply. The semiconductor space charge must compensate the surface charge. The dipole length is of the order of the Debye length of the semiconductor, which is very large (1000 nm). The dipole potential (~1 eV) is obtained with a small interface density (10^{12} cm^{-2}) and a large distance. This density of interface states can be provided by defects; acceptors will pin the Fermi level for *n*-type semiconductor and donors for *p*-type.

3.7.1.2 Thick Coverages, $\theta \gg 1$ ML. The compensating charge to the interface charge in the semiconductor is provided by the metal. The screening length is about the Thomas–Fermi wavelength—that is, a few angstroms. A high density of interface states (10^{14} cm^{-2}) is then needed. This density is easily reached within the MIGS model where at least one dangling bond per surface atom is available. On the other hand, 10^{14} cm^{-2} is a very large quantity of defects, which cannot exist in many interfaces of high crystallographic quality.

Submonolayer measurements need careful analysis. Photovoltage effects can induce a difference in the pinning positions for *n*- and *p*-type semiconductors. The growth mode is important too. Island growth can create metallic islands for small coverages, and MIGS can exist. Another problem can come from the fact that part of the semiconductor surface remains free and the Fermi level is set at the semiconductor surface pinning position instead of the metal–semiconductor interface one.

3.7.2 Comparison of MIGS and Defects Models

The success and failure of MIGS and defects models are summarized in Table 3.

Schottky-barrier heights with Si and GaAs do not show linear variation in Φ_m (as predicted by the MIGS model) (Figs. 5 and 6), but the MIGS model description seems to be predominant. The electronegativity difference between the metal and the semiconductor influences the charge transfer at the interface and so the variations of E_F about Φ_0. In general, there must be a competition between the defect density and the MIGS density. When the MIGS pinning is not efficient enough, defects must play a role. Strain could have some effects too.

Schottky, Bardeen, MIGS, and defects are macroscopic models, based on macroscopic values, and therefore they cannot take into account the microscopic structure of the interface. This requires the use of complex

Table 3. Comparison of MIGS and defects models

	Defects	MIGS
Different pinning positions for *n*- and *p*-type semiconductor for small coverage	Yes—different defects are involved for *n* and *p* types	No—but experimental difference can come from the photovoltage effect
Pinning for nonpolar *III-V* semiconductor surfaces (no surface state in the gap)	Yes, no need of surface states	Yes—pinning to the average dangling-bond level if the resonance is broad enough
Pinning for small coverage $\theta <$ 1 ML	Yes	No—MIGS do not exist
Pinning for thick coverage	Yes, with a defect density of 10^{14} cm^{-2}, difficult to imagine	Yes—pinning by MIGS
Relative independence of the SBH with respect to the metal used	Yes—the energy given by the metal deposition creates native defects that are a bulk property of the semiconductor	Yes—if the MIGS density is high enough, they can absorb any supplementary charge due to a modification of the Fermi level in the metal
Band lineup in heterojunctions	No	Yes—IDIS model

electronic-structure calculation methods, which are briefly discussed in the following section.

4. MICROSCOPIC ASPECT: THE ELECTRONIC-STRUCTURE CALCULATION POINT OF VIEW

The microscopic geometry of the interface might have some importance in the Schottky-barrier height determinations. The strength of metal–semiconductor bonds at the interface and the positions of the neighbors are important parameters that are ignored in macroscopic models.

Electronic band-structure calculations within the local-density approximation take the system geometry into account. They are known to give very good results for bulk metals. Problems arise when they are applied to semiconductors. These methods are based on ground-state theory, and in principle, they cannot reproduce the semiconductor gap. They give gap widths about half the experimental values. In practice, the top of the valence band and even the bottom of the conduction band are well described: only the gap width is reduced.

Dealing with interfaces brings out another problem: the loss of translational symmetry. These methods only apply to periodic systems. A supercell technique is then needed: An infinity of interfaces are juxtaposed to restore periodicity.

As an example, let us consider the $NiSi_2$–Si (111) interface. The system exists with two orientations:

1. $NiSi_2$ and Si axes are identical; and
2. $NiSi_2$ is rotated by 180° with respect to Si.

This results in a change of the third neighbors of Ni interfacial atoms, and experimental barrier heights differ by 0.1–0.15 eV. Whether this difference is due to defects or geometry was an important question in the late 1980s. Even though results (linear muffin-tin orbitals method) from different authors disagree somewhat, they always give different barrier heights for the two orientations. This confirms most experimental results that the difference in barrier heights is more probably due to the difference in geometry than to defects.

Despite all the difficulties of *ab initio* electronic-structure calculation methods, they can give very interesting information on the role of the interface geometry because they allow a microscopic study. Other methods have been recently developed to take into account the energetics of the interface; for example, quantum molecular dynamics would allow the determination of the most stable geometry together with its electronic structure. The use of these methods for complicated systems such as Schottky barriers is a challenge for the next few years.

5. CONCLUSION

More than one century after the first report of the rectifying properties of the metal–semiconductor interface and despite the huge amount of experimental and theoretical work, the phenomena involved in the Schottky barrier are not yet completely understood. Correlations between the barrier height and some bulk properties of its constituents give interesting information; for example, the relation to the heat of formation indicates that interface bonds play a role. Macroscopic models highlight the role of the states in the semiconductor band gap at the interface. Whether these states are due to defects or are induced by the metal is not yet finally established, but MIGS seems to be predominant. Variations of the Schottky-barrier height with respect to the MIGS prediction can come from defects or strains, but the nature of the bonds and the interface geometry can play a role too. Electronic-structure calculations are needed to go beyond the models, up to the microscopic aspect of the interface.

The large spread in experimental results, which can even be contradictory, does not give a general view of the problem. This comes from the sensitivity of the barrier to the conditions of the experiment: temperature, reactivity, mode of growth. Measurement-related problems add further difficulties. Nonetheless, general trends can be drawn for most metal–semiconductor interfaces: The barrier height does not depend on the doping for thick interfaces. It is not very sensitive to the nature of the metal, and it reaches its thick-limit value for metal coverages as small as a few monolayers. The development of new techniques like BEEM gives access to new types of information which may explain many discrepancies.

More experimental and theoretical studies are thus needed to improve the understanding of the mechanism of the Schottky barrier.

ACKNOWLEDGMENT

The authors wish to thank P. Muret, O. Symko, and J. Y. Veuillen for very fruitful discussions.

GLOSSARY

Accumulation Zone: Region of the semiconductor where the top of the valence band (bottom of the conduction band) is closer to or above (below) the Fermi level.

Charge-Neutrality Level (CNL): In the semiconductor, surface states are filled up to this CNL. It corresponds to a local neutrality of the surface.

Depletion Zone, Space-Charge Region: Region in the semiconductor where ionized donors or acceptors (depending on the doping) are not compensated. No carriers subsist in this region.

Electronic Affinity: Energy difference between the bottom of the conduction band of a semiconductor and the vacuum level.

MIGS: "Metal-induced gap states," model that assumes Fermi-level pinning by metal-induced states in the semiconductor gap. It is applied to Schottky barriers and heterojunctions and is also called IDIS, "induced density of interface states."

Ohmic Contact: Metal–semiconductor contact that lets the current flow in both directions.

Schottky-Barrier Height Φ_b: Barrier electrons (holes) have to jump at a metal–semiconductor junction. For an *n*-type semiconductor, $\Phi_{bn} = E_c - E_F$ and for a *p*-type $\Phi_{bp} = E_F - E_v$.

Schottky Lowering: Barrier-height lowering due to image force effect.

UDM: "Unified defect model," model that assumes the Fermi-level pinning by native defects in the semiconductor.

Vacuum Level: Energy level of a resting electron.

Work Function: Energy needed to extract an electron from the Fermi level and take it to infinity (to the vacuum level).

Works Cited

Andrews, J. M., Phillips, J. C. (1975), *Phys. Rev. Lett.* **35,** 56–59.

Bardeen, J. (1947), *Phys. Rev.* **71,** 717–727.

Braun, F. (1874), *Ann. Phys. (Leipzig)* **153,** 556.

Brillson, L. J. (1978), *Phys. Rev. Lett.* **40,** 260–263.

Cowley, A. M., Sze, S. M. (1965), *J. Appl. Phys.* **36,** 3212–3220.

Delerue, C., Lannoo, M., Langer, J. M. (1988), *Phys. Rev. Lett.* **61,** 199–202.

Duboz, J. Y., Badoz, P. A., Arnaud d'Avitaya, F., Rosencher, E. (1989), *Phys. Rev. B* **40,** 10607–10610.

Freeouf, J. L. (1980), *Solid State Commun.* **33,** 1059–1061.

Freeouf, J. L., Woodall, J. M. (1981), *Appl. Phys. Lett.* **39,** 727–729.

Heine, V. (1965), *Phys. Rev.* **138,** A1689–A1696.

Hollinger, G. (1987), in: *Semiconductor Interfaces: Formation and Properties,* Springer Proceedings in Physics 22, New York: Springer-Verlag, pp. 210–231.

Kaiser, W. J., Bell, L. D. (1988), *Phys. Rev. Lett.* **60,** 1406–1409.

Bell, L. D., Kaiser, W. J. (1988), *Phys. Rev. Lett.* **61,** 2368–2371.

Lannoo, M., Friedel, P. (1991), in: *Atomic and Electronic Structure of Surfaces,* Springer Series in Surface Science 16, New York: Springer-Verlag, Chap. 7.

Louie, S. G., Chelikowsky, J. R., Cohen, M. L. (1977), *Phys. Rev.* **15,** 2154–2162.

Mönch, W. (1990), *Rep. Prog. Phys.* **53,** 221–278.

Mott, N. F. (1938), *Proc. Cambridge Philos. Soc.* **34,** 568–572.

Muret, P., Elguennouni, D., Missous, M., Rhoderick, E. H., Baptist, R., Pellissier, A. (1992), *Appl. Surf. Sci.* **56–57,** 341–347.

Ortega, J., Sanchez-Dehesa, J., Flores, F. (1988), *Phys. Rev. B* **37,** 8516–8518.

Rhoderick, E. H., Williams, R. H. (1988), in: *Metal Semiconductor Contacts,* Oxford: Clarendon, Chap. 5.

Schottky, W. (1938), *Naturwissenschaften* **26,** 843.

Schottky, W. (1939), *Z. Phys.* **113,** 367–414.

Schottky, W. (1940), *Phys. Z.* **41,** 570–573.

Schottky, W. (1942), *Z. Phys.* **118,** 539–592.

Spicer, W. E., Chye, P. W., Skeath, P. R., Su, C. Y., Lindau, I. (1979), *J. Vac. Sci. Technol.* **16,** 1427–1433.

Spicer, W. E., Kendelewicz, T., Newman, N., Cao, R., McCants, C., Miyano, K., Lindau, I. (1988), *Appl. Surf. Sci.* **33/34,** 1009–1029.

Tejedor, C., Flores, F., Louis, E. (1977), *J. Phys. C.* **10,** 2163–2177.

Tersoff, J. (1984), *Phys. Rev. Lett.* **52,** 465–468.

Tersoff, J., Harrison, W. (1987), *Phys. Rev. Lett.* **58,** 2367–2370.

Tung, R. T, Ng, K. K., Gibson, J. M., Levi, A. F. J. (1986), *Phys. Rev. B* **33,** 7077–7090.

Yndurain, Y. (1971), *J. Phys. C.* **4,** 2849–2858.

Zur, A., McGill, T. C., Smith, D. L. (1983), *Phys. Rev. B* **28,** 2060–2067.

Further Reading

Electrical Behavior, Schottky Lowering, Experimental Methods

Rhoderick, E. H., Williams, R. H. (1988), *Metal Semiconductor Contacts,* Oxford: Clarendon, Chap. 5.

Sze, S. M. (1981), *Physics of Semiconductor Devices,* New York: Wiley-Interscience.

Correlations between Barrier Height and Semiconductor Dangling-bond Level

Lannoo, M., Friedel, P. (1991), *Atomic and Electronic Structure of Surfaces,* Springer Series in Surface Science No. 16, New York: Springer-Verlag, Chap. 7.

Phenomenological Models and Correlations

Mönch, W. (1990), *Rep. Prog. Phys.* **53,** 221–278.

Comparison MIGS/Defects

Flores, F., Ortega, J. (1992), *Appl. Surf. Sci.* **56–58,** 301–310.

Spicer, W. E., Cao, R., Miyano, K., Kendelewicz, T., Lindau, I., Weber, E., Liliental-Weber, Z., Newman, N. (1988), *Appl. Surf. Sci.* **41/42,** 1–16.

SCATTERING OF LIGHT

See RAMAN SCATTERING

SCIENTIFIC COMPUTING BY NUMERICAL METHODS

CHRISTINA C. CHRISTARA AND KENNETH R. JACKSON, *Computer Science Department, University of Toronto, Toronto, Ontario, Canada*

Introduction 3
1. **Floating-Point Arithmetic** 3
1.1 The IEEE Standard 3
1.2 Rounding Errors 4
1.3 The Effects of Inexact Arithmetic: Some Illustrative Examples 5
2. **The Direct Solution of Linear Algebraic Systems** 7
2.1 Gaussian Elimination 8
2.2 Back Substitution 9
2.3 The *LU* Factorization 9
2.4 Forward Elimination 10
2.5 Scaling and Pivoting 10
2.6 The Cholesky Factorization 12
2.7 Banded and Sparse Matrices .. 14
2.8 Rounding Errors, Condition Numbers, and Error Bounds .. 16
2.9 Iterative Improvement 18
3. **The Iterative Solution of Linear Algebraic Systems** 18
3.1 Basic Iterative Methods 18
3.2 The Conjugate-Gradient Method 22
4. **Overdetermined and Underdetermined Linear Systems** 24
4.1 The Normal Equations for Overdetermined Linear Systems 24
4.2 The Normal Equations for Underdetermined Linear Systems 25
4.3 Householder Transformations and the *QR* Factorization 25
4.4 Using the *QR* Factorization to Solve Overdetermined Linear Systems 26
4.5 Using the *QR* Factorization to Solve Underdetermined Linear Systems 26
4.6 The Gram–Schmidt Orthogonalization Algorithm .. 26
4.7 Using Gram–Schmidt to Solve Overdetermined Linear Systems 27
4.8 Using Gram–Schmidt to Solve Underdetermined Linear Systems 27
5. **Eigenvalues and Eigenvectors of Matrices** 27
5.1 The Power Method 28
5.2 The *QR* Method 28
5.3 Transforming a Symmetric Matrix to Tridiagonal Form ... 29
5.4 Inverse Iteration 29
5.5 Other Methods 30
6. **Nonlinear Algebraic Equations and Systems** 30
6.1 Fixed-Point Iteration 31
6.2 Newton's Method for Nonlinear Equations 31
6.3 The Secant Method 31
6.4 The Bisection and *Regula Falsi* Methods 31
6.5 Convergence 32
6.6 Rate of Convergence 32
6.7 Newton's Method for Systems of Nonlinear Equations 33
6.8 Modifications and Alternatives to Newton's Method 33
6.9 Polynomial Equations 33
6.10 Horner's Rule 34
7. **Unconstrained Optimization** 34
7.1 Some Definitions and Properties 34
7.2 The Fibonacci and Golden-Section Search Methods 35
7.3 The Steepest-Descent Method 36
7.4 Conjugate-Direction Methods 36
7.5 The Conjugate-Gradient Method 37
7.6 Newton's Method 37
7.7 Quasi-Newton Methods 37

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

8. **Approximation** 37
8.1 Polynomial Approximation 38
8.2 Polynomial Interpolation 38
8.2.1 Monomial Basis 38
8.2.2 Lagrange Basis 38
8.2.3 Newton Basis and Divided Differences 38
8.3 Polynomial Interpolation with Derivative Data 39
8.4 The Error in Polynomial Interpolation 39
8.5 Piecewise Polynomials and Splines 40
8.5.1 Constant Splines 40
8.5.2 Linear Splines 40
8.5.3 Quadratic Splines 41
8.5.4 Quadratic Piecewise Polynomials 41
8.5.5 Cubic Splines 41
8.5.6 Cubic Hermite Piecewise Polynomials 41
8.6 Piecewise Polynomial Interpolation 42
8.6.1 Linear Spline Interpolation 42
8.6.2 Quadratic Spline Interpolation 42
8.6.3 Cubic Spline Interpolation 42
8.6.4 Cubic Hermite Piecewise Polynomial Interpolation 43
8.7 Least-Squares Approximation 43
8.7.1 Orthogonal Polynomials 44
8.7.2 The Gram–Schmidt Orthogonalization Algorithm .. 44
8.7.3 Constructing the Least-Squares Polynomial Approximation 44
9. **Numerical Integration—Quadrature** 44
9.1 Simple Quadrature Rules 45
9.1.1 Some Definitions 45
9.1.2 Gaussian Quadrature Rules 46
9.1.3 Translating the Interval of Integration 46
9.1.4 Comparison of Gaussian and Newton–Cotes Quadrature Rules 47
9.2 Composite (Compound) Quadrature Rules 47
9.3 Adaptive Quadrature 47
9.4 Romberg Integration and Error Estimation 48
9.5 Infinite Integrals and Singularities 49
9.6 Monte-Carlo Methods 50
10. **Ordinary Differential Equations** 50
10.1 Initial-Value Problems (IVPs) 50
10.1.1 Two Simple Formulas 50
10.1.2 Stiff IVPs 51
10.1.3 Solving Implicit Equations 52
10.1.4 Higher-Order Formulas 53
10.1.5 Runge–Kutta Formulas 54
10.1.6 Linear Multistep Formulas 55
10.1.7 Adams Formulas 55
10.1.8 Backward Differentiation Formulas 56
10.1.9 Other Methods 56
10.1.10 Adaptive Methods 56
10.2 Boundary-Value Problems (BVPs) 57
10.2.1 Shooting Methods 57
10.2.2 One-Step Methods 58
10.2.3 Other Methods 58
11. **Partial Differential Equations (PDEs)** 59
11.1 Classes of Problems and PDEs 59
11.1.1 Some Definitions 59
11.1.2 Boundary Conditions 60
11.2 Classes of Numerical Methods for PDEs 60
11.2.1 Analysis of Numerical Methods for PDEs 61
11.2.1.1 Convergence Analysis (for BVPs and IVPs) 61
11.2.1.2 Stability Analysis (for IVPs) ... 61
11.2.1.3 Time (Computational) Complexity Analysis 61
11.2.1.4 Memory Complexity Analysis 62
11.2.1.5 Overall Efficiency Analysis 62
11.3 Finite-Difference Methods for BVPs 62
11.3.1 An Example of a Finite-Difference Method in One Dimension 63
11.3.2 An Example of a Finite-Difference Method in Two Dimensions 63
11.4 Finite-Element Methods for BVPs 65
11.4.1 The Galerkin Method 65
11.4.2 The Collocation Method 67
11.5 Finite-Difference Methods for IVPs 68
11.5.1 An Example of an Explicit One-Step Method for a Parabolic IVP 69

11.5.2 An Example of an Implicit One-Step Method for a Parabolic IVP 69
11.5.3 An Example of an Explicit Two-Step Method for a Hyperbolic IVP 70
11.6 The Method of Lines 71
11.7 Boundary-Element Methods ... 71
11.8 The Multigrid Method 71
12. Parallel Computation 72
12.1 Cyclic Reduction 73
13. Sources of Numerical Software 74
Glossary 75
Mathematical Symbols Used 76
Abbreviations Used 77
Works Cited 77
Further Reading 78

INTRODUCTION

Numerical methods are an indispensable tool in solving many problems that arise in science and engineering. In this article, we briefly survey a few of the most common mathematical problems and review some numerical methods to solve them.

As can be seen from the table of contents, the topics covered in this survey are those that appear in most introductory numerical methods books. However, our discussion of each topic is more brief than is normally the case in such texts and we frequently provide references to more advanced topics that are not usually considered in introductory books.

Definitions of the more specialized mathematical terms used in this survey can be found in the Glossary at the end of the article. We also list some mathematical symbols and abbreviations used throughout the survey in the two sections following the Glossary.

1. FLOATING-POINT ARITHMETIC

In this section, we consider the representation of floating-point numbers, floating-point arithmetic, rounding errors, and the effects of inexact arithmetic in some simple examples. For a more detailed discussion of these topics, see Wilkinson (1965) and Goldberg (1991) or an introductory numerical methods text.

1.1 The IEEE Standard

The approval of the IEEE (Institute of Electrical and Electronics Engineers) Standard for Binary Floating-Point Arithmetic (IEEE, 1985) was a significant advance for scientific computation. Not only has this led to cleaner floating-point arithmetic than was commonly available previously, thus greatly facilitating the development of reliable, robust numerical software, but, because many computer manufacturers have since adopted the standard, it has significantly increased the portability of programs.

The IEEE standard specifies both single- and double-precision floating-point numbers, each of the form

$$(-1)^s \times b_0.b_1b_2 \cdots b_{p-1} \times 2^E, \qquad (1)$$

where $s = 0$ or 1, $(-1)^s$ is the *sign* of the number, $b_i = 0$ or 1 for $i = 0, \ldots, p-1$; $b_0.b_1b_2 \cdots b_{p-1}$ is the *significand* (sometimes called the *mantissa*) of the number, and the *exponent* E is an integer satisfying $E_{\min} \le E \le E_{\max}$. In single-precision, $p = 24$, $E_{\min} = -126$ and $E_{\max} = +127$, whereas, in double-precision, $p = 53$, $E_{\min} = -1022$, and $E_{\max} = +1023$. We emphasize that a number written in the form (1) is binary. So, for example, $1.100 \cdots 0 \times 2^0$ written in the format (1) is equal to the decimal number 1.5.

A *normalized* number is either 0 or a floating-point number of the form (1) with $b_0 = 1$ (and so it is not necessary to store the leading bit). In single-precision, this provides the equivalent of 7 to 8 significant decimal digits with positive and negative numbers having magnitudes roughly in the range $[1.2 \times 10^{-38}, 3.4 \times 10^{+38}]$. In double-precision, this is increased to about 16 significant decimal digits and a range of roughly $[2.2 \times 10^{-308}, 1.8 \times 10^{+308}]$.

An *underflow* occurs when an operation produces a nonzero result in the range $(-2^{E_{\min}}, +2^{E_{\min}})$. In the IEEE standard, the default is to raise an underflow exception flag and to continue the computation with the result correctly rounded to the nearest *denormalized* number or zero. A denormalized floating-point number has the form (1) with $E = E_{\min}$ and $b_0 = 0$. Because denormalized numbers use some of the leading digits from the significand to represent the magnitude of the number, there are fewer digits available to represent its significant digits. Using denormalized numbers in this way for underflows is sometimes referred to as *gradual underflow*. Many older non-IEEE machines do not have denormalized numbers and, when an underflow occurs, they either abort the computation or replace the result by zero.

An *overflow* occurs when an operation produces a nonzero result outside the range of floating-point numbers of the form (1). In the IEEE standard, the default is to raise an overflow exception flag and to continue the computation with the result replaced by either $+\infty$ or $-\infty$, depending on the sign of the overflow value. Many older non-IEEE machines do not have $+\infty$ or $-\infty$ and, when an overflow occurs, they usually abort the computation.

The IEEE standard also includes at least two NaNs (Not-a-Number) in both precisions, representing indeterminate values that may arise from invalid or inexact operations such as $(+\infty) + (-\infty)$, $0 \times \infty$, $0/0$, ∞/∞, or $\sqrt{x}$ for $x < 0$. When a NaN arises in this way, the default is to raise an exception flag and to continue the computation. This novel feature is not available on most older non-IEEE machines.

It follows immediately from the format (1) that floating-point numbers are discrete and finite, whereas real numbers are dense and infinite. As a result, an arithmetic operation performed on two floating-point numbers may return a result that cannot be represented exactly in the form (1) in the same precision as the operands.

A key feature of the IEEE standard is that it requires that the basic arithmetic operations $+$, $-$, $\times$, $/$, and $\sqrt{\ }$ return *properly rounded* results. That is, we may think of the operation as first being done exactly and then properly rounded to the precision of the result. An operation with $\pm\infty$ is interpreted as the limiting case of the operation with an arbitrary large value in place of the ∞, when such an interpretation makes sense; otherwise the result is a NaN. An operation involving one or more NaNs returns a NaN.

The default rounding mode is *round-to-nearest;* that is, the exact result of the arithmetic operation is rounded to the nearest floating-point number, where, in the case of a tie, the floating-point number with the least significant bit equal to 0 is selected. The standard also provides for directed roundings (round-towards- $+\infty$, round-towards- $-\infty$, and round-towards-0), but these are not easily accessed from most programming languages.

Another important feature of the IEEE standard is that comparisons are exact and never overflow or underflow. The comparisons $<$, $>$, and $=$ work as expected with finite floating-point numbers of the form (1) and $-\infty < x < +\infty$ for any finite floating-point number x. NaNs are unordered and the comparison of a NaN with any other value—including itself—returns false.

The IEEE standard also provides for *extended* single- and double-precision floating-point numbers, but these are not easily accessed from most programming languages, and so we do not discuss them here.

As noted above, the IEEE standard has been widely adopted in the computer industry, but there are several important classes of machines that do not conform to it, including Crays, DEC Vaxes, and IBM mainframes. Although their floating-point numbers are similar to those described above, there are important differences. Space limitations, though, do not permit us to explore these systems here.

1.2 Rounding Errors

Since IEEE standard floating-point arithmetic returns the correctly rounded result for the basic operations $+$, $-$, $\times$, $/$, and $\sqrt{\ }$, one might expect that rounding errors would never pose a problem, particularly in double-precision computations. Although rounding errors can be ignored in many cases, the examples in the next subsection show that they may be significant even in simple calculations. Before considering these examples, though, we need to define an important machine constant and describe its significance in floating-point computation.

Machine epsilon, often abbreviated *mach-eps,* is the distance from 1 to the next larger floating-point number. For numbers of the form (1), mach-eps $= 2^{1-p}$. So for IEEE single- and double-precision numbers, mach-eps is $2^{-23} \approx 1.19209 \times 10^{-7}$ and $2^{-52} \approx 2.22045 \times 10^{-16}$, respectively.

A common alternative definition of machine epsilon is that it is the smallest positive floating-point number ϵ such that $1 + \epsilon > 1$ in floating-point arithmetic. We prefer the definition given in the previous paragraph, and use it throughout this section, because it is independent of the rounding mode and so is characteristic of the floating-point number system itself, while the alternative definition given in this paragraph depends on the rounding mode as well. We also assume throughout this section that round-to-nearest is in effect. The discussion below, though, can be modified easily for the alternative definition of mach-eps and other rounding modes.

It follows immediately from the floating-point format (1) that the absolute distance between floating-point numbers is not uniform. Rather, from (1) and the definition of mach-eps above, we see that the spacing between floating-point numbers in the intervals $[2^k, 2^{k+1})$ and $(-2^{k+1}, -2^k]$ is mach-eps $\times\ 2^k$ for $E_{\min} \le k \le E_{\max}$. Thus, the absolute distance between neighboring nonzero normalized floating-point numbers with the same exponent is uniform, but the floating-point numbers near $2^{E_{\min}}$ are much closer together in an absolute sense than those near $2^{E_{\max}}$. However, the *relative* spacing between all nonzero normalized floating-point numbers does not vary significantly. It is easy to see that, if x_1 and x_2 are any two neighboring nonzero normalized floating-point numbers, then

$$\frac{\text{mach-eps}}{2} \le \left|\frac{x_1 - x_2}{x_1}\right| \le \text{mach-eps}. \qquad (2)$$

As a result, it is more natural to consider relative, rather than absolute, errors in arithmetic operations on floating-point numbers, as is explained in more detail below.

For $x \in \mathbb{R}$, let $fl(x)$ be the floating-point number nearest to x, where, in the case of a tie, the floating-point number with the least significant bit equal to 0 is selected. The importance of mach-eps stems largely from the observation that, if $fl(x)$ does not overflow or underflow, then

$$fl(x) = x(1 + \delta) \quad \text{for some } |\delta| \le u, \qquad (3)$$

where u, the *relative roundoff error bound,* satisfies u = mach-eps/2 for round-to-nearest and u = mach-eps for the other IEEE rounding modes. Rewriting (3) as

$$\delta = [fl(x) - x]/x,$$

we see that δ is the relative error incurred in approximating x by $fl(x)$, and so (3) is closely related to (2).

If op is one of $+$, $-$, $\times$, or / and x and y are two floating-point numbers, let $fl(x \text{ op } y)$ stand for the result of performing the arithmetic operation x op y in floating-point arithmetic. If no arithmetic exception arises in the floating-point operation, then

$$fl(x \operatorname{op} y) = (x \operatorname{op} y)(1 + \delta) \quad \text{for some } |\delta| \le u, \qquad (4)$$

where the $(x \text{ op } y)$ on the right side of (4) is the exact result of the arithmetic operation. Similarly, if x is a nonnegative normalized floating-point number, then

$$fl(\sqrt{x}) = \sqrt{x}(1 + \delta) \quad \text{for some } |\delta| \le u. \qquad (5)$$

Again, the u in either (4) or (5) is the relative roundoff error bound and the δ is the relative error incurred in approximating x op y by $fl(x \text{ op } y)$ or $\sqrt{x}$ by $fl(\sqrt{x})$, respectively.

Although the relations (4)–(5) are not quite as tight as the requirement that the basic operations $+$, $-$, $\times$, /, and $\sqrt{\ }$ return the correctly rounded result, they are very useful in deriving error bounds and explaining the effects of rounding errors in computations.

1.3 The Effects of Inexact Arithmetic: Some Illustrative Examples

As noted at the start of the last subsection, since IEEE standard floating-point arithmetic returns the correctly rounded result for the basic operations $+$, $-$, $\times$, /, and $\sqrt{\ }$, one might expect that rounding errors would never pose a problem, particularly in double-precision computations. This, though, is not the case. In this subsection we consider a simple example that illustrates some of the *pitfalls* of numerical computation.

Suppose that we compute the expression

$$1 + 10^{10} - 10^{10} \qquad (6)$$

in single precision from left to right. We first compute $fl(1 + 10^{10}) = 10^{10}$, the correctly rounded single-precision result. Then we use this value to compute $fl(10^{10} - 10^{10}) = 0$, without committing an additional rounding error. Thus, $fl((1 + 10^{10}) - 10^{10}) = 0$, whereas the true result is 1.

The key point to note here is that $fl(1 + 10^{10}) = 10^{10} = (1 + 10^{10})(1 + \delta)$, where $|\delta| = 1/(1 + 10^{10}) < 10^{-10} < u = 2^{-24}$. So the rounding error that we commit in computing $1 + 10^{10}$ is small relative to $1 + 10^{10}$, the true result of the first addition, but the absolute error of 1 associated with this addition is not small compared with the true final answer, 1, thus illustrating the **Rule for Sums:**

> Although a rounding error is always small relative to the result that gives rise to it, it might be large relative to the true final answer if intermediate terms in a sum are large relative to the true final answer.

The Rule for Sums is important to remember when computing more complex expressions such as the truncated Taylor series $T_k(x) = 1 + x + x^2/2 + \cdots + x^k/k! \approx e^x$, where k is chosen large enough so that the *truncation error*

$$e^x - T_k(x) = \sum_{i=k+1}^{\infty} x^i/i!$$

is insignificant relative to e^x. It is easy to prove that, if $x \geq 0$ and k is large enough, then $T_k(x)$ is a good approximation to e^x. However, if $x < 0$ and of moderate magnitude, then the rounding error associated with some of the intermediate terms $x^i/i!$ in $T_k(x)$ might be much larger in magnitude than either the true value of $T_k(x)$ or e^x. As a result, $fl(T_k(x))$, the computed value of $T_k(x)$, might be completely erroneous, no matter how large we choose k. For example, $e^{-15} \approx 3.05\,902 \times 10^{-7}$, while we computed $fl(T_k(x)) \approx 2.12\,335 \times 10^{-2}$ in IEEE single-precision arithmetic on a Sun Sparcstation.

A similar problem is less likely to occur with multiplications, provided no overflow or underflow occurs, since from (4)

$$fl(x_1 \cdot x_2 \cdots x_n) = x_1 \cdot x_2 \cdots x_n(1 + \delta_1) \cdots (1 + \delta_{n-1}), \tag{7}$$

where $|\delta_i| \leq u$ for $i = 1, \ldots, n - 1$. Moreover, if $nu \leq 0.1$, then $(1 + \delta_1) \cdots (1 + \delta_{n-1}) = 1 + 1.1n\delta$ for some $\delta \in [-u,u]$. Therefore, unless n is very large, $fl(x_1 \cdot x_2 \cdots x_n)$ is guaranteed to be a good approximation to $x_1 \cdot x_2 \cdots x_n$. However, it is not hard to find examples for which $nu \gg 1$ and $fl(x_1 \cdot x_2 \cdots x_n)$ is a poor approximation to $x_1 \cdot x_2 \cdots x_n$.

The example (6) also illustrates another important phenomenon commonly called *catastrophic cancellation:* all the digits in the second sum, $fl(10^{10} - 10^{10})$, cancel, signaling a catastrophic loss of precision. Catastrophic cancellation refers also to the case that many, but not all, of the digits cancel. This is often a sign that a disastrous loss of accuracy has occurred, but, as in this example when we compute $fl(1 + 10^{10}) = 10^{10}$ and lose the 1, it is often the case that the accuracy is lost before the catastrophic cancellation occurs.

For an example of catastrophic cancellation in a more realistic computation, consider calculating the roots of the quadratic $ax^2 + bx + c$ by the standard formula

$$r_{\pm} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{8}$$

We used this formula to compute the roots of the quadratic $x^2 - 10^4x + 1$ in IEEE single-precision arithmetic on a Sun Sparcstation. The computed roots were 10^4 and 0, the larger of which is accurate, having a relative error of about 10^{-8}, but the smaller one is completely wrong, the true root being about 10^{-4}. A similar result usually occurs whenever $|ac|/b^2 \ll 1$. The root of larger magnitude is usually computed precisely, but the smaller one is frequently very inaccurate as a result of catastrophic cancellation, since $b^2 - 4ac \approx b^2$ and so $|b| - \sqrt{b^2 - 4ac} \approx 0$. Although the second of these relations signals catastrophic cancellation, the loss of precision occurs in the first.

There is an easy remedy for the loss of precision due to catastrophic cancellation in this case. Use (8) to compute r_1, the root of larger magnitude, and then use the alternative formula $r_2 = c/ar_1$ to compute the smaller one. The relative error in r_2 is at most $(1 + u)^2$ times larger than the relative error in r_1, provided that no overflows or underflows occur in computing c/ar_1.

Another point to note is that, if we compute (6) from right to left, instead of left to right, then

$$fl(1 + (10^{10} - 10^{10})) = fl(1 + 0) = 1.$$

It is not particularly significant that this computation gives the correct answer, but what is important is that it illustrates that floating-point addition is not associative, although it is commutative, since $fl(a + b)$ and $fl(b + a)$ are both required to be the correctly rounded value for $a + b = b + a$. Similar results hold for multiplication and division.

Many other fundamental mathematical relations that we take for granted do not hold for floating-point computations. For example, the result that $\sin(x)$ is strictly increasing for $x \in (0,\pi/2)$ cannot hold in any floating-point system in which x and $\sin(x)$ are in the same precision, since there are more floating-point numbers in the domain $(0,\pi/2)$ than there are in $(0,1)$, the range of $\sin(x)$.

Finally, we end this section by noting that overflows and underflows often cause problems in computations. After an overflow, $\pm\infty$ or NaN frequently propagates through the computation. Although this can sometimes yield a useful result, it is more often a signal of an error in the program or its input. On the other hand, continuing the computation with denormalized numbers or zero in place of an underflow can often yield a useful numerical result. However, there are cases when this can be disastrous. For example, if x^2 underflows, but y^2 is not too close to the underflow limit, then $fl(\sqrt{x^2 + y^2})$, the computed value of $\sqrt{x^2 + y^2}$, is still accurate. However, if both x^2 and y^2 underflow to 0, then $fl(\sqrt{x^2 + y^2}) = 0$, although $\sqrt{x^2 + y^2} \geq \max(|x|,|y|)$ may be far from the underflow limit.

It is often possible to ensure that overflows do not occur and that underflows are harmless. For the example considered above, note that $\sqrt{x^2 + y^2} = s\sqrt{(x/s)^2 + (y/s)^2}$ for any scaling factor $s > 0$. If we choose $s = 2^k$ for an integer $k \approx \log_2[\max(|x|,|y|)]$, then neither $(x/s)^2$ nor $(y/s)^2$ can overflow, and any underflow that occurs is harmless, since one of $(x/s)^2$ or $(y/s)^2$ is close to 1. Moreover, in IEEE floating-point arithmetic, multiplying and dividing by $s = 2^k$ does not introduce any additional rounding error into the computation.

A similar problem with overflows and underflows occurs in formula (8) and in many other numerical computations. Overflows can be avoided and underflows can be rendered harmless in computing $\sqrt{b^2 - 4ac}$ in (8) by scaling in much the same way as described above for $\sqrt{x^2 + y^2}$.

2. THE DIRECT SOLUTION OF LINEAR ALGEBRAIC SYSTEMS

In this section, we consider the *direct* solution of linear algebraic systems of the form $Ax = b$, where $A \in \mathbb{R}^{n\times n}$ (or $\mathbb{C}^{n\times n}$) is a nonsingular matrix and x and $b \in \mathbb{R}^n$ (or $\mathbb{C}^n$). A numerical method for solving $Ax = b$ is direct if it computes the exact solution of the system when implemented in exact arithmetic. Iterative methods for nonsingular linear systems and methods for overdetermined and underdetermined systems are considered in Secs. 4 and 5, respectively.

The standard direct methods for solving $Ax = b$ are based on, or closely related to, *Gaussian elimination* (GE), the familiar variable elimination technique that reduces the original system $Ax = b$ to an *upper-triangular system,* $Ux = \tilde{b}$, which has the same solution x. We present a simple form of GE in Sec. 2.1 and show how $Ux = \tilde{b}$ can be solved easily by *back substitution* in Sec. 2.2. We then explain how the simple form of GE presented in Sec. 2.1 for $Ax = b$ relates to the *LU* factorization of the coefficient matrix A in Sec. 2.3 and to *forward elimination* in Sec. 2.4. Enhancements to this simple form of GE to make it an efficient, robust, reliable numerical method for the solution of linear systems are outlined in Sec. 2.5. The closely related Cholesky factorization for symmetric positive-definite matrices is presented in Sec. 2.6. We consider how to adapt these methods to banded and sparse linear systems in Sec. 2.7. We end with a discussion of the effects of rounding errors on the direct methods in Sec. 2.8 and of *iterative improvement,* a technique to ameliorate these effects, in Sec. 2.9. See Sec. 13 for a discussion of sources of high-quality numerical software for solving systems of linear algebraic equations.

GE can also be applied to singular systems of linear equations or overdetermined or underdetermined linear systems of m equations in n unknowns. However, it is not as robust as the methods discussed in Sec. 5 for these problems, and so we do not present these generalizations of GE here. In addition, we note that there are several mathematically equivalent, but computationally distinct, implementations of GE and the factorizations discussed here. The reader interested in a more comprehensive treatment of these topics should consult an advanced text, such as Golub and Van Loan (1989).

Finally we note that it is generally inadvisable to solve a system $Ax = b$ by first computing A^{-1} and then calculating $x = A^{-1}b$. The techniques discussed in this section are usually both more reliable and more cost effective than methods using A^{-1}. We also note that, although Cramer's rule is a useful theoretical tool, it is an extremely ineffective computational scheme.

2.1 Gaussian Elimination

First, to unify the notation used below, let $A_0 = A$ and $b_0 = b$.

GE for $Ax = b$ proceeds in $n - 1$ stages. For $k = 1, \ldots, n - 1$, we begin stage k of GE with the reduced system $A_{k-1}x = b_{k-1}$, where columns $1, \ldots, k - 1$ of A_{k-1} contain 0's below the main diagonal. That is, for $A_{k-1} = [a_{ij}^{(k-1)}]$, $a_{ij}^{(k-1)} = 0$ for $j = 1, \ldots, k - 1$ and $i = j + 1, \ldots, n$. This corresponds to the variables $x_1, \ldots, x_{i-1}$ having been eliminated from equation i of $A_{k-1}x = b_{k-1}$ for $i = 2, \ldots, k - 1$ and the variables $x_1, \ldots, x_{k-1}$ having been eliminated from the remaining equations $k, \ldots, n$. Moreover, x is the unique solution of both $A_{k-1}x = b_{k-1}$ and $Ax = b$. Note that all the assumptions above hold vacuously for $k = 1$, since, in this case, the "reduced" system $A_0x = b_0$ is just the original system $Ax = b$ from which no variables have yet been eliminated.

During stage k of GE, we further the reduction process by eliminating the variable x_k from row i of $A_{k-1}x = b_{k-1}$ for $i = k + 1, \ldots, n$ by multiplying row k of this system by $m_{ik} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ and subtracting it from row i.

Note that the multipliers m_{ik} are not properly defined by $m_{ik} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ if the *pivot element* $a_{kk}^{(k-1)} = 0$. In exact arithmetic, this can happen only if the $k \times k$ leading principal minor of A is singular. We consider how to deal with zero (or nearly zero) pivots in Sec. 2.5. For now, though, we say that this simple form of GE "breaks down" at stage k and we terminate the process.

After the last stage of GE, the original system $Ax = b$ has been reduced to $Ux = \tilde{b}$, where $\tilde{b} = b_{n-1}$ and $U = A_{n-1}$. Note that $U = [u_{ij}]$ is an upper-triangular matrix (i.e., $u_{ij} = 0$ for $1 \le j < i \le n$) with $u_{kk} = a_{kk}^{(k-1)}$ for $k = 1, \ldots, n$. Therefore, if A and all its leading principal minors are nonsingular, then $u_{kk} = a_{kk}^{(k-1)} \neq 0$ for $k = 1, \ldots, n$, so U is nonsingular too. Moreover, in this case, x is the unique solution of both $Ax = b$ and $Ux = \tilde{b}$. The latter system can be solved easily by back substitution, as described in Sec. 2.2.

The complete GE process can be written in pseudocode as shown in Table 1. Note that at stage k of GE, we know the elements $a_{ik}^{(k)} = a_{ik}^{(k-1)} - m_{ik} \cdot a_{kk}^{(k-1)} = 0$ for $i = k + 1, \ldots, n$, and so we do not need to perform this calculation explicitly. Consequently, instead of j running for $k, \ldots, n$ in Table 1, as might be expected, j runs from $k + 1, \ldots, n$ instead.

To reduce the storage needed for GE, the original matrix A and vector b are often overwritten by the intermediate matrices A_k and vectors b_k, so that at the end of the GE process, the upper-triangular part of A contains U and b contains $\tilde{b}$. The only change required to the algorithm in Table 1 to implement this reduced storage scheme is to remove the superscripts from the coefficients a_{ij} and b_i. As explained below, it is also important to store the multipliers m_{ik}. Fortunately, the $n - k$ multipliers $\{m_{ik} : i = k + 1, \ldots, n\}$ created at stage k of GE can be stored in the $n - k$ positions in column k of A below the main diagonal that are elimi-

Table 1. Gaussian elimination (GE) for the system $Ax = b$.

```
for $k = 1, \ldots, n - 1$ do
    for $i = k + 1, \ldots, n$ do
        $m_{ik} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$
        for $j = k + 1, \ldots, n$ do
            $a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik} \cdot a_{kj}^{(k-1)}$
        end
        $b_i^{(k)} = b_i^{(k-1)} - m_{ik} \cdot b_k^{(k-1)}$
    end
end
```

nated in stage k. Thus, in many implementations of GE, the upper-triangular part of A is overwritten by U and the strictly lower part of A is overwritten by the multipliers m_{ik} for $1 \leq k < i \leq n$.

A straightforward count of the operations in Table 1 shows that GE requires $n(n-1)/2 \approx n^2/2$ divisions to compute the multipliers m_{ik}, $n(2n-1)(n-1)/6 \approx n^3/3$ multiplications and subtractions to compute the coefficients of U, and $n(n-1)/2 \approx n^2/2$ multiplications and subtractions to compute the coefficients of $\tilde{b}$. Since multiplications and subtractions (or additions) occur in pairs so frequently in matrix calculations, we refer to this pair of operations as a *flop*, which is short for *floating-point operation*. Thus, the computational work required to reduce a system $Ax = b$ of n equations in n unknowns to $Ux = \tilde{b}$ is about $n^3/3$ flops. We show in Sec. 2.2 that $Ux = \tilde{b}$ can be solved by *back substitution* using n divisions and about $n^2/2$ flops.

2.2 Back Substitution

Let $U = [u_{ij}]$ be an $n \times n$ nonsingular upper-triangular matrix. That is, $u_{ij} = 0$ for $1 \leq j < i \leq n$ and $u_{ii} \neq 0$ for $1 \leq i \leq n$. Then the linear algebraic system $Ux = \tilde{b}$ can be solved easily by back substitution, as shown in Table 2.

It is easy to see from Table 2 that back substitution requires n divisions and $n(n-1)/2 \approx n^2/2$ multiplications and subtractions. So the computational work is about $n^2/2$ flops.

Table 2. Back substitution to solve $Ux = \tilde{b}$.

```
for $i = n, \ldots, 1$ do
    $x_i = \left(\tilde{b}_i - \sum_{j=i+1}^{n} u_{ij}x_j\right) \Big/ u_{ii}$
end
```

2.3 The *LU* Factorization

Applying Gaussian elimination (GE) as described in Sec. 2.1 to solve the linear algebraic system $Ax = b$ of n equations in n unknowns is closely related to computing the *LU* factorization of the matrix A, where $L_1 = [l_{ij}^{(1)}]$ is a unit lower-triangular matrix (i.e., $l_{ii}^{(1)} = 1$ for $i = 1, \ldots, n$ and $l_{ij}^{(1)} = 0$ for $1 \leq i < j \leq n$) and $U_1 = [u_{ij}^{(1)}]$ is an upper-triangular matrix (i.e., $u_{ij}^{(1)} = 0$ for $1 \leq j < i \leq n$) satisfying

$$A = L_1 U_1. \tag{9}$$

The factorization (9) exists and is unique if and only if all the leading principal minors of A are nonsingular. In this case, it can be shown that the matrix U_1 in (9) is the same as the upper-triangular matrix U produced by GE, and the elements in the strictly lower-triangular part of $L_1 = [l_{ij}^{(1)}]$ satisfy $l_{ij}^{(1)} = m_{ij}$ for $1 \leq j < i \leq n$, where the m_{ij} are the multipliers used in GE. Moreover, the U_1 in (9) is nonsingular if A is; L_1 is always nonsingular if the factorization exists.

From the discussion in Sec. 2.1, it follows that computing the *LU* factorization of an $n \times n$ matrix A in this way requires $n(n-1)/2 \approx n^2/2$ divisions and $n(2n-1)(n-1)/6 \approx n^3/3$ multiplications and subtractions. Thus, the computational work required to calculate it is about $n^3/3$ flops.

If we need to solve $m > 1$ systems $Ax_i = b_i$, $i = 1, \ldots, m$, (or $AX = B$, where X and $B \in \mathbb{R}^{n\times m}$ or $\mathbb{C}^{n\times m}$), we may obtain significant computational savings by computing the *LU* factorization of A once only and using the factors L_1 and U_1 to solve each system $Ax_i = b_i$ for $i = 1, \ldots, m$ by first solving $L_1\tilde{b}_i = b_i$ for $\tilde{b}_i$ by *forward elimination*, as described in Sec. 2.4, and then solving $U_1 x_i = \tilde{b}_i$ for x_i by back substitution, as outlined in Sec. 2.2. This procedure is essentially the same as performing GE to reduce A to U once only, saving the multipliers $\{m_{ik}\}$ used in the process, and then, for each system $Ax_i = b_i$, using the multipliers to perform the same transformation on b_i to produce $\tilde{b}_i$ and solving $Ux_i = \tilde{b}_i$ for x_i by back substitution. With either of these procedures, the computational work required to solve all m systems $Ax_i = b_i$ is about $n^3/3 + mn^2$ flops, whereas, if we apply GE as outlined in Sec. 2.1 to each system $Ax_i = b_i$, recomputing U each time, the computational work required to solve all m systems $Ax_i = b_i$ is about $m(n^3/3 + n^2)$ flops, which is much greater if m and/or n is large.

Finally, note that we intentionally used the same symbol $\tilde{b}_i$ for the solution of $L\tilde{b}_i =$

b_i and the transformed right-side vector produced by GE, since these vectors are identical.

2.4 Forward Elimination

Let $L = [l_{ij}]$ be an $n \times n$ lower-triangular matrix (i.e., $l_{ij} = 0$ for $1 \leq i < j \leq n$). If L is nonsingular too, then $l_{ii} \neq 0$ for $1 \leq i \leq n$ and so the linear algebraic system $L\tilde{b} = b$ can be solved easily by forward elimination, as shown in Table 3.

It is easy to see from Table 3 that forward elimination requires n divisions and $n(n - 1)/2 \approx n^2/2$ multiplications and subtractions. So the computational work is about $n^2/2$ flops.

If $L = [l_{ij}]$ is unit lower-triangular (i.e., $l_{ii} = 1$ for $i = 1, \ldots, n$ as well as L being lower-triangular), as is the case for the L produced by the LU factorization described in Sec. 2.3, then the division by l_{ii} in Table 3 is not required, reducing the operation count slightly to about $n^2/2$ flops. However, we have presented the forward-elimination procedure in Table 3 with general $l_{ii} \neq 0$, since other schemes, such as the Cholesky factorization described in Sec. 2.6, produce a lower-triangular matrix that is typically not unit lower-triangular.

The name "forward elimination" for this procedure comes from the observation that it is mathematically equivalent to the forward-elimination procedure used in GE to eliminate the variables $x_1, \ldots, x_{i-1}$ from equation i of the original system $Ax = b$ to produce the reduced system $Ux = \tilde{b}$.

2.5 Scaling and Pivoting

As noted in Sec. 2.1, the simple form of Gaussian elimination (GE) presented there may "break down" at stage k if the pivot $a_{kk}^{(k-1)} = 0$. Moreover, even if $a_{kk}^{(k-1)} \neq 0$, but $|a_{kk}^{(k-1)}| \ll |a_{ik}^{(k-1)}|$ for some $i \in \{k + 1, \ldots, n\}$, then $|m_{ik}| = |a_{ik}^{(k-1)}/a_{kk}^{(k-1)}| \gg 1$. So multiplying row k of A_{k-1} by m_{ik} and subtracting it from row i may produce large elements in the resulting row i of A_k, which in turn may produce still larger elements during later stages of the GE process. Since, as noted in Sec. 2.8, the bound on the rounding errors in the GE process is proportional to the largest element that occurs in A_k for $k = 0, \ldots, n - 1$, creating large elements during the GE reduction process may introduce excessive rounding error into the computation, resulting in an unstable numerical process and destroying the accuracy of the LU factorization and the computed solution x of the linear system $Ax = b$. We present in this section *scaling* and *pivoting* strategies that enhance GE to make it an efficient, robust, reliable numerical method for the solution of linear systems.

Table 3. Forward elimination too solve $L\tilde{b} = b$.

```
for $i = 1, \ldots, n$ do
    $\tilde{b}_i = \left(b_i - \sum_{j=1}^{i-1} l_{ij}\tilde{b}_j\right) \Big/ l_{ii}$
end
```

Scaling, often called *balancing* or *equilibration,* is the process by which the equations and unknowns of the system $Ax = b$ are scaled in an attempt to reduce the rounding errors incurred in solving the problem and improve its conditioning, as described in Sec. 2.8. The effects can be quite dramatic.

Typically, scaling is done by choosing two $n \times n$ diagonal matrices $D_1 = [d_{ij}^{(1)}]$ and $D_2 = [d_{ij}^{(2)}]$ (i.e., $d_{ij}^{(1)} = d_{ij}^{(2)} = 0$ for $i \neq j$) and forming the new system $\hat{A}\hat{x} = \hat{b}$, where $\hat{A} = D_1AD_2^{-1}$, $\hat{x} = D_2x$, and $\hat{b} = D_1b$. Thus, D_1 scales the rows and D_2 scales the unknowns of $Ax = b$, or, equivalently, D_1 scales the rows and D_2^{-1} scales the columns of A. Of course, the solution of $Ax = b$ can be recovered easily from the solution of $\hat{A}\hat{x} = \hat{b}$, since $x = D_2^{-1}\hat{x}$. Moreover, if the diagonal entries $d_{11}^{(1)}, \ldots, d_{nn}^{(1)}$ of D_1 and $d_{11}^{(2)}, \ldots, d_{nn}^{(2)}$ of D_2 are chosen to be powers of the base of the floating-point number system (i.e., powers of 2 for IEEE floating-point arithmetic), then scaling introduces no rounding errors into the computation.

One common technique is to scale the rows only by taking $D_2 = I$ and choosing D_1 so that largest element in each row of the scaled matrix $\hat{A} = D_1A$ is about the same size. A slightly more complicated procedure is to scale the rows and columns of A so that the largest element in each row and column of $\hat{A} = D_1AD_2^{-1}$ is about the same size.

These strategies, although usually helpful, are not foolproof: it is easy to find examples for which row scaling or row and column scaling as described above makes the numerical solution worse. The best strategy is to scale on a problem-by-problem basis depending on what the source problem says about the significance of each coefficient a_{ij} in $A = [a_{ij}]$. See an advanced text such as Golub and Van Loan (1989) for a more detailed discussion of scaling.

For the remainder of this section, we assume that scaling, if done at all, has already been performed.

The most commonly used pivoting strategy is *partial pivoting*. The only modification required to stage k of GE described in Sec. 2.1 to implement GE with partial pivoting is to first search column k of A_{k-1} for the largest element $a_{ik}^{(k-1)}$ on or below the main diagonal. That is, find $i \in \{k, \ldots, n\}$ such that $|a_{ik}^{(k-1)}| \ge |a_{\mu k}^{(k-1)}|$ for $\mu = k, \ldots, n$. Then interchange equations i and k in the reduced system $A_{k-1}x = b_{k-1}$ and proceed with stage k of GE as described in Sec. 2.1.

After the equation interchange described above, the pivot element $a_{kk}^{(k-1)}$ satisfies $|a_{kk}^{(k-1)}| \ge |a_{ik}^{(k-1)}|$ for $i = k, \ldots, n$. So, if $a_{kk}^{(k-1)} \neq 0$, then the multiplier $m_{ik} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ must satisfy $|m_{ik}| \le 1$ for $i = k + 1, \ldots, n$. Thus no large multipliers can occur in GE with partial pivoting.

On the other hand, if the pivot element $a_{kk}^{(k-1)} = 0$, then $a_{ik}^{(k-1)} = 0$ for $i = k, \ldots, n$, whence A_{k-1} is singular and so A must be too. Thus, GE with partial pivoting never "breaks down" (in exact arithmetic) if A is nonsingular.

Partial pivoting adds a little overhead only to the GE process. At stage k, we must perform $n - k$ comparisons to determine the row i with $|a_{ik}^{(k-1)}| \ge |a_{jk}^{(k-1)}|$ for $j = k, \ldots, n$. Thus, GE with partial pivoting requires a total of $n(n - 1)/2 \approx n^2/2$ comparisons. In addition, we must interchange rows i and k if $i > k$, or use some form of indirect addressing if the interchange is not performed explicitly. On the other hand, exactly the same number of arithmetic operations must be executed whether or not pivoting is performed. Therefore, if n is large, the added cost of pivoting is small compared with performing approximately $n^3/3$ flops to reduce A to upper-triangular form.

Complete pivoting is similar to partial pivoting except that the search for the pivot at stage k of GE is not restricted to column k of A_{k-1}. Instead, in GE with complete pivoting, we search the $(n - k) \times (n - k)$ lower right block of A_{k-1} for the largest element. That is, find i and $j \in \{k, \ldots, n\}$ such that $|a_{ij}^{(k-1)}| \ge |a_{\mu\nu}^{(k-1)}|$ for $\mu = k, \ldots, n$ and $\nu = k, \ldots, n$. Then interchange equations i and k and variables j and k in the reduced system $A_{k-1}x_{k-1} = b_{k-1}$ and proceed with stage k of GE as described in Sec. 2.1. Note that the vector x_{k-1} in the reduced system above is a reordered version of the vector of unknowns x in the original system $Ax = b$, incorporating the variable interchanges that have occurred in stages $1, \ldots, k - 1$ of GE with complete pivoting.

After the equation and variable interchanges described above, the pivot element $a_{kk}^{(k-1)}$ satisfies $|a_{kk}^{(k-1)}| \ge |a_{ik}^{(k-1)}|$ for $i = k, \ldots, n$. So, if $a_{kk}^{(k-1)} \neq 0$, the multiplier $m_{ik} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ must satisfy $|m_{ik}| \le 1$ for $i = k + 1, \ldots, n$, as is the case with partial pivoting. However, with complete pivoting, the multipliers tend to be even smaller than they are with partial pivoting, since the pivots tend to be larger, and so the numerical solution might suffer less loss of accuracy due to rounding errors, as discussed further in Sec. 2.8.

After the row and column interchanges in stage k, the pivot element $a_{kk}^{(k-1)} = 0$ only if $a_{ij}^{(k-1)} = 0$ for $i = k, \ldots, n$ and $j = k, \ldots, n$ in which case A_{k-1} is singular and so A must be too. Thus, like GE with partial pivoting, GE with complete pivoting never "breaks down" (in exact arithmetic) if A is nonsingular.

Moreover, if A is singular and $a_{ij}^{(k-1)} = 0$ for $i = k, \ldots, n$ and $j = k, \ldots, n$, then the GE process can be terminated at this stage and the factorization computed so far used to advantage in determining a solution (or approximate solution) to the singular system $Ax = b$. However, this is not as robust a technique as the methods discussed in Sec. 4 for overdetermined problems, and so we do not discuss this further here. The reader interested in this application of GE should consult an advanced text such as Golub and Van Loan (1989).

Complete pivoting, unlike partial pivoting, adds significantly to the cost of the GE process. At stage k, we must perform $(n - k + 1)^2 - 1$ comparisons to determine the row i and column j with $|a_{ij}^{(k-1)}| \ge |a_{\mu\nu}^{(k-1)}|$ for $\mu = k, \ldots, n$ and $\nu = k, \ldots, n$. Thus, GE with partial pivoting requires a total of $n(n - 1)(2n + 5)/6 \approx n^3/3$ comparisons. On the other hand, exactly the same number of arithmetic operations must be executed whether or not pivoting is performed. So the cost of determining the pivots is comparable to the cost of performing approximately $n^3/3$ flops required to reduce A to upper-triangular form. In addition, we must interchange rows i and k if $i > k$ and columns j and k if $j > k$ or use some form of indirect addressing if the interchange is not performed explicitly. Thus even though GE with complete pivoting has better roundoff-error properties than GE with partial pivoting, GE with partial pivoting is used more often in practice.

As is the case for the simple version of GE presented in Sec. 2.1, GE with partial or complete pivoting is closely related to computing the *LU* factorization of the matrix A. However, in this case, we must account for the row or row and column interchanges by extending (9) to

$$P_2A = L_2U_2 \tag{10}$$

for partial pivoting and

$$P_3AQ_3^{\mathrm{T}} = L_3U_3 \tag{11}$$

for complete pivoting, where P_2 and P_3 are permutation matrices that record the row interchanges performed in GE with partial and complete pivoting, respectively; Q_3^{T} is a permutation matrix that records the column interchanges performed in GE with complete pivoting; L_2 and L_3 are unit lower-triangular matrices with the $n - k$ multipliers from stage k of GE with partial and complete pivoting, respectively, in column k below the main diagonal, but permuted according to the row interchanges that occur in stages $k + 1, \ldots, n - 1$ of GE with partial and complete pivoting, respectively; and U_2 and U_3 are the upper-triangular matrices produced by GE with partial and complete pivoting, respectively.

A permutation matrix P has exactly one 1 in each row and column and all other elements equal to 0. It is easy to check that $PP^{\mathrm{T}} = I$ so that P is nonsingular and $P^{\mathrm{T}} = P^{-1}$. That is, P is an orthogonal matrix. Also note that we do not need a full $n \times n$ array to store P: The information required to form or multiply by $P = [P_{ij}]$ can be stored in an n-vector $p = [p_i]$, where $p_i = j$ if and only if $P_{ij} = 1$ and $P_{ik} = 0$ for $k \neq j$.

The factorizations (9), (10), and (11) are all called *LU factorizations;* (10) is also called a *PLU factorization.* Unlike (9), the *LU* factorizations (10) and (11) always exist. P_2, P_3, Q_3, L_2, and L_3 are always nonsingular; U_2 and U_3 are nonsingular if and only if A is nonsingular. Moreover, the factorizations (10) and (11) are unique if there is a well-defined choice for the pivot if more than one element of maximal size occurs in the search for the pivot and if there is a well-defined choice for the multipliers in L if the pivot is zero.

The *LU* factorization (10) can be used to solve the linear system $Ax = b$ by first computing $\hat{b}_2 = P_2b$, then solving $L_2\tilde{b}_2 = \hat{b}_2$ by forward elimination, and finally solving $U_2x = \tilde{b}_2$ by back substitution. The steps are similar if we use the *LU* factorization (11) instead of (10), except that the back substitution $U_3\hat{x}_3 = \tilde{b}_3$ yields the permuted vector of unknowns $\hat{x}_3$. The original vector of unknowns x can be recovered by $x = Q_3^{\mathrm{T}}\hat{x}_3$. We have used $\tilde{b}_2$ and $\tilde{b}_3$ for the intermediate results above to emphasize that this is the same as the vector $\tilde{b}$ that is obtained if we perform GE with partial and complete pivoting, respectively, on the original system $Ax = b$.

As noted earlier for (9), if we need to solve $m > 1$ systems $Ax_i = b_i$, $i = 1, \ldots, m$ (or $AX = B$, where $X, B \in \mathbb{R}^{n\times m}$ or $\mathbb{C}^{n\times m}$), we may obtain significant computational savings by computing the *LU* factorization of A once only. The same observation applies to the *LU* factorizations (10) and (11).

We end by noting that not having to pivot to ensure numerical stability can be a great advantage in some cases—for example, when factoring a banded or sparse matrix, as described in Sec. 2.7. Moreover, there are classes of matrices for which pivoting is not required to ensure numerical stability. Three such classes are complex Hermitian positive-definite matrices, real symmetric positive-definite matrices, and diagonally dominant matrices.

2.6 The Cholesky Factorization

In this subsection, we present the *Cholesky factorization* of a real symmetric positive-definite $n \times n$ matrix A. It is straightforward to modify the scheme for complex Hermitian positive-definite matrices.

Recall that $A \in \mathbb{R}^{n \times n}$ is symmetric if $A = A^{\mathrm{T}}$, where A^{T} is the transpose of A, and it is positive definite if $x^{\mathrm{T}}Ax > 0$ for all $x \in \mathbb{R}^n$, $x \neq 0$. The Cholesky factorization exploits these properties of A to compute a lower-triangular matrix L satisfying

$$A = LL^{\mathrm{T}}. \tag{12}$$

The similar *LDL factorization* computes a unit lower-triangular matrix $\tilde{L}$ and a diagonal matrix D satisfying

$$A = \tilde{L}D\tilde{L}^{\mathrm{T}}. \tag{13}$$

We present the *dot product* form of the Cholesky factorization in Table 4. It is derived by equating the terms of $A = [a_{ij}]$ to those of LL^{T} in the order (1,1), (2,1), . . . , $(n,1)$, (2,2), (3,2), . . . , $(n,2)$, . . . , (n,n) and using the lower-triangular structure of $L = [l_{ij}]$ (i.e., $l_{ij} = 0$ for $1 \leq i < j \leq n$). Other forms of the Cholesky factorization are discussed in advanced texts such as Golub and Van Loan (1989).

It can be shown that, if A is symmetric positive-definite, then

$$a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 > 0$$

(in exact arithmetic) each time this expression is computed in the Cholesky factorization. Therefore, we may take the associated square root to be positive, whence $l_{jj} > 0$ for $j = 1, \ldots, n$. With this convention, the Cholesky factorization is unique.

Moreover, it follows from

$$l_{jj} = \left(a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2\right)^{1/2}$$

that

$$a_{jj} = \sum_{k=1}^{j} l_{jk}^2.$$

So the elements in row j of the Cholesky factor L are bounded by $\sqrt{a_{jj}}$ even if we do not pivot. Consequently, as noted in Sec. 2.5, it is customary to compute the Cholesky factorization without pivoting.

Table 4. The Cholesky factorization of a symmetric positive-definite matrix A.

```
for $j = 1, \ldots, n$ do
    $l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}$
    for $i = j + 1, \ldots, n$ do
        $l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk}\right) \Big/ l_{jj}$
    end
end
```

Note that the method in Table 4 accesses the lower-triangular part of A only, and so only those elements need to be stored. Moreover, if we replace l_{jj} and l_{ij} by a_{jj} and a_{ij}, respectively, in Table 4, then the modified algorithm overwrites the lower-triangular part of A with the Cholesky factor L.

A straightforward count of the operations in Table 4 shows that the Cholesky factorization requires n square roots, $n(n - 1)/2 \approx n^2/2$ divisions, and $n(n - 1)(n + 1)/6 \approx n^3/6$ multiplications and subtractions. This is approximately half the arithmetic operations required to compute the *LU* factorization of A. Of course, the storage required for the Cholesky factorization is also about half that required for the *LU* factorization.

The Cholesky, *LDL*, and *LU* factorizations are closely related. Let $D_1 = [d_{ij}^{(1)}]$ be the diagonal matrix with the same diagonal elements as $L = [l_{ij}]$ (i.e., $d_{jj}^{(1)} = l_{jj}$ for $j = 1, \ldots, n$ and $d_{ij}^{(1)} = 0$ for $i \neq j$). D_1 is nonsingular, since, as noted above, $l_{jj} > 0$ for $j = 1, \ldots, n$. Moreover, $\tilde{L} = LD_1^{-1}$ is unit lower-triangular. If we also let $D = D_1D_1 = D_1D_1^{\mathrm{T}}$, then

$$A = LL^{\mathrm{T}} = (\tilde{L}D_1)(\tilde{L}D_1)^{\mathrm{T}} = \tilde{L}D_1D_1^{\mathrm{T}}\tilde{L}^{\mathrm{T}} = \tilde{L}D\tilde{L}^{\mathrm{T}},$$

where $\tilde{L}D\tilde{L}^{\mathrm{T}}$ is the *LDL* factorization of A. Furthermore, if we let $U = D\tilde{L}^{\mathrm{T}}$, then $\tilde{L}U$ is the *LU* factorization of A.

The *LDL* factorization can be computed directly by equating the terms of A to those of $\tilde{L}D\tilde{L}^{\mathrm{T}}$, just as we did above for the Cholesky factorization. This leads to a scheme similar to that shown in Table 4, but without any square roots, although it has $n(n - 1)/2 \approx n^2/2$ more multiplications. Thus the cost of computing the factorization remains about $n^3/6$ flops and the storage requirement remains about $n^2/2$.

An advantage of the *LDL* factorization is that it can be applied to a symmetric indefinite matrix. The Cholesky factorization is not applicable in this case, since LL^T is always symmetric positive-semidefinite. However, since LDL^T is always symmetric, pivoting must be restricted to ensure that the reordered matrix PAQ^T is symmetric. The simplest way to maintain symmetry is to use *symmetric pivoting* in which $Q = P$. This, though, restricts the choice of the pivot at stage k of the LDL factorization to a_{jj} for $j = k, \ldots, n$. As a result, in some cases, the LDL factorization may incur much more rounding error than GE with partial or complete pivoting.

2.7 Banded and Sparse Matrices

Significant savings in both computational work and storage can often be obtained in solving $Ax = b$ by taking advantage of zeros in the coefficient matrix A. We outline in this subsection how these savings may be realized for banded and more general sparse matrices.

To begin, note that an $n \times n$ matrix A is said to be *sparse* if the number of nonzero elements in A is much less than n^2, the total number of elements in A. *Banded matrices* are an important subclass of sparse matrices in which the nonzero elements of the matrix are restricted to a band around the main diagonal of the matrix. The *lower bandwidth* of a matrix $A = [a_{ij}]$ is the smallest integer p such that $a_{ij} = 0$ for $i - j > p$, the *upper bandwidth* of A is the smallest integer q such that $a_{ij} = 0$ for $j - i > q$, and the *bandwidth* of A is $1 + p + q$. Clearly, if p and $q \ll n$, then A is sparse, since A has at most $(1 + p + q)n \ll n^2$ nonzero elements.

Banded and more general sparse matrices arise in many important applications. For example, quadratic spline interpolation, as described in Sec. 8.6.2, requires the solution of a linear system $Tc = g$, where T is a tridiagonal matrix (i.e., banded with $p = q = 1$) and c is the vector of coefficients for the quadratic spline interpolant. Banded matrices also arise in the solution of boundary-value problems for ordinary differential equations. See Sec. 11.3.1 for an example of a system $Tu = g$, where T is a symmetric positive-definite tridiagonal matrix. More general sparse matrices arise in the numerical solution of partial differential equations. See Sec. 11.3.2 for an example of the matrix associated with the standard five-point difference scheme for Poisson's equation.

If A is large and sparse, it is common to store only the nonzero elements of A, since this greatly reduces the storage requirements. This is easy to do if A is banded, since we can map the elements from the band of A to a $(1 + p + q) \times n$ array representing A in a *packed format*, several of which are commonly used in practice. If $A = [a_{ij}]$ is a general sparse matrix, then we require a more general *sparse-matrix data structure* that stores each nonzero element a_{ij} of A along with some information used to recover the indices i and j.

We consider Gaussian elimination (GE) for a banded matrix first. To begin, note that the reduction process described in Sec. 2.1 maintains the band structure of A. In particular, $a_{ik}^{(k-1)} = 0$ for $i - k > p$, whence the multipliers $m_{ik} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ in Table 1 need to be calculated for $i = k + 1, \ldots \min(k + p, n)$ only and the i loop can be changed accordingly. Similarly $a_{kj}^{(k-1)} = 0$ for $j - k > q$, and so the reduction $a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ij}a_{kj}^{(k-1)}$ in Table 1 needs to be calculated for $j = k + 1, \ldots, \min(k + q, n)$ only and the j loop can be charged accordingly. It therefore follows from a straightforward operation count that GE modified as described above for banded matrices requires $np - p(p + 1)/2 \approx np$ divisions to compute the multipliers m_{ik}, either $npq - p(3q^2 + 3q + p^2 - 1)/6 \approx npq$ (if $p \leq q$) or $npq - q(3p^2 + 3p + q^2 - 1)/6 \approx npq$ (if $p \geq q$) multiplications and subtractions to compute the coefficients of U, and $np - p(p + 1)/2 \approx np$ multiplications and subtractions to compute the coefficients of $\tilde{b}$. Furthermore, note that, if we use this modified GE procedure to compute the LU factorization of A, then the lower-triangular matrix L has lower bandwidth p and the upper-triangular matrix U has upper bandwidth q. As noted in Sec. 2.1, it is common to overwrite A with L and U. This can be done even if A is stored in packed format, thereby achieving significant reduction in storage requirements.

The back-substitution method shown in Table 2 and the forward-elimination method shown in Table 3 can be modified similarly so that each requires n divisions, the back-substitution method requires $nq - q(q + 1)/2 \approx nq$ multiplications and subtractions, while the forward-elimination method requires $np - p(p + 1)/2 \approx np$ multiplications and subtractions. In addition, recall that the n divisions are not needed in forward elimination if L is unit lower-triangular, as is the case for the modified LU factorization described here.

A similar modification of the Cholesky method shown in Table 4 results in a procedure that requires n square roots, $np - p(p + 1)/2 \approx np$ divisions, and $(n - p)p(p + 1)/2 + (p - 1)p(p + 1)/6 \approx np^2/2$ multiplications and subtractions. In deriving these operation counts, we used $p = q$, since the matrix A must be symmetric for the Cholesky factorization to be applicable. Moreover, the Cholesky factor L has lower bandwidth p. Thus, as for the general case, the Cholesky factorization of a band matrix requires about half as many arithmetic operations and about half as much storage as the LU factorization, since packed storage can also be used for the Cholesky factor L.

If partial pivoting is used in the LU factorization of A, then the upper bandwidth of U may increase to $p + q$. The associated matrix L is a permuted version of a lower-triangular matrix with lower bandwidth p. Both factors L and U can be stored in packed format in a $(1 + 2p + q) \times n$ array. The operation count for the LU factorization, forward elimination, and back solution is the same as though A were a banded matrix with upper bandwidth $p + q$ and lower bandwidth p. Thus, if $p > q$, both computational work and storage can be saved by factoring A^{T} instead of A and using the LU factors of A^{T} to solve $Ax = b$. If complete pivoting is used in the LU factorization of A, then L and U may fill in so much that there is little advantage to using a band solver. Consequently, it is advantageous not to pivot when factoring a band matrix, provided this does not lead to an unacceptable growth in rounding errors. As noted in Sec. 2.5, it is not necessary to pivot for numerical stability if A is complex Hermitian positive definite, real symmetric positive definite, or column-diagonally dominant. If pivoting is required, it is advantageous to use partial, rather than complete, pivoting, again provided this does not lead to an unacceptable growth in rounding errors.

Extending GE to take advantage of the zeros in a general sparse matrix is considerably more complicated than for banded matrices. The difficulty is that, when row k of A_{k-1} is multiplied by m_{ik} and added to row i of A_{k-1} to eliminate $a_{ik}^{(k-1)}$ in stage k of GE, as described in Sec. 2.1, some zero elements in row i of A_{k-1} may become nonzero in the resulting row i of A_k. These elements are said to *fill in* and are collectively referred to as *fill-in* or *fill.*

However, pivoting can often greatly reduce the amount of fill-in. To see how this comes about, the interested reader may wish to work through an example with the *arrowhead* matrix

$$A = \begin{pmatrix} 5 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Since A is a real symmetric positive-definite matrix, there is no need to pivot for numerical stability. It is easy to see that the Cholesky and LU factors of A completely fill in. Interchanging the first and last rows and the first and last columns of A corresponds to forming the permuted matrix $B = PAP^{\mathrm{T}}$, where $P = I - [1,0,0,0,-1]^{\mathrm{T}}[1,0,0,0,-1]$ is a permutation matrix. Since B is also a real symmetric positive-definite matrix, there is no need to pivot for numerical stability when factoring B. However, in this case, the Cholesky and LU factors of B suffer no fill-in at all. This small example can be generalized easily to arbitrarily large arrowhead matrices having the similar properties that the LU factors of A completely fill in while those of the permuted matrix $B = PAP^{\mathrm{T}}$ suffer no fill-in at all.

If we use an appropriate sparse-matrix data structure to store only the nonzeros elements of a sparse matrix A and its LU factors, then reducing the fill-in reduces both the storage needed for the LU factors and the computational work required to calculate them. It also reduces the computational work required for forward elimination and back substitution, since the methods shown in Tables 2 and 3 can be modified easily so that they use the nonzero elements in L and U only, thereby avoiding multiplications by zero and the associated subtractions.

Therefore, the goal in a sparse LU or Cholesky factorization is to reorder the rows and columns of A to reduce the amount of fill-in.

If we need to pivot to ensure numerical stability, then this might conflict with pivoting to reduce fill-in. Unfortunately, even without this complication, finding the optimal reordering to minimize the fill-in is computationally too expensive to be feasible in general. There are, though, many good *heuristic* methods to reorder the rows and columns of A that greatly reduce the fill-in and computational work in many important cases. The reader seeking a more complete description of sparse-matrix factorizations should consult a text on this topic, such as Duff *et al.* (1986) or George and Liu (1981).

We end this subsection with an example illustrating the importance of exploiting the zeros in a sparse matrix A. Consider the linear system derived in Sec. 11.3.2 by discretizing Poisson's equation on an $m \times m$ grid. A is an $n \times n$ symmetric positive-definite matrix with $n = m^2$ and $5m^2 - 4m \approx 5m^2 = 5n$ nonzero elements, out of a total of n^2 elements in A. So, if m is large, A is very sparse. Moreover, if we use the *natural ordering* for the equations and variables in the system, then A, as shown in Sec. 11.3.2, is a banded matrix with lower and upper bandwidth $m = \sqrt{n}$.

The computational work and storage required to solve this linear system are shown in Table 5. We consider three cases:

1. dense, the zeros in A are not exploited at all;
2. banded, we use a band solver that exploits the band structure of A;
3. sparse, we use a general sparse solver together with the *nested dissection ordering* (see George and Liu, 1981) for the equations and variables in the system.

The columns labeled factor and solve, respectively, give the approximate number of flops needed to compute the Cholesky factorization of the matrix A and to solve the linear system given the factors. The columns labeled store A and store L, respectively, give the approximate number of storage locations needed to store A and its Cholesky factor L.

Table 5. Computational work and storage required to solve the linear system $Ax = b$ derived from discretizing Poisson's equation on an $m \times m$ grid.

	Factor	Solve	Store A	Store L
Dense	$m^6/6$	m^4	$m^4/2$	$m^4/2$
Banded	$m^4/2$	$2m^3$	m^3	m^3
Sparse	$O(m^3)$	$O(m^2 \log m)$	$5m^2$	$O(m^2 \log m)$

2.8 Rounding Errors, Condition Numbers, and Error Bounds

In this subsection, we consider the effects of rounding errors in solving $Ax = b$. In doing so, we use vector and matrix norms extensively. Therefore, we recommend that, if you are not acquainted with norms, you familiarize yourself with this topic before reading this subsection. Most introductory numerical methods texts or advanced books on numerical linear algebra contain a section on vector and matrix norms.

The analysis of the effects of rounding errors in solving $Ax = b$ usually proceeds in two stages. First we establish that the computed solution $\tilde{x}$ is the exact solution of a perturbed system

$$(A + E)\tilde{x} = b + r \tag{14}$$

with bounds on the size of E and r. Then we use (14) together with bounds on the size of A and A^{-1} to bound the error $x - \tilde{x}$.

The first step is called a *backward error analysis,* since it casts the error in the solution back onto the problem and allows us to relate the effects of rounding errors in the computed solution $\tilde{x}$ to other errors in the problem, such as measurement errors in determining the coefficients of A and b. Another advantage of proceeding in this two-stage fashion is that, if $\tilde{x}$ is not sufficiently accurate, it allows us to determine whether this is because the numerical method is faulty or whether the problem itself is unstable.

Throughout this section we assume that A is an $n \times n$ nonsingular matrix and $nu < 0.1$, where u is the *relative roundoff error bound* for the machine arithmetic used in the computation (see Sec. 1.2).

If we use Gaussian elimination (GE) with either partial or complete pivoting together with forward elimination and back substitution to solve $Ax = b$, then it can be shown that the computed solution $\tilde{x}$ satisfies (14) with $r = 0$ and

$$\|E\|_\infty \le 8n^3\gamma\|A\|_\infty u + O(u^2), \tag{15}$$

where

$$\gamma = \max_{i,j,k} \frac{|a_{ij}^{(k)}|}{\|A\|_\infty}$$

is the *growth factor* and $A_k = [a_{ij}^{(k)}]$ for $k = 0, \ldots, n - 1$ are the intermediate reduced matrices produced during the GE process (see Secs. 2.1 and 2.5). It can be shown that $\gamma \leq 2^{n-1}$ for GE with partial pivoting and $\gamma \leq [n(2 \times 3^{1/2} \times 4^{1/3} \times \cdots n^{1/(n-1)})]^{1/2} \ll 2^{n-1}$ for GE with complete pivoting. Moreover, the former upper bound can be achieved. However, for GE with both partial and complete pivoting, the actual error incurred is usually much smaller than the bound (15) suggests: it is typically the case that $\|E\|_\infty \propto \|A\|_\infty u$. Thus, GE with partial pivoting usually produces a computed solution with a small backward error, but, unlike GE with complete pivoting, there is no guarantee that this will be the case.

If A is column-diagonally dominant, then applying GE without pivoting to solve $Ax = b$ is effectively the same as applying GE with partial pivoting to solve this system. Therefore, all the remarks above for GE with partial pivoting apply in this special case.

If we use the Cholesky factorization without pivoting together with forward elimination and back substitution to solve $Ax = b$, where A is a symmetric positive-definite matrix, then it can be shown that the computed solution $\tilde{x}$ satisfies (14) with $r = 0$ and

$$\|E\|_2 \leq c_n \|A\|_2 u, \tag{16}$$

where c_n is a constant of moderate size that depends on n only. Thus the Cholesky factorization without pivoting produces a solution with a small backward error in all cases.

Similar bounds on the backward error for other factorizations and matrices with special properties, such as symmetric or band matrices, can be found in advanced texts, such as Golub and Van Loan (1989).

It is also worth noting that we can easily compute an *a posteriori* backward error estimate of the form (14) with $E = 0$ by calculating the *residual* $r = A\tilde{x} - b$ after computing $\tilde{x}$. Typically $\|r\|_\infty \propto \|b\|_\infty u$ if GE with partial or complete pivoting or the Cholesky factorization is used to compute $\tilde{x}$.

Now we use (14) together with bounds on the size of A and A^{-1} to bound the error $x - \tilde{x}$. To do so, we first introduce the *condition number* $\kappa(A) = \|A\| \, \|A^{-1}\|$ associated with the problem $Ax = b$. Although $\kappa(A)$ clearly depends on the matrix norm used, it is roughly of the same magnitude for all the commonly used norms, and it is the magnitude only of $\kappa(A)$ that is important here. Moreover, for the result below to hold, we require only that the matrix norm associated with $\kappa(A)$ is submultiplicative (i.e., $\|AB\| \leq \|A\| \, \|B\|$ for all A and $B \in \mathbb{C}^{n \times n}$) and that it is consistent with the vector norm used (i.e., $\|Av\| \leq \|A\| \, \|v\|$ for all $A \in \mathbb{C}^{n \times n}$ and $v \in \mathbb{C}^n$). It follows immediately from these two properties that $\kappa(A) \geq 1$ for all $A \in \mathbb{C}^{n \times n}$. More importantly, it can be shown that, if $\|E\|/\|A\| \leq \delta$, $\|r\|/\|b\| \leq \delta$, and $\delta\kappa(A) = r < 1$, then $A + E$ is nonsingular and

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \frac{2\delta}{1 - r} \kappa(A). \tag{17}$$

Moreover, for any given A, there are some b, E, and r for which $\|x - \tilde{x}\|/\|x\|$ is as large as the right side of (17) suggests it might be, although this is not the case for all b, E, and r. Thus we see that, if $\kappa(A)$ is not too large, then small relative errors $\|E\|/\|A\|$ and $\|r\|/\|b\|$ ensure a small relative error $\|x - \tilde{x}\|/\|x\|$, and so the problem is *well-conditioned*. On the other hand, if $\kappa(A)$ is large, then $\|x - \tilde{x}\|/\|x\|$ might be large even though $\|E\|/\|A\|$ and $\|r\|/\|b\|$ are small, and so the problem is *ill-conditioned*. Thus, as the name suggests, the condition number $\kappa(A)$ gives a good measure of the conditioning—or stability—of the problem $Ax = b$.

Combining the discussion above with the earlier observation that typically $\|r\|_\infty \propto \|b\|_\infty u$ if GE with partial or complete pivoting or the Cholesky factorization is used to compute $\tilde{x}$, we get the general rule of thumb that, if $u \approx 10^{-d}$ and $\kappa_\infty = \|A\|_\infty \|A^{-1}\|_\infty \approx 10^q$, then $\tilde{x}$ contains about $d - q$ correct digits.

Many routines for solving linear systems provide an estimate of $\kappa(A)$, although most do not compute $\|A\| \, \|A^{-1}\|$ directly, since they do not compute A^{-1}.

There are many other useful inequalities of the form (17). The interested reader should consult an advanced text, such as Golub and Van Loan (1989).

2.9 Iterative Improvement

The basis of *iterative improvement* is the observation that, if x_1 is an approximate solution to $Ax = b$, we can form the residual $r_1 = b - Ax_1$, which satisfies $r_1 = A(x - x_1)$, and then solve $Ad_1 = r_1$ for the difference $d_1 = x - x_1$ and finally compute the improved solution $x_2 = x_1 + d_1$. In exact arithmetic, $x_2 = x$, but, in floating-point arithmetic, $x_2 \neq x$ normally. So we can repeat the process using x_2 in place of x_1 to form another improved solution x_3, and so on. Moreover, if we have factored A to compute x_1, then, as noted in Secs. 2.3 and 2.5, there is relatively little extra computational work required to compute a few iterations of iterative improvement.

The catch here is that, as noted in Sec. 2.8, typically $\|r\|_\infty \propto \|b\|_\infty u$ if GE with partial or complete pivoting or the Cholesky factorization is used to compute $\tilde{x}$. So, if we compute $r_1 = b - Ax_1$ in the same precision, then r_1 will contain few if any correct digits. Consequently, using it in iterative improvement usually does not lead to a reduction of the error in x_2, although it may lead to a smaller E in (14) in some cases. However, if we compute $r_k = b - Ax_k$ in double precision for $k = 1, 2, \ldots,$ then iterative improvement may be quite effective. Roughly speaking, if the relative roundoff error bound $u \approx 10^{-d}$ and the condition number $\kappa(A) = \|A\|\,\|A^{-1}\| \approx 10^q$, then after k iterations of iterative improvement the computed solution x_k typically has about $\min(d, k(d - q))$ correct digits. Thus, if $\kappa(A)$ is large, but not too large, and, as a result, the initial computed solution x_1 is inaccurate, but not completely wrong, then iterative improvement can be used to obtain almost full single-precision accuracy in the solution.

The discussion above can be made more rigorous by noting that iterative improvement is a basic iterative method of the form (19)–(20) for solving $Ax = b$ and applying the analysis in Sec. 3.1.

3. THE ITERATIVE SOLUTION OF LINEAR ALGEBRAIC SYSTEMS

In this section, we consider *iterative* methods for the numerical solution of linear algebraic systems of the form $Ax = b$, where $A \in \mathbb{R}^{n\times n}$ (or $\mathbb{C}^{n\times n}$) is a nonsingular matrix, x, and $b \in \mathbb{R}^n$ (or $\mathbb{C}^n$). Such schemes compute a sequence of approximations $x_1, x_2, \ldots$ to x in the hope that $x_k \to x$ as $k \to \infty$. Direct methods for solving $Ax = b$ are considered in Sec. 2.

Iterative methods are most frequently used when A is large and sparse, but not banded with a small bandwidth. Such matrices arise frequently in the numerical solution of partial differential equations (PDEs). See, for example, the matrix shown in Sec. 11.3.2 that is associated with the standard five-point difference scheme for Poisson's equation. In many such cases, iterative methods are more efficient than direct methods for solving $Ax = b$: They usually use far less storage and often require significantly less computational work as well.

We discuss *basic iterative methods* in Sec. 3.1 and the *conjugate-gradient acceleration* of these schemes in Sec. 3.2. A more complete description and analysis of these methods and other iterative schemes are provided in Axelsson (1994), Golub and Van Loan (1989), Hageman and Young (1981), Young (1971), and Varga (1962). See Sec. 13 for a discussion of sources of high-quality numerical software for solving systems of linear algebraic equations.

We discuss multigrid methods in Sec. 11.8, because these iterative schemes are so closely tied to the PDE that gives rise to the linear system $Ax = b$ to which they are applied.

3.1 Basic Iterative Methods

Many iterative methods for solving $Ax = b$ are based on splitting the matrix A into two parts, M and N, such that $A = M - N$ with M nonsingular. M is frequently called the *splitting matrix.* Starting from an initial guess x_0 for x, we compute $x_1, x_2, \ldots$ recursively from

$$Mx_{k+1} = Nx_k + b. \tag{18}$$

We call such a scheme a *basic iterative method,* but it is often also referred to as a *linear stationary method of the first degree.*

Since $N = M - A$, (18) can be rewritten as

$$Mx_{k+1} = (M - A)x_k + b = Mx_k + (b - Ax_k),$$

which is equivalent to

$$Md_k = r_k, \tag{19}$$

$$x_{k+1} = x_k + d_k, \tag{20}$$

where $r_k = b - Ax_k$ is the residual at iteration k. Although (18) and (19)–(20) are mathematically equivalent, it might be computationally more effective to implement a method in one form than the other.

Clearly, for either (18) or (19)–(20) to be effective,

1. it must be much easier to solve systems with M than with A, and
2. the iterates $x_1, x_2, \ldots$ generated by (18) or (19)–(20) must converge quickly to x, the solution of $Ax = b$.

To address point **2**, first note that, since $Ax = b$ and $A = M - N$, $Mx = Nx + b$. So, if the sequence $x_1, x_2, \ldots$ converges, it must converge to x. To determine whether the sequence $x_1, x_2, \ldots$ converges and, if so, how fast, subtract (18) from $Mx = Nx + b$ and note that the error $e_k = x - x_k$ satisfies the recurrence $Me_{k+1} = Ne_k$, or equivalently $e_{k+1} = Ge_k$, where

$$G = M^{-1}N = I - M^{-1}A$$

is the associated *iteration matrix.* So

$$e^k = G^k e_0. \tag{21}$$

Using (21), we can show that, starting from any initial guess x_0, the sequence $x_1, x_2, \ldots$ generated by (18) converges to x if and only if $\rho(G) < 1$, where

$$\rho(G) = \max\{|\lambda| : \lambda \text{ an eigenvalue of } G\}$$

is the *spectral radius* of G. Moreover, $\rho(G)$ is the "asymptotically average" amount by which the error e_k decreases at each iteration. Consequently, $(\log \epsilon)/\log \rho(G)$ is a rough estimate of the number of iterations of (18) required to reduce the initial error e_0 by a factor ϵ. Thus, it is common to define

$$R(G) = -\log \rho(G) \tag{22}$$

to be the *rate of convergence* (sometimes called the *asymptotic rate of convergence* or the *asymptotic average rate of convergence*) of the iteration (18).

One useful general result is that, if $A = M - N$ is Hermitian positive-definite and if the Hermitian matrix $M^{\mathrm{H}} + N$ is positive-definite too, then $\rho(G) < 1$ and the associated iteration (18) converges.

Possibly the simplest iterative scheme is the *RF method* (a variant of Richardson's method) for which $M = I$ and $N = I - A$, whence (18) reduces to

$$x_{k+1} = x_k + r_k,$$

where $r_k = b - Ax_k$ is the residual at iteration k. From the general discussion above, it follows that this scheme converges if and only if $\rho(I - A) < 1$. Because of this severe constraint on convergence, this scheme is not often effective in its own right, but it can be used productively as the basis for polynomial acceleration, as discussed in Sec. 3.2.

We describe the Jacobi, Gauss–Seidel, SOR, and SSOR methods next, and then consider their convergence. In describing them, we use the notation $A = D - L - U$, where D is assumed to be nonsingular and consists of the diagonal elements of A for the point variant of each method or the diagonal submatrices of A for the block variant. L and U are the negatives of the strictly lower- and upper-triangular parts of A, respectively, either point or block, as the case may be. Typically, the block variant of each method converges faster than the point version, but requires more computational work per iteration. Thus, it is usually not clear without additional analysis which variant will be more effective.

The *Jacobi iteration* takes $M_{\mathrm{J}} = D$ and $N_{\mathrm{J}} = L + U$, resulting in the recurrence

$$Dx_{k+1} = (L + U)x_k + b. \tag{23}$$

The associated iteration matrix is $G_{\mathrm{J}} = D^{-1}(L + U)$. The *Gauss–Seidel iteration* takes $M_{\mathrm{GS}} = D - L$ and $N_{\mathrm{GS}} = U$, resulting in the recurrence

$$(D - L)x_{k+1} = Ux_k + b. \tag{24}$$

The associated iteration matrix is $G_{\mathrm{GS}} = (D - L)^{-1}U$. Although (24) may at first appear a little more complicated than (23), it is in fact easier to implement in practice on a sequential machine, since one can overwrite x_k when computing x_{k+1} in (24), whereas this is generally not possible for (23). However, for the Gauss–Seidel iteration, the jth component of x_{k+1} might depend on the ith component of x_{k+1} for $i < j$, because of the factor L on the left side of (24). This often inhibits vectorization and parallelization of the Gauss–Seidel iteration. Note that the Jacobi iteration has no such dependence, and so might be more effective on a vector or parallel machine.

Relaxation methods for $Ax = b$ can be written in the form

$$x_{k+1} = x_k + \omega(\hat{x}_{k+1} - x_k), \tag{25}$$

where $\omega \neq 0$ is the relaxation parameter and $\hat{x}_{k+1}$ is computed from x_k by some other iterative method. The best known of these schemes is *successive overrelaxation* (SOR) for which

$$D\hat{x}_{k+1} = Lx_{k+1} + Ux_k + b. \tag{26}$$

Equations (25) and (26) can be combined to give

$$\left(\frac{1}{\omega}D - L\right)x_{k+1} = \left(\frac{1-\omega}{\omega}D + U\right)x_k + b, \tag{27}$$

which is an iteration of the form (18) with $M_{\text{SOR}}(\omega) = (1/\omega)D - L$ and $N_{\text{SOR}}(\omega) = [(1 - \omega)/\omega]D + U$. It follows immediately from (24) and (27) that the SOR iteration reduces to the Gauss–Seidel method if $\omega = 1$. Moreover, because of the similarity between (24) and (27), the SOR iteration shares with the Gauss–Seidel method the implementation advantages and disadvantages noted above.

Overrelaxation corresponds to choosing $\omega > 1$ in (25) or (27), while underrelaxation corresponds to choosing $\omega \in (0,1)$. Historically, $\omega > 1$ was used in SOR for the solution of elliptic PDEs—hence the name successive *over*relaxation—but underrelaxation is more effective for some problems. See for example Young (1971) for a more complete discussion.

The *symmetric SOR* (SSOR) method takes one half step of SOR with the equations solved in the standard order followed by one half step of SOR with the equations solved in the reverse order:

$$\left(\frac{1}{\omega}D - L\right)x_{k+1/2} = \left(\frac{1-\omega}{\omega}D + U\right)x_k + b, \tag{28}$$

$$\left(\frac{1}{\omega}D - U\right)x_{k+1} = \left(\frac{1-\omega}{\omega}D + L\right)x_{k+1/2} + b. \tag{29}$$

These two half steps can be combined into one step of the form (18) with

$$M_{\text{SSOR}}(\omega) = \frac{\omega}{2-\omega}\left(\frac{1}{\omega}D - L\right)D^{-1}\left(\frac{1}{\omega}D - U\right)$$

and

$$N_{\text{SSOR}}(\omega) = \frac{(\omega-1)^2}{\omega(2-\omega)}D + \frac{1-\omega}{2-\omega}(L+U) + \frac{\omega}{2-\omega}LD^{-1}U.$$

Note that, if $A = D - L - U$ is a real, symmetric, positive-definite matrix and $\omega \in (0,2)$, then $M_{\text{SSOR}}(\omega)$ is a real, symmetric, positive-definite matrix too, since, in this case, $\omega/(2 - \omega) > 0$, both D and D^{-1} are symmetric positive-definite, and $(1/\omega)D - L = [(1/\omega)D - U]^{\mathrm{T}}$ is nonsingular. This property of $M_{\text{SSOR}}(\omega)$ plays an important role in the effective acceleration of the SSOR iteration, as discussed in Sec. 3.2.

We now consider the convergence of the Jacobi, Gauss–Seidel, SOR, and SSOR methods. It is easy to show that the Jacobi iteration (23) converges if A is either row- or column-diagonally dominant. It can also be shown that the Gauss–Seidel iteration (24) converges if A is Hermitian positive-definite. Furthermore, if A is *consistently ordered,* a property enjoyed by a large class of matrices, including many that arise from the discretization of PDEs [see, for example, Young (1971) for details], it can be shown that the Gauss–Seidel iteration converges twice as fast as the Jacobi iteration if either one converges.

The iteration matrix associated with the SOR iteration (27) is

$$\begin{aligned} G_{\text{SOR}}(\omega) &= [M_{\text{SOR}}(\omega)]^{-1}N_{\text{SOR}}(\omega) \\ &= \left(\frac{1}{\omega}D - L\right)^{-1}\left(\frac{1-\omega}{\omega}D + U\right). \end{aligned}$$

For any nonsingular A with nonsingular D, it can be shown that $\rho(G_{\mathrm{SOR}}(\omega)) \geq |\omega - 1|$ with equality possible if and only if all eigenvalues of $G_{\mathrm{SOR}}(\omega)$ have magnitude $|\omega - 1|$. So, if $\omega \in \mathbb{R}$, as is normally the case, a necessary condition for the convergence of SOR is $\omega \in (0,2)$. It can also be shown that, if A is Hermitian, D is positive-definite, and $\omega \in \mathbb{R}$, then the SOR iteration converges if and only if A is positive-definite and $\omega \in (0,2)$. Furthermore, if A is consistently ordered and all the eigenvalues of $G_{\mathrm{J}} = D^{-1}(L + U)$ are real and lie in $(-1,1)$, then the optimal choice of the SOR parameter ω is

$$\omega_0 = \frac{2}{1 + \sqrt{1 - [\rho(G_{\mathrm{J}})]^2}} \in (1,2)$$

and

$$\begin{aligned}\rho(G_{\mathrm{SOR}}(\omega_0)) &= \omega_0 - 1 = \left(\frac{\rho(G_{\mathrm{J}})}{1 + \sqrt{1 - [\rho(G_{\mathrm{J}})]^2}}\right)^2 \\ &= \min_{\omega} \rho(G_{\mathrm{SOR}}(\omega)) < \rho(G_1) \\ &= \rho(G_{\mathrm{GS}}) = [\rho(G_{\mathrm{J}})]^2 < \rho(G_{\mathrm{J}}).\end{aligned} \tag{30}$$

In many cases, though, it is not convenient to calculate the optimal ω. Hageman and Young (1981) discuss heuristics for choosing a "good" ω.

If A is a real, symmetric, positive-definite matrix, then the SSOR iteration (28)–(29) converges for any $\omega \in (0,2)$. Moreover, determining the precise value of the optimal ω is not nearly as critical for SSOR as it is for SOR, since, unlike SOR, the rate of convergence of SSOR is relatively insensitive to the choice of ω. However, for SSOR to be effective,

$$\rho(D^{-1}LD^{-1}U) \leq \tfrac{1}{4}$$

should be satisfied—or nearly so. If this is the case, then a good value for ω is

$$\omega_1 = \frac{2}{1 + \sqrt{2[1 - \rho(G_{\mathrm{J}})]}}$$

and

$$\rho(G_{\mathrm{SSOR}}(\omega_1)) \leq \frac{1 - \{[1 - \rho(G_{\mathrm{J}})]/2\}^{1/2}}{1 + \{[1 - \rho(G_{\mathrm{J}})]/2\}^{1/2}}, \tag{31}$$

where $G_{\mathrm{SSOR}}(\omega) = [M_{\mathrm{SSOR}}(\omega)]^{-1}N_{\mathrm{SSOR}}(\omega)$ is the SSOR iteration matrix.

For a large class of problems, including many that arise from the discretization of elliptic PDEs, $\rho(G_{\mathrm{J}}) = 1 - \epsilon$ for some ϵ satisfying $0 < \epsilon \ll 1$. For such problems, if (30) is valid and (31) holds with $=$ in place of $\leq$, then

$$\rho(G_{\mathrm{J}}) = 1 - \epsilon,$$

$$\rho(G_{\mathrm{GS}}) = (\rho(G_{\mathrm{J}}))^2 \approx 1 - 2\epsilon,$$

$$\rho(G_{\mathrm{SSOR}}(\omega_1)) \approx 1 - 2\sqrt{\epsilon/2},$$

$$\rho(G_{\mathrm{SOR}}(\omega_0)) \approx 1 - 2\sqrt{2\epsilon},$$

whence the rates of convergence for these schemes are

$$R(G_{\mathrm{J}}) \approx \epsilon,$$

$$R(G_{\mathrm{GS}}) = 2R(G_{\mathrm{J}}) \approx 2\epsilon,$$

$$R(G_{\mathrm{SSOR}}(\omega_1)) \approx 2\sqrt{\epsilon/2},$$

$$R(G_{\mathrm{SOR}}(\omega_0)) \approx 2\sqrt{2\epsilon} \approx 2R(G_{\mathrm{SSOR}}(\omega_1)),$$

showing that the Gauss–Seidel iteration converges twice as fast as the Jacobi iteration, the SOR iteration converges about twice as fast as the SSOR iteration, and the SOR and SSOR iterations converge much faster than either the Gauss–Seidel or Jacobi iteration. However, the SSOR iteration often has the advantage, not normally shared by the SOR method, that it can be accelerated effectively, as discussed in Sec. 3.2.

Another class of basic iterative methods is based on *incomplete factorizations*. For brevity, we describe only the subclass of *incomplete Cholesky factorizations* (ICFs) here; the other schemes are similar. See an advanced text, such as Axelsson (1994), for details.

For an ICF to be effective, A should be symmetric positive-definite (or nearly so), large, and sparse. If A is banded, then the band containing the nonzeros should also be sparse, as is the case for the discretization of Poisson's equation shown in Sec. 11.3.2. The general idea behind the ICFs is to compute a lower-triangular matrix L_{ICF} such that $M_{\mathrm{ICF}} = L_{\mathrm{ICF}}L_{\mathrm{ICF}}^{\mathrm{T}}$ is in some sense close to A and L_{ICF} is much sparser than L, the true Cholesky factor of A. Then employ the iteration (19)–(20) to compute a sequence of approximations $x_1, x_2, \ldots$ to x, the solution of $Ax =$

b. Note that (19) can be solved efficiently by forward elimination and back substitution as described in Secs. 2.4 and 2.2, respectively, since the factorization $M_{\mathrm{ICF}} = L_{\mathrm{ICF}} L_{\mathrm{ICF}}^{\mathrm{T}}$ is known. This scheme, or an accelerated variant of it, is often very effective if it converges rapidly and L_{ICF} is much sparser than L.

A simple, but often effective, way of computing $L_{\mathrm{ICF}} = [l_{ij}]$ is to apply the Cholesky factorization described in Table 4, but to set $l_{ij} = 0$ whenever $a_{ij} = 0$, where $A = [a_{ij}]$. Thus, L_{ICF} has the same sparsity pattern as the lower-triangular part of A, whereas the true Cholesky factor L of A might suffer significant fill-in, as described in Sec. 2.7. Unfortunately, this simple ICF is not always stable.

As noted in Sec. 2.9, iterative improvement is a basic iterative method of this form, although, for iterative improvement, the error $N = M - A$ is due entirely to rounding errors, whereas, for the incomplete factorizations considered here, the error $N = M - A$ is typically also due to dropping elements from the factors of M to reduce fill-in.

For a more complete discussion of ICFs and other incomplete factorizations, their convergence properties, and their potential for acceleration, see an advanced text such as Axelsson (1994).

We end this section with a brief discussion of *alternating-direction implicit* (ADI) methods. A typical example of a scheme of this class is the Peaceman–Rachford method

$$(H + \alpha_n I)x_{n+1/2} = b - (V - \alpha_n I)x_n, \tag{32}$$

$$(V + \alpha_n' I)x_{n+1} = b - (H - \alpha_n' I)x_{n+1/2}, \tag{33}$$

where $A = H + V$, $\alpha_n > 0$, $\alpha_n' > 0$, and A, H, and V are real, symmetric, positive-definite matrices. For many problems, it is possible to choose H, V, and $\{\alpha_n, \alpha_n'\}$, so that the iteration (32)–(33) converges rapidly, and it is much cheaper to solve (32) and (33) than it is to solve $Ax = b$. For example, for the standard five-point discretization of a separable two-dimensional elliptic PDE, H and V can be chosen to be essentially tridiagonal and the rate of convergence of (32)–(33) is proportional to $1/\log h^{-1}$, where h is the mesh size used in the discretization. In contrast, the rate of convergence for SOR with the optimal ω is proportional to h. Note that $h \ll 1/\log h^{-1}$ for $0 < h \ll 1$, supporting the empirical evidence that ADI schemes converge much more rapidly than SOR for many problems. However, ADI schemes are not applicable to as wide a class of problems as SOR is.

For a more complete discussion of ADI schemes, see an advanced text such as Varga (1962) or Young (1971).

3.2 The Conjugate-Gradient Method

The *conjugate-gradient* (CG) *method* for the solution of the linear system $Ax = b$ is a member of a broader class of methods often called polynomial acceleration techniques or Krylov-subspace methods. (The basis of these names is explained below.) Although many schemes in this broader class are very useful in practice, we discuss only CG here, but note that several of these schemes, including Chebyshev acceleration and GMRES, apply to more general problems than CG. The interested reader should consult an advanced text such as Axelsson (1994), Golub and Van Loan (1989), Hageman and Young (1981) for a more complete discussion of polynomial acceleration techniques and Krylov-subspace methods. The close relationship between CG and the Lanzcos method is discussed in Golub and Van Loan (1989).

The *preconditioned conjugate-gradient* (PCG) *method* can be viewed either as an acceleration technique for the basic iterative method (18) or as CG applied to the preconditioned system $M^{-1}Ax = M^{-1}b$, where the splitting matrix M of (18) is typically called a preconditioning matrix in this context. We adopt the second point of view in this subsection.

An instructive way of deriving PCG is to exploit its relationship to the minimization technique of the same name described in Sec. 7.5. To this end, assume that the basic iterative method (18) is *symmetrizable*. That is, there exists a real, nonsingular matrix W such that $S = WM^{-1}AW^{-1}$ is a real, symmetric, positive-definite (SPD) matrix. Consider the quadratic functional $F(y) = \frac{1}{2}y^{\mathrm{T}}Sy - y^{\mathrm{T}}\hat{b}$, where $\hat{b} = WM^{-1}b$ and $y \in \mathbb{R}^n$. It is easy to show that the unique minimum of $F(y)$ is the solution $\hat{x}$ of $S\hat{x} = \hat{b}$. It follows immediately from the relations for S and $\hat{b}$ given above that $x = W^{-1}\hat{x}$ is the solution of both the preconditioned system $M^{-1}Ax = M^{-1}b$ and the original system $Ax = b$. If we take $M = W = I$, then $S = W(M^{-1}A)W^{-1} = A$ and, if we assume that A is a real SPD matrix, PCG reduces to the standard (unpreconditioned) CG method.

It is easy to show that if both A and M are real SPD matrices, then $M^{-1}A$ is symmetrizable. Moreover, many important practical problems, such as the numerical solution of self-adjoint elliptic PDEs, give rise to matrices A that are real SPD. Furthermore, if A is a real SPD matrix, then the splitting matrix M associated with the RF, Jacobi, and SSOR [with $\omega \in (0,2)$] iterations is a real SPD matrix too, as is the M given by an incomplete Cholesky factorization, provided it exists. Hence, these basic iterative methods are symmetrizable in this case and so PCG can be used to accelerate their convergence. In contrast, the SOR iteration with the optimal ω is generally not symmetrizable. So PCG is not even applicable and more general Krylov-subspace methods normally do not accelerate its convergence.

We assume throughout the rest of this subsection that A and M are real SPD matrices, because this case arises most frequently in practice and also because it simplifies the discussion below. Using this assumption and applying several mathematical identities, we get the computationally effective variant of PCG shown in Table 6. Note that W does not appear explicitly in this algorithm. Also note that, if we choose $M = I$, then $\tilde{r}_k = r_k$ and the PCG method reduces to the unpreconditioned CG method for $Ax = b$.

Many other mathematically equivalent forms of PCG exist. Moreover, as noted above, a more general form of PCG can be used if $M^{-1}A$ is symmetrizable, without either A or M being a real SPD matrix. For a discussion of the points, see an advanced text such as Hageman and Young (1981).

Table 6. The preconditioned conjugate-gradient (PCG) method for solving $Ax = b$.

```
choose an initial guess $x_0$
compute $r_0 = b - Ax_0$
solve $M\tilde{r}_0 = r_0$
set $p_0 = \tilde{r}_0$
for $k = 0, 1, \ldots$ until convergence do
    $\alpha_k = r_k^T\tilde{r}_k / p_k^T A p_k$
    $x_{k+1} = x_k + \alpha_k p_k$
    $r_{k+1} = r_k - \alpha_k A p_k$
    solve $M\tilde{r}_{k+1} = r_{k+1}$
    $\beta_k = r_{k+1}^T\tilde{r}_{k+1} / r_k^T\tilde{r}_k$
    $p_{k+1} = \tilde{r}_{k+1} + \beta_k p_k$
end
```

Before considering the convergence of PCG, we introduce some notation. The *energy norm* of a vector $y \in \mathbb{R}^n$ with respect to a real SPD matrix B is $\|y\|_{B^{1/2}} = (y^T B y)^{1/2}$. The *Krylov subspace* of degree $k - 1$ generated by a vector v and a matrix W is $\mathcal{K}_k(v,W) = \text{span}\{v, Wv, \ldots, W^{k-1}v\}$.

It is easy to show that the r_k that occurs in Table 6 is the residual $r_k = b - Ax_k$ associated with x_k for the system $Ax = b$ and that $\tilde{r}_k = M^{-1}r_k = M^{-1}b - M^{-1}Ax_k$ is the residual associated with x_k for the preconditioned system $M^{-1}Ax = M^{-1}b$. Let $e_k = x - x_k$ be the error associated with x_k. It can be shown that the x_k generated by PCG is a member of the shifted Krylov subspace $x_0 + \mathcal{K}_k(\tilde{r}_0, M^{-1}A) \equiv \{x_0 + v : v \in \mathcal{K}_k(\tilde{r}_0, M^{-1}A)\}$. Hence, PCG is in the broader class of Krylov-subspace methods characterized by this property. Moreover, it can be shown that the x_k generated by PCG is the unique member of $x_0 + \mathcal{K}_k(\tilde{r}_0, M^{-1}A)$ that minimizes the energy norm of the error $\|e_k\|_{A^{1/2}} = (e_k^T A e_k)^{1/2}$ over all vectors of the form $e'_k = x - x'_k$, where x'_k is any other member of $x_0 + \mathcal{K}_k(\tilde{r}_0, M^{-1}A)$. Equivalently, $e_k = P_k^*(M^{-1}A)e_0$, where $P_k^*(z)$ is the polynomial that minimizes $\|P_k(M^{-1}A)e_0\|_{A^{1/2}}$ over all polynomials $P_k(z)$ of degree k that satisfy $P_k(0) = 1$. This result is the basis of the characterization that PCG is the optimal polynomial acceleration scheme for the basic iterative method (18).

In passing, note that the iterate x_k generated by the basic iteration (18) is in the shifted Krylov subspace $x_0 + \mathcal{K}_k(\tilde{r}_0, M^{-1}A)$ also and that, by (21), the associated error satisfies $e_k = (I - M^{-1}A)^k e_0$. So (18) is a Krylov-subspace method too and its error satisfies the polynomial relation described above with $P_k(z) = (1 - z)^k$. Thus, the associated PCG method is guaranteed to accelerate the convergence of the basic iteration (18)—at least when the errors are measured in the energy norm.

The characterization of its error discussed above can be very useful in understanding the performance of PCG. For example, it can be used to prove the *finite termination property* of PCG. That is, if $M^{-1}A$ has m distinct eigenvalues, then $x_m = x$, the exact solution of $Ax = b$. Since $M^{-1}A$ has n eigenvalues, $m \leq n$ always and $m \ll n$ sometimes. We caution the reader, though, that the argument used to prove this property of PCG assumes exact arithmetic. In floating-point arithmetic, we rarely get $e_m = 0$, although we frequently get $e_{\tilde{m}}$ small for some $\tilde{m}$ possibly a little larger than m.

The proof of the finite termination property can be extended easily to explain the rapid convergence of PCG when the eigenvalues of $M^{-1}A$ fall into a few small clusters. So a preconditioner M is good if the eigenvalues of $M^{-1}A$ are much more closely clustered than those of the unpreconditioned matrix A.

Because of the finite termination property, both CG and PCG can be considered direct methods. However, both are frequently used as iterative schemes, with the iteration terminated long before $e_m = 0$. Therefore, it is important to understand how the error decreases with k, the iteration count. To this end, first note that the characterization of the error ensures that $\|e_{k+1}\|_{A^{1/2}} \le \|e_k\|_{A^{1/2}}$ with equality only if $e_k = 0$. That is, PCG is a *descent* method in the sense that some norm of the error decreases on every iteration. Not all iterative methods enjoy this useful property.

The characterization of the error can also be used to show that

$$\|e_k\|_{A^{1/2}} \le 2\left(\frac{\sqrt{\lambda_n/\lambda_1} - 1}{\sqrt{\lambda_n/\lambda_1} + 1}\right)^k \|e_0\|_{A^{1/2}}, \tag{34}$$

where λ_n and λ_1 are the largest and smallest eigenvalues, respectively, of $M^{-1}A$. (Note that $\lambda_n, \lambda_1 \in \mathbb{R}$ and $\lambda_n \ge \lambda_1 > 0$ since $M^{-1}A$ is symmetrizable.) It follows easily from the definition of the energy norm and (34) that

$$\|e_k\|_2 \le 2\sqrt{\kappa_2(A)}\left(\frac{\sqrt{\lambda_n/\lambda_1} - 1}{\sqrt{\lambda_n/\lambda_1} + 1}\right)^k \|e_0\|_2, \tag{35}$$

where $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2$ is the condition number of A in the 2-norm (see Sec. 2.8). Although (35) is generally not as tight as (34), it may be more relevant to the practitioner. It follows from either (34) or (35) that, in this context, a preconditioner M is good if λ_n/λ_1 is (much) closer to 1 than is the ratio of the largest to smallest eigenvalues of A.

Unlike many other iterative methods, such as SOR, PCG does not require an estimate of any parameters, although some stopping procedures for PCG require an estimate of the extreme eigenvalues of A or $M^{-1}A$. See an advanced text such as Axelsson (1994), Golub and Van Loan (1989), Hageman and Young (1981) for details.

4. OVERDETERMINED AND UNDERDETERMINED LINEAR SYSTEMS

A linear system $Ax = b$, with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ given and $x \in \mathbb{R}^n$ unknown, is called *overdetermined* if $m > n$. Such a system typically has no solution. However, there is always an x that minimizes $\|b - Ax\|_2$. Such an x is called a *least-squares* solution to the overdetermined linear system $Ax = b$. Moreover, if rank$(A) = n$, then the least-squares solution is unique.

A linear system $Ax = b$, as above, is called *underdetermined* if $m < n$. Such a system typically has infinitely many solutions. If $Ax = b$ has a solution, then there is an x of minimum Euclidean norm, $\|x\|_2$, that satisfies $Ax = b$. Such an x is called a *least-squares* solution to the underdetermined linear system $Ax = b$. It can be shown that, if rank$(A) = m$, then the least-squares solution is unique.

In the following subsections, we describe methods to compute the least-squares solution of overdetermined and underdetermined linear systems, assuming that the matrix A has full rank, i.e., rank$(A) = \min(m,n)$. For the more general case of rank$(A) < \min(m,n)$, see an advanced text such as Golub and Van Loan (1989).

4.1 The Normal Equations for Overdetermined Linear Systems

Let $Ax = b$ be an overdetermined linear system with m equations in n unknowns and with rank$(A) = n$. The x that minimizes $\|b - Ax\|_2$ satisfies

$$\frac{\partial[(b - Ax)^{\mathrm{T}}(b - Ax)]}{\partial x_j} = 0 \quad \text{for } j = 1, \dots, n.$$

These relations can be rewritten in matrix form as $A^{\mathrm{T}}(b - Ax) = 0$ or equivalently

$$A^{\mathrm{T}}Ax = A^{\mathrm{T}}b, \tag{36}$$

which is called the system of *normal equations* for the overdetermined linear system $Ax = b$. It can be shown that the matrix A^TA is symmetric and positive-definite if rank(A) $= n$. So the system (36) has a unique solution, which is the least-squares solution to $Ax = b$.

Note that the matrix A^TA is $n \times n$. Computing the matrix A^TA requires about $mn^2/2$ flops (floating-point operations), while computing A^Tb requires about mn flops, and solving (36) requires about $n^3/6$ flops, if we assume that the Cholesky factorization (see Sec. 2.6) is used.

The condition number of the matrix A^TA is often large. More specifically, it can be shown that if A is square and nonsingular, then the condition number of A^TA is the square of the condition number of A (see Sec. 2.8). As a result, the approach described above of forming and solving the normal equations (36) often leads to a serious loss of accuracy. Therefore, in Secs. 4.4 and 4.7, we discuss more stable alternatives for solving least-squares problems.

4.2 The Normal Equations for Underdetermined Linear Systems

Let $Ax = b$ be an underdetermined linear system with m equations in n unknowns and with rank(A) $= m$. It can be shown that the least-squares solution x to $Ax = b$ can be written in the form A^Ty for some $y \in \mathbb{R}^m$ that satisfies

$$AA^Ty = b. \tag{37}$$

This is a linear system of size $m \times m$, called the system of *normal equations* for the underdetermined linear system $Ax = b$. It can be shown that the matrix AA^T is symmetric and positive-definite if rank(A) $= m$. So the system (37) has a unique solution. The unique least-squares solution to $Ax = b$ is $x = A^Ty$.

Computing the matrix AA^T requires about $nm^2/2$ flops, while computing A^Ty requires about mn flops, and solving (37) requires about $m^3/6$ flops, if the Cholesky factorization (see Sec. 2.6) is used.

As is the case for overdetermined systems, the method of forming and solving the normal equations (37) to compute the least-squares solution to $Ax = b$ is numerically unstable in some cases. In Secs. 4.5 and 4.8, we discuss more stable alternatives.

4.3 Householder Transformations and the *QR* Factorization

An *orthogonal transformation* is a linear change of variables that preserves the length of vectors in the Euclidean norm. Examples are a rotation about an axis or a reflection across a plane. The following is an example of an orthogonal transformation $y = (y_1,y_2)$ to $x = (x_1,x_2)$:

$$x_1 = 0.6y_1 + 0.8y_2,$$

$$x_2 = 0.8y_1 - 0.6y_2.$$

It is easy to see that $\|x\|_2 = \|y\|_2$.

An *orthogonal matrix* is an $m \times n$ matrix Q with the property $Q^TQ = I$. Note that, if (and only if) Q is square, i.e., $m = n$, the relation $Q^TQ = I$ is equivalent to $QQ^T = I$ and so $Q^T = Q^{-1}$. An orthogonal transformation of y to x can be written as $x = Qy$, where Q is an orthogonal matrix. In the above example, the corresponding orthogonal matrix is

$$Q = \begin{pmatrix} 0.6 & 0.8 \\ 0.8 & -0.6 \end{pmatrix}.$$

If Q is an orthogonal matrix, then

$$\|x\|_2^2 = \|Qy\|_2^2 = (Qy)^T(Qy) = y^T(Q^TQ)y$$
$$= y^Ty = \|y\|_2^2,$$

whence $\|x\|_2 = \|y\|_2$. This property is exploited in the methods described below to solve least-squares problems. Moreover, the numerical stability of these schemes is due at least in part to the related observation that orthogonal transformations do not magnify rounding error.

A *Householder transformation* or *Householder reflection* is an orthogonal matrix of the form $H = I - 2ww^T$, where $\|w\|_2 = 1$. Note that, when H is 2×2, the effect of H on a vector x is equivalent to reflecting the vector x across the plane perpendicular to w and passing through the origin of x. A Householder reflection can be used to transform a nonzero vector into one containing mainly zeros.

An $m \times n$ matrix $R = [r_{ij}]$ is *right trian-*

gular if $r_{ij} = 0$ for $i > j$. Note that if (and only if) R is square, i.e., $m = n$, the terms right triangular and upper triangular are equivalent.

Let A be an $m \times n$ matrix with $m \geq n$. The QR factorization of A expresses A as the product of an $m \times m$ orthogonal matrix Q and an $m \times n$ right-triangular matrix R. It can be computed by a sequence $H_1, H_2, \ldots, H_n$ of Householder transformations to reduce A to right-triangular form R. More specifically, this variant of the QR factorization proceeds in n steps (or $n - 1$ if $n = m$). Starting with $A_0 = A$, at step k for $k = 1, \ldots, n$, H_k is applied to the partially processed matrix A_{k-1} to zero the components $k + 1$ to m of column k of A_{k-1}. $Q = H_1H_2 \ldots H_n$ and $R = A_n$, the last matrix to be computed. For the details of the QR factorization algorithm using Householder transformations or other elementary orthogonal transformations, see Hager (1988) or Golub and Van Loan (1989).

4.4 Using the *QR* Factorization to Solve Overdetermined Linear Systems

Assume that A is an $m \times n$ matrix, with $m \geq n$ and $\text{rank}(A) = n$. Let $Ax = b$ be the linear system to be solved (in the least-squares sense, if $m > n$). Let $A = QR$ be the QR factorization of A. Note that

$$\begin{aligned}\|Ax - b\|_2 &= \|QRx - b\|_2 = \|Q(Rx - Q^{\mathrm{T}}b)\|_2 \\ &= \|Rx - Q^{\mathrm{T}}b\|_2.\end{aligned}$$

Therefore, we can use the QR factorization of A to reduce the problem of solving $Ax = b$ to that of solving $Rx = Q^{\mathrm{T}}b$, a much simpler task. We solve the latter by first computing $y = Q^{\mathrm{T}}b$. Let $\hat{y}$ be the vector consisting of the first n components of y and $\hat{R}$ the upper-triangular matrix consisting of the first n rows of R. Now solve $\hat{R}x = \hat{y}$ by back substitution (see Sec. 2.2). Then x is the least-squares solution to both $Rx = Q^{\mathrm{T}}b$ and $Ax = b$.

It can be shown that the QR factorization algorithm applied to an $n \times n$ linear system requires about $2n^3/3$ flops, which is about twice as many as the LU factorization needs. However, QR is a more stable method than LU and it requires no pivoting. The QR factorization algorithm applied to an $m \times n$ linear system requires about twice as many flops as forming and solving the normal equations. However, QR is a more stable method than solving the normal equations.

4.5 Using the *QR* Factorization to Solve Underdetermined Linear Systems

Assume that A is an $m \times n$ matrix, with $m < n$ and $\text{rank}(A) = m$. Let $Ax = b$ be the linear system to be solved (in the least-squares sense). Obtain the QR factorization of A^{T} by the QR factorization algorithm: $A^{\mathrm{T}} = QR$, where Q is an $n \times n$ orthogonal matrix and R is an $n \times m$ right-triangular matrix. Let $\hat{R}$ be the upper-triangular matrix consisting of the first m rows of R. Solve $\hat{R}^{\mathrm{T}}\hat{y} = b$ by back substitution (see Sec. 2.2) and let $y = (\hat{y}^{\mathrm{T}}, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^n$. Note that y is the vector of minimal Euclidean norm that satisfies $R^{\mathrm{T}}y = b$. Finally compute $x = Qy$ and note that x is the vector of minimal Euclidean norm that satisfies $R^{\mathrm{T}}Q^{\mathrm{T}}x = b$ or equivalently $Ax = b$. That is, x is the least-squares solution to $Ax = b$.

4.6 The Gram–Schmidt Orthogonalization Algorithm

The *Gram–Schmidt orthogonalization* algorithm is an alternative to QR. The modified version of the algorithm presented below is approximately twice as fast as QR and more stable than solving the normal equations.

Assume that $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ and that A has n linearly independent columns $\{a_j\}_{j=1}^{n}$. The Gram–Schmidt algorithm applied to A generates an $m \times n$ orthogonal matrix Q and an $n \times n$ upper-triangular matrix R satisfying $A = QR$. In Table 7, we present a stable version of the algorithm, often called the *modified Gram–Schmidt algorithm*.

Table 7. The modified Gram–Schmidt algorithm.

```
for $j = 1, \ldots, n$ do                      ($q_j$ and $a_j$ are
    $q_j = a_j$                                    columns $j$ of $Q$
    for $i = 1, \ldots, j - 1$ do                  and $A$,
        $r_{ij} = q_i^{\mathrm{T}} q_j$                          respectively)
        $q_j = q_j - r_{ij}q_i$
    end
    $r_{jj} = \|q_j\|_2$
    $q_j = q_j/r_{jj}$
end
```

4.7 Using Gram–Schmidt to Solve Overdetermined Linear Systems

Assume that A is an $m \times n$ matrix, with $m \geq n$ and $\operatorname{rank}(A) = n$. Let $Ax = b$ be the linear system to be solved (in the least-squares sense, if nonsquare). The method for solving $Ax = b$ using the Gram–Schmidt algorithm is similar to that described in Sec. 4.4 for the QR factorization.

First compute the QR factorization of A by the Gram–Schmidt algorithm. Next compute $y = Q^T b$ and solve $Rx = y$ by back substitution (see Sec. 2.2). Then x is the least-squares solution to $Ax = b$.

It can be shown that the Gram–Schmidt algorithm applied to an $n \times n$ linear system requires about $n^3/3$ flops, which is about the same number of arithmetic operations as the LU factorization algorithm needs and about half of what the QR factorization algorithm requires. Moreover, the modified Gram–Schmidt algorithm, as presented in Table 7, is relatively stable. The Gram–Schmidt algorithm applied to an $m \times n$ linear system requires about the same number of flops as that needed to form and to solve the normal equations, but the Gram–Schmidt algorithm is more stable.

4.8 Using Gram–Schmidt to Solve Underdetermined Linear Systems

Assume that A is an $m \times n$ matrix, with $m < n$ and $\operatorname{rank}(A) = m$. Let $Ax = b$ be the linear system to be solved (in the least-squares sense). The method for solving $Ax = b$ using the Gram–Schmidt algorithm is similar to that described in Sec. 4.5 for the QR factorization.

First compute the QR factorization of A^T by the Gram–Schmidt algorithm. Next solve $R^T y = b$ by back substitution (see Sec. 2.2) and set $x = Qy$. Then x is the least-squares solution to $Ax = b$.

5. EIGENVALUES AND EIGENVECTORS OF MATRICES

Given an $n \times n$ matrix A, $\lambda \in \mathbb{C}$, and $x \in \mathbb{C}^n$ satisfying $Ax = \lambda x$, λ is called an *eigenvalue* of A and x is called an *eigenvector* of A. The relation $Ax = \lambda x$ can be rewritten as $(A - \lambda I)x = 0$, emphasizing that λ is an eigenvalue of A if and only if $A - \lambda I$ is singular and that an eigenvector x is a nontrivial solution to $(A - \lambda I)x = 0$. It also follows that the eigenvalues of A are the roots of $\det(A - \lambda I) = 0$, the *characteristic equation* of A. The polynomial $p(\lambda) = \det(A - \lambda I)$ of degree n is called the *characteristic polynomial* of A and plays an important role in the theory of eigenvalues.

An $n \times n$ matrix A has precisely n eigenvalues, not necessarily distinct. It also has at least one eigenvector for each distinct eigenvalue. Note also that, if x is an eigenvector of A, then so is any (scalar) multiple of x and the corresponding eigenvalue is the same. We often choose an eigenvector of norm one in some vector norm, often the Euclidean norm, as the representative.

Two matrices A and B are similar if $A = W^{-1}BW$ for some nonsingular matrix W. The matrix W is often referred to as a *similarity transformation*. It is easy to see that, if $Ax = \lambda x$, then $B(Wx) = \lambda(Wx)$. Thus, similar matrices have the same eigenvalues and their eigenvectors are related by the similarity transformation W. Similarity transformations are often used in numerical methods for eigenvalues and eigenvectors to transform a matrix A into another one B that has the same eigenvalues and related eigenvectors, but whose eigenvalues and eigenvectors are in some sense easier to compute than those of A.

Eigenvalues play a major role in the study of convergence of iterative methods (see Sec. 3). Eigenvalues and eigenvectors are also of great importance in understanding the stability and other fundamental properties of many physical systems.

The matrix A is often large, sparse, and symmetric. These properties can be exploited to great advantage in numerical schemes for calculating the eigenvalues and eigenvectors of A.

A common approach to calculate the eigenvalues (and possibly the eigenvectors) of a matrix A consists of two stages. First, the matrix A is transformed to a similar but simpler matrix B, usually tridiagonal, if A is symmetric (or Hermitian), or Hessenberg, if A is nonsymmetric (or non-Hermitian). Then, the eigenvalues (and possibly the eigenvectors) of B are calculated. An exception to this approach is the power method (see Sec. 5.1).

A standard procedure for computing the eigenvectors of a matrix A is to calculate the eigenvalues first then use them to compute the eigenvectors by inverse iteration (see Sec. 5.4). Again, an exception to this approach is the power method (see Sec. 5.1).

Before describing numerical methods for the eigenvalue problem, we comment briefly on the sensitivity of the eigenvalues and eigenvectors to perturbations in the matrix A, since a backward error analysis can often show that the computed eigenvalues and eigenvectors are the exact eigenvalues and eigenvectors of a slightly perturbed matrix $\hat{A} = A + E$, where E is usually small relative to A. In general, if A is symmetric, its eigenvalues are well-conditioned with respect to small perturbations E. That is, the eigenvalues of $\hat{A}$ and A are very close. This, though, is not always true of the eigenvectors of A, particularly if the associated eigenvalue is a multiple eigenvalue or close to another eigenvalue of A. If A is nonsymmetric, then both its eigenvalues and eigenvectors may be poorly conditioned with respect to small perturbations E. Therefore, the user of a computer package for calculating the eigenvalues of a matrix should be cautious about the accuracy of the numerical results. For a further discussion of the conditioning of the eigenvalue problem, see Wilkinson (1965).

5.1 The Power Method

The power method is used to calculate the eigenvalue of largest magnitude of a matrix A and the associated eigenvector. Since matrix–vector products are the dominant computational work required by the power method, this scheme can exploit the sparsity of the matrix to great advantage.

Let λ be the eigenvalue of A of largest magnitude and let x be an associated eigenvector. Also let z_0 be an initial guess for some multiple of x. The power method, shown in Table 8, is an iterative scheme that generates a sequence of approximations $z_1, z_2, \ldots$ to some multiple of x and another sequence of approximations $\mu_1, \mu_2, \ldots$ to λ. In the scheme shown in Table 8, z_k is normalized so that the sequence $z_1, z_2, \ldots$ converges to an eigenvector x of A satisfying $\|x\|_\infty = 1$. Normalizations of this sort are frequently used in eigenvector calculations.

Table 8. The power method.

```
Pick $z_0$
for $k = 1, 2, \ldots$ do
    $w_k = Az_{k-1}$
    Choose $m \in \{1, \ldots, n\}$ such that
        $|(w_k)_m| \geq |(w_k)_i|$ for $i = 1, \ldots, n$
    $z_k = w_k/(w_k)_m$
    $\mu_k = (w_k)_m/(z_{k-1})_m$
    test stopping criterion
end
```

The power method is guaranteed to converge if A has a single eigenvalue λ of largest magnitude. The rate of convergence depends on $|\lambda_2|/|\lambda|$, where λ_2 is the eigenvalue of A of next largest magnitude. With some modifications the power method can be used when A has more than one eigenvalue of largest magnitude.

After the absolutely largest eigenvalue of A has been calculated, the power method can be applied to an appropriately deflated matrix to calculate the next largest eigenvalue and the associated eigenvector of A, and so on. However, this approach is inefficient if all or many eigenvalues are needed. The next sections describe more general-purpose methods for eigenvalue and eigenvector computations. For an introduction to the power method, see Atkinson (1989). For further reading, see Golub and Van Loan (1989) or Wilkinson (1965).

5.2 The *QR* Method

The *QR* method is widely used to calculate all the eigenvalues of a matrix A. It employs the *QR* factorization algorithm presented briefly in Sec. 4.4. Here, we recall that, given an $n \times n$ matrix A, there is a factorization $A = QR$, where R is an $n \times n$ upper-triangular matrix and Q an $n \times n$ orthogonal matrix.

The *QR* method, shown in Table 9, is an iterative scheme to compute the eigenvalues of A. It proceeds by generating a sequence

Table 9. The *QR* method.

```
Set $A_0 = A$
for $k = 1, 2, \ldots$ do
    Compute the QR factorization of $A_{k-1} = Q_kR_k$
    Set $A_k = R_kQ_k$
    test stopping criterion
end
```

$A_1, A_2, \ldots$ of matrices, all of which are similar to each other and to the starting matrix $A_0 = A$. The sequence converges either to a triangular matrix with the eigenvalues of A on its diagonal or to an almost triangular matrix from which the eigenvalues can be calculated easily.

For a real, nonsingular matrix A with no two (or more) eigenvalues of the same magnitude, the *QR* method is guaranteed to converge. The iterates A_k converge to an upper-triangular matrix with the eigenvalues of A on its diagonal. If A is symmetric, the iterates converge to a diagonal matrix. The rate of convergence depends on $\max\{|\lambda_{i+1}/\lambda_i| : i = 1, \ldots, n-1\}$, where $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$.

If two or more eigenvalues of A have the same magnitude, the sequence $A_1, A_2, \ldots$ may not converge to an upper-triangular matrix. If A is symmetric, the sequence converges to a block-diagonal matrix, with blocks of order 1 or 2, from which the eigenvalues of A can be calculated easily. If A is nonsymmetric, the problem is more complicated. See Wilkinson (1965) or Parlett (1968) for details.

In many cases, the *QR* method is combined with a technique known as *shifting* to accelerate convergence. A discussion of this procedure can be found in Atkinson (1989), Golub and Van Loan (1989), Parlett (1980), and Wilkinson (1965).

The *QR* method for eigenvalues requires the computation of the *QR* factorization of a matrix and a matrix–matrix multiplication at each iteration. For a large matrix A, this is an expensive computation, making the method inefficient. To improve its efficiency, the matrix A is normally preprocessed, reducing it to a simpler form. If A is symmetric, it is reduced to a similar tridiagonal matrix, as described in Sec. 5.3. If A is nonsymmetric, it is reduced to a similar upper Hessenberg matrix by a comparable algorithm. It can be shown that, if $A_0 = A$ is in Hessenberg or tridiagonal form, then all the iterates A_k generated by the *QR* method will also be in Hessenberg or tridiagonal form, respectively. With appropriate implementation techniques, the *QR* factorization and the matrix–matrix products applied to matrices with these special structures require fewer operations than would be needed for arbitrary dense matrices.

With the modifications discussed above, the *QR* method is an efficient, general-purpose scheme for calculating the eigenvalues of a dense matrix. The eigenvectors can also be calculated if all the similarity transformations employed in the *QR* process are stored. Alternatively, the eigenvectors can be calculated by inverse iteration, as described in Sec. 5.4.

5.3 Transforming a Symmetric Matrix to Tridiagonal Form

As noted above, the eigenvalues of a symmetric (or Hermitian) matrix A are often calculated by first transforming A into a similar tridiagonal matrix T. Householder transformations (see Sec. 4.3) are usually employed to perform this task.

The algorithm to obtain the matrix T, given A, resembles the *QR* factorization algorithm described briefly in Sec. 4.3. It proceeds in $n - 2$ steps and generates a sequence $H_1, H_2, \ldots, H_{n-2}$ of Householder transformations to reduce A to tridiagonal form T. Starting with $A^{(0)} = A$, at step k for $k = 1, \ldots, n-2$, we form $A_k = H_k A_{k-1} H_k$, where H_k is chosen to zero the components $k + 2$ to n of both row k and column k of A_{k-1}. T is A_{n-2}, the last matrix computed by this process. Note that $T = H_{n-2} \cdots H_1 A H_1 \cdots H_{n-2}$ is similar to A because each H_k is symmetric and orthogonal.

It can be shown that the reduction to tridiagonal form by Householder transformations is a stable computation in the sense that the computed T is the exact T for a slightly perturbed matrix $\hat{A} = A + E$, where E is usually small relative to A. As a result, it can be shown that the eigenvalues of A and T differ very little. For a brief introduction to tridiagonal reduction, see Atkinson (1989). For further reading, see Golub and Van Loan (1989) or Wilkinson (1965). For other methods to reduce A to tridiagonal form, such as planar rotation orthogonal matrices, see Golub and Van Loan (1989).

Similar schemes can be used to reduce a nonsymmetric (or non-Hermitian) matrix to Hessenberg form.

5.4 Inverse Iteration

Inverse iteration is the standard method to calculate the eigenvectors of a matrix A, once its eigenvalues have been calculated.

This scheme can be viewed as the power method (see Sec. 5.1) applied to the matrix $(A - \tilde{\lambda}I)^{-1}$ instead of A, where $\tilde{\lambda}$ is an (approximate) eigenvalue of A.

To see how inverse iteration works, let $\tilde{\lambda}$ be an approximation to a simple eigenvalue λ of A and let x be the associated eigenvector. Also let z_0 be an initial guess to some multiple of x. Inverse iteration, shown in Table 10, is an iterative method that generates a sequence of approximations $z_1, z_2, \ldots$ to some multiple of x. In the scheme shown in Table 10, z_k is normalized so that the sequence $z_1, z_2, \ldots$ converges to an eigenvector x of A satisfying $\|x\|_\infty = 1$. As noted already in Sec. 5.1, normalizations of this sort are frequently used in eigenvector calculations.

Note that, if $\tilde{\lambda} = \lambda$, then $A - \tilde{\lambda}I$ is singular. Moreover, if $\tilde{\lambda} \approx \lambda$, then $A - \tilde{\lambda}I$ is "nearly" singular. As a result, we can expect the system $(A - \tilde{\lambda}I)w_k = z_{k-1}$ to be very poorly conditioned (see Sec. 2.8). This, though, is not a problem in this context, since any large perturbation in the solution of the ill-conditioned system is in the direction of the desired eigenvector x.

The sequence of approximate eigenvectors $z_1, z_2, \ldots$ is typically calculated by first performing an LU decomposition of $A - \tilde{\lambda}I$, often with pivoting, before the start of the iteration and then using the same LU factorization to perform a forward elimination followed by a back substitution to solve $(A - \tilde{\lambda}I)w_k = z_{k-1}$ for $k = 1, 2, \ldots$ (see Sec. 2). We emphasize that only one LU factorization of $A - \tilde{\lambda}I$ is needed to solve all $(A - \tilde{\lambda}I)w_k = z_{k-1}, k = 1, 2, \ldots$.

For a brief discussion of the inverse iteration, including its stability and rate of convergence, see Atkinson (1989). For further reading, see Wilkinson (1965).

5.5 Other Methods

Another way to calculate the eigenvalues of A is to compute the roots of the characteristic polynomial $p(\lambda) = \det(A - \lambda I)$. The techniques discussed in Sec. 6.9 can be used for this purpose. However, this approach is often less stable than the techniques described above and it is usually not more efficient. Therefore, it is not normally recommended.

Table 10. Inverse iteration.

```
Pick $z_0$
for $k = 1,2,\ldots$ do
    Solve $(A - \tilde{\lambda}I)w_k = z_{k-1}$
    $z_k = w_k/\|w_k\|_\infty$
    test stopping criterion
end
```

An obvious method for calculating an eigenvector x, once the corresponding eigenvalue λ is known, is to solve the system $(A - \lambda I)x = 0$. Since this system is singular, one approach is to delete one equation from $(A - \lambda I)x = 0$ and replace it by another linear constraint, such as $x_j = 1$ for some component j of x. However, it can be shown [see Wilkinson (1965)] that this method is not always stable and can lead to very poor numerical results in some cases. Thus, it is not recommended either.

Several other methods for calculating the eigenvalues and eigenvectors of a matrix have been omitted from our discussion because of space limitations. We should, though, at least mention one of them, the Jacobi method, a simple, rapidly convergent iterative scheme, applicable to symmetric matrices, including sparse ones.

The eigenvalue problem for large sparse matrices is a very active area of research. Although the QR method described in Sec. 5.2 is effective for small- to medium-sized dense matrices, it is computationally very expensive for large matrices, in part because it does not exploit sparsity effectively. The Lanczos and Arnoldi methods are much better suited for large sparse problems. For a discussion of these methods, see Cullum and Willoughby (1985), Parlett (1980), Saad (1992), and Scott (1981).

6. NONLINEAR ALGEBRAIC EQUATIONS AND SYSTEMS

Consider a nonlinear algebraic equation or system $f(x) = 0$, where $f: \mathbb{R}^n \to \mathbb{R}^n$, $n \geq 1$. A *root* of $f(x)$ is a value $\alpha \in \mathbb{R}^n$ satisfying $f(\alpha) = 0$. A nonlinear function may have one, many, or no roots.

Most numerical methods for computing approximations to roots of a nonlinear equation or system are *iterative* in nature. That is, the scheme starts with some initial guess x_0 and computes new successive approximations x_k, $k = 1, 2, \ldots$, by some formula until a *stopping criterion* such as

$$\|f(x_k)\| \le \epsilon,$$

$$\|f(x_k)\| \le \epsilon\|f(x_0)\|,$$

$$\|x_k - x_{k-1}\| \le \epsilon,$$

$$\|x_k - x_{k-1}\| \le \epsilon\|x_k\|, \quad \text{or}$$

$$k > \text{maxit}$$

is satisfied, where ϵ is the error tolerance and maxit is the maximum number of iterations allowed.

6.1 Fixed-Point Iteration

A *fixed point* of a function $g(x)$ is a value α satisfying $\alpha = g(\alpha)$. A *fixed-point iteration* is a scheme of the form $x_k = g(x_{k-1})$ that uses the most recent approximation x_{k-1} to the fixed point α to compute a new approximation x_k to α. In this context, the function g is also called the *iteration function.*

One reason for studying fixed-point iterations is that given a function $f(x)$, it is easy to find another function $g(x)$ such that α is a root of $f(x)$ if and only if it is a fixed point of $g(x)$. For example, take $g(x) = x - f(x)$. Many root-finding methods can be viewed as fixed-point iterations.

Given an iteration function g, a fixed-point scheme starts with an initial guess x_0 and proceeds with the iteration as follows:

```
for $k = 1, 2, \ldots$ do
    $x_k = g(x_{k-1})$
    test stopping criterion
end
```

6.2 Newton's Method for Nonlinear Equations

We consider scalar equations (i.e., $n = 1$) first, and extend the results to systems of equations (i.e., $n > 1$) in Sec. 6.7.

Newton's method is a fixed-point iteration based on the iteration function $g(x) = x - f(x)/f'(x)$, where $f'(x)$ is the first derivative of f. More specifically, the new approximation x_k to the root α of f is computed by the formula

$$x_k = x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})},$$

which uses the previous approximation x_{k-1} to α. Geometrically, Newton's method approximates the nonlinear function $f(x)$ by its tangent (a straight line) at the current approximation x_{k-1} and takes x_k to be the intersection of the tangent with the x axis. That is, x_k is the root of the local linear approximation to $f(x)$. Newton's method is applicable if and only if f is differentiable and f' is nonzero at the point of approximation.

6.3 The Secant Method

The secant method is applicable to scalar equations only and is not a fixed-point iteration. The new approximation x_k to the root α of f is computed using two previous approximations x_{k-1} and x_{k-2} by the formula

$$x_k = x_{k-1} - f(x_{k-1})\frac{x_{k-1} - x_{k-2}}{f(x_{k-1}) - f(x_{k-2})}.$$

The secant method can be viewed as a variant of Newton's method in which $f'(x_{k-1})$ is approximated by $[f(x_{k-1}) - f(x_{k-2})]/[x_{k-1} - x_{k-2}]$. Geometrically, the secant method approximates the nonlinear function $f(x)$ by the chord subtending the graph of f at the two most recently computed points of approximation x_{k-1} and x_{k-2} and takes x_k to be the intersection of the chord with the x axis. That is, x_k is the root of the local linear approximation to $f(x)$. The secant method is applicable if and only if $f(x)$ is continuous and takes different values at x_{k-1} and x_{k-2}. To start, the secant method requires initial guesses for x_0 and x_1. These are usually chosen close to each other and must satisfy $f(x_0) \neq f(x_1)$.

6.4 The Bisection and *Regula Falsi* Methods

The bisection method is not a fixed-point iteration. It is applicable to a scalar equation $f(x) = 0$ if and only if $f(x)$ is continuous and there are two points L and R for which $f(L)f(R) \le 0$. These conditions guarantee the existence of at least one root of $f(x)$ in the interval $[L,R]$. Without loss of generality, let $L < R$. To start, the bisection method approximates the root by the midpoint $M = (L + R)/2$ of $[L,R]$ and halves the interval at each iteration as follows.

```
forever do
    $M = (L + R)/2$
    if $f(L)f(M) \le 0$ then $R = M$
        else $L = M$
    test stopping criterion
end
```

Note that this iteration maintains the property $f(L)f(R) \le 0$, as L and R are changed. So, when the algorithm terminates, a root of f is guaranteed to be in $[L,R]$. $M = (L + R)/2$ is often taken as the approximation to the root.

Several root-finding methods are similar to bisection. For example, *regula falsi* chooses

$$M = \frac{f(R)L - f(L)R}{f(R) - f(L)} \tag{38}$$

but is otherwise the same as bisection. Note that the M computed from (38) is the intersection of the chord subtending the graph of $f(x)$ at L and R with the x axis and so is guaranteed to lie in $[L,R]$, since the property $f(L)f(R) \le 0$ is maintained throughout the iteration even though L and R are changed.

6.5 Convergence

Iterative methods for nonlinear equations can be guaranteed to converge under certain conditions, although they may diverge in some cases.

The bisection method converges whenever it is applicable, but if $f(x)$ has more than one root in the interval of application, there is no guarantee which of the roots the method will converge to.

The convergence of a fixed-point iteration depends critically on the properties of the iteration function $g(x)$. If g is smooth in an interval containing a fixed point α and $|g'(\alpha)| < 1$, then there is an $m \in [0,1)$ and a neighborhood I around the fixed point α in which $|g'(x)| \le m < 1$. In this case, the fixed-point iteration $x_k = g(x_{k-1})$ converges to α if $x_0 \in I$. To give an intuitive understanding why this is so, we assume that $x_0, \ldots, x_{k-1} \in I$ and note that

$$x_k - \alpha = g(x_{k-1}) - g(\alpha) = g'(\xi_k)(x_{k-1} - \alpha),$$

where we have used $x_k = g(x_{k-1})$, $\alpha = g(\alpha)$ and, by the mean-value theorem, ξ_k is some point in I between x_{k-1} and α. Thus, $|g'(\xi_k)| \le m < 1$ and so $|x_k - \alpha| \le m|x_{k-1} - \alpha| \le m^k|x_0 - \alpha|$, whence $x_k \in I$ too and $x_k \to \alpha$ as $k \to \infty$.

A more formal statement of this theorem and other similar results giving sufficient conditions for the convergence of fixed-point iterations can be found in many introductory numerical methods textbooks. See for example Conte and de Boor (1980); Dahlquist and Björck (1974); Johnson and Riess (1982); Stoer and Bulirsch (1980).

Newton's method converges if the conditions for convergence of a fixed-point iteration are met. [For Newton's method, the iteration function is $g(x) = x - f(x)/f'(x)$.] However, it can be shown that Newton's method converges quadratically (see Sec. 6.6) to the root α of f if the initial guess x_0 is chosen sufficiently close to α, f is smooth, and $f'(x) \ne 0$ close to α. It can also be shown that Newton's method converges from any starting guess in some cases. A more formal statement of these and other similar results can be found in many introductory numerical methods textbooks. See for example Conte and de Boor (1980); Dahlquist and Björck (1974); Johnson and Riess (1982); Stoer and Bulirsch (1980). For a deeper discussion of this topic, see Dennis and Schnabel (1983).

6.6 Rate of Convergence

The rate of convergence of a sequence $x_1, x_2, \ldots$ to α is the largest number $p \ge 1$ satisfying

$$\|x_{k+1} - \alpha\| \le C\|x_k - \alpha\|^p \quad \text{as } k \to \infty$$

for some constant $C > 0$. If $p = 1$, we also require that $C < 1$. The larger the value of p the faster the convergence, at least asymptotically. Between two converging sequences with the same rate p, the faster is the one with the smaller C.

A fixed-point iteration with iteration function g converges at a rate p with $C = |g^{(p)}(\alpha)|/p!$ if $g \in \mathcal{C}^p$, $g^{(i)}(\alpha) = 0$, for $i = 0, 1, 2, \ldots, p - 1$, and $g^{(p)}(\alpha) \ne 0$, where $g^{(i)}(x)$ is the ith derivative of $g(x)$. Thus, Newton's method usually converges quadratically, i.e., $p = 2$ with $C = |g''(\alpha)|/2$, where $g(x) = x - f(x)/f'(x)$. If $f'(\alpha) = 0$, Newton's method typically converges linearly. If $f'(\alpha) \neq 0$ and $f''(\alpha) = 0$, Newton's method converges at least cubically, i.e., $p \geq 3$. The secant method converges at a superlinear rate of $p = (1 + \sqrt{5})/2 \approx 1.618$, i.e., faster than linear but slower than quadratic. Bisection converges linearly, i.e., $p = 1$ and $C = 1/2$.

6.7 Newton's Method for Systems of Nonlinear Equations

Newton's method for a scalar nonlinear equation (see Sec. 6.2) can be extended to a system of nonlinear equations with $f : \mathbb{R}^n \to \mathbb{R}^n$, for $n > 1$. In this case, the new iterate x_k is computed by

$$x_k = x_{k-1} - [J(x_{k-1})]^{-1}f(x_{k-1}),$$

where $J(x)$ is the *Jacobian* of f, an $n \times n$ matrix with its (i,j) entry equal to $\partial f_i(x)/\partial x_j$. To implement this scheme effectively, the matrix–vector product $z = [J(x_{k-1})]^{-1}f(x_{k-1})$ should *not* be calculated by first computing the inverse of the Jacobian and then performing the matrix–vector multiplication, but rather by first solving the linear system $[J(x_{k-1})]z_k = f(x_{k-1})$ by Gaussian elimination (see Sec. 2.1) or some other effective method for solving linear equations (see Secs. 2 and 3) and then setting $x_k = x_{k-1} - z_k$.

6.8 Modifications and Alternatives to Newton's Method

Solving the linear system $[J(x_{k-1})]z_k = f(x_{k-1})$ requires first evaluating all the partial derivatives of all components of f at the point x_{k-1} and then solving the linear system by performing Gaussian elimination (see Sec. 2.1) or some other effective method for solving linear equations (see Secs. 2 and 3). In the unmodified Newton's method, this procedure is repeated at every iteration, requiring $O(n^3)$ flops (floating-point operations) if Gaussian elimination is used. There exist variants of Newton's method that reduce the computational work per iteration significantly. Even though these schemes typically converge more slowly, they often dramatically reduce the cost of solving a nonlinear system, particularly if $n \gg 1$.

The *chord Newton* method, often called the *simplified Newton method*, holds $J(x_{k-1})$ fixed for several steps, thus avoiding many Jacobian evaluations and *LU* factorizations. However, it still requires one f evaluation and both a forward elimination and a back substitution (see Secs. 2.4 and 2.2, respectively) at each iteration.

Some other variants approximate the Jacobian by matrices that are easier to compute and simpler to solve. For example, the Jacobian may be approximated by its diagonal, giving rise to a *Jacobi-like Newton's* iteration, or by its lower-triangular part, giving rise to a *Gauss–Seidel-like Newton's* scheme. (See Sec. 3 for a discussion of Jacobi and Gauss–Seidel iterations for linear equations.)

Quasi-Newton schemes, which avoid the computation of partial derivatives, are alternatives to Newton's method. Some quasi-Newton schemes approximate partial derivatives by finite differences. For example,

$$[J(x)]_{ij} = \frac{\partial f_i}{\partial x_j}(x) \approx \frac{f_i(x + \delta e_j) - f_i(x)}{\delta},$$

where δ is a small nonzero number and $e_j \in \mathbb{R}^n$ is the jth unit vector, with a 1 in component j and 0's in all other entries.

Possibly the best-known quasi-Newton scheme is *Broyden's* method. It does not require the computation of any partial derivatives nor the solution of any linear systems. Rather, it uses one evaluation only of f and a matrix–vector multiply, requiring $O(n^2)$ flops, per iteration. Starting with an initial guess for the inverse of the Jacobian, $J(x_0)$, it updates its approximation to the inverse Jacobian at every iteration. Broyden's method can be viewed as an extension of the secant method to $n > 1$ dimensions. For a brief description of the algorithm, see Hager (1988).

6.9 Polynomial Equations

Polynomial equations are a special case of nonlinear equations. A polynomial of degree k has exactly k roots, counting multiplicity. One may wish to compute all roots of a polynomial or only a few select ones. The methods for nonlinear equations described above can be used in either case, although more effective schemes exist for this special class of problems. The efficient evaluation of the polynomial and its derivative is discussed below in Sec. 6.10.

Deflation is often used to compute roots of polynomials. It starts by locating one root r_1 of $p(x)$ and then proceeds recursively to compute the roots of $\hat{p}(x)$, where $p(x) = (x - r_1)\hat{p}(x)$. Note that, by the fundamental theorem of algebra, given a root r_1 of $p(x)$, $\hat{p}(x)$ is uniquely defined and is a polynomial of degree $k - 1$. However, deflation may be unstable unless implemented carefully. See an introductory numerical methods textbook for details.

Localization techniques can be used to identify regions of the complex plane that contain zeros. Such techniques include the Lehmer–Schur method, Laguerre's method, and methods based on Sturm sequences [see Householder (1970)]. Localization can be very helpful, for example, when searching for a particular root of a polynomial or when implementing deflation, since in the latter case, the roots should be computed in increasing order of magnitude to ensure numerical stability.

The roots of a polynomial $p(x)$ can also be found by first forming the *companion matrix* A of the polynomial $p(x)$—the eigenvalues $\{\lambda_i : i = 1, \ldots, n\}$ of A are the roots of $p(x)$—and then finding all, or a select few, of the eigenvalues of A. See Sec. 5 for a discussion of the computation of eigenvalues and an introductory numerical methods book, such as Hager (1988), for a further discussion of this root-finding technique.

6.10 Horner's Rule

Horner's rule, also called *nested multiplication* or *synthetic division*, is an efficient method to evaluate a polynomial and its derivative. Let $p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ be the polynomial of interest, and α the point of evaluation. The following algorithm computes $p(\alpha)$ in z_0 and $p'(\alpha)$ in y_1 efficiently.

```
$z_n = a_n$
$y_n = a_n$
for $j = n - 1, \ldots, 1$ do
    $z_j = \alpha z_{j+1} + a_j$
    $y_j = \alpha y_{j+1} + z_j$
end
$z_0 = z_1\alpha + a_0$
```

7. UNCONSTRAINED OPTIMIZATION

The *optimization problem* is to find a value $x^* \in \mathbb{R}^n$ that either *minimizes* or *maximizes* a function $f : \mathbb{R}^n \to \mathbb{R}$. We consider only the minimization problem here, since maximizing $f(x)$ is equivalent to minimizing $-f(x)$.

Sometimes the minimizer must satisfy constraints such as $g_i(x) = 0$, $i = 1, \ldots, m_1$, or $h_j(x) \geq 0$, $j = 1, \ldots, m_2$, where g_i and h_j: $\mathbb{R}^n \to \mathbb{R}$. Thus, the general minimization problem can be written as

$$\min_{x \in \mathbb{R}^n} f(x)$$

subject to

$$g_i(x) = 0, \quad i = 1, \ldots, m_1$$

$$h_j(x) \geq 0, \quad j = 1, \ldots, m_2.$$

If any of the functions f, g_i, or h_j are nonlinear, then the minimization problem is *nonlinear;* otherwise, it is called a *linear programming* problem. If there are no constraints, the minimization problem is called *unconstrained;* otherwise, it is called *constrained.*

In this section, we present numerical methods for nonlinear unconstrained minimization problems (NLUMPs) only. For a more detailed discussion of these schemes, see an advanced text such as Dennis and Schnabel (1983).

7.1 Some Definitions and Properties

Let $x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}} \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$. The *gradient* ∇f of the function f is the vector of the n first partial derivatives of f:

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right)^{\mathrm{T}}.$$

Note $\nabla f(x) : \mathbb{R}^n \to \mathbb{R}^n$.

A point x^* is a *critical point* of f if $\nabla f(x^*) = 0$. A point x^* is a *global minimum* of f if $f(x^*) \leq f(x)$ for all $x \in \mathbb{R}^n$. A point x^* is a *local minimum* of f if $f(x^*) \leq f(x)$ for all x in a neighborhood of x^*. If x^* is a local minimum of f, then it is also a critical point of f, assuming $\nabla f(x^*)$ exists, but the converse is not necessarily true.

The *Hessian* $\nabla^2 f(x) = H(x)$ of f is an $n \times n$ matrix with entries

$$H_{ij}(x) = \frac{\partial^2 f}{\partial x_i \partial x_j}(x).$$

If f is smooth, then

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i},$$

and so the Hessian is symmetric. A critical point x^* is a local minimum of f if $H(x^*)$ is symmetric positive-definite.

Numerical methods for solving NLUMPs are all iterative in character. Some try to compute roots of the gradient, i.e., critical points of f, which are also local minima. These methods are related to schemes used to solve nonlinear equations described in Sec. 6.

Minimization methods can be classified into three main categories:

- *direct-search methods,* which make use of function values only,
- *gradient methods,* which make use of function and derivative values, and
- *Hessian methods,* which make use of function, derivative, and second-derivative values.

Methods from each class are discussed below.

7.2 The Fibonacci and Golden-Section Search Methods

The Fibonacci and golden-section search methods belong to the class of direct-search methods. They are similar to the bisection method for nonlinear equations described in Sec. 6.4, although important differences exist. They are used in one-dimensional minimization only.

The Fibonacci and golden-section search methods are applicable if $f(x)$ is continuous and unimodal in the interval of interest $[a,b]$, where by *unimodal* we mean that $f(x)$ has exactly one local minimum $x^* \in [a,b]$ and $f(x)$ strictly decreases for $x \in [a,x^*]$ and strictly increases for $x \in [x^*,b]$.

The main idea behind these methods is the following. Let x_1 and x_2 be two points satisfying $a < x_1 < x_2 < b$. If $f(x_1) \geq f(x_2)$, then x^* must lie in $[x_1,b]$. On the other hand, if $f(x_1) \leq f(x_2)$ then x^* must lie in $[a,x_2]$. Thus, by evaluating f at a sequence of test points $x_1, x_2, \ldots, x_k$, we can successively reduce the length of the interval in which we know the local minimum lies, called the *interval of uncertainty.*

The Fibonacci and golden-section search methods differ only in the way the sequence of test points is chosen. Before describing the two methods, we give two more definitions.

The *coordinate* v of a point x *relative to an interval* $[a,b]$ is $v = (x - a)/(b - a)$.

The *Fibonacci numbers* are defined by the initial condition $F_0 = F_1 = 1$ and the recurrence $F_k = F_{k-1} + F_{k-2}$ for $k \geq 2$.

The Fibonacci search method applied to a function f on an interval $[a,b]$ starts with two points x_k and x_{k-1} satisfying $(x_k - a)/(b - a) = F_{k-1}/F_k$ and $(x_{k-1} - a)/(b - a) = F_{k-2}/F_k = 1 - F_{k-1}/F_k$, whence $a < x_{k-1} < x_k < b$. It then computes the sequence $x_{k-2}, x_{k-3}, \ldots, x_1$ as follows. After evaluating $f(x_k)$ and $f(x_{k-1})$, the interval of uncertainty is either $[a,x_k]$ or $[x_{k-1},b]$.

- If the interval of uncertainty is $[a,x_k]$, then x_{k-1} belongs to it and $(x_{k-1} - a)/(x_k - a) = F_{k-2}/F_{k-1}$. In this case, x_{k-2} is chosen to satisfy $(x_{k-2} - a)/(x_k - a) = F_{k-3}/F_{k-1}$. The method proceeds to the next iteration with the two points x_{k-1} and x_{k-2} in the interval $[a,x_k]$, with relative coordinates F_{k-2}/F_{k-1} and F_{k-3}/F_{k-1}, respectively. Note that $f(x_{k-1})$ is already computed, so that only $f(x_{k-2})$ needs to be evaluated in the next iteration.
- If the interval of uncertainty is $[x_{k-1},b]$, then x_k belongs to it and $(x_k - x_{k-1})/(b - x_{k-1}) = F_{k-3}/F_{k-1}$. In this case, x_{k-2} is chosen to satisfy $(x_{k-2} - x_{k-1})/(b - x_{k-1}) = F_{k-2}/F_{k-1}$. The method proceeds to the next iteration with the two points x_k and x_{k-2} in the interval $[x_{k-1},b]$, with relative coordinates F_{k-3}/F_{k-1} and F_{k-2}/F_{k-1}, respectively. Note that $f(x_k)$ is already computed, so that only $f(x_{k-2})$ needs to be evaluated in the next iteration.

Thus, at the start of the second iteration, the situation is similar to that at the start of the first iteration, except that the length of the interval of uncertainty has been reduced. Therefore, the process described above can be repeated.

Before the last iteration, the interval of uncertainty is $[c,d]$ and x_1 is chosen to be $(c + d)/2 + \epsilon$, for some small positive number ϵ.

As noted above, the Fibonacci search method requires one function evaluation per iteration. The main disadvantage of the method is that the number of iterations k must be chosen at the start of the method. However, it can be proved that, given k, the length of the final interval of uncertainty is the shortest possible. Thus, in a sense, it is an optimal method.

The golden-section search method also requires one function evaluation per iteration, but the number of iterations k does not need to be chosen at the start of the method. It produces a sequence of test points $x_1, x_2, \ldots$ and stops when a predetermined accuracy is reached.

Let $r = (\sqrt{5} - 1)/2 \approx 0.618$ be the positive root of the quadratic $x^2 + x - 1$. Note that $1/r = (\sqrt{5} + 1)/2 \approx 1.618$ is the famous *golden ratio*. For the Fibonacci search method, it can be proved that, for large k, the coordinates of the two initial points x_k and x_{k-1} relative to $[a,b]$ are approximately r and $1 - r$, respectively. Thus, if, at each iteration, the two points are chosen with these coordinates relative to the interval of uncertainty, the resulting method, called the golden-section search method, is an approximation to the Fibonacci search. Moreover, if a point has coordinate $1 - r$ relative to $[a,b]$, then it has coordinate r relative to $[a, a + (b - a)r]$. Similarly, if a point has coordinate r relative to $[a,b]$, then it has coordinate $1 - r$ relative to $[a + (b - a)(1 - r), b]$. This property is exploited in the golden-section search method, enabling it to use one function evaluation only per iteration.

Both methods described above are guaranteed to converge whenever they are applicable and both have a linear rate of convergence (see Sec. 6.6). On the average, the length of the interval of uncertainty is multiplied by r at each iteration. For a further discussion of these methods, see an introductory numerical methods text such as Kahaner *et al.* (1989).

7.3 The Steepest-Descent Method

The steepest-descent (SD) method belongs to the class of gradient schemes. It is applicable whenever the partial derivatives of f exist and f has at least one local minimum. Whenever it is applicable, it is guaranteed to converge to some local minimum if the partial derivatives of f are continuous. However, it may converge slowly for multidimensional problems. Moreover, if f possesses more than one local minimum, there is no guarantee to which minimum SD will converge.

At iteration k, the SD method performs a search for the minimum of f along the line $x_k - \alpha\nabla f(x_k)$, where α is a scalar variable and $-\nabla f(x_k)$ is the direction of the steepest descent of f at x_k. Note that, since α is scalar, minimizing $f(x_k - \alpha\nabla f(x_k))$ with respect to (w.r.t.) α is a one-dimensional minimization problem. If α^* is the minimizer, x_{k+1} is taken to be $x_k - \alpha^*\nabla f(x_k)$. A brief outline of SD is given in Table 11. See Buchanan and Turner (1992), Johnson and Riess (1982), or Ortega (1988) for further details.

7.4 Conjugate-Direction Methods

The definition of conjugate directions is given w.r.t. a symmetric positive-definite (SPD) matrix A: the vectors (directions) u and v are A-conjugate if $u^{\mathrm{T}}Av = 0$. Thus, conjugate directions are orthogonal or perpendicular directions w.r.t. an inner product $(u,v) = u^{\mathrm{T}}Av$ and, as such, are often associated with some shortest-distance property.

The conjugate-direction (CD) methods form a large class of minimization schemes. Their common characteristic is that the search direction at every iteration is conjugate to previous search directions. Proceeding along conjugate search directions guarantees, in some sense, finding the shortest path to the minimum.

The CD methods are guaranteed to converge in at most n iterations for the SPD quadratic function $f(x) = c + b^{\mathrm{T}}x - \frac{1}{2}x^{\mathrm{T}}Ax$, where A is an $n \times n$ SPD matrix.

There are several techniques to construct conjugate directions, each one giving rise to

Table 11. The steepest-descent method.

```
Pick an initial guess $x_0$ and a tolerance $\epsilon$
for $k = 1, \ldots,$ maxit do
    $s_{k-1} = -\nabla f(x_{k-1})$
    if $\|s_{k-1}\| \le \epsilon$ exit loop
    find $\alpha^* \in \mathbb{R}$ that minimizes $f(x_{k-1} + \alpha s_{k-1})$
    $x_k = x_{k-1} + \alpha^* s_{k-1}$
end
```

a different CD method. The best-known is *Powell's* method [see Buchanan and Turner (1992) or Ortega (1988)].

7.5 The Conjugate-Gradient Method

As the name implies, at each iteration, the conjugate-gradient (CG) method takes information from the gradient of f to construct conjugate directions. Its search direction is a linear combination of the direction of steepest descent and the search direction of the previous iteration. A brief outline of CG is given in Table 12.

The CG method is guaranteed to converge in at most n iterations for the SPD quadratic function $f(x) = c + b^{\mathrm{T}}x - \frac{1}{2}x^{\mathrm{T}}Ax$, where A is a $n \times n$ SPD matrix. See Sec. 3.2, Golub and Van Loan (1989), or Ortega (1988) for a more detailed discussion of this minimization technique applied to solve linear algebraic systems.

There exist several variants of the CG method, most of which are based on slightly different ways of computing the step size β.

7.6 Newton's Method

Newton's method for minimizing f is just Newton's method for nonlinear systems applied to solve $\nabla f(x) = 0$. The new iterate x_k is computed by $x_k = x_{k-1} - [H(x_{x-1})]^{-1} \times \nabla f(f_{k-1})$, where $H(x_{k-1})$ is the Hessian of f at x_{k-1}. See Sec. 6.7 for further details.

If f is convex, then the Hessian is SPD and the search direction generated by Newton's method at each iteration is a descent (downhill) direction. Thus, for any initial guess x_0, Newton's method is guaranteed to converge quadratically, provided f is sufficiently smooth.

For a general function f, there is no guarantee that Newton's method will converge for an arbitrary initial guess x_0. However, if started close enough to the minimum of a sufficiently smooth function f, Newton's method normally converges quadratically, as noted in Sec. 6.6. There exist several variants of Newton's method that improve upon the reliability of the standard scheme.

Table 12. The conjugate-gradient (CG) method.

```
Pick an initial guess $x_0$ and a tolerance $\epsilon$
Initialize $s_0 = 0$ and $\beta = 1$
for $k = 1, \ldots,$ maxit do
    if $\|\nabla f(x_{k-1})\| \leq \epsilon$ exit loop
    $s_k = -\nabla f(x_{k-1}) + \beta s_{k-1}$
    find $\alpha^* \in \mathbb{R}$ that minimizes $f(x_{k-1} + \alpha s_k)$
    $x_k = x_{k-1} + \alpha^* s_k$
    $\beta = \|\nabla f(x_k)\|^2 / \|\nabla f(x_{k-1})\|^2$
end
```

7.7 Quasi-Newton Methods

At every iteration, Newton's method requires the evaluation of the Hessian and the solution of a linear system $[H(x_{k-1})]s_k = -\nabla f(x_{k-1})$ for the search direction s_k. Quasi-Newton methods update an approximation to the inverse Hessian at every iteration, thus reducing the task of solving a linear system to a simple matrix–vector multiply. The best-known of these schemes are the *Davidon–Fletcher–Powell* (DFP) and the *Broyden–Fletcher–Goldfarb–Shanno* (BFGS) methods. See Buchanan and Turner (1992) or Dennis and Schnabel (1983) for further details.

8. APPROXIMATION

It is often desirable to find a function $f(x)$ in some class that approximates the data $\{(x_i, y_i) : i = 1, \ldots, n\}$. That is, $f(x_i) \approx y_i$, $i = 1, \ldots, n$. If f matches the data exactly, that is, f satisfies the *interpolation relations* or *interpolation conditions* $f(x_i) = y_i$, $i = 1, \ldots, n$, then f is called the *interpolating function* or the *interpolant* of the given data.

Similarly, it is often desirable to find a simple function $f(x)$ in some class that approximates a more complex function $y(x)$. We say that *f interpolates y*, or f is the *interpolant* of y, at the points x_i, $i = 1, \ldots, n$, if $f(x_i) = y(x_i)$, $i = 1, \ldots, n$. The problem of computing the interpolant f of y at the points x_i, $i = 1, \ldots, n$, reduces to the problem of computing the interpolant f of the data $\{(x_i, y_i) : i = 1, \ldots, n\}$, where $y_i = y(x_i)$.

An interpolant does not always exist and, when it does, it is not necessarily unique. However, a unique interpolant does exist in many important cases, as discussed below.

A standard approach for constructing an interpolant is to choose a set of *basis functions* $\{b_1(x), b_2(x), \ldots, b_n(x)\}$ and form a *model*

$$f(x) = \sum_{j=1}^{n} a_j b_j(x),$$

where the numbers a_j are unknown coefficients. For f to be an interpolant, it must satisfy $f(x_i) = y_i$, $i = 1, \ldots, n$, which is equivalent to

$$\sum_{j=1}^{n} a_j b_j(x_i) = y_i, \quad i = 1, \ldots, n.$$

These n conditions form a system $B\mathbf{a} = \mathbf{y}$ of n linear equations in n unknowns, where $\mathbf{a} = (a_1, a_2, \ldots, a_n)^T$, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, and $B_{ij} = b_j(x_i)$, $i, j = 1, \ldots, n$. If B is nonsingular, then the interpolant of the data $\{(x_i, y_i) : i = 1, \ldots, n\}$ w.r.t. the basis functions $b_1(x)$, $b_2(x), \ldots, b_n(x)$ exists and is unique. On the other hand, if B is singular, then either the interpolant may fail to exist or there may be infinitely many interpolants.

8.1 Polynomial Approximation

Polynomial approximation is the foundation for many numerical procedures. The basic idea is that, if we want to apply some procedure to a function, such as integration (see Sec. 9), we approximate the function by a polynomial and apply the procedure to the approximating polynomial. Polynomials are often chosen as approximating functions because they are easy to evaluate (see Sec. 6.10), to integrate, and to differentiate. Moreover, polynomials approximate well more complicated functions, provided the latter are sufficiently smooth. The following mathematical result ensures that arbitrarily accurate polynomial approximations exist for a broad class of functions.

Weierstrass Theorem: If $g(x) \in \mathscr{C}[a,b]$, then, for every $\epsilon > 0$, there exists a polynomial $p_n(x)$ of degree $n = n(\epsilon)$ such that $\max\{|g(x) - p_n(x)| : x \in [a,b]\} \leq \epsilon$.

8.2 Polynomial Interpolation

Techniques to construct a polynomial interpolant for a set of data $\{(x_i,y_i) : i = 1, \ldots, n\}$ are discussed below. Here we state only the following key result.

Theorem: If the points $\{x_i : i = 1, \ldots, n\}$ are distinct, then there exists a unique polynomial of degree at most $n - 1$ that interpolates the data $\{(x_i,y_i) : i = 1, \ldots, n\}$. (There are no restrictions on the y_i's.)

8.2.1 Monomial Basis One way to construct a polynomial that interpolates the data $\{(x_i, y_i) : i = 1, \ldots, n\}$ is to choose as basis functions the monomials $b_j(x) = x^{j-1}$, $j = 1, \ldots, n$, giving rise to the model $p_{n-1}(x) = a_1 + a_2 x + a_3 x^2 + \cdots + a_n x^{n-1}$. As noted above, the interpolation conditions take the form $B\mathbf{a} = \mathbf{y}$. In this case, B is the *Vandermonde matrix* for which $B_{ij} = x_i^{j-1}$, $i, j = 1, \ldots, n$, where we use the convention that $x^0 = 1$ for all x.

It can be shown that the Vandermonde matrix B is nonsingular if and only if the points $\{x_i : i = 1, \ldots, n\}$ are distinct. If B is nonsingular, then, of course, we can solve the system $B\mathbf{a} = \mathbf{y}$ to obtain the coefficients a_j, $j = 1, \ldots, n$, for the unique interpolant of the data.

It can also be shown that, although the Vandermonde matrix is nonsingular for distinct points, it can be ill-conditioned, particularly for large n. As a result, the methods described below are often computationally much more effective than the scheme described here.

8.2.2 Lagrange Basis An alternative to the monomial basis functions discussed above is the Lagrange basis polynomials

$$b_j(x) = l_j(x) = \prod_{\substack{i=1 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i} \quad j = 1, \ldots, n,$$

which are of degree $n - 1$ and satisfy

$$b_j(x_i) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Hence $B = I$ and the system $B\mathbf{a} = \mathbf{y}$ of interpolation conditions has the obvious unique solution $a_j = y_j$, $j = 1, \ldots, n$.

Note that changing the basis from monomials to Lagrange polynomials does not change the resulting interpolating polynomial $p_{n-1}(x)$, since the interpolant is unique. It only affects the representation of $p_{n-1}(x)$.

8.2.3 Newton Basis and Divided Differences Another useful basis is the set of Newton polynomials

$$b_j(x) = \prod_{i=1}^{j-1} (x - x_i) \quad j = 1, \ldots, n.$$

The coefficients a_j of the interpolating polynomial written with the Newton basis are relatively easy to compute by a recursive algorithm using *divided differences.* Before describing this form of the interpolating polynomial, though, we must introduce divided differences.

Given a function f with $f(x_i) = y_i$, $i = 1, \ldots, n$, define the divided difference with one point by

$$f[x_i] = y_i \quad i = 1, \ldots, n.$$

If $x_{i+1} \neq x_i$, define the divided difference with two points by

$$f[x_i, x_{i+1}] = \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \quad i = 1, \ldots, n - 1.$$

If $x_i \neq x_{i+k}$, define the divided difference with $k + 1$ points by

$$f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{f[x_{i+1}, x_{i+2}, \ldots, x_{i+k}] - f[x_i, x_{i+1}, \ldots, x_{i+k-1}]}{x_{i+k} - x_i} \quad i = 1, \ldots, n - k.$$

We can extend this definition of divided differences to sets $\{x_i : i = 1, \ldots, n\}$ with repeated values by noting that

$$\lim_{x_{i+1} \to x_i} f[x_{i+1}, x_i] = \lim_{x_{i+1} \to x_i} \frac{y_{i+1} - y_i}{x_{i+1} - x_i} = f'(x_i).$$

So, if $x_i = x_{i+1}$, we define $f[x_i, x_{i+1}] = f'(x_i)$. Similarly, it can be shown that

$$\lim_{\substack{x_{i+1} \to x_i \\ \vdots \\ x_{i+k} \to x_i}} f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{f^{(k)}(x_i)}{k!}.$$

So, if $x_i = x_{i+1} = \cdots = x_{i+k}$, we define $f[x_i, x_{i+1}, \ldots, x_{i+k}] = f^{(k)}(x_i)/k!$.

Using divided differences, the coefficients a_j can be computed as $a_j = f[x_1, \ldots, x_j]$. An advantage of the divided-difference form is that it extends easily to the interpolation of derivatives as well as function values, as discussed below. Another advantage is that if the coefficients a_j, $j = 1, \ldots, n$, are computed from n data points, it is easy to add more data points and construct a higher-degree interpolating polynomial without redoing the whole computation, since the new data can be added easily to the existing divided-difference table and the interpolating polynomial extended.

8.3 Polynomial Interpolation with Derivative Data

In this section, we consider *Hermite* or *osculatory interpolation,* which requires that a function and its first derivative be interpolated at the points $\{x_i : i = 1, \ldots, n\}$. The key result is stated below.

Theorem: Given the data $\{(x_i, y_i, y_i') : i = 1, \ldots, n\}$, where the points $\{x_i : i = 1, \ldots, n\}$ are distinct, there exists a unique polynomial interpolant $p_{2n-1}(x)$ of degree at most $2n - 1$ that satisfies $p_{2n-1}(x_i) = y_i$, $i = 1, \ldots, n$, and $p_{2n-1}'(x_i) = y_i'$, $i = 1, \ldots, n$.

The techniques used to construct such a polynomial are similar to those described in Sec. 8.2. The following choices of basis functions are often used.

1. Monomials: $b_j(x) = x^{j-1}$, $j = 1, \ldots, 2n$;
2. Generalized Lagrange basis polynomials: $b_j(x) = [1 - 2(x - x_j)l_j'(x_j)][l_j(x)]^2$, $j = 1, \ldots, n$, and $b_{n+j}(x) = (x - x_j)[l_j(x)]^2$, $j = 1, \ldots, n$;
3. Newton basis polynomials:

$$b_j(x) = \prod_{i=1}^{j-1} (x - x_i)^2, \quad j = 1, \ldots, n,$$

and

$$b_{n+j}(x) = \left(\prod_{i=1}^{j-1} (x - x_i)^2\right)(x - x_j), \quad j = 1, \ldots, n.$$

More general forms of polynomial interpolants are discussed in some numerical methods books. See for example Davis (1975).

8.4 The Error in Polynomial Interpolation

Two key results for the error in polynomial interpolation are given in the following two theorems. For their proofs, see an introductory numerical methods text such as Dahlquist and Björck (1974), Johnson and Riess (1982), or Stoer and Bulirsch (1980). Before stating the results, though, we must define $\mathrm{spr}[x,x_1,x_2,\ldots,x_n]$ to be the smallest interval containing $x, x_1, x_2, \ldots, x_n$.

Theorem: Let $g(x) \in \mathscr{C}^n$, and let $p_{n-1}(x)$ be the polynomial of degree at most $n - 1$ that interpolates $g(x)$ at the n distinct points $x_1, x_2, \ldots, x_n$. Then, for any x,

$$g(x) - p_{n-1}(x) = \frac{g^{(n)}(\xi)}{n!} \prod_{i=1}^{n} (x - x_i),$$

where $g^{(n)}(x)$ is the nth derivative of $g(x)$ and ξ is some point in $\mathrm{spr}[x,x_1,x_2,\ldots,x_n]$.

Theorem: Let $g(x) \in \mathscr{C}^{2n}$, and let $p_{2n-1}(x)$ be the polynomial of degree at most $2n - 1$ that interpolates $g(x)$ and $g'(x)$ at the n distinct points $x_1, x_2, \ldots, x_n$. Then, for any x,

$$g(x) - p_{2n-1}(x) = \frac{g^{(2n)}(\xi)}{(2n)!} \prod_{i=1}^{n} (x - x_i)^2,$$

where $g^{(2n)}(x)$ is the $2n$th derivative of $g(x)$ and ξ is some point in $\mathrm{spr}[x,x_1,x_2,\ldots,x_n]$.

Note that there is a close relationship between the error in polynomial interpolation and the error in a Taylor series. As a result, polynomial interpolation is normally effective if and only if g can be approximated well by a Taylor series.

More specifically, the polynomial interpolation error can be large if the derivative appearing in the error formula is big or if $\mathrm{spr}[x,x_1,x_2,\ldots,x_n]$ is big, particularly if the point x of evaluation is close to an end point of $\mathrm{spr}[x_1,x_2,\ldots,x_n]$ or outside this interval.

8.5 Piecewise Polynomials and Splines

Given a set of *knots* or *grid points* $\{x_i : i = 1, \ldots, n\}$ satisfying $a = x_0 < x_1 < \cdots < x_n = b$, $s(x)$ is a *piecewise polynomial* (PP) of degree N w.r.t. the knots $\{x_i, i = 0, \ldots, n\}$ if $s(x)$ is a polynomial of degree N on each interval (x_{i-1},x_i), $i = 1, \ldots, n$. A polynomial of degree N is always a PP of degree N, but the converse is not necessarily true.

A *spline* $s(x)$ of degree N is a PP of degree N. The term spline usually implies the continuity of $s(x), s'(x), \ldots, s^{(N-1)}(x)$ at the knots $\{x_0,x_1,\ldots,x_n\}$. In this case, $s \in \mathscr{C}^{N-1}$, the space of continuous functions with $N - 1$ continuous derivatives. Sometimes, though, the terms PP and spline are used interchangeably.

Let $s(x)$ be a PP of degree N, and assume that $s(x), s'(x), \ldots, s^{(K)}(x)$ are continuous at the knots $\{x_0,x_1,\ldots,x_n\}$. Since $s(x)$ is a polynomial of degree N on each of the n subintervals (x_i,x_{i-1}), $i = 1, \ldots, n$, it is defined by $D = n(N + 1)$ coefficients. To determine these coefficients, we take into account the continuity conditions that s and its K derivatives must satisfy at the $n - 1$ interior knots $\{x_1,\ldots,x_{n-1}\}$. There are $K + 1$ such conditions at each interior knot, giving rise to $C = (n - 1)(K + 1)$ conditions. Thus, we need $M = D - C = n(N - K) + K + 1$ properly chosen additional conditions to determine the coefficients of s.

In the following we give examples of PPs and splines and their associated basis functions.

8.5.1 Constant Splines The constant PP model function

$$\phi(x) = \begin{cases} 1 & \text{for } 0 \le x \le 1, \\ 0 & \text{elsewhere,} \end{cases}$$

is a constant spline w.r.t. the knots $\{0,1\}$. Note that it is not continuous at the knots, so that $\phi \in \mathscr{C}^{-1}$. The constant PP functions

$$\phi_i(x) = \phi((x - a)/h - i + 1), \quad i = 1, \ldots, n,$$

are constant splines w.r.t. the evenly spaced knots $\{x_i = a + ih : i = 0, \ldots, n\}$, where $h = (b - a)/n$, and form a set of basis functions for the space of constant splines w.r.t. these knots.

8.5.2 Linear Splines The linear PP model function

$$\phi(x) = \begin{cases} x & \text{for } 0 \le x \le 1, \\ 2 - x & \text{for } 1 \le x \le 2, \\ 0 & \text{elsewhere,} \end{cases}$$

is a linear spline w.r.t. the knots $\{0,1,2\}$. Note that ϕ is continuous, but ϕ' does not exist at the knots, so that $\phi \in \mathscr{C}^0$. The linear PP functions

$$\phi_i(x) = \phi((x - a)/h - i + 1), \qquad i = 0, \ldots, n, \quad (39)$$

are linear splines w.r.t. the evenly spaced knots $\{x_i = a + ih : i = 0, \ldots, n\}$, where $h = (b - a)/n$, and form a set of basis functions for the space of linear splines w.r.t. these knots.

8.5.3 Quadratic Splines The quadratic PP model function

$$\phi(x) = \begin{cases} x^2/2 & \text{for } 0 \le x \le 1, \\ [x^2 - 3(x-1)^2]/2 & \text{for } 1 \le x \le 2, \\ [x^2 - 3(x-1)^2 + 3(x-2)^2]/2 & \text{for } 2 \le x \le 3, \\ 0 & \text{elsewhere,} \end{cases}$$

is a quadratic spline w.r.t. the knots $\{0,1,2,3,\}$. Note that ϕ and ϕ' are continuous, but ϕ'' does not exist at the knots, so that $\phi \in \mathscr{C}^1$. The quadratic PP functions

$$\phi_i(x) = \phi[(x - a)/h - i + 2], \quad i = 0, \ldots, n + 1 \qquad (40)$$

are quadratic splines w.r.t. the evenly spaced knots $\{x_i = a + ih : i = 0, \ldots, n\}$, where $h = (b - a)/n$, and form a set of basis functions for the space of quadratic splines w.r.t. these knots.

8.5.4 Quadratic Piecewise Polynomials The quadratic PP model functions

$$\phi(x) = \begin{cases} x(2x - 1) & \text{for } 0 \le x \le 1, \\ (x - 2)(2x - 3) & \text{for } 1 \le x \le 2, \\ 0 & \text{elsewhere,} \end{cases}$$

$$\psi(x) = \begin{cases} -4x(x - 1) & \text{for } 0 \le x \le 1, \\ 0 & \text{elsewhere,} \end{cases}$$

are continuous, but neither ϕ' nor ψ' exists at their knots $\{0,1,2\}$ and $\{0,1\}$, respectively. So ϕ and $\psi \in \mathscr{C}^0$. The functions

$$\phi_i(x) = \begin{Bmatrix} \phi((x - a)/h - i/2 + 1) & \text{for } i \text{ even} \\ \psi((x - a)/h - (i + 1)/2 + 1) & \text{for } i \text{ odd} \end{Bmatrix} \quad i = 0, \ldots, 2n$$

are quadratic PPs w.r.t. the evenly spaced knots $\{x_i = a + ih : i = 0, \ldots, n\}$, where $h = (b - a)/n$. Note that ϕ_i is continuous, but ϕ_i' does not exist at the knots, and so $\phi_i \in \mathscr{C}^0$. These functions form a basis for the space of quadratic PPs in $\mathscr{C}^0$ w.r.t. these knots.

8.5.5 Cubic Splines The cubic PP model function

$$\phi(x) = \begin{cases} x^3/6 & \text{for } 0 \le x \le 1, \\ [x^3 - 4(x-1)^3]/6 & \text{for } 1 \le x \le 2, \\ [x^3 - 4(x-1)^3 + 6(x-2)^3]/6 & \text{for } 2 \le x \le 3, \\ [x^3 - 4(x-1)^3 + 6(x-2)^3 - 4(x-3)^3]/6 & \text{for } 3 \le x \le 4, \\ 0 & \text{elsewhere,} \end{cases}$$

is a cubic spline w.r.t. the knots $\{0,1,2,3,4\}$. Note that ϕ, ϕ', and ϕ'' are continuous, but ϕ''' does not exist at the knots, and so $\phi \in \mathscr{C}^2$. The cubic PP functions

$$\phi_i(x) = \phi((x - a)/h - i + 2), \quad i = -1, \ldots, n + 1 \qquad (41)$$

are cubic splines w.r.t. the evenly spaced knots $\{x_i = a + ih : i = 0, \ldots, n\}$, where $h = (b - a)/n$, and form a set of basis functions for the space of cubic splines w.r.t. these knots.

8.5.6 Cubic Hermite Piecewise Polynomials The cubic PP model functions

$$\phi(x) = \begin{cases} x^2(3 - 2x) & \text{for } 0 \le x \le 1, \\ (2 - x)^2(2x - 1) & \text{for } 1 \le x \le 2, \\ 0 & \text{elsewhere,} \end{cases}$$

$$\psi(x) = \begin{cases} x^2(x - 1) & \text{for } 0 \le x \le 1, \\ (2 - x)^2(x - 1) & \text{for } 1 \le x \le 2, \\ 0 & \text{elsewhere,} \end{cases}$$

are in $\mathscr{C}^1$ since ϕ, ϕ', ψ, and ψ' are all continuous at the knots $\{0,1,2\}$, but neither ϕ'' nor ψ'' exists at the knots. The functions

$$\phi_i(x) = \begin{Bmatrix} \phi((x - a)/h - i/2 + 1) & \text{for } i \text{ even} \\ \psi((x - a)/h - (i - 1)/2 + 1) & \text{for } i \text{ odd} \end{Bmatrix}, \quad i = 0, \ldots, 2n + 1 \qquad (42)$$

are cubic PPs w.r.t. the evenly spaced knots $\{x_i = a + ih : i = 0, \ldots, n\}$, where $h = (b - a)/n$. Note that ϕ_i and ϕ_i' are continuous, but ϕ_i'' does not exist at the knots, and so $\phi_i \in \mathscr{C}^1$. These functions form a basis for the space of cubic PPs in $\mathscr{C}^1$ w.r.t. these knots.

8.6 Piecewise Polynomial Interpolation

Piecewise polynomials, including splines, are often used for interpolation, especially for large data sets. The main advantage in using PPs, instead of polynomials, to interpolate a function $g(x)$ at n points, where $n \gg 1$, is that the error of a PP interpolant does not depend on the nth derivative of $g(x)$, but rather on a low-order derivative of $g(x)$. Usually, if $g(x) \in \mathscr{C}^{N+1}[a,b]$ and $s_N(x)$ is a PP of degree N in $\mathscr{C}^K$ that interpolates $g(x)$ at $n(N - K) + K + 1$ properly chosen points, then the interpolation error at any point $x \in [a,b]$ satisfies

$$|g(x) - s_N(x)| \leq Ch^{N+1} \max_{a\leq\xi\leq b} |g^{(N+1)}(\xi)|$$

for some constant C independent of $h = \max\{x_i - x_{i-1} : i = 1, \ldots, n\}$. Another advantage of PP interpolation is that it leads either to simple relations that define the PP interpolant or gives rise to banded systems that can be solved easily for the coefficients of the PP interpolant by the techniques described in Sec. 2.7. In the following subsections, we give examples of PP interpolation. We assume throughout that $h = (b - a)/n$ and $x_i = a + ih$ for $i = 0, \ldots, n$, but these restrictions can be removed easily.

An introduction to PP interpolation is presented in many introductory numerical analysis textbooks, such as Johnson and Riess (1982). For a more detailed discussion, see an advanced text such as Prenter (1975) or de Boor (1978).

8.6.1 Linear Spline Interpolation Let $\phi_i(x)$ be defined by (39). The function

$$s_1(x) = \sum_{i=0}^{n} c_i\phi_i(x)$$

with $c_i = g(x_i)$ is the unique linear spline that interpolates $g(x)$ at the knots $\{x_i : i = 0, \ldots, n\}$.

8.6.2 Quadratic Spline Interpolation Let $\phi_i(x)$ be defined by (40). The function

$$s_2(x) = \sum_{i=0}^{n+1} c_i\phi_i(x)$$

is the unique quadratic spline that interpolates $g(x)$ at the $n + 2$ points x_0, $(x_{i-1} + x_i)/2$, $i = 1, \ldots, n$, and x_n, if the coefficients c_i, $i = 0, \ldots, n + 1$, are chosen so that they satisfy the $n + 2$ conditions $s_2(x_0) = g(x_0)$, $s_2((x_{i-1} + x_i)/2) = g((x_{i-1} + x_i)/2)$, $i = 1, \ldots, n$, and $s_2(x_n) = g(x_n)$. These conditions give rise to a tridiagonal linear system of $n + 2$ equations in the $n + 2$ unknowns c_i, $i = 0, \ldots, n + 1$:

$$\begin{pmatrix} 0.5 & 0.5 & & & & \\ 0.125 & 0.75 & 0.125 & & & \\ & 0.125 & 0.75 & 0.125 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 0.125 & 0.75 & 0.125 \\ & & & & 0.5 & 0.5 \end{pmatrix} \times \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_n \\ c_{n+1} \end{pmatrix} = \begin{pmatrix} g(x_0) \\ g((x_0 + x_1)/2) \\ g((x_1 + x_2)/2) \\ \vdots \\ g((x_{n-1} + x_n)/2) \\ g(x_n) \end{pmatrix}.$$

Since this system is column diagonally dominant, it is nonsingular and has a unique solution. Moreover, if the first and last equations are multiplied by 0.25, the resulting system is symmetric positive-definite. So the coefficients $\{c_i\}$ of the associated quadratic spline interpolant can be computed easily by the techniques described in Sec. 2.7.

8.6.3 Cubic Spline Interpolation Let $\phi_i(x)$ be defined by (41). The function

$$s_{32}(x) = \sum_{i=-1}^{n+1} c_i\phi_i(x)$$

is a cubic spline that interpolates $g(x)$ at the $n + 1$ points x_i, $i = 0, \ldots, n$, if the coefficients c_i, $i = -1, \ldots, n + 1$, are chosen so that they satisfy the $n + 1$ conditions $s_{32}(x_i) = g(x_i)$, $i = 0, \ldots, n$. These conditions give rise to a linear system of $n + 1$ equations in the $n + 3$ unknowns c_i, $i = -1, \ldots, n + 1$. To determine the unknown coefficients uniquely, two more appropriately chosen conditions (equations) are needed. The standard choices are

1. $s''_{32}(x_0) = 0$ and $s''_{32}(x_n) = 0$, giving rise to the *natural* or *minimum-curvature* cubic spline interpolant;
2. $s'_{32}(x_0) = g'(x_0)$ and $s'_{32}(x_n) = g'(x_n)$, giving rise to a more accurate cubic spline interpolant, but one that requires knowledge of $g'(x)$ (or a good approximation to it) at both end points;
3. the *not-a-knot* conditions, which force $s_{32}(x)$ to have a continuous third derivative at the knots x_1 and x_{n-1}.

In all three cases, the resulting linear system of $n + 3$ equations in $n + 3$ unknowns is almost tridiagonal. More specifically, it is tridiagonal with the exception of the two rows corresponding to the extra conditions. So the coefficients $\{c_i\}$ of the associated cubic spline interpolant can be computed easily by the techniques described in Sec. 2.7.

8.6.4 Cubic Hermite Piecewise Polynomial Interpolation Let $\phi_i(x)$ be defined by (42). The function

$$s_{31}(x) = \sum_{i=0}^{2n+1} c_i\phi_i(x)$$

with $c_{2i} = g(x_i)$, $i = 0, \ldots, n$, and $c_{2i+1} = g'(x_i)$, $i = 0, \ldots, n$, is the unique Hermite PP that interpolates $g(x)$ and its derivative at the $n + 1$ points x_i, $i = 0, \ldots, n$.

8.7 Least-Squares Approximation

It is possible to construct a function $f(x)$ that approximates the data $\{(x_i, y_i) : i = 0, \ldots, m\}$ in the sense that $f(x_i) - y_i$ is "small" for $i = 0, \ldots, m$, but $f(x_i) \neq y_i$ in general. One way is to construct an f in some class of functions that minimizes

$$\sum_{i=0}^{m} w_i[f(x_i) - y_i]^2$$

for some positive *weights* w_i, $i = 0, \ldots, m$. Such an f is called a *discrete least-squares* approximation to the data.

Similarly, it is possible to construct a function $f(x)$ that approximates a continuous function $y(x)$ on an interval $[a,b]$ in the sense that $|f(x) - y(x)|$ is "small" for all $x \in [a,b]$. One way is to construct an f in some class of functions that minimizes

$$\int_a^b w(x)[f(x) - y(x)]^2dx$$

for some positive and continuous *weight function* $w(x)$ on (a,b). Such an f is called a *continuous least-squares* approximation to y.

Often, f is chosen to be a polynomial of degree $n < m$. (For the discrete least-squares problem, if f is a polynomial of degree $n = m$, then f interpolates the data.)

Let f and $g \in \mathscr{C}[a,b]$. We denote the *discrete inner product* of f and g at distinct points x_i, $i = 0, \ldots, m$, w.r.t. the weights w_i, $i = 0, \ldots, m$, by

$$(f,g) = \sum_{i=0}^{m} w_i f(x_i)g(x_i)$$

and the *continuous inner product* of f and g on $[a,b]$ w.r.t. weight function $w(x)$ by

$$(f,g) = \int_a^b w(x)f(x)g(x)dx.$$

Given an inner product $(\cdot,\cdot)$, as above, we denote the *norm* of f by $\|f\| = (f,f)^{1/2}$. Thus we have the *discrete norm*

$$\|f\| = \left(\sum_{i=0}^{m} w_i f(x_i)^2\right)^{1/2}$$

and the *continuous norm*

$$\|f\| = \left(\int_a^b w(x)f(x)^2dx\right)^{1/2}.$$

Therefore, the problem of constructing a least-squares approximation $f(x)$ to a given set of data $\{(x_i,y_i)\}$ or to a given function $y(x)$ is to construct an $f(x)$ that minimizes the norm, discrete or continuous, respectively, of the error $\|f - y\|$.

We note that the "discrete inner product" is not strictly speaking an inner product in all cases, since it may fail to satisfy the property that $(f,f) = 0$ implies $f(x) = 0$ for all x. However, if we restrict the class of functions to which f belongs appropriately, then $(\cdot,\cdot)$ is a true inner product. For example, if we restrict f to the class of polynomials of degree at most n and if $n < m$, then $(f,f) = 0$ implies $f(x) = 0$ for all x. Similar remarks apply to the discrete norm.

Before giving the main theorem on how to construct the least-squares polynomial approximation to a given set of data or to a given function, we introduce orthogonal polynomials and the Gram–Schmidt process to construct them.

8.7.1 Orthogonal Polynomials A set of $n + 1$ polynomials $\{q_i(x) : i = 0, \ldots, n\}$ is *orthogonal* w.r.t. the inner product $(\cdot,\cdot)$ if $(q_i,q_j) = 0$ for $i \neq j$. A set of $n + 1$ orthogonal polynomials $\{q_i(x) : i = 0, \ldots, n\}$ is *orthonormal* w.r.t. the inner product $(\cdot,\cdot)$ if in addition $(q_i,q_i) = 1$ for $i = 0, \ldots, n$.

8.7.2 The Gram–Schmidt Orthogonalization Algorithm The Gram–Schmidt algorithm applied to a set of $n + 1$ linearly independent polynomials $\{p_j : j = 0, \ldots, n\}$ generates a set of $n + 1$ orthonormal polynomials $\{q_i(x) : i = 0, \ldots, n\}$ and a set of $n + 1$ orthogonal polynomials $\{s_i(x): i = 0, \ldots, n\}$. Often the set of $n + 1$ linearly independent polynomials $\{p_j : j = 0, \ldots, n\}$ is chosen to be the set of monomials $\{x^j : j = 0, \ldots, n\}$.

The Gram–Schmidt algorithm for polynomials is similar to the Gram–Schmidt algorithm for matrices described in Sec. 4.6. The reader may refer to an introductory numerical methods text such as Johnson and Riess (1982) for more details. Here we note only that the role of the inner product of vectors in the algorithm in Sec. 4.6 is replaced by the inner product, discrete or continuous, of functions as defined in Sec. 8.7.1.

8.7.3 Constructing the Least-Squares Polynomial Approximation The following result is proved in many introductory numerical methods books. See for example Johnson and Riess (1982).

Theorem: Assume that we are given either a continuous function $y(x)$ on $[a,b]$ or a data set $\{(x_i,y_i) : i = 0, \ldots, m\}$. Let $\{q_j : j = 0, \ldots, n\}$ be a set of orthonormal polynomials w.r.t. an inner product $(\cdot,\cdot)$ appropriate for the given data and assume that $\{q_j : j = 0, \ldots, n\}$ spans $\{x^i : i = 0, \ldots, n\}$, where $n < m$ for the discrete problem. Then

$$p^*(x) = \sum_{j=0}^{n} (y,q_j)q_j(x)$$

is the least-squares polynomial approximate of degree at most n. It is optimal in the sense that, if $p(x)$ is any other polynomial of degree at most n, then $\|y - p^*\| < \|y - p\|$, where $\|\cdot\|$ is the norm associated with the inner product $(\cdot,\cdot)$.

As noted in Sec. 8.7.2, a set of orthonormal polynomials $\{q_j : j = 0, \ldots, n\}$ that spans $\{x^i : i = 0, \ldots, n\}$ can be constructed by the Gram–Schmidt algorithm applied to the monomial basis polynomials $\{x^i : i = 0, \ldots, n\}$.

9. NUMERICAL INTEGRATION—QUADRATURE

In this section, we consider formulas for approximating integrals of the form

$$I(f) = \int_a^b f(x)dx.$$

Such formulas are often called quadrature rules. We assume that a and b are finite and that f is smooth in most cases, but we briefly discuss infinite integrals and singularities in Sec. 9.5.

In many practical problems, $f(x)$ is given either as a set of values $f(x_1), \ldots, f(x_n)$ or $f(x)$ is hard or impossible to integrate exactly. In these cases, the integral may be approximated by numerical techniques, which often take the form

$$Q(f) = \sum_{i=1}^{n} w_i f(x_i).$$

Such a formula is called a *quadrature rule,* the $\{w_i\}$ are called *weights,* and the $\{x_i\}$ are called *abscissae* or *nodes.*

Most quadrature rules are derived by first approximating f by a simpler function, frequently a polynomial, and then integrating the simpler function. Thus, the area under the curve f, which is the exact value of $I(f)$, is approximated by the area under the curve of the simpler function.

For a more detailed discussion of the topics in this section, see an introductory numerical methods text such as Conte and de

Boor (1980), Dahlquist and Björck (1974), Johnson and Riess (1982), or Stoer and Bulirsch (1980).

9.1 Simple Quadrature Rules

Several simple quadrature rules follow:

- The *rectangle rule* approximates $I(f)$ by the area under the constant $y = f(a)$ or $y = f(b)$.
- The *midpoint rule* approximates $I(f)$ by the area under the constant $y = f((a + b)/2)$.
- The *trapezoidal rule* approximates $I(f)$ by the area under the straight line joining the points $(a,f(a))$ and $(b,f(b))$.
- *Simpson's rule* approximates $I(f)$ by the area under the quadratic that interpolates $(a,f(a))$, $(m,f(m))$ and $(b,f(b))$, where $m = (a + b)/2$.
- The *corrected trapezoidal rule* approximates $I(f)$ by the area under the cubic Hermite that interpolates $(a,f(a),f'(a))$ and $(b,f(b), f'(b))$.
- Newton–Cotes rules are discussed in Sec. 9.1.1 below.
- Gaussian rules are discussed in Sec. 9.1.2 below.

The formula $Q(f)$ and the associated error $I(f) - Q(f)$ for each quadrature rule listed above are given in Table 13, where n is the number of function and derivative evaluations, d is the polynomial degree of the quadrature rule (see Sec. 9.1.1 below), η is an unknown point in $[a,b]$ (generally different for each rule), $m = (a + b)/2$ is the midpoint of the interval $[a,b]$, and C and K are some constants. For the derivation of these quadrature rules and their associated error formulas, see an introductory numerical methods text such as Conte and de Boor (1980); Dahlquist and Björck (1974); Johnson and Riess (1982); Stoer and Bulirsch (1980).

9.1.1 Some Definitions A quadrature rule that is based on integrating a polynomial interpolant is called an *interpolatory rule.* All the simple quadrature rules listed in Table 13 are interpolatory. Writing the polynomial interpolant in Lagrange form and integrating it, we see immediately that the weights w_i do not depend on the function f, but only on the abscissae $\{x_i : i = 1, \ldots, n\}$.

Quadrature rules are, in general, not ex-

Table 13. Simple quadrature rules.

Quadrature rule	n	d	Interpolant	$Q(f)$	$I(f) - Q(f)$
Rectangle	1	0	Constant	$(b - a)f(a)$	$\frac{(b-a)^2}{2} f'(\eta)$
Midpoint	1	1	Constant	$(b - a)f(m)$	$\frac{(b-a)^3}{24} f''(\eta)$
Trapezoidal	2	1	Linear	$\frac{b-a}{2}[f(a) + f(b)]$	$-\frac{(b-a)^3}{12} f''(\eta)$
Simpson's	3	3	Quadratic	$\frac{b-a}{6}[f(a) + 4f(m) + f(b)]$	$-\frac{(b-a)^5}{2880} f^{(4)}(\eta)$
Corrected trapezoidal	4	3	Cubic	$\frac{b-a}{2}[f(a) + f(b)] + \frac{(b-a)^2}{12}[f'(a) - f'(b)]$	$\frac{(b-a)^5}{720} f^{(4)}(\eta)$
Newton-Cotes	n	$\geq n - 1$	Deg. $n - 1$	See Sec. 9.1.1	$\frac{(b-a)^{d+2}}{K} f^{(d+1)}(\eta)$
Gaussian	n	$2n - 1$	Deg. $n - 1$	See Sec. 9.1.2	$\frac{(b-a)^{d+2}}{C} f^{(d+1)}(\eta)$

act. An error formula for an interpolatory rule can often be derived by integrating the associated polynomial interpolant error. Error formulas for some simple quadrature rules are listed in Table 13.

A quadrature rule that is exact for all polynomials of degree d or less, but is not exact for all polynomials of degree $d + 1$, is said to have *polynomial degree d.* An interpolatory rule based on n function and derivative values has polynomial degree at least $n - 1$.

Quadrature rules that include the end points of the interval of integration $[a,b]$ as abscissae are called *closed rules,* while those that do not include the end points are called *open rules.* An advantage of open rules is that they can be applied to integrals with singularities at the end points, whereas closed rules usually can not.

Interpolatory rules based on equidistant abscissae are called *Newton–Cotes* rules. This class includes the rectangle, midpoint, trapezoidal, corrected trapezoidal, and Simpson's rules. Both open and closed Newton–Cotes quadrature rules exist.

9.1.2 Gaussian Quadrature Rules A quadrature rule

$$Q(f) = \sum_{i=1}^{n} w_i f(x_i)$$

is fully determined by n, the abscissae $\{x_i : i = 1, \ldots, n\}$, and the weights $\{w_i : i = 1, \ldots, n\}$. Gauss showed that, given n and the end points a and b of the integral,

1. there exists a unique set of abscissae $\{x_i\}$ and weights $\{w_i\}$ that give a quadrature rule—called the Gaussian quadrature rule—that is exact for all polynomials of degree $2n - 1$ or less;
2. no quadrature rule with n abscissae and n weights is exact for all polynomials of degree $2n$;
3. the weights $\{w_i\}$ of the Gaussian quadrature rule are all positive;
4. the Gaussian quadrature rule is open;
5. the abscissae $\{x_i\}$ of the Gaussian quadrature rule are the roots of the shifted Legendre polynomial $q_n(x)$ of degree n, which is the unique monic polynomial of degree n that is orthogonal to all polynomials of degree $n - 1$ or less w.r.t. the continuous inner product

   $$(f,g) = \int_a^b f(x)g(x)dx$$

   (see Sec. 8.7.1);
6. the Gaussian quadrature rule is interpolatory, i.e., it can be derived by integrating the polynomial of degree $n - 1$ that interpolates f at the abscissae $\{x_i\}$.

Thus, Gauss derived the class of open interpolatory quadrature rules of maximum polynomial degree $d = 2n - 1$. As noted above, these formulas are called *Gaussian quadrature rules* or *Gauss–Legendre quadrature rules.*

9.1.3 Translating the Interval of Integration The weights and abscissae of simple quadrature rules are usually given w.r.t. a specific interval of integration, such as [0,1] or [−1,1]. However, the weights and abscissae can be transformed easily to obtain a related quadrature rule appropriate for some other interval.

One simple way to do this is based on the linear change of variables $\hat{x} = \beta(x - a)/(b - a) + \alpha(b - x)/(b - a)$, which leads to the relation

$$\int_\alpha^\beta f(\hat{x})d\hat{x} = \int_a^b f\left(\beta\frac{x - a}{b - a} + \alpha\frac{b - x}{b - a}\right) \times \frac{\beta - \alpha}{b - a}dx. \qquad (43)$$

So, if we are given a quadrature rule on $[a,b]$ with weights $\{w_i : i = 1, \ldots, n\}$ and abscissae $\{x_i : i = 1, \ldots, n\}$, but we want to compute

$$\int_\alpha^\beta f(\hat{x})d\hat{x},$$

we can apply the quadrature rule to the integral on the right side of (43). An equivalent way of viewing this is that we have developed a related quadrature rule for the interval of integration $[\alpha,\beta]$ with weights and abscissae

$$\hat{w}_i = \frac{\beta - \alpha}{b - a}w_i \quad \text{and}$$

$$\hat{x}_i = \beta\frac{x_i - a}{b - a} + \alpha\frac{b - x_i}{b - a}$$

for $i = 1, \ldots, n$. Note that, because we have used a linear change of variables, the original rule for $[a,b]$ and the related one for $[\alpha,\beta]$ have the same polynomial degree.

9.1.4 Comparison of Gaussian and Newton–Cotes Quadrature Rules

We list some similarities and differences between Gaussian and Newton–Cotes quadrature rules:

1. The weights of a Gaussian rule are all positive, which contributes to the stability of the formula. High-order Newton–Cotes rules typically have both positive and negative weights, which is less desirable, since it leads to poorer stability properties.
2. Gaussian rules are open, whereas there are both open Newton–Cotes rules and closed Newton–Cotes rules.
3. Gaussian rules attain the maximum possible polynomial degree $2n - 1$ for a formula with n weights and abscissae, whereas the polynomial degree d of a Newton–Cotes rule satisfies $n - 1 \leq d \leq 2n - 1$ and the upper bound can be obtained for $n = 1$ only.
4. The abscissae and weights of Gaussian rules are often irrational numbers and hard to remember, but they are not difficult to compute. The abscissae of a Newton–Cotes rule are easy to remember and the weights are simple to compute.
5. The set of abscissae for an n-point Gaussian rule and for an m-point Gaussian rule are almost disjoint for all $n \neq m$. Thus we can not reuse function evaluations performed for one Gaussian rule in another Gaussian rule. Appropriately chosen pairs of Newton–Cotes rules can share function evaluations effectively.
6. Both Gaussian and Newton–Cotes rules are interpolatory.

9.2 Composite (Compound) Quadrature Rules

To increase the accuracy of a numerical approximation to an integral, we could use a rule with more weights and abscissae. This often works with Gaussian rules, if f is sufficiently smooth, but it is not advisable with Newton–Cotes rules, for example, because of stability problems associated with high-order formulas in this class.

Another effective way to achieve high accuracy is to use *composite* quadrature rules, often also called *compound* quadrature rules. In these schemes, the interval of integration $[a,b]$ is subdivided into *panels* (or subintervals) and on each panel the same simple quadrature rule is applied. If the associated simple quadrature rule is interpolatory, then this approach leads to the integration of a piecewise polynomial (PP) interpolant. Thus, using a composite quadrature rule, instead of a simple one with high polynomial degree, leads to many of the same benefits that are obtained in using PP interpolants compared with high-degree polynomial interpolants (see Sec. 8.5).

Table 14 summarizes the composite quadrature rules and the associated error formulas. In the table, n is the number of function and derivative evaluations, d is the polynomial degree of the quadrature rule, η is an unknown point in $[a,b]$ (in general different for each rule), $h = (b - a)/s$ is the step size of each panel, s is the number of panels, and PP stands for piecewise polynomial. Composite quadrature rules based on Gaussian or Newton–Cotes rules can also be used, although they are not listed in Table 14.

9.3 Adaptive Quadrature

We see from Table 14 of composite quadrature rules that the smaller the step size h of a panel, the smaller the expected error. It is often the case that an approximation to the integral

$$I(f) = \int_a^b f(x)dx$$

is needed to within a specified accuracy ϵ. In this case, *adaptive quadrature* is often used. Such schemes refine the grid (or collection of panels) until an estimate of the total error in the integration is within the desired precision ϵ. Adaptive quadrature is particularly useful when the behavior of the function f varies significantly in the interval of integration $[a,b]$, since the scheme can use a fine grid where f is hard to integrate and a coarse grid where it is easy, leading to an efficient and accurate quadrature procedure.

Table 15 gives a simple general recursive procedure for adaptive quadrature. We as-

Table 14. Composite quadrature rules.

Quadrature Rule	n	d	PP interpolant	Formula	Error
Rectangle	s	0	Constant	$h \sum_{i=0}^{s-1} f(a + ih)$	$\frac{h}{2}(b - a)f'(\eta)$
Midpoint	s	1	Constant	$h \sum_{i=1}^{s} f[a + (i - 1/2)h]$	$\frac{h^2}{24}(b - a)f''(\eta)$
Trapezoidal	$s + 1$	1	Linear	$\frac{h}{2}[f(a) + f(b) + 2\sum_{i=1}^{s-1} f(a + ih)$	$-\frac{h^2}{12}(b - a)f''(\eta)$
Simpson's	$2s + 1$	3	Quadratic	$\frac{h}{6}\left[f(a) + f(b) + 2\sum_{i=1}^{s-1} f(a + ih) + 4\sum_{i=1}^{s} f(a + (i - 1/2)h)\right]$	$-\frac{h^4}{2880}(b - a)f^{(4)}(\eta)$
Corrected trapezoidal	$s + 3$	3	Cubic Hermite	$\frac{h}{2}\left[f(a) + f(b) + 2\sum_{i=1}^{s-1} f(a + ih)\right] + \frac{h^2}{12}[f'(a) - f'(b)]$	$\frac{h^4}{720}(b - a)f^{(4)}(\eta)$

Table 15. Adaptive quadrature procedure.

```
subroutine $AQ(a,b,\epsilon)$
    $(Q,E) = LQM(a,b)$
    if ($E \le \epsilon$) then
        return $(Q,E)$
    else
        $m = (a + b)/2$
        return $AQ(a,m,\epsilon/2) + AQ(m,b,\epsilon/2)$
    end
end
```

sume that we can make use of a routine LQM (Local Quadrature Module) that implements a quadrature rule in some interval $[a,b]$ and returns Q, an approximation to the integral, and E, an estimate of the error. In the next section, we discuss how an error estimate may be obtained.

We note that the adaptive quadrature procedure shown in Table 15 does not illustrate how to reuse function evaluations where possible. An effective adaptive quadrature routine should do this, since function evaluations are often the most computationally expensive part of the procedure.

9.4 Romberg Integration and Error Estimation

As discussed in the last subsection, adaptive quadrature requires an error estimate. As an illustration, we consider how one may be obtained for the composite trapezoidal rule.

Let $T_s(f)$ denote the composite-trapezoidal-rule approximation to

$$I(f) = \int_a^b f(x)dx$$

using s panels and let $E_s = I(f) - T_s(f)$ be the associated error. On the basis of the error formula in Table 13, we expect E_{2s} to be about four times smaller than E_s, assuming that $f''(x)$ does not vary too much. That is,

$$I(f) - T_s(f) = E_s,$$

$$I(f) - T_{2s}(f) = E_{2s} \approx \tfrac{1}{4}E_s.$$

Subtracting these two equations, we get

$$T_{2s}(f) - T_s(f) = E_s - E_{2s} \approx \tfrac{3}{4}E_s.$$

So $E_s \approx 4[T_{2s}(f) - T_s(f)]/3$ and $E_{2s} \approx [T_{2s}(f) - T_s(f)]/3$. Thus, by applying the composite trapezoidal rule first with s and then with $2s$ panels, we obtain estimates of the error in both $T_s(f)$ and $T_{2s}(f)$.

If we add the estimate of the error $E_s = I(f) - T_s(f)$ to $T_s(f)$ we often obtain a better approximation to $I(f)$ than either $T_s(f)$ or $T_{2s}(f)$. That is, $\hat{T}_s(f) = T_s(f) + 4[T_{2s}(f) - T_s(f)]/3 = [4T_{2s}(f) - T_s(f)]/3$ is often a better approximation to $I(f)$ than either $T_s(f)$ or $T_{2s}(f)$, particularly if the function f is smooth and the grid is fine. Thus, by applying the compound trapezoidal rule with s and $2s$ panels and taking an appropriate linear combination of the two approximations, we construct a better approximation to $I(f)$. To be more specific, it can be shown that this eliminates the lowest-order term in the error E_s or E_{2s}. This process can be repeated to eliminate the next higher-order term in the error and so on. In addition, it can be generalized easily to other quadrature rules.

This is the basic idea behind *Romberg integration.* By applying a quadrature rule repeatedly with more panels each time, we can eliminate the leading terms of the error expansion, and thereby obtain better and better approximations to $I(f)$.

9.5 Infinite Integrals and Singularities

If one or both end points of the interval of integration $[a,b]$ are infinite, the integral is called infinite. We restrict the discussion of infinite integrals to the case

$$I(f) = \int_a^\infty f(x)dx,$$

sometimes called a semi-infinite integral, since only one end point is infinite. Other cases can be handled similarly.

Under the assumption that

$$I(f) = \int_a^\infty f(x)dx$$

exists, one way to approximate the integral is to *truncate* $I(f)$, and compute instead

$$\hat{I}(f) = \int_a^b f(x)dx,$$

for some sufficiently large b, by a standard quadrature rule $Q(f)$. The error in this approach is $I(f) - Q(f) = [I(f) - \hat{I}(f)] + [\hat{I}(f) - Q(f)]$. It is often possible to choose b so that

$$I(f) - \hat{I}(f) = \int_b^\infty f(x)dx$$

is small and to choose Q so that $\hat{I}(f) - Q(f)$ is also small.

$I(f)$ can also be approximated by first performing a change of variables to *transform* the infinite integral to a standard one. More specifically, let $x = g(t)$. Then

$$I(f) = \int_a^\infty f(x)dx = \int_{g^{-1}(a)}^{g^{-1}(\infty)} f(g(t))\cdot g'(t)dt,$$

where $g^{-1}(x)$ is the inverse of $g(x)$. If we can choose g so that $g^{-1}(a)$ and $g^{-1}(\infty)$ are both finite, then $I(f)$ is transformed to a finite integral that can be evaluated by a standard quadrature rule. However, this procedure may introduce singularities (discussed below). If so, it might not lead to a computationally easier problem to solve.

Infinite integrals can also be approximated by special quadrature formulas that are directly applicable to infinite intervals of integration. For further details concerning this approach, see an introductory numerical methods text such as Conte and de Boor (1980), Dahlquist and Björck (1974), Johnson and Riess (1982), or Stoer and Bulirsch (1980).

A singular integral

$$I(f) = \int_a^b f(x)dx$$

is one in which f is singular (i.e., becomes infinite) at some point in $[a,b]$. Singular and infinite integrals are closely related: A change of variables often transforms one into the other.

In the computing of singular integrals by a quadrature rule, the value of f might be required at or close to a point of singularity, and so the quadrature rule may be either inapplicable or inaccurate. It often happens that the singularity in f occurs at the end point a or b, in which case, an open formula, such as a Gaussian rule, may be effective.

The performance of a quadrature rule applied to a singular integral might be improved by a change of variables. A transformation $x = g(t)$ that often helps to remove or lessen the effect of a singularity is $g(t) = b - (b - a)u^2(2u + 3)$ for $u = (t - b)/(b - a)$.

9.6 Monte-Carlo Methods

Monte-Carlo methods (*q.v.*) are of a statistical nature. For the sake of simplicity, we briefly present them for one-dimensional integrals only, but they can be extended easily to multidimensional integrals and are most useful in this context.

To begin, choose n random points $\{U_i : i = 1, \dots, n\} \subset [0,1]$ and scale each U_i to $[a,b]$ by $u_i = a + U_i(b - a)$. Then

$$Q_n(f) = \frac{(b-a)}{n} \sum_{i=1}^{n} f(u_i)$$

is a Monte-Carlo approximation to

$$I(f) = \int_a^b f(x)dx.$$

If we consider $Q_n(f)$ to be a random variable, then it can be shown that its mean is $I(f)$ and its standard deviation is $|b - a| \times \rho(f)/\sqrt{n}$, where $\rho(f)$ is a constant that depends on f, but not n. Assuming that $Q_n(f)$ is close to being normally distributed, we are led to statistical statements about the error, such as that $|Q_n(f) - I(f)| \leq 2|b - a|\rho(f)/\sqrt{n}$ nineteen times out of twenty.

Note that the error bound above decreases like $1/\sqrt{n}$, much more slowly than the bounds for the standard compound quadrature rules given in Table 14. This suggests that Monte-Carlo methods are not very effective for one-dimensional integrals of smooth functions. However, an error formula similar to that given above continues to hold for multidimensional integrals, while extensions of standard methods become increasingly less efficient as the dimension of the integral to be approximated increases. As a result, Monte-Carlo methods are among the best schemes available for approximating high-dimensional integrals.

10. ORDINARY DIFFERENTIAL EQUATIONS

In this section, we consider numerical methods for the solution of ordinary differential equations (ODEs). We begin by introducing some simple schemes for the initial-value problem (IVP)

$$y'(x) = f(x,y(x)) \quad x \in [a,b],$$
$$y(a) = y_0, \tag{44}$$

where $y : \mathbb{R} \to \mathbb{R}^m$ and $f : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$. We assume throughout that the IVP (44) has a unique solution for $x \in [a,b]$. We also discuss more sophisticated adaptive methods and explain the terms *stiff* and *nonstiff* problems and how to choose methods appropriate for these two classes of problems. We then briefly consider the boundary-value problem (BVP)

$$y'(x) = f(x,y(x)) \quad x \in [a,b],$$
$$g(y(a), y(b)) = 0, \tag{45}$$

which we assume has a locally unique solution.

For both IVPs and BVPs, we consider systems of first-order ODEs only, since most commonly available codes are for first-order systems and any higher-order ODE can be reduced to a system of first-order equations. However, using a method designed for higher-order equations directly may lead to a more efficient solution of the problem.

Because of space constraints, we do not discuss many important related problems such as differential-algebraic equations or delay differential equations. For a more comprehensive discussion of these and other related topics, see an advanced text such as Ascher *et al.* (1988), Butcher (1987), Hairer *et al.* (1987), Hairer and Wanner (1991), Lambert (1991), or Shampine (1994).

10.1 Initial-Value Problems (IVPs)

10.1.1 Two Simple Formulas Most standard numerical methods for the IVP (44) start with the initial value y_0 at $x_0 = a$ and then compute approximations $y_n \approx y(x_n)$ for $n = 1, \dots, N$ on a discrete grid $a = x_0 < x_1 < \dots < x_N = b$. The distance between adjacent grid points, $h_n = x_{n+1} - x_n$, $n =$

$0, \ldots, N-1$, is referred to as the *step size* at step $n+1$. Schemes are often presented with a constant step size $h = h_n$ for all n, but this is generally not required. Moreover, as discussed below, variable–step-size methods are often much more efficient.

Possibly the simplest numerical scheme for (44) is *Euler's method,* sometimes called the *forward Euler method:*

$$y_{n+1} = y_n + h_n f(x_n, y_n). \tag{46}$$

This formula is motivated from the observation that the true solution of a scalar IVP of the form (44) satisfies

$$\begin{aligned} y(x_{n+1}) &= y(x_n) + h_n y'(x_n) + \frac{h_n^2}{2} y''(\eta_n) \\ &= y(x_n) + h_n f(x_n, y(x_n)) + \frac{h_n^2}{2} y''(\eta_n) \end{aligned} \tag{47}$$

for some point $\eta_n \in [x_n, x_{n+1}]$, which follows from standard Taylor-series theory. Thus, the true solution of the IVP (44) satisfies an equation that is very similar to (46). This argument can be extended easily to systems.

The approximations $y_n \approx y(x_n)$ are computed in the order $n = 1, \ldots, N$ using the formula (46). To be more specific, on the first step of Euler's method from x_0 to $x_1 = x_0 + h_0$, we substitute the initial value (x_0, y_0) into the right side of (46) to compute $y_1 \approx y(x_1)$. Thus, at the end of the first step, we have (x_1, y_1). On the second step of Euler's method from x_1 to $x_2 = x_1 + h_1$, we substitute (x_1, y_1) into the right side of (46) to compute $y_2 \approx y(x_2)$. The procedure continues in a similar way for $n = 2, \ldots, N$. Note that this evaluation process applies equally well to systems of equations (i.e., $m > 1$). In this case, h_n is a scalar, but y_{n+1}, y_n and $f(x_n, y_n)$ are all m-vectors.

Euler's method is an *explicit formula* in the sense that the evaluation process described above does not require the solution of any linear or nonlinear algebraic equations. The backward Euler formula

$$y_{n+1} = y_n + h_n f(x_{n+1}, y_{n+1}) \tag{48}$$

is a typical example of an *implicit formula.* It can be motivated from a Taylor-series expansion of the true solution $y(x)$ of the IVP (44) about x_{n+1} similar to the expansion (47) above for $y(x)$ about x_n. Moreover, we again compute the approximations $y_n \approx y(x_n)$ in the order $n = 1, \ldots, N$. However, note that, on step $n+1$ from x_n to $x_{n+1} = x_n + h_n$, we start with (x_n, y_n) and must solve Eq. (48) for y_{n+1}.

We will return to the question of how to solve for y_{n+1} shortly, but first we explain briefly in the next subsection why we may wish to use an implicit scheme (such as the backward Euler formula) rather than an explicit one (such as the forward Euler formula) even though the former clearly requires more work per step than the latter.

10.1.2 Stiff IVPs Roughly speaking, a stiff IVP is one in which some terms in the solution decay rapidly with respect to the length of the integration, while others vary slowly on this time scale. To illustrate this concept, consider the linear constant-coefficient problem $y' = Ay$, $y(0) = y_0$, for $x \in [0,1]$, where

$$A = \frac{1}{2}\begin{pmatrix} -10^6 - 1 & 10^6 - 1 \\ 10^6 - 1 & -10^6 - 1 \end{pmatrix}, \quad y_0 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

If we let $z = Py$, where

$$P = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \quad P^{-1} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix},$$
$$D = PAP^{-1} = \begin{pmatrix} -10^6 & 0 \\ 0 & -1 \end{pmatrix},$$

then $z' = Py' = PAy = PAP^{-1}z = Dz$ and $z(0) = Py(0) = (1,1)^T$. Therefore, $z(x) = (e^{-10^6 x}, e^{-x})^T$ and so $y(x) = P^{-1}z(x) = (e^{-x} + e^{-10^6 x}, e^{-x} - e^{-10^6 x})^T$. The term $e^{-10^6 x}$ that occurs in both $z(x)$ and $y(x)$ gives rise to an initial transient that decays rapidly with respect to the interval of integration [0,1], while the term e^{-x} is associated with a slowly varying smooth term on this scale. The fast and slow terms occur in different components of $z(x)$, but they are mixed in $y(x)$, which is often the case in practice. After the $e^{-10^6 x}$ term dies out, both components of $y(x)$ vary smoothly like e^{-x}.

If we apply the forward Euler formula to $y' = Ay$, we get $y_{n+1} = y_n + h_n A y_n = (I + h_n A)y_n$. If we multiply the last equation through by P and perform the change of variables $z_n = Py_n$, we get $z_{n+1} = (I + h_n D)z_n$. The two components of z_n satisfy $z_{n+1}^{(1)} = (1 - 10^6 h_n)z_n^{(1)}$ and $z_{n+1}^{(2)} = (1 - h_n)z_n^{(2)}$. If we let $h_n > 2 \times 10^{-6}$ at any point during the integration, then $(1 - 10^6 h_n) < -1$ and $z_n^{(1)}$ will grow in magnitude and oscillate in sign as n increases. Since $y_n = P^{-1} z_n$, this will cause both components of y_n to oscillate about e^{-x_n} with growing amplitude as n increases. Since this instability is undesirable, we must restrict $h_n < 2 \times 10^{-6}$ throughout the integration, even though, after the initial transient, this is likely a much smaller step size than would be required to integrate the slowly varying e^{-x} component of the solution accurately.

On the other hand, if we apply the backward Euler formula to $y' = Ay$, we get $y_{n+1} = y_n + h_n A y_{n+1}$ and so $y_{n+1} = (I - hA)^{-1} y_n$. If we multiply this equation through by P and perform the change of variables $z_n = P y_n$, we get $z_{n+1} = (I - h_n D)^{-1} z_n$. Therefore, $z_{n+1}^{(1)} = z_n^{(1)}/(1 + 10^6 h_n)$ and $z_{n+1}^{(2)} = z_n^{(2)}/(1 + h_n)$. Hence, no matter how large $h_n > 0$ is, $z_n^{(1)}$ decays as n increases. Consequently, after the initial transient, we can choose $h_n > 0$ to integrate $z_n^{(2)}$ accurately without fear of $z_n^{(1)}$ becoming unstable. Since $y_n = P^{-1} z_n$, the same conclusion applies to y_n.

The example above can be generalized to larger systems of equations $y' = Ay$. If A is diagonalizable, then the performance of the method on $y' = Ay$ can be deduced from its performance on the scalar test problems $y' = \lambda y$, where the λ's range over the eigenvalues of A. If the real part of each λ is negative, then $y(x) \to 0$ as $x \to \infty$. We would like the numerical solution to have the same behavior without having to restrict h_n outside the transient region. Methods with this property are called *A-stable*. Generalizing the example above, it is easy to see that the backward Euler formula is A-stable, while the forward Euler formula is not.

The importance of the example above and the scalar test problem $y' = \lambda y$ in particular is that the performance of methods on these simple problems is indicative of their behavior on more general nonlinear stiff problems. A nonrigorous, but intuitive, justification of this follows from the local linearization of $y' = f(x,y)$ at (x_n, y_n):

$$y'(x) \approx f(x_n,y_n) + f_x(x_n,y_n)(x - x_n) + f_y(x_n,y_n)(y - y_n),$$

where $f_y(x,y) = \partial f(x,y)/\partial y \in \mathbb{R}^{m \times m}$ is the Jacobian of f. This problem is usually stiff if

1. some eigenvalue of $f_y(x_n,y_n)$ has a large negative real part with respect to the interval of integration,
2. no eigenvalue of $f_y(x_n,y_n)$ has a large positive real part with respect to the interval of integration, and
3. no eigenvalue of $f_y(x_n,y_n)$ has a large imaginary part unless it also has a relatively large negative real part.

An IVP that is not stiff is called *nonstiff*.

Stiff IVPs arise in many applications, such as chemical kinetics and electrical circuits. As noted earlier, they are characterized by components that vary on vastly different time scales: Some terms in the solution decay rapidly to steady state while others vary slowly.

The observation above that the explicit forward Euler formula is not appropriate for a stiff problem, while the implicit backward Euler formula is, can be generalized. All commonly used formulas that are suitable for stiff problems are implicit in some sense.

See the survey article of Shampine and Gear (1979) or an advanced text such as Butcher (1987), Hairer and Wanner (1991), Lambert (1991), or Shampine (1994) for a more detailed discussion of stiffness.

10.1.3 Solving Implicit Equations We return now to methods for solving for y_{n+1} in an implicit formula such as (48). One common approach is the predictor–corrector technique, which is just a fixed-point iteration as described in Sec. 6.1. For this scheme (and most others), we need in initial approximation $y_{n+1}^{(0)}$ to y_{n+1}. This could be computed, for example, from the forward Euler formula or some other explicit scheme, or simply by taking $y_{n+1}^{(0)} = y_n$. In the terminology of predictor–corrector techniques, the formula for computing $y_{n+1}^{(0)}$ is referred to as the *predictor* formula. For a predictor–corrector method based on the backward Euler formula (48), the corrector formula would be

$$y_{n+1}^{(l+1)} = y_n + h_n f(x_{n+1}, y_{n+1}^{(l)}), \quad l = 0, 1, \ldots . \qquad (49)$$

We substitute $y_{n+1}^{(0)}$ into the right side of (49) to compute $y_{n+1}^{(1)}$, which we in turn substitute into the right side of (49) to compute $y_{n+1}^{(2)}$, and so on. It is easy to show that $y_{n+1}^{(l)} \to y_{n+1}$ as $l \to \infty$ if f satisfies the Lipschitz condition

$$\|f(x_{n+1},y) - f(x_{n+1},z)\| \leq L\|y - z\|$$

for some constant L and all y, z in a convex domain containing y_{n+1} and $y_{n+1}^{(l)}$ for $l = 0, 1, \ldots$ and $h_nL < 1$. In most codes, one or two corrections only are needed, since the initial guess $y_{n+1}^{(0)}$ is normally a good approximation to y_{n+1}. Consequently, this scheme is not much more expensive to implement than the explicit forward Euler formula. However, a predictor–corrector implementation of an A-stable method (such as the backward Euler formula) is not A-stable. On the contrary, because of the requirement $h_nL < 1$, it will suffer a step size restriction on a stiff problem similar to that of an explicit formula.

Alternatively, we could rewrite the backward Euler formula (48) as

$$F(y) = y - y_n - h_n f(x_{n+1},y) = 0, \qquad (50)$$

where we have replaced the unknown y_{n+1} by y, and then apply one of the other techniques described in Sec. 6 for finding roots of equations to compute the solution $y = y_{n+1}$ of (50). The most commonly used root-finding technique in this context is Newton's method—or a variant of it. As noted in Sec. 6.7, for systems of equations, Newton's method takes the form

$$\begin{aligned}[I - h_n f_y(x_{n+1},y_{n+1}^{(l)})]\Delta_l \\ = y_n + h_n f(x_{n+1},y_{n+1}^{(l)}) - y_{n+1}^{(l)}, \qquad (51)\end{aligned}$$

$$y_{n+1}^{(l+1)} = y_{n+1}^{(l)} + \Delta_l, \qquad (52)$$

where $f_y(x,y) = \partial f(x,y)/\partial y \in \mathbb{R}^{m\times m}$ is the Jacobian of f. Note that we must solve a linear system of m equations in m unknowns to compute the Newton update vector Δ_l in (51). Typically, Gaussian elimination with partial pivoting (see Sec. 2.5) is used to solve such linear systems. A band or sparse solver (see Sec. 2.7) may dramatically decrease the cost of solving (51) if $I - h_n f_y(x_{n+1},y_{n+1}^{(l)})$ is large and sparse. Similarly, iterative methods, such as the preconditioned conjugate-gradient method (see Sec. 3.2), may significantly reduce the cost of solving some large sparse problems. See Sec. 13 for a discussion of sources of high-quality numerical software, including stiff-ODE solvers that incorporate sparse and iterative linear equation solvers.

The computational work required to solve (51) can be decreased significantly by using a chord Newton method, often called a simplified Newton method, that holds the Newton iteration matrix $I - h_n f_y(x_{n+1},y_{n+1}^{(l)})$ constant over several iterations and possibly several steps of the integration, thus avoiding the necessity to factor the Newton iteration matrix on each iteration (see Sec. 6.8). However, even with this savings, the cost per step of the Newton iteration may be much larger than a predictor–corrector method. However, it has the advantage for formulas appropriate for stiff problems that it avoids the step size restriction associated with the predictor–corrector technique or explicit formulas. Thus, even though the Newton iteration might make the scheme much more expensive per step, the step sizes that can be used might be so much larger that the total cost of the integration is significantly less. Finally note that, as for predictor–corrector methods, an initial guess for $y_{n+1}^{(0)}$ is required. It can be computed by the techniques described above.

10.1.4 Higher-Order Formulas A numerical method for ODEs is said to be *of order p* or *p*th *order* or *p*th-*order convergent* if $y_n = y(x_n) + O(h^p)$ for some integer $p \geq 1$, where $O(h^p)$ is any quantity (in this case, the global error) that can be bounded by h^p times a constant that is independent of h, but that may depend on the IVP and the numerical method. Most standard texts on the numerical solution of ODEs show that both the forward and backward Euler methods are first-order convergent.

Higher-order methods are frequently used in practice because they offer the potential of significantly reducing the computational work required to generate an accurate solution to the IVP (44). To get an intuitive feeling for this, suppose that the length of integration $b - a = 1$, that we use a constant step size h throughout the numerical integration, that the global error for a pth-order method satisfies $y_n - y(x_n) = h^p$, and that we require this error to be of size 10^{-10}. Under these assumptions, the optimal step size for the method is $h = 10^{-10/p}$, and the resulting number of steps needed to integrate from a to b is $N = 10^{+10/p}$. To be more specific, for $p = 1, 2, 5, 10$, the number of steps required is $N = 10^{10}, 10^5, 10^2, 10^1$, respectively. Thus, even though a higher-order method may require more work per step than a lower-order scheme, the dramatic reduction in the number of steps required frequently makes a higher-order method much more efficient than a lower-order one—particularly for problems with stringent error tolerances.

Two common second-order formulas are the *trapezoidal rule*

$$y_{n+1} = y_n + \tfrac{1}{2}h_n[f(x_n,y_n) + f(x_{n+1},y_{n+1})] \tag{53}$$

and the *implicit midpoint rule*

$$y_{n+1} = y_n + f(x_n + h_n/2, (y_n, + y_{n+1})/2), \tag{54}$$

each of which is implicit, since one clearly needs to solve for y_{n+1}. Note that neither formula requires much more work per step than the backward Euler formula. Moreover, both formulas are A-stable and effective for solving stiff problems at relaxed error tolerances.

10.1.5 Runge–Kutta Formulas Runge–Kutta (RK) formulas are a general class of methods containing many higher-order schemes. The general form of an s-stage RK formula is

$$\begin{aligned} k_i &= f\Big(x_n + c_i h_n, y_n + h_n \sum_{j=1}^{s} a_{ij}k_j\Big), \qquad i = 1, \dots, s, \\ y_{n+1} &= y_n + h_n \sum_{i=1}^{s} b_i k_i. \end{aligned} \tag{55}$$

That is, we must first compute the s function values k_i and then form a weighted average of the k_i's to compute y_{n+1} from y_n.

RK formulas are one-step schemes in the sense that all the information required to compute y_{n+1} from y_n is generated on the current step from x_n to x_{n+1}. That is, unlike multistep formulas discussed in the next subsection, a RK formula does not require any information from past steps.

If the stages of the RK formula can be ordered so that $a_{ij} = 0$ for all $j \geq i$, then the formula (55) is explicit in the sense that the k_i's can be computed in the order $i = 1, \dots, s$ without having to solve any linear or nonlinear equations. In what follows, we assume that the formula has been so ordered if possible. If the RK formula (55) is not explicit, then it is implicit and at least one linear or nonlinear equation must be solved to compute the k_i's.

The coefficients of a RK formula are frequently displayed in a tableau as

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \dots & b_s \end{array}$$

If all the elements in the tableau on the diagonal and above it are zero, then the associated formula is explicit.

All the methods considered so far are in fact RK formulas. The RK tableaux for the forward Euler formula, backward Euler formula, implicit midpoint rule, and trapezoidal rule are listed below in that order:

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}, \qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}, \qquad \begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array},$$

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}.$$

Note that the first three are one-stage RK formulas and the final one, the trapezoidal rule, is a two-stage RK formula. The tableau for the classical four-stage fourth-order explicit RK formula is

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

This formula has been widely used since it was published by Kutta in 1901. In the days of hand calculation, the zero coefficients below the diagonal were a distinct benefit, but this is no longer a significant advantage on a modern computer. Moreover, there are now many better formulas, both of order four and of higher order. The interested reader should consult an advanced text such as Butcher (1987), Hairer *et al.* (1987), Hairer and Wanner (1991), Lambert (1991), or Shampine (1994) for further details.

As the sample formulas above suggest, a high-order RK formula requires more stages than a low-order one. The minimum number of stages that an explicit RK formula requires to attain orders 1 to 8 are listed below.

Order	1	2	3	4	5	6	7	8
Stages	1	2	3	4	6	7	9	11

On the other hand, implicit s-stage RK formulas of order $2s$ exist for all $s \geq 1$. Moreover, it can be shown that this is the maximal order possible.

Explicit RK formulas are frequently used to solve nonstiff IVPs. Some implicit RK formulas are A-stable (or nearly so) and are suitable for solving stiff IVPs. Formulas with four or fewer stages are quite effective for problems with relaxed error tolerances, while formulas with five or more stages are suitable for problems with more stringent accuracy requirements. Since RK formulas are one-step schemes, unlike linear multistep formulas (LMFs) discussed in the next subsection, RK formulas are more suitable than LMFs for problems that require rapid changes in step size, such as problems with discontinuities. See Sec. 13 for a discussion of sources of high-quality numerical software, including routines based on RK formulas.

10.1.6 Linear Multistep Formulas

Linear multistep formulas (LMFs) can be written in the form

$$y_{n+1} = \sum_{i=1}^{k} \alpha_i y_{n+1-i} + h \sum_{i=0}^{k} \beta_i f(x_{n+1-i}, y_{n+1-i}), \qquad (56)$$

where we assume that at least one of α_k or β_k is nonzero [otherwise we can reduce k in (56)]. Formula (56) is in fact a k-step method, since it uses values over k steps to compute y_{n+1}. Therefore, we assume that, at the start of step $n + 1$, we have $y_{n+1-k}, \ldots, y_n$ and our task is to compute y_{n+1} by (56). In this case, if $\beta_0 = 0$, then the $f(x_{n+1}, y_{n+1})$ term can be dropped from the right side of (56), and so we can evaluate the right side of (56) to compute y_{n+1} without having to solve any linear or nonlinear equations. That is, if $\beta_0 = 0$, the formula is explicit. On the other hand, if $\beta_0 \neq 0$, then y_{n+1} occurs on both sides of (56), and so a linear or nonlinear equation must be solved to compute y_{n+1}. Therefore, the formula is implicit. Depending on the context, a predictor–corrector method or some variant of Newton's method is typically used to solve for y_{n+1}.

Of course, at the start of the integration, $n = 0$ and $y_{1-k}, \ldots, y_{-1}$ are typically not available for $k > 1$. One solution to this problem is to compute $y_1, \ldots, y_{k-1}$ by a one-step method (such as a RK formula) of the same order as the k-step LMF and then start using the k-step LMF at step k. Alternatively, we could use a one-step LMF on step 1, a two-step LMF on step 2, and so on, until we reach step k, after which we can use (56) on that step and all subsequent steps. Most LMF codes employ the latter strategy and adjust the step size so that the accuracy obtained by the formulas with smaller k is comparable to that obtained by formulas with larger k.

If we ignore stability, then it is possible to obtain a k-step LMF of order $2k$ for any $k \geq 1$. Moreover, it is easy to show that this is the maximal order possible. However, these maximal-order formulas are unstable for $k \geq 3$ and so are not useful in practice. It can be shown that, for any $k \geq 1$, the maximal order of a stable k-step LMF is $k + 1$ if k is odd and $k + 2$ if k is even.

We have presented the LMFs in this section for fixed step size only, since variable–step-size formulas are considerably more complicated. However, a variable–step-size variable-order scheme may be far more efficient in practice. Such schemes are discussed in advanced texts on the numerical solution of ODEs, such as Hairer *et al.* (1987), Hairer and Wanner (1991), Lambert (1991), and Shampine (1994), but not usually in introductory numerical methods books.

10.1.7 Adams Formulas

Adams formulas are a subclass of LMFs that have the form

$$y_{n+1} = y_n + h \sum_{i=0}^{k} \beta_i f(x_{n+1-i}, y_{n+1-i}). \qquad (57)$$

In the explicit Adams–Bashforth formulas, β_0

$= 0$ and the remaining k β_i's are chosen to obtain the maximal possible order k. In the implicit Adams–Moulton formulas, $\beta_0 \neq 0$ and the $k + 1$ β_i's are chosen to obtain the maximal possible order $k + 1$. These coefficients are listed in most advanced texts on numerical methods for ODEs and in many introductory numerical methods books. It turns out that the forward Euler formula is the one-step Adams–Bashforth formula, and the trapezoidal rule is the one-step Adams–Moulton formula. Moreover, note that the order of the Adams–Moulton formulas is optimal for k odd and nearly optimal for k even.

The implicit Adams–Moulton formulas have somewhat better numerical characteristics than the explicit Adams–Bashforth formulas. Consequently, Adams formulas are usually implemented in a predictor–corrector fashion, with the k- or $(k + 1)$-step Adams–Bashforth formula used for the predictor and a k-step Adams–Moulton formula used for the corrector. Adams predictor–corrector schemes are the basis for several very effective variable–step-size variable-order codes for nonstiff IVPs. See Sec. 13 for a discussion of sources of high-quality numerical software, including routines based on Adams formulas.

10.1.8 Backward Differentiation Formulas Backward differentiation formulas (BDFs), sometimes called *Gear formulas,* are another subclass of LMFs that have the form

$$y_{n+1} = \sum_{i=1}^{k} \alpha_i y_{n+1-i} + h\beta_0 f(x_{n+1}, y_{n+1}), \tag{58}$$

where $\beta_0 \neq 0$ and so the BDFs are implicit. The $k + 1$ coefficients of a k-step BDF are chosen to obtain the maximal possible order k. However, the BDFs are stable for $1 \le k \le 6$ only. They are A-stable for $k = 1$ and 2 and nearly A-stable for $k = 3$, 4, and 5, with the loss of A-stability increasing with k. For $k = 6$ the loss of A-stability increases to such an extent that this formula is frequently excluded from use.

The coefficients for the BDFs are listed in most advanced texts on numerical methods for ODEs and in many introductory numerical methods books. It turns out that the backward Euler formula is the one-step BDF.

Because the BDFs are usually used to solve stiff problems, the implicit equation is normally solved by Newton's method or some variant of this root-finding scheme.

BDFs are the basis for several very effective variable–step-size variable-order codes for stiff IVPs. See Sec. 13 for a discussion of sources of high-quality numerical software, including routines based on BDFs.

10.1.9 Other Methods Taylor-series methods and extrapolation schemes are two other classes of formulas that are sometimes used in practice, but much less frequently than Runge–Kutta or linear multistep formulas. See an advanced text such as Butcher (1987), Hairer *et al.* (1987), Hairer and Wanner (1991), Lambert (1991), or Shampine (1994) for a discussion of these and other classes of methods.

10.1.10 Adaptive Methods Most good programs for the numerical solution of ODEs vary their step size—and possibly their order—in an attempt to solve the problem as efficiently as possible subject to a user-specified error tolerance, *tol.* The error that is controlled is usually the local error on each step, rather than the global error $y_n - y(x_n)$ that the user might at first expect. However, in most good programs the global error is at least roughly proportional to *tol,* so that reducing *tol* usually reduces the global error. A few codes report an estimate of the global error as well. If such an estimate is available, it is often optional, since estimating the global error frequently increases the cost of the integration significantly.

A useful way to interpret *tol* and the associated local error is in the *backward error* sense. (See Sec. 2.8 for a discussion of backward error analysis in the context of solving linear algebraic systems $Ax = b$.) When called to solve the IVP (44), many good programs generate a numerical solution that is the exact solution of the slightly perturbed problem

$$z'(x) = f(x, z(x)) + \delta(x), \quad z(a) = y_0, \tag{59}$$

where $\|\delta(x)\| \lessapprox tol$. A few codes compute $\delta(x)$ explicitly and attempt to ensure that it is bounded by *tol,* but most satisfy (59) indirectly (some less reliably than others) by controlling some measure of the local error. For a more complete discussion of global errors, local errors, and their relationship to the perturbed equation (59), see an advanced text such as Butcher (1987), Hairer *et al.* (1987), Hairer and Wanner (1991), Lambert (1991), or Shampine (1994).

We believe that this backward error approach is often the most natural way to view the error in the numerical integration of an IVP. In many practical problems, we know $f(x,y)$ approximately only, possibly because of measurement errors or neglected terms in the model. Therefore, the true solution of the system satisfies an equation of the form (59), where in this case $\delta(x)$ is the error in the model. So, any solution to an IVP of the form (59) may be equally good provided $\|\delta(x)\|$ is less than the error in the model.

Programs for the numerical solution of ODEs often contain many other useful features. For example, some routines for nonstiff IVPs warn the user if the problem is stiff, while others automatically switch between stiff and nonstiff methods depending on the characteristics of the problem. Some programs contain sophisticated strategies to integrate problems with discontinuities in f or its derivatives much more efficiently and reliably than programs that do not attempt to detect discontinuities. Also, some programs return an interpolant for the numerical solution or allow the user to evaluate the numerical solution at very closely spaced points more efficiently than if the integration method itself produced all these output points. This can greatly increase the efficiency of codes when used to produce graphical output or to detect when the numerical solution satisfies some condition [such as $y(x) = c$ for some constant c].

10.2 Boundary-Value Problems (BVPs)

10.2.1 Shooting Methods *Shooting* is conceptionally one of the simplest numerical techniques for solving the boundary-value problem (BVP) (45). In its simplest form, often called *simple shooting,* we guess an initial condition $y(a) = y_0$ for the IVP (44) for the same ODE as the BVP (45), solve the IVP (44), and test whether the boundary condition $g(y_0,y(b;a,y_0)) = 0$ is satisfied, or nearly so, where $y(b;a,y_0)$ is the solution at $x = b$ of the IVP (44) with the initial condition $y(a) = y_0$. In most cases, the first guess for the initial condition $y(a) = y_0$ does not yield a $g(y_0,y(b;a,y_0))$ that is close enough to 0. So we must apply some root-finding technique to adjust the initial condition $y(a) = y_0$ until $g(y_0,y(b;a,y_0)) = 0$ is satisfied, or nearly so, assuming that there is a solution to the BVP.

Each time we adjust the initial condition, we must solve the IVP (44) again with the new initial condition $y(a) = y_0^{(l)}$ to compute $y(b;a,y_0^{(l)})$ and then $g(y_0^{(l)},(b;a,y_0^{(l)}))$. For a scalar ODE ($m = 1$), we could try a simple technique such as bisection (see Sec. 6.4) to solve $g(y_0,y(b;a,y_0)) = 0$, but this converges slowly and so requires many solutions of the IVP (44) with different initial conditions $y_0^{(l)}$. Moreover, bisection is not applicable to systems of ODEs ($m > 1$).

The usual approach is to apply a variant of Newton's method (see Sec. 6.7) to solve $g(y_0,y(b;a,y_0)) = 0$. However, this requires that we compute an approximation to the Newton iteration matrix

$$\frac{dg(y_0^{(l)},y(b;a,y_0^{(l)}))}{dy_0} = \frac{\partial g(y_0^{(l)},y(b;a,y_0^{(l)}))}{\partial y_a} + \frac{\partial g(y_0^{(l)},y(b;a,y_0^{(l)}))}{\partial y_b}\frac{\partial y(b;a,y_0^{(l)})}{\partial y_0},$$

where $\partial g(y_0^{(l)},y(b;a,y_0^{(l)}))/\partial y_a$ is the partial derivative of g with respect to its first argument, $\partial g(y_0^{(l)},y(b;a,y_0^{(l)}))/\partial y_b$ is the partial derivative of g with respect to its second argument, and $\partial y(b;a,y_0^{(l)})/\partial y_0$ is the partial derivative of $y(b;a,y_0^{(l)})$ with respect to the initial condition $y(a) = y_0^{(l)}$. It can be shown that $\partial y(b;a,y_0^{(l)})/\partial y_0 = Y_l(b)$ for $Y_l : \mathbb{R} \to \mathbb{R}^{m\times m}$ the solution of the *variational equation*

$$Y_l'(x) = f_y(x,y_l(x))Y_l(x), \quad Y_l(a) = I, \tag{60}$$

where $y_l(x)$ is the solution of the associated IVP (44) with initial condition $y(a) = y_0^{(l)}$ and $f_y(x,y) = \partial f(x,y)/\partial y \in \mathbb{R}^{m\times m}$ is the Jacobian of f. Therefore, on each iteration of Newton's method, we must solve the IVP (44) with initial condition $y(a) = y_0^{(l)}$ for $y_l(x)$ as well as the variational equation (60) associated with $y_l(x)$. Since it may take many iterations before we find a $y_0^{(l)}$ for which $g(y_0^{(l)},y(b;a,y_0^{(l)}))$ is sufficiently close to 0, this is often a computationally expensive process.

Moreover, the associated IVP (44) may be unstable even though the BVP (45) is stable. As a result, simple shooting may break down or perform poorly. One way around this difficulty is to employ *multiple shooting.* In this scheme, we choose $N + 1$ shooting points $\{x_i : i = 0, \ldots, N\}$ satisfying $a = x_0 < x_1 < \cdots < x_{N-1} < x_N = b$, guess at N initial conditions s_i, $i = 0, \ldots, N - 1$, and solve the N IVPs

$$y_i' = f(x,y_i), \quad x \in [x_i,x_{i+1}] \quad i = 0, \ldots, N-1, \qquad y_i(x_i) = s_i. \tag{61}$$

These IVPs are completely independent and so could be integrated simultaneously. Hence, this scheme is often called *parallel shooting.*

We need to adjust the initial conditions s_i, $i = 0, \ldots, N - 1$, so that

$$y_i(x_{i+1}) = s_{i+1}, \quad i = 0, \ldots, N-2, \tag{62}$$

$$g(s_0, y_N(b)) = 0, \tag{63}$$

where the first set of conditions (62) ensures $y_i(x_{i+1}) = y_{i+1}(x_{i+1})$ at the $N - 1$ interior shooting points $x_1, \ldots, x_{N-1}$, thus allowing us to patch the functions $y_i(x)$ together into a continuous function $y(x)$ on $[a,b]$, and the second condition (63) enforces the boundary condition for the BVP (45).

A variant of Newton's method (see Sec. 6.7) is usually used to solve (62)–(63). The solution process is similar to, but somewhat more complicated than, that described above for simple shooting. It should be noted that the linear systems associated with Newton's method for (62)–(63) have a very special structure that can be exploited to great computational advantage. See an advanced text such as Ascher *et al.* (1988) for details.

Both simple and multiple shooting simplify significantly when applied to a linear ODE $y' = A(x)y + b(x)$. Newton's method converges in one iteration and the resulting scheme is equivalent to what is frequently called the method of *superposition.* If the boundary conditions are separated, this scheme simplifies still further. See an advanced text such as Ascher *et al.* (1988) for details.

Good shooting programs contain heuristics for choosing the shooting points and adjusting the tolerance for the IVP solver in an attempt to solve the BVP to within a user-specified tolerance. They also contain many other components, similar to those described in Sec. 10.1.10 for IVPs.

10.2.2 One-Step Methods It is common to apply a one-step method, such as a Runge–Kutta (RK) formula, to solve the BVP (45). Since a collocation method applied to an ODE often reduces to a RK formula, this class of methods is broader than it might at first appear.

To simplify the discussion, assume that the one-step method can be written in the form

$$y_{n+1} = y_n + h_n \phi(x_n,y_n,h_n), \tag{64}$$

where $y_n \approx y(x_n)$, $h_n = x_{n+1} - x_n$, and $a = x_0 < x_1 < \cdots < x_N = b$. Note that the RK formula (55) is of this form with

$$\phi(x_n,y_n,h_n) = \sum_{i=1}^{s} b_i k_i,$$
$$k_i = f\left(x_n + c_i h_n, y_n + h_n \sum_{j=1}^{s} a_{ij} k_j\right). \tag{65}$$

To apply the one-step formula (64) to the BVP (45), we simply combine the equations (64) together with the boundary conditions to get a large system of equations

$$\Phi(y_0, \ldots, y_N) = \begin{Bmatrix} y_{n+1} - y_n - h_n\phi(x_n,y_n,h_n), \ n = 0, \ldots, N-1, \\ g(y_0,y_N) \end{Bmatrix} = 0. \tag{66}$$

It is usual to apply a variant of Newton's method (see Sec. 6.7) to solve (66). As for shooting, the main difficulty here is to compute the $(N + 1)m \times (N + 1)m$ Newton iteration matrix $\partial\Phi(y_0, \ldots, y_N)/\partial(y_0, \ldots, y_N)$ and solve the associated linear system for the update to the approximate solution $y_0^{(l)}, \ldots, y_N^{(l)}$ to (66). See an advanced text such as Ascher *et al.* (1988) for a more complete discussion of this important point.

Good BVP codes contain heuristics for choosing the grid points to solve the BVP to within a user-specified tolerance. They also contain many other components, similar to those described in Sec. 10.1.10 for IVPs.

10.2.3 Other Methods There are several other classes of numerical methods for BVPs for ODEs. Some of these are discussed in Sec. 11 as numerical methods for BVPs for partial differential equations. An important class of methods, not discussed there, consists of *defect correction* schemes, including *deferred correction* as a special case. The basic idea behind these schemes is to apply a simple technique, possibly in the class discussed in the last subsection, and then estimate the *defect* or truncation error in the discretization and solve a related problem again with the same simple technique in an attempt to eliminate the error. See an advanced text such as Ascher *et al.* (1988) for further details.

11. PARTIAL DIFFERENTIAL EQUATIONS (PDEs)

A partial differential equation (PDE) is an equation in which the partial derivative of some order of the unknown function w.r.t. some independent variable occurs. For example,

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = g(x,y) \tag{67}$$

is a PDE, where $u(x,y)$ is an unknown function, $\partial^2 u/\partial x^2$ and $\partial^2 u/\partial y^2$ denote the partial second derivatives of u w.r.t. x and y, respectively (often denoted by u_{xx} and u_{yy}, respectively), and $g(x,y)$ is a given function. The terms that involve u and its derivatives define the (partial differential) *operator* L, where, for example, $L = \partial^2/\partial x^2 + \partial^2/\partial y^2$ in (67), and the rest of the terms (usually the right side of the equation) form the *source term.*

11.1 Classes of Problems and PDEs

PDEs describe many important physical and technological phenomena. These phenomena can be divided into two basic types, which in turn are associated with two basic classes of problems for PDEs.

1. Equilibrium phenomena, elliptic PDEs, boundary-value problems. In steady-state phenomena, the equilibrium configuration u often satisfies

$$Lu = g \quad \text{in } \Omega, \tag{68}$$

$$Bu = \gamma \quad \text{on } \partial\Omega, \tag{69}$$

where Ω is a spatial N-dimensional domain, $\partial\Omega$ is the boundary of Ω, u is the unknown function of N variables, g and γ are known functions of N variables, and L and B are partial differential operators. Such problems are called *boundary-value problems* (BVPs). Often L is an elliptic operator. Equation (69) is frequently referred to as the *boundary condition* (BC). The definition of an elliptic operator in the general case is beyond the scope of this article, but some typical examples are given below.

2. Propagation phenomena, parabolic and hyperbolic PDEs, initial-value problems. In phenomena of a transient nature, the initial state is often given and we wish to predict the subsequent behavior. The function u at some point $t \in (0,T)$ frequently satisfies

$$Lu = g \quad \text{in } \Omega \times (0,T), \tag{70}$$

$$Bu = \gamma \quad \text{on } \partial\Omega \times (0,T), \tag{71}$$

while the initial configuration satisfies

$$Iu = g_0 \quad \text{in } \Omega \cup \partial\Omega, \tag{72}$$

where $(0,T)$ is the time interval of interest, Ω is a spatial N-dimensional domain, $\partial\Omega$ is the boundary of Ω, u is the unknown function of N spatial variables and one-time variable t, g and γ are known functions of N spatial variables and t, g_0 is a known function of N spatial variables, and L, B, and I are partial differential operators. Such problems are called *initial-value problems* (IVPs). L is often either a parabolic or a hyperbolic operator (see below). Equation (72) is often referred to as an *initial condition* (IC).

11.1.1 Some Definitions The *dimension* of a PDE is the number of independent variables in the PDE. The *order* of a PDE is the order of the highest derivative of the unknown function occurring in the PDE. A PDE is called *linear* if there are no nonlinear terms in the equation involving the unknown function or its derivatives; otherwise it is called *nonlinear.*

For example,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{j=1}^{N} b_j(x)\frac{\partial u}{\partial x_j} + c(x)u = d(x) \tag{73}$$

is N-dimensional, second order, and linear, where $x = (x_1, \ldots, x_N)$ is an N-dimensional vector of independent variables, $u(x)$ is the unknown function, and $\{a_{ij} : i,j = 1, \ldots, N\}$, $\{b_j : j = 1, \ldots, N\}$, c, and d are given functions of x. If any of the functions a_{ij}, b_j, or c depends on u or its derivatives, or if d is nonlinear in u or its derivatives, then the PDE is nonlinear.

A linear operator L is called *positive-definite* if $(Lu,u) > 0$ for all $u \neq 0$, where $(\cdot,\cdot)$ denotes an inner product (see Sec. 8.7). In addition, L is called *self-adjoint* if $(Lu,v) = (u,Lv)$ for all u and v in the associated function space. For example, the two-dimensional second-order linear PDE

$$au_{xx} + bu_{xy} + cu_{yy} + du_x + eu_y + fu = g \tag{74}$$

is self-adjoint if $d(x,y) = \partial a/\partial x$, $e(x,y) = \partial c/\partial y$ and $b(x,y) = 0$.

Consider the linear differential equation $Lu = g$, with L self-adjoint and positive-definite. Consider also the quadratic functional $F(u)$ defined by $F(u) = (Lu,u) - 2(u,g)$. The *minimum functional theorem* states that the solution of the differential equation $Lu = g$ coincides with the function u that minimizes $F(u)$. In general, when a solution to a differential equation corresponds to an extremum of a related functional, we have a *variational principle.* Numerical methods that construct approximations to the solution of a differential equation by using such a relationship are called *variational methods, Ritz methods, Rayleigh–Ritz methods,* or *energy methods.* The latter term comes from the observation that many variational methods are based on the physical principle of energy minimization.

The two-dimensional second-order linear PDE (74) is *elliptic, parabolic,* or *hyperbolic* if $D = b^2 - 4ac < 0, = 0,$ or > 0, respectively. For the definitions of these terms in the general case, the reader is referred to any introductory PDE book, such as Ames (1992), Celia and Gray (1992), or Hall and Porsching (1990). Typical examples of elliptic, parabolic, and hyperbolic PDEs are

- Laplace's equation, $u_{x_1x_1} + u_{x_2x_2} + \cdots + u_{x_{N-1}x_{N-1}} + u_{x_Nx_N} = 0$, which is elliptic;
- the heat equation, $u_{x_1x_1} + u_{x_2x_2} + \cdots + u_{x_{N-1}x_{N-1}} - u_{x_N} = 0$, which is parabolic; and
- the wave equation $u_{x_1x_1} + u_{x_2x_2} + \cdots + u_{x_{N-1}x_{N-1}} - u_{x_Nx_N} = 0$, which is hyperbolic.

Two other classical elliptic PDEs (given in two dimensions) are

- Poisson's equation, $u_{xx} + u_{yy} = f(x,y)$, and
- the Helmholtz equation, $u_{xx} + u_{yy} + \kappa u = f(x,y)$, where κ is a constant.

The *normal derivative* of a surface $u(x,y)$ is the rate of change of u along the direction of the outward normal (i.e., the direction perpendicular to the surface). Let α be the angle that the direction of the outward normal makes with the x axis at a point (x,y) on u. Then the normal derivative $\partial u/\partial n$ of u (often denoted by u_n) at the point (x,y) is $u_n = u_x \times \cos\alpha + u_y \sin\alpha$. The normal derivative can also be written as the inner product of the gradient of u, ∇u, with the unit outward normal vector, n. That is, $u_n = \nabla u \cdot n$. This definition of the normal derivative for two dimensions can be generalized easily to N dimensions.

11.1.2 Boundary Conditions We now list some common types of boundary conditions (BCs) corresponding to the partial differential operator B in (69) or (71).

- *Dirichlet:* $Bu = u$.
- *Neumann:* $Bu = u_n$.
- *General* (for second-order PDEs): $Bu = \alpha(x)u + \beta(x)u_n$.
- *Mixed* (for second-order PDEs): Often, on parts of the boundary we have Dirichlet BCs and on the other parts Neumann ones. (Also, the term "mixed" may sometimes refer to more general types of BCs.)
- *Essential:* For PDEs of order $2m$, essential BCs involve u and its derivatives of order up to $m - 1$.
- *Natural:* For PDEs of order $2m$, natural BCs involve the derivatives of u from order m to $2m - 1$.

For further reading on the classification of PDE problems, operators, and boundary conditions, see Ames (1992), Celia and Gray (1992), or Hall and Porsching (1990).

11.2 Classes of Numerical Methods for PDEs

The two most commonly used methods for approximating the solution of PDEs are next described briefly.

Finite-Difference Methods (FDMs) have the following main steps:

1. Choose a finite-difference (FD) approximation of the derivatives involved in the PDE, BCs, and ICs. The result is a discretized PDE, BCs, and ICs.
2. Choose a set of n data points in the domain and on the boundary, on which the discretized PDE, BCs, and ICs must be satisfied. The result is a set of n equations w.r.t. the approximate values of u at the n data points.
3. Write the n equations of step **2** as a system and solve the system (discrete model). (If the PDE is linear, the system will usually be linear.) The solution is the approximate value of u at each of the n data points.
4. Evaluate the approximation to u at some point(s) of the domain (if needed).

Finite-Element Methods (FEMs) have the following main steps:

1. Choose a finite-element (FE) space, say n-dimensional, in which the approximation u_Δ is constrained to belong, and a set of basis functions that span the space, say $\{\phi_i : i = 1, \ldots, n\}$. Then write

$$u_\Delta(x) = \sum_{i=1}^{n} \alpha_i \phi_i(x).$$

   The unknown scalars α_i, $i = 1, \ldots, n$, are often called the *degrees of freedom* (DOF), or *coefficients*, of the FE representation of u_Δ.
2. Choose a set of n conditions that the approximation u_Δ must satisfy. The result is a set of n equations w.r.t. the n coefficients of u_Δ.
3. Write the n equations of step **2** as a system and solve the system (discrete model). (If the PDE is linear, the system will usually be linear.) The solution is the vector of coefficients of u_Δ.
4. Evaluate the approximation to u at some point(s) of the domain.

11.2.1 Analysis of Numerical Methods for PDEs Some common techniques used to analyze numerical methods for PDEs are discussed below. The analysis can be used to evaluate a method w.r.t. some chosen criteria or measures. In the following discussion, we use u_Δ to denote the approximation to u computed by the method.

11.2.1.1 Convergence Analysis (for BVPs and IVPs). We study the behavior of the error $u - u_\Delta$ as n increases. Assuming $\|u - u_\Delta\| \to 0$ as $n \to \infty$, we can write $\|u - u_\Delta\| \leq C(1/n)^\alpha$ for some constants C and α. The largest constant α for which this inequality holds is called the *order of convergence* of the method. As a first rough measure, the larger the α the better the method, as $(1/n)^\alpha$ will converge to 0 faster as $n \to \infty$ for larger α. To estimate the order of convergence of a method experimentally, we often devise PDE problems with known solutions and then solve them using the PDE method under investigation, first using n DOF, then $2n$ DOF, etc. We then plot $\|u - u_\Delta\|$ versus n on a log–log scale. The slope of the plotted line is an approximation to the order of convergence of the method.

11.2.1.2 Stability Analysis (for IVPs). We study the behavior of the error $u - u_\Delta$ as a function of t for increasing t. We often say that a method is *stable* if $\|u - u_\Delta\|$ remains bounded as $t \to \infty$. Otherwise, it is called *unstable*. Or, we study how the error at some point in time propagates to the next point in time. In a stable method, the error is not amplified.

11.2.1.3 Time (Computational) Complexity Analysis. We study the time that the method takes to compute the approximate solution to the PDE as a function of the n DOF. The time is usually proportional to the number of floating-point operations, although it also depends on the implementation and the hardware (computer) used. The most time-consuming part of a FDM or FEM is usually the third step (solution of the system), while the second most time-consuming part is usually the second step (generation of the system). For FDMs, the fourth step can also be time consuming, particularly if the value of the approximation at arbitrary points of the domain is required, since this computation requires interpolation, often using piecewise polynomials (PPs) or splines. The data to be interpolated are the approximate values of u at the grid points. Interpolation is not required in step **4** of a FEM, since the approximate solution can be evaluated at any point of the domain by the formula

$$u_\Delta(x) = \sum_{i=1}^{n} \alpha_i \phi_i(x).$$

By studying the particular implementation of a method, we are usually able to derive an approximate formula, such as time ≈ Kn^{β}, for some constant K, or time = $O(n^{\beta})$, relating the computational complexity to n. The smaller the β, the faster the method, and, among methods with the same β, the smaller the K, the faster the method.

11.2.1.4 Memory Complexity Analysis. We study the memory (storage) requirements of a method as a function of the n DOF. These requirements depend on the storage scheme used for the matrix arising in step **2** and the solver used in step **3**.

11.2.1.5 Overall Efficiency Analysis. Often, the most practical way of comparing two methods is to ask

1. if the methods were to run for the same length of time, which one would give the least error, or
2. given a certain error tolerance, which method satisfies that tolerance faster.

To test the overall efficiency of methods, we usually plot the error versus the time required to compute the approximate solution on a log–log scale. The method with the steepest slope is the most efficient.

11.3 Finite-Difference Methods for BVPs

A FD approximation to a derivative of a function u at a point x is a linear combination of values of u at points near x (often including x). Usually, a FD approximation is first derived for some derivative of a function of one variable, and then it is extended to partial derivatives of functions of several variables.

Let x be the point of interest and h, h_E, h_W small step sizes. The following are several examples of FD approximations in one dimension:

$$u_x(x) = \frac{u(x+h) - u(x)}{h} + O(h), \tag{75}$$

$$u_x(x) = \frac{u(x) - u(x-h)}{h} + O(h), \tag{76}$$

$$u_x(x) = \frac{u(x+h) - u(x-h)}{2h} + O(h^2), \tag{77}$$

$$u_x(x) = \frac{h_W^2 u(x+h_E) + (h_E^2 - h_W^2)u(x) - h_E^2 u(x-h_W)}{h_E(h_E+h_W)h_W} + O(h_E h_W), \tag{78}$$

$$u_{xx}(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + O(h^2), \tag{79}$$

$$u_{xx}(x) = \frac{2h_W u(x+h_E) - 2(h_E+h_W)u(x) + 2h_E u(x-h_W)}{h_E(h_E+h_W)h_W} + O(h_E - h_W) + O([\max(h_E,h_W)]^2). \tag{80}$$

Let (x,y) be the point of interest and h, h_E, h_W, h_N, h_S small step sizes. The following are several examples of FD approximations in two dimensions.

$$u_{xx}(x,y) = \frac{u(x+h,y) - 2u(x,y) + u(x-h,y)}{h^2} + O(h^2), \tag{81}$$

$$u_{xx}(x,y) + u_{yy}(x,y) = \frac{u(x+h,y) + u(x,y+h) - 4u(x,y) + u(x,y-h) + u(x-h,y)}{h^2} + O(h^2), \tag{82}$$

$$u_{xy}(x) = \frac{u(x+h,y+h) - u(x-h,y+h) - u(x+h,y-h) + u(x-h,y-h)}{4h^2} + O(h^2), \tag{83}$$

$$u_{xy}(x) = \frac{u(x+h_E,y+h_N) - u(x-h_W,y+h_N) - u(x+h_E,y-h_S) + u(x-h_W,y-h_S)}{(h_E+h_W)(h_S+h_N)} + O(\max(h_E,h_W,h_S,h_N)). \tag{84}$$

Note the following.

1. Each FD approximation listed above includes an error term. The actual FD approximation is the right side of the equation excluding the error term.
2. Approximations (75)–(79) and (81)–(83) use *uniform* step sizes, while the rest use *nonuniform* step sizes.
3. Approximations (75), (76), (80), and (84) are of *first order,* while the rest are of *second order.* The order refers to the (lowest) exponent of the step size(s) in the error term.
4. All FD approximation formulas are derived by using appropriate Taylor-series expansions around the point of approximation.
5. Approximations (81)–(84) are derived by using combinations of one-dimensional Taylor series and make use of values of u at points on a *rectangular grid.*
6. It is possible to derive two-dimensional FD approximations that make use of values of u at points on a *triangular, quadrilateral* (but not rectangular), *polygonal,* or *irregular grid* with points positioned arbitrarily. Such approximations can be derived by using two-dimensional Taylor's series.

11.3.1 An Example of a Finite-Difference Method in One Dimension Consider the problem

$$u_{xx} = g(x) \quad \text{in } (0,1), \tag{85}$$

$$u = \gamma(x) \quad \text{at } x = 0 \quad \text{and} \quad x = 1. \tag{86}$$

Using the FD approximation (79), we transform (85) to

$$[u(x+h) - 2u(x) + u(x-h)]/h^2 = g(x) + O(h^2). \tag{87}$$

Let $\{x_i = ih : i = 0, \ldots, n\}$ with $h = 1/n$ be the set of *grid points* and let $U_i \approx u(x_i)$ for $i = 1, \ldots, n$. Without the $O(h^2)$ error term, the discretized PDE (87) at the grid point x_i becomes

$$(U_{i+1} - 2U_i + U_{i-1})/h^2 = g(x_i). \tag{88}$$

From (86), we have for the point x_1

$$(U_2 - 2U_1)/h^2 = g(x_1) - \gamma(x_0)/h^2. \tag{89}$$

By writing equation (89) first, then (88) for $i = 2, \ldots, n-2$, and finally a relation similar to (89) for the point x_{n-1}, and then multiplying each equation by h^2, we get the following linear system:

$$\begin{pmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_{n-2} \\ U_{n-1} \end{pmatrix} = h^2 \begin{pmatrix} g(x_1) \\ g(x_2) \\ g(x_3) \\ \vdots \\ g(x_{n-2}) \\ g(x_{n-1}) \end{pmatrix} - \begin{pmatrix} \gamma(x_0) \\ 0 \\ 0 \\ \vdots \\ 0 \\ \gamma(x_n) \end{pmatrix}. \tag{90}$$

Note that this system is symmetric, diagonally dominant in all rows, and strictly diagonally dominant in the first and last rows. Therefore it is also positive-definite and has a unique solution. By solving it, we obtain $U_i \approx u(x_i)$, $i = 1, \ldots, n-1$. Using interpolation, we can approximate u at any other point of the domain.

It can be proved that $\max\{|u(x_i) - U_i| : i = 1, \ldots, n-1\} = O(h^2)$. That is, the approximation is second order at the grid points.

The computational complexity of the method described is $O(n)$, since the linear system (90) is tridiagonal and of size $n - 1$ (see Sec. 2.7).

11.3.2 An Example of a Finite-Difference Method in Two Dimensions Consider the problem

$$u_{xx} + u_{yy} = g(x,y) \quad \text{in } (0,1) \times (0,1), \tag{91}$$

$$u = \gamma(x,y) \quad \text{on } x = 0, \ x = 1, \quad y = 0, \ y = 1. \tag{92}$$

Using the FD approximation (82), we transform (91) to

$$\frac{u(x+h,y) + u(x,y+h) - 4u(x,y) + u(x,y-h) + u(x-h,y)}{h^2} = g(x,y) + O(h^2). \tag{93}$$

Let $\{(x_i,y_j) : x_i = ih,\ y_j = jh,\ i, j = 0, \ldots, n\}$ with $h = 1/n$ be the set of grid points and let $U_{ij} \approx u(x_i,y_j)$. Without the $O(h^2)$ error term, the discretized PDE (93) at the grid point (x_i,y_j), $i = 1, \ldots, n-1$, $j = 1, \ldots, n-1$, becomes

$$(U_{i+1,j} + U_{i,j+1} - 4U_{i,j} + U_{i,j-1} + U_{i-1,j})/h^2 = g(x_i,y_j). \tag{94}$$

From (92), we have for the point (x_1,y_1)

$$(U_{1,2} + U_{2,1} - 4U_{1,1})/h^2 = g(x_1,y_1) - [\gamma(x_0,y_1) + \gamma(x_1,y_0)]/h^2. \tag{95}$$

Similar relations hold at the three other corners of the domain. Also, for the points (x_1,y_j), $j = 2, \ldots, n-2$, we have

$$(U_{1,j+1} - 4U_{1,j} + U_{1,j-1} + U_{2,j})/h^2 = g(x_1,y_j) - \gamma(x_0,y_j)/h^2. \tag{96}$$

Similar relations hold for other grid points one grid line away from the boundary.

One way to number the grid points, and also the equations and unknowns, is bottom up then left to right: (1,1), (1,2), (1,3), . . . , $(1, n-2)$, $(1, n-1)$, (2,1), (2,2), (2,3), . . . , $(2, n-2)$, $(2, n-1)$, ... That is, first (95), then (96) for $j = 2, \ldots, n-2$, then relations similar to (95) for the points (x_1,y_{n-1}), (x_2,y_1), then (94) for $i = 2$, $j = 2, \ldots, n-2$, etc. Using this ordering, we get a linear system $AU = \mathbf{g}$, where, after multiplying each equation by h^2, the matrix A has the form

$$\left[\begin{array}{cccccccccccccccc}
-4 & 1 & & & & 1 & & & & & & & & & & \\
1 & -4 & 1 & & & & 1 & & & & & & & & & \\
 & \ddots & & \ddots & & & & \ddots & & & & & & & & \\
 & & 1 & -4 & 1 & & & & 1 & & & & & & & \\
 & & & 1 & -4 & & & & & 1 & & & & & & \\
1 & & & & & -4 & 1 & & & & 1 & & & & & \\
 & 1 & & & & 1 & -4 & 1 & & & & 1 & & & & \\
 & & \ddots & & & & \ddots & \ddots & \ddots & & & & \ddots & & & \\
 & & & \ddots & & & & \ddots & \ddots & \ddots & & & & \ddots & & \\
 & & & & 1 & & & & 1 & -4 & 1 & & & & 1 & \\
 & & & & & 1 & & & & 1 & -4 & & & & & 1 \\
 & & & & & & 1 & & & & & -4 & 1 & & & \\
 & & & & & & & 1 & & & & 1 & -4 & 1 & & \\
 & & & & & & & & \ddots & & & & \ddots & \ddots & \ddots & \\
 & & & & & & & & & 1 & & & & 1 & -4 & 1 \\
 & & & & & & & & & & 1 & & & & 1 & -4
\end{array}\right].$$

This can be rewritten as

$$\begin{pmatrix}
T & I & & & \\
I & T & I & & \\
 & \ddots & \ddots & \ddots & \\
 & & I & T & I \\
 & & & I & T
\end{pmatrix},$$

where T is a tridiagonal matrix of size $n-1$ with -4's on the diagonal and 1's on the super- and subdiagonal and I is the identity matrix of size $n-1$. Thus the matrix is *block-tridiagonal* of size $(n-1)^2$. The vector of unknown is

$$\begin{aligned}
U = (&U_{1,1}, U_{1,2}, \ldots, U_{1,n-2}, U_{1,n-1}, \\
&U_{2,1}, U_{2,2}, \ldots, U_{2,n-2}, U_{2,n-1}, \ldots \\
&U_{n-2,1}, U_{n-2,2}, \ldots, U_{n-2,n-2}, U_{n-2,n-1}, \\
&U_{n-1,1}, U_{n-1,2}, \ldots, U_{n-1,n-2}, U_{n-1,n-1})^{\mathrm{T}}.
\end{aligned}$$

The right-side vector is

$$\begin{aligned}\mathbf{g} = (&h^2g_{1,1} - \gamma_{0,1} - \gamma_{1,0}, h^2g_{1,2} - \gamma_{0,2}, \ldots,\\ &h^2g_{1,n-2} - \gamma_{0,n-2}, h^2g_{1,n-1} - \gamma_{0,n-1} - \gamma_{1,n},\\ &h^2g_{2,1} - \gamma_{2,0}, h^2g_{2,2}, \ldots, h^2g_{2,n-2},\\ &h^2g_{2,n-1} - \gamma_{2,n}, \ldots, h^2g_{n-2,1} - \gamma_{n-2,0},\\ &h^2g_{n-2,2}, \ldots, h^2g_{n-2,n-2},\\ &h^2g_{n-2,n-1} - \gamma_{n-2,n}, h^2g_{n-1,1} - \gamma_{n,1}\\ &- \gamma_{n-1,0}, h^2g_{n-1,2} - \gamma_{n,2}, \ldots, h^2g_{n-1,n-2}\\ &- \gamma_{n,n-2}, h^2g_{n-1,n-1} - \gamma_{n,n-1} - \gamma_{n-1,n})^{\mathrm{T}},\end{aligned}$$

where $g_{ij} = g(x_i,y_j)$ and $\gamma_{ij} = \gamma(x_i,y_j)$. Note that this system is symmetric, diagonally dominant in all rows, and strictly diagonally dominant in all rows corresponding to grid points one grid line away from the boundary. Therefore, it is positive-definite and has a unique solution. By solving the system $AU = \mathbf{g}$, we obtain $U_{ij} \approx u(x_i,y_j)$ for $i = 1, \ldots, n - 1$ and $j = 1, \ldots, n - 1$. Using interpolation, we can approximate the value of u at any other point of the domain.

It can be proved that $\max\{|u(x_i, y_j) - U_{i,j}| : i, j = 1, \ldots, n - 1\} = O(h^2)$. That is, the approximation is second order at the grid points.

The computational complexity of the method described above depends on the method used to solve the linear system $AU = \mathbf{g}$. Note that A has at most five nonzero entries per row, it is banded with lower and upper bandwidth $n - 1$, and its size is $(n - 1)^2$. If a direct band solver is used to solve $AU = \mathbf{g}$, then the computational complexity of the method is $O(n^4)$, but sparse direct solvers are more efficient (see Sec. 2.7). In addition, there exist iterative methods (e.g., multigrid; see Sec. 11.8 and Briggs, 1987) that can solve this system much more efficiently, reducing the computational complexity of the method to almost $O(n^2)$.

Note that the properties of the matrix A, such as symmetry, diagonal dominance, positive-definiteness, and the sparsity pattern (block-tridiagonal with at most five nonzero entries per row), are highly dependent on the simplicity of the differential operator associated with (91) and boundary conditions (92), the choice of uniform and rectangular grid, and the FD approximation (93). For a differential operator with first-order derivative terms and/or Neumann BCs, symmetry is lost. Symmetry may also be lost if a nonuniform grid is chosen, even if it is rectangular. Diagonal dominance depends on the coefficients of the differential operator and on the absence of first-order derivative terms. The block-tridiagonal form will most likely be affected if an irregular grid is chosen. Fast linear solvers, such as multigrid and FFT (fast Fourier transform) solvers, work well on the matrix A, but may not perform as well on more general systems. The development of fast linear solvers for such matrices is an open and active area of research. See, for example, Van Loan (1992) or Hackbusch (1994) and the references therein. For further reading on FDMs, see Strikwerda (1989).

11.4 Finite-Element Methods for BVPs

The first step in a FEM is to choose a FE approximation space and a basis for it. The most commonly used spaces are piecewise polynomials (PPs) or splines (see Sec. 8.5). Let n be the dimension of the approximation space and let $\{\phi_j(x) : j = 1, \ldots, n\}$ be a set of basis functions for the space.

Consider the problem (68)–(69). Let

$$u_\Delta(x) = \sum_{j=1}^{n} \alpha_j\phi_j(x)$$

be the approximation to u. The next step in a FEM is to choose n conditions that the approximation must satisfy. A FEM is characterized by these conditions. The most common FEMs are the *Galerkin method* and the *collocation method.*

11.4.1 The Galerkin Method Given an inner product $(\cdot,\cdot)$, usually defined by

$$(f,g) = \int_\Omega f(x)g(x)dx,$$

we require that u_Δ satisfies

$$(\phi_i, Lu_\Delta - g) = 0, \quad i = 1, \ldots, n, \tag{97}$$

forcing the *residual* $Lu_\Delta - g$ to be orthogonal to the approximation space, and making it, in a sense, as "small" as possible. If L is a linear operator, then the relations (97) are equivalent to

$$\sum_{j=1}^{n} \alpha_j(\phi_i,L\phi_j) = (\phi_i,g), \quad i = 1, \ldots, n, \tag{98}$$

which can be written in the form $A\boldsymbol{\alpha} = \mathbf{g}$, where A is an $n \times n$ matrix with entries $A_{ij} = (\phi_i, L\phi_j)$, $i = 1, \ldots, n$, $j = 1, \ldots, n$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ is the vector of coefficients, and $\mathbf{g}$ is a vector with entries $g_i = (\phi_i, g)$, $i = 1, \ldots, n$. We usually use numerical integration to compute the entries of A and $\mathbf{g}$ (see Sec. 9).

As an example of the Galerkin method in one dimension, consider the problem (85)–(86) and the set of grid points $\{x_i = ih : i = 0, \ldots, n\}$ with $h = 1/n$. Let $\{\phi_i : i = 0, \ldots, n\}$ be the set of linear spline basis functions w.r.t. the knots (grid points) $\{x_i\}$, as defined by (39). Then

$$u_\Delta(x) = \sum_{i=0}^{n} \alpha_i \phi_i(x)$$

is the linear spline approximation to u. From the BC (86), we get $u_\Delta(x_0) = \gamma(x_0)$, which implies $\alpha_0 = \gamma(x_0)$. Similarly, $\alpha_n = \gamma(x_n)$. The remaining unknowns $\{\alpha_i : i = 1, \ldots, n-1\}$ are determined by the $n - 1$ Galerkin conditions

$$\sum_{j=0}^{n} \alpha_j (\phi_i, L\phi_j) = (\phi_i, g), \quad i = 1, \ldots, n-1,$$

which, for the particular L associated with (85), are equivalent to

$$\sum_{j=0}^{n} \alpha_j (\phi_i, \phi_j'') = (\phi_i, g), \quad i = 1, \ldots, n-1.$$

Writing the inner product (ϕ_i, ϕ_j'') as an integral and applying integration by parts, these conditions reduce to

$$\sum_{j=1}^{n-1} \alpha_j \int_0^1 \phi_i' \phi_j' dx = \left[\phi_i \sum_{j=0}^{n} \alpha_j \phi_j' \right]_0^1 - \int_0^1 \phi_i g dx - \alpha_0 \int_0^1 \phi_i' \phi_0' dx - \alpha_n \int_0^1 \phi_i' \phi_n' dx, \quad i = 1, \ldots, n-1. \tag{99}$$

Note that the term

$$\left[\phi_i \sum_{j=0}^{n} \alpha_j \phi_j' \right]_0^1 = 0,$$

since $\phi_i(0) = \phi_i(1) = 0$ for $i = 1, \ldots, n-1$. Relations (99) form a linear system of size $n - 1$; the associated matrix A has elements

$$A_{i,j} = \int_0^1 \phi_i' \phi_j' dx.$$

Since the basis functions $\{\phi_i\}$ are nonzero on at most two subintervals, A is tridiagonal with elements

$$A_{i,i} = \int_{x_{i-1}}^{x_{i+1}} \phi_i' \phi_i' dx, \quad i = 1, \ldots, n-1,$$

$$A_{i,i-1} = \int_{x_{i-1}}^{x_i} \phi_i' \phi_{i-1}' dx, \quad i = 2, \ldots, n-1,$$

$$A_{i,i+1} = \int_{x_i}^{x_{i+1}} \phi_i' \phi_{i+1}' dx, \quad i = 1, \ldots, n-2.$$

It can be proved that this matrix is also symmetric positive-definite. Thus, the associated system has a unique solution. By solving the system, we obtain the coefficients $\{\alpha_i : i = 0, \ldots, n\}$ of u_Δ, which we can evaluate at any point of the domain (0, 1).

It can be proved that $\max\{|u(x) - u_\Delta(x)| : x \in [0, 1]\} = O(h^2)$. That is, the approximation is second order on the whole domain.

The computational complexity of the method described is $O(n)$, since the linear system that has to be solved is tridiagonal and of size $n - 1$ (see Sec. 2.7).

Relations (99) can also be derived using a variational method (see Sec. 11.1.1). Thus, for problem (85)–(86), there is a variational method equivalent to the Galerkin method. This is true for every differential equation problem with a self-adjoint positive-definite operator. There exist differential equation problems, though, that are not characterized by variational principles. In such cases, the Galerkin method is applicable, while the variational method is not.

As an example in two dimensions, consider problem (91)–(92) with the grid points $\{(x_i, y_j) : x_i = ih, y_j = jh, i, j = 0, \ldots, n\}$ for $h = 1/n$. A common way to define an approximation space for two-dimensional problems is to choose a tensor product of approximation spaces in each dimension. Let $\{\phi_i(x) : i = 0, \ldots, n\}$ be the linear spline basis functions w.r.t. the knots $\{x_i : i = 0, \ldots, n\}$ and let $\{\phi_j(y) : j = 0, \ldots, n\}$ be the linear spline basis functions w.r.t. the knots $\{y_j : j = 0, \ldots, n\}$, as defined in (39). Then

$$u_\Delta(x,y) = \sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_{ij}\phi_i(x)\phi_j(y)$$

is the bilinear spline approximation to u. Continuing as in the one-dimensional case, we derive a system of $(n + 1)^2$ equations in $(n + 1)^2$ unknowns. The associated matrix A is block-tridiagonal, with at most nine nonzero entries per row and bandwidth $n + 2$. It is also symmetric positive-definite. Thus, the associated system has a unique solution. Moreover, it can be proved that the approximation u_Δ is second order.

Note that, if instead of a rectangular subdivision of the domain and bilinear elements, we choose a triangular subdivision and linear elements (w.r.t. x and y), we would get a system similar to that of Sec. 11.3.2.

An important property of the Galerkin method is that, for any self-adjoint positive-definite differential operator, the resulting matrix is symmetric positive-definite, even if the grid is irregular. As stated before, for every differential equation problem with a self-adjoint and positive-definite operator, there is a variational method equivalent to the Galerkin method. This holds for higher-dimension problems too. Thus, large, sparse, symmetric, positive-definite matrices arise from the application of variational methods.

For an introduction to the FEM, including its computer implementation, see Becker *et al.* (1981). An error analysis is carried out in Strang and Fix (1973).

11.4.2 The Collocation Method We first pick n *collocation points* $\{t_i : i = 1, \ldots, n\}$ in Ω and on $\partial\Omega$. We then require that u_Δ satisfies

$$Lu_\Delta(t_i) - g(t_i) = 0, \quad \text{if } t_i \in \Omega, \tag{100}$$

$$Bu_\Delta(t_i) - \gamma(t_i) = 0, \quad \text{if } t_i \in \partial\Omega, \tag{101}$$

forcing the residuals $Lu_\Delta - g$ and $Bu_\Delta - \gamma$ to be zero at the collocation points, and making them, in a sense, as "small" as possible. If L and B are linear, relations (100)–(101) are equivalent to

$$\sum_{j=1}^{n} \alpha_j L\phi_j(t_i) = g(t_i), \quad \text{if } t_i \in \Omega, \tag{102}$$

$$\sum_{j=1}^{n} \alpha_j B\phi_j(t_i) = \gamma(t_i), \quad \text{if } t_i \in \partial\Omega, \tag{103}$$

which can be written in the form $A\boldsymbol{\alpha} = \mathbf{g}$, where A is an $n \times n$ matrix with entries $A_{ij} = L\phi_j(t_i)$, $j = 1, \ldots, n$, for all $t_i \in \Omega$, and $A_{ij} = B\phi_j(t_i)$, $j = 1, \ldots, n$, for all $t_i \in \partial\Omega$; $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ is the vector coefficients; and $\mathbf{g}$ is a vector with entries $g_i = g(t_i)$ for all $t_i \in \Omega$ and $g_i = \gamma(t_i)$ for all $t_i \in \partial\Omega$.

The choice of collocation points is critical to the success of the method. It affects not only the solvability and other properties (such as symmetry, diagonal dominance, bandedness) of the matrix A but also the accuracy of the approximation u_Δ. Depending on the FE approximation space that u_Δ belongs to, some standard choices of collocation points in one dimension are listed below.

1. If the FE approximation space is the space of quadratic splines (quadratic PPs in $\mathscr{C}^1$), the collocation points are chosen to be the midpoints of the subintervals (x_{i-1}, x_i), $i = 1, \ldots, n$, and the two boundary points. The same choice of collocation points is effective if the FE approximation space is composed of any other even-degree splines, with the exception that some additional collocation conditions may be required at boundary points or points close to the boundary.
2. If the FE approximation space is the space of cubic splines (cubic PPs in $\mathscr{C}^2$), the collocation points are chosen to be the grid points $\{x_i : i = 0, \ldots, n\}$. At each of the boundary points, x_0 and x_n, both conditions (100) and (101) are imposed. The same choice of collocation points is effective if the FE approximation space is composed of any other odd-degree splines, with the exception that some additional collocation conditions may be required at boundary points or points close to the boundary.
3. If the FE approximation space is the space of cubic PPs in $\mathscr{C}^1$ (cubic Hermite PPs), the collocation points are chosen to be the two Gauss points $x_{i-1} + (3 \pm \sqrt{3}) \times (x_i - x_{i-1})/6$ in each subinterval (x_{i-1}, x_i), $i = 1, \ldots, n$, and the two boundary grid points.

As an example in one dimension, consider problem (85)–(86) and the set of grid points $\{x_i = ih : i = 0, \ldots, n\}$ with $h = 1/n$. Let the collocation points be the midpoints $t_i = (x_{i-1} + x_i)/2$, $i = 1, \ldots, n$, and the end points $t_0 = x_0$ and $t_{n+1} = x_n$. Let $\{\phi_i : i = 0, \ldots, n + 1\}$ be the quadratic spline basis functions w.r.t. the knots (grid points) $\{x_i\}$, as defined in (40). Then

$$u_\Delta(x) = \sum_{i=0}^{n+1} \alpha_i \phi_i(x)$$

is the quadratic spline approximation to u. Relation (100) for the PDE (85) becomes $u''_\Delta(t_i) = g(t_i)$, and so relation (102) becomes

$$\alpha_{i-1}\phi''_{i-1}(t_i) + \alpha_i\phi''_i(t_i) + \alpha_{i+1}\phi''_{i+1}(t_i) = g(t_i),$$

which reduces to

$$(\alpha_{i-1} - 2\alpha_i + \alpha_{i+1})/h^2 = g(t_i), \quad i = 1, \ldots, n. \tag{104}$$

Relation (101) for the BC (86) becomes $u_\Delta(t_0) = \gamma(t_0)$, and so relation (103) becomes

$$\alpha_0\phi_0(t_0) + \alpha_1\phi_1(t_0) = \gamma(t_0),$$

which reduces to

$$(\alpha_0 + \alpha_1)/2 = \gamma(t_0). \tag{105}$$

Similarly, the collocation condition at $t_{n+1} = 1$ reduces to

$$(\alpha_n + \alpha_{n+1})/2 = \gamma(t_{n+1}). \tag{106}$$

Writing (105) first, then (104) for $i = 1, \ldots, n$, and finally (106), we get a tridiagonal system of equations w.r.t. the coefficients $\{\alpha_i : i = 0, \ldots, n + 1\}$. The system is diagonally dominant and it can be proved that it has a unique solution. It can also be scaled so that it is symmetric positive-definite.

It can be proved that $\max\{|u(x) - u_\Delta(x)| : x \in [0, 1]\} = O(h^2)$. That is, the approximation is second order on the whole domain. There exists a variant of this method, though, that is fourth order at the grid points and midpoints and third order on the whole domain (Houstis *et al.*, 1988).

The computational complexity of the method described above is $O(n)$, since the linear system that must be solved is tridiagonal and of size $n + 1$ (see Sec. 2.7).

As an example in two dimensions, consider problem (91)–(92) and the set of grid points $\{(x_i, y_j) : x_i = ih, y_j = jh, i, j = 0, \ldots, n\}$ with $h = 1/n$. A common approximation space for two-dimensional problems is a tensor product of approximation spaces in each dimension. Let $\{\phi_i(x) : i = 0, \ldots, n + 1\}$ be the quadratic spline basis functions w.r.t. the knots (grid points) $\{x_i : i = 0, \ldots, n\}$ and let $\{\phi_j(y) : j = 0, \ldots, n + 1\}$ be the quadratic spline basis functions w.r.t. the knots $\{y_j : j = 0, \ldots, n\}$, as defined in (40). Then

$$u_\Delta(x,y) = \sum_{i=0}^{n+1}\sum_{j=0}^{n+1} \alpha_{ij}\phi_i(x)\phi_j(y)$$

is the biquadratic spline approximation to u. Continuing as in the one-dimensional case, we derive a system of $(n + 2)^2$ equations and unknowns. The associated matrix is block-tridiagonal, with at most nine nonzero entries per row, and has bandwidth $n + 3$. It can be proved that this system has a unique solution and that the approximation u_Δ is second order. With appropriate modifications, though, the order can be improved as in the one-dimensional case (Christara, 1994).

For a general introduction to collocation methods, see Prenter (1975).

11.5 Finite-Difference Methods for IVPs

Consider the problem (70)–(72). Let the temporal grid points be $\{t_j = jh_t : j = 0, \ldots, m\}$ with $h_t = T/m$. Starting with the initial values of u at t_0, given by (72), most FDMs for IVPs compute approximate values of u at each subsequent temporal grid point t_j, in the order $j = 1, \ldots, m$, using previous and/or current approximate values of u at neighboring space points.

If at each temporal grid point t_j a method uses only approximations from previous temporal grid points, it is called *explicit*, as it does not require the solution of a system of equations to proceed from one temporal grid point to the next. If at some temporal grid point t_j a method uses approximations from the current temporal grid point t_j, it is called *implicit*, as it requires the solution of a system of equations to proceed from one temporal grid point to the next. If at the time step from t_{j-1} to t_j a method uses approximations from t_{j-1} and t_j only, it is called *one-step*. Likewise, we can define *two-step* methods, etc. These definitions for PDEs are similar to those given in Sec. 10 for ODEs.

11.5.1 An Example of an Explicit One-Step Method for a Parabolic IVP Consider the problem

$$u_t = u_{xx} \quad \text{in } (0,1) \times (0,T), \tag{107}$$

$$u = \gamma_0(t) \quad \text{on } x = 0,\ t \in (0,T), \tag{108}$$

$$u = \gamma_1(t) \quad \text{on } x = 1,\ t \in (0,T), \tag{109}$$

$$u = g(x) \quad \text{on } t = 0,\ x \in [0,1]. \tag{110}$$

Using the FD approximations (79) for u_{xx} and (75) for u_t, we transform (107) to

$$\begin{aligned}&[u(x, t + h_t) - u(x, t)]/h_t \\ &\quad = [u(x + h, t) - 2u(x, t) + u(x - h, t)]/h^2 \\ &\qquad + O(h_t + h^2).\end{aligned} \tag{111}$$

Let $\{x_i = ih : i = 0, \ldots, n\}$ with $h = 1/n$ be the set of spatial grid points and $\{t_j = jh_t : j = 0, \ldots, m\}$ with $h_t = T/m$ be the set of temporal grid points. Also let $U_{i,j} \approx u(x_i,t_j)$ for $i = 0, \ldots, n$ and $j = 0, \ldots, m$. Then the discretized PDE (111) at the point (x_i,t_j), $i = 1, \ldots, n - 1$, $j = 1, \ldots, m$, becomes

$$(U_{i,j+1} - U_{i,j})/h_t = (U_{i+1,j} - 2U_{i,j} + U_{i-1,j})/h^2.$$

Letting $r = h_t/h^2$, we can rewrite this relation as

$$U_{i,j+1} = rU_{i+1,j} + (1 - 2r)U_{i,j} + rU_{i-1,j} \tag{112}$$

for $i = 2, \ldots, n - 2$. For $i = 1$, we have from (108)

$$U_{1,j+1} = rU_{2,j} + (1 - 2r)U_{1,j} + r\gamma_0(t)_j. \tag{113}$$

Similarly, for $i = n - 1$, we have from (109)

$$U_{n-1,j+1} = r\gamma_1(t_j) + (1 - 2r)U_{n-1,j} + rU_{n-2,j}. \tag{114}$$

For $j = 1$, we have from (110)

$$U_{i,1} = rg(x_{i+1}) + (1 - 2r)g(x_i) + rg(x_{i-1}). \tag{115}$$

Thus, we can compute $U_{i,j} \approx u(x_i,t_j)$ from a linear combination of three neighboring spatial approximations at t_{j-1}.

It can be proved that if $r < \frac{1}{2}$, then $\max\{|u(x_i, t_j) - U_{i,j}| : i = 1, \ldots, n,\ j = 1, \ldots, m\} = O(h^2 + h_t)$; thus the order of convergence is one w.r.t. to h_t and two w.r.t. h. It can also be proved that if $r < \frac{1}{2}$, the method is stable. However, the restriction $r < \frac{1}{2}$ may be impractical for many problems, since it forces h_t to be very small if h is small and so the method must take many steps to integrate the problem.

The computational complexity of the method is $O(nm)$, since for each grid point (x_i,t_j) a constant number of floating-point operations must be performed.

11.5.2 An Example of an Implicit One-Step Method for a Parabolic IVP Consider the problem (107)–(110) once more. Using the FD approximations (79) for u_{xx} and (76) for u_t, we transform (107) to

$$\begin{aligned}&[u(x, t + h_t) - u(x, t)]/h_t = [u(x + h, t + h_t) \\ &\quad - 2u(x, t + h_t) + u(x - h, t + h_t)]/h^2 \\ &\quad + O(h_t + h^2).\end{aligned} \tag{116}$$

Again let $\{x_i = ih : i = 0, \ldots, n\}$ with $h = 1/n$ be the set of spatial grid points and $\{t_j = jh_t : j = 0, \ldots, m\}$ with $h_t = T/m$ be the set of temporal grid points. Also let $U_{i,j} \approx u(x_i,t_j)$ for $i = 0, \ldots, n$ and $j = 0, \ldots, m$. Then, the discretized PDE (116) at the point (x_i,t_j), $i = 1, \ldots, n - 1$, $j = 1, \ldots, m$, becomes

$$\begin{aligned}&(U_{i,j+1} - U_{i,j})/h_t \\ &\quad = (U_{i+1,j+1} - 2U_{i,j+1} + U_{i-1,j+1})/h^2.\end{aligned}$$

Letting $r = h_t/h^2$ again, we can rewrite this relation as

$$-rU_{i-1,j+1} + (1 + 2r)U_{i,j+1} - rU_{i+1,j+1} = U_{i,j} \tag{117}$$

for $i = 2, \ldots, n - 2$. For $i = 1$, we have from (108)

$$(1 + 2r)U_{1,j+1} - rU_{2,j+1} = U_{1,j} + r\gamma_0(t_{j+1}). \tag{118}$$

Similarly, for $i = n - 1$, we have from (109)

$$-rU_{n-2,j+1} + (1 + 2r)U_{n-1,j+1} = U_{n-1,j} + r\gamma_1(t_{j+1}). \quad (119)$$

For $j = 1$, we have from (110)

$$-rU_{i-1,1} + (1 + 2r)U_{i,1} - rU_{i+1,1} = g(x_i). \quad (120)$$

Thus, at the jth time step, a tridiagonal linear system must be solved to compute $U_{i,j} \approx u(x_i,t_j)$. The diagonal entries of the associated matrix are all equal to $1 + 2r$, while the off-diagonal entries are all equal to $-r$. The system is symmetric positive-definite and strictly diagonally dominant; thus it has a unique solution.

It can be proved that $\max\{|u(x_i, t_j) - U_{i,j}| : i = 1, \ldots, n, j = 1, \ldots, m\} = O(h^2 + h_t)$; thus the order of convergence is one w.r.t. to h_t and two w.r.t. h. It can also be proved that the method is stable without any restrictions on r (except $r > 0$). The computational complexity of the method is $O(nm)$, since at each time step we must solve a tridiagonal linear system of size $n - 1$ (see Sec. 2.7).

Note that for the problem (107)–(110), which is one-dimensional w.r.t. to space, both the explicit and implicit methods have the same computational complexity. This is not true for problems in more space dimensions. For such problems, the solution of a linear system at each time step can be very time consuming, making an implicit method much more expensive per step than an explicit one. However, because there is no restriction on r for some implicit schemes, while there always is for an explicit one, some implicit schemes may be able to take far fewer time steps than an implicit one. As a result, an implicit method may be computationally more efficient than an explicit one.

11.5.3 An Example of an Explicit Two-Step Method for a Hyperbolic IVP Consider the problem

$$u_{tt} = u_{xx} \quad \text{in } (0,1) \times (0,T), \quad (121)$$

$$u = \gamma_0(t) \quad \text{on } x = 0, \quad t \in (0,T), \quad (122)$$

$$u = \gamma_1(t) \quad \text{on } x = 1, \quad t \in (0,T), \quad (123)$$

$$u = \gamma_0(x) \quad \text{on } t = 0, \quad x \in (0,1), \quad (124)$$

$$u_t = \gamma_1(x) \quad \text{on } t = 0, \quad x \in (0,1). \quad (125)$$

Using the FD approximation (79) for u_{xx} and u_{tt}, we transform (121) to

$$[u(x, t + h_t) - 2u(x, t) + u(x, t - h_t)]/h_t^2 = [u(x + h, t) - 2u(x, t) + u(x - h, t)]/h^2 + O(h_t^2 + h^2). \quad (126)$$

Again let $\{x_i = ih : i = 0, \ldots, n\}$ with $h = 1/n$ be the set of spatial grid points and $\{t_j = jh_t : j = 0, \ldots, m\}$ with $h_t = T/m$ be the set of temporal grid points. Also let $U_{i,j} \approx u(x_i,t_j)$ for $i = 0, \ldots, n$ and $j = 0, \ldots, m$. Then, the discretized PDE (126) at the point (x_i,t_j), $i = 1, \ldots, n - 1, j = 1, \ldots, m$, becomes

$$(U_{i,j+1} - 2U_{i,j} + U_{i,j-1})/h_t^2 = (U_{i+1,j} - 2U_{i,j} + U_{i-1,j})/h^2.$$

Letting $r = h_t/h$, we can rewrite this relation as

$$U_{i,j+1} = r^2U_{i-1,j} + 2(1 - r^2)U_{i,j} + r^2U_{i+1,j} - U_{i,j-1} \quad (127)$$

for $i = 2, \ldots, n - 1$. For grid points close to the boundary, the approximate values of U are replaced by the values of the functions γ_0 and γ_1 at the appropriate points, as in Secs. 11.5.1 and 11.5.2. Thus, we can compute $U_{i,j+1} \approx u(x_i,t_{j+1})$ from a linear combination of three neighboring spatial approximations at time t_j and one approximation at time t_{j-1}.

Since (127) is a two-step formula, at the initial time point t_0 it cannot be applied as is. At that point, we use the ICs (124)–(125) and the FD approximation (75) to get

$$U_{i,1} = g_0(x_i) + h_t g_1(x_i). \quad (128)$$

It can be proved that, if $r < 1$, then the method is stable. It can also be shown that $\max\{|u(x_i, t_j) - U_{i,j}| : i = 1, \ldots, n, j = 1, \ldots, m\} = O(h^2 + h_t^2)$; thus the order of convergence is two w.r.t. to both h_t and h. Note that the restriction $r < 1$ is not impractical in this case, since it requires only that $h_t < h$. The computational complexity of the method is $O(nm)$, since for each grid point (x_i,t_j) we apply a formula with a constant number of floating-point operations.

11.6 The Method of Lines

The general idea behind the *method of lines* (MOL) is to use an ODE solver along one of the dimensions of the PDE, while using a PDE discretization across the other dimensions. In its most common form for the solution of IVPs for PDEs, an ODE solver is used along the temporal dimension, while a PDE discretization is employed across the spatial dimensions, transforming an IVP for a PDE into a system of IVPs for ODEs.

To see how this is done, consider the problem (107)–(110) again. Let

$$u_\Delta(x,t) = \sum_{j=1}^{n} \alpha_j(t)\phi_j(x)$$

be a FE approximation to the true solution $u(x,t)$. Now apply a FEM condition to u_Δ to discretize the PDE (107) w.r.t. the spatial dimension. For example, collocation at the points x_i, $i = 1, \ldots, n$, yields

$$\sum_{j=1}^{n} \alpha_j'(t)\phi_j(x_i) = \sum_{j=1}^{n} \alpha_j(t)\phi_j''(x_i).$$

Let $\boldsymbol{\alpha}(t) = (\alpha_1(t), \ldots, \alpha_n(t))^{\mathrm{T}}$, Φ be the matrix with entries $\Phi_{ij} = \phi_j(x_i)$, and A be the matrix with entries $A_{ij} = \phi_j''(x_i)$. Then the PDE (107) is approximated by the system of ODEs $\Phi\boldsymbol{\alpha}'(t) = A\boldsymbol{\alpha}(t)$.

To obtain an IC for the ODE, we construct an interpolant g_Δ of g in the same space as that spanned by $\{\phi_j : j = 1, \ldots, n\}$. Let

$$g_\Delta(x) = \sum_{j=1}^{n} \beta_i\phi_i(x)$$

be the FE representation of g_Δ in that space and set $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^{\mathrm{T}}$. Then

$$\Phi\boldsymbol{\alpha}'(t) = A\boldsymbol{\alpha}(t), \tag{129}$$

$$\boldsymbol{\alpha}(0) = \boldsymbol{\beta} \tag{130}$$

is a well-defined IVP for ODEs. Thus, the PDE problem (107)–(110) is converted to an IVP for a system of n ODEs. The latter can be solved by the techniques described in Sec. 10.

Note that applying an ODE method to discretize the IVP (129)–(130) results in a discretization for the PDE (107)–(110). That is, the MOL produces a discretization for a PDE. However, it is generally agreed that standard software for ODEs is more highly developed than for PDEs. Thus, using the MOL to decouple the discretization of the spatial and temporal variables allows us to exploit easily sophisticated time-stepping techniques. As a result, the MOL is often the simplest effective method to solve a PDE.

11.7 Boundary-Element Methods

The general idea behind *boundary-element methods* (BEMs) is to transform the PDE to an integral equation in which the integrations take place along the boundary only of the PDE domain, thus eliminating the need for domain discretization and reducing the dimension of the PDE by one. For example, a one-dimensional integral equation is solved instead of an equivalent two-dimensional PDE. The BEM is applicable to BVPs for Laplace's or Poisson's equation, and many other simple PDEs. If applicable, this approach is often very effective, especially when the PDE domain is highly irregular.

11.8 The Multigrid Method

The multigrid method (MM) exploits the connection between a physical problem and its matrix analog to accelerate the convergence of an iterative method (see Sec. 3). For simplicity, we describe the MM for the one-dimensional problem (85)–(86), although the merits of the scheme become apparent for two- and higher-dimensional problems (see Sec. 11.3.2). Also, we illustrate the technique using Jacobi's method as the basic iterative scheme. The MM, though, can be used with many other iterative methods as a preconditioning technique.

Let A be the matrix in (90). Apply Jacobi's method (see Sec. 3.1) with an extra damping factor of 2 to the linear system (90). The associated iteration matrix is $G = I - A/4$. It can be shown that the eigenvalues of G are $\mu_i = \cos^2[i\pi/(2n)]$, $i = 1, \ldots, n - 1$, and that the components of the eigenvector v_i associated with μ_i are $\sin[i\pi(j/n)]$, $j = 1, \ldots, n - 1$. These are also the eigenvectors of A. Since $\{v_i : i = 1, \ldots, n - 1\}$ spans $\mathbb{R}^{n-1}$, we can write the error e_0 associated with the initial guess for the damped Jacobi iteration as

$$e_0 = \sum_{i=1}^{n-1} \alpha_i v_i$$

for some scalars $\{\alpha_i : i = 1, \ldots, n - 1\}$. It then follows easily from the discussion in Sec. 3.1 that the error at iteration k is

$$e_k = \sum_{i=1}^{n-1} \alpha_i \mu_i^k v_i.$$

The terms of the sum corresponding to small values of i are called *low-frequency components,* while those corresponding to large values of i are called *high-frequency components.* Note that $0 < \mu_{n-1} < \mu_{n-2} < \cdots < \mu_2 < \mu_1 < 1$. Moreover, $\mu_1 \approx 1 - (\pi/2n)^2$, while $\mu_{n-1} \approx (\pi/2n)^2$. Consequently, $e_k \to 0$ as $k \to \infty$, but the low-frequency components of the error converge slowly, while the high-frequency components converge rapidly.

To accelerate the convergence of the low-frequency components of the error, consider solving the problem (90) on a coarse grid with $\hat{n} = n/2$ subintervals and $\hat{n} + 1$ grid points, assuming for simplicity that n is even. Although the coarse grid has about half the number of grid points of the fine one, the $\hat{n} - 1$ eigenvectors of the matrix $\hat{A}$ for the coarse grid provide a good representation of the low- to middle-frequency eigenvectors of A. As a result, the solution to the problem (90) on the coarse grid provides a good approximation to the low- to middle-frequency components of the fine-grid solution. This suggests that the coarse-grid solution can be used to provide good approximations to the low- to middle-frequency components of the fine-grid solution, while the damped Jacobi iteration on the fine grid can be used to provide good approximations to the middle- to high-frequency components of the fine-grid solution.

This is the motivation behind the MM. The term "multigrid" refers to the use of several levels of grids (possibly a fine grid, several intermediate-level grids, and a coarse grid), so that each level damps certain components of the error fast.

To view the MM as a preconditioning technique, consider the linear system $Au = g$ corresponding to the discretization of problem (85)–(86) on some fine grid. Apply one (or a few) damped Jacobi iteration(s) to $Au = g$ to obtain an approximate solution vector $\tilde{u}$. Let $r = g - A\tilde{u}$ be the *residual* vector. Project r to a coarse grid to obtain the *coarse-grid residual* vector $\hat{r}$. This can be done by appropriately interpolating the components of r and evaluating the interpolant at the points of the coarse grid (see Sec. 8). Let $\hat{A}$ be the matrix corresponding to the discretization of problem (85)–(86) on the coarse grid. Solve (or approximately solve) $\hat{A}\tilde{r} = \hat{r}$. This can be done by applying a few damped Jacobi iterations, or by recursively applying the MM to $\hat{A}\tilde{r} = \hat{r}$, or by using a direct solver (see Sec. 2), since $\hat{A}$ is a smaller matrix than A. Now $\tilde{r}$ is the *preconditioned coarse-grid residual* vector. Extend $\tilde{r}$ to the fine grid to obtain the *preconditioned fine-grid residual* vector $\bar{r}$. This can be done by appropriately interpolating the components of $\tilde{r}$ and evaluating the interpolant at the points of the fine grid. Add $\bar{r}$ to $\tilde{u}$ to obtain a new approximate solution vector. Repeat the process until convergence. Usually, only a few iterations are needed. Note that this scheme has some similarities to iterative improvement, as described in Sec. 2.9.

The power of the MM lies in the fact that the coarse grid, which acts as a preconditioner, allows the information to pass from a point of the problem domain to another point in a few steps, while the fine grid maintains the accuracy required. Note that the interpolation and the evaluation of the interpolant needed for the projection of a fine-grid vector to a coarse-grid vector and for the extension of a coarse-grid vector to a fine-grid vector often reduce to simple relations, such as averaging neighboring vector components. A practical introduction to the MM, including an error analysis, can be found in Briggs (1987).

12. PARALLEL COMPUTATION

The increasing demand by scientists and engineers to solve larger and larger problems constitutes the primary motivation for parallel computation. Another impetus is the cost effectiveness of computers consisting of many standard CPUs compared with those based on one very fast CPU.

A parallel computer has the ability to execute simultaneously many different processes by having several independent processors concurrently perform operations on different data. A parallel computer in which all processors perform the same operation on different data is called a *single-instruction, multiple-data* (SIMD) machine, while one in which each processor has the ability to perform different operation(s) on data is called a *multiple-instruction, multiple-data* (MIMD) machine. Sometimes a processor may use data computed by another processor, either by exchanging messages with it or by sharing some common memory area. The former is characteristic of *distributed-memory* or *message-passing* machines, while the latter is characteristic of *shared-memory* machines.

A vector computer may be viewed as a restricted form of a parallel machine. On such a computer, one arithmetic operation is performed by several (usually a few) processors, which cooperate in a pipelined manner. Each processor receives an input operand from another processor, performs a part of the operation on it, and passes the result on to the next processor. While a processor performs its part of the operation on an operand, the next processor performs its part of the operation on the previous operand, etc. This pipelining technique is very effective when the same operation is performed on many data—for example, on each component of a long vector.

To utilize parallel computers effectively, we must be able to split the computation into *parallel processes* to be assigned to different processors. To this end, it is desirable to have fully independent computations in each process. For many problems, though, this is impossible, but there are techniques to minimize the dependence of the computation of one process on another. It is also desirable to have almost equal amounts of computation in each process for *load balancing.* This is also hard to accomplish in many cases, but again there exist techniques to achieve reasonable load balancing.

Many computational scientists study the parallelization of existing numerical methods, as well as the development of new methods appropriate for parallel machines. In the next subsection, we discuss, as a simple example, a numerical method that is effective for solving tridiagonal linear systems on parallel computers.

12.1 Cyclic Reduction

Consider solving the tridiagonal system

$$\begin{pmatrix} a_1 & c_1 & & & & \\ b_2 & a_2 & c_2 & & & \\ & b_3 & a_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & b_{n-1} & a_{n-1} & c_{n-1} \\ & & & & b_n & a_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix}$$

by the following *LU* factorization algorithm for tridiagonal matrices:

```
for $k = 1, \ldots, n - 1$ do
    $b_{k+1} = b_{k+1}/a_k$
    $a_{k+1} = a_{k+1} - b_{k+1}c_k$
end
```

At the end of the computation, the modified *b*'s form the subdiagonal of the unit lower-triangular matrix *L* and the modified *a*'s and *c*'s form the diagonal and superdiagonal, respectively, of the upper-triangular matrix *U* (see Secs. 2.3 and 2.7).

Note that the computation proceeds in the order $b_2, a_2, b_3, a_3, \ldots$. Each value computed depends on the previous one. Therefore, it seems that the computation is purely sequential and that there is no easy way to parallelize it. However, there are other ways to solve tridiagonal linear systems, and some of them can be implemented effectively on a parallel machine.

Assume, for simplicity, that $n = 2^q - 1$, where q is a positive integer. Multiply row 1 by b_2/a_1 and subtract it from row 2, eliminating x_1 from row 2. Also multiply row 3 by c_2/a_3 and subtract it from row 2, eliminating x_3 from row 2. The new row 2 involves variables x_2 and x_4 only.

Repeat the process described above for the $(n - 1)/2$ groups of rows (3,4,5), (5,6,7), etc. This eliminates the odd unknowns from the even equations. The even equations form a new tridiagonal linear system of about half the size, $(n - 1)/2 = 2^{q-1} - 1$, called the *reduced* system. This technique is often called *odd–even reduction.*

Now apply odd–even reduction to the reduced system to obtain another reduced system that is again about half as big as the first reduced system. The recursive application of odd–even reduction continues for $q = \log_2(n + 1)$ steps. At each step, the even equations of the previous step form a reduced tridiagonal system of about half the size of the system from the previous step. At the end of step q, one equation in one unknown remains, so that unknown can be computed easily. This recursive technique is often called *cyclic reduction.*

Then the computation continues in the reverse order with a process called *back substitution.* At each step of back substitution, the even variables are known from the solution of the associated reduced system of about half the size. Substituting these values back into the odd equations of the larger system, we can easily compute all the odd variables.

Both the cyclic-reduction algorithm and the back-substitution algorithm require $q = \log_2(n + 1)$ steps each. In cyclic reduction, the number of floating-point operations is divided by 2 at each step, starting with $O(n)$ floating-point operations in the first step. Thus, it requires $O(n \log n)$ arithmetic operations. Similarly, back substitution also requires $O(n \log n)$ arithmetic operations. Therefore, the computational complexity of the full solution is $O(n \log n)$.

Observe that the algorithm described above is highly parallel. The elimination operations applied to a group of three rows to obtain the reduced system at each step are independent of the elimination operations applied to any other group of three rows, and so can be carried out in parallel. Similarly, the substitution operations to compute the odd unknowns in a reduced system given the even ones are independent of each other. Thus, the unknowns of each back-substitution step can also be computed in parallel.

Assume that we have $p = (n - 1)/2$ processors. Initially, processor 1 is assigned rows (1,2,3), processor 2 is assigned rows (3,4,5), processor 3 is assigned rows (5,6,7), and so on. After the first odd–even reduction step, processor 2 will use equation 2 from processor 1, equation 4 from itself, and equation 6 from processor 3. Similarly, processor 4 will use equation 6 from processor 3, equation 8 from itself, and equation 10 from processor 5, and so on. Only the even processors will continue. The procedure is repeated. The final reduced system is solved by one processor. For the back substitution, one processor works first, then two, then four, and so on.

Thus, the algorithm requires $2 \log_2(n + 1)$ steps with a constant amount of computation done on each processor per step, So the parallel computational complexity of the algorithm is $O(\log n)$, which is a factor of $O(n) = O(p)$ improvement over the $O(n \times \log n)$ computational complexity of the serial version of the algorithm, and a little less than $O(n)$ improvement over the $O(n)$ serial computational complexity of the standard *LU* factorization algorithm for tridiagonal systems. So we can say that, asymptotically, the algorithm has perfect *speedup.*

Note that when a processor uses rows computed by another processor, some communication and/or synchronization must take place between processors. This may degrade the parallel performance of the algorithm from the perfect asymptotic performance. The time spent in communication and/or synchronization depends heavily on the way the processors cooperate. More specifically, it depends on the interconnection network between processors and on the implementation of specific hardware instructions.

For an introduction to parallel numerical methods, see Bertsekas and Tsitsiklis (1989), Ortega (1988), or Van de Velde (1994).

13. SOURCES OF NUMERICAL SOFTWARE

Although most of this article has dealt with elementary numerical methods, we strongly recommend that readers do not program these schemes themselves. High-quality software incorporating these—or more sophisticated—numerical methods is readily available. In addition, good library routines often contain many additional strategies and heuristics (not discussed here) to improve their efficiency and reliability. Using such routines, rather than attempting to reprogram them, will likely save readers a significant amount of time as well as produce superior numerical results.

We highly recommend that readers familiarize themselves with the *Guide to Available Mathematical Software* (GAMS) recently developed by the National Institute of Standards and Technology (NIST). GAMS is both an on-line cross-index of available mathematical software as well as a repository of some 9000 high-quality problem-solving modules from more than 80 software packages. It provides centralized access to such items as abstracts, documentation, and source code of the software modules that it catalogs. Most of this software represents FORTRAN subprograms for mathematical problems that commonly occur in computational science and engineering, such as solution of systems of linear algebraic equations, computing matrix eigenvalues, solving nonlinear systems of differential equations, finding minima of nonlinear functions of several variables, evaluating the special functions of applied mathematics, and performing nonlinear regression. Among the packages cataloged in GAMS are

- the IMSL, NAG, PORT, and SLATEC libraries;
- the BLAS, EISPACK, FISHPAK, FNLIB, FFTPACK, LAPACK, LINPACK, and STARPAC packages;
- the DATAPLOT and SAS statistical analysis systems;
- the netlib routines, including the Collected Algorithms of the ACM (see below).

Note that although GAMS catalogs both public-domain and proprietary software, source code of proprietary software is not available through GAMS, although related items such as documentation and example programs often are. Software can be found either by browsing through a decision tree or performing a key-word search. GAMS can be accessed in several ways: telnet gams.nist.gov; gopher gams.nist.gov; or ⟨www browser⟩ http://gams.nist.gov, where ⟨www browser⟩ is a World Wide Web browser such as Mosaic or Netscape. Report any questions or problems to gams@cam.nist.gov. For more details, log in to the system or see Boisvert *et al.* (1985) and Boisvert (1990).

Included in the software cataloged by GAMS are many high-quality public-domain routines available by electronic mail (e-mail) from *netlib*. These routines are now also available through Xnetlib, a more sophisticated X interface to netlib and the NA-Net Whitepages, or through the World Wide Web at the address http://www.netlib.org/index.html. For more information on netlib, see Dongarra and Grosse (1987) or Dongarra *et al.* (1995), send the message "send index" by e-mail to either netlib@ornl.gov or netlib@research.att.com, or access http://www.netlib.org/index.html through the World Wide Web.

The ACM Transactions on Mathematical Software publishes refereed public-domain software. These high-quality routines, covering a broad range of problem areas, are included in the Collected Algorithms of the ACM, available through both GAMS and netlib.

Not mentioned above are the commercial interactive packages MATLAB, Maple, and Mathematica. MATLAB is built upon a foundation of sophisticated matrix software and includes routines for solving many standard mathematical and statistical problems. In addition, "toolboxes" for several application areas, such as control theory, are available. Both Maple and Mathematica are primarily symbolic algebra packages, but contain many high-quality numerical routines as well. For more information on MATLAB, contact The MathWorks Inc., 24 Prime Park Way, Natick, MA 01760; phone: (508) 653-1415; FAX: (508) 653-2997; e-mail: info@mathworks.com. For more information on Maple, contact Waterloo Maple Software, 450 Phillip St., Waterloo, Ontario, Canada, N2L 5J2; phone: (519) 747-2373; FAX: (519) 747-5284; e-mail: info@maplesoft.on.ca. For more information on Mathematica, contact Wolfram Research Inc., 100 Trade Center Dr., Champaign, IL 61820-7237; phone: (217) 398-0700; FAX: (217) 398-0747; e-mail: info@wri.com.

GLOSSARY

Banded Matrix: An $m \times n$ matrix all of whose nonzero elements occur in a band around its main diagonal.

Block Diagonal: See **Diagonal.**

Characteristic Equation of a Matrix: Defined only for an $n \times n$ matrix A, the equation $\det(a - \lambda I)$.

Characteristic Polynomial of a Matrix: Defined only for an $n \times n$ matrix A, the polynomial $p(\lambda) = \det(A - \lambda I)$, of degree n.

Column-Diagonally Dominant: Describing a matrix A if A^T is **row-diagonally dominant.**

Diagonal: Describing an $m \times n$ matrix $D = [d_{ij}]$ for which $d_{ij} = 0$ for $i \neq j$. The matrix D is *block-diagonal* if each d_{ij} in the preceding definition is a submatrix rather than a single number.

Diagonally Dominant: Describing a matrix A if either A or A^T is **row-diagonally dominant.**

Eigenvector: Defined only for an $n \times n$ matrix A, a (possibly complex) number λ such that for some nonzero vector x, $Ax = \lambda x$. Any vector satisfying this equation is an *eigenvector* associated with the eigenvalue λ.

Euclidean Norm (of a vector x): The quantity $\|x\|_2 = (x^H x)^{1/2}$, where x^H is the complex-conjugate transpose of the vector x. Often called the *2-norm.*

Flop: A floating-point operation on a computer; a multiplication and either an addition or a subtraction.

Hermitian: Describing an $n \times n$ matrix A for which $A^H = A$, where A^H is the complex-conjugate transpose of A.

Hessenberg: Either **upper** or **lower Hessenberg.** A symmetric Hessenberg matrix is tridiagonal.

Leading Principal Minor of a Matrix A: The $k \times k$ submatrix in the top left corner of A.

Lower Hessenberg: Describing a matrix A for which $a_{ij} = 0$ for $i < j - 1$. That is, it is lower-triangular except for a single nonzero superdiagonal.

Lower Triangular: Describing an $n \times n$ matrix $L = [l_{ij}]$ for which $l_{ij} = 0$ for $1 \le i < j \le n$. It is *strictly lower-triangular* if $l_{ij} = 0$ for $1 \le i \le j \le n$, and *block lower-triangular* or *block strictly lower-triangular,* respectively, if each l_{ij} in the preceding definitions is a submatrix rather than a single number.

Orthogonal: Describing an $m \times n$ matrix Q for which $Q^T Q = I$.

Permutation Matrix: An $n \times n$ matrix with exactly one element in each row and column equal to 1 and all other elements equal to 0.

Rank of a Matrix: The maximal number of independent rows (or columns) of the matrix.

Right Triangular: Describing an $m \times n$ matrix $R = [r_{ij}]$ such that $r_{ij} = 0$ for $i > j$. If $m = n$, the terms *right triangular* and *upper triangular* are equivalent.

Row-Diagonally Dominant: Describing an $m \times n$ matrix, with $m \le n$, for which

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| < |a_{ii}|$$

for $i = 1, 2, \ldots, m$.

Similar Matrices: Any two matrices A and B such that $B = WAW^{-1}$ for some nonsingular matrix W, which is called the *associated similarity transformation.*

Sparse: Describing a matrix in which the number of nonzero elements is much less than the total number of elements in the matrix.

Spectral Radius of a (Square) Matrix A: The quantity $\rho(A) = \max\{|\lambda| : \lambda \text{ an eigenvalue of } A\}$.

Strictly Lower Triangular: See **Lower Triangular.**

Strictly Upper Triangular: See **Upper Triangular.**

Symmetric: Describing an $n \times n$ matrix for which $A^T = A$, where A^T is the transpose of A.

Symmetric Indefinite: Describing a real symmetric matrix A for which $x^T A x > 0$ for some real n-vector x and $y^T A y < 0$ for some real n-vector y.

Symmetric Positive (Negative) Definite: Describing a real symmetric matrix A for which $x^T A x > 0$ ($x^T A x < 0$) for all real n-vectors $x \neq 0$.

Symmetric Positive (Negative) Semidefinite: Describing a real symmetric matrix A for which $x^T A x \ge 0$ ($x^T A x \le 0$) for all real n-vectors $x \neq 0$.

Transpose: For an $m \times n$ matrix $A = [a_{ij}]$, the $n \times m$ matrix $A^T = [a^t_{ij}]$ where $a^t_{ij} = a_{ji}$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. The transpose of a column (row) vector is a row (column) vector.

Unit Lower Triangular: Describing a lower-triangular matrix $L = [l_{ij}]$ for which $l_{ii} = 1$ for $i = 1, 2, \ldots, n$.

Upper Hessenberg: Describing a matrix $A = [a_{ij}]$ such that $a_{ij} = 0$ for $i > j + 1$. That is, it is upper-triangular except for a single nonzero subdiagonal.

Upper Triangular: An $n \times n$ matrix $U = [u_{ij}]$ such that $u_{ij} = 0$ for $1 \le j < i \le n$. It is *strictly upper-triangular* if $u_{ij} = 0$ for $1 \le j \le i \le n$, and *block upper-triangular* or *block strictly upper-triangular,* respectively, if each u_{ij} in the preceding definitions is a submatrix rather than a single number.

Tridiagonal: Describing an $m \times n$ matrix $A = [a_{ij}]$ such that $a_{ij} = 0$ for $|i - j| > 1$.

MATHEMATICAL SYMBOLS USED

A^H: The complex-conjugate transpose of the matrix A.

A^T: The transpose of the matrix A.

$\mathbb{C}$: The set of complex numbers.

$\mathbb{C}^n$: The set of complex vectors with n components.

$\mathbb{C}^{m \times n}$: The set of complex $m \times n$ matrices.

$\mathscr{C}$: The set of continuous functions.

$\mathscr{C}[a,b]$: The set of continuous functions on the interval $[a,b]$.

$\mathscr{C}^p$: The set of continuous functions with p continuous derivatives.

$\mathscr{C}^p[a,b]$: The set of continuous functions with p continuous derivatives on the interval $[a,b]$.

$O(h^p)$: Any quantity that depends on h that can be bounded above by Ch^p for some constant C and all $h \in (0,H]$ for some $H > 0$.

$O(n^p)$: Any quantity that depends on n that can be bounded above by Cn^p for some constant C and all positive integers n.

$\mathbb{R}$: The set of real numbers.

$\mathbb{R}^n$: The set of real vectors with n components.

$\mathbb{R}^{m \times n}$: The set of real $m \times n$ matrices.

x^T: The transpose of the vector x.

$\|x\|$ and $\|A\|$: Norms of the vector x and the matrix A, respectively.

$\|x\|_2$ and $\|A\|_2$: Euclidean norms (also called two-norms) of the vector x and the matrix A, respectively.

z^H: The complex-conjugate transpose of the vector z.

$\rho(A)$: The spectral radius of a square matrix A.

ABBREVIATIONS USED

ACM: Association for Computing Machinery.

ADI: Alternating direction implicit.

BC: Boundary condition.

BVP: Boundary-value problem.

CD: Conjugate direction.

CG: Conjugate gradient.

DOF: Degrees of freedom.

FD: Finite difference.

FDM: Finite-difference method.

FE: Finite element.

FEM: Finite-element method.

GAMS: Guide to Available Mathematical Software.

GE: Gaussian elimination.

IC: Initial condition.

ICF: Incomplete Cholesky factorization.

IEEE: Institute of Electrical and Electronics Engineers.

IVP: Initial-value problem.

LMF: Linear multistep formula.

MM: Multigrid method.

ODE: Ordinary differential equation.

PCG: Preconditioned conjugate gradient.

PDE: Partial differential equation.

PP: Piecewise polynomial.

RK: Runge–Kutta.

SD: Steepest descent.

SOR: Successive overrelaxation.

SPD: Symmetric positive-definite.

SSOR: Symmetric successive overrelaxation.

w.r.t.: With respect to.

Works Cited

Ames, W. F. (1992), *Numerical Methods for Partial Differential Equations,* New York: Academic.

Ascher, U. M., Mattheij, R. M. M., Russell, R. D. (1988), *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations,* Englewood Cliffs, NJ: Prentice-Hall.

Atkinson, K. E. (1989), *An Introduction to Numerical Analysis,* 2nd ed., New York: Wiley.

Axelsson, O. (1994), *Iterative Solution Methods,* Cambridge, U.K.: Cambridge Univ. Press.

Becker, E. B., Carey, G. F., Oden, J. T. (1981), *Finite Elements,* Vol. I, Englewood Cliffs, NJ: Prentice-Hall.

Bertsekas, D. P., Tsitsiklis, J. N. (1989), *Parallel and Distributed Computation: Numerical Methods,* Englewood Cliffs, NJ: Prentice-Hall.

Boisvert, R. (1990), "The Guide to Available Mathematical Software Advisory System," in: E. Houstis, J. Rice, R. Vichnevetsky (Eds.), *Intelligent Mathematical Software Systems,* Amsterdam: North-Holland, pp. 167–178.

Boisvert, R. F., Howe, S. E., Kahaner, D. K. (1985), "GAMS: A Framework for the Management of Scientific Software," *ACM Trans. Math. Software* **11** (4), 313–355.

Briggs, W. L. (1987), *A Multigrid Tutorial,* Philadelphia: SIAM.

Buchanan, J. L., Turner, P. R. (1992), *Numerical Methods and Analysis,* New York: McGraw-Hill.

Butcher, J. C. (1987), *The Numerical Analysis of Ordinary Differential Equations,* New York: Wiley.

Celia, M. A., Gray, W. G. (1992), *Numerical Methods for Differential Equations,* Englewood Cliffs, NJ: Prentice-Hall.

Christara, C. C. (1994), "Quadratic Spline Collocation Methods for Elliptic Partial Differential Equations," *BIT* **34** (1), 33–61.

Conte, S. D., de Boor, C. (1980), *Elementary*

Numerical Analysis, 3rd ed., New York: McGraw-Hill.

Cullum, J., Willoughby, R. (1985), *Lanczos Algorithms for Large Symmetric Eigenvalue Computations,* Vol. 1, *Theory,* and Vol. 2, *Programs,* Boston: Birkhäuser.

Dahlquist, G., Björck, Å. (1974), *Numerical Methods,* Englewood Cliffs, NJ: Prentice-Hall.

Davis, P. J. (1975), *Interpolation and Approximation,* New York: Dover.

de Boor, C. (1978), *A Practical Guide to Splines,* New York: Springer-Verlag.

Dennis, J. E., Schnabel, R. B. (1983), *Numerical Methods for Unconstrained Optimisation and Nonlinear Equations,* Englewood Cliffs, NJ: Prentice-Hall.

Dongarra, J., Grosse, E. (1987), "Distribution of Mathematical Software via Electronic Mail," *Commun. ACM* **30** (5), 403–407.

Dongarra, J., Rowan, T., Wade, R. (1995), "Software Distribution Using XNETLIB," *ACM Trans. Math. Software* **21** (1), 79–88.

Duff, I., Erisman, A., Reid, J. (1986), *Direct Methods for Sparse Matrices,* Oxford, U.K.: Oxford Univ. Press.

George, A., Liu, J. (1981), *Computer Solution of Large Sparse Positive Definite Systems,* Englewood Cliffs, NJ: Prentice-Hall.

Goldberg, D. (1991), "What Every Computer Scientist Should Know about Floating-Point Arithmetic," *ACM Comput. Surveys* **23,** 5–48.

Golub, G. H., Van Loan, C. F. (1989), *Matrix Computations,* 2nd ed., Baltimore: John Hopkins Univ. Press.

Hackbusch, W. (1994), *Iterative Solution of Large Sparse Systems of Equations,* New York: Springer-Verlag.

Hageman, L. A., Young, D. M. (1981), *Applied Iterative Methods,* New York: Academic.

Hager, W. W. (1988), *Applied Numerical Linear Algebra,* Englewood Cliffs, NJ: Prentice-Hall.

Hairer, E., Nørsett, S. P., Wanner, G. (1987), *Solving Ordinary Differential Equations I: Nonstiff Problems,* New York: Springer-Verlag.

Hairer, E., Wanner, G. (1991), *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems,* New York: Springer-Verlag.

Hall, C. A., Porsching, T. A. (1990), *Numerical Analysis of Partial Differential Equations,* Englewood Cliffs, NJ: Prentice-Hall.

Householder, A. (1970), *The Numerical Treatment of a Single Nonlinear Equation,* New York: McGraw-Hill.

Houstis, E. N., Christara, C. C., Rice, J. R. (1988), "Quadratic Spline Collocation Methods for Two-Point Boundary Value Problems," *Int. J. Numer. Methods Eng.* **26,** 935–952.

IEEE (1985), *IEEE Standard for Binary Floating-Point Arithmetic,* New York: American National Standards Institute, ANSI/IEEE Std. 754-1985.

Johnson, L. W., Riess, R. D. (1982), *Numerical Analysis,* Reading, MA: Addison-Wesley.

Kahaner, D., Moler, C., Nash, S. (1989), *Numerical Methods and Software,* Englewood Cliffs, NJ: Prentice-Hall.

Lambert, J. D. (1991), *Numerical Methods for Ordinary Differential Equations,* New York: Wiley.

Ortega, J. M. (1988), *Introduction to Parallel and Vector Solution of Linear Systems,* New York: Plenum.

Parlett, B. N. (1968), "Global Convergence of the Basic QR Algorithm on Hessenberg Matrices," *Math. Comput.* **22,** 803–817.

Parlett, B. N. (1980), *The Symmetric Eigenvalue Problem,* Englewood Cliffs, NJ: Prentice-Hall.

Prenter, P. M. (1975), *Splines and Variational Methods,* New York: Wiley.

Saad, Y. (1992), *Numerical Methods for Large Eigenvalue Problems,* New York: Manchester Univ. Press (Wiley).

Scott, D. (1981) "The Lanczos Algorithm," in: I. S. Duff (Ed.), *Sparse Matrices and Their Uses,* London: Academic, pp. 139–160.

Shampine, L. F. (1994), *Numerical Solution of Ordinary Differential Equations,* New York: Chapman & Hall.

Shampine, L. F., Gear, C. W. (1979), "A User's View of Solving Stiff Ordinary Differential Equations," *SIAM Rev.* **21,** 1–17.

Stoer, J., Bulirsch, R. (1980), *Introduction to Numerical Analysis,* New York: Springer-Verlag.

Strang, G., Fix, G. J. (1973), *An Analysis of the Finite Element Method,* Englewood Cliffs, NJ: Prentice-Hall.

Strikwerda, J. C. (1989), *Finite Difference Schemes and Partial Differential Equations,* Pacific Grove, CA: Wadsworth and Brooks/Cole.

Van de Velde, E. F.. (1994), *Concurrent Scientific Computing,* New York: Springer-Verlag.

Van Loan, C. F. (1992), *Computational Frameworks for the Fast Fourier Transform,* Philadelphia: SIAM.

Varga, R. S. (1962), *Matrix Iterative Analysis,* Englewood Cliffs, NJ: Prentice-Hall.

Wilkinson, J. H. (1965), *The Algebraic Eigenvalue Problem,* Oxford, U.K.: Oxford Univ. Press.

Young, D. M. (1971), *Iterative Solution of Large Linear Systems,* New York: Academic.

Further Reading

Ciarlet, P. G. (1989), *Introduction to Numerical Linear Algebra and Optimization,* Cambridge, U.K.: Cambridge Univ. Press.

Forsythe, G. E., Malcolm, M. A., Moler, C. B. (1977), *Computer Methods for Mathematical Computations,* Englewood Cliffs, NJ: Prentice-Hall.

Forsythe, G. E., Moler, C. B. (1967), *Computer*

Solution of Linear Algebraic Systems, Englewood Cliffs, NJ: Prentice-Hall.

Golub, G. H., Ortega, J. M. (1992), *Scientific Computing and Differential Equations,* New York: Academic.

Golub, G. H., Ortega, J. M. (1993), *Scientific Computing: An Introduction with Parallel Computing,* New York: Academic.

Isaacson, E., Keller, H. B. (1966), *Analysis of Numerical Methods,* New York: Wiley.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T. (1986), *Numerical Recipes: The Art of Scientific Computing,* Cambridge, U.K.: Cambridge Univ. Press.

Schultz, M. H. (1973), *Spline Analysis,* Englewood Cliffs, NJ: Prentice-Hall.

Stewart, G. W. (1973), *Introduction to Matrix Computations,* New York: Academic.

Wilkinson, J. H. (1963), *Rounding Errors in Algebraic Processes,* Englewood Cliffs, NJ: Prentice-Hall.

SCINTILLATION DETECTORS

See DETECTORS, SCINTILLATION

SEA ICE

STEPHEN F. ACKLEY, *Cold Regions Research and Engineering Laboratory, Hanover, New Hampshire, U.S.A.*

	Introduction	81
1.	**Sea-Ice Formation and Growth**	83
1.1	Atomic Structure of Ice	83
1.2	Freezing of Sea Water	85
1.2.1	Constitutional Supercooling	85
1.2.2	Kinetic Crystal Growth	85
1.2.3	Brine Entrapment	85
1.3	Growth of Sea-Ice Sheets	86
1.4	Incidental Processes of Sea-Ice Growth	88
1.4.1	Snow Ice	88
1.4.2	Platelet Ice	89
1.4.3	Ice Stalactites	89
1.5	Phase Relations in Sea Ice	89
2.	**Sea-Ice Biology**	90
2.1	Sea Ice as Habitat	90
2.1.1	Surface Communities	91
2.1.2	Interior Communities	92
2.1.3	Bottom Communities	93
3.	**Distribution of Sea Ice**	93
3.1	Ice-Thickness Distribution	93
3.2	Perennial Ice	97
3.3	Seasonal Sea Ice	98
3.3.1	Fast-Ice Zones	98
3.3.2	Marginal-Ice Zone	98
3.4	Geophysical Effects on Ice Distribution	98
3.4.1	Ice Circulation and Deformation in the Arctic and Antarctic	99
4.	**Conclusion—Sea Ice, Complex Response in a Changing Climate**	101
	Glossary	102
	Works Cited	102
	Further Reading	103

INTRODUCTION

Sea ice is one form of natural ice in the environment, along with glaciers, seasonal snow, frozen ground, and the ice formed on lakes and rivers. It is distinguished by its formation from bodies of water that contain significant concentrations of salts, the oceans and seas. Ice formed in salt water is significantly different from fresh-water ice in both its growth behavior and structure. Sea ice is more akin to a multiphase alloy than it is to the single–solid-phase, purer ice formed in other environments.

The area covered by sea ice, seasonally or perennially, is restricted to higher latitudes of the northern and southern hemispheres (Fig. 1). In lower latitudes, isolated areas are also found in estuaries, embayments, and sounds when exceptionally cold winters occur. The ice covers about 10% (35×10^6 km^2) of the area of the ocean at its maximum seasonal extent, in a relatively thin veneer varying in thickness from a few tens of centimeters to several meters. Deformed ice can occasionally be tens of meters thick, and in actively deforming regions deformation can alter the mean thickness by several meters of ice.

The climatic and oceanic influences of the ice covers are even greater than might be expected, even considering their substantial area, since the ice can strongly change the air–sea interaction in its presence. For example, the ice albedo reflects 80% or more of the incoming solar radiation compared with 10% for non-ice-covered water, and contributes to lower air temperatures. Sea-ice formation influences oceanic processes: Salts are rejected into the underlying water when the ice forms, changing its salinity and density and causing convective overturning. These denser waters then flow outward from the polar regions at great depths and are replaced by other water masses flowing in from temperate latitudes. Sea-ice growth therefore plays a significant role in the large-scale thermohaline circulation of the world ocean (called the thermohaline conveyor belt). Sea ice can influence biological pro-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

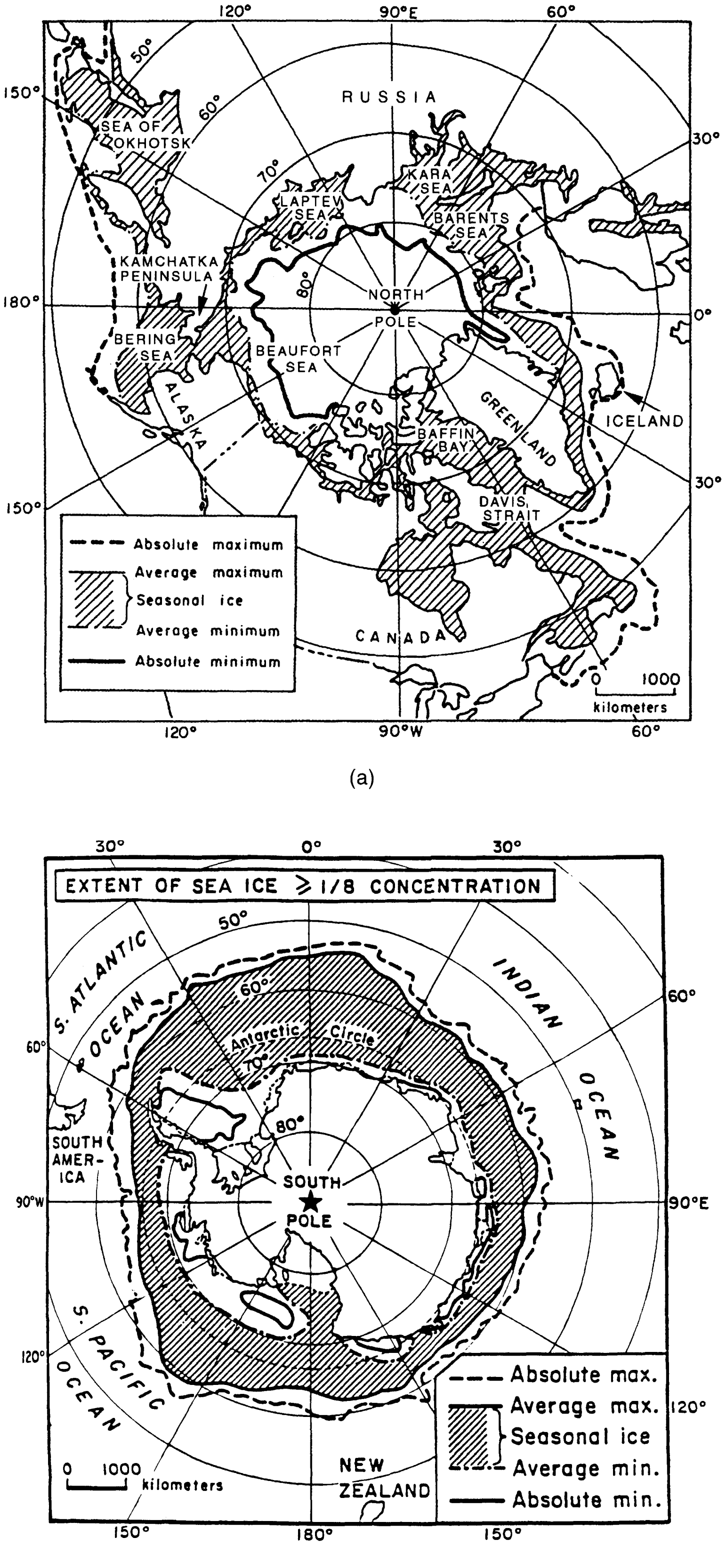
120°
90°E
60°
150°
30°
180°
0°
150°
30°
120°
90°W
60°
50°
60°
70°
80°
RUSSIA
SEA OF OKHOTSK
LAPTEV SEA
KARA SEA
BARENTS SEA
KAMCHATKA PENINSULA
NORTH POLE
BERING SEA
BEAUFORT SEA
ALASKA
GREENLAND
ICELAND
BAFFIN BAY
DAVIS STRAIT
CANADA
0 1000 kilometers
Absolute maximum
Average maximum
Seasonal ice
Average minimum
Absolute minimum
30°
0°
30°
EXTENT OF SEA ICE ≥ 1/8 CONCENTRATION
50°
60°
Antarctic Circle
70°
80°
S. ATLANTIC OCEAN
INDIAN OCEAN
60°
60°
SOUTH AMERICA
90°W
SOUTH POLE
90°E
S. PACIFIC OCEAN
120°
120°
0 1000 kilometers
NEW ZEALAND
150°
180°
150°
Absolute max.
Average max.
Seasonal ice
Average min.
Absolute min.

(a)

(b)

FIG. 1. Sea-ice extents in the (a) northern and (b) southern hemispheres (after Barry, 1986).

duction in the ocean as well. Ice melt is a mechanism for stabilizing the upper layer of the ocean by lowering the salinity and, therefore, the density of the surface layer. This stabilizing effect allows biological organisms to remain longer in this upper layer, or photic zone, where photosynthesis is greater, enhancing primary production in these ice edge areas.

The properties of sea ice are sensitive to the growth conditions and, after formation of the initial ice cover, its thermal and dynamic history. These conditions are both spatially and temporally variable, leading to significant differences in sea-ice behavior from location to location and season to season. Our theme is to develop the relationship between these processes and the resulting ice properties and ice distribution. We then review how the sea ice interacts with the geophysical and biological environments.

1. SEA-ICE FORMATION AND GROWTH

Sea ice, at temperatures generally seen in the natural environment, is dominated by two phases: salt-free ice, and liquid brine with a high concentration of salts. Depending on the temperature, history of development, and constituents of the sea water, it can also contain varying amounts of gases (primarily air), solid salts, organic inclusions (such as phytoplankton and bacteria), and inorganic inclusions (such as bottom sediments and wind-transported dust).

The two-phase nature of sea ice arises from the relative insolubility of salts in ice, as opposed to water, and a process of brine entrapment during ice growth. The equilibrium solute partition coefficient (k_0) for most solutes in ice (the ratio of the mole fraction of solute in the solid to that in the liquid phase) is about $k_0 < 10^{-4}$. This value suggests that the number of salt impurities that enter substitutionally into the ice lattice is small, so that the ice can be considered a pure phase (Weeks and Ackley, 1986) for most applications. This relative purity and the properties of the ice in the sea-ice matrix arise from its lattice structure.

1.1 Atomic Structure of Ice

A well-known attribute of water, unusual relative to most substances, is that its solid phase is less dense than its liquid phase: i.e., ice floats. Other properties, such as ice's unusually high melting point, and its high strength at high homologous temperatures (temperature relative to its melting point) compared with hydrogen compounds of similar molecular weight (HF, H_2S, CH_4), all result from the hydrogen bonds between water molecules in the ice-crystal lattice. Water is a polar molecule consisting of two relatively mobile protons (hydrogen atoms) located at the two positively charged sites on the molecule and two negative charge clouds associated with the oxygen atom, all distributed around the oxygen atom at roughly equal angles to each other. Four bonds, corresponding to these charge clouds, "point" to the vertices of a tetrahedron with the oxygen atom at the center (Fig. 2), an arrangement identified as tetrahedral coordination (Hobbs, 1974). In the ice lattice, the protons, because of their small size and mobility, can arrange to point to the negative ends of adjacent water molecules, forming hydrogen bonds. Hydrogen bonds are intermediate in bond strength, falling between a high-strength covalent bond (where electrons, rather than protons, are shared between atoms) and a weak ionic bond (where electrons are given up to the other atom rather than shared).

Ice, as a hydrogen-bonded structure, obeys the following rules:

1. Each of the four bonds in ice that connect it to other water molecules has only one hydrogen on it; and
2. of the four hydrogens around an oxygen

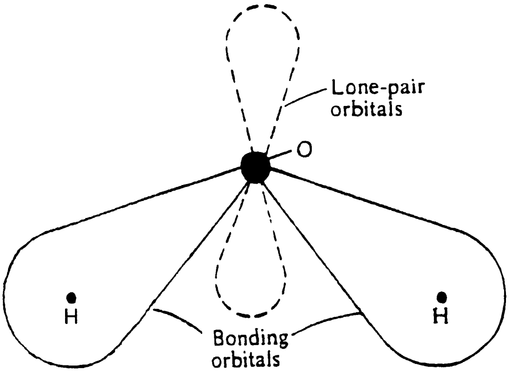

FIG. 2. Tetrahedral bond structure of the water molecule (after Hobbs, 1974).

atom, only two are close to the oxygen atom, so at each lattice site it "looks" like a water molecule (Weeks and Ackley, 1986; Hobbs, 1974).

The hydrogen bond is linear, giving ice its open crystal-lattice structure and its lower density relative to water, in which the hydrogen bonds are broken and the molecules can pack more closely. The tetrahedral coordination gives ice a crystal structure with hexagonal symmetry, affecting many of the large-scale characteristics of both pure and sea ice.

While several crystal structures (polymorphs) of ice exist, the structure of the ice in sea ice, and the other ices that exist on the Earth's surface, is known as ordinary ice (ice I). The ice I crystal structure is such that the oxygen atoms are concentrated close to a series of parallel planes referred to as the basal or (0001) plane in Miller indices. A vector normal to the basal plane is referred to as the crystallographic or *c* axis. Within the basal plane, three equivalent axes (six directions ⟨11$\bar{2}$0⟩) form the *a* axes, which correspond to three equivalent close-packed rows of atoms (Fig. 3). The *a*-axis directions correspond to the arms of snowflakes growing from the vapor and, for our application, the arms of dendritic ice crystals growing in unstable or kinetic growth from a salt solution.

The relative insolubility of impurities in ice can be understood as a result of hydrogen bonding. Hydrogen bonding applies to the sharing of a proton between pairs of electronegative atoms. These are found to be oxygen, nitrogen, or fluorine (Hobbs, 1974). Because the electronegative impurities found in sea water, e.g., the chloride ion, do not partake in hydrogen bonds, they cannot substitute for oxygen in the ice lattice, and are rejected during the phase transition. Other electropositive impurities (e.g., the sodium ion) are much less mobile than the hydrogen atom and cannot bond between oxygen atoms, and so they too are rejected. The inclusion of gases in the interstices between the ice-lattice sites only occurs under much higher pressures than are found during the freezing of sea water. Air, for example, concentrates and nucleates into bubbles at a

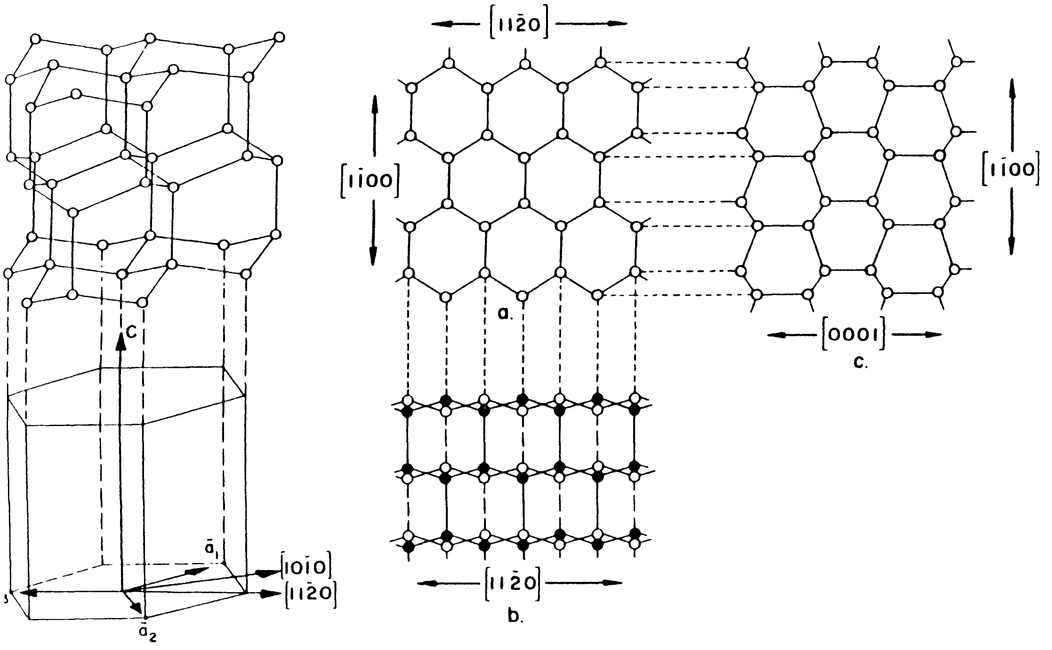

FIG. 3. Structure of ice; crystallographic axes are indicated. Views (a), (b), and (c) are planar sections of the lattice along the directions and planes indicated (after Weeks and Ackley, 1986).

growing ice interface and is not found within the ice lattice.

The impurities in sea water are therefore relatively insoluble in ice at lattice scales, and so the presence of liquid brine in sea ice arises from a macroscopic trapping process controlled by the growth behavior of ice.

1.2 Freezing of Sea Water

Ice growth from the liquid phase can take place under two thermodynamic conditions: stable, equilibrium, or planar growth; and unstable, kinetic, or dendritic growth. Kinetic growth commonly takes place in a supercooled condition of the liquid phase. It occurs in an episodic rapid way, triggered by a nucleation event, and uses the local supercooling as a heat sink for the latent heat released in the transition from the liquid to the solid phase. Dendritic growth is the dominant process in sea-ice growth, while planar growth occurs without supercooling and is seen only in fresh-water bodies. Because of the rapid growth of ice as dendrites and the relatively pure nature of the ice phase, spaces containing the liquid phase are also created, trapping brine within sea ice as it grows.

1.2.1 Constitutional Supercooling In a salt solution the equilibrium freezing point is a function of its composition; for example, sea water at 35 ppt (35 parts per thousand, or 35 g dissolved solids per kg of sea water) of salt has a freezing point of −1.8 °C. But the solution is supercooled if it is cooled to below its freezing point without nucleating ice crystals or changing its composition. At a growing ice interface, most of the salts are rejected. In the solution itself, however, the rate at which heat can be taken away is about ten times that at which salt can diffuse from a region of high to a region of low concentration. While the thermal diffusion lowers the temperature, the salt diffusion, necessary to lower the equilibrium freezing point of the rest of the solution, cannot keep pace with the thermal cooling effect, so that the solution becomes supercooled. Constitutional supercooling is, therefore, inherent in salt solutions.

1.2.2 Kinetic Crystal Growth The crystal anisotropy in ice, because of its structure as a hexagonal material, leads to a geometric selection in its growth from the melt. Crystal-growth theory (Weeks and Ackley, 1986) predicts that atoms arriving from solution at a crystal surface will have a much higher probability of remaining if they arrive at a step in the surface than if they arrive at an atomistically smooth face. Steps are common in ice where line lattice-defect sites (screw dislocations) intersect the crystal surface (Weeks and Ackley, 1986). Dislocations in ice are much more likely to occur in the basal plane (the plane normal to the *c* axis), serving as the step sites for preferential growth. In the early stages of ice growth, disks quickly form with their *c* axes normal to the plane of the disk, rather than as spheres that we might expect if the growth took place isotropically. The lateral growth rate of the disc is about 100 times greater than that which thickens it. These disks can then develop dendritic arms, still in the basal plane, pointing in the close-packed or *a* directions. Supercooling in salt solutions ensures that ice growth takes place kinetically and results in the preferred orientations of the disks and dendrites. This behavior occurs whether the ice forms as free-floating disks or is attached to the bottom of an existing ice sheet. The increase in thickness of a sea-ice sheet is by addition or extension of these dendrites into the supercooled solution below. If these dendrites are attached to an existing ice sheet, the resulting system is one of platelike structures with their basal planes in the growth direction or, equivalently, with their *c* axes perpendicular to the growth direction (Fig. 4).

1.2.3 Brine Entrapment As these ice platelets grow by "stabbing" into the supercooled solution nearby, salt is rejected by the growing pure-ice phase. The resulting solution near the ice crystal is increased in concentration by the rejection process. Since these salty regions are more dense than the surroundings, they descend (as plumes or blobs) and are replaced by upwelled water at the lower ambient concentration, and at the equilibrium freezing point for the solution. This fluid fills the spaces between the dendritic plates and is trapped as an inclusion, either when a new dendrite grows between the existing plates or as the ice plates gradually thicken and neck off the brine inclusion. These brine pockets, if there is a pre-

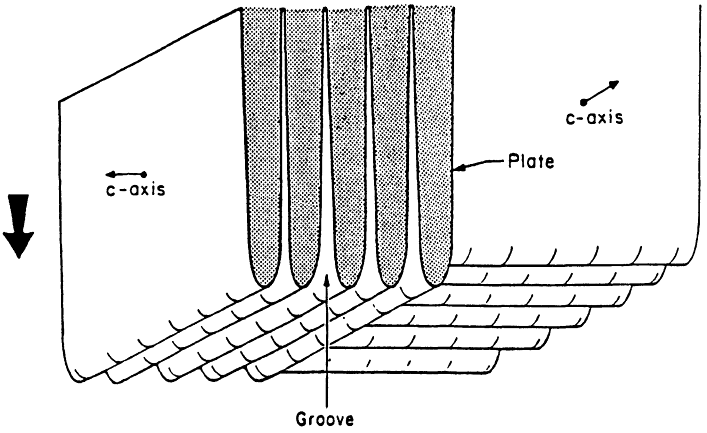

FIG. 4. Dendritic structure of columnar sea-ice crystals (after Gow and Tucker, 1991).

ferred orientation of the ice platelets, can then develop into parallel rows and reflect the local crystal structure in their orientation (Fig. 5). Once trapped, the brine pockets freeze or melt according to the phase relationships of the salt solution in ice. For example, the salt concentration in brine pockets in sea ice at −10 °C is approximately 115 ppt compared with 35 ppt at the freezing point for sea water (−1.8 °C).

1.3 Growth of Sea-Ice Sheets

The majority of sea-ice sheets observed in nature are formed by the kinetic growth of ice in sea water, either initially as free-floating disks, or as dendrites attached to an existing ice sheet, but usually as a combination of the two.

The initial ice formation (Fig. 6) that occurs at or near the surface of the sea water in the form of disks, needles, and platelets is called *frazil.* Continued freezing results in a soupy mixture—grease ice—consisting of sea water and unconsolidated frazil crystals. Wind-induced turbulence or wave action inhibits the immediate development of a solid ice cover. For example, in the presence of a sustained wave field, pancake ice (Fig. 7) forms by the packing and freezing together of the unconsolidated frazil crystals. The constant bumping of the pancakes then pushes frazil crystals up around their perimeters. When the wave action dies off, the pancakes consolidate by freezing together to form a continuous ice sheet. Pancake ice consolidation, up to about 1 m thickness, accounts for about 60% of the sea-ice cover at maximum extent in the Antarctic seas (Weeks and Ackley, 1986). The necessary turbulence at the seasonally advancing ice edge is provided by waves generated in the stormy oceans surrounding Antarctica.

The resulting ice structure of the consolidated sheet retains the crystal sizes and shapes from the initial turbulent formation. When viewed in thin section using polarized light, the texture of the ice sample, showing the individual crystals, exhibits a granular appearance (Fig. 8). Although each crystal probably grew as a disk in its basal plane, the final arrangement is relatively randomly oriented since the disks are pushed into their final configuration in no particular order or orientation. As the individual crystals freeze together, they trap off pockets of liquid between the individual crystal grains of pure ice, at separations corresponding to the crystal diameters, varying from a few tenths of a millimeter to a few millimeters (Weeks and Ackley, 1986). Granular ice texture, as a product of pancake-ice formation, therefore accounts for 60% of the volume of sea ice seen around Antarctica, corresponding to the area of pancake growth.

Under calm conditions, however, the frazil crystals in the grease ice quickly freeze together to form a solid, continuous ice cover with thickness between 1 and 10 cm. Once a continuous ice sheet forms, the underlying ocean is isolated from the cold air. The latent heat of freezing is lost upward through the ice sheet by thermal conduction, so that the growth rate is now determined by

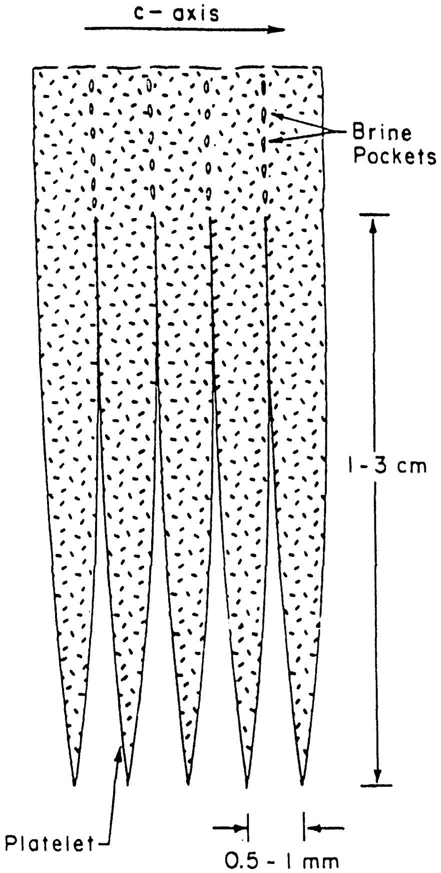

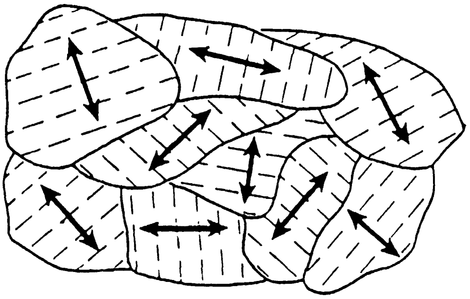

FIG. 5. Necking off of brine pockets from initial brine layers (grooves shown in Fig. 4) as ice growth and cooling take place (after Horner, 1985).

the temperature gradient in the ice and the ice's effective thermal conductivity. In addition, once a continuous sheet has formed, ice crystals lose a degree of growth freedom, i.e. they are attached to the sheet, and so thickening proceeds by freezing at the crys-

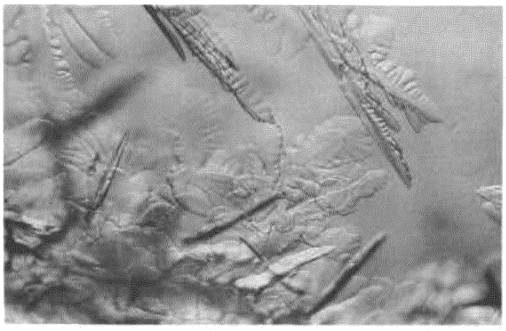

FIG. 6. Initial disks during the freezing of sea water (after Weeks and Ackley, 1986).

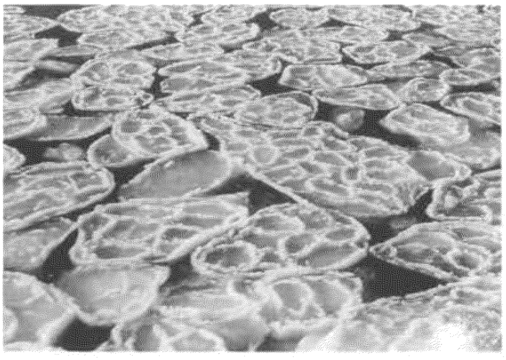

FIG. 7. Pancake ice forming in a wave field (after Gow and Tucker, 1991).

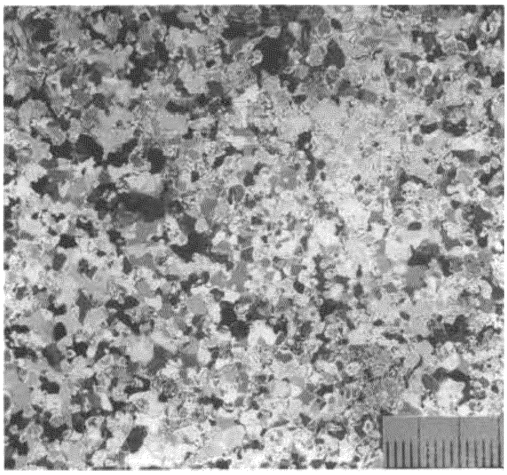

FIG. 8. Horizontal thin sections of granular ice. Scale divisions are mm. (After Gow and Tucker, 1991.)

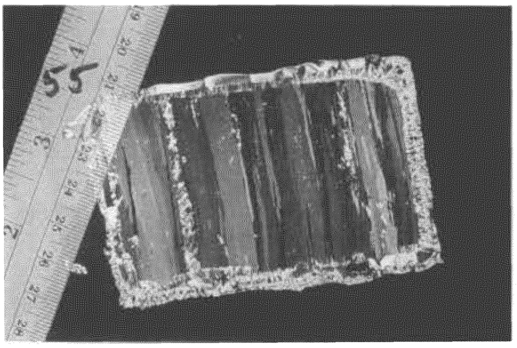

FIG. 9. Vertical thin section of columnar ice. Scale divisions are mm.

tals' dendritic ice-growth interfaces at the lower boundary (Fig. 4). Since the direction of growth is defined, the dendrites are oriented to the maximum growth velocity, i.e., so that their basal planes penetrate edge-on into the supercooled water below. Unlike the frazil ice in the upper layers, large sections of the plates (crystal domains) are oriented in the same direction, with layers of brine interspersed between them. In the vertical this yields regions of columnar structure (Fig. 9), each of the columns consisting of an array of ice platelets with a common crystal orientation, alternating with brine layers (Fig. 10). This substructure is sometimes called "intragranular" (within grain) as opposed to the "intergranular" (between grain) location of the brine substructure in frazil or granular ice. The spacing of these plates, or equivalently, the brine-layer separation, varies from a few tenths of a millimeter to about 1 mm, with the wider spacings occurring at the slowest growth rates (Weeks and Ackley, 1986).

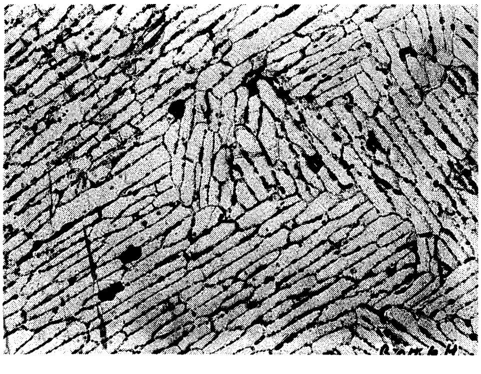

FIG. 10. Brine layers alternating with ice platelets, shown in horizontal thin section. The platelets are of the order of mm sizes, while the brine layers are ~0.1 mm.

In the crystal sense, the size of the individual ice plates is similar for sea ice of frazil or columnar origin, since the distance between the brine inclusions, representing the breakdown of the ice-crystal lattice spacing, is of the order of a few tenths to 1 or 2 mm in both structures. However, the regions over which arrays of these ice plates are ordered vary significantly between these two ice-growth regimes, as comparison of Figs. 8 and 9 indicates. Examination of ice structures by use of these sections therefore is a convenient way to determine whether the growth regime was turbulent (granular structure) or quiescent (columnar structure).

Sea ice in the Arctic Ocean, because of its isolation from surrounding seas by continental land masses, develops under calmer conditions than in the Antarctic seas. Structurally, therefore, the frazil or granular ice is confined to the top few centimeters of the sea ice, and the rest of the ice sheet (about 90%) consists of columnar ice. Because the Arctic Ocean is centered at the pole, while the Antarctic seas surround a continent at the pole, thermal conditions for sea ice formation are generally more severe in the Arctic, resulting in greater mean thickness: 2 m annually, compared with about 1 m in the Antarctic.

1.4 Incidental Processes of Sea-Ice Growth

Together the frazil and columnar structures account for 80%–90% of the initial growth of sea ice in the world's oceans. Other structures, identified as snow ice, platelet ice, and ice stalactites, account for the remaining 10%–20%. In some regions, however, these other forms can account for 30% or more of the ice structure, as, for example, platelet ice does within 50 km of the Antarctic coast.

1.4.1 Snow Ice Snow ice on sea-ice covers develops from the loading of the ice sheet by a snow cover. The snow cover depresses the ice-sheet surface below sea level and, if a crack occurs, sea water can laterally intrude at the snow–ice interface and flood the base of the snow cover. The snow–sea-water slurry or slush can then refreeze as a layer of imposed snow ice on the existing ice surface. Since the refreezing occurs in the slurry, the texture of snow ice is granular. It bears some strong similarities to frazil ice and is difficult to differentiate from frazil ice on structural grounds alone. However, oxygen isotope ratios in snow are quite different than in sea ice because snow forms by vapor diffusion while sea ice forms by freezing of sea water. A measurement of the relative oxygen isotope content in sea ice can therefore provide a well-defined measure of whether a sample is derived from snow or purely from sea water. These measurements in Antarctic sea ice have shown that snow ice accounts for a few percent of the total volume of the ice cover. However, it is estimated that 50% of the surface area of Antarctic sea ice is flooded at some time during its history, and so the surface layer of the sea ice there is in many instances dominated by snow-ice formation. Since remote sensing from space is heavily influenced by the surface condition, determination of ice properties from satellite imagery depends strongly on the degree of surface flooding at the time of the measurement.

1.4.2 Platelet Ice Platelet ice is formed at depth in the ocean and arises from the circulation of ocean waters, primarily under continental ice shelves, and their upwelling at the edges of the shelves. Water circulated at depth can be cooled to temperatures below the surface freezing point, since the freezing point is depressed under pressure. If the ice shelf bottom is sloping, dynamic effects can cause this water to upwell. As the water is brought to a shallower depth adiabatically, it is potentially supercooled, i.e., its actual temperature is lower than the freezing point for water of that composition at its new, shallower depth. If nucleation can occur, then ice-crystal growth can take place since the temperature difference between the higher freezing point and the actual temperature is a heat sink for the latent heat released by crystallization. These ice crystals develop as disks up to several centimeters in diameter and float upward. The disks either consolidate into a sheet at the surface or form cloudlike billow accumulations, up to several meters in thickness, at the base of an existing ice cover. These accumulations occur beneath both the ice shelves and adjacent sea-ice sheets. Their loosely stacked, unconsolidated nature gives them a relatively high proportion of water to ice (3 to 1) and provides a more freely exchanging environment, both for ocean water and biological organisms, than the more consolidated ice overhead. The association with potentially supercooled waters, however, confines platelet-ice development to near-shore regions, particularly in the Antarctic, where waters at depth can flow out from underneath the numerous ice shelves.

1.4.3 Ice Stalactites During the consolidation of an ice cover, brine at a temperature below the freezing point of the surrounding ocean water is rejected from the growing ice sheet. The brine is also more saline than the ocean water, as some of its fresh water has been removed by conversion to ice. As this cold plume descends into the ocean water, both saline and thermal diffusion take place between the plume and the surrounding water. Since the diffusivities for heat and salt are different, water surrounding the plume is cooled faster than it is compositionally changed, leading to the constitutional supercooling observed at the base of an ice sheet. Ice formation therefore takes place in the area adjacent to the plume. Continued rejection of the cold brine leads to the development of a feature similar in appearance to stalactites in caves. Ice stalactites are long features (10–100 cm) with a hollow core through which the brine plume flows, attached to the base of the ice sheet. They have been observed in new ice growth in cracks and under existing ice sheets in both the Arctic and Antarctic (Weeks and Ackley, 1986). While their contribution to the total ice growth is probably small, their appearance is striking (Fig. 11), and they may be dynamically important for fluid exchange from within the ice to the ocean mixed layer.

1.5 Phase Relations in Sea Ice

Once the ice cover has formed, the relationships between the ice, liquid brine, and solid salts are governed by the thermodynamic relationships shown in the phase diagram in Fig. 12 (Weeks and Ackley,

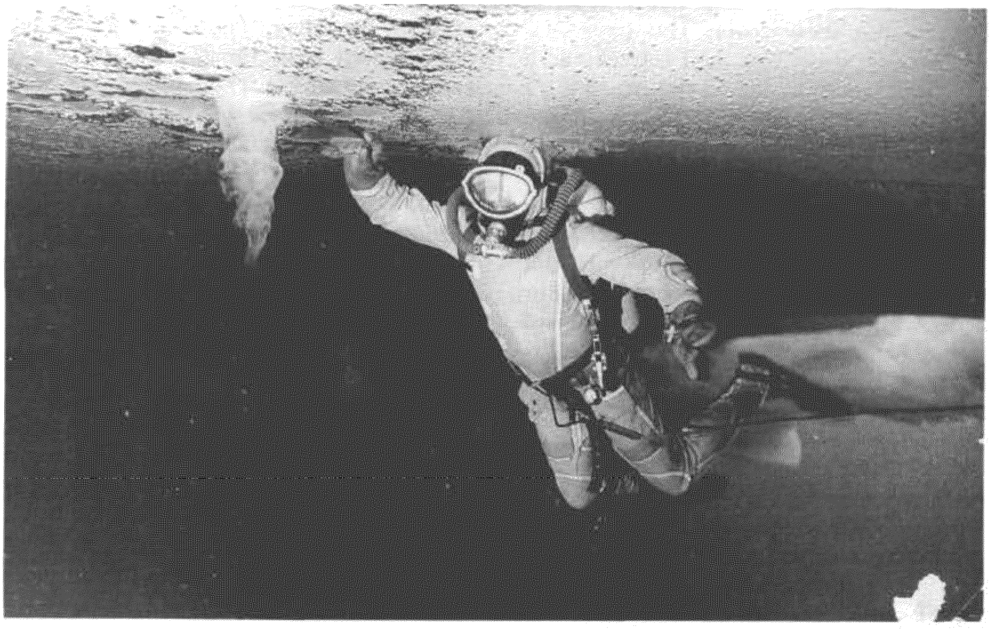

FIG. 11. Photograph of an ice stalactite.

1986). The phase diagram is a useful construct since it allows a computation of, for example, the amount of liquid in ice of a given temperature and composition. Since many of the properties of ice—mechanical, thermal, electromagnetic, and optical—depend on the liquid fraction, its computation is necessary to compare with measurements of these properties or to predict them. Since most naturally occurring sea ice is of lower salinity than the "standard" base for the phase diagram, the amounts of brine and ice are usually computed using equations that fit the slope of the phase diagram by a polynomial, and are then scaled by the ratio of the actual salinity of the ice sample to that of the standard. The measured density of the sample is also used to compute simultaneously the air volume, as the phase diagram gives the masses of ice, brine, and salts in a given mass of sea ice, and so the density is required to convert to volumetric units. The density, temperature, texture, and salinity of the ice sample, with use of the relationships given by the phase diagram, therefore provide a relatively complete set of information on its physical-chemical state.

2. SEA-ICE BIOLOGY

2.1 Sea Ice as Habitat

While our superficial view suggests that ice in general is an inhospitable environment, a closer examination shows otherwise, at least for sea ice. The two-phase nature of sea ice, with sea water, usually enriched in salts and nutrients, entrapped within it, actually provides a microenvironment. Organisms that can fit, tolerate the cold, and adapt to the saline and light changes typical of the ice cover will survive and sometimes thrive there. Most common are marine algae and diatoms (the photosynthetically active organisms), but protists and zooplankton that feed on algae, and the bacteria that decompose dead algal cells and other detritus, are also significant components of the biology in sea ice (Horner, 1985). Especially in the Antarctic, ice habitat has developed in a number of forms falling into three basic categories: those that are found on the top surface, in the interior, or at the bottom surface of the ice (Fig. 13) (Horner *et al.*, 1992). The type of habitat depends on the physical conditions during ice-cover formation and evolu-

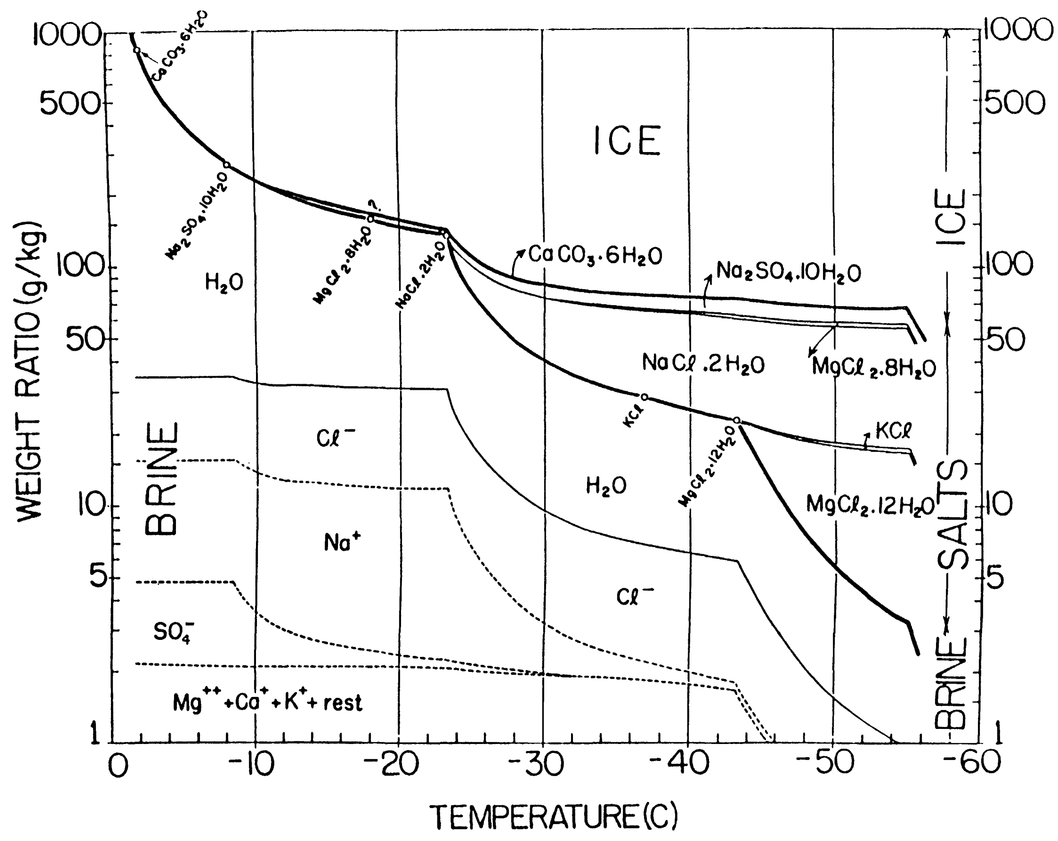

FIG. 12. Phase diagram for sea ice (after Weeks and Ackley, 1986).

tion; so there are significant relationships between the ice properties and the specific biology associated with it.

Advantages of ice habitat, relative to the under-ice marine environment, include closer proximity to the light available for photosynthesis; the presence of solid ice surfaces to attach to near the light, so that energy is not consumed by the organism to stay near the surface; and some protection from predation by larger organisms that cannot fit into the interstices in the ice.

2.1.1 Surface Communities Surface communities develop in water or slush pools created by the flooding of the top surface of the ice cover. Three processes have been identified that contribute to surface flooding. These are melt-pond formation from summer snow melt, snow-load depression of the ice cover so that it floods, and flooding induced by ice-ridge loading on the surface.

While melt ponds are a widespread feature on the Arctic pack, they are less important biologically than sea-water flooding since snow melt, being fresh water, is relatively inhospitable to marine organisms.

Depression of the ice surface, either by snow loading or by the weight of ice blocks produced during ice deformation, can lower the top surface below sea level, which is followed by flooding by sea water through cracks or pores in the ice. The sea water can contain a seed population of organisms as well as the nutrients necessary to sustain photosynthesis. These communities then bloom in the higher-light environment on the top surface of the ice, giving it a distinct color because of the high concentration of organisms (Horner, 1985). Typically, the concentration of marine algae is several hundred times higher per unit volume of flooded snow than it is in the ocean waters directly below the ice. Snow ice and the related bio-

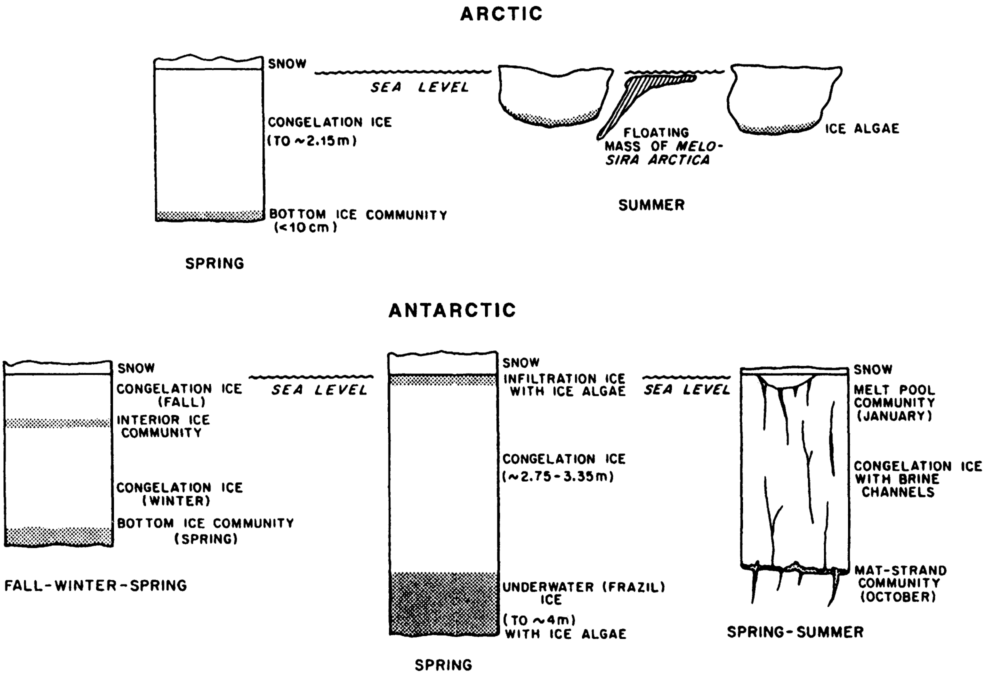
ARCTIC
SNOW
SEA LEVEL
CONGELATION ICE (TO ~2.15m)
BOTTOM ICE COMMUNITY (<10cm)
SPRING
FLOATING MASS OF MELOSIRA ARCTICA
ICE ALGAE
SUMMER
ANTARCTIC
SNOW
CONGELATION ICE (FALL)
SEA LEVEL
INTERIOR ICE COMMUNITY
CONGELATION ICE (WINTER)
BOTTOM ICE COMMUNITY (SPRING)
FALL-WINTER-SPRING
SNOW
INFILTRATION ICE WITH ICE ALGAE
SEA LEVEL
CONGELATION ICE (~2.75-3.35m)
UNDERWATER (FRAZIL) ICE (TO ~4m) WITH ICE ALGAE
SPRING
SNOW
MELT POOL COMMUNITY (JANUARY)
CONGELATION ICE WITH BRINE CHANNELS
MAT-STRAND COMMUNITY (OCTOBER)
SPRING-SUMMER

FIG. 13. Habitats identified in sea ice, divided into surface, interior, and bottom (after Horner, 1985).

logical community, resulting from the flooding of the top surface, are common over large areas of the Antarctic sea ice. The thicker ice in the Arctic requires relatively deeper snow to initiate flooding, and so the flooding-produced biological community occurs less frequently there.

2.1.2 Interior Communities Two main interior communities have been identified: frazil-ice communities that develop during ice formation, and the freeboard-layer community that develops in older ice.

The frazil-ice communities were first discovered in newly forming pancake ice in Antarctica, where concentrations of marine organisms were ten to one hundred times those found in the underlying water. It was also discovered, by examination of the nutrient levels in this newer ice, that the organisms were usually there at these higher concentrations before any significant photosynthesis occurred in the ice. Unlike the surface communities, they were physically incorporated at high levels rather than multiplying after incorporation. Mechanisms have been identified that concentrate organisms within the ice, including water pumping and scavenging by floating ice particles. Pumping, where sea water containing organisms is surged through a layer of frazil or grease ice, is thought to concentrate the organisms. As sea water is brought upward into the layer of grease ice, the organisms stick preferentially to the ice surfaces, perhaps because of some biological adhesion. Wave passage or water circulation in closed-loop cells can bring new water containing organisms to the interior of the grease-ice layer where some organisms are deposited, so that over even a few hours, significant enrichment of the ice with organisms can take place. Rafting of the ice (where two adjacent sheets slide over one another) can distribute organisms to the interior of the ice. Since frazil ice is particularly common in the Antarctic, these frazil-ice–associated interior communities are found widely there.

Scavenging of particles by freely floating frazil crystals as they move through the water column, either horizontally or vertically, also can concentrate particles. This process may account, more than wave pumping, for the high concentrations of sediments in ice seen in near-shore regions of the Arctic where wave action is not as common as in the open-ocean regions of the Antarctic.

Freeboard-layer interior communities are generally found between 5 and 10 cm below the ice surface in ice that lasts through the summer period. The condition of this layer, which is usually rich in biology, is a honeycomblike structure of partially melted ice filled with sea water and fresher ice melt. Communication of this layer with the surface sea water is shown by the presence of larger zooplankton, such as the shrimplike krill, that are often found in this layer. The layer may develop through a biophysical feedback process, in which organisms bloom and absorb solar radiation at this near-surface level. Some ice melting, caused by the trapped solar radiation, is followed by intrusion of sea water, containing the larger organisms, through the holes and gaps created by the melting. This layer has been observed in the limited area of Antarctic perennial ice at the end of the summer period. As the ice enters its second year, freezing-induced convection can also lead to nutrient enrichment of the pore water and an autumn bloom of algae inside the ice.

2.1.3 Bottom Communities Bottom communities have been identified in both the Arctic and Antarctic, although the platelet-layer habitat, below the bottom surface of the consolidated ice, appears to be restricted to the Antarctic. While the interior community shows a strong association with frazil ice, the bottom community forms in the platelike dendritic structure seen at the base of columnar ice. Since the overlying ice limits the light penetrating to the base, the bottom community develops in the fall and spring seasons when incident total light levels are so high that the small fraction that is transmitted is sufficient to sustain photosynthesis by algae that attach to the ice dendrites at the base. Even during these periods, the level of snow cover on the ice is a critical factor. It appears that the maximum productivity is attained when a thin snow cover (rather than no snow or a thick snow cover) is present, perhaps optimizing the light condition for these somewhat shade-adapted organisms. Because the bottom surface is in constant communication with the under-ice sea water, nutrient exchange is constantly taking place, unlike at higher levels, within the ice, where fluid flow is more restricted. As a result, the bottom ice communities are able to sustain production at hundreds of times the levels seen elsewhere within the ice. This results in the accumulation of biomass, in the lower 20 cm of the ice, as high as 56 g/m^2 of carbon, a figure similar to hayfield production on land surfaces, and rarely found elsewhere in oceanic waters.

A community also develops in the platelet layer seen in the under-ice coastal region of Antarctica. Even higher levels of biologic activity are seen here than in the overlying ice. Factors contributing to this production are the 3-to-1 sea water-to-ice ratio (which optimizes the space available for colonization), the stability of the layer relatively close to the light, and a high capacity for nutrient exchange over the several-meter thickness of the layer.

In both the Arctic and Antarctic, a thick snow cover impedes the development of these bottom communities, since light limitation restricts their photosynthetic capacity. Bottom communities are relatively sparse, for example, in some drifting pack-ice areas where the snow cover is apparently too thick to allow light penetration sufficient to sustain photosynthesis.

3. DISTRIBUTION OF SEA ICE

At the largest scales, sea ice exhibits features that depend on the full set of geophysical conditions for its formation. The lifetime of the ice cover, for example, can vary from days in the ice's furthest extension from the high latitudes, to a season in its most common form, to several years in the highest latitudes. The ice-thickness distribution and how it evolves best describes these processes and their effects. The categories described here are additionally classified as perennial ice zones and seasonal ice zones, the latter subdivided into fast-ice zones and marginal-ice zones.

3.1 Ice-Thickness Distribution

Plate tectonics changes the visible topographic features of the Earth's crust, cracking and roughening the original, more level, land or ocean-bottom surfaces initially formed by sedimentary or accretionary processes. Similarly, though at different scales, the sea-ice cover undergoes a form of tectonic activity and changes from a relatively more uniform slab, formed by freezing or frazil accumulation, to a rough, uneven surface characterized by ice pileups (pressure ridges) (Fig. 14) and cracks (leads) (Fig. 15). Relative ice motion and stress in the ice arise from the forces exerted on the upper and lower surfaces by wind and water drag, the motion of ocean currents, and the Coriolis force on the ice (Hibler, 1986). If the stress is high enough, the ice breaks, creating either the open-water areas that can then freeze quickly under subfreezing atmospheric conditions, or broken ice that is piled into thick ice ridges as the ice is shoved together. Coastal boundaries can also be an "immovable object" that moving ice breaks and piles against, resulting in thick ridges (Fig. 16). Or coasts may act as source regions for new ice, as fast-freezing open-water regions (polynyas) form when offshore winds, ocean currents, or tides pull the ice cover away instead of pushing it into the coast (Fig. 17).

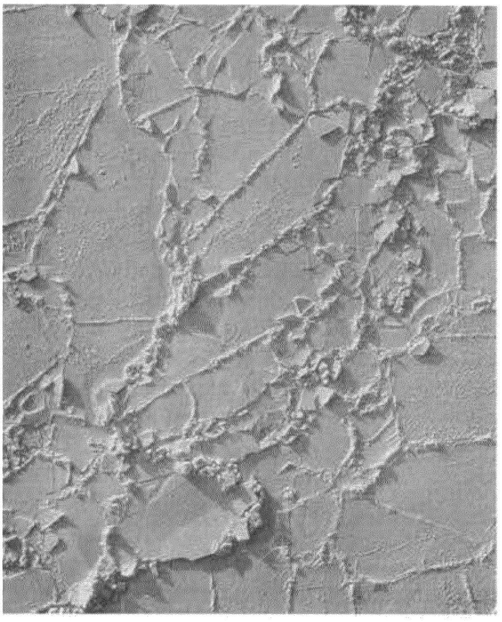

FIG. 14. Photograph of pressure ridges on sea ice. Area pictured is 100×120 m^2. (From Steffen, 1986.)

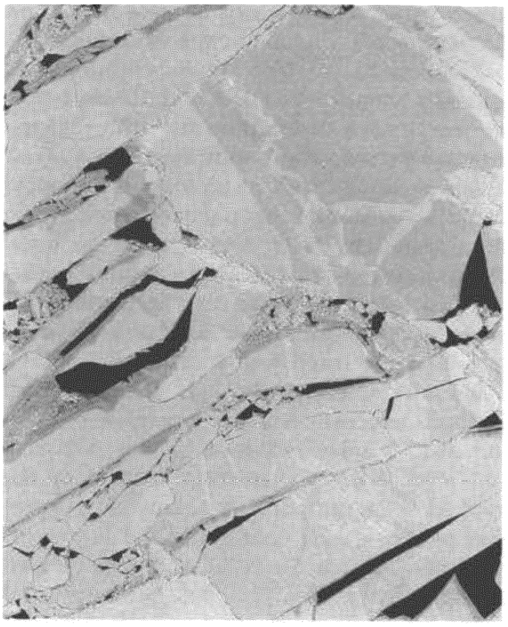

FIG. 15. Photograph of cracks, or leads, in sea ice. Area pictured is 145×180 m^2. (From Steffen, 1986.)

The ice-thickness changes or, more correctly, the ice-thickness distribution resulting from these processes is the most important property of the ice cover at scales from centimeters (vertically) to hundreds of kilometers (horizontally). One reason for this importance is that the thickness distribution tells us the mass of the ice cover (Rothrock, 1986). The thickness variations also govern, to a large degree, the interaction of the ocean and atmosphere. On ice-covered oceans, the rate of energy transfer (heat loss or gain) to the atmosphere and the ice strength (its resistance to motion and deformation) both depend directly on ice thickness. Ice-growth rate, which governs the salt flux to the ocean, depends itself directly on ice thickness.

Regions of modest size can contain ice of many different thicknesses. In a statistical sense we can represent this variation by the ice-thickness distribution $g(h)$ (Rothrock, 1986; Thorndike, 1986). This function can be used to represent the fraction of a region R with thickness between h and $h + dh$ as

$$g(h)dh = \frac{dA(h, h + dh)}{R}, \tag{1}$$

where $dA(h, h + dh)$ is the area within R

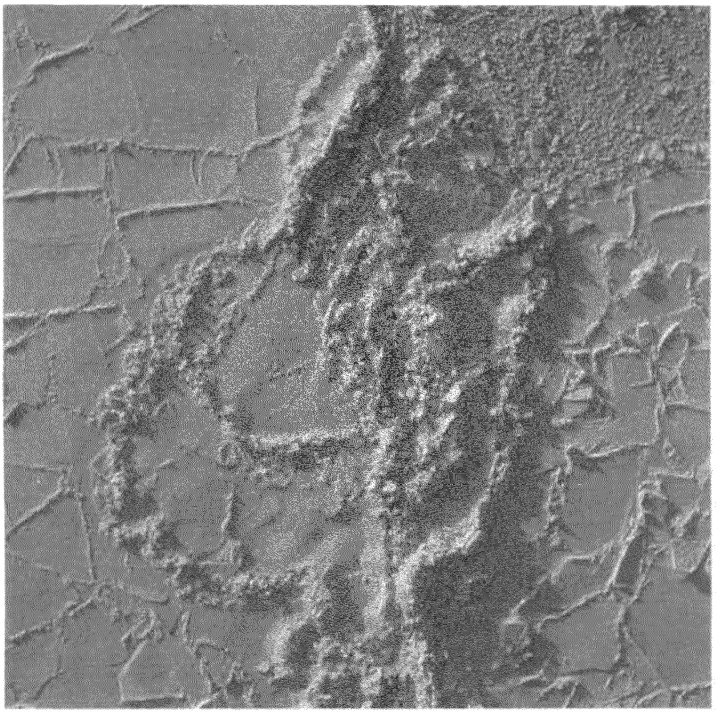

FIG. 16. Photograph of sea ice piling in compression, as in against a coastal boundary. Area pictured is 180×220 m^2. (From Steffen, 1986.)

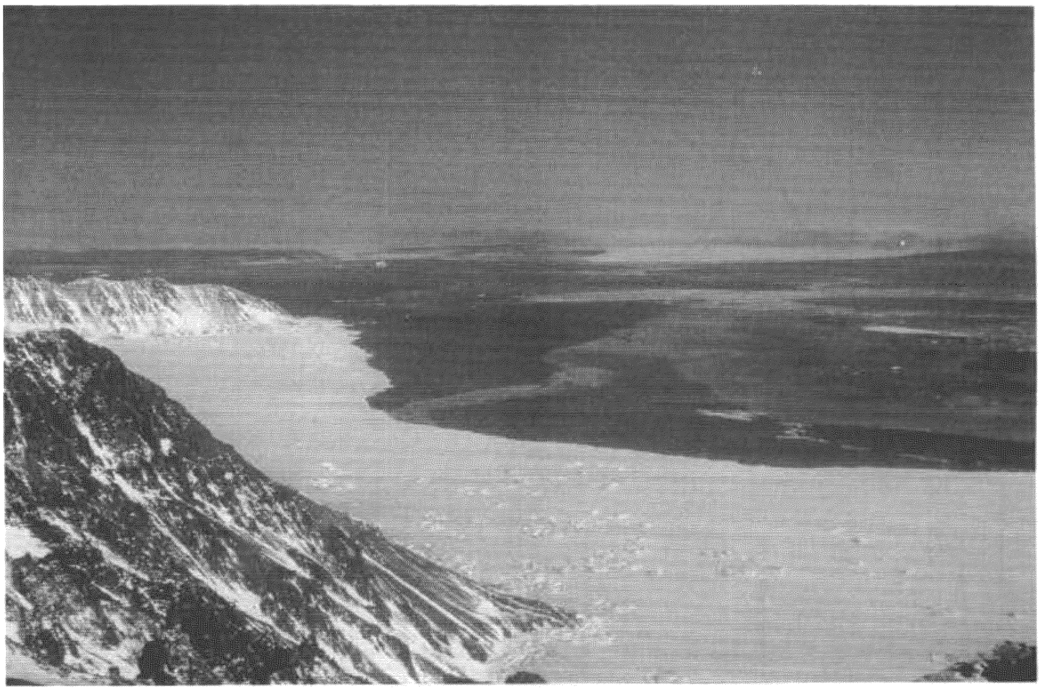

FIG. 17. Photograph of a coastal polynya. Bottom width of picture is 200 m. (From Steffen, 1986.)

covered by ice in the thickness range (h, $h + dh$). $g(h)$ itself has the properties of a probability density function with dimensions of length^{-1}.

Note also that when a finite fraction of the region R is ice free, g will have a delta function at $h = 0$, and so g includes information about ice concentration [defined as the areal fraction of ice cover, or $1 - dA(0,dh)/R$] as well as thickness. The mean ice thickness in R is

$$h_m = \int_0^\infty hg(h)dh \qquad (2)$$

and includes areas of open water.

Taking into account spatial and temporal variations, we can write the ice-thickness distribution more generally as $g(h,\mathbf{x},t)$.

The general thickness distribution is governed by the equation (Thorndike, 1986; Rothrock, 1986) that relates the time and space evolution of the statistical function g to physical forcing terms as

$$\frac{\partial g}{\partial t} = \mathrm{div}(\mathbf{v}g) - \frac{\partial(fg)}{\partial h} + \Psi, \qquad (3)$$

where $f(h,\mathbf{x},t) = \partial h/\partial t$ is the thermodynamic growth rate of ice of thickness h, $\mathbf{v}$ is the horizontal velocity vector, div($\mathbf{v}g$) is the flux divergence, and Ψ is a function that accounts for the mechanical redistribution of ice from one thickness to another, modeling the formation of both leads and pressure ridges. Figure 18 shows the behavior of each of these terms (arrows represent delta functions at zero thickness) and the resulting ice thickness distribution. The shapes of the distributions reflect the importance of the various processes. For example, a high thickness tail on the distribution is only possible through the mechanical redistribution that converts thin ice into thick ice (pressure ridging). Purely thermodynamic processes, shown by f, tend to thicken thin ice quickly but only up to an equilibrium thickness, where the growth rate f becomes zero. For thick ice, the growth rate is negative, indicating that melting occurs below very thick ice. This thermodynamic effect is caused by the poor heat conduction through thick ice and the presence of a small, but significant, heat flux from the ocean in most ice-covered regions.

Limited thickness-distribution information

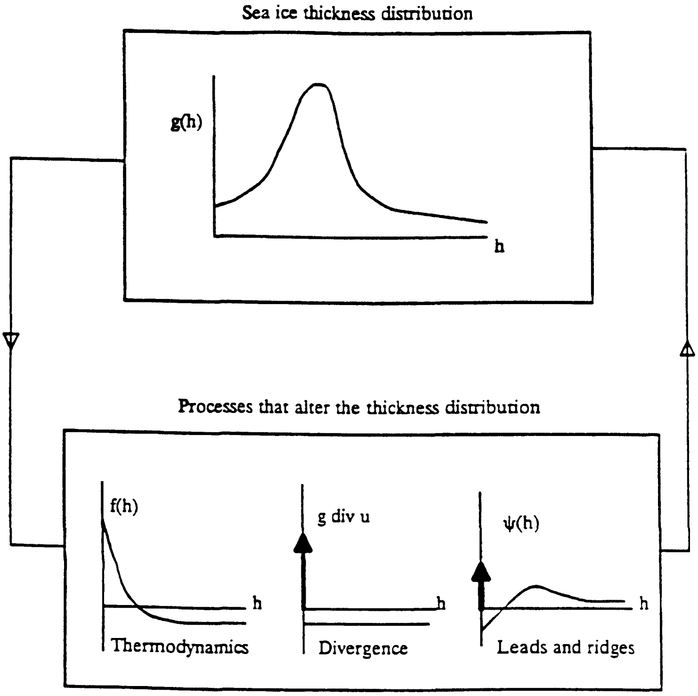
Sea ice thickness distribution
g(h)
h
Processes that alter the thickness distribution
f(h)
h
Thermodynamics
g div u
h
Divergence
ψ(h)
h
Leads and ridges

FIG. 18. Ice-thickness distribution and its dependence on growth [$f(h)$], deformation (Ψ), and divergence [div(vg)] (after Thorndike *et al.*, 1992).

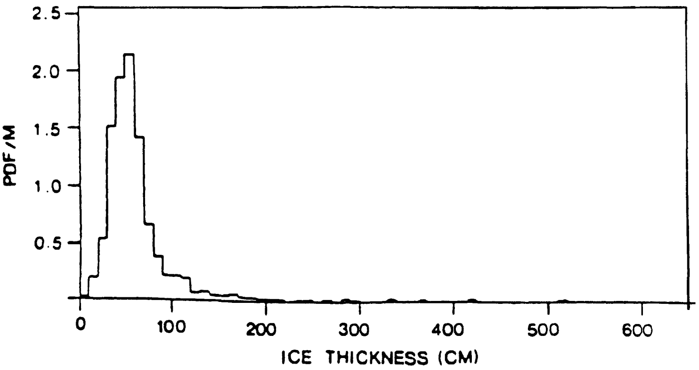

FIG. 19. Antarctic sea-ice–thickness distribution (after Gow and Tucker, 1991).

is available from the polar regions. One from the Antarctic sea ice is shown in Fig. 19, and one for the Arctic is shown in Fig. 20. Even the few distributions available, however, reveal the dominant ice processes and main characteristics of the ice cover on a large scale. The Antarctic ice thickness has a single peak (one mode) and a mean thickness of about 0.6 m. This distribution reflects the generally warm conditions under which the Antarctic pack ice grows, and its formation by primarily thermodynamic processes, shown by the absence of a deformation tail in the distribution. Low divergence, by the absence of a peak at the thicknesses less than the mean, is also indicated. In contrast, a high tail in the Arctic thickness distribution indicates a more robust redistribution of thin ice into thick ice, and significant ridging is in fact present. Similarly, the bimodal (two peaks) distribution indicates that divergence effects have occurred more recently resulting in the thin-ice peak at 0.8 m, a measure of ongoing dynamic activity. The high mean ice thickness (3 m) indicates a generally cold climate, rapid thermodynamic thickening of thin ice, and little ice-thickness loss in the summer.

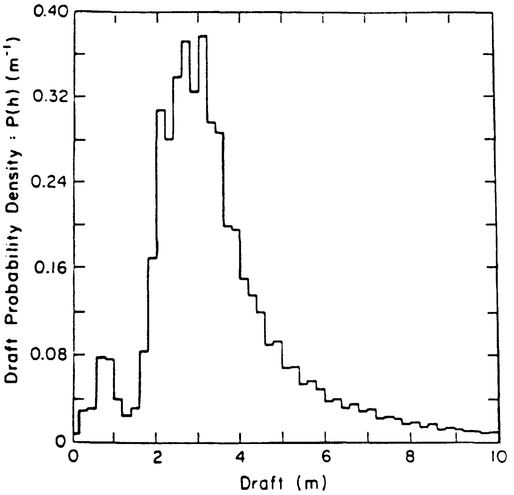

FIG. 20. Arctic sea-ice–thickness distribution. (After Gow and Tucker, 1991.)

3.2 Perennial Ice

Ice covers that last through at least one summer season and into the next growing period are described as perennial ice. Figure 1 shows the distribution of perennial ice in the Arctic and Antarctic, given as the regions inside the seasonal minimum ice edge.

The amounts and distributions of perennial ice in the two regions are highly contrasted; two-thirds of the ice cover in the Arctic is perennial compared with only one-fifth in the Antarctic. The two major influences that lead to this difference are the higher latitudes of the ice-covered ocean in the Arctic relative to the ice-covered ocean in the Antarctic, and the complete open-ocean sea-ice boundary in the Antarctic compared with the Mediterranean character (ocean surrounded by land masses) of the Arctic basin.

The higher latitudes of the Arctic Basin give a longer winter season, and thicker ice develops that only partially ablates away during the summer season. In the Antarctic, the ice is thinner and the open-ocean boundary provides a broad melting zone that extends around the full circumference of the sea-ice zone. Ice is advected into the open ocean from higher latitudes and warmer water destroys it quickly. In contrast, the advected ice from the Arctic can pass only through a relatively narrow passage to be destroyed in the warmer waters of the North Atlantic. This passage at Fram Strait amounts to 10% of the Arctic Basin's circumference. The life of Arctic Basin ice is of the order of 3 to 7 years before it is advected out of the Basin and destroyed, while Antarctic perennial ice is only two years old because the motion through the perennial zone (e.g., the Western Weddell Sea) is relatively quick. Through exposure to several summers with more intense radiation, Arctic perennial ice is relatively low in salinity compared with the Antarctic, where only one summer exposure is common. Flushing by surface snow-melt water also helps desalinate the upper layers of Arctic ice, while the sea-water–flooded snow cover remains relatively intact on the Antarctic ice through the single summer period, ensuring that little salt is lost from the Antarctic ice in its second year.

3.3 Seasonal Sea Ice

3.3.1 Fast-Ice Zones Fast ice is the sea ice that grows seaward from a coast and stays in place (attached, or "fast," to the coast) during winter. In spring and summer it can either melt in place or break away from the coast and then melt (Wadhams, 1986). A small fraction lasts through the summer as perennial fast ice. Fast ice is found in both the Arctic and Antarctic. In some areas, primarily Arctic coasts, the fast ice is stabilized by grounded ice-deformation features (pressure ridges), and can extend out to the 20–30-m depth contour that is the draft limit of such ridges. In constricted channels, such as in the Canadian Archipelago or McMurdo Sound, Antarctica, the fast ice fills the whole width, regardless of water depth, but can be subjected to some movement and deformation if strong winds occur before the ice is stabilized. Fast ice in Antarctica is only 1% of the total area of ice-covered ocean, but is important in the local ecology, providing habitat for some varieties of seals and penguins. It also modifies the oceanic light regime in the coastal region, important in the overall biological productivity, and the ice itself also supports an abundant biological community.

3.3.2 Marginal-Ice Zone The marginal-ice zone is the part of the ice cover that is affected by the open-ocean boundary. Several cold drift currents that contain ice can be considered part of the marginal-ice zone (East Greenland and Labrador) as is ice lying within 150–200 km of the northern Antarctic ice edge. The majority of seasonal ice that drifts is in the marginal-ice zone in its formation period, in its decay period, or both. Generally, ice in the marginal zone has much greater freedom of movement than ice in the central pack and consequently a number of ice–ocean–atmosphere interaction phenomena occur that are peculiar to this zone.

A principal effect on ice processes in this zone is that of ocean waves. As described before, the interaction of the wave field with a growing ice cover accounts for the high proportion of frazil or granular ice seen in the Antarctic. Wave fields also have important influences on the decay of seasonal ice. As large-diameter (kilometer-scale) floes are advected into the zone where waves can penetrate, they are broken by flexure into pieces only a few to 10 m in diameter, greatly increasing the perimeter length that is in contact with warm water and accelerating the ice decay. The lower ice concentration and smaller floes seen here also respond differently to wind and ocean-current forcing than the interior pack, influencing the decay of the ice cover as well.

Some marginal-ice zones, such as the East Greenland Current, coincide with frontal zones in the ocean. Satellite data have shown that complex ocean-eddy structures are revealed by the pattern of the ice floes on the surface. Modeling studies and measurements have shown that the ice is an active participant in ocean dynamics, influencing the kinematics of the eddies and how they form, move, and decay as well as providing a tracer of the eddy motions.

3.4 Geophysical Effects on Ice Distribution

The presence of an ice cover in a region is a product of several forces or factors. Some of these are the surface-energy balance (temperature and radiation) (Maykut, 1986); the ocean thermal structure; and the wind and ocean currents that drive the motion and deformation of the ice cover (Hibler, 1986). Interactions that add levels of complexity to ice conditions include the wave-field effects that cause pancake-ice formation and the sub–ice-shelf ocean dynamics that lead to platelet-ice formation. Wind-driven effects and the deformation of the ice cover can, for example, change the ice-thickness distribution so that the thickest ice occurs in lower latitudes than expected from the location of the coldest conditions (Hibler, 1986). And open water is often found continuously at the far southern boundary of the pack ice nearest the Antarctic continent because of the strong winds that blow ice away from the continent.

3.4.1 Ice Circulation and Deformation in the Arctic and Antarctic Two dominant features of the ice circulation in the Arctic Basin are the Beaufort Sea Gyre, a clockwise circulation pattern in the Beaufort Sea, and the transpolar drift stream, a basin-wide feature that transports ice from the western to the eastern side of the basin (Fig. 21). The ice circulation mimics the wind field that drives it, caused by the mean sea-level pressure field. In the mean, a large high-pressure area exists over the Beaufort Sea and isobars over the remainder of the basin trend west to east. With large daily and seasonal variability, mean velocities of the ice are about 2 km day^{-1} (±7 km day^{-1}). As a result of this circulation pattern, ice floes can remain in the basin, especially the Beaufort Sector, for many years. The ice dynamics also lead to a buildup of ice along the coasts of Greenland and the Canadian Archipelago. Ice with average thickness in excess of 6 m is found along these coasts in contrast to the 3- to 4-m ice found over the North Pole. Figure 22 shows ice-thickness contours obtained from submarine up-looking sonar data. Model results, where no ice dynamics is allowed, have shown that the thickest ice would be expected over the

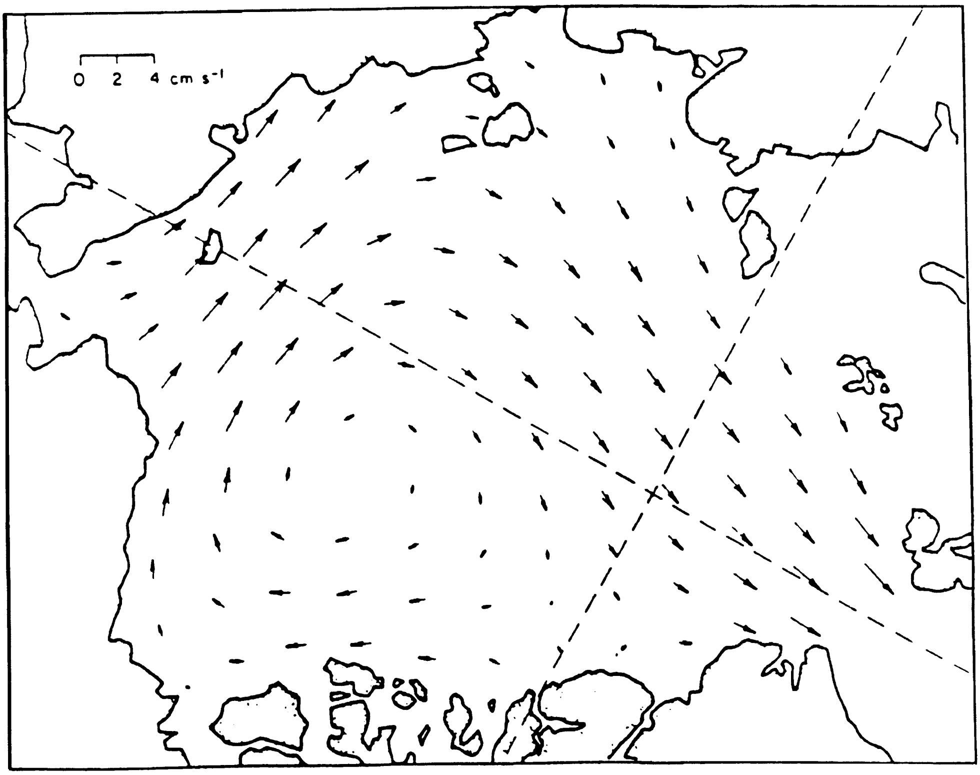

FIG. 21. Mean field of ice motion in the Arctic Basin (after Gow and Tucker, 1991). Greenland is at the bottom right.

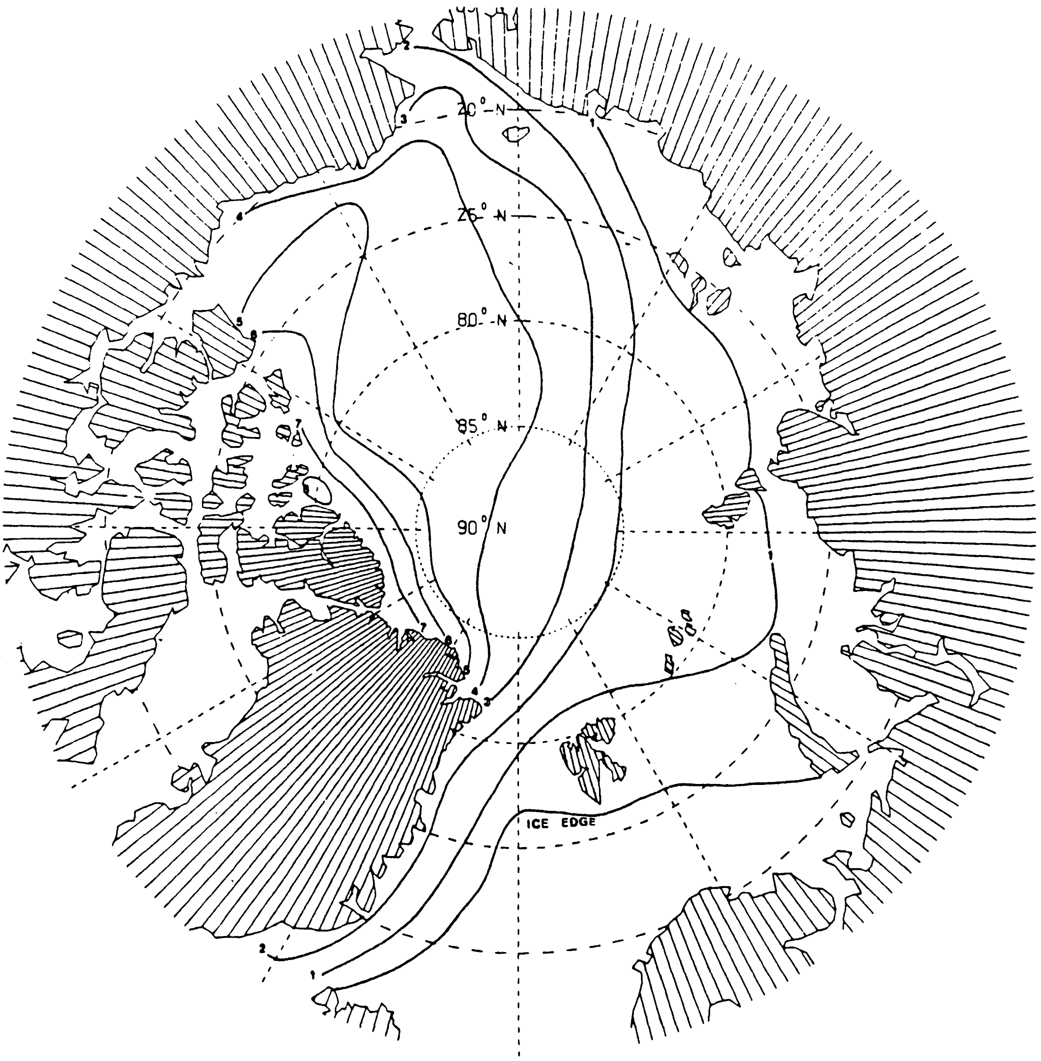

FIG. 22. Contours of Arctic mean ice draft obtained from submarine sonar (after Gow and Tucker, 1991).

North Pole where the coldest conditions prevail. As the data show, however, the thickest ice actually forms elsewhere, as a result of ice dynamics (Hibler, 1986).

In the Antarctic the major drift features are the wind-driven circulation at the north near the boundary with the Antarctic Circumpolar Current, and the East Wind Drift flowing in the opposite direction in the south immediately adjacent to the continent (Fig. 23) (Gow and Tucker, 1991). This coastal flow is directed north by protrusions of the continent in the Weddell and Ross Seas. The ice in the northward flow joins the Circumpolar Current, which transports it into warmer water where it eventually melts.

Because of the northward flow in the Weddell Sea, the ice at the northern end lasts longer than ice elsewhere, thickened by deformation during its journey. It is thicker than ice further south, despite generally colder conditions in the south. In this case, the ice circulation and deformation invert the thickness relationships expected from thermodynamics alone. Mean drift rates (4–6 km day^{-1}) are two to three times those of Arctic sea ice, and ice-drift rates up to 20 km day^{-1} have also been observed, reflecting the

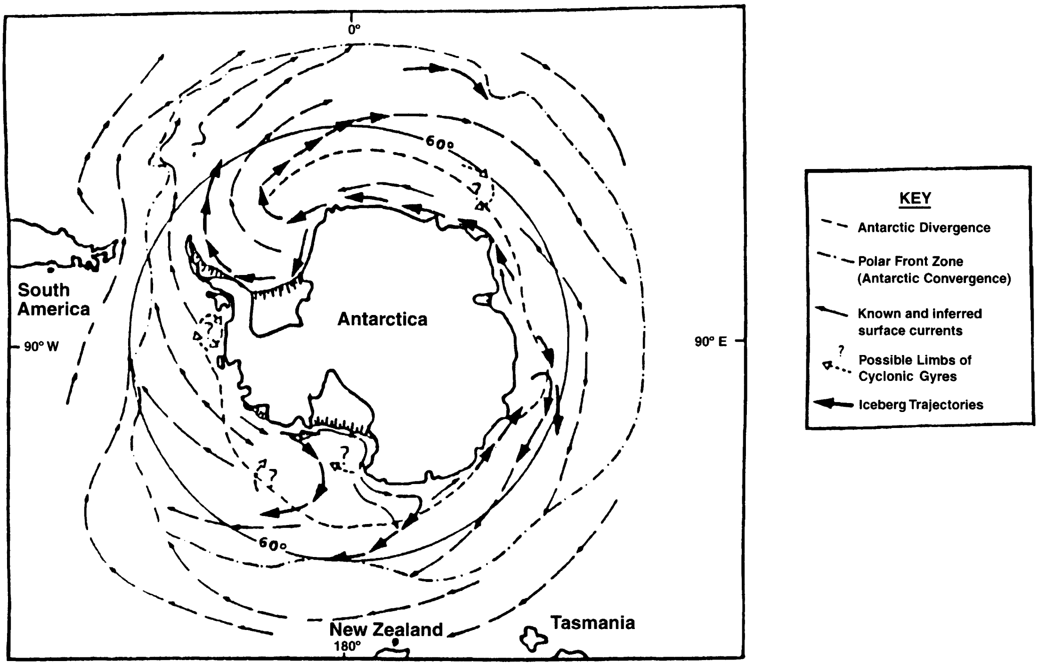
0°
60°
South America
90° W
Antarctica
90° E
60°
New Zealand
180°
Tasmania
KEY
Antarctic Divergence
Polar Front Zone (Antarctic Convergence)
Known and inferred surface currents
Possible Limbs of Cyclonic Gyres
Iceberg Trajectories

FIG. 23. Surface circulation (of ice and water) around Antarctica (after Gow and Tucker, 1991).

more energetic atmospheric and ice circulations in the Southern Hemisphere.

4. CONCLUSION—SEA ICE, COMPLEX RESPONSE IN A CHANGING CLIMATE

While cold conditions are necessary for ice-cover formation, many other factors in the atmosphere and ocean modify the distribution, thickness, and properties of the ice cover. Some of these include the snow cover, the ocean thermal state, and the wind and ocean currents. Because of the ice cover's capability to influence also the ocean and atmosphere, through feedback processes, the response of the ice cover to climate-change effects would be difficult to predict, even assuming the climate-change effects are themselves predictable.

In the outer areas of the pack ice—in the seas surrounding the Arctic Basin or at the northern limit of the Antarctic pack ice—atmospheric and oceanic thermal effects may determine the presence or absence of the ice covers. Even there, however, certain regions are affected by the advection of ice formed in more polar latitudes. Ice-cover changes may in that case be more sensitive to changes in the wind and ocean circulation fields than they would be to atmospheric temperature changes. It appears, for example, that the paleoclimatic extension of the sea-ice cover into the North Atlantic was in response to a decrease in salinity of the upper ocean there, affecting the deep convection that brings heat to the surface. The expanded sea-ice field then cooled the atmosphere near it and then influenced how weather systems moved across the North Atlantic. Lower temperatures in northern Europe may then have resulted from the sea-ice effects, rather than causing them. Studies on Antarctic sea-ice conditions have also suggested that local temperature changes at the Antarctic coast are related to sea-ice extent, while the ice extent itself is more controlled by the wind fields prevailing at the time than by the temperature.

The sea-ice fields also interact strongly with the biological environment. Changes in the ice cover can drive populations to different regions as the food supply associated with, for example, melting ice edges, or sea-ice biology, responds to external changes.

Deep-ocean areas are generally low in productivity except in special zones, like the ice edges, where melt, wave damping, and other processes occur that enhance biological production. A reduction in sea-ice area or a decrease in the circumference of marginal ice would then potentially reduce the ice edge available for biological production.

The sea-ice system combines the vagaries of atmospheric and oceanic circulation with nuances unique to its state as a more solid rather than strictly fluid medium. Predictions of those circulations are somewhat tractable in their full dimension only by numerical modeling with the full computational power of the world's fastest and largest computers. Further computational development, as well as a better understanding of sea-ice processes, is necessary before the combined atmosphere–ice–ocean–biosphere interaction can be fully understood.

GLOSSARY

Brine: Sea water enriched in salts by ice freezing.

Columnar Ice: Ice grown by one-dimensional cooling by thermal conduction, which produces ice crystals with the appearance of long columns parallel to the growth direction.

Fast Ice: Ice, formed in a water body, that is attached to a shoreline.

Frazil Ice: Ice formed as small disks, plates, and dendritic arms in suspension in a water body, usually under turbulent flow conditions.

Ice Stalactite: An elongated tube of ice formed at and attached to the base of an ice sheet, similar in appearance to a cave stalactite.

Leads: Cracks in a sea-ice cover, of the order of 5 to 100 m in width, and sometimes exceeding many kilometers in length.

Marginal-Ice Zone: The region of the ice-covered ocean that is close to the open ocean where no sea-ice cover is present. Ice floes in the marginal-ice zone are usually broken up by the wave action.

Perennial Ice: Ice that has survived one or more summer seasons.

Platelet Ice: A type of ice crystal formed in and associated with the outflow of cold supercooled water from beneath large continent-based ice shelves.

Polynyas: Large irregularly shaped bodies of open water, either surrounded by sea ice or bounded on one side by coast and on the other by sea ice ("coastal polynyas"). (Originally Russian.)

Pressure Ridges: Ice pileups formed by ice-floe collisions; can be many meters thicker than the original ice sheet.

Seasonal Ice: Ice that forms during the winter and melts by the end of summer.

Snow Ice: Ice formed by the freezing of water-saturated snow.

Works Cited

Barry, R. G. (1986), in: N. Untersteiner (Ed.), *The Geophysics of Sea Ice,* NATO Advanced Study Institutes Series B, Physics, Vol. 146, New York: Plenum, Chap. 15.

Gow, A. J., W. B. Tucker III (1991), *Physical and Dynamic Properties of Sea Ice in the Polar Oceans,* CRREL Monograph 91-1, Hanover, NH: Cold Regions Research and Engineering Laboratory.

Hibler, W. D. III, (1986), in: N. Untersteiner, (Ed.), *The Geophysics of Sea Ice,* NATO Advanced Study Institutes Series B, Physics, Vol. 146, New York: Plenum, Chap. 9.

Hobbs, P. (1974), *Ice Physics,* Oxford: Clarendon Press, p. 13.

Horner, R. (1985), *Sea Ice Biota,* Boca Raton, FL: CRC Press.

Horner, R., S. F. Ackley, G. S. Dieckmann, B. Gulliksen, T. Hoshiai, I. A. Melnikov, W. S. Reeburgh, M. Spindler, C. W. Sullivan (1992), "Ecology of Sea Ice Biota. 1. Habitat and Terminology," *Polar Biol.* **12,** 417–427.

Maykut, G. A. (1986), in: N. Untersteiner (Ed.), *The Geophysics of Sea Ice,* NATO Advanced Study Institutes, Series B, Physics, Vol. 146, New York: Plenum, Chap. 5.

Rothrock, D. A. (1986), in: N. Untersteiner (Ed.), *The Geophysics of Sea Ice,* NATO Advanced Study Institutes Series B, Physics, Vol. 146, New York: Plenum, Chap. 8.

Steffen, K. (1986), *Atlas of the Sea Ice Types, Deformation Processes and Openings in the Ice,* Eidgenössische Technische Hochschule, Zurich, Switzerland, Report No. 20.

Thorndike, A. S. (1986), in: N. Untersteiner (Ed.), *The Geophysics of Sea Ice,* NATO Advanced Study Institutes Series B, Physics, Vol. 146, New York: Plenum, Chap. 7.

Thorndike, A. S., Parkinon, C., Rothrock, D. A. (Eds.) (1992), Report of the Sea Ice Thickness Workshop, Polar Science Center, Univ. of Washington, Seattle, WA.

Wadhams, P. (1986), in: N. Untersteiner (Ed.), *The Geophysics of Sea Ice,* NATO Advanced Study Institutes, Series B, Physics, Vol. 146, New York: Plenum, Chap. 14.

Weeks, W. F., S. F. Ackley (1986), in: N. Untersteiner, (Ed.), *The Geophysics of Sea Ice,* NATO Advanced Study Institutes Series B, Physics, Vol. 146, New York: Plenum, Chap. 1.

Further Reading

Horner, R. (Ed.) (1985), *Sea Ice Biota,* Boca Raton, FL: CRC Press.

Untersteiner, N. (Ed.) (1986), *The Geophysics of Sea Ice,* NATO Advanced Study Institutes, Series B, Physics, Vol. 146, New York: Plenum.

SEDIMENTARY BASIN EVOLUTION

CHRISTINE E. KOLTERMANN, *Pegasus Geoscience, Santa Clara, California, U.S.A.*

	Introduction	105
1.	**The Source Area**	107
2.	**Classification of Sedimentary Basins**	109
3.	**Basin Formation**	109
3.1	Tectonics	110
3.2	Loading of the Lithosphere	113
3.3	Compaction	114
4.	**Basin Filling**	114
4.1	Depositional Environments and Facies Models	115
4.1.1	Nonmarine Sediments	115
4.1.2	Coastal and Shallow-Sea Sediments	118
4.1.3	Marine Sediments	119
4.1.4	Volcanic Sediments	122
4.2	Basin-Filling Processes	122
4.3	Modeling of Basin Filling	124
4.3.1	Process-Based Models	124
4.3.2	Geometric Models	129
5.	**Diagenesis**	131
6.	**Geologic History Analysis**	132
7.	**Conclusions**	134
	Glossary	134
	Works Cited	135
	Further Reading	137

INTRODUCTION

A sedimentary basin is a large topographic depression in the Earth's crust that accumulates a substantial thickness of sediment plus pore fluids over geologic time. The sinking or vertical downward displacement of the Earth's surface that forms sedimentary basins is called *subsidence.* Subsidence initially is caused by tectonic mechanisms that create vertical and horizontal motions in the Earth's crust and later is augmented by the weight of accumulating sediments, water, and ice, and sediment compaction. Principal causative mechanisms for tectonic subsidence include (Ingersoll, 1988; Busby and Ingersoll, 1995)

1. *extension* (stretching or pulling apart) of the Earth's crust and *rifting* (large-scale faulting and tilting);
2. thermal cooling and contraction of heated crust; or
3. *tectonic loading,* which is depression of the lithosphere adjacent to mountains formed by crustal compression.

As a result of the unique tectonic setting of each basin, basins vary widely in size and shape. However, basins are typically at least tens of kilometers from side to side and often have a circular, semicircular, or elongated trough-shaped geometry in plan view (Einsele, 1992; Miall, 1984).

Important aspects of sedimentary basins in a continental setting are illustrated in Fig. 1. The inset shows basin formation mechanisms, while the larger figure shows a mountainous source area that provides fluid and sediment to an adjacent sedimentary basin. Processes operating in the source area include erosion and the chemical and mechanical *weathering,* or breakdown, of rock into smaller particles. Climate variation in the source area affects the intensity of sediment weathering as well as the magnitude and frequency of storms. Wind, water, and mass movements transport sediment from highland source areas into basins. Additional sediment is produced by biologic activity (the growth of organisms) or chemically precipitated from sea water or lakes. Depositional processes, chemical and biological activity, environmental conditions, and the rise and fall of sea level all affect the spatial patterns of sediment grain size, mineralogical composition, porosity, and permeability that develop within the sedimentary basin fill.

Figure 1 also illustrates the sedimentary rock cycle, a continuous cycle in which the Earth's crust is *uplifted,* or raised in eleva-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

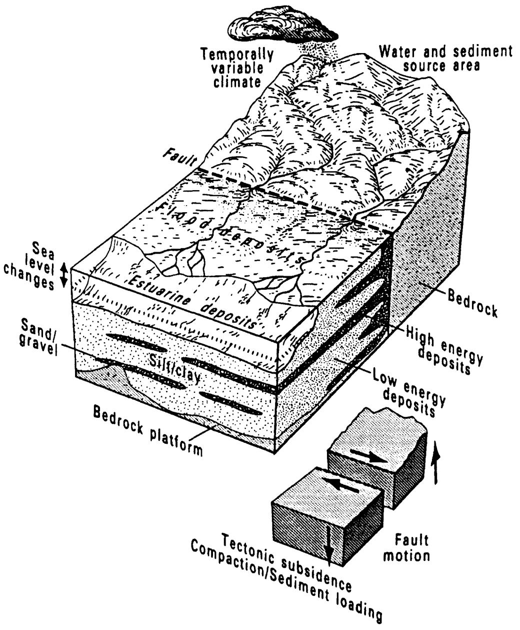

FIG. 1. Example of sedimentary basin-formation and basin-filling processes. Three sets of processes must be considered in the formation and filling of sedimentary basins. Processes in the source area (shown as the mountains) cause weathering and erosion of rock to provide sediment to the receiving basin, into which sediments are transported and accumulate (on the coastal plain and below sea level). Basin-formation processes (shown in the inset) such as tectonics (here shown as motion along faults), sediment loading, and compaction cause subsidence to form a topographic depression. After Koltermann and Gorelick (1992). Reprinted by permission of the American Association for the Advancement of Science.

tion, to form source-area mountains; rocks are weathered and eroded; and sediment is transported and deposited in basins. Sedimentation is not continuous in time. In any given basin, there may be gaps in the sedimentary record of long or short duration, in which either no deposition took place or erosion removed previously deposited sediments. As basin-fill sediments are continually buried by younger deposits, compaction and increasing temperatures as well as precipitation of dissolved minerals from pore fluids cause the sediment grains to cement together and become rock. Physical, chemical, and biological changes to sediments after deposition, collectively called *diagenesis,* include precipitation of cements from circulating subsurface fluids in fractures and pore spaces, and compaction of sediment grains. Diagenetic changes cause sediment particles to *lithify,* that is, to become rock. Folding and faulting of basin-fill sediments can occur during sedimentation as well as after sedimentation ceases. The rock cycle eventually comes full circle, as sedimentary rocks are uplifted into mountains.

Sedimentary rocks contain most of the world's metals, minerals, fossil fuels, and potable underground water supplies. To exploit and manage the Earth's natural geologic resources (groundwater, hydrocarbons, coal, minerals, etc.), we must know the distribution and properties of oil reservoirs, aquifers, and rocks of economic value in the subsurface. Understanding the controls on sedimentary basin formation and filling is critical for successful exploration and production of the Earth's mineral and energy resources as well as prediction of subsurface fluid flow and contaminant migration. Sedimentary rocks also tell us much about the Earth's geologic history, from the progression of life through time to the forces that cause earthquakes.

This article discusses how sedimentary basins evolve; how they are classified, form, and fill with sediment; the typical environments in which sediment is deposited; and common patterns in basin fill. The article is organized around the three components affecting a sedimentary basin and the rock cycle, as illustrated in Fig. 1: the source area, basin-formation mechanisms, and basin-filling processes. First, a brief source-area section discusses the factors affecting fluid and sediment inputs to sedimentary basins. Then, the ways in which sedimentary basins have been classified on the basis of basin-formation tectonics are described. The basin-formation section focuses on mechanisms of basin creation. In the basin-filling section, common depositional environments are first described. Then, the section addresses the transport and accumulation of *clastic* sediment, defined as particles eroded from a previously existing geologic formation, as well as sediment of biological or chemical origins. Chemical and physical processes affecting sediments after deposition, called diagenesis, are then described. A discussion of geologic history analysis—which uses local and regional geologic and climatic data to constrain the structural, sea-level, climatic, and hydrologic histories of a sedimentary basin—is presented. The ways in which sedimentary basin formation and filling and source areas have been modeled mathematically are discussed within each section. Finally, some conclusions are presented. (See also PLATE TECTONICS; STRUCTURAL AND PHYSICAL GEOLOGY; EARTH, INTERIOR STRUCTURE OF THE; VISUALIZATION, SIMULATION, AND MODELING BY COMPUTER.)

1. THE SOURCE AREA

A *source area* is a feature of positive relief that constitutes the drainage area for fluid and sediment transported into a basin. As the source area is uplifted, it is simultaneously eroded, although uplift rates are almost never exactly balanced by erosion. Physical, chemical-mineralogical, and biological weathering of source-area rocks at and just beneath the land surface breaks rocks down into smaller particles, alters minerals into more stable minerals such as clays, forms soils, and dissolves some minerals into water. Source-area *denudation* is defined as the combination of natural processes that result in the progressive lowering of the Earth's surface by weathering, erosion, mass wasting, and transport by wind and water. The rate at which fluid and sediment is transported from a source area into a sedimentary basin is a function of the source-area size, climate, hydrology, and vegetation, and the rock composition. In continental settings, runoff of surface water into streams transports most of the particulate matter and dissolved minerals, but wind and ice can also transport sand and debris, and groundwater transports some dissolved minerals out of source areas and into sedimentary basins. The rocks of the source area are known as the *parent rocks,* or *provenance,* of the basin sediments. The grain sizes and minerals present in the parent rocks are directly reflected in the basin sediments. An extensive discussion of source-area denudation and techniques for estimation of sediment production from source areas is provided by Slingerland *et al.* (1994).

Source-area mountains form when the Earth's crust is deformed as driven by the motion of the giant plates that cover the surface of the Earth, and where volcanic activity creates individual mountains or mountain ranges. Situations in which mountains are built include the coming together or convergence of two continents; the convergence of an oceanic tectonic plate with a continental plate that results in *subduction,* or a diving underneath, of oceanic crust beneath continental crust; the faulting and tilting of large blocks of rock due to crustal extension (stretching); volcanic activity; and the faulting and folding of continental crust in the interior of a continental plate due to compression.

Physical and chemical processes operating in source areas can be explicitly represented to predict the development of stream patterns (called *drainage networks*), the magnitude and frequency of stream flows, patterns of weathering and erosion, and the amount of sediment transported out of the source area. The source area can be mathematically defined through representing the physics of mountain building, called uplift models, as well as through the climatic and hydrologic processes that create drainage networks, cause precipitation in source areas, and transport sediment. These modeling approaches are discussed in Slingerland *et al.* (1994), Koltermann (1993), and Turcotte and Schubert (1982). Models of uplift are mathematical models of tectonic forces that predict areas of high elevation relative to surrounding areas. An example of a model of uplift due to convergence has been presented by Beaumont *et al.* (1991), who explored the role of wind direction and precipitation patterns on erosion of mountain ranges during their uplift. They modeled the physics of two continental plates converging in which rocks were folded and faulted to form a mountain range, which then exerts a tectonic load on the Earth's crust and is subject to fluvial erosion by precipitation that collects on the land surface as streams. They considered the dynamic interaction between tectonic compression, wind (storm) direction, orographic effects, stream drainage-network patterns, erosion, and isostatic compensation (Fig. 2). Modeling the deformation of the continental plates requires assumptions about the relations between compressional forces (the applied stress) and the rate of deformation (the *strain rate*). It was assumed that the lithosphere is a rigid-plastic material that does not deform until a critical threshold stress is

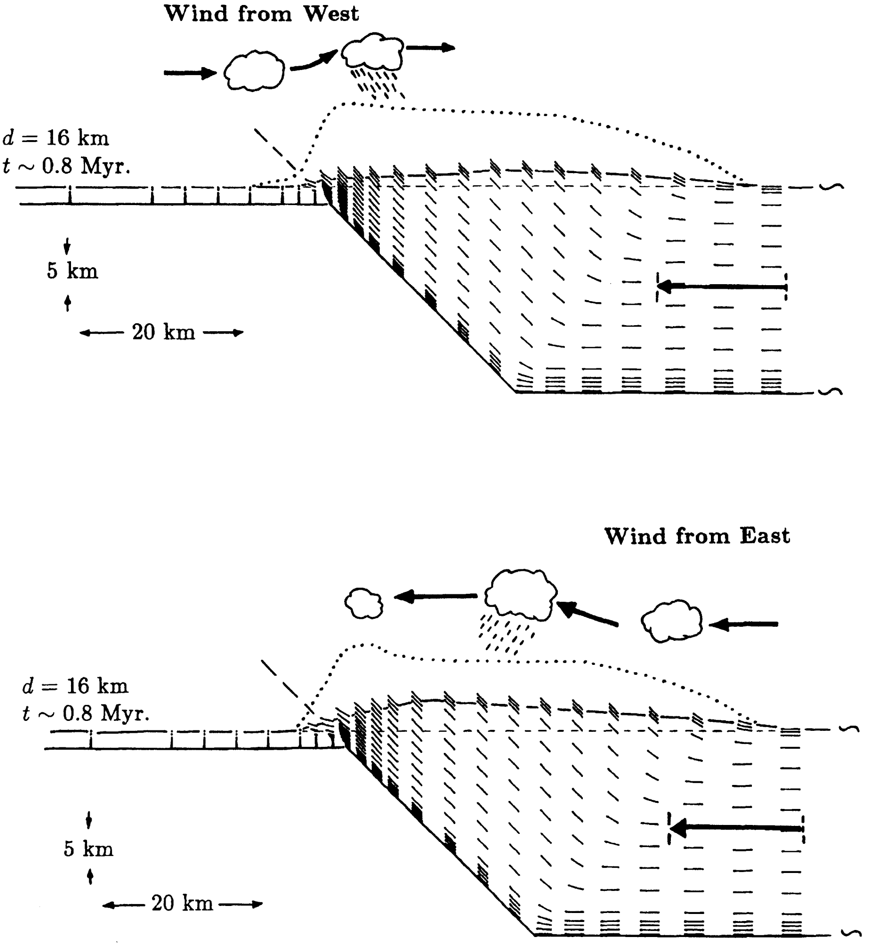

FIG. 2. Model of uplift caused by compressive forces. Results from two simulation experiments with different wind directions (west or east) that affect the spatial distribution of precipitation and therefore the location of fluvial erosion. Both simulations were conducted for 0.8 million years (t = 0.8 Myr) and with the same amount of crustal shortening from tectonic compression (d = 16 kms), shown by the bold horizontal arrows. Small dashed lines are velocity vectors. The dotted lines beneath the clouds show the projected position of the land surface if denudation had not taken place. Model results predict stream-channel network patterns, mountain-range topography, and erosional unloading of the Earth's crust. After Beaumont *et al.* (1991). Reprinted by permission of Chapman and Hall, ITPS Ltd.

achieved. For stress values greater than the critical value, a plastic deformation relation is used, in which the strain rate is proportional to the applied stress. Beaumont *et al.* (1991) represented this situation with the Navier–Stokes equation for conservation of momentum for viscous fluid flow, neglecting the inertial term because over geologic time scales, deformation of rocks is an extremely slow process:

$$\eta_e \nabla^2 \mathbf{v} - \nabla p + \rho_L \mathbf{g} = 0. \tag{1a}$$

Deformation is assumed to be incompressible:

$$\nabla \cdot \mathbf{v} = 0. \tag{1b}$$

Here η_e is an effective viscosity, $\mathbf{v}$ is the velocity field (three components), p is pressure, ρ_L is the density of the lithosphere, and $\mathbf{g}$ is gravitational acceleration. Results of this modeling effort show that the mountain topography, orographic effects, and wind direction produce spatially variable precipitation, which preferentially erodes the mountain range and collects in an evolving stream drainage network. Fluvial erosion leads to erosional "unloading" of the mountain range and additional uplift through isostatic rebound, in which the Earth's crust responds to a lightening of the rock "load" on the underlying mantle. The combination of tectonic compression and rebound produces spatial and temporal changes in uplift rates.

A detailed treatment of source-area hydrologic processes is provided by *rainfall-runoff* models, also known as *watershed* or *catchment* models, that represent the physics of hydrologic processes. All aspects of the hydrologic (water) cycle can be considered, including weathering, bedrock and soil erosion, mass movements, hillslope processes, infiltration of water into the subsurface, the uptake of water by plants, overland flow, and flow within stream channels. The amount of precipitation that infiltrates into the soil versus the amount that becomes stream flow is affected by the initial topography, soil types, soil moisture conditions, soil infiltration capacity, soil depth, groundwater flow conditions, groundwater storage properties, and the sequence of storm patterns, durations, and rainfall intensities. An example of a rainfall-runoff model is provided by Willgoose *et al.* (1991), who developed a source-area model that focuses on the long-term evolution of source-area topography and drainage-network development. Other methods of estimation of ancient stream flows from source areas, without explicit modeling of hydrologic processes or uplift, called *paleohydrologic* techniques, are discussed in Koltermann (1993).

2. CLASSIFICATION OF SEDIMENTARY BASINS

Sedimentary basins vary in size, shape, age, morphology, cause of formation, and patterns of basin fill. The location of fluid and sediment sources, rates of subsidence, patterns of sediment deposition, and areas of thickest sediment accumulation also vary. Some basins have been structurally deformed (folded and faulted), while others have been uplifted and eroded to form mountains. Despite the numerous factors that can cause variations between basins, geologists have tended to classify sedimentary basins on the basis of tectonic mechanisms of basin formation, proximity to tectonic plate boundaries, and type of plate boundaries (Dickinson, 1974, 1976; Ingersoll, 1988; Mitchell and Reading, 1986; Busby and Ingersoll, 1995). Table 1 contains an abbreviated basin classification system based on tectonic origin to illustrate some of the numerous basin types, some examples of which are shown in Fig. 3. Basin classification is useful for purposes of comparing common characteristics of basins and causes of basin subsidence, and for understanding the features observed in basins affected by basin-formation mechanisms.

Einsele (1992) points out that while the basin shape and size are controlled by tectonic processes, the morphology of the basin-fill sediment surface is strongly controlled by both tectonic and sedimentation processes. Factors such as locations, rates, and grain-size populations of sediment input and depositional, chemical, and hydrodynamic processes affect the spatial distribution of sediment types and ocean or lake water depths. To address these factors, studies of sedimentary basins account for depositional processes and environmental conditions affecting sedimentation as well as tectonic processes. These latter issues are discussed in the section on basin filling.

3. BASIN FORMATION

Sedimentary basins are formed by a downward sinking of the land surface known as *subsidence.* Three factors contribute to subsidence: tectonics, sediment loading, and compaction. Tectonic subsidence initiates basin formation. Tectonic subsidence is caused primarily by forces that thin the crust and by heating followed by thermal cooling and contraction, and to a lesser extent by compressive forces. As a basin evolves, the weight of accumulating sediments, ice, and water, called *loading,* becomes an additional driving force on subsidence. Tectonics and sediment loading cause subsidence of the basement rock that forms the floor of sedimentary basins as well as subsidence of the sediment surface. *Compaction* operates on the sediments themselves, causing subsidence of the sediment surface but not the

Table 1. Classification of sedimentary basins based on tectonic style [modified from Kingston *et al.* (1983), Mitchell and Reading (1986), and Einsele (1992)].

Basin category	Type of crust	Style of tectonics	Common basin characteristics
Continental or interior sag basins[a]	Continental[b]	Divergence[c]	Basins of regional extent, slow subsidence
Continental or interior fracture basins	Continental	Divergence	Narrow basins, fault-bounded valleys, rapid initial rift-stage subsidence
Basins on passive continental margins, margin sag basins	Transitional[d]	Divergence and shear	Asymmetric basins, outbuilding of sediment, slow subsidence during later stages
Oceanic sag basins	Oceanic[e]	Divergence	Asymmetric basins of regional extent, slow subsidence
Basins related to subduction	Oceanic or transitional/ Oceanic	Convergence[f] or dominantly divergence	Basins can be asymmetric, widely varying depth and subsidence rates
Basins related to collision	Oceanic or continental	Convergence or crustal flexure[g] with local convergence or transform motions[h]	Tendency to increasing subsidence through time, asymmetric basins, rapid sedimentary loading
Strike-slip/wrench basins[i]	Continental and/or oceanic	Transform motion, possible divergence or convergence	Relatively small elongated basins, rapid subsidence

[a] A sag basin is a broad, shallow basin of regional extent, with gently sloping sides.
[b] Continental crust underlies continents and continental shelves.
[c] Divergence occurs when two lithospheric plates move away from one another.
[d] Transitional crust occurs at the boundary between the oceans and the continents.
[e] Oceanic crust underlies oceans. It is thinner, denser, and of lower seismic velocity than continental crust.
[f] Convergence occurs when two lithospheric plates move toward one another.
[g] Flexure is regional bending of the lithosphere due to compression or loading.
[h] Transform motion occurs along transform faults, where plates slide past one another.
[i] A strike-slip or wrench basin occurs along transform margins.

basin floor. Subsidence occurs both above and below sea level, forming nonmarine and marine sedimentary basins, respectively. Subsidence rates can range from 10 to 1000 meters per million years.

3.1 Tectonics

Tectonic (or thermomechanical) subsidence mechanisms are driven by heat production from radioactive elements within the Earth. Horizontal motions of the Earth's crust drive vertical motions caused by changes in crustal thickness, temperature, and loading or unloading of the crust. Vertical motions form sedimentary basins and uplift source areas, and alter the path of sediment transported into and throughout basins. Tectonic mechanisms of basin formation are (Busby and Ingersoll, 1995)

1. *crustal thinning,* primarily caused by crustal extension (stretching or pulling apart) and rifting (large-scale faulting and tilting), but also caused by magmatic withdrawal;
2. *cooling and contraction* of lithosphere that was previously heated by underlying hot asthenosphere;
3. *tectonic loading,* in which subsidence occurs as the lithosphere compensates for the weight of mountains formed by crustal compression;

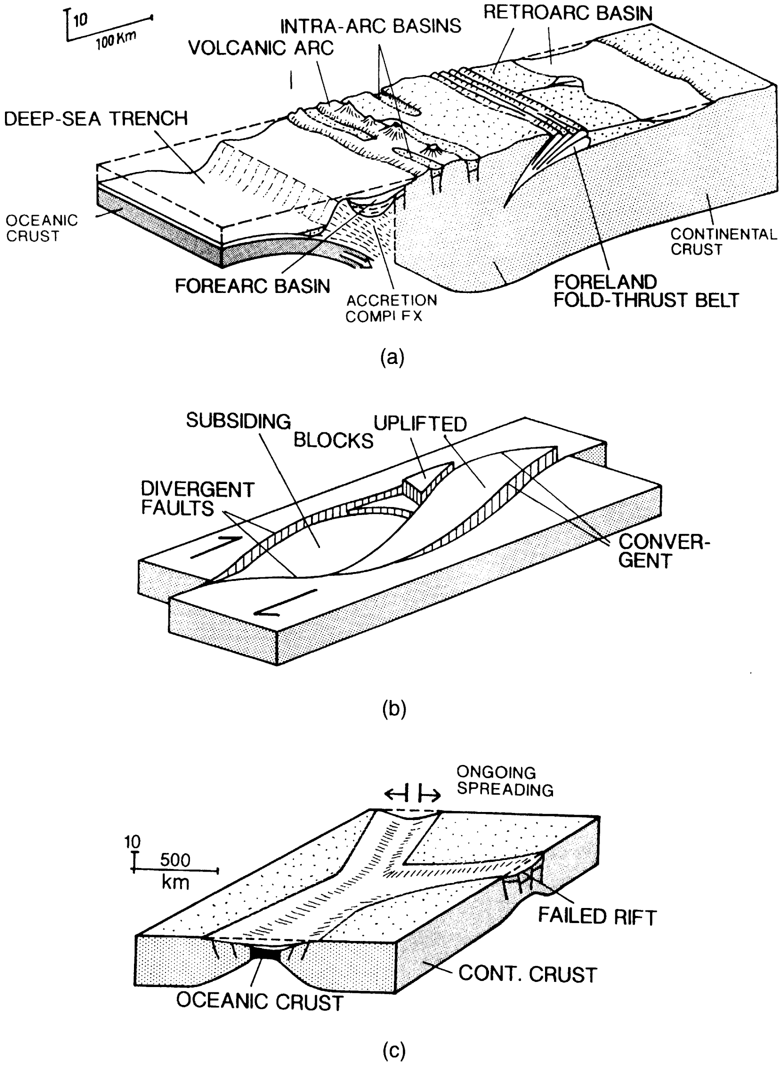
10
100 Km
INTRA-ARC BASINS
RETROARC BASIN
VOLCANIC ARC
DEEP-SEA TRENCH
OCEANIC CRUST
CONTINENTAL CRUST
FOREARC BASIN
ACCRETION COMPLEX
FORELAND FOLD-THRUST BELT
(a)
SUBSIDING
BLOCKS
UPLIFTED
DIVERGENT FAULTS
CONVERGENT
(b)
ONGOING SPREADING
10
500 km
FAILED RIFT
CONT. CRUST
OCEANIC CRUST
(c)

FIG. 3. Tectonic basin classification examples. (a) Continental margin basins formed at a *convergent* tectonic plate margin. This margin is subject to subduction as an oceanic plate dives beneath a continental plate, forming a deep-sea trench offshore, a string of volcanos (called a *volcanic arc*) onshore and parallel to the trench, and numerous associated sedimentary basins. A basin forms in front of the volcanic arc system (*forearc* basin). Sediments beneath this basin (the *accretion complex*) have been scraped off of the subducting oceanic plate and pile up (accrete) against the continental plate. Small basins form within the volcanic arc system from faulting, called *intra-arc* basins. Farther inland, convergence has caused faulting and folding, forming a mountain range known as a *foreland fold-thrust belt.* An elongate basin develops adjacent and parallel to the foreland fold-thrust belt, called a *foreland basin,* or here, a *retroarc* basin because it is behind the volcanic arc system. (b) Strike slip/wrench basins form along *transform* margins, where two tectonic plates slide past one another along a fault or a system of closely spaced faults. Any kinks in the fault(s) or splitting or joining of faults produces compressive or extensional forces that form sedimentary basins, where some blocks are uplifted while others subside. (c) A rift-zone basin, formed at a *divergent* tectonic plate margin. In this example, a new ocean basin floored by hot oceanic crust, shown in black, is being formed at a *spreading center,* along two arms of the rift zone, while a third arm has ceased to spread, and so is known as a *failed rift.* After Einsele (1992). Reprinted by permission of Springer-Verlag New York, Inc.

4. *volcanic loading,* in which subsidence occurs to compensate for the weight of a growing volcano;
5. *subcrustal loading,* in which the lithosphere experiences subsidence due to bending (*flexure*) from thickening due to faulting in the lithosphere caused by compressive forces;
6. *asthenospheric flow,* primarily caused by the descent of a subducting plate; and
7. *crustal densification,* in which the density of the crust increases because of changing pressure and temperature conditions, and the increased density causes sinking.

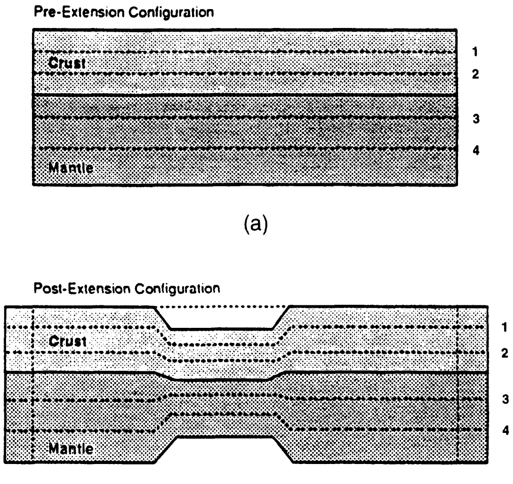

(a)

(b)

FIG. 4. Extension, rifting, and subsidence. (a) Pre-extension configuration of crust overlying mantle. Temperature isotherms are labeled 4 (hottest) to 1 (coolest). (b) Post-extension configuration of crust overlying mantle. Extension caused initial subsidence, an upwelling of mantle material, and hotter temperatures brought closer to the land surface. Gradual cooling of the heated rocks causes contraction and additional subsidence. Modified from Slingerland *et al.* (1994). Reprinted by permission of Prentice-Hall, Upper Saddle River, New Jersey.

Each plate tectonic setting has its own combination of controls on basin formation and subsidence (Dickinson, 1974, 1976). Extension and thermal contraction and cooling are characteristic of locations where tectonic plates diverge, such as rift zones. Tectonic and subcrustal loading is characteristic of locations where tectonic plates converge, such as subduction zones. In plate interiors and where plates slide past one another along large-scale faults, called transform faults, subsidence is caused by a combination of extensional and compressional mechanisms.

Crustal extension that produces rifting is the first stage in the formation of an ocean basin [Fig. 3(c)]. Rifts are long, narrow troughs bounded by faults that form because crustal extension ruptures the crust. Rifting occurs on continents or on thinned continental crust beneath oceanic waters. Extension of the lithosphere causes thinning of the crust, an upward vertical movement (*upwelling*) of upper mantle material, and a change in the lithosphere temperature profile (Fig. 4). The result of mantle upwelling is high temperatures brought close to the mantle–crust boundary and the land surface, an increased temperature gradient, some thermal expansion of thinned crustal rocks, and a downward isostatic adjustment of the crust. The thinned crust is riddled with faults and subject to rapid initial subsidence, forming a sedimentary basin. After extension ceases, heat is conducted upward and lost to the atmosphere, and the rocks cool, increase in density, and contract, causing thermal subsidence in addition to the subsidence caused by extension and rifting.

Thermal contraction and cooling causes thermal subsidence, controlled in part by the thermal properties of the *lithosphere,* a layer of strength that includes the Earth's crust and the upper part of the mantle, approximately 100 km in thickness. Thermal subsidence always follows heating of rocks. For example, the eruption of molten rock at plate boundaries, such as the mid-Atlantic ridge, forms a mountain range of high-temperature rocks. As the rocks cool, they contract, increase in density, and consequently subside. Cooling and contraction takes tens of millions of years; therefore, thermal subsidence is a very long-term process, commonly represented in basin-formation models by a decay-type equation.

A more rigorous modeling approach was taken by McKenzie (1978) and Jarvis and McKenzie (1980), who developed time-dependent analytical models for the dependence of heat flow and thermal subsidence on the rate and amount of crustal stretching during extension and rifting. The time history of the thermal anomaly developed beneath the rifted area is represented by a heat-flow equation:

$$\frac{\partial T}{\partial t} + G(Z - z)\frac{\partial T}{\partial z} = k\frac{\partial^2 T}{\partial z^2}, \qquad (2)$$

where T is temperature, t is time, k is the thermal diffusivity, Z is the elevation of the land surface, z is the elevation above the lithosphere–asthenosphere boundary, and G is the vertical velocity gradient.

Crustal *compression* results in an increase in crustal thickness as rocks are folded and faulted on top of their neighbors to form mountains. This increased weight of rocks on the Earth's crust, called *tectonic loading* of the crust, causes subsidence in front of the mountain belt, which, together with subsidence caused by crustal compression, forms a basin adjacent to the mountain range. Subsidence adjacent to loads of rock rather than just beneath such loads is known as *flexural* subsidence. The long, trough-shaped sedimentary basins formed in front of mountain ranges as a result of tectonic loading of rocks in the mountain range are known as *foreland* basins. Jordan (1981) used a model of flexural subsidence to show that tectonic loading coupled with erosion and redistribution of the mountain range can produce foreland basins that are hundreds of kilometers in width. Similarly, subcrustal loading occurs because of a thickening of the lithosphere due to faulting caused by compressive forces. The increased weight of the lithosphere is compensated for by subsidence. Regardless of the cause of the loading, models of loading account for both local and regional subsidence, as discussed in the next section.

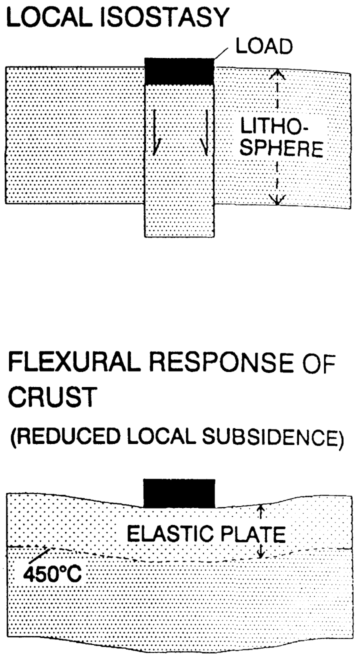

FIG. 5. Isostatic and flexural loading. Local isostatic response of the lithosphere to a load occurs directly beneath the load. Regional flexural response of the lithosphere to a load causes less subsidence beneath the load but greater subsidence over a broad area. After Einsele (1992). Reprinted by permission of Springer-Verlag New York, Inc.

3.2 Loading of the Lithosphere

Subsidence due to loading on the Earth's crust has been represented as either local *isostatic compensation* or *flexural bending* of the lithosphere. Mathematical models of isostatic and flexural subsidence are presented by Turcotte and Schubert (1982). Isostatic compensation assumes that the mass of sediments, lithosphere, and water above a certain depth of compensation within the mantle are balanced by an equal pressure from below (Fig. 5). The lithosphere, sediments, water, and ice are treated as one-dimensional vertical loads directly above the corresponding mantle. The isostatic response of the crust can be calculated with an Airy model, which assumes that the sediment load is entirely compensated at the location of the load when the load occurs (Turcotte and Schubert, 1982):

$$s = \frac{(\rho_m - \rho_w)}{(\rho_m - \rho_s)}(W_0 - W_f), \qquad (3)$$

where ρ_m is the density of the mantle, ρ_s is the density of sediment, ρ_w is the density of water, s is the thickness of the sediment load, and W_0 and W_f are the initial and final water depths.

In addition to the local deformation of the lithosphere beneath loads, the lithosphere adjacent to loads responds to either loading or unloading by subsidence or uplift, respectively. This is because the lithosphere has a strength, and loads from accumulating sediments or growing highlands are compensated regionally by flexure (Fig. 5). The strength of the lithosphere and the manner in which it deforms affect the distance over which loads are compensated and the resulting sedimentary basin geometry. Flexural models mathematically describe the strength of the crust to propagate and support loads of sediment and rock regionally. Flexural models typically represent the lithosphere as an elastic beam that has a specified flexural rigidity, in which deformation is recoverable and the mantle (treated as a denser fluid) supports the lithosphere and modifies its downward flexure (Turcotte and Schubert, 1982):

$$D\frac{d^4w}{dx^4} = q(x) - P\frac{d^2w}{dx^2}, \tag{4a}$$

$$D = Eh^3/12(1 - n), \tag{4b}$$

where D is the flexural rigidity, a measure of the elastic strength of the lithosphere and a function of the elastic constants of the rock, i.e., E (Young's modulus) and n (Poisson's ratio); h is the elastic thickness of the lithosphere; $q(x)$ is the downward force per unit area; P represents horizontal forces; and w is the vertical deflection. The thickness of the elastic lithosphere, h, is estimated either by a strength envelope, which designates the depths at which the lithosphere behaves in a brittle and ductile manner with increasing stress, or by the depth of the 450 °C isotherm, which has been found to produce a good approximation to the base of the elastic lithosphere. The strength and flexural rigidity of the lithosphere are temperature dependent, with both increasing as the lithosphere cools.

3.3 Compaction

The third subsidence mechanism, compaction, occurs as sediments are buried and subject to pressure from the increasing weight of overlying sediments. Compaction is the reduction in bulk volume, thickness, and void space of a volume of sediments. It is a post-depositional, or diagenetic, change to the original sediments. The amount of compaction varies with grain size, shape, and composition as well as with the initial porosity and with increasing pressure as sediments are buried under increasing loads of younger sediments. For example, clays compact more than sand grains. Clays often have extremely high porosities of 70–80% when first deposited, but they can compact to have porosities that are less than half of the original value. Although compaction is typically the least important subsidence mechanism in terms of magnitude, differential subsidence in sands and clays due to differing amounts of compaction strongly affects the land surface topography and consequently affects stream migration patterns. Compaction mechanisms include grain rearrangement to make a tighter particle packing; dissolution of minerals at grain-to-grain contacts because of pressure, called *pressure solution;* grain deformation and fracturing; and the release of water bound to clay minerals. These mechanisms are typically represented empirically in basin-formation models with curves showing how porosity decreases with increasing depth of burial for different sediment types.

Subsurface fluid pressures also exert control on compaction. Most curves of porosity versus depth of burial are developed from field data under hydrostatic pressures; i.e., pore fluids expelled during burial are allowed to escape through the overlying sediments, so that there is an equilibrium between increasing pressure from sediment accumulation, compaction, and pore-water expulsion. However, rapid deposition of low-permeability fine-grained sediments inhibits pore-fluid expulsion. Then, the pore fluid bears some of the weight of the overlying sediment pile, producing an overpressured zone, where pore-water pressures are greater than hydrostatic. In these zones, compaction is reduced from expected values. This situation has been modeled by Bethke *et al.* (1990), who focused on overpressured zones in the Gulf Coast of the United States.

4. BASIN FILLING

Basin filling is strongly controlled by factors governing basin formation, as well as the source area, sea-level changes, energy conditions within the basin, and conditions conducive to the formation of chemical and biological sediments. As a basin fills, the horizontal and vertical land-block movements caused by thermal or mechanical deformation of the Earth's crust, sediment loading, and compaction provide a continuously moving land surface upon which sediments accumulate. The size, shape, and subsidence rates of the basin sediment container control how much of the sediment influx is trapped within a basin and where it accumulates. The basin geometry and rates of subsidence versus sedimentation dictate whether the sediment pile is preserved or whether there are erosion-caused gaps in the sedimentary record. The basin-fill characteristics strongly depend on the rate at which sediment is received and the composition of the sediment. The source area controls the volume and grain-size distribution of sediment and the dissolved mineral load supplied to a basin. In depositional systems above sea level, the dissolved mineral load that infiltrates into the groundwater flow system plays an important role in post-depositional, or diagenetic, changes to sediments, such as the chemical precipitation of minerals in the pore spaces. The location of the shoreline and energy conditions above and below sea level affect the transport and deposition of different sediment grain sizes and biological activity. While the basin-formation processes of uplift and subsidence are critical to the production of clastic sediments and development of the basin container, the basin-filling processes of sediment erosion, transport by water and wind, deposition, water circulation, and chemical and biological activity control most processes occurring within a basin. All of these factors work together to create spatial patterns in different sediment types throughout basin fill.

The spatial patterns and the fossil content of basin-fill sediments are diagnostic of the depositional and environmental conditions under which sediments were deposited, called *depositional environments*. Each depositional environment has a unique set of physical (especially hydrodynamic), chemical, and biological conditions and particular depositional processes, thus producing a characteristic sedimentary deposit (Reineck and Singh, 1980). Within each depositional environment, sedimentary *facies* can be distinguished, defined as the "sum of all the primary characteristics of a sedimentary unit" produced as a result of deposition under given environmental conditions (Reineck and Singh, 1980). These common characteristics include fossil content, sediment chemistry, grain-size distribution, and small-scale sedimentary features. Common characteristics of rocks and sediments are compiled in *facies models*, which visually show, through maps, cross sections, block diagrams, and one-dimensional vertical columns, the typical configuration of sediments deposited within each depositional environment. Facies models are used to interpret and predict the grain size, mineralogical composition, and spatial arrangement of stratigraphic units. Of special interest to earth scientists interested in extraction of hydrocarbons, groundwater, and contaminants from the subsurface are the sand- and gravel-body geometries predicted by facies models, because interconnected sands and gravels produce connected high-permeability pathways for fluid flow.

4.1 Depositional Environments and Facies Models

Depositional environments and facies models have been described in detail by many authors (for example, Reineck and Singh, 1980; Blatt *et al.*, 1980; Reading, 1986; Einsele, 1992). A brief description of common environments based on these references is provided here. Here, depositional environments are divided into the following general categories:

1. nonmarine (fluvial, glacial, desert, and lacustrine);
2. coastal and shallow-sea sediments (beaches, estuaries, deltas, and continental shelf); and
3. marine (continental margin, slope, and deep ocean basins).

It should be noted that sediment can be transported from one depositional environment to another, such as from source-area highlands on continents to marine basins. In addition, sediments of volcanic origin occur in many depositional environments.

4.1.1 Nonmarine Sediments Nonmarine depositional environments are those occurring on land or in surface-water bodies located above sea level. These sediments can be of stream, wind, lake, volcanic, or swamp origin. Sediments are deposited under fluvial, glacial, lacustrine, and desert conditions. *Fluvial* sediments are produced by the action of streams or rivers that erode, transport, and deposit sediment from highlands to lowlands. *Glacial* sediments are deposited by the action of ice or glaciers, while *lacustrine* deposits are deposited in or formed in a lake. Within desert environments, *eolian* sediments are transported and deposited by the wind. Commonly, fluvial, glacial, and eolian sediments are not conducive to the preservation of fossils. A remarkable exception to this statement is the recent discovery of a fossil nesting dinosaur, whose preservation is attributed to burial in a sandstorm (Norell *et al.*, 1995).

Rivers are the most important geomorphic agent in transporting fluid and sediment from highland source areas on the continents to lakes and the ocean. The most notable feature of fluvial systems is the channel pattern, that is, the geometry of channels in map view. The number of channels and their width, depth, slope, sinuosity, and type of sediment load are diagnostic. *Sinuosity* is a measure of the length of a stream channel versus the length of the valley it flows through. Sinuous channels have many bends; straight channels have few or no bends. Sediment load is composed of bedload and suspended load, the proportions of which are affected by sediment availability and stream hydrodynamics. Streams transport the larger, heavier grains, such as gravel and sand, as *bedload*, by rolling, bouncing, and sliding on the stream bottom. In contrast, finer-grained sediments, such as silts and clays, are transported in suspension within the water column by turbulence and called *suspended load.* Fluvial depositional environments include meandering rivers, braided streams, anastomosing rivers, and alluvial fans (Fig. 6). Within each fluvial depositional environment, there are subenvironments, including the channel, its levees, and flood-plain sediments.

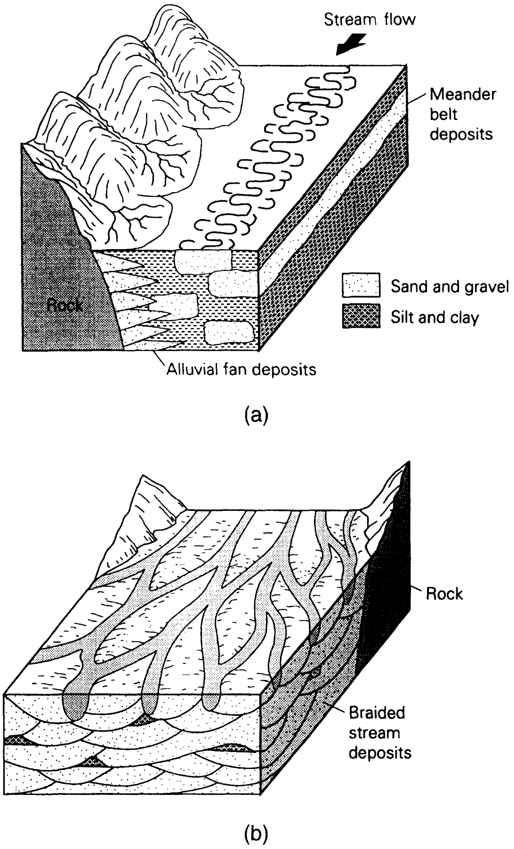

FIG. 6. Fluvial depositional environments. (A) Sands deposited within alluvial fans and by a meandering stream. The wedge-shaped alluvial-fan sands are adjacent to the mountain range, while the sands deposited from the meandering stream lie within and are parallel to the valley axis. (B) Sands from braided-stream deposits with little fine-grained sediment. Sands appear as sheets, ribbons, wedges, or tabular elongate bodies. The legend to the right of (A) also applies to (B); that is, sand and gravel are shown as lightly stippled, while silt and clay are shown as dark-stippled areas. Modified from Allen (1965). Reprinted by permission of Blackwell Science Ltd.

Meandering rivers have one highly sinuous primary channel. These rivers typically form where the land surface slope is low, the stream banks are composed of cohesive materials, the flow discharge is relatively steady, and the suspended-load/bedload ratio is high. When a meandering stream overflows its banks, suspended load in the flood waters is transported out of the stream channel to settle out over the flood plain. Coarse-grained sediments remain behind, transported on the stream channel bottom. The sediments deposited by meandering streams are coarse-grained stream channel deposits embedded in fine-grained flood-plain deposits. In areal view, a *braided* stream looks like a wide, shallow channel containing many interwoven small channels separated by small sand and gravel islands. Channels repeatedly divide to flow around the islands and then rejoin. Channels shift position frequently. Braided streams typically transport coarse-grained bedload sediments, have a steep stream-bottom gradient, and are relatively straight and wide. In sedimentary deposits formed by fluvial processes, the channel type and proportions of bedload versus suspended load affect the geometry of the channel deposits preserved in the sedimentary record (Fig. 7). *Alluvial fans* occur where a stream flows out from a steep, narrow mountain canyon onto a lower-gradient valley or coastal plain, depositing sediment in a characteristic shape that looks like a segment of a cone, where the apex of the cone is at the highest elevation. The fan shape is produced by the decrease in stream power as the gradient abruptly decreases, causing rapid deposition of transported sediments. The flows can be either sediment-laden water flows or debris flows. An alluvial fan that extends into an ocean or a lake is called a *fan delta*.

Glacial sediments are formed by the action of ice and occur both on the continents and in shallow marine waters. The most common sediments in this environment are produced by continental ice sheets, which transport debris from upland source areas, erode sediment at the ice sheet base, and deform the underlying rock. When the ice melts and the glacier retreats, the transported debris is dropped. In contrast to fluvial deposits, the processes of sediment transport within, beneath, and in front of ice sheets do not separate sediments by grain size. Deposits from glaciers are a mixture of grain sizes ranging from boulders to clay-sized sediments, with large boulders or *clasts* appearing to occur at random in a matrix of silt and clay-sized material. Sediments deposited by glacial action are called *till.* The morphology of glacial deposits is indicative of depositional processes. For example, *moraines* are elongated hills that occur at the sides and front of glaciers resulting from debris pushed at the edges of the ice. *Drumlins* and *eskers* form beneath a glacier where there are fissures in the ice. They are teardrop-shaped hills and narrow ridges on the land surface, respectively, left behind after a glacier has melted. The large volume of water produced by a melting glacier forms *meltwater streams* in front of glaciers that transport sediment. *Glacial outwash* deposits occur adjacent to ice sheets and are primarily composed of gravels and coarse sands transported short distances. As meltwater streams flow farther from a glacier, they deposit *glaciofluvial* sediments, with characteristics similar to alluvial fans and braided streams.

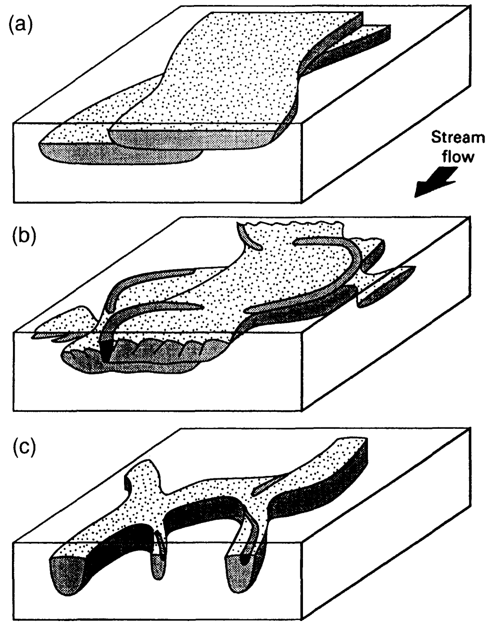

FIG. 7. Fluvial sand-body geometries. A schematic of the three-dimensional geometry of sand bodies typically seen in (A) bedload, (B) mixed-load, and (C) suspended-load fluvial systems. The channel type and proportions of bedload versus suspended load affect the geometry of the channel deposits preserved in the sedimentary record. Modified from Galloway and Hobday (1983). Reprinted by permission of Springer-Verlag New York, Inc.

Lacustrine sediments, or sediments deposited in lakes, include both clastic sediments transported into lakes and sediments created by biological or chemical activity within lakes. Lakes typically contain fresh water, but some have salinities greater than that of sea water. Incoming streams carrying clastic sediment are affected by the density difference between the lake water and the incoming mixture of water and sediment transported by a stream. Flows denser than lake water occur along the lake bottom, flows of similar density mix within the water column, and less dense flows float at the lake water surface. Coarse-grained materials are primarily deposited near the shoreline, with finer-grained materials deposited successively farther from the shoreline. *Varves* are alternating, horizontal, thin sequences of lacustrine sediment reflecting settling from suspension of fine-grained particles during seasonal changes in temperature or from seasonal precipitation of minerals caused by differences in evaporation. Typically, summer layers are light-colored, and winter layers are dark-colored. The proportion of lake sediments derived from biological activity is strongly controlled by the lake water chemistry and the influx of nutrients. In particular, the salinity content and chemical composition of lakes exert strong control on biological activity and therefore on sediments composed of fossil plant and animal remains. Lacustrine sediments of chemical origin provide deposits of salts that are mined for their economic value, including carbonates, sulfates, chlorides, borates, and nitrates.

Desert environments occur where rainfall is insufficient to produce much vegetation, mean annual evaporation far exceeds precipitation, and sediment transport by wind is significant. Eolian deposits are composed of wind-blown particles, typically fine to medium sand and silt. During transport, silt and some sand particles are carried in suspension within turbulent air, while larger, heavier sand particles may be bounced or rolled along the land surface in a process called *saltation.* During deposition, the action of wind blowing on dry sediment particles produces ripples, dunes, ridges, and large hills characteristic of eolian sediments (Fig. 8). While eolian deposits are typically thought of as occurring only in deserts and on sandy beaches, they also occur within fluvial environments. When rainstorms occur in the desert, they are often violent downpours. Coupling of intense rain with a lack of vegetation produces sediment-laden flash floods and debris flows that produce fan-shaped sedimentary deposits in desert valleys.

4.1.2 Coastal and Shallow-Sea Sediments Coastal and shallow-sea environments include beaches, tidal flats, lagoons, estuaries, deltas, and continental shelves. Sediments deposited in shallow coastal waters and on the continental shelf are affected by ocean energy conditions as well as by processes that control the location, timing, volume, and grain-size distribution of sediment influx. Waves, tidal action, and bottom and surface currents serve to redistribute and rework sediments pouring in from continental sources, removing finer-grained ma-

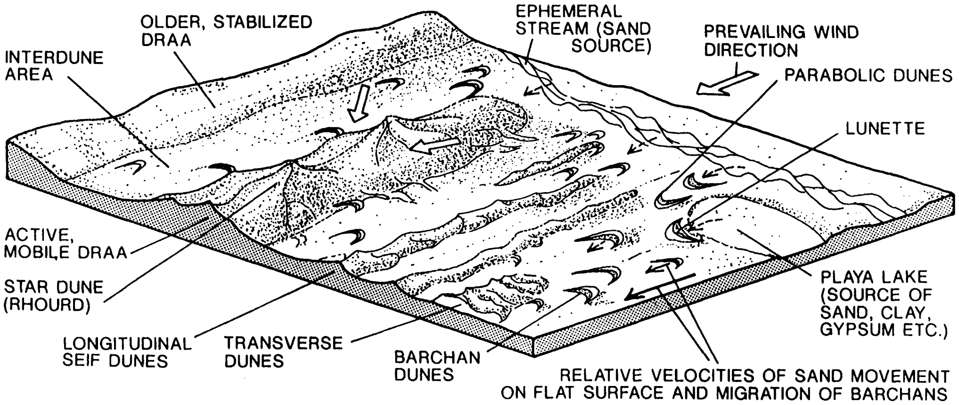
OLDER, STABILIZED DRAA
INTERDUNE AREA
EPHEMERAL STREAM (SAND SOURCE)
PREVAILING WIND DIRECTION
PARABOLIC DUNES
LUNETTE
ACTIVE, MOBILE DRAA
STAR DUNE (RHOURD)
LONGITUDINAL SEIF DUNES
TRANSVERSE DUNES
BARCHAN DUNES
PLAYA LAKE (SOURCE OF SAND, CLAY, GYPSUM ETC.)
RELATIVE VELOCITIES OF SAND MOVEMENT ON FLAT SURFACE AND MIGRATION OF BARCHANS

FIG. 8. Eolian depositional environment. The action of wind transporting sediment particles produces many different types of dunes and *draas,* which are very large-scale dunes. In eolian environments, infrequent rains produce infrequent (ephemeral) streams and *playa lakes* that produce evaporitic minerals as they dry up. *Barchan dunes* are isolated, crescent-shaped sand dunes that lie transverse to the prevailing wind direction and are convex in the upwind direction. The wings or horns of barchan dunes point downwind. Barchan dunes can grow to heights of more than 30 m and widths up to 350 m from horn to horn. They form where the sand supply is limited and the wind is constant and of moderate velocity. In contrast, *parabolic dunes* have a long, scoop-shaped (parabolic) form that is convex in the downwind direction, so that its horns point upwind. Parabolic dunes are commonly covered with sparse vegetation and occur along coastal areas where strong onshore winds supply abundant sand. After Einsele (1992). Reprinted by permission of Springer-Verlag New York, Inc.

terials and transporting them into deeper water. Wave energy and the *longshore currents* produced by waves (currents along the coastline) produce commonly observed features, such as cliff erosion, barrier islands, spits, and lagoons. Storms, particularly those with large waves, alter the sediments and sediment geometries produced under normal depositional conditions.

The coast forms the transition between marine ocean and nonmarine continental conditions. Coastal sediments are typically sand but occasionally gravel. Finer-grained silts and clays are transported in suspension into the ocean, whereas the coarser-grained sand and gravel bedload transported by streams is deposited at and near the shoreline. Wave energy and shallow marine currents along the coast redistribute sediment deposited by rivers and exert control on beach morphology. Beach sand can also be reworked by the wind into coastal dunes.

Estuaries are the funnel-shaped seaward end of rivers where fresh water mixes with sea water. Estuary salinity and water level is affected on a daily basis by tidal fluctuations and mixing between fresh water and sea water. The water in an estuary is often stratified, with denser sea water flowing beneath less dense fresh water. Sediments deposited in estuaries are typically clays to sands of both marine and fluvial origin mixed with decomposed organic matter.

Deltas occur where a stream enters a body of standing surface water, depositing sediment that forms a large expanse of land of low relief that builds out into a lake or the ocean. Formation of a delta requires a high sediment load transported in the stream plus subsidence in the location of deltaic sediment deposition. A portion of the delta is above lake or sea level, while the rest is below. Commonly, the areal shape of a delta resembles a segment of a cone, or a triangle; however, the shape is affected by tidal action, coastline shape, the degree of sediment transport along the coastline, and wave action. The stream flowing over the delta branches into numerous smaller streams, called *distributaries*. Marine deltas occur in shallow ocean waters along coastlines where the sediment supply is greater than the amount of sediment waves and currents can erode. When the volume of sediment supplied to a delta is large, the delta builds out (*progrades*) into the standing body of surface water, producing a characteristic sediment geometry in vertical section. The geometry of delta bedding includes relatively flat-lying, usually coarse-grained sediments at and above the lake or ocean (*topset* beds); steeply inclined sandy sediments from near the water surface to near the base of the slope (*foreset* beds); and gently inclined fine-grained sediments at the bottom (*bottomset* beds). Sea-level fluctuations affect the position of the shoreline and, thus, the position of finer-grained marine versus coarser-grained nonmarine deposition and sediments of biological origin. The relative contributions of energy from waves, tides, storms, and marine currents affect the geometry of coarse- versus fine-grained sediments within deltas (Fig. 9).

The common features and depositional environments of a common continental margin are shown in Fig. 10. The *continental shelf* refers to the gently sloping underwater zone from the shoreline down to about 200 m ocean water depth. The magnitude of shelf currents caused by tides, density differences, and waves is highly variable. Currents can distribute high volumes of sediment on the shelf and produce large, migrating ripple structures on the sea floor. Large storms also play a significant role in distributing shelf sediment. Commonly, sediment grain size varies as the water depth increases, from sands at the coast, to clayey silt to silty sand in a transition zone that starts just below the wave base, to mud in the deepest waters of the shelf. Where currents are not strong, there is biological activity. Sediments are *bioturbated* (churned up) by the action of bottom-dwelling organisms, often obliterating sedimentary features formed during sediment deposition.

4.1.3 Marine Sediments There are numerous marine depositional environments, affected by water depth, temperature, salinity, and energy conditions. Beyond the continental shelf lies the *continental slope*, which has a much steeper gradient than does the shelf. The base of the slope lies in deep water and is called the *continental rise*. Sediments on the continental rise are largely from material that has slumped off the continental slope, called *slope apron deposits*, or sediments transported as density currents

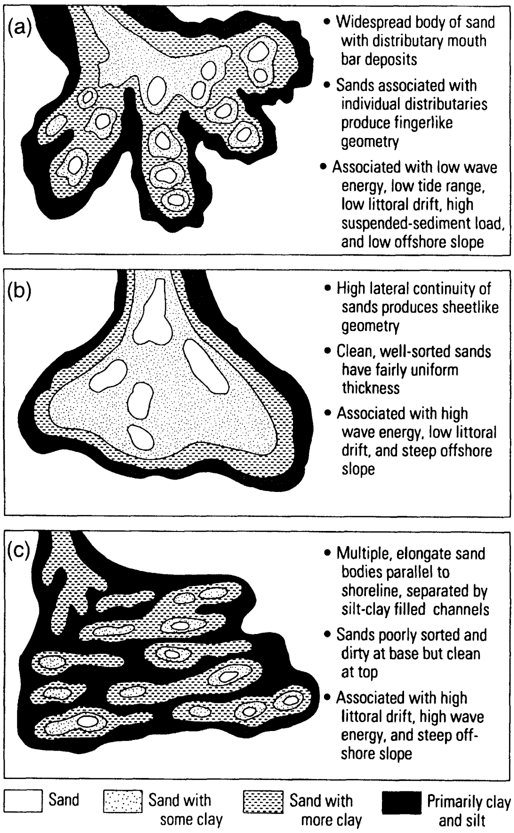

(a)

- Widespread body of sand with distributary mouth bar deposits
- Sands associated with individual distributaries produce fingerlike geometry
- Associated with low wave energy, low tide range, low littoral drift, high suspended-sediment load, and low offshore slope

(b)

- High lateral continuity of sands produces sheetlike geometry
- Clean, well-sorted sands have fairly uniform thickness
- Associated with high wave energy, low littoral drift, and steep offshore slope

(c)

- Multiple, elongate sand bodies parallel to shoreline, separated by silt-clay filled channels
- Sands poorly sorted and dirty at base but clean at top
- Associated with high littoral drift, high wave energy, and steep off-shore slope

FIG. 9. Deltaic depositional environments. The relative contributions of energy from waves, tides, storms, and marine currents affects the geometry of coarse- versus fine-grained sediments within deltas. Modified from Coleman and Wright (1975). Reprinted by permission of Houston Geological Society.

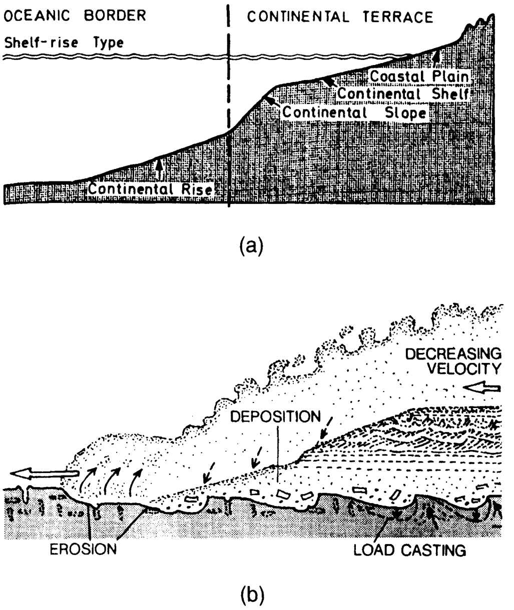

FIG. 10. Marine depositional environments and processes. (A) The common features and depositional environments of a common continental margin include the continental shelf, continental slope, continental rise, and abyssal plains (or ocean basins). After Curray (1969). Reprinted by permission of the American Geological Institute. (B) Turbidity currents are a common sediment transport mechanism in the ocean. As a turbidity current travels down the steeply dipping continental slope and reaches the more gently sloped continental rise and ocean basin, its velocity decreases and deposition starts, here shown as the classic depositional sequence (*Bouma* sequence). Deposition from a turbidity current onto unconsolidated mud produces *load casts,* formed by partial sinking of the deposited materials into the mud. After Einsele (1992). Reprinted by permission of Springer-Verlag New York, Inc.

through submarine canyons to form deep-sea fans. The deepest depositional environments are the *abyssal plains,* or *ocean basins,* which lie beyond the continental rise. Ocean basins are monotonous in their extent and usual lack of topographic features, except for occasional volcanic cones. Their depth is exceeded only by the depth of deep-sea trenches, which are about 2 km deeper than the surrounding ocean floor. Areas of higher elevation within marine environments include sediments deposited on top of seamounts and submarine ridges or platforms. In these locations, sediment sequences are typically thin and composed largely of sediments of biological or chemical origin. Although the continental shelf, slope, and rise type margin has been discussed here, it should be noted that the margins of some continents do not have a well-developed shelf, slope, and rise. Rather, deep-sea trenches, folded and faulted rocks, or submerged steep cliffs are present.

Oceanic sediments are composed of materials transported into the sea by water, wind, or mass movements as well as by sediments of chemical or biological origin produced within the sea. By far, rivers are the greatest contributors of sediment to the oceans. However, continental sediments delivered to the oceans by rivers can be efficiently transported and deposited far from their source areas and at substantial depths within the oceans. While continental margins accumulate sediment largely from deposition by flowing sediment/water mixtures, the portions of ocean basins farthest from the continents accumulate sediments primarily by vertical settling of very fine-grained sediment. Sediment can be redistributed into deeper water by the force of gravity or transported long distances by ocean currents. For example, slope failure on the continental shelf produces mass movements that travel at high velocities down submarine canyons to form deep-sea fans. Transport of coarse sediment long distances into deep ocean waters occurs by *turbidity currents,* which are bottom-flowing density currents containing high concentrations of suspended sediment that travel at high velocities and spread areally when not confined by an underwater canyon. Turbidity currents occur in lakes as well as the ocean. High volumes of sediment are also transported downslope by large-scale gravity slides and slumps. Salinity, temperature, and water pressure affect the density of ocean waters and thus exert control on ocean circulation. Denser water sinks and flows beneath less-dense water to form an ocean bottom current. The result of these density currents is that a global pattern of ocean circulation is caused by the temperature difference between the equator and the poles. The other strong control on ocean circulation is global wind patterns that control wind-driven surface currents.

Sediments of biological origin commonly occur in sedimentary basins with low sedimentation rates, because the growth of organisms is adversely affected by high volumes of clastic sediments. Sediments of biological origin are produced either within the water column (*planktonic*) or on the sea floor (*benthonic*). To thrive, each organism has its own set of required environmental conditions regarding water salinity, temperature, depth, and energy conditions. For example, formation of *carbonate deposits* of biological origin depends on conditions favorable for carbonate-producing marine life, including warm water temperatures, sunlight, and a limited influx of gravel, sand, silt, and clay. Warm, calm, shallow marine waters provide a good environment for the development of carbonates. Carbonates of biological origin are the skeletons or shells of carbonate-producing animals that live in warm, shallow marine waters, typically at depths less than 10 to 20 m and in low latitudes. A reef is an example of a carbonate deposit formed from the structures of frame-building animals such as corals. Sediments formed by slow vertical settling of particles of biological origin include muds containing significant amounts of continental, volcanic, or calcium carbonate materials, oozes, chalk, limestones, and clays. For example, carbonates can result from the skeletons of carbonate-producing animals such as *foraminifera* that settle to the sea floor after the animal dies. Other sediments of biological origin produced in the ocean include the siliceous skeletons of organisms such as diatoms, sponges, and radiolaria. These siliceous skeletons accumulate to form bands and layers of a colorful mineral called *chert.* The proportion of deep marine sediments attributed to the continents or volcanic origins versus the proportion derived from organic skeletal remains of micro-organisms is diagnostic of depositional environments. Sedimentary deposits in which less than 25% of the coarse fraction is composed of continental, volcanic, or larger organisms of continental shelf origin are typically located beyond the continental margin and on the abyssal plain. In contrast, sedimentary deposits in which more than 25% of the coarse fraction is predominantly composed of continental, volcanic, and skeletons of shallow-water organisms usually accumulate near the continental margin.

Shallow seas and other areas of shallow marine waters are the typical sites for sediment of chemical precipitation. Sediments of chemical origin are formed under specific temperature, salinity, *p*H, and oxidation-reduction (Eh) conditions. Carbonate minerals of chemical origin are inorganically precipitated from sea water and include a suite of common rock-forming minerals, such as calcium and magnesium carbonate. Carbonates formed by chemical precipitation require particular environmental conditions causing evaporation and chemical supersaturation of carbonate constituents. The production of carbonate minerals in the ocean is affected by the amount of calcium transported into the oceans from rivers, the ability of animals to extract calcium from sea water, and the availability of other nutrients required by carbonate-producing organisms (Gross, 1980).

Small oceanic basins connected to the world oceans by restricted inlets/outlets can be subject to conditions that differ from those of the open ocean. Examples include the Gulf of California and the Persian Gulf. The salinity of these small oceanic basins can be the same as, greater than, or less than that of the open ocean, depending on the cross-sectional area of the opening to the ocean, the amount of fresh-water influx, and climate conditions, such as conditions amenable to evaporation. Where there is little freshwater influx from precipitation and streams, these basins can attain salinities far greater than average seawater salinities and are therefore good candidates for sediments of chemical origin. In contrast, in areas of large freshwater influx, these basins can have salinities less than that of sea water.

4.1.4 Volcanic Sediments Sediments of volcanic origin can contribute substantially to the total volume of basin fill. The size of volcanic rock fragments ranges from microscopic dust and ash particles to huge blocks. Initial transport and deposition of volcanic rocks is accomplished by fallout and flows. *Fallout* occurs from an ash cloud produced by an eruption. *Flows* are hot mixtures of melted volcanic rock, gas, and solid particles that flow downhill under the influence of gravity and collect, cool, and solidify in topographic depressions.

4.2 Basin-Filling Processes

In addition to an understanding of depositional environments, study of basin filling requires an understanding of how sediment is transported and deposited on a small scale, as well as how large-scale patterns in sedimentary basin fill are produced. For example, knowledge of how sediment is dispersed throughout a basin by wind, water, and mass movements is the key to understanding many features observed in the rock record. Sedimentation patterns, particularly cycles of sedimentation, can be controlled by processes that operate within a sedimentary basin, called internal or *autocyclic* processes, and by processes external to a sedimentary basin, called *allocyclic* processes (Miall, 1984; Einsele, 1992). External factors, such as tectonics, climate change, and sea-level change, affect sedimentation patterns over a large areal extent and therefore produce regional patterns in basin-fill sediments. Tectonic forces exert strong control over the locations of source areas versus receiving basins and rates of uplift and subsidence. Tectonics affects the basin geometry, the elevation difference between the source area and receiving basin, and can influence hydrologic processes through the source-area size, elevation, and topography. Sea level controls the location of marine versus nonmarine depositional environments and associated sediment grain sizes, depositional features such as cross bedding, and fossil content, while the basin energy conditions control how sediment is distributed once it arrives. Climate change affects local lake levels, the magnitude and frequency of stream flows, the amount and composition of sediment transported by streams, weathering and sediment production in the source area, and rates of erosion in the source area. In contrast, internal processes, such as the switching of a stream channel to a lower elevation within a stream valley (*avulsion*) and switching of areas of lobe-shaped deposition on marine deltas (*delta lobes*), contribute to variability in sedimentation patterns over a limited areal extent.

Sediment is transported into a sedimentary basin by water, wind, or the force of gravity. For mixtures of water and sediment, there is a spectrum of transport processes in which the proportion of fluid to sediment grades from relatively clear streams to flows dominated by mud and debris. One must distinguish between *subaqueous* flows, in which sediment is transported beneath the surface of a standing body of water (a lake or an ocean), and *subaerial* flows, in which the flow is in contact with the atmosphere. In subaerial flows, the same fluvial processes that transported sediment into the basin also serve to distribute sediment throughout the basin. In contrast, while fluvial processes may transport sediment into a lake or ocean, subaqueous processes take over the transport and distribution of sediment beneath the water surface. In subaqueous flows, sediment transport is also affected by gravity, density, wind-produced waves and currents, and ocean circulation.

When water is the transporting medium in a mixture of fluid and sediment, the flow is known as a *fluid gravity flow*. These flows obey Newton's law of viscosity. The sediment concentration is small enough (less than 5%) that it does not appreciatively affect the rheological properties of the mixture, and both grain-to-grain interactions and the displacement of fluid by falling grains within the water column are not significant (Blatt *et al.*, 1980). Therefore, the effects of sediment concentration on fluid density, viscosity, particle settling velocities, flow dynamics, and sedimentation mechanics need not be considered in modeling of fluid gravity flows (Tetzlaff and Harbaugh, 1989).

The processes that transport sediment are discussed in detail in Allen (1985) and Blatt *et al.* (1980). In fluid gravity flows, the processes by which sediment is transported are quite different for coarser sediment transported on the stream bed, called bedload, and for finer sediment transported in the water column, called suspended load. Bedload grains are transported by rolling, bouncing, and sliding on the stream bed, whereas suspended-load grains remain in transport by turbulence in the water column. Particles that can be supported in suspension by turbulence are transported from the bedload upward into the suspended load. Deposition occurs when the flow can no longer support particles in suspension or can no longer move particles on the bed. Stream valley deposits of gravels and sands embedded in silts and clays are an example of a sedimentary deposit produced by fluid gravity flows.

When a stream flows into a lake or ocean, the flow cross-sectional area suddenly expands, the stream velocity drops, and the flow turbulence decreases, reducing the flow's capacity to transport both suspended sediment and bedload. Consequently, bedload sediment is rapidly deposited at the mouth of the stream. Clay particles start to flocculate as they are affected by the change in water chemistry from fresh stream water to saline (alkaline) ocean water. *Flocculation* is the process by which small particles suspended in the water column clump together to make bigger particles with higher settling velocities.

Rapid deposition seaward of the stream mouth causes oversteepened slopes on the floor of the lake or ocean. Such slopes are unstable and therefore provide sources of sediment to deeper portions of basins through mechanisms such as slumping, sliding, and flows of dense sediment–water mixtures. The offshore, steeply sloped foreset beds of marine deltas are an example of structures formed by these processes.

A viscous mixture of fluid and sediment that moves downslope under the force of gravity rather than by fluid transport is called a *mass movement.* Mass movements typically occur where there are both a steep topographic gradient and sufficient rock or sediment for slope failure. Rock falls, slides, slumps, creep, and sediment gravity flows form the basic types of mass movements (Blatt *et al.*, 1980). In *rock falls,* a pile of angular rock fragments is deposited adjacent to the source of the fragments. Rock falls occur on steep slopes either on dry land or beneath the sea. *Slides* are characterized by the quick downslope movement of a mass of soil and rock over a surface at the base of the slide, called a *shear surface.* The mass is subject to little internal deformation but may rotate. *Slumps* are quick downslope mass movements that experience rotation and much internal deformation. If the sliding or slumping mass experiences *liquefaction,* in which the solid sediment mass is transformed to a fluid mass because of increased pore pressures and loss of strength, a debris or mudflow develops. *Creep* is the slow downslope movement of soils and unconsolidated sediments.

The mass movements most important in terms of total volume of sediments observed in the geologic record are dense mixtures of fluid and sediment called *sediment gravity flows.* In contrast to fluid gravity flows, in sediment gravity flows the sediment concentration and grain interactions affect the particle settling velocities and the fluid–sediment mixture density and viscosity. Four transport mechanisms have been identified (Blatt *et al.*, 1980). First, turbulence within the fluid–sediment mixture exerts an upward force on sediment grains. Second, grain-to-grain collisions cause grains to support each other. Third, as grains settle under the force of gravity, fluid is displaced, rises, and exerts an upward force on other sediment grains. Fourth, a mixture of fluid and fine-grained sediment has high strength and can therefore support larger grains within the flow. As a result, coarse-grained particles are transported well above the base of the flow, not just at the base of the flow as in fluid gravity flows. Two common forms of sediment gravity flows are *debris* or *mud flows* and *turbidity currents* (Fig. 10). Debris flows are composed of rock fragments, soil, and mud, with large fragments dispersed in and supported by a finer-grained cohesive matrix. Mud flows are primarily fine-grained materials, with few larger fragments. Turbidity currents are high-velocity, short-duration, underwater, turbulent suspension currents that occur as a result of either suspended-sediment–laden floodwaters or, more commonly, a failed sediment mass that entrains ocean or lake water as it proceeds downslope. Sediments deposited by turbidity currents are known as *turbidites*. The grain size and composition of turbidites preserved in the geologic record is diagnostic of source-area parent sediment type, turbidity-current density, sediment transport mechanisms, and proximity to source area. Sediment gravity flows commonly form fan-shaped deposits located seaward of submarine canyons and large rivers, known as submarine fans. Deltas and deep-sea fans are common settings for widespread creep, slumps, debris and mud flows, and turbidity currents.

4.3 Modeling of Basin Filling

Models of basin filling represent spatial and temporal changes in the volume and distribution of sediments accumulating in a basin. These models seek to reproduce sedimentary geometries, locations of depositional environments, and facies relations within a subsiding sedimentary basin. To untangle the relative contribution of each geologic process to the observed geology, all relevant external and internal forcing factors as well as fluid flow and sediment transport processes must be included. There are two categories of basin-fill models: process-based and geometric. Process-based models incorporate the physics of fluid dynamics and sedimentation mechanics to predict sediment erosion, transport, and deposition. In contrast, geometric models are based on algebraic constraints and empirical rules developed from conceptual depositional (facies) models. These models help assess the effects of multiple geologic, hydrologic, and climatic processes on observed basin-fill patterns. Basin-fill models must be linked to source-area and basin-formation models. Source-area models serve to estimate the influx of fluid and sediment into a basin. Basin-formation models predict the basin geometry and subsidence, and assist in predicting the topography.

4.3.1 Process-Based Models Process-based basin-fill models use a combination of governing equations for conservation of mass, momentum, and energy for fluid; conservation of mass for sediment; relations for fluid and sediment rheological properties; empirical relations for sediment transport and entrainment; and relations between rates of erosion or deposition and the size distribution of sediments transported. The equations predict how water, wind, and mass movements transport sediment into and throughout basins and how sediments are distributed during deposition.

A complete account of unsteady flow in three dimensions for an isotropic, Newtonian fluid is provided by the mass-conservation and the Navier–Stokes force-momentum equations (Bird *et al.*, 1960). Conservation of fluid mass mathematically describes the rate of change of density at a fixed point resulting from changes in the mass velocity vector $\rho\mathbf{v}$:

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \rho\mathbf{v}. \tag{5a}$$

If the fluid is assumed to be incompressible,

$$\nabla \cdot \mathbf{v} = 0, \tag{5b}$$

where t is time, $\mathbf{v}$ is the fluid-flow velocity vector (three components), and ρ is the flow density.

Conservation of fluid momentum mathematically describes how the fluid accelerates or decelerates. The motion of the fluid changes because of the forces (pressure, viscous, and gravitational) acting on it:

$$\begin{aligned}\frac{\partial(\rho\mathbf{v})}{\partial t} = {} & -(\mathbf{v}\cdot\nabla)\rho\mathbf{v} - \nabla p - [\nabla\cdot\boldsymbol{\tau}] \\ & + \mu\nabla^2\mathbf{v} + \rho\mathbf{g},\end{aligned} \tag{6}$$

where $\mathbf{g}$ is gravitational acceleration, p is pressure, $\boldsymbol{\tau}$ is the Navier–Stokes stress tensor, a function of flow viscosity and flow velocity gradients (nine components), and μ is the flow viscosity.

The equation for conservation of energy

must be included when vertical flow-velocity variations, such as flow turbulence, are important. Conservation of fluid energy mathematically describes how the accumulation of internal and kinetic energy changes because of changes in the flow rate, heat conduction, and work done on the fluid by forces (gravitational, pressure, and viscous) acting on it:

$$\begin{aligned}\frac{\partial}{\partial t}\left(\tfrac{1}{2}\rho \mathbf{v}^2\right) = {} & -(\mathbf{v}\cdot\nabla)\tfrac{1}{2}\rho\mathbf{v}^2 + \rho(\mathbf{v}\cdot\mathbf{g}) - \nabla\cdot p\mathbf{v} \\ & - p(-\nabla\cdot\mathbf{v}) - \nabla\cdot[\boldsymbol{\tau}\cdot\mathbf{v}] \\ & - (-\boldsymbol{\tau}:\nabla\mathbf{v}),\end{aligned} \tag{7}$$

where all variables are as previously defined.

The Navier–Stokes equations as shown are suitable for subaerial flows. In contrast, subaqueous flows require consideration of flow density, buoyancy, and additional frictional forces. First, the density difference between sediment-laden fresh water and sea water caused by differences in sediment concentration, salinity, and temperature must be considered. This density difference is accounted for by multiplying gravity by the relative density of the flow with respect to the surrounding fluid, $(\rho_{ws} - \rho_{\text{sea}})/\rho_{ws}$, where ρ_{ws} is the density of the water–sediment mixture, and ρ_{sea} is the density of the surrounding fluid (sea water or lake water). The density difference affects where the flow is located, the location beneath sea level. For flows that float at sea level, friction occurs at the freshwater–seawater interface at the bottom of the flow. In contrast, for submarine flows denser than sea water, friction occurs between the flow and the sea floor as well as at the freshwater–seawater interface at the top of the flow. Second, entrainment of sea water into the sediment-laden flow by turbulent mixing must be considered. Additional important considerations below sea level include the transport of sediment by longshore currents and ocean circulation patterns. Martinez and Harbaugh (1993) provide examples of simulation of longshore transport and sediment transport by wind-driven waves.

Solving the Navier–Stokes equations in three dimensions is computationally intractable for the large spatial and temporal domains of sedimentary basins. However, with a few simplifying assumptions, realistic large-scale problems have been addressed (Tetzlaff and Harbaugh, 1989; Koltermann and Gorelick, 1992). Typically, all flow variables are vertically integrated to eliminate vertical flow-velocity variations and internal shear within the flow. Assuming a homogeneous, incompressible, isothermal fluid allows fluid density and viscosity to be treated as constant for fluid gravity flows. Coriolis forces are neglected unless the effects of ocean currents must be simulated.

Sediment-transport models require an equation of conservation of sediment mass as well as equations relating the fluid mechanics to the ability of the flow to transport sediment, whether the transporting medium is wind or water. Allen (1985) and Blatt *et al.* (1980) provide discussions of sediment transport. Conservation of sediment mass mathematically describes topographic changes due to erosion or deposition and the resulting changes in the sediment concentration in the flow:

$$h_f\left(\frac{\partial c}{\partial t} + [\bar{\mathbf{v}}\cdot\nabla c]\right) = -\frac{\partial z_s}{\partial t}(1 - \phi), \tag{8}$$

where ϕ is porosity, h_f is flow depth, z_S is the sediment surface elevation, $\bar{\mathbf{v}}$ is the vector of depth-averaged flow velocity, and c is the sediment concentration. The thickness of sediment deposited or eroded is reflected in the change in the land surface elevation multiplied by the packing concentration $(1 - \phi)$ of the deposited sediment.

Sediment-transport equations developed from field and laboratory studies have been based either on statistical regression fits to data or on theoretical analysis of the forces required to initiate particle motion and *entrain* particles into the flow. The *entrainment* of particles occurs when grains are dislodged from a position of rest. The critical or threshold shear stress required to initiate particle motion and entrain particles into the flow is related to particle size, density, shape, and cohesion, and to the roughness of the surface the grain rests upon (Komar, 1988; Blatt *et al.*, 1980).

Many equations have been proposed to relate the bedload transport capacity of a stream to flow hydraulic characteristics and the physical properties of both fluid and sediment. Reviews are provided by the American Society of Civil Engineers (ASCE, 1976) and Dawdy and Vanoni (1986). These equations typically estimate bedload transport capacity as a function of the difference between the *boundary shear stress* (the shear stress at the interface between the fluid and the stream bottom) and the *critical shear stress:*

$$q_s = K(\tau_0 - \tau_c)^{3/2}, \qquad (9)$$

where q_s is the bedload transport rate, τ_0 is the boundary shear stress [the product of the streambed slope, the stream hydraulic radius (area divided by wetted perimeter), and the fluid specific weight], τ_c is the critical shear stress, and K is an empirically derived function of particle characteristics.

In contrast to bedload, the amount of suspended sediment transported often is supply limited. That is, the capacity of the stream to transport sediment is greater than the supply available in the source area. For a given magnitude of flood, the suspended sediment concentration can vary widely as a result of factors affecting the source area, such as vegetation, precipitation, temperature, parent rock type, and soil moisture conditions. For example, Koltermann and Gorelick (1992) showed that in the coastal mountain ranges of California, streams with similarly sized drainage basins but different climate conditions had suspended sediment concentrations that varied over several orders of magnitude. These kinds of studies show that suspended sediment loads must be measured rather than predicted on the basis of flow hydraulic variables.

In all cases, the particle settling velocities are critical. Settling velocities are greatest in clear water and decrease with increasing sediment concentration because particle collisions and interactions hinder particle settling (Mirza and Richardson, 1979). In particular, the presence of a high concentration of fine-grained particles impedes the settling of coarser particles.

In fluid gravity flows, the spatial patterns of different grain sizes throughout a sedimentary basin is partially controlled by *downstream fining,* the processes that cause the average grain size within a stream channel to diminish with distance from the source area. Two mechanisms control downstream fining. First, particle abrasion breaks bedload particles down into smaller sizes as they are bounced, slid, and rolled along a stream bottom. Second, selective deposition of heavier and larger grains with higher settling velocities results in preferential deposition of these grains closer to the source area. The division of sediment into different grain-size populations during transport or deposition is called *fractionation.* To date, in basin-fill modeling particle abrasion is neglected, while selective deposition is enforced by a hierarchy of grain-size classes in order of decreasing settling velocities.

Examples of fluid–gravity-flow and sediment-transport models used to make predictions about spatial patterns in sedimentation are not numerous but do exist. To predict spatial variations in mean grain size and sorting in one bend of a meandering stream, Bridge (1977, 1992) developed a model of nonuniform stream flow, bedload transport, and changes in bed geometry with erosion and deposition (Fig. 11). Simulation results reproduce the trends in grain size observed in a natural channel. Bitzer and Pflug (1990) simulated steady-state flow and suspended-sediment transport in small ocean basins. An example simulation shows the geometry and distribution of grain sizes in terms of percentages of gravel, sand, and shale predicted from flow through a harbor (Fig. 12). Fluid-flow and sediment-transport models for purposes of geologic predictions can get quite complicated. For example, Ericksen *et al.* (1990) simulated wind- and tide-driven circulation, sedimentation, and erosion in oceanic basins by accounting for continuity of fluid, salt, heat, and sediment as well as sediment transport and density-dependent fluid flow. They applied their model to the late Devonian Catskill Sea to estimate ancient wind and tide conditions.

To date, most process-based basin-fill modeling has been of fluid gravity flows because modeling of sediment gravity flows is more complex and difficult. As a result of the dependence of the flow hydraulic characteristics on the flow rheology, the governing equations for fluid and sediment in sediment gravity flows must be solved simultaneously at each time step, rather than sequentially. In addition, the equation for conservation of energy must be included to represent turbulence. This type of modeling requires consideration of sedimentation from highly concentrated multicomponent suspensions, and prediction of sediment-concentration–dependent fluid viscosity, density, and particle set-

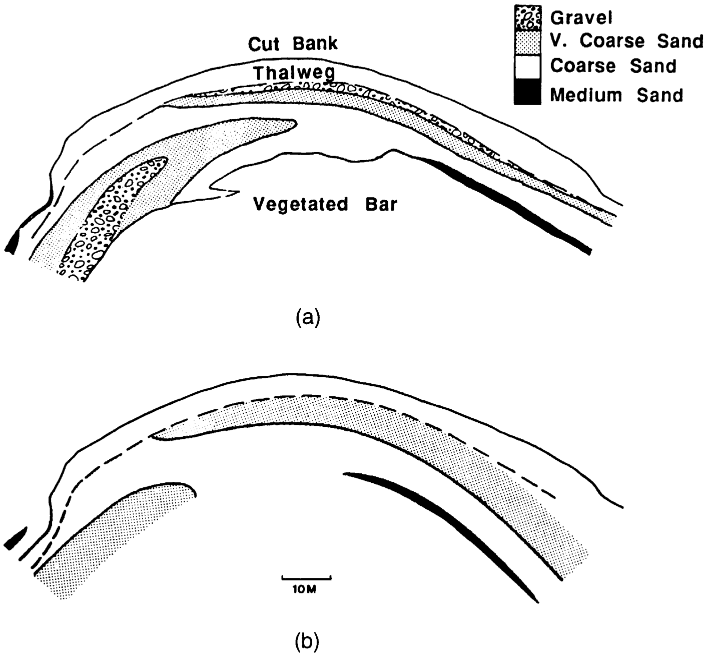

FIG. 11. Process-based image of a meander (stream channel bend). The predicted grain size (lower figure) compares favorably to the observed grain size (upper figure) in a natural channel. After Bridge (1977). Reprinted by permission of John Wiley and Sons, Ltd.

tling velocities (Zeng and Lowe, 1992). In sediment gravity flows such as turbidity currents, the bulk of the coarse sediment is transported within the flow rather than as bedload, sediment size fractionation is affected by particle interactions and fluid displacements by falling grains, deposition is dominated by deposition from suspended load rather than bedload, and sediment size fractionation occurs not during transport but during deposition. Unlike fluid gravity flows in which water entrains sediment particles, the suspended load in a turbidity current results from entrainment of water into a failed sediment mass. Therefore, the sediment concentration in the flow and the volume of the flow are highly variable over short time periods and distances. Modeling of subaqueous sediment gravity flows and density currents has been performed primarily to analyze

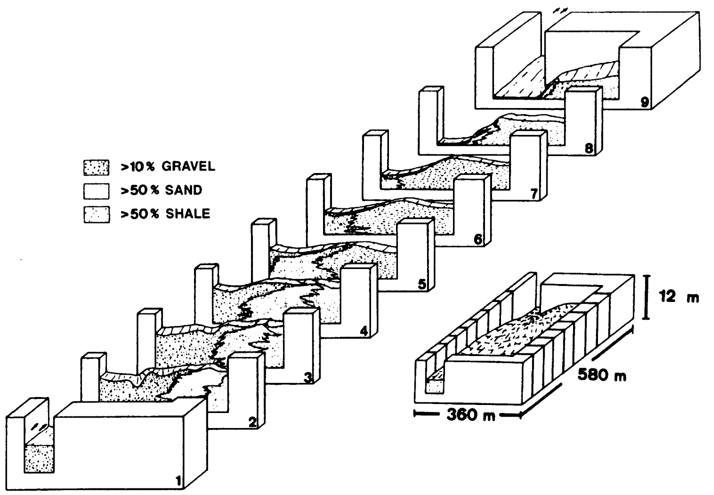

FIG. 12. Process-based image of harbor sedimentation. The geometry and distribution of grain sizes in terms of percentages of gravel, sand, and shale predicted from flow through a harbor is shown. Flow is from lower left to upper right. Downstream fining is evident in the distribution of grain sizes in the simulated harbor, where coarser grain sizes are deposited closer to the harbor inlet, and finer grain sizes are deposited closer to the outlet. The inset shows the size and geometry of the harbor. After Bitzer and Pflug (1990). Reprinted by permission of Prentice Hall, Upper Saddle River, New Jersey.

highly erosive turbidity currents in submarine canyons (Fukushima *et al.*, 1985), rather than to examine sedimentation patterns, although research is progressing in the latter area.

The other important sediment–gravity-flow category is debris flows, the physics of which differs from turbidity currents. Because a debris flow has strength that must be overcome, a debris flow starts moving only after the critical shear stress required to initiate motion is exceeded (Johnson, 1970). The strength of a debris flow is so great that transport of even boulder-size rock fragments occurs. A debris flow remains in motion until the shear stress at the base of the flow drops below the critical shear stress, at which time the entire mass ceases to move. The base of the debris flow deforms as it flows, while the top flows as a rigid mass. Debris-flow sediments show that the upper parts of a debris flow contain coarser materials dispersed in a finer-grained matrix, without the grain-size sorting, stratification, and small sedimentary features produced by flowing fluid or sediment seen in fluvial flows and turbidity currents. Although grain-size variations within debris-flow deposits occur, they are not important to fluid-flow properties such as permeability, which is controlled by the fine-grained matrix. Often, the predicted location and geometry of debris-flow deposits are of greatest concern. Therefore, they have been numerically modeled using a Coulomb continuum approach (Savage and Hutter, 1989), in which a debris flow is represented as one undifferentiated mass that changes shape as it flows from the initiation of movement to its final location. Results of simulations compare favorably to both wet and dry experimental debris flows (Iverson and LaHusen, 1992).

Models of fluid gravity flows have been used in process-based basin-fill models that include external forcing factors, links to a source-area model, and a model of basin formation. The first large-scale process-based model to represent basin fill in three dimensions was that of Tetzlaff and Harbaugh (1989). Their model predicts three-dimensional grain-size variations at a large scale produced in fluvial and deltaic environments. Figure 13 shows a plan and cross-sectional view from a simulation of braided-stream deposits. The general form of the model has lent it to simulation of alluvial fans, deltas, and braided-stream deposits. Realistic sedimentary sequences have been produced for the Golden Oil field, Louisiana, the National Petroleum Reserve, Alaska, and the Ivishak formation, Alaska. They began with a simple model of tectonic subsidence. More recent work has included sediment transport by waves (Martinez and Harbaugh, 1993) and links to simulation of hydrocarbon migration in sedimentary deposits produced by the basin-fill model (Wendebourg and Harbaugh, in press).

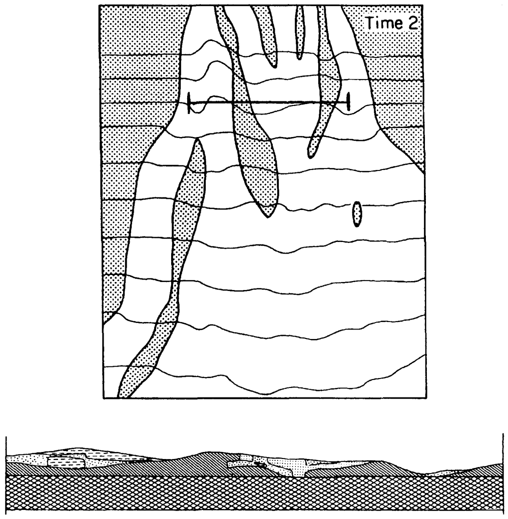

FIG. 13. Process-based image of braided-stream deposits. Areal and cross-sectional views of simulation results from four source streams flowing over a planar surface. The thick horizontal bar located 1/4 of the way down from the top of the areal view shows the location of the cross-sectional view. Lines crossing the figure from left to right are elevation contour lines. Coarser deposits are stippled. In the cross-sectional view, bedrock is shown as cross-hatched, while changes in stippling indicate sediments deposited at successively later times. Modified from Tetzlaff and Harbaugh (1989). Reprinted by permission of Van Nostrand Reinhold publishers.

Koltermann and Gorelick (1992) extended the model of Tetzlaff and Harbaugh (1989) to represent sea-level change, horizontal and vertical fault motion, sediment loading, numerous tectonic subsidence mechanisms, compaction, paleoclimate-driven fluctuations in floods and sediment loads, and realistic sequences of ancient floods. The model was adapted to a supercomputer to simulate sedimentary basin evolution over geologic time scales (hundreds of thousands to millions of years). Simulation of the geologic history of the Alameda Creek alluvial fan in north-central California showed how paleoclimate fluctuations create large-scale patterns in sedimentation by affecting ancient flood magnitudes and the volume and composition of sediment loads transported into a basin. These large-scale patterns produced highly permeable gravel layers separated by thick, low-permeability clay layers (Fig. 14). Results show that basin evolution models can be used to distinguish between large-scale spatial patterns in sedimentary basin fill caused by tectonics, sea-level change, and climate change. The grain-size distributions were used to develop a realistic model of porosity and permeability variations (Koltermann and Gorelick, 1995) and were linked to a groundwater flow simulator (Koltermann, 1993) to illustrate how sedimentary-basin evolution models can be useful predictive tools for subsurface fluid flow.

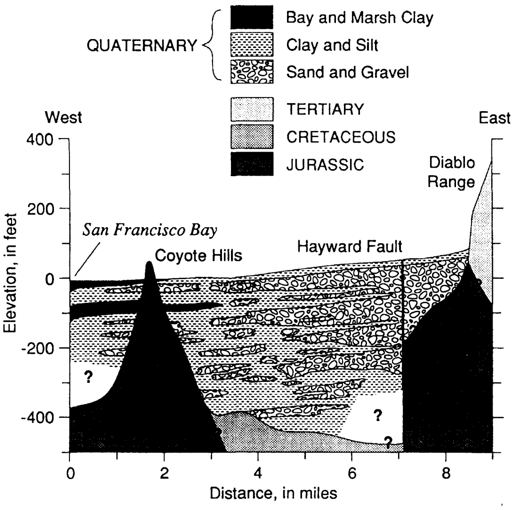

FIG. 14. Alluvial-fan cross section showing cycles of sedimentation. Simulation of the geologic history of the Alameda Creek alluvial fan in north-central California showed how paleoclimate fluctuations create large-scale patterns in coarse, high-permeability gravels and sands versus fine, low-permeability silts and clays by affecting ancient flood magnitudes and the volume and composition of sediment loads transported into a basin. Results show that coupling of a source-area model, a basin-formation model, and a basin-fill model helps geologists distinguish between spatial patterns in sedimentary basin fill due to tectonics, sea-level change, and climate change.

4.3.2 Geometric Models Geometric models predict large-scale grain-size patterns in sedimentary basin fill using rules consistent with observed arrangements of depositional environments and geologic laws, such as the *principle of superposition,* which states that younger sediments are successively deposited on top of older sediments and rock. The rules constrain patterns in sediment grain size, the location of marine versus nonmarine sediments, the location of carbonate minerals, and the slope of the land surface and sea floor. Some geometric methods produce results at fine spatial detail, corresponding to the width of braided channels, whereas others produce coarser results, corresponding to the large-scale sediment geometries resulting from the rise and fall of sea level.

Geometric modeling methods include stratigraphy, analytical, random-walk, and random-avulsion (stream-channel migration) approaches. The most useful and widely used of the geometric modeling methods are stratigraphy models. These models produce cross-sectional images of large-scale lithologic sequences that show depositional environments, grain-size distributions, ages of deposition, and *unconformities,* which are gaps in the rock record resulting from a cessation of deposition or from erosion (Fig. 15). *Stratigraphy* models are commonly applied in shallow marine settings to decipher the effects of subsidence mechanisms, sediment-supply fluctuations, basin geometry, and sea-level changes on the geologic record. These models have received a lot of attention in the petroleum industry and have been used to examine the geometry, extent, and continuity of sand bodies in petroleum reservoirs. Stratigraphy models include an extensive set of rules to achieve a realistic result. These rules

1. constrain the flow direction to be from one side of a cross section (a vertical image) to the other, forcing sediment to move downhill;
2. control topographic gradients;
3. control where sedimentation and erosion occur, often as a function of the location of the shoreline;
4. use downstream fining rules to deposit coarse sediment near the source area and progressively finer sediment away from the source area;

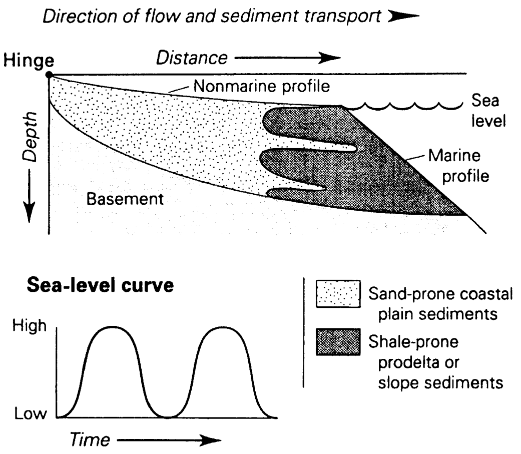
Direction of flow and sediment transport
Hinge
Distance
Depth
Nonmarine profile
Sea level
Marine profile
Basement
Sea-level curve
High
Low
Time
Sand-prone coastal plain sediments
Shale-prone prodelta or slope sediments

FIG. 15. Stratigraphy model image. Stratigraphy predicted for steady subsidence, two cycles of a sinusoidal fluctuation in sea level, and a constant sediment supply. Results show that sea-level fluctuations produce cyclicity in coarse-grained deposits, with high sea-level stands trapping coarse materials close to the shoreline. Modified from Jervey (1988). Reprinted by permission of the Society for Sedimentary Geology.

5. control the amount of sediment deposited in any given location during an interval of time, usually as a simple function of the distance from the shoreline; and
6. consider the production of carbonate minerals as empirical functions of water depth.

Results produced by these models look like geologic interpretations from seismic sections.

The other geometric methods are of less importance in terms of practical applications. *Analytical* models predict the lateral extent and thickness of high-permeability gravels given a basin size, subsidence rate, and sediment influx. Paola *et al.* (1992) developed an analytical model to examine the effects of four external factors on gravel geometries: sediment influx, subsidence, gravel fraction in the sediment supply, and water flux (Fig. 16). Their work showed that a variety of mechanisms produce similar-looking cycles in sedimentary deposits forming groundwater aquifers and petroleum reservoirs.

Random-walk models predict braided-stream channel network configurations by allowing particles of fluid to travel over a to-

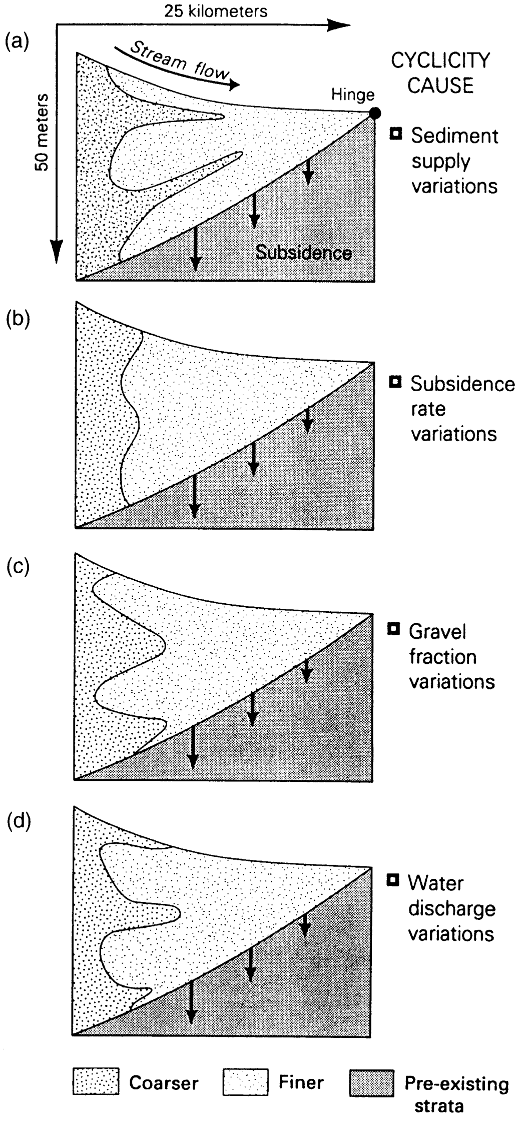
(a)
25 kilometers
50 meters
Stream flow
Hinge
Subsidence
CYCLICITY CAUSE
Sediment supply variations
(b)
Subsidence rate variations
(c)
Gravel fraction variations
(d)
Water discharge variations
Coarser
Finer
Pre-existing strata

FIG. 16. Analytical images of cyclicity mechanisms. Hypothetical cross sections of a basin showing the effects of rapid variation in (a) sediment supply, (b) tectonic subsidence, (c) gravel fraction in the sediment supply, and (d) water flux. Each of the four cyclicity mechanisms is varied independently in a sinusoidal fashion. Variations follow a sinusoidal curve through time. Results show that multiple mechanisms can be responsible for observed cyclicity in alluvial basins. Modified from Paola *et al.* (1992). Reprinted by permission of Blackwell Science, Ltd.

pography from a source area to a discharge point using probabilistic rules for the location of fluid particles through time (Webb, 1995). Given hydraulic relations between flow discharge, velocity, and channel cross-sectional area, a channel topology can be predicted. The type of sediment deposited throughout the channels can be predicted by applying rules for sediment transport given the flow hydraulics.

Random-avulsion models mimic the abrupt change of location (avulsion) of a meandering stream channel to the lowest topographic elevation within a subsiding valley, using probabilistic rules to govern when avulsion occurs. Images show the large-scale cross-sectional architecture of channel sands encased in clays (Fig. 17). Random-avulsion models have been used to examine the distribution of sand bodies deposited by meandering streams as functions of subsidence rate and the proportions of coarse- versus fine-grained sediment transported by streams (Bridge and Leeder, 1979).

5. DIAGENESIS

Important changes occur after sediments are deposited. As loose, unconsolidated sediments are buried, they are subject to increasing pressure and temperatures, bathed by fluids, subject to chemical reactions, compacted, and finally *lithified,* or turned into rock. As these processes occur, the sediments experience a tighter packing, an increase in density, and a decrease in both porosity and permeability. The post-depositional processes causing chemical/mineralogical, biological, and physical changes to sediments and rocks are collectively referred to as *diagenesis.* Readers are referred to Blatt *et al.* (1980) for a discussion of how diagenesis affects different minerals and to Einsele (1992) for a chapter on diagenesis and fluid flow, thermal histories, and hydrocarbon generation. Diagenetic processes cause compaction and expulsion of pore fluids, precipitation of mineral cements and concretions, growth of minerals such as clays in pore spaces, replacement of primary minerals by secondary minerals, reactions (such as dissolution) of minerals with pore fluids, recrystallization and formation of new minerals, incorporation of water into a mineral or release of water from a mineral, and breakdown of minerals due to the action of bacteria. Chemical processes cause soluble or thermodynamically unstable minerals to dissolve and be reprecipitated or recrystallize as more stable minerals. Diagenesis can overprint original depositional features, such as fossils and small-scale primary sedimentary structures such as cross bedding, as well as larger features such as cycles of sedimentation. Through time, diagenetic processes tend to reduce drastically the original porosity and permeability of sediments. However, fracturing of rock can increase the porosity and drastically increase the permeability. This is referred to as *secondary* porosity and permeability.

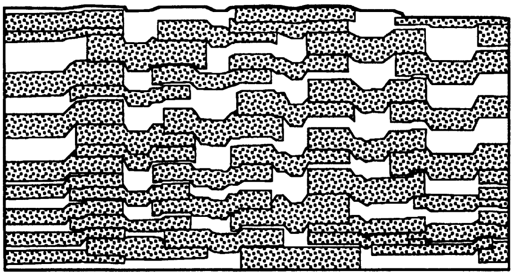

FIG. 17. Random-avulsion image of a meander belt. Cross section from a random-avulsion simulation in which the channel-belt sands (stippled) are encased in fine-grained flood-plain deposits. Compaction produces the distortion of the originally horizontal channel sands. Modified from Bridge and Leeder (1979). Reprinted by permission of Blackwell Science, Ltd.

Near the land surface, an important post-depositional change is the mechanical and chemical weathering of surface sediments and rock and the formation of soil horizons, complete with roots, biological activity, and breakdown of sediment grains into clay minerals. This part of the sedimentary rock cycle produces weathered materials for erosion and sediment transport.

Sediments of chemical and biological origin are particularly susceptible to diagenesis. Carbonate and siliceous minerals experience dramatic changes after deposition. Minerals dissolve, creating additional pore space. New carbonate minerals precipitate, forming cement in the voids of existing rocks. The originally deposited carbonate minerals can be replaced by other, more stable, carbonate minerals, such as the commonly observed calcite (calcium carbonate) and dolomite (calcium magnesium carbonate). Sediments of biological origin rich in silica are highly soluble. Dissolution of one siliceous mineral and reprecipitation as another siliceous mineral is common. Finally, as the temperature and pressure increase with depth of burial, silicon oxide is recrystallized to become the familiar mineral *quartz.*

Geologists tend to distinguish between diagenetic processes that occur when sediments have not been buried very deeply, and processes that occur after deep burial when temperatures have substantially increased. Changes in the temperature of basin sediments cause diagenetic changes because where sediments of organic origin are present, the thermal history of a basin exerts control on the formation of hydrocarbons and coal. This thermal history is recorded by temperature-sensitive organic and inorganic indicators within the sediments. Thermal histories record increasing depth of burial as well as uplift and erosion (Fig. 18). Recall that in basins formed by extension and rifting, high-temperature mantle material is brought closer to the overlying crust, causing increased temperatures in crustal rocks that then cool slowly. These higher temperatures affect the sediments in terms of diagenetic processes. Deep burial also is important to the diagenetic processes that produce ore deposits. Deep subsurface fluids can be of sufficiently high salinity and metal concentrations to precipitate ore minerals (Einsele, 1992). It is important to note that as the temperature increases beyond about 200 °C, diagenesis gives way to low-grade metamorphism (Einsele, 1992).

Modeling of diagenesis includes representation of the chemistry of subsurface waters in contact with sediments and rock, as well as the flow of subsurface fluids and transport of solutes. Freeze and Cherry (1979) and Anderson and Woessner (1992) provide texts on groundwater flow and modeling. Wendebourg and Harbaugh (in press) treat modeling of hydrocarbon migration in sedimentary basins. Readers are referred to Bethke *et al.* (1990) for examples of modeling basin thermal histories, fluid flow, and diagenesis on a large scale in sedimentary basins. Models of diagenesis have not yet been coupled with models of sedimentary basin formation and filling. The only diagenetic process currently commonly represented in sedimentary basin formation and filling models is compaction.

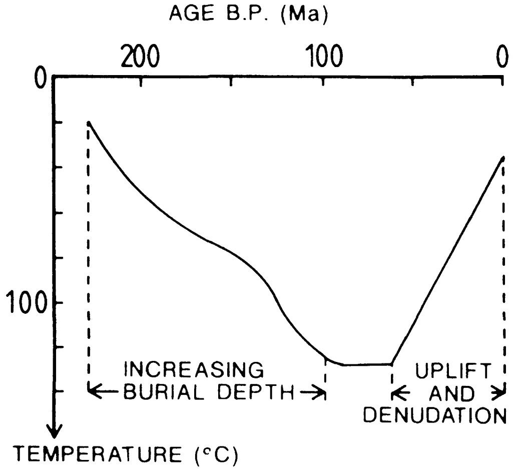

FIG. 18. Thermal history diagram. At this location in the Colorado Plateau, the thermal history, as reflected by temperature-sensitive indicators in the sediments, records an increasing depth of burial until 100 million years (Ma) before present (B.P.), as well as more recent uplift and denudation. After Meyer *et al.* (1989). Reprinted by permission of Springer-Verlag New York, Inc.

6. GEOLOGIC HISTORY ANALYSIS

We have both direct and indirect views of the Earth's subsurface that provide data used in the analysis of a sedimentary basin's history. Direct measurements of rock or sediment properties can be obtained by "field" observation of outcrops, such as in valleys, cliffs, and road cuts, mapping, and the drilling of wells into the sedimentary basin fill. Specific information can include lithologic logs, soil borings, cores, paleontologic information, age data, and well-pressure test data. Indirect measurements include geophysical data, such as seismic profiles and borehole geophysical logs. Seismic data provide images of large-scale geometric relations in sediments, whereas borehole data produce small-scale information on petrophysical properties used to infer lithology. All of the data are combined to produce interpretive cross sections and maps of a basin.

All available data must be compiled and analyzed through techniques of sedimentary basin analysis, which examine the structural, climatic, and sea-level history of a sedimentary basin (Miall, 1984). Through such analysis one can understand the basin tectonic style and geometry; estimate the rates of subsidence, sedimentation, and sea-level change; determine whether subsidence is accelerating or decelerating; estimate the climatic conditions under which the basin formed; and develop facies relations. Several authors provide a list of questions to guide a comprehensive analysis of the geologic history of a basin (Table 2). Basin evolution models require the following information from a geologic history analysis:

1. the paleogeography, including the timing of basin formation, relation to plate boundary and margin type, basin type, the subsidence history, location and motion on faults, reconstruction of horizontal and vertical land-block movements, and the topography through time;
2. the paleobathymetry, including the history of sea-level change, shoreline location through time, and paleocurrents indicating past flow directions;
3. the paleohydrology, including a magnitude–frequency relation for paleofloods, a relation for flood velocity versus discharge, and a relation for the effects of paleoclimate change on paleohydrology; and
4. the paleosediment discharge, including a relation for the amount of bedload and suspended load carried by each flood, a function of climate and hydrology as well as the parent rock type, and the grain-size distribution in the source sediment supply.

The quantity and quality of the geologic history data base is limited by the infrequent or recent nature of our observations (such as sparse well data and stream-flow records that seldom have more than 100 years of record) and difficulties in age dating. The age of rocks and sediments is established on the basis of radiometric dating techniques, dated periodic reversals in the Earth's mag-

Table 2. Questions about the geologic and climatic history of a basin (Modified from Miall, 1984; Pettijohn *et al.*, 1972; Reading, 1986).

1. Basin age:
   - How old is the basin under study?
   - What does this tell us about the regional history of subsidence and uplift?
2. Tectonism:
   - What type of basin is under consideration, in terms of plate-tectonic theory?
   - Where within the basin are the rocks or sediments (aquifer or reservoir) of interest?
   - What was the tectonic style of the basin during sedimentation?
   - Does this information or fossil evidence or radiometric age dates suggest the length of time the basin was active and its rate of sediment accumulation?
   - Are there gaps in the rock record, or folds or faults that occurred during deposition that provide evidence of variations in the rate of basin subsidence?
   - Are there widespread lithologic units or cyclic sequences that could have been caused by episodic basin subsidence or episodic source area uplift or climatic changes?
3. Climate:
   - What were the global and regional climate conditions during basin formation and filling?
   - Is there any internal evidence of climate change (abundance or type of flora and fauna, diagnostic minerals, or presence of glacial features)?
   - What does this evidence tell us about probable controls on river discharge (perennial or ephemeral flow), bank cohesiveness (vegetation cover), or grain size (chemical vs mechanical weathering)?
   - Are there local variations in climate within the sedimentary basin or between the sedimentary basin and source area?
4. Regional depositional trends:
   - What are the main facies variations in the basin, and over what scale do the variations occur, both vertically and laterally?
   - What do the spatial patterns in regional depositional trends look like?
   - What are the patterns of deposition above sea level versus below sea level?
5. Local depositional trends:
   - What is the thickness range, lateral variability, and grain-size composition of cycles in the sediments?
   - Can a pattern or sequence of stream channel migration or avulsion be detected?
   - Can specific morphological units be identified?
   - What small-scale sedimentary structures are observed?

netic field, and fossil animal and plant remains. Age dating is particularly difficult in terrestrial sediments (land-based sediments deposited above sea or lake water), where the fossil content may be negligible and where volcanic ash beds may be few and far between. Given some age dates in a basin, the ages of the remaining strata can be estimated relative to the dated strata. Climate data come from cores of sediments from the ocean floor, from modern and ancient lakes, and from ice caps. These data provide records of the isotopic composition of sea water, sedimentation rates, the magnitude and frequency of fluctuations in global ice volume, and records of pollen and plant assemblage fluctuations associated with changes in precipitation and temperature.

7. CONCLUSIONS

Because the spatial and temporal scales of sedimentary basin evolution are far beyond the scales we can examine in the laboratory or within a human lifespan, earth scientists have turned to numerical models to examine indirectly and make predictions about that which we cannot observe directly. In recent years, efforts have been made to quantify the processes through which sedimentary basins evolve, and to integrate descriptive geologic studies with mathematical models of source areas, basin formation and filling mechanisms, and post-depositional diagenetic changes. Such modeling studies have three objectives:

1. to increase our understanding and test our hypotheses about the Earth;
2. to predict the effect of geologic processes for engineering or environmental purposes; and
3. to provide information for resource management decisions, such as three-dimensional quantitative descriptions of porosity and permeability fields used in computer models of subsurface fluid flow to predict well locations for hydrocarbon extraction or groundwater contamination cleanup.

Basin evolution modeling provides a quantitative means to evaluate the effects of basin geometry, factors governing basin formation, source-area fluid and sediment supply histories, and sea-level and climatic histories on the architecture of sedimentary deposits. This approach provides a rigorous framework to help interpret the relative effects of competing forcing factors affecting basin fill and improves our ability to interpret the geologic record and make predictions about basin genesis and subsurface fluid flow. Models allow us to study geologic processes in areas that are relatively inaccessible, such as in submarine canyons, and over extremely long time periods.

Basin evolution studies must address four major components. First, models of the source area predict the time sequence and volume of fluid and sediment transported into the basin. Second, models of basin formation predict the tectonic and loading controls on basin geometries and rates of subsidence. Third, models of basin filling predict the grain-size distribution of the sediments deposited within a basin, the location and type of sediments of chemical or biological origin, and the distribution of depositional environments, particularly related to ocean water depths. Fourth, models of diagenesis predict post-depositional changes. Much research has been devoted to studying and modeling each of these components in isolation rather than as part of an entire system. For example, many geodynamic models have been constructed to examine tectonic subsidence alone and neglected basin filling and the additional subsidence caused by sediment loading. However, quantitative basin evolution modeling is in its infancy because earth scientists have only just begun to merge descriptive geologic studies with available quantitative tools. This approach to model simultaneously all components of basin evolution appears destined to become more common as issues involving complex and interrelated tectonic, sedimentologic, hydrologic, climatic, and diagenetic histories are increasingly addressed.

GLOSSARY

Avulsion: The abrupt switching in location of a stream channel, usually during flooding.

Clastic sediment: Particles eroded from a previously existing geologic formation.

Convergence: The motion of two lithospheric plates toward each other to join together at a common boundary.

Denudation: The combination of natural processes that result in the progressive lowering of the Earth's surface by weathering, erosion, mass wasting, and transport by wind and water.

Depositional environment: The set of physical (especially hydrodynamic), chemical, and biological conditions, and depositional processes, occurring in a given location to produce a characteristic sedimentary deposit.

Diagenesis: The physical, chemical, and biological changes to sediments after deposition, including precipitation of cements from circulating subsurface fluids in fractures and pore spaces, and compaction of sediment grains into a tighter packing arrangement.

Extension: The stretching or pulling apart of the Earth's crust that causes the crust to thin and fracture.

Facies: Geologic units of common characteristics that can be distinguished by the "sum of all the primary characteristics of a sedimentary unit" produced as a result of deposition under given environmental conditions.

Flexural Subsidence: Depression of the Earth's surface adjacent to loads of rock and sediment rather than just beneath such loads. It is a consequence of the fact that the Earth's crust has a strength and can propagate loads.

Geodynamic Models: Models of basin formation.

Geologic History Analysis: A collection of techniques that use local and regional geologic and climatic data to constrain the structural, sea level, climatic, and hydrologic histories of a sedimentary basin.

Isostatic Compensation: The vertical adjustment of the lithosphere to maintain an equilibrium among rocks, sediments, and water as loads are either added to or subtracted from the lithosphere.

Lithification: The processes by which recently deposited, loose sediments are transformed into rock.

Lithosphere: A layer of strength from the surface of the Earth downward. It includes the crust and part of the upper mantle and is approximately 100 km thick.

Paleowater Depths: The depths of ancient seas at an earlier point in time.

Rifting: The large-scale faulting and tilting of the Earth's crust occurring after extension of the crust.

Source Area: A feature of positive relief that constitutes the drainage area for fluid and sediment transported into a basin.

Subduction: The process by which one lithospheric plate descends beneath another.

Subsidence: The sinking or vertical downward displacement of the Earth's surface.

Tectonic Loading: Depression of the lithosphere adjacent to mountains formed by crustal compression.

Uplift: The positive vertical motion of the Earth's surface by tectonic forces or isostatic compensation.

Upwelling: An upward vertical movement of upper mantle material.

Weathering: The group of chemical and physical processes that cause decomposition of geologic materials *in situ* into smaller particles, individual minerals, and dissolved constituents, making materials available for transport by wind and water.

Works Cited

Allen, J. R. L. (1965), "A review of the origin and characteristics of recent alluvial sediments," *Sedimentology* **5,** 89–191.

Allen, J. R. L. (1985), *Principles of Physical Sedimentology,* London: George Allen and Unwin, 272 pp.

American Society of Civil Engineers (ASCE) (1976), in V. A. Vanoni (Ed.), *Sedimentation Engineering,* Manuals and Reports on Engineering Practice No. 54, New York: American Society of Civil Engineers, 745 pp.

Anderson, M. P., Woessner, W. (1992), *Applied Groundwater Modeling. Simulation of Flow and Advective Transport,* San Diego: Academic Press, 381 pp.

Beaumont, C., Fullsack, P., Hamilton, J. (1991), "Erosional Control of Active Compressional Orogens," in: K. R. McClay (Ed.), *Thrust Tectonics,* New York: Chapman and Hall, 447 pp.

Bethke, C. M., Harrison, W. J., Upson, C., Altaner, S. P. (1990), "Supercomputer Analysis of Sedimentary Basins," *Science* **239,** 261–267.

Bird, R. B., Stewart, W. E., Lightfoot, E. N. (1960), *Transport Phenomena,* Chichester: Wiley and Sons, 780 pp.

Bitzer, K., Pflug, R. (1990), "DEPO3D: A Three-Dimensional Model for Simulating Clastic Sedimentation and Isostatic Compensation in Sedimentary Basins," in: T. A. Cross (Ed.), *Quantitative Dynamic Stratigraphy,* Englewood Cliffs, NJ: Prentice Hall, pp. 335–348.

Blatt, H., Middleton, G., Murray, R. (1980), *Origin of Sedimentary Rocks,* Englewood Cliffs, NJ: Prentice Hall, 782 pp.

Bridge, J. S. (1977), "Flow, Bed Topography, Grain Size and Sedimentary Structure in Open Channel Bends: A Three-Dimensional Model," *Earth Surf. Proc.* **2,** 401–416.

Bridge, J. S. (1992), "A Revised Model for Water Flow, Sediment Transport, Bed Topography, and Grain Size Sorting in Natural River Bends," *Water Resources Res.* **28** (4), 999–1013.

Bridge, J. S., Leeder, M. R. (1979), "A Simulation Model of Alluvial Stratigraphy," *Sedimentology* **26,** 617–644.

Busby, C. J., Ingersoll, R. V. (1995), *Tectonics of Sedimentary Basins,* Cambridge, MA: Blackwell Science, 579 pp.

Coleman, J. M., Wright, L. D. (1975), "Modern River Deltas: Variability of Process and Sand Bodies," in: M. L. Broussard (Ed.), *Deltas, Models for Exploration,* Houston: Houston Geological Society, pp. 99–149.

Curray, J. R. (1969), "Shallow Structure of the Continental Margin," in: D. J. Stanley (Ed.), *The New Concepts of Continental Margin Sedimentation,* JC-12, Washington, DC: American Geological Institute, pp. 1–22.

Dawdy, D. R., Vanoni, V. A. (1986), "Modeling Alluvial Channels," *Water Resources Res.* **22** (9), 71S–81S.

Dickinson, W. R. (1974), "Plate Tectonics and Sedimentation," in: W. R. Dickinson (Ed.), *Tectonics and Sedimentation,* SEPM Special Publication 22, Tulsa, OK: Society of Economic and Paleontologic Mineralogists, p. 1–27.

Dickinson, W. R. (1976), *Plate Tectonic Evolution of Sedimentary Basins,* AAPG Continuing Education Course Notes Series No. 1, Tulsa, OK: American Association of Petroleum Geologists, 62 pp.

Einsele, G. (1992), *Sedimentary Basins,* New York: Springer-Verlag, 628 pp.

Ericksen, M. C., Masson, D. S., Slingerland, R., Swetland, D. W. (1990), "Numerical Simulation of Circulation and Sediment Transport in the Late Devonian Catskill Sea," in: T. A. Cross (Ed.), *Quantitative Dynamic Stratigraphy,* Englewood Cliffs, NJ: Prentice Hall, pp. 293–305.

Freeze, R. A., Cherry, J. A. (1979), *Groundwater,* Englewood Cliffs, NJ: Prentice-Hall, 604 pp.

Fukushima, Y., Parker, G., Pantin, H. M. (1985), "Prediction of Ignitive Turbidity Currents in Scripps Submarine Canyon," *Marine Geol.* **67,** 55–81.

Galloway, W. E., Hobday, D. K. (1983), *Terrigenous Clastic Depositional Systems,* New York: Springer-Verlag, 423 pp.

Gross, M. G. (1980), "Variations in the O^{18}/O^{16} and C^{13}/C^{12} ratios of diagenetically altered limestones in the Bermuda Islands," *J. Geol.* **72,** 170–194.

Ingersoll, R. (1988), "Tectonics of Sedimentary Basins," *Geol. Soc. Am. Bull.* **100,** 1704–1719.

Iverson, R. M., LaHusen, R. G. (1992), "Momentum Transport in Debris Flows: Large-Scale Experiments," *EOS, Trans. Am. Geophys. Union* **73** (43), 227.

Jarvis, G. T., McKenzie, D. P. (1980), "Sedimentary Basin Formation with Finite Extension Rates," *Earth Planet. Sci. Lett.* **48,** 42–52.

Jervey, M. T. (1988), "Quantitative geological modeling of siliciclastic rock sequences and their seismic expression," in: C. K. Wilgus, B. S. Hastings, C. G. Kendall, H. M. Posamentier, C. A. Ross, J. C. Van Wagoner (Eds.) *Sea Level Changes—An Integrated Approach,* SEPM Special Publication 42, Tulsa, OK: Society of Economic and Paleontologic Mineralogists, pp. 47–69.

Johnson, A. (1970), *Physical Processes in Geology,* Geology, San Francisco: Freeman, Cooper, and Co.

Jordan, T. E. (1981), "Thrust Loads and Foreland Basin Evolution, Cretaceous, Western United States," *Am. Assoc. Pet. Geol. Bull.* **65,** 2506–2520.

Kingston, D. R., Dishroon, C. P., Williams, P. A. (1983), "Global basin classification system," *Am. Assoc. Pet. Geol. Bull.* **67,** 2175–2193.

Koltermann, C. E., Gorelick, S. M. (1992), "Paleoclimate Signature in Terrestrial Flood Deposits," *Science* **256,** 1775–1782.

Koltermann, C. E., Gorelick, S. M. (1995), "Fractional packing model for prediction of hydraulic conductivity in sediment mixtures," *Water Resources Res.* **31,** 3283–3297.

Koltermann, C. E. (1993), *Geologic Modeling of Spatial Variability in Sedimentary Environments for Groundwater Flow Simulation,* Ph.D. Dissertation, Stanford University, 314 pp.

Komar, P. D. (1988), "Sediment Transport by Floods," in: V. R. Baker, R. C. Kochel, P. C. Patton (Eds.), *Flood Geomorphology,* New York: Wiley, pp. 97–111.

Martinez, P. A., Harbaugh, J. W. (1993), *Simulating Nearshore Processes,* New York: Pergamon, 265 pp.

McKenzie, D. P. (1978), "Some Remarks on the Development of Sedimentary Basins," *Earth Planet. Sci. Lett.* **40,** 25–32.

Meyer, A. J., Landais, P., Brosse, E., Pagel, M., Carisey, J. C., Krewdl, K. (1989), "Thermal History of the Permian Formations from the Breccia Pipes Area (Grand Canyon Region, Arizona)," *Geol. Rundsch.* **78,** 427–438.

Miall, A. D. (1984), *Principles of Sedimentary Basin Analysis,* Heidelberg: Springer-Verlag, 490 pp.

Mirza, S., Richardson, J. F. (1979), "Sedimentation of Suspension of Particles of Two or More Sizes," *Chem. Eng. Sci.* **34,** 447–454.

Mitchell, A. H. G., Reading, H. G. (1986), Sedimentation and Tectonics, in: H. G. Reading (Ed.), *Sedimentary Environments and Facies,* 2nd ed., Oxford: Blackwell, pp. 471–519.

Norell, M. A., Clark, J. M., Chiappe, L. M., Dashzeveg, D. (1995), *Science* **378,** 774–776.

Paola, C., Heller, P. L., Angevine, C. L. (1992), "The Large-Scale Dynamics of Grain-Size Variation in Alluvial Basins, 1: Theory," *Basin Res.* **4,** 73–90.

Pettijohn, F. J., Potter, P. E., Siever, R. (1972), *Sand and Sandstone,* New York: Springer-Verlag, 618 pp.

Reading, H. G. (Ed.) (1986), *Sedimentary Environments and Facies,* New York: Elsevier, 557 pp.

Reineck, H. E., Singh, I. B. (1980), *Depositional Sedimentary Environments,* Heidelberg: Springer-Verlag, 549 pp.

Savage, S. B., Hutter, K. (1989), "The motion of a finite mass of granular material down a rough incline," *J. Fluid Mech.* **199,** 177–215.

Slingerland, R., Harbaugh, J. W., Furlong, K. (1994), *Simulating Clastic Sedimentary Basins,* Englewood Cliffs, NJ: Prentice Hall, 220 pp.

Tetzlaff, D. M., Harbaugh, J. W. (1989), *Simulating Clastic Sedimentation,* New York: Van Nostrand Reinhold, 202 pp.

Turcotte, D. L., Schubert, G. (1982), *Geodynamics: Applications of Continuum Physics to Geological Problems,* Wiley: New York, 450 pp.

Webb, E. K. (1995), "Simulation of Braided Channel Topology and Topography," *Water Resources Res.* **31,** 2603–2611.

Wendebourg, J., Harbaugh, J. (in press), *Simulating Oil Migration and Accumulation in Stratigraphic Traps,* Oxford: Pergamon, 350 pp.

Willgoose, G. R., Bras, R. L., Rodriguez-Iturbe, I. (1991), "A Coupled Channel Network Growth and Hillslope Evolution Model," *Water Resources Res.* **27** (7), 1671–1702.

Zeng, J., Lowe, D. R. (1992), "A Numerical Model for Sedimentation from Highly-Concentrated Multi-sized Suspensions," *Math. Geol.* **24,** 393–415.

Further Reading

Harbaugh, J. W., Bonham-Carter, G. (1970), *Computer Simulation in Geology,* New York: Wiley, 575 pp.

Kendall, C. G. St. C., Strobel, J., Moore, P., Cannon, R., Bezdek, J., Biswas, G. (1989), "Simulation of the Sedimentary Fill of Basins," in: E. K. Franseen, W. L. Watney (Eds.), *Sedimentary Modeling: Computer Simulation of Depositional Sequences,* Kansas Geological Survey Subsurface Geology Series 12, Lawrence, KS: Kansas Geological Survey, pp. 1–5.

Kleinspehn, K. L., Paola, C. (Eds.) (1988), *New Perspectives in Basin Analysis,* New York: Springer-Verlag.

Lawrence, D. T., Doyle, M., Aigner, T. (1990), "Stratigraphic Simulation of Sedimentary Basins—Concepts and Calibrations," *Am. Assoc. Pet. Geol. Bull.* **74,** 273–295.

Ritter, D. F. (1986), *Process Geomorphology,* New York: Wm. C. Brown, 579 pp.

Simpson, J. E. (1987), *Gravity Currents: In the Environment and the Laboratory,* New York: Wiley, 244 pp.

Walker, R. G. (1984), "General Introduction: Facies, facies sequences and facies models," in: R. G. Walker (Ed.), *Facies Models,* 2nd ed., Geoscience Canada Reprint Series 1, St. John's, Newfoundland: Geological Association of Canada, pp. 1–9.

Wilgus, C. K., Hastings, B. S., Kendall, C. G., Posamentier, H. W., Ross, C. A., Van Wagoner, J. C. (Eds.) (1988), *Sea-Level Changes: An Integrated Approach,* SEPM Special Publication 42, Tulsa, OK: Society of Economic and Paleontologic Mineralogists.

SEISMOGRAPHS

JONATHAN BERGER AND DUNCAN C. AGNEW, *Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California, San Diego, San Diego, California, U.S.A.*

Introduction 139
1. **Seismic Sensors** 140
1.1 Sensor Requirements: Dynamic Range, Sensitivity, Bandwidth 141
1.2 Suspension Designs 142
1.3 Position Sensors 143
1.4 Feedback 143
1.5 Instrumental Noise Sources 144
1.6 Calibration 144
1.7 Examples of Modern Seismometers 145
1.8 Strain Seismometers 146
2. **Recording Systems** 146
2.1 Recording Systems—Historical 146
2.2 Modern Data-Acquisition Systems 147
2.2.1 Analog-to-Digital Conversion 147
2.2.2 Timing 148
2.2.3 Digital Processing 149
2.2.4 Digital Recording 150
2.2.5 Digital Telemetry 150
3. **Networks and Arrays** 150
3.1 Local and Regional Networks 150
3.2 Global Networks 150
3.3 Arrays 151
Glossary 154
Works Cited 154
Further Reading 154

INTRODUCTION

Seismographs are systems for recording the motion of the ground; very similar principles apply to systems for measuring the vibrations of structures. The interest in making such records comes primarily from the study of the waves radiated from a sudden source of stress change, such as an earthquake or explosion. These waves provide information, often the only information, both about the source and about the material properties of the parts of the Earth through which they propagate. Using artificially generated seismic waves to study Earth structure has led to a vast commercial application of seismic methods in the exploration for hydrocarbons and other mineral resources. The development of products for use by that industry has had some influence on what is used in academic studies, but less than might be supposed: Most academic studies are of waves with lower frequency (which travel to greater distances and depths), and from earthquake sources. We focus here on the instruments and procedures used primarily outside the exploration industry.

The seismic system consists of two parts: the seismometer, or seismic sensor, a transducer for converting ground motion into some other recordable quantity (almost always an electrical signal), and a recording system. (Note that in the exploration industry the customary term for seismometer is "geophone.") While the recording system shares some properties with other types of data recording, the nature of seismology imposes special requirements. When earthquake-generated waves are used, the times at which these will occur are unknown. This means that seismographs must be operated continuously; but since the signals of interest occupy only a small fraction of the recording time, the recorder needs a high capacity.

Because each earthquake is a unique and nonreproducible event, geophysicists often use data from instruments that would now be obsolete, extracting new results from old data. It is therefore appropriate to consider the historical development of seismic systems in some detail. The first instruments capable of producing data of useful quality were developed around 1900 (Dewey and Byerly, 1967); studies of earthquakes before

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

this time are characterized as "preinstrumental." These early systems, like those in use today, measured the motion of a suspended mass relative to the ground, but were limited in their ability to magnify this motion, since this was done through a system of levers, either mechanical or optical. Such "direct-recording" systems predominated until the 1950s; one of them, the Wood–Anderson seismometer (Anderson and Wood, 1925), was the basis for the earthquake magnitude scale. An advance on this was the introduction of electromagnetic sensors (discussed below) driving galvanometer recorders. Although developed soon after 1900 by B. Golitzin, these were not used extensively in global seismology until the introduction of the World-Wide Standardized Seismographic Network in 1963. The latest advance, in the mid-1970s, was the introduction of active electronics and feedback into the sensor to shape the response and reduce noise.

The mid-1970s also saw the beginning of routine digital recording in earthquake seismology. Almost all systems before this (except in exploration geophysics) recorded a time record of the signal as a "wiggly" line, purely in analog form. The rapid advance of computer technology since 1970 has meant that digital storage devices have reached the capacity needed for seismology, and virtually all new seismographs are digital in their recording, though the sensors use analog electronics.

1. SEISMIC SENSORS

Except for the limited class of instruments discussed in Sec. 1.8, all seismic sensors are inertial: They measure the motion of a suspended mass relative to the instrument frame, the motion of which can usually be taken to be the same as that of the ground. In all modern instruments the mass is free to travel only in one direction, giving a single degree of freedom, and a single direction of ground motion to which the sensor is (ideally) sensitive. The suspension must also exert some restoring force to keep the mass in a fixed position in the absence of shaking. This restoring force is, again ideally, proportional to the distance of the mass from its stable position, and this makes the equation of motion of the mass that of a standard harmonic oscillator:

$$m\ddot{q} + d\dot{q} + kq = m\ddot{x}, \tag{1}$$

where q is the displacement of the mass relative to the frame and x the motion of the frame in "inertial" space (actually, to a frame corotating with the Earth). The mass is m, the restoring force is kq, and $d\dot{q}$ is a velocity-dependent force included to account for damping. This can be written in a more standard form as

$$\ddot{q} + 2\gamma\omega_0\dot{q} + \omega_0^2 q = \ddot{x}, \tag{2}$$

where ω_0 is the frequency of the undamped mechanical mass/suspension combination and γ is the damping factor.

This equation shows that what a seismic sensor measures is ground acceleration (with some complications to be discussed below). If the frequency of ground motion is much less than ω_0, the mass motion q will increase as ω_0^{-1}. In early instruments, when magnifying the motion was difficult, this made long mechanical periods attractive, which was achieved by minimizing the restoring force. The difficulty this introduces is that the lower the restoring force the more susceptible the instrument is to forces introduced from other sources (such as temperature effects), and the more fragile it becomes: Seismometers for use under rough conditions thus have to have shorter periods than those fixed in observatories. Note that if the frequency of ground motion is much greater than ω_0, the mass displacement is simply proportional to the ground displacement, so that the sensor can be termed a "displacement meter."

One complication that can be of importance is that horizontal accelerations caused by local translation of the ground cannot be distinguished from the apparent accelerations caused by rotations about a horizontal axis (tilts). These accelerations appear because any tilt gives the local gravity vector a component along any originally horizontal instrument axis. At long periods, the acceleration caused by a given translation becomes smaller and the importance of tilt-induced accelerations larger. While it is possible to measure tilts separately using a gyroscope or suspended bar to define a fixed direction, these techniques do not have adequate stability for geophysical use. All devices constructed to measure tilt in fact use the gravity vector to define a direction and are thus affected by accelerations. Whether a device is called a seismometer or a tiltmeter depends on the emphasis placed on one aspect or another: What is measured is the same for both (Agnew, 1986). Similarly, since a vertical seismometer measures changes in vertical acceleration, it is in some sense equivalent to a gravimeter for measuring changes in the acceleration of gravity: Which name is used depends upon the application.

1.1 Sensor Requirements: Dynamic Range, Sensitivity, Bandwidth

What has made seismographs such demanding instruments to construct has been the very wide range, in both frequency and amplitude, in what is to be measured. Figure 1 shows the range of some signals and the background noise against which they must be measured.

Partly because of the peak in the noise spectrum (Fig. 1), known as the microseism peak, and partly for historical reasons, seismic waves (and seismic sensors) have been divided into short period (SP) and long pe-

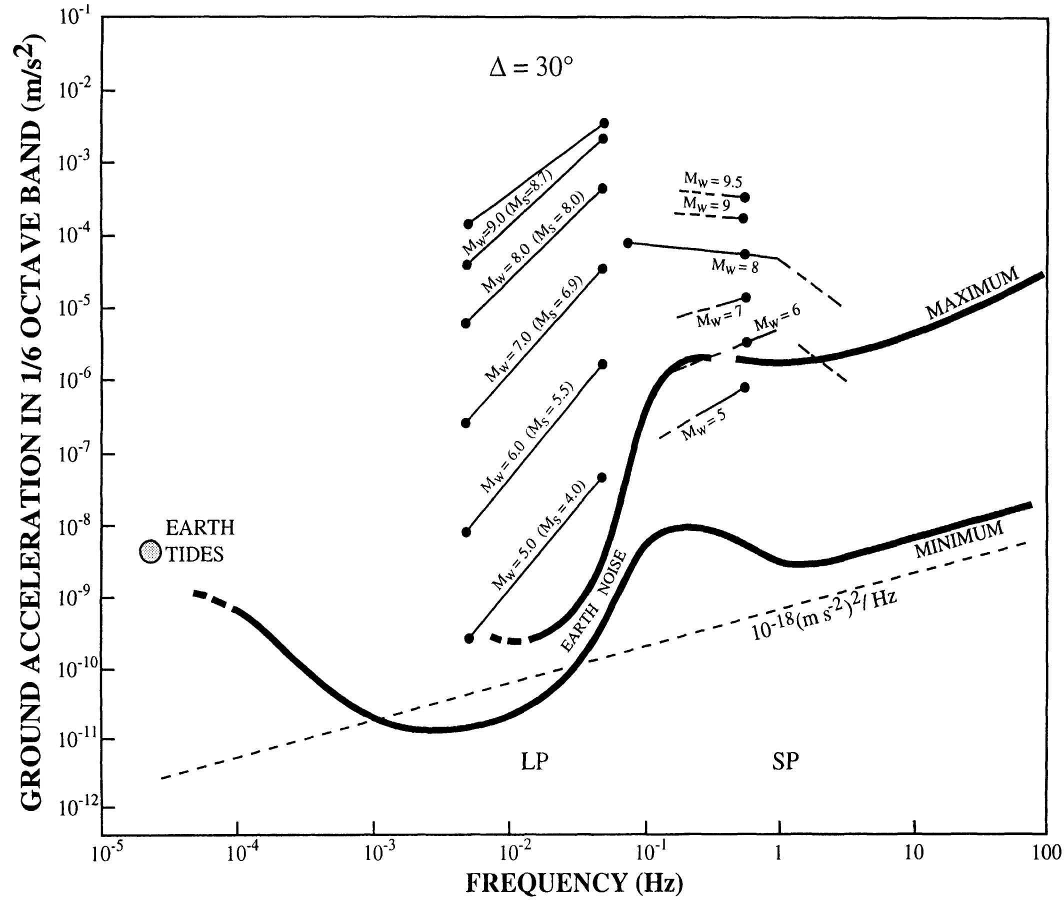

FIG. 1. Ambient seismic noise. The heavy lines show the range of expected ambient noise at a continental site. The light lines show the signals expected from earthquakes of different magnitudes at a distance of 30°. The dashed line indicates a level of constant acceleration with a rms value of 3×10^{-9} m s^{-2} over a 10-Hz bandwidth.

riod (LP), with the dividing period being about 6 s. This is, however, a relative term: In exploration geophysics, a long-period sensor might be sensitive to wave periods longer than 0.5 Hz.

Earthquake signals that propagate as elastic waves extend over a bandwidth from about 3×10^{-4} Hz, the frequency of the gravest normal mode of the whole Earth, to frequencies in excess of 50 Hz. The attenuation in the Earth limits the maximum frequency present to about 100 Hz within tens of kilometers of the source and 1–10 Hz at distances beyond 1000 km.

Ground motions generated by earthquakes can reach from nearly 1 $\mathrm{m\ s^{-2}/Hz^{1/2}}$ in the near field (on the order of an earthquake fault dimension) down to the smallest detectable signals just above the ambient background noise. At teleseismic distances (1000 km or more from the source) or greater, peak accelerations will rarely exceed 5×10^{-3} $\mathrm{m\ s^{-2}/Hz^{1/2}}$, and those will be for surface waves with a frequency of about 5×10^{-2} Hz. Figure 1 illustrates the range of signals expected at about 30° (3300 km) from an earthquake source. The background noise, representing the limit of detectability of earthquake signals, is a composite of the quietest readings observed at different sites around the world (Peterson, 1993). The dynamic range of teleseismic signals to be measured peaks at some 160 dB at about 3×10^{-2} Hz.

1.2 Suspension Designs

As noted above, the mass in an inertial seismometer is attached to the frame through some mechanical system, called the "suspension," which has to allow it to move with only one degree of freedom and provide a restoring force to keep it in position in the absence of acceleration. This is especially difficult for instruments that measure vertical motion (in whole or in part) because the restoring force must then be, not zero, but an amount equal to the local force of gravity upon the mass. The difficulty is then in keeping this nonzero restoring force constant to 10^{-10} (equivalent to the ambient ground noise over a 10-Hz bandwidth). Nearly all inertial seismometers use some combination of elastic materials and geometrical constraints on the mass to provide a suspension. The simplest example of an elastic suspension would be a mass hanging from a coil spring; a simple suspension for horizontal motion would be a mass on the end of a rod, forming a simple pendulum. In the second case the restoring force is provided by gravity. Most actual suspensions are elaborations and combinations of these.

For measurements of horizontal motion the mass may be constrained to move in a straight line, with the restoring force provided by springs: This is the design used in most short-period sensors. For longer-period instruments, an old, but still common, suspension remains the horizontal pendulum (or "garden gate"), in which the mass is constrained to move in a nearly horizontal circle, making the gravitational restoring force small (Rodgers, 1968). Still another way to achieve this is the inverted pendulum, in which the mass rotates above its pivot point; this is, of course, unstable, so that springs are used to provide the necessary restoring forces.

For the shortest-period vertical sensors, the mass hanging on a spring is a satisfactory design, but if the period is (in practice) much more than 1 s, the equilibrium extension of the spring under gravity becomes too long. The usual solution to this is the LaCoste "zero-length spring" suspension (LaCoste, 1934), in which a specially designed spring is used in a particular geometry that can give an arbitrarily small restoring force linear in mass displacement. Other designs are possible; for example the STS-1 seismometer (Wielandt and Streckeisen, 1982) used in many modern installations uses a design based on the elastica (the shape assumed by a thin strip of metal bent elastically to large deflection).

One complication in an actual seismometer design is that whatever constraints are applied to the mass must not apply any additional restoring force to it; thus, for example, if the mass pivots about some point, the pivots must not apply any forces (unless designed to). As the period increases the limits on allowable extraneous forces diminish, and the need for careful design increases. Most seismometers thus require fine mechanical design and construction. A further complication is that many such forces are produced by thermal expansion of the materials of which the sensor is constructed, so that the best operation of a seismometer will take place only in a stable environment.

1.3 Position Sensors

The mechanical design of the sensor provides a mass whose displacement with respect to the instrument frame indicates Earth motion according to Eq. (1). The next step, in any modern instrument, is to convert this motion into an electrical signal that can be recorded. This is the purpose of the position sensor. While a very large number of such sensors have been designed for industrial use, only a few types have been employed in seismometers.

By far the commonest sensor is the moving-coil system, in which a coil moving through the field of a permanent magnet produces a voltage whose magnitude is proportional to the velocity of the mass. This system is in common use for sensors with periods of 1 s and less, and is simple, robust, and linear. In conjunction with modern electronic amplifiers such a sensor can resolve ambient ground motion for periods from 10 s to many hertz (Rodgers, 1994). It does not, however, have adequate sensitivity at longer periods. For longer-period systems it is necessary to use a position sensor whose output is proportional to mass displacement; essentially the only systems used in seismometers are capacitive and inductive sensors [both reviewed in Agnew (1986)]. The capacitive sensor is capable of resolving much smaller motions, but is environmentally less robust than the inductive systems. Both types of sensors require specialized electronics with phase-sensitive detection in order to achieve low noise, and also need careful design to assure a linear response over their full range of motion.

1.4 Feedback

One way of greatly improving sensor performance, now utilized in nearly all "observatory-grade" seismometers (though not in geophones), is to incorporate feedback of some sort. What this means is that the output of the mass-position sensor is not just recorded, but is also used to generate a force that is applied to the mass [in addition to those described by Eq. (1)]. There are a variety of ways of doing this; the most common is to make the force proportional (over some frequency range) to the mass displacement away from a null position: This is called force feedback, or sometimes a force-balance system. This requires low-noise electronics, but has the merit of reducing the motion of the mass significantly, thus reducing the linearity required of the mechanical system and replacing it by that needed in the electronics (and the voltage-to-force converter). The two main methods of applying a feedback force (electromagnetically, through a coil in a magnetic field, and electrostatically, through a capacitor) can both be made very linear.

Figure 2 illustrates schematically the principles of force feedback. In the Fourier domain, the quantities $A_g(s)$ and $E_o(s)$ are the ground acceleration and the output voltage, respectively. The quantities T_p, T_1, and T_2 represent the responses of the seismometer pendulum described, the response of the displacement transducer, and the feedback portion of the system, respectively. The feedback signal, converted to a force, is summed with the force of the ground acceleration by the mass itself with the positive and negative signs indicating that these forces are in opposition.

The overall response of the system is given by

$$E_o(s) = \frac{T_p T_1}{1 + T_p T_1 T_2} A_g(s), \tag{3}$$

where

$$T_p(s) = \frac{1}{s^2 + 2\gamma\omega_0 s + \omega_0^2} \tag{4}$$

from Eq. (2).

By careful design of the responses of the displacement transducer and feedback circuit elements, it is possible to alter the frequency response of the system, that is $E_o(s)/$

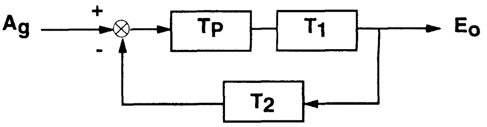

FIG. 2. Block diagram of feedback seismometer.

$A_g(s)$, from that available from the purely mechanical system described by Eq. (2). For example, with a mechanical system of relatively short period, one can obtain a system response that is flat with respect to ground velocity over an extended frequency range (Fels and Berger, 1994).

1.5 Instrumental Noise Sources

When seismologists refer to ground noise, they usually mean motions of the ground caused by something other than the waves of interest. Distinct from this is noise in the seismic sensor itself. Two important sources of noise that are peculiar to such a sensor are the thermal Brownian noise, and apparent tilts.

The thermal Brownian noise is the irreducible minimum noise level of a sensor, and arises from the motion of the mass caused by Brownian motion of the surrounding materials. It can be shown (Melton, 1976) that, for a freely suspended mass whose motion is governed by Eqs. (1) and (2), the noise level is flat in acceleration power spectral density, and is equal to

$$8kT\gamma\omega_0/m, \qquad (5)$$

where T is the absolute temperature of the system and k is Boltzmann's constant. As an example, a mass of 0.1 kg with a restoring force that gives a period of 2 s, and critical damping ($\gamma = 1$), will have a noise level at room temperature of 10^{-18} m^2 s^{-4} Hz^{-1}, comparable to the lowest level of ground noise. Since the temperature cannot easily be lowered, the two main ways to reduce the noise are to increase the mass or to decrease the damping force. The latter can usefully be done if the system uses feedback to provide damping, since this does not increase noise; this approach is followed in several modern sensors.

Human activities also are a source of seismic noise, and sensors should be located in an area removed from sources of cultural noise such as towns or major roads. For island settings, the site should be chosen as far from the coast or breaking reef as practical to avoid noise caused by the surf.

As noted above, horizontal seismometers depend on the direction of local gravity for a reference, and thus cannot distinguish between tilts and horizontal motions. At periods greater than about 10 s, the tilts of the ground caused by loading from atmospheric turbulence created by wind produce a noise level on horizontal instruments that exceeds that present in the vertical by a considerable amount. Thus, the seismometer should be well insulated and shielded from the effects of temperature-induced tilts and from the effects of wind- and pressure-generated Earth noise at the surface (Sorrels and Goforth, 1973). Additionally, most sensors are temperature sensitive, and a stable temperature environment is required.

Traditionally, seismometers have been installed in basements, vaults, tunnels, or other places where a good coupling to the local rocks can be achieved, where the sensors can be isolated from common sources of man-made vibrations, and far enough below the surface to reduce the temperature-induced tilts.

In the 1960s much research was devoted to studying the efficacy of installing seismometers in boreholes. It was quickly discovered that while the background noise decreased with depth, so too did the signal strength as a result of destructive interference from surface reflections. Empirically, it was found that not much improvement was achieved beyond a depth of 100 m, and this is now the accepted standard for quiet continental sites.

1.6 Calibration

In order to obtain useful information from the recorded seismogram, it is necessary to have the seismic system calibrated: It must be known, for all frequencies of interest, how large a ground motion in a given direction will produce a given recorded signal. This problem, in most modern systems, is divided into two parts. Relative calibration is the problem of determining the frequency response of the system, which may be done by injecting a known signal into the feedback loop (or an auxiliary forcing system) and performing a cross-spectral analysis between this known input and the recorded output (Berger *et al.*, 1976; Rodgers *et al.*, 1995). This may, with modest care, provide relative responses to better than 1%.

The more difficult problem of absolute calibration is then to relate the output to actual ground motion (perhaps only at one frequency). Providing a known displacement (from putting the seismometer on a shake table) is possible at short periods but becomes difficult at longer ones: Most long-period systems are calibrated by tilting them (so that the change in gravity mimics an apparent acceleration) or by moving masses around them.

Another method of calibrating seismometers is to provide a way to induce a force on the seismometer mass electromagnetically, usually by the application of a voltage to a forcing coil. In practice, it is considerably easier to calibrate the forcing coil than it is to calibrate the overall seismometer and hence this technique is widely used.

Finally there is the apparently simple problem of establishing the actual orientation of the seismometer. Since this often has to be done in boreholes or other underground locations this is not as straightforward as it might appear; and experience has shown that it is often done incorrectly.

1.7 Examples of Modern Seismometers

Two examples of modern three-component seismometers are shown in Figs. 3 and 4; one is designed for quick and simple installation in a vault or other environment accessible to the installer and the other is designed to be emplaced in an underground borehole. Both instruments employ force feedback and have electronics to shape the response so that the output is proportional to ground velocity over most of their bandwidth, which extends from less than 10^{-2} to about 50 Hz.

The STS-2 manufactured by G. Streckeisen AG of Pfungen, Switzerland, is a second generation of the leaf-spring design of Wielandt and Streckeisen (1982). It is meant to be easily portable and capable of operation over a wide temperature range. The unit employs three identical sensors, which are obliquely oriented, with their sensitive axes 45° to the vertical. Analog circuitry combines the sensor outputs to provide signals proportional to the ground motion in the coordinate system of the local vertical. The sensors and control electronics are mounted in a vacuum-tight cylindrical case of 24-cm diam designed to minimize the effects of distortion by changes in barometric pressure. The top and bottom covers are secured to a massive base plate in such a way as to let the covers compress without stressing the entire package. As with other modern feedback seismometers, it will resolve the minimum ground noise (Fig. 1) over a frequency band from about 2×10^{-3} to 15 Hz.

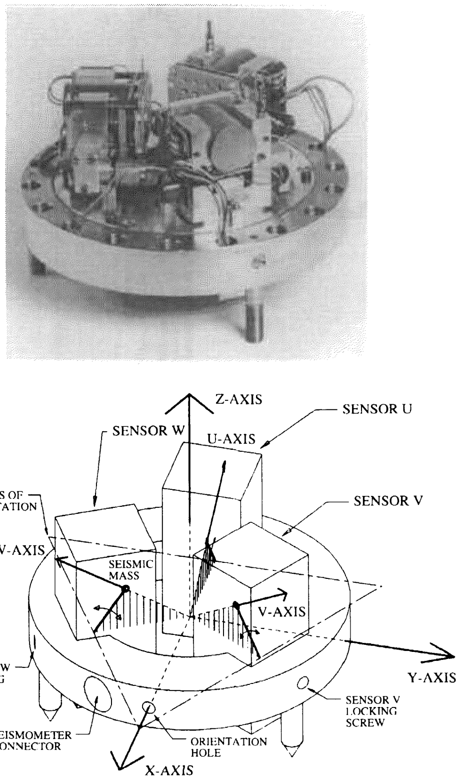

FIG. 3. Streckeisen STS-2 three-component seismometer.

The CMG-1, manufactured by Guralp Systems, Ltd, of Reading, England, is a borehole seismometer designed to be installed in a casing of at least 15-cm inside dimension. It consists of a vertically oriented and two horizontally oriented sensors housed, together with control electronics, in a package of 12-cm diam by 1.55-m length. Two external cylindrical metal bars are affixed to the casing to fit the inside diameter of the borehole into which the package is installed. After lowering the unit to the desired depth, a spring-loaded arm is released by motor drive, which locks the package firmly against the borehole walls.

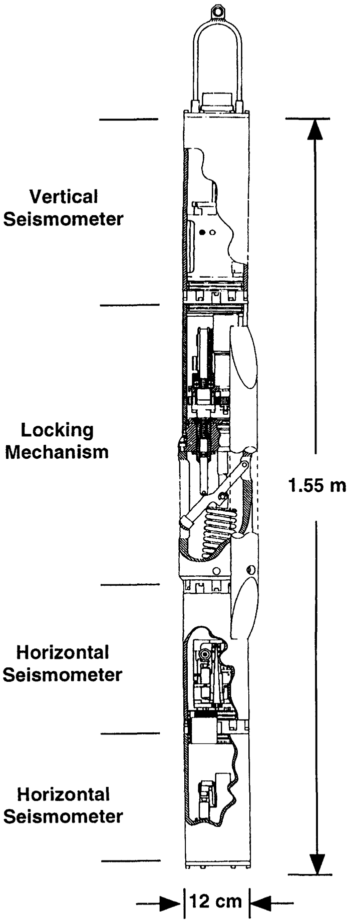

FIG. 4. Guralp CMG-1 three-component borehole seismometer.

The CMG-1 will level itself in boreholes tilted up to 3.5° from vertical. The horizontal components can be rotated relative to the borehole *in situ* to a precision of 0.2°. It has an output that is proportional to ground velocity from about 2.8×10^{-3} to 50 Hz with an internal noise lower than the lowest ambient ground noise (Fig. 1) from about 1.5×10^{-3} to 20 Hz.

1.8 Strain Seismometers

One method of recording seismic waves does not rely on measuring the ground motion, but rather the strains associated with it. One way of doing this is to measure the displacement between two relatively closely spaced points: Since this is (for seismic waves with wavelengths longer than this distance) proportional to the extensional strain, such instruments are also called extensometers. The first strain seismometer of sufficient sensitivity was developed by Benioff in 1935: This used a velocity transducer, one part of which was attached to one point and the other to a 30-m rod extending to the other point; such solid-rod strainmeters have been built in a variety of modes (Agnew, 1986). A greater separation between fiducial points can be achieved by using a laser system to measure the relative displacements. An alternative way to measure strain is to cement a flexible capsule in the ground, and measure the change in volume through the motion of a liquid in and out of it; such systems, called dilatometers, are in some use for looking for very-long-period strain changes. Neither extensometers nor dilatometers have seen much use in recording seismic waves; it does not appear that strain measurements usually provide much information not available from measuring displacements with inertial sensors.

2. RECORDING SYSTEMS

2.1 Recording Systems—Historical

In the earliest seismometers the motion of the mass was magnified mechanically and recorded directly, one of the lowest-friction ways of doing this being to have a stylus scratching paper coated with a layer of soot. Direct optical magnification, recording on photographic paper or film, came into common use somewhat later. Both methods remain in occasional use, the former being an inexpensive and robust method for field operations. The usual way of providing an immediate visual record remains a pen-and-ink recorder, almost always recording on a drum, but the majority of seismic recorders (before the introduction of digital recording) used photographic recording of the light spot reflected off a mirror attached to a galvanometer. The galvanometer provided a low-noise method for detecting the small voltages generated by most seismic sensors (especially those using a moving-coil position sensor), as well as some flexibility in determining the frequency response (Willmore, 1979).

2.2 Modern Data-Acquisition Systems

The general requirements of the seismographic data-acquisition system are dictated by the purpose of the station: Those employed for seismic exploration will differ considerably from those employed for teleseismic monitoring, for example. Many elements are common, however, and the spectra of signals to be recorded and background noise will determine the details. For illustration, we describe the requirements for a teleseismic station. Figure 1 depicts the spectrum of signals to be measured and a model of the ambient ground noise. From this figure, the required bandwidth and dynamic range can be determined.

The main functions of the data-acquisition system are

1. amplification or electrical isolation of each input analog signal,
2. antialias filtering,
3. analog-to-digital conversion,
4. formatting and time stamping of the data,
5. signal processing, mainly filtering and decimation,
6. event detection,
7. digital data storage on some high-capacity medium,
8. data transmission, via a modem or hard wire, to a data-collection center, and
9. digital-to-analog conversion for generating calibration signals or to monitor the digital signals.

Because the sensors are usually placed in a relatively inaccessible location such as an underground vault or borehole, the data-acquisition systems are usually divided into two parts. The amplifiers and analog-to-digital conversion units (ADC) are located close to the sensors to minimize possible electrical interference, while the control and recording units are located in a place with convenient access for operators. A generic seismological data-acquisition system designed primarily for use at earthquake observatories is outlined in Fig. 5.

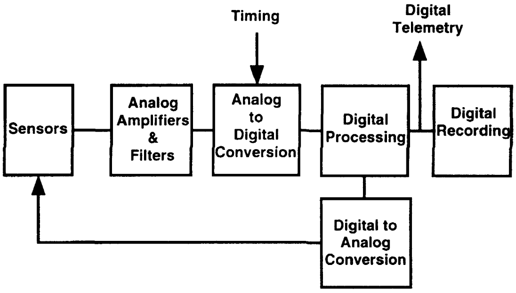
Timing
Digital Telemetry
Sensors
Analog Amplifiers & Filters
Analog to Digital Conversion
Digital Processing
Digital Recording
Digital to Analog Conversion

FIG. 5. Generic seismic data-acquisition system.

The analog amplifiers and filtering subsystem provide the interface between the sensors and the ADC. Their purpose is to provide any analog amplification that may be required to match the sensor output levels with the available ADC noise level. Additionally, the unit must low-pass filter the analog signals to insure there will be no aliasing in the analog-to-digital conversion process. In general the ambient ground noise passed through the sensor should produce a voltage slightly greater than the combined noise level of the amplifiers, analog filters, and ADC in order to maximize dynamic range.

Quite often, a variety of sensors will be used to measure the entire spectrum of seismic signals and other interesting variables. A typical unit might be designed to accept the following inputs:

1. three axis broadband channels,
2. three axis high-frequency channels,
3. three axis broadband-seismometer mass-position channels.
4. three axis strong-motion channels,
5. one channel for barometric pressure,
6. one channel for temperature sensor, and
7. three axis anemometer channels.

2.2.1 Analog-to-Digital Conversion

The analog-to-digital conversion subsystem is the heart of the data-acquisition system. As the name implies, it performs the conversion of the signal from its continuous analog form into its sampled digital representation. Additionally, it must "attach" to each digital sample the time of sampling.

The bandwidth of the seismic signals of interest (~100 Hz) is not a challenge to modern ADCs. The dynamic range required, however, is available only in the highest-resolution digitizers currently available. As discussed in Sec. 1.1, the dynamic range of teleseismic signals to be measured peaks at about 0.03 Hz at some 160 dB.

The dynamic range can be defined as the ratio of

full scale (FS) to rms noise in $\frac{1}{6}$ decade band ($\Delta f/f = 0.386$).

If the spectral density of the system noise is $N(f)$ per f_n, where f_n is the digitized Nyquist frequency, then the dynamic range per $\frac{1}{6}$ decade band will be *frequency dependent* and given by

$$\text{Dynamic range} = \frac{\text{FS}}{\sqrt{0.386N(f)f/f_n}}. \qquad (6)$$

The process of analog-to-digital conversion introduces an uncertainty into the signal as a result of the finite resolution. Assuming that there is a uniform probability of the actual signal voltage being anywhere in the interval $-\frac{1}{2}\Delta V$ to $+\frac{1}{2}\Delta V$, where ΔV is the resolution or least count (lc) of the digitizer, this is equivalent to adding a noise of variance $\frac{1}{12}(\text{lc})^2$ evenly distributed over the recorded bandwidth. Thus, the perfect digitizer adds a noise of power spectral density

$$N(f) = (\text{lc})^2/12f_n, \qquad (7)$$

where f_n is the digitized bandwidth. (From this formula, it can be seen that the noise of the digitizer can be reduced by increasing the digitizer bandwidth—that is, sampling faster.) For an n-bit ADC, where $\text{FS} = 2^n$, the dynamic range for a perfect n-bit encoder in decibels (dB) is then

$$\text{Dynamic range} = 6.02n + 17.4 - 10\log(f/f_n)\ \text{dB}. \qquad (8)$$

In Table 1, the dynamic range of perfect n-bit encoders is given at the Nyquist frequency and at 1/2000 of the Nyquist frequency. For a unit sampling at 200 samples per second these frequencies would be 100 and 5×10^{-2} Hz respectively. Also listed is the sample rate required to produce 160-dB dynamic range at 20 Hz, matching the performance of the seismometer illustrated in Fig. 1. From this table it can be seen that to achieve the desired dynamic range with a single digitizer, 24-bit precision is required.

Table 1. Dynamic range for perfect n-bit encoders.

Number of bits	Dynamic range at f_n (dB)	Dynamic range at $5\times10^{-2}f_n$ (dB)	Sample rate for 160 dB at 20 Hz (s^{-1})
12	89.64	122.64	4.58×10^8
16	113.72	146.72	1.82×10^6
24	161.88	194.88	40

In practice, the noise of the digitizer usually exceeds these theoretical figures as a result of a variety of internal electronic sources. A typical noise spectrum of a digitizer is illustrated in Fig. 6 (Given *et al.*, 1992). In general the shape is common to most such devices: a flat spectrum down to some frequency f_0, and below that a $1/f$ behavior:

$$N(f) = N_0 \quad \text{for } f > f_0 \qquad (9)$$

and

$$N(f) = N_0 f_0/f \quad \text{for} < f_0. \qquad (10)$$

In this case, for $f < f_0$ the dynamic range will be a constant at

$$\text{Dynamic range} = 6.02n + 6.6 - 10\log N_0 - 10\log(f_0/f_n)\ \text{dB}. \qquad (11)$$

2.2.2 Timing Timing of seismographic data is a critical element of station design, as the analysis of the data depends upon a common time base for widely separated sites. Generally speaking, the time *accuracy* should be a small fraction of a sample interval so that time series from across the network can be aligned. For data sampled at 20 Hz, this translates to a desired accuracy specification of about 5 ms.

There are a number of systems that broadcast Universal Coordinated Time (UTC) of sufficient precision for seismographic purposes. A number of national authorities support radio transmissions such as those broadcast on WWVB in the United States and MSF in Northern Europe, but the coverage of these systems is quite limited (for a global network) and reception is often poor. The OMEGA navigational system offers practically global coverage but sometimes reception is adversely affected by local topography. The system provides an accuracy estimated to 5 ms by clock manufacturers, but UTC can only be determined to within ±5 s because of the nature of the broadcast signals. An operator must set the time to within this bound manually. The recently introduced Global Positioning System (GPS), however, offers a very accurate global timing

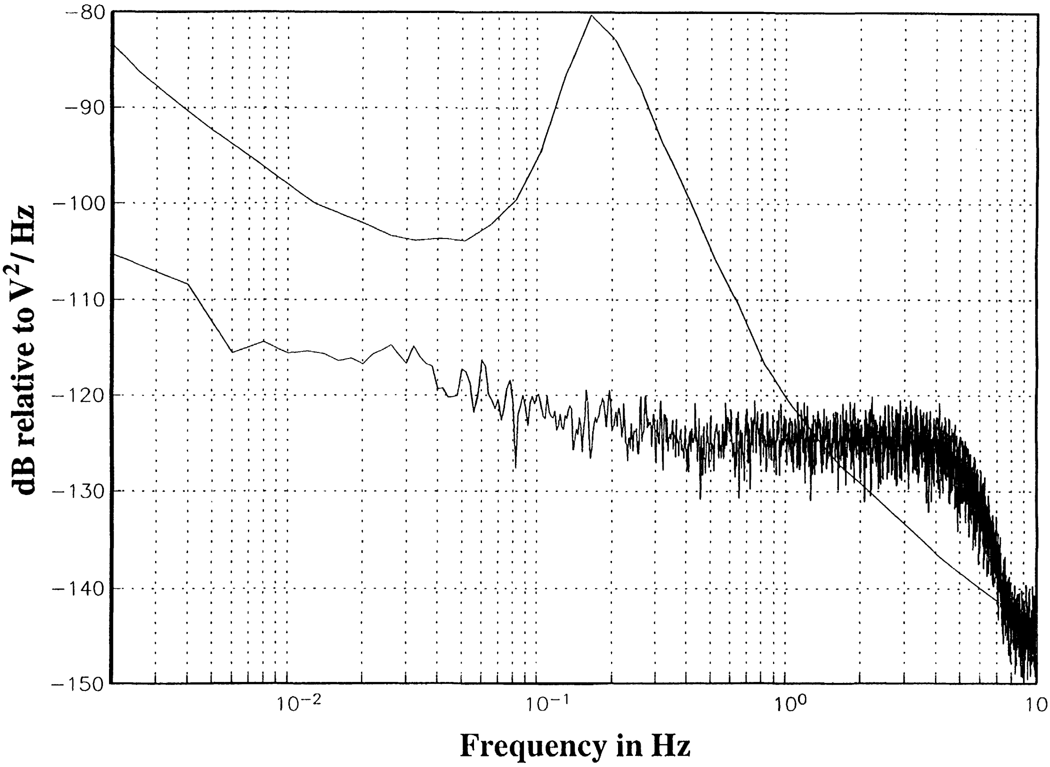

FIG. 6. Performance of a 24-bit Hewlett-Packard analog-to-digital converter. Full scale is 40 V and the sample rate is 20 samples s^{-1}. For frequencies above 0.15 Hz, the noise figure is about 6 dB higher than a perfect 24-bit encoder. For comparison, the smooth curve represents a model spectrum of ambient ground noise that has been passed through a hypothetical seismometer with a conversion factor of 2400 V m^{-1} s^2.

system that is rapidly replacing the other systems completely.

The clocks typically produce two signals: a 1-Hz pulse precisely synchronized to UTC and a time string containing the day, hour, minute, and second, and a quality factor character. The quality factor indicates if the clock is locked to the broadcast standard or, if not, how long it has been out of lock.

2.2.3 Digital Processing As discussed above, it is common to oversample the analog data to achieve better resolution through averaging. The averaging is done as part of the data-acquisition process by an internal processor, which performs digital filtering and decimation. For data-analysis purposes, it is desirable to have a variety of data streams recorded from a single sensor. Typically, sample rates of 100, 40, 20, 1, and 0.1 Hz are recorded.

In addition to the filtering and decimation, digital processing is often required to provide a "detection" filter for selected data streams. This allows, for example, strong-motion sensors to be recorded at a high rate only when a signal is detected. Data are bandpass filtered in a frequency range appropriate for the sensors, and then passed to a trigger algorithm.

A typical detector for strong-motion sensors is based on the comparison between a short-term average (STA) and a long-term average (LTA) of the sensor signal (or channel) passed through a 1.0–30.0-Hz bandpass filter. The LTA is actually computed as an average of STAs and therefore "lags" the STA by a period of time. When the ratio of the STA to the LTA is above a threshold level then the detector triggers for that channel. A trigger is maintained until the STA/LTA ratio drops below some fraction of the trigger ratio, or until the maximum event time is exceeded. Multiple channels may be used as detector inputs; an "event" occurs when a predefined number of input channels trigger (referred to as the "voting threshold"). This helps to decrease the number of false triggers due to a noisy channel. While an event is occurring any streams designated as "triggered" will be saved (otherwise they are kept only temporarily as pretrigger records). A record is usually kept of event start and end times (and approximate amplitudes). Typical event detector parameters might be as follows:

1. STA length = 0.05 s,
2. LTA length = 10 s,
3. trigger ratio = 5,
4. maximum event time = 4 min,
5. detector input channels = three short-period channels, and
6. voting threshold = 2.

2.2.4 Digital Recording It is usually required to record all data streams at the station as a local archive. The data rates from a typical teleseismic station often amount to about 2000 bytes/s of raw data, and so the daily volume of data to be recorded amounts to nearly 175 megabytes. That rate can be reduced by about a factor of 3 by various forms of data compression. A variety of digital recording media such as digital audio tape (DAT) or helical-scan 8-mm tape are readily available to accommodate a week's data on a single volume.

2.2.5 Digital Telemetry It is common to require a method of near real-time recovery of at least some of the data via telecommunication circuits. While data rates recorded at the station may amount to 2000 bytes per second, when required this rate can easily be reduced. Table 2 indicates that the overall data rate can be reduced to about 200 bytes per second for each station.

Table 2. Typical data rates.

Data streams	Date rates
Continuous seismic	3 channels × 40 samples s^{-1} × 4 bytes $sample^{-1}$ = 480 bytes s^{-1}
Other data	20 channels × 1 samples s^{-1} × 4 bytes $sample^{-1}$ = 80 bytes s^{-1}
Total data	= 560 bytes s^{-1}
Total with compression	Assume factor of ~3 = 200 bytes s^{-1}

Using modern error-correcting modems, the capacities of two types of commercial telecommunication circuits are indicated in Table 3. The lower-speed circuits are available virtually worldwide via the commercial "dial-up" telephone system. The higher-speed circuits are becoming more available as services such as the Integrated Services Digital Network spread globally.

3. NETWORKS AND ARRAYS

3.1 Local and Regional Networks

Many areas of the world are subject to frequent damaging earthquakes. In these areas, it is common to deploy a local or regional network for the purpose of locating earthquakes within a small, well-defined region.

An example of a regional network is the Southern California Network (Fig. 7), which has the prime purpose of detecting and locating earthquakes in southern California, and particularly in the urban areas around the Los Angeles Basin. Over 100 stations consisting of short-period, vertical-component seismometers are telemetered to a central site in real time, where the data are processed to detect and locate earthquakes in the region.

3.2 Global Networks

The first truly global seismographic network was the World-Wide Standardized Seismographic Network (WWSSN), deployed in the early 1960s (Oliver and Murphy, 1971) primarily to study the problems associated with detection and identification of underground nuclear testing (though not to monitor such tests). The network data, however,

Table 3. Telecommunication circuit capacities.

Nominal circuit bandwidth (bits s^{-1})	Nominal circuit bandwidth (bytes s^{-1})	One-minute volume (kilobytes)	All data compressed (min day^{-1})
9 600	1200	72	240
64 000	8000	480	36

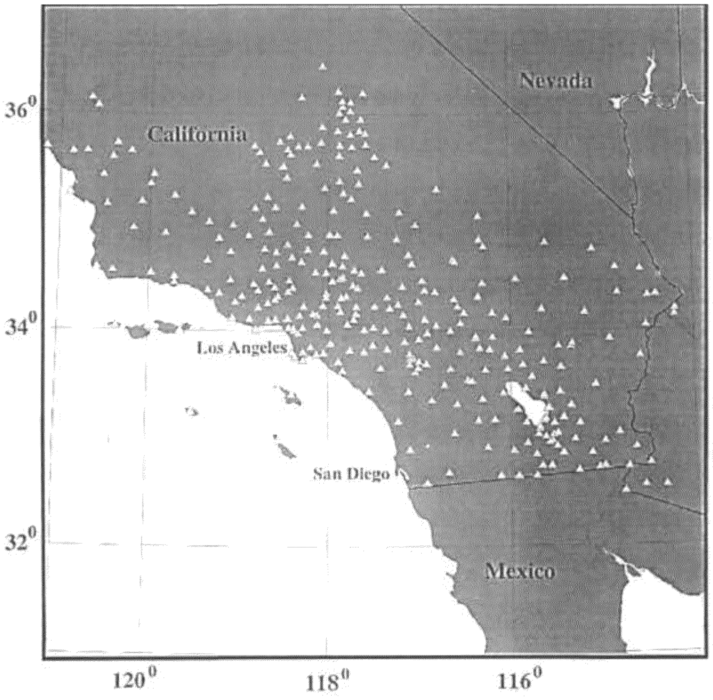

FIG. 7. The Southern California Network.

led to a revolution in solid-earth geophysics. Not only did these data lead to a new understanding of earthquake source mechanisms, but they also led to the seismological evidence that was important to the development of plate tectonics.

Among other things that were learned by an analysis of the WWSSN data were better ways of detecting and identifying clandestine nuclear testing, and this led to a decline in the support of the network and its gradual deterioration. Other less dense networks were deployed, such as GEOSCOPE, the French worldwide digital network, but it was not until the mid-1980s that a concerted plan to replace the WWSSN with modern instrumentation was developed by a consortium of university seismologists called the Incorporated Research Institutions for Seismology (IRIS).

By 1995, some 92 of a planned 147 stations of the Global Seismographic Network were operational (Fig. 8). The stations are not all equipped with identical equipment but they all meet common standards for bandwidth, dynamic range, and linearity. Data from all stations are collected and archived in a common format at a central data center, which distributes data to users upon request.

3.3 Arrays

Seismic arrays consist of two-dimensional groups of sensors deployed over dimensions of a few to tens of kilometers as illustrated in Fig. 9. The basic idea of a seismic array is that the background noise tends to be incoherent over spatial dimensions for which the signals are coherent, and hence processing of the sensor data can enhance the signal-to-noise ratio of such signals (Green *et al.*, 1965). Additionally, processing can "steer" the array in both azimuth and velocity, which can be used to help locate the source of the signals.

If a plane wave in the surface plane impinges upon the array with a velocity c in a direction ϑ, the arrival time at the ith seismometer of the array will be

$$t_i = t_0 + \frac{\cos\vartheta}{c}(x_i - x_0) + \frac{\sin\vartheta}{c}(y_i - y_0), \tag{12}$$

where t_0 is the arrival time at a reference

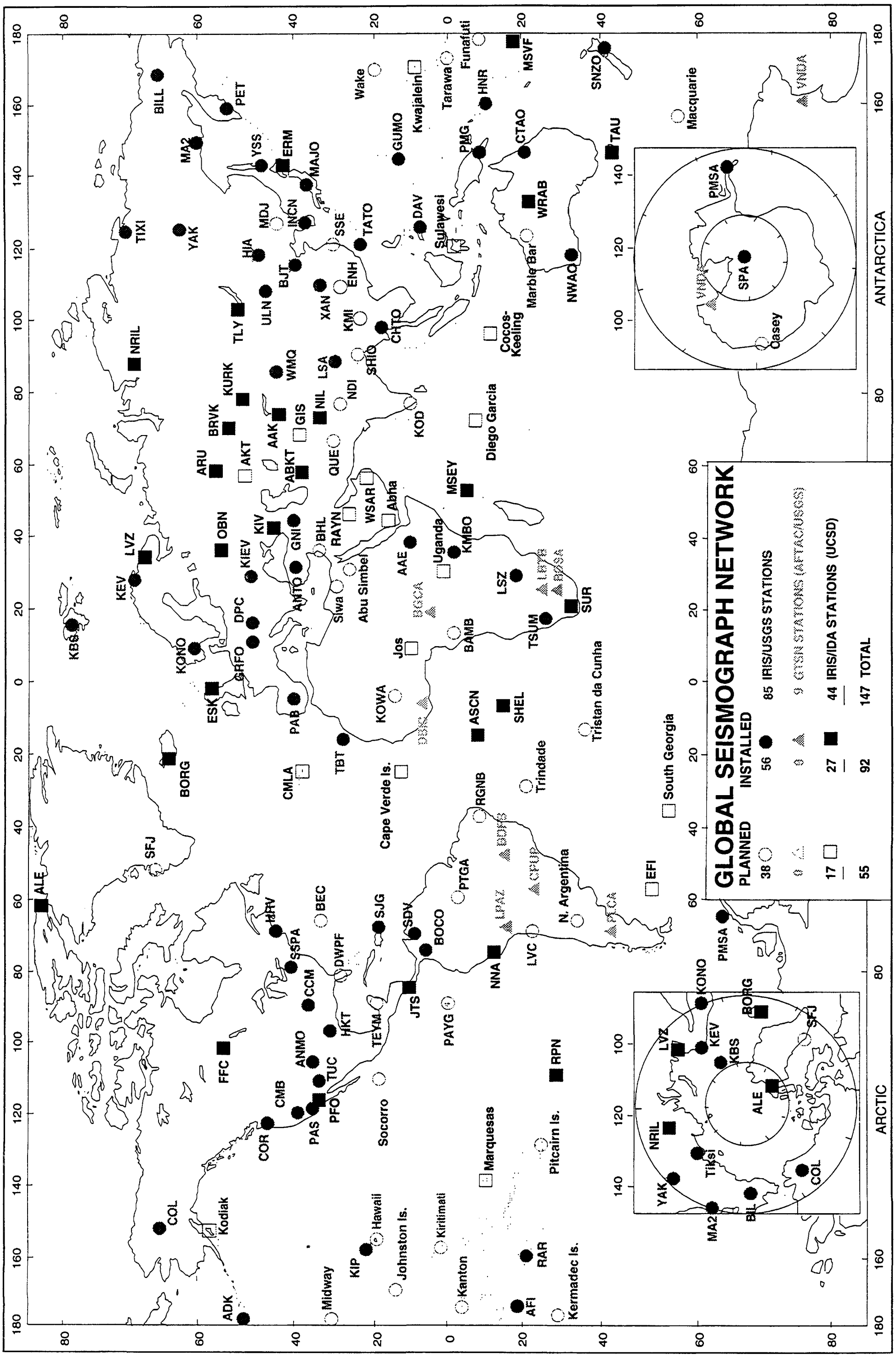
GLOBAL SEISMOGRAPH NETWORK
PLANNED 38 ○ INSTALLED 56 ● 85 IRIS/USGS STATIONS
0 △ 0 ▲ 9 GTSN STATIONS (AFTAC/USGS)
17 □ 27 ■ 44 IRIS/IDA STATIONS (UCSD)
55 92 147 TOTAL
ARCTIC
ANTARCTICA
ALE
SFJ
BORG
KBS
KEV
LVZ
KONO
ESK
GRFO
DPC
KIEV
OBN
ARU
BRVK
AKT
KURK
NRIL
TIXI
YAK
BILL
MA2
PET
YSS
ERM
MAJO
INCN
MDJ
HIA
BJT
ULN
TLY
XAN
ENH
KMI
SSE
TATO
DAV
GUMO
Wake
Kwajalein
Tarawa
Funafuti
MSVF
HNR
PMG
CTAO
WRAB
Marble Bar
NWAO
TAU
SNZO
Macquarie
VNDA
Sulawesi
CHTO
SHIO
LSA
NDI
WMQ
AAK
NIL
GIS
ABKT
QUE
KOD
Diego Garcia
Cocos-Keeling
KIV
GNI
ANTO
BHL
RAYN
WSAR
Abha
Siwa
Abu Simbel
AAE
Uganda
KMBO
MSEY
BGCA
BAMB
LSZ
LBTB
BOSA
TSUM
SUR
Tristan da Cunha
PAB
TBT
KOWA
Jos
DBIC
ASCN
SHEL
Trindade
South Georgia
CMLA
Cape Verde Is.
RGNB
PTGA
BDFB
LPAZ
CPUP
N. Argentina
PLCA
EFI
LVC
NNA
BOCO
SDV
SJG
BEC
HRV
SSPA
DWPF
CCM
JTS
PAYG
TEYM
HKT
ANMO
TUC
FFC
CMB
PAS
PFO
COR
Socorro
RPN
Pitcairn Is.
Marquesas
COL
Kodiak
ADK
Midway
KIP
Hawaii
Johnston Is.
Kiritimati
Kanton
AFI
RAR
Kermadec Is.
PMSA
SPA
Casey
Tiksi
BIL

FIG. 8. The Global Seismographic Network. The IRIS/USGS stations are operated on behalf of IRIS by the United States Geological Survey (USGS) while the IRIS/IDA stations are operated by the University of California, San Diego (UCSD). The GTSN refers to stations of the Global Telemetered Seismographic Network operated on behalf of the Air Force Technical Applications Command (AFTAC) by the USGS.

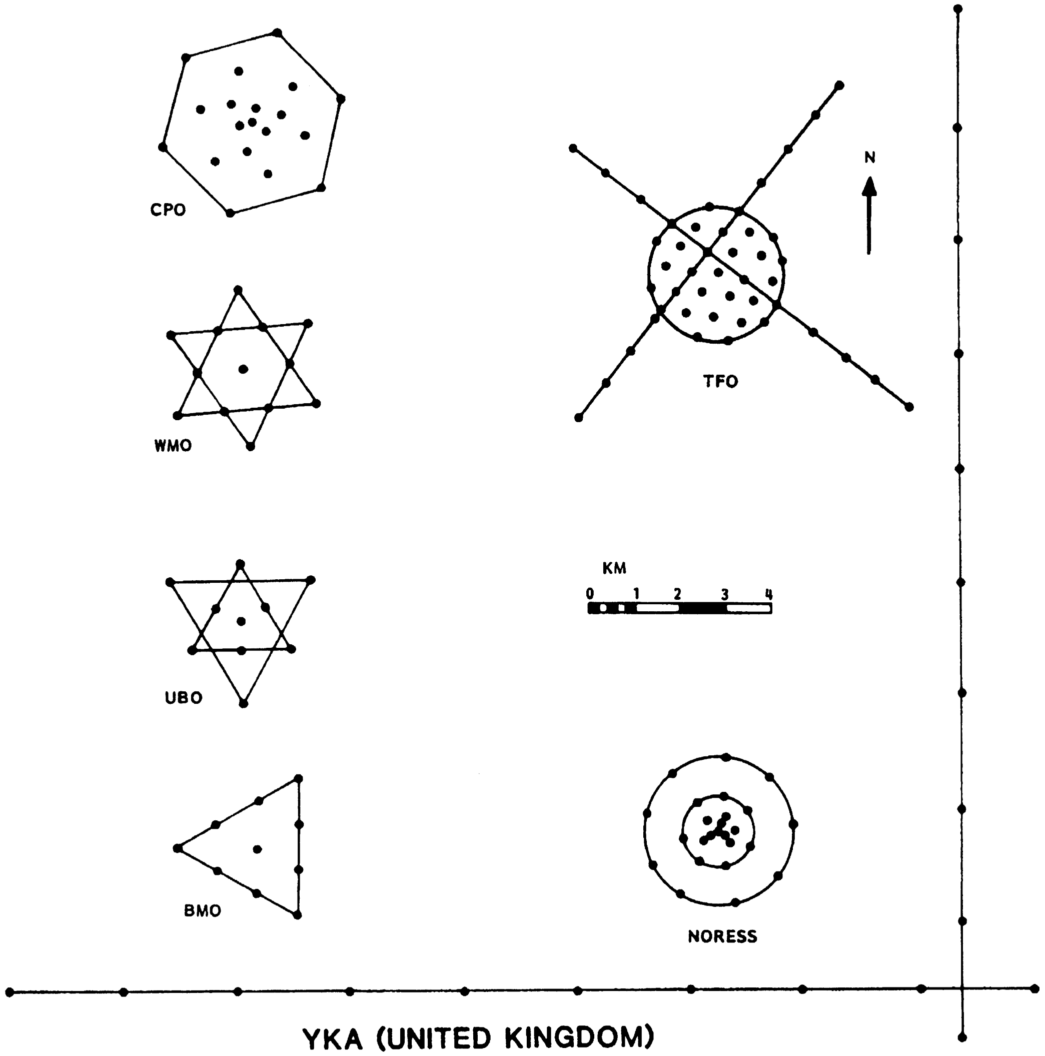

FIG. 9. Some array layouts (after Farrell, 1985). The arrays in the left column all had apertures of about 4 km. TFO was a larger experimental array and NORESS is a modern array of about 3-km aperture and 25 elements. The YKA array is a typical crossed-line array, but in this figure the intersection has been moved.

point (x_0,y_0). The array can be "steered" by delaying each seismometer output by an amount $t_i(\vartheta c)$ and summing the signals.

Plate 1 shows an example of such processing for an array of aperture about 1.5 km shown in the upper left of the figure. The plots of beam amplitude in "slowness" $(\cos\vartheta/c, \sin\vartheta/c)$ space—essentially the reciprocal velocity across the array—are shown for two different signals identified in the beam. Circles indicate constant velocity, with infinite velocity (a vertically traveling wave) at the center. From this analysis it can be seen that the first signal comes from an azimuth of 53° with an apparent velocity of 2.5 km/s—most likely a noise pulse from the nearby city of Ashgabat—while the second comes from an azimuth of 211° with a velocity of 8.6 km/s, appropriate for an earthquake at the location shown.

GLOSSARY

ADC: Analog-to-digital converter.

Geophone: The term used for a seismometer in exploration geophysics. Typically a small unit with a natural frequency of several hertz.

lc: Least count. One digital unit output from an ADC.

LP: Long period—referring to a wave period greater than about 6 s.

LTA: Long-term average.

SP: Short period—referring to a wave period less than about 6 s.

STA: Short-term average.

Teleseismic: Referring to source-to-receiver distances greater than about 2000 km.

UTC: Universal Coordinated Time.

WWSSN: World-Wide Standardized Seismographic Network.

Works Cited

Agnew, D. C. (1986), "Strainmeters and Tiltmeters," *Rev. Geophys.* **24,** 579–624.

Anderson, J. A., Wood, H. O. (1925), "Description and Theory of the Torsion Seismometer," *Bull. Seism. Soc. Am.* **15,** 1–72.

Berger, J., Agnew, D. C., Parker, R. L., Farrell, W. E. (1976), "Seismic System Calibration: 2. Cross-Spectral Calibration Using Random Binary Signals," *Bull. Seism. Soc. Am.* **69,** 271–288.

Dewy, J., Byerly, P. (1967), "The Early History of Seismometry," *Bull. Seism. Soc. Am.* **59,** 183–227.

Farrell, W. E. (1985), "Sensors, Systems and Arrays," in: Ann V. Kerr (Ed.), *Seismic Instrumentation under VELA-Uniform. The VELA Program, A Twenty-Five Year Review of Basic Research,* Washington, DC: Defense Advanced Research Projects Administration.

Fels, J.-F. Berger, J. (1994), "Parametric Analysis and Calibration of the STS-1 Seismometer of the IRIS/IDA Seismographic Network," *Bull. Seism. Soc. Am.* **84,** 1580–1592.

Given, H. K., Berger, J., Fels, J.-F., Horwitt, D., Winther, C. (1992), "The IRIS-3 High Resolution Data Acquisition System," *Technical Report,* Scripps Institution of Oceanography Reference Series No. 92-32.

Green, P. E., Frosch, R. A., Romney, C. F. (1965), "Principals of an Experimental Large Aperture Seismic Array (LASA)," *Proc. IEEE* **53,** 1821–1833.

LaCoste, L. J. B. (1934), "A New Type, Long Period Vertical Seismograph," *Physics* **5,** 178–180.

Melton, B. (1976), "The Sensitivity and Dynamic Range of Inertial Seismographs," *Rev. Geophys.* **14,** 93–116.

Oliver, J., Murphy, L. (1971), "WWSSN: Seismology's Global Network of Observing Stations," *Science* **174,** 254–261.

Peterson, J. (1993), "Observations and Modeling of Seismic Background Noise," *Technical Report,* United States Department of Interior Geological Survey Open-File Report No. 93-322.

Rodgers, P. W. (1968), "The Response of the Horizontal Pendulum Seismometer to Rayleigh and Love Waves, Tilt, and Free Oscillations of the Earth," *Bull. Seism. Soc. Am.* **58,** 1384–1406.

Rodgers, P. W. (1994), "Self-Noise Spectra for 34 Common Electromagnetic Seismometer/Preamplifier Pairs," *Bull. Seism. Soc. Am.* **84,** 222–228.

Rodgers, P. W., Martin, A., Robertson, M., Hsu, M. (1995), "Signal-Coil Calibration of Electromagnetic Seismometers," *Bull. Seism. Soc. Am.* **85,** 845–850.

Sorrels, G. G., Goforth, T. T. (1973), "Low-Frequency Earth Motion Generated by Slowly Propagating Partially Organized Pressure Fields," *Bull. Seism. Soc. Am.* **63,** 1583–1601.

Weilandt, E., Streckeisen, G. (1982), "The Leaf-Spring Seismometer: Design and Performance," *Bull. Seism. Soc. Am.* **72,** 2349–2367.

Willmore, P. L. (1979), *Technical Report, Manual of Seismological Observatory Practice,* World Data Center A Report No. SE-20, Boulder, CO: U.S. Department of Commerce.

Further Reading

Agnew, D. C. (1986), "Strainmeters and Tiltmeters," *Rev. Geophys.* **24,** 579–624.

Aki, K., Richards, P. (1980), *Quantitative Seismology: Theory and Methods,* San Francisco, CA: W. H. Freeman.

Carpenter, E. W. (1965), "An Historical Review of Seismometer Array Development," *Proc. IEEE* **53,** 1816–1821.

Lee, W. H. K., Stewart, S. W. (1981), *Principles and Applications of Microearthquake Networks,* New York: Academic.

Usher, M. J., Burch, R. F., Guralp, C. (1979), "Wide-Band Feedback Seismometers," *Phys. Earth Planet. Inter.* **18,** 38–50.

Willmore, P. L. (1960), "The Detection of Earth Movements," in: S. K. Runcorn (Ed.), *Methods and Techniques of Geophysics,* Vol. 1, New York: Interscience, pp. 230–276.

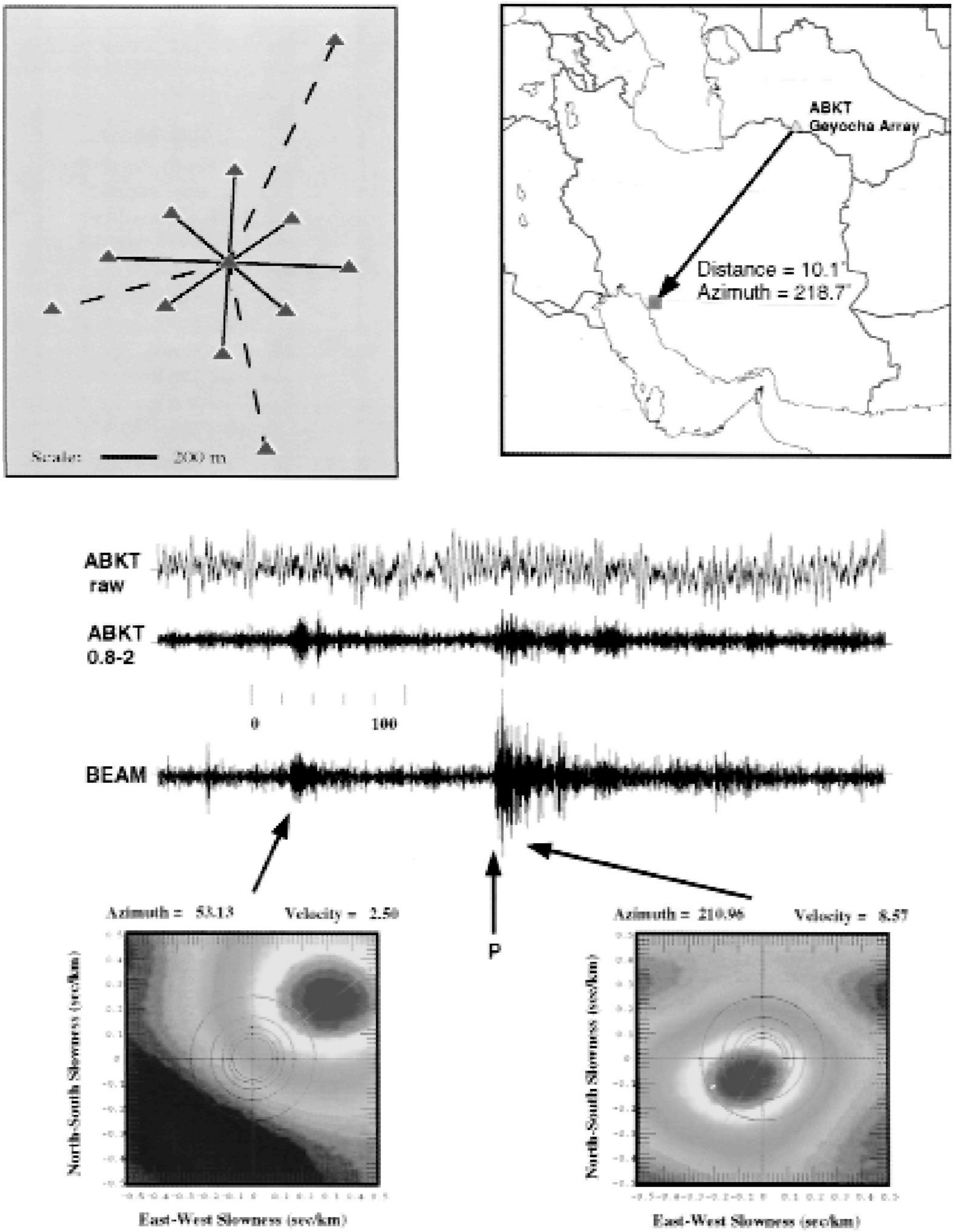

Vol. 17 – PLATE 1. Example of array processing. These results were obtained from an it experimental broadband array near the city of Ashgabat, Turkmenistan. The time series labeled ABKT is the signal on a single sensor. ABKT 0.8–2 is the same signal bandpass filtered between 0.8 and 2 Hz. The time series labeled BEAM is simply the undelayed sum of all elements.

SEISMOLOGY

Carl Kisslinger, *Department of Geological Sciences and Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, U.S.A.*

Introduction 155
1. **Seismic Source Mechanisms** 156
1.1 Geologic Faults 156
1.2 Energy Flow in an Earthquake: The Elastic-Rebound Model 157
2. **Seismic Waves** 157
2.1 Body Waves 157
2.2 Generalized Snell's Law 158
2.3 Surface Waves 159
2.4 Free Oscillations 160
2.5 Attenuation of Seismic Waves 161
2.6 Locating Earthquakes 161
3. **Focal Mechanisms from the Wave Radiation Pattern** 162
4. **Measures of Earthquake Size** 164
4.1 Intensity 164
4.2 Magnitude and Seismic Moment 164
5. **Earthquake Geography** 166
6. **Earthquake Hazard Assessment and Earthquake Prediction** 167
6.1 Hazard Assessment 168
6.2 Earthquake Prediction 169
7. **Applications of Seismology to Societal Needs** 171
Glossary 172
Works Cited 173
Further Reading 174

INTRODUCTION

Seismology is the study of ground shaking (from Greek: *seien* = to shake + *ology*). Ground vibrations from a variety of sources, natural (earthquakes, storms) and artificial (explosions, rock bursts in mines, machinery), have been investigated as part of this science. The term is now used primarily for the study of earthquakes and related phenomena. Modern seismology consists of two principal aspects: the study of earthquakes as a geophysical phenomenon and the study of the internal structure of the Earth on the basis of data provided by the observations of waves generated by seismic sources (see Earth, Interior Structure of). For studies of earth structure, the data may come from natural earthquakes or from artificial sources, such as explosions.

Theoretical seismology is based primarily on the mechanics of continuous media, especially solid materials (see Mechanical Properties of Solids; Mechanical and Elastic Waves in Solids; Acoustics, Linear). The mechanics of failure in solids under stress and the theory of wave propagation in solids and liquids provide the foundation. The materials, in the general case, are lossy and possibly elastically anisotropic.

An earthquake may be defined as a vibration of the Earth's surface caused by a transient disturbance of the equilibrium of rocks near or beneath the surface (Macelwane, 1947, p. 68; Kostrov and Das, 1988). This definition excludes more or less steady-state sources of observable vibrations, such as storms over the ocean. However, the study of the continuous waves generated by such sources, called microseisms, is a traditional topic in seismology. The definition admits sources like explosions as "artificial earthquakes" (Richter, 1958, p. 151).

The point within the Earth at which the disturbance occurs is called the *hypocenter* or *focus* of the earthquake. Because earthquake sources are spatially extended, and in view of the way in which earthquakes are routinely located from instrumental data, the term hypocenter is usually interpreted as the point at which the disturbance initiates. The point on the Earth's surface vertically above

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

the hypocenter is the *epicenter.* The *epicentral distance* to another point on the surface is measured as the arc of the great circle connecting the two points and may be expressed in either linear units (km) or angular units (deg).

1. SEISMIC SOURCE MECHANISMS

As understanding of Earth physics evolved, various energetic transient phenomena were proposed as the cause of earthquakes. Collapse of underground cavities, impacting meteorites, and processes within the chambers of volcanoes (see VOLCANICS) all produce measurable ground motion that fits the definition of "earthquake." The total energy released globally in such processes is negligible compared with that from the dominant cause, the sudden slip of rock masses along geologic faults. Earthquakes caused by fault slip are called *tectonic,* i.e., caused by the forces acting in the interior of the Earth to form geologic structures. More than a century of analysis of field and instrumental observations supports the hypothesis that almost all earthquakes, and all strong ones, are of tectonic origin (Scholz, 1990).

1.1 Geologic Faults

A fault is the interface between two rock masses along which there has been tangential displacement of one side relative to the other. The fault surface is pictured as a plane that intersects the Earth's surface along a line called the *fault trace.* Faults in nature may have complex geometries, which can be represented by a combination of planes with varying orientations. The spatial orientation of a fault surface is given by two angles: the dip, the angle measured downward between the horizontal plane and the fault plane, and the strike, the azimuth of the fault trace measured from the north. The mode of faulting is described by the direction of the slip vector, which lies in the fault plane and describes the motion of the block on one side of the fault relative to the other side. *Dip slip* occurs when the slip vector is parallel to the dip direction, and may be *normal* faulting, in which the block above the fault plane (the hanging wall) moves down relative to the block below (the foot wall), or *reverse,* in which the hanging wall moves up. Reverse faulting on a fault with small dip is called *thrust* faulting. *Strike slip* occurs when the slip vector is horizontal, parallel to the strike of the fault. The fault slip during an earthquake is often a combination of dip and strike slip.

The mode of faulting is directly related to the orientation of the ambient principal tectonic stresses. When the maximum compressive stress (the P axis) is vertical and the minimum compressive stress (the T axis) is horizontal (extensional tectonics), normal faulting dominates. Reverse faulting occurs when the maximum compressive stress is horizontal and the minimum compressive stress is vertical (compressional tectonics). When the maximum and minimum principal stresses are both horizontal, strike-slip faulting results. The axis of the intermediate principal stress lies in the fault plane. It follows that if the orientation of the plane on which slip occurred and the slip direction can be determined from either field or instrumental observations of an earthquake, important information about the regional stress field can be recovered.

A set of values of strike, dip, and slip direction is called a *focal-mechanism solution* for the earthquake. The determination of these basic geometric parameters is the first step in the analysis of the process that caused an event. Other characteristics of the fault can be derived from the recorded ground motion, such as the length and width of the surface that ruptured, the speed with which the rupture propagated across the surface, and the distribution of slip on the surface.

Few earthquakes, usually only strong, shallow ones on land, produce surface evidence of the faulting from which a focal mechanism can be developed from observations in the field. Several procedures have been developed for extracting this information from the records of ground motion produced by the earthquake. They all depend on the known relations between the geometric properties of the source and the pattern of radiation of seismic waves. The fault-slip model of earthquake origin is accepted as valid, but variations of this model that include displacements other than simple shearing can be accommodated. It is necessary, therefore, to examine the properties of the waves generated by an earthquake before considering how they may be used to retrieve its focal mechanism. Because waves represent one form in which the energy released in an earthquake is accounted for, a consideration of the flow of energy in an earthquake is discussed first.

1.2 Energy Flow in an Earthquake: The Elastic-Rebound Model

The original form of the elastic-rebound model of earthquake occurrence was formulated by Reid (1910, pp. 29–31). Working from the detailed data provided by the investigation of the 1906 California earthquake, he proposed the following description of the flow of energy in the process of producing fault rupture:

1. The energy released in an earthquake is present just before the event in the form of elastic strain energy in the rocks around the fault.
2. This energy has accumulated slowly as the rocks deform under the action of tectonic forces. The origin of these forces was not known to Reid (1910, pp. 27–28) nor understood until the development of the theory of plate tectonics (*q.v.*).
3. The strain accumulates until the resulting stresses on the fault exceed the shear strength of the fault surface, at which time sudden slip occurs to produce the earthquake. The rocks surrounding the fault move toward a position of minimum strain (rebound).

The fundamental concept that seismic energy is stored slowly by elastic deformation and released suddenly by fault rupture has held up well with time. The details of the model have been modified by better understanding of the material properties in fault zones and advances in theoretical fracture mechanics (Kostrov and Das, 1988; Scholz, 1990).

If the strain field can be represented by homogeneous simple shear, the energy per unit volume available to produce an earthquake is $e = \frac{1}{2}\mu\gamma^2$, where μ is the shear modulus and γ is the angle of shear. For typical crustal values of the shear modulus and of the shear strains associated with earthquakes, e is of the order of 100 J/m^3.

Several mechanisms are available to absorb the energy released in the event. Energy goes into the creation of a newly fractured surface and possibly the uplift of rock masses, energy appears as heat due to the friction on the sliding surfaces, and energy is radiated away from the source in the form of waves. The energy budget of an earthquake is not well determined, but a large fraction of the released energy goes into waves.

2. SEISMIC WAVES

2.1 Body Waves

The theory of waves in solids demonstrates that two types of waves can exist. The well-known differential equation expressing motion with infinitesimal strains in an unbounded perfectly elastic, isotropic body governs the behavior of these waves:

$$\frac{\partial^2 \mathbf{u}}{\partial t^2} = \mathbf{f} + \frac{(k + \frac{4}{3}\mu)}{\rho}\nabla\nabla\cdot\mathbf{u} - \frac{\mu}{\rho}\nabla \times (\nabla \times \mathbf{u}), \tag{1}$$

where $\mathbf{u}(\mathbf{x},t)$ is the particle displacement, $\mathbf{f}$ is the body force per unit mass, k is the incompressibility modulus, μ is the shear modulus, and ρ is the density of the medium (Aki and Richards, 1980).

The two wave types are the manifestation of the propagation from the source of the two fundamental types of deformation, volumetric strain, given by $\nabla\cdot\mathbf{u}$, and shear strain, given by $\nabla \times \mathbf{u}$. The former, termed compressional waves, gives rise to waves that are longitudinally polarized in an isotropic medium, labeled P waves (for primary) in seismology. Shear deformation propagates as shear waves, transversely polarized, labeled S waves (for secondary). In a perfectly elastic, isotropic solid, these waves travel with the velocities

$$V_P = \sqrt{(k + \tfrac{4}{3}\mu)/\rho} \qquad V_S = \sqrt{\mu/\rho}. \tag{2}$$

It may be seen from these relations that the P-wave velocity is greater than the S-wave velocity in any material; hence the designations primary and secondary to describe the order of their arrival at an observing point. The ratio V_P/V_S in common crystalline rocks is about 1.8. It is also seen that shear waves travel only in solid materials, not in fluids, for which the shear modulus is zero.

The situation is more complicated for anisotropic media, in that the polarization is no longer strictly longitudinal or transverse, the wave velocity depends on the direction of propagation, and the velocity of *S* waves depends on the polarization as well as the propagation direction (Musgrave, 1970). Anisotropy does exist in parts of the Earth's crust and upper mantle, and may exist at greater depths. In the following, we assume isotropy.

2.2 Generalized Snell's Law

A complexity encountered in seismology that does not occur in acoustics or electromagnetic wave studies is the conversion of part of the energy in an incident wave of either *P* or *S* type into the other type at boundaries between materials of different properties. The converted waves may be both reflected back from the boundary and transmitted through it. Thus, a wave of either type will, in general depending on the angle of incidence and the contrast in properties of the two materials, give rise to four waves at a boundary, a reflected and transmitted wave of the same type as well as the converted waves. For a *P* wave incident in medium 1, angle of incidence i_{inc}, the directions of propagation of the four derived waves are given by a generalization of Snell's law (see MECHANICAL AND ELASTIC WAVES IN SOLIDS):

$$\frac{\sin i_{\text{inc}}}{V_{P1}} = \frac{\sin i_1}{V_{P1}} = \frac{\sin j_1}{V_{S1}} = \frac{\sin i_2}{V_{P2}} = \frac{\sin j_2}{V_{S2}}. \qquad (3)$$

In Eq. (3), all angles are measured between the normal to the boundary and the ray. The angle i_1 is the angle of reflection of the *P* wave and is equal to i_{inc}; j_1 is the angle of reflection of the *S* wave converted from the incident *P* wave. The angle i_2 is the angle of refraction of the transmitted *P* wave; j_2 is the angle of refraction of the transmitted *S* wave. For an incident *S* wave, i_{inc} is replaced by j_{inc} and V_{P1} by V_{S1} in the first term. The reflection and transmission coefficients for each of the four derived waves, as functions of the angle of incidence and contrast in material properties, are given in Aki and Richards (1980, pp. 133–151).

One consequence of these relations is that a critical angle of incidence exists whenever the velocity of the incident wave is less than the velocity of the derived ray. Because $V_S < V_P$ in all materials, a critical angle for *S*-to-*P* conversion by reflection always exists, for which $\sin j_{\text{inc}} = V_S/V_P$. A frequently occurring case in the Earth is $V_{P1} < V_{P2}$ for waves incident on a boundary from above. Here, for $i_{\text{inc}} = \sin^{-1}(V_{P1}/V_{P2})$, the critically refracted energy travels along the boundary with velocity V_{P2} and radiates energy back into the first medium at the critical angle. These *head waves* can be quite prominent and provide a powerful means of mapping subsurface structure.

In an unbounded, homogeneous, perfectly elastic Earth, the signal from an earthquake would consist of a *P* wave followed by an *S* wave, and nothing else. As seen in Fig. 1, an actual seismogram is much more complex. Much of the complexity is due to the reflection, refraction, and conversion from one type to the other of the original *P* and *S* waves at boundaries between layers of material with differing properties. The dominant wave arrivals on seismograms depend on the epicentral distance of the station. Local and regional records show primarily high-frequency phases that have traveled in the layers of the crust. At intermediate distances, 20° to 80°, body waves that have traveled through the mantle along various paths dominate. At greater distances, waves that have traveled through the core become important.

The development of the wave train with increasing distance is illustrated in Fig. 2, in which a few of the many distinct arrivals are identified. The initial *P* arrival at station ZOBO is very small because at this distance the wave has encountered the core of the Earth and the arrival seen has been diffracted at the core–mantle boundary. The arrival at station CTAO marked SKS is a wave that left the source as an *S* wave, entered the core as a *P* wave, then left the core after conversion back to an *S* wave. An epicentral distance of about 80° is the shortest at which a wave that has traveled part of its path in the core arrives. The travel time of the direct *P* wave through the center of the Earth to the antipodes is about 20 minutes.

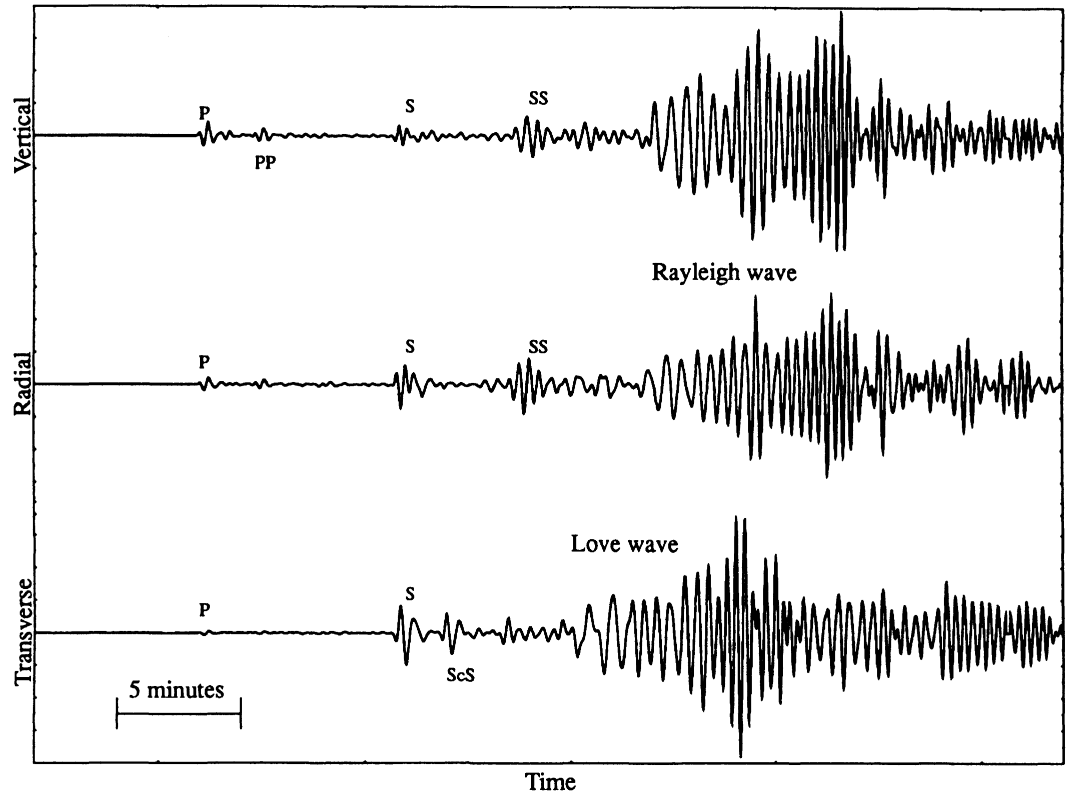

FIG. 1. Long-period seismograms recorded at Glen Almond, Quebec, Canada, from an earthquake in the Fox Islands, Alaska. The earthquake occurred on 5 January 1987, with body-wave magnitude m_b = 6.0, surface-wave magnitude M_S = 6.6. The epicentral distance to the station was 57.6° or 6400 km. The horizontal components have been rotated to correspond to radial and transverse motion relative to the great circle from the epicenter to the station. The body-wave arrivals labeled are as follows: the direct *P* and *S* wave from the source; PP and SS, *P* and *S* waves reflected one time from the Earth's surface at the midpoint between the epicenter and the station; ScS, *S* wave reflected from the core–mantle boundary. The dispersion in the surface-wave trains is seen. The Rayleigh-wave train is seen on the vertical and radial components, the Love-wave train on the transverse component. Digital wave-form data provided by the National Earthquake Information Center, U.S. Geological Survey.

The boundaries marking the most pronounced contrasts in properties are at the Earth's outer surface, rock against air (PP,SS in Fig. 1), and at the core–mantle boundary, solid silicates against melted iron alloy (ScS in Fig. 1) (see EARTH, INTERIOR STRUCTURE OF THE). Strong signals are received as reflections, including multiple reflections, from these boundaries, as well as from waves that have been refracted through the core. Because no *S* waves that have traveled directly through the core can be detected on seismograms, it is concluded that the core is, at least in part, fluid. Numerous other boundaries marking less sharp changes in properties have been identified from their signatures in the seismic wave train. The most detailed information we have about the internal constitution of the Earth has come from the determination of *P*- and *S*-wave velocities at all depths within the Earth by analysis of the observed travel times of wave energy along the many possible paths.

2.3 Surface Waves

In addition to the body waves that travel through the crust, mantle, and core, wave energy also becomes trapped in waveguides and spreads out in two dimensions. Prominent waveguides are formed by the Earth's outer surface in combination with layers in

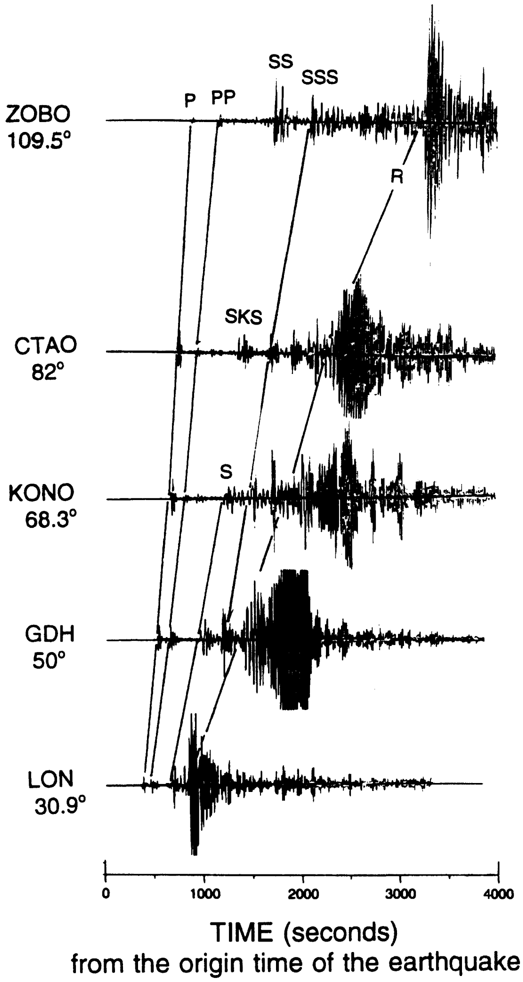

FIG. 2. The long-period vertical-component seismograms from the same earthquake shown in Fig. 1 over a large range of epicentral distances, given in degrees of arc below the station designator. The stations are Longmire, Washington, LON; Godhaven, Greenland, GDH; Kongsberg, Norway, KONO; Charter Towers, Australia, CTAO; and Zongo Valley, Bolivia, ZOBO. The amplitude scale at each station is arbitrary and has been adjusted for convenience in plotting. Only a few of the many phase arrivals are identified. The small-*P* arrival at ZOBO has been diffracted at the core–mantle boundary. CTAO is at about the shortest distance from the epicenter at which a wave that has traveled part of its path in the core, SKS, arrives. The time between phase arrivals at a station is a function of its epicentral distance. Digital wave-form data provided by the National Earthquake Information Center, U.S. Geological Survey.

the crust and upper mantle. The resulting waves, called *surface waves,* produce the largest signals on typical seismograms of earthquakes with hypocentral depths less than 100 km, Figs. 1 and 2. Surface waves are of two general types, one polarized in the plane of incidence (vertical and radial components of displacement), the other polarized horizontally, with displacement perpendicular to the plane of incidence. The former are called *Rayleigh waves,* the latter *Love waves.* Surface waves propagate with little attenuation in the waveguide, and die off exponentially with depth outside the waveguide (Bullen and Bolt, 1985).

Surface waves are *dispersed,* so that the wave velocities depend on the wavelength. Observations of either phase or group velocity as a function of wave period are useful for deriving the Earth structure that controls the propagation along the path from source to seismograph station. Each surface-wave type propagates in an infinite number of modes, manifested as the number of nodal planes of the motion within the waveguide. Love-wave velocities are controlled by the densities and shear-wave velocities and the thickness of the layers making up the waveguide. Rayleigh-wave velocities depend on these variables, as well as on the *P* wave velocities.

2.4 Free Oscillations

In addition to the propagating body and surface waves, earthquakes may also generate *free oscillations* of the Earth, i.e., the normal modes of vibration of the planet as a bounded, layered elastic body (Bullen and Bolt, 1985). Free oscillations were first detected only from extremely powerful earthquakes, but with modern broadband seismographs (see SEISMOGRAPHS) they may be seen from events as small as magnitude 6 (see Magnitude, Sec. 4.2).

Free oscillations are of two fundamental types, spheroidal and toroidal, corresponding to the two fundamental types of surface waves. Spheroidal modes may be recorded on instruments sensitive to the vertical component of motion, including gravimeters, while toroidal modes are found only on the horizontal components. Free oscillations are studied by computing the frequency spectrum of a long sample of the ground motion following an earthquake. They show up as a sequence of spectral lines at well-defined frequencies.

The frequencies of many identified modes have been measured. The theory of normal modes of an elastic body is used to calculate the expected frequencies for a given model of the Earth, specified in terms of a distribution with radius of the density and elastic moduli. The model can then be adjusted to give the best fit to the suite of measured frequencies.

The width of a given spectral line is determined by the anelastic attenuation in the Earth of the particular mode. This provides a powerful tool for exploring anelastic properties within the Earth. The spectral lines are converted into multiplets by the rotation of the Earth and by its ellipticity, and so these effects are taken into account in the calculations.

2.5 Attenuation of Seismic Waves

The amplitudes of seismic waves decrease with distance traveled, primarily because of geometric spreading. For all except plane waves, the amplitude in a perfectly elastic medium must decrease as the wave front expands, in such a way that the total energy flux through the wave front is conserved. Because the energy flux per unit area of the wave front varies as the square of the amplitude, this purely geometrical effect is easily accounted for. The amplitude of a spherically spreading wave decreases as r^{-1}, while a cylindrically spreading wave amplitude decreases as $r^{-1/2}$.

Other causes of seismic-wave attenuation are elastic scattering from inhomogeneities in the medium and anelastic attenuation. Shear-wave energy removed from the advancing wave front by scattering is thought to be the principal source of the short-period coda on seismograms of local and regional earthquakes. Anelastic attenuation is of fundamental importance, because the departures from elasticity of the materials within the Earth provide an important clue to composition and ambient conditions at depth (Aki and Richards, 1980, pp. 167–185).

A number of processes have been proposed as the physical basis for anelastic attenuation in the Earth (Jackson and Anderson, 1970). Attenuation is measured by the quality factor, Q, defined by the fractional loss of energy per cycle as

$$2\pi Q^{-1} = -\Delta E/E. \tag{4}$$

Q is directly related to the viscous and elastic parameters in any of the various models of viscoelastic materials that have been analyzed (e.g., Maxwell solid, Voigt solid, standard linear solid).

The decrease in amplitude with distance traveled, for a wave traveling with velocity V, with geometric spreading removed, may be expressed as $\exp[-\pi fr/VQ]$. The decrease of amplitude with time for a standing wave, such as a free oscillation, is $\exp[-\pi ft/Q]$. Q may be measured either from observations of amplitudes of propagating waves, or from the linewidths in the spectra of free oscillations. A measure of attenuation can also be derived from the rate of decay of the coda of earthquakes recorded at small distances. On the assumption that this coda consists of single backscattered shear waves, a measure called the coda Q is calculated.

The frequency dependence of Q is a fundamental question of seismic-wave propagation (Anderson and Given, 1982). Laboratory and field studies of the attenuation of P and S body waves, surface waves, and free oscillation provide the basic data. Q for both P and S waves is essentially independent of frequency over a wide portion of the seismic spectrum and increases outside of that band, so that the Earth behaves as an absorption-band filter. Anderson and Given (1982) presented evidence that the characteristics of the absorption band vary with depth within the Earth.

Anelastic attenuation introduces frequency dependence in the velocities of seismic waves. Therefore, Earth models derived from the velocities of waves with greatly different frequencies can be reconciled only by taking into account the dispersion introduced by anelasticity.

2.6 Locating Earthquakes

Earthquakes are located by analysis of the time of arrival of the various waves at seismograph stations. Hundreds of such stations have been established at known positions on the Earth. The mean travel times of seismic waves along the many possible paths through the Earth as functions of epicentral distance and hypocentral depth have been established after more than a century of observations. The location process consists of adjusting the latitude, longitude, and depth of the hypocenter and the origin time of the event so that the observed arrival times, especially of the direct *P* and *S* waves, at all of the stations that recorded the event are in the best agreement, in the least-squares sense, with the times calculated from the standard travel-time curves. Adjustments to the standard travel-time curves to take into account local or regional variations in wave velocities from the standard Earth model at the source or at the recording stations may be required for accurate solutions. Solutions are compiled into global earthquake catalogs, such as the Bulletin of the International Seismological Center and the Preliminary Determinations of Epicenters of the U.S. Geological Survey, as well as national and regional catalogs (Bullen and Bolt, 1985, Chap. 10).

3. FOCAL MECHANISMS FROM THE WAVE RADIATION PATTERNS

Slip on a fault is accepted as the cause of almost all earthquakes, yet very few earthquakes produce surface slip that can be observed in the field. The fault origin is accepted because the pattern of radiation of seismic waves, as recorded by seismographs, can, for most earthquakes, be quantitatively explained by this model. The polarity and amplitude of the waves recorded at a station depend on the distance and azimuth of the station from the epicenter, which are known once the earthquake is located, and the dip, strike, and slip direction of the fault, the unknowns for which a solution is sought. The equations describing the radiation patterns of both body and surface waves are given in Aki and Richards (1980, Chaps. 4, 7).

The standard method, still widely used for routine investigations, is based on the distribution of polarities of the first motion of *P* waves at many points on the Earth's surface. Because it is the first-arriving wave, the polarity of the *P* wave (compression or dilatation) from sufficiently strong earthquakes can be read with less ambiguity than for later arrivals.

The analysis begins by approximating the ruptured fault as a point source of the waves. The length of the actual fault and the speed of rupture propagation can later be estimated by extending the analysis. The point source can be modeled as a dislocation or by the equivalent distribution of body forces. For a pure shear dislocation, the body force equivalent is a double couple without moment. The pattern of radiation of the *P*-wave displacement u_p at an observing point at $\mathbf{R}$ from this type of source located at the origin is given by

$$u_p(\mathbf{R},t) = \frac{2(\boldsymbol{\delta}\cdot\mathbf{n})(\boldsymbol{\delta}\cdot\mathbf{V})\mu A\delta}{4\pi\rho V_p^3 r}, \tag{5}$$

where $\boldsymbol{\delta}$ is the unit vector in the direction from the source to the observing point, $\mathbf{n}$ is unit normal to the fault plane, $\mathbf{V}$ is the average vector velocity of slip on the fault plane, A is the area of the fault, and r is the distance to the point of observation (Aki and Richards, 1980, p. 113).

For an observing point in the $(\mathbf{n},\mathbf{V})$ plane, the first factor in the radiation pattern is $\sin\alpha$, where α is the angle between the slip vector and the direction to the station; the second factor is proportional to $\cos\alpha$. The result is a four-lobed radiation pattern of alternating compressions and dilatations, separated by nodal lines at $\alpha = 0$ (along the strike of the fault) and $\alpha = 90°$. When embedded in the Earth with an arbitrary orientation, the pattern divides the planet into four segments, separated by two nodal planes. One of the planes is the fault, the other, the auxiliary plane, normal to the first.

To apply this result, the earthquake must first be located. Then, a small "focal sphere" is imagined to surround the (point) source. The earthquake energy leaves the focal sphere along ray paths that are refracted through the Earth's interior to emerge at points distributed over its surface. From the known distribution of wave velocity with depth, the ray emerging at any point on the Earth's surface can be traced back to the point at which it left the focal sphere. The corresponding direction of the first motion, compression or dilatation, is then assigned to that point. By convention, the distribution of first motions on the lower hemisphere of the focal sphere is mapped by equal-area stereographic projection. The analyst then decides the position of the two nodal curves, constrained to be orthogonal, that separate the quadrants of compressions and dilatations. The two nodal curves are maps of the intersection of the two nodal planes with the focal sphere. The strike, dip, and direction of slip on each of these two planes can then be read from the resulting graph (Lee and Stewart, 1981, pp. 139–148). The entire process is now usually done by computer.

The resulting focal-mechanism solution is fundamentally ambiguous, in that there is no way, in principle, to decide from these data which of the two nodal planes is the fault plane and which the auxiliary plane. Independent information, such as knowledge of the dominant structural trends in the region and the distribution of aftershocks of the event, can help resolve the ambiguity.

Examples of focal-mechanism solutions obtained for one month, August 1992, by the U.S. Geological Survey are plotted at the positions of the earthquake epicenters in Fig. 3. The type of faulting and regional stress patterns associated with various tectonic settings can be inferred from such maps.

Fault-plane solutions do yield useful information about the orientation of the principal stress axes that correspond to the solution. The axis of maximum compression (*P* axis)

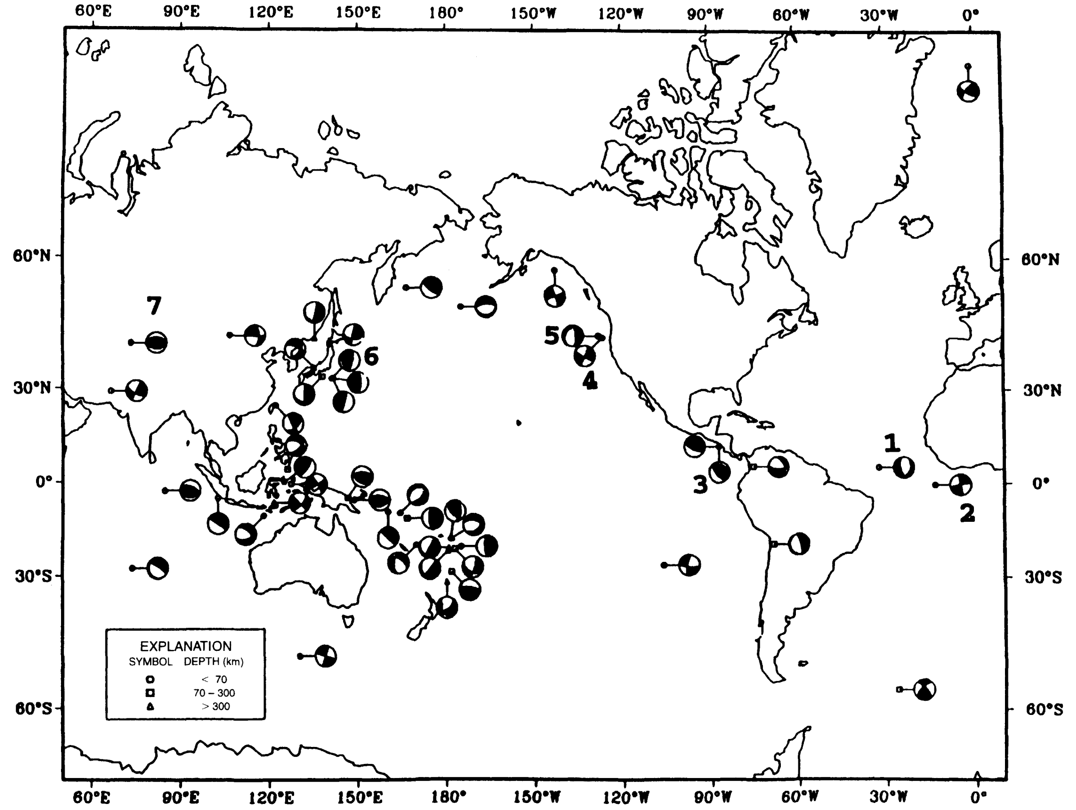

FIG. 3. A world map with focal mechanisms of earthquakes that occurred during August 1992. Each solution is connected to the epicenter of the earthquake by a line segment. The circles are equal-area stereographic maps of the lower hemisphere of the focal sphere. The two nodal planes of each solution, one of which is the fault plane, and the quadrants of compressional first motions (black) and dilatational first motions (white) are shown. A few typical mechanisms are marked with numbers: 1 is a normal fault on the Mid-Atlantic Ridge, 2 is a strike-slip earthquake on a transform fault that offsets segments of the Mid-Atlantic Ridge, 3 is a thrust event at the Middle America subduction zone, 4 and 5 are a strike-slip and normal fault, respectively, associated with the undersea ridge offshore of northern California, 6 is a thrust event associated with subduction of the Pacific Plate under Eurasia at the Japan Trench, and 7 is an intraplate thrust event, reflecting the north–south compression resulting from the collision of the Indian subcontinent into the Asian land mass. From U.S. Geological Survey, Preliminary Determination of Epicenters, Monthly Listing, August 1992.

lies in the center of the quadrant of dilatations, the axis of minimum compression (T axis) lies in the center of the quadrant of compressional first motions, and the axis of intermediate stress lies along the intersection of the two nodal planes. The directions may not be the same as the orientation of the current regional tectonic stress field, as the rupture may have occurred along some previously established plane of weakness in the rocks at an arbitrary angle to that stress field. Techniques have been developed by which the focal-mechanism solutions of many earthquakes in a region can be inverted for the contemporary stress field that best fits the ensemble of data (Gephardt and Forsyth, 1984).

Occasionally the distributions of first motions cannot be separated into four quadrants by great circles on the focal sphere. The implication is that the radiation is not described by Eq. (5), that is, the source was not equivalent to a simple double couple, but also contained mutually orthogonal components of extensional and compressional displacement. Solutions are obtained in this case by extended analysis, on the assumption that the net volumetric change at the source is zero.

4. MEASURES OF EARTHQUAKE SIZE

4.1 Intensity

The size or strength of an earthquake may be measured in several ways. The oldest measure is *intensity,* which is expressed in terms of the visible effects of the earthquake on natural and manmade features (Richter, 1958, Chap. 11). Descriptions of earthquake strength in terms of effects such as human perception, landslides, ground displacement, and damage to buildings were a natural preinstrumental approach. Intensity is still determined and mapped because of its direct applicability to earthquake-resistant design of structures. Several intensity scales have been devised, each of which assigns a numerical value to each level of damage or other observable effects. The most commonly used scale in the United States is the Modified Mercalli scale, which ranges from I (felt by only a few persons under favorable conditions) to XII (total destruction). Slight damage occurs at MM V (Roman numerals are conventionally used for intensities to avoid confusion with earthquake magnitude). The distribution of intensities is often plotted on an *isoseismal map,* on which areas in which the intensity is judged to be constant are separated by contours called *isoseismals.*

Intensity is a useful measure for engineering purposes and because it is possible to assign distributions of intensities to earthquakes from the preinstrumental era from written descriptions of the effects. Intensity is, however, of limited value in the scientific study of earthquakes. Earthquakes in unpopulated areas, including the numerous earthquakes under the oceans, cannot be assigned an intensity on the basis of observed effects. Also, the assignment of an intensity value is often subjective, so that different observers may report somewhat different values. For these reasons and because of the need for an objective universal measure of earthquake strength that could be used for statistical analysis of earthquake catalogs, C. F. Richter developed the magnitude scale in 1935.

4.2 Magnitude and Seismic Moment

The *magnitude* of an earthquake is a measure of its strength based on the recorded amplitude of a specified wave type (Richter, 1958, Chap. 22). The original Richter scale was based on the maximum wave amplitude produced by the earthquake, as recorded on a standard seismograph (fixed magnification, period, and damping) then used in the network of stations operated by the California Institute of Technology. The scale was intended for use only for local and regional earthquakes (epicentral distances less than about 600 km) in southern California, but was so successful that the basic idea has been extended for global use.

The logarithm (base 10) of the amplitude in micrometers of the largest wave on the seismograms was taken as the measure. The logarithm was used because of the vast range of strengths of earthquakes. Because wave amplitudes decrease with distance from the source, it was necessary to adjust the observed values to what they would have been at some standard epicentral distance, chosen as 100 km. Richter developed an empirical attenuation curve of wave amplitude vs distance for southern California, so that this adjustment could be made. There remained only to set the zero value on the scale. Richter decided that an event that produced a maximum recorded amplitude on the standard seismograph of 10^3 μm at 100 km would be a magnitude 3 earthquake.

Magnitudes determined by the original process, using data equivalent to those from the original standard seismograph, are called *local magnitudes,* M_L. The concept was so successful that scales applicable to seismograms recorded with any calibrated seismograph, at any distance from the epicenter, were subsequently developed. The *surface-wave magnitude,* M_S, is calculated from the logarithm of the largest amplitude in the Rayleigh-wave train at periods near 20 s. The recorded amplitude is converted into true ground amplitude by use of the known magnification of the instrument. Standard correction is made for epicentral distance.

Because surface waves are not strongly excited by earthquakes deeper than about 100 km, the surface-wave magnitude scale is of limited applicability. Therefore, Gutenberg and Richter, in 1956, developed a scale based on the seismic body waves, which are generated by all earthquakes. The most widely used scale is the *body-wave magnitude,* m_b, based on the amplitude of *P* waves with periods near 1 s. This scale is based on the logarithm of the ratio of amplitude of the ground motion (in micrometers) to period (in seconds) of the largest amplitude in the first few cycles of the recorded *P* wave. A correction is made for epicentral distance and hypocentral depth.

All of the magnitude scales discussed so far saturate at large magnitudes. That is, the energy release in the earthquake increases without a corresponding increase in the measured magnitude. Saturation is a consequence of the shape of the typical earthquake amplitude–frequency spectrum. The spectrum, shown schematically in Fig. 4, is characterized by a flat segment at low frequencies, a corner frequency, which depends on the physical dimensions of the fault segment that ruptured, and a high-frequency falloff. The corner frequency moves to lower values as the size of the fault and the energy released in the earthquake increase. If the earthquake is so big that the corner frequency is below the frequency of the waves used to calculate magnitude, saturation occurs. A solution to this problem is the use of the flat, low-frequency part of the spectrum, as proposed by Kanamori (1977). The very low frequencies involved were not well recorded by earlier seismographs, but are by modern broadband instruments. The level of the low-frequency part of the spectrum is proportional to the seismic moment (see below), and the magnitude scale based on the moment, called M_W by Kanamori, is now widely used.

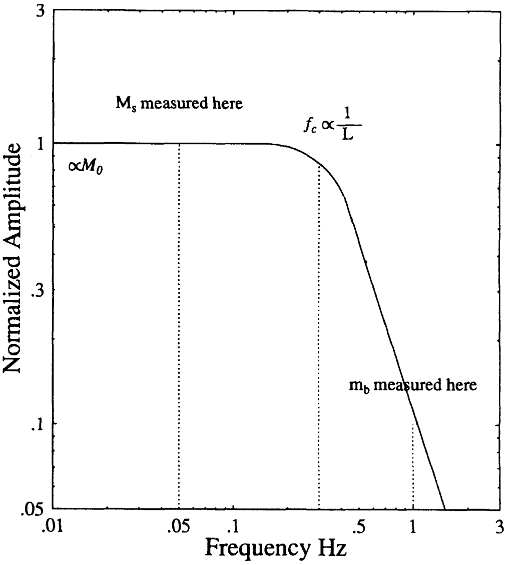

FIG. 4. Schematic earthquake displacement spectrum. The details of the theoretical spectrum are model dependent, but the main features are a flat low-frequency part that is proportional to the seismic moment, a corner frequency f_c that is inversely proportional to the characteristic dimension *L* of the rupture surface, and a high-frequency falloff, usually taken as ω^{-2}, though ω^{-3} is preferred by some seismologists. The body-wave magnitude is measured at about 1 Hz, the surface-wave magnitude at about 0.05 Hz. As the earthquake dimension increases, the corner frequency moves to lower frequencies. If the corner frequency falls below the frequency at which a particular magnitude is measured, the magnitude will not increase in accord with the increase in energy or moment. That magnitude scale is saturated.

An increase of one unit on any of the magnitude scales signifies an increase by a factor of 10 of the amplitude of the wave used. However, studies of the relation of wave amplitude and energy release have led to the empirical relation $\log E$ (joules) = $1.5M_S + 4.8$, so that an increase in M_S of one unit corresponds to an increase in released seismic energy by a factor of about 30. Thus, it takes almost 1000 M_S = 4 earthquakes to release the energy in one M_S = 6.

No maximum or minimum possible magnitude follows from the definition of the scale. Nature seems to have an upper limit of about M_W = 9, presumably governed by the maximum amount of strain energy rocks are capable of storing and the maximum dimensions of fault rupture surfaces.

The *seismic moment* is a measure of earthquake strength based on the fault-slip model of earthquake generation that is more directly related to the physics of the source than either intensity or magnitude. The seismic moment is defined as $M_0 = \mu AD$, where μ is the shear modulus of the material in which the fault occurs, A is the area of the rupture surface, and D is the average displacement on the surface during the earthquake (Aki and Richards, 1980, p. 49). The seismic moment, in newton meters, is related to the moment magnitude, M_W (Hanks and Kanamori, 1979), by $\log M_0 = 1.5M_W + 9.05$, so that typical values would be a seismic moment of 1.1×10^{18} N m for a moment magnitude 6 earthquake.

5. EARTHQUAKE GEOGRAPHY

The geographic distribution of earthquakes is basic information about the Earth that is to be interpreted in terms of the dynamic processes that shape major geologic features. Figure 5 shows the global distribution of earthquakes for the years 1981–1990. One conclusion from this map is that earthquakes are not randomly or uniformly distributed over the planet, but that most occur in distinct, well-defined seismic zones or belts. This information is essential in assessing earthquake hazards, in that we know where the probability of occurrence of strong earthquakes is high and where they are rare or may never occur. The broad features of the distribution are revealed by observations during a short time, like ten years. However, a short interval will not yield information about some significant characteristics. The strongest earthquake to be expected at any site and the mean time between events of a given magnitude cannot be determined. Also, some areas of infrequent, but significant seismicity may not show activity during the interval.

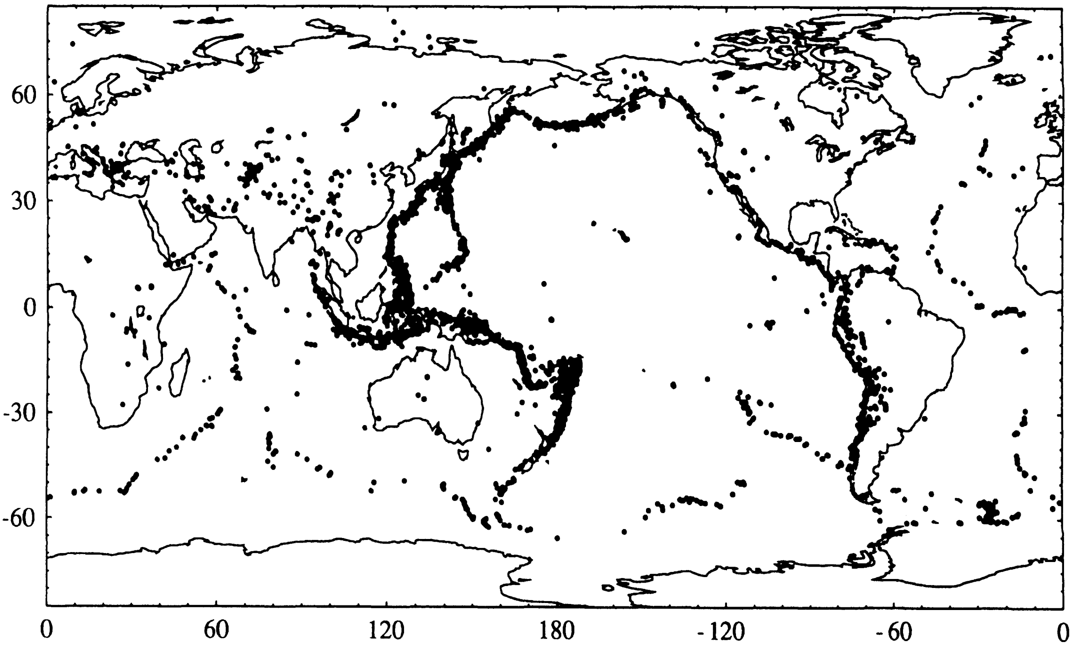

FIG. 5. Global distribution of earthquakes, 1981–1990. All events (3950) with magnitude greater than 5.5 are plotted. The data are taken from the catalog, Preliminary Determination of Epicenters, produced by the U.S. Geological Survey. At this rather high minimum magnitude some areas of lower-level activity, such as the eastern United States, show no activity during the selected decade.

The classical, pre-plate-tectonics designation of the principal concentrations of seismicity was as the Circum-Pacific, Alpide-Himalaya, and Oceanic Active Belts (Gutenberg and Richter, 1949). We now understand that those concentrations exist at the active boundaries between mobile lithospheric plates (Gubbius, 1990) (see PLATE TECTONICS). Some of the most compelling evidence in support of the plate-tectonics model came from the analysis of seismological observations. For example, focal-mechanism solutions of the strong earthquakes show that subduction zones are compressive regimes (thrust faulting) and undersea rifts are tensional regimes (normal faulting), Fig. 3. The sense of strike-slip on transform faults is that expected from the plate-tectonics model. Earthquakes at plate boundaries are called *interplate* earthquakes; those away from plate boundaries are *intraplate.*

The forces driving the plates and plate interactions provide the tectonic forces called for by the elastic-rebound theory. It is presumed that plate motions are the ultimate source of the stresses driving both interplate and intraplate earthquakes, but the relation is not as obvious for the latter. In particular, the reasons for the localization of intraplate events are often obscure. For example, the diffuse distributions of seismicity in eastern North America and in western China are almost certainly the consequence of the westward motion of North America from the Mid-Atlantic Ridge in the first case and the continuing intrusion of the Indian subcontinent into Asia in the second. The reasons for the local details of the stress distribution and the relation of the active sites to local and regional geologic features are, however, not well understood.

Events along the Pacific coast of the coterminous 48 United States are clearly related to the interaction between the Pacific/Juan de Fuca and North American plates. Events farther inland have a variety of focal mechanisms associated with the geologically diverse settings. Activity in the western intermountain zone, in the central Mississippi Valley, and at apparently isolated sites along the east coast do not easily fit into a plate-tectonics explanation. They are interpreted as the result of complex plate interactions over long periods of geologic time. The configuration of the faulting responsible for great earthquakes in the central Mississippi Valley in 1811–1812, and numerous smaller events since, has been clarified by detailed geophysical and geological studies. The problem is more difficult along the Atlantic coast because even strong earthquakes cannot be readily associated with individual faults or fault systems, even when faults in the vicinity of the activity have been mapped.

Earthquakes in Alaska, the largest within the territory of the United States, are directly caused by the relative motion of the Pacific and North American plates, primarily subduction along the Aleutian Trench and under southern Alaska. Earthquakes in Hawaii are concentrated under the "big island," Hawaii, and have a different kind of origin. Some are apparently caused by stresses in the upper part of the crust induced by pressure within the chambers and conduits of the volcanoes. Those on the south flank of Kilauea volcano are thought to be related to the seaward motion of the volcanic pile over the buried oceanic crust, also under the influence of volcano-induced stress and probably gravity (Klein *et al.,* 1987).

The geographic distribution of the zones of high seismicity is an obvious starting point for the assessment of earthquake hazards, one important application of seismology.

6. EARTHQUAKE HAZARD ASSESSMENT AND EARTHQUAKE PREDICTION

The quest for methods to foretell the occurrence of destructive earthquakes goes back to antiquity, but has been the subject of a major international scientific effort only since the mid-1960s. The distinction between hazard assessment and prediction is primarily with regard to the specificity of the projected time of occurrence. An earthquake hazard assessment is a statement of the probability of occurrence in a defined area within a given time interval of either an earthquake with magnitude equal to or exceeding a specified value, or the occurrence of earthquake-generated ground motion exceeding a specified level at a given site. In addition to general concern for public safety, efforts to develop improved hazard-assessment methodologies have been motivated by the need for earthquake-resistant design and construction of critical facilities, such as dams, hospitals, and power plants, the failure of which because of an earthquake might produce catastrophic effects.

An earthquake prediction is a statement of the time, place, and magnitude of a specific future earthquake, with estimates of the uncertainty of each of those parameters. No reliable, generally applicable methods for formulating such predictions have yet been developed, though many approaches have been proposed and tested, with some limited success. Because of differences in the kinds of observations used, it is convenient to classify predictions by the time to the predicted event: long-term (decades to a few years); intermediate-term (a few years to a few weeks); short-term (a few weeks to a few hours). Long-term prediction efforts have become almost the same as hazard assessments, except for the implication that the event is expected to occur near the end of the long interval rather than with equal likelihood at any time within the interval.

6.1 Hazard Assessment

Earthquake hazard assessment starts from knowledge of the geographic distribution of earthquakes (Reiter, 1990). From catalogs of earthquakes located by instruments or from historical records, and geological evidence of prehistorical earthquakes, estimates can be made of the most likely sites for future earthquakes and the largest earthquakes that can be expected in each seismic source region. Uncertainties in this process arise from the possible incompleteness of catalogs for the time interval they cover, the short historical record in many parts of the world, and limited understanding of earthquake genesis, particularly in intraplate settings.

The next step is the estimation of the mean time between occurrences of earthquakes of a given magnitude in a particular seismic source region. The empirical Gutenberg–Richter relation between $N(M)$, the number of earthquakes with magnitude equal to or greater than M, and the magnitude is used for this purpose: $\log_{10}N(M) = a - bM$ (Reiter, 1990, Chap. 10). This relation approximately describes seismicity distributions on scales from global to local, with the value of b near 1. The parameters a and b are calculated as the maximum-likelihood estimates from the available catalog. If the data are normalized with respect to time to 1 year, the reciprocal of $N(M)$ is the average time in years between events at or above magnitude M within the specified area. The longer the time interval covered by the catalog and the wider the range of included magnitudes, the more reliable are the estimates of the parameters.

A problem with the application of this form is the recent discovery that the relation is most likely not a straight line over all magnitudes. Extrapolation from the frequent smaller events to estimates of the recurrence interval of the largest ones may be in serious error. One kind of behavior is the leveling off of the relation at larger magnitudes, with a relatively large number of stronger events with about the same magnitude, and few or none larger (the characteristic earthquake case). In this case extrapolation leads to an underestimate of the frequency of the largest events. Another possibility is a sharp decrease in the number of larger events, so that extrapolation overestimates the frequency of the largest magnitudes. This type of behavior has been explained by the geometry of the faults, with the falloff occurring for events larger than the smallest one that breaks through the entire brittle, seismogenic layer.

The estimation of the strongest earthquake, $M_{\max}$, to be expected in a source region is a difficult problem (Reiter, 1990, Chap. 5). At what magnitude should the Gutenberg–Richter relation be terminated at its upper end? Experience demonstrates that taking the strongest known earthquake of the past as a measure of the strongest to be expected in the future is dangerous, especially in intraplate settings with short earthquake histories of only a few centuries. Mapping of the geologic structures in the region may reveal the dimensions of the largest units, the longest faults in particular, that are likely to participate in a single earthquake. For very long, through-going fault systems of 1000 km or more, it becomes necessary to identify barriers to slip, such as bends in the fault or offsets, that may serve to limit the length of any single break. A variety of statistical methods for estimating the extreme value of future magnitude from the known past events have been tested, but none has been proven to give reliable estimates.

Because of the limitations on hazard assessment imposed by the brevity of recorded human history relative to geologic time, efforts have been made to extend the record through *paleoseismic* investigations. Geologic evidence of prehistoric great earthquakes takes the form of preserved effects of such events. Under favorable circumstances, the amount of slip in the event and the time of its occurrence can be estimated.

A variety of evidence has been invoked for paleoseismic studies, from the offset of ancient shorelines of lakes to the interruption of the growth cycle of old trees on faults by tearing of their roots. Systematic searches for paleoseismic evidence on accessible faults are made by trenching. By careful mapping of the distribution of materials in the walls of a trench dug across or along a fault, the offset of layered beds may be detected, the beds disrupted in a single event identified, and direct evidence of strong ground shaking, such as deposits of sand liquefied by an ancient earthquake, seen. If the disturbed material in the trench walls includes organic matter, such as peat, the age of the disturbance can be determined radiometrically.

Trenching has been especially successful for the southern San Andreas fault in California (Sieh, 1984). The mean time interval between great earthquakes during the past 2000 years is now well determined. However, important for efforts at hazard assessment, clear evidence for clustering of large earthquakes in time has been found (Sieh *et al.*, 1989). Several groups of earthquakes have happened on the southern San Andreas fault with the time between individual events much shorter than the mean interevent time. Then a long time may pass before the next cluster begins. The implication is that the long-term mean time interval may not be a reliable input to probabilistic assessments of the occurrence of the next great earthquake, at least on this fault. The approach taken is to use the best-determined value of the mean time interval and incorporate the statistical uncertainty of this value in the assessment.

Once the probability of occurrence of an earthquake of a given magnitude within a seismic source region has been calculated, the estimation of the damage potential of that event in its surroundings is the final step in hazard assessment. Experience shows that damage depends strongly on local geologic conditions, especially the thickness and properties of poorly consolidated materials near the surface. For general hazard assessments, the peak value of some ground-motion parameter, such as peak acceleration or particle velocity, is estimated for hard-rock conditions as a function of distance from the event of specified magnitude. These estimates, which embody their own uncertainties, are made from empirical curves of ground motion vs distance that have been assembled for various geographic regions.

If recordings of strong ground motion from previous damaging events in the region are available, or if the use of such recordings from other regions can be justified, a response spectrum for the anticipated earthquake can be computed. This measure is the peak response (usually acceleration or velocity) of a simple oscillator with specified viscous damping at any time during the event as a function of the period of the oscillator. The response spectrum contains much more information about the damage potential than a single value of, say, peak acceleration.

The important effects of local ground conditions must be incorporated into the assessment process on a site-specific basis. A well-known example of the effects of near-surface geology in controlling damage is the pattern of destruction in Mexico City as a result of the Michoacán earthquake, 19 September 1985. The heavy damage and loss of life were concentrated hundreds of kilometers from the epicenter, with almost no damage in between. The severe effects were the direct result of the amplification of the ground motion in the unconsolidated lake-bottom sediments on which the damaged part of Mexico City is built. The heavy damage caused in the Marina District of San Francisco by the Loma Prieta, California, earthquake, 18 October (UTC) 1989, is further evidence of local-site effects. The damage at a distance of 100 km from the epicenter was due to a variety of responses, including soil liquefaction and ground-motion amplification, in the artificial fill, mostly sand, on which the structures were built.

6.2 Earthquake Prediction

Research on earthquake prediction has been organized as a serious scientific effort only since about 1960 (Asada, 1982; Scholz, 1990, Chap. 7). Early work was done in the former Soviet Union, Japan, and the People's Republic of China. A formal program began in the United States in 1977 as part of the National Earthquake Hazards Reduction Program. This broadly based international effort has not yet resulted in any reliable techniques for establishing the time of occurrence of a specific earthquake at a given place. Enough cases of observations of anomalies in geophysical parameters prior to earthquakes have been reported to lend support to the concept that at least some earthquakes in some geological situations may be predictable on the basis of physical precursors. On the other hand, developments in nonlinear continuum mechanics leave open the question as to whether it is, in principle, possible to know the conditions within a fault zone and the surrounding region at any time well enough to be able to predict the future state of that zone more than a short time into the future.

The search for prediction methods is conducted within the framework provided by the plate-tectonics model of geologic processes and the elastic-rebound model of earthquake energetics. According to the latter, the goal should be to observe evidence that the rocks in a seismic source zone are in a highly strained state, with the corresponding stresses approaching the shear strength of the fault-zone materials. The possibility of making the necessary observations is limited by the location of the fault, with the best chances for shallow faults on land. Because many of the world's greatest earthquakes happen under the sea in subduction zones (see PLATE TECTONICS), the attempts to predict these events must depend on the interpretation of data collected at a distance from the source region.

Two principal approaches to prediction are the search for specific precursors of earthquakes and the application of pattern-recognition techniques. A precursor is an anomaly in some geophysical observable that precedes an earthquake and can be interpreted as diagnostic of the approaching occurrence. Anomalies in many physical variables that are potentially indicative of a state of high strain or stress have been noted and tested as possible precursors (Asada, 1982; Wyss, 1991). Many of these reported anomalies either have not been reproduced in reevaluation of the supporting data or in additional cases, have been poorly documented, or have not been demonstrated in a convincing way to be truly related to the subsequent earthquake.

The fully developed case for a claimed precursor must include not only the evidence that the phenomenon preceded some earthquakes, but also tests of the false-alarm rate resulting from the use of this phenomenon (anomaly with no earthquake) and the failure-to-predict rate (earthquake with no anomaly). In addition, a quantitative definition of what constitutes an "anomaly" is required. Because some attempts at prediction have been offered on the basis of pseudoscience, it is essential that rigorous standards be applied to the evaluation of claims of success.

The most promising phenomena for further investigation appear to be crustal deformation in the vicinity of faults and changes in the spatial and temporal distributions of background seismicity. Crustal deformation is appealing because it is directly related to the accumulation of strain in rocks. Crustal deformation may be monitored by direct observations of uplift and subsidence relative to some base, changes in the distances between points a few kilometers or tens of kilometers apart along lines that may or may not cross active faults, local measurements of tilt and strain in the rocks, and creep along active faults. Indirect evidence of crustal deformation can be derived from gravity measurements, which are sensitive to elevation changes (but also to possible variations in mass beneath the observing point), and changes in water level in wells in response to strain changes. Pronounced elevation changes have been reported as occurring years to months prior to some strong earthquakes. Tilt and strain anomalies have also been reported as precursors. Observation systems in regions selected for prediction attempts routinely monitor such parameters.

Changes in the distribution of background seismicity are indicators of changes in the stress state or strength distribution on the fault. Patterns of both seismic activation and seismic quiescence have been suggested as intermediate-term precursors of a strong earthquake. If earthquakes above some minimum magnitude for which the catalog is thought to be complete are counted as a function of time over some extended period, an average rate of occurrence for the region covered by the catalog may be computed. Intervals of clustering and decreased activity are interspersed through the time series, but a meaningful long-term mean rate can be established.

Seismic quiescence occurs when the rate of background seismicity becomes significantly lower, in a statistical sense, than the long-term mean rate, and stays lower for a time long compared with the typical fluctuations that characterize the sequence. Quiescence in the future epicentral region has been identified in a number of cases as occurring months to years before strong earthquakes. The subject remains controversial, however, because of imperfections in earthquake catalogs. Some changes in the rate of events derived from catalogs have been demonstrated as artificial, due to changes in the monitoring network or data-analysis procedures. As with all searches for precursors, only the observation of many more cases of carefully documented quiescence in clear association with a subsequent strong earthquake can lead to the acceptance of this phenomenon as reliably diagnostic of an impending event.

Many strong earthquakes are preceded by foreshocks. The final warning in the most famous case of a successful prediction, Haicheng, China, 4 February 1975 (Asada, 1982, pp. 255–259), was based on the observations of foreshocks that began three days before, increased to a peak of about 60 events per hour on the morning of the event, then suddenly stopped some eight hours before the mainshock. However, no characteristics of foreshocks have been identified that can reliably distinguish them as such until after the main shock has occurred. One approach is to assume that every moderate (magnitude 4 to 6) event in a selected area is a foreshock whose occurrence increases the probability of a stronger earthquake during the next few days. This approach is useful for alerting emergency response services and public administrators, as long as it does not lead to an unacceptable false-alarm rate.

Seismicity patterns are among the traits that are considered in pattern-recognition approaches to prediction. As currently implemented, pattern-recognition/prediction algorithms are almost entirely the products of Russian geophysicists, led by Keilis-Borok (Keilis-Borok *et al.,* 1988). The investigator may define as traits any number of phenomena that may conceivably be associated with a coming earthquake. Each trait is then scored in terms of its association with a set of known past earthquakes. Those traits that have high scores are retained as diagnostic. The resulting algorithm can then be tested against another set of known past earthquakes, not part of the original set on which the learning was done. The finally accepted version is then applied to the prediction of future events by yielding Times of Increased Probability (TIPs) in prescribed geographic regions. No attempt is made to understand the physical basis for the success of a particular trait; the approach is purely phenomenological. Although the output of the procedure is the identification of TIPs in those places selected by the algorithm, the amount by which the probability of occurrence is increased over the long-term average probability is not stated quantitatively. The approach is now being evaluated by a number of investigators.

In addition to crustal deformation and seismicity changes, changes in several stress- and strain-dependent properties of rocks have been examined as possible intermediate- or short-term precursors. These include changes in the compressional-wave velocity due to the formation of microcracks in highly stressed materials followed by the filling of the cracks with ground water, changes in electrical conductivity and direction of magnetization, and changes in ground-water chemistry, especially the concentration of radon gas in the water (Scholz, 1990, Chap. 7). Although none of these anomalies has been verified as a reliable precursor, research continues, and it is conceivable that some combination of anomalies, each of which by itself increases the probability of a strong earthquake only slightly, may be found that can be used as a predictive tool.

7. APPLICATIONS OF SEISMOLOGY TO SOCIETAL NEEDS

The reduction of the disastrous effects of earthquakes in populated areas is one direct application of knowledge gained through seismological research. In addition to providing the input to the hazard-assessment procedures described above, seismology also provides structural engineers with data about the properties of strong motion that are needed for the development of building codes and design criteria. The ground-motion properties of primary engineering interest are the dominant frequencies in the motion, the peak values of the motion, typically of ground acceleration, and the duration of the motion.

The mitigation of the disastrous effects of extreme natural events, including earthquakes, is not purely, perhaps even primarily, a technical problem. Effective mitigation measures require actions governed by powerful economic and sociological factors. Strict land-use regulation (avoidance of the hazard) and rigorously enforced building codes can reduce future losses. However, the best technological decisions on such matters are often modified to accommodate other societal needs, including the need to accommodate ever-increasing populations.

The application of seismological techniques to mapping in detail the geological structure of the upper 10 km or so of the Earth's crust is the principal tool for exploration for oil and gas. The principles employed are the same as those used to investigate the planet on a global or regional scale; the difference is that the frequency band of the waves observed and the source–receiver distances are selected to yield the high resolution needed. Controlled sources, such as charges of high explosives, compressed-air guns, and mechanical vibrators, provide the wave energy. Large numbers of receivers are deployed and grouped to provide high signal-to-noise ratios and spatial filtering that enhances the wave arrivals that are useful. Modern techniques of seismic exploration use extensive processing of digitally recorded signals, including both travel-time and amplitude information, to produce the cross sections of subsurface structure that can be interpreted in terms of potential for yielding hydrocarbons (see GEOPHYSICS).

Another application of seismology is monitoring underground nuclear explosions. The problem consists of several parts. An unknown source must be detected by suitably deployed instruments, recognized as a seismic event, and located. It must then be identified as a natural event, an earthquake, or an explosion. This is the discrimination problem. If the decision is that it was an explosion, the yield must be estimated. Support for research to develop reliable monitoring technology has led to major advances in both theoretical and observational seismology. Although the end of the Cold War has eliminated incentives for the known nuclear powers to monitor test programs of each other, seismological monitoring of clandestine explosions has an important role to play in oversight of the nonproliferation of nuclear weapons.

GLOSSARY

Body Wave: A seismic wave that travels through the interior of the Earth, propagating in three dimensions.

Dip of a Fault: The angle between the horizontal plane and a fault plane, usually taken as less than 90°.

Earthquake: Vibration of the Earth's surface due to a transient disturbance of the equilibrium of the rocks in the interior.

Epicenter: Point on the surface vertically above the point of initiation of an earthquake rupture.

Fault: Surface separating two blocks of rock, along which there has been tangential relative motion of the blocks.

Fault Trace: Line of intersection of a fault plane, projected upward if necessary, with the Earth's surface.

Focal Mechanism: A description of the geometric properties of the fault rupture that generated an earthquake. The essential elements are the strike, dip, and direction of slip on the fault surface.

Focal Sphere: An imaginary sphere of small radius surrounding an earthquake source, on which the direction of first motion is mapped.

Hypocenter: The point in the Earth at which the energy in an earthquake is released. It is the point of initial release in the case of an extended fault rupture.

Interplate Earthquake: An earthquake at the interface between two lithospheric plates.

Intraplate Earthquake: An earthquake in the interior of a lithospheric plate.

Intensity: A measure of earthquake strength based on effects on natural features, manmade structures, and human reactions.

Isoseismal: A curve separating zones of equal intensity on a map of earthquake intensities.

Love Wave: A surface wave that is a horizontally polarized shear wave. Love waves exist only in layered systems.

Magnitude: A measure of earthquake strength based on the recorded amplitude of selected waves generated by the earthquake.

Moment, Seismic: A measure of earthquake strength based on the area of the ruptured surface and the mean relative displacement across that surface.

Microseisms: Continuous vibrations of the surface produced by sources such as wind or storms at sea.

Normal Fault: A fault on which the block above the fault plane moves downward in the direction of dip.

Precursor: An anomalous change in some observable, that is diagnostic of an impending earthquake.

***P* Wave:** An elastic wave related to propagating volumetric strain. It is longitudinally polarized in an isotropic medium.

Quiescence: A significant decrease in the rate of occurrence of background seismicity from the long-time average rate.

Rayleigh Wave: A surface wave for which the particle motion is polarized in the plane of incidence. In the absence of layering, Rayleigh waves are nondispersive.

Reverse Fault: A fault on which the block above the fault plane moves upward in the direction of dip.

Strike of a Fault: The angle between the fault trace and geographic north, conventionally defined so that the fault dips down to the right (<90°) when one looks in the direction of the strike.

Strike-Slip Fault: A fault on which the relative motion of the blocks is horizontal.

Surface Wave: A seismic wave that travels as a guided wave in layers defined by the Earth's outer surface and prominent boundaries in the interior. Surface waves propagate in two dimensions and exhibit geometric dispersion. They die off exponentially with distance away from the waveguide.

***S* Wave:** An elastic wave related to propagating shear strain. It is transversely polarized in an isotropic medium.

Tectonic Earthquake: An earthquake driven by the forces in the Earth that form geologic structures. Tectonic earthquakes are caused by sudden slip on a fault.

Works Cited

Aki, K., Richards, P. G. (1980), *Quantitative Seismology,* San Francisco: W. H. Freeman.

Anderson, D. L., Given, J. W. (1982), *J. Geophys. Res.* **87,** 3893–3904.

Asada, T. (Ed.) (1982), *Earthquake Prediction Techniques,* Tokyo: University of Tokyo Press.

Bullen, K. E., Bolt, B. A. (1985), *An Introduction to the Theory of Seismology,* 4th ed., Cambridge, U.K.: Cambridge Univ. Press.

Gephardt, J. W., Forsyth, D. W. (1984), *J. Geophys. Res.* **89,** 9305–9320.

Gubbins, D. (1990), *Seismology and Plate Tectonics,* Cambridge, U.K.: Cambridge Univ. Press.

Gutenberg, B., Richter, C. F. (1949), *Seismicity of the Earth,* Princeton: Princeton Univ. Press.

Hanks, T., Kanamori, H. (1979), *J. Geophys. Res.* **84,** 2348–2352.

Jackson, D. D., Anderson, D. L. (1970), *Rev. Geophys. Space Phys.* **8,** 1–63.

Kanamori, H. (1977), *J. Geophys. Res.* **82,** 2981–2987.

Keilis-Borok, V. I., Knopoff, L., Rotwain, I. M., Allen, C. R. (1988), *Nature* **335,** 690–694.

Klein, F. W., Koyanagi, R. Y., Nakata, J. S., Tanigawa, W. R. (1987), in R. W. Decker, T. L. Wright, P. H. Stauffer (Eds.), *Volcanism in Hawaii,* U.S. Geological Survey Professional Paper 1350, Washington, D.C.: U.S. GPO, pp. 1019–1185.

Kostrov, B. V., Das, S. (1988), *Principles of Earthquake Source Mechanics,* Cambridge, U.K.: Cambridge Univ. Press.

Lee, W. H. K., Stewart, S. W. (1981), *Principles and Applications of Microearthquake Networks,* New York: Academic.

Macelwane, J. B. (1947), *When the Earth Quakes,* Milwaukee: Bruce.

Musgrave, M. J. P. (1970), *Crystal Acoustics,* San Francisco: Holden-Day.

Reid, H. F. (1910), *The California Earthquake of April 18, 1906,* Vol. II, *The Mechanics of the Earthquake,* Report of the State Earthquake Investigation Commission, Washington, DC: Carnegie Institution of Washington.

Reiter, L. (1990), *Earthquake Hazard Analysis,* New York: Columbia Univ. Press.

Richter, C. F. (1958), *Elementary Seismology,* San Francisco: W. H. Freeman.

Scholz, C. H. (1990), *The Mechanics of Earthquakes and Faulting,* Cambridge, U.K.: Cambridge Univ. Press.

Sieh, K. (1984), *J. Geophys. Res.* **89,** 7641–7670.

Sieh, K., Stuiver, M., Brillinger, D. (1989), *J. Geophys. Res.* **94,** 603–624.

Wyss, M. (Ed.) (1991), *Evaluation of Proposed Earthquake Precursors,* Washington, DC: American Geophysical Union.

Further Reading

Bolt, B. A. (1993), *Earthquakes,* New York: W. H. Freeman.

Kennett, B. L. N. (1983), *Seismic Wave Propagation in Stratified Media,* Cambridge, U.K.: Cambridge Univ. Press.

Mogi, K. (1985), *Earthquake Prediction,* Tokyo: Academic.

SEMANTIC VIEW OF THEORIES

R. I. G. HUGHES, *Department of Philosophy, Logic and Scientific Method, The London School of Economics and Political Science, London, U.K.*

Introduction 175

1. **The Syntactic View** 175

2. **The Set-Theoretic Approach** 176

3. **The State-Space Approach** ... 177

4. **The Representational Approach** 178

Works Cited 179

Further Reading 180

INTRODUCTION

The use of "the semantic method" in "the logical analysis of theories" was proposed by the Dutch philosopher Evert Beth in 1949. The method he advocated was that developed in Tarski's investigation of formal languages in the 1930s. The phrases *the semantic view* and *the semantic conception* now cover a variety of approaches to the study of physical theories. They have in common the rejection of an earlier conception of theories, variously called *the syntactic view, the statement view,* and (to mark its predominance during the first half of this century) *the received view.*

1. THE SYNTACTIC VIEW

Between about 1930 and 1960 the statement view was articulated by the logical empiricists, notably Hans Reichenbach, Rudolf Carnap, and Carl Hempel, who took as a model for their work Hilbert's (1971) formalization of the foundations of geometry. On their view, as adumbrated by Carnap (1939), a physical theory consists of a formal calculus together with a (partial) interpretation of that calculus. The formal calculus is a theory in the logician's sense: it is the deductive closure of a set of axioms. The language in which the axioms are stated and the principles governing deductions are to be those of modern logic. The language contains connectives (e.g., "→" for "If . . . , then . . ."), quantifiers (e.g., "$\forall x$" for "For any x"), variables ("x", "y", etc.), and predicates ("P", "Q", etc.); the sentence "$\forall x(Px \rightarrow Qx)$" may be read as "For any x, if x is a P, then x is a Q", or, more succinctly, as "All Ps are Qs." This abstract syntactical structure is given content by an interpretation. Only the elements of the language that can be interpreted as referring to observable quantities and their magnitudes (the *observable predicates*) can be "cashed out," and so the formal language is only partially interpreted. Although theoretical predicates receive no direct interpretation, the theory contains correspondence rules that relate them to observational terms.

It was no part of the logical empiricists' program that this articulation of a theory should correspond to the way that it would be presented and used by a practicing physicist. Rather, their aim was to isolate and display the logical structure and the cognitively significant features of a theory. To this end, one of the few fully developed analyses of this kind, Reichenbach's (1957) axiomatization of relativity, "was aimed at exhibiting the roles of experience and of convention in physical theorizing about space, time, and motion" (Hempel, 1977). A detailed account of the development of the received view, and also of the criticisms leveled against it, is provided by Suppe (1977).

The crucial move made by proponents of the semantic view is to reject the thesis that a theory is primarily a linguistic entity. Instead of concentrating, as the logical empiricists did, on the language in which the theory is articulated, they focus attention on the structures that the axioms of a theory define. In the formal version of the semantic approach these structures are models of the theory, in the sense in which that word is

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

used in logic and formal semantics. Hence the phrase, "the semantic view." While an emphasis on the models of a theory is common to all versions of the semantic view, different versions contain different proposals concerning the way these models should be specified. To some extent these differences also correspond to different views about the function that the analysis should serve. In the examples that follow, the demand that the specifications should have a canonical form is progressively relaxed. At the same time, not only do the analyses move closer to scientific practice, but a close match between the two becomes a desideratum.

2. THE SET-THEORETIC APPROACH

The set-theoretic approach was initiated by Suppes (1969) in the 1960s and taken up in the next decade by Sneed (1979) and Stegmüller (1979). Like the logical empiricists, Suppes aimed at clarifying the foundations of physical theories, rather than at an analysis that matched scientific practice. He too emphasized the axiomatic method, but he differed from the empiricists, first, by using this method to define set-theoretical structures and, second, by allowing informal set-theoretical methods of proof where the empiricists had insisted on a rigid adherence to first-order logic. (A logic is *first order* if it allows quantification over individuals but not over predicates. In the sentence, "Every light bulb was defective in some way or other," for example, "every" involves quantification over individuals, but "some," quantification over predicates.) Instead of looking to Hilbert's work in mathematics as a model, Suppes and those who followed him looked to the group of French mathematicians collectively known as Bourbaki.

Sneed (1979) claims that the approach yields a "logical clarification of theories"; his account of classical particle mechanics offers an example of a Bourbaki-style presentation of a scientific theory. Sneed here uses the set-theoretic predicates "is a PM" (is a particle mechanics) and is a "CPM" (is a classical particle mechanics) where other writers would talk of a structure being a system of a particular kind. The predicates are defined as follows.

x is a PM if and only if there exists a P, T, $\mathbf{s}$, m, $\mathbf{f}$, such that

1. $x = \langle P,T,\mathbf{s},m,\mathbf{f}\rangle$;
2. P is a nonempty finite set;
3. T is an interval of real numbers;
4. $\mathbf{s}$ is a function from $P \times T$ into the set of ordered triples of real numbers such that, for all $p \in P$, $t \in T$, $D^2\mathbf{s}(p,t)$ exists;
5. m is a function from P into the real numbers such that, for all $p \in P$, $m(p) > 0$; and
6. $\mathbf{f}$ is a function from $P \times T \times I$ into the set of ordered triples of real numbers such that, for $p \in P$, $t \in T$,

$$\sum_{i \in I} \mathbf{f}(p,t,i) \text{ is absolutely convergent.}$$

A system P of point masses, moving in three-dimensional space during a time interval T under the influence of a collection of forces, constitutes a realization of a PM; clause **4** (in which the operator D is the time derivative) stipulates that the acceleration of each particle is at all times well defined; clause **5**, that the mass m of each particle is positive and never varies; clause **6**, that the total (vector) sum of the forces acting on each particle is well defined.

The class of models is then further restricted by adding Newton's second law, to define a classical particle mechanics.

x is a CPM if and only if

1. x is a PM; and
2. for all $p \in P$ and $t \in T$,

$$m(p)D^2\mathbf{s}(p,t) = \mathbf{f}(p,t,i)$$

Sneed goes on to define the subclass of CPMs that are also governed by Newton's third law (Newtonian classical particle mechanics). In any theory the specification of set-theoretic structures is accompanied by a set of intended applications; an application is a physical system (e.g., our solar system) that may be treated as a particle mechanics. Sneed is unwilling, however, to regard statements of the form "Q is a CPM" as expressing the empirical content of the theory (where Q is an intended application), because of difficulties involving theoretical terms. The problem, in the case of particle mechanics, lies with the mass and force functions. If Q is a putative application of classical particle mechanics, there may be no way to determine the values of these functions for Q that does not already presuppose that Q is a CPM. Sneed acknowledges that he gives only a partial resolution of this problem. In a later work co-authored with Wolfgang Balzer and Ulises Moulines (Balzer *et al.*, 1987), Sneed suggests that the occurrence of the same theoretical term (e.g., mass) in more than one theory offers a different solution.

In this work Balzer *et al.* show how versatile the set-theoretic approach is by applying it, not only to classical particle mechanics, classical collision mechanics, and relativistic collision mechanics, but also to decision theory, Daltonian stoichemistry, simple equilibrium thermodynamics, Lagrangian mechanics, and pure exchange economics. The authors also apply the set-theoretic approach to a general theory of theories, as does Stegmüller (1976, 1979).

3. THE STATE-SPACE APPROACH

The state-space approach stays much closer to orthodox formulations of physical theories than does the set-theoretic approach. The most influential presentation of it was made by van Fraassen (1983), who acknowledged debts to Hermann Weyl and (especially) to Evert Beth. The approach is also implicit in the work of von Neumann (1955) and Mackey (1963) on quantum mechanics and suggests an approach to the interpretive problems raised by that theory. It also offers a way of comparing theories as different as classical and quantum mechanics, and these theories will serve as examples. (See Hughes, 1989.) It has also been successfully applied to evolutionary theory in biology; see Thompson (1989) and Lloyd (1988).

On the state-space approach a theory is characterized by three things:

1. the way in which the state of a system is specified;
2. the way in which the state evolves in time; and
3. the relation the theory postulates between the state of a system and the values of physical quantities associated with the system.

In the Hamilton-Jacobi formulation of classical mechanics, a classical system consists of a set of n point masses. The state of the system at a given time is specified by a point ω in a space Ω of $6n$ dimensions. The coordinates of ω are the three components q_x, q_y, q_z, of position and the three components p_x, p_y, p_z, of momentum, for each of the n particles. The evolution of the state ω through time is determined by the Hamiltonian function $H{:}\Omega \rightarrow R$ for the system, whose value for a given state is the system's total energy in that state. The evolution of ω through time is then given by the $3n$ pairs of equations

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \qquad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}.$$

In contrast, the state space of a quantum system is a Hilbert space $\mathcal{H}$, possibly infinite-dimensional. A pure state of the system is represented by a normalized vector ψ in $\mathcal{H}$; more generally, a state is representable by a density operator D (a trace-class operator of trace 1) on $\mathcal{H}$. Here, however, only pure states will be considered. In quantum mechanics, the energy of a system is associated with an operator H on $\mathcal{H}$ (the Hamiltonian operator); the evolution of ψ through time is given by Schrodinger's equation

$$ih\, \partial\psi/\partial t = H\psi.$$

Radical differences between classical and quantum mechanics appear in the relations the two theories postulate between a system's state and the measurable values of physical quantities associated with the system. As in the special case of energy, whereas in classical mechanics each physical quantity A is associated with a real-valued function $f_A{:}\Omega \rightarrow R$, in quantum mechanics a physical quantity is associated with a Hermitian operator $A{:}\mathcal{H} \rightarrow \mathcal{H}$. In the classical case, the value of f_A for a given state gives the value of the corresponding physical quantity. These values, obviously, will be restricted to the range of f_A. The predictions of the theory take the form "In state ω the measured value of A will be $f_A(\omega)$."

In quantum mechanics, on the other hand, the possible values of a physical quantity are restricted to the spectrum of the corresponding Hermitian operator, A. This spectrum may be discrete (as in the case of a spin operator), in which case the spectrum consists of the eigenvalues of A, or continuous (as in the case of the momentum operator p). In the discrete case, to each eigenvalue a_i of the operator is associated a projection operator P_i from the spectral decomposition of A, so that

$$A = \sum a_i P_i.$$

P_i projects onto the ray spanned by the eigenvectors corresponding to the eigenvalue a_i.

In the continuous case, to the Hermitian operator A we assign a spectral measure $P(x)$, such that

$$A = \int_{-\infty}^{\infty} x dP(x).$$

We then define, for each semi-closed interval $\Delta = \{x: a < x \leq b\}$ of $\mathbb{R}$, the operator

$$P^A_\Delta = P(b) - P(a).$$

In both cases, the theory predicts, not the value of the quantity that measurement will yield, but the probability that a particular outcome will result. In the discrete case the probability p_i that a measurement of A will yield the result a_i is given by

$$p_i = \langle \psi | P_i \psi \rangle.$$

In the continuous case, the probability p that a measurement of A will yield a result in the interval Δ is given by

$$p_i = \langle \psi | P^A_\Delta \psi \rangle.$$

In both the classical and quantum cases, van Fraassen (1983) suggests that the observable magnitudes are to be identified with elements in the "empirical substructures" within the models defined by the theory. For example, in quantum mechanics the set of subspaces of the Hilbert space (each associated with a value, or set of values, of a physical quantity) is the empirical substructure of the Hilbert space.

In this way, the state-space approach may suggest a specific interpretive stance with respect to a theory. For, in the case of quantum mechanics, the set of subspaces has the algebraic structure of a coherent orthomodular lattice (equivalently a transitive partial Boolean algebra: see Hughes, 1989; QUANTUM LOGIC). The "logic" of these structures is nonstandard; within them the Boolean distributive law does not hold, in general. Thus the state-space representation of quantum theory leads naturally to the "quantum logical" approach to its interpretation.

4. THE REPRESENTATIONAL APPROACH

The set-theoretic and state-space versions of the semantic account emphasize foundational theories of physics, like classical mechanics and quantum mechanics. They offer alternative ways of representing the internal structure of these theories. They do not examine the question of how a theory is applied. In contrast, the representational approach accommodates both foundational theories and their applications. For this reason, of the three versions of the semantic view examined, it is the one that most nearly conforms to scientific practice.

On this account, a physical theory involves two things:

1. the specification of a class of models (the *theoretical definition*); and
2. the postulate that a particular part of the world, or of the world as we describe it, can be adequately represented by a model from that class (the *theoretical hypothesis*).

This two-part analysis is due to Giere (1984). It is explicitly endorsed by van Fraassen (1989) in his recent writings and implicitly relied on by (e.g.) Suppe (1989) and Cartwright (1983). These four authors differ, however, in their description of the relation between the model and its subject, the part of the world that it represents. Van Fraassen (1989) takes the subject–model relation to be "some embedding, or approximate embedding." This accords with his suggestion that the empirical content of the model is provided by the empirical substructures of the theory; he describes these as "candidates for the direct representation of observable phenomena." In doing so he limits himself to those comparatively rare phenomena to which a foundational theory can be directly applied. Of the four authors mentioned above, he alone pays no attention to the theoretical models used in specific applications of a theory. While the behavior of such "local models," be they models of atomic structure or of the phenomenon of superconductivity, may be governed by the laws of a foundational theory, the model itself is crafted to fit the individual phenomenon.

As an example of a local model, Giere uses the ideal harmonic oscillator whose behavior conforms to the laws of classical mechanics. No "real oscillating system," no weight suspended from a spring, behaves exactly like the ideal oscillator. Giere (1985) describes the subject–model relation in terms of similarity; the subject is "similar to the proposed model in specified respects, and to a specified degree."

For Suppe (1989) the model is a "kind of replica" of its subject; he goes on to distinguish between replicas produced by *abstraction* and those produced by *idealization*. In abstraction we ignore certain influences that affect actual systems (e.g., friction, air resistance); in these cases it is causally possible (Suppe's term) to realize the conditions under which the behavior of the subject and model agree. In idealization we postulate abstract, unrealizable models, as in particle mechanics where the model consists of a set of point masses.

In terms of the relative importance they assign to foundational theory, Cartwright (1983) stands at the opposite end of the spectrum from van Fraassen. In criticizing van Fraassen's version of the semantic account, she writes: "Schrödinger's equation, even when coupled with principles which tell us what Hamiltonians to use for square-well potentials, two-body Coulomb interactions, and the like, does not constitute a theory of anything. To have a theory of the ruby laser, or of bonding in the benzene molecule, one must have models for these phenomena which tie them to descriptions in the mathematical theory." These models, which Cartwright calls *simulacra* to emphasize their fictional nature, do the work of physics. They are models of *phenomena*. On this view, theorizing is a four-level process: The behavior of phenomena is described in terms of simulacra (models of phenomena); that of the simulacra, in terms of models of a theory (infinite potential wells, linear harmonic oscillators, and the like); and that of these "local" models, in terms of the foundational theory in question. As an example she cites Louisell's treatment of a gas laser. The laser, and its behavior, is the phenomenon to be modeled. It is represented as a collection of three-level atoms in interaction with a quantized electromagnetic field; both these components also interact with a damping reservoir. The behavior of this simulacrum is in turn modeled by representing it in a way that is amenable to treatment by the relevant foundational theory, in this case quantum theory.

Physics offers examples of theoretical representation that conform to the views of each of the four authors cited. Indeed, one virtue of this approach to the analysis of theory is that it draws attention to the many different ways in which theories are applied. It also offers a new approach to vexed philosophical issues such as the status of scientific laws and the nature of scientific explanation (see van Fraassen, 1989; Hughes, 1993). For these reasons it underlies much contemporary work in the philosophy of science.

Works Cited

Balzer, W., Moulines, C. U., Sneed J. D. (1987), *An Architectonic for Science*, Dordrecht: Reidel.

Carnap, R. (1939), *Foundations of Logic and Mathematics*, Chicago: Univ. of Chicago Press.

Cartwright, N. (1983), *How the Laws of Physics Lie*, Oxford: Clarendon Press.

Giere, R. (1984), *Understanding Scientific Reasoning*, 2nd ed., Fort Worth, TX: Holt, Rinehart and Winston.

Giere, R. (1985), "Constructive Realism," in: P. M. Churchland, C. A. Hooker (Eds.), *Images of Science*, Chicago: Univ. of Chicago Press.

Hempel, C. G. (1977), "Formulation and Formalization of Scientific Theories," in: F. Suppe (Ed.), *The Structure of Scientific Theories*, 2nd ed., Urbana, IL: Univ. of Illinois Press.

Hilbert, D. (L. Under, transl.) (1971), *The Foundations of Geometry*, La Salle, IL: Open Court (orig. pub. 1899).

Hughes, R. I. G. (1989), *The Structure and Interpretation of Quantum Mechanics*, Cambridge, MA: Harvard University Press.

Hughes, R. I. G. (1993), "Theoretical Explanation," in: P. A. French, T. E. Uehling, H. K. Wettstein (Eds.), *Philosophy of Science*, Midwest Studies in Philosophy Vol. XVIII, Notre Dame, IN: Univ. of Notre Dame Press.

Lloyd, E. A. (1988), *The Structure and Confirmation of Evolutionary Theory*, Westport, CT: Greenwood Press.

Mackey, G. W. (1963), *The Mathematical Foundations of Quantum Mechanics,* New York: Benjamin.

Reichenbach, H. (M. Reichenbach, J. Freund, transl.) (1957), *Axiomatization of the Theory of Relativity,* Berkeley, CA: Univ. of California Press (orig. pub. 1924).

Sneed, J. D. (1979), *The Logical Structure of Mathematical Physics,* 2nd ed., Dordrecht: Reidel.

Stegmüller, W. (1979), *The Structuralist View of Theories,* New York: Springer Verlag.

Suppe, F. (Ed.) (1977), *The Structure of Scientific Theories,* 2nd ed., Urbana, IL: Univ. of Illinois Press.

Suppe, F. (1989), *The Semantic Conception of Theories and Scientific Realism,* Urbana, IL: Univ. of Illinois Press.

Suppes, P. (1969), *Studies in the Methodology and Foundations of Science,* Dordrecht: Reidel.

Thompson, P. (1989), *The Structure of Biological Theories,* Albany: SUNY Press.

van Fraassen, B. C. (1983), *The Scientific Image,* Oxford: Clarendon Press.

van Fraassen, B. C. (1989), *Laws and Symmetries,* Oxford: Clarendon Press.

von Neumann, J. (R. T. Beyer, transl.) (1955), *Mathematical Foundations of Quantum Mechanics,* Princeton: Princeton University Press (orig. pub. 1932).

Further Reading

The Syntactic View

Carnap, R. (1939), *Foundations of Logic and Mathematics,* Chicago: Univ. of Chicago Press.

Hempel, C. G. (1977), "Formulation and Formalization of Scientific Theories," in: F. Suppe (Ed.), *The Structure of Scientific Theories,* 2nd ed., Urbana, IL: Univ. of Illinois Press.

Hilbert, D. (L. Under, transl.) (1971), *The Foundations of Geometry,* La Salle, IL: Open Court (orig. pub. 1899).

Nagel, E. (1961), *The Structure of Science,* New York: Harcourt, Brace, and World.

Reichenbach, H. (M. Reichenbach, J. Freund, transl.) (1957), *Axiomatization of the Theory of Relativity,* Berkeley, CA: Univ. of California Press (orig. pub. 1924).

Critical Perspectives on the Syntactic View

Grünbaum, A., Salmon, W. C. (Eds.) (1988), *The Limits of Deductivism,* Berkeley: Univ. of California Press.

Suppe, F. (Ed.) (1977), *The Structure of Scientific Theories,* 2nd ed., Urbana, IL: Univ. of Illinois Press. The editor's introduction offers a detailed account and critique of the syntactic view.

The Set-Theoretic Approach

Balzer, W., Moulines, C. U., Sneed, J. D. (1987), *An Architectonic for Science,* Dordrecht: Reidel.

Sneed, J. D. (1979), *The Logical Structure of Mathematical Physics,* 2nd ed., Dordrecht: Reidel.

Stegmüller, W. (1976), *The Structure and Dynamics of Theories,* New York: Springer Verlag.

Stegmüller, W. (1979), *The Structuralist View of Theories,* New York: Springer Verlag. Provides a good introduction to this approach.

Suppes, P. (1969), *Studies in the Methodology and Foundations of Science,* Dordrecht: Reidel.

The State-Space Approach

Hughes, R. I. G. (1989), *The Structure and Interpretation of Quantum Mechanics,* Cambridge, MA: Harvard University Press.

Lloyd, E. A. (1988), *The Structure and Confirmation of Evolutionary Theory,* Westport, CT: Greenwood Press.

Mackey, G. W. (1963), *The Mathematical Foundations of Quantum Mechanics,* New York: Benjamin.

Thompson, P. (1989), *The Structure of Biological Theories,* Albany: SUNY Press.

van Fraassen, B. C. (1991), *Quantum Mechanics: An Empiricist View,* Oxford: Clarendon Press.

von Neumann, J. (R. T. Beyer, transl.) (1955), *Mathematical Foundations of Quantum Mechanics,* Princeton: Princeton University Press (orig. pub. 1932).

All the above apply the state-space approach, explicitly or implicitly, to particular theories. A more general account appears in van Fraassen, B. C. (1983), *The Scientific Image,* Oxford: Clarendon Press.

The Representational Approach

Cartwright, N. (1983), *How the Laws of Physics Lie,* Oxford: Clarendon Press.

French, P. A., Uehling, T. E., Wettstein, H. K. (Eds.) (1993), *Philosophy of Science,* Midwest Studies in Philosophy Vol. XVIII, Notre Dame, IN: Univ. of Notre Dame Press.

Giere, R. (1984), *Understanding Scientific Reasoning,* 2nd ed., Fort Worth, TX: Holt, Rinehart and Winston. Chapter 5 gives a straightforward introduction to this approach.

Hughes, R. I. G. (1993), "Theoretical Explanation," in: French, P. A., Uehling, T. E., Wettstein, H. K. (Eds.) (1993), *Philosophy of Science,* Midwest Studies in Philosophy Vol. XVIII, Notre Dame, IN: Univ. of Notre Dame Press.

Suppe, F. (1989), *The Semantic Conception of Theories and Scientific Realism,* Urbana, IL: Univ. of Illinois Press.

van Fraassen, B. C. (1989), *Laws and Symmetries,* Oxford: Clarendon Press.

SEMICONDUCTOR-DEVICE INTEGRATION

Hideo Sunami, *Memory Business Operation, Semiconductor & Integrated Circuits Division, Hitachi, Ltd., Kodaira, Tokyo, Japan*

Introduction 181
1. **Elementary Processes and Devices for Integration** 184
1.1 Elementary Processes 184
1.2 Patterning Technology 184
1.3 Planarization Technology 186
2. **Process Integration Technologies** 188
2.1 Isolation Structure 188
2.2 Device Structure 189
2.3 CMOS Process Integration 189
2.3.1 DRAM Process Integration 189
2.3.2 Logic Process Integration 191
2.4 Bipolar Process Integration 193
2.5 GaAs Process Integration 195
3. **Further Integration Technologies** 195
4. **Conclusion** 197
Glossary 197
Works Cited 197
Further Reading 198

INTRODUCTION

Since the transistor was invented in the late 1940s, various kinds of semiconductor devices have been developed. They are rectifiers, amplifiers using small-signal or power transistors, microwave-generating diodes, light-emitting diodes (LEDs), laser diodes (LDs), line and area light sensors, infrared sensors, pressure sensors, and *p*H sensors. Recently, very tiny mechanical devices have been fabricated also using silicon process techniques.

Since an electronic circuit integrating more than two devices was invented and demonstrated in 1958 by J. S. Kilby (Kilby, 1976), the integrated-circuit (IC) industry has grown very rapidly at an average annual rate of greater than 10%. It is predicted that worldwide IC sales will surpass \$200 × 10^9 by the year 2000. The growth rate of the IC industry is greater than that of the whole electronics industry. This is partly because the proportion of IC components in products has been gradually increasing. There are very few electronic products that do not take advantage of ICs today.

Even though there are several other potential semiconductor materials, such as gallium arsenide, gallium phosphide, germanium, and silicon carbide, most integrated circuits commercially available in volume production have been made from silicon. The major reasons are that

1. silicon is a very stable material;
2. there exist appropriate dopant atoms for forming *n*-type and *p*-type regions;
3. the silicon-dioxide film formed by oxidation of the silicon itself is chemically and physically very stable; and
4. a planar technology (Hoerni, 1960) has been invented that is very effective for fabrication of monolithic integrated circuits.

It is also very fortunate for the silicon IC industry that silicon material is plentiful on Earth.

Silicon integrated circuits are categorized according to device use or function. As regards devices, the leading technology has been shifting from bipolar transistor, to *p*-channel MOSFET (metal-oxide-semiconductor field-effect transistor), to *n*-channel MOSFET, and then to CMOS (complementary MOS).

Bipolar devices are still used in ultrahigh-speed logic and memory for supercomputers and main-frames, in linear ICs for consumer products, as microwave transistors for telecommunication, and for high-power devices. BiCMOS (bipolar CMOS) devices are now available combining the advantages of bipolar and CMOS transistors.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

As for their functions, there are memories, logics, and microprocessors. The memories include DRAM (dynamic random-access memory, SRAM (static RAM), Mask ROM (read-only memory), and nonvolatile ROM such as EPROM (erasable and programmable ROM) and EEPROM (electrically erasable PROM). Flash memory is one kind of EEPROM. The logics include G/A (gate array), signal processors, and standard logics such as TTL (transistor-transistor logic). Most microcomputers incorporate both logics and very high-speed cache memory. As the integration level increases, various kinds of devices are inevitably integrated in response to the new functions required.

The basic configuration and operation of major integrated circuits that are now widely manufactured in semiconductor industry are briefly described in the following.

DRAM (Dynamic Random-Access Memory): A 4-megabit DRAM incorporates 4 194 304 (= 2^{22}) memory cells. The memory cell consists of one MOS transistor and one capacitor. Stored charge in the capacitor is continuously leaking out, resulting in loss of stored "memory"; therefore, a restoring operation is required typically every 100-ms interval. Thus this is called "dynamic" RAM. Since the memory access time is around 100 ns and relatively slower than that of SRAM, DRAM is typically used for the main memory of the computer.

SRAM (Static RAM): Its memory cell principally consists of six transistors configuring a flip-flop. Since the memory cell is several times larger than that of DRAM, one-quarter as much memory volume is available in the same technology generation. As no additional restoring operation is required, this is called "static" RAM. The access time is generally much faster than that of DRAM; thus, SRAM is used for cache memory of the computer and main memory of the supercomputer. At the same time, extremely low-power SRAM is also available making full use of no need for a restoring operation. Beside MOS SRAM, bipolar SRAM is still being produced for ultrahigh-speed operation.

EPROM (Erasable and Programmable Read-Only Memory): Even though there are various kinds of EPROM and EEPROM, the memory cell principally consists of one or two MOS transistors having a storage site of a floating gate or an interface of different materials inside a gate oxide. The writing operation is done by charging the storage site, while erasing is performed by UV-light exposure. Since the erasing time is typically around 10 min with UV exposure, EPROM is in general used as one-time writable ROM.

EEPROM (Electrically Erasable and Programmable Read-Only Memory): The memory cell is principally the same as that of EPROM, while erasing is done by hot-electron injection or tunneling instead. More than 10^4 cycles of erasing is guaranteed. EEPROM is used as user-programmable ROMs in various consumer electronics.

Flash EEPROM: One type of EEPROM. This is so called because erasing proceeds simultaneously across all memory cells like a "flash."

Mask ROM: This is the first read-only memory. It consists of a memory cell with one simple MOS transistor. The "write" operation is no more than a patterning process that makes or does not make an electrical connection to certain transistors at the memory site. This is the highest density memory of all resulting in the least expensive and highest volume usage.

Logic LSI: A generic term for LSI incorporating logic circuits mainly used for logic operation. Usually it contains certain ROM and SRAM required for the logic operation. There are various kinds of logic LSIs in the semiconductor market.

MPU (Micro Processor Unit): One kind of logic LSI incorporating various kinds of logic and memory circuits used for calculation for computing. The memories are principally very fast SRAM, used as a level-1 cache memory, and ROM. The relative area of total memories has become larger and close to almost one-half of total chip area in advanced MPUs.

The market trend in these silicon devices is shown in Fig. 1. The growth rates of memory and microprocessors are bigger than those of others. In 1994, the market share of DRAM was around 60% in the MOS memory category. The rapid market growth from 1992 to 1994 is mainly due to the growth of the personal computer market.

Figure 2 shows device counts of DRAM and microprocessor chips. Fourfold higher packing has been realized every three years since 1970 when the 1-kilobit DRAM was introduced. As of 1995, a leading-edge, 32-bit

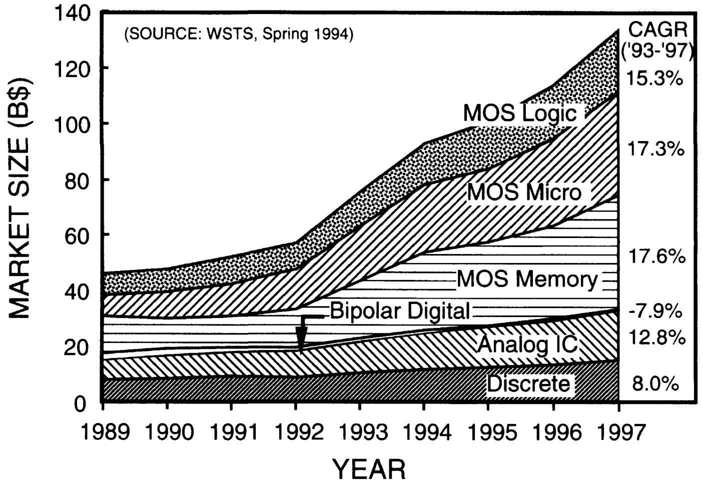

FIG. 1. Market trends in semiconductor products. Trends from 1994 are projected.

microprocessor integrates 3 million transistors on a 300-mm^2 die with 0.8-μm feature size, while a 4-Mbit DRAM has 5 million transistors and 4 million capacitors on an 80-mm^2 die with the same feature size. The packing is limited by memory-cell size for DRAM and by metal interconnect wiring in microprocessors.

A cross section of a typical sub-μm CMOS device identifying critical processing components is shown in Fig. 3. This is representative of a large variety of products, some of which may not use some of the components depending on the function and performance requirements and the manufacturer's experience. Individual elementary processes are not described in this article except for a few selected processes referred to in the next section.

The drive current I_{drive} in the saturation region of an MOS transistor is fundamentally expressed as

$$I_{\text{drive}} = \mu\epsilon_i W_{ch}(V_{gs} - V_t)^2/2T_iL_{ch}, \qquad (1)$$

where μ, ϵ_i, W_{ch}, V_{gs}, V_t, T_i, and L_{ch} are carrier mobility, permittivity of gate insulator, channel width, gate-to-source voltage, threshold voltage, gate-insulator thickness, and channel length, respectively. Thus, a thinner gate insulator and a shorter channel length lead to better performance. This is one of

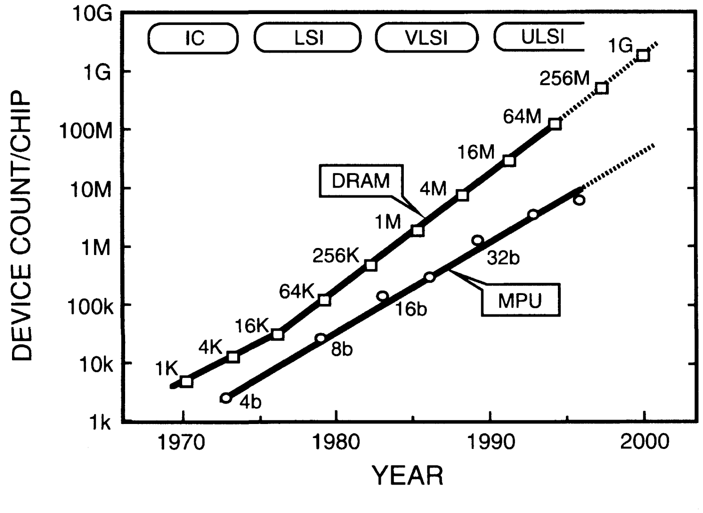

FIG. 2. Device counts of DRAMs and microcomputers. Current DRAM always shows the largest device count among all ICs because of its simple and tiny memory-cell structure.

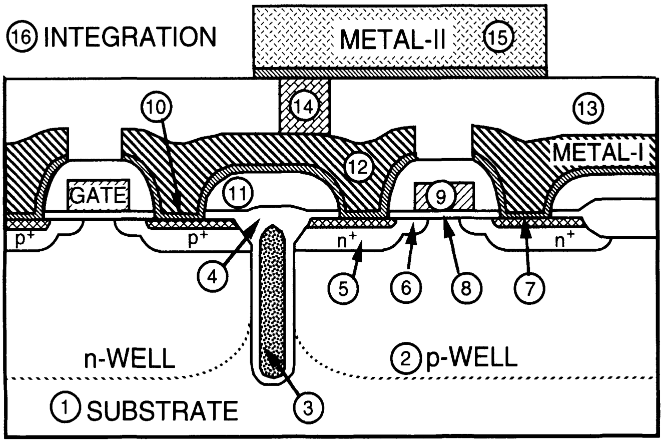

FIG. 3. Elementary components and key issues for submicron CMOS devices. (1) Substrate: epitaxial growth, SOI (Si-on-insulator); (2) well formation: retrograde well, low resistance; (3) deep-trench isolation: etching, side-wall doping, insulator filling; (4) recessed isolation: short bird's beak; (5) junction: low-leakage *p-n* junction, low sheet resistance, silicide; (6) impurity profile: well-defined doping; (7) junction contact: low resistance, barrier material, silicon-cluster precipitation; (8) gate insulator: thinner insulator with good integrity, low defect density, high reliability; (9) gate: low resistance, silicide and/or polycide, refractory metal; (10) self-aligned contact: barrier materials, e.g., titanium nitride, TiN; (11) intermediate insulator I; low-temperature reflow; (12) metallization I; low resistance, heat resistance, free from migration; (13) intermediate insulator II: planarization, less water absorption; (14) via hole: plug, low contact resistance, self-aligning; (15) metallization II: heat resistance, low resistance; (16) fine patterning: delineation accuracy, aligning, and etching.

the biggest motivating forces driving toward smaller devices and resulting in more functions on a die. These characteristics are summarized by the "scaling law" (Dennard *et al.*, 1974).

Since DRAM's greater integration started in 1970, technology innovation has continuously been achieved for these 25 years. The results and projected techniques are shown in Fig. 4 up to the year 2000. Techniques shown in the figure are selected to represent the major components of a typical process flow.

Although a transistor of 0.04-μm gate length has already been demonstrated operating at room temperature, many manufacturing issues remain to be solved for production of 1-gigabit DRAM-class devices and beyond. These are discussed in Sec. 2.3.

1. ELEMENTARY PROCESSES AND DEVICES FOR INTEGRATION

1.1 Elementary Processes

In terms of functions, the elementary processes in wafer fabrication can be categorized into four technology groups as

1. patterning,
2. isolation,
3. device formation, and
4. interconnection.

Each technology group consists of combinations of elementary processing techniques such as crystal growth and wafer preparation, epitaxy, lithography, wet and dry etching, ion implantation and annealing, oxidation, diffusion, dielectric-film deposition, silicidation, and metallization.

In addition to these fabrication techniques, failure analyses and process simulations are essential for realizing higher yield and shortening the design cycle, especially during the development phase. Following fabrication and test, packaging technique (see PACKAGING TECHNOLOGY) is required to realize reliable IC products, which must be resistant to severe environments such as high temperature, high humidity, and mechanical stress. These topics are beyond the scope of this article, which is limited to wafer-fabrication techniques.

1.2 Patterning Technology

As previously shown in Fig. 4, one IC generation of three years duration begins at a feature size that is only 70% of that in the

		1970		1980				1990		FORECAST		2000
Feature Size (μm)		12	8	5	3	2	1.3	0.8	0.5	0.3	0.2	0.13
DRAM (bit)		1K	4K	16K	64K	256K	1M	4M	16M	64M	256M	1G
Memory Cell	STRUCTURE	3T	1T, PLANAR					STACK or TRENCH				(3D)
	INSULATOR THICK.(nm)*1	120	100	50	35	20	10	8	5	4	3	2
Device	POWER SUPPLY(V)	20	12		5				Int.3.3	3.3 - 1.5		
	TRANSISTOR	pMOS	nMOS					CMOS			(BiCMOS)	
	DRAIN STRUCTURE *2	SD				DD		LDD	(IMPROVED)		(VERTICAL)	
	OXIDE THICK.(nm)	120	100	75	50	35	25	20	15	12	8	5
	CHANNEL LENGTH(μm)	8	5	3	2	1.3	0.8	0.8	0.5	0.3	0.2	0.15
	JUNCTION DEPTH (μm)	1.5	0.8	0.5	0.35	0.3	0.25	0.2	0.15	0.12	0.1	0.08
Process	LITHOGRAPHY *3	CONTACT			1/1PJ	1/10PJ	1/5PJ	(g-LINE)	(i-LINE)	(EXCIMER)		(X/EB)
	ETCHING *4	SOLUTION				PLASMA, RIE			ECR, RIE		HDP	
	ISOLATION	PLANAR	LOCOS				(IMPROVED)		RECESSED		TRENCH	
	GATE MATERIAL	Al	POLYSi	DOUBLE POLY		POLYCIDE		(Al-SHUNT)		REFRACTORY		
	METALLIZATION	Al		Al-Si			Al-Si-Cu (TiW, TiN)				W	Cu
	WAFER SIZE ("φ)	2	2.5	3	4	5	6	6-8	8	8-12 (SOI)		

*1: SiO$_2$ equivalent, *2: SD(Single Drain) / DD(Double Drain) / LDD(Lightly-Doped Drain),
*3: PJ(Projection) / X(X-ray) / EB(Electron Beam), *4:RIE(Reactive-Ion Etching) /
ECR(Electron Cyclotron Resonance) / HDP(High-Density Plasma)

FIG. 4. Technology road map mainly for MOS ICs. Selected techniques beyond 64-Mbit DRAM are forecast.

previous generation. Therefore patterning is an enabling process for each generation.

Figure 5 shows practical resolutions as a function of wavelength λ for three major lithographies (Okazaki, 1993): optical, x-ray, and EB (electron beam). Since the feature size of integrated circuits that are now under volume production is greater than 0.35 μm, optical reduction projection steppers (step-and-repeat–type projection aligner) with *g*-line (λ = 0.436 μm) and *i*-line (λ = 0.365 μm) light from a mercury-arc source offer enough performance for production.

Two of the most important fundamental performance factors of optical lithography are resolution R and depth of focus DOF. They are expressed as

$$R = k_1\lambda/\mathrm{NA} \qquad (2)$$

and

$$\mathrm{DOF} = k_2\lambda/\mathrm{NA}^2, \qquad (3)$$

where λ and NA are wavelength of incident light and numerical aperture of lens, respectively, and k_1 and k_2, which are factors defined by photoresist and exposure conditions, are typically 0.6 and 0.5, respectively. It must be noted that increasing NA for the purpose of improving resolution gives rise to an accelerated decrease in DOF. The most advanced stepper using *i*-line light achieves a NA of 0.6 in production today.

Beyond 0.35 μm, a stepper with excimer-laser light source of KrF (λ = 0.248 μm) or ArF (λ = 0.193 μm) should be a potential candidate. Further beyond the 0.2-μm region, F_2 (λ = 0.154 μm), x ray, or EB has a potential performance. Even though an x-ray exposure system with synchrotron-radiation light source has great advantages such as high resolution, high throughput, and large DOF, open issues still remain regarding difficulty in precise mask making and high investment cost. Despite its excellent resolution capability, an x-ray reduction projection system suffers from technical difficulty in implementation of adequate reduction lens and mirror. The electron-beam system has been

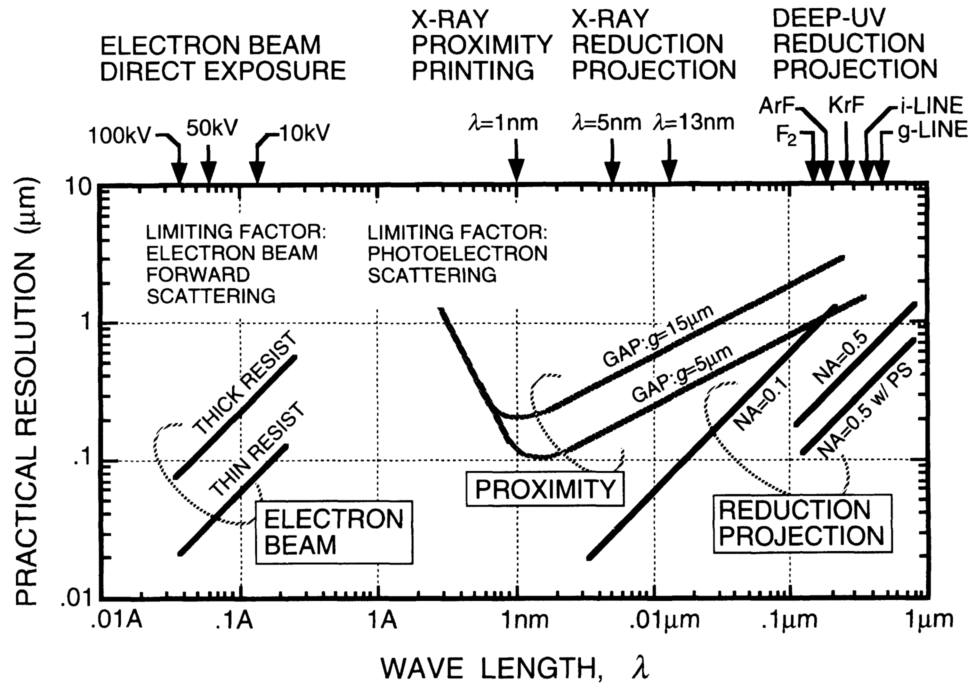

FIG. 5. Practical resolutions of electron-beam, x-ray, and deep-UV lithographies (Okazaki, 1993). Resolution of patterning R is practically proportional to λ/NA for reduction projection system, and to $(\lambda g)^{1/2}$ for proximity printing system, where λ, NA, and g are wavelength of incident light, numerical aperture of lens, and gap length between mask and resist to be exposed. Ultimate attainable resolutions are fundamentally limited by photoelectron scattering for x ray and by forward scattering of electron beam for EB. PS in reduction projection system denotes phase-shift technology, which is discussed later.

widely used to fabricate photomasks for the optical stepper, but its very low throughput is still not capable of volume production.

Besides the technical performance, the cost/performance factor is also very important from the economic viewpoint. Empirically, the practical resolution R of aligners is hyperinversely proportional to its price. For example, ratios of R/price are (2 μm)/(\$0.2 × 10^6), (0.3 μm)/(\$$10^6$), (0.2 μm)/(\$2 × 10^6), (0.1 μm)/(\$4 × 10^6), and 0.1 μm/(10 × 10^6) for a 1:1 stepper, an *i*-line stepper, an excimer stepper, an EB machine, and a synchrotron-radiation x-ray system, respectively. Therefore, it is a key issue to extend the resolution limit without changing and/or improving the present aligner. Potential methods are shown in Fig. 6. They are candidates for extending the resolution limit to the next generation. Multilayer resist (MLR), Fig. 6(a), makes use of the better resolution with thin and flat photoresist; contrast-enhanced layer (CEL), Fig. 6(b), accelerated generation in transparent area; focus-latitude enhancement exposure (FLEX), Fig. 6(c), multiple exposures with different focal points for thick and stepped photoresist; and phase-shifting mask (PSM), Fig. 6(d), contrast enhancement by interference between adjacent beams.

Simulated resolution enhancements using a phase-shifting mask are shown in Fig. 7. While keeping a DOF value of ±0.75 μm, resolution limit size can be reduced to 70% compared with a conventional mask. The coherence factor σ of the incident light is chosen to be 0.5 in this simulation.

1.3 Planarization Technology

As the integration level goes further, multilevel metal wirings have become increasingly important. At present three- or four-level metallization is essential to logic ICs. Since deposition and patterning of metal wires are difficult to carry out on a stepped

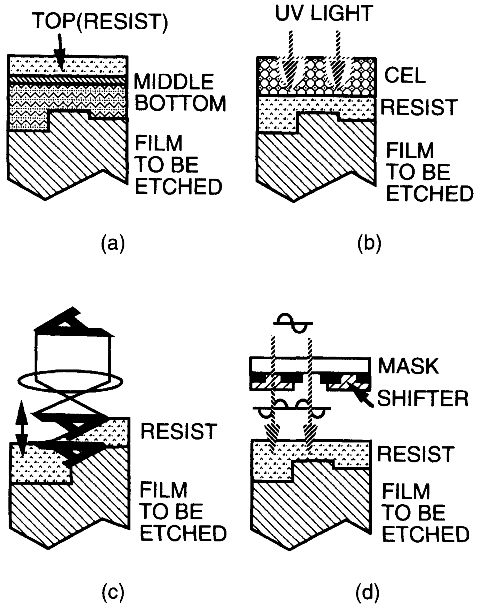
TOP(RESIST)
MIDDLE
BOTTOM
FILM
TO BE
ETCHED
(a)
UV LIGHT
CEL
RESIST
FILM
TO BE
ETCHED
(b)
RESIST
FILM
TO BE
ETCHED
(c)
MASK
SHIFTER
RESIST
FILM
TO BE
ETCHED
(d)

FIG. 6. Proposed methods to enhance resolution: (a) multilayer resist; (b) contrast-enhancement layer; (c) focus-latitude–enhancement exposure; and (d) phase-shifting mask.

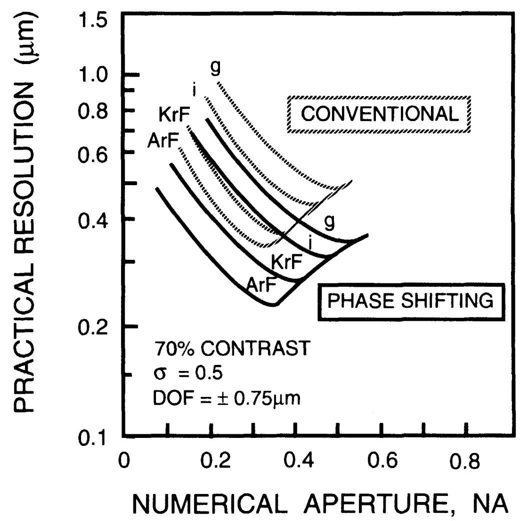

FIG. 7. Simulated resolution enhancement with phase-shifting mask for *g* line of Hg (wavelength λ = 436 nm), *i* line of Hg (λ = 365 nm), KrF excimer (λ = 248 nm), and ArF excimer (λ = 193 nm).

surface, planarization of the insulating film underneath the metal wires is essential. Planarization is also very effective for the finer lithography with its decreased DOF in an advanced stepper having a high-NA lens.

Trends in the number of metallization levels and the aspect ratio of patterns delineated are shown in Fig. 8. As the level increases, the step height on the die surface also increases, resulting in the degradation of focus latitude in lithography and metal thinning due to poor step coverage. To cope with these problems, several techniques are developed to fill narrow gaps and deep vias

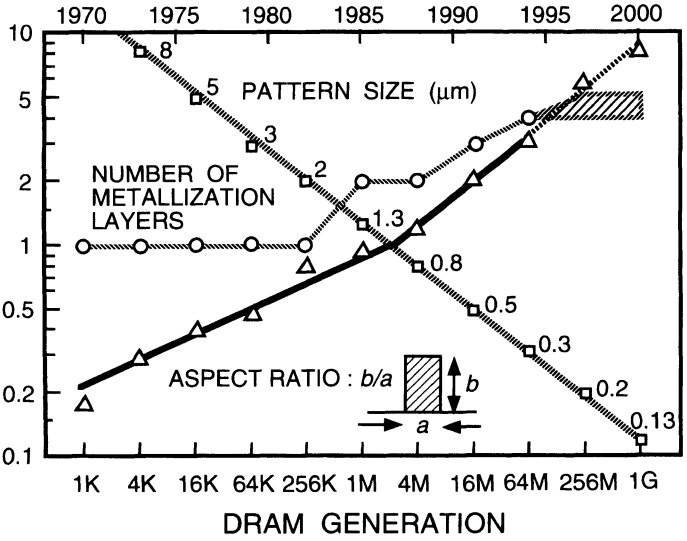

FIG. 8. Increased difficulty in patterning in terms of metallization layers and aspect ratio of patterns to be delineated.

with dielectric films and/or metals. They are the following:

1. ECR (electron cyclotron resonance) sputtering, which results in improved step coverage by taking advantage of high surface mobility of high-energy particles;
2. conformal glass-film deposition with LP-CVD (low-pressure, high-temperature chemical vapor deposition), taking advantage of long mean free path at low pressure;
3. SOG (spin-on-glass), in which an organic or inorganic silicon-compound material in an alcoholic solvent is spin-coated providing excellent gap-filling performance;
4. reflow of BPSG (boro- and phosphosilicate glass, with mole concentrations of boron and phosphorus around 20% and reflow temperature ranging from 800 to 900 °C); and
5. CMP (chemical mechanical polishing).

An example of planarization using CMP, which has increasingly become popular in LSI manufacturing, is shown in Fig. 9. Protruding areas are preferentially polished and simultaneously etched by chemical etchant as shown in the etching-model drawings in the figure. The planarization result is outstanding.

2. PROCESS INTEGRATION TECHNOLOGIES

2.1. Isolation Structure

The isolation structure in LSI has the role of electrical isolation for adjacent transistors. This is "passive performance" as compared with "active performance" of a transistor. The size of a transistor is defined by the performance required on the basis of its drive current, previously described in Eq. (1). While the smaller the better for the isolation region, the isolation region usually becomes bigger than the transistor region.

The most popular isolation technique for more than twenty years has been LOCOS (local oxidation of silicon). As shown in Fig. 10(a), a silicon-nitride (Si_3N_4) film works as a mask against high-temperature oxidation. Oxidation in steam at around 1000 °C for a few tens of minutes produces a several-hundred-nanometer-thick field oxide. Implanted boron forms a p^+ region working as a channel stopper between adjacent transistors.

One major problem associated with LOCOS is "encroachment" of field oxide beneath the nitride mask, causing narrowing of the transistor channel. Since width of the narrowed region, denoted by W_{eff}, is the channel width of the transistor, a pre-enlarged width of W_{mask} should be designed to maintain the minimum channel width. The structures obtained, shown in Fig. 10(c) with "isoplanar" technique, are named "bird's beak" and "bird's head." The isoplanar technique utilizes a recessed silicon surface to keep the entire surface relatively flat as compared with LOCOS.

When the pad oxide beneath the nitride becomes thinner, the bird's beak can be reduced. However, it frequently causes dislocations along field-oxide edges during high-temperature field oxidation due to the huge tensile stresses in the nitride films. The pad oxide works as a stress buffer region between the nitride and the silicon substrate.

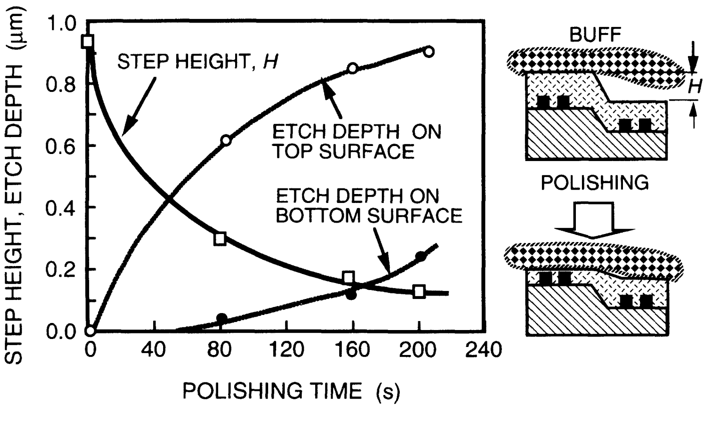

FIG. 9. Step-height reduction with chemical mechanical polishing (CMP).

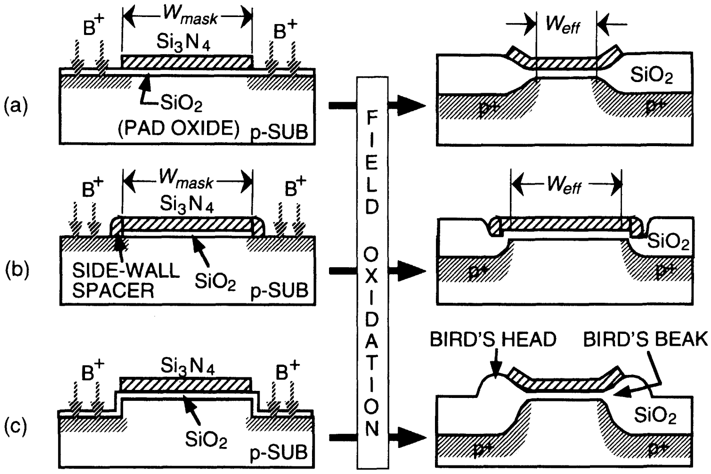

FIG. 10. LOCOS (local oxidation of silicon) structures: (a) conventional LOCOS formation technique; (b) improved LOCOS structure utilizing silicon-nitride side-wall spacer; and (c) conventional isoplanar technique.

An improved LOCOS structure is shown in Fig. 10(b). A side-wall spacer of silicon nitride works as an oxidation stopper resulting in a larger channel width. As the process window of the improved LOCOS is narrower than that of the conventional LOCOS, processing conditions must be carefully chosen to control dislocation formation.

A much improved isolation structure is the deep and narrow trench isolation, previously shown in Fig. 3. There still exist a lot of issues to be solved, such as deep and narrow trench etching, filling materials and techniques, channel-stop implantation into the trench, etc.

2.2 Device Structure

Figure 11 shows a typical process sequence for an advanced submicron *n*-channel MOS transistor. Key issues are realization of larger effective channel width with an advanced isolation technique and the increase in maximum applicable voltage across source to drain.

Its isolation structure is realized by a technique already shown in Fig. 10(b). The thicker side wall spacer and thicker field oxide cause dislocations beneath the field oxide giving rise to anomalously increased leakage current. To cope with its relatively narrow process window, thicknesses of the pad oxide beneath the side-wall spacer and field oxide have to be controlled.

This device structure is realized with a lightly doped drain (LDD) (Ogura *et al.*, 1980). The relatively lower impurity concentration of the n^+ layer reduces the electric field strength, resulting in increased breakdown voltage across source and drain. A role of the side-wall spacer, shown in Fig. 11(f), is to keep high–impurity-concentration layers of n^{++} away from the front edges of source and drain. Even though this LDD structure improves the breakdown voltage, this brings an adverse effect on the drivability of the transistor as if a series resistance is inserted in the equivalent circuit. It is a trade-off between breakdown voltage and drivability.

2.3 CMOS Process Integration

CMOS ultralarge-scale integration (ULSI) has played an increasingly important role across the entire electronics industry. Major products are memories and microprocessors as previously shown in Fig. 1. Some modern microcomputers incorporate logic and high-speed memory in the same chip. To meet the needs of specific applications, nonvolatile memory can be included to achieve a user-programmable microcomputer.

2.3.1 DRAM Process Integration The equivalent circuit of a DRAM memory cell is very simple, consisting of one MOS transistor and one storage capacitor. The most important performance factor is expressed as C_s/C_b, where C_s is the storage capacitance and C_b is the bit-line capacitance. Ordinarily, 128 or 256 memory cells are connected to one bit line. Historically the value of C_s/C_b

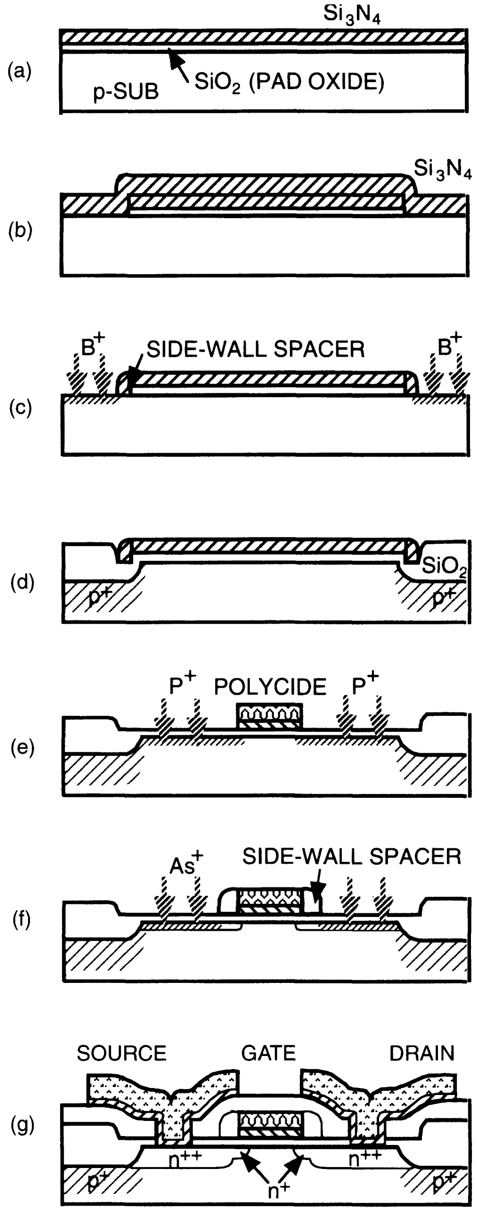

FIG. 11. Process sequence for an advanced n-channel MOS transistor: (a) pad oxide and first silicon-nitride film deposition on p-type, (100)-oriented silicon substrate; (b) second silicon-nitride film deposition subsequent to delineation of the first for active transistor region; (c) anisotropic dry etching of the second nitride film to leave nitride side-wall spacer and channel-stop implant of boron; (d) field-oxide formation; (e) polycide (silicide/polysilicon) gate delineation and phosphorus implant to form n^+ lightly doped drain; (f) arsenic implant to form n^{++} source and drain with oxide side-wall mask; and (g) glass film is deposited and aluminum/barrier material contacts are formed.

and the signal-voltage swing of the bit line have been chosen to be about 0.1 and 100–200 mV, respectively, over all DRAM generations including considerations for operational margin, soft-error immunity, refresh performance, etc. The storage capacitance C_s is expressed as

$$C_s = \epsilon_i A / T_i, \tag{4}$$

where ϵ_i, A, and T_i are permittivity of the capacitor insulator, area of the capacitor electrodes, and thickness of the capacitor insulator, respectively. Even though the memory cell has been reduced in size, namely, the capacitor area A has been reduced, the storage capacitance cannot be reduced proportionally from operational margin and reliability points of view.

Alpha particles can also be fatal to DRAM operation by causing soft errors. Almost 200 fC of charge, which is generated by an alpha particle having a maximum energy of 5 MeV, is dispersed 25 μm deep into the Si substrate. To keep a sufficiently low soft-error rate, a storage capacitance greater than 20 fF is needed unless an innovative structure is adopted.

Trends in DRAM structural parameters are shown in Fig. 12. From 4 kbit to 1 Mbit, planar capacitors were adequate, whereas three-dimensional (3D) capacitors are needed to keep sufficient storage capacitance beyond 4 Mbit. As clearly shown in the figure, the capacitor area surpasses the cell size after the 4-Mbit generation. A cross-sectional SEM photograph is shown in Fig. 13 for an advanced 16-Mbit stacked-capacitor DRAM cell having double-fin–type CUB (capacitance under bit line) scheme. BPSG reflow under the bit line and SOG over the bit line are utilized for planarization.

The 3D capacitor was introduced in response to the inability to make adequate quality dielectric films that scale with area A according to Eq. (4). Proposed 3D capacitors for a 64-Mbit DRAM are shown in Fig. 14. A simple 3D capacitor with ordinary SiO_2–Si_3N_4–SiO_2 capacitor insulator may not be large enough for the 64-Mbit DRAM and beyond.

Meanwhile, in the 256 Mbit chip and the future, rapid reduction in the thickness is supposed to be required again. One approach is the adoption of a novel insulator

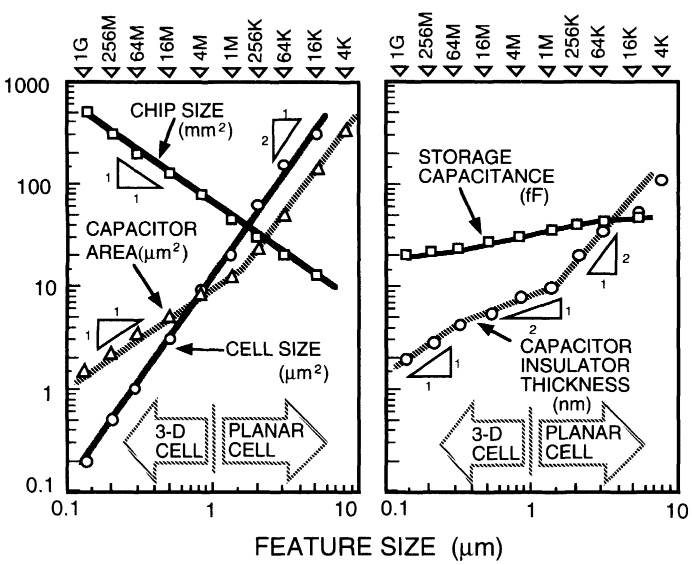

FIG. 12. Device-parameter trends in DRAM. They are divided into two regimes: planar cells before 1 Mbit and three-dimensional cells after 1 Mbit. Reduction of capacitor-insulator thickness is categorized into three regimes in terms of keeping sufficient storage-capacitance value: (1) 4 K to 1 M, capacitance is achieved mainly by the thickness decrease; (2) 1 to 64 M, it is done mainly by structure innovation, such as stacked-capacitor and trench-capacitor cells; (3) 64 M and beyond, both thickness decrease and structure innovation tend to be less effective, and material innovation is likely to be required. Capacitor-insulator thickness stands for SiO_2-equivalent thickness.

having much higher permittivity. Ta_2O_5 and $(Ba,Sr)TiO_3$ films are two of the potential candidates as shown in Fig. 15. The product of dielectric constant and breakdown field strength is nothing but the maximum storable charge per unit area, Q_{max}. It is to be noticed that the relative Q_{max} varies from 1 to 3 over the various kinds of dielectric films plotted inside the shaded region in the figure.

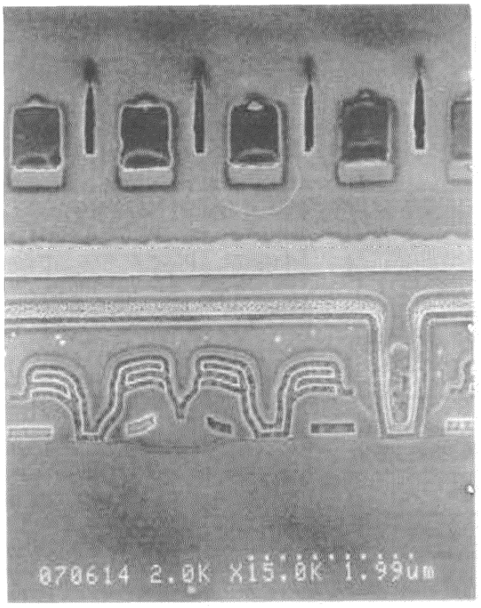

FIG. 13. Cross section of an advanced 16-Mbit DRAM cell that uses double-fin CUB (capacitance under bit line) structure.

A typical R&D-type process sequence for 16-Mbit DRAM is shown in Figs. 16–20. It is no exaggeration to say that almost all manufacturers have been developing process sequences different from each other based on their past experiences. Regarding memory-cell structures, both stacked and trench capacitors are commercially available today. Therefore, there may not be a unique and/or standard solution. Furthermore, a couple of process sequences have been developed for various generations in the same DRAM product, as shown later in Fig. 24.

In this case, eighteen photolithographic steps are defined to complete the sequence. In actual DRAM products, it is natural that more photolithographic steps may be added to produce different kinds of transistors having different threshold voltages and additional metal wiring.

2.3.2 Logic Process Integration Logic ICs have different aspects in their performance requirements compared with memories. In the same 0.8-μm feature-size generation, an average device density is 100 μm^2/device for logics and 10 μm^2/device for DRAM. Wiring density is a key issue to make dense packing in logic LSIs, while memory needs a straightforward increase in number of memory cells. The wiring complexity of a

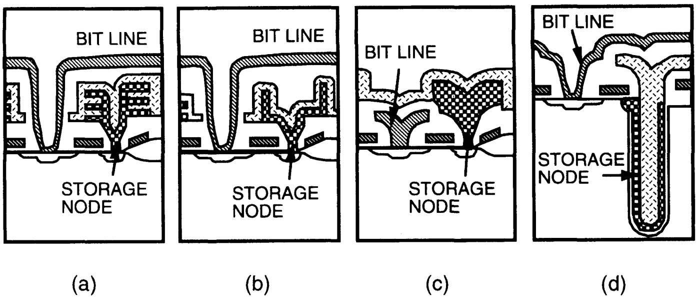
BIT LINE
STORAGE NODE
(a)
(b)
(c)
(d)

FIG. 14. Proposed 64-Mbit DRAM cells: (a) triple-fin type with a CUB (capacitor under bit line) structure; (b) crown type for CUB; (c) rugged polysilicon with COB (capacitor over bit line); and (d) shielded storage-node–type trench capacitor.

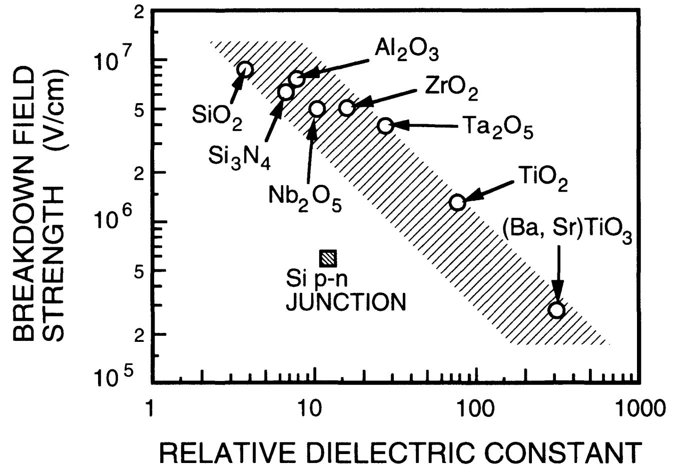

FIG. 15. Potential high-permittivity materials to be used for capacitor insulator in DRAM cells.

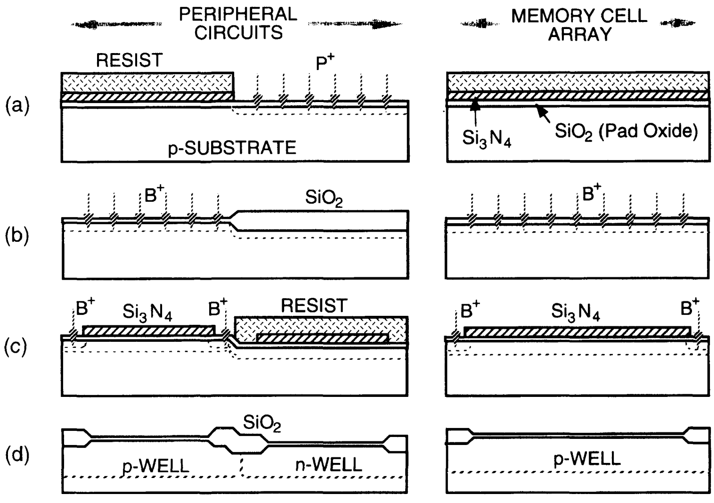
PERIPHERAL CIRCUITS
MEMORY CELL ARRAY
RESIST
P+
p-SUBSTRATE
Si3N4
SiO2 (Pad Oxide)
B+
SiO2
Si3N4
RESIST
p-WELL
n-WELL
(a)
(b)
(c)
(d)

FIG. 16. Twin-well formation step of typical R&D-type process sequence for 16-Mbit DRAM. (a) Selective phosphorus implant into *p*-type, (100)-oriented substrate using photoresist mask (the first photolithography process), for *n*-well formation; (b) field-oxide formation and nonselective boron implant for *p*-well formation; (c) selective nitride deposition with the second photolithography and selective boron implant to form channel stopper using photoresist mask (the third photolithography); and (d) field-oxide formation, *p*- and *n*-wells drive-in, and gate-oxide formation.

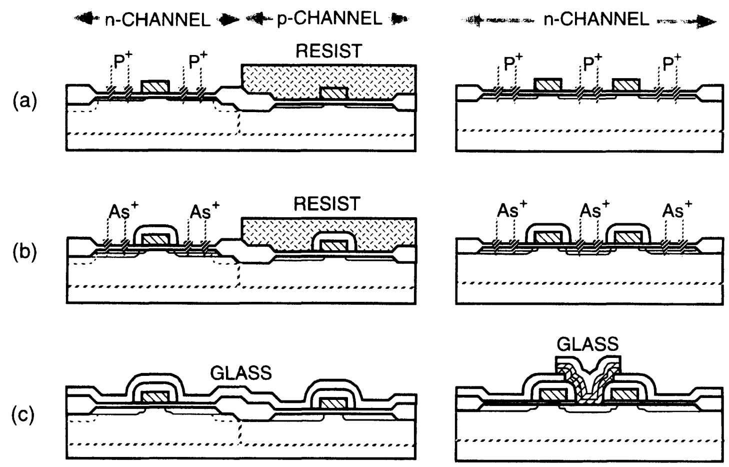

FIG. 17. Transistor-formation step of typical R&D-type process sequence for 16-Mbit DRAM: (a) selective polycide (silicide/phosphorus-doped polysilicon) deposition (the fourth photolithography), selective boron implant into *p*-channel region to form lightly doped drain (fifth photoetching), and selective phosphorus implant into *n*-channel region to form lightly doped drain (the sixth photolithography); (b) selective arsenic implant to form *n*-channel source and drain with oxide sidewall spacer using photoresist mask (the seventh photolithography); and (c) glass-film deposition, the eighth photolithography to form bit-line contact, and selective deposition of polycide bit line (the ninth photolithography).

logic IC theoretically increases with the number of gates as $N^{1.5}$. Therefore, multilayer metallization is very efficient to reduce the die size. For example, the die area of some logic LSIs is reduced to 0.45 by four-level metals and 0.6 by three-level metals as compared with 1 for double-level metals. This die shrinkability can easily overcome the cost increase incurred by more-level metallization.

The key issues for logic IC processes are

1. planarization (see Sec. 1.3) of the underlying insulator and metal films, and
2. low-ε dielectric films resulting in higher switching speed and lower power consumption.

Films such as fluorine-incorporated CVD (chemical vapor deposited) SiO_2, organic SOG (spin-on-glass), and polyimide resins are potential candidates for low-ε dielectric films.

2.4 Bipolar Process Integration

A typical sequence for the present self-aligned bipolar IC is shown in Fig. 21. One of the key objectives in achieving higher speed is to decrease the base resistance

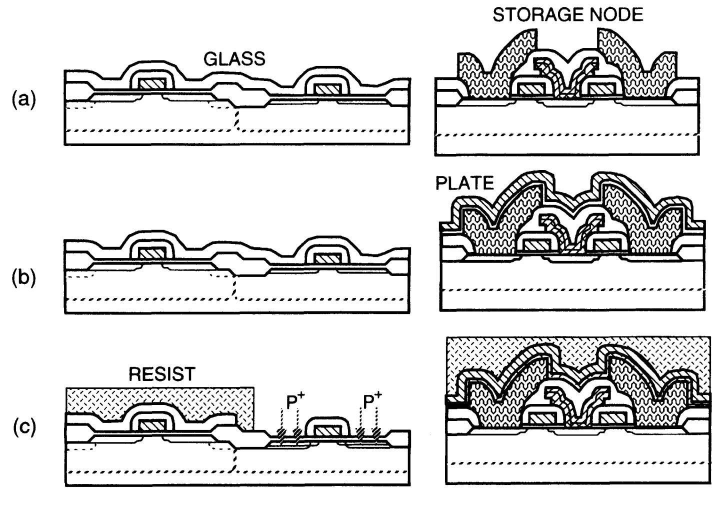

FIG. 18. Memory-cell–formation step of typical R&D-type process sequence for 16-Mbit DRAM: (a) glass-film deposition, storage-node contact formation with tenth photoetching, selective arsenic-doped polysilicon deposition (the eleventh photolithography) to form storage node; (b) very thin capacitor-insulator deposition and selective phosphorus-doped polysilicon plate formation (the twelfth photolithography); and (c) selective boron implant to form *p*-channel source and drain using photoresist mask.

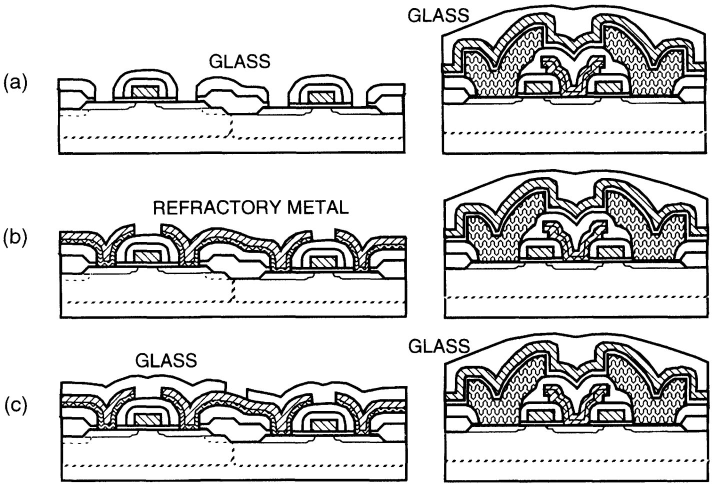

FIG. 19. Metal-wiring formation step of typical R&D-type process sequence for 16-Mbit DRAM: (a) glass-film formation with boro- and phosphosilicate glass (BPSG) reflow and the thirteenth photolithography delineates contact holes to source and drain region; (b) selective first metal wiring of tungsten/titanium-nitride electrode (the fourteenth photolithography); and (c) phosphosilicate glass/spin-on-glass/phosphosilicate glass deposition and subsequent annealing for the densification and first via-hole formation (the fifteenth photolithography).

while keeping the base region thin. Today, bipolar base width can reach below 50 nm and cutoff frequencies approaching 80 GHz have been realized. This high-speed performance is getting rather close to the GaAs microwave transistor. Furthermore, a champion performance of 113-GHz cutoff frequency has recently been achieved with a leading-edge heterobipolar transistor having a SiGe base (Crabbe *et al.*, 1993).

As the demand for downsized computers becomes stronger, very high-speed bipolar ICs have been losing their market share as previously shown as "Bipolar Digital" in Fig. 1. This is partly because they consume huge power to achieve ultrahigh speed and require a very expensive liquid cooling system. Moreover, the high power, sometimes reaching 100 W/chip, does not allow high-density integration, leading to the use of costly multichip modules. Such multichip modules suffer from signal delays due to the relatively long wiring between chips. So, speed and power interact adversely in a high-cost technology, and densely packed CMOS ULSI is expected to outperform the bipolar ICs in the future even if the MOS transistor itself is less speedy than the bipolar transistor.

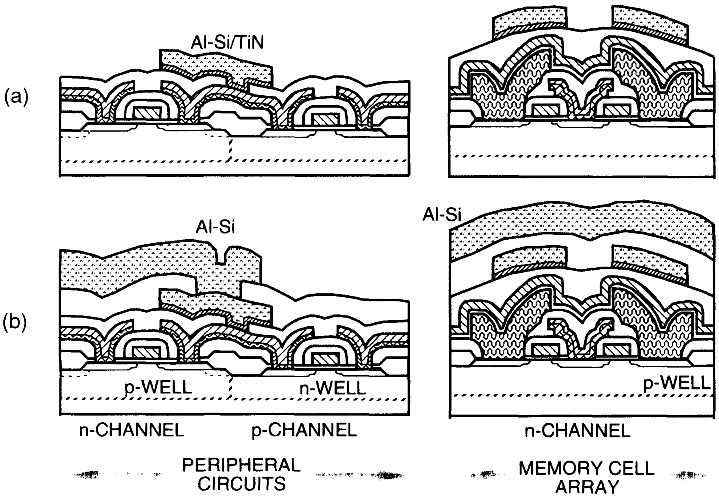

FIG. 20. Multilayer metallization step of typical R&D-type process sequence for 16-Mbit DRAM: (a) Selective deposition of second metal wiring of silicon-contained aluminum/titanium nitride (the sixteenth photolithography); and (b) plasma-oxide/spin-on-glass/plasma-oxide intermediate insulator formation, second via-hole formation with the seventeenth photolithography, and selective aluminum-wiring formation with the eighteenth photolithography.

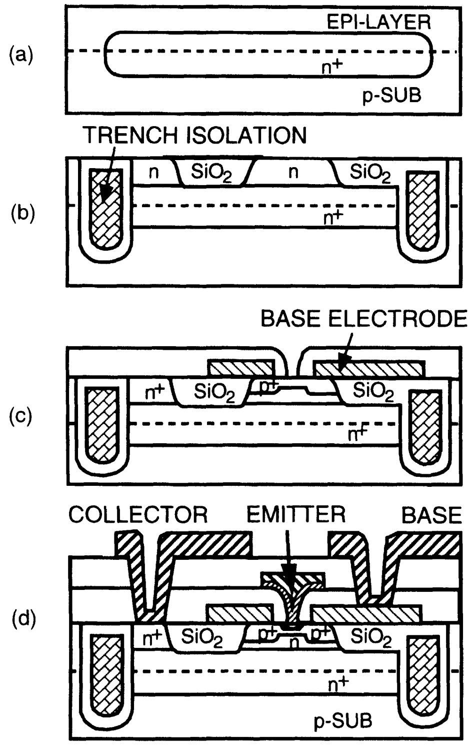

FIG. 21. Process sequence for a self-aligned bipolar IC: (a) forming n^+ buried layer on *p*-type substrate and *n*-type epitaxial growth; (b) etching deep trench, filling with semi-insulating polysilicon, and LOCOS (local oxidation of Si) isolation; (c) boron-doped polysilicon patterning to form the base region and base electrode, glass-film deposition, and anisotropic dry etching to define the emitter window; and (d) arsenic-doped polysilicon deposition to form the emitter electrode, contact-hole etching, and aluminum-electrode formation.

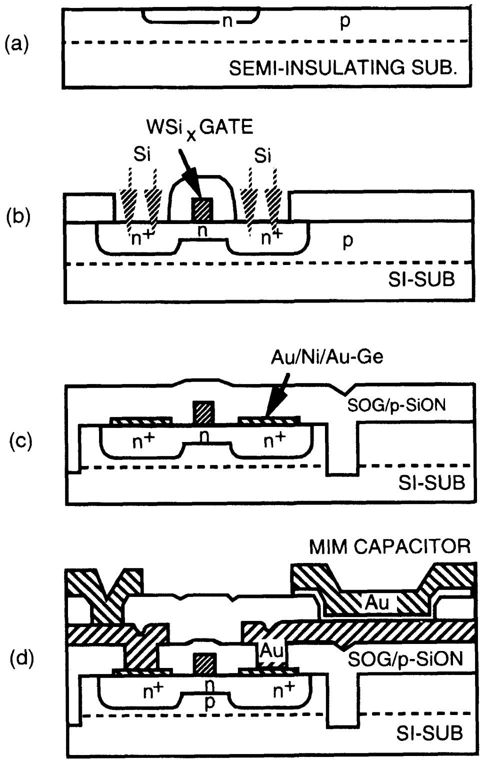

FIG. 22. Process sequence for GaAs MESFET IC: (a) selective *n*-type channel region formation subsequent to Mg-implanted *p*-type layer formation on semi-insulating substrate; (b) transistor region definition, side-wall spacer formation for WSi_x gate, source and drain region formation with Si implantation; (c) Au/Ni/Au–Ge contact formation on source and drain regions, trench etching for isolation, and Au-electrode formation; (d) intermediate insulator deposition, via-hole etching, and simultaneous formation of metal-insulator–metal capacitor and subsequent metallization.

2.5 GaAs Process Integration

Even if a leading bipolar device has already reached the cutoff-frequency region of 100 GHz, the GaAs FET still possesses the advantages of low-noise and ultrahigh-frequency operation. About double the operation speed is realized compared with silicon bipolar LSIs. Figure 22 shows an example of a process sequence for a GaAs MES (metal semiconductor) FET IC.

Unlike silicon, crystalline GaAs neither has a stable insulator nor can it withstand high temperature. Except for a maximum temperature of around 800 °C at impurity activation, ordinary processing temperature is around 400 °C. This is a distinct disadvantage in terms of integrity, reliability, and planarization. For example, a CVD SiO_2 film formed at around 400 °C causes poor step coverage, less resistance to chemical etchant due to its enhanced porosity, and very low dielectric field strength, compared with those of that formed at 800 °C. However, small-scale GaAs ICs have been used in a very limited area where ultrahigh-speed operation is required.

3. FURTHER INTEGRATION TECHNOLOGIES

During the past thirty years, continuous innovation has been a characteristic of silicon IC technology. Despite four times

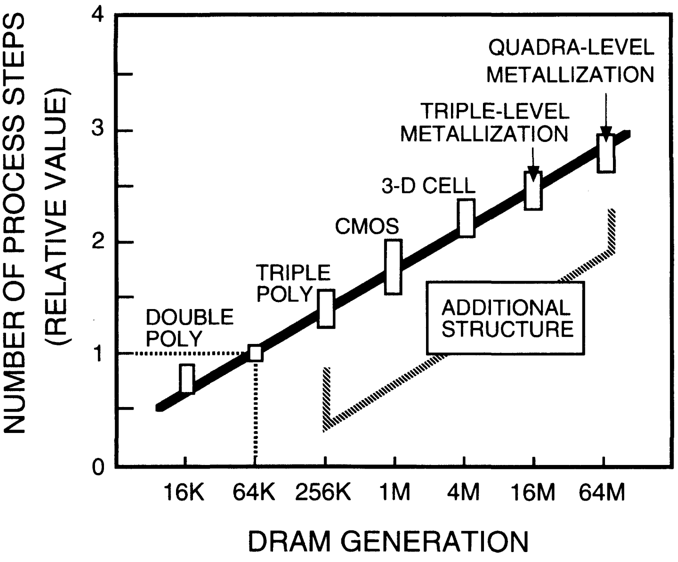

FIG. 23. Increase in number of process steps for DRAMs in response to their structural innovations.

greater integration per generation, prices of DRAMs have remained nearly constant throughout several generations. As a result, bit price has been reduced to 1/1000 from 1 kbit to 1 Mbit. This has certainly led to very rapid growth of the semiconductor industry because of very extensive usage of semiconductor products in various industries.

Increased chip cost due to enlarged die size has been compensated by wafer-size increases. Therefore, if the wafer size ceases to increase, there may not exist any method to compensate for the increase in wafer-fabrication cost caused by process-step increase and the growing investment in new fabrication facilities.

An example of process-step increase is shown in Fig. 23 for the case of DRAM. Difficulty in preventing a decrease in storage capacitance will become fatal in the future. In that era, memories that operate on the basis of different mechanisms will be dominant and may take over the DRAM market assuming that they will be produced less expensively than DRAM.

The candidates might be FRAM (ferroelectric RAM) and/or innovative memories unknown today. Since FRAM utilizes a

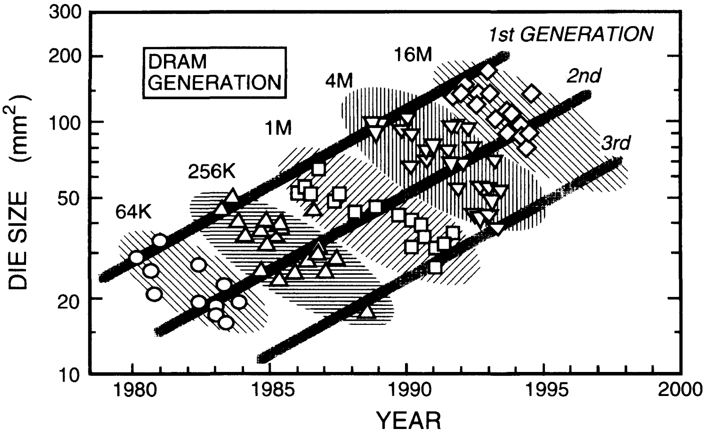

FIG. 24. Die-size reduction for three groupings within a DRAM generation. The third generation delivers almost one-half of the die size as compared with the first.

mechanism of polarization inside a ferroelectric film such as PZT and PLZT, it inherently works as both RAM and ROM. Furthermore, it is switchable by applying a voltage to the plate.

Flash EEPROM has a potential for low-cost memory by virtue of its simple one-transistor memory cell. However, it is not random-access memory but still read-only memory. If its price is low enough, it will be able to take over some part of the mass-storage market. Its operation mechanism is based on charging and discharging of high-energy electrons across thin SiO_2. Macroscopic and microscopic oxide integrities will be key issues for volume production.

4. CONCLUSION

As the semiconductor industry has continuously been growing for these 30 years, device processing has become more and more important. Device count in an IC product has also increased very rapidly resulting in more functions.

Investment in production facilities has increased more rapidly than semiconductor prices. To cope with this dilemma, wafer size has been periodically increased. To get more chips from a wafer, cut-down or die-shrink products have become popular for all IC products. A typical trend is shown in Fig. 24 for DRAMs. Three die-size groupings within each respective DRAM generation are found for products available in the market.

GLOSSARY

Chip Shrink (Model): Reduction of chip size with optically shrunk photomasks to get more chips on a wafer.

Cut Down (Model): Manufacture of a present-level product utilizing redesign and the integration process of the next generation—for example, 4-Mbit DRAM (ordinary with 0.8-μm feature size) manufactured by 0.5 μm (primarily used for 16-Mbit DRAM production). Generally smaller than "chip-shrink" model.

Feature Size: Minimum dimension, which generally denotes the smallest gate length of MOS transistor. It also stands for a level of process generation.

Integration Process: Process sequence to fabricate a semiconductor product, consisting of a specific combination of various elementary processes.

Isolation: Recessed regions on a substrate surface filled with dielectric films such as SiO_2 to isolate active and passive devices in integrated circuits. Typical techniques are LOCOS and trench isolations.

Lithography: Methods to delineate patterns using one of several kinds of aligners such as contact, projection, step-and-repeat, and step-and-scan with ultraviolet light, x rays, or electron beam.

Planarization: Methods to planarize dielectric films underlying metal wires. They include bias sputtering, chemical mechanical polishing (CMP), reflow of CVD oxide, spin-on-glass (SOG), etc.

Plug: Method to fill a via hole with tungsten or other conductive materials to form a contact between two-level conductive layers. Also a metal piece left in the hole.

Polycide: Stacked electrode consisting of a silicide over polysilicon. Polysilicon plays the role of a reliable gate material, and silicide, a low-resistance material.

Process Generation: Representative integration process designated by a feature size such as 8, 5, 3, 2, 1.3, 0.8, or 0.5 μm. DRAM has been a technology driver for 25 years.

Silicidation: Method to form low-resistive layer of silicide consisting of silicon and a metal such as W, Mo, Co, Ti, Ta, etc. A self-aligned silicidation is often called "salicide."

Soft Error: Volatile operation failure caused by an alpha-particle hit. Substantially pronounced in DRAMs; however, the potential also exists in other LSIs.

Three-Dimensional Cell: Vertically formed storage cell for DRAMs such as stacked capacitor and trench capacitor.

Works Cited

Crabbe, E. F., Meyerson, B. S., Stork, J. M. C., Harame, D. L. (1993), "Vertical Profile Optimization of Very High Frequency Epitaxial Si- and SiGE-Base Bipolar Transistor," in: *IEEE International Electron Devices Meeting Technical Digest,* New York: IEEE, pp. 83–86.

Dennard, R. H., Gaensslen, F. H., Yu, H.-N., Rideout, V. L., Bassous, E., LeBlanc, A. R. (1974), "Design of Ion-Implanted MOS FET's with Very Small Physical Dimensions," *IEEE J. Solid-State Circ.* **SC-9,** 256.

Hoerni, J. A. (1960), "Planar Silicon Transistor and Diodes," presented at IRE Electron Devices Meeting, Washington, D.C.

Kilby, J. S. (1976), "The Invention of the Integrated Circuit," *IEEE Trans. Electron Dev.* **ED-23,** 648.

Ogura, S., Tsang, P. J., Walker, W. W., Critchlow, P. L., Shepard, J. F. (1980), "Design and Characteristics of the Lightly Doped Drain-Source (LDD) Insulated Gate Field-Effect Transistors," *IEEE Trans. Electron Dev.* **ED-27,** 1359.

Okazaki, S., (1993), "Lithographic Technologies for Future ULSI," *Appl. Surf. Sci.* **70/71,** 609–612.

Further Reading

IBM J. Res. Develop. (1995), Special Issue, "The IBM CMOS Technology," **39** (1/2).

Kooi, Else (1991), *The Invention of LOCOS,* New York: IEEE.

Sze, S. M. (1988), *VLSI Technology,* New York: McGraw-Hill.

SEMICONDUCTOR-DEVICE MODELING

Toru Toyabe, *Department of Information and Computer Sciences, Faculty of Engineering, Toyo University, Kawagoe, Saitama, Japan*

Introduction 199
1. **Basic Equations for Classical Device Modeling** 200
2. **Analytic Models for Semiconductor Devices** 201
2.1 *p-n*–Junction Diode 201
2.2 Bipolar Transistor 202
2.3 MOSFET 202
3. **Numerical Models for Semiconductor Devices** 204
4. **Physical Parameters** 204
4.1 Carrier Mobility 204
4.2 Generation–Recombination Rate 205
4.3 Auger-Recombination Rate 206
4.4 Impact-Ionization Rate 206
4.5 Band-Gap Narrowing Effect ... 207
5. **An Example Using the Drift-Diffusion Model** 207
6. **Energy-Transport Model** 207
7. **Monte Carlo Model** 210
7.1 Motivation 210
7.2 Electric Field Calculation 210
7.3 Free-Flight Calculation 211
7.4 Scattering Mechanisms 211
8. **Other Solution Methods to the Boltzmann Transport Equation** 212
8.1 Scattering-Matrix Approach ... 212
8.2 Cellular-Automaton Method ... 212
9. **Conclusions** 213
Glossary 213
Works Cited 214
Further Reading 214

INTRODUCTION

Since the bipolar transistor was invented by Bardeen, Brattain, and Shockley in 1947, semiconductor devices have developed at an incredible rate. The junction field-effect transistor (JFET) was introduced by Shockley in 1952, and the metal-oxide–semiconductor field-effect transistor (MOSFET) was proposed and fabricated by Kahng and Atalla in 1960. Moreover, the metal-semiconductor field-effect transistor (MESFET), the light-emitting diode, the laser diode, and the charge-coupled device (CCD) were proposed and developed in the succeeding decade.

The advent of the MOSFET was largely based on planar process technology. This has been undoubtedly the most important technology for the development of the integrated circuit in these past 30 years. A fourfold higher packing density has been realized every three years since 1970. As a result the present-day very large-scale integration (VLSI) integrates more than ten million transistors per single chip. Thanks to great advances of fine-pattern technology, 16-Mbit dynamic random-access memory (DRAM) is now in mass production with fabrication technology of minimum feature size of 0.5 μm.

A guiding principle of the MOS LSI development has been the scaling law (Dennard *et al.,* 1974). In early stages of development, the electrical behavior of MOSFETs obeyed well the scaling law that is based on early analytical device modeling, because of their large device size. For miniaturized devices, however, the deviation from the scaling law and classical device modeling has been remarkable. The short-channel effect is a typical example. The device miniaturization has also provoked problems relating to the reliability of LSIs, such as Complementary MOSFET (CMOS) latchup, characteristics degradation due to hot carriers, and alpha-particle–induced soft error (May and Woods, 1979). To cope with these problems, numerical device modeling based on fundamental physical equations for semiconductor devices has become necessary. At present, two-dimensional numerical simulation is commonly used in the development stage of device prototypes.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

The approach taken in early device modeling was analytic in the sense that closed-form solutions of the basic semiconductor equations are derived for different regions of the interior of the device and combined with boundary conditions between the regions to get a global model. Shockley derived an analytical model for bipolar transistors based on his *p-n*–junction theory in 1949. The bipolar transistors were divided into the emitter, base, and collector regions, and closed-form solutions of the diffusion equations for minority carriers were solved in each region and combined to get an entire bipolar transistor model.

An analytical model for junction field-effect transistors (JFETs) was also given by Shockley in 1952. The operation mechanism of JFETs is two-dimensional, since the direction of the channel current from source to drain is orthogonal to the direction of the controlling gate field. Shockley reduced this two-dimensional problem into two one-dimensional problems. This approach is called the gradual-channel model, since it is assumed that the electric field in the direction of channel is much smaller than the one in the direction of the gate-substrate. A gradual-channel model was also presented for MOSFETs by Ihantolla and Moll (1964). These analytical models were transparent and very useful thanks to the compact forms of their expressions.

The gradual-channel model, however, cannot be applied to recent miniaturized devices, because of the failure of the gradual-channel approximation. Thus, more accurate modeling based on numerical computation is indispensable for understanding current device operation and the design of modern devices. Numerical simulation of semiconductor devices started from the one-dimensional model of bipolar transistors by Gummel in 1964. In his work, it was shown that the coupled system of Poisson and current-continuity equations can be solved by an iterative computation within realistic CPU time. One-dimensional numerical simulation was applied to Gunn diodes and IMPATT diodes to analyze their operational mechanisms in the 1960s.

Two-dimensional numerical simulations that started in the 1970s have successively been applied to MOSFETs, JFETs, MESFETs, and bipolar transistors. Two-dimensional effects on the electrical characteristics, such as short-channel effects in small-geometry MOSFETs and high-level injection phenomena in bipolar transistors, were successfully analyzed. Moreover, three-dimensional numerical simulations became feasible in the 1980s in part thanks to the advent of supercomputers, and have been applied to analysis of narrow-channel effects in miniaturized MOSFETs and alpha-particle–induced soft-error phenomena in DRAM cells.

Silicon VLSI devices have been miniaturized to such a level that devices with the minimum feature size of 0.1 μm will be realized in the near future; however, limits of the conventional device structure and operation principle have become evident. To investigate the relevant physical phenomena, such as hot-carrier, nonlocal impact-ionization, and ballistic effects, and the device limits, various advanced models have been proposed and developed—e.g., the hydrodynamic model, the thermal hydrodynamic model, the Monte Carlo model, and the full-band Monte Carlo model. Moreover, more advanced treatments such as the scattering-matrix method, the cellular-automaton method, and quantum-mechanical analysis are applied to investigate the behavior of electrons in the crucial parts of future devices. In the following, these various device models are described.

1. BASIC EQUATIONS FOR CLASSICAL DEVICE MODELING

The classical analytic device models and the majority of the numerical device models commonly used are based on the drift–diffusion (DD) model. The carrier transport is expressed by the drift current due to the electric field and the diffusion current due to the spatial variation of the carrier density. Thus, the electron current density J_n and the hole current density J_p are expressed as

$$\mathbf{J}_n = qn\mu_n\mathbf{E} + qD_n\nabla n \tag{1}$$

and

$$\mathbf{J}_p = qp\mu_p\mathbf{E} - qD_p\nabla p, \tag{2}$$

where q is the electron charge, $\mathbf{E}$ is the electric field, and n, p, μ_n, μ_p, D_n, and D_p are the carrier densities, mobilities, and diffusion constants for electrons and holes, respectively. It is assumed that the Einstein relation holds for the mobility and diffusion constant:

$$D_{n,p} = (k_B T/q)\mu_{n,p}, \tag{3}$$

where T is the lattice temperature.

The basic equations in the DD model are the current-continuity equations for electrons and holes,

$$q\frac{\partial n}{\partial t} = \nabla \cdot \mathbf{J}_n + qG - qR, \tag{4}$$

$$q\frac{\partial p}{\partial t} = -\nabla \cdot \mathbf{J}_p + qG - qR, \tag{5}$$

and the Poisson equation

$$\epsilon \Delta \psi = q(n - p - N_D + N_A), \tag{6}$$

where G and R are the generation rate and the recombination rate of carriers, respectively, ϵ is the dielectric constant of the semiconductor, ψ is the electrostatic potential, and N_D and N_A are the ionized donor and acceptor concentrations, respectively. Physical mechanisms producing G and R are described in Sec. 4.

2. ANALYTIC MODELS FOR SEMICONDUCTOR DEVICES

Analytic models for representative devices such as diodes, bipolar transistors, and MOSFETs, which are clearly the basis for understanding the operational mechanisms, are derived by analytically solving the basic equations given above with appropriate simplifying assumptions.

2.1 *p-n*-Junction Diode

An ideal *p-n*-junction diode has a simple structure as shown in Fig. 1. The diode has rectifying characteristics in which it carries a large current in the forward-biased condition ($V > 0$) and nearly zero current in the reverse-biased condition ($V < 0$). To derive an analytical model for the diode, the basic equations of the drift–diffusion model are solved under the following assumptions. In the electrically neutral n and p regions outside of the depletion layer (with no mobile carriers) of the p-n junction, the injected minority carriers flow only by diffusion under the forward-biased condition. Then, the current-continuity equations (4) and (5) are simplified into simple diffusion equations by substitution using Eqs. (1) and (2), as

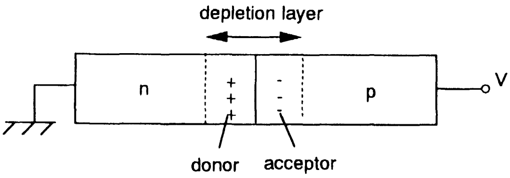

FIG. 1. *p-n*-junction diode.

$$D_n \frac{d^2 n}{dx^2} = \frac{n - n_{p0}}{\tau_n} \quad \text{in the } p \text{ region}, \tag{7}$$

$$D_p \frac{d^2 p}{dx^2} = \frac{p - p_{n0}}{\tau_p} \quad \text{in the } n \text{ region}, \tag{8}$$

where τ_n and τ_p are electron and hole lifetimes characterizing the recombination rate, and n_{p0} and p_{n0} are the equilibrium minority-carrier densities, respectively. The injected minority-carrier densities at the edge of the depletion layer are given by Shockley's law as

$$n = n_{p0} \exp(qV/k_B T), \tag{9}$$

$$p = p_{n0} \exp(qV/k_B T). \tag{10}$$

Solving the diffusion equations (7) and (8) under the boundary conditions (9) and (10), the current density is obtained as

$$J = J_s[\exp(qV/k_B T) - 1], \tag{11}$$

with the saturation current density

$$\mathbf{J}_s = qD_n n_{p0}/L_n + qD_p p_{n0}/L_p, \tag{12}$$

where $L_n = (D_n \tau_n)^{1/2}$ and $L_p = (D_p \tau_p)^{1/2}$ are the diffusion lengths for electrons and holes, respectively. Equation (11) with Eq. (12) is the celebrated Shockley equation, which is the ideal diode law.

2.2 Bipolar Transistor

A one-dimensional *npn* bipolar transistor is shown in Fig. 2. When a positive base–emitter voltage V_{BE} is applied, electrons are injected from the emitter to the base. When a positive collector–emitter voltage V_{CE} is applied, the injected electrons in the base reach the edge of the collector depletion layer and are driven to the collector by the strong electric field in the depletion layer. To derive an analytical model, the following assumptions are made:

1. uniform impurity profile in each region (emitter, base, and collector);
2. low-level injection;
3. constant mobility;
4. no generation–recombination in the depletion layer.

Then, the diffusion equation for the neutral base, emitter, and collector regions under the boundary conditions at the depletion-layer edges given by Shockley's law can be solved, and the emitter, base, and collector current densities can be derived as

$$J_E = q\left(\frac{D_n n_{B0}}{W_B} + \frac{D_p p_{E0}}{L_E}\right)\exp\left(\frac{qV_{BE}}{k_B T}\right), \tag{13}$$

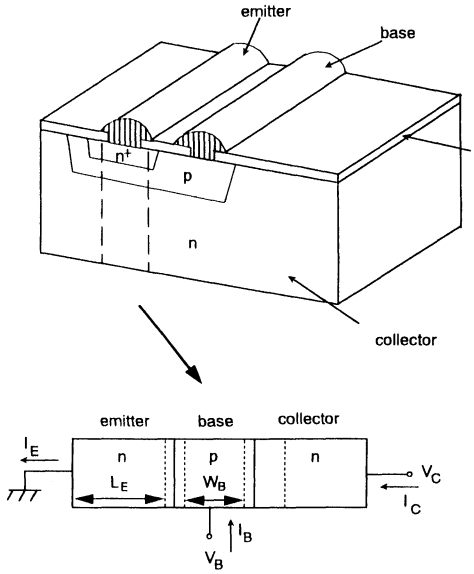

FIG. 2. Schematic structure (upper) and one-dimen-

$$J_B = q\frac{D_p p_{E0}}{L_E}\exp\left(\frac{qV_{BE}}{k_B T}\right), \tag{14}$$

$$J_C = q\frac{D_n n_{B0}}{W_B}\exp\left(\frac{qV_{BE}}{k_B T}\right), \tag{15}$$

where W_B is the neutral base width, L_E is the neutral emitter width, p_{E0} is the equilibrium minority-carrier (hole) density in the emitter (equal to n_i^2/N_E), and n_{B0} is the equilibrium minority-carrier (electron) density in the base (equal to n_i^2/N_B). A formula for the current gain is given as

$$h_{FE} = \frac{I_C}{I_B} = \frac{D_n L_E N_E}{D_p W_B N_B}. \tag{16}$$

It should be noted, however, that actual current gain is smaller than that given in Eq. (16) as a result of band-gap narrowing in the emitter, which is ignored in the above derivation.

2.3 MOSFET

An ideal metal-oxide–semiconductor structure is shown in Fig. 3, where the work-function difference, $q\phi_{ms}$, is equal to zero:

$$q\phi_{ms} = q\phi_m - (q\chi + \tfrac{1}{2}E_g + q\psi_B) = 0, \tag{17}$$

where $q\phi_m$ is the work function of the metal, $q\chi$ is the electron affinity, E_g is the band-gap energy, and $q\psi_B$ is the difference between the Fermi potential and the midgap. The energy-band diagrams for negative and positive voltages applied on the metal are shown in

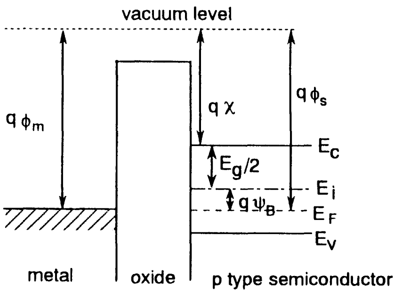

FIG. 3. Energy-band diagram of an ideal MOS diode for zero biasing voltage, where $q\phi_m$ is the metal work function, $q\phi_s$ the semiconductor work function, and χ the semiconductor electron affinity.

Fig. 4. For negative voltage, an accumulation layer of holes appears at the interface; for low positive voltage, a depletion layer extends from the interface; and for high positive voltage, the electron inversion layer appears at the interface. The MOSFET shown in Fig. 5 utilizes these MOS diode characteristics. When the gate voltage is zero or low, a depletion layer appears under the gate and no current flows between the source and the drain. When a high gate voltage is applied, an inversion layer called the channel appears under the gate, and current flows through this channel between the source and the drain for nonzero drain voltage. To derive an analytic model for MOSFET I-V characteristics, the following assumptions are made:

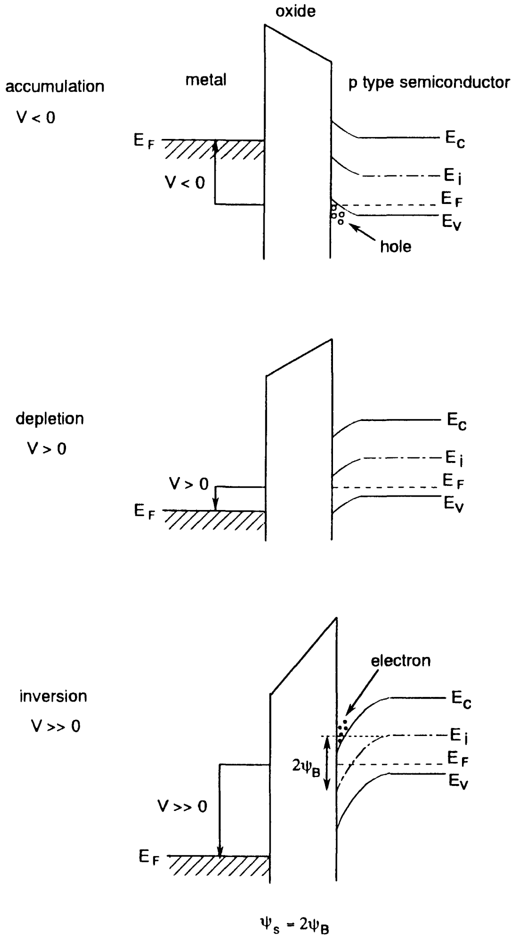

FIG. 4. Energy-band diagram of an ideal MOS diode for nonzero biasing voltage.

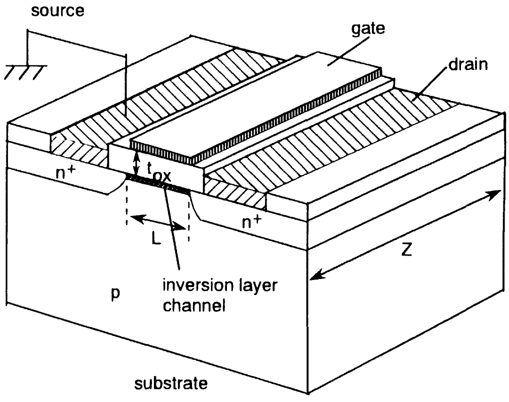

FIG. 5. Structure of a MOSFET.

1. The MOS diode is ideal.
2. Only drift current is considered (the diffusion current is ignored).
3. Electron mobility in the inversion layer is constant.
4. The impurity density of the substrate is constant.
5. Reverse saturation current is ignored.
6. The electric field component parallel to the channel direction is much smaller than the electric field component normal to the channel direction.

Assumption **6** is called the gradual-channel approximation.

On the basis of these assumptions the following analytic expression for the drain current can be derived:

$$I_D = \frac{W}{L}\mu_n C_{ox}\left\{\left(V_G - 2\psi_B - \frac{V_D}{2}\right)V_D - \frac{2}{3} \times \frac{(2\epsilon_s q N_A)^{1/2}}{C_{ox}}\left[(V_D + 2\psi_B)^{3/2} - (2\psi_B)^{3/2}\right]\right\}, \tag{18}$$

where L is the channel length, W is the channel width, C_{ox} is the gate-oxide capacitance ($= \epsilon_{ox}/t_{ox}$), and N_A is the substrate doping density. The drain current [Eq. (18)] of the gradual-channel model is plotted in Fig. 6. When the drain voltage V_D is low, the drain current is given as

$$I_D = (W/L)\mu_n C_{ox}(V_G - V_T)V_D, \tag{19}$$

where V_T is a threshold voltage given by

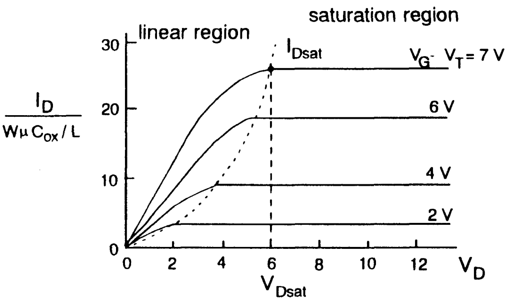

FIG. 6. Drain-current characteristics of a MOSFET derived by the gradual-channel model. The dashed curve indicates the locus of the saturation points (V_{Dsat}, I_{Dsat}).

$$V_T = 2\psi_B + [2\epsilon_s q N_A(2\psi_B)]^{1/2}/C_{ox} + V_{FB}. \tag{20}$$

When the gate voltage is higher or lower than the threshold voltage, the drain current flows or does not flow, respectively. Nonideality of the MOS diode is taken into consideration in Eq. (20) by the flat-band voltage V_{FB}, which is given as

$$V_{FB} = \phi_{ms} - Q_{ox}/C_{ox}, \tag{21}$$

where Q_{ox} is the oxide charge.

3. NUMERICAL MODELS FOR SEMICONDUCTOR DEVICES

The analytical models described in Sec. 2 are based on various assumptions. These assumptions, such as uniform impurity density, constant carrier mobility, or the gradual-channel approximation, inevitably restrict the models' applicability. In contrast, numerical models are general, since they derive device characteristics from solving numerically the basic semiconductor equations, e.g., the drift–diffusion equations in Sec. 1, without the severe assumptions mentioned above. Therefore, the numerical models can be used for quantitative analysis or design of realistic devices. Moreover, the hot-carrier phenomena and the nonlocal impact ionization, which crucially affect current semiconductor devices, and even nonstationary microscopic phenomena such as the ballistic effect, can be treated by the numerical approach based on the appropriate basic equations, e.g., the hydrodynamic equation and the Boltzmann equation.

However, numerical computation techniques to solve the basic semiconductor equations are needed. These techniques should be powerful, since solving partial differential equations or transport equations tends to be computer-expensive. To solve partial differential equations numerically, the finite-difference method (FDM) using orthogonal grid lines, and the control-volume method, using triangular elements in 2D and tetrahedron elements in 3D, are generally used (see SCIENTIFIC COMPUTATION BY NUMERICAL METHODS). A resulting linear system that is expressed by a large sparse matrix can be efficiently solved using recently developed methods such as CGS (conjugate gradient squared) and Bi-CGSTAB (van der Vorst, 1992). At present, device simulation using 10^4 grid points can be executed quite easily in routine design jobs.

At present, the numerical models are divided into several classes: drift–diffusion model, hydrodynamic model, Monte Carlo model, etc. In addition, different approaches to solve the Boltzmann transport equation and various quantum-mechanical treatments have been reported.

The drift–diffusion model is still a standard model from the point of view of the application. For practical use, modeling of physical parameters such as carrier mobility is very important. In the numerical approach, physical-parameter models can be incorporated even if they are rather complicated. Thus, most practical DD model simulators are equipped with sophisticated physical-parameter models.

4. PHYSICAL PARAMETERS

4.1 Carrier Mobility

Carrier mobilities are related to the scattering mechanisms that deflect the carrier motion. Electrons and holes in a semiconductor are scattered by thermal lattice vibrations (phonons), ionized impurities, neutral impurities, and the surface. As a result, carrier mobilities depend on the temperature through phonon scattering, on the impurity density through ionized impurity scattering, and on the electric field through optical phonon scattering.

In the DD model, the carrier mobility is commonly expressed in empirical form, as a function of the temperature, impurity density, and electric field. The temperature dependence of the carrier mobility in Si can be expressed as

$$\mu_L = \mu_0(T/300\ \mathrm{K})^{-\alpha}, \tag{22}$$

where α is about 2.5 and 2.3 for electron and hole, respectively, and μ_0 is the low-field mobility for low impurity density (1450 and 500 $\mathrm{cm^2/V \cdot s}$ for electrons and holes, respectively). The mobility μ_L is regarded as the one for low impurity density and for low electric field.

An empirical model including impurity-density dependence (for Si) is given by Caughey and Thomas (1967) as

$$\mu_{LI} = \mu_{\min} + \frac{\mu_L - \mu_{\min}}{1 + [(N_D + N_A)/N_0]^{\alpha}}, \tag{23}$$

where $\mu_{\min}$, N_0, and α are 80 $\mathrm{cm^2/V \cdot s}$, $1 \times 10^{17}\ \mathrm{cm^{-3}}$, and 0.7, respectively.

For electric fields higher than 5×10^4 V/cm in Si, the carrier drift velocity is nearly constant. This velocity is called the saturation velocity v_{sat}, which is about 1×10^7 cm/s. An empirical expression including this high-field effect is given by Scharfetter and Gumbel (1969) and Thornber (1980) as

$$\mu_{LIE} = \mu_{LI}\left\{1 + \left(\frac{\mu_{LI}E}{v_c}\right)^2 \left(\frac{\mu_{LI}E}{v_c} + F\right)^{-1} + \left(\frac{\mu_{LI}E}{v_{\mathrm{sat}}}\right)^2\right\}^{-1/2}, \tag{24}$$

where E is the electric field and v_c and F_0 are parameters.

Carriers in an inversion layer of a MOSFET experience surface scattering besides phonon and impurity scattering. Surface scattering is related to surface roughness, interface charge, and quantum-mechanical effects of a two-dimensional gas. The surface mobility has been measured as an effective mobility determined from the drain conductance for the MOSFET in the low-drain-voltage region (linear region). It has been shown that the surface mobilities follow a universal curve, independent of the substrate impurity concentration, when they are plotted as a function of the effective normal field. Experimental results by Takagi *et al.* (1988) are shown in Fig. 7, which shows a remarkable universality in the wide range from 10^{15} to $10^{18}\ \mathrm{cm^{-3}}$. For MOSFET simulation, implementation of the surface-mobility model is important.

4.2 Generation–Recombination Rate

When a system of electrons and holes deviates from thermal equilibrium, generation–recombination processes take place to restore thermal equilibrium. For indirect-band-gap semiconductors such as Si, the recombination process takes place through indirect transition of electrons and holes between the conduction band and the valence band via trap levels in the band gap.

The single-level recombination model (Shockley *et al.*, 1952) can be derived from four processes: electron capture, electron emission, hole capture, and hole emission. The recombination rate R_{SRH} is given as

$$R_{SRH} = (pn - n_i^2)\left\{\tau_p\left[n + n_i \exp\left(\frac{E_t - E_i}{k_B T}\right)\right] + \tau_n\left[p + n_i \exp\left(-\frac{E_t - E_i}{k_B T}\right)\right]\right\}^{-1}, \tag{25}$$

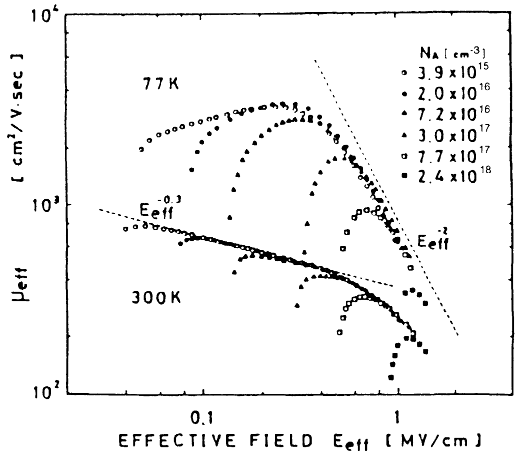

FIG. 7. Effective electron mobility μ_{eff} in *n*-channel MOSFETs vs normal field E_{eff} with substrate acceptor concentration as a parameter (Takagi *et al.*, 1988).

where τ_n and τ_p are the electron and hole lifetimes, respectively, and E_t is the trap energy level. The carrier lifetimes are determined by the carrier capture cross sections and the trap density. When electrons are injected into p-type Si as minority carriers, the recombination rate R_{SRH} is given approximately as

$$R_{SRH} = (n - n_{p0})/\tau_n, \tag{26}$$

where n_{p0} is the equilibrium minority-electron density. The carrier lifetime is found empirically to depend on the doping density as

$$\tau_{n,p} = \frac{\tau_{n,p0}}{1 + (N_D + N_A)/N_r} \tag{27}$$

or

$$\tau_{n,p} = \tau_{n,p0}[N_r/(N_D + N_A)]^{\alpha}, \tag{28}$$

with α in the range 0.3–0.6.

4.3 Auger-Recombination Rate

Auger recombination involves three-particle transitions. There are three processes:

1. An electron in the conduction band recombines with a hole in the valence band and excites another electron in the conduction band with the generated energy;
2. the recombination energy excites an electron in the light valence band to the heavy valence band; and
3. the recombination energy excites an electron in the spin-orbit–split band to the valence band.

The Auger-recombination rate is given as

$$R_{\mathrm{AUG}} = (C_n n + C_p p)(np - n_i^2), \tag{29}$$

where $C_n = 2.8 \times 10^{-31}$ cm^6/s and $C_p = 9.9 \times 10^{-32}$ cm^6/s, at 300 K. Since Auger recombination is dominant for high impurity density, such as the emitter layer in a bipolar transistor, it affects the current gain of the bipolar transistor.

4.4 Impact-Ionization Rate

Impact ionization is a carrier-generation process that takes place in a region of high electric field. In the process, carriers generate electron–hole pairs in their transit, with the energy acquired from the electric field. Since this process is related to the band structure and various scattering mechanisms, accurate analytical treatment is not easy. Empirical models are generally used.

The generation rate of impact ionization, G_{ii}, defined as the number of electron–hole pairs per unit volume and per unit time, can be expressed using the ionization rates α_n and α_p as

$$G_{ii} = \alpha_n |J_n|/q + \alpha_p |J_p|/q. \tag{30}$$

The electron ionization rate α_n is defined as the number of generated electron–hole pairs per unit length of travel and per electron. An empirical model for the ionization rate α_n is given as

$$\alpha_n = A_n \exp(-b_n/E^m), \tag{31}$$

where E is the electric field and A_n, b_n, and m are parameters. The exponent m is equal to 1 or 2. Empirically $m = 1$ is used for Si and Ge, and $m = 2$ is used for GaAs. Different values for A_n and b_n have been reported in the literature. For example, A_n is 6.2×10^5 cm^{-1} and b_n is 1.08×10^5 V/cm for 2.4×10^5 V/cm $< E < 5.3 \times 10^5$ V/cm. For holes, the same expression as Eq. (31) holds.

Solving the Boltzmann transport equation numerically, Baraff (1962) obtained a universal plot for the ionization rate in three-parameter form:

$$\alpha = \frac{1}{\lambda} f\left(\frac{E_r}{E_i}, \frac{E_i}{q\lambda E}\right), \tag{32}$$

where λ is the optical phonon mean free path, E_r is the average energy loss per phonon collision, and E_i is the ionization threshold energy.

For the latest submicron devices, the electric field profile is so steep that nonstationary transport effects are prominent and the impact ionization is not directly determined by the local field. For these cases, the ionization rate is considered as a function of carrier energy or to be dependent on the trajectory of the carrier movement. The impact-ionization rate as a function of the electron energy has recently been calculated on the basis of first principles using realistic band structures (Sano and Yoshii, 1992).

4.5 Band-Gap Narrowing Effect

Highly doped diffusion layers, such as the emitter region in a bipolar transistor, are often used in semiconductor devices. Impurity concentrations in these regions are higher than 10^{20} cm^{-3}. In such cases, the impurity bands are formed below the conduction band and above the valence band and merge with the respective bands. Therefore, the band gap is effectively narrower than for the intrinsic case (very pure with a negligibly small amount of impurities). Band-gap narrowing ΔE_g can be measured and an empirical expression can be given (Slotboom and de Graaff, 1976) as

$$\Delta E_g = qV_1\left\{\ln\left(\frac{N_D + N_A}{N_0}\right) + \left(\left[\ln\left(\frac{N_D + N_A}{N_0}\right)\right]^2 + C\right)^{1/2}\right\} \qquad (33)$$

with

$$V_1 = 9\,\text{mV},\quad N_0 = 10^{17}\,\text{cm}^{-3},\quad C = 0.5. \qquad (34)$$

The band-gap narrowing effect brings about an increase in the intrinsic carrier density,

$$n_{ie} = n_i \exp(\Delta E_g/2k_BT), \qquad (35)$$

where n_{ie} is called the effective intrinsic carrier density. This effect causes a decrease in the current gain of bipolar transistors from the ideal value, and modeling this effect is indispensable for simulation.

5. AN EXAMPLE USING THE DRIFT–DIFFUSION MODEL

A typical two-dimensional simulation of an n-channel MOSFET is shown. A schematic cross section of a LDD (lightly doped drain) NMOSFET is shown in Fig. 8. The impurity profile was calculated using a process simulator and its bird's-eye-view plot is shown in Fig. 9. The number of grid points is 58 × 61 = 3538. The channel region and the junction region are divided by a fine mesh. The device parameters are as follows:

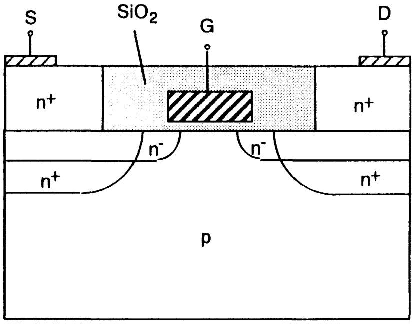

FIG. 8. Cross section of a LDD (lightly doped drain) MOSFET. The source and the drain contacts are replaced by n^+ diffusion layers because of a restriction of the simulator used.

gate-oxide thickness 105 Å
gate length 0.6 μm
substrate impurity density 2×10^{16} cm^{-3}
channel impurity density 3×10^{17} cm^{-3}
junction depth 0.2 μm.

Calculated drain-current–drain-voltage characteristics are shown with gate voltage as a parameter in Fig. 10. Increase in the drain current due to partial avalanche breakdown is observed for drain voltage larger than 5 V. Potential and electron-density distributions for a typical biasing point of V_G = 3 V and V_D = 3 V are shown in Figs. 11 and 12, respectively.

6. ENERGY-TRANSPORT MODEL

In miniaturized devices, electric fields are high and vary sharply with respect to position. The electrons in these devices acquire high energy from the electric fields, and hot-electron transport plays an important role. In this case the electron energy distribution deviates significantly from the equilibrium distribution, and the drift–diffusion model loses validity. The Boltzmann transport theory presents a way to cope with this problem. Using the energy-transport model, which is also called the hydrodynamic model (Forghieri *et al.*, 1988), is one of the practical approaches to solve the Boltzmann transport equation.

In the energy-transport model, the basic equations for electrons are given as

$$\frac{\partial n}{\partial t} = -\nabla \cdot n\mathbf{v} + G, \qquad (36)$$

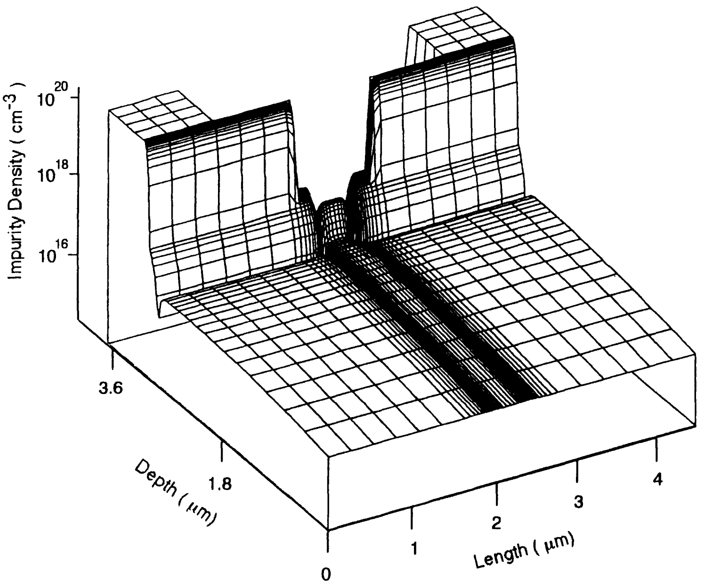

FIG. 9. Impurity-density distribution of the LDD MOSFET.

$$\frac{\partial(m^*n\mathbf{v})}{\partial t} = qn\nabla\psi - \nabla nk_BT_n - \frac{m^*n\mathbf{v}}{\tau_m}, \qquad (37)$$

$$\frac{\partial(nw)}{\partial t} = -\nabla\cdot n\mathbf{v}w + qn\mathbf{v}\cdot\nabla\psi - \nabla\cdot n\mathbf{v}k_BT_n + \nabla\cdot\kappa\nabla T_n - \frac{n(w-w_0)}{\tau_w} + Gw, \qquad (38)$$

where G is the net generation rate, which is composed of various generation–recombination mechanisms mentioned in the previous section; m^* is the electron effective mass; T_n is the effective electron temperature; τ_m and τ_w are the momentum and energy relaxation times, respectively; w is the average electron energy, which is related to the mean velocity v and the effective electron temperature T_n as

$$w = \tfrac{1}{2}m^*v^2 + \tfrac{3}{2}k_BT_n; \qquad (39)$$

and κ is thermal conductivity. Equations (36), (37), and (38) are derived from the moments of the Boltzmann transport equation, and represent mass conservation, momentum conservation, and energy conservation,

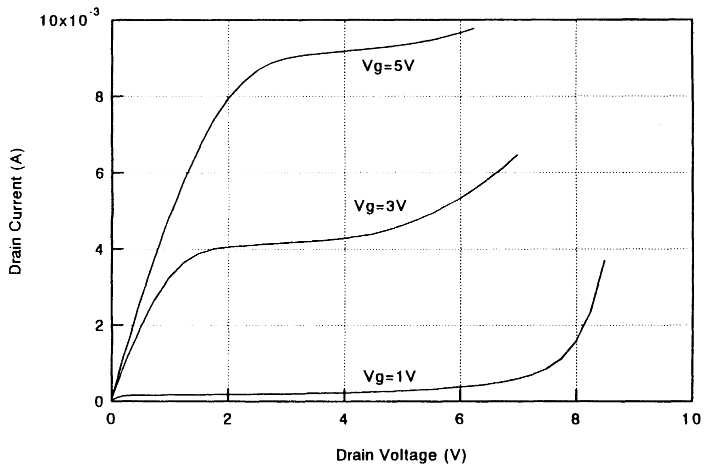

FIG. 10. Drain-current–drain-voltage characteristics of the LDD MOSFET.

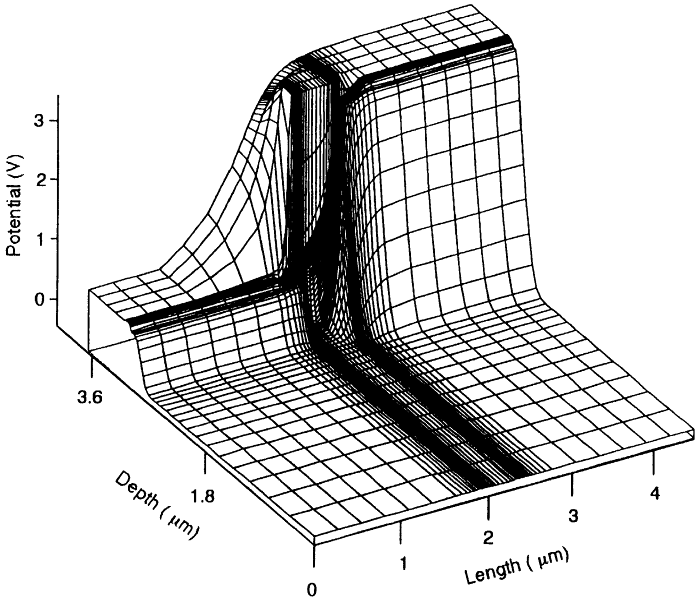

FIG. 11. Potential distribution of the LDD MOSFET for V_D = 3 V and V_G = 3 V.

respectively. The sequence of the moments is truncated at the second moment [Eq. (38)] by assuming that heat flow can be described by $-\kappa \nabla T_n$. It is assumed that the effects of the various scattering mechanisms on momentum and energy are expressed by the relaxation times and can be given as functions of the energy. This approach is also called the relaxation-time approximation. The momentum relaxation time τ_m is related to the mobility by $\mu = (q/m^*)\tau_m$. It is hard to measure the energy relaxation time, and it is often treated as a fitting parameter, though it should be derived from scattering theory.

Hot-carrier phenomena are reflected by high average energy of electrons, resulting in the electron temperature being higher than the lattice temperature, causing a change in

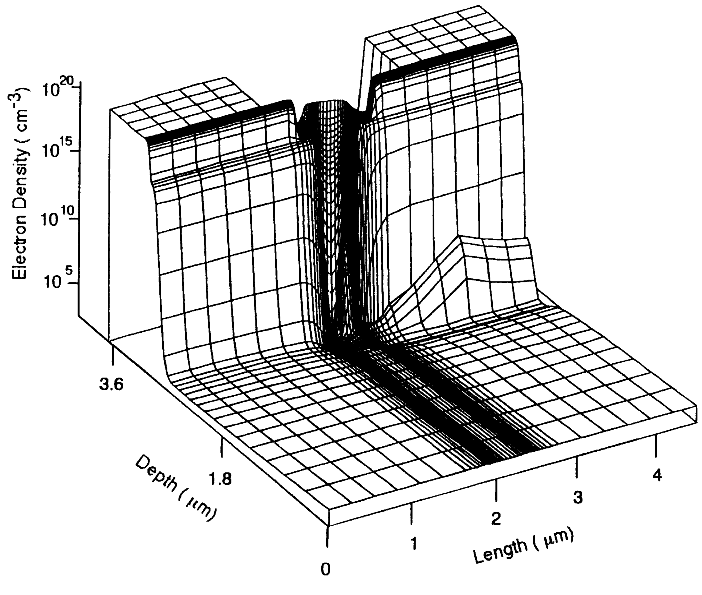

FIG. 12. *Electron-density distribution of the LDD MOSFET for* V_D = 3 V and V_G = 3 V.

mobility and impact-ionization rate values. The diffusion constant given as $D = (k_BT_n/q)\mu$ indicates a larger diffusion in the "hot" region than in the "cold." This model predicts channel spreading near the drain junction of a miniaturized MOSFET under high drain voltages. Substrate currents in MOSFETs are successfully modeled by the energy-transport model including the energy-dependent impact-ionization rates as shown in Fig. 13 (Katayama and Toyabe, 1989).

7. MONTE CARLO MODEL

7.1 Motivation

As devices shrink into deep submicron and sub–0.1-μm regions, the DD and the hydrodynamic models are no longer valid, since nonequilibrium transport of highly energetic electrons, which affects strongly the electrical characteristics of the devices, must be treated rigorously. For this problem, it is necessary to solve the Boltzmann transport equation (BTE). It is impossible in practice to solve directly the BTE coupled with the Poisson equation self-consistently even by the fastest supercomputer, since the equation is a complicated integro-differential equation with respect to the electron distribution function in 2D or 3D real space and 3D wave-vector space. The Monte Carlo method (*q.v.*), however, is a viable means to solve the BTE.

The Monte Carlo (MC) model can accurately handle various scattering mechanisms and the full band structures of semiconductors. The impact of the full band structures on device physics has been thoroughly investigated by Fischetti and Laux (1988). Since high-field effects predominantly observed in small devices are governed by electrons in the high-energy portion of the bands, not by the ones near the band minima, only the MC model based on the full band structure can give reliable predictions for the electrical characteristics of small devices. Thus, a comparative study has shown that the transconductance of an *n*-GaAS MOSFET is almost the same as that of an *n*-Si MOSFET for channel lengths between 0.05 and 0.25 μm, in contrast to the superiority of the low-field characteristics of *n*-GaAs devices resulting from the high electron mobility at low fields (Fischetti and Laux, 1991).

7.2 Electric Field Calculation

The computation of the Monte Carlo model is composed of the electric field computation, the free-flight computation, and the

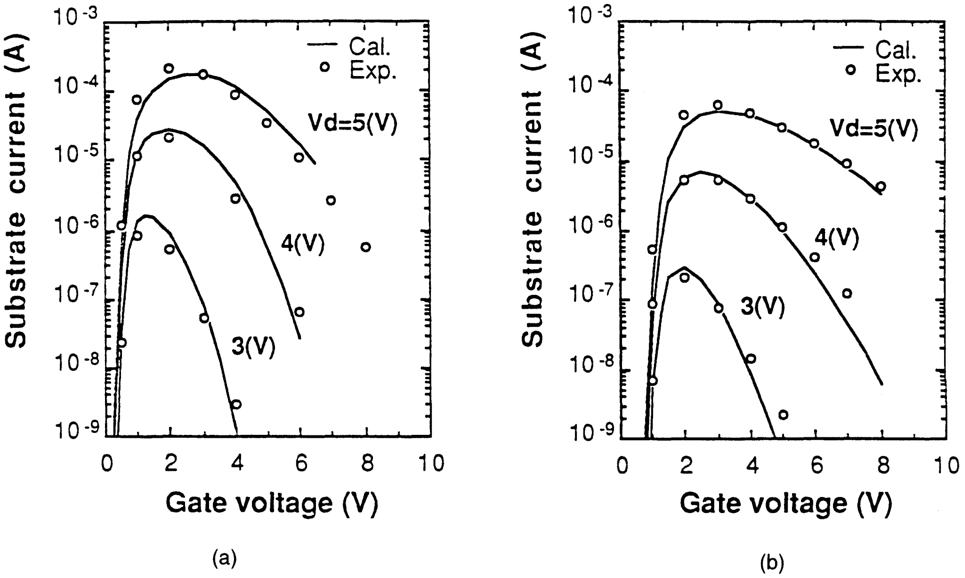

FIG. 13. Substrate-current characteristics for *n*-channel MOSFETs, where $L_{\text{eff}} = 0.99$ μm, $W = 100$ μm. (a) Oxide thickness $t_{ox} = 21$ nm; (b) $t_{ox} = 41$ nm. (Katayama and Toyabe, 1989.)

scattering computation as shown in Fig. 14. The electric field can be obtained by solving the Poisson equation, which is usually solved using the finite-difference method (FDM), the finite-element method (FEM), or the fast-Fourier-transform (FFT) method. In these methods, real space is divided by grid points and the Poisson equation is discretized on the grid points.

The particles distributed in space have to be assigned to grid points to form the space-charge–density term of the Poisson equation. A simple way is using the nearest-grid-point method in which each particle is assigned to the nearest grid point. This method produces a discontinuous charge distribution when the number of particles is not large. Another method, called the cloud-in-cell method, assigns particles to neighboring grid points with appropriate weights. This method produces a smoother charge distribution than with the nearest-grid-point method.

7.3 Free-Flight Calculation

Particles travel under the influence of the electric field between collisions. When analytic approximations of the bands are adopted, each free-flight time t_f is determined by a random number r_f that is uniform between 0 and 1 as

$$t_f = -\log(r_f)/\Gamma, \tag{40}$$

where the constant Γ is the total scattering probability including the so-called "self-scattering."

In the MC model adopting the full bands (Fischetti and Laux, 1988), however, the high efficiency of this technique vanishes since the equations of motion,

$$\frac{d\mathbf{r}}{dt} = \frac{1}{\hbar}\nabla_k E_\nu(\mathbf{k}), \tag{41}$$

$$\frac{d\mathbf{k}}{dt} = \frac{q}{\hbar}\nabla_r \psi(\mathbf{r}) = -\frac{q\mathbf{F}(\mathbf{r})}{\hbar}, \tag{42}$$

cannot be integrated analytically, where $E_\nu(\mathbf{k})$ is the νth energy band and $\mathbf{F}$ is the electric field at the electron position $\mathbf{r}$. In this case, a prefixed time step Δt of the order of 0.1 fs is preferable. The band structure of technological interest can be obtained from empirical-pseudopotential calculation. The entire Brillouin zone is divided into irreducible wedges, and each wedge is discretized by mesh points. The energy, group velocity, and effective mass for each mesh point are stored in a lookup table (Fischetti and Laux, 1988).

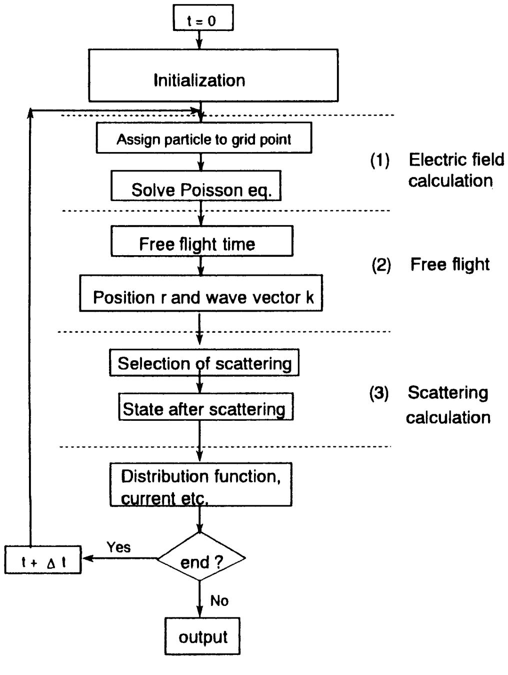
t = 0
Initialization
Assign particle to grid point
Solve Poisson eq.
(1) Electric field calculation
Free flight time
Position r and wave vector k
(2) Free flight
Selection of scattering
State after scattering
(3) Scattering calculation
Distribution function, current etc.
Yes
t + Δ t
end ?
No
output

FIG. 14. Flow of Monte Carlo method.

7.4 Scattering Mechanisms

The scattering probability S from a state $|i\rangle$ to a state $|f\rangle$ per unit time is given using the Fermi "golden rule" as

$$S = \frac{2\pi}{\hbar}|\langle i|H'|f\rangle|^2\delta(E_f - E_i), \tag{43}$$

where H' is a perturbation Hamiltonian. In the case of phonon scattering, the scattering probability from a point $\mathbf{k}$ in the νth band to a point $\mathbf{k}'$ in the ν'th band $S(\nu,\mathbf{k},\nu',\mathbf{k}')$ is given as

$$\begin{aligned} S(\nu,\mathbf{k},\nu',\mathbf{k}') &= \frac{2\pi}{\hbar}M(|\mathbf{k} - \mathbf{k}'|)^2|I(\nu,\nu';\mathbf{k},\mathbf{k}')|^2 \\ &\times (n_\mathbf{q} + \tfrac{1}{2} \pm \tfrac{1}{2}) \\ &\times \delta(E_{\nu',\mathbf{k}'} - E_{\nu,\mathbf{k}} \pm \hbar\omega_\mathbf{q}), \end{aligned} \tag{44}$$

where M depends on various parameters of the crystal such as sound velocity, density, deformation potential, etc., I is the overlap integral, $n_{\mathbf{q}}$ is the number of phonons with the wave vector $\mathbf{q}$, and signs $+$ and $-$ correspond to the emission and absorption of a phonon, respectively.

Scattering mechanisms are electron–phonon scattering (acoustic phonon and optical phonon); particle–particle scattering (electron–electron and electron–hole); and other scattering (impurity, alloy, and impact ionization).

In the Monte Carlo method, the successive scattering for each electron is modeled through the selection of a final state after a scattering using random numbers. As can be imagined from the complexity of the band-structure data and the scattering probability expressions, the Monte Carlo model needs a huge amount of CPU time and computer memory. Nevertheless, a large amount of exploring work has been reported by the IBM group, e.g., comparative studies of Si, GaAs, InP, InGaAs, and Ge FET performance (Fischetti and Laux, 1991).

8. OTHER SOLUTION METHODS TO THE BOLTZMANN TRANSPORT EQUATION

8.1 Scattering-Matrix Approach

The scattering-matrix approach (Das and Lundstrom, 1990) begins by dividing a device into a number of small elements (thin slabs) and defining scattering matrices, which relate the fluxes incident upon each element to the emerging fluxes. For a slab shown in Fig. 15, the emerging carrier fluxes are related to those incident on it by

$$\begin{bmatrix} b^+ \\ a^- \end{bmatrix} = \begin{bmatrix} t & r' \\ r & t' \end{bmatrix} \begin{bmatrix} a^+ \\ b^- \end{bmatrix} = S \begin{bmatrix} a^+ \\ b^- \end{bmatrix}, \tag{45}$$

where S is the scattering matrix for the slab. By cascading the individual scattering matrices for the slabs composing a semiconductor device, the composite scattering matrix for the entire device is evaluated.

For actual simulations, the incident and emerging fluxes are resolved into M modes that represent different three-dimensional $\mathbf{k}$ vectors. For this case, the elements of the scattering matrix, t, r, t', and r', become $M \times M$ submatrices. By assuming local thermodynamic equilibrium at the contacts, the incident fluxes from contacts are evaluated using the Fermi–Dirac distributions. Starting from these fluxes, all the other fluxes can be obtained from the computation using scattering matrices. The carrier current, density, and distribution function in momentum space can be derived from the fluxes. This approach is free from the statistical noise that appears in the MC model. Almost all the computation burden exists in the scattering-matrix computation, which is usually carried out by Monte Carlo simulation in the slabs. Once the scattering matrices are calculated, the subsequent device simulation is carried out very efficiently. The method is capable of treating quantum interference effects as well. Demonstration of this approach was given for high-field transport in bulk Si and nonstationary transport in 1D Si devices with good agreement with Monte Carlo simulations (Das and Lundstrom, 1990).

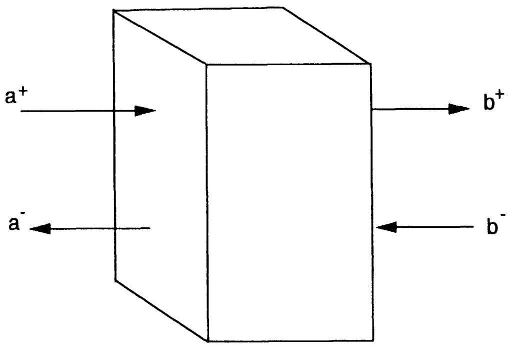

FIG. 15. Characterization of electron transport across a slab of semiconductor in terms of incident and emerging fluxes (Das and Lundstrom, 1991).

8.2 Cellular-Automaton Method

In the cellular-automaton method (Kometer *et al.*, 1992), lattices in real ($\mathbf{x}$) and momentum ($\mathbf{k}$) space are defined with the cells. The $\mathbf{x}$ cells and $\mathbf{k}$ cells are populated by fictitious particles—the lattice-gas particles—which obey nondeterministic cellular-automaton (CA) rules. It is shown that the Boltzmann equation for semiclassical transport can be transformed into a Boolean master equation, which represents a cellular automaton with nearest-neighbor interaction in position space. The translation term in **x** and **k** space in the BTE is transformed to an effective collision term resulting in the CA rules. Boolean cell occupancies evolve in time obeying the master equation. The average over an ensemble of CA states (the average over a large number of cells) yields macroscopic quantities such as particle densities. The CA simulations have been verified, with agreement with Monte Carlo results (Kometer *et al.*, 1992). This approach is well suited for carrier systems with pronounced spatial inhomogeneities, large density variations, and two-particle scattering interactions.

9. CONCLUSIONS

Semiconductor-device modeling already has a history of 40 years. From early analytical modeling it has evolved into sophisticated numerical modeling. The classic drift–diffusion model is now a commonly used tool in the design world of the semiconductor industry. For modern devices of small dimensions, advanced approaches of the hydrodynamic model, at least, and the Monte Carlo model are indispensable for understanding the operating mechanism. The full-band Monte Carlo model has been realized to attain rigorous understanding of small devices of 0.1 μm or even smaller. Furthermore, to go beyond the MC model, approaches using the scattering-matrix and cellular-automaton methods have been proposed. The quantum-mechanical treatment will be more important for smaller devices. The continuing increase in the cost of semiconductor development makes numerical modeling of semiconductor devices, with the help of the decrease in the cost of computation, all the more important.

GLOSSARY

Alpha-Particle–Induced Soft Error: Volatile operation failure in memory or logic circuits in LSI caused by an alpha-particle hit.

CG Method: A numerical solution method for linear systems expressed by symmetric positive-definite matrices. CGS and Bi-CGSTAB are extended versions applicable to nonsymmetric matrices.

Drift–Diffusion Model: A traditional model used in device modeling where device operation is described by the potential and the carrier-density distributions governed by the Poisson and current-continuity equations, respectively.

Electron Temperature: The electron temperature is introduced to express a rise of the electron mean energy above the thermal value $3k_BT/2$ in the nonlinear transport caused by a high electric field, assuming the displaced Maxwellian distribution.

Energy-Transport Model: One of the advanced device models, also called hydrodynamic model, where the carrier temperature is introduced to describe the state of the system besides the potential and the carrier-density distributions. The equations for the model are derived by taking the moments of the Boltzmann transport equation.

FET (Field Effect Transistor): A kind of transistor in which the current flowing in the semiconductor is controlled by the electric field produced by the biasing voltage on the gate electrode. Representative types of FETs are junction FET (JFET), metal-oxide–semiconductor FET (MOSFET), and metal-semiconductor FET (MESFET).

Gunn Diode: A microwave solid-state oscillator utilizing the effect of electron transfer in the conduction band from low-energy, high-mobility states to high-energy, low-mobility states in GaAs in a high electric field.

Hot-Carrier Effects: Adverse effects caused by hot electrons or hot holes such as nonlinear conduction, parasitic current of substrate current, and device-performance degradation in MOSFETs.

IMPATT (Impact Avalanche Transit Time) Diode: A microwave solid-state oscillator utilizing the avalanche process in a reverse-biased *p-n* junction for generation of energetic carriers.

Planar Process Technology: Fabrication technology of semiconductor devices with electrodes and active region on the same semiconductor surface.

Relaxation Time: Time constants appearing in the approximation of the collision term of the Boltzmann transport equation. The momentum and energy relaxation times are used in the energy-transport model.

Works Cited

Baraff, G. A. (1962), *Phys. Rev.* **128,** 2507–2517.

Caughey, D. M., Thomas, R. E. (1967), *Proc. IEEE* **52,** 2192–2193.

Das, A., Lundstrom, M. S. (1990), *Solid-State Electron.* **33,** 1299–1307.

Dennard, R. H., Gaensslen, F. H., Yu, H. N., Rideout, V. L., Bassous, E., LeBlanc, A. (1974), *IEEE J. Solid-State Circuits* **SC-9,** 256–268.

Fischetti, M. V., Laux, S. E. (1988), *Phys. Rev. B* **38,** 9721–9745.

Fischetti, M. V., Laux, S. E. (1991), *IEEE Trans. Electron Dev.* **ED-38,** 650–660.

Forghieri, A., Guerrieri, R., Ciampolini, P., Gnudi, A., Rudan, M., Baccarani, G. (1988), *IEEE Trans. Comput. Aided Design* **7,** 231–242.

Gummel, H. K. (1964), *IEEE Trans. Electron Dev.* **ED-11,** 455–465.

Ihantola, H. K. J., Moll, J. L. (1964), *Solid-State Electron.* **7,** 423–430.

Katayama, K., Toyabe, T. (1989), in: *International Electron Devices Meeting Technical Digest,* New York: IEEE Publishing Services, pp. 135–138.

Kometer, K., Zandler, G., Vogl, P. (1992), *Phys. Rev. B* **46,** 1382–1394.

May, T. C., Woods, M. H. (1979), *IEEE Trans. Electron Dev.* **ED-26,** 2–9.

Sano, N., Yoshii, A. (1992), *Phys. Rev. B* **45,** 4171–4180.

Scharfetter, D. L., Gummel, H. K. (1969), *Electron Devices* **ED-16,** 64–77.

Shockley, W., Read, W. T. (1952), *Phys. Rev.* **87,** 835–842.

Slotboom, J. W., de Graaff, H. C. (1976), *Solid-State Electron.* **19,** 857–862.

Takagi, S., Iwase, M., Toriumi, A. (1988), in: *International Electron Devices Meeting Tech. Digest,* New York: IEEE Publishing Services, pp. 398–401.

Thornber, K. K. (1980), *J. Appl. Phys.* **51,** 2127–2136.

van der Vorst, H. A. (1992), *SIAM J. Sci. Statist. Comput.* **13,** 631–644.

Further Reading

Jacoboni, C., Lugli, L. (1989), *The Monte Carlo Method for Semiconductor Device Simulation,* New York: Springer-Verlag.

Selberherr, S. (1984), *Analysis and Simulation of Semiconductor Devices,* New York: Springer-Verlag.

Snowden, C. M. (1986), *Introduction to Semiconductor Device Modelling,* Singapore: World Scientific.

SEMICONDUCTOR DETECTORS

See DETECTORS, SEMICONDUCTOR

SEMICONDUCTOR LASERS

See LASERS, SEMICONDUCTOR

SEMICONDUCTORS TO SUPERCONDUCTORS: ORGANIC LOWER-DIMENSIONAL SYSTEMS

ROBERT MELVILLE METZGER, *Department of Chemistry, University of Alabama, Tuscaloosa, Alabama, U.S.A.*

Introduction 215
1. One-Dimensional Band Theory of Solids 216
2. One-Dimensional Instabilities 217
3. Conductivity and Superconductivity 218
4. Mulliken Charge-Transfer Complexes in Solution 219
5. Charge-Transfer Crystals 220
6. Optics of Polyenes 221
7. Proposed Excitonic Superconductivity 221
8. Inorganic Superconductors 221
9. TTF-TCNQ and Other Quasi One-Dimensional Metals 222
10. $(TMTSF)_2X$ or Bechgaard Salts: the First Organic Superconductors 224
11. $(BEDT\text{-}TTF)_2X$ and Other Organic Superconductors 224
12. Conducting Polymers 229
13. Conducting Langmuir–Blodgett Films 230
14. Design Rules for Crystalline Organic Metals and Superconductors 231
14.1 Molecular Requirements 231
14.2 Crystallographic Requirements 231
14.3 Band Requirements 232
14.4 Solid-State Electronic Requirements 233
14.5 Electron–Phonon Coupling Requirement 233
15. Design Rules for Polymeric Organic Conductors 233
15.1 Small Band Gaps 234
15.2 Fiber Alignment 234
15.3 Polymer Solubilization 234
15.4 Stereoregular Polymerization 234
15.5 Facilitate Interchain Hopping 234
16. Applications 234
17. Conclusion 235
Acknowledgment 235
Glossary 235
Works Cited 236
Further Reading 238

INTRODUCTION

Physicists tend to think of organic molecules as large, bland, boring, usually colorless assemblages of covalently bonded atoms that are electrically neutral overall, needed for that occasional headache (acetylsalicylic acid), for that pesky mosquito (naphthalene), to dye blue jeans (indigo), to make "plastics," i.e., structural polymers, or to enable the complicated chemical processes of life.

For decades it was unusual to be excited about the solid-state properties of organic molecules and their ions. In part this is because organic solids usually have only covalent chemical bonds within each molecule and van der Waals or London interactions between molecules; partial or complete formal charges on constituent cations and anions, and therefore Coulomb interactions between organic molecules in crystals, are the exception, rather than the rule. However, metallic conduction in organic systems had been dreamed of as early as 1911. Until the mid-1960s the number of serious researchers of the organic solid state remained small. The names of Hideo Akamatsu, Noel S. Bayliss, Melvin Calvin, Daniel D. Eley, Helmuth

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Kainer, Aleksandr I. Kitaigorodskii, Jan Kommandeur, Yoshio Matsunaga, Harden M. McConnell, Albert Szentgyörgyi, A. N. Terenin, and A. T. Vartanyan come to mind, but the list should include their many students and co-workers, many of whom are still active today.

The discovery of superconductivity in potassium-intercalated graphite by Hannay *et al.* (1965), the development of quasimetallic organic donor–acceptor crystals (charge-transfer, or "two-chain" systems) in the 1960s to early 1970s, with metallic conductivity along the direction of stacking, the discovery of highly conducting doped polyacetylene by Shirakawa *et al.* (1977), the discovery of superconductivity in the polymer polythiazyl, $(SN)_x$, by Greene *et al.* (1975), in organic "one-chain" crystals by Jérome *et al.* (1980), and in alkali-metal salts of C_{60} (buckminsterfullerides) by Hebard *et al.* (1991), and the development of light-emitting diodes based on conducting polymers by Burroughes *et al.* (1990) have all helped to educate the solid-state physics community to the possibilities of organic systems, whose molecular orbitals can be finely "tuned" by chemical synthesis to enhance desired physical properties.

This article reviews the progress made in organic crystalline and polymeric conductors and superconductors, by first introducing relevant issues from solid-state theory and molecular-orbital theory, then summarizing the history of the advances in crystalline conductors, polymers, and superconductors. After that, summaries of molecules, polymers, crystals, and superconductors are presented, and design rules for new systems are discussed.

We start our discussion with concepts from the band theory of solids, discuss what can break the symmetry of one-dimensional systems, introduce electrical conductivity and superconductivity, present the Mulliken charge-transfer theory for solution complexes and its extension to solids, then discuss briefly the simple π-electron theory for long polyenes.

1. ONE-DIMENSIONAL BAND THEORY OF SOLIDS

Consider a one-dimensional (usually almost infinite) set of N atoms, molecules, or point masses, all equally spaced at interparticle distances d along the real-space coordinate x, with Born–von Kármán periodic boundary conditions for the potential energy:

$$V(x) = V(x + Nd). \tag{1}$$

By the theorems of Floquet (in one dimension; 1883) and Bloch (in three dimensions; 1928) the periodicity of the lattice requires that the eigenfunction of the appropriate Hamiltonian must satisfy

$$\psi(x + Nd) = \exp(ikNd)\psi(x), \tag{2}$$

where k is the wave vector. Multiplying this wave vector by Planck's constant h divided by 2π yields the crystal momentum $p = (h/2\pi)k$. Free electrons in vacuum obey the Schrödinger equation

$$\begin{aligned} H\psi(x) &= (p^2/2m)\psi(x) \\ &= (h^2k^2/8\pi^2m)\psi(x) = E\psi(x), \end{aligned} \tag{3}$$

where H is the Hamiltonian, p is the momentum, m is the electron mass, and E is the energy. The free-electron eigenvectors $\psi(x)$ are the standing-wave sine functions $\sin(kx) = \sin(n\pi x/Nd)$ and cosine functions $\cos(n\pi x/Nd)$, or the traveling-wave functions $\exp(\pm in\pi x/Nd)$, where $n = -(N - 1)$ to N.

For the periodic linear lattice one can start from N simple free-electron traveling-wave functions $\exp(\pm ikx)$; the interactions between the N atoms or molecules will break the N-fold degeneracy of the energy E, and create narrowly spaced energy levels that can be considered a semicontinuous energy band of width W; these can fill, lowest energy first, up to the "band edges" $k = \pm\pi/d$. The traveling-wave functions must undergo Bragg scattering at these band edges, and the quadratic dependence of E on k in Eq. (3) must be softened at the band edges, so that an energy gap opens up and $dE/dk = 0$ at $k = \pm\pi/d$. A more realistic dispersion relation $E(k)$ for a periodic linear lattice is given within the tight-binding approximation (Bloch, 1928):

$$E(k) = U - 2t\cos(kd), \tag{4}$$

where U is the on-site energy,

$$U = \langle\psi_i|H|\psi_i\rangle, \tag{5}$$

for the electron on site i, H is the one-electron Hamiltonian, ψ_i is the wave function for the electron localized at site i, and t is the Mulliken transfer integral for an electron moving from site i to the adjacent site $i + 1$:

$$t = \langle \psi_i | H | \psi_{i+1} \rangle. \tag{6}$$

Chemists will recognize Eq. (4) from the simple Hückel molecular-orbital theory for aromatic π-electron systems; t is akin to the Hückel resonance integral β. The bandwidth W is given by

$$W = 4t. \tag{7}$$

This can be verified by defining $W = \max[E(k)] - \min[E(k)]$ in Eq. (4). This band can be filled by electrons (or holes) symmetrically up to the maximum (minimum) Fermi wave vectors k_F $(-k_F)$, either with only one electron per site (if the Coulomb electron–electron repulsion discourages more than one electron per site), or with two electrons (spin up and spin down) per site. If the band is filled up to the band edge, then

$$k_F = \pi/d. \tag{8}$$

If the band is only partially filled, then k_F will be some fraction of π/d. Bragg (Umklapp) x-ray scattering can be observed between the minimum $(-k_F)$ and the maximum $(+k_F)$ of band filling, at the reciprocal wave vector $2k_F$.

For free electrons the Fermi energy ϵ_F is given by

$$\epsilon_F = h^2 k_F^2 / 8\pi^2 m^*, \tag{9}$$

where m^* is the effective electron mass. For bound electrons in an energy band a different expression for ϵ_F can be crafted, e.g., from Eq. (4), as $\epsilon_F = U - 2t\cos(k_F d)$. A "filled band" has two electrons (or holes) per site (with spin up and spin down); a "half-filled band" has only one electron (or hole) per site; a "quarter-filled band" has one electron (or one hole) per two sites.

Very useful for the systems discussed here are the Hamiltonian named after John Hubbard (1963a, 1963b, 1964), but first discussed by Van Vleck (1932):

$$\begin{aligned} H = {} & -t \sum_{i,\sigma} [a^\dagger_{i,\sigma} a_{(i+1),\sigma} + a_{i,\sigma} a^\dagger_{(i+1),\sigma}] \\ & + U \sum_i a^\dagger_{i,\sigma} a_{i,\sigma} a^\dagger_{i,\sigma} a_{i,\sigma} \end{aligned} \tag{10}$$

and the extended Hubbard Hamiltonian (Hubbard, 1978):

$$\begin{aligned} H = {} & -t \sum_{i,\sigma} [a^\dagger_{i,\sigma} a_{(i+1),\sigma} + a_{i,\sigma} a^\dagger_{(i+1),\sigma}] \\ & + U \sum_i a^\dagger_{i,\sigma} a_{i,\sigma} a^\dagger_{i,\sigma} a_{i,\sigma} \\ & + V \sum_i a^\dagger_{(i+1)} a_{(i+1)} a^\dagger_i a_i, \end{aligned} \tag{11}$$

where $a^\dagger_{i,\sigma}$ creates an excitation of spin σ at site i, and $a_{i,\sigma}$ annihilates it, t is the Mulliken transfer integral [Eq. (6)], U is the on-site Coulomb energy [akin to the U of Eq. (5), but here it is a two-particle on-site energy], and V is the nearest-neighbor Coulomb interaction energy. Despite the phenomenological attractiveness of these two Hamiltonians, they cannot be solved for the general case. They have been solved exactly only for one-dimensional systems and for the two-state system, and numerically for a restricted set of conditions.

2. ONE-DIMENSIONAL INSTABILITIES

An organic crystal, or an organic thin film, is a three-dimensional object, but some of its properties are "quasi one-dimensional," i.e., resemble, but cannot coincide with, the predictions of one-dimensional physics, whose characteristics are simple but often peculiar.

An infinite one-dimensional "regular" chain of particles (atoms, electrons, or molecules), separated by equal distances d, is thermodynamically unstable; for instance, entropy considerations imply that it cannot exist as two or more distinct phases, or have a phase transition at finite temperatures (Landau and Lifschitz, 1958).

Peierls (1955) showed that a different, energetic instability of a one-dimensional electron gas, driven by electron–phonon interactions, causes nonuniformities in the electron distribution, and leads to a structural distortion and to a (usually second-order) phase transition, at a temperature T_P (the Peierls temperature). Below T_P either a dimerization into two sets of unequal interparticle distances d' and d'' (such that $d' + d'' = 2d$) or some other structural distortion must occur. The electronic energy of the metallic chain is lowered by the formation of a charge-density wave (CDW) of amplitude $\rho(x)$:

$$\rho(x) = \rho_0[1 + \alpha \cos(2k_F x + \phi)], \tag{12}$$

where x is the coordinate along the chain, ρ_0 is the uniform charge density, and $\alpha\rho_0$ is the charge-modulation amplitude. This phase transition opens up a "Peierls" energy gap 2Δ in the dispersion relation for the energy. Chemists are familiar with a conceptually similar Jahn–Teller distortion in organometallic systems. The Peierls distortion disrupts the periodicity, and can transform a metal (above T_P) to an electrical semiconductor or insulator (below T_P).

Below T_P the static (locked) CDW of the conduction electrons at $2k_F$ (or $4k_F$) couples with the other atomic or molecular electrons in the lattice; the resulting slight lattice distortions give rise to extra x-ray reflections. Above T_P, the CDWs are mobile excitations, with no phase locking between excitations on nearby chains: their x-ray signatures are diffuse reflections, similar to thermal diffuse scattering streaks in reciprocal space. These streaks sharpen as the temperature is lowered and T_P is approached. Below T_P, the static lattice distortion can be observed as new reflections between reciprocal-lattice layer lines. When the band filling is a rational fraction ($\frac{1}{4}$, $\frac{1}{2}$, $\frac{2}{3}$, 1, etc.) then the CDW excitations coincide with certain Bragg reflections of the background lattice, and are more difficult to detect.

When the band filling is not rational, the CDW reflections at wavelength λ are incommensurate with the background lattice of periodicity d, and the extent of charge transfer ρ (defined as the ratio of N_e, the net number of electrons, to N, the total number of sites) can be measured directly by the equation

$$\rho = N_e/N = 2d/j\lambda = 2dk_F/\pi, \tag{13}$$

where $j = 1$ for $2k_F$ and $j = 2$ for $4k_F$ scattering.

A similar magnetic ordering in the spin system can transform an ordered uniform paramagnetic chain ($S = \frac{1}{2}$ per site, $\rho = 1$) above a "spin-Peierls" transition temperature T_{SP} into a chain of spin-paired singlet "dimers" ($S = 0$ per dimer) below T_{SP}. One can also have a spin-density–wave (SDW) instability: many $\rho = \frac{1}{2}$ "one-chain" salts are paramagnetic above a SDW ordering temperature T_{SDW}, and become antiferromagnetic with a trapped SDW state below T_{SDW}. SDW instabilities can also have x-ray signatures.

The theory of CDW instabilities has received much attention: it differentiates between the weak-coupling limit ($U \ll t$), the intermediate-coupling limit, and the strong-coupling limit ($U > t$). There are other possible instabilities. Instabilities in a 1-D system, if driven by a strong on-site electron–electron Coulomb repulsion U, lead to a Mott–Hubbard insulator, particularly for $\rho = 1$ systems: here charge localization ensues, and the crystal becomes an insulator. For ρ values with rational fractions ($\rho = \frac{1}{2}$, $\frac{2}{3}$), a Wigner crystal can occur: the charges alternate regularly in the crystal, so that one site has $\rho = 0$, the next $\rho = 1$, and so on. Finally, an Anderson metal–insulator transition is driven by even a weak random field due to structural disorder, which can then localize the electronic states.

3. CONDUCTIVITY AND SUPERCONDUCTIVITY

The electrical conductivity σ (for electrons or holes) is defined by

$$\sigma = L/RA, \tag{14}$$

where the sample resistance R is measured between parallel planes L apart for a crystal of cross-sectional area A. For metals σ decreases with increasing temperature, as the mean free path of carriers is increasingly limited by scattering from lattice phonons. For semiconductors, σ increases exponentially with increasing temperature, usually following an Arrhenius-type temperature dependence:

$$\sigma = \sigma_0 \exp(-\Delta E/kT); \tag{15}$$

here ΔE is the experimental activation barrier, T is the temperature, and k is Boltzmann's constant. The conductivity is related to the carrier (electron or hole) mobility μ by

$$\sigma = Ne\mu, \tag{16}$$

where e is the electronic charge and N is the carrier concentration.

For superconductors, some of the electrons close to the Fermi surface form "Cooper pairs" of electrons with opposite spins and opposite crystal momenta, thanks to a "critical" coupling of both electrons to certain lattice phonons. This occurs below a critical temperature T_c, which is affected by both applied pressure and applied magnetic field. For external magnetic fields between zero and a critical field H_c, there is a partial or complete flux exclusion within the superconductor (Meissner–Ochsenfeld effect), so that a bulk superconductor is diamagnetic. In addition to the critical parameters T_c and H_c, there is also the critical current j_c above which the superconductor returns to the normal state.

In the Bardeen–Cooper–Schrieffer (BCS) theory (Bardeen *et al.*, 1957a, 1957b) T_c for the phase transition between the "normal" metal state and the lower-temperature superconducting state of Cooper pairs with infinite electrical conductivity is given by

$$T_c = 1.14\theta_D \exp[-1/U_{ep}D(\epsilon_F)], \tag{17}$$

where U_{ep} is the electron–phonon coupling energy, $D(\epsilon_F)$ is the density of states at the Fermi energy ϵ_F, and θ_D is the Debye temperature of the lattice. For elemental superconductors the product $U_{ep}D(\epsilon_F)$ is typically in the range 0.1 to 0.5, and so T_c is one to five orders of magnitude smaller than θ_D (see SUPERCONDUCTIVITY, LOW-TEMPERATURE). One "chemical" issue is how T_c could be raised, if BCS theory were applicable to new classes of superconductors.

4. MULLIKEN CHARGE-TRANSFER COMPLEXES IN SOLUTION

One of the early concerns was the formation of intense coloration in crystals of weak stoichiometric (1:1) complexes [benzene (**1**) (Fig. 1) with iodine (I_2), or naphthalene (**2**) with trinitrobenzene (TNB, **3**)] and other similar systems. This intense color is formed when the components are mixed in solution, or when they are co-crystallized as stoichiometric solid-state complexes. These solutions and crystals showed, almost unchanged, the full optical absorption spectrum of the neutral components, plus an extra broad, intense absorption band with little or no vibrational structure.

The phenomenon was not understood until Mulliken (1950) proposed his charge-transfer theory. He explained that these 1:1 solution complexes have ground-state wave functions $\Psi_{GS}(D,A)$ and excited-state wave functions $\Psi_{ES}(D,A)$ that are linear combinations of the wave functions ϕ_0 of the neutral donor D and the neutral acceptor A, and of the charge-transfer wave functions ϕ_{CT} of the donor cation D^+ and the acceptor anion A^-, as follows:

$$\Psi_{GS}(D,A) = a\phi_0(DA) + b\phi_{CT}(D^+A^-), \tag{18}$$

$$\Psi_{ES}(D,A) = a\phi_{CT}(D^+A^-) - b\phi_0(DA). \tag{19}$$

Typical values for the coefficients are $a \ll b$, e.g., $a = 0.05$, $b = 0.95$. For the benzene–iodine complex D = one-electron donor = benzene, A = one-electron acceptor = I_2. The charge-transfer transition is between the states $\Psi_{GS}(D,A)$ and $\Psi_{ES}(D,A)$, with charge-transfer band

$$h\nu_{CT} = \langle \Psi_{ES}(D,A)|H|\Psi_{ES}(D,A)\rangle - \langle \Psi_{GS}(D,A)|H|\Psi_{GS}(D,A)\rangle, \tag{20}$$

where H is the appropriate Hamiltonian.

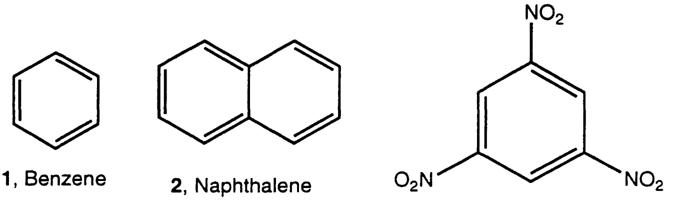

1, Benzene **2**, Naphthalene **3**, TNB

FIG. 1. Molecular structures of the weak one-electron donors benzene (**1**) and naphthalene (**2**), and of the weak one-electron acceptor 1,3,5-trinitrobenzene (**3**).

This transition is strongly allowed, because the D^+A^- state (both molecules at typical van der Waals separations, 3.3–3.8 Å apart) has $S = \frac{1}{2}$ for each ion, but $S = 0$ overall, thanks to the mixing of states in Eqs. (18) and (19). One can quantify the charge transfer ρ ($0 \le \rho \le 1$) between adjacent D and A molecules from the amount of mixing of ionic states with the neutral states:

$$\rho = a/(a^2 + b^2)^{1/2}. \tag{21}$$

Nothing in the Mulliken theory prevents a large value of ρ.

The 1:1 complex of benzene (or naphthalene) with trinitrobenzene is called a Mulliken charge-transfer complex, but is also often called a $\pi{:}\pi$ complex, because it is considered to be due to the weak intermolecular interaction of the π electrons of benzene or naphthalene with the π electrons of trinitrobenzene. When discussing molecular orbitals (molecular wave functions), chemists distinguish between σ electrons (which build up electron density along the internuclear axis, and have stronger localization between bonding centers) and π electrons (which have nodes along the internuclear axis, high electron density in the direction perpendicular to the internuclear axis, and greater delocalization between the atoms of the molecule). In planar aromatic molecules, like benzene and naphthalene, there are several π electrons (6 for benzene, 10 for naphthalene) whose wave functions have amplitudes above and below the plane formed by the C and H atoms, and which are considered to be much more mobile than the σ electrons, and almost "separate" from the σ-electron network.

5. CHARGE-TRANSFER CRYSTALS

The arguments used for solution and gas-phase charge-transfer complexes could be extended to a charge-transfer crystal, which Mulliken called an "∞:∞" complex. The first solid charge-transfer crystals were colored, but had low thermodynamic stability, relative to the separate neutral donor and acceptor crystals. However, if the donor D had low gas-phase ionization potential I_D, and if the acceptor had high gas-phase electron affinity A_A—

$$D(g) \rightarrow D^+(g) + e^-, \quad \Delta U = I_D; \tag{22}$$

$$e^- + A(g) \rightarrow A^-(g), \quad \Delta U = -A_A \tag{23}$$

—then the solid complexes tended to show much higher electrical conductivity, and to be paramagnetic. McConnell *et al.* (1965) explained that, if the stacking in the solid state was of planar molecules D atop planar molecules A (later this was called the "mixed-stack" case), then the crystals tended to fall into two distinct classes: the almost ionic case ($\rho \approx 1$), when the crystal Coulomb or Madelung energy E_M was more negative than the difference $I_D - A_A$ was positive, i.e.,

$$E_M + I_D - A_A < 0 \quad \text{(ionic case, } \rho \approx 1), \tag{24}$$

or the cost of ionizing the lattice was more than repaid by the ionic binding energy of the crystal; and the almost neutral case ($\rho \approx 0$) when the Madelung energy was not enough to counterbalance the cost of ionizing the lattice:

$$E_M + I_D - A_A > 0 \quad \text{(neutral case, } \rho \approx 0). \tag{25}$$

The (almost) ionic complexes would be paramagnetic and have moderate (semi)-conductivity. For "strong" one-electron donors $I_D < 6.5$ eV; for strong one-electron acceptors $A_A > 2.5$ eV. Most strong donors are poor acceptors, and vice versa. A plot of I_D and A_A as a function of number of π electrons converges asymptotically at $I_D = A_A = 4.4$ eV, the value for graphite, which is as good an electron donor as it is an electron acceptor.

In the exact case $\rho = 1$ (one electron per molecule) the electrons (or holes) will be localized, since any transport of charge will put two electrons on the same site, at a huge cost in on-site Coulomb repulsion energy; thus a $\rho = 1$ system is known as a Mott–Hubbard insulator.

The discovery of strong cyanocarbon acceptors TCNE (tetracyanoethylene, **4**) (Fig. 2) by Cairns *et al.* (1957), and TCNQ (7,7,8-8-tetracyanoquinodimethane, **5**) by Acker *et al.* (1960) opened the door to strongly ionic complexes, with interesting magnetic properties. These ion-radical salts ("one-chain" salts) were based either on the the strong

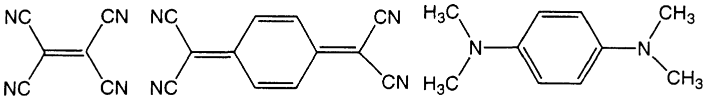

4, TCNE **5**, TCNQ **6**, TMPD

FIG. 2. Molecular structures of the strong one-electron acceptors TCNE (**4**) and TCNQ (**5**), and of the strong one-electron donor TMPD (**6**).

electron acceptor TCNQ, or on the strong electron donor TMPD (*N,N,N′,N′*-tetramethyl-para-phenylenediamine, **6**). The perchlorate salt of TMPD, known as Wurster's blue perchlorate ($TMPD^+ClO_4^-$), led to an understanding of the paramagnetism of linear-chain systems (Thomas *et al.*, 1963), which, at a spin-Peierls ordering temperature T_{SP}, undergo a spin-Peierls transition to a diamagnetic low-temperature ground state.

6. OPTICS OF POLYENES

Kuhn (1949) analyzed the lowest-energy optical absorption band of long linear polyenes, or oligomers of conjugated linear polymers, by free-electron molecular-orbital theory (a slight modification of the particle-in-a-box problem). $E_{\max}$, the energy of maximum absorbance for the lowest optical absorbance band, can be related to n_π, the number of π electrons, to L, the length of the linear "box," and to L_O, the "effective length" of the π-electron chain in the oligomer, defined so that $L = n_\pi L_O$:

$$E_{\max} \approx h^2 n_\pi / 8mL^2 = h^2/8mL_O^2 n_\pi. \qquad (26)$$

Thus a plot of $E_{\max}$ versus $1/n_\pi$ is linear for oligomeric substrands of known conducting polymers, and should extrapolate to $E_{\max} = 0$ for the perfectly degenerate and conjugated infinite linear polymer, which will be metallic. For instance, $E_{\max} = 0$ for graphite and for $(SN)_x$.

7. PROPOSED EXCITONIC SUPERCONDUCTIVITY

Little (1964) made a controversial but very stimulating proposal. If the electron–phonon coupling U_{ep} in Eq. (17) is replaced by the much larger electron–electron coupling matrix element V_{ee}, then T_c should become two orders of magnitude larger, and superconductivity at room temperature becomes plausible. This proposal for "excitonic superconductivity" became a mantra for the hopes for superconductivity at higher T_c in organic systems. No systems to date have been proven to implement the mechanism implied in Little's proposal.

Having discussed the general physics of the subject, we now present brief overviews of inorganic superconductors (the "competition" for the organic case) and of organic quasi one-dimensional metals.

8. INORGANIC SUPERCONDUCTORS

Elemental metallic superconductors have a maximum $T_c = 9.2$ K for Nb, and systems of two metals, such as the A15 crystal Nb_3Ge, reached $T_c = 23.2$ K (Gavaler, 1973). A search for novel superconductors first concentrated on the highly conducting inorganic "Krogmann salts," such as $K_3Pt(CN)_4Br_{0.3}$ $2.3H_2O$, which are stacks of cyanoplatinate ions with Pt ··· Pt nonbonded distances of 2.77 Å (same as in bulk metallic Pt) and "quasi one-dimensional" metallic conduction along the Pt chains (Krogmann, 1969). The CDW states were first found experimentally in the Krogmann salts, but no superconductors were found among them.

As mentioned above, $T_c = 0.55$ K was found in the stage-1 layered graphitic compound KC_8, and superconductivity with $T_c = 0.3$ K was also found for the crystalline polymer polythiazyl, $(SN)_x$.

In the 1980s ceramic perovskite defect oxide superconductors drove T_c very high indeed: 25–40 K in the "214" system $La_{2-x}Sr_xCuO_{4-y}$ (Bednorz and Müller, 1986), 93 K in the "123" system $YBa_2Cu_3O_{7-x}$ (Wu *et al.*, 1987), 110 K in the Bi-1223 system (Maeda *et al.*, 1988), 125 K in the Tl-2134 system (Parkin *et al.*, 1988), and 150 K under pressure in the Hg-1223 system (Chu *et al.*, 1993). Applications are at present limited by relatively low values of the critical current j_c, but intense work is under way to overcome this limitation by better processing and increase of contact between superconducting grains.

The discovery of buckminsterfullerene C_{60} (Kroto *et al.*, 1985) and its practical synthesis (Krätschmer *et al.*, 1990) were followed by the finding that alkali-metal salts of buckminsterfullerenes A_3C_{60} (A = K, Rb, Cs) superconduct from T_c = 18.6 to 40 K (Hebard *et al.*, 1991; Palstra *et al.*, 1995), as does one alkaline-earth salt. These compounds are, however, extremely susceptible to water vapor.

9. TTF-TCNQ AND OTHER QUASI ONE-DIMENSIONAL METALS

After the synthesis of tetrathiafulvalene (TTF, **7a**; see Fig. 3) (Wudl *et al.*, 1970), the first "organic metal," TTF TCNQ, was characterized by Cowan and co-workers (Ferraris *et al.*, 1973). From erroneous measurements in some samples of TTF TCNQ Heeger and co-workers claimed "superconducting fluctuations" at 58 K (Coleman *et al.*, 1973). This stimulated a veritable explosion of interest in organic ionic solids; careful measurement showed that there is no superconductivity in TTF TCNQ (Thomas *et al.*, 1976). TTF TCNQ belongs to a new class of metallic conductors, the quasi one-dimensional metals. These molecular conductors have decreasing electrical conductivity at increasing temperature, and large directional anisotropy in this conductivity: along the molecular and ionic stacks, where there is reasonable intermolecular (interionic) π–π overlap, the conductivity is typically 10 to 1000 times larger than in the transverse directions. These are almost one-dimensional systems, which became real laboratory examples instead of mere theoretical curiosities. Many variants on TTF were studied, and an understanding of what it takes to make organic metals, and what may be needed for organic superconductivity, evolved gradually (Cowan, 1989; Cowan *et al.*, 1990).

Figure 4 shows the room-temperature crystal structure of TTF TCNQ (Kistenmacher *et al.*, 1974), and Fig. 5 shows its conductivity as a function of temperature. At high temperature TTF TCNQ is metallic, with $\sigma(T) \propto T^{-2.3}$; since it has a reasonably high coefficient of thermal expansion, as do most molecular crystals, therefore a more meaningful quantity to consider is the conductivity at constant volume, for which $\sigma_V(T) \propto T^{-1.29}$: This means that one-phonon scattering processes are dominant (Conwell, 1980). It has been determined that a CDW

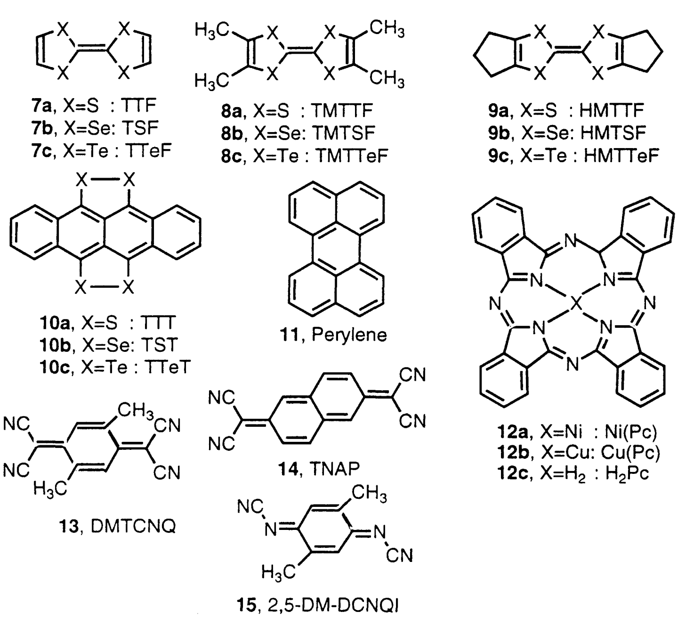

FIG. 3. Molecular structures of TTF (**7a**) and other chalcogen-containing electron donors (**7b–10c**), of perylene (**11**) and phthalocyanines (**12**), and of three electron acceptors (**13–15**) similar to TCNQ.

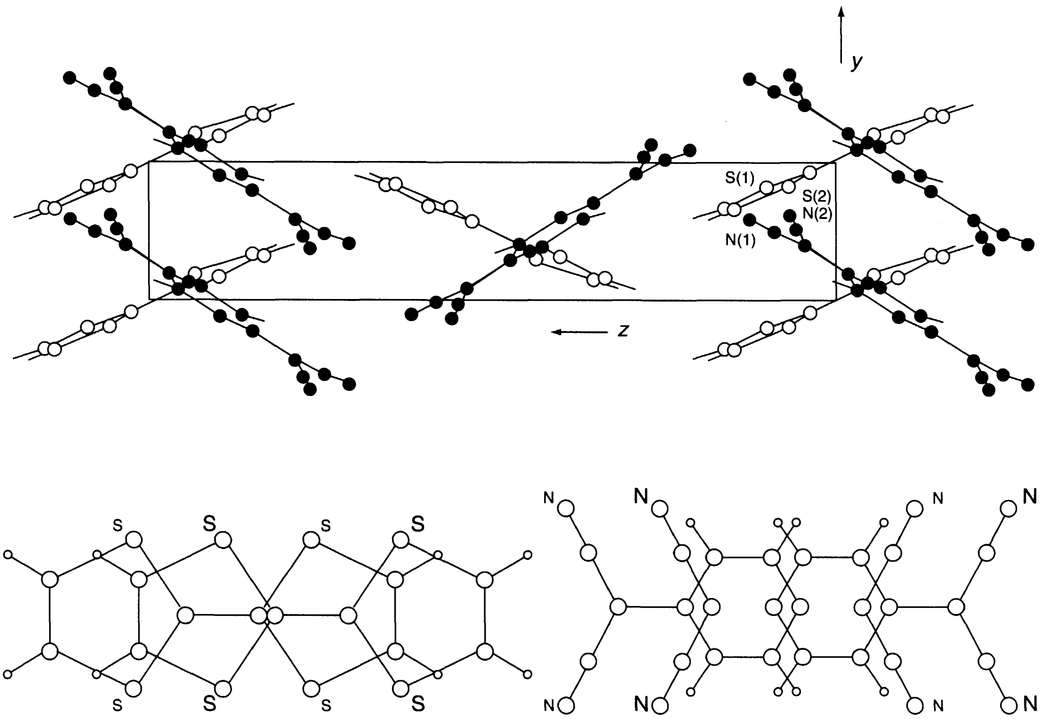

FIG. 4. Room-temperature crystal structure of TTF TCNQ (Kistenmacher *et al.*, 1974). The upper diagram is a projection along [100], with the unit cell axes *a* vertical and *c* horizontal, which shows the stacks of TTF (open circles for the atom positions) and TCNQ (filled circles for the atom positions). The lower diagram shows the intermolecular overlap, projected normal to the least-squares molecular planes, along the TTF stack (left), and the TCNQ stack (right).

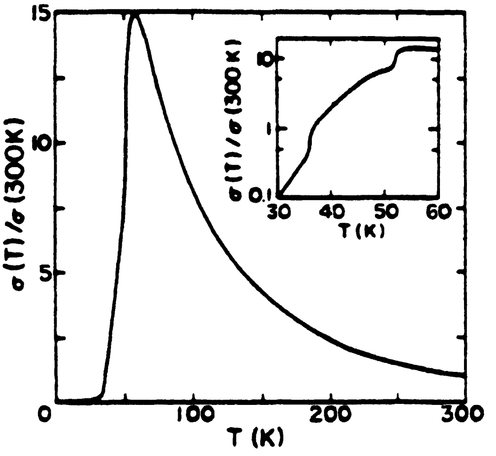

FIG. 5. Temperature dependence of the conductivity σ of TTF TCNQ (Ferraris *et al.*, 1973).

starts on the TCNQ stacks at about 160 K; at 54 K a coupling occurs between CDWs on different TCNQ chains; at 49 K a CDW starts on the TTF stacks, and by 38 K a complete Peierls state is observed. The TTF molecules exhibit a slip of only about 0.034 Å along their long molecular axis, relative to the ordered structure (Coppens *et al.*, 1987); thus the Peierls CDW state involves a very subtle structural distortion.

Table 1 summarizes (Cowan *et al.*, 1990) the electrical conductivity of several charge-transfer crystals, most of which have a definite Peierls transition to a semiconducting state below a temperature $T_{\max}$. In several cases, however, there is instead a very broad maximum in the plot of conductivity versus temperature, and the compound retains its high conductivity to the lowest temperatures measured, without going superconducting: one salt, Cu(2,5-DM-DCNQI)$_2$, even reaches at 0.5 K a conductivity of about 5×10^5 S cm^{-1}. Figure 6 shows the temperature de-

Table 1. Selected organic metals (Cowan *et al.*, 1990). None of these compounds superconduct. The temperature of maximum conductivity is in parentheses if it is a broad conductivity maximum. The molecular structures are given in Figs. 2 (molecule **5**) and 3 (molecules **7–15**).

Compound	Molecular structure		Charge transfer (ρ)	Room-temp. conductivity (σ_{RT}/S cm^{-1})	Maximum conductivity (σ_{max}/S cm^{-1})	Temp. of max. cond. (T_{max}/K)
	D	*A*				
TTF TCNQ	**7a**	**5**	0.59	500	2×10^4	59
TMTTF TCNQ	**8a**	**5**	0.65	350	5×10^3	60
HMTTF TCNQ	**9c**	**5**	0.72	500	2×10^3	75
TSF TCNQ	**7b**	**5**	0.63	800	1×10^4	40
TMTSF TCNQ	**8b**	**5**	0.57	1200	7×10^3	61
TMTSF DMTCNQ	**8b**	**13**	0.50	500	5×10^3	42
HMTSF TCNQ	**9b**	**5**	0.74	2000	7×10^3	(32)
TTeF TCNQ	**7c**	**5**	0.71	1800	2.5×10^4	<4
HMTTeF TCNQ	**9c**	**5**	n.a.	550	9×10^2	(73)
HMTTeF DMTCNQ	**9c**	**13**	n.a.	460	1×10^3	(83)
HMTSF TNAP	**9b**	**14**	n.a.	2900	2×10^4	50
$(TTT)_2I_3$	**10a**		0.50	1000	3×10^3	(60)
$(TST)_2I$	**10b**		0.50	3900	1×10^4	(35)
$(TST)_2Cl$	**10b**		0.50	2100	2×10^4	(26)
(Perylene)$_2$(PF$_6$)$_{1.1}$	**11**		0.55	900	1×10^3	200
Ni(Pc)I	**12a**		0.33	550	5×10^3	(25)
Cu(Pc)I	**12b**		0.33	900	7×10^3	$\leqslant$(30)
H$_2$(Pc)I	**12c**		0.33	750	4×10^3	15
Cu(2,5-DM-DCNQI)$_2$	**15**		0.50	800	5×10^5	(3.5)

pendence of the resistivity of four TCNQ salts; HMTSF TCNQ does not undergo a Peierls transition.

10. (TMTSF)$_2$*X* OR BECHGAARD SALTS: THE FIRST ORGANIC SUPERCONDUCTORS

By using the donor TMTSF (tetramethyltetraselenafulvalene, **8b**) and electro-crystallization, Bechgaard, Jérome, and co-workers found that (TMTSF)$_2$PF$_6$ was the first organic superconductor, with T_c = 0.9 K at an applied pressure of 10 kbar (Jérome *et al.*, 1980). Here the lateral two-dimensional Se–Se interactions and the applied pressure together defeat the Peierls transition. The first ambient-pressure superconductor, (TMTSF)$_2$ClO$_4$ (T_c = 1.4 K), followed quickly; Figure 7 shows its crystal structure.

Table 2 summarizes all the known organic superconductors, plus the stage-1 intercalated graphite KC$_8$, the polymeric (SN)$_x$ discussed above, and the fullerene superconductors. Additional relevant molecular structures of donors and acceptors are shown in Fig. 8.

Eight superconducting "Bechgaard" salts (TMTSF)$_2$*X* (*X* = inorganic anions) were found, but almost all have $T_c \approx$ 1 K; the highest T_c is 2.1 K for *X* = FSO$_3^-$. For the Bechgaard salts, an ordering in the tetrahedral anions (ReO$_4^-$, ClO$_4^-$ under rapid cooling past 24 K) or in the octahedral anions (PF$_6^-$) can occur; this doubles the unit cell size, opens up a small band gap, and causes antiferromagnetic ordering of the spins in the TMTSF species to a new collective spin-density wave (SDW) state. This ordering makes these salts insulating, not superconducting. In many cases, hydrostatic pressure can prevent the anions from ordering, and a smooth transition from the metallic state to the superconducting state can occur. Figure 9 shows what is understood about the temperature–pressure phase diagram of several TMTSF salts.

11. (BEDT-TTF)$_2$*X* AND OTHER ORGANIC SUPERCONDUCTORS

Cava and coworkers prepared the "all-sulfur" donor BEDT-TTF (bis-ethylenedithiolene-tetrathiafulvalene, or ET for short, **16a**) (Mizuno *et al.*, 1978). Saito and co-workers found that in some (BEDT-TTF)$_2$*X* salts the usual Peierls-driven metal-to-insula-

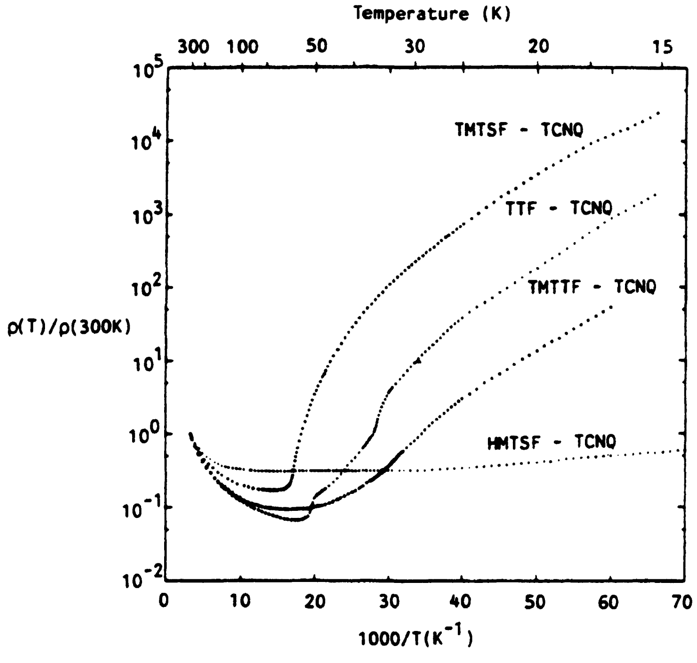

FIG. 6. Temperature dependence of the electrical resistivity ($1/\sigma$) of four TCNQ salts (Bloch *et al.*, 1977).

tor phase transition was suppressed (Saito *et al.*, 1982); then Parkin and co-workers found in $(ET)_2ReO_4$ the first sulfur-based superconductor (Parkin *et al.*, 1983). These salts with formula $(ET)_2X$ (X = inorganic anions) crystallize in a wide variety of crystallographically distinct phases, sometimes with several phases growing from the same solution (Williams *et al.*, 1992). This polymorphism, rare among organic systems, is baffling; for some anions X, certain phases are insulating, while others are highly conducting, and some are superconducting. There seem to be several closely ranked cohesive-energy minima that are available to the growing microcrystal.

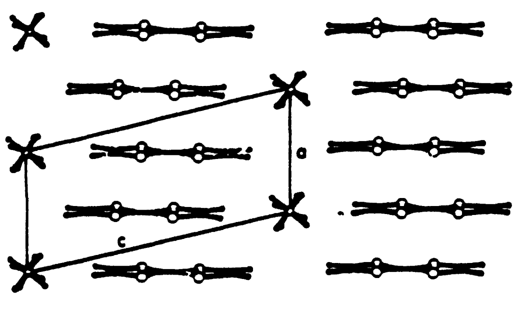

FIG. 7. Crystal structure of $(TMTSF)_2ClO_6$ (Thorup *et al.*, 1981), projected along [010]. The unit cell is shown; the TMTSF stacking is vertical; the stacks are separated by perchlorate anions, which are at the corners of the unit cell.

There are 20 organic superconductors of the ET_2X family; the present record critical temperature is T_c = 12.8 K at 0.3 kbar for κ-(BEDT-TTF)$_2$Cu[N(CN)$_2$]Cl; close ET cousins are one superconducting salt of BEDT-TSF (**16b**), and one superconducting salt of BEDO-TTF (**16c**). Other organic "hole" superconductors are salts of the electron donors MDT-TTF (**17**) (one superconductor), of EDT-TTF (**18**), of MET (**19**), of DMET (**20**) (seven superconducting salts, maximum T_c = 1.9 K), and of TMTTF (**8a**) (one superconductor). Electron superconductors are formed with the electron acceptor anions Ni(dmit)$_2$ (**21a**) (three superconductors, maximum T_c = 5.0 K at 7 kbar) and Pd(dmit)$_2$ (**21b**). Ten alkali-metal salts of C_{60} and one calcium salt of C_{60} are superconductors, but with much higher critical temperatures.

So far, all TMTSF and ET-type superconductors have true mixed valence and $\rho = \frac{1}{2}$ [except $\rho = \frac{2}{3}$ for (BEDT-TTF)$_2$Cl$_3$·2H$_2$O]. The TMTSF salts, and some ET salts, are quasi one-dimensional, but with enough warping of the Fermi surface to make them pseudo 2-D, and to defeat the Peierls transition. The κ-phase (BEDT-TTF) salts are really 2-D systems; this phase, shown in Fig.

Table 2. Molecular structure of donor or acceptor, and superconducting critical temperatures T_c (/kelvin, at 1 bar) or T_c (/kelvin, at applied pressure P/kbar) for organic, intercalated graphite, $(SN)_x$, and fullerene superconductors (updated from Williams *et al.*, 1992). The molecular structures are given in Figs. 3 (molecules **7–8**) and 8 (molecules **16–22**).

Compound	Molecular structure	T_c/K @ P/kbar	Compound	Molecular structure	T_c/K @ P/kbar
KC_8 (stage-1 graphite)		0.55	$(ET)_4Hg_{3-x}Cl_8$	**16a**	1.8 @ 12
$(SN)_x$		0.3	$(ET)_4Hg_{2.89}Br_8$	**16a**	4.3
$(TMTSF)_2PF_6$	**8b**	0.9 @ 10	$(BEDO\text{-}TTF)_3Cu(NCS)_3$	**16c**	1.1
$(TMTSF)_2ClO_4$	**8b**	1.2	$\lambda\text{-}(BEDT\text{-}TSF)_2GaCl_4$	**16b**	8.0
$(TMTSF)_2AsF_6$	**8b**	1.1 @ 12	$(MDT\text{-}TTF)_2AuI_2$	**17**	4.5
$(TMTSF)_2SbF_6$	**8b**	0.4 @ 11	$(DMET)_2I_3$	**20**	0.47
$(TMTSF)_2TaF_6$	**8b**	1.4 @ 12	$(DMET)_2IBr_2$	**20**	0.59
$(TMTSF)_2ReO_4$	**8b**	1.3 @ 9.5	$(DMET)_2AuCl_2$	**20**	0.83
$(TMTSF)_2FSO_3$	**8b**	2.1 @ 6.5	$(DMET)_2AuBr_2$	**20**	1 @ 1.5
$(TMTTF)_2Br$	**8a**	0.8 @ 26	$\kappa\text{-}(DMET)_2AuBr_2$	**20**	1.9
$(ET)_2ReO_4$	**16a**	2.0 @ 4.5	$(DMET)_2AuI_2$	**20**	0.55 @ 5
$\alpha_t\text{-}(ET)_2I_3$	**16a**	7–8	$(DMET)_2Au(CN)_2$	**20**	0.8 @ 5
$\alpha\beta\text{-}(ET)_2I_3$	**16a**	2.5–6.9	$(TTF)[Ni(dmit)_2]_2$	**7a,21a**	5.0 @ 7
$\beta\text{-}(ET)_2I_3$	**16a**	1.4	$[(CH_3)_4N][Ni(dmit)_2]_2$	**21a**	3.0 @ 3.2
$\beta^*\text{-}(ET)_2I_3$	**16a**	8.0 @ 0.5	$\alpha\text{-}(TTF)[Pd(dmit)_2]_2$	**21b**	1.7 @ 22
$\gamma\text{-}(ET)_3(I_3)_{2.5}$	**16a**	2.5	$\alpha'\text{-}(TTF)[Pd(dmit)_2]_2$	**21b**	6.42 @ 20.7
$\theta\text{-}(ET)_2(I_3)_{0.98}(AuI_2)_{0.02}$	**16a**	3.6	$\alpha\text{-}(EDT\text{-}TTF)[Ni(dmit)_2]$	**18,32a**	1.3
$\kappa\text{-}(ET)_2(I_3)$	**16a**	3.6	Na_2CsC_{60}	**22**	10.5
$\beta\text{-}(ET)_2IBr_2$	**16a**	2.8	K_3C_{60}	**22**	19.6
$\beta\text{-}(ET)_2AuI_2$	**16a**	4.98	K_2RbC_{60}	**22**	22.5
$(ET)_3Cl_2\cdot 2H_2O$	**16a**	2 @ 16	$K_1Rb_2C_{60}$	**22**	28.0
$\alpha\text{-}(ET)_2NH_4Hg(SCN)_4$	**16a**	1.15	$K_{1.5}Rb_{1.5}C_{60}$	**22**	25.1
$\beta(ET)_{1.96}(MET)_{0.04}I_3$	**16a,19**	4.6	Rb_3C_{60}	**22**	29.8
$\kappa\text{-}(ET)_2Cu(NCS)_2$	**16a**	10.4	Rb_2CsC_{60}	**22**	31.3
$\kappa\text{-}(ET)_2Cu(N(CN)_2)Br$	**16a**	11.6	$Rb_1Cs_2C_{60}$	**22**	33
$\kappa\text{-}(ET)_2Cu(N(CN)_2)Cl$	**16a**	12.8 @ 0.3	Cs_3C_{60}	**22**	29.5
$\kappa\text{-}(ET)_2Ag(CN)_2\cdot H_2O$	**16a**	5.0	Cs_3C_{60}	**22**	40 @ 10
$\kappa_L\text{-}(ET)_2Cu(CF_3)_4\cdot TCE$	**16A**	4.0	Ca_5C_{60}	**22**	8.4

10, has isolated dimers connected by dispersion interactions to form rough two-dimensional sheets. Figure 11 shows the temperature dependence of the conductivity of κ-(BEDT-TTF)$_2$Cu(NCS)$_2$.

Alas, structural correlations of T_c with chemical formula or structure type are limited. The β-(ET)$_2X$ salts with linear anions have a linear dependence of T_c with anion length (but this correlation fails for very long

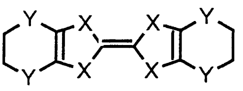

16a, X=S , Y=S: BEDT-TTF = ET
16b, X=Se, Y=S: BEDT-TSF
16c, X=S ,Y=O: BEDO-TTF

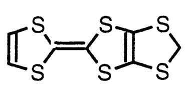

17, MDT-TTF

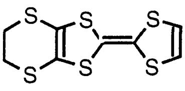

18, EDT-TTF

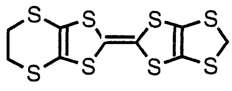

19, MET

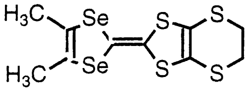

20, DMET

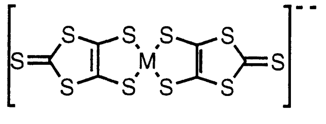

21a, M=Ni: Ni(dmit)$_2$ $^{--}$
21b, M=Pd: Pd(dmit)$_2$ $^{--}$

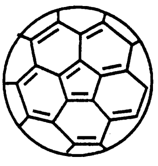

22, C_{60}

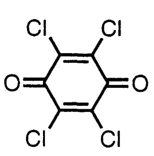

23, pChl

FIG. 8. Molecular structures of more donors (**16–20**) and acceptors (**21–23**).

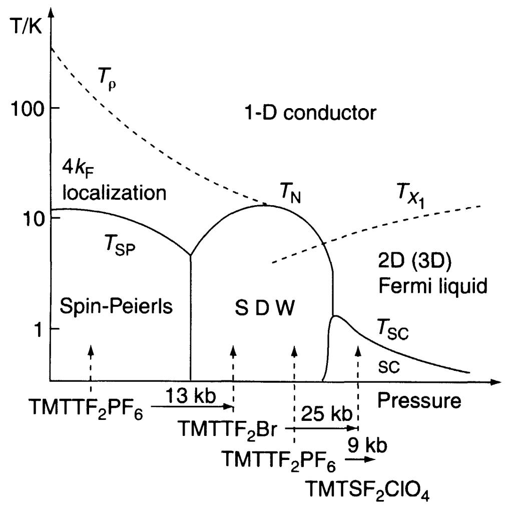
T/K
100
10
1
T_p
1-D conductor
$4k_F$ localization
T_N
T_{X_1}
T_{SP}
2D (3D) Fermi liquid
Spin-Peierls
S D W
T_{SC}
SC
Pressure
$TMTTF_2PF_6$
13 kb
25 kb
$TMTTF_2Br$
9 kb
$TMTTF_2PF_6$
$TMTSF_2ClO_4$

FIG. 9. Temperature-pressure phase diagram of TMTSF salts (Jérome, 1990).

anions, as other phases are formed). The $(TMTSF)_2X$ salts with tetrahedral anions X show a linear dependence of the Peierls metal-to-insulator phase transition temperature with tetrahedral anion radius.

All entries in Table 2 are "hole" superconductors, except for the metal $(dmit)_2$ salts, where the carriers are electrons. The room-temperature conductivities of these compounds are usually about one or two orders of magnitude smaller than those shown in Table 1.

Included in Table 2 are the superconducting (but very air-sensitive) alkali-metal and alkaline-earth fullerides: these are compounds with three-dimensional superconductivity, where the alkali-metal ions are just counterions tucked into tetrahedral and octahedral holes in the cubic fullerene crystal structure (Fig. 12). One can idly debate whether C_{60} is an organic molecule, or just a molecular allotrope of elemental carbon: in fact it is the largest molecule ever "synthesized" by astrophysicists. The critical temperatures of ET salts seem to be stuck at about 13 K, while the fullerene salts have reached 40 K, and the ceramic superconductors are far ahead, at 150 K. For amusement, one can plot the critical temperature T_c as a function of year of discovery (Fig. 13). Such plots do not predict scientific success, but stimulate our hopes.

The coherence length in the organic superconductors is of the order of a lattice constant (as it is for the ceramic-oxide superconductors). The organic superconductors are of type II (there are two critical fields). The "dimensionality" of the (super)conductivity is between 1 and 2 for the organic

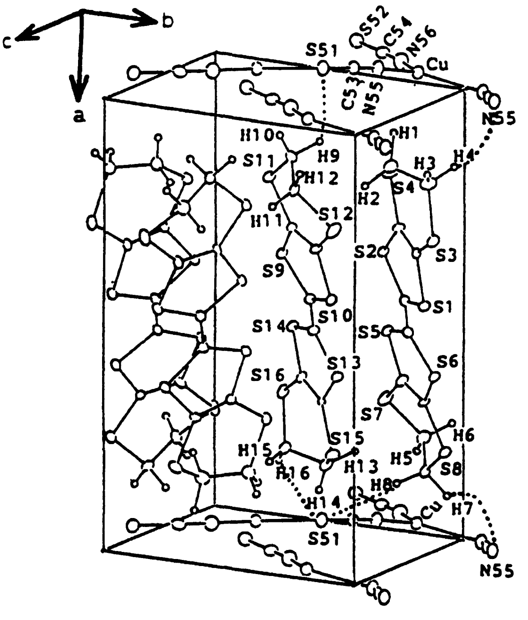
c
b
a
S51
C52
C54
N56
Cu
C53
N55
H10
H1
S11
H9
H3
H4
H12
S4
H2
S12
H11
S2
S3
S9
S1
S10
S14
S5
S13
S6
S16
S7
S15
H6
H15
H13
H5
S8
H16
H8
H14
H7
S51
N55

FIG. 10. Crystal structure of κ-$(BEDT\text{-}TTF)_2Cu(NCS)_2$ at 104 K (Oshima *et al.*, 1988).

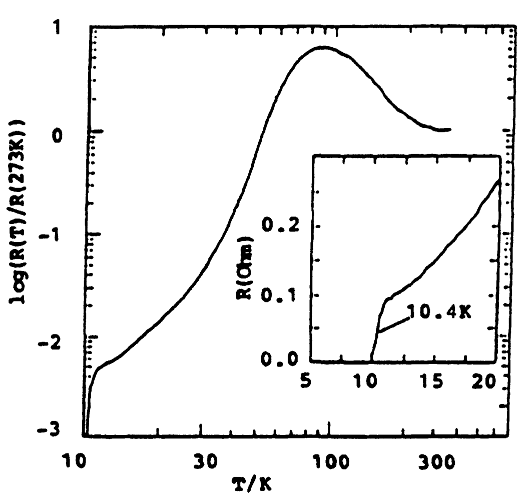

FIG. 11. Temperature dependence of the resistivity ($1/\sigma$) of κ-$(BEDT\text{-}TTF)_2Cu(NCS)_2$ (Saito, 1990).

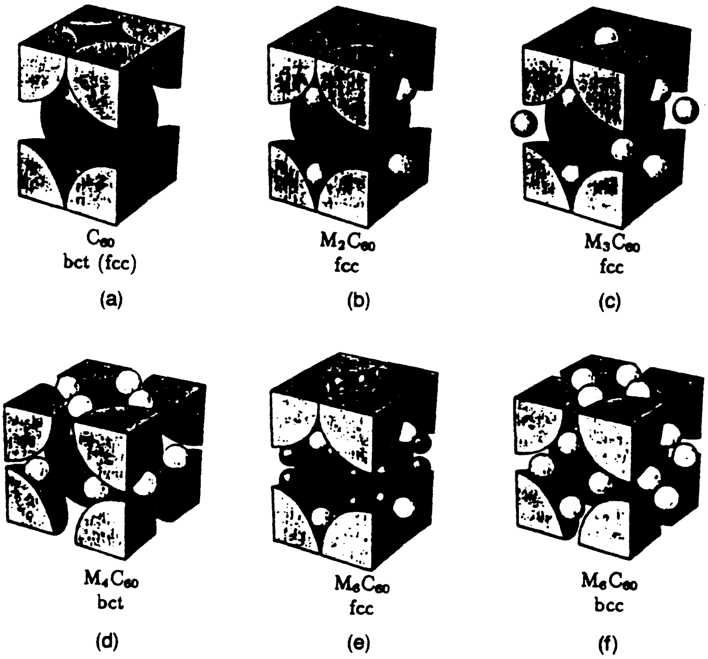
C_{60} bct (fcc) (a)

M_2C_{60} fcc (b)

M_3C_{60} fcc (c)

M_4C_{60} bct (d)

M_6C_{60} fcc (e)

M_6C_{60} bcc (f)

FIG. 12. Crystal structures of alkali fullerides (Dresselhaus *et al.*, 1993).

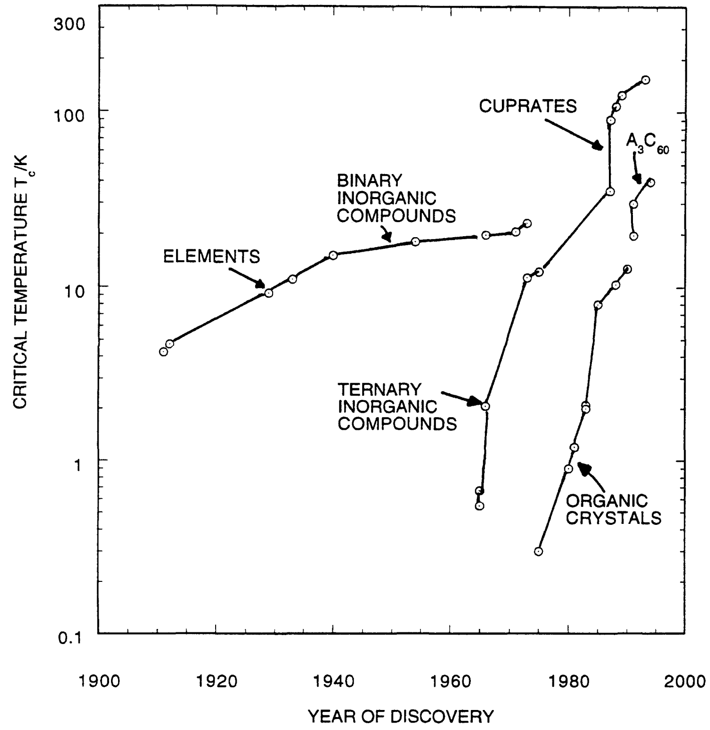

FIG. 13. Logarithm of the critical temperatures T_c of superconductors, plotted against the year of their discovery (adapted from Day, 1990).

superconductors (it is 3 for the fullerides). For organic superconductors, the isotope-effect results are unclear: It is not certain which phonon modes are important for the superconductivity.

For the ceramic-oxide superconductors the mechanism of superconductivity is difficult to explain by standard BCS theory. For the organic superconductors, the superconductivity mechanism is probably of the BCS type.

12. CONDUCTING POLYMERS

After the short-lived interest in polythiazyl, $(SN)_x$ (**24**), and its superconductivity, organic conducting polymers were found when the Shirakawa synthesis (1971) of *cis*- and *trans*-polyacetylene, $(CH)_x$ (**25**), first yielded insulating films of $(CH)_x$, and then Shirakawa and co-workers (1977) doped polyacetylene with electron donors (e.g., sodium) and acceptors (e.g., iodine) to form the first high-conductivity organic polymer. This was followed (see Fig. 14; for a review, see Skotheim, 1986) by the discovery of other *n*-doped or *p*-doped conducting polymers, such as poly-paraphenylene, $(C_6H_4)_x$ (**26**), poly-(para-phenylene vinylene), $(C_8H_6)_x$ (**27**), and the air-stable polymers, such as polypyrrole, $(C_4H_3N)_x$ (**28**), and polythiophene, $(C_4H_2S)_x$ (**29**), and the rediscovery of the properties of emeraldine salt, $(C_{24}H_{22}N_4)_x$, the conducting form of polyaniline (**30**).

The discovery of organic conducting polymers such as polyacetylene raises the question of how long a conducting strand can be, and when the strand can be conducting. When first formed, polyacetylene is in the *cis* form; this can be converted to the all-*trans* form by heating, or by doping with electron donor or acceptor counterions (sodium cation, iodide anion, etc.). Although the band gap in *trans*-$(CH)_x$ is large (4.7 eV), the existence of midgap infrared absorption led to the suggestion that the relevant excitations were neutral solitons (**33**; see Fig. 15) or po-

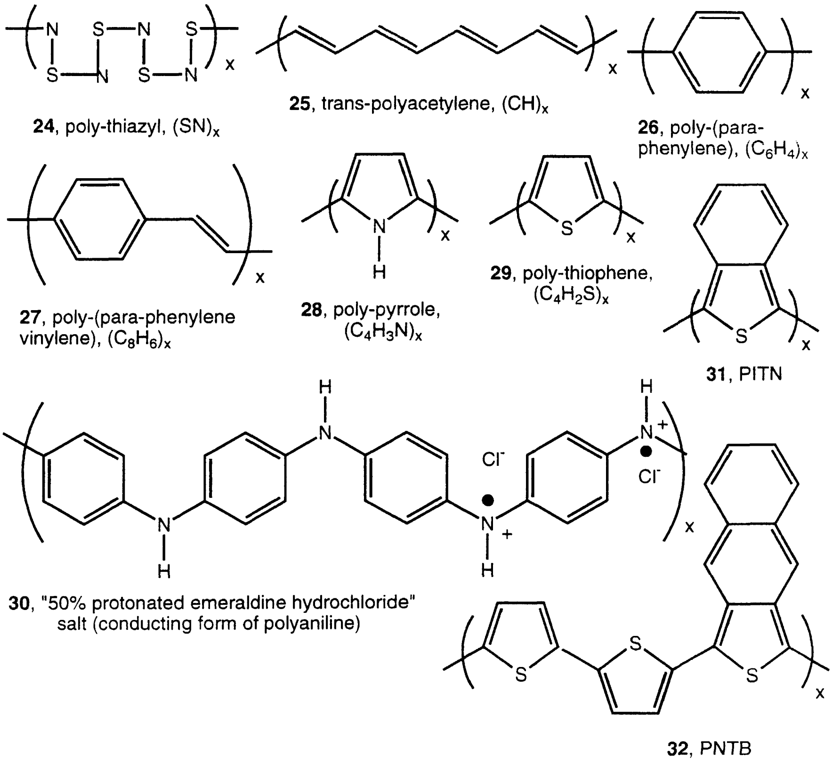

24, poly-thiazyl, $(SN)_x$

25, trans-polyacetylene, $(CH)_x$

26, poly-(para-phenylene), $(C_6H_4)_x$

27, poly-(para-phenylene vinylene), $(C_8H_6)_x$

28, poly-pyrrole, $(C_4H_3N)_x$

29, poly-thiophene, $(C_4H_2S)_x$

31, PITN

30, "50% protonated emeraldine hydrochloride" salt (conducting form of polyaniline)

32, PNTB

FIG. 14. Molecular structure of conducting polymers (**24–32**).

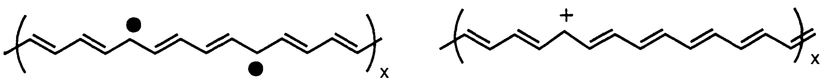

33, Neutral soliton-anti-soliton pair on $(CH)_x$

34, Positive polaron on I-doped $(CH)_x$

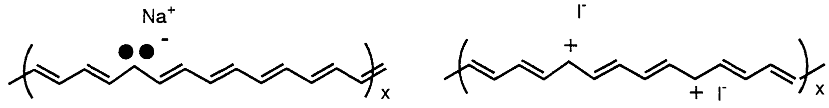

35, Negative polaron on Na-doped $(CH)_x$

36, Positive bipolaron on heavily I-doped $(CH)_x$

FIG. 15. Depiction of solitons (**33**), polarons (**34–35**), and bipolarons (**36**) in doped and undoped polyacetylene.

larons (**34, 35**). These solitons are breaks in the conjugation, which occur in pairs (soliton–antisoliton pairs), which can be treated theoretically, and whose experimental existence was proven.

For most (semi)conducting polymers, the "effective conjugation length" L [Eq. (26)] can be estimated from the optical absorption spectrum. For all known conducting polymers, except for graphite and for $(SN)_x$, there is a finite optical gap $E_{\max}$, with contributions from both the σ skeleton and the π electrons. The optical gap is usually close to the band gap: It is assumed that the charge carriers (electrons or holes) have higher mobility in the excited state than in the ground state. The undoped-polymer band gaps range from $E_{\max}$ = 4.7 eV for polyacetylene, to 3.8 eV for the emeraldine base form of polyaniline, 3.6 eV for poly-paraphenylene, 3.0 eV for polypyrrole, and 2.58 eV for polythiophene. The single- to double-bond length alternation, which indicates a nonzero $E_{\max}$, is conceptually related to a Peierls gap, and is of the order of 0.08–0.10 Å in polyacetylene.

The band gaps $E_{\max}$ are lowered dramatically after doping with electron donors (sodium, potassium, etc.) or electron acceptors (iodine, ferric chloride, arsenic pentafluoride, nitrosonium hexafluorophosphate, etc.); most conducting polymers are dark or black in their conducting state.

For conducting polymers, the monomer site M must accommodate at least one electron (M^-) or hole (M^+), thus forming a mobile polaron (**34,35**). A second electron or hole can easily go onto a neighboring site, thus allowing for a new polaron state. If this second electron or hole goes onto the same monomer site, forming M^{++} or M^{--}, then a bipolaron is formed (**36**), which tends to localize the electron transfer and freeze the transport. In conducting polymers, however, the mobility of the polaron is limited by the need to accommodate the movement of the counterion, which is typically a small inorganic ion loosely associated with the polaron site, which must move with the moving charge. This is why conducting polymers have, typically, low mobilities.

Most conducting polymers have room-temperature conductivities σ that vary from 10^{-3} to 10^3 S cm^{-1}; the values depend strongly on the details of the processing. For extremely uniform polyacetylene $\sigma = 5 \times 10^5$ cm^{-1} has been achieved (Naarmann and Theophilou, 1987). For a stereoregular alkylated polythiophene conductivities up to 5.5×10^3 S cm^{-1} have been reported (McCullough and Williams, 1993).

The conductivity is almost always thermally activated—i.e., semiconducting, not metallic. It is thought that the activation energy ΔE [Eq. (15)] is mostly due to interchain hopping barriers. The early conducting polymers were all insoluble, and so molecular-weight determinations and processability were not feasible. With the introduction of solubilizing side chains, the solubility and processability increased dramatically, but at the price of the conductivity, which was presumably lowered by the increased interchain distance. The molecular weights of typical soluble polymers range between 10^3 and 10^4 g mol^{-1}.

13. CONDUCTING LANGMUIR–BLODGETT FILMS

Irving Langmuir and Kathleen Blodgett showed in the 1930s that one can transfer monomolecular layers of amphiphilic molecules (such as cadmium arachidate) from the air–water interface onto solid supports, molecular layer by molecular layer. These LB films are thus very nice, thin, two-dimensional crystalline materials (Roberts, 1990). Large anisotropic conductivities in LB films were demonstrated when Barraud and coworkers showed that a multilayer film of *N*-docosylpyridium TCNQ, deposited on an inert substrate, when doped with iodine vapor, reorganized and exhibited in-plane conductivities of 0.1 S cm^{-1}, and out-of plane conductivities 10 orders of magnitude smaller (Ruaudel-Teixier *et al.*, 1985). Most conducting LB films have activated conductivity, but some are metallic. Recently, Metzger and co-workers showed that K-doped 50-layer LB films of C_{60} superconduct below about 8 K, about 10 K lower than the bulk crystal (Wang *et al.*, 1993).

14. DESIGN RULES FOR CRYSTALLINE ORGANIC METALS AND SUPERCONDUCTORS

Having presented a summary of the recent advances in several areas, we attempt to codify design rules for newer and better conductors and superconductors.

Several semiempirical rules have developed (Cowan, 1989; Cowan *et al.*, 1990) to find new high-conductivity compounds and superconductors. These are summarized and updated as the following five requirements.

14.1 Molecular Requirements

The molecular constituents (electron donors D *and/or electron acceptors* A*) must have stable open-shell (free-radical) states* D^{+} *or* A^{-}*, and stable dication* (D^{++}) *and dianion* (A^{--}) *states.*

For any candidate molecule, one must consider the relative location of the highest occupied molecular orbital (HOMO) of the neutral donor (D) and the lowest unoccupied molecular orbital (LUMO) of the neutral acceptor (A) molecule. The "parent" neutral molecule and the "daughter" anion radical A^{-} or cation radical D^{+} must be both thermodynamically and kinetically stable, and accessible (not too high in energy). Since organic chemists must (or prefer) to work in solution, the species should not oxidize or reduce the solvent. In addition, the second ionization state (D^{++} or A^{--}) must also be stable, since an electron or a hole must also be given a chance to reside on the same, already ionized center. The most convenient experimental test of this stability of radicals is the measurement of the solution cyclic voltammogram. Many relatively stable organic free radicals are known. There are, at present, four methods useful for stabilizing these systems:

1. Force high spin density onto heteroatoms; this works well for TTF and TCNQ.
2. Use the steric hindrance of substituents, so that radical species are stable once made, as in diphenylpicrylhydrazyl, or in nitroxide spin labels.
3. Use push–pull substituent effects.
4. Use derivatives of odd-alternant aromatic hydrocarbons.

Of these four methods, only **1** has been seriously exploited for organic metals; methods **2**, **3**, and **4** do not necessarily stabilize the dication/dianion states.

A representative list of donors and acceptors is given in Figs. 2, 3, and 8. The synthetic variability of these structures is great, but the donors and acceptors must be stable and soluble in reasonable solvents if crystals are to be grown, and so the molecular sizes cannot be increased without limit. Adding solubilizing groups (alkyl chains) to an otherwise insoluble core molecule can help, but often these groups will prevent the efficient packing of molecules in crystals.

14.2 Crystallographic Requirements

Planar molecules D *or* A *with delocalized* π *molecular orbitals are best. Molecular components should be of appropriate or compatible size. One-dimensional metals need segregated stacks of radicals, and not mixed stacks.*

The planarity requirement for the donors and the acceptors is a logical inheritance from the charge-transfer complexes, and caused by the need for reasonable overlap between adjacent centers, to facilitate larger Mulliken transfer integrals t [Eq. (6)], and therefore good electron transport. However, compact molecules or clusters, which have small van der Waals distances between them and large overlaps, may have a moderate amount of intramolecular nonplanarity.

The requirement that components (electron donors D and electron acceptors A) be of compatible size is a crystal-packing requirement. The *horror vacui* that Kitaigorodskii (1973) tried to put on a computational basis in crystal-packing calculations means that, with the best-designed donors and acceptors, if they are of very different sizes, good compact structures with charge transfer are not achieved. This is why there are few interesting salts with TCNE, and so many with TCNQ: TCNE is too small an electron acceptor, compared with most donors, while hundreds of salts are known with TCNQ.

To make good electron acceptors, electron-rich substituent groups like $-SF_5$ or $-SeF_5$ could be attached to molecules, but these bulky groups will usually impede efficient crystal packing. This restricts strong electron acceptors to cyanocarbons; thus there are comparatively few good electron acceptors.

The issue of the "right" organic lattice is important. At one extreme, with very weak donors (high I_D) and very weak acceptors (low A_A), one has slightly colored crystals with so-called "mixed stacks" of alternating D and A molecules spaced at van der Waals distances apart (3.5 Å or higher), forming infinite "*DADA*" stacks, where the π electrons of D overlap with the π electrons of A; these have low ρ values. At the other extreme, with very strong donors (low I_D) and strong acceptors (high A_A), the tendency is to form a so-called mixed-stacking lattice, the organic chemical equivalent of sodium chloride, i.e., a $\rho \approx 1$ lattice with D^+ atop A^-; these crystals with $D^+A^-D^+A^-$ stacking are very deeply colored or shiny black, paramagnetic, and semiconducting.

Torrance pointed out that the segregated stacking that seems essential for high conductivity can be achieved by combining donors of intermediate I_D and acceptors of intermediate to high A_A (Torrance, 1979). Thus, TMPD TCNQ has a mixed stack (I_D = 6.25 eV for TMPD), while TTF TCNQ has a segregated stack (I_D = 6.83 for TTF). Since gas-phase electron affinities A_A are particularly difficult to measure, and gas-phase ionization potentials I_D are not routinely measured, one can seek guidance and solace from solution half-wave reduction potentials $E_{1/2}$, measured typically by cyclic voltammetry. Torrance, and also Saito and Ferraris (1980), have given rules of thumb for which are promising combinations of $E_{1/2}$ values. The interface between neutral and ionic mixed-stack *DA* complexes has been found: The 1:1 crystal complex of TTF (**7a**) with chloranil (**23**) was a *DADA* "neutral" ($\rho \approx$ 0.1) crystal at room temperature, which undergoes a collective transition to an ionic state ($\rho \approx 0.4$) at either low temperature or high pressure (Torrance *et al.*, 1981).

Black mixed-stack crystals and black segregated-stack crystals cannot be told apart by the naked eye, but only by electrical measurements or a solved crystal structure.

An important distinction to be made is between the two-stack systems (like TTF TCNQ, where electrons can migrate along the TCNQ stacks, and holes can migrate along the TTF stacks) and the one-stack systems, or ion-radical salts (or radical-ion salts), such as the $(TMTSF)_2X$ salts, the alkali TCNQ salts M_n(TCNQ), and the (BEDT-TTF)$_2X$ salts: The holes are localized on the TMTSF or BEDT-TTF sublattice, and the electrons on the TCNQ sublattice, and the rest of the crystal consists of diamagnetic, passive counterions.

However, this business of good stacking should not be exaggerated: The 12.5-K superconductor κ-(BEDT-TTF)$_2$Cu[N(CN)$_2$]Cl has a layer of donors, which form van der Waals "dimers" of two BEDT-TTF species with charge $\rho = \frac{1}{2}$ each, with good intradimer overlap, but bad interdimer overlap, and yet the system is superconducting!

Finally, the finite crystal size is a problem: Typical organic conductors and superconductors form needle-shaped crystals of reasonable length (0.5 to 2 cm) but very small widths (usually 0.1 to 1 mm, or even as small as 0.025 mm).

14.3 Band Requirements

The HOMO–LUMO gap should be as small as possible, and the bandwidths should be as large as possible. This can be understood in terms of band theory, and of the Hubbard Hamiltonian, Eq. (10). The width of an energy band is $4t$; the effective repulsion of electrons within the same band is U_{eff} (akin to U, but with screening): One wants $U_{\text{eff}} < 4t$, and an incompletely filled band. Typical U_{eff} are 1 to 4 eV, while t is typically only

0.1–0.25 eV. U_{eff} can be decreased by increasing the molecular polarizability, by introducing atoms with many electrons (Se, Te, I) into the structure. One could decrease the HOMO–LUMO gap, while increasing the bandwidth, so as to force the LUMO- and HOMO-derived bands to overlap directly: This would yield a semimetallic regime, as for bismuth, but has not yet been realized for organic metals. The "tunability" of molecular orbitals within a certain class of related structures is finite, and the bandwidth of these systems remains relatively small.

14.4 Solid-State Electronic Requirements

Mixed valence, or fractional charge transfer ρ, *should be achieved. Inhomogeneous charge and spin distribution are preferable. A Peierls distortion should be avoided, by hydrostatic pressure, if necessary, or by efficient interchain interactions. Disorder should be avoided in the crystal.*

It was quite clear that intermediate values of the ionicity ρ were possible, indeed necessary for metallic conduction, within "segregated-stack" structures (D atop D, etc., and A atop A). One must achieve partial charge transfer (certainly $0 < \rho < 1$, but so far not yet $1 < \rho < 2$), and these species must be mixed valent, i.e., if one of two sites is a cation and the other a neutral molecule, then there must be only one crystallographically distinct species (so that the Franck–Condon rearrangement necessary to move the electron from the cation to the neutral molecule is not large). The theoretical classification into neutral and ionic crystals [Eqs. (24), (25)] by comparing Madelung energies to the cost of ionization works well for mixed-stack crystals, but such calculations did not yield the desired energy minimum for the partial ionicity of mixed-valence crystals (Metzger, 1981), because higher-order polarization and ρ-dependent dispersion energies need to be included in the calculation. TTF TCNQ is thermodynamically stable (Metzger, 1977).

The intermolecular (interionic) distances must be regular. This "mixed valency" requires that there be only one crystallographically unique molecular site, which must share its partial valency with the nearest-neighbor sites along the stack. The many "complex" stoichiometry TCNQ salts, e.g., $Cs_2(TCNQ)_3^{--}$ or triethylammonium$(TCNQ)_2^-$, which exhibit "trimeric" or "tetrameric" units of several crystallographically distinct TCNQ molecules and $TCNQ^-$ anions held at van der Waals separations, cannot conduct well.

14.5 Electron–Phonon Coupling Requirement

The electrons must couple with the crystal phonons so as to enable the right magnitude of coupling for Cooper–pair formation.

This requirement is the most mysterious. It is not known why certain crystals [e.g., HMTSF TCNQ or $Cu(2,5\text{-DM-DCNQI})_2$] conduct very well at low temperatures, but do not form Cooper pairs. One can only speculate that certain intramolecular or intermolecular vibrations or rigid-body librational modes must be "right" for superconductivity. Kuzmany (1993) has suggested that strained molecules, e.g., C_{60}, have intramolecular Raman modes that may somehow be very important for superconductivity.

15. DESIGN RULES FOR POLYMERIC ORGANIC CONDUCTORS

Nobody knows how far one can push the conductivity by careful processing, aligning the polymer strands, and removing defects. The tremendous advantage of conducting polymers, in comparison with crystals of organic conductors or superconductors, is that larger samples can be made, at the price of residual catalyst impurities, structural defects, and dependence of physical properties on the processing methods.

It has been suggested (Heeger, 1990) that in a "dirty polymer" such as 10% doped polyacetylene, the charge-carrier concentration can be $N = 4 \times 10^{21}$ cm^{-3}, and the mobility $\mu < 0.2$ cm^2 V^{-1} s^{-1}, whence by Eq. (16) $\sigma < 30$ S cm^{-1}. By increasing the mean free path of charge carriers from about one lattice constant to several hundred lattice constants, the theoretical σ can be increased to hundreds of S/cm. In the "clean limit" of conductivity limited only by phonon scattering, $\sigma = 2 \times 10^6$ S cm^{-1}, i.e., four times higher than that of Cu, was predicted (Heeger, 1991).

Five design rules emerge for metallic conducting polymers.

15.1 Small Band Gaps

Reduce the band gaps by condensing large aromatic rings onto the backbone. The optical band gap $E_{\max}$ is reduced if there is a large quinoid contribution (see Fig. 16) to the ground state of the conjugated polymer. This has been accomplished in part in **31** and **32**: A band gap of 0.65 eV has been found for **32** (Lakshmikantham *et al.*, 1993).

15.2 Fiber Alignment

Stress-align the polymer strands. This is a traditional technique in fibrous polymers, and has given excellent results for polyacetylene (Naarmann and Theophilou, 1987). It turns out that for polyacetylene the chain lengths are probably quite long (100 or so C atoms), while for polyaniline and polythiophene 10 to 20 oligomers may be the individual chain length.

15.3 Polymer Solubilization

Increase the polymer solubility by adding long-chain alkyl substituents to the monomer. This brings one gain (solubility and processability of the polymer), but usually with two drawbacks (decreased effective conjugation length because of increased rotational disorder, and decreased interchain overlap). But for the newer soluble polymers one can determine molecular weights, viscosity, and other polymer chemical parameters unavailable, e.g., for polyacetylene.

15.4 Stereoregular Polymerization

Chemically protect all "unwanted" side positions of monomer (McCullough and Lowe, 1992; McCullough and Williams, 1993). This eliminates undesirable branched polymerization at unwanted side positions. For instance, thiophene is naively thought to polymerize preferentially at the α (or 2,2′) positions, but in reality a considerable amount of polymerization at the β (or 3,3′) positions occurs. If one chemically blocks the β positions in thiophene by a bulky substituent, then the polymerization is forced to occur only at the α positions, and to become stereoregular, and the conductivities increase dramatically (McCullough and Lowe, 1992; McCullough and Williams, 1993).

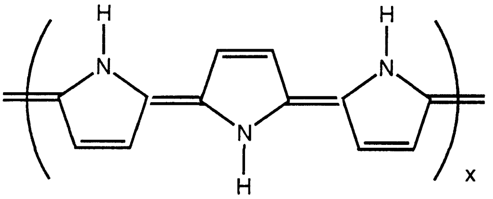

Quinoid state of poly-pyrrole

FIG. 16. Structure of the quinoid state for polypyrrole.

15.5 Facilitate Interchain Hopping

Reduce the interchain barrier to hopping by bringing the chains closer together. The conductivity in most polymers is thermally activated, and limited by hopping between conducting strands: The macroscopic behavior is semiconductive, not metallic. If interstrand communication can be improved, maybe metallic conductivity will be achieved. This requires careful molecular modeling, to ensure good packing by exploiting van der Waals forces.

Design rules for superconductivity in inorganic conducting polymers are somewhat premature: So far, $(SN)_x$ remains the only polymeric superconductor. It is likely that the monomer needs to have two stable oxidation or reduction states (see design rule 1 for crystals, above): Then bipolaronic insulating states are neatly avoided.

16. APPLICATIONS

A huge commercial success was registered when International Business Machines Corp. demonstrated in 1964 that a charge-transfer polymeric complex between trinitrofluorenone and polyvinylcarbazole could replace the high-cost selenium disk as the electrostatic imaging surface in electrostatic copiers. The early promise of lightweight batteries based on polyacetylene encountered insurmountable shelf-life problems, because of the intrinsic instability of the polymer. Conducting polymers have been used as electromagnetic shielding, and may have found applications, shrouded in military secrecy, for the wing surfaces in "stealth" bombers. More pedestrian applications, such as thermochromic temperature dating of stored foods in supermarkets and secondary batteries based on polyaniline, have entered the marketplace.

17. CONCLUSION

The fascinating world of organic conductors, superconductors, and polymers has been shown to exhibit not only interesting fundamental lower-dimensional physics, but also some promising commercial applications.

ACKNOWLEDGMENT

This work was supported in part by USAFOSR Grant No. F49620-92-J-0529DEF.

GLOSSARY

Bandwidth: Energy spacing between lowest level in a band and the highest level in the same band.

Bechgaard Salts: The TMTSF (tetramethyltetraselenafulvalene, **8b**) salts $TMTSF_2X$, where X is an inorganic anion; historically the first organic superconductors.

Bipolaron: Pseudoparticle that denotes two close dipoles in a lattice, and that has a diffusive coherent motion in the lattice.

Charge-Density Wave (CDW): Deformation of charge density in a crystal lattice from its equilibrium distribution (which is periodic with the lattice). Can be dynamic or static.

Charge Transfer ρ: Estimate of contribution of ionic state ($\rho = 1$) that replaces, partially or totally, the neutral state ($\rho = 0$). Can be calculated indirectly from $2k_F$ and $4k_F$ CDW x-ray scattering [Eq. (13)], from infrared spectral shifts, or from chemical shifts in nuclear quadrupole resonance spectra, x-ray photoelectron spectra, or (rarely) nuclear magnetic-resonance spectra.

Conduction Band: Energy band that is partially filled, because there are more states available than electrons (or electron pairs) to fill them; the top of the occupied levels is the Fermi level. The electrons at the top of the conduction band are relatively mobile, and can contribute to the electrical conductivity of the crystal.

Cooper Pairing: Pairing of electron states of opposite spin and opposite momentum in a crystal, mediated by coupling of both electrons to a particular lattice phonon.

Counterion: An anion or cation that is not involved in the process of interest (here, conductivity or superconductivity), but is needed as "counterbalance" to achieve electrical neutrality for the stoichiometric unit of a salt.

Effective Conjugation Length: Equivalent length of the π-electron network in a linear polyene, used in a particle-in-a-box model to calculate its lowest-energy optical absorption.

Energy Band: Set of energies for a set of N identical particles (atoms or molecules) that are very closely spaced, because their N-fold degeneracy is lifted by the interparticle interactions, e.g., for molecules in a crystalline solid.

Fermi Energy: Highest occupied energy band in the ground state of a free-electron gas at $T = 0$ kelvin.

Fermi Level: Chemical potential, or partial molar Gibbs free energy, for the electrons in a solid at finite temperatures.

Fermi Wave Vector: Wave vector for electron in the crystal with highest crystal momentum.

Highest Occupied Molecular Orbital (HOMO): Molecular wave function associated with the highest energy level that is occupied by an electron (or electron pair) in a molecule.

Hubbard Hamiltonian: Phenomenological Hamiltonian for N sites in an ordered array (atoms or molecules in a crystal or other lattice), which includes the Mulliken transfer integral t for electron transfer between sites, and the on-site energy U for electron–electron repulsion on the same site.

Lowest Unoccupied Molecular Orbital (LUMO): Molecular wave function associated with the lowest energy level that is not occupied by an electron in a molecule.

Mulliken Charge-Transfer Complex: Weak intermolecular complex, in solution or in the solid state, which mixes the ionic and neutral states of two components, so as to preserve the optical spectrum of these components, yet allow a *strong optical transition* between the mostly neutral and mostly ionic states.

Mulliken Transfer Integral: Energy involved in moving an electron from one site in a crystal to a nearest-neighbor site.

One-Electron Acceptor: A molecule with relatively high vacuum electron affinity (2 to 3.5 eV), which easily accepts an electron (gets reduced).

One-Electron Donor: A molecule with relatively low vacuum ionization energy (5 to 7 eV), which easily donates an electron (gets oxidized).

Peierls Instability: Structural instability of the one-dimensional uniform lattice toward distortion.

Peierls Transition: Transition between uniform high-temperature lattice and CDW-distorted low-temperature lattice; often separates high-temperature metallic regime from low-temperature semiconducting regime.

π Electron: Electron in a molecular orbital with *p* (perpendicular) symmetry and a nodal plane in the internuclear region between atoms.

Polaron: Pseudoparticle that denotes a dipole in a lattice, and that has diffusive coherent motion in the lattice.

Quinoid: State of a conjugated system with localized double bonds, as in benzoquinone, rather than delocalized double bonds, as in benzene (opposite: benzenoid).

σ Electron: Electron in a molecular orbital with *s* (longitudinal) symmetry and finite amplitude along the internuclear axis between atoms.

Soliton: Pseudoparticle representing a localized structural distortion that can diffuse coherently in a lattice.

Spin-Density Wave (SDW): Deformation of spin density in a crystal lattice from its equilibrium distribution (which has the periodicity of the lattice). Can be dynamic or static.

Spin-Peierls Transition: Transition between uniform paramagnet at higher temperature and distorted (lightly dimerized) diamagnet at lower temperature.

Superconductor: Solid with infinite electrical conductivity and diamagnetic flux exclusion.

Tight-Binding Approximation: Assumption that, at a lattice point or its immediate neighborhood, the crystal wave functions can be approximated as the local wave functions of the atom or molecule at that lattice point.

Works Cited

Acker, D. S., Harder, R. J., Hertler, W. R., Mahler, W., Melby, L. R., Benson, R. E., Mochel, W. E. (1960), *J. Am. Chem. Soc.* **82,** 6408–6409.

Bardeen, J., Cooper, L. N., Schrieffer, J. R. (1957a), *Phys. Rev.* **106,** 162–164.

Bardeen, J., Cooper, L. N., Schrieffer, J. R. (1957b), *Phys. Rev.* **108,** 1175–1204.

Bednorz, J. G., Müller, K. A. (1986), *Z. Phys. B* **64,** 189–193.

Bloch, F. (1928), *Z. Phys.* **52,** 555–600.

Bloch, A. N., Carruthers, T., Poehler, T. O., Cowan, D. O. (1977), in: H. J. Keller (Ed.), *Chemistry and Physics of One-Dimensional Metals,* NATO Advanced Study Institutes, Series B: Physics, Vol. 25, New York: Plenum, pp. 47–85.

Burroughes, J. H., Bradley, D. D. C., Brown, A. R., Marks, R. N., Mackay, K., Friend, R. H., Burns, P. L., Holmes, A. B. (1990), *Nature* **347,** 539–541.

Cairns, T. L., Carboni, R. A., Coffman, D. D., Engelhardt, V. A., Heckert, R. E., Little, E. L., McGeer, E. G., McKusick, B. C., Middleton, W. J. (1957), *J. Am. Chem. Soc.* **79,** 2340–2341.

Chu, C. W., Gao, L., Chen, F., Huang, Z. J., Meng, R. L., Xue, Y. Y. (1993), *Nature* **365,** 323–325.

Coleman, L. B., Cohen, M. J., Sandman, D. J., Yamagishi, F. G., Garito, A. F., Heeger, A. J. (1973), *Solid State Commun.* **12,** 1125–1132.

Conwell, E. (1980), *Phys. Rev. B* **22,** 1761–1780.

Coppens, P., Petricek, V., Levendis, D., Larsen, F. K., Paturle, A., Yan, G., LeGrand, A. D. (1987), *Phys. Rev. Lett.* **59,** 1695–1697.

Cowan, D. O. (1989), in: Z. Yoshida, T. Shiba, Y. Ohshiro (Eds.), *New Aspects of Organic Chemistry,* Tokyo: Kodansha.

Cowan, D. O., Fortkort, J. A., Metzger, R. M. (1990), in: R. M. Metzger, P. Day, G. C. Papavassiliou (Eds.), *Lower-Dimensional Systems and Molecular Electronics,* NATO Advanced Study Institutes, Series B: Physics, Vol. 248, New York: Plenum, pp. 1–22.

Day, P. (1990), in: R. M. Metzger, P. Day, G. C. Papavassiliou (Eds.), *Lower-Dimensional Systems and Molecular Electronics,* NATO Advanced Study Institutes, Series B: Physics, Vol. 248, New York: Plenum, pp. 115–128.

Dresselhaus, M. S., Dresselhaus, G., Eklund, P. C. (1993), *J. Mater. Res.* **8,** 2054–2097.

Ferraris, J., Cowan, D. O., Walatka, V. J., Perlstein, J. H. (1973), *J. Am. Chem. Soc.* **95,** 948–949.

Floquet, G. (1883), *Ann. Sci. Ecole Norm. Sup.* **12,** 47.

Gavaler, J. R. (1973), *Appl. Phys. Lett.* **23,** 480–482.

Greene, R. L., Street, G. B., Sutter, L. J. (1975), *Phys. Rev. Lett.* **34,** 577–579.

Hannay, N. B., Geballe, T. H., Matthias, B. T., Endres, K., Schmidt, P., Macnair, D. (1965), *Phys. Rev. Lett.* **14,** 225–226.

Hebard, A. F., Rosseinsky, M. J., Haddon, R. C., Murphy, D. W., Glarum, S. H., Palstra, T. T. M., Ramirez, A. P., Kortan, A. R. (1991), *Nature* **350,** 600–601.

Heeger, A. J. (1990), in: R. M. Metzger, P. Day, G. C. Papavassiliou (Eds.), *Lower-Dimensional Systems and Molecular Electronics,* NATO Advanced Study Institutes, Series B: Physics, Vol. 248, New York: Plenum, pp. 293–301.

Hubbard, J. (1963a), *Proc. R. Soc. London A* **276,** 238–257.

Hubbard, J. (1963b), *Proc. R. Soc. London A* **277,** 237–259.

Hubbard, J. (1964), *Proc. R. Soc. London A* **281,** 401–419.

Hubbard, J. (1978), *Phys. Rev. B* **17,** 494–505.

Jérome, D. (1990), in: R. M. Metzger, P. Day, G. C. Papavassiliou (Eds.), *Lower-Dimensional Systems and Molecular Electronics,* NATO Advanced Study Institutes, Series B: Physics, Vol. 248, New York: Plenum, pp. 85–89.

Jérome, D., Mazaud, A., Ribault, M., Bechgaard, K. (1980), *J. Phys. (Paris) Lett.* **41,** L95–L98.

Kistenmacher, T. J., Phillips, T., Cowan, D. O. (1974), *Acta Crystallogr. B* **30,** 763–768.

Kitaigorodsky, A. I. (1973), *Molecular Crystals and Molecules,* New York: Academic.

Krätschmer, W., Lamb, L. D., Fostiropoulos, K., Huffman, D. R. (1990), *Nature* **347,** 354–358.

Krogmann, K. (1969), *Angew. Chem. Int. Ed. Engl.* **8,** 35–42.

Kroto, H. W., Heath, J. R., O'Brien, S. C., Curl, R. F., Smalley, R. E. (1985), *Nature* **318,** 162–163.

Kuhn, H. (1949), *J. Chem. Phys.* **17,** 1198–1212.

Kuzmany, H. (1993), private communication.

Lakshmikantham, M. V., Lorcy, D., Scordilis-Kelley, C., Wu, X.-L., Parakka, J. P., Metzger, R. M., Cava, M. P. (1993), *Adv. Mater.* **5,** 723–726.

Landau, L. D., Lifschitz, L. M. (1958), *Statistical Physics,* Vol. 5 of *Course of Theoretical Physics,* London: Pergamon, p. 482.

Little, W. A. (1964), *Phys. Rev. A* **134,** 1416–1424.

Maeda, H., Tanaka, Y., Fukutomi, M., Asano, T. (1988), *Jpn. J. Appl. Phys. Lett.* **27,** L207–L210.

McConnell, H. M., Hoffman, B. M., Metzger, R. M. (1965), *Proc. Natl. Acad. Sci. U.S.A.* **53,** 46–50.

McCullough, R. D., Lowe, R. D. (1992), *J. Chem. Soc. Chem. Commun.,* 70–72.

McCullough, R. D., Williams, S. P. (1993), *J. Am. Chem. Soc.* **115,** 11608–11609.

Metzger, R. M. (1977), *J. Chem. Phys.* **66,** 2525–2533.

Metzger, R. M. (Ed.) (1981), *Crystal Cohesion and Conformational Energies,* Springer Topics in Current Physics Vol. 26, Berlin: Springer.

Mizuno, M., Garito, A. F., Cava, M. P. (1978), *J. Chem. Soc. Chem. Commun.,* 18–19.

Mulliken, R. S. (1950), *J. Am. Chem. Soc.* **72,** 600–608.

Naarmann, H., Theophilou, N. (1987), *Synth. Met.* **22,** 1–8.

Oshima, K., Mori, T., Inokuchi, H., Urayama, H., Yamochi, H., Saito, G. (1988), *Phys. Rev. B* **38,** 938–941.

Palstra, T. T. M., Zhou, O., Iwasa, Y., Sulewski, T. E., Fleming, R. M., Zagorski, B. R. (1995) in: L. Y. Chiang, P. Bernier, D. S. Bethune, T. W. Ebbesen, R. M. Metzger, J. W. Mintmire (Eds.), *Science and Technology of Fullerene Materials,* Materials Research Society Symposium Proceedings Vol. 359, Pittsburgh, PA: Materials Research Society, pp. 285–288.

Parkin, S. S. P., Engler, E. M., Schumaker, R. R., Lagier, R., Lee, V. Y., Scott, J. C., Greene, R. L. (1983), *Phys. Rev. Lett.* **50,** 270–273.

Parkin, S. S. P., Lee, V. Y., Engler, E. M., Nazzal, A. I., Huang, T. C., Gorman, G., Savoy, R., Beyers, R. (1988), *Phys. Rev. Lett.* **60,** 2539–2542.

Peierls, R. E. (1955), *Quantum Theory of Solids,* Oxford: Clarendon, p. 108.

Roberts, G. (Ed.) (1990), *Langmuir-Blodgett Films,* New York: Plenum.

Ruaudel-Teixier, A., Vandevyver, M., Barraud, A. (1985), *Mol. Cryst. Liq. Cryst.* **120,** 319–322.

Saito, G., Ferraris, J. P. (1980), *Bull. Chem. Soc. Jpn.* **53,** 2141–2145.

Saito, G., Enoki, T., Inokuchi, H. (1982), *Chem. Lett.,* 1345–1348.

Saito, G. (1990), in: R. M. Metzger, P. Day, G. C. Papavassiliou (Eds.), *Lower-Dimensional Systems and Molecular Electronics,* NATO Advanced Study Institutes, Series B: Physics, Vol. 248, New York: Plenum, pp. 67–84.

Shirakawa, H., Ikeda, S. (1971), *Polym. J.* **2,** 231–244.

Shirakawa, H., Louis, E. J., MacDiarmid, A. G., Chiang, C. K., Heeger, A. J. (1977), *J. Chem. Soc. Chem. Commun.,* 578–580.

Skotheim, T. A. (Ed.) (1986), *Handbook of Conducting Polymers,* Vols. 1, 2, New York: Dekker.

Thomas, D. D., Keller, H., McConnell, H. M. (1963), *J. Chem. Phys.* **39,** 2321–2329.

Thomas, G. A., Schafer, D. E., Wudl, F., Horn, P. M., Rimai, D., Cook, J. W., Glocker, D. A., Skove, M. J., Chu, C. W., Groff, R. P., Gillson, J. L., Wheland, R. C., Melby, L. R., Salamon, M. G., Craven, R. A., DePasquali, G., Bloch, A. N., Cowan, D. O., Walatka, V. V., Pyle, R. E., Gemmer, R., Poehler, T. O., Johnson, G. R., Miles, M. G., Wilson, J. D., Ferraris, J. P., Finnegan, T. F., Warmack, R. J., Raaen, V. F., Jérome, D. (1976), *Phys. Rev. B* **13,** 5105–5100.

Thorup, N., Ringdorf, G., Soling, H., Bechgaard, K. (1981), *Acta Crystallogr. B* **37,** 1236–1240.

Torrance, J. B. (1979), *Acc. Chem. Res.* **12,** 79–86.

Torrance, J. B., Girlando, A., Mayerle, J. J., Crowley, J. I., Lee, V. Y., Batail, P., LaPlaca, S. J. (1981), *Phys. Rev. Lett.* **47,** 1747–1750.

Van Vleck, J. H. (1932), *The Theory of Electric and Magnetic Susceptibilities,* Oxford, U.K.: Oxford Univ. Press.

Wang, P., Metzger, R. M., Bandow, S., Maruyama, Y. (1993), *J. Phys. Chem.* **97,** 2926–2927.

Williams, J. M., Ferraro, J. R., Thorn, R. J., Carlson, K. D., Geiser, U., Wang, H. H., Kini, A. M., Whangbo, M.-H. (1992), *Organic Superconductors (Including Fullerenes) Synthesis, Structure, Properties, and Theory,* Englewood Cliffs, NJ: Prentice-Hall.

Wu, M. K., Ashburn, J. R., Torng, C. J., Hor, P. H., Meng, R. L., Gao, L., Huang, Z. J., Wang, Y. Q., Chu, C. W. (1987), *Phys. Rev. Lett.* **58,** 908–910.

Wudl, F., Smith, G. M., Hufnagel, E. J. (1970), *J. Chem. Soc. Chem. Commun.,* 1453–1454.

Further Reading

The interested reader can glean more detail from the several thousand research articles since the 1970s and from the following representative list of references:

Review Articles

Bryce, M. R., Murphy, L. C. (1984), *Nature* **309,** 119–126.

Garito, A. F., Heeger, A. J. (1974), *Acc. Chem. Res.* **7,** 232–240.

Greene, R. L., Street, G. B. (1984), *Science* **226,** 651–656.

Heeger, A. J., Kivelson, S., Schrieffer, J. R., Su, W. P. (1988), *Rev. Mod. Phys.* **60,** 781–850.

Kommandeur, J. (1985), *Nouv. J. Chim.* **9,** 341–345.

Miller, J. S., Epstein, A. J. (1987), *Angew. Chem. Int. Ed. Engl.* **26,** 287–293.

Roth, S., Bleier, H. (1985), *Adv. Phys.* **36,** 385–462.

Schukat, G., Fanghänel, E. (1987), *Sulfur Rep.* **14,** 245–300.

Sólyóm, L. (1979), *Adv. Phys.* **28,** 201–303.

Soos, Z. G. (1983), *Isr. J. Chem.* **23,** 37–48.

Williams, J. M., Schultz, A. J., Geiser, U., Carlson, K. D., Kini, A. M., Wang, H. H., Kwok, W.-K., Whangbo, M.-H., Schirber, J. E. (1991), *Science* **252,** 1501–1508.

Wudl, F. (1984), *Acc. Chem. Res.* **17,** 227–232.

Edited Topical Volumes and Monographs

Billups, W. E., Ciufolini, M. A. (Eds.) (1993), *Buckminsterfullerenes,* New York: VCH.

Briegleb, G. (1960), *Elektronen-Donator-Acceptor Systeme,* Berlin: Springer.

Conwell, E. M. (Ed.) (1988), *Highly Conducting Quasi-One-Dimensional Organic Crystals,* Boston: Academic.

Devreese, J. T., Evrard, R. P., Van Doren, V. E. (Eds.) (1977), *Highly Conducting One-Dimensional Solids,* New York: Plenum.

Ferraro, J. R., Williams, J. M. (1987), *Introduction to Synthetic Electrical Conductors,* San Diego: Academic.

Lieb, E. H., Mattis, D. C. (1964), *Mathematical Physics in One Dimension,* New York: Academic.

Miller, J. S. (Ed.) (1981, 1982, 1983), *Extended Linear-Chain Compounds,* Vols. 1, 2, 3, New York: Plenum.

Monceau, P. (Ed.) (1985), *Electronic Properties of Inorganic Quasi-One-Dimensional Compounds,* Vols. 1, 2, Dordrecht, Holland: Reidel.

Symposium Proceedings:

Ann. N.Y. Acad. Sci. **313** (1978); *Chem. Scripta* **17** (1981); *Mol. Cryst. Liq. Cryst.* **77, 79, 81, 83, 85, 86** (1982); *J. Phys. (Paris) Coll.* **44 C3** (1983); *Mol. Cryst. Liq. Cryst.* **117–120** (1985); *Phys. B* **143** (1986); *Synth. Met.* **17–19** (1987), **27–29** (1989), **41–43** (1991), **57–59** (1993), **69–71** (1995); *Springer Lect. Notes in Phys.* **65** (1977), **95–96** (1979), *Springer Proc. in Phys.* **51** (1990); *Springer Ser. in Solid State Sci.* **88** (1990), **117** (1993); *Mater. Res. Soc. Symp. Proc.* **173** (1990), **247** (1992); *Am. Chem. Soc. Symp. Ser.* **5** (1974), **481** (1992).

Hatfield, W. E. (Ed.) (1979), *Molecular Metals,* New York: Plenum.

Kresin, V., Little, W. A. (Eds.) (1990), *Organic Superconductivity,* New York: Plenum.

Masuda, K., Silver, M. (Eds.) (1974), *Energy and Charge Transfer in Organic Semiconductors,* New York: Plenum.

Ovchinnikov, A. A., Ukrainskii, I. I. (Eds.) (1991), *Low-Dimensional Conductors and Superconductors,* Berlin: Springer.

North American Treaty Organization Advanced Study Institutes (NATO ASI)

Alcácer, L. (Ed.) (1980), *The Physics and Chemistry of Low Dimensional Solids,* NATO ASI Series C: Chemistry, Vol. 56, Dordrecht, Holland: Reidel.

Beeby, J. L. (Ed.) (1991), *Condensed Systems of Low Dimensionality,* NATO ASI Series B: Physics, Vol. 253, New York: Plenum.

Brédas, J.-L. (Ed.) (1991), *Conjugated Polymeric Materials: Opportunities in Electronics, Optoelectronics and Molecular Electronics,* NATO ASI Series E: Materials Science, Vol. 182, Dordrecht, Holland: Kluwer.

Delhaès, P., Drillon, M. (Eds.) (1988), *Organic and Inorganic Low-Dimensional Crystalline Materials,* NATO ASI Series B: Physics, Vol. 168, New York: Plenum.

Jérome, D., Caron, L. G. (Eds.) (1987), *Low-Dimensional Conductors and Superconductors,* NATO ASI Series B: Physics, Vol. 155, New York: Plenum.

Keller, H. J. (Ed.) (1977), *Chemistry and Physics of One-Dimensional Metals,* NATO ASI Series B: Physics, Vol. 25, New York: Plenum.

Metzger, R. M., Day, P., Papavassiliou, G. C. (Eds.) (1990), *Lower-Dimensional Systems and Molecular Electronics,* NATO ASI Series B: Physics, Vol. 248, New York: Plenum.

For lack of space, this review did not discuss certain ancillary areas; the reader is therefore asked to read elsewhere about:

Nonlinear Optical Properties

Prasad, P. N., Williams, D. J. (1990), *Introduction to Nonlinear Optical Effects in Molecules and Polymers,* New York: Wiley.

The Promise of Molecular Electronic Devices

Ashwell, G. J. (Ed.) (1992), *Molecular Electronics,* Taunton, Somerset: Research Studies Press.

Aviram, A. (Ed.) (1992), *Molecular Electronics—Science and Technology,* New York: American Institute of Physics.

Birge, R. R. (Ed.) (1994), *Biomolecular Electronics, American Chemistry Society Advances in Chemistry Series,* Vol. 240, Washington, DC: American Chemical Society.

Carter, F. L. (Ed.) (1982, 1987, 1988), *Molecular Electronic Devices,* Vol. 1, New York: Dekker; Vol. 2, New York: Dekker; Vol. 3, Amsterdam: North-Holland.

Organic and Organometallic Ferromagnets

Gatteschi, D. (1994), *Adv. Mater.* **6,** 635–645.

Miller, J. S., Epstein, A. J. (1994), *Angew. Chem. Int. Ed. Engl.* **33,** 385–415.

SEMICONDUCTORS, AMORPHOUS—ELECTRONIC PROPERTIES

H. Fritzsche, *James Franck Institute and Department of Physics, University of Chicago, Chicago, Illinois, U.S.A.*

Introduction 242
1. **Electronic States** 243
1.1 Chemical Bonding 243
1.2 Extended and Localized States 243
1.3 Electronically Active Defects and Dangling Bonds 244
1.4 Positive- and Negative-*U* Defects 245
1.5 Thermodynamic Equilibrium of Defects 246
1.5.1 Defect Pool 246
1.5.2 Doping Equilibration 246
1.5.3 Valence-Alternation Defects 247
1.6 Inhomogeneities, Potential Fluctuations, Percolation 247
1.7 Determination of Electronic States 248
1.7.1 Optical Absorption 248
1.7.2 Photoemission Spectroscopies 249
1.7.3 Electron Spin Resonance (ESR) 250
1.7.4 Transient Emission Spectroscopies 251
2. **Electronic Transport** 252
2.1 Hopping Transport in Localized States 253
2.2 dc Conductivity in Extended States 254
2.2.1 Chalcogenide Glasses 254
2.2.2 Hydrogenated Amorphous Silicon 254
2.2.3 Amorphous In_2O_{3-x} 256
2.3 Hall Effect 257
2.4 Thermopower 257
2.5 Conductance Fluctuations in Hydrogenated Amorphous Silicon 258
3. **Photoexcitation and Recombination** 259
3.1 Radiative and Nonradiative Recombination 259
3.2 Geminate Recombination 260
3.3 Photoluminescence (PL) 260
3.3.1 Steady-State PL 260
3.3.2 Time-Resolved PL 261
3.4 Photoconductivity 261
3.4.1 Primary and Secondary Photoconductivity 261
3.4.2 Steady-State Photoconductivity 261
3.4.3 Photoconductivity Decay 262
3.4.4 Modulated Photoconductivity 263
3.5 Drift Mobility 263
3.5.1 Deep Trapping 264
3.6 Spin-Dependent Recombination 265
4. **Metastability** 266
4.1 Chalcogenide Glasses 266
4.1.1 Photostructural Changes 266
4.1.2 Metastable Defects 267
4.1.3 Optical Anisotropies 267
4.2 *a*-Si:H and Related Materials 267
4.2.1 Light-Induced Defects 267
4.2.2 Metastabilities Produced by Rapid Quenching 268
4.2.3 Metastabilities Produced by Space Charges 269
4.3 In_2O_{3-x} 269
5. **Interfaces and Multilayers** 270
5.1 Oxide Interface and Gas Adsorption 270
5.2 Multilayers 270
5.2.1 Compositional Multilayers 271
5.2.2 Doping-Modulated Multilayers 271
6. **Conclusions** 271
Acknowledgment 272
Glossary 272
Works Cited 273
Further Reading 276

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

During the past 15 years amorphous semiconductors have spawned a rapidly growing multibillion-dollar industry. Thin-film transistors made of amorphous silicon and amorphous silicon nitride control the pixels in large-area liquid-crystal displays used for flat-panel computer and television screens. Amorphous silicon and its alloys cover the drums of optoelectronic copying machines and are the active ingredients in thin-film photovoltaic modules that have achieved over 10% efficiency in converting sunlight into electricity. Compact-disk optical memories utilize the optical contrast between the amorphous and crystalline phases of chalcogenide materials. In many of these optoelectronic-device applications, use is made of the fact that amorphous semiconductors can be deposited as very thin films on large or curved flexible substrates.

Solid materials that lack the periodic symmetry and regular spatial arrangement of the atoms that are characteristic of crystals are called amorphous or vitreous or more generally noncrystalline. It is useful to distinguish between vitreous and amorphous materials. As vitreous materials (glasses) are cooled from the melt their viscosity increases as the atoms become frozen in place. Below a certain temperature, called the glass-transition temperature, the viscosity is so large that the vitreous material is considered a solid. Its atomic arrangement even though frozen is very similar to that of its liquid state above the glass-transition temperature. Amorphous materials, in contrast, can normally not be prepared by cooling their melt. They can be prepared only in the form of thin films by deposition on substrates that are kept below the crystallization temperature. This article deals with noncrystalline semiconductors, i.e., both vitreous and amorphous materials.

Semiconductors are materials whose electrical conductivity falls between those of metals and of insulators. Moreover, their conductivity can often be changed appreciably by altering the materials composition or adding specific elements that act as electron donors or acceptors. The electron concentration and conductivity of semiconductors can also be changed and controlled by charge injection or by exposure to light.

The study of the electronic properties that specifically distinguish noncrystalline semiconductors from insulators and ionic conductors started with the discovery by Goryunova and Kolomiets (1955) of the semiconductor properties of chalcogenide glasses, the exploration of amorphous Ge and Si by Tauc *et al.* (1966), and the bipolar switching and memory effects found by Ovshinsky (1968) in thin films of chalcogenide semiconductors.

The lack of long-range order has been an obstacle as well as a challenge for scientists trying to understand the electronic and optical properties of noncrystalline semiconductors. Between perfect order and disorder there are innumerable configurational and topological arrangements of atoms for which we do not have even a descriptive terminology. Nonetheless, there are structural similarities between crystalline and noncrystalline semiconductors, particularly in the nearest-neighbor sphere around each atom, because the same chemical forces and the same chemical bonds hold both kinds of solids together. Since the electronic properties of materials are largely governed by the chemical nature of the constituent atoms and the type of bonding between atoms, the same materials that are metals, semiconductors, or insulators as crystals remain so in the noncrystalline state. This leads to a useful classification shown in Table 1, which distinguishes covalent and ionic materials and separates the covalent ones into tetrahedrally bonded semiconductors and chalcogenide glasses. The latter contain as major constituents elements of group VI of the periodic table that are called chalcogens. Unlike crystalline semiconductors, noncrystalline semiconductors may contain many elements in a wide range of compositions since they do not constitute a thermodynamic equilibrium phase. As a consequence, noncrystalline semiconductors tend to be compositionally heterogeneous and contain spatial fluctuations of their band gap or their internal potential, which complicate a description of their physical properties. We restrict the following discussion to a few prototypical materials in each category of Table 1.

Table 1. Classification and examples of noncrystalline semiconductors.

1. Covalently bonded materials
   - A. Tetrahedral amorphous films
     $Si:H$, $Ge:H$, $SiGe_x:H$, $SiC_x:H$, $SiN_x:H$, GaAs, GaSb, ...
   - B. Tetrahedral glasses, $A^{II}B^{IV}C_2^{V}$
     $CdGe_xAs_2$, $CdSi_xP_2$, $ZnSi_xP_2$, $CdSn_xAs_2$, . . .
   - C. Chalcogenide glasses
     - (i) Elements and compounds
       Se, S, Te, As_2Se_3, As_2S_3, . . .
     - (ii) Multicomponent glasses
       Ge–Sb–Se Si–Ge–As–Te
       Ge–As–Se As_2Se_3–As_2Te_3
       As–Se–Te Tl_2Se–As_2Te_3, . . .
   - D. Others
     B, As, P, . . .
2. Quasi-ionic materials
   - A. Metal-oxide films
     In_2O_{3-x}, $In_2O_{3-x}:Sn$, SnO_{2-x}, ZnO_x, $AgSbO_3$, . . .
   - B. Oxide glasses
     V_2O_5–P_2O_5 MnO–Al_2O_3–SiO_2
     V_2O_5–GeO_2–BaO TiO_2–B_2O_3–BaO, . . .
3. Dielectric films
   SiO_{2-x}, Al_2O_3, Si_3N_4, BN, . . .

1. ELECTRONIC STATES

1.1 Chemical Bonding

Most amorphous semiconductors are covalently bonded, and the coordinations of their constituent atoms follow simple chemical rules. The solid has the lowest energy when the maximum number of valence electrons participates in covalent bonds, which means that they occupy bonding states and no electrons occupy antibonding states. Each state has two electrons with opposite spins according to Pauli's exclusion principle. Group IV elements form four hybridized sp^3 orbitals and are tetrahedrally coordinated. Group V elements form three covalent bonds with their p orbitals and leave a pair of s electrons nonbonded. Group VI elements, the chalcogens, use only two p electrons for bonding in twofold coordination, leaving two s electrons and two p electrons in nonbonding orbitals. This leads to the widely observed $8 - N$ rule, which relates the coordination number C with the column number N of the element:

$$C = 8 - N \quad \text{for} \quad N \geq 4, \tag{1}$$

$$C = N \quad \text{for} \quad N < 4. \tag{2}$$

The bond energies are further optimized by favoring stronger heteropolar over weaker homopolar bonds. Deviations of the covalent bond lengths and bond angles from the ideal values and stronger interactions of the nonbonding electrons contribute to the excess energy of the amorphous state relative to the energy of the crystalline state.

Materials that easily form glasses have average covalent coordinations C_{av} between 2 and 3. Phillips (1979) suggested that the optimal value is 2.45 for glass forming because the short-range order imposed by bond-stretching and bond-bending forces is just sufficient to exhaust the local degrees of freedom. Materials with C_{av} between 3 and 4 are overconstrained and can only be deposited as amorphous films, except for some compounds like $CdSi_xP_2$, which are called tetrahedral glasses; see Table 1.

1.2 Extended and Localized States

The bonding and nonbonding orbitals that hold all the valence electrons spread into a band of states, the valence band, as the atoms are brought close together to form a covalently bonded solid. The band is several eV wide and the spreading is caused by the interaction of neighboring orbitals, which overlap and form electron wave functions that extend over the whole volume of the solid. Likewise, the antibonding orbitals interact, overlap, and form extended wave functions. Their states spread into a band that is called the conduction band. Because the noncrystalline solid lacks translational symmetry, these extended wave functions are very different from the Bloch functions of periodic structures.

In materials that contain a dominant concentration of group VI elements, i.e., chalcogens, the valence band is formed by the two nonbonding (or lone-pair) p orbitals. These noncrystalline solids are called chalcogenide glasses or lone-pair semiconductors (Kastner, 1972).

The valence and conduction bands of states as well as the energy gap separating them are a consequence of chemical bonding. These bands do not have sharp energetic boundaries as they do in crystalline semiconductors. Instead, the variations in bond angles and bond lengths and the presence of weaker bonds as well as of topological bond structures that exist in noncrystalline solids, such as loops with odd numbers of covalent bonds, produce electronic states that tail from both bands into the gap. These tail states are localized in space with a typical localization radius of a few bond lengths. The energy that separates the localized tail states from the extended band states is called the mobility edge because at very low temperatures an electron should still be mobile in an extended state while it is immobile in a localized state. There is a mobility edge E_V separating band and tail states at the top of the valence band and a mobility edge E_C separating these two kinds of states at the bottom of the conduction band. The conductivity of a semiconductor at a given temperature is governed by the small concentration of electrons that are in the conduction band and the small concentration of missing electrons in the valence band. These missing electrons behave like positive charges and are called holes just as in crystalline semiconductors (see CONDUCTIVITY IN INSULATORS AND SEMICONDUCTORS). Since it costs less energy to produce a missing electron at the top compared with the bottom of the valence band, the energy of the hole is least at the top of the valence band.

The valence and conduction bands are approximately 5–10 eV wide and contain about twice as many electron states as there are atoms in the solid. The factor 2 comes from the two spin directions of the electron. Less than 0.1% of these states are localized tail states. Nonetheless, these tail states are important because they outnumber the concentrations of electrons and holes participating in conduction. One not only needs to know the energy of the states, but the concentration of states that exists in an energy interval dE at an energy E. This quantity is called the density of states $g(E)$ and given the units [$\mathrm{cm}^{-3}\ \mathrm{eV}^{-1}$]. This density-of-states function is sketched against electron energy for a typical noncrystalline semiconductor in Fig. 1. Since this is the energy range of interest, we only show the top of the valence band with its mobility edge E_V and the valence-band tail and the bottom of the conduction band with its tail and mobility edge E_C. Values of the gap $E_C - E_V$ range between 0.5 and 3 eV for semiconductors. Amorphous materials having a larger gap are called insulators. Near the gap center a peak in $g(E)$ is sketched in Fig. 1. This peak is due to localized states originating from defects, which are discussed in the following.

1.3 Electronically Active Defects and Dangling Bonds

A noncrystalline semiconductor whose atoms satisfy their valences by following the 8 − N rule of coordination can be considered defect free. Atoms that deviate from the 8 − N coordination rule are electronically active defects. If they are overcoordinated, they donate an electron to the host material, thereby changing their number of valence electrons to obey the 8 − N rule. For example, a P atom bonded to four neighbors donates an

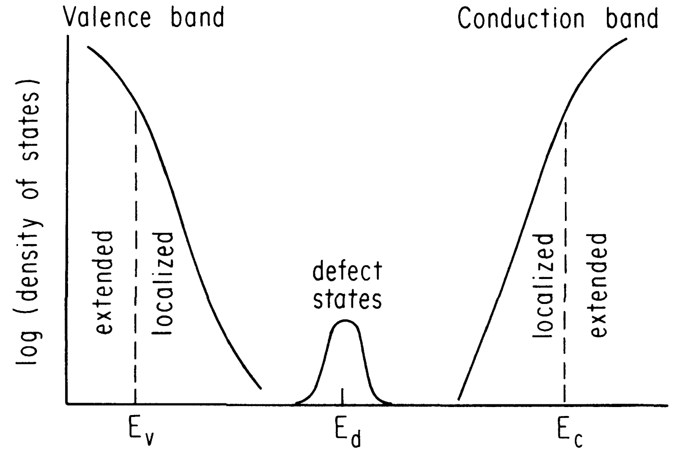

FIG. 1. Density of states distribution in a noncrystalline semiconductor. $E_c - E_v$ is the mobility gap.

electron and becomes positively charged, such that the number of valence electrons comes to agree with $N = 4$. Likewise an undercoordinated atom adjusts its number of valence electrons by accepting an electron and becoming negatively charged.

Coordination defects of the host material are often amphoteric defects that can exist in three charge states, D^+, D^0, and D^-, as they are occupied by zero, one, or two electrons. In chalcogenide glasses, these are undercoordinated (dangling-bond) defects and also overcoordinated chalcogen or pnictogen atoms. In *a*-Si:H, the dominant defect is believed to be a three-fold coordinated Si atom whose fourth sp^3 hybrid orbital remains unbonded. Such a dangling bond is often written as Si_3 where the subscript stands for the coordination number. After accepting an electron and becoming a negatively charged Si_3^-, its tetrahedral bond angles (~109°) might relax toward a p^3 configuration (~90°), whereas the positively charged Si_3^+ might prefer a sp^2 planar bond configuration (~120°). Alternatively, the dominant defect in *a*-Si:H has been pictured as a floating bond, i.e., a fivefold coordinated Si where the wave function of the gap state is distributed on the five neighbors of the central atom (Pantelides, 1986).

1.4 Positive- and Negative-*U* Defects

According to Pauli's exclusion principle, every electronic state can be occupied by up to two electrons. Doubly and singly occupied levels of localized states will not have the same energy because of the repulsive Coulomb energy U_C of the two electrons. Moreover, a change in charge state is usually accompanied by an adjustment of the neighboring bonds, which contributes a relaxation energy W. The two-electron (2*e*) level is then separated from the one-electron (1*e*) level by an effective correlation energy

$$U = U_C - W. \tag{3}$$

The possibility of a negative U occurring for large relaxations, $W > U_C$, was pointed out by Anderson (1975) and such defects were identified by Street and Mott (1975) in chalcogenide glasses.

Fermi statistics are not applicable to correlated electron states. Instead, the occupation probabilities f_0, f_1, f_2 for zero, one, and double occupancy are obtained from the grand partition function as

$$f_0 = z^{-1} = \{1 + 2\exp[-\beta(E - E_F)] + \exp[-\beta(2E - 2E_F + U)]\}^{-1}, \tag{4}$$

$$f_1 = z^{-1}\, 2\exp[-\beta(E - E_F)], \tag{5}$$

$$f_2 = z^{-1}\exp[-\beta(2E - 2E_F + U)]. \tag{6}$$

Here the zero of energy E is placed at the level E_d of the defect and $\beta = (kT)^{-1}$. One finds that for $+U$ defects the 2*e* level exists only when the 1*e* level is occupied. For $-U$ defects, however, the 1*e* level lies above the 2*e* level and remains unoccupied for $kT \ll |U|$. The important difference between $+U$ and $-U$ defects is illustrated in Fig. 2, which shows the shift of the Fermi energy E_F as the defect occupancy n/N of N defects is varied as, for instance, by doping or by inducing a space charge (Adler and Joffa, 1976). The difference in behavior of E_F in Figs. 2(a) and 2(b) can be understood as follows: $E_C - E_F$

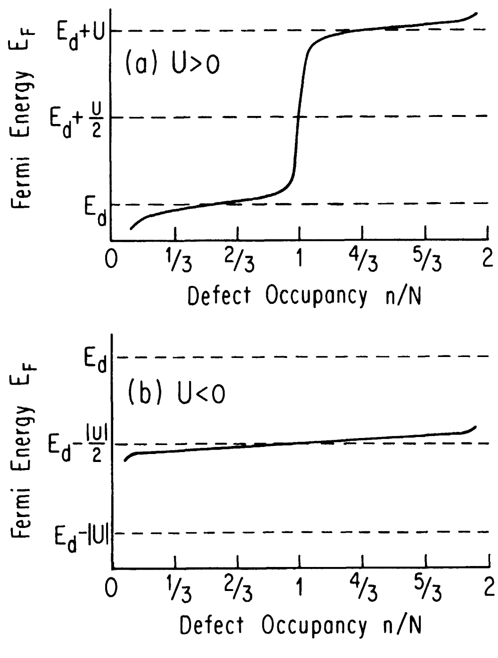

FIG. 2. Fermi energy as a function of electronic occupancy *n*/*N* of defects having (a) positive correlation energy *U* and (b) negative correlation energy, $-|U|$. After Adler and Joffa (1976).

represents the average thermal energy for exciting electrons to the conduction band. This energy remains the same for $-U$ defects; it is the average of the first and second ionization energy, $E_C - E_F = \frac{1}{2}(E_1 + E_2)$, as long as there are some $-U$ defects in the D^- and D^+ states. Hence E_F remains pinned for $0 < n/N < 2$. For $+U$ defects E_F moves quickly from the 1*e* to the higher lying 2*e* level at $n = N$. Figure 1 shows the density of one-electron states. Therefore, the 2*e* level of D^- does not appear in this figure.

1.5 Thermodynamic Equilibrium of Defects

The observation that coordination defects produce electronic levels in the gap or act as donors or acceptors holds for both crystalline and noncrystalline semiconductors. There are important differences, however. The translational symmetry of crystals imposes the same coordination environment on all equivalent atomic sites. The energies of defects of each kind are therefore the same and the doping efficiency of the doping elements is 100 percent. In noncrystalline semiconductors, the coordination is not confined by symmetry, and foreign atoms can achieve their preferred $(8 - N)$-rule coordination. The number of coordination defects is therefore governed above a thermodynamic equilibration temperature T_E by minimizing the free energy, and the doping efficiency can be considerably smaller than unity. In glasses, T_E is the glass-transition temperature, which usually lies between 100 and 300 °C. In hydrogenated amorphous silicon and silicon alloys, the defect reactions are mediated by diffusing hydrogen. The defect reaction can therefore attain thermodynamic equilibrium on a time scale set by hydrogen diffusion and T_E can be taken as the temperature below which the defect equilibration takes longer than about 10 min. This temperature has been found to lie between 150 and 200 °C.

1.5.1 Defect Pool The native dangling-bond defect concentration N_D often lies in the low 10^{15}-cm^{-3} range in undoped *a*-Si:H because of the presence of about 10 at.% hydrogen, which bonds to most potential dangling bonds. In addition to passivating preexisting dangling bonds, H atoms break strained Si–Si bonds, relieving strain in the material. Why is there still such residue of dangling bonds in the presence of about $5 \times 10^{21}\ \text{cm}^{-3}$ hydrogen atoms? The residual defect concentration may be the equilibrium value N_D established above T_E from a pool of potential defects, N_{Pool},

$$N_D = N_{\text{Pool}} \exp(-E_D/kT_E). \tag{7}$$

The effective defect-formation energy E_D depends on the position of the Fermi energy E_F because of the energy $E_d - E_F$ gained by transferring an electron from D^0 to E_F or the energy $E_F - (E_d + U)$ gained by transferring an electron from E_F to D^0. The energy in the gap E_d of a dangling bond depends on the bonding environment. It is possible to conceive of a pool of potential defects that, if picked at random, would cover energies over a wide range in the gap. The point is that they are not picked at random but with preference for a low formation energy. Hence, defects with small E_d are favored in *n*-type material where E_F lies high and defects with large E_d are favored in *p*-type material where E_F lies low in the gap. According to this defect-pool model, defects having small or large E_d equilibrate at $T \geq T_E$ in response to the energy position of E_F. This happens regardless of whether E_F is determined by doping or by an induced space charge as for instance in a field-effect transistor.

1.5.2 Doping Equilibration We restrict the discussion to amorphous, tetrahedrally coordinated hydrogenated materials because they are essentially the only noncrystalline semiconductors that can be doped. Choosing *a*-Si:H as an example one finds the following doping and defect reactions (Street, 1985):

$$\text{Si}_4^0 + \text{P}_3^0 \rightleftarrows \text{P}_4^+ + D^-, \tag{8}$$

$$\text{Si}_4^0 + \text{B}_3^0 \rightleftarrows \text{B}_4^- + D^+, \tag{9}$$

$$\text{Si}_4^0 + \text{Si}_4^0 \rightleftarrows D^+ + D^-. \tag{10}$$

The left-hand sides are the normal $(8 - N)$-rule coordinations. Hence, P_3^0 and B_3^0 do not act as donors and acceptors, respectively. The right-hand sides require the formation energies for dangling bonds and for changing the dopant coordinations to P_4 and B_4. These energies depend on the position of E_F

as a result of the energy gain of transferring the electrons to and from E_F as discussed above, and are smallest when E_F lies near the gap center. As a consequence, the doping efficiency is highest in compensated *a*-Si:H that contains equal concentrations of P and B. The reactions (8)–(10) are mediated by hydrogen diffusion and hence come to rest at $T < T_E$. The P_4^+ donors in *n*-type material are compensated by dangling-bond defects D^- according to Eq. (8). As a consequence, E_F remains between the donor and D^- gap states and cannot move into the conduction band as it does in heavily doped degenerate crystalline semiconductors. The same defect compensation occurs upon *p*-type doping according to Eq. (9).

The final charge state of dopants and defects depends on E_F, i.e., the combined effect of all defect reactions and the presence of induced space charges.

1.5.3 Valence-Alternation Defects It sometimes costs less energy to form a pair of coordination defects than a single one. This happens in chalcogenide glasses where a charge transfer from one defect to another leads to a bonding change and a net gain in energy. This means that the reaction

$$2D^0 \rightarrow D^+ + D^- \tag{11}$$

is exothermic and D is a $-U$ defect (Street and Mott, 1975). Kastner *et al.* (1976) proposed that the D^+-D^- pair of defects derives from alternating the valences of the group VI (chalcogen) and group V (pnictogen) atoms. The $8 - N$ rule is then obeyed by the charged defects in the following manner: The D^+ defects are overcoordinated and the D^- defects are undercoordinated chalcogen or pnictogen atoms. Since the number of covalent bonds is not changed by creating such D^+-D^- valence-alternation pairs (VAP), the formation energy E_{VAP} of a defect pair should be small. $E_{VAP} \sim 0.5$–0.7 eV is approximately the positive Coulomb energy E_C of placing the second electron on the undercoordinated D^- center. The equilibrium concentrations $[D^-]$ and $[D^+]$ of the defects near the glass-transition temperature T_g follow from the law of mass action and are

$$[D^-][D^+] = N_A^2 \exp(-E_{VAP}/kT_g) = N_0^2. \tag{12}$$

N_A is the concentration of atoms capable of undergoing valence alternations by utilizing the nonbonding orbitals of adjacent atoms. An appreciable density, $N_0 \sim 10^{17}$ cm^{-3}, of intrinsic VAP defects is therefore expected to exist even in well-annealed chalcogenide glasses.

The discussion of the pinning of E_F by $-U$ defects of concentration N (see Fig. 2) suggests that foreign dopant atoms exceeding $2N$ can unpin E_F and produce *n*-type or *p*-type chalcogenide glasses. Such attempts fail because Eq. (12) demands that the product $[D^-][D^+]$ reach an equilibrium value N_0^2 at T_g. Large concentrations of charged acceptors $[A^-]$ added to the glass above T_g will force a compensating increase in $[D^+]$ to satisfy the neutrality condition

$$[D^+] = [D^-] + [A^-]. \tag{13}$$

As a consequence, the total number $[D^+] + [D^-]$ of $-U$ defects increases and always exceeds $[A^-]$ while the product $[D^-][D^+]$ remains fixed obeying Eq. (13) (Fritzsche and Kastner, 1978). This self-compensation mechanism of VAP defects can be circumvented by adding modifying additives to the glass below T_g where the VAP concentration remains frozen (Ovshinsky, 1977).

1.6 Inhomogeneities, Potential Fluctuations, Percolation

Amorphous semiconductors tend to be inhomogeneous. Only glasses having an average coordination close to $C_{av} = 2.4$ can be prepared rather strain-free from the melt (Phillips, 1979). Tetrahedrally coordinated semiconductors, including those that are hydrogenated, form overconstrained networks whose internal strain is often relaxed by the formation of voids and clusters of certain constituents. If the length scale of the inhomogeneities exceeds the tunneling distance of charge carriers, then fluctuations of the band gap or of the internal potential create classical turning points for carriers, which restrict the carrier motion to percolation paths (Zallen, 1983). Compositional inhomogeneities that lead to band-gap fluctuations require analysis on mesoscopic scales. Long-range potential fluctuations arising from random distributions of charged coordination defects can be estimated as follows.

For a concentration N of charged defects, the statistical deviation ΔZ of their average number in a volume L^3 is

$$\Delta Z = (NL^3)^{1/2}. \tag{14}$$

On a length scale L, these cause potential variations

$$\Delta V = (NL^3)^{1/2} e/4\pi\epsilon\epsilon_0 L. \tag{15}$$

The physics of the semiconductor determines either ΔV or L as illustrated by the following examples. In an n-type sample containing a concentration of N shallow donors and according to Eq. (8) an equal number of compensating $+U$ defects, the length scale in Eq. (15) is the screening length

$$L = (8\pi\epsilon\epsilon_0/ne)^{1/2}, \tag{16}$$

where n is the concentration of trapped and free charges. In a sample compensated with equal concentrations N of shallow donors and acceptors, screening occurs when the fluctuations move E_F into the band tails. Hence $\Delta V \leq E_g/2$ and L follows from Eq. (15). In an undoped semiconductor containing only $+U$ defects, $\Delta V \approx U/2$, and L follows again from Eq. (15). Chalcogenide glasses containing $-U$ defects should be free of potential fluctuations because E_F remains pinned. ΔZ of Eq. (14) is kept zero by charge transfer between the $-U$ defects.

Compositional heterogeneities lead to band-gap and potential fluctuations, which affect the processes that depend on the position of the Fermi energy with respect to the band edges, such as the defect concentration, their charge state, and optical absorption (Bar-Yam *et al.*, 1986; Branz and Silver, 1990; Fritzsche, 1994). Moreover, long-range potential fluctuations affect electronic transport, in particular the activation energy of the thermopower (Overhof and Beyer, 1984) (see Sec. 2.4) and the drift mobility of electrons and holes (Street *et al.*, 1988b; Howard and Street, 1991).

1.7 Determination of Electronic States

The basis for understanding a semiconductor is the knowledge of the concentration and energies of the energy eigenstates. They are difficult to extract from experiments on amorphous semiconductors because there is a lack of sharp demarcation energies such as band edges to serve as reference points. Moreover, defect energies depend on the position of E_F if they arise from a defect pool and are subject to relaxations. In addition, defect and dopant concentrations exhibit a number of metastabilities, discussed in a later section.

1.7.1 Optical Absorption

The optical absorption spectra of different noncrystalline semiconductors are very similar except for the magnitude of the optical gap. Because of the lack of long-range order and translational symmetry, there are no momentum-conservation selection rules governing optical transitions. All energy-conserving transitions from occupied to empty states in Fig. 1 are allowed. This explains why materials exhibit stronger absorption when they are amorphous than when they are crystalline. The change in optical properties is used for data storage in phase-change compact-disk optical memories.

A typical absorption spectrum is shown in Fig. 3. The high absorption region I ($\alpha \geq 10^4$

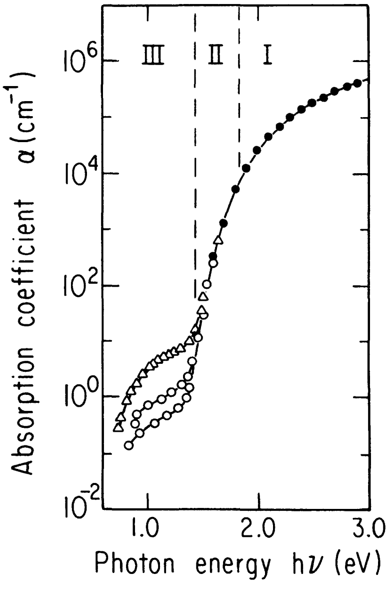

FIG. 3. Absorption spectrum of *a*-Si:H: interband-absorption region I, Urbach-edge region II, and defect-absorption region III. The open symbols are data taken with the photothermal-deflection method. The different curves in region III represent samples with different defect concentrations.

cm^{-1}) reflects the joint density of extended band states with the above assumptions. Assuming parabolic densities of states, as for free particles, Tauc (1974) obtained the relation

$$(\alpha h\nu)^{1/2} = B(h\nu - E_g), \tag{17}$$

which defines the optical Tauc gap E_g. The constant B contains the optical matrix element and the effective electron and hole masses. Its magnitude has been found to lie between 300 and 1000 $(\text{eV cm})^{-1/2}$ when the energies are measured in eV and the absorption coefficient in cm^{-1}. Toward lower photon energies $h\nu$ and below $\alpha = 5 \times 10^3$ cm^{-1} the absorption is exponential over one to three decades and is described by

$$\alpha = \alpha_0 \exp(h\nu/E_0). \tag{18}$$

This Urbach-edge region II arises from transitions between the band tails and the opposite bands. E_0 is the exponential slope of the density of states of the valence- or the conduction-band tail, whichever is larger. This is usually the valence-band tail, and the magnitude ranges between $E_0 \sim 0.05$ and 0.08 eV.

Absorption region III toward lower $h\nu$ is due to transitions to or from coordination defects in the gap. Their concentration N can be obtained by integrating over the defect absorption,

$$N = A \int \alpha d(h\nu). \tag{19}$$

The coefficient A is obtained by determining N by other methods or by assuming the same dipole-matrix element as for other optical transitions. The value of A depends on the method used for determining α in region III. For photothermal-deflection spectroscopy $A \approx 8 \times 10^{15}$ cm^{-2} eV^{-1} in a-Si:H (Jackson and Amer, 1982), and for the constant-photocurrent method one often uses $A \approx 1.6 \times 10^{16}$ cm^{-2} eV^{-1} because only half of the transitions yield electron photocarriers. The correct value depends on doping and the position of the Fermi energy.

1.7.2 Photoemission Spectroscopies

Direct information about the valence- and conduction-band density of states is obtained from two photoemission spectroscopies (Ley, 1984). The first probes the occupied valence-band states. Photons of known energy excite electrons above the vacuum level so that they can leave the sample to be counted and energy analyzed. The second probes the empty conduction-band states. Electrons of known energy bombard the sample. As they fall into unoccupied states they emit x rays. The conduction band $g(E)$ is obtained by measuring the x-ray intensity at fixed wavelength as the electron-beam energy is varied.

Figure 4 compares the valence- and conduction-band spectra of amorphous hydrogenated silicon with those of crystalline silicon. The similarity of these spectra emphasizes that the chemical bonds have a much stronger effect on the energy states than the presence of long-range order (Jackson *et al.*, 1985).

Optical absorption spectra are essentially the product of the square of the optical dipole matrix element and the joint densities of the valence- and conduction-band states.

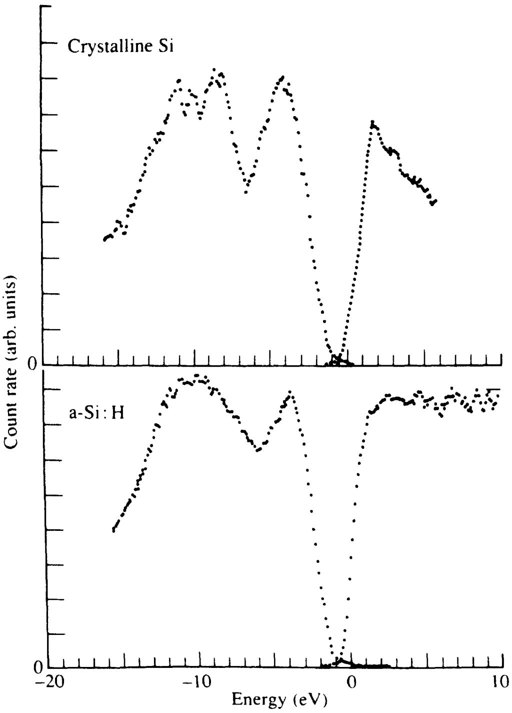

FIG. 4. The valence and conduction bands of crystalline silicon and a-Si:H measured by x-ray photoemission spectroscopy. After Jackson *et al.* (1985).

By combining the information from optical absorption and their photoemission spectra of Fig. 4, Jackson *et al.* (1985) were able to determine the dipole matrix element for *a*-Si:H. They found a magnitude of 10 Å times the electronic charge independent of energy below ~3.4 eV. This magnitude is the same for transitions between extended states and between localized and extended states. Transitions between localized states have vanishingly small matrix elements since most of them are spatially uncorrelated. These findings support the interpretation of optical absorption spectra as reflecting joint densities of states, and they indicate that the random-phase approximation is valid for *a*-Si:H.

Further details of $g(E)$ near the valence-band edge are revealed by a photoemission technique that measures the total yield $Y(h\nu)$ of emitted electrons as a function of photon energy (Winer and Ley, 1987):

$$Y(h\nu) \propto M^2(h\nu) \int_{E_{\mathrm{vac}}}^{\infty} g_V(E) g_c(E + h\nu) dE. \qquad (20)$$

The density of states of the valence band $g_V(E)$ is proportional to $dY/d(h\nu)$ because the probability of photoemission is nearly constant above the work-function energy E_{vac} and the transition-matrix element $M(h\nu)$ is assumed to be a constant. Figure 5 shows the results for *a*-Si:H. One observes an exponential tail region with $E_0 = 0.051$ eV and a band of defect states centered about 0.7 eV above E_V. These measurements cannot identify the mobility edge E_V. Other evidence suggests $g_V(E_V) \approx 10^{21}$ cm^{-3} eV^{-1}.

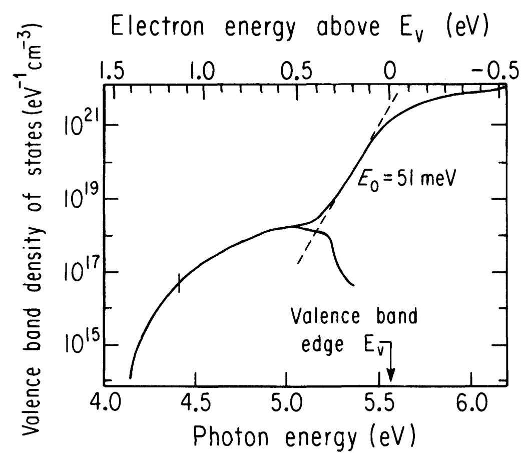

FIG. 5. Photoemission total yield measurements of the valence-band density of states of *a*-Si:H. After Winer and Ley (1987).

1.7.3 Electron Spin Resonance (ESR)

ESR yields important information about the concentration, the structure, and the localization radius of singly occupied localized states. Since the local defects are randomly oriented in an amorphous material, one measures the powder average of the gyromagnetic g tensor and of the nuclear hyperfine interaction. Modern ESR spectrometers can detect as few as 10^{10} spins per gauss of linewidth. By measuring ESR signals as a function of sample thickness, one can separate spin contributions of the bulk of the sample from those of the surface and of interfaces. Moreover, spin orientations play an important role in hopping of charge carriers between localized states and in recombination processes as discussed in Sec. 3.6. As a result, photoconductivity and luminescence become spin dependent and serve as sensitive tools to detect ESR optically (Solomon, 1979; Morigaki, 1984).

The lack of any ESR signal in chalcogenide glasses (Agarwal, 1975) indicates that the localized valence-band-tail states are doubly occupied and the conduction-band tail states are empty and thus have the parentage of their respective bands. Moreover, the general lack of paramagnetic states in these materials suggests that the defects have a negative effective correlation energy (Anderson, 1975) and exist in the D^+ and D^- charge states, which are diamagnetic. By illuminating these glasses at low temperatures, Bishop *et al.* (1977) created the metastable neutral D^0 states and measured the light-induced ESR (LESR). The concentrations (10^{16}–10^{17} cm^{-3}) of the defects and the observed anisotropies of the g tensors and hyperfine interactions agree with those expected for valence-alternation defects (Bishop *et al.*, 1977; Gaczi, 1982).

In evaporated or sputtered *a*-Si, which does not contain hydrogen, the spin signal is always very large, originating from 10^{19}–10^{20} cm^{-3} defects. These have a gyromagnetic ratio $g = 2.0055$ and have been identified by Brodsky and Title (1969) with silicon dangling bonds. In hydrogenated material, *a*-Si:H, the defect concentration is reduced to the low 10^{15}-cm^{-3} range. Even though the defect concentration increases with doping according to Eqs. (8) and (9), the ESR signal vanishes as the Fermi level is shifted with doping because the defects become charged and hence diamagnetic. From these observations one deduces a positive correlation energy $U \approx 0.3$ eV. By using light-induced ESR and creating metastable D^0 centers, one can determine the defect concentration in doped *a*-Si:H.

ESR signals at $g = 2.0080$ and $g = 2.0040$, respectively, are detected for valence-band tail states and conduction-band tail states that are singly occupied either by doping or by exciting electrons and holes into the tail states by light at low temperatures (Stuke, 1977). In strongly doped *a*-Si:H, one also finds an ESR signal of neutral phosphorus or arsenic donors in *a*-Si:H. From the hyperfine splitting, one deduces a localization radius of 10 ± 1 Å for the extra electron bound to the phosphorus and 9 ± 1 Å for the arsenic donor (Stutzmann and Street, 1985). In a similar manner, one has identified the spin signals of the Ge dangling bonds and of the singly occupied tail states in *a*-Ge:H (Stutzmann *et al.*, 1987).

1.7.4 Transient Emission Spectroscopies Several transient emission techniques have been developed to explore the energies and concentrations of deep electronic gap states in semiconductors. The principle of junction transient measurements is illustrated in Fig. 6. A semiconductor is shown with one discrete level of deep states in the space-charge region of a Schottky barrier, which is produced by making a metal contact to an *n*-type semiconductor. In Fig. 6(a) the sample is at equilibrium at a certain reverse bias. In Fig. 6(b) a voltage filling pulse allows gap states in the depletion region to capture electrons from the conduction band. When the bias is returned to its original value, as shown in Fig. 6(c), occupied gap states above E_F remain occupied until their electrons are excited to the conduction band by thermal or optical transition processes. The emitted electrons are then rapidly swept out by the high electric field of the depletion region. The thermal release time is

$$\tau_e = \nu_0^{-1} \exp(\Delta E/kT), \qquad (21)$$

which increases exponentially with the energy depth ΔE of the gap state from E_C. The frequency ν_0 is related by detailed balance to the trapping cross sections of the gap states. Its magnitude must be determined independently; $\nu_0 \approx 10^{12}$ to 10^{13} s^{-1} (Lang *et al.*, 1982a; Johnson, 1983), but values near $\nu_0 \approx 10^8$ s^{-1} have also been reported (Okushi *et al.*, 1982). As the extra carriers are released, the original depletion region gets reestablished.

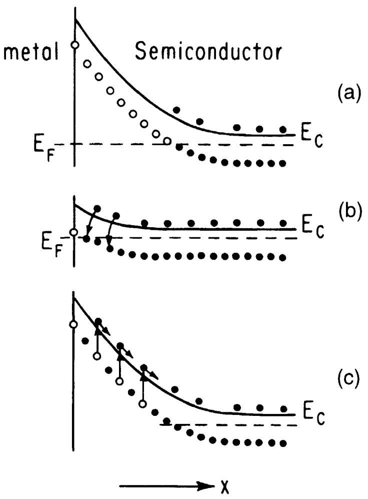

FIG. 6. Sketch of a metal–semiconductor Schottky barrier. (a) Reverse bias at equilibrium, (b) after voltage filling pulse, and (c) return to original reverse bias and thermal emission of nonequilibrium carriers.

The thermal-emission transient can be monitored by measuring the change in the Schottky-barrier capacitance $\Delta C(t)$ as a function of time, which occurs because the depletion region will contract as negative charge is lost and the positive charge density increases. $\Delta C(t)$ can be analyzed to yield $g(E)$, the energy dependence of the density of gap states as shown for an *n*-type *a*-Si:H sample in Fig. 7. According to Eq. (21) extra electrons in deep gap states take a long time to be released. It is therefore often found convenient to study $\Delta C(t_0)$ after a time window (t_0) as a function of temperature. The analysis of such thermal-emission spectra gave the first evidence for the existence of a broad band of defects in *n*-type *a*-Si:H located near the gap center (Cohen *et al.*, 1980; Lang *et al.*, 1982a).

Another variation is the use of subgap light in place of the purely thermal excita-

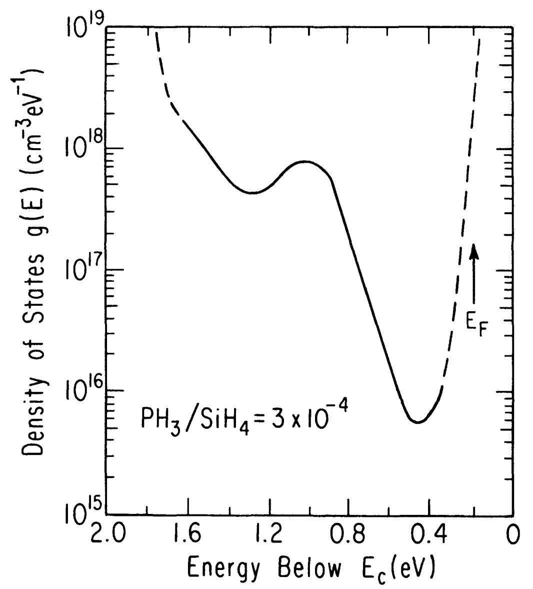

FIG. 7. Density of states in n-type a-Si:H measured by transient capacitance spectroscopy. The central peak is associated with dangling-bond defects. After Lang *et al.* (1982a).

tion process. This method is called photocapacitance spectroscopy. The photocapacitance signal, defined as the difference in the capacitance transient signals with and without subgap light exposure, reveals the distribution of optical transitions for the nonequilibrium electron population produced by the filling pulse. This spectrum is very similar to that obtained by photothermal-deflection spectroscopy (Johnson and Biegelsen, 1985; Gelatos *et al.*, 1986, 1988). However, because the change in capacitance is sensitive to the sign of the charge left behind in the depletion region, these measurements can be used to distinguish minority- and majority-carrier excitation processes (Gelatos *et al.*, 1988).

Several useful variations of this basic deep-level transient spectroscopy (DLTS) have been developed. For example, by changing the duration of the filling pulse, structural relaxations of the dangling-bond defect following the trapping of an extra electron can be studied (Cohen *et al.*, 1993), though the interpretation of these results remains controversial. By combining ESR measurements with DLTS, the paramagnetic nature of the gap state, either before or after receiving an extra electron, can be identified (Cohen *et al.*, 1982; Johnson and Biegelsen, 1985). In this manner, one can demonstrate that the peak in $g(E)$ located about 0.9 eV below E_C in n-type a-Si:H shown in Fig. 7 originates from D^- dangling-bond states. A deep minimum in $g(E)$ was found between these D^- states and the exponential rise in $g(E)$ of the conduction-band tail.

A final variation on such capacitance methods that has been used to study deep defects in a-Si:H with good success has been the drive-level capacitance profiling method (Michelson *et al.*, 1985). This method is an extension of the more familiar *C-V* profiling method that is used to determine the spatial variation of dopants in crystalline semiconductors. In the drive-level method, however, the variation of the junction capacitance is determined not only as a function of the dc bias but also as a function of the measurement frequency (or temperature) as well as the amplitude of the alternating voltage. In this manner it is possible to determine the distribution of deep defects as a function of both thermal energy depth and their spatial position within the sample (Unold and Cohen, 1991).

The transient emission spectroscopies require a manipulation of the band bending near a Schottky contact or a junction. This is not possible in chalcogenide glasses, whose Fermi level is pinned by the $-U$ defect centers (see Secs. 1.4 and 1.5.3).

2. ELECTRONIC TRANSPORT

In dealing with the transport properties of noncrystalline semiconductors one faces some major problems. Many studies are made on thin films whose surfaces and interfaces can have different properties from that of the bulk on account of surface adsorbates, out-diffused hydrogen, or interface states. Furthermore, electronic transport is quite sensitive to material heterogeneities that depend on preparation conditions.

Even when care is taken to alleviate these material problems, it is very difficult to extract microscopic transport parameters from experiments because of the influence of the large number of localized tail states and the fact that the anomaly of the Hall coefficient makes it impossible to learn about the band mobilities of the carriers and their temperature dependence.

Finally one is faced with profound theoretical problems. The disorder potential, which originates from the lack of long-range order in noncrystalline semiconductors, prevents us from solving the Schrödinger equation and from treating electronic transport within the effective-mass approximation. From Anderson's theory of localization (Anderson, 1958) one expects regions of localized states at the edges of the conduction and valence bands as sketched in Fig. 1 and verified experimentally as discussed in previous sections. Even though the distinction between localized and extended states may be conceptually sharp, the electronic transport usually involves both kinds of states.

In this spirit one may associate a differential conductivity $\sigma(E)$ with each energy slice dE at E by using an energy-dependent mobility $\mu(E)$ and write

$$\sigma(E) = e\mu(E)g(E)f(E)[1 - f(E)], \tag{22}$$

$$\sigma = \int \sigma(E)dE. \tag{23}$$

Even this approach (Kubo, 1957; Greenwood, 1958) is based on assumptions that are not easily justified in amorphous semiconductors. First, the use of a single-particle density of states $g(E)$ and the Fermi function $f(E)$ neglects correlation effects of electrons in localized states. Second, Eqs. (22) and (23) assume that conduction in each energy shell can be treated separately, a doubtful procedure when charge motion involves energy exchange with the lattice as in the case of phonon-assisted hopping and polaron conduction. Finally, the above equations assume that the material is homogeneous and that conduction is not spatially restricted to percolation paths.

2.1 Hopping Transport in Localized States

In contrast to extended states, the wave functions of localized states are confined in space, decaying exponentially as $\exp(-r/a)$. The localization radius a presumably depends on the binding energy $E_c - E$ of the state. A reasonable estimate is

$$a(E) = [h^2/m(E_c - E)]^{1/2}. \tag{24}$$

The tunneling or hopping rate of a charge carrier from a state at energy E_i to an unoccupied state at E_F over a separation R is

$$\nu(R) = \nu_0 \exp(-2R/a) \exp[-(E_F - E_i)/kT] \quad \text{for} \quad E_F > E_i, \tag{25a}$$

$$\nu(R) = \nu_0 \exp(-2R/a) \quad \text{for} \quad E_F < E_i. \tag{25b}$$

Both the density of localized tail states and the localization radius a increase as one approaches the mobility edge from the gap. As a consequence, the hopping probability rises rapidly so that the mobility of the charge carriers may lack the sharp discontinuity that gave the mobility edge its name (Cohen *et al.*, 1969).

At low temperatures very few charge carriers are excited into the extended states and conduction proceeds by phonon-assisted hopping first to neighboring localized states and at even lower temperatures to more distant localized states that are close in energy (variable-range hopping) (see Sec. 8 in CONDUCTIVITY IN INSULATORS AND SEMICONDUCTORS). A particular energy level in the band tail of localized states, called the transport energy, plays a crucial role in hopping transport of carriers in equilibrium as well as nonequilibrium conditions for both steady-state and transient phenomena. For a density-of-states distribution that falls off exponentially with a logarithmic slope E_0 as in Fig. 5, the distance of the transport energy E_t from the nearest mobility edge is

$$E_t = 3E_0 \ln\left[\frac{3E_0 a(4\pi N_0/3)^{1/3}}{2kT}\right], \tag{26}$$

where N_0 is the concentration of localized tail states. The transport energy maximizes the hopping rate and is the most probable final energy of hopping carriers regardless of their initial energy (Grünewald and Thomas, 1979; Shapiro and Adler, 1985; Monroe, 1985).

The hopping processes have been studied in a-In_2O_{3-x} at low temperatures (Ovadyahu, 1986, 1990; Bellingham *et al.*, 1991). Evaporated or sputtered films of a-Ge or a-Si also contain a sufficient concentration of defect states near E_F to support hopping conduction at low temperatures; however, lack of homogeneity of these materials makes a quantitative analysis difficult. The photoconductivity in both chalcogenide glasses and *a*-Si:H, at low temperatures (see Sec. 3.4), is due to hopping of photocarriers from higher-energy to lower-energy localized gap states before they recombine.

2.2 dc Conductivity in Extended States

For conduction in extended states Eqs. (22) and (23) yield for electrons

$$\sigma(T) = e\mu_c g(E_C)kT \exp[-(E_C - E_F)/kT] \tag{27}$$

and an equivalent expression for holes. In general both E_F and the mobility edges E_C and E_v depend on temperature. Above 100 K the dependence is linear so that

$$E_C - E_F = E_a - (\delta_c + \delta_F)T. \tag{28}$$

The conductivity follows the Arrhenius law

$$\sigma(T) = \sigma_0 \exp(-E_a/kT) \tag{29}$$

with the prefactor

$$\sigma_0 = e\mu_c g(E_C)kT \exp[(d_c + \delta_F)/k]. \tag{30}$$

The disorder potential severely limits the phase coherence of the extended states. The carrier mobility in the extended states is between 1 and 10 cm^2/(V s). This corresponds to a mean free path of the order of 5–7 Å at 300 K. Simple transport equations that assume weak scattering conditions are obviously not valid for these small scattering lengths. Using the random-phase approximation, Friedman (1971) obtained for the band mobility

$$\mu = \frac{2\pi}{3}\frac{ea^3}{h} C_{\text{av}} \frac{J}{kT} a^3 J g(E_V), \tag{31}$$

where $C_{\text{av}} = 3$ is an average coordination number, $J \approx 1$ eV is the two-site transfer integral, and $a = 0.25$ nm is the interatomic spacing. Another estimate (Cohen, 1970) treats the charge transport in classical terms as diffusion of carriers jumping between neighboring sites with jump rate $\nu = J/h$:

$$\mu = \frac{1}{6}\frac{ea^2}{kT}\nu. \tag{32}$$

Both expressions (31) and (32) yield $\mu \sim 5$ cm^2/(V s) at 300 K.

2.2.1 Chalcogenide Glasses The conductivity of chalcogenide glasses follows the Arrhenius law (28) with an activation energy E_a that is close to one-half the optical gap: $E_a = E_g/2$. This quite general relationship between E_a and E_g in chalcogenide glasses suggests that E_F is pinned near midgap and that the optical gap E_g is approximately equal to the mobility gap $E_C - E_V$. Figure 8 shows the logarithm of the room-temperature conductivity against E_a for a large number of chalcogenide glasses. The straight lines have the slope $(kT)^{-1}$ with $T = 300$ K, and the intercept with the $E_a = 0$ axis is σ_0. For most of these materials one finds $10^2 < \sigma_0 < 10^4$ V^{-1} cm^{-1}. The positive sign of the thermopower indicates that conduction is by holes in the valence band. From the mid-range value $\sigma_0 = 10^3$ V^{-1} cm^{-1} and Eq. (30) we can estimate the hole mobility μ_v in the following manner. In these glasses $\delta_F = 0$ because E_F is pinned by the coordination defects, which have a negative correlation energy (see Sects. 1.4 and 1.5.3). The temperature dependence of the hole mobility edge E_V is approximately half of that of the gap, which is for these glasses is

$$\frac{dE_g}{dT} \approx 8 \times 10^{-4} \text{ eV/K}. \tag{33}$$

Substituting these values into Eq. (30) and using a reasonable estimate $g(E_V) = 10^{21}$ cm^{-3} eV^{-1} we obtain at 300 K a hole mobility $\mu_v = 2$ cm^2/(V s). This agrees well with the estimates from Eqs. (31) and (32). This shows that between, say, 200 K and the glass-transition temperature T_g the conduction takes place in the extended states of the valence band. Only some Bi-containing chalcogenide glasses show *n*-type conduction.

2.2.2 Hydrogenated Amorphous Silicon The importance of *a*-Si:H arises from the fact that the Fermi energy can be shifted through the gap by *n*-type or *p*-type doping

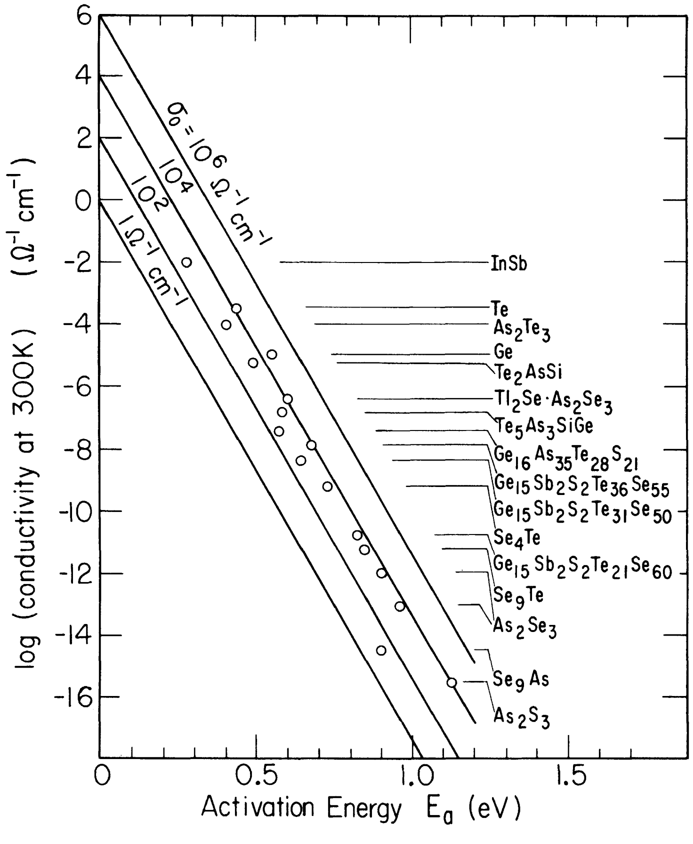

FIG. 8. Relation between conductivity at 300 K and activation energy E_a defined by Eq. (29) for several chalcogenide glasses.

(Spear and LeComber, 1975). Figure 9 shows the dramatic change of the room-temperature conductivity as small fractions of dopant gases B_2H_6 (for *p*-type) and PH_3 (for *n*-type) are added to the SiH_4 gas used for plasma-assisted chemical vapor deposition of these materials.

Over a large temperature range 150 < T < 450 K the conductivity follows the Arrhenius law (29). In contrast to chalcogenide

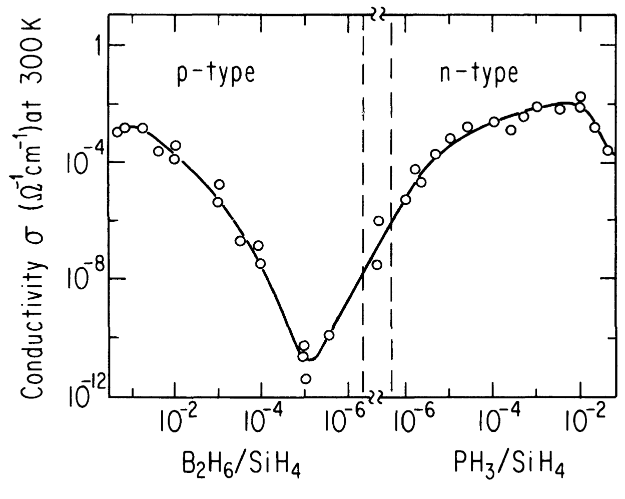

FIG. 9. Conductivity at 300 K of *a*-Si:H as a function of dopant concentration in the gas phase. After Beyer and Overhof (1984).

glasses, however, σ_0 is not constant but changes over five orders of magnitude as E_a is changed by *n*-type or *p*-type doping. This is shown in Fig. 10. This relation can approximately be described by

$$\ln\sigma_0 = \ln\sigma_{00} + E_a/kT_m, \tag{34}$$

where $\sigma_{00} \approx 0.1\ \text{V}^{-1}\ \text{cm}^{-1}$ and $T_m \approx 600$ K. T_m is called the Meyer–Neldel temperature because relation (34) was first observed in a number of polycrystalline materials by Meyer and Neldel (1937). Overhof and Thomas (1989) tried to explain this relation in *a*-Si:H in the following manner. In strongly doped material, E_F lies in the steeply rising $g(E)$ of the band tails. E_F moves in general toward lower $g(E)$ with increasing T in order to conserve charge neutrality. This so-called statistical shift of E_F contributes therefore a term δ_F to Eq. (28), which can be of either sign depending on the shape of $g(E)$ near E_F. As a consequence, σ_0 given by Eq. (30) is expected to depend on E_a. The solid curve in Fig. 10 was calculated by them for a reasonable model for $g(E)$. At high doping levels (small E_a) the small values of σ_0 may be due to hopping conduction in the localized tail states.

2.2.3 Amorphous In_2O_{3-x} In most amorphous semiconductors, E_F cannot be shifted into the extended states to produce a degenerate semiconductor. In chalcogenide glasses, E_F is pinned by $-U$ defect centers and in *a*-Si:H doping is limited by the self-compensating increase in dangling-bond defect concentration. In amorphous (and microcrystalline) In_2O_{3-x}, however, the electron concentration is governed by the oxygen deficiency x and E_F can be shifted through the mobility edge E_c to produce a metallic state. A series of $\sigma(T)$ curves obtained by stepwise oxidizing a conducting film of *a*-In_2O_{3-x} is shown in Fig. 11. The lower conductivity curves exhibit the Meyer–Neldel relation with $\sigma_{00} \approx 0.1\ \text{V}^{-1}\ \text{cm}^{-1}$ and $T_m \approx 600$ K similar to *a*-Si:H. The films are metallic for $\sigma \geq 150\ \Omega^{-1}\ \text{cm}^{-1}$ (Bellingham *et al.*, 1991; Pashmakov *et al.*, 1993).

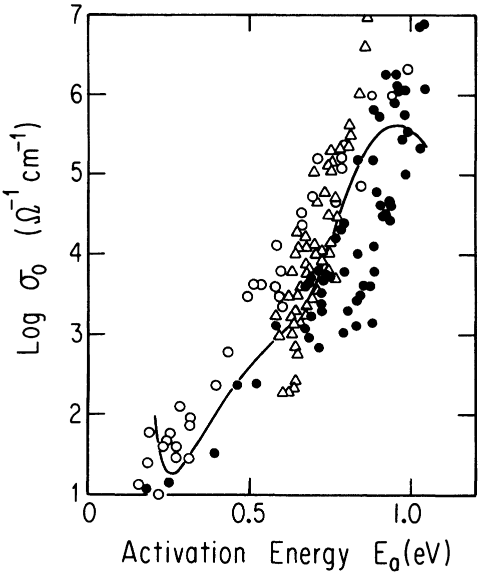

FIG. 10. Meyer–Neldel relation between conductivity prefactor σ_0 and activation energy E_a as defined by Eq. (29) in *a*-Si:H. Data from different laboratories. After Tanielian (1982). Solid curve was calculated by Overhof and Thomas (1989).

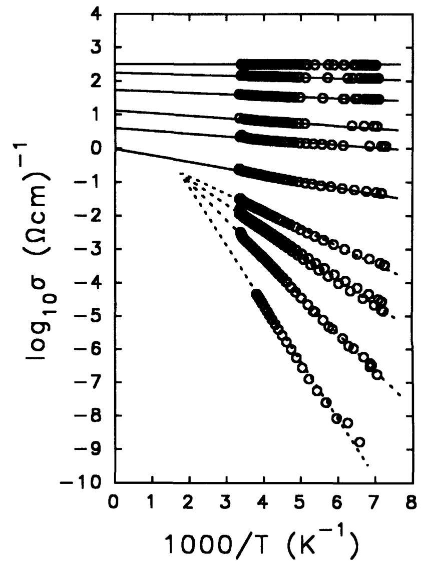

FIG. 11. Temperature dependence of the conductivity of a 25-nm-thick In_2O_{3-x} sample, first made metallic by photoreduction and then reoxidized in steps. Lower σ curves extrapolate to a common point at T_m = 590 K. After Pashmakov *et al.* (1993).

2.3 Hall Effect

When a magnetic field B_y is applied normal to a current density j_x, an electric Hall field E_z appears perpendicular to both the current and the magnetic field driven by the Lorentz force. The Hall coefficient R relates these quantities:

$$E_z = Rj_xB_y. \tag{35}$$

Since the conductivity is proportional to the product of the carrier concentration and mobility, studying both $\sigma(T)$ and $R(T)$ yields the temperature dependence of the carrier concentration and the mobility. In amorphous semiconductors (with one exception) the Hall effect is anomalous and does not give the desired information (Fritzsche, 1974; LeComber *et al.*, 1977; Kakalios, 1989). It has the opposite sign from what is expected from doping and thermopower measurements and the magnitude is about 50 times smaller than expected from conductivity measurements. The origin of this anomaly is still a mystery. The Hall-effect anomaly does not seem to be caused by the small value of the carrier mobilities in the extended band states because crystalline semiconductors with similar or smaller mobilities behave normally.

In contrast to the other noncrystalline semiconductors, the metal-oxide films listed under 2.A. in Table 1 show a normal sign and magnitude of the Hall effect. The carriers are electrons and their mobility μ = 5–20 cm^2/(V s) is similar to that of electrons in *a*-Si:H or of holes in chalcogenide glasses. This shows that the Hall-effect sign anomaly of the latter is not due to the small value of the mobility. The carrier concentration determined from R in *a*-In_2O_{3-x} agrees with the optical free-carrier absorption (Bellingham *et al.*, 1990). Moreover, the Hall mobilities of *a*-In_2O_{3-x} and of $AgSbO_3$ are nearly the same in the amorphous and in the crystalline states (Bellingham et al., 1990; Yasukawa *et al.*, 1995).

2.4 Thermopower

A temperature gradient across a conductor causes the charge carriers at one end to be more energetic than at the other. This causes a drift or diffusion of the energetic carriers from the hotter to the cooler end until an opposing electric field, the thermopower, prevents further charge flow and establishes a steady state. The sign of the potential at the cold end is therefore the sign of the dominant charge carriers.

The thermopower or Seebeck coefficient is $S = \Delta V/\Delta T$ where ΔV is the potential difference measured at two points of the sample having a temperature difference ΔT. The interpretation of the thermopower is the same for amorphous and crystalline materials. The Onsager relation associates the thermopower S with the Peltier coefficient:

$$\Pi = ST. \tag{36}$$

This coefficient has the simple meaning that it is the energy transported by the charge carriers weighted by their contribution to the electrical conductivity. The energy transported is measured relative to the Fermi energy E_F. One therefore can write quite generally (Fritzsche, 1971)

$$\Pi = ST = -\frac{1}{e}\frac{\int (E - E_F)\sigma(E)dE}{\int \sigma(E)dE}. \tag{37}$$

All contributions to Eq. (37) from energy states above E_F are from electrons; contributions from states below E_F have the opposite sign and hence are from holes. S has the sign of the dominant charge carrier. The validity of Eq. (37) is limited by the same assumptions mentioned in connection with Eqs. (22) and (23).

For electron conduction above the electron mobility edge E_c, Eq. (37) reduces to

$$S = -k/e\left[\frac{E_C - E_F}{kT} + A\right]. \tag{38}$$

The first term in this expression describes the thermopower when all electrons have the same energy E_c. However, since the electrons have an average energy of about $\frac{3}{2}kT$ above E_c, there is the correction term $A \approx \frac{3}{2}$.

For hole conduction near the mobility edge E_V, the sign of S is positive as E_C in Eq. (38) is replaced by E_V. According to Eqs. (27) and (38), the temperature dependences of $S(T)$ and of $\ln\sigma(T)$ are both governed by the energy separation of E_F from the nearest mobility edge. Hence, the quantity $Q = \ln\sigma + (e/k)S$ should be a constant. Experimentally one finds, however,

$$Q = Q_0 - E_Q/kT \quad (39)$$

with $0.05 < E_Q < 0.2$ eV for *a*-Si:H, *a*-Ge:H, and the chalcogenide glasses. This additional E_Q term depends on preparation parameters and heat treatment and may be due to long-range random potential fluctuations (Beyer and Overhof, 1979; Overhof and Beyer, 1984).

2.5 Conductance Fluctuations in Hydrogenated Amorphous Silicon

Studies of conductance fluctuations have been employed to investigate aspects of defect kinetics and electronic properties in amorphous semiconductors that are not revealed in bulk transport measurements. Observations of random telegraph switching noise (RTSN) in semiconducting devices with small effective volumes and constraining geometries at cryogenic temperatures, which limit the number of free charge carriers in the material, have been attributed to carrier capture and emission processes of individual defects. The recent observation of RTSN, that is, sharp jumps between discrete resistance levels, in coplanar and transverse current measurements in macroscopic volumes of *a*-Si:H at and above room temperature was therefore quite surprising (Rogers *et al.*, 1986; Arce *et al.*, 1989; Choi *et al.*, 1990; Parman *et al.*, 1991). Figure 12 shows a time trace of the resistance of an *n*-type (10^{-3} PH_3/SiH_4) doped *a*-Si:H film with coplanar electrodes (width 0.2 mm, separation 1 mm) at 400 K (Parman *et al.*, 1991). Switching events have been observed for time scales ranging from a few milliseconds to several tens of seconds, with fractional resistance jumps varying from $\Delta R/R \sim 10^{-2}$ to 10^{-4}. Comparison of two-probe and four-probe measurements confirms that the RTSN is not a contact effect.

If the RTSN in *a*-Si:H arose from charge trapping in defects, then for the macroscopic film in Fig. 12 the magnitude of the resistance jumps implies that 10^5 electrons are simultaneously trapped and released. Alternatively, it has been suggested that the RTSN reflects the presence of inhomogeneous current filaments, which arise from the long- or medium-range compositional disorder or potential fluctuations in the amorphous semiconductor. The switching noise indicates that the resistance of these microchannels is time dependent, possibly as a result of hydrogen or electronic motion. If the fractional resistance jump is associated with a change in the volume of a microchannel available to carry current, then the cross-sectional diameter of a current filament is estimated to be on the order of ~1–10 μm, which is consistent with microchannel size estimates based on RTSN studies of transverse current measurements (Arce *et al.*, 1989; Choi *et al.*, 1990; Teuschler *et al.*, 1993). However, it is not obvious how hydrogen motion could influence the conductance of a channel this large, nor how a smaller filament could influence the electronic conductivity of a larger region. If the conductance fluctuations in amorphous silicon do result from inhomogeneous current filaments, then, unlike other

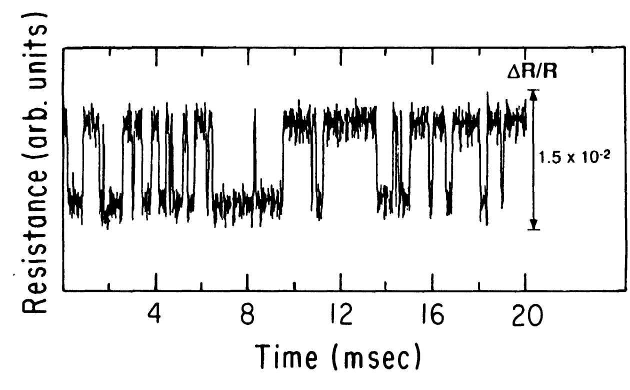

FIG. 12. Resistance fluctuations in *a*-Si:H at 400 K. After Parman *et al.* (1991).

transport measurements, noise studies (in essence an electronic measurement) can reveal information concerning disorder over length scales that are much longer than the inelastic scattering length.

3. PHOTOEXCITATION AND RECOMBINATION

Illumination with band-gap light, $h\nu \geq E_g$, creates a nonequilibrium condition in which excess electrons and holes populate extended and localized states. Radiative as well as nonradiative recombination of excess electron–hole pairs tends to restore equilibrium. The excess carriers give rise to a photoconductivity that can exceed the dark conductivity by a factor of many million, which is used for light detectors, for detector arrays, and in electrophotographic copiers. The excess carriers convert light into electricity in photovoltaic cells and solar panels. The process of light emission during radiative recombination is called photoluminescence.

The thermalization and recombination processes involve a sequence of steps, illustrated in Fig. 13, which are distinguished by different time scales. Thermalization in extended states proceeds with emission of phonons at a cooling rate of about 2 eV/ps, much faster than in crystals, because the crystal-momentum selection rules do not apply to noncrystalline materials. Thermalization into localized states occurs at a rate $\nu_0 \sim 10^{13}$ s^{-1}, a phonon frequency. Transitions between localized states separated by R occur by phonon-assisted hopping at a rate described by Eq. (25).

3.1 Radiative and Nonradiative Recombination

Figure 13 illustrates a number of recombination processes. They occur from the extended states or from the localized tail states. Since the recombining electron–hole pairs have energies between $E_g/2$ and E_g, an appreciable energy has to be dissipated in the process. This occurs either radiatively by emission of a photon or nonradiatively by an Auger process or the emission of many phonons. Radiative recombination (photoluminescence) is favored at low temperatures and for larger recombination energies. However, the process cannot be faster than $\tau_0 \sim 10^{-8}$ s, the dipole radiation lifetime. Since the carriers get trapped by localized tail states in about $\nu_0 \sim 10^{-13}$ s, one does not observe recombination between extended states. Radiative recombination between localized electrons and holes separated a distance R occurs by tunneling at a rate

$$\nu_r = \tau_0^{-1} \exp(-2R/a). \tag{40}$$

As this rate is relatively slow, nonradiative recombination processes dominate in materials with large electron–phonon interactions. In hydrogenated amorphous silicon, for instance, the Si dangling-bond defects are effective nonradiative recombination centers. In chalcogenide glasses it is possible that the recombination energy is dissipated by a brief

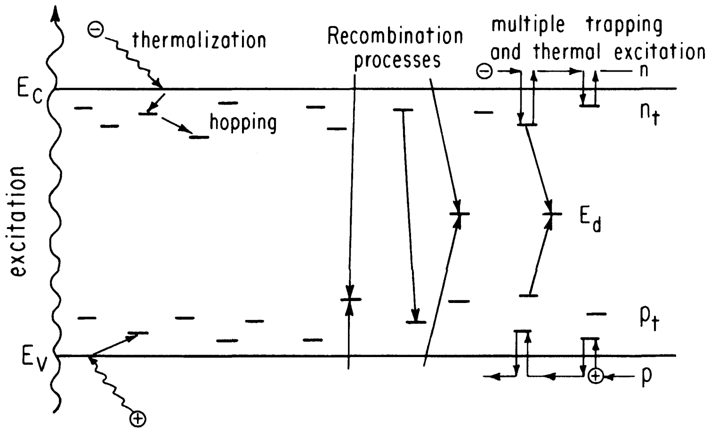

FIG. 13. Schematic representation of excitation, thermalization, and recombination of charge carriers.

configuration change, which is modeled by Street (1977) and illustrated in Fig. 14. This model suggests that an electron or a hole localized in tail states changes temporarily the valency of nearby atoms. This causes a distortion toward the configuration of an intimate valence-alternation pair (see Sec. 1.5.3). In this example the pair consists of overcoordinated and undercoordinated Se atoms. Path I in Fig. 14 corresponds to recombination of the transient exciton, path II to the photoinduced creation of a metastable defect pair, a process to be discussed in Sec. 4.1. Auger processes are probably responsible for the fast nonradiative recombination (a few picoseconds) of very high photocarrier concentrations (10^{19}–10^{21} cm^{-3}) excited by ultrashort laser pulses.

3.2 Geminate Recombination

Recombination is called geminate when the very same electron–hole pair recombines as was created by photoexcitation. It occurs when the pair does not drift far apart, i.e., at low temperatures and low light intensities and in materials having small thermal-carrier concentrations. The geminate recombination lifetime is independent of excitation intensity because the fate of each pair is independent of the others. This distinguishes it from nongeminate recombination.

Onsager (1938) calculated the probability for an electron to escape the attractive Coulomb force of its hole. This model is less applicable to amorphous semiconductors because the electron need not overcome the Coulomb potential but can escape by hopping to localized states in the Coulomb well.

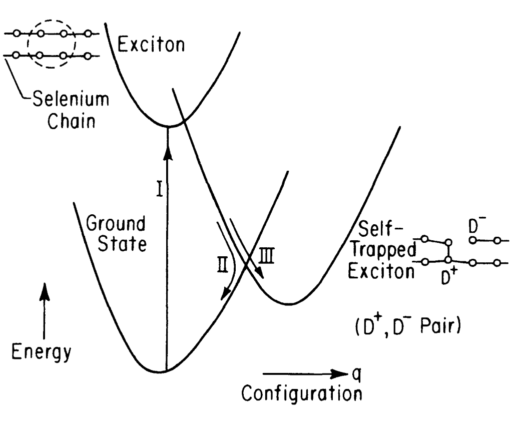

FIG. 14. Energy-configuration diagram showing (I) photoexcitation, (II) recombination via a transient exciton, and (III) the formation of a self-trapped exciton. After Street (1977).

3.3 Photoluminescence

3.3.1 Steady-State PL The light emitted by the radiative recombination processes is called photoluminescence. Figure 15 shows for three chalcogenide glasses the photoluminescence spectrum (PL) as well as the spectrum of photoexcitation light (PLE) that produces the photoluminescence. It is understandable that the PLE spectrum starts with the spectrum of optical absorption α and falls off at higher photon energies where the excitation is limited to a surface layer. More remarkable is the rather large Stokes shift of the PL to about one-half the energy gap. This Stokes-shift energy is understood to be dissipated by a lattice relaxation that is closely related to the relaxation that produces a negative effective correlation energy

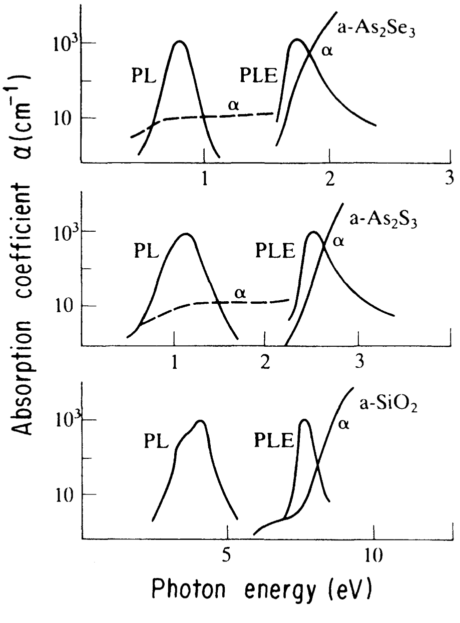

FIG. 15. Photoluminescence (PL) and photoluminescence excitation (PLE) spectra of three chalcogenide glasses. The solid curve α represents light-induced metastable absorption. After Gee and Kastner (1979).

($-U$) of the coordination defects in these glasses (see Secs. 1.4 and 1.5.3).

Prolonged light exposure of chalcogenide glasses decreases the PL efficiency at low temperatures. This PL fatigue is accompanied by the growth of a subgap absorption shown by the dashed lines in Fig. 15. These light-induced effects and the appearance of an ESR signal disappear upon warming the samples to room temperature. These effects arise at low temperatures when an increasing number of normally charged $-U$ defects become neutral and paramagnetic by trapping photoexcited carriers.

In *a*-Si:H, excitation with band-gap light produces a PL peak at low temperatures near $h\nu = 1.3$ eV with a full width of 0.3 eV. Although the PL energy is again lower than the excitation energy, the difference is believed to be the energy loss of the carriers as they hop down into tail states before recombining radiatively. The defects are predominantly nonradiative recombination centers and produce only a weak PL at 0.9 eV.

3.3.2 Time-Resolved PL After a short (less than 10^{-9} s) excitation pulse, the decay of PL proceeds with diminishing intensity to about 10^{-2} s. The PL decay of *a*-Si:H is shown as an example in Fig. 16. The lower panel shows the distribution of radiative lifetimes calculated from the upper-panel data. One notices a broad (two orders of magnitude wide) distribution of lifetimes centered near $\tau = 10^{-4}$ s or 10^{-5} s depending on the temperature. The broad distribution of lifetimes suggests that the electron–hole pairs recombine radiatively by tunneling over separations R according to Eq. (39). With an increase in temperature the PL intensity decreases because of thermally assisted tunneling of carriers to nonradiative recombination centers. This affects in particular long-lived pairs causing in Fig. 16 the shift of the lifetime distribution to shorter times (Tsang and Street, 1979).

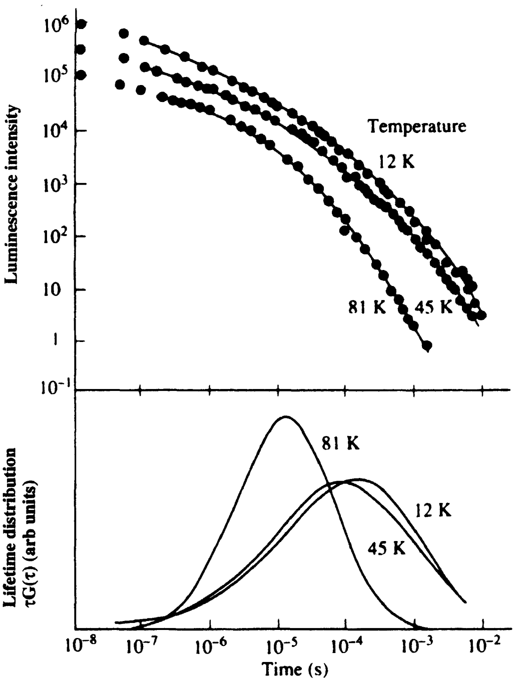

FIG. 16. Decay of photoluminescence of *a*-Si:H following a short excitation pulse. The lower curves are the distribution of luminescence lifetimes calculated from the data. After Tsang and Street (1979).

3.4 Photoconductivity

3.4.1 Primary and Secondary Photoconductivity Ordinary photoconductivity is measured by using Ohmic electrical contacts to the semiconductor, which pass both electrons and holes. The material is supposed to be free of space charge and the photocurrent is proportional to the applied electric field. This is the secondary photoconductivity and the topic of the following section. If the contacts are blocking, one deals with the primary photocurrent. Examples are charge collection in *p-i-n* solar cells or structures with Schottky-barrier metal contacts. The primary photocurrent is mentioned in Sec. 3.5 on time-of-flight drift mobility. The recombination processes are best studied using the secondary photoconductivity because space charges, which distort the applied field, are absent.

3.4.2 Steady-State Photoconductivity In steady state the photoconductivity does not change with time because the constant excitation rate is exactly balanced by the recombination rate. A steady-state nonequilibrium distribution function can be defined (Simmons and Taylor, 1971), which when substituted into Eqs. (22) and (23) yields the photoconductivity σ_p. For electrons in extended states having a mobility μ_C one can write (with equivalent expressions for holes)

$$\sigma_p = en\mu_C, \tag{41}$$

$$n = \eta_n G\tau_n. \tag{42}$$

Here G is the electron–hole pair-generation rate (in units of $\text{cm}^{-3}\ \text{s}^{-1}$) and τ_n is the electron recombination lifetime. The quantum efficiency η_n is the fraction of electrons that not only get excited to conduction-band states above E_c in Fig. 1 or 13 but that also recombine nongeminately. The quantum efficiency is smaller than unity when carriers get excited to localized states or when pairs recombine geminately and hence are unable to transport a net charge.

The electron recombination lifetime τ_n usually gets shorter with increasing generation rate G because more holes become available for recombination. Consequently Eqs. (41) and (42) yield a sublinear dependence on light intensity,

$$\sigma_p \propto G^\gamma \quad \text{with} \quad \gamma \leq 1. \tag{43}$$

As is illustrated in Fig. 17 the temperature dependence of the photoconductivity has essentially the same shape for many amorphous semiconductors, even though the magnitudes differ. The small values found in chalcogenide glasses may be attributed to a small quantum efficiency due to geminate recombination. The rapid rise of σ_p with increasing T is attributed to the rise in free-carrier concentration n relative to the trapped-carrier concentration n_t. Major factors that decrease the magnitude of σ_p for a given generation rate G are the defect concentration, the concentration of localized tail states, geminate recombination, potential fluctuations, and compositional heterogeneities.

The unusual decrease and minimum in $\sigma_p(T)$ observed in Fig. 17 for *a*-Si:H near room temperature is called thermal quenching of photoconductivity (Rose, 1963). In this material it can be attributed to the opening of a new recombination channel as, with increasing T, the trapped holes become more mobile and enhance recombination through defect states. In this region one observes supralinearity, $\gamma \geq 1$ in Eq. (43).

The temperature-independent photoconductivity at very low T cannot be described by Eqs. (41) and (42). It is due to the hopping of photocarriers down to lower-energy localized tail states. This energy-loss hopping attains an overall drift component in the applied field, which results in photoconductivity (Shklovskii *et al.*, 1989).

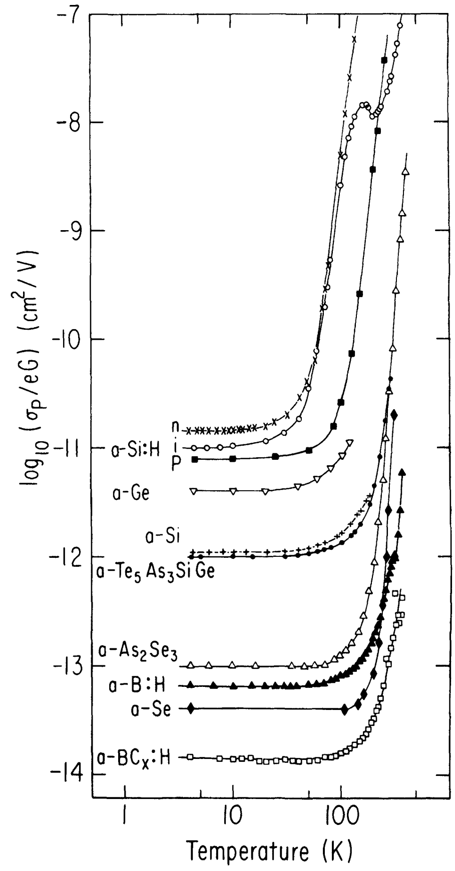

FIG. 17. Temperature dependence of the photoconductivity normalized by the electron charge and the excitation rate for different noncrystalline semiconductors. After Johanson *et al.* (1989).

3.4.3 Photoconductivity Decay Upon terminating illumination the photocurrent decays from its steady-state value as the nonequilibrium distribution of the carrier population returns via thermal excitation and recombination to the equilibrium distribution $f(E)$. The initial logarithmic decay defines the decay time τ_D as

$$\left.\frac{d\ln\sigma_p}{dt}\right|_{t=0} = \frac{1}{\tau_D}. \tag{44}$$

A measurement of τ_D and of $\mu_C\tau_n$ from the steady-state photoconductivity allows one to calculate the electron-drift mobility μ_d using the relation (Hoheisel *et al.*, 1988)

$$\mu_C \tau_n = \mu_d \tau_D. \tag{45}$$

By measuring the decay of the photocurrent after a short laser pulse (1–10 ns), instead of the steady-state photocurrent, one gains information about the product of the band mobility and the deep trapping or recombination lifetime of the dominant photocarriers (Antoniadis and Schiff, 1992).

3.4.4 Modulated Photoconductivity

The modulated photocurrent method (Oheda, 1981) is used to determine the density of localized states in a certain energy region of the mobility gap. The modulated photoconductivity $\sigma_m(\omega)$ is measured in response to a sinusoidal component of the electron–hole pair-generation rate $G_m(\omega)$. The density of states $g(\Delta E)$ at the energy

$$\Delta E(\omega) = kT \ln(\nu_0/\omega) \tag{46}$$

below the nearest mobility edge is obtained from the magnitude of $\sigma_m(\omega)$ and the phase shift ϕ of the photocurrent relative to the light modulation. According to Brüggemann *et al.* (1990) the relation is

$$g(\Delta E) = \frac{2}{\pi} g(E_C) \frac{\omega}{\nu_0} \left[\frac{e\mu_C G_m(\omega)}{\omega \sigma_m(\omega)} \sin\phi - 1 \right], \tag{47}$$

where $g(E_C)$ and μ_C are the density of states and the mobility at the electron mobility edge E_C. Here it is assumed, as in Eq. (27), that electrons contribute more to the photocurrent than holes. Equation (46) is derived from Eq. (21) by relating the modulation frequency ω with the inverse thermal release time τ_e^{-1}. Their relation is the basis of this experiment because photocarriers at energy $\Delta E(\omega)$ are thermally released and retrapped synchronously with the modulation frequency albeit with a phase lag ϕ.

The gap energy ΔE of Eq. (46) can be scanned by varying either the frequency ω or the temperature T. A larger energy range is explored by changing both. An example of the density of gap states of an *a*-Si:H sample determined by the modulated photoconductivity method is shown in Fig. 18. The energy separation ΔE from the electron mobility edge was determined for each T and ω from Eq. (46). The temperature scale for the 8-Hz modulation measurements is shown on top of the figure. The nearly identical density-of-states distributions obtained from the three modulation frequencies indicate the correctness of the basic assumptions. The peak at ΔE = 0.65 eV is due to dangling-bond defect states.

3.5 Drift Mobility

As sketched on the right-hand side of Fig. 13, each carrier is frequently trapped into localized tail states and thermally emitted back

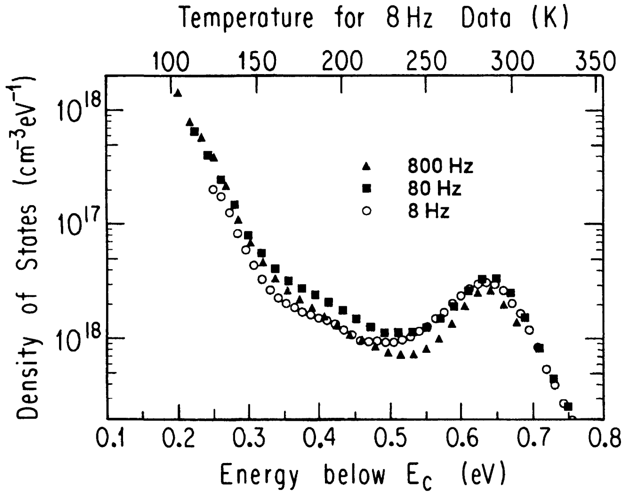

FIG. 18. Density of gap states of *a*-Si:H determined with the modulated photoconductivity method by scanning the temperature for three modulation frequencies. The energy distance ΔE from the electron mobility edge was obtained from Eq. (46) by accounting for a small temperature variation of ν_0. After Zhong and Cohen (1993).

into the extended states. The long-time average mobility of an electron, for example, is not its extended-state mobility μ_C but a smaller drift mobility μ_d which depends on the ratio of the time the electron is free τ_f to the time it is trapped τ_t as

$$\mu_d = \mu_C \tau_f/(\tau_f + \tau_t). \tag{48}$$

One way of measuring μ_d is with the time-of-flight apparatus sketched in Fig. 19(a). An excitation light pulse creates electron–hole pairs near the left transparent contact to the sample. Its photon energy is sufficiently high to have a short absorption depth. Depending on the sign of the applied field F, either electrons or holes are drawn through the sample. In the example of Fig. 19(a) the photoexcited holes drift with an average velocity

$$v_d = \mu_d F \tag{49}$$

through the length L of the sample toward the negative electrode. The drift current I_d is recorded as a function of time. The current I_d should persist until the carriers arrive at the negative electrode after a transit time

$$\tau_T = L/v_d, \tag{50}$$

which with Eq. (49) determines μ_d. In most cases, however, the observed $I_d(t)$ is so featureless that τ_T cannot be identified. Scher and Montroll (1975) discovered that this is to be expected when the motion of the carriers is subjected to a broad distribution of trap release times as is the case here.

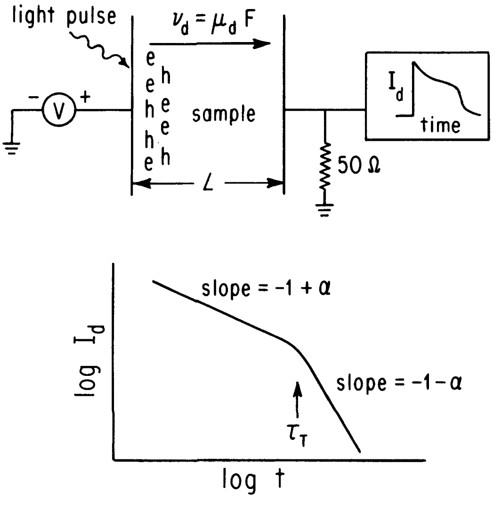

FIG. 19. (a) Sketch of time-of-flight apparatus and (b) time dependence of dispersive drift current.

They further found that τ_T is defined by the position of a break in the curve when $I_d(t)$ is plotted on a logarithmic plot as shown in Fig. 19(b). This transport is called dispersive and the parameter α obtained from the slopes of the $\log I_d$ vs $\log t$ curves is called the dispersion parameter.

A characteristic feature of dispersive transport is that μ_d depends on the measurement time. The reason is the following. The trapping time τ_t in Eq. (48) increases exponentially with increasing binding energy of the tail-state traps,

$$\tau_t = \nu_0^{-1} \exp[(E_C - E)/kT]. \tag{51}$$

As a consequence, when the measuring time is increased, more deep traps can participate in trapping and releasing drifting electrons. Hence, τ_t increases thereby reducing μ_d. In the time-of-flight experiment, the measurement time is τ_T, which depends on the field F according to Eqs. (49) and (50). This in brief explains the observed field dependence of μ_d shown by the data in Figs. 20(a) and 20(b). Plotted here is the temperature dependence of μ_d of electrons and holes in *a*-Si:H. These curves and the dispersion parameter α can be fitted with assumed density of states distributions $g(E)$ of localized tail states. Good fits are obtained with $g(E)$ increasing exponentially toward E_C and E_V, respectively, as sketched in Fig. 1. The logarithmic slopes of these tails are $E_0 \sim 0.05$ eV for holes and 0.025 eV for electrons (Tiedje and Rose, 1980; Orenstein and Kastner, 1981).

3.5.1 Deep Trapping

In the absence of deep traps or defect states one should eventually collect the total charge Q_0 generated by the light pulse in a time-of-flight experiment. On the other hand, when carriers are lost to deep traps at a rate $1/\tau$, a smaller Q is collected. According to Hecht (1932)

$$\frac{Q}{Q_0} = \frac{\mu_d \tau F}{L}\left[1 - \exp\left(-\frac{L}{\mu_d \tau F}\right)\right]. \tag{52}$$

A measurement of charge collection Q/Q_0 as a function of applied field F in Fig. 19(a)

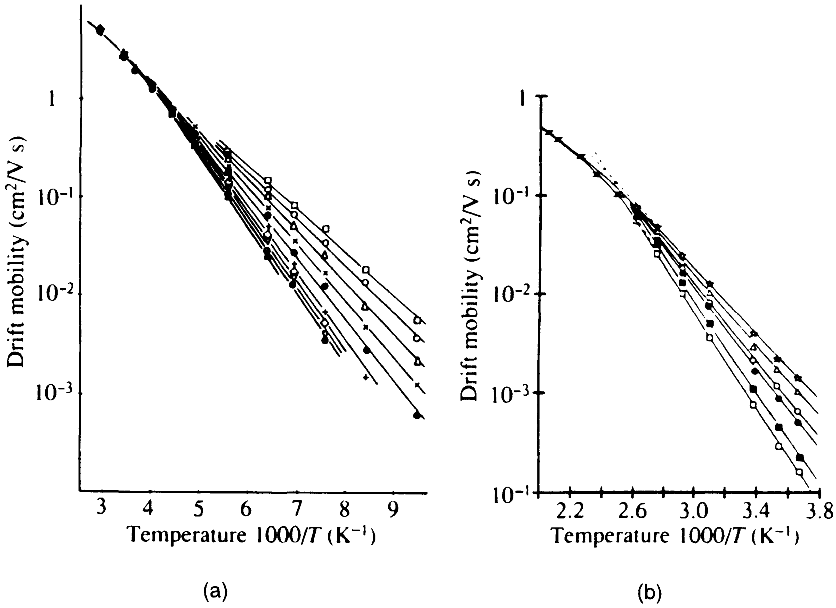

FIG. 20. Temperature dependence of the (a) electron and (b) hole drift mobilities of *a*-Si:H at different electric fields increasing from the bottom from 5 × 10^2 to 5 × 10^4 V/cm. The field dependence is caused by dispersion. After Street (1991b).

yields the product $\mu_d\tau$ of drift mobility and deep trapping time. Figure 21 shows $\mu_d\tau$ at 300 K for electrons and holes in both *n*-type and *p*-type as well as compensated and intrinsic *a*-Si:H (Street *et al.*, 1983). The results confirm the following picture. The dangling-bond defect concentration increases with *n*-type or *p*-type doping according to Eqs. (8) and (9). In *n*-type samples these defects are negatively charged D^- and hence can trap holes but not electrons. The results of Fig. 21 on *n*-type samples confirm this: $\mu_d\tau$ of holes decreases with *n*-type doping but $\mu_d\tau$ of electrons remains constant at 10^{-6} cm² V⁻¹. The opposite happens in *p*-type samples where the defects become D^+ and can trap electrons but not holes. In intrinsic *a*-Si:H, in contrast, many defects are neutral so that $\mu_d\tau$ of both electrons and holes are reduced with increasing defect concentration. The results also indicate that the drift mobility μ_d and thus the tail-state slope E_0 are not affected by doping.

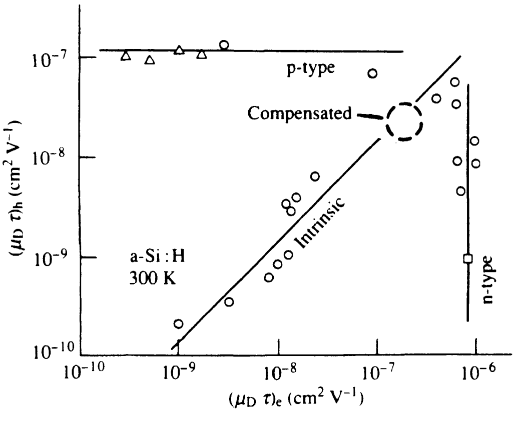

FIG. 21. Product of band mobility and deep trapping lifetime of electrons and holes for undoped, doped, and compensated *a*-Si:H. After Street *et al.* (1983).

3.6 Spin-Dependent Recombination

Recombination of an electron–hole pair is allowed when the spins of the electron and the hole are antiparallel and forbidden when they are parallel because the final state of the recombination process must have zero spin. Geminate recombinations proceed unhindered unless spin-flip processes occur during thermalization. Nongeminate pairs, on the other hand, have random spin alignments. The recombination rates can, therefore, be changed by perturbing the thermal spin alignments by a magnetic field or by changing the population of Zeeman-split levels resonantly with microwave fields (ESR-induced transitions). Since different g values are associated with tail-state electrons, tail-state holes, and neutral defect states, the role of each of these entities in the recombination event can be studied in detail.

A magnetic field produces a spin polarization P_e and P_h of the electrons and holes, respectively, which decreases their recombination rate W:

$$W = W_0(1 - P_e P_h). \tag{53}$$

At resonance, the microwave field tends to equalize the populations of the spin states and to reduce the spin polarizations. This enhances the photoluminescence and reduces the photoconductivity. At 300 K a spin polarization of $P_e P_h \sim 10^{-6}$ is expected for $g \sim 2.0$ spins and a field of 3000 G. Experimentally one finds, however, changes in photoluminescence, photoconductivity, and photoinduced absorption that are several orders of magnitude larger than this value. This enhancement is caused by the fact that a net surplus of spin-aligned electron–hole pairs accumulates because their recombination is forbidden. No macroscopic magnetization occurs, however, because the orientations of the spin-aligned pairs remain random. A spin flip of either partner of the spin-aligned pair at ESR resonance induces recombination. This large spin correlation between recombining electron–hole pairs enhances the sensitivity of optically detected magnetic resonance far beyond the sensitivity of ordinary ESR (Solomon, 1979; Morigaki, 1984).

4. METASTABILITY

Essentially all noncrystalline semiconductors exhibit changes in their electronic and optical properties when they are subjected to electron bombardment, strong light, or other excitations. These changes are metastable and reversible in that the original properties can be restored by annealing the material at a characteristic equilibration temperature T_E that usually lies between 150 and 300 °C.

Metastabilities are well known in glasses whose structure, density, and general properties can be changed by quenching from high temperatures where they are soft or liquid to room temperature where they are solid. The original properties can be restored by annealing near the glass-transition temperature. The metastabilities of noncrystalline semiconductors are quite different, however. They are produced by electronic excitations below the equilibration temperature and others occur in materials such as a-Si:H and a-In_2O_3, which are amorphous films and do not exhibit a softening or glass transition. It appears that the local dissipation of the recombination energy of an electron–hole pair may give rise to a metastable change in bonding configuration.

4.1 Chalcogenide Glasses

4.1.1 Photostructural Changes Exposure to light or other radiation that excites electron–hole pairs produces metastable structural changes in nearly all chalcogenide glasses (DeNeufville *et al.*, 1974; Ka. Tanaka, 1976, 1981; Owen *et al.*, 1985; Ke. Tanaka, 1990; Pfeiffer *et al.*, 1991). As a consequence of light-induced changes in atomic configurations, a variety of physical and chemical properties are altered such as density, hardness, elastic constants, electronic transport and optical properties, and chemical solubility, which allows the use of these materials for photolithography.

A striking light-induced effect is called photodarkening, a shift ΔE of the optical gap E_g and the Urbach edge (see Sec. 1.7.1) toward lower energies. The magnitude of ΔE produced at different illumination temperatures T_i relative to T_g is shown for several glasses in Fig. 22. Glasses containing S show the largest and those containing Te the smallest shift.

Photostructural changes are largest at low temperatures and are not present for illumination near T_g. They are caused by electron–hole pair-producing radiation and are attributed to the accumulated effects of local bonding changes caused by nonradiative electron–hole pair recombinations. More than 10^{26} cm^{-3} recombinations are needed to bring the photostructural changes to saturation. The metastable changes can be eliminated by incorporating certain elements such as Cu into the glass (Liu and Taylor,

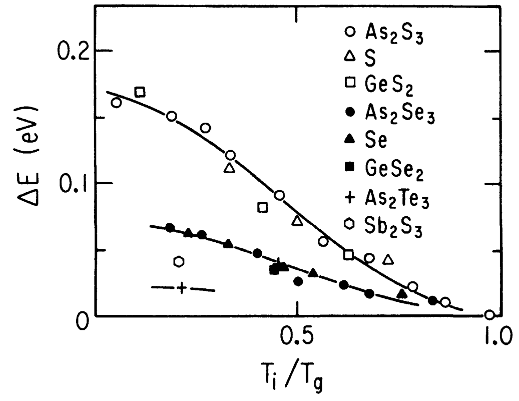

FIG. 22. Photoinduced shift of the absorption edge as a function of the illumination temperature T_i normalized by the glass-transition temperature for several chalcogenide glasses. After Ke. Tanaka (1983).

1987), presumably because these elements provide recombination channels that do not yield structural changes.

Some models relate photostructural changes with specific sites or defects (Ke. Tanaka, 1990), while other models attribute them to bonding changes of the network and a resulting decrease in medium-range order (Elliott, 1986; Pfeiffer *et al.*, 1991; Fritzsche, 1993a). The photodarkening shift ΔE of Fig. 22 is attributed to a broadening of the valence band due to increased lone-pair orbital interactions in the metastable configuration. This accounts for the dependence of the photodarkening ΔE on the chalcogen element (see Fig. 22). The broadening of the valence band giving rise to photodarkening has been observed by x-ray photoelectron spectroscopy (Kolobov *et al.*, 1991).

4.1.2 Metastable Defects The concentration of $-U$ defects in some chalcogenide glasses can be increased more than tenfold by exposure to band-gap light (Biegelsen and Street, 1980). This results in an increase in the subgap absorption and in the light-induced ESR signal, and a decrease in photoluminescence. The equilibrium concentration of $-U$ defects is reestablished by annealing at T_g. The metastable defects are valence-alternation pairs, which may be thought of being produced via path III of the nonradiative recombination scheme of Fig. 14 (Street, 1977).

4.1.3 Optical Anisotropies Exposure of chalcogenide glasses to linearly or circularly polarized band-gap light produces an anisotropy of the dielectric tensor that results in linear or circular dichroism as well as in linear or circular birefringence, respectively (Zhdanov *et al.*, 1979; Lyubin and Tikhomirov, 1989, 1991). These photoinduced anisotropies can be reversibly reproduced after annealing, and the axes defining the optical anisotropy can be turned by turning the polarization direction of the exposing light. The induced dichroism has normally the sign $\alpha_{\parallel} < \alpha_{\perp}$ and the induced birefringence yields $n_{\parallel} < n_{\perp}$, where $\parallel$ and $\perp$ refer to the relative polarizations of the probe light and inducing light. These effects have been attributed to geminate nonradiative recombination events that produce local changes in bond configurations (Fritzsche, 1993b, 1995).

4.2 *a*-Si:H and Related Materials

4.2.1 Light-Induced Defects In 1977 Staebler and Wronski discovered that prolonged exposure to band-gap light decreases both the dark conductivity and the photoconductivity in doped and undoped *a*-Si:H as shown in Fig. 23. The degraded conductivity state is stable at room temperature but the original values are restored by annealing between 150 and 200 °C. Electron-spin-resonance (ESR) experiments demonstrated that this effect is due to the creation of metastable dangling-bond defects (Dersch *et al.*, 1980), which move the Fermi level toward midgap and act as recombination centers reducing the photoconductivity and photoluminescence. Irradiation with ions (Street *et al.*, 1979), electrons (Voget-Grote *et al.*, 1980), or x rays (Pontuschka *et al.*, 1982) also produces these metastable defects. It is believed that they are created during the recombination of irradiation-produced electron–hole pairs (Guha *et al.*, 1983), because the recombination events produced by a forward-bias current in *a*-Si:H pin devices create the same metastable defects (Street, 1991a). The concentration of excess defects ΔN_D saturates near 10^{17} cm^{-3} in undoped *a*-Si:H. The defect creation is relatively inefficient as more than 10^{26} cm^{-3} recombination events are needed to reach that value. ΔN_D can reach values greater than 10^{18} cm^{-3} in

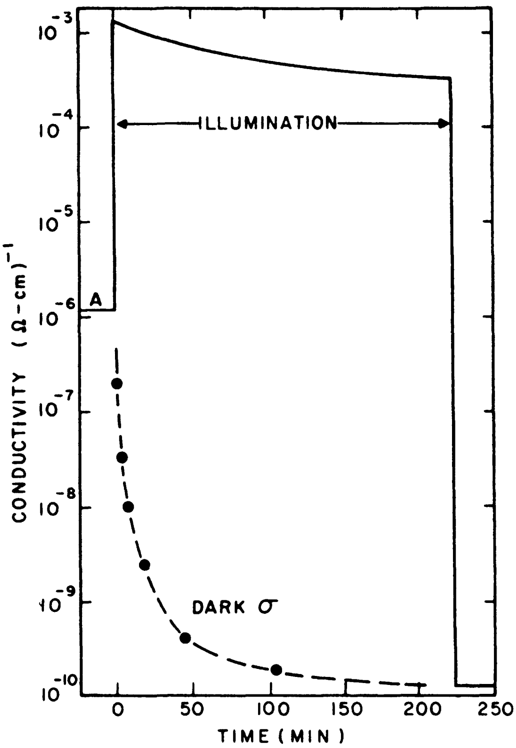

FIG. 23. Decrease of the photoconductivity (solid curve) and of the dark conductivity (dashed curve) of *a*-Si:H with time of exposure to 200-mW/cm^2 filtered tungsten light. After Staebler and Wronski (1980).

doped *a*-Si:H and with short-pulse laser irradiation.

The defect-creation kinetics has been studied by Stutzmann *et al.* (1985). They found that the defect density N_D increases as

$$\Delta N_D(t) = \text{const} \times G^{0.6} t^{0.33} \tag{54}$$

when ΔN_D is larger than twice the concentration of N_D of native defects. For small ΔN_D, a slower growth is observed. This nonlinear dependence on time and electron–hole pair generation-rate G in Eq. (54) was explained by Stutzmann and co-workers (1985) with the following model. Defects are created by nonradiative recombination of electron–hole pairs at spatially correlated tail states that correspond to the bonding and antibonding states of weak Si–Si bonds. The recombination energy of about 1.4 eV, released locally while the bond breaks, allows a bonded hydrogen to insert itself such that the weak bond cannot reform. These are rare events as most recombinations are assumed to proceed via dangling-bond defects. This creation mechanism yields not single defects but pairs of defects. Such pairing has not been confirmed by ESR, but a defect pair may get separated by further bond switching.

At elevated temperatures, annealing of light-induced defects follows a stretched-exponential time dependence (Jackson and Kakalios, 1988),

$$\Delta N_D = N_D(0) \exp[-(t/\tau)^{\beta}]. \tag{55}$$

The parameters of the decay, the dispersion parameter β, and the temperature dependence of the decay time τ are found to be consistent with the hydrogen-glass model in which hydrogen diffusion governs relaxation toward equilibrium (Street *et al.*, 1987; Kakalios, 1991). According to this model, however, light-induced defect creation and annealing should cease with the diffusion of hydrogen. This is not the case. The defects are created with comparable efficiencies at temperatures between 4.2 and 400 K, and annealing of defects created at low temperatures commences near 150 K where hydrogen cannot diffuse (Stradins and Fritzsche, 1994).

There are many open questions regarding the details of the creation mechanism, the role of impurities, and whether native and light-induced defects have the same energy in the gap and whether defect formation occurs preferentially at internal surfaces of voids. The elimination of light-induced degradation is of great importance for the long-term stability of solar cells.

4.2.2 Metastabilities Produced by Rapid Quenching Dangling-bond defects may be created not only by recombination of photoexcited electron–hole pairs but also by recombination of thermally excited carriers at high temperatures (Smith and Wagner, 1985). This process might explain the presence of $N_D \sim 10^{15}$ cm^{-3} native defects in high-quality *a*-Si:H. More defects should be observed when the material is quenched from higher temperatures. Indeed, rapid quenching from above 250 °C gave the first experimental evidence for a thermodynamic equilibration of the defect and dopant reactions Eqs. (8)–(10) (Ast and Brodsky, 1979; Street *et al.*, 1986; Kakalios *et al.*, 1987).

As shown in Fig. 24, the conductivity of doped *a*-Si:H below an equilibration temperature T_E depends on the thermal history, the quench temperature T_Q, and the rate of cooling. This is typical behavior for a glass, which falls out of equilibrium below a glass-transition temperature. The conductivity of rapidly quenched samples is higher than that of slowly cooled samples because reactions (8) and (9) are endothermic with an activation energy for changing the coordinations of the dopant and Si atoms. Hence, a higher dopant concentration is quenched from higher T_Q. The nonequilibrium excess carrier concentration relaxes below T_E. The relaxation process follows the stretched exponential of Eq. (55) with relaxation parameters that suggest that hydrogen diffusion controls the relaxation process (Kakalios *et al.*, 1987; Street *et al.*, 1988a). Further evidence of the role of hydrogen diffusion in the relaxation process is the observation that samples that have a large H-diffusion rate have a low value of T_E and slower stretched-exponential relaxation. The H-diffusion rate increases as E_F moves away from the gap center and is larger for *p*-type than for *n*-type *a*-Si:H, while the opposite trend is observed for T_E. Hydrogen-glass equilibration above a temperature T_E and metastabilities below T_E not only occur in *a*-Si:H but are generally observed in hydrogenated tetrahedral materials and alloys with N, C, and O.

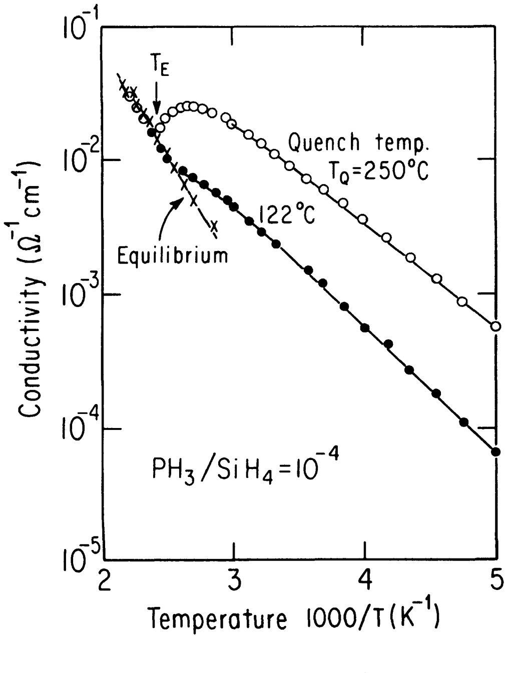

FIG. 24. Temperature dependence of the conductivity of *n*-type *a*-Si:H after annealing (equilibrium) and quenching from different temperatures. The data were taken during subsequent warming. After Street *et al.* (1988b).

4.2.3 Metastabilities Produced by Space Charges The presence of a space-charge layer in junction devices or field-effect transistors produces a change in the defect concentration relative to the equilibrium concentration in the space-charge–free material. Similar effects arise from space charges produced by injected currents (Kruhler *et al.*, 1984), and a metastable, field-induced increase in the doping efficiency was reported by Lang *et al.* (1982b). These effects provide convenient means for studying below T_E the temperature dependence of the creation and annealing kinetics since the space charges can be manipulated with external voltages and measured capacitively (Hepburn *et al.*, 1986; Powell *et al.*, 1987, Jackson and Kakalios, 1989). One finds that the creation and anneal processes are described by stretched exponentials, Eq. (55), with relaxation parameters appropriate for H diffusion in either *n*-type or *p*-type material (Jackson and Kakalios, 1989).

A possible cause for these effects can be that at a given temperature, the equilibrium concentrations of defects and of dopants depend on the formation energies that govern the reactions (8)–(10). These formation energies depend in turn on the Fermi energy E_F since the energy involved in charging the centers is measured relative to E_F. Any change in space charge produces a local change in E_F and causes a relaxation to a new equilibrium state.

4.3 In_2O_{3-x}

Amorphous layers of In_2O_{3-x} exhibit metastabilities that may be related to light-induced changes in the oxygen deficiency *x*, which determines the electron concentration. The family of conductivity curves shown in Fig. 11 was obtained on one sample. The different conductivity states can be reached by photoreduction of the oxide with ultraviolet band-gap light or by reoxidation (Pashmakov *et al.,* 1993). These effects need to be considered when indium oxide or indium tin oxide layers (ITO) are used as transparent conducting electrodes.

5. INTERFACES AND MULTILAYERS

The bulk properties of an amorphous semiconductor can be extracted from measurements on thin films only when the effects of contacts, surfaces, and interfaces are understood. The effects of the latter extend to a depth of the diffusion length or the width of the space-charge region. They can dominate the conduction properties of thin semiconductor films whose defect density is small. Surface states can dominate the electron-spin-resonance signal, measurements of subgap absorption by photothermal-deflection spectroscopy, or the photoemission spectra. For amorphous semiconductors there is no known method to obtain a clean surface equivalent to cleaving crystals in ultrahigh vacuum. Moreover, the atomic layers near the surface and the substrate often differ from the bulk in composition and structure.

These concerns do not apply to measurements on chalcogenide glasses. The failure of attempts to make Schottky-barrier contacts or to bend their bands by a gate voltage in a field-effect geometry confirms the effectiveness of $-U$ defects in pinning the Fermi level (see Sec. 1.4). The following sections refer therefore to high-quality *a*-Si:H.

5.1 Oxide Interface and Gas Adsorption

By measuring the thickness dependence of the spin density and of the subgap defect absorption, Jackson *et al.* (1983) found a surface defect density of about 10^{12} cm^{-2}, which is two orders of magnitude less than the density on freshly cleaved crystalline Si surfaces. This is attributed to a saturation of Si bonds with hydrogen. This surface passivation greatly reduces the oxidation rate of *a*-Si:H relative to crystalline Si (Ponpon and Bourdon, 1982). At 300 K, it takes about 10 hours to grow a 2-Å-thick oxide layer on *a*-Si:H in air. A native oxide layer is negatively charged. A small positive space-charge layer exists adjacent to the surface in *a*-Si:H, which can dominate the sheet conductance of *p*-type samples (Aker and Fritzsche, 1983), while the corresponding electron depletion in *n*-type samples does not significantly alter the film's conductance.

The passivation of surface states with hydrogen makes *a*-Si:H particularly sensitive to effects of gas exposure. Chemisorption of molecules involves a charge transfer to the adsorbing substance. In the absence of surface states, these charges produce a space-charge layer in the amorphous semiconductor, which can dominate its conductance. Adsorbed layers of water and ammonia act as electron donors and increase the conductance of undoped or slightly *n*-type *a*-Si:H by several orders of magnitude. In *p*-type *a*-Si:H these adsorbates compensate for the effect of a negatively charged oxide layer and decrease the conductance manyfold (Tanielian, 1982).

The enhancement of the conductance by many orders of magnitude by controlling the space-charge layer with a gate voltage across a thin insulator is the desired function of *a*-Si:H thin-film transistors, which are used as control elements in flat-panel liquid-crystal displays. Here great care is taken to avoid interface states that would reduce the charge induced in the semiconductor.

5.2 Multilayers

Multilayer structures consist of a large number of ultrathin layers of different semiconductors. A certain sequence of materials can be repeated with any desired periodicity. Because of the layer periodicity they are known as superlattice structures in the field of crystalline semiconductors (see HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR). A much larger variety of multilayers can be prepared with amorphous compared with crystalline materials because no restriction for matching lattice constants of adjacent layers limits the choice of amorphous materials. For example, the Si–N bond in *a*-SiN$_x$:H is 1.7 Å long and the Si–Si bond in *a*-Si:H is 2.3 Å. Nevertheless, these materials can be multilayered with interface state densities less than 10^{11} cm^{-2}. Similarly, *a*-Si:H/*a*-Ge:H multilayers have fewer than 10^{10} cm^{-2} interface states. As a consequence of the periodic change in composition there is a corresponding change in the band gap in these multilayers. In contrast, doping multilayers are made by alternating *n*-type and *p*-type doping in the same semiconductor, *npnp* multilayers, or by inserting an undoped layer in between, producing what are called *nipi* structures. Charge transfer between the differently doped layers causes the internal potential to fluctuate with the periodicity of the doping while the band gap remains constant.

5.2.1 Compositional Multilayers The sharpness of x-ray diffraction patterns indicates that compositional multilayers can be prepared with remarkable flatness and regularity for hundreds of layers (Abeles and Tiedje, 1983). The abruptness of the interfaces depends on the growth sequence; the Si-on-SiO$_x$ interface, for instance, is atomically abrupt while the SiO$_x$-on-Si interface has a 5-Å-wide region graded in oxygen concentration, according to photoemission measurements (Yang *et al.*, 1989). Optical absorption measurements show an increase in the optical gap of *a*-Si:H/*a*-SiN$_x$ multilayers when the silicon-layer thickness is less than 30 Å, suggesting that the electron coherence length at room temperature is of this order if interpreted as a quantum confinement effect. Although a blue-shift absorption edge can also arise from unintentional alloying of N in the *a*-Si:H layers, Hattori *et al.* (1987) reported the observation of steps in the differential absorption spectra of *a*-Si:H/*a*-SiC$_x$:H multilayers with silicon quantum well thicknesses between 20 and 50 Å. A coherence length this long is surprising in view of the short mean free path of 5–7 Å suggested by transport measurements (see Sec. 2.2).

5.2.2 Doping-Modulated Multilayers When *npnp* or *nipi* *a*-Si:H multilayers are illuminated, the photoexcited electron–hole pairs find themselves in regions of strong electric fields due to the series of back-to-back *pn* junctions. If the drift length of the carriers is greater than the layer thickness, these charge carriers are spatially separated to different layers. When illumination is stopped, recombination of the excess charge carriers is inhibited by the built-in electric fields, and the excess electrons in the *n* layers and holes in the *p* layers increase the parallel conductance of the multilayer film. This field-induced enhanced recombination lifetime was predicted by Döhler (1972) and confirmed at low temperatures by Hundhausen *et al.* (1984). Room-temperature exposure creates a persistent photoconductivity that takes days to decay (Kakalios and Fritzsche, 1984). This effect appears to be due to the different rates of metastable defect creation in the *n*-type and *p*-type layers, which raises the Fermi level of the multilayer (Hamed, 1991).

6. CONCLUSIONS

Noncrystalline semiconductors are playing an increasing practical role in applications such as large-area displays, compact optical computer memories, electrophotographic processing, photovoltaics, and optical readout devices. This is partly due to their inexpensive thin-film processing, which extends and complements conventional chip processing by allowing multilayer stacking.

Amorphous and vitreous semiconductors exhibit fascinating new properties that differ fundamentally from those of crystals. Whereas the dopant and defect concentrations remain essentially fixed in crystals after they are grown, these concentrations are subject to thermodynamic equilibration laws above the equilibration or glass-transition temperatures in noncrystalline semiconductors. Moreover, while the electronic energies of defect and dopant states are fixed and have definite values in crystals, these energies depend on the position of the Fermi level in noncrystalline semiconductors because the electrochemical potential affects the creation energy of the defects and dopants. Furthermore, because of the greater structural flexibility of noncrystalline materials and the shorter localization radius of charge carriers in their defect and dopant states, there is generally a larger lattice relaxation associated with a change in charge state in these materials than in crystals. The strength of this relaxation can be so strong in chalcogenide glasses that their defects have a negative effective correlation energy. The thermodynamic equilibrium of defects established in noncrystalline semiconductors can easily be disturbed by external electron–hole pair excitations. The ensuing recombination events produce metastable structural changes that produce a variety of new phenomena rarely observed in crystals. Among these are light-induced metastable defects, light-induced changes in the optical gap, and optical anisotropies induced by linearly or circularly polarized light. These phenomena are uncommon in the more constrained structures of crystals.

We have moved far from viewing glasses and amorphous materials as highly imperfect crystals. The study of noncrystalline semiconductors has greatly enriched and will continue to challenge our understanding of materials.

ACKNOWLEDGMENT

Preparation of this article was supported by the National Science Foundation Grant NSF DMR 9408670.

GLOSSARY

Acceptor: Impurity that can accept an electron from the host semiconductor.

Amphoteric Defect: A defect that can either release or accept an electron from the host semiconductor.

Chalcogen: Element of group VI of the periodic table: O, S, Se, Te.

Conduction Band: The lowest unfilled band of electron states.

Correlation Energy: The Coulomb repulsion energy of two electrons of opposite spin in one state. The effective correlation is the Coulomb repulsion minus the relaxation energy of the surrounding in response to the addition of the second electron.

Dangling Bond: Unsatisfied bond or free radical. In amorphous silicon, for example, the paramagnetic defect state associated with a threefold-coordinated silicon atom.

Defect Pool: The reservoir of potential defect sites.

Density of States: Number of states available for an electron per unit volume and per unit energy range.

Donor: Impurity that can give up an electron.

Doping: Addition of impurities that act as either donors or acceptors and hence can increase the conductivity of a semiconductor sample.

Drift Mobility μ_d: The average velocity of a charge carrier in a unit electric field when the carrier spends a fraction of its time immobilized in a trap.

Extended States: Electron states whose wave functions extend over the volume of the sample.

Geminate Recombination: Recombination of the very same electron–hole pair that was excited by a photon.

Hole: A missing electron in a normally occupied state.

Hopping Conduction: Motion of electron between localized states as a result of the finite overlap of the wave functions of these states. Phonons are emitted or absorbed when the states differ in energy.

Hydrogen-Glass Model: Description of a hydrogenated semiconductor in which hydrogen diffusion governs the kinetics of defect and dopant equilibration.

Localized States: Electron states whose wave function describes an electron localized in a small region of the semiconductor.

Lone-Pair Semiconductor: A semiconductor whose valence band originates from filled nonbonding orbitals.

Mobility Edge: The energy that separates localized from extended states near the band edges of a noncrystalline semiconductor.

Mobility μ: The net velocity developed by an electron or hole as a result of one unit of electric field.

Persistent Photoconductivity: A metastable light-induced increase in the dark conductivity.

Phonon: Quantum of vibrational energy of one of the normal modes into which the random thermal motion of the atoms in a solid may be decomposed.

Photoconductivity: The increase in electrical conductivity induced by absorption of light.

Photodarkening: Light-induced metastable change in the structure of chalcogenide materials that causes a red shift of the optical absorption edge.

Photoluminescence: Secondary light emitted from a sample that is excited by primary light.

Pnictogen: Element of group V of the periodic table: N, P, As, Sb.

Staebler–Wronski Effect: Light-induced creation of metastable defects in amorphous silicon and related alloys.

Statistical Shift of Fermi Level: The temperature dependence of the Fermi energy that is demanded by charge conservation in the presence of a density of states changing with energy.

Stokes Shift: Energy difference between the exciting primary light and the emitted secondary light.

Telegraph Noise: Sudden random changes of the conductance of a sample between two or more discrete values.

Thermopower: The potential difference per degree centigrade between the warm end and the cold end of a sample placed in a temperature gradient.

Transport Energy: The energy of predominant hopping conduction in an energy-dependent density of localized states.

Urbach Edge: Region of the optical absorption spectrum in which the absorption coefficient increases exponentially with photon energy.

Valence-Alternation Defects: A defect pair of which one partner is positive and an overcoordinated atom and the other is negative and an undercoordinated atom. They may be close (intimate pair) or distant.

Valence Band: The uppermost filled energy band of electron states.

Works Cited

Abeles, B., Tiedje, T. (1983), *Phys. Rev. Lett.* **51,** 2003–2006.

Adler, D., Joffa, E. J. (1976), *Phys. Rev. Lett.* **36,** 1197–1200.

Agarwal, S. C. (1975), *Phys. Rev. B* **7,** 685.

Aker, B., Fritzsche, H. (1983), *J. Appl. Phys.* **54,** 6628–6633.

Anderson, P. W. (1958), *Phys. Rev.* **109,** 1492–1505.

Anderson, P. W. (1975), *Phys. Rev. Lett.* **34,** 953–955.

Antoniadis, H., Schiff, E. A. (1992), *Phys. Rev. B* **46,** 9482–9492.

Arce, R., Ley, L., Hundhausen, M. (1989), *J. Non-Cryst. Solids* **114,** 696–698.

Ast, D. G., Brodsky, M. H. (1979), in: B. L. H. Wilson (Ed.), *Proceedings of the 14th International Conference on the Physics of Semiconductors,* IOP Conference Series No. 43, Bristol: Institute of Physics, p. 1159.

Bar-Yam, Y., Adler, D., Joannopoulos, J. D. (1986), *Phys. Rev. Lett.* **57,** 467–470.

Bellingham, J. R., Graham, M., Adkins, C. J., Phillips, W. A. (1991), *J. Non-Cryst Solids* **137&138,** 519–532.

Bellingham, J. R., Phillips, W. A., Adkins, C. J. (1990), *J. Phys.: Condensed Matter* **2,** 6207–6221.

Beyer, W., Overhof, H. (1979), *Solid State Commun.* **31,** 1–4.

Beyer, W., Overhof, H. (1984), in: J. I. Pankove (Ed.), *Semiconductors and Semimetals,* Vol. 21C, New York: Academic, p. 257.

Biegelsen, D. K., Street, R. A. (1980), *Phys. Rev. Lett.* **44,** 803–806.

Bishop, S. G., Strom, U., Taylor, P. C. (1977), *Phys. Rev. B* **15,** 2278–2294.

Branz, H. M., Silver, M. (1990), *Phys. Rev. B* **42,** 7420–7428.

Brodsky, M. H., Title, R. S. (1969), *Phys. Rev. Lett.* **23,** 581–585.

Brüggemann, R., Main, C., Berkin, J., Reynolds, S. (1990), *Philos. Mag. B* **62,** 29–45.

Choi, W. K., Owen, A. E., LeComber, P. G., Rose, M. J. (1990), *J. Appl Phys.* **68,** 120–123.

Cohen, J. D., Harbison, J. B., Wecht, K. W. (1982), *Phys. Rev. Lett.* **48,** 109–112.

Cohen, J. D., Lang, D. V., Harbison, J. P. (1980), *Phys. Rev. Lett.* **45,** 197–200.

Cohen, J. D., Leen, T. M., Rasmussen, R. J. (1993), *Phys. Rev. Lett.* **69,** 3358–3361.

Cohen, M. H. (1970), *J. Non-Cryst. Solids* **4,** 391–409.

Cohen, M. H., Fritzsche, H., Ovshinsky, S. R. (1969), *Phys. Rev. Lett* **22,** 1065–1068.

DeNeufville, J. P., Moss, S. C., Ovshinsky, S. R. (1974), *J. Non-Cryst. Solids* **13,** 191–223.

Dersch, H., Stuke, J., Beichler, J. (1981), *Appl. Phys. Lett.* **38,** 456–458.

Döhler, G. H. (1972), *Phys. Status Solidi B* **52,** 79–92.

Elliott, S. R. (1986), *J. Non-Cryst. Solids* **81,** 71–98.

Friedman, L. (1971), *J. Non-Cryst. Solids* **6,** 329–341.

Fritzsche, H. (1971), *Solid State Commun.* **9,** 1813–1815.

Fritzsche, H. (1974), in: J. Tauc (Ed.), *Amorphous and Liquid Semiconductors,* London: Plenum, p. 221.

Fritzsche, H. (1993a), *Philos. Mag. B* **68,** 561–572.

Fritzsche, H. (1993b), *J. Non-Cryst. Solids* **164–166,** 1169–1172.

Fritzsche, H. (1994), *Appl. Phys. Lett.* **65,** 2824–2826.

Fritzsche, H. (1995), *Phys. Rev. B* **52,** 15854–15861.

Fritzsche, H., Kastner, M. (1978), *Philos. Mag. B* **37,** 285–292.

Gaczi, P. (1982), *Philos. Mag. B* **45,** 241–259.

Gee, C. M., Kastner, M. (1979), *Phys. Rev. Lett.* **42,** 1765–1769.

Gelatos, A. V., Cohen, J. D., Harbison, J. P. (1986), *Appl. Phys. Lett.* **49,** 722–724.

Gelatos, A. V., Mahavadi, K. K., Cohen, J. D., Harbison, J. P. (1988), *Appl. Phys. Lett.* **53,** 403–405.

Goryunova, N. A., Kolomiets, B. T. (1955), *Zh. Tekh. Fiz.* **25,** 984–994 and 2069–2078.

Greenwood, D. A. (1958), *Proc. Phys. Soc. London* **71,** 585–596.

Grünewald, M., Thomas, P. (1979), *Phys. Status Solidi B* **94,** 125–133.

Guha, S., Yang, T., Czubatyj, W., Hudgens, S. J., Hack, M. (1983), *Appl. Phys. Lett.* **42,** 588–589.

Hamed, A. (1991), *Phys. Rev. B* **44,** 5585–5602.

Hattori, K., Mori, T., Okamoto, H., Hamakawa, Y. (1987), *Appl. Phys. Lett.* **51,** 1259–1261.

Hecht, K. (1932), *Z. Phys.* **77,** 235–245.

Hepburn, A. R., Marshall, J. M., Main, C., Powell, M. J., van Berkel, C. (1986), *Phys. Rev. Lett.* **56,** 2215–2218.

Hoheisel, M., Fuhs, W. (1988), *Philos. Mag. B* **57,** 411–419.

Howard, J. A., Street, R. A. (1991), *Phys. Rev. B* **44,** 7935–7940.

Hundhausen, M., Ley, L., Carius, R. (1984), *Phys. Rev. Lett.* **53,** 1598–1601.

Jackson, W. B., Amer, N. M. (1982), *Phys. Rev. B* **25,** 5559–5562.

Jackson, W. B., Kakalios, J. (1988), *Phys. Rev. B* **37,** 1020–1023.

Jackson, W. B., Kakalios, J. (1989), in: H. Fritzsche (Ed.), *Amorphous Silicon and Related Materials,* Singapore: World Scientific, p. 247.

Jackson, W. B., Biegelsen, D. K., Nemanich, R. J., Knights, J. C. (1983), *Appl. Phys. Lett.* **42,** 105–107.

Jackson, W. B., Kelso, S. M., Tsai, C. C., Allen, J. W., Oh, S.-J. (1985), *Phys. Rev. B* **31,** 5187–5198.

Johanson, R. E., Fritzsche, H., Vomvas, A. (1989), *J. Non-Cryst. Solids* **114,** 274–276.

Johnson, N. M. (1983), *Appl. Phys. Lett.* **42,** 981–983.

Johnson, N. M., Biegelsen, D. K. (1985), *Phys. Rev. B* **31,** 4066–4069.

Kakalios, J. (1989), *J. Non-Cryst. Solids* **114,** 372–374.

Kakalios, J. (1991), in: J. I. Pankove (Ed.), *Semiconductors and Semimetals,* Vol. 34, New York: Academic, p. 381.

Kakalios, J., Fritzsche, H. (1984), *Phys. Rev. Lett.* **53,** 1602–1605.

Kakalios, J., Street, R. A., Jackson, W. B. (1987), *Phys. Rev. Lett.* **59,** 1037–1040.

Kastner, M. (1972), *Phys. Rev. Lett.* **28,** 355–357.

Kastner, M., Adler, D., Fritzsche, H. (1976), *Phys. Rev. Lett.* **37,** 1504–1507.

Kolobov, A. V., Kostikov, A. A., Yu, P., Lantratova, S. S., Lyubin, V. M. (1991), *Fiz. Tverd. Tela* **33,** 781–785 [*Sov. Phys. Solid State* **33,** 444–447].

Kruhler, W., Pfleiderer, H., Plattner, R., Stetter, W. (1984), in: P. C. Taylor, S. G. Bishop (Eds.), *Optical Effects in Amorphous Semiconductors,* AIP Conference Proceedings No. 120, New York: American Institute of Physics, p. 311.

Kubo, R. (1957), *J. Phys. Soc. Jpn.* **12,** 570–586.

Lang, D. V., Cohen, J. D., Harbison, T. P. (1982a), *Phys. Rev. B* **25,** 5285–5320.

Lang, D. V., Cohen, J. D., Harbison, J. P. (1982b), *Phys. Rev. Lett.* **48,** 421–424.

LeComber, P. G., Jones, D. I., Spear, W. E. (1977), *Philos. Mag.* **35,** 1173–1187.

Ley, L. (1984), in: J. D. Joannopoulous, G. Lucovsky (Eds.), *The Physics of Hydrogenated Amorphous Silicon II,* Berlin: Springer-Verlag, Chap. 3.

Liu, J. Z., Taylor, P. C. (1987), *Phys. Rev. Lett.* **59,** 1938–1941.

Lyubin, V. M., Tikhomirov, V. K. (1989), *J. Non-Cryst. Solids* **114,** 133–135.

Lyubin, V. M., Tikhomirov, V. K. (1991), *J. Non-Cryst. Solids* **137&138,** 993–996.

Meyer, W., Neldel, H. (1937), *Z. Tech. Phys.* **12,** 588–593.

Michelson, C. E., Gelatos, A. V., Cohen, J. D. (1985), *Appl. Phys. Lett.* **47,** 412–414.

Monroe, D. (1985), *Phys. Rev. Lett.* **54,** 146.

Morigaki, K. (1984), in: J. I. Pankove (Ed.), *Semiconductors and Semimetals,* Vol. 21C, Orlando, FL: Academic, p. 155.

Oheda, H. (1981), *J. Appl. Phys.* **52,** 6693–6700.

Okushi, H., Tokumaru, Y., Yamasaki, S., Oheda, H., Tanaka, K. (1982), *Phys. Rev. B* **25,** 4313–4316.

Onsager, L. (1938), *Phys. Rev.* **54,** 554–557.

Orenstein, J., Kastner, M. (1981), *Phys. Rev. Lett.* **46,** 1421–1424.

Ovadyahu, Z. (1986), *J. Phys. C* **19,** 5187–5213.

Ovadyahu, Z. (1990), in: H. Fritzsche, M. Pollak (Eds.), *Hopping and Related Phenomena,* Singapore: World Scientific, p. 193.

Overhof, H., Beyer, W. (1984), *Philos. Mag. B* **49,** L9–L14.

Overhof, H., Thomas, P. (1989), *Electronic Transport in Hydrogenated Amorphous Semiconductors,* Springer Tracts in Modern Physics Vol. 114, Berlin, Heidelberg, New York: Springer-Verlag.

Ovshinsky, S. R. (1968), *Phys. Rev. Lett.* **21,** 1450–1453.

Ovshinsky, S. R. (1977), in: W. E. Spear (Ed.), *Proceedings of the 7th International Conference on Amorphous Semiconductors,* Edinburgh: Centre for Industrial Consultancy and Liaison, Univ. of Edinburgh, p. 519.

Owen, A. E., Firth, A. P., Ewen, P. J. S. (1985), *Philos. Mag. B* **52,** 347–362.

Pantelides, S. T. (1986), *Phys. Rev. Lett.* **57,** 2979–2982.

Parman, C. E., Israeloff, N. E., Kakalios, J. (1991), *Phys. Rev. B* **44,** 8391–8394.

Pashmakov, B., Claflin, B., Fritzsche, H. (1993), *J. Non-Cryst. Solids* **164–166,** 441–444.

Pfeiffer, G., Paesler, M. A., Agarwal, S. C. (1991), *J. Non-Cryst. Solids* **130,** 111–143.

Phillips, J. C. (1979), *J. Non-Cryst. Solids* **34,** 153–181.

Ponpon, J. P., Bourdon, B. (1982), *Solid-State Electron.* **25,** 875–876.

Pontuschka, W. M., Carlos, W. E., Taylor, P. C., Griffith, R. W. (1982), *Phys. Rev. B* **25,** 4362–4376.

Powell, M. J., van Berkel, C., French, I. D., Nicholls, D. H. (1987), *Appl. Phys. Lett.* **51,** 1242–1244.

Rogers, C. T., Buhrman, R. A., Korger, H., Smith, L. N. (1986), *Appl. Phys. Lett.* **49,** 1107–1109.

Rose, A. (1963), *Concepts in Photoconductivity and Allied Problems,* New York: Interscience.

Scher, H., Montroll, E. W. (1975), *Phys. Rev. B* **12,** 2455–2477.

Shapiro, F. R., Adler, D. (1985), *J. Non-Cryst. Solids* **77&78,** 139–142.

Shklovskii, B. I., Fritzsche, H., Baranovskii, S. D. (1989), *Phys. Rev. Lett.* **62,** 2989–2992.

Simmons, J. G., Taylor, G. W. (1971), *Phys. Rev. B* **4,** 502–511.

Smith, Z., Wagner, S. (1985), *Phys. Rev. B* **32,** 5510–5513.

Solomon, I. (1979), in: M. H. Brodsky (Ed.), *Amorphous Semiconductors,* Topics in Applied Physics Vol. 36, Berlin: Springer, Chap. 7.

Spear, W. E., LeComber, P. G. (1975), *Solid State Commun.* **17,** 1193–1196.

Staebler, D. L., Wronski, C. R. (1977), *Appl. Phys. Lett.* **31,** 292–294.

Staebler, D. L., Wronski, C. R. (1980), *J. Appl. Phys.* **51,** 3262–3268.

Stradins, P., Fritzsche, H. (1994), *Philos. Mag. B* **69,** 121–139.

Street, R. A. (1977), *Solid State Commun.* **24,** 363–365.

Street, R. A. (1985), *J. Non-Cryst. Solids* **77&78,** 1–16.

Street, R. A. (1991a), *Appl. Phys.* **59,** 1084.

Street, R. A. (1991b), *Hydrogenated Amorphous Silicon,* Cambridge, U.K.: Cambridge Univ. Press.

Street, R. A., Biegelsen, D. K., Stuke, J. (1979), *Philos. Mag B* **40,** 451–464.

Street, R. A., Mott, N. F. (1975), *Phys. Rev. Lett.* **35,** 1293–1296.

Street, R. A., Zesch, J., Thompson, M. J. (1983), *Appl. Phys. Lett.* **43,** 672–674.

Street, R. A., Kakalios, J., Hayes, T. M. (1986), *Phys. Rev. B* **34,** 3030–3033.

Street, R. A., Kakalios, J., Tsai, C. C., Hayes, T. M. (1987), *Phys. Rev. B* **35,** 1316–1333.

Street, R. A., Hack, M., Jackson, W. B. (1988a), *Phys. Rev. B* **37,** 4209–4224.

Street, R. A., Kakalios, J., Hack, M. (1988b), *Phys. Rev. B* **38,** 5603–5609.

Stuke, J. (1977), in: W. E. Spear (Ed.), *Proceedings of the 7th International Conference on Amorphous and Liquid Semiconductors,* Edinburgh: Centre for Industrial Consultancy and Liaison, Univ. of Edinburgh, p. 406.

Stutzmann, M., Street, R. A. (1985), *Phys. Rev. Lett.* **54,** 1836–1839.

Stutzmann, M., Jackson, W. B., Tsai, C. C. (1985), *Phys. Rev. B* **32,** 23–47.

Stutzmann, M., Biegelsen, D. K., Street, R. A. (1987), *Phys. Rev. B* **35,** 5666–5701.

Tanaka, Ka. (1976), in: G. Lucovsky, G. L. Galeener (Eds.), *Structure and Excitations in Amorphous Solids,* AIP Conference Proceedings No. 31, New York: American Institute of Physics, p. 148.

Tanaka, Ka. (1981), in: F. Yonezawa (Ed.), *Fundamental Physics of Amorphous Semiconductors,* New York: Springer, p. 104.

Tanaka, Ke. (1983), *J. Non-Cryst. Solids* **59&60,** 925–928.

Tanaka, Ke. (1990), *Rev. Solid State Sci.* **2,** (1990), ibid. **3,** 641–659.

Tanielian, M. (1982), *Philos. Mag. B* **45,** 435–462.

Tauc, J. (1974), *Amorphous and Liquid Semiconductors,* London and New York: Plenum.

Tauc, J., Grigorivici, R., Vancu, A. (1966), *Phys. Status Solidi* **15,** 627–637.

Teuschler, T., Hundhausen, M., Ley, L., Arce, R. (1993), *Phys. Rev. B* **47,** 12687–12695.

Tiedje, T. (1984), in: J. I. Pankove (Ed.), *Semiconductors and Semimetals,* Vol. 21C, Orlando, FL: Academic, Chap. 6.

Tiedje, T., Rose, A. (1980), *Solid State Commun.* **37,** 49–52.

Tsang, C., Street, R. A. (1979), *Phys. Rev. B* **19,** 3027–3040.

Unold, T., Cohen, J. D. (1991), *Appl. Phys. Lett.* **58,** 723–725.

Voget-Grote, U., Kümmerle, W., Fischer, R., Stuke, J. (1980), *Philos. Mag. B* **41,** 127–140.

Winer, K. (1989), *Phys. Rev. Lett.* **63,** 1487–1490.

Winer, K., Ley, L. (1987), *Phys. Rev. B* **36,** 6072–6078.

Yang, L., Abeles, B., Eberhardt, W., Slasiewski, H. E., Sondericker, D. (1989), *Phys. Rev.* **39,** 3801–3816.

Yasukawa, M., Hosono, H., Ueda, N., Kawazoe, H. (1995), *Jpn. J. Appl. Phys.* **34,** L281–L284.

Zallen, R. (1983), *The Physics of Amorphous Solids,* New York: Wiley.

Zhdanov, V. G., Kolomiets, B. T., Lyubin, V. M., Malinovskii, V. K. (1979), *Phys. Status Solidi A* **52,** 621–626.

Zhong, F., Cohen, J. D. (1993), *Phys. Rev. Lett.* **71,** 597–600.

Further Reading

Adler, D., Fritzsche, H. (1984), *Tetrahedrally-Bonded Amorphous Semiconductors,* New York: Plenum.

Adler, D., Fritzsche, H., Ovshinsky, S. R. (1984), *Physics of Disordered Materials,* New York: Plenum.

Brodsky, M. H. (Ed.) (1979), *Amorphous Semiconductors,* Topics in Applied Physics Vol. 36, Berlin: Springer.

Elliott, S. R. (1990), *Physics of Amorphous Materials,* 2nd ed., London: Longmans.

Fritzsche, H. (Ed.) (1989), *Amorphous Silicon and Related Materials,* Advances in Disordered Materials Vol. 1, Singapore: World Scientific.

Fritzsche, H. (1990), *Transport, Correlation and Structural Defects,* Advances in Disordered Materials Vol. 3, Singapore: World Scientific.

Kastner, M. A., Thomas, G. A., Ovshinsky, S. (Eds.) (1987), *Disordered Semiconductors,* New York: Plenum.

Mott, N. F., Davis, E. A. (1979), *Electronic Processes in Non-Crystalline Materials,* Oxford: Clarendon.

Pankove, J. I. (Ed.) (1984), *Hydrogenated Amorphous Silicon,* Semiconductors and Semimetals Vol. 21, Orlando, FL: Academic.

Street, R. A. (1991), *Hydrogenated Amorphous Silicon,* Cambridge, U.K.: Cambridge Univ. Press.

Zallen, R. (1983), *The Physics of Amorphous Semiconductors,* New York: Wiley.

SEMICONDUCTORS, AMORPHOUS—MATERIAL PROPERTIES

S. R. Elliott, *Department of Chemistry, University of Cambridge, Cambridge, U.K.*

Introduction 277
1. **Preparation** 277
1.1 Vapor–Solid Amorphization ... 277
1.1.1 Evaporation 278
1.1.2 Sputtering 278
1.1.3 Chemical Vapor Deposition 278
1.2 Liquid–Solid Amorphization ... 279
1.3 Solid–Solid Amorphization 279
2. **Classification of Amorphous Semiconductors** 281
2.1 Structural Classification 282
2.2 Electronic Classification 283
3. **Optical Properties** 284
3.1 Electronic Structure 284
3.2 Interband Transitions 286
3.3 Urbach Edge 287
3.4 Subgap Transitions and Defect States 290
4. **Vibrational and Thermal Properties** 290
4.1 Vibrational Density of States 290
4.2 Propagating and Localized Phonons 293
4.3 Soft Configurations and Two-Level (Tunneling) Systems 294
4.4 Heat Capacity 295
4.5 Thermal Conductivity 297
Glossary 300
Works Cited 301

INTRODUCTION

Amorphous semiconductors have no long-range translational order (periodicity), by definition, and this structural disorder confers a number of unique and characteristic properties in their physical behavior. Thus, both electronic and vibrational states can become spatially localized under certain circumstances, and the transport and other behavior associated with such localized states in amorphous solids can be very different from that exhibited by the corresponding crystalline material counterparts. The influence of structural disorder on electronic transport and related properties is covered by the article Semiconductors, Amorphous: Electronic Properties. Here, attention is concentrated on optical and vibrational (thermal) properties of amorphous semiconductors, together with a discussion on methods of preparation and classification schemes for such materials.

1. PREPARATION

Amorphous materials differ from their crystalline counterparts by possessing a greater degree of configurational entropy and free energy. In general, such excess entropy and free energy has to be incorporated either sufficiently rapidly in the course of formation of the solid that crystallization is precluded or *post facto* by means of damage mechanisms. Because of the predominantly covalent nature of the bonding in amorphous semiconductors, certain amorphization methods that work for, say, metals are not applicable—e.g., solid-state diffusional amorphization. Nevertheless, there are many general techniques for producing amorphous semiconductors, and these can be divided into three categories, depending on whether amorphization commences in the vapor, liquid, or solid phase.

1.1 Vapor–Solid Amorphization

Vapor deposition is a commonly used technique for preparing amorphous semiconductors in thin-film form. The effective cooling rate when a vapor is condensed on a cold substrate is extremely high, and usually high enough to preclude crystallization of the growing film unless the substrate temperature is sufficiently elevated to give rise to

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

appreciable surface diffusion. A number of different techniques exist for ensuring that the elements needed to produce the amorphous film exist in the vapor phase prior to condensation on a substrate. In addition, the deposition process may simply be physical (collision with, and adhesion on, the substrate by the vapor species), or it may involve a (heterogeneous) chemical reaction at the growing vapor–solid interface.

1.1.1 Evaporation This is perhaps the simplest technique, conceptually. A charge of the material to be deposited is heated *in vacuo* to form a vapor, and a thin film of the material condenses on a substrate held in the vapor stream. For relatively low–melting-point materials, such as chalcogenides, the material may simply be placed in a resistively heated boat. For higher–melting-point materials, e.g., Si or Ge, electron-beam heating may be required in order to vaporize enough material.

Evaporation of multicomponent alloy systems may be troublesome because of preferential vaporization of high–vapor-pressure species (atoms or molecules), leading to deposited films with a composition different from that of the starting material. "Flash" evaporation, whereby grains of the charge are heated extremely rapidly (e.g., by dropping on a preheated element), can obviate such disproportionation problems to a certain extent.

1.1.2 Sputtering An alternative method of producing material in the vapor phase is by sputtering a solid target with charged ions from a plasma established by electrical discharge in a low-pressure gas. Thin-film deposition occurs on a substrate held on the far side of the plasma from the target. When the gas used is inert, e.g., Ar, "physical" sputtering occurs: ablation of the target takes place with subsequent deposition of the ablated material on a substrate, perhaps containing some physically entrained sputtering gas. This process is better at preserving the stoichiometry of multicomponent compounds than is simple evaporation. Alternatively, a reactive gas may be added to the sputtering gas, resulting in its becoming chemically incorporated in the growing film: an example is the use of hydrogen in producing hydrogenated amorphous silicon, *a*-Si:H.

1.1.3 Chemical Vapor Deposition Chemical vapor deposition (CVD) is a general term for the deposition of thin solid films following the chemical reaction between vapor-phase precursor species, usually by means of a heterogeneous reaction at the growing film–vapor interface. In the simplest case, the heterogeneous reaction is promoted by heating the substrate upon which the film is to be deposited. However, the very high temperatures required to cause the reaction to take place often lead to crystal formation; hence, straightforward CVD is not particularly useful for the preparation of films of amorphous semiconductors. Nevertheless, a variant termed the "hot-wire" technique has proved useful in producing device-quality (low defect-density) *a*-Si:H films that have rather low H contents (0.1–1 at.%) (Mahan *et al.*, 1991). In this, a metallic filament (e.g., tungsten) is heated to temperatures in the range 1000–2000 °C, upon which the silane (SiH_4) feedstock gas decomposes catalytically (Matsumura, 1989). The resulting free radicals (SiH_n,H) then impinge on a substrate positioned nearby, and held at a much lower temperature (~200 °C), thereby forming a film of *a*-Si:H.

CVD can also be promoted using energy sources other than thermal to initiate reaction. Perhaps the most common technique is the use of an electrical plasma in so-called plasma-enhanced (PE) CVD or glow-discharge (GD) decomposition. The plasma causes dissociation of the feedstock gas into ions and free radicals that then impinge on a moderately heated substrate (~200 °C) to form a thin film. The processes occurring in the plasma and at the surface are complex and interrelated (Roca i Cabarrocas, 1993). The primary electron-initiated decomposition of, say, silane, in the formation of *a*-Si:H, creates free radicals (e.g., SiH_3). There is a subsequent competition between surface-related processes (deposition, surface mobility, and reaction) and secondary gas-phase reactions leading to clusters, polymers, or even powder depending on the rf power, gas pressure, flow rate, and reactor geometry used. The composition of the feedstock gas is also critical: high dilution of SiH_4 with H_2, for example, leads to the formation of microcrystalline, not amorphous, silicon. The PECVD process is important because not only can it be used to produce device-quality films of *a*-

Si:H (with about 10–15 at.% hydrogen), but such films can be controllably doped *n*-type or *p*-type by incorporating into the gas-flow phosphine (PH_3) or diborane (B_2H_6), respectively (Spear and Le Comber, 1975). A disadvantage of conventional PECVD is that bombardment of the growing film surface by ions from the plasma can be deleterious to film quality. To avoid this effect, the plasma can be operated remote from the growth chamber, in which case only one of the gaseous species (e.g., hydrogen) is excited by the plasma, the resulting ions being extracted and reacting with other gases (e.g., silane) in the growth chamber (Lucovsky and Tsu, 1987).

Finally, an alternative energy source that can be used to initiate CVD is light. Photo-CVD uses photons (typically in the vacuum-ultraviolet) directly to photolyze precursor molecules in the gas prior to deposition. For example, photo-CVD of disilane (Si_2H_6) has been shown to produce high-quality *a*-Si:H films (Yoshida *et al.*, 1990).

1.2 Liquid–Solid Amorphization

The formation of glasses by means of rapid quenching of a melt is perhaps the simplest and most widespread method of producing amorphous materials. Although commonly used for the production of conventional oxide (e.g., silicate) glasses, the melt-quenching technique can also be used to produce semiconducting chalcogenide (S-, Se-, Te-based) glasses, as well as (with difficulty) the tetrahedrally coordinated "chalcopyrite" [II–IV–$(V)_2$] glasses, e.g., $CdGeAs_2$. It is necessary that the cooling rate be sufficiently rapid that crystallization is precluded (i.e., faster than the homo/heterogeneous crystalline nucleation rate). For many chalcogenide glasses, quenching of an ampoule containing the melt (giving a cooling rate of a few hundred degrees per second) is sufficient; however, for those compositions more prone to crystallization (e.g., tellurides or silicon chalcogenides), considerably faster quenching rates are required, necessitating the use of techniques such as twin-roller quenching.

Generally, a prerequisite for vitrification by melt quenching is that the atomic structures of the amorphous solid and of the melt be very similar so that vitrification is favored. This is the case for conventional glass formers, such as oxides and chalcogenides, but not for Si or Ge, where, although the semiconducting solid forms (both amorphous and crystalline) have tetrahedral coordination, the metallic liquid state is characterized by octahedral coordination. Although this change in coordination prevents the amorphous state being formed from the liquid by conventional (relatively slow) quenching methods, nevertheless it has been demonstrated that illumination of crystalline Si by picosecond laser pulses causes melting of the surface layer, and subsequent ultrafast ($\sim 10^{14}$ K s^{-1}) quenching results in an amorphous Si product (Tsu *et al.*, 1979; Liu *et al.*, 1979; Cullis *et al.*, 1982; Smirl *et al.*, 1986). It is not clear at present to what extent the very thin layers of *a*-Si (often associated with *c*-Si) produced by "laser glazing" resemble material produced by vapor-deposition techniques (e.g., evaporation, sputtering, GD decomposition); it is likely that the *a*-Si resulting from laser-quenching experiments is considerably more disordered and defective.

1.3 Solid–Solid Amorphization

There are two methods, with application to semiconductors, by which crystalline materials can be rendered amorphous while remaining in the solid state, namely, pressure quenching and radiation damage. The former topic will be considered first since it bears some passing resemblance to liquid–solid amorphization discussed in the previous section.

The formation of noncrystalline solids by means of pressure quenching is based on the behavior whereby a (high-coordination, often metallic) high-pressure crystalline phase becomes mechanically unstable at a particular (high) pressure and temperature, and the transition to the crystalline modification stable under such circumstances (usually semiconducting with a lower coordination number) is kinetically hindered on release of the pressure and/or reduction in temperature, resulting in a disordered, often amorphous, phase being produced.

This technique is particularly appropriate for the group-IV materials Si and Ge and their III–V analogs, e.g., GaSb, where the melt is octahedrally coordinated, as is a high-pressure phase, although the crystalline (and amorphous) phase stable at ambient pressure and temperature is tetrahedrally coordinated (see Fig. 1). Pressure quenching of the GaSb II octahedral crystalline phase from a pressure of 90 kbar at room temperature produces bulk amorphous GaSb with crystalline inclusions, the proportion of which depends on the melt temperature from which the initial isobaric temperature quench to form the high-pressure crystalline phase (II) is commenced (Demishev *et al.*, 1994). Similar bulk (~mm^3 volume) samples of *a*-Ge and *a*-Si, again with 20–50 vol % crystallite inclusions, can also be made by pressure quenching (at ~100 K) from a high-pressure crystalline phase (II); for *a*-Ge, the

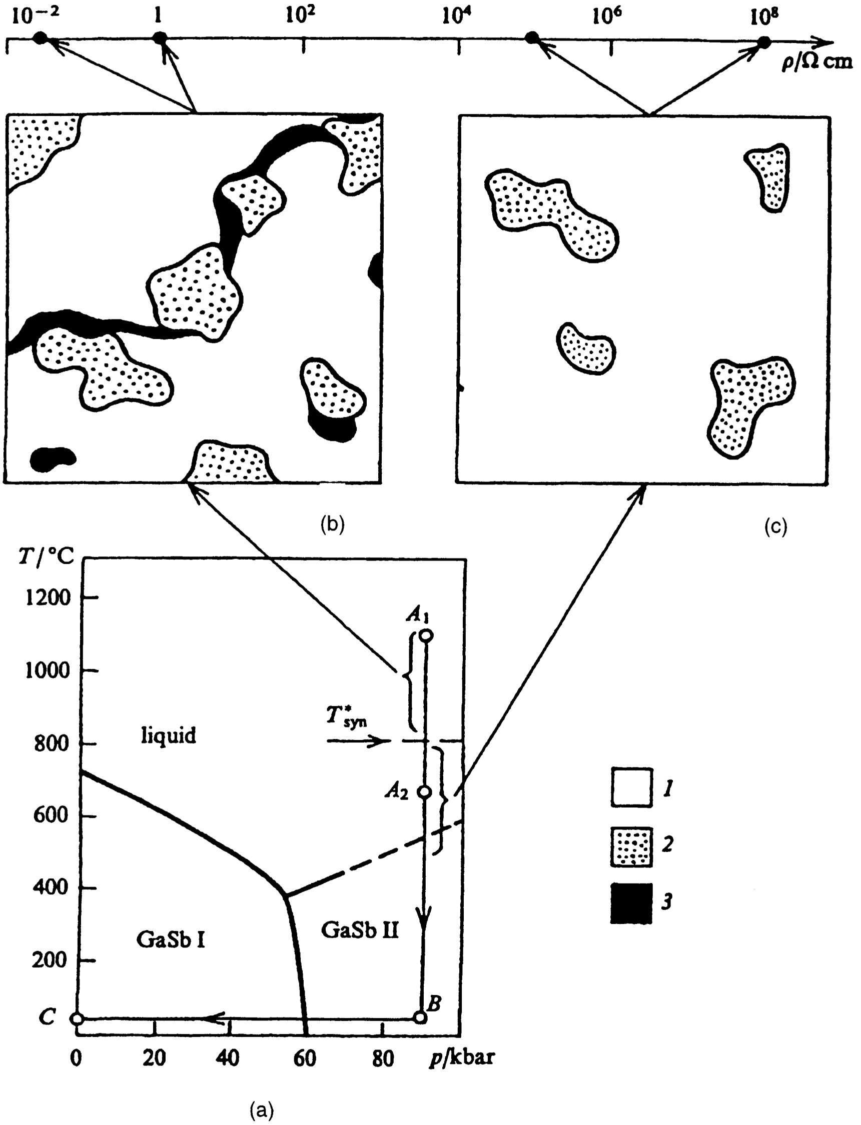

FIG. 1. Illustration of the formation of bulk GaSb by pressure quenching, A → B → C (Demishev *et al.*, 1994), and typical morphologies of the samples obtained by thermal quenching from two temperatures in the melt, A_1 and A_2 (material 1 is bulk *a*-GaSb, 2 is *c*-GaSb, and 3 is a crystalline Ga_xSb_{1-x} phase). Phase I is the low-pressure, tetrahedrally coordinated form, and II, the high-pressure, octahedrally coordinated form, of crystalline GaSb.

transition pressure needed to transform to the amorphous phase is ~1 GPa, and for *a*-Si, ~5 GPa (Brazhkin *et al.*, 1992). The advantage of this technique is that bulk samples of the amorphous phase can be produced that otherwise can only be formed in thin-film form by vapor deposition; the disadvantage is that currently it seems impossible to obtain such pressure-quenched samples completely free of minority crystallite inclusions.

An alternative approach to creating amorphous materials via a solid–solid transformation is by means of radiation damage—e.g., associated with ion implantation. For example, irradiation of crystalline silicon by 2-MeV Si^+ ions causes the material to become amorphous (Hecking *et al.*, 1986). Sufficient energy is deposited into the "spike" of the target material traversed by the impinging ion that the material locally melts. If the surrounding sample temperature is high enough, the atomic mobility is sufficiently large that the melt can recrystallize. However, recrystallization is kinetically hindered at low temperatures, and at sufficiently low temperatures ($T \leq 125$ K for Si), amorphization in the collision-cascade region is complete (Hecking *et al.*, 1986).

Finally, a novel form of radiation-induced amorphization is the athermal photoinduced amorphization found to occur in thin films of crystalline $As_{50}Se_{50}$ (Elliott and Kolobov, 1991; Kolobov and Elliott, 1995). This process is reversible: illumination of a thermally crystallized film of $As_{50}Se_{50}$, having the realgar structure composed of As_4Se_4 molecules, by low-intensity band-gap light causes complete amorphization; subsequent thermal annealing causes the photoamorphized film to crystallize once more, whereupon the whole process can be repeated (see Fig. 2). Although the precise mechanism of photoamorphization is not yet clear, it is believed that the action of the light is to break strained intramolecular As–As bonds with the subsequent reformation of bonds between molecules, thereby forming a (partially) cross-linked amorphous structure. Mechanical strain associated with the substrate–film interface may also play a role (Kolobov and Elliott, 1995).

2. CLASSIFICATION OF AMORPHOUS SEMICONDUCTORS

The definition of a semiconductor is rather imprecise, and the distinction between a semiconductor and an insulator is merely

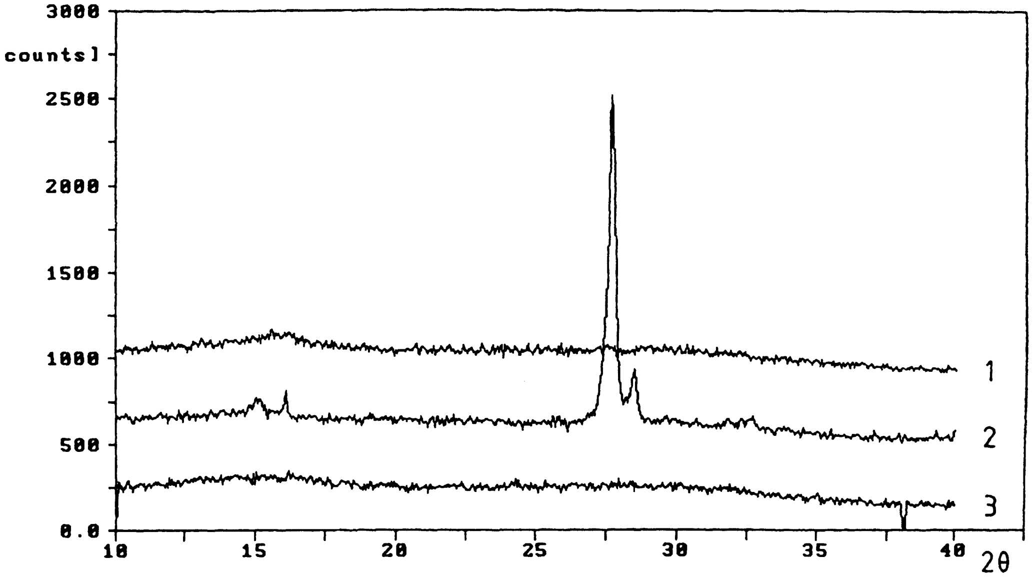

FIG. 2. Reversible photoamorphization of $As_{50}Se_{50}$ (Kolobov and Elliott, 1995). The x-ray diffractograms refer respectively to (curve 1) as-evaporated amorphous film; (curve 2) crystalline film produced by annealing at 185 °C for 24 h; (curve 3) photoamorphized film (after illumination for 4 h.

one of degree, in either magnitude of the band gap between valence and conduction bands or the value of the dc conductivity at, say, room temperature. For the purpose of discussion in this article, a material will be regarded as a semiconductor if its band gap is less than 4 eV and its electrical conduction is intrinsic; i.e., the charge carriers result from thermal excitation across the gap. This arbitrary definition has the advantage that whole classes of material are thereby completely included/excluded. For example, with this definition, all amorphous chalcogenide materials (alloys of S, Se, Te with Si, Ge, As, B, etc.) are deemed to be semiconductors, whereas all oxide glasses are regarded as insulators. [The exceptions to this otherwise appealingly simple picture are those oxide glasses containing mixed-valence (transition-metal) ions, e.g., Fe^{2+}/Fe^{3+}, V^{4+}/V^{5+}, etc.; in such cases, "small-polaron" transport of electrons/holes, involving lattice distortion, can occur between ions with differing charge states (Mott and Davis, 1979), giving rise to a value of intrinsic electronic conductivity comparable to that of a material having a much smaller band gap.]

It is also useful, for the purpose of discussing their properties, to classify amorphous semiconductors further, on the basis of either their atomic or electronic structure, although to some extent these two classifications are linked.

2.1 Structural Classification

It has been conventional in the past, perhaps because of prejudice emanating from the *crystalline* semiconductor field, to divide amorphous semiconductors into two categories, i.e., those that are tetrahedrally coordinated and those that are not. In addition to Si and Ge, other tetrahedrally coordinated amorphous materials include the III–V compounds (e.g., GaAs, GaSb) and the so-called chalcopyrite glasses based on the II–IV–$(V)_2$ composition, e.g., $(CdAs_2)_{1-x}Si_x$ (Houserova and Hruby, 1972). While there is some sense in such a division based on electronic-structure (sp^3-bonding) arguments (see Sec. 2.2), there is also merit in considering a division based on structural/mechanical considerations but, instead, where the average atomic coordination number, r, is greater than, or less than, a critical value, $r_c = 2.67$, rather than equal to, or less than, a value of 4.

Such a structural/mechanical classification is based on the Phillips–Thorpe idea of "rigidity percolation" wherein, for a covalent network constrained by bond-stretching and bond-bending interactions, a mechanical critical point occurs when the number of constraints per atom, n_c, is equal to the dimensionality n_d of the embedding space (Phillips, 1981a):

$$n_c = n_d. \tag{1}$$

The constraints are counted as follows (Thorpe, 1995). A single constraint is associated with each bond, and correspondingly, $r/2$ constraints with each r-coordinated atom. In addition, there are (a maximum of) $2r - 3$ angular constraints per atom. If the total (maximum) number of constraints per atom

$$n_c = r/2 + (2r - 3) \tag{2}$$

is equated to $n_d = 3$ [Eq. (1)], then the critical average coordination so obtained is

$$r_c = 2.4. \tag{3}$$

For average coordinations higher than this value, the structure is said to be overconstrained; for lower coordinations the structure is underconstrained or floppy. Should a fraction m/N of twofold-coordinated (anion) sites have the angular constraint lifted, then the rigidity-percolation threshold occurs at the average coordination number (Zhang and Boolchand, 1994)

$$r_c = 2.4 + 0.4m/N. \tag{4}$$

For AX_2-type glasses with 4:2 coordination (A = Si, Ge; X = O, S, Se), if *all* anions have the angular constraint broken, then $m/N = 2/3$ and Eq. (4) yields for the threshold average coordination number

$$r_c = 2.67 \tag{5}$$

in accord with the stoichiometry AX_2 for chemical stability.

The significance of this approach is that a simple mechanical criterion dictates whether glass formation is to be expected or not. Overconstrained materials, with $r > r_c$, are *not* expected to be glass formers because of the high degree of (homogeneous) mechanical strain expected to be incorporated if they were to form, whereas underconstrained materials, with $r < r_c$, are expected to form glasses readily. This picture has some validity, since chalcogenide and oxide materials, with $r < 2.67$, are indeed good glass formers, whereas As ($r = 3$), Si, and Ge ($r = 4$) do not readily form glasses but can nevertheless be produced in amorphous form by vapor quenching (see Sec. 1.1). However, the approach is somewhat simplistic since it completely ignores weaker interactions, such as van der Waals forces, and the fact that tetrahedrally coordinated materials such as $(CdAs_2)_{1-x}Si_x$, $0 < x < 0.25$, do form glasses, with difficulty, by rapid quenching from the melt (Houserova and Hruby, 1972) [and even *a*-Si is formed by ultrafast glazing (Sec. 1.3)] poses problems for the model. Examples of amorphous semiconductors categorized by this classification scheme are given in Table 1.

2.2 Electronic Classification

Another useful type of classification of amorphous semiconductors is in terms of electronic structure. It has been customary in the past to categorize materials into those that have sp^3-hybrid bonding (i.e., are tetrahedrally coordinated) and those characterized by having lone-pair (i.e., nonbonding) electrons. Examples of lone-pair semiconductors are chalcogenides, where the twofold-coordinated group-VI chalcogen atoms, having an atomic electronic configuration of s^2p^4, in fact give rise to two lone-pair bands in the solid, disparate in energy: one, the *s*-lone-pair band, lies deep within the valence band, and the other, the *p*-lone-pair band, lies at the very top of the valence band, with the intervening band comprising *p*-bonding states. The valence-band density-of-states profiles for these two types of amorphous semiconductors, obtained from photoemission experiments, are shown in Figs. 3(a) and 3(b).

Table 1. Structural classification of amorphous semiconductors in terms of mechanical rigidity.

Material	Average coordination number	Overconstrained (O)/ Underconstrained (U)
Si, Ge	4	O
GaAs	4	O
$CdGeAs_2$	4	O
As	3	O
$GeSe_2$, GeS_2	2.67	—
As_2Se_3, As_2S_3	2.4	U
Se	2	U

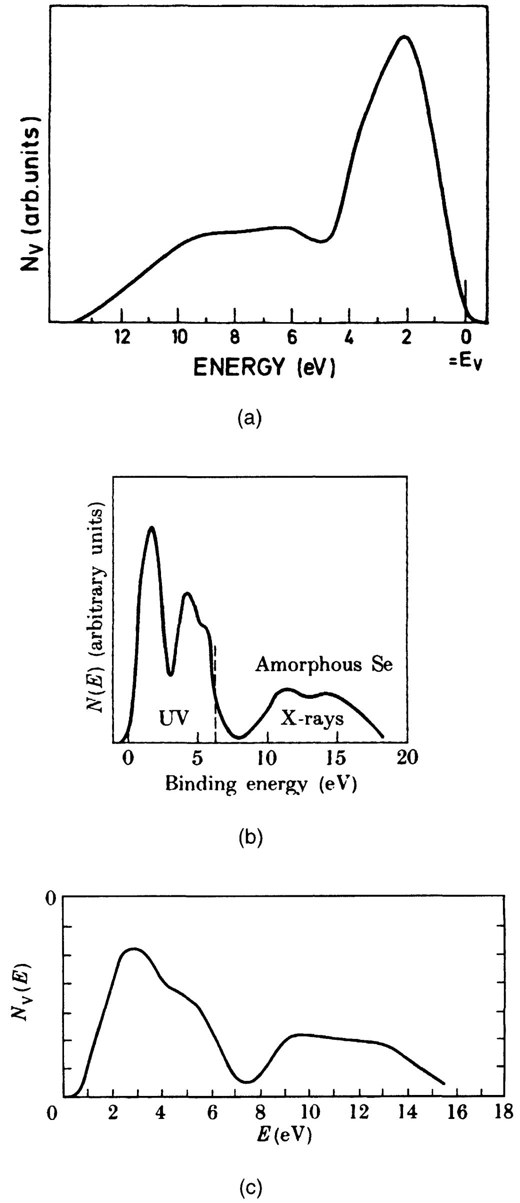

FIG. 3. Valence-band density of states obtained from photoemission for (a) *a*-Si (Ley, 1984), (b) *a*-Se (Shevchik *et al.*, 1973), (c) *a*-As (Ley *et al.*, 1973).

In fact, the key discriminator is the existence, or otherwise, of p-like lone-pair states at the top of the valence band, since these can be preferentially photoexcited and give rise to a plethora of photoinduced phenomena in the case of amorphous chalcogenides. In the absence of lone-pair states at the top of the valence band, from an electronic point of view it makes little difference, in fact, whether the bonding states forming the valence band derive from sp^3-hybrid orbitals (e.g., Si) or from other states, e.g., p orbitals (i.e., As). Thus, it is perhaps preferable to devise a classification scheme of amorphous semiconductors, based on electronic-structure considerations, in which one category includes materials that have (p-like) lone pairs comprising the top of the valence band, and the other includes materials where the top of the valence band, at least, comprises bonding orbitals. In this way, the awkward position of the group-V semiconductors, As and P, is resolved. With an atomic electronic configuration of s^2p^3, these materials have a deep-lying s-like lone pair but with the upper part of the valence band composed of p-like σ bonding orbitals associated with the threefold covalent coordination characteristic of these materials [see Fig. 3(c)] and so should not properly be regarded as lone-pair semiconductors. Examples of amorphous semiconductors categorized by this classification scheme are given in Table 2.

3. OPTICAL PROPERTIES

3.1 Electronic Structure

The absence of long-range order (translational periodicity) in amorphous semiconductors has a number of ramifications in relation to their electronic structure and cognate properties, e.g., optical behavior. The fact that there is no real-space periodic lattice means that correspondingly there is no reciprocal lattice either for amorphous materials, and so the electron wave vector **k** is no longer a good quantum number and cannot be used as a label for electron states, as in the case of crystalline materials. However, the density of states (DOS) *is* still valid as a description of electron states, even in the presence of disorder. Although the densities of states of amorphous and crystalline forms of the same (semiconducting) material are superficially similar—for example, the bandwidths and band gaps are often comparable (reflecting the marked similarities in the local bonding and short-range order in such cases)—nevertheless there can also be pronounced differences between the densities of states for amorphous and crystalline forms as a result of disorder. In particular, the sharp peaks in the DOS of crystalline semiconductors resulting, say, from Van Hove singularities (where $\nabla_{\mathbf{k}}E = 0$) are absent in the case of amorphous materials (see Fig. 4).

Table 2. Electronic classification of amorphous semiconductors

Material	Lone pair (L)/Bonding (B)
Si, Ge	B
GaAs	B
$CdGeAs_2$	B
As	B
$GeSe_2$, GeS_2	L
As_2Se_3, As_2S_3	L
Se	L

The lack of a reciprocal lattice, and the corresponding inutility of **k** as a label for electron states in amorphous materials, means that the additional selection rule for optical transitions in crystalline semiconductors, based on **k** conservation between initial and final states, breaks down in the case of amorphous semiconductors. In effect, *all* electronic states are accessible in an optical transition for an amorphous semiconductor possessing complete structural disorder. In practice, however, remnants of structural order—e.g., medium-range order—may cause partial reinstatement of the **k**-selection rule for optical transitions, the degree of which is still the subject of some debate.

An additional effect of the presence of structural disorder on electron states and related optical behavior is the broadening caused to the density of states at the band edges, bordering the forbidden gap. Fluctuations in bond lengths and angles result in the formation of "band tails," which appear to have an exponential dependence on energy for most amorphous semiconductors. A model for the ubiquitous existence of exponential band tails is the thermal-equilibration model of Bar-Yam *et al.* (1986), in which the atomic and electronic structures of an amor-

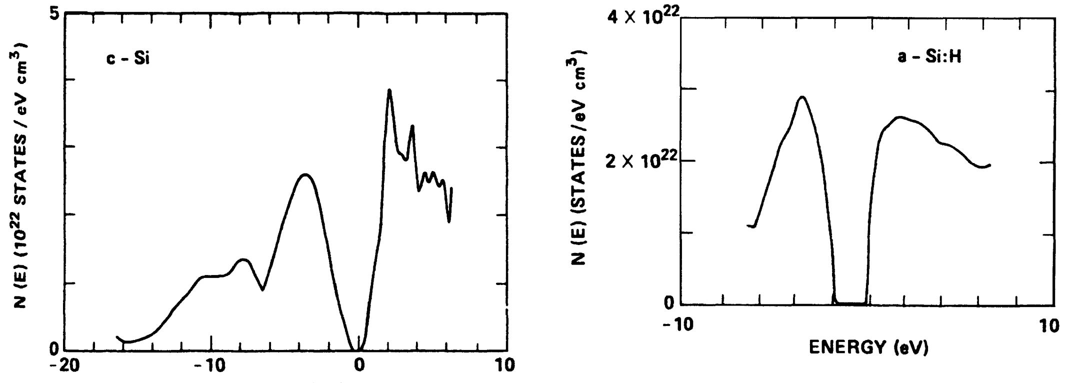

(a)

(b)

FIG. 4. Valence- and conduction-band DOS of (a) crystalline Si and (b) amorphous Si:H measured by x-ray photoemission (XPS) and inverse XPS, respectively (Jackson *et al.*, 1985).

phous solid are assumed to be interrelated: energetics determine the structure, which in turn controls the electronic density of states. The starting point is an ideal disordered network, containing no defects, with optimally coordinated atoms and equal bond lengths and angles. Fluctuations in bond lengths and angles, and defects, are then assumed to be generated thermally in such an ideal network, being in thermal equilibrium above, and frozen in below, a critical temperature T^*, which might be the glass-transition temperature of a glass or the film-deposition temperature for an amorphous thin film. The Boltzmann factor involved in the probability of creating above T^*, say, an extra filled electronic state at an energy $E - E_v$ above the valence-band mobility edge at E_v naturally leads to an exponential form for the valence-band–tail density of states (Bar-Yam *et al.*, 1986):

$$N_v(E) = N_0 \exp[-(E - E_v)/kT_v], \qquad (6)$$

where

$$T_v = T^*/(1 + \delta_v), \qquad (7)$$

and a similar expression for the conduction-band–tail state density T_c, where

$$T_c = T^*/\delta_c, \qquad (8)$$

and where δ_v and δ_c are parameters relating the changes in system energy and the disorder-induced raising and lowering of valence- and conduction-band states, respectively. An estimate for the ratio $\alpha = \delta_c/(1 + \delta_v) = T_c/T_v \simeq 0.5$ for Si obtained from tight-binding considerations (Bar-Yam *et al.*, 1986) is in accord with the experimental value of T_v = 500 K and T_c = 300 K for *a*-Si:H (Tiedje, 1984).

The final consequence of the absence of long-range order on the properties of electron states and associated optical behavior to be discussed concerns the *character* of the states: structural disorder can cause electron states to become spatially *localized* in the Anderson (1958) sense. A critical amount of disorder is necessary to localize electron states, with the consequence that there exist at the edges of the valence and conduction bands demarcation energies, so-called mobility edges, separating extended and localized states, where the magnitude of the electron mobility changes drastically (see, e.g., Elliott, 1990). Although there is no change in the DOS at a mobility edge, the change in character of the states (localized → extended) does have an influence on the optical behavior through the optical transition matrix element, being different for extended–extended (or extended–localized) and localized–localized transitions. With this in mind, it is convenient to divide, somewhat arbitrarily, the discussion about optical properties of amorphous semiconductors into three sections, dealing respectively with interband (valence–conduction band) transitions, transitions involving states in the band tails (the so-called Urbach-edge region), and transitions involving localized defect (and other) states deep in the gap between valence and conduction bands.

3.2 Interband Transitions

The optical properties of amorphous and crystalline semiconductors are almost entirely determined by the imaginary part (ϵ_2) of the complex dielectric constant ϵ^*,

$$\epsilon^* = \epsilon_1 + i\epsilon_2, \tag{9}$$

where ϵ_2 is related to the optical absorption coefficient $\alpha(\omega)$ by

$$\epsilon_2(\omega) = nc\alpha(\omega)/\omega \tag{10}$$

and where n is the refractive index and c is the speed of light. In a conventional optical-absorption experiment, α is obtained from the transmitted light intensity T and the reflectivity R via the Beer–Lambert expression

$$T = (1 - R)^2\exp(-\alpha d), \tag{11}$$

where d is the sample thickness.

For an amorphous semiconductor, where **k** conservation no longer holds because of the presence of structural disorder, ϵ_2 is related to a convolution of occupied valence and unoccupied conduction-band densities of states (the so-called "joint" DOS) and the momentum matrix element $P(\omega)$ (Connell, 1979):

$$\epsilon_2(\omega) = \frac{8\pi^2e^2a^3}{m^2\omega^2}P^2(\omega)\int N_v(E)N_c(E + \hbar\omega)\,dE, \tag{12}$$

where N_v and N_c are the valence-band and conduction-band DOS, respectively, and a is the interatomic spacing. A comparison of experimental curves of $\epsilon_2(\omega)$ for crystalline and amorphous Si is given in Fig. 5. Although superficially rather similar, the curve for *a*-Si is featureless, whereas that for *c*-Si exhibits a number of sharp features (Van Hove singularities) associated with the joint DOS. Furthermore, ϵ_2 for *c*-Si lies below that for *a*-Si at energies lower than the peak because optical transitions in *c*-Si at energies ~3 eV are "indirect," i.e., involve absorption/emission of a phonon to conserve wave vector (the band gap of *c*-Si at 1.1 eV is indirect), and correspondingly have a lower transition probability than the effectively "direct" transitions of *a*-Si.

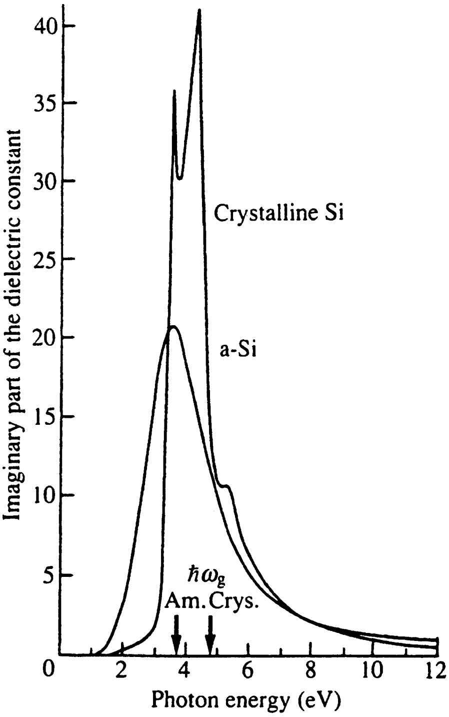

FIG. 5. Comparison of the imaginary part of the dielectric constant, $\epsilon_2(\omega)$, for crystalline and amorphous (*a*-) Si (Pierce and Spicer, 1972). The energy $\hbar\omega_g$ is the energy of the peak in ϵ_2.

The momentum matrix element $P^2(\omega)$ is related to the dipole-moment matrix element $R^2(\omega)$ via the following relation (Jackson *et al.,* 1985):

$$P^2(\omega) = (m\omega/\hbar)^2R^2(\omega). \tag{13}$$

The energy dependence of $R^2(\hbar\omega)$ can be obtained from the measured spectrum of $\epsilon_2(\omega)$, together with a knowledge of the DOS of valence and conduction bands, from Eq. (12). Figure 6 shows a comparison of $R^2(\hbar\omega)$ for *c*-Si and *a*-Si:H (Jackson *et al.,* 1985). The matrix element for *c*-Si increases rapidly for energies above the direct gap at ~3.4 eV,

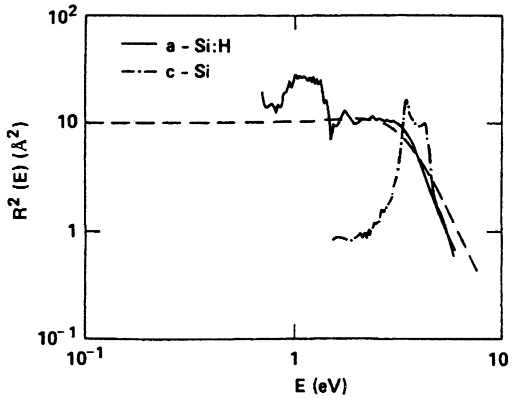

FIG. 6. Energy dependence of the dipole-moment matrix element $R^2(\omega)$ for *a*-Si:H and *c*-Si (Jackson *et al.*, 1985). The dashed line is a fit to the data for *a*-Si:H assuming a damped harmonic-oscillator model.

whereas that for *a*-Si:H is approximately constant below this energy, and decreases rapidly above, and can be fitted quite well with the form of the matrix element expected for a classical damped harmonic oscillator

$$R^2(\hbar\omega) \sim 1/[(\omega^2 - \omega_0^2) + \Gamma^2\omega^2] \tag{14}$$

with $\hbar\omega_0 \simeq 3.5$ eV and $\Gamma = 4$ eV.

The quantity of primary interest obtainable from interband optical transitions (leading to absorption coefficients in the range $10^3 < \alpha < 10^6$ cm^{-1}) is the optical gap. However, for the case of amorphous semiconductors there is no unique (nor commonly accepted) definition for this quantity. One type of definition is based on an arbitrary level set for the optical absorption coefficient, α, or the density of states, namely, the photon energy corresponding to $\alpha = 10^3$ cm^{-1} (E_{03}) or 10^4 cm^{-1} (E_{04}), or to the transition between the valence- and conduction-band densities of states equal to 10^{20} states/eV cm^3 (E_g^{20}).

Alternatively, an estimate for the optical gap can be obtained by extrapolation, assuming a particular form for the shape of the density of the band edges. For instance, if it is assumed that the conduction-band DOS scales as $N(E) \sim (E - E_c)p$ from the conduction-band edge at E_c, with a similar form for the valence-band DOS with exponent q, then Eq. (12) implies that

$$\omega^2\epsilon_2(\omega) \propto \omega\alpha(\omega) \propto P^2(\omega)(E - E_g)^{p+q+1} \propto \omega^2R^2(\omega)(E - E_g)^{p+q+1}, \tag{15}$$

where E_g is the (optical) gap defined in terms of the separation of the two extrapolated band edges. With the assumption that the momentum matrix element P^2 is independent of frequency and that the DOS at the band edges have a parabolic energy dependence $N(E) \propto E^{1/2}$, characteristic of free-electron states, then the so-called Tauc plot (Tauc *et al.*, 1966) for the dependence of the optical absorption coefficient on photon energies in the vicinity of the gap is obtained:

$$[\hbar\omega\alpha(\hbar\omega)]^{1/2} = A(\hbar\omega - E_g), \tag{16}$$

which is widely used to obtain estimates for E_g. However, as seen from Fig. 6 for the case of *a*-Si:H, it is instead the dipole-moment matrix element R^2 that is more nearly constant, and furthermore the energy dependence of the band tails is more nearly linear than parabolic, resulting in an appreciably different estimate for E_g. Table 3 contains a number of estimates for E_g obtained for *a*-Si:H using a variety of assumptions.

3.3 Urbach Edge

The part of the optical absorption tail corresponding to absorption coefficients in the range $1 \leq \alpha \leq 10^3$ cm^{-1} has, for all amorphous semiconductors, a universal exponential energy dependence

$$\alpha(\omega) = \alpha_0 \exp[(E - \hbar\omega)/E_0], \tag{17}$$

Table 3. Estimates for the energy gap for *a*-Si:H obtained from optical and density-of-states data under a variety of assumptions (Jackson *et al.*, 1985).

Energy gap or extrapolation	DOS energy dependence	Constant matrix element	Gap (eV)
E_g^μ (mobility gap)	—	—	1.93
E_g^{20}	—	—	1.61
E_{04}	—	—	1.85
E_{03}	—	—	1.68
$(\epsilon_2)^{1/2}$	$E^{1/2}$	R^2	1.64
$(\epsilon_2)^{1/3}$	E	R^2	1.49
$(E^2\epsilon_2)^{1/2}$	$E^{1/2}$	P^2	1.86
$(E^2\epsilon_2)^{1/3}$	E	P^2	1.55

where the energy E is comparable to the threshold energy obtained from interband optical absorption at higher photon energies (Sec. 3.2) and the slope parameter E_0 typically has values in the range 50–100 meV. This ubiquitous behavior is similar in its energy dependence to that found in alkali and silver halide crystals (Urbach, 1953) and is named after its discoverer; however, the temperature dependence is somewhat different.

The exponential character of the photon-energy dependence of $\alpha(\omega)$ is clearly evident in the data for *a*-Si:H shown in Fig. 7 and is now believed simply to reflect the exponen-

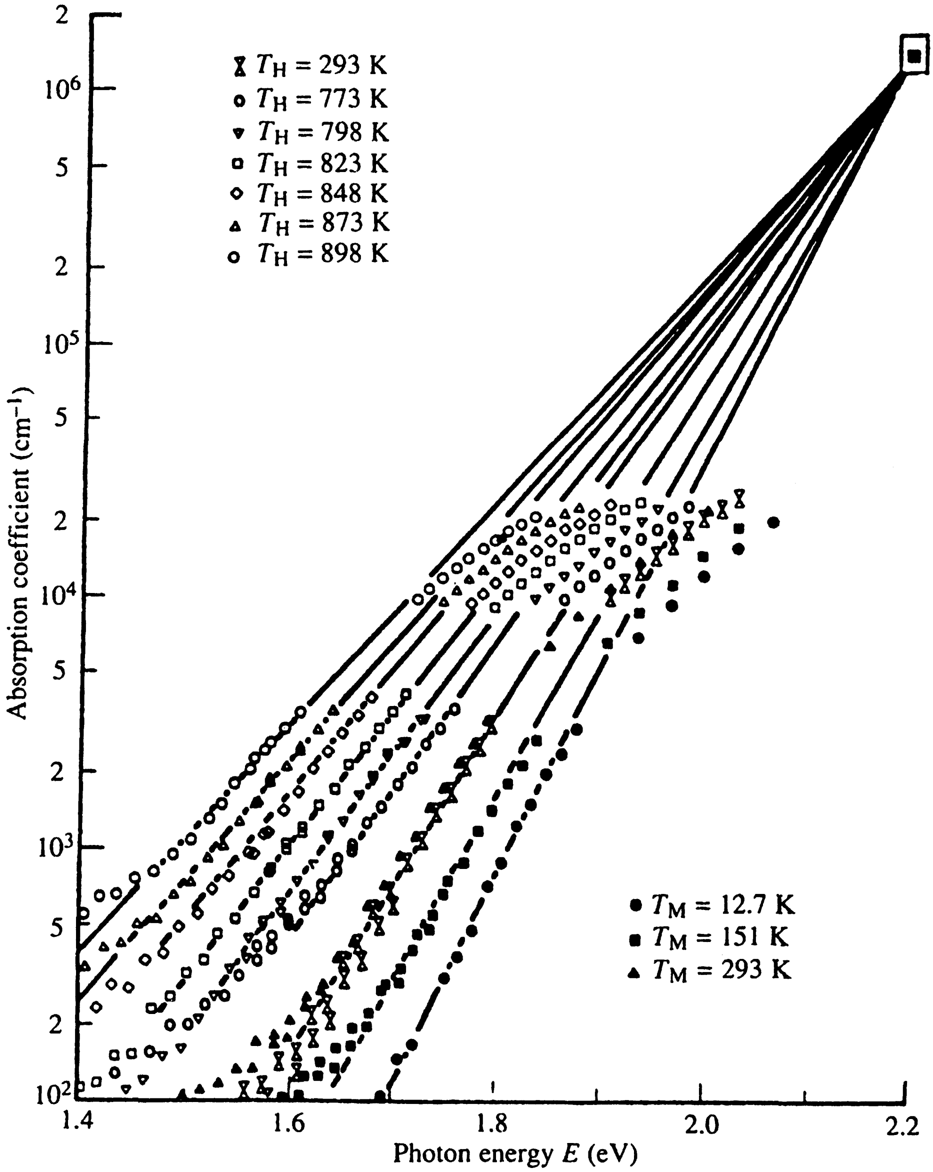

FIG. 7. Optical absorption edge of *a*-Si:H, showing Urbach-type behavior, measured at various temperatures T_M, and also measured at 300 K after annealing at the temperatures T_H shown (Cody, 1984).

tial energy dependence of the tail-state density of states [Eq. (6)], through the joint density of states [Eq. (12)], caused by disorder (static structural and thermal). However, in the case of *a*-Si:H, the Urbach slope reflects mostly the valence-band–tail state density since this is broader than the conduction-band–tail width.

The temperature dependence of the Urbach absorption edge can be understood theoretically in terms of a model involving both thermal and static disorder (Cody *et al.*, 1981; Street, 1991). The temperature dependence of the optical gap E_g can be regarded as arising predominantly from the electron–phonon interaction via the deformation potential, B (neglecting the thermal-expansion contribution of the network), viz.,

$$E_g(T) = E_g(0) - B[u^2(T) + u_D^2 - u^2(0)], \tag{18}$$

where $u^2(T)$ is the mean square displacement of the atoms due to thermal vibrations, $u^2(0)$ is that due to zero-point motion, and u_D^2 is that due to static disorder. The Urbach theory of optical absorption in crystals proposes that the slope parameter E_0 is proportional to $u^2(T)$. For amorphous semiconductors, there will be an additional contribution due also to static disorder u_D^2, so that

$$E_0(T) = K[u^2(T) + u_D^2], \tag{19}$$

where K is a constant. Combining Eqs. (18) and (19) gives

$$E_g(T) = E_g(0) - Bu^2(0) \times \left[\frac{E_0(T)}{E_0(0)}\left(1 + \frac{u_D^2}{u^2(0)}\right) - 1\right]. \tag{20}$$

Note that Eq. (20) predicts a linear correlation between the magnitude of the optical gap E_g and the Urbach slope parameter E_0 at any given temperature that is borne out experimentally [see Fig. 8(c)].

On the assumption of an Einstein model for the lattice vibrations, characterized by the Einstein temperature θ_E (which is $\frac{3}{4}$ of the Debye temperature θ_D), then

$$u^2(T)/u^2(0) = \coth(\theta_E/2T) \tag{21}$$

when M is the (effective) atomic mass. The

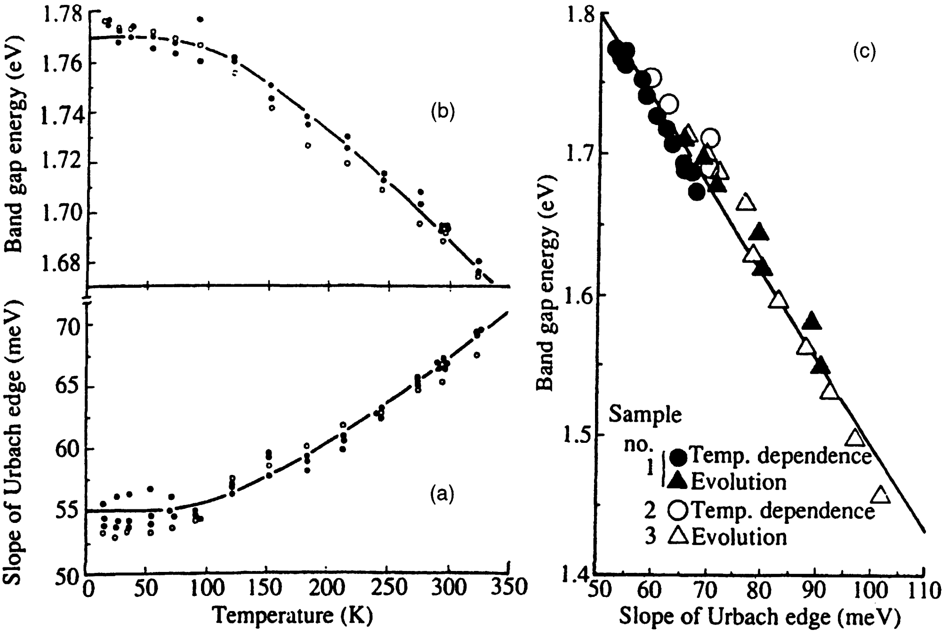

FIG. 8. Behavior of Urbach-edge parameters for *a*-Si:H (Cody *et al.*, 1981). (a) Temperature dependence of the Urbach slope parameter, E_0; (b) temperature dependence of the optical gap, E_g; (c) correlation between E_g and E_0.

explicit form for the temperature dependence of the optical gap, $E_g(T)$, is then obtained by substituting Eq. (21) into Eq. (18). For the case of *a*-Si:H, a fit to the data of $E_g(T)$ [obtained from the Tauc relation, Eq. (16)] is obtained with $\theta_D = 536$ K [see Fig. 8(b)], comparable to the literature value of $\theta_D = 625$ K for *c*-Si. The temperature dependence of the Urbach slope parameter, $E_0(T)$, is obtained in a similar fashion, with the added assumption (to eliminate the constant *K*) that $E_0 \to kT$ as $T \to \infty$ (Cody *et al.*, 1981), giving

$$E_0(T) = k\frac{\theta_E}{2}\left[\coth\left(\frac{\theta_E}{2T}\right) + \frac{u_D^2}{u^2(0)}\right]. \qquad (22)$$

Use of the same value of θ_D (or θ_E) as obtained from the fit to $E_g(T)$ also accounts for the data of $E_0(T)$ remarkably well [see Fig. 8(a)].

3.4 Subgap Transitions and Defect States

Amorphous semiconductors invariably contain structural defects, e.g., dangling bonds, whose electron states lie deep within the forbidden energy gap. These strongly localized states can also be probed optically, albeit with somewhat more difficulty than is the case for transitions involving band-edge or tail states (see Secs. 3.2 and 3.3). Since the absorption coefficient for transitions involving gap states is so low ($\alpha \leq 10$ cm^{-1}), reflecting the low level of the defect density of states, conventional absorption measurements to obtain values of α are precluded, particularly for thin-film amorphous semiconductors (e.g., *a*-Si:H). Instead, techniques such as photothermal deflection spectroscopy (PDS) can be used (see Amer and Jackson, 1984, for a review). The PDS detects, as heat, all the optical energy absorbed in a sample, which is proportional to αd when $\alpha d \ll 1$ (*d* being the sample thickness). Light from a chopped pump beam is absorbed by the sample, and the associated heat is detected by the deflection of a low-intensity probe beam, at grazing incidence to the film surface, caused by the thermally induced modulation of the refractive index of a medium (e.g., a liquid) in thermal contact with the sample.

Results of PDS measurements of defect gap states in *a*-Si:H, made by PECVD and deposited at different rf powers, are shown in Fig. 9(a); the correlation between the integrated gap-state absorption and the dangling-bond defect density as determined from electron-spin resonance is evident in Fig. 9(b).

4. VIBRATIONAL AND THERMAL PROPERTIES

4.1 Vibrational Density of States

In the case of crystalline materials, vibrational excitations can be described in terms of dispersion curves, $\omega(\mathbf{q})$, giving the dependence of the frequency of propagating collective modes (phonons) on the wave vector $\mathbf{q}$. The character of these modes can be distinguished most simply in the long-wavelength limit: "acoustic" branches are characterized by their frequency going to zero as $\mathbf{q}$ tends to 0, whereas "optic" branches have finite frequencies in the $\mathbf{q} = 0$ limit.

For the case of amorphous materials, the picture is complicated by the lack of translational periodicity, as it is for electronic states (Sec. 3.1). Thus, $\mathbf{q}$ is no longer a good quantum number nor a good label of vibrational states in general. Nevertheless, in the low-frequency, long-wavelength limit, amorphous solids appear structurally isotropic and act as elastic solids; as a consequence, vibrational excitations can still be described as acoustic phonons, albeit more heavily damped (with a shorter lifetime) than for crystals. At higher $\mathbf{q}$ values, however, the $\omega(\mathbf{q})$ relation becomes increasingly smeared out (see Fig. 10), and the modes will no longer be plane-wave states (as for crystals) and will be more localized.

However, the vibrational density of states (VDOS) is, like its electronic counterpart (Sec. 3.1), still valid as a description of vibrational states in amorphous materials. The VDOS may be determined experimentally by inelastic neutron scattering (see Elliott, 1990), although IR absorption and Raman scattering spectra can, under certain circumstances, also give an insight into the VDOS; however, the existence of frequency-dependent matrix elements in the latter two cases often precludes a quantitative determination of the VDOS by these means. Figure 11

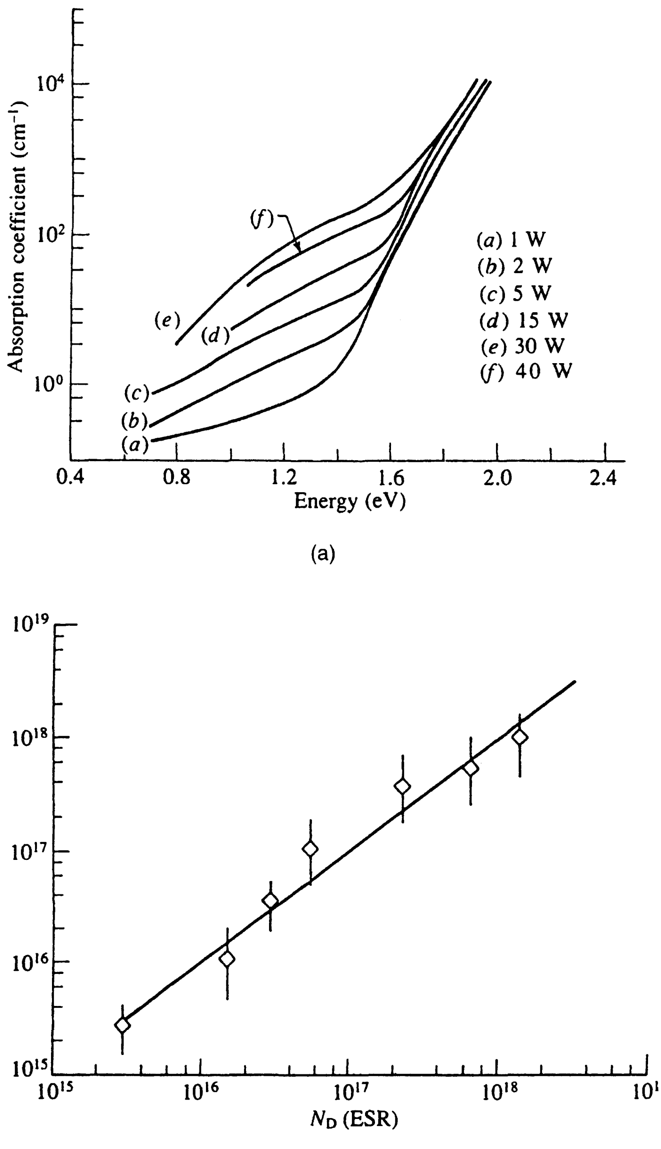

(a)

(b)

FIG. 9. (a) PDS measurements of the Urbach edge and gap-state transitions in PECVD *a*-Si:H deposited at the various rf powers shown (Jackson and Amer, 1982). (b) Correlation between the integrated gap-state absorption and the ESR spin density for various samples of *a*-Si:H (Jackson and Amer, 1982).

shows the experimental (inelastic-neutron) VDOS for *a*-Si with, for comparison, that calculated for *c*-Si, as well as Raman spectra for *a*- and *c*-Si. It can be seen that the density-of-states distributions are rather similar for crystalline and amorphous forms, except that, as for the case of the electronic density of states (Sec. 3.1), the sharp cusplike features (Van Hove singularities) characteristic of the crystalline density of states are smeared out in the amorphous case. This overall similarity is not too surprising since the short-range order is the same (tetrahedral) in both cases, with very similar bond lengths and angles, and hence they have correspondingly similar bond-stretching and bond-bending force constants. Note also that, because of the lack of translational periodicity in the amorphous phase, and the corresponding absence of a well-defined reciprocal lattice, the wave vector **q** is no longer a good quantum number and that the

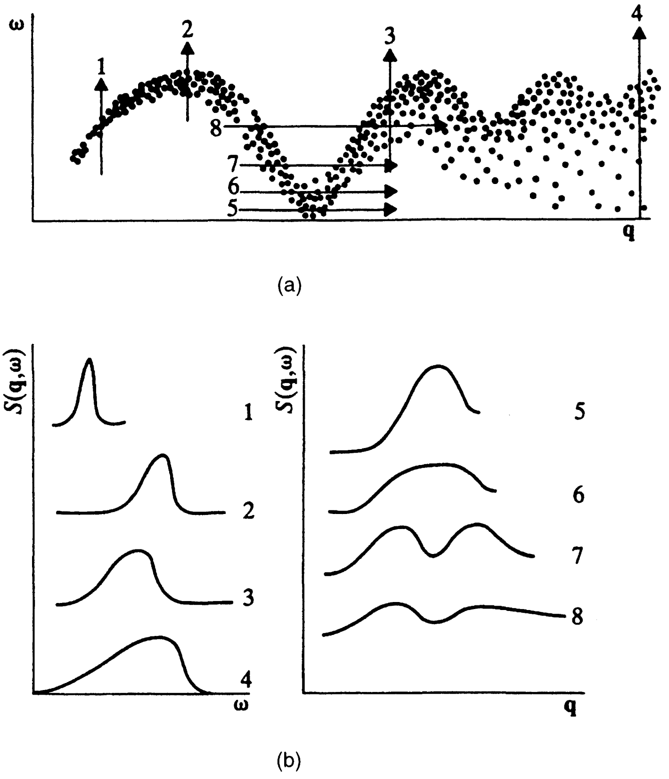

(a)

(b)

FIG. 10. (a) Schematic illustration of a phonon dispersion relation $\omega(\mathbf{q})$ for an acoustic branch in an amorphous material. The form of the scattering law, i.e., the dynamic structure factor $S(\mathbf{q},\omega)$, is shown in (b) for various cuts through the curve in (a) (Leadbetter, 1973).

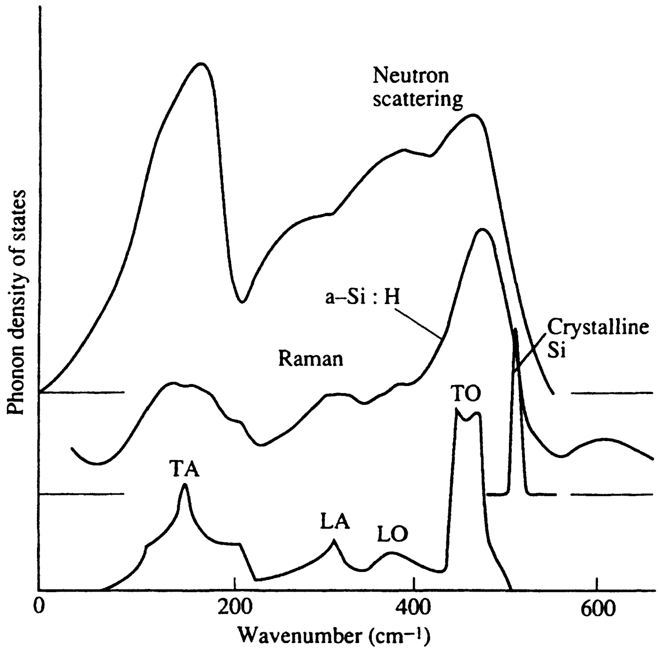

FIG. 11. Vibrational density of states for *a*-Si:(:H) measured by inelastic neutron scattering (Kamitakahara *et al.*, 1984) and Raman scattering (Tsai and Nemanich, 1980), compared with the Raman spectrum for *c*-Si and a calculated VDOS for *c*-Si, where the transverse (T) and longitudinal (L) modes of the acoustic (A) or optic (O) phonons are marked (Street, 1991).

selection rule $\Delta\mathbf{q} = 0$ governing optical transitions between vibrational states in crystals is relaxed in the amorphous state. Thus, although the transition in the case of the Raman spectrum of the crystal is restricted to zone-center phonons (thereby having a frequency corresponding to a transition, between lowest acoustic and highest optic bands, equal to the maximum band width of the VDOS), in the case of amorphous semiconductors, in principle, transitions between *all* vibrational states are allowed because of the breakdown of the **q**-selection rule, and as a result, the Raman spectrum more nearly represents the VDOS (see Fig. 11).

4.2 Propagating and Localized Phonons

Phonons, like electron states, are extended, propagating modes in the case of single crystals (without impurities, defects, etc.). However, the structural disorder characteristic of amorphous materials can serve to localize phonons, in the same way as for electron states, giving rise to a mobility edge at a frequency ω_c, below which phonons are extended and above which they are localized. The existence of such a mobility edge is signaled by the precipitous decrease in the phonon mean free path l (obtained from thermal-conductivity measurements; see Sec. 4.5) as the phonon wavelength decreases, or alternatively as the temperature increases (Zeller and Pohl, 1971)—see Fig. 12. This behavior is observed for all types of nonmetallic amorphous solid, semiconductors (Se) and insulators (GeO_2 and SiO_2) alike. The mean free path of the glasses reaches a minimum saturation value when the inverse phonon wave vector, $k = 2\pi/\lambda$, is comparable to l, the so-called Ioffe–Regel (1960) limit for localization,

$$kl \simeq 1. \tag{23}$$

Additional evidence for the existence of phonon localization comes from the anomalously long phonon lifetime (longer by three orders of magnitude than that for *c*-Si) found in *a*-Si:H for frequencies $\omega \geq 100$ cm^{-1} from time-resolved Raman experiments (Scholten *et al.*, 1993). This result can be understood if, say for the case of localized TO phonons (480 cm^{-1}), decay into two high-frequency extended acoustic phonons cannot take place as in the crystalline case, but instead a low-probability "hopping" process (Orbach, 1992) occurs, whereby decay can only take place via another high-frequency localized phonon and a low-frequency extended phonon.

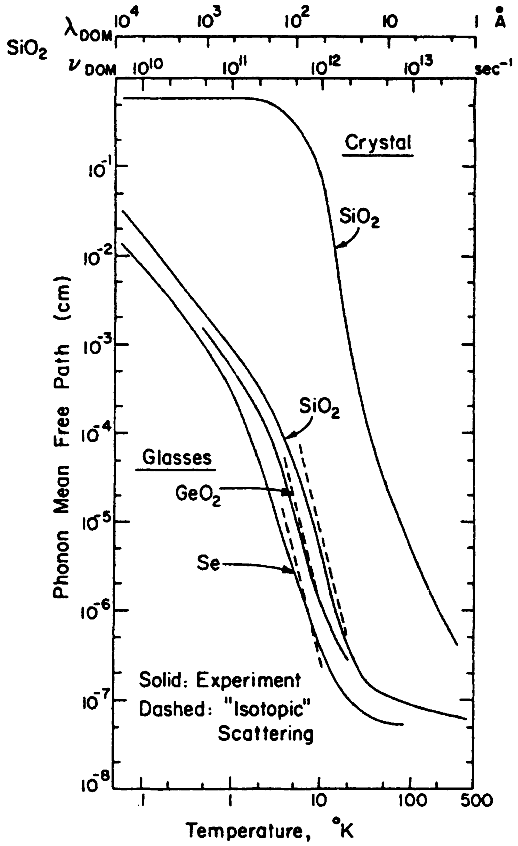

FIG. 12. Variation of phonon mean free path with temperature (or equivalently phonon wavelength) for glassy Se, GeO_2, and SiO_2, compared with that for crystalline SiO_2 (Zeller and Pohl, 1971).

The precise cause of the phonon localization is still controversial, although the frequency dependence $l \propto \omega^{-4}$ deduced from Fig. 12 is indicative of Rayleigh scattering controlling the mean free path, and a model invoking elastic Rayleigh scattering from atomic-density fluctuations in amorphous solids on a length scale corresponding to short- and medium-range order has been proposed by Elliott (1992). Scattering of phonons by "soft configurations" (Karpov *et al.*, 1983), groups of atoms having anomalously low-frequency harmonic or anharmonic potentials (see Sec. 4.3), has also been suggested (Buchenau *et al.*, 1992) to be responsible for the behavior of the mean free path for frequencies below the mobility edge.

Although not of direct relevance to *fully dense* amorphous semiconductors, mention should also be made briefly of the localized vibrational excitations known as "fractons" (Orbach, 1986) that are supported by self-similar fractal structures for which the mass scales with distance as

$$M(r) \propto r^{D_{\mathrm{H}}}, \tag{24}$$

where the Hausdorff dimension $D_{\mathrm{H}} \leq d$ (d is the normal Euclidean dimension), over the range of length scales $a < r < \zeta$, where a is a typical interatomic distance and ζ marks the distance beyond which the material stops being self-similar and becomes structurally homogeneous. Thus, at frequencies lower than a crossover frequency ω_{CO}, normal phonon modes exist; at frequencies higher than ω_{CO}, localized fracton modes exist instead, for which the vibrational density of states has the frequency dependence

$$\rho(\omega) \propto \omega^{\tilde{d}-1} \tag{25}$$

where $\tilde{d}(\leq D_{\mathrm{H}} \leq d)$ is the so-called fracton dimension, conjectured to have the universal value of $\frac{4}{3}$ (Alexander and Orbach, 1982), compared with the Debye-type behavior $\rho(\omega) \propto \omega^2$ for plane-wave phonons.

There is no experimental evidence for fractallike structures in fully dense amorphous semiconductors in either bulk-glass or thin-film forms, and so fractons cannot be responsible for the anomalous behavior observed in either the heat capacity (Sec. 4.4) or thermal conductivity (Sec. 4.5) in such materials. Nevertheless, in those materials, such as silica aerogels, that *do* possess a fractallike structure, fracton behavior does appear to be observed (Courtens *et al.*, 1988).

4.3 Soft Configurations and Two-Level (Tunneling) Systems

Analysis of the vibrational response of crystalline materials is almost invariably carried out within the harmonic approximation. However, in the case of amorphous materials, local variations in the atomic structure associated with disorder can lead to varying degrees of anharmonicity in the interatomic potentials, which can have a profound effect on the thermal properties of glasses, such as heat capacity (Sec. 4.4) and thermal conductivity (Sec. 4.5).

A general, albeit phenomenological, model accounting for such anharmonicity is the so-called soft-configuration model (Karpov *et al.*, 1983; Galperin *et al.*, 1989) that considers soft local modes involving an atom, or group of atoms, having the following empirical form for the potential as a function of displacement x in terms of a generalized coordinate:

$$V(x) = E[\eta(x/a)^2 + \xi(x/a)^3 + (x/a)^4]. \tag{26}$$

Here a is a distance of the order of the interatomic spacing, and E sets the energy scale and is given approximately by $E \simeq mv_s^2$, where m is the atomic mass and v_s is the sound velocity. The coefficients η and ξ are assumed to be randomly distributed in amorphous materials and, depending on their values, Eq. (26) describes harmonic or anharmonic single-well, and double-well, potentials.

In this sense, the soft-configuration model is a generalization of the two-level (or tunneling-state) one proposed earlier by Anderson *et al.* (1972) and Phillips (1972) to account for the low-temperature ($T < 1$ K) thermal anomalies of glasses (see Secs. 4.4 and 4.5). The two-level model, too, is a phenomenological model and assumes that certain groups of atoms in amorphous solids can occupy one of two equilibrium positions corresponding to a double-well potential (Fig. 13). At sufficiently low temperatures, atomic tunneling can occur through the potential barrier separating the two potential minima, and the resulting tunnel-split energy levels could be one origin of two-level systems. Although, as will be seen later (Secs. 4.4 and 4.5), the two-level model accounts well for the low-temperature thermal behavior of glasses, nevertheless it is still a phenomenological approach lacking a proper microscopic (atomistic) explanation. Although a motion of coupled (corner-shared) tetrahedra has been proposed to be the mode responsible in the case of v-SiO$_2$ (Buchenau *et al.*, 1986), this cannot be generally the case for all amorphous materials

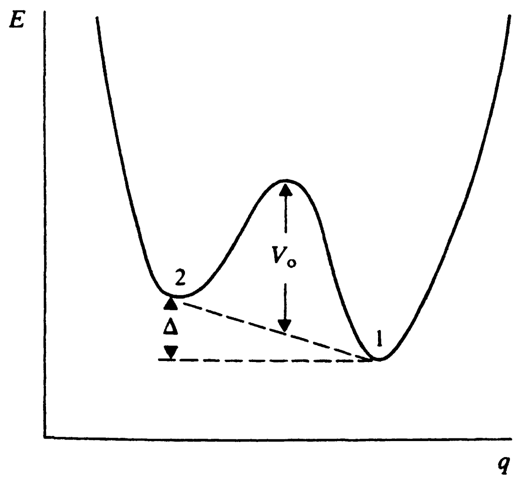

FIG. 13. Schematic illustration of the double-well potential characterizing a two-level system plotted as a function of some configuration coordinate, q. The parameters Δ (asymmetry energy) and V_0 (barrier height) are assumed to be randomly distributed in a glass.

(metals, semiconductors, and insulators) that exhibit the ubiquitous anomalous low-temperature thermal behavior.

4.4 Heat Capacity

The lattice-vibrational contribution to the heat capacity of crystalline solids at low temperatures (less than the Debye temperature θ_D) is well accounted for by the famous Debye law, $C_v \propto T^3$. In the same temperature regime, the behavior of *all* glasses (metallic and nonmetallic, alike) is different (see Phillips, 1981b, for a review): the nonelectronic, atomic-vibrational contribution to the temperature dependence of the heat capacity can be represented as (Zeller and Pohl, 1971)

$$C_v = c_1 T^\alpha + c_3 T^3, \qquad (27)$$

where the exponent α has a value close to unity. This behavior (for the case where $\alpha = 1$) can be confirmed most simply by plotting heat-capacity data in the form C_v/T vs T^2: the intercept gives the coefficient c_1 and the gradient gives c_3.

Such a plot for the case of samples of As_2S_3, with varying impurity concentrations, is shown in Fig. 14 (Stephens, 1976). It is apparent that there is a marked extrinsic effect manifested in the heat capacity, although no one-to-one correspondence with the impurity content is evident. Nevertheless, the highest-purity glass does indeed show a temperature dependence for the heat capacity of the form given by Eq. (27), and it is apparent, as is generally the case, that the coefficient c_3 for the cubic term is appreciably larger than the value expected from simple Debye theory. Indeed, at temperatures of the order of 10 K, all glasses exhibit a peak in the temperature dependence of the heat

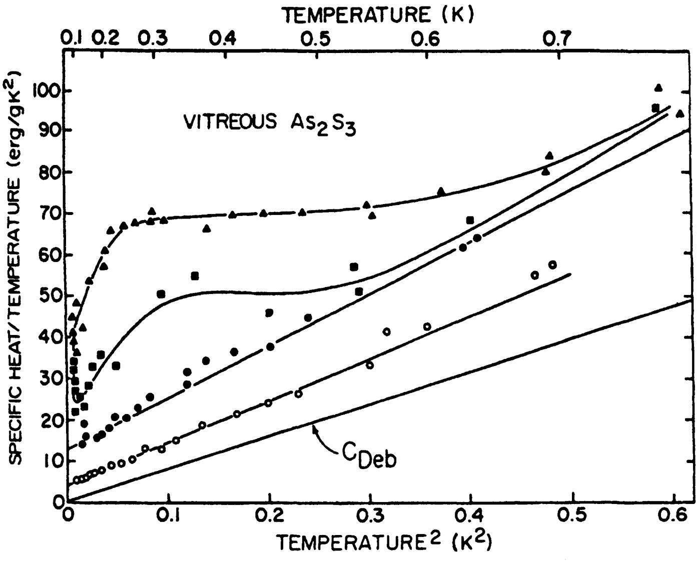

FIG. 14. Specific heat at low temperatures for four samples of a-As_2S_3 with varying impurity contents, plotted as C_v/T vs T^2 (Stephens, 1976). The straight line through the origin is the heat capacity predicted by the Debye model; the presence of the anomalous linear term in the specific heat is marked by the nonzero intercept of the data in such a plot. The purity of the samples (in terms of both metal and water content) increases from the upper to the lower set of data.

capacity normalized to the Debye (T^3) law (see Fig. 15), indicative of the existence of a low-frequency peak in the VDOS (again normalized to the corresponding Debye ω^2 form) that is also widely observed in low-frequency Raman scattering spectra of glasses (the so-called "boson peak").

The anomalous, almost linearly temperature-dependent specific heat observed in glasses at low temperatures (<1 K) was first accounted for by the two-level system model (Anderson *et al.*, 1972; Phillips, 1972). The heat capacity in this case is given by

$$C_v = k_B \int_0^\infty n(E)\left(\frac{E}{2k_BT}\right)^2 \operatorname{sech}^2\left(\frac{E}{2k_BT}\right) dE, \qquad (28)$$

where E is the energy separation between the two levels and $n(E)$ is the corresponding density of states. For the case of the tunneling model, the energy is given by

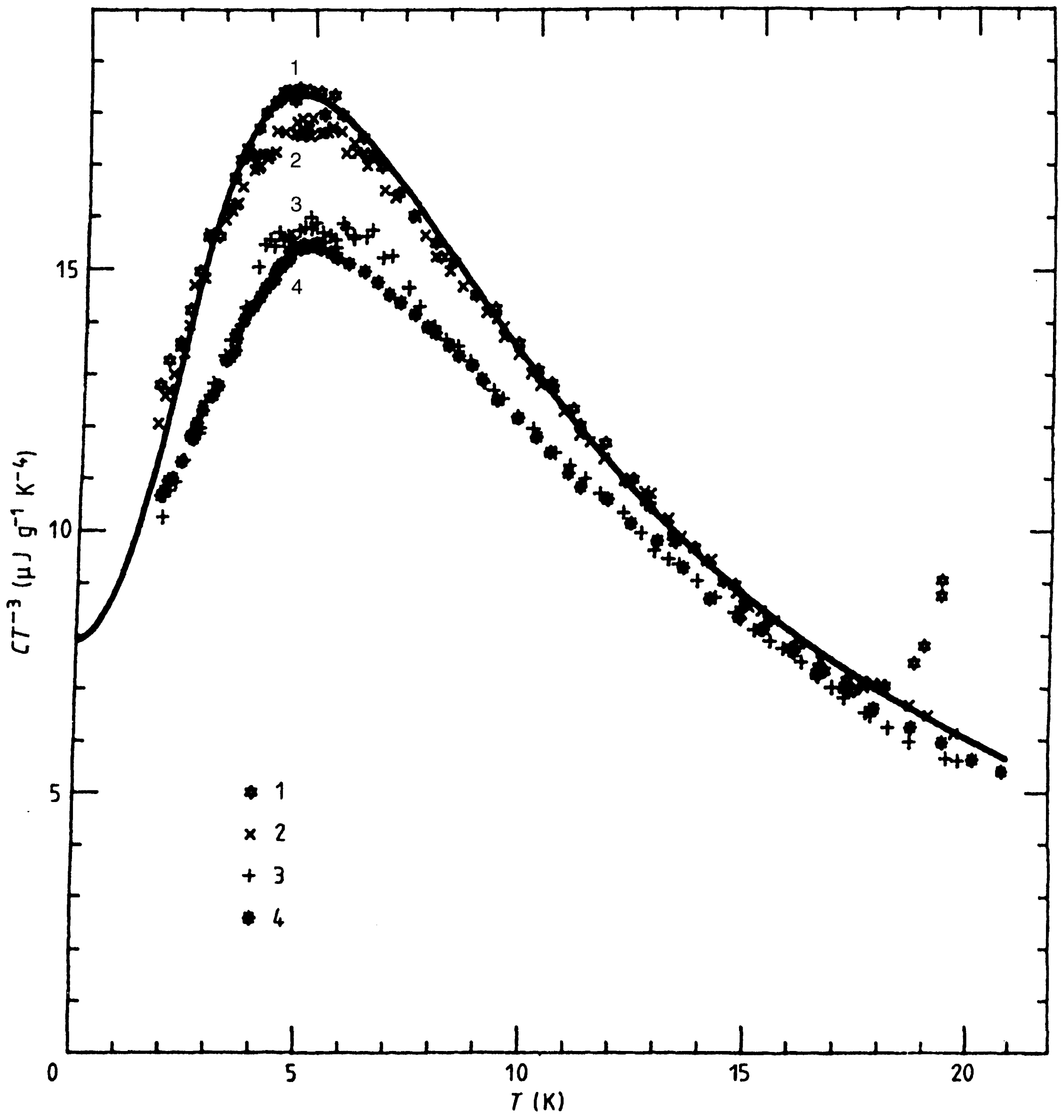

FIG. 15. Heat capacity of four samples of bulk glassy As_2S_3 (as quenched, samples 1 and 2; annealed, samples 3 and 4) normalized with respect to the Debye T^3 form (Ahmad *et al.*, 1986).

$$E^2 = \Delta^2 + \Delta_0^2, \tag{29}$$

where Δ_0^2 is the energy difference between the two lowest tunnel-split states in the symmetric-well case ($\Delta = 0$, see Fig. 13), and

$$\Delta_0 = \hbar\omega_0 e^{-\lambda}, \tag{30}$$

where

$$\lambda = (2mV_0/\hbar^2)^{1/2}d \tag{31}$$

with ω_0 being the oscillation frequency in an individual well. Evaluation of Eq. (28), assuming that the density of states is constant in energy, $n(E) = n_0$, over the energy range of interest (10^{-5}–10^{-4} eV), gives

$$C_v = (\pi^2/6)n_0 k_B^2 T. \tag{32}$$

If, instead, a weak energy dependence of $n(E)$ is admitted [$n(E) \propto E^m$, with $m = 0.1$–0.3], then a slightly superlinear temperature dependence of the heat capacity results, $C_v \propto T^{1+m}$, in accord with many experimental data (Stephens, 1976).

The more general soft-configuration model (Karpov *et al.*, 1983) treats the low-temperature (<1 K) linear-temperature-dependence regime of the heat capacity in terms of anharmonic excitations (identical to tunneling states), but it can also account for data at higher temperatures (<10 K) in terms of excitations involving soft harmonic potentials (Buchenau *et al.*, 1991).

An interpretation of the low-temperature heat capacity in terms of extra excitations associated with two-level systems or soft configurations is understandable for the case of the amorphous semiconductors *a*-Se or As_2S_3 and the glassy insulators SiO_2 and B_2O_3, since they all contain two-fold coordinated group-VI atoms (O,S,Se) that it is easy to envisage could be involved in the soft modes responsible. However, it is more difficult to understand in this manner that the notionally tetrahedrally coordinated material *a*-Si:H also shows a quasi-linear temperature dependence of the heat capacity, with a value of the coefficient c_1 (Eq. 27) almost equal to that for *v*-SiO_2 (Graebner *et al.*, 1984). It is conceivable that singly coordinated hydrogen, or low-coordinated atoms in the vicinity of the voids responsible for the 14% density deficit (relative to *c*-Si), could be the origin of the tunneling states, but it remains a remarkable coincidence that the densities of two-level systems are the same as between *a*-Si:H and SiO_2. A proper, microscopic, understanding of the heat capacity of amorphous semiconductors (and insulators) is still lacking.

4.5 Thermal Conductivity

Thermal transport in crystalline nonmetallic materials occurs by means of propagating phonons. The thermal conductivity may be written, from kinetic theory, as

$$\kappa = \tfrac{1}{3}\sum_i \int_0^{\omega_D} C_i(\omega)v_i(\omega)l_i(\omega)\,d\omega, \tag{33}$$

where the sum is over the three longitudinal and transverse phonon modes, v and l are the phonon velocity and mean free path, respectively, and $C(\omega)$ is the contribution to the specific heat from phonons of frequency ω. The temperature dependence of κ for a crystal typically exhibits a peak at low temperatures; above, say 10 K, phonons are scattered by intrinsic (phonon–phonon umklapp) processes, and hence l and, consequently, κ decrease with increasing temperature since the number of phonons increases; at lower temperatures, l increases with decreasing temperature and eventually becomes equal to the sample dimensions and hence is constant, and thereafter $\kappa(T)$ follows the temperature dependence of $C_v(T)$ and is therefore proportional to T^3 [see Fig. 16(a)].

In contrast, the behavior of all amorphous nonmetallic materials is completely different in two respects (Zeller and Pohl, 1971; Freeman and Anderson, 1986): no peak in $\kappa(T)$ is evident, but instead there is a plateau in the range 10–100 K for most amorphous materials [see Fig. 16(b)], although this appears to start at 100–200 K for the case of *a*-Si:H (Cahill *et al.*, 1994). The plateau is followed by a further rise at higher temperatures; this is not evident in Fig. 16(b) but is shown in Fig. 17 for the case of the amorphous semiconductor *a*-Se and some insulating oxide glasses. Furthermore, $\kappa(T)$ for amorphous materials also differs from that of crystals at low temperatures: $\kappa(T) \propto T^2$ for amorphous materials in general (although the limiting low-temperature behavior of *a*-Ge shown in

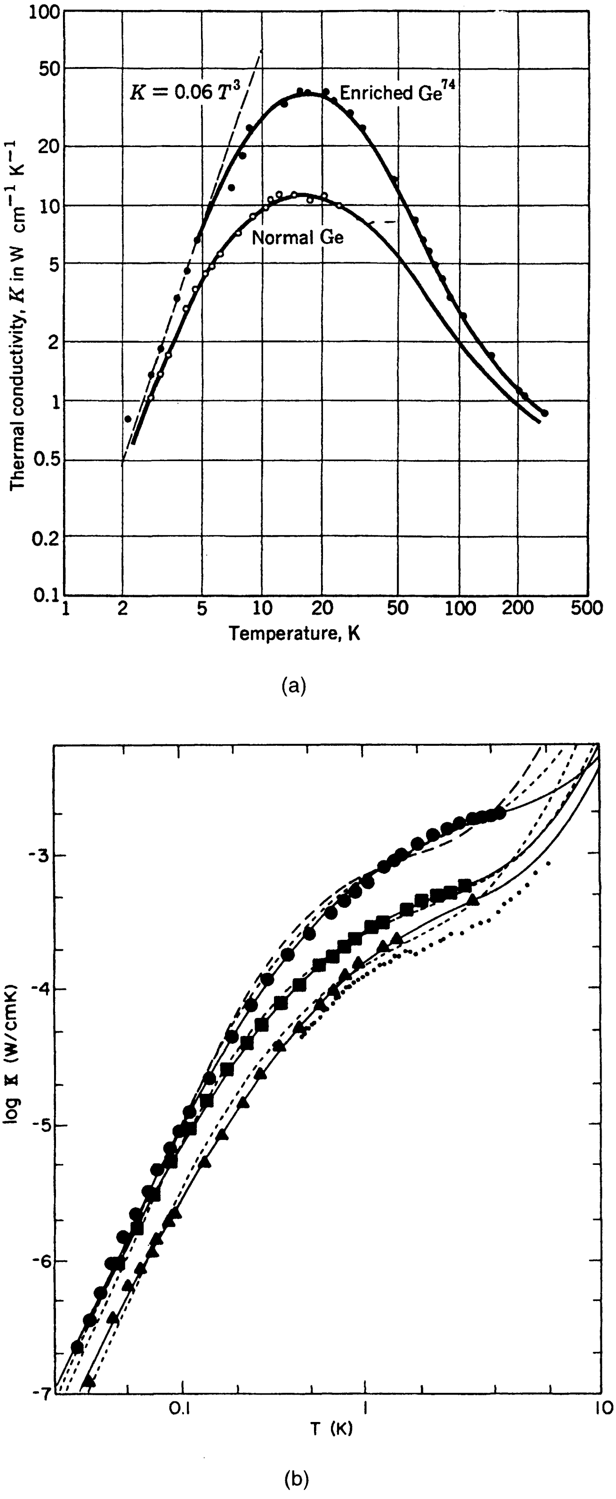

(a)

(b)

FIG. 16. Temperature dependence of the thermal conductivity of Ge: (a) *c*-Ge (Kittel, 1976); (b) *a*-Ge (Graebner and Allen, 1983). The three data sets in (b) refer to different samples of *a*-Ge with, from top to bottom, decreasing densities.

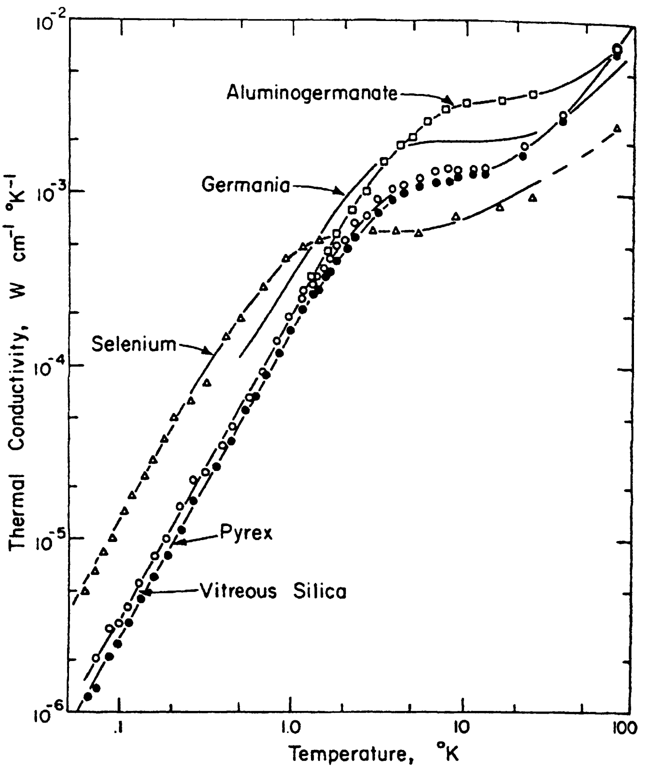

FIG. 17. Temperature dependence of the thermal conductivity of some semiconducting and insulating glasses (Zeller and Pohl, 1971).

Fig. 16(b) appears to be proportional to T^3 as for crystals).

The low-temperature behavior of $\kappa(T)$ for amorphous materials has been explained in terms of the additional phonon-scattering centers controlling the mean free path at low temperatures (<1 K), i.e., two-level systems (see, e.g., Phillips, 1981b; Freeman and Anderson, 1986) or, more generally, soft configurations (Buchenau *et al.*, 1992). Phonons having an energy $\hbar\omega = E$, the tunnel splitting, can be scattered resonantly by a process involving excitation from the ground state of the two-level system, followed by the subsequent decay and emission of an incoherent phonon, for which the mean free path is (Phillips, 1981b)

$$l_{res} = \frac{\rho v^3}{\pi n_0 M^2 \omega} \coth\left(\frac{\hbar\omega}{2k_B T}\right) \qquad (34)$$

where M is the coupling constant relating E to the strain. In the dominant-phonon approximation, $\hbar\omega \sim k_B T$, and so Eq. (34) predicts that $l_{res} \propto T^{-1}$, which, taken together with $C_v \propto T^3$ (because the heat is carried by phonons at temperatures below the onset of the plateau in κ), gives from Eq. (33) $\kappa \propto T^2$, as found experimentally (Zeller and Pohl, 1971; Freeman and Anderson, 1986). However, although the phenomenological tunneling-state model can account for the functional form of $\kappa(T)$ for $T \leq 1$ K, it is difficult to understand thereby the reason for the empirical relation

$$l/\lambda = \rho v^2 / M^2 n_0 \simeq 150 \qquad (35)$$

that is apparently valid for all glasses (Freeman and Anderson, 1986).

At higher temperatures ($T > 1$ K) or higher frequencies ($\omega \geq 10^{11}$ s^{-1}), the phonon mean free path decreases precipitously according to

$$l \propto \omega^{-n} \quad (n = 3\text{–}4). \qquad (36)$$

This behavior therefore accounts for the temperature independence of $\kappa(T)$ in the so-called plateau region. Possible physical mechanisms that could give rise to relation (36) are Rayleigh scattering of phonons by density fluctuations, for which $n = 4$ (see Elliott, 1992, for a recent discussion), or alternatively scattering by soft configurations with an assumed density of quasiharmonic states $n(\omega) \propto \omega^4$ (Buchenau *et al.*, 1992).

At the upper frequency/temperature end of the plateau in κ, the mean free path has become reduced to a length of the order of interatomic spacings, $l \sim a$, and also to the dimension of the phonon wavelength (see Fig. 12), suggesting, from Eq. (23), that phonon localization is taking place there. The upturn in κ observed above the plateau region (Fig. 17) is therefore surprising and indicative of another mechanism being responsible for heat transport in this high-temperature/frequency region since the propagating phonons are in the Dulong–Petit limit for which $\kappa(T)$ = constant. Neither the two-level system nor the soft-configuration model can account for this behavior (Buchenau *et al.*, 1992). Orbach (1992, 1993) has proposed that a "phonon-hopping" process occurs in this regime, analogous to the case of variable-range hopping of localized electrons (see, e.g., Mott and Davis, 1979; Elliott, 1990), in which, as a result of *anharmonic* interactions, a localized phonon with frequency ω "hops" to a neighboring site, thereby producing a localized phonon of frequency ω' there, mediated by a (low-frequency) extended phonon of frequency ω_0 ($\omega = \omega' + \omega_0$). A calculation of $\kappa(T)$ appropriate for this situation (taken from a treatment for fractons in fractal structures) gives (Orbach, 1992, 1993)

$$\kappa \propto T \tag{37}$$

for temperatures greater than the plateau region, in qualitative agreement with experimental data (Zeller and Pohl, 1971).

GLOSSARY

Amorphous Solid: A noncrystalline material lacking long-range translational or orientational structural order.

Band Tail: Electron states extending into the forbidden energy (band) gap of an amorphous (*q.v.*) semiconductor, either up from the valence band or down from the conduction band, resulting from the presence of disorder.

Boson Peak: A peak at low frequencies in the vibrational density of states (*q.v.*) of amorphous solids (*q.v.*), most often detected by Raman scattering, whose temperature dependence follows the bose factor.

Chalcogenide: An alloy of a chalcogen (an element from group VI of the periodic table, i.e., S, Se, or Te) with other elements [often Si, Ge, B, As, P, etc. in the case of amorphous materials (*q.v.*)].

Chemical Vapor Deposition (CVD): Method of producing solid materials in thin-film form by means of chemical reactions between precursors in the vapor phase and a solid substrate.

Dangling Bond: A defect comprising an undercoordinated atom, ubiquitous in covalent amorphous materials (*q.v.*).

Density of States (DOS): The number of states (electronic or vibrational) at a particular energy (frequency) per unit energy (frequency) interval.

Fractal: A self-similar structure for which the mass scales with distance as a power law [Eq. (24)] with an exponent less than that (3) characteristic of a fully dense material.

Fracton: A localized (*q.v.*) vibrational excitation supported by a fractal (*q.v.*) structure.

Glass: An amorphous solid (*q.v.*), often prepared by melt quenching, that exhibits a glass transition (discontinuity in the heat capacity).

Glow-Discharge (GD) Decomposition: Synonym for plasma-enhanced chemical vapor deposition (*q.v.*).

Localized State: An electron (or vibrational) state whose wave function is confined to a particular spatial region for all times, as a result of disorder.

Lone Pair: Nonbonding pair of electrons in an orbital.

Medium-Range Order: Structural organization in an amorphous solid (*q.v.*) at distances beyond that characteristic of short-range order (*q.v.*), i.e., 5–20 Å.

Mobility Edge: Demarcation energy separating localized (*q.v.*) from extended (electron or vibrational) states.

Polaron: A charge carrier (electron or hole) associated with lattice distortion.

Short-Range Order: Structural organization in an amorphous solid (*q.v.*) at the shortest length scale (2–5 Å), associated with bond lengths and angles and coordination numbers.

Soft Configuration: A group of atoms in an amorphous solid (*q.v.*) supporting anomalously low-frequency vibrational modes and associated with strongly anharmonic vibrational potentials [Eq. (26)].

Sputtering: Method of thin-film deposition based on ablation of a target by ions from a neighboring gas-phase plasma.

Two-Level System: Double-well vibrational potential for a "soft" configuration (*q.v.*) of a group of atoms, for which atomic tunneling through the potential barrier separating the two wells can take place at low temperatures.

Variable-Range Hopping: Tunneling motion of a localized particle (electron or phonon) between sites with varying separation, such that more distant coupled sites have smaller energy separations, and vice versa.

Urbach Edge: Exponential dependence of optical absorption coefficient on photon energy [Eq. (17)] associated with electronic transitions between band-tail (*q.v.*) electron states distributed exponentially in energy in the case of semiconducting amorphous materials (*q.v.*).

Works Cited

Ahmad, N., Hutt, K. W., Phillips, W. A. (1986), *J. Phys. C* **19,** 3765–3773.

Alexander, S., Orbach, R. (1982), *J. Phys. (Paris) Lett.* **43,** L625–L631.

Amer, N. M., Jackson, W. B. (1984), in: J. Pankove (Ed.), *Semiconductors and Semimetals,* New York: Academic, Vol. 21B, p. 83.

Anderson, P. W., Halperin, B. I., Varma, C. M. (1972), *Philos. Mag.* **25,** 1–9.

Anderson, P. W. (1958), *Phys. Rev.* **109,** 1492–1505.

Bar-Yam, Y., Adler, D., Joannopoulos, J. D. (1986), *Phys. Rev. Lett.* **57,** 467–470.

Brazhkin, V. V., Lyapin, A. G., Popova, S. V., Voloshin, R. N. (1992), *Pis'ma Zh. Eksp. Teor. Fiz.* **56,** 156–159 (*JETP Lett.* **56,** 152–156).

Buchenau, U., Galperin, Yu. M., Gurevich, V. L., Schober, H. R. (1991), *Phys. Rev.* **B43,** 5039–5045.

Buchenau, U., Galperin, Yu. M., Gurevich, V. L., Parshin, D. L., Ramos, M. A., Schober, H. R. (1992), *Phys. Rev.* **B46,** 2798–2808.

Buchenau, U., Prager, M., Nücker, N., Dianoux, A. J., Ahmad, N., Phillips, W. A. (1986), *Phys. Rev.* **B34,** 5665–5673.

Cahill, D. G., Katiyar, M., Abelson, J. R. (1994), *Phys. Rev.* **B50,** 6077–6081.

Cody, G. D. (1984), in: J. Pankove (Ed.), *Semiconductors and Semimetals,* New York: Academic, Vol. 21B, p. 11.

Cody, G. D., Tiedje, T., Abeles, B., Brooks, B., Goldstein, Y. (1981), *Phys. Rev. Lett.* **47,** 1480–1483.

Connell, G. A. N. (1979), in: M. H. Brodsky (Ed.), *Amorphous Semiconductors,* New York: Springer-Verlag, Vol. 36, p. 73.

Courtens, E., Vacher, R., Pelous, J., Woignier, T. (1988), *Europhys. Lett.* **6,** 245–250.

Cullis, A. G., Webber, H. C., Chew, N. G., Poate, J. M., Baeri, P. (1982), *Phys. Rev. Lett.* **49,** 219–222.

Demishev, S. V., Kosichkin, Yu. V., Sluchanko, N. E., Lyapin, A. G. (1994), *Usp. Fiz. Nauk.* **164,** 195–230 (*Sov. Phys. Usp.* **36,** 185–217).

Elliott, S. R., Kolobov, A. V. (1991), *J. Non-Cryst. Sol.* **128,** 216–220.

Elliott, S. R. (1992), *Europhys. Lett.* **19,** 201–206.

Elliott, S. R. (1990), *Physics of Amorphous Materials,* 2nd ed., New York: Longman.

Freeman, J. J., Anderson, A. C. (1986), *Phys. Rev.* **B34,** 5684–5690.

Galperin, Yu, M., Karpov, V. G., Kozub, V. I. (1989), *Adv. Phys.* **38,** 669–737.

Graebner, J. E., Allen, L. C. (1983), *Phys. Rev. Lett.* **51,** 1566–1569.

Graebner, J. E., Golding, B., Allen, L. C., Knights, J. C., Biegelsen, D. K. (1984), *Phys. Rev.* **B29,** 3744–3746.

Hecking, N., Heidemann, K. F., Te Kaat, E. (1986), *Nucl. Instrum. Meth. Phys. Res.* **B15,** 760–764.

Houserova, J., Hruby, A. (1972), *Czech J. Phys.* **B22,** 861–863.

Ioffe, A. F., Regel, A. R. (1960), *Prog. Semicond.* **4,** 237–291.

Jackson, W. B., Amer, N. M. (1982), *Phys. Rev.* **B25,** 5559–5562.

Jackson, W. B., Kelso, S. M., Tsai, C. C., Allen, J. W., Oh, S.-J. (1985), *Phys. Rev.* **B31,** 5187–5198.

Kamitakahara, W. A., Shanks, H. R., McClelland, J. F., Buchenau, U., Gompf, F., Pintchovious, L. (1984), *Phys. Rev. Lett.* **52,** 644–647.

Karpov, V. G., Klinger, M. I., Ignat'ev, F. N. (1983), *Zh. Eksp. Teor. Fiz.* **84,** 760–765 (*Sov. Phys. JETP* **57,** 439–448).

Kittel, C. (1976), *Introduction to Solid State Physics,* 5th ed., New York: Wiley.

Kolobov, A. V., Elliott, S. R. (1995), *Philos. Mag.* **B71,** 1–10.

Leadbetter, A. J. (1973), in: B. T. M. Willis (Ed.), *Chemical Applications of Thermal Neutron Scattering,* Oxford: Oxford Univ. Press, p. 146.

Ley, L. (1984), in: J. D. Joannopoulos and G. Lucovsky (Eds.), *The Physics of Hydrogenated Amorphous Silicon,* Vol. II, New York: Springer Verlag, p. 61.

Ley, L., Pollak, R. A., Kowalczyk, S., McFeely, R., Shirley, D. A. (1973), *Phys. Rev.* **B8,** 641–646.

Liu, P. L., Yen, R., Bloembergen, N., Hodgson, R. T. (1979), *Appl. Phys. Lett.* **34,** 864–866.

Lucovsky, G., Tsu, D. V. (1987), *J. Non-Cryst. Sol.* **97–98,** 265–268.

Mahan, A. H., Carapella, J., Nelson, B. P., Crandall, R. S., Balberg, I. (1991), *J. Appl. Phys.* **69,** 6728–6730.

Matsumura, H. (1989), *J. Appl. Phys.* **65,** 4396–4402.

Mott, N. F., Davis, E. A. (1979), *Electronic Processes in Non-Crystalline Solids,* 2nd ed., Oxford: Clarendon Press.

Orbach, R. (1993), *J. Non-Cryst. Sol.* **164–166,** 917–922.

Orbach, R. (1992), *Philos. Mag.* **B65,** 289–301.

Orbach, R. (1986), *Science* **231,** 814–819.

Phillips, J. C. (1981a), *J. Non-Cryst. Sol.* **43,** 37–77.

Phillips, W. A. (Ed.) (1981b), *Amorphous Solids: Low-Temperature Properties,* New York: Springer-Verlag.

Phillips, W. A. (1972), *J. Low Temp. Phys.* **7,** 351–360.

Pierce, D. T., Spicer, W. E. (1972), *Phys. Rev.* **B5,** 3017–3029.

Roca i Cabarrocas, P. (1993), *J. Non-Cryst. Sol.* **164–166,** 37–42.

Scholten, A. J., Akimov, A. V., Verleg, P. A. W. E., Dijkhuis, J. I., Meltzer, R. S. (1993), *J. Non-Cryst. Sol.* **164–166,** 923–926.

Shevchik, N. J., Cardona, M., Tejeda, J. (1973), *Phys. Rev.* **B8,** 2833–2841.

Smirl, A. L., Boyd, I. W., Boggers, T. F., Moss, S. C., van Driel, H. M. (1986), *J. Appl. Phys.* **60,** 1169–1182.

Spear, W. E., Le Comber, P. G. (1975), *Solid State Comm.* **17,** 1193–1196.

Stephens, R. B. (1976), *Phys. Rev.* **B13,** 852–865.

Street, R. A. (1991), *Hydrogenated Amorphous Silicon,* Cambridge, U.K.: Cambridge Univ. Press.

Tauc, J., Grigorovici, R., Vancu, A. (1966), *Phys. Status Solidi* **15,** 627–637.

Thorpe, M. F. (1995), *J. Non-Cryst. Sol.* **182,** 135–142.

Tiedje, T. (1984), in: J. Pankove (Ed.), *Semiconductors and Semimetals,* New York: Academic, Vol. 21C, p. 207.

Tsai, C. C., Nemanich, R. J. (1980), *J. Non-Cryst. Sol.* **35–36,** 1203–1208.

Tsu, R., Hodgson, R. T., Tan, T. Y., Baglin, J. E. (1979), *Phys. Rev. Lett.* **42,** 1356–1358.

Urbach, F. (1953), *Phys. Rev.* **92,** 1324.

Yoshida, A., Inoue, K., Ohashi, H., Saito, Y. (1990), *Appl. Phys. Lett.* **57,** 484–486.

Zeller, R. C., Pohl, R. O. (1971), *Phys. Rev.* **B4,** 2029–2041.

Zhang, M., Boolchand, P. (1994), *Science* **266,** 1355–1357.

SEMICONDUCTORS, COMPOUND—ELECTRONIC PROPERTIES

SANDIP TIWARI, *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, New York, U.S.A.*

	Introduction	303
1.	**Band Structures**	304
1.1	Mathematical Formulations	304
1.2	Conduction-Band Structure	304
1.3	Valence-Band Structure	307
1.4	Band Gaps	307
2.	**Heterojunctions and Discontinuities**	308
2.1	Unstrained Systems	308
2.2	Strained Systems	311
3.	**Scattering and Transport**	311
3.1	Phonon Dispersion and Scattering	311
3.2	Majority-Carrier Transport	315
3.3	High-Field Transport of Electrons	320
4.	**Minority-Carrier Transport Properties**	321
4.1	Diffusion Coefficients	323
5.	**Confinement Effects**	324
5.1	Transport Consequences of Confinement	324
6.	**Hot-Carrier Properties**	325
6.1	Avalanche and Other Generation Effects	326
6.2	Off-Equilibrium Transport	328
7.	**Recombination**	329
7.1	Defects and Hall–Shockley–Read Recombination	329
7.2	Auger Recombination	330
8.	**Tunneling**	331
9.	**Heavy Doping Effects**	332
10.	**Piezoelectric Effects**	332
	Glossary	333
	Work Cited	334
	Further Reading	334

INTRODUCTION

The term compound semiconductor usually refers to the semiconducting materials formed by compounds of group IIIB elements (Al, Ga, In, e.g.) and group VB elements (N, P, As, Sb, e.g.), group IIB elements (Zn, Cd, Hg, e.g.) and group VIB elements (S, Se, Te, e.g.), and combinations of them. GaAs, $Ga_{1-x}Al_xAs$, $Ga_{1-x}In_xAs$, $Ga_xIn_{1-x}As_yP_{1-y}$, $Al_{1-x}In_xAs$, InP, $Ga_{1-x}In_xP$, GaN, ZnSe, and $Cd_xZn_{1-x}Se$ are some of the more popular examples—binary, ternary, and quaternary—and appealing in device structures because of their electronic or optical properties. The electronic properties that attract the interest are usually the superior transport of carriers—faster than in silicon, the existence of semi-insulating substrates and large breakdown voltages in some of the larger-band-gap semiconductors—each of which is particularly useful in high-frequency microwave applications or even high-speed applications. Many compound semiconductors have direct energy band gaps allowing efficient radiative recombination. Lattice-matched growth allows stacking of different compound semiconductors or grading of their compositions with changes in band gaps, refractive indexes, etc., while maintaining crystalline and interfacial quality. These allow for confinement of carriers and/or photons (wave guiding) to be achieved, resulting in compact lasers as well as very high-mobility structures where donors can be separated from carriers and ionized-impurity scattering minimized. Compound semiconductors have thus been attractive for both engineering and scientific reasons. The group II–VI compound semiconductors have poor mobilities and hence have not been applied to electronic devices in any significant way. They have been employed in optical devices—particularly light-emitting diodes and lasers in the short visible wavelength region. This article will

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

therefore concentrate on the III–V compound semiconductors, and Table 1 is a compilation of properties that we refer to throughout.

1. BAND STRUCTURES

1.1 Mathematical Formulations

Describing the behavior of electrons and holes—the energy-momenta relationship—in a solid such as a compound semiconductor crystal is inherently complex, a many-body problem. The lattice potential is large and thus cannot be treated as a perturbation, and the electrons can best be described by plane waves with lattice periodicity (Bloch functions) whose modulation is very strong (like free atoms) near the ions and very weak (like free electrons) in between. An approximate description of the behavior of electrons in this periodic potential requires a solution to the Schrödinger equation using the crystal Hamiltonian. This crystal Hamiltonian contains energy terms due to kinetic, potential, and magnetic contributions arising from electrons in a lattice structure of the nuclei. The solution can be accomplished in a number of ways using perturbation theory: using tight-binding approximation (i.e., using functions that are linear combinations of either atomic or molecular orbitals), using pseudopotentials, and using various other functional approaches such as Green's-function techniques. Perturbation theory allows the determination of the eigenenergies using the smaller Hamiltonian terms as perturbing functions. Pseudopotentials allow the most convenient means to solve the problem at the present time. The influence of each ion and the environment on electrons in the conduction and the valence bands is incorporated through a potential that approximates realistically, although empirically through experimental matching of band-structure results, the influence of the ion in the important region away from the core while taking into account the crystal symmetry in the form of a Fourier series over the reciprocal lattice vector and the nonlocal effects such as due to different angular momentum components of electrons (for $|s\rangle$ and $|p\rangle$, e.g.) through variations in the pseudopotential. Since the basis functions, involving calculations based on using atomic orbitals near the cores, are $|s\rangle$ and $|p\rangle$ like, the resulting functions are combinations of these at various important points of the Brillouin zone. The band structure is the interesting result of the calculation that describes the relationship between the energy and the wave vector of electrons—the specific values that they can have. There are bands in energy that are allowed and bands that are not allowed. Semiconductors in their pure state are compounds that have these allowed bands either completely filled or completely empty at absolute zero. At a finite temperature, thermal energy allows carriers to be transferred from the fully occupied band to the empty band because the gap between the two, the band gap, is small enough (on the order of electron volts). This makes them semiconducting and allows for carrier transport to take place in both the partially occupied band (electrons in the conduction band) and the partially empty band (holes representing the absence of electrons in the valence band).

1.2 Conduction-Band Structure

Figure 1 shows the conduction- and valence-band structure for GaAs over a large energy range, and for comparison and discussion we include the conduction-band structure of three other popular compound semiconductors: InP (Fig. 2), InAs (Fig. 3), and AlAs (Fig. 4). GaAs is perhaps the most common compound semiconductor in use as a substrate for field-effect transistors and utilized together with AlAs, i.e., in the ternary form $Ga_{1-x}Al_xAs$, in formation of heterojunction discontinuities that have wide applications. Alloy compositions of GaAs with InAs, $Ga_{1-x}In_xAs$, have also been quite common and utilize discontinuities with $Al_{1-x}In_xAs$, InP, and in strained structures with GaAs and its ternary compounds. At the composition of $Ga_{0.47}In_{0.53}As$, it is lattice matched to InP and grown on it. The band structures of these materials also point out a number of interesting features of the variety of compound semiconductors.

Some general observations: GaAs is a direct–band-gap semiconductor with the ordering of minima from low to high energies in the conduction band as Γ ($\langle 000\rangle$), L ($\langle 111\rangle$ and its equivalents), and X ($\langle 100\rangle$ and its equivalents). The lowest constant-energy sur-

Table 1. A compilation of some of the major parameters of compound semiconductors.

Property	AlN	AlP	AlAs	AlSb	GaN	GaP	GaAs	GaSb	InN	InP	InAs	InSb
Lattice[a]	ZB	ZB	ZB	ZB	WZ	ZB	ZB	ZB	WZ	ZB	ZB	ZB
a, Å	*a* axis 3.112 *c* axis 4.980	5.467	5.661	6.136	*a* axis 3.160 *c* axis 5.125	5.451	5.653	6.096	*a* axis 3.544 *c* axis 5.703	5.869	6.058	4.479
ρ, g/cm^3		2.40	3.76	4.26		4.138	5.36	5.614		4.81	5.667	5.775
ϵ_0	9.1	9.8	10.06	12.04	10.4_c	11.1	12.91	15.69		12.61	15.15	16.8
ϵ_∞	4.8	7.54	8.16	10.24	5.8_c	9.075	10.1	14.44	9.3	9.61	12.25	15.68
E_g, eV, 300 K	6.2	2.45	2.14	1.63	3.44	2.268	1.423	0.70	2.09	1.3511	0.356	0.18
ΔE_Γ, eV	0	1.237	0.767	0.507		0.496	0	0		0	0	0
ΔE_L, eV		0.824	0.322	0.292		0.415	0.323	0.217		0.832	1.078	1.038
ΔE_X, eV		0	0	0		0	0.447	0.491		1.492	1.607	1.569
E_g, eV, 77 K			2.223	1.666		2.338	1.512		2.21	1.4135	0.414	0.228
$\hbar\omega_{LO}$, meV	112.8	62.1	50.1	42.1	92.4	50.0	35.4		86.1	42.8	29.6	23.7
$\hbar\omega_{TO}$, meV	82.7	54.5	44.8	39.5	69.4	45.5	33.2		59.3	37.7	26.9	22.3
n_i, cm^{-3}				5.0×10^6		3.0×10^6	1.8×10^6	4.3×10^{12}		1.2×10^8	1.3×10^{15}	2.0×10^{16}
m_Γ		0.166	0.149	0.126		0.122	0.067	0.048		0.083	0.031	0.018
m_{Ll}		1.556	1.456	1.601		1.498	1.538	1.400		1.893	1.590	1.700
m_{Lt}		0.162	0.165	0.142		0.358	0.126	0.139		0.425	0.326	0.096
m_{Xl}		1.118	1.664	2.350		1.868	1.987	1.630		0.927	2.499	2.000
m_{Xt}		0.256	0.223	0.255		0.249	0.229	0.212		0.277	0.274	0.268
E_{so}, eV		0	0.29	0.645		0.08	0.340	0.750		0.110	0.410	0.810
m_{hh}		0.630	0.760	0.940		0.790	0.510	0.290		0.850	0.600	0.440
m_{lh}		0.200	0.150	0.290		0.140	0.082	0.047		0.089	0.027	0.016
μ_{el}, cm^2/V·s, 300 K				200		200	8×10^3	5×10^3		5×10^3	3×10^4	
($<10^{15}$ cm^{-3}), 77 K						2.8×10^3	1.5×10^5			6×10^4	3×10^5	4×10^5
μ_{el}, cm^2/V·s, 300 K		50		250	600	160	4.8×10^3			3.2×10^3	2×10^4	8×10^4
($\approx 10^{17}$ cm^{-3}), 77 K		30		800		1.2×10^3	5.2×10^3			4×10^3	3.5×10^4	1.5×10^5
μ_{el}, cm^2/V·s, 300 K			50		180	95	2.2×10^3			1.8×10^3	1.3×10^4	1.0×10^4
($\approx 10^{19}$ cm^{-3}), 77 K			18			450	3×10^3				2×10^4	8×10^4
μ_h, cm^2/V·s, 300 K				450		150	400	1500		180	480	1500
($<10^{15}$ cm^{-3}), 77 K							9×10^3			1×10^3		9×10^3
μ_h, cm^2/V·s, 300 K				375		120	320	880		150		
($\approx 10^{17}$ cm^{-3}), 77 K				1.8×10^3				2.4×10^3		600		3×10^3
μ_h, cm^2/V·s, 300 K				175		30	50	300		28		
($\approx 10^{19}$ cm^{-3}), 77 K								1×10^3		18		700

---

[a] ZB: zincblende; WZ: wurtzite.

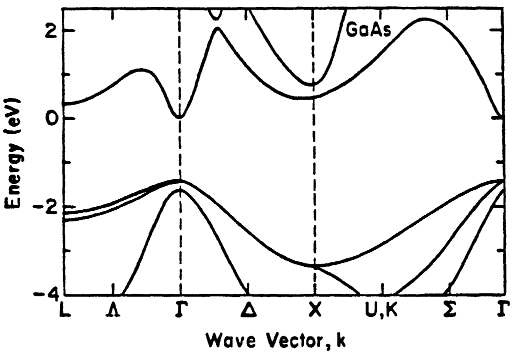

FIG. 1. Conduction- and valence-band structure of GaAs over major points and directions in the first Brillouin zone.

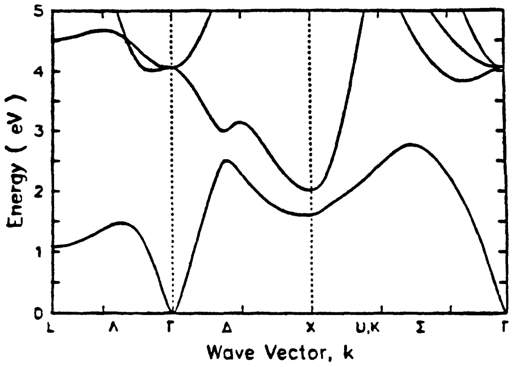

FIG. 3. Conduction-band structure of InAs over major points and directions in the first Brillouin zone.

face, due to the *Γ* valley, is spherically symmetric. The band gap of intrinsic GaAs, at 300 K, is 1.42 eV; the *L* valley is 0.29 eV, and the *X* valley is about 0.48 eV, above the *Γ* minimum. InP and InAs are also direct–band-gap compound semiconductors with an ordering of conduction-band minima similar to that of GaAs. The band gap of intrinsic InP is 1.35 eV, and that of InAs is significantly smaller, ≈0.36 eV, at 300 K. AlAs and GaP are two of the few compound semiconductors with indirect energy-band gap.

AlAs is an indirect–band-gap semiconductor with an intrinsic band gap of 2.14 eV; the minimum in conduction-band energy occurs nearly 85% of the way to the *X* point from the *Γ* point (similar to silicon). This is followed by the *L* valley. It exhibits a behavior similar to that of silicon, and some of these characteristics may be traced to aluminum, which is a light atom like silicon. Constant low-energy surfaces around this valley, loosely referred to as the *X* valley, are sixfold degenerate and aligned along orthogonal axes.

In indirect semiconductors, such as AlAs (silicon and germanium are the two common elemental semiconductor examples), the lowest conduction bands have multiple valleys associated with them. A consequence of this is the availability of a large number of states to scatter into. A large change in momentum is required for this process and is available through zone-edge phonons. The result of this is that, at low energies, these semiconductors experience higher scattering than direct–band-gap compound semiconductors.

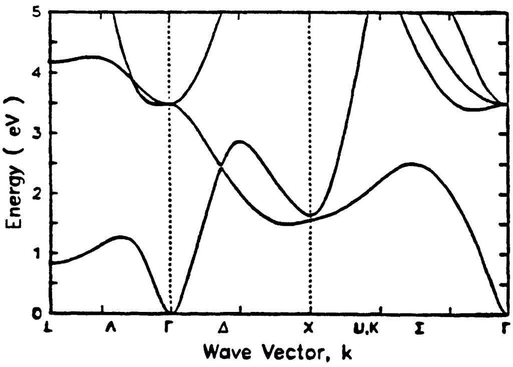

FIG. 2. Conduction-band structure of InP over major points and directions in the first Brillouin zone.

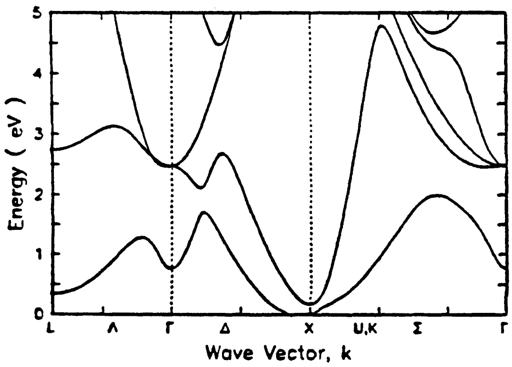

FIG. 4. Conduction-band structure of AlAs over major points and directions in the first Brillouin zone.

1.3 Valence-Band Structure

All compound semiconductors exhibit a universal feature of a degenerate light- and heavy-hole band at zone center, and a split-off band. The split-off band comes about because of the perturbation from the magnetic energy associated with the spin and orbit of the electron. This coupling is generally referred to as spin–orbit coupling or $L \cdot S$ coupling, and the split-off band arising from it is 0.34 eV below the valence-band maximum in GaAs. Silicon, an elemental semiconductor, does not exhibit a large separation of the split-off band. The origin of this is in the spin–orbit energy that gives rise to the formation of the split-off band: Light atoms exhibit small spin–orbit interaction. In silicon, spin–orbit splitting is only 0.044 eV. Spin–orbit energy, however, rises rapidly with atomic mass. Germanium has a spin–orbit splitting of 0.29 eV and does exhibit the split-off band. In InP a split-off band occurs 0.78 eV below. A consequence of multihole bands, split-off bands, large effective masses, etc., is a large hole-scattering rate. This together with the large mass results in poor hole mobilities. Only GaSb, of all the compound semiconductors, exhibits any reasonable hole mobility although still not comparable to the electron mobility.

1.4 Band Gaps

One of the unique attributes of compound semiconductors, exploited in many specialized device structures, is the rich selection of various semiconductors that form heterojunctions of the same lattice constant yet differing in band gap, thus allowing for a broader selection and tuning of desired characteristics. Figure 5 shows the variation of band gap as function of lattice constant for various ternary semiconductors with their binary end points. A broad range of band gaps is possible; for optical devices it is useful in achieving lasing at a variety of useful wavelengths—from infrared to blue at the present time. In electronic devices, its major significance has been in growing single-crystal materials of differing band gap and achieving either discontinuities in band edges at the interfaces or a grading of band gap in the interface region. Discontinuities allow for unique transport and charge control properties that are useful in field-effect transistors. As an example, $Ga_{1-x}Al_xAs$/GaAs and $Al_{1-x}In_xAs/Ga_{1-x}In_xAs$ allow large electron densities in accumulation/inversion layers at the interfaces in the smaller–band-gap material while the doping is located in the larger–band-gap material. This leads to superior mobilities for carriers primarily because of a reduction in ionized-impurity scattering. Grading of band gaps allows for placement of the change in band gap in the conduction band edge or valence band edge. An *n-p-n* $Ga_{1-x}Al_xAs$/GaAs/GaAs bipolar junction transistor with a graded interface at the emitter–base junction allows for different barriers for electrons and holes, a smaller barrier for electrons than for holes leading to superior injection efficiencies.

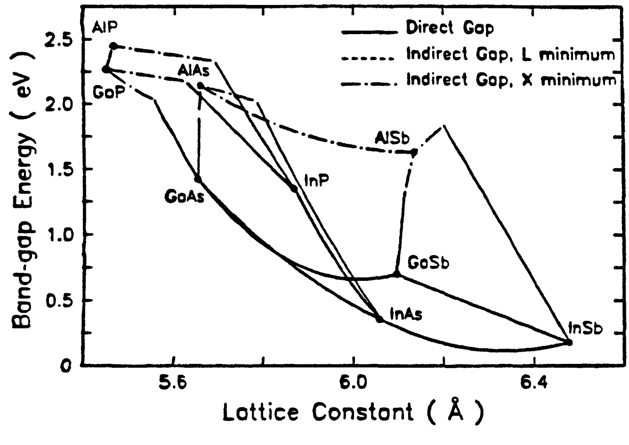

FIG. 5. Band gap versus lattice constant for compound semiconductors. Legend in the figure describes the key to the lowest valley in the various combinations of materials.

2. HETEROJUNCTIONS AND DISCONTINUITIES

2.1 Unstrained Systems

Figure 6 summarizes very approximately the conduction- and valence-band discontinuity of compound semiconductors as a function of the lattice constant, assuming that a combination can be grown in good crystalline quality and that the members are lattice matched, or slightly mismatched to allow pseudomorphic* growth and, hence, allow the fabrication of structures where band-gap changes take place. Figure 6 is an approximate description, based on observed band-gap discontinuities, of the expected conduction- and valence-band discontinuities if one were to make an ideal interface between two semiconductors shown in the figure. Tables 2 and 3 show the discontinuity reference values and bowing and other parameters relevant to the figures of band-gap variation and band-discontinuity variation. The connecting lines in Fig. 6 have been drawn to follow the behavior for the conduction-band edge. The figure is approximate and should be used with caution.

In heterojunctions, the semiconductor band structure may be considered undisturbed right up to the interface. At the interface, in the abrupt heterojunctions, discontinuities occur in the conduction band (ΔE_c) and valence band (ΔE_v), whose magnitude is experimentally determined. Many *ab initio* models and calculations of band discontinuities have been presented, some of which are in fair agreement with experiments. Histori-

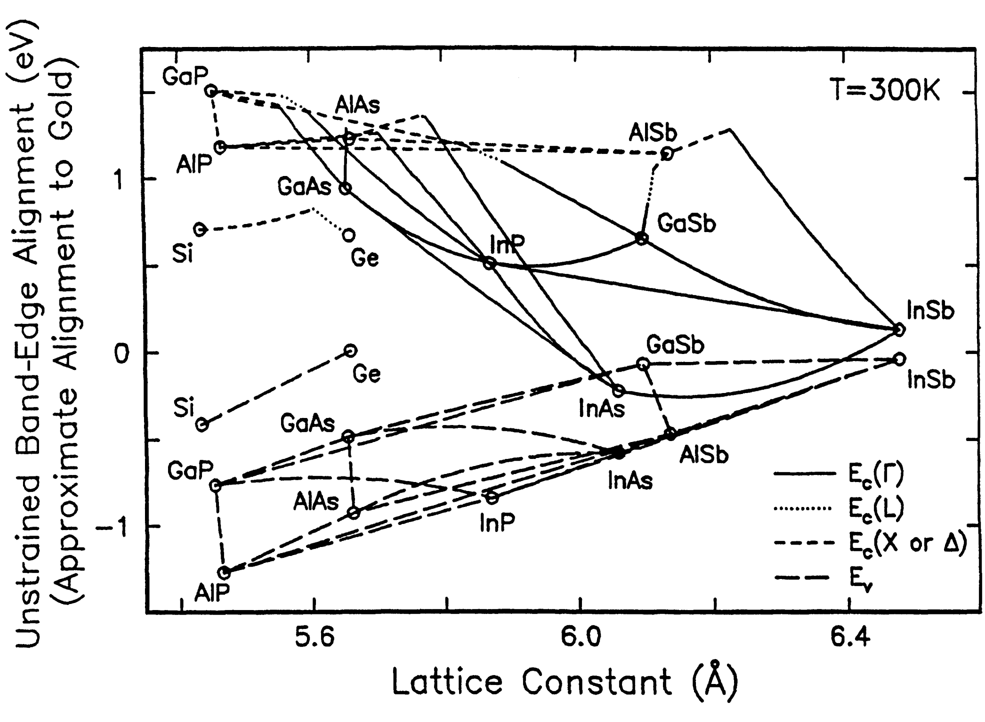

FIG. 6. Discontinuities at heterostructure interfaces for various compound semiconductor systems plotted as a function of lattice constant. This figure is approximate and should be used with great caution. It has been assembled using available and reliable discontinuity data, most of which is for unstrained interfaces. The strained data have also been incorporated approximately, and hence the behavior described by this figure should be considered an approximate description of what one might expect if one were to make an ideal interface of two sets of materials.

---

*A is a *pseudomorph* of B if it possesses the external form characteristic of B. When a semiconductor is grown on another semiconductor, the grown semiconductor, if it is crystalline, takes the in-plane lattice periodicity of the basis semiconductor. It is said to be pseudomorphic.

Table 2. Discontinuity references at 300 K.

Lattice constant of group (Å)	Material system	ΔE_g (eV)	ΔE_c (eV)	ΔE_v (eV)
5.44	GaP/Si	1.148	0.798	0.35
5.66	GaAs/Ge	0.758	0.268	0.49
	AlAs/Ge	1.48	0.55	0.93
	$Ga_{0.7}Al_{0.3}As/GaAs$	0.395	0.263	0.132
	AlAs/GaAs	0.73	0.29	0.44
	$Ga_{0.51}In_{0.49}P/GaAs$	0.463	0.223	0.24
5.87	$Al_{0.48}In_{0.52}As/InP$	0.106	0.30	−0.194
	$InP/Ga_{0.47}In_{0.53}As$	0.600	0.20	0.40
	$Al_{0.48}In_{0.52}As/Ga_{0.47}In_{0.53}As$	0.706	0.50	0.206
6.10	AlSb/GaSb	0.887	0.487	0.40
	GaSb/InAs	0.370	0.88	−0.51
	AlSb/InAs	1.257	1.367	−0.11

cally, the first of the models was Anderson's electron affinity rule, which assumed continuity of the vacuum level. This predicts $\Delta E_c = \chi_2 - \chi_1$, where χ's are the electron affinities in the two semiconductors. This is in error for most compound semiconductors. For the SiO_2/Si interface, it is relatively close, but does not predict the observed orientation dependence. Atomic-orbital–based calculations predict that when the anion species are common at heterojunctions, such as at $Ga_{1-x}Al_xAs/GaAs$, most of the discontinuity should occur in the conduction band. The argument is that valence-band states are *p*-orbital like and come from the anion that is common. In practice conduction-band discontinuity does not always dominate, showing that significant mixing takes place. For $Ga_{1-x}Al_xAs/GaAs$, the conduction-band discontinuity is $\approx 0.65\Delta E_g$ for aluminum mole fractions of less than 0.4. Semiempirical techniques are therefore usually employed.

Table 3. Bowing parameters for band edges at 300 K.

Material system	c_{Γ_v} (eV)	$c_{\Gamma_c-\Gamma_v}$ (eV)	$c_{L-\Gamma_v}$ (eV)	$c_{X-\Gamma_v}$ (eV)
$Ga_{1-x}Al_xP$	—	—	—	—
$In_{1-x}Al_xP$	—	—	—	—
$In_{1-x}Ga_xP$	0.098	0.77	1.15	0.207
$Ga_{1-x}Al_xAs$	0	0.37	0.055	0.245
$In_{1-x}Al_xAs$	0.479	0.675	—	—
$In_{1-x}Ga_xAs$	0.460	0.431	0.5	0.08
$Ga_{1-x}Al_xSb$	—	0.47	0.7	—
$In_{1-x}Al_xSb$	—	0.43	—	—
$In_{1-x}Ga_xSb$	—	0.413	0.33	0.13
$AlAs_{1-x}P_x$	—	—	—	—
$AlSb_{1-x}As_x$	—	—	—	—
$AlSb_{1-x}P_x$	—	—	—	—
$GaAs_{1-x}P_x$	—	0.19	0.16	0.21
$GaSb_{1-x}As_x$	—	1.15	—	—
$GaSb_{1-x}P_x$	—	—	—	—
$InAs_{1-x}P_x$	—	0.32	—	—
$InSb_{1-x}As_x$	—	0.53	—	—
$InSb_{1-x}P_x$	—	—	—	—
$Ge_{1-x}Si_x$	—	—	0.33	0.17

The discontinuity is, of course, with respect to a specific band minimum. Different minima have different discontinuities that follow different trends, represented, e.g., in the bowing parameters of the alloy. At different specific compositions, different minima may be the lowest. At a heterojunction, which minimum should be considered to be important, and hence which discontinuity should be used, depends on the dominant physical process. All the discontinuities *Γ-Γ*, *L-L*, *X-X*, etc., exist at the heterojunction interface. As an example, consider tunneling through a barrier of large–band-gap material such as $Ga_{1-x}Al_xAs$ from GaAs. Electrons in GaAs are largely in the *Γ* valley; if the $Ga_{1-x}Al_xAs$ barrier is relatively thin, these *Γ* electrons are most likely to tunnel via *Γ*-like states in the barrier. However, if the barrier is thick, they may do so through *X*-like states, a process that will require phonon interaction in order to allow for change in momentum. Thus, for thick barriers, the difference in energy between the *Γ* valley of GaAs and the *X* valley of $Ga_{1-x}Al_xAs$ may be more pertinent. Consider the same for transfer of carriers from donors in $Ga_{1-x}Al_xAs$ to GaAs. The transfer of carriers takes place from whichever is the lowest valley in $Ga_{1-x}Al_xAs$, irrespective of the width of the barrier. This interesting behavior of the discontinuity (Fig. 7) has implications for the charge-transfer process and for perpendicular current flow.

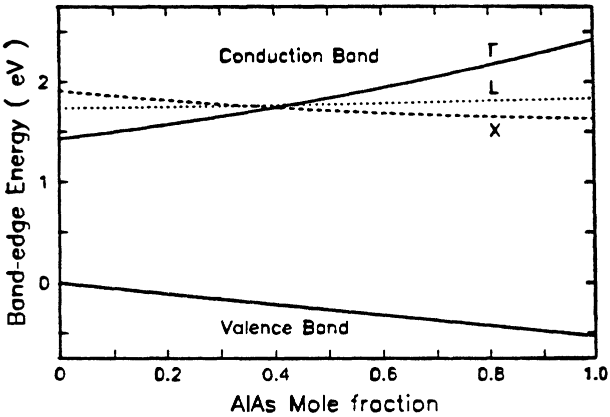

FIG. 7. The conduction- and valence-band discontinuities in the (Ga,Al)As/GaAs system as functions of aluminum mole fraction for the Γ and X valleys.

At aluminum mole fractions of greater than 0.45, the X band is the minimum and forms the lower barrier. So, for $x > 0.45$, the barrier between the X valley in $Ga_{1-x}Al_xAs$ and the Γ valley in GaAs is lower, and in fact for charge transfer, and for tunneling in larger barrier widths, the relevant maximum discontinuity occurs at $x \approx 0.45$. For field-effect transistors based on heterostructures, this results in a larger leakage current between the gate and the channel as mole fractions go toward the AlAs end of the semiconductor composition, counterintuitive to the simplistically expected trend. Suffice it to say, the barrier thickness and barrier height for each of the valleys need to be considered together with elastic and nonelastic tunneling processes to determine the dominant barrier in the current-transport mechanism; for charge-transfer calculations, these are independent of the geometric characteristics of the barrier.

$Ga_{1-x}Al_xAs$ remains closely lattice matched at all mole fractions of aluminum and gallium arsenide (lattice mismatch remains less than 0.1% to both GaAs and AlAs). $Ga_{1-x}In_xP$, GaP_xSb_{1-x}, etc., are lattice matched only at specific mole fractions. For limited thickness of the mismatched layers, however, the resulting strain can be accommodated. Thus, materials with close match only at specific mole fractions can still be grown with a fair degree of reproducibility. Examples of these, of interest, are $Ga_{0.47}In_{0.53}As$ and $Al_{0.48}In_{0.52}As$ to InP, and $Ga_{0.51}In_{0.49}P$ to GaAs. A consequence of this is a tunability of characteristics of the material to achieve desired objectives. A differing barrier height than that of metal-semiconductor junction can be achieved—small heights for microwave detection and large heights for suppressing gate-leakage current in field-effect transistors. The changes in band gap can be used in suppressing injection characteristics of one particular carrier through a suitable use of grading and doping as in heterostructure bipolar transistors. A major usage of the changes in band gap at a spatially abrupt interface has been in obtaining two-dimensional channels of carriers for operation in field-effect transistors.

Practical limitations quite often arise from our ability to grow lattice-mismatched semiconductors. Lattice strain prevents growth of semiconductors with large mismatch because of energy considerations. Thermal energy provides the requisite barrier energy for this relaxation process, and the lattice regains its unstrained lattice constant. A lattice with high strain relaxes by generating defects such as dislocations, etc. A low temperature of growth allows one to make structures that are metastable and have a significant strain due to large lattice mismatch, but they may relax by radiative or nonradiative means during operation of a device. In any case, there is a critical thickness beyond which, for given growth conditions, one cannot grow a suitably high-quality semiconductor. The thicknesses required for heterostructure field-effect transistors are usually sufficiently small that quite significant mismatch can be accommodated. An example of this is the considerable work in the $Ga_{1-x}Al_xAs/Ga_{1-x}In_xAs$ system where low mole fractions of indium arsenide have been used to obtain improvement in amount of charge and transport. The effect of strain in such structures is both in the band structure itself and in scattering due to changes in band structure.

2.2 Strained Systems

During the growth of a structure, the grown epitaxial layer assumes the lattice constant of the crystal and is under strain of $\epsilon = \Delta a/a$, where a is the lattice constant of the unstrained crystal. This strain has an effect on the band structure because it introduces a spatially periodic perturbation differing from that of the unstrained crystal. The strain introduced due to the growth, caused by lattice in the plane perpendicular to the growth orientation, is a biaxial strain.

The symmetry of the band structure has an important relationship with the polarity of the effect on band gap. For direct materials, such as $Ga_{1-x}In_xAs$ grown on GaAs, this biaxially compressive strain leads to a decrease in the band gap. The effect of compressive strain in an indirect-gap crystal—e.g., $Al_xIn_{1-x}P$/AlP—is to increase the band gap from the unstrained situation. So, the polarity of this is dependent on the symmetry of the band-gap minima.

These changes in band gap occur together with changes in the band curvature at the band minima of the conduction band and band maxima of the valence bands, with the changes in the latter being particularly significant. The constant-energy surface for the valence bands is a warped surface. It exhibits differing curvature in different directions. Also, at these energies, we need to consider both the light-hole and the heavy-hole bands. The primary effect of the strain in these is to lift the degeneracy. For the example of $Ga_{1-x}In_xAs$ grown on GaAs, it causes, at the zone center, the heavy-hole band to become the lower-hole energy band, and the light-hole band is pushed significantly farther away. The heavy-hole band also changes considerably in curvature at the zone center, resulting in a lower effective mass for holes at the zone center. This effective-mass change is orientation dependent because the constant-energy surface is warped. Significant complications have now been introduced in the description of the electronic structure by introduction of this strain. Since the masses are different in different orientations and degeneracy has been lifted at the zone center, we would expect interesting consequences for hole transport due to the effect on mass and scattering and for electron–hole interactions, such as in optical processes.

For transport, the primary effect is the lifting of the light-hole and heavy-hole degeneracy, which reduces scattering and hence improves the hole-transport properties. The effect on distortion of the conduction band is smaller than that for hole bands in these examples. The in-plane distortion of the crystal due to the biaxial strain is fourfold symmetric. Band-edge symmetries can therefore be destroyed by this strain. For the examples considered, for electrons, the out-of-plane transport of electrons improves, but the in-plane transport deteriorates.

3. SCATTERING AND TRANSPORT

3.1 Phonon Dispersion and Scattering

At finite temperatures, atoms vibrate around their mean position. The periodic vibration leads to perturbation of the stationary periodic potential and causes changes in the electron wave function in time. The lattice vibrations can be described in a framework that is analogous to the description of electrons. Phonons that represent these vibrations occurring at a frequency ω_q have an energy that is $\hbar\omega_q$, where $\hbar$ is the reduced Planck's constant, and have a momentum that is $\hbar\mathbf{q}$. Phonons describe vibrations just as photons describe electromagnetic waves.

In a three-dimensional crystal, movement of atoms about their mean position may be described by three independent modes of oscillations: one unique longitudinal mode along the wave vector of the oscillation and any two orthogonal directions for transverse oscillations. In the primitive cell, if there are N different types of atoms either of differing mass or ordering in space, $3N$ vibration modes will result. In general, three of these branches, the acoustic branches, will disappear at zone center. The remaining $3N - 3$ branches will be optical branches. For cubic crystals—GaAs, e.g.—$N = 2$, and hence there are three acoustic (low-energy) branches and three optical (high-energy) branches, each

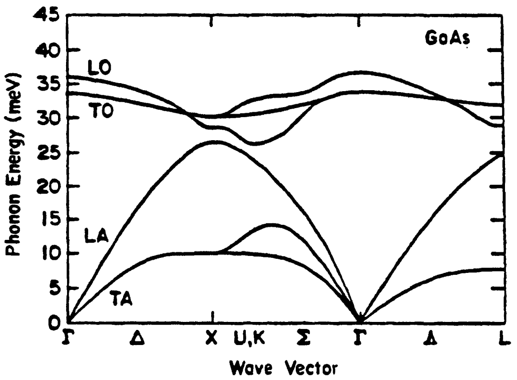

FIG. 8. Phonon dispersion curves for GaAs along major directions in the Brillouin zone.

three comprising one longitudinal and two transverse modes. Scattering resulting from optical and acoustic phonons is commonly referred to as optical-phonon scattering and acoustic-phonon scattering, respectively. Since the energy of the acoustic-phonon mode is smaller, it produces less perturbation than the optical phonons. However, for ⟨100⟩, ⟨110⟩, or ⟨111⟩ directions, the two transverse directions are indistinguishable. So, in the cubic crystal, the two transverse oscillations are degenerate.

Examples of phonon dispersion curves, which describe the energy or frequency dependence as a function of momentum, are shown in Figs. 8 and 9 for GaAs and InAs. The order of magnitude of the optical-phonon energy is between 25 and 50 meV for most semiconductors of interest, on the same order of magnitude as the thermal energy. Heavier ions should be more sluggish; the acoustic branch then should have a lower peak energy. This is reflected in the ascending ordering of peak acoustic energies of InAs, Ge, GaAs, and Si in inverse proportion to the weight of the heaviest element. A similar trend holds for the lowest optical-branch energy for these materials. An additional interesting feature of these figures is that, for compound semiconductors, the longitudinal optical phonon is slightly more energetic than the transverse optical phonon at low momenta because of the presence of ionic charge. At high momenta, in certain directions, they may even cross over because the higher momenta results in a larger reduction in energy of the longitudinal mode due to the effect of the dipole. These figures also show that at specific large momenta, in directions differing from the major crystal axis directions ⟨100⟩, ⟨110⟩, or ⟨111⟩, the degenerate transverse modes split because of the breakdown of symmetry. This occurs for both optical and acoustic modes.

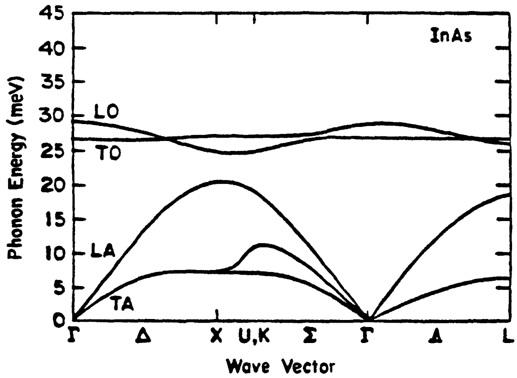

FIG. 9. Phonon dispersion curves for InAs along major directions in the Brillouin zone.

The carrier transport in a medium occurs under the influence of applied forces, which accelerate the carriers in a preferred direction, and carrier interactions with their environment, whose general effect is to render the motion random. Scattering is a general term for the interaction of the carrier with the surroundings.

An electron in the crystal under the influence of an external force, such as due to the electric field, picks up its energy during its acceleration, but it changes energy and/or momentum by various scattering mechanisms that come about because of lattice vibrations, other carriers in its surroundings, and defects of the lattice. The various scattering mechanisms are events of interaction between the carrier and its surroundings that cause it to change its momentum and/or energy. It therefore represents the individual and aggregate effect of the carrier's surroundings that influences the carrier's individual and aggregate transport behavior. If the scattering processes were not to occur, the application of an electric field would increase the velocity linearly with time, an occurrence generally described as ballistic motion. In reality, for long sample lengths, it reaches a limiting value. This limiting value is proportional to the field at low fields (v =

$\mu\mathscr{E}$ where v is the velocity, μ the mobility, and $\mathscr{E}$ the electric field, a constant-mobility region) and saturates to the value v_s at high fields (the saturation-velocity limit). The division of these scattering mechanisms is in three general classes:

1. scatterings that come about from imperfections in the periodicity of the lattice, which we call defect scattering;
2. carrier-related scattering due to interactions between the carriers, a process that becomes important at high carrier concentrations and that can couple with other modes of scattering; and
3. scattering due to interaction of carriers with the perturbing potential produced by lattice vibrations.

Imperfections in the periodic potential of a crystal come about as a result of crystal defects such as vacancies, dislocations, interstitials, etc. These are quite uncommon and the density of them is less than that of unintentional substitutional impurities. Defect scattering also arises from impurity atoms themselves that are used to dope the crystal; this is generally referred to as impurity scattering and originates in the deviation of the local potential around an impurity from the local and periodic potential of the atoms of the host crystal. If the impurity atom is ionized (as is almost always the case), it is called ionized-impurity scattering, and this particular scattering is strong, as shown in Fig. 10 for GaAs. If the impurity atom is not ionized, i.e., it is neutral, it is referred to as neutral-impurity scattering. Neutral-impurity scattering is quite weak because of the small perturbation in potential from a neutral atom on a lattice site. Consider the scattering due to impurity atoms in a crystal as a function of temperature. We assume that impurity density is not so large so as to cause formation of an impurity band and its coalescence with the conduction band. Such a material, therefore, shows freeze-out of carriers; i.e., at low temperatures, the carriers do not have sufficient thermal energy to make the transition to the conduction or valence bands. At temperatures below the freeze-out temperature, the impurities are largely neutral, and hence there is a weak impurity scattering. As the temperature is raised, impurities ionize, and impurity scattering becomes stronger, and hence the carrier mobility decreases. As the temperature is raised further, the thermal energy of the carriers continues to increase, their velocities are larger, and the impurity scattering process becomes weak again. An analogy may be drawn between this and the Rutherford scattering used to analyze the scattering of light atoms by heavier nuclei. A high velocity and energy of the particle causes both a smaller time of interaction and a smaller effect on the velocity of the carrier. Because of the large disparity in masses between the atoms and the moving carriers, this scattering conserves energy and is an example of elastic scattering. At high temperatures, other scattering processes (see Fig. 10) become important, and impurity scattering plays an increasingly smaller role. It is generally important in the 77 K (liquid nitrogen temperature) to 300 K (room temperature) range. The third defect scattering that is im-

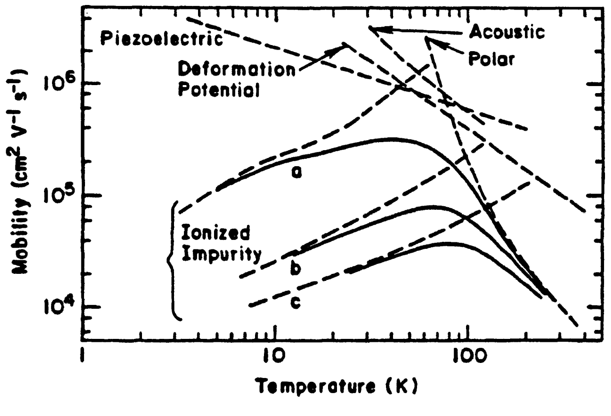

FIG. 10. Electron mobility in *n*-type lightly doped GaAs as a function of temperature together with contributions from the various sources of scattering. After Wolfe *et al.* (1970). The polar scattering referred to in this figure is the component of scattering arising from the perturbation arising from the polar nature of GaAs.

portant occurs in mixed crystals such as $Ga_{1-x}In_xAs$ and $Ga_{1-x}Al_xAs$: alloy scattering. Because of the random variation of the atoms on either the cation or the anion sites of the lattice, a nonperiodic perturbation occurs naturally in such material even when no defect has been added. The perturbations are sufficiently small that alloy scattering generally becomes important only below 300 K.

The last of the defect scattering processes important to devices is related to interactions at the surfaces and interfaces of the lattice (interface scattering). This is a crystal-defect scattering process and is certainly very important at insulator–semiconductor interfaces. It can be important if interfaces are rough and have steps such as in misoriented compound semiconductor growth. Since carriers in two-dimensional systems are confined very close to the interface, any surface-potential perturbation such as through surface atomic steps leads to a decrease in the mobility. The silicon surface mobility is almost half as much as the corresponding bulk mobility.

Steps are not the only way this scattering can cause a net energy or momentum change of the carriers. If there are a number of surface states (surface states result from a number of causes, principal among which are surface reconstruction and nonterminated bonds at the surface), they may cause perturbation effects of their own even for a perfectly smooth surface. For an ideal interface, an incident carrier goes through a reflection without losing its energy. Its momentum parallel to the interface is maintained while its momentum perpendicular to the interface is reversed. An example of this is the $Ga_{1-x}Al_xAs$/GaAs interface, which can be made both atomically smooth as well as with negligible interface state density. A low-energy electron that cannot penetrate the $Ga_{1-x}Al_xAs$/GaAs barrier loses no energy and maintains the momentum parallel to the interface. From the scattering perspective, a nonideal interface is the SiO_2/Si interface. Quite a large fraction of incident carriers lose energy and momentum at the interface because of interface states and atomic roughness. The result of such scattering, called diffuse scattering, is a thermalization of carrier energy distribution and decrease of mobility.

Carrier–carrier scattering conserves the total energy and momentum of the interacting particles. This scattering, also sometimes referred to as plasma scattering, does cause redistribution of the energy and momentum among the interacting particles, and hence its primary effect is in influencing the carrier distribution as a function of energy and momentum. Thus, the scattering influences those parameters that depend on the shape of the distribution of carriers in energy and momentum. The relaxation rate of energy and transfers to other valleys or bands that depend strongly on energy tails do therefore get influenced directly by the carrier–carrier interactions. An additional important way that carrier–carrier scattering influences transport is through its coupling with other scattering mechanisms—the coupling of plasma scattering together with phonons, referred to as plasmon scattering, is an important scattering mechanism in hot-electron devices. In inversion or accumulation layers at high charge densities, this can be quite significant.

Lattice scattering is a broad term used for those scattering mechanisms that arise from the existence of the lattice itself such as due to optical and acoustic phonons. An electron that scatters via optical-phonon scattering can have a large change in momentum since optical phonons can have large momentum. Large momentum changes, most often, occur with changes in the valleys or bands that the carriers occupy. Holes, with their degenerate bands, can easily change bands—e.g., between light-hole and heavy-hole; thus corresponding scattering processes, termed interband scattering processes, can occur even at low energies. These interband scattering processes are unlikely for electrons at low energies. High-energy electrons, such as during an avalanche process, may change bands. All such processes involving changes in bands are called interband scattering processes, while if a process occurs within a band, it is called an intraband scattering process.

Electrons need ≈36 meV in GaAs before they can emit an optical phonon, i.e., lose energy in an optical scattering process, since the optical-phonon energy in GaAs is ≈36 meV. With lowering of temperature, the consequence of this high energy of optical phonons is a decrease in their numbers, referred to as freeze-out of optical phonons. Corresponding to this smaller likelihood of occurrence of optical phonons, there is a smaller likelihood of absorption of an optical phonon. Hence, at low temperatures there is a reduction in optical-phonon scattering. The emission process can still occur and continues to be the leading cause of loss of energy at high electric fields. Both of these factors are important to the observed velocity–field behavior and operation of high-mobility and hot-electron structures as a function of temperature and bias.

Now consider the nature of the perturbing potential in acoustic mode. A change in atomic spacing, i.e., deformation of the lattice, leads to a disturbance in the potential. The magnitude of the deformation potential is proportional to the strain induced by vibrations, and the corresponding scattering is called deformation-potential scattering. In addition to this spacing effect, if the atoms are slightly ionic as in all of the compound semiconductors, then the displacement of the atoms leads to a potential perturbation due to the charge on the ions. This is called the piezoelectric potential, and the resulting scattering process is called piezoelectric scattering. It is important in compound semiconductors at low temperatures in relatively pure materials (see Fig. 10). Recall that optical scattering is weak at low temperatures and low fields on account of optical-phonon freeze-out, etc., and impurity scattering is weak because of purity and carrier freeze-out, leaving acoustic-phonon scattering as the dominant process. The deformation-potential scattering is weaker than the piezoelectric scattering, and hence piezoelectric scattering dominates.

Optical vibrations also lead to potential perturbation in two ways similar to the case of acoustic vibrations. For one, the perturbation effect of the strain, which is nonpolar in nature, leads to a scattering that is sometimes referred to as nonpolar optical scattering. The second, due to the polarization from the ionic charge, is referred to as polar optical scattering. In GaAs, AlAs, etc., both these scattering processes are important. In general, for compound semiconductors, the polar optical-phonon scattering is the dominant optical scattering mechanism.

3.2 Majority-Carrier Transport

We will now discuss the relationship of these various scattering mechanisms with the steady-state transport behavior of carriers. The operation of a device is interlinked with the transport behavior of carriers. The velocity–field curves for carriers in a semiconductor (Fig. 11) describe the steady-state behavior of the carriers for both low fields, where perhaps an equally useful parameter is the mobility, and high fields. The transport of the carrier is related to the scattering of the carriers. Minority-carrier transport is important in the base of a bipolar transistor. It can be significantly different from a majority-carrier device at comparable background doping and identical temperatures. This comes about because of changes in scattering behavior that result from changes in screening effects of the perturbations. Scattering can also change when the band occupation behavior is changed such as through

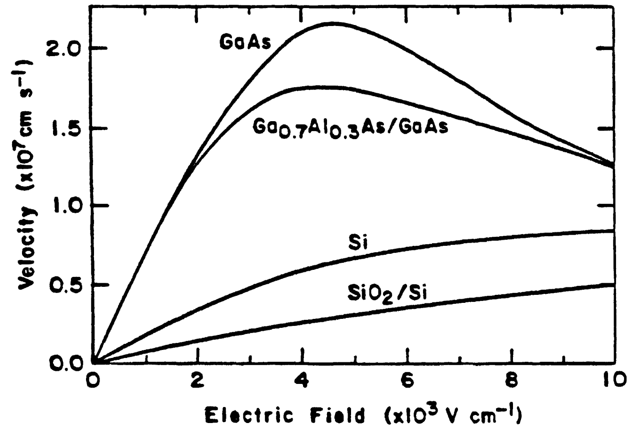

FIG. 11. Steady-state velocity–field characteristics of electrons, in lightly doped bulk Si and at the SiO_2/Si interface, and in lightly doped GaAs and at the $Ga_{0.7}Al_{0.3}As$/GaAs interface at 300 K.

the quantum-mechanical effects of two-dimensional carrier gas systems.

We now consider scattering rates of some common compound semiconductors in order to relate them to the transport characteristics of interest in the operation of devices. Figure 12 shows some of the scattering rates due to phonon processes, at 300 K and low energies, for GaAs, InP, $Ga_{0.47}In_{0.53}As$, and InAs. The minimum in the scattering rates, occurring at very low energies, is due to acoustic scattering. At the low-energy threshold for the optical-phonon modes, scattering rises rapidly because of the large rate of optical-phonon scattering at 300 K. At 77 K and lower temperatures, this rate would be lower. At low energies the scatterings are intravalley events. At higher energies, comparable with or higher than the secondary valley separation, a large density of states is available for intervalley scattering. So, these scattering rates show another threshold where the secondary valley occurs in GaAs, InP, and $Ga_{0.47}In_{0.53}As$. InAs has a secondary valley at higher than 1 eV and does not, therefore, show such a threshold. Note that InP has the highest low-energy scattering; it has both a higher effective mass that leads to higher density of states to scatter into and a stronger coupling coefficient for scattering. Both $Ga_{0.47}In_{0.53}As$ and InAs show favorable low scattering-rate characteristics over the energy range, a behavior partly due to the larger secondary valley separation, which inhibits intervalley scattering over a significant energy range.

The two divisions of lattice scattering that we have considered are intervalley and intravalley scattering. In intravalley scattering, low-momentum phonons are involved. At low fields, this intravalley scattering is important in materials where the conduction-band minimum occurs at the Γ point, e.g., GaAs, InP, and most other compound semiconductors. Intervalley scattering involves large-momentum phonons and is important at low fields in materials with conduction-band minima at nonzero crystal momentum, i.e., non-Γ-minimum semiconductors.

The phonons that take part in intervalley scattering are sometimes referred to as intervalley phonons even though they may be either of the acoustic or of the optical type. When fields are high, electrons pick up enough energy between scattering events that even where the central valley is the minimum energy valley, intervalley scattering becomes important. The negative differential mobility of GaAs, e.g., arises from intervalley scattering from Γ to L and X valleys. Figure 13 shows the velocity–field characteristics for GaAs (≈mid-10^{16} cm^{-3} doping) as a function of temperature. GaAs, InP, etc., actually show a significant region of negative differential mobility arising from the transfer of carriers to secondary valleys at the higher energies. Transfer to the secondary valleys occurs into a high density of states and therefore occurs with a large scattering rate. The resulting pronounced negative differential characteristic has been utilized in achieving negative differential resistance for gener-

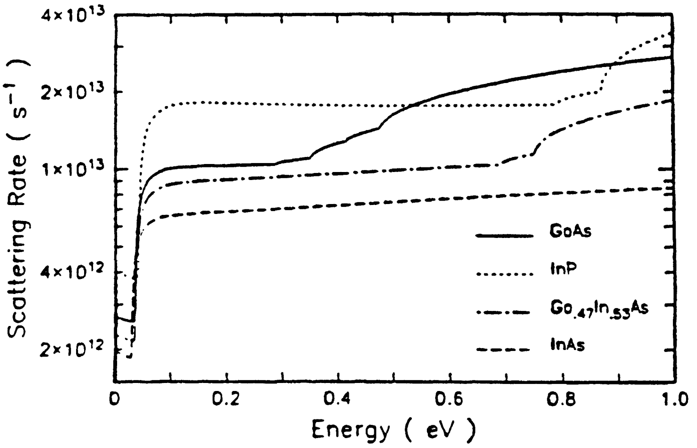

FIG. 12. Steady-state phonon scattering rates as functions of energy, at low energies, in GaAs, InP, $Ga_{1-x}In_xAs$, and InAs at 300 K.

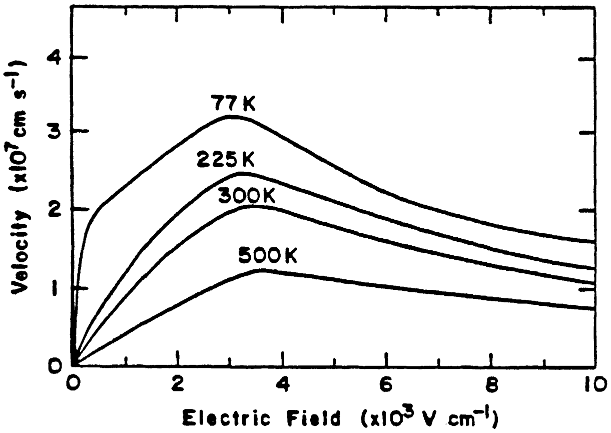

FIG. 13. Velocity–field characteristics for electrons in lightly doped GaAs as a function temperature.

ation of microwave power. The decrease in mobility at low fields is due to intravalley scattering, the decrease in velocity at higher fields is due to intervalley scattering, and the saturation in velocity at very high fields is due to a balance between the energy gain between scattering events and the energy loss due to high intervalley scattering.

In thermal equilibrium, the energy and momentum exchange between electrons and the lattice is balanced out by phonon absorption and emission. With field, the energy and momentum is lost at a higher rate to the lattice, and at moderate fields, the intervalley optical-phonon scattering processes dominate. The momentum loss causes randomization of the electron velocity distribution in momentum space and consequently results in a finite mobility. Note that the momentum-loss rate is higher than the energy-loss rate (also recall that optical phonons involve a loss of ≈36 meV of energy per emission event). These losses are fitted phenomenologically by time constants called relaxation times (τ_p for momentum and τ_w for energy).

We may now describe some general features of the relative importance of different scattering mechanisms in the transport of carriers. Since ionized-impurity scattering is inefficient at higher carrier energies, it becomes important in the 77- to 300-K temperature range. Alloy scattering is important for transport in mixed crystals, in the temperature range below 300 K. Carrier–carrier scattering preserves both the energy and the momentum if it is between carriers of the same type. It therefore causes a substantial difference between the transport of a majority carrier and a minority carrier. Lattice scattering becomes dominant at high temperatures and high carrier energies. Optical phonons, with their large momentum and energy compared to the acoustic phonons, are dominant when carriers have large energies and are dominant in intervalley transfers. Acoustic phonons establish a lower limit in scattering since they can exchange very low energies also. At very low temperatures, therefore, while optical-phonon absorption can be frozen out because of their reduced density, acoustic phonons continue to be important. At low temperatures and low fields, both the acoustic-phonon and the impurity scattering are important. At high fields at very low temperatures, impurity scattering becomes weak because of the higher carrier energy, but optical-phonon emission becomes important. Acoustic-phonon scattering should also be considered. At high temperatures phonon scattering prevails, and both absorption and emission processes become dominant.

Now, let us consider the temperature dependence of the low-field and high-field behavior in the velocity–field curves of Fig. 13. At high fields, as we have already discussed, this velocity is dominated by intervalley scattering processes occurring largely through optical phonons, which are polar in compound semiconductors and nonpolar in silicon and germanium. At low fields, other mostly intervalley scattering processes dominate. The low-field behavior is equally well characterized by the mobility, the drift mobility, which relates the velocity with the low electric field. While drift mobility is the real mobility of interest in most electronic devices, one quite often also uses Hall mobility, which is a far easier and accurate transport parameter to measure.

Drift mobility, which we have discussed thus far, depends on the carrier scattering embodied in the phenomenological carrier relaxation time. To calculate the mobility, we need to determine this relaxation time, including its functional dependence on energy. The average value of this relaxation time then is obtained by averaging over the energy states that the carrier occupies. For electrons, using a parabolic band approximation, we obtain

$$\tau_\mu = \bar{\tau}. \tag{1}$$

The averaging is over energy, which is related to the momentum in the parabolic approximation through a square law. The relaxation time, therefore, is also often written as

$$\tau_\mu = \frac{\langle (E - E_c)\tau \rangle}{\langle E - E_c \rangle} = \frac{\langle v^2 \tau \rangle}{\langle v^2 \rangle}. \tag{2}$$

The time constant for mobility can be calculated given knowledge of what scattering process is dominant. Alternately, if more than one is important, they all have to be evaluated and the resultant determined as a geometric mean. The geometric mean comes about because scattering rates add, and the time constant is inversely proportional to this rate. Thus, the time constant τ_μ is related to the time relating the inverse of scattering rates $\tau_1, \tau_2, \ldots$

$$\frac{1}{\tau_\mu} = \frac{1}{\tau_1} + \frac{1}{\tau_2} + \cdots, \tag{3}$$

and the drift mobility is related to this time constant through*

$$\mu_{\text{drift}} = \frac{q\tau_\mu}{m^*} = \frac{q}{m^*}\frac{\langle v^2 \tau \rangle}{\langle v^2 \rangle}. \tag{4}$$

If one associates $\mu_1, \mu_2, \ldots$ as the mobilities with the various scattering processes corresponding to $\tau_1, \tau_2, \ldots$, then

$$\frac{1}{\mu} = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \cdots. \tag{5}$$

For a given momentum, the velocity of carriers is inversely related to the effective mass. A larger effective mass results in smaller velocity for identical forces on the carrier. Since, in general, the effective mass is a tensor, it has a directional dependence. Let us now consider how the occupation of the bands, the motion of carriers due to nonisotropy of masses, and occupation of more than one degenerate bands are related. In thermal equilibrium, the carriers occupy bands following Pauli's exclusion principle. Constant-energy surfaces are used to show those momentum values that have equal energy and hence an equal likelihood of occupation in thermal equilibrium. The shapes of such surfaces are dependent on the band anisotropy. In cubic crystals, at thermal equilibrium, three types of constant-energy surfaces are usually encountered. They are a spherical surface encountered for Γ valley minima in the conduction band, an ellipsoidal surface encountered for L and X valley minima in the conduction band, and a warped surface encountered for two degenerate valence bands. A parabolic band has an isotropic mass; its constant-energy surface is spherical. L- and X-valley minima, however, have a rotational symmetry around the crystallographic directions and show a differing effective mass in the transverse direction to it than in the longitudinal direction. The constant-energy surface is an ellipsoid. Valence bands are usually doubly degenerate at zone center. For any energy in excess of this zone-center energy, the constant-energy surface is warped.

Thus, even for moderate doping, in cases where the band structure is not a simple parabola, where the band minimum does not occur at the zone center, and where degenerate bands exist, the effective mass for use in conduction calculations has to be derived as a suitable average that reflects the net effect. We first consider the conduction band. For the lowest conduction-band minimum in the Γ band, the bands are quite isotropic, and we may use the usual mass. At higher energy, however, e.g., when a large field is applied and intervalley transfer occurs, we may not use this constant isotropic effective mass. Transfer of carriers, in GaAs, occurs first to the L valley and then to the X valley with increasing energy. An L valley has eight equivalent minima in eight different directions. However, since the minimum always occurs at the Brillouin-zone edge, only four equivalent complete valleys need to be considered. For X valleys, the minimum occurs slightly inside the Brillouin zone, and hence all six equivalent minima and six equivalent valleys should be considered. The L and X minima are ellipsoidal surfaces of constant energy; i.e., they can be represented by

---

*This relationship, Mathiessen's rule, has limited, not universal, validity since it is based on the relaxation-time approximation.

$$E - E_c = \frac{\hbar^2}{2m_l^*}(k_l - k_{ol})^2 + \frac{\hbar^2}{2m_t^*}(k_{t1} - k_{t2})^2, \tag{6}$$

where the subscripts l and t denote longitudinal and transverse directions. This is simply a translation of axes from the original Cartesian coordinate in the reciprocal space to a new Cartesian coordinate that is aligned along the longitudinal axis, and because of symmetry, the transverse variations are identical. In terms of the effective mass, if the transformed x axis is now aligned with the longitudinal axis, then the new effective mass tensor is

$$\frac{1}{m^*} = \begin{pmatrix} 1/m_l^* & 0 & 0 \\ 0 & 1/m_t^* & 0 \\ 0 & 0 & 1/m_t^* \end{pmatrix}. \tag{7}$$

If the electric field is applied in the $\langle 100 \rangle$ direction, for X minima, there are four ellipsoid surfaces for which the transverse effective mass is the relevant mass in the determination of the response of the carrier to the field. There are two ellipsoidal surfaces for which the longitudinal mass is relevant. Hence, in the process of approximating into the parabolic band picture, i.e., an expression of $E - E_c = \hbar^2 k^2/2m_c^*$, we must use the mass

$$m_c^* = \tfrac{1}{6}[2m_t^* + 4(m_l^* m_t^*)^{1/2}]. \tag{8}$$

For the L-valley minima with a field in the $\langle 100 \rangle$ direction, all the valleys are equivalent with the effective conduction mass

$$m_c^* = \tfrac{1}{4}(\tfrac{1}{3} m_t^{*2} + \tfrac{2}{3} m_l^* m_t^*)^{1/2}. \tag{9}$$

The valence band is centered in the Brillouin zone, and the light- and heavy-hole bands are degenerate. The equivalent mass can be derived as

$$\frac{1}{m_c^*} = \left(\frac{g_{hh}(E)}{m_{hh}^*} + \frac{g_{lh}(E)}{m_{lh}^*}\right)\frac{1}{g_{hh}(E) + g_{lh}(E)}, \tag{10}$$

where $g_{hh}(E)$ and $g_{lh}(E)$ are the heavy-hole and light-hole densities of states. The ratio of densities of states in the expression above represents the probability of finding the hole in a state with that mass. Since the density of states varies as $\frac{3}{2}$ power of the mass, the conduction mass for holes is

$$\frac{1}{m_c^*} = \frac{m_{hh}^{*1/2} + m_{lh}^{*1/2}}{m_{hh}^{*3/2} + m_{lh}^{*3/2}}. \tag{11}$$

Hall mobility is a mobility parameter that results from Hall-effect measurement. It is a simple and hence very common measurement to gauge the low-field transport characteristic of semiconductors. Our discussion here is to emphasize its relationship with the drift mobility, which is of more immediate interest from a device perspective. The Hall effect is a phenomenon that results from the application of a magnetic field on a sample. Consider an n-type rectangular semiconductor sample with a magnetic field applied perpendicular to the sample, a voltage bias applied to force a current to flow in the plane of the sample, and a voltage being measured perpendicular to this flow of current. A voltage, called the Hall voltage, develops in the perpendicular direction because of the Lorentzian force acting on the charged particle. For the electron, this force is given by

$$\mathbf{F} = -q(\mathbf{v} \times \mathbf{B}). \tag{12}$$

In steady state, carriers accumulate and deplete in the transverse direction causing a transverse electric field, which opposes and cancels the field due to the Lorentzian force, to appear. This allows us to write the electric field in the transverse direction, related to the Hall voltage developed across the transverse direction, in terms of the magnetic field, the current, and a parameter called the Hall coefficient. This field is related as

$$\mathscr{E}_z = R_H J_x B_y, \tag{13}$$

where R_H is the Hall coefficient. The current itself is related to the drift mobility of carriers. The Hall coefficient is related to this drift mobility by

$$R_H = r\left[\frac{n\mu_{dn}^2 - p\mu_{dp}^2}{(n\mu_{dn} + p\mu_{dp})^2}\right] \qquad (14)$$

for a sample with both electron and hole conduction. The prefactor r, the Hall factor, usually varies between 1 and 2 for most common scattering processes. This factor is related to the relaxation time for Hall mobility determination, τ_H (analogous to τ_μ, the relaxation time for drift mobility determination), by

$$\tau_H = \langle\tau^2\rangle/\langle\tau\rangle. \qquad (15)$$

Note the difference between this and the relaxation time associated with the drift mobility. If only acoustic scattering dominates, and the conduction and valence bands can be assumed to be parabolic, then $r = 3\pi/8$. The motion of carriers in the transverse direction occurs as a result of the applied magnetic field and is determined by the velocity in the longitudinal direction. The proportionality of constant in the longitudinal direction is the drift mobility; however, the proportionality in the transverse direction is a mobility term that depends on the same prefactor as in the Hall factor. Hence, the Hall mobility ratio in samples with unipolar conductivity is

$$\mu_H = r\mu_d. \qquad (16)$$

Because the ratio r varies between 1 and 2, the Hall mobility is always larger than the drift mobility. Its significance and ubiquity is largely a result of the ease of its measurement.

The Hall and drift mobility are, of course, related to each other through the characteristics of the dominant scattering processes, i.e., through the ratio r above. Figure 14 shows the Hall mobility as a function of temperature for various dopings of GaAs.

3.3 High-Field Transport of Electrons

The mobility of carriers decreases as a function of particle energy because the scattering rate continues to increase as a function of the energy. Thus, drift and Hall mobilities become negligible at high energies because carriers move short distances between scattering events.

It is the velocity of carriers that is of direct interest to us. At high energies, even though the mobility becomes negligible, the velocity for a high applied field is still high. At low energies, both the velocity and the mobility are useful parameters because they are directly related to each other. However, mobility is used more commonly, both because of the ease of measurement and because of simplifications in many steady-state drift–diffusion calculations in devices.

The saturation of velocity at high fields occurs for the following reason. Carriers drift in the high field between scattering. Any increase in energy resulting from this transport in high field causes an increase in scattering rate also. Increase in electric field,

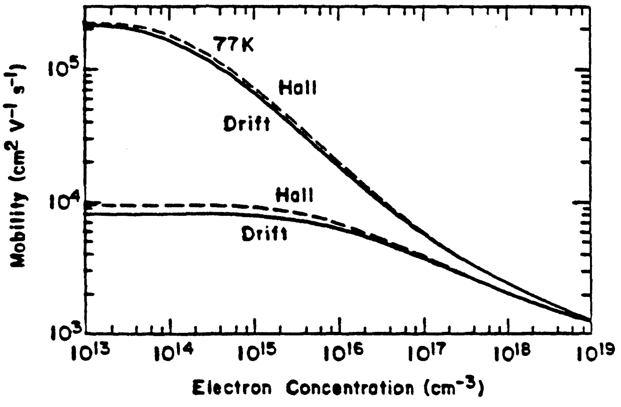

FIG. 14. Drift and Hall mobilities as functions of background electron density at 300 and 77 K for GaAs.

therefore, does not result in significant additional increase in the velocity, and the mobility becomes negligible. An alternative way of looking at this behavior in high fields, where the carrier is in equilibrium with its local conditions, is to note that the mobility is a function of the carrier energy. At low fields, the average carrier energy increases proportionally with the electric field, the scattering rate increases proportionally, and hence the carrier mobility remains constant and independent of the energy and the electric field. At high fields, however, the carrier energy increase causes the mobility to decrease inversely with the electric field and hence velocity saturation.

Optical-phonon scattering is the principal source of scattering in the high-field region. Since the occupation of the optical modes decreases strongly as the temperature decreases, energy relaxation is not as strong, and hence absorption processes are less likely. The drift velocity at high fields also shows a temperature dependence. An example of this is given in Figure 15, where there is nearly a 20% change in the drift velocity at high fields (approximately the saturated velocity) between 300 and 77 K.

4. MINORITY-CARRIER TRANSPORT PROPERTIES

An electron as a majority carrier suffers Coulomb scattering from the donors, other electrons, and any other residual impurities. Usually, the latter two are small; only at high carrier densities, such as at high dopings or in two-dimensional electron gas, do the carrier–carrier interactions also become strong. An electron as a minority carrier suffers Coulomb scattering from the acceptors, the holes, other electrons, and any other residual impurities. The last two, as in the case of electron as a majority carrier, are usually small. Holes are the new twist to the environment of the electron.

The behavior of the electron as a minority carrier depends on the semiconductor and on any residual field in the semiconductor device. It depends on the semiconductor because any interaction between electron and hole—rather the heavy hole since it is far more common—involving exchange of momentum and energy is a function of their relative effective masses. The screening of acceptors by holes (i.e., reduction of Coulomb perturbation by holes in the vicinity of the ionized acceptor) may allow for an increase in the carrier mobility. Holes, if they are considerably heavier than the electron, may cause scattering very similar to that due to ionized acceptors. A slight residual field in the structure causes motion of carriers in opposite directions, which may result in a drag effect on each carrier by the other. The magnitude of the effect on each carrier is a function of their effective mass and the disparity in the effective mass. Scattering is a strong function of temperature; hence, different considerations may apply to different temperatures.

Figure 16 shows a plot, for electrons, of minority- and majority-carrier mobilities in

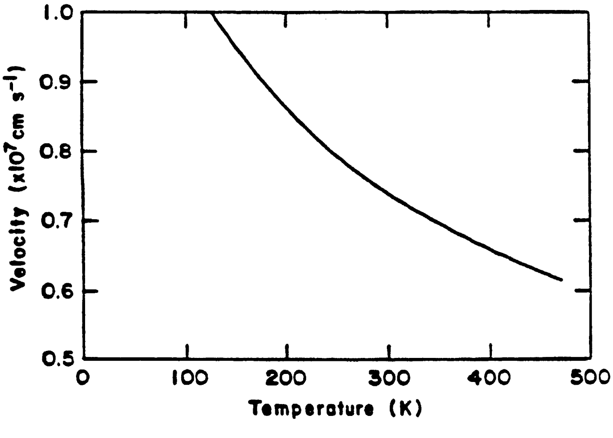

FIG. 15. Velocity of electrons at a field of 50 kV/cm in lightly doped GaAs as a function temperature.

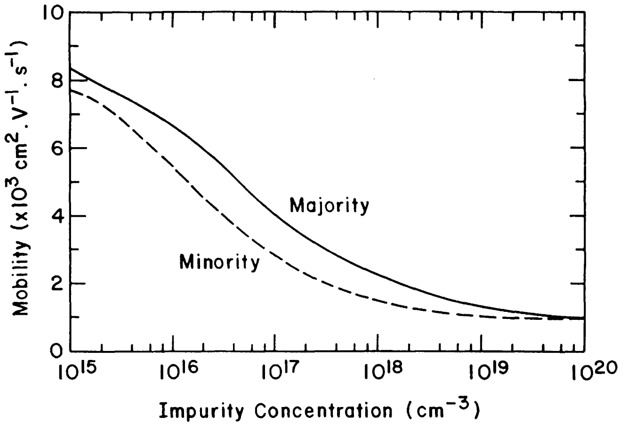

FIG. 16. Majority- and minority-carrier mobilities for electrons for GaAs.

GaAs as functions of background donor and acceptor density. This figure is for mobility at 300 K. GaAs shows a minority-carrier mobility that is nearly half the value of the majority-carrier mobility. Comparatively, silicon shows similar mobility for electrons as a majority and minority carrier. In silicon, the effective masses for electrons and holes are relatively similar. Exchanges involving electrons and holes, and screening of impurities by holes, together contribute to this similarity of mobilities. In GaAs, electrons and holes have widely differing masses. Holes can be considered static for scattering purposes and lead to Coulomb scattering of a similar magnitude as the ionized acceptor. This behavior may change at a different temperature. At 77 K, e.g., for GaAs in the 10^{15}-cm^{-3} doping range, the minority-carrier mobility for electrons is higher than that as a majority carrier.

Minority-carrier devices usually have fields in regions where the electron transits. In the base of the bipolar transistor, e.g., a small field exists even for uniformly doped base structures. This field is the result of the gradient in the hole distribution established to compensate for the electron distribution. It causes a rapid decrease in mobility with the electric field, even though the mobility at zero field is quite comparable to the majority-carrier mobility. The majority-carrier mobility, of course, does not vary substantially. This difference in behavior is usually ascribed to the electron–hole drag effects, which are quite significant for Si because of comparable effective masses. The carrier–carrier scattering conserves momentum. In the absence of electric field, the average drift momentum is zero. In the presence of drift field, electrons and holes move in opposite directions. The interacting electrons and holes that are in the vicinity of each other and are, on an average, moving in opposite directions, exert a Coulomb attractive force on each other. This reduces the effect of the drift field, and externally, appears as the lowering of mobility. The effect is strong and shows up at a low electric field. At higher fields, the drag effect does not remain as important because the Coulomb scattering rate decreases (the electron now has a higher energy).

The consequence of such an effect can be strong in devices such as a bipolar transistor, even if no electric field is built into the structure. In the presence of current, there exists a gradient of electron density in the base. In order to maintain charge neutrality, the hole density also changes as a function of position to compensate for the excess charge due to electrons. The excess hole population is equal to the electron population under quasistatic conditions. The diffusion current due to the gradient in hole density is nearly compensated by the drift current resulting from a small electric field that is established in the base. The magnitude of this electric field can be, approximately, found by setting the hole transport current equal to zero (it is substantially smaller than the electron current for the *n-p-n* bipolar transistor). The current-density equation in the drift–diffusion approximation can be used to model this condition. For an electric field $\mathscr{E}$ and diffusivity $\mathscr{D}_p$ for holes, the absence of hole current implies

$$J_p = q\mu_n \mathscr{E} p - q\mathscr{D}_p \frac{dp}{dz} \approx 0; \tag{17}$$

i.e.,

$$\mathscr{E} \approx \frac{kT}{q}\frac{1}{p}\frac{dp}{dz}. \tag{18}$$

Because the magnitude of dp/dz is negative, the electric field is pointed toward the electron-injecting electrode.

This is a very substantial simplification of a considerably complex many-body problem. It has been argued that under certain conditions, this drag effect should be balanced by the effect of majority-carrier distribution. Hole current being small, the holes are nearly stationary. Thus, they should not transfer much momentum to electrons. Electrons, however, do transfer momentum to holes, resulting in the space charge and electric field to prevent the resulting hole current. This field aids the motion of electrons, just as in the high injection effect, and hence should compensate.

4.1 Diffusion Coefficients

The phenomenon of diffusion is one of the very direct consequences of scattering. Net diffusion takes place from regions of high carrier concentrations to regions of low carrier concentrations. Regions of high concentrations have a higher carrier population, resulting in a higher number of scattering events. Some of these lead to carrier motion toward the low–carrier-concentration region. However, it is not completely compensated for by scattering events that lead to the motion of carriers from the low–concentration-region to the high–concentration-region. The net result is a flow, or diffusion, of carriers determined by the gradient of the carrier concentration and characterized by the diffusion coefficient ($\mathscr{D}_n$ for electrons and $\mathscr{D}_p$ for holes).

Diffusion coefficients are related to mobilities because their underlying basis is in scattering. This relationship, the Einstein relationship, was derived for Brownian motion in the Maxwell–Boltzmann distribution limit of classical gases. In thermal equilibrium (i.e., with constant quasi-Fermi levels E_{Fn} for electrons and E_{Fp} for holes) in an n-type sample,

$$J_n = qn\mu_n \mathscr{E} + q\mathscr{D}_n \frac{dn}{dz} = 0, \tag{19}$$

and its converse. The carrier concentration is given by

$$n = N_c F_{1/2}\left(\frac{E_{Fn} - E_c}{kT}\right), \tag{20}$$

where $F_{1/2}(\ )$ represents the Fermi integral of order $\frac{1}{2}$. This allows us to show that

$$\frac{\mathscr{D}_n}{\mu_n} = \frac{kT}{q}\frac{F_{1/2}((E_{Fn} - E_c)/kT)}{F_{-1/2}((E_{Fn} - E_c)/kT)}, \tag{21}$$

the Einstein relationship for electrons, and following similar arguments,

$$\frac{\mathscr{D}_p}{\mu_p} = \frac{kT}{q}\frac{F_{1/2}((E_v - E_{Fp})/kT)}{F_{-1/2}((E_v - E_{Fp})/kT)}, \tag{22}$$

the Einstein relationship for holes. In the Maxwell–Boltzmann limit, these can be simplified to

$$\frac{\mathscr{D}_n}{\mu_n} = \frac{\mathscr{D}_p}{\mu_p} = \frac{kT}{q}. \tag{23}$$

The diffusion coefficients of electrons in GaAs and other compound semiconductors are larger than in silicon for the same reason as for mobility—largely because of a smaller effective mass. The high diffusion coefficient leads to shorter base transit time in bipolar transistors in GaAs and other compound semiconductors for comparable base widths. It also has implications for noise behavior; a larger diffusion coefficient usually results in smaller magnitude of noise. Since, in general, conduction-band structure can be anisotropic, scattering processes are anisotropic, and since fields exist along specific directions, the diffusion coefficient shows anisotropy. Examples of magnitude of diffusion coefficient with field are given in Fig. 17 for the longitudinal and transverse directions with respect to the field.

The anisotropic nature of polar scattering

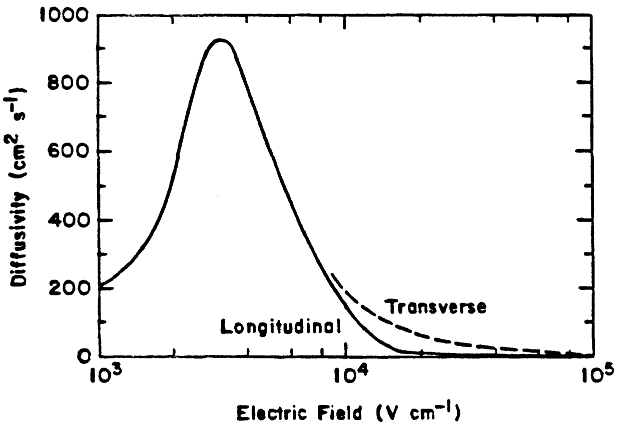

FIG. 17. Diffusion coefficients in the longitudinal and transverse directions for GaAs at 300 K for low doping.

is a particularly strong contributor to this anisotropy. Since polar scattering occurs across the field range so long as it leads to carrier energies above the optical-phonon threshold, the diffusion coefficient also exhibits similar behavior. Anisotropy exists at both low and high fields. At high fields, the *L* and *X* valleys are mostly occupied, the longitudinal effective mass is large, and hence longitudinal diffusion is suppressed. It decreases to very low values compared to the zero-field value, just as in the case of mobility. The pinch-off regions of a field-effect transistor are regions of high fields where the carrier gradient can become large for large–gate-length devices. The low diffusion coefficient reduces the diffusive current in such regions even if carriers move with finite velocity (the saturated velocity) as a result of drift effects.

5. CONFINEMENT EFFECTS

5.1 Transport Consequences of Confinement

In two-dimensional electron gases, scattering due to ionized impurities occurs through a remote process because donors and carriers are separated. The perturbation in potential due to ionized donors (sometimes this is referred to as Coulomb scattering also) is separated from the electron gas, as shown in Fig. 18. Decrease of the ionized-impurity scattering and increase of the carrier screening of the ionized impurities because of large carrier densities in the two-dimensional channel result in a significantly higher mobility. Now, scattering processes such as the piezoelectric scattering become important, at least in part of the temperature range, as shown in Fig. 19. At the lowest temperatures, the residual background doping at the high-mobility interface is still important, although because this is low, the mobility is very high. It is interesting to compare this figure with Fig. 10, which shows the mobility versus temperature for low-doped GaAs. Note that over a fairly broad range of temperature, near and above 77 K, the behavior is very similar. Below that temperature, the improvements due to carrier screening make the mobilities of the two-dimensional electron-gas system significantly larger.

The added complication in the two-di-

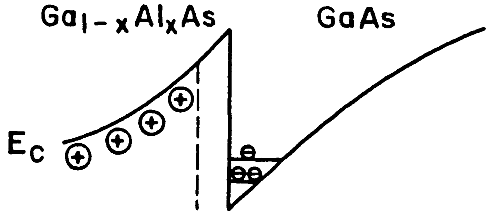

FIG. 18. Separation of two-dimensional electron gas from the ionized donors in a two-dimensional electron gas structure. The larger–band-gap material ($Ga_{1-x}Al_xAs$) from which electron transfer occurs is shown on the left. The right-hand side shows the smaller–band-gap material (GaAs) to which the transfer occurs. An undoped spacer layer, used to decrease remote Coulomb scattering, is also shown in the figure.

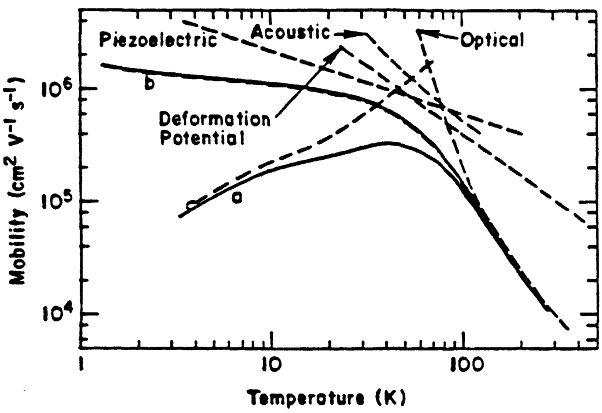

FIG. 19. Theoretical mobility in a two-dimensional electron gas at the $Ga_{0.7}Al_{0.3}As/GaAs$ interface as a function of temperature. The figure also shows components of mobility arising from different scattering mechanisms in bulk pure GaAs.

mensional electron-gas structures is that one should also consider the quantization and its effects on scattering, as well as the effect of carrier–carrier interaction and carrier–dopant interaction. The former leads to a decrease in mobility; the latter tends to screen the perturbation potential associated with the dopant. Thus, there is actually an optimum in carrier concentration in the channel where the two balance and allow the highest mobilities at low temperatures. Note the large increases in low field mobility that can be achieved by a decrease or removal of the ionized-impurity scattering. At 300 K, since ionized-impurity scattering is only one of the two dominant scattering processes, the increase is not very large. At 77 K, mobilities of greater than 70 000 cm^2 V^{-1} s^{-1} are quite common. The behavior of two-dimensional electron gases is quite complicated; it has increased carrier–carrier scattering whose effect, as we have discussed, is to change the distribution function but not to affect the energy and momentum.

The interaction of carrier–carrier scattering processes with other processes are also important in these structures. Coupling between optical phonons and the carrier plasma—the plasmon—becomes stronger as a function of carrier density in GaAs. Significant deviation and large energy effects begin to occur at mid-10^{17} cm^{-3} carrier density. Carrier densities higher than this exist in two-dimensional electron gases in device structures, and therefore mobility is fairly sensitive to carrier scattering. Too low a carrier density leads to reduction of screening of Coulomb scattering from remote and local donors, and too high a density leads to increased carrier–carrier scattering effects.

6. HOT-CARRIER PROPERTIES

There are a number of device-related phenomena that can be intuitively related to the band structures. One example of this is the relationship between scattering rate and density of states. Scattering is proportional to the coupling between the two states connected by a scattering process and the number of states available for scattering into. Thus, at the onset of secondary valley transfer, i.e., when carriers have kinetic energy approximately equal to the intervalley separation, a rapid rise in scattering rate occurs, as seen in Fig. 12 for GaAs, InP, and $Ga_{1-x}In_xAs$. Also note that at low energies, InP has the highest scattering rate. InP has the highest effective mass of the four semiconductors in the figure; it therefore has the highest density of states.

Effective mass, which is inversely proportional to the second derivative of the energy with respect to the wave vector, i.e., related to the band structure, is central to a lot of transport parameters of interest. Mobility is inversely proportional to it; a lower effective mass usually implies faster low-field transport characteristics. It also results in improvement in diffusion coefficient and higher tunneling probability. A smaller effective mass means lower density of states, and hence carriers have a higher velocity for the same energy, and so long as the scattering rate is lower, the nonequilibrium effects are stronger. The highest velocities, the maximum group velocity, that the carriers can attain under the most ideal of circumstances is also limited by the band structure.

Another good example of a strong band-structure relationship exists in the phenomenon of avalanche breakdown due to impact ionization. This is a process in which a carrier picks up enough energy between scattering events to cause an electron to jump from the valence band into the conduction band during a scattering event. The process is both band-structure and scattering-process dependent.

6.1 Avalanche and Other Generation Effects

The breakdown phenomenon of most interest occurs in depletion regions of devices. In bipolar transistors, it is the breakdown in the base–collector junction region. In field-effect transistors, it is the breakdown in the drain-to-substrate and gate-to-drain depletion regions. These occur because there exists a high field, due to the reverse bias at the junction, and this high electric field is capable of accelerating the carrier and imparting to it a large kinetic energy. Indeed, this kinetic energy can become sufficiently large to cause an electron, or even more than one electron under the right conditions, to be removed from the valence band and transferred to the conduction band.

The process involving a single electron–hole pair generation consists of a transfer of energy from the hot carrier to an electron in the valence band, which jumps into the conduction band leaving a hole behind. Also, for small–band-gap semiconductors, such as InAs, even multiple such pairs may be generated.

With complicated band structure, electrons can be from different bands, holes can be from different bands, phonons—particularly optical phonons with their large momentum—may be involved, etc. Indeed, since the band structure is orientation dependent, and impact ionization occurs as a consequence of large energy of the carrier from rapid acceleration in a field, in the event of insufficient randomization, it may even be orientation dependent.

The ionization process is characterized in a semiconductor by the parameter ionization rate. Ionization rate is the relative increase in carrier density per unit length of carrier travel. Thus, the electron ionization rate (α_n) is the rate of increase in number of electrons per unit length of the travel of an electron.

For GaAs, the impact ionization occurs largely from carriers that are not in the lowest bands. Electrons cause ionization from higher bands, and holes generally cause it from the split-off band. At low fields, the ionization rate is also orientation dependent because of lack of randomization. Examples of ionization rates as a function of electric field are given in Figs. 20 and 21.

For most compound semiconductors with greater than 1 eV of band gap, the hole ionization rate is larger than the electron ionization rate. The hole effective mass being larger, for similar velocities in the high-field regions, the hole has significantly higher kinetic energy. At lower fields, electrons are more likely to have a higher energy since they usually have a smaller scattering rate than holes. Thus at low fields, the electron ionization rate tends to be larger. For smaller–band-gap semiconductors, InAs and $Ga_{1-x}In_xAs$, e.g., the hole ionization rate is

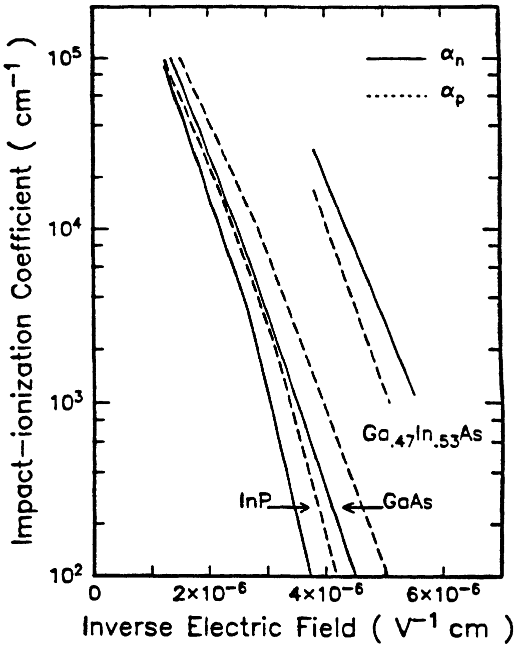

FIG. 20. Impact-ionization rates for electrons and holes for some compound semiconductors.

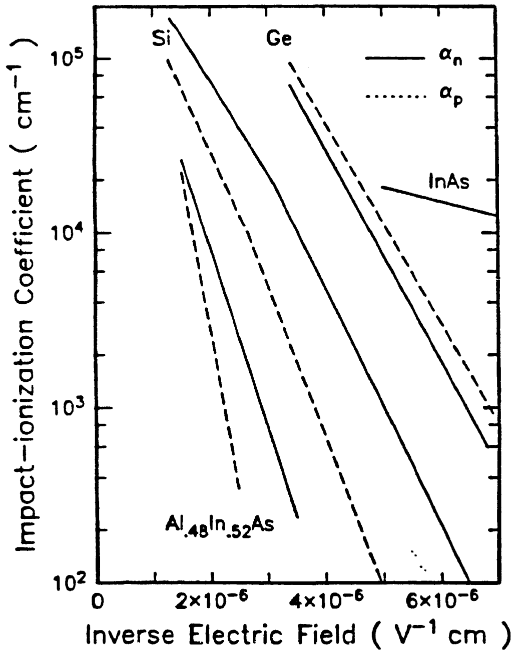

FIG. 21. Ionization rates for electrons and holes for some compound semiconductors, silicon, and germanium.

always low because the threshold energy for ionization is lower.

Impact ionization does not necessarily need an electric field to show a large generation rate for electrons and holes. In compound-semiconductor heterostructures, there exist many smaller–band-gap compounds that can be lattice matched to other semiconductors with a band-gap discontinuity for conduction or valence band that is larger than the band gap. Even in the absence of an electric field, high-energy carriers injected into the smaller–band-gap semiconductor can cause impact ionization. This process is often referred to as Auger generation.

When impact ionization occurs in the presence of an electric field, such as in a *p-n* junction, both electrons and holes get accelerated in opposite directions, and both cause further impact ionization. The process of electron–hole pair creation, acceleration of carriers, and creation of more electron–hole pairs sets up an avalanche, and hence the generation of electron–hole pairs in an electric field is also known as an avalanche process, and the related breakdown of a *p-n* junction, an avalanche breakdown.

The increase in electron or hole current resulting from the ionization caused by electron and hole current flow is

$$\frac{dJ_n}{dz} = \frac{dJ_p}{dz} = -(\alpha_n J_n + \alpha_p J_p), \qquad (24)$$

where J_n is the electron current density and J_p is the hole current density. A simple expression for multiplication factor for electrons ($\mathcal{M}_n$), the ratio between electron current at the collecting edge to that at the injecting edge of the high field region, results in the form

$$1 - \frac{1}{\mathcal{M}_n} = \int_{-z_p}^{z_n} \alpha_n \exp\left[-\int_{\zeta}^{z_n} (\alpha_n - \alpha_p)\, d\eta\right] d\zeta, \qquad (25)$$

where z_n and z_p are the coordinates of the edges of the depletion region.

The condition for breakdown caused by electron-initiated avalanche is, therefore, the condition where $\mathcal{M}_n$ goes to infinity. This occurs when the integral in this equality goes to unity. Avalanche breakdown, therefore, occurs when

$$\int_{-z_p}^{z_n} \alpha_n \exp\left[-\int_{\zeta}^{z_n} (\alpha_n - \alpha_p)\, d\eta\right] d\zeta = 1. \qquad (26)$$

The condition for breakdown with a hole-initiated process can be found following a similar treatment with an analogous mathematical expression of the process.

While generally, at the high fields that dominate avalanche process, the hole ionization coefficient dominates, this is not true for small–band-gap semiconductors, such as $Ga_{0.47}In_{0.53}As$ and InAs. Here hole-initiated processes are less likely, and so even though by virtue of a smaller band gap they are more easily likely to avalanche, because of suppression of hole avalanching, the decrease in breakdown voltage is not as pronounced.

Quite often, the integral on the right of the above equations for multiplication factors is approximated by the relationship $(V/BV)^{\nu}$, where BV is the breakdown voltage; i.e.,

$$\mathcal{M} = \frac{1}{1 - (V/BV)^{\nu}}. \qquad (27)$$

This relationship, depending on the magnitude of the power ν, shows a soft breakdown or a hard breakdown. Soft breakdown is meant to imply a slow and gradual increase in multiplication factor with increasing bias. It occurs when the ionization rates are slowly varying functions of the electric field. A hard breakdown, with $\nu \geq 4$, leads to a rapid change in multiplication factor and comes about when the ionization rate changes rapidly with electric field. Small-band-gap materials usually have a higher magnitude of the ionization rate and a slow change with electric field in it. They therefore exhibit relatively softer breakdown characteristics.

6.2 Off-Equilibrium Transport

When carriers traverse a region where there is a significant gradient in the electric field, they undergo a velocity overshoot in parts of the region. Velocity overshoot is a term that implies that the local velocity v at a local field $\mathscr{E}$ becomes higher than what it would reach for a slowly varying field $\mathscr{E}$ of the same magnitude. A local off-equilibrium occurs. It arises because the carrier energy and the local electric field are no longer in equilibrium with each other. An example of this local equilibrium at high fields is the existence of velocity saturation. The mobility of carriers is a function of energy. Because of the parallel between the energy and the electric field, the mobility can be written as a function of electric field. When the electric field becomes large, higher carrier energy results. Higher carrier energy causes a larger amount of scattering, which reduces the mobility. Thus, velocity, a product of mobility and electric field with mobility varying inversely with field, becomes a constant. When a large gradient of electric field exists, the carrier does not immediately achieve the energy that would correspond to the local field if the field were slowly varying. It actually has a lower energy and hence suffers less scattering, and yet it is accelerated by the high field. The carrier, therefore, is said to have a velocity overshoot. This large velocity occurs in spite of the fact that it has a lower energy. Another way of saying this is that the relaxation rate for momentum is larger than that for energy. This behavior has become increasingly important as device dimensions have shrunk and is particularly important in compound semiconductors because of their lower scattering rates at low energies. Thus, velocity overshoot occurs in field-effect transistors in the rapidly varying field region near the drain where channel pinch-off is usually said to have occurred. It also occurs in bipolar junction transistors in the base–collector region.

An example of such a velocity overshoot, for the case of GaAs, InP, and $Ga_{1-x}In_xAs$, for a gradient in electric field is shown in Fig. 22. In the case of GaAs, the reduction in kinetic energy occurs when the carriers begin transferring to the L valley. Such a transfer can occur after the carrier has acquired an energy of 0.36 eV (the Γ-L intervalley transfer threshold energy). When such transfer can occur, because the carrier has acquired the requisite energy, the carrier scattering rate becomes larger, and the carriers begin to lose energy. The Γ-L transfer of car-

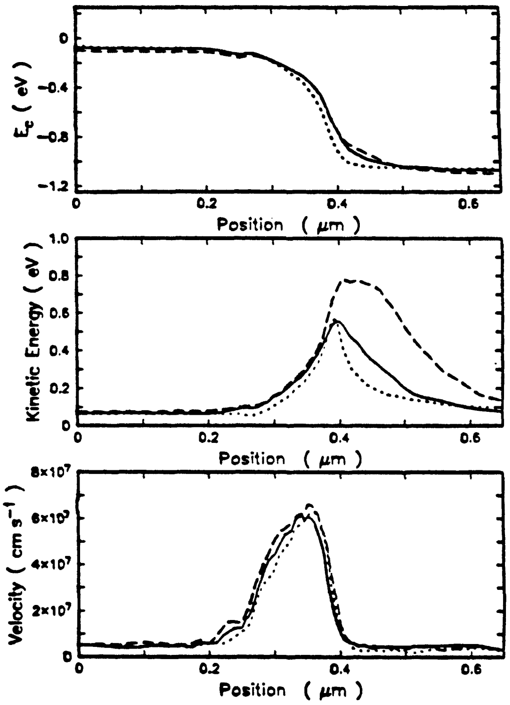

FIG. 22. Conduction-band–edge energy, kinetic energy, and velocity in a short field-effect structure, where electric field (the gradient of the band edge) changes rapidly. The solid lines are for GaAs, the short dashed lines for InP, and the long dashed lines for $Ga_{1-x}In_xAs$.

riers takes place in a time scale of at most 100 fs, and hence the carriers can still transit distances corresponding to this, ≈300 Å, before their energy begins to decrease. InP has a higher scattering rate and higher effective mass at lower energies than either GaAs or $Ga_{1-x}In_xAs$, and hence its velocities are lower toward the source end. However, at higher carrier energies, its velocity overshoot is higher than that of GaAs, because GaAs exhibits a larger scattering rate corresponding to the intervalley transfer. $Ga_{1-x}In_xAs$ has the most favored overshoot characteristics because of its lower scattering rate and large intervalley transfer threshold.

7. RECOMBINATION

The interactions between carriers of opposite types are particularly important in compound semiconductors because most of them are direct–band-gap semiconductors. Additional interactions between electrons and holes through levels within the band gap, due to traps, are also important in compound semiconductors. So, several forms of excitation processes can result either in the recombination of electron and hole pairs or in their generation. Radiative recombination, usually involving band-to-band transitions, with the energy released as photons is the basis of light-emitting diodes and lasers. Radiative recombination can also occur sometimes through discrete levels—e.g., in the Zn–O center combination in GaP. This is a unique example where the Zn–O pair forms an isoelectronic recombination center by replacing a Ga–P pair with an energy level in the band gap. Recombination can also occur without the involvement of radiative processes, such as through traps, through other processes where momentum conservation involves phonons, as well as through Auger processes where excess energy of annihilation or creation of an electron–hole pair is associated with a third carrier. All these processes are important under different conditions in devices.

In order to understand the statistics of these processes and their significance, we use the principle of detailed balance that states that, for a system in thermal equilibrium, the rate of any process and that of its inverse balance each other. The adjective "detailed" emphasizes that this balance occurs in all details of the process. Consider radiative recombination in thermal equilibrium. Detailed balance requires that, in thermal equilibrium, the same fraction of electron–hole recombination is radiative as is the electron–hole pair generation, and with the same spectral characteristics as the radiation that occurs with electron–hole pair recombination.

7.1 Defects and Hall–Shockley–Read Recombination

Traps, because they can exist in more than one charge state, have the ability to act as recombination-generation centers. This ability comes about because these centers can either capture an electron and/or a hole from the conduction band or emit them to the bands. Thus, electrons and holes can recombine or lead to a generation process at these traps. The statistics for this were first analyzed by Hall and by Shockley and Read, and it is known as Hall–Shockley–Read (HSR) recombination.

We will consider a donorlike trap, of density N_T, in an n-type crystal. Being a donorlike trap, it exists either in a positively charged state, and we consider a donor that exists only as a singly ionized donor of density N_T^+, or in an un-ionized state with density N_T^0. The singly ionized state of the trap exists in the band gap at an energy E_T. The net rate of change of excess density can be written as

$$-\frac{dp'}{dt} = -\frac{dn'}{dt} = (np - n_i^2) \times \left\{\left[n + N_c \exp\left(-\frac{E_c - \xi_f}{kT}\right)\right]\frac{1}{c_h N_T} + \left[p + N_v \exp\left(-\frac{E_T - E_v}{kT}\right)\right]\frac{1}{c_e N_T}\right\}^{-1}. \tag{28}$$

This expression relates the net rate of change in carrier concentration with time, i.e., $\mathcal{U} = \mathcal{G} - \mathcal{R} = dp'/dt = dn'/dt$ where $\mathcal{U}$ is the rate of change in carrier concentration and $\mathcal{G}$ and $\mathcal{R}$ are the generation and recombination rates.

Usually, one defines parameters τ_{p0} and τ_{n0} as

$$\tau_{p0} = 1/c_h N_T$$
$$\tau_{n0} = 1/c_e N_T, \qquad (29)$$

and hence the minority-carrier lifetime for the n-type semiconductor, for a donorlike trap with a single state, is

$$\begin{aligned}\tau_p &= -\frac{p'}{dp'/dt}\\ &= \frac{p'}{np - n_i^2}\left\{\left[n + N_c \exp\left(-\frac{E_c - \xi_f}{kT}\right)\right]\tau_{p0}\right.\\ &\quad \left. + \left[p + N_v \exp\left(-\frac{E_T - E_v}{kT}\right)\right]\tau_{n0}\right\}, \qquad (30)\end{aligned}$$

and majority-carrier lifetime is

$$\begin{aligned}\tau_n &= -\frac{n'}{dn'/dt}\\ &= \frac{n'}{np - n_i^2}\left\{\left[n + N_c \exp\left(-\frac{E_c - \xi_f}{kT}\right)\right]\tau_{p0}\right.\\ &\quad \left. + \left[p + N_v \exp\left(-\frac{E_T - E_v}{kT}\right)\right]\tau_{n0}\right\}. \qquad (31)\end{aligned}$$

These equations assume a single-level donorlike trap in the limit of Maxwell–Boltzmann approximation. Traps can be acceptorlike; the equations have the same form as above. A considerably more complex situation occurs when a trap has multiple levels in the band; e.g., transition metals with incomplete d orbitals usually have several levels associated with different states of ionization. Chromium, e.g., has three levels within the band gap of GaAs, and it is among the impurities used to obtain semi-insulating GaAs. Statistics in such situations can be quite complicated, because the presence of one state excludes the other. Figure 23 shows a plot of electron lifetime as a function of acceptor doping for GaAs, showing regions where HSR and Auger recombination, which we discuss next, dominate.

7.2 Auger Recombination

Electrons and holes can recombine, without the HSR-type interaction at a trap or the involvement of photon emission, by releasing the excess energy to another carrier. Such processes involving more than two carriers are usually referred to as Auger processes. Impact ionization, discussed earlier, is actually one example of an Auger process, a generation process that involves the creation of an additional electron–hole pair as a result of the excess energy of a hot carrier.

The number of Auger recombination processes is large. Auger transitions may take place through band-to-band, band–to–shallow-levels, or shallow-levels–to–shallow-levels transitions involving excitons, etc. In all these cases the excess energy released in the transition is picked up by a mobile carrier. Because of the involvement of carriers, shallow levels, and transition across much of the band gap, the process becomes stronger both

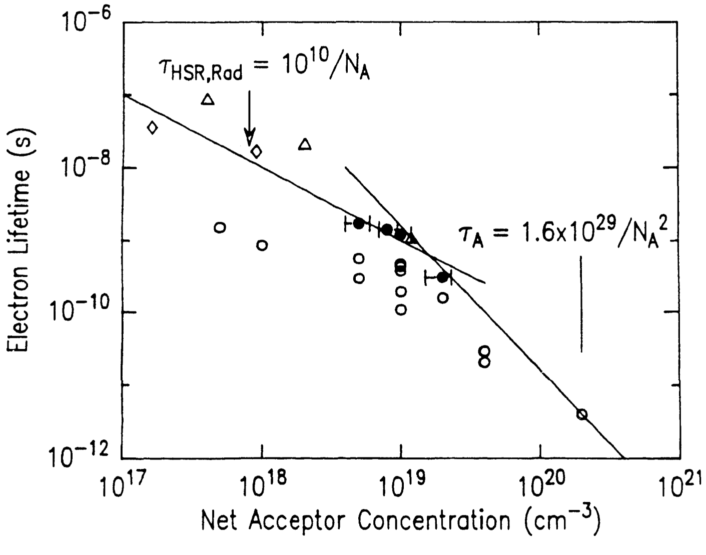

FIG. 23. Lifetime as a function of acceptor concentration in GaAs.

with increases in doping and by reduction in band gap. In small–band-gap materials, it is also strongly temperature dependent. Electrons are also more easily likely to get energetic since they usually have a lighter mass. Light holes are also likely to pick up excess energy; however, most holes exist as heavy holes because of the larger density of states for the heavy-hole band. Band structure is therefore very central to how and which Auger recombination processes are important, as it is for its inverse, the Auger generation process.

The net Auger recombination rate due to electrons and holes is given by

$$\begin{aligned}\mathcal{U}_{\mathrm{A}} &= \mathcal{R}_{\mathrm{A}e} - \mathcal{G}_{\mathrm{A}e} + \mathcal{R}_{\mathrm{A}h} - \mathcal{G}_{\mathrm{A}h}\\ &= (\gamma_n n + \gamma_p p)(np - n_i^2). \qquad (32)\end{aligned}$$

Under low-level injection, these expressions reduce to a form where the net recombination rate is proportional to the excess carrier concentration. For *p*-type material, such as the base of a bipolar transistor, the net recombination rate is

$$\mathcal{U}_{\mathrm{A}} = \gamma_p N_{\mathrm{A}}^2 n', \qquad (33)$$

where n' is the excess carrier concentration. Using this, in a way similar to that of HSR recombination, an Auger lifetime can be defined whose magnitude is

$$\tau_{\mathrm{A}} = 1/\gamma_n N_{\mathrm{A}}^2. \qquad (34)$$

For most semiconductors, where the band-to-band Auger processes dominate, in the limit of heavy doping, the lifetime does appear to follow an inverse dependence with the square root of the doping.

8. TUNNELING

Tunneling is the important example of the wave-particle duality. The de Broglie wavelength *vis-à-vis* the device dimension characterizes an estimate of the importance of the wave aspect. At device dimensions smaller than or comparable to the de Broglie wavelength, it is more appropriate to discuss electrons in crystals as a wave, and the classical picture that assumes a negligible time scale of scattering, drift, and diffusion averaged from a number of scattering events, etc., is increasingly poor. Tunneling, e.g., cannot be explained in the particle description; it requires the wave description. The de Broglie wavelength is related to the effective mass, which because it is generally small in the direct–band-gap compound semiconductors leads to large de Broglie wavelength, and increasing wave effects, tunneling being the most prominent of them.

Tunneling is a manifestation of the wave nature of electrons. If an overlap between the states in the valence band and the conduction band is large, then electrons can tunnel from the valence band to the conduction band in the junction. One may treat the forbidden gap as a potential barrier of the classical barrier-tunneling problem. Zener tunneling is a classical example of this and occurs in a variety of tunnel diodes, and it is a source of breakdown in *p-n* junctions. Zener breakdown is a form of internal emission that takes place by tunneling.

Zener tunneling will be substantial whenever the potential barrier height is small, the potential barrier width is short, or the effective mass is small. Any of these allow for a large overlap of the electron and hole wave functions. For a given band structure, the potential barrier height and width, related as the electric field, are conducive to tunneling at large dopings. Small–band-gap semiconductors have intrinsically a smaller potential barrier, leading to larger Zener tunneling probabilities, and it may take place even at low doping. The process thus has an effect on breakdown voltages both for highly doped semiconductors and for small–band-gap semiconductors. The breakdown voltages can be obtained by determination, again, of the transmission function and the source function, the two characterizing the probability of tunneling and the availability of carriers for tunneling.

Like our discussion of avalanche breakdown, Zener processes may also involve phonon processes. In the classical problem of tunneling through a barrier we ignore it; i.e., we consider only energy-conserving elastic processes. Phonon-assisted tunneling, an inelastic process, however, can be quite strong. During such a process, a longitudinal or transverse optical or acoustic phonon is emitted or absorbed in order to conserve momentum. In fact, for indirect semiconductors, this is the dominant form of tunneling because it can accommodate the large differences in momenta between the bottom of the conduction band and the top of the valence band.

9. HEAVY DOPING EFFECTS

In the presence of a large number of electrons and holes, the strong carrier–carrier interaction leads to a reduction in the band gap. Thus, under heavy doping, the band gap decreases. Figure 24 summarizes ΔE_g at 300 K for GaAs derived from measurements on transistor structures. A parametric fit to these data, optimized for higher dopings, is given by the relation

$$\Delta E_g = 2.00 \times 10^{-11} N_A^{0.50}, \tag{35}$$

where the unit of ΔE_g is eV and that of N_A is cm^{-3}. The magnitude of this shrinkage is quite substantial. At an acceptor doping of 1 × 10^{19} cm^{-3}, this shrinkage is ≈0.07 eV, nearly 5% of the band gap of GaAs. It is obviously of substantial significance in highly doped base regions of bipolar transistors as well as source and drain regions of field-effect transistors. Figure 24 also includes the effective band-gap shrinkage from a theoretical calculation. We emphasize that none of these is an actual band parameter; they are quantitative fits.

10. PIEZOELECTRIC EFFECTS

An issue unique to compound semiconductors is related to their ionic character and noncentrosymmetry. Stress-induced polarization results in polarization charge, whose effect shows up significantly through electrostatic effects in the operation of the device. The polarization can be viewed to result in additional charge because

$$\nabla \cdot (\epsilon_s \mathscr{E} + \mathscr{P}) = \rho, \tag{36}$$

where $\mathscr{P}$ is the polarization vector and ρ is the charge density.

The polarization results in additional charge and hence device-parameter variation because of stress in the material. This stress can occur in devices because of different expansion coefficients of the materials employed; it may result from the dielectric films employed on semiconductor surfaces or the metalizations of the gate and Ohmic contacts that are in proximity to the device active region.

When a compressive or tensile stress occurs, the angle between bonds for the atoms away from the center increases or decreases,

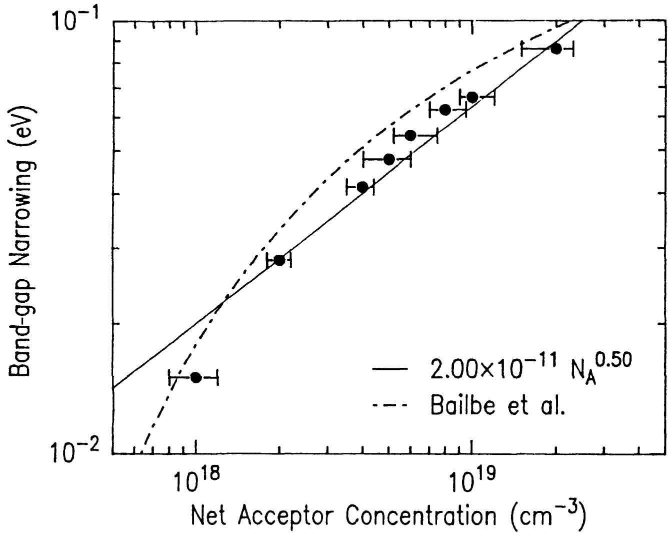

FIG. 24. Band-gap change with doping in GaAs at 300 K.

resulting in a net displacement. Consequently, there is a net dipole moment in the crystal. This polarization, caused by stress, can be determined in general by tensor analysis as

$$\mathcal{P} = \boldsymbol{d}:\boldsymbol{\sigma}, \tag{37}$$

where $\boldsymbol{d}$ is the piezoelectric tensor and $\boldsymbol{\sigma}$ is the stress tensor.

There is an additional requirement for this polarization to lead to a cumulative charge effect and hence a macroscopic variation with position. This is that the crystal be noncentrosymmetric. Noncentrosymmetry causes the polarization of the unit cells to cumulate because the polarizations are additive and have the same direction. Centrosymmetric crystals do not cumulate the polarizations of unit cells since they oppose in alternate unit cells.

Most of the elements of $\boldsymbol{d}$, except six, are zero for compound semiconductors and have identical value. This results in a polarization vector

$$\begin{bmatrix} \mathcal{P}_x \\ \mathcal{P}_y \\ \mathcal{P}_z \end{bmatrix} = d \begin{bmatrix} \sigma_{xy} \\ \frac{1}{2}(\sigma_{xx} - \sigma_{zz}) \\ \sigma_{yz} \end{bmatrix}, \tag{38}$$

where d is the magnitude of the equal nonzero components, the x subscript denotes $\langle 01\bar{1}\rangle$ direction, the y subscript denotes $\langle \bar{1}00\rangle$ direction, and the z subscript denotes $\langle 011\rangle$ direction. The strongest effect arises from the surface dielectric and metal films in close vicinity of the active region of the devices. In a metal-gate–controlled transistor, stress occurs because of both the gate film and the dielectric films used on the surface. The presence of polarization charge can cause electric-parameter (e.g., threshold voltage) change of both signs. In addition, depending on the orientation of the gate or dielectric lines, the polarization results are different. Si_3N_4, one of the dielectric films employed, is usually under compressive stress, while amorphous SiO_2, another dielectric employed, is usually under tensile stress. The resulting polarization from the dielectric film is therefore of the opposite sign, and hence threshold voltage variation is opposite.

In transistors made on (100) surfaces with gates in the $\langle 011\rangle$ or $\langle 0\bar{1}1\rangle$ direction, the piezocharge is at either its maximum or the minimum under the gate. Near the tail of the doping density, its effect can be substantial because of reduction in dopant charge. It can, therefore, significantly disturb the low-current behavior of devices—i.e., it affects parameters such as threshold voltage and conductance at low biases with the effect gate-orientation dependent. The problem of changes in threshold voltage in devices that are oriented differently on the same wafer places a significant design constraint in their use.

The threshold shift also becomes more pronounced with a decrease in dimensions, i.e., shortening of the gate length. Shortening of gate length naturally leads to an increase in the stress underneath the gate due to stress from dielectric films. This gate-length effect is in addition to the effect of short-channel phenomenon.

GLOSSARY

Acoustic Phonons: The low-energy mode of lattice vibrations.

Avalanche Process: The cascade process where an energetic carrier loses some of its excess energy by creating an electron–hole pair.

Auger Recombination: The inverse process of avalanche process.

Band Gap: The difference between minimum of conduction-band edge and maximum of valence-band edge.

Discontinuity: The change in conduction- or valence-band edge at an abrupt transition from one composition to another composition of a crystal.

Drift Mobility: The constant of proportionality relating the drift velocity with local electric field.

Hall Mobility: A mobility derived from the Hall voltage that develops transverse to the current flow in the presence of a magnetic field.

Heterojunction: The junction formed between two dissimilar materials.

Majority Carrier: The carrier type that is in majority in thermal equilibrium in a semiconductor.

Minority Carrier: The carrier type that is in minority in thermal equilibrium in a semiconductor.

Optical Phonons: The high-energy modes of lattice vibrations.

Overshoot or Off-Equilibrium Effect: The condition when the local velocity of a carrier is larger than the velocity obtained from the product of mobility and electric field.

Piezoelectric Effect: The effects arising from accumulation of polarization charge in presence of strain in a noncentrosymmetric crystal.

Scattering: The interaction between a carrier and its surroundings.

Work Cited

Wolfe, C. M., Stillman, G. E., Lindley, W. T. (1970), *J. Appl. Phys.* **41,** 3088–3091.

Further Reading

Adachi, S. (1985), "GaAs, AlAs, and $Al_xGa_{1-x}As$: Material Parameters for Use in Research and Device Applications," *J. Appl. Phys.* **58,** R1–R29.

Blakemore, J. S. (1982), "Semiconducting and Other Major Properties of Gallium Arsenide," *J. Appl. Phys.* **53,** R123–R181.

Burstein, E., Lundqvist, S. (Eds.) (1969), *Tunneling Phenomena in Solids,* New York: Plenum.

Conwell, E. M. (1978), "High Field Transport in Semiconductors," in: F. Seitz, D. Turnbull, H. Ehrenreich (Eds.), *Solid State Physics,* Vol. 9, New York: Academic.

Duke, C. B. (1969), "Tunneling in Solids," in: F. Seitz, D. Turnbull, H. Ehrenreich (Eds.), *Solid State Physics,* Vol. 10, New York: Academic.

Harrison, W. A. (1969), *Solid State Theory,* New York: McGraw-Hill.

Hess, K. (1982), "Aspects of High-field Transport in Semiconductor Heterolayers and Semiconductor Devices," in: *Advances in Electronic and Electron Physics,* Vol. 59, New York: Academic.

Jaros, M. (1985), "Electronic Properties of Semiconductor Alloy Systems," *Rep. Prog. Phys.* **48,** 1091–1154.

Nag, B. R. (1980), *Electron Transport in Compound Semiconductors,* Berlin: Springer.

Paul, W. (Ed.) (1982), *Handbook on Semiconductors,* Amsterdam: North-Holland.

Pearsall, T. P. (Ed.) (1990), *Strained Layer Superlattices: Physics,* Semiconductors and Semimetals Vol. 32, San Diego: Academic.

Pearsall, T. P. (1991), *Strained Layer Superlattices: Materials Science and Technology,* Semiconductors and Semimetals Vol. 33, San Diego: Academic.

Ridley, B. K. (1988), *Quantum Processes in Semiconductors,* Oxford: Clarendon.

Sze, S. M. (1981), *Physics of Semiconductor Devices,* New York: Wiley.

Tiwari, S. (1992), *Compound Semiconductor Device Physics,* New York: Wiley.

Wolfe, C. M., Holonyak, N., Jr., Stillman, G. E. (1989), *Physical Properties of Semiconductors,* Englewood Cliffs, NJ: Prentice-Hall.

SEMICONDUCTORS, COMPOUND—MATERIAL PROPERTIES

V. Swaminathan, *Bell Laboratories, Lucent Technologies, North Andover, Massachusetts, U.S.A.*

P. Capper, *GEC-Marconi Infra-Red Ltd., Southampton, United Kingdom*

Introduction 335
1. **III–V Compounds** 336
1.1 Binary Compounds 336
1.1.1 Bonding and Crystal Structure 336
1.1.1.1 Fractional Ionic Character in Bonding 336
1.1.1.2 Brillouin Zones 337
1.1.2 Energy-Band Structure 338
1.1.2.1 Effective Mass 339
1.1.3 Impurities and Defects 340
1.1.4 Diffusion 343
1.1.5 Optical Properties 343
1.1.6 Mechanical Properties 345
1.1.6.1 Deformation Potential 346
1.1.7 Heterojunctions 346
1.2 Alloys 347
1.2.1 Energy-Band Structure of Alloys 347
1.2.2 Material Parameters of Alloys 349
1.2.2.1 Lattice Constant 350
1.2.2.2 Band Gap 351
1.2.2.3 Quarternary Alloys 351
1.2.2.4 Heteroepitaxial Strain and Band Gap 351
1.2.2.5 Effective Mass in Ternary and Quarternary Alloys 353
1.2.2.6 Phonons in Ternary and Quarternary Alloys 354
1.2.2.7 Refractive Index 355
1.2.2.8 Thermal Resistivity 356
1.2.2.9 Thermal Expansion Coefficient 356
2. **II–VI Compounds** 357
2.1 Binary Compounds 357
2.1.1 Bonding and Crystal Structure 357
2.1.2 Energy-Band Structure 358
2.1.3 Impurities and Defects 360
2.1.4 Diffusion 363
2.1.5 Optical Properties 364
2.1.6 Mechanical Properties 364
2.1.7 Heterojunctions 365
2.2 Alloys 365
2.2.1 Energy-Band Structure of Alloys 365
2.2.2 Material Properties of Alloys 366
3. **Conclusion** 368
Glossary 369
Works Cited 369
Further Reading 371

INTRODUCTION

Semiconductors are classified into two broad categories as elemental semiconductors and compound semiconductors. Si and Ge are examples of elemental semiconductors. The compounds are grouped into families depending on the chemical nature of the constituents. III–V compounds represent semiconductor compounds made from elements of group III and group V of the periodic table (e.g., GaAs, InP, GaP, InAs, InSb, GaN). Similarly, II–VI compounds are made from elements of group II and group VI (e.g., CdTe, ZnTe, ZnSe, CdSe, CdS).

While the early research was devoted to the physical properties of the III–V compounds, the emphasis was quickly shifted to technology and applications. Phenomenal progress has been made in the preparation and processing of these compounds, and some of them (e.g., GaAs, InP) are used to make technologically important electronic and photonic devices. In this latter aspect, they have an advantage over Si which, being an indirect-band-gap semiconductor, is not an efficient photonic material. Besides, they (e.g., GaAs, InP) are attractive for high-speed devices since the electron saturation velocity in GaAs is 1.5 times that of Si. The low-field

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

electron mobility in GaAs or InP is also higher than that in Si. The advantages of high speed and low power dissipation of GaAs devices appeal to the communications industry. When compared to Si, III–V compounds have higher radiation hardness and are capable of operating over a wider temperature range from −196 to 300 °C. Because of this, GaAs devices are of special interest for space and military applications.

The major use of GaAs at present is in the area of microwave devices, high-speed digital integrated circuits, and as substrates for epitaxial layer growth to fabricate photonic and electronic devices. For InP, the use is almost exclusively as substrates for growing lattice-matched epitaxial films of alloy semiconductors such as GaInAs and GaInAsP. These ternary and quaternary compounds are the materials of choice for making light sources and detectors for the present-day fiber-optic communication systems. Among other III–V compounds, GaP has been the material of choice for visible-light–emitting diodes. InSb is used for infrared detectors. Both InSb and GaSb are used as substrate material for epitaxial growth of Sb-based compounds. Recently, interest in wide–band-gap (>2 eV) III–V nitrides has increased on account of progress in their growth by epitaxy.

Improvements in both bulk and particularly epitaxial growth techniques in recent years, combined with a deeper understanding of defects and contacting/passivation issues, have led to a resurgence in II–VI compound studies and use. Electrical, optical, and magnetic properties have been utilized to make a wide range of emission and detector devices. Wide-gap II–VI compounds, such as ZnSe, are used in lasers and LEDs, while narrow-gap compounds, such as CdTe and HgCdTe, are used in solar cells, optical devices, room-temperature γ- and x-ray detectors, and infrared detectors and emitters. It is probable that HgCdTe is the third most widely studied semiconductor system, after Si and GaAs, because of the effort expended over the last 30 years or so, with most support coming from military sources. Zinc oxide is used as an acoustic-wave, electro-optic, and acousto-optic material and in some solar-cell structures. IIa–VIa compounds, such as CaS, are indirect-gap semiconductors and are used as electroluminescent materials and as phosphors, when doped with rare-earth elements (e.g., Eu, Ce). Optical modulators and isolators are made from diluted magnetic semiconductors (*q.v.*) (DMS) based on II–VI compounds, such as CdMnTe and HgFeSe. The attraction of these DMS materials lies in the ability to tune—by varying the composition, the energy gap, and the lattice parameter—the purely magnetic properties, including paramagnetic, spin-glass, and antiferromagnetic properties, and those properties, such as giant Faraday rotation of visible and near-infrared light and giant negative magnetoresistance in the semiconductor–semimetal transition region, that are unique to DMS materials.

The following sections summarize the relevant data for the more important binary and ternary III–V and II–VI compounds, which have been distilled from the references. There is insufficient space in this article to provide detailed discussion of the various properties and effects, particularly in view of the ranges given in the literature for many of the parameters. However, a significant bibliography is given at the end, and the reader is referred to books in that list.

1. III–V COMPOUNDS

1.1 Binary Compounds

1.1.1 Bonding and Crystal Structure Most III–V compound semiconductors crystallize in the cubic zinc-blende (sphalerite) structure. The wide–band-gap nitrides (e.g., AlN, GaN, InN) crystallize in the hexagonal wurtzite structure. In the tetrahedral arrangement of atoms in the sphalerite structure, each group-III atom is bonded to four group-V atoms in the tetrahedron and vice versa. The cubic unit cell of a sphalerite structure is shown in Fig. 1. There are four molecules of the compound *AB* per unit cell.

1.1.1.1 Fractional Ionic Character in Bonding. Although the bonding in the sphalerite structure consists of sp^3 hybridized orbitals, there is some charge transfer between the two types of atoms, giving rise to a partial ionic character to the bonding, unlike in diamond cubic semiconductors. If f_i and f_h denote the fractions, respectively, of ionic or heteropolar character and covalent or homopolar character in the bond, then

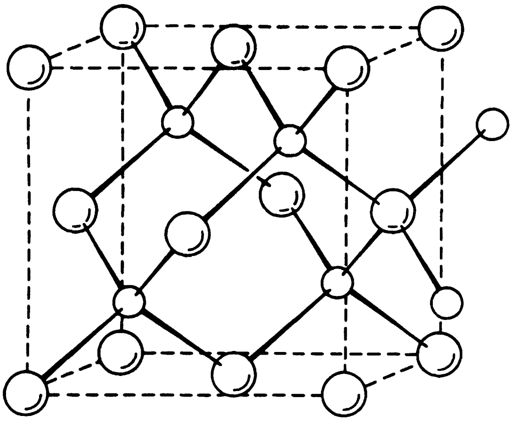

FIG. 1. The cubic unit cell of a sphalerite structure.

Table 1. Fractional ionic character of bonds and cohesive energy of III–V semiconductors (Phillips, 1973).

Crystal	Fractional ionic character, f_i	Cohesive energy[a] (kcal/mole)
BAs	0.002	
BP	0.006	
AlSb	0.250	
BN	0.256	−293.9
GaSb	0.261	−118.3
AlAs	0.274	−157.3
AlP	0.307	−169.5
GaAs	0.310	−135.4
InSb	0.321	−108.5
GaP	0.327	−152.4
InAs	0.357	−125.6
InP	0.421	−132.9
AlN	0.449	
GaN	0.500	

---

[a] Gibbs free energy of sublimation into neutral atoms at STP.

the sum $f_i + f_h$ equals unity. In elemental semiconductors $f_h = 1$ and $f_i = 0$. A semi-empirical theory of fractional ionic or covalent character of the bond has been developed by Phillips (1973), which involves estimating the ionic contribution to the energy gap of the AB semiconductor. Table 1 lists the f_i values for III–V compounds (Phillips, 1973). Also listed in Table 1 are the cohesive energies, denoted by the Gibbs free energies of atomization at standard temperature and pressure (STP), of the respective compounds.

1.1.1.2 Brillouin Zones. The electronic and vibronic states in crystals are best described in reciprocal space or **k** space. The smallest unit cell in reciprocal space is called the first Brillouin zone. While $(a/2)[110]$, $(a/2)[101]$, and $(a/2)[011]$ are the primitive translation vectors of the face-centered cubic (fcc) lattice, $(2\pi/a)[11\bar{1}]$, $(2\pi/a)[1\bar{1}1]$, and $(2\pi/a)[\bar{1}11]$ are the primitive translation vectors in the reciprocal lattice. Incidentally, the latter are the primitive translation vectors of a body-centered cubic (bcc) lattice, so that the bcc lattice is the reciprocal lattice of the fcc lattice. The volume of the primitive cell in the reciprocal fcc lattice is $4(2\pi/a)^3$.

The first Brillouin zone for the cubic semiconductors is the truncated octahedron shown in Fig. 2. The principal symmetry points and lines in **k** space are labeled, using the standard nomenclature. The Cartesian coordinates of these symmetry points are $\Gamma = (0,0,0)$; $X = (2,0,0)$; $L = (1,1,1)$; $\Lambda = (x,x,x)$; $\Delta = (2x,0,0)$; $W = (2,1,0)$; $K = U = (3/2,3/2,0)$; $\Sigma = (3/2x, 3/2x,0)$; where x denotes a number between 0 and 1 and the coordinates are given in units of π/a where a is the lattice constant.

Table 2 lists the lattice constant, density, melting point, Debye temperature, coefficient of thermal expansion, and thermal conductivity of the binary compounds. For the last two parameters, the values at or close to 300 K are given (Madelung *et al.*, 1982, 1984).

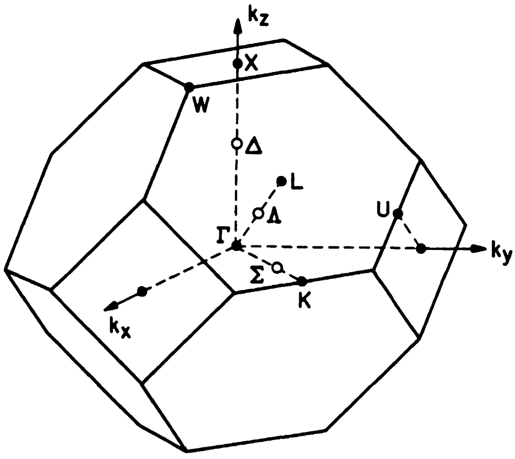

FIG. 2. The first Brillouin zone for the cubic semiconductor showing the symmetry points and axes (from Zallen, 1982).

Table 2. Lattice constant, density, melting point, Debye temperature, coefficient of thermal expansion, and thermal conductivity of III–V compounds.

Compound	Lattice constant (Å)	Density (gm cm^{-3})	Melting point (K)	Debye Temperature^{a,b} (K)	Coefficient of thermal expansion^{c} (10^{-6}/°C)	Thermal conductivity^{a,b,c} (W cm^{-1} K^{-1})
AlN	3.11^{e} 4.98^{f}	3.255	3273		5.27 (⊥) 4.15 (‖)	
AlP	5.467	2.4	2823	588	4.5	0.9
AlAs	5.66	3.7	2013	417	4.9	0.8
AlSb	6.136	4.26	1338	292	4.0	0.57
GaN	3.16^{e} 5.125^{f}	6.095	1973	600	3.17 (⊥) 5.59 (‖)	1.3
GaP	5.4512	4.138	1740	456	4.5	0.77
GaAs	5.6532	5.3161	1513	344	6.86	0.46
GaSb	6.0959	5.6137	985	266	7.75	0.39
InN	3.5446^{e} 5.7034^{f}	6.89	1373		3.9 (⊥) 2.95 (‖)	
InP	5.8687	4.81	1335	321	4.75	0.68
InAs	6.0583	5.667	1215	249	4.52	0.273
InSb	6.4794	5.7747	800	203	5.37	0.166

---

[a] From Holland (1966).

[b] From Madelung *et al.* (1982).

[c] Values near or at 300 K.

[e] Lattice parameter = a.

[f] Lattice parameter = c.

1.1.2 Energy-Band Structure The **k**·**p** method is used to calculate the band structure in the vicinity of the most important symmetry points in the Brillouin zone in great detail (Kane, 1982). Semiconductors with tetrahedral coordination have valence-band maxima at the Γ point, and many have conduction-band minima also at the Γ point (e.g., GaAs, InP). The representation of the valence band is Γ_{15}. When electron spin is included, there are altogether six valence-band states. The six states are further split by spin–orbit interaction into a fourfold degenerate $J = 3/2$ and a twofold degenerate $J = 1/2$ states, the $J = 1/2$ band being the lowest band. The $J = 3/2$ states are further characterized by the azimuthal quantum number of J relative to the k axis, $m_j = 3/2$ or $1/2$. They are labeled as the heavy-hole (twofold degenerate, $J = 3/2$, $m_j = 3/2$) band and light-hole (twofold degenerate, $J = 3/2$, $m_j = 1/2$) band. The $J = 1/2$ band is called the spin–orbit split-off valence band. The value of the spin–orbit splitting for III–V semiconductors is given in Table 3 (Phillips, 1973). The conduction band is derived from s-like orbitals and denoted by the representation Γ_1. It is very much like the mirror image of the spin–orbit split-off valence band.

The calculated energy-band structure of GaAs is shown in Fig. 3 (Cohen and Chelikowsky, 1988). The various symmetry points in the Brillouin zone are also indicated. The lowest conduction-band minimum Γ_1 occurs at $k = 0$ with higher minima occurring in the $\langle 100 \rangle$ (Δ) and $\langle 111 \rangle$ (Λ) directions. The L and X minima are respectively 0.31 and 0.52 eV above the Γ_1 minimum (Cohen and Chelikowsky, 1988). The valence band consists of heavy-hole and light-hole bands degenerate at $k = 0$ and a spin–orbit-split band at 0.34 eV lower energy. The various critical-point energies are also indicated in Fig. 3. These critical points satisfy the condition $\nabla_k E(k) = 0$.

The parameters pertinent to the electronic

Table 3. Spin–orbit splittings in III–V compounds (Phillips, 1973).

Compound	Δ_0 (eV)
GaP	0.127
GaAs	0.34
GaSb	0.80
InP	0.11
InAs	0.38
InSb	0.82

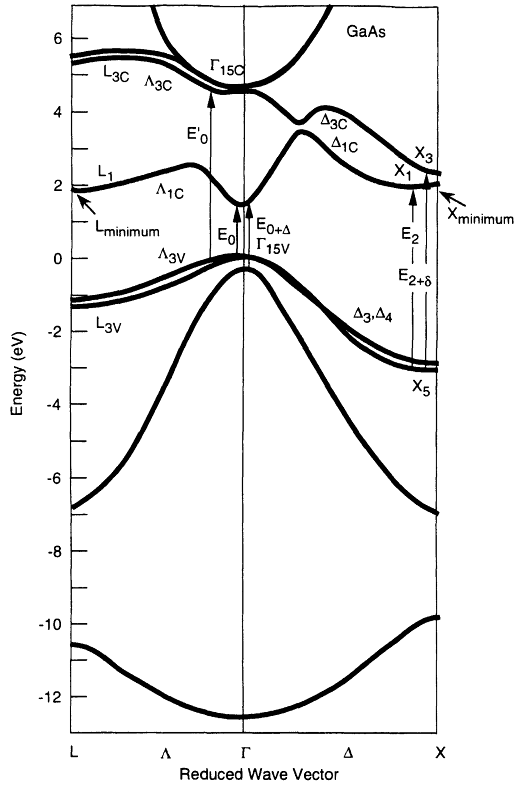

FIG. 3. Calculated energy band structure of GaAs in two principal directions in the Brillouin zone. The various critical point energies are indicated. (From Cohen and Chelikowsky, 1988.)

band structure are summarized in Tables 4 and 5. The direct and indirect energy gaps, their temperature dependences, and the pressure dependence of the minimum energy gap are given in Table 4. For AlP, AlAs, AlSb, and GaP, the conduction-band minimum is located at the Δ axes near the Brillouin-zone boundary. For AlAs the k value of the conduction band minimum is $0.087k_{\max}$ from the X point, where $k_{\max} = 2\pi/a$, a being the lattice constant. For GaP the corresponding value is $0.043k_{\max}$ from the X point.

The temperature dependence of the band gap is generally expressed by (Varshni, 1967)

$$E_g(T) = E_g(0) - \alpha T^2/(T + \beta), \qquad (1)$$

where $E_g(0)$ is the energy gap at 0 K, and α and β are constants, with β having a value close to the Debye temperature. Near room temperature and above, $E_g(T)$ varies linearly with temperature. The values of the linear temperature coefficient of the energy gap are given in Table 4. Table 6 gives the α and β values for obtaining E_g below room temperature (Casey and Panish, 1978).

1.1.2.1 Effective Mass. Table 7 lists the effective masses at conduction- and valence-band edges. Semiconductors that have conduction-band minimum at $\Gamma = 0$ are characterized by one mass only. Those that have minimum near the X point are characterized

Table 4. Energy-band structure of III–V compounds.

Compound	Direct energy gap (eV) $\Gamma_{15} - \Gamma_1$	Indirect energy gap (eV) $\Gamma_{15} - X_1$	Indirect energy gap (eV) $\Gamma_{15} - L_1$	$\frac{dE_d}{dT}$ (10^{-4} eV K^{-1})[a]	$\frac{dE_{\text{ind}}}{dT}$ (10^{-4} eV K^{-1})[b]	$\frac{dE}{dP}$ (10^{-6} eV/bar)[c]
AlN	6.2					
AlP	3.62 (77 K)	2.45			−3.6	
AlAs	3.14	2.14[d]		−5.2	−4.0	
AlSb	2.22	1.63[d]			−3.5	−1.5
GaN	3.44			−6.7		4.2
GaP	2.78	2.268[d]		−4.5	−5.2	−1.6
GaAs	1.424	1.804	1.81 (110 K)	−3.9	−2.4[e]	12.0
GaSb	0.7	1.25 (10 K)	0.81	−3.7		14.7
InN	2.09					
InP	1.34	2.04	1.74	−2.9	−3.7[e]	8.8
InAs	0.356			−3.5		10.6
InSb	0.18			−2.8		15.9

---

[a] Temperature dependence of direct energy gap.
[b] Temperature dependence of indirect energy gap.
[c] Pressure dependence of minimum energy gap.
[d] Minimum situated at the Δ axes near the boundary: $k = (0.903, 0, 0)$ for AlAs; $k = (0.95, 0, 0)$ for GaP.
[e] Temperature dependence of the energy separation between Γ and X minima.

Table 5. Critical-point energies (eV) of III–V compounds. $E_0 = \Gamma_{15v} - \Gamma_{15c}$, $E_1 = \Lambda_{3V} - \Lambda_{1C}$, $E_1 + \Delta = \Lambda_{4V} - \Lambda_{1C}$, $E_2 = X_{5V} - X_{1C}$, $E_2 + \delta = X_{5V} - X_{3C}$.

Compound	E_0	E_1	$E_1 + \Delta$	E_2	$E_2 + \delta$
AlAs	4.34			4.54	4.89
AlSb	3.7 (77 K)	2.78	3.18		
GaN	8.5	6.25		7.65	
GaP	4.8 (80 K)	3.7 (80 K)	3.9 (80 K)		
GaAs	4.49 (4.2 K)	2.9	3.17		
GaSb	3.2 (80 K)				
InP	4.8 (77 K)	3.15	3.3	5.04	5.6
InAs	4.5 (77 K)	2.5	2.75	4.72	5.3
InSb	3.16	1.88	2.38	4.08	4.6 (110 K)

by longitudinal and transverse effective masses, m_l and m_t. Because of lack of inversion symmetry, the degenerate bands at the X point are split into a higher-lying X_3 band and a lower-lying X_1 band. For the indirect–band-gap semiconductors, the density-of-states effective mass is given by $\nu^{2/3}(m_l m_t^2)^{1/3}$, where ν is the number of equivalent conduction-band minima. When the minimum occurs at the X point, $\nu = 3$. The valence band is characterized by three effective masses corresponding to the heavy-hole, light-hole, and spin–orbit bands. The density-of-states effective mass for holes is given by $(m_{hh}^{3/2} + m_{lh}^{3/2})^{2/3}$.

1.1.3 Impurities and Defects Crystalline imperfections play an important role in determining the material quality of III–V compound semiconductors; they include point defects, dislocations, stacking faults, grain boundaries, and interfaces. Point defects can be native defects, such as vacancies and interstitials of the group-III or group-V atoms, as well as foreign atoms. The latter are generally intentionally added to the material to obtain the desired electrical properties. They can also be present as background impurities in the crystals because of contamination from the crystal growth systems. Dislocations and stacking faults are examples of line defects and two-dimensional defects, respectively. They can be present in the as-grown material or can be introduced during device processing steps, such as diffusion and ion implantation. Their presence is, in general, harmful to the devices. Interfaces between semiconductors, as in multilayer structures, or between semiconductors and dielectric films that are deposited on them as part of device processing can be a source or sink for defects. As a result, defect-free interfaces are desired.

Table 6. Temperature dependence of the minimum energy gap. Parameters α and β as defined in Eq. (1) are obtained from Casey and Panish (1978).

Compound	$E_g(0)$ (eV)	α (10^{-4} eV K^{-1})	β (K)
AlP	2.52	3.18	588
AlAs	2.239	6.0	408
AlSb	1.687	4.97	213
GaP	2.338	5.771	372
GaAs	1.519	5.405	204
GaSb	0.810	3.78	94
InP	1.421	3.63	162
InAs	0.420	2.50	75
InSb	0.236	2.99	140
$Ga_{0.47}In_{0.53}As$/InP	0.822	4.5	327[a,b]

[a] From Pearsall *et al.* (1983).

[b] The linear temperature coefficient obtained from absorption measurement is -3.48×10^{-4} eV/K (Zielinski *et al.*, 1986).

In III–V components, the electrical properties are controlled to a large extent by foreign impurities, which are either intentionally added or present as contaminants. Impurities give rise to two kinds of energy levels, which are often denoted as shallow and deep energy levels. Shallow impurities have energy levels that are placed on the order of 0.1 eV or less with respect to the band edge. On the other hand, deep impurities have energy levels toward the middle of the energy gap. The deep impurities are also sometimes referred to as traps.

Shallow impurities are further classified as donors or acceptors depending upon whether they produce electron (n-type) or hole (p-type) conductivity. For impurities that substitute for the cation or anion site in

Table 7. Effective-mass parameters of III–V compounds.

Compound	Conduction band (units of m_0)			Valence band (units of m_0)		
	m_l	m_t	m_e	m_{hh}	m_{lh}	m_{hso}
AlP				0.63	0.2	0.29
AlAs	1.5	0.19	0.79[a]	0.76	0.15	0.24
AlSb	1.64	0.23	0.92[a]	0.94	0.14	0.29
GaN			0.27	0.8		
GaP	7.25	0.313	1.86[b]	0.54	0.16	0.24
GaAs			0.067	0.49	0.08	0.15
GaSb			0.044	0.34	0.044	0.13
InN			0.12	0.5	0.17	
InP			0.075	0.56	0.12	0.12
InAs			0.024	0.37	0.025	0.14
InSb			0.014	0.39	0.016	0.47

[a] For three equivalent minima, m_e is given by $3^{2/3}(m_l m_t^2)^{1/3}$.

[b] Because of the presence of a camel's-back conduction-band structure and extremely high nonparabolicity, there is considerable scatter in the effective-mass values reported in the literature.

the III–V compound, donor or acceptor action is realized depending on the chemical valence difference between the host and the impurity. If the valence difference is greater than zero, donors are created, and if it is less than zero, acceptors are created. If there is no valence difference, then the impurity is called isovalent—eg., N substituting for P in GaP.

For the III–V compounds, group-VI impurities such as S, Se, and Te substitute for group-V atoms and behave as shallow donors. Likewise, group-II impurities such as Be, Mg, Zn, and Cd substitute for group-III atoms and give rise to shallow acceptors. Group-IV impurities like Si, Ge, and Sn, on the other hand, show amphoteric behavior. That is, a group-IV impurity on the anion site behaves as an acceptor, while one on the cation site behaves as a donor.

The energy level associated with a donor (acceptor) represents the binding energy of an electron (hole) to the impurity in the semiconductor host. For shallow impurities, the binding energy can be obtained using the effective-mass approximation (Kohn, 1957) and is given by

$$E_n = \frac{1}{n^2}\left(\frac{m^*}{m_0}\right)\left(\frac{1}{\varepsilon^2}\right)\mathfrak{R}, \tag{2}$$

where $\mathfrak{R}$ is 13.6 eV, the ionization energy of the free hydrogen atom, ε is the dielectric constant of the semiconductor, m^* is the effective mass, and m_0 is the free-electron mass. For electrons (donors) m^* is the electron effective mass, and the energies are measured from the bottom of the conduction band. For holes (acceptors) m^* is the hole effective mass and the energies are measured from the top of the valence band. The Bohr radius corresponding to the bound states is given by

$$a = a_{\mathrm{H}}\frac{\varepsilon}{(m^*/m_0)}, \tag{3}$$

where a_{H} is the Bohr radius of the hydrogen atom..

For direct-gap III–V compounds, such as GaAs and InP with nondegenerate conduction band and isotropic electron effective mass, the effective-mass approximation is a good description for the binding of electrons to donors. The binding energies and Bohr radii of donors in GaAs and InP are respectively $E_D = 5$ meV, $a_D = 102$ Å, and $E_D = 7$ meV, $a_D = 84$ Å, with $\varepsilon = 12.9$ and $m_e = 0.067m_0$ for GaAs and $\varepsilon = 12.6$, $m_e = 0.08m_0$ for InP. For acceptors, the situation is somewhat complicated because of the complex valence-band structure (Kohn, 1957), but qualitatively it can be expected that the larger effective mass of the holes compared to that of the electrons gives rise to larger acceptor binding energy and smaller Bohr radius compared to that of donor.

The effective-mass theory predicts the same energy levels for donors and acceptors irrespective of their chemical nature. But in reality, the binding energies vary depending on the impurity, and the variation is significant, especially for the acceptors. The difference between the effective-mass binding energy and the experimental value is referred to as a "chemical shift" (Pantelides, 1975). Phillips (1970) proposed a dielectric model to explain the chemical shifts in terms of the differences in electronegativity between the substitutional impurity and the host atom that it has replaced. When the electronegativity difference is large, the deviation of the binding energy from the effective-mass value is large.

The energy levels of donors and acceptors are usually obtained by observing the transitions of electrons from the impurity state to one of the band continua or to another impurity state or, conversely, from one of the band continua to the impurity state. The ground state and/or the excited states of the impurity may also be involved in such transitions. A large number of electrical, optical, and magnetic resonance techniques have been employed to characterize defects and obtain information of their binding energies and microscopic structure. Many chemical analysis techniques have been used to assess impurity concentrations. Often, special analysis tools with sensitivity in the parts per billion range are required. Secondary-ion mass spectrometry, otherwise known as SIMS, is one of the popular chemical analysis techniques used in the compound semiconductor industry.

Impurities from columns other than II, IV, and VI of the periodic table can also be incorporated in a III–V crystal, and they may have energy level(s) deep within the band gap. The electronic wave functions centered on these impurities are, in general, more localized than is the case for simple donors and acceptors. Further, the effective-mass approximation is not a good description of the binding energy of the electronic particle. 3*d* transition metal atoms such as Ti, Cr, Mn, Fe, Co, Ni, and Cu are examples of deep impurities in III–V compounds. They are used to obtain semi-insulating material with carrier concentrations and resistivities similar to that of the intrinsic semiconductor. Also, these impurities can act as efficient recombination centers and limit the efficiency of light-emitting devices. In some cases, the 3*d* elements have been found to give doubly ionizable donors or acceptors. A junction technique known as deep-level transient spectroscopy, DLTS, is widely used to probe deep impurities and defects and obtain information on ionization energies and emission/capture cross sections for electrons and holes.

There are six possible native defects in an AB compound: two types of vacancies (V_A,V_B), interstitials (A_i,B_i), and misplaced atoms (A_B,B_A). Since the A/B site ratio required by the crystal structure has to be maintained, the native defects occur in sets containing at least two types in a stoichiometric compound (Kroger, 1974). These sets are the Schottky disorder involving V_A and V_B, interstitial disorder involving A_i and B_i, and antistructure disorder involving A_B and B_A. There is also the possibility of combinations of these three disorder types. A disorder involving vacancies and interstitials—for example, $V_A + A_i$ or $V_B + B_i$—is called Frenkel disorder.

Native defects play an important role in determining the electronic properties of III–V compounds. Their concentrations are affected by the conditions of preparation of the crystal (e.g., partial pressure of constituents) that represent the equilibrium between the solid phase and the external liquid or gas phase. Incorporation of substitutional impurities, which include many of the donor and acceptor atoms, depends on the concentration of native defects. Native defects themselves can exist in neutral or charged form, and as charged species they sometimes directly control the conductivity type of the semiconductor. A well-known example is the EL2 deep level in GaAs, which gives rise to the semi-insulating characteristics of undoped material (Martin and Makram-Ebied, 1986). Semi-insulating GaAs required for fabricating integrated circuits such as field-effect transistors can be obtained by doping with Cr. However, Cr redistributes or outdiffuses during annealing steps in processing. Hence, undoped semi-insulating GaAs is very attractive for fabricating integrated circuits, and this is achieved by the deep donor EL2 that has been associated with an As_{Ga} defect. Other examples where native defects directly affect the electronic properties are GaN and GaSb. In undoped form the former is *n* type, and the latter is *p* type, believed to be caused by V_N and V_{Ga}, respectively.

1.1.4 Diffusion Diffusion data in III–V compounds can be separated into two sections, one dealing with the diffusion of the main constituents of the crystal, self-diffusion, and another with the diffusion of impurities. There are two self-diffusion coefficients in a III–V compound: one corresponding to the group-III atom and another to the group-V atom. The self-diffusion coefficients are functions of the deviation of the stoichiometry of the compound, which is controlled by the partial pressure of the constituents during diffusion. Self-diffusion coefficients are generally determined using radioactive tracers. Impurity diffusion occurs under conditions of concentration gradients and is described by a chemical diffusion coefficient. In addition to the component pressures during diffusion, other factors such as ionization state of the impurity, electric fields, mechanical stresses, complex interactions of impurities with other point defects, etc., affect impurity diffusion coefficients. For this reason, data for impurity diffusion coefficients reported in the literature tend to vary over a wide range. Interdiffusion of constituent elements taking place in the presence of a concentration gradient is of importance in superlattice structures (e.g., GaAs/AlAs superlattice).

Diffusion constants are customarily described by the Arrhenius expression, $D = D_0 \exp(-Q/kT)$, where D is the diffusion constant, Q is the activation energy for diffusion, k is the Boltzmann constant, and T is the absolute temperature. In compound semiconductors, D also depends on the component pressure. Table 8 summarizes the self-diffusion data for some of the III–V compounds, and Table 9 summarizes the diffusion coefficients for some common impurities (Madelung *et al.*, 1984). The wide range of values for Q and D_0 for a given constituent of the compound as in self-diffusion coefficients or for a given impurity in a host is caused, to a large extent, by the different diffusion conditions such as component pressure, distribution of defects in the crystal existing prior to the diffusion, etc. For this reason the data in Tables 8 and 9 should be used only to obtain order-of-magnitude estimates of diffusion coefficients.

Table 8. Self-diffusion constants in III–V compounds.

Compound	Element	D_0 ($cm^2\ s^{-1}$)	Q (eV)
GaAs	Ga	1.0×10^7	5.6
	Ga	4×10^{-5}	2.6
	As		10.2
	As	5.5×10^{-4}	3.0
GaSb	Ga	3.2×10^3	3.15
	Sb	3.4×10^4	3.45
	Sb	8.7×10^2	1.13
InP	In	1×10^5	3.85
	P	7×10^{10}	5.65
InAs	In	6×10^5	4.0
	As	3.0×10^7	4.45
InSb	In	1.8×10^{-9}	0.28
	In	5.0×10^{-2}	1.82
	In	1.8×10^{13}	4.3
	Sb	1.4×10^{-6}	0.75
	Sb	5.0×10^{-2}	1.94
	Sb	3.1×10^{13}	4.3

1.1.5 Optical Properties Table 10 lists the optical properties—refractive index near the band-gap wavelength, its temperature dependence, the static and high-frequency dielectric constants, and the frequencies of the zone-center LO and TO phonons. The optical absorption of solids can be described either by the complex index of refraction $n + ik$ or by the complex dielectric function $\epsilon + i\epsilon_2$, where n, k, ϵ_1, and ϵ_2 are all real functions of the frequency ω, and

$$\epsilon_1 = n^2 - k^2, \quad \epsilon_2 = 2nk. \tag{4}$$

The functions $\epsilon_1(\omega)$ and $\epsilon_2(\omega)$ are connected by the Kramers–Kronig relations. The absorption coefficient α is related to k by

$$\alpha = 2\omega k/c, \tag{5}$$

where k, the imaginary part of the refractive index, is the extinction coefficient. Values of k and n over a wide range of photon energies for many III–V semiconductors have been compiled by Seraphin and Bennett (1967).

The static and high-frequency dielectric constants are related by

$$\epsilon(0) = \epsilon(\infty) + 4\pi \mathrm{N} e_T^2/\omega_{TO}^2 M, \tag{6}$$

where M is the reduced mass, e_T is the effective charge, ω_{TO} is the TO phonon frequency, and N is the number of ion pairs. Note from Table 10 that both $\epsilon(0)$ and $\epsilon(\infty)$ show an inverse dependence on E_g, decreasing with increasing E_g.

Table 9. Chemical diffusion coefficients for common impurities in III–V compounds.

Impurity	Host	D_0 (cm^2 s^{-1})	Q (eV)	Remarks
Au	GaP	8	2.5 ± 0.3	From A face
		20	2.4 ± 0.2	From B face
	GaAs	1.0×10^{-3}–29	1–2.6	
	InP	1.3×10^{-5}	0.48	
	InAs	5.8×10^{-4}	0.65	
	InSb	7×10^{-4}	0.32	
Be	GaAs	7.3×10^{-6}	1.2	
Cd	GaAs	5.0×10^{-2}	2.43	
	GaSb	1.5×10^{-6}	0.72	
	InP	1.8	1.9	
	InAs	7.4×10^{-4}	1.15	
	InSb	$(1.0\text{–}13) \times 10^{-5}$	1.1–11.2	
Cu	GaAs	3×10^{-2}	0.53	
	InP	3.8×10^{-3}	0.69	
	InAs	3.6×10^{-3}	0.52	
	InSb	9.0×10^{-4}	1.08	
Mn	GaAs	6.5×10^{-1}	2.49	
O	GaAs	2.0×10^{-3}	1.1	
S	GaAs	$(1.6\text{–}12) \times 10^{-5}$	1.63–1.86	
	GaP	3.2×10^{3}	4.7	
	InAs	6.78	2.2	
Se	GaAs	3×10^{3}	4.16	
	InAs	12.6	2.20	
	InSb	1.6	1.87	
Si	GaAs	0.11	2.5	
Zn	GaAs			Complex profiles; extensive studies
	GaSb	4×10^{-2}	1.6	
	InP			Complex profiles; extensive studies
	InAs			Complex profiles; extensive studies

Table 10. Optical properties of III–V compounds.

Compound	Energy gap E_g (eV)	Refractive index n near E_g	$\frac{1}{n}\frac{dn}{dT}$ (K^{-1})	Dielectric constant[a]		Wave number of Raman phonon (cm^{-1})	
				$\epsilon(0)$	$\epsilon(\infty)$	LO	TO
AlN	6.28			9.14[c]	4.84[c]		
AlP	2.45	3.03	3.5×10^{-5}	9.8	7.54		
AlAs	2.14	3.18	4.6×10^{-5}	10.06	8.16	404.1	360.9
AlSb	1.63	3.4	3.5×10^{-5}[b]	12.04	10.24	339.6	318.8
GaN	3.44	2.6	2.6×10^{-5}[c]	12.2	5.2		
GaP	2.27	3.45	2.5×10^{-5}[b]	11.1	9.08	403.0	367.3
GaAs	1.42	3.65	4.5×10^{-5}	12.91	10.9	291.9	268.6
GaSb	0.70	3.82	8.2×10^{-5}	15.69	14.44	233.0	224.0
InN	2.09	3.12			9.3[c]		
InP	1.34	3.41	2.7×10^{-5}	12.61	9.61	345.0	303.7
InAs	0.36	3.52	6.5×10^{-5}[b]	15.15	12.25	238.6	217.3
InSb	0.18	4.00	1.2×10^{-4}	17.7	15.68	190.8	179.8

---

[a] In units of ϵ_0, the permittivity of free space 8.85×10^{-14} F cm^{-1}.
[b] From Ghosh *et al.* (1986).
[c] From Madelung *et al.* (1982).

In the study of transport and optical properties of polar semiconductors, the coupling between the electron and LO phonons becomes important. The electron–LO-phonon interaction is given by the well-known Fröhlich coupling constant (De Vreese, 1972):

$$\alpha_F = \frac{e^2}{4\pi\epsilon_0}\left(\frac{m^*}{2\hbar^3\omega_{LO}}\right)^{1/2}\left(\frac{1}{\epsilon(\infty)} - \frac{1}{\epsilon(0)}\right), \tag{7}$$

where ω_{LO} is the LO phonon frequency. Equation (7) shows that α_F depends on the ionic polarizability of the crystal, which is related to $\epsilon(\infty)$ and $\epsilon(0)$.

1.1.6 Mechanical Properties The elastic compliances S_{ij} in units of 10^{-12} cm² dyne^{-1} are listed in Table 11. For a cubic crystal, there are three independent elastic constants: S_{11}, S_{12}, and S_{44}. The elastic compliance relates strain to stress,

$$\epsilon_i = \sum_j S_{ij}\sigma_j, \tag{8}$$

where the strain ϵ_i and stress σ_j are the components of second-rank tensors and S_{ij}'s are components of a fourth-rank tensor. Equation (8) can be expressed in its covariant form

$$\sigma_i = \sum_j C_{ij}\epsilon_j, \tag{9}$$

where C_{ij}'s are the elastic stiffness constants. The compliance tensor is the reciprocal of the stiffness tensor. For a cubic crystal, C_{ij}'s are related to S_{ij}'s according to (Nye, 1957)

$$\begin{aligned} C_{11} &= (S_{11} + S_{12})/S, \\ C_{12} &= -S_{12}/S, \\ C_{44} &= 1/S_{44}, \end{aligned} \tag{10}$$

where

$$S = (S_{11} - S_{12})(S_{11} + 2S_{12}).$$

From the point of practical utility, the elastic constant of interest is the Young's modulus E. Another elastic property of interest is the Poisson's ratio ν. In cubic crystals both E and ν are anisotropic. The expressions for E and ν for an arbitrary crystallographic direction in a cubic crystal are given by Brantley (1973):

$$\begin{aligned} \frac{1}{E} &= S_{11} - 2(S_{11} - S_{12} - \tfrac{1}{2}S_{44})(l_1^2l_2^2 + l_2^2l_3^2 + l_1^2l_3^2), \\ \nu &= -\frac{S_{12} + (S_{11} - S_{12} - \tfrac{1}{2}S_{44})(l_1^2m_1^2 + l_2^2m_2^2 + l_3^2m_3^2)}{S_{11} - 2(S_{11} - S_{12} - \tfrac{1}{2}S_{44})(l_1^2l_2^2 + l_2^2l_3^2 + l_1^2l_3^2)}, \end{aligned} \tag{11}$$

where l is the longitudinal stress axis, m is an orthogonal direction to l, and the l's and m's are the direction cosines for l and m with respect to the cube axes. For the ⟨100⟩ axes

$$\begin{aligned} 1/E &= S_{11} \\ \nu &= -S_{12}/S_{11}. \end{aligned} \tag{12}$$

From Eq. (11) it follows that the factor $2(S_{11} - S_{12})/S_{44}$ is a measure of deviation from isotropy. For the biaxial plane stress conditions associated with thin films, often encountered in the case of heterostructures, longitudinal stresses and strains parallel to the film substrate are related to each other by $E/(1 - \nu)$. The expressions for this composite elastic constant are easily obtained from Eq. (11).

The bulk modulus for zinc-blende–type crystals is given by

Table 11. Elastic compliances of III–V compounds, in units of 10^{-12} cm² dyne^{-1}.

Compound	S_{11}	S_{12}	S_{44}
AlP	1.090	−0.350	1.630
AlAs	1.070	−0.320	1.840
AlSb	1.696	−0.562	2.453
GaP	0.973	−0.298	1.419
GaAs	1.176	−0.365	1.684
GaSb	1.582	−0.495	2.314
InP	1.650	−0.594	2.170
InAs	1.945	−0.685	2.525
InSb	2.443	−0.863	3.311

Table 12. Deformation-potential constants of III–V compounds.

Compound	a (eV) Direct gap	a (eV) Indirect gap	b (eV)	d (eV)
AlP				
AlAs				
AlSb	−5.9	2.2	−1.35	−4.3
GaP	−9.6		−1.65	−4.5
GaAs	−9.77		−1.70	−4.55
GaSb	−8.28		−2.0	−4.7
InP	−6.35		−2.0	−5.0
InAs	−6.0		−1.8	−3.6
InSb	−7.7		−2.0	−4.9

$$B = 1/3(S_{11} + 2S_{12}). \tag{13}$$

For $\langle 100 \rangle$ axes $B = E/3(1 - 2\nu)$. Knowing the density ρ and C_{ij}, one can obtain the sound velocity v using the relation

$$v = (C_{ij}/\rho)^{1/2}. \tag{14}$$

For a cubic crystal, using $C_{ij} = C_{11}$ in Eq. (14) gives the velocity of a longitudinal wave propagating along a cube axis, and using $C_{ij} = C_{44}$ gives the velocity of a shear wave.

1.1.6.1 Deformation Potential. Application of an external stress has a profound effect on the band structure. Hydrostatic stress causes a shift of the energy states. Uniaxial stress lowers the symmetry of the lattice and splits some degenerate states. In cubic semiconductors, the fourfold degeneracy of the valence band is lifted, resulting in, respectively, $J = 3/2$, $m_j = 3/2$ and $J = 3/2$, $m_j = 1/2$ heavy-hole and light-hole bands. The shift of E_g with hydrostatic stress and the splitting of the valence band δE_v are given by the deformation-potential constants. The values of these constants are listed in Table 12. Variation of direct energy gap with hydrostatic pressure p is described by a constant a given by

$$a = -\tfrac{1}{3}(C_{11} + 2C_{12})\frac{dE_g}{dp}. \tag{15}$$

The splitting of the valence band under [100] stress is described by the deformation-potential constant b and the splitting under [111] stress by constant d, according to the relations

$$\begin{aligned} \delta E_v &= b(S_{11} - S_{12})\sigma \\ \delta E_v &= (d/2\sqrt{3})S_{44}\sigma, \end{aligned} \tag{16}$$

where σ is the applied uniaxial stress.

1.1.7 Heterojunctions Heterostructures have attracted much attention because of scientific and technological interest (Milnes and Feucht, 1982; Kroemer, 1985). The primary reason for the consideration of heterostructures for semiconductor device applications is the abrupt change in the energy-band structure at the heterointerface. This leads to discontinuities, or offsets, in the conduction- and valence-band edges. Such offsets act as potential steps in addition to pure electrostatic potential variation that may be present and as such can be used to control the flow and distribution of electrons and holes in device structures.

Heterojunctions can be classified by their band lineup. The types of band offsets that occur at abrupt semiconductor heterojunctions are illustrated in Fig. 4. In the strad-

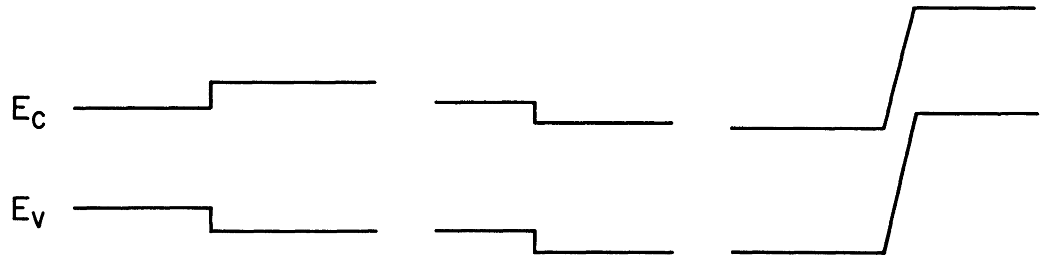

(a) (b) (c)

FIG. 4. Schematic band-gap diagrams showing the different heterojunction band lineups: (a) straddling lineup, (b) staggered lineup, and (c) broken-gap lineup.

dling band lineup shown in Fig. 4(a), the band offsets in both the conduction and valence bands act as potential barriers and keep electrons and holes in the smaller–band-gap semiconductors. Heterostructures that have straddling lineups are called type I heterostructures. Notable examples of these are the GaAs/Al$_x$Ga$_{1-x}$As and the InP/GaInAsP heterojunctions. Some heterojunctions have staggered [Fig. 4(b)] or broken-gap [Fig. 4(c)] band lineups. These are called type II heterostructures. The heterojunction InAs/GaSb is of the broken-gap type in which the top of the valence band of GaSb is higher than the bottom of conduction band of InAs. In this situation, the electrons in the top region of the valence band of GaSb can freely enter the conduction band of InAs when the two semiconductors are brought into contact with each other.

The parameter of utmost importance in heterostructures for device applications is the band-edge offset. The band-edge offsets in the conduction and valence bands are related to the difference in the band gaps of the two components of the heterojunction as

$$\Delta E_V + \Delta E_C = \Delta E_g. \tag{17}$$

Thus, if one of the offsets is calculated, the other is easily obtained since ΔE_g is reasonably well known. The calculation of ΔE_V, for example, requires knowledge of the absolute energies of the valence-band edges of the semiconductors. Further, when the two semiconductors are brought into contact with each other, the charge redistribution in the interface region and the associated electronic interface dipole effects may be significant. Given these two facts, ΔE_V may be written as

$$\Delta E_V = (E_{V0}^A - E_{V0}^B) - \delta E_V, \tag{18}$$

where E_{V0}^A and E_{V0}^B are the energies with respect to the vacuum level of the valence-band maxima of the two semiconductors A and B, respectively, and δE_V is the energy shift in the band edges at the interface caused by the formation of the interface dipole. Apart from charge redistribution at the interface, the interface dipole effect may also be affected by the presence of localized interface defects, particularly in mismatched heterostructures (Tejedor and Flores, 1978). It is also important to include in the calculations of band offset the effect of strain caused by the lattice mismatch of the constituents of the heterostructure.

As shown schematically in Fig. 5, ΔE_C is equal to the difference in electron affinities of A and B (Anderson, 1962),

$$\Delta E_C = \chi_A - \chi_B. \tag{19}$$

ΔE_V is obtained from Eq. (17), neglecting the contribution of δE_V. This is the electron affinity rule (EAR) to calculate the band offset. From Fig. 5 the difference in the ionization energies of A and B is related to the valence-band offset (Harrison, 1977):

$$\Delta E_V = \phi_B - \phi_A = E_{V0}^B - E_{V0}^A. \tag{20}$$

Note that δE_V is neglected in determining ΔE_V. Equation (20) represents the Harrison atomic-orbital (HAO) model to obtain the band offset. Table 13 lists the valence-band offsets for some of the heterostructures in III–V compounds.

1.2 Alloys

1.2.1 Energy-Band Structure of Alloys

III–V alloy semiconductors have assumed an important role in the semiconductor industry because of their utility in many electronic and photonic devices such as heterostructure transistors, light-emitting diodes, and lasers. Let us consider two III–V compounds AC and BC that form a cation alloy $A_xB_{1-x}C$ (e.g., Ga$_x$In$_{1-x}$As). Similarly, compounds AC

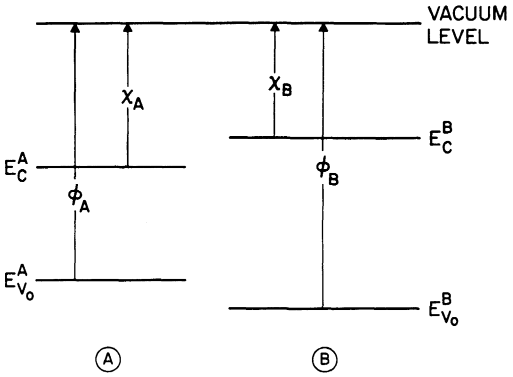

FIG. 5. Schematic band diagram showing the differences in electron affinity and ionization potential of semiconductors *A* and *B*.

Table 13. Calculated and experimental valence-band offsets (eV) for GaAs- and InP-based heterostructures[a].

Heterostructure	Experiment	EAR[b]	HAO[c]	Tersoff's model	Effective-dipole model[a]	Total-energy minimization
Ge/InP	0.64	0.90	0.64	0.58	0.67	
Ge/GaAs	0.25–0.65	0.70	0.41	0.52	0.51	
Si/InP	0.57	0.58	0.24	0.40	0.36	
Si/GaAs	0.05	0.38	0.03	0.34	0.22	
GaAs/AlAs	0.19–0.50	0.52	0.04	0.35	0.36	
GaAs/InAs	0.17	0.19	0.32	0.20	0.16	
GaAs/ZnSe	0.96–1.10	1.31	1.05	—	0.95	
InP/CdS	1.63	1.40	1.36	—	0.97	
InSb/InP	—	0.87			0.71	
InSb/GaAs	—	0.67			0.53	
InAs/InP		0.39			0.29	
GaAs/InP	—	0.20			0.13	
$Ga_{0.47}In_{0.53}As$/InP	0.346 ± 0.01[d] 0.39[e]					0.41
$Al_{0.48}In_{0.52}As$/InP	0.36 0.46[f]					0.25
$Ga_{0.47}In_{0.53}As$/$Al_{0.48}In_{0.52}As$	0.13, 0.20 0.21[f]					0.17

---

[a] Ruan and Ching (1987).
[b] Electron affinity rule model.
[c] Harrison atomic orbital model.
[d] Lang (1987).
[e] Forrest (1987).
[f] Hybertsen (1990a, 1990b).

and *AD* form an anion alloy AC_xD_{1-x} (e.g., $InAs_xP_{1-x}$). Mixed compounds of this type are called pseudobinary alloys. If the two compounds do not share a cation or an anion, one gets a quaternary alloy—for example, $Ga_xIn_{1-x}As_yP_{1-y}$.

It has been found experimentally that in many semiconductor alloys the direct energy gap ($\Gamma^v_{15} - \Gamma^c_1$) varies nonlinearly with composition. For an $A_xB_{1-x}C$ compound, the energy gap can be expressed in the form (Van Vechten and Bergstresser, 1970)

$$E_0(x) = a + bx + cx^2, \tag{21}$$

where c is called the bowing parameter and is four times the deviation of $E_0(x)$ from linearity at $x = 0.5$. The other two parameters, a and b, are determined by the values of E_0 observed in the pure binary compounds. The bowing parameter has been calculated using the dielectric method of band-structure calculation (Phillips and Van Vechten, 1969; Van Vechten, 1969). The total bowing parameter has been taken to be the sum of two terms: an intrinsic bowing and an extrinsic bowing. The latter is the result of the aperiodic crystal potential caused by disorder in the alloys and is essentially a short-range (within one unit cell) effect. The calculated values of these two bowing parameters and the experimental value pertaining to the minimum direct gap are given for some III–V compounds in Table 14 (Van Vechten and Bergstresser, 1970).

Figure 6 shows the lattice parameter versus energy gap at room temperature for various III–V compounds and their alloys (Ma-

Table 14. Comparison of theory and experiment for the bowing parameter pertaining to the minimum direct energy gap (Van Vechten and Bergstresser, 1970). c_i and c_e are the intrinsic (virtual crystal) and extrinsic (disorder) bowing parameters. $c_{cal} = c_i + c_e$.

Alloy	c_i (eV)	c_e (eV)	c_{cal} (eV)	c_{exp} (eV)
GaAs-P	0.21	0.09	0.30	0.21
InAs-P	0.15	0.08	0.23	0.20, 0.26
Ga-InSb	0.12	0.24	0.36	0.43
Ga-InAs	0.28	0.29	0.57	0.33, 0.56
Ga-AlAs	0.0	0.03	0.03	~0.2
Ga-InP	0.39	0.31	0.70	0.88

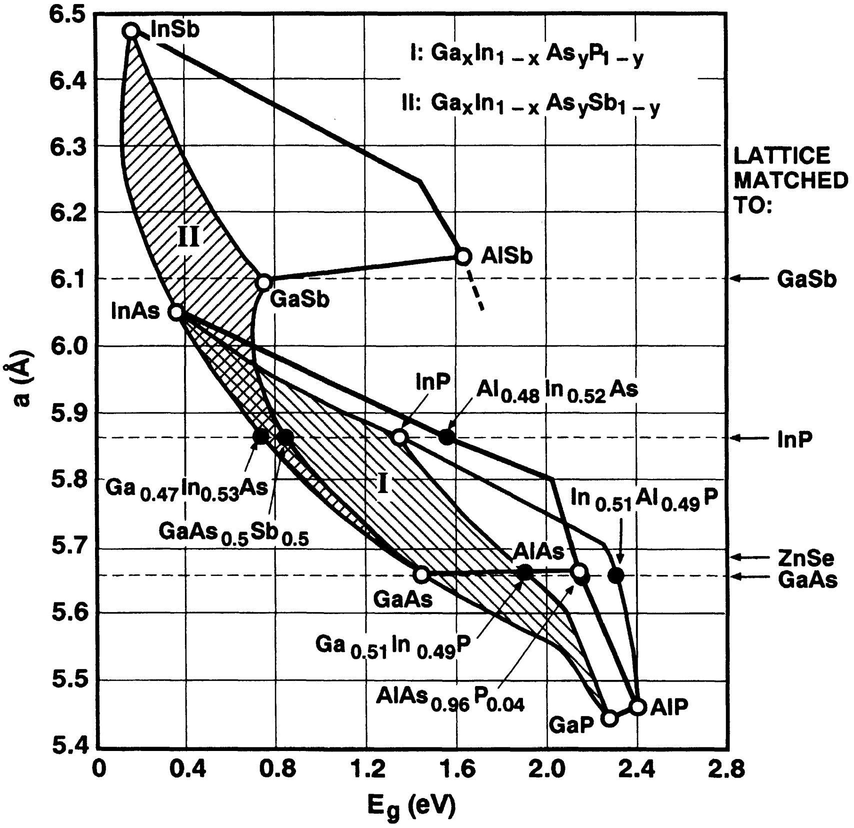

FIG. 6. Lattice parameter versus energy gap at room temperature for various III–V compounds and their alloys (from Madelung and Schulz, 1987). The lines connecting the binary end points represent the effect of the bowing parameter.

delung and Schulz, 1987). The composition of the lattice-matched ternary alloys and the corresponding substrates on which they are grown are indicated in Fig. 6. The availability of good-quality substrates dictates to a large extent which of the lattice-matched heterostructures would be of use for device structures. At present, technologically important III–V heterostructures are limited to those that are grown on GaAs and InP substrates.

The shaded region I in Fig. 6 represents the quaternary alloy $Ga_xIn_{1-x}As_yP_{1-y}$, which can be grown lattice matched on GaAs or InP substrates. When GaAs is used as the substrate, the lattice-matched GaInAsP alloy covers the energy range 1.42–1.91 eV, more or less the same energy range as that covered by the ternary alloy AlGaAs. The GaInAsP alloy grown on InP substrate spans the energy range 0.75–1.35 eV between the band gap of $Ga_{0.47}In_{0.53}As$ (0.75 eV) and that of InP (1.35 eV). The ternary alloy $Al_xIn_{1-x}As$ is lattice matched to InP at $x \approx 0.48$ and has a direct energy gap of 1.45 eV at room temperature. The $Al_xGa_yIn_{1-x-y}As$ alloy covers the spectral range 0.75–1.45 eV and is lattice matched to InP substrate for $x + y \approx 0.47$. This alloy, containing only one group-V atom, covers a wider spectral range than $Ga_xIn_{1-x}As_yP_{1-y}$ and may be advantageous for many heterostructure devices.

1.2.2 Material Parameters of Alloys

In deriving many physical parameters of the alloys, an interpolation scheme is generally adopted using the values of the related binary compounds. For a ternary compound $A_xB_{1-x}C$, the parameter P is derived from

$$P(A_xB_{1-x}C) = xP_{AC} + (1 - x)P_{BC}. \qquad (22)$$

Material parameters such as lattice constant vary linearly with composition as per Eq. (22). As noted in the last section, some parameters such as the energy gap deviate from linearity, and in that case, Eq. (21) is more appropriate.

The parameter of a quaternary compound such as $A_xB_{1-x}C_yD_{1-y}$ can be obtained from the respective values of the four binaries, AC, AD, BC, and BD:

$$P(A_xB_{1-x}C_yD_{1-y}) = xyP_{AC} + x(1-y)P_{AD} + (1-x)yP_{BC} + (1-x)(1-y)P_{BD}. \tag{23}$$

If the parameters for the ternary alloy $A_xB_{1-x}C$, $A_xB_{1-x}D$, AC_yD_{1-y}, and BC_yD_{1-y} are available, then Eq. (22) can be modified to give (Glisson *et al.*, 1978)

$$P(A_xB_{1-x}C_yD_{1-y}) = \frac{x(1-x)[(1-y)P_{ABD} + yP_{ABC}] + y(1-y)[xP_{ACD} + (1-x)P_{BCD}]}{x(1-x) + y(1-y)}. \tag{24}$$

The ternary parameters P_{ABD} and so on may include a quadratic dependence given by Eq. (21). This interpolation equation reduces to the average of the four ternary parameters at $x = y = 0.5$. In general, when specific experimental data are unavailable, most physical parameters of alloy semiconductors may be derived using a linear interpolation scheme.

1.2.2.1 Lattice Constant. The lattice constant of ternary III–V alloys generally varies linearly with composition (Vegard's law). This is illustrated in Fig. 7 for the $Ga_xIn_{1-x}As$ alloy (Hocking *et al.*, 1966). Lattice-matched composition for growth on InP substrates occurs at $x \approx 0.468$.

Vegard's law can also be assumed for the quaternary alloys. For the $Ga_xIn_{1-x}As_yP_{1-y}$ alloys, Vegard's law is given by

$$\begin{aligned} a(x,y) &= 0.1896y - 0.4175x + 0.0124xy + 5.8687 \\ &= a_{\mathrm{InP}}[1 + (\Delta a/a)_0], \end{aligned} \tag{25}$$

where $\Delta a/a$ is the relaxed lattice mismatch relative to the substrate. The quaternary alloys grown lattice matched on InP substrates can be considered part of a pseudobinary system between $Ga_{0.47}In_{0.53}As$ and InP. The lattice-matching relation between x and y can be expressed as

$$x \cong 0.47y \qquad (0 < y < 1.0)$$

or, using Eq. (23) more rigidly, as

$$x = \frac{0.1896y}{0.4175 - 0.012y} \qquad (0 < y < 1). \tag{26}$$

The quantity $(\Delta a/a)_0$ in Eq. (25) is related to the measured $(\Delta a/a)_\perp$ for the case of thin epitaxial layers (Hornstra and Bartels, 1978):

$$\left(\frac{\Delta a}{a}\right)_\perp = \frac{1+\nu}{1-\nu}\left(\frac{\Delta a}{a}\right)_0. \tag{27}$$

For {100} crystals according to Eq. (10), Eq. (27) reduces to

$$\left(\frac{\Delta a}{a}\right)_{100} = \frac{S_{11} - S_{12}}{S_{11} + S_{12}}\left(\frac{\Delta a}{a}\right)_0, \tag{28}$$

where S_{ij} is the elastic compliance (see Sec. 1.1.6). Using the S_{ij} values in Table 11, we obtain

$$(\Delta a/a)_{100} = 2.125(\Delta a/a)_0. \tag{29}$$

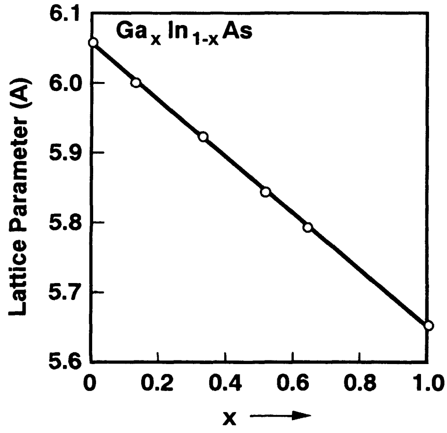

FIG. 7. Room-temperature lattice parameter versus composition for $Ga_xIn_{1-x}As$ (from Hocking *et al.*, 1966).

1.2.2.2 Band Gap. For the alloys, a parameter that is often of interest is the variation of E_g with composition. Besides the variation of the minimum energy gap, the variation of the energy separation between the different conduction-band minima is also of interest. The composition dependence of the energy gap is often represented by Eq. (21). Table 15 gives E_g versus composition for some important ternary alloy semiconductors (Madelung *et al.*, 1982, 1984; Madelung and Schulz, 1987).

1.2.2.3 Quaternary Alloys. For determining the band gap of the quaternary alloy, Moon *et al.* (1974) have used the following equation:

$$E_g(A_xB_{1-x}C_yD_{1-y}) = xE_{ACD} + (1-x)E_{BCD} - \Delta, \tag{30}$$

where Δ is the bowing-parameter term for the quaternary alloy and is given by

$$\Delta = x(1-x)[(1-y)c_{ABD} + yc_{ABC}] + y(1-y)[xc_{ACD} + (1-x)c_{BCD}], \tag{31}$$

where c_{ABD} and so on are the ternary bowing parameters of Eq. (21). For $x = y = 0.5$, Δ is $\frac{1}{8}$ of the sum of the four ternary bowing parameters, which is twice as large as the quaternary bowing parameter given by Eq. (24). A comparison of the two interpolation schemes given by Eqs. (24) and (30) with measured values of the band gap in $Ga_xIn_{1-x}As_yP_{1-y}$ alloys showed that both schemes give comparable errors (Glisson *et al.*, 1978). However, for calculation of the band gap, Eq. (30), which is the quaternary version of Eq. (21), is preferred since it has some theoretical basis. Using the band gaps and lattice constants of the binary compounds and Eqs. (22), (24), and (30), Glisson *et al.* (1978) have derived the lowest-energy band gap and lattice constant of several quaternary alloys of the III–V semiconductors at room temperature. Their calculations are reproduced in Figs. 8 and 9. The composition dependences of energy gaps for the quaternary compounds $Ga_xIn_{1-x}As_yP_{1-y}$ and $Al_xGa_yIn_{1-x-y}As$ are also given in Table 15.

1.2.2.4 Heteroepitaxial Strain and Band Gap. One of the aims in growing device-quality heterostructures is to reduce the lattice mismatch between the epitaxial layer and the substrate as much as possible. Since the layers are usually much thinner than the substrate, the strain caused by lattice mismatch can be accommodated elastically or plastically by the generation of dislocations when the thickness of the layers exceeds a

Table 15. Compositional dependence of the energy gap in III–V ternary and quaternary alloy semiconductors at 300 K.

Alloy	Direct energy gap E_Γ	Indirect energy gap E_X	Indirect energy gap E_L
$Al_xIn_{1-x}P$	$1.34 + 2.23x$	$2.24 + 0.18x$	
$Al_xGa_{1-x}As$	$1.424 + 1.247x (x < 0.45)$ $1.424 + 1.087x + 0.438x^2$	$1.905 + 0.10x + 0.16x^2$	$1.705 + 0.695x$
$Al_xIn_{1-x}As$	$0.36 + 2.35x + 0.24x^2$	$1.8 + 0.4x$	
$Al_xGa_{1-x}Sb$	$0.73 + 1.10x + 0.47x^2$	$1.05 + 0.56x$	
$Al_xIn_{1-x}Sb$	$0.172 + 1.621x + 0.43x^2$		
$Ga_xIn_{1-x}P$	$1.34 + 0.511x + 0.6043x^2$ $(0.49 < x < 0.55$, VPE layer)		
$Ga_xIn_{1-x}As$	$0.356 + 0.7x + 0.4x^2$		
$Ga_xIn_{1-x}Sb$	$0.172 + 0.165x + 0.413x^2$		
GaP_xAs_{1-x}	$1.424 + 1.172x + 0.186x^2$		
$GaAs_xSb_{1-x}$	$0.73 - 0.5x + 1.2x^2$		
InP_xAs_{1-x}	$0.356 + 0.675x + 0.32x^2$		
$InAs_xSb_{1-x}$	$0.18 - 0.41x + 0.58x^2$		
$Ga_xIn_{1-x}As_yP_{1-y}$	$1.35 + 0.668x - 1.068y + 0.758x^2 + 0.078y^2 - 0.069xy - 0.322x^2y + 0.03xy^2$		
$Al_xGa_yIn_{1-x-y}As$	$0.36 + 2.093x + 0.629y + 0.577x^2 + 0.436y^2 + 1.013xy - 2.0xy(1 - x - y)$		

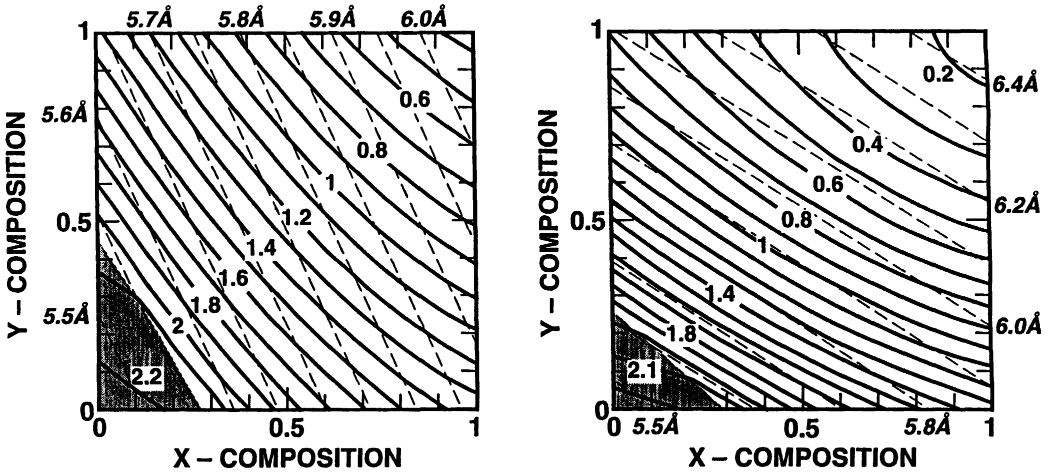

(a) (b)

FIG. 8. Energy-band gap (solid curves) and lattice-constant contours (dashed curves) at 300 K for (a) $Ga_{1-x}In_xP_{1-y}As_y$ and (b) $Ga_{1-x}In_xP_{1-y}Sb_y$ alloys. The shaded region shows the compositional range over which the material has an indirect band gap. (From Glisson *et al.*, 1978.)

certain critical value (Asai and Oe, 1983). Both types of strain have a profound influence on device characteristics.

In the elastic regime, for coherent growth (in-plane lattice parameter of the epitaxial layer equals that of the substrate) the lattice mismatch is accommodated by a tetragonal distortion of the layer (Hornstra and Bartels, 1978). For this case, the lattice parameter of the unstrained epitaxial layer is given by Eq. (27). When $(\Delta a/a)_\perp$ is positive (the layer has a larger lattice parameter than the sub-

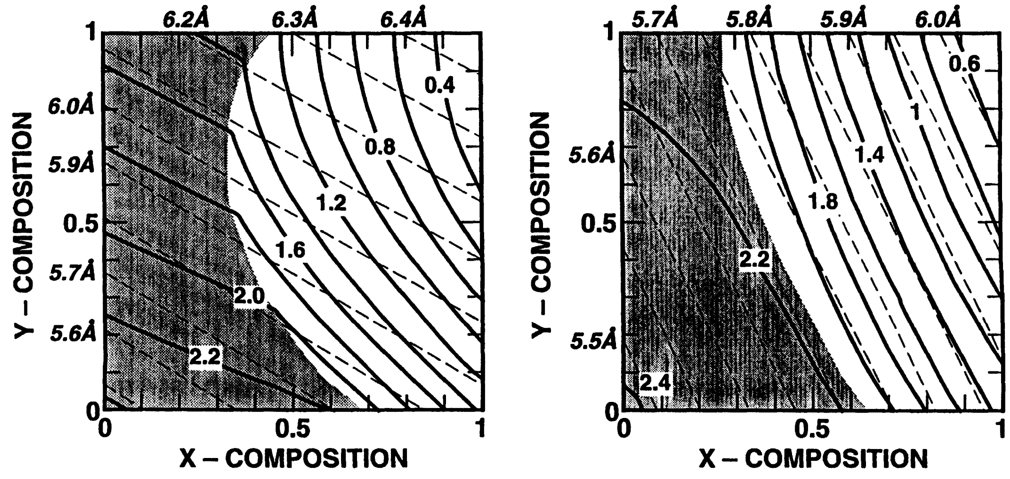

(a) (b)

FIG. 9. Energy-band gap (solid curves) and lattice-constant contours (dashed curves) at 300 K for (a) $Al_{1-x}In_xP_{1-y}Sb_y$ and (b) $Al_{1-x}In_xP_{1-y}As_y$ alloys. The shaded region shows the compositional range over which the material has an indirect band gap. (From Glisson *et al.*, 1978.)

strate), there is in-plane compressive strain in the layer and tensile strain normal to the interface. When $(\Delta a/a)_\perp$ is negative, there is in-plane tension and compression normal to the layer. The hydrostatic component of the biaxial strain shifts the band gap while the shear component lifts the degeneracy of the valence band, as shown in Fig. 10. For biaxial compression the valence band closer to the conduction band is the heavy-hole band, whereas it is the light-hole band that is closer for biaxial tension. A similar situation is realized for externally applied uniaxial strain of tension and compression, respectively. The wafer bowing resulting from the biaxial strain is also shown in Fig. 10.

Asai and Oe (1983) derived the transition energies to the $J = 3/2$, $m_j = 3/2$ and $J = 3/2$, $m_j = 1/2$ valence bands using the stress-dependent Hamiltonian (Bir and Pikus, 1974), and they are as follows:

$$E_{(3/2,3/2)} = \left[-2a\frac{C_{11} - C_{12}}{C_{11}} + b\frac{C_{11} + 2C_{12}}{C_{11}} \right]\varepsilon, \quad (32)$$

$$E_{(3/2,1/2)} = \left[-2a\frac{C_{11} - C_{12}}{C_{11}} - b\frac{C_{11} + 2C_{12}}{C_{11}} \right]\varepsilon, \quad (33)$$

where ε is the mismatch strain, taken to be positive for compression.

From Eqs. (32) and (33), it follows that for in-plane compressive strain (positive mismatch), the lowest-energy optical transition involves the $m_j = 3/2$ valence band, and for in-plane tensile strain (negative mismatch) it involves the $m_j = 1/2$ valence band. Figure 11 shows the band gap, derived from photoluminescence measurements, of $\mathrm{Ga}_x\mathrm{In}_{1-x}\mathrm{As}$ grown on InP for lattice mismatch varying from -0.2% to $+0.2\%$ (Bassignana *et al.*, 1989). The data clearly illustrate the strain effect on band gap. There is good agreement between the experimental points and the solid lines derived from Eqs. (32) and (33) using the extrapolated values of elastic and deformation-potential constants for $\mathrm{Ga}_{0.47}\mathrm{In}_{0.53}\mathrm{As}$ from the binary values.

1.2.2.5 Effective Mass in Ternary and Quaternary Alloys. According to the **k**·**p** approximation, the effective mass of the car-

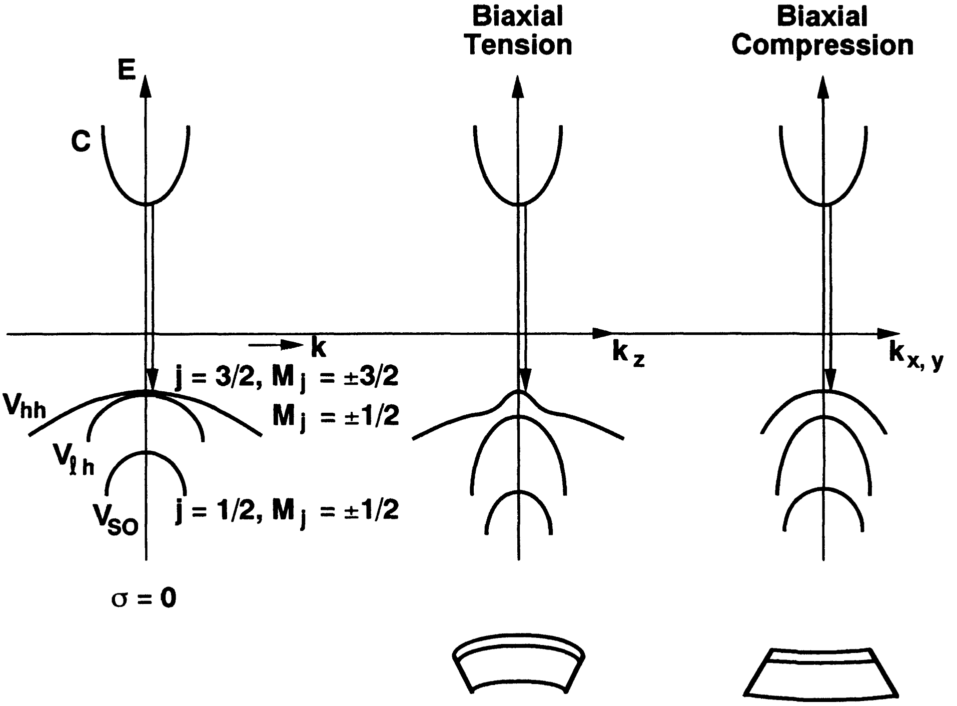

FIG. 10. Valence-band structure of cubic semiconductors under biaxial strain. The fourfold degenerate valence band is split into two twofold degenerate bands. The valence band closest to the conduction band is different depending on the nature of the strain. Wafer bowing resulting from biaxial strain is also shown.

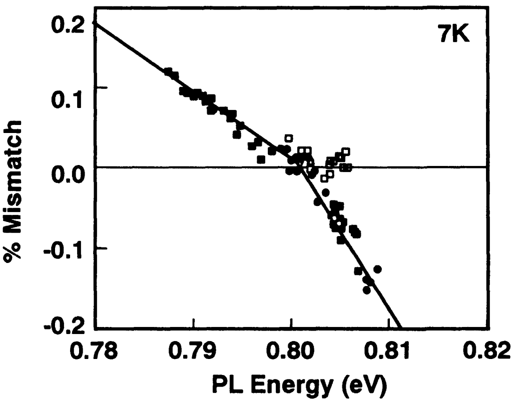

FIG. 11. Photoluminescence peak energy at 7 K of $Ga_xIn_{1-x}As/InP$ as a function of lattice mismatch. The effect of biaxial strain on peak energy is explained on the basis of Eqs. (32) and (33). (From Bassignana *et al.*, 1989.)

rier is related to the square of the matrix element P connecting the conduction band and the light-hole and spin–orbit-split valence bands. It has been found in many ternary alloys that the effective mass calculated under the assumption that P^2 varied linearly between the values for the binary compounds does not agree well with the experimental values. The origin of this discrepancy has been attributed to disorder-induced mixing of the conduction and valence bands at the Γ point (Berolo *et al.*, 1973). The effect of this mixing is to reduce P^2 and thus to increase the effective mass above that which would be obtained by the $\mathbf{k}\cdot\mathbf{p}$ calculation.

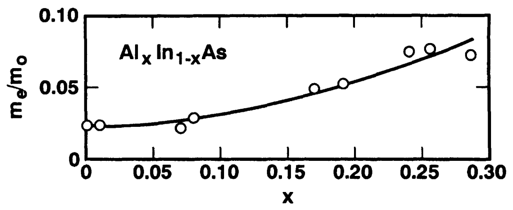

FIG. 13. Electron effective mass versus composition in $Al_xIn_{1-x}As$ (from Matyas and Karoza, 1982.)

Figures 12 and 13 show the electron effective mass for $Ga_xIn_{1-x}As$ (Thomas and Woolley, 1971) and $Al_xIn_{1-x}As$ (Matyas and Karoza, 1982), respectively. Figures 14 and 15 show, respectively, the electron effective mass and hole effective masses for lattice-matched $Ga_xIn_{1-x}As_yP_{1-y}$ (Pearsall, 1982). The heavy-hole and split-off band masses can be obtained from the values of the binary compounds using Eq. (23).

1.2.2.6 Phonons in Ternary and Quaternary Alloys. In ternary or quaternary alloys, one can expect phonon frequencies due to the different binary compounds. Such multiple-mode behavior has been observed in the Raman-scattering and infrared-reflectivity measurements on alloys. Figure 16 shows the phonon frequencies in lattice-matched $Ga_xIn_{1-x}As_yP_{1-y}$ (Soni *et al.*, 1986) as functions of y. The low-frequency (220–260 cm^{-1}) phonons are assigned to GaInAs modes. At $y = 1$, they are separated into InAs-type and GaAs-type vibrations. The high-frequency (300–360 cm^{-1}) band ap-

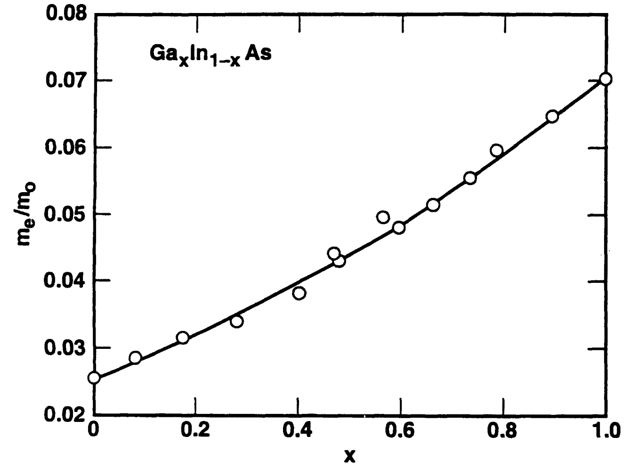

FIG. 12. Electron effective mass versus composition in $Ga_xIn_{1-x}As$ (from Thomas and Woolley, 1971.)

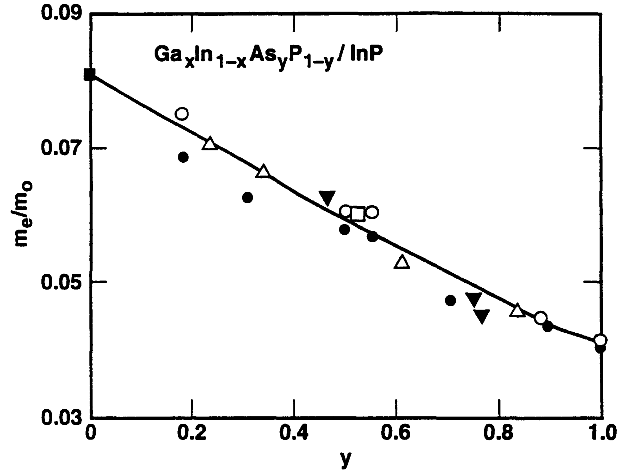

FIG. 14. Electron effective mass versus composition in $Ga_xIn_{1-x}As_yP_{1-y}$/InP (from Pearsall, 1982.)

pears for all compositions, except for $y = 1$, and is attributed to InP and GaP vibrations.

1.2.2.7 Refractive Index. One of the important factors in the design of heterostructure lasers and optoelectronic devices is the refractive index n. Often, knowledge of n as a function of photon energy near and below the band gap is required. Adachi (1982) proposed a generalized model of dielectric constants of semiconductors based on models of the intraband transitions. According to this model, the real part of the dielectric constant ϵ is expressed as a function of photon energy E as

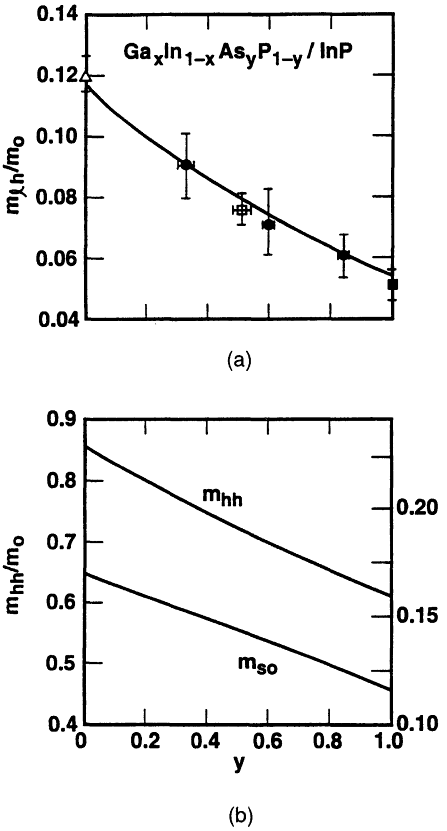

FIG. 15. (a) Light-hole effective mass and (b) heavy-hole effective mass and effective mass of the spin–orbit-split-off valence band versus composition in $Ga_xIn_{1-x}As_yP_{1-y}$ on InP substrates (from Pearsall, 1982.)

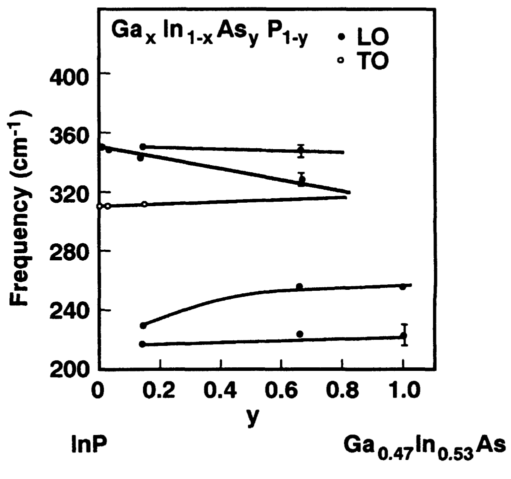

FIG. 16. Measured phonon frequencies of various Raman modes in $Ga_xIn_{1-x}As_yP_{1-y}$ as a function of y. The solid lines emphasize the experimental trends. (From Soni *et al.*, 1986.)

$$\epsilon(E) = A_0\left\{f(x) + \tfrac{1}{2}\left[\frac{E_g}{E_g + \Delta_0}\right]^{3/2} f(x_0)\right\} + B_0, \tag{34}$$

where A_0 and B_0 are constants obtained by fitting experimental data to Eq. (32), and

$$\begin{aligned} f(x) &= x^{-2}[2 - (1 + x)^{1/2} - (1 - x)^{1/2}], \\ x &= E/E_g, \\ x_0 &= E/(E_g + \Delta_0). \end{aligned} \tag{35}$$

Figure 17 shows the experimental values of $\epsilon(\approx n^2)$ for $Ga_xIn_{1-x}As_yP_{1-y}$ as a function of wavelength (Chandra *et al.*, 1981) for several compositions. The dispersion curves calculated based on Eqs. (34) and (35) are also shown in Fig. 17. The agreement with experimental curves is quite good over most of the wavelength range.

1.2.2.8 Thermal Resistivity. Thermal resistivity of semiconductors is an important parameter in the design of power-dissipating devices such as semiconductor lasers. It is also important in calculating the figure of merit of thermoelectric devices (e.g., Peltier devices). The thermal resistivity of ternary and quaternary alloys follows a quadratic relation the same as E_g. For the ternary alloy $A_xB_{1-x}C$ it can be written as

$$W(x) = xW_{AC} + (1 - x)W_{BC} + c_{A-B}x(1 - x), \tag{36}$$

where W_{AC} and W_{BC} are the thermal resistivities of the binary compounds. Values of the bowing parameter c_{A-B} for several ternary alloys have been obtained by Adachi (1983) and are given in Table 16. Note that c_{As-P} values of InAsP and GaAsP alloys are considerably smaller than c_{In-Ga} of InGaAs and that $c_{Ga-Al} \approx c_{As-P}$. The mean atomic weight of In-Ga is about two times than that of As-P, which is nearly the same as that of Al-Ga. The anharmonic contribution to thermal resistivity resulting from mass difference between the atoms is supposed to be responsible for the differences and similarities in the c values.

The thermal resistivity of a quaternary alloy $A_xB_{1-x}C_yD_{1-y}$ can be expressed in the same way as Eq. (35) but with two bowing parameters c_{A-B} and c_{C-D}. The thermal resistivity of $Ga_xIn_{1-x}As_yP_{1-y}$ lattice-matched to InP calculated using $c_{In-Ga} = 72$ W^{-1} deg cm and $c_{As-P} \simeq 25$ W^{-1} deg cm is shown in Fig. 18 (Adachi, 1983). The thermal resistivity increases with increasing As content and reaches a maximum value of 24 W^{-1} deg cm at $y \sim 0.75$.

1.2.2.9 Thermal Expansion Coefficient. In double heterostructures consisting of compositionally different layers, differences in the thermal expansion coefficient of the layers can generate elastic stresses during cooling from the growth temperature to room temperature or during thermal cycling of heterostructure devices. In Fig. 19

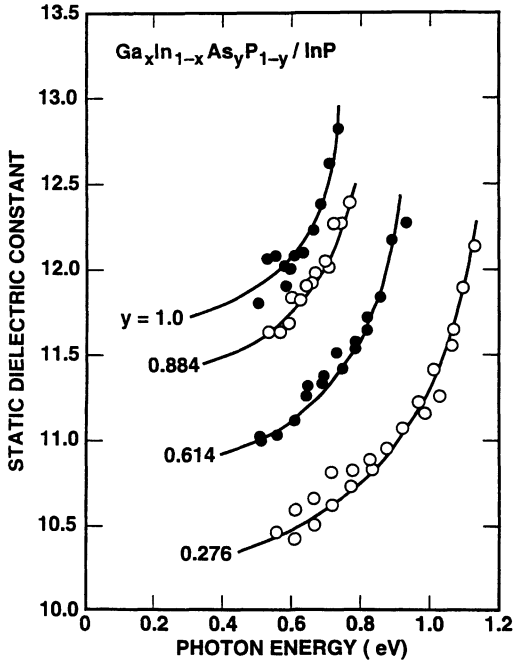

FIG. 17. Dielectric constant versus photon energy for $Ga_xIn_{1-x}As_yP_{1-y}$/InP at various compositions (from Chandra *et al.*, 1981.)

Table 16. The alloy disorder bowing parameter c_{A-B} denoting the deviation from linearity of the thermal resistivity of ternary alloys. The bowing parameter is obtained by fitting experimental data by Eq. (36), From Adachi (1982).

Ternary alloy	c_{A-B} (W^{-1} deg cm)
InGaAs	72
InAsP	30
GaAsP	20
AlGaAs	30
InGaP	72 (est)
InGaSb	72 (est)

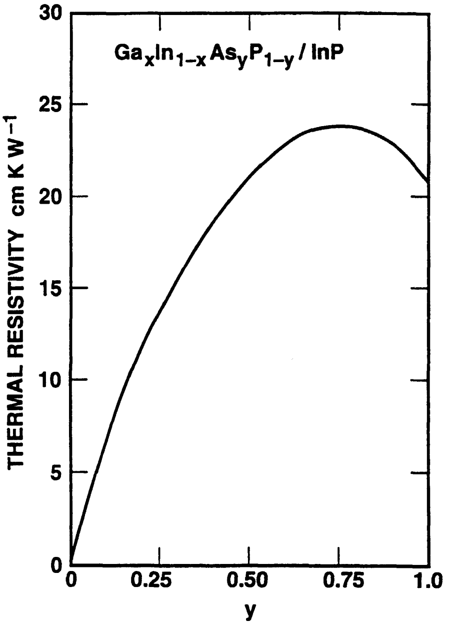

FIG. 18. Thermal resistivity versus composition calculated for $Ga_xIn_{1-x}As_yP_{1-y}$/InP (from Adachi, 1983.)

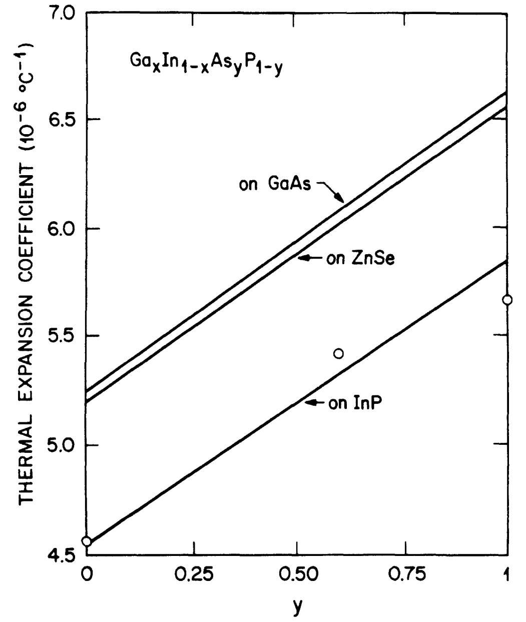

FIG. 19. Thermal expansion coefficient of $Ga_xIn_{1-x}As_yP_{1-y}$ lattice matched to InP, ZnSe, and GaAs as a function of composition. The curves are calculated and the points are experimental. (From Bisaro *et al.*, 1979.)

the thermal expansion coefficient of $Ga_xIn_{1-x}As_yP_{1-y}$ alloy lattice-matched to InP obtained from linear interpolation is shown together with experimental data at selected y values (Bisaro *et al.*, 1979). Pietsch and Marlow (1986) found that the thermal expansion coefficient of the relaxed lattice is about 30% smaller than the values shown in Fig. 19.

2. II–IV COMPOUNDS

2.1 Binary Compounds

2.1.1 Bonding and Crystal Structure II–VI compounds cover the entire range from fully ionic (NaCl structure, in CdTe at high pressure) to fully covalent (cinnabar, α-HgS) bonding. In all the tables of data it will be noted that several gaps occur; this is particularly so for HgS, which has received relatively little study, and ZnO. Table 17 gives the fractional ionic character (Phillips) and the cohesive energy. Table 18 summarizes lattice constants, densities, melting points, thermal expansion coefficients, and thermal conductivities at ~300 K for the binary compounds.

Most mercury chalgocenides crystallize in the cubic zinc-blende (sphalerite) structure with tetrahedral coordination, although the stable phase of HgS, α-HgS, has the hexagonal cinnabar structure. Some of the mercury chalgocenide alloys with other II–VI compounds crystallize in the hexagonal (wurtz-

Table 17. Fractional ionic character and cohesive energy of II–VI binaries.

Compound	Phillips ionicity	Cohesive energy[a] (kcal mole^{-1})
HgS	0.79	—
HgSe	0.68	—
HgTe	0.65	—
CdS	0.69	−120
CdSe	0.7	—
CdTe	0.72	−80
ZnS	0.63	−130
ZnSe	0.68	−110
ZnTe	0.55	−95
ZnO	0.62	−140

[a] Gibbs free energy of sublimation into neutral atoms at STP.

Table 18. Basic properties of the binary II–VI compounds. Values in parentheses are for hexagonal structures.

Compound	Lattice constant (a, c = hexagonal) (Å)	Density (g cm^{-3})	Melting point T_m (K)	Thermal expansion coefficient (10^{-6} °C^{-1})	Thermal conductivity (W cm^{-1} K^{-1})
HgS	5.85	7.7	1723	7.7	—
HgSe	6.086	8.2	1073	8.3	0.016
HgTe	6.462	8.05	943	5.2	0.031
CdS	5.82 (4.13 a, 6.74 c)	(4.82)	1678–1748	(4.2)	(0.2)
CdSe	6.082 (4.3 a, 7.01 c)	5.7 (5.81)	1523	(4.8)	(0.043)
CdTe	6.481 (4.57 a, 7.47 c)	5.67 (5.9)	1365–1371	5.1 (5.9)	0.01
ZnS	5.41 (3.82 a, 6.26 c)	(4.1)	1991–2123	(6.2)	0.026
ZnSe	5.67 (4.00 a, 6.54 c)	5.42 (5.26)	1788–1798	6.8 (6.2–6.5)	0.19
ZnTe	6.13 (4.27 a, 6.99 c)	5.72 (5.7)	1563	(8.2)	0.18
ZnO	(3.25 a, 5.21 c)	(5.28)	—	(7.2)	—

ite) structure when immiscibility is found, but complete miscibility exists if both binaries are of the sphalerite structure. The tetrahedrally coordinated bonds between the metal atoms and the four neighboring chalcogen atoms are formed by sp^3 hybrids, and the bonding is partly covalent and partly ionic. The band structure of the mercury chalcogenides is of the symmetry-induced zero-gap type. The Γ_6 level, which is usually in the conduction band, is below the Γ_8 level, usually in the valence band, giving negative fundamental band gaps; i.e., the compounds are semimetals.

Cadmium-based and zinc-based compounds tend to crystallize in the cubic zincblende structure, at atmospheric pressure, but CdSe, CdS, and ZnS can form in the hexagonal wurtzite structure. At high pressures, CdTe transforms into the NaCl structure and then, at higher pressures still, into the white tin structure. CdTe is the most ionic of the Cd and Zn chalcogenides (see Table 17) with ZnTe the least ionic.

The IIa–VIa binary compounds have the NaCl crystal structure, and these refractory compounds are prepared in a high-temperature floating melt-zone Xe-arc furnace.

2.1.2 Energy-Band Structure

The Brillouin zone of cubic crystals was described in Sec. 1.1.1.2. The lattice of wurtzite crystals consists of two interpenetrating hexagonal close-packed lattices, one containing anions and the other cations. The translation vectors are

$$(a/2)(1,-\sqrt{3},0),\ (a/2)(1,\sqrt{3},0),\ \text{and}\ c(0,0,1),$$

where the vectors are given with respect to Cartesian axes. The reciprocal lattice is hexagonal having the translation vectors

$$(2\pi/a)(1,-3^{-1/2},0), (2\pi/a)(1,3^{-1/2},0), (2\pi/c)(0,0,1).$$

The Brillouin zone is shown in Fig. 20. The calculated energy-band structures for the cubic ZnSe and CdTe are given in Figs. 21(a) and 21(b), while the hexagonal ZnS diagram is given in Fig. 21(c) and that of the semimetal HgTe is given in Fig. 21(d). The critical-point energies are shown in Figs. 21(a)–21(d).

The spin–orbit splittings for the II–VI binaries are given in Table 19. Table 19 also shows the energy gaps of the binaries and their temperature and pressure dependences, where known.

Figure 22 shows a plot of the energy gap against lattice parameter for selected II–VI compounds. The lines are simply tie lines,

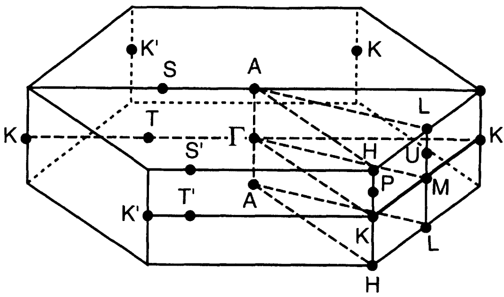

FIG. 20. Brillouin zone for the wurtzite lattice with lines and points of special symmetry (from Aven and Prenner, 1967.)

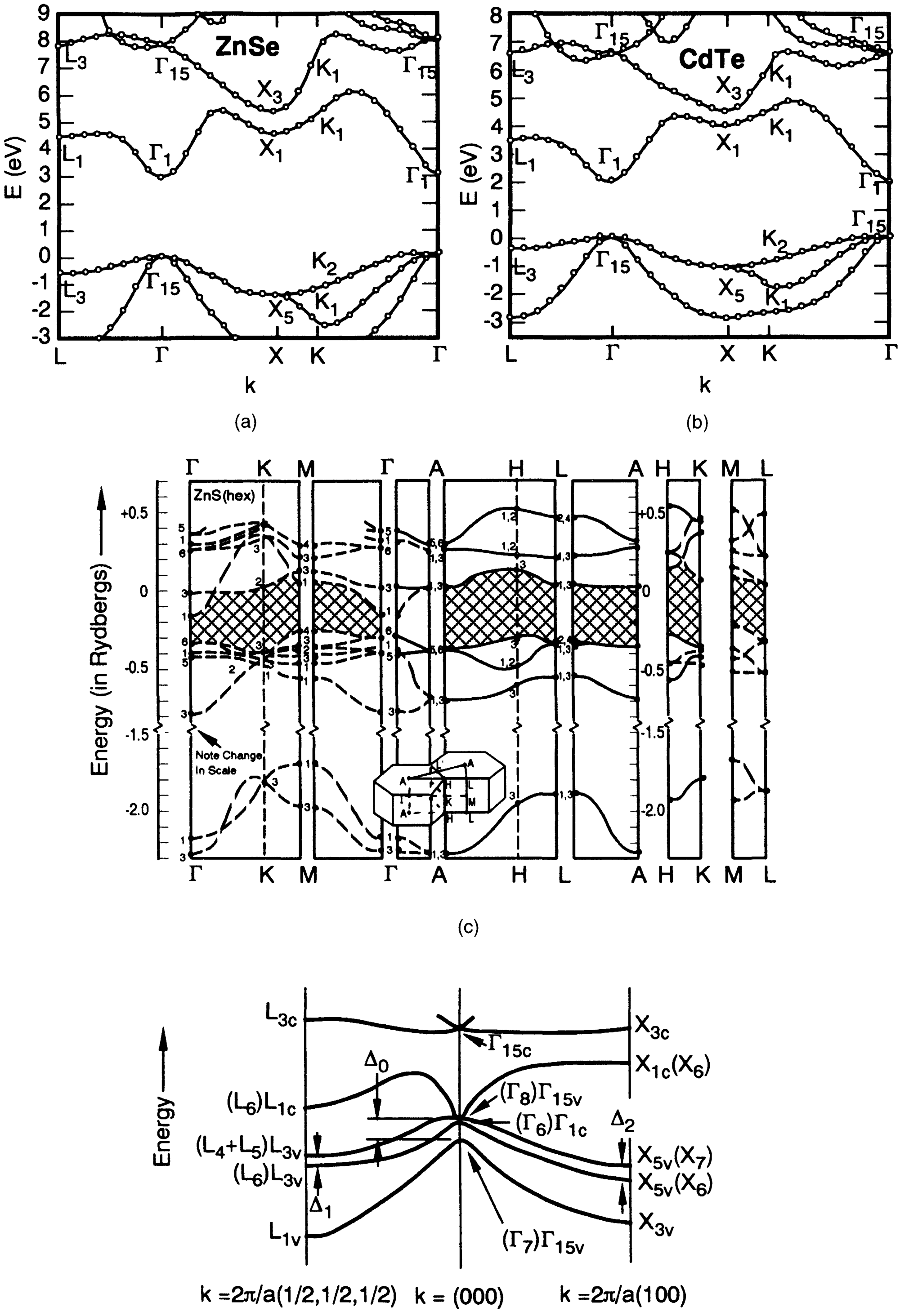

(a) (b)

(c)

(d)

FIG. 21. Energy bands for (a) ZnSe, (b) CdTe, (c) ZnS, and (d) HgTe (from Aven and Prenner 1967.)

Table 19. Energy-band structure parameters for binary II–VI compounds. Values in parentheses are for hexagonal structures.

Compound	Electron affinity	Spin–orbit splitting (e = exp) (eV)	Energy gap (eV)	$\frac{dE_g}{dP}$ (10^{-6} eV kbar^{-1})	$\frac{dE_g}{dT}$ (10^{-4} eV K^{-1})
HgS	—	—	−0.2	—	—
HgSe	—	0.35	−0.22	—	−7
HgTe	—	1.08 (e), 0.78	−0.3	7–14	—
CdS	(4.79)	0.065	2.42, (2.53)	—	−5.2
CdSe	(4.95)	0.42	1.67, (1.9)	—	−9
CdTe	4.28	0.9 (e), 0.86	1.5	8	−3.6
ZnS	3.9	0.27	3.6, (3.78)	—	−5 to −10
ZnSe	4.09	0.43	2.67	—	−7.2
ZnTe	3.53	0.91 (e), 0.89	2.2	—	−5
ZnO	(4.57)	—	(3.2)	—	−10

but, as Vegard's law applies in all the systems, the divisions indicate the relevant proportions of the binaries required to produce a given energy gap. Table 20 gives the effective masses of the binary compounds.

Table 21 gives the energy gaps (at 2 K) of the IIa–VIa binaries, and Fig. 23 shows the theoretical band structure for the sulfide binaries; see Kaneko and Koda (1988). Most of these compounds are predicted to be indirect-gap semiconductors, except for BaO.

2.1.3 Impurities and Defects Point defects are very important in II–VI compounds while line defects, such as dislocations, are probably not significant in affecting transport properties. The most important point defects are vacancies and interstitials of the constituents, either of which can be monovalent or divalent. In CdTe and CdS, interstitials are present in much lower concentrations than residual impurities but could still be responsible for deep trap levels. There is

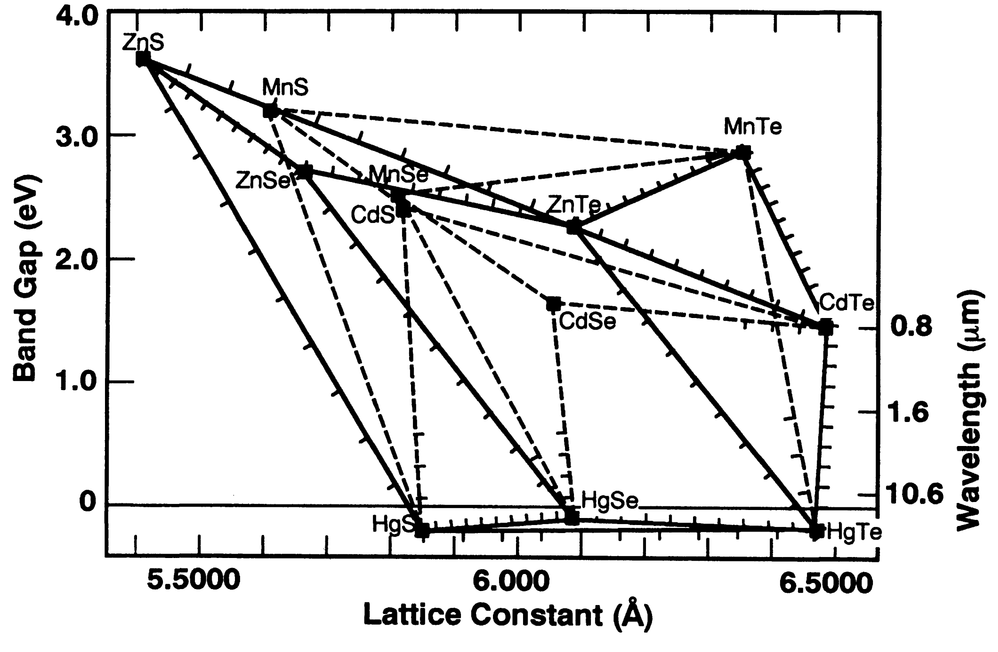

FIG. 22. Energy gap versus lattice parameter for selected II–VI binaries.

Table 20. Effective electron and hole masses of binary II–VI compounds. Values in parentheses are for hexagonal structures.

Compound	m_e/m_0	m_h/m_0
HgSe	0.05	0.02–0.08
HgTe	0.03	0.6
CdS	(0.2)	(0.8)
CdSe	(0.13)	(0.45)
CdTe	(0.14) 0.09	(0.37)
ZnS	0.31	0.5–1.0
ZnSe	(0.17)	(0.6)
ZnTe	(0.16) 0.12	(0.1–0.3)

Table 21. Band-structure parameters (in eV) of binary IIa–VIa compounds.

Compound	Indirect gap X_c–Γ_v (eV)	Direct gap $E_g(X)$ (eV)	Direct gap $E_g(\Gamma)$ (eV)
CaO	—	6.875	—
CaS	4.434	5.343	5.8
CaSe	3.85	4.898	—
SrO	—	5.793	6.08
SrS	4.32	4.831	5.387
SrSe	3.813	4.475	4.57
BaO	—	3.985	8.3
BaS	3.806	3.941	5.229
BaSe	3.421	3.658	4.556

no evidence of S, Se, or Te interstitials, and isolated vacancies of S, Se, and Te appear to be inactive electrically. Vacancies of Zn and Cd have been the most widely studied point defects, either isolated or complexed with impurities. The Zn-vacancy–shallow-donor complex (*A* center) is the major compensating defect in *n*-type ZnSe and ZnS. Similar complexes (Cd-vacancy–shallow donor) occur in CdS and CdSe heat-treated in S and Se atmospheres, respectively. Vacancies of Zn are double acceptors in ZnTe. In general, the

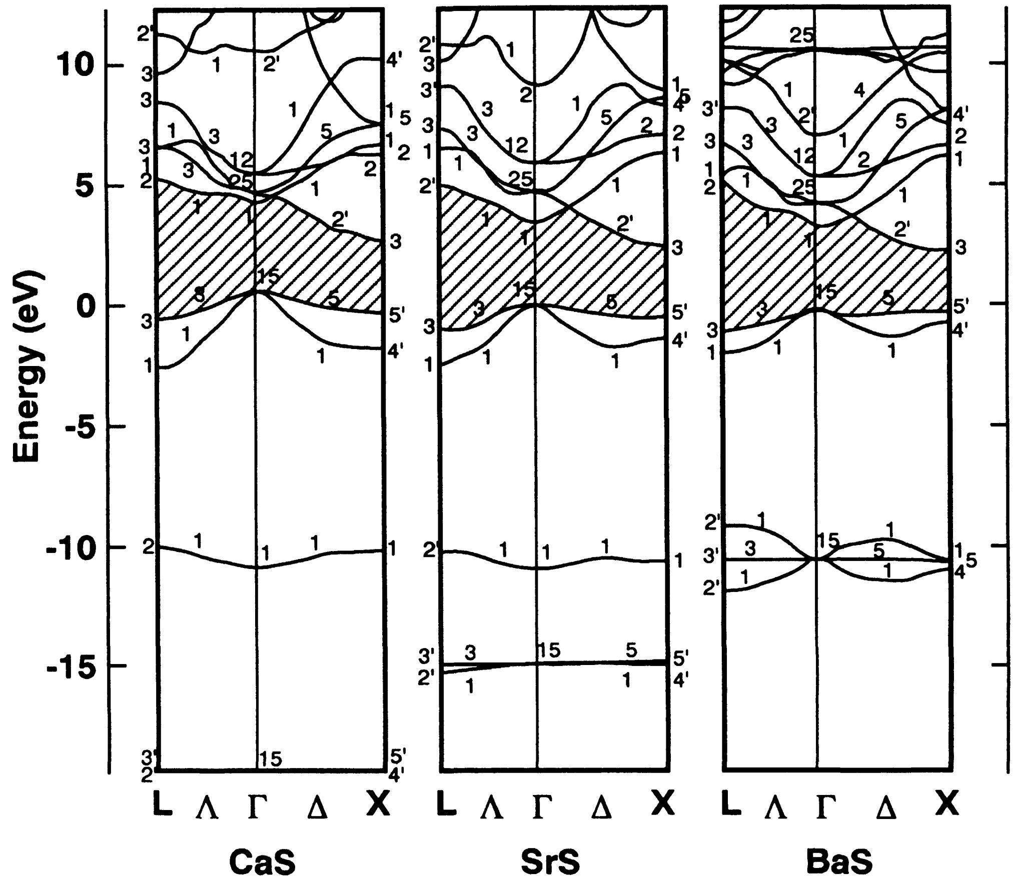

FIG. 23. Energy-band structure for IIa–VIa binaries (from Kaneko and Koda, 1988).

concentration of isolated interstitials and vacancies is quite low, except for cation vacancies in the tellurides.

In early work on wide-gap II–VI compounds, it was found impossible to produce *p*-type ZnSe, CdS, CdSe, or ZnS (even after annealing in the chalcogen vapor), while ZnTe could only be prepared *n* type. This was due to the low solubility of impurities and to compensation problems (probably together with deep trap levels). In general, the higher the ionicity of the compound, the more easily *n*-type material forms and the more difficult it is to obtain *p*-type material. With the advent of the more advanced nonequilibrium epitaxial techniques of metal-organic vapor-phase epitaxy (MOVPE), photo-assisted (PA) MOVPE, molecular beam epitaxy (MBE), PAMBE, atomic-layer epitaxy (ALE), and plasma-assisted epitaxy (PAE), *p*-type doping in ZnSe has been achieved using Li, N, and O, despite the latter being isoelectronic with Se. *n*-type doping has also been accomplished using Al, Cl, Br, and iodine. ZnSe has received more attention than the other wide-gap compounds because of the drive to produce blue lasers.

On account of the lack of good-quality bulk II–VI substrates, most growth is on III–V compound substrates such as GaAs (which is mismatched by ~0.27% with ZnSe and matched to ZnS_xSe_{1-x} with $x = 0.05$), InSb, or InP, or on Si or Ge. The orientation of the substrate can affect the phase of CdS and CdSe epitaxial layers grown. A major step has been the recent production of purer Se from which to grow purer, undoped, high–electrical-resistivity, uncompensated ZnSe.

Low-temperature photoluminescence is the main tool used in the study of impurities and defects, and there is a plethora of complex structures involving two or more defects and extrinsic impurities. Transmission electron microscopy (TEM) studies have revealed dislocation networks at substrate/layer interfaces and the presence of III_2VI_3 interfacial compounds. Stacking faults and microtwins are also often observed [particularly on the (111) orientation]. GaAs epitaxial buffer-layer growth prior to the deposition of the ZnSe can reduce these markedly.

Now that controlled bipolar doping has been achieved in ZnSe, the emphasis is switching to studies of the deep-level centers, which may be limiting the radiative recombination efficiency, and those structural defects that also limit device performance. Clearly reproducibility of both growth and device technologies are being addressed currently. Strained-layer superlattices (SLSs) of ZnSe/ZnSSe and ZnSe/ZnMnSe are thought to show promise for laser structures.

CdTe has received a great deal of attention, and defects can render the material *n* or *p* type, depending on the heat treatment used. Such treatments do, however, affect the amount of second-phase precipitation, the dislocation density, and the strain in the material. Extrinsic doping has met with the same problems of low solubility and compensation as in the wide-gap compounds. Once again, use of nonequilibrium low-temperature growth, such as PAMBE, has led to *n*-type doping using In and iodine and to *p*-type doping with As and Sb. Additions of Zn and/or Se to CdTe hardens the lattice, reduces the dislocation density, and decreases the x-ray rocking-curve widths. Such materials are extensively used as substrates in the epitaxial growth of HgCdTe layers. Defects in epitaxial CdTe grown on GaAs substrates have been delineated with TEM and cathodoluminescence, where misfit dislocations are seen to be numerous. Various buffer layers can be used to decrease these dislocations. Deep-level centers have been studied with deep-level transient spectroscopy (DLTS), electron paramagnetic resonance (EPR), and optically detected magnetic resonance (ODMR), while native metal vacancies have recently come under investigation using positron annihilation techniques.

There is very little information on defects in the narrow-gap Hg compounds, particularly for HgS. HgSe can only be prepared *n* type, as the usual acceptor species from groups I and V are inactive. HgTe can be prepared both *n* and *p* type. HgTe has a very high electron mobility at low temperatures as a result of its low effective mass and high purity. Doping of HgTe with acceptors from group I and donors from groups III and IV has been accomplished. "Doping" of HgSe with Fe produces drastic changes in both the electrical and magnetic properties of the binary. Similar addition of Mn to HgTe produces interesting magnetic properties. These diluted magnetic semiconductor effects were originally studied in CdMnTe but now include Co^{2+}, Fe^{2+}, and Eu^{2+} as "dopants."

The ability to produce zinc-blende structure binaries of MnSe, MnTe, and FeSe by MBE led to the increased interest in these compounds. See SEMICONDUCTORS, DILUTED MAGNETIC.

2.1.4 Diffusion

Diffusion data are important in both material and device technologies, particularly in the areas of control of stoichiometry, impurity doping, and compositional interdiffusion. Most data to date have been gathered in bulk samples, as epitaxial material needs much lower temperatures and the material is available in much smaller volumes, making quantitative diffusion studies more difficult. There is a need to monitor or control temperature, nonstoichiometry, and doping/impurity levels during diffusion work. Diffusion occurs in conditions of either chemical equilibrium, with no gradients in chemical potential, or chemical disequilibrium, where diffusion occurs down gradients in chemical potential. Atoms can migrate via nondefect (ring or direct exchange processes) or defect (vacancy, interstitial, or complexes) mechanisms. Diffusion coefficients are related to concentration profiles and are measured using radio-tracer, Rutherford back-scattering (RBS), resistivity or capacitance-voltage (C–V) profiling, and *p*/*n*-junction depth measurements. Diffusion measurements in multiquantum-well (MQW) and SLS structures are becoming more important, but problems are severe because of the small dimensions.

The mechanism of chalcogen self-diffusion is via neutral self-interstitials, under chalcogen-rich conditions, and via chalcogen vacancies at high metal partial pressures. In general, diffusion coefficients $D(X)$ are $\sim 10^{-8}$ cm^2 s^{-1} at 1 atm but fall rapidly with temperature; e.g., for CdTe at $T_m/2$, $D(X) \sim 10^{-12}$ cm^2 s^{-1}. Metal self-diffusion is more complex as a result of nonstoichiometry and doping effects. In ZnSe, ZnTe, and CdTe, $D(M) \sim 10^{-8}$ cm^2 s^{-1} at T_m but falls to $\sim 10^{-16}$ cm^2 s^{-1} at $T_m/2$ and is believed to be due to either neutral associates $[(M_iV_m)$ or $(V_mV_x)]$ or metal interstitials (M_i). Doping has been seen to increase metal self-diffusion in CdTe. In ZnS, CdS, and CdSe diffusion varies with partial pressure in the wurtzite structure but not in the zinc-blende structures. Despite the wurtzite structure of CdS no anisotropy of diffusion is observed. Data in Hg-containing compounds is scarce, but metal diffusion is much faster in HgTe than in CdTe, at a given temperature. Short-circuit path diffusion has been suggested to explain this difference. Chalcogen diffusion rates in HgTe were similar to those in CdTe. Information on chemical self-diffusion is scarce but diffusion coefficients $D(\Delta)$ are in the range 10^{-4}–10^{-6} cm^2 s^{-1} at T_m and $\sim 10^{-12}$ cm^2 s^{-1} at $T_m/3$. In CdTe, $D(\Delta)$ is independent of Cd partial pressure, decreases in doped material, is purity related, but is independent of dislocation density. It is thought to be due to metal interstitials $(\mathrm{Cd}_i)^{2-}$ and is important in controlling stoichiometry at the annealing temperatures used and during cooldown from the growth temperatures.

Impurity-diffusion profiles in II–VI compounds are very often complex and do not follow simple erfc or Gaussian distributions. They often have two sections, arbitrarily designated as fast and slow, with the former possibly linked to short-circuit paths—e.g., dislocations or grain boundaries. Elements from groups I, III, IV, V, and VII have all been studied, mainly because of their doping behavior. A vast range of diffusion coefficients D_i exists at a given temperature together with a variety of mechanisms. For Au in CdTe, $D_i \sim 10^{-6}$ cm^2 s^{-1} at T_m, while for Cu in CdTe a value of 10^{-13} cm^2 s^{-1} at $T_m/4$ has been determined. For group-III elements in ZnTe, CdTe, and ZnSe, values of 10^{-6} cm^2 s^{-1} at T_m are typical, decreasing to 10^{-11}–10^{-13} cm^2 s^{-1} at $T_m/2$. Data for other groups are scarce, but it should be noted that diffusion of group-I elements and iodine in CdTe has been reported to be significant even at room temperature. Few data exist in HgTe, or other Hg-containing compounds, but where they do, diffusion coefficients appear to be higher than equivalent ones in CdTe.

Interdiffusion $[\bar{D}(x)]$ in ternary compounds is a function of composition (x), nonstoichiometry, doping, and temperature. Typically $\bar{D}(x)$ values lie in the range 10^{-10}–10^{-13} cm^2 s^{-1} at 600–1100 °C. In common-metal ternaries $\bar{D}(x)$ should be independent of doping, whereas donor and possibly acceptor doping enhances $\bar{D}(x)$ in common-chalcogen systems. The latter could have important consequences for stability in doped MQW and SLS structures, for which there is a dearth of information. Impurity diffusion

Table 22. Optical properties of binary II–VI compounds.

Compound	Refractive index (wavelength in μm)	ε_0	ε_∞	LO phonon ($\mathrm{cm^{-1}}$)	TO phonon ($\mathrm{cm^{-1}}$)
HgS	n_e = 3.2–2.8 (0.6–11) n_0 = 2.9–2.6 (0.6–11)	—	11.4	—	—
HgSe	—	23	12.2	180	130
HgTe	—	20.8	15.1	148	148
CdS	2.7	9.1 ($E \parallel c$) 8.58 ($E \perp c$)	5.26 ($E \perp c$) 5.31 ($E \parallel c$)	304	240
CdSe	—	8.41 9.38 ($E \parallel c$)	5.98 ($E \perp c$) 6.07 ($E \parallel c$)	210	170
CdTe	2.65	10.36	7.18	170	140
ZnS	2.4–2.3 (0.5–1)	8.32	5.17	352	320
ZnSe	2.43–2.39 (3–12)	8.53	5.7	252	200
ZnTe	2.63 (10.6)	9.1	6.75	200	177
ZnO	—	~8	—	—	—

coefficients in MQW structures are typically $\sim 10^{-18}\ \mathrm{cm^2\ s^{-1}}$ at 300 °C.

2.1.5 Optical Properties Refractive indices, static and high-frequency dielectric constants, and the frequencies of zone-center LO and TO phonons of the binaries are given in Table 22. See Sec. 1.1.5 for a full description of the pertinent equations. Both the static and high-frequency dielectric constants decrease with increasing energy gap, as seen in the III–V compounds.

Optical bistability has been observed in CdTe, CdS, ZnSe, and ZnS. Optical nonlinearities for applications in optical-bistability and all-optical circuits have been studied in wide- and narrow-gap II–VI compounds (Miller, 1987, 1992). The polarization P of a material varies in a nonlinear manner with electric field E, as

$$P = \varepsilon_0(\chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \text{etc.}), \qquad (37)$$

where ε_0 = permittivity of free space and χ's are susceptibilities. $\chi^{(2)}$ is responsible for second-harmonic generation, sum- and difference-frequency generation, paramagnetic oscillation, optical rectification, and the linear electrolyte effect (Pockels). $\chi^{(3)}$ gives rise to third-harmonic generation, the quadratic electrooptical effect (Kerr), stimulated Brillouin and Rayleigh scattering, stimulated Raman scattering, four-photon mixing, and optical phase conjugation. Measured values of nonlinear optical coefficients range from a nonresonant value of $\sim 5 \times 10^{-12}$ e.s.u. in CdS to >1 e.s.u. in CdHgTe under resonant conditions.

2.1.6 Mechanical Properties Bulk modulus (B), elastic stiffness constants (C_{11},C_{12},C_{44}), and elastic compliances (S_{11},S_{12}) for the binary compounds are given in Table 23. Section 1.1.6 gives the relevant basic equations and the relationships be-

Table 23. Mechanical properties of binary II–VI compounds. Values in parentheses are for hexagonal structures.

Compound	B^a (10^{11} Pa)	C_{11} (10^{11} Pa)	C_{12} (10^{11} Pa)	S_{11} ($10^{-11}\ \mathrm{Pa^{-1}}$)	S_{12} ($10^{-12}\ \mathrm{Pa^{-1}}$)	C_{44} (10^{11} Pa)
HgTe	4.76	0.585	0.408	—	—	0.222
CdS	6.64 (6.61)	0.858	0.5334	—	—	1.49
CdSe	5.57 (5.32)	0.749	0.461	—	—	0.132
CdTe	4.45	0.535	0.367	4.25	−1.7	0.2
ZnS	8.02 (8.21)	1.07	0.67	1.89	−0.72	0.46
ZnSe	6.67	0.826	0.498	2.26	−0.85	0.44
ZnTe	5.09	0.722	0.409	—	—	0.308
ZnO	(13.2)	—	—	—	—	—

---

[a] For cubic crystals, bulk modulus $B = 1/(3S_{11} + 2S_{12})$.

tween these parameters and Young's modulus E and Poisson's ratio ν ($\sim -S_{12}/S_{11}$). The elastic stiffness constants, together with lattice parameter and hardness values for the binaries, are used to calculate the microhardness of alloys. The formation of alloys is often accompanied by an increase in hardness, e.g., Zn into HgTe or HgCdTe. This is termed solid-solution hardening. Hardness of a material can be related to the type of chemical bonding and can serve as a measure of binding strength.

2.1.7 Heterojunctions Four considerations apply to the choice of a heterostructure system. These are energy gap, lattice mismatch, doping type, and band offsets. Figure 22 gives the relationship between the first two parameters, and Sec 2.1.3 outlined the doping behavior. Strain energy between mismatched binary pairs increases with thickness until a critical value is reached and dislocations are introduced to reduce the strain. Metastable structures, where a layer is sufficiently strained to match the previous layer but not enough to result in dislocations, also exist. Typical values reported for the valence-band offsets for the binary pairs indicated are given in Table 24, along with the electron affinities. It is clear that the differences in electron affinities between binary pairs do not normally agree with the experimental or theoretical (Tersoff) values of valence-band offset, in disagreement with Eq. (18) and Fig. 5. This offset is required to be low in solar-cell structures also, and this is thought to be one of the critical parameters in device performance. Figure 4 shows the three types of band offset that occur in heterojunctions. In II–VI compounds type I heterojunctions include ZnTe/ZnSe, while type II heterojunctions include ZnTe/CdSe. Type III semimetal/semiconductor superlattices consist of a semimetal, such as HgTe, and a semiconductor, such as CdTe. The wavelength of the infrared detector made from such structures can be controlled by the layer thickness rather than the composition uniformity, and the larger effective masses reduce the dark currents. The high diffusion coefficients in these systems, however, may lead to interdiffusion, even at the low growth temperature of ~200 °C used in MBE or on high-temperature storage of devices.

Table 24. Valence-band offsets (VBO) and electron affinities for pairs of binary II–VI compounds.

Binaries	VBO (exp), theory	Electron affinities
HgTe/CdTe	(0.35), 0.37	
HgSe/CdSe	0.57	
HgTe/ZnTe	(0.25), 0.26	
HgSe/ZnTe	0.03	
HgSe/ZnSe	0.6	
CdTe/ZnTe	(0.1), 0.13	4.28/3.53
CdS/CdSe	0.83	4.8/4.95
CdSe/MnSe	0.25	
CdTe/MnTe	0.25	
ZnS/ZnSe	0.82	3.9/4.09
ZnTe/ZnSe	1.0	3.53/4.09
ZnTe/CdSe	0.61	3.53/4.95
ZnSe/CdS	0.54	4.09/4.8
ZnS/CdSe	1.05	3.9/4.95
ZnS/CdS	0.28	3.9/4.8

These heterojunctions form the basis for numerous SL, SLS, and MWQ structures with useful electrical, optical, and magnetic properties. The last field has benefited from the availability of zinc-blende structure Mn-containing compounds, for use as barrier layers, via nonequilibrium growth techniques.

2.2 Alloys

2.2.1 Energy-Band Structure of Alloys Ternary alloys enable tuning of the band gap and lattice constants by varying the alloy composition, making these compounds very versatile. In pseudobinary alloys and wide-gap II–VI compounds, the fundamental band edge $E_g = \Gamma_6 - \Gamma_8$ varies, in general, nonlinearly with composition and/or temperature. Most measurements are fitted by

$$E_g = [xE_g(AC) + (1 - x)E_g(BC)] - bx(1 - x), \tag{38}$$

where $E_g(AC)$ and $E_g(BC)$ are the band gaps for the binaries and b is a "bowing parameter," but significant temperature dependences are also seen. Empirical relationships (both theoretical and experimental) of $E_g(x,T)$ (with T in K and E_g in eV) are given in Table 25.

The lattice parameter versus energy gap at room temperature for the II–VI binaries is shown in Fig. 22. The availability of good-

Table 25. Variation of energy gap with composition and temperature in ternaries and quaternaries.

Compound	$E_g(x, T)$
$Hg_{1-x}Cd_xTe$	$E_g = -0.313 + 1.787x + 0.444x^2 - 1.237x^3 + 0.932x^4 + (0.667 - 1.714x + 0.76x^2)T/1000 \quad (T > 70\ K)$
$Hg_{1-x}Cd_xSe$	$E_g = -0.209(1 - 7.172x - 2.174x^2) + 0.000737\,(1 - 1.277x - 0.151x^2)T + 2.001 \times 10^{-9}\,(1 + 23.45x - 599.4x^2)T^2$
$Hg_{1-x}Zn_xTe$	$E_g = -0.27 + 1.3x + 1.27x^2\ (T = 77\ K)$
	$E_g = -0.17 + 1.1x + 1.27x^2\ (T = 300\ K)$
	$E_g(x, T) = -0.3 + 0.0324x^{1/3} + 2.731x - 0.629x^2 + 0.533x^3 + 0.00053T\,(1 - 0.76x^{1/2} - 1.29x)$
$HgSe_{1-x}Te_x$	$E_g = -0.0687 - 0.484x + 0.408x^2\ (T = 300\ K)$
$Hg_{1-x}Mn_xSe$	$E_g = -0.27 + 0.044x\ (x\ 0.37)\ (T = 10\ K)$
$Zn_{1-x}Hg_xSe$	$E_g = 2.781 - 0.31x + 2.3x^2$
$Cd_{1-x}Zn_xTe$	$E_g = 1.4637 + 0.49613x + 0.2289x^2\ (T = 300\ K)$
	$E_g = 1.586 + 0.5006x + 0.29692x^2\ (T = 77\ K)$
$CdTe_xSe_{1-x}$	$E_g = 1.78 + 0.937x + 0.755x^2\ (T = 5\ K)$
$Cd_{1-x}Mn_xTe$	$E_g = 1.59 + 1.4x\ (T = 4\ K)$
ZnS_xSe_{1-x}	$E_g = [xE_g(ZnSe) + (1 - x)E_g(ZnS)] - 0.456x(1 - x)$
$ZnSe_xTe_{1-x}$	$E_g = [xE_g(ZnSe) + (1 - x)E_g(ZnTe)] - 1.23x(1 - x)$
	$E_g = 2.67 - 1.64x + 1.23x^2\ (T = 300\ K)$
ZnS_xTe_{1-x}	$E_g = [xE_g(ZnS) + (1 - x)E_g(ZnTe)] - 3x(1 - x)$
$Zn_xCd_{1-x}S$	$E_g = 2.43 + 1.28x\ (T = 300\ K)$
$Zn_xCd_{1-x}Se$ $(x < 0.6)$	$E_g = E_g(CdSe) + [E_g(ZnSe) - E_g(CdSe) - 0.82]x + 0.82x^2\ (300\ K)$
(W, $x < 0.5$)	$E_g = E_g(CdSe) + [E_g(ZnSe) - E_g(CdSe) - 0.87]x + 0.87x^2\ (80\ K)$
(ZB, $x > 0.7$)	$E_g = 1.832 + 0.688x + 0.35x^2\ (T = 77\ K)$
$Zn_{1-y}Cd_ySe_{1-x}Te_x$	$E_g(x, y) = 2.82 - 1.39y - 1.935x + 1.182xy + 0.35y^2 + 1.507x^2 - 0.184xy^2 - 0.752yx^2\ (T = 5\ K)$

quality, lattice-matched substrates of sufficient size and purity/perfection is still a limitation in the growth of well-developed ternary systems (see Sec. 2.2.2).

2.2.2 Material Properties of Alloys In general, the II–VI binaries form complete solid solutions with each other. Exceptions to this rule are where one of the binaries crystallizes in the hexagonal wurtzite form. In these cases, miscibility gaps exist. Diluted magnetic semiconductor materials, based on Mn, show these effects when Mn(Te,Se,S) combine with the other II–VI binaries. Figure 24 shows an overview of the crystal structure with zinc-blende (Cub) and wurtzite (Hex) regions highlighted. Many ternary alloys follow Vegard's law (see Sec. 1.2.2.1); i.e., the lattice parameter varies linearly with composition between the binaries—e.g., the HgTe/CdTe, HgTe/ZnTe, ZnS/ZnSe, and ZnS/CdS systems. Where lattice-parameter differences are large, the choice of substrate, and its quality and size, is difficult, and recourse is often then taken to growing a II–VI compound on a III–V substrate—e.g., ZnSe on GaAs. Some systems show a bowing in lattice parameter between the end binaries. This general rule applies to many other physical properties of ternary alloys. There is insufficient space in this article to give much detail of ternaries and quaternaries, and so illustrative examples from the $Hg_{1-x}Cd_xTe$ and $Cd_{1-x}Zn_xTe_{1-y}Se_y$ systems will be given.

As mentioned in the Introduction, materials from the CdTe, CdZnTe, and CdTeSe series are used extensively as substrates for the epitaxial growth of $Hg_{1-x}Cd_xTe$, which is then used to fabricate infrared detectors. The properties that make the Cd/Zn/Te/Se materials suitable include high transmission (~65%), lattice matching to the HgCdTe (for x or $y \sim 0.04$ in $Cd_{1-x}Zn_xTe_{1-y}Se_y$), low levels of large-size second-phase precipitation, low dislocation densities (mid-10^4 cm^{-2}), high electrical resistivity (10^9 Ω cm), similar thermal and mechanical properties to HgCdTe, and availability in sufficient single-crystal, twin-free quantities. The addition of Zn and Se to CdTe is known to harden the lattice (as does the addition of Zn to HgTe—see Sec. 2.1.6) as well as improve the lattice matching. Problems remaining in these materials include reproducibility, purity, and uniformity of lattice matching in the Zn-containing compounds due to the strong segre-

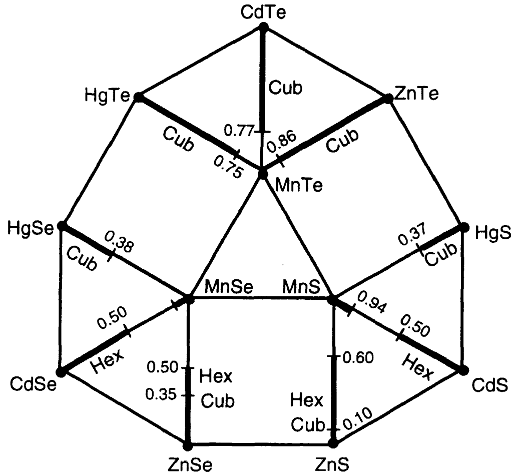

FIG. 24. Overview of $II_{1-x}Mn_xVI$ alloys and their crystal structure. "Hex" and "Cub" indicate wurtzite and zinc blende, respectively. (From Willardson and Beer, 1978.)

gation of Zn during growth. Despite these residual difficulties, many suitable high-quality substrates are produced and used, mainly in LPE growth of HgCdTe.

The electron effective mass in CdHgTe varies according to

$$(m_e/m_0) = 1 + 2F + 0.33E_p[2/E_g + 1/(E_g + \Delta)], \tag{39}$$

where F is a constant (~ -0.8), E_p = interband energy (~19 meV), Δ = spin–orbit splitting (~1 eV), and E_g = energy gap, which is given in Table 25 as a function of composition x and temperature. For the heavy hole, effective-mass values lie in the range $(0.4\text{–}0.7)m_0$. Extrapolation from CdTe to CdZnTe and CdTeSe can be made using

$$(m_e/m_0)(2/E_g) + 1/(E_g + \Delta) = 0.187. \tag{40}$$

In CdHgTe, polar optical phonons arise from CdTe-type and HgTe-type modes and fine structure due to vibration modes of the various combinations of anions and cations in the lattice (from far-IR and Raman-scattering studies). The variations with composition, at 77 and 300 K, of these modes are shown in Fig. 25 (Capper, 1994).

The dielectric constants of CdHgTe vary with composition as

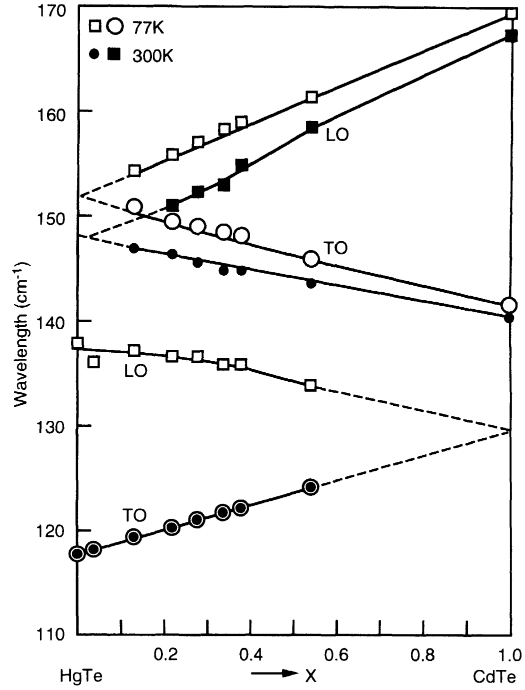

FIG. 25. Longitudinal (LO) and transverse (TO) optical phonon frequencies in HgCdTe at T = 77 and 300 K (from Capper, 1994.)

$$\varepsilon_\infty = 15.2 - 15.6x + 8.2x^2 \tag{41}$$

$$\varepsilon_0 = 20.5 - 15.6x + 5.7x^2. \tag{42}$$

ε_0 for CdZnTe varies nonlinearly between the binaries, increasing as Zn content increases up to 0.3, then decreasing. The dispersion of the refractive index of CdHgTe has been fitted by

$$n^2 = a_1 + \frac{a_2\lambda^2}{\lambda^2 - (1/a_3)^2} + \frac{a_4\lambda^2}{\lambda^2 - (a_5)^2}, \tag{43}$$

where a's are constants and λ = wavelength. Experimental data and the fits by this equation are shown in Figs. 26(a) and 26(b) (Capper, 1994). There are no experimental data on n for CdZnTe and CdTeSe, and various extrapolation procedures similar to the above expression have been proposed, but for CdTeSe these are problematic on account of the phase change that occurs as the Se content increases.

Linear thermal expansion coefficients for HgTe and CdTe can be approximated by

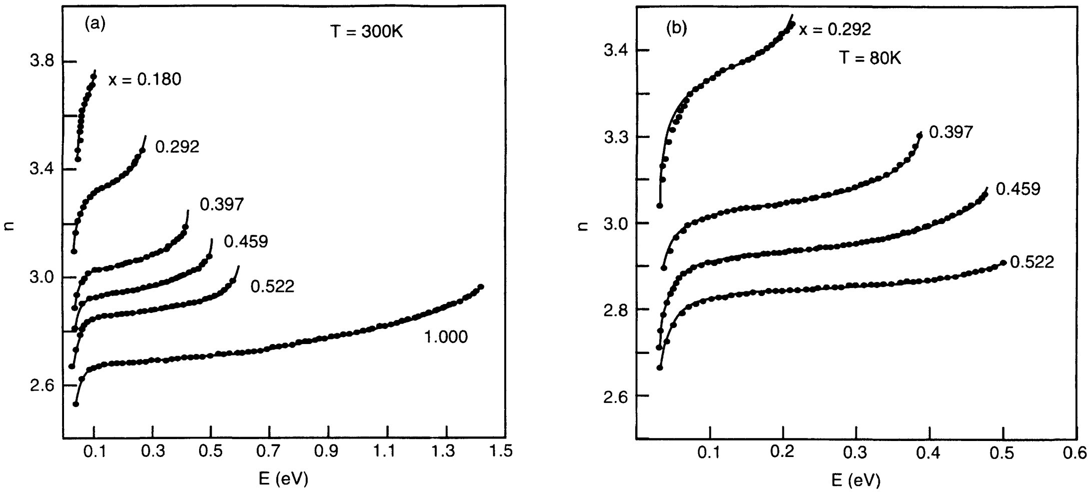

(a) (b)

FIG. 26. Spectral dependences of refractive index of HgCdTe at (a) $T = 300$ K and (b) $T = 80$ K (from Capper, 1994.)

$$\alpha = 4 \times 10^{-6} + 4 \times 10^{-9}T \tag{44}$$

and

$$\alpha = 4 \times 10^{-6} + 3.3 \times 10^{-9}\,T, \tag{45}$$

respectively. Values for $x = 0.2$ and 0.7 CdHgTe are 4.26×10^{-6} and 4.5×10^{-6}, respectively, at 290 K. Negative values are found for temperatures below about 60 K. No data exist for CdZnTe and CdTeSe.

The x range of interest for infrared detectors made from $Hg_{1-x}Cd_xTe$ is mainly 0.2–0.3, covering the 3–5 and 8–12 μm bands (Brice and Capper, 1987; Capper, 1994). Most relevant properties vary strongly with both x and temperature (see Capper, 1994, for details). The fundamental properties of high optical absorption coefficient, low intrinsic carrier concentration, high minority-carrier lifetime, and high electron/hole mobility ratio, together with the capability for band-gap engineering (see Table 25), make it almost ideal for a wide range of infrared detector types including photoconductors, Signal Processing In The Element (SPRITEs; see Glossary), photodiodes, and charge-transfer devices. Bipolar doping has been achieved using the lower diffusing elements in most growth techniques following the recognition of the importance of controlling the stoichiometry at the growth temperature. This control is achieved after a low-temperature Hg anneal to minimize the concentration of metal vacancies, which are the principal defects and act as acceptors (possibly doubly ionized). However, the physical properties of the material, which include high mercury vapor pressure, a widely separated phase diagram, high diffusion coefficients, softness, and brittleness, have presented challenging problems to both the material preparation and device fabrication areas. Challengers to its preeminent position include Schottky barriers on silicon, SiGe heterojunctions, SiGe SLs, GaAlAs MQWs, III-V *nipi* structures, high-temperature superconductors, and thermal detectors. It is noteworthy, however, that none of these competitors can really compete in terms of fundamental properties.

3. CONCLUSION

This review has given an overview of some of the main material properties of a wide range of III–V and II–VI binary, ternary, and quaternary compounds. In addition to basic bonding and crystal and energy-band structures, various optical, mechanical, and defect properties have been described. Some of the multitude of device types have been outlined and suitable material combinations detailed. Representatives of the two types of compounds, i.e., GaAs and HgCdTe, are the second and third most studied semiconductors after silicon.

GLOSSARY

Band-Edge Offset: Difference in energy of conduction or valence bands in heterojunctions, which act as potential barriers to carrier movement.

Brillouin Zone: The polyhedron in reciprocal space over which the wave vector **k** is allowed to vary.

Bowing Parameter: Quantity describing the deviation from linearity of an alloy property variation with composition.

Carrier Effective Mass: Related to the matrix element P connecting the conduction band and the light-hole and spin–orbit-split valence bands.

Debye Temperature: A measure of the temperature above which all vibrational modes in a crystal are excited and below which they are frozen out.

Diluted Magnetic Semiconductors (DMS): Semiconductors with a fraction of constituent ions replaced by magnetic ions (e.g., Mn^{+2}, Fe^{2+}, etc.).

Fröhlich Coupling Constant: Constant describing the interaction of electrons with the LO phonon.

Heterojunction: Junction between two binaries with an abrupt change in energy-band structure.

Kramers–Kronig Relation: Relation between the static and high-frequency dielectric constants.

LO: Longitudinal optical phonon.

Multiquantum Well (MQW): Material with many QWs essentially producing a bulk solid with two-dimensional properties.

Phillips Ionicity: Fractional ionic character of binary compound estimated semiempirically from the ionic contribution to the energy gap.

Poisson's Ratio: Elastic constant derived from elastic compliances relating transverse to axial strain.

Quantum Well (QW): A three-layer (minimum) structure producing carrier confinement giving rise to a "two-dimensional" semiconductor.

Signal Processing in the Element (SPRITE): A type of infrared detector where the scene is scanned across the detecting element at the same rate as carriers drift along the element.

Spin–Orbit Splitting: Splitting of orbitally degenerate energy levels in crystals by the spin–orbit interaction.

Strained-Layer Superlattice (SLS): As for SL, but where the binaries have very different lattice parameters.

Superlattice (SL): As the separation of wells in a MQW structure is reduced, the energy levels in one well start to interact with those in neighboring wells, producing an SL.

TO: Transverse optical phonon.

Vegard's Law: Linear variation of lattice parameter with composition in alloy systems.

Young's Modulus: Elastic constant derived from elastic compliances relating stress to strain in the elastic region.

Works Cited

Adachi, S. (1982), *J. Appl. Phys.* **53,** 5863–5869.

Adachi, S. (1983), *J. Appl. Phys.* **54,** 1844–1848.

Anderson, R. L. (1962), *Solid State Electron.* **5,** 341–351.

Asai, H., Oe, K. (1983), *J. Appl. Phys.* **54,** 2052–2056.

Bassignana, I. C., Miner, C. J., Puetz, N. (1989), *J. Appl. Phys.* **65,** 4299–4305.

Berolo, O., Woolley, J. C., Van Vechten, J. A. (1973), *Phys. Rev.* **B8,** 3794–3798.

Bir, G. L., Pikus, G. E. (1974), *Symmetry and Strain Induced Effects in Semiconductors,* New York: Wiley.

Bisaro, R., Mevenda, P., Pearsall, T. P. (1979), *Appl. Phys. Lett.* **34,** 100–102.

Brantley, W. A. (1973), *J. Appl. Phys.* **44,** 534–535.

Brice, J. C., Capper, P. (Eds.) (1987), *Properties of Mercury Cadmium Telluride,* EMIS Datareview Series No. 3, London: INSPEC/IEE.

Capper, P. (Ed.) (1994), *Properties of Narrow-Gap Cadmium-based Compounds,* EMIS Datareview Series No. 10, London: INSPEC/IEE.

Casey, H. C., Jr., Panish, M. B. (1978), *Heterostructure Lasers Part B: Materials and Operating Characteristics,* New York: Academic, p. 9.

Chandra, P., Coldren, L. A., Strege, K. E. (1981), *Electron. Lett.* **17,** 6–7.

Cohen, M. L., Chelikowsky, J. R. (1988), *Electronic Structure and Optical Properties of Semiconductors,* Berlin: Springer-Verlag.

DeVreese, J. T. (Ed.) (1972), *Polarons in Ionic Crystals and Polar Semiconductors*, Amsterdam: Elsevier Science Publishers.

Forrest, S. R. (1987), in: F. Capasso, G. Margaritondo (Eds.), *Heterojunction Band Discontinuities*, Amsterdam: North-Holland, p. 311.

Ghosh, D. K., Samanta, L. K., Bhar, G. C. (1986), *Infrared Phys.* **26,** 111–114.

Glisson, T. H., Hauser, J. R., Littlejohn, M. A., Williams, C. K. (1978), *J. Electron. Mat.* **7,** 1–16.

Harrison, W. A. (1977), *J. Vac. Sci. Technol.* **14,** 1016–1021.

Hocking, E. F., Kudman, I., Seidel, T. E., Schmalz, C. M., Steigmeier, E. F. (1966), *J. Appl. Phys.* **37,** 2879–2887.

Holland, M. G. (1966), in: R. K. Willardson, A. C. Beer (Eds.), *Semiconductors and Semimetals*, Vol. 2, New York: Academic, p. 3.

Hornstra, J., Bartels, W. J. (1978), *J. Cryst. Growth* **44,** 513–517.

Hybertsen, M. (1990a), *J. Vac. Sci. Technol.* **B8,** 773–778.

Hybertsen, M. (1990b), *Phys. Rev. Lett.* **64,** 555–558.

Kane, E. O. (1982), in: W. Paul (Ed.), *Handbook on Semiconductors*, Vol. 1, Amsterdam: North Holland, Chap. 4A.

Kaneko, Y., Koda, T. (1988), *J. Cryst. Growth* **86** 72–78.

Kohn, W. (1957), in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics—Advances in Research and Applications*, Vol. 5, New York: Academic, p. 257.

Kroemer, H. (1985), "Surface and Interface Effects in VLSI," in: N. G. Einspruch, R. S. Bauer (Eds.), *VLSI Electronics Microstructure Science*, Vol. 10, Chap. 4, New York: Academic.

Kroger, F. A. (1974), *The Chemistry of Imperfect Crystals*, Vol. 2, Amsterdam: North-Holland, p. 237.

Lang, D. V. (1987), in: F. Capasso, G. Margaritondo (Eds.), *Heterojunction Band Discontinuities*, Amsterdam: North-Holland, p. 377.

Madelung, O., Schulz, M., Weiss, H. (Eds.) (1982), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, Group III, *Crystal and Solid State Physics*, Vol. 17a, *Physics of Group IV Elements and III–V Compounds*, Berlin: Springer.

Madelung, O., Schulz, M., Weiss, H. (Eds.) (1984), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, Group III, *Crystal and Solid State Physics*, Vol. 17d, *Technology of III–V, II–VI and Non-Tetrahedrally Bonded Compounds*, Berlin: Springer.

Madelung, O., Schulz, M., (Eds.) (1987), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, Group III, *Crystal and Solid State Physics*, Vol. 22a, *Intrinsic Properties of Group IV Elements and III–V, II–VI and I–VII Compounds*, Berlin: Springer.

Martin, G. M., Makram-Ebeid, S. (1986), "The Mid-Gap Donor Level EL2 in GaAs," in: S. Pantelides (Ed.), *Deep Centers in Semiconductors*, Chap. 6, New York: Gordon and Breach.

Matthews, J. W., Blakeslee, A. E. (1974), *J. Cryst. Growth* **27,** 118–125.

Matyas, E. E., Karoza, A. G. (1982), *Phys. Status Solidi B* **111,** K45–K48.

Miller, A. (1987), in: T. C. Brice, P. Capper (Eds.), *Properties of Mercury Cadmium Telluride*, EMIS Datareview Series No. 3, London: INSPEC/IEE, p. 35.

Miller, A. (1992), in: H. E. Ruda (Ed.), *Widegap II–VI Compounds for Opto-Electronic Applications*, London: Chapman & Hall.

Milnes, A. G., Feucht, D. L. (1982), *Heterostructures and Metal Semiconductor Junctions*, New York: Academic.

Moon, R. L., Antypas, G. A., James, L. W. (1974), *J. Electron. Mater.* **3,** 635–644.

Nye, J. F. (1957), *Physical Properties of Crystals*, Oxford: Clarendon Press.

Pantelides, S. T. (1975), in: H. J. Queisser (Ed.), *Festkörperprobleme*, Vol. XV, Berlin: Springer, p. 149.

Pearsall, T. P. (1982), in: T. P. Pearsall (Ed.), *GaInAsP Alloy Semiconductors*, New York: Wiley, p. 295.

Pearsall, T. P., Eaves, L., Portal, J. C. (1983), *J. Appl. Phys.* **54,** 1037–1047.

Phillips, J. C. (1970), *Phys. Rev B* **1,** 1540–1544.

Phillips, J. C. (1973), *Bonds and Bands in Semiconductors*, New York: Academic.

Phillips, J. C., Van Vechten, J. A. (1969), *Phys. Rev. Lett.* **22,** 705–708.

Pietsch, U., Marlow, D. (1986), *Phys. Status Solidi A* **93,** 143–149.

Ruan, Y. C., Ching, W. Y. (1987), *J. Appl. Phys.* **62,** 2885–2897.

Seraphin, B. O., Bennett, H. E. (1967), in: R. K. Willardson, A. C. Beer (Eds.), *Semiconductors and Semimetals*, New York: Academic, p. 499.

Soni, R. K., Abbi, S. C., Jain, K. P., Balkanski, M., Slempkes, S., Benchimol, J. C. (1986), *J. Appl. Phys.* **59,** 2184–2188.

Tejedor, C., Flores, F. (1978), *J. Phys.* **C11,** L19–L23.

Thomas, M. B., Woolley, J. C. (1971), *Can. J. Phys.* **49,** 2052–2060.

Van Vechten, J. A. (1969), *Phys. Rev.* **187,** 1007–1020.

Van Vechten, J. A., Bergstresser, T. K. (1970), *Phys. Rev.* **B1,** 3351–3358.

Varshni, Y. P. (1967), *Physica* **34,** 149–154.

Zallen, R. (1982), in: W. Paul (Ed.), *Handbook on Semiconductors*, Amsterdam: Elsevier Science Publishers, Vol. 1, Chap. 1.

Zielinski, E., Schweizer, H., Streubel, K., Eisek, H., Weimann, G. (1986), *J. Appl. Phys.* **59,** 2196–2204.

Further Reading

Adachi, S. (1985), *J. Appl. Phys.* **58,** R1.

Advanced Materials for Optics and Electronics—European Workshop II–VI (1994), **3.**

Aven, M., Prener, J. S. (Eds.) (1967), *Physics and Chemistry of II–VI Compounds,* Amsterdam: North-Holland.

Bartoli Jr., F. J., Schaake, H. F., Schetzina, J. F. (Eds.) (1990), *Properties of II–VI Semiconductors, Bulk Crystals, Epitaxial Films, Quantum Well Structures and Dilute Magnetic Systems, Mater. Res. Soc. Symp. Proc.* **161.**

Bhargava, R. N., Ruth, R. P., Yno, T., Nurmikko, A. V. (1994), Special issue on II–VI compounds, *J. Cryst. Growth* **138,** 1–1120.

Biefeld, R. M., Gunshor, R. L., Malik, R. J. (Eds.) (1991), *Long-Wavelength Semiconductor Devices, Materials and Processes, Mater. Res. Soc. Symp. Proc.* **216.**

Casselman, T. N. (Ed.) (1986), Proceedings of the 1985 U.S. Workshop on Physics and Chemistry of Mercury Cadmium Telluride, *J. Vac. Sci. Technol.* **A4,** 1955–2238.

Cheung, D. T. (Ed.) (1982), Proceedings of the 1st U.S. Workshop on Physics and Chemistry of Mercury Cadmium Telluride, *J. Vac. Sci. Technol.* **21,** 109–263.

Cheung, D. T. (Ed.) (1983), Proceedings of the 1982 U.S. Workshop on Physics and Chemistry of Mercury Cadmium Telluride, *J. Vac. Sci. Technol.* **A1,** 1586–1764.

Doe, John (Ed.) (1985), Proceedings of the 1984 U.S. Workshop on Physics and Chemistry of Mercury Cadmium Telluride, *J. Vac. Sci. Technol.* **A3,** 55–284.

Dornhaus, R., Nimtz, G. (Eds.) (1985), *Narrow-Gap Semiconductors,* Berlin: Springer-Verlag.

Elliott, C. T., Gordon, N. T. (1993), in: T. S. Moss (Ed.), *Handbook on Semiconductors,* Amsterdam: Elsevier Science Publishers B. V., Vol. 4 (Ed.) C. Hilsum.

Farrow, R. F., Schetzina, J. F., Cheung, J. T. (Eds.) (1987), *Materials for Infrared Detectors and Sources, Mater. Res. Soc. Symp. Proc.* **90.**

Fonash, A. (1981), *Solar Cell Device Physics,* New York: Academic Press.

J. Cryst. Growth **59** (1982); **72** (1985), **86** (1988), **117** (1991).

J. Electron. Mater. **22/8** (1993), p. 801–1112.

Katz, A. (Ed.) (1991), *Indium Phosphide and Related Materials: Processing, Technology, and Devices,* London: Artech House.

Madelung, O., Schulz, M., Weiss, H. (Eds.) (1982), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology,* Group III, *Crystal and Solid State Physics,* Vol. 17a, *Physics of Group IV Elements and III–V Compounds,* Berlin: Springer.

Løvold, S., Mullin, B. (Eds.) (1992), Proceedings of the NATO Workshop on Narrow-Gap Semiconductors, *Semicond. Sci. Technol.* **6,** C1–C136.

Madelung, O., Schulz, M., Weiss, H. (Eds.) (1984), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology,* Group III, *Crystal and Solid State Physics,* Vol. 17d, *Technology of III–V, II–VI and Non-Tetrahedrally Bonded Compounds,* Berlin: Springer.

Madelung, O., Schulz, M., (Eds.) (1987), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology,* Group III, *Crystal and Solid State Physics,* Vol. 22a, *Intrinsic Properties of Group IV Elements and III–V, II–VI and I–VII Compounds,* Berlin: Springer.

Mater. Sci. Eng. **B16** (1993).

McGill, T. C., Sotomayor-Torres, C. M., Gebhardt, W. (Eds.) (1989), *Growth and Optical Properties of Wide-Gap II–VI Low-Dimensional Semiconductors, NATO Advanced Study Institutes, Series B: Physics,* vol. 200, New York: Plenum.

Mullin, B., Stradling, T. (Eds.) (1993), Special issue on narrow-gap semiconductors, *Semicond. Sci. Technol.* **8,** S1–S456.

Neuberger, M. (1972), *Handbook of Electronic Materials,* Vol. 7, *III–V Ternary Semiconducting Compounds—Data Tables,* New York: Plenum.

Ruda, H. E. (Ed.) (1992), *Widegap II–VI Compounds for Opto-Electronic Applications,* London: Chapman & Hall.

Schaake, H. F. (Ed.) (1987), Proceedings of the 1986 U.S. Workshop on Physics and Chemistry of Mercury Cadmium Telluride, *J. Vac. Sci. Technol.* **A5,** 2997–3210.

Schaake, H. F. (Ed.) (1988), Proceedings of the 1987 U.S. Workshop on Physics and Chemistry of Mercury Cadmium Telluride, *J. Vac. Sci. Technol.* **A6,** 2589–2839.

Schaake, H. F. (Ed.) (1990), Proceedings of the 1989 U.S. Workshop on Physics and Chemistry of Mercury Cadmium Telluride, *J. Vac. Sci. Technol.* **A8,** 989–1259.

Seiler, D. G. (Ed.) (1991), Proceedings of the 1990 U.S. Workshop on Physics and Chemistry of Mercury Cadmium Telluride and Novel Infrared Detector Materials, *J. Vac. Sci. Technol.* **B9,** 1611–1901.

Seiler, D. G. (Ed.) (1992), Proceedings of the 1991 U.S. Workshop on Physics and Chemistry of Mercury Cadmium Telluride and Other II–VI Compounds, *J. Vac. Sci. Technol.* **B10,** 1345–1662.

Swaminathan, V., Macrander, A. T. (1991), *Materials Aspects of GaAs and InP Based Structures,* New Jersey: Prentice Hall.

Willardson, R. K., Beer, A. C. (Eds.) (1978), *Semiconductors and Semimetals,* Vols. 13, 25, New York: Academic.

Willardson, R. K., Beer, A. C. (Eds.) (1981), *Semiconductors and Semimetals,* Vols. 16, 18, New York: Academic.

SEMICONDUCTORS, DILUTED MAGNETIC

J. K. FURDYNA, M. DOBROWOLSKA, AND H. LUO, *Department of Physics, University of Notre Dame, Notre Dame, Indiana, U.S.A.*

Introduction 373
1. **Overview of Basic Properties** 374
1.1 Crystal Structure of Bulk DMS Alloys 374
1.2 Band Structure of DMS Crystals in the Absence of a Magnetic Field 376
1.2.1 General Band-Structure Features 376
1.2.2 The Energy Gap and Its Dependence on Composition 377
1.2.3 Intra-ionic Optical Transitions in $A^{II}_{1-x}Mn_xB^{VI}$ Alloys 379
1.3 Epitaxial DMS Systems 380
2. **Exchange Interactions** 381
2.1 The *sp-d* Exchange Interaction 381
2.2 The *d-d* Exchange Interaction 383
3. **Effect of *sp-d* Exchange on Electronic Properties** 384
3.1 Bulk DMS Alloys 384
3.1.1 Exchange Splitting of Exciton Transitions in DMSs 384
3.1.2 Giant Faraday Rotation in DMS Alloys 386
3.1.3 Acceptors and Negative Magnetoresistance in *p*-Type $Hg_{1-x}Mn_xTe$ 387
3.1.4 Magnetic-Field–Induced Metal-Insulator Transition 387
3.1.5 Magnetic Polarons 389
3.2 Quantum Wells and Superlattices 389
3.2.1 Spin Segregation 389
3.2.2 Wave-Function Mapping 391
3.2.3 Spin Lifetimes 392
4. **Magnetic Properties of DMS Systems** 393
4.1 Bulk DMS Alloys 393
4.1.1 Magnetic Phases 393
4.1.2 Magnetic Susceptibility of Paramagnetic Phase 394
4.1.3 Role of Antiferromagnetically Coupled Pairs 395
4.1.4 Spin-Glass Transition and Short-Range Antiferromagnetic Ordering 395
4.2 Magnetic Properties of DMS Layers and Multilayers 398
5. **Applications of DMSs** 399
Glossary 400
Works Cited 401
Further Reading 402

INTRODUCTION

Diluted magnetic semiconductors (DMSs) are semiconducting alloys whose lattice is made up in part of substitutional magnetic atoms. The most extensively studied and most thoroughly understood materials of this type are the $A^{II}_{1-x}Mn_xB^{VI}$ alloys, in which a fraction of the group-II sublattice is replaced at random by Mn. The worldwide scientific interest and intense research activity that this group of materials enjoys is evidenced by over two thousand papers published on DMSs to date, by an exceptionally large number of invited papers devoted to this subject at international conferences, and by a number of books (see, e.g., Furdyna and Kossut, 1988; Jain, 1991) and survey articles (see, e.g., Furdyna, 1988; Kossut and Dobrowolski, 1993) that have appeared in the last decade.

In recent years, studies involving II-VI semiconductors containing other transition-metal ions (e.g., iron and cobalt) also began to appear, as methods to prepare these new systems became perfected. However, the amount of research carried out on these new compounds, although steadily increasing, is still considerably smaller than that on the "traditional" manganese-based DMSs. Apart from the II-VI materials alloyed with transi-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

tion-metal chalcogenides, also IV-VI DMS alloys (e.g., $Pb_{1-x}Eu_xTe$, $Pb_{1-x}Mn_xTe$) are beginning to be investigated. Finally, there is a small body of literature already in existence on II-V semiconductor alloys involving transition metals [e.g., $(Cd_{1-x}Mn_x)_3As_2$].

The intense research activity in DMSs has its origin in the fact that the subject represents an interface of two well-established disciplines—semiconductor physics and magnetism—and the "cross products" of these otherwise independent areas of research lead to a host of new electrical, optical, and magnetic phenomena. Specifically, these "cross products" hold promise of modifying the electronic properties of a semiconductor by introducing magnetic interactions. In the $A^{II}_{1-x}Mn_xB^{VI}$ alloys, this change is brought about by the *sp-d* exchange interaction between the spins of band electrons and the localized moments of magnetic ions. The consequences, as we will detail later, are quite dramatic: *g* factors are effectively enhanced by as much as two orders of magnitude, Faraday rotations become very large, and the magnetoresistance can become negative, reaching gigantic values and leading to an insulator-to-metal transition induced by increasing magnetic field. Further, the DMS alloys are also of fundamental relevance to contemporary problems in disordered magnetism, since they display a low-temperature magnetic phase that strongly resembles a spin glass, as well as various types of antiferromagnetism. These magnetic properties arise from the exchange interaction between the *d* electrons of the magnetic ions (the *d-d* exchange). Finally, the hope is to harness some of the magnetic-field–related effects in DMS materials for magneto-optical device applications.

Recently the interest in DMS properties has been further increased by the development of methods to prepare DMSs in layer and multilayer forms with a high degree of perfection, including new crystal phases that do not exist in the bulk. Combining the physics of quantum wells and superlattices with the broad range of electronic and magnetic phenomena that exist in DMS materials promises an unusually rich spectrum of entirely new physical phenomena. Alternatively, extending magnetic studies to DMSs at concentrations of magnetic ions not accessible by bulk growth (e.g., $Zn_{1-x}Mn_xTe$ for $x \gtrsim 1$) has led to discoveries of novel forms of long-range antiferromagnetic order. All this is reflected in intense—and a rapidly growing amount of—research activity around the world devoted specifically to layered DMS structures.

Because of its multicomponent nature—semiconductor physics, magnetism, bulk and thin-film phenomena, combined with a large number of possible DMS alloys—a comprehensive presentation of the subject presents a challenge. In addressing this task, our strategy is as follows. We focus primarily on Mn-based DMSs, which by now have been most extensively studied, and whose behavior is representative of the phenomena that we consider most interesting. We also begin with a thorough discussion of bulk DMS systems. Having established the basic DMS properties in this way, it is then a simple matter to extend the discussion also to DMS-based layered structures.

1. OVERVIEW OF BASIC PROPERTIES

1.1 Crystal Structure of Bulk DMS Alloys

In all cases that are of present interest the "magnetic" transition-metal ions are substituted for the cations in a host semiconductor. Since transition metals do not belong to the same column of the periodic table as the atoms that they replace, this fact limits the atomic fraction of magnetic ions that can be incorporated in a stable form of various bulk DMSs.

Table 1 lists the range of molar fractions in which various DMS alloys were obtained, by bulk crystal growth, as well as the crystallographic structure of these alloys. One can see from the table, for example, that $Cd_{1-x}Mn_xTe$ forms a ternary alloy of zinc-blende structure with x up to 0.77, while $Zn_{1-x}Mn_xSe$ exhibits zinc-blende structure for $x < 0.30$ and wurtzite structure for $0.30 < x < 0.57$. Given the fact that the stable crystal structures of MnTe, MnSe, and MnS are neither zinc-blende nor wurtzite, it is remarkable that the $A^{II}_{1-x}Mn_xB^{VI}$ solid solutions shown in Table 1 can reach such high values of x. For example, MnTe itself crystallizes in the NiAs structure, yet in the case of $Cd_{1-x}Mn_xTe$ the zinc-blende structure of the

Table 1. Crystal structures and ranges of composition of bulk DMS alloys (for references see for example Furdyna, 1988, and Kossut and Dobrowolski, 1993).

Compound	Crystal structure	Composition range
$Zn_{1-x}Mn_xS$	Zinc blende	$0 < x < 0.1$
	Wurtzite	$0.1 < x < 0.45$
$Zn_{1-x}Mn_xSe$	Zinc blende	$0 < x < 0.30$
	Wurtzite	$0.30 < x < 0.57$
$Zn_{1-x}Mn_xTe$	Zinc blende	$0 < x < 0.86$
$Cd_{1-x}Mn_xS$	Wurtzite	$0 < x < 0.45$
$Cd_{1-x}Mn_xSe$	Wurtzite	$0 < x < 0.50$
$Cd_{1-x}Mn_xTe$	Zinc blende	$0 < x < 0.77$
$Hg_{1-x}Mn_xS$	Zinc blende	$0 < x < 0.37$
$Hg_{1-x}Mn_xSe$	Zinc blende	$0 < x < 0.38$
$Hg_{1-x}Mn_xTe$	Zinc blende	$0 < x < 0.75$
$(Cd_{1-x}Mn_x)_3As_2$	Tetragonal	$0 < x < 0.12$
$(Zn_{1-x}Mn_x)_3As_2$	Tetragonal	$0 < x < 0.15$
$Pb_{1-x}Mn_xS$	Rocksalt	$0 < x < 0.05$
$Pb_{1-x}Mn_xSe$	Rocksalt	$0 < x < 0.17$
$Pb_{1-x}Mn_xTe$	Rocksalt	$0 < x < 0.40$
$Zn_{1-x}Fe_xS$	Zinc blende	$0 < x < 0.26$
$Zn_{1-x}Fe_xSe$	Zinc blende	$0 < x < 0.21$
$Zn_{1-x}Fe_xTe$	Zinc blende	$0 < x < 0.01$
$Cd_{1-x}Fe_xSe$	Wurtzite	$0 < x < 0.20$
$Cd_{1-x}Fe_xTe$	Zinc blende	$0 < x < 0.03$
$Hg_{1-x}Fe_xSe$	Zinc blende	$0 < x < 0.20$
$Hg_{1-x}Fe_xTe$	Zinc blende	$0 < x < 0.02$
$Zn_{1-x}Co_xS$	Zinc blende	$0 < x < 0.15$
$Zn_{1-x}Co_xSe$	Zinc blende	$0 < x < 0.05$
$Cd_{1-x}Co_xSe$	Wurtzite	$0 < x < 0.08$

"parent" CdTe survives for x as high as 0.77, and even higher for $Zn_{1-x}Mn_xTe$.

The miscibility of manganese with II-VI hosts is particularly large, as compared with other transition-metal ions. The possibility of obtaining crystals in which the amount of the magnetic component can be varied in a continuous fashion within wide limits is one of the reasons why DMSs acquired such great interest in the scientific community.

The study of lattice parameters has been carried out most systematically in the case of DMSs based on II-VI compounds. The lattice parameters were found to vary linearly in proportion with the molar fraction of the magnetic component. (Such linear dependence on concentrations of respective constituents of an alloy is frequently referred to as Vegard's law.) For example, in the case of the zinc-blende structure, the lattice parameter can be expressed in the form

$$a = (1 - x)a_{\text{II-VI}} + x a_{\text{TM-VI}}, \qquad (1)$$

where x is the molar fraction of the transition-metal chalcogenide, and $a_{\text{TM-VI}}$ and $a_{\text{II-VI}}$ represent the lattice constants of the end-point binary compounds: a transition-metal chalcogenide in the (hypothetical) zinc-blende phase and the II-VI "parent," respectively.

In view of the diversity of possible lattice structures (zinc-blende and wurtzite) that various II-VI DMSs can acquire, it is useful to introduce the notion of the cation–cation distance d, in terms of which lattice dimensions can be given without specifying the explicit structure of the material. Figure 1 shows the variation of the cation–cation distance in II-VI–based DMSs containing manganese. For extensive information and references on this subject, see Chapter 1 in Furdyna and Kossut (1988).

The usefulness of such a plot is made clear by considering, for example, the linear behavior of d observed for the selenides. Note from Table 1 that $Hg_{1-x}Mn_xSe$ is cubic,

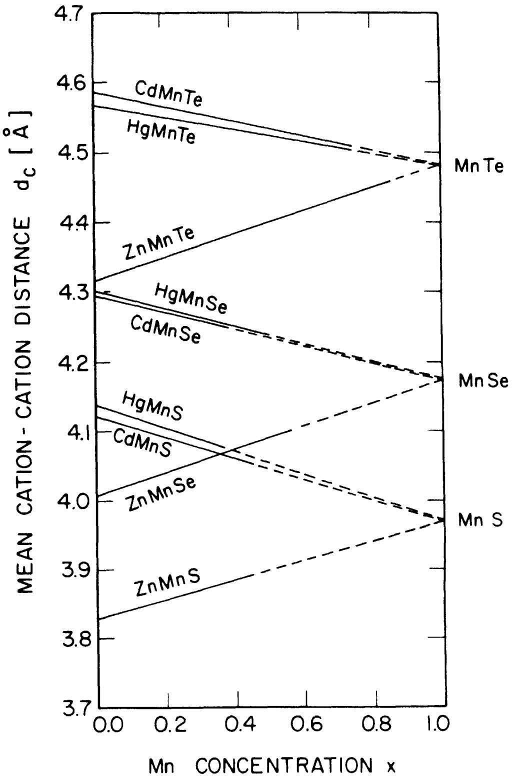

FIG. 1. Mean cation–cation distances d as a function of Mn mole fraction x for $A^{II}_{1-x}Mn_xB^{VI}$ alloys (after Furdyna, 1988).

$Cd_{1-x}Mn_xSe$ is hexagonal, and $Zn_{1-x}Mn_xSe$ changes its crystal structure "in stride" as x increases. However, having established the parameter d for MnSe by extrapolating, e.g., the lattice-constant data for the hexagonal $Cd_{1-x}Mn_xSe$ to $x = 1$, we can predict the lattice constant for the cubic $Hg_{1-x}Mn_xSe$, or for either of the two phases of $Zn_{1-x}Mn_xSe$. Two additional points are worthy of note in Fig. 1. First, it is interesting that the mutual relationship of d for HgSe, CdSe, and ZnSe in this figure closely resembles that for HgTe, CdTe, and ZnTe, despite the differences in structure. We also note that the value of d for $Zn_{1-x}Mn_xSe$ increases smoothly through the cubic-to-hexagonal transition, suggesting that it is the increase in the lattice spacing that forces the onset of hexagonal structure, rather than the other way around.

The actual lattice parameters can be obtained from the parameter d as follows:

$$\text{zinc blende: } a = \sqrt{2}d, \tag{2}$$

$$\text{wurtzite: } a = d, \quad c = (\tfrac{8}{3})^{1/2}d. \tag{3}$$

It should be noted that, apart from its fundamental importance, the precise knowledge of the lattice parameter is of considerable practical interest in that it provides a convenient determination of crystal composition in ternary alloys. It is a fortunate circumstance in this respect that in the case of all $A^{II}_{1-x}Mn_xB^{VI}$ alloys the linear variation of the lattice parameter with x is very large, making this approach to the determination of composition quite reliable.

1.2 Band Structure of DMS Crystals in the Absence of a Magnetic Field

1.2.1 General Band-Structure Features

The band structure of the $A^{II}_{1-x}Mn_xB^{VI}$ DMS alloys can be presented in a unified way by stressing four underlying features:

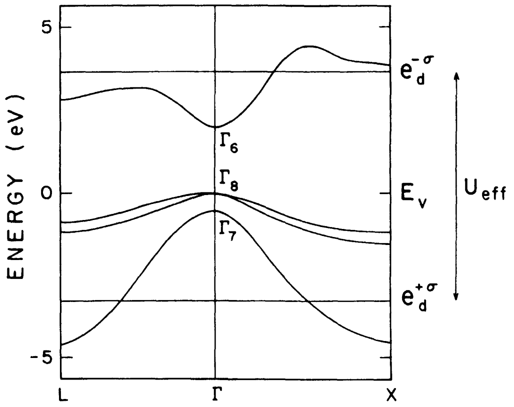

FIG. 2. Schematic band structure for a zinc-blende $A^{II}_{1-x}Mn_xB^{VI}$ alloy. The behavior of the *sp* bands is qualitatively the same as in zinc-blende $A^{II}B^{II}$ semiconductors. The Mn levels are spin split at each site by the energy U_{eff} ($\approx$7 eV). The effect of *p-d* hybridization and of the tetrahedral crystal field on the Mn levels is neglected in this schematic presentation, since the effect is small on the scale of the figure. For wurtzite $A^{II}_{1-x}Mn_xB^{VI}$ alloys the band structure would be qualitatively the same except that the degeneracy seen at Γ_8 in the figure would be removed, resulting in a slight valence-band splitting ($\approx$20 meV) at the Γ point.

1. Qualitatively, the band structure of the $A^{II}_{1-x}Mn_xB^{VI}$ alloys closely resembles that of the $A^{II}B^{VI}$ materials having the same crystal structure. In particular, these alloys are direct-gap semiconductors, with the band extrema occurring at the Γ point. This is illustrated by Fig. 2, where the curves show the characteristic energy (E) vs wave vector k dispersion typical for a zinc-blende $A^{II}B^{VI}$ semiconductor. These bands, common to the $A^{II}B^{VI}$ compounds and to the $A^{II}_{1-x}Mn_xB^{VI}$ alloys alike, originate from the s and p orbitals of the constituent atoms, and we shall refer to them as *sp* bands.
2. Quantitatively, the band structure of the $A^{II}_{1-x}Mn_xB^{VI}$ material transforms in a smooth manner from that of the parent $A^{II}B^{VI}$ compound to the band structure of the tetrahedrally bonded MnB^{VI} "hypothetical" binary compound. Although this evolution is not strictly linear with x, a linear variation (as would be predicted by the virtual-crystal approximation; see, e.g., Hass and Ehrenreich, 1983), is a good first-order guide for a quantitative description of the dependence of the energy gap and other band parameters on x. The band structures of the parent $A^{II}B^{VI}$ binary compounds are well known. Although tetrahedrally bonded MnB^{VI} compounds either do not exist in nature (e.g., MnTe), or are of such poor quality that electronic measurements on these com-

pounds are not reliable, band structures of zinc-blende or wurtzite MnS, MnSe, and MnTe can be established by the various computational schemes commonly used in band-structure calculations. Furthermore, with the advent of epitaxy (see Sec. 1.3), it is now possible to grow some transition-metal chalcogenides (e.g., MnTe, FeSe) in thin-film form using this nonequilibrium growth technique, giving access to band-structure information of these materials. The band structures thus obtained can then serve as the $x = 1$ limit for interpolation between the $A^{II}B^{VI}$ and MnB^{VI}. This procedure, although not always precise, nevertheless serves as a very useful guide for estimating the properties of $A^{II}_{1-x}Mn_xB^{VI}$ alloys for $0 < x < 1$.

3. Superimposed on the above band-structure characteristic of the $A^{II}B^{VI}$ semiconductors is the effect of the $3d^5$ shell of Mn. The energy of the narrow band originating from the ground state of $3d^5$ electrons is approximately 3.5 eV below the valence-band edge (see, for example, Franciosi *et al.*, 1985), i.e., relatively far from the band edges that determine basic properties of a semiconductor. This level is marked as $e_d^{+\sigma}$ in Fig. 2. It is interesting to note that the position of this manganese-derived state is rather insensitive to the II-VI semiconductor "parent." The location of an excited state that corresponds to placing an extra electron on the $3d$ shell is less certain ($e_d^{-\sigma}$ in Fig. 2). Inverse photoemission experiments (i.e., adding a sixth electron to a given Mn site) appear to indicate that it lies some 3.5 eV above the top of the valence band in $Cd_{1-x}Mn_xTe$ (Wall *et al.*, 1989), but more work is needed to establish this conclusively.
4. While the knowledge of the positions of both $e_d^{+\sigma}$ and $e_d^{-\sigma}$ bands, described above, is extremely important for formulating *p-d* hybridization (which underlies both *sp-d* and *d-d* exchange interactions, to be discussed later), it must be emphasized that these levels are determined in the context of a removal of an electron ($3d^5 \rightarrow 3d^4$) from, or of an addition of an electron ($3d^5 \rightarrow 3d^6$) to, a given Mn site. These transitions should not be confused with intra-ion transitions, involving a spin flip within the $3d^5$ level itself. The latter transitions, occurring at energies around 2.1 eV, are of fundamental importance in the optical properties of the $A^{II}_{1-x}Mn_xB^{VI}$ alloys, as will be discussed in Sec. 1.2.3.

In the remainder of this section we focus on the direct-gap and on the intra-Mn transitions, since these features are of primary importance in determining the optical and electrical properties of DMSs in zero magnetic field. For a discussion of the general band structure, including points far from the Γ point and deep-lying bands, we refer the reader to the theoretical and experimental papers dealing explicitly with the subject (see, for example, Larson *et al.*, 1988).

1.2.2 The Energy Gap and its Dependence on Composition All $A^{II}_{1-x}Mn_xB^{VI}$ alloys are direct-gap semiconductors, like their $A^{II}B^{VI}$ parent materials, as is shown in Fig. 3 for three groups of alloys: wurtzite, zinc-blende open gap, and zinc-blende zero gap. As in nonmagnetic semiconductors, the valence band for the zinc-blende case consists of light and heavy holes. For open-gap zinc-blende materials, the heavy-hole and the light-hole bands are degenerate at $k = 0$. The closed-gap zinc-blende case corresponds to a degeneracy of the heavy hole and the conduction band at $k = 0$. For the wurtzite crystals, these degeneracies are lifted by the lower symmetry of this structure.

The effect on the energy gap of replacing the atoms of the group-II element by Mn is illustrated by the case of $Cd_{1-x}Mn_xTe$. As Mn is substituted for Cd in the CdTe lattice, the energy gap increases, to the point that the originally opaque material eventually becomes transparent to visible light. Figure 4 shows the values of the fundamental optical energy gap determined experimentally in CdMnTe at several temperatures. The saturation observed for $x > 0.4$ results from very efficient optical transitions within the Mn^{2+} ions, whereas the gap variation is given by the position of the feature labeled A, observed by Lee and Ramdas (1984) in piezoreflectivity experiments. This technique permits detection of valence-to-conduction-band transitions even in the presence of strong intra-Mn^{++} absorption lines. The latter optical transitions within Mn^{++} ions of course also take place for $x < 0.40$, but for those concentrations they are completely obscured by the fundamental band-edge absorption. As

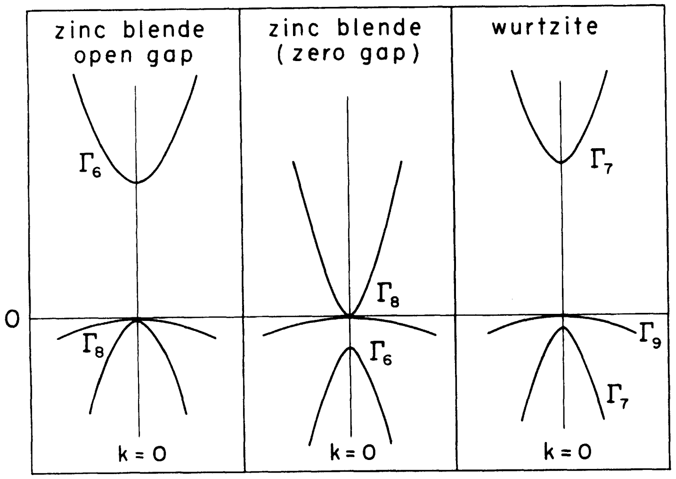

FIG. 3. Band structure near the Γ point for three cases of interest: an open-gap zinc-blende semiconductor (e.g., $Cd_{1-x}Mn_xTe$, or $Hg_{1-x}Mn_xTe$ for $x > 0.07$); a zero-gap zinc-blende semiconductor (e.g., $Hg_{1-x}Mn_xTe$ for $x < 0.05$); and a wurtzite semiconductor (e.g., $Cd_{1-x}Mn_xSe$). In the wurtzite case, the $\Gamma_9 \rightarrow \Gamma_7^{(c)}$ and $\Gamma_7^{(v)} \rightarrow \Gamma_7^{(c)}$ transitions determine the optical absorption edge for the **E** $\perp$ **c** and the **E** $\parallel$ **c** polarizations, respectively, where **E** is the electric field of incident light and **c** is the hexagonal axis. This is the basis of dichroism observed in wurtzite crystals.

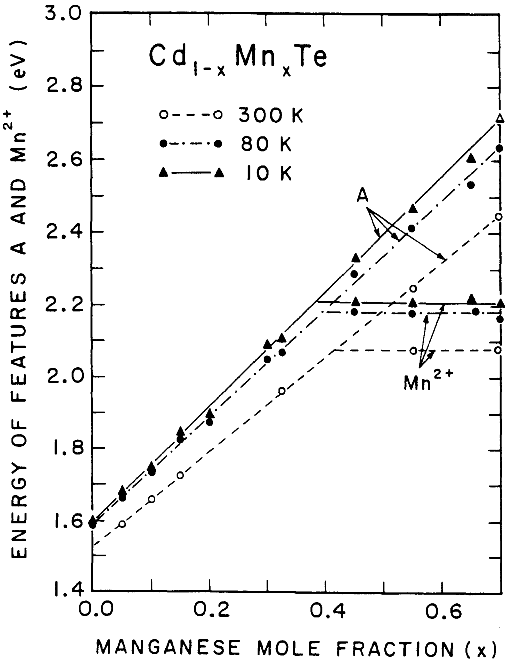

FIG. 4. The energy gap (obtained from the free-exciton transition, marked "A" in the figure) and the $^6A_1 \rightarrow {}^4T_1$ intra-Mn transition (marked "Mn^{2+}") as functions of Mn concentration x for $Cd_{1-x}Mn_xTe$ for three temperatures. Note the qualitative similarity in the temperature dependence of both types of transitions (from Lee and Ramdas, 1984).

we can see from the figure, the variation of the fundamental band gap with concentration in diluted magnetic semiconductors can be described, in first approximation, by a linear dependence. When a linear extrapolation to $x = 1$ is made, the value of the energy gap in hypothetical zinc-blende MnTe is obtained. Its value at 4.2 K is 3.2 eV. The same extrapolated value is obtained in the case of $Zn_{1-x}MnTe$ and $Hg_{1-x}Mn_xTe$.

Such "tunability" of the energy gap by composition has an especially profound effect in the case of Hg compounds. Consider $Hg_{1-x}Mn_xTe$ as an example. In this case the role of Mn is similar to that of Cd in $Hg_{1-x}Cd_xTe$. For low values of x this alloy is a zero-gap semiconductor, like the "parent" HgTe. At several atomic percent of Mn ($x \approx 0.07$ at 4.2 K; $x \approx 0.05$ at room temperature) the energy gap of $Hg_{1-x}Mn_xTe$ "opens," and for larger values of x the alloy becomes an ordinary (i.e., positive gap) semiconductor, the gap being extremely sensitive to x, very much like in $Hg_{1-x}Cd_xTe$. A great deal of interest exists in this feature because of potential applications of variable narrow-gap semiconductors as infrared detectors.

The energy gap in DMSs—as in other semiconducting compounds—exhibits considerable temperature dependence, such that the energy gap tends to open wider as the temperature decreases. This behavior is similar to that characteristic of the parent $A^{II}B^{VI}$ compounds, with the result that in the general vicinity of the optical band edge the $A^{II}_{1-x}Mn_xB^{VI}$ alloys tend to be more transparent to shorter and shorter wavelengths as the temperature is decreased. The temperature variation for Cd_xMn_xTe is illustrated in Fig. 4. Detailed studies indicate this variation to be roughly linear between 77 and 300 K, a rather useful fact for making interpolations to intermediate values of T (Furdyna and Kossut, 1988, Chap. 2). In Table 2 we summarize the energy-gap values for the $x = 0$ and $x = 1$ limits of the II-VI DMS alloys for three temperatures. The possibility of tuning the band gap of DMSs, a feature characteristic for many other semiconducting alloy systems, represents an advantageous feature, since by adjusting the composition one can achieve a gap value required by particular application needs. For an extensive discussion of the band structure and references, see Chapter 2 in Furdyna and Kossut (1988).

1.2.3 Intra-ionic Optical Transitions in $A^{II}_{1-x}Mn_xB^{VI}$ Alloys The $A^{II}_{1-x}Mn_xB^{VI}$ systems differ from the $A^{II}B^{VI}$ compounds, and from "orthodox" II-VI ternaries such as $Cd_{1-x}Zn_xTe$, in that the 3*d* shell of the Mn^{++} ions is only half-filled. This leads to new intra-Mn^{++} electronic transitions, generally in the vicinity of 2.1 eV, which dominate the optical properties of $A^{II}_{1-x}Mn_xB^{VI}$ alloys at high values of x, and are responsible for their characteristic deep-red color in that range of compositions. Furthermore, these transitions within the 3*d* shell make wide-gap $A^{II}_{1-x}Mn_xB^{VI}$ materials, such as $Zn_{1-x}Mn_xSe$ and $Zn_{1-x}Mn_xS$, of considerable interest in the context of electroluminescent display devices.

Although the Mn^{++} *d* levels shown in Fig. 2 hybridize with the *sp* bands, and thus broaden into *d* bands, the degree of such broadening is in fact quite small. The description of 3*d* electronic states, therefore, can often be made in terms of atomic orbitals, modified by the presence of a crystal field of an appropriate symmetry. In particular, manganese with its exactly half-filled *d* shell forms a ground state that is an orbital singlet (orbital moment $L = 0$) denoted by ^6S, which does not split further in the presence of the crystal field nor by the spin-orbit interaction. In the presence of these interactions the ground state of *d* electrons in Mn ions is traditionally labeled as $^6A_1$. It is, of course, a spin sextet ($S = \frac{5}{2}$) (↑↑↑↑↑). The lowest-lying excited states, which correspond to flipping the spin of one of the five *d* electrons of manganese (↑↑↑↑↓), are subject to splitting by the crystal-field interaction (see ELECTRON LEVEL SPLITTING), as shown schematically in Fig. 5. The relative position of the crystal-field–split states was studied intensively in the past for various symmetries

Table 2. Energy gaps for $x = 0$ and $x = 1$ limits of the $A^{II}_{1-x}Mn_xB^{VI}$ alloys for liquid helium (LHe), liquid nitrogen (LN_2), and room (R.T.) temperatures (for references see for example Furdyna, 1988, and Kossut and Dobrowolski, 1993).

Compound	T	E_g (eV)	Compound	T	E_g (eV)
ZnS	LHe		HgS	LHe	<0
	LN_2			LN_2	~0
	R.T.	3.8–3.9		R.T.	>0
ZnSe	LHe	2.8	HgSe	LHe	−0.27
	LN_2			LN_2	−0.205
	R.T.			R.T.	+0.10
ZnTe	LHe	2.38	HgTe	LHe	−0.300
	LN_2	2.375		LN_2	−0.252
	R.T.	2.27		R.T.	−0.146
CdS	LHe		MnS	LHe	~3.3
	LN_2			LN_2	~3.0
	R.T.	2.45		R.T.	~3.0
CdSe	LHe	1.83	MnSe	LHe	3.3
	LN_2	1.82		LN_2	3.14
	R.T.	1.75		R.T.	2.9
CdTe	LHe	1.595	MnTe	LHe	3.2
	LN_2	1.586		LN_2	3.05
	R.T.	1.528		R.T.	2.9
			FeSe	LHe	~3.02
				LN_2	
				R.T.	

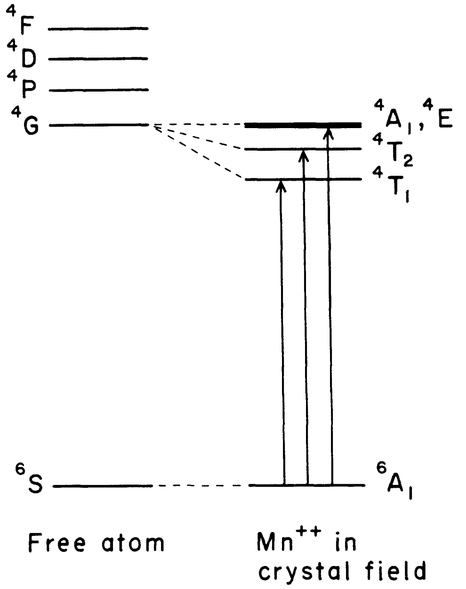

FIG. 5. A schematic diagram of the splitting of the lowest excited state of the $3d^5$ level (^4G) relative to the ground state (^6S) for a Mn^{++} ion in the presence of a tetrahedral crystal field. The arrows indicate possible intra-Mn transitions.

of the host material (see, e.g., Abragam and Bleaney, 1970).

Although optical transitions from the ground ^6S state to lowest excited states are forbidden for a free Mn ion, the spin-orbit interaction and the absence of the inversion symmetry may relax this selection rule when Mn ions are placed into a crystalline host matrix. In fact, optical transitions from the ground state ($^6A_1$) to excited states originating from the ^4G multiplet of the free Mn ion are very probable and correspond to very strong absorption of light. Of these, the $^6A_1 \rightarrow {}^4T_1$ is the lowest-energy (and therefore the most important) transition. It corresponds to approximately 2.1 eV, and constitutes in effect an optical absorption edge in crystals that contain a relatively high concentration of Mn ($x > 0.1$).

The energy of the $^6A_1 \rightarrow {}^4T_1$ transitions in relation to the energy gap is shown for $Cd_{1-x}Mn_xTe$ in Fig. 4, illustrating that either the energy gap or the $^6A_1 \rightarrow {}^4T_1$ transition determines the absorption edge (i.e., the onset of opacity), depending on which of these occurs at the lower energy. It is interesting to note that the temperature dependence of the energy gap and the intra-Mn^{++} transitions has the same qualitative character.

The transitions within the Mn 3*d* shell are the source of an efficient emission of light observed in the electroluminescence in $Zn_{1-x}Mn_xSe$ and $Zn_{1-x}Mn_xS$ (for a review of the energy-transfer mechanisms responsible for the electroluminescent properties of zinc-based DMSs see, e.g., Gumlich, 1981). Because of this intense emission, an application in flat-panel displays has been found for these materials.

Turning to Fe as the constituent of DMSs, let us note that—in contrast to the case of Mn—now even the ground state of the 3*d* electrons is subject to splitting by the crystal field of the $A^{II}B^{VI}$ matrix. This is because Fe, with its six *d* electrons, possesses a nonvanishing orbital moment $L = 2$. The lowest-lying state is in this case a singlet. Thus the $A^{II}_{1-x}Fe_xB^{VI}$ materials display paramagnetic behavior of the Van Vleck type. The existence of low-lying excited states of Fe completely changes the absorption spectra in the infrared of wide-gap DMSs containing iron, as compared with their Mn counterparts. While, e.g., $Cd_{1-x}Mn_xTe$ is transparent in these ranges of photon energies, compounds containing Fe ions possess very characteristic absorption bands occurring in the near infrared as well as in the far infrared. These absorption bands are due to optical transitions between various Fe levels split by the crystal-field and spin-orbit interactions. From such spectroscopic investigations of absorption by transition-metal impurities in various semiconductor matrices, a great deal of information concerning the crystal-field and spin-orbit interaction parameters can thus be obtained. More details about Fe as the constituent of DMSs can be found, e.g., in the review by Kossut and Dobrowolski (1993).

1.3 Epitaxial DMS Systems

As may be seen from Table 1, there are limits to the composition of DMS alloys in bulk crystal form (i.e., those grown by equilibrium crystal-growth methods). For example, the alloy $Zn_{1-x}Mn_xTe$ is zinc blende and can be obtained in single phase for $x < 0.86$, while $Cd_{1-x}Mn_xSe$ is wurtzite and has a composition range of $x < 0.50$. The higher composition ranges for such materials are not accessible in equilibrium bulk growth, because the equilibrium phases of the end-point materials (for example, rocksalt MnSe and NiAs-structure MnTe) are incompatible with the crystal structure of the parent semiconductor.

The development of epitaxial growth techniques—such as molecular-beam epitaxy (MBE)—allows us to transcend such limitations because of three factors: the nonequilibrium nature of the growth process, the low growth temperature, and the effect of the substrate. These factors allow the growth of crystal phases that would be unstable under typical bulk-growth conditions. The first example of such epitaxial "stabilization" was the growth of zinc-blende $Zn_{1-x}Mn_xSe$ epilayers over the composition range $0 < x \leq 0.66$ (see Kolodziejski *et al.*, 1985). (Bulk-grown $Zn_{1-x}Mn_xSe$ crystallizes in the zinc-blende phase up to $x = 0.3$ and is wurtzite for higher Mn content.) Further, the zinc-blende phase of pure MnSe was also stabilized in the form of thin, strained layers sandwiched between ZnSe layers (see Kolodziejski *et al.*, 1986). Perhaps the most striking example of MBE stabilization of new phases is the growth of single-phase zinc-blende $Cd_{1-x}Mn_xSe$ over the composition range $0 < x \leq 1.00$, even though both the end-point materials CdSe and MnSe have different stable structures for bulk growth (wurtzite and rocksalt, respectively; see Samarth *et al.*, 1989). Finally, both the binary magnetic semiconductors MnTe (Durbin *et al.*, 1989) and FeSe (Jonker *et al.*, 1988) have been grown in the zinc-blende phase by MBE.

2. EXCHANGE INTERACTIONS

As pointed out in the Introduction, DMSs such as $A^{II}_{1-x}Mn_xB^{VI}$ or $A^{II}_{1-x}Co_xB^{VI}$ differ from "normal" semiconductor alloys by the existence of two types of exchange interaction associated with the presence of the magnetic constituent in the lattice. The strong Kondo-type *sp-d* exchange between the spins of band electrons and the localized moments of the magnetic ions has a significant influence on electronic properties, while the weaker Heisenberg inter-ion exchange interaction (the *d-d* exchange) underlies the static and dynamic magnetic properties. Theoretical calculations have shown how the exchange interactions in the $A^{II}_{1-x}Mn_xB^{VI}$ alloys can be related to the band parameters $e_d^{+\sigma}$, U_{eff} (see Fig. 2), and a parameter V_{pd} that describes the degree of hybridization of the *p* and *d* orbitals of the alloy system (Ehrenreich *et al.*, 1987; Larson *et al.*, 1988). Similar calculations for exchange interactions in Co- and Fe-based DMSs are not yet available.

2.1 The *sp-d* Exchange Interaction

When a semiconductor contains localized magnetic moments (e.g., $Zn_{1-x}Mn_xTe$), its band structure will be modified by the exchange interaction of these moments (i.e., the $3d^5$ electrons) with band (i.e., *sp*) electrons. It is this *sp-d* exchange coupling that is responsible for most of the electronic peculiarities of the DMSs, e.g., for the greatly enhanced spin splitting of the band states in the presence of a magnetic field (see ELECTRON LEVEL SPLITTING). We start with a description of the physical origin and the form of the *sp-d* exchange interaction, and then relate the modifications of the band structure that result from this mechanism.

The *sp-d* exchange contribution to the band structure can be expressed formally by adding a new exchange term H_{ex} to the original Hamiltonian H_0, so that the total Hamiltonian H_T now reads

$$\begin{aligned} H_T &= H_0 + H_{ex} \\ &= H_0 + \sum_{\mathbf{R}_i} [J^{sp\text{-}d}(\mathbf{r} - \mathbf{R}_i)]\mathbf{S}_i \cdot \sigma, \end{aligned} \tag{4}$$

where $\mathbf{S}_i$ and σ are the spin operators for Mn^{++} and for the band electrons, respectively, $J^{sp\text{-}d}(\mathbf{r} - \mathbf{R}_i)$ is the electron–ion *sp-d* exchange-coupling constant, and $\mathbf{r}$ and $\mathbf{R}_i$ are the coordinates of the band electron and of the Mn^{++} ion, respectively. The summation is only over the lattice sites occupied by the Mn^{++} ions.

Two convenient approximations can be made to simplify H_{ex}. First, since the electronic wave function is very extended, so that the electron "sees" a large number of Mn^{++} ions at any time, we can make use of the molecular-field approximation, replacing $\mathbf{S}_i$ by the thermal average $\langle \mathbf{S} \rangle$ taken over all Mn^{++} ions. For paramagnetic systems, if the applied field $\mathbf{H}$ is in the z direction, we have

$\langle \mathbf{S} \rangle = \langle S_z \rangle$, a quantity intimately related to the magnetization of the system (see Sec. 4).

Second, and again because the electronic wave function spans a large number of lattice sites, we can replace $J^{sp\text{-}d}(\mathbf{r} - \mathbf{R}_i)$ by $xJ^{sp\text{-}d}(\mathbf{r} - \mathbf{R})$, where now $\mathbf{R}$ denotes the coordinate of every site of the fcc (or hcp) cation sublattice, with the summation now carried out over all R. We are thereby expressing the exchange interaction in terms of the virtual-crystal approximation. With these approximations, we have

$$H_{ex} = \sigma_z \langle S_z \rangle x \sum_{\mathbf{R}} J^{sp\text{-}d}(\mathbf{r} - \mathbf{R}), \tag{5}$$

with the summation extending over all cation sites. The major advantage of Eq. (5) is that in that approximation H_{ex} has the periodicity of the lattice. This allows us to use the same wave functions for solving the exchange-interaction problem as those used for diagonalizing H_0.

The *sp-d* exchange modification of the band structure will thus be determined by the values of the exchange constants $J^{sp\text{-}d}$, which are different for the Γ_6 and the Γ_8 bands. Traditionally, the exchange constant characterizing the interaction between the *d* electrons and the *s*-like Γ_6 electrons at the center of the Brillouin zone (the Γ point) is designated by the symbol α, and the exchange constant corresponding to the interaction between *d* electrons and the *p*-like Γ_8 electrons, also at the Γ point, is designated by β. The constant α originates only from ferromagnetic potential exchange and is hence positive. The parameter β has a larger magnitude than α and is negative. This has contributions mainly from antiferromagnetic kinetic exchange arising from the strong *p-d* hybridization of valence-band states. Calculations for the $A^{\text{II}}_{1-x}\text{Mn}_x B^{\text{VI}}$ alloys show that the parameter β is related to the band-structure parameters defined in the previous section through the Schrieffer–Wolff expression (Larson *et al.*, 1988)

$$N_0\beta = -32(V_{pd})^2[(e_d^{+\sigma} + U_{\text{eff}} - E_v)^{-1} + (E_v - e_d^{+\sigma})^{-1}], \tag{6}$$

where V_{pd} is the hybridization parameter and N_0 is the number of cations per unit volume.

The experimentally determined values of α and β for various DMSs are listed in Table 3. For comparison, we have also included the latest measurements of these constants in the Fe- and Co-based DMSs. Note that in both latter cases the *sp-d* exchange is significantly larger than in the $A^{\text{II}}_{1-x}\text{Mn}_x B^{\text{VI}}$ alloys.

The consequences of Eq. (5) have been extensively discussed in the literature. It is easy to see that the major consequence of the *s-d* and *p-d* interactions is to modify substantially the band structure of the parent semiconductor in the presence of an external magnetic field. We will illustrate this using the example of wide-gap $A^{\text{II}}_{1-x}\text{Mn}_x B^{\text{VI}}$ alloys. In this case the Zeeman splitting due to H_{ex} is much greater than the orbital (Landau) splitting and the spin splitting predicted by ordinary *sp* band theory, because in wide-gap materials the effective masses m^* are large and the "*sp* band" *g* factors are of the order of unity. The resulting Zeeman splitting of the conduction- and valence-band edges at the center of the Brillouin zone is schematically depicted in Fig. 6. Note that the splitting is qualitatively identical to "ordinary" spin splitting that takes place in nonmagnetic semiconductors (two spin levels for Γ_6, four for Γ_8), and in this sense the exchange contribution can be naturally absorbed into the concept of an effective *g* factor whenever the concept of a *g* factor is meaningful. The situation differs from the case of nonmagnetic semiconductors primarily by the fact that the magnitude of the spin splitting is now extremely large, that its be-

Table 3. Experimental *sp-d* exchange constants $N_0\alpha$ and $N_0\beta$ for the $A^{\text{II}}_{1-x}T_x B^{\text{VI}}$ alloys (in electronvolts) (N_0 is the number of cations per unit volume).

Alloy	$N_0\alpha$	$N_0\beta$	N_0 $(\alpha - \beta)$
$\text{Zn}_{1-x}\text{Mn}_x\text{S}$	. . .	. . .	. . .
$\text{Zn}_{1-x}\text{Mn}_x\text{Se}$	0.26	−1.11	1.37
$\text{Zn}_{1-x}\text{Mn}_x\text{Te}$	0.18	−1.05	1.23
$\text{Cd}_{1-x}\text{Mn}_x\text{S}$	0.22	−1.80	2.02
$\text{Cd}_{1-x}\text{Mn}_x\text{Se}$	0.26	−1.11	1.37
$\text{Cd}_{1-x}\text{Mn}_x\text{Te}$	0.22	−0.88	1.10
$\text{Hg}_{1-x}\text{Mn}_x\text{S}$	. . .	. . .	. . .
$\text{Hg}_{1-x}\text{Mn}_x\text{Se}$	0.40	−0.70	1.10
$\text{Hg}_{1-x}\text{Mn}_x\text{Te}$	0.40	−0.60	1.00
$\text{Zn}_{1-x}\text{Co}_x\text{Se}$	. . .	. . .	2.42
$\text{Cd}_{1-x}\text{Co}_x\text{Se}$	0.32	. . .	. . .
$\text{Zn}_{1-x}\text{Fe}_x\text{Se}$	0.22	−1.74	1.96
$\text{Cd}_{1-x}\text{Fe}_x\text{Se}$	0.225	−1.90	2.12

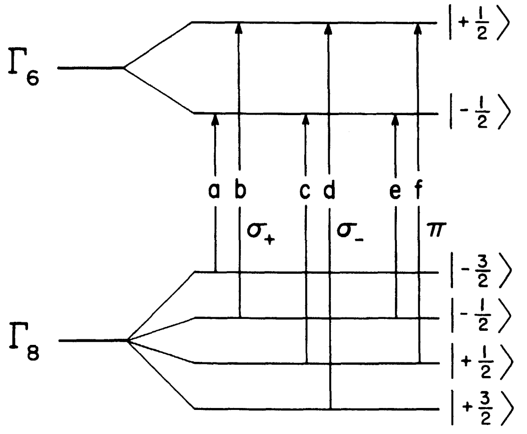

FIG. 6. A schematic picture of the spin splitting of the top of the valence band (Γ_8) and the bottom of the conduction band (Γ_6) for a wide-gap $A^{II}_{1-x}Mn_xB^{VI}$ alloy in a magnetic field. In the fourfold-degenerate valence band, $\pm\frac{3}{2}$ indicates heavy holes, $\pm\frac{1}{2}$ indicates light-hole states. The diagram also shows the electric-dipole–allowed transitions (arrows) for the two circular polarizations rotating transverse to the applied field (usually observed in the Faraday geometry, designated by σ_+ and σ_-), and for the linear polarization parallel to the field (observed in the ordinary Voigt geometry, designated by π). The Faraday-geometry transitions marked *a* and *d* are significantly stronger than those marked *b* and *c*.

havior is determined primarily by the properties of the magnetization (i.e., that it varies with temperature and is not necessarily linear in *H*), and that—in this approximation—it is almost completely independent of the band structure. However, the selection rules governing transitions between the levels shown in Fig. 6 remain the same as between spin levels of a nonmagnetic semiconductor.

As seen in the figure—and as already anticipated in the preceding discussion of α and β—the splitting of the heavy-hole band (indicated by $\pm\frac{3}{2}$) is considerably larger than for the conduction band, and is opposite in sign (i.e., the energy for spin-up electrons increases, and for spin-up heavy holes it decreases, when a magnetic field is applied). On the other hand, the behavior of light holes (indicated by $\pm\frac{1}{2}$ in the fourfold-degenerate Γ_8 band) is similar to that of electrons, in sign and in magnitude.

The Zeeman splittings in Fig. 6 can be conveniently represented by an effective *g* factor in the form

$$\Delta E = g_i^{\mathrm{eff}}\mu_{\mathrm{B}}B, \tag{7}$$

where the subscript *i* refers to the conduction or the heavy-hole band and g_i^{eff} is given by

$$g_e^{\mathrm{eff}} = g_e^* - N_0\alpha x\langle S_z\rangle/\mu_{\mathrm{B}}B, \tag{8}$$

$$g_{hh}^{\mathrm{eff}} = g_{hh}^* - N_0\beta x\langle S_z\rangle/\mu_{\mathrm{B}}B. \tag{9}$$

It can further be shown that for light holes

$$g_{lh}^{\mathrm{eff}} = g_{lh}^* + \tfrac{1}{3}N_0\beta x\langle S_z\rangle/\mu B. \tag{10}$$

Here g_i^* is the *g* factor obtained from the band structure without the exchange contribution (typically of the order of unity), *B* is the applied magnetic field, *x* is the atomic fraction of cations replaced by the magnetic atoms, and $\langle S_z\rangle$ is the thermal average of the *z* component of Mn^{++} spin. $\langle S_z\rangle$ is related to the dc magnetization *M* by

$$M = xN_0g_0\mu_{\mathrm{B}}\langle S_z\rangle, \tag{11}$$

where μ_{B} is the Bohr magneton and g_0 is the *g* factor of the magnetic ion (e.g., $g_{\mathrm{Mn}} = 2$). The terms containing α and β are strongly temperature dependent because of the presence of $\langle S_z\rangle$ [see Eq. (5)], and can be as large as 100 or more at low temperatures.

2.2 The *d-d* Exchange Interaction

The interaction between the magnetic constituents of the DMS lattice (e.g., the Mn–Mn interaction in the $A^{II}_{1-x}Mn_xB^{VI}$ family of ions) takes place via the *d* electrons localized on the magnetic ions, and is hence termed the *d-d* exchange. This interaction is expressed via the exchange integral J^{dd}, which determines the magnetic properties of the given DMS system (see Sec. 4).

Experimental measurements of the *d-d* exchange integral have been made using a wide variety of techniques—high-field magnetization, neutron scattering, Raman scattering, and low-field magnetic susceptibility. The results of measurements of the nearest-neighbor exchange *J* are summarized in Table 4. Once again, note the substantially larger values of the exchange integral in the Co–DMS alloys.

The *d-d* exchange in the $A^{II}_{1-x}Mn_xB^{VI}$ alloys has been calculated perturbatively by Larson *et al.* (1988), who showed that J^{dd} has

Table 4. Nearest-neighbor *d-d* exchange integrals for the $A^{II}_{1-x}T_xB^{VI}$ alloys (in kelvins) (for references, see for example Furdyna, 1988, and Kossut and Dobrowolski, 1993).

Alloy	J^{dd}
$Zn_{1-x}Mn_xSe$	−13.3, −13
$Zn_{1-x}Mn_xTe$	−9.5, −10.0
$Cd_{1-x}Mn_xS$	−10.6
$Cd_{1-x}Mn_xSe$	−7.9, −8.3
$Cd_{1-x}Mn_xTe$	−6.3, −6.9
$Hg_{1-x}Mn_xSe$	−10.9
$Hg_{1-x}Mn_xTe$	−7.15
$Zn_{1-x}Co_xS$	−47 ± 6
$Zn_{1-x}Co_xSe$	−54 ± 8
$Zn_{1-x}Fe_xSe$	−22
$Cd_{1-x}Fe_xSe$	−11.25
$Hg_{1-x}Fe_xSe$	−19
$Hg_{1-x}Fe_xSe$	−18

contributions from three classes of antiferromagnetic interactions: superexchange, the Bloembergen–Rowland interaction (in which the Mn–Mn exchange is mediated by valence-band holes), and RKKY-type interaction between the Mn^{++} *d* electrons and the conduction electrons. In these alloys, the dominant contribution arises from the hybridization-induced superexchange, which involves only the anion-derived upper valence-band states and the Mn *d* states. It may thus be viewed as being mediated by the *sp-d* exchange process (i.e., *d-d* superexchange can be symbolized as a *d-sp-d* interaction), which accounts for the fact that the α's and β's discussed in the preceding section are significantly larger than the J^{dd}'s. Surprisingly, this superexchange seems to be the dominant exchange mechanism even in the narrow-gap Hg compounds where one might have expected the Bloembergen–Rowland interaction to play a significant role (Lewicki *et al.*, 1988). A useful parametrization of the superexchange interaction can be made in terms of the band-structure parameters shown in Fig. 2 (Larson *et al.*, 1988):

$$J^{dd}(R) = -2(V_{pd})^4[(e_d^{+\sigma} + U_{\text{eff}} - E_v)^{-2}(U_{\text{eff}})^{-1} + (e_d^{+\sigma} + U_{\text{eff}} - E_v)^{-3}]f(R). \qquad (12)$$

In the wide-gap $A^{II}_{1-x}Mn_xB^{VI}$ alloys, the spatial dependence of the Mn–Mn exchange, $f(R)$, is well described by a Gaussian decay. The *d-d* interaction is primarily a nearest-neighbor interaction: for instance, in $Cd_{1-x}Mn_xTe$, the next-nearest-neighbor exchange constant J_2 is typically about 5 to 10 times weaker than the nearest-neighbor exchange J.

3. EFFECT OF *sp-d* EXCHANGE ON ELECTRONIC PROPERTIES

3.1 Bulk DMS Alloys

The *sp-d* exchange interactions influence physical phenomena that involve electrons in the conduction and valence bands (e.g., magnetotransport, interband and intraband magneto-optics), exciton levels (e.g., Faraday rotation), and impurity levels. This interaction has led to a multitude of new, often spectacular electrical and optical effects. We illustrate this with several representative examples. For a more comprehensive treatment of these effects, as well as for other illustrations, the reader is referred to Furdyna and Kossut (1988) or one of the review articles now available (e.g., Kossut and Dobrowolski, 1993).

3.1.1 Exchange Splitting of Exciton Transitions in DMSs Zeeman splitting of the free-exciton ground state provides one of the most direct and convenient situations for measuring exchange effects in wide-gap DMSs. For simplicity, we shall consider excitons in zinc-blende $A^{II}_{1-x}Mn_xB^{VI}$ alloys. The most important interband optical transitions are those between the band-edge states. They are shown in Fig. 6, together with electric-dipole–allowed transitions for various polarizations of light indicated by vertical arrows. Here σ_- and σ_+ denote transitions elicited by light that is circularly polarized in the plane perpendicular to the applied magnetic field (usually realized in the so-called Faraday geometry), and π corresponds to linear polarization, with the electric field of the photon parallel to the magnetic field (the ordinary Voigt geometry).

The experimentally observed magnetic field dependence of the exciton transition energies is shown in Fig. 7 for $Zn_{1-x}Mn_xTe$. Qualitatively, there is nothing unusual about the nature of the spectrum, in the sense that the same selection rules are involved as in a nonmagnetic semiconductor. What is unique is the very size of the effect—splittings of the order of 100 meV or more, to observe which

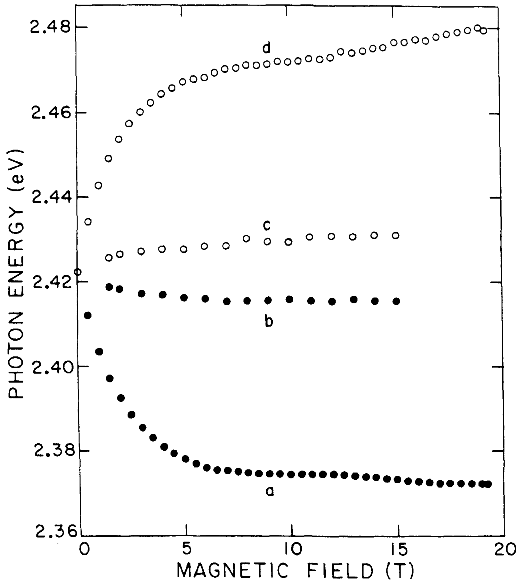

FIG. 7. Magnetic field dependence of the energies of transitions *a*, *b*, *c*, and *d* (see Fig. 6) of the 1*s* exciton in $Zn_{1-x}Mn_xTe$ ($x = 0.05$), observed at 1.4 K in magnetoreflectance in the Faraday configuration. Solid and open circles correspond to transitions observed with the σ_+ and σ_- circular polarizations, respectively (after Aggarwal *et al.*, 1986).

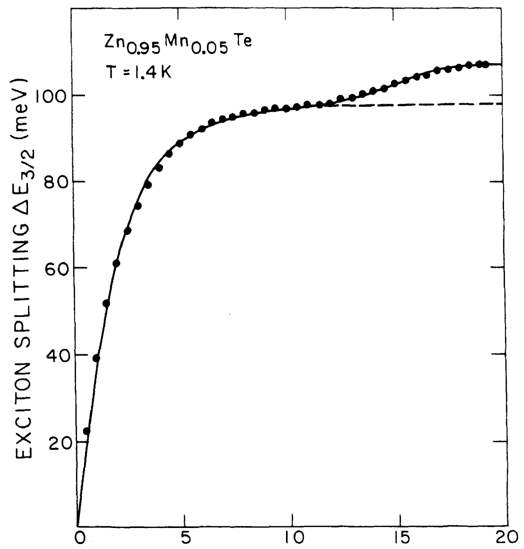

FIG. 8. Magnetic field dependence of the energy difference between the *d* and *a* transitions of the 1s exciton in $Zn_{1-x}Mn_xTe$ ($x = 0.05$), corresponding to the data in Fig. 7. Experimental data are shown by the points. The solid curve is the best theoretical fit to the data, including the contribution of Mn^{++}–Mn^{++} pairs. The dashed curve shows the Brillouin-function fit without the pairs (after Aggarwal *et al.*, 1986).

in a nonmagnetic semiconductor one would require fields of the order of a megagauss. Second, in practical terms the entire Zeeman splitting is caused by the exchange term H_{ex}. Because of the size of the splitting and the ease of measuring sharp exciton lines, this feature provides an exceptionally convenient experimental technique for determining α, β, and the details of magnetization M. This one-to-one correspondence between the splitting of excitonic features and the magnetization enables us to use optical measurements to study magnetic properties. This often is of considerable advantage, since optical measurements are very precise. Moreover, such measurements can be performed on samples whose magnetic properties are otherwise difficult to assess (e.g., because of their small size or mass, as in the case of ultrathin layers). This optical access to M has been exploited, for example, for observing the high-field magnetization steps (see Sec. 4.1.2) using exciton splitting as the measuring technique, as shown in Fig. 8. Measurements of the optical properties were also employed by researchers studying the spin-glass phase in DMSs, in this case exploiting the phenomenon of Faraday rotation (see, e.g., Rigaux *et al.*, 1986).

Zeeman splitting in wurtzite $A^{\mathrm{II}}_{1-x}\mathrm{Mn}_x B^{\mathrm{VI}}$ alloys behaves similarly, with the added minor complication introduced by the fact that the top of the valence band is no longer degenerate at the Γ point. This introduces a small splitting of the exciton spectrum at $H = 0$. Except for this, everything that has been said about zinc-blende $A^{\mathrm{II}}_{1-x}\mathrm{Mn}_x B^{\mathrm{VI}}$ materials applies equally well to the wurtzite members of the DMS family. A similar splitting is also present in DMSs in layer form, again because of the lowering of the cubic symmetry that automatically takes place in such structures as a result of tetragonal distortion via compressive or tensile strain within the layers.

For further details concerning excitons in DMSs, the reader is referred to the excellent and comprehensive treatment of this subject in Chapter 7 of Furdyna and Kossut (1988).

3.1.2 Giant Faraday Rotation in DMS Alloys The exceedingly large Zeeman splitting of the absorption edge and of the exciton level in wide-gap DMSs also results in a proportionately large difference in dispersion (i.e., a large difference in the index of refraction) for light of opposite circular polarizations in the Faraday geometry, particularly in the immediate vicinity of the band edge. This in turn provides the mechanism for Faraday rotation, which attains extremely large values (of the order of 1000°/cm-kG) at liquid-helium temperatures. The immense size of the effect is illustrated in Fig. 9. Because of its size and ease of measurement, Faraday rotation in DMSs—like the spin splitting of the exciton, to which it is closely linked—has frequently been used as a diagnostic tool to map the details of magnetization, both static and dynamic. Its very large size also holds promise of magneto-optical device applications, such as isolators, circulators, and other nonreciprocal devices.

Faraday rotation θ_F can be expressed in a convenient form for the temperature range where Curie–Weiss law holds [see Eq. (18) in Sec. 4.1.2]. This range includes room temperature, which is of particular importance from the point of view of quantitatively assessing the usefulness of this effect in nonreciprocal devices. Faraday rotation can then be expressed as

$$\theta_F = \frac{xPHl}{(T + \theta_0 x)} \frac{N_0(\beta - \alpha)\hbar^2\omega^2}{\{[E_g(0) + x\Delta E_g]^2 - \hbar^2\omega^2\}^{3/2}}, \tag{13}$$

where θ_0 is the Curie–Weiss temperature defined in section 4.1.2, ΔE_g equals $E_g(1) - E_g(0)$, $E_g(1)$ and $E_g(0)$ are the energy gaps for $x = 0$ and $x = 1$, respectively (see Table 2), and P is a constant. In Eq. (13), θ_F is in degrees, T and θ_0 in kelvins, the energies $N_0(\beta - a)$, $\hbar\omega$, and E_g in electronvolts, H in gauss, and l in centimeters. The constant P has the value

$$P = (35 \pm 2) \text{ K deg/G cm},$$

obtained empirically, and is in principle in-

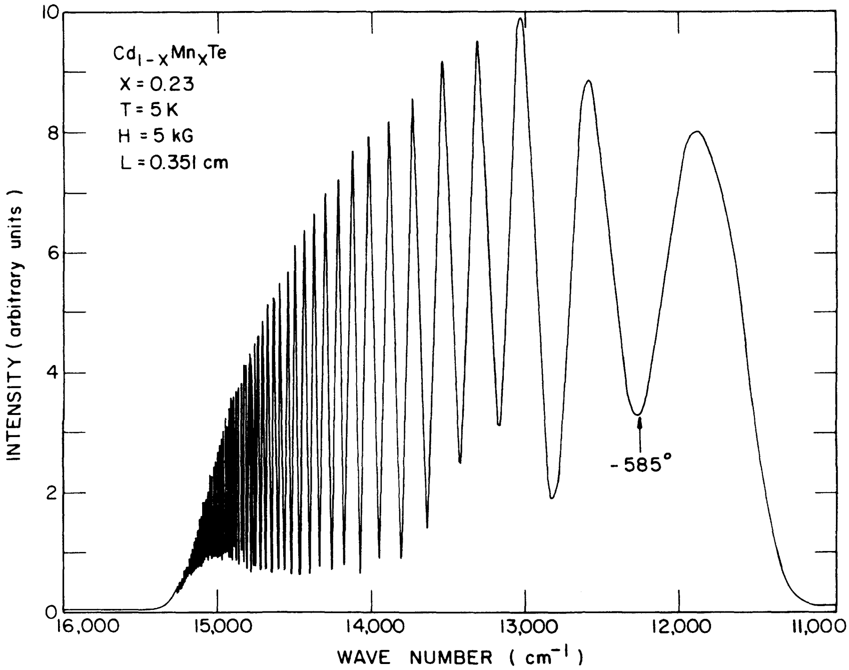

FIG. 9. Transmission as a function of wave number of initially linearly polarized light through a $Cd_{1-x}Mn_xTe$ ($x = 0.23$) slab, sandwiched between two linear polarizers, displaying giant Faraday rotation. At 15 000 cm^{-1} the rotation exceeds −8000 degrees. The experiment is carried out at 5 K and 5 T (after Bartholomew *et al.*, 1986).

dependent of the material. Note, however, that an additional dependence of θ_F on temperature will follow from the temperature dependence of E_g (see Table 2).

3.1.3 Acceptors and Negative Magnetoresistance in *p*-Type $\mathbf{Hg_{1-x}Mn_xTe}$ One of the most striking effects of the *sp-d* exchange is the negative magnetoresistance exhibited by *p*-type $Hg_{1-x}Mn_xTe$ (Mycielski and Mycielski, 1980). This phenomenon is due to the mixing of light- and heavy-hole band contributions to the acceptor wave function in the presence of a magnetic field. The admixture arises as follows. In the presence of a magnetic field, the exchange interaction leads to an anisotropy of the constant-energy surfaces associated with the four spin-split components of the valence band near the Γ point. A $\mathbf{k}\cdot\mathbf{p}$ perturbation calculation yields

$$E_{\pm 3/2}(k) = -\tfrac{1}{2}\hbar^2\left[\left(\frac{3}{4m_{lh}} + \frac{1}{4m_{hh}}\right)(k_x^2 + k_y^2) + \frac{1}{m_{hh}}k_z^2\right] \pm 3B, \qquad (14)$$

$$E_{\pm 1/2}(k) = -\tfrac{1}{2}\hbar^2\left[\left(\frac{1}{4m_{lh}} + \frac{3}{4m_{hh}}\right)(k_x^2 + k_y^2) + \frac{1}{m_{lh}}k_z^2\right] \pm B, \qquad (15)$$

where m_{lh} and m_{hh} are the effective masses of heavy and light holes and $B = \frac{1}{6}N_0\beta x\langle S_z\rangle$.

The acceptor ionization energy in $Hg_{1-x}Mn_xTe$ depends on x, and for concentrations of interest here ($x \approx 0.15$) is about 10 meV in the absence of magnetic field (see, e.g., Mycielski, 1983; Furdyna and Kossut, 1988, Chap. 8). For B negative, the valence subband that is closest to the acceptor level is the $E_{-3/2}(k)$ sublevel. This spin sublevel is characterized by a large longitudinal mass (m_{hh}) and a small transverse mass ($\frac{4}{3}m_{lh}$). As the magnetic field is increased, the other subbands move further away from the acceptor level (spin splittings between the subbands can be considerably larger than 10 meV), so that the acceptor wave function becomes increasingly dominated by the properties of $E_{-3/2}(k)$. Thus, with increasing magnetic field, the acceptor Bohr orbit acquires an increasing admixture of the light-hole mass. Because of the decrease in the mass determining the Bohr orbit, the acceptor wave function expands spatially in the direction perpendicular to the applied magnetic field. This leads to a decrease in the ionization energy of the acceptor, hence increasing the free-carrier concentration. The increased spatial extent of the acceptor wave function also results in greater impurity conduction (i.e., hopping). As a result of the disklike shapes of the acceptor wave functions, the effects arising from the swelling of the Bohr orbit are anisotropic. These consequences are manifested in the form of a dramatic negative magnetoresistance of nearly an order of magnitude per tesla at low temperatures, which is anisotropic with respect to the direction of H (Fig. 10) (Wojtowicz and Mycielski, 1983).

3.1.4 Magnetic-Field–Induced Metal–Insulator Transition The negative magnetoresistance described above is also exhibited by *n*-type $Cd_{1-x}Mn_xSe$ (see, e.g., Shapira, 1987). The effect here is not as well understood as in the case of *p*-type $Hg_{1-x}Mn_xTe$, but is obviously very different, since it involves shallow donors and the conduction band. The ionization energy of the donors is in this case typically about 20 meV, and the Zeeman splitting of the donor level is expected to follow closely the behavior of the conduction-band edge. It is likely, however, that the negative magnetoresistance in $Cd_{1-x}Mn_xSe$ arises from the expansion of the impurity wave function in increasing magnetic field due to the destruction of magnetic polarons.

In certain ranges of impurity concentration, the negative magnetoresistance in *p*-$Hg_{1-x}Mn_xTe$ and *n*-$Cd_{1-x}Mn_xSe$ leads to a magnetic-field–induced insulator-to-metal transition (Mott transition) (Wojtowicz *et al.*, 1986). Unlike the phenomenon in ordinary semiconductors, the transition from an insulating phase to a metallic phase occurs in increasing magnetic field. Above a critical field H_c, the conductivity displays the form (Wojtowicz *et al.*, 1986)

$$\sigma(T,H) = \sigma(0,H) + A(H)T^{1/2}. \qquad (16)$$

In other words, the conductivity at absolute zero remains finite, indicating metallic behavior. The values of $\sigma(0,H)$ for *p*-$Hg_{1-x}Mn_xTe$ and *n*-$Cd_{1-x}Mn_xSe$ are shown in Fig. 11. In both these systems, the metal–insulator transition is the direct result of the

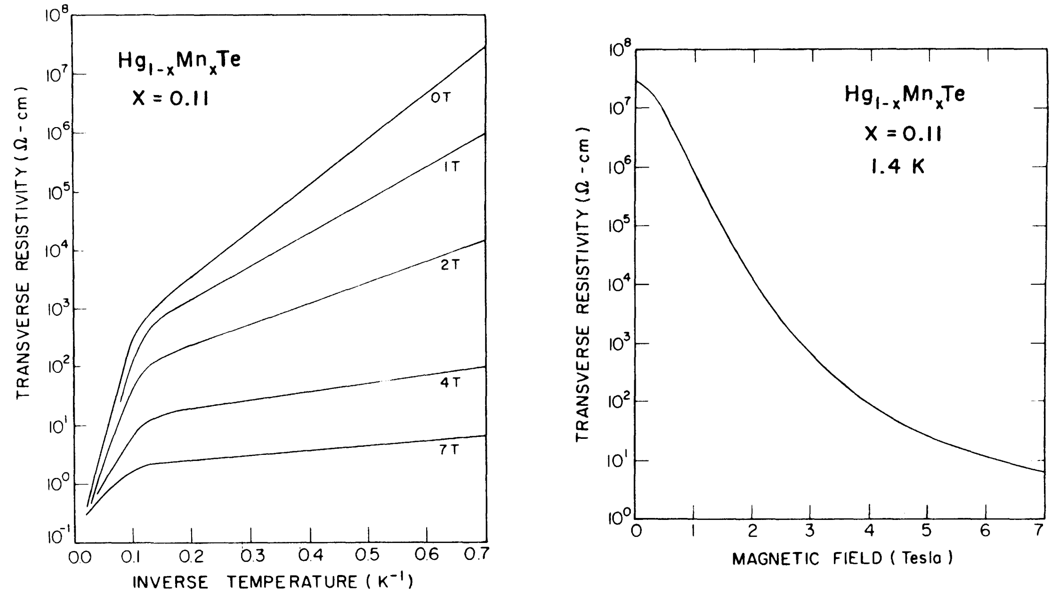

FIG. 10. Large negative magnetoresistance in the impurity-conduction region of p-type $Hg_{1-x}Mn_xTe$ ($x \approx 0.11$, acceptor concentration $\approx 10^{16}$ cm^{-3}) (after Wojtowicz and Mycielski, 1983).

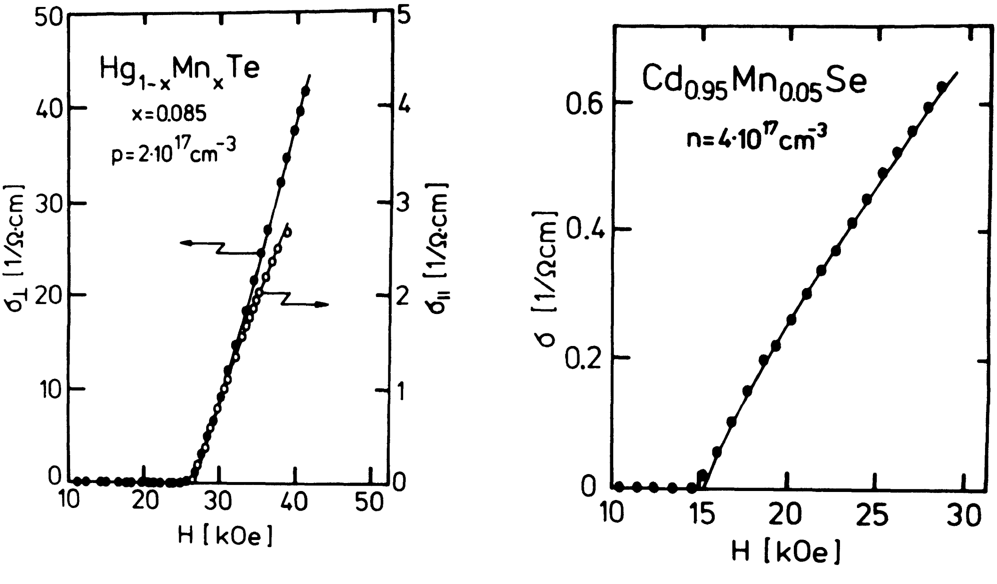

FIG. 11. Magnetic field dependence of the extrapolated zero-temperature conductivity for p-type $Hg_{1-x}Mn_xTe$ and n-type $Cd_{1-x}Mn_xSe$. The $Hg_{1-x}Mn_xTe$ shows that the transverse and longitudinal magnetoconductivities, although different, yield the same value of critical field. (After Wojtowicz *et al.*, 1986.)

giant spin splitting produced by the *sp-d* exchange. Above the critical field H_c, this splitting produces a redistribution of carriers between spin subbands that are effectively decoupled because of a Coulomb gap. The metal–insulator transition is then determined entirely by delocalized carriers within a single spin-polarized subband. This is a unique property of the metal-insulator transition in DMSs, and has led to the definition of a new universality class (Wojtowicz *et al.*, 1986).

3.1.5 Magnetic Polarons All the above effects occur in the presence of external magnetic fields. We finally mention one that is observed when the magnetic field is zero: the bound magnetic polaron (BMP). The BMP phenomenon can be understood as follows: Consider, as an example, an electron bound in a donor level. The hydrogenlike orbit of the electron encloses a large number of Mn^{++} ions (~300 for $x \approx 0.05$), with which the electron interacts through spin–spin exchange. This interaction polarizes the Mn^{++} spins, leading to a finite magnetization on the scale of the donor orbit. In addition, the spontaneous fluctuations of magnetization can also lead to a finite magnetic moment on the same local scale. As a result of both contributions the bound electron perceives a finite magnetization, and energy is therefore required to flip the spin of the donor electron even when there is no external magnetic field. Since the first observation of bound magnetic polarons in *n*-type $Cd_{1-x}Mn_xSe$ (donors) and in $Cd_{1-x}Mn_xTe$ (acceptors), a great deal of attention has been given to this subject in the DMS context.

For additional details and a full theoretical treatment, the interested reader is directed to the review of the bound magnetic polaron problem in DMSs in Chapter 10 of Furdyna and Kossut (1988).

3.2 Quantum Wells and Superlattices

Semiconductor heterostructures owe most of their spectacular properties to our ability to deposit in an atomically coherent fashion a sequence of alternating ultrathin layers of different materials, characterized by different energy gaps (see, e.g., Bastard, 1988; HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR). Quantum wells and superlattices represent an extremely important example of such heterostructures. Because of the difference in the energy gaps E_g of the constituent materials, as well as the relative alignment of the band edges (the "band offsets"), electrons and holes become confined along the direction perpendicular to the layer plane. Among numerous semiconductor heterostructures fabricated so far, those incorporating DMSs are in a class by themselves, since they combine many fascinating aspects of semiconductor quantum structures (quantum confinement, efficient radiative recombination, etc.), with the unusual magnetic signatures characteristic of DMS systems discussed in the foregoing sections. For example, the *sp-d* interaction has exciting implications for quantum wells and superlattices consisting of nonmagnetic and DMS layers. Clearly most of the important properties of semiconductor heterostructures—such as localization of electron and hole states (see ELECTRON STATES: LOCALIZED), as well as optical and electrical properties involving those states—arise in one way or another from the existence of bandgap differences (band offsets) between adjacent layers. In DMS/non-DMS heterostructures the large Zeeman splitting of the band edges in the DMS layers thus provides a handle for "tuning" this relative band alignment over a significant scale (typically over several tens of meV) simply by varying an applied magnetic field. By tuning the band alignment, one then automatically has the means of dramatically changing the properties of the heterostructure during the course of the experiment. In this section we review several striking examples of such "Zeeman tuning," unique to DMS-based heterostructures.

3.2.1 Spin Segregation The ability to tune the band alignment by the application of a magnetic field underlies the idea that—under favorable circumstances—a magnetic field can be used to induce a spin-dependent potential in a DMS/non-DMS multilayer system, so as to form a "spin superlattice."

A "spin superlattice" (SSL) is a superlattice in which carriers with opposite spin states are confined in different layers. In order to achieve such spin modulation, we look for a structure in which the energy gaps of the constituent layers are initially (i.e., in the absence of magnetic field) equal, and the band offsets at the interfaces are initially zero, as shown at the top of Fig. 12 (where DMS regions are designated by shading). When a magnetic field is applied, the large Zeeman splitting of the band edges in the DMS layers results in induced band offsets and, consequently, in a spatial separation of the spin-up and spin-down states, as shown in Fig. 12 (bottom). The SSL phenomenon has already been observed experimentally in several DMS/non-DMS multiple-quantum-well systems (see, for example, Dai *et al.*, 1991).

One particularly attractive DMS system for achieving the structure described above—which will be used as an example in this section—is the ZnSe/$Zn_{1-x}Mn_xSe$ superlattice. The energy gap of $Zn_{1-x}Mn_xSe$ exhibits a rather striking bowing with Mn concentration x at low temperatures, first decreasing with x, and then rapidly increasing. It is thus possible to find a value of x at which $Zn_{1-x}Mn_xSe$ has the same energy gap as ZnSe (this occurs for $x \approx 0.04$ at 1.5 K). Furthermore, since the valence-band offsets at ZnSe/$Zn_{1-x}Mn_xSe$ interfaces are close to zero by the common-anion rule (which works reasonably well in wide-gap II-VIs), the value of x that gives equal energy gaps in ZnSe and $Zn_{1-x}Mn_xSe$ automatically leads to zero (or, in practice, to very small) band offsets in the conduction and the valence bands, thus satisfying the properties stipulated in the preceding paragraph.

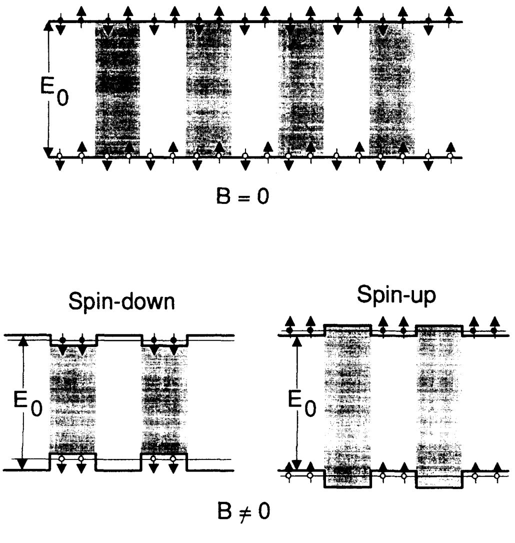

FIG. 12. A schematic diagram of the band structure of a magnetic-field–induced spin superlattice, consisting of DMS (shaded) and non-DMS (unshaded) layers with equal energy gaps at $B = 0$ (upper panel). The lower panel shows the band structure at $B \neq 0$ for spin-down (left) and spin-up (right) states of electrons and heavy holes. The Zeeman splitting in non-DMS layers is negligible. The spin states are indicated by arrows.

Examining Fig. 6 we note that, when the magnetic field is applied, the band edge of $Zn_{1-x}Mn_xSe$ for spin-down electrons will move down, and it will move up in energy for spin-down holes. The $Zn_{1-x}Mn_xSe$ layers thus become wells for holes and electrons of the spin-down orientation. Similarly, for spin-up states the conduction band moves up in energy, and the heavy-hole band moves down, so that for this spin orientation the ZnSe layers become the wells. This magnetic-field–induced band offset has therefore the consequence of spatially separating electron and hole states of different spin orientations. It is particularly convenient to observe the formation of such spin separation in magnetoabsorption experiments with the magnetic field applied perpendicular to the SSL layers (the Faraday geometry), in which only spin-conserving transitions are allowed ($\Delta S = 0$). Using left- and right-handed circular polarizations (designated by σ_L and σ_R), one can then observe separately transitions between spin-up valence- and conduction-band states (corresponding to σ_R), and those between spin-down states (σ_L).

Figure 13 (left) shows the magnetic field dependence of excitonic transitions observed with the σ_L and σ_R polarization in a ZnSe/$Zn_{0.96}Mn_{0.04}Se$ SSL, with DMS and non-DMS layers 112 Å wide. The figure illustrates very nicely the spatial separation of spins occurring in the SSL. Considering first the spin-down (σ_L) transitions, we see the rapid red-shift of the σ_L absorption line, since the initial and final states of the transitions are both in the DMS layers at finite magnetic fields, and follow the field dependence of the valence- and conduction-band edges of the DMS material (see Fig. 6). In contrast, the spin-up transitions take place between states localized in the ZnSe layers, which become the wells for this spin orientation when the field is applied. This is responsible for the flatness of the transition energy at higher fields—i.e., once the offset becomes well defined. The small initial rise in energy seen in

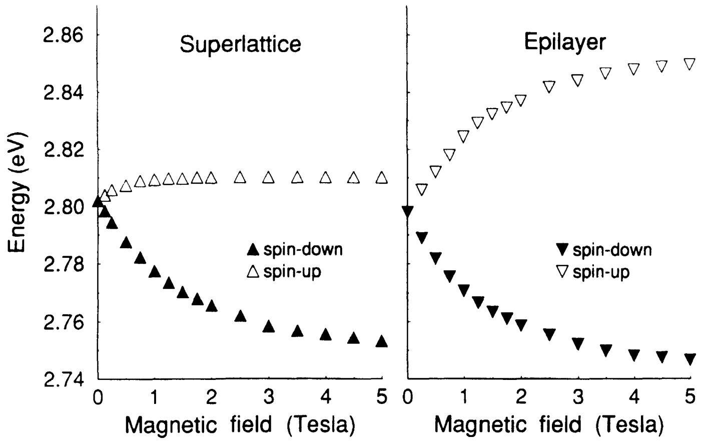

FIG. 13. Experimental energies of the Zeeman-split excitonic ground-state transitions in a $Zn_{0.96}Mn_{0.04}Se$/ZnSe (112 Å/112 Å) SSL for the two spin orientations (left panel). Note the striking asymmetry in the Zeeman shift for the spin-up and spin-down transitions relative to $B = 0$. The right panel shows the energies of the Zeeman splitting of the exciton transition observed on a $Zn_{2.96}Mn_{0.04}Se$ epilayer (after Dai *et al.*, 1991).

Fig. 13 at low fields (i.e., when the offset is just beginning to form) arises from the fact that in that region of extremely shallow wells the state in the ZnSe well is much more sensitive to the height of the (DMS) barrier, which increases with the field. As the barrier continues to increase, the energy of the state confined in the ZnSe well becomes less and less sensitive to the barrier height, and eventually flattens out, reflecting the behavior of the band edges in the nonmagnetic quantum well in which that state is now confined.

For comparison, we show on the right of Fig. 13 the position of free-exciton magnetoabsorption data observed for σ_L and σ_R polarizations in an epilayer grown with the same Mn concentration as the $Zn_{1-x}Mn_xSe$ layers of the SSL. Note that, in contrast with the SSL data, the σ_L and σ_R branches for the epilayer are quite symmetric. The asymmetry seen on the left-hand side is a characteristic signature observed in spin-separating structures.

3.2.2 Wave-Function Mapping One of the most important properties of semiconductor heterostructures is their ability to localize electrons and holes in space. Localization of carriers (as measured by their spatial wave-function distributions) is fairly obvious for low-lying states in deep quantum wells. It is an altogether different matter, however, for states at energies above the barriers, which may be confined in well layers or in barrier layers, or may not be localized at all. Furthermore, below-barrier states with energies close to the barriers (i.e., states in very shallow wells, or high-lying excited states) exhibit significant leakage into the barrier regions. Determination of the degree of leakage in the latter case, and of the region of localization in the case of above-barrier subbands, presents a major challenge.

One can meet this challenge by exploiting the fact that, in heterostructures consisting of DMS and non-DMS layers, the Zeeman splitting of a given state will reflect its weighted probability distribution over these two media. In other words, the Zeeman splitting will be determined effectively by how many localized magnetic moments the electron "sees". In this section we emphasize the application of this idea further, specifically as a tool for mapping the wave function (or probability) distribution in heterostructures.

Figure 14 shows photoluminescence from two single nonmagnetic quantum wells ($Zn_{1-x}Cd_xSe$, $x = 0.12$) of different widths, sandwiched between DMS ($Zn_{1-x}Mn_xSe$, $x = 0.20$) barriers and grown in a monolithic structure. The figure compares the photoluminescence peak positions for the two wells in $B = 0$, $B = 3$ T, and $B = 6$ T for the σ_L (spin-down) polarization. Note that for the narrower well (where the states lie nearer the top) the Zeeman redshift is considerably greater, indicating that the closer the energy of a given state is to the top of the well, the more its wave function penetrates into the barriers. By comparing the observed splitting with that exhibited by an epilayer grown from the barrier material itself, one can thus extract rather precise information on the de-

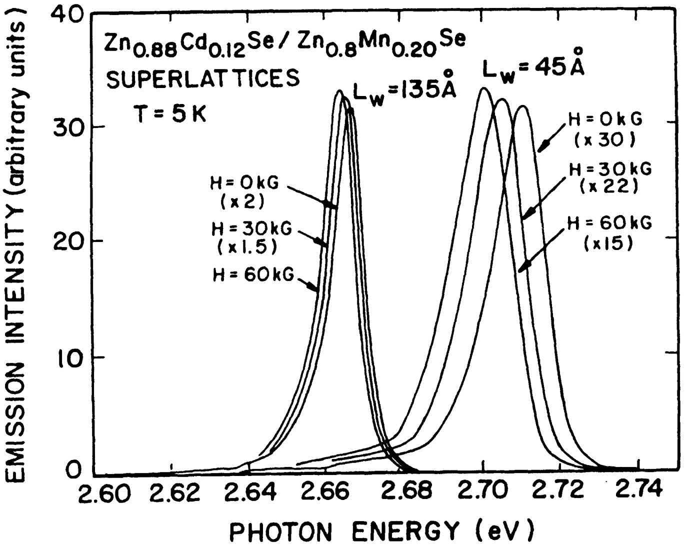

FIG. 14. Photoluminescence signal obtained from two $Zn_{0.88}Cd_{0.12}Se$ quantum wells of different thicknesses (135 and 45 Å) grown in a monolithic structure between thick $Zn_{0.8}Mn_{0.20}Se$ barriers (after Alonso *et al.*, 1991).

gree of penetration of the wave function into the DMS barrier.

This technique has been applied, for example, to pinpoint the localization of above-barrier subbands in superlattices. Using a type-I superlattice (i.e., a structure in which the same layer serves as the potential well for both electrons and holes), Zhang *et al.* (1992) have compared the Zeeman splitting of transitions between below-barrier states (which of course occur in the well layers) and between above-barrier states. On the basis of this comparison, they demonstrated that above-barrier states are localized in the barriers. Type-II superlattices are structures with a potential profile such that a given layer serves as the well for one type of carrier (say, electrons) and as the barrier for the other (say, holes). The ground-state transitions in such a structure are therefore spatially indirect. Using similar techniques it was also shown, however (Zhang *et al.*, 1993), that there exist spatially direct (i.e., type-I-like) excitons in type-II superlattices (formed by one state in the well and one above the barrier). The mechanism for localization of the above-barrier states has been theoretically shown (Luo and Furdyna, 1993) to arise from constructive interference of free-carrier waves at above-barrier energies, but the unambiguous experimental demonstration of how these states are distributed in space could only be accomplished by exploiting the differences in Zeeman splittings in the different layers, as described here.

3.2.3 Spin Lifetimes So far we have focused on the tunability of the band alignments in DMS/non-DMS heterostructures by means of an external magnetic field, and on the effect of such tuning on optical transition energies. In particular, focusing on spin-conserving transitions, we have discussed the band structure for the spin-up states separately from that for spin-down states, as if they represented physically different media. Of course the spin-split energy schemes coexist within the same sample (as in Fig. 12, lower panel), and this provides a unique opportunity for studying the lifetimes of excited spin states, as follows.

Consider an electron in the spin-up state (such as that residing in the non-DMS wells in the lower panel of Fig. 12). This electron can recombine radiatively with a spin-up hole in accord with the spin-conserving selection rule, but it can also make a nonradiative transition to a DMS well by means of a spin flip. The latter process for electrons—and the analogous relaxation of the holes to the energetically favored spin-down states in the valence band—thus compete with the radiative transitions. It is easy to see that, as the band offset increases, the nonradiative transition will become more and more probable, at the expense of the radiative spin-conserving process.

This process manifests inself dramatically in photoluminescence. If the sample is photoexcited so as to populate both spin states in the conduction band (as by linear polarization), the photoluminescence observed at $B = 0$ will contain equal intensities of the two circular polarizations. However, as the field is increased—i.e., as the field-induced energy difference between the spin-up and spin-down states becomes larger—the emitted photoluminescence will contain less and less of the higher-energy photons (the σ_R component). By analyzing the rate of decrease of this component as a function of spin splitting, one can obtain information about the spin lifetimes and, in appropriate geometries, also about the effect of spatial separation on these lifetimes (since in a heterostructure such as the SSL, a spin flip also involves a transfer to an adjacent layer). For more details we refer the reader to Jonker *et al.* (1993).

While photoluminescence contains such information in principle, extracting the values of the lifetimes is in this case rather indirect, involving a number of assumptions. An alternative—and more direct—approach is to use time-resolved methods for this purpose. In either case, the important point to emphasize is that Zeeman tuning provides a unique laboratory for such experiments, by its ability to separate the energies of the two spin states continuously, and over a significant range.

4. MAGNETIC PROPERTIES OF DMS SYSTEMS

We now present an overview of magnetic properties of the DMS alloys as they are presently understood. While the magnetic properties are interesting in their own right, it should be clear from the preceding section that they also bear on the electrical and optical properties of these DMS materials through the mechanism of the *sp-d* exchange, which implicitly involves the magnetization M (i.e., the parameter $\langle S_z \rangle$). This gives additional impetus for the need of obtaining a clear picture of magnetism in DMSs at both microscopic and macroscopic levels. The main role in the determination of magnetic properties of DMSs is played by electrons from the partially filled 3*d* shells of transition-metal atoms. It has to be remembered, however, that semiconductors that constitute host matrices of DMSs are themselves weak diamagnets. Although weak, this diamagnetism has to be included in a quantitative analysis of magnetic data.

4.1 Bulk DMS Alloys

4.1.1 Magnetic Phases Low-field static magnetic susceptibility measurements indicate that the $A^{\mathrm{II}}_{1-x}\mathrm{Mn}_x B^{\mathrm{VI}}$ alloys exist in different magnetic phases, depending on the Mn concentration x and on the temperature T. Figure 15 shows the magnetic phase dia-

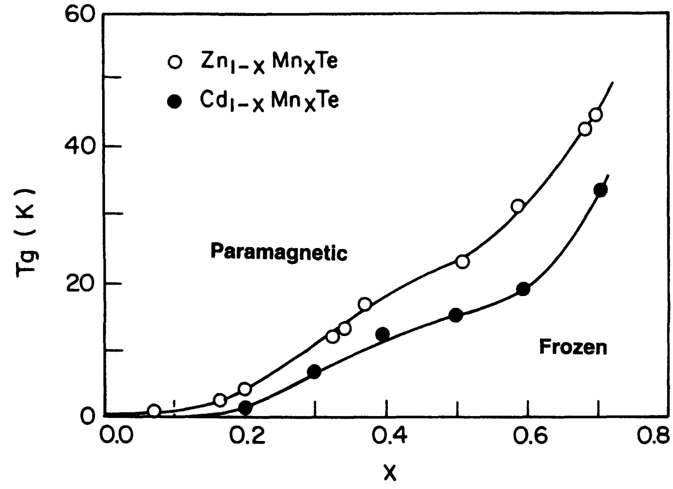

FIG. 15. Magnetic phase diagram for $Zn_{1-x}Mn_xTe$ and $Cd_{1-x}Mn_xTe$ (after Furdyna, 1988). The area above and to the left of each curve denotes the paramagnetic region for the respective materials; below and to the right corresponds to the frozen phase. In mapping out this diagram, the phase boundary is based on the occurrence of a cusp in the temperature dependence of the low-field static magnetic susceptibility. T_g denotes the temperature at which this cusp is observed.

gram for $Zn_{1-x}Mn_xTe$ and $Cd_{1-x}Mn_xTe$, which is qualitatively typical of all DMSs. The phase diagram consists of two regions: a high-temperature paramagnetic phase and a low-temperature frozen phase. The latter phase generally occurs when $x > 0.2$ (as suggested by the major portion of Fig. 15), but recent work has shown that spin freezing can also occur for lower values of x at very low temperatures.

In mapping this diagram, the phase boundary is based upon the occurrence of a cusp in the temperature variation of the low-field static magnetic susceptibility. Neutron-diffraction measurements on bulk DMS crystals (i.e., typically for $x < 0.6$) show that the low-temperature phase does not exhibit any long-range magnetic order in that range of concentrations. Assuming that the anomaly observed in the susceptibility corresponds to a genuine phase transition, we have to conclude that the low-temperature phase is a disordered but "frozen" state, such as a spin glass, with the possibility of short-range antiferromagnetic clusters.

We now proceed to discuss the susceptibility and magnetization behavior in the two phases shown in Fig. 15.

4.1.2 Magnetic Susceptibility of Paramagnetic Phase

In the very dilute limit the transition-metal ions can be regarded as isolated entities. In this range of crystal compositions magnetic properties of DMSs are governed by the sequence and the distances between the energy levels of $3d$ electrons in an environment of a specified symmetry. Since the ground state of d electrons in Mn ions is a spin sextet, $S = \frac{5}{2}$ (see Sec. 1.2.3), Mn-containing DMSs are simple paramagnets. However, the lowest-lying state of Fe d electrons in a semiconducting matrix is a singlet. Therefore in Fe-containing DMSs Van Vleck paramagnetism is observed. For example, in $A^{\mathrm{II}}_{1-x}Mn_xB^{\mathrm{VI}}$ the static magnetic susceptibility displays a rapid variation with temperature, which roughly corresponds to a $1/T$ dependence given by the Curie law

$$\chi = C_0 x/T, \tag{17}$$

where $C_0 = N_0(g\mu_B)^2 S(S+1)/(3k_B)$. However, the same amount of Fe in exactly the same semiconducting host is characterized by a susceptibility that is practically independent of temperature—a behavior typical for Van Vleck paramagnetism.

When the magnetic transition-metal atoms are incorporated in DMSs in larger proportions, deviations from such simple magnetic behavior become observable. They are connected with the fact that the probability of finding an isolated atom, i.e., with no nearest magnetic neighbors, becomes smaller and smaller, and an increasing number of magnetic ions must be viewed as being members of clusters. The mean size of such clusters grows with the amount of transition-metal atoms present in the crystal. However, at low field, the magnetization M is found to be linear in H, as in the dilute case, again allowing us to define the static susceptibility χ ($M = \chi H$). As shown in Fig. 16, experimental measurements of χ indicate that at high temperatures χ displays a Curie–Weiss behavior,

$$\chi = C(x)/[T - \theta(x)]. \tag{18}$$

The Curie–Weiss temperature $\theta(x)$ obtained

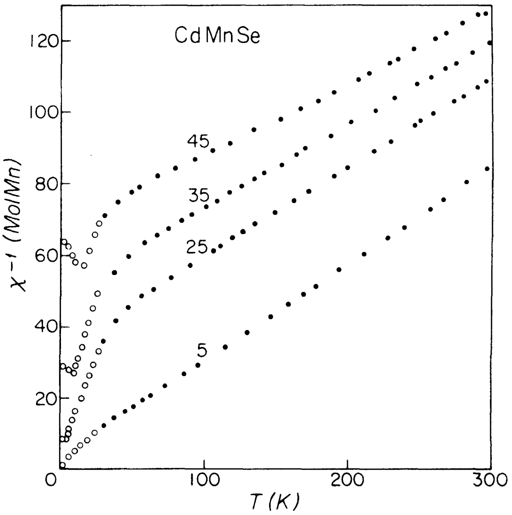

FIG. 16. Inverse magnetic susceptibility as a function of temperature for $Cd_{1-x}Mn_xSe$ of various mole fraction (given in atomic percent in the figure). The low-temperature data (open circles) were taken at 30 gauss for increasing temperature, after zero-field cooling. The high-temperature data (solid circles) were obtained at 8.5 kG. Note the linear Curie–Weiss behavior above 40 K, the characteristic downturn of χ^{-1} at lower temperatures, and the cusp seen in all but the $x = 0.05$ data, signaling the spin-glass transition (after Oseroff, 1982).

from experiment is negative, indicating antiferromagnetic interactions between the spins.

The high-temperature behavior of χ can be analytically derived by representing $A^{\mathrm{II}}_{1-x}\mathrm{Mn}_x B^{\mathrm{VI}}$ alloys as randomly dilute magnetic systems with antiferromagnetic Heisenberg interactions. The calculation of χ in the high-temperature limit yields (Spalek *et al.*, 1986) $C(x) = C_0 x$ [where C_0 is the Curie constant already defined in Eq. (17)], and

$$\theta(x) = x\theta_0 = -\tfrac{2}{3}xS(S+1)ZJ/k_{\mathrm{B}}. \qquad (19)$$

Here Z is the number of nearest neighbors ($Z = 12$ for zinc-blende and wurtzite DMSs), and J is the nearest-neighbor d-d exchange integral discussed in Sec. 2.2. Note, then, that J can be estimated from experimental measurements of $\theta(x) = \theta_0 x$; and, conversely, that the high-temperature χ can be calculated analytically for any x if J is known for a given alloy family.

For higher values of x at low temperatures, experimental results show a departure from the Curie–Weiss law, in the form of a downturn in the plot of χ^{-1} vs T, as seen in Fig. 16. This downturn is observed in all $A^{\mathrm{II}}_{1-x}\mathrm{Mn}_x B^{\mathrm{VI}}$ alloys. There has been no quantitative analytical explanation of this behavior. However, it can be qualitatively understood by including higher-order expansion terms in the calculation of χ.

4.1.3 Role of Antiferromagnetically Coupled Pairs At very high fields ($B > 10$ T) the magnetization of relatively dilute DMSs ($x < 0.15$) shows a steplike behavior as a function of field (Shapira *et al.*, 1986). This is a direct consequence of the existence of antiferromagnetically coupled nearest-neighbor pairs. In the case of Mn pairs in DMSs interacting via Heisenberg-type exchange, the energy-level scheme is particularly simple. It consists of a ladder of levels, each characterized by the total spin of the pair S_T ($S_T = 0, 1, 2, 3, 4, 5$). The separations between the levels are determined by the magnitude of the exchange constant J. Analytically, the energy-level scheme is given by

$$E = -J[S_T(S_T+1) - \tfrac{35}{2}] + g\mu_{\mathrm{B}}mH, \qquad (20)$$

where J is the nearest-neighbor Mn^{++}–Mn^{++} exchange integral; and for each value of S_T, m takes values $-S_T, -S_T+1, \ldots, S_T$. For $J < 0$, the ground state is nonmagnetic at zero field, with $S_T = 0$, and it remains the ground state as long as $g\mu_{\mathrm{B}}H/|J|$ is less than 2.

While this condition holds, pairs do not contribute to the total magnetization. However, once the value of $g\mu_{\mathrm{B}}H/|J|$ exceeds 2, the energy for $S_T = 1$, $m = -1$ becomes the ground state, as shown in the upper part of Fig. 17. Suddenly there is a contribution to the magnetization from all the pairs, and this results in a step in the magnetization (lower part of Fig. 17). Other such steps are similarly predicted for higher fields, as also shown in the figure. The magnetic fields at which these steps occur hence provide a direct measure of J. Determination of the exchange constants from the position of the magnetization steps so far has been done only for Mn atoms in various DMSs.

For completeness, it should be remarked that similar information from the energy-level spectrum of a pair of magnetic ions can be obtained from inelastic neutron scattering (Giebultowicz *et al.*, 1987). In fact, in the case of very large exchange constants (such as those in Co^{++} systems), where magnetic fields required to observe magnetization steps would be prohibitively large, inelastic neutron scattering is the primary and most precise source of such information.

4.1.4 Spin-Glass Transition and Short-Range Antiferromagnetic Ordering As we remarked earlier, the low-field magnetic susceptibility of the $A^{\mathrm{II}}_{1-x}\mathrm{Mn}_x B^{\mathrm{VI}}$ DMSs shows a well-defined cusp at a critical temperature T_g (as seen in Fig. 16), signaling a possible phase transition. This serves to define a magnetic phase diagram, showing the boundary between the high-temperature paramagnetic phase and the low-temperature frozen phase, shown in Fig. 15. The low-temperature phase is macroscopically disordered and displays many of the characteristics of a spin-glass state. In particular, the low-temperature specific heat has a linear temperature dependence and—at least in samples with $x < 0.65$—does not show any anomaly at T_g. Furthermore, remanence effects and irreversible phenomena are observed in magnetization studies below T_g. Although the precise nature of this phase remains a subject of

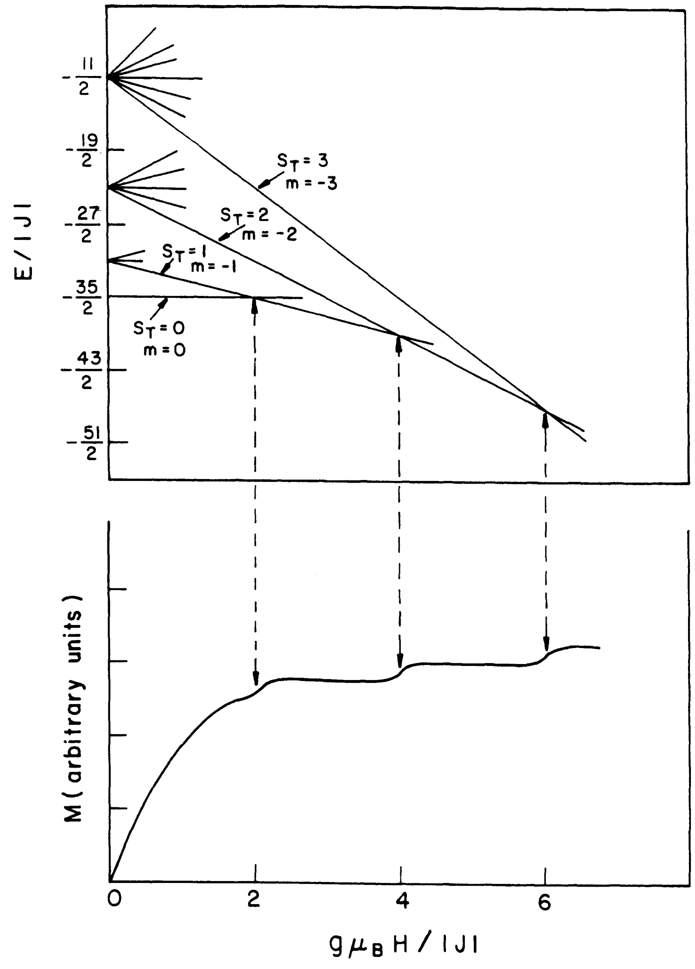

FIG. 17. Energy-level scheme for a pair of Mn^{++} ions, showing how steps occur in the magnetization at high fields. The relative size of the steps in *M* is exaggerated for clarity.

some controversy, we will for the present adhere to the tradition of calling it a spin glass, with the understanding that the rigorous magnetic classification of DMS alloys is yet to come.

The spin-glass behavior in DMSs has frequently been attributed to frustration of the antiferromagnetic interactions between the Mn^{++} ions, arising from the lattice topology of the $A^{II}_{1-x}Mn_xB^{VI}$ alloys. This is based on theoretical arguments showing that, for a diluted magnetic fcc lattice with nearest-neighbor interactions only, a spin-glass state becomes possible when the magnetic ion concentration exceeds a percolation threshold of $x \cong 0.19$. Indeed, an extrapolation in Fig. 15 of all experimental results obtained for temperatures above ~4.2 K tends toward $x \approx 0.2$. This feature has been traditionally taken as evidence that frustration is the underlying mechanism of spin-glass formation in DMSs. However, it is now clear that the spin-glass phase can form for arbitrarily small x at sufficiently low temperatures, as has indeed been observed for $Cd_{1-x}Mn_xSe$, $Zn_{1-x}Mn_xTe$, $Zn_{1-x}Mn_xSe$, and $Hg_{1-x}Mn_xTe$. For this reason, and because fcc and hcp lattices do in fact allow for antiferromagnetic order of the third kind (see below), we shall deemphasize the role of frustration as a primary mechanism of the spin-glass formation.

For the highest concentrations of transition-metal atoms attainable in bulk DMSs there are clear indications of magnetically ordered phases appearing at low temperatures. These conclusions were originally inferred from measurements of the susceptibility and the specific heat (where an anomaly

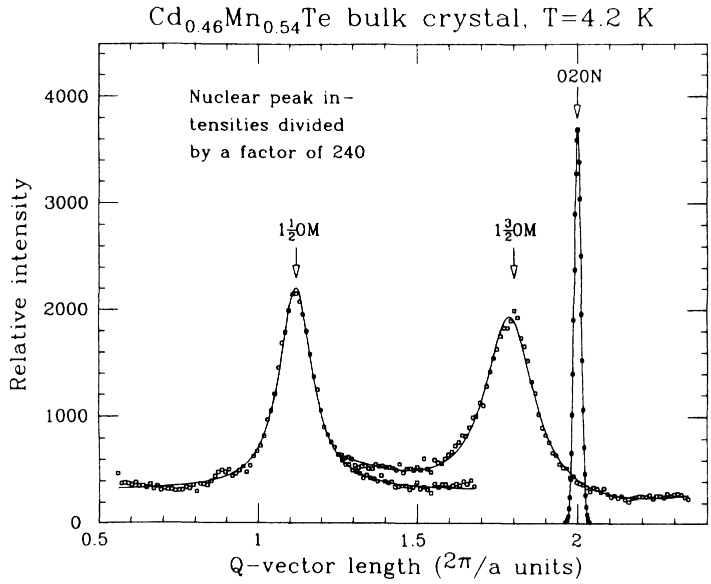

FIG. 18. Neutron-scattering data for a bulk $Cd_{1-x}Mn_xTe$ single crystal. Note that the magnetic peaks are much broader than the nuclear peaks, indicating short-range order.

is found at the transition temperature), and have been conclusively established by neutron-diffraction studies, where magnetic peaks were observed in $Cd_{1-x}Mn_xTe$, $Zn_{1-x}Mn_xTe$, and $Zn_{1-x}Mn_xSe$ (see Chap. 4 in Furdyna and Kossut, 1988). An example of such peaks is presented in Fig. 18 obtained for $Cd_{0.46}Mn_{0.54}Te$. The figure shows a sharp nuclear Bragg peak corresponding to the zinc-blende structure of the alloy, and diffuse magnetic peaks that arise from a magnetically ordered array of Mn spins. The positions of these diffuse peaks are a signature of type-III antiferromagnetic order (AFM-III), where the Mn spins are arranged as shown in Fig. 19.

Neutron-diffraction studies of $Zn_{1-x}Mn_xTe$ and $Zn_{1-x}Mn_xSe$ yield much the same results as for $Cd_{1-x}Mn_xTe$, showing spin-cluster formation, AFM-III short-range order, and satu-

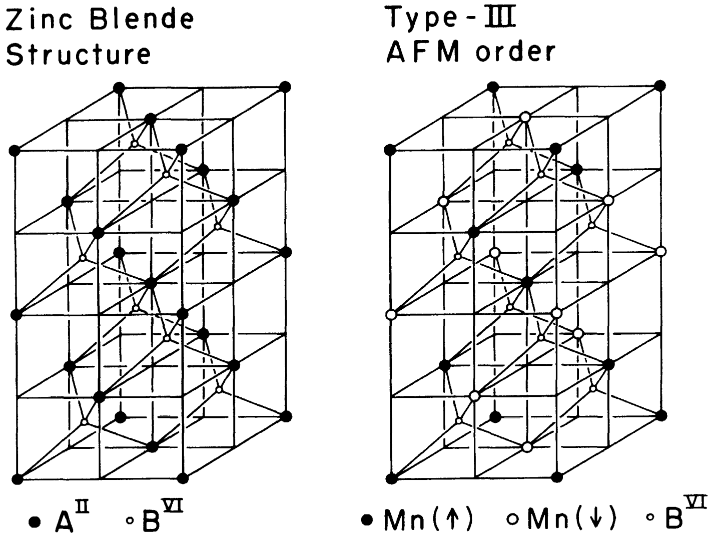

FIG. 19. The tetragonal unit cell for the AFM-III structure of zinc-blende crystals shown alongside two zinc-blende unit cells.

ration of the diffuse magnetic peak widths at low temperatures. The case of $Zn_{1-x}Mn_xSe$ is especially interesting, in that for $x > 0.30$ this material crystallizes in the wurtzite structure. Low-temperature data then reveal the existence of spin clusters correlated in the wurtzite version of the AFM-III antiferromagnetic order (i.e., that displayed by wurtzite β-MnS).

4.2 Magnetic Properties of DMS Layers and Multilayers

The magnetic properties of epitaxial DMS layers and DMS superlattices show remarkable differences compared with the behavior observed in the bulk. Magnetism in these systems has been investigated by direct magnetic probes such as neutron scattering (see, e.g., Giebultowicz *et al.*, 1992), SQUID magnetization measurements (see, e.g., Awschalom *et al.*, 1987), and spin-flip Raman scattering (Arora *et al.*, 1987), as well as indirect methods (Zeeman shift of photoluminescence peaks; Kolodziejski *et al.*, 1986). We begin by addressing the question of magnetic ordering in DMS epilayers and superlattices. Neutron-scattering experiments were performed on epitaxial layers of $Cd_{1-x}Mn_xSe$ ($x \approx 0.7$) and $Zn_{1-x}Mn_xTe$ ($0.70 < x \leq 1.0$), as well as ZnSe/MnSe, ZnTe/MnTe, and ZnTe/MnSe superlattices grown by molecular-beam epitaxy. (Although MnTe and MnSe can hardly be referred to as "diluted," we include them in this discussion because they represent the end points of the DMS families of alloys, and because their study "evolved" in the context of DMS investigations.) The epitaxial layers were 1 to 2 μm thick, and the individual superlattice layers were between typically 10 and 100 Å. All materials were in the zinc-blende phase.

The neutron diffraction observed on $Cd_{0.3}Mn_{0.7}Se$ is shown in Fig. 20 (Giebultowicz *et al.*, 1990). As in lower concentrations investigated on bulk specimens, we again observe clear signatures of AFM-III order [the presence of $(1,\frac{1}{2},0)$ and $(1,\frac{3}{2},0)$ peaks]. We note immediately, however, that—unlike the bulk data—these magnetic peaks are as sharp as the nuclear (0,2,0) peak, indicating long-range AFM-III order. Furthermore, the transition to long-range AFM-III order occurs at a sharply defined Néel temperature, above which there is no sign of any magnetic order. This again is in strong contrast to the bulk case, where the magnetic correlation length increases gradually as the temperature is decreased.

In a cubic crystal (and thus in the epilayers just discussed) the AFM-III structure exists in three "domains" corresponding to the three equivalent orientations of the unit cell shown in Fig. 19, allowed by the cubic symmetry. Studies involving the ZnSe/MnSe and

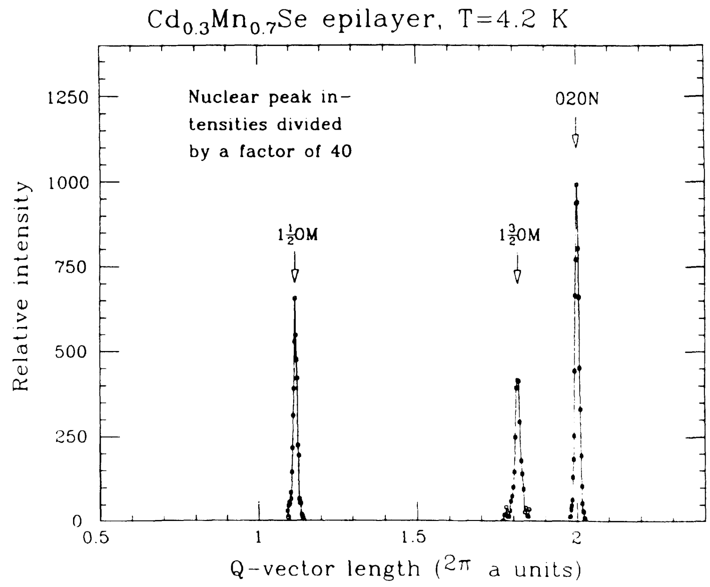

FIG. 20. Neutron-scattering data for a $Cd_{1-x}Mn_xSe$ epilayer (x = 0.70). Note that the magnetic peaks and nuclear peaks have similar widths, indicating that the magnetic order is long range.

ZnTe/MnTe superlattices show similar long-range AFM-III order, but with a major difference: Only one type of AFM-III domain is seen, with the tetragonal axis perpendicular to the plane of the epilayer. This is caused by the fact that the magnetic layer is in both these cases under compression, which lowers the originally cubic symmetry to tetragonal. In both of these superlattice families the magnetic order is long-ranged within the plane of the epilayer, while the correlation range normal to the epilayer corresponds exactly to the MnSe layer thickness. The long-range AFM-III order persists even when the individual MnSe layers are made as small as 20 Å. A totally surprising situation evolves when the magnetic layers are subjected to tension rather than compression, as is the case in ZnTe/MnSe superlattices. In that case AFM-III order disappears and, instead, a long-range helimagnetic phase sets in that is incommensurate with the lattice (Giebultowicz *et al.*, 1992). This system of magnetic semiconductors thus represents a unique system in which the ordering of spins that characterize the magnetic sublattice can be "engineered" by strain.

We mention, finally, that earlier studies of ZnSe/MnSe and $Cd_{1-x}Mn_xTe$/CdTe superlattices—albeit preliminary—have also suggested that dramatic changes should be expected as the DMS layers approach two-dimensionality. Magnetoluminescence and SQUID magnetization measurements in ZnSe/MnSe superlattices have shown that the MnSe layers are antiferromagnetically ordered when layer thicknesses are greater than a few monolayers. However, as the layer thickness approaches the monolayer limit, the MnSe layers remain paramagnetic, at least down to the lowest temperatures investigated so far (2 K). Similarly, SQUID magnetization measurements in CdTe/$Cd_{1-x}Mn_xTe$ superlattices (x in the vicinity of 0.2) have indicated a clear spin-glass transition in relatively thick $Cd_{1-x}Mn_xTe$ layers (80 Å), signaled by a cusp in the susceptibility. However, the cusp disappears when the layers are of the order of 20 Å, suggesting that the layers do not undergo a spin-freezing transition. All these studies strongly indicate that a crossover from three to two dimensions inhibits the formation of both spin-glass and antiferromagnetic order.

5. APPLICATIONS OF DMSs

DMS alloys exhibit large changes of band structure with composition, and thus constitute important candidates for those device applications where band-structure engineering is involved. Thus, for example, $Hg_{1-x}Mn_xTe$ is a viable candidate for infrared detectors, having properties very similar to those of $Hg_{1-x}Cd_xTe$. Since in this context of band-gap engineering the properties of DMSs closely parallel those of their nonmagnetic counterparts, we will not consider this issue beyond pointing out the range of band-gap adjustability implicit in Table 2.

Instead, we focus on the *sp-d* exchange, whereby effects associated with electronic spins—normally ignored in semiconductors—become sufficiently large in DMSs to merit consideration in the device context. Specifically, we will focus on Faraday rotation.

The large size of Faraday rotation and associated dispersion of circularly polarized modes in the alloys offers three types of applications to be considered. Since a rotation of the order of 45° is readily obtained at room temperature in moderate magnetic fields with a high figure of merit (see Dillon *et al.*, 1990), one may contemplate nonreciprocal optical devices (isolators and circulators) as a practical possibility. Faraday rotation may also be readily made use of in optical modulation and switching. Current efforts in this direction have examined the effect of an electric field on Faraday rotation in $Cd_{1-x}Mn_xTe$ films ($x \approx 0.5$) fabricated by ionized-cluster–beam deposition. These experiments have shown a large electric-field–induced change in the Verdet constant for photon energies in the vicinity of the energy gap. Finally, the large size of the Faraday effect might be exploited in highly sensitive optical-fiber–fed dielectric magnetometers, for measuring minute static as well as dynamic (up to gigahertz frequencies; Butler *et al.*, 1986) magnetic fields. Preliminary investigations have indicated that sensors with small dimensions (1 mm diameter) are feasible, and that static magnetic fields of 10 G can easily be resolved.

The large Faraday rotation involves, of course, large phase shifts of two counterrotating circular polarizations. If, instead of working with linear polarization, one employs a single circular polarization traveling along the applied magnetic field, one can then construct a magnetic-field–tunable phase shifter. This is of interest in interferometry as well as in optical cavity tuning (e.g., to optimize laser output).

We should note here that in the area of magneto-optical devices—especially of isolators—the use of ferromagnetic garnets is at a fairly advanced stage. These materials (e.g., Bi-enriched yttrium iron garnets) manifest immense Faraday rotations (and have a built-in magnetic field), but their use is restricted to wavelengths longer than 0.9 μm because of prohibitive absorption at shorter wavelengths. For $\lambda < 0.9$ μm DMS alloys (especially $Cd_{1-x}Mn_xTe$ and $Cd_{1-x}Mn_xSe$) are by far the best candidates for magneto-optical applications, showing larger Faraday rotations and better figures of merit (rotation/loss) than their competitors, such as paramagnetic garnets or rare-earth glasses. It should also be noted that even for $\lambda > 0.9$ μm, garnets suffer from the disadvantages that it is (at least at this stage) not possible to incorporate them monolithically in integrated optoelectronic circuits, the unit cell of a garnet consisting of 124 atoms. DMS alloys, on the other hand, are readily grown on substrates such as GaAs, making DMS films highly suitable for integrated optical-circuit structures. In this context the study of Faraday-effect phenomena in planar geometries, suitable for nonreciprocal devices based on DMS films, still requires considerable attention, since practically all magneto-optical work in DMS materials has been focused on bulklike phenomena.

GLOSSARY

Domain: In a magnetically ordered system, a region that contains a single configuration of the ordered structure.

Donor and Acceptor Levels: Typically refers to a substitutionally introduced atom, which has one more (or less) electron in the outermost shell than the atom that it replaces. The energy position of the extra (or missing) electron is called the donor (or the acceptor) level.

Epilayers: Layers that are grown epitaxially and are sufficiently thick for quantum effects to be negligible.

Exciton: A combination of an electron and a hole (the latter having a positive charge) bound together through Coulomb interaction. Both the form of the interaction and the eigenstates are very similar to those in a hydrogen atom.

Helimagnetic Phase: An ordered magnetic phase in which the spins of nearest neighbors are at an angle different from 180° and, therefore, form a helical pattern along the direction perpendicular to the spins.

Heterostructure: Consists of different materials, grown to form a multilayer coherent monolithic structure.

Magnetoresistance: A change in resistance arising from the application of a magnetic field.

Monolithic Structure: Consists of materials that are grown together without having to break the atomic arrangement of the original crystal structures. They are said to be atomically coherent.

Optical Isolators and Modulators: Devices that can be used to block reflected signals from returning to the source (isolators) and to modulate light intensity (modulators).

Radiative and Nonradiative Recombination: An electron in the conduction band can recombine with a hole in the valence band by emitting a photon, which is called radiative recombination. When the emitted energy is not in the form of photons, the process is referred to as nonradiative recombination.

Spin-Glass Phase: A magnetic phase in which orientations of spins are randomly distributed, but are nevertheless highly correlated. In other words, it behaves as a disordered spin system with individual spins "frozen."

SQUID: Acronym for superconducting quantum interference device, which has extremely high sensitivity in measuring magnetic susceptibility.

Superlattice: A monolithic structure with two or more materials grown as an alternating sequence of layers.

Universality Class: A common behavior (for very different physical systems) that typically depends on only a few pure numbers, such as the dimensionality of the space in which the phenomenon takes place.

Zeeman Splitting: The splitting of the energy levels associated with different spin states in an applied magnetic field.

Works Cited

Abragam, A., Bleaney, B. (1970), *Electron Paramagnetic Resonance of Transition Ions,* Oxford, U.K.: Clarendon.

Aggarwal, R. L., Jasperson, S. N., Becla, P., Furdyna, J. K. (1986), *Phys. Rev. B* **34,** 5894–5896.

Alonso, R. G., Oh, E., Ramdas, A. K., Luo, H., Samarth, N., Furdyna, J. K., Ram-Mohan, L. R. (1991), *Phys. Rev. B* **44,** 8009–8016.

Arora, A. K., Bartholomew, D. V., Peterson, D. L., Ramdas, A. K. (1987), *Phys. Rev. B* **35,** 7966–7972.

Awschalom, D. D., Hong, J. M., Chang, L. L., Grinstein, G. (1987), *Phys. Rev. Lett.* **59,** 1733–1736.

Bartholomew, D. V., Furdyna, J. K., Ramdas, A. K. (1986), *Phys. Rev. B* **34,** 6943–6950.

Bastard, G. (1988), *Wave Mechanics Applied to Semiconductor Heterostructures,* New York: Halsted.

Butler, M. A., Martin, S. J., Baughman, R. J. (1986), *Appl. Phys. Lett.* **49,** 1053–1055.

Dai, N., Luo, H., Zhang, F. C., Samarth, N., Dobrowolska, M., Furdyna, J. K. (1991), *Phys. Rev. Lett.* **67,** 3824–3827.

Dillon, J. F. Jr., Furdyna, J. K., Debska, U., Mycielski, A. (1990), *J. Appl. Phys.* **67,** 4917–4919.

Durbin, S. M., Han, J., Sungki, O., Kobayashi, M., Menke, D. R., Gunshor, R. L., Fu, Q., Pelakanos, N., Nurmikko, A. V., Li, D., Gonsalves, J., Otsuka, N. (1989), *Appl. Phys. Lett.* **55,** 2087–2089.

Ehrenreich, H., Hass, K. C., Larson, B. E., Johnson, N. F. (1987), in: R. L. Aggarwal, J. K. Furdyna, S. Von Molnar (Eds.), *Diluted Magnetic (Semimagnetic) Semiconductors,* Materials Research Society Symposia Proceedings Vol. 89, Pittsburgh: Materials Research Society, pp. 187–196.

Franciosi, A., Chang, S., Reifenberger, R., Debska, U., Reidel, R. (1985), *Phys. Rev. B* **32,** 6682–6687.

Furdyna, J. K. (1988), *J. Appl. Phys.* **64,** R29–R64.

Furdyna, J. K., Kossut, J. (Eds.) (1988), *Diluted Magnetic Semiconductors,* Vol. 25 of R. K. Williamson, A. C. Beer (Eds.), *Semiconductors and Semimetals,* New York: Academic.

Giebultowicz, T. M., Rhyne, J. J., Furdyna, J. K. (1987), *J. Appl. Phys.* **61,** 3537–3539.

Giebultowicz, T. M., Klosowski, P., Samarth, N., Luo, H., Rhyne, J. J., Furdyna, J. K. (1990), *Phys. Rev. B* **42,** 2582–2585.

Giebultowicz, T., Samarth, N., Luo, H., Furdyna, J. K., Klosowski, P., Rhyne, J. J. (1992), *Phys. Rev. B* **46,** 12076–12079.

Gumlich, H. E. (1981), *J. Lumin.* **23,** 73–99.

Hass, K. C., Ehrenreich, H. (1983), *J. Vac. Sci. Technol. A* **1,** 1678–1682.

Jain, M. (Ed.) (1991), *Diluted Magnetic Semiconductors,* Singapore: World Scientific.

Jonker, B. T., Krebs, J. J., Qadri, S. B., Prinz, G. A., Volkening, F. A., Koon, N. C. (1988), *J. Appl. Phys.* **63,** 3303–3305.

Jonker, B. T., Fu, L. P., Yu, W. Y., Chou, W. C., Petrou, A., Warnock, J. (1993), *J. Electron. Mater.* **22,** 489–495.

Kolodziejski, L. A., Gunshor, R. L., Bonsett, T. C., Venkatasubramaniam, R., Datta, S., Byslma, R. B., Becker, W. M., Otsuka, N. (1985), *Appl. Phys. Lett.* **47,** 169–171.

Kolodziejski, L. A., Gunshor, R. L., Otsuka, N., Gu, B. P., Hefetz, Y., Nurmikko, A. V. (1986), *Appl. Phys. Lett.* **48,** 1482–1484.

Kossut, J., Dobrowolski, W. (1993), in: K. H. J. Buschow (Ed.), *Handbook of Magnetic Materials,* Vol. 7, Amsterdam: Elsevier, Chap. 4.

Larson, B. E., Hass, K. C., Ehrenreich, H., Carlsson, A. E. (1988), *Phys. Rev. B* **37,** 4137–4154.

Lee, Y. K., Ramdas, A. K. (1984), *Solid State Commun.* **51,** 861–863.

Lewicki, A., Spalek, J., Furdyna, J. K., Galazka, R. R. (1988), *Phys. Rev. B* **37,** 1860–1863.

Luo, H., Furdyna, J. K. (1993), *Mod. Phys. Lett. B* **7,** 299–305.

Mycielski, A., Mycielski, J. (1980), *J. Phys. Soc. Jpn.* **49,** Suppl. **A,** 807–810.

Mycielski, J. (1983), *Applications of High Magnetic Fields in Semiconductors,* Berlin: Springer.

Oseroff, S. B. (1982), *Phys. Rev. B* **25,** 6584–6594.

Rigaux, C., Mycielski, A., Barilero, G., Menant, M. (1986), *Phys. Rev. B* **34,** 3313–3318.

Samarth, N., Luo, H., Furdyna, J. K., Qadri, S. B., Lee, Y. R., Ramdas, A. K., Otsuka, N. (1989), *Appl. Phys. Lett.* **54,** 2680–2682.

Shapira, Y., Foner, S., Becla, P., Domingues, D. N., Naughton, M. J., Brooks, J. S. (1986), *Phys. Rev. B* **33,** 356–365.

Shapira, Y. (1987), in: R. L. Aggarwal, J. K. Furdyna, S. von Molnar (Eds.) *Diluted Magnetic (Semimagnetic) Semiconductors,* Materials Research Society Symposia Proceedings Vol. 89, Pittsburgh: Materials Research Society, pp. 209–218.

Spalek, J., Lewicki, A., Tarnawski, Z., Furdyna, J. K., Galazka, R. R., Obuszko, Z., (1986), *Phys. Rev. B* **33,** 3407–3418.

Wall, A., Franciosi, A., Gao, Y., Weaver, J. H ., Dow, J. D., Kasowski, R. V. (1989), *J. Vac. Sci. Technol. A* **7,** 656–662.

Wojtowicz, T., Mycielski, A. (1983), *Physica B* **117 & 118,** 476–478.

Wojtowicz, T., Dietl, T., Sawicki, M., Plesiewicz, W., Jaroszynski, J. (1986), *Phys. Rev. Lett.* **56,** 2419–2422.

Zhang, F. C., Dai, N., Luo, H., Samarth, N., Dobrowolska, M., Furdyna, J. K., Ram-Mohan, L. R. (1992), *Phys. Rev. Lett.* **68,** 3220–3223.

Zhang, F. C., Luo, H., Dai, N., Samarth, N., Dobrowolska, M., Furdyna, J. K. (1993), *Phys. Rev. B* **47,** 3806–3810.

Further Reading

Aggarwal, R. L., Furdyna, J. K., von Molnar, S. (Eds.) (1987), *Diluted Magnetic (Semimagnetic) Semiconductors,* Materials Research Society Proceedings Vol. 89, Pittsburgh: Materials Research Society.

Brandt, N. B., Moshchalkov, V. V. (1984), *Adv. Phys.* **33,** 193–256.

Goede, O., Heimbrodt, W. (1988), *Phys. Status Solidi B* **146,** 11–62.

Shahzad, K., Samarth, N., Furdyna, J. K., Permogorov, S. (Eds.) (1992), *J. Lumin.* **52,** 183–191.

Taguchi, T., Schulman, J. N. (Eds.) (1993), *Physica B, Condensed Matter* **191.**

SEMICONDUCTORS, ELEMENTAL—ELECTRONIC PROPERTIES

Franco Bassani and Giuseppe La Rocca, *Scuola Normale Superiore, Pisa, Italy*

Introduction 403
1. **Symmetry Properties of the Diamond Crystal** 404
2. **Electronic Energy Bands** 406
2.1 Electronic-Structure Calculations 407
2.2 Electronic Bands 408
3. **Phonon Bands** 408
3.1 Lattice-Dynamics Calculations 409
3.2 Phonon Bands and Density of States 409
4. **Band Extrema and Effective-Mass Approximation** 410
4.1 Electron Pockets 411
4.2 Hole Pockets 412
5. **Density of States and Critical-Point Analysis** 412
6. **Optical Properties** 413
6.1 General Concepts and Methods 414
6.2 Direct Interband Transitions 416
6.3 Indirect Interband Transitions 417
6.4 Inelastic Scattering 418
6.5 Excitonic and Polaritonic Effects 418
6.6 Core-State Excitations 419
7. **Photoelectron Spectroscopy** 419
7.1 Photoemission 420
7.2 Inverse Photoemission 421
8. **Dielectric Screening** 421
8.1 General Theory 422
8.2 Simplified Model Calculations 422
9. **Impurity Levels** 422
9.1 Shallow Levels 423
9.2 Resonant Levels 424
9.3 Deep Levels 424
10. **External Perturbations** 425
10.1 Hydrostatic Pressure 425
10.2 Uniaxial Stress 426
10.3 Electric Field 426
10.4 Magnetic Field 427
11. **Electrical Conductivity** 428
11.1 Free-Carrier Conductivity 428
11.2 Carrier Density 429
11.3 Carrier Mobility 430
11.4 Hall Effect and Magnetoresistance 431
11.5 High-Field Effects 432
11.6 Impurity-Band Conductivity 434
Works Cited 435
Further Reading 436

INTRODUCTION

The existence of solids whose resistivity is intermediate between that of metals and of insulators was known already in the 19th century. This class of materials is very large, since their resistivities at room temperature can vary between 10^{-4} and 10^{10} Ω cm. It was also recognized that the value of the resistivity is not the best criterion to define semiconductors because at absolute zero temperature they are all insulators; moreover, their resistivity is extremely sensitive to the presence of impurities. Semiconductors can rather be defined by the property that their resistivity decreases with increasing temperature in a way that is exponential with the inverse temperature. Another property of semiconductors, opposite to that of metals, is to be transparent to electromagnetic radiation of low frequency, and to absorb totally light above a critical frequency ω_0, which can be in the visible or infrared region.

These were all mysterious properties for a long time, because the classical model of free electrons, which was able to explain the metallic properties, could not be adopted, and the alternative model of electrons bound

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

to their atom in the solid would produce insulators. But even more mysterious was the fact, already known at the beginning of this century, that positive-charge carriers were responsible for electrical currents in some semiconductors, as shown by the sign of the Hall coefficient, and that the number of carriers greatly varied from sample to sample. The clue to an understanding of the semiconductor properties came between the two world wars with the application to solids of the concepts of quantum mechanics and the notion of energy band states. The Pauli exclusion principle led to the introduction of the concept of "hole" and explained the negative Hall constant of some semiconductors. Furthermore, the crucial role of impurities (doping) was understood.

During World War II, much interest was devoted to silicon, particularly to point-contact rectifiers for radar technology. After the war, detailed studies were undertaken of the transport properties of group-IV elements, which crystallize in the diamond structure and are basic semiconductors. Germanium in particular was of interest because its conductivity could be easily controlled by adding impurities. This led to the discovery of the transistor by Bardeen, Brattain, and Shockley in 1947, which can be singled out as the most important milestone in the field of semiconductor physics and technology; since then, solid-state electronics has had a dramatic growth. The development of microdevices also stimulated a great improvement in the growth techniques of monocrystals and in the control of their impurity content (e.g., Czochralski technique, floating-zone refinement; see CRYSTAL GROWTH), so that totally reproducible results could be obtained. A tremendous advance was the production of silicon monocrystals, of good-quality silicon surfaces, and of interfaces of silicon with metals and oxide insulators. In the late 1950s, the control of film formation, lithography, and etching led to the introduction of integrated circuits by Kilby and Noyce. After the 1960s, large-scale integrated circuitry was developed, with the consequences that are now apparent in computer and communications microelectronics, and in optoelectronics.

The theoretical understanding of the phenomena has proceeded parallel to the experimental progress. The group-IV elements are simple crystals, and their electronic structure was understood first. Their band structure has been computed since the 1960s, and it is continuously improved upon. The detailed knowledge of the Ge and Si band structures in fact preceded their verification by optical and photoemission experiments and has given the ideas for new phenomena like the Gunn oscillations.

At present we feel that the field is in full expansion, particularly in the optoelectronic effects: the transformation of light into electrical current (solar cells) and the transformation of electricity into laser radiation (semiconductor lasers). New progress in this field is forthcoming, due to the possibility of growing layers of atoms, or strings and dots of atoms, of one type on a substrate of different material by molecular-beam epitaxy. Also in this field Si and Ge may be the leading materials, because they are so abundant and form crystal alloys in all proportions. Ge_xSi_{1-x} alloys of different thickness and shape can be deposited on a Si substrate to form completely new microstructures, tailored to obtain desired properties.

1. SYMMETRY PROPERTIES OF THE DIAMOND CRYSTAL

The group-IV elemental semiconductors (germanium, silicon, and gray tin) are crystals with the symmetry of diamond; i.e., they belong to the space group $Fd3m$ (O_h^7). The space-group symmetry, along with time-reversal symmetry, determines completely the symmetry properties of a crystal at the macroscopic as well as at the microscopic level. Symmetry profoundly affects all the properties of a material, as will be shown in the following sections. Other elemental semiconductors (such as boron or selenium) have more complex crystal structures; as they are not related to Si and Ge (by far the most important materials in this class), they will not be considered in this article.

The diamond structure can be described as two interpenetrating face-centered cubic (fcc) lattices displaced by $\sqrt{3}a/4$ along the main diagonal of the conventional cubic cell of side a, as shown in Fig. 1. The lattice constants a of diamond (C), silicon (Si), germanium (Ge), and gray tin (α-Sn) are 3.56, 5.43, 5.65, and 6.46 Å, respectively. The transla-

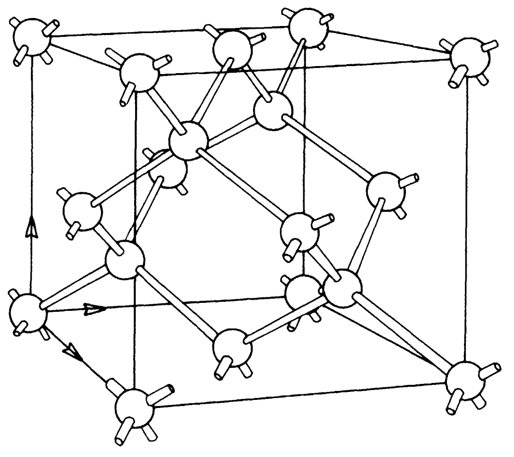

FIG. 1. Conventional (nonelementary) cubic cell of the diamond structure.

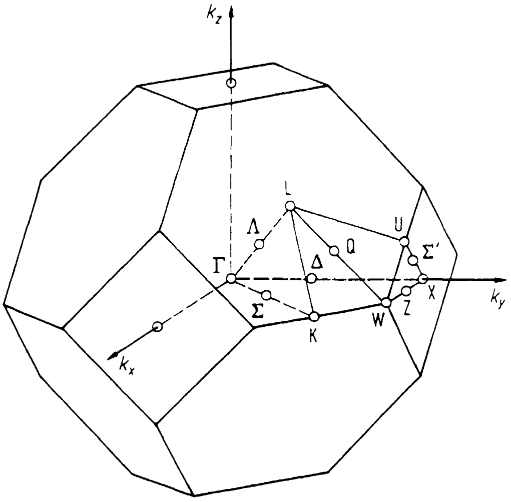

FIG. 2. Brillouin zone of the diamond crystal; the capital letters label lines and points of high symmetry.

tional symmetry is, therefore, described by a Bravais fcc lattice of primitive translations $\boldsymbol{\tau}_1 = (a/2)(0,1,1)$, $\boldsymbol{\tau}_2 = (a/2)(1,0,1)$, and $\boldsymbol{\tau}_3 = (a/2)(1,1,0)$, the crystal being invariant for translations by any vector $\mathbf{R} = n_1\boldsymbol{\tau}_1 + n_2\boldsymbol{\tau}_2 + n_3\boldsymbol{\tau}_3$ with integral values of n_1, n_2, and n_3. The reciprocal lattice is body-centered cubic (bcc) with primitive vectors $\mathbf{h}_1 = (2\pi/a)(\bar{1},1,1)$, $\mathbf{h}_2 = (2\pi/a)(1,\bar{1},1)$, and $\mathbf{h}_3 = (2\pi/a)(1,1,\bar{1})$; the reciprocal-lattice vectors, $\mathbf{G} = m_1\mathbf{h}_1 + m_2\mathbf{h}_2 + m_3\mathbf{h}_3$, with m_1, m_2, and m_3 integers, have the property $e^{i\mathbf{R}\cdot\mathbf{G}} = 1$, for any lattice translation $\mathbf{R}$. The primitive (symmetric) cell of the reciprocal lattice, i.e., the first Brillouin zone (BZ), of the diamond structure is shown in Fig. 2. The physical significance of the reciprocal lattice and of the BZ is that, as a consequence of the periodic translational symmetry (Bloch theorem), all elementary excitations in a crystal (such as phonons and electrons) can be described by wave functions of the form $\Psi(\mathbf{k},\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}u(\mathbf{r})$, where the wave vector $\mathbf{k}$ belongs to the BZ and the function u is invariant under lattice translations $[u(\mathbf{r} + \mathbf{R}) = u(\mathbf{r})]$; furthermore, when two or more such elementary excitations interact, their total wave vector is conserved modulo a reciprocal lattice vector; i.e., $\mathbf{k}_i = \mathbf{k}_f + \mathbf{G}$, $\mathbf{k}_i$ and $\mathbf{k}_f$ being the total initial and final wave vectors (processes in which $\mathbf{G} \neq 0$ are called umklapp processes).

Besides translational symmetry, the point-group symmetry characterizes the properties of the diamond structure under rotations and reflections; the point group of diamond is the full cubic group (O_h). At the microscopic level, though, half of the 48 symmetry operations of O_h are accompanied by a fractional translation such that the two interpenetrating fcc lattices are exchanged (the space group of diamond being a nonsymmorphic space group with a basis of two atoms). The point group determines the macroscopic symmetry of a crystal, in particular the form of all the material tensors (such as the polarizability tensor), the relations among elastic constants, etc. Furthermore, at the microscopic level, the point-group symmetry determines the degeneracies of elementary excitations as a function of their wave vector $\mathbf{k}$ in the BZ zone. In particular, elementary excitations at points and lines of high symmetry in the BZ, marked in Fig. 2, are characterized by their transformation properties under point-group symmetry operations that leave their wave vector unchanged (i.e., the little group of their wave vector). Finally, the O_h point-group symmetry allows one to restrict oneself to 1/48 only of the Brillouin zone (irreducible wedge), as the entire zone can be obtained from it by applying rotation and reflection symmetry operations. Standard group-theoretical methods are employed to obtain systematically all the wealth of information that symmetry alone provides (Bir and Pikus, 1974; Bassani and Pastori Parravicini, 1975). Besides space-group symmetry, time-reversal symmetry also affects the properties of a crystal. At a macroscopic level, it constrains the form of material tensors (Onsager's relations); for instance, it ensures that the conductivity tensor is symmetric: $\sigma_{ij} = \sigma_{ji}$. When external perturbations break time-reversal symmetry, such relations are correspondingly modified; for instance, when a magnetic field $\mathbf{B}_0$ is present, $\sigma_{ij}(\mathbf{B}_0) = \sigma_{ji}(-\mathbf{B}_0)$. At a microscopic level, time-reversal symmetry brings about additional degeneracies for the elementary excitation spectra. The most important with regard to electrons is that, when their spins are taken into account, in a crystal with inversion symmetry (as for diamond) at each wave vector **k** the electronic states are (at least) doubly degenerate, because the state $|\mathbf{k},\uparrow\rangle$ with spin up is degenerate with the state $|-\mathbf{k},\uparrow\rangle$ by inversion symmetry, and the latter is degenerate with the state $|\mathbf{k},\downarrow\rangle$ by time-reversal symmetry. The spin degeneracy between the states $|\mathbf{k},\uparrow\rangle$ and $|\mathbf{k},\downarrow\rangle$ is, in general, absent in materials lacking an inversion center (such as in III–V compound semiconductors) or when time-reversal symmetry is lifted, e.g., by an external magnetic field (Kim *et al.*, 1989).

2. ELECTRONIC ENERGY BANDS

Virtually all properties (transport, optical, vibrational) of a semiconductor depend on its electronic structure (see ELECTRON STRUCTURE OF SOLIDS); the latter is described by the energy bands $E_n(\mathbf{k})$, i.e., the energy levels of electronic states having a wave vector **k** and definite transformation properties under the symmetry operations of the little group of **k** (see Sec. 1). The high symmetry of the crystal enormously simplifies the problem of calculating the energy bands and provides rigorous results, such as the degeneracies at special lines and points of the Brillouin zone and the compatibility relations among different lines and points (Bassani and Pastori Parravicini, 1975). Nevertheless, a quantitative computation of the electronic structure of a given semiconductor is a very difficult task, even when basic approximations are made, such as the adiabatic approximation considering the nuclei fixed at their equilibrium sites (see Sec. 3) and the one-electron approximation (neglecting many-body correlations among the electrons). Besides, in the so-called band approximation, each electron moves in the presence of a local potential $V(\mathbf{r})$, including the interactions with the ions as well as with all other electrons in an average (mean-field) sense; the crystal potential $V(\mathbf{r})$ can be either empirically determined or calculated *ab initio*. In any case, $V(\mathbf{r})$ has the full symmetry of the crystal, and the corresponding solutions for the electronic states are Bloch functions $\Psi(\mathbf{k},\mathbf{r})$ with a wave vector **k** in the BZ (see Sec. 1). Ψ indicates a two-component wave function (spinor), one component for each of the two electron-spin orientations. In the band approximation, the fundamental equation for calculating the electronic eigenenergies $E_n(\mathbf{k})$ and eigenfunctions $\Psi_n(\mathbf{k},\mathbf{r})$ is

$$\left[\frac{p^2}{2m} + V(\mathbf{r}) + \frac{\hbar}{4m^2c^2}\boldsymbol{\sigma}\cdot(\nabla V \times \mathbf{p})\right]\Psi_n(\mathbf{k},\mathbf{r}) = E_n(\mathbf{k})\Psi_n(\mathbf{k},\mathbf{r}), \tag{1}$$

where **p** is the momentum operator ($\mathbf{p} = -i\hbar\nabla$) and the last term on the left-hand side is the spin–orbit interaction (σ_j are the 2×2 Pauli spin matrices). The spin–orbit coupling is a relativistic correction, which induces energy splittings according to the symmetry properties of spin-$\frac{1}{2}$ particles, i.e., the double-group symmetry (Bassani and Pastori Parravicini, 1975); spin–orbit effects are significant for heavy elements but, in a first approximation, can be neglected in light elements. For each **k**, the equation above has a series of solutions labeled by the band index n, and, for each n, the energy band $E_n(\mathbf{k})$ is a continuous function of **k** (see Figs. 3–5). At zero temperature, the electrons present in a semiconductor fill the levels $E_n(\mathbf{k})$ in increasing order of energy, according to the Pauli exclusion principle, up to the topmost fully occupied band (valence band), which is separated by an energy gap E_g from the lowest empty band (conduction band). For typical semiconductors, the energy gap is of the order of 1 eV: for Ge $E_g = 0.774$ eV, for Si $E_g = 1.17$ eV (at room temperature, respectively 0.66 and 1.1 eV); diamond has $E_g = 5.4$ eV and would rather be considered an insulator instead of a semiconductor; gray tin has a vanishing energy gap (zero-gap semi-

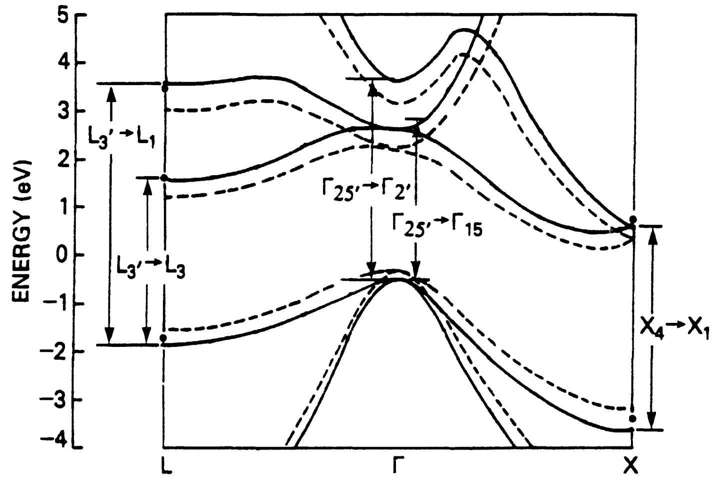

FIG. 3. Band structure of silicon: the dashed lines are from a local-density approximation *ab initio* calculation; the solid lines are from a quasiparticle *ab initio* calculation including many-body correlation effects (Pickett and Wang, 1984); the arrows indicate experimental transition energies (Madelung, 1982).

conductor or ideal semimetal). From the theoretical point of view, an energy gap of the order of 1 eV is the hallmark of semiconductors, from which all their characteristic properties depend.

2.1 Electronic Structure Calculations

The methods employed to calculate the electronic energy bands of semiconductors are numerous and diverse; for instance, while some consider a crystal as a collection of atoms (localized orbital methods), others treat it as a perturbed uniform fluid (plane-wave methods); while some are to a large extent empirical, others instead are *ab initio*. A very popular localized orbital approach is the tight-binding method in which the wave functions are expanded in a basis of linear combination of atomic orbitals (LCAO); usually, only a limited number of atomic orbitals are included, and only the interactions among a limited number of neighbors are considered. This method, at least in its sim-

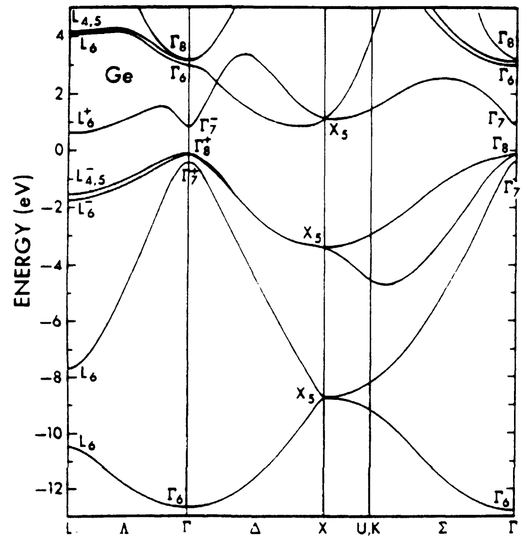

FIG. 4. Band structure of germanium from an empirical pseudopotential calculation (Chelikowsky and Cohen, 1976).

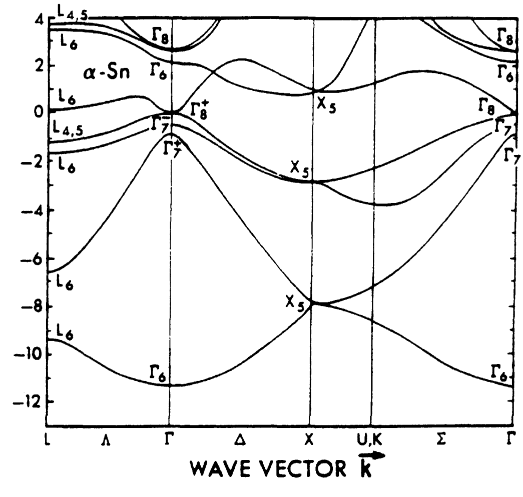

FIG. 5. Band structure of α-Sn from an empirical pseudopotential calculation (Chelikowsky and Cohen, 1976).

pler semiempirical versions, is transparent and economic; it has been very successful in calculating both electronic and structural properties of semiconductors and establishing their chemical trends (Harrison, 1989).

Among the plane-wave methods, the most important is the pseudopotential approach. In a crystal, the valence (or conduction) -band wave functions are rapidly varying near each atomic site, as they are orthogonal to those of the core states; therefore, their expansions in plane waves are very slowly convergent. It is possible, though, to replace them with pseudowave functions that are smooth (nodeless) near the atomic sites while replacing the ionic potential with a much smaller pseudopotential, in such a way that the energies $E_n(\mathbf{k})$ of valence and conduction-band states can be computed using a limited number of plane waves (Bassani and Celli, 1961); furthermore, from the pseudowave functions, the electronic charge density outside the cores can also be obtained. The pseudopotential itself (or only its first few Fourier components) can be empirically chosen to reproduce a set of data (mostly optical transition energies) or can be calculated *ab initio* and self-consistently (Bachelet *et al.*, 1982). The pseudopotential method has been very successful in conjunction with density-functional theory in the local-density approximation (LDA), i.e., using an exchange-correlation potential dependent only on the local electronic charge density. The local-density approximation, though, is not quite satisfactory; in particular, the energy gaps of semiconductors are usually significantly underestimated (even though the shapes of both valence and conduction bands are well described; see Fig. 3). This shortcoming can be mended by going beyond the one-electron approximation and calculating the quasiparticle self-energies including (approximately) many-body effects (Strinati *et al.*, 1980; Hybertsen and Louie, 1985).

2.2 Electronic Bands

Group-IV elemental semiconductors are characterized by tetrahedral bonding with sp^3 hybridization of their four outer-shell electrons. Going from diamond to gray tin, as the ionic cores become larger, the lattice constants increase and the energy gaps decrease. The general appearance of the energy bands (see Figs. 3–5) of elemental semiconductors presents close analogies, especially for valence bands, which are due to the similar chemistry of the bonds and the identical symmetry properties of the crystals. Diamond has the top of the valence band at $\mathbf{k} = 0$ (symmetry Γ'_{25}) and the bottom of the conduction band along the Δ line (⟨100⟩ axis). The band structure of silicon, shown in Fig. 3, is very similar, but the conduction band Γ_{15} p-like states are pushed to higher energies relative to the conduction band Γ'_2 s-like states through orthogonalization with respect to the core p levels (absent in C). This effect is further enhanced as one goes from Si to Ge where also orthogonalization to the core d levels pushes the Γ_{15} states (partly d-like) even higher relative to s-like bands. We show in Fig. 4 the band structure of germanium: The effects of spin–orbit coupling (neglected for the light elements above, for which are qualitatively similar) are larger in this case, because the increase in atomic number magnifies all relativistic effects. We notice that the topmost valence band (Γ'_{25} three times degenerate without spin) splits into a Γ_8^+ fourfold degenerate and a Γ_7^+ twofold degenerate band separated by a spin–orbit energy gap $\Delta = 0.3$ eV (for Si, Δ is only 0.04 eV). Besides, the Γ'_2 conduction band (Γ_7^- in double-group notation) is lower than the Γ_{15} band (split into Γ_8^- and Γ_6^-) and, consequently, the lowest conduction state (symmetry L_6^+) is at the L point (boundary of the BZ along the ⟨111⟩ direction). The band structure of gray tin is shown in Fig. 5: Relativistic effects are even more important, and the Γ_7^- band is lower than the Γ_8^+ band (inverted gap), which turns out to be both the topmost valence band and the lowest conduction band (Groves and Paul, 1963).

All the main features of the band structures presented above are firmly established from both the theoretical and the experimental points of view. These results represent one of the most important achievements in the physics of semiconductors and provide a framework within which almost all material properties (from phonon spectra to optical and transport properties) can be understood.

3. PHONON BANDS

The vibrations of the nuclei around their sites in the perfect crystal affect both the equilibrium properties of a solid and its response to external probes. The phonons are the quantized traveling eigenmodes of vibration. As in the case of electrons, the phonon energies as functions of the wave vector **k** are arranged in bands according to the space-group symmetry of the crystal; for the diamondlike semiconductors, there are three acoustic and three optical phonon branches (Kittel, 1966).

3.1 Lattice-Dynamics Calculations

Since the early days of solid-state physics, the calculation of the lattice dynamics of semiconductors has attracted much attention for both its theoretical and its practical relevance. From the microscopic point of view, the basic problem is the evaluation of the dynamical matrix by diagonalization of which the phonon eigenenergies and eigendisplacements are obtained (Born and Huang, 1954). For this purpose, many phenomenological models have been developed, which successfully yield physical insight in the interpretation of experimental data. With modern computing capabilities, also accurate *ab initio* calculations have become feasible.

The valence-force–field model explains the lattice dynamics on the basis of a set of empirical force constants similar to those employed in the description of molecular vibrations. The potential energy is given in terms of internal coordinates such as bond lengths, bond angles, and their combinations involving a limited number of neighbors. Within such a framework, Tubino *et al.* (1972) have been able to describe well the phonon-dispersion relations of elemental semiconductors using as fitting parameters six force constants corresponding to the reaction of the valence electrons to deformations of the network of covalent bonds (e.g., stretching, bending). Another important phenomenological approach is the adiabatic bond-charge model proposed by Weber (1977) in which, besides the ions, also massless, pointlike bond charges appear in the dynamical equations. Four kinds of forces are considered: long-range Coulomb forces, ion–ion central forces (as for metals), ion–bond-charge central forces (which stabilize the bond charges at the equilibrium positions at the middle of the bonds), and, finally, bond-bending forces (of the valence-force–field type). With only four fitting parameters, the phonon-dispersion curves for Si, Ge, and α-Sn are well described; in particular, the typical flattening of the transverse acoustic branch can be explained. Many other semiempirical models have been developed (very popular are the shell models), but most of them employ a large number of free parameters (ten or more), the interpretation of which is not always transparent; therefore, they provide rather a neat interpolation of the data than physical insight.

Modern *ab initio* methods relating the vibrational properties of solids to their electronic structure are very accurate and computationally demanding. An efficient scheme, based on density-functional linear-response theory, has been developed by Giannozzi *et al.* (1991). It is a self-consistent perturbative approach in which the dynamical matrix is derived from the response of the electronic system to an external static potential representing the ion displacements. The electronic system is described using a plane-wave expansion, *ab initio* pseudopotentials, and the local-density approximation for the exchange-correlation potential (see Sec. 2); then, to compute the response to the perturbation, a Green's-function technique is employed so that the computational burden for the perturbed system is not significantly heavier than that for the unperturbed one. Excellent results have been obtained for the phonon eigenenergies and eigendisplacements of elemental (as well as polar) semiconductors; they are displayed for Si and Ge in Figs. 6 and 7, respectively.

3.2 Phonon Bands and Density of States

From the experimental point of view, the most useful technique to measure the phonon-dispersion curves $\omega_\nu(\mathbf{k})$ is the coherent scattering of thermal neutrons whereby both energy and (pseudo)momentum are exchanged with the lattice with the participation of phonons having any wave vector in the Brillouin zone. One-phonon Raman scattering can only involve phonons with a negligible wave vector (i.e., at the Γ point of the BZ). However, from the folded phonon peaks of a superlattice structure, the bulk phonon dispersion can be reconstructed (Jusserand and Cardona, 1989). The two-phonon Raman scattering spectrum, on the other hand, mimics the phonon density of states. Impu-

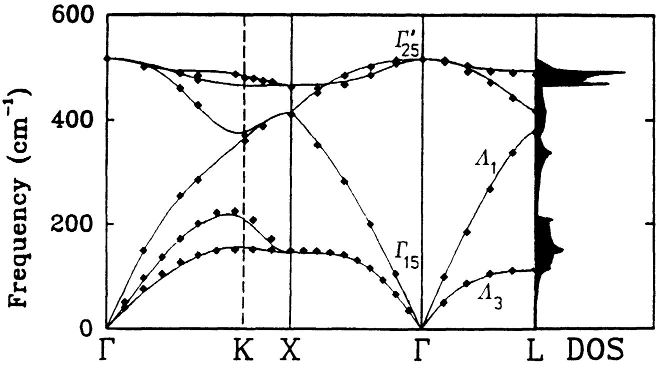

FIG. 6. Phonon-dispersion curves of Si: the data points are from inelastic neutron scattering (Nilsson and Nelin, 1972), and the lines are from an *ab initio* density-functional-perturbation theory (Giannozzi *et al.*, 1991); the rightmost panel shows the calculated phonon density of states.

rity-activated infrared absorption also gives information on the phonon density of states, but it is not always easily interpreted. Figures 6 and 7 show the phonon-dispersion curves of silicon and germanium as measured by neutron scattering, compared to modern *ab initio* calculations (also shown are the corresponding densities of states). Just as in the case for electrons, the phonon-dispersion curves are labeled according to the space-group symmetry of the diamond lattice, and the corresponding critical points clearly show up as Van Hove singularities in the phonon density of states (see Sec. 5).

4. BAND EXTREMA AND EFFECTIVE-MASS APPROXIMATION

Many of the transport and optical properties of semiconductors are dominated by carriers (electron and holes introduced by doping or excited across the energy gap) that occupy the bottom of the conduction band or the top of the valence band. An accurate description of the electronic energy bands near these extrema is given for the unperturbed crystal by the $\mathbf{k}\cdot\mathbf{p}$ method. Then, the response of the carriers to external perturbations, such as electric and magnetic fields, can be dealt with using the effective-mass approximation.

The $\mathbf{k}\cdot\mathbf{p}$ method consists of a perturbative calculation of the electronic states in the proximity of a given $\mathbf{k}_0$ point in the Brillouin zone (usually of high symmetry) corresponding to a maximum in the valence band (hole pocket) or a minimum in the conduction band (electron pocket). If we write a Bloch function at $\mathbf{k}$ in terms of all those at $\mathbf{k}_0$,

$$\psi_j(\mathbf{k},\mathbf{r}) = \sum_n c^j_n(\mathbf{k})\psi_n(\mathbf{k}_0,\mathbf{r})e^{i(\mathbf{k}-\mathbf{k}_0)\cdot\mathbf{r}}$$

the energy eigenvalues $E_j(\mathbf{k})$ at $\mathbf{k}$ as well as the expansion coefficients $c^j_n(\mathbf{k})$ for the respective eigenstates are given by the solution of the secular problem:

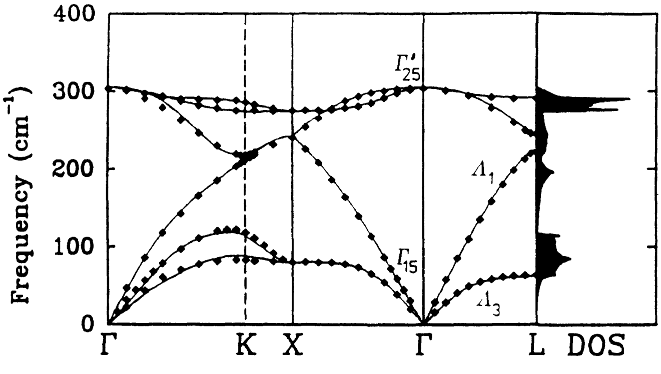

FIG. 7. Phonon-dispersion curves of Ge: the data points are from inelastic neutron scattering (Nilsson and Nelin, 1971), and the lines are from an *ab initio* density-functional-perturbation theory (Giannozzi *et al.*, 1991); the rightmost panel shows the calculated phonon density of states.

$$\sum_{n'}\left\{\left[\frac{\hbar^2}{2m}(\mathbf{k}-\mathbf{k}_0)^2 + E_n(\mathbf{k}_0) - E_j(\mathbf{k})\right]\delta_{nn'} + \frac{\hbar}{m}(\mathbf{k}-\mathbf{k}_0)\cdot\mathbf{p}_{nn'}(\mathbf{k}_0)\right\}c^j_{n'}(\mathbf{k}) = 0, \tag{2}$$

where

$$\mathbf{p}_{nn'}(\mathbf{k}_0) = \int \psi_n^*(\mathbf{k}_0,\mathbf{r})\mathbf{p}\psi_{n'}(\mathbf{k}_0,\mathbf{r})\, d^3\mathbf{r}. \tag{3}$$

For $\mathbf{k} \simeq \mathbf{k}_0$, a limited number of bands with energies $E_n(\mathbf{k}_0)$ close to the energy $E_{n0}(\mathbf{k}_0)$ of the extremum of interest are included in the summations above. For a nondegenerate band, a single-band approximation is usually employed, the interaction with all other bands being treated in perturbation theory up to second order in $\mathbf{k} - \mathbf{k}_0$. In this case, one gets simply

$$\begin{aligned} E_{n0}(\mathbf{k}) &\simeq E_{n0}(\mathbf{k}_0) + \frac{\hbar^2}{2m}(\mathbf{k}-\mathbf{k}_0)^2 + \frac{\hbar^2}{m^2}\sum_{n'\neq n0}\frac{[(\mathbf{k}-\mathbf{k}_0)\cdot\mathbf{p}_{n0n'}][\mathbf{p}_{n'n0}\cdot(\mathbf{k}-\mathbf{k}_0)]}{E_{n0}(\mathbf{k}_0) - E_{n'}(\mathbf{k}_0)} \\ &\equiv E_{n0}(\mathbf{k}_0) + \frac{\hbar^2}{2}(\mathbf{k}-\mathbf{k}_0)\cdot\mathbf{m}^{-1}\cdot(\mathbf{k}-\mathbf{k}_0), \end{aligned} \tag{4}$$

where the inverse effective-mass tensor $\mathbf{m}^{-1}$ for the extremum of interest has been introduced. Finally, measuring the energy from $E_{n0}(\mathbf{k}_0)$ and the wave vector from $\mathbf{k}_0$, and choosing the $\hat{\mathbf{1}}$, $\hat{\mathbf{2}}$, and $\hat{\mathbf{3}}$ directions along the principal axes of $\mathbf{m}^{-1}$, the standard expression for ellipsoidal pockets is obtained:

$$E(\mathbf{k}) = \frac{\hbar^2}{2}\left(\frac{k_1^2}{m_1} + \frac{k_2^2}{m_2} + \frac{k_3^2}{m_3}\right), \tag{5}$$

where the effective-mass parameters can be positive, negative, or infinite. From Eq. (4), it follows that small gaps are in general accompanied by small effective masses.

Then, in the presence of slowly varying external perturbations, the dynamics of electrons and holes can be described as that of free carriers (i.e., without considering the rapidly varying crystal potential) with the (anisotropic) effective mass just discussed. The effective-mass approximation can be generalized to deal with degenerate bands and to consider the spin–orbit interaction (Luttinger and Kohn, 1955; Bir and Pikus, 1974).

4.1 Electron Pockets

The symmetry of the dispersion relation of electron (or hole) pockets as well as the values of the corresponding effective-mass parameters can be measured by cyclotron resonance, i.e., the absorption of radio-frequency radiation in the presence of an external magnetic field $\mathbf{B}$. The resonance frequency is $\omega_c = eB/m^*(\hat{\mathbf{B}})c$, with

$$m^*(\hat{\mathbf{B}}) = \left[\frac{m_1m_2m_3}{m_1(\hat{\mathbf{B}}\cdot\hat{\mathbf{1}})^2 + m_2(\hat{\mathbf{B}}\cdot\hat{\mathbf{2}})^2 + m_3(\hat{\mathbf{B}}\cdot\hat{\mathbf{3}})^2}\right]^{1/2}. \tag{6}$$

From the dependence of the observed absorption on the magnitude and direction of the magnetic field, the symmetry of the pockets of carriers as well as their effective-mass parameters can be determined.

The conduction band of Ge has four electron pockets at the L point of the BZ (see Fig. 4; the half pockets at $\mathbf{k}$ and $-\mathbf{k}$ are to be combined together); they are ellipsoids of revolution elongated along the ⟨111⟩ directions. The values of the longitudinal mass $m_L = m_3 \simeq 1.6\ m$ and of the transverse mass $m_T = m_1 = m_2 \simeq 0.08\ m$ have been determined from cyclotron resonance measurements such as those shown in Fig. 8 (Dresselhaus *et al.*, 1955): with $\mathbf{B}$ along $[\sqrt{3},\sqrt{3},\sqrt{2}]/2$, the four ⟨111⟩ electron pockets split into three groups inequivalently oriented with respect to the magnetic field, each one of these giving rise to a distinct resonance according to Eq. (6). The conduction band of Si has six electron pockets along the ⟨100⟩ directions (see Fig. 3); they are ellipsoids of revolution elongated along the same directions with a longitudinal mass $m_L = m_3 \simeq 1.0\ m$ and a transverse mass $m_T = m_1 = m_2 \simeq 0.2\ m$ (measured from cyclotron resonance; see Fig. 9 where because of the choice of magnetic-field orientation two distinct electronic resonances show up).

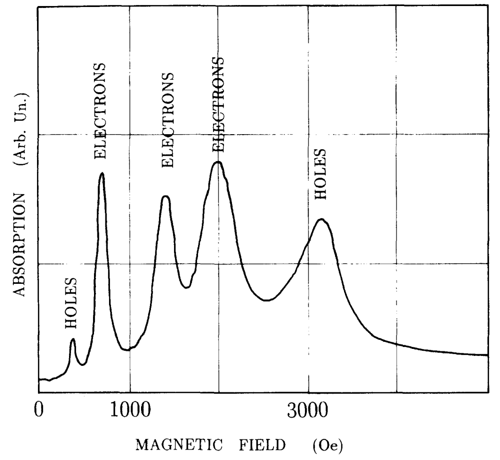

FIG. 8. Cyclotron resonance in Ge (at a frequency of 24 GHz and a temperature of 4 K) with magnetic field in the (110) plane at 60° from the [001] axis (Dresselhaus *et al.*, 1955).

4.2 Hole Pockets

The top of the valence band in both Si and Ge is at the center of the BZ: Γ'_{25} without spin, Γ_8^+, and Γ_7^+ including spin (see above). The Γ_7^+ split-off hole states can be described as a simple band with an effective mass $m_{so} \simeq 0.25m$ in Si and $m_{so} \simeq 0.08m$ in Ge. In a rough approximation, the Γ_8^+ hole states can be described as belonging to two distinct spherical pockets each characterized by an isotropic mass: the heavy-hole pocket (with mass $m_{hh} \simeq 0.5m$ for Si and $m_{hh} \simeq 0.3m$ for Ge) and the light-hole pocket (with mass $m_{lh} \simeq 0.16m$ for Si and $m_{lh} \simeq 0.044m$ for Ge). Also the hole masses have been determined from cyclotron resonance measurements such as those shown in Figs. 8 and 9. This simple picture is not quite correct; even when the hole kinetic energies are larger than $\hbar\omega_c$ and only two fundamental hole cyclotron resonances appear (classical regime), the heavy-hole pocket is significantly anisotropic. Furthermore, when the hole kinetic energies are comparable to $\hbar\omega_c$, a very complex hole cyclotron-resonance spectrum appears (quantum regime), as shown in Fig. 10 (Hensel and Suzuki, 1974). A complete description of the hole states can be given using the multiband effective-mass approach (Luttinger and Kohn, 1955), in which the valence-band degeneracy is fully accounted for.

5. DENSITY OF STATES AND CRITICAL-POINT ANALYSIS

Many material properties are determined to a large extent by the density of states $g(E)$, i.e., the number of phonon (electron, etc.) states per unit volume and unit energy. Since the elementary excitations in a crystal have energy spectra arranged in bands $E_n(\mathbf{k})$, the density of states can be written as

$$g(E) = \sum_n \int_{\mathrm{BZ}} \frac{d^3k}{(2\pi)^3}\, \delta(E_n(\mathbf{k}) - E), \qquad (7)$$

where the integral is over the Brillouin zone and the summation is over all bands (and includes the multiplicity of degenerate bands,

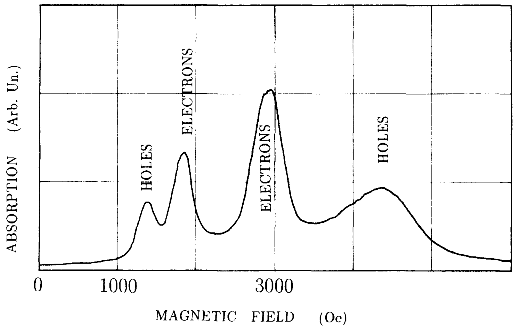

FIG. 9. Cyclotron resonance in Si (at a frequency of 24 GHz and a temperature of 4 K) with magnetic field in the (110) plane at 30° from the [001] axis (Dresselhaus *et al.*, 1955).

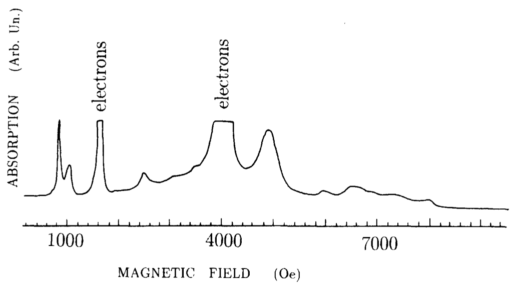

FIG. 10. Cyclotron resonance in Ge (at a frequency of 52.9 GHz and a temperature of 1.2 K) with magnetic field along the [111] axis; the complex spectrum of hole transitions is evident (Hensel and Suzuki, 1974).

at least 2 for electronic states in elemental semiconductors). An equivalent expression is

$$g(E) = \sum_n \frac{1}{(2\pi)^3} \int_{S(E)} \frac{dS}{|\nabla_{\mathbf{k}} E_n(\mathbf{k})|}, \tag{8}$$

where the two-dimensional integral is over a surface in **k** space for which $E_n(\mathbf{k}) = E$. The latter expression for $g(E)$ clearly shows that a large contribution to the density of states comes from points where the gradient of $E_n(\mathbf{k})$ with respect to **k** vanishes (critical points).

Topological considerations indicate that each band has at least a minimum set of critical points: one maximum, one minimum, and six saddle points (three of type M_1 and three of type M_2). Around a critical point $\mathbf{k}_0$, the energy band $E_n(\mathbf{k})$ has in general a quadratic expansion in $\mathbf{k} - \mathbf{k}_0$, as given by Eq. (5): if all the coefficients (i.e., the effective masses for the electronic case) are positive, the critical point is a minimum (type M_0); if one only is negative, it is a saddle point of type M_1; if two are negative, it is a saddle point of type M_2; if all are negative, it is a maximum (type M_3). The energies of the critical points increase going from those of type M_0 to M_1, M_2, and M_3; the three critical points of type M_1 may correspond to the same energy, and likewise those of type M_2. A critical point having energy E_0 is accompanied by a singularity in the density of states around E_0 having a characteristic shape (shown in Fig. 11) depending on the type of the critical point (van Hove singularities):

$$\begin{aligned}
M_0 \Rightarrow g(E) &= \begin{cases} B + O(E - E_0) & \text{if } E < E_0 \\ B + A(E - E_0)^{1/2} + O(E - E_0) & \text{if } E > E_0 \end{cases} \\
M_1 \Rightarrow g(E) &= \begin{cases} B - A(E_0 - E)^{1/2} + O(E - E_0) & \text{if } E < E_0 \\ B + O(E - E_0) & \text{if } E > E_0\,, \end{cases} \\
M_2 \Rightarrow g(E) &= \begin{cases} B + O(E - E_0) & \text{if } E < E_0 \\ B - A(E - E_0)^{1/2} + O(E - E_0) & \text{if } E > E_0 \end{cases} \\
M_3 \Rightarrow g(E) &= \begin{cases} B + A(E_0 - E)^{1/2} + O(E - E_0) & \text{if } E < E_0 \\ B + O(E - E_0) & \text{if } E > E_0 \end{cases}
\end{aligned}$$

where B is a constant and $O(E - E_0)$ is a term of order $E - E_0$ when E tends to E_0. In some cases, the analytical behavior of the density of states at a critical point is different from that just described because for the corresponding $E_n(\mathbf{k})$ the quadratic expansion does not hold; in particular, for the acoustic-phonon branch around the minimum at $E = 0$, the dispersion is linear and $g(E)$ increases as E^2 (Debye spectrum).

The location of critical points in the BZ is mainly (but not completely) determined by symmetry considerations. For the elemental semiconductors with the diamond lattice, the points Γ, X, L, and W in the BZ (see Fig. 2) are always critical points; other critical points usually occur along lines or planes of symmetry, but critical points at a generic wave vector **k** cannot be excluded. Photoemission spectra (see Sec. 7) often give direct information on $g(E)$; optical spectroscopies (see Sec. 6) give access to the joint density of states, which has analogous singularities. For phonons, $g(E)$ is obtained from inelastic neutron scattering (see Sec. 3). In an experimental phonon or electron density of states, the van Hove singularities may be difficult to identify; using modulation techniques (see Sec. 10), the derivative spectra can be obtained that show very clearly the stronger singularities in the first (and higher-order) derivatives of $g(E)$. The study of van Hove singularities has been instrumental in establishing the quite detailed picture of phonon and electronic energy bands of elemental semiconductors (see Secs. 2 and 3) on a sound experimental basis (Brust *et al.*, 1962; Cohen and Chelikowsky, 1988).

6. OPTICAL PROPERTIES

The most precise way to investigate the atomic and electronic structures of all substances is to study their response to electromagnetic radiation. This general principle, however, has held for semiconductors only since the 1960s, when the optical quality of

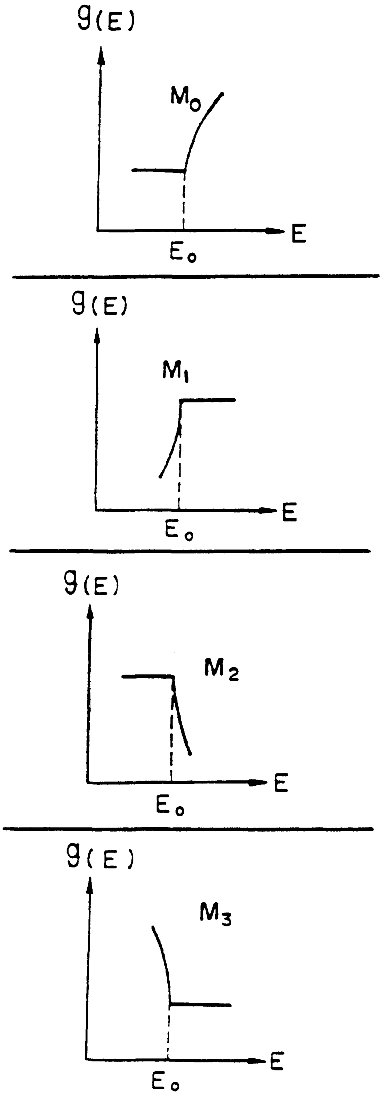

FIG. 11. Schematic representation of the three-dimensional density of states $g(E)$ near critical points of type M_0, M_1, M_2, and M_3; see text for details.

the surfaces made it possible to measure the real and the imaginary parts of the dielectric function as functions of the light frequency. Before, it was not possible to measure light dispersion and absorption because the absorption is so strong when the frequency is higher than the threshold frequency that no radiation goes through the semiconductor, even for very thin layers. This difficulty was finally overcome by measuring the reflectivity over a large frequency range with sufficient accuracy so that application of the Kramers–Kronig (K.K.) relations made it possible to determine the real and the imaginary parts of the reflection coefficient. From this the dispersion and the absorption could be obtained, and the bases for the experimental verification of the electronic structure of semiconductors were established (Philipp and Ehrenreich, 1963; Greenaway and Harbeke, 1968; Bassani and Pastori Parravicini, 1975).

6.1 General Concepts and Methods

We recall that the normal-incidence reflection coefficient, defined as the ratio between the electric fields of reflected and incident waves, is related to the complex index of refraction $\tilde{n}$ by the Fresnel relation

$$r = \frac{E_r}{E_i} = \frac{\tilde{n} - 1}{\tilde{n} + 1} = |r|e^{i\theta} \tag{9}$$

and, taking the logarithm,

$$\log r = \log |r| + i\theta. \tag{10}$$

Since this is an analytic function of the frequency ω in the upper complex plane $\omega + i\eta$, the following K.K. relation holds true:

$$\theta(\omega) = -\frac{2\omega}{\pi} \mathrm{P} \int_0^\infty \frac{\log |r(\omega')|}{\omega'^2 - \omega^2} d\omega', \tag{11}$$

where P denotes the principal value; and from the measured reflectivity $|r(\omega')|^2$, $\theta(\omega)$ can be obtained, and consequently real and imaginary parts of $\tilde{n} = n + ik$ can be derived from Eq. (9). This amounts to the knowledge of dispersion and absorption since the complex dielectric function is related to $\tilde{n}$ by

$$\tilde{\epsilon} = \epsilon_1 + i\epsilon_2 = \tilde{n}^2. \tag{12}$$

The dissipation is related to the imaginary parts, since the absorption coefficient is given by

$$\alpha(\omega) = \frac{2k\omega}{c} = \frac{\epsilon_2\omega}{nc}, \tag{13}$$

and the real part $\epsilon_1 = n^2 - k^2$ gives the dispersion.

To relate the measured optical constants to the electronic structure, we must consider the coupling dipolar interaction between the electromagnetic radiation and the electrons,

$$H'(\mathbf{r}) = e\mathbf{E}\cdot\mathbf{r}, \tag{14}$$

where $\mathbf{E} = \mathbf{E}_0 e^{-i\omega t} + \mathbf{E}_0^* e^{i\omega t}$ denotes the electric field of the radiation and $\mathbf{r}$ is the position of the electron that can make a transition between different electronic states. We recall that the transition probability per unit time between initial state $|i\rangle$ and final state $|f\rangle$ is given by the Fermi "golden rule"; in first order

$$P_{i\to f}^{(1)} = \frac{2\pi}{\hbar}|\langle f|H'|i\rangle|^2\delta(E_f - E_i \pm \hbar\omega), \tag{15}$$

and in second order

$$P_{i\to f}^{(2)} = \frac{2\pi}{\hbar}\left|\sum_n \frac{\langle f|H'|n\rangle\langle n|H'|i\rangle}{E_n - E_i \pm \hbar\omega}\right|^2 \times \delta(E_f - E_i \pm \hbar\omega \pm \hbar\omega), \tag{16}$$

where in the energy-conserving delta functions we include the energies of the photons absorbed ($-\hbar\omega$ in first order) or emitted ($+\hbar\omega$ in first order). In the case of semiconductors we must sum over all occupied initial states and empty final states in the unit volume to compute the electromagnetic power dissipated at a given frequency. In first order,

$$W(\omega) = \frac{1}{V}\sum_{i,f} P_{i\to f}^{(1)}\hbar\omega(f_i^0 - f_f^0), \tag{17}$$

where V is the crystal volume and f_n^0 is the Fermi distribution function for state n:

$$f_n^0 = \frac{1}{\exp[(E_n - \mu)/kT] + 1}. \tag{18}$$

The absorption coefficient is by definition

$$\alpha(\omega) = \frac{\text{dissipated power per unit volume}}{\text{incoming flux}} = \frac{W(\omega)}{(n^2E^2/2\pi)(c/n)}. \tag{19}$$

We can observe that it is better to consider the imaginary part of the dielectric function because it does not depend on the real part of the index of refraction. From the above expressions, we obtain in first order for the dissipative function

$$\epsilon_2(\omega) = 4\pi^2\frac{1}{V}\sum_{i,f}|\langle f|e\mathbf{r}|i\rangle|^2\delta(E_f - E_i - \hbar\omega), \tag{20}$$

where the summation is over all occupied initial states and empty final states. The real part $\epsilon_1(\omega)$ is then obtained from the K.K. dispersion relation

$$\epsilon_1(\omega) = 1 + \frac{2}{\pi}\mathrm{P}\int_0^\infty \frac{\omega'\epsilon_2(\omega')}{\omega'^2 - \omega^2}d\omega'. \tag{21}$$

In second order we can have two-photon absorption or emission, or Raman processes where one photon is absorbed and one emitted.

It is also possible to have second-order processes where only one photon is involved, provided we consider the additional interaction with the vibrations of the lattice, which may be seen as a phonon field, which contributes an additional interaction to (14) of the type

$$H'(\mathbf{r}) = \sum_{\mathbf{q}\nu}(V_{\mathbf{q}}^{(\nu)}e^{i\mathbf{q}\cdot\mathbf{r}}a_{\mathbf{q}\nu} + V_{\mathbf{q}}^{(\nu)*}e^{-i\mathbf{q}\cdot\mathbf{r}}a_{\mathbf{q}\nu}^+), \tag{22}$$

where $a_{\mathbf{q}\nu}$ ($a_{\mathbf{q}\nu}^+$) are destruction (creation) operators, which refer to phonons of mode ν and with wave vector $\mathbf{q}$, such that

$$\begin{aligned} a_{\mathbf{q}\nu}|n_{\mathbf{q}\nu}\rangle &= \sqrt{n_{\mathbf{q}\nu}}|n_{\mathbf{q}\nu} - 1\rangle, \\ a_{\mathbf{q}\nu}^+|n_{\mathbf{q}\nu}\rangle &= \sqrt{n_{\mathbf{q}\nu} + 1}|n_{\mathbf{q}\nu} + 1\rangle, \end{aligned} \tag{23}$$

$n_{\mathbf{q}\nu}$ being the phonon number, given by the statistical distribution function of Bose–Einstein:

$$n_{\mathbf{q}\nu} = \frac{1}{\exp[\hbar\omega_\nu(\mathbf{q})/k_BT] - 1}, \tag{24}$$

where we have denoted with $\hbar\omega_\nu(\mathbf{q})$ the phonon energy. Evidence has been found of all the above processes, which has provided a verification of the electronic and also of the vibrational structure of elemental semiconductors.

6.2 Direct Interband Transitions

Most of the optical absorption in crystals is due to interband transitions, i.e., to promotion of electrons from the valence bands $E_v(\mathbf{k})$ to the conduction bands $E_c(\mathbf{k})$, the values of **k** being the same because the photon wave vector is very small (direct transitions). This means that the summations on initial and final states in (20) can be replaced by an integration on the **k** vector space, on which the electron energies depend. Because of the δ-function condition in (20), we obtain a singular behavior in the vicinity of critical points of the Brillouin zone, where

$$\nabla_k[E_c(\mathbf{k}) - E_v(\mathbf{k})] = \mathbf{0}. \tag{25}$$

These give different types of singularities in the joint density of states (Van Hove singularities), analogous to those in the density of states discussed in Sec. 5. These structures have been found in all elemental semiconductors, and this has allowed the identification of the critical-point transition peaks and the verification of the band structure.

As a typical example, we show in Fig. 12 the optical excitation spectrum of Ge, measured as explained above and computed using the band-structure calculations described in Sec. 2 and Eq. (20) (Brust *et al.*, 1962). Similar results are obtained for Si and C, but in the case of gray tin the energy gap between valence and conduction bands vanishes because the Γ_7^- state falls below the Γ_8^+ state, which is half full (see Sec. 2). We stress that not only have the critical points been identified but also the full band structure has been put to test by the described analysis of the optical properties. An immediate verification comes from a detailed study of the edge and of the E_1 peaks at 2.1 eV. They turn out to be split by the spin–orbit separation as expected from the band structure described in Sec. 2 ($\Delta \simeq 300$ meV in Ge at the edge and about 2/3 of it at E_1). A further verification comes from the calculation of the dipole matrix elements at all points of the Brillouin zone and for all possible transitions to higher conduction bands. This justifies the fact that almost all of the possible transitions take place to the lowest empty conduction bands, which can also be proved by verifying the *f*-sum rule in the optical range (Bassani and Altarelli, 1983):

$$\int_0^{\omega_M} \omega' \epsilon_2(\omega')\, d\omega' \simeq \frac{2\pi^2 e^2 n_{\text{eff}}}{m}, \tag{26}$$

where in this case the frequency ω_M is just above the spectrum in the visible region and n_{eff} is about equal to the number of valence electrons that take part in the optical transitions (eight per unit cell in elemental semiconductors).

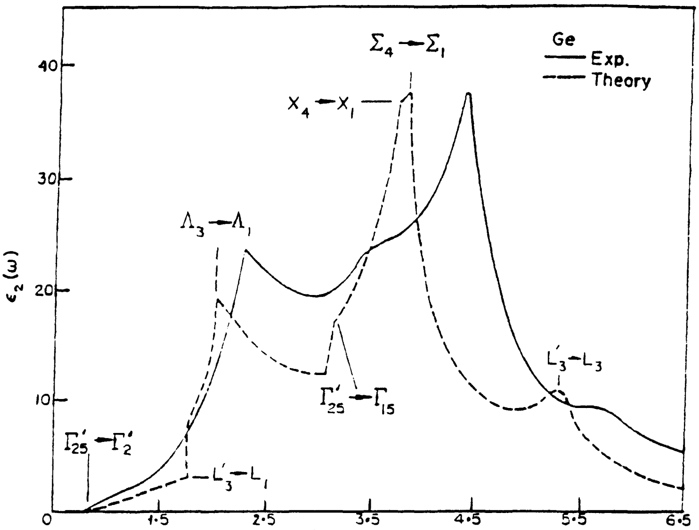

FIG. 12. Imaginary part of the dielectric function of Ge: the solid line is experimental, and the dashed one, theoretical; optical transitions at critical points are explicitly indicated (Brust *et al.*, 1962).

6.3 Indirect Interband Transitions

The band structures of C, Si, and Ge clearly show that, while the top of the valence band is at the center of the Brillouin zone (Γ point), the minimum of the conduction band occurs away from Γ (close to the X point along the Δ line in diamond and Si, at the L point in Ge). This has been verified by cyclotron resonance experiments, but it also shows in the optical properties, because the phonon interaction (22) makes possible an optical transition between different points of the Brillouin zone, the missing (pseudo)momentum being provided by the phonon emitted or adsorbed. We make use of the second-order transition probability (16), taking the perturbation interaction to be the sum of Eqs. (14) and (22), and obtain, assuming a constant matrix element squared (C) involving the states near the extrema that give the strongest contribution,

$$\epsilon_2(\omega) = \frac{4\pi^2 e^2 C}{m^2\omega^2}\left(n_{\mathbf{q}_0\nu_0} + \tfrac{1}{2} \pm \tfrac{1}{2}\right) \times \int_{\mathrm{BZ}}\int_{\mathrm{BZ}} \frac{2d^3k_1 d^3k_2}{(2\pi)^6}\,\delta(E_c(\mathbf{k}_1) - E_v(\mathbf{k}_2) \pm \hbar\omega_{\nu 0}(\mathbf{q}_0) - \hbar\omega), \qquad (27)$$

where + (−) refers to creation (destruction) of a phonon of mode ν_0 with wave vector $\mathbf{q}_0$ connecting the valence-band maximum and the conduction-band minimum. The summation over the final electronic states above the edge ($\hbar\omega_0$) gives an energy dependence that is different from that of direct transitions, and precisely

$$\epsilon_2(\omega) = 0 \text{ when } \omega < \omega_0, \quad \epsilon_2(\omega) \sim (\omega - \omega_0)^2 \text{ when } \omega > \omega_0.$$

Furthermore, different edges occur in correspondence to each phonon, and their values depend on the fact that the phonons are absorbed or emitted. At low temperature we can only create phonons, and consequently the edge occurs at higher energy:

$$\hbar\omega_0 = E_g + \hbar\omega_{\nu 0}(\mathbf{q}_0). \qquad (28)$$

At higher temperature we can also absorb phonons, with a temperature-dependent probability given by the Bose–Einstein factor, and the edge is shifted to lower energy:

$$\hbar\omega_0 = E_g - \hbar\omega_{\nu 0}(\mathbf{q}_0). \qquad (29)$$

Indirect transitions are responsible for the weak absorption tail in C, Si, and Ge, which reflects their conduction-band structures. We show in Fig. 13 the absorption tail of Si; a fit to the profile of the indirect edge gives the frequency of the X phonons that make the transition possible.

Under conditions of electronic excitations, the initial state may be the minimum of the conduction band, and the electron–hole recombination produces luminescence. The recombination luminescence is indirect in ele-

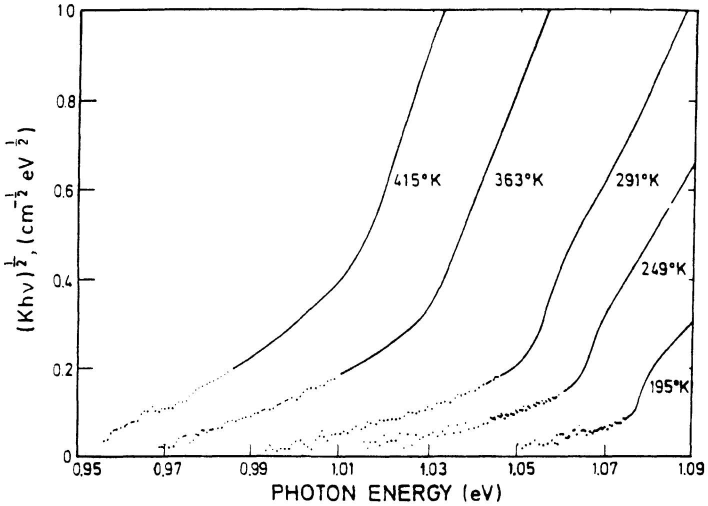

FIG. 13. Indirect absorption edge in Si at different temperatures (MacFarlane *et al.*, 1958).

mental semiconductors and shifted to lower energy at low temperatures because the conservation of energy requires

$$\hbar\omega = E_g - \hbar\omega_{\nu 0}(\mathbf{q}_0). \tag{30}$$

6.4 Inelastic Scattering

Besides the indirect transitions, the phonons are involved in the inelastic scattering of light (Raman scattering for optical phonons and Brillouin scattering for acoustic phonons) whereby one photon is absorbed and a second photon is emitted, the energy difference between the two being due to the phonon created (Stokes process) or absorbed (anti-Stokes process). The Stokes process, being proportional to $n + 1$, is present at all temperatures, while the anti-Stokes process, proportional to n (the Bose–Einstein factor for the relevant phonon), is significant only at a sufficiently high temperature. The Raman spectrum of diamond, from which the energy of the zone-center optical phonon can be precisely measured, is shown in Fig. 14; a typical Brillouin-scattering spectrum of Ge, involving acoustic phonons of appropriate wave vector, is illustrated in Fig. 15. Brillouin- and, especially, Raman-scattering data provide valuable information on the electron-phonon coupling that mediates these processes (resonant conditions with electronic transitions are also frequently employed). Therefore, they are very useful for understanding the electron and phonon band structures (see Sects. 2 and 3) as well as the scattering mechanisms limiting the free-carrier mobilities (see Sec. 11).

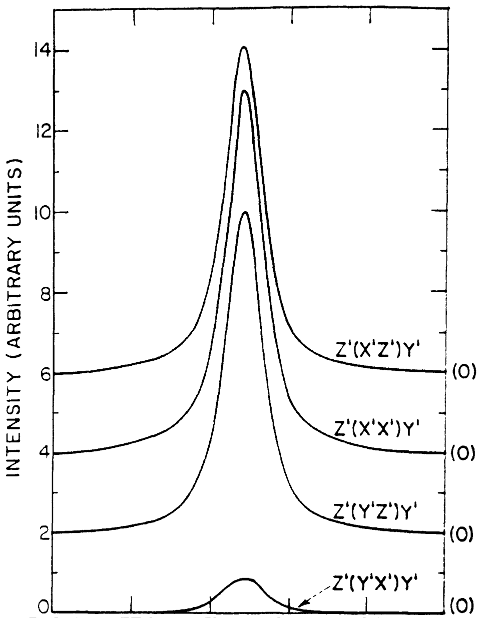

FIG. 14. One-phonon Raman spectrum of diamond; the polarization features correspond to the Γ'_{25} symmetry of the zone-center optical phonon (Solin and Ramdas, 1970).

6.5 Excitonic and Polaritonic Effects

As a result of exchange and correlation effects within the many-electron system, the excitation states are not simply obtained as the sums of the occupied band-state energies of the electrons, as assumed in the preceding sections, but depend on an additional two-body interaction between the electron promoted to the conduction band and the hole that is left in the valence band. This interaction consists of two terms (Bassani and Pastori Parravicini, 1975), an electron–hole Coulomb attraction screened by the dielectric function of the medium $\epsilon(\mathbf{k})$ and a spin-dependent electron–hole exchange of the form

$$2J\delta_S, \tag{31}$$

where $\delta_S = 1$ for singlet states and $\delta_S = 0$

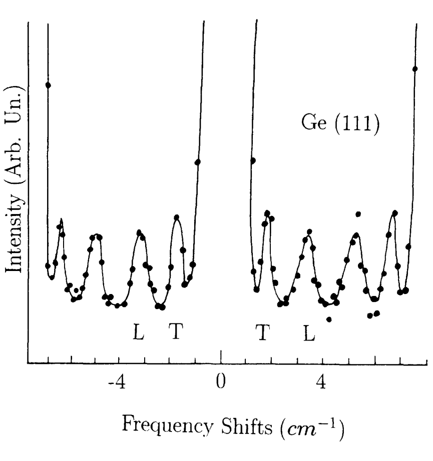

FIG. 15. Brillouin spectrum of Ge showing both transverse acoustic (T) and longitudinal acoustic (L) phonons (Sandercock, 1972).

for triplets. The term J is a two-body contribution, which can be computed from the wave functions of valence and conduction states. It consists of a short-range term J_s that corresponds to a local field correction, and a long-range contribution J_l that is singular in the limit $\mathbf{k} \to 0$, depending on the fact that the exciton is longitudinal or transverse with respect to the direction of motion. As a consequence we obtain a hydrogen-type series of states below the gap, of the form

$$E_n = E_g - \frac{R^*}{n^2} + 2J_s\delta_S + 2J_l\begin{pmatrix} \frac{8}{3}\pi \\ -\frac{4}{3}\pi \end{pmatrix}\delta_S, \qquad (32)$$

where the coefficient $+8\pi/3$ refers to longitudinal excitons and $-4\pi/3$ to transverse excitons. Since $R^* = (\mu^*/m)(R/\epsilon^2)$ is the effective Rydberg, modified with respect to the atomic value by the reduced effective mass μ^* and by the dielectric screening, and both J_s and J_l are much smaller than R^*, the excitonic series is hydrogen-like with only the lowest states appearing in the excitation spectrum. If we consider the free motion of the center of mass, we obtain a dispersion law for the exciton states of the type

$$E_n(k_{\text{ex}}) = E_n + \frac{\hbar^2 k_{\text{ex}}^2}{2(m_e + m_h)}, \qquad (33)$$

where for indirect gaps $\mathbf{k}_{\text{ex}}$ is measured with respect to $\mathbf{q}_0$, the wave-vector difference between the two band extrema. This produces a frequency dependence above the exciton states that is of the type $[\hbar\omega - E_n \pm \hbar\omega_{\nu 0}(\mathbf{q}_0)]^{1/2}$, rather than the dependence typical of indirect interband transitions. The excitonic effects are almost negligible in elemental semiconductors because of the large values of ϵ but have been duly observed. They become much more important in polar compounds.

Associated to the excitonic states there is also a modification of the electromagnetic waves in the crystal, which produces anomalous dispersion in the frequency region close to the exciton. This is obtained from the usual dispersion law of propagating waves in media

$$\omega^2 = [c^2/\epsilon(\omega,\mathbf{k})]k^2, \qquad (34)$$

where the dielectric function has poles in correspondence to the excitons (33). The solution of (34) for the transverse modes gives an upper polariton and a lower polariton, upper and lower being referred to the transverse exciton energy (Hopfield, 1958). Though polaritonic effects are more important in polar materials, they have been nevertheless observed in elemental semiconductors.

6.6 Core-State Excitations

The sum rule (26) that we mentioned before is strictly valid only when the upper frequency limit is infinite and the number n refers to the total number of electrons. This means that additional absorption occurs as the frequency increases, with edges corresponding to transitions due to electrons from inner shells (1*s* in C; 1*s*, 2*s*, 2*p* in Si; etc.). The theory of inner-shell excitation spectra is similar to that for valence electrons, but account must be taken of the fact that the energy bands of the "core" states are flat. Consequently, only direct transitions occur, and the structures in the interband transition spectrum depend only on the conduction-band density of states. The exciton states depend only on the electron effective masses and are dispersionless. Measurements of the core-state excitations require frequencies from the vacuum ultraviolet region to the hard x-ray region and can be performed using synchrotron radiation; they fully confirm the conduction-band structures described in Sec. 2 (Bassani and Altarelli, 1983).

7. PHOTOELECTRON SPECTROSCOPY

Besides optical spectroscopy, the most useful experimental technique to measure the electronic energy bands is photoelectron spectroscopy (see PHOTOEMISSION AND PHOTOELECTRON SPECTRA). The dispersion curves of electronic states that are below the Fermi level and, in some cases, above the vacuum level can be studied by a number of variations of photoemission spectroscopy (Cardona and Ley, 1978) in which the energy, wave vector, and polarization (spin) of the incident photon (emitted electron) can be selectively chosen or averaged over. The electronic states with energies between the Fermi level and the vacuum level are not involved in photoemission processes as they can be neither initial states (because they are empty in the ground state) nor final states (because the electrons are unable to escape from the solid once excited). These levels can be investigated by inverse photoemission, which is, in a sense, the time-reversed process of photoemission: An incoming electron falls into an electronic level available in the solid, emitting a photon, which is detected. The main advantage of photoemission and inverse photoemission with respect to optical spectroscopy is the capability of providing absolute energy levels rather than energy differences between levels. Besides lower resolution and (especially for inverse photoemission) lower statistics, the main disadvantage for the study of bulk properties is that photoelectron spectroscopy is surface sensitive because of the short escape or penetration depths of electrons (of the order of 10 Å). The progress in ultrahigh-vacuum and sample-preparation techniques as well as the availability of synchrotron radiation have led to significant developments in photoemission.

7.1 Photoemission

A simple physical picture of the photoemission process is given by the three-step model: First an electron in the solid is promoted to an empty band by absorption of the incoming photon, then it travels toward the surface and, finally, escapes into the vacuum to be detected. The electrons may (secondary electrons) or may not (primary electrons) undergo inelastic scattering processes in the solid before escaping through the surface. The yield I of primary electrons emitted with energy E by photons of energy $\hbar\omega$ is determined by three factors: the probability P of optical excitation, T of reaching the surface, and D of leaving the solid $[I(E,\hbar\omega) = P(E,\hbar\omega)T(E)D(E)]$. T is a smooth function of the electron energy and so is D, beyond the low-energy cutoff due to the potential barrier at the solid-to-vacuum interface. Therefore, the structures in the energy distribution curves measured in angle-integrated photoemission reflect mainly the dependence of P on incoming photon energy and emitted electron energy. The behavior of P is analogous to that of the imaginary part of the dielectric constant ϵ_2 (see Sec. 6) and is determined by two factors: the optical excitation matrix element $M_{fi}(\mathbf{k})$, which depends on the wave functions of initial (valence) and final (conduction) electronic states in the semiconductor, and the energy distribution of the joint density of states, which depends on the dispersion relations of the initial $E_i(\mathbf{k})$ and final $E_f(\mathbf{k})$ levels:

$$P(E,\hbar\omega) \propto \sum_{f,i} \int d^3k M_{fi}(\mathbf{k})\delta(E_f(\mathbf{k}) - E_i(\mathbf{k}) - \hbar\omega)\delta(E - E_f(\mathbf{k})). \quad (35)$$

The approximation of a constant matrix element is often made and, for photon energies higher than about 20 eV, the structures in P correspond to those in the density of states of the occupied states only (the density of final states being practically structureless); in this case, the photoemission spectra mimic the density of states of the valence band.

Photoemission spectra (corrected for the secondary-electron contribution) of diamond, Si, and Ge are shown in Fig. 16 (Cavell *et al.*, 1973). The three main peaks correspond in each case to prominent peaks in the density of states of the valence band: From higher to lower binding energies, they are associated, respectively, to predominantly s-like, mixed

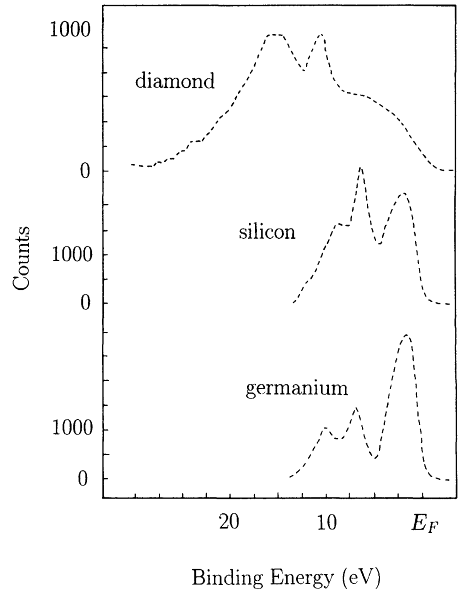

FIG. 16. X-ray photoelectron spectra of diamond, Si, and Ge (Cavell *et al.*, 1973).

s- and *p*-like, and predominantly *p*-like states. Accurate information on the dispersion relations of individual bands can be obtained by angle-resolved photoelectron spectroscopy in which the angle of emission of the electrons is measured. Representative results for Ge are shown in Fig. 17 (Hsieh *et al.*, 1984).

7.2 Inverse Photoemission

Low-lying empty states can be investigated by inverse photoemission; in particular, combining photoemission (whereby an electron is extracted from the valence band) and inverse photoemission (whereby an electron is injected into the conduction band) measurements, the energy gap between single-particle valence and conduction states can be directly determined. In this respect, it is to be noted that the gaps measured by optical spectroscopy (whereby an electron is excited from the valence into the conduction band) are in general smaller than the difference between single-particle states because of the attractive electron–hole interaction. Figure 18 shows the energy dispersion of conduction-band states in Si as derived from angle-resolved inverse photoemission data (Straub *et al.*, 1985) along with the topmost valence band as derived from angle-resolved photoemission (Uhrberg *et al.*, 1984); the energy gaps at the L point of the Brillouin zone are compared with the optical E_1 and E_1' gaps (Madelung, 1982): a marked lowering of the E_1 gap (by about 0.5 eV) due to the excitonic effect is evident.

8. DIELECTRIC SCREENING

We have already assumed, when dealing with excitons, that the Coulomb attraction between charges is screened by a wave-vector–dependent dielectric function $\epsilon(\mathbf{k})$. The usual dielectrical constant ϵ_0 corresponds to the screening of a fixed charge at large distance and coincides with the limit of the optical dielectric function (see Sec. 6) $\epsilon(\omega)$ for $\omega \to 0$.

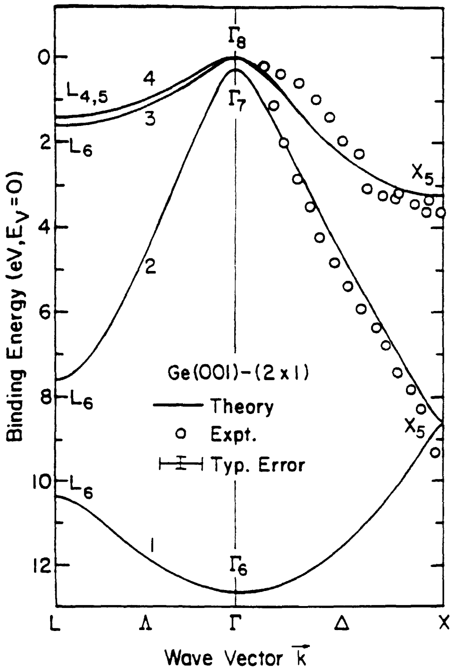

FIG. 17. Valence-band dispersion curves of Ge: the circles are obtained from angle-resolved photoemission (Hsieh *et al.*, 1984), and the solid lines are from a pseudopotential calculation (Chelikowsky and Cohen, 1976).

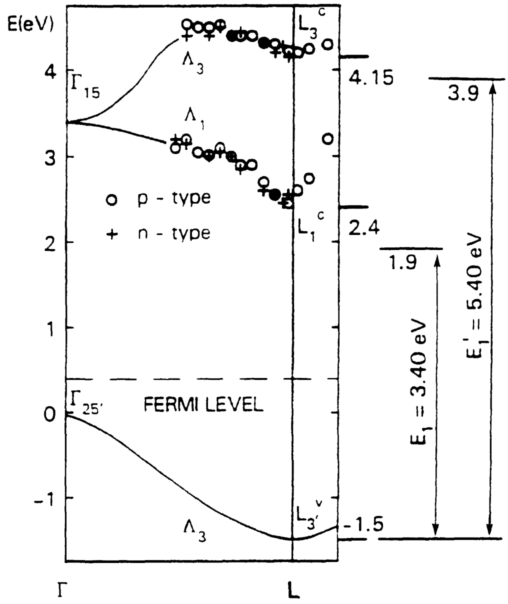

FIG. 18. Energy dispersion of conduction-band states of Si from inverse photoemission (Straub *et al.*, 1985). Also shown are the valence band from photoemission (Uhrberg *et al.*, 1984) and, for comparison, the optical gaps at *L* (Madelung, 1982).

8.1 General Theory

Since in elemental semiconductors we do not have polarization due to phonons, the origin of the dielectric screening is all in the modification of the electronic functions of the crystal produced by the electric charge. This, in principle, can be computed from the detailed knowledge of the band structure, taking into account correlation effects. A general expression (Lindhardt formula), which can be derived from the quantum definition of polarizability (in the random-phase approximation), for the diagonal expression of the dielectric tensor [i.e., neglecting umklapp processes that would produce off-diagonal tensor terms $\epsilon(\mathbf{k}, (\mathbf{k} + \mathbf{G}))$, with $\mathbf{G} \neq 0$] is given by

$$\epsilon(\mathbf{k},\omega) = 1 + \frac{4\pi e^2}{k^2} \sum_{v,c} \int_{\mathrm{BZ}} \frac{2d^3k'}{(2\pi)^3} \frac{|\langle \Psi_{c\mathbf{k}'}|e^{-i\mathbf{k}\cdot\mathbf{r}}|\Psi_{v\mathbf{k}'+\mathbf{k}}\rangle|^2}{E_c(\mathbf{k}') - E_v(\mathbf{k}' + \mathbf{k}) - \hbar\omega - i\hbar\epsilon^+}, \tag{36}$$

ϵ^+ denoting that the positive zero limit is considered (see Bassani and Pastori Parravicini, 1975). It can be shown that for $\mathbf{k} = 0$, the above expression reduces to that described in Sec. 6 for the optical transitions. We can also observe that the static limit ($\omega \rightarrow 0$) has a finite value in semiconductors because of the energy gap, and its asymptotic limit for large wave vectors is 1. It also follows from (36) that the values of the dielectric screening increase with decreasing band gaps, which explains the large values of the dielectric constants in semiconductors and their trend from C to Sn ($\epsilon_0 = 12$ in Si, 16 in Ge, 24 in gray tin).

8.2 Simplified Model Calculations

Many authors have performed calculations of the static screening function $\epsilon(\mathbf{k},0)$, using approximate models for the band structure, and also improving on expression (36) by adding off-diagonal terms with $\mathbf{G} \neq 0$ to the polarization propagator. The first of such simplified models is the "free electron with a gap" model due to Penn (1962), who fitted the numerical results with a simple formula that contains all the relevant features:

$$\epsilon(k) \simeq 1 + \left(\frac{\hbar\omega_p}{E_g}\right)^2 \left(1 - \frac{E_g}{4E_F}\right) \times \left[1 + \frac{E_F}{E_g}\left(\frac{k}{k_F}\right)^2 \left(1 - \frac{E_g}{4E_F}\right)^{1/2}\right]^{-2}, \tag{37}$$

E_F being the Fermi energy given by the electron density and $\hbar k_F$ the corresponding Fermi momentum, and E_g is the gap introduced between the Fermi energy and the empty states.

For the purpose of illustration we show in Fig. 19 the computed values of the static dielectric function for silicon, in a number of approximations. It can be observed that, in the case of excitons or of shallow impurity states (see Sec. 9) in elemental semiconductors, the Bohr radius is so large that the use of the approximation $\mathbf{k} = 0$, appropriate to charges at infinite distance, is justified. This will not be the case in strong insulators, or organic materials, where the interaction is localized inside the unit cell and the limit of large wave vectors is more appropriate.

9. IMPURITY LEVELS

Foreign atoms, in interstitial or substitutional sites, are the most important defect in semiconductor crystals. The ability to control

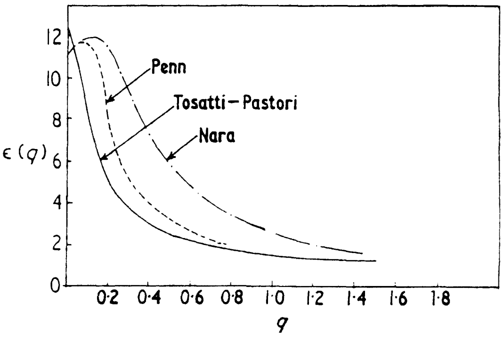

FIG. 19. Dielectric screening function of silicon as a function of the wave vector (in arbitrary units): the dashed-dotted line is from Nara (1965), the dashed one from Penn (1962), and the solid one from Tosatti and Pastori Parravicini (1971).

the content of impurities, and their beneficial as well as pernicious effects, is a key issue in semiconductor technology: The concentration of undesired dopants in Si is typically kept below 10^{12} cm^{-3}. An impurity breaks the translational symmetry of the perfect crystal and may give rise to electronic states, more or less localized around the defect site, with energies lying in the forbidden energy gap of the host material. The main character of these impurity levels is whether they are close to the conduction- or valence-band edges (shallow donor or acceptor states) or rather are close to the middle of the gap (deep states). In the first case, they dominate the transport properties, providing free carriers to the conduction or valence band; in the second case, they mainly affect the recombination of free electron–hole pairs by trapping them at the same site. Furthermore, besides shallow and deep states, an impurity may bring about resonant states that are still partly localized around the defect site but have energy coinciding with that of extended states of the host crystal.

All the properties of the electronic states associated with impurity levels depend on the electronic band structure of the host semiconductor as well as on the impurity potential (for substitutional sites, the difference between the foreign- and replaced-atom potentials). General theoretical schemes have been developed to handle all types of impurity states in a unified way (Bassani *et al.*, 1969); in the following, though, the hydrogenic model based on the effective-mass approximation (see Sec. 4) will be mostly referred to, because it is physically transparent and well suited for the shallow states that are used for doping (see Sec. 11).

9.1 Shallow Levels

In group-IV elemental semiconductors, such as Si and Ge, substitutional group-III impurities, such as Al and B, behave as shallow acceptors, and group-V ones, such as P and As, as shallow donors (see Tables 1 and 2). The extra hole or electron, with respect to the fully occupied valence band, is bound to the defect site by the long-range Coulomb potential due to the charge difference between the impurity and host nuclei. Because of the low effective masses ($m^* \simeq 0.2\ m$) and the high dielectric constants ($\epsilon \gtrsim 10$) typical of semiconductors, the corresponding localized states extend over many unit cells, and these impurities can be described as a solid-state analog of the hydrogen atom with scaled-down Rydberg energy $R^* = m^*e^4/2\hbar^2\epsilon^2$ ($\simeq$20 meV) and Bohr radius $a^* = \hbar^2\epsilon/m^*e^2$ ($\simeq$50 Å).

Table 1. Group-III acceptor levels in silicon and germanium (measured from the valence band in meV).

	B	Al	Ga
Si	45	67	72
Ge	10.4	10.2	10.8

This simple picture can be easily extended to include details of the host band structure, such as the many-valley conduction-band minima and the degeneracy of the valence-band maximum (Ramdas and Rodriguez, 1981). In particular, because the electron pockets in Si and Ge are ellipsoids of revolution with a longitudinal mass different from the transverse mass (see Sec. 4), the three *p* levels are no longer degenerate but split into a singlet p_0 level and a doublet $p_\pm$ level, as observed in absorption spectroscopy when electric dipole transitions from the 1*s* ground state of the donor impurity are excited (see Fig. 20). Furthermore, as a result of the interaction between different equivalent valleys, the 1*s* donor ground states split into one A_1 singlet, one E doublet, and one T_2 triplet in Si (which has six electron pockets) and into one A_1 singlet and one T_2 triplet in Ge (which has four electron pockets).

Aside from complications due to details of the host band structure, the simple hydrogenic model cannot account for significant shifts in binding energies of different impurities (chemical shifts). These are usually important only for the ground state (more localized) and are due to the departure of the impurity potential from a Coulomb potential at short distances from the defect site. This

Table 2. Group-V donor levels in silicon and germanium (measured from the conduction band in meV).

	P	As	Sb
Si	45	54	39
Ge	12	12.7	9.6

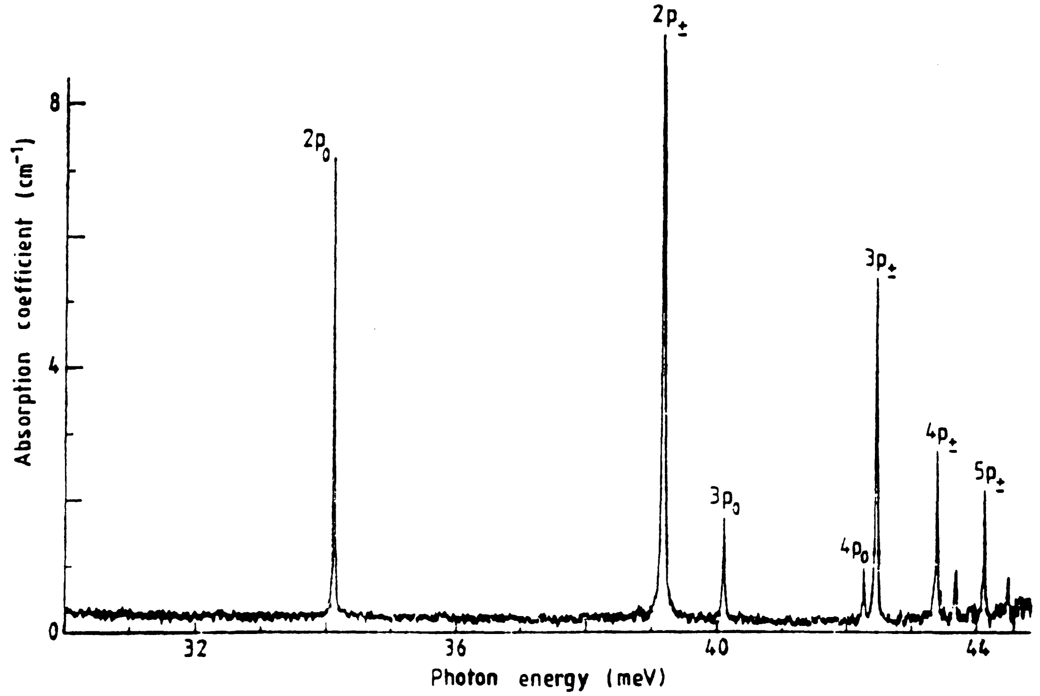

FIG. 20. Excitation spectrum of P donors in Si at liquid-He temperature (Jagannath *et al.*, 1981).

central-cell correction is marked for nonisocoric impurities (i.e., those with a core electron distribution different from that of the host), for which a pseudopotential approach has been used (Pantelides, 1978). Furthermore, at short distances (large wave vectors) the use of a (large) constant dielectric function is questionable as the screening efficiency is decreased (see Sec. 8).

9.2 Resonant Levels

If the host band structure has a secondary extremum (in either the conduction or valence band), the impurity potential may give rise to a localized state associated with it. Such a level is degenerate with extended Bloch states of the host crystals with values of **k** corresponding to the principal extremum. Therefore, the impurity level close to the secondary extremum is broadened by the interaction with the degenerate continuum of the band states belonging to the principal extremum. Such a situation occurs, for instance, for acceptor states at the top of the valence band of Si and Ge with the split-off hole band (Γ_7^+) playing the role of the secondary maximum, as depicted in Fig. 21.

9.3 Deep Levels

When the wave function of an impurity state is localized over a small number of unit cells around the defect site, the effective-mass approximation breaks down, and no simple theoretical description can be given. Such impurities have energy levels deep in the forbidden energy gap. For instance, in silicon, Pd has an acceptor level (at 0.34 eV from the valence band) and O has two donor levels (at 0.16 and 0.51 eV from the conduction band) and two acceptor levels (at 0.41 and 0.74 eV from the valence band); in germanium, Cu has two acceptor levels (at 0.33 and 0.53 eV from the valence band) and Mn has two acceptor levels (at 0.16 and 0.42 eV from the valence band). The corresponding wave functions, when expanded in Bloch states, comprise a large part of the Brillouin zone; they may therefore act as efficient radiative recombination centers for electron–hole pairs in indirect semiconductors, partic-

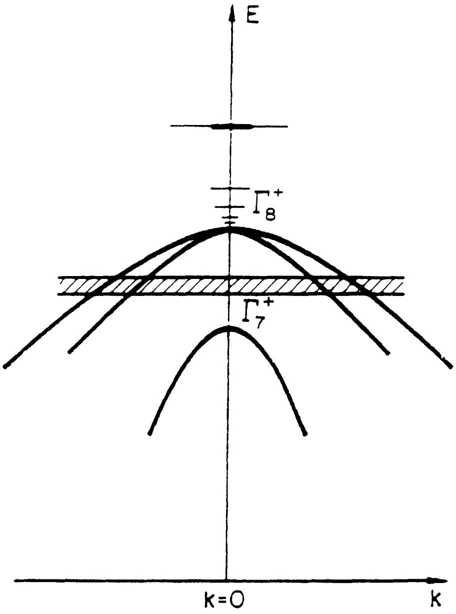

FIG. 21. Acceptor states near the top of the valence band, the shaded level representing a resonant state associated with the secondary maximum of the Γ_7^+ band.

ularly when they have an energy level close to the middle of the band gap. Such is the case, for example, of Au in Si: increasing the Au concentration from 10^{14} to 10^{18} cm^{-3} reduces the lifetime of nonequilibrium free carriers from 1 μs to 0.1 ns. A special case of localized levels is the one due to isoelectronic impurities, such as Ge in Si, for which there is no long-range Coulomb potential, but only an impurity potential localized in the central cell. Such impurities may or may not split a bound state from a band of extended states depending on the ratio between the strength of the impurity potential and the band width: for a larger band width (dependent on the host), a stronger central-cell potential (dependent on the impurity-host difference) is required to introduce a localized state.

10. EXTERNAL PERTURBATIONS

From the experimental point of view, much information on the properties of semiconductors has been obtained by measuring changes induced by externally applied perturbations. The latter can preserve the symmetry of the crystal (as for hydrostatic pressure) or lower it (as for uniaxial stress or electric and magnetic fields), in which case detailed information on the symmetry properties of, say, the critical points of the electronic bands can be obtained. A special case of symmetry-lowering perturbation is the presence of surfaces or interfaces that reduce the translational symmetry of a crystal (Stella, 1993); surface electronic levels in Ge were first observed by Chiarotti *et al.* in 1968.

The possibility of employing a phase-sensitive detection technique in conjunction with a periodic modulation of the external perturbation improves the signal-to-noise ratio by several orders of magnitude on the one hand, and, on the other hand, allows one to enhance the sharp features of, say, the reflectivity with respect to a smoothly varying background. Among the pioneering works in modulation spectroscopy, there are the studies of surface traps in Ge (Chiarotti *et al.*, 1962) and of the Franz–Keldysh effect in Ge and Si (Seraphin and Hess, 1965; Frova and Handler, 1965).

10.1 Hydrostatic Pressure

The application of hydrostatic pressure brings about a lattice-constant change that, through a modification of the crystal (pseudo)potential, may significantly affect the electronic band structure. In particular, the optical gaps between valence and conduction states at various points in the Brillouin zone of semiconductors such as Ge and Si shift in a characteristic way determined by symmetry: the gap between the top of the valence band at Γ (Γ'_{25}) and the conduction band at Γ (Γ'_2) increases at a rate of about 10 meV/kbar, the one between Γ'_{25} and the conduction band along the [111] direction (Λ_1) increases at a rate of about 5 meV/kbar, whereas the one between Γ'_{25} and the conduction band along the [100] direction (Δ_1) decreases at a small rate of about -1.5 meV/kbar. Thus, at elevated pressures (higher than about 50 kbar), the minimum of the conduction band in Ge moves from L_1 to Δ_1; as a consequence, the many-valley structure at the bottom of the conduction band becomes analogous to that of silicon with ellipsoidal electron pockets along the cubic axes.

10.2 Uniaxial Stress

Uniaxial stress lowers the symmetry of cubic semiconductors and, therefore, induces qualitative changes in their properties. Piezospectroscopy has been instrumental in establishing the main features of the electronic bands of Si and Ge. For instance, reflectivity of Ge around 2.2 eV is dominated by interband transitions involving states along the Λ line in the BZ: uniaxial stress along the [001] direction does not split the degeneracy among the eight [111] valleys, whereas uniaxial stress along the [111] direction induces a difference between the two of them along the stress axis and the other six; the latter stress brings about a marked polarization dependence of the reflectivity (Gerhardt, 1965). More striking features are observed in the piezospectroscopy of shallow impurity levels, whereby the bound states closely follow the shifts of the respective band extrema (Ramdas and Rodriguez, 1981).

10.3 Electric Field

The application of an electric field is another powerful technique to probe the electronic structure of semiconductors, particularly in conjunction with modulation spectroscopy (Cardona, 1969). An electric field **E** drastically alters the symmetry of a crystal, breaking the translational invariance of the potential along the direction of the field; as a consequence, the component of the wave vector **k** along the field is no longer a good quantum number, but changes with time according to

$$\mathbf{k}(t) = \mathbf{k}(0) - e\mathbf{E}t/\hbar. \qquad (38)$$

Considering the corresponding change in the character of the wave functions, the effects of an electric field on the optical properties can be calculated in the vicinity of a critical point (where they are particularly important). For an absorption edge (M_0 critical point), the main changes in $\epsilon_2(\omega)$ are an exponential tail in the gap and an oscillatory behavior above threshold (Franz–Keldysh effect), as sketched in Fig. 22. The exponential tail is proportional to

$$\epsilon_2(\omega) \propto \exp\left[-\frac{4}{3}\sqrt{\frac{(\omega_0 - \omega)^3 2\mu\hbar}{e^2E^2}}\right], \qquad (39)$$

where μ is the reduced mass and ω_0 is the frequency of excitations at the critical point. The changes in optical constants measured by electroreflectance modulation spectros-

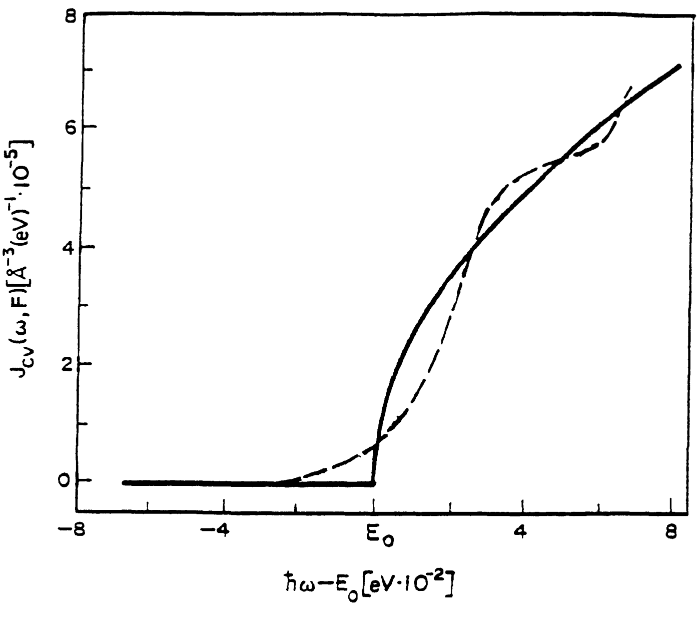

FIG. 22. Electric field effects on the behavior of the joint density of states in the vicinity of an absorption edge (M_0); the solid line represents the zero-field case.

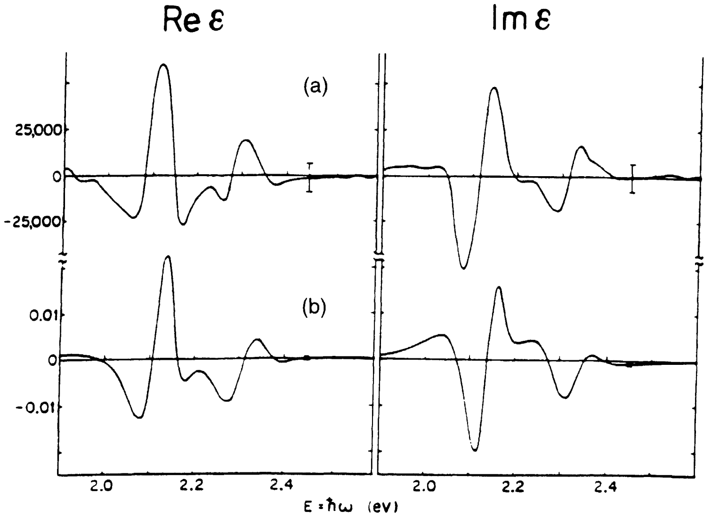

FIG. 23. Optical constants of Ge at the E_1 and $E_1 + \Delta_1$ gaps from low-field electroreflectance data (b) compared to third-order derivatives (a) calculated from ellipsometric data (Aspnes, 1972).

copy have a very simple interpretation in the regime of low applied fields. In this case, the spectra are simply proportional to the third derivative of the respective optical constant with respect to frequency according to

$$\Delta\epsilon(\omega) \propto \frac{e^2E^2}{\mu\hbar}\frac{1}{\omega^2}\frac{d^3}{d\omega^3}[\omega^2\epsilon(\omega)], \qquad (40)$$

as shown in Fig. 23 (Aspnes, 1972).

10.4 Magnetic Field

A magnetic field alters the electronic states drastically: The motion in the plane perpendicular to the field is quantized into discrete levels (Landau levels) arranged in a regular ladder (at least for simple bands) with a spacing given by the cyclotron frequency ω_c, whereas the dispersion along the field remains unaffected. Therefore, the three-dimensional density of states collapses into a series of one-dimensional densities of states (each corresponding to a Landau level), as shown (for an absorption edge) in Fig. 24. The Landau quantization is responsible for the cyclotron resonance described in Sec. 4 and for many other phenomena, such as an oscillatory behavior of the magnetization (de Haas–van Alphen effect) or of the conductivity (Shubnikov–de Haas effect) as a function of magnetic field (Kittel, 1966).

Magneto-optical experiments demonstrate very directly the presence of Landau levels. Figure 25 (Aggarwal *et al.*, 1967) shows a

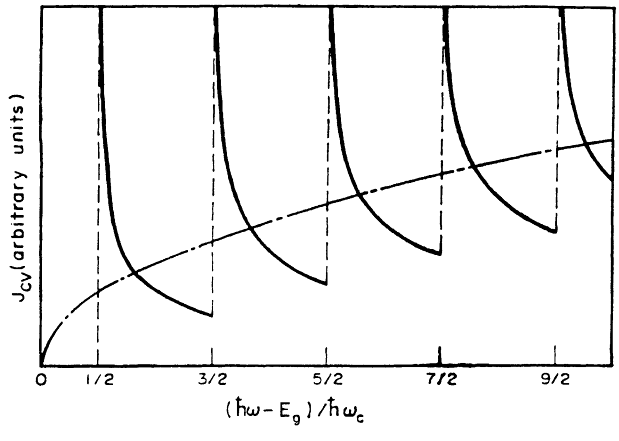

FIG. 24. Joint density of states in the presence of a magnetic field for a critical point M_0 (the dot-dashed line represents the zero-field density of states); the one-dimensional singularities do not take into account broadening effects.

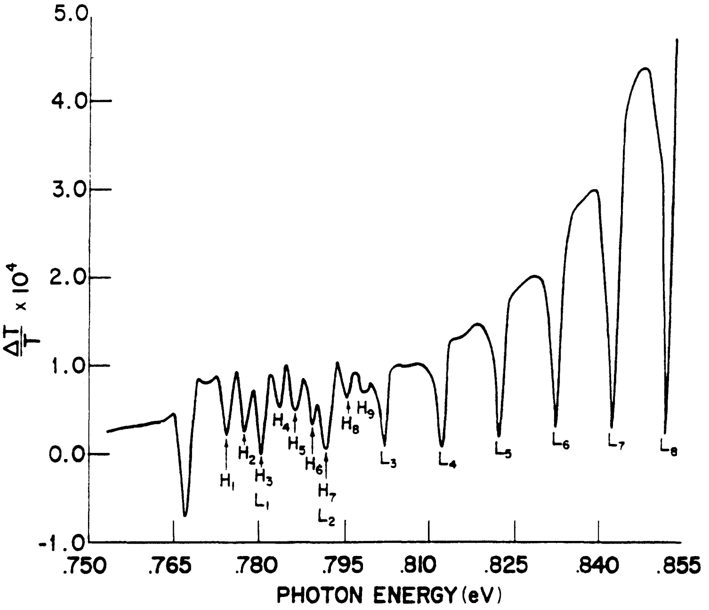

FIG. 25. Magnetopiezotransmission spectrum in Ge (Aggarwal *et al.*, 1967); see text for details.

piezomodulated transmission spectrum of Ge in the presence of a magnetic field along the [110] axis. The two series of sharp features (L_1, L_2, . . .; H_1, H_2, . . .) correspond to transitions between the same level at the top of the valence band and the Landau levels associated with the four electron pockets along the [111] directions: Two of these are perpendicular to the magnetic field and have a higher effective mass and a lower Landau level spacing ($H_{n+1} - H_n$) than the other two pockets (associated to the L series). The Landau level energies increase (approximately linearly) with magnetic field, and the results of magneto-optical experiments are usually presented as a "fan" diagram from which important material parameters (such as effective masses) can be derived (Madelung, 1982).

11. ELECTRICAL CONDUCTIVITY

The most typical property of pure semiconductors is the temperature dependence of their resistivity. From the technological point of view, on the other hand, the possibility of controlling their resistivity through the introduction of a suitable concentration of shallow impurities (doping) is of paramount importance (Sze, 1985). The main factors determining the dc electrical conductivity of homogeneous bulk Si and Ge, pure as well as doped, are discussed in the following.

11.1 Free-Carrier Conductivity

In a cubic semiconductor, the low-field conductivity tensor reduces to a scalar σ given (see INSULATORS AND SEMICONDUCTORS, CONDUCTIVITY IN) by

$$\sigma = e(n\mu_n + p\mu_p), \tag{41}$$

where n and p are the densities of conduction-band electrons and valence-band holes, respectively, and μ_n and μ_p are the corresponding mobilities given by

$$\mu_{n,p} = (e/m^{(\sigma)}_{n,p})\tau_{n,p}, \tag{42}$$

where $m^{(\sigma)}$ is the appropriate effective mass and τ is the appropriate momentum relaxation time. For a degenerate semiconductor, as in the case of a metal, τ corresponds to the momentum relaxation time of carriers at the Fermi energy, whereas for a nondegenerate semiconductor it reduces to

$$\tau = \langle v^2\tau(v)\rangle/\langle v^2\rangle, \tag{43}$$

where the averages are taken over a Maxwell–Boltzmann distribution, v is the carrier thermal velocity, and the scattering is assumed to be energy dependent, but isotropic. For a many-valley band extremum, such as for the electron pockets in Si and Ge, the conductivity effective mass to be used above is an average given by

$$1/m_n^{(\sigma)} = (2/m_T + 1/m_L)/3, \qquad (44)$$

and n is the total density of electrons equally distributed among the equivalent valleys. From the values of the electron pocket transverse and longitudinal masses (see Sec. 4), one obtains for Si $m_n^{(\sigma)} \simeq 0.27\ m$ and for Ge $m_n^{(\sigma)} \simeq 0.12\ m$. When more than one electron or hole bands are populated, e.g., light- and heavy-hole bands, their contributions add up in the above expression for σ.

The conductivity of a semiconductor, in contrast to that of a metal, varies strongly (see below) because of the dramatic dependence of the electron and hole densities on the temperature and the doping level; a smoother variation is also brought about by mobility changes.

11.2 Carrier Density

In a nondegenerate semiconductor, the density of electrons in the conduction band, n, and of holes in the valence band, p, can be expressed as

$$n(T,\mu) = N_c(T)\exp[(\mu - E_c)/k_BT], \qquad (45)$$

$$p(T,\mu) = N_v(T)\exp[(E_v - \mu)/k_BT], \qquad (46)$$

μ being the chemical potential and

$$N_c = \left(\frac{k_BT}{\pi\hbar^2}\right)^{3/2} \frac{(m_c^{(D)})^{3/2}}{\sqrt{2}}, \qquad (47)$$

$$N_v = \left(\frac{k_BT}{\pi\hbar^2}\right)^{3/2} \frac{(m_v^{(D)})^{3/2}}{\sqrt{2}}, \qquad (48)$$

where $m^{(D)}$ is the density-of-states effective mass given for the conduction band by an average over the equivalent valleys,

$$m_c^{(D)} = g^{2/3}(m_T^2 m_L)^{1/3}, \qquad (49)$$

g being the valley degeneracy (6 for Si, 4 for Ge), and for the valence band by an average over heavy and light holes,

$$m_v^{(D)} = (m_{lh}^{3/2} + m_{hh}^{3/2})^{2/3}. \qquad (50)$$

The densities of electrons and holes in thermal equilibrium at temperature T satisfy the law of mass action

$$np = N_cN_ve^{-E_g/k_BT}. \qquad (51)$$

In a pure (intrinsic) semiconductor, the density of electrons n_i in the conduction band equals the density of holes p_i in the valence band, and they are given by

$$n_i = p_i = \sqrt{N_cN_v}\exp(-E_g/2k_BT);$$

the intrinsic concentration of carriers is dominated by the exponential dependence on $E_g/2k_BT$. From the effective-mass and energy-gap values (see Secs. 2 and 4), one obtains at room temperature (300 K) for Si $N_c \simeq 3 \times 10^{19}$ cm^{-3}, $N_v \simeq 1 \times 10^{19}$ cm^{-3}, $n_i = p_i \simeq 1.4 \times 10^{10}$ cm^{-3}, and for Ge $N_c \simeq 1 \times 10^{19}$ cm^{-3}, $N_v \simeq 5 \times 10^{18}$ cm^{-3}, $n_i = p_i \simeq 2.3 \times 10^{13}$ cm^{-3}. For the nondegenerate regime here considered, the intrinsic Fermi level (chemical potential) is given by

$$\mu_i = \tfrac{1}{2}(E_c + E_v) + \tfrac{3}{4}k_BT\ln(m_v^{(D)}/m_c^{(D)}) \qquad (52)$$

and is close to the center of the gap [i.e., $(E_c + E_v)/2)$].

In a (partially compensated) doped semiconductor with a concentration of donors N_d and of acceptors N_a, if $N_d > N_a$ (n type) the electron (majority-carrier) density is approximately $n = N_d - N_a$ and the hole (minority-carrier) density is correspondingly $p = n_ip_i/n$; if $N_d < N_a$ (p-type), similar formulas hold with the role of majority and minority carriers interchanged. In n-type semiconductors the chemical potential μ is shifted a bit closer to E_c with respect to μ_i, as given by

$$\frac{N_d - N_a}{2n_i} = \sinh\left(\frac{\mu - \mu_i}{k_BT}\right); \qquad (53)$$

the converse is true for p-type semiconductors. In the above, the temperature has been assumed to be high enough to ionize all the impurities but not so high that the concentration of intrinsic carriers (i.e., those ther-

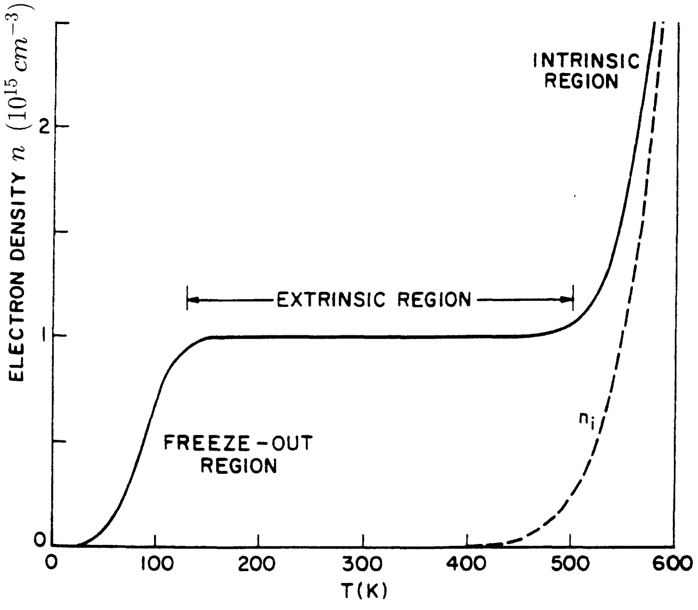

FIG. 26. Typical temperature dependence of majority-carrier concentration in a doped semiconductor. The intrinsic carrier concentration is shown by the dashed line for comparison. The numerical values are for Si with $N_d = 10^{15}$ cm^{-3}.

mally excited across the energy gap) should be comparable to the majority-carrier concentration. In this case, the carrier concentration is independent of temperature and determined by the doping level (extrinsic regime). At lower temperatures, the concentration of ionized impurities decreases because more and more electrons fall into the impurity bound states, and the concentration of free carriers drops (freeze-out regime). The temperature dependence of the majority-carrier concentration in a doped semiconductor over a wide temperature range is depicted in Fig. 26.

In all of the above, it has been assumed that the chemical potential is within the energy gap, so that the thermal distribution of electrons in the conduction band and holes in the valence band is given by the (classical) Boltzmann distribution (nondegenerate regime). When the level of doping is extremely high, the number of electrons may be larger than $(m_c^{(D)}k_BT/\hbar^2)^{3/2}$ or the number of holes larger than $(m_v^{(D)}k_BT/\hbar^2)^{3/2}$, in which case the majority carriers are in the degenerate regime (like the electrons in a metal) and must be described using the Fermi distribution with a chemical potential (possibly) higher than E_c or lower than E_v.

For the case of a *p-n* junction at equilibrium, the chemical potential is the same on both sides, being closer to the conduction-band edge in the *n*-type material and to the valence-band edge in the *p*-type material. This situation implies a bending of the electron energy bands (shown in Fig. 27), which is associated to the electrostatic field appearing in the depletion layer at the junction. The corresponding potential acts as a barrier for the diffusion of the majority carriers; lowering this potential barrier by means of an applied forward bias leads to an exponentially increasing current. The rectifying properties of *p-n* junctions are discussed at length in the various articles on semiconductor devices.

11.3 Carrier Mobility

The carrier mobility depends on the temperature and the doping level through the average momentum relaxation time τ,

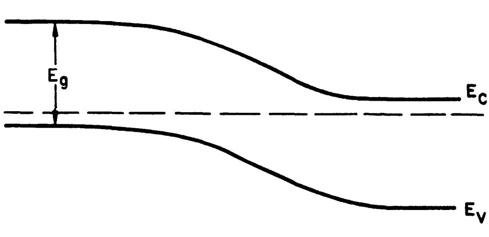

FIG. 27. Energy-band diagram of a *p-n* junction.

which is determined by the various scattering mechanisms. At room temperature, in pure or moderately doped elemental semiconductors, the mobility is limited mainly by longitudinal acoustic-phonon deformation-potential scattering, whereas at lower temperatures or heavy doping levels, the ionized impurity scattering is dominant (see INSULATORS AND SEMICONDUCTORS, CONDUCTIVITY IN). The acoustic-phonon scattering rate $1/\tau_{ac}$ depends on temperature like $T^{3/2}$. In fact, from perturbation theory (Fermi's "golden rule"), it is proportional to the product of the interaction matrix element squared and the density of final states; the former contains a factor of T coming from the number of phonons present (the energy of an acoustic phonon being much smaller than the thermal energy) and the latter a factor of $T^{1/2}$ as the density of states of the deflected electron scales with the square root of its kinetic energy. Experimentally (Ziman, 1960), the phonon-scattering–limited mobility scales like $T^{-1.6}$ in n-type Ge, $T^{-2.3}$ in p-type Ge, $T^{-2.6}$ in n-type Si, and $T^{-2.3}$ in p-type Si; apart from the first case, there are significant deviations from the $T^{3/2}$ behavior, mostly due to the mixture of acoustic and optical phonon scattering (and, for n-type Si, intervalley scattering). The mobility values in pure materials at room temperature are for Si $\mu_n \simeq 1500$ cm^2/(V s), $\mu_p \simeq 500$ cm^2/(V s), and for Ge $\mu_n \simeq 3900$ cm^2/(V s), $\mu_p \simeq 1900$ cm^2/(V s).

The impurity Coulomb scattering rate $1/\tau_{im}$, instead, is proportional to $T^{-3/2}N_{im}$, where N_{im} is the density of ionized impurities. In fact, it is given by N_{im} times the product of the Rutherford cross section and the electron velocity; a factor of T^{-2} comes from the energy dependence of the former (i.e., faster electrons are deflected less) and a factor of $T^{1/2}$ from the latter. This result does not change much when one considers the free-carrier screening (which essentially cuts off the small-angle scattering). Experimentally, it is not easy to see the $T^{3/2}$ dependence of the impurity-scattering–limited mobility as the temperature range in which this mechanism is dominant is restricted at low temperatures by the carrier freeze-out and at high temperatures by the onset of phonon scattering. However, when more sources of scattering are competing, the total resistivity is approximately given by the sum of the resistivities obtained from each single-scattering mechanism at a time (Matthiessen's rule). As a result, the mobility of doped semiconductors exhibits a maximum at the temperature corresponding to the crossover from impurity scattering ($\mu_{n,p} \propto T^{3/2}/N_{im}$) to phonon scattering ($\mu_{n,p} \propto T^{-3/2}$). Such behavior is evident in Fig. 28, which shows the electron mobility in Si versus temperature for various doping levels (Sze, 1985).

The temperature dependence of the conductivity of Si at different doping levels is shown in Fig. 29 (Morin and Maita, 1954). The intrinsic regime is evident over most of the temperature range; the flattening of some of the curves at about room temperature corresponds to the extrinsic regime, in which the conductivity is determined by the doping level. The temperature dependence of σ is dominated by that of the carrier concentration in the intrinsic regime, whereas its weaker variation in the extrinsic regime is due to the temperature dependence of the mobility.

11.4 Hall Effect and Magnetoresistance

Measuring the Hall coefficient is a widely used method to characterize the concentration (and type) of free carriers (see INSULATORS AND SEMICONDUCTORS, CONDUCTIVITY IN); the Hall coefficient $\mathrm{R_H}$ is given by

$$\mathrm{R_H} = \frac{p\mu_p^2 - n\mu_n^2}{ce(p\mu_p + n\mu_n)^2}. \tag{54}$$

In weak fields, the mobilities appearing in the numerator of the above expression are to be multiplied by corrective factors of order unity (i.e., the ratios of Hall mobilities to drift mobilities: μ_H/μ) that depend on the scattering mechanism and the band-structure details. For instance, for an n-type many-valley semiconductor such as Si or Ge:

$$\frac{\mu_\mathrm{H}}{\mu} = \frac{\langle v^2\tau^2\rangle\langle v^2\rangle}{\langle v^2\tau\rangle^2}\,\frac{3\gamma(\gamma + 2)}{(2\gamma + 1)^2}, \tag{55}$$

where $\gamma = m_L/m_T$ and the factor $\langle v^2\tau^2\rangle\langle v^2\rangle/\langle v^2\tau\rangle^2$ is 1.18 for acoustic deformation potential scattering and 1.93 for ionized impurity scattering; the mass anisotropy factor is also close to unity: about 0.8 for $\gamma \simeq 20$, which is appropriate for Ge (about 0.9 for Si).

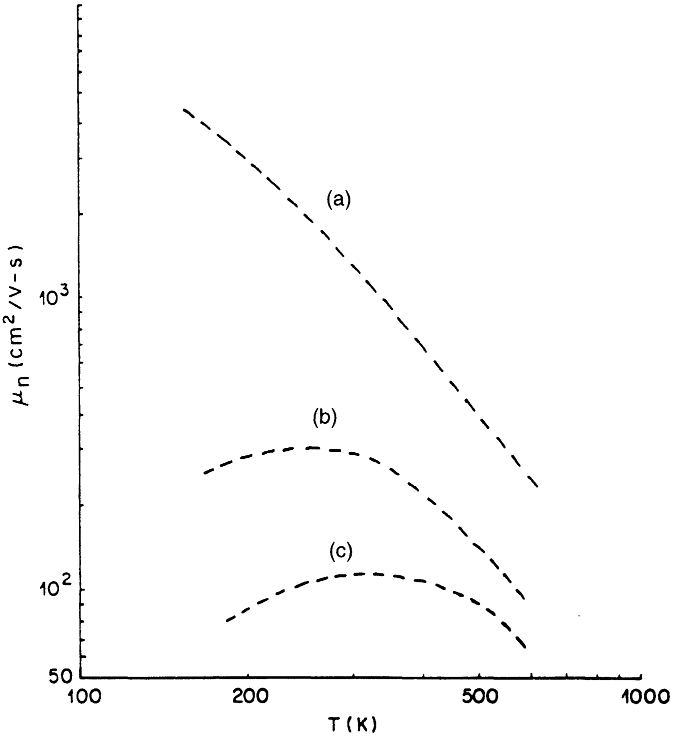

FIG. 28. Electron mobility in Si versus temperature for various doping levels (after Sze, 1985).

The magnetoresistance is much more sensitive to the mass anisotropy and other band-structure features; in particular its strong dependence on the current and magnetic-field directions with respect to the cubic axes (see Fig. 30) allows the identification of the symmetry of the carrier pockets involved (Seeger, 1991). In weak fields, the magnetoresistance $\Delta\rho$ is proportional to the square of the magnetic-field strength and varies with the direction cosines (i,j,k) and (l,m,n), respectively, of the current and the magnetic field according to the inverted Seitz equation

$$\begin{aligned}\Delta\rho/\rho B^2 &= b + c(il + jm + kn)^2 \\ &\quad + d(i^2l^2 + j^2m^2 + k^2n^2), \qquad (56)\end{aligned}$$

where b, c, and d are material-dependent coefficients.

11.5 High-Field Effects

So far we have considered the low-field conductivity corresponding to the Ohmic regime. With increasing field strength, the free-carrier drift velocity approaches their thermal velocity and hot-carrier effects become important (see INSULATORS AND SEMICONDUCTORS, CONDUCTIVITY IN). In particular, in Si and Ge, the high-field mobility (limited mainly by inelastic phonon scattering) eventually decreases linearly with the electric field and the drift velocity saturates, as shown in Fig. 31 (Levinshtein and Simin, 1992). The room-temperature saturation velocities are in Ge, for both electrons and holes, about 5 × 10^6 cm/s, and in Si, for both electrons and holes, about 10^7 cm/s, whereas the corresponding saturation fields are about 5 kV/cm in n-type Ge, about 15 kV/cm in p-type Ge, about 15 kV/cm in n-type Si, and about 60 kV/cm in p-type Si (but a strongly non-Ohmic behavior is already evident at smaller fields). In the presence of hot electrons, the equivalent valleys need not be equally populated, and, as a consequence, even in a cubic material like n-type Ge, the current need not be parallel to the electric field (Sasaki–Shibuya effect). Moreover, also inequivalent, higher-energy valleys may become populated: For instance, dramatic changes in the Hall coefficient of n-type Ge at electric fields higher than 1 kV/cm are brought about by the transfer of elec-

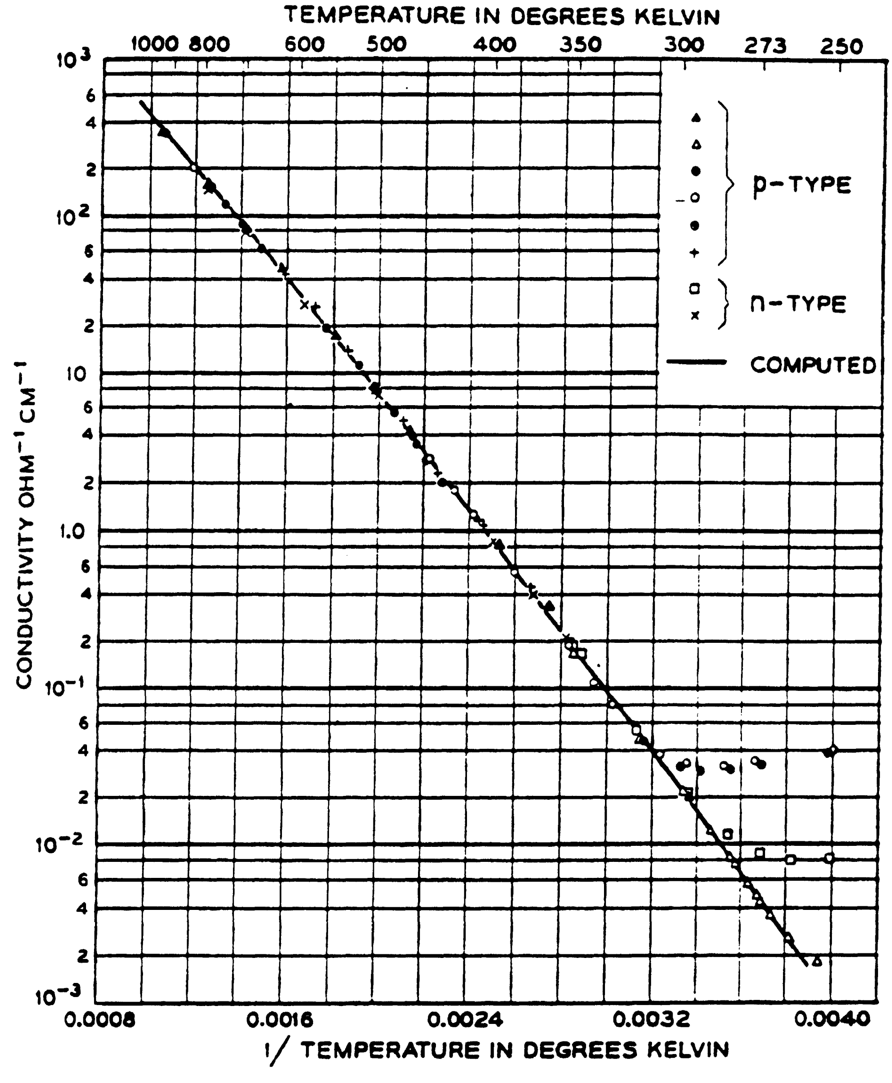

FIG. 29. Temperature dependence of conductivity in Si at various doping levels (Morin and Maita, 1954).

trons from the ⟨111⟩ lowest minima to the ⟨100⟩ higher minima (Heinrich *et al.*, 1970); the Hall coefficient is dominated by the contribution of the ⟨111⟩ high-mobility valleys and starts to increase as their population starts to decrease. In several compound semiconductors such as GaAs, such a carrier transfer to inequivalent valleys is responsible for a decrease of drift velocity at high electric fields, which leads to peculiar current oscillations (Gunn effect). In the presence of hot holes, complex features of the degenerate valence band become important—in particular, the nonparabolicity of the hole dispersion, i.e., the variation with kinetic energy of the effective mass (Costato and Reggiani, 1970). If one increases the electric field strength even beyond the saturation regime, a sudden "runaway" growth of the conductivity is eventually observed (breakdown). This is due to the avalanche multiplication of the number of free carriers: The field is so strong that a carrier can gain between two collisions enough kinetic energy to knock out an electron from the valence to the conduction band (impact ionization). Typical break-

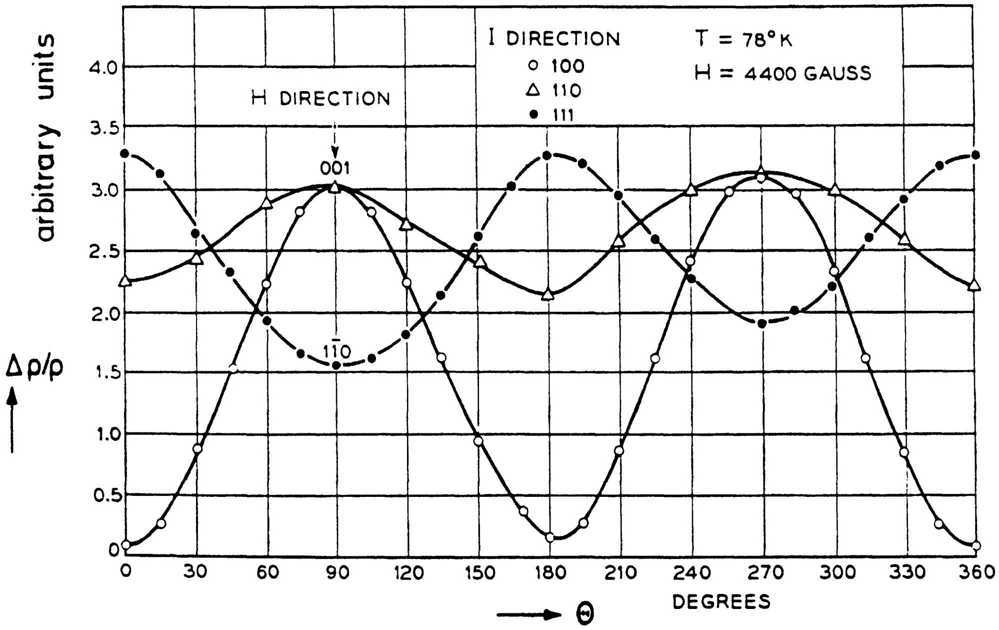

FIG. 30. Magnetoresistance of *n*-type Si as a function of the angle between the current and the magnetic field (in a [010] plane) at liquid-nitrogen temperature (after Pearson and Herring, 1954).

down electric fields are 10^5 V/cm for Ge and 3×10^5 V/cm for Si.

11.6 Impurity-Band Conductivity

In doped semiconductors, when the temperature is lowered so that from the extrinsic regime one enters into the freeze-out regime, the conductivity drops very rapidly because the concentration of free carriers becomes negligible. If the doping level is not too low, though, the resistivity increase flattens out because of the contribution of impurity-

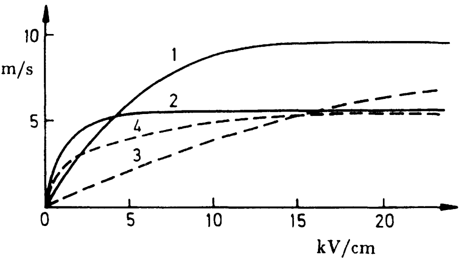

FIG. 31. Drift velocity versus electric field for electrons (solid lines) and holes (dashed lines) in Si (curves 1 and 3) and Ge (curves 2 and 4) at room temperature (after Levinshtein and Simin, 1992).

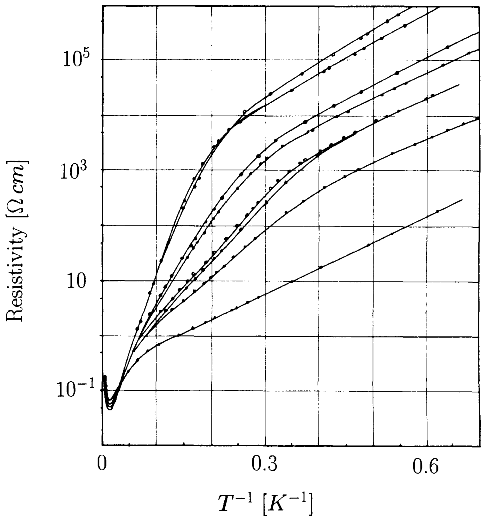

FIG. 32. Impurity-band conductivity in *p*-type Ge; the doping level increases from the top curve ($N_t \simeq 2.1 \times 10^{16}$ cm^{-3}) to the bottom one ($N_t \simeq 7.3 \times 10^{16}$ cm^{-3}) (Fritzsche, 1955).

bound carriers to the conductivity (see Fig. 32). This so-called impurity-band conductivity can be described by two types of fixed-range (nearest neighbor) hopping processes with two different activation energies:

$$\sigma = c_1 e^{-E_1/k_BT} + c_2 e^{-E_2/k_BT}; \tag{57}$$

the one with lower activation energy dominates σ at very low temperatures and is due to the hopping from neutral to ionized impurities (in partially compensated materials), while the other (not always evident in the curves of Fig. 32) involves the double occupancy of impurities. When the doping level is very high ($\simeq 10^{18}$ cm^{-3}), the wave functions pertaining to different impurities overlap each other to the point that their energy levels broaden into an impurity band that merges with the conduction band (impurity metal regime).

Works Cited

Aggarwal, R. L., Zuteck, M. D., Lax, B. (1967), *Phys. Rev. Lett.* **19,** 236–238.

Aspnes, D. E. (1972), *Phys. Rev. Lett.* **28,** 168–171.

Bachelet, G. B., Hamann, D. R., Schlüter, M. (1982), *Phys. Rev. B* **26,** 4199–4228.

Bassani, F., Altarelli, M. (1983), in: N. Koch (Ed.), *Handbook on Synchrotron Radiation,* Vol. 1A, Amsterdam: North Holland.

Bassani, F., Celli, V. (1961), *J. Phys. Chem. Solids* **20,** 64–75.

Bassani, F., Iadonisi, G., Preziosi, B. (1969), *Phys. Rev.* **186,** 735–746.

Bassani, F., Pastori Parravicini, G. (1975), *Electronic States and Optical Transitions in Solids,* Oxford: Pergamon Press.

Bir, G. L., Pikus, G. E. (1974), *Symmetry and Strain-Induced Effects in Semiconductors,* New York: Wiley.

Born, M., Huang, K. (1954), *Dynamical Theory of Crystal Lattices,* Oxford: Oxford University Press.

Brust, D., Phillips, J. C., Bassani, F. (1962), *Phys. Rev. Lett.* **9,** 94–97.

Cardona, M. (1969), "Modulation Spectroscopy," in: F. Seitz, H. Ehrenreich (Eds.), *Solid State Physics,* Suppl. 11, New York: Academic Press.

Cardona, M., Ley, L. (Eds.) (1978), *Photoemission in Solids I and II,* Topics in Applied Physics Vol. 26 and Vol. 27, Berlin: Springer-Verlag.

Cavell, R. G., Kowalczyk, S. P., Ley, L., Pollak, R. A., Mills, B., Shirley, D. A., Perry, W. (1973), *Phys. Rev. B* **7,** 5313–5316.

Chelikowsky, J. R., Cohen, M. L. (1976), *Phys. Rev. B* **14,** 556–582.

Chiarotti, C., Del Signore, G., Frova, A., Samoggia, G. (1962), *Nuovo Cimento* **26,** 403–406.

Chiarotti, C., Del Signore, G., Nannarone, S. (1968), *Phys. Rev. Lett.* **21,** 1170–1172.

Cohen, M. L., Chelikowsky, J. R. (1988), *Electronic Structure and Optical Properties of Semiconductors,* Berlin: Springer-Verlag.

Costato, M., Reggiani, L. (1970), *Lett. Nuovo Cimento Ser. I,* **3,** 239–245.

Dresselhaus, G., Kip, A. F., Kittel, C. (1955), *Phys. Rev.* **98,** 368–384.

Fritzsche, H. (1955), *Phys. Rev.* **99,** 406–419.

Frova, A., Handler, P. (1965), *Phys. Rev. Lett.* **14,** 178–180.

Gerhardt, U. (1965), *Phys. Rev. Lett.* **15,** 401–403.

Giannozzi, P., De Gironcoli, S., Pavone, P., Baroni, S. (1991), *Phys. Rev. B* **43,** 7231–7242.

Greenaway, D. L., Harbeke, H. (1968), *Optical Properties and Band Structures of Semiconductors,* Oxford: Pergamon Press.

Groves, S., Paul, W. (1963), *Phys. Rev. Lett.* **11,** 194–196.

Harrison, W. A. (1989), *Electronic Structure and the Properties of Solids,* New York: Dover.

Heinrich, H., Lischka, K., Kriechbaum, M. (1970), *Phys. Rev. B* **2,** 2009–2016.

Hensel, J. C., Suzuki, K. (1974), *Phys. Rev. B* **9,** 4219–4257.

Hopfield, J. J. (1958), *Phys. Rev.* **112,** 1555–1567.

Hsieh, T. C., Miller, T., Chiang, T. L. (1984), *Phys. Rev. B* **30,** 7005–7008.

Hybertsen, M. S., Louie, S. G. (1985), *Phys. Rev. Lett.* **55,** 1418–1421.

Jagannath, C., Grabowski, Z. W., Ramdas, A. K. (1981), *Phys. Rev. B* **23,** 2082–2098.

Jusserand, B., Cardona, M. (1989), in: M. Cardona, G. Güntherodt (Eds.), *Light Scattering in Solids V,* Berlin: Springer-Verlag.

Kim, N., La Rocca, G. C., Rodriguez, S., Bassani F. (1989), *Rivista Nuovo Cimento* **12** (2), 1–67.

Kittel, C. (1966), *Introduction to Solid State Physics,* New York: Wiley.

Levinshtein, M. E., Simin, G. S. (1992), *Getting to Know Semiconductors,* Singapore: World Scientific.

Luttinger, J. M., Kohn, W. (1955), *Phys. Rev.* **97,** 869–883.

MacFarlane, G. G., McLean, T. P., Quarringron, J. E., Roberts, E. V. (1958), *Phys. Rev.* **111,** 1245–1254.

Madelung, O. (Ed.) (1982), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology,* Group 3, *Semiconductors,* Vol. 17a, *Physics of Group IV Elements and III-V Compounds,* Berlin: Springer.

Morin, E. J., Maita, J. P. (1954), *Phys. Rev.* **94,** 1525–1529.

Nara, H. (1965), *J. Phys. Soc. Jpn.* **20,** 778–784.

Nilsson, G., Nelin, G. (1971), *Phys. Rev. B* **3,** 364–369.

Nilsson, G., Nelin, G. (1972), *Phys. Rev. B* **6,** 3777–3786.

Pantelides, S. T. (1978), *Rev. Mod. Phys.* **50,** 797–858.

Pearson, G. L., Herring, C. (1954), *Physica* **20,** 975–978.

Penn, D. R. (1962), *Phys. Rev.* **128,** 2093–2097.

Philipp, H. R., Ehrenreich, H. (1963), *Phys. Rev.* **129,** 1550–1560.

Pickett, W. E., Wang, C. S. (1984), *Phys. Rev. B* **30,** 4719–4733.

Ramdas, A. K., Rodriguez, S. (1981), *Rep. Prog. Phys.* **44,** 1297–1387.

Sandercock, J. R. (1972), *Phys. Rev. Lett.* **28,** 237–240.

Seeger, K. (1991), *Semiconductor Physics,* New York: Springer-Verlag.

Seraphin, B. O., Hess, R. B. (1965), *Phys. Rev. Lett.* **14,** 138–140.

Solin, S. A., Ramdas, A. K. (1970), *Phys. Rev. B* **1,** 1687–1698.

Stella, A. (1993), in: A. Stella, L. Miglio (Eds.), *Semiconductor Superlattices and Interfaces,* Proceedings of the International School of Physics "Enrico Fermi", course CXVII, New York: North-Holland.

Straub, D., Ley, L., Himpsel, F. J. (1985), *Phys. Rev. Lett.* **54,** 142–145.

Strinati, G., Mattausch, H. J., Hanke, W. (1980), *Phys. Rev. Lett.* **45,** 290–294.

Sze, S. M. (1985), *Semiconductor Devices,* New York: Wiley.

Tosatti, E., Pastori Parravicini, G. (1971), *J. Phys. Chem. Solids* **32,** 623–626.

Tubino, R., Piseri, L., Zerbi, G. (1972), *J. Chem. Phys.* **56,** 1022–1039.

Uhrberg, R. I. G., Hansson, G. V., Karlsson, U. O., Nicholls, I. M., Persson, P. E. S, Flodström, S. A., Engelhardt, R., Koch, E. E. (1984), *Phys. Rev. Lett.* **52,** 2265–2268.

Weber, W. (1977), *Phys. Rev. B* **15,** 4789–4803.

Ziman, J. M. (1960), *Electrons and Phonons,* Oxford: Oxford University Press.

Further Reading

Besides the works cited, further references are listed in the selected bibliographies of Bassani and Pastori Parravicini (1975) and Cohen and Chelikowsky (1988). Finally, an excellent advanced textbook is Yu, P. Y., Cardona, M. (1995), *Fundamentals of Semiconductor Physics,* New York: Springer-Verlag.

SEMICONDUCTORS, ELEMENTAL—MATERIAL PROPERTIES

Howard R. Huff, *SEMATECH, Austin, Texas, U.S.A.*

	Introduction	437
1.	**Chemical Bonding and Semiconductors**	439
1.1	Atomic Structure	439
1.2	Group-IVa Crystalline Semiconductors	439
1.2.1	Crystal Structure	439
1.2.2	Electronic Structure	441
1.2.3	Thermodynamic Properties	443
2.	**The Defect Solid State**	444
2.1	Point Defects	444
2.1.1	Vacancies and Interstitials	444
2.2	Line Defects	447
2.2.1	Dislocations	447
2.2.2	Misfit Dislocations	450
2.3	Planar/Interface Defects	450
2.3.1	Stacking Faults	450
2.3.2	Twins	451
2.3.3	Surfaces	451
3.	**Crystalline Material Properties**	452
3.1	Electronic Properties	452
3.1.1	Electrons and Holes	452
3.1.2	Donors and Acceptors	455
3.1.3	Metal Impurities	458
3.2	Thermal Properties	459
3.2.1	Thermal Expansion	459
3.2.2	Thermal Conductivity	460
3.3	Mechanical Properties	460
3.4	Optical Properties	465
4.	**Silicon–Germanium**	466
5.	**β-Silicon Carbide**	468
6.	**Device/Integrated-Circuit Implications**	468
7.	**Conclusions**	470
	Acknowledgments	470
	Glossary	471
	Works Cited	471
	Further Reading	475

INTRODUCTION

Electronics is the world's largest industry with 1995 global yearly revenue of approximately \$1.15 × 10^{12} (U.S.). This is expected to reach approximately \$2.0 × 10^{12} (U.S) by the end of the century. Elemental semiconductor materials constitute a minor component of the electronics-based revenue stream, accounting for about \$5.1 × 10^9 (U.S.) in 1995 and projected to reach about \$9.4 × 10^9 (U.S.) by the year 2000. Although elemental-semiconductor-materials revenue constitutes only about 0.5% of the electronics industry revenue, elemental semiconductor materials are, nevertheless, the cornerstone upon which the industry resides.

The unique characteristic of crystalline semiconducting materials, compared to metals, is their reduction of resistivity with increasing temperature (negative temperature coefficient of resistivity) over selected regions of temperature as schematically illustrated in Fig. 1 (Teal *et al.,* 1985). The electrical resistivity in these regions decreases exponentially with absolute temperature. The decrease of resistivity with increasing absolute temperature in the low-temperature range is due to the thermal excitation of free carriers (i.e., electrons or holes) from the dopant impurities and is referred to as extrinsic conduction. That is, unlike metals where the addition of impurities causes the resistivity to increase, specific dopant impurities in semiconductors can result in drastic decreases in resistivity. The increase in resistivity with further increasing temperature in the low-temperature range is caused by the decrease of the free-carrier mobility. Finally, the decrease in resistivity at higher temperatures is due to the thermal excitation of electron–hole pairs (i.e., intrinsic conduction) dominating the resistivity. The temperature at which intrinsic conduction begins increases

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

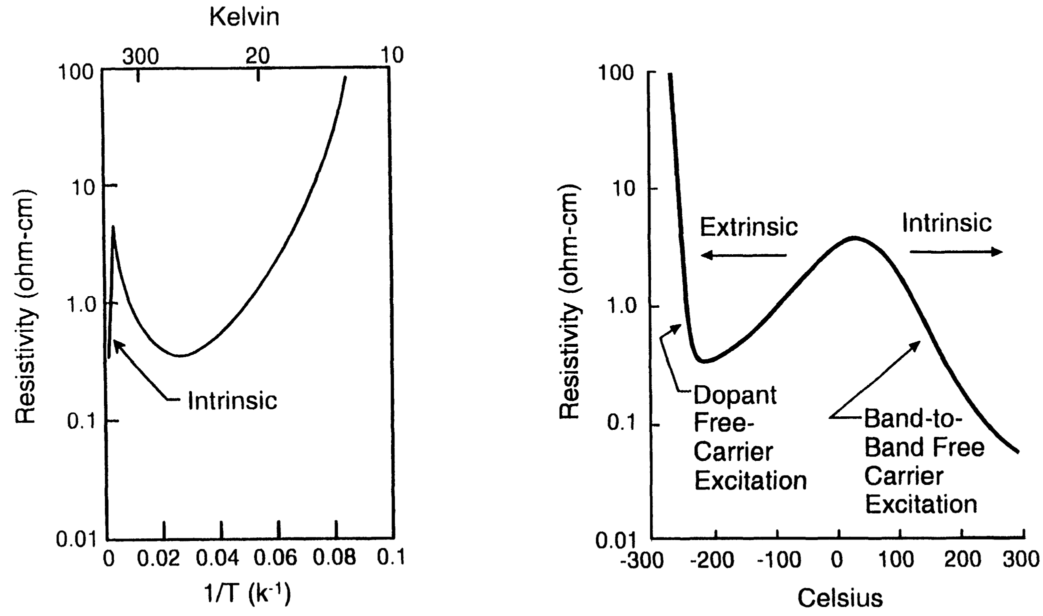

FIG. 1. Effect of temperature on resistivity. The portion of the curve between about −200 °C and 50 °C shifts up and down with the concentration of impurities in the semiconductor. These data are for Ge (following Teal *et al.*, 1985).

as the semiconductor energy gap increases. The mid-temperature, extrinsic range, where the impurities are fully ionized and before intrinsic conduction begins, is the region where most semiconductor devices operate. The resistivity in this region can readily be varied by the introduction of a properly selected dopant impurity concentration (Teal *et al.*, 1985). For example, the resistivity in silicon can be varied by seven orders of magnitude from 10^{-4} to 10^{3} Ω cm (Irvin, 1962; Thurber *et al.*, 1980a, 1980b). This article reviews, in addition to the electrical properties, a wide range of material properties for the group-IVa crystalline semiconductors diamond, silicon, germanium, and gray tin, as well as silicon–germanium alloys and β-SiC, important semiconducting material combinations derived from the group-IVa elements. The variation and correlation of their structural, thermodynamic, electrical, thermal, mechanical, and optical properties are emphasized, rather than an exhaustive analysis of the data *per se*. Although these properties are dopant-type, dopant-density, and temperature dependent, the characteristics presented are mainly for lightly doped materials at room temperature.

The atomic structure of the group-IVa elements carbon, silicon, germanium, and tin is introduced in Sec. 1. This is followed by an examination of their condensed-state crystalline semiconducting properties including crystal structure and chemical bonding, electronic structure, and thermodynamic properties, as well as kinetic considerations where appropriate. Appropriate data will also be included for silicon–germanium alloys and β-SiC where available. Comprehension of the defect solid state and the consequences of deviations from perfect periodicity, central to the understanding of semiconductor material properties, is discussed in Sec. 2. Semiconductor fabrication processes have a critical, complementary influence on material microstructure and, thereby, material properties. The process–structure–property methodology is critical, in conjunction with device and integrated circuit (IC) fabrication, to facilitate tailoring the semiconductor material properties to ensure optimal device/IC performance, reliability, and yield. Accordingly, the variation of the significant electrical, thermal, mechanical, and optical properties of the group-IVa crystalline semiconductors in Sec. 3, silicon–

germanium alloys in Sec. 4, and β-SiC in Sec. 5 should be considered in that context. Section 6 discusses the emergence of silicon, and silicon-based materials such as silicon–germanium, as the premier materials driving the IC microelectronics revolution. Section 6 also cautions against undue optimism for diamond, which is often cited as the ultimate semiconducting device material. Section 7 presents several conclusions. Finally, a further reading list complements the list of works cited.

1. CHEMICAL BONDING AND SEMICONDUCTORS

1.1 Atomic Structure

Carbon, silicon, germanium, and tin are elements in group IVa of the periodic table (Lide, 1993–1994). Their atomic numbers, gram atomic weights, electronic configurations, atomic radii, and ionization energies are listed in Table 1. Each element has four valence electrons, two in *s* states and two in *p* states, for the ground-state configuration, which is designated by the spectroscopic notation $^3P_0$ (Landau and Lifshitz, 1958). The monotonic decrease of the ionization energy with increasing atomic number of the elements will be seen to exhibit analogs with the variation in a number of crystalline semiconducting material properties.

1.2 Group-IVa Crystalline Semiconductors

1.2.1 Crystal Structure The group-IVa semiconductors crystallize in the tetrahedrally coordinated diamond crystal structure. The diamond crystal structure is conveniently described as two interpenetrating face-centered cubic (fcc) crystals displaced along the ⟨111⟩ direction by one-fourth the diagonal length (see Fig. 2) (Harrison, 1980; Bullis and Huff, 1994). The unit cell of this structure contains eight sites at the points (0,0,0), $(0,\frac{1}{2},\frac{1}{2})$, $(\frac{1}{2},0,\frac{1}{2})$, $(\frac{1}{2},\frac{1}{2}0)$, $(\frac{1}{4},\frac{1}{4},\frac{1}{4})$, $(\frac{1}{4},\frac{3}{4},\frac{3}{4})$, $(\frac{3}{4},\frac{1}{4},\frac{3}{4})$, and $(\frac{3}{4},\frac{3}{4},\frac{1}{4})$. The length of the unit cell in the ⟨100⟩ direction, referred to as the lattice constant a_0, is the distance required to translate the unit cell to the position of the adjacent unit cell along the three ⟨100⟩ directions. This replication of the unit cell results in the crystal structure. A useful characteristic is the covalent radius, r_c, defined as one-half the distance between nearest neighbors in a crystal of the pure element, in terms of the cube edge a_0. Alternatively, the closest separation of atomic sites along the ⟨111⟩ direction or bond length d (twice the value of r_c) is conveniently utilized. The bond length is a critical parameter affecting a number of physical properties involving the elastic stiffness coefficients such as the bulk modulus (see Sec. 3.3). The bond energy is a useful

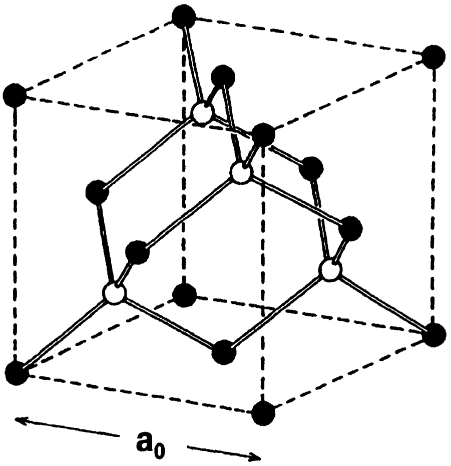

FIG. 2. Silicon crystal structure; the open circles indicate substitutional sites interior to the unit cell (following Kittel, 1996).

Table 1. Characteristics of several group-IVa elements.

Element	Atomic number	Gram atomic weight (g)	Electronic configuration	Atomic radius (nm)	Ionization energy (eV)
Carbon	6	12.01	$1s^22s^22p^2$	0.091	11.3
Silicon	14	28.09	$1s^22s^22p^63s^23p^2$	0.132	8.1
Germanium	32	72.59	$1s^22s^22p^63s^23p^63d^{10}4s^24p^2$	0.137	7.9
Tin	50	118.69	$1s^22s^22p^63s^23p^63d^{10}4s^24p^64d^{10}5s^25p^2$	0.162	7.3

Table 2. Variation of several structural/bonding characteristics for the group-IVa and β-silicon carbide semiconductors.

Element	Metallicity parameter (α_m)	Lattice constant a_0(nm)	Bond length d(nm)	Cohesive energy (eV/bond)
Diamond[a]	0.34	0.356683	0.154	3.68
β-Silicon carbide	0.45	0.43596	0.189	3.17
Silicon	0.66	0.543075	0.235	2.32
Germanium	0.81	0.565754	0.245	1.90
Gray tin[b]	0.87	0.64912	0.281	1.56

[a] Carbon crystallizes as diamond under conditions discussed in Sec. 1.2.3. p-type semiconducting diamond is referred to as type IIb (see DIAMOND AND DIAMONDLIKE CARBON).

[b] Also referred to as α-tin.

parameter denoting the strength of the chemical bond (Pauling, 1960). The cohesive energy (enthalpy of sublimation) is a measure of the energy per bond required to separate the solid into isolated atoms. It has been a useful parameter historically and is included for completeness, although it is not currently a fundamental parameter describing variations in material properties (Harrison, 1980). The values of a_0, d, and the cohesive energy for the group-IVa semiconductors are summarized in Table 2. The increase in a_0 and d and the concurrent decrease in cohesive energy from diamond to gray tin in Table 2 can be explained in terms of the decrease in the strength of the covalent bond from diamond through gray tin. β-SiC's parameters are between those of silicon and carbon, as expected. These trends may also be correlated with the decrease in the energy gap E_g and the increase in the metallicity parameter α_m, to be discussed in Sec. 1.2.2.

Each atomic site in the diamond crystal is surrounded by four nearest-neighbor atomic sites (coordination number 4), which reside at the points of a tetrahedron with bond angles $\theta = \arccos(-\frac{1}{3}) = 109.47°$. The relative positions of the four nearest neighbors have different directions for atoms in the two fcc structures. Therefore, the two fcc structures are not equivalent and the diamond crystal structure cannot be reduced to a simpler crystal structure. The covalent bond formed between nearest neighbors is described by sp^3 hybridization (see Sec. 1.2.2.) Each atom of the pair contributes one electron, which bond with opposite spins as described by the Pauli principle. As a result, each atom on an atomic site in a perfect crystal is surrounded by eight electrons: four of "its own" and one from each of its four nearest neighbors. This octet of electrons, similar to that of a noble gas, is chemically very stable (Pauling, 1960). No electrons are free to move through the crystal unless sufficient energy is applied to break the bond.

The sp^3 hybridized bond is highly directional and accounts for the diamond crystal structure exhibiting the smallest packing density (i.e., volume fraction of the atoms in the unit cell) among the cubic crystal classes, which include the simple cubic, body-centered cubic (bcc), and fcc structures (see Table 3). The variation of the packing density correlates with the geometric coeffi-

Table 3. Variation of structural parameters for the cubic crystal structures.

Crystal structure	Coordination number	Covalent radius (r_c)	Atoms per unit cube	Packing density	Geometric coefficient (α)
fcc	12	$(a_0/4)\sqrt{2}$	4	$(\pi/12)\sqrt{8} = 0.7405$	1.79
bcc	8	$(a_0/4)\sqrt{3}$	2	$(\pi/8)\sqrt{3} = 0.6802$	1.79
Simple cubic	6	$a_0/2$	1	$\pi/6 = 0.5236$	1.76
Diamond	4	$(a_0/8)\sqrt{3}$	8	$(\pi/16)\sqrt{3} = 0.3401$	1.67

cient α characterizing the electrostatic environment of the cubic crystal classes (Harrison, 1980). Alpha depends only on the geometry of the atomic arrangements; charge magnitudes and distances have been factored out. The diamond crystal structure exhibits the smallest geometric constant. Silicon and germanium can also exhibit a phase transformation at room temperature to a body-centered tetragonal crystal structure without macroscopic plastic deformation (Roos *et al.*, 1995a).

An equivalent but more fundamental description of the diamond crystal structure utilizes the concept of an fcc lattice with basis vectors (0,0,0) and $(\frac{1}{4},\frac{1}{4},\frac{1}{4})$. The symmetry operations required to describe the arrangement of atoms associated with the lattice (i.e., the space group) is $Fd3m$ using the International notation (Burns, 1985). The F refers to the fcc lattice and $d3m$ refers to the point symmetry, glide, and screw operations. The diamond crystal structure is an example of a nonsymmorphic space group, which requires a translation by a fraction of a unit cell length in order to specify the space group. The point group is $m3m$, since only lattice translations are involved (the lattice symbol F is ignored) and all translations by a fraction of a unit cell length are set equal to zero. β-SiC and the III–V compounds such as gallium arsenide (zinc-blende structure) are characterized by $F\bar{4}3m$ and $\bar{4}3m$ for the corresponding space and point groups, respectively. In particular, the zinc-blende structure does not have a screw axis or a center of symmetry inversion between the (0,0,0) and $(\frac{1}{4},\frac{1}{4},\frac{1}{4})$ positions because of the dissimilarity of the atoms. The origin of their polarity effects in the ⟨111⟩ directions is due to this lack of a center of symmetry inversion. Silicon–germanium alloys, however, are not ordered wherein every silicon and germanium atom, for example, is precisely locked onto a (0,0,0) and $(\frac{1}{4},\frac{1}{4},\frac{1}{4})$ lattice site, respectively (or vice versa). Accordingly, one does not expect to see a polarity effect in silicon–germanium alloys, although, in principle, $Si_{0.5}Ge_{0.5}$ should be ordered.

1.2.2 Electronic Structure Consider a representative group-IVa element arranged in the diamond crystal structure at 0 K. At large internuclear distances, the two s and two p valence electrons are still associated with each atom. As the atoms approach each other, using the conventional one-electron model, the discrete atomic energy s and p levels are each broadened into a band as a result of coupling between the neighboring atomic states, as described by the tight-binding theory (Harrison, 1980) (see Fig. 3). These bands may be regarded as a quasicontinuum of energy states propagating throughout the material. The s bands are completely filled with two electrons donated per atom. The p bands, however, can accommodate six electrons per atom and, therefore, are only partially filled with two electrons donated per atom. This partial filling of bands is characteristic of a metal. As the atoms are brought still closer together, the broadening bands reach each other. At this point, the s and p valence electron states reshuffle to form bonding and antibonding bands. This process has also been described as the formation of the prototypical sp^3 hybridized, tetrahedrally bonded purely covalent bonds (Harrison, 1980). This reshuffling, or sp^3 hybridization process, results in an energy gap, E_g, with four energy states below and four energy states above the energy gap, with no propagating energy states within the energy gap. The original four valence electrons of each atom completely fill the lower four states, which are collectively referred to as the bonding states or valence band. The four states above the energy gap are completely empty and are collectively referred to as the antibonding states or conduction band. Each of the four bonding states is doubly degenerate on account of the presence of one bonding state contributed by each of the four neighboring atoms, as described in Sec. 1.2.1. The valence band, therefore, is completely filled with a stable octet of electrons per atom. Although the original valence electrons belong to the crystal system, there is, nevertheless, a high concentration of electronic charge pileup in the regions between the atoms. This is interpreted as the covalent bond between adjacent atoms formed between two oppositely spinning electrons. The system has no free electronic carriers at 0 K and is an insulator. This description of the band model is consistent with the bond model for crystalline materials described in Sec. 1.2.1. Interesting theoretical challenges are posed, however, in reconciling the band and bond approaches in

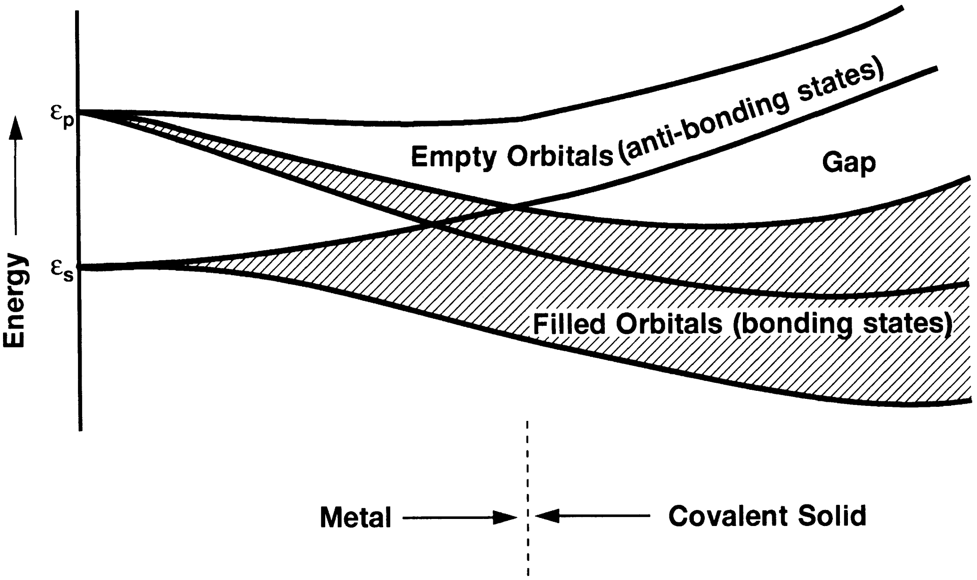
Energy
ε_p
ε_s
Empty Orbitals (anti-bonding states)
Gap
Filled Orbitals (bonding states)
Metal
Covalent Solid

FIG. 3. Formation of energy bands in a homopolar tetrahedral semiconductor as the atoms are brought together. The internuclear distance decreases to the right. (Following Harrison, 1980.)

the absence of long-range order (Ziman, 1979).

The energy gap is the defining characteristic of an insulator and is typically greater than 5–6 eV. A pure semiconductor, however, has a sufficiently small value of E_g relative to kT to facilitate the thermal excitation of electrons from valence-band states to conduction-band states, resulting in free electrons and holes at finite temperatures (see Sec. 3.1). The value of E_g slightly decreases with increasing temperature (Lide, 1993–1994) for the group-IVa semiconductors since the increased temperature increases the amplitude of atomic vibrations, thereby increasing the effective internuclear distance. The value of E_g decreases (increases) slightly with increasing pressure for silicon (germanium) (Lide, 1993–1994). Germanium exhibits greater changes, compared to silicon, in E_g with changes in temperature and pressure because of its weaker covalent bonding, as noted in Sec. 1.2.1. The magnitude of E_g varies with different directions in the crystal and is generally expressed with respect to specific directions in reciprocal (**k**) space (Herman, 1955; Sze, 1981).

The fundamental parameter distinguishing the elemental semiconductors is the metallicity parameter α_m. This parameter is defined as the ratio of the metallic energy V_1 to the covalent energy V_2 within a factor of 1.1 (Harrison, 1980):

$$\alpha_m = 1.1(V_1/V_2). \tag{1}$$

The metallic energy relates to the band broadening and is determined principally by the separation in energy between the atomic p and s states $[\frac{1}{4}(\epsilon_p - \epsilon_s)]$. The sp splitting (and therefore V_1) does not change greatly as one proceeds from diamond to gray tin (i.e., with increasing atomic number) as seen in Table 4. The separation in energy between the valence and conduction bands arises from the coupling strength between states on neighboring atoms, which resulted in the bonding–antibonding splitting (Harrison, 1980). Inasmuch as the bond length significantly increases as one proceeds from diamond to gray tin, the covalent bond becomes weaker and, likewise, the covalent energy V_2 decreases relative to the band broadening. Accordingly, the energy gap E_g decreases and α_m increases as one proceeds from diamond to gray tin (see Table 4) (Harrison, 1980). β-SiC's energy gap is between diamond and silicon as expected. The energy gap for silicon–germanium alloys can be tuned to intermediate values between silicon and germa-

Table 4. Metallic energy, covalent energy, metallicity, and energy gap for the group-IVa and β-silicon carbide semiconductors.

Element	Metallic energy V_1 (eV)	Covalent energy V_2 (eV)	Metallicity α_m	Energy gap at 0 K (eV)	Energy gap at 300 K (eV)
Diamond	2.13	6.94	0.34	5.48	5.45
β-Silicon carbide	1.91	4.66	0.45	2.6	2.3
Silicon	1.76	2.98	0.66	1.17	1.12
Germanium	2.01	2.76	0.81	0.74	0.66
Gray tin	1.64	2.10	0.87	0.075	0.0

nium by varying the germanium content (see Sec. 4). When α_m approaches unity, semiconducting behavior disappears as the valence and conduction bands approach each other, and metallic behavior (i.e., partially filled bands) is approached. This indeed occurs for gray tin (Harrison, 1981), at least at the wave vector $\mathbf{k} = 0$, where E_g approaches zero. The metallicity parameter is the parameter *sine qua non* by which the distinct variations in the group-IVa crystalline semiconductor material characteristics may be correlated.

1.2.3 Thermodynamic Properties The first and second laws of thermodynamics are usefully expressed in terms of the Gibbs free energy ΔG of a chemical reaction at constant temperature and pressure (Swalin, 1972; Kubaschewski and Alcock, 1973):

$$\Delta G = \Delta H - T\Delta S, \qquad (2)$$

where ΔH is the change in enthalpy and ΔS is the change in entropy associated with the formation of the product at absolute temperature T. The enthalpy of formation is a constant for every substance at a given temperature and pressure relative to the reference or standard state. The reference state usually selected for crystalline solids is the most stable form of the element at 298 K and one atmosphere pressure. Therefore, the enthalpy of formation for carbon, silicon, germanium, and white tin is zero. Several thermodynamic data for the group-IVa elements and β-SiC are tabulated in Table 5 (Kubaschewski and Alcock, 1973). Negative values for ΔH^0_{298} refer to an exothermic process whereas a positive value for ΔH^0_{298} refers to an endothermic process. The positive value of ΔH^0_{298} for diamond indicates it is not stable at room temperature in comparison to the graphitic form. The negative value of ΔH^0_{298} for gray tin, compared with white tin, might indicate that the former is the stable form of tin at 298 K. However, white tin is the stable form at 298 K (Swalin, 1972); the temperature at which gray tin transforms to white tin at one atmosphere pressure is 286 K (Swalin, 1972). Clearly, the enthalpy change is not a sufficient criterion to deduce the direction of a chemical reaction.

The complete description of a chemical reaction requires utilization of the entropy term in Eq. (2). Entropy may be interpreted

Table 5. Selected thermodynamic properties for the group-IVa and β-silicon carbide semiconductors at 298 K.

Element	ΔH^0_{298} (kJ/mole)	S^0_{298} (J/mole K)	ΔG^0_{298} (kJ/mole)	c_{p298} (J/mole K)
Carbon	0	5.69	−2.87	10.62
Diamond	1.90	2.44	—	7.56
β-Silicon carbide	−66.88	16.51	−71.80	29.54
Silicon	0	18.81	—	20.73
Germanium	0	42.22	—	26.07
White tin	0	51.41	−0.03	26.32
Gray tin	−1.96	44.73	—	25.28[a]

[a] From Lide (1993–1994).

as the decrease in atomic order of a system undergoing any process. It is quantitatively calculated by the enthalpy exchanged during an isothermal and isobaric reversible process divided by the absolute temperature at which the process takes place. The entropy of formation at a given temperature may be determined for every substance relative to a reference state. Since the third law of thermodynamics indicates that the entropy of any ordered crystalline solid approaches the same value at 0 K (Swalin, 1972; Kubaschewski and Alcock, 1973), the common practice is to select 0 K as the reference state and arbitrarily choose the value zero for the entropy at 0 K (and one atmosphere). Entropy, therefore, can be determined "absolutely," unlike enthalpy, which was seen to be determined relative to the value of enthalpy at 298 K. β-SiC's entropy is significantly closer to that of silicon, rather than simply between diamond and silicon's values (see Table 5). The entropy of fusion decreases as one proceeds from silicon to gray tin (with increasing metallicity) (Kubaschewski and Alcock, 1973). This is reasonable since the strength of the covalent bond and the Debye temperature decrease with increasing metallicity and the group-IVa semiconductors are metallic in the liquid state.

The Gibbs free energy of reaction favors white tin over gray tin at temperatures greater than 286 K (Swalin, 1972) and graphitic carbon over diamond at room temperature (Kubaschewski and Alcock, 1973) (see Table 5). The observation of diamond at room temperature, therefore, is indicative of its nonequilibrium thermodynamic status. Bulk crystalline diamond is thermodynamically stable with respect to graphite at high temperature (>1000 °C) and high pressure (6×10^{11} dyn/cm^2); the reaction rate is enhanced by the utilization of a catalyst such as nickel (see DIAMOND AND DIAMONDLIKE CARBON). Crystalline diamond has also been grown via nonequilibrium chemical vapor deposition (CVD) processes, most of which have in common the production of atomic hydrogen and hydrocarbon radicals in regimes where solid carbon is expected to be the stable product. Two mechanisms have been proposed whereby crystalline diamond, rather than graphite or vitreous carbon, is observed. The first suggests that graphite is etched by atomic hydrogen at a faster rate than diamond, and hence diamond is kinetically stable with respect to graphite. The second mechanism suggests that diamond surfaces are stabilized by termination with atomic hydrogen. This topic continues to be an important area of research, and the reader is referred to the literature for further details (Davis *et al.*, 1988; Yarbrough, 1990) (see CARBON MATERIALS; DIAMOND AND DIAMONDLIKE CARBON; GRAPHITE).

The molar specific heat c_{p298} is the energy required to raise the temperature of one molecular weight of the material by one degree at constant pressure. The classical Dulong–Petit law indicates that c_{p298} for one molecular weight of a monatomic solid is about 26.33 J/mole K (6.3 cal/mole K) at room temperature (De Launay, 1956). The specific heat at 300 K for silicon, and especially diamond, however, is significantly less than 26.33 J/mole K because the Debye temperature θ_D for these elements is significantly larger than 300 K (see Sec. 3.2 and Tables 5 and 14). That is, the Dulong–Petit law is valid only when the atoms are sufficiently loosely bound that the equipartition of energy is valid (T/θ_D approximately greater than 80%) (De Launay, 1956). β-SiC's specific heat of 29.54 J/mole K, however, is only about 6% less than the sum of c_{p298} for carbon and silicon. This additivity of the specific heat of the elements to obtain the specific heat of the compound in a chemical reaction is known as Neumann and Kopp's rule (Kubaschewski and Alcock, 1973).

2. THE DEFECT SOLID STATE

2.1 Point Defects

2.1.1 Vacancies and Interstitials At 0 K, silicon is a perfect insulator inasmuch as the isotopes Si^{28}, Si^{29}, and Si^{30} exhibit only zero-point (weakly coupled) atomic vibrations. At any nonzero temperature, however, thermodynamic requirements introduce a degree of disorder (entropy) into the crystal structure so as to minimize the Gibbs free energy [Eq. (2)] at thermal equilibrium. In this case, ΔH is the enthalpy, and ΔS is mainly the configurational (i.e., mixing) entropy associated with the formation of each type of point defect at absolute temperature T (Swalin, 1972; Seeger, 1976).

Atoms transfer off the substitutional sites onto interstitial sites creating Frenkel (vacancy–interstitial) point defects (see POINT AND EXTENDED DEFECTS IN SOLIDS). The interstitial is also referred to as a self-interstitial. Five equivalent tetrahedral interatomic voids, in which interstitial atoms are easily accommodated in silicon, are located along the body diagonals at the points $(\frac{1}{2},\frac{1}{2},\frac{1}{2})$, $(\frac{1}{4},\frac{1}{4},\frac{3}{4})$, $(\frac{1}{4},\frac{3}{4},\frac{1}{4})$, $(\frac{3}{4},\frac{1}{4},\frac{1}{4})$, and $(\frac{3}{4},\frac{3}{4},\frac{3}{4})$. A description of additional interstitial configurations has recently been summarized (Hu, 1994). Vacancies are also injected from the surface into the crystal, creating Schottky point defects. The Frenkel and Schottky defect systems are described by the Boltzmann statistics; the temperature dependence of the Gibbs free energy for these reactions, however, is still an active area of investigation (Brown *et al.*, 1994). The point-defect concentrations and the size of their clusters may be experimentally determined, however, by such techniques as diffuse x-ray scattering (Schulz, 1982), positron annihilation (Asoka-Kumar *et al.*, 1994), or the Bond x-ray method (Bond, 1960). The Bond method can precisely measure the change in lattice constant and deduce the point-defect concentrations. The concentrations are extremely low, probably $<10^5$ cm^{-3} at temperatures approximately 77 K for silicon. The point defects are in the range of 10^{13}–10^{16} cm^{-3} from approximately 800 to 1400 °C for silicon; their presence and concentration are deduced by comparing experimental dopant diffusion profiles in devices and ICs with model calculations. The temperature-dependent concentration and diffusion coefficient for vacancies and interstitials in silicon remains a significant theoretical challenge (Taylor *et al.*, 1993; Brown *et al.*, 1994).

The creation of a vacancy in an otherwise rigid crystal should increase the crystal volume by one atomic volume (Seeger, 1976). Likewise, the creation of one interstitial should decrease the crystal volume by one atomic volume. In reality, the crystal elastically deforms in response to the introduction of point defects, which results in additional changes of the crystal volume. The elastic deformation, however, affects the lattice constant and the macrosopic dimensions equally. Thermal expansion also affects the unit cell and the crystal volume equally. By subtracting the relative change of volume of the unit cell from the relative change of the crystal volume, one obtains the difference between the vacancy and the interstitial concentrations. For a cubic crystal, one may write (Seeger, 1976)

$$C_V - C_I = 3[(\Delta l/l_0) - (\Delta a/a_0)] + \dots, \qquad (3)$$

where Δl refers to the length change relative to the sample length, l_0, in a low-temperature state essentially free of vacancies and interstitials and Δa and a_0 are the corresponding quantities for the lattice constant (Seeger, 1976).

A measure of the type of point defect present in metals such as molybdenum and tungsten and the group-IVa semiconductors is illustrated in Fig. 4 (Slack and Bartram, 1975). The linear thermal expansion coefficient α is represented by the continuous line and is related to $\Delta l/l_0$. The dashed line represents the linear extrapolation of α from approximately 70% of the absolute melting temperature and is related to $\Delta a/a_0$. Accordingly, the dominance of vacancies or interstitials may be determined from the sign of the deviation of the experimental quantities in Eq. (3). Vacancies are the dominant intrinsic point defect in metals such as molybdenum and tungsten. The evidence is not so clear for germanium and diamond, although self-diffusion data in germanium is compatible with the presence of vacancies (Seeger and Swanson, 1968; Holwigle *et al.*, 1985). There has been no evidence of interstitials in germanium (Gösele and Tan, 1991). The thermal expansion coefficients of silicon and SiC are significantly different from those of metals, germanium, and diamond. The rather constant value of the thermal expansion coefficient of silicon at temperatures greater than approximately 725 °C has been interpreted as evidence that interstitials are as significant as vacancies. Collaborative evidence supporting the importance of the interstitial in silicon at high temperature comes from self-diffusion data (Seeger and Swanson, 1968; Holwigle *et al.*, 1985; Gösele and Tan, 1991) and transmission electron microscopy (TEM) analysis, indicating that A-type swirl defects are interstitial-type microscopic dislocation loops (Foll *et al.*, 1977; Chikawa and Shirai, 1979) (see Sec. 2.2.1). Swirl is a generic term referring to a class of defects in silicon distributed in a swirl pat-

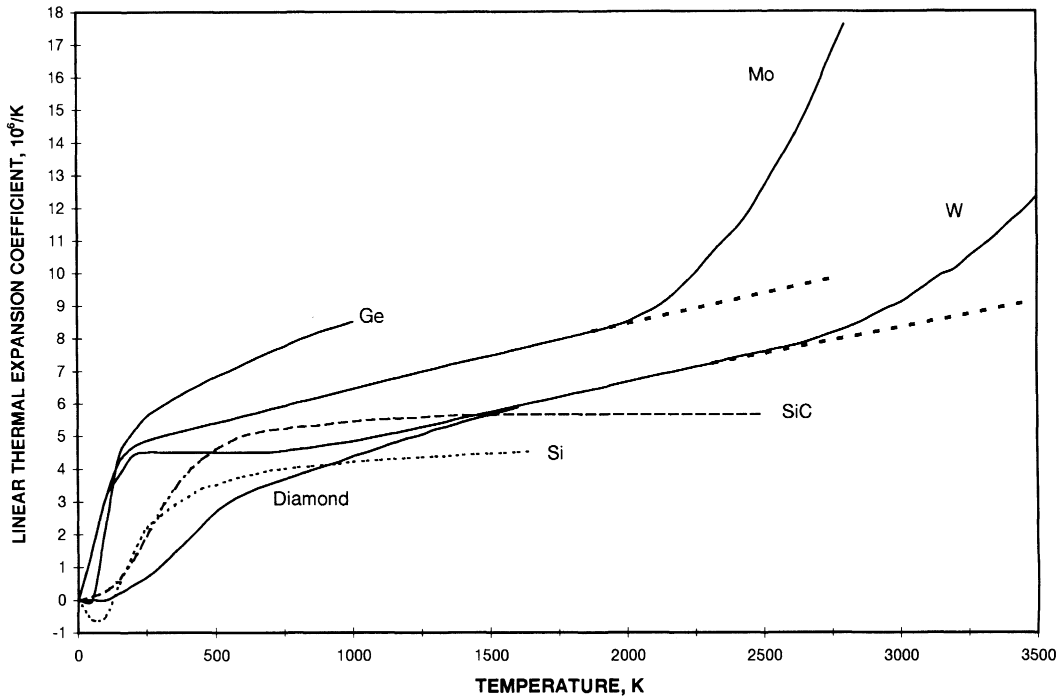

FIG. 4. Thermal expansion coefficients of several diamondlike solids (following Slack and Bartram, 1975).

tern reflecting melt convection, rotation of the crystal, and temperature fluctuations at the crystal–melt interface during crystal growth. The defects develop as a result of the condensation of point-defect complexes formed via solid-state reactions in the absence of readily available line and/or surface sinks. It should be emphasized, however, that the observed defects reflect conditions subsequent to the complete crystal growth process, resulting from the complicated set of microscopic defect reactions in silicon (Huff, 1992; Brown *et al.*, 1994), and require consideration of the initially solidifying silicon at the melt interface during crystal growth (Dornberger and von Ammon, 1996).

Defects such as vacancies and interstitials, therefore, are thermodynamically required at any nonzero temperature in order to minimize the Gibbs free energy ΔG in Eq. (2). On the other hand, one- and two-dimensional defects, such as dislocations (line defects) and stacking faults or surfaces (planar/interface defects), respectively, have high Gibbs free energies of formation because of the large number of atoms involved in the defects and a reduced configurational entropy due to the large number of defect sites. Consequently, line and planar defects are introduced only in environments that are far away from thermal equilibrium. Nonequilibrium defects can be minimized, in principle, by bulk reorganization and surface reconstruction processes during thermal annealing. Since diamond is not thermodynamically stable at room temperature and pressure, however, but is controlled by kinetic considerations, achieving relatively defect-free diamond is challenging.

Any deviation from perfect periodicity, furthermore, may be regarded as a defect. This includes free electronic carriers, particularly in semiconductors and insulators, phonons, impurities, isotopic variations, etc. Controlling point-defect impurities such as oxygen, carbon, hydrogen, and nitrogen are, in particular, of immense importance for silicon IC process technologies (Bullis and Huff, 1994). The wave functions describing defects are nonpropagating solutions to Schrödinger's equation; that is, they are spatially localized in the crystal and often exhibit localized (and deep) energy states within the energy gap. Defects that introduce localized energy levels within the energy gap may or may not influence the fabrication or performance of ICs. For example, fabrication of IC structures at reduced process temperatures (≤800 °C) may accentuate the effects of different charge states of localized defects through the promotion of unique solid-state defect reactions compared to high-temperature processing and/or significantly different rates for the same reactions. The resulting changes in the silicon microstructure will affect the subsequent device and IC performance (Huff, 1992).

2.2 Line Defects

2.2.1 Dislocations Dislocations are (nonequilibrium) line imperfections in the crystal that extend over many atomic spacings (see POINT AND EXTENDED DEFECTS IN SOLIDS). They can move (slip) under stress fields in the {111} planes (glide plane) and ⟨110⟩ directions (slip directions) in the diamond crystal, once they have been activated from the bound state. Their movement provides the mechanism for plastic deformation; that is, the permanent displacement of atomic planes past each other as the result of an external stress (Cottrell, 1961; Hirth and Lothe, 1982). The {111} planes in the silicon crystal are unequally spaced in the ratio 1:3 (see Fig. 2). Therefore, two sets of {111} glide planes, the *shuffle* and *glide*, exist. The *shuffle* set is observed if glide occurs between the widely spaced planes. The *glide* set is observed if glide occurs between the narrowly spaced planes.

When an extra plane of atoms occurs on one side of the slip plane, the line of atoms at the end of this extra half plane is called an edge dislocation (see Fig. 5a). The motion of edge dislocations in the glide plane occurs at kink sites (Hirth and Lothe, 1982) and is along the direction of the shearing force. The effect of having two atoms per lattice site in the group-IVa semiconductor is to double the number of atomic half planes associated with an edge dislocation (Hobstetter, 1960). Under some conditions the crystal may twist. In this case, there are no extra half planes of atoms. This type of dislocation, known as a screw dislocation, occurs along the line where the planes above and below the slip plane return to their original relationship. Screw dislocations move through the crystal in the direction perpendicular to the shearing force (Cottrell, 1961; Hirth and Lothe, 1982). The stress required to move dislocations decreases and their velocity increases exponentially with increasing temperature (Alexander and Haasen, 1968).

Dislocations are characterized by a net displacement, the Burgers vector **b**, and a unit vector **I** that is parallel to the dislocation line. Dislocations in real crystals may be pure edge or pure screw, where the dot

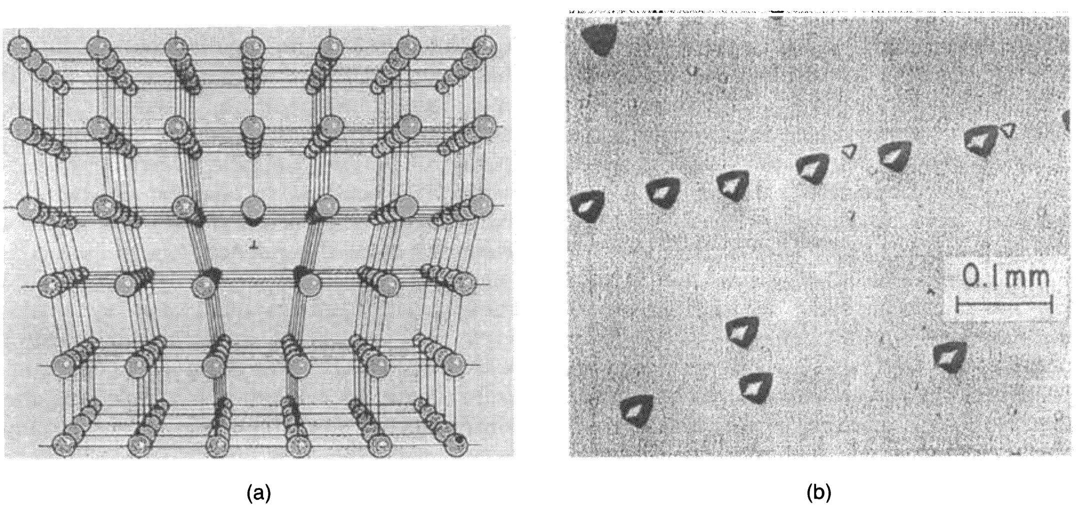

(a) (b)

FIG. 5. (a) Atomistic representation of an edge dislocation (after Gay, 1960). (b) Photomicrograph of a preferentially etched (111) silicon wafer illustrating etch pits.

product of **b** and **l** equals zero and **b**, respectively, or more commonly, a combination of both configurations. In silicon and germanium, a more common configuration than the double half-plane edge dislocation occurs. This particular structure requires only one extra half plane and is referred to as the 60° dislocation (Hobstetter, 1960; Hirth and Lothe, 1982). In this case, the angle between the dislocation line and the Burgers vector is 60°. Electron and x-ray diffraction methods, which resolve the orientation of **b** relative to **l**, are routinely employed for structural determination. The motion of the edge and screw dislocations described above are confined to a glide plane that contains both **b** and **l**. Glide motion does not increase the dislocation line length. Motion normal to the glide plane is referred to as climb and is nonconservative. That is, climb requires the addition or subtraction of point defects and results in an increase of the dislocation line length.

The atomic arrangement and peculiar features of dislocations in silicon have been described (Hornstra, 1958; Hirth and Lothe, 1982). A dislocation cannot end within a single crystal except at another dislocation; thus, dislocations form closed loops or interconnecting networks. Unbound dislocations can multiply as a result of thermal stress fields during IC thermal processing, penetrate active regions of devices and ICs, and lead to electrical malfunctions (Lawrence and Huff, 1982; Huff and Shimura, 1985; Huff and Goodall, 1995). It was originally believed that dislocations could not form at low temperatures in brittle solids such as silicon (Allen, 1957). However, it has been found that (somewhat bound) dislocations could be produced in a pure cracking event (i.e., "crack dislocation") during room-temperature semiconductor material shaping processes (Allen, 1957; Wise *et al.*, 1994) and by nanoindentation at room temperature (Page *et al.*, 1992). These dislocations are not necessarily in a glide configuration (i.e., lying entirely in {111} planes), but at elevated temperatures such as during IC processing and with sufficient thermal stress, these crack dislocations can act as sources for extensive slip. That is, although silicon wafers utilized for the fabrication of ICs are macroscopically dislocation free, latent sources often exist at the wafer edges and back surface, reducing the critical resolved shear stress to activate dislocation movement. On the other hand, bound dislocations are incapable of glide, or they form loops too small to be activated into macroscopic slip (Wise *et al.*, 1994). Historical reviews of the effects of dislocations on semiconductor devices and IC performance and the efforts to obtain and maintain dislocation-free material during IC fabrication include extensive bibliographies (Hu, 1986; Huff and Goodall, 1995).

The correspondence between the intersection of an edge dislocation line with the surface and the etch pit formed by using a preferential etchant is extensively utilized to assess the crystallographic perfection of semiconductors (see Fig. 5b) (Dash, 1960; Gatos and Lavine, 1965). Substantiation of this correlation was obtained by the use of copper decoration and direct imaging of the dislocations (Dash, 1956). The presence of screw dislocations has also been deduced by etchants, although this correlation posed a slightly more complicated situation as regards their clear-cut etching characteristics.

The activation enthalpy required to move an edge dislocation to an adjacent configuration, neglecting dislocation charge effects and reconstruction at the dislocation core, is called the Peierls energy (Chen *et al.*, 1992). The nonmonotonic decrease of the Peierls energy with increasing metallicity, however, highlights the simplicity of neglecting dislocation core reconstruction effects and related structure-sensitive factors (see Table 6). The elementary process (i.e., kink energy U_K), by which an atom in an edge dislocation moves, has been correlated with the energy gap at the Tamm temperature (Siethoff, 1994), taken as 0.7 times the absolute melting temperature T_M. Although the resistance to motion of dislocations in covalent crystals has been suggested to decrease with excitation of electrons from the valence band to the conduction band, the effects of dislocation core reconstruction effects must also be considered. The activation enthalpy for vacancy self-diffusion, Q^{SD}, was observed to be a constant factor of 1.5 times U_K, reflecting the similarity of breaking and reconstructing bonds with the kink mechanism (Siethoff, 1994). The hardness—breaking and rearrangement of electronic bonds—of many materials scales with their atomic density; diamond, with the highest atomic density, is the

Table 6. Metallicity parameter, Peierls energy, kink energy, energy gap at 0.7 T_M, hardness, atomic density, and mass density for the group-IVa and β-silicon carbide semiconductors.

Element	Metallicity parameter α_m	Peierls energy (eV)	Kink energy U_K (eV)	Energy gap at 0.7 T_M (eV)	Hardness (kg/mm^2)	Atomic density (10^{23} atoms/cm^3)	Mass density (g/cm^3)
Diamond	0.34	11.35	10.7	5.35	8375	1.76	3.51525
β-SiC	0.45	—	—	—	3100–3475	0.97	3.21
Silicon	0.66	5.95	2.35	0.93	1125	0.50	2.3283
Germanium	0.81	6.67	1.89	0.44	750	0.44	5.3234
Gray tin	0.87	5.66	—	—	—	0.29	5.765

hardest material (see Sec. 3.3 for a related discussion of the bulk modulus, i.e., the resistance of atomic electron shells being pushed closer together). The general monotonic decreases in the theoretical Peierls energy, kink energy U_K, energy gap at $0.7T_M$, hardness, and atomic density with increasing metallicity are summarized in Table 6.

A fairly low concentration of point defects such as oxygen, (2–5) × 10^{17}/cm^3 (4–10 ppma) in silicon, where the oxygen concentration is based on the "old" ASTM calibration factor, F121-79, Annual Book of ASTM Standards, Part 43, Electronics (1979), can impede dislocation multiplication by dislocation locking (solution hardening) (Sumino, 1983). A somewhat higher concentration of oxygen facilitates the formation of small ($\leq$500 Å) dislocation loops and half loops as a result of the condensation of oxygen agglomerates into small silicon oxide precipitates SiO_x during IC fabrication. That is, as the wafer cools down during IC fabrication, the localized stress, caused by the difference in thermal expansion coefficients between SiO_x and silicon, induces dislocations and silicon interstitials to form around the oxide precipitates. The dislocations are bound to the precipitate, however, and are small enough not to catastrophically multiply. These dislocations and associated precipitate dislocation complexes (PDCs) (Tan *et al.*, 1977) for 20–30 ppma of oxygen also impede additional dislocation movement or recombination by work-hardening the wafer. Precipitate dislocation complexes can exhibit either a positive or a negative influence on IC electronic characteristics. If the PDCs are located sufficiently far away from the device active region (typically >30 μm below the active device area), they can provide sites, due to the localized crystal disruption, for gettering metallic impurities inadvertently introduced during IC fabrication. Metallic impurities become trapped in the strain fields around the PDCs and/or may form metallic silicides by reaction of the metals with silicon interstitials. Internal (intrinsic) gettering requires the controlled incorporation of oxygen within the crystal and the subsequent formation and growth of the PDCs within the wafer to achieve high-performance ICs (Tan *et al.*, 1977; Lawrence and Huff, 1982; Huff *et al.*, 1983; Shimura, 1994). The importance of reducing the carbon concentration to less than 1 × 10^{16}/cm^3 (0.2 ppma) to minimize its catalytic effects on oxygen precipitation continues to receive attention (Fukuda, 1994).

An oxygen concentration of approximately 1.5 × 10^{18}/cm^3 (>30 ppma), in conjunction with no bulk dislocations to assist in dispersing in the oxygen, results in large numbers of silicon oxygen precipitates (size approximately $\geq$2000 Å) during thermal processing, which serve as dislocation generation sources. Dislocation loops can expand beyond a critical stage during subsequent thermal processes resulting in catastrophic multiplication and can lead to a large fraction of ICs (especially bipolar ICs) fabricated in such material exhibiting nonfunctional electrical performance (Taylor *et al.*, 1960; Ravi, 1981; Lawrence and Huff, 1982; Kolbesen and Strunk, 1985). Control of dislocations and macroscopic slip during epitaxial layer deposition and IC fabrication has been extensively investigated (Dyer *et al.*, 1971; Rossi *et al.*, 1985; Wise *et al.*, 1994). The formation of dislocations is an extremely process- and structure-sensitive parameter, dependent on the wafer diameter, interstitial and precipitated oxygen content, carbon content, thermal history of the wafer, latent sources of damage such as periphery damage (e.g., edge chips, crack dislocations), and strain rate (rate of heating/cooling in the furnace), as well as the detailed IC thermal process conditions (and sequence) such as insert (idle) and final temperature, location of wafer support points, wafer spacing, and type of boat material during thermal processing.

The synergistic interaction of the oxygen concentration in the silicon wafer, the wafer-to-wafer spacing in the quartz boat during thermal processing, and the detailed thermal process conditions (including sequence), for example, significantly control the growth of the oxygen precipitates. Excessive oxygen precipitation can result in unacceptable warpage of the silicon wafer precluding effective lithographic printing and reducing the critical stress for plastic deformation or slip by the generation of internal stresses around the precipitates. Process engineers continue aggressively to pursue identification of the optimal oxygen concentration and associated IC thermal-process sequences to utilize internal gettering, perhaps in conjunction with back-surface external gettering, while concurrently preventing uncontrolled dislocation generation, propagation, and the attendant wafer warpage (Lawrence and Huff, 1982; Huff and Goodall, 1995).

The microscopic nature and electrical effect of the various dislocation structures (Kimerling and Patel, 1985) continue to form an important area of investigation. Plastic deformation in silicon at 700 °C was shown to introduce a deep generation–recombination (g–r) energy level, which generally anneals out at 900 °C, leaving behind energy levels much closer to both the conduction- and valence-band edges. Thermal annealing of "freshly formed" dislocations and associated point defects is clearly evidence of bulk reconstruction processes and is reflected in the change of the energy-level spectrum (Kimerling and Patel, 1979). It is now known that dislocation motion produces a whole range of deep energy levels, mostly residual defects from dislocation interactions and defective locations in the dislocation core. Here, also, thermal annealing simplifies the energy level spectrum (Kisielowski and Weber, 1991). Correlation of the structural and electronic properties of dislocations with their photoluminescence *D*-line spectra constitutes a major area of research (Sauer *et al.*, 1985; Weber and Alonso, 1990; Weber, 1994).

2.2.2 Misfit Dislocations. Misfit dislocations, resulting from the mismatched lattice constants between a (generally) lightly doped epitaxial layer and a heavily doped substrate, can partially relieve the misfit-induced strain. The misfit dislocations are localized at the epitaxial layer–substrate interface and are visible to the eye as a cross-hatched array. The cross-hatch lines appear at 90° angles to each other on {100} oriented silicon and at 60° angles on {111} oriented silicon (Wise *et al.*, 1994). Misfit dislocations can provide gettering sites very close to the device active regions, which may be useful in low-temperature IC processes (Bean *et al.*, 1988). However, if misfit dislocations are too close to or terminate at the surface, rather than the wafer edges, they might degrade IC performance by penetrating the device active regions.

2.3 Planar/Interface Defects

2.3.1 Stacking Faults Stacking faults are a local error in the order of the sequential layering of crystallographic planes. The diamond crystal structure normally exhibits an *ABCABCABC*, etc., stacking sequence in the direction of the {111} planes (Bullis and Huff, 1994). A stacking sequence *ABCABABC*, etc., however, does not change the nearest-neighbor (*AB*) configuration. Indeed, the localized energy level associated with the stacking fault lies close to the conduction-band edge in silicon (Kimerling *et al.*, 1977). The second-nearest-neighbor configuration, however, is changed. Stacking faults are surrounded by partial dislocations, which facilitate the faults joining smoothly with the nonfaulted regions within the crystal structure (Cottrell, 1961; Hirth and Lothe, 1982). The partial dislocations exhibit deep energy levels (Kimerling *et al.*, 1977). Decoration of the partial dislocations by metals can cause electrical degradation of the device if they terminate within *p-n* junctions (Huff and Goodall, 1995). Since the stacking-fault nuclei occur in the silicon crystal, significant effort has been expended to understand, control, and minimize the occurrence of these nuclei during crystal growth and subsequent epitaxial deposition.

2.3.2 Twins A twin is a portion of a crystal that contains atoms stacked in one order (e.g., *ABCAB*) on one side of a boundary (the twin plane) and in the mirror image order (e.g., *BACBA*) on the other. Although these structures can occur during Czochralski crystal growth (see Sec. 3.1.2), the crystals or parts of crystals containing them are not dislocation free and, therefore, are recycled or discarded. Twinned structures can also occur during epitaxial silicon deposition. With improvements in current silicon wafer preparation technology and contamination-free manufacturing methodologies, however, twin-plane defects rarely occur in epitaxial silicon (nor germanium) films (Wise *et al.*, 1994). The state of the art for diamond and α- and β-SiC (see Sec. 5), however, is not as advanced.

2.3.3 Surfaces Surfaces and interfaces are significant crystal defects, exhibiting unique physicochemical characteristics and phenomena. The surface energy γ is defined as one-half the energy required to rupture the bonds per unit area of surface at 0 K. Theoretical surface energies for the major crystallographic orientations are listed in Table 7 for the group-IVa semiconductors (Clarke, 1992). The experimental fracture surface energies F for the {111} orientation are comparable to the theoretical estimates (Clarke, 1992). The decrease in γ and fracture surface energy with increasing semiconductor metallicity (weaker covalent bonding), for a given surface orientation, is not unexpected. The variation in γ for the various surface orientations of silicon appears to correlate with the bond density (see Table 8), but more sophisticated models of surface reconstruction are required for quantitative comparison with experiments (Ligenza, 1961; Shimura and Huff, 1985; Shimura, 1989).

Table 7. Major crystallographic theoretical surface energies γ (J/m^2) and experimental fracture surface energy F (J/m^2) for the group-IVa semiconductors.

Element	$\gamma\{100\}$	$\gamma\{110\}$	$\gamma\{111\}$	$F\{111\}$
Diamond	9.82[a]	—	5.35[b]	~6.0
Silicon	2.13	1.51	1.46	1.23
Germanium	1.84	1.30	1.07	1.06
Gray tin	—	—	—	—

[a] Theoretical surface energy $\gamma\{100\}$ (Adamson, 1990).

[b] Alternative theoretical surface energy $\gamma\{111\}$ equals 5.65 J/m^2 (Adamson, 1990).

Table 8. Properties of the major crystallographic surfaces for silicon.

Property	{100}	{110}	{111}
Spacing (nm)	0.543	0.384	0.313
Atomic density (10^{14}/cm^2)	6.78	9.59	7.83
Bond density (10^{14}/cm^2)	13.56	9.59	7.83

Because of its importance to the semiconductor industry, the silicon surface has been extensively studied; both polished and epitaxial silicon surfaces exhibit a plethora of defects, which must be controlled to achieve high-yielding IC performance. These defects include particles, minute shallow pits and/or surface microroughness, residual surface chemical residues, or structural defects such as the occasional epitaxial stacking fault. The measurement, counting, composition, morphology, removal, and prevention of these defects is a state-of-the-art challenge in silicon wafer technology. Laser scanning instruments are widely utilized to monitor these defects (Huff *et al.*, 1995). An inspection technique that has also become widely used in the examination of polished wafers is collimated light. The light illuminates the wafer surface, is reflected to a light-receiving screen, and is transformed to a visual image on a television monitor. By adjusting the contrast and focal depth, a variety of surface anomalies can be qualitatively observed.

In particular, surface microroughness is an imperfection of both the polished and epitaxial silicon surface and may degrade metal oxide semiconductor (MOS) gate oxide integrity (Ohmi *et al.*, 1992; Rajagopalan *et al.*, 1993; Zhong *et al.*, 1996). The magnitude and spatial frequency of surface microroughness, as well as the configuration (i.e., bandwidth) of the measuring instrument used to collect the data (i.e., pixel density, image area), must be considered in characterizing surface microroughness (Bullis, 1994; Strausser *et al.*, 1994). Today's high-quality polished or epitaxial silicon surfaces typically exhibit root mean square (rms) microroughness values from approximately 0.05 to 0.5 nm in the spatial frequency range from approximately 0.1 to 10 μm^{-1}.

3. CRYSTALLINE MATERIAL PROPERTIES

A number of crystalline material properties may conveniently be described in terms of the flux generated in response to an externally applied field. The proportionality constant relating the field to the resulting flux is referred to as a constitutive material property. The isotropic or anisotropic behavior of the constitutive property can be predicted on the basis of the tensor rank of the property and the crystal symmetry (Nye, 1964; Kelly and Groves, 1970). The electrical conductivity, thermal conductivity, and dielectric constant, for example, are second-rank tensors relating two vectors. Each of these tensors is reduced to a single number, and the resulting material properties are isotropic, in the case of the diamond crystal structure. This result is especially important in the case of the electrical conductivity.

The elastic stiffness coefficients, however, form a fourth-rank tensor that relates two second-rank tensors. The elastic properties of the group-IVa semiconductors, therefore, are anisotropic since the elastic stiffness tensor is reduced to three principal (diagonal) coefficients for the diamond crystal structure. Micromachined silicon structures exhibit unique mechanical and electronic characteristics due to their anisotropic mechanical characteristics. A number of interesting phenomena are observed by examining the material properties in the presence of various external fields (see Sec. 3.1.1).

There are experimental difficulties, however, in determining the tensor rank of a number of material properties such as the yield stress, cleavage strength, refractive index (even though the tensor property of the dielectric constant is known), and a host of surface properties such as surface energy, surface hardness, chemical etch rate, and crystal growth rate (Nye, 1964). The diffusion coefficient is an especially interesting example. Although diffusion inside the diamond crystal is isotropic, the surface itself exhibits a potential barrier dependent on the surface orientation. Additionally, very thin regions of a crystal at interfaces or surfaces may exhibit two-dimensional properties and thus display a different symmetry from the bulk (three-dimensional) crystal. The electron mobility in silicon inversion layers, for example, is anisotropic, whereas the bulk mobility is isotropic.

3.1 Electronic Properties

3.1.1 Electrons and Holes Distributed within the vibrating system of atoms and vacancy–interstitial point defects at finite temperatures are free carriers generated by collisions between the atoms and phonons. Electrons excited out of the valence band with energies greater than E_g populate the multiple conduction-band minima at nonzero values of the wave vector **k** and are referred to as "intrinsic" electrons. Concurrently, an equal number of (light and heavy) holes are left resident at the valence-band maximum at the wave vector $\mathbf{k} = 0$ (i.e., zone center) along with holes resident at the split-off valence band, also at $\mathbf{k} = 0$. The total sum of these holes is referred to as "intrinsic" holes, although the split-off holes are often ignored. The offset in **k** space of the valence-band maximum and conduction-band minima accounts for the designation of the group-IVa semiconductors (except gray tin) as indirect energy-gap semiconductors.

The lowest energy of the bands above E_g and the highest energy of the bands below E_g as a function of the reduced wave vector **k** are plotted along the [100] (*X*) and [111] (*L*) directions in Fig. 6 for silicon (Chelikowsky and Cohen, 1976). The minima at the bottom of the lowest conduction band are located at about 80% of the distance between the center ($\mathbf{k} = 0$) and the edge of the Brillouin zone along the six ⟨100⟩ axes. The six equivalent minima, 1.12 eV above the valence-band maximum, are denoted by the Brillouin-zone symmetry point *X* (Harrison, 1980; Ridley, 1982). The surfaces of constant energy are ellipsoids of revolution, rather than spheres, as in the case of GaAs and InSb, so that at each (equivalent) minimum the band curvature varies with direction in **k** space. The effective mass is an important parameter characterizing the energy bands; it is inversely proportional to the curvature of the energy band. The longitudinal electron

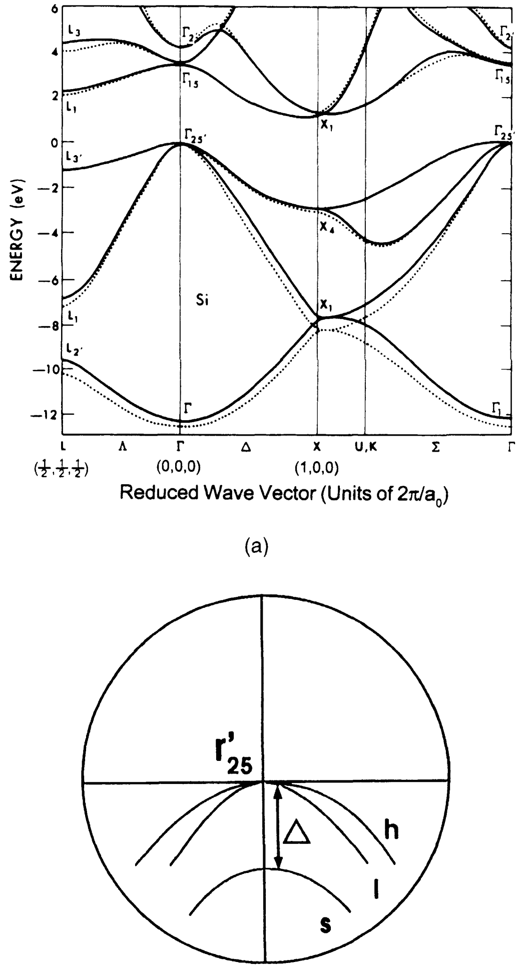

(a)

(b)

FIG. 6. Energy bands of silicon: (a) in the vicinity of the energy gap (following Chelikowsky and Cohen, 1976) and (b) near the valence-band maximum (following Anderson, 1989). The solid (dashed) line in (a) represents a nonlocal (local) pseudopotential analysis.

effective mass normalized by the free electron mass, m_l^*/m_0, describes the variation of energy with wave number measured from the minimum of X in the direction of Γ ($\mathbf{k}$ = 0 symmetry point). The normalized transverse electron effective mass, m_t^*/m_0, describes the variation of energy with wave number in the perpendicular **k**-space direction. Typically, m_l^* is of the order of the free-electron mass, while m_t^* is of the order of a tenth of the free-electron mass (see Table 9). The isotropic density-of-states electron effective mass is defined as (Smith, 1961; Green, 1990)

$$m_{nd}^*/m_0 = [(m_l^*/m_0) \times (m_t^*/m_0)^2]^{1/3}. \tag{4}$$

There are two additional conduction bands in silicon with lowest energy gaps at the Brillouin-zone symmetry points L and Γ. The former is located at the edge of the Brillouin zone along the eight ⟨111⟩ directions. The L and Γ energy gaps are 2.2 and 4.08 eV, respectively, above the valence-band maximum. Transitions between the valence and the L or Γ conduction bands can be observed under optical excitation (see Sec. 3.4). The energy gaps for silicon at the X, L, and Γ symmetry points are summarized in Table 10 (Pollak *et al.,* 1970; Chelikowsky and Cohen, 1976; Ridley, 1982).

In the case of germanium, the minima at the bottom of the lowest conduction band are located at the edge of the Brillouin-zone boundary along the eight ⟨111⟩ directions at the L symmetry point. This behavior in germanium, compared with silicon, is related to the increased lattice constant a_0 (i.e., decreased strength of the interatomic potential) and electronic core effects. The energy gaps for germanium at the X, L, and Γ symmetry points are also summarized in Table 10 (Ridley, 1982). The influence of *d*-band mixing in germanium reduces the L and Γ energy gaps, relative to Si, so that the L indirect gap is smaller than the X indirect gap. The influence of *d*-band mixing is even more dominant in the case of gray tin. In this case, Γ (direct-gap) undercuts the energy gap at the L symmetry point and, accordingly, gray tin is a direct-gap semiconductor (see Table 10) (Pollak *et al.,* 1970). The gray-tin band structure, however, is further complicated by crossing of the lowest conduction band and the upper valence band, resulting in a complicated set of effective masses (Pollak *et al.,* 1970). Fuller comprehension of the variation of these energy gaps requires an examination of the interatomic matrix elements.

The valence bands for the group-IVa elements have a single extremum at the center of the Brillouin zone ($\mathbf{k}$ = 0) (see Fig. 6) (Chelikowsky and Cohen, 1976; Anderson, 1989). The two upper valence bands are degenerate at the maximum and the constant-energy surfaces are warped spheres. The va-

Table 9. Several electronic characteristics for the group-IVa and β-silicon carbide semiconductors at 300 K.[a] See Sec. 4 for a discussion of the variation of the energy gap and mobility for $Si_{1-x}Ge_x$.

Element	m_n^*/m_0	m_p^*/m_0	μ_n[k] (cm^2/V s)	μ_p[k] (cm^2/V s)	n_i[o] (cm^{-3})	σ_i[o] (Ω cm)$^{-1}$
Diamond	1.4[b] 0.36[c] 0.57[d]	0.7[e] 2.18[f] 1.06[g] 2.78[h]	2000	1600	—	$<10^{-15}$
β-Silicon carbide	0.65[b] 0.24[c] 0.33[d]	—	900	40	—	$<10^{-6}$
Silicon	0.92[b] 0.19[c] 0.32[d]	0.153[e] 0.537[f] 0.234[g] 0.590[h]	1450	500	1.07×10^{10}	3.3×10^{-6}
Germanium	1.588[b] 0.0815[c] 0.22[d]	0.043[e] 0.352[f] 0.095[g] 0.36[h]	3900	1800	2.33×10^{13}	2.0×10^{-2}
Gray tin	0.024[i] 0.21[j]	0.2–0.45[e]	1.2×10^5 [l] 3000[m]	3000[n]	—	$\sim 2 \times 10^{-4}$

---

[a] References: Lide, 1993–1994; Anderson, 1989; Madelung, 1991; EMIS Datareview series, 1988; DIAMOND AND DIAMONDLIKE CARBON; GERMANIUM; King, 1994; Davis, 1994; Sze, 1981; Green, 1990; Smith, 1961; Morkoc *et al.*, 1994.

[b] Longitudinal electron effective mass, m_l^*/m_0

[c] Transverse electron effective mass, m_t^*/m_0

[d] Isotropic density-of-states electron effective mass, m_{nd}^*/m_0; see Eq. 4

[e] Light-hole effective mass, m_{lh}^*/m_0

[f] Heavy-hole effective mass, m_{hh}^*/m_0

[g] Split-off hole effective mass, m_{so}^*/m_0

[h] Isotropic density-of-states hole effective mass, m_{pd}^*/m_0; see Eq. 5

[i] Light electron effective mass, m_{le}^*/m_0

[j] Heavy electron effective mass, m_{he}^*/m_0

[k] Mobilities are for lightly doped material.

[l] Light electron, 100 K.

[m] Heavy electron, 270 K.

[n] 270 K.

[o] See discussion associated with Eqs. 6a through 9.

Table 10. Energy gaps for the group-IVa and β-silicon carbide semiconductors at *L*, *X*, and *Γ* symmetry points and split-off valence-band energy (*Δ*) at 300 K.

Element	*L* (eV)	*X* (eV)	*Γ* (eV)	*Δ* (eV)
Diamond	5.77[a]	5.45	13.04	0.006
β-Silicon carbide	—	2.2	7.75	—
Silicon	2.2[b]	1.12	4.08	0.044
Germanium	0.66	0.96	0.89	0.29
Gray tin	0.3	0.35[c]	0.082	0.64

---

[a] Theoretical calculation (Ridley, 1982).

[b] Theoretical calculation (Chelikowsky and Cohen, 1976).

[c] Pollak *et al.* (1970).

lence band exhibiting the smaller curvature is denoted as the "heavy-hole" band while the higher curvature valence band is the "light-hole" band. A third valence band is split off from the other two by relativistic spin–orbit interactions. The split-off valence band, *Δ*, is 0.044 eV below the maximum of the light- and heavy-hole bands for silicon. The curvature (and, therefore, effective mass) of this approximately spherical band is intermediate between the curvature at the maxima of the heavy- and light-hole bands for silicon and germanium. The light-hole effective mass m_{lh}^*/m_0, heavy-hole effective mass m_{hh}^*/m_0, and split-off hole effective mass m_{so}^*/m_0, are summarized in Table 9.

The isotropic density-of-states hole effective mass is defined as (Green, 1990)

$$m_{pd}^*/m_0 = \{(m_{lh}^*/m_0)^{3/2} + (m_{hh}^*/m_0)^{3/2} + [(m_{so}^*/m_0)\exp(-\Delta/kT)]^{3/2}\}^{2/3}. \quad (5)$$

The split-off valence-band energies (Ridley, 1982) are also summarized in Table 10.

The number of intrinsic holes (p_i) and electrons (n_i) per cm^3 at thermal equilibrium, for the nondegenerate case, are described by the Boltzmann statistics (Smith, 1961):

$$p_i = N_V \exp(-|E_{F,i}^V|/kT), \quad (6a)$$

$$n_i = N_C \exp(-|E_{F,i}^C|/kT), \quad (6b)$$

where N_v is the effective density of states in the valence band, equal to $2(2\pi m_0 kT/h^2)^{3/2} \times (m_{pd}^*/m_0)^{3/2}$; N_C is the effective density of states in the conduction band, equal to $2(2\pi m_0 kT/h^2)^{3/2}\ M(m_{nd}^*/m_0)^{3/2}$; M is the number of minima in the conduction band [i.e., six for Si and four for Ge (Smith, 1961)]; $|E_{F,i}^V|$, $|E_{F,i}^C|$ are the absolute values of the intrinsic Fermi energy relative to the valence-band edge and conduction-band edge, respectively; and m_{pd}^*, m_{nd}^*, and m_0 are the isotropic density-of-states hole effective mass, the isotropic density-of-states electron effective mass, and the free-electron mass, respectively.

The intrinsic electron–hole excitation phenomenon is described by the mass-action relationship (Smith, 1961):

$$n_i p_i = n_i^2. \quad (7)$$

Utilizing Eqs. (6a), (6b), and (7), the intrinsic electron and hole concentrations per cm^3 may be determined:

$$n_i = p_i = (N_C N_V)^{1/2} \exp(-E_g/2kT), \quad (8)$$

where E_g is the temperature-dependent activation enthalpy (i.e., the minimum energy gap) equal to $|E_{F,i}^V| + |E_{F,i}^C|$. The intrinsic concentration of electrons and holes, calculated for silicon and germanium at 300 K utilizing Eq. (8) and the electronic characteristics summarized in Tables 9 and 10, is a fundamental material electrical characteristic. The corresponding intrinsic conductivity is determined at 300 K using Table 9 and the relation

$$\sigma_i = p_i|e|\mu_p + n_i|e|\mu_n, \quad (9)$$

where μ_p and μ_n are the temperature-dependent hole and electron mobilities, respectively, in the case of phonon-limited scattering at 300 K (Sze, 1981).

The coefficients of electrical conductivity, σ_{kl}, a second-rank tensor, relate the electric field component E_l to the component of electric charge flow per unit area per second j_k (i.e., the current per unit area) in the kth direction. The electrical conductivity tensor σ_{kl} can be transformed to its principal axes and, therefore, principal conductivity coefficients, σ_1, σ_2, and σ_3, which are equal (isotropic) for the diamond crystal structure (see Eq. 9). Although the individual conduction bands for the group-IVa semiconductors are anisotropic, the macroscopic conductivity is isotropic as a result of the symmetry of the diamond crystal structure. An external mechanical stress or magnetic field, however, can lift the degeneracy of the conduction bands, and the conductivity will no longer be isotropic, leading to a number of useful electronic device applications.

3.1.2 Donors and Acceptors The fundamental property of a crystalline semiconductor that results in its technological importance is the easily controllable and spatially variable concentration and type of free carriers (Irvin, 1962; Thurber *et al.*, 1980a, 1980b). According to G. K. Teal (Teal, 1976; Teal *et al.*, 1985):

> "I reasoned that polycrystalline germanium, with its variations in resistivity and its randomly occurring grain boundaries, twins and crystal defects that acted as uncontrolled resistances, electron or hole emitters and traps would affect transistor operation in uncontrolled ways. . . . My general aims for the single crystal research were:
>
> 1. to produce a conducting medium in which a high degree of crystal perfection, of uniformity of structure and of chemical purity is attained; and
> 2. to build into this highly perfect medium in a controlled way the re-

> quired resistivities and electrical boundaries to give a variety of device possibilities by control of the chemical composition (i.e., donor and acceptor concentration) along the direction of single crystal growth."

Significant perturbation of the intrinsic electron–hole concentrations is accomplished by the selective (extrinsic) doping of the semiconductor. Group-Va donor impurities such as phosphorus, arsenic, or antimony result in free electrons in the conduction band (*n*-type semiconductor) while the group-IIIa acceptor impurities such as boron result in free holes in the valence band (*p*-type semiconductor). These dopants occupy substitutional sites in the group-IVa semiconductors with shallow ground-state energy levels in the energy gap. The energy levels are described by the hydrogenic approximation in the effective-mass theory, taking into account the dielectric constant of the semiconductor and the effective mass of the free carrier (i.e., electron or hole in the semiconductor) (Harrison, 1980). The dopant wave functions are localized (i.e., nonpropagating) solutions to Schrödinger's equation for doping densities less than the isotropic density of states. The ground states are typically 40–50 meV from the appropriate band edge for silicon (donors below the conduction-band edge, acceptors above the valence-band edge) and are readily ionized at 300 K. Table 11 summarizes the energy levels of several common dopants in diamond, β-SiC, silicon, and germanium (Sze, 1981; Beadle *et al.*, 1985).

The dopant impurities are introduced into the growing silicon (or germanium) crystals during growth from the melt by the modified Czochralski (Teal–Little) technique (Teal *et al.*, 1985; Runyan and Huff, 1994). The concentration $C_s(g)$ along the direction of crystal growth, for the group-IIIa or group-Va dopant impurity, is described by the equation for normal freezing (Pfann, 1966):

$$C_s(g) = k_e C_1 (1 - g)^{k_e(d_l/d_s) - 1}, \qquad (10)$$

where g is the fraction of crystal solidified (seed end, $g = 0$; bottom end, g approaches 1); C_l is the initial dopant concentration in the melt [$C_s(0) = k_e C_l$]; k_e is the effective distribution coefficient, equal to the ratio of the dopant concentration in the crystal to the dopant concentration in the bulk melt (i.e., far away from the crystal–melt interface); and d_l/d_s is the ratio of the melt density to the crystal density. This ratio, about 1.1 and 1.05 for silicon and germanium, respectively, is typically taken as unity since the experimental uncertainty in k_e is generally greater than 10%.

The effective distribution coefficient k_e is dependent on the crystal growth rate and melt flow conditions at the crystal–melt interface and is related to the equilibrium distribution coefficient k_0 by the Burton–Prim–Slichter relationship (Burton *et al.*, 1953):

Table 11. Energy levels (meV) for several common dopants in group-IVa and β-silicon carbide semiconductors.[a]

Dopant	Diamond[b]	β-Silicon carbide[c]	Silicon	Germanium	Gray Tin
B	370	—	45	10	—
Al	—	200	67	10	—
Ga	—	—	72	11	—
P	—	—	44	12	—
As	—	—	49	13	—
Sb	—	—	39	10	—

---

[a] Energy levels are measured from the conduction-band edge for *n*-type dopants and the valence-band edge for *p*-type dopants.

[b] Nitrogen and phosphorus are *n*-type dopants. Phosphorus is believed to be too large to fit within the diamond crystal (DIAMOND AND DIAMONDLIKE CARBON), but donor energy levels about 1.7 and 4.0 eV below the conduction band edge have been identified (Davies, 1994; Collins, 1990).

[c] For nitrogen concentrations $\leq 10^{17}/\text{cm}^3$, the electrical ionization energy approximately equals 45 meV (Pensl and Troffer, 1995). For $10^{17}/\text{cm}^3$–$10^{19}/\text{cm}^3$, the electrical ionization energy decreases to as low as approximately 20 meV for $10^{19}/\text{cm}^3$ (Pensl and Choyke, 1993).

Table 12. Equilibrium distribution coefficients for several dopants at the melting points of silicon[a] and germanium.

Dopant	Silicon	Germanium
B	0.8	17
Al	0.002	0.073
Ga	0.008	0.087
Si	1	5.5
Ge	0.33	1
P	0.35	0.08
As	0.30	0.02
Sb	0.023	0.003

---

[a] The equilibrium distribution coefficient of oxygen, an important impurity in silicon, is about 0.25 (Lin and Hill, 1983).

$$k_e = k_0[k_0 + (1 - k_0)\exp(-f\delta/D)]^{-1}, \tag{11}$$

where f is the microscopic crystal growth rate, δ is the thickness of the stagnant boundary layer in the melt at the crystal–melt interface, and D is the dopant diffusivity in the bulk melt.

Equilibrium distribution-coefficient data for several dopants in silicon and germanium are summarized in Table 12 (Trumbore, 1960). It is of particular interest that k_0 for acceptor impurities is greater in germanium than for silicon while k_0 for donor impurities is greater in silicon than for germanium. These observations may correlate with the experimental observation that donor dopants diffuse faster in germanium and acceptor dopants diffuse faster in silicon (Kooi, 1991). The interpretation of Fig. 4 that interstitials are more predominant in silicon than in germanium may be consistent with this observation. That is, a dopant with a larger (smaller) value of k_0 might preferentially be (not be) incorporated within the crystal and, therefore, exhibit a smaller (larger) diffusion coefficient. These ideas are qualitatively summarized in Table 13, although detailed variations occur between each dopant for both silicon and germanium. The defect-mediated diffusion by point defects in Table 13 refers to the dopants in Table 12.

The donor (acceptor) impurity becomes positively (negatively) charged when the free electron (hole) resides in the conduction (valence) band. Charge neutrality requires

$$n + N_A^- = p + N_D^+, \tag{12}$$

where N_A^- and N_D^+ are the ionized acceptor and donor concentrations, respectively.

One may utilize the generalized mass-action relationship (applicable in the case of nondegenerate statistics at thermal equilibrium) (Smith, 1961),

$$np = n_i^2, \tag{13}$$

with the charge neutrality condition from Eq. (12) to relate the electron and hole concentrations.

For example, an (essentially) uncompensated p-type silicon wafer at 300 K with an acceptor concentration N_A of 1.3×10^{15} boron atoms per cm^3 (10 Ω cm) and an intrinsic concentration of 1.07×10^{10} cm^{-3} (Green, 1990) has hole and electron concentrations

$$p \sim N_A^- = 1.3 \times 10^{15}\,\mathrm{cm}^{-3}, \tag{14a}$$

$$n = n_i^2/p = 8.81 \times 10^4\,\mathrm{cm}^{-3}. \tag{14b}$$

The corresponding extrinsic conductivity σ equals 9.88×10^{-2} (Ω cm)$^{-1}$, calculated from Eqs. (9), (14a), and (14b) with p_i and n_i replaced by p and n, respectively, in Eq. (9). For this dopant concentration, the phonon-limited mobilities at 300 K are valid approximations to the usual dopant-density–dependent mobilities (Sze, 1981) (see Table 9).

Table 13. Qualitative correlation among equilibrium distribution coefficient, predominant point defect, and diffusion coefficient at high temperature in silicon and germanium.

Element	k_0 (p type)	k_0 (n type)	$[I]$[a]	$[V]$[a]	D (p type)	D (n type)
Silicon	Less[b]	More	More	Less	More	Less
Germanium	More[b]	Less	Less	More	Less	More

---

[a] Interstitial and vacancy concentrations, respectively.

[b] Less and more refer to a comparison between silicon and germanium.

In this case, the holes are referred to as majority carriers, and the electrons are minority carriers. Under thermal equilibrium conditions, the electron (minority carrier) contribution to the extrinsic conductivity is negligible in comparison to the hole (majority carrier) contribution. The hole and electron concentrations for the nondegenerate extrinsic case are related to the Fermi energy by analogy with Eqs. (6a) and (6b):

$$p = N_V \exp(-|E_F^V|/kT), \tag{15a}$$

$$n = N_C \exp(-|E_F^C|/kT), \tag{15b}$$

where $|E_F^V|$ and $|E_F^C|$ are the absolute values of the Fermi energy relative to the valence-band edge and conduction-band edge, respectively.

Alternatively, one may combine Eqs. (6a), (6b), (15a), and (15b) to yield

$$p = p_i \exp(+|\varphi_F|), \tag{16a}$$

$$n = p_i \exp(-|\varphi_F|), \tag{16b}$$

where $|\varphi_F|$ is the absolute value of the Fermi energy relative to the intrinsic Fermi energy normalized by kT. It should be noted that, although the Fermi energy, or alternatively the normalized Fermi energy, is spatially constant at thermal equilibrium, it may nevertheless be spatially variable relative to the band edges, even at thermal equilibrium. In this case, the free-carrier flux due to the spatially varying internal electric field set up by the energy-band bending [resulting in a space-charge region (SCR)] is exactly balanced by the free-carrier diffusion flux. The intrinsic Fermi energy, however, follows the energy-band bending. Similar concepts may be utilized to describe the surface energy-band bending due to surface states and the resulting surface space-charge region (SSCR) at thermal equilibrium.

3.1.3 Metal Impurities Transition metals in silicon exhibit a significant component of interstitial, as well as substitutional, solubility. This results in a mixed diffusion process, with a fast interstitial component. The *3d* transition metals in silicon, such as iron and chromium, mainly occupy and diffuse via interstitial sites. They accordingly exhibit high diffusion coefficients. Cobalt, Ni, and Cu diffuse sufficiently fast, even at 300 K, that these elements transfer from interstitial sites onto dislocation sites as well as form clusters and precipitates. The *3d* metals in germanium, on the other hand, occupy and diffuse via substitutional sites (vacancies are the dominant point defect in germanium). The *4d* metals in silicon, such as Ag, exhibit a solubility on substitutional sites comparable to interstitial sites and, therefore, diffuse more slowly than the *3d* transition metals. The *5d* metals, such as Au and Pt, exhibit a dominant substitutional solubility. The *5d* metals accordingly take longer until their substitutional solubility is attained throughout the wafer. The U-shaped diffusion profile of Au, however, is clear evidence that there is also a fast-diffusing interstitial component. The lack of sufficient interstitial sinks in dislocation-free silicon limits the kinetics of the transfer of Au onto substitutional sites (kick-out mechanism) during diffusion (Taylor *et al.*, 1993). The temperature-dependent solubility and diffusion coefficients, taking into account kinetic considerations during cool-down from high temperatures, are especially important for transition metals such as Au in determining their state of aggregation at room temperature. The properties of transition metals in silicon have been reviewed with particular emphasis on their diffusion, solubility, and electrical activity (Weber, 1988; Lemke, 1994). Size effects and electronegativity differences between the impurity and host atoms are also important considerations and must be taken into account for these analyses.

Although metallic impurities can affect the electrical resistivity of silicon if they are present in large enough concentration, they are more important as traps or generation–recombination (g–r) centers that can degrade the carrier lifetime (Schroder, 1990; Bullis and Huff, 1995). The carrier lifetime is an extremely important example of the structure sensitivity of material properties in semiconductors. Metallic elements exhibit deep energy levels (i.e., far from the band edges) in silicon with ground states typically 0.4–0.6 eV from the nearest band edge (Sze, 1981). In general, metals are considered undesirable in silicon even at levels approaching 1 part per trillion atomic (1 ppta = 5×10^{10} atoms/cm^3). Since most transition elements have very small distribution coefficients (10^{-5}) in silicon (Trumbore, 1960), their direct incorporation from the melt during crystal growth is unlikely if normal precautions are taken. Metallic contamination by Fe, Ni, and Cu in the range of 10^{11}–10^{13} cm^{-2}—for example, during ion implantation, dry etching, and plasma CVD—is much more likely to be introduced onto the wafer surface during wafer preparation or IC fabrication (Huff and Goodall, 1996a). The surface metals diffuse into the wafer bulk during subsequent thermal processing. The surface and bulk descriptions, however, cannot be simply transposed via the wafer thickness because of surface–bulk segregation during thermal processing. In view of this, prescriptions are under development to determine the equivalent bulk concentration from a given surface concentration during subsequent thermal processing (Helms, 1995). An extensive methodology has developed for the removal of surface metals by cleaning with suitable solutions of high-purity chemicals (Kern, 1993). The sensitivity of surface diagnostic techniques varies but ranges between $\approx 10^8$ and 10^{10} atoms/cm^2 for a variety of transition metals (Diebold *et al.*, 1992).

3.2 Thermal Properties

The Debye temperature θ_D is related to the maximum phonon (quantized normal mode of vibration) frequency ν_D that can propagate in the crystalline material at a given temperature (De Launay, 1956):

$$k\theta_D = h\nu_D. \tag{17}$$

The value for θ_D at 300 K decreases from diamond to gray tin, because of the decrease in the bond energy (cohesive energy) of the group-IVa semiconductors and the increase in the interatomic distance, making it easier to populate higher-energy phonon modes. β-SiC's Debye temperature is virtually intermediate between diamond and silicon's values. The Debye temperature is a key parameter in thermal characteristics, including the melting point, coefficient of linear thermal expansion α, and thermal conductivity κ. These data are listed in Table 14 for 300 K (De Launay, 1956; Slack and Bartram, 1975; Field, 1991).

3.2.1 Thermal Expansion The coefficients of thermal expansion α_{ij}, a second-rank tensor, relate a scalar ΔT to a second-rank strain tensor ϵ_{ij}. When a small temperature change ΔT is uniformly applied throughout the crystal, all the components of ϵ_{ij} are proportional to ΔT. The thermal expansion tensor α_{ij} can be transformed to its principal axes and, therefore, principal thermal expansion coefficients, α_1, α_2, and α_3. These coefficients are equal (isotropic) for the diamond crystal structure (see Table 14). In heavily doped silicon at 300 K, α is expected to increase above 2.57×10^{-6}/K (Bullis and Huff, 1994), but in all cases it is smaller than the value for germanium (see Fig. 4). The values of α for silicon and silicon carbide are significantly different from those of metals, germanium, and diamond. In relatively pure silicon, α is very small and positive below 18 K. It is negative between 18 K and about 120 K, reaching a minimum of about -0.7×10^{-6}/K at about 80 K. Above approximately 120 K, α becomes positive and increases monotonically to a nearly constant value beyond 1200 K of about 4.4

Table 14. Several thermal characteristics of the group-IVa and β-silicon carbide semiconductors at 300 K.

Element	Debye temperature (K)	Melting temperature (°C)	Coefficient of linear thermal expansion, α (10^{-6} K^{-1})	Thermal conductivity, κ (W/cm K)
Diamond	1862	3827[a]	1.05	20
β-Silicon carbide	1430	>1827[b]	2.77	5.0
Silicon	645	1412	2.57	1.5
Germanium	374	958	5.90	0.6
Gray tin[c]	260	—	~5.5	—

---

[a] Diamond-graphite-liquid eutectic at 1.23×10^5 atm.

[b] Sublimes at 1 atm pressure; melts at 2830 °C at 35 atm.

[c] Gray tin transforms to white tin at 13 °C; white tin melts at 232 °C.

$\times$ 10^{-6}/K (Slack and Bartram, 1975). At temperatures above 1000 K, silicon exhibits the smallest value of α for a number of diamondlike solids (Slack and Bartram, 1975). As noted in Sec. 2.1.1, this contributed to the interpretation that interstitials are as significant as vacancies in silicon. The high-temperature dependence of α for silicon carbide also appears consistent with a similar interpretation. Although α for SiC is similar to that of silicon at 300 K, its saturated value of 5.65 $\times$ 10^{-6}/K is larger than that for silicon.

The difference in α between silicon and various films such as silicon dioxide or silicon nitride, which are used in silicon IC technology, is particularly important. It is responsible for stresses that can generate and propagate dislocations in silicon resulting in plastic deformation and subsequent degradation of IC performance (Lawrence and Huff, 1982; Huff and Shimura, 1985; Huff and Goodall, 1995). These effects are particularly important at corners in windows of patterned films and trench structures (Kooi, 1991).

3.2.2 Thermal Conductivity The coefficients of thermal conductivity κ_{ij}, a second-rank tensor, relate the negative of the temperature gradient component $\partial T/\partial x_j$ to the component of heat flow per unit area per second (h_i) in the ith direction. The thermal conductivity tensor κ_{ij} can be transformed to its principal axes and, therefore, principal thermal conductivity coefficients, κ_1, κ_2, and κ_3. These coefficients are equal (isotropic) for the diamond crystal structure. The thermal transport mechanism in semiconductors is predominantly by phonons, rather than by free electrons (as in metals) or holes. Thus, the Wiedemann-Franz law, which relates the electronic thermal conductivity to the electrical conductivity in metals, does not apply to semiconductors.

Natural diamond consists of 98.892% of C^{12} and 1.108% C^{13}. Enriched diamond with 99.93% C^{12} has a large room-temperature value of κ of about 33 W/cm K (Field, 1991; DIAMOND AND DIAMONDLIKE CARBON). Silicon, germanium, and tin also exhibit a number of natural isotopic species (see Table 15). The isotope effect on the thermal conductivity for these elements, however, is significantly smaller compared to diamond because of their much larger atomic mass. For example, the mass differential between C^{12} and C^{13} is 8.3% while the mass differential between Si^{28} and Si^{29} is 3.6%. The room-temperature value of κ for silicon is 1.5 W/cm K (see Table 14). As the temperature increases above 300 K, κ decreases on account of increased phonon scattering until it reaches a value of about 0.22 W/cm K at the melting point of silicon (Runyan, 1965). The thermal conductivity increases for silicon below room temperature, reaching a maximum at about 20–30 K as a result of reduced phonon scattering. Below this temperature range, κ decreases because of increased boundary (umklapp) scattering. In this region, κ is also reduced by increased impurity concentration. The room-temperature value of κ for silicon is essentially independent of impurity concentration, except for concentrations greater than about $10^{18}/cm^3$ (Bullis and Huff, 1994). At these free-carrier concentrations, κ decreases as a result of increased phonon–electron (or hole) intravalley and intervalley scattering processes. The potential increase in κ due to electron (or hole) transport is apparently more than offset by these increased scattering processes (Runyan, 1965).

Table 15. Natural isotopic species and percent for the group-IVa elements.

Element	Isotopic species/Percent
Carbon	C^{12}, 98.892%; C^{13}, 1.108%
Silicon	Si^{28}, 92.23%; Si^{29}, 4.67%; Si^{30}, 3.10%
Germanium	Ge^{70}, 20.4%; Ge^{72}, 27.4%; Ge^{73}, 7.8%; Ge^{74}, 36.6%; Ge^{76}, 7.8%
Tin	Sn^{112}, 1.01%; Sn^{114}, 0.68%; Sn^{115}, 0.35%; Sn^{116}, 14.28%; Sn^{117}, 7.67%[a]; Sn^{118}, 23.84%; Sn^{119}, 8.68%[a]; Sn^{120}, 32.75%; Sn^{122}, 4.74%; Sn^{124}, 6.01%

[a] Mass number of "stable" radioactive isotopes.

3.3 Mechanical Properties

The variation of the group-IVa semiconductor mechanical properties generally correlates with the variation of the metallicity parameter α_m and Debye temperature θ_D. These properties include the hardness number (see Sec. 2.2.1), elastic stiffness coefficients, Young's modulus Y, bulk modulus K, shear modulus G, Poisson's ratio ν, and the theoretical fracture strength σ_{th}. The reciprocal of Young's modulus Y, which relates an applied stress σ to the resulting strain ϵ, [Hooke's law, $\epsilon = (1/Y)\sigma$], is an especially important material characteristic.

The mechanical properties are often expressed via an alternative, but equivalent, representation of the generalized form of Hooke's law:

$$\sigma_{ij} = c_{ijkl}\epsilon_{kl}, \tag{18}$$

where the elastic stiffness coefficients c_{ijkl}, composing a fourth-rank tensor, relate the strain ϵ_{kl} to the stress σ_{ij}. Both the strain and the stress are second-rank tensors. The symmetry of both the strain tensor and, in the absence of body torques, the stress tensor implies symmetry of the c_{ijkl} in the first two and last two suffixes, respectively. This reduces the 81 c_{ijkl} coefficients to 36 c_{mn} coefficients ($c_{11} \ldots c_{66}$), thereby facilitating utilization of matrix notation. That is, the first two suffixes are abbreviated into a single one running from 1 to 6, and the last two suffixes are abbreviated in the same way. Accordingly, the tensor notation 11, 22, 33, 23 and 32, 31 and 13, and 12 and 21 for σ_{ij} and ϵ_{kl} are reduced to 1, 2, 3, 4, 5, and 6, respectively (Nye, 1964). The assumption of a parabolic crystal potential (i.e., $c_{mn} = c_{nm}$) results in 21 independent coefficients. Finally, symmetry considerations for cubic crystals such as the diamond crystal structure (i.e., cyclic interchange of the x, y, and z coordinates) results in three independent elastic stiffness coefficients, c_{11}, c_{12}, and c_{44}, and anisotropic elastic properties (see Table 16) (McSkimin *et al.*, 1951; McSkimin, 1953; Nye, 1964; Runyan, 1965; Henisch and Roy, 1969; Kelly and Groves, 1970; Madelung, 1991; Chen *et al.*, 1992; Tu *et al.*, 1992). It should be noted that the c_{mn} are not the components of and, therefore, do not transform like the components of a second-rank tensor. To transform the c_{mn} to other axes, it is necessary to go back to tensor notation. The elastic properties form an important example in which the family of cubic crystal structures exhibit anisotropic behavior of a material property.

Inasmuch as the elastic stiffness coefficients are determined by the interatomic forces in the crystal, the general decrease in the elastic stiffness coefficients in Table 16 with increasing metallicity and increased lattice constant a_0 (see Table 2) is expected. The composition dependence of the elastic stiffness coefficients of the silicon–germanium alloys and superlattices is not well understood (see Table 16). However, the most reliable data suggest a linear interpolation between the elastic stiffness coefficients of silicon and germanium, despite the 4.2% difference in the bond lengths of silicon and germanium (see Table 2) (Mendik *et al.*, 1991; Baker and Arzt, 1995). A more systematic study of the elastic properties of silicon–germanium alloys is required (Roos *et al.*, 1995a, 1995b).

Although microscopic models are generally expressed in terms of the c_{ij} coefficients, computation of the engineering parameters Y, K, G, and ν is facilitated by utilizing the

Table 16. Elastic stiffness coefficients for the group-IVa, β-silicon carbide, and several silicon–germanium alloy and superlattice semiconductors at 300 K.

Element	c_{11} (10^{11} dyn/cm^2)	c_{12} (10^{11} dyn/cm^2)	c_{44} (10^{11} dyn/cm^2)
Diamond	107.640	12.520	57.740
β-Silicon carbide	35.23	14.04	23.29
Silicon	16.772	6.498	8.036
SiGe (5084 Å)[a]	14.7 ± 0.36	5.90 ± 0.42	7.30 ± 0.22
SiGe (6795 Å)[b]	14.29 ± 0.34	5.64 ± 0.41	6.81 ± 0.20
$Si_{0.49}Ge_{0.51}$[c]	13.77 ± 0.33	5.12 ± 0.4	6.66 ± 0.19
Germanium	13.112	4.923	6.816
Gray tin	6.90	2.93	3.62

---

[a] Material density is 2.79 g/cm^3; the Å data refer to the superlattice thickness (Mendik *et al.*, 1991).
[b] Material density is 3.27 g/cm^3; the Å data refer to the superlattice thickness (Mendik *et al.*, 1991).
[c] Material density is 3.85 g/cm^3 (Mendik *et al.*, 1991).

Table 17. Elastic compliance coefficients for the group-IVa, β-silicon carbide, and several silicon–germanium alloy and superlattice semiconductors at 300 K.

Element	s_{11} $(10^{13}\ \mathrm{dyn/cm^2})^{-1}$	s_{12} $(10^{13}\ \mathrm{dyn/cm^2})^{-1}$	s_{44} $(10^{13}\ \mathrm{dyn/cm^2})^{-1}$
Diamond	0.952	−0.0992	1.73
β-Silicon carbide	3.67	−1.05	4.29
Silicon	7.61	−2.12	12.44
SiGe (5084 Å)[a]	8.83	−2.53	13.70
SiGe (6795 Å)[b]	8.95	−2.53	14.68
$\mathrm{Si_{0.49}Ge_{0.51}}$[c]	9.05	−2.45	15.02
Germanium	9.59	−2.62	14.67
Gray tin	19.4	−5.78	27.62

[a] Material density is 2.79 g/cm^3; the Å data refer to the superlattice thickness (Mendik *et al.*, 1991).
[b] Material density is 3.27 g/cm^3; the Å data refer to the superlattice thickness (Mendik *et al.*, 1991).
[c] Material density is 3.85 g/cm^3 (Mendik *et al.*, 1991).

elastic compliance coefficients s_{ijkl}. The s_{ijkl} relate a sufficiently small applied stress σ_{kl} to strain ϵ_{ij} (Hooke's law). Note that the s_{ijkl} tensor is the reciprocal of the c_{ijkl} tensor. Here also, the 81 s_{ijkl} coefficients reduce to three s_{mn} coefficients for the diamond crystal structure. The s_{mn} coefficients are related to the c_{ij} by

$$s_{11} = (c_{11} + c_{12})/[(c_{11} - c_{12})(c_{11} + 2c_{12})], \tag{19a}$$

$$s_{12} = -c_{12}/[(c_{11} - c_{12})(c_{11} + 2c_{12})], \tag{19b}$$

$$s_{44} = 1/c_{44}, \tag{19c}$$

and are summarized in Table 17 (Runyan, 1965). The c_{mn} coefficients can be expressed in terms of the s_{mn} by interchanging the c's and s's of Eqs. (19) (Nye, 1964; Runyan, 1965).

The expression for Young's modulus Y is

$$[Y(l_1l_2l_3)]^{-1} = s_{11} - 2(s_{11} - s_{12} - 0.5s_{44}) \times [(l_1^2l_2^2 + l_2^2l_3^2 + l_3^2l_1^2)]. \tag{20}$$

Young's modulus is anisotropic and varies with direction via the expression in brackets containing the direction cosine terms, l_1, l_2, l_3, with respect to the unit vectors along the ⟨100⟩ axes, with indices 1, 2, and 3 corresponding to the x, y, and z (⟨100⟩) axes, respectively (Nye, 1964). The direction cosine term is zero for the ⟨100⟩ directions, 1/4 for the ⟨110⟩ directions, and 1/3 for the ⟨111⟩ directions. Accordingly, if $s_{11} - s_{12} - 0.5s_{44}$ is positive (as is the case for the group-IVa semiconductors), Young's modulus is a maximum in the ⟨111⟩ directions, a minimum in the ⟨100⟩ directions, and intermediate for the ⟨110⟩ directions. For an applied force lying in an arbitrary direction in the {100} plane, the value of Young's modulus is between the ⟨100⟩ and ⟨110⟩ values. For an applied force lying along an arbitrary direction in a {111} plane, Young's modulus is constant and equal to the value for the ⟨110⟩ direction (Riney, 1961; Runyan, 1965; Bullis and Huff, 1994). Experimental data determined for Y are summarized in Table 18 for the major crystallographic orientations for the group-IVa and β-SiC semiconductors (Tu *et al.*,

Table 18. Experimental Young's modulus data for the major crystallographic directions for the group-IVa and β-silicon carbide semiconductors compared to expected values (in parentheses) at 300 K.

Element	⟨100⟩ $(10^{12}\ \mathrm{dyn/cm^2})$	⟨110⟩ $(10^{12}\ \mathrm{dyn/cm^2})$	⟨111⟩ $(10^{12}\ \mathrm{dyn/cm^2})$
Diamond	— (10.504)	— (11.64)	11.60 (12.077)
β-Silicon carbide	— (2.725)	— (3.85)	4.80 (4.46)
Silicon	1.302 (1.314)	1.689 (1.708)	1.875 (1.897)
Germanium	1.037 (1.043)	1.380 (1.398)	1.551 (1.569)
Gray tin	— (0.515)	— (0.729)	— (0.845)

1992; Wentorf, 1992; Schlichting and Riley, 1992; Bullis and Huff, 1994). These data compare rather favorably with the values calculated from Eqs. (20) (parenthesized data in Table 18) using the elastic compliance coefficients in Table 17. The decrease in Young's modulus, for a given direction, with increasing metallicity (weaker covalent bonding) appears reasonable.

The elastic stiffness coefficients in silicon have been noted to decrease fairly linearly with temperature between approximately 27 and 700 °C; the quantity $c_{ij}^{-1}\, dc_{ij}/dT$ varies as -9.4×10^{-5}/K, -9.8×10^{-5}/K, and -8.3×10^{-5}/K for ij equal to 11, 12, and 44, respectively (Nielsen, 1988). The temperature dependence of Young's modulus, in an isotropic approximation over approximately 800–1200 °C, has been fitted as $Y \sim 5.8 \times 10^{-2} \exp[(1\ \text{eV})/kT]$ dyn/cm^2 (Wijaranakula, 1993).

The reciprocal of the bulk modulus K (the reciprocal of the volume compressibility) is an isotropic parameter relating an applied three-dimensional unit hydrostatic pressure to the fractional change in volume. The reciprocal of the linear compressibility β relates an applied unit hydrostatic pressure to the fractional change in length in the same direction and, in general, depends on direction. The reciprocal of the shear modulus G relates an applied shear force to the fractional change in angular displacement and also varies with direction. Poisson's ratio ν is the negative ratio of the strain perpendicular to an applied stress to the strain in the direction of the applied stress. Poisson's ratio depends on both the direction of the applied stress and the measured perpendicular direction of contraction. These mechanical characteristics may be expressed in terms of the elastic compliance coefficients, for the diamond crystal structure, by (Riney, 1961; Nye, 1964)

$$K = \{3(s_{11} + 2s_{12})\}^{-1}, \tag{21}$$

$$\beta = s_{11} + 2s_{12}, \tag{22}$$

$$G = 1/s_{44}, \tag{23}$$

$$\nu = -s_{12}/s_{11}. \tag{24}$$

The values of K, Y, and ν are related in the isotropic case by (Nye, 1964)

$$K = Y\{3(1 - 2\nu)\}^{-1}. \tag{25}$$

The values of Y, K, G, and ν for an isotropic polycrystalline aggregate, computed from single-crystal data by a Green's-function method, are summarized in Table 19 (Ruoff, 1979; Field, 1991). Young's modulus and the bulk-modulus data compare rather favorably with the single-crystal values derived from Eqs. (20) (⟨110⟩ case for Young's modulus) and (21), respectively (see parenthesized data in Table 19), using the elastic compliance coefficients in Table 17. The decrease in K with increasing metallicity correlates with the similar trend in hardness observed in Table 6. Since the bulk modulus is an energy per unit volume and the energy varies inversely with the second power of the bond length d, K scales inversely with the fifth power of the bond length (Harrison, 1980).

Poisson's ratio varies considerably with direction from 0.01 to 0.2, 0.048 to 0.403,

Table 19. Experimental isotropic polycrystalline aggregate mechanical data compared to single-crystal values (in parentheses) for the group-IVa semiconductors at 300 K.

Element	Y^a (10^{12} dyn/cm^2)	$K^{b,c}$ (10^{12} dyn/cm^2)	G^d (10^{12} dyn/cm^2)	ν^e
Diamond	11.41 (11.64)	4.42 (4.42)	5.53 (5.78)	0.07 (0.104)
Silicon	1.629 (1.708)	0.979 (0.990)	0.666 (0.800)	0.223 (0.279)
Germanium	1.316 (1.398)	0.76 (0.766)	0.545 (0.680)	0.207 (0.273)
Gray tin	— (0.729)	— (0.426)	— (0.362)	— (0.298)

[a] ⟨110⟩ case.
[b] Additional bulk modulus data in (Harrison, 1980).
[c] Computed from Eq. (21).
[d] Computed from Eq. (23).
[e] Computed from Eq. (24).

and 0.022 to 0.403 for diamond (Field, 1991), silicon (Wortman and Evans, 1965), and germanium (Wortman and Evans, 1965), respectively. It usually has a value slightly greater than 0.25 and is 0.29 for any direction in a {111} plane for silicon (Bullis and Huff, 1994). Poisson's ratio for germanium has been reported to be 0.278 (McSkimin, 1953). The nonmonotonic behavior of G and ν in Table 19 may be due to the method of averaging (Ruoff, 1979; Field, 1991). The often poor agreement for the experimental G and ν values in Table 19, in comparison with the calculations using Eqs. (23) and (24), respectively, may arise because the equations are applicable for specific crystallographic directions, rather than for an isotropic polycrystalline aggregate.

As noted earlier, the elastic stiffness coefficients are determined by the interatomic forces in the crystal. A measure of the elastic anisotropy in group-IVa semiconductors may be assessed by determining the deviation from unity of the expression

$$s = (c_{11} - c_{12})/2c_{44}. \qquad (26)$$

Values of s for diamond, β-SiC, silicon, germanium, and gray tin are 0.824, 0.455, 0.639, 0.601, and 0.548, respectively (Harrison, 1980). The elastic stiffness coefficients may be combined to yield the radial and angular rigidity force constants, C_0 and C_1, respectively (Harrison, 1980):

$$C_0 = (3a_0^3/16)(c_{11} + 2c_{12}), \qquad (27)$$

$$C_1 = (a_0^3/32)(c_{11} - c_{12}). \qquad (28)$$

The degree to which the group-IVa semiconductors are not described by central forces (i.e., ionic bonds), but rather by highly directional covalency forces, may be determined by examining the deviation of c_{11} from c_{44} (i.e., the Cauchy relation) (Harrison, 1980). The crystal is "softened" by those distortions that can be accommodated by the hybridized bonds; this "softening" is reflected in a lower value of c_{44}. The c_{44} elastic stiffness coefficient can be expressed in terms of the force constants C_0 and C_1 (Harrison, 1980):

$$c_{44} = (32/a_0^3)(C_0C_1/[C_0 + 8C_1]). \qquad (29)$$

The three independent elastic stiffness constants reflect the three important physical effects summarized in the force constants C_0 and C_1 and the elastic stiffness coefficient c_{44}. Table 20 illustrates that the radial and angular rigidity force constants C_0 and C_1 and the Cauchy relation decrease with increasing metallicity as expected, using the experimental elastic stiffness coefficients from Table 16. Inclusion of the a_0^3 term in Eqs. (27) and (28) in Table 20 does not change this trend. The macroscopic force constants C_0 and C_1 derived from the experimental elastic stiffness coefficients can be compared with those derived from the characteristic vibrational frequencies (see Sec. 3.4). The speed of propagation of sound waves can be determined from the elastic stiffness coefficients and the material density

Table 20. Several elastic stiffness characteristics related to group-IVa, β-silicon carbide, and silicon–germanium alloy and superlattice semiconductors at 300 K.

Element	Metallicity parameter (α_m)	$(c_{11} + 2c_{12})$ (10^{12} dyn/cm^2)	$(c_{11} - c_{12})$ (10^{12} dyn/cm^2)	$(c_{11} - c_{44})$ (10^{12} dyn/cm^2)
Diamond	0.34	13.26	9.51	4.99
β-Silicon carbide	0.45	6.33	2.12	1.19
Silicon	0.66	2.98	1.03	0.87
SiGe (5084 Å)[a]	—	2.65	0.88	0.74
SiGe (6795 Å)[b]	—	2.56	0.87	0.75
$Si_{0.49}Ge_{0.51}$[c]	—	2.40	0.87	0.71
Germanium	0.81	2.30	0.82	0.63
Gray tin	0.87	—	0.44[d]	—

---

[a] Material density is 2.79 g/cm^3; the Å data refer to the superlattice thickness (Mendik *et al.*, 1991).
[b] Material density is 3.27 g/cm^3; the Å data refer to the superlattice thickness (Mendik *et al.*, 1991).
[c] Material density is 3.85 g/cm^3 (Mendik *et al.*, 1991).
[d] Theoretical data (Harrison, 1980).

(Boer, 1992). Finally, third-order elastic stiffness coefficients reflecting nonlinear effects such as thermal expansion and the interaction of phonons have also been summarized (Boer, 1992).

The theoretical fracture strength, σ_{th}, is expressed by (Davies, 1994)

$$\sigma_{th} = (Y\gamma/d)^{1/2}. \tag{30}$$

Utilizing the experimental $Y\{111\}$ data from Table 18, the $\gamma\{111\}$ data from Table 7, and the bond-length d data from Table 2, the values of σ_{th} decrease with increasing metallicity from diamond, silicon, and germanium in the order 2.01×10^{12}, 0.34×10^{12}, and 0.26×10^{12} dyn/cm^2, respectively. The theoretical fracture strength is approximately 0.2 the Young's modulus and approximately 0.4 the bulk modulus. Microscopic cracks and other material defects, however, reduce the fracture strength in real materials. Accordingly, the plastic deformation of semiconductors during device and IC fabrication is extremely structure sensitive. Extensive investigations on both semiconductor material preparation and the retention of the improved material microstructure during device and IC fabrication is an extremely active area of research.

3.4 Optical Properties

The coefficients of the dielectric constant ϵ_{ij}, a second-rank tensor, relate the electric field component E_j to the displacement component D_i in the ith direction. The dielectric constant tensor ϵ_{ij} can be transformed to its principal axes and, therefore, principal dielectric constant coefficients ϵ_1, ϵ_2, and ϵ_3, which are equal (isotropic) for the diamond crystal structure. The refractive index n, the ratio of the speed of light in a vacuum to the speed of light in a material, is related to the dielectric constant ϵ, at the same frequency, from Maxwell's equations:

$$n = (\epsilon\mu_m)^{1/2}, \tag{31}$$

where μ_m is the magnetic permeability. The extinction coefficient is sufficiently small in the long-wavelength (static) regime that the imaginary part of both the dielectric constant and the index of refraction can be neglected. The magnetic permeability may be taken as unity since the magnetic susceptibility χ is very small for the group-IVa semiconductors ($\mu_m = 1 + \chi$). The experimental data for ϵ, n, and χ are summarized in Table 21. The trend in the properties is consistent with previous material trends for the group-IVa semiconductors. For example, the increase in the index of refraction with increasing metallicity and decreasing Debye temperature (see Table 14) indicates that the semiconductor is becoming "less stiff" to the excitation and propagation of vibrational modes in the crystal. The optical excitation of electrons from the valence band to higher-order conduction-band minima, noted in Sec. 3.1.1, can be qualitatively described by the Phillips ionicity number (Phillips, 1973). The theoretical values of the Phillips ionicity are also summarized in Table 21.

Phonons in the group-IVa semiconductors facilitate indirect energy-gap electronic transitions via absorption or emission as appropriate. The phonons are required in order to conserve crystal momentum during optical transitions. The acoustic modes (i.e., sound waves) exhibit zero energy (other than zero-point motion) at zero wave number. Optical modes, however, can be excited into oscilla-

Table 21. Experimental data for the static dielectric constant, index of refraction, magnetic susceptibility, and Phillips ionicity number for the group-IVa and β-silicon carbide semiconductors.

Element	Static dielectric constant (ϵ)	Index of refraction (n)	Magnetic susceptibility (χ) (10^{-6}) (cgs)	Phillips ionicity number
Diamond	5.7	2.42	−5.9	2.50
β-Silicon carbide	9.72	3.12	−12.8	—
Silicon	11.9	3.42	−3.9	1.41
Germanium	16.0	4.00	−0.12	1.35
Gray tin	24.0	4.90[a]	−37.0	1.15

---

[a] Theoretical value (Harrison, 1980).

tory motion, even at zero wave number. Both acoustic and optical modes of vibration exhibit longitudinal and transverse motion. The transverse acoustical-mode frequencies at the Brillouin-zone boundary and the transverse optical-mode frequencies at $k = 0$ are summarized in Table 22 (Harrison, 1980); the reduction in the frequency of the transverse acoustical mode with increasing metallicity is observed. The force constants derived from these normal modes of vibration are not in agreement with the values derived from the elastic stiffness coefficients in Sec. 3.3, indicating the limitation of current models. An extensive listing of a number of phonon modes at crystal symmetry points has been summarized (Boer, 1992).

Finally, many semiconductors such as silicon are optically almost indistinguishable from metals because of their high reflectivity. For example, approximately 30% of the visible to far-infrared incident electromagnetic energy radiation is lost by reflection at each air–silicon interface (Smith, 1961; Bullis and Huff, 1994). The transmission of energy in silicon can be increased by the use of an antireflecting (AR) coating with a thickness equal to 1/4 of the incident radiation's wavelength and an index of refraction equal to $(n_{\mathrm{Si}}n_{\mathrm{air}})^{1/2}$ ($n_{\mathrm{air}} = 1$) (Bullis and Huff, 1994). In the visible region of the spectrum, semiconductors, like metals, generally absorb strongly, having absorption coefficients approximately 10^4–10^5 cm^{-1} (Smith, 1961). Over the range 700 to 1040 nm, the absorption coefficient of silicon, α_{Si}, is given by the empirical relationship (Nartowitz and Goodman, 1985)

$$\alpha_{\mathrm{Si}} = (8.4732 \times 10^{-3}\,\nu - 76.417)^2, \qquad (32)$$

where α_{Si} is in cm^{-1} and ν is the wave number in cm^{-1}. A significant increase in absorption occurs below about 400 nm due to the onset of direct transitions from the valence band to the higher-energy conduction band at $k = 0$.

Table 22. Transverse acoustical mode frequencies at the Brillouin zone and transverse optical mode frequencies at zero wave number.

Element	Transverse acoustic mode frequency (Brillouin zone) (10^{13}/s)	Transverse optical mode frequency ($k = 0$) (10^{13}/s)
Diamond	2.42	0.392
Silicon	0.449	0.155
Germanium	0.0242	0.899
Gray tin	0.0126	0.060

On the other hand, pure diamond is a good "window" for all wavelengths longer than 227 nm. Diamond is unique in that it exhibits almost no optical absorption from extremely low frequencies, approaching dc, to energies just below the energy gap of 5.45 eV (227 nm) (see DIAMOND AND DIAMONDLIKE CARBON).

4. SILICON–GERMANIUM

There are basically two classes of transistors: bipolar, in which both electrons and holes play an active role, and unipolar, such as MOS devices, based on the modulation of the conductance of a (surface) layer. The limitations of silicon electronic devices have warranted the development of compound semiconductors such as gallium arsenide for specialized, mainly high-frequency, electronic applications. The limited device integration possible with compound semiconductors, however, and the complexity of processing compound semiconductors have driven the extension of silicon technology by forming alloys of silicon with germanium (Kasper and Bean, 1988; Madelung, 1991; Meyerson, 1992). The silicon energy gap ($E_g = 1.12$ eV) at the X symmetry point can be monotonically decreased with the addition of germanium ($E_g = 0.66$ eV). The flexibility of tuning the energy gap in silicon–germanium alloys can improve the performance of specialized, high-performance electronic devices.

In some cases, the germanium content is changed abruptly at a certain distance from the surface of the semiconductor to form an abrupt "heterojunction," with a discontinuity in the energy gap. Because silicon and germanium have similar electron affinities, the energy required to remove an electron from the bottom of the conduction band to outside the solid, of 4.05 and 4.00 eV, respectively, the conduction-band edge is nearly continuous at the silicon–germanium boundary, while the valence-band edge has a discontinuity large enough to modify device performance (Kamins *et al.*, 1989; Kasper *et* *al.*, 1993). This valence-band discontinuity is especially useful in bipolar transistors. The gain of an *n-p-n* bipolar transistor is related to the ratio of the electron current injected across the emitter–base junction to the sum of the electron and hole currents crossing that junction. The valence-band discontinuity creates a barrier that decreases the injection of holes from the base into the emitter, thereby increasing the gain of the transistor. A further approach to improving bipolar transistor performance is accomplished by increasing the germanium content gradually with position in the base layer (Patton *et al.*, 1989; Cressler *et al.*, 1994). In this case, the energy gap varies as a function of position across the base region of the transistor. The varying energy gap creates an internal electric field that accelerates the free carriers (i.e., electrons in the case of an *n-p-n* transistor) across the base region, thereby increasing the maximum frequency response f_t of the bipolar transistor toward 100 Ghz.

A *p*-channel silicon MOS transistor has a small hole mobility (see Table 9). The hole mobility in silicon–germanium alloys, however, can be greater than in pure silicon (Bean, 1988). The increased carrier mobility enhances the transconductance of the *p*-channel MOS transistor. Furthermore, in silicon–germanium *p*-channel MOS transistors (Garone *et al.*, 1991), a thin layer of silicon–germanium, slightly separated from the oxide–semiconductor interface by a silicon layer, can spatially confine holes. These confined holes travel in the silicon–germanium layer and, accordingly, their velocity is not reduced by scattering from the oxide–semiconductor interface.

Optimal electrical properties are achieved by ensuring that the crystal bonding is uninterrupted across the silicon/silicon–germanium interface. Table 2, however, indicates that the lattice constant of germanium is about 4% greater than that of silicon. The lattice mismatch increases with increasing germanium content in silicon–germanium alloys. Modern growth techniques using CVD can form a thin layer of silicon–germanium alloy on silicon with a continuous crystal structure across the interface (i.e., the crystal structure is "commensurate"). The alloy layer, however, is compressed in the plane of the film. With increasing layer thickness, the energy associated with compressing the alloy layer to fit the silicon crystal increases. Above a "critical" thickness (Matthews and Blakeslee, 1974), the strain energy exceeds the energy necessary to form misfit dislocations, discussed in Sec. 2.2.2. As the bonds across the interface break to form the misfit dislocations, the strain in the system is reduced, thereby lowering the total energy. Accordingly, although silicon–germanium layers thicker than the "equilibrium" critical thickness can be formed, they are unstable (Matthews and Blakeslee, 1974). When these "metastable" layers are heated above the stability temperature appropriate to the layer thickness and germanium content, any irregularity already present can nucleate a misfit dislocation, taking into account kinetic considerations. The dislocation then propagates along the interface, creating a line of broken bonds. While misfit dislocations at the interface may be detrimental to device performance, interactions between misfit dislocations may more seriously degrade the electrical performance of devices built in the layer. The limited stable thickness of silicon–germanium, therefore, constrains the application of silicon–germanium layers. Since very high-performance devices require thin layers ($\leq$100 nm) to limit carrier transit times, however, this restriction is not serious. Thin silicon–germanium layers are indeed compatible with device requirements.

The strain in silicon–germanium layers also exhibits the beneficial effect of increasing the energy-gap difference relative to pure silicon (People, 1985). For example, while an unstrained silicon–germanium layer containing 20% germanium has an energy gap about 0.09 eV smaller than that of silicon, the energy gap of a strained layer with the same germanium content is about 0.17 eV smaller than that of silicon. The required electrical performance can, therefore, be obtained with a smaller germanium content in a strained, rather than an unstrained, layer.

Carrier mobilities in silicon–germanium layers are expected to differ from the silicon values for several reasons. First, carrier mobilities in germanium are greater than for silicon (see Table 9). Second, the random arrangement of germanium in the silicon crystal structure leads to a reduction in mobility due to an additional scattering mechanism, referred to as "alloy" scattering. Third, the strain in the silicon–germanium layer modifies the curvature of the energy bands, changing the carrier effective mass and, thereby, the carrier mobility (Smith, 1961). Finally, carrier lifetime is strongly reduced by impurities such as oxygen in the film. Since oxygen is more readily incorporated into silicon–germanium alloys than in silicon, a higher-purity ambient is required when forming silicon–germanium alloys compared to elemental silicon films.

While most applications of silicon–germanium focus on high-performance transistors, optical devices may also become important (Bean, 1988). Since the energy gap of silicon–germanium alloys is smaller than for silicon, the optical absorption edge corresponds to a longer wavelength (i.e., smaller energy), and silicon–germanium is useful as a photodetector at longer wavelengths compared to silicon.

5. β-SILICON CARBIDE

Silicon carbide can crystallize in the diamond crystal structure, where the silicon and carbon atoms are arranged at the (0,0,0) and $(\frac{1}{4},\frac{1}{4},\frac{1}{4})$ basis vector positions, respectively (or vice versa). This configuration is referred to as β-SiC. Silicon carbide can also occur in a variety of modifications that can be described as different stacking sequences of the closest-packed {111} layers (see Sec. 2.3.1). Since these polymorphs only differ in their stacking sequences, the term polytypes instead of polymorphs is used to describe them (Azaroff, 1960). These structures occur in the hexagonal and rhombohedral crystal structures and are collectively referred to as α-SiC. The most commonly observed polytypes are referred to as 4*H*, 6*H*, and 15*R*, where the numerical coefficient designates the number of double layers required to replicate the crystal structure. The letter *H* denotes that the hexagonal cell is primitive, and the letter *R* signifies a rhombohedral cell (i.e., a hexagonal cell that is centered). Additional structures such as 21*R*, 33*R*, and 51*R* are also observed. The reason only particular polytypes are observed is not yet fully comprehended, but it appears to be the rule rather than the exception and is determined by the conditions during their growth (Tairova and Tsvetkov, 1983). It has been observed, however, that 6*H* is most often formed in the absence of impurities, whereas polytypes such as 51*R*, 141*R*, and others may occur when the melt contains a relatively large number of impurities. The role of impurities has previously been suggested to aid in the formation of screw dislocations, which facilitate the formation of complex polytypes (Azaroff, 1960) and spurious nucleation effects (Davis, 1994). This topic continues to be an important area of research, and the reader is referred to the literature for further details (Davis *et al.,* 1988; Pirouz *et al.,* 1991; Morkoc *et al.,* 1994; Pensl and Troffer, 1995).

The physicochemical properties of α- and β-SiC are different. In particular, α-SiC exhibits significantly different properties as a result of the effects of differing stacking sequences (Henisch and Roy, 1969). Several structural, electrical, thermal, mechanical, and optical properties of β-SiC have been summarized in earlier tables. α-SiC, however, especially 4*H*, has been reported to exhibit superior device performance compared to β-SiC (Spencer, 1993; Itoh *et al.,* 1994) because of its high electron mobility, low anisotropy of the electron effective mass, small nitrogen donor electrical ionization energy (Pensl and Troffer, 1995), and low micropipe-density epitaxial layers. Nevertheless, the usefulness of β-SiC as a semiconducting device material (Davis *et al.,* 1990) is briefly discussed in Sec. 6 because of the emphasis in this review on semiconductor material configurations with the diamond crystal structure.

6. DEVICE/INTEGRATED-CIRCUIT IMPLICATIONS

Silicon rapidly replaced germanium in device technology on account of its larger energy gap, which allowed higher-temperature operation and lower reverse current. In addition, silicon's oxide is stable, whereas germanium's oxide is water soluble. Accordingly, silicon was favored for IC fabrication since the stable oxide facilitated its utilization in the planar process as

1. a diffusion mask for *p-n* junction fabrication,
2. passivation of the silicon surface and *p-n* junctions intersecting the surface, and
3. a dielectric layer for supporting and sep-

Table 23. Material and device characteristics utilized in figures of merit.

Material	κ (W/cm K)	E_{bd} (V/cm)	V_{sat} (cm/s)	ϵ
Diamond	20	1.0×10^7	2.7×10^7	5.7
β-Silicon carbide	5.0	4.0×10^6	2.5×10^7	9.72
Silicon	1.15	3.0×10^5	1.0×10^7	11.9
Gallium arsenide	0.46	3.5×10^7	2.0×10^7	13.1

arating metallic conductor overlayers (Adcock, 1969; Kooi, 1991; Huff, 1996).

Alternative materials to silicon that might exhibit "better" device parameters have, of course, been proposed. These materials include gallium arsenide, silicon carbide, and diamond (Collins, 1990; Davis *et al.*, 1990). Several important material and device characteristics utilized in figures of merit include the thermal conductivity κ, the breakdown electric field E_{bd}, and the saturation (limiting) electron velocity v_{sat}. These characteristics, as well as the dielectric constant ϵ, are summarized in Table 23.

The square of the product of breakdown electric field E_{bd} and the saturation (limiting) electron velocity v_{sat} has been proposed as a measure of the power handling capability (V/s) of a transistor (Johnson, 1965). This figure of merit is

$$J = (E_{bd} v_{sat}/2\pi)^2. \qquad (33)$$

An additional figure of merit relates to the thermally limited switching speed of transistors (Keyes, 1972) in ICs for computer logic applications. The ability of a semiconductor to dissipate the thermal energy generated by the transport of electrons or holes from the source to drain during operation of the MOS transistor is critically important for dense MOS circuitry. This figure of merit is

$$K = \kappa(v_{sat}/\epsilon)^{1/2}, \qquad (34)$$

where a smaller dielectric constant ϵ (i.e., reduced amount of stored charge) results in a faster switching device. These figures of merit are summarized in Table 24. Diamond exhibits the best characteristics, followed by β-SiC (Davis *et al.*, 1990). Gallium arsenide exhibits a poorer thermally limited switching characteristic compared with silicon, on account of its smaller thermal conductivity (Meindl, 1987). Although these figures of merit are useful for comparison, they are theoretical values that assume that the semiconductor material is being operated at its physical limits (Collins, 1990). In practice, carrier mobilities (see Table 9) and, especially, device geometry appear to be more influential on performance than might be inferred from figure-of-merit values (Collins, 1990). Diamond's high thermal conductivity makes it especially useful as a heat sink material to mitigate excessive temperature gradients during the operation of dense IC components.

The potential benefit of gallium arsenide due to its electron mobility of 8500 cm^2/V s has been extensively discussed. This benefit, however, may not be realized when one considers not only the thermally limited switching time as in Table 24 but ULSI system-level limitations such as interconnection resistance (Meindl, 1987). It has also been suggested that the only diamond device that appears to offer some promise is an enhancement-mode *p*-type MOSFET operating at room temperature (Collins, 1990). The challenge for the future is to fabricate such a device with submicron dimensions necessary for operation at microwave frequencies (Collins, 1990). β- and α-SiC will also continue to

Table 24. Summary of several device figures of merit.

Material	J (V/s)^2	Ratio to Si	K (W/K cm$^{1/2}$ s$^{1/2}$)	Ratio to Si
Diamond	1.85×10^{27}	8103	4.35×10^4	31.52
β-Silicon carbide	2.53×10^{26}	1107	8.03×10^3	5.82
Gallium arsenide	1.24×10^{24}	5.44	5.68×10^2	0.41
Silicon	2.28×10^{23}	1.0	1.38×10^3	1.0

draw attention. Nevertheless, crystalline silicon and silicon-based materials, such as silicon–germanium, are expected to continue, through sophisticated device design and processing technologies, to be the main engine driving the microelectronics revolution to the approximate regime of 50-nm design rules and, perhaps, beyond (Huff, 1993; Semiconductor Industry Association, 1994; Huff and Goodall, 1996). Improved control of the magnitude, tolerance, and uniformity of silicon material characteristics, coupled with an improved understanding of the interrelationships among IC performance, device parameters, IC fabrication processes, and silicon material characteristics, continue to be the key to superior IC product performance, reliability, and yield. The silicon material trends anticipated to ensure the required performance characteristics for gigabit-memory ICs are summarized in the National Technology Roadmap for Semiconductors (Semiconductor Industry Association, 1994).

7. CONCLUSIONS

The group-IVa semiconductors form the prototypical sp^3-hybridized, tetrahedrally bonded, purely covalent bonds. The bonds are highly directional in the diamond crystal structure, which is conveniently described as two interpenetrating face-centered cubic crystals displaced along the ⟨111⟩ direction by one-fourth the diagonal length. The diamond crystal structure was seen to exhibit the smallest packing fraction in the cubic crystal class. This large open space can accommodate a number of interstitial impurities and accounts for the extreme structure sensitivity of diverse material properties such as the yield stress and carrier lifetime.

The fundamental parameter distinguishing the group-IVa semiconductors was seen to be the metallicity parameter, the parameter *sine qua non* by which the distinct variations in the crystalline semiconductor characteristics may be correlated. The variation of the energy gap, lattice constant, Debye temperature, hardness, atomic density, bond energy, radial and angular force constants, and dielectric constant, for example, correlate with the metallicity parameter. β-SiC generally follows the correlation with the metallicity parameter. Although silicon–germanium also follows some of these trends, alloy effects complicate the analysis. The electrical conductivity, thermal conductivity, and dielectric constant are each second-rank tensors that can be reduced to a single number and, therefore, the resulting properties are isotropic. The elastic stiffness coefficients, however, form a fourth-rank tensor that can be reduced to three principal (diagonal) coefficients; therefore, most of the elastic properties of the group-IVa semiconductors were seen to be anisotropic.

Silicon–germanium alloys and β-SiC are important semiconducting material combinations derived from the group-IVa elements. Semiconductor properties and device considerations for these materials as well as diamond were briefly examined. Nevertheless, crystalline silicon and silicon-based materials, such as silicon–germanium, are expected to continue, through sophisticated device design and processing technologies, to be the main engine driving the microelectronics revolution to the approximate regime of 50-nm design rules and, perhaps, beyond.

Crystalline silicon and silicon-based materials will continue to be the most intensively studied material systems in the scientific literature. The opportunities for further silicon-based advancements appear boundless, certainly quite sufficient to support numerous research and development thrusts well into the 21st century (Huff and Goodall, 1996b). In conjunction with IC process/device/circuit innovations for cost-effective production of ICs (Ohmi, 1995; Koike, 1995), electronics will offer unsurpassed opportunities for an improved quality of life for the world's citizens.

ACKNOWLEDGMENTS

The critical reading of this manuscript by Dr. W. M. Bullis, Dr. R. K. Goodall, Dr. E. Kooi, Dr. B. S. Roos, Dr. W. R. Runyan, Prof. H. Berk, Prof. U. Gösele, Prof. W. A. Harrison, Prof. R. Helms, Prof. L. C. Kimerling, Prof. J. Meindl, Prof. F. H. Pollack, and Prof. E. Weber as well as selected portions by Dr. J. Hersener, Dr. G. Pensl, Dr. E. Wijaranakula, Prof. A. L. Ruoff, and Prof. E. A. Fitzgerald is appreciated. Dr. A. Burk supplied a number of references to β-SiC material characteristics. Dr. T. Kamins kindly contributed the section on silicon–germanium, and Mr. V. Bhat replicated several figures for the manuscript. Finally, I thank Linda Ofshe for compiling and formatting the manuscript.

GLOSSARY

Conduction Band: The antibonding states or band, where the electrons "belong" to the whole crystal.

Crystal: A periodic three-dimensional structure.

Defect: A deviation from perfect periodicity of the crystalline structure.

Diamond Crystal Structure: The group-IVa semiconductor crystal structure, often described as two interpenetrating face-centered cubic crystals displaced along the ⟨111⟩ direction by one-fourth the diagonal length.

Dopant: A group-Va (group-IIIa) element that resides on a substitutional site, introducing an electron (hole) into the conduction (valence) band.

Effective Distribution Coefficient: Ratio of the impurity concentration incorporated into a growing crystal at the crystal-melt interface to the impurity concentration in the bulk of the liquid.

Energy Gap: The minimum energy separating the valence-band maximum from the conduction-band minimum. When the energy extrema are located at different values in momentum (**k**) space, the energy gap is denoted as indirect.

Interstitial Atom: An atom residing on a crystal site other than a substitutional site.

Line Defect: A line imperfection in the crystal, such as a dislocation, that extends over many atomic spacings.

Metallicity Parameter: The ratio of the metallic energy to the covalent energy for a semiconductor.

Point Defect: A defect that extends only over a few atomic spacings such as a vacancy, interstitial atom, oxygen atom, or carbon atom.

Semiconductor: A material that exhibits a reduction in resistivity with increasing temperature over selected regions of temperature.

***sp*^{3} hybridization:** Interaction between *s* and *p* valence electrons to form bonding and antibonding states or bands in the diamond crystal structure.

Substitutional Site: A crystal site where the atoms of the crystal normally reside.

Vacancy: A crystal site where a substitutional atom has been removed.

Valence Band: The bonding states or band, where the electrons "belong" to the whole crystal, but statistically are spatially localized between the atoms.

Works Cited

Adamson, A. W. (1990), *Physical Chemistry of Surfaces,* 5th ed. New York: Wiley.

Adcock, W. (1969), in: R. R. Haberecht, E. L. Kern, (Eds.), *Semiconductor Silicon,* Pennington, NJ: The Electrochemical Society, pp. 36–54.

Alexander, H., Haasen, P. (1968), in: F. Seitz, D. Turnbul, H. Ehrenreich (Eds.), *Solid State Physics,* Vol. 22, New York: Academic, pp. 27–158.

Allen, J. W. (1957), *Philos. Mag.* **2,** 1475–1481.

Anderson, H. L. (Ed.) (1989), *A Physicist's Desk Reference,* 2nd ed., New York: American Institute of Physics.

Asoka-Kumar, P., Lynn, K. G., Welch, D. O. (1994), *J. Appl. Phys.* **76,** 4935–4982.

Azaroff, L. V. (1960), *Introduction to Solids,* New York: McGraw-Hill.

Baker, S. P., Arzt, E. (1995), in: E. Kasper (Ed.), *Properties of Strained and Relaxed Silicon Germanium,* Exeter, U.K.: INSPEC, pp. 67–69.

Beadle, W. E., Tsai, J. C. C., Plummer, R. D. (1985), *Quick Reference Manual for Silicon Integrated Technology,* New York: Wiley.

Bean, J. C. (1988), in: E. Kasper, J. C. Bean (Eds.), *Silicon-Molecular Beam Epitaxy,* Vol. 2, Pennington, NJ: The Electrochemical Society, pp. 65–110.

Bean, K., Lindberg, K., Rozgonyi, G. (1988), in: M. Scott, Y. Akasaka, R. Reif (Eds.), *Epitaxy in Advanced Materials for ULSI,* Pennington, NJ: The Electrochemical Society, pp. 3–24.

Boer, K. W. (1992), *Survey of Semiconductor Physics,* New York: Van Nostrand Reinhold.

Bond, W. L. (1960), *Acta Crystallogr.* **13,** 814–818.

Brown, R. A., Maroudas, D., Sinno, T. (1994), *J. Cryst. Growth* **137,** 12–25.

Bullis, W. M. (1994), in: H. R. Huff, W. Bergholz, K. Sumino (Eds.) *Semiconductor Silicon/94,* Pennington, NJ: The Electrochemical Society, pp. 1156–1169.

Bullis, W. M., Huff, H. R. (1994), in: D. Bloor, R. J. Brook, M. C. Flemings, S. Mahajan (Eds.), *The Encyclopedia of Advanced Materials,* London: Pergamon, pp. 2478–2496.

Bullis, W. M., Huff, H. R. (1996), *J. Electrochem. Soc.* **143,** 1399–1405.

Burns, G. (1985), *Solid State Physics,* New York: Academic.

Burton, J. A., Prim, R. C., Slichter, W. P. (1953), *J. Chem. Phys.* **21,** 1987–1991.

Chelikowsky, J. R., Cohen, M. L. (1976), *Phys. Rev. B* **14,** 556–581.

Chen, A. B., Sher, A., Yost, W. T. (1992), in: K. T. Faber, K. Malloy (Eds.) *The Mechanical Properties of Semiconductors,* New York: Academic, pp. 1–77.

Chikawa, J., Shirai, S. (1979), *Japan J. Appl. Phys.* **18** (Suppl. 18-1), 153–164.

Clarke, D. R. (1992), in: K. T. Faber, K. Malloy (Eds.), *The Mechanical Properties of Semiconductors,* New York: Academic, pp. 79–142.

Collins, A. T. (1990), in: J. T. Glass, R. Messier, N. Fujimori (Eds.), *Diamond, Silicon Carbide and Related Wide Bandgap Semiconductors,* MRS Symposium Proceedings Vol. 162, Pittsburgh, PA: Materials Research Society, pp. 3–14.

Cottrell, A. H. (1961), *Dislocations and Plastic Flow in Crystals,* Oxford: Oxford University Press.

Cressler, J. D., Harame, D., Comfort, J. H. (1994), in: *IEEE International Solid State Circuits Conference,* New York: IEEE, pp. 24–27.

Dash, W. C. (1956), *J. Appl. Phys.* **27,** 1193–1195.

Dash, W. C. (1960), in: H. C. Gatos, (Ed.), *Properties of Elemental and Compound Semiconductors,* Metallurgical Society Conferences, Vol. 5, New York: Interscience, pp. 195–209.

Davies, G. (Ed.) (1994), *Properties and Growth of Diamond,* EMIS Datareviews Series, London: INSPEC, p. 9.

Davis, R. F. (1994), in: D. Bloor, R. J. Brook, M. C. Flemings, S. Mahajan (Eds.), *The Encyclopedia of Advanced Materials,* London: Pergamon, pp. 2465–2469.

Davis, R. F., Palmour, J. W., Edmond, J. A. (1990), in: J. T. Glass, R. Messier, N. Fujimori (Eds.), *Diamond, Silicon Carbide and Related Wide Bandgap Semiconductors,* MRS Symposium Proceedings Vol. 162, Pittsburgh: Materials Research Society, 463–474.

Davis, R. F., Sitar, Z., Williams, B. E., Kong, H. S., Kim, H. J., Palmour, J. W., Edmond, J. A., Ryu, J., Glass, J. T., Carter, C. H., Jr. (1988), *Materials Science and Engineering, B1,* Lausanne: Elsevier, pp. 77–104.

De Launay, J. (1956), in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics,* Vol. 2, New York: Academic, pp. 219–303.

Diebold, A. C., Maillot, P., Gordon, M., Baylis, J., Chacon, J., Witowski, R., Arlinghaus, H. F., Knapp, J. A., Doyle, B. L. (1992), *J. Vac. Sci. Technol. A.* **10,** 2945–2952.

Dornberger, E., von Ammon, W. (1996), *J. Electrochem. Soc.* **143,** 1648–1653.

Dyer, L. D., Huff, H. R., Boyd, W. W. (1971), *J. Appl. Phys.* **42,** 5680–5688.

EMIS Datareviews (1988), *Properties of Silicon,* EMIS Datareviews Series Vol. 4, London: INSPEC, p. 4.

Field, J. E. (1991), in: R. E. Clausing, L. L. Horton, J. C. Angus, P. Koidl (Eds.), *Diamond and Diamond-Like Films and Coatings,* NATO Advanced Study Institutes Series B, Physics, Vol. 266, New York: Plenum, pp. 17–35.

Foll, H., Gösele, U., Kolbesen, B. O. (1977), in: H. R. Huff, E. Sirtl (Eds.) *Semiconductor Silicon/77,* Pennington, NJ: The Electrochemical Society, pp. 565–574.

Fukuda, T. (1994), *Appl. Phys. Lett.* **65,** 1376–1378.

Garone, P. M., Venkataraman, V., Sturm, J. C. (1991), *IEEE Electron Device Lett.* **12,** 230–232.

Gatos, H. C., Lavine, M. C. (1965), in: A. F. Gibson, R. E. Burgess (Eds.) *Progress in Semiconductors,* Vol. 9, New York: Wiley, pp. 3–45.

Gay, A. G. (1960), *Elements of Physical Metallurgy,* Reading, MA: Addison-Wesley.

Gösele, U., Tan, T. Y. (1991), *MRS Bull.* **16** (11), 42–46.

Green, M. A. (1990), *J. Appl. Phys.* **67,** 2944–2954.

Harrison, W. A. (1980), *Electronic Structure and The Properties of Solids,* San Francisco: W. H. Freeman.

Harrison, W. A. (1981), *Phys. Rev. B* **24,** 5835–5843.

Helms, R. (1996), in: W. M. Bullis, D. Seiller, A. C. Diebold (Eds.), *Semiconductor Characterization: Present Status and Future Needs,* New York: American Institute of Physics, pp. 110–117.

Henisch, K. H., Roy, R. (Eds.) (1969), *Silicon Carbide—1968,* New York: Pergamon.

Herman, F. (1995), *Proc. IRE* **43,** 1703–1732.

Hirth, J. P., Lothe, J. (1982), *Theory of Dislocations,* 2nd ed., New York: Wiley.

Hobstetter, J. N. (1960), in: Hannay, N. B. (Ed.), *Effect of Imperfections on Germanium and Silicon in Semiconductors,* New York: Reinhold, pp. 508–540.

Holwigle, N. A., Frank, W., Hobel, J., Pearton, P., Haller, E. E. (1985), *J. Appl. Phys.* **57,** 5211–5219.

Hornstra, J. (1958), *J. Phys. Chem. Solids* **5,** 129–141.

Hu, S. M. (1994), in: F. Shimura (Ed.), *Oxygen in Silicon,* New York: Academic, pp. 153–190.

Huff, H. R. (1992), in: S. Mahajan, L. C. Kimerling (Eds.), *Concise Encyclopedia of Semiconducting Materials & Related Technologies,* New York: Pergamon, pp. 478–492.

Huff, H. R. (1993), in: G. K. Celler, E. Middlesworth, K. Hoh (Eds.), *Proceedings of the Fourth International Symposium on ULSI Science and Technology,* Pennington, NJ: The Electrochemical Society, pp. 103–132.

Huff, H. R. (1996), *Electrochem. Soc. Extended Abstracts,* **96-1,** 341–343.

Huff, H. R., Goodall, R. K. (1995), in: R. B. Fair, B. Lojek (Eds.), *RTP '95,* Rapid Thermal Processing Conference, Round Rock, TX: RTP '95, pp. 9–40.

Huff, H. R., Goodall, R. K. (1996a), in: W. M. Bullis, D. Seiller, A. C. Diebold (Eds.), *Semiconductor Characterization: Present Status and Future Needs,* New York: American Institute of Physics, pp. 67–96.

Huff, H. R., Goodall, R. K. (1996b), *Interface* **5** (2), 31–35.

Huff, H. R., Shimura, F. (March, 1985), *Solid State Tech.* **28** (3), 103–118.

Huff, H. R., Schaake, H. F., Robinson, J. T., Baber, S. C., Wong, D. (1983), *J. Electrochem. Soc.* **130,** 1551–1555.

Huff, H. R., Goodall, R. K., Williams, E., Woo, K.-S., Liu, B. Y. H., Warner, T., Hirleman, D., Gildersleeve, K., Bullis, W. M., Scheer, B. W., Stover J. (1995), in: E. M. Middlesworth, H. Z. Massoud (Eds.), *Proceedings of the Fifth International Symposium on ULSI Science and Technology,* Pennington, NJ: The Electrochemical Society, pp. 365–383.

Irvin, J. C. (1962), *Bell Syst. Tech. J.* **41,** 387–412.

Itoh, A., Akita, H., Kimoto, T., Matsunami, H. (1994), *Appl. Phys. Lett.* **65,** 1400–1402.

Johnson, E. O. (1965), *RCA Rev.* **26,** 163–177.

Kamins, T. I., Nauka, K., Camnitz, L. H., Kruger, J. B., Turner, J. E., Rosner, S. J., Scott, M. P., Hoyt, J. L., King, C. A., Noble, D. B., Gibbons, J. F. (1989), *International Electron Devices Meeting, Washington, D.C.,* Piscataway, NJ: IEEE, pp. 647–650.

Kasper, E., Bean, J. C. (Eds.) (1988), *Silicon-Molecular Beam Epitaxy,* Pennington, NJ: The Electrochemical Society.

Kasper, E., Gruhle, A., Kibbel, H. (1993), *International Electron Devices Meeting, Washington, D.C.,* Piscataway, NJ: IEEE, pp. 79–81.

Kelly, A., Groves, G. W. (1970), *Crystallography and Crystal Defects,* Reading, MA: Addison-Wesley.

Kern, W. (Ed.) (1993), *Handbook of Semiconductor Wafer Cleaning Technology,* Park Ridge, NJ: Noyes Publications.

Keyes, R. W. (1972), *Proc. IEEE* **60,** 225.

Kimerling, L. C., Patel, J. R. (1985), in: N. G. Einspruch, H. R. Huff (Eds.), *VLSI Electronics: Microstructure Science,* Vol. 12, New York: Academic, pp. 223–267.

Kimerling, L. C., Leamy, H. J., Patel, J. R. (1977), *Appl. Phys. Lett.* **30,** 217–219.

Kimerling, L. C., Patel, J. R. (1979), *Appl. Phys. Lett.* **38,** 73–75.

King, C. A. (1994), in: D. Bloor, R. J. Brook, M. C. Flemings, S. Mahajan (Eds.), *The Encyclopedia of Advanced Materials,* London: Pergamon, pp. 2496–2500.

Kisielowski, C., Weber, E. R. (1991), *Phys. Rev. B* **44,** 1600–1612.

Kittel, C. (1996), *Introduction to Solid State Physics,* 7th ed., New York: Wiley.

Koike, A., Shimoyashura, S., Kubota, K., Suzuki, Y., Kiguchi, A., Fujisawa, A., Takamatsu, A., Okabe, T. (1995), in: *International Symposium on Semiconductor Manufacturing,* Piscataway, NJ: IEEE, pp. 239–242.

Kolbesen, B. O., Strunk, H. P. (1985), in: N. G. Einspruch, H. R. Huff (Eds.), *VLSI Electronics: Microstructure Science,* Vol. 12, New York: Academic, pp. 143–222.

Kooi, E. (1991), *The Invention of LOCOS, IEEE Case Histories of Achievement in Science and Technology,* Vol. 1, Parsipanny, NJ: IEEE Press.

Kubaschewski, O., Alcock, C. B. (1973), *Metallurgical Thermochemistry,* 5th ed., New York: Pergamon.

Landau, L. D., Lifshitz, E. M. (1958), *Quantum Mechanics, Non-Relativistic Theory,* New York: Pergamon.

Lawrence, J. E., Huff, H. R. (1982), in: N. G. Einspruch (Ed.), *VLSI Electronics: Microstructure Science,* Vol. 5, New York: Academic, pp. 51–102.

Lemke, H. (1994), in: H. R. Huff, W. Bergholz, K. Sumino (Eds.), *Semiconductor Silicon/1994,* Pennington, NJ: The Electrochemical Society, pp. 695–710B.

Lide, D. R. (Ed.) (1993–1994), *CRC Handbook of Chemistry and Physics,* 74th ed., Boca Raton, FL: CRC Press.

Ligenza, J. R. (1961), *J. Phys. Chem.* **65,** 2011–2014.

Lin, W., Hill, D. W. (1983), *J. Appl. Phys.* **54,** 1082–1085.

Madelung, O. (Ed.) (1991), *Data in Science and Technology, Semiconductors, Group IV Elements and III-V Compounds,* Berlin: Springer-Verlag.

Matthews, J. W., Blakeslee, A. E. (1974), *J. Cryst. Growth* **27,** 118–125.

McSkimin, H. J. (1953), *J. Appl. Phys.* **24,** 988–997.

McSkimin, H. J., Bond, W. L., Buckler, E., Teal, G. K. (1951), *Phys. Rev.* **83,** 1080.

Meindl, J. (1987), in: S. Broydo and C. M. Osburn (Eds.), *ULSI Science and Technology/1987,* Pennington, NJ: The Electrochemical Society, pp. 3–17.

Mendik, M., Ospelt, M., Von Kanel, H., Wachter, P. (1991), *Appl. Surface Sci.* **50,** 303–307.

Meyerson, B. S. (1992), *Proc. IEEE* **80,** 1592–1608.

Morkoc, H., Strite, S., Gao, G. B., Lin, M. E., Sverdlov, B., Burns, M. (1994), *J. Appl. Phys.* **76,** 1363–1398.

Nartowitz, E. S., Goodman, A. M. (1985), *J. Electrochem. Soc.* **132,** 2992–2997.

Nielsen, O. H. (1988), in: *Properties of Silicon,* EMIS Datareviews Series Vol. 4, London: INSPEC, p. 15.

Nye, J. F. (1964), *Physical Properties of Crystals,* Oxford: Oxford University Press.

Ohmi, T., Miyashita, M., Itano, M., Imaoka, T., Kawanabe, I. (1992), *IEEE Trans. Electron Devices* **39,** 537–545.

Ohmi, T. (1995), *Ultra Clean Soc. Proc.* **26,** 91.

Page, T. F., Oliver, W. C., McHargue, C. J. (1992), *J. Mater. Res.* **7,** 450–473.

Patton, G. L., Harame, D. L., Stork, J. M. C., Meyerson, B. S., Scilla, G. J., Ganin, E. (1989), *IEEE Electron Device Lett.* **10,** 534–536.

Pauling, L. (1960) *The Nature of the Chemical Bond,* Ithaca, NY: Cornell Univ. Press.

Pensl, G., Choyke, W. J. (1993), *Physica B* **185,** 264–283.

Pensl, G., Troffer, Th. (1995), in: H. Richter, M. Kittler, C. Claeys (Eds.), *GADEST '95,* Zurich: Scitec Publications, pp. 115–126.

People, R. (1985), *Phys. Rev. B* **32,** 1405–1408.

Pfann, W. G. (1966), *Zone Melting,* 2nd ed. New York: Wiley.

Phillips, J. C. (1973), *Bonds and Bands in Semiconductors,* New York: Academic.

Pirouz, P., Yang, J. W., Powell, J. A., Ernst, F. (1991), in: A. G. Cullis, A. R. Long (Eds.), *Microscopy of Semiconducting Materials 1991,* IOP Conference Proceedings No. 117, Bristol, U.K.: IOP Publishing, Sec. 3, 149–154.

Pollak, F. H., Cardona, M., Higginbotham, C. W., Herman, F., Van Dyke, J. P. (1970), *Phys. Rev. B* **2,** 352–363.

Rajagopalan, S., Mitra, U., Pan, S., Gupta, K., Lin, C. M., Sery, G., Mittal, S., Hasserjian, K., Lo, W. J., Neubauer, G. (1993), in: *31st Annual Proceedings of Reliability Physics/1993,* New York: IEEE, pp. 28–31.

Ravi, K. V. (1981), *Imperfections and Impurities in Semiconductor Silicon,* New York: Wiley.

Ridley, B. K. (1982), *Quantum Processes in Semiconductors,* Oxford: Oxford University Press.

Riney, T. D. (1961), *J. Appl. Phys.* **32,** 454–460.

Roos, B. S., Richter, H., Wollweber, J. (1995a), in: H. Richter, M. Kittler, C. Clayes (Eds.), *GADEST '95,* Zurich: Scitec Publications, pp. 509–515.

Roos, B. S., Richter, H., Morgenstern, Tillack, B. (1995b) in: S. P. Baker, P. Borgesen, B. H. Townsend, C. A. Ross, C. A. Volkert (Eds.), *Thin Films: Stresses and Mechanical Properties V,* MRS Symposium Proceedings Vol. 356, Pittsburgh, PA: Materials Research Society, pp. 277–282.

Rossi, J. A., Dyson, W., Hellwig, L. G., Haley, T. M. (1985), *J. Appl. Phys.* **58,** 1798–1802.

Runyan, W. R. (1965), *Silicon Semiconductor Technology,* New York: McGraw-Hill.

Runyan, W. R., Huff, H. R. (1994), in: J. J. McKetta (Ed.), *Encyclopedia of Chemical Processing and Design,* Vol. 49, New York: Marcel Dekker, pp. 340–365.

Ruoff, A. L. (1979), in: K. D. Timmerhaus, M. S. Barber (Eds.), *High Pressure Sci. Technol.* **2,** 525–548.

Sauer, R., Weber, J., Stolz, J., Weber, E. R., Kusters, K.-H., Alexander, H. (1985), *Appl. Phys. A* **36,** 1–13.

Schlichting, J., Riley, F. L. (1992), in: S. Mahajan, L. C. Kimerling (Eds.), *Concise Encyclopedia of Semiconducting Materials & Related Technologies,* London: Pergamon, pp. 451–455.

Schroder, D. K. (1990), *Semiconductor Material and Device Characterization,* New York: Wiley.

Schulz, H. (1982), in: E. Kaldis (Ed.), *Current Topics in Materials Science,* Vol. 8, Amsterdam: North-Holland, pp. 276–379.

Seeger, A. (1976), in: J. Treusch (Ed.) *Festkörperprobleme XVI—Advances in Solid State Physics,* Braunschweig: Vieweg, pp. 149–178.

Seeger, A., Swanson, M. L. (1968) in: R. R. Hasiguti (Ed.), *Lattice Defects in Semiconductors,* Tokyo: Univ. of Tokyo Press, pp. 93–130.

Semiconductor Industry Association (November, 1994), *1994 National Technology Roadmap for Semiconductors,* San Jose, CA: Semiconductor Industry Association.

Shimura, F. (1989), *Semiconductor Silicon Crystal Technology,* New York: Academic.

Shimura, F., Huff, H. R. (1985), in N. G. Einspruch (Ed.), *VLSI Handbook,* New York: Academic, pp. 191–269.

Shimura, F. (Ed.) (1994), *Oxygen in Silicon,* New York: Academic.

Siethoff, H. (1994), *Appl. Phys. Lett.* **65,** 174–176.

Slack, G. A., Bartram, S. F. (1975), *J. Appl. Phys.* **46,** 89–98.

Smith, R. A. (1961), *Semiconductors,* Cambridge: Cambridge Univ. Press.

Spencer, M. G. (Ed.) (1993), *Proceedings Fifth International Conference on SiC and Related Materials,* IOP Conference Series No. 137, Bristol: IOP Publishing.

Strausser, Y. E., Doris, B., Diebold, A. C. Huff, H. R. (1994), *Electrochem. Soc. Extended Abstracts* **94-1,** 461–462.

Sumino, K. (1983), J. Nishizawa (Ed.), *Semiconductor Technologies* **8,** 23–44.

Swalin, R. A. (1972), *Thermodynamics of Solids,* 2nd ed. New York: Wiley.

Sze, S. M. (1981), *Physics of Semiconductor Devices,* 2nd ed. New York: Wiley.

Tairova, Y. M., Tsvetkov, V. F. (1983), in: P. Krishna (Ed.), *Growth and Characterization of Polytypic Crystals,* New York: Pergamon, pp. 111–162.

Tan, T. Y., Gardner, E. E., Tice, W. K. (1977), *Appl. Phys. Lett.* **30,** 175–176.

Taylor, W., Gösele, U., Tan, T. Y. (1993), in: G. R. Srinivasan, K. Taniguchi, C. S. Murthy (Eds.), *Process Physics and Modeling in Semiconductor*

Technology, Pennington, NJ: The Electrochemical Society, pp. 3–19.

Taylor, W. E., Dash, W. C., Miller, L. E., Mueller, C. W. (1960), H. C. Gatos (Ed.), *Properties of Elemental and Compound Semiconductors,* Metallurgical Society Conferences, New York: Interscience, Vol. 5, pp. 327–335.

Teal, G. K. (1976), *IEEE Trans. Electron Devices* **ED-23,** 621–639.

Teal, G. K., Runyan, W. R., Bean, K. E., Huff, H. R. (1985), in: J. F. Young, R. S. Shane (Eds.), *Materials and Processes,* 3rd ed., New York: Marcel Dekker, pp. 219–312.

Thurber, W. R., Mattis, R. L., Liu, Y. M., Filliben, J. J. (1980a), *J. Electrochem. Soc.* **127,** 2291–2294.

Thurber, W. R., Mattis, R. L., Liu, Y. M., Filliben, J. J. (1980b), *J. Electrochem. Soc.* **127,** 1807–1812.

Trumbore, F. A. (1960), *Bell Syst. Tech. J.* **39,** 205–233.

Tu, K.-N., Mayer, J. W., Feldman, L. C. (1992), *Electronic Thin Film Science,* New York: Macmillan.

Weber, E. R. (1988), *Properties of Silicon,* EMIS Datareviews Series Vol. 4, London: INSPEC, pp. 409–451.

Weber, J. (1994), *Diffus. Defect Data* **37–38,** 13–24.

Weber, J., Alonso, M. I., (1990), in: K. Sumino (Ed.), *Defect Control in Semiconductors, 1990,* Amsterdam: Elsevier, pp. 1453–1457.

Wentorf, R. H. (1992), in: S. Mahajan, L. C. Kimerling (Eds.), *Concise Encyclopedia of Semiconducting Materials & Related Technologies,* New York: Pergamon, pp. 91–94.

Wijaranakula, W., (1993), *J. Electrochem. Soc.* **140,** 3306–3316.

Wise, R. L., Lindberg, K. J., Dyer, L. D., Huff, H. R. (1994), in: D. Bloor, R. J. Brook, M. C. Flemings, S. Mahajan (Eds.) *The Encyclopedia of Advanced Materials,* London: Pergamon, 2469–2478.

Wortman, J. J., Evans, R. A. (1965), *J. Appl. Phys.* **36,** 153–156.

Yarbrough, W. A. (1990), in: J. T. Glass, R. Messier, N. Fujimori (Eds.), *Diamond, Silicon Carbide and Related Wide Bandgap Semiconductors,* MRS Symposium Proceedings No. 162, Pittsburgh, PA: Materials Research Society, pp. 75–83.

Ziman, J. M. (1979), *Models of Disorder,* London: Cambridge University Press.

Further Reading

Cerofolini, G. F., Meda, L. (1989), *Physical Chemistry Of, In and On Silicon,* Berlin: Springer-Verlag.

Graff, K. (1994), *Metal Impurities in Silicon-Device Fabrication,* Berlin: Springer-Verlag.

Harbeke, G., Schulz, M. J. (1989), *Semiconductor Silicon,* Berlin: Springer-Verlag.

Monch, W. (1993), *Semiconductor Surfaces and Interfaces,* Berlin: Springer-Verlag.

Runyan, W. R., Bean, K. E. (1990), *Semiconductor Integrated Circuit Processing Technology,* Reading, MA: Addison-Wesley.

Sze, S. M. (Ed.) (1988), *VLSI Technology,* 2nd ed., New York: McGraw-Hill.

Tarui, Y. (Ed.) (1986), *VLSI Technology: Fundamentals and Applications,* Berlin: Springer-Verlag.

Wolf, S., Tauber, R. N. (1986), *Silicon Processing for the VLSI Era,* Vol. 1, *Process Technology,* Sunset Beach, CA: Lattice Press.

Wolf, S. (1990), *Silicon Processing for the VLSI Era,* Vol. 2, *Process Integration,* Sunset Beach, CA: Lattice Press.

Wolf, S. (1995), *Silicon Processing for the VLSI Era,* Vol. 3, *The Submicron MOSFET,* Sunset Beach, CA: Lattice Press.

Yu, P. Y., Cardona, M. (1995), *Fundamentals of Semiconductors, Physics and Materials Properties,* Berlin: Springer-Verlag (1996).

Zangwill, A. (1988), *Physics at Surfaces,* Cambridge: Cambridge University Press.

SEMICONDUCTORS, ORGANIC—ELECTRONIC PROPERTIES

MARTIN POPE, *Chemistry Department, New York University, New York, New York, U.S.A.*

CHARLES E. SWENBERG, *Physics Department, George Mason University, Fairfax, Virginia, U.S.A.*

Introduction 478
1. **Molecular Structure of Typical Organic Semiconductors** 479
1.1 π-Electron Molecules 479
1.2 π-Electron Solids 481
1.2.1 Organic Crystals 481
1.2.2 Polymers 481
1.2.3 Molecularly Doped Polymers .. 481
1.2.4 Langmuir–Blodgett (LB) Films 482
1.3 σ-Bonded Solids 482
1.3.1 Polysilanes 482
1.4 Fullerenes, C_{60} 482
2. **Electronically Excited States** 482
2.1 Excited States of Aggregates of Molecules 482
2.2.1 Excitons 482
2.2.1.1 Frenkel Excitons 483
2.2.1.2 Wannier–Mott Excitons 483
2.2.1.3 Charge-Transfer (CT) Excitons 483
2.2.2 Polarons 483
2.2.3 Solitons 483
2.2.4 Bipolarons 484
3. **Carrier Generation** 484
3.1 Band Theory 484
3.2 Energy Levels 485
3.2.1 Valence and Conduction Bands 485
3.3 Single Positive or Negative Carriers in Organic Solids 485
3.3.1 Generation Mechanisms 485
3.3.1.1 Dark Injection from an Electrode 485
3.3.1.2 Injection by Optically Excited Electrodes 487
3.3.2 Exciton Interactions 487
3.3.2.1 Exciton Dissociation at an Electrode 488
3.3.2.2 Exciton Dissociation at Impurities 488
3.3.2.3 Exciton Dissociation at Surface Molecules 488
3.3.3 Photovoltaic Effect 488
3.4 Carrier Generation of Pairs in the Bulk 489
3.4.1 Thermal Mechanism 490
3.4.2 Photonic Mechanisms 490
3.4.2.1 Autoionization 490
3.4.2.2 Band-to-Band Transitions 490
3.4.2.3 Excitonic Mechanisms 491
4. **Carrier Recombination** 491
4.1 Geminate Recombination—Onsager Theory 491
4.2 Geminate Recombination—Time-Dependent Theory 494
4.3 Bimolecular Recombination ... 494
4.3.1 Electroluminescence 494
5. **Carrier Transport** 495
5.1 Types of Transport 496
5.1.1 Band Motion 496
5.1.2 Electric-Field Effects 496
5.1.3 Hopping Motion 496
5.1.4 Transport in Polymeric Materials 497
5.1.5 Dispersive Transport 500
5.1.6 Bässler Formalism 500
5.1.7 Positron Mobility 502
5.2 Steady Current Flow 502
5.2.1 Basic One-Carrier Theory 502
5.2.2 Space-Charge–Limited Currents (SCLC) 503
5.2.3 Carrier Trapping 503
5.2.3.1 Thermally Activated Mobility 505
5.2.3.2 Thermally Stimulated Currents (TSC) and Isothermal Decay Currents (IDC) 505
5.2.3.3 Optically Stimulated Current 506
5.2.3.4 Detrapping by Excitons 506
5.2.3.5 Trap Distributions 506

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

5.3 Transient Currents (Single Carrier) 507
5.3.1 Space-Charge–Free (SCF) Transients 507
5.3.2 SCLC Transients 507
6. Photoemission 509
6.1 Energy-Level Analysis 510
6.2 Multiphoton Studies 511
7. Magnetic-Field Effects 511
7.1 Hall Mobility 511
7.2 Magnetic-Field Effects on Spin–Spin Interactions 511
8. Applications 513
8.1 Electrophotography 513
8.2 Electroluminescence (EL) 513
8.3 Sensors 514
8.4 Semiconductor Devices 514
8.5 Photorefractive Polymers 515
Acknowledgments 516
Glossary 516
Works Cited 517
Further Reading 519

INTRODUCTION

Semiconductors play a vital role in society. They are at the base of all computers, all modern communication systems, all controls for modern transportation systems, and in fact all modern control systems in general. They constitute a multibillion-dollar industry. The vast majority of semiconductors (and as will be explained shortly, all photoconducting insulators are included in this group) are inorganic, being composed of such compounds as silicon, germanium, gallium arsenide, and others (see INSULATORS AND SEMICONDUCTORS, CONDUCTIVITY IN). These materials are used in both crystalline and amorphous forms. The theoretical treatment of the crystalline material is well developed and is capable of providing both a rationalization of observed behavior and a predictive capacity for as yet uncharted areas. The theory of amorphous inorganic semiconductors is not as well developed.

Alongside the vast electronics industry based on inorganic semiconductors is the huge polymer industry, based on organic compounds. The synthetic genius of chemists in this field is well known. The ability of nature to fashion biological structures capable of thought, sensation, and communication, and using light to grow food provides an existence proof of the feasibility of synthesizing polymers that can mimic these functions. There is sufficient evidence that this goal is not a dream. As will be described in Sec. 8, examples of such developments have been realized. Most electrographic copiers have polymeric films that capture and transfer the image to the page. The development of polymeric electronic materials, with all the benefits of magnitude of production capacity and product versatility, depends on an adequate theoretical understanding of the fundamental structural and electronic properties of organic solids. A brief description of progress towards the end will be presented here.

This article will be devoted entirely to organic molecular solids. An organic compound is one that contains carbon in combination with other elements. Of the more than two million known compounds, about 90% are organic. Some compounds such as metal carbonates are usually considered inorganic compounds. However, the number of carbon compounds considered as inorganic is relatively small, and a minimal error is committed by accepting the definition given above. A molecular solid is one in which the individual molecules are electrically neutral, but are attracted to neighboring molecules by the weak van der Waals forces (see COHESIVE ENERGY). Examples of molecular solids are the rare gases at sufficiently low temperatures, polymers like polyethylene, and many organic compounds, like anthracene. Typical intermolecular bond energies are ≤ 0.1 eV, but the intramolecular bond energies can be quite high; in polyethylene, the energy is about 3.4 eV. Most of the organic compounds of particular interest to this discussion are called conjugated hydrocarbons. Essentially all electronic phenomena that have been observed in inorganic compounds have also been found in organic compounds. The differences are in degree only.

As a further subdivision, the intramolecular bonding of the organic systems to be discussed will be conjugated in nature. These are the only compounds to date that have been extensively studied. Conjugated compounds have delocalized electronic orbitals, which lead to facile intramolecular carrier transport and to a system of lower-lying ionization levels than would be the case if the orbitals were highly localized. In addition, the overlap of orbitals on neighboring molecules facilitates electron mobility. The greater ease of generating carriers and the increased carrier mobility of conjugated compounds make them better semiconductors than other compounds such as sugars or plastics like polyethylene or Teflon. The description of a conjugated bond will be found in Sec. 1. Aside from polymeric systems, single-crystal solids will be reported on whenever possible. Finally, only those processes that involve the generation, transport, and recombination of free charges in organic semiconductors (OSC) will be discussed. Generally speaking, a semiconductor is a material that exhibits a thermally activated dark conductivity in the range 10^{-3} to 10^{-10} Ω^{-1} m^{-1}. This definition will be broadened herein to include molecular materials that can efficiently transport photoexcited or dark-injected carriers (in some cases with negligible loss to deep traps) even though these materials are insulators in the dark. This excludes a significant group of organic materials that exhibit metallic conductivity, and even superconductivity (Ramirez, 1994). It also excludes phenomena based on lattice polarization rather than current flow, such as pyroelectricity and piezoelectricity (Mort and Pfister, 1982), and ferroelectricity (Kepler and Anderson, 1992). In addition, recently developed organic paramagnetic solids will not be discussed (see Miller and Epstein, 1994).

The plan of this article is to introduce the reader to the types of molecular structures that will be encountered. This is followed by a brief description of the photoexcited states that appear in the more detailed exposition. Then comes a discussion of the various techniques that are available for generating the charge carriers that form the basis of all the anticipated functions of organic semiconductors (OSC). Following this comes a discussion of carrier recombination, which is one of the most characteristic features of OSC. Carrier recombination can be very efficient, and presents a problem if a high quantum efficiency of free-carrier generation is desired. If the carriers survive the recombination obstacle, they must move through the solid, and this process is not adequately understood; in fact, it represents a major unsolved problem.

An obstacle to carrier transport consists of chemical and physical defects in the solid that trap the carriers; this is discussed next. The following two sections deal with special techniques for probing the electronic structure of solids, and the last section describes several applications in which organic materials can seriously challenge inorganic materials.

Note: In view of the extensive referencing that will be made to the monographs authored by Pope and Swenberg (1982) and by Silinsh and Capek (1994), these works will be referred to by the initials of the authors and the relevant page number in each book, e.g., PS, p. 450 or SC, p. 265.

1. MOLECULAR STRUCTURE OF TYPICAL ORGANIC SEMICONDUCTORS

1.1 π-Electron Molecules

The electronic configuration of the ground state of the carbon atom is $1s^2 2s^2 2p^2$. There are four electrons in the outer electronic orbitals. The two *s* electrons are paired, and the two *p* electrons are unpaired in each of two of the three mutually perpendicular *p* orbitals. Another possible electronic configuration for the carbon atom may be realized by mixing the 2*s* and the three 2*p* orbitals to create a set of four equivalent degenerate orbitals, referred to as sp^3 hybrid orbitals. These orbitals are oriented so that each points in the direction of one corner of an equivalent tetrahedron. Methane, CH_4, provides an example of sp^3 bonding. It is also possible for one *s* and two *p* orbitals (e.g., p_x, p_y) to combine to form three sp^2 orbitals (sp^2 hybridization); these are referred to as planar trigonal orbitals. In this case, one of the three original mutually perpendicular atomic *p* orbitals, the p_z orbital, remains unaltered. The three sp^2 orbitals are all coplanar and directed about 120° apart from each other; the bonds formed from these orbitals are called σ bonds; the p_z is perpendicular to the plane of the sp^2 orbitals. The hybrid-orbital concept provides a convenient model for un-

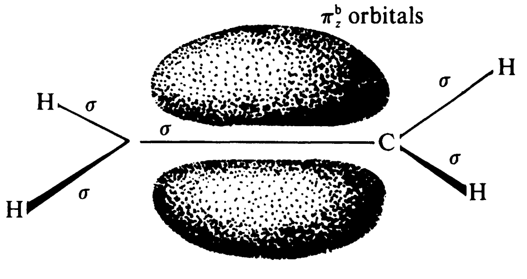
π_z^b orbitals
H
H
H
H
C
σ
σ
σ
σ
σ

FIG. 1. Double-bond character of ethylene, C_2H_4. All σ bonds are represented by straight lines. The p_z atomic orbitals of the C atoms overlap to form the bonding molecular π_z^b orbital.

derstanding the formation of such double-bonded carbon structures as ethylene (see Fig. 1) and polyacetylene [see Fig. 2(a)]. More important for this work is the type of structure found in benzene, anthracene, and the entire family known as conjugated aromatic compounds. The p_z orbital of each carbon atom in the anthracene molecule, as an example, is perpendicular to the plane containing all the carbon atoms, and the spacing between the carbon atoms is such that there is an overlap between neighboring p_z orbitals, creating what is called a molecular π bond. The π bond establishes a delocalized electron density above and below the plane of the carbon atoms with zero π-electron density in the nodal plane coinciding with the plane of the molecule. The C–C interatomic distance in anthracene and other conjugated molecules is small enough so that the π-electron system is delocalized over the entire molecule. In contrast, the sp^2 orbitals generate highly localized electron density in the plane of the ring, and between the carbon and hydrogen atoms. The new form of carbon referred to as buckminsterfullerene, or fullerene, C_{60}, is neither a linear polyacene nor a hydrocarbon. Nevertheless, in its crystalline form, it behaves like a typical organic molecular solid, with electronic prop-

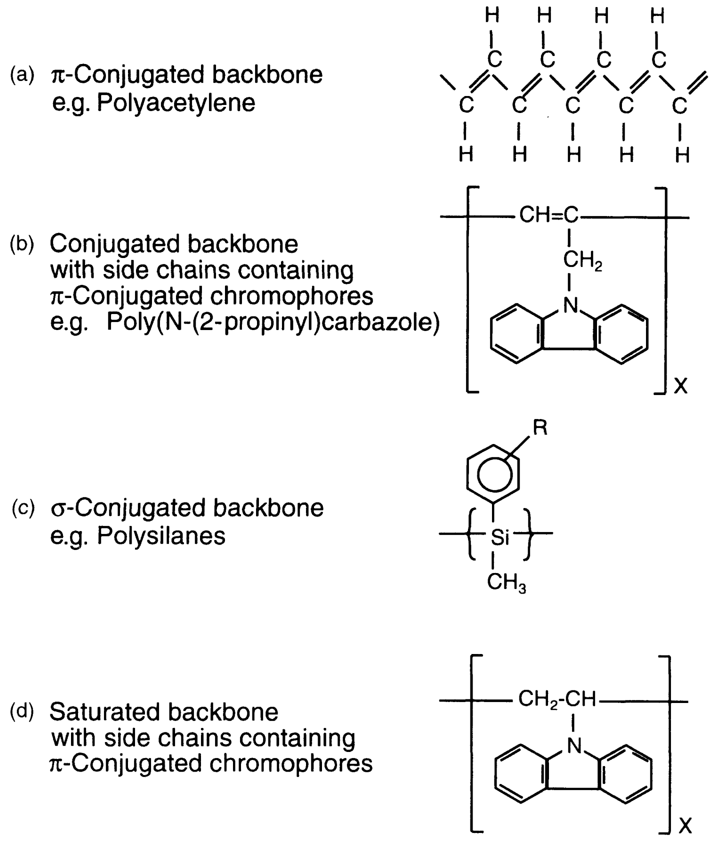
(a) π-Conjugated backbone
e.g. Polyacetylene
(b) Conjugated backbone
with side chains containing
π-Conjugated chromophores
e.g. Poly(N-(2-propinyl)carbazole)
(c) σ-Conjugated backbone
e.g. Polysilanes
(d) Saturated backbone
with side chains containing
π-Conjugated chromophores
H
C
CH=C
CH2
N
X
R
Si
CH3
CH2-CH
N
X

FIG. 2. Three characteristic groups of polymeric organic semiconductors. (a), (b) Polymers with completely π-conjugated backbones, and some with π-conjugated pendant groups as well; (c) polymers with σ-conjugated backbones like the polysilanes and some with σ-saturated backbones and π-conjugated pendant groups; (d) polymers with a saturated backbone with π-conjugated pendant groups, such as polyvinylcarbazole (PVK). (Redrawn from Nešpůrek, 1993.)

erties similar to those of anthracene (see Sec. 3.4.2.1). It is therefore included in the class of organic semiconductors.

In addition to the ringed aromatic compounds, there is the important class of conjugated polymeric compounds, an example of which is polyacetylene, the simplest linear conjugated polymer [see Fig. 2(a)]. In studying the electronic properties of these carbon-backbone conjugated compounds, it will be sufficient to focus on the properties of the π electrons, to the exclusion of the σ-electron network, since σ electrons are most tightly bound and thus do not participate in transport. The π electrons are in the highest-energy occupied molecular orbitals (HOMO) and are the most easily excited or ionized without disrupting the σ bonds that hold the molecule together. Since the lowest-energy unoccupied molecular orbitals (LUMO) are also π orbitals, the electronic transitions from the HOMO to the LUMO are referred to as π–π^* transitions. These transitions can result in singlet or triplet excited states, depending on whether the spin of the electron remains the same or changes orientation upon excitation.

1.2 π-Electron Solids

1.2.1 Organic Crystals As a crystal, anthracene is the most thoroughly studied conjugated aromatic hydrocarbon insofar as its electronic properties are concerned. This crystal is a typical molecular crystal; it is soft, easily deformed, and has a relatively low melting point. However, despite the weak interaction between the molecules in the crystal, there is sufficient interaction between neighboring molecules to create a band of crystal energy levels. This band of states is referred to as an exciton band (see EXCITONS) and is depicted graphically in Fig. 3. Anthracene has two inequivalent molecules per unit cell, and the interactions among these molecules produce a divided band of crystal states referred to as a Davydov splitting (see EXCITONS).

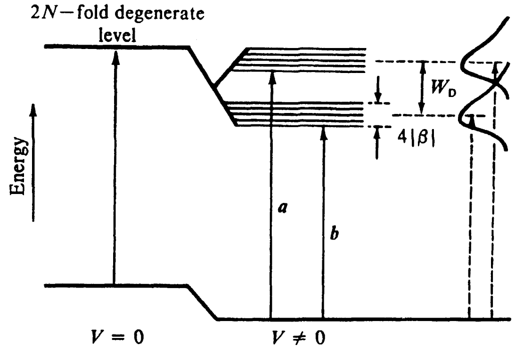

FIG. 3. Splitting of the 2*N*-fold degenerate level for a crystal with two inequivalent molecules per unit cell into two distinct Davydov bands. The polarizations of each band are denoted by *a* and *b*. The energy separating the two $k = 0$ states is W_D, the Davydov splitting. Its magnitude arises from inequivalent intermolecular interactions only.

1.2.2 Polymers The polymers that are most important as semiconductors and photoconductors can be classified into three groups (Nešpůrek, 1993). These are shown in Fig. 2; in Figs. 2(a) and 2(b) are polymers with completely π-conjugated backbones, and some with π-conjugated ligands as well. These become highly conductive when doped with donors or acceptors. As photoconductors, they operate in the visible and near infrared. Shown in Fig. 2(c) are polymers with σ-conjugated backbones, such as the polysilanes. These are good photoconductors, and have hole mobilities of the order 10^{-8} m^2 V^{-1} s^{-1}. Polymers like poly(*N*-vinylcarbazole) with a saturated (σ-bonded) backbone and conjugated side chains [Fig. 2(d)] are normally photoconductive in the UV.

1.2.3 Molecularly Doped Polymers Polymers have highly desirable mechanical and chemical properties such as flexibility, strength, and chemical stability. It is difficult, however, to synthesize polymers at will with any of a list of desirable electronic properties. On the other hand, there are monomeric molecules with desirable electronic properties, but with none of the properties of polymers. By dissolving or dispersing short-chain molecules in a stable polymeric matrix, it is possible to combine some of the best properties of both polymer and dopant. The great advantage of the molecularly doped polymers is that they can be prepared in the form of thin films, and their mobilities can be adjusted from 10^{-13} to 10^{-7} m^2 V^{-1} s^{-1} by varying the concentration of the active species. More significantly, they can be rendered effectively trap free for the majority carrier by judicious choice of the dopant (see Sec. 5.3).

As a host polymer, a favorite is a polycarbonate (commercial name Lexan). The dopant can be an electron acceptor, an electron donor, or a combination of both. A typical example of a molecularly doped polymer is a dispersion of *N*-isopropyl carbazole (NIPCA) in Lexan. Lexan by itself is an excellent insulator. When NIPCA is dispersed in the Lexan in a concentration of $\sim 10^{26}$ molecules m^{-3}, then, under the action of light, hole generation and transport are observed; the photoexcited dopant ionizes into a hole and electron but only the hole is mobile. Hole transport takes place by hopping from molecule to molecule of NIPCA.

1.2.4 Langmuir–Blodgett (LB) Films The LB film technique makes possible the construction of thin films with a thickness varying from that of one or two molecules to many tens of molecules, such that each layer can possess carefully tailored properties (see Zasadzinski *et al.*, 1994). The production of thin films is of considerable technological importance (see THIN FILMS). By use of the LB technique, a multilayer structure can be assembled in which essentially two-dimensional (2-D) organic semiconductors separated by insulating spacers form a sequence of quantum wells (see COHESIVE ENERGY) that behave like those in inorganic systems (SC, p. 195; Roberts, 1990).

1.3 σ-Bonded Solids

1.3.1 Polysilanes Unlike the conjugated aromatic π-bonded organic compounds that constitute almost the entire group that possesses semiconducting properties, there is the group of polysilanes (and polygermanes). In the polysilanes, an example of which is shown in Fig. 2(c), the σ bonds along the Si–Si backbone are delocalized. These σ-bonded polymers exhibit many of the properties of the π-bonded compounds. These include a high fluorescence quantum efficiency ($\sim$0.5 in some compounds) and excellent photoconductivity. A simplified picture of the basis for delocalization is the following: In a polysilane polymer there are approximately sp^3 orbitals facing each other on neighboring Si atoms (called vicinal Si orbitals) and weaker interactions between approximately sp^3 orbitals on the same Si atom (called germinal Si orbitals). Ignore for the present the sp^3 orbitals that are attached to the ligands on each Si atom. The interacting vicinal orbitals split into strongly localized σ bonding and antibonding orbitals. The weaker geminal interactions connect the entire linear chain of localized bond orbitals with each other, creating delocalized orbitals over the entire Si backbone (see also Sec. 5.1.5) (Miller and Michl, 1989; Kepler and Soos, 1993).

1.4 Fullerenes, C_{60}

In 1985, a remarkable discovery was made of a new allotropic form of carbon, called buckminsterfullerene or fullerene, having the molecular formula C_{60}. It has a structure resembling a soccer ball; it is a truncated icosahedron consisting of 32 faces, of which 20 are hexagonal and 12 are pentagonal. The compound is aromatic in its chemical properties. The solid is a typical molecular crystal and the crystalline material is an organic semiconductor. The molecule has a large electron affinity, and so it can trap electrons when added as a dopant to other systems (see Sec. 3.4.2.1 for additional data). When alkali-metal atoms are diffused into a C_{60} film, ionic compounds are formed, some of which are superconductors at the relatively high temperature of 33 K and ambient pressure (e.g., $RbCs_2C_{60}$) (see Ramirez, 1994).

2. ELECTRONICALLY EXCITED STATES

A complete discussion of optically excited states in organic molecular crystals is outside the bounds of this article. However, an important characteristic feature of these solids is the ability of the optical excitation to migrate inside the crystal in the form of an exciton. The ground state of the molecule is usually a singlet state. Nevertheless, there is great interest in organic molecules that have triplet ground states, with the view of preparing ferromagnetic organic solids (Miller and Epstein, 1994).

2.1 Excited States of Aggregates of Molecules

2.2.1 Excitons Many, but not all, spectral properties of small aggregates of molecules and crystals are directly traceable to properties of individual molecules. Weak as the interaction between molecules may be, it imposes a communal response upon the aggregate; the collective response to electronic excitation is embodied in a quasiparticle called an exciton (see EXCITONS). In conjugated organic solids, exciton properties are greatly affected by the size of the exciton, defined as the average distance between the electron and hole. These limiting states are called Frenkel and Wannier excitons, with charge-transfer (CT) excitons occupying an intermediate position.

2.2.1.1. Frenkel Excitons. In a solid containing a high concentration of identical chromophores (see the Glossary), the electronic excitation of any particular chromophore (neglecting surface effects) produces the same electronic state of the solid. This final crystal state is called a Frenkel exciton. Frenkel excitons are electrically neutral and represent the limiting case where the electron–hole pairs are located primarily on the same molecule, moving as a unit from molecule to molecule. In the case of polymers, the molecule is so large that distinctions among the various excitonic types blur. When pairs are far apart, a Frenkel exciton can behave like a polaron pair (see POLARONS). The exciton can have singlet or triplet spin character, depending on the excited state of the molecule. The exciton bandwidth ($4|\beta|$ in Fig. 3) is determined by the intermolecular forces. The bandwidth of a singlet exciton in anthracene is about 200 cm^{-1}, whereas for the lowest triplet exciton, it is $\sim$20 cm^{-1}. These bandwidths are quite anisotropic as a result of the anisotropy of the lattice. Since they are mobile, these excitons are capable of dissociation at an electrode, a defect, or a surface to produce a charge carrier. They are thus important in photoconductivity. As for polymers, the backbone of the π-conjugated polymers such as substituted poly(p-phenylenevinylene) (PPV) and σ-bonded polysilanes also supports excitonic motion, as do the pendant π-conjugated groups, known as chromophores; fullerene, C_{60}, also exhibits excitonic properties (Vardeny and Rothberg, 1994, p. 187).

2.2.1.2 Wannier–Mott Excitons. In contrast to the Frenkel exciton, the Wannier–Mott exciton is not confined to an individual molecule, but consists of a widely separated hole and electron pair that are still bound to each other by the Coulomb interaction. The energy levels and corresponding absorption bands associated with the Wannier–Mott states resemble Rydberg transitions in the hydrogen atom. These modified Rydberg levels form an energy-level series (E_n) described by the equation

$$E_n = E_g - G/n^2 \quad n = 1, 2, \ldots,$$

where E_g is the band gap and G is the binding energy of the exciton. The dissociation energy of the exciton is given by the value $n = 1$. Wannier–Mott excitons are important in inorganic solids such as Si, Ge, II–VI compounds, and rare-gas solids; they may play a role in the polysilane polymeric semiconductors. For a discussion of excitons in inorganic solids, see Knox (1963).

2.2.1.3 Charge-Transfer (CT) Excitons. Charge-transfer excitons can be considered as intermediate in size between small (Frenkel) and large (Wannier-Mott) excitons. The hole and its paired electron are on different molecules, but are bound Coulombically to each other and move as a unit through the solid. CT excitons play an important role in carrier generation and recombination in organic semiconductors. These excitons also exhibit a series of excited states consisting of increased electron–hole separations. The energies of these states can be fitted by a Rydberg-type series for electron–hole separations >1.0 nm.

2.2.2 Polarons In addition to neutral quasiparticles, there are charged states that can differ significantly from that of a free hole or electron by virtue of their interaction with the lattice phonons (see PHONONS). This interaction results in the polarization of the neighboring electrons and nuclei, and produces a lattice distortion. If the distortion is confined to nearest neighbors, it is referred to as a small polaron (see POLARONS). If the polarization covers a region with a radius much larger than a lattice constant, it is called a large polaron. As discussed in Sec. 5, the lattice distortion that accompanies the carrier modifies the temperature dependence of the mobility.

2.2.3 Solitons In the context of organic semiconductors, a soliton is a mobile chemical structural defect. In *trans*-polyacetylene (*trans*-PA), an interchange of single and double bonds does not change the energy of the system. This is not true for *cis*-PA or with most conjugated polymers. There are therefore two degenerate forms of the ground state of *trans*-PA. In the process of forming *trans*-PA by isomerization of *cis*-PA, the transformation can take place at different points of the *cis*-PA molecule, and so it is possible for the two degenerate forms (A and B in Fig. 4) to meet, producing the defect shown in the figure. This defect consists of two adjoining single bonds with an unpaired electron at the point of junction, and is called a soliton. This is not a completely appropriate appellation, since unlike the rigorously defined soliton, these topological defects cannot pass through each other unchanged. The soliton can move along the chain almost isoenergetically, and it can be charged or uncharged. It can also hop from chain to chain (Heeger *et al.*, 1988). The mechanism of charge transport through other polymers lacking a degenerate ground state cannot depend upon charged soliton transport, or intersoliton hopping (see Sec. 5.1.4).

2.2.4 Bipolarons Another model for transport in polymers proposes the existence of like-charged solitons as pairs on a single chain that remain near each other despite the presence of Coulomb repulsion. The repulsive interaction is overcome by the energy gained by each charge as it polarizes the lattice in its vicinity. These charged pairs are referred to as bipolarons, and transport from one chain to another can take place by the transfer of a pair of charges (bipolaron) from a charged chain to a neutral chain, just as occurs for single charges in normal semiconductors (see Sec. 5.1.4) (Brédas and Street, 1985).

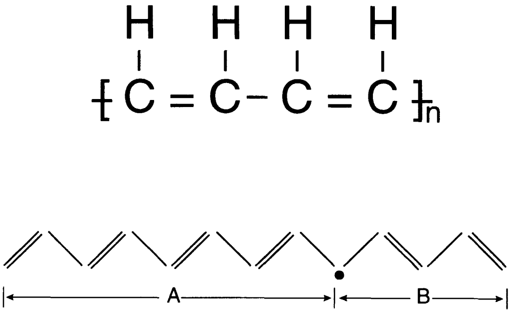

$$\left[\overset{H}{\underset{}{C}}=\overset{H}{C}-\overset{H}{C}=\overset{H}{C}\right]_n$$

FIG. 4. Bond-alternation pattern of *trans*-polyacetylene showing the connection of the two degenerate forms (A and B) of the molecule. The soliton defect is the junction of the two single bonds. The defect is a radical (unpaired electron) and is neutral as shown. Hydrogen bonds are omitted for clarity.

3. CARRIER GENERATION

In conjugated organic solids, the intrinsic conductivities can vary continuously from that of excellent insulators to superconductors. The mechanism of carrier generation is therefore not general for these solids. The organic carrier-transport solids discussed herein consisting of a single component are essentially insulators in the dark. Unless the band gap is less than 0.5 eV, the density of free carriers produced thermally at room temperature will be less than the number of trapped carriers generated by impurities. This section is devoted to those solids in which the intrinsic dark conductivity is low, below $10^{-3}\ \Omega^{-1}\ \mathrm{m}^{-1}$.

Carriers can be generated in organic solids by a variety of methods. It is possible to use special electrodes that will inject carriers of either sign into the solid in the dark. It is also possible to photogenerate carriers by creating excitons that dissociate at an impurity, or react with an electrode. In this case, free carriers of either sign can be generated inside the solid, but not both at the same time in the same place. The examples just cited describe what is referred to as extrinsic carrier generation. Intrinsic photogeneration of carriers in a molecular crystal requires light of energy at least equal to that of the band gap. In a single-component polymer, the individual polymer molecule must be ionized for intrinsic generation; in multicomponent polymers, photogeneration involves the participation of the dopant and is classified as extrinsic.

3.1 Band Theory

In typical inorganic semiconductors, where there is a crystalline lattice and intermolecular interactions are large, there is a well-defined forbidden energy gap E_g separating broad ($\gg kT$) valence and conduction bands. Carrier generation in these solids takes place when an absorbed photon energy exceeds E_g, the quantum efficiency of carrier generation is a weak function of the electric field and temperature, and the mobility μ varies as T^{-n}, $n > 1$ (see Sec. 5).

In typical organic molecular-crystal semiconductors at room temperature, the carrier bandwidths are narrow, in many cases $\ll kT$. This is a consequence of their weak intermolecular interaction energy and low dielectric constants. The Coulomb binding energy between the hole and electron is therefore large and long range. As a consequence, the quantum efficiency of carrier generation is a function of the electric field and temperature. However, Karl (1990) has shown that as the temperature is lowered to liquid He temperatures in ultrapure naphthalene, the one-electron band picture is valid (bandwidths $\gg kT$), at least insofar as the mobility is concerned (see sec. 5).

In linear conjugated-backbone polymers, broad 1-D carrier bands exist along the chain at room temperature. In *trans*-PA, the π bandwidth is ~10 eV. However in the direction perpendicular to the chain, the interaction energy is ~0.1 eV, and in view of the lack of long-range order, an intermolecular band theory is not tenable. The quantum efficiency of carrier generation is greatly decreased by geminate recombination (sec. 4.1). The product of the quantum efficiency of carrier production and the probability of escaping fast geminate recombination is 0.01 (Něspůrek, 1993). The effective generation efficiency therefore depends markedly on the electric field and temperature, as it does in other linear-chain polymers such as poly(*p*-phenylenevinylene) polymers (Gailberger and Bässler, 1991).

The question of whether carrier generation takes place as a result of a valence-to-conduction-band transition (referred to as a band-to-band or BB transition) in organic semiconductors is taken up in Sec. 3.4.2.2.

3.2 Energy Levels

3.2.1 Valence and Conduction Bands

In view of the weak intermolecular interactions, the ionic energy levels in a molecular crystal bear a close relationship to those in the isolated molecule, modified by the polarization energy gained by bringing the molecular ions into the solid phase. In its simplest form, the energy E_g required to produce a pair of free carriers in organic molecular crystals is given by

$$E_g = I_g - A_g - P^+ - P^-, \tag{1}$$

where I_g and A_g are respectively the ionization energy and electron affinity of the isolated molecule, and P^+ and P^- are the polarization energies of the positive and negative carriers inside the solid, respectively; they are typically between 1 and 2 eV (Seki, 1989). Although Eq. (1) is only an approximation, it nevertheless is still a useful guide.

Generally, the polarization energies for the positive and negative carriers have been taken as equal to each other. However because there is an asymmetry in the charge-quadrupole interactions of electrons and holes, the total effective polarization energy of the hole is greater than that of the electron, varying from 0.08 eV for benzene to 0.48 eV for pentacene (see SC, p. 116).

The adiabatic values of E_g for anthracene, tetracene, and pentacene are respectively 4.1, 3.1, and 2.5 eV. These are the energies of the relaxed electrons and holes with respect to each other. The optical band gaps are 4.4, 3.4, and 2.8 eV; these are the optical excitation energies required to produce the separated holes and electrons in their unrelaxed polarization states.

As for polymers, Eq. (1) is not useful. Carrier generation is observed at the absorption edge in the carbon-backbone polymers. However, in these systems, where the generation efficiency and carrier mobility are strong functions of the field and temperature, it is difficult and perhaps irrelevant to claim that a BB transition rather than an extrinsic process is responsible for the conductivity (see Sec. 3.4.2.2).

3.3 Single Positive or Negative Carriers in Organic Solids

3.3.1 Generation Mechanisms

3.3.1.1 Dark Injection from an Electrode. When an electrode is placed in contact with an organic semiconducting solid, there is a redistribution of charge that accompanies the equalization of the Fermi energies of the two phases. Let Φ_M be the work function of the electrode, Φ_I the work function of the insulator, and χ (also referred to as A_C) the electron affinity of the insulator, as is shown in Fig. 5. If the ionization energy of the solid is I_c, the energy required to move an electron from the insulator into the electrode is $I_c - \Phi_M$, whereas the energy required for the opposite process is $\Phi_M - \chi$. If $I_c - \Phi_M < \Phi_M - \chi$, then surface electrons will shift from the insulator into the electrode, making the electrode negatively charged with respect to the insulator.

The energy difference $\phi_0 = \Phi_M - \chi$ is the smallest barrier that must be overcome by an electron in the electrode if it is to be injected into the insulator. For holes, there is an analogous value, $\phi_0^+ = I_c - \Phi_M$. Electrons injected into the surface region of the insulator must still overcome a potential barrier in order to enter the bulk of the crystal; this barrier is field dependent, and is shown in Fig. 6 as $\phi_m - \phi_0$.

For anthracene, $E_g \approx 4$ eV, $I_c = 5.8$ eV,

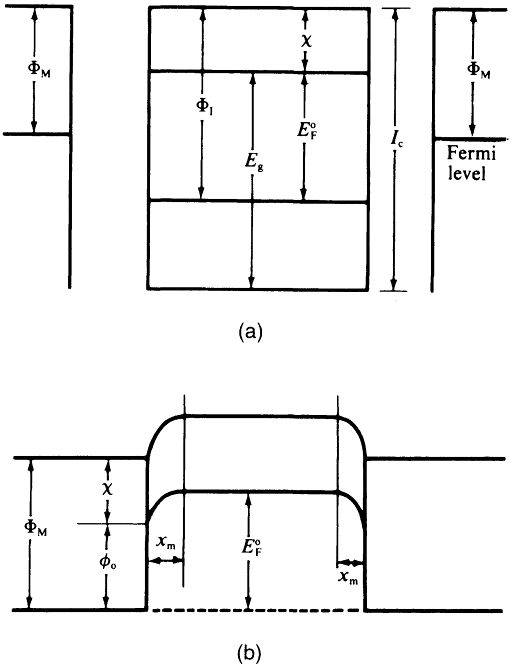

FIG. 5. Energy-level diagram for an insulator with identical metal contacts showing the barrier shape at zero bias (a) before the electrodes are attached to the insulator, (b) after the electrodes are attached. Symbols are defined in the text. (Redrawn from Frank and Simmons, 1967.)

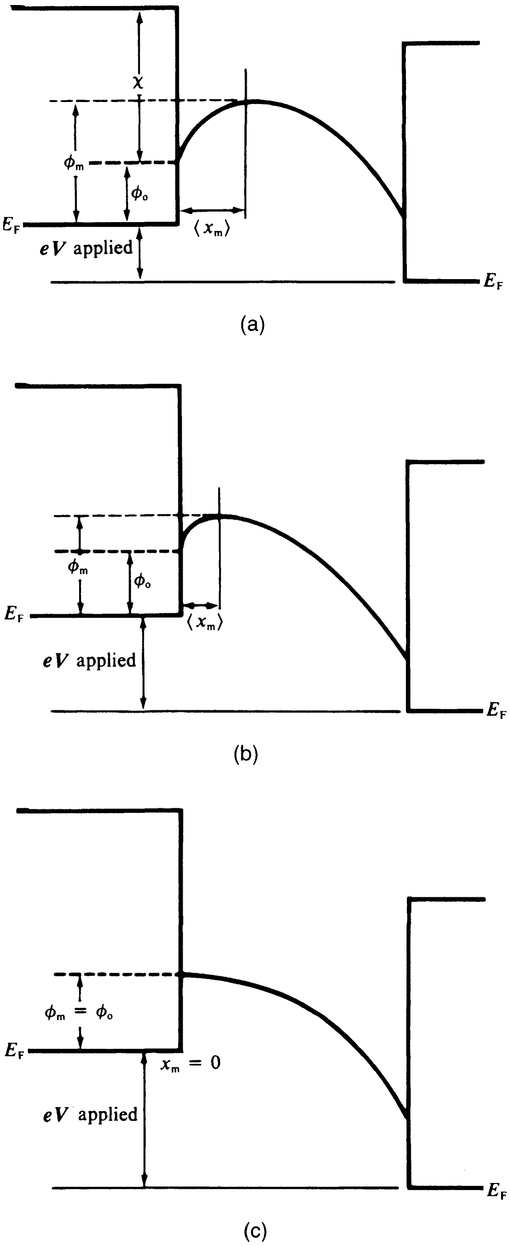

FIG. 6. Effect of increasing the electrical bias on the height and position of the barrier maximum at x_m; (a) small bias; (b) intermediate bias; (c) bias sufficient for the onset of emission-limited current. (Redrawn from Frank and Simmons, 1967.)

and $\chi(A_C) \sim 1.8$ eV; so Φ_I for anthracene should be ~3.8 eV. A good choice for an electrode for electron injection is one with a small work function Φ_M. The alkali metals satisfy this requirement. Sodium (Na) has a Φ_M of ~2.8 eV, and so the barrier to electron injection would be ~1 eV. This is further reduced by a chemical reaction between the Na and anthracene. For hole injection, large-work-function electrodes are required. The noble metals, such as gold (Au), are suitable. The Φ_M of Au is ~5.4 eV, providing a barrier of only 0.4 eV. By using both Au and Na as opposing electrodes on anthracene, electroluminescence is produced (see Sec. 4.3.1).

If the contacting electrode has a low work function, electrons will be injected thermionically into the surface region of the OSC. Equilibrium is established when the rate of injection equals the rate of surface recombination. Upon application of an external field, the Fermi levels of the electrodes are displaced by an amount equal to the external voltage V, and the potential barrier to electrons drops eventually from ϕ_m to ϕ_0 [see Fig. 6(c)]. The surface region of extent x_m serves as an excess-charge reservoir that supplies carriers upon demand. As the applied electric field increases, carriers are forced into the solid at the injecting contact to a maximum determined by the buildup of space charge inside the crystal. This space charge creates an electric field in the opposite direction to that of the external field at the injecting contact, reducing the field strength there virtually to zero. The current that flows under these conditions is called a space-charge–limited current (SCLC), and is the maximum current that can flow in the semiconductor at the applied external field and temperature. A contact that can inject this amount of space charge into the reservoir at the surface is referred to as an Ohmic contact (see Sec. 5.2.1).

In addition to metal electrodes, it has been found that oxidizing electrolyte solutions, such as Ce^{4+}, form excellent hole-injecting contacts (Pope *et al.*, 1962; see Gerischer and Willig, 1976, for a more complete discussion of electrolyte contacts). Currents as high as 1 A m^{-2} were injected in the dark into an anthracene crystal 15 μm thick with 40 V applied.

In addition to the Ohmic contacts just discussed, a contact can be non-Ohmic; i.e., it can be neutral or blocking. In a neutral contact, the energy bands in the insulator remain flat up to the point of contact with the electrode. The concentration of carriers is the same at the interface near the electrode as in the bulk. For a defect-free insulator without surface states, this occurs when $\Phi_M = \Phi_I$. A blocking contact that effectively prevents electron injection, as an example, would be created if in Fig. 5, $\Phi_I < \Phi_M$; in this case, electrons would spill over into the contact, charging the interface negative. This creates an added potential-energy barrier and results in a depletion of electrons in the solid at the interface, relative to the concentration of electrons deeper into the solid.

3.3.1.2 Injection by Optically Excited Electrodes. It follows from Fig. 6 that the barrier to injection of carriers into organic insulators can be overcome if the electrode in contact with the insulator is optically excited. Under these conditions, an electron in the electrode can be excited into the conduction level of the insulator, for electron conductivity, or an electron in the insulator can be excited into the Fermi level of the electrode, enabling hole conductivity in the insulator. The electrode can be a metal, an adsorbed dye, an electrolyte, or a solid mixture of a dye in a polymer.

The condition for electron injection is

$$E_{\text{ext}} + A_C - \chi_e > 0, \tag{2}$$

where E_{ext} is the externally supplied energy, A_C is the electron affinity of the crystal, and χ_e is the electron affinity of the electrode. For metals, χ_e is also the work function Φ_M. The condition for hole injection is

$$E_{\text{ext}} + \chi_e - I_C > 0, \tag{3}$$

where I_C is the ionization energy of the crystal.

If the electrode (e.g., a metal) is optically excited, photoemission of electrons from the metal into the organic solid may be induced in the same way as into a vacuum, except that much lower energies are required.

3.3.2 Exciton Interactions Excitons represent mobile excited states that transport the potential energy of optically excited molecular states. The singlet and triplet excitons are particularly well studied. In anthracene, the singlet exciton transports 3.1 eV and the triplet exciton transports 1.8 eV; hence, this amount of energy can be transferred in exciton dissociation at electrodes, impurities, and surface molecules (see Sec. 3.4.2.2). The same processes can take place in polymers that support exciton transport, which includes all the important single-component systems.

3.3.2.1 Exciton Dissociation at an Electrode. Exciton dissociation is a major mechanism for photoconduction when an electrode is placed in direct contact with an organic semiconductor. The energy-level scheme shown in Fig. 7 shows that both electron and hole injection are energetically possible in anthracene by the reaction of excitons with the electrode. The exciton can transfer its energy via a Förster dipole–dipole mechanism to an electron inside the metal. This can take place over a range of about 3 nm. For triplet excitons, energy can be transferred over a distance of 1.8 nm by electron exchange. Similar behavior is noted in most polymers that support exciton motion, as in polysilanes (Kepler and Soos, 1993).

3.3.2.2 Exciton Dissociation at Impurities. The presence of electron or hole acceptors at the surfaces provides additional sites for exciton dissociation. Oxygen is a well-known electron acceptor, and both singlet and triplet excitons can dissociate at these sites, delivering an electron to the O_2 and liberating a hole for transport. The exciton dissociation rate at an O_2 site on anthracene is about 10^{13} s^{-1}. Iodine is also effective in dissociating excitons at a free surface. The oxidation product anthraquinone on anthracene creates a 1-eV electron trap and provides a ready site for singlet-exciton dissociation to release a hole.

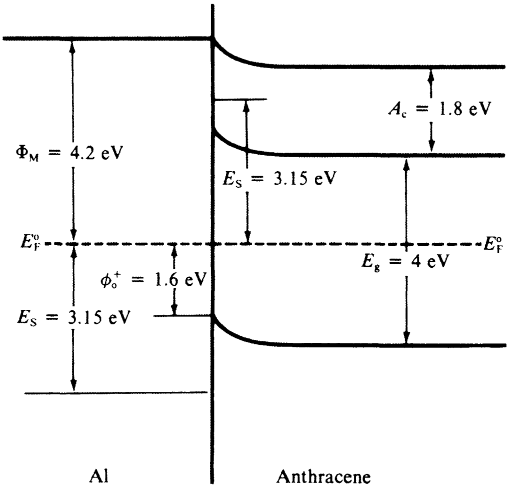

FIG. 7. Energy relationships at the interface of an Al electrode and an anthracene crystal. The singlet energy E_S is shown to be sufficient to cause the injection of either a hole or an electron.

3.3.2.3 Exciton Dissociation at Surface Molecules. Surface-adsorbed dyes are effective as hole injectors into anthracene when they react with singlet excitons. The injection efficiency of a dye depends on the degree of overlap between the singlet fluorescence spectrum and the absorption spectrum of the dye. This overlap determines the efficiency of Förster dipole–dipole transfer (Knox, 1963). The sensitization of charge injection into organic semiconductors finds application in photography and electrophotography. For a more complete discussion of dyes at surfaces, see Gerischer and Willig (1976).

3.3.3 Photovoltaic Effect Consider a system consisting of a layer of a semiconducting material or materials, sandwiched between two opposing electrodes, at least one of which is partially transparent to light. If light is shone through the transparent electrode in the absence of an external field and a voltage difference is produced thereby, then a photovoltaic effect has been demonstrated. One type of photovoltage can result from differences in the diffusivity of the carriers. This is known as the Dember effect and historically was the mechanism for the photovoltaic effect first observed in an anthracene crystal. More important are two other mechanisms for generating a photovoltage. In both cases, advantage is taken of a potential barrier that is present in the dark as a result of differences in the work functions of contacting materials. One barrier is called a Schottky barrier, and is set up between the electrode and the semiconductor (see Schottky Barriers). The other is a *p*/*n*-junction barrier (see Light-Emitting Diodes), set up between two semiconductors in which the majority carriers are holes and electrons respectively.

When all the contacts are made, the Fermi levels equalize and the band bending seen in Fig. 8 is produced. Light absorbed at the Schottky contact causes electrons to be excited into the conduction level, and the built-in potential forces the electrons into the Schottky contact and the holes into the Ohmic contact. The maximum photovoltage obtainable is the potential difference between the Fermi level and the conduction-band edge. In the *p*/*n* junction, metals A and B are Ohmic contacts, and the potential difference at the interface is caused by the depletion of majority carriers that flow into the regions of opposite majority-carrier conductivity. The maximum photovoltage in this case is the potential difference between the upper edge of the valence band and the lower edge of the conduction band.

Typical examples of normally *p*-type organic semiconductors are phthalocyanines and quinacridone pigments. For *n*-type materials, there are pyrylium dyes, and perylene–3,4,9,10 tetracarboxylic acid. The origin of the basic *p* or *n* character of the organic compounds is not well understood. By doping, it is a straightforward matter to create the required majority-carrier type. Unlike inorganic materials, *p*- and *n*-type organic semiconductors are not generated by the introduction of minute amounts of impurities into the pristine semiconductor. The doping is produced by adding mol% amounts of acceptor or donor molecules into the host matrix. Typical acceptor molecules are O_2, halogens, and tetracyanoquinodimethane. For donor molecules, there are NH_3 and phenothiazine (see Borsenberger and Weiss, 1993).

For a Schottky configuration, the cell would be glass/(Au, Ag, or ITO electrode)/(organic semiconductor)/(In or Al electrode). For *p*/*n* cells, the configuration would be glass/ITO/*p*-type/*n*-type)/(Au or Ag). Overall power conversion efficiencies of 2% have been achieved with *p*/*n* junctions and 1% with a Schottky-type cell. In inorganic systems, efficiencies of more than 20% have been attained.

3.4 Carrier Generation of Pairs in the Bulk

As previously stated, the selection of organic semiconductors to be discussed is based on an arbitrary demarcation level of conductivity of $10^{-3}\ \Omega^{-1}\ m^{-1}$. Thus, all the organic compounds with conductivities $>10^{-3}\ \Omega^{-1}\ m^{-1}$ would fall into the class of

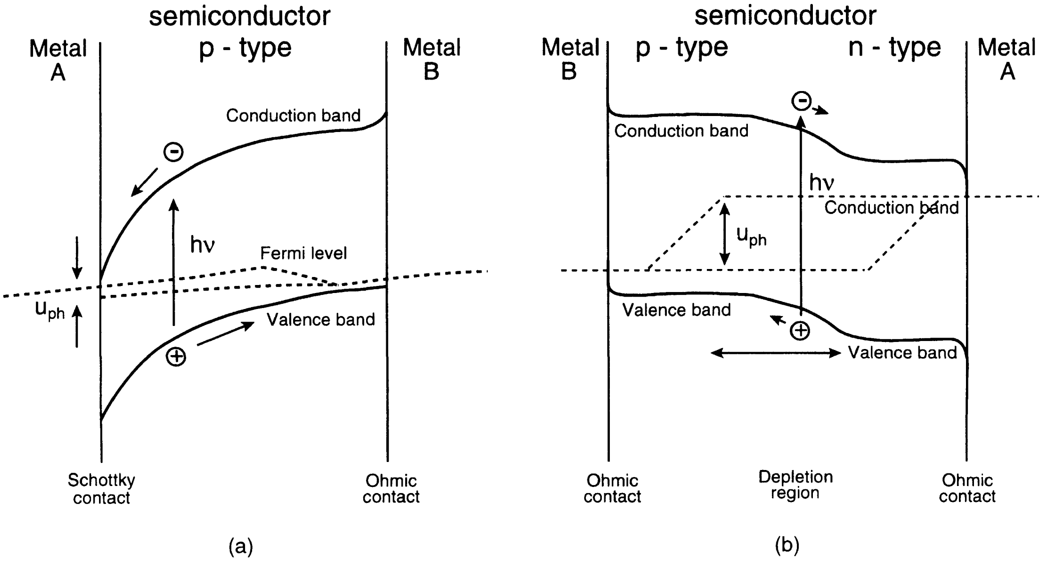

FIG. 8. Schematic of photovoltaic mechanisms. (a) A Schottky contact is formed between metal *A* of low work function and a *p*-type semiconductor. The metal *B* forms an Ohmic contact. (b) A typical *p*-*n* junction. Both metals form Ohmic contacts. The maximum photovoltage in each case is given by U_{ph}. (Redrawn from Wöhrle and Meissner, 1991.)

organic conductors and will not be discussed further. This class includes high-conductivity donor–acceptor complexes such as TTF–TCNQ and doped polymers such as I_2-doped polyacetylene.

3.4.1 Thermal Mechanism All the materials reported on in this section are essentially insulators in the dark, and so intrinsic thermal generation of carrier pairs is negligible.

3.4.2 Photonic Mechanisms

3.4.2.1 Autoionization. The mechanism of bulk photogeneration of carrier pairs in molecular organic crystals has been most extensively studied on anthracene. In the threshold region of photogeneration there is evidence for the direct generation of bound hole–electron pairs, which require thermal activation energy and/or a high electric field for complete separation into pairs of carriers. For photon energies larger than the band gap, a higher excited molecular state is formed, which autoionizes to create a pair of carriers that move apart from each other with excess kinetic energy. However, the efficiency of free-carrier generation is small, $\sim 10^{-4}$, because of efficient carrier recombination (see Sec. 4.1). In passing, it may be noted that the efficiency of carrier generation in crystalline fullerene, C_{60}, is also 10^{-4}. The band gap of this new form of carbon is ~2.2 eV; its hole mobility is 1.74×10^{-4} m^2 V^{-1} s^{-1} and its electron mobility is 0.52×10^{-4} in the same units. The mobility is independent of temperature from 46 to 250 K; this and its other electronic properties are similar to those of other organic molecular crystals.

In addition to carrier generation by the action of a single photon, it is also possible to generate carrier pairs by a multiphoton process involving low-energy photons. The final excited state produced by multiphoton excitation autoionizes to generate the carriers as in the case with single-photon excitation.

In polymers consisting of a saturated backbone and π-conjugated pendant chromophores, e.g., poly(*N*-vinylcarbazole), or a conjugated backbone with π-conjugated pendant chromophores, e.g., poly(*N*-2 propinyl) carbazole, photogeneration proceeds as with monomeric crystals, namely by autoionization followed by Onsager-type geminate recombination (see Sec. 4.1) (for a review, see Nešpůrek, 1993, and Borsenberger and Weiss, 1993).

3.4.2.2 Band-to-Band Transitions. In molecular crystals of the polyacene type, electron and hole bands in ultrapure materials have been demonstrated to exist at $T \sim 4$ K by Karl (1990). At higher temperatures, the thermal energy exceeds the bandwidth and band theory is not applicable. In some polymers, like *trans*-PA and polydiacetylenes (PDA), there is evidence of a direct valence-to-conduction-band transition. While the lowest π–π^* transition in *trans*-PA along the chain may be BB (Sariciftci and Heeger, 1994), in which an intrachain widely separated electron–hole pair is formed, the lowest π–π^* transition in PDA is excitonic. In *trans*-PA, in ≤ 1 ps, the carriers polarize their surroundings, becoming charged solitons, which neutralize and decay into phonons in <1 ps. In PDA, the oppositely charged species undergo geminate recombination in subpicosecond times. Since the lowest π–π^* excitation in polydiacetylene is excitonic, any conductivity that is associated with the absorption edge must be attributed to an external agency that supplies the binding energy of the singlet exciton, which is about 0.5 eV. This could be an impurity or a large electric field.

The lowest π–π^* transition in most polymers is not generally a BB transition (see Bässler *et al.,* 1995). In PPV and derivative polymers, the lowest π–π^* transition is a neutral intrachain exciton. This exciton can become an interchain exciton of CT character, dissociating into carriers under the action of an external field and thermal energy. Geminate recombination in all cases limits the carrier-generation efficiency (see Sec. 4).

In poly(methylphenylsilane), the lowest σ–σ^* transition at 3.5 eV generates a singlet exciton. There is no evidence of polaron formation. The threshold for direct carrier generation is at ~4.6 eV, producing holes and electrons. Only the holes are mobile, and the transport is nondispersive. The efficiency of carrier generation is $\sim 10^{-2}$ (Kepler and Soos, 1993).

There is as yet no unanimity in the field regarding the relative importance of BB (Sariciftci and Heeger, 1994) as opposed to excitonic mechanisms for carrier generation. However, if by a BB transition is meant the generation of an uncorrelated hole–electron pair that is free to dissociate even under the influence of a small external field, and with a weak temperature dependence, this has not yet been observed in organic semiconductors at room temperature.

3.4.2.3 Excitonic Mechanisms. When two singlet excitons come within a critical distance ~1 nm from each other (known as the Förster transfer distance), a transfer of energy from one exciton to the other can occur (see EXCITONS) The final state is a highly excited exciton that possesses sufficient energy to ionize. In anthracene, the singlet energy is 3.15 eV, so that the exciton fusion produces 6.3 eV. This is greater than the band gap of ~4 eV, and is even greater than the ionization energy of the crystal, 5.8 eV. The exciton–exciton fusion rate constant is $\sim 10^{-14}$ m^3 s^{-1} (diffusion-limited) and the probability of ionization is $\sim 10^{-4}$. A remarkably similar excitonic behavior has been observed in the polysilanes.

A singlet exciton can also excite a triplet exciton, yielding an energy of ~4.9 eV. The triplet–singlet collision rate constant is $\sim 5 \times 10^{-15}$ m^3 s^{-1}, with about the same probability of ionization as in the singlet–singlet case. Furthermore, photons can ionize either singlet or triplet excitons. The ionization cross section for the photon–singlet is $\sim 10^{-23}$ m^2, and for the photon–triplet is $\sim 10^{-25}$ m^2.

A direct two-photon ionization process can take place starting at a photon energy of 2 eV in anthracene. For two 3.1-eV photons, the cross section for carrier generation is $\sim 10^{-32}$ m s. These cross sections are typical of other organic molecular semiconductors.

4. CARRIER RECOMBINATION

In organic semiconductors, carrier recombination is a highly efficient process, and it limits the overall efficiency of carrier generation to a degree not approached in inorganic semiconductors. The reason for this is twofold. First, because of the generally low dielectric constant of most organic semiconductors, there is very little shielding of the Coulombic force between the electron and hole of the newly generated *e–h* pair. The attractive force is therefore potent at separation distances of as much as 17 nm in the *e–h* pair. Second, the intermolecular bandwidths for the carriers are narrow, implying low mobilities and frequent carrier-scattering events. This scattering facilitates energy loss in the initially created *e–h* pair with excess kinetic energy. The task of complete separation thus will depend on the external field strength and the temperature.

4.1 Geminate Recombination—Onsager Theory

Geminate recombination is the recombination of an electron with a hole, both of which came from the same molecule. Such recombination in OSCs dominates the recombination process for all but the highest concentration of *e–h* pairs. In contrast to bimolecular recombination between uncorrelated *e–h* pairs, geminate recombination is kinetically first order (see CHEMICAL KINETICS).

Geminate recombination was given a firm theoretical basis by Onsager (1938), who treated the problem as it existed in a liquid, and the theory was applied in a definitive manner to organic solids by Chance and Braun (1973). The Onsager theory gives the probability that a thermalized charge pair will escape geminate recombination and dissociate under the influence of an external electric field of strength E.

The Onsager theory requires a knowledge of the thermalization distance r_{th}, the temperature, and the dielectric constant of the medium. With these parameters and a knowledge of the initial quantum yield of electron–hole pairs, the field and temperature dependence of the free-carrier yield can be calculated. An important quantity is the Coulomb-capture distance r_c. This is the distance of separation of the *e–h* pair at which point the electrical potential energy is equal to the thermal energy. This distance is given by

$$r_c = e^2/4\pi\epsilon\epsilon_0 kT \equiv 2q. \tag{4}$$

The average thermalization distance $\langle r_{\mathrm{th}} \rangle$ is defined as the mean distance traversed by the hot carrier as it comes to thermal equilibrium with its surroundings, and is usually not known. There is a distribution of thermalization distances and no single distribution function fits all occasions. The energy-loss mechanisms for a highly excited initial state vary. Ionization may take place from the initial state, in which case $\langle r_{\mathrm{th}} \rangle$ will increase with the energy of excitation, until an energy-loss mechanism is encountered in the medium, such as a vibrational excitation. The initial state may undergo an internal conversion to some lower-lying state, which then dissociates, in which case $\langle r_{\mathrm{th}} \rangle$ will be independent of the initial energy of excitation.

The Onsager relationship for the low-field case (applicable for most situations) is

$$f(r_{\mathrm{th}},\theta) = [\exp(-A)](1 + AB), \tag{5a}$$

where $A = 2q/r_{\mathrm{th}}$, $B = \beta\, r_{\mathrm{th}}(1 + \cos\theta)$, $q \equiv e^2/8\pi\epsilon\epsilon_0 kT$, $\beta = eE/2kT$, θ is the angle between the radius vector $\mathbf{r}$ and the applied field vector $\mathbf{E}$, and $f(r_{\mathrm{th}},\theta)$ is the probability that the charge pair will dissociate as a function of the separation distance and the angle between the applied field and the radius vector. For an isotropic medium of dielectric constant $\epsilon = 3.02$, and an applied field of 10^5 V m^{-1}, r_{th} would have to be 45 nm to have a probability of 0.7 for escaping geminate recombination. At 10^7 V m^{-1}, r_{th} could be 8 nm and still yield the same probability (0.7) for escape.

To calculate the quantum yield as a function of the external field $\Phi(E)$ using Eq. (5a), it is necessary to have a proper distribution function for r_{th}. If $g(r_{\mathrm{th}},\theta)$ is the probability per unit volume of finding the ejected electron in a volume element $d\tau$ at (θ,r_{th}), then, assuming an isotropic medium and a spherically symmetric $g(r_{\mathrm{th}},\theta)$, the quantum yield would be

$$\Phi(E) = 4\pi\Phi_0(1 + 2\beta q) \times \int_0^\infty [\exp(-2q/r_{\mathrm{th}}]g(r_{\mathrm{th}})r_{\mathrm{th}}^2 dr_{\mathrm{th}}, \tag{5b}$$

where Φ_0 is the primary quantum yield of carrier pairs per absorbed photon and is generally not known. It is possible to get an estimate of Φ_0, which can help distinguish among various functions for $g(r_{\mathrm{th}})$ (see SC, p. 259, and PS, p. 494). The most used functions are the delta function, the Gaussian, and the exponential. For anthracene, in the photon-energy range 4.4–5.2 eV, the values of Φ_0 based on various assumptions in the use of the Onsager equation (see PS, p. 494) are 5×10^{-3} for a delta-function $g(r_{\mathrm{th}})$, 10^{-2} for a Gaussian, and 1.4×10^{-1} for an exponential. For a photon range of 5.4–6.2 eV, the values of Φ_0 are 3.4×10^{-5} for the delta function, 5.2×10^{-3} for the Gaussian, and 3.4×10^{-2} for the exponential. The constancy of Φ_0 over a wide photon-energy range is indicative of a fixed lower-lying state that autoionizes to yield the carriers.

The Onsager theory predicts a value of the ratio of the slope to the intercept (S/I) of a curve such as that shown in Fig. 9. This is given by the expression

$$\begin{aligned} S/I &= 2\beta q/E = e^3/8\pi\epsilon\epsilon_0 k^2T^2 \\ &= 3.38 \times 10^{-7}(298/T)^2\ \mathrm{m\,V^{-1}}. \end{aligned} \tag{6}$$

Taking $\epsilon = 3.23$, the experimental value of the S/I ratio for electrons is $(3.21 \pm 0.12) \times 10^{-7}$ m V^{-1}, and for holes it is $(3.02 \pm 0.08) \times 10^{-7}$ m V^{-1}. The temperature dependence of the value of S/I at a fixed excitation wavelength is in agreement with the Onsager prediction.

The evaluation of $\langle r_{\mathrm{th}} \rangle$ was based on the model shown in Fig. 10 and an assumption of a delta function for $g(r_{\mathrm{th}})$. As may be seen, when using light of energy less than E_g, it is possible to populate intermediate e–h pair states either directly or by autoionization from some higher-lying level. It follows then that there will be a thermal energy of activation for photogeneration defined as E_a^{ph} that must be supplied in order to produce free carriers. If measured as a function of photon energy, one obtains

$$\langle r_{\mathrm{th}} \rangle(h\nu) = e^2/4\pi\epsilon\epsilon_0 E_a^{\mathrm{ph}}(h\nu). \tag{7}$$

The results for anthracene are shown in Fig. 11. A more sophisticated modification of the Onsager theory was made by SC (1994), who included the effect of electric field and temperature on the thermalization distance. The Onsager theory, however, is still in wide use.

In dealing with geminate recombination in some polymers, it is found that another theory propounded by Onsager in 1934 is more appropriate (Braun, 1984). This theory gives the same low-field S/I ratio as the 1938 theory, but it predicts the increased dissociation efficiency of a CT state in the presence of an electric field. Such near-neighbor CT states are often the initial photogenerated

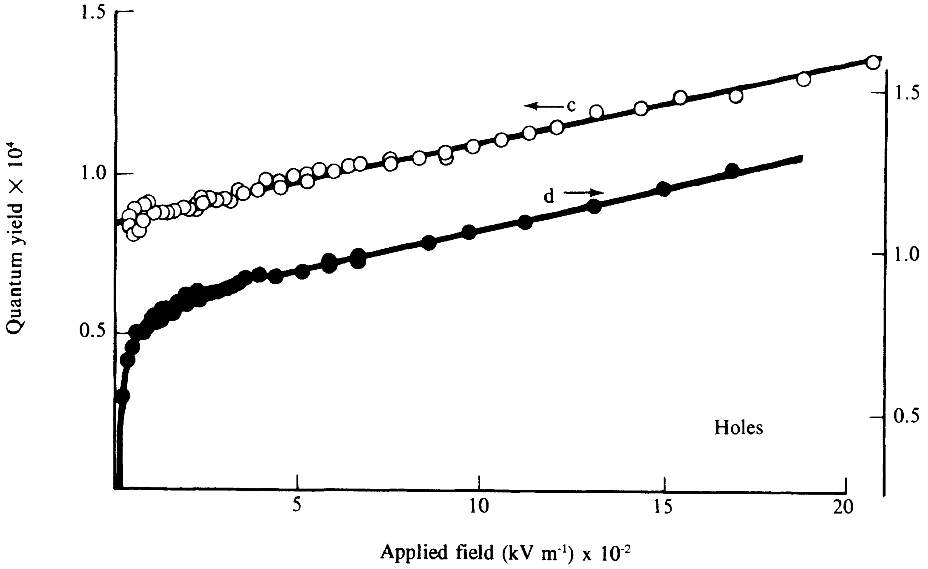

FIG. 9. Hole-carrier quantum yield at 255 nm as a function of the applied field. Curve *c* is for a previously unirradiated crystal sample; in curve *d*, the trapped electrons, left behind from the earlier curve-*c* experiments, reduce the low-field free-hole yield. Note that for clarity of presentation, the origins of curves *c* and *d* are different. (Redrawn from Chance and Braun, 1973.)

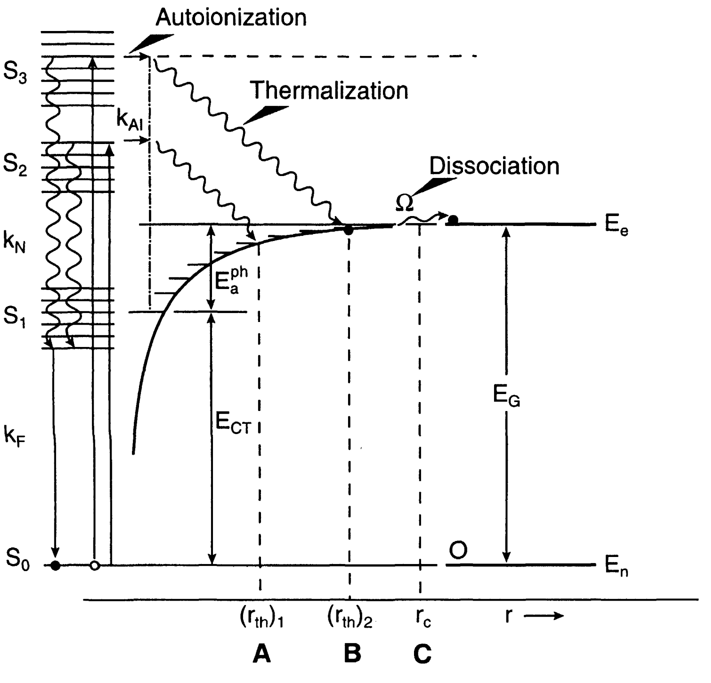

FIG. 10. Schematic diagram of the multistep photogeneration process in organic molecular crystals. Molecular levels in the crystal are shown at the left for scale purposes. The ground state S_0 is shown with one electron in some excited state S_n. The various processes of autoionization (AI), thermalization, and final dissociation of an electron–hole pair state are depicted. The curve depicts the Rydberg-type energy levels of the CT exciton. These levels can be populated directly in the subthreshold region, and by thermalization as shown. E_{CT} is the energy of the CT state, and E_a^{ph} is the activation energy required to dissociate the CT state. *A*, *B* represent various thermalization distances; *C* is at the Coulomb-capture distance. E_G is the band gap. (Adapted from SC, p. 256.)

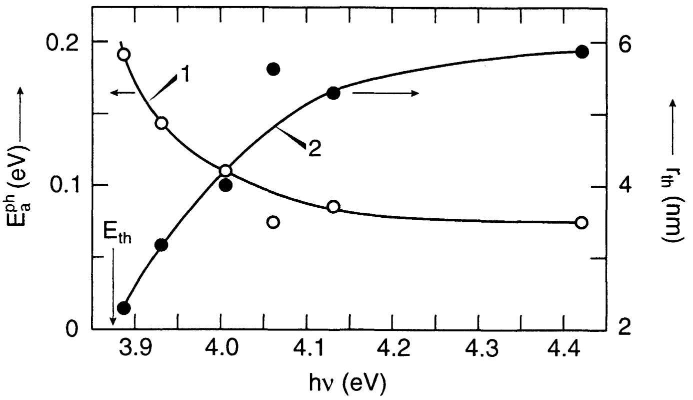

FIG. 11. Experimental $E_a^{ph}(h\nu)$ dependence in an anthracene crystal (curve 1) and the corresponding $r_{th}(h\nu)$ (curve 2) from Eq. (7) (redrawn from SC, p. 265).

states (see Borsenberger and Weiss, 1993, p. 91).

The original Onsager theory was based on a condensed-phase continuum. However, a crystal is not a continuum, and other models have been proposed (for references, see Borsenberger and Weiss, 1993, p. 93). In addition, the original Onsager theory was developed for 3D materials, whereas in polymers such as *t*-PA and PPV, the initial carrier generation takes place within a single 1D polymeric chain. The thermalization distance for the carriers in 1D may exceed the Coulomb capture radius and still the carriers will not escape geminate recombination. For such situations, a 1D Onsager theory has been derived (see Gailberger and Bässler, 1991, for discussion and references).

4.2 Geminate Recombination—Time-Dependent Theory

The Onsager (1938) theory of geminate recombination is a steady-state theory, and has as one of its boundary conditions that recombination occurs with unit efficiency when the distance separating the *e–h* pair reaches a specific minimum value. Both of these conditions were removed by Hong and Noolandi (1978; see also Noolandi, 1982), who derived an analytical time-dependent theory with variable recombination efficiencies at the origin. In addition to the analytical theory, there have been computer simulations of time-dependent geminate recombination by Ries *et al.* (1983).

4.3 Bimolecular Recombination

Bimolecular recombination following intrinsic carrier photogeneration in organic semiconductors is relatively unimportant compared with geminate recombination, except under conditions of extremely high light intensity ($\sim 10^{30}$ m^{-2} s^{-1}). However, bimolecular recombination takes place when electrons and holes are injected separately into the same organic semiconductor (double injection), and brought together by the electric field. The rate of the recombination is given by $\gamma_{eh}np$, where γ_{eh} is the bimolecular rate constant, and n, p are the concentrations of holes and electrons, respectively.

Bimolecular recombination was treated theoretically by Langevin under the restriction that the mean free path of the carriers be less than the Coulomb-capture radius r_c. This condition is amply met in organic semiconductors. The calculated value for γ_{eh} in anthracene is 10^{-12} m^3 s^{-1}, in good agreement with the experimental result, $(3 \pm 2) \times 10^{-12}$ m^3 s^{-1}.

4.3.1 Electroluminescence The most dramatic result of bimolecular recombination is the production of light. This radiative recombination is referred to as electroluminescence (EL) (see ELECTROLUMINESCENCE). The first observation of EL caused by radiative recombination in organic crystals was made by Pope *et al.* (PS, p. 93). Double-injection EL in anthracene was produced by Helfrich and Schneider (1965), who used solutions of negative and positive anthracene ions as cathode and anode respectively. The EL is essentially identical with the normal fluorescence.

In addition to the organic molecular crystals, EL has been observed in composite polymeric systems consisting of a dye (e.g., squarylene) dispersed in a plastic matrix (e.g., polycarbonate), and in homogeneous polymers (e.g., polyphenylenevinylene). An outstanding feature of many organic materials is their high fluorescence efficiencies. Charge-injecting electrodes are required to generate the *e–h* pair that will recombine to fluoresce. Electron-injecting electrodes are chosen from low–work-function materials (e.g., Na, Mg, Ca) and hole-injecting electrodes are chosen from high–work-function materials [e.g., Au, CuI, indium tin oxide (ITO)]. At least one transparent electrode is necessary to permit the transmission of the luminescence.

A dominant characteristic of *e–h* recombination in organic semiconductors is that the *e–h* pair passes through an intermediate exciton or other intergap state, which then decays radiatively into a singlet state, or nonradiatively into any one of three triplet states. Thus the maximum EL efficiency per injected carrier pair is 25%, which is a serious limitation (see also Sec. 7.2.)

5. CARRIER TRANSPORT

Transport in inorganic semiconductors generally is band-type; the electron and hole wave functions are delocalized; holes in Ge have a mean free path (at RT) before scattering of approximately 100 nm, a value considerably larger than nearest-neighbor lattice spacing of 0.25 nm. If the carrier wave functions are highly localized, carriers are strongly scattered and motion through the lattice is random-walk–like with short hopping distances.

For each extreme type of motion, the mobility, defined as the drift velocity per unit electric field, has a different magnitude and temperature dependence. In wide conduction bands, the mobility $\mu > 10^{-4}$ m^2 V^{-1} s^{-1} and $\mu \propto T^{-n}$, $n > 1$. For strongly localized carriers, $\mu \ll 10^{-4}$ m^2 V^{-1} s^{-1} and $\mu \propto \exp(-E/kT)$ where E is an activation energy; the band model stresses the collective nature of the conducting states whereas the hopping, or diffusive, model emphasizes the molecular character. Molecular crystals, such as the polyacenes, have mobilities between these limiting cases and exhibit transport characteristics influenced strongly by the anisotropy of the crystal lattice. In polyacenes near room temperature, $\mu \propto T^{-n}$, $n > 1$, whereas the carrier's mean free path λ inferred from both theory and experiment is only of the order of a lattice constant, a limit where temperature-activated hopping is required.

Only recently has a phenomenological model been available capable of quantitatively accounting for such a carrier dependence on temperatures (from 4 to 300 K) and for electric fields $\approx 10^5$ up to 10^8 V m^{-1}, thereby explaining both hot- and thermalized-carrier transport in polyacene crystals. In this model, the quasiparticle effective mass depends significantly on temperature (varying from a few times the electron's free mass m_e at near zero temperature to well over $400m_e$ at room temperature), the sign of the carrier, the direction of transport, and the electric field, at least in the high-field regime (SC, p. 318).

In most polymers, molecular glasses [such as amorphous films of tetracene and other polyacenes (Bässler, 1981)], and doped polymers, charge transport is described by a formalism based on disorder. Disorder is characterized by the terms *diagonal* and *off-diagonal,* as well as *static* and *dynamic.* Diagonal disorder owes its origin to the nonhomogeneous local environment around any given molecular site due, as examples, to chemical impurities, chain termination, or pendant-group orientations, and to thermally induced motion. This introduces a distribution of site energies. This situation in turn leads to a distribution of intersite distances, and therefore transfer energies J. Such effects are termed off-diagonal disorder. It is sometimes convenient to distinguish static (diagonal or off-diagonal) disorder, due to environmental inhomogeneities, from that produced by thermal fluctuations, which is referred to as dynamic (diagonal or off-diagonal) disorder. As a result of these types of disorder, carrier hopping occurs through a manifold of localized states having different positional disorder and different localized energies.

A fundamental distinction between polaron models and disorder formalism is that in polaron transport, the activation energy reflects the deformation energy (large electron–phonon coupling) whereas in the disorder formalism, the electron–phonon coupling is weak and the activation energy arises from transport between sites characterized by static disorder.

5.1 Types of Transport

5.1.1 Band Motion When the nearest-neighbor overlap energy J is large relative to other interactions for the carrier, the eigenfunctions are approximated by Bloch-band states. In this case, in the constant–relaxation-time (τ_{rel}) approximation,

$$\mu = e\tau_{\mathrm{rel}}/m^*, \tag{8}$$

where m^* is the constant effective mass and need not be equal to m_e because the lattice phonons renormalize the bare bands. In some solids this can enhance m^* relative to m_e as in Li metal where $m^*/m_e = 1.4$; and in others, it gives an $m^* < m_e$, as in InSb where m^* is isotropic and $\sim 0.014\ m_e$. The temperature dependence of μ is reflected primarily in those microscopic processes that determine τ_{rel}. For acoustic-phonon scattering, $\tau_{\mathrm{rel}} \propto T^{-n}$ ($n > 1$), whereas for optical-phonon scattering $\tau_{\mathrm{rel}} \propto n^{-1}(\omega)$ where $n(\omega)$ is the Bose–Einstein distribution. In general, for band-type behavior, the temperature dependence of μ is of the form aT^{-n} (see PHONONS IN CRYSTAL LATTICES).

5.1.2 Electric-Field Effects Equation (8) predicts a carrier drift velocity linear in applied field E, provided m^* and τ_{rel} are independent of E. However, low-temperature measurements in ultrapure polyacene crystals have shown a nonlinear field dependence for the drift velocity v_d. For $E < 10^6\ \mathrm{V\ m^{-1}}$, $v_d \propto \sqrt{E}$, whereas at higher field, v_d saturates. This initial field-dependent decrease in mobility can be rationalized as due to field-enhanced acoustic-phonon–carrier scattering. For band transport, Shockley has shown that

$$\tau_{\mathrm{rel}}^{-1}(E) = \tau_{\mathrm{rel}}^{-1}(0) \times \{\tfrac{1}{2} + [\tfrac{1}{4} + (3\pi/32)(\mu_0 E/\mu_1)^2]^{1/2}\}^{1/2}, \tag{9}$$

where μ_0 is the low-field mobility and μ_1 is the longitudinal sound velocity, typically of the order of (0.2 to 0.5) × $10^3\ \mathrm{m\ s^{-1}}$. At higher field strengths, the energy the carrier acquires from the external field between successive scattering events can exceed the crystal optical-phonon energies, $\hbar\omega_{\mathrm{ph}}$; hence the emission of crystal phonons imposes a saturation velocity on the charge carriers,

$$(V_d)_s = (\hbar\omega_{ph}/2m^*)^{1/2}, \tag{10}$$

as has been demonstrated in polyacene crystals.

5.1.3 Hopping Motion When static disorder is negligible, and when J (the nearest-neighbor interaction energy) is comparable to the site-energy differences due to dynamic disorder, the effect of carrier–lattice interactions needs to be considered in forming the first-order transport states. Because of electron–lattice interactions (polarization), a carrier's neighboring atoms are displaced from their equilibrium positions, and these distortions form a potential well around the carrier in which it is (partially) self-trapped. In organic molecular solids, J is of the order of 10^{-2} eV and is small compared with the polarization energy. The composite entity consisting of the carrier and the local distortion is called a polaron (see POLARONS). The size of the polaron will be strongly reflected in the temperature dependence (and field dependence) of hopping models, and many polaron-transport models have been formulated and reported in the literature.

In the adiabatic limit, the effects of lattice polarization impart to the carrier an effective mass

$$m^* = m_e \exp[S(T)], \tag{11}$$

where S is a Huang–Rhys factor that depends on temperature and the coupling constants of the carrier to the phonons. $S(T)$ increases with increasing temperature, and correspondingly, the polaron bandwidth narrows. At room temperature, transport is more appropriately considered as a sequence of hops through the lattice; in this limit $\mu \propto \exp(-E/kT)$ where E depends on temperature and crystallographic direction, and is approximately one-half the polaron binding energy (E_p); i.e., $E \approx E_p/2 - J$. However, in crystalline molecular solids, it is frequently found that the mobility varies as T^{-n}, and in some molecular crystals such as naphthalene and anthracene, the electron mobility in the high-T region is practically T independent in the c' direction. This behavior is not consistent with band theory or polaron transport.

A rigorous microscopic theory for the dependence of μ on T is still lacking. However, the μ-vs-T behavior has been rationalized (SC, 1994) by considering the carrier as a "nearly small" polaron that moves via tunneling from one lattice site to another. In this picture, the activation energy decreases with increasing temperature, and the dependence of μ on T is modeled by the variation of m^* and $\tau_{\rm rel}$ with T and E; m^* varies from $\sim 10m_e$ to $\sim 400m_e$ from $T \sim 40$ K to room temperature. The physical model for transport is therefore one of strong scattering ($\lambda \sim a$) coupled with tunneling through a highly anisotropic local-potential barrier.

5.1.4 Transport in Polymeric Materials

Polymers generally do not form crystals that are large enough to study as such, and exhibit transport properties considerably more diverse than single crystals. A notable exception is the class of polydiacetylenes. These polymers are prepared by forming a single crystal of the monomer, after which the crystal is polymerized to produce the final, essentially single-crystal polymer. The mobility of the carriers is typical of those measured in the polyacenes, namely, of the order of 10^{-4} m^2 V^{-1} s^{-1}.

Although small regions of polymers may be crystalline, the overall response of polymers is typical of an amorphous system having a manifold of localized site energies and intersite distances within a highly anisotropic electronic structure. The chainlike structure provides for strongly covalent bonding along the chains and considerably weaker van der Waals interactions between chains. For polyacetylene, the bandwidth of conducting states along the chain is of the order of 10 eV but is only one hundredth of this perpendicular to the chain. A distinct feature of most polymers is their lack of long-range order [static and dynamic diagonal and off-diagonal disorder (see Sec. 5.1.3)], small intermolecular interaction energies, and large polarization energies (~ 1 eV). Thus the charge-transporting sites (which can be either molecules or a polymer-chain segment) have site energies ϵ, and transfer energies J, that are characterized by distribution functions. Most theoretical studies of charge transport in disordered molecular solids and molecularly doped polymers have assumed that the inhomogeneous broadening of the hopping-site energies ϵ has the form

$$\rho(\epsilon) = (2\pi\sigma)^{-1/2} \exp(-\epsilon^2/2\sigma^2), \qquad (12)$$

as this is the shape usually observed in the absorption (and fluorescence) bands of disordered organic solids. Here, ϵ is measured relative to the center of the distribution. Equation (12) neglects polaronic effects and assumes that adjacent site energies are uncorrelated. This distribution-function form for the transfer energies is admittedly oversimplified but contains a large measure of validity.

If site-energy fluctuations are sufficiently large that they overcome the delocalizing effects of the (average) transfer energy $\langle J \rangle$, then the ordinarily distinct boundaries of the valence and conduction bands become diffuse, and energy states appear as a tail extending into the forbidden gap. These gap states overlap each other weakly if at all, and injected charge becomes localized. The quantitative localization criterion (Duke, 1982) can be written as

$$\Delta > cz\langle J \rangle, \qquad (13)$$

where Δ denotes the rms deviation of $\rho(\epsilon)$, z is the number of nearest neighbors, and c is a dimensionless constant of order unity that depends on the dimensionality of the system. For polymers and molecular glasses the static polarization-induced fluctuations in ϵ nearly always produce localized states. The term Fermi glass is often used to denote systems satisfying the criteria given by Eq. (13), i.e., molecular solids in which all electronic states are localized. Experimental evidence supporting the Fermi-glass description of polymeric systems comes from an analysis of the gaseous- and solid-phase valence photoemission spectra (see Duke, 1982).

A characteristic of molecular glasses at short times (Bässler, 1981) and molecularly doped polymers is the strong electric-field and temperature dependence of their carrier mobilities. Despite the wide diversity in both chemical and physical structure of Fermi glasses, most display a mobility μ for $T_0 > T$ varying as

$$\mu(E,T) = \mu_0 \exp[-(\Delta_0 - \beta E^{1/2})/kT_e], \tag{14}$$

where Δ_0 is a low-field activation or barrier energy, reflecting the system static diagonal disorder, β is a constant, k is the Boltzmann constant, and $T_e^{-1} = T^{-1} - T_0^{-1}$; T_0 is a characteristic temperature, typically of the order of the glass-transition temperature.

It appears that in polysilanes (and polygermanes) the dependence of μ on E is insensitive to the nature of the pendant groups (either aromatic or aliphatic) as well as the molecular weight of the polymer. This observation has been taken to mean that only the long-chain backbone is involved in the transport. The observation of hopping behavior, despite the delocalized σ orbitals, indicates that the transport states do not extend over the entire chain, but consist of a population of short segments interspersed among longer segments. These short segments provide relatively localized sites. A fit of lnμ(hole) versus $E^{1/2}$ for films of poly(methylphenylsilylene) (PMPS), a saturated-backbone polymer with pendant groups (CH_3 and benzene), is shown in Fig. 12. (*Note:* The nomenclature in the literature concerning the Si-backbone polymers can be confusing. The above-named compounds are referred to by other authors as silanes instead of silylenes; the same equivalence exists between the germylenes and the germanes.) The square-root dependence of μ on E is also characteristic of hole mobilities in poly(methyl-cyclo-hexylsilylene), poly(methyl-*n*-propylsilylene), and poly(methyl-*p*-methoxy-silylene).

The functional dependence of ln μ on E and T given in Eq. (14), initially used for poly(*N*-vinylcarbazole) (PVK), is by no means universal, and there is a lack of consensus as to the origin of this field and temperature behavior. However, to discriminate unequivocally between square-root and linear field dependence requires a span of E measurements greater than one sees in the μ-vs-E plots. Similarly, in most studies, the temperature range investigated is narrow and thus the exp(T^{-1}) dependence can easily be interpreted as an exp(T^{-2}) dependence. In particular, the disorder formalism of Bässler (see Sec. 5.1.6) predicts a mobility $\mu \propto \exp(-T_0^2/T^2)$ for $T_0 < T < T_g$, where T_g is the glass-transition temperature and T_0 is the temperature associated with the onset of dispersive transport (see Sec. 5.1.5). In doped polymers, the range of temperatures is narrow, so that the distinction between T^{-1} and T^{-2} behavior is small. The temperature dependences of the slopes of lnμ vs $E^{1/2}$ for poly(methylphenylsilylene) (PMPS) plotted against T^{-1} are illustrated in Fig. 13. The intercept (at 408 K) is the characteristic temperature T_0 for PMPS.

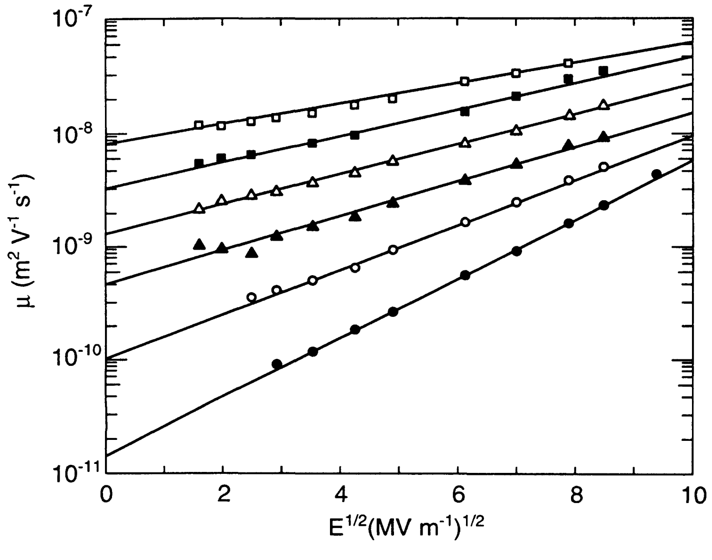

FIG. 12. Hole mobility vs $E^{1/2}$ in PMPS films at various temperatures: ●, 214 K; ⊙, 234 K; ▲, 254 K; △, 274 K; ■, 294 K; □, 314 K (redrawn from Abkowitz *et al.*, 1991).

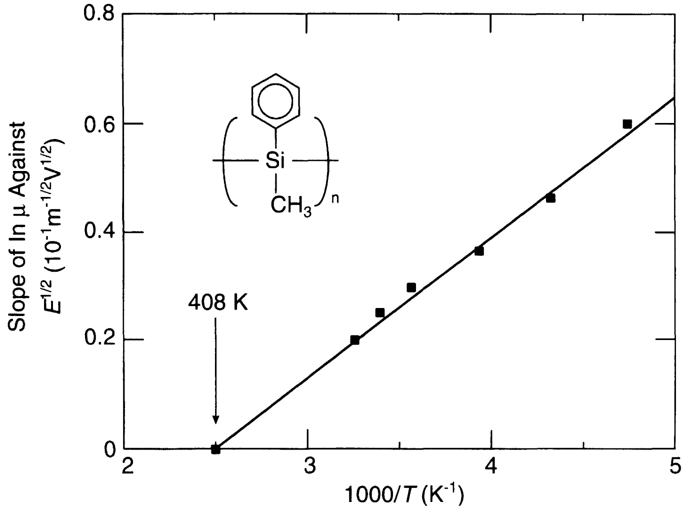

FIG. 13. Temperature dependence of the slopes of lnμ vs $E^{1/2}$. The intercept at 408 K is the characteristic temperature T_0. PMPS shown in insert. (Redrawn from Abkowitz *et al.*, 1991.)

There is as yet no satisfactory theory to rationalize the functional form $\Delta_0 - \beta E^{1/2}$ in Eq. (14), and this is a major unsolved problem in this field.

In contrast to the extensive data on hole transport, there are considerably fewer studies of electron transport because of the difficulties associated with trapping; 0_2, an electron trap, is an almost unavoidable chemical impurity. In many systems, the field dependence is similar to that of hole mobilities, e.g., mixtures of 3,3′-dimethyl-5,3′-di-*t*-butyl diphenoquinone (DPQ) in a polymer host, such as polystyrene or polycarbonate.

In addition to holes and electrons, carriers in polymers can consist of charged solitons and bipolarons. In *trans*-PA, the neutral-soliton (see Sec. 2.1.1.2) diffusion coefficient is estimated to be 10^{-10} to 10^{-12} m^2 s^{-1}. It has a spatial extent or size (i.e., a range of lattice distortions associated with the defect) of the order of 10–14 lattice constants, and has an effective mass $\sim 5m_e$. The neutral soliton has an associated spin of $\frac{1}{2}$. Upon doping with donors or acceptors (residing between polymer chains) a soliton can become negatively or positively charged by the transfer of an electron to (or from) its unpaired orbital. The charged soliton is spinless. The literature values of the π–π^* band gap vary from 1.4 to 1.8 eV and the neutral-soliton level is situated in the middle of the gap. Figure 14(a) illustrates the mid-gap positions of a soliton in the neutral, positively, and negatively charged states. When Coulomb interactions are included, the charged-soliton level is no longer at the center of the gap.

In addition to solitons, both *trans*- and *cis*-PA can support polaron states when either electrons or holes are injected into the polymer. Figure 14(b) illustrates the position of polaron states relative to the bands as two states symmetrically disposed about the center of the gap. For the hole polaron, the

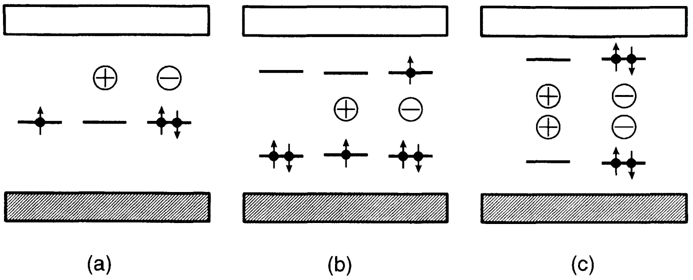

FIG. 14. Positions of the energy levels of (a) solitons in the neutral, positively, and negatively charged states; (b) polarons in the positively and negatively charged states; (c) bipolarons charged positively and negatively within the band gap.

lower level is singly occupied and the higher level is empty. For the electron polaron, the lower level is doubly occupied by antiparallel spins and the upper level is singly occupied. In *trans*-PA, it is energetically favorable for two neighboring polarons in the same chain to interact and form two like-charged solitons (kink and antikink charged states). However, in materials with a nondegenerate ground state, such as *cis*-PA, a widely separated kink–antikink pair is energetically unfavorable (Brédas and Street, 1985); instead, a correlated pair of polarons can form called a bipolaron. Their energies are located within the gap as illustrated in Fig. 14(c). The bipolaron is stabilized by the presence of the strong local lattice distortion, which screens the repulsive Coulomb force in a manner analogous to that exhibited by the Cooper electron pairs responsible for superconductivity (see SUPERCONDUCTIVITY, LOW-TEMPERATURE). In *trans*-PA, the interchain hopping of a single charged soliton necessitates surmounting a relatively large potential-energy barrier because of the large number of carbon atoms on each chain that must relax to new equilibrium positions. For this reason, a pair of closely spaced solitons can more easily hop from one chain to another chain since a smaller number of carbon atoms on each chain have to relax. This transport process is illustrated in Fig. 15.

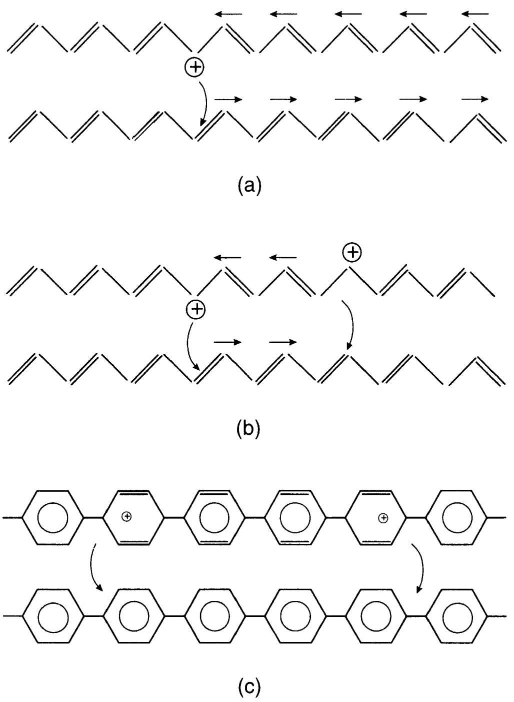

FIG. 15. Schematic of motion of a soliton and bipolaron between polymer chains. (a) Interchain hopping of a soliton in *trans*-polyacetylene; arrows represent those carbon atoms that move. (b) Interchain hopping of a bipolaron. (c) Interchain hopping of a bipolaron in poly(*p*-phenylene).

5.1.5 Dispersive Transport When a delta-function distribution of charge drifts across a sample under the influence of an electric field, its shape alters through thermal spreading and also as a result of the presence of molecular disorder. In the latter instance, the transition probabilities from site to site vary. The arrival of the leading edge of the charge packet at the collecting electrode results in a decrease in the measured current, and a plot of current vs time can exhibit a well-defined inflection point (see Fig. 16). Transport exhibiting this behavior is termed nondispersive. If, however, the time required for the mean velocity of the carriers to reach steady state is comparable to or exceeds the mean transit time, then the plot of current vs time does not show a well-defined inflection point. This transport behavior is called dispersive (see Borsenberger and Weiss, 1993). At sufficiently low temperature, all molecular glasses and doped polymers display dispersive transport on account of the enhanced importance of site-energy distributions and spatial-disorder energy variations relative to kT.

It has been found that by plotting the logarithm of the photocurrent vs the logarithm of the time, an inflection point can be distinguished that can be associated with the arrival time of a characteristic segment of the carrier packet. A few models have been proposed to rationalize this dispersive behavior, the most significant being those referred to as Scher–Montroll, multiple trapping, and Bässler formalism (see Borsenberger and Weiss, 1993). A brief description of the Bässler formalism is now presented.

5.1.6 Bässler Formalism Transport in disordered molecular solids and molecularly doped polymers is strongly influenced by the effects of diagonal and off-diagonal disorder.

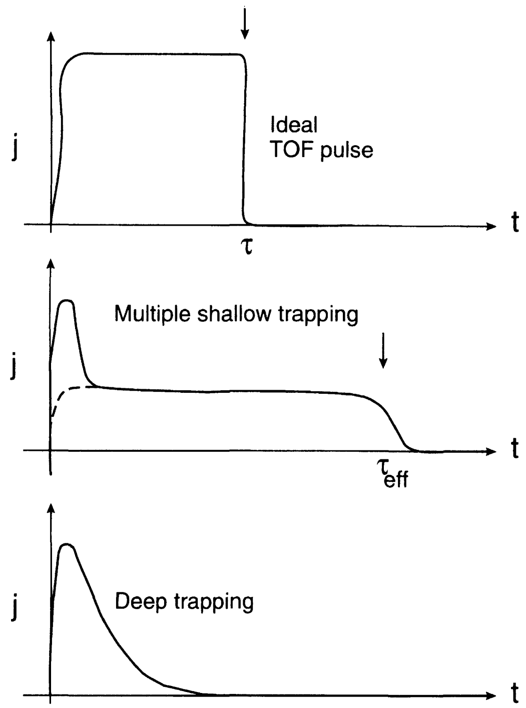

FIG. 16. Top curve: ideal time-of-flight (TOF) current-pulse shape; arrow shows the arrival (transit) time τ_t of the carriers at the collecting electrode. Middle curve: multiple–shallow-trapping case with the initial spike decay representing the trap-filling process, followed by the equilibrium trapping–detrapping process. Bottom curve: deep trapping, with no defined transit time. For all curves, the ordinate is the current and the abscissa is time, in arbitrary units. (Redrawn from Karl, 1990.)

Lacking analytical solutions, most investigators have resorted to Monte Carlo simulation (Bässler, 1993; see also Monte-Carlo Methods). In this approach, one adopts a Gaussian distribution for the hopping-site energies [as given by Eq. (12)] and assumes the jump rate between sites i and j separated by a distance R_{ij} with energies ϵ_i and ϵ_j to be of the form

$$\nu_{ij} = \nu_0 W_{ij} \times \begin{cases} \exp[-(\epsilon_j - \epsilon_i)/kT] & \text{for} \quad \epsilon_j \geq \epsilon_i, \\ 1 & \text{for} \quad \epsilon_j \leq \epsilon_i, \end{cases} \tag{15}$$

where ν_0 is a frequency factor. Implicit in Eq. (15) are the assumptions of weak electron–phonon coupling (i.e., polaronic effects are negligible) and that downward jumps in energy are unaffected by an electric field. Calculations assume that the variation in the intersite electronic wave-function overlap factor W_{ij} is exponential,

$$W_{ij} = \exp(-\gamma_{ij} R_{ij}/a), \tag{16}$$

where a is the lattice constant. The parameter γ_{ij} is a dimensionless random variable describing off-diagonal disorder effects (i.e., a measure of the inverse wave-function decay constant). Monte Carlo calculations (see MONTE CARLO METHODS) based on the assumption that γ_{ij} is a simple sum of uncorrelated Gaussian distributions, $\Gamma_i + \Gamma_j$, each characterized by a variance Σ, predict mobilities described well by the analytical expression

$$\mu = \begin{cases} \mu_0 \exp(-2\hat{\sigma}^2/3) \exp[C(\hat{\sigma}^2 - \Sigma^2)E^{1/2}] & \text{for} \quad \Sigma > 1.5, \\ \mu_0 \exp[C(\hat{\sigma}^2 - 2.25\Sigma^2)E^{1/2}] & \text{for} \quad \Sigma < 1.5, \end{cases} \tag{17}$$

where $\hat{\sigma} = \sigma/kT$; σ is the variance for the site-energy distribution function [Eq. (12)], $\hat{\sigma}$ is referred to as the disorder parameter, and C is a constant with the value 3×10^{-4} $\text{m}^{1/2}$ $\text{V}^{1/2}$. Figure 17 illustrates the good agreement between the calculated constant C and that inferred from the slope of β ($= C\hat{\sigma}^2$) vs $\hat{\sigma}^2$ for DPQ-doped polystyrene (PS) or polycarbonate (PC), C(experiment) = 2.9×10^{-5} $\text{m}^{1/2}$ $\text{V}^{1/2}$. The decrease in mobility of the carrier in PC compared with PS can be attributed to the increased disorder associated with the carbonyl groups in PC. The parameter $\hat{\sigma}$ is of central importance to this theory because together with a knowledge of the sample thickness, it can be used to predict the temperature at which the mobility transients become dispersive. The nondispersive transport regime is one in which the mean carrier-arrival time in a time-of-flight experiment is independent of the sample thickness; this is no longer true in the dispersive transport regime.

The assumption that $\gamma_{ij} = \Gamma_i + \Gamma_j$ implies

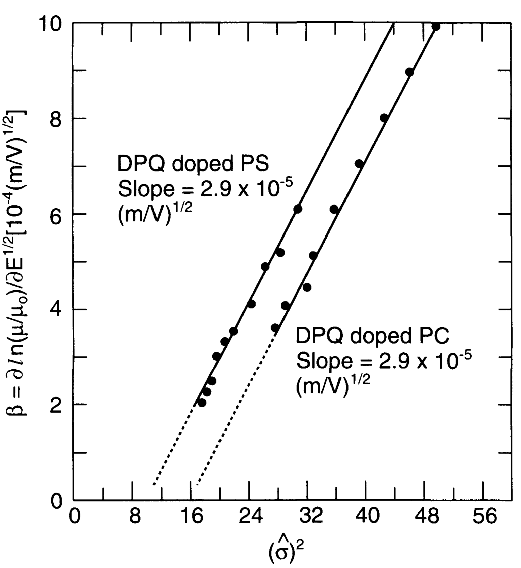

FIG. 17. β vs $\hat{\sigma}^2$ for a mixture of two asymmetrical diphenoquinone (DPQ) compounds in polystyrene (PS) and a polycarbonate (PC). The DPQ concentrations were 30%. Here, $\hat{\sigma} = \sigma/kT$ and $\beta = \partial \ln(\mu/\mu_0)/\partial E^{1/2}$. (Redrawn from Borsenberger and Rossi, 1992.)

a strong correlation between all transitions to and from a given site and omits factors of probable importance. The proper inclusion and estimate of the effects of off-diagonal disorder and, in addition, polaronic effects in carrier transport in disordered molecular solids and molecularly doped polymers constitute a major task (Gartstein and Conwell, 1994).

Nevertheless, the Bässler formalism does account for a large body of information on transport in polymers. In particular are the following:

1. its correct predictions of the field dependence of the mobility in the high-field regime;
2. the observed value of the universal constant C [see Eq. (17)], which is the slope of the plot of $\ln\mu/E^{1/2}$ vs $\hat{\sigma}^2$;
3. the temperature dependence of the mobility; and
4. the transition from nondispersive to dispersive transport (the shape of the log current vs log carrier transit time), thereby generalizing the Scher–Montroll formalism.

5.1.7 Positron Mobility As noted, the motion of an electron through a lattice involves interactions with local perturbations, including those produced by impurity molecules or lattice defects. A useful method for obtaining an estimate of the microscopic mobility in insulators with a high density of electron traps is provided by the mobility of positrons (PS, p. 369; Mills and Karl, 1993). That the positron drift mobility is an upper bound to the electron mobility can be rationalized as follows: For positrons, a chemical impurity does not act as a trap because there is no stable empty orbit that the positron can occupy to form a temporary ion, as would be the case with an electron. The chemical impurity would act as a scattering center because it distorts the local electric field. Just as is the case with electrons, positrons can become trapped in vacancies and structural imperfections if polarization energy is gained. Thus if the positron mobility is significantly larger than the electron mobility, it can be concluded that the electron mobility is chemical-impurity dominated; this is the most frequently encountered case in organic semiconductors.

That μ_p can be extremely large in aromatic solids was shown by the Mills and Karl (1993) measurements in high-purity monoclinic α-perylene single crystals. Over a temperature range 100–350 K, $\mu_p \propto T^{-1}$, a clear departure from the $T^{-3/2}$ dependence expected if the mobility was acoustic-phonon limited and relatively independent of temperature for $T < 50$ K. They obtained a value of ~0.1 $m^2 V^{-1} s^{-1}$, quite similar to their measured value of 0.09 $m^2 V^{-1} s^{-1}$ measured for the electron mobility in naphthalene at 10 K. The limit on μ_p was ascribed to the presence of residual impurities. In anthracene at 300 K in highly purified material, $\mu_p \sim 0.03$ $m^2 V^{-1} s^{-1}$, 300 times the room-temperature mobility. This large value of μ_p can be shown to imply that positrons move in extended states with a mean free path of ~8.5 nm, approximately 15–20 times the lattice constant.

5.2 Steady Current Flow

5.2.1 Basic One-Carrier Theory A simplified theory of space-charge–limited currents (SCLC) was introduced into the field of organic crystals by Helfrich and Mark (1963).

The total current that flows in an insulator is composed of two components, one driven by the electric field and the other by carrier-concentration gradients. These components are referred to as drift and diffusion currents, respectively. In the simplified theory, consideration of diffusion currents is omitted. For a more complete theory including diffusion, the work of Bonham and Jarvis (1978) should be consulted.

The nature of the contact to the organic semiconductor (OSC) is of major importance in determining the qualitative and quantitative response of the current to an external field, both in the dark and in the presence of absorbed light in the contact region (see Sec. 3.3). A contact can change its characteristics under the influence of an electric field. At high field strength, an Ohmic contact can be exhausted, leading to an emission-limited contact (limited by the rate of injection from the electrode). The energetic description of an Ohmic contact and its depletion is shown in Fig. 6.

5.2.2 Space-Charge–Limited Currents (SCLC) By definition, the voltage dependence of the SCLC current in all materials provided with Ohmic contacts is the same. For a trap-free insulator (or semiconductor), or for one containing shallow traps (trap depth from the transport level ≤ 0.3 eV) in which only one type of carrier is mobile, and with neglect of diffusion current, this SCLC is

$$J = (9/8)\theta\mu\epsilon\epsilon_0 V^2/L^3, \qquad (18)$$

where the trapping ratio $\theta = n_f/(n_f + n_t) = 1$ for a trap-free solid.

If, as is usually the case, the carriers encounter deeper traps during their passage through the sample, then J is reduced, sometimes by factors of 10^{-8}. It is the aim of SCLC theory to be able to deduce, from the J–V response of the material as a function of its temperature and thickness, such parameters as the concentration of traps, and their energetic distribution in the forbidden gap.

By making specific assumptions about the energetic distribution of traps (still neglecting diffusion), it is possible to derive more complicated expressions for the SCLC, which reveal a more detailed microscopic picture of the material under study. By providing a solid with hole- and electron-injecting contacts on opposite sides, it is possible to inject SCLC for both carriers simultaneously. This is referred to as double injection. The presence of both types of carrier introduces a new process, namely carrier recombination. Carrier recombination plays a major role in limiting current in two-carrier injection systems, and is the physical basis for electroluminescence (see Sec. 8.2). A good description of double injection may be found in Lampert and Mark (1970) and in PS, pp. 413 *et seq.*

5.2.3 Carrier Trapping The transport of carriers through organic crystals is dominated by the presence and distribution of carrier-trapping sites. An excellent discussion of the effects of traps on carrier mobility is given by Silinsh (1980, p. 250), SC, p. 372, Karl (1990), and PS, pp. 227 *et seq.* Traps can be situated at one specific energy level within the forbidden gap. This would be the case for a chemical impurity. Traps can also be produced by structural defects, such as vacancies, interstitials, grain boundaries, and dislocations. In anthracene, there are about 2×10^{20} m^{-3} lattice vacancies at room temperature. The number of dislocations depends on crystal preparation and handling. For melt-grown crystals, the total density can reach 10^{25} m^{-3}. Some defects such as lattice expansions do not trap carriers, but do cause scattering; such defects are called antitraps.

Chemical impurities can provide sites of lower potential energy for a carrier, causing localization for periods of time determined by the trap depth from the appropriate conduction level; chemical impurities can also act as antitraps. A hole trap is produced when an impurity has a lower ionization energy than the host material. The hole-trap depth E_t^h is given to a first approximation by the relation

$$E_t^h = I_c^h - I_c^g, \qquad (19)$$

where I_c^h and I_c^g are the ionization energies of the host and guest, both in crystal form. If $E_t^h > 0$, a hole trap is formed. If $E_t^h < 0$, an antitrap is formed. An electron trap is produced when an impurity has a larger electron affinity than that of the host. The electron-trap depth E_t^e is given to a first approximation by the relation

$$E_t^e = A_c^g - A_c^h, \tag{20}$$

where A_c^g and A_c^h are the electron affinities of the host and guest, respectively, both in the crystal form. If $E_t^e > 0$, an electron is trapped, and for the inverse, an antitrap is formed.

As is the case with inorganic semiconductors, ultrapurification of organic crystals has produced not only quantitative increases in the mobility of carriers, but qualitative differences in the sense that at low temperature (≈5 K) it was possible to demonstrate band mobility as opposed to the hopping motion that dominates at room temperature, and also to detect hot conduction electrons and holes (Karl, 1990).

The chief technique for determining whether trapping is taking place, and whether the traps are shallow or deep, is to measure the drift mobility of the carriers in a TOF method. In this method, the sample is placed between two blocking electrodes and a strongly absorbed light pulse is used to generate carriers close to one of the electrodes in the presence of an external field. Forced by the external field, the carriers of one sign drift across the bulk of the sample until they reach the opposite surface. By reversing the direction of the external field, the drift mobility of the oppositely charged carrier may be determined. As the carriers drift across the sample, they induce a current flow in the external circuit, which provides the TOF curve.

In Fig. 16 are shown the shapes of TOF pulses and what they indicate about trapping. In (a) is shown an ideal TOF pulse indicative of no significant trapping. The arrow marks the arrival time of the sheet of injected carriers at the back surface of the sample. In (b) is shown the effect of multiple shallow trapping, thermally activated detrapping, and retrapping. There is still a distinct arrival time for the injected carriers but the transit time has been increased because of the time spent in traps by the carriers. In (c) is shown a TOF curve where deep trapping is encountered. Here, there is no definite arrival time because the spread in arrival times exceeds the experimental sampling time.

In Fig. 18 is an illustration of a classic example of trap-free TOF curves for anthracene under external fields varying from 0 to 4.5 × 10^5 V m^{-1}. The broadening of the arrival tail is due to the mutual repulsion of the electrons as they drift, and also to the finite width of the charge-generation region. For a more complete discussion of methods of measuring and interpreting drift mobility, see PS, pp. 709 *et seq.* In order to obtain data such as are shown in Fig. 18, it is necessary to employ ultrapure single crystals, or, as was shown by Abkowitz *et al.* (1993), to use materials that have small ionization energies, making it difficult for an impurity to trap a hole by ionizing into the host hole.

It is practically impossible to determine the complete trap constellation in an organic semiconductor using any single experimental technique. The dominant methods depend on filling the various traps with carriers, and then emptying the traps to produce a characteristic pattern of enhanced conductivity in

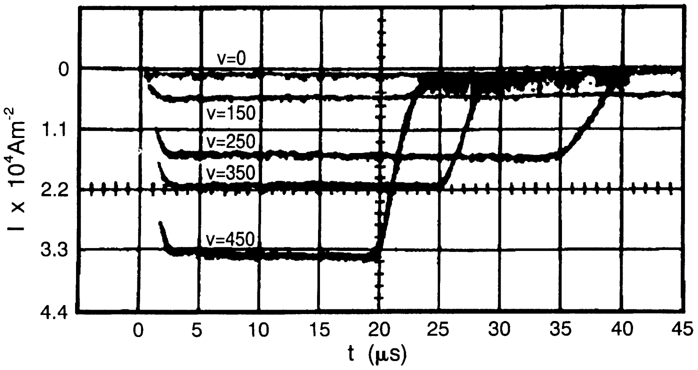

FIG. 18. Trap-free TOF curves for a 1.01-mm-thick anthracene crystal under applied voltages ranging from 0 to 450 V. The ordinate shows the current and the abscissa shows the time. The light pulse had a 1-μs halfwidth. (Redrawn from Karl *et al.*, 1970.)

the external circuit. These methods are mainly

1. thermally activated mobility;
2. thermal release, referred to as thermally stimulated currents (TSC);
3. photoexcitation of charged traps using specific wavelengths of light in order to deduce the depth of the traps;
4. excitation of trapped charge by energy transfer from excitons;
5. space-charge–limited current measurements as a function of temperature.

5.2.3.1 Thermally Activated Mobility. This method is most effective for the determination of shallow trap depths, if the concentration or the influence of deep traps is small enough to permit the carrier to traverse the crystal while interacting multiply only with shallow traps. When a single level of shallow trap at a density N_t and depth E_t is present then

$$\mu = \mu_0[1 + (N_t/N_0)\exp(E_t/kT)]^{-1}. \qquad (21)$$

From this equation, it follows that at high temperature ($kT \gg E_t$), shallow trapping is virtually absent and $\mu \propto \mu_0$, whereas at lower temperature,

$$\mu \approx \mu_0(N_t/N_0)\exp(-E_t/kT). \qquad (22)$$

In Fig. 19 is shown an Arrhenius plot of the electron and hole mobilities in an anthracene crystal doped with tetracene. Although the concentration of impurity was only 4 parts in 10^7, the effect on the mobility is considerable. The trap depths were calculated from these curves, using the above equations.

5.2.3.2 Thermally Stimulated Currents (TSC) and Isothermal Decay Currents (IDC). In the TSC technique, the traps in a solid are filled by passing a current through the solid at low temperatures; the voltage is then reversed and the temperature of the solid is increased linearly by means of a controlled heat input. As the traps empty, a current is measured and characteristic peaks are observed, indicative of traps lying at specific levels below the conduction band (in the case of electrons). In the IDC technique, the current through an insulator is abruptly increased, as with a pulse of exciting light, and the time decay of the current is measured at constant temperature. Both techniques involve observing current changes produced by the return of the system to thermal equilibrium after a nonequilibrium trap occupancy has been established. This initial trap loading can be due to either electrons or holes. It is possible to load the traps with carriers of only one sign if an injecting contact is used. The temperature range over which TSC and IDC experiments are performed determines the lower and upper limits to trap depths capable of being probed; generally traps shallower than 0.3 eV or deeper than 0.9 eV are not detectable by TSC techniques in the lower–melting-point polyacenes.

The analysis of TSC spectra is beset with difficulties stemming from overlapping peaks, an unknown spatial distribution of

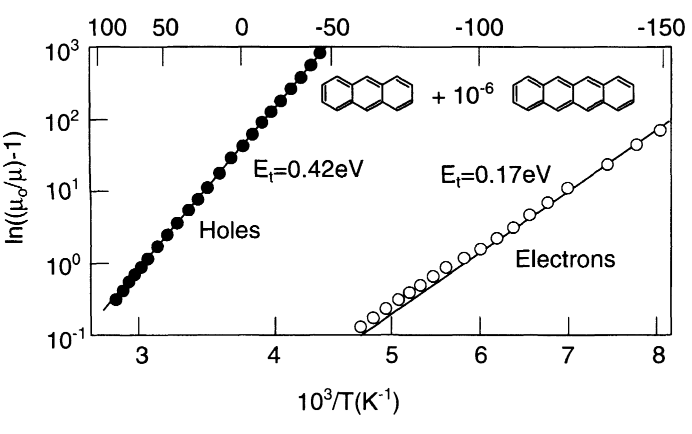

FIG. 19. Plot of $\ln[(\mu_0/\mu) - 1]$ vs $1/T$ for tetracene-doped anthracene, according to Eq. (21). Tetracene is a hole and electron trap and is present in a concentration of 0.4 ppm. Trap depths for holes and electrons are shown. (Redrawn from Karl, 1990.)

filled traps, and lack of temperature control across the sample. Also, the data are markedly affected by heating rate, applied electric field, efficiency of trap filling, and, in the case of polymers, the thickness of the sample (Stasiak *et al.*, 1980). In short, these experiments must be carried out with full awareness of the extensive literature that exists (see PS pp. 231 and 428 for references).

The lifetime of trapped carriers, τ_t, is related to the trap depth E_t (for the specific case of electrons) by

$$\tau_t^{-1} = \nu \exp[-(E_c - E_t)kT], \tag{23}$$

where ν is the escape-frequency factor representing the effective rate of lattice collisions with the trapped carrier. The parameter τ_t can be determined by measuring the IDC at several temperatures near the temperature peak in the TSC spectrum. The values inferred for the hole trap in anthracene are $E_t = 0.63$ eV and $\nu = 1.5 \times 10^9\ \mathrm{s}^{-1}$. The escape frequency is low compared with typical lattice frequencies of $\approx 10^{13}\ \mathrm{s}^{-1}$. A high retrapping rate would produce such a small escape frequency.

5.2.3.3 Optically Stimulated Current. The experimental arrangement for optically stimulated currents is the same as for TSC, except that the sample is not subjected to a rising temperature schedule. Instead, monochromatic light is focused on the sample and the photocurrent is measured as a function of the wavelength of light. The spectra show the electronic radical-ion transitions of the local trap. The electronic transitions of the radical ions can be identified from solution studies.

By combining optically stimulated currents with TSC, it becomes possible to identify the nature of the trap. This was done with tetracene-doped anthracene, and a hole trap at 0.45 eV was identified.

It is also possible to carry out a direct trap-to-band transition using light of the proper wavelength. Thus, for the case of tetracene in anthracene, it is possible to use IR light of energy close to the trap energy of 0.45 eV and detrap the trapped charge with an efficiency that outlines the vibrational states of the trap.

5.2.3.4 Detrapping by Excitons. Singlet and triplet excitons migrating through the solid can detrap trapped carriers when the exciton encounters the trap, and energy transfer to the trapped carrier is sufficient for detrapping. Neither the identity nor the depth of the trap can be determined by this method. However, important information about the spatial distribution of traps can be obtained.

The quenching of excitons by trapped carriers can, in addition, be used to distinguish between Ohmic and non-Ohmic contacts. When a contact is Ohmic and a SCLC is flowing through a solid, the total charge Q injected into the solid obeys the simple relationship $Q \approx CV$. In most crystals, Q consists mainly of trapped charge. At relatively low light intensity, the quenching of excitons depends linearly on the concentration of carriers (mostly trapped). Triplet excitons have a long natural lifetime, and so they are sensitive to relatively low concentrations of trapped carriers. The quenching of triplet excitons by trapped carriers reduces the exciton lifetime and concentration. The decrease in exciton concentration can be monitored by measuring the delayed singlet-exciton fluorescence produced by the fusion of two triplet excitons. It is thus possible to follow the quenching of the triplet excitons by trapped charge as a function of the voltage applied to the crystal. When an Ohmic contact is operative, then $Q \approx CV$ and the quenching is linear with applied voltage. With an Ohmic CuI contact applied to an anthracene crystal, the linear quenching curve persists up to about 1000 V on a 225-μm-thick crystal. When the contact is depleted, the photocurrent saturates and the quenching is essentially constant as the voltage is increased. An example of a contact that is Ohmic to begin with and then becomes emission-limited is provided by Ce^{4+} in aqueous solution (PS, p. 399).

5.2.3.5 Trap Distributions. Theoretically, SCLC and its temperature dependence can be used to deduce the concentration and energetic distribution of traps in a solid. As was discussed in Sec. 5.2.3, this can be validly accomplished only if the correct distribution function is chosen for the traps in the solid. An example of the use of an exponential distribution of traps is given by Bässler *et al.* (1969), who studied single-crystal tetracene. From the J-vs-V dependence and from measurements of the temperature dependence of J at various values of V, it was possible to determine the position of the quasi-Fermi level E_F, and hence the level of filled traps. It was concluded that in the energy range $0.6 < E_t < 0.9$ eV, the traps are distributed exponentially. It was also possible to determine the concentration of traps as $\sim 3 \times 10^{21}$ m^{-3}.

For a Gaussian distribution of traps, the diagnostic criteria are completely different. The SCLC equations must be solved with a Gaussian distribution function

$$N(E_t) = (N_t/\sigma\sqrt{2\pi}) \exp[-(E - E_t)^2/2\sigma^2], \tag{24}$$

where E is the mean trap depth, N_t is the total trap density, and σ is a measure of dispersion of trap levels on either side of the mean. These equations can be solved analytically only by dividing the observed SCLC regime into about four sections. When this is done, the results for various materials are as shown in Table 1.

5.3 Transient Currents (Single Carrier)

5.3.1 Space-Charge–Free (SCF) Transients An important application of space-charge–free transients is to measure the time of flight (TOF) of a sheet of charge injected into the solid. From the shape of the transient pulse amplitude as a function of time, it is possible to deduce the mobility of the carrier, and whether it had encountered a shallow or deep trap en route to the collecting electrode. The shape of the response may be seen in Fig. 16. From the transit time and the electric field, the TOF can be calculated. In addition, the electric field is large enough to make the transit time of the carriers much shorter than the dielectric relaxation time of the solid, i.e., $\tau < \epsilon\epsilon_0/\sigma$. This is necessary in order to see a field effect on the transit time. In high-resistivity materials, this condition is easily realized.

In the ideal case, the arrival time of the charge at the collecting electrode is sharply delineated. In the case of deep trapping, there is no obvious transit time at all, as is shown in Fig. 16.

5.3.2 SCLC Transients A current transient brought about by the injection of an amount of charge $Q = CV$, where C is the capacitance of the sample, has been thoroughly analyzed theoretically, and is discussed in detail in Lampert and Mark (1970). A good example of the utility of trap-free SCLC transients (TFSCLC) is given by Abkowitz *et al.* (1993). They outlined a tech-

Table 1. Typical values of Gaussian trap-distribution parameters E_t, σ, N_t, and N_e for a number of aromatic crystals (anthracene, tetracene, pentacene, and perylene) in different crystalline states. N_e is the density of states at the conductivity level, and the other parameters are defined in the text. SC = single crystal; OL = oriented polycrystalline layers obtained by evaporation in vacuo; QA = quasiamorphous obtained by evaporation in vacuo on a cooled surface; PC = polycrystalline. (From Silinsh, 1980).

Compound	Crystalline state	E_t (eV)	σ (eV)	N_t (10^6 m^{-3})	N_e (10^6 m^{-3})
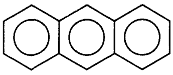	SC	0.5–0.8	0.03–0.13	2×10^{11}–8×10^{13}	10^{16}–10^{20}
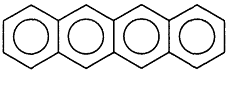	SC	0.6	0.05–0.14	10^{13}–10^{14}	10^{18}–10^{21}
	OL	0.3–0.4	0.05–0.14	6×10^{13}–2×10^{14}	10^{12}–10^{15}
		0.1	0.07–0.13	4×10^{14}–6×10^{15}	
	QA	0.26	0.11	2×10^{14}	
		0.1	0.05	4×10^{15}	
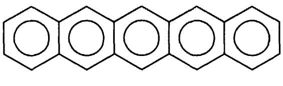	OL	0.35	0.11	8×10^{13}	10^{14}–10^{15}
		0.08	0.07	8.5×10^{14}	
	QA	0.35	0.06	8×10^{13}	10^{14}–10^{15}
		0.04	0.07	7×10^{15}	
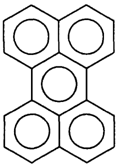	PC	0.4	0.03	4×10^{13}	10^{11}–10^{15}
		0.3	0.05–0.09	(4–7) $\times 10^{15}$	

nique for determining whether a contact is Ohmic, using trap-free polymers. The polymer was a thin film (9 μm) of poly-(tetraphenylbenzidine) (PTPB). This polymer is a hole conductor and has an ionization energy so low that there are very few impurities with lower ionization energies that can act as hole traps (see Sec. 5.2.3). The experimental configuration is shown in Fig. 20.

For a trap-free solid, the maximum current is given by the relation

$$J = (9/8)\mu\epsilon\epsilon_0 V^2/L^3 = (9/8)CV/At_{tr}, \qquad (25)$$

where C is the capacitance of the experimental configuration, t_{tr} is the transit time of the injected carrier, and A is the area of the contact. Thus, by measuring t_{tr} and C, it should be possible to calculate the steady-state SCLC at the same electric field, if the contact is Ohmic. A TFSCLC transient current plotted against time is shown as the top curve in Fig. 21(a). The time t_p at which there is a peak in the response curve, according to the theory of TFSCLC, is related to the transit time of the carriers simply as $t_{tr} = t_p/0.8$. The transient TFSCLC is generated by applying a positive field step of 10^7 V m^{-1} at $T = 323$ K. In a TOF experiment made on the same film at the same applied field, the bottom curve in Fig. 20(b) is obtained. The mean hole-transit time is taken at the point where the TOF signal has dropped to one-half of the plateau amplitude; within experimental accuracy, it is the same as is shown in the top curve of the same figure.

The steady-state TFSCLC is attained after the elapse of several transit times, and is shown in the top curve of Fig. 21(b). It is possible to calculate the steady-state TFSCLC, I_{SS}, using the expression

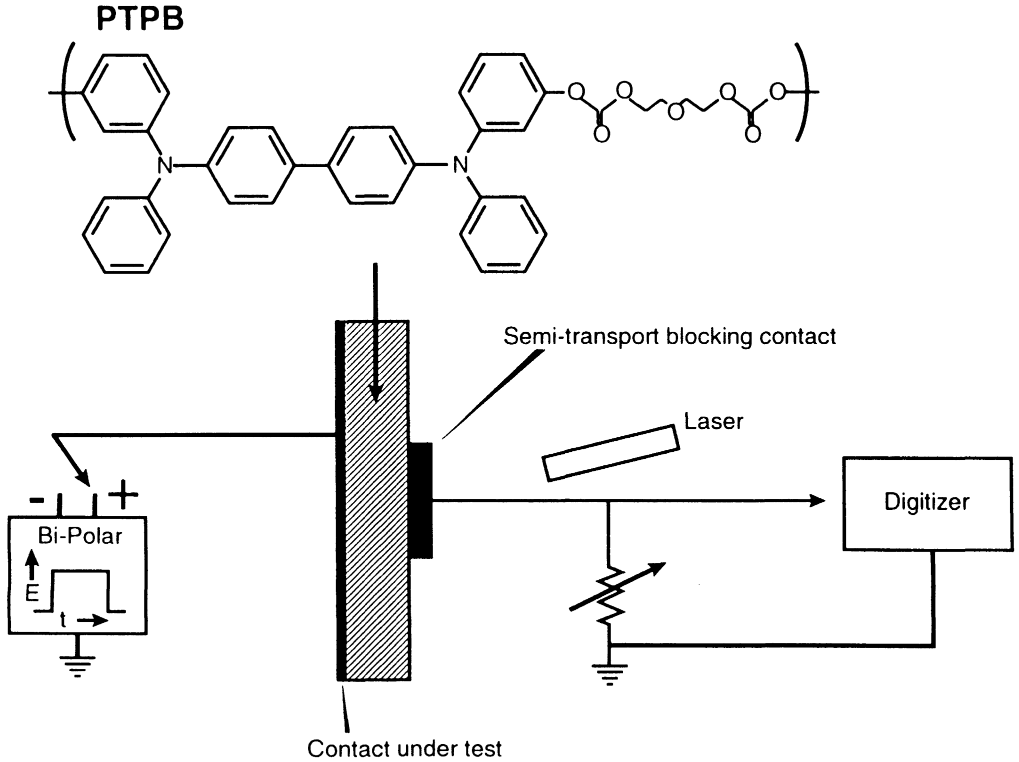

FIG. 20. Schematic of experimental configuration. Structure of the trap-free polymer PTPB is displayed. During TOF measurements, the semitransparent contact is positively biased and the field is applied 15 ms prior to the light flash. During the transient dark-injection measurements (see Fig. 21), the contact under test (substrate) is excited by a positive voltage pulse. (Redrawn from Abkowitz *et al.*, 1993.)

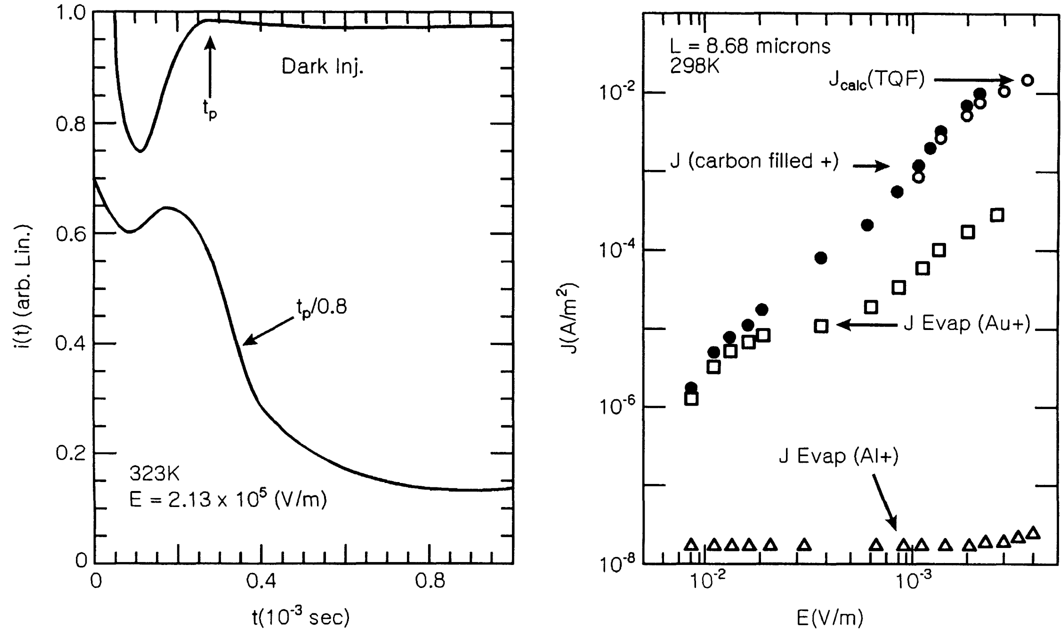

(a) (b)

FIG. 21. (a) Comparison of a dark hole-injection transient and the TOF hole transient in the same PTPB film (T = 323 K, E = 2.13 × 10^7 V m^{-1}). (b) Steady-state J-vs-E plots on the same PTPB film fitted with various contacts. Solid circles are hole currents sustained by the carbon-filled polymer contact. Open squares are hole currents sustained by a gold contact. Triangles are hole currents injected from an evaporated aluminum contact. Open circles are steady-state dark currents computed from TOF transit times using TFSCLC theory. Departure of the slope from the theoretical value of 2 is due to field dependence of the mobility. (Redrawn from Abkowitz *et al.*, 1993.)

$$I_{SS} = (9/8)CV/t_{tr} \tag{26}$$

and the value of t_{tr} taken from the TOF measurement. If the contact is Ohmic, the calculated and measured I_{SS} should agree.

In Fig. 21(b) are shown the results for several different contacts. As is seen from the equivalence of the calculated and measured steady-state TFSCLC, the carbon-filled polymer contact is Ohmic. The evaporated Au contact approaches Ohmicity at low field strength, but becomes emission-limited at higher field strength. The Al contact is strictly emission-limited. The departure of the slope of the curve of ln$J$ vs ln$V$ from the theoretical value of 2 is due to the field dependence of the mobility that is characteristic of transport in molecularly doped polymers (see Sec. 5.2.3).

This rare demonstration of TFSCLC is made possible because the film is thin and the hole conductor has an ionization energy so low that there are few impurity species with a lower ionization energy. Thus there are no hole traps in the film, and the hole is the majority carrier.

6. PHOTOEMISSION

Photoemission takes place from and provides information about the surface region of a solid; this region is of the order of a few nanometers in depth. However, unlike typical inorganic materials, organic semiconductors do not exhibit dangling bonds at the surface, and so the surface properties do not differ markedly from those in the bulk.

Photoemission is used mainly to obtain such information as the ionization energy of the solid, the work function, the energy-level structure in the valence band, and the density of states in the valence band. By using multiphoton photoemission, in which two or more quanta of low-energy light add their energies and succeed in providing a total energy equal to or greater than the ionization energy of the solid, it is possible to measure exciton levels, the band gap, and trap levels.

The yield of photoelectrons per incident photon, which in the case of anthracene is about 10^{-4} for a photon energy about 1 eV above the threshold, is too low to be of practical interest.

6.1 Energy-Level Analysis

In Fig. 22 is shown a series of photoemission curves plotted against the binding energy for a series of polyacenes. Included in the figure are the photoemission data for the corresponding gas-phase molecules, shifted by an arbitrary energy interval to bring the peaks of the yields into coincidence. The binding energy represents the location of an energy level in the valence band of the compound, relative to the position of the Fermi level of the solid, E_F. Thus, for anthracene, the first peak appears about 1.8 eV below the Fermi level. This is the level of the highest occupied orbitals. In anthracene, using the known values for the band gap of 4.0 eV, and the ionization energy of the crystal, the Fermi level should lie at 2.0 eV above the highest occupied valence level. If the valence and conduction bands remained flat up to the surface (no surface states to pin the Fermi level at some other position) then in Fig. 22, the first peak should appear at 2.0 eV instead of 1.8 eV. Thus E_F is closer to the valence band, indicating the presence of charge between the Fermi level and the valence band, due perhaps to the presence of some negatively charged surface states.

As may also be seen in Fig. 22, there is a close correspondence between the vapor-phase and solid-phase spectra. This is a further manifestation of the weak intermolecular interaction in the van der Waals solids that allows the individual molecular properties to dominate the solid-state spectrum. The shift in energy required to put the vapor-phase and solid-phase spectra into coincidence is a measure of the cohesive-energy contribution to that transition. The broadening of each peak in the solid-phase spectrum can be traced to fluctuations in the intermolecular electronic polarization. This illuminates one of the most significant differences

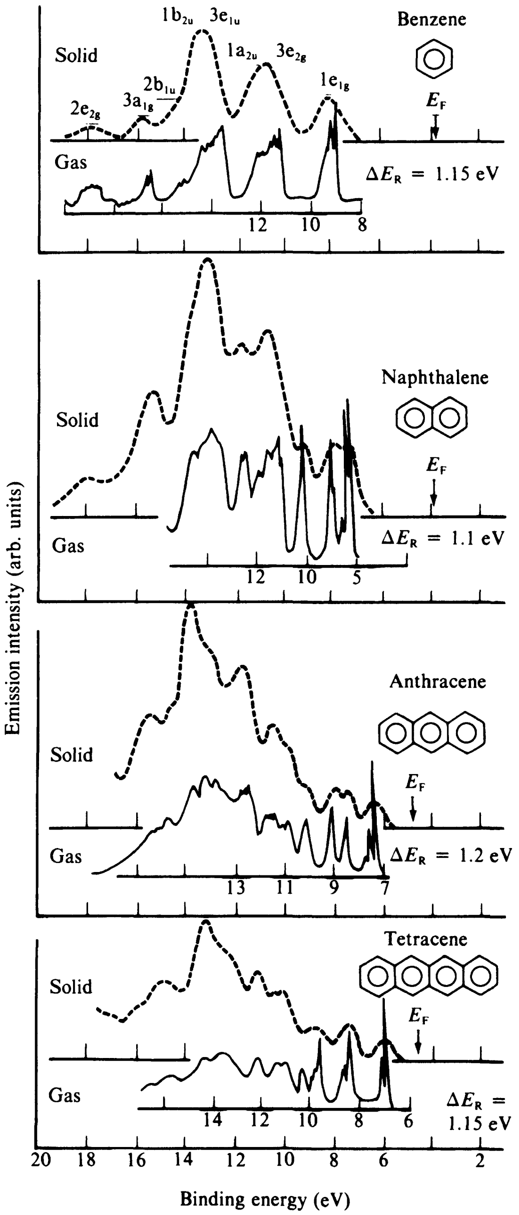

FIG. 22. Comparison of gas-phase photoemission spectra with photoemission yield from polycrystalline films using He I radiation ($h\nu$ = 21.22 eV). ΔE_R gives the shift resulting from the cohesive-energy contribution to the positions of the energy levels in going from the gaseous to the solid state in order to obtain coincidence for most of the prominent bands. The broadening of each peak in the solid phase is due to the polarization fluctuations caused by the presence of defects and phonon interactions. E_F denotes the position of the Fermi level. (Redrawn from Grobman and Koch, 1979.)

between organic crystals and inorganic materials, namely the importance of electron–phonon interactions. These fluctuations (0.1 to 0.5 eV) are caused by defects, termination of the crystal at the surface, and thermal vibrations. Electrons originating from monomolecular layers at different depths from the surface show slightly different energies, since the polarization energy is a function of the distance from the surface.

To obtain information about the density of states in the valence and conduction bands, accurate photoemission and optical data must be available. It is also based on other assumptions, most of which can be justified. For more details, see PS, pp. 552 *et seq.*

Since autoionization plays a large role in the photogeneration of carriers, it is not surprising that it also plays a role in photoemission. This is exemplified by the fact that absorption and photoemissive yields vary in the same way (though not proportionally) (PS, p. 568). The effects of AI are more evident in multiphoton photoemission.

6.2 Multiphoton Studies

The first multiquantum photoemission study of organic semiconductors was carried out on anthracene (Pope *et al.*, 1965), and was shown to be due to a fusion of two Frenkel singlet excitons, each at 3.15 eV, to generate an excited state at 6.3 eV, which autoionized to emit an electron. The bimolecular rate constant for this process was calculated to be $\sim 10^{-14}$ m^3 s^{-1}.

In addition to Frenkel exciton–exciton fusion reactions, it is also possible to determine a wide variety of lower-lying energy states, such as trapped-charge and *e–h* pair states, and also photon–photon photoemission (see PS, p. 477).

7. MAGNETIC-FIELD EFFECTS

7.1 Hall Mobility

Measurement of the Hall effect (PS, p. 374) allows the free-carrier density and trap-free mobility to be determined. In the aromatic solids anthracene and naphthalene, Hall mobilities are of the order of 10^{-4} m^2 V^{-1} s^{-1} at room temperature. In simple metals, $\mu_H = \mu_D$, while usually in semiconductors $\mu_H > \mu_D$, the exact value being determined by the energy dependence of the carrier-scattering mechanism. In general, dark Hall measurements on most organic semiconductors are too difficult to perform because their conductivity is too small at the low fields necessary to have Ohmic currents. In these cases, the Hall measurements are made using the photocurrents. The ratio of the photo-Hall hole mobility in the *b* direction to the drift mobility is about 8 in anthracene (PS, p. 376). Recently, however, the Hall effect for the organic semiconductor *bis* (1,2,5-thiadiazolo)-*p*-quinobis (1,3-dithiole) was found to be 4×10^{-4} m^2 V^{-1} s^{-1} at room temperature with the sign of the carrier as positive, suggestive of a band model for the holes, since $\mu_H > 10^{-4}$ m^2 V^{-1} s^{-1}, the lower bound for band mobility.

7.2 Magnetic-Field Effects on Spin–Spin Interactions

Magnetic-field effects in organic solids were initially observed in delayed fluorescence in anthracene single crystals due to the fusion of two mobile triplet excitons to create a final fluorescent singlet state (Merrifield, 1968). Subsequently, magnetic-field effects in organic semiconductors were observed in photoconductivity, electroluminescence, charge-detrapping processes, scintillation, fluorescence, and geminate recombination, in both crystalline and polymeric systems.

Figure 23 illustrates the high-field anisotropy of the photoenhanced bulk conductivity in anthracene as reported by Geacintov *et al.* (1970). The dips in the photocurrent owe their origin to a reduction in the efficiency of detrapping of holes by triplet excitons. The common theoretical elements of these field effects can be summarized as follows. Consider two species, *A* and *B*, initially spatially separated, which through their motion enter a correlated state (*AB*); these could represent triplet or singlet excitons, trapped or free carriers, or paramagnetic defects. This correlated state evolves in time according to a Hamiltonian H_S that does not include spatial variables and is therefore called a spin Hamiltonian. After a time of the order of nanoseconds, two or more final products, *C* and

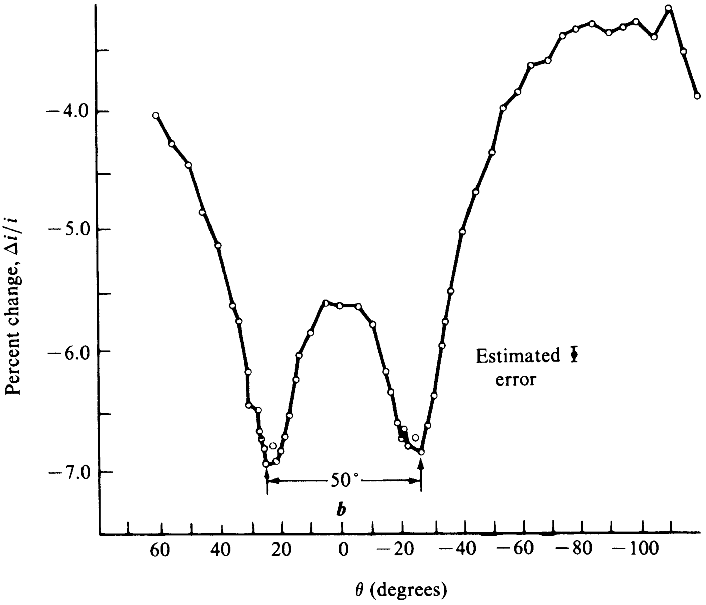

FIG. 23. Effect of magnetic field $H = 3400$ G on hole photocurrent i. H was rotated in the ab crystal plane. Photocurrents were measured with 24 V across the crystal; current with $H = 0$ was 1.3×10^{-5} A m^{-2} at 366 nm. The b axis is at 0 and the resonances (arrows) are calculated from theory. $T = 298$ K. (Redrawn from Geacintov *et al.*, 1970.)

D, are formed. Viewed as a chemical reaction, the overall process is usually written as

$$A + B \underset{k_{-1}}{\overset{k_1}{\leftrightarrows}} (AB) \overset{k_2}{\rightarrow} C + D, \qquad (27)$$

where k_1 denotes the pair-formation rate constant and k_{-1} the pair-dissociation rate constant, and k_2 represents the transition to a new state of the system. The basic assumptions that allow the process shown in Eq. (27) to have a magnetic-field effect are that

1. the overall reaction is spin conserving, and
2. the thermalization times τ_t of the pair spin states are longer than the pair-state lifetimes τ_p.

Since H_S does not commute with the total spin of the pair state, the external magnetic field H can impart a different percentage of spin character to each of the pair states ψ_l. If the final states $C + D$ in Eq. (27) have a total spin S, and P_S denotes the S projection operator, then only pair states ψ_l for which $|\langle S|P_S\psi_l\rangle|^2 \neq 0$ contribute to the overall reaction since $k_2 \propto |\langle S|P_S\psi_l\rangle|^2$.

In geminate recombination, the relevant interactions are the electron and hole hyperfine terms; generally one includes only the contact term $\mathbf{I}\cdot\mathbf{S}$, the Zeeman interaction, and the exchange interaction $J(\mathbf{r})\mathbf{S}_e\cdot\mathbf{S}_h$. As an example of how field effects provide insight into fundamental processes, measurements by Frankevich and Lymarev (1992) on the magnetic-field effects on photocurrent in polymers PTS, PPV, PPPV, and PMOP-PPV have permitted estimates of the size of the geminate electron–hole (CT) state; these vary from 1.5 to 3.5 nm.

Triplet-exciton pairs can fuse to produce not only a fluorescent singlet state, but also free carriers provided that twice the triplet energy exceeds the band gap; and in such cases, the intrinsic current should be magnetic-field dependent. Evidence for exciton fusion as a mechanism for the intrinsic generation of free carriers in molecularly doped polymer consisting of solid solutions of the amine TTA [tri-(p-tolyl)amine] in the polymer Lexan was observed by Orlowski and Scher (1983). It was found that the photogeneration yield Y depended on light intensity as I^2 at low intensity and as I at high light intensity and at some intermediate value in between with I depending on the electric-field strength. The fusion of triplet excitons in this case resulted initially in a charge-separated state in which the polymer contributed an acceptor state and a hole was created on the TTA molecule. The overall rate of carrier generation was diffusion limited and had a rate constant of the order of

10^{-17} m^3 s^{-1}, which is typical of organic materials.

8. APPLICATIONS

8.1 Electrophotography

By far the greatest industrial application of organic semiconductors is in the field of electrophotography. A schematic of a conventional organic photoreceptor is shown in Fig. 24. It consists of a two-layer system: a charge-generation layer, between 1 and 3 μm in thickness, and a charge-transport layer, between 15 and 30 μm in thickness. Charge-transport layers are typically mixtures of low–molecular-weight donor or acceptor compounds in a polymer. The carrier moves by thermally assisted hopping along the path created by molecules in contact with each other. The charge-generation layer is chosen for its sensitivity to the light used for exposure, and for its quantum efficiency of charge generation. This layer plays the role of the photosensitive electrode, injecting charge into the charge-transport layer. The charge-transport layer is chosen for its superior carrier mobility, interlayer compatibility, absence of interfacial trapping, mechanical properties, and injected-carrier range.

The quantum efficiency of a two-layer system consisting of a metal-free phthalocyanine charge-generating layer and a hydrazone-doped polycarbonate charge-transport layer was measured, yielding a primary quantum yield of 0.5. See Borsenberger and Weiss (1993) and Schein (1988) for more complete details regarding electrophotography.

8.2 Electroluminescence (EL)

The industrial potential for electroluminescent organic materials, particularly polymers, is great. Not only does this provide the possibility of large-area, flexible illumination, with significant energy saving, but it makes possible flat video screens. EL was first reported in polymers by Burroughes *et al.* (1990) in *p*-phenylene vinylene (PPV). This thin film has a low turnon voltage (~3 V) and relatively high quantum efficiency (1% photons out/electrons in), and shows green and orange light using different polymers. They are bright, and can be seen in room lighting. In addition, a light-emitting diode (LED) made out of oriented polymers will emit polarized light; this opens another possibility for transferring information. Oriented photoluminescence has already been demonstrated (Heeger, 1992). Recent improvements in PPV involve preparing block copolymers consisting of alternating sequences of conjugated and nonconjugated segments (Holmes *et al.*, 1992). This mimics the GaAs quantum-well structure. By varying the degree of conjugation and copolymerization, the colors can be varied from green-blue to orange-red; the light is also brighter. More efficient LEDs can be made with transparent polyaniline as a hole-injecting electrode, instead of indium and tin oxide (Yang and Heeger, 1994).

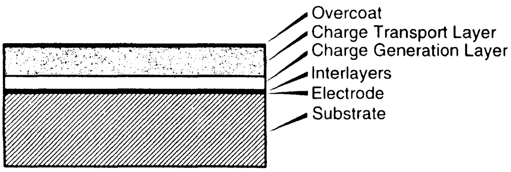

FIG. 24. A cross-section schematic of a conventional dual-layer photoreceptor (redrawn from Borsenberger and Weiss, 1993.)

The same device that emits EL can, under reverse bias, act as a photodetecting diode, with a yield of 0.2 electron/photon (Yu *et al.*, 1994). Electron injection is usually the difficult function to accomplish because it requires efficient chemical reducing power. Such materials are susceptible to attack by ubiquitous oxidants, such as air. Alloys of Mg and Ag were used by Tang and Van Slyke (1987), who ushered in the modern era of EL by introducing the scheme of using a bilayer structure, in which the functions of charge generation, transport, and luminescence were accomplished by using materials best suited for each function. This structure consists of a hole-transport layer and a luminescent layer. The hole-transport layer is an amorphous diamine film, and the luminescent layer consists of 8-hydroxyquinoline aluminum (Alq_3), which mostly transmits electrons. An intermediate layer of coumarine-doped Alq_3 improves EL by a factor of 2. A typical multilayer EL configuration is shown in Fig. 25.

The main problem with organic EL devices is operational stability. At present, 10 000 hours of continuous operation has been achieved.

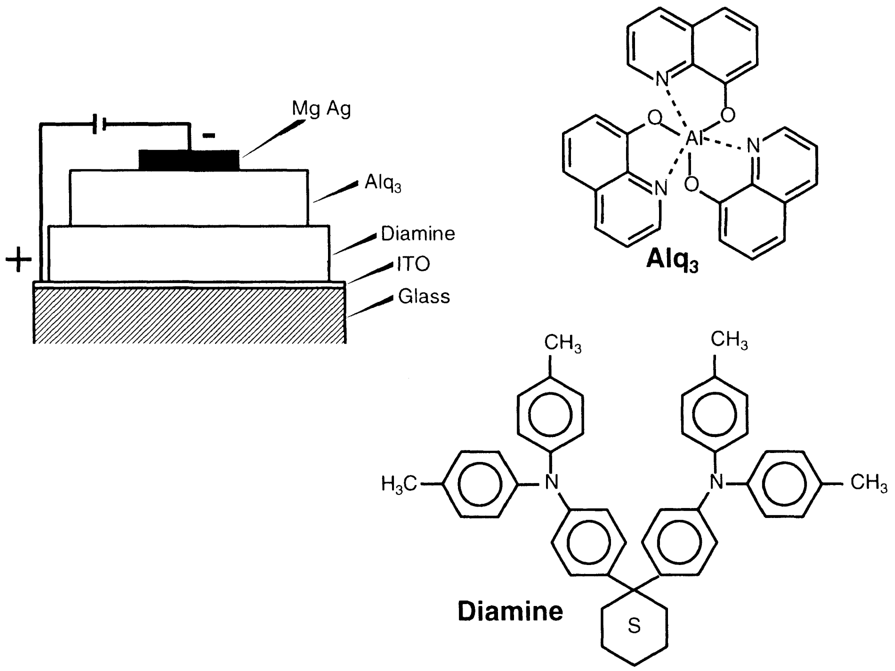

FIG. 25. Configuration of a dual-layer electroluminescent cell and molecular structures of the hole-transport layer (diamine) and the electron and luminescent layers (Alq_3). The top electrode is an electron injector and the ITO layer injects holes. This organic diode acts as a rectifier, with the forward bias represented by a positive voltage on the ITO. The unit can emit visible light at 2.5 V. (Redrawn from Tang and Slyke, 1987.)

8.3 Sensors

Considerable work has been carried out on the phthalocyanines with the goal of preparing chemically stable sensors for various gases. A reproducible sensor sensitive to nitrogen dioxide at relatively low temperature (313 K) was prepared from tetra-*t*-butyl-substituted copper phthalocyanine, deposited by the Langmuir–Blodgett (LB) technique on a glass substrate (Chyla *et al.*, 1993) and aged. The change in conductivity ΔJ of the LB film exposed to NO_2 relative to the conductivity J_0 of a film free of NO_2 varied from $\Delta J/J_0 \sim 0.2$ to $\Delta J/J_0 \sim 0.8$ as the NO_2 concentration was changed from 0.28 to 1.22 ppm.

Sensors based on conductive polymers, capable of detecting biologically important compounds such as glucose, are discussed by Heller (1992).

8.4 Semiconductor Devices

Three types of devices that have been fabricated are a Schottky-barrier diode, a metal-insulator–semiconductor (MIS) device, and a metal-insulator–semiconductor field-effect transistor (MISFET). The major advance in the use of organic semiconductor devices is the ability to prepare a solution of the precursor polymer that can be deposited in layers in the correct geometry, after which the precursor is transformed into the final conjugated polymer. Polyacetylene is one important polymer that can be used in this manner; it can be doped *n* type with Na and *p* type with I_2.

To prepare a Schottky-barrier diode, a layer of gold is evaporated on a substrate followed by a coating of polyacetylene, on top of which is placed the low-work-function layer of aluminum. The ratio of forward to reverse current is about 5×10^5 at a bias voltage of 1.5 V with a saturation current of about 10 A m^{-2}.

A MIS device has been prepared by depositing a layer of gold (20 nm) on a silica substrate, followed by a polyacetylene layer 20 nm thick, upon which in turn is deposited by spin-coating a layer of polymethylmethacrylate 150 nm thick, upon which is evaporated another 20-nm layer of gold. This device permits the creation of accumulation, depletion, and inversion layers in the polyacetylene, and thus a greater control of the current passing through the device.

In Fig. 26 is shown an example of a MISFET device. The source (emitter) and the drain (collector) are supplied with *n*-Si contacts, which are intended to make Ohmic contact to *n*-type polyacetylene film. The channel conductance changes by a factor of 10^5 when the gate-to-source voltage is changed from -40 to $+10$ V.

The MISFET device demonstrates that the charge at the polyacetylene/insulator interface in the MIS device is mobile, since, as is seen in Fig. 26, it is possible to pass current from source to drain and thus measure the conductance of the surface-charge layer. From these conductance measurements, it is possible to determine the carrier mobility, which comes out to 10^{-8} m^2 V^{-1} s^{-1}. Similar values were found for the extrinsic carriers in undoped polyacetylene.

8.5 Photorefractive Polymers

A photorefractive (PR) material is one in which a spatial modulation of its refractive index occurs as a result of a light-induced spatial modulation of an internal charge distribution. The material must therefore contain an optically nonlinear component. Such a material has potentially important applications, such as high-density optical storage, phase conjugation, and programmable optical interconnection. At present most if not all commercially available photorefractive materials are inorganic, such as $LiNbO_3$, $(Sr,Ba)Nb_2O_6$, GaAs, and $BaTiO_3$. The discovery of a photorefractive effect in polymers (see Moerner and Silence, 1994) is of great potential importance, since polymers are inexpensive, easy to process, and versatile.

The basic requirements for a PR material are

1. a photoionizable charge-generating component,
2. a charge-transporting medium,
3. trapping sites, and
4. a nonlinear optical (NLO) material whose index of refraction can be altered under the influence of a space-charge field.

The principle of operation is as follows: A polymer that contains all of the components listed above is prepared in the form of a thin film (~100 μm) sandwiched between transparent electrodes. A large electric field is applied to the sandwich, at a temperature appropriate for the alignment of the NLO chromophores with the external field. The optimal NLO optical materials are conjugated organic compounds with an asymmetric charge distribution as would be produced by having an electron-donating substituent at one end of the molecule and an electron-accepting molecule at the other end. Matters are improved if the distance between these groups is increased judiciously. The orientation of the NLO chromophores by means of an external electric field breaks the isotropy of the sample making possible a maximum electro-optic effect.

To write a hologram, two intersecting coherent beams of light are focused on the surface of the film, using a wavelength that addressed the charge-generating component only. The interference between the beams of

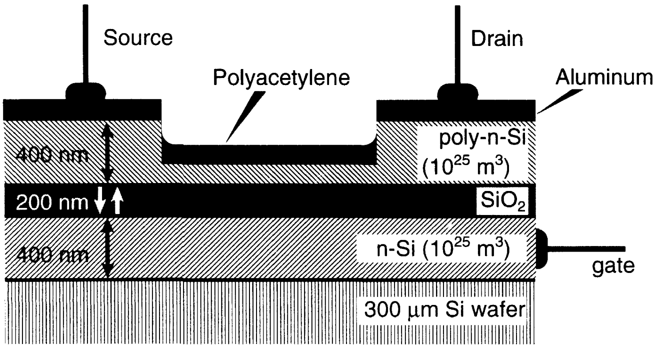

FIG. 26. Schematic diagram for a polyacetylene MISFET structure. Dimensions shown are to scale, except the channel width (20 μm) and length (1.5 × 10^{-2} m). (Redrawn from Burroughes *et al.*, 1990.)

light produces a stationary pattern of light and dark fringes. In the illuminated regions, charge generation takes place, and the external electric field produces charge separation. The charge-generation component is dispersed in a matrix that is chosen to support the movement of only one sign of carrier, usually the hole. The electron thus remains stationary and the positive charge is displaced into traps in the dark region. The light source is removed, leaving a pattern of displaced electrons and holes similar to the interference pattern, but the phase of the space-charge electric field is shifted in phase from the light-intensity pattern by about 90°. The displacement of the trapped holes from the immobile electrons produces an electrostatic restoring field that is high enough to alter the refractive index of the NLO chromophore. Interestingly, the presence of traps, which is the bane of photoconductors, is a basic necessity in the PR effect.

If the material has a linear electro-optic effect, the magnitude of the refractive-index modulation Δn is related to the space-charge field E_{sc} by the relation

$$\Delta n = -\tfrac{1}{2}E_{sc}n^3r_e, \qquad (28)$$

where r_e is the effective electro-optic coefficient for the geometry under consideration (about 1–20 for organics). Using a figure of merit $Q = n^3r_c/\epsilon_r$ where ϵ_r is the dc dielectric constant, the value for the inorganic material is 19.3; for present polymeric materials it is 2.2 with a potential value of 31.1 in sight.

The phase-shifted refractive-index modulation forms a grating that can diffract light and permit the utilization of two-beam coupling whereby one of the writing beams gains energy at the expense of the other, a property that is made use of in photonic devices. A sample composition for a PR polymer is polyvinylcarbazole (PVK) functioning as the hole transporter, doped with a NLO chromophore compound, 3-fluoro-4-(*N*,*N*-diethylamino)-β nitrostyrene (FDEA-NST) and sensitized for charge generation with 2,4,7-trinitro-9-fluorene (TNF). Typical poling field strengths are about 10^7 V m^{-1}, and space-charge fields about half that. Carrier mobilities are in the range 10^{-11} to 10^{-10} m^2 V^{-1} s^{-1}, and quantum efficiencies of charge generation are $\sim 10^{-2}$. A grating efficiency η, defined as the intensity ratio of the diffracted beam to that of the incoming reading beam, of the order of 100% has been reported (Meerholz *et al.*, 1994).

Although the readout requires the use of light, which tends to destroy the information residing in the charge pattern, it is possible to minimize this effect (Silence *et al.*, 1994). This greatly increases the utility of the PR effect as an information-storage method. For a complete introduction to and survey of PR methodology, see Moerner and Silence (1994).

ACKNOWLEDGMENTS

We have benefited from the support of Department of Energy grant No. DE-FG02-86ER-60405 and from discussions with Heinz Bässler, Charles B. Duke, R. Glen Kepler, William E. Moerner, Lewis J. Rothberg, Lawrence B. Schein, Ching W. Tang, and Marek Zielinski, and especially with Martin A. Abkowitz and Nicholas E. Geacintov.

GLOSSARY

Aliphatic: Containing no aromatic rings.

Amphiphilic: Having both hydrophobic and hydrophilic substituents on the same molecule.

Aromatic Compound: An unsaturated hydrocarbon containing planar ring systems stabilized by delocalized electrons, as in benzene.

Autoionization: The spontaneous ionization of an electronically excited molecular state.

Bipolaron: A quasiparticle composed of two correlated like-charged polarons.

Chromophore: An atom or group of atoms that acts as a unit in light absorption.

cis: A molecular geometry; thus in the structure shown, *A* and *B* are *cis* with respect to the double bond, which does not permit independent rotation of *A* or *B* around the bond.

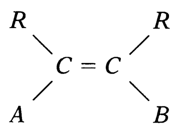

Conjugated: A system of alternating covalent single and double bonds; electron delocalization ensues.

Coulomb-Capture Radius: The mean distance at which the Coulomb-interaction energy between oppositely charged particles equals the thermal energy.

DPQ: 3,3′-dimethyl-5,3′-di-*t*-butyldiphenoquinone.

Exciton: A mobile electronically excited neutral nonconducting state.

Fermi Glass: A molecular solid in which there are no bandlike states, i.e., all electronic states are localized.

Geminate Recombination: Fusion of a hole–electron pair formed from the same molecule.

Glass-Transition Temperature: A small range of temperature over which the viscosity of the melt increases steeply, viz., several orders of magnitude within 10 K.

HOMO: Highest occupied molecular orbital.

ITO: Indium–tin oxide transparent contact.

LUMO: Lowest unoccupied molecular orbital.

OSC: Organic semiconductor.

PA: Polyacetylene.

PC: Polycarbonate.

Peierls Instability: A lattice instability in a one-dimensional system due to electron–phonon (and/or Coulomb) interactions, having rational-fractionally filled orbitals.

Pendant Group: A group of atoms attached as a unit to a polymeric backbone.

π Bond: A chemical bond formed by the overlap of neighboring *p* orbitals that has one and only one nodal plane, which includes the internuclear axis.

PMPS: Poly(methylphenylsilylene).

Polaron: A quasiparticle consisting of an excess charge and its associated polarization field.

Primary Quantum Yield: The yield per photon of carriers before any recombination of any kind has occurred.

PS: Polystyrene.

PVK: Polyvinylcarbazole.

Schottky Barrier: A potential barrier formed by a space-charge double layer whose width and magnitude are dependent on the applied voltage.

Sigma (σ) Bond: A chemical bond made by the overlap of orbitals that have cylindrical symmetry around the bonding axis.

Soliton: A mobile bond-alternation defect in a π-conjugated chain; if charged, has a spin of zero, and if uncharged, has a spin of one-half.

Symbatic: In phase with the absorption spectrum.

Thermalization Distance: The distance traveled by an ionized and excited electron before it comes into thermal equilibrium with its surroundings. Distance is measured in a straight line between starting and final points.

TNF: 2,4,7-trinitro-9-fluorene.

TOF: Time of flight.

trans: A type of molecular geometry; thus in the structure shown, *A* and *B* are *trans* with respect to the double bond, which does not permit independent rotation of *A* or *B* around the bond.

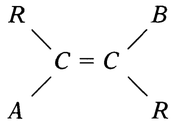

Works Cited

Abkowitz, M., Bässler, H., Stolka, M. (1991), *Philos. Mag. B* **63,** 201–220.

Abkowitz, M., Facci, J. S., Stolka, M. (1993), *Appl. Phys. Lett.* **63,** 1892–1894.

Bässler, H., Brandl, V., Deussen, M., Göbel, E. O., Kersting, R., Kurz, H., Lemmer, U., Mahrt, R. F., Ochse, A. (1995), *Pure Appl. Chem.* **67,** 377–385.

Bässler, H. (1993), *Phys. Status Solidi B* **175,** 15–56.

Bässler, H. (1981), *Phys. Status Solidi B* **107,** 9–54.

Bässler, H., Herrmann, G., Riehl, N., Vaubel, G. (1969), *J. Phys. Chem. Solids* **30,** 1579–1585.

Bonham, J. S., Jarvis, D. H. (1978), *Aust. J. Chem.* **31,** 2103–2115.

Borsenberger, P., Rossi, M. (1992), *J. Chem. Phys.* **96,** 2390–2394.

Borsenberger, P. M., Weiss, D. S. (1993), *Organic Photoreceptors for Imaging Systems,* New York: Dekker.

Braun, C. L. (1984), *J. Chem. Phys.* **80,** 4157–4161.

Brédas, J. L., Street, G. B. (1985), *Acc. Chem. Res.* **18,** 309–315.

Burroughes, J. H., Bradley, D. D. C., Brown, A. R., Marks, R. N., Mackey, K., Friend, R. H., Burns, P. L., Holmes, A. B. (1990), *Nature* **347,** 539–541.

Burroughes, J. H., Jones, C. A., Lawrence, R. A., Friend, R. H. (1990), in: J. L. Brédas, R. R.

Chance (Eds.), *Conjugated Polymeric Materials: Opportunities in Electronics, Optoelectronics, and Molecular Electronics,* Dordrecht, the Netherlands: Kluwer, p. 221.

Chance, R. R., Braun, C. L. (1973), *J. Chem. Phys.* **59,** 2269–2272.

Chyla, A., Sworakowski, J., Szczurek, A., Brynda, E., Nešpůrek, S. (1993), *Mol. Cryst. Liq. Cryst.* **230,** 1–6.

Duke, C. B. (1982), "The Electronic Structure of Semiconducting Polymers," in: Joel S. Miller (Ed.), *Extended Linear Chain Compounds,* Vol. 2, New York: Plenum, pp. 59–125.

Frank, R. I., Simmons, J. G. (1967), *J. Appl. Phys.* **38,** 832–840.

Frankevich, E. L., Lymarev, A. A. (1992), *Mol. Cryst. Liq. Cryst.* **218,** 103–108.

Gailberger, M., Bässler, H., (1991), *Phys. Rev. B.* **44,** 8643–8651.

Gartstein, Y. N., Conwell, E. M. (1994), *Chem. Phys. Lett.* **217,** 41–47.

Geacintov, N. E., Pope, M., Fox, S. (1970), *J. Phys. Chem. Solids* **31,** 1375–1379.

Gerischer, H., Willig, F. (1976), *Top. Curr. Chem.* **61,** 31–84.

Grobman, W. D., Koch, E. E. (1979), in: L. Ley, M. Cardona (Eds.), *Photoemission in Solids II,* Topics in Applied Physics Vol. 27, New York: Springer-Verlag.

Heeger, A. J., Kivelson, S., Schrieffer, J. R., Su, W.-P. (1988), *Rev. Mod. Phys.* **60,** 781–850.

Heeger, A. J. (1992), *Conjugated Polymers and Related Materials: Interconnection of Chemical and Electronic Structure,* Oxford, U.K.: Oxford Univ. Press.

Helfrich, W., Mark, P. (1963), *Z. Phys.* **171,** 527–536.

Helfrich, W., Schneider, W. G. (1965), *Phys. Rev. Lett.* **14,** 229–231.

Heller, A. (1992), *J. Phys. Chem.* **96,** 3579–3587.

Holmes, A. B., Burn, P. L., Kraft, A., Friend, R. H., Bradley, D. D. C. (1992), *Nature* **356,** 47–49.

Hong, K. M., Noolandi, J. (1978), *J. Chem. Phys.* **69,** 5026–5039.

Karl, N., Schmid, E., Seeger, M. (1970), *Z. Naturforsch.* **25a,** 382–391.

Karl, N. (1990), in: K. Sumino (Ed.), *Defect Control in Semiconductors,* North-Holland: Elsevier, pp. 1725–1746.

Kepler, R. G., Anderson, R. A. (1992), *Adv. Phys.* **42,** 1–57.

Kepler, R. G., Soos, Z. G. (1993), in: T. Kobayashi (Ed.), *Relaxation in Polymers,* Singapore: World Scientific, p. 100–133.

Knox, R. S. (1963), *Theory of Excitons,* New York: Academic.

Lampert, M. A., Mark, P. (1970), *Current Injection in Solids,* New York: Academic.

Meerholz, K., Volodin, B. L., Sandalphon, Kippelen, B., Peyghambarian, N. (1994), *Nature* **371,** 497–500.

Merrifield, R. E. (1968), *Acc. Chem. Res.* **1,** 129–135.

Miller, J. S., Epstein, A. J. (1994), *Angew. Chem. Int. Ed. Engl.* **33,** 385–415.

Miller, R. B., Michl, J. (1989), *Chem. Rev.* **89,** 1359–1410.

Mills, A. P., Jr., Karl, N. (1993), *Phys. Rev. B.* **48,** 7050–7056.

Moerner, W. E., Silence, S. M. (1994), *Chem. Rev.* **94,** 127–155.

Mort, J., Pfister, G. (1982) *Electronic Properties of Polymers.* New York: Wiley.

Nešpůrek, S. (1993), *Synth. Met.* **61,** 55–60.

Noolandi, J. (1982), *J. Electrostat.* **12,** 13–25.

Onsager, L. (1938), *Phys. Rev.* **54,** 554–557.

Orlowski, T. E., Scher, H. (1983), *Phys. Rev. B* **27,** 7691–7702.

Pope, M., Kallmann, H., Giachino, J. (1965), *J. Chem. Phys.* **42,** 2540–2543.

Pope, M., Kallmann, H. P., Chen, A., Gordon, P. (1962), *J. Chem. Phys.* **36,** 2486–2492.

(PS) Pope, M., Swenberg, C. E. (1982), *Electronic Processes Crystals,* New York: Oxford Univ. Press; (1997) *Electronic Processes in Organic Crystals and Polymers,* New York: Oxford Univ. Press.

Ramirez, A. P. (1994), *Superconductiv. Rev.* **1,** 1–101.

Ries, B., Schönherr, G., Bässler, H., Silver, M. (1983), *Philos. Mag.* **48,** 87–106.

Roberts, G. (1990), *Langmuir-Blodgett Films,* New York: Plenum.

Sariciftci, N. S., Heeger, A. J. (1994), *Int. J. Mod. Phys. B* **8,** 237–274.

Schein, L. B. (1988), *Electrophotography and Development Physics,* Heidelberg: Springer-Verlag.

Seki, K. (1989), *Mol. Cryst. Liq. Cryst.* **171,** 255–270.

Silence, S. M., Twieg, R. J., Bjorklund, G. C., Moerner, W. E. (1994), *Phys. Rev. Lett.* **73,** 2047–2050.

Silinsh, E. A. (1980), *Organic Molecular Crystals; Their Electronic State,* Berlin: Springer-Verlag.

(SC) Silinsh, E. A., Capek, V. (1994), *Organic Molecular Crystals: Interaction, Localization, Transport Phenomena,* New York: American Institute of Physics.

Stasiak, M., Jeszka, J. K., Zielinski, M., Plans, J., Kryszewski, M. (1980), *J. Phys. D.: Appl. Phys.* **13, L221–L224.**

Tang, C. W., Van Slyke, S. A. (1987), *Appl. Phys. Lett.* **51,** 913–915.

Vardeny, Z. V., Rothberg, L. J. (Eds.) (1994), Proceedings of 2nd International Conference on Optical Probes of Conjugated Polymers and Fullerenes, *Mol. Cryst. Liq. Cryst.* **256,** 1–932.

Wöhrle, D., Meissner, D. (1991), *Adv. Mater.* **3,** 129–138.

Yang, Y., Heeger, A. J. (1994), *Appl. Phys. Lett.* **64,** 1245–1247.

Yu, G., Zhang, C., Heeger, A. J. (1994), *Appl. Phys. Lett.* **64,** 1540–1542.

Zasadzinski, J. A., Viswanathan, R., Madsen, L., Garnaes, J., Schwartz, D. K. (1994), *Science* **263,** 1726–1733.

Further Reading

Borsenberger, P. M., Weiss, D. S. (1993), *Organic Photoreceptors for Imaging Systems,* New York: Dekker.

Brédas, J. L., Chance, R. R., (Eds.) (1990), *Conjugated Polymeric Materials: Opportunities in Electronics, Optoelectronics, and Molecular Electronics,* Dordrecht, the Netherlands: Kluwer.

Duke, C. B. (1982), "The Electronic Structure of Semiconducting Polymers," in: Joel S. Miller (Ed.), *Extended Linear Chain Compounds,* Vol. 2, New York: Plenum, pp. 59–125.

Gerischer, H., Willig, F. (1976), *Top. Curr. Chem.* **61,** 31.

Grobman, W. D., Koch, E. E. (1979), in: L. Ley, M. Cardona (Eds.), *Photoemission in Solids II,* Topics in Applied Physics Vol. 27, New York: Springer-Verlag.

Heeger, A. J. (1992), *Conjugated Polymers and Related Materials: Interconnection of Chemical and Electronic Structure,* Oxford, U.K.: Oxford Univ. Press.

Karl, N. (1990), in: K. Sumino (Ed.), *Defect Control in Semiconductors,* North-Holland: Elsevier, 1725–1746.

Knox, R. S. (1963), *Theory of Excitons,* New York: Academic.

Lampert, M. A., Mark, P. (1970), *Current Injection in Solids,* New York: Academic.

Miller, J. S., Epstein, A. J. (1994), *Angew. Chem. Int. Ed. Engl.* **33,** 385–415.

Moerner, W. E., Silence, S. M. (1994), *Chem. Rev.* **94,** 127–155.

Mort, J., Pfister, G. (1982), *Electronic Properties of Polymers,* New York: Wiley.

Mott, N. F., Davis, E. A. (1979), *Electronic Processes in Non-Crystalline Materials,* Oxford, U.K.: Clarendon.

Noolandi, J. (1987), "Stochastic Theory of Electron-Hole Transport and Recombination in Amorphous Materials," in: G. R. Freeman (Ed.), *Kinetics of Non-Homogeneous Processes,* New York: Wiley-Interscience, pp. 465–525.

Pope, M., Swenberg, C. E. (1984), *Annu. Rev. Phys. Chem.* **35,** 613–656.

(PS) Pope, M., Swenberg, C. E. (1982), *Electronic Processes in Organic Crystals,* New York: Oxford Univ. Press.

Schein, L. B. (1988), *Electrophotography and Development Physics,* Heidelberg: Springer-Verlag.

Silinsh, E. A. (1980), *Organic Molecular Crystals; Their Electronic State,* Berlin: Springer-Verlag.

(SC) Silinsh, E. A., Capek, V. (1994), *Organic Molecular Crystals: Interaction, Localization, Transport Phenomena,* New York: American Institute of Physics.

Vardeny, Z. V., Rothberg, L. J. (Eds.) (1994), Proceedings of 2nd International Conference on Optical Probes of Conjugated Polymers and Fullerenes, *Mol. Cryst. Liq. Cryst.* **256,** 1–932.

SENSORS, ACOUSTIC

LAWRENCE C. LYNNWORTH, *Panametrics, Inc., Waltham, Massachusetts, U.S.A.*

	Introduction	521
1.	**Principal Measurands in Process Control**	523
1.1	Flow	523
1.1.1	Flow Profile	523
1.1.2	Single-Phase Fluids	523
1.2	Level of Liquids	526
1.2.1	Nonintrusive Methods	526
1.2.2	Intrusive Methods	528
1.3	Temperature	529
1.3.1	Average Temperature	529
1.3.2	Temperature Profile along One Path	531
1.3.3	Tomographic Reconstruction of Profile from Averages Obtained over Many Scanned Paths	531
1.4	Pressure	534
1.4.1	Fluids	534
1.4.2	Solids under Stress	536
2.	**Other Measurands**	536
2.1	Density of Gases; Density of Liquids	536
2.1.1	Gas Density	536
2.1.2	Liquid Density	537
2.2	Concentration Analysis of Binary Gas Mixtures	537
2.3	Designing Pipes and Tanks to Accommodate Sensors, Particularly Ultrasonic Sensors	539
	Works Cited	539
	Further Reading	540

INTRODUCTION

Acoustic sensors may be understood to mean transducers, devices, or systems that generate, detect, or measure mechanical vibrations. These vibrations may be characterized in terms of frequency, bandwidth, information content, intensity, polarization, speed of propagation, and the rate at which they are attenuated. Mechanical vibrations or elastic waves interact with, influence, or are influenced by other elastic waves and also by other forms of energy (thermoacoustic or optoacoustic interactions), a material's elastic properties and stress, and metallurgical and chemical processes (ultrasonic welding, sonochemistry). Acoustic waves are influenced by gradients and inhomogeneities in, and by the motion of, the medium in which they are propagating (turbulent or particulate scattering, beam drift, Doppler shift), and by the boundaries of that medium, or boundary conditions.

Ultrasonic refers to elastic waves of frequency $f \geq 20$ kHz. Ultrasonic waves are thus vibrations at a frequency above human hearing. With respect to sensing and measurements, high frequency avoids interference from many audible, low-frequency noises due to wind, machinery, pumps, and vibration of large bodies. High frequency allows resolution of "the small" in both the temporal and spatial senses. Attenuation often imposes the upper limit on the maximum usable frequency. *Audible* frequencies, $16 \text{ Hz} < f < 20$ kHz, are sometimes selected to overcome attenuation or to limit propagation to one mode in a plate. The parameter that is to be measured is called the *measurand.*

The *sensor* may be a system that includes one or more transducers and the medium itself or an existing internal or bounding structure. The sensor system may also be a foreign sensor, intrusive, added to the existing process to make the required measurement. The four main measurands in process control are flow, pressure, temperature, and level. (To the extent that flow measurements include mass flow rate and flow profile, the rheological properties of fluids such as density and viscosity may be viewed as falling occasionally under the "flow" umbrella.) To sense these measurands by ultrasonic pulse techniques, one can measure propagation in the medium itself (e.g., speed of sound in air

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

is proportional to the square root of the air's absolute temperature), or in a "sensor" that touches the medium. Rather than perturb the situation, one prefers to exploit an existing structural element as the sensor, or as a reflector that bounds the sensed medium at a known location. A pipe wall or an entire section of pipe, for example, can sometimes be used as a temperature sensor, a fluid-density sensor, or an acoustic window that allows one to interrogate the fluid within the pipe noninvasively. The fluid then serves as its own "sensor." On the other hand, to avoid uncertainties attributed to variability in existing structural elements not expressly designed nor quality controlled for sensing, one may need to introduce a foreign sensor [or foreign reflector(s)] for measurement purposes. This invasive sensor is often a resonator or waveguide—some form of probe—in which propagation is a unique and repeatable function of the measurand in the medium. The invasive or "wetted" sensor may also consist of one or more transducers that are each acoustically isolated from the medium's boundary, and hence from one another if two transducers are used, so that the medium can be interrogated or listened to without interference from boundary-borne noises.

The discussion of acoustic sensors is limited here to technical equipment and practical applications in the frequency range 10^3 to 10^9 Hz, and to "small" amplitudes (particle displacement $\ll$ wavelength). States of matter will include gas, liquid, and solid media and multiphase mixtures. Emphasis is placed on

1. two decades of the ultrasonic spectrum, between 50 kHz and 5 MHz;
2. measurement of the four principal measurands in the field of industrial process control, with emphasis on flow and liquid level in closed conduits and tanks, respectively; and
3. sensing of such measurands based on their influence on the propagation of ultrasound.

Waves of interest include longitudinal, transverse shear, extensional, torsional, and plate waves (Lamb, Rayleigh). It will be seen that appropriate utilization of transducers often allows one to interpret wave speed, reflections, transmission, and attenuation in terms of one or more measurands. "Appropriate utilization" includes selecting the proper piezoelectric material (Gualtieri *et al.*, 1994), vibration mode, transducer type and location, coupling means, probing means, electronic excitation, detection, and display or output means, etc.

Acoustical waves are often selected as the means to measure a particular parameter. One example is measuring the distance to a moving web or sheet. The sound waves can accomplish this task without contacting the web. Ultrasound often gets the job done with no penetration of the boundary (e.g., clamp-on transducers) or with minimal disturbance of the measurand even when the boundary must be penetrated (e.g., wall-mounted transducers for an across-the-stack flow measurement). Another example is downward-looking "air sonar" to measure the level of liquids in a tank or solids in a bin. Process-control engineers may select acoustical waves to sense process measurands when they need

- noninvasive, noncontact, or minimally invasive measurement;
- high accuracy (measurands usually transformed to time or frequency measurements);
- reliability (no moving parts, in the usual sense);
- fast response (can be <1 ms);
- remote sensing, sometimes with no physical contact;
- average reading over an extended region;
- profile information (point by point, or small-path average);
- computer compatibility of time, frequency, or, say, 8-bit amplitude data;
- low cost, especially for multiplexed and/or mass-produced sensors;
- small size, small mass;
- avoidance of problem(s) associated with competing (nonultrasonic) technologies;
- data or results unobtainable any other way; and
- useful sensing of two or more parameters simultaneously, e.g., flow velocity and composition, density and viscosity, or flow velocity and density.

The seven basic elements in a typical active acoustical sensing system may be represented as [Electrical Energy Source] → [Electronic Transmitter] → [Electroacoustic

Transmitting Transducer] → [Wave–Measurand Interaction Zone] → [Acoustoelectric Receiving Transducer] → [Electronic Receiver] → [Information Output]. These [items] or [functions] will become clear in relation to measuring specific measurands. An abbreviated representation of an active acoustical sensing system is [Transmit] → [Interact] → [Receive].

Much of the material in this article is condensed from Lynnworth (1989, 1994a). That book and chapter, respectively, or the references cited therein, may be referred to for more technical or historical details than could be included here. The present article, however, includes some new developments, new data, and references to works published in the period 1992 to early 1995. See also Papadakis (1997). In this article, space restrictions generally make it impractical to identify every concept with its inventor or author. Manufacturers are identified and specific products and model numbers for measuring flow, liquid level, temperature, pressure, etc., by ultrasonic as well as nonultrasonic sensors are indexed or reviewed annually in several periodicals, e.g., in the U.S. journals *m&c (Measurements & Control)* and *Sensors,* and in archival periodicals such as *Journal of the Acoustical Society of America, Journal of Sound and Vibration, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control,* and *Acoustica.* Journals of interest that are published in the United Kingdom include *Measurement Science & Technology* and *Ultrasonics.*

1. PRINCIPAL MEASURANDS IN PROCESS CONTROL

1.1 Flow

All four principal process-control measurands have been analyzed and measured by acoustics, but acoustic measurement of flow is probably the subject of more studies than pressure, temperature, and level combined. A study of the issues involved in measuring flow by transmission and reflection methods provides a base for understanding many other acoustic/ultrasonic measurements.

1.1.1 Flow Profile Flow velocity is theoretically zero at the walls of a conduit and maximum at the center or at some other point. One is usually interested in *total* volumetric flow $Q = V_A A$ or total *mass* flow $M_f = \rho Q$, where V_A = area-averaged flow velocity, A = duct area, and ρ = fluid density. Now V_A can be thought of as the area-weighted average of V_i's in annuli or pixels of areas A_i. One can try to measure V_i in many small regions of area dA and then integrate over A, or one can measure over one or more paths and relate the path averages V_P to V_A. Range-gated Doppler measurement is one way to get to $Q = \int_A V_i dA$. Use of a meter factor K leads to $V_A = KV_P$. Special paths or quadrature methods yield V_A largely independent of flow profile. Examples include midradius paths and Gauss–Chebyshev paths. If the various paths that are to be interrogated differ in transit time by more than the ringdown time of the transducers, it is possible to interrogate a number of these paths simultaneously by connecting sets of transmitting or receiving transducers in parallel electrically (Lynnworth, 1994c, 1995). In some cases interrogation is possible over the entire area A all at once. Examples of recent European research on the influence of flow disturbances are Fleury (1994) and Holm (1995). If the flow is irrotational, the circulation Γ can be measured along a closed path such as the inscribed equilateral triangle in a round pipe (Smith, 1994; Smith *et al.*, 1995).

In general, flow velocity is not only a function of position but also a function of time t. Acoustic flowmeters tend to be linear and so yield the correct time-averaged value for $V(t)$, unlike square-law devices that yield the rms value. Acoustic flowmeters are also inherently faster in response than sensor elements having inertia. [See Hamidullin (1993).] Regarding clamp-on, see, for example, Baumoel (1994).

1.1.2 Single-Phase Fluids In the contrapropagation method, ultrasonic (or sometimes audible) waves are transmitted upstream and downstream. From the transit times t_1 and t_2 in each direction, and knowledge of the path and flow profile, the average flow velocity V_A is determined. A rather simple derivation of the basic flow-sensing equation is possible if one imagines a fluid of sound speed c flowing at a uniform velocity $V < c$ in a duct of cross-sectional area A, interrogated by two point sensors on the axis spaced a distance L apart. The transit times in the upstream and downstream directions, respectively, are

$$t_1 = L/(c - V) \quad \text{and} \quad t_2 = L/(c + V). \tag{1}$$

The reciprocals of these transit times, multiplied by L, are

$$L/t_1 = c - V \quad \text{and} \quad L/t_2 = c + V. \tag{2}$$

Accordingly,

$$V = \frac{L}{2}\left(\frac{1}{t_2} - \frac{1}{t_1}\right) = \frac{L}{2}\left(\frac{\Delta t}{t_1 t_2}\right) \quad \text{and}$$
$$c = \frac{L}{2}\left(\frac{1}{t_2} + \frac{1}{t_1}\right) = \frac{L}{2}\left(\frac{\Sigma t}{t_1 t_2}\right). \tag{3}$$

The upstream–downstream time difference can be obtained from Eq. (1) as $\Delta t = 2LV/(c^2 - V^2)$. This can be expressed in terms of the Mach number $M_s = V/c$ for $M_s \ll 1$:

$$\Delta t = \frac{2LV/c^2}{1 - M_s^2} = (2LV/c^2)(1 + M_s^2 + M_s^4 + \cdots). \tag{4}$$

At sufficiently small Mach numbers, the following approximations are valid:

$$V \approx c^2 \Delta t/2L \tag{5}$$

and

$$\Delta t \approx 2LV/c^2. \tag{6}$$

For liquids, V is usually $\ll c$, but for gases in the tall stacks of 100-MW fossil-fueled electric utilities (Matson and Davis, 1994) or in offshore flare lines (Folkestad and Mylvaganam, 1993), it is not uncommon to find $V/c > 0.1$ at the time that flow measurements are required. If the flow is interrogated obliquely, or if refraction or beam drift occurs, the above equations must be re-examined.

Differences between ultrasonic measuring systems for gas flow and liquid flow stem from differences in kinematic and acoustic properties of these fluids. The densities of air and water, for example, are 1.3 and 1000 kg/m^3, respectively. Examples of contrapropagation flowmeters for liquids and gases are shown in Figs. 1 through 3. (See Liu *et al.*, 1996.)

Prior to about 1988, contrapropagation was limited to single-phase fluids. However, the use of cross-correlation detection has extended this method to two-phase cases (Jacobson *et al.*, 1988; Hayward, 1994). Conversely, improvements in reflection methods have enabled such equipment to measure "clear" liquids in which the dominant scatterers might be only turbulent eddies. By 1994 contrapropagation had been applied to single-phase liquids and gases up to 260 °C and to cryofluids such as liquid N_2 down to approximately −200 °C. The flow of liquids, even at these temperature extremes, can be measured using the *clamp-on* method, if one has the appropriate couplant and coupling method. Soft-metal foil and high coupling pressure constitute one of the solutions at temperature extremes. Measuring the flow of gases, however, generally requires an opening in the pipe, unless the pipe is plastic, or the gas pressure is $\gtrsim$100 bar. Exceptions may occur for special pipe materials and geometries, or where opportunities exist to subtract electronically the pipe-borne acoustic cross talk that normally would overwhelm the receiver if one attempted to build a gas clamp-on flowmeter.

As explained by Pollard (1977, pp. 139–140), for signals in the form of a pulse, the cross-correlation coefficient is given by

$$R_{12}(\tau) = \int_{-\infty}^{\infty} f_1(t) f_2(t - \tau) dt. \tag{7}$$

The cross-correlation coefficient is a measure of the correlation or similarity between two (nonidentical) wave forms as one of the signals, $f_2(t)$, is shifted along the time axis past the other signal $f_1(t)$. To understand how cross-correlation can find a signal buried in noise, consider that if the form of the transmitted signal is known, then it may be cross-correlated with the received sum of signal and noise. Since the cross-correlation coefficient for random noise eventually averages out to zero, there remains only the signal in the form of its autocorrelation function. Furthermore, since phase information is retained in computing the cross-correlation coefficient, the value of τ for which $R_{12}(\tau)$ has its peak value is a measure of the transit time of the signal.

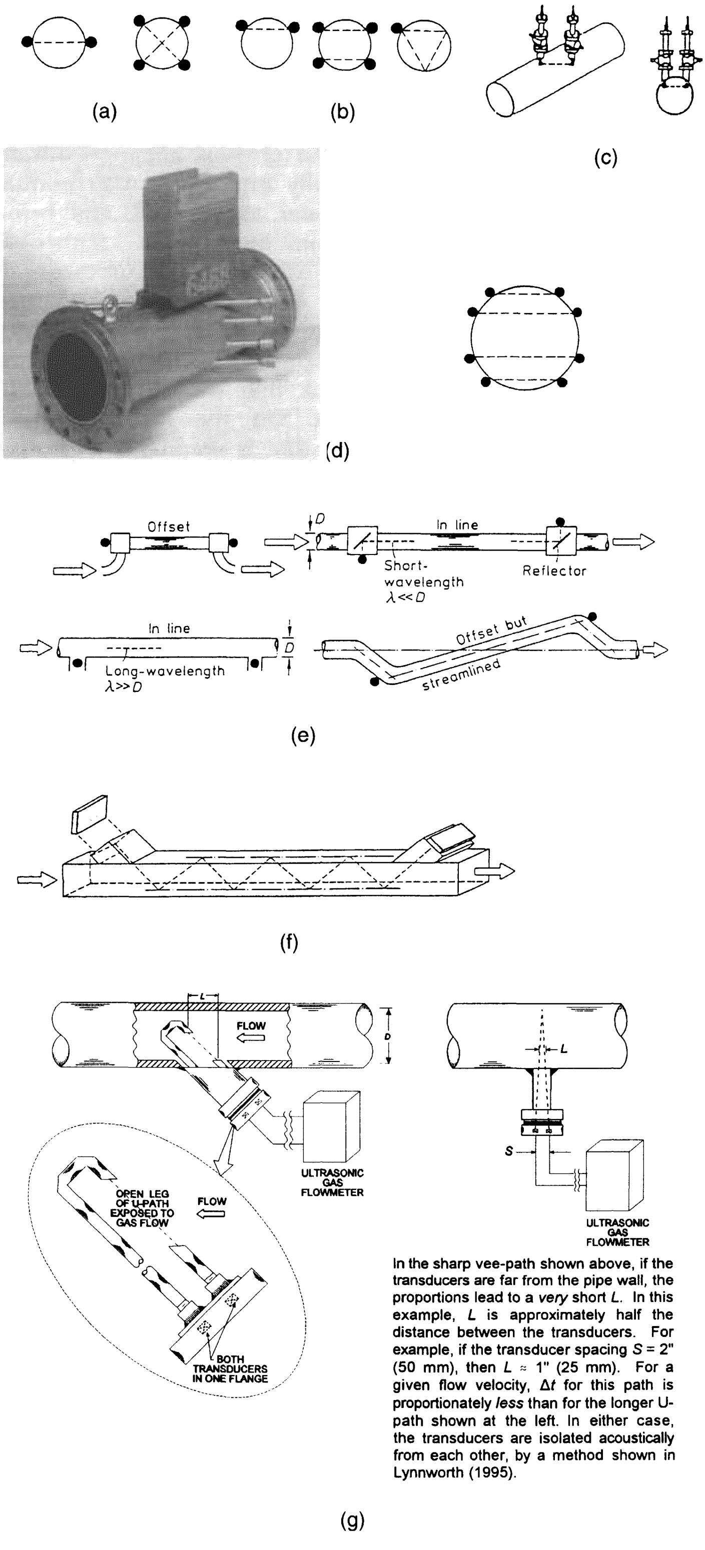
(a)
(b)
(c)
(d)
Offset
In line
D
Short-wavelength
λ<<D
Reflector
In line
D
Long-wavelength
λ>>D
Offset but streamlined
(e)
(f)
L
FLOW
D
OPEN LEG OF U-PATH EXPOSED TO GAS FLOW
FLOW
BOTH TRANSDUCERS IN ONE FLANGE
ULTRASONIC GAS FLOWMETER
L
S
ULTRASONIC GAS FLOWMETER
(g)

In the sharp vee-path shown above, if the transducers are far from the pipe wall, the proportions lead to a *very* short *L*. In this example, *L* is approximately half the distance between the transducers. For example, if the transducer spacing $S = 2"$ (50 mm), then $L \approx 1"$ (25 mm). For a given flow velocity, Δt for this path is proportionately *less* than for the longer U-path shown at the left. In either case, the transducers are isolated acoustically from each other, by a method shown in Lynnworth (1995).

FIG. 1. Examples of transducer locations (●) and interrogation paths (---) for contrapropagation flow measurements along (a) tilted diameter(s), (b) one, two, or three midradius chords, (c) chord segment, (d) Gauss–Chebyshev chords, (e) axial paths that are offset, in-line, and offset but using an uninterrupted streamlined cell, (f) 45° zigzag interrogation of 100% of flowing cross section in a square conduit, and (g) paths requiring only one port. Copyright credits: (a), (b), (e), and (f) VCH, 1994; (c) Texas A&M University, 1984; (d) and (g) Panametrics, 1992 and 1994.

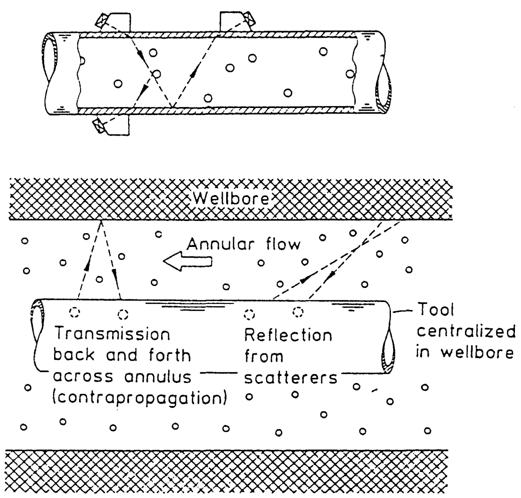
Wellbore
Annular flow
Tool centralized in wellbore
Transmission back and forth across annulus (contrapropagation)
Reflection from scatterers

FIG. 2. Example of transducer (⊠,⊠) locations to measure the flow velocity of a fluid that is intermittently two-phase. Top, ordinary pipe. Bottom, downhole flow tool.

1.2 Level of Liquids

The level of liquids in closed vessels such as tanks or pipes can be measured with audible and ultrasonic elastic waves. The methods currently in use in industry may be divided into nonintrusive and intrusive categories, then divided according to whether the measurement is at one or more discrete levels or continuous, further divided according to the type of wave used (longitudinal, transverse shear, torsional, Lamb, Rayleigh, etc.), further divided according to whether propagation is measured in the liquid, in the liquid's container wall, or in a sensor immersed in the liquid, and lastly divided according to whether the effect of the liquid is primarily an attenuation effect or a transit-time or sound-speed effect. Some special solutions may not fit this categorization exactly; some of the following remarks may apply to measuring the level of solids too, and also to liquids in open channels.

Figures 4 and 5 compare a number of different liquid-level measuring approaches that are organized according to the above categories. As with flow, one would think that the nonintrusive measurement would be preferred over intrusive. The large number of intrusive liquid-level sensors in use, however, suggests that there must be important limitations on nonintrusive methods, or that there must be important advantages to intrusive sensors.

1.2.1 Nonintrusive Methods Nonintrusive methods (clamp-on, bond-on, press-on, weld-on) have the obvious apparent advantages of minimally interfering with normal operations in order to be used, and being transportable from one site to another so that one sensor can yield measurements in different places. For example, a fuel-truck driver could temporarily attach a magnetically clamped high-level sensor to each customer's steel tank that was to be filled, to avoid overfilling. On the other hand, a clamp-on device that is seemingly so simple to use might be misused or abused. Permanent intrusive sensors can be mass produced largely independent of the tank-wall details, as they are typically threaded into a standard pipe fitting.

The nonintrusive techniques that are well known include liquid-presence detecting at a discrete point, based on detecting an echo from the far wall; sensing the change in ringdown when a probe is slid up and down a vessel wall; and sending the sound wave vertically up and timing the round trip. A hybrid case occurs when the nonintrusive transducer is outside the vessel but a stillwell or reflectors are placed inside the vessel to improve the measurement. (A stillwell is a tube of relatively small diameter compared with length, so that sloshing and wavy action are smoothed out or stilled. The stillwell sometimes provides the further advantage of guiding the wave. In other words, the stillwell prevents attenuation caused by beam spread, and avoids spurious echoes from off-axis reflectors. A potential disadvantage is that the density of liquid inside the stillwell could end up being different from that outside the stillwell.)

Nonintrusive techniques that have been overlooked in numerous surveys of liquid-level sensors are those using plate waves and transverse shear waves. One plate-wave solution, due to Royer *et al.* (1992), is attenuation based and makes use of the leakage of symmetric plate waves. Another plate-wave or flexural-wave solution is sound-speed based and makes use of the retarding effect of a liquid adjacent to a tank wall (Liu and Lynnworth, 1993, 1995). This latter plate-

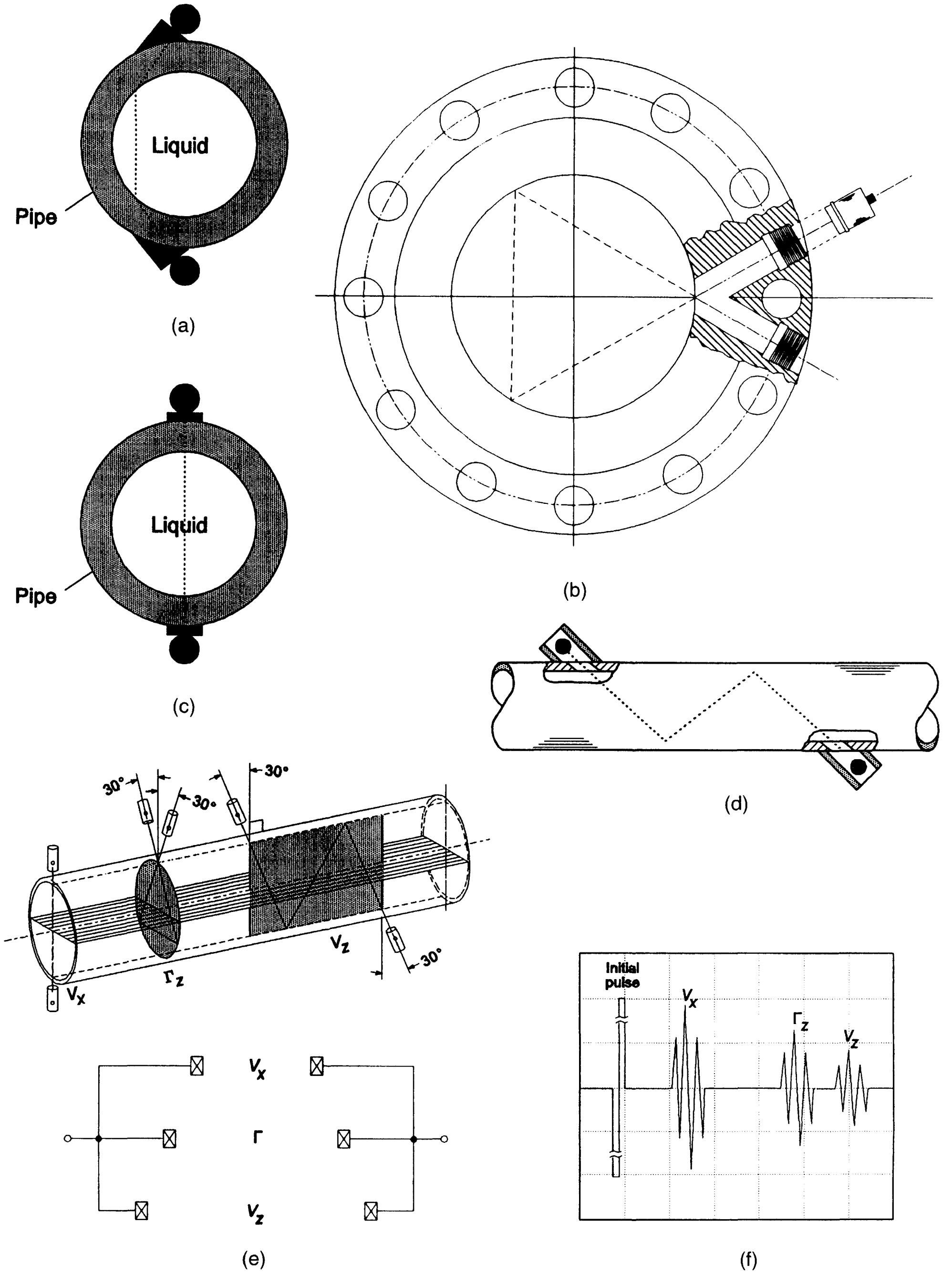
Liquid
Pipe
(a)
(b)
Liquid
Pipe
(c)
(d)
30°
30°
30°
30°
Vz
Γz
Vx
Vx
Γ
Vz
(e)
Initial pulse
Vx
Γz
Vz
(f)

FIG. 3. Transducer locations (●) and paths (---) for measuring (a) liquid swirl presence by clamp-on, (b) circulation Γ_z of a vortex, (c) liquid cross flow V_x by clamp-on, (d) axial flow V_z, (e) the simultaneous measurement of Γ_z, V_x, and V_z in air, using pairs of electrically paralleled, acoustically isolated transducers and paths differing in transit time by more than the ringdown time of the transducers, (f) schematic of oscillogram corresponding to (e). © 1994 Panametrics.

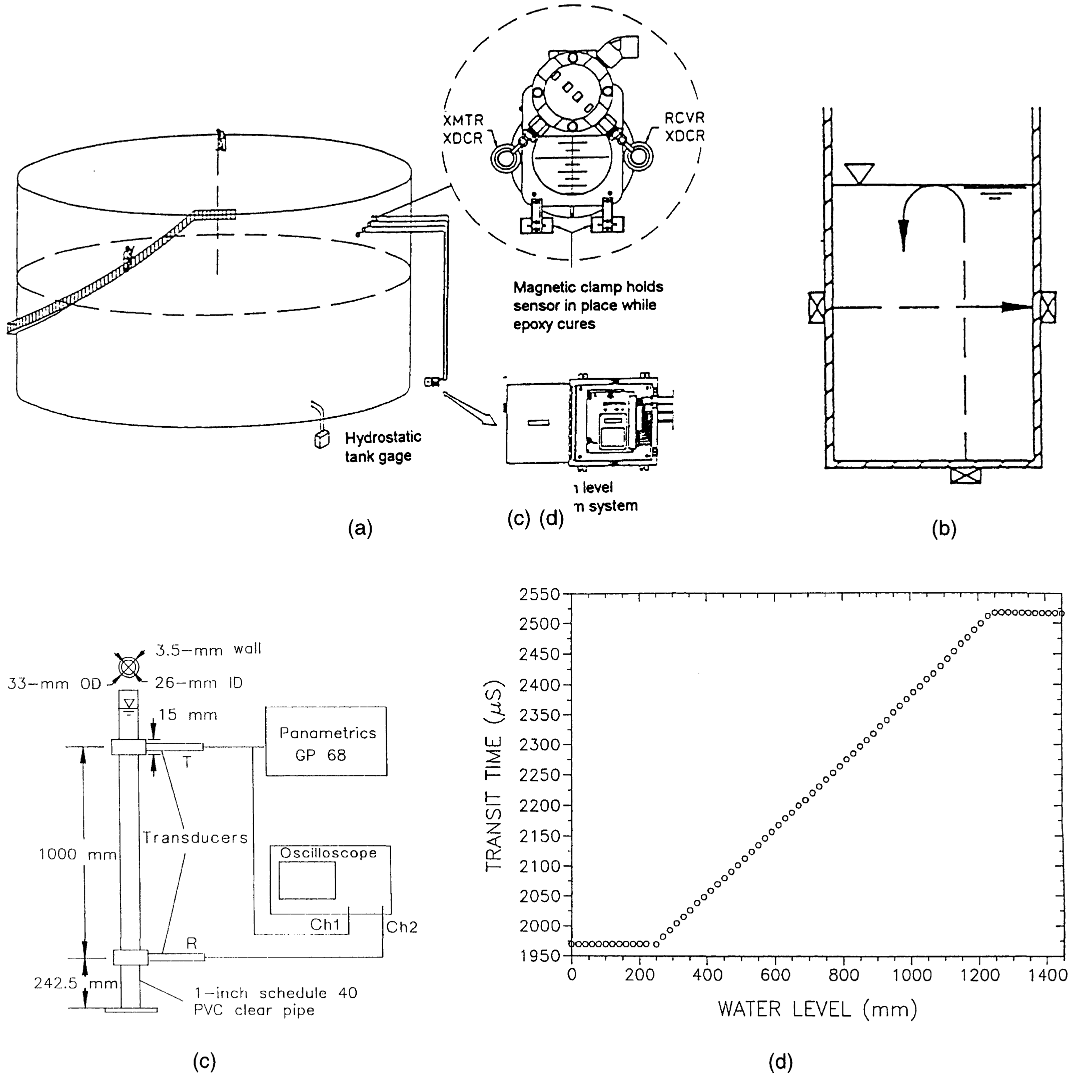
XMTR XDCR
RCVR XDCR
Magnetic clamp holds sensor in place while epoxy cures
Hydrostatic tank gage
ı level
·n system
(a)
(c) (d)
(b)
3.5–mm wall
33–mm OD
26–mm ID
15 mm
Panametrics GP 68
T
Transducers
1000 mm
Oscilloscope
Ch1
Ch2
R
242.5 mm
1–inch schedule 40 PVC clear pipe
TRANSIT TIME (μS)
2550
2500
2450
2400
2350
2300
2250
2200
2150
2100
2050
2000
1950
0
200
400
600
800
1000
1200
1400
WATER LEVEL (mm)
(c)
(d)

FIG. 4. Nonintrusive liquid-level techniques (a) for detecting liquid presence near the top of large fuel-storage tanks, and (b) for discrete and continuous measurements on a relatively small-diameter vessel whose bottom is accessible and not covered with deposits. Method in (c) uses flexural or bending waves, and yielded the data plotted in (d), after Liu and Lynnworth (1993, 1995). © 1993 Panametrics. Although such waves propagate dispersively (Brillouin, 1960) they nevertheless may prove practical.

wave method can be applied with the path horizontal or vertical. In some cases, the choice depends on the objective being discrete or continuous measurements.

1.2.2 Intrusive Methods In the situation shown in Fig. 5(a), the downward-looking sonars may operate free-field or in a multiple-notch stillwell. In either case the transducers do not contact the product, but a hole in the roof is necessary. The hole is sealed by a flange when the sensor or transducer is installed. Recent advances in transducers for use in air, as may be used in downward-looking sonars, are discussed or reviewed by Dabirikhah and Turner (1994), Hutchins (1994), and Mágori (1994). If the product beneath the vapor is a settled liquid that is clean and not bubble- nor residue-bearing, a wetted probe such as a torsional waveguide may be used (Kim *et al.,* 1993). [Some versions of this waveguide may be used to sense liquid density ρ and viscosity η. Whether this is better than alternative and especially noninvasive solutions based on the complex reflection coefficient (Sheen, 1994) or streaming (Hertz *et al.*, 1991) needs to be determined in each case.] Sometimes a magnetic float rides up and down a magnetostrictive waveguide, using the Wiedemann effect to achieve positional or liquid-level accuracy of a fraction of 1 mm over long spans (manufacturer: MTI Corp.).

Tuning forks and gap sensors are in wide use, installed by threading or flanging. Liquid dampens the tuning-fork vibration, an "attenuation" effect, but liquid increases the gap signal compared with the near-zero ultrasonic signal strength transmitted across the vapor. On the other hand, an airborne signal across a wall-to-wall "gap" will be interrupted (attenuated) by the presence of solids piling up in the path. SAW (surface acoustic wave) and plate-wave sensors are often proposed to sense the presence of particular substances (Wenzel *et al.*, 1994).

The upward-looking transducer runs the risk of being buried in sludge, but if cleaned or if used only in clean liquids, it can provide a reliable echo from the liquid–vapor interface. A stillwell stabilizes the interface and can contain reference reflectors to compensate for sound-speed gradients, the importance of which increases if a temperature gradient exists. This need for multiple reflectors is readily appreciated if one recognizes that, for water and some other liquids, sound speed changes by about $\frac{1}{4}\%$ per °C. A 45° internal (immersed) reflector sometimes allows a sidewall-mounted transducer to interrogate vertically. The difficult problem of measuring the level of molten aluminum is discussed by Nygaard and Mylvaganam (1993).

1.3 Temperature

This section covers average temperature, temperature profile, and tomographic reconstruction. In most cases, the temperature is derived from a measurement of sound speed [Fig. (6a)].

1.3.1 Average Temperature In contrast to thermocouples, resistance temperature detectors, integrated-circuit temperature sensors, and other small devices that sense temperature essentially at one point, by using acoustic or ultrasonic waves one can measure the speed of sound over an extended path to obtain an average temperature reading. At sufficiently low carrier frequency (57 Hz), and with a sufficiently intense (221 dB) coded source, this has been done over a distance of 18 000 km, using the ocean's 1- to 2-km-thick *SOFAR* channel, achieving a range that reached halfway around the world (Baggeroer and Munk, 1992; see also Holing, 1994). On a somewhat smaller scale, as in combustors where gas temperature may be of the order of 1000 °C (Sec. 1.3.3), the path lengths may be 10 to 20 m, and the frequency may be 1500 Hz (Kleppe, 1989, 1991; Yori, 1992). In smokestacks at somewhat milder conditions of only 200 °C and flow velocities up the stack of Mach 0.1, the frequency can be higher, reaching up to the lowest octave of ultrasound, 20 to 40 kHz. A frequency as high as 100 kHz was used to measure sound speed and flow velocity across a 3-m path in a smokestack duct at 150 °C (Matson and Davis, 1994).

If a *wire waveguide* is used as the temperature sensor, the frequency used in the past has typically been 100 kHz. A wire waveguide is simply a wire that guides acoustic or ultrasonic waves, something like the string in a "string telephone" stretched between paper cups. A wire can guide sound something like the way a fiber-optic waveguide guides light, or a microwave waveguide guides electromagnetic microwaves. Starting around 1960 various investigators proposed using wire-waveguide sensors as replacements for thermocouples (tc's) in order to reduce the number of sensor elements, to avoid tc errors attributed to insulator failure at very high temperature, to avoid tc diffusion errors, and to achieve a more rugged design. In some nuclear-fuel pin studies, some or all of these potential advantages were realized. However, at the time those studies were conducted, no practical way was found to extend the results to analogous industrial applications in an economic fashion. In recent times, however, the widespread use of microprocessors and of ultrasonic high-precision intervalometer instruments for other purposes, principally flow and liquid level, means that it may soon be practical to extend to temperature measurements the use of electronic instruments that

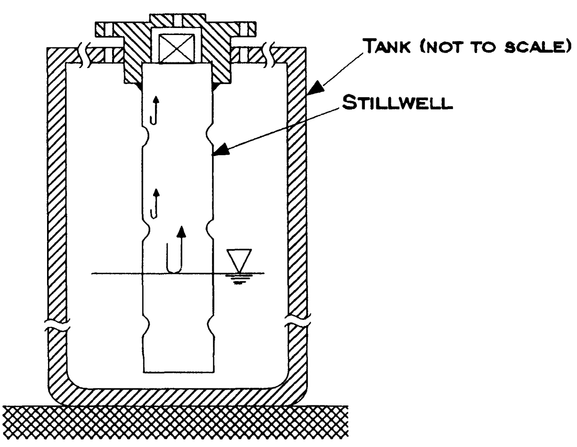

(i) Notched stillwell guides the waves, in a design manufactured by Saab Tank Control

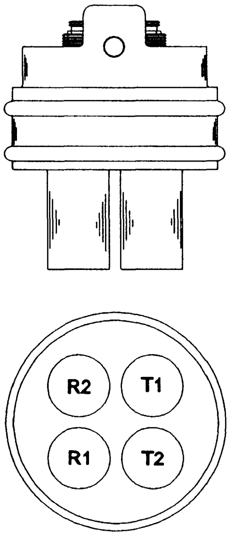

(ii) Transducer array transmits unguided sound beam, then waits for echoes. This type transducer was originally developed for wall-to-wall flow measurements in large smokestacks (Matson and Davis, 1994).

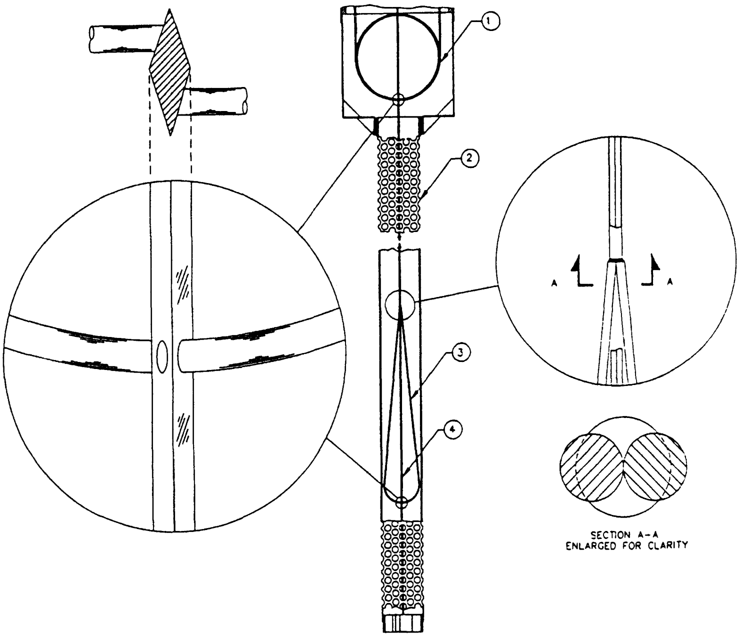

(iii) Torsional waveguide, after Kim et al. (1993). Legend: 1 -Top lead-in/out wires; 2 -Perforated stillwell; 3 -Bottom lead-in/out wires; 4 - Diamond cross-section waveguide.

(a)

FIG. 5. Intrusive liquid-level techniques. (a) Downward-looking sonar using (i) guided or (ii) unguided waves, generated and detected using an array of four transducers operating in pairs, and (iii) torsional waveguide top-mounted sensors. (Details in Kim *et al.*, 1993.) [Regarding (i) an early downward-looking sonar with reference reflectors is due to Tomioka (1968).] (b) Side-mounted gap probe. (c) Upward-looking transducers with or without reflector.

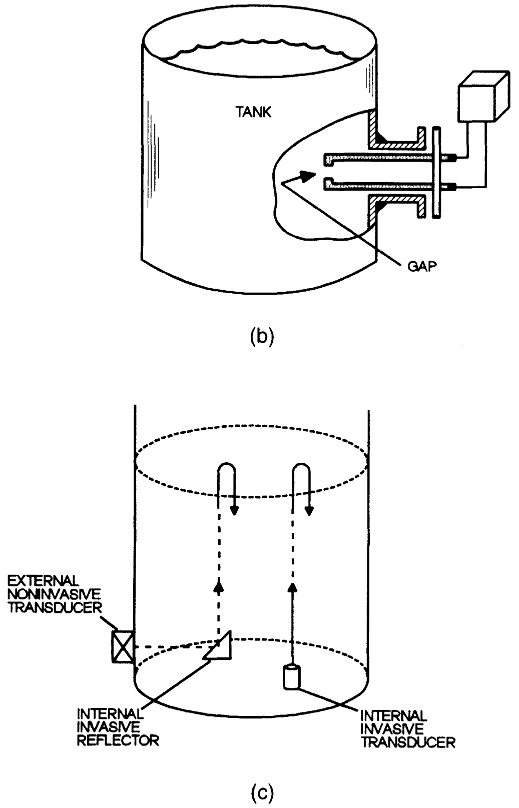
TANK
GAP
(b)
EXTERNAL NONINVASIVE TRANSDUCER
INTERNAL INVASIVE REFLECTOR
INTERNAL INVASIVE TRANSDUCER
(c)

FIG. 5. Continued.

were originally developed for one of the other major process-control measurands. Sometimes the sensor is a waveguide added to the system; sometimes an existing structural element or process boundary can be the sensor [Figs. 6(b)–6(g)].

1.3.2 Temperature Profile along One Path Shortly after the wire waveguide was introduced into the United States as a candidate sensor to measure temperature in certain extreme environments, it was recognized that by intentionally creating a series of discontinuities along the waveguide, one could measure temperature profile. By 1970 this had been demonstrated in laboratory prototype sensors (Lynnworth and Patch, 1970). In other words, the "average" temperature measurement of Sec. 1.3.1, utilizing the sound-velocity–temperature data represented by the thin-wire curves in Figure 6(a), can be applied to the smaller path lengths of a multizone sensor such as that in Fig. 6(b). It is also possible to "tap" a multizone wire waveguide at *n* points to extract profile information (Kim *et al.*, 1993), somewhat analogous to picking up signals at stations along the 18 000-km *SOFAR* path mentioned above.

As an example of how an ultrasonic "flowmeter" can be applied to multizone and multisensor measurements of sound velocity in wire waveguides, see Collard (1990). While Collard's 1991 dissertation was concerned with determining elastic moduli at high temperature, as was Bell's pioneering work (1957), there are many similarities between such work and measurements of temperature. Such similarities, in fact, were recognized in the United Kingdom by 1960 and provided the stimulus for the earliest wire-waveguide (average) temperature-sensor developments by Bell, Hub, Thorne, Jacques, and colleagues in that decade. Those investigators, particularly Thorne, also demonstrated how melting-point indicators (e.g., solder in a small hole, or solder between the waveguide and a surrounding cylindrical mass) could be built into a waveguide, to provide a precise measurement of when a particular temperature was reached at a particular point along a waveguide. Fendrock and Varela (1995) describe a waveguide thermometer instrument used to measure furnace temperature profiles.

1.3.3 Tomographic Reconstruction of Profile from Averages Obtained over Many Scanned Paths Audible sound and ultrasound offer potential advantages over competing technologies in applications that require long-path averages and in temperature-profile determinations based on a multizone path or tomographic reconstruction. The "sensor" may be gas, liquid, or solid [Fig. 6(a)]. A number of multizone waveguide applications are reviewed in Lynnworth (1989). Many of these used a waveguide similar to that shown in Fig. 6(b). Methods of isolating the electrical part of the assembly or the sensor from the environment are shown in Figs. 6(c) and 6(d).

Since the mid-1980s, CAT scanning of large combustors and boilers has been proven practical. Early proposals from about 1970, as well as one of the first large-scale working examples, due to Green (1985), are shown in Fig. 7. In the first two schematics of Fig. 7(a), lines with arrows represent the

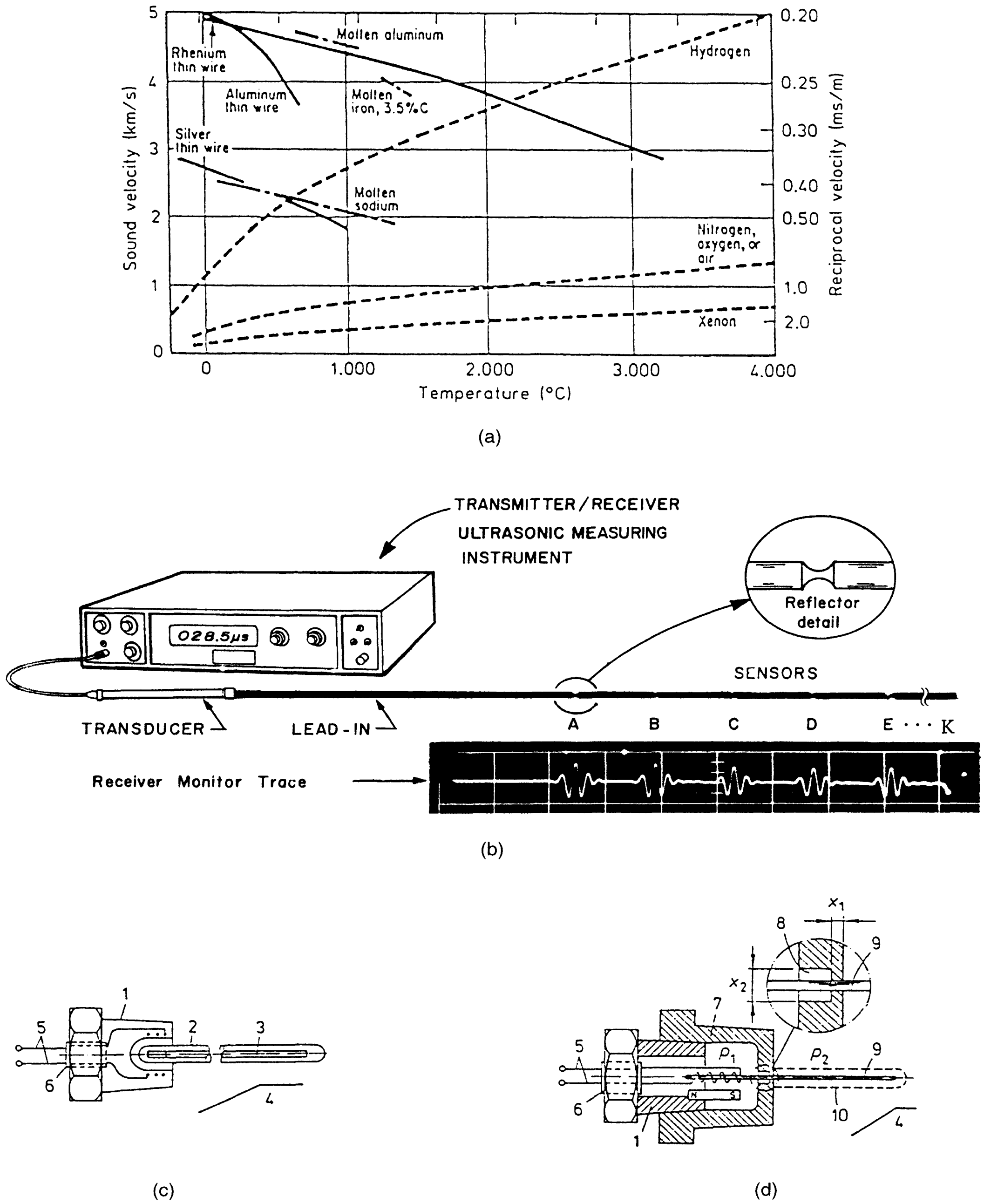

(a)

(b)

(c)

(d)

FIG. 6. (a) Sound speed *c* as a function of temperature *T* in gases, liquids, and solids. (b) An ultrasonic *T* profiler waveguide comprised of ten zones, after Lynnworth and Patch (1970). Echoes are obtained from reflective grooves A, B, . . . , K. The time between echoes depends on sound speed in the segment from whose ends the echoes are generated. The data in (a) are used to convert sound speed measured in each zone to the average temperature in that zone. (c) Magnetic feedthrough: 1, secondary seal plug; 2, sealed all-metal sheath protecting sensor (3) from environment (4); 5, leadwires passing through glass-to-metal or other feedthrough (6). (d) Diaphragm feedthrough, after Lynnworth (1994a), © 1994 VCH. Legend: 7, plug contains thin-wall cavity (8) where x_1 and x_2 are chosen to withstand pressure difference $\Delta P = P_2 - P_1$ and other stresses; sensor (9) passes through diaphragm, to which it is welded or brazed; 10, optional sheath. The bar magnet (N–S) biases the magnetostrictive transducer material into its linear range (analogous to poling of ferroelectric material) and orients its domains so that the generated wave will be purely the extensional mode (no torsion). Thermometer designs proposed for bakery: (e) Moving hearth plate is its own sensor. (f) Rolling drum is its own sensor. (g) Multizone waveguide sensor. After Lynnworth (1992). © 1992 John Wiley & Sons.

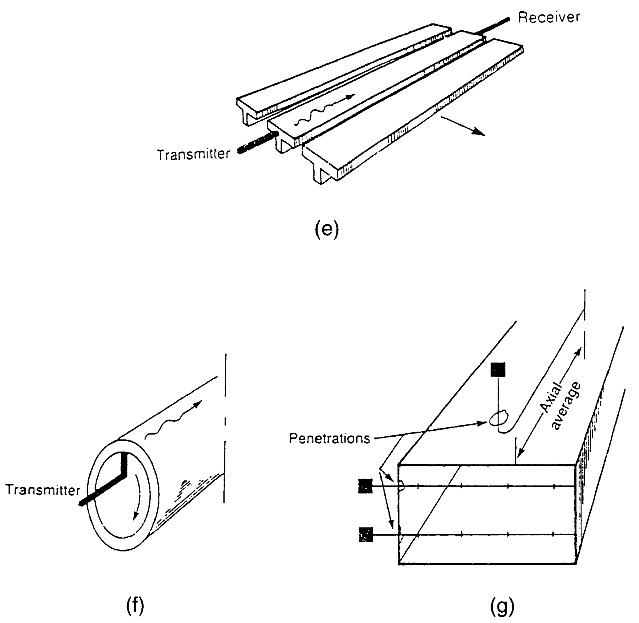
Receiver
Transmitter
(e)
Transmitter
(f)
Penetrations
Axial average
(g)

FIG. 6. Continued.

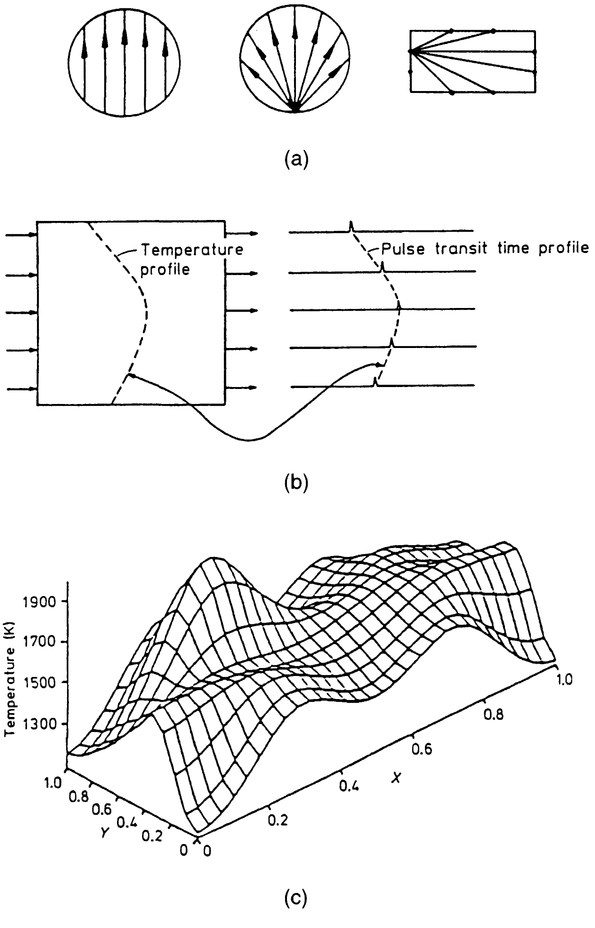
(a)
Temperature profile
Pulse transit time profile
(b)
Temperature (K)
1900
1700
1500
1300
1.0
0.8
0.6
0.4
0.2
0
Y
X
(c)

FIG. 7. (a) Tomographic thermometry proposed for gases by Lynnworth and Carnevale (1966) and for hot steel billets by Lynnworth and Patch (1970). In first and second drawings, lines with arrows represent the central rays of beams that sample the temperature profile in the circular ducts. In practice, each beam actually spreads; see text. Third drawing, ray paths launched from one transducer in one wall of scheme described in text for rectangular duct. (b) Ultrasonic determination of temperature distribution inside large metal billets, proposed by Lynnworth (1970). Measuring the transit time vs elapsed (real) time was also proposed, to monitor the approach to equilibrium. (c) Tomographically reconstructed temperature profiles in the exit plane of a large furnace, after Green (1985). © 1985 AIP; reproduced with permission.

central rays of beams that sample the temperature profile in the circular ducts. In practice, each beam actually spreads. Beam spread, or *diffraction*, depends mainly on the diameter d of the transducer compared with the wavelength λ transmitted into the fluid medium. For $d/\lambda = 10$, the most intense part of the beam's energy is confined to a cone having an included angle $\theta_{3\,\mathrm{dB}} = 0.1$ radian $= 5.7°$. Suppose that along each wall of a rectangular duct, two transducers are installed, such that each transducer communicates with all the other transducers except the one on its own wall; see the last schematic in Fig. 7(a). Then there are 24 independent paths. More recent work along these lines is reported by Kleppe (1989, 1991) and Basarab-Horwath and Dorozhevets (1994). To overcome attenuation in the hot, sooty, turbulent gas, Kleppe used intense waves at audible, not ultrasonic, frequencies, e.g., ~1.5 kHz for paths of 10 to 20 m.

1.4 Pressure

1.4.1 Fluids The speed of sound c generally increases as pressure P increases in liquids, but the effect is too small to be of much practical use for sensing P, especially as the P effect on c would probably be masked by effects of temperature T or composition. In gases near atmospheric pressure, the sign of dc/dP can be plus or minus, depending on the type of gas (Bhatia, 1967, p. 15). The increase in c caused by P up to approximately 300 bar in several inorganic gases is shown in Lynnworth (1989), p. 232, based on work by Carey *et al.* (1969). The attenuation coefficient α and density ρ do not change much with pressure in a liquid. But in a gas the effects of P on α and Z are large enough to provide a basis for using the gas as its own P sensor (Fig. 8).

Intrusive P-sensing probes are available commercially that utilize the change in resonant frequency of a quartz crystal. See Figs. 9(a) and 9(b), which include contributions from researchers at Quartztronics and Schlumberger, respectively [EerNisse *et al.* (1988); Gordon and Wiggins (1990); Besson *et al.* (1993)].

One way of measuring fluid pressure by exploiting pressure coupling across the dry interface between two solids contained in a sealed two-crystal transducer assembly is shown in Fig. 10, after a proposal by Lynnworth (1995). [For pressure-coupling data see Crecraft (1964).] In Fig. 10, the cylindrical object between gaps G1 and G2 is an insulated solid rod such as fused silica. If it is metallized, at least one end must be electrically insulated. The rod is supported by soft O-rings or by other means, in a way that allows electrical lead wires to pass from the transducers to the connectors at the right.

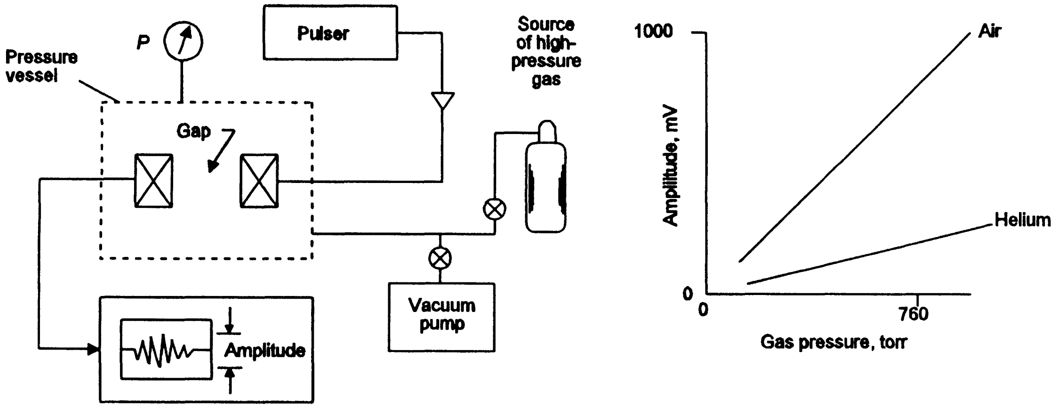

FIG. 8. Amplitude of an ultrasonic pulse transmitted across the gap is a nearly linear function of gas pressure up to ~10 bar for inorganic gases such as air or helium. The idealized graph covers the bottom portion of that range. The slope for air is greater than that for He by approximately the ratio $r = (M_{\mathrm{air}}\,\gamma_{\mathrm{air}}/M_{\mathrm{He}}\gamma_{\mathrm{He}})^{1/2} = 2.5$, where the Ms are molecular weights and γ's are specific heat ratios. Beam spread and attenuation also need to be taken into account, in a more accurate calculation. Virial coefficients are used to quantify departures from ideal-gas behavior.

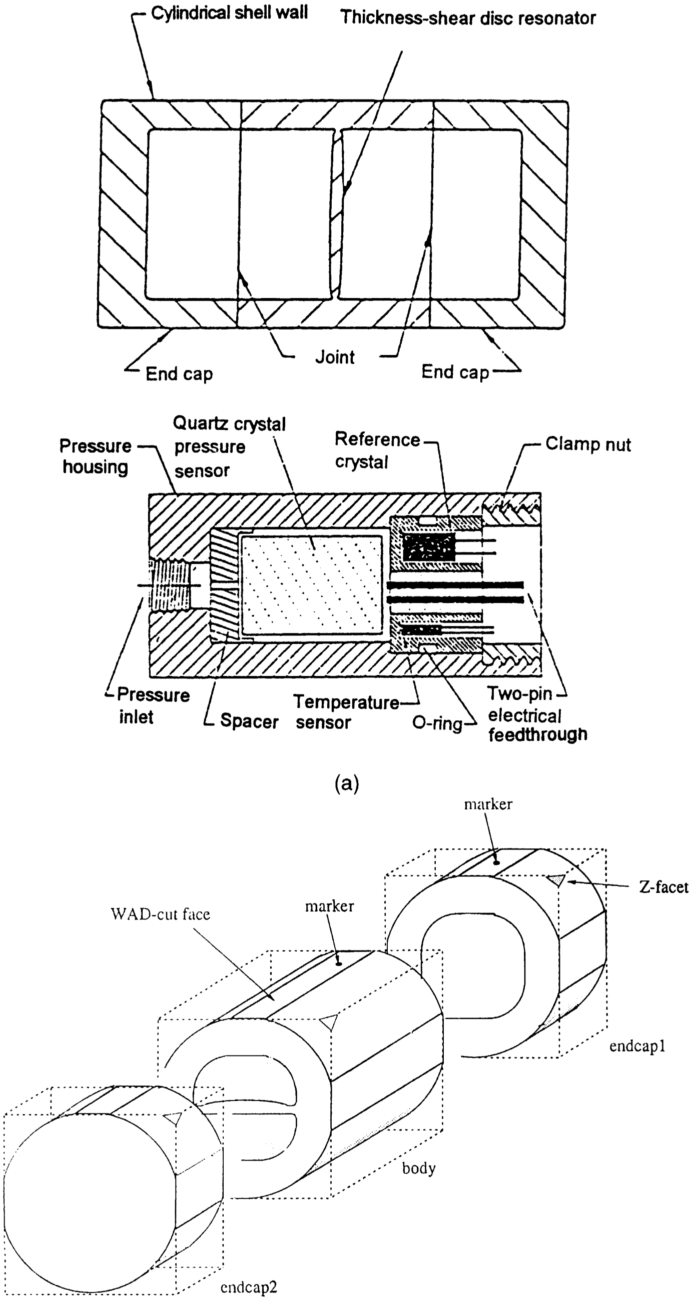
Cylindrical shell wall
Thickness-shear disc resonator
End cap
Joint
End cap
Pressure housing
Quartz crystal pressure sensor
Reference crystal
Clamp nut
Pressure inlet
Spacer
Temperature sensor
O-ring
Two-pin electrical feedthrough
(a)
marker
Z-facet
WAD-cut face
marker
endcap1
body
endcap2
(b)

FIG. 9. Quartz-cystal *P* sensors. (a) Design manufactured by Quartztronics, Inc., after EerNisse *et al.* (1988) and Gordon and Wiggins (1990), courtesy of and © 1990 Quartztronics. (See also Clayton and EerNisse, 1992.) The upper part represents a zoomed enlargement of the quartz-crystal pressure sensor shown below it. (b) Schematic diagram of the three elements of a dual-mode thickness-shear SPA design, after Besson *et al.* (1993), courtesy of and © 1994 Schlumberger, Ltd. This design has been developed by Schlumberger, and is being manufactured by them for downhole pressure measurements in oil fields. It is referred to as CQG™ (Combinable Quartz Gauge) in the oil-field industry.

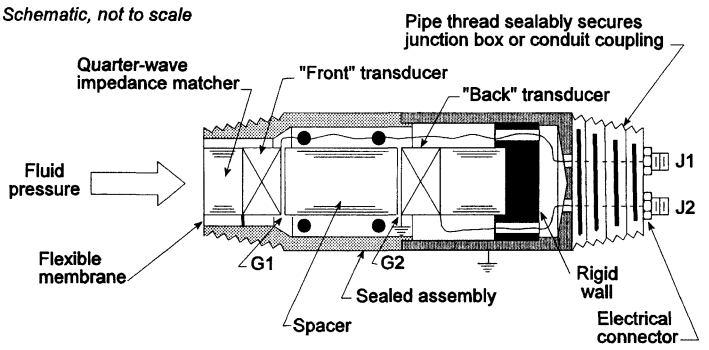

FIG. 10. Pressure coupling within a sealed transducer assembly. The strength of the signal internally coupled between "front" and "back" transducers is a function of fluid pressure bearing against the assembly. Gaps G1, G2, exaggerated in this schematic, close up and couple more strongly when fluid pressure P increases. Accordingly, a signal would be transmitted more efficiently from one transducer to the other as P increases. The front transducer is intended to be part of a trans-fluid interrogation system—for example, part of an ultrasonic flowmeter where transducers are diagonally positioned on opposite sides of a pipe as in Fig. 2 or in Fig. 3d. The two O-rings (●) symbolize one way of concentrically centralizing the spacer. The ungrounded lead wire from the front transducer passes through these O-rings, then through a small hole in the rigid wall, and then to electrical connector J1. The ungrounded lead from the back transducer passes through a second small hole in the rigid wall to J2. After a proposal by Lynnworth (1995).

Compared with a ceramic rod such as fused silica, a metal rod will normally exhibit a much bigger change in sound speed as temperature changes. In this case, the time delay in the metal can be interpreted, with the help of data like those in Fig. 6(a), in terms of the average temperature in the rod. To the extent that the rod is at the same temperature as the fluid, then the fluid *pressure* sensor also acts as a fluid *temperature* sensor. It should be noted that while "pressure coupling" provides a means to sense wide *ranges* in pressure, it has not yet (1995) been demonstrated to do so with high accuracy.

1.4.2 Solids under Stress Provided textural effects can be eliminated, the small change in sound speed that is associated with elastic nonlinearity or higher-order elastic constants can be used to measure residual stress. Measurements are made orthogonally or along crystal axes, preferably from the no-load to full-load condition to eliminate texture or anisotropy effects that could mask the stress effect that is sought. Recent work by Shaikh (1995) suggests that a surface-skimming longitudinal wave may be particularly sensitive and convenient for measuring *relative* stress.

2. OTHER MEASURANDS

2.1 Density of Gases; Density of Liquids

2.1.1 Gas Density For a one-component ideal gas at pressure $P < 10$ bar, the sound speed is $c = (\gamma RT/M_W)^{1/2}$ where γ = specific heat ratio C_P/C_V, $C_P = C_V + R$, R = universal gas constant, T = absolute temperature, and M_W = molecular weight. For a mixture of two ideal gases, c is calculated by weighting the specific heats and M_W's according to the mole fractions X_1 of the first gas and $X_2 = 1 - X_1$ of the second gas; see Polturak *et al.* (1986) or Valdes and Cadet (1991). This means replacing γ by the weighted average $\bar{\gamma}$ and similarly replacing M_W by $\bar{M}_W$. We now have for the binary mixture $c = (\bar{\gamma}RT\bar{M}_W)^{1/2}$ where

$$\bar{\gamma} = \frac{X_1 C_{P1} + (1 - X_1) C_{P2}}{X_1 C_{V1} + (1 - X_1) C_{V2}} = 1 + \left(\frac{X_1}{\gamma_1 - 1} + \frac{1 - X_1}{\gamma_2 - 1} \right)^{-1} \tag{8}$$

and

$$\bar{M}_W = X_1 M_{W1} + (1 - X_1) M_{W2}. \tag{9}$$

Generalizing to a multicomponent gas yields

$$\bar{M}_W = \sum_{i=1}^{N} X_i M_{Wi}, \tag{10}$$

where

$$\sum_{i=1}^{N} X_i = 1. \tag{11}$$

If the specific heats and M_W's are known for the individual gases, measurement of c and T yields the average molecular weight $\bar{M}_W$ of the binary gas mixture. Additionally, if the pressure P is known, the gas mixture's density ρ_{mix} can be calculated. In units of g/L,

$$\rho_{\text{mix}} = \frac{\bar{M}_W}{22.4} \frac{P}{P_0} \frac{T_0}{T}, \tag{12}$$

where P_0 and T_0 are standard pressure and temperature, 760 mm and 273 K. If the gases are not ideal, one or more virial coefficients (or supercompressibility factors) must be introduced into the equations relating sound speed to molecular weight and density.

2.1.2 Liquid Density For special cases such as some pure liquids, liquid density ρ_{liq} can be determined from sound speed c_{liq} in the liquid (the "medium as its own sensor" method). For water, c_{liq} is not uniquely associated with T unless T is only above or only below 74 °C. Salinity, pressure, and contaminants also influence c_{liq}.

Two other approaches to determining ρ_{liq} are

1. use of a foreign sensor such as a torsional waveguide, provided the cross section is noncircular;
2. use of a tube, or duct wall, as the sensor, interrogated by flexural or bending waves.

2.2 Concentration Analysis of Binary Gas Mixtures

By using the equation for sound speed in an ideal gas, $c = (\gamma RT/M_W)^{1/2}$, one can readily calculate c if one knows γ, R, T, and M_W. Conversely, if c is measured, one can determine T (Mayer, 1873), γ (Sherratt and Griffiths, 1934), or M_W, provided the other terms are known. [In flare gases (Smalling *et al.*, 1986) and in some other multicomponent gas mixtures $\bar{M}_W$ and $\bar{\gamma}$ are not independent of one another. Therefore, c, T, and the theoretical or empirical relationship between $\bar{M}_W$ and $\bar{\gamma}$ suffice to determine the mixture's average $\bar{M}_W$. Another multicomponent case in which c and T measurements can yield $\bar{M}_W$ is the *pseudobinary* gas mixture. This is a mixture of two gases, one of which is itself a mixture, such as air, that has a relatively stable composition and presumably known M_W and γ.]

Although the concept of determining M_W from c is contained in the works of Herzfeld and Litovitz (1959) and Bhatia (1967), and demonstrated by data in their books or in other publications, the idea did not gain much attention until the 1980s. Since the 1980s, there has been a growing interest in determining M_W for binary mixtures of MOCVD (metal-organic chemical-vapor deposited) gases and some other binary mixtures; see Valdes and Cadet (1991) and Hallewell and Lynnworth (1994). This 1994 reference deals with the special case of two gases A and B of widely different molecular weights, say $M_A \gg M_B$, and where one of the gases, say A, is present at very low concentrations, <1%. In this special case X_A, the concentration of gas A can be computed using the approximation

$$X_A \approx \left[2 \frac{M_B}{M_A} \right] \left[1 - \frac{c_{\text{mixture}}}{c_B} \right]. \tag{13}$$

As a numerical example, consider a dilute (<1%) solution of TMIn (trimethylindium) in H_2. Approximating the molecular weights of these gases by 160 and 2, respectively, and taking c in H_2 to be 1284 m/s at 20 °C, Eq. (13) yields

$$X_{TMI} \approx \frac{2 \times 2}{160} \left(1 - \frac{c_{\text{mixture}}}{1284} \right) = \frac{1}{40} \left[1 - \frac{c_{\text{mixture}}}{1284} \right]. \tag{14}$$

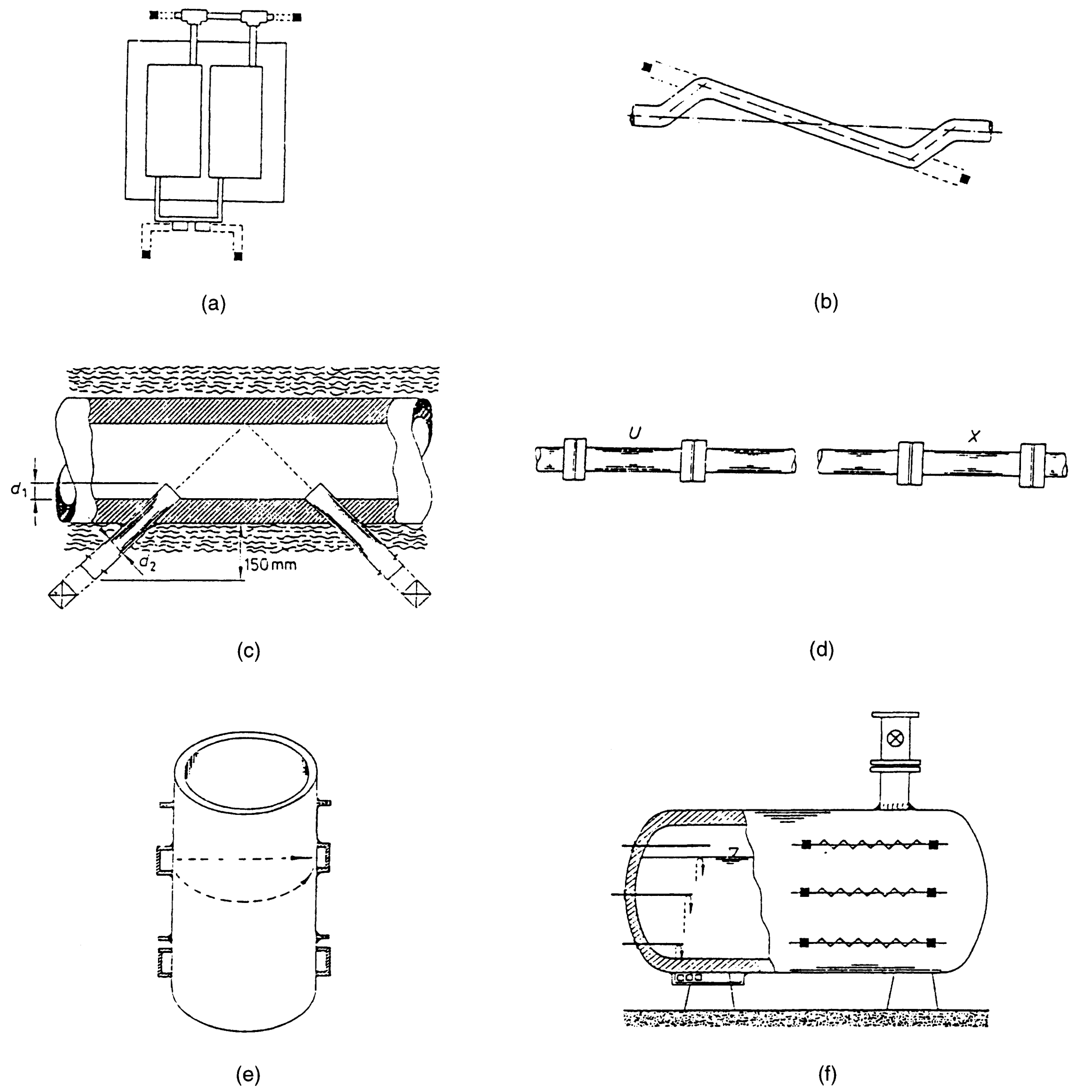

(a)

(b)

(c)

(d)

(e)

(f)

FIG. 11. Examples of adapting a process boundary at the time of manufacture to accommodate ultrasonic sensors that are to be installed later. (a) Clamps for clamp-on vee-path flowmeters; plugged tees for hybrid axial-path flow interrogation. (b) Tube bent smoothly, for clamp-on axial-path interrogation. (c) Buffer rods welded into a high-temperature insulated pipe. (d) In a long straight pipe, two similar or identical dummy spoolpieces U and X have been installed as spacers. Spoolpiece U is to be replaced later by a precision-bore ultrasonic flowmeter like that in Fig. 1(d) having one or more paths. Spoolpiece X is to be replaced by a nonultrasonic flowmeter [turbine, orifice plate, Venturi, magmeter (magnetic field flowmeter), or Coriolis sensor, for example] to establish the installed accuracy of U relative to X. Some designers would put U and X in a bypass line so that replacement can occur without shutting down the main process line. Spoolpiece X can also be thought of as representing a general-purpose sensor port to accommodate whatever sensor is "best" as a function of new requirements and ongoing improvements in sensor technology. (e) Clamps and threaded studs welded up along wall of a tall storage tank. (f) Clamps, reference reflectors, and isolation-valved port welded on horizontal-axis thick-walled storage tank. © 1992 Panametrics.

Note that γ and T do not appear in Eq. (13). Gas A is hardly present and so its γ is not important here. But T must be known, in order to enter the proper value for c_B in Eq. (13).

In practical applications, considerable care may be required to bring the gas to a stable temperature and, if pulse techniques are used, to measure over a path long enough to achieve adequate precision in c for a given instrument-limited resolution of transit time. Resonant cavities or interferometers provide a way to achieve high precision in a short path.

2.3 Designing Pipes and Tanks to Accommodate Sensors, Particularly Ultrasonic Sensors

Just as one essential design parameter of a welded structure or weldment is its inspectability, so too should be the *sensorability* of a process. Sensorability, the ability to be sensed, requires that the designer select the sensor generically or specifically, as part of the design process, or at least provide a sensor nozzle, port, or mounting bracket. If clamp-on sensors are to be used, there are cases where it will be more economical to weld the clamps or other sensor-mounting fixtures onto the pipe or tank when that pipe or tank is being fabricated (Fig. 11), rather than as an afterthought. At the design stage, rather than after the pipe or tank is in operation, if one can anticipate or select the sensors that will be used later, one can enhance their performance, and simplify their installation, e.g., avoid subsequent hot-tapping or plant shutdown merely to install a sensor. Sensor mounts and reference reflectors, for example, could be welded safely and easily prior to filling a pipe or tank with an explosive hydrocarbon. Once the container is filled, welding is forbidden in some areas (e.g., above or within ~1 m of the liquid level) and certainly is more complicated, even in areas where welding is allowed, than if the necessary provisions for sensors could have been created earlier.

Works Cited

Baggeroer, A., Munk, W. (1992), *Phys. Today* **45** (9), 22–30. See also Holing, D. (1994), *Amicus J.* **16** (3), 19–23.

Basarab-Horwath, I., Dorozhevets, M. M. (1994), "Measurement of the Temperature Distribution in Fluids Using Ultrasonic Tomography," in: M. Levy, S. C. Schneider, B. R. McAvoy (Eds.), *1994 IEEE Ultrasonics Proceedings*, New York: IEEE, pp. 1891–1894.

Baumoel, J. (1994), U.S. Patent No. 5,343,737.

Bell, J. F. W. (1957), *Philos. Mag.* **2,** 1113–1120.

Besson, R. J., Boy, J. J., Glotin, B., Jinzaki, Y., Sinha, B., Valdois, M. (1993). *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **40,** 584–591.

Bhatia, A. B. (1967), *Ultrasonic Absorption*, London: Oxford Univ. Press.

Brillouin, L. (1960), *Wave Propagation and Group Velocity*, Pure and Applied Physics, Vol. 8, New York: Academic.

Carey, C. A., Carnevale, E. H., Uva, S., Marshall, T. (1969), *Experimental Determination of Gas Properties at High Temperatures and/or Pressures*, Arnold Engineering Development Center Report No. AEDC-TR-69-78, Springfield, VA: National Technical Information Service.

Clayton, L. D., EerNisse, E. P. (1992), *Sensors* **9** (11), 37–42.

Collard, S. M. (1990), *Scripta Metall. Mater.* **24,** 623–627. See also Collard, S. M. (1991), Ph.D. Diss., Rice University.

Crecraft, D. I. (1964), *J. Sound Vib.* **1,** 381–387.

Dabirikhah, H., Turner, C. W. (1994), *Electron. Lett.* **30** (18), 1549–1550.

Dogan, A., Yoshikawa, S., Uchino, K., Newnham, R. E. (1994), in: M. Levy, S. C. Schneider, B. R. McAvoy (Eds.), *1994 IEEE Ultrasonics Proceedings*, New York: IEEE, pp. 935–939.

EerNisse, E. P., Ward, R. W., Wiggins, R. B. (1988), *IEEE Trans. Utrason. Ferroelec. Freq. Control* **35,** 323–330.

Fendrock, C., Varela, D. W. (1995), *Sensors* **12** (1), 23–37.

Fleury, G. (1994), *Reconstruction de Profils de Vitesse à Partir de Données Lacunaires et Optimisation d'Instrument—Application à la Débitmétrie Ultrasonore*, Orsay, France: Université de Paris-Sud, École Supérieure d'Électricité.

Folkestad, T., Mylvaganam, K. S. (1993), *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **40,** 193–215.

Gordon, J., Wiggins, R. B. (1990), *Sensors* **7** (8), 30–34.

Green, S. F. (1985), *J. Acoust. Soc. Am.* **77,** 759–763.

Gualtieri, J. G., Kosinski, J. A., Ballato, A. (1994), *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **41,** 53–59.

Hallewell, G. D., Lynnworth, L. C. (1994), in: M. Levy, S. C. Schneider, B. R. McAvoy (Eds.), *1994 IEEE Ultrasonics Proceedings*, New York: IEEE, pp. 1311–1316.

Hamidullin, V. K. (1993), in: M. Levy, S. C. Schneider, B. R. McAvoy (Eds.), *1994 IEEE Ultrasonics Proceedings*, New York: IEEE, pp. 395–399.

Hayward, G. (Ed.) (1994), special issue on Correlation Techniques and Applications in Ultrasonics, *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **41,** 577–663.

Hertz, T. G., Dymling, S. O., Lindström, K., Persson, H. W. (1991), *Rev. Sci. Instrum.* **62** (2), 457–462.

Herzfeld, K. F., Litovitz, T. A. (1959), *Absorption and Dispersion of Ultrasonic Waves,* New York: Academic.

Holm, M. (1995), *Simulation of Flowmeter Installation Effects,* LUTHMDN/TMVK—7019-SE, Lund, Sweden: Lund Institute of Technology.

Hutchins, D. A. (1994), in: M. Levy, S. C. Schneider, B. R. McAvoy (Eds.), *1994 IEEE Ultrasonics Proceedings,* New York: IEEE, Paper II-3.

Jacobson, S. A., Lynnworth, L. C., Korba, J. M. (1988), U.S. Patent No. 4,787,252.

Kim, J. O., Bau, H. H., Liu, Y., Lynnworth, L. C., Lynnworth, S. A., Hall, K. A., Jacobson, S. A., Korba, J. M., Murphy, R. J., Strauch, M. A., King, K. G. (1993), *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **40** (5), 563–576.

Kleppe, J. A. (1989), *Engineering Applications of Acoustics,* Artech. See also Kleppe, J. A. (1991), *I & CS (Instruments & Control Systems)* **64** (6), 35–40; Yori, L. G. (1992), *I & CS (Instruments & Control Systems)* **65** (11), 61–62.

Lalande, F., Chaudhry, Z., Rogers, C. A. (1995), *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **42,** 21–27.

Liu, Y., Lynnworth, L. C. (1993), in: *1993 IEEE Ultrasonics Symposium Proceedings,* pp. 385–390; see also Liu, Y., Lynnworth, L. C. (1995), U.S. Patent No. 5,456,114.

Liu, Yi, Lynnworth, L. C., Nguyen, T. H., Xiao, D., Walters, J. (1996), in: *Proc. International Pipeline Conference,* ASME.

Lynnworth, L. C. (1970), *Nav. Res. Rev.* **23** (7), 1–22, 32.

Lynnworth, L. C. (1989), *Ultrasonic Measurements for Process Control,* New York: Academic.

Lynnworth, L. C. (1992), in: P. H. Sydenham, R. Thorn (Eds.), *Handbook of Measurement Science,* Vol. 3, New York: Wiley, Chap. 38, 1655–1689.

Lynnworth, L. C. (1994a), in: H. H. Bau, N. F. deRooij, B. Kloeck (Eds.), *Mechanical Sensors,* Vol. 7, Weinheim: VCH, Chap. 8, pp. 285–329.

Lynnworth, L. C. (1994b), U.S. Patent No. 5,275,060.

Lynnworth, L. C. (1994c), in: M. Levy, S. C. Schneider, B. R. McAvoy (Eds.), *1994 IEEE Ultrasonics Proceedings,* New York: IEEE, pp. 1317–1321.

Lynnworth, L. C. (1995), U.S. Patent No. 5,437,194.

Lynnworth, L. C., Carnevale, E. H. (1966), *NASA CR-54979.*

Lynnworth, L. C., Patch, D. R. (1970), *MTRSA (Materials Research and Standards)* **10** (8), Cover, 6–11, 40.

Lynnworth, L. C., Hallewell, G. D., Bragg, M. I. (1994), in: *Proceedings of the "FLOMEKO '94" Conference on Flow Measurement in the Mid 90s,* Session 7, Paper #3, pp. 1–15.

Mágori, V. (1994), in: M. Levy, S. C. Schneider, B. R. McAvoy (Eds.), *1994 IEEE Ultrasonics Proceedings,* New York: IEEE, pp. 471–481.

Matson, J., Davis, R. (1994), *I&CS (Instruments & Control Systems)* **67** (2), 67–69.

Mayer, A. M. (1873), *Philos. Mag.* Ser. 4, **45,** 18–22.

Nygaard, O. G. H., Mylvaganam, K. S. (1993), *Technisches Messen* **60** (1), 4–14.

Papadakis, E. P. (1997) (Ed.), *Ultrasonic Instruments and Devices: Reference for Modern Instrumentation, Techniques and Technology,* Academic, to be published.

Pollard, H. F. (1977), *Sound Waves in Solids,* Pion, pp. 139–140.

Polturak, E., Garrett, S. L., Lipson, S. G. (1986), *Rev. Sci. Instrum.* **57,** 2837–2841.

Royer, D., Dieulesaint, E., Legras, O. (1992), *J. Phys. III* **2,** 145–168.

Shaikh, N. (1995), private communication.

Sheen, S. H. (1994), *Machine Design* **66** (23), 60 and 62.

Sheen, S. H., Chien, H.-T., Raptis, A. C. (1995), private communication.

Sherratt, G. G., Griffiths, E. (1934), *Proc. R. Soc. London, Ser. A* **147,** 292–308.

Smalling, J. W., Braswell, L. D., Lynnworth, L. C., Wallace, D. R. (1984), in: *Proceedings of the 39th Annual Symposium on Instrumentation for the Process Industries,* Research Triangle Park, NC: Instrument Society of America, pp. 27–38.

Smalling, J. W., Braswell, L. D., Lynnworth, L. C. (1986), U.S. Patent No. 4,596,133.

Smith, R. H., Jr. (1994), *The Direct Measurement of Circulation in Free Surface Vortices,* MS Thesis, Worcester Polytechnic Institute, Worcester, Massachusetts; see also Smith, R. H., Jr., Durgin, W. W., Johari, H. (1995), *Proceedings of AIAA Conference,* AIAA Paper 95-0104, American Institute of Aeronautics and Astronautics, Inc., 9 pp.

Sunthankar, Y., (1973), *IEEE Trans. Son. Ultrason.,* **SU-20,** 274–278.

Tomioka, G. (1968), U.S. Patent No. 3,394,589.

Valdes, J. L., Cadet, G. (1991), *Anal. Chem.* **63,** 366–369.

Wenzel, S. W., Costello, B. J., White, R. M. (1994), *Sensors* **11** (12), 47–49.

Further Reading

Babikov, O. I. (1960), *Ultrasonics and Its Industrial Applications,* Consultants Bureau, New York: Plenum.

Ballato, A. (1983), "Piezoelectric Resonators," in: B. Parzen (Ed.), *Design of Crystal and Other Harmonic Oscillators,* New York: Wiley, Chap. 3, pp. 66–122 and pp. 432–436.

Brekhovsikh, L. M. (1980), *Waves in Layered Media,* Applied Mathematics and Mechanics Vol. 16, New York: Academic.

Duncombe, E. (1984), *J. Phys. E.: Sci. Instrum.* **17,** 17–18.

Frederick, J. R. (1965), *Ultrasonic Engineering,* New York: Wiley.

Gerber, E. A., Ballato, A. (Eds.) (1985), in: *Precision Frequency Control,* Vol. 1, *Acoustic Resonators and Filters,* 448 pp.; Vol. 2, *Oscillators and Standards,* 480 pp.; New York: Academic.

Hamidullin, V. K. (1989), *Ultrasonic Control Instrumentation Systems* (in Russian), Leningrad (St. Petersburg) University, 248 pp.

Hamidullin, V. K. (1992), *Ultrasonic Transducers for Flow and Liquid Level Measurements* (in Russian), Leningrad (St. Petersburg) University, 81 pp.

Hayward, G. (Ed.) (1994), special issue on Correlation Techniques and Applications in Ultrasonics, *IEEE Trans. Ultrason. Ferroelec. Freq. Control.* **41,** 577-663.

Kažys, R.-J. (1986), *Ultrasonic Information Measuring Systems,* Vilnius Mokslas (in Russian).

Krautkrämer, J., Krautkrämer, H. (1983), *Ultrasonic Testing of Materials,* 3rd ed., New York: Springer.

Leighton, T. G. (1994), *The Acoustic Bubble,* New York: Academic, 613 pp.

Malghan, S. G. (Ed.) (1993), *Electroacoustics for Characterization of Particulates and Suspensions,* National Institute of Standards and Technology Special Publication No. 856, Washington, DC: U.S. GPO.

Mason, W. P. (1950), *Piezoelectric Crystals and Their Application to Ultrasonics,* New York: Van Nostrand.

Meyers, R. A. (Ed.) (1992), *Encyclopedia of Physical Science and Technology,* 18-volume set, New York: Academic, 13,788 pp.

Papadakis, E. P. (1997) (Ed.), *Ultrasonic Instruments and Devices: Reference for Modern Instrumentation, Techniques and Technology,* Academic, to be published.

Pierce, A. D. (1981), *Acoustics,* New York: McGraw-Hill, 642 pp.

Pierce, A. D., Thurston, R. N. (Eds.) (1993), *Physical Acoustics,* Vol. 22, New York: Academic, 384 pp.

Rosa, D. (Ed.), (1994), "1995 Buyer's Guide," *Sensors,* Vol. 11, No. 9, Part 2, Peterborough, NH: Helmers.

Schlichting, H. (1955), *Boundary Layer Theory,* London: Pergamon.

Viktorov, I. A. (1967), *Rayleigh and Lamb Waves,* New York: Plenum.

SENSORS, INFRARED

Joseph S. Accetta, *Industrial Technology Institute, Ann Arbor, Michigan, U.S.A.*

	Introduction	543
1.	**Physical and Technical Principles**	544
1.1	Principles of Operation	546
1.1.1	Infrared Detectors and Arrays	546
1.1.2	Collection Optics	547
1.1.3	Filters	548
1.1.4	Cryogenics	549
1.1.5	Electronics	549
1.1.6	Scanners	549
1.1.7	Displays and Signal Processing	549
2.	**Design, Function, and Operation**	551
2.1	Sensor Figures of Merit	551
2.2	Scanning Systems	551
2.3	Staring Systems	551
2.4	Infrared Phenomenology	552
2.4.1	Object Radiation	552
2.4.2	Backgrounds	553
2.4.3	Atmospherics	553
3.	**Manufacturing and Economics**	553
4.	**Major Applications**	554
4.1	Infrared Radiation Pyrometry	555
4.2	Search and Tracking Systems	556
4.3	Imaging Systems	556
4.3.1	Forward-Looking Infrared (FLIR) Systems	556
4.3.2	Medical and Industrial Thermal Imagers	556
4.3.3	Astronomical Imagers	557
4.4	Remote Sensing and Environmental Applications	558
4.4.1	Airborne and Spaceborne Spectral Systems	558
4.4.2	Infrared Remote Sensing of the Atmosphere	559
4.5	Radiometers	560
4.6	Infrared Laser Radar	560
4.7	LIDAR and DIAL	560
5.	**Further Research**	561
	Glossary	561
	Works Cited	562
	Further Reading	562

INTRODUCTION

The infrared (IR) region of the electromagnetic spectrum lies just above the visible region, as measured in wavelengths, and extends out to submillimeter wavelengths. Human vision is adapted to the spectral output of the sun and does not respond to infrared radiation. The sensation of heat is a familiar biological sensory response to infrared radiation common to most living species and is particularly well developed in some. An infrared sensor is defined as a device to transform infrared radiation in the spectral region of 0.75 to 1000 μm into interpretable information. As a result of specialization, modern infrared sensor technology is complex and sophisticated.

As described in a brief account of the history of infrared sensors given by Hudson (1969), IR sensing began with the discovery by Herschel in 1800 of "invisible rays." While measuring the heating effects of the refracted spectral components of sunlight with a prism, Herschel discovered that a portion of the spectrum beyond red that was invisible nevertheless elicited a response from the thermometer. This very first infrared detector, the thermometer, was followed by the thermocouple, the bolometer, and a number of other novel detection schemes including infrared-sensitized photographic film. In 1917 Case employed the first infrared detector based upon the photoconductive effect (Dereniak and Crowe, 1984). Although early infrared sensors were little more than primitive detectors, the characteristic exploitation of a number of related and enabling technologies such as the addition of optical elements and electronic amplifiers followed rap-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

idly. As is frequently but sadly the case, much of the success of modern infrared sensors can be attributed to the massive investment of military research since WW II directed at night warfare and surveillance.

One of the most familiar examples of an infrared sensor is the thermal imager that converts radiation emitted from an object or scene into a visually interpretable image. The variations in gray scale or color frequently represent relative temperature differences. A typical example of the imagery from one of these devices is shown in Fig. 1. The relative contrast of the author's face is displayed as function of temperature. In this case the lighter regions correspond to higher temperature, although it could be presented in reverse as well. On occasion images such as these are coded in false color, each color representing a range of temperature. With certain assumptions about the material properties of the object, a temperature range can be assigned to the various regions. Versions of the thermal imager are widely used in applications ranging from astronomy to medicine. The greatest scientific and technological impact of the development of infrared sensing devices is in infrared astronomy, environmental remote sensing, and industrial process control. In addition infrared sensors have clearly played a revolutionary role in modern warfare. Complexity and usage range from the exotic, including satellite-borne astronomical, environmental, and military reconnaissance systems, to numerous commercial and industrial remote temperature-sensing applications.

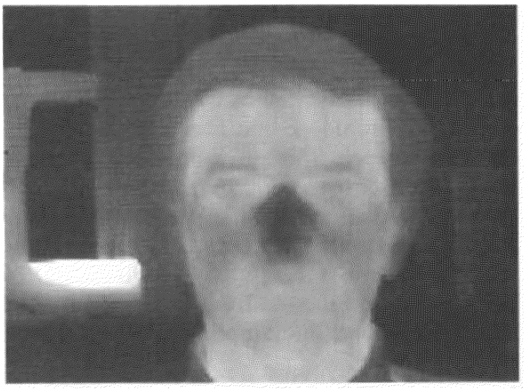

FIG. 1. Infrared image of the author taken with a Cincinnati Electronics IRC 160 infrared imager. Darker regions correspond to lower relative temperatures.

A generally accepted taxonomy of infrared sensors has yet to be established. However, Hudson (1969) made one of the earliest attempts at application-dependent classification (see Sec. 3). Characteristic components and technology afford yet another means of categorization. Distinctions can be made on the basis of imaging versus point detection, the means by which contiguous imagery is generated, and active versus passive systems. In the active case the employment of illuminators is the distinguishing feature, while in the passive case natural illumination or emission is relied on. A multitude of highly specific sensor configurations have been developed to meet the demand for specialized usage. To some extent categorization depends upon the accepted jargon and practices of specific applications.

In its most elementary form, an infrared sensor is a device that responds in a measurable way to incident infrared radiation. Although it is common practice to regard thermometers, thermocouples, and other infrared-sensitive devices as infrared sensors, they are more properly categorized as infrared detectors. In modern usage an infrared sensor is comprised of a thermal or photon solid-state detector element combined with other application-dependent components, as described below. Although sensor configurations vary widely, the basic building blocks are shown in Fig. 2 with additions or deletions as necessary to meet certain requirements.

1. PHYSICAL AND TECHNICAL PRINCIPLES

All objects at finite temperature and in thermal equilibrium radiate energy. The radiation per unit wavelength is given by the Planck radiation law (Reif, 1965):

$$M_e(\lambda,T) = \epsilon(\lambda)(2\pi hc^2/\lambda^5) \times \exp[hc/kT - 1]^{-1}\ (\mathrm{W/m^3})$$

where $\epsilon(\lambda)$ = spectral emissivity, c = speed of light in vacuum = 2.997925 × 10 m/s, h = Planck constant = $6.6260755 \times 10^{-34}$ J s, k = Boltzmann constant = 1.38054

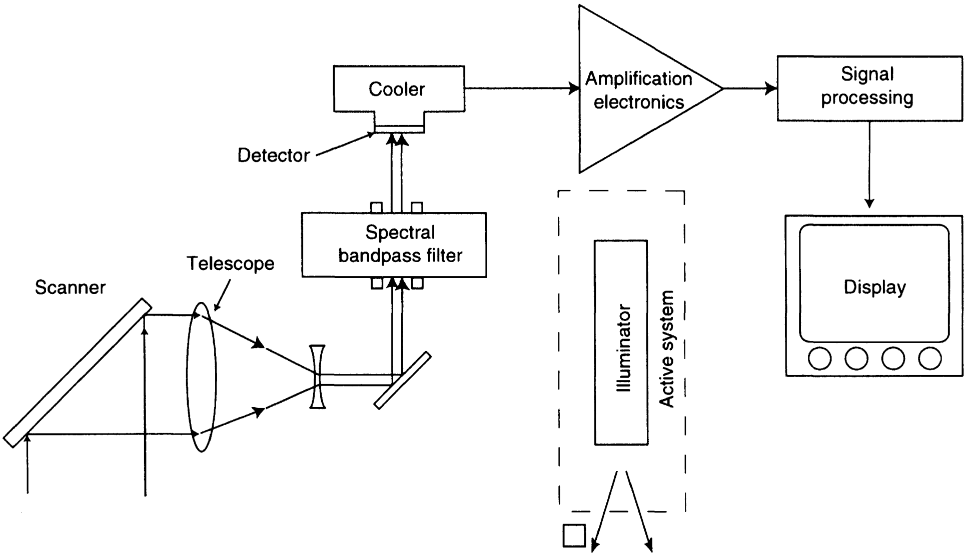
Cooler
Detector
Amplification electronics
Signal processing
Telescope
Scanner
Spectral bandpass filter
Illuminator
Active system
Display

FIG. 2. Generic components of a modern, complex infrared sensor system. The laser is used as a source of monochromatic illumination in certain environmental and military radar applications.

$\times\ 10^{-23}$ J/K, T = absolute temperature in kelvins, and λ = radiation wavelength in meters.

This expression is shown graphically in Fig. 3. We note that for a given temperature, the emitted radiation has a maximum at a specific wavelength. As the temperature increases, the wavelength at maximum emission shifts progressively to shorter wavelengths, and the relative emission rises. In accordance with Wien's law, the wavelength at maximum emission varies inversely with temperature. For objects near room temperature, for example, the maximum occurs in

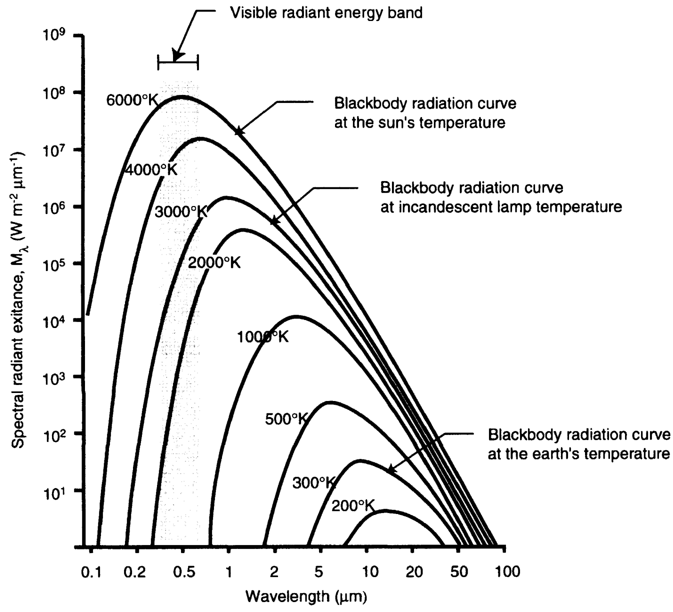

FIG. 3. Spectral distribution of energy emitted per unit wavelength radiated from blackbodies of various temperatures (Lillesand and Kiefer, 1994). As the temperature rises, the emission both increases and shifts to shorter wavelengths.

the long-wave IR region near 10 μm. In general the position of this maximum has a strong influence on the selection of the sensor wavelength band of operation.

Spectral emissivity $\epsilon(\lambda)$ is a property of macroscopic matter that is a measure of the efficiency of emission of absorption of radiation relative to a perfect blackbody, for which $\epsilon(\lambda) = 1.0$. For a "graybody," that is an object whose emissivity is independent of wavelength, only the magnitude (but not the relative spectral distribution) of the radiation is altered. For nonblackbody radiation where the emissivity is a function of wavelength, the distribution of the radiation may be radically altered from the Planck distribution. In some substances such as gases, vibrational-rotational molecular transitions give rise to characteristic emission and absorption lines in the infrared. These gases can thus be identified by analysis of their corresponding spectra. In general, because of energy-level broadening in the solid state, narrow-line emission or reflectance spectra are not usually observed in solid materials. Relatively broad spectral features are much more common. High-temperature line broadening produces the same effect in the radiation spectra of gases at high temperature. It is the analysis of such spectral features of materials and even living organisms that allows remote discrimination and identification of various materials constituting the observed scene. This technique forms the basis of the science of multispectral remote sensing (see Sec. 4.4). Integrating the Planck expression over all wavelengths yields the Stefan–Boltzmann equation, which, under certain assumptions about the object's emissivity, enables an inference of object temperature. Remote infrared temperature-sensing devices use this principle to measure object temperature indirectly. The accuracy of this technique depends largely upon on how well the emissivity is known.

1.1 Principles of Operation

A modern infrared sensor comprises an infrared-sensitive element, or detector, to convert the collected radiation to electrical signals and at least one or more of the following:

1. optical elements to collect and focus radiation;
2. a scanning device to image the scene of interest;
3. a spectral filter to eliminate extraneous radiation; and
4. some form of display device or subsequent signal processing to convert the electrical signal into useful information.

In some cases an illuminator may be used to provide a controlled source of radiation. Additional components could include cryogenics for detector cooling, internal temperature references, mechanical gimbals, servo control for pointing and tracking, and a host of other components in specialized applications. The physical and technical principles of the typical major components of an infrared sensor are discussed elsewhere in this work. In many instances strong physical analogies exist between sensors operating in the optical region of the spectrum and the infrared region, and some devices are capable of operating in both.

Complex infrared sensors contain a composite of diverse technologies including solid-state detectors, optical design and fabrication techniques, optical materials, mechanical scanning devices, electronics, and display and signal-processing electronics. These are briefly discussed below. For an extended discussion of the components of infrared systems reviewed below see Accetta and Shumaker (1993), Vol. 3.

1.1.1 Infrared Detectors and Arrays

The responsive element in all modern infrared sensors is the detector that converts the incoming radiation into electrical signals by one of several possible physical mechanisms. These signals are subsequently amplified by accompanying electronics. The physical mechanisms that have been employed in detectors include bolometric, photoconductive, photoelectromagnetic, photovoltaic, pyroelectric, thermopneumatic, and thermovoltaic processes. The majority of modern infrared sensors depend upon photon detectors using photoconductive and photovoltaic processes and thermal detectors using bolometric and pyroelectric processes. For an extended discussion of detectors see RADIATION DETECTORS, INFRARED.

In general, photon detectors employ some form of cooling to approach their ultimate limits of sensitivity. In contrast, thermal detectors are generally not cooled and are used in applications requiring modest levels of sensitivity. (A prominent exception is the infrared bolometer employed in astronomy, which is sometimes cooled to liquid helium temperatures to achieve very high sensitivity and wide spectral response). Figure 4 shows several specific detector types and the spectral ranges commonly employed. The parameter commonly used to describe the detector figure of merit is referred to as D* ("D star") and is widely employed in sensor performance predictions (Dereniak and Crowe, 1984).

Detectors are often fabricated in linear, rectangular, or square arrays as an alternative to creating images by mechanical scanning. These focal-plane arrays (FPAs) spatially sample the optical image in the focal plane of the sensor to create nearly contiguous imagery. Each individual detector in the FPA senses the radiation falling within its active boundaries. The detector element and its corresponding contribution to the image is referred to as a "pixel" (short for picture element). FPAs are employed in some imaging sensors and possess certain technical advantages over systems that mechanically scan the field of view over a single detector. For example, rapid scanning over the sensor scene of interest may have mechanical limitations that may be circumvented with the use of FPAs. The choices here depend on the particular application.

Modern FPAs may consist of thousands of individual detector elements. An associated amplifier and output wires from each detector are thus impractical. A means for amplifying and multiplexing (handling many detector outputs over a few wires) the signals was required. To solve this problem the FPA is now integrated with multiplexing electronics on a single chip and commonly referred to as a charge-coupled device (CCD) or infrared CCD (IRCCD). The term CCD really comes from the specific means by which the outputs are multiplexed but has become generic in usage.

The majority of infrared detector elements in use are uncooled single-element types. Detectors suitable for this purpose include pyroelectrics, which create a temperature-dependent change in effective capacitance, and bolometers, whose resistance is temperature dependent. There are many industrial applications where sufficient signal to overcome inherent detector noise is readily available, especially those employing active illuminators or intense thermal sources. Thermal detectors are the standard of usage in many industrial applications. Because of the attendant advantages in simplicity and cost, a great deal of recent work in detector development has concentrated on high-sensitivity uncooled operation (Rogalski, 1994). See also RADIATION DETECTORS, INFRARED for an extended discussion of modern infrared detector technology.

1.1.2 Collection Optics The optical design principles used in the design of IR sensors are strongly analogous to those used in

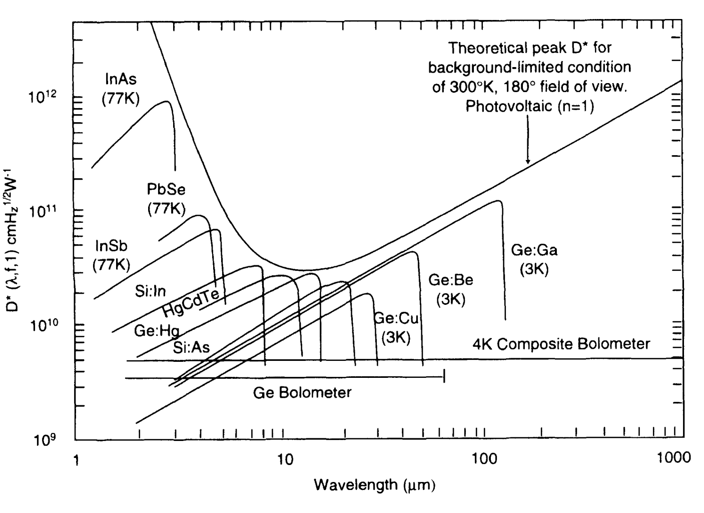

FIG. 4. The spectral response for various types of detectors expressed in terms of detectivity D* (Dereniak and Crowe, 1984). D* is a commonly used parameter for the figure of merit that reflects the measured signal-to-noise ratio per watt of incident flux under specific conditions including the incident wavelength, spectral bandwidth, and temporal frequency. The measurement is further normalized by the detector area and electrical bandwidth of the measurement.

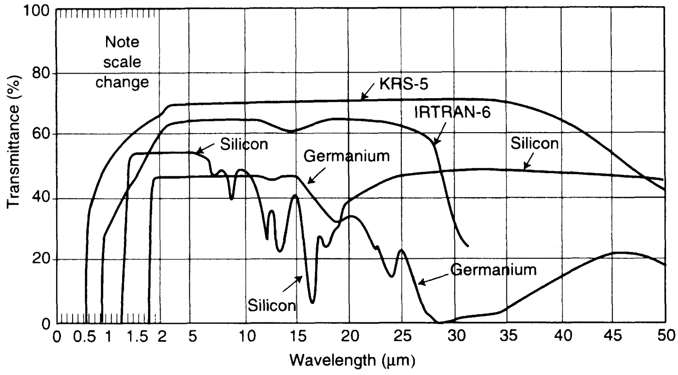

FIG. 5. Relative transmittance of various materials used in infrared refractive elements (2.0 mm thickness) (Spiro and Schlessinger, 1995).

the visible region of the spectrum. Departures from conventional techniques in high-sensitivity applications include the necessity of accounting for thermal radiation from the optical elements themselves. The collection optics for a typical infrared sensor function in the same manner as elements in the optical region of the spectrum, i.e., to collect and focus the radiation on the detector. To avoid unacceptable signal attenuation, highly transmissive refractive elements and highly reflective mirrors are required. Although these considerations are seemingly elementary, the practicalities of achieving these properties with conventional optical materials are often problematic. In some cases exotic materials have been developed that have the necessary mechanical and thermal properties as well. The spectral transmittance of some typical IR optical materials are shown in Fig. 5. Reflective elements are typically metal substrates with protective coatings to achieve environmental stability. The spectral reflectance of various metals used for reflective elements are shown in Fig. 6. See OPTICS, LINEAR; OPTICS, GEOMETRICAL.

1.1.3 Filters Spectral filters serve to limit the spectral range of the IR radiation falling on the detector. Limiting the observable spectral bandpass pays large dividends in signal detectivity by restricting the extent of undesirable contributions from extraneous backgrounds. Filters may employ interference coatings or transmissive elements exhibiting selective absorption. Other selectively reflective devices and gratings can be used as well (Wolfe and Zissis, 1985). These devices include Christiansen filters, made up of small, closely packed particles of an infrared-transparent substance suspended in a liquid or a gas; wavelength-selective reflection filters made up of various crystalline materials; refraction filters that depend on the

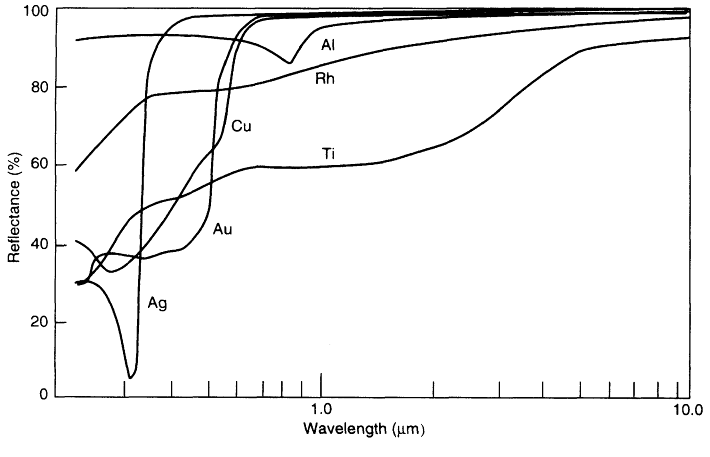

FIG. 6. Reflectance of various metallic films used in infrared reflective elements (Wolfe and Zissis, 1985).

wavelength dependence of the index of refraction to focus radiation selectively at different points along the optical axis; and polarization interference filters.

1.1.4 Cryogenics To achieve high sensitivity and in some cases simply to effect operation over a given spectral region, detectors must be cooled well below room temperature. Cooling reduces inherent detector and preamplifier noise and is nearly essential in most astronomical, environmental, and military applications. In applications beyond 10 μm, additional cooling of the optics, filter, and other components is necessary to reduce internally generated background radiation. These devices are sometimes called coolers but are more properly termed cryostats. Cryostats may use electrical means, such as the thermoelectric effect, or compressed gases, as in the Joule–Thompson effect. Other types include liquid or solid cryogens as the refrigerant such as liquid nitrogen, liquid helium, or solid CO_2. Cryostats must be well insulated from the environment, holding the detector temperature constant within a very narrow range. They must be highly reliable and compact, especially for spaceborne applications, and must possess a considerable degree of mechanical integrity under harsh environmental conditions. A typical cryostat is shown in Fig. 7.

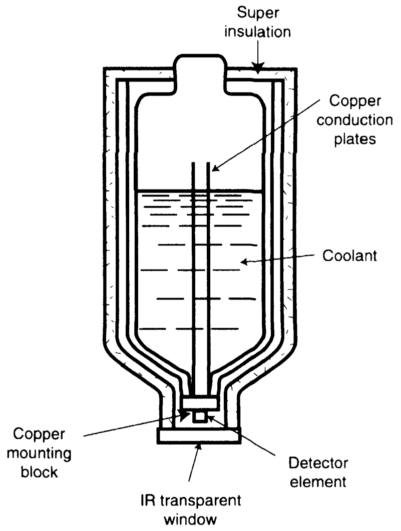

FIG. 7. A conventional Dewar flask used for holding liquid cryogens to cool the detector. The detector is in physical contact with a copper block immersed in the coolant. Typical coolants are liquid nitrogen at 77 K or helium at 4 K. (Wolfe and Zissis, 1985.)

1.1.5 Electronics Modern IR sensors employ sensitive electronics to amplify the extremely weak signals generated by the IR detectors. Such IR-detector electronics are quite similar to the circuitry used with visible detectors. The desirable characteristics are low noise, high gain, and impedance matched to the detector for efficient power transfer. Often the preamplifier electronics are cooled along with the detector to reduce noise. The idealized electronics design is such that the overall detector/preamplifier performance is detector limited. High-impedance detectors are very sensitive to electromagnetic interference and external noise, and amplifiers must be located in close physical proximity to the detectors.

In early sensor designs, preamplifiers built from discrete components were placed in close proximity to the detectors on the cold focal-plane mounting. The advent of detector arrays with thousands of separate detectors requires a different approach that integrates the amplifier and readout electronics with the detector. The CCD or change-coupled device integrates all of these functions on a single chip and has become the most prevalent device used for infrared imaging. A typical CCD is shown in Fig. 8.

1.1.6 Scanners To create images of infrared scenes, a means must be provided to scan the detector field of view over a much larger area than that addressed by a single detector. As described above, one way to achieve this is with focal-plane arrays imaging simultaneously the entire scene of interest on the detector array. In single-detector designs, electromechanical scanning devices are employed to move one or more optical elements physically to achieve the desired effect. Regular and precise motion is achieved by optical-component rotation or oscillation or by various electromagnetic means. A typical mechanical scanner is shown in Fig. 9. Piezoelectric, electro-optical, and acousto-optical effects are used to effect motion of the optical axis. Finally, motion of a spacecraft or aircraft platform may be used to create contiguous imagery and is often employed when such motion is regular and predictable.

1.1.7 Displays and Signal Processing The electrical signals from the sensor may ultimately be used to drive various cathode-

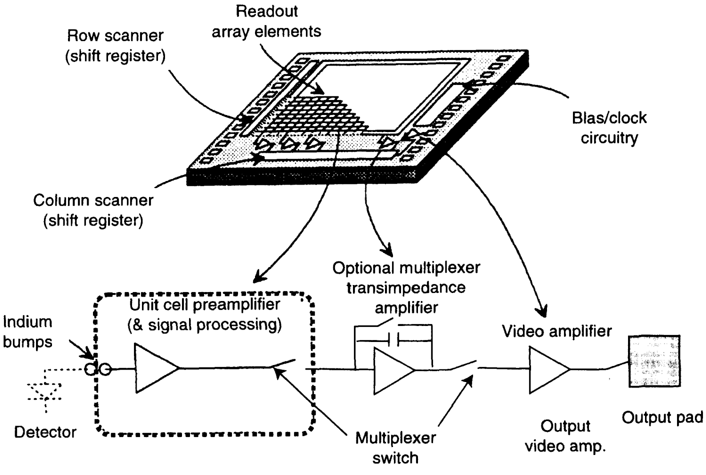
Readout array elements
Row scanner (shift register)
Bias/clock circuitry
Column scanner (shift register)
Optional multiplexer transimpedance amplifier
Unit cell preamplifier (& signal processing)
Indium bumps
Video amplifier
Detector
Multiplexer switch
Output video amp.
Output pad

FIG. 8. Focal-plane array and readout electronics integrated onto a single chip (Accetta and Shumaker, 1993). Charge is transferred synchronously from the individual detector elements to shift registers and subsequently transformed into electrical impulses by the video amplifiers. This technique reduces the number of output leads in an $n \times n$ array from n^2 to $2n$.

ray tubes (CRTs) or solid-state displays for visual interpretation, transferred to photographic film for archival purposes, or processed for additional information. Further processing may take the form of noise or background reduction or call upon a multitude of image-processing techniques for feature enhancement and detection. The nature of this additional information and the processing required to extract it is application specific and has a myriad of possibilities. These include electrical signals indicating the presence or absence of a particular object, utilization of infrared spectral information in the assignment of various biological or geological forms as in remote sensing, and the location and nature of emitting objects as in infrared astronomy. The spatial distribution

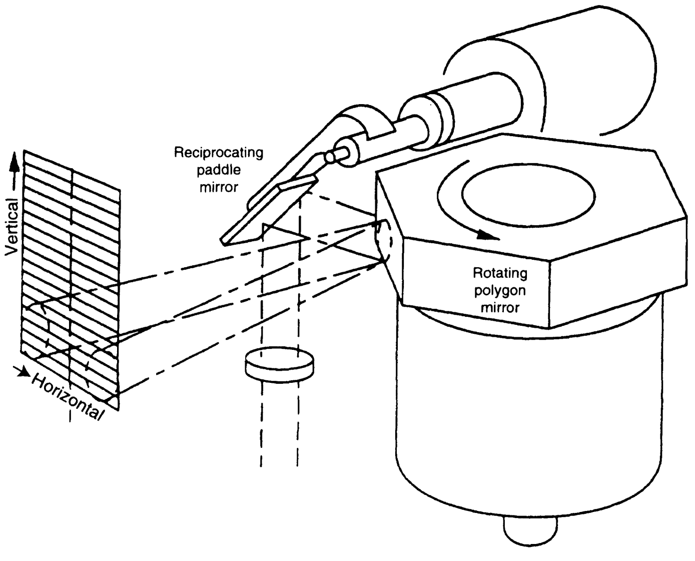
Vertical
Horizontal
Reciprocating paddle mirror
Rotating polygon mirror

FIG. 9. Rotating-polygon and nodding-mirror scanner configured for rectangular scanning (Accetta and Shumaker, 1993). The reciprocating paddle mirror moves the field of view in the vertical direction while the rotating polygon mirror creates motion in the horizontal direction.

of temperature is used in diagnostic medicine, process control, and industrial inspection. Other applications include surveillance and guidance and control as in military and comparable industrial operations.

2. DESIGN, FUNCTION, AND OPERATION

2.1 Sensor Figures of Merit

The figure of merit of an IR sensor often depends on the specific application. In industrial applications the minimum detectable temperature change may serve as a figure of merit. In certain military applications the figure of merit is quite frequently the probability of detecting a given target in a given background, given a certain tolerable false alarm rate. In imaging applications spatial resolution in combination with sensitivity is often the determining factor, i.e., the degree to which fine detail can be resolved by a given sensor. In an infrared radiometer (discussed below) the absolute radiometric accuracy of the sensor is the overriding concern. Some formalisms commonly employed in the assessment of infrared sensor performance include noise equivalent irradiance (NEI), noise equivalent temperature difference (NETD), and minimum resolvable temperature difference (MRTD). In complex cases such as forward-looking infrared systems (FLIRs) discussed below, a human operator is involved. Thus psychophysical factors are an intrinsic part of the figure of merit, invoking the statistical performance of a large number of human operators. In environmental systems minimum detectable concentration of a given molecular species can be a figure of merit.

As implied by the foregoing, figures of merit are quite specific to a given sensor type. However, in almost all applications the signal-to-noise ratio is an underlying consideration in determining sensor performance. The probability of detection of an infrared signal is a strong function of the signal-to-noise ratio. In imaging applications it affects the quality of the displayed image. Noise may result from the detector, preamplifier, or background, and noise mitigation is the major preoccupation of the sensor designer. Information extraction in noisy conditions constitutes a voluminous portion of the modern literature on signal processing.

2.2 Scanning Systems

Historically, the first infrared sensors were simple point detectors with the field of view limited by the detector geometry. To create infrared scenes or maps, the point detectors were mechanically scanned over the scene to create infrared images. The essence of the scanning system is that a single detector element may be used to form an entire image if the optics in the system scans over the object point by point. An infrared line scanner is an imaging device that forms images, usually by successive scans of a rotating mirror in a specific direction. The perpendicular scan direction needed for a two-dimensional image is provided by forward motion of the sensor platform or alternatively by motion of the scene of interest across the field of the sensor as in Fig. 10. In environmental remote-sensing applications, the platform could be an aircraft or satellite. Early "forward-looking infrared systems" (FLIRS), as they were called (described below), were adaptations of line scanners with the addition of a mechanical means to do the scanning in the other dimension. A modern version of this system is the satellite-based LANDSAT and SPOT sensors described in Lillesand and Kiefer (1994).

2.3 Staring Systems

Staring systems, as the name implies, simply stare. There is a strong analogy to the human eye. The entire scene is imaged onto a two-dimensional array of detectors (FPA) as previously discussed. Because of the small but finite spacing of the individual detector elements, the image is essentially sampled at regularly spaced intervals. The ultimate spatial resolution of the image so obtained is determined by the individual detector geometry, the array geometry, the resolving capability of the optics, and other factors. In the visible region of the spectrum, the largest commercial and military applications of staring systems are conventional TV cameras that operate over a very wide range of lighting conditions. With the availability of infrared focal-plane arrays, the TV-camera con-

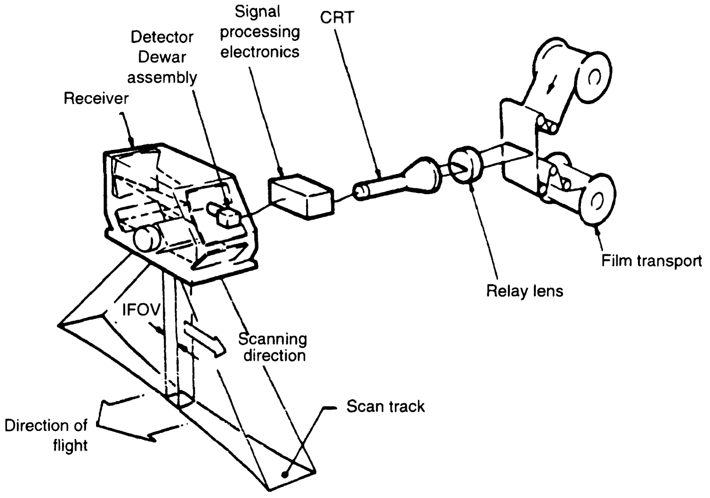

FIG. 10. Typical IR line scanner used for environmental and military data gathering from a moving platform (Accetta and Shumaker, 1993). The image is recorded in synchronism with the motion of the platform and with the scanning device. Generally, the imagery must be further corrected for small perturbations in the flight path of the aircraft.

cept has been extended into the infrared regions of the spectrum as with the thermal imager discussed previously.

2.4 Infrared Phenomenology

It is the combined phenomenology of the source, background, and atmosphere that will drive the design of a sensor. Other factors such as economics, the environment, and available technology may play a significant role in the choice as well. A proper understanding of the phenomenological relationships between these factors is essential to achieving satisfactory sensor performance.

As shown in Fig. 11, the radiation from a typical object of interest embedded in a terrestrial background is actually a superposition of a number of sources of radiation including the background radiation from the terrain within the sensor field of view, the background radiation reflected from the object into the field of view, the radiation from the object itself, and the radiation from the intervening atmospheric path. These sources are in turn affected by the atmospheric absorption along the path from the object to the sensor. In the general case the sensor-design process requires a detailed understanding of all relevant sources of radiation. The determination of these effects is often difficult and highly dependent on local environmental conditions. The proper choice of the spectral band that optimizes sensor performance demands a consideration of the above factors. The degree of optimization of a specific sensor design depends on the extent to which these factors are known and understood.

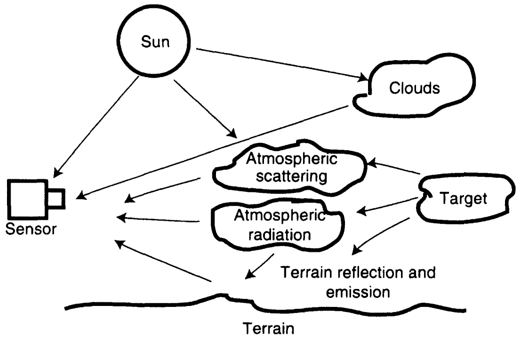

FIG. 11. Naturally occurring phenomenology associated with infrared sensing. There are multiple sources of direct and reflected natural radiation that interfere with the target signal.

2.4.1 Object Radiation

The essential infrared characteristics of an object are temperature, spectral reflectance, and emissivity. From these quantities, the emitted and reflected radiation from an object or scene in a given environmental geometry can be determined. In the general case object radiation is not blackbody in nature, and thus, its spectral properties have to be accounted for. Reflectance spectra from some natural materials are shown in Fig. 12. The spectral characteristics of the materials constituting the scene or object of interest will again af-

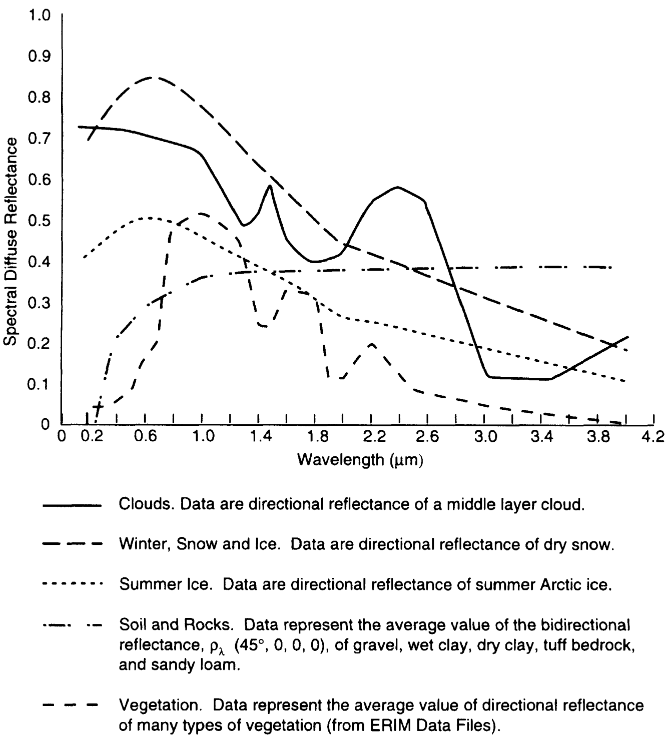

FIG. 12. Diffuse reflectance of earth-atmosphere components (Wolfe and Zissis, 1985). The spectral dependence of these components may be used to discriminate against undesirable effects on sensor performance.

fect the choice of sensor spectral band of operation.

2.4.2 Backgrounds Backgrounds represent an undesirable component of incident radiation and are a serious source of interference in the detection problem. In Fig. 11 there are several sources of radiation including reflected solar and emitted radiation from surrounding terrain or clouds; thermal radiation from the atmosphere itself; and solar, lunar, and stellar sources. The distinction between objects and background is arbitrary depending on the interests of the observer. Sensors are generally designed to maximize object signal and minimize background interference by selection of the spectral band that optimizes the ratio of signal power to background power.

2.4.3 Atmospherics The atmosphere can play a significant role in the absorption of IR radiation over long observation distances. The absorption is both wavelength and distance dependent; thus sensors operating over long distances must use regions of the spectrum with relatively low absorption to minimize signal loss. Most infrared absorption in the atmosphere is due to water vapor and CO_2 with a minor dependence on other gaseous species. Scattering also plays a role as a potential signal attenuator. A typical atmospheric transmission curve is shown in Fig. 13. Note that the "windows" of relatively high transmission occur in the 3.0–5.5- and 8.0–14.0-μm regions of the spectrum. In certain astronomical applications observation at wavelengths outside of these windows requires exoatmospheric-based instrumentation using balloons, high-altitude aircraft, satellites, or sounding rockets as platforms.

3. MANUFACTURING AND ECONOMICS

Because of the diversity in both technology and complexity of infrared sensors, the economic and manufacturing factors are not readily discussed in detail. In the simplest cases—e.g., an industrial sensor employing an uncooled detector—the major components of the sensor including the detector element, electronics, and housing can be readily manufactured in-house for very low cost. Optics, if required, are rarely manufactured

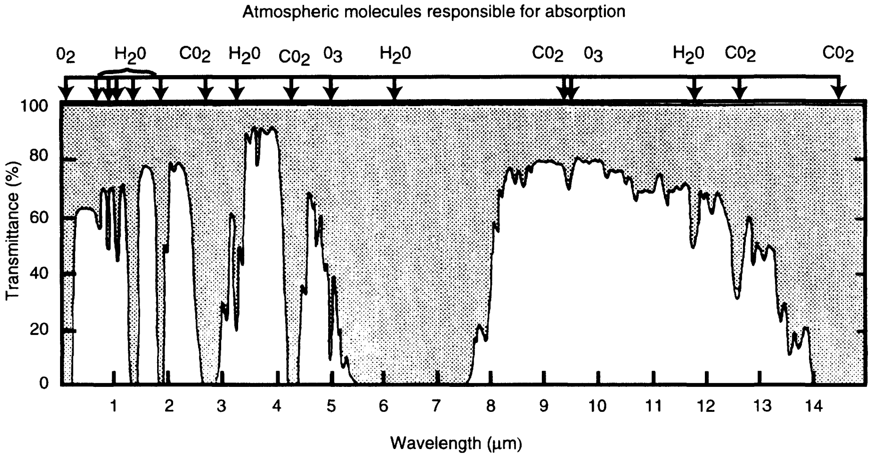

FIG. 13. Spectral transmittance of the atmosphere from 0.0 to 15.0 μm. Note the presence of "windows" at 3.0–5.5 μm and 8.0–14.0 μm. (Lillesand and Kiefer, 1994).

in-house. In contrast, most manufacturers of complex sensors typically buy required specialized and sophisticated components when possible rather than attempting to manufacture them. "Build or buy" decisions generally lean in the "buy" direction. The potential reduction in revenue resulting from nonvertically integrated manufacturing is offset by relieving the manufacturer of the necessity of keeping up with the state of the art in a number of diverse technologies. Other expenses are reduced as well.

It is not uncommon to purchase detectors, cryogenics, and electronic and optical components from several outside sources. Manufacturers typically rely on in-house sources for mechanical design, electronic design, final assembly, and testing. It would be extremely rare to find all the necessary technology under one roof. Complex sensor manufacturers are thus highly dependent on external sources of supply.

As might be expected, the overall economic aspects depend upon the attributes of the sensor components and volume produced. Infrared detectors and arrays that depend upon semiconductor technology and cryostats are usually much more expensive than their uncooled pyroelectric and thermopile equivalents. Infrared optical materials and components are also more expensive than their visible counterparts. For a comparison with components that operate in the visible region of the spectrum, a modern visible TV camera employing an uncooled silicon detector array can be currently purchased for a few hundred dollars in quantity. However, a sensitive infrared equivalent will likely cost tens of thousands of dollars and could exceed millions in some extremely specialized cases. As usage increases, prices will drop accordingly, especially in those cases highly amenable to mass production and the use of inexpensive and uncooled detectors or FPAs, plastic optics, and application-specific integrated circuits. Significant production quantities should strongly drive price reductions.

Although we might conjecture that many additional commercial and industrial activities could benefit greatly from the mass application of infrared sensors, the overall cost–benefit relationship must be understood by the user before such large-scale application would occur. There is a great deal of promise that over the near term the cost of sensitive, rugged, high-quality infrared sensors will be significantly reduced.

4. MAJOR APPLICATIONS

There are a multitude of scientific, industrial, and military applications for IR sensors. Some of the major applications are shown in Table 1.

Table 1. Major infrared applications (adapted from Hudson, 1969).

Military	Industrial	Medical	Scientific
Intrusion detection	Fire detection	Obstacle detection for the blind	Satellite detection
Bomber defense	Fuel ignition monitor	Skin-temperature measurement	Space-vehicle navigation and flight control
Missile guidance	Law enforcement	Cancer detection	Horizon sensors
Navigation and flight control	Remote dimensioning	Monitoring of healing and infection	Solar trackers
Ship, aircraft, ICBM, mine detection	Process control	Remote biosensors	Optical structure of horizon
Fire control	Temperature measurement of brake linings, power lines, cutting tools, welding and soldering operation, ingots	Studies of skin heating	Lunar, planetary, and stellar temperatures
Proximity fuses	Turbulence detection	Detection and monitoring of pollution	Remote sensing of weather conditions
Aircraft collision warning	Analysis of organic chemicals and gases	CO_2 levels in blood and breath	Heat transfer in plants
Target signatures	Blood-alcohol levels	Localization of placental sites	Earth heat balance
Terrain analysis	Pipeline leaks	Clothing efficiency	Detection of extraterrestrial life
Poison gas detection	Water contaminants and pollution	Diagnosis of incipient stroke	Terrain analysis
Fuel vapor detection	Display of heating patterns	Measurement of pupillary diameter	Monitor spacecraft atmospheres
Detection of contaminants	Efficiency of thermal insulation	Location of vein and arterial blockages	Zero-*g* liquid-level gage
Reconnaissance	Crime prevention	Monitoring eye movements	Magnetic field measurement
Surveillance	Crop-disease detection	Nocturnal habits of animals	Earth resource survey
Thermal mapping	Agricultural assessments	Corneal opacities	Mapping ocean flow
Underground facilities detection	Automatic camera focus	Heat therapy	Forest fire detection
Damage assessment	Intrusion detection		Water pollution studies
Night driving	Auto-collision prevention		Location of crevasses
Weapon location	Traffic monitoring		Sea-ice reconnaisance
Camouflage detection	Radiant heating and drying		Petroleum exploration
Station keeping	Data and voice communications		Forgery detection
Docking and loading	Intervehicle speed sensing		Epitaxial film thickness
Communications	Aircraft landing aids		Planetary constituents
Countermeasures	Material bonding inspection		Gem identification
Range finding	Forgery detection		Water quality analysis
Command links	Personal identification		Detection of diseased crops
			Animal communications

4.1 Infrared Radiation Pyrometry

Infrared radiation pyrometry is a generic term that is often used to describe remote temperature measurement by observation of emitted thermal flux in one or more infrared spectral bands. An infrared pyrometer is analogous to an optical pyrometer whose complexity ranges from a simple single-detector and amplifier combination to more complex units such as a hand-held optically sighted device operating in multiple spectral bands. Except for the hand-held variety, fields of view and directions of observation are generally fixed. Pyrometers generally require modest sensitivity, simplicity, rugged construction, and low cost. The characteristics of pyroelectric and thermopile detectors are consistent with these requirements. Pyrometers may be calibrated to measure temperature or relative flux. Numerous examples of relevant applications are cited in Table 1, for many of which detailed discussions are given in Hudson (1969) and Kaplan (1993). Of particular importance are industrial applications such as the monitoring and control of processes requiring continuous remote measurement of temperature. An illustrative example is the temperature control of moving steel in a rolling mill or extrusion line. Another is the monitoring of transformer and insulator temperature in power-distribution systems. The general inaccessibility of these components requires portable remote measurement systems. There are several models of this type of sensor in current production.

Infrared radiation pyrometry can be applied to many other industrial problems such as intrusion alarms or appliance control where the measured temperature may cross certain pre-established thresholds triggering a control signal or other event.

4.2 Search and Tracking Systems

An infrared search and tracking system (IRST) is designed to detect reliably, to locate, and continuously to track infrared-emitting objects and targets in the presence of background radiation and other disturbances. Principally employed by the military, IRST systems are usually confined to unresolved point-source objects in highly structured backgrounds and at ranges beyond 5 km. They generally use mechanical scanning in both dimensions to create a two-dimensional image or "frame." The spectral bands employed are the usual atmospheric windows at 3.0–5.5 and 8–12 μm. Psychophysical factors are not a consideration in these systems since the search and track processes are highly automated. Most of the current research activity in IRST systems is in the signal-processing domain to extract targets from severe background clutter. A typical military IRST is shown in Fig. 14.

4.3 Imaging Systems

4.3.1 Forward-Looking Infrared (FLIR) Systems An FLIR is an infrared analog of a visible TV camera. The term FLIR is an acronym for "forward-looking infrared" that originated from the modification of an early 1960s U.S. Air Force downward-looking infrared mapper to a forward-looking real-time infrared imaging system. The name has remained, although inappropriate, and is usually used in a military context. Thermal or infrared imager would be a better name. The FLIR systems are operated nearly exclusively in the 3.0–5.5- or 8.0–14.0-μm region of the spectrum corresponding to atmospheric windows of high transmissivity. The FLIR systems are found in both scanning and staring configurations and generally do not rely on platform motion for the creation of an image frame. High-performance military systems are used in targeting, navigation, and surveillance applications for the imaging of extended targets. A block diagram of a typical FLIR system is shown in Fig. 15.

4.3.2 Medical and Industrial Thermal Imagers Medical and industrial thermal imagers use the same basic technology as FLIR systems. In these applications infrared imaging and analysis is often referred to as thermography. The instruments used for this

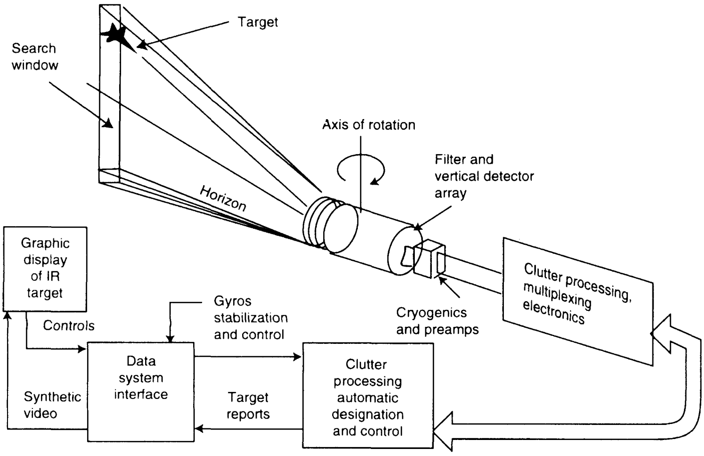
Target
Search window
Axis of rotation
Filter and vertical detector array
Horizon
Graphic display of IR target
Controls
Synthetic video
Gyros stabilization and control
Data system interface
Target reports
Cryogenics and preamps
Clutter processing automatic designation and control
Clutter processing, multiplexing electronics

FIG. 14. Block diagram of a typical infrared search and track system (IRST) (Accetta and Shumaker, 1993). The sensor axis is continuously swept about the horizon. When a target detection occurs, the system reverts to a tracking mode of operation maintaining continuous surveillance of the target.

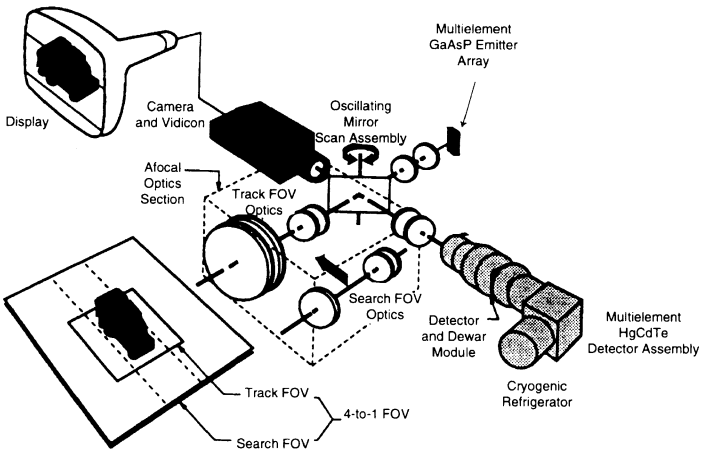

FIG. 15. First-generation forward looking infrared system (FLIR) (Accetta and Shumaker, 1993). Rather than driving the display directly in this modular system, the electrical signals from the detector array drive a solid-state GaAsP light-emitter array. The output of the emitter array is used as a strong optical excitation source for the Vidicon. "FOV" stands for field of view.

purpose are generally less expensive than their military counterparts. The infrared image of the author in Fig. 1 is characteristic of medical infrared imagery. Temperature anomalies may be indicative of certain pathological conditions such as breast tumors, abnormal arterial and venous functions, and infections. Industrial applications of IR imagers include the remote inspection of industrial facilities, and process monitoring and control. For example, utility companies often use mobile industrial thermography units to locate overheating components on inaccessible utility poles. Electronics companies use imagers to detect overheating components, and biologists use these systems to gauge cellular metabolism.

4.3.3 Astronomical Imagers The importance of infrared sensors in astronomy lies in their ability to detect the emissions of otherwise invisible dust grains and compact interstellar sources heated by stellar sources or shock waves. Prototypical stars at their earliest stages of formation and not yet at the threshold of visible emission are also detectable (Zeilik and Smith, 1987). Additional advantages provided by observation in the infrared regions of the spectrum are less attenuation by interstellar dust than in the visible region and the ability to observe in daylight. Analogous to their visible photometry counterparts, the earliest infrared astronomical applications simply focused the radiation from a stellar source onto a calibrated lead sulfide infrared detector. The emitted infrared intensity from stellar sources was computed after correcting for atmospheric and interstellar absorption and the detector spectral response.

Modern infrared astronomical sensors range from sensitive bolometers consisting of a tiny chip of doped germanium operated at very low temperature to the employment of sophisticated InSb (indium antimonide) or HgCdTe (mercury-cadmium-telluride, also known as MCT) infrared focal-plane arrays to produce detailed imagery. Infrared astronomy has become a major subdiscipline of observational astronomy.

A vast majority of optical telescopes have been used for infrared observations at one time or another. Early adaptations used infrared-sensitive photometers. There are now a number of astronomical telescopes specifically designed for infrared use and equipped with sophisticated focal-plane arrays (McLean, 1994). One of the many major triumphs of infrared astronomy is the Infrared Astronomical Satellite (IRAS). Designed to circumvent the limitations of endoatmospheric observation due to extinction and turbulence (also known as "seeing"), this satellite-borne infrared imager generated extensive survey maps revealing a multitude of previously unknown IR celestial sources (Beichman *et al.*, 1988). The telescope, shown in Fig. 16, was a two-mirror Ritchey–Chretien configuration mounted inside a superfluid helium tank mounted within an evacuated main shell. The liquid helium supply cooled the entire telescope including the focal-plane array to achieve extremely sensitive performance. The major specifications of

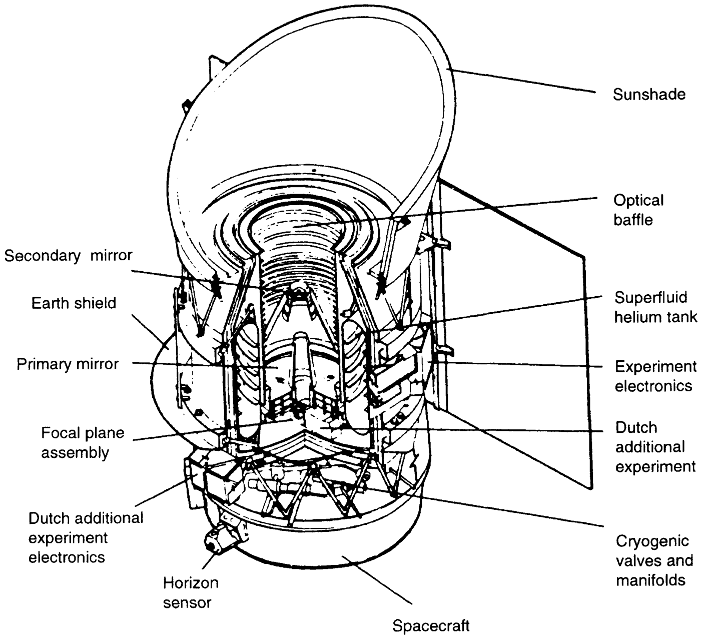

FIG. 16. The Infrared Astronomical Satellite (IRAS) infrared sensor configuration (Beichman *et al.*, 1988). This instrument is a successful example of a modern, extremely complex infrared sensor.

this remarkable testimony to modern infrared-sensor technology include a collecting aperture of 57.0 cm, a focal-plane operating temperature of 2.6 K, eight infrared and two visible detector arrays, and operation in the 12-, 25-, 60-, and 100-μm wavelength bands.

4.4 Remote Sensing and Environmental Applications

4.4.1 Airborne and Spaceborne Spectral Systems The earliest environmental applications of infrared sensors used photographic film sensitized in the near-infrared region. As sensor technology evolved, the spectral region of observance was extended to longer wavelengths and multiple spectral bands. Analogous to its counterpart in the visible region of the spectrum, infrared multispectral sensing is a relatively new airborne or spaceborne technique to exploit the information content of infrared images by intercomparisons of identical imagery in multiple spectral bands. Images are collected simultaneously in multiple spectral bands, and good image registration is critical to achieve optimum performance. One means of collection is line scanning, as shown in Fig. 10. Useful information arises from differences in spectral reflectance properties of the terrain that can be correlated to various material types after correcting for source temperatures, background radiation sources, path radiance, and transmission factors. The basic technique described is employed in satellites and aircraft sensor platforms for pollution detection and monitoring, land-use assessment, treaty monitoring, change detection, soil and moisture content, crop inventories, snow cover, bathymetry, mineral exploration, and disaster-damage assessment (Lillesand and Kiefer, 1994).

The principle behind multispectral analysis is that the ratio of the intensities in two or more spectral bands is nearly constant and unique for a given material. Objects within a given scene can be classified on the basis of the observed ratio. The spectral bands that optimize separation between various materials are carefully chosen from large trial data sets for consistency. This category of infrared sensing is further subdivided into multispectral, hyperspectral, and ultraspectral, depending on the spectral resolution used. Multispectral sensors operate with spectral bandwidths on the order of 10% of the center wavelength and typically use from 5 to 20 distinct bands. Hyperspectral systems operate at about 1% of the center wavelength and may use 100–200 bands. Ultraspectral sensors have bandwidths on the order of 0.1% of the center wavelength and may have up to 1000 separate spectral bands. The LANDSAT series is perhaps the most well known of the spaceborne multispectral sensor systems. Depending on the configuration, several different sensor configurations were employed, including both infrared and visible sensors operating in multiple bands. The French SPOT satellite offers similar capabilities.

Thermal scanners operate in the long-wave infrared region (LWIR) of the spectrum and can be used to infer surface-temperature differences. A good example of the use of this technique is to monitor surface water temperature near nuclear power plants. Heating and cooling rates over diurnal periods offer information about the type, condition, and properties of observed materials as well. The NASA Thermal Infrared Multispectral Scanner operates exclusively in the thermal infrared region using six spectral bands.

4.4.2 Infrared Remote Sensing of the Atmosphere A special class of infrared instruments has arisen in the context of atmospheric remote sensing for either pollutant or natural composition studies. The primary properties of the atmosphere usually observed for this purpose are absorption, aerosol scattering, and fluorescence. A given gaseous species has a unique absorption-band structure, and thus with some modification the techniques of molecular spectroscopy are applicable to infrared remote sensing of the atmosphere.

Rayleigh or elastic molecular scattering by gases involves no shift in the incident wavelength and is strongly wavelength dependent. Mie scattering, also elastic, is an observable aerosol characteristic and may exhibit wavelength dependence depending on the size of the particle. Inelastic scattering involves a shift between incident and scattered radiation. Raman scattering is an example of inelastic scattering at the molecular level. By observation of the scattering and absorption properties of the atmosphere with infrared instruments, significant information can be obtained regarding the molecular and aerosol species and concentration. Scattering is often investigated with atmospheric light detection and ranging (LIDAR) systems (see Sec. 4.7).

A prominent instrument and technique for the observation of the absorptive properties of the atmosphere is the Fourier-transform spectrometer (FTS). Operating in the infrared regions of the spectrum, this application is usually referred to as an FT-IR (Simpson, 1994). In its elementary form an infrared spectrometer is an infrared sensor with a highly selective wavelength-filtering device attached. (See SPECTROMETERS, INFRARED.) The wavelength-selection device used in the typical FT-IR is based upon the Michelson interferometer and has superior sensitivity over conventional spectrometers for the observation of relatively weak spectral features (see INTERFEROMETERS AND INTERFEROMETRY). The FT-IR is usually employed in the long-path configuration, as shown in Fig. 17, using a hot background source as the source of radiation and observing molecular absorption lines from the intervening atmosphere. The preferable configuration for this instrument is termed "monostatic" where a wide-banded infrared illumination source is collocated with the receiver. The "bistatic" mode may be used over extremely long paths to maximize signal return. In this configuration the illuminator and receiver are located at opposite ends of the region under observation.

FT-IR systems have been employed in a

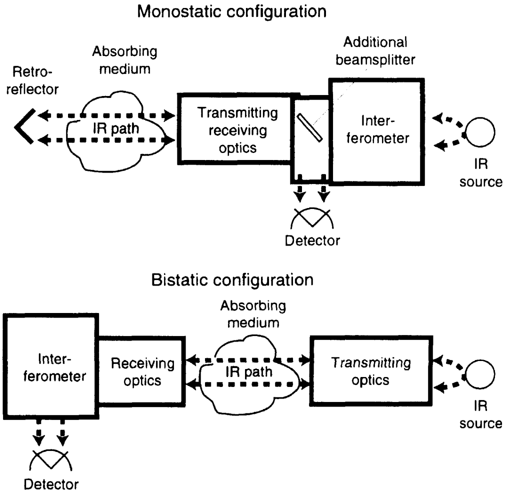

FIG. 17. Monostatic and bistatic FT-IR long-path configurations (Russwurrm and Childers, 1995). Because of superior noise canceling ability, the monostatic configuration is preferred.

variety of data-gathering experiments, including limb sounding of atmospheric absorption from satellites and balloon platforms, absorption of the atmosphere using the moon as a source, and earthbound measurements of absorption against the sun. In the context of pollution monitoring, FT-IR instruments have been used for stratospheric emission measurements from aircraft and balloons, smokestack emission monitoring at a distance, and analysis of forest fire emissions. Recently, FT-IR techniques have emerged as a technique for the generation of spectrally resolved images for use in medical, industrial, and astronomical applications.

4.5 Radiometers

Radiometers are calibrated infrared sensors and are used as laboratory or field instruments for quantitative measurement of emitted infrared flux from sources of interest (Wolfe and Zissis, 1985). There is a direct and quantifiable relationship between the observed signal level and the apparent radiance of the source. In addition to wide-scale research applications, such as the measurement of terrestrial heat transfer and solar output, there are a number of commercial applications, including plastic, glass, and thin-film process monitoring, and laser power meters. A number of special- and general-purpose commercial units are available. More sophisticated devices are spectroradiometric insofar as they may use a number of discrete spectral bands or be continuously tunable over a given region of the infrared spectrum and are thus akin to the conventional laboratory infrared spectrometer. Some radiometers employ focal-plane arrays to yield calibrated imagery.

4.6 Infrared Laser Radar

Infrared laser radar in its simplest form is simply the optical analog of the conventional microwave radar. It is an example of an active infrared sensor that employs its own source of (generally narrow-band) illumination and has the additional capability of measuring the range to an object of interest. More complex versions may be employed in imaging modes. Laser radar instruments are used to measure range-dependent quantities such as aerosol concentration and molecular composition (see Sec. 4.7). Laser radars may utilize repetitively pulsed, amplitude-modulated or frequency-modulated lasers to determine range as in conventional radars. Laser radar is not confined to the infrared region but is conveniently employed there because of the fortuitous existence of both atmospheric windows and high-power lasers operating at 0.8, 1.06, and 10.6 μm using GaAs, Nd-YAG, and CO_2 lasers, respectively. (See LASERS, GAS; LASERS, SOLID STATE; LASER TECHNOLOGY.) Increased levels of sophistication include heterodyne detection and synthetic-aperture imaging techniques. Infrared laser radar is employed in both military and industrial applications for tasks ranging from the elementary, such as proximity detectors and communications systems, to the highly complex, such as target detection, robotic guidance, and inspection of manufactured parts.

4.7 LIDAR and DIAL

Light detection and ranging (LIDAR) is an example of a relatively sophisticated active infrared sensor. Although the following definition is not universally accepted, LIDAR is a variant of laser radar used by the environmental science community to glean atmospheric information by analysis of backscattered range-dependent radiation from atmospheric molecular or aerosol constituents (Measures, 1985). The objects of interest for these systems are atmospheric particulates, aerosols, and absorptive species. The observable is the range-resolved backscattered radiation in which the intensity of the backscattered component is a function of the aerosol concentration. The LIDAR systems operate in the IR region but are also employed at the shorter wavelengths where the scattering effects are greater. Frequency shifts in the backscattered component of high-power visible or UV lasers due to Raman scattering can be characteristic of molecular rotational-vibrational transitions. Thus, they can be used to identify the presence of certain gases including those which, by homonuclear symmetry, do not radiate in the infrared.

Differential absorption LIDAR (DIAL) refers to a subset of LIDAR systems that use differences in atmospheric backscatter be-

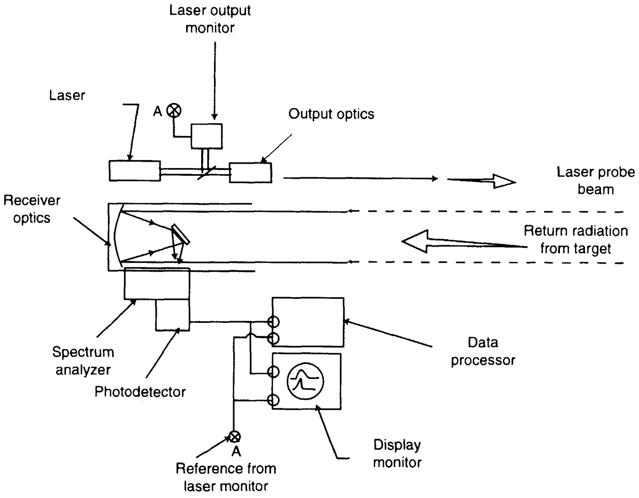

FIG. 18. A block diagram of a typical laser environmental sensor. The laser may operate in a comparative mode at multiple frequencies, as in the case of DIAL.

tween two closely spaced transmitted laser wavelengths to deduce concentrations of atmospheric components. The wavelengths are chosen so that one lies within the absorption line of the molecule of interest and the other lies just off the line. The range-resolved difference in intensity between the two returns is a measure of the species concentration along the path. Figure 18 illustrates a typical infrared LIDAR.

5. FURTHER RESEARCH

Infrared sensors could benefit greatly from further research in several associated technology areas. Clearly there is a fairly rapid migration from military applications to civilian and industrial uses. The reader is referred to the applicable sections in this Encyclopedia for discussions of recent developments in infrared sensor components. Of particular importance is the development of sensitive and rugged detector elements free from the encumbrances of cryogenics or thermoelectric cooling. Advances here could substantially reduce cost and complexity of IR detectors and imaging arrays, thus yielding wider applications in industrial, transportation, and medical applications. Increased yields in the semiconductor detector manufacturing process would be a contributing factor to lower cost.

In the imaging domain a number of new techniques under development such as aperture-plane interferometry, passive interferometric range-angle imaging, and range-Doppler imaging (Accetta and Shumaker, 1993, Vol. 8) are expected to expand the versatility of imaging processes useful in the infrared.

The environmental sciences would benefit greatly from the availability of low-cost, high-power tunable laser diodes for atmospheric investigations (see LASERS). More spectral versatility in laser sources for DIAL and LIDAR systems would increase the general utility of these instruments.

Although many of the early advances in infrared technology were heavily driven by military funding, current commercial interests have sparked a considerable amount of civilian research and development that will undoubtedly lead to further usage of infrared sensors in many new applications.

GLOSSARY

Bathymetry: Mapping of the sea bottom.

Bolometer: A very sensitive, wide–spectral-band detector based on an element whose resistance varies as a function of absorbed radiation.

CCD: Charge-coupled device. A generic term for a focal-plane array device to multiplex outputs of the many elements over a few data lines.

Clutter: Highly inhomogeneous infrared backgrounds.

Cryostat: A cooling device.

DIAL: Differential absorption LIDAR. A laser radar system used for atmospheric investigations. The system transmits multiple laser frequencies near an absorption-line edge and measures the differential backscattered radiation between on-line and off-line measurements to deduce species concentration.

FLIR: Forward-looking infrared. Military term for an infrared imaging system used in navigation and targeting applications. Civilian versions are in use in industrial and medical applications.

FPA: Focal-plane array. A square or rectangular array of detectors located in the focal plane of an optical system.

FT-IR: Fourier-transform interferometer operating in the infrared.

IRAS: Infrared imaging astronomical satellite, launched by NASA in 1983.

IRCCD: A CCD operating in the infrared region of the spectrum.

IRST: Infrared search and track system. A passive infrared system primarily used by the military to locate and track infrared emitting targets.

LIDAR: Light detection and ranging. A research laser radar system used for atmospheric investigations using backscattered radiation from aerosols.

Long-Wave Infrared (LWIR or Thermal): 8.0 to 14.0 μm.

Mid-Wave Infrared (MWIR): 3.0 to 5.5 μm.

Near Infrared (NIR): 0.7 to 1.1 μm.

Short-Wave Infrared (SWIR): 1.1 to 2.5 μm.

Thermography: A term for infrared imaging and analysis.

Thermopile: A detector comprised of a number of thermocouples connected in series.

Works Cited

Accetta, J. S., Shumaker, D. L. (Eds.) (1993), *Infrared and Electro-Optical Systems Handbook,* Ann Arbor, MI: Environmental Research Institute of Michigan (ERIM), and Bellingham, WA: SPIE Press.

Beichman, C. A. *et al.* (1988), *Infrared Astronomical Satellite (IRAS),* NASA RP-1190, Vol. 1.

Dereniak, E. L., Crowe, D. G. (1984), *Optical Radiation Detectors,* New York: Wiley.

Hudson, R. D. Jr. (1969), *Infrared System Engineering,* New York: Wiley.

Kaplan, H. (1993), *Practical Applications of Infrared Thermal Sensing and Imaging Equipment,* Bellingham, WA: SPIE Press.

Lillesand, T. M., Kiefer, R. W. (1994), *Remote Sensing and Image Interpretation,* New York: Wiley.

McLean, I. S. (Ed.) (1994), *Infrared Astronomy with Arrays,* Boston: Kluwer Academic Publishers.

Measures, R. M. (1985), *Laser Remote Sensing,* New York: Wiley.

Reif, F. (1965), *Fundamentals of Statistical and Thermal Physics,* New York: McGraw-Hill, p. 381.

Rogalski, A. (1994), *Infrared Physics and Technol.* **40** (1), 1-218.

Russwurrm, G. M., Childers, J. W. (1995), "Environmental Sensing," *Laser Focus World* **31** (4), 79–84.

Simpson, O. A. (Chair) (1994), *Proceedings of the Symposium on Optical Sensing for Environmental Monitoring, (Sp-89),* Pittsburgh: Air And Waste Management Association.

Wolfe, W. L., Zissis, G. J. (Eds.) (1985) *The Infrared Handbook,* Ann Arbor: Environmental Research Institute of Michigan (ERIM).

Zeilik, M., Smith, E. v. P. (1987), *Introductory Astronomy and Astrophysics,* Philadelphia: Saunders College Publishing.

Further Reading

Barnes, N. P. (1995), "Laser Remote Sensors," *Laser Focus World* **31** (4), 87–94.

Killinger, D. K., Mooradian A., *Optical and Laser Remote Sensing,* Berlin: Springer-Verlag.

Lloyd, M. (1975), *Thermal Imaging Systems,* New York: Plenum Press.

Russwurrm, G. M., Childers, J. W. (1995), "Environmental Sensing," *Laser Focus World* **31** (4), 79–84.

Seyrafi, K. (1985), *Electro-Optical Systems Design,* Los Angeles: Electro-Optical Research Company.

Spiro, I. J., Schlessinger, M. (1995), *Infrared Technology Fundamentals,* New York: Marcel Dekker.

SEPARATION PROCESSES

James Zhou, *Research and Technology, Allied Signal Inc., Des Plaines, Illinois, U.S.A.*

W. S. Winston Ho, *Corporate Research, Exxon Research and Engineering Company, Annandale, New Jersey, U.S.A.*

Norman N. Li, *Research and Technology, Allied Signal Inc., Des Plaines, Illinois, U.S.A.*

Introduction 600
1. **Distillation** 600
1.1 Introduction and Process Description 600
1.2 Distillation-Process Equipment 600
1.3 Distillation-Process Design and Operations 601
1.4 Important Industrial Applications 603
2. **Absorption** 603
2.1 Introduction and Process Description 604
2.2 Process Equipment 604
2.3 Process Design and Operations 604
2.4 Important Industrial Applications 605
3. **Adsorption** 605
3.1 Introduction and Process Description 605
3.2 Adsorption Equilibria 606
3.2.1 Langmuir Isotherm 606
3.2.2 BET Isotherm 606
3.3 Process Design and Operations 607
3.4 Important Industrial Applications 607
4. **Ion Exchange** 609
4.1 Introduction and Process Description 609
4.2 Ion-Exchange Resins 609
4.2.1 Polymeric Ion-Exchange Resins 609
4.2.2 Inorganic Ion-Exchange Materials 610
4.2.3 Other Ion-Exchange Materials 610
4.3 Ion-Exchange Operation 610
4.4 Important Industrial Applications 610
4.4.1 Water Treatment 610
4.4.2 Metals Recovery 611
4.4.3 Other Applications 611
5. **Extraction and Leaching** 611
5.1 Introduction and Process Description 611
5.2 Principles and Selection of Solvents 611
5.3 Mass-Transfer Coefficients 612
5.4 Process Equipment 612
5.5 Extraction and Leaching Operations 613
5.6 Important Industrial Applications 613
6. **Crystallization** 614
6.1 Introduction and Process Description 614
6.2 Process Equipment 614
6.3 Crystallizer Operations 616
6.4 Important Industrial Applications 617
7. **Filtration** 617
7.1 Introduction and Process Description 617
7.2 Process Equipment 618
7.3 Filtration Operations 620
7.4 Important Industrial Applications 621
8. **Sedimentation** 621
8.1 Introduction and Process Description 621
8.2 Process Equipment 621
8.2.1 Thickeners 621
8.2.2 Clarifier 622
8.3 Sedimentation Operations 623
8.4 Important Industrial Applications 624
9. **Membrane Separations** 625
9.1 Introduction and Process Description 625
9.2 Process Equipment 630
9.3 Membrane Separation Operations 632

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

9.4 Important Industrial Applications 635
9.4.1 Gas Permeation 635
9.4.2 Pervaporation 635
9.4.3 Dialysis 635
9.4.4 Electrodialysis 635
9.4.5 Reverse Osmosis 635
9.4.6 Ultrafiltration 636
9.4.7 Microfiltration 636
9.4.8 Emulsion Liquid Membrane .. 636
Glossary 636
Works Cited 638

INTRODUCTION

Separation processes are widely used for the recovery and purification of desired products in manufacturing operations in various industrial sectors including the chemical, petrochemical, petroleum, pharmaceutical, medical, food, dairy, beverage, paper, textile, metallurgical, and electronic industries, as well as for the removal of pollutants in pollution control and environmental protection. This article covers the major separation techniques employed in these processes, namely, distillation, absorption, adsorption, ion exchange, extraction and leaching, crystallization, filtration, sedimentation, and membrane separations. Included for the description of each separation technique are introduction and process description, process equipment, process operations, and important industrial applications. While in-depth description is not possible because of the scope of the article, references are given for further reading.

1. DISTILLATION

1.1 Introduction and Process Description

Distillation is the most commonly used method for separating liquid mixtures in chemical industries and the most studied and understood method (Fair, 1987; Seader, 1984). Distillation effects separation of the liquid mixtures by the relative difference of the effective vapor pressure, or volatility, of liquid components. Thus, a large volatility difference between components in a liquid mixture means that they will be easily separated via distillation. When such difference does not exist, or is very small, separation by distillation becomes impossible or costly. On the basis of the mode of operation, distillation can be classified as continuous or batch. The term fractional distillation is often used interchangeably with distillation.

The overall separation efficiency of the distillation process is determined by the relative volatility, the number of contacting stages (trays), and the liquid and vapor flow rates.

1.2 Distillation-Process Equipment

The equipment used to carry out distillation can be divided into two large categories, i.e., tray-type columns and packed-bed–type columns. The former provides a stagewise contacting, whereas the latter provides a countercurrent differential contacting between the vapor and the liquid. A tray-type distillation column is shown schematically in Fig. 1, and a packed-bed–type distillation column is shown in Fig. 2.

The feed, which is to be separated into fractions, is introduced into the column at one or more places along the column. Because of density difference, the liquid flows downward, while the vapor flows upward. Equilibrium is assumed to exist between the liquid and vapor streams on all trays. Liquid reaching the bottom is partially vaporized in the reboiler, and the vapor is pushed upward. The heavier fraction left at the bottom is withdrawn as the bottom product. The vapor reaching the top is condensed in the overhead condenser. Part of the condensate is returned to the column as reflux to provide liquid overflow, and the remainder is the top product.

The trays in tray columns generally contain some type of vapor dispersion device such as liftable valves, bubble caps, or simple perforations. The packed-type column can have either random packing or ordered packing. Figure 3 shows some of these random packings used in the industry. The choice of which packing to use depends on the efficiency of contact required for the separation and on the pressure drop from top to bottom of the column (Strigle, 1994; Fair, 1987). Ordered packing, also known as struc-

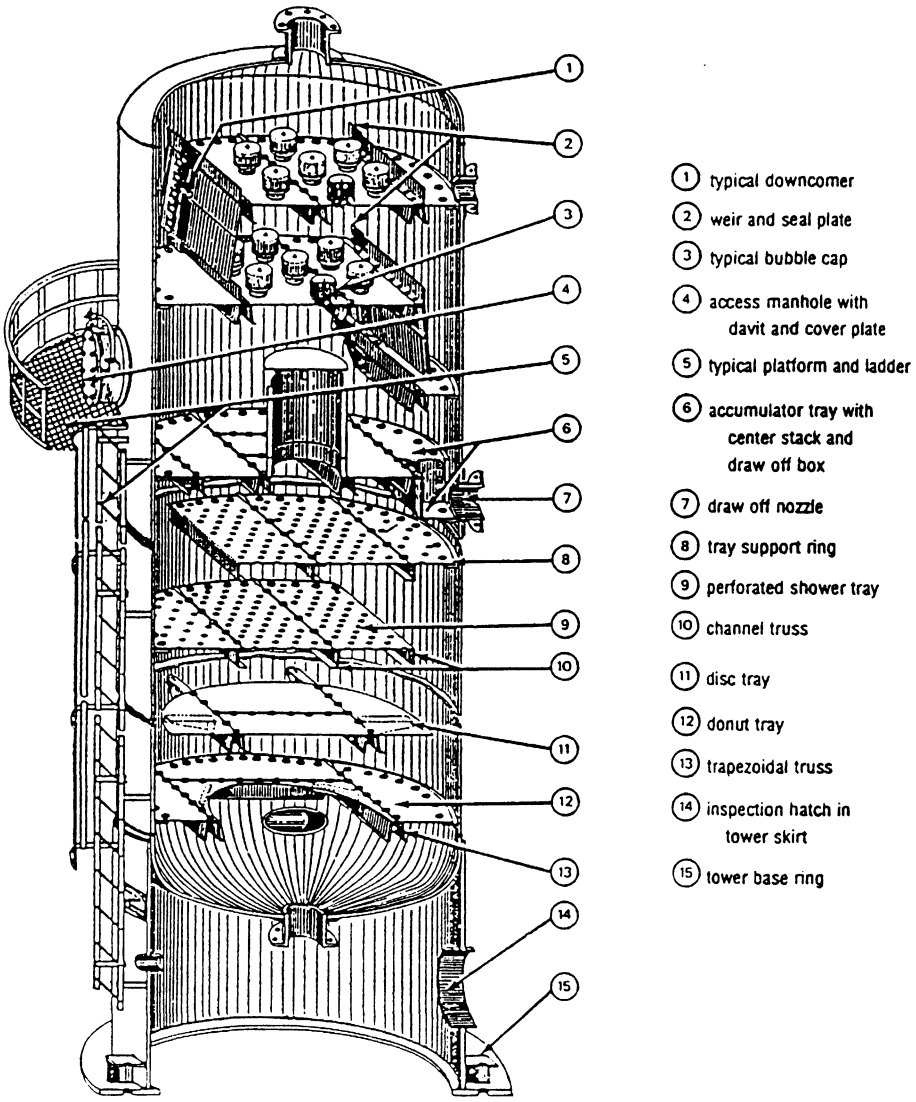
① typical downcomer
② weir and seal plate
③ typical bubble cap
④ access manhole with davit and cover plate
⑤ typical platform and ladder
⑥ accumulator tray with center stack and draw off box
⑦ draw off nozzle
⑧ tray support ring
⑨ perforated shower tray
⑩ channel truss
⑪ disc tray
⑫ donut tray
⑬ trapezoidal truss
⑭ inspection hatch in tower skirt
⑮ tower base ring

FIG. 1. Schematic diagram of a typical tray column [reprinted from Paruit (1982) with permission].

tured packing, provides high mass-transfer efficiency but also at a higher cost of manufacturing.

1.3 Distillation-Process Design and Operations

The basic design equations for the distillation process are obtained from mass and energy balances and from thermodynamic principles. For a generalized equilibrium stage shown in Fig. 4, the mass-balance equation is

$$F^n z_i^n + L^{n+1} x_i^{n+1} + V^{n-1} y_i^{n-1} = (V^n + V^{n'}) y_i^n + (L^n + L^{n'}) x_i^n. \qquad (1)$$

The summation of mole fractions must be unity:

$$\sum_i x_i^n = \sum_i y_i^n = \sum_i z_i^n = 1. \qquad (2)$$

Since each stream is also associated with an energy (enthalpy), thus, the energy balance equation of the stages is

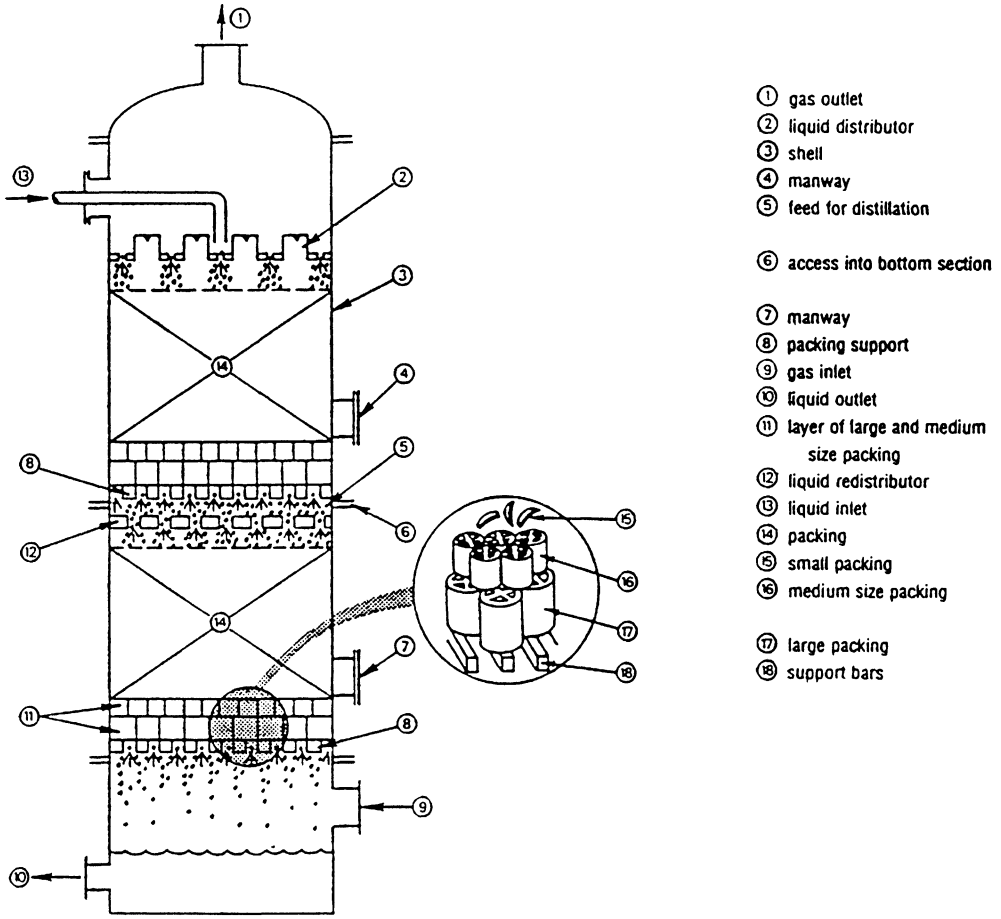
①
⑬
②
③
⑭
④
⑧
⑤
⑥
⑫
⑮
⑯
⑭
⑰
⑦
⑱
⑪
⑧
⑨
⑩
① gas outlet
② liquid distributor
③ shell
④ manway
⑤ feed for distillation
⑥ access into bottom section
⑦ manway
⑧ packing support
⑨ gas inlet
⑩ liquid outlet
⑪ layer of large and medium size packing
⑫ liquid redistributor
⑬ liquid inlet
⑭ packing
⑮ small packing
⑯ medium size packing
⑰ large packing
⑱ support bars

FIG. 2. Schematic diagram of a typical packed column [reprinted from Paruit (1982) with permission].

$$Q^n + F^n H_f^n + L^{n+1}h^{n+1} + V^{n-1}H^{n-1}$$
$$= (V^n + V^{n'})H^n + (L^n + L^{n'})h^n, \qquad (3)$$

where Q^n is the heat input into stage n, and the h's and H's are the liquid and vapor enthalpies of the streams, respectively.

The other important equation is supplied by equilibrium thermodynamics, which relates the liquid and vapor phase compositions by

$$K_j^n = y_i^n/x_i^n, \qquad (4)$$

where K_i^n is the relative volatility.

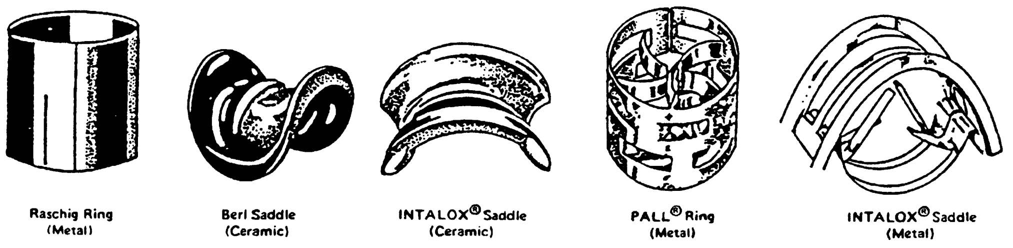
Raschig Ring (Metal)
Berl Saddle (Ceramic)
INTALOX® Saddle (Ceramic)
PALL® Ring (Metal)
INTALOX® Saddle (Metal)

FIG. 3. Some of the random packings used in packed columns [reprinted from Fair (1987) with permission].

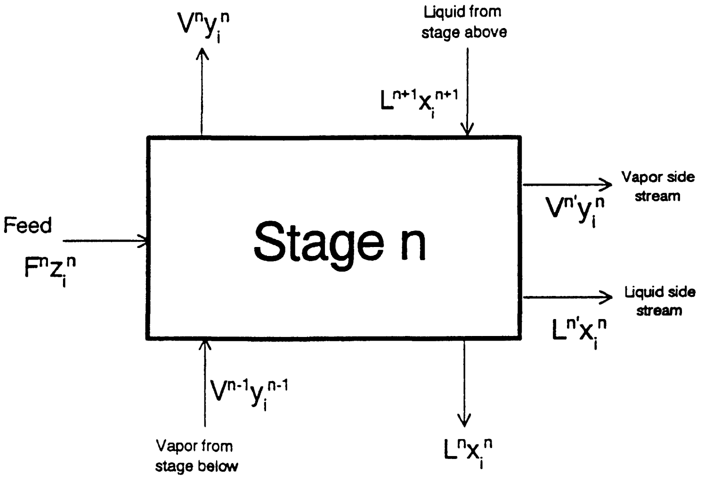

FIG. 4. Generalized equilibrium stage.

The above relations have been incorporated into many process simulation software packages available from several sources, for example, ASPEN-PLUS (1994) and ChemCAD (1993). These software packages also contain a large amount of thermodynamic data for pure substances and for mixtures. Furthermore, they can also be used to estimate the vapor–liquid equilibrium (VLE) behavior based on certain theoretical models such as UNIFAC.

For a given separation, the above relationships are used to determine the key operating parameters such as the number of theoretical stages required, the reflux ratio, the maximum operational capacity, the minimum operating liquid or vapor rate, the efficiency as compared with the theoretical value, and, most importantly, the operating pressure of the column. The choice of operating pressure can generally be made on the basis of the relative volatility between the components.

1.4 Important Industrial Applications

The single largest application of the distillation process is in oil refining where crude oil is separated into many fractions before and after catalytic reforming. The other applications are numerous in the chemical and food industries. It is safe to say that if there is a chemical plant, there must be one or more distillation columns. In addition to the application in refining, the following is a short list of some of the most important industrial applications:

1. Production of styrene monomer: vacuum distillation of styrene monomer from a mixture of ethylbenzene, water, benzene, and toluene using packed columns.
2. Ethylene glycol (EG) purification: vacuum distillation of EG using tray columns.
3. Ethanol production: azeotropic distillation of ethanol from water using benzene as an entrainer (a third component added to the system to effect the separation of an otherwise inseparable mixture by distillation).
4. Ethylene/ethane separation: pressure distillation of ethylene from ethane and propylene using packed or tray columns.
5. Cryogenic fractional distillation: a special case of separating gaseous mixtures by fractional distillation of liquid air to produce high-purity nitrogen, oxygen, and argon.

2. ABSORPTION

Absorption and adsorption can be generally termed sorption, which refers to processes of selective transfer of one or more solutes from a fluid phase to a *liquid phase* where absorption occurs, or to a batch of *solid particles* where adsorption occurs. The absorption process involves contacting a gas phase with a liquid absorbent, whereas the adsorption process consists of contacting either a gas or a liquid phase with solid particles known as adsorbent. Both of these processes have grown into essential unit operations and are widely practiced in chemical, petrochemical, and many other industries. We briefly summarize both of these processes separately in the following sections.

2.1 Introduction and Process Description

Absorption can be classified into three general types based on the interaction between the absorbent liquid and the absorbed components (absorbates) of the gas phase (Zarzycki, 1993; Kohl, 1987).

1. *Reversible reaction:* A reversible chemical reaction occurs between the absorbates and a component of the absorbent. In this case, the absorbates are taken up by the absorbent under one set of conditions and released under another set of conditions. This type of absorption is the most important in consequence of large-scale applications in natural-gas sweetening processes (removal of acid gases such as CO_2 and H_2S from light hydrocarbons) (Campbell, 1976; Kohl and Riesenfeld, 1979).
2. *Physical solution:* The absorbates simply dissolve in the liquid absorbent. Separation of components from the gas stream occurs as a result of the difference in solubility between components; the absorbate having a higher solubility preferentially dissolves in the absorbent.
3. *Irreversible reaction:* An irreversible chemical reaction occurs between the absorbates and a component of the absorbent. Since the absorbates are irreversibly captured by the absorbent, the rate of mass transfer of this type of process tends to be quite high as compared with the other two types of absorption processes.

The inverse operation, called stripping or desorption, involves the transfer of dissolved volatile components from the liquid to the gas phase. The same basic principles apply to both absorption and desorption operations.

2.2 Process Equipment

Absorption processes are similar to distillation processes; both involve a gas–liquid contacting, and both are equilibrium processes. Therefore, the process equipment used in these two processes is also similar. The three major types of equipment used in absorption are tray-type, packed, and spray columns (Kohl, 1987; Edwards, 1984). The first two types of columns are identical to those used for distillation. The spray column is unique to the absorption process.

Tray columns are well suited for clean, noncorrosive, and nonfoaming liquid absorbents, and low to medium absorbent flow rates. The advantages of tray columns are their simple design, capability of internal cooling, and, most importantly, large number of mass-transfer units due to the absence of gas channeling. The main disadvantages are that the tray columns are difficult to operate, and cannot adapt easily to changing operating conditions such as feed flow-rate change.

The packed columns have either random packing or structured packing. The random packings used are shown in Fig. 3. These packings are dumped into the column and can generally be changed easily when column operating parameters change. Packed columns are used for corrosive liquid, under foaming conditions, at high liquid-to-gas ratios, and at low pressure drops. Structured packing offers higher column efficiency but at an added cost.

Spray columns are used when solids are present in either the liquid or the gas stream and when a higher pressure drop cannot be tolerated. Spray columns offer a very limited number of mass-transfer units (theoretical number of trays) and hence are not suitable for difficult separations. The main application of spray columns is for the absorption of SO_2, H_2S, and other impurities from flue gas where low pressure drop is essential.

One other type of gas–liquid contacting device is the bubble column, where the gas is bubbled through a column filled with the liquid absorbent. This type of column has very low efficiency and is most often used as a reactor.

2.3 Process Design and Operations

Design of absorption processes includes the choice of absorbents, the choice of liquid–gas contactors, and, finally, operating conditions. The operation of the absorption is generally straightforward. The gas stream is introduced from the bottom of the tray, packed, or spray column and allowed to flow upward. The liquid stream is introduced from the top of the column and flows downward.

Selection of absorbents is primarily based on experimental data and on economic considerations. Thus, the absorbent should have a high solubility for the absorbates, and a low vapor pressure for reduced vapor loss and effluent contaminations, and to the extent possible it should be inexpensive, stable, noncorrosive, nonfoaming, and nonflammable.

The next step in the design process is the determination of the VLE behavior of the absorbent–absorbate system. The VLE represents the limiting conditions of the system, yielding the minimum liquid-to-gas ratio and the maximum gas purity. The VLE data are generally presented as the solubility of absorbates in the absorbents. When this solubility is very low, it can usually be expressed by the well-known Henry's law:

$$\begin{aligned} p_A &= Hx_A, \\ p_A &= H'c_A, \end{aligned} \tag{5}$$

where p_A is the partial pressure of component A in the gas phase; x_A and c_A are mole fraction and molar concentration, respectively, of A in the liquid phase; and H and H' are Henry's-law constants in appropriate units.

Vapor–liquid equilibrium data for many gas–liquid systems are available in *Perry's Chemical Engineers' Handbook* (Perry and Green, 1984), and in other engineering data books (Ohe, 1990, 1989; Cheremisinoff, 1995; Gas Processors Suppliers Association, 1987). When experimental data are not available, prediction can be made using several techniques, such as those discussed in Sec. 1. Comprehensive discussion on this subject is given by Reid *et al.* (1987).

Selection of the mass transfer equipment is based on the operating requirements, economics, and the advantages and disadvantages of each type of equipment, as discussed in Sec. 2.2. Once the selection is made, detailed mass and energy balances in conjunction with the mass-transfer efficiency considerations determine the number of trays needed or the height of the packing and the physical size of the column. We do not discuss the computational methods in detail here, but the readers are referred to the discussion in Secs. 14 and 18 of *Perry's Chemical Engineers' Handbook* (Perry and Green, 1984). Computer simulations of absorption processes are also available in ASPEN-PLUS (1994), ChemCAD (1993), etc.

2.4 Important Industrial Applications

Large-scale industrial applications of absorption processes include the absorption of SO_3 with sulfuric acid to make fuming sulfuric acid, and the absorption of HCl with water to produce hydrochloric acid. More recently, the use of lime/limestone slurries for the removal of SO_2 from flue gases has developed. Many other applications using absorption processes are given in the literature (Kohl, 1987).

As mentioned earlier, the largest industrial application of absorption processes is in the natural- and synthetic-gas sweetening processes, where acid gases such as H_2S and CO_2 are removed from gaseous hydrocarbons consisting mostly of methane and small amounts of ethane and higher hydrocarbons (Kohl and Riesenfeld, 1979; Campbell, 1976). This process is based on reversible chemical reactions of the acid gases with the basic liquid absorbents such as monoethanolamine, diglycolamines, diethanolamine, methyldiethanolamine, and sterically hindered amines (Sartori *et al.*, 1987), as well as mixtures of two or more of the amines. The absorbents react with CO_2 and H_2S either selectively or nonselectively under absorption conditions and release the absorbed species under desorption/stripping conditions (usually higher temperature, vacuum, or use of stripping gas). The regenerated absorbents are returned to the absorber, and the cycle continues.

3. ADSORPTION

3.1 Introduction and Process Description

Adsorption occurs when a gas or liquid stream is brought into contact with a batch of solid particles (adsorbent). Specific component(s) of the fluid stream can then be concentrated on the adsorbent surface and thus separated from the remaining components. Adsorption is heterogenous in nature and classified by the interaction between the adsorbents and adsorbates. Three types of adsorption exist based on the interactions, which are physical, chemical, and specific (Suzuki, 1990; Keller *et al.*, 1987; Vermeulen *et al.*, 1984).

1. *Physical adsorption* occurs through weak van der Waals forces between the adsorbents and the adsorbates. These forces include the dispersion force and classical electrostatic forces. Physical adsorption constitutes the most important type of adsorption because this weak interaction makes regeneration of adsorbents relatively straightforward.
2. *Chemical adsorption* involves chemical reactions between adsorbents and adsorbates, resulting in new species being formed on the adsorbent surface. The chemical bond formed is generally much stronger than the van der Waals interaction. Chemical adsorption is typically identified with monolayer coverage.
3. When adsorbates interact with specific functional groups on the adsorbent surface without chemical transformation, the resulting adsorption process is termed *specific adsorption.* The forces associated with these specific interactions range from values common to physical adsorption to the stronger forces common to chemical adsorption.

We concentrate our discussion on physical adsorption because of its importance to industrial separations. Ion-exchange processes are discussed in more detail in a later section. A related process known as chromatography is not discussed here. The interested reader is referred to Brown and Grushka (1993) for chromatographic analysis and to Ganetsos and Barker (1993) for industrial applications; see also CHROMATOGRAPHY.

The most often used process equipment for adsorption is the packed column, otherwise known as the fixed-bed adsorber. Its design and operation are governed by the adsorption equilibria. Therefore, an understanding of adsorption equilibria provides one with the tool for understanding adsorption processes.

3.2 Adsorption Equilibria

The distribution of the adsorbates in the fluid phase and in the adsorbent solid phase at thermodynamic equilibrium is described by the adsorption isotherm, which relates fluid-phase adsorbate concentration to the amount of adsorbate per unit weight of adsorbent. A wide variety of equilibrium isotherms exists. The most common two types are presented here.

3.2.1 Langmuir Isotherm This isotherm was derived by Langmuir on the basis of the assumptions that at most a monolayer of adsorbate could cover the solid surface and that the adsorption energy is constant and independent of surface coverage. For gas–solid systems, the Langmuir isotherm is expressed as

$$Q = Q_{\max} K_a P/(1 + K_a P), \tag{6}$$

where Q is the adsorbate loading (wt. of adsorbate/wt. of adsorbent), $Q_{\max}$ is the maximum loading at monolayer coverage, K_a is the adsorption equilibrium constant, and P is the partial pressure of the adsorbate in the gas phase at equilibrium. For liquid–solid systems, P is replaced by concentration C of the adsorbate in the liquid phase at equilibrium.

3.2.2 BET Isotherm The Brunauer-Emmett-Teller (Brunauer *et al.*, 1938) isotherm is based on the assumptions that the adsorption is multilayer and that the energy associated with adsorption beyond the first layer is the same as for condensation. The simplest form of the BET isotherm is

$$\frac{Q}{Q_{\text{mono}}} = \frac{KP/P^s}{[1 + (K - 1)P/P^s](1 - P/P^s)}, \tag{7}$$

where Q_{mono} is the adsorbate loading for monolayer coverage, P^s is the vapor pressure of pure adsorbate at the adsorption temperature, and K is a constant related to the energy of adsorption. For liquid systems, P^s is replaced by the saturation concentration C^s

of the pure adsorbate in the liquid phase, and P by concentration C.

The adsorption of many gas–solid and liquid–solid systems can be fitted by the Langmuir or the BET isotherm. However, these isotherms are phenomenological in nature, and their representation of the adsorption data does not constitute positive proof that the assumptions used to derive these isotherms are operative in the specific adsorption process in question.

3.3 Process Design and Operations

Keys to the design of adsorption processes are the choice of adsorbent and choice of adsorption and regeneration procedures. In general, it is desirable to use adsorbents with as small a particle size as conditions for efficient operation permit, because the rate of adsorption is controlled by the transport of adsorbate molecules on the surface or within the pores of adsorbents. The pressure drop through a packed column of adsorbent particles is usually inversely proportional to the particle size. Therefore, tradeoffs must be made sometimes between the mass-transfer rate and pressure drop. To avoid such tradeoffs, compound-size columns packed with two different particle sizes, usually arranged in series, are often employed.

The adsorbents of commercial importance are zeolite molecular sieves (ZMSs), activated alumina, silica gel, and activated carbon. The ZMS is crystalline, and the adsorption takes place within the pores formed by the crystal lattice. The ZMS differentiates species by size, shape, and polarity. Activated alumina and silica gel adsorbents are amorphous Al_2O_3 and SiO_2, respectively. Both types of adsorbent are used extensively for drying gases and liquids. Activated carbon adsorbents are microcrystalline, nongraphitic carbon with internal pores and large surface area ranging from 300 to 2500 m^2/g. Activated carbon adsorbents are most often used for removing organic compounds from liquid and gas streams.

Adsorption processes most often used in industrial applications are pressure-swing adsorption for gas purification, temperature-swing adsorption for gas or liquid purification, and inert-purge adsorption for gas or liquid purification. These three different processes are shown schematically in Fig. 5.

In the *pressure-swing adsorption* (PSA) cycle, a stream containing a small amount of an adsorbate at partial pressure P_1 is passed through the adsorbent column. The equilibrium loading X_1 is expressed in units of weight or moles of adsorbate per unit weight or volume of adsorbent. After equilibrium is reached, more feed is passed through the column (usually from the opposite direction) at an adsorbate partial pressure P_2. Desorption occurs and a new equilibrium loading, X_2 is established. The difference between X_1 and X_2 is known as the delta loading that determines the maximum removal capacity of the adsorbent column. After desorption, new feed is introduced at P_1 and the cycle continues. The time required to load, depressurize, regenerate, and repressurize a packed column is usually from a few seconds to a few minutes.

The *temperature-swing* cycle uses temperature to lower the equilibrium partial pressure of the adsorbate, thus regenerating the column. The time required to heat, desorb, and cool the adsorbent column for regeneration can range from a few hours to a few days. The long regeneration time makes the temperature-swing cycle unsuitable for removing high concentrations of adsorbates.

The *inert-purge* cycle uses an inert (nonadsorbing) fluid containing little or no adsorbate to lower the partial pressure or concentration of the adsorbate in the column. If enough inert fluid is passed through the column, the adsorbent can be depleted of adsorbate and completely regenerated. Thus, the equilibrium loading is X_1 at pressure P_1.

The selection of adsorption processes for a given task is often based on process economics as compared to various schemes of adsorption processes and competing processes such as distillation, absorption, and membrane separation.

3.4 Important Industrial Applications

Adsorption process is widely used in gas and liquid purification where small amounts of impurities are removed by adsorbents from feed streams. Drying of gas and liquid streams using adsorbents (ZMS, activated alumina, silica gel) can be found in almost all sectors of chemical, petrochemical, petroleum, natural gas, and pharmaceutical indus-

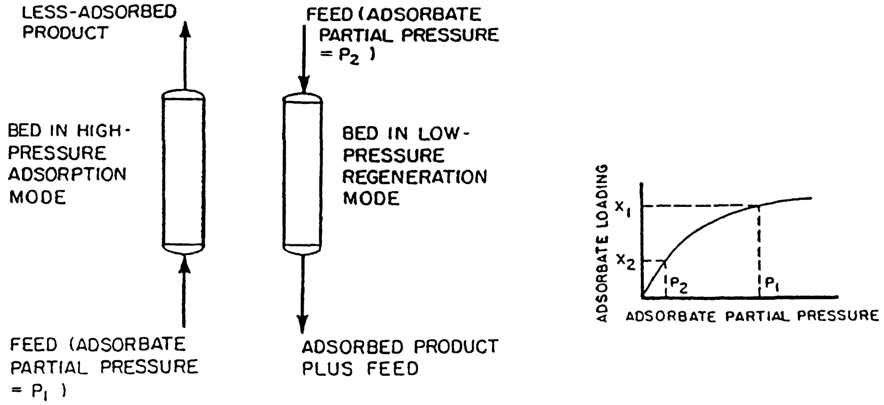
LESS-ADSORBED PRODUCT
FEED (ADSORBATE PARTIAL PRESSURE = P2)
BED IN HIGH-PRESSURE ADSORPTION MODE
BED IN LOW-PRESSURE REGENERATION MODE
FEED (ADSORBATE PARTIAL PRESSURE = P1)
ADSORBED PRODUCT PLUS FEED
ADSORBATE LOADING
X1
X2
P2
P1
ADSORBATE PARTIAL PRESSURE

Pressure-swing adsorption cycle

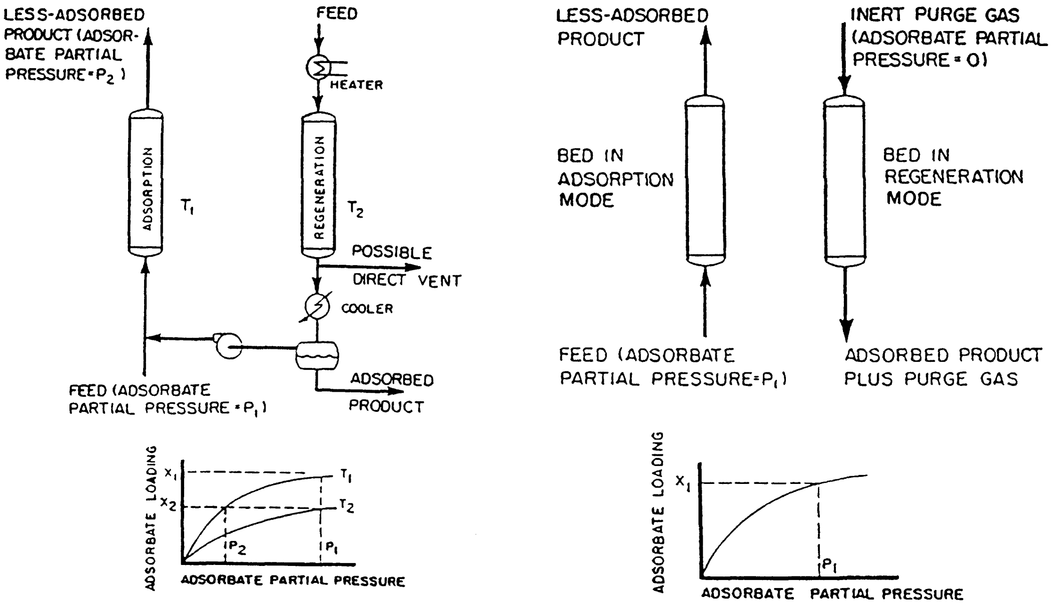
LESS-ADSORBED PRODUCT (ADSORBATE PARTIAL PRESSURE=P2)
FEED
HEATER
ADSORPTION
T1
REGENERATION
T2
POSSIBLE DIRECT VENT
COOLER
FEED (ADSORBATE PARTIAL PRESSURE=P1)
ADSORBED PRODUCT
ADSORBATE LOADING
X1
X2
T1
T2
P2
P1
ADSORBATE PARTIAL PRESSURE
LESS-ADSORBED PRODUCT
INERT PURGE GAS (ADSORBATE PARTIAL PRESSURE=0)
BED IN ADSORPTION MODE
BED IN REGENERATION MODE
FEED (ADSORBATE PARTIAL PRESSURE=P1)
ADSORBED PRODUCT PLUS PURGE GAS
ADSORBATE LOADING
X1
P1
ADSORBATE PARTIAL PRESSURE

Temperature-swing adsorption cycle

Inert-purge adsorption cycle

FIG. 5. Typical adsorption process schemes [reprinted from Keller *et al.* (1987) with permission].

tries. Other large-scale applications include removal of volatile organic compounds (VOC) from air streams, removing SO_2 from effluent gases, and removing color and odor from water and organic liquid streams (Noll *et al.*, 1991).

A large-scale PSA process is used for O_2 and N_2 production using ZMS or using a type of adsorbent known as carbon molecular sieve. Examples of liquid bulk separation using adsorbents include separation of fructose and glucose, normal paraffin from aromatics and isoparaffins, and *p*-xylene from other C_8 aromatics (Ruthven, 1984).

4. ION EXCHANGE

4.1 Introduction and Process Description

Ion exchange is intimately related to adsorption in that both involve a fluid phase and a solid particulate adsorbent phase and both use the same types of process equipment (Streat and Cloete, 1987; Vermeulen *et al.*, 1984). Moreover, the basic fundamentals involved in liquid adsorption and ion exchange are similar. The adsorbents used in the ion-exchange process are often referred to as ion-exchange resins because these adsorbents are often made of cross-linked polymeric materials, such as polystyrene cross-linked with divinyl benzene (DVB) and polyacrylic acid cross-linked with DVB.

The processes used are batch process, fixed-bed (column) process, continuous process, and fluidized-bed (slurry-bed) process (Dorfner, 1991; Streat and Cloete, 1987). In all these processes, the electrolyte solution is brought into contact with the ion exchanger, and certain ions in the solution are replaced by ions from the ion-exchange material. Similarly, all processes involve not only an ion take-up step (also known as a service step) but also a regeneration step.

On account of this similarity between adsorption and ion exchange, we do not discuss the process equipment used for ion exchange but concentrate our discussion on ion-exchange resins and on applications of ion-exchange processes.

4.2 Ion-Exchange Resins

Ion-exchange resins are naturally divided into two categories, cation- and anion-exchange resins. These resins are microporous and have either cations or anions chemically bonded to the resin backbone. For cation-exchange resin, the anions are bound to the solids. For anion-exchange resins, the cations are bound to the solids. Since electric neutrality is always preserved, ions having the opposite charge to the fixed ions, known as counter ions, must be present in sufficient quantity. These counter ions are free to move in and out of the micropores of the ion-exchange resin either by diffusion or under the influence of an electrical field. They can be replaced by other ions of the same charge from solution in the ion-exchange process.

The concentration of fixed charges on the ion-exchange resin defines the ion-exchange capacity, which is often measured in milliequivalents (meq.) of charge per unit weight of dry resin. The ion-exchange process is always stoichiometric to preserve electroneutrality. The ion-exchange capacity is independent of the nature of the counter ions; e.g., a divalent counter ion will replace two univalent ions. The rate of the ion-exchange process is governed by the diffusion rate of the counter ions in the micropores and in the stagnant layer adjacent to the ion-exchange resin.

4.2.1 Polymeric Ion-Exchange Resins Polymeric ion-exchange resins are widely used in the industry. These resins essentially come from either cross-linked polystyrene or cross-linked polyacrylic acid with the desired functional groups attached to the polymer backbone for ion exchange.

Among cation-exchange resins, polystyrene sulfonic acid (PSSA) is by far the most common cation-exchange resin. It is made by sulfonating DVB cross-linked polystyrene (PS) in chlorosulfuric or concentrated sulfuric acid at about 80 °C. The sulfonic acid group may be placed in one of three positions on the benzene ring of PS, the most probable being meta or para. The resulting resin acidity depends on the position of the sulfonic acid group and on the matrix structure of the resin. The total capacity of the PSSA resin is about 4.0 to 5.5 meq/dry gram. Polystyrene sulfonic acid is a strong acid ion-exchange resin.

Polyacrylic acid cation-exchange resin is the second most commonly used resin. It is made by hydrolysis of cross-linked polyacrylates or polyacrylonitrile. The polyacrylic acid resin is characterized by higher capacity (about 8.0 to 13.0 meq/dry gram).

Anion-exchange resins are generally based on an amine or ammonium group. Strong base anion-exchange resins use quaternary ammonium compounds, whereas weak base products employ tertiary amine groups. The backbone polymers used to make the anion-exchange resins are exactly the same as those used for the cation exchangers. The PS polymer is made into the anion-exchange resin in two steps: cloromethylation and subsequent addition reaction with ammonium or primary, secondary, or tertiary amine. The common strong base anion-exchange resins are made from trimethylamine or diethanolamine. The available amine groups are numerous; therefore, a large number of anion-exchange resins can be made to suit the special requirements.

4.2.2 Inorganic Ion-Exchange Materials Inorganic ion-exchange materials differ from polymeric ion-exchange resins in that they may have extremely high selectivity for one ion or a group of similar ions, that they may be used at temperatures much higher than the operating temperature limit of the polymeric resins, and that they are stable under high radiation dose (Dorfner, 1991). The most commonly used inorganic ion-exchange materials are natural and synthetic zeolite; hydrous oxides—for example, zirconia and stannic oxide—; insoluble salts—for example, molyphosphate and zirconium phosphate—; ferrocyanides; and clays and pillared clays.

4.2.3 Other Ion-Exchange Materials There are other types of ion-exchange materials that do not fit into the above-mentioned two large categories—for example, liquid ion-exchange materials and ion-exchange membranes (Dorfner, 1991).

Liquid ion exchangers are formed by attaching ion-exchange groups to oligomers, uncross-linked polymers, or organic molecules. These materials are most often in the liquid state and insoluble in aqueous solutions but soluble in organic solvents. Liquid ion exchangers can be used where solid ion exchangers can be used. The largest industrial application for liquid ion exchangers is in the production of citric acid (Code of Federal Regulations, 1977).

4.3 Ion-Exchange Operation

As mentioned earlier, the ion-exchange operation can be divided into two steps: the exchange step and the regeneration step. There are four basic techniques used for the ion-exchange process, i.e., batch, fixed bed, continuous, and fluidized bed.

Batch operation involves contacting the ion-exchange material with the electrolyte solution in any suitable vessel with agitation until equilibrium. The ion exchanger is then separated from the solution by filtration, centrifugation, or settling. The separated ion exchanger is regenerated by mixing it with a suitable regeneration chemical.

The fixed-bed or column process involves packing the solid ion exchanger into a column and flowing the electrolyte solution in a downward or upward manner. The ions that are to be exchanged always contact the fresh ion exchanger until the ion-exchange capacity of the column is fully utilized and breakthrough occurs. At that point, the column is washed free of the feed electrolyte mixture with water, and elution and regeneration take place subsequently.

The continuous ion-exchange process employs either fixed-bed or fluidized-bed operations. When fixed-bed operation is used, a bank of ion-exchange beds is used, and each bed in the bank is in a different process step at any given time because feed solution to the bed changes with time.

The fluidized-bed process is similar to the batch process, except that the exchanger material is continuously cycled in and out of the main process step to be washed and regenerated.

4.4 Important Industrial Applications

Ion-exchange processes find their largest applications in water treatment, i.e., water deionization and water softening. More recently, ion-exchange processes have also been used for pollution control. Other important applications include sugar processing, noble-metal recovery from spent catalysts, leach solutions, uranium recovery from low-grade mineral leach solutions, and purification and recovery of pharmaceutical products.

4.4.1 Water Treatment Water deionization basically removes cation and anion species such as Ca^{2+}, Mg^{2+}, K^+, Na^+, and other metal ions using H^+ from a cation-exchange resin and Cl^-, SO_4^{2-}, NO_3^-, etc., using OH^- from an anion-exchange resin. The process generally uses a fixed bed of intimately mixed cation- and anion-exchange resins in the hydrogen and hydroxide forms, respectively. Ultrapure water produced by this process is used in electronic, chemical, and pharmaceutical industries.

Water-softening processes essentially remove Ca^{2+} and Mg^{2+} ions from water using Na^+ from cation-exchange resins. Most cation-exchange resins can be used for this process because the selectivity for divalent ions is much higher. Regeneration of the resin is accomplished using concentrated sodium chloride solution.

Water desalination using ion-exchange membranes in an electrodialysis process is also practiced industrially in large scale.

4.4.2 Metals Recovery Ion-exchange processes have been used to recover various metals from dilute solutions. These metals include noble metals such as gold, silver, platinum, and palladium. Uranium ore upgrade from low-grade ore is practiced on a large scale using an ion-exchange process in conjunction with a liquid–liquid extraction process. In general, the ion-exchange process is highly suited for recovering high-value products from dilute streams.

4.4.3 Other Applications Ion-exchange processes are routinely used in effluent treatment to remove and recover pollutants. Typical examples include chromium removal from an electroplating effluent, ammonia and nitrate removal from groundwater, and cesium and strontium removal from radioactive wastewater. Ion-exchange processes are also used in sugar processing plants to soften sugar juices by removing scale-forming elements such as Ca^{2+} before concentration, to decolorize, to catalyze inversion of sucrose to fructose and glucose, and finally to separate fructose from glucose. Ion-exchange processes are also used extensively in the pharmaceutical industry for separation and analysis.

5. EXTRACTION AND LEACHING

5.1 Introduction and Process Description

The extraction and leaching processes are among the most utilized ones for separating one or more constituents from a mixture. This separation is based on the solubility difference of the components in a second phase formed by the extraction or leaching solvents (Skelland and Tedder, 1987; Robbins, 1984). The term extraction generally refers to the separation of components (solutes) from a liquid mixture by a second liquid (solvent) phase immiscible or only partially miscible with the original liquid mixture, whereas leaching refers to the separation from a solid mixture.

The extraction and leaching processes can be classified into physical and chemical processes. The physical process is based on difference in solubility as expressed by the partition coefficient. The chemical process depends on chemical changes or chemical interaction between the solute and the extraction solvent to increase the solute solubility in the solvent. The solute extracted can be separated from the solvent by distillation, evaporation, or back-extraction (stripping) with a different solvent. This allows the solvent to be recycled to the extraction/leaching process. Industrial success of extraction/leaching hinges on the economics of solvent and solute recovery.

5.2 Principles and Selection of Solvents

Extraction and leaching are also equilibrium-based processes. The thermodynamic principles suggest that in order for phases to be in equilibrium, the chemical potential of each species in all phases must be uniform. The partition coefficient m of a solute in the two liquid phases at equilibrium can be expressed as

$$m = y_A/x_A, \tag{8}$$

where y_A and x_A are the molar (or mass) fractions of species A in the two liquid phases.

Liquid–liquid extraction in practice rarely involves only three components. For such systems, if the two liquid phases are immiscible, the linear distribution law can be applied to solutes 1, 2, . . . i, separately:

$$m_i = y_i/x_i. \tag{9}$$

The ease of separation is represented by the selectivity or separation factor α, defined as the ratio of the partition coefficients:

$$\alpha_{AB} = m_A/m_B = y_A x_B/x_A y_B. \tag{10}$$

The chemical extraction process depends on a chemical change to effect mass transfer from the mixture to the chemical extractant. The transferring species can associate, dissociate, or react with the solvent. However, chemical and physical extraction processes have in common the need to maintain a high interfacial area for ultimate contact between the two phases and a quick and clean phase separation after the transfer of species has been completed. Chemical extraction finds many applications, especially in the separation of metals (Chapman, 1987). The metal ion of interest is converted selectively to an organometallic compound or complex soluble in an organic carrier solvent.

The selection of solvent depends on equilibrium considerations, physical properties, cost, ease of recovery, and safety and environmental concerns. Solvent selectivity is determined by physical and chemical equilibria and is the determining factor to the purity of the recovered products. A higher selectivity can reduce processing costs substantially by reducing solvent and solute recovery costs.

Physical properties such as viscosity, density, boiling point, and surface tension determine the extent of phase dispersion and separation, the solvent and solute diffusion rate, the wetting of solid surfaces, and recovery conditions. In chemical extraction and leaching, reaction conditions such as *p*H and redox potential must also be considered.

5.3 Mass-Transfer Coefficients

Equilibrium considerations define the ultimate extent of separation. However, the rate of the processes is determined by the mass-transfer considerations (Rydberg *et al.*, 1992; Skelland and Tedder, 1987). Accordingly, the mass-transfer rate depends on the interfacial contact area. For both liquid–liquid extraction and liquid–solid extraction, the mass flux at the interface is given via Fick's first law by the following equation derived on the basis of the film model:

$$J = D^*(C_s - C)/\delta, \tag{11}$$

where D^* is a diffusion coefficient; C_s and C are concentrations of solute at the interface and in the bulk liquid, respectively; and δ is the effective thickness of the liquid film at the interface.

The mass-transfer rate based on molecular diffusion through the interfacial film can be expressed phenomenologically by mass-transfer coefficients k_x and k_y and the concentration differences for each phase:

$$J = k_x(x_A - x'_A) = k_y(y'_A - y_A), \tag{12}$$

where x_A and y_A are the bulk concentrations of species A and x'_A and y'_A are concentrations of A adjacent to the interface. The concentration adjacent to the interface, in turn, is usually represented by the equilibrium partition relationship $y'_A = mx'_A$. Thus, an overall mass-transfer coefficient K_A can be defined and the mass flux expressed as

$$J = K_x(x_A - x^*_A), \tag{13}$$

where

$$\frac{1}{K_x} = \frac{1}{k_x} + \frac{1}{mk_y}; \tag{14}$$

x^*_A is the concentration of A at equilibrium with y_A. Equation (14) defines the overall mass-transfer coefficient and is known as the resistance additivity rule; i.e., the overall resistance to mass transfer is the sum of the resistances from the two phases. Furthermore, K_x is also a function of the partition coefficient m, which could be large depending on the nature of the solvent.

The rate of mass transfer depends strongly on agitation and temperature, and on the degree of mobility of the droplet surface. When two immiscible liquids are mixed together by agitation, a dispersion of droplets is created to increase the contact area. Increasing the degree of agitation results in smaller droplets, hence larger contact area, and higher mass-transfer rate.

5.4 Process Equipment

Continuous countercurrent extraction is usually carried out in a vertical column. The contactor is designed to disperse one phase into droplets and achieve the desired mass transfer in a given number of stages. The most commonly used contactor devices are mixer-settler, packed column, perforated-plate column, and various rotary agitated columns (Robbins, 1984; Skelland and Tedder, 1987).

The various packed and perforated-plate columns are similar to the ones discussed in previous sections, for distillation, absorption, adsorption, and ion exchange. For the packed column, the packing may be different for liquid–liquid extraction than for other purposes discussed earlier.

In addition, there are several types of mechanically agitated columns. The agitation can be provided by a centrifugal device, a stator-rotor type design, or a reciprocating-plate design. For example, the rotating-disk contactor (RDC) and the Karr reciprocating-plate column (RPC) are shown in Fig. 6.

Leaching is carried out in equipment designed to handle solid particles. The often used types of equipment are moving-bed percolation extractor and rotary-tray extractor (Wadsworth, 1987; Schwartzberg, 1987; Miller, 1984).

5.5 Extraction and Leaching Operations

There are three basic methods in extraction and leaching processes for contacting solvent(s) with the feed: concurrent, crosscurrent, and countercurrent.

Concurrent stagewise contact is not usually necessary because equilibrium is established or approached in stage 1. Therefore, a single-stage contactor is generally used, and the two phases are mixed to establish equilibrium and allowed to settle, after which the extracted phase containing solvent plus extracted solute and the raffinate phase composed of feed depleted of solute are separated. Single-stage leaching only differs slightly from single-stage liquid–liquid extraction in its operation. The extent of separation by a single-stage contactor is limited by equilibrium attainable under the operating conditions. To enhance the degree of separation, multistage contacting in a crosscurrent or countercurrent manner becomes necessary.

FIG. 6. Two types of mechanically agitated columns.

The crosscurrent arrangement in which fresh solvent is added at each stage enhances to a certain extent the separation obtainable for a given solvent-to-feed ratio over the single-stage arrangement. To increase further the degree of separation, the solvent-to-feed ratio has to be increased.

Countercurrent extraction represents the most efficient process for a given solvent-to-feed ratio. The feed and solvent flow countercurrently and emerge at opposite ends of the contacting stages (cascade). The maximum concentration driving force is maintained during the course of countercurrent flow. The feed having the highest solute concentration, entering stage 1, contacts solvent that has already passed through all other stages, whereas the raffinate leaving the last stage has been contacted with fresh solvent. Because of the economic advantages, continuous countercurrent operation is the choice for commercial-scale operations whenever possible.

5.6 Important Industrial Applications

Liquid–liquid extraction finds its largest applications in the petroleum industry for the separation of aromatic and aliphatic hydrocarbons, for removing aromatics from lubricating oils and jet fuels, and for desulfurization of crude oil. It is also used in large-scale production of anhydrous acetic acid and for polymer-grade ϵ-caprolactam.

Liquid–liquid extraction is also used as a major separation process in the pharmaceutical industry for the recovery of antibiotics and extraction of vitamins from natural products. For example, penicillin is extracted from fermentation broth using a batch extraction process and amyl acetate or *n*-butyl acetate as a solvent at low *p*H's from 2 to 2.5 and then back-extracted using an aqueous buffer solution at a *p*H of about 7 to 7.5.

The extraction of metals is also a large area of application for solvent extraction. Nuclear fuel processing is almost exclusively carried out using solvent extraction. For example, uranium is obtained from ore leach liquors by extraction with tributyl phosphate. Spent nuclear fuel is also reprocessed using the same process.

Leaching finds its largest application in food- and metal-processing industries—for example, the extraction of coffee from ground roasted coffee beans, oil from oil seeds, and natural flavor and color from nuts, leaves, and flowers, and in the metal-processing industry, the mining of gold by leaching with sodium cyanide solutions.

6. CRYSTALLIZATION

6.1 Introduction and Process Description

Crystallization is the process of producing crystals in the conversion of a chemical species from an amorphous solid, liquid, or gaseous state to the crystalline state. Crystallization is important as a thermal separation process because a large number of highly purified species, which are and can be marketed in the form of crystals, can be obtained from relatively impure solutions, melts, or the gaseous phase in a single processing step. Crystallization from a solution involving a solvent is called solution crystallization, whereas crystallization from a melt in the absence of a solvent is termed melt crystallization. Solution crystallization involves most industrial applications of the crystallization operation, and it is the subject of this section. For melt crystallization, the reader is referred to the literature (Moyers, 1984; Toyokura *et al.*, 1994).

Solution crystallization involves two distinct steps: crystal formation (nucleation) and crystal growth. The crystals must first form and then grow. The formation of a new solid phase in the bulk of a fluid phase without the involvement of a solid–fluid interface is called homogeneous nucleation; in the presence of surfaces, other than that of the crystal itself, such as those of the container, agitator, or foreign particles, is called heterogeneous nucleation; and in the presence of crystals of the crystallizing species itself is termed secondary nucleation. Crystal growth is the increase in size of the nucleus with a layer-by-layer addition of solute.

Both nucleation and crystal growth require that the solution be supersaturated. Supersaturation is a measure of the quantity of solute present in the solution compared with the quantity that would be present if the solution were kept for a very long period of time with the solid phase of the solute in contact with the solution. The latter quantity is the equilibrium solubility at the temperature and pressure under consideration. In general, supersaturation is expressed as a coefficient:

$$S = \frac{\text{parts solute/100 parts solvent}}{\text{parts solute at equilibrium/100 parts solvent}} \geq 1.0. \qquad (15)$$

Generally, the rate of nucleation is directly proportional to the supersaturation coefficient, and it has been modeled (Randolph and Larson, 1971; Singh, 1979; Bennett, 1984; Mersmann, 1944b). Crystal growth involves two steps:

1. diffusion of the solute to the solution-crystal interface and
2. surface reaction to absorb the solute into the crystal lattice.

The crystal growth rate has also been modeled (Singh, 1979; Bennett, 1984; Mersmann, 1994b).

Supersaturation can be generated by the following five methods (Singh, 1979; Mersmann, 1994a):

1. evaporation: by evaporating the solvent from the solution;
2. cooling: by cooling the solution via heat exchange;
3. vacuum cooling: by vacuum flashing the solution adiabatically to a lower temperature, leading to inducing crystallization by simultaneous cooling of the solution and evaporation of the solvent;
4. reaction: by chemical reaction (e.g., molecular-sieve zeolites); and
5. salting out: by adding a third species to change the solubility relationship.

6.2 Process Equipment

According to the means of suspending the growing crystals, four basic types of crystallizers (Singh, 1979; Bennett, 1984; Mersmann, 1994a) may be classified:

1. Circulating magna crystallizers: All growing crystals are circulated through the zone of the crystallizer where supersaturation is generated. Circulating magna (slurry) crystallizers are the most important crystallizers in use today, and they include the forced-circulation (FC) and draft-tube-baffle (DTB) types. In the FC crystallizer shown in Fig. 7, the feed solution enters the circulating pipe below the product discharge at a point sufficiently below the free liquid surface to prevent flushing. The slurry of crystals together with the feed is circulated through a heat exchanger to raise its temperature by about 2 to 6 °C. The heated slurry enters the crystallizer slightly below the liquid surface to cause boiling. The vaporization of the solvent and the consequent cooling result in the supersaturation needed for nucleation and crystal growth. The DTB crystallizer is shown in Fig. 8. It has an inner baffle forming a partitioned settling area in a closed vessel. Inside the settling area is a top-entering agitator located near the bottom. The agitator operates at low speeds and gives axial flow. The draft tube prevents body swirl and minimizes turbulence in the circulating slurry.
2. Circulating-liquor crystallizers: In this type of crystallizer, only the liquor is circulated while the crystals are retained in the suspension vessel by the upflowing solution. By changing the liquor flow rate, the output and crystal-size distribution are controlled.
3. Scraped-surface crystallizers: The supersaturation in this type of crystallizer is generated by direct-contact heat exchange between the slurry and a jacket or double wall containing a cooling medium. The cooling medium circulates through the shell side, while an internal agitator fitted

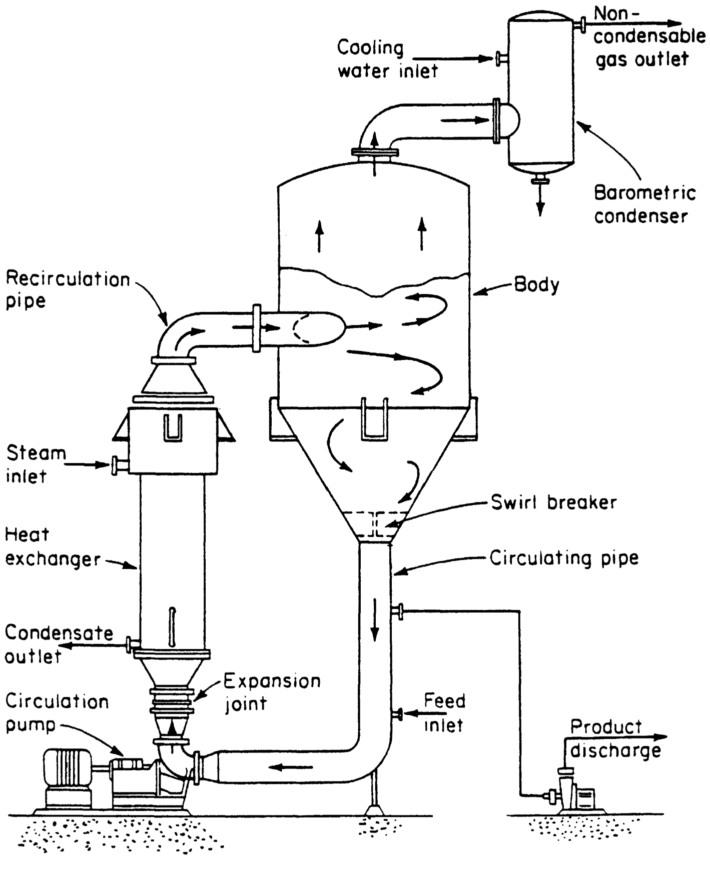
Cooling water inlet
Non-condensable gas outlet
Barometric condenser
Recirculation pipe
Body
Steam inlet
Heat exchanger
Swirl breaker
Circulating pipe
Condensate outlet
Expansion joint
Circulation pump
Feed inlet
Product discharge

FIG. 7. Forced-circulation crystallizer [Swenson Division, Whiting Corporation, reprinted from Bennett (1984) with permission].

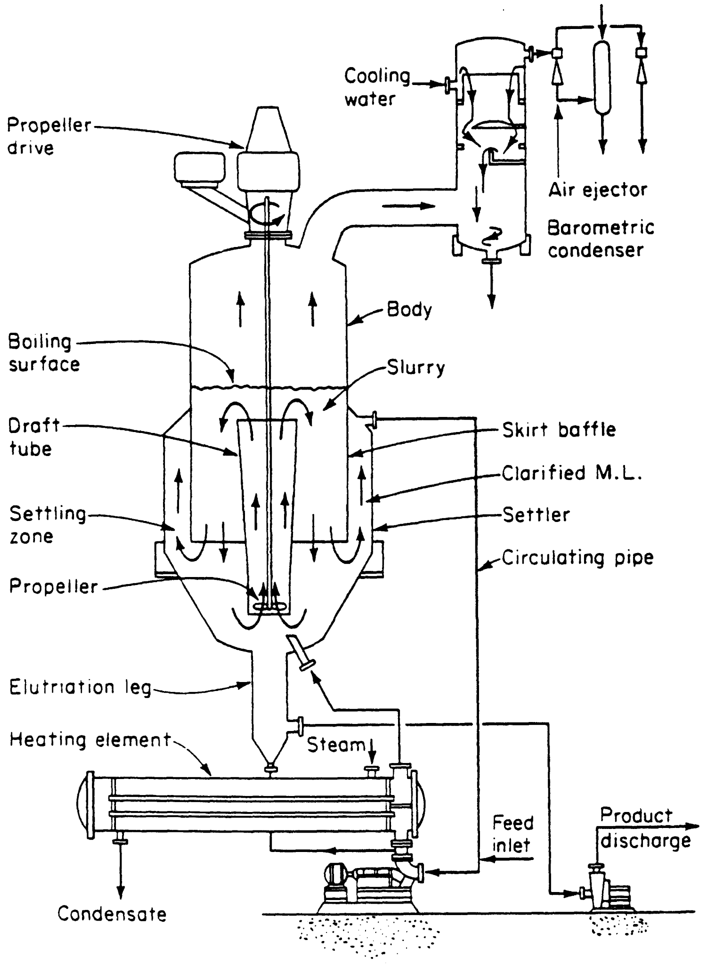
Cooling water
Propeller drive
Air ejector
Barometric condenser
Body
Boiling surface
Slurry
Draft tube
Skirt baffle
Clarified M.L.
Settling zone
Settler
Circulating pipe
Propeller
Elutriation leg
Heating element
Steam
Feed inlet
Product discharge
Condensate

FIG. 8. Draft-tube-baffle crystallizer [Swenson Division, Whiting Corporation, reprinted from Bennett (1984) with permission].

with spring-loaded scrapers scrapes the heat-transfer surface to prevent solids buildup. These crystallizers are best suited for handling high-viscosity solutions.

4. Tank crystallizers: The hot feed solution with the solute(s) of normal solubility is pumped into the crystallizer and allowed to cool for supersaturation to occur either by natural convection and radiation or by surface cooling through coils in the tank or in a jacket on the outside of the tank. This type of crystallizer is generally used for small-capacity applications including fine chemicals, pharmaceuticals, and food products in which the liquor, not the crystals, is generally the desired product.

6.3 Crystallizer Operations

The key operations are maintaining crystallizer stability, controlling crystal-size distribution, minimizing fouling, and separating solids from the liquor (Singh, 1979; Mersmann, 1994a). Steadiness of operation is much more important in crystallization by maintaining crystallizer stability than in other separation processes. Crystal growth is a layer-by-layer process, and the retention time required in most commercial crystallizers to produce crystals of the size normally desired is on the order of 2 to 6 h. On the other hand, nucleation can be generated in a fraction of a second. The influence of any upsets in operating conditions, in terms of the excess nuclei produced, is very short term in comparison with the total growth period of the crystal product. In general, four to six retention periods will pass before the effects of any upset will be damped out. Thus, the recovery period may be from 8 to 36 h.

The keys to controlling crystal-size distribution are the rate of nucleation and slurry density. The rate of nucleation required to produce crystals of a given size decreases exponentially as the crystal size increases. For example, 1 lb of 150-mesh seed crystals is enough to generate 1 ton of 14-mesh product crystals. Thus, extreme care must be exercised to prevent prenucleation in the incoming feed stream or unwanted nuclei being introduced through recycle streams. The slurry density is primarily controlled by regulating the flow rate through the crystallizer.

Fouling (Krause, 1993; Mersmann, 1994b) can be controlled to a great extent by determining the optimum slurry density and circulation rates. Variations in liquor purity affect fouling and must be controlled.

The product stream from crystallizers generally contains less than 30 wt % solids and is usually thickened to 50 wt % or greater prior to the solid–liquid separation step, for which either conical settlers or hydrocyclones can be used. In general, the final solid–liquid separation is achieved by filters or centrifuges. Filters are preferred for high-capacity applications, and a separate washing step is needed, whereas centrifuges are used for a more thorough separation of solids and liquor.

6.4 Important Industrial Applications

Important industrial applications include the manufacture of the following products:

1. Industrial chemicals: Ammonium sulfate, sodium and potassium chlorides, potassium nitrate, trisodium phosphate, sodium ferrocyanide, and sodium tetraborate are produced via crystallization. For these products, circulating-magna crystallizers are typically used for the production of large crystals. For ammonium sulfate, circulating-liquor crystallizers are sometimes employed for the production of small, uniform crystals (Singh, 1979; Mersmann, 1994a).
2. Inorganic chemicals: Gypsum, sodium nitrate, ammonium nitrate, silver nitrate, urea, sodium sulfate, inorganic chlorates, dichromates and sulfites are produced in small, uniform crystals by the use of circulating-liquor crystallizers (Singh, 1979; Mersmann, 1994a).
3. Organic compounds: Organic dyes, *p*-xylene, chlorobenzene, and organic acids (benzoic, propionic, butyric, adipic, sorbic, etc.) are processed via scraped-surface crystallizers (Singh, 1979; Mersmann 1994a).
4. Food products: Vegetable oils, sugar, and fatty acids are prepared in batch operations via tank crystallizers (Singh, 1979; Mersmann, 1994a).
5. Molecular-sieve zeolites: The zeolites are processed via circulating-liquor or tank crystallizers (Singh, 1979).

7. FILTRATION

7.1 Introduction and Process Description

Filtration is the process of removing particles from a fluid by forcing the fluid through a filtering medium (septum), on which the particles are deposited. The fluid passing through the septum is called filtrate, and it may be a liquid or a gas. The solid particles may be coarse or fine, rigid or soft (or slimy), separate individuals or aggregates. The valuable product is either the fluid, the solid, or both.

The operation of filtration depends on the mechanism for arrest and accumulation of particles, and it can be divided into two classes, deep-bed filtration and cake filtration (Jacobs, 1984; Chiang and He, 1993). Deep-bed filtration is usually preferred when the solids content of the slurry is very low (about 0.1%). A deep bed of porous media such as diatomaceous earth (precoat) or fibrous material supported on a coarse filter is used to remove fine particles from the slurry. The particles to be removed are often considerably smaller than the pores of the filter medium and can penetrate a considerable depth before they are captured. The particles can be captured via the following mechanisms: sieving, impingement, interception at solid–liquid interfaces, gravity settling, Brownian diffusion, and electrostatic forces

(Grace, 1956; Tien and Payatakes, 1979; Chiang and He, 1993).

Most liquid filtration processes operate on the mechanism of cake filtration. The filtered solids are deposited on the surface of the filter medium, and the solid cake thus deposited quickly becomes the actual filter medium.

For the relationship between pressure drop and filtration rate in cake filtration and in deep-bed filtration through a precoat filter, the Kozeny–Carman equation can be used in the following form (Donald, 1958; Chiang and He, 1993; Pinheiro and Cabral, 1993a):

$$\frac{\Delta p}{L} = \frac{k\mu u(1 - \epsilon)^2(s_p/V_p)^2}{\epsilon^3}, \tag{16}$$

where Δp is the pressure drop across the cake at thickness L, μ is the viscosity of the filtrate, u is the velocity of the filtrate based on the filter area (the volumetric flux of the filtrate), ϵ is the porosity of the cake, s_p is the surface area of a single particle, V_p is the volume of a single particle, and k is a constant. The constant k has a value of 5 for uniform, spherical particles, and it ranges between 3.5 and 5.5 for other particles. Wide particle size distributions tend to decrease k, whereas high porosity ($\epsilon > 0.8$) tends to increase it.

7.2 Process Equipment

As shown in Eq. (16), liquid flows through a filter medium via a pressure differential across the medium. The pressure differential is achieved by application of a pressure on the slurry side or a vacuum on the filtrate side. Most industrial filters are either pressure filters or vacuum filters. The filters include the following types (McCabe and Smith, 1956; Jacobs, 1984; Pinheiro and Cabral, 1993a).

1. Filter press: A filter press is a discontinuous pressure filter, and it is commonly in the plate-and-frame type shown in Fig. 9. The filter press consists of alternating plates and frames mounted on two parallel horizontal bars and clamped together. The filter cloths are stretched against the plates, which have channeled surfaces to allow filtrate drainage. The slurry material under pressure is fed into the hollow frame space, and the filtrate passing through the filter cloth drains via the channels on the plate. The cake is accumulated inside the frame until it fills the frame space. Washing the cake can be done by introducing the washing fluid through washing plates. The washing fluid flows through the entire cake thickness. The cake is then discharged by separating the plates and frames.
2. Rotary-drum filter: This is the most common type of continuous vacuum filter, shown schematically in Fig. 10 (Davis and Grant, 1992). A horizontal cylindrical drum (with a radius r_d) covered by a filtering medium (e.g., canvas cloth) turns at a constant angular velocity Ω (0.1 to 2 rpm), and it is partially submerged in an agitated slurry trough containing the suspension (slurry) to be filtered. To the interior space of the drum, a vacuum is applied. As the drum dips under the surface of the suspension (with an angle Θ_f subtended by the portion of the drum's surface that is in contact with the suspension), the vacuum sucks permeate perpendicularly through the drum surface. A layer of solids, i.e., the cake, builds up on the filter medium as the filtrate is sucked through the filter medium. As the cake leaves the suspension, it enters the washing and drying zones, in which the applied vacuum draws wash liquid and air through the cake. When the cake leaves the drying zone, the cake is discharged normally by scraping off with a doctor blade. Once the cake is removed, that part of the drum reenters the suspension and the operation cycle repeats. In addition, applying pressure on the suspension can be achieved by enclosing the filter in a shell.
3. Precoat filter: This is a modified rotary-drum filter with a layer of filter aid, e.g., diatomaceous earth, applied to the filter surface. The cake discharge mechanism is an advancing scraper knife to remove a thin layer of the precoat along with the cake in order to present a clean surface to the slurry. This filter is particularly useful for filtering small amounts of fine or gelatinous solids and deformable flocs, which normally plug a filter medium.
4. Cartridge filter: This type consists of a re-

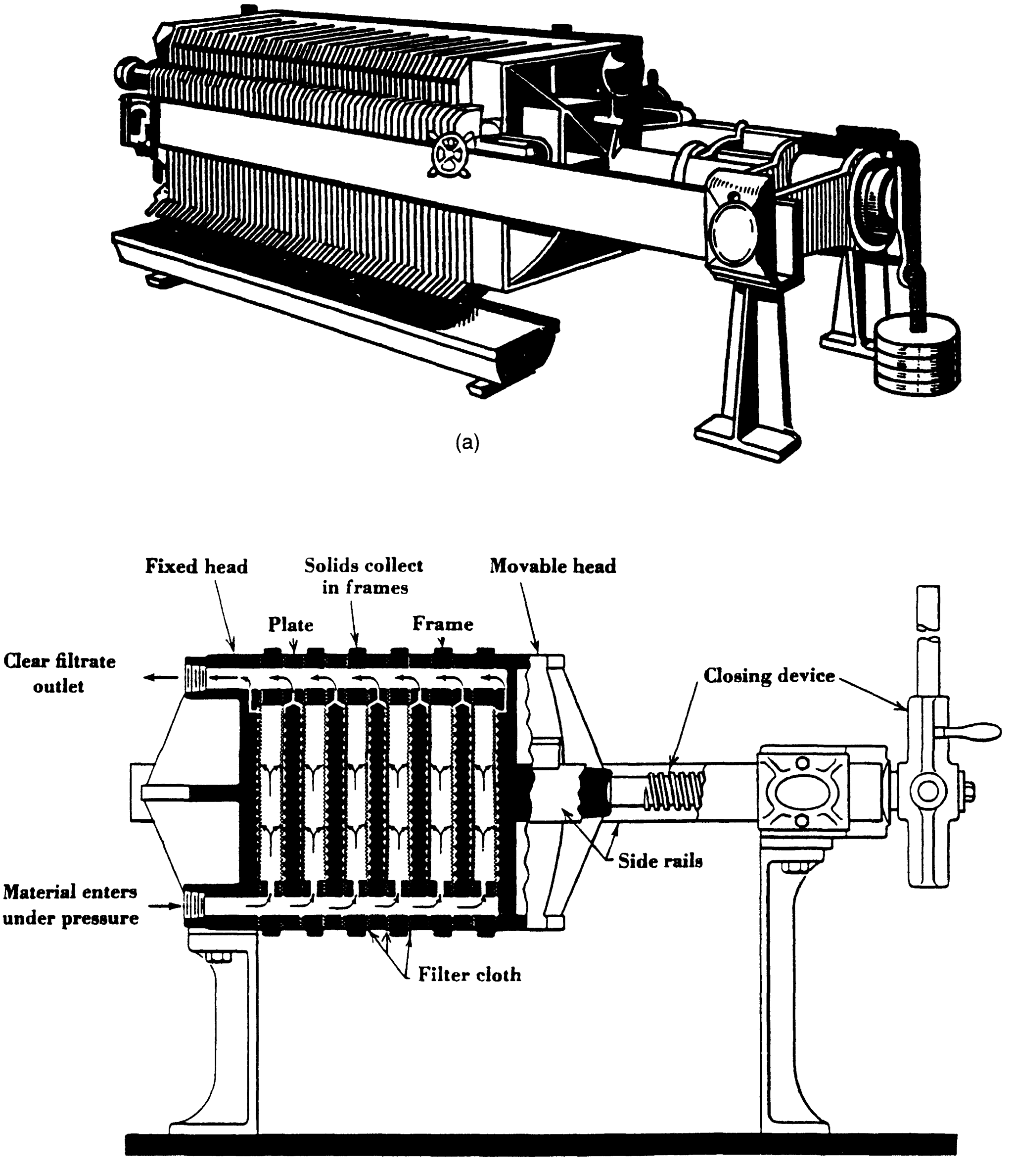
Fixed head
Solids collect in frames
Movable head
Plate
Frame
Clear filtrate outlet
Closing device
Side rails
Material enters under pressure
Filter cloth

(a)

(b)

FIG. 9. Plate-and-frame filter presses: (a) outside view [Eimco Process Equipment Co., reprinted from Jacobs (1984) with permission] and (b) inside view [T. Shriver and Co., Inc., reprinted from Bennett and Meyers (1982) with permission].

newable filter element of fiber, resin-impregnated filter paper, or porous stainless steel of controlled porosity. The filter is usually used for removing small amounts of solids from process fluids, and it is placed in a line carrying the fluid to be clarified; i.e., clarification occurs while the fluid is in transit.

In all filters, filtering media (septa) are used to effect fluid-particle separations. In industrial applications, the most common fil-

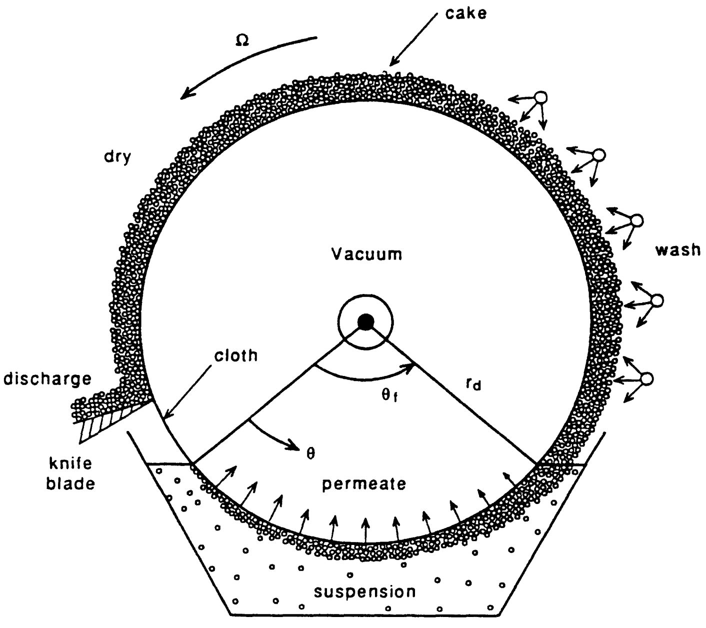

FIG. 10. Rotary-drum filter [reprinted from Davis and Grant (1992) with permission].

tering medium is canvas cloth. Other filtering media such as stainless steel cloth, monel metal cloth, glass cloth, and paper are used for corrosive fluids. Some synthetic fabrics, e.g., polyvinylidenechloride, nylon, and polyester, are also highly chemically resistant. Materials used in rigid porous media include sintered stainless steel and other metals, alumina, silica, porcelain, graphite, and some plastics. These media offer a wide range of chemical and temperature resistances, and they are mainly used for clarification.

Filter aids, e.g., diatomaceous earth, perlite (aluminum alkali silicate), and cellulosic fibers, are frequently used for filtration operations with problems of slow filtration rate and rapid medium plugging. The filter aid is mixed with the slurry before filtration, and it forms a highly permeable filter that can trap very fine or gelatinous particles and deformable flocs.

7.3 Filtration Operations

Filtration operations may be classified into two categories: batch and continuous. In batch operations, filter presses and cartridge filters are generally used, whereas in continuous operations, rotary-vacuum filters and precoat filters are commonly employed. The operations may include the following three modes:

1. constant-pressure filtration is the most frequent mode of operation, in which compressed air or vacuum is used as the driving force,
2. constant-rate filtration is usually achieved by the use of positive displacement pumps, and
3. variable-pressure, variable-rate filtration is a result of using centrifugal pumps (McCabe and Smith, 1956; Jacobs, 1984; Pinheiro and Cabral, 1993a).

As shown in Eq. (16), the rate of filtration can be improved by

1. decreasing the viscosity of filtrate: this can be achieved by heating the slurry or diluting it with a fluid of lower viscosity;
2. increasing the pressure drop across the filter cake, but there is an upper limit on the pressure drop in vacuum filtration, and the pressure drop can increase the filter cake resistance; and
3. decreasing filter cake resistance, i.e., increasing the porosity of the cake and decreasing the thickness of the cake: increasing the porosity of the cake can be

achieved by the use of filter aids or by coagulation via using flocculants or changing the *p*H of the slurry, whereas decreasing the cake thickness can be done by employing moving blades to remove the filter cake mechanically (Chiang and He, 1993).

The filtrate retained in the pores of the filter cake is removed via washing. The maximum wash ratio between wash fluid volume and cake void volume is usually 1.5–2.0 for removing about 90% of the retained material (Pinheiro and Cabral, 1993a).

7.4 Important Industrial Applications

The industrial applications include the following (McCabe and Smith, 1956; Gregor *et al.*, 1992; Pinheiro and Cabral, 1993a; Combest and Corporation, 1994):

1. Filtration and washing of pigments and dyes: Plate-and-frame filter presses are usually used for these applications in batch operation.
2. Dewaxing of petroleum: This is carried out by the use of plate-and-frame filter presses.
3. Pharmaceutical processes for the recovery of fermentation broths, ascorbic acid, and antibiotics (albamycin, erythromycin, penicillin, streptomycin, etc.): For these processes, vacuum precoat filters are typically employed in continuous operation.
4. Dewatering of sludges: This is conducted via rotary-drum pressure filters in continuous operation.
5. Filtration and washing of fine crystals: Rotary-drum vacuum filters are usually employed in continuous operation.
6. Clarification of process liquids and cooling water: Cartridge filters are typically used in batch operation.
7. Automotive and truck filters: Cartridge filters are generally employed for fuel injection and air intake.

8. Sedimentation

8.1 Introduction and Process Description

Sedimentation, similar to filtration, is the process of separating a fluid from particles suspended in it by allowing the particles to settle out from the fluid. The fluid above the bed of solid particles is called supernatant, and it can then be removed from the particles. In the beginning of the sedimentation process, the particles in a slurry (suspension) behave independently. But as the particles fall closer to the bottom of the settling vessel, the influences of neighboring particles become pronounced, which is called hindered settling. To remove relatively large, heavy particles such as coarse sands, which have reasonably high settling velocities, gravity sedimentation under free or hindered settling is satisfactory. However, to remove fine particles, flocculation to increase the effective diameter of the settling particle and centrifugation to replace the force of gravity can be used to enhance the fluid-particle separation (Baczek *et al.*, 1988; McCabe and Smith, 1956; Pinheiro and Cabral, 1993b).

Sedimentation may be divided into two functional operations: thickening and clarification, both operations taking place simultaneously in sedimentation. However, the primary objective of thickening is to increase the concentration of suspended particles in a slurry stream, whereas the main purpose of clarification is to remove relatively small amounts of particles and produce a clear effluent. In general, thickeners are designed for the heavy-duty requirements imposed by highly concentrated slurry, whereas clarifiers have features to ensure essentially complete removal of particles, including greater depth, special provision for coagulation and flocculation, and greater overflow-weir length (Dahlstrom *et al.*, 1984).

8.2 Process Equipment

As described above, sedimentation equipment includes thickeners and clarifiers. Industrially important thickeners and clarifiers are for continuous operations.

8.2.1 Thickeners There are two types of continuous thickeners: conventional and high-rate, both with circular horizontal cross section since the rectangular alternative tends to give higher maintenance costs and poor performance for high underflow concentrations (Baczek *et al.*, 1988; Dahlstrom *et al.*, 1984; Pinheiro and Cabral, 1993b). A conventional thickener is shown in Fig. 11. The slurry feed is introduced to the center of

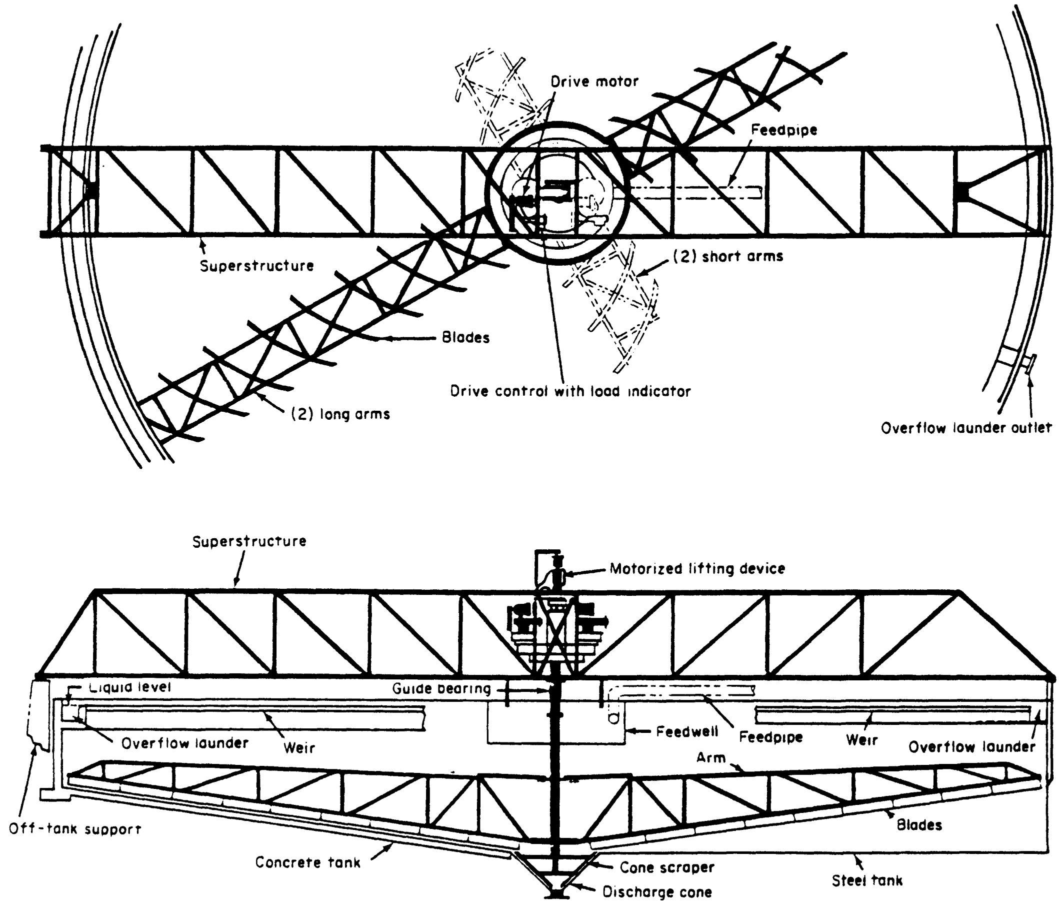
Drive motor
Feedpipe
Superstructure
(2) short arms
Blades
Drive control with load indicator
(2) long arms
Overflow launder outlet
Superstructure
Motorized lifting device
Liquid level
Guide bearing
Overflow launder
Weir
Feedwell
Feedpipe
Arm
Weir
Overflow launder
Off-tank support
Blades
Concrete tank
Cone scraper
Discharge cone
Steel tank

FIG. 11. Conventional thickener [Eimco Process Equipment Co., reprinted from Dahlstrom *et al.* (1984) with permission].

the thickener in a pipe or an open launder and enters the feed well. The feed well is designed to minimize the turbulence caused by the feed entry and to introduce the feed below the clear-liquid surface. The solid particles settle to the bottom where they are directed to the underflow part by a revolving raking mechanism, while clear liquid is removed from the top of the thickener in a peripheral launder. The conventional thickener can be used with or without flocculants. If flocculants are utilized, they are usually added to the feed launder or in the feed well.

A high-rate thickener is designed specifically to maximize the efficiency of flocculation. As shown in Fig. 12, the feed stream enters the feed well, and the flocculant is added also at the feed well but at various locations to optimize mixing with the feed stream. A mechanical mixer is used to disperse the flocculant. A deaeration step must be employed to remove entrained air from the feed stream so that the turbulence in the thickener can be minimized and the flotation of solid particles will not occur. This thickener can decrease the equipment area and volume by a factor of about 5 in comparison with the conventional thickener.

8.2.2 Clarifier Clarifiers are basically identical to thickeners in design and layout except that the former equipment uses a structure of lighter construction and a drive head with a lower torque capability (Baczek *et al.*, 1988; Dahlstrom *et al.*, 1984; Pinheiro and Cabral, 1993b; Springer, 1993). These

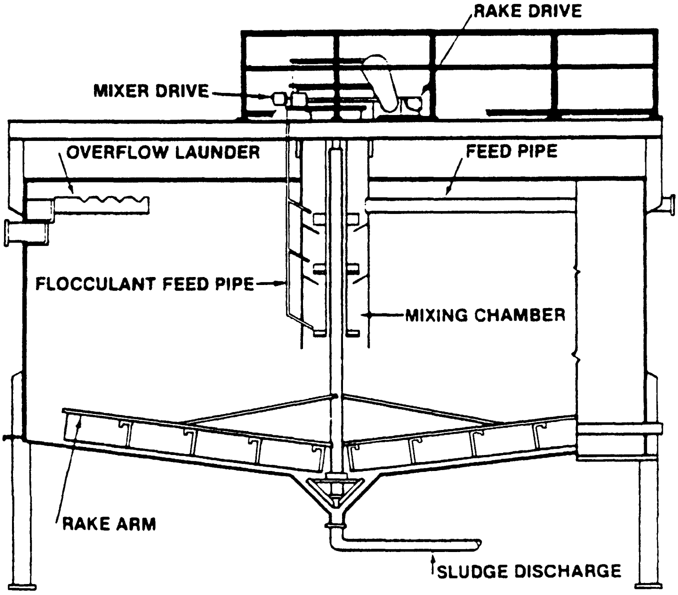
RAKE DRIVE
MIXER DRIVE
OVERFLOW LAUNDER
FEED PIPE
FLOCCULANT FEED PIPE
MIXING CHAMBER
RAKE ARM
SLUDGE DISCHARGE

FIG. 12. High-rate thickener [Eimco Process Equipment Co., reprinted from Dahlstrom *et al.* (1984) with permission].

differences are allowed because in clarifiers the thickened pulp (concentrated solid particles) produced is considerably less in volume and lower in concentration. Both circular and rectangular horizontal cross-section geometries are used. As shown in Fig. 13 for a rectangular clarifier, the influent (feed) enters along one of the short sides, and the clear liquid overflows along the opposite side. The chain-type drag moves the deposited pulp to a sludge hopper located at the influent end by means of scrapers fixed to the chains. During their return, the scrapers travel near the liquid level and thus act as skimming devices for removing surface scum.

8.3 Sedimentation Operations

Batch sedimentation operation allows the slurry (suspension) to settle for the required period of time, after which the sludge and

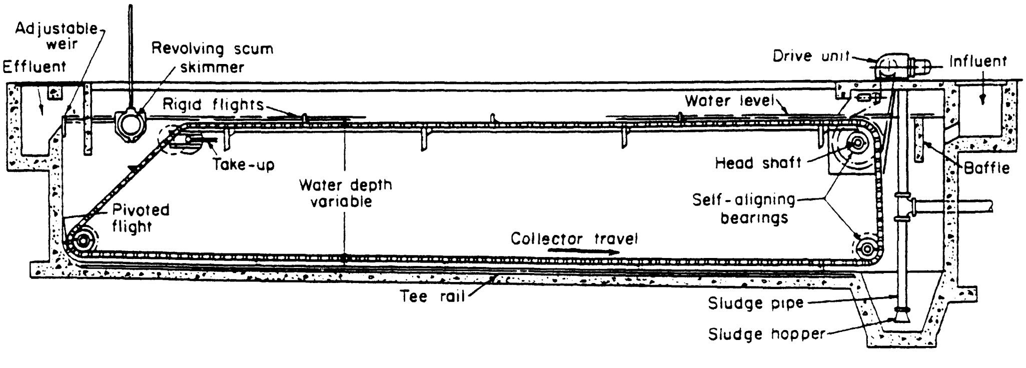
Adjustable weir
Effluent
Revolving scum skimmer
Rigid flights
Take-up
Pivoted flight
Water depth variable
Collector travel
Tee rail
Drive unit
Influent
Water level
Head shaft
Baffle
Self-aligning bearings
Sludge pipe
Sludge hopper

FIG. 13. Rectangular clarifier [Rexnord Inc., reprinted from Dahlstrom *et al.* (1984) with permission].

supernatant are removed separately. In this operation, the retention time is the major design parameter. But this operation is expensive in labor requirements and limited in capacity. Thus, industrial applications are mainly in continuous operation (Dahlstrom *et al.*, 1984; Pinheiro and Cabral, 1993b).

In continuous operation, a solids vertical concentration profile in the thickener or clarifier is kept constant to ensure that both thickening and clarification occur (Dahlstrom *et al.* 1984; Pinheiro and Cabral, 1993b). Such a concentration profile is shown schematically in Fig. 14 (Pinheiro and Cabral, 1993b). In the local thickening (compression) zone, a scraper or a rake is used to direct the sludge to the bottom outlet and to aid in solids compression. The solids content and the degree of thickening depend on the time of retention of the solids in the thickening zone. The retention time is, in turn, proportional to the depth of this zone, which can be increased only at the expense of a deeper and more costly thickener. A clear overflow effluent can be obtained if the upward flow velocity of the liquid in the dilute zone is less than the minimum terminal velocity of the solid particle. The liquid velocity is directly proportional to the overflow rate. If the overflow rate is too high, a cloudy overflow liquid will result. Thus, the overflow rate needs to be controlled in order to obtain a clean overflow effluent (McCabe and Smith, 1956; Pinheiro and Cabral, 1993b).

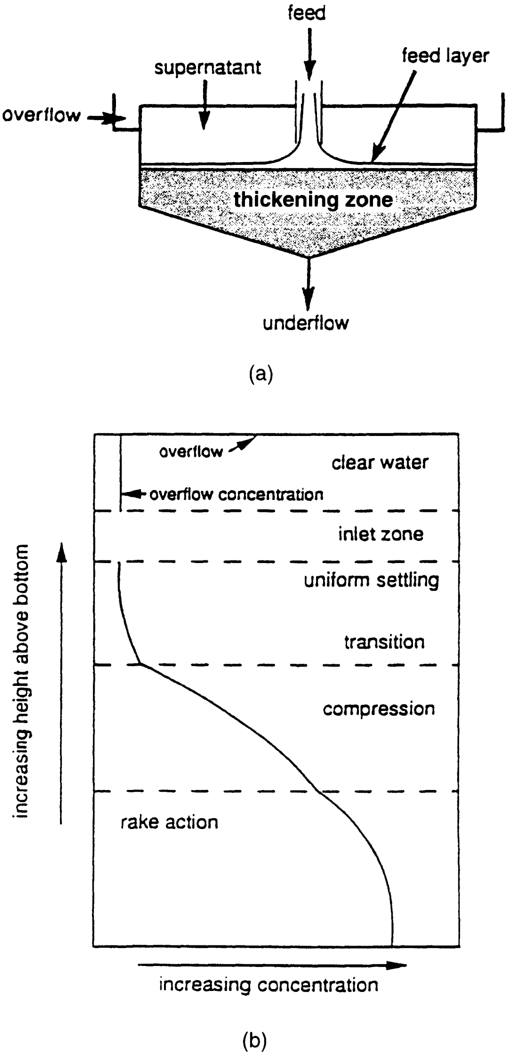

FIG. 14. Schematic of a solids vertical concentration profile: (a) thickener and (b) solids vertical concentration profile [reprinted from Pinheiro and Cabral (1993) with permission].

8.4 Important Industrial Applications

The applications include the following (Dahlstrom *et al.*, 1984; Pinheiro and Cabral, 1993b):

1. Municipal wastewater treatment: Both thickeners and clarifiers are used for these applications. Thickeners are employed for primary sludge, waste-activated sludge, and anaerobically digested sludge.
2. Drinking-water treatment: Clarification is normally carried out after 30-min. flocculation. It is also conducted with softening treatment using lime.
3. Metallurgical operations: Sedimentation is used in processing the concentrates and tailings of ores including copper, iron, lead, nickel, zinc, etc.
4. Pulp and paper: Clarifiers are used for processing green and white liquors. Thickeners are employed for handling kraft waste, deinking waste, and paper-mill waste.
5. Flue-gas desulfurization sludge: For treating the sludge, sedimentation is used.
6. Oil–water separations in refineries: For preliminary oil–water separations in refineries, rectangular clarifiers are usually employed.

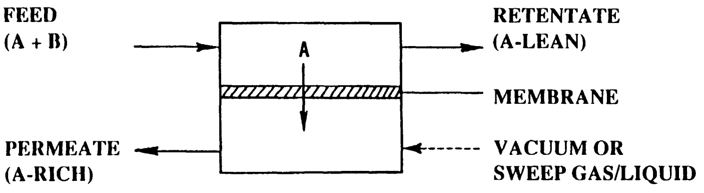

FIG. 15. Schematic of a membrane process.

9. MEMBRANE SEPARATIONS

9.1 Introduction and Process Description

A membrane process, shown schematically in Fig. 15, requires two bulk phases physically separated by a third phase, the membrane. In all membrane processes, the feed is separated into a stream that goes through the membrane, the permeate, and a portion of the feed that is retained by the membrane, the retentate (the reject or the concentrate). In some processes [e.g., pervaporation and vapor (gas) permeation], a vacuum is applied in the permeate side, whereas in a few processes (e.g., dialysis, electrodialysis, and emulsion liquid membrane), a sweep, wash, or strip stream is needed on the permeate side. The membrane controls the transport of mass between the two bulk phases. One of the species in the feed mixture, i.e., A in this figure, is allowed to pass through the membrane in preference to others, that is, the membrane is selective to species A. The permeate phase is enriched in the species A, while the retentate phase is depleted of it (Ho and Li, 1984; Ho and Sirkar, 1992a, 1992b).

The membrane is either a homogeneous phase or a heterogeneous collection of phases, and it may be any one or a combination of the following: nonporous solid, microporous or macroporous solid with a fluid (liquid or gas) in the pores, a liquid phase with or without a second phase, or a gel. The membrane phase is almost always thin when compared with the dimensions of the bulk phases in at least two other directions. Hollow fibers and emulsion liquid membranes provide exceptions where the membrane thickness is of the order of the dimensions of one of the bulk phases (Ho and Sirkar, 1992b; Lonsdale, 1989). In general, a commercial membrane consists of a thin, selective, active layer or skin (about 0.1–5 μm) on a porous support layer (about 30–200 μm), which provides mechanical strength. If the active and support layers are formed in a single operation from a given material, the membrane is referred to as having an integrally skinned structure. If the active layer is a coating on the support layer, or if a coating is put on the active layer, the membrane is said to have a composite structure. Two geometries of membranes are usually used, namely, flat films and hollow fibers (outside diameters of about 50–1000 μm and walls of about 10–100 μm thick). Figure 16 shows an integrally skinned hollow-fiber membrane (Kesting, 1985).

The transport of any species across the membrane is caused by one or more driving forces. These driving forces arise from a gradient of chemical potential or electrical potential. A gradient in chemical potential may be due to concentration gradient or pressure gradient or both. The transmembrane flux of

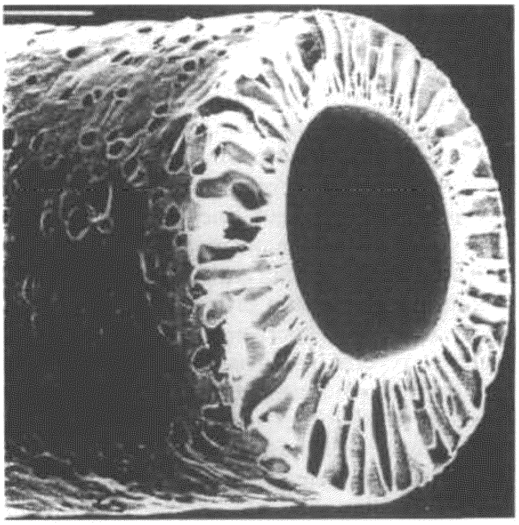

FIG. 16. Integrally skinned hollow-fiber membrane with a dense, active layer on a porous support structure [Amicon Corporation, reprinted from Kesting (1985) with permission].

any species per unit driving force is proportional to the permeability of the species. If the driving force is described by the use of a partial pressure difference (Δp_i) or a concentration difference (Δc_i) across the membrane for species i (if there is an osmotic pressure difference $\Delta\pi$, the pressure gradient driving force is $\Delta p_i - \Delta\pi$), then

$$\text{Transmembrane Flux of Species } i = \left(\frac{\text{Permeability of Species } i}{\text{Effective Membrane Thickness}}\right)(\Delta p_i \text{ or } \Delta c_i). \tag{17}$$

The ratio of permeability of species i to effective membrane thickness is sometimes called the permeance (which is also called permeability in some special processes, e.g., dialysis and microfiltration). The membrane selectivity between any two species can be defined in a number of ways. A common definition called separation factor α_{ij} for two species i and j is

$$\alpha_{ij} = \frac{c_i''/c_j''}{c_i'/c_j'} \tag{18}$$

where the prime and double prime superscripts denote the upstream bulk phase (feed or retentate) and the downstream bulk phase (permeate), respectively. The separation factor is equal to the ratio of the permeabilities of the two species under conditions where the downstream pressure or concentration is negligible in comparison with the upstream pressure or concentration. Such selectivities are reduced or destroyed if the membrane has gross defects through which nonselective hydrodynamic flow of one bulk phase into the other bulk phase occurs.

There have been eight commercialized membrane separation processes; they are gas permeation, pervaporation, dialysis, electrodialysis, reverse osmosis, ultrafiltration, microfiltration, and emulsion liquid membrane (the nanofiltration process is included in reverse osmosis) (Ho and Sirkar, 1992a, 1992b; Belfort, 1984; Bungay *et al.*, 1986; Rautenbach and Albrecht, 1989; Porter, 1990). Gas permeation involves gases on the high-pressure side of the membrane permeating through the membrane to its low-pressure side. The membranes currently used in most commercial gas separations are polymeric. The transport mechanism through polymeric membranes is solution-diffusion, and it consists of three steps:

1. solution of the permeating molecules from the feed at the upstream side of the membrane,
2. diffusion of these molecules through the membrane, and
3. desorption at the downstream side of the membrane to become the permeate (Ho and Li, 1984; Zolandz and Fleming, 1992; Koros and Pinnau, 1994; Paul and Yampol'skii, 1994).

In a pervaporation process, a liquid feed contacts one side of a membrane while a vacuum is drawn on the other side of the membrane to produce a permeate vapor, which is subsequently condensed to a liquid. The pervaporation membranes currently employed in most commercial applications are polymeric, and the transport mechanism through the membranes is solution-diffusion. The separation of feed components at given operating conditions is based on the selectivity of the membrane employed. In turn, the membrane selectivity comes from the different solubilities and diffusivities of the components in the membrane (Huang, 1991; Fleming and Slater, 1992; Wijmans and Baker, 1993).

Dialysis is the transfer of solute molecules across a membrane by diffusion from a concentrated solution to a dilute solution. A simultaneous diffusion of solvent molecules through the membrane occurs in the opposite direction—a phenomenon called osmosis. The ratio of mass of water transported to mass of dissolved solute dialyzed is defined as the water-transport number. The membranes used can be either porous or nonporous. The remaining solution on the feed side is called retentate, and the solution on the other side to receive the solutes to be removed from the feed is called dialyzate. Separation of solutes in dialysis is due to differences in their diffusion rates. The process of mass transfer from the feed compartment to the dialyzate compartment involves diffusion through a liquid film on each side of the membrane as well as through the membrane itself. Diffusion of solute molecules is governed by a number of factors, such as the inherent mobility of the molecules and the restrictive effect or drag exerted by the membrane pores on both solute and solvent molecules (Ho and Li, 1984; Kessler and Klein, 1992).

The electrodialysis process uses an electrodialysis stack such as shown schematically in Fig. 17 consisting of a series of anion- and cation-exchange membranes arranged in an alternating pattern between an anode and a cathode to form individual cells (Strathmann, 1992). A cell is composed of a volume with two adjacent membranes (about 0.5–2 mm between two membranes), i.e., anion- and cation-exchange membranes. If an ionic solution, e.g., an aqueous salt solution, is pumped through these cells and an electrical potential established between the anode and cathode, the negatively charged anions move toward the anode while the positively charged cations migrate toward the cathode. The anions pass easily through the positively charged anion-exchange membrane but are rejected by the negatively charged cation-exchange membrane. Similarly, the positively charged cations migrate through the cation-exchange membrane but are retained by the anion-exchange membrane. The overall result is an increase in the ion concentration in alternate compartments, while the ion concentration in the other compartments decreases. The solution with the reduced ion concentration is usually called the diluate, and the concentrated solution is the brine or the concentrate. The driving force for the ion transport is the applied electrical potential between the anode and cathode.

Reverse osmosis separates the solvent from a solution by forcing it to pass through a membrane by applying a pressure greater than the normal osmotic pressure. The solvent molecules are of about the same size as that of solute molecules. The solute rejection efficiency R_S is defined as

$$R_s = 1 - c_s''/c_s', \tag{19}$$

where c_S' is the solute concentration in the upstream retentate and c_S'' is the solute concentration in the downstream permeate (Ho and Li, 1984). Reverse osmosis processes may be classified into three types:

1. high-pressure reverse osmosis (5.6–10.5 MPa, i.e., 813–1524 psi), e.g., seawater desalination with very high rejection of inorganics (95–99.9% NaCl rejection);
2. low-pressure reverse osmosis (1.4–4.2 MPa, i.e., 203–610 psi), e.g., brackish water desalination and moderate to high rejection of low molecular weight organics; and

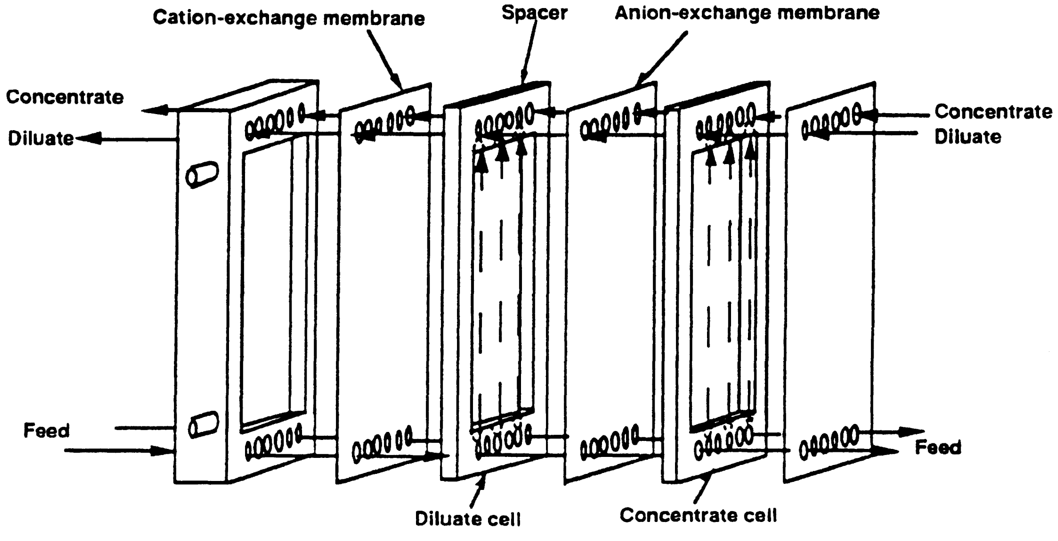

FIG. 17. Exploded view of components in an electrodialysis stack [reprinted from Strathmann (1992) with permission].

3. nanofiltration or loose reverse osmosis (0.3–1.4 MPa, i.e., 44–203 psi), e.g., divalent ion removal, partial demineralization, or 0–20% NaCl rejection.

Most transport models for reverse osmosis assume the solution-diffusion mechanism, the sieving mechanism, and the surface force–pore flow mechanism (Bhattacharyya *et al.*, 1992; Ho and Li, 1984; Sourirajan and Matsuura, 1985).

Similar to reverse osmosis, ultrafiltration is a pressure-driven membrane process capable of separating solution components on the basis of molecular size and shape. Under an applied pressure difference across an ultrafiltration membrane, solvent and small solute species pass through the membrane and are collected as permeate while larger solute species are retained by the membrane and recovered as a concentrated retentate. Ultrafiltration involves solutes whose molecular dimensions are 10 or more times larger than those of the solvent and are usually below 0.02 μm (200 Å) in size. The solutes or the materials to be separated usually have molecular weights greater than 500, such as macromolecules (proteins, polymers, starches, natural gums, enzymes, etc.), colloidal dispersions (clays, pigments, minerals, latex particles, and microorganisms), and emulsions (grease–detergent and oil–water emulsions). As a consequence of relatively high molecular weight, the osmotic pressures of the solutes are usually low. Thus, the operating pressures of ultrafiltration range only from 34 to 689 kPa (5 to 100 psi). The lower pressures reduce the energy requirement for pumping and compression and the equipment cost by a considerable margin over reverse osmosis. Ultrafiltration can be used to concentrate, purify, and fractionate macrosolutes or materials in the feed liquid. The solute rejection efficiency for ultrafiltration is defined in the same way as that for reverse osmosis, Eq. (19). Transport in ultrafiltration is through the sieving mechanism. When solvent migrates toward the membrane surface, it carries solute, which is rejected at the membrane surface, resulting in an accumulation of solute on the membrane. This accumulation usually leads to the formation of a gel layer, i.e., secondary membrane (Ho and Li, 1984; Cheryan, 1986; Kulkarni *et al.*, 1992; Sourirajan and Matsuura, 1985).

Similar to reverse osmosis and ultrafiltration, microfiltration is a pressure-driven membrane process, and it separates micron-sized particulates from a fluid. In general, microfiltration is defined as the filtration of a suspension containing colloidal or fine particulates with linear dimensions in the approximate range of 0.02 to 10 μm. These particulates are usually larger than the solutes separated by reverse osmosis and ultrafiltration. Thus, the osmotic pressure for microfiltration is negligible, and the transmembrane pressure drop is relatively small (typically, 7 to 345 kPa, i.e., 1 to 50 psi). The membrane pore size and permeate flux for microfiltration are generally larger than those for reverse osmosis and ultrafiltration. Similar to ultrafiltration, microfiltration is based on the transport through the sieving mechanism (Davis and Grant, 1992).

The emulsion liquid membrane (ELM) process is unique and different from the membrane processes discussed above. The membrane is a liquid phase involving an emulsion configuration. Emulsion liquid membranes, also called surfactant liquid membranes or liquid surfactant membranes, are essentially double emulsions, i.e., water/oil/water (W/O/W) systems or oil/water/oil (O/W/O) systems. For the W/O/W systems, the oil phase separating the two aqueous phases is the liquid membrane. For the O/W/O systems, the liquid membrane is the water phase that is between the two oil phases. Figure 18 shows a schematic of a continuous ELM process. This process includes four steps:

1. emulsification;
2. dispersion of the emulsion in contact with the external, continuous phase for extraction;
3. settling to separate the emulsion from the external phase, which is the raffinate if the internal phase becomes the extract; and
4. breaking the emulsion to recover the internal phase as the extract and the membrane phase for recycle.

The ELMs are usually prepared by first forming an emulsion between two immiscible phases and then dispersing the emulsion in a third (continuous) phase by agitation for extraction. The membrane phase is the liquid phase that separates the encapsulated,

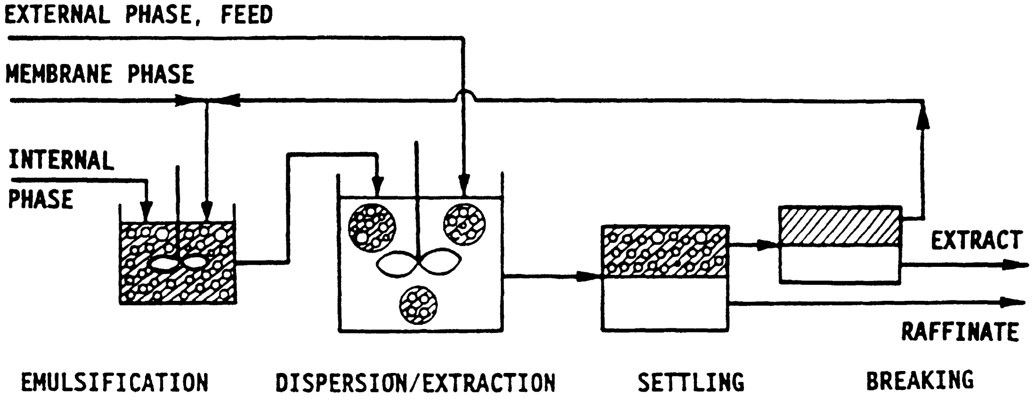

FIG. 18. Schematic of a continuous emulsion liquid membrane process [reprinted from Ho and Li (1992) with permission].

internal droplets in the emulsion from the external, continuous phase, as shown schematically in Fig. 19. In general, the internal, encapsulated phase and the external, continuous phase are miscible. However, the membrane phase must not be miscible with either of these two phases in order to be stable. To maintain the integrity of the emulsion during the extraction process, the membrane phase generally contains some surfactant(s) and additive(s) as stabilizing agents, and it also contains a base material that is solvent for all the other ingredients.

Separation of mixtures is via the solution-diffusion mechanism by selective transport of one component through the membrane phase into the receiving phase of lower equivalent concentration. Surfactant(s), additive(s), and extractant(s) included in the membrane phase can control the selectivity and permeability of the membrane. An individual component can be trapped and concentrated in the internal phase for later disposal or recovery. Once separation is achieved, the emulsion and external, continuous phases are separated, usually by settling as in conventional solvent extraction. The extracted component can be recovered from the "loaded" internal phase of the emulsion by breaking the emulsion, usually by the use of an electrostatic coalescer. From the broken emulsion, the membrane phase recovered can then be recycled to the emulsification step for the preparation of the emulsion with a regenerated or fresh internal reagent phase. Since their discovery by Li (1968), emulsion liquid membranes have demonstrated considerable potential as effective tools for a wide variety of separations (Ho and Li, 1992; Gu *et al.*, 1992).

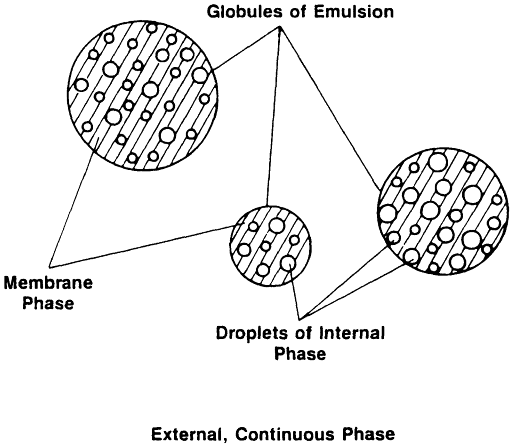

FIG. 19. Schematic of an emulsion liquid membrane system [reprinted from Ho and Li (1992) with permission].

Large-scale or widespread commercial uses were developed earliest for microfiltration, dialysis (hemodialysis), electrodialysis, and reverse osmosis. Ultrafiltration was adopted for commercial uses immediately afterward. The first large-scale gas-separation modules were developed by Du Pont as early as 1970; but the first successful commercial membrane gas-separation processes were announced by Monsanto in the late 1970s. Many more have appeared since then. Significant commercialization of pervaporation separation began during the period 1985–1990. Commercialization of emulsion liquid membranes for the removal of zinc and phenol from wastewaters was also achieved in Austria and China, respectively, around 1986.

All commercialized membrane separation processes, except gas permeation and microfiltration of gaseous feeds, are capable of separating liquid solutions or liquid-phase feeds, and virtually all of them are principally used for separating aqueous solutions. In the aqueous solutions, the solutes may be microsolutes (inorganic or organic), macrosolutes, proteins, polymers, colloids, whole cells, precipitates, dirt particles, etc. An exception is pervaporation, whose principal commercial use at this time is for removing small amounts of water from organic liquids such as alcohols. Also, a new pervaporation technology based on a polyimide–polyester copolymer membrane (Ho *et al.*, 1990, 1991) has been developed recently to separate heavy catalytically cracked naphtha into an aromatics-rich permeate and an aromatics-lean retentate (Exxon, 1994). The higher aromatic content of the permeate improves its octane and therefore makes it an excellent gasoline blendstock, while the retentate becomes a much better distillate (jet fuel and diesel) blend component because of its reduced aromatics level. In addition, ultrafiltration has been used for treating waste aqueous streams with a very high volume fraction of organics, and microfiltration separation of organic liquids is practiced.

In general, staging in membrane processes is economically unfavorable. Thus, membrane processes are usually best suited for bulk separation rather than for purification. Exceptions include membrane separations with very high selectivities, e.g., reverse osmosis for desalination, pervaporation for dehydration of ethanol and isopropanol, pervaporation for removal of sparingly soluble organics from wastewater and some ultrafiltration and microfiltration applications, and the emulsion liquid membrane process where staging and facilitated transport mechanisms with chemical reactions (Li, 1978; Ho and Li, 1992; Gu *et al.*, 1992) can be incorporated (Ho and Sirkar, 1992a).

Although most membranes used in the commercialized processes are polymeric, other membranes, e.g., ceramic, metal, etc., are also used or being studied. However, a few common features of membranes are dictated by the mechanism for transport/selectivity, e.g., only nonporous or solvent membranes can give the solution-diffusion mechanism. Therefore, the selective layer in gas permeation and pervaporation membranes is solventlike, i.e., nonporous. Sieving by a membrane presupposes the existence of pores, and thus membranes in dialysis, ultrafiltration, and microfiltration have pores. Membranes for ultrafiltration and microfiltration are generally identified as microporous (Ho and Sirkar, 1992b).

9.2 Process Equipment

In commercial processes, membranes must be packaged into modules with as high surface area per unit volume as possible but still allow good flow distribution and efficient contact of the feed with the membrane. Module housings are normally fabricated from standard-sized pipe, and they range from about 4 to 12 in. (10 to 30 cm) in diameter and from 4 to 20 ft (1.2 to 6.1 m) in length. The housing materials depend on applications, and carbon steel and aluminum are usually suitable. Membrane geometries lead to different module designs. Membrane modules include the following:

1. Spiral-wound modules: A spiral-wound module constructed from flat-film membranes is shown schematically in Fig. 20 (Zolandz and Fleming, 1992). It is fabricated by taking two membrane sheets with a permeate spacer in between and sealing them together with glue at three edges to form a membrane leaf. The open end of the leaf is glued to a perforated tube. The membrane leaf and a feed spacer are then rolled around the perforated tube and glued with an outer wrap to form the module. When the module is used, the feed is brought to the feed spacer and in contact with outer surfaces of the membrane leaf. Permeant passes through the walls of the membrane leaf and then flows along the permeate spacer to the perforated, collection tube. The modules typically contain about 1000 ft^2 (93 m^2) of membrane area per cubic foot (0.028 m^3) volume, i.e., 3281 m^2/m^3, for

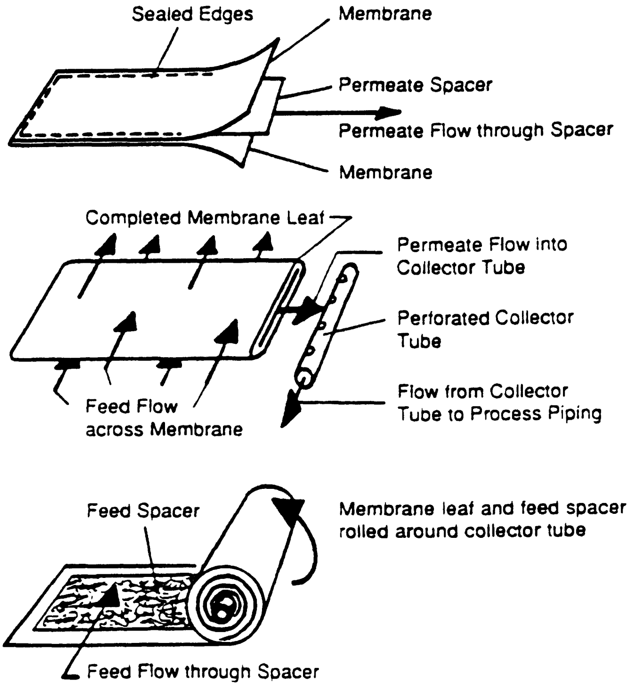
Sealed Edges
Membrane
Permeate Spacer
Permeate Flow through Spacer
Membrane
Completed Membrane Leaf
Permeate Flow into Collector Tube
Perforated Collector Tube
Flow from Collector Tube to Process Piping
Feed Flow across Membrane
Feed Spacer
Membrane leaf and feed spacer rolled around collector tube
Feed Flow through Spacer

FIG. 20. Schematic of a spiral-wound module assembly [reprinted from Zolandz and Fleming (1992) with permission].

gas permeation and about 300 ft^2/ft^3 (984 m^2/m^3) for liquid separations.

2. Hollow-fiber modules: The modules are fabricated in a similar manner as shell-and-tube heat exchangers. The fibers are arranged parallel to one another and glued together to form tube sheets at either one or both ends of the module. There is a seal between the tube-sheet exterior and the pressure vessel in order to isolate the feed from the permeate. Figure 21 shows a schematic of a cross-flow feed hollow-fiber module. Hollow-fiber modules can provide the highest membrane area per unit volume among all membrane modules. The hollow-fiber modules for gas permeation can have about 3000 ft^2 per cubic foot volume, i.e., 9843 m^2/m^3 (Zolandz and Fleming, 1992).
3. Tubular modules: Similar to hollow-fiber modules, tubular models are like shell-and-tube heat exchangers. But the tube diameters are generally 2.5 to 25 mm, and they are larger than hollow-fiber diameters. In general, the feed enters the tubes at one end of the module, is concentrated, and exits at the other end, while the permeate comes out from the shell side. Recently, ceramic modules have become available. Figure 22 shows the Ceraflo™ ceramic module containing nineteen 3-mm-i.d. tubes. The feed flows through the tubes while the permeate passes through the microporous walls of the module into the surrounding shell for withdrawal (Kulkarni *et al.,* 1992; Mir *et al.,* 1992). The ceramic modules are particularly suited for high-temperature applications (about 200 °C and higher).
4. Plate-and-frame modules: Flat-film membranes can be assembled in a plate-and-frame module similar to the filter presses shown in Fig. 9. A new variation, i.e., flat-stack membrane module, has been developed recently and is shown in Fig. 23 (Ohlrogge *et al.,* 1995). In this new module, two round, flat-film membranes are thermally welded at the cutting edges forming a membrane envelope. The envelopes are placed on a perforated, central tube, and the stack of the envelopes is divided into asymmetrical compartments by means of baffle plates. In each compartment, the number of the envelopes is reduced progressively in the direction of the feed flow with the decrease of feed vol-

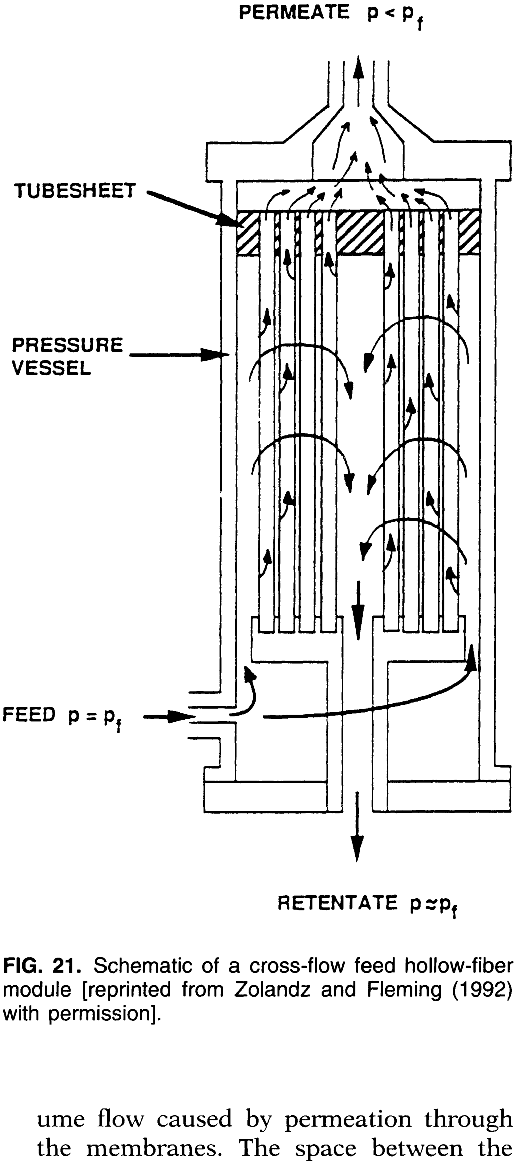

FIG. 21. Schematic of a cross-flow feed hollow-fiber module [reprinted from Zolandz and Fleming (1992) with permission].

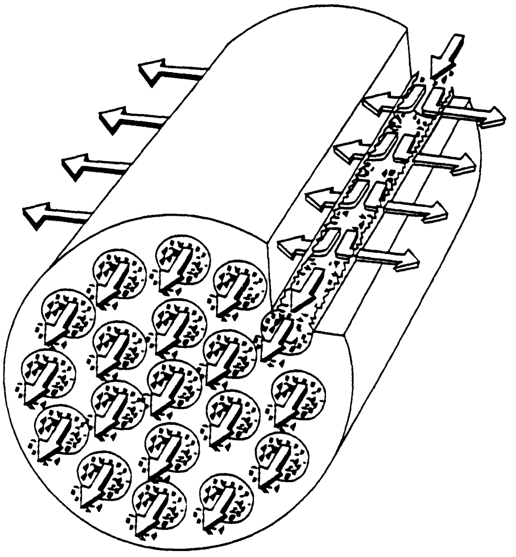

FIG. 22. Ceraflo™ ceramic module (courtesy Millipore Corporation, Literature No SD117, May 1989).

ume flow caused by permeation through the membranes. The space between the membranes ranges from 0.25 to 2.5 mm (Mir *et al.*, 1992). As described earlier, an electrodialysis stack similar to the plate-and-frame arrangement but using rectangular membranes is shown in Fig. 17.

5. Rotary-drum filters: Rotary-drum vacuum filters are widely used for difficult filtrations including microfiltrations for the removal of microbial cells and other particulates that form compressible cakes with large resistances (Davis and Grant, 1992). As described earlier, the filter is shown schematically in Fig. 10.

For gas permeation, pervaporation, and reverse osmosis, both spiral-wound and hollow-fiber modules are used. Plate-and-frame modules are sometimes used for gas permeation and pervaporation. Dialysis processes usually employ hollow-fiber and plate-and-frame modules. In ultrafiltration and microfiltration, tubular, spiral-wound, and plate-and-frame modules are normally used. As mentioned above, electrodialysis stacks are employed for electrodialysis, and rotary-drum filters are sometimes used for microfiltration. For emulsion liquid membranes, the contacting devices are the same as those for solvent extraction, including mixer settlers and mechanically agitated column extractors (Gu *et al.*, 1992).

9.3 Membrane Separation Operations

Membrane process operations with compact modular systems are usually the simplest among all separation processes. As long as the driving force for the membrane process exists, the membrane separation can take place. The driving force for gas permeation, reverse osmosis, ultrafiltration, and microfiltration processes comes from the pressure differential across the membrane for the species transported through the membrane. Thus, these processes are called

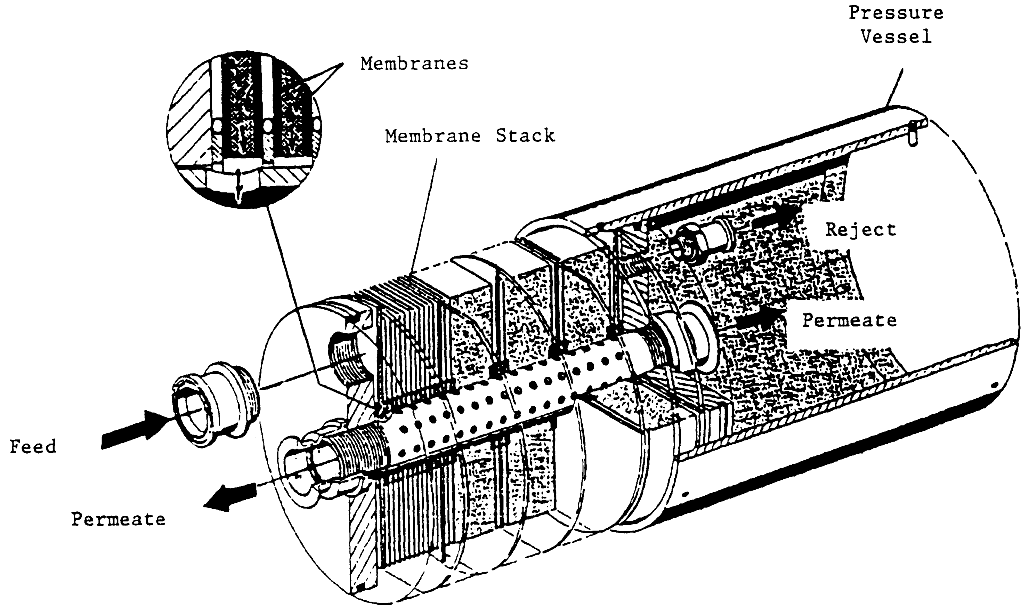

FIG. 23. Flat-stack membrane module [reprinted from Ohlrogge *et al.* (1995) with permission].

pressure-driven processes. The driving force for the concentration-driven processes, namely, pervaporation, dialysis, and emulsion liquid membrane, is the concentration difference across the membrane for the species diffusing through membrane. For pervaporation, temperature can also be used to increase the concentration differential for a given vacuum in the permeate side. As mentioned earlier, the driving force for ion transport in electrodialysis is the applied electrical potential between the anode and cathode. Industrial membrane processes are generally continuous in operation.

Membrane modules are designed to operate within a range of feed flow rates, and it is a good practice to stay within the range. Operating the module below the minimum design flow rate can give poor distribution of the feed over the available membrane area, whereas operating above the design maximum flow rate can result in excessive pressure drop for the feed stream flowing through this module. Operation with poor distribution of the feed results in both poorer product (e.g., permeate) purity and lower module productivity. In a given membrane stage, the proper feed rate is achieved by piping modules in parallel and dividing the feed stream equally among the modules. If the feed rate will change with time, a membrane stage can consist of parallel modules, and the required number of modules is then employed as needed (Zolandz and Fleming, 1992).

The purity of the permeate stream, e.g., the product stream in gas permeation, reduces from its maximum at a stage cut (the ratio of the permeate flow rate to the feed flow rate) of zero as the recovery of the permeate increases. The trade-off between purity and recovery is intrinsic to membrane processes. To increase product purity beyond what can be achieved by once-through operation, the membrane process must be operated with recycle. Recycle of permeate to the feed side increases the component concentration in the feed side, resulting in higher product purity at a given recovery. However, recompression of the recycled permeate in a pressure-driven process, e.g., gas permeation, adds to both the investment and operating costs (Zolandz and Fleming, 1992).

In pervaporation, the process is mainly driven by condensation of the permeate, resulting in a significant vacuum and lower temperatures on the permeate site of the membrane. Even though a vacuum pump is still needed for startup, it is normally not used in steady-state operation, and it is small. The enthalpy of vaporization for the feed is consumed, and interstage heaters to reheat the feed are employed (Fleming and Slater, 1992).

In dialysis (Kessler and Klein, 1992), advantage should be taken of natural convection currents in dialysis cells in order to maintain a maximum concentration gradient across the membrane. The membrane may be mounted vertically with the feed and solvent phases flowing countercurrently on opposite sides of the membrane. The directions of flow of these two phases should be determined by their densities. For example, if the feed has a higher density than the solvent, it should be introduced at the bottom of the feed compartment and the solvent should be introduced at the top of the dialyzate compartment. The solvent in the feed compartment, entering by osmosis through the membrane, will produce a rising stream of low-density weak liquor, whereas the diffusing solute in the dialyzate compartment will form a stream of high-density solution flowing down the membrane surface. This density-streaming effect can reduce greatly the thickness of the liquid films at both sides of the membrane and thus lower the overall resistance to dialysis (Ho and Li, 1984).

In electrodialysis, concentration polarization effects result in the depletion of ions in the laminar boundary layer at the membrane surfaces in the cell containing the diluate flow stream and in the increase of ions in the laminar boundary layer at the membrane surfaces in the cell containing the brine solution. Concentration polarization decreases separation efficiency. The magnitude of concentration polarization can be controlled by the electrical current density, the spacer design to decrease the boundary-layer thickness, and the flow velocities of the diluate and brine solutions (Strathmann, 1992).

In reverse osmosis, ultrafiltration, and microfiltration, concentration polarization decreases fluxes. When solvent passes through a membrane, the solute (or particulates in microfiltration) left behind concentrates in a solution next to the membrane surface. Accompanying the increase in solute concentration at the interface are an increase in the osmotic pressure of the solution at the interface and the formation of a secondary membrane, resulting in flux reduction. The solute layer built up near the membrane surface may reach a constant thickness as a result of a balance of two opposing factors, i.e., the connective transport of the solute toward the membrane by the bulk motion of the solvent and the back diffusion of the solute away from the membrane surface due to the concentration gradient established near the phase boundary. Concentration polarization can be minimized by increasing the fluid shear rate to decrease the boundary-layer thickness at the membrane surface and to promote the mass transport of the solute away from the membrane surface. Turbulence promoters (e.g., feed spacers in spiral-wound modules and static mixers in tubular modules) have been used to decrease concentration polarization (Ho and Li, 1984; Bhattacharyya *et al.*, 1992; Kulkarni *et al.*, 1992; Davis and Grant, 1992).

As mentioned earlier, an emulsion liquid membrane process consists of four steps:

1. emulsification,
2. dispersion of the emulsion in contact with the external, continuous phase for extraction,
3. settling to separate the emulsion from the external phase, and
4. breaking the emulsion to recover the internal phase and the membrane phase.

Emulsification is carried out by the use of colloidal mills or static homogenizers. Dispersion and settling of the emulsion in contact with the external phase are conducted with mixer settlers and mechanically agitated column extractors followed by settlers, which are the same as those used conventionally for solvent extraction. An electrostatic coalescer is usually employed to break the emulsion (Ho and Li, 1992; Gu *et al.*, 1992).

Pretreatment of feed, such as the use of micron-sized filtration (about 0.5–5 μm), is usually needed to minimize solids buildup in membrane modules. Fouled membranes are normally cleaned by clean-in-place procedures that minimize downtime. Periodic reversal in flow direction may help to prevent particulates from buildup in membrane modules. Cleaning solutions, e.g., detergent and surfactant solutions and some solvents (e.g., isopropanol), are generally employed. A typical cleaning cycle may use 1 to 3 module volumes of the cleaning solution at an elevated temperature (about 50–80 °C) to circulate through the membrane system. The cycle may be carried out once every 1–4 weeks (Fleming and Slater, 1992; Ho and Sirkar, 1992a).

9.4 Important Industrial Applications

Applications include the following for eight commercialized membrane processes.

9.4.1 Gas Permeation Gas separation applications include hydrogen recovery, air (oxygen and nitrogen) separation, acid gas removal, and helium recovery. Hydrogen is recovered from hydroprocessing purge streams in refineries, from ammonia-plant purge streams and other petrochemical plant streams, and from synthesis-gas (H_2/CO) ratio adjustment. Air is separated into oxygen- and nitrogen-enriched streams. Oxygen-enriched air has usually been limited to medical uses, whereas commercial-scale uses (e.g., enhanced combustion) have yet to reach their full potential. The purity of the nitrogen-enriched stream ranges from 95 to 99% nitrogen, which is suitable for use as inert gas. Acid gases, e.g., CO_2 and H_2S, are removed from natural gas. A potentially large market for CO_2 separation membrane systems is the recovery of CO_2 from gas produced in oil fields where enhanced oil recovery via CO_2 injection is used. Helium is recovered from natural gas (Zolandz and Fleming, 1992).

9.4.2 Pervaporation Its industrial uses are primarily for the dehydration of organics such as ethanol and isopropanol and the removal of organics—e.g., ethanol from dealcoholization of beers, wines, and liquors, and trichloroethane from groundwaters and wastewaters (Fleming and Slater, 1992). As mentioned earlier, a new pervaporation technology has been developed recently to separate heavy catalytically cracked naphtha into an aromatics-rich permeate for mogas blending and an aromatics-lean retentate for distillate blending (Ho *et al.*, 1990, 1991; Exxon, 1994).

9.4.3 Dialysis The artificial kidney, hemodialysis, for purifying human blood is the major application of dialysis in the medical field. The principal solutes to be removed for chronic patients are urea, uric acid, creatinine, phosphates, and excess amounts of chloride (Ho and Li, 1984; Kessler and Klein, 1992). Industrial applications include the recovery of caustic soda in rayon processing and the recovery of spent acids from metallurgical liquors (Ho and Li, 1984).

9.4.4 Electrodialysis Its main applications include (Strathmann, 1992) these:

1. Desalination of brackish water. This application competes directly with reverse osmosis and multistage flash evaporation. For brackish water with relatively low salt concentration (less than 5000 ppm), electrodialysis is usually the most economic process.
2. Production of table salt from seawater. This application to concentrate sodium chloride up to 200 g/L prior to evaporation is used nearly exclusively in Japan.
3. Wastewater treatment. The main application is in processing rinse waters from the electroplating industry.

9.4.5 Reverse Osmosis Its applications include the following (Ho and Li, 1984; Bhattacharyya *et al.*, 1992):

1. Desalination of sea and brackish waters. Reverse-osmosis units have successfully desalted sea and brackish waters to potable quality (Ko and Guy, 1988).
2. Demineralization. Reverse-osmosis units have been integrated into the ultrapure water systems that supply water for electronics-manufacturing processes (Rautenbach and Albrecht, 1989).
3. Recycle operations in electroplating and electrocoating plants. This application not only recovers valuable materials, such as nickel, chromium, and cyanide, but at the same time eliminates water-discharge problems.
4. Processing wheys, juices, and effluents in the food industry. Cheese and cottage-cheese wheys are processed by the use of reverse osmosis, which is combined with ultrafiltration to separate the solutes into fractions containing mostly protein, lactose, and lactic acid as well as to concentrate each of these fractions. Reverse osmosis is also used to concentrate juices, milk, and syrup.
5. Wastewater treatment. This includes applications to treat dilute processing

streams in the pulp and paper and textile industries for recycling water and decreasing color, biochemical oxygen demand, and other objectionable components in effluents.

6. Water treatment in the power-generation industry. The primary uses are to produce boiler-quality water for steam generation and to treat wastewater effluents from cooling-water blowdown (Patra *et al.*, 1987).

9.4.6 Ultrafiltration Its applications include the following (Ho and Li, 1984; Cheryan, 1986; Kulkarni *et al.*, 1992):

1. Electrophoretic paint recovery. Ultrafiltration is used to process the paint by retaining the polymer resins and pigmented solids while allowing inorganic salts, water, and solvent to permeate through the membrane. The retained species are returned to the electropaint tank. The permeate is then used to rinse the freshly painted components as they emerge from the paint and to recover the drag-out excess paint (Michaels, 1981).
2. Protein extraction in food and dairy industries. Large protein molecules from cheese, casein whey, or skim milk are concentrated.
3. Juice clarification. Haze components are removed from fruit juices, e.g., apple juice (Blanck and Eykamp, 1986).
4. Waste-machining and metal-rolling emulsions. Ultrafiltration is used to concentrate emulsified machining and metal-rolling oils from an initial 2 to 10% to a level of 25 to 50%. The concentrated emulsion can be disposed of by burning and the water permeate discharged to the drain.

9.4.7 Microfiltration Its applications include the following:

1. Sterilization of pharmaceuticals: Microfiltration is used to sterilize pharmaceuticals, including water for injection, ophthalmic solutions, antibiotics, vaccines, etc. (Goel *et al.*, 1992).
2. Stabilization of beers and wines: Microfiltration is used increasingly in the beverage industry mainly for clarification and biological stabilization of beers and wines (Goel *et al.*, 1992).
3. Purification of process fluids in semiconductor manufacture: Both gaseous and liquid process fluids used to manufacture semiconductors and integrated circuits are purified by microfiltration to remove particles (Goel *et al.*, 1992).
4. Clarification of antibiotics: Microfiltration is used for the recovery and purification of antibiotics from fermentation broths (Mir *et al.*, 1992).

9.4.8 Emulsion Liquid Membrane Applications for ELM processes include these (Ho and Li, 1992):

1. Zinc removal from wastewater in the viscose fiber industry: This was first commercialized at Lenzing AG in Austria with a wastewater feed rate of 75 000 L/h (11 300 bbl/d), and zinc was removed from 200 to 0.3 mg/L (Marr and Draxler, 1992).
2. Phenol removal from wastewater in the chemical industry: This was commercialized at the Nanchung Plastic Factory in Guangzhou, China with a feed rate of 250 L/h, and phenol was reduced from 1000 to 0.5 mg/L (Marr and Draxler, 1992).
3. Cyanide removal from waste liquors in gold processing: This is being commercialized at the Huang-hua Mountain Gold Plant, near Tian-jin, China with a feed rate of about 600 L/h, and cyanide can be reduced from 130 to 0.5 mg/L (Jin and Zhang, 1990; Jin *et al.*, 1994).
4. Well control fluid for preventing well blowout and sealing loss zones in wells: The well control fluid is a pumpable water-in-oil emulsion containing clay particles in the oil phase. High shear at drill-bit nozzles breaks up the oil film and allows clay particles to contact water droplets, resulting in thickening the emulsion to very viscous paste for controlling the well in the drilling operation.

GLOSSARY

Absorbate: The solute component that is absorbed.

Absorbent: The liquid that absorbs solute components from the feed.

Adsorbate: The solute component that is adsorbed.

Adsorbent: The solid that adsorbs solute components from the feed.

Adsorption Isotherm: The isotherm that relates the fluid-phase adsorbate concentration to the amount of the adsorbate adsorbed per unit weight of the adsorbent.

BET: Abbreviation for the Brunauer-Emmett-Teller thermodynamic isotherm.

Boundary Layer: The region near a solid surface, e.g., a membrane, where fluid motion is affected by the surface.

Cake Filtration: Filtration in which the filtered solids are deposited on the surface of the filter medium, and the solid cake thus deposited quickly becomes the actual filter medium.

Chemical Potential: A measure of how the Gibbs free energy for a system changes when a specific amount of a component is added to the system. For membrane separation, the driving forces for transport through a membrane arise from a gradient of chemical potential or electrical potential. A gradient in chemical potential may be due to concentration gradient or pressure gradient or both.

Clarification: Process that removes relatively small amounts of particles and produces a clear effluent.

Composite Structure: Membrane structure in which the active layer is a coating on the support layer, or a coating is put on the active layer.

Concentrate: The portion of the feed solution that is retained on the high-pressure or -concentration side of the membrane. The terms "concentrate" and "retentate" are used interchangeably.

Concentration Polarization: When solvent transports through a membrane, the solute (or particulates, in microfiltration) left behind concentrates in a solution next to the membrane surface. Accompanying the increase in solute concentration at the interface are an increase in the osmotic pressure of the solution at the interface and the formation of a secondary membrane, resulting in flux reduction.

Counter Ions: The ions having the opposite charge to the fixed ions of an ion-exchange resin.

Deep-Bed Filtration: A deep bed of porous media such as diatomaceous earth (precoat) or fibrous material supported on a coarse filter is used to remove fine particles from a slurry. The particles to be removed are often considerably smaller than the pores of the filter medium and can penetrate a considerable depth before they are captured.

Dialyzate: The solution on the side of the membrane to receive the solutes to be removed from the feed on the other side of the membrane in dialysis.

Diffusion Coefficient (Diffusivity): A measure of the diffusive mobility of a solute in a solvent [see Eq. (11)].

Entrainer: A third component added to the system to effect the separation of an otherwise inseparable mixture by distillation.

Equilibrium: A state from which there is no tendency to depart spontaneously. An equilibrium state still has certain permissible changes or processes, e.g., heat transfer, work of volume displacement, and mass transfer across a phase boundary.

Extract: The phase comprising a solvent and the solute extracted from the feed.

Fluidized Bed: A bed in which the particles, e.g., ion-exchange resin beads, are held in suspension by the gas phase in contact with them.

Flux: Amount of fluid passing through the membrane.

Fouling: Phenomenon in which the membrane or equipment adsorbs or interacts in some manner with solutes in the feed stream, resulting in a decrease in performance (reduction in the membrane flux or equipment productivity). This phenomenon is usually irreversible and time dependent. In general, fouling effects can only be remedied by shutting down the system and cleaning the membrane or equipment by chemical means.

Integrally Skinned Structure: Membrane structure in which the active and support layers are formed in a single operation from a given material.

Ion-Exchange Capacity: Milliequivalents (meq.) of charge in an ion-exchange resin per unit weight (gram) of the dry resin.

Mass-Transfer Coefficient: A measure of the mobility of a solute, which is defined as the ratio of the solute flux to the solute concentration difference [see Eq. (12)].

Nucleation: Crystal formation initiated by the presence of a foreign body in the fluid or an irregularity in the wall of the container. The formation of a new solid phase in the bulk of a fluid phase without the involvement of a solid–fluid interface is called homogeneous nucleation; in the presence of surfaces, other than that of the crystal itself, such as those of the container, agitator, or foreign particles is called heterogeneous nucleation; or in the presence of the crystals of the crystallizing species itself is termed secondary nucleation.

Packed Column: Mass-transfer device filled with ceramic, metal, or plastic packings to facilitate mass transfer by increasing the contacting area. The packings are classified as random packings or structured packings. Random packings are dumped into the column onto a support plate whereas structured packings are shaped in the same size as the column with a regular structure.

Partition Coefficient: The ratio of the mole (or mass) fraction of the solute in the extract phase to its mole (or mass) fraction in the raffinate phase [see Eq. (8)].

Permeability: The ratio of the flux through the membrane to the concentration or pressure gradient across the membrane [see Eq. (17)].

Permeate: The portion of the feed that passes through the membrane.

PSA: Abbreviation for pressure-swing adsorption.

PSSA: Abbreviation for polystyrene sulfonic acid.

Reflux Ratio: The ratio of the condensate returned to the distillation column to the total condensate.

Rejection Efficiency: A measure of how well a membrane retains or allows passage of a solute [see Eq. (19)].

Relative Volatility: The ratio of the vapor pressure of two components.

Retentate: The portion of the feed that is retained on the high pressure or concentration side of the membrane. The terms "retentate" and "concentrate" are used interchangeably.

Separation Factor: The selectivity between species *i* and *j* is defined as the ratio of species *i* and *j* mole or mass fractions (or concentrations) in the extract (for solvent extraction) or in the permeate (for membrane separation), divided by their corresponding ratio of mole or mass fractions (or concentrations) in the raffinate, retentate, or feed [see Eqs. (10) and (18)].

Stage Cut: The ratio of the permeate flow rate to the feed flow rate.

Supersaturation: A measure of the quantity of solute present in the solution compared with the quantity, i.e., equilibrium solubility, which would be present if the solution were kept for a very long period of time with the solid phase of the solute in contact with the solution.

Theoretical Stage: A theoretical stage (defined at the existing or intended reflux rate) is considered as the unit stage that will produce a stage of equilibrium between the vapor and the liquid leaving the given stage.

Thickening: Process that increases the concentration of suspended particles in a slurry stream.

Tray Column: Gas–liquid contacting device, also known as mass-transfer device (usually in cylindrical shape) equipped with trays (plates). The tray columns are usually operated in countercurrent mode.

UNIFAC: Abbreviation for UNIQUAC (universal quasi-chemical activity coefficient) functional group activity coefficient, which is a thermodynamic group contribution method.

VLE: Abbreviation for vapor–liquid equilibrium.

VOC: Abbreviation for volatile organic compound.

Water-Transport Number: The ratio of mass of water transported per mass of dissolved solute dialyzed.

ZMS: Abbreviation for zeolite molecular sieve.

Works Cited

ASPEN-PLUS Reference Manual (1994).

Baczek, F. A., Emmett, R. C., Kominek, E. G. (1988), in: P. A. Schweitzer (Ed.), *Handbook of Separation Technologies for Chemical Engineers,* 2nd ed., New York: McGraw-Hill, p. 4-121.

Belfort, G. (1984), *Synthetic Membrane Processes,* New York: Academic.

Bennett, C. O., Meyers, J. E. (1982), *Momentum, Heat and Mass Transfer,* 3rd ed., New York: McGraw-Hill, pp. 224–246.

Bennett, R. C. (1984), in: R. H. Perry, D. W. Green (Eds.), *Perry's Chemical Engineers' Handbook,* 6th ed., New York: McGraw-Hill, pp. 19-24 to 19-40.

Bhattacharyya, D., Williams, M. E., Ray, R. J., McCray, S. B. (1992), in: W. S. Ho, K. K. Sirkar (Eds.), *Membrane Handbook,* New York: Chapman & Hall, pp. 263–390.

Blanck, R. G., Eykamp, W. (1986), *AIChE Symp. Ser.* **82** (250), 59.

Brown, P. R., Grushka, E. (1993), *Advances in Chromatography,* New York: Marcel Dekker.

Brunauer, S., Emmett, P. H., Teller, E. (1938), *J. Am. Chem. Soc.* **60,** 309.

Bungay, P. M., Lonsdale, H. K., dePinho, M. N. (1986), *Synthetic Membranes: Science, Engineering and Applications,* Dordrecht, Holland: D. Reidel Publishing Co.

Campbell, J. M. (1976), *Gas Conditioning and Processing: Absorption and Fractionation, Pumping, Compression and Expansion, Refrigeration, Hydrate Inhibition, Dehydration and Process Control,* Vols. 1 and 2, Norman OK: Campbell Petroleum Series.

Chapman, T. W. (1987), in: R. W. Rousseau (Ed.), *Handbook of Separation Process Technology,* New York: Wiley.

ChemCAD III Users' Guide (1993).

Cheremisinoff, P. N. (1995), *Processing Engineering Data Book,* Lancaster, PA: Technomic Publishing Co.

Cheryan, M. (1986), *Ultrafiltration Handbook,* Lancaster, PA: Technomic Publishing Co.

Chiang, S.-H., He, D. (1993), *Fluid/Particle Sep. J.* **6,** 64.

Code of Federal Regulations (1977), Title 21, §173.280, Washington, DC: U.S. Government Printing Office.

Combest, J., Corporation, K. (1994), *Fluid/Particle Sep. J.* **7,** 131.

Dahlstrom, D. A., Emmett, R. C., Klepper, R. P., Silverblatt, C. E. (1984), in: R. H. Perry, D. W. Green (Ed.), *Perry's Chemical Engineers' Handbook,* 6th ed., New York: McGraw-Hill, p. 19-53.

Davis, R. H., Grant, D. C. (1992), in: W. S. Ho, K. K. Sirkar (Eds.), *Membrane Handbook,* New York: Chapman & Hall, pp. 455–505.

Donald, M. B. (1958), *Chem. Eng. Practice* **6,** 473.

Dorfner, K. (Ed.) (1991), *Ion Exchangers,* Berlin: Walter de Gruyter.

Edwards, W. M. (1984), in: R. H. Perry, D. W. Green (Eds.), *Perry's Chemical Engineers' Handbook,* 6th ed., New York: McGraw-Hill, p. 14-1.

Exxon Research & Engineering Co. (1994), The ER&E Membrane Process, Refining Technology Brochure.

Fair, J. R. (1987), in: R. W. Rousseau (Ed.), *Handbook of Separation Process Technology,* New York: Wiley.

Fleming, H. L., Slater, C. S. (1992), in: W. S. Ho, K. K. Sirkar (Eds.), *Membrane Handbook,* New York: Chapman & Hall, pp. 103–159.

Ganetsos, G., Barker, P. E. (1993), *Preparative and Production Scale Chromatography,* New York: Marcel Dekker.

Gas Processors Suppliers Association (1987), *Engineering Data Book,* Tulsa, OK: Gas Processors Suppliers Association.

Goel, V., Accomazzo, M. A., DiLeo, A. J., Meier, P., Pitt, A., Pluskal, M., Kaiser, R. (1992), in: W. S. Ho, K. K. Sirkar (Eds.), *Membrane Handbook,* New York: Chapman & Hall, pp. 506–570.

Grace, H. P. (1956), *AIChE J.* **2,** 316.

Gregor, E. C., Walt, C., Mollet, J. R. (1992), *Fluid/Particle Sep. J.* **5,** 163.

Gu, Z. M., Ho, W. S., Li, N. N. (1992), in: W. S. Ho, K. K. Sirkar (Eds.), *Membrane Handbook,* New York: Chapman & Hall, pp. 656–700.

Ho, W. S., Li, N. N. (1984), in: R. H. Perry, D. W. Green (Eds.), *Perry's Chemical Engineers' Handbook,* 6th ed., New York: McGraw-Hill, pp. 17-14 to 17-35.

Ho, W. S., Li, N. N. (1992), in: W. S. Ho, K. K. Sirkar (Eds.), *Membrane Handbook,* New York: Chapman & Hall, pp. 595–655.

Ho, W. S., Sartori, G., Thaler, W. A., Dalrymple, D. C. (1990), Polyimide/Aliphatic Polyester Copolymers, U. S. Patent 4,944,880.

Ho, W. S., Sartori, G., Thaler, W. A., Dalrymple, D. C. (1991), Polyimide/Aliphatic Polyester Copolymers, U. S. Patent 4,990,275.

Ho, W. S., Sirkar, K. K. (Eds.) (1992a), *Membrane Handbook,* New York: Chapman & Hall.

Ho, W. S., Sirkar, K. K. (1992b), in: W. S. Ho, K. K. Sirkar (Eds.), *Membrane Handbook,* New York: Chapman & Hall, pp. 3–15.

Huang, R. Y. M. (Ed.) (1991), *Pervaporation Membrane Separation Processes,* Amsterdam: Elsevier.

Jacobs, L. J. (1984), in: R. H. Perry, D. W. Green (Eds.), *Perry's Chemical Engineers' Handbook,* 6th ed., New York: McGraw-Hill, p. 19-65.

Jin, M., Wen, T., Lin, L., Liu, F., Liu, L., Zhang, Y., Zhang, C., Deng, P., Song, Z. (1994), *Mo Kexue Yu Jishu* **14,** 16.

Jin, M., Zhang, Y. (1990), *Proc. International Congress on Membranes and Membrane Processes,* Chicago, IL, August 20–24, 1990, Vol. I, pp. 676–678.

Keller, G. E., II, Anderson, R. A., Yon, C. M. (1987) in: R. W. Rousseau (Ed.), *Handbook of Separation Process Technology,* New York: Wiley, pp. 644–696.

Kessler, S. B., Klein, E. (1992), in: W. S. Ho, K. K. Sirkar (Eds.), *Membrane Handbook,* New York: Chapman & Hall, pp. 161–215.

Kesting, R. E. (1985), *Synthetic Polymeric Membranes: A Structural Perspective,* 2nd ed., New York: Wiley, pp. 1–21.

Ko, A., Guy, D. (1988), in: B. Parekh (Ed.), *Reverse Osmosis Technology,* New York: Marcel Dekker, Chap. 5.

Kohl, A. L. (1987), in: R. W. Rousseau (Ed.), *Handbook of Separation Process Technology,* New York: Wiley.

Kohl, A. L., Riesenfeld, F. C., (1979), *Gas Purification,* 3rd ed., Houston: Gulf Publishing Co.

Koros, W. J., Pinnau, I. (1994), in: D. R. Paul, Y. Yampol'skii (Eds.), *Polymeric Gas Separation Membranes,* Boca Raton, FL: CRC Press, pp. 209–271.

Krause, S. (1993), *Int. Chem. Eng.* **33,** 355.

Kulkarni, S. S., Funk, E. W., Li, N. N. (1992), in: W. S. Ho, K. K. Sirkar (Eds.), *Membrane Handbook,* New York: Chapman & Hall, pp. 391–453.

Li, N. N. (1968), Separating Hydrocarbons with Liquid Membranes, U. S. Patent 3,410,794.

Li, N. N. (1978), *J. Membr. Sci.* **3,** 265.

Lonsdale, H. K. (1989), *J. Membr. Sci.* **43,** 1.

Marr, R. J., Draxler, J. (1992), in: W. S. Ho, K. K. Sirkar (Eds.), *Membrane Handbook,* New York: Chapman & Hall, pp. 701–724.

McCabe, W. L., Smith, J. C. (1956), *Unit Operations of Chemical Engineering,* New York: McGraw-Hill, pp. 324–376.

Mersmann, A. (1994a), *Crystallization Technology Handbook,* New York: Marcel Dekker.

Mersmann, A. (1994b), in: A. Mersmann (Ed.), *Crystallization Technology Handbook,* New York: Marcel Dekker, p. 79.

Michaels, A. S. (1981), *Chemtech* **(January),** 36.

Miller, S. A. (1984), in: R. H. Perry, D. W. Green (Eds.), *Perry's Chemical Engineers' Handbook,* 6th ed., New York: McGraw-Hill, p. 19-1.

Mir, L., Michaels, S. L., Goel, V., Kaiser, R. (1992), in: W. S. Ho, K. K. Sirkar (Eds.), *Membrane Handbook,* New York: Chapman & Hall, pp. 571–594.

Moyers, G. C. (1984), in: R. H. Perry, D. W. Green (Eds.), *Perry's Chemical Engineers' Handbook,* 6th ed., New York: McGraw-Hill, p. 17-3.

Noll, K. E., Gounaris, V., Hou, W. S. (1991), *Adsorption Technology for Air and Water Pollution Control,* Chelsea, MI: Lewis Publishers.

Ohe, S. (1989), *Vapor-Liquid Equilibrium Data,* New York: Elsevier.

Ohe, S. (1990), *Vapor-Liquid Equilibrium Data at High Pressure,* New York: Elsevier.

Ohlrogge, K., Wind, J., Behling, R.-D. (1995), *Sep. Sci. Technol.* **30,** 1625.

Paruit, B. H. (Ed.) (1982), *Illustrated Glossary of Process Equipment,* Houston, TX: Gulf Publishing Co.

Patra, R., Prabhakar, S., Misra, B., Ramani, M. (1987), *Desalination* **67,** 507.

Paul, D. R., Yampol'skii, Y. (Eds.) (1994), *Polymeric Gas Separation Membranes,* Boca Raton, FL: CRC Press.

Perry, R. H., Green, D. W. (Eds.) (1984), *Perry's Chemical Engineers' Handbook,* 6th ed., New York: McGraw-Hill.

Pinheiro, H., Cabral, J. M. S. (1993a), in: J. F. Kennedy, J. M. S. Cabral (Eds.), *Recovery Processes for Biological Materials,* New York: Wiley, pp. 67–95.

Pinheiro, H., Cabral, J. M. S. (1993b), in: J. F. Kennedy, J. M. S. Cabral (Eds.), *Recovery Processes for Biological Materials,* New York: Wiley, pp. 97–131.

Porter, M. C. (1990), *Handbook of Industrial Membrane Technology,* Park Ridge, NJ: Noyes Publications.

Randolph, A. D., Larson, M. A. (1971), *Theory of Particulate Processes,* New York: Academic.

Rautenbach, R., Albrecht, R. (1989), *Membrane Processes,* New York: Wiley.

Reid, R. C., Prausnitz, J. M., Polling B. E. (1987), *The Properties of Gases and Liquids,* New York: McGraw-Hill.

Robbins, L. A. (1984), in: R. H. Perry, D. W. Green (Eds.), *Perry's Chemical Engineers' Handbook,* 6th ed., New York: McGraw-Hill Book Co., p. 15–1.

Ruthven, D. M. (1984), *Principles of Adsorption and Adsorption Processes,* New York: Wiley-Interscience.

Rydberg, J., Musikas, C., Choppin, G. R. (1992), *Principles and Practices of Solvent Extraction,* New York: Marcel Dekker.

Sartori, G., Ho, W. S., Savage, D. W., Chludzinski, G. R., Wiechert, S. (1987), *Separ. Purif. Methods* **16,** 171.

Schwartzberg, H. G. (1987), in: R. W. Rousseau (Ed.), *Handbook of Separation Process Technology,* New York: Wiley.

Seader, J. D. (1984), in: R. H. Perry, D. W. Green (Eds.), *Perry's Chemical Engineers' Handbook,* 6th ed., New York: McGraw-Hill, p. 13-1.

Singh, G. (1979), in: P. A. Schweitzer (Ed.), *Handbook of Separation Technologies for Chemical Engineers,* New York: McGraw-Hill, p. 2-151.

Skelland, A. H. P., Tedder, D. W. (1987), in: R. W. Rousseau (Ed.), *Handbook of Separation Process Technology,* New York: Wiley.

Sourirajan, S., Matsuura, T. (1985), *Reverse Osmosis/Ultrafiltration Process Principles,* Ottawa, Canada: National Research Council Canada Publications.

Springer, A. M. (1993), in: A. M. Spriger (Ed.), *Industrial Environmental Control,* 2nd ed., Atlanta, GA: TAPPI, pp. 344–360.

Strathmann, H. (1992), in: W. S. Ho, K. K. Sirkar (Eds.), *Membrane Handbook,* New York: Chapman & Hall, pp. 217–262.

Streat, M., Cloete, F. L. D. (1987), in: R. W. Rousseau (Ed.), *Handbook of Separation Process Technology,* New York: Wiley.

Suzuki, M. (1990), *Adsorption Engineering,* New York: Elsevier.

Tien, C., Payatakes, A. C. (1979), *AIChE J.* **25,** 737.

Toyokura, K., Wintermantel, K., Hirasawa, I., Wellinghoff, G. (1994), in: A. Mersmann (Ed.), *Crystallization Technology Handbook,* New York: Marcel Dekker, p. 459.

Vermeulen, T., LeVan, M. D., Hiester, N. K., Klein, G. (1984), in: R. H. Perry, D. W. Green (Eds.), *Perry's Chemical Engineers' Handbook,* 6th ed., New York: McGraw-Hill, p. 16-1.

Wadsworth, E. M. (1987), in: R. W. Rousseau

(Ed.), *Handbook of Separation Process Technology,* New York: Wiley.

Wijmans, J. G., Baker, R. W. (1993), *J. Membr. Sci.* **79,** 101.

Zarzycki, R. (1993), *Absorption: Fundamentals & Applications,* New York: Pergamon Press.

Zolandz, R. R., Fleming, G. K. (1992), in: W. S. Ho, K. K. Sirkar (Eds.), *Membrane Handbook,* New York: Chapman & Hall, pp. 17–101.

SERVO AND CONTROL DEVICES

JAMES P. TREVELYAN, *Department of Mechanical and Materials Engineering, The University of Western Australia, Nedlands, Western Australia*

Introduction 1
1. **Introduction to Control Systems** 2
1.1 History of Control Engineering 2
2. **Thinking about Control Systems—the Basic Principles (Defining Standard Terms)** 2
2.1 Subtle Effects of Feedback Control—Is the Air Conditioning Colder in Summer? 4
2.2 Self-Filling Water Tank 4
2.3 Hydraulic Actuator 6
2.4 Electric Motor Positioner 7
3. **Control-System Performance** 9
3.1 Bandwidth 9
3.2 Transient Response—Simulating a dc Servo Drive .. 11
3.3 Proportional Control 12
3.4 PID Controller 12
3.5 Limits of PID Control—When to Use Other Methods 14
3.6 Fuzzy-Logic Control 14
3.7 Self-Tuning and Adaptive Controllers 14
3.8 Advanced Control Techniques 15
4. **Computers—from Performance Prediction to Control** 15
4.1 Computation Delay Time 15
4.2 Sensor Resolution 17
4.3 Output Resolution 17
5. **Some Important Limitations** 18
5.1 Steady Torque and Thermal Load on a Motor 18
5.2 Speed Limitations 19
5.3 Power-Supply Limitations 20
5.4 Sensor Location 20
6. **Faults in Control Systems—Unexpected behavior** 20
6.1 Runaway 20
6.2 Persistent, Steady Error 21
6.3 Unstable Transient Response 22
7. **Discrete Control Systems** 22
Appendix—Excel Spreadsheet Formulas for dc Servo Simulation 24
Acknowledgments 24
Glossary 24
Works Cited 26
Further Reading 27

INTRODUCTION

To understand how servo and control devices work, we need to think about the machines or processes they regulate. We can then think of a *control system** that comprises the control device and the regulated machine or process. When we drive a car, we form a control system comprising ourselves (controller) and the car on the road (process).

Control engineering is the technology of control systems; it is a pervasive technology. It surrounds us and we rely on it for many of our basic needs. Control systems keep electric generators turning at exactly the right speed to provide a constant frequency of alternating current in our electricity supply. Hydraulic and electrical control systems keep aircraft and transport systems running. Other systems ensure the continuity of our water supply, control the temperature and environment inside our buildings, and regulate the production processes that produce

---

*Important technical terms are introduced *in italics.*

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

the material goods we use in our everyday lives.

Yet control is even more pervasive than this, for our own bodies contain control systems. They maintain our internal and chemical environment at precisely the conditions needed to maintain our intricate biochemistry. Other systems keep us balancing upright, allow us to walk where we want to and to follow rapidly moving objects with our eyes, and even keep us breathing when we forget to.

James Lovelock in his now classic work on the Gaia theory has shown how biological control systems can be seen to regulate the temperature and physical composition of the Earth's environment. His work has created considerable controversy, as he claims that this degree of control is the essence of life itself. However, the controversy revolves around the meaning of the word "life" and the facts are not in dispute (Lovelock, 1987).

It is surprising, therefore, that some relatively simple concepts can be applied to understand the behavior of all of these diverse systems.

1. INTRODUCTION TO CONTROL SYSTEMS

1.1 History of Control Engineering

Much of our knowledge of control systems was developed during and immediately after the Second World War in the 1940s. Stimulated by the need for accurate and effective weapons, the advent of radar and navigation devices opened the possibility of automatic control for guns, missiles, and aircraft. The classical control theory of linear systems was essentially complete by the early 1950s. A *linear system* is one for which the response to two simultaneous inputs is the sum of the responses to each input presented separately. Few real control systems behave exactly like this, but linear theory can give a useful explanation for many systems. In the years since, control theory has been modified and extended to cope with more complex systems and some nonlinear systems. However, a comprehensive theory covering nonlinear systems in the way that classical control covers linear systems has so far been elusive.

In the 1960s control theorists turned their attention to other applications. As the study of economics and business became more and more based on quantitative methods, control theory seemed to become more and more applicable. Governments started to use these ideas to control their economies. Unfortunately, control theorists knew little about the human values intrinsic to government and economic activity, and neither politicians nor economists cared much for the systematic analysis and mathematics that provide the basis for control theory. We are still encumbered with the consequences of these misunderstandings in the way our governments attempt to control economies. However, in their attempts to deal with the uncertainties of economic systems, control theorists developed some remarkably useful tools for systems with random variability.

Since the 1970s, computers have been used to extend greatly the performance and reliability of simple control systems. Since computers are so reliable, complex control systems can now be built. However, in the 1990s there is a big gap between control theory and application. Complex control systems are feasible in theory, but are seldom used in practice, particularly where safety is critical. We do not yet understand how to design complex control systems that are absolutely safe. Therefore, designers often aim to simplify control systems, at the expense of performance, so that safety can be assured.

2. THINKING ABOUT CONTROL SYSTEMS—THE BASIC PRINCIPLES (DEFINING STANDARD TERMS)

Most electric ovens have a control knob to set the internal temperature. The control device is often a temperature-sensitive (thermostat) switch that simply turns the heating element on if the temperature is too low and off when the temperature is too high (Fig. 1).

Many controllers work on a similar principle. The "switch on" temperature is a little less than the "switch off" temperature. If this difference is small, given constant heating and cooling rates, the power will need to be switched on and off rapidly. Usually the difference is a few degrees so the power is switched every minute or so. The power will

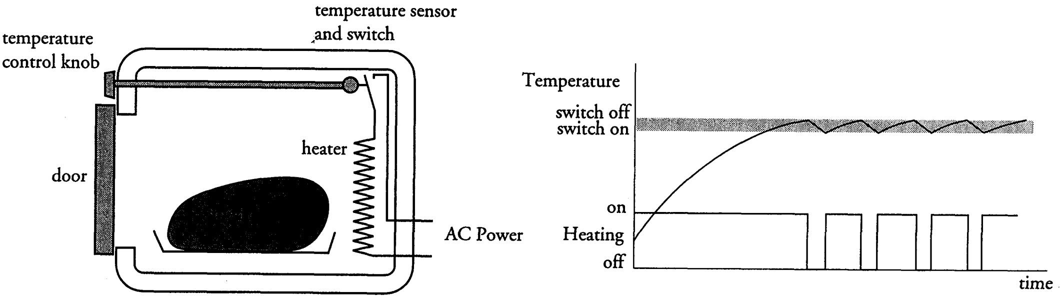
temperature sensor
and switch
temperature
control knob
heater
door
AC Power
Temperature
switch off
switch on
on
Heating
off
time

(a) Oven schematic

(b) Oven temperature changes

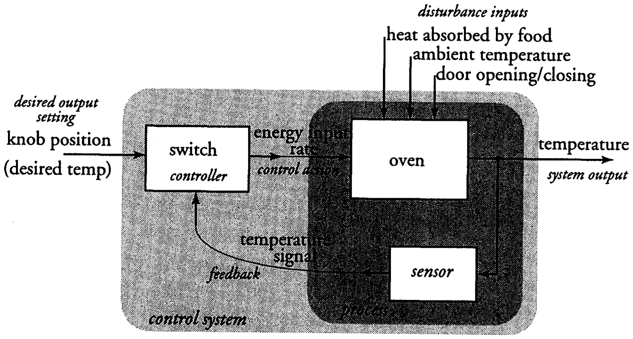
disturbance inputs
heat absorbed by food
ambient temperature
door opening/closing
desired output
setting
knob position
(desired temp)
switch
controller
energy input
rate
control action
oven
temperature
system output
temperature
signal
feedback
sensor
control system
process

(c) Block diagram of control system

FIG. 1. An electric oven: a common example of a control system. The graph (top right) shows how the temperature will change with time after the oven is switched on. The block diagram (lower) represents the action of the control system in a systematic format explained in the text.

be on for a greater proportion of the time if a high temperature has been set, since more heat is lost through the walls.

This is a simple and common example of a control device, but it illustrates some important principles.

1. In any control system, we must define which system property we need to regulate. In this case it is temperature. We might need to regulate more than one property. Each property we want to regulate is usually called an *output state variable,* or *system output.*
2. Some machines require only two, or a small number of, possible system-output values. These control systems are often called *discrete control systems.*
3. We must be able to observe or measure the system output. This is known as the *observability* principle. Sometimes, we can only take indirect measurements of the system output. For example, the temperature of materials heated to extremely high temperature might only be observable by measuring heat transferred to nearby objects. In other situations, measurements of the system output might include large random errors.
4. There must be some way of changing the state of the machine or process in order to change the system output. In this example, we change the oven state by adjusting the electric power to the heating element. This is called the *control action* (again, there may be more than one control action). This is a fundamental principle known as *controllability.* A process may not always be controllable. For example, if the temperature outside the oven is higher than the set temperature, the oven temperature cannot be controlled by the control action (electric heating).
5. There are usually several factors other than the control actions that affect the system output. In the case of the oven, heat absorbed by the food being cooked,

heat lost through the door opening or through the wall insulation, and the outside temperature all affect the system output, and cannot be controlled. These are called *disturbance inputs.*

As in the oven example, many controllers adjust the control action in response to measurements of the system output. This is known as *feedback control.* The oven controller adjusts the control action (electrical heating) in response to the difference between the system-output measurement (internal temperature sensed by the thermostat switch) and the *desired output* setting (position of oven temperature knob). A stable control system must use negative feedback; thus the control action must be decreased if the system output is too high.

We can visualize a control system with an abstract diagram that shows the process, the controller, and the feedback sensor as separate blocks. The links between them are abstract variables and do not necessarily represent material or energy flow, as a process flow diagram might portray. From this we can see that a feedback-control system forms a loop. Thus feedback control is also known as *closed-loop control.*

Most control systems use feedback. However, this is not necessarily the case for all control systems. Some control systems use *open-loop control* where there is no feedback during the operation of the control system. There is no need to measure the system state while the system is operating. However, such a system can only be designed if we know the process characteristics accurately, and the disturbance inputs either are known or affect the system output only slightly. Open-loop control systems can be simpler and more stable, but feedback control is often necessary to achieve the required accuracy.

2.1 Subtle Effects of Feedback Control—Is the Air Conditioning Colder in Summer?

Feedback-control systems sometimes behave contrary to common sense. They require very careful thinking and systematic methods of analysis. One example will illustrate this.

Many of us have experienced the frustration of feeling so cold inside an air-conditioned building in summer that we need extra clothes or even a heater to feel comfortable. Why is this so? Common sense says that it must be warmer on the inside if the weather is hotter on the outside, and if you feel cold it is just because you have become adjusted to the hot weather outside.

The air conditioner pumps cold air into the living space until the measured air temperature reaches the desired level set by the thermostat knob. The air enters the living space and absorbs heat coming through the walls and ceiling, and heat given off by people in the building. Even with air circulation, the temperature varies from one part of the building to another. The temperature sensor, located high on the wall, is in one of the warmer spots. Therefore, the temperature at the sensor is slightly higher than the temperature where the people are, say 1° in winter.

In summer, with more heat coming through the walls, the temperature difference between the center of the room and the sensor on the wall is greater than in winter—the sensor may be 5° warmer in summer. Therefore, when the room has been cooled so that the measured temperature is correct, the actual temperature at the center of the room will be 4° lower than in winter. So your senses may not be deceiving you.

We can learn from this that the *sensor location* is of critical importance. This lesson will be reinforced later.

2.2 Self-Filling Water Tank

Almost every toilet has a self-filling water tank that uses a simple control device to regulate the level of water stored in the tank (see Fig. 2). The water inlet valve is operated by a lever with a float on the end. As the water level rises, lifting the float, the lever turns and gradually closes the valve. Eventually, the water level rises high enough to close the valve and no more water enters the tank.

We can predict the behavior of this kind of tank by making some simple assumptions about the valve. Given the water level h and the full level $h_{\max}$, assume that water inflow is proportional to the drop in level below $h_{\max}$ until it is fully open at h_1 below $h_{\max}$. Thus,

$$q_{\text{in}} = q_{\max} \quad \text{if} \quad h < h_1$$

(valve fully open), and

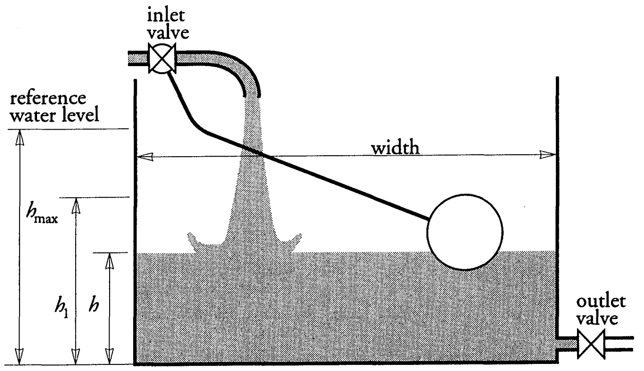

FIG. 2. Self-filling water tank used for toilets and animal drinking troughs (refer to text).

$$q_{in} = k(h_{max} - h) \quad \text{if} \quad h \geq h_1$$

(valve part open), where

$$k = q_{max}/(h_{max} - h_1).$$

Knowing the surface area A of the water in the tank (assume vertical sides so that the area is constant), calculate the level change over a short time interval δt by assuming the net inflow does not change:

$$\delta h \approx \delta t(q_{in} - q_{out})/A.$$

Given an initial water level h_0 and the outflow q_{out} as a function of time, we can repeat these computations to calculate how the water level in the tank changes at each time step. A spreadsheet is convenient for this and a drawing package can graph the results. Excel calculated the results drawn by Illustrator in Fig. 3 (Microsoft, 1994; Adobe, 1994). A sample of an Excel spreadsheet for simulation is in the Appendix.

Initially the tank is empty, and for a while no water leaves the tank. It fills at a constant rate until the level reaches h_1 and then at a decreasing rate until the tank is full. If all the water is flushed from the tank, the level drops again. Later we see what happens in the case of a leak—the steady-state level is less than h_{max}. A control system like this exhibits a *transient* response, when the level is changing with time, and later a *steady-state* response when the level is effectively constant.

Calculating the behavior of the tank like this is known as *simulation*. The equations represent a mathematical model of the tank and, in effect, we have solved a differential equation that governs its behavior.

We can build a block diagram of the tank using the model (Fig. 4). The system output that we want to control is the water level in the tank. The desired output is represented by the arrangement of the float on the lever. If we bend the float lever, we will change the desired output. This method is often adopted by plumbers or farmers who maintain these tanks. The process is the tank itself and, perversely, the outflow of water is a disturbance

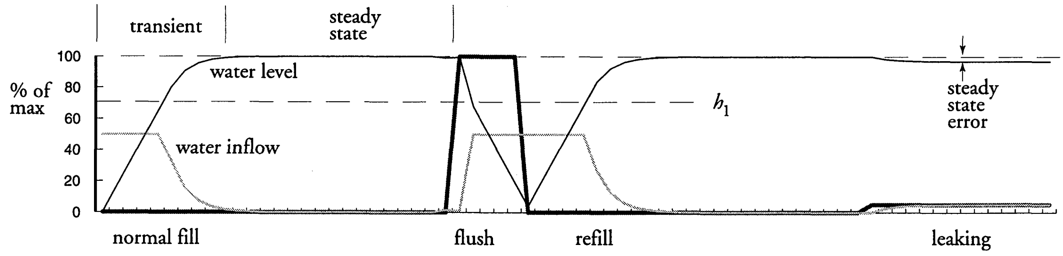

FIG. 3. Graph of tank behavior, predicted by simulation. The graph shows how the water inflow, outflow, and level vary with time. Each is expressed as a percentage of maximum level or flow rate.

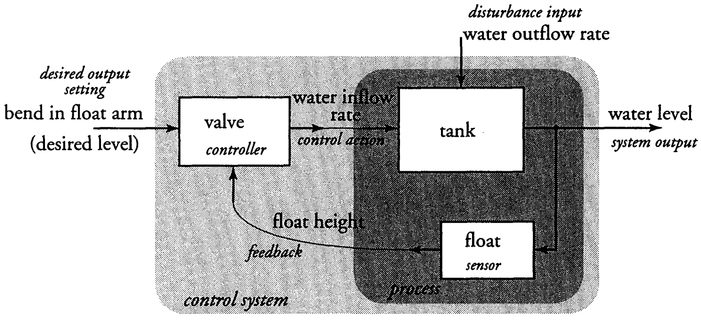

FIG. 4. Control block diagram of self-filling water tank in same systematic format as Fig. 1(c). Notice that the system output is not the water from the tank, as a process flow diagram might indicate. The system output is what we want to control, i.e., the water level.

input. This is because the rate of outflow is set outside the system, but it has an effect on the tank (process). The feedback sensor is the float, and the lever and control valve can be considered as the controller. The control action, therefore, is the water inflow rate.

Most introductory texts show how systems can be modeled using block diagrams and differential equations, and then how the solutions of these equations can be characterized in general terms (see any control textbook such as D'Azzo and Houpis, 1981). Linear control theory stems from this approach, and not only predicts the response of a linear system, but also shows how to design a controller to provide the desired response.

However, most real control systems are nonlinear, including this simple example when the water level is less than h_1 or greater than $h_{\max}$. Practical performance limits must be taken into account, and the most effective way to do this is simulation as shown above. Later we will see when linear analysis can be helpful and when simulation is needed.

2.3 Hydraulic Actuator

Hydraulic servo-actuators are used extensively in aircraft to move the flight controls because they are fast, light, powerful, and, under the right conditions, 100% reliable. When the correct seals are used, friction is low and the oil does not leak. Similar actuators move robots in response to instructions from a control computer; the sheep-shearing robot is a good example (Trevelyan, 1992). A simple actuator can develop controlled forces of up to several tonnes in a few thousandths of a second. The position of a load weighing several hundred kilograms can be changed within a few tens of milliseconds.

The diagram (Fig. 5) shows the basic principles. A voltage proportional to the required position of the piston is supplied to an amplifier. Another voltage is generated by the potentiometer—this voltage is proportional to the actual position of the piston. The difference between the two voltages (which is proportional to the error in the piston position) is amplified and a current proportional to the voltage difference is supplied to the servo valve. The servo valve directs oil flow to either end of the cylinder to move the piston. The oil flows at a rate proportional to the current in the valve coils—thus the piston moves toward the required position at a velocity that is proportional to the error between the required position and its current position.

The servo valve is the key to understanding this impressive device, for it uses an electric current of approximately 10 mA to regulate an oil flow of up to 120 liters per minute at a pressure of up to 35 MPa. (The high-pressure oil is supplied from a separate pump unit.) Typical high-performance servo valves respond in 2 to 4 ms.

If we were to look more closely at a servo valve, we would find that it contains a control system of its own. This control system causes the valve spool to move in proportion to the electric current, thus influencing the oil flow. Again, the amplifier contains a controller to regulate the electric current so that it is proportional to the voltage difference. Thus one control system might consist of several control systems in its component parts. Typical costs are shown in Fig. 5.

Parts of the valve must be built to extremely fine tolerances—most notably the

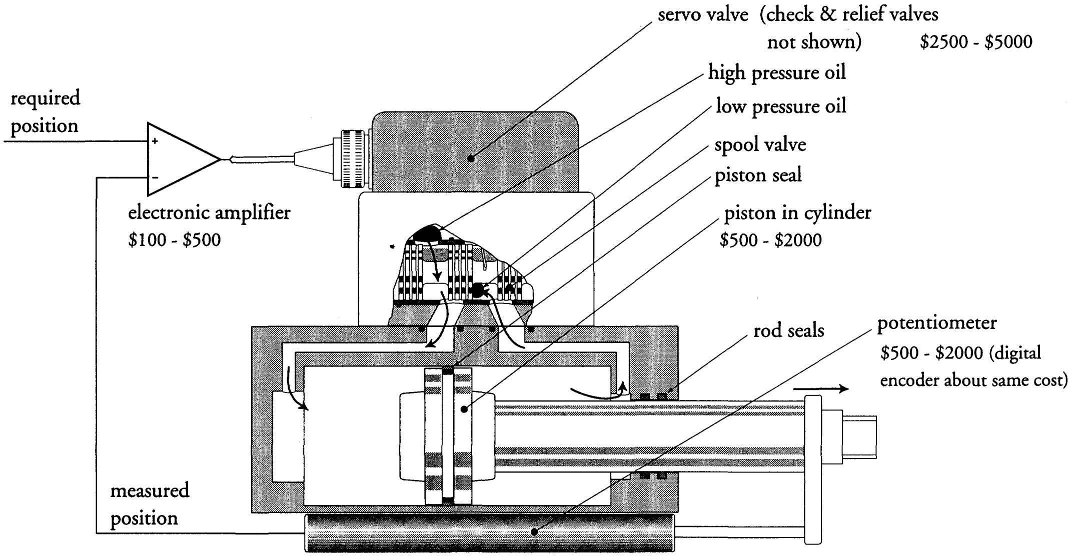

FIG. 5. Hydraulic servo-actuator. High-pressure oil supplied by a pump (not shown) is directed to one end of the cylinder by the servo valve, and the oil from the other end returns through the same valve to the oil-storage tank. A linear potentiometer measures the piston position, but there are many alternatives such as electromagnetic sensors (LVDT), magnetostrictive sensors, and digital optical encoders. Refer to text for further explanation.

clearance between the spool and the valve liner insert. The mating surfaces are highly polished, and most valves incorporate some form of filtration to minimize the chance of contaminating particles in the oil causing damage to the surfaces. The slightest damage can cause a dramatic failure of a hydraulic control system; see Sec. 6.1. However, by following appropriate design approaches, one can eliminate contamination and achieve very high reliability.

2.4 Electric Motor Positioner

Small electric motors are widely used for moving parts of a machine, such as the lens in an autofocus camera, which is moved to keep the image sharply focused. Electric motors, using digital encoders for position and speed feedback, are used for some of the most precise position controllers in computer-controlled instruments, machine tools, and robots (Fig. 6). Electric drives are popular because they are extremely reliable, repeatable, and precise and are easy to maintain. The power-to-weight ratio of hydraulic actuators can be 10 times greater, but they are less accurate unless expensive noncontact sealing arrangements are used.

The most direct way to control an electric motor exploits the relationship between current, force, and magnetic field strength. In a constant field, the sideways force is proportional to the electric current in a wire. Thus, the current in a permanent-magnet motor (a common type used for servo applications) generates a proportional torque.

If we use a transistor to regulate current directly, considerable heat is liberated at the transistor. Designing transistors that can withstand the effects of this heat is difficult, and so most motor controllers use a *switching controller* (or more fully, a pulse-width–modulated switching controller) (Bradley *et al.*, 1991). An electronic semiconductor switch does not dissipate much heat at all. A dc power source is rapidly switched electronically from one direction to the opposite in response to pulses (Fig. 7). If the "on" time is the same as the "off" time, the average current is zero. If the "on" time is longer than the "off" time, then the average current will be positive in the direction shown; otherwise it will be negative. Because the inductance of the motor is large, the current takes some time to change. Thus, switching the supply at a frequency of 15–40 kHz provides effective control of current without dissipat-

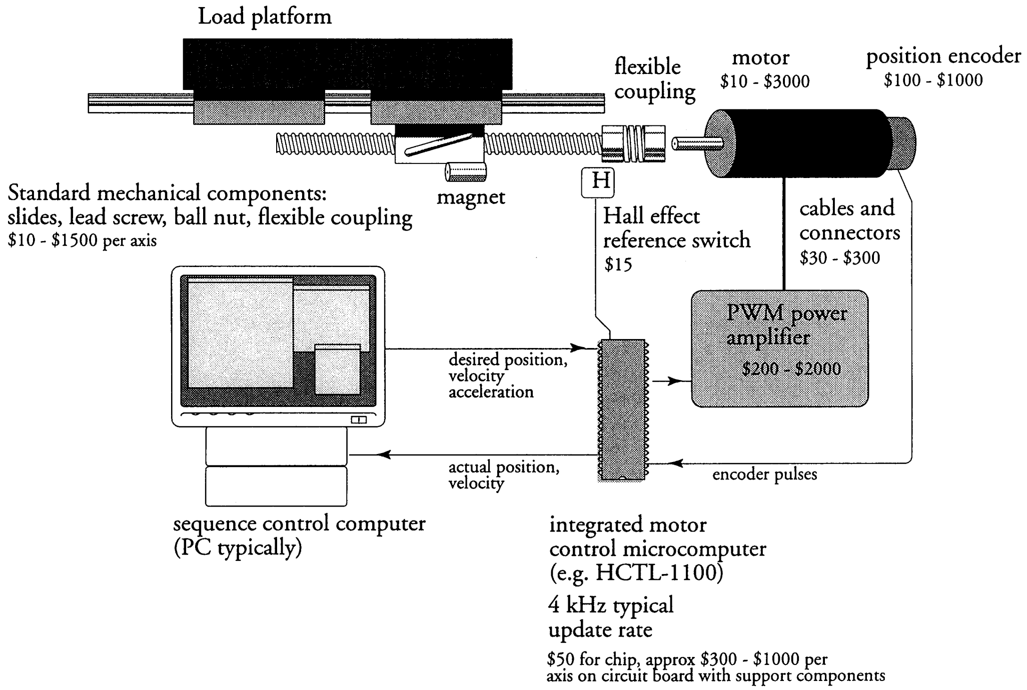
Load platform
flexible coupling
motor
$10 - $3000
position encoder
$100 - $1000
Standard mechanical components:
slides, lead screw, ball nut, flexible coupling
$10 - $1500 per axis
magnet
H
Hall effect
reference switch
$15
cables and
connectors
$30 - $300
PWM power
amplifier
$200 - $2000
desired position,
velocity
acceleration
actual position,
velocity
encoder pulses
sequence control computer
(PC typically)
integrated motor
control microcomputer
(e.g. HCTL-1100)
4 kHz typical
update rate
$50 for chip, approx $300 - $1000 per
axis on circuit board with support components

FIG. 6. Typical electric servo positioner built from standard components as used in robots and machine tools.

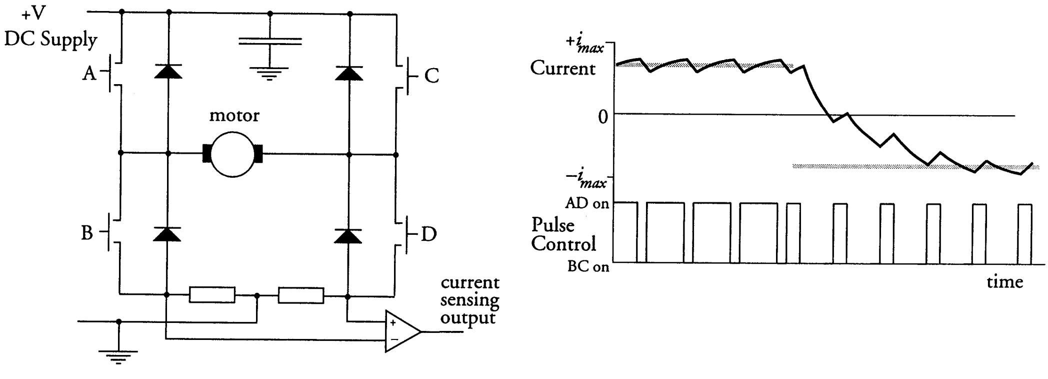
+V
DC Supply
A
B
C
D
motor
current
sensing
output
+
-
$+i_{max}$
Current
0
$-i_{max}$
AD on
Pulse
Control
BC on
time

(a) PWM Amplifier Circuit

(b) Pulse width current control: changing the pulse width changes the average current.

FIG. 7. Pulse-width–modulated power amplifier for a permanent-magnet electric motor. Electronic switches alternate the voltage applied to the motor. We control the average motor current by changing the relative duration of the forward and reverse voltage pulses. The graph shows a typical current variation with time. The pulse frequency is typically 15 kHz or greater. The average current is measured and used as a feedback signal for current control. Diodes protect the electronic switches from induced voltage surges. This circuit configuration is often called "H bridge" because the layout resembles a letter H.

ing heat in the switching transistors. The average current can be measured electronically, and then used to adjust the pulse width, thus the term *pulse-width modulation* (PWM).

Notice how there is close similarity between pulse-width modulation and the simple switching thermostat control of an electric oven. The essential difference is that the frequency of pulse-width modulation is usually constant. We achieve apparently continuously variable control of current (in this instance) by using simple on/off control at a "microscopic" level.

The most accurate and convenient means of measuring the position of a motor shaft is an optical encoder (Fig. 8). In its most common form, it provides two pulse streams as the shaft turns (usually from 100 up to 20 000 pulses or more per revolution). One pulse stream leads the other by one-quarter of a cycle (phase quadrature), and so the direction of motion can be detected. By first moving the load to a reference position (usually a reference proximity switch) and then counting the pulses, allowing for direction changes, we can measure position to within (at least) one-quarter of a pulse period. Absolute encoders are available at greater cost for comparable resolution and are increasingly used in machine tools and robots because there is no need to move to a reference position first. Resolvers are also used and provide comparable accuracy.

A single microcomputer can be used to receive the desired motor position and/or speed from a master control computer, count the pulses, and produce the pulse train used to switch the power supply.

The cost of electric servo systems ranges from a few dollars (for radio-control models) to thousands of dollars, depending on speed of response, torque, power, and accuracy. A typical accurate instrumentation-positioning servo might cost between $2000 and $10 000. An instrumentation servo can be built for less if skilled electronics and mechanical technicians are available to help, but 2–6 months full-time effort should be allowed to build one from components without extensive prior experience. Unfortunately, few common standards have been adopted by component suppliers and so linking them together is not easy.

3. CONTROL-SYSTEM PERFORMANCE

3.1 Bandwidth

One of the most important concepts to understand is *bandwidth*, or, in more detailed terms, the *frequency response* of a con-

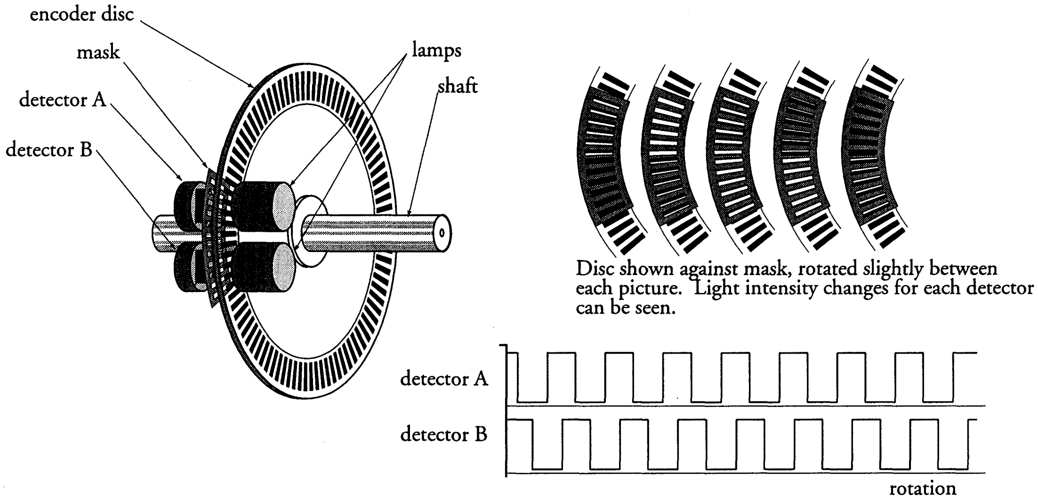

Disc shown against mask, rotated slightly between each picture. Light intensity changes for each detector can be seen.

FIG. 8. Incremental optical encoder of a type often used for measuring the motor shaft position and speed in low-cost servos. The two detectors produce pulses as the shaft rotates and the light intensity changes. The mask has two sets of slots arranged such that pulses from the detectors are 90° out of phase (see sequence at top left). The detectors incorporate threshold logic to convert light intensity changes to an on/off electrical signal. Many encoders also provide an additional detector to provide a reference pulse once per revolution.

trol system, process, sensor, or controller. One way to visualize this is to imagine what happens if you drive a car over a road with regular undulations of a given height and spacing (say 3-cm height, 30-cm spacing). If you drive regularly on unsealed roads, you may have encountered this yourself.

Driving very slowly, the car rides up and down slowly and follows the undulations.

Driving fast, the wheels follow the undulations but the springs absorb the wheel motion and the car body hardly moves at all (but you will feel the throbbing of the wheels going up and down).

At some speed between these extremes, the car bounces *in resonance* with the undulations, and may bounce as much as two or three times the undulation height if the shock absorbers are not working well.

The frequency response of a process defines the *amplitude ratio* (ratio of output amplitude to input) and *phase angle* for a range of input frequencies. A frequency response is usually plotted logarithmically. For a typical control system, the frequency response resembles the graph in Fig. 9, which shows the idealized linear response predicted by theory, and a typical measured response. Frequency-response measurements can be made with specialized equipment such as a frequency analyzer, commonly used for experimental studies of vibrations or audio equipment (D'Azzo and Houpis, 1981).

A typical control system responds with a unit amplitude response and zero response lag (i.e., the output almost exactly follows the desired value) up to a certain frequency limit, beyond which the system responds less and less. The phase angle (negative, meaning a response lag) rapidly increases too. The bandwidth, $\omega_{\max}$, is usually interpreted as the range of frequencies for which the response is no less than 70% of the input and the phase lag is less than 90°.

The *response time* of a system is the time needed for the system output to travel the first 63% of the desired change and is roughly equivalent to $1/\omega_{\max}$ where $\omega_{\max}$ is in radians/second (2π radians per second $=$ 1 hertz).

Thinking about bandwidth is important because it is difficult to design a control system for which the response time of the system is significantly less than its components.

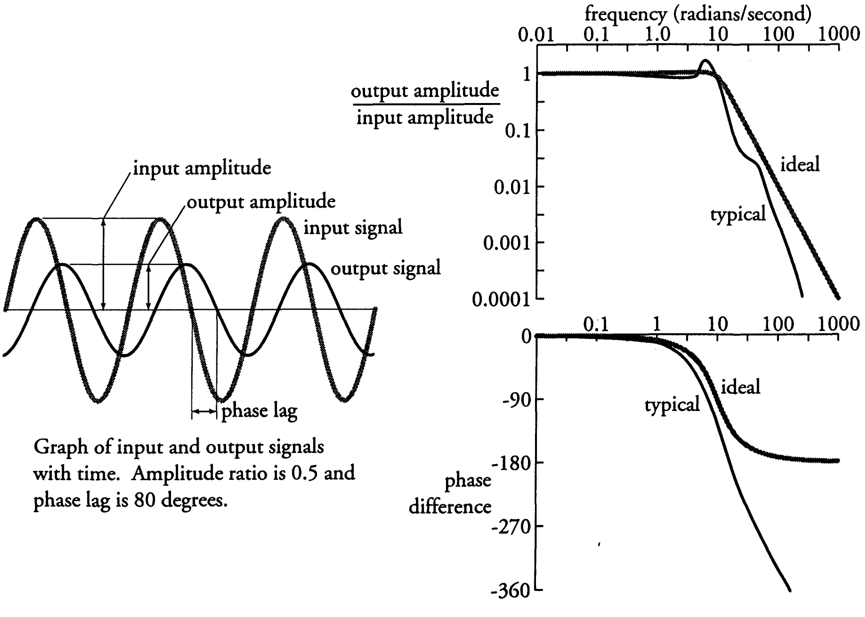

Graph of input and output signals with time. Amplitude ratio is 0.5 and phase lag is 80 degrees.

FIG. 9. Graph of typical control-system frequency response. The phase of a real system lags more than theory predicts because of factors such as computation delay in the controller (which is typically a small microprocessor), transport delays in fluids or even solid members where motion is limited to the speed of sound, mechanical backlash, and friction.

This is only possible if the characteristics of the components can be accurately predicted, either from theory or experimental studies.

Take, for example, the control of a space-robot manipulator. Factory robots are built with rigid links that do not deform significantly under load. Therefore, the position of the end of the robot can be controlled by controlling the relative positions of the links, typically using electric servos to drive rotary joints to known angles. Rigid links are usually heavy, and so robots carry loads that are small compared with their weight. Robots with flexible links may be much lighter, and very large robots for use in space cannot be made rigid. To control a flexible robot, we must compensate for deflection under load and eliminate oscillations.

A large space robot might oscillate at a natural frequency of one cycle every 10 or 20 seconds when carrying a large load (0.3–0.6 rad/s). This implies that it will be difficult to achieve a control bandwidth of 0.03–0.06 rad/s and therefore a realistic response time might be 30 s. Faster control of flexible robots has only been achieved under idealized laboratory conditions, and requires accurate modeling of the arm and intricate computations.

The sheep-shearing robot provides a useful illustration of the importance of bandwidth. This gives us an insight into a simple performance measure that predicts the maximum speed for tracking an unknown surface using a sensor on the robot's tool.

For a robot to follow an unknown surface, a robot has to guide its end effector over the surface using information from one or more sensors. Figure 10 shows some simulations of surface following with a position-control system. The robot end effector has a distance sensor, and approaches the surface at a constant speed, traversing the surface toward the right. The gray horizontal line shows the desired surface tracking distance δ. If we are not to hit the surface, this represents the maximum trajectory error allowed. This arrangement performs best if the sensed displacement error is limited to $\pm\delta$. For stable behavior the approach velocity must be limited.

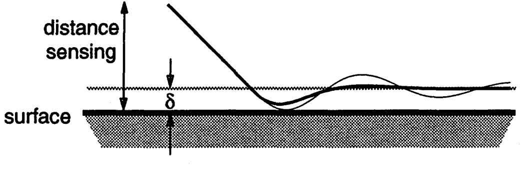

FIG. 10. A robot approaches an unknown surface using a sensor mounted on its tool or end effector to measure the distance error. The maximum approach speed must be limited to ensure that the robot does not hit the surface.

A remarkably simple relationship predicts the minimum bandwidth required of the control system, ω (rad/s):

$$\omega \geq v/\delta,$$

where v is the maximum approach speed perpendicular to the surface. This simple measure can be applied in many different contexts where sensor outputs modify planned robot movements.

3.2 Transient Response—Simulating a dc Servo Drive

We have seen how control systems exhibit two different types of behavior after a change in the desired output setting. First there is a transient while the process is adjusted to the new setting, and then the system settles to a new steady state. Clearly, the transient should be as quick as possible and *stable:* There will be little or no tendency for the output to overshoot or oscillate about the desired value.

Not all processes are easy to control and, as we saw earlier, correct analysis of a feedback-control system requires great care in thinking. It helps to follow the basic principles outlined in Sec. 2. A dc motor servo is a useful example.

Let us take, for example, a typical small rare-earth permanent-magnet motor. These specifications are based on a small dc servo motor (Maxon, 1994): 10 W rating, 9 V nominal voltage, 585 rpm/V back EMF, 16.3 mN m/A torque constant, max continuous current 1.5 A, 1.2 Ω coil resistance, rotor inertia 9.38 g cm^2 ($\times 10^{-7}$ to convert to kg m^2), mechanical time constant 4.24 ms, inductance 0.12 mH.

For a given current input i (amperes) the motor produces torque $\tau = 0.0163i$ N m. The inertia of the motor is I kg m^2 (1 kg m^2 $= 10^7$ gm cm^2) and the load equivalent inertia is I_L, and so the acceleration is $\alpha =$

$(0.0163i - \tau_L)/(I + I_L)$, where τ_L is the load torque (friction, steady load, etc.).

The model of the motor can be set up in a sequence of three steps that calculate the behavior over a small time interval δt. The calculation is repeated over enough time steps to calculate the response of the control system. At each given time step (j),

$$\text{shaft acceleration } \alpha_j = (0.0163i - \tau_L)/(I + I_L); \quad (1)$$

$$\text{shaft speed} \quad \omega_j \approx \omega_{j-1} + \alpha_j \delta t; \quad (2)$$

$$\text{shaft position} \quad \theta_j \approx \theta_{j-1} + \omega_j \delta t. \quad (3)$$

Step (2) above calculates the change in shaft speed over the time interval, $\alpha_j \delta t$, and adds this to the previous value of shaft speed to obtain the new speed. Step (3) does the same for shaft position. These calculations are not exact but good enough for the purpose here. More precise methods are only useful if the parameter values are known exactly, which is seldom the case.

There are some excellent simulation and design packages for control systems (such as MATLAB with a Control Toolkit, or Matrix-X) but an exposition of their use is beyond the scope of this article. A spreadsheet, once again, is quite sufficient to explore some stability problems.

3.3 Proportional Control

The most common controller arrangement is simply to use the difference between the system output and desired setting, or error, multiplied by a fixed constant called *gain*. The desired shaft position is θ_d and we arrange our simple controller to determine the value of motor current by

$$\text{current} \quad i_j = k(\theta_d - \theta_{j-1}). \quad (4)$$

Notice the negative sign for θ_j—this is the negative-feedback term. The term $\theta_d - \theta_{j-1}$ is the position error. The gain is k. Notice that the gain is also a unit conversion from radians to amperes.

If, for the sake of this example, we set the initial position and speed of the motor at position zero, the gain 5.0, the desired position 0.5, and the load torque 0.002, we obtain the response shown in Fig. 11. Notice how the shaft position oscillates in an unstable manner. Usually we can improve the system stability by decreasing the gain. However, this type of control system is still unstable, but with a decreased frequency. Not all control systems behave like; this many are quite stable with proportional control.

There is one important detail. The length of the time step, δt, needs to be chosen with some care. It should be short enough to represent the shortest response time in the system by about 20 steps. If there are large changes in each variable from one step to the next, reduce the time step. Check that the results do not change much if the time-step value is altered.

3.4 PID Controller

The behavior of the proportional motor-control system above can be modified by using a different controller. The most common form of process controller used industrially is a PID (proportional + integral + deriva-

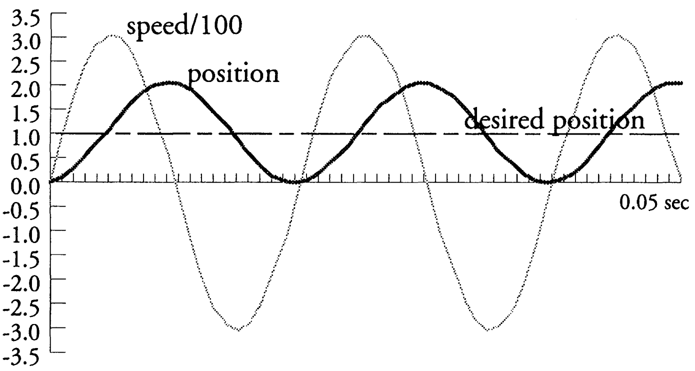

FIG. 11. Simulated response to a step change in desired position with a proportional controller.

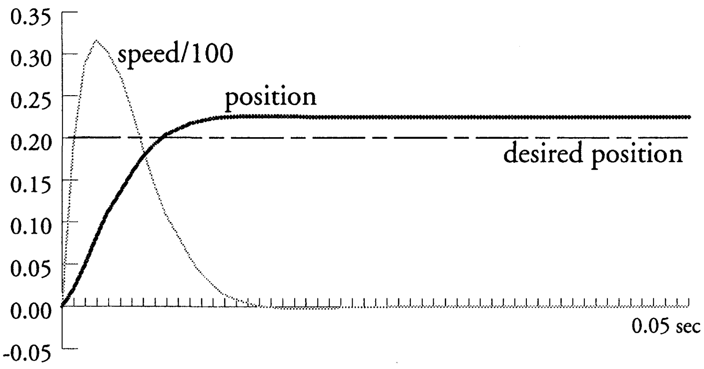

FIG. 12. Proportional + derivative control for an ideal system. Notice the steady-state error remaining after the transient.

tive) controller that combines three forms of the error variable used above:

1. a proportional component (as used above),
2. the derivative of the error, and
3. the integral of the error.

To do this, step (4) above is rewritten as two steps: integrate the error,

$$e_j = e_{j-1} + \delta t(\theta_d - \theta_{j-1}); \quad (4a)$$

$$i_j = k_p(\theta_d - \theta_{j-1}) \quad \text{(proportional term)}$$

$$k_d(\omega_d - \omega_{j-1}) \quad \text{(derivative term)}$$

$$k_i e_j \quad \text{(integral term).} \quad (4b)$$

Then, the gains k_p, k_d, and k_i are adjusted to give the best result, most often by trial and error.

In practice the magnitude of e_j, the integrated error, is limited to a range $\pm e_{\max}$ to prevent *integral windup,* which can occur during sudden large disturbance or setting changes. Windup limiting:

$$\text{if } e_j > e_{\max} \text{ then } e_j = e_{\max};$$
$$\text{if } e_j < -e_{\max} \text{ then } e_j = -e_{\max}. \quad (4c)$$

The usual procedure for adjusting the gains goes like this. First, the proportional gain is increased with $k_d = 0$ and $k_i = 0$ until the process responds fast enough (Fig. 11). Second, if unwanted oscillations occur, k_d is increased to help reduce them (0.025 in Fig. 12). (The effect of introducing the derivative term by using k_d has, in this case, an effect similar to adding mechanical damping.) When the quickest stable response is obtained, k_i is used to remove unwanted steady-state error remaining after the transient response has died away ($k_d = 0.03$, $k_i = 350$ in Fig. 13). (In this case the steady-state error is due to the load torque; Sec. 6.2 describes other causes.)

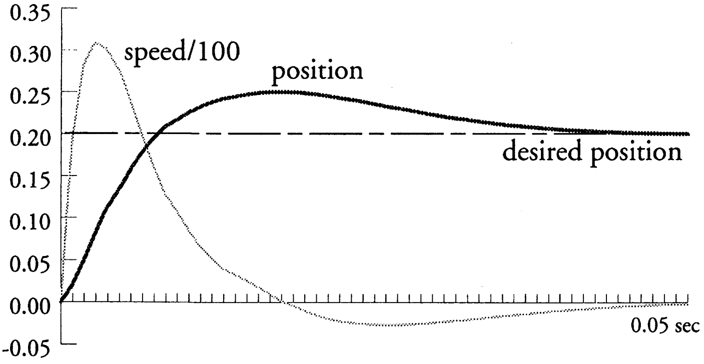

FIG. 13. Proportional + derivative + integral control. The steady-state error has been removed, but the transient lasts longer. In practice, many such simulation studies would be performed using control-system inputs representative of the actual application. Friction, time delays, power limits, sensor resolution, and other practical limitations need to be taken into account.

3.5 Limits of PID Control—When to Use Other Methods

We have seen how PID control works for a simple motor controller. PID control has been successfully applied to many processes and machines, but it has its limitations.

1. It is difficult to control nonlinear systems in which the gains need to change. For example, some positioning servos perform much better using error-squared control. Here, the control signal is proportional to the error squared, and thus the gain is higher when the error is large. When the error is small, the gain decreases, so that stability for final approach to the desired position is better.

   If the inertia of a motor shaft changes when the load changes, then a PID controller with fixed gains may go into unstable oscillations.
2. If the dynamics of the process are more complex, then a PID controller may not provide a satisfactory stable response. For example, if a motor drives a load on a long shaft that can twist by a significant amount, the simple model [steps (1)–(3) above] may not be accurate. A more complex model of the dynamic behavior will be needed, and a more complex controller may be necessary to control the torsional vibrations of the load on the long, compliant shaft.

   Advanced control techniques are available to design such a controller. These include cascade controllers, where extra feedback loops are used to stabilize parts of the system, making the whole system easier to control, and adaptive controllers that automatically adjust themselves to changing dynamics.

3.6 Fuzzy-Logic Control

No discussion of servos would be complete today without some mention of fuzzy logic. Fuzzy logic has become very popular with designers of industrial control systems, and many commercial design packages and hardware systems are available.

Fuzzy logic does not perform as well as most other control techniques can, but it is much easier to use for industrial designers. Also, a fuzzy-logic controller can incorporate the expertise of skilled process operators in the form of explicit rules (Johnston, 1994).

Each input to the controller is given a number of possible states. For example, a positioning error might be classified as follows: above 5 mm, "large positive"; 5 to 1 mm, "medium positive"; 1 to 0.1 mm, "small positive"; 0.1 to -0.1 mm, "zero"; -0.1 to -1 mm, "small negative"; -1 to -5 mm, "medium negative"; below -5 mm, "large negative." The controller is programmed with a particular rule for each possible combination of input states. Thus with three inputs and seven states each, 7^3 rules are needed.

Although the input classification might seem like coarse resolution, the control action from the controller is usually a nonlinear continuous function of the inputs with apparently infinite resolution. The controller is also programmed with "membership functions" so that a particular input value, say 3.5 mm, can be classified as "large positive," "medium positive," "small positive," etc. with different "degrees of membership" between 0 and 1. In this case, "medium positive" would have the highest membership value. Then the results of the rules for all input states with nonzero membership values are combined to give some optimum weighted output value.

Fuzzy-logic controllers have provided excellent results with processes that behave in a simple manner. Fuzzy control often beats PID control because the nonlinear nature of real processes and controllers can be taken into account. However, more conventional control techniques are far superior to fuzzy control for processes with complex dynamic behavior (such as a flexible robot arm), *provided* the behavior is observable and can be accurately predicted.

3.7 Self-Tuning and Adaptive Controllers

To control some processes, we need to adjust the gains of a simple PID controller to keep the response stable and eliminate unwanted errors. Several techniques have been developed to do this automatically. A number of companies offer self-tuning controllers, which adjust their gains automatically in response to changing conditions. Broadly speaking, there are two types of controllers:

1. gain-scheduling controllers, in which a table of gain values for particular conditions has been programmed in advance;
2. true adaptive controllers, which adjust the gains automatically on line to maintain optimum performance. This technique works well when relatively slow changes in gains are sufficient. This type of controller needs to be able to observe the system behavior for a little while before attempting to change the gain settings, and it is useful when gain scheduling is impractical (see Borrie, 1986).

3.8 Advanced Control Techniques

So far, we have looked entirely at single-output control systems. Many industrial process and machine controllers must work with several controlled outputs. For example, an aircraft autopilot must control aircraft speed, height, attitude (angle to the ground), and compass heading. Separate controllers cannot be used because all these aspects of aircraft behavior are "tightly coupled" together. Changing height, for example, requires an attitude change and engine adjustment, and will affect speed.

Advanced control techniques have been extensively developed for this and several other applications. Standard textbooks on control used for university engineering courses provide the theoretical background on these methods. However, most advanced control methods need to be applied with great care, and it is essential to seek advice from people with application experience (Borrie, 1986, discusses several different approaches).

4. COMPUTERS—FROM PERFORMANCE PREDICTION TO CONTROL

In Sec. 3.3, we predicted the performance of a control system by careful analysis and simulation. Step (4) represented the action of the controller.

In many, if not most, control systems today, the controller is a microcomputer. Computers have replaced analog electronic circuits used widely until the 1970s, because they are more reliable, more versatile, and easier to maintain.

The program in the control computer performs the same calculations as shown in step (4). Thus, in the case of a motor controller, the shaft position and velocity would be taken from input registers set by encoder pulse-counting logic, and the current setting would be sent to an output register, or converted to a PWM pulse train by the same microcomputer. Microcomputers have been specially designed for many typical applications. Hewlett-Packard, for example, produces the HCTL-1100 integrated circuit, which counts encoder pulses and controls a variety of electric motors or hydraulic actuators.

Many control and instrumentation systems are based on IBM personal computers or clones (PCs). Specialist companies supply interfaces to allow sensors to input feedback signals and the computer outputs to operate the process directly.

There are some special practical issues that must be considered. The most fundamental is the computation delay time.

4.1 Computation Delay Time

Computer control is based on the notion that the computer will regularly input process measurements and recalculate the control action to regulate the process. These operations are repeated indefinitely. The elapsed time between measurement at the feedback sensor and delivery of the newly calculated control-action signal to the process is the *computation delay time.*

Looking at steps (4a), (4b), and (4c) (Sec. 3.4 above), we can see that the computations are very simple indeed, and a modest computer could easily perform these in a few microseconds. We might therefore conclude that computation delay is not a significant issue. However, there are many other factors that lengthen the delay time.

First, the feedback sensor may itself incorporate a microcomputer. Often known as "smart" sensors, they are used because they are usually self-calibrating, and diagnose their own fault conditions. However there will be some computation delay because of this.

If the sensor signal is analog, it must be converted to digital form (A/D converter). To do this it is sampled, and then converted. The delay between sampling and arrival of the converted value in the computer memory must be taken into account.

If the sensor signal is transmitted to the control computer in digital form, then there will be data-transmission delay. Digital-data networks are increasingly used in control systems to reduce the cost and complexity of wiring. However, transmission delays can be much longer than the data-transmission rate (bits/second) might suggest, and most suppliers do not specify transmission delay times.

Similar delays may occur in outputting the computed control-action signal, particularly if it must be transmitted through a network again.

Next, the control computer will most likely need to check the validity of input signals to make sure that fault conditions are detected promptly. Usually this is more complex than just checking to make sure a signal is within a given range. Fault conditions are not easy to detect, especially if they are to be detected in time to take useful safety precautions, with sufficient reliability to reduce false alarms to an acceptably low level. While specialized "real-time" expert systems are being developed to do this, we do not yet understand how to design such systems in a straightforward manner.

Experience suggests that delays resulting from all these factors can far exceed the time required to perform the control calculations. Further, it is difficult if not impossible, in most cases, to determine the actual delay times from hardware-manufacturer data sheets. If delay times are stated, they are typically for only part of the measurement process.

For example, an interface data sheet may specify the maximum conversion rate for analog signals as 20 kHz. However, it may be difficult to achieve a computation delay of less than 1 ms in practice using DOS, even though the control calculations might take a few microseconds. Using a faster processor to reduce input delay may not help as much as expected.

All this assumes that the software running in the control computer has complete control over the computer's operating system.

Modern computer systems are making increasing use of networks to link computers together, and of graphical user interfaces (GUIs) such as Windows or OSF Motif. However, networks and GUIs can take over a computer so that a process-control program running in the computer may suspend operations for 100 ms or more. Delays such as this usually make direct control impractical using this type of operating system. However, Labview is an example of GUI software that can be used for direct control of processes with slower response times.

Because of this, specialized real-time operating systems are available for process control. Examples are QNX and Lynx for PCs and VX-Works for Unix computers. These operating systems have restricted features for general use, but ensure that delays caused by the operating system are minimized. Even so, total delays of less than 100 μs are difficult to achieve.

An interesting example is the transputer chip introduced for high-speed parallel computing in the 1980s. While each chip is fast and apparently ideal for control, communication across an array of chips incurs a series of cumulative delays that dominates the total computation delay.

One paradox of modern computing is that computer-speed increases have had only a marginal effect on response time. One reason for this is that input/output (I/O) operations take their timing from a standardized I/O bus compatible with both slower and faster machines. Thus delays will correspond more to the speed of the slowest (and probably the oldest) computer in a given range. Many I/O devices for PCs synchronize their timing with the original XT model introduced in the early 1980s. Another reason is that operating systems are designed to handle more and more options, networks, GUIs, etc., and so they consume an ever-increasing proportion of the computing power of the central processor.

As a rule of thumb, the computation delay in a feedback-control system should be less than 1/20 of the desired response time (Fig. 14). Thus, if a system-response time of 20 ms is adequate (approximately 10-Hz bandwidth), a PC computer could serve as controller. Otherwise specialized hardware and/or operating systems should be considered essential.

As a second rule of thumb, the computation delay should be measured with representative equipment and software, rather than being estimated from performance

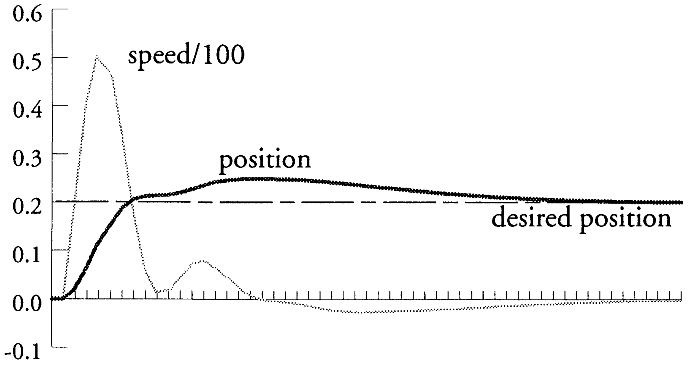

FIG. 14. Simulated behavior of the same PID control system as the preceding graphs, but with a 1-ms time delay in the controller. The reduction in stability is apparent from oscillation in the speed trace. A delay of 2 ms is sufficient to cause unstable behavior.

specifications. Measurement is not difficult. Apply a known-frequency signal to the sensor and observe the response lag in the control action.

The sheep-shearing robot was controlled by a computer capable of 3 million instructions per second (MIPS), which is modest by today's standards. With special software and hardware, the total delay was reduced to about 900 μs for the fastest actuator in the robot. Of this, the operating system contributed about 100 μs, control took about 300 μs, and the rest was needed for communication delays, A/D sampling, scaling, and fault detection.

Typical dc motors have response times of between 4 and 20 ms; this is the time needed for the motor shaft to accelerate to 63% of maximum speed with the rated voltage, assuming no load, and a zero-resistance power supply. Thus the computation delay of a motor controller should be less than 200 μs.

4.2 Sensor Resolution

The next factor in using a computer for control is the resolution of the input signal. An analog signal is said to have infinite resolution, which means that tiny signal changes can, in theory, be measured. However, when an analog signal is converted to digital form, the level must be represented by a finite number of bits, typically 8, 10, 12, 14, or 16. 8-bit resolution provides a maximum of 1 part in 256. However, a typical sensor might operate within a limited voltage range, say, 1 to 6 V. An A/D converter providing 8 bits of resolution usually has a range of −10 to +10 V, and so only 64 different values of sensor output can be represented between +1 and +6 V: The useful resolution is only 1 part in 64. For this reason, we aim to provide as much A/D resolution as possible within price, speed, size, and power restrictions. The useful resolution limits the accuracy that a control system can achieve, and if it is poor, it may affect stability.

Digital position encoders present a different resolution problem. While the resolution is theoretically limited by the number of pulses on the encoder, in practice the maximum pulse rate that can be reliably transmitted and counted may be the limiting factor. As we have seen, velocity feedback is essential for stable control. Velocity can be obtained from the pulse frequency, but at small velocities this information loses resolution because the time between pulses becomes longer. In some important situations, the loss of velocity information at low speed can seriously affect stability.

Digital encoders are often preferred because they are reliable, accurate, and repeatable. Resolution can be a problem, particularly if, for example, an encoder is being used to measure a relatively small position or angle change. If the resolution is poor, such that accurate positioning to within 1 encoder count is essential, a special design may be needed. A normal control system is likely to oscillate between one count and the next.

4.3 Output Resolution

In the same way that resolution of the input signal is a problem, output resolution can also be a problem, but the requirements are usually less stringent. For example, in the case of the electric motor, resolution of the output may not be fine enough to generate a current (and hence torque) to balance the steady load exactly. Once again, the effects can be explored by simulation (Fig. 15).

In this instance, mechanical friction can be helpful. If the mechanical-friction torque corresponds to a range of three or more possible output values (and hence motor-current values), output resolution should not cause oscillation problems.

In Excel, the effects of resolution can easily be computed. Given a quantity in, say, cell F7 that is to be measured to a resolution of x, the measurement can be computed by the formula

INT(F7/x + 0.5) * x.

5. SOME IMPORTANT LIMITATIONS

5.1 Steady Torque and Thermal Load on a Motor

The calculations shown so far are an approximate model of an ideal control system and electric motor. Real systems have important limitations that must be taken into account. In an ideal system the gains can be increased indefinitely to speed up the response of the motor. However, in doing so, the current i will far exceed what a practical power supply can provide and what a real motor could absorb without serious damage through heating, mechanical distortion of the windings, or demagnetization. The motor used as a case study can take a current of as much as 15 A for a brief instant, but the maximum continuous current is only 1.5 A.

The maximum continuous current is limited by the thermal capacity of the motor, particularly the rotor. Brushless dc motors, ac motors, and switched-reluctance motors can provide much more power for a given size and weight, but require more complex controllers (Kenjo, 1991).

If the controller can supply the peak current (implying that the dc power voltage is considerably higher than the nominal rating of the motor), then the motor may be damaged by a sustained high current that may occur during testing, or through a fault condition. The motor must be protected by a special circuit breaker to prevent damage.

Figure 16 illustrates a typical problem that must be considered in designing an electric servo. A motor is required to position a load vertically, and carry the weight of the load. This is a simplified version of an important problem in designing robot manipulators with several motors at the joints in the arm.

The first problem is that the load has inertia, which increases the apparent inertia of the motor shaft, slowing the response. However, in most applications, there is a gear reduction head on the motor so that the winding shaft rotates more slowly than the motor. The effective inertia is reduced by the square of the gear ratio, and so the motor inertia is usually the dominant component.

A more significant problem is the steady load torque experienced by the motor at rest. This must be balanced by a motor current unless the gear ratio is so high that friction in the mechanism can resist the weight of the load. This is unlikely if the load is to have a fast response. A steady current liberates heat in the windings through resistance

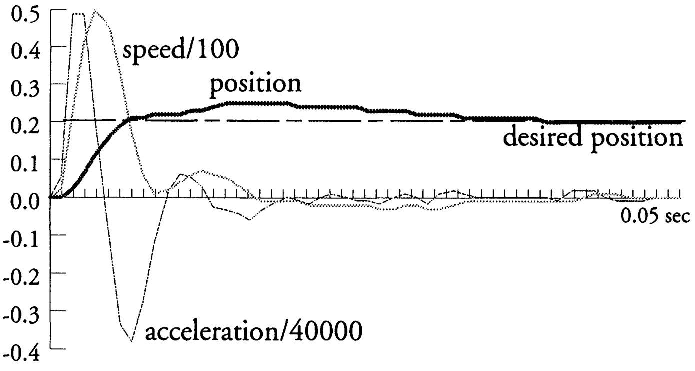

FIG. 15. Simulated PID controller from previous graphs, with 0.01 unit position resolution and 0.02 mA output resolution. The erratic acceleration trace shows how limited resolution introduces noise into a system, which will cause audible vibrations and noise in the servo. A simple simulation such as this is sufficient to confirm that noise and stability may be a problem. Experimental investigation would be necessary to measure the levels of noise and effects on accuracy.

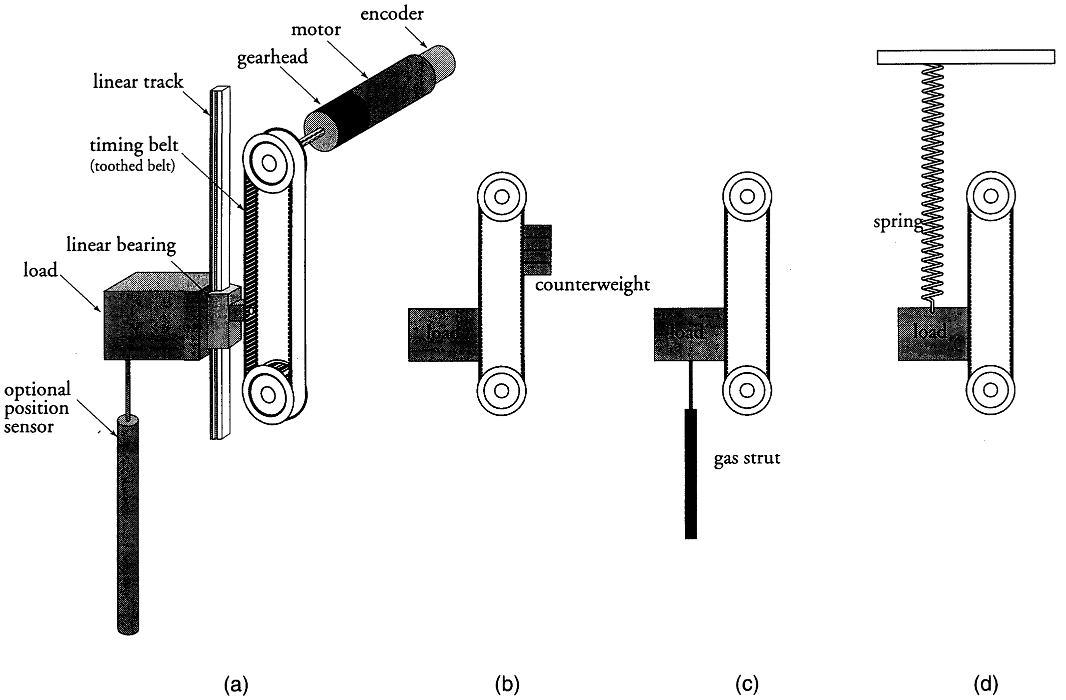

FIG. 16. (a) Typical position servo arrangement where load must be moved vertically against gravity force. Servo limitations influence important mechanical design choices. (b) Counterweight balances load weight. (c) Gas strut balances load weight. (d) Tension spring balances load weight. A weight, as in (b), can provide a constant counterbalance. This works well unless the added inertia slows the response too much. Further, the load mass may vary during operation so that a weight may not always balance the load.

effects, and the steady current that can be supplied to a typical motor is much less than the peak current (see Sec. 3.2). Therefore, most designers would provide some form of counterbalance in this instance (Bradley *et al.,* 1991).

A spring or gas strut (often used to balance the weight of the rear door of a car) provides balance without added mass. However, the force is often dependent on position and, in the case of a gas strut, speed.

Thus, the rate at which heat can be removed from a motor is often a major performance limitation, and indirectly affects several aspects of mechanical design. If the motor is close to heat-sensitive components (such as electronics or plastic materials), thermal loadings may have to be kept well below the maximum specification for the motor. Higher temperatures may change the motor characteristics enough to affect control performance.

The maximum continuous torque is limited by the heat dissipated in the motor windings. ac motors provide a higher power output for a given size, but require more complex controllers.

5.2 Speed Limitations

As the speed of a permanent-magnet motor increases, the windings generate a reverse voltage referred to as *back EMF* proportional to shaft speed. This opposes current flow, and for the motor in Sec. 3.2 is 1 V for each 585 rpm. Therefore, since the dc power-supply voltage is limited, the maximum current will be limited as the motor speed increases, and hence the maximum forward acceleration. The maximum reverse acceleration will increase at the same time, however, provided the power supply is designed to take the increased current. Braking with excessive reverse current can cause mechanical damage to some motors.

5.3 Power-Supply Limitations

Power supply can be a limitation in both electric and hydraulic systems. In the former, it is usually most important for battery-driven systems. Here there are two considerations. First, the overall drain on the battery needs to be minimized, and second, the supply current and voltage will drop as the battery loses charge. This may affect stability of the servo. A good computer control system can monitor battery charge and make allowances.

A hydraulic system is seldom self-contained, and so energy conservation is seldom so important. In fact, a hydraulic servo is very energy inefficient as the energy consumed on each stroke of the piston is independent of the load (power used is supply pressure × piston velocity × piston area). Further, servo valves require a constant oil flow to operate, called a pilot flow, typically 1%–3% of full rated flow. So the major consideration in a hydraulic system is pressure loss in the pipes and fittings between the pump and the servo valve, and then between the valve and the cylinder.

Note that the characteristics of both electric motors and hydraulic actuators change with temperature.

5.4 Sensor Location

Normally the feedback sensor is placed where the most direct output measurement can be obtained. This helps to minimize accuracy problems discussed in Sec. 2.1. If we look again at Fig. 16, we see that there is a complex mechanical transmission between the motor and the sensor if we use the load position sensor. This will introduce some compliance, stiction, and backlash between the motor and the sensor. Backlash is caused by clearance between adjacent parts: The motor must turn through some angle before the load moves. The term *stiction* describes a friction effect that is common in mechanical systems. When the friction coefficient decreases as sliding speed increases, motion tends to be intermittent or jerky and the moving parts seem "stuck" between motions. This effect is referred to as stiction and, like backlash, affects the ultimate accuracy that can be achieved.

If a slow response is sufficient, the control system may be stable. However, if a fast response is needed, these mechanical effects may destabilize control. For this reason, it is common to place the position encoder on the motor shaft, and use a stiff mechanical transmission. Figure 16 shows a timing belt being used. In Fig. 6, a lead screw and recirculating ball nut are shown. These components provide a stiffer and more precise mechanical transmission, but cost more and add inertia.

A good designer will place the feedback sensor and actuator (or motor) as close to the point of action as possible, and for the fastest response, the sensor and motor should be colocated. If the encoder is on the motor shaft, a reduction gearhead or lead screw helps to increase the effective position resolution.

Two sensors could be used if the mechanical transmission cannot be made rigid: one on the motor shaft and another at the load. Alternatively, a tachometer could be used at the motor and used to control motor speed with one controller, and an encoder could be used at the load for position control. However, more complex control is needed.

6. FAULTS IN CONTROL SYSTEMS—UNEXPECTED BEHAVIOR

When control systems work well for a while, they tend to be taken for granted. Failure is often marked by a sudden and dramatic change in behavior rather than a gradual deterioration seen in other machinery. This often occurs because the feedback that compensates for process variation can also mask progressive failure of a system component. The failure then only becomes evident when the controller can no longer compensate for this and the process variation combined.

While each type of control system (mechanical, electronic, hydraulic, etc.) has characteristic failures, there are failures common to all in principle.

6.1 Runaway

One of the most alarming and dangerous failures occurs when a position controller drives its load to one end of its range at maximum speed, possibly causing mechanical damage.

This type of failure emphasizes the need to cushion the load at each end, no matter how reliable the controller may be. Hydraulic cylinders are provided with optional hydraulic cushions as part of their design. Electric drives require limit switches to disable the motor if it is driving the load toward the end, and these must be placed far enough from the ends to allow the motor to run down and stop. With large loads or inertias, rubber cushions or shock absorbers are essential.

Possible explanations for runaway are the following:

1. Feedback-sensor failure. For example, if the mechanical coupling to the position sensor fails, then the measured position will not change, even if the motor shaft rotates. When a new position is requested, the motor turns to drive the load toward the desired position. However, the feedback signal does not change, and so the controller drives the motor faster and faster until the load reaches the end of its allowed motion . . . *Bang!* No external stimulus is needed—small natural perturbations can start a runaway some time after the component failure occurs. Apart from mechanical failure, a break in a sensor wire, a faulty connector, or failure of the sensor itself can cause similar effects. If the failure is interpreted by the controller as a large position change, runaway will occur immediately. Otherwise, it may happen later.

   A controller can be designed to shut off power automatically if there is no detected motion when the motor power level exceeds a given level. Usually this is detected by monitoring the error signal: Failure is presumed if the error exceeds a predetermined limit.
2. Amplifier failure. Failure of the electronic switches or equivalent devices in an electrical amplifier (Fig. 7), or of the valve in a hydraulic or pneumatic controller (Fig. 5), can cause similar effects.

   Once again, a controller can be designed to shut off power when the error limit is exceeded. Note that the power-shutoff mechanism or circuit must be independent of the amplifier. Hydraulic systems usually include an emergency shutoff valve that closes off the high-pressure supply to the servo valves. Some hydraulic servos also include check valves between the servo valve and the cylinder. These valves are held open during normal operation, but close if there is a power or controller failure. Electric and pneumatic servos may need a brake to stop motion safely.
3. Mechanical failure. If the load is not directly driven by the motor, failure of the transmission or counterbalance can also cause runaway.

Experience suggests that wiring faults form a large proportion of all failures. Particularly with robots, wires and connectors are subjected to repeated flexure and vibration, and will fail with intermittent faults eventually.

Normal flexible connecting wire consists of seven copper strands. Special cables are made to withstand repeated flexure, and each wire may consist of 100 or more extremely fine strands in specially chosen sleeving. Stress-relieving entry glands should be used where free cables enter rigid boxes or connectors, and where possible, mechanical "cable-track" chains should be used to constrain moving wires and prevent accidental kinking.

6.2 Persistent, Steady Error

1. Feedback-sensor failure. For example, if the mechanical coupling to the position-measurement encoder has slipped, then the measured position will be wrong. Thus, the motor turns to drive the load toward the desired position as measured by the sensor. If the error is large, this can cause a collision with the end stops. (Check the measured position when the load is at a known position. If the encoder has not slipped, check that the encoder mounting is secure.)
2. In the case of an incremental encoder, false pulses may be generated by electromagnetic interference (EMI). The interference may come from sources such as flickering fluorescent tubes or other motors starting or stopping. It can also come from the switching pulses of the servo motor—this may be the hardest to detect

and confirm. Absolute encoders help to prevent this problem. (Run a length of wire alongside the sensor wire and measure the EMI picked up by the wire using an oscilloscope. Compare the observed level with the voltage generated by the sensor.)

EMI often causes obscure faults that can be transient and difficult to observe. It is best to take basic precautions:

a. Amplify weak sensor signals close to their source, and where possible use transmission methods that are resistant to EMI such as a current loop, differential line drivers/receivers, or fiber optics.

b. Place dc power-supply regulators close to all sensitive circuits. dc isolation can be achieved by using a high-frequency ac power supply and converting to dc power at each sensor electronics unit. A high-frequency, low-power supply can be very compact.

c. Shield cables (sensor, motor, and power cables), and pass the shield drain wire through connectors back to a reliable common earth point for the whole system.

d. Use shielded twisted-pair cable designed for signal transmission, and if necessary, also for repeated flexure.

e. Use electronic filters to prevent EMI from entering boxes along wires or power supplies.

f. If possible, avoid running sensor cables parallel or alongside power cables.

Note that some EMI precautions (such as line filters or line drivers) can attenuate pulses from an incremental encoder if the pulse frequency is too high. Thus pulses may be lost during rapid motion. This can also result in a steady persistent error.

3. Mechanical overload—the load on the motor is too great for the controller to compensate. Try removing or reducing the load.
4. An excessive offset in amplifier or servo valve can result in a steady error. (An offset exists if the amplifier or servo-valve input is nonzero when the motor is not moving.)

6.3 Unstable Transient Response

After some time in operation, the control system may become less stable, with more overshoot after large position changes and more oscillation.

1. If the motor drive power is reduced (or hydraulic pressure is reduced) the maximum drive power is reduced. This may cause unstable behavior. (Measure maximum torque/force/acceleration and speed available.)
2. Change in load inertia or stiffness. Usually, increased inertia causes more overshoot. However, if a controller is "tuned" to the characteristics of a particular load, any change in load may reduce the controller performance. (Test controller with different loads.)
3. Often the friction in mechanical systems decreases with use, and this can reduce the energy dissipated by friction. If a bearing or lubrication fails, friction can increase with age and this can also cause instability. (Measure friction changes.)
4. Controller settings (such as gain) may have been accidentally changed following maintenance. Controllers should be designed such that gain settings do not have to be changed for maintenance or testing procedures. (Always record gain settings.)
5. Backlash (mechanical slack movement) will occur in joints and bearings of mechanisms subject to wear. As backlash increases, unstable behavior of the controller may increase. Earlier we noted the importance of choosing an appropriate sensor location. If the backlash occurs between the actuator and the sensor, it will almost certainly cause instability. Otherwise, it will cause steady errors. However, even when the sensor is mounted on the actuator, backlash can cause unstable behavior if it occurs between the actuator and the load, because it changes the apparent load characteristics. (Make provision for mechanical wear—provide adjustment for backlash, use wear-resistant components, keep contamination away from joints and seals, and measure backlash.)

7. DISCRETE CONTROL SYSTEMS

So far we have talked about control systems with variables that change more or less continuously. For example, the potentiometer of the hydraulic actuator provides a voltage output that varies continuously with the position of the piston. The servo valve generates an oil flow to the cylinder that is continuously variable.

There is another class of control systems where these variables have a small number of possible values. The simplest of these are on/off control systems. Figure 17 shows an on/off hydraulic actuator controlled with a four-way valve. Although the valve spool is very similar to that of a servo valve, its operation is quite different. If neither the "A" solenoid nor the "B" solenoid is activated electrically, the spool remains at the center position and no oil can flow to the cylinder. If the "A" solenoid is activated, the spool moves to the left, causing the piston also to move to the left at a constant speed.

The position of the piston is sensed with microswitches, which provide a signal when the piston has reached either end of its travel. Thus we can only be sure about the position of the actuator when the piston has reached the position of either microswitch.

There are countless examples of on/off control systems because they are often much simpler and cheaper than servo systems. Most automated factory machines use this kind of control. Pneumatic machines of this type are in widespread use around the world. Several manufacturers supply a vast range of standardized components (such as Festo, SMC, Parker, ABB, Siemens, Texas Instruments, and Rexroth, to name just a few). The controller (usually a microcomputer) is usually known as a programmed logic controller (PLC).

While discrete control systems operate quite differently from continuous control systems the same basic principles apply. The system must be observable—the system state we wish to control must be able to be measured. The system must also be controllable. Even though the system input is either on or off, it must still affect the system state in a predictable manner for the control system to work.

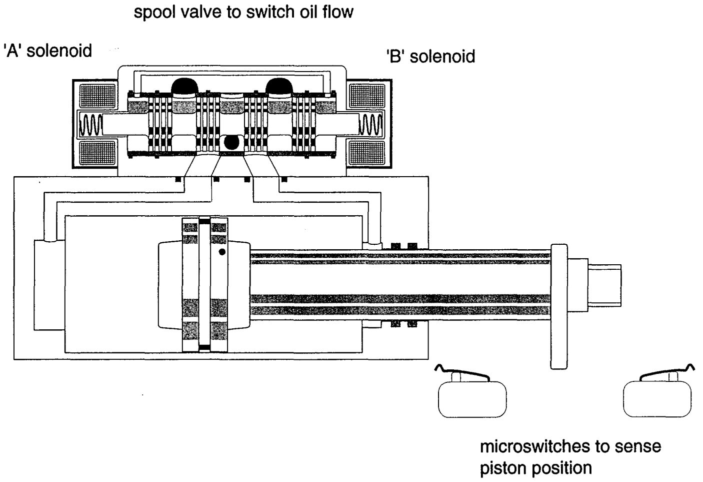
spool valve to switch oil flow
'A' solenoid
'B' solenoid
microswitches to sense piston position

FIG. 17. A simple on/off hydraulic actuator for positioning—refer to text.

APPENDIX

Excel Spreadsheet Formulas for dc Servo Simulation (Proportional Control, Fig. 11):

A	B	C	D	E
1) Time step	0.001		Desired posn	1
2) Gain	5		Inertia	9.38
3)				
4) Time	Current	Accn	Speed	Shaft posn
5)				
6) 0			0	0
7) =A6 + B1	= B2* (E1 − E6)	= (0.0163* B7 + 0.002)/(E2* 0.0000001)	= D6 + C7* B1	= E6 + D7* B1
8) = A7 + B1	= B2* (E1 − E7)	= (0.0163* B8 + 0.002)/(E2* 0.0000001)	= D7 + C8* B1	= E7 + D8* B1
9) = A8 + B1	= B2* (E1 − E8)	= (0.0163* B9 + 0.002)/(E2* 0.0000001)	= D8 + C9* B1	= E8 + D9* B1
10) = A9 + B1	= B2* (E1 − E9)	= (0.0163* B10 + 0.002)/(E2 0.0000001)	= D9 + C10* B1	= E9 + D10* B1
11) = A10 + B1	= B2* (E1 − E10)	= (0.0163* B11 + 0.002)/(E2* 0.0000001)	= D10 + C11* B1	= E10 + D11* B1

Row 7 represents the formulas in steps 1–4, and is copied down as far as necessary to obtain the desired number of results.

Time step	0.0010		Desired posn	1.000
Gain	5.000		Inertia	9.380
Time	Current	Accn	Speed	Shaft posn
0.0000			0.000	0.000
0.0010	5.000	89019.190	89.019	0.089
0.0020	4.555	81284.580	170.304	0.259
0.0030	3.703	66487.397	236.791	0.496
0.0040	2.519	45913.325	282.704	0.779
0.0050	1.106	21349.981	304.054	1.083

Plot columns D and E against time (shaft speed and position).

ACKNOWLEDGMENTS

The author would like to acknowledge the contributions of Michael Ong, who worked with the author as a research officer on the robotic sheep-shearing project and more recently developing low-cost servo-system components. Michael explored many of the practical aspects of servo systems and not only mastered them, but recorded his designs and calculations in great detail. He has not published his reports. However, many of the insights presented here stem from his work. David Elford and Jan Baranski, colleagues in the sheep-shearing project, also have contributed useful insights and design details.

GLOSSARY

Actuator (Sec. 2.3): A device that moves in response to an external command signal (Figs. 5, 6).

Adaptive Control (Sec. 3.7): Control system in which the controller behavior is adjusted automatically as the system properties change with time.

Amplitude Ratio (Sec. 3.1): Part of a frequency response, the ratio of amplitude of output to the amplitude of the input signal (Fig. 9).

Back EMF (Sec. 5.2): Reverse voltage generated in a spinning electric motor.

Backlash (Sec. 5.4): Clearance between adjacent mechanical parts that transmit motion.

Bandwidth (Sec. 3.1): Range of frequencies over which a device operates in a certain manner.

Closed-Loop Control (Sec. 2): A control system that adjusts the control action in response to measurements of the system output (Fig. 1).

Compliance (Sec. 5.4): Inverse of stiffness; the position of a mechanical transmission with compliance will be affected by the applied force.

Computation Delay Time (Sec. 4.1): Time between measurement and change in computer output signal, needed to transmit measurement to computer, process the measurement, and transmit the calculated output signal to the process.

Control Action (Sec. 2): Means of changing the state of a machine or process in order to change the system output (Fig. 1).

Control Engineering (Sec. 1.1): Technology of control systems.

Control System (Sec. 1.1): A control device and a regulated machine or process connected together (Fig. 1).

Controllability (Sec. 2): Degree to which the state of a machine or process can be changed.

Derivative Control (Sec. 3.4): Control system in which the control action is proportional to the rate of change of error (difference between the measured system output and desired output).

Desired Output Setting (Sec. 2): Control-system input that defines the desired value of the control-system output (Fig. 1).

Discrete Control Systems (Secs. 2, 7): Control systems in which variables have a finite number of possible values, usually a small number, often just two possible values or states (Fig. 16).

Disturbance Inputs (Sec. 2): Factors other than the control actions that affect the system output (Fig. 1).

Electromagnetic Interference (Sec. 6.2): False signals caused by wires acting as radio aerials, which pick up local electrical fields radiated from other wires.

Encoder (Sec. 2.4): An optical/electronic sensor for measuring position of an actuator or machine (Fig. 8).

Excel (Sec. 2.2): Software by Microsoft for manipulating spreadsheets.

Feedback Control (Sec. 2): A control system that adjusts the control action in response to measurements of the system output (Fig. 1).

Frequency Response (Sec. 3.1): Response of a device to a time-changing input that varies with a given frequency (Fig. 9).

Gain (Sec. 3.3): Factor by which an error, or other signal, is multiplied in a control system. Usually constant, but in some systems can be varied automatically.

Graphical User Interface (Sec. 4.1): Means of human–computer interaction by using a pictorial or graphical display and possibly a device (such as a mouse) to designate a point on the display, as well as a keyboard.

Illustrator (Sec. 2.2): Software by Adobe for creating illustrations and graphs.

Integral Control (Sec. 3.4): Control system in which the control action is proportional to the accumulated sum (or integral) of the error (difference between the measured system output and desired output).

Integral Windup (Sec. 3.4): An undesirable by-product of using integral control when large changes in the desired output setting occur; the error integrator accumulates a large unwanted offset during the transient response period. To avoid this, the integrator output is limited to a certain range.

Linear System (Sec. 1.1): One for which the response to two simultaneous inputs is the sum of the responses to each input presented separately.

Matrix-X (Sec. 3.2): Software package used for control-system simulation and design.

MATLAB (Sec. 3.2): Software package used for control-system simulation and design.

Observability (Sec. 2): The degree to which certain properties or states of a system can be measured.

On/Off Control Systems (Secs. 2, 7): Control systems in which variables have just two possible values or states (Figs. 1, 16).

Open-Loop Control (Sec. 2): Control system that adjusts the control action without measuring the system output; absence of feedback.

Output State Variable (Sec. 2): Property of control system that needs to be regulated (Fig. 1).

Phase Angle (Sec. 3.1): Part of a frequency response; the degree to which a system output either lags behind the system input, or leads ahead of it (Fig. 9).

PID Controller (Sec. 3.4): Controller that uses combined proportional, integral, and derivative control.

Potentiometer (Sec. 2.3): An electrical sensor often used to measure the position of an actuator (Fig. 5).

Programmed Logic Controller (Sec. 7): Small computer that can be used as a controller for a process, typically used for on/off or discrete control systems.

Proportional Control (Sec. 3.3): Control system in which the control action is proportional to the error (difference between the measured system output and desired output).

Pulse-Width–Modulation (PWM) Amplifier (Sec. 2.4): Common electronic-circuit arrangement for regulating current in an electric motor (Fig. 7).

Real-Time Operating System (Sec. 4.1): A computer-operating system designed to keep computation-delay time within specified limits.

Resolution (Sec. 4.2): Least detectable change in a signal.

Resonance (Sec. 3.1): Peak in amplitude-ratio part of frequency response (Fig. 9).

Response Time (Sec. 3.1): Time for the system output to reach the first 63% of the desired change.

Self-Tuning Control (Sec. 3.7): Control system in which the controller settings (such as proportional gain) are adjusted automatically as the system properties change with time.

Servo (Sec. 2.3): A device that changes its setting automatically in response to an external command.

Servo Valve (Sec. 2.3): An electromechanical valve that regulates the flow of fluid in a hydraulic control system (Fig. 5).

Simulation (Sec. 2.2): Calculation of behavior from mathematical equations that represent the properties of a system (Fig. 3).

Stability (Sec. 3.2): The degree to which the transient response disappears quickly.

Steady-State Response (Sec. 2.2): Response of control system after transient responses have stopped when the process or machine state no longer changes with time (Fig. 3).

Stiction (Sec. 5.4): A friction effect, common in mechanical systems, which occurs when the friction coefficient decreases as sliding speed increases so that motion tends to be intermittent or jerky and the moving parts seem "stuck" between motions.

Switching Controller (Sec. 2.4): Common electronic-circuit arrangement for regulating current in an electric motor (Fig. 7).

System Output (Sec. 2): Property of control system that needs to be regulated (Fig. 1).

Thermal Capacity (Sec. 5.1): Ability of electric motor to dissipate excess heat.

Transient Response (Sec. 2.2): Response of control system that changes with time but that stops after some time (Fig. 3).

Works Cited

Adobe (1994), *Adobe Illustrator Users Guide,* Adobe Systems, Inc.

Borrie, J. A. (1986), *Modern Control Systems: A Manual of Design Methods,* Englewood Cliffs, NJ: Prentice-Hall.

Bradley, D. A., Dawson, D., Burd, N. C., Loader, A. J. (1991), *Mechatronics—Electronics in Products and Processes,* London: Chapman and Hall.

D'Azzo, J. J., Houpis, C. H. (1981), *Linear Control System Analysis and Design,* New York: McGraw-Hill.

Johnston, R. (1994), "Fuzzy Logic Control," *GEC J. Res.* **11** (2), 99–109.

Kenjo, Tak (1991), *Electric Motors and Their Controls,* Oxford: Oxford Univ. Press.

Lovelock, J. (1987), *The Practical Science of Planetary Medicine,* London: Pan Books.

Maxon (1994), *Maxon Motor Catalog,* 1994/1995 edition, Sacshseln, Switzerland: Interelectric AG, p. 79, model RE 025 055 34EAA200A.

Microsoft (1994), *Excel Spreadsheet Software Package User Manual Version 4. for Windows,* Microsoft, Inc.

Trevelyan, J. P. (1992), *Robots for Shearing Sheep,* Oxford: Oxford Univ. Press.

Further Reading

Almost all of the thousands of learned papers published each year on control and servo systems focus exclusively on mathematical control theory, and so information on practical servo systems can be hard to find. The following is a small selection of books, articles, and catalogs to which the reader could progress for more detail. Many commercial catalogs have good application notes from which useful information can be gleaned.

Control Theory

Borrie, J. A. (1986), *Modern Control Systems: A Manual of Design Methods,* Englewood Cliffs, NJ: Prentice-Hall.

D'Azzo, J. J., Houpis, C. H. (1981), *Linear Control System Analysis and Design,* New York: McGraw-Hill.

Electric Motors

Bradley, D. A., Dawson, D., Burd, N. C., Loader, A. J. (1991), *Mechatronics—Electronics in Products and Processes,* London: Chapman and Hall.

Katebi, M. R., McIntyre, J., Lee, T., Grimble, M. J. (1993), "Integrated Process and Control Design for Fast Coordinate Measurement machine," *Mechatronics* **3,** 335–342.

Kenjo, Tak (1991), *Electric Motors and Their Controls,* Oxford: Oxford Univ. Press.

Maxon (1994), *Maxon Motor Catalog,* 1994/1995 edition, Sacshseln, Switzerland: Interelectric AG, p. 79, model RE 025 055 34EAA200A.

Yoshiaki, Hachisuka (1991), "Compact and High Torque Accelerators for Locomotion Robot," in: *Proceedings of the International Symposium on Advanced Robotics Technology (ISART),* Tokyo: Japan Industrial Robot Association, pp. 343–368.

Hydraulic and Pneumatic Systems

Boulet, B., Daneshmend, L., Hayward, V., Nemri, C. (1991), "System Identification and Modelling of a High Performance Hydraulic Actuator," in: *Proceedings of Second International Symposium on Experimental Robotics,* Berlin: Springer.

Daley, S. (1990), "Self Tuning Control of a Hydraulic Test Rig Using a Personal Computer," *GEC J. Res.* **7** (3), 157–167.

Dransfield, P. (1990), "PC Controlled Air Powered Servodrives—Their Response Potential," in: *Proceedings of the 3rd International Fluid Power Workshop,* Boston: IFS.

Jacobsen, S. C., Iverson, E. K., Knutti, D. F., Johnson, R. T., Biggers, K. B. (1986), "Design of the Utah/MIT Dextrous Hand," in: *Proceedings of the IEEE International Conference on Robotics and Automation, San Francisco,* New York: IEEE, pp. 1520–1532.

Trevelyan, J. P. (1992), *Robots for Shearing Sheep. Shear Magic,* Oxford: Oxford Univ. Press.

Uusijarvi, R. W., Torngren, E. (1994), "Introducing Distributed Control in Mobile Machines Based on Hydraulic Actuators," *Mechatronics* **4,** 139–157.

There are relatively few references on modern hydraulic servo systems. Good information is available from leading suppliers such as Moog, Parker, and Rexroth through local distributors.

Mechanical Design of Servo Systems and Transmission Elements

Rivin, E. I. (1988), *Mechanical Design of Robots,* New York: McGraw-Hill.

Stiffler, A. K. (1992), *Design with Microprocessors for Mechanical Engineers,* New York: McGraw-Hill.

Several bearing manufacturers such as THK supply specialized components such as linear tracks, recirculating ball bearings, and ball nuts.

Robotics

McKerrow, P. J. (1991), *Introduction to Robotics,* Reading, MA: Addison-Wesley.

Sharkey, P. M., Murray, D. W., Vanevelde, S., Reid, I. D., McLauchlan, P. F. (1993), "Modular Head/Eye Platform for Real-Time Reactive Vision," *Mechatronics* **3,** 517–535.

Trevelyan, J. P. (1992), *Robots for Shearing Sheep. Shear Magic,* Oxford: Oxford Univ. Press.

Trevelyan, J. P. (in press), "Robot Force Control without Stability Problems," in: *Proceedings of 3rd International Symposium on Experimental Robotics, Tokyo,* Berlin: Springer.

Fuzzy Control

Johnston, R. (1994), "Fuzzy Logic Control," *GEC J. Res.* **11,** 99–109. A good general introduction.

Other

Lovelock, J. (1987), *The Practical Science of Planetary Medicine,* London: Pan Books. Gives some interesting comments on the relevance of feedback systems to environmental science.

Catalogs

Major suppliers include Sankyo Denki, Escap, ABB, and GEC/Alsthom.

Motorola, National Semiconductor, and Hewlett-Packard manufacture a range of motion-control and servo components, including encoders and control microprocessors. Most semiconductor companies supply switching circuits in various capacities.

SHOCK WAVES

Y. Horie, *North Carolina State University, Raleigh, North Carolina, U.S.A.*

Akira B. Sawaoka, *Tokyo Institute of Technology, Nagatsuta, Yokohama, Japan*

Introduction 29
1. **Fundamental Features of Shock Waves** 30
2. **Elementary Wave Interactions** 32
2.1 Characteristic Equations and Riemann Invariants 33
2.2 Examples of Wave Interactions 34
3. **Various Problems** 35
4. **Experimental and Diagnostic Techniques** 36
5. **Numerical Techniques** 37
6. **Applications** 37
6.1 Aerodynamics and Propulsion 38
6.2 Materials Synthesis 39
6.3 Medical Applications 39
6.4 Metalworking 40
6.5 Inertial-Confinement Fusion 40
7. **Conclusion** 41
Glossary 41
Works Cited 42
Further Reading 42

INTRODUCTION

In physical science, the term "shock wave" refers to a narrow region of compression (wave front) moving at supersonic speed across which the pressure, velocity, and other related quantities of the medium change in a nearly discontinuous manner. The term is also used to signify the entire wave including the wave's trailing expansion phase, such as the boom from a supersonic aircraft, or the blast wave from an explosion.

Shock waves arise from many sources: a sudden release of various forms of energy in a limited space; supersonic motion of a body in a fluid medium; or the impact of gases, liquids, and solids at high speed. Physically, the formation of a shock wave is attributed to the fact that in most materials the sound speed increases with increasing pressure. Thus, when a finite-amplitude wave is created in a medium, its profile becomes steeper during propagation and develops into a near discontinuity as a result of a high-pressure regime overtaking a low-pressure regime. The maximum steepness of the wave front is limited by the dissipative processes of viscosity and heat conduction.

The phenomena of shock waves are truly ubiquitous. They occur on scales ranging from atomic to cosmic. Electrons moving in a tiny semiconductor device may cause such a wave, as may a massive star collapsing under gravity at the end of its evolution. The shock wave from the latter is the basis of the supernova explosion that is thought to be the origin of most heavy elements in the galaxies. Life, then, is born of seeds scattered by cosmic shock waves (Glass, 1974).

Shock waves are man-made as well as natural. A sonic boom is a conical shock wave from the passage of a supersonic aircraft through the atmosphere. Shock waves are often invisible without the assistance of a special diagnostic device, but the devastating damage they can cause, e.g., from the detonation of chemical or nuclear explosives, is often witnessed. However, even nuclear explosions are minute when compared with the shock waves generated by the impact of large meteorites with planets. Such an impact on Earth is thought to have caused the demise of dinosaurs many millions of years ago. The recent comet impact on Jupiter provided a spectacular astronomical event readily observable on Earth.

Chemical explosives are routinely used in mining and civil engineering for excavation

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

and demolition. The almost instantaneous reaction of explosives is indispensable for the deployment of air bags that often save lives in auto accidents. High-pressure shock waves are used to synthesize diamond and other hard materials for applications in grinding and polishing. Shock-synthesized diamond powders are said to have unique polishing properties that do not occur with naturally made diamond. Their lack of sharp cutting edges avoids the creation of deep grooves. A focused shock wave from an electric discharge or a small amount of explosives has been used to fracture gallstones to facilitate their removal for several years now. Shock waves associated with supersonic flight are to be avoided because of their disturbing qualities to humans and animals, but, in the case of reentry vehicles, shock waves help to increase drag forces; e.g., by increasing the gas density, they aid in deceleration through the Earth's atmosphere.

There are also many exciting examples of potential applications that are currently under investigation. For instance, scramjets (supersonic combustion ramjet) are being considered for the proposed hypersonic National Space Plane that can carry passengers from New York to Tokyo in less than three hours. Laser fusion, one of the several techniques of obtaining a practically inexhaustible source of energy, uses an imploding shock to heat and compress a target with the aim of initiating and controlling the thermonuclear-fusion reaction therein.

Obviously, shock waves are not limited to industrial applications. They have been used to develop our understanding of materials response to extreme conditions of high pressure and temperature (see Equations of State; Solids, Dynamic High-Pressure Effects in) and also to test the laws of the physical world. There have been many fundamental contributions in the studies of gas dynamics, chemical kinetics of gases at high temperatures, high-pressure and high-temperature equations of state and other properties of solids, and the planetary interiors. The study of cosmic shock-wave phenomena has been a key to understanding of the creation of the universe and the evolution of stars, the interstellar medium, our solar system, and even the Earth. This article only touches a very small portion of truly ubiquitous shock-wave phenomena.

1. FUNDAMENTAL FEATURES OF SHOCK WAVES

Theorizing about shock waves dates as far back as Rankine's memoir (Rankine, 1870). Mathematical formulation varies from a simple physical approach in one space dimension to the approach based on weak solutions of hyperbolic differential equations in multiple spatial dimensions. In this article, discussion will be limited to an idealized, one-dimensional shock wave in a perfect fluid. This is a simple model, but it describes basic shock-wave relationships. A more general derivation and the inclusion of such effects as chemical reactions, ionization, and electromagnetic fields may be found in references provided at the end of the article.

As shown schematically in Fig. 1, a pair of vertical lines represents a narrow region of a plane shock wave propagating from left to right in a perfect fluid across which the flow variables change almost suddenly. The conservation laws of mass, momentum, and energy applied across this near discontinuity yield (Courant and Friedrichs, 1948)

$$\rho_1 u_1 = \rho_0 u_0, \tag{1}$$

$$p_1 + \rho_1 u_1^2 = p_0 + \rho_0 u_0^2, \tag{2}$$

$$\begin{aligned} &\rho_1 u_1(E_1 + \tfrac{1}{2}u_1^2) + p_1 u_1 \\ &\quad = \rho_0 u_0(E_0 + \tfrac{1}{2}u_0^2) + p_0 u_0, \end{aligned} \tag{3}$$

where p is pressure, ρ is density, E is specific internal energy, and u is flow velocity. The subscripts 0 and 1 denote the values of variables upstream and downstream of the discontinuity, respectively. These equations are known to hold in the normal direction for even an arbitrary discontinuous surface.

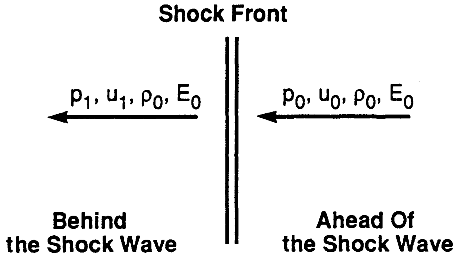

FIG. 1. A schematic diagram of a normal plane shock.

By rearranging variables in Eqs. (1)–(3), one can transform them into the more useful form

$$m^2 = (p_1 - p_0)/(V_0 - V_1), \tag{4}$$

$$u_1 = u_0 \pm (p_1 - p_0)/m, \tag{5}$$

$$E_1 - E_0 = \tfrac{1}{2}(p_1 + p_0)(V_0 - V_1), \tag{6}$$

where $m = \rho_0 u_0 = \rho_1 u_1$ and $V = 1/\rho$. The last equation is known as the Hugoniot equation. In many applications, enthalpy is a more useful variable than internal energy. Then, the Hugoniot equation becomes

$$H_1 - H_0 = \tfrac{1}{2}(p_1 - p_0)(V_0 + V_1). \tag{7}$$

The speed of the shock, D, is

$$D = V_0[(p_1 - p_0)/(V_0 - V_1)]^{1/2} + u_0. \tag{8}$$

The locus of final states is called the Hugoniot curve (or shock adiabat) and represents the states that can be reached from the initial state (p_0, V_0) by a single shock. It does not represent the successive states of compression such as an isentropic compression curve (see Fig. 2). Since the shock adiabat involves two parameters, the initial state and the shock strength, two successive shocks cannot, in general, be effected by a single shock wave. Equations (4) and (5) have useful geometric meaning in the p-V plane. For instance, $-m^2$ is the slope of the chord called the Rayleigh line, drawn from (V_0, p_0) to (V_1, p_1). Then, the shock in the internal energy is the area under this Rayleigh line (see Fig. 2).

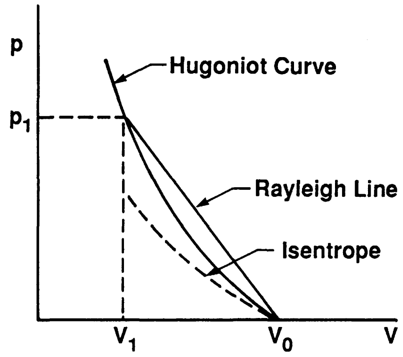

FIG. 2. Hugoniot curve in the p-V plane.

There are four variables in the shock equations, and so a fourth equation is required to determine the Hugoniot curve. This is typically accomplished by hypothesizing an equation of state such as the ideal-gas law or by introducing experimentally obtained data on the D-u relationship.

Consider the ideal-gas law

$$E = pV/(\gamma - 1), \tag{9}$$

where γ is the ratio of the specific heats at constant pressure and at constant volume. Then, it may be found that

$$\frac{V_1}{V_0} = \frac{(\gamma + 1)p_0 + (\gamma - 1)p_1}{(\gamma - 1)p_0 + (\gamma + 1)p_1}, \tag{10}$$

$$\frac{T_1}{T_0} = \left(\frac{p_1}{p_0}\right)\frac{(\gamma + 1)p_0 + (\gamma - 1)p_1}{(\gamma - 1)p_0 + (\gamma + 1)p_1}. \tag{11}$$

This shows that an arbitrarily strong shock cannot compress the gas by a factor greater than $(\gamma + 1)/(\gamma - 1)$. For a monatomic gas of $\gamma = 5/3$, this factor is 4, a large density change mentioned previously regarding the increase of drag forces on reentry vehicles.

A linear D-u relationship is known to hold for a variety of condensed matter over a wide range of pressures, even up to a megabar range for some materials; $D = C_0 + gu_p$ where C_0 is comparable to the bulk sound speed at ambient pressures, u_p is the particle velocity of shock running into an initially stationary state $(u_0 - u_1)$, and g is a constant. Then, the Hugoniot in the p-V plane is given by

$$p_1 = \frac{\rho_0 C_0 (1 - V_1/V_0)}{[1 - g(1 - V_1/V_0)]^2}. \tag{12}$$

Computations of other thermodynamic variables such as temperature require mechanical and thermal equations of state. For instance,

$$p(V,E) = p_i(V) + b(E - E_i), \tag{13}$$

$$E = E_i + C_V(T - T_i), \tag{14}$$

where p_i and E_i are pressure and internal energy on the isotherm, C_v is the specific heat at constant volume, and b is a constant. Then, the pressure and temperature on the Hugoniot are given by

$$p_1 = \frac{p_i - b(E_i - E_0)}{1 - b(V_0 - V_1)/2}, \tag{15}$$

$$T_1 = T_0 + \frac{p_1 - p_i}{bC_V}, \tag{16}$$

where $p_0 = 0$.

For a normal fluid for which

$$(\partial^2 p/\partial \rho^2)_S > 0,$$

where S is the entropy, it is known (Zel'dovich and Raizer, 1965) that this entropy change is of the third order in terms of the shock strength $p_1 - p_0$ or $V_0 - V_1$. This underlies the important fact that the motion of a shock wave, even in an ideal fluid, is an irreversible process and leaves a permanent change in the medium after the wave has passed. In fact, this condition is used to select the physically admissible solutions from among those that satisfy the shock conditions. Since the entropy increases only in compression for normal materials, an expansion shock wave does not exist in such a medium. A typical exception is a medium that involves a phase transition or a chemical reaction. Other noteworthy consequences of the entropy increase are that

1. the shock is compressive ($p_1 > p_0$ for $V_0 > V$), and
2. the shock wave is supersonic relative to the upstream material, but the downstream flow is subsonic.

Since a true discontinuity is physically impossible in a real material, the shock transition occurs in a thin layer (shock structure). But the entropy change across the shock does not depend on the mechanisms of entropy production. These mechanisms control only the structure and the thickness. In a gas, this dissipation is due to the collision of atoms and molecules. On the macroscopic level, if high-temperature effects such as ionization can be ignored, the collisional process may be approximated, model-wise, by viscosity and heat conduction. Then, the structure of gasdynamic shock can be adequately described by the Navier–Stokes equations (Hayes, 1960).

The Navier–Stokes equations have proven useful, not only for investigating numerous inert dynamic problems, but also for problems that involve complicating effects such as dissociation, chemical reactions, and ionization (an extreme example of which is the compression of targets in laser fusion). With any of these processes, the shock conditions require no modification; the only complications that require change are the equations of state or the details of the structure (Hayes, 1960).

If temperatures exceed a few tens of thousands of degrees, there will be an additional complication resulting from radiative heat transfer. Since the mean free paths of light are much greater than those of gas particles, the radiative transfer controls the thickness of the shock structure. This requires the modification of the basic conservation equations. Furthermore, if an external magnetic field exists, then shock waves must be reconstructed on the basis of the equations of magnetohydrodynamics. However, for many technological applications, shock waves with an infinitely thin discontinuity provide an adequate first-order description.

The notion of viscosity and heat conduction has also been used to represent the macroscopic shock structure of solids, but viscous and conductive behaviors require different types of descriptions because of their underlying heterogeneous and anisotropic properties of shear strength (Graham, 1993). Likewise, multiphase heterogeneous systems such as gas–solid mixtures will also require new formulations (Hetsroni, 1982).

There are even collisionless shocks in space plasma where the structure is no longer due to collisions, but is created by electromagnetic interactions. The resulting shock thickness is known to stretch over a distance of many thousands of kilometers. The Earth's bow shock, one of the best examples, is caused by the supersonic flow of ionic particles (solar wind) past Earth's magnetosphere (Stone and Tsurutani, 1985; see also MAGNETOSPHERE).

2. ELEMENTARY WAVE INTERACTIONS

In a physical world, a shock wave is rarely steady or solitary. It interacts with other types of waves as well as material bodies. Without support, a shock wave attenuates quickly and "degenerates" into waves dispersed over a wider region through decompression. In this section, elementary examples of plane shock-wave interactions in an ideal compressible fluid are considered. First, the relationships for propagation of continuous disturbances, referred to as simple waves, are developed.

2.1 Characteristic Equations and Riemann Invariants

The continuum time-dependent equations for conservation of mass, momentum, and energy can be written in one space dimension as

$$\frac{d\rho}{dt} + \rho\frac{\partial u}{\partial x} = 0, \tag{17}$$

$$\rho\frac{du}{dt} + \frac{\partial p}{\partial x} = 0, \tag{18}$$

$$\frac{dE}{dt} + p\frac{dV}{dt} = 0 \quad \text{or} \quad \frac{dS}{dt} = 0, \tag{19}$$

where t = time, x = Eulerian space coordinate, S = entropy, and

$$\frac{d}{dt} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x}. \tag{20}$$

Since $\rho = \rho(p,S)$ for an ideal fluid, Eqs. (17) and (18) can be transformed into the characteristic form (Lieberman and Velikovich, 1986)

$$\left[\frac{\partial p}{\partial t} + (u \pm C)\frac{\partial p}{\partial x}\right] \pm \rho C\left[\frac{\partial u}{\partial t} + (u \pm C)\frac{\partial u}{\partial x}\right] = 0, \tag{21}$$

where

$$C = (\partial p/\partial \rho)_s^{1/2}. \tag{22}$$

Recalling the physical meaning of the total time derivative along a particle path, Eq. (21) can be reduced to

$$\rho C\frac{du}{dt} \pm \frac{dp}{dt} = 0, \quad \text{along}\,\frac{dx}{dt} = u \pm C. \tag{23}$$

The directions specified by dx/dt in the x-t plane are called characteristics ($C_\pm$ respectively) and define a system of new coordinates. Physically, they represent the local propagation paths of all analytic disturbances (simple waves). Equation (19) is also the characteristic equation (C_0) along particle motion:

$$\frac{dS}{dt} = 0, \quad \text{along}\,\frac{dx}{dt} = u. \tag{24}$$

Since the sound speed is a function of density alone, Eq. (23) can be further simplified by functions called Riemann invariants,

$$I_\pm = u \pm \int \frac{dp}{\rho C}. \tag{25}$$

The dynamic equations become

$$I_\pm = \text{constant}, \quad \text{along}\,\frac{dx}{dt} = u \pm C. \tag{26}$$

The flow is now completely described by two variables: u and p (or any one of the other variables: ρ or C) along the characteristics. The Riemann invariants provide the conservation conditions along grids of the new coordinate systems.

For illustration, consider a small disturbance propagating in one direction. On setting $u = u_0 + \Delta u$, $p = p_0 + \Delta p$, it may be seen that (Zel'dovich and Raizer, 1965)

$$I_\pm = \Delta u \pm \Delta p/\rho_0 C_0 = \text{const.} \tag{27}$$

If the perturbation is a forward-facing wave (a simple wave moving to the right), then I_- is constant since no disturbance is propagating in the C_- direction. Likewise, if it is backward facing, moving to the left, then I_+ is constant. In general, a disturbance originating at some point p in the space-time coordinate (see Fig. 3) can only influence events that lie in the region demarked between the C_+ and C_- through p advancing in time. Likewise, the state at p can be affected by a certain region of events in the past (domain of dependence).

It may now be seen that if the speed of acoustic waves increases with pressure, a continuous, finite pressure wave propagating in one direction steepens its profile with increasing time because of converging C_+

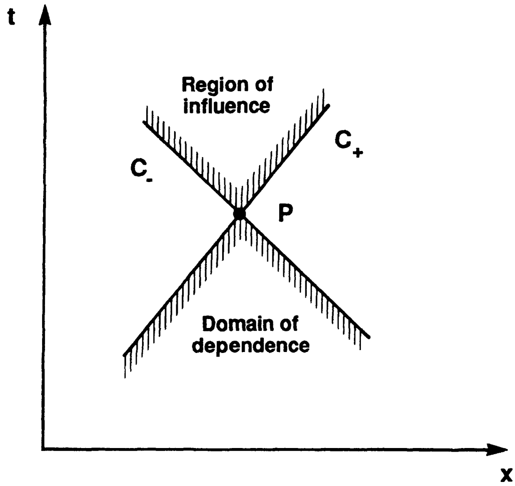

FIG. 3. Region of influence and domain of dependence defined by $C_\pm$ characteristics.

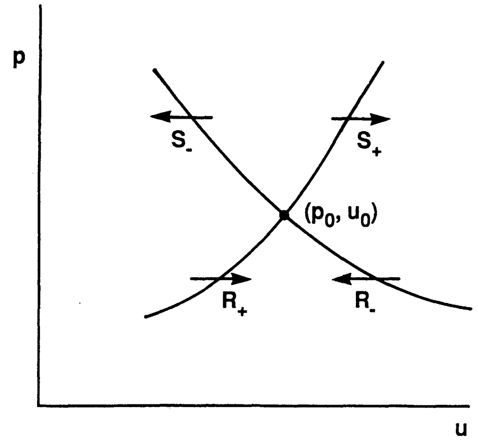

FIG. 4. Four types of wave transition in the *p-u* plane.

grids and that after some time it develops into a situation in which the two characteristic curves intersect. This is physically impossible because the flow field becomes multiple-valued. The result is the formation of a discontinuous shock wave. It may also be seen that diverging expansion waves (rarefaction waves) do not develop the nonuniqueness problem.

Generalization of the characteristics of problems involving more than two independent variables is found in advanced texts (e.g., Courant and Friedrichs, 1948). Here, note only that the characteristic grids in two dimensions become multidimensional surfaces (e.g., a cone in two space variables) and integration of the governing equations still involves partial differential equations. Therefore, engineering applications of these equations are limited in comparison with numerical approaches. However, their solid foundation in mathematics makes them an indispensable tool, not only in mathematical analysis of wave propagation, but also in the numerical solution of engineering problems.

2.2 Examples of Wave Interactions

The essential point of the previous discussion is that the change of an arbitrary state created by wave interactions is connected by specific paths described by Eqs. (5) and (25) in the *p-u* plane as illustrated in Fig. 4. In this figure, the labels *S* and *R* indicate a shock and a simple expansion wave originating at the initial state (p_0, u_0). The signs + and − designate forward- and backward-facing waves, respectively. To illustrate the use of these fundamental relationships, two applications (Duvall, 1971; Zel'dovich and Raizer, 1965) will be considered.

Figure 5 sketches a solution to the Riemann problem, where at an initial time, a semi-infinite region of a high-pressure gas on the left-hand side is in contact with a low-pressure gas on the other side (an idealized shock-tube problem). At subsequent times, a uniform shock travels to the right (a forward-facing shock S_+) and a family of expansion waves travel to the left (backward-facing, simple expansion waves R_-). When viewed in the *p-u* plane, the continuity of *p* and *u* for the state between the separating waves requires that this state must be given by the intersection of S_+ originating at $(p_0, u_0 = 0)$ and R_- at $(p_1, u = 0)$.

A special application of the above solution is a plane shock induced by explosive detonation. If the detonation process is assumed to be an adiabatic expansion of an ideal gas from a high-pressure state p_1 to an ambient pressure p_0, then the velocity of the products is given by (Schall, 1971)

$$u = \int_{p_0}^{p_1} \frac{dp}{\rho C} = \frac{2C_1}{\gamma - 1}\left[1 - \left(\frac{p_0}{p_1}\right)^{(\gamma-1)/2\gamma}\right], \qquad (28)$$

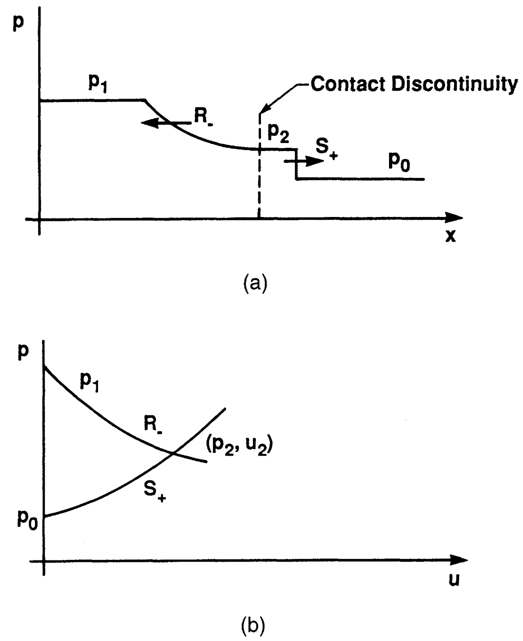

FIG. 5. Schematic diagrams of wave interactions in a shock tube: (a) a spatial wave profile and (b) wave transition in the *p-u* plane.

where C_1 is the sound speed of the detonation product. If this shock interacts with an inert medium like a metal plate without impacting it, then the strength of the shock transmitted in the medium can be located using the same procedure as previously given. If the gas impacts with a finite velocity, then it belongs to the type considered next.

The second case is concerned with the transmission of a shock through a material interface. See Fig. 6. The incident shock S^{I}_{+} propagates from left (material I) to right (material II). It may be assumed, without loss of generality, that impedances satisfy $\rho_{02}D_2 > \rho_{01}D_1$. Then, both the transmitted and reflected waves are shock waves. Again the continuity of p and u on the material interface indicates that the state separating the two waves is found at B, the intersection of S^{I}_{+} and S^{I}_{-}. The latter originates at A, the state for the incident shock. If the second material is a rigid wall, it can be shown that

$$\frac{p_2 - p_0}{p_1 - p_0} = \frac{\rho_1 D'_{21} + \rho_0 D_1}{\rho_0 D_1} \underset{p1\to p0}{\longrightarrow} 2, \tag{29}$$

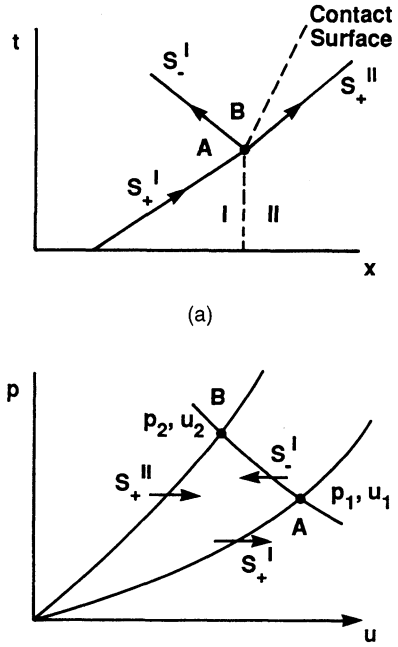

FIG. 6. Schematics of shock-wave transmission through a material interface: (a) shock trajectories in the *x-t* plane and (b) shock transition in the *p-u* plane.

where D_1 = velocity of the incident shock, D'_{21} = the velocity of the reflected shock relative to the material ahead, and $p_B > p_A$. It may be seen that if the magnitudes of shock impedances are reversed, then B is related to A by R_{-}.

3. VARIOUS PROBLEMS

As briefly mentioned in the Introduction, there are many interesting problems regarding shock waves. The following is a short list of such topics that are beyond the scope of the present article. They are complex because

1. the mathematics of shock-wave propagation is nonlinear;
2. the shock wave and materials response are interdependent—the shock wave affects materials behavior and vice versa; and

3. the wave motion occurs in multiple dimensions.

In the real world, the normal shock waves described previously are a special case of a more general family of oblique waves in which the mass flow across the shock front has a component parallel to the discontinuity. These shock waves are obviously two- and three-dimensional in nature. The discussions of these oblique waves are found in most textbooks of gas dynamics. Nevertheless, so far as the shock conditions are concerned, the obliquity does not change the shock conditions in the normal direction. For plane waves, the change required is the condition that the tangential component of the flow velocity is preserved. One often-illustrated oblique shock is the supersonic flight of a wedge- or cone-shaped object. Many of these shocks are illustrated in the collections of Van Dyke (1982) and Glass (1974).

When temperatures are high enough, shock waves inevitably induce chemical reactions in their wake. They occur not only in gases, but also in solids and even nuclear substances (nucleosynthesis on a cosmological scale). If the amount of heat is large enough and it is deposited fast enough, the reactions can lead to a unique phenomenon called detonation. This mode of combustion is distinguished from ordinary combustion in that the reacted materials move in the same direction as that of the propagating reaction wave (Fickett and Davis, 1979).

Converging and diverging shock waves are difficult problems to solve even in one dimension. They are fundamentally transient problems and require either a numerical method or a special analytical technique such as the similarity method (Sedov, 1959).

Beyond the simple condition mentioned earlier, the question of stability is a difficult issue because it could arise from geometry, or material properties, or both (Whitham, 1974).

4. EXPERIMENTAL AND DIAGNOSTIC TECHNIQUES

In some areas such as astrophysics, it is difficult to conduct experiments of observed phenomena under controlled conditions. But in gas dynamics and aerodynamics, shock tubes, wind tunnels, or combinations of both have been used to study shock-wave phenomena for many decades. According to Anderson (1990), Vieille used the former in 1899 and Busemann conducted experiments with the latter in the early 1930s.

A wind tunnel consists essentially of a channel through which a uniform stream of gas is driven at controlled velocities. For subsonic tests, the system is typically driven by a fan or propeller, but it requires a tank of compressed gas and a properly shaped nozzle to obtain supersonic streams at a high Mach number (u/C) of 4 or 5. Obviously, the latter are run only for short periods of time.

Tests in wind tunnels typically involve a scaled model of a real object and require a careful matching of the Mach and the Reynolds numbers (the ratio of momentum to viscous forces). As a result, a hypersonic tunnel such as the one at the NASA Langley Research Center is required to employ cryogenic nitrogen gas. Induced shock temperatures will reach 20 000 K and above.

In shock tubes, shock waves are produced in a closed tube by a sudden deposition of energy. Some of the most common sources of energy are compressed gas, electric discharge, explosives, and combustion. Each covers a different range of loading conditions. No single facility can cover the entire range of pressures and temperatures required to study diverse topics ranging from a simple shock in an inert gas, to weak shocks in automobile exhaust pipes, to shocks involving chemical reactions and ionizations (Green, 1964).

The scope and breadth of shock-tube and shock-tunnel applications may be seen by the following chapter titles excerpted from the proceedings of the 17th International Symposium on Shock Waves and Shock Tubes, *Current Topics in Shock Waves* (Kim, 1990). The chapter titles include shock formation; focusing and implosion; shock reflection and diffraction; turbulence, ionization, and plasma interaction; chemical kinetics, pyrolysis, and soot formation; ignition of detonation and combustion; particle entrainment and shock propagation through particle suspension; boundary layers; and blast simulation.

A variety of diagnostic techniques have

been developed to study the properties of shock waves. The measurements typically monitor the shock velocity and other quantities such as density, pressure, and temperature, and chemical compositions. Generally, they are classified as optical and nonoptical techniques. The optical techniques include Schlieren and shadow photography, interferometry, spectroscopy, and x-ray absorption. Particularly noteworthy is the rapid development in laser technology and its applications to spectroscopic measurements. Examples of nonoptical techniques are piezoelectric and ionization gauges.

Techniques for producing shocks in condensed matter are numerous. They include chemical explosives, high-speed impact of projectiles, and the deposition of energy by laser or charged-particle irradiation. When compared with gases, shock-induced changes of temperature and compression are, in general, small with the obvious exceptions of porous and powder materials, though pressures as high as several terapascals can be achieved in these materials.

Diagnostic techniques for solids, again, include optical and nonoptical techniques; however, with opaque solids, optical techniques are limited to surface measurements such as interferometric velocity measurements, luminescence, and surface temperatures. There exist not only stress gauges such as manganin and quartz gauges, but also electromagnetic gauges for particle-velocity measurements. The measurements of other physical properties such as electrical, magnetic, and optical properties have also provided valuable information, though the measurement of chemical compositions is limited at the present time to post-shock samples as exemplified in diamond synthesis.

5. NUMERICAL TECHNIQUES

It may not be an overstatement to say that, in the technical and industrial applications of shock waves, the single most important development over the last several decades is that of numerical calculations based on finite differences. The progress has been truly explosive. The solution of fluid-dynamic problems with shocks is no longer restricted to specialists, but it is becoming (if it is not already) an integral part of routine design tasks for engineers.

What is also significant is that with the advent of fast workstations, access to supercomputers is no longer critical to solving problems. For instance, the calculation of shock-free airfoils described in the following requires many thousands of cells and equations, but can be approached by anyone with a workstation and an idea. This has leveled the playing field and created a robust community of diverse ideas. For a historical review of the subject, with special focus on computational fluid dynamics, refer to two excellent articles by MacCormack (1985, 1993).

Before the advent of fast digital computers, the method of characteristics was a reasonably accurate means of numerically integrating hyperbolic systems of partial differential equations. For instance, the first plan of a supersonic wind tunnel was designed by a numerical method based on the characteristic equations by Prandtl and Busemann [a pedagogical account is found in Anderson (1990)]. But it is a cumbersome method because the characteristic net deforms as the flow changes (see Sec. 2.1) and may introduce inordinate numerical errors. In contrast, finite-difference techniques are more straightforward than the method of characteristics and are less prone to errors associated with grids. Hence, most contemporary approaches to numerical solutions of shock problems are based on finite-difference methods.

6. APPLICATIONS

There are no distinct boundaries between the scientific and technological applications of shock waves, but there is a useful distinction. In the scientific application, shock waves alone are the object of research and this research, in turn, provides information about the source or the medium in which shock waves propagate. Astronomical and interstellar shock waves are typical examples. Shock-wave research in condensed matter, motivated mostly by weapons development, has provided an understanding of shock-compression behavior under extreme conditions of high pressures, high temperatures, and high rates of loading.

In industrial and technological applications, shock waves or the force of shock waves are used either constructively or destructively, or both. Explosive metal forming and demolition in civil engineering are typical examples of the constructive/destructive use of shock waves. Metal forming exploits the force of high-pressure shock waves to do some useful work that may not be possible by other means. Shock waves may also form the metal at a much faster rate or more efficiently than other means. Scramjets, as described in the following, are another example. Extracorporeal lithotripsy is an example in which the destructive power of shock waves is used for a constructive purpose.

The subjects of lithotripsy, aerodynamics, diamond synthesis, and inertial-confinement fusion are treated in detail in other articles in the *Encyclopedia*.

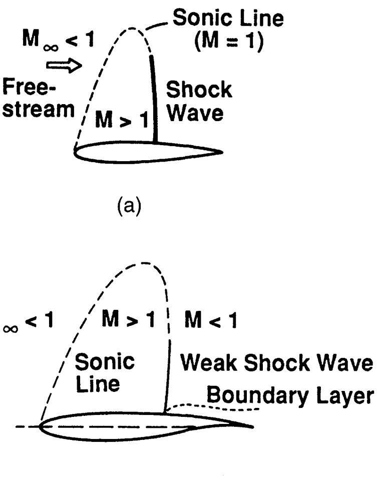

FIG. 7. A sketch of flow pattern near airfoils: (a) a section with a strong shock and (b) a section with no shock or a weak shock.

6.1 Aerodynamics and Propulsion

It may not be entirely wrong to state that the study of shock waves in the early days was motivated by the desire to fly at supersonic speed. The literature is too vast to conduct even the review of a single topical area. Therefore, the following examples of shock-free airfoil and supersonic combustion are meant to convey just the flavor of current interest.

A shock-free airfoil is not an application of shock waves in the strictest sense, but it is directly tied to shock-wave research because the goal is to eliminate shocks that occur even in subsonic flight as a result of the acceleration of air flow above the wing to generate lift (see Fig. 7) (see AERODYNAMICS). Such airfoils reduce the drag generated by shocks (about 5% of the lift force) to near zero. In practice, however, the complete elimination of shock waves is found to be very difficult because any small perturbation in a flow results in the appearance of relatively weak shocks. The design of such airfoils is a prime example of the application of numerical techniques through computer power in solving aerodynamic problems.

The concept of air-breathing supersonic combustion (scramjets) is an old idea that dates back to the 1960s (Waltrup *et al.*, 1976). It is proposed for both jets at high Mach numbers (>6) and booster rockets. Recent revival of interest has been triggered by President Reagan's proposal (1985) for hypersonic planes such as the National Aerospace Plane that will fly from New York to Tokyo in less than three hours. These planes are proposed to cruise at a speed of about 8 km/s in an Earth orbit.

There are numerous advanced ideas of scramjets, but there is not yet an engine in practical production. Many technical problems need to be overcome (Billing, 1986). One such issue is the problem of slow mixing in supersonic flow. This requires a long combustion section that is counter to the enhanced thrust from supersonic combustion. The gas dynamics of shock formation from the release of chemical energy at the entrance of the combustion section is another problem. Nevertheless, technical arguments supporting the use of scramjets for hypersonic flight are very strong. Scramjet designers today have many advantages over the pioneers of the 1960s. They have a critical tool, the computer, for large-scale computations of complex reacting flow from shocks. It is not unrealistic to expect to see a prototype space plane in the near future.

Connected with scramjets, there is also a revival of the concept of a supersonic combustion ramjet-in-tube to accelerate a projectile to a very high speed. In this scheme, a projectile flies through a premixed fuel–oxidizer mixture and uses energy through its supersonic combustion (Hertzberg *et al.*, 1988).

6.2 Materials Synthesis

Excluding military applications, the early study of shock (detonation) waves was carried out on solids. In the 18th century, there were attempts to exploit the high-pressure shock waves generated by impacting rifle bullets on graphite powder in a steel cavity to manufacture diamond through phase transition. Although the actual discovery and subsequent commercialization of shock-synthesized diamond did not take place until the early 1960s, present understanding of the shock synthesis of diamond shows that diamond powders probably were produced in the early experiments but were unidentifiable because of their microscopic size.

Because of the very short duration of the shock pulse (on the order of microseconds), the product powders are small, ranging typically from 0.01 μm to several microns. However, these particles are aggregates of microcrystallites and have a blocky shape with thousands of microcrystallite cutting edges that have unique cutting and polishing properties. As a result, they are used in special superprecision applications in the microelectronics industry. In Russia, shock-synthesized diamond powders are being used as an additive lubricant in oil as well as a raw material for manufacturing hard compounds.

Diamond powder is not the only substance that is commercially manufactured by the shock-compression technique. Wurtzite-type and zincblende-type boron nitride have also been produced on a commercial scale and used for the manufacture of cutting and grinding tools.

6.3 Medical Applications

The use of explosive shock waves is common in demolishing structures, breaking boulders, and removing earth. Now the same destructive power is used for noninvasive crushing of bladder renal calculi. The technique is named extracorporeal shock-wave lithotripsy (ESWL) (Chaussy *et al.*, 1980). Since the first clinical application of one such device in 1980, its effectiveness has been recognized worldwide (20 000 cases were reported in 1985) and ESWL is now considered the preferred treatment for urinary stones.

Figure 8 is a schematic of such a device. Box S represents the generator of focusing shock waves. The coil represents the high-voltage source or the detonator of a microexplosion. Shock waves are generated at the first focus of an ellipsoidal cavity and are refocused at the second focus point where a target stone is located (Takayama, 1993). The mechanisms of stone disintegration are not yet well understood, but are thought to be caused by the cavitation of water on the stone's surface, hammering from water jet, and compressive and tensile waves in stones. Efforts to understand the mechanisms better are hampered by the fact that there is no adequate device to measure high pressures in living tissues. Existing gauges are made inoperable because of ionic, chemical, and electromagnetic activities occurring in living tissues.

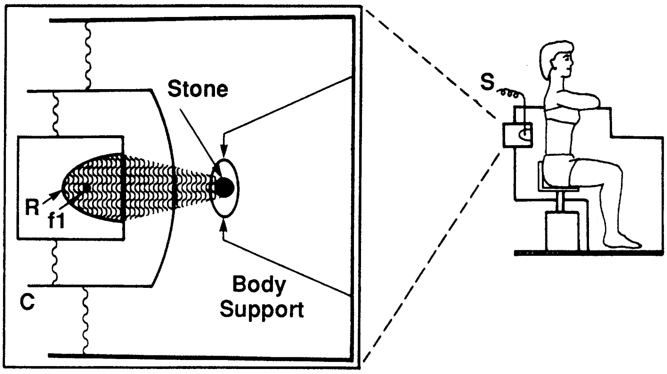

FIG. 8. A sketch of ESWL: R = reflector, C = a wave-coupling device, f1 = shock-wave source located at first focal point [after Takayama (1993)].

There are many new studies of focused shock waves in medical applications. One application that has already passed the stage of clinical tests is extracorporeal shock-wave osteoclasis (ESWO). In this technique, shock waves are used for the treatment of pseudoarthosis and delayed (and controlled) union of the fractures by creating microfissures and bony fragments (Schlecbergur and Sengo, 1992). Another application is the shock destruction of cancer cells.

6.4 Metalworking

The use of high-intensity shock waves for a variety of metal processes such as forming, cutting, embossing, and joining is a mature subject; nevertheless, new applications are far from being exhausted. The earliest patent on the expansion of bicycle tubes with low explosives dates back to 1898. High explosives are probably the most common source of energy today, but a high-voltage discharge and a magnetic field are also used for forming.

Techniques of explosive forming and welding do not require sophisticated equipment. In forming, an explosive is detonated and its energy (in the form of a shock wave) is transmitted through a fluid medium, commonly water, to a workpiece that is typically held on a female die. In the 1960s, large structural elements with a diameter as large as 3–5 m were fabricated by the explosive technique. Since then, for a variety of reasons such as noise, precision, and cost, the technology has been on the wane except for the following few bright areas:

1. the cladding of corrosion-resistant titanium on steel plate, and
2. the consolidation of amorphous metal.

The latter is still in its infancy, however, because of the problem of cracking during the expansion of shock waves. Nevertheless, the coating of thin amorphous film or powder is known to have passed the experimental stage (Murr, 1988).

An interesting new trend in metal forming is the redirection of the technology from mass production to the production of unique pieces that require precision. One example is the fabrication of denture plates made of pure Ti (Kani, 1977). These plates have several advantages over existing plastic pieces: thermal conductivity, thickness, and dimensional accuracy. Shock hardening of pure Ti helps to develop subsequent dimensional stability. Another area of experimentation is the creation of artwork. Again, the aim is not the mass production of ornamental pieces as in the past, but a unique artwork that is akin to painting and demands a close collaboration between the artist and the engineer in charge of production. An example is shown in Plate 1.

6.5 Inertial-Confinement Fusion

The role of shock waves in the quest for fusion power through inertial confinement has elements that are analogous to those in metal forming and propulsion. A small fuel pellet is heated and compressed by laser or ion beams irradiated uniformly over the pellet surface for a few nanoseconds (see Fig. 9).

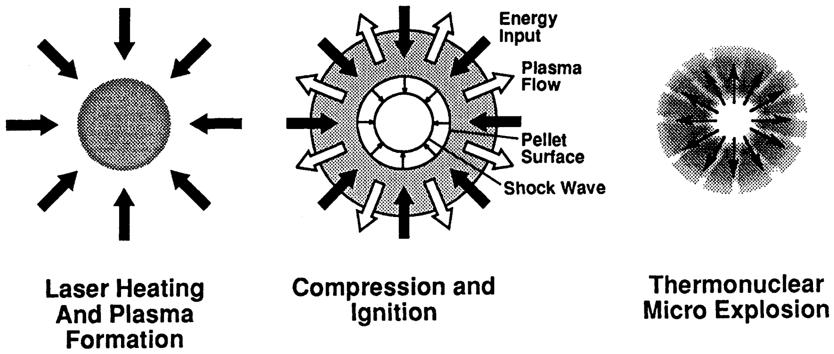

Laser Heating And Plasma Formation

Compression and Ignition

Thermonuclear Micro Explosion

FIG. 9. A sketch of laser fusion.

A spherically heated outer layer becomes an ablating plasma, moves outward, exerts through recoil a pressure of 10^8 to 10^9 bars on the surface, and creates a radially converging, strong shock wave as well as an imploding mass motion. When the core reaches thermonuclear conditions (on the order of 2×10^5 kg/m^3 and 100 million degrees centigrade for deuterium–tritium fuel), it ignites and burns. If there are enough fusion reactions and sufficient inertial force to hold the compressed pellet, the reactions propagate outward, releasing much more energy, and produce an overall positive energy gain. Someday microscopic imploding shock waves may bring us abundant, clean energy through man-made, controlled nuclear reactions.

On a cosmic scale, shock waves generated in the final throes of collapsing, massive stars (supernovas) are known to play a crucial role in the creation of elements heavier than helium, the origin of life, in some sense, in the universe. In type I supernova having solar masses of 1.44 to about 8, the shock wave is driven by explosive nuclear reactions (Prange and Girvin, 1990). But in type II, a supernova with a solar mass greater than 8, the shock wave is said to be generated by the impact of an outer layer, collapsing as a result of gravity, on a core that has become a nuclear matter whose density is of the order of 2×10^7 kg/m^3! Type II is epitomized by the recent case of the 1987A supernova that captivated the attention of not only astronomers and physicists, but also laypeople. The scale of a supernova event is beyond any human experience—e.g., the gravitational energy release of over 10^{46} joules in less than a tenth of second and the initial temperature of 200 billion degrees at the center. Study of such a shock is not limited to the interior evolution of supernovas. As the shock bursts out of a supernova, it provides a large mass of hot gas and interacts with the surrounding media. Thus, the shock waves become a rich source of x rays and infrared emissions and provide a valuable means of investigating interstellar media.

7. CONCLUSION

Shock waves differ from other mechanical waves in that

1. they move at supersonic speed,
2. their amplitudes are determined by the energy of the wave, and
3. entropy increases across their near discontinuous wave fronts.

Technological applications are mostly based on the high-pressure and high-temperature effects of shocks under high rates of loading as exemplified by scramjets, inertial fusion, and diamond synthesis. Scientific applications, due to the ubiquitous nature of shock phenomena on microscopic to cosmic scales, have contributed to the understanding of such diverse subjects as solid state, interstellar structure, and supernova.

GLOSSARY

Detonation: A supersonic, self-sustained mode of combustion. The reaction region is preceded by a shock wave.

Hugoniot Equations: Equations that express the changes of flow properties across shock waves, based on the conservation of mass, momentum, and energy.

Laser Fusion: An approach to liberation of nuclear energy by heating and compressing a small pellet by the irradiation of laser light.

Mach Number: A dimensionless quantity (M) given by u/C. If $M > 5$, the motion is said to be hypersonic.

Navier–Stokes Equations: Partial differential equations governing the motion of a viscous, heat-conducting fluid.

Ram Jet: A compressorless jet engine. Oxidizer air in an engine duct is compressed by the forward motion of the aircraft, mixed with fuel, and burned. The exhaust from the rear of the engine produces thrust.

Rarefaction Wave: An expansion wave across which the pressure decreases.

Reynolds Number: A dimensionless parameter (Re) given by the ratio of inertial forces to viscous forces.

Schlieren (and Shadow) Photography: An optical technique for photographing density gradient (of disturbances) in fluid motion under a special lighting arrangement.

Scramjet: An air-breathing engine in which combustion occurs at supersonic air velocity.

Solar Wind: Motion of ionized particles from the sun's corona with velocities of 300–1000 km/s.

Supernova: A massive star that suddenly explodes into great brilliance by a factor of 10^{10}.

Works Cited

Anderson, A. D., Jr. (1990), *Modern Compressible Flow,* New York: McGraw-Hill, Chaps. 9 and 11.

Billing, F. S. (1986), *Ramjets with Supersonic Combustion,* AGARD LS-136, Neuilly-Sur-Seine, France: North Atlantic Treaty Organization, pp. 8.1–8.29.

Chaussy, C., Brendel, W., Schmiedt, E. (1980), *Lancet* **2,** 1265–1268.

Courant, R., Friedrichs, K. O. (1948), *Supersonic Flow and Shock Waves,* New York: Interscience.

Duvall, G. E. (1971), in: P. Caldirola and H. Knoeppel (Eds.), *Physics of High Energy Density,* New York: Academic, pp. 7–47.

Fickett, W., Davis, W. C. (1979), *Detonation,* Berkeley: Univ. of California Press.

Glass, I. I. (1974), *Shock Waves & Man,* Toronto: Univ. of Toronto Press.

Graham, R. A. (1993), *Solids under High Pressure Shock Compression,* New York: Springer.

Green, E. F. (1964), *Chemical Reactions in Shock Waves,* New York: Academic.

Hayes, W. D. (1960), *Gasdynamic Discontinuities,* Princeton: Princeton Univ. Press.

Hertzberg, A., Bruckner, A. P., Bogdanoff, D. W. (1988), *AIAA J.* **26,** 195–203.

Hetsroni, G. (1982), *Handbook of Multiphase Systems,* New York: McGraw-Hill.

Kani, K. (1977), in *Proceedings of 6th International Conference on High Energy Rate Fabrication, Essen, West Germany,* Denver: Denver University Press, pp. 3.8–3.815.

Kim, Y. W. (1990), *Current Topics in Shock Waves,* New York: American Institute of Physics.

Lieberman, M. A., Velikovich, A. L. (1986), *Physics of Shock Waves in Gases and Plasmas,* New York: Springer-Verlag, Chap. 1.

MacCormack, R. W. (1985), AIAA Paper 85-0032, New York: American Institute of Aeronautics and Astronautics.

MacCormack, R. W. (1993), AIAA Paper 93-3291-CP, New York: American Institute of Aeronautics and Astronautics.

Murr, L. E. (1988), *Shock Waves for Industrial Applications,* Philadelphia: Noyes.

Prange, R. E., Girvin, S. M. (1990), *Supernovae,* New York: Springer.

Rankine, W. J. M. (1870), *Trans. R. Soc. (London),* 160–277.

Schall, R. (1971), in: P. Caldirola, H. Kneppel (Eds.), *Physics of High Energy Density,* New York: Academic, pp. 239–240.

Schlecbergur, R., Sengo, T. (1992), *Arch. Orthop. Trauma Surg.* **111,** 224–227.

Sedov, L. L. (1959), *Similarity and Dimensional Methods in Mechanics,* New York: Academic.

Stone, R. G., Tsurutani, B. T. (1985), *Collisionless Shocks in the Heliosphere: a Tutorial Review,* Washington, DC: American Geophysical Union.

Takayama, K. (1993), *Jpn. J. Appl. Phys.* **32,** 2192–2198.

Van Dyke, M. D. (1982), *An Album of Fluid Motion,* Stanford, CA: Parabolic Press.

Waltrup, P. J., Anderson, G. Y. Stull, F. D. (1976), in *Proceedings of 3rd International Symposium on Air Breathing Engines,* pp. 835–861.

Whitham, G. B. (1974), *Linear and Nonlinear Waves,* New York: Wiley.

Zel'dovich, Ia. B., Raizer, Yu. P. (1965), *Physics of Shock Waves and High Temperature Hydrodynamics,* New York: Academic.

Further Reading

Journals/Proceedings

Shock Waves; The Journal of Fluid Mechanics; Physics of Fluids; Combustion, Explosion, and Shock Waves (tr. of Fiz. Goreniya Vzryva); International Journal of Impact Engineering.

International Symposium on Shock Waves and Shock Tubes (Biannual since 1957), *Symposium International on Detonation.*

Mathematics

Lax, P. D. (1973), *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves,* Philadelphia: SIAM

Le Veque, R. J. (1992), *Numerical Methods for Conservation Laws,* Boston: Birkhauser.

Liberman, M. A., Velikovich, A. L. (1986), *Physics of Shock Waves in Gases and Plasmas,* New York: Springer.

Menikoff, R., Plohr, B. J. (1989), *Rev. Mod. Phys.* **61,** 75–130.

Oran, E. S., Boris, J. P. (1987), *Numerical Simulation of Reactive Flow,* New York: Elsevier.

Potter, D. (1973), *Computational Physics,* New York: Wiley.

Richtmyer, R. D., Morton, K. W. (1967), *Difference Methods for Initial-Valued Problems,* New York: Wiley.

Scramjets

Avery, W. H., Dugger, G. L. (1964), *Astronaut. Aeronaut.* **2,** 42–47.

Northam, G. B., Anderson, G. Y. (1986), AIAA

Paper 86-0159, New York: American Institute of Aeronautics and Astronautics.

Medical Applications

Kandel, L. B., Harrison, L. H., McCullough, D. L. (1987), *State of the Art Extracorporeal Shock Wave Lithotripsy,* New York: Futura.

Inertial-Confinement Fusion

Verlarde, G., Ronen, Y., Martinez-Val, J. M. (Eds.) (1993), *Nuclear Fusion by Inertial Confinement,* Boca Raton, FL: CRC Press.

Yamanaka, C. (1991), *Introduction to Laser Fusion,* New York: Harwood Academic Publishers.

Collisionless Shocks

Stone, R. G., Tsurutani, B. T. (1985), *Collisionless Shocks in the Heliosphere: Review of Current Research,* Washington, DC: American Geophysical Union.

Supernovae and Interstellar Shock Waves

Bethe, H. (1986), in: S. L. Shapiro, S. A. Teukolsky (Eds.), *Highlights of Modern Astrophysics,* New York: Wiley, Chap 3.

McKee, C. F., Draine, R. T. (1991), *Science* **252,** 397–403.

Ostriker, J. P., McKee, C. F. (1988), *Rev. Mod. Phys.* **60,** 1–68.

Material Synthesis and Metal Working

Crossland, B. (1982), *Explosive Welding of Metals and Its Applications,* Oxford: Clarendon.

Horie, Y., Sawaoka, A. B. (1993), *Shock Compression Chemistry of Materials,* Tokyo: KTK Scientific Publishers.

Sawaoka, A. B. (Ed.) (1993), *Shock Waves in Materials Science,* Tokyo: Springer.

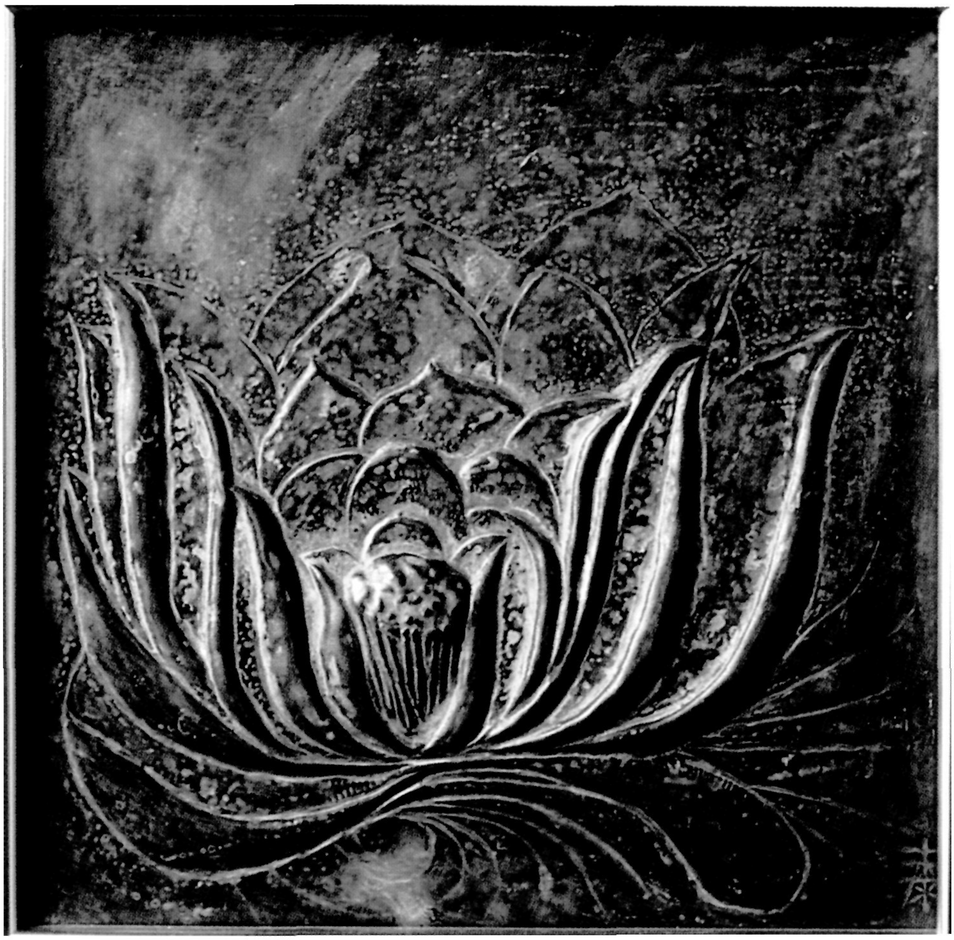

Vol. 18 – PLATE 1. An explosively formed art work by Yasuko Sawaoka (by permission of the artist).

Si–SiO$_2$ INTERFACE, ELECTRONIC PROPERTIES

ATSUSHI HIRAIWA, *Central Research Laboratory, Hitachi, Ltd., Kodaira, Tokyo, Japan*

Introduction 45
1. **Preparation of SiO$_2$** 46
1.1 Thermal Oxidation 46
1.2 Formation of SiO$_2$ Other Than by Oxidation 46
1.2.1 Chemical Vapor Deposition (CVD) 46
1.2.2 Liquid-Phase Deposition (LPD) 47
2. **Characterization of the SiO$_2$–Si Interface** 48
2.1 Morphology 48
2.2 Bonding Structure 48
2.3 Electrical Characterization 48
2.3.1 Theoretical Evaluation of Interface Traps 48
2.3.2 Measurement of Interface Traps 50
2.3.2.1 Quasistatic Method 50
2.3.2.2 High-Low–Frequency Capacitance Method 51
2.3.2.3 Charge Pumping 52
2.3.2.4 Electron-Spin Resonance 53
3. **The SiO$_2$–Si Interface and Device Performance** 53
3.1 Operational Principle of MOS Transistors 53
3.2 Degradation of MOS Transistor Performance 55
3.2.1 Increase in Subthreshold Swing 55
3.2.2 Reduction of Drain Current (1): Loss of Surface Carriers 56
3.2.3 Reduction of Drain Current (2): Decrease in Mobility 56
3.2.3.1 Screened Coulomb Scattering 56
3.2.3.2 Phonon Scattering 56
3.2.3.3 Surface-Roughness Scattering 57
3.3 Excessive Junction Leakage Current 57
3.3.1 Requirements of Low-Power DRAMs 57
3.3.2 Shockley–Read–Hall Process 58
3.3.3 Other Current Generation Processes 59
3.3.3.1 Local Zener Effect 59
3.3.3.2 Diffusion 59
4. **Generation and Passivation of Interface Traps** 59
4.1 Physical Origins 59
4.1.1 Photon Beams 59
4.1.1.1 Below 8 eV 59
4.1.1.2 From 8 eV to 470 keV 60
4.1.1.3 Above 470 keV 61
4.1.2 Electron and Hole Flux 61
4.1.2.1 Below 10 eV 61
4.1.2.2 Above 10 eV 63
4.1.3 Atomic and Molecular Beams 64
4.2 Chemical Origins 64
4.2.1 Thermal Annealing 64
4.2.2 Hydrogen Annealing 65
5. **Breakdown of SiO$_2$** 66
6. **Summary** 67
Acknowledgements 68
Glossary 68
Works Cited 68
Further Reading 70

INTRODUCTION

Information technology today owes much to the progress of large-scale integration (LSI) technology. This has been achieved by scaling down and integrating devices. The high reliability and high production yield indispensable to the widespread use of LSI, on the other hand, are due to the excellent properties of the SiO$_2$–Si interface. This is because most current LSI uses metal–oxide–semiconductor (MOS) transistors as switching devices.

Although the idea of MOS transistors had

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

been proposed in the 1920s (Lilienfeld, 1926), premature technology prevented the device operation. The study of the SiO$_2$/Si interface started in the 1940s in search of the origin of the shielding layer that obstructed the switching operation of MOS transistors, and during these studies bipolar transistor action was discovered in 1947 (Bardeen and Brattain, 1948). This discovery shifted the research activities from the interface to the Si substrate.

During the development of bipolar transistors, anomalous properties of junctions emerged as reliability problems. As researchers realized that these anomalies were not a sole result of bulk properties, the study of the interface was reactivated in the 1950s in an effort to passivate the Si surface. Thus the development of a high-quality SiO$_2$ surface led to a successful implementation of MOS transistors in 1960 (Kahng and Atalla, 1960).

In the early MOS transistors, the most important issue was mobile charges in the SiO$_2$ because these charges (which were later attributed to Na contamination) degraded the devices as a result of operation. The 1964 invention of MOS stabilization treatment using P$_2$O$_5$ solved this problem. In the 1960s, the interface trap density was reduced to the low 10^{10}-cm^{-2} range by annealing the MOS transistors in an atmosphere containing hydrogen. Because such a low trap level causes negligible device degradation, since the mid-1970s interest in the interface focused not on the characteristics of the as-made interface, but on its degradation after fabrication. One kind of degradation is induced by the device operation and is known as the hot-carrier effect (Ning *et al.*, 1975). Another kind is due to exposure to ionizing radiation (see, for example, Ma and Dressendorfer, 1989). The interest in this problem followed the failure of the telecommunication satellite Telstar 1, which passed through the Van Allen belt in 1962. Although the Telstar failure was in a bipolar transistor, MOS transistors were also found to be vulnerable to such radiation. These two kinds of postfabrication degradation spurred the investigation of carrier traps at the SiO$_2$–Si interface as well as in bulk SiO$_2$. For more details of the history, refer to the publication by Nicollian and Brews (1982).

From a practical viewpoint the hot-carrier effect is still a crucial issue in device development. Furthermore, a better SiO$_2$–Si interface has recently proved essential for making low-power dynamic random-access memories (DRAMs), and for making nonvolatile memories of flash type. These devices require the interface to be controlled far more strictly and will thus stimulate new activities in this field.

1. PREPARATION OF SiO$_2$

1.1 Thermal Oxidation

A high-quality SiO$_2$–Si interface is usually made by oxidizing a Si substrate in an atmosphere containing oxygen at a temperature ranging from 800 to 1100 °C. A steam ambient is widely used in the oxidation both of thin oxides for gate isolation (10–20-nm thick) and of thick oxides for device isolation (about 500-nm thick). This is because the steam oxide grows faster than the dry oxide. Another reason is the empirically superior breakdown tolerance of the steam oxide. The oxidation is carried out in a high-purity quartz tube oriented vertically in order to avoid the influx of open air by convection. Figure 1 shows an example of the kind of furnaces used industrially. Before the Si substrate is oxidized, it is cleaned to reduce particulate and metallic contamination to or below the levels of 0.1 particle/cm^2 and 1×10^{10} atoms/cm^2.

The oxide growth at the SiO$_2$–Si interface is due to the diffusion of oxidizing molecules, such as O$_2$ and H$_2$O, through the oxide (see Grove, 1967). This means that the interface is continuously renewed during the oxidation and that the properties of the interface are determined mostly by the final step of the oxidation. In the case of SOI (silicon-on-insulator) substrates, the top Si is oxidized at the interfaces with both the surface and buried oxide. The oxide that grows at the top-Si/buried-oxide interface increases the thickness of the buried oxide, as shown in Fig. 2.

1.2 Formation of SiO$_2$ Other Than by Oxidation

1.2.1 Chemical Vapor Deposition (CVD) A SiO$_2$ film can also be formed by the thermal reaction of monosilane (SiH$_4$) or

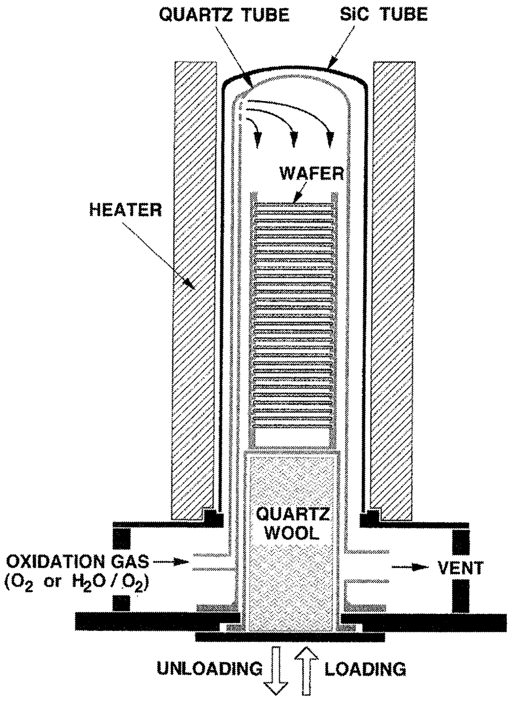

FIG. 1. Vertical oxidation furnace.

dichlorosilane (SiH_2Cl_2) with oxidizing gas in a reactor kept at 10–30 Pa to get a uniform and conformal film. This is called chemical vapor deposition [see CVD (CHEMICAL VAPOR DEPOSITION)]. Various combinations of reactant gases are used depending on the deposition temperature, and one of the criteria considered in selecting the gases is their thermal stability. A mixture of SiH_2Cl_2 and N_2O is used in the deposition at temperatures over 850 °C, below which SiH_4 is used instead of SiH_2Cl_2. Deposition at a temperature below 550 °C is done using a gas mixture of SiH_4 and O_2 to increase the growth rate. It is usually at atmospheric pressure.

The CVD oxides applied to poly-Si gate MOS capacitors have an interface quality comparable to that of thermal oxides. MOS transistors fabricated on polycrystalline Si films in static random access memory (SRAM) use a CVD oxide as the gate oxide, because the thermal oxide formed on poly-Si films has poor insulation characteristics due to the crystal defects.

1.2.2 Liquid-Phase Deposition (LPD)

This is an extreme case of low-temperature preparation developed by Nagayama *et al.* (1988). Oxide film is deposited in a silica-supersaturated solution prepared by dissolving silica powder in hydrofluorosilicic acid (H_2SiF_6). The deposition proceeds, even at room temperature, by the reaction

$$H_2SiF_6 + 2H_2O \Leftrightarrow 6HF + SiO_2$$

and is facilitated by continuously adding boric acid, which pushes this reaction toward SiO_2 formation by consuming the hydrofluoric acid:

$$H_3BO_3 + 4HF \Leftrightarrow BF_4^- + H_3O^+ + 2H_2O.$$

The LPD oxide capacitor has an interface

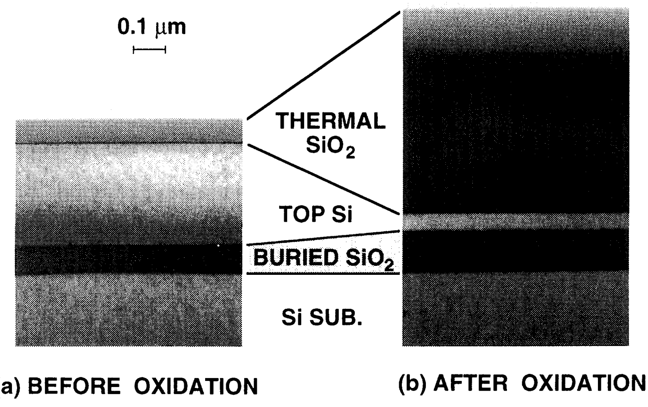

(a) BEFORE OXIDATION (b) AFTER OXIDATION

FIG. 2. Cross-sectional TEM images of SOI (silicon-on-insulator) wafers. The thickness of the buried oxide increased by 40 nm after oxidation in a steam ambient at 1350 °C. After Nakashima *et al.* (1994); © 1994 IEEE.

trap density of about 1×10^{11} cm^{-2} eV^{-1} at the middle of the Si band gap, one order of magnitude greater than that of the thermal oxide. For repeated gate-voltage scans, the capacitor exhibits different capacitance-vs-voltage curves, indicating the existence of bulk-oxide traps (Yeh *et al.,* 1994).

2. CHARACTERIZATION OF THE SiO$_2$–Si INTERFACE

2.1 Morphology

High-resolution transmission electron microscope (TEM) images of the SiO$_2$–Si interface indicate the existence of an undulation with an amplitude corresponding to one or two atomic layers of Si. We should note, however, that the actual interface is not necessarily as flat as it seems because TEM images are formed by electronic diffraction and scattering through the samples. This means that for any structure to be detected it must have sufficient extent along the direction of electron-beam incidence. Theoretical analyses of TEM imaging indicate that the actual roughness corresponds to more than four atomic layers (Goodnick *et al.,* 1985; Ohdomari *et al.,* 1986). The results of atomic force microscope (AFM) surface profiling are consistent with the larger undulation.

The SiO$_2$ is mostly amorphous even near the interface, but there is some evidence that a crystalline phase exists in the thermal oxide. Rochet *et al.* (1989) observed epitaxial crystalline growth of SiO$_2$ along the [1$\bar{1}$0] ledges on a vicinal (001) Si surface by using reflection high-energy electron diffraction (RHEED) together with TEM. Takahashi *et al.* (1993) also observed the crystalline phase on a (001) Si surface by using x-ray crystal truncation rod (CTR) scattering analysis, and they concluded that the crystalline phase not only exists at the interface but that it also makes up a few percent of the volume of the bulk SiO$_2$.

2.2 Bonding Structure

Grunthaner and Grunthaner (1986) and Himpsel *et al.* (1988) have used x-ray photoelectron spectroscopy (XPS) to investigate the chemical bonding structure of the SiO$_2$–Si interface, and an electron spectroscopy for chemical analysis (ESCA) spectrometer is widely used in XPS. The x ray of the Al $K\alpha$ line (1486 eV) is monochromatized and used to irradiate the samples in a chamber kept at a pressure of 1×10^{-7} Pa or less. The energy spectrum of photoelectrons from the Si $2p$ core level is measured with a multichannel analyzer. The sensitivity with which the interfacial Si atoms are detected can be improved by using photons with an energy of about 130 eV radiated from a synchrotron. This is because the escape depth of the photoelectrons is about 3 Å in Si at this photon energy, whereas the escape depth in the conventional spectrometer is about 15 Å. In the case of 130-eV photon irradiation, only a small fraction of the photoelectrons from the Si substrate are detected because of the shorter escape depth. This suppression of the Si signal makes the signal from the interface more visible. The XPS measurement is limited to interfaces covered by an oxide whose thickness is less than 10 times the escape depth.

Figure 3 shows an example of the XPS analysis. The spectrum consists of a number of peaks. These peaks correspond to the shifted core-level binding energies that reflect the valence charge redistribution induced by the change in the local atomic arrangements. This redistribution is known as chemical shift. The peaks are associated with the intermediately oxidized Si atoms—Si$^+$ (Si$_3$≡Si–O), Si^{2+} (Si$_2$=Si=O$_2$), and Si^{3+} (Si–Si≡O$_3$)—that form the transition region at the SiO$_2$–Si interface.

The total amount of the intermediately oxidized Si atoms indicates that the interfacial transition region extends for about one (Grunthaner and Grunthaner, 1986) or two (Himpsel *et al.,* 1988) layers of Si. As shown in Fig. 3, for each of the Si surface orientations the intensity of Si^{3+} is far greater than expected from the ideal image of the local atomic geometry at the interface. This greater intensity of Si^{3+} also suggests that the interface is rougher than is observed by TEM.

2.3 Electrical Characterization

2.3.1 Theoretical Evaluation of Interface Traps Carrier traps at the SiO$_2$–Si interface exchange charges with the Si surface,

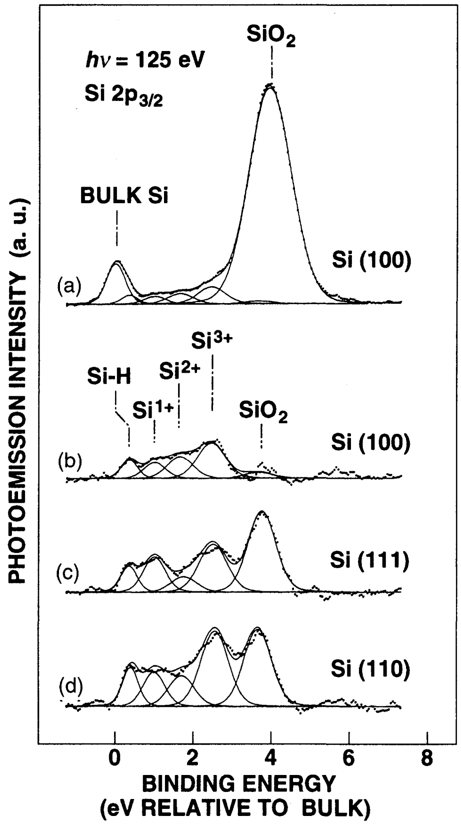

FIG. 3. Photoelectron spectra of a 2-nm-thick layer of oxide on Si. (a) The original spectrum on (100) Si. (b)–(d) The spectra after the removal of the bulk-Si and SiO$_2$ components. (After Niwano *et al.*, 1991.)

and this leads to an important property: Their charges change depending on the energy level relative to the Fermi potential. This property results from the fact that the trap energy level is located in the Si band gap. The interface traps are classified into two groups according to their charging characteristics. One consists of acceptor-type traps that are negatively charged when the energy level is below the Fermi level and are otherwise neutral. The other group consists of donor-type traps that are neutral when the level is below the Fermi level and are otherwise positive.

Sakurai and Sugano (1981) theoretically investigated the trap states assuming the Bethe lattice structure and obtained the following results, which are consistent with experiments. A perfect interface has no states in the Si band gap even if the angle of adjacent Si–O–Si bonds is varied. A dangling bond of oxygen has no band-gap state either. It is the dangling bond of Si backbonded to three other Si atoms that becomes a trap located at the middle of the Si band gap. Besides this bond, extended Si–Si bonds (with a bonding orbital) between $Si_3{\equiv}Si\cdot$ and $\cdot Si{\equiv}O_3$ produce trap states in the lower half of the Si band gap. They are donorlike because of the bonding orbitals. They might be regarded as oxygen vacancies. On the other hand, weak Si–O bonds (with an antibonding orbital) between $Si_3{\equiv}Si\cdot$ and $\cdot O{-}Si$ produce trap states in the upper half of the band gap. They are acceptorlike because of the antibonding orbitals. As shown in Fig. 4, the energy levels of both types approach the midgap when the bond length increases. Assuming a practical distribution of bond lengths, Sakurai and Sugano reproduced the U-shaped distribution of trap states that is

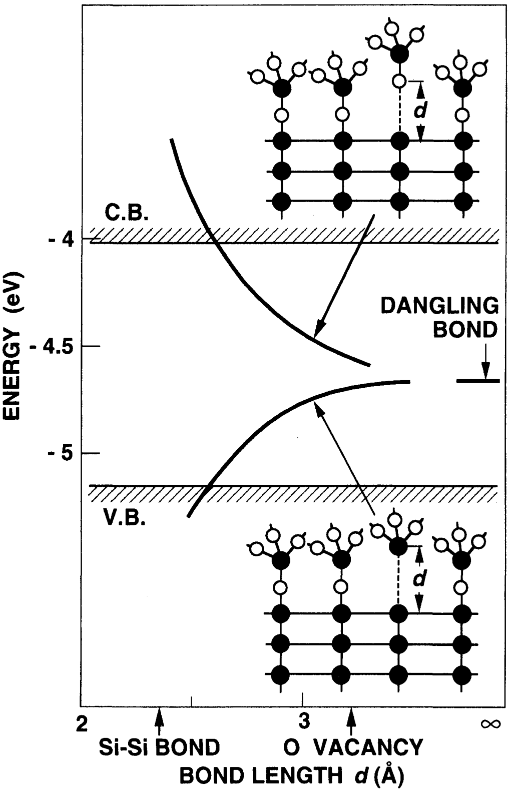

FIG. 4. Theoretical energy levels of Si–O and Si–Si weak bonds at the SiO$_2$–Si interface. Insets show the atomic configuration model used in the calculation. (After Sakurai and Sugano, 1981.)

widely observed. They also showed by calculation that the trap levels of both types go into either the valence band or the conduction band using H, Cl, F, or OH to terminate the dangling bonds.

2.3.2 Measurement of Interface Traps

2.3.2.1 Quasistatic Method A voltage applied to an electrode formed on a SiO$_2$–Si structure controls the potentials (relative to the Fermi potential) of interface-trap levels and changes the interface-trap charges (see Sec. 2.3.1). This change is detected by various capacitance methods and helps us characterize the traps. The quasistatic method developed by Berglund (1966) is the most popular among the capacitance methods because of its simplicity. In this method a low-frequency small modulation is superimposed on the voltage applied to the electrode (called gate after the naming in MOS transistors; see Sec. 3.1). The interface-trap and Si surface charges follow the slow modulation, and the entire MOS capacitance C_{LF} is given by $C_{LF} \equiv dQ_T/dV_G$, where Q_T stands for the areal density of total charges and V_G is the gate voltage. From the charge neutrality requirement and Maxwell's equation we get $Q_T = -(Q_{it} + Q_s + Q_{ox})$ and $Q_T = C_{ox}V_{ox}$, where Q_{it} is the areal density of interface-trap charges, Q_s is the areal density of Si surface charges, Q_{ox} is the areal density of bulk-oxide charges, C_{ox} is the oxide capacitance, and V_{ox} is the voltage across the oxide layer. We should take care that Q_{ox} is not a simple integration of oxide charge density $\rho(x)$, but a moment with respect to the gate given by

$$Q_{ox} = \frac{1}{d}\int_0^d x\rho(x)\,dx, \tag{1}$$

where d is the thickness of the oxide layer (Fig. 5). The voltage V_{ox} is given by $V_{ox} = (V_G - \psi_s) - (V_{G0} - \psi_{s0})$, where ψ_s is the band bending at the Si surface, and V_{G0} and ψ_{s0} are the values of V_G and ψ_s when $V_{ox} = 0$. Differentiating the above equations with respect to ψ_s, we get $C_{LF} = C_{ox}(C_S + C_{it})/[C_{ox} + (C_S + C_{it})]$, where $C_{it}(\psi_s) \equiv -dQ_{it}/d\psi_s$ and $C_s(\psi_s) \equiv -dQ_s/d\psi_s$ are the interface-trap capacitance and the Si surface capacitance. The interface-trap capacitance is related to the trap level density D_{it} (cm^{-2} eV^{-1}) at the Fermi level ϕ_s through the equation $D_{it}(\phi_s) = C_{it}(\psi_s)/q^2$, where q is the electronic charge. The Fermi level ϕ_s is equal to $\phi_s = \psi_B - \psi_s$, where ψ_B is the bulk Fermi level. Both ϕ_s and ψ_B are measured from the intrinsic level E_i ($\psi_B > 0$ for p-type Si).

The capacitance meter fails to provide accurate results in low-frequency modulation measurements because the signals are noisy.

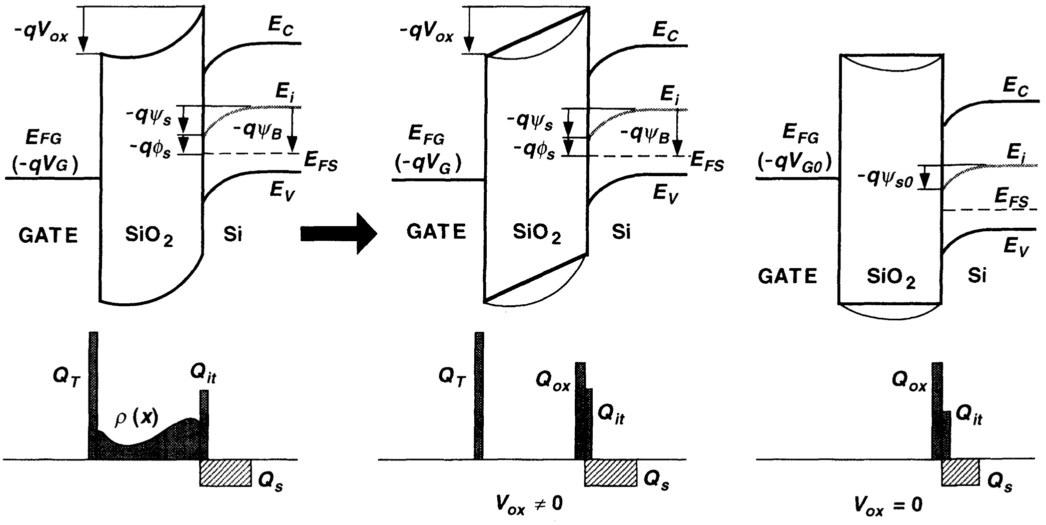

(a) **ACTUAL DISTRIBUTION** (b) **DISTRIBUTION ASSUMED IN CAPACITANCE MEASUREMENT**

FIG. 5. Energy-band diagrams of MOS capacitors with bulk-oxide charges.

Accordingly, C_{LF} is usually obtained by measuring the displacement current $I_G = C_{LF}dV_G/dt$ induced by varying the gate voltage slowly, from which this method is named *quasistatic.*

The Si surface capacitance is theoretically calculated as (see Sze, 1981; Nicollian and Brews, 1982)

$$C_s(\psi_s) = \left|\frac{\epsilon_s}{2L_{Di}} \frac{[e^{\beta\psi_B}(1 - e^{-\beta\psi_s}) + e^{-\beta\psi_B}(e^{\beta\psi_s} - 1)]}{F(\beta\psi_s,\beta\psi_B)}\right| \quad \left(\beta \equiv \frac{q}{kT}\right), \qquad (2)$$

where ϵ_s is the dielectric constant of Si, $L_{Di} = \sqrt{\epsilon_s kT/2q^2 n_i}$ is the intrinsic Debye length, n_i is the intrinsic carrier density of Si, k is Boltzmann's constant, T is temperature, and $F(\beta\psi_s,\beta\psi_B)$ is the normalized electric field at the Si surface defined by

$$F(\beta\psi_s,\beta\psi_B) = [e^{\beta\psi_B}(e^{-\beta\psi_s} + \psi_s - 1) + e^{-\beta\psi_B}(e^{\beta\psi_s} - \psi_s - 1)]^{1/2}.$$

Thus we can get the interface-trap density by measuring C_{LF} and using the following equation:

$$D_{it}(\phi_s) = \frac{1}{q^2}\left[\frac{C_{ox}C_{LF}(V_G)}{C_{ox} - C_{LF}(V_G)} - C_s(\psi_s)\right]. \qquad (3)$$

Figure 6 shows examples of quasistatic capacitance curves of MOS capacitors fabricated on p-type Si. The samples, under a positive gate bias (4.3 V), were irradiated with x rays from a Mo tube. As the x-ray dosage increases, the capacitance curve is distorted by the interface-trap buildup (see Sec. 4.1.1).

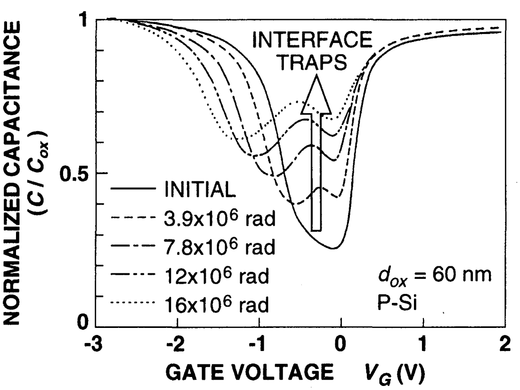

FIG. 6. Quasistatic capacitance curves of MOS capacitors under a positive bias V_G = 4.3 V and irradiated with x rays from a Mo-cathode tube (50 keV). (After Watanabe *et al.*, 1988.)

Berglund (1966) developed an experimental method to get the bandbending ψ_s as a function of V_G. He differentiated the equations $Q_T = C_{ox}V_{ox}$ and $V_{ox} = (V_G - \psi_s) - (V_{G0} - \psi_{s0})$ with respect to V_G and obtained

$$\psi_s = \int_{V_{G1}}^{V_G}\left[1 - \frac{C_{LF}(V)}{C_{ox}}\right]dV + \psi_s|_{V=V_{G1}}, \qquad (4)$$

where V_{G1} is the initial gate voltage. If we start from the strong accumulation or inversion, $\psi_s|_{V=V_{G1}}$ is approximately equal to $(E_V - E_{FS})/q$ or $(E_C - E_{FS})/q$ (where E_{FS} is the Fermi level in the bulk Si) in p-type Si. By changing the gate voltage, we can control ψ_S and get the interface-trap density as a function of the trap level in the Si band gap.

2.3.2.2 High-Low-Frequency Capacitance Method When using the quasistatic method, we need to know the Si surface capacitance as a function of the gate voltage, and this capacitance is usually calculated theoretically using Eq. (2). This requires us to know the dopant distribution in the Si substrate, and this knowledge is not always easy to get. Castagné and Vapaille (1971) therefore developed the high-low-frequency method, in which the surface capacitance is obtained experimentally from a high-frequency capacitance measurement in the following way. The interface-trap level shifts relative to the Fermi level as the gate voltage changes. Those traps that crossed the Fermi level exchange carriers with a Si conduction or valence band near the interface. The exchange of carriers has a noninfinitesimal time constant τ given by $\tau = 1/(\sigma_n v_{th} n_s + \sigma_p v_{th} p_s)$, where σ_n and σ_p are, respectively, the electron and hole capture cross sections; n_s and p_s are the electron and hole concentrations near the interface; and v_{th} is the carrier thermal velocity (Iunovich, 1958). Because of this time constant, the traps respond to the gate-voltage modulation like a resistor R_{it} and a capacitor C_{it} in series: $\tau = R_{it}C_{it}$.

Because of the resistive element R_{it} of the interface traps, the entire circuit responds to the high-frequency (about 1 MHz) modulation as if there were no interface traps. This is an ideal case (no traps), and the impedance of the entire MOS system is given by a capacitance $C_{HF} = C_{ox}C_S/(C_{ox} + C_S)$. In the lower limit of the modulation frequency, we can regard the interface traps as a pure capacitance, and the entire system is reduced to the capacitor C_{LF}. Because C_S is the same for these two cases if the gate voltage is the same, the capacitance associated with the interface-trap charges is given by

$$C_{it} = \left(\frac{1}{C_{LF}} - \frac{1}{C_{ox}}\right)^{-2} - \left(\frac{1}{C_{HF}} - \frac{1}{C_{ox}}\right)^{-1}. \tag{5}$$

Figure 7 shows the high-frequency capacitance curves of the same samples as in Fig. 6. The negative shift of the *CV* curves is due to positive charges produced by the x-ray irradiation (see Sec. 4.1.1). Slight distortions caused by the interface traps are also seen in the curves. Figure 8 shows the interface-trap level distribution in the Si band gap derived from Figs. 6 and 7 by using the method of this section.

2.3.2.3 Charge Pumping Brugler and Jespers (1969) developed the charge-pumping method as a probe of SiO₂–Si interface traps. As shown in Fig. 9, the source and drain of MOS transistors are connected together and reversely biased with respect to the substrate. A pulse applied to the gate drives the surface between inversion and accumulation. The interface traps in *p*-type Si are filled with carriers during the inversion

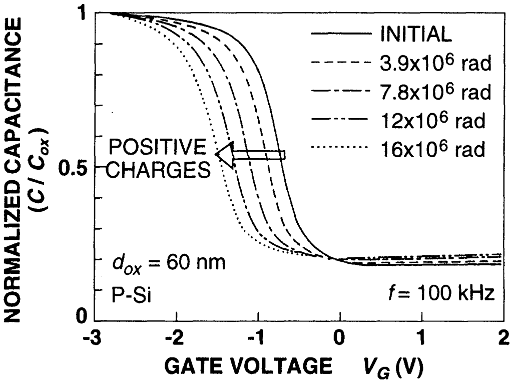

FIG. 7. High-frequency capacitance curves of the same samples as in Fig. 6. Measured at 100 kHz. (After Watanabe, unpublished.)

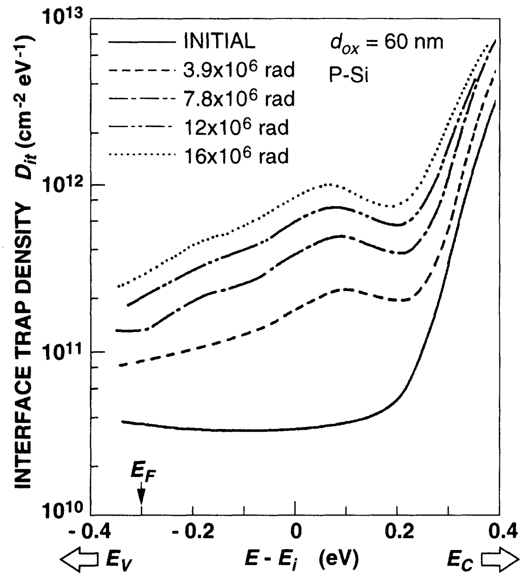

FIG. 8. Interface-trap density of the same samples as in Figs. 6 and 7. (After Watanabe *et al.*, 1988.)

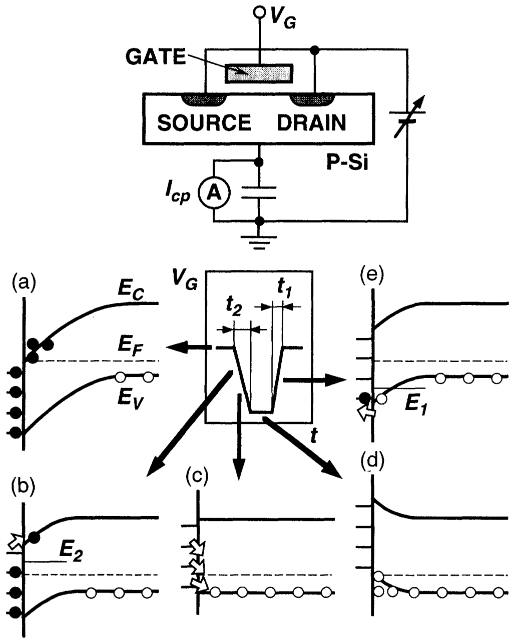

FIG. 9. Experimental setup and band diagrams illustrating the charge-pumping method.

period [Fig. 9(a)]. When the surface turns from inversion to accumulation, the inversion-layer carriers flow into the source and drain first. Most of the trapped carriers subsequently recombine with the majority carriers in the substrate [Fig. 9(c)]. They produce a dc component of the substrate current given by

$$I_{cp} = qS_G f_{cp} \int_{E_1}^{E_2} D_{it} dE_{it}, \qquad (6)$$

where S_G is the gate area, f_{cp} is the pulse frequency, and D_{it} is the interface-trap density (cm^{-2} eV^{-1}). Some of the trapped carriers ($E > E_2$) are emitted into the conduction band (the edge of which is E_C) and do not contribute to the substrate current [Fig. 9(b)]. On the other hand, some traps ($E < E_1$) capture electrons from the valence band [the edge of which is E_V, Fig. 9(e)] and reduce the substrate current. These electron-emission and -capture processes ($E_2 \rightarrow E_C$, $E_V \rightarrow E_1$) are suppressed by keeping the transition times t_2 (inv. → acc.) and t_1 (acc. → inv.) short. When t_2 is too short, however, some of the inversion-layer carriers do not flow into the source or drain but instead recombine with the substrate majority carriers. This effect produces an additional substrate current ("geometric component") that has nothing to do with the interface traps. The use of short-channel MOS transistors (less than 1 μm) is effective in avoiding this effect.

Charge pumping is a direct method of measuring the interface traps. It is applicable to MOS transistors with a gate area as small as 1 μm^2. The capacitance method, in contrast, requires a gate area of at least 10^4 μm^2 when the commercially available capacitance meter is used. The basic charge-pumping method [Eq. (6)] provides only the average value of D_{it}, and many subsequent studies have tried to extract the distribution of D_{it} by changing the high or low levels of the gate pulse that controls the values of E_2 and E_1 (Haddara and Christoloveanu, 1986).

2.3.2.4 Electron-Spin Resonance The microscopic nature of the interface traps was revealed by electron-spin-resonance (ESR) measurements (see ELECTRON PARAMAGNETIC RESONANCE). ESR detects electronic spins associated with dangling bonds, and from the resonance absorption line we can get information about their structure. Dangling bonds, named P_b centers, which have a resonance g value of 2.008, are found at the SiO$_2$–Si interface. Poindexter *et al.* (1984) showed experimentally that the P_b center is a dangling bond of an interfacial Si atom backbonded to three other Si atoms, just as predicted by Sakurai and Sugano (1981). The experiments using unannealed samples (Poindexter *et al.*, 1984) and radiation-damaged samples (Lenahan *et al.*, 1981) showed that some of the interface traps indeed are P_b centers.

Not all of the interface traps are P_b centers, however. Cartier *et al.* (1993) reported a case in which only 10% of the generated interface traps were P_b centers. This result is another support for the extended bond model of interface traps (see Sec. 2.3.1).

3. THE SiO$_2$–Si INTERFACE AND DEVICE PERFORMANCE

This section reviews the ways in which the SiO$_2$–Si interface affects the performance of MOS transistors, which are the most important components of current LSI. There are two ways for the interface traps to degrade the performance of MOS transistors: by degrading switching performance and by increasing junction leakage current.

3.1 Operational Principle of MOS Transistors

First we briefly review the basic operational principle of MOS transistors, shown schematically in Fig. 10. The details are given in the textbook by Sze (1981). The MOS transistors are composed of three electrodes. Two of them, called source and drain, are heavily doped n-type regions in the lightly doped p-type Si substrate. The source is usually used as a voltage reference, and the electrical connection between the source and drain is controlled by a third electrode (gate) that is placed above the Si substrate and insulated from it by SiO$_2$ (gate oxide). A positive bias applied to the gate induces a two-dimensional electron layer, called an inversion layer, in the surface region of the Si. Thus electrons driven by the positive bias applied to the drain flow from the source to the drain. This electron flow is called drain current. Without the positive

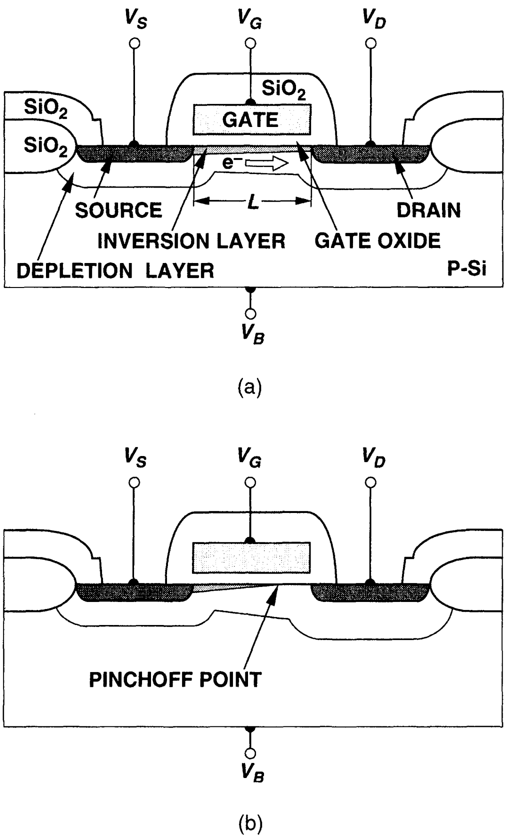

FIG. 10. Schematic cross sections of MOS transistors operated in (a) linear and (b) saturation regions.

gate bias, the drain is reversely biased to the substrate and electrically isolated from the source. A MOS transistor of this type is called an *n*-channel transistor, because the conduction layer (channel) is fed by negative carriers (electrons). We treat only *n*-channel MOS transistors in any detail in this article.

The speed of the LSI circuit is determined by two factors, one of which is the speed with which the transistor itself is driven into the "on" state. Before the formation of the inversion layer, a space-charge layer is formed below the gate, and most of the interface traps are filled by carriers. Because of these processes only a small number of carriers are induced in the Si surface during the buildup period of gate voltage. The drain current I_D depends exponentially on the gate voltage in this period, and we can use a subthreshold swing S, which is defined as the gate-voltage swing necessary to increase the drain current by a factor of 10, as a measure of the switching speed of MOS transistors. In the presence of interface traps, S is given by

$$S \equiv \frac{dV_G}{d\log_{10}I_D} = S_0\frac{1+(C_D+C_{it})/C_{ox}}{1+C_D/C_{ox}}, \qquad (7)$$

where $S_0 \approx (kT/q)\ln 10(1 + C_D/C_{ox})$ is the subthreshold swing in the absence of interface traps and C_D is the depletion-layer capacitance, which is equal to ϵ_s/x_D (where x_D is the width of the depletion layer). Equation (7) shows that the traps have an effect of adding a capacitance C_{it} to the system in parallel with the depletion capacitance, thus retarding the formation of the inversion layer. The smaller the subthreshold swing is, the faster the MOS transistor is driven into the "on" state.

The other factor determining the circuit speed is the magnitude of the drain current. The signal of component circuits (such as inverters) is output after charging or discharging stray or intentionally prepared capacitors through MOS transistors. Accordingly, the magnitude of current that the transistor can drive (drain current) determines the circuit speed. The drain current depends on the applied voltages in a different way for gate voltages $V_G - V_T > V_D$ (linear region) and $V_G - V_T < V_D$ (saturation region). Here V_D is the drain voltage and V_T is the minimum gate voltage for formation of the inversion layer (called threshold voltage). In the linear region, an almost uniform inversion layer is formed, and the carrier drifts along the channel driven by the field $F_C = V_D/L$ (where L is the channel length). If μ is the carrier mobility, the drift velocity v is expressed as $v = \mu F_C$. The inversion charge density is equal to the oxide capacitance multiplied by the potential difference between the gate and the inversion layer. If we put qN_I as the charge density near the source, the charge density near the drain is equal to $qN_I - C_{ox}V_D$ because the potential of the inversion layer is lower than that near the source by the amount V_D. Thus the average charge density is given by $qN_I - C_{ox}V_D/2$, and the drain current I_D is given by

$$I_D = W(qN_I - C_{ox}V_D/2)\mu V_D/L, \qquad (8)$$

where W is the channel width and q is the electron charge. For a small value of V_D, this equation is reduced to a linear function of V_D.

In the saturation region, on the other hand, both N_I and v vary along the conduction path. In the lowest-order approximation, $N_I(x)$ and $v(x)$ may be replaced by the averages $N_I/2$ (N_I is the carrier density at the source) and v. The factor $\frac{1}{2}$ in the average carrier density is due to the density being N_I at the source and about zero at the pinchoff point, where the surface potential is $V_G - V_T$. Thus the drain current is expressed as

$$I_D = \tfrac{1}{2}WqN_I v.$$

The drift velocity is again expressed as $v = \mu F_C$, where F_C is the average field, which is equal not to V_D/L (V_D is the drain voltage), but to $(V_G - V_T)/L$. This is because the drain current remains constant even when the drain voltage V_D increases beyond the pinchoff voltage $V_P = V_G - V_T$. The excess voltage $V_D - V_P$ is imposed on the depleted region near the drain and has little effect on the enhancement of the current. Thus we get

$$I_D = W\mu(qN_I)^2/2LC_{ox}, \tag{9}$$

where $qN_I = C_{ox}(V_G - V_T)$. We note that the value given by this equation does not depend on the value of V_D and is equal to the maximum of Eq. (8) (at $V_D = qN_I/C_{ox} = V_G - V_T$). The term *saturation* comes from these results.

From a practical point of view, the saturation region is much more important than the linear region. In order to show this we briefly review the operational principle of an inverter, which is one of the elementary components of an LSI circuit. The inverter consists of an MOS transistor and a load connected in series (Fig. 11). The signal (high or low voltage) is input to the gate of the transistor, and the inverted signal is output to the drain. Now we consider the case where the input voltage turns from low to high. When the input voltage is low, the drain voltage is high because the load is connected to the power supply, and no current flows through the transistor. During the buildup period of the input voltage, drain current begins to flow, and the transistor is in the saturation region because the gate voltage is not high enough. The drain current discharges stray capacitors connected to the drain and decreases the drain voltage. The operational state of the transistor then turns into the linear region. Accordingly, the switching time is mostly determined by the saturation drain current.

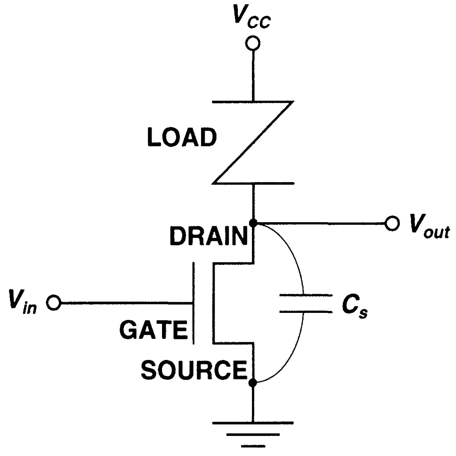

V_{CC} : POWER SUPPLY VOLTAGE
C_s : STRAY CAPACITOR

FIG. 11. Inverter.

The mobility is physically defined only for low fields where the velocity is proportional to the field. In practical applications, however, the mobility in the nonlinear region is extensively used for the analysis of device characteristics. In this case we regard the mobility as defined by the equation $\mu = v/F_C$. Then μ is a function of F_C, expressed by Thornber (1980) as

$$\mu = \mu_0\left[1 + \frac{(\mu_0 F_C/u_l)^2}{G + \mu_0 F_C/u_l} + \left(\frac{\mu_0 F_C}{v_{sat}}\right)^2\right]^{-1/2} \tag{10}$$

where μ_0 is the mobility at low fields, u_l is the longitudinal acoustic phonon velocity, G is a fitting parameter, and v_{sat} is the saturation velocity. It is easily seen that $\mu \approx \mu_0$ for $F_C \ll u_l/\mu_0$, whereas $\mu \approx v_{sat}/F_C$ for $F_C \gg u_l/\mu_0$. Typical values of the parameters used in the simulation are listed in Table 1.

3.2 Degradation of MOS Transistor Performance

3.2.1 Increase in Subthreshold Swing

The subthreshold swing S in current MOS transistors is from 70 to 80 mV, and Eq. (7)

Table 1. Parameters of mobility. After Yamaguchi (1983).

	u_l (cm/s)	G	v_{sat} (cm/s)	μ_0 (cm^2/Vs)
Electron	4.9×10^6	8.8	1.04×10^7	1400
Hole	2.93×10^6	1.6	1.0×10^7	480

shows that S increases appreciably when the interface-trap capacitance C_{it} becomes comparable to the depletion-layer capacitance $C_D \approx 3 \times 10^{-8}$ F/cm^2. The trap density of as-fabricated transistors is little more than 10^{10} cm^{-2} eV^{-1} and corresponds to $C_{it} \approx 5 \times 10^{-9}$ F/cm^2, which is an order of magnitude less than C_D. Thus, the generation of interface traps after fabrication is the dominant factor that increases the subthreshold swing.

3.2.2 Reduction of Drain Current (1): Loss of Surface Carriers In n-channel MOS transistors under operation, the Fermi level is close to the conduction-band edge in the channel. Accordingly, the acceptor-type interface traps (areal density $N_{it,C}$) are negatively charged and reduce the inversion carrier density from N_I to $N_I - N_{it,C}$, leading to I_D degradation as seen From Eq. (9). Similar phenomena are observed in p-channel MOS transistors where the Fermi level approaches the valence-band edge, and donor-type interface traps that are positively charged reduce N_I. Because the inversion carrier density in current devices is about 8×10^{12} cm^{-2}, the interface traps of as-fabricated devices ($<5 \times 10^{10}$ cm^{-2}) have a negligible effect. Device operation, however, generates trap levels beyond the 4×10^{11}-cm^{-2} level that cause a 10% reduction of the drain current.

3.2.3 Reduction of Drain Current (2): Decrease in Mobility The interface traps reduce the low-field mobility μ_0 in Eq. (10) through Coulomb scattering, phonon scattering, and surface-roughness scattering. Assuming Matthiessen's rule, we can express the mobility as

$$\frac{1}{\mu_0} = \frac{1}{\mu_C} + \frac{1}{\mu_{ph}} + \frac{1}{\mu_{SR}},$$

where μ_C, μ_{ph}, and μ_{SR} are the mobilities determined by the three scattering processes.

3.2.3.1 Screened Coulomb Scattering This process limits the mobility when the inversion carrier density is low. The scattering of carriers is caused by charged centers: space charges (ionized dopants) in the depletion layer (qN_{depl}), charges in the gate oxide (qN_{ox}), and interface charges ($qN_{it,C}$). It should be noted that only the charged interface traps ($N_{it,C}$), not the interface traps as a whole (N_{it}), contribute to this process. Sah *et al.* (1972) expressed the results of their theoretical and experimental analysis as follows:

$$\mu_C \propto TN_I^{\alpha}/N_C \qquad (\alpha \approx 1.2\text{–}1.4), \tag{11}$$

where T is the temperature and N_C is the total density of scattering centers (given by $N_C = N_{depl} + N_{ox} + N_{it,C}$). The term N_I^{α} (missing in their original report) represents the screening effect by the carriers themselves, which effect was grounded theoretically by Stern and Howard (1967). When N_I is large, the mobility is less sensitive to N_I and is not approximated by such a simple formula.

By reducing N_I, the interface traps reduce the mobility, as seen from Eq. (11). They further reduce the mobility by increasing N_C. In as-fabricated practical devices, both the oxide and interface charges are little more than 10^{10} cm^{-2}, and their contribution to N_C is negligible when compared to that of the depletion-layer charges, $N_{depl} \approx 10^{12}$ cm^{-2}. Thus, the reduction of mobility by Coulomb scattering is mostly caused by the decrease in the number of inversion carriers.

3.2.3.2 Phonon Scattering This scattering is caused mostly by acoustic phonons. Kawaji (1969) considered the two-dimensional nature of the inversion carriers and obtained the result

$$\mu_{ph} = \frac{\hbar^3 q \rho c_l^2 \langle z \rangle}{\Xi^2 k_B T m_t^2} \propto \frac{1}{TF_{eff}^{1/3}}, \tag{12}$$

where $\hbar$ is Planck's constant divided by 2π, ρ is the mass density, c_l is the longitudinal sound velocity, $\langle z \rangle$ is the effective thickness of the inversion layer (proportional to $F_{eff}^{-1/3}$; Kawaji, 1969), Ξ is the deformation potential (12 eV; Ezawa *et al.*, 1971), m_t is an electron effective mass parallel to the interface, and $F_{eff} = q(N_{depl} + \eta_{ph} N_I)/\epsilon_S$ and corresponds to the average field in the inversion layer. The value of η_{ph} was reported as $\frac{11}{32}$.

In this process, the interface traps increase the mobility by reducing F_{eff}. This effect is small, however, and as seen from Eqs. (8) and (9), the drain current itself decreases because of the reduction of inversion carrier density.

3.2.3.3 Surface-Roughness Scattering This process becomes dominant when the inversion carrier density is high, and it generally is the most important process because the device is usually operated under such a condition. The SiO$_2$–Si interface is not completely flat. The potential barrier that confines the inversion carriers outside SiO$_2$ fluctuates because of this roughness and scatters the inversion carriers in the same way as in Coulomb scattering. The variation ϕ of the barrier height is given as $\phi \approx F_{eff}\Delta$, where F_{eff} is the effective field given by $F_{eff} = q(N_{depl} + \eta_{SR}N_I)/\epsilon_S$, and Δ is the average displacement of the interface from flatness. Matsumoto and Uemura (1974) and Ando (1977) theoretically obtained

$$\mu_{SR} = \hbar^3\left\{F_{eff}^2 q m_t^2 \Lambda^2 \Delta^2 \int_0^{\pi} d\theta(1 - \cos\theta) \times \exp[-\tfrac{1}{2}k_{\mathrm{F}}^2\Lambda^2(1 - \cos\theta)]\right\}^{-1}, \tag{13}$$

and $\eta_{SR} = \frac{1}{2}$, where k_{F} is the wave number of the Fermi level and Λ is the correlation length of spatial variation of the interface undulation and is equal to $\frac{1}{3}$ for Gaussian-correlated undulations. Matsumoto and Uemura (1974) obtained $\Delta \approx 5$ Å and $\Lambda \approx 15$ Å for (100) electrons at 1.7 K.

In the surface-roughness scattering, the interface affects the mobility through the asperity Δ of the interface. The interface traps increase the mobility through the decrease in the effective field similarly to the way they do in phonon scattering.

Recent experiments by Takagi *et al.* (1990) give us a good view of the mobility as a function of inversion carrier density covering the phonon and surface-roughness scattering (Fig. 12). In the region of a moderate field and at room temperature, the mobility changes in proportion to $1/TF_{eff}^{1/3}$. As a matter of course we cannot observe this component at temperatures below 100 K. On the other hand, mobilities are less sensitive to the effective field than they are to the theoretical F_{eff}^{-2} for a higher field at high temperatures. This discrepancy is ascribed to the phonon scattering.

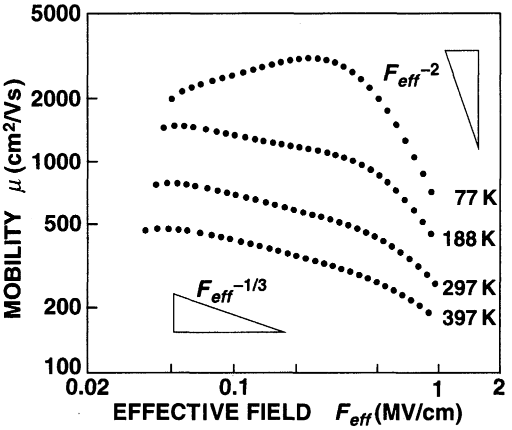

FIG. 12. Electron mobility in the inversion layer on (100) Si. Acceptor concentration of the substrate is 3.9×10^{15} cm^{-3}, and F_{eff} is defined by setting $\eta = 1/2$. (After Takagi *et al.*, 1990.)

3.3 Excessive Junction Leakage Current

3.3.1 Requirements of Low-Power DRAMs A *p-n* junction is formed between the source (or drain) and Si substrate of MOS transistors and is reversely biased. Leakage current, driven by the reverse bias, flows through the junction and causes LSI failures when it is generated excessively. DRAM is one of the LSIs that are sensitive to the leakage current. It stores one-bit information in a unit (called memory cell) that consists of a MOS transistor and a capacitor (Fig. 13). The capacitor consists of a pair of electrodes separated by an insulator and fabricated extending over the MOS transistor to reduce the memory-cell area in most DRAMs of 4 Mbits or more. When the logical "1" state is stored in a cell, there are more positive charges in the lower capacitor electrode (storage electrode) than in the case when the "0" state is stored. The potential of the plate electrode is kept constant. The MOS transistor switches the charging and discharging of the capacitor. The storage electrode is electrically isolated from the Si substrate by the *p-n* junction. Because some leakage current is inevitable in the junction, the stored charges are lost in the course of time. In DRAMs the charges are read from the capacitor before they are lost, and the full charges are again stored in the capacitor. This operation for data retention (called refreshing)

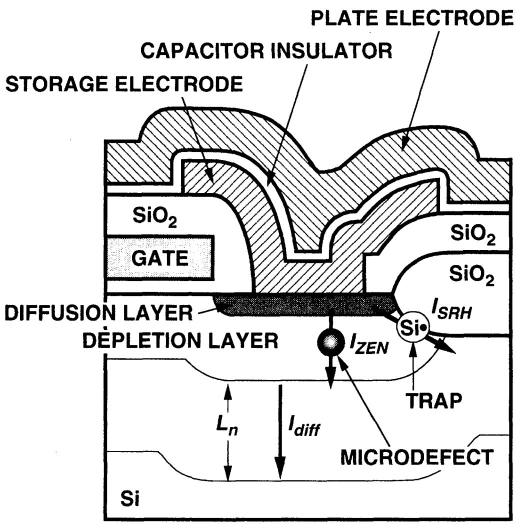

FIG. 13. Schematic cross section showing possible junction leakage paths in a DRAM cell. The cell is of a stacked type used in most DRAMs of 4 Mbits or more.

consumes much power. Accordingly, to meet the market need for low-dissipation-power DRAMs, it is important to make the refreshing interval as long as possible, namely, to reduce the junction leakage current. Figure 13 shows typical paths of the junction leakage: SRH (Shockley–Read–Hall) process, local Zener effect, and diffusion. Although only the SRH process is related to the SiO_2–Si interface (Kircher, 1975), we look over all three of these paths.

3.3.2 Shockley–Read–Hall Process A space-charge layer (called depletion layer) exists between the neutral *n*-type and *p*-type regions that form the junction. An electric field is generated by the space charge and sweeps the carriers (electrons and holes) in the depletion layer toward the neutral regions. The carriers thus swept lead to the leakage current and are generated by thermal excitation of Si valence electrons into the conduction band. This excitation has an activation energy of the Si band gap (E_g = 1.12 eV) and leads to only a negligible leakage current near room temperature. Carrier traps have energy levels in the Si band gap (see Sec. 2.3.1) and assist the thermal carrier generation by serving as stepping stones to the excitation. This process was investigated in detail by Shockley and Read (1952) and Hall (1952), and is called the SRH process. The leakage current produced by this process is expressed as

$$I_{\mathrm{SRH}} = \int_{depl} qN_t G_{\mathrm{SRH}}\, dV, \tag{14}$$

where N_t is the trap density and the integral should be carried out over the depletion region. G_{SRH} is the carrier generation rate given by

$$G_{\mathrm{SRH}} = \sigma_p \sigma_n v_{th} n_i \left[\sigma_n \exp\left(\frac{E_t - E_i}{kT}\right) + \sigma_p \exp\left(\frac{E_i - E_t}{kT}\right)\right]^{-1}, \tag{15}$$

where σ_n and σ_p are the electron and hole capture cross sections, n_i is the intrinsic carrier density, E_t is the trap energy level, and E_i is the Fermi level of nondoped Si (close to the middle of the Si band gap). According to this equation, most of the leakage current is caused by the traps close to E_i. For these traps the SRH current has almost the same temperature dependence as n_i, which has an activation energy of $E_g/2$.

The electric field in the present LSI appreciably enhances this SRH current (Hurkx *et al.*, 1992). In the absence of the field, the trapped electrons can be emitted over the entire trap depth $E_C - E_t$ only by thermal excitation. Under an electric field, they can be emitted by a combination of thermal excitation over a part of the trap depth and tunneling through the remaining potential barrier along the field. Hole emission is similar, and these processes enhance the carrier generation rate as follows:

$$G_{\mathrm{SRH}}(F) = G_{\mathrm{SRH}}\left[1 + 2\sqrt{3\pi}\,\frac{F}{F_\Gamma} \exp\left(\frac{F}{F_\Gamma}\right)^2\right], \tag{16}$$

where F is the field, $F_\Gamma = [24m^*(kT)^3]^{1/2}/q\hbar$, and m^* is the effective mass of carriers. Hurkx *et al.* (1992) experimentally obtained $m^* = 0.25\ m_0$, where m_0 is the electron mass.

Experiments using junctions of different areas and edge lengths show that there is a leakage-current component that is proportional to the edge length and has an activation energy close to $E_g/2$. This component is attributed to the SHR current generated by the traps at the interface between the Si substrate and the SiO$_2$ that is formed surrounding the devices for electric isolation. As the device size is reduced, the edge component dominates the areal component because the edge length is scaled by a constant K while the area is scaled by K^2 ($K < 1$). The contribution from the bulk traps in Si is negligibly small because their density in the current LSI is less than 10^{12} cm^{-3}.

3.3.3 Other Current Generation Processes

3.3.3.1 Local Zener Effect The local Zener effect is the mechanism of anomalous current generation in defective junctions (Ohyu and Hiraiwa, 1992). Crystalline microdefects, such as oxide and metallic precipitates, locally enhance the electric field in the depletion layer around them because the dielectric constant of a defect differs from that of the crystalline Si. Circular metallic precipitates enhance the field by a factor of as much as 3; oxides, by a factor of as much as 1.4. The enhanced electric field causes band-to-band (valence band to conduction band) tunneling of electrons, which is known as the Zener effect (Zener, 1934). The current generated is expressed as

$$I_{\text{Zener}} = AV\exp(-B/\xi F), \tag{17}$$

where V is the volume of the microdefect, F is the external field at the microdefect, and ξ is the field enhancement factor. B is a physical constant that depends on the Si band gap and tunneling effective mass and is experimentally given as 1.9×10^7 V/cm (Hurkx *et al.*, 1992). The local Zener current largely depends on the electric field at the defect and only slightly depends on temperature through the term $E_g^{2/3}$ involved in B.

3.3.3.2 Diffusion The diffusion current is fed by minority carriers (electrons in the case of n^+-p^- junction) that are generated in the bulk Si (p^- in the case of n^+-p^- junction) within a diffusion length L_n from the depletion-layer edge and flow into the depletion layer. It is given by Shockley (1949) as

$$I_{diff} = S_j q \frac{L_n}{\tau_n}\frac{n_i^2}{N_A}, \tag{18}$$

where S_j is the area of the depletion layer, N_A is the dopant concentration in the bulk Si, and τ_n is the minority-carrier lifetime in the bulk Si within L_n from the depletion-layer edge, given by $\tau_n = 1/(\sigma_n v_{th} N_t)$.

The diffusion current as a function of temperature T is mostly determined by the term $n_i^2 \propto \exp(-E_g/kT)$ of Eq. (18) because the terms L_n and τ_n change slowly with T. Accordingly, the activation energy of the diffusion current is almost equal to that of n_i^2 ($E_g = 1.12$ eV). The magnitude of this component is proportional to the area of the junction and becomes larger than that of the SRH-process current at temperatures above 100°C.

4. GENERATION AND PASSIVATION OF INTERFACE TRAPS

4.1 Physical Origins

Interface traps are produced and annihilated by various processes during LSI fabrication and device operation. The processes are classified into two groups: physical and chemical. The former is driven by energetic beams. From a quantum mechanical point of view, the energy is carried by particles that are classified according to their mass as follows; photon (no mass), electron or hole, and atom or molecule. The results depend not only on the kind of particles but also on their energy, and the results are described separately for the respective particle and energy.

4.1.1 Photon Beams

4.1.1.1 Below 8 eV Photons of this energy range excite electrons in Si and inject them into SiO$_2$ (internal photoemission, IPE), thus causing a negative oxide charging similar to that caused by electron injection (Williams, 1965). Emptying of the trapped carriers is also observed. Current–voltage measurements of the IPE electrons through the oxide (called the photo I-V method) are effective for investigating the charge distribution in the film (Berglund and Powell, 1971; DiMaria, 1976). The amount of IPE electrons is mostly limited by the local field near the electron-emitting electrode. Because the field is affected by the oxide charge, we can derive the charge distribution from the difference between the gate voltages that correspond to the same IPE current before and after the operation of interest. Letting ΔV_G^+ and ΔV_G^- be the respective voltage shifts for the electron emissions from the Si (positive V_G) and the gate (negative V_G), we can express the actual charge change ΔQ and the center $\bar{x}$ (centroid) of the charge distribution averaged by the moment by

$$\Delta Q = \frac{\epsilon_S}{d}(\Delta V_G^- - \Delta V_G^+), \tag{19}$$

and

$$\bar{x} = d\Delta V_G^+/(\Delta V_G^+ - \Delta V_G^-). \tag{20}$$

The charge change ΔQ and centroid $\bar{x}$ are related to the difference of distributions $\Delta\rho(x)$ (equal to $\rho_{\text{after}} - \rho_{\text{before}}$) by the equations

$$\Delta Q = \int_0^d \Delta\rho(x)\,dx,$$

and

$$\bar{x} = \frac{1}{\Delta Q}\int_0^d x\Delta\rho(x)\,dx.$$

4.1.1.2 From 8 eV to 470 keV Vacuum ultraviolet light (VUV, 5 eV to 1 keV), x rays (100 eV to 100 keV), and some γ rays (>10 keV) are classified here. VUV photons are produced by the plasma used for depositing films and for etching the selectively masked films to produce LSI patterns. Gamma rays are emitted from nuclear reactors and from the outer walls of artificial satellites irradiated by cosmic rays (beta rays) through the process of bremsstrahlung.

The most fundamental process involved is electron–hole pair generation in the bulk SiO$_2$. The holes thus generated are trapped at oxide defects (presumably oxygen vacancies), producing paramagnetic defects associated with positive charges (Lenahan and Dressendorf, 1984):

$$\mathrm{O_3{\equiv}Si{-}Si{\equiv}O_3} + h^+ \rightarrow \mathrm{O_3{\equiv}Si\cdot} + {}^{+}\mathrm{Si{\equiv}O_3}. \tag{21}$$

These defects are called E'_γ centers (g = 2.0004) and are shown schematically in Fig. 14(b). When they result from a hard x ray that penetrates the entire film, both the E'_γ centers and positive charges are located close to the SiO$_2$–Si interface (Lenahan and Dressendorfer, 1984). This supports the hypothesis that the hole-trapping defects are O vacancies whose density is enhanced by strain (Ohmameuda *et al.*, 1991). A soft x ray, on the other hand, produces the paramagnetic defects only in the penetrated surface region of the film (Yokogawa *et al.*, 1990). In this case the defects show up only after an extended x-ray irradiation, suggesting that there are fewer O vacancies in the region far from the SiO$_2$–Si interface. The

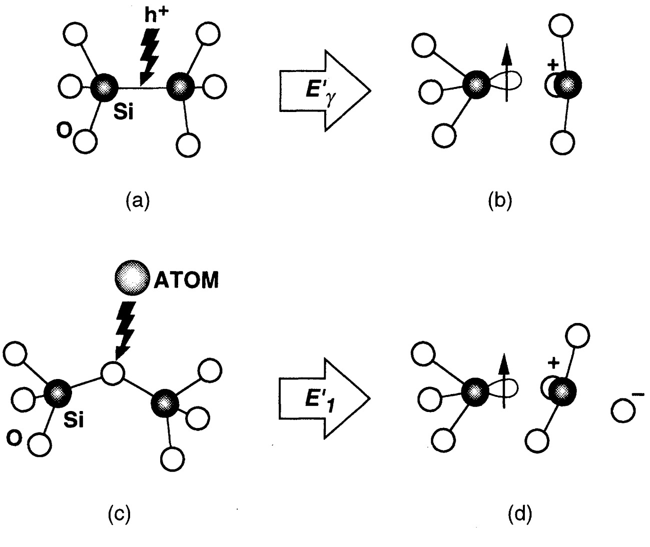

FIG. 14. Schematic illustrations of (a), (c) the precursors and (b), (d) the ESR-active states of the E' centers in SiO$_2$. (b) corresponds to the E'_γ center, and (d) to the E'_1 center. (After Warren *et al.*, 1992.)

positive charges caused the negative shift of the high-frequency *CV* curves shown in Fig. 7. The positive charging increases almost linearly with the photon dosage, and more charges were produced for the positive gate bias. This positive bias effect is ascribed to the enhanced drifting of holes to the interface.

Interface-trap generation is closely connected with E'_γ-center formation. Atomic hydrogen is released following electron trapping at water-related traps (O$_3$Si–H in "dry" oxide and O$_3$Si–OH in wet oxide) by the reaction

$$O_3{\equiv}Si{-}H + e^- \rightarrow O_3{\equiv}Si{:}^- + H^0$$

or

$$O_3{\equiv}Si{-}OH + e^- \rightarrow O_3{\equiv}Si{-}O{:}^- + H^0. \qquad (22)$$

We should notice that a nominally dry oxide contains an appreciable amount of water (from 10^{17} to 10^{20} H atoms/cm^3 in the form of ≡Si–OH or ≡Si–H) as reported by Revesz (1979). The atomic hydrogen H^0 dimerizes by the reaction

$$H^0 + H^0 \rightarrow H_2 + 4.5\ \text{eV} \qquad (23)$$

and forms hydrogen molecules that diffuse to the interface. The H_2 molecule reacts with the E'_γ centers, presumably releasing H^0 again:

$$O_3{\equiv}Si\cdot + H_2 \rightarrow O_3{\equiv}Si{-}H + H^0. \qquad (24)$$

Reactions (22) to (24) lead to hydrogen migration from the bulk SiO$_2$ to the interface, which was demonstrated by Gale *et al.* (1983) in the case of avalanche injection of electrons.

The atomic hydrogen finally produces the interface traps. Considering that more interface traps are generated under a positive bias, an electronic involvement by the reaction

$$Si_3{\equiv}Si{-}H + H^0 + e^- \rightarrow Si_3{\equiv}Si{:}^- + H_2 \qquad (25)$$

is likely in the interface-trap formation (Griscom, 1985). Some hole involvement other than reaction (21), however, cannot be ruled out (see Sec. 4.1.2.1).

In the case of soft (nonpenetrating) x rays, on the other hand, the atomic hydrogen seems to be supplied to the interface both by the direct reaction (22) throughout the oxide and by reactions (21) to (24) in the surface region penetrated by the x rays. Winokur and Sokoloski (1976) reported a similar interface-trap buildup for both hard and soft x rays.

The number of interface traps thus produced is comparable to the number of positive charges. As shown in Fig. 8, the trap distribution has a peak 0.1–0.3 eV above the midgap.

For photons of energies close to the resonance absorption, the diffusion of excitons also seems responsible for the interface-trap generation (Weinberg and Rubloff, 1978). At higher energies up to 50 keV, the photoelectric effect is predominant, and electrons are scattered away from the bound atoms, completely absorbing the photons. For photons with energies larger than 50 keV, though, Compton scattering is the predominant effect. In this process, photons only lose part of their energy in the first scattering event and continuing scattering until all of it is absorbed. Thus the resulting phenomenon is essentially the same as in the energy range below 50 keV. The electron scattering is followed by emission of Auger electrons or characteristic x rays.

4.1.1.3 Above 470 keV In this region, at least a part of the electrons scattered by the Compton effect subsequently displace the atoms in SiO$_2$. For energies greater than 1 MeV, electron–positron pair generation takes place, although its frequency is much less than that of the Compton effect. Before the phenomena resulting from atomic displacement or pair generation become dominant, interface degradation caused by electron–hole pair generation occurs. Accordingly, from the practical point of view, the device degradation mechanism is the same as that with photons of lower energies.

4.1.2 Electron and Hole Flux

4.1.2.1 Below 10 eV The second group of physical origins that affect the interface is electrons and holes. The drift electrons of MOS transistors (or holes in *p*-MOS transistors) are confined outside of the SiO$_2$ because of the SiO$_2$–Si potential barrier: 3.1 eV (3.8 eV for holes). Energetic electrons and holes (called hot carriers) are produced by impact ionization near the drain region of the Si substrate, where a large electric field exists (see Sec. 3.1). These carriers pass the barrier, are trapped in the SiO$_2$, and generate interface traps. Additionally, in flash memories where information is stored as negative charges in an electrically isolated gate, electrons are forced to flow through the oxide by a high field.

To study this trapping phenomenon, Nicollian *et al.* (1969) developed the avalanche-injection method. They repeatedly applied a large-amplitude ac voltage to the gate of MOS capacitors. This voltage capacitively induces a deep depletion layer accompanied by a high field in the Si substrate, causing impact ionization and subsequent injection (driven by the dc component of the gate voltage) of hot carriers into the oxide. This method has advantages over other techniques, such as high-field stressing (HFS): Only electrons or holes are injected into the film, and the field applied to the film may be suppressed below 3 MV/cm.

The bulk trapping and interface-trap generation depends on the kind of carriers (holes or electrons) and on the way they are introduced. As shown in Fig. 15, avalanche-injected holes produce numbers of positive charges that increase monotonically with increase in the number of holes injected into the SiO$_2$. The efficiency of positive-charge generation is independent of the magnitude of the applied field (Schwerin *et al.*, 1990). Postoxidation annealing (POA) enhances the positive charging, and this enhancement is ascribed to an increase in the number of O vacancies. The number of interface traps is also increased by hole injection, and there is a linear relationship between the number of traps generated and the number of positive charges. Different values of the ratio of interface traps to holes are reported: 1 (Al gate) and 0.4 (poly-Si gate) by Thanh *et al.* (1986) and 0.05 (poly-Si gate) by Schwerin *et al.* (1990). The interface traps are mostly located from 0.1 to 0.3 eV above the midgap in the Si band gap. We cannot understand this interface-trap generation in the framework of the mechanism described in Sec. 4.1.1.2 because the negligible electron fluence involved means that atomic or molecular hydrogen is not supplied to the interface. The two-step model by Lai (1983) seems more likely to apply here. A strained Si–O–Si bond first captures a hole, which breaks the Si–O bond, and the oxygen atom thus freed moves to a position of lower strain. The trapped hole eventually captures an electron, restoring the charge neutrality, but the oxygen atom does not return to its original position. This leads to a dangling bond or a weak bond, resulting in an interface trap as depicted by Sakurai and Sugano (1981). This process might be at least partly responsible for the interface traps generated by photon irradiation.

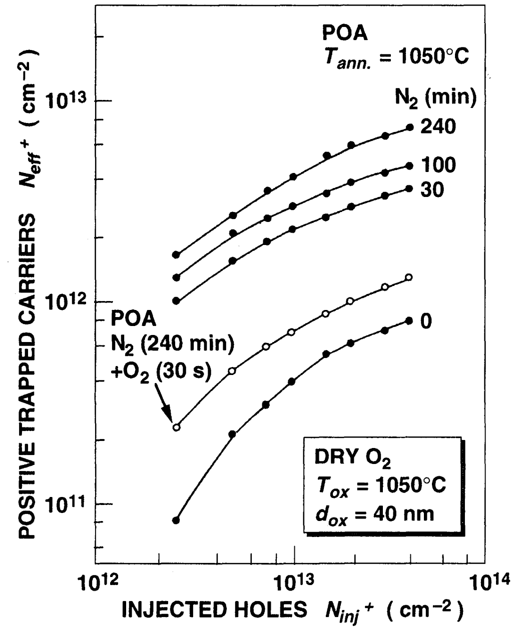

FIG. 15. Density of positive trapped carriers as a function of avalanche-injected hole fluence. The samples were treated by POA in N$_2$ for different periods. Reprinted from Thanh *et al.* (1986), with kind permission from Elsevier Science Ltd.

In the case of avalanche injection of electrons, negative charges first build up and then decrease with extended injection (Fig. 16). This is the so-called turnaround effect. The fluence of electron injection necessary to produce an appreciable charging is about three orders of magnitude larger than that necessary when holes are injected. The efficiency of the electron trapping is appreciably enhanced by oxide fields above 4 MV/cm (Heyns *et al.*, 1989). This is in contrast to

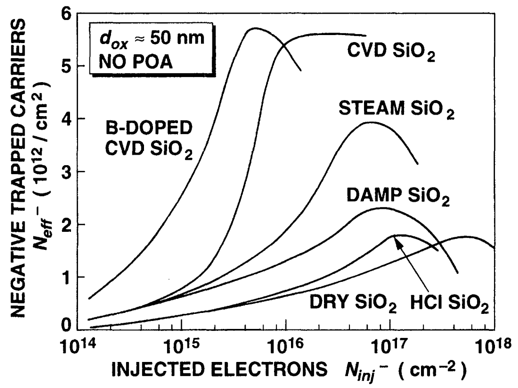

FIG. 16. Density of negative trapped carriers as a function of avalanche-injected electron fluence. Oxidation was performed at 1000 °C in the ambient of dry O$_2$, dry O$_2$ + 3% HCl (HCl SiO$_2$), dry O$_2$ + 3% H$_2$O (DAMP SiO$_2$), or steam. CVD oxide films were deposited using SiH$_4$ + O$_2$ at 800 °C. B-doped films were doped with 5 wt % B$_2$O$_3$. (After Gdula, 1976.)

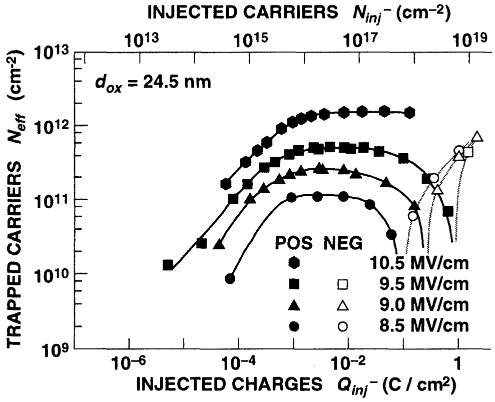

FIG. 17. Density of trapped carriers as a function of injected electron fluence. The electrons were introduced into SiO$_2$ driven by a high field. Closed symbols are for positive net charge, and open symbols are for negative net charge. (After DiMaria *et al.*, 1993.)

what is observed in hole injection where the trapping efficiency was not enhanced (Schwerin *et al.*, 1990). A field of 4 MV/cm in the oxide produces hot electrons with the average energy larger than the 2.3-eV threshold for hydrogen release (DiMaria and Stasiak, 1989). The negative charging is more remarkable in wet oxides. From this we can see that the water-related species, such as Si–H and Si–OH, are possible electron traps. The decreased negative charging after an extended injection is due to the buildup of positive charges that are presumably injected from the electrode. Here again the relationship between the generated interface traps and negative charges is almost linear. In poly-Si gates the ratios of traps to charges are from 0.2 to 1 (Thanh *et al.*, 1986; Heyns *et al.*, 1989). The interface-trap distribution reaches a peak from 0.1 to 0.3 eV above the Si midgap for fields above 4 MV/cm, below which a broad distribution in the Si band gap is reported.

In the case of high-field stressing, most of the current that flows through the SiO$_2$ is fed by electrons, regardless of the biasing polarity. Nonetheless, the net charge is initially positive and turns negative only after an extended stressing (Fig. 17). This is due to the trapping of holes that are generated by band-to-band ionization in the oxide. The electron fluence where the positive charging begins to saturate agrees well with that of the appreciable negative charging shown in Fig. 16. This coincidence indicates that the saturation of positive charging is due to charge compensation by the buildup of negative charges. Higher fields increase the fraction of holes in the current and enhance the efficiency of the positive charging. DiMaria *et al.* (1993) reported that the number of interface traps is about 1/10 of that of the positive charges in the early stage of trap generation.

4.1.2.2 Above 10 eV Plasma etching for LSI pattern formation is a source of electrons with energies from 10 eV to a few hundred electron volts. The thick oxide for device isolation is exposed to this irradiation. Electron-beam (EB) lithography is another source of high-energy electron irradiation (20–30 keV). Because the oxide is covered by an EB-sensitive polymer film during the irradiation, the interface degradation in this process is mostly caused by the low-energy secondary electrons produced by the incident electrons.

Electrons with energies from 10 eV to a few hundred electron volts excite electron–hole (e-h) pairs that generate interface traps and bulk positive charges in the same way that photons do, and the interface-trap distribution in the Si band gap again peaks from 0.1 to 0.3 eV above the Si midgap. Yunogami and Mizutani (1993) reported the following processes specific to the electronic irradiation: plasmons first produced at the surface or in the bulk of SiO_2 decay, resulting in e-h pair generation. All the e-h pairs are not produced by plasmons, but oscillations specific to the plasmon energies are clearly observed when the number of generated interface traps or oxide charges is plotted as a function of the incident electron energy.

Electrons with energies greater than a few hundred kiloelectron volts displace the constituent atoms of the oxide, resulting in phenomena similar to the atomic displacement in Sec. 4.1.3.

4.1.3 Atomic and Molecular Beams Atoms introduced by ion implantation displace (elastic scattering) and ionize (inelastic scattering) the atoms in SiO_2. The elastic scattering produces paramagnetic (ESR active) defects called E'_1 centers. These defects are the products of O recoil from the network $O_3{\equiv}Si{-}O{-}Si{\equiv}O_3$, and consist of ESR-active $O_3{\equiv}Si\cdot$ and positively charged $^+Si{\equiv}O_3$ [Fig. 14(d)]. The positive charges that are observed in the implanted SiO_2 even after a thermal annealing might be due to these E'_1 centers. The inelastic scattering, on the other hand, produces electron–hole pairs that cause charging and trap generation in a way similar to the way photon, electron, and hole injection does.

The ion implantation also produces interface traps, and the number of traps remaining after 900 °C annealing is determined solely by the energy dissipated in the nuclear stopping events (Fig. 18). This figure clearly shows that the interface traps are produced by the displacement of host (SiO_2) atoms by nuclear collisions with the impinging and recoiled atoms. We also found that the implanted oxygen atoms restore the broken bonds at the interface to the state before the implantation (Hiraiwa, to be published). But because this repairing process was not observed when Ar, Ne, and C ions were implanted, it seems to be a chemical process.

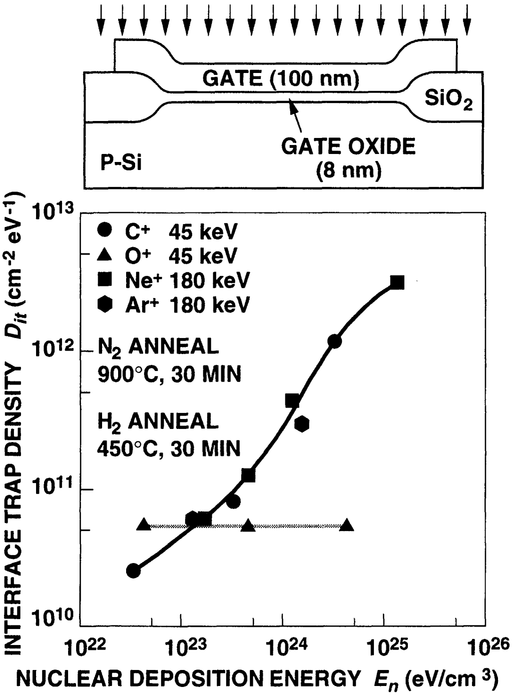

FIG. 18. Midgap interface-trap density of ion-implanted SiO_2 after N_2 annealing at 900 °C. The samples received the ion implantation after the formation of phosphorus-doped Si film as a gate. The ion dose is normalized by the total energy that the ions dissipated through a nuclear stopping process. (After Hiraiwa, to be published.)

4.2 Chemical Origins

4.2.1 Thermal Annealing The results of thermal annealing are somewhat controversial, probably because the purity of the annealing ambient and of the sample itself is not always controlled. The traditional furnace usually contains not a small amount of moisture (as impurity) from the environment, and the source-gas purity is easily degraded if the piping is not prepared and operated carefully. And because even a trace amount of moisture (10 ppm) passivates the interface traps, we should take special care to ensure the purity of the ambient when we study annealing. Figure 19 shows that in a high-purity nitrogen annealing, the number of P_b centers (some of the interface traps, see Sec. 2.3.2.4) increases during annealing times shorter than 5 min. This increase is attributed to the desorption of hydrogen atoms from Si–H bonds that are easily broken above 500 °C by the reaction

$$Si{-}H \rightarrow Si\cdot + H^0. \qquad (26)$$

Brower *et al.* (1990) obtained the rate equation as

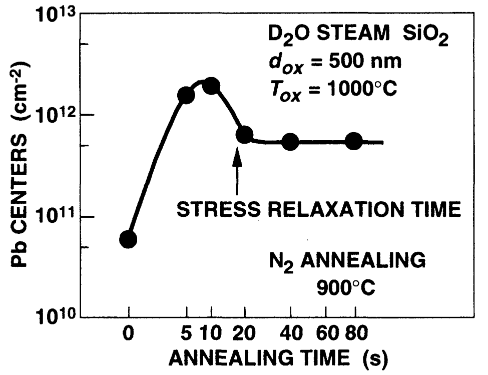

FIG. 19. Density of P_b centers as a function of annealing time. After Ohji *et al.* (1990); © 1990 IEEE.

$$\frac{d[\text{Si–H}]}{dt} = -k_d[\text{Si–H}], \tag{27}$$

where

$$k_d = (1.2 \times 10^{12}\ \text{s}^{-1}) \exp\left(-\frac{2.56\ \text{eV}}{kT}\right).$$

The number of P_b centers decreases during an extended annealing, showing thermal passivation in the absence of hydrogen-related species. As a result of this thermal passivation effect, a plot of the number of P_b centers as a function of annealing temperature has a peak (near 800 °C in Fig. 20). Montillo and Balk (1971) reported the direct thermal passivation of the interface traps, but some involvement of hydrogen (or deuterium) supplied from the wet or nominally dry oxide itself cannot be ruled out.

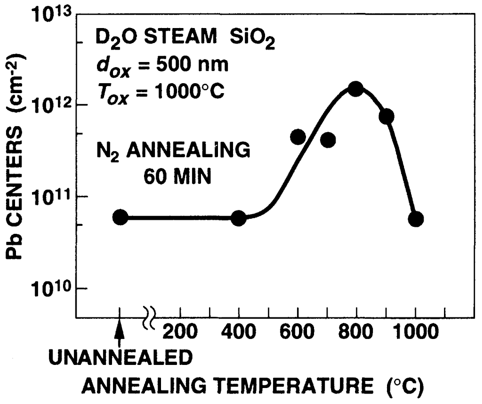

FIG. 20. Density of P_b centers as a function of annealing temperature. Reproduced from Ohji *et al.* (1990); © 1990 IEEE.

Not only hydrogen desorption but also the thermal decomposition of SiO$_2$ seems to contribute to the generation of interface traps. This decomposition is enhanced by the catalytic action of metallic impurities, such as Cu, W, and Pt. In extreme cases, portions of contaminated oxide films disappear during high-temperature ultrahigh-vacuum annealing (Tromp *et al.*, 1985). The decomposition proceeds at the SiO$_2$–Si interface by the reaction Si + SiO$_2$ → 2 SiO ↑. The feasibility of interface-trap generation by this process is supported by the result of an experiment in which one Cu atom present during the oxidation produced one interface trap on the average (Hiraiwa and Itoga, 1994).

4.2.2 Hydrogen Annealing In LSI processes the interface traps are electrically passivated by annealing in the final step of device fabrication. This annealing must be done below 450 °C, however, because the Al wirings are already formed, and we cannot expect thermal passivation to occur at these low temperatures (Sec. 4.2.1). Thus, we usually terminate the interfacial dangling bonds by exposing the samples to a hydrogen-containing ambient.

The results of hydrogen annealing depend on whether a gate electrode is formed or not. Figure 21 shows the case without the gate, which is a kind of postoxidation annealing. According to Brower and Myers (1990) again, the hydrogen molecules passivate the interface traps by the reaction

$$\text{Si}\cdot + \text{H}_2 \rightarrow \text{Si–H} + \text{H}^0, \tag{28}$$

where

$$\frac{d[\text{Si}\cdot]}{dt} = -k_f[\text{Si}\cdot][\text{H}_2],$$

$$k_f = (1.94 \times 10^{-6}\,\text{cm}^3\,\text{s}^{-1}) \exp\left(-\frac{1.66\,\text{eV}}{kT}\right). \tag{29}$$

We should take care that reactions (26) and (28) are reversible. In the case of H$_2$ annealing, H^0 is so scarce that reverse reactions are not likely. In ample atomic hydrogen, on the other hand, reverse reactions take place. Thus, we observe both passivation and de-

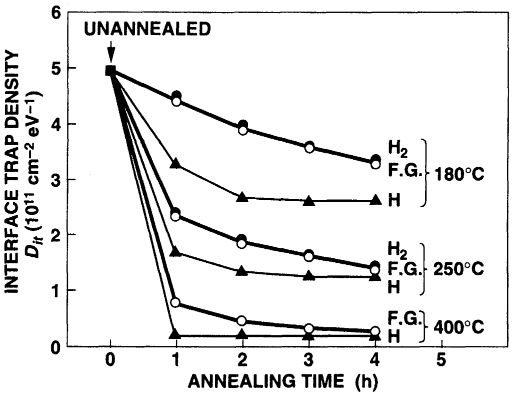

FIG. 21. Interface-trap density at midgap as a function of annealing time at various temperatures in various ambients: H_2, molecular hydrogen; F. G., forming gas (10% H_2 in N_2); and H, atomic hydrogen produced by a high-voltage discharge. (After Thanh and Balk, 1988.)

passivation of the interface traps depending on the temperature and initially passivated state (Thanh and Balk, 1988). The result is more interface traps at lower temperatures, which in turn leads to more interface degradation occurring as a result of radiation or carrier injection (where the atomic hydrogen is released from the bulk SiO$_2$, as described in Sec. 4.1) because the interface is kept at low temperatures, except during annealings.

The results of postmetallization annealing differ from those of POA because of the presence of the gate. For both polycrystalline Si (poly-Si) and metals such as Al, even annealing in an ambient without any hydrogen-containing molecules reduces the number of interface traps. This reduction is ascribed to hydrogen atoms bound to crystalline defects (such as grain boundaries and dislocations) in the poly-Si gate. In an Al gate, the number of interface traps is reduced much more than in a poly-Si gate, possibly because the Al electrode reduces Si–OH that was incorporated into SiO$_2$ during oxidation or adsorbed onto the film before Al metallization (Reed and Plummer, 1988).

Fluorine atoms also terminate the interface traps. They are much larger than hydrogen atoms and require a higher temperature to reach the SiO$_2$–Si interface. Because the energy of the Si–F bond is greater than that of the Si–H bond, the F passivation is more stable during thermal treatments. Accordingly, the introduction of F atoms before oxidation is also effective in passivating the interface traps. The F passivation also makes the interface more resistant to the effects of radiation and hot-carrier injection (Nishioka *et al.*, 1989). We should take care, however, that we do not increase the number of traps and decrease the dielectric constant of the oxide by introducing too many F atoms.

5. BREAKDOWN OF SiO$_2$

Destructive breakdown of the SiO$_2$ film occurs when a large electric field is applied to MOS capacitors. The breakdown field differs among capacitors (Yamabe and Taniguchi, 1985). When the capacitor area is smaller, we observe more breakdown events at the maximum field (greater than 10 MV/cm). From these results many researchers consider that SiO$_2$ has its own (intrinsic) dielectric strength and that the breakdown is triggered by some structural defects in the SiO$_2$.

Even if the SiO$_2$ field is smaller, the breakdown occurs when the field is applied for an extended period. In the estimation of breakdown reliability, we usually accelerate the breakdown by increasing the field far beyond the daily-use level because we cannot wait for 10 years (certified lifetime of the devices). To predict the daily-use reliability from the accelerated testing, an understanding of the breakdown mechanism is indispensable. There is little doubt about the view that the intrinsic breakdown is caused by the current that is induced by the field and flows through the SiO$_2$. Although most of the current is carried by electrons, it seems that holes are deeply involved in the breakdown event. This is because the total amount of holes that pass through the SiO$_2$ until breakdown occurs is constant independent of the applied field, whereas the electron fluence varies (Chen *et al.*, 1986). The holes are considered to be generated in the oxide by impact ionization or to be injected into the oxide from the anode by tunneling with possible assistance of excitation by electrons (incident from SiO$_2$). DiMaria *et al.* (1993), on the other hand, reported that the breakdown is correlated with the bulk-SiO$_2$ and SiO$_2$–Si-interface trapping. In any case, however, the fundamental processes of the intrinsic breakdown are still beyond our understanding.

From a practical point of view, the intrinsic dielectric strength is enough for MOS transistor operation because the SiO$_2$ field is no more than 4 MV/cm. Thus, most of the electronics engineers are concerned about how to reduce the oxide defects. There are many studies that have investigated oxide-defect formation in relation to MOS capacitor processing conditions. Through these studies Si-crystal defects (such as oxide precipitates in Czochralski-grown Si) are found to cause the breakdown (Itsumi *et al.*, 1992). It is also reported that metallic contamination degrades the SiO$_2$ by lowering the potential barrier of SiO$_2$ (Takiyama *et al.*, 1994). However, it is still unclear what properties make the oxide defects different from the intrinsic region.

6. SUMMARY

The transition from bulk Si to SiO$_2$ seems to be sharp, extending for only one or two atomic layers, but it is not as flat as it appears in TEM images: the actual morphology of the interface presumably has undulations corresponding to more than four atomic layers. Defects at the interface act as carrier traps, some of which are paramagnetic and are detected as P_b centers by ESR measurement. The P_b center is ascribed to a dangling bond of an interfacial Si atom that is backbonded to three other Si atoms. These centers, however, form only some of the interface traps, and the rest of the traps have not been identified.

The properties of the SiO$_2$–Si interface are crucial to MOS transistor performance that is characterized by the subthreshold swing and the drain current. The increase in subthreshold swing is caused by the interface traps and reduces the switching speed of the MOS transistor itself. The drain current, on the other hand, determines the switching time of an LSI circuit after the onset of the "on" state of MOS transistors. The interfacial undulation reduces the drain current because surface-roughness scattering reduces the carrier mobility. The interface traps, as well as bulk-oxide traps, reduce the drain current mostly by reducing the number of surface carriers. On the other hand, the interface traps generate a current by the SRH process in the depleted region of junctions, and this leads to an excessive junction leakage current that causes data-retention errors of DRAMs.

The interface traps are generated by ionizing radiation and carrier injection, and thermal annealing not only passivates but also generates the traps. Many researchers consider the generation kinetics in the following way, although other explanations cannot be ruled out. Some of the interface traps are generated by the release of terminating hydrogen from Si–H bonds. This proceeds either thermally (Si–H → Si· + H^0) or by a reaction with atomic hydrogen ($Si_3{\equiv}Si{-}H + H^0 \rightarrow Si_3{\equiv}Si\cdot + H_2$). One of the three processes that produce the atomic hydrogen is the electron trapping at water-related traps, such as Si–H and Si–OH bonds. Another is the thermal decomposition of these bonds. The other is the cracking of hydrogen molecules at oxide defects known as E' centers. The hydrogen molecules are supplied by the dimerization of atomic hydrogen or by diffusion from the ambient. The E' centers are formed by hole trapping at oxygen vacancies that are concentrated where the SiO$_2$ network is strained, such as in the interfacial region. Interface traps might also be generated by a two-step process in which a hole is first trapped at the strained Si–O–Si bond and then captures an electron, thus leaving a weak or dangling bond. The electrons and holes thus involved in the trap generation are injected from the Si substrate or electrode. They are also generated in pairs by ionizing radiation, or band-to-band ionization under a high field. The passivation of the interface traps proceeds by reactions that are the reverse of those generating the traps, and these passivation reactions are facilitated by high temperature and by ambient atmospheres containing hydrogen molecules. The annealing of the traps at high temperatures is also observed even when hydrogen is not present.

Destructive breakdown of SiO$_2$ occurs when a high field is applied. Even a lower field causes breakdown after an extended application. The breakdown is caused by the current that is induced in the SiO$_2$ by the field. Although most of the current is carried by electrons, it seems that holes are deeply involved in the breakdown. The correlation of the breakdown with bulk-SiO$_2$ and SiO$_2$–Si-interface trapping is also reported. Structural defects existing in the SiO$_2$ trigger the breakdown and reduce the lifetime. The physical reality of the defects, however, is still unclear.

Many problems concerning the SiO$_2$–Si interface remain unsolved despite many studies. The increasing number of damage processes, and the growing need for a better interface (such as is required for low-power DRAMs), will spur further studies of the SiO$_2$–Si interface.

ACKNOWLEDGMENTS

The author acknowledges valuable suggestions on MOS transistor operation from Shin'ichiro Kimura and Ryo Nagai. The section on junction leakage current owes much to discussions with Kiyonori Ohyu, and many thanks are owed to Kikuo Watanabe, Masataka Kato, Takashi Yunogami, Yuzuru Ohji, and Ken'etsu Yokogawa for their discussions of the oxide traps.

GLOSSARY

Avalanche Injection: One of the methods to introduce carriers from the Si substrate into the SiO$_2$ film. The carriers are produced by impact ionization in the Si depletion layer, induced by a large-amplitude ac gate voltage. They are subsequently driven into the oxide by the dc component of the gate voltage. The field across the oxide can be kept below 3 MV/cm.

***CV* Method:** A method to quantify bulk oxide charges and interface traps by measuring the capacitance of MOS capacitors as a function of the gate voltage.

ESR: Electron-spin resonance measurement. From the resonance absorption line, we can get information about the structure of dangling bonds.

Interface Trap: Carrier trap that exists at the SiO$_2$–Si interface and exchanges a charge with the Si surface. Some are ascribed to the dangling bond of an interfacial Si back-bonded to three other Si atoms.

LSI: Large-scale integration or large-scale–integrated circuit. Currently more than 30 million components can be integrated in a single chip.

MOS: A structure consisting of stacked metal, oxide, and semiconductor layers (substrate). In the strict sense of the word, most present devices do not have the MOS structure, but they do have the SOS (semiconductor-oxide-semiconductor) structure because they have polycrystalline silicon gates heavily doped—with P in *n*-channel devices or with B in some *p*-channel devices.

SRH Process: The process that restores the carrier density to its equilibrium value, assisted by traps that are energetically located in the Si band gap. This process generates carriers and causes an excess leakage current in a reversely biased junction. Named after Shockley, Read, and Hall, who formulated the process.

Works Cited

Ando, T. (1977), *J. Phys. Soc. Jpn.* **43,** 1616–1626.

Bardeen, J., Brattain, W. H. (1948), *Phys. Rev.* **74,** 230.

Berglund, C. N. (1966), *IEEE Trans. Electron Devices* **ED-13,** 701–705.

Berglund, C. N., Powell, R. J. (1971), *J. Appl. Phys.* **42,** 573–579.

Brower, K. L., Myers, S. M. (1990), *Appl. Phys. Lett.* **55,** 162–164.

Brugler, J. S., Jespers, P. G. A. (1969), *IEEE Trans. Electron Devices* **ED-16,** 297–302.

Cartier, E., Stathis, J. H., Buchanan, D. A. (1993), *Appl. Phys. Lett.* **63,** 1510–1512.

Castagné, R., Vapaille, A. (1971), *Surface Sci.* **28,** 157–193.

Chen, I.C., Holland, S., Young, K. K., Chang, C., Hu, C. (1986), *Appl. Phys. Lett.* **49,** 669–671.

DiMaria, D. J. (1976), *J. Appl. Phys.* **47,** 4073–4077.

DiMaria, D. J., Stasiak, J. W. (1989), *J. Appl. Phys.* **65,** 2342–2356.

DiMaria, D. J., Cartier, E., Arnold, D. (1993), *J. Appl. Phys.* **73,** 3367–3384.

Ezawa, H., Kawaji, S., Nakamura, K. (1971), *Jpn. J. Appl. Phys.* **13,** 126–155.

Gale, R., Feigl, J., Magnee, C. W., Young, D. R. (1983), *J. Appl. Phys.* **54,** 6938–6942.

Gdula, R. A. (1976), *J. Electrochem. Soc.* **123,** 42–47.

Goodnick, S. M., Ferry, D. K., Wilmsen, C. W., Liliental, Z., Fathy, D., Krivanek, O. L. (1985), *Phys. Rev. B* **32,** 8171–8186.

Griscom, D. L. (1985), *J. Appl. Phys.* **58,** 2524–2533.

Grunthaner, F. J., Grunthaner, P. J. (1986), *Materials Science Reports 1,* Amsterdam: North Holland, pp. 65–160.

Haddara, H., Christoloveanu, S. (1986), *Solid-State Electron.* **29,** 767–772.

Hall, R. N. (1952), *Phys. Rev.* **87,** 387.

Heyns, M. M., Krishna Rao, D., De Keersmaecker, R. F. (1989), *Appl. Surf. Sci.* **39,** 327–338.

Himpsel, F. J., McFeely, F. R., Taleb-Ibrahimi, A., Yarmoff, J. A., Hollinger, G. (1988), *Phys. Rev. B* **38,** 6084–6096.

Hiraiwa, A., Itoga, T. (1994), *IEEE Trans. Semicond. Manufact.* **7,** 60–67.

Hurkx, G. A. M., Graaff, H. C., Kloosterman, W. J., Knuvers, M. P. G. (1992), *IEEE Trans. Electron Devices* **ED-39,** 2090–2098.

Itsumi, M., Nakajima, O., Shiono, N. (1992), *J. Appl. Phys.* **72,** 2185–2191.

Iunovich, A. E. (1958), *Sov. Phys. Tech. Phys.* **3,** 646–650.

Kahng, D., Atalla, M. M. (1960), *IRE-IEEE Solid-State Device Research Conference, Carnegie Institute of Technology, Pittsburgh, PA,* New York: IEEE.

Kawaji, S. (1969), *J. Phys. Soc. Jpn.* **27,** 908–910.

Kircher, C. J. (1975), *J. Appl. Phys.* **46,** 2167–2173.

Lai, S. K. (1983), *J. Appl. Phys.* **54,** 2540–2546.

Lenahan, P. M., Brower, K. L., Dressendorfer, P. V., Johnson, W. C. (1981), *IEEE Trans. Nucl. Sci.* **NS-28,** 4105–4106.

Lenahan, P. M., Dressendorfer, P. V. (1984), *J. Appl. Phys.* **55,** 3495–3499.

Lilienfeld, J. E. (1926), U.S. Patent 1,745,175 (filed in 1926), 1,877,140 (filed in 1928), 1,911,018 (filed in 1928).

Ma, T. P., Dressendorfer, P. V. (Eds.) (1989), *Ionization Radiation Effects in MOS Devices and Circuits,* New York: Wiley.

Matsumoto, Y., Uemura, Y. (1974), *Jpn. J. Appl. Phys.* **Suppl. 2,** Pt. 2, 367–370.

Montillo, F., Balk, P. (1971), *J. Electrochem. Soc.* **118,** 1464–1468.

Nagayama, H., Honda, H., Kawahara, H. (1988), *J. Electrochem. Soc.* **135,** 2013–2016.

Nakashima, S., Katayama, T., Miyamura, Y., Matsuzaki, A., Imai, M., Izumi, K., Ohwada, N. (1994), "Thickness Increment of Buried Oxide in a SIMOX Wafer by High-Temperature Oxidation," *Proceedings of the 1994 IEEE International SOI Conference, Nantucket Island, Massachusetts,* New York: IEEE, pp. 71–72.

Nicollian, E. H., Brews, J. R. (1982), *MOS (Metal Oxide Semiconductor) Physics and Technology,* New York: Wiley.

Nicollian, E. H., Goetzberger, A., Berglund, C. N. (1969), *Appl. Phys. Lett.* **15,** 174–177.

Ning, T. H., Osburn, C. M., Yu, H. N. (1975), *Appl. Phys. Lett.* **26,** 248–250.

Nishioka, Y., Ohyu, K., Ohji, Y., Natsuaki, N., Mukai, K., Ma, T. P. (1989), *J. Appl. Phys.* **66,** 3909–3912.

Niwano, M., Katakura, H., Takeda, Y., Takakuwa, Y., Miyamoto, N., Hiraiwa, A., Yagi, K. (1991), *J. Vac. Sci. Technol.* **A9,** 195–200.

Ohdomari, I., Mihara, T., Kai, K. (1986), *J. Appl. Phys.* **60,** 3900–3904.

Ohji, Y., Nishioka, Y., Yokogawa, K., Mukai, K., Qiu, Q., Arai, E., Sugano, T. (1990), "Effects of Minute Impurities (H, OH, F) on SiO$_2$/Si Interface as Investigated by Nuclear Resonant Reaction and Electron Spin Resonance," *IEEE Trans. Electron Devices* **ED-37,** 1635–1642.

Ohmameuda, T., Miki, H., Asada, K., Sugano, T., Ohji, Y. (1991), *Jpn. J. Appl. Phys.* **30,** L1993–L1995.

Ohyu, K., Hiraiwa, A. (1992), *Ext. Abst. 1992 Int. Conf. Solid State Devices and Materials,* Tsukuba: The Japan Society of Applied Physics, pp. 73–75.

Poindexter, E. H., Gerardi, G. J., Rueckel, M.-E., Caplan, P. J., Johnson, N. M., Biegelsen, D. K. (1984), *J. Appl. Phys.* **56,** 2844–2849.

Reed, M. L., Plummer, J. D. (1988), *J. Appl. Phys.* **63,** 5776–5793.

Revesz, A. G. (1979), *J. Electrochem. Soc.* **126,** 122–130.

Rochet, F., Froment, M., D'Anterroches, C., Roulet, H., Dufour, G. (1989), *Philos. Mag. B,* **59,** 339–363.

Sah, C. T., Ning, T. H., Tschopp, L. L. (1972), *Surf. Sci.* **32,** 561–575.

Sakurai, T., Sugano, T. (1981), *J. Appl. Phys.* **52,** 2889–2896.

Schwerin, A. V., Heyns, M. M., Weber, W. (1990), *J. Appl. Phys.* **67,** 7595–7601.

Shockley, W. (1949), *Bell Syst. Tech. J.* **28,** 435–489.

Shockley, W., Read, W. T. (1952), *Phys. Rev.* **87,** 835–842.

Stern, F., Howard, W. E. (1967), *Phys. Rev.* **163,** 816–835.

Sze, S. M. (1981), *Physics of Semiconductor Devices,* 2nd ed., New York: Wiley, Chaps. 1, 2, 7, and 8.

Takagi, S., Iwase, M., Toriumi, A. (1990), *Ext. Abst. 22nd Conf. Solid State Devices and Materials,* Sendai: The Japan Society of Applied Physics, pp. 275–278.

Takahashi, I., Shimura, T., Harada, J. (1993), *J. Phys.: Condens. Matter* **5,** 6525–6536.

Takiyama, M., Ohtsuka, S., Sakon, T., Tachimori, M. (1994), *IEICE Trans. Electron.* **E77-C,** 464–472.

Thanh, L. Do, Aslam, M., Balk, P. (1986), "Defect Structure and Generation of Interface States in MOS Structure," *Solid-State Electron.* **29,** 829–840.

Thanh, L. Do, Balk, P. (1988), *J. Electrochem. Soc.* **135,** 1797–1801.

Thornber, K. K. (1980), *J. Appl. Phys.* **51,** 2127–2136.

Tromp, R., Rubloff, G. W., Balk, P., LeGoues, F. K. (1985), *Phys. Rev. Lett.* **55,** 2332–2335.

Warren, W. L., Poindexter, E. H., Offenberg, M., Muller-Warmuth, W. (1992), *J. Electrochem. Soc.* **139,** 872–880.

Watanabe, K., Kato, M., Nagata, M., Okabe, T. (1988), *Electron. Commun. Jpn.* **71,** P. 2, 83–92.

Weinberg, Z. A., Rubloff, G. W. (1978), *Appl. Phys. Lett.* **32,** 184–186.

Williams, R. (1965), *Phys. Rev.* **140,** A569–575.

Winokur, P. S., Sokoloski, M. M. (1976), *Appl. Phys. Lett.* **28,** 627–630.

Yamabe, K., Taniguchi, K. (1985), *IEEE Trans. Electron Devices* **ED-32,** 423–428.

Yamaguchi, K. (1983), *IEEE Trans. Electron Devices* **ED-30,** 658–663.

Yeh, C. F., Lin, S. S., Yang, T. Z., Chen, C. L., Yang, Y. C. (1994), *IEEE Trans. Electron Devices* **ED-41,** 173–179.

Yokogawa, K., Yajima, Y., Mizutani, T., Nishimatsu, S., Suzuki, K. (1990), *Jpn. J. Appl. Phys.* **29,** 2265–2268.

Yunogami, T., Mizutani, T. (1993), *J. Appl. Phys.* **73,** 8184–8188.

Zener, C. (1934), *Proc. Roy. Soc.* **A145,** 523–529.

Further Reading

Ando, T., Fowler, A. B., Stern, F. (1982), *Rev. Mod. Phys.* **54,** 437–672.

Grove, A. S. (1967), *Physics and Technology of Semiconductor Devices,* New York: Wiley.

Schroder, D. K. (1990), *Semiconductor Material and Device Characterization,* New York: Wiley.

SIGNAL PROCESSING, ACOUSTICAL

See UNDERWATER AND ATMOSPHERIC ACOUSTIC SIGNAL PROCESSING

SILICON DIOXIDE

R. Brückner, *Institut für Nichtmetallische Werkstoffe, Technische Universität Berlin, Garmisch-Partenkirchen, Germany.*

	Introduction	101
1.	**The Various Phases of Silicon Dioxide**	101
1.1	Phases at Atmospheric Pressure, Properties	101
1.2	High-Pressure Phases, Properties	102
1.3	Morphology and Main Applications	102
1.3.1	Quartz and Cristobalite	102
1.3.2	Vitreous Silica, Silica Gel, Silica-Glass Fibers	103
1.4	The System Si–O, Bond Nature and Structural Aspects	105
2.	**Properties of Quartz and Primarily of Vitreous Silica, Special Applications**	106
2.1	The Vitreous State of Glasses and the Amorphous State of Silicon Dioxide	107
2.2	Mechanical Properties	108
2.2.1	Elastic Constants and Internal Friction	108
2.2.2	Strength and Fracture, Fracture Toughness and Fatigue	109
2.3	Thermal Properties	110
2.3.1	Density and Thermal Expansion	110
2.3.2	Heat Capacity and Heat Conduction	111
2.3.3	Diffusion	112
2.3.4	Viscosity	114
2.4	Optical Properties	115
2.4.1	Refractive Index, Reflectance, Transmittance, Luminescence, Light Scattering	116
2.4.2	Optical Waveguides	118
2.4.3	Nonlinear Properties of Optical Silica-Glass Fibers; Optical Fiber Sensors	119
2.5	Electrical and Electronic Properties, Structural Defects	121
2.5.1	Conduction and Losses	121
2.5.2	Electronic Structure and Properties; Density of States	122
2.5.3	Structural Defects, Paramagnetic and Color Centers	123
2.5.4	SiO_2 Films in Microelectronic Devices	125
2.6	Applications of Vitreous Silica	128
3.	**Silica Gel, Aerogels**	129
4.	**Summary**	129
	Acknowledgments	129
	Glossary	129
	Works Cited	130
	Further Reading	131

INTRODUCTION

Silicon dioxide is a very simple chemical compound that is widely spread over the earth's crust in the form of quartz, the most stable modification at normal temperature and pressure conditions. It serves as a basic raw material for silicate glasses and silicate ceramics. As simple as this compound is on the one hand, its crystallographic modifications and their structures are manyfold and interesting on the other. However, the most interesting version with respect to properties and physical applications is the noncrystalline state of SiO_2, that is, silica glass and silica gel. Therefore, in this chapter the many varieties of silica glass will be discussed preferably.

1. THE VARIOUS PHASES OF SILICON DIOXIDE

1.1 Phases at Atmospheric Pressure, Properties

At normal atmospheric pressure, the following modifications of silica exist (Fig. 1): α-quartz is the most stable low-temperature

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

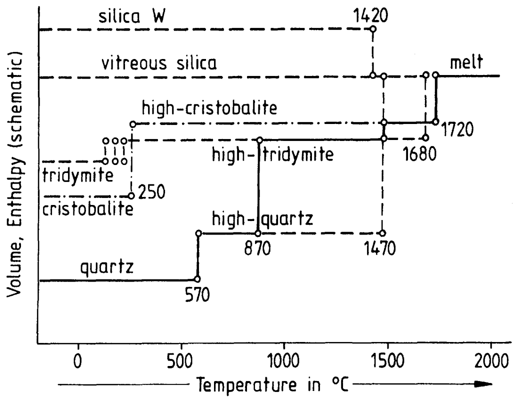

FIG. 1. Silica phases at atmospheric pressure (semi-schematic).

trigonal phase. At 573 °C the structure changes to the high-temperature hexagonal one, and at 1470 °C this phase converts to cubic high-cristobalite that has a melting temperature of $T_m = 1713$ °C. Upon cooling, a very viscous melt from above T_m undercooling is obtained very easily without crystallization that is frozen to a glass below about 1200 °C (Hetherington *et al.,* 1964), the so-called glass transition temperature *Tg*.

In impure SiO_2, with impurities such as alkali oxide, hexagonal high tridymite occurs at 867 °C from quartz, and then at 1470 °C it converts into the cubic high cristobalite. Because of impurities, the orthorhombic low tridymite has numerous variations, called tridymite M with inversions I to III, S with I to VI, and tridymite U. In a strict sense tridymite does not belong to the pure SiO_2 system (Flörke, 1956a). All low–high inversions of quartz, cristobalite, and tridymite are reversible displacive transformations that take place rapidly, whereas the conversions to another phase, e.g., quartz to cristobalite or cristobalite to tridymite and vice versa, are reconstructive transformations with very slow rates. These transformation rates are one reason, among others, that SiO_2 is a very good glass former.

Another low-pressure modification of SiO_2 occurs at high temperatures under reducing, water-vapor-free conditions and under conditions at which oxidation of SiO occurs: the so-called silica W that crystallizes in long needles with a chain structure such as (Sosman, 1965)

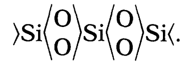

$$\rangle Si\langle{}^{O}_{O}\rangle Si\langle{}^{O}_{O}\rangle Si\langle. \qquad (1)$$

This modification is unstable and will change in the presence of water vapor.

On the basis of density, the following series comprise approximately the low-pressure low-temperature modifications, in addition to the diagram of Fig. 1: quartz, 2.65; cristobalite, 2.33; tridymite, 2.46; fused silica, 2.20; and silica W, 1.96 g cm^{-3}. The refractive indices of these modifications change parallel to the changes in densities.

1.2 High-Pressure Phases, Properties

Keatite is created hydrothermally at about 380–540 °C and at 88–130 MPa. The density is about 2.50 g cm^{-3}, and its thermal expansion is negative. The densest phase of silica (4.35 g cm^{-3}) is stishovite, which is produced approximately at 16 GPa and at 1200–1400 °C. It is the only modification of SiO_2 with a sixfold coordination of Si^{4+} (Sosman, 1965).

Coesite, the second heaviest phase of silica (density 3.01 g cm^{-3}), is formed at temperatures of 300–1700 °C and under pressures of 1.5–4.0 GPa.

1.3 Morphology and Main Applications

Although silicon dioxide exists in a large number of modifications and is able to maintain the various fundamental phases at normal pressure and room temperature, even the high-pressure modifications, only a few of them are suitable for applications. The reason is that the crystals of many of the modifications have only been prepared so far in sizes <1 mm in their maximum dimension (Sosman, 1965). Quartz and fused silica are the great exceptions. Cristobalite is of importance in the high-temperature furnace industry (Jebsen-Marwedel and Brückner, 1980).

1.3.1 Quartz and Cristobalite The appearance of quartz in nature is manyfold. Relatively large crystals are rock quartz, smoky quartz, rose quartz, amethyst, mild quartz, cat's eye, tiger's eye, etc. Fine crystalline quartz are chalcedon, onyx, sardonyx, jade, agate, etc. Quartz crystals as well-developed large pieces can be found in nature as rock crystals that were grown up hydrothermally from solutions. Also the artificially produced quartz crystals can be obtained as large crystals and are produced via the hydrothermal route. Usually the crystals are of excellent transparency that may be reduced by γ radiation (smoky rock crystals) or that will be reduced selectively resulting in amethyst or in rose quartz. While the smoky rock crystals can be decolored by heating up to about 250–300 °C where the paramagnetic color centers or the structural defects are regenerated (Sosman, 1965; Griscom, 1986), the violet or purple and red color of the amethyst and rose quartz, respectively, cannot be removed because they are produced by iron, manganese, and titanium impurities. Other remarkable properties of quartz are the birefringence with an optical indicatrix of a rotational ellipsoid with the two refractive indices $n_\omega = 1.544$ and $n_\epsilon = 1.553$; the rotary power for plane-polarized light along the *c* axis of the hexagonal low-temperature modification in a dextro- or levo-oriented sense depends on the twisted structure of the SiO_4 tetrahedra along the *c* axis. This property is less pronounced in quartz crystals with lower quality due to twinning. Amethyst has an "alternating dextro- and levostructure with such a remarkable—never artificially reached—symmetry of twinning that no rotary power for plane-polarized light exists in it" (Sosman, 1965).

One of the applications of the optical rotary power of quartz is an optical element, called half-shadow plate, for dextro- and levocircularly polarized light. Well known are the $\lambda/10$, $\lambda/20$, and $\lambda/4$ plates for optical compensators in order to measure birefringence of other substances or the so-called quartz wedges for continuously produced birefringence and, in conjunction with crossed Nicol prisms, to get optimum contrast [Nurmarski method for the phase-contrast microscope (see OPTICAL MICROSCOPY)].

Another very important application of well-developed quartz crystals is based on the piezoelectric effect. This property is the basis of ultrasonic sources usually in the megahertz range for longitudinal and transversal waves and is also the basis of steering elements in chronometry (for more details see QUARTZ).

Cristobalite can be obtained only as small crystals usually by a reconstructive transformation at high temperatures (~1300–1650 °C) from quartz or from fused silica. The main reason for the small crystal sizes is the displacive high–low (cubic-tetragonal) transformation with a volume contraction of ≤4% at temperatures between 180 and 270 °C, which depends on impurities and on the temperature of the reconstructive transformation. The lower the latter temperature and the larger the impurity concentration, the smaller are the displacive transformation temperatures and the volume changes. The reason is the increasing disorder of the structure (Flörke, 1956b). The main application of cristobalite is the use as refractory bricks, particularly in glass industry furnaces, particularly as a light material (ρ = 2.33 g cm^{-3}) for the furnace ceiling (Jebsen-Marwedel and Brückner, 1980).

1.3.2 Vitreous Silica, Silica Gel, Silica-Glass Fibers These varieties are the glassy versions of SiO_2 and respectively the amorphous versions of SiO_2 gels. The glassy version is also called vitreous silica, silica glass, fused quartz, or quartz glass. It can be obtained in two ways: first of all by melting (see Fig. 1) and recently by the sol-gel and the colloid methods; likewise, in principle, the silica glass fibers. Because of the noncrystalline condition, these SiO_2 varieties can be formed into all geometries that are possible by the usual glass-forming principles. This formability is the main reason for the larger range of and potential for applications as compared to the crystalline versions of SiO_2 (Sec. 1.3.1).

The properties that are unique for the noncrystalline versions of SiO_2 are the low coefficient of thermal expansion ($\sim 5 \times 10^{-7}$ K^{-1}), the high glass transition temperature (~1100–1200 °C), the high tensile strength of the fibers (~10 GPa), the excellent elastic constants, the low optical absorption (losses <0.5 dB/km), and the excellent chemical stability.

These excellent properties, which are useful for many important applications (e.g., the use of SiO_2 crucibles in semiconductor production, the use in optical waveguides, and the use in microelectronic elements on the basis of silicon, see Sec. 2.6), are possible only through special preparation techniques that have lead to many kinds of fused silica. These may be divided into a few groups or types, listed in Table 1. The different impu-

Table 1. Characteristic types of vitreous silica developed over decades for special properties and applications and from different sources.

		Typical impurities in ppm				Characteristic optical and other properties, remarks
Type	Method of manufacture	Al	Na	OH	Cl	
I	Electric melting of quartz crystals in crucibles under vacuum	10–100	4	≤5	Not detected (n.d.)	Good transmittance in the near infrared, absorption band at 247 nm
II	H_2/O_2 flame fusion of quartz crystals	<30	1	150–300	(n.d.)	Good UV transmittance
III	Hydrolysis of $SiCl_4$ (or other silanes) in H_2/O_2 flame, deposition and direct consolidation in the same step	⩽1	⩽1	1000	<50	Excellent UV transmittance, absorption bands at 1.4, 2.2, and 2.7 μm
IIIa	Hydrolysis of $SiCl_4$ (or other silanes) in H_2/O_2 flame, deposition of soot particles to form a porous intermediate body, consolidtaion to full density in a separate step; no dehydration	⩽1	⩽1	200	<50	Excellent UV transmittance, better IR transmittance than type III
IIIb	As IIIa wtih dehydration of intermediate soot body by Cl treatment	⩽1	⩽1	<0.1	1000	Excellent transmittance from UV to IR
IV	Oxidation of $SiCl_4$ in water-vapor–free plasma flame	⩽1	⩽1	<5	<50	Purest type so far
Va	Electric melting of very pure pegmatitic quartz in Mo or W crucibles in H_2-containing atmosphere as melted	10–20	~1	100	(n.d.)	OH absorption at 2.7 μm and small SiH IR-band, large absorption around 245-nm; high gas content caused by physically dissolved H_2
Vb	As Va but out-gassed 10 h at 1080 °C in vacuum or air	10–20	~1	<1–15	(n.d.)	Absorption at 247 nm higher viscosity than types II–IV; volume-wise the most importnat type at present, nearly H_2 free
VI	Sintering of colloidal SiO_2 glass powder preforms after chloride reaction purification	⩽1	⩽1	⩽1	100–500	Fabricated on pilot scale so far
VII	Sol-gel route of Si-alkoxide, drying and sintering	<1	<1	<1	(n.d.)	Fabricated on laboratory scale so far

rities of the various silica-glass types effect the various properties and will be discussed in Sec. 2. The principal preparation techniques are the electrical or flame (H_2/O_2 or plasma) melting of (rock) quartz crystals, the hydrolysis and pyrolysis of $SiCl_4$, the sol-gel route starting from Si-alkoxide solution, and sintering of a silica gel or sintering together with colloidal amorphous SiO_2 powder (Clasen, 1987; Hench, 1993). The advantage of the sol-gel route is the high purity of the silica glass (even without OH^- and Cl^-; Hench, 1993). The advantage of the colloidal particle sintering technique is the high cleaning effect from impurities such as metal ions and water and the possibility of the introduction of specific ions for coloring and of colloidal coloring metals (Clasen, 1993).

The preparation of silica gel can be done in an additional way to the sol-gel route.

Starting material is the so-called water glass, an alkali silicate glass with ≤50 wt % Na_2O. The Na_2O content is leached out by water; the water is exchanged by alcohol or by fluid CO_2, which is transformed to a temperature and pressure above the critical point of CO_2 where the surface tension cannot contract the porous silica gel. In this manner silica gels with a large porosity can be obtained in large pieces without shrinking cracks. These are the so-called aerogels with densities as small as 0.005 g cm^{-3}. The main properties of these are their ultrasonic properties (see Sec. 3) and excellent thermal isolation.

Silica-glass fibers have much higher tensile strength than the bulk silica glass. Additionally, in the case of types IIIb, IV, and Vb silica glasses (Table 1), application is usual as optical waveguides with high optical transparency or low optical losses in the near-infrared part of the spectrum. The production is possible usually via a so-called preform and pulling out the heated end to a fiber of the desired diameter. In the case of simple fibers for reinforcing purposes of composite materials and thermal isolating, the preform consists only of a silica-glass rod of about 8 mm in diameter. If d_r is the diameter of the rod, d_f that of the fiber, and V_f the drawing velocity of the fiber, the velocity V_r of the rod results from the relation

$$V_r{:}V_f = d_f^2{:}d_r^2.$$

The drawing temperatures are well above 2000 °C. The silica-glass fibers as optical waveguides, on the other hand, are built up as step-index or gradient-index fibers, where the refractive index of the core is larger than that of the slad. More details are given in Sec. 2.4.2.

1.4 The System Si–O, Bond Nature and Structural Aspects

The total system Si–O is not well known, except for the two extreme components, Si and SiO_2. The meta composition, SiO, is stable only at high temperatures and decomposes into SiO + SiO_2 and Si + SiO at somewhat lower temperatures and into Si + SiO_2 at still lower temperatures (Sosman, 1965). Silicon monoxide, SiO, is formed as one of the gaseous products when silica is heated above 1500 °C or is in contact with silicon, carbon, hydrogen, or hydrocarbon at or above temperatures of about 1000 °C. Figure 2 shows the partial pressures of the decomposition of SiO_2 in a neutral atmosphere between 1500 and 4000 K (Fanderlik, 1991).

The building unit of crystalline and vitreous silicon dioxide is the SiO_4 tetrahedron, or better, the $SiO_{4/2}$ tetrahedron. The only exception is stishovite with its SiO_6 octahedron. The SiO_4 tetrahedron establishes the short-range order and is based on the excited state of the covalent part of the silicon atom forming the sp^3 hybride: $1s^2\ 2s^2\ 2p^6\ 3s^1\ 3p^3$. This arrangement corresponds to four valencies with σ bonds directed toward the four corners of the regular tetrahedron occupied by four oxygen atoms. However, the Si–O bond is, to a relatively large extent (~40%), also ionic or heteropolar with respect to the Pauling scale of electronegativity data. This mixed bonding character was established among others by determination of the electron density of low-temperature quartz by means of x-ray techniques (Brill *et al.*, 1942).

The three-dimensional linkage of adjacent $SiO_{4/2}$ tetrahedra at the oxygen corners determines the long-range order of the different crystalline silica modifications. Thus, the Si–

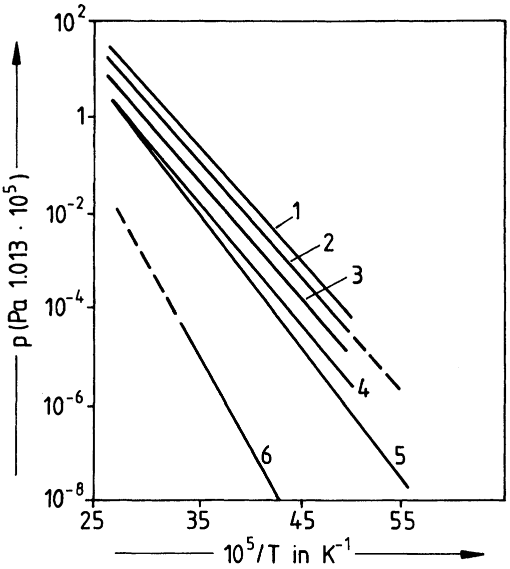

FIG. 2. Partial pressure of the reaction products of SiO_2 in the range of 1500–4000 K in neutral atmosphere. 1 = p_{total}, 2 = p_{SiO}, 3 = p_{O_2}, 4 = p_O, 5 = p_{SiO_2}, 6 = $p_{Si_2O_2}$ (reproduced with kind permission of Elsevier Science from Fanderlik, 1991).

O–Si bond angle is 144° for the trigonal low quartz, 146° for the hexagonal high quartz, 148° for the tetragonal low cristobalite, and 152° for the cubic high cristobalite. In contrast to the crystalline modifications, this bond angle has a wide distribution in vitreous silica: 120° to 180° with a maximum frequency at 145°. The reason for this broad distribution is the absence of long-range order in vitreous silica that is responsible for the large flexibility of the silica-glass network.

While the structures of the crystalline phases of SiO_2 are well determined by x-ray and neutron diffraction methods, the investigations and discussions on silica-glass structure and glass structures in general are still going on. There is the random-network hypothesis of Zachariasen (1932), Warren *et al.* (1936), and Warren (1991) on the one hand and the crystallite hypotheses of Lebedev (1921) and Valenkov and Porai-Koshits (1936) on the other. The discussions had been very controversial until Lebedev (1940) himself wrote that there is no significant difference between the two hypotheses because one can find ordered regions in ideal disordered systems, and it is possible to produce such small crystals that they are practically no longer crystals (paracrystals, x-ray amorphous). From the overwhelming amount of literature about the structure of silica glasses, the following references may be cited: Wright and Leadbetter (1976), Porai-Koshits (1977), Konnert *et al.* (1982), Zarzycki (1982), Wright *et al.* (1982), and Hosemann *et al.* (1986). Because of the uncertainties of the diffraction methods when applied to amorphous structures, many other structural analysis methods had been applied such as NMR (nuclear magnetic resonance), EPR (electron paramagnetic resonance), Raman and infrared spectroscopy, XPS (x-ray photoelectron spectroscopy), and STEM (scanning transmission electron microscopy). Representatively for those investigations may be cited the monograph of Wong and Angell (1976).

It is beyond the scope of this article to go into more detail of the numerous, partly very selective, and sometimes controversial interpretations and discussions. With respect to the silica glasses, the structural concept of a nearly complete continuous three-dimensional network of $SiO_{4/2}$ tetrahedra with various selective short-range orders, with the lack of long-range order, and with a not yet exactly known intermediate-range order appears to be generally accepted, as may be concluded also from the excellent physical properties of the various silica glasses (Table 1) described in the following paragraphs of this contribution.

The longer-range order of the crystalline modifications as well as the short-range order of the silica glass can be changed by irradiation with energetic neutrons (1 MeV, $>10^{19}$ n/cm^2). Thus, quartz loses its rotary power and its piezoelectric ability and reaches a density of 2.26 g cm^{-3}. When exposed to similar doses, vitreous silica, on the other hand, has an increase of density from 2.20 to 2.26 g cm^{-3}; i.e., the same final density value as the irradiated crystalline modifications (Primak, 1958). In the structure of this disordered version of SiO_2, there is a decrease in the Si–Si distances and a shift of the Si–O–Si bond-angle distribution to smaller angles. From the view of the vitreous silica structure, the disordered version shows also a shift to more crystallinity, which is evident from increase in the density and refractive index and also from increases of thermal conductivity at low temperature as compared with an unirradiated silica glass (Cohen, 1958; Evans *et al.*, 1982; see also Sec. 2.3.2).

2. PROPERTIES OF QUARTZ AND PRIMARILY OF VITREOUS SILICA, SPECIAL APPLICATIONS

A typical behavior of the vitreous state is a sensitive dependence of the properties on the mechanical and thermal prehistory in a defined manner. Vitreous silica (and other glasses to a certain degree, too) can be densified permanently even at room temperature up to about 15% of specific volume or density after application of high pressure (various authors; see Brückner, 1970/71, pp. 148–152). The density of a glass is also dependent on the cooling rates from temperatures above T_g to at least 100 K below T_g. This behavior of glasses is a consequence of the lack of long-range order, and this lack affects other glass properties such as viscosity, refractive index, density, and glass-fiber properties and, in the case of silica glasses, also the thermal expansion. Therefore, it is very important to know the thermal and mechanical prehistory of any glass, here of the silica glass; for instance: the smaller the cooling rate, the smaller is the density of vitreous silica. These effects are the consequence of the glass transition range, characterized by the glass transition temperature T_g (or by the fictive temperature T_f; see Sec. 2.1.1) at which the structure of the melt will be frozen in and at which the coefficient of thermal expansion changes its amount reversibly if the cooling and heating rate are equal. Therefore, they are not observed in crystals with or without displacive and reconstructive transformations.

2.1 The Vitreous State of Glasses and the Vitreous and Amorphous State of Silicon Dioxide

The vitreous state of usual, multicomponent glasses as compared with their crystalline state is shown in Fig. 3. If crystallization is prevented by fast enough cooling, the structure of the melt (above or below T_m, the melting point) is maintained by freezing to a solid body, called glass, at a fictive temperature $T_f > T_g$. If the cooling rate is low enough, the metastable (undercooled) glass melt freezes in at a lower temperature than T_g (at $T_f < T_g$) and at a smaller specific volume with a less open structure that is determined by the metastable equilibrium curve (solid-dashed straight line in Fig. 3). The points of intersection between the two straight lines with the lower temperature coefficient and the metastable straight line with the larger temperature coefficient indicate the fictive temperatures T_f, which depend strongly on the cooling (or quenching) rate. Therefore, the term "fictive temperature" is a very practical measure for the thermal prehistory. The concept of fictive temperature in glass science (Tool, 1946) is that temperature T_f to which the structural state corresponds after rapid cooling or rapid heating, where "rapid" means a rate $v = |\Delta T/\Delta t| > |\Delta T/\tau|$, with τ being the relaxation time of the structural rearrangement or that time during which a property (volume, enthalpy, etc.) relaxes to the e^{-1} amount of the deviation from the metastable equilibrium curve (see arrows from the curves of fast cooling or that of fast heating to the metastable equilibrium curve at a temperature $T = T_{\text{meas.}}$ in Fig. 3, where $T_{\text{meas.}}$ means the temperature at which a certain property is measured, and is time dependent). Thus, many properties of glasses depend on T_f. In the case of fibers, very high T_f values are reached. The so-called glass transition temperature T_g is obtained by cooling and heating a glass with a rate of $v = |dT/dt| = 2$ to 5 K/min. In this case $T_g = T_f$ by definition.

Silica glass is unusual in that the metastable equilibrium curve does not have a positive temperature coefficient as in the case of all known glasses, but it does have a negative one between 1000 and about 1500 °C that

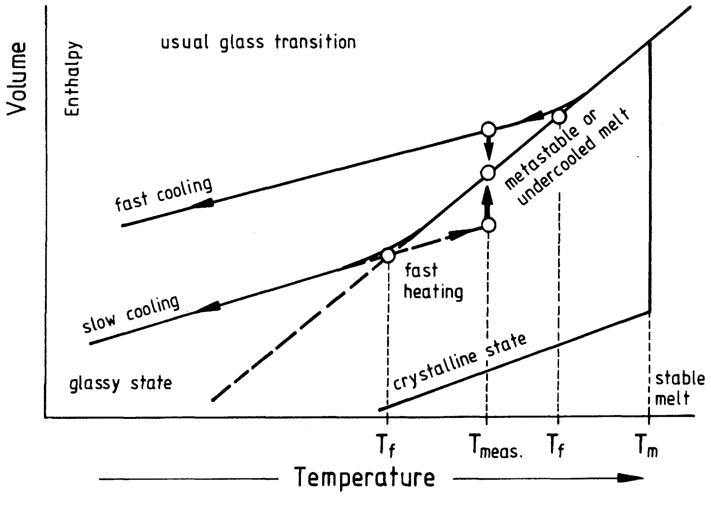

FIG. 3. Enthalpy and/or specific volume of multicomponent silicate glasses versus temperature for two different cooling and heating rates and fictive temperatures, $T_f \gtrless T_g$. T_g can be between T_{meas} and T_f or at T_{meas}. Comparison with the crystalline state (schematic). The partly solid and dashed straight line is the metastable equilibrium curve for a usual (multicomponent) oxide glass.

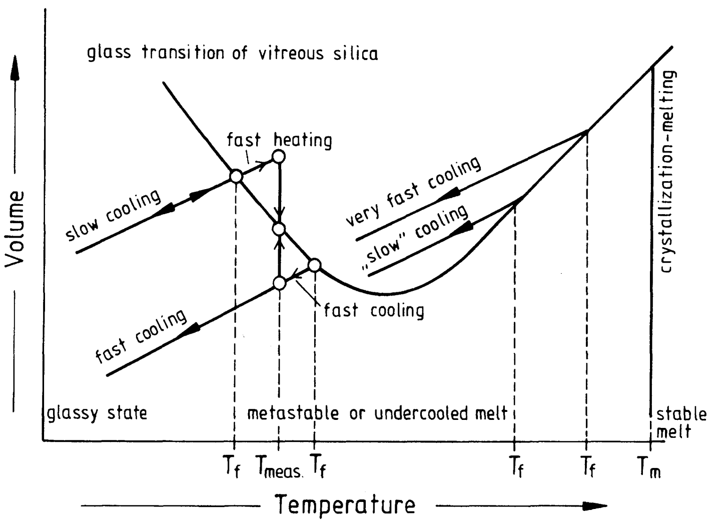

FIG. 4. Specific volume of vitreous silica versus temperature in the metastable and glassy range for various fictive temperatures (schematic), different cooling and heating rates, and fictive temperatures, $T_f \gtrless T_{\mathrm{meas}}$. T_g can be between T_{meas} and T_f or at T_{meas}. The arrows indicate the route of relaxation to the metastable equilibrium at T_{meas} for samples with two different fictive temperatures T_f (schematic).

changes to a positive one with increasing OH^- content (Figs. 4 and 5) and/or network modifier concentration (e.g., alkali oxide) (Brückner, 1970/71). This change means that the metastable range of hydroxyl-free and -poor silica glasses have a negative temperature coefficient at least partly in the temperature range of 1000–1500 °C. There is also a negative coefficient in the temperature range below 140 or 210 K, depending on both thermal history and hydroxyl content (see Sec. 2.3.1).

The amorphous state of silica gel (see Secs. 1.3.2 and 3) is due to its low-temperature treatment (far below T_g) and due to its preparation technique (Fricke, 1988). If silica gel (or aerogel) is heated to the T_g range, a sintering process with a large volume contraction leads to vitreous silica. Thin films of SiO_2 (or $2SiO + O_2$) produced by vaporization on a cool target are initially amorphous, not glassy or vitreous, because, similar to the gel state, they do not have the characteristic T_g, but they will be transformed also to the vitreous state by heating to the T_g range or beyond it in neutral (or oxidizing, O_2) atmosphere. In this way the glassy state is more stable than the amorphous state, although the glassy state itself is thermodynamically unstable and only preserved by the large viscosity (see Sec. 2.3.4).

2.2 Mechanical Properties

2.2.1 Elastic Constants and Internal Friction The elastic moduli also have unusual properties: the temperature coefficients

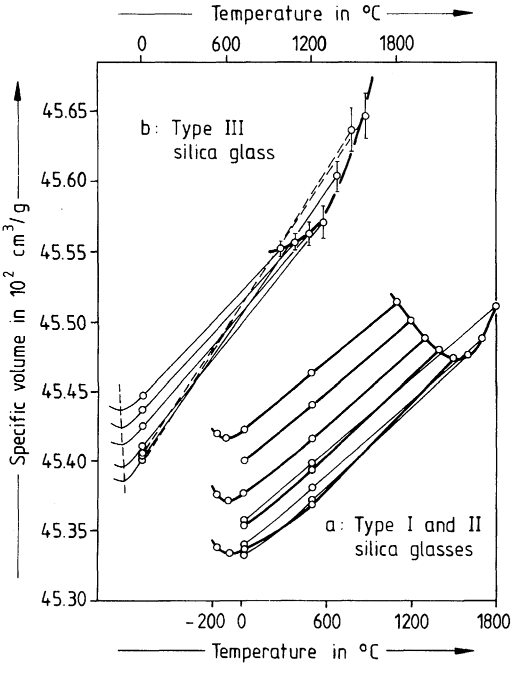

FIG. 5. Volume–temperature behavior of hydroxyl-poor (types I, II, and IV) and hydroxyl-rich (type III) silica glasses in the metastable range (thick curves). The more or less straight lines (thin curves) are mean thermal volume expansion curves of the glasses with different fictive temperatures. Upper and lower temperature scales belong to type-III and type -I, -II, and -IV silica glasses, respectively (Brückner, 1970/71).

of the elastic moduli and of the longitudinal elastic strain (of fibers and rods) are positive. The elastic moduli may be expressed by experimentally determined relations: $E = 73.3 \times (1 + 5.75\epsilon)$ GPa, $G = 32.1(1 + 3.06\epsilon)$ GPa, where E is Young's modulus, G is the rigidity modulus, and ϵ is the strain. The consequence is the positive pressure coefficient of the compressibility, β_m (p = compressive stress):

$$\beta_m = a(t) + b(T)p = \frac{1}{V_0}\frac{V_0 - V_p}{p - p_0} = 1/K = \frac{3(1 - 2\nu)}{E} \qquad (3)$$

with

$$a(T) = (26.43 - 0.0025T) \times 10^{-12}\ \mathrm{Pa}^{-1}, \qquad (3a)$$

$$b(T) = (43.6 - 0.080T) \times 10^{-17}\ \mathrm{Pa}^{-1}, \qquad (3b)$$

where ν is Poisson's ratio and T is in °C. These data are valid in the temperature range of 22–260 °C and in the pressure range of 0–300 MPa (see Brückner, 1970/71). Thus, the compressibility increases with pressure, whereas those of quartz and of most other glasses and substances decrease with pressure. Above 350 MPa the compressibility has a normal dependence on pressure for fused silica.

The elastic properties measured at room temperature at acoustic frequencies are a function of the thermal history (see Sec. 2.1) because the density depends also on it. As will be shown in Sec. 2.3.1, the density increases with the fictive temperature in the temperature range 1000–1550 °C before a normal negative temperature coefficient is obtained (see Brückner, 1970/71). In this way the ultrasonic shear velocity, the shear modulus, and the compressibility decrease with increasing density (increasing fictive temperature), whereas the dilatational velocity, the Lamé parameter, and Poisson's ratio increase with increasing density. Increasing hydroxyl and chloride content makes vitreous silica "softer" (lower moduli, lower viscosity).

The deviation of Young's modulus from linear positive temperature dependence at 800–1000 °C is connected to an increase of internal friction (see Brückner, 1970/71). Structurally, these friction losses are related to the initial breakdown of the polymer network from a rigid glassy state to an undercooled liquid metastable state. As in the case of other properties, the hydroxyl content influences the internal friction coefficient; the hydroxyl-rich silica glasses (e.g., type III) have a steep increase in internal friction at lower temperatures than the hydroxyl-free or -poor silica glasses (types I, II, IV, to VII), a behavior that is closely connected to the viscosity (see Sec. 2.3.4).

At low temperatures (<60 K) and high frequencies (ultrasonic longitudinal waves in the kilohertz and megahertz range), another internal-friction peak is observed together with a minimum of rigidity due to a relaxation mechanism (see Brückner, 1970/71). It is remarkable that this kind of absorption does not occur in the crystalline quartz structure. The low activation energy of 4.2 kJ/mole indicates that the losses may be due to Si–O–Si bond deformations of those oxygen atoms having alternative metastable equilibrium positions of equal energies. Again the maximum of the attenuation of the intensity of ultrasonic waves is shifted by the hydroxyl content: from 43.5 K for hydroxyl-poor types, I and II, to 47.5 K for type-III fused silica.

2.2.2 Strength and Fracture, Fracture Toughness and Fatigue Strength is a statistically influenced and surface-dependent property for an ideal elastic brittle body such as silica glass. Small mechanical damage causes a large weakening. Therefore, bulk material with surface cracks has strengths ranging from 60 to 100 MPa. The highest observed tensile strengths reported for silica fibers are about 10–15 GPa, close to the theoretical strength of about 30 GPa (see Brückner, 1970/71) and for which the microcracks (so-called Griffith cracks) are on the order of magnitude of nanometers. In addition to a much smaller statistical distribution of cracks within the fibers, the elongation and thinning of the cracks during the fiber-drawing process might explain the large difference of the strength between the fiber and bulk material.

On account of brittle fracture, the fracture toughness of silica glass, measured by the critical stress-intensity factor, is 0.79 MPa $\mathrm{m}^{1/2}$ at 300 K.

The static fatigue behavior of silica-glass fibers (20–40-μm diameters) was studied extensively. While in vacuo at 78 K no fatigue was observed during a time of days, only half of the strength at 78 K was observed in vacuo at room temperature with a decrease from about 900 to 700 MPa during a month. In air, at room temperature, a decrease from 600 to about 300 MPa during a year was measured (Proctor *et al.*, 1967; see Brückner, 1970/71).

2.3 Thermal Properties

2.3.1 Density and Thermal Expansion

One of the most interesting anomalies of silica glasses as compared with multicomponent glasses (see GLASSES) is the density or specific volume and its thermal behavior. As shown in Fig. 4 (Sec. 2.1), the metastable equilibrium curve has partly (between 1000 and 1500 °C) a negative temperature coefficient. This coefficient is a consequence of the decrease of specific volume within the metastable range, which has a minimum at about 1500–1550 °C for the hydroxyl-poor or hydroxyl-free silica glasses and which is not the case for the silica glass with 0.12% OH^- content (type III) (Fig. 5). The low-temperature minimum around −100 °C is still present; it is only shifted to lower temperatures. Both minima are influenced also by the fictive temperature; thus, the minimum temperatures for the hydroxyl-poor silica glass is shifted from −133 to −95 °C, and that of the hydroxyl-rich version is shifted from −120 to −57 °C (see also Fig. 7) with increasing fictive temperature.

The thermal expansion of silica glasses is also dependent on the fictive temperature and on hydroxyl content. Figure 6 shows thermal expansion data for type-II silica-glass samples with differing fictive temperatures. Hydroxyl-poor samples with a fictive temperature corresponding to the minimum of the specific volume at 1550 °C and hydroxyl-rich silica glasses at 1460 °C have the highest thermal expansion (curve 4); those of still higher or lower fictive temperatures have lower thermal expansion (e.g., curve 3). These effects are consistent with the Grüneisen relation:

$$\beta c_v/\alpha V = \text{constant}, \tag{4}$$

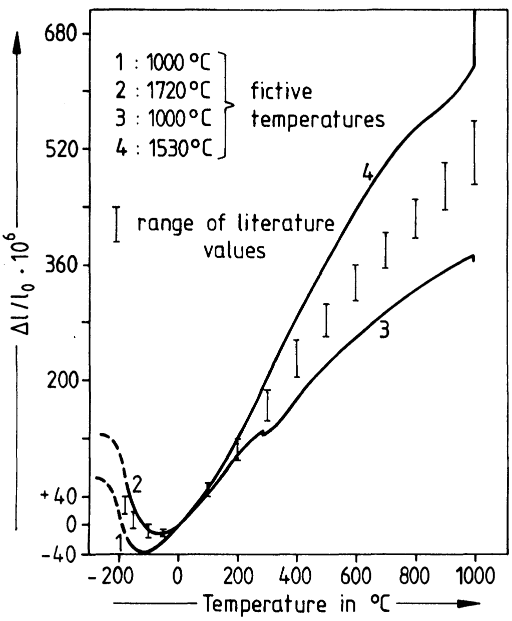

FIG. 6. Thermal expansion below T_g of type-II silica-glass samples with various fictive temperatures versus temperature (Brückner, 1970/71).

where β = compressibility, c_v = specific heat, α = coefficient of thermal expansion, and V = specific volume. If V has a minimum, α should have a maximum, and if β has a minimum, c_v should have a maximum. These effects are at least qualitatively the case for fused silica when the fictive temperature or when the temperature itself is changed from low (1000 °C) to high (1800 °C) (Brückner, 1970/71).

The low-temperature minimum is shifted to higher temperatures with increasing fictive temperature, i.e., it is independent of the high-temperature minimum. There is a linear relationship between the minimum temperature and fictive temperature (Fig. 7). Again a difference between the groups of hydroxyl-poor and hydroxyl-rich silica glasses is found (Brückner, 1970/71).

It is interesting that the crystalline high-temperature modifications, quartz, and crystobalite have a change from positive to negative thermal expansion between 800 and 1000 °C, indicating that structural fragments or preordered regions exist in the vitreous state that are responsible for the high-temperature minimum of silica glass. The low-

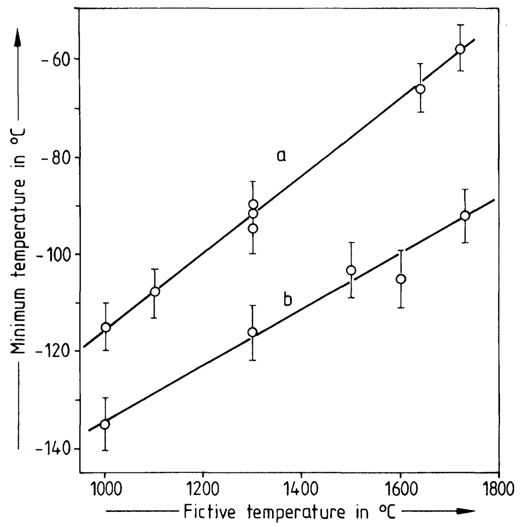

FIG. 7. Shift of the low-temperature minimum with fictive temperature for type-II (curve a) and type-III (curve b) vitreous silica (Brückner, 1970/71).

temperature minimum on the other hand is not found for quartz and crystobalite, indicating that it results from randomly disordered regions in which the preordered regions are embedded. Its shift with fictive temperature may be caused by structural stresses produced by quenching and freezing-in the structure at the fictive temperatures (Brückner, 1970/71). The effect of negative thermal expansion of vitreous silica, as well as that of high cristobalite and high quartz, may be understood qualitatively with special emphasis to the transversal vibrations of the oxygen atoms and with the concept that the major contribution to the thermal behavior is produced by the transversal oxygen vibrations.

2.3.2 Heat Capacity and Heat Conduction According to the classical expression for the temperature dependence of the specific heat,

$$c_v = 3RE(x) \tag{5}$$

in which E is the Einstein function,

$$E(x) = x^2e^x/(e^x - 1)^2, \tag{6}$$

with $x = h\nu/kT = \Phi/T$, where h = Planck's constant, ν = frequency of proper oscillation, k = Boltzmann's constant, R = gas constant, and Φ = the Debye temperature, the specific heat of vitreous silica was calculated, and a relatively good agreement with experimental values was obtained (see Brückner, 1970/71). If the following three Debye temperatures are chosen: $\Phi_S = 1100$ K for each of the three silicon vibrations, $\Phi_T = 370$ K for each of the three transversal oxygen vibration modes, and $\Phi_L = 1220$ K for each of the two longitudinal oxygen oscillations, the Φ_i correspond to vibrations in the IR range of 23×10^{12}, 7.7×10^{12}, and 25.4×10^{12} Hz.

The deviations at low temperatures <100 K are only seemingly small when plotted on an approximate scale; however, they contain an excess of specific heat not only for silica glasses but for all glassy materials, even for metallic glasses, an effect that does not exist in the corresponding crystalline phases, e.g., quartz where the T^3 law is valid. Hunklinger (1982) pointed out that the excess heat is a consequence of relaxation processes that "originates from the modulation of the thermodynamic equilibrium of a subsystem in a sample by the strain and stress field of the sound waves" (phonons). With the introduction of only two possible configurational states, the double-well potential, and of resonant transition between the energies of a two-level system (TLS) for phonons, models have been developed to explain the large difference, a factor of 1000, for the specific heat between the SiO_2 glass and the crystalline state (quartz) at 25 mK. Low-energy excitations (phonons) interact strongly with the disordered network structure. Since the relaxation time of these interactions is in the range of 1 μs up to 10^4 s, the specific heat depends on the duration of a measurement, as was shown by various authors (see Hunklinger, 1982). It is suggested that the nature of these low-energy excitations, which also determine the ultrasonic properties of glasses, is based on clusters of atoms that undergo structure relaxation even at the lowest temperatures by tunneling between the states of the two-level system and/or by rotation of SiO_4 tetrahedra.

Although no exact explanation can be given so far for the excess specific heat at temperatures below 20 K, especially below 5 K, it is important that, in a qualitative way (inconsistent with lattice dynamics), a model is given with the two-level concept in which the Si–O–Si bending vibrations can be made responsible for all the low-temperature specific heat. All of the frequencies required to account for this excess heat capacity are to be found in the region of optical modes of very low frequencies with a continuum extending from 560 cm^{-1} down to 8 cm^{-1} as can be shown directly by the Raman spectrum and the Brillouin scattering spectra of fused silica (see Hunklinger, 1982).

A characteristic difference between the vitreous and crystalline state is very evident through the temperature dependence of heat conduction (Fig. 8). According to the regular atomic arrangements in crystal lattices, the crystals have large values of the mean free path for the phonons at low temperatures and therefore large values of heat conduction. The lack of periodicity and symmetry of a lattice in the glass network causes large interchange energies between thermoelastic waves in such a way that the mean free path, Λ, is significantly shorter. It is assumed that in case of glasses, Λ is independent of temperature, as is also the average sound velocity v; thus, the heat conduction,

$$\lambda_c = cv\Lambda/3 \tag{7}$$

is predominantly proportional to the specific heat c at most temperatures. For temperatures well above room temperature, the radiation heat conductivity (photon conduction),

$$\lambda_r = 16\sigma n^2 T^3/3\epsilon, \tag{8}$$

is superimposed, where σ is the Stefan-Boltzmann constant, n is the refractive index, ϵ is the optical absorption coefficient, and T is the absolute temperature.

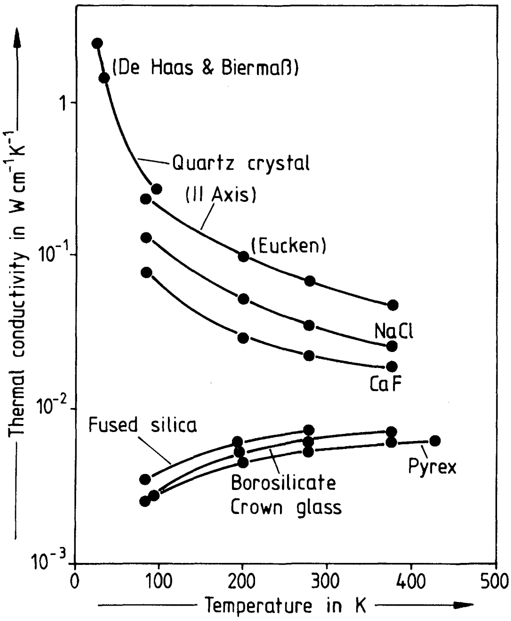

FIG. 8. Heat conduction of crystalline and glassy substances, particularly that of quartz and silica glass, versus temperature (gathered from various authors; see Brückner, 1970/71).

The proportionality between λ_c and c is valid to temperatures as low as liquid oxygen temperature (Hunklinger and Schickfus, 1981). At lower temperatures the conductivity decreases less than that expected for a much longer mean free path for the longitudinal than for the transversal waves caused by a small fraction of three-phonon processes at lower frequencies and temperatures. This deficiency leads to completely uncoupled longitudinal waves with lengths very much longer than those of the transversal waves, and this decoupling produces an additional conductivity (curve I in Fig. 9) proportional to T in the range 0–5 K in addition to curve II, resulting by superposition to the experimental curve III. Above 5 K a partial coupling causes a decrease of curve I that leads to the "knee" of curve III.

Besides thermal fluctuations, defects and voids with low-frequency mechanical resonances may have spin–lattice interactions and relaxations and, hence, unusual scattering of phonons that influence the short mean free path, especially those of the transversal phonons. Therefore, a dependence of thermal conductivity on thermal prehistory and on hydroxyl-poor and hydroxyl-rich silica glasses is to be expected.

2.3.3 Diffusion The gas permeation and the mobility of network-forming (oxygen, silicon) and network-modifying (alkali ions, OH^- groups) species are important in applications of vitreous silica as, e.g., high-temperature recipients. Figure 10 gives an impression of the large range of diffusivities of various atomic species as a function of temperature. Helium and hydrogen have high diffusivities due to their small sizes and due to the open network structure of vitre-

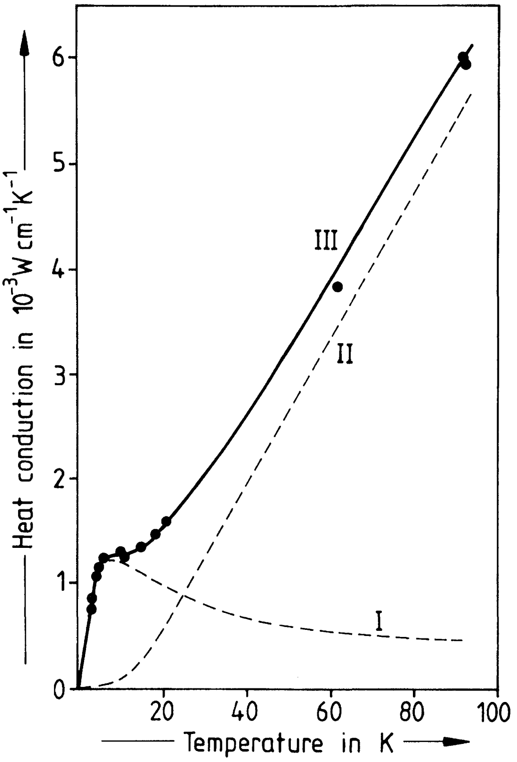

FIG. 9. Heat conduction of vitreous silica at low temperature. Curves I and II, calculated curves; curve III, sum of curves I and II. Symbols: measured values (gathered from Kittel, 1949, and Berman, 1951; see Brückner, 1970/71).

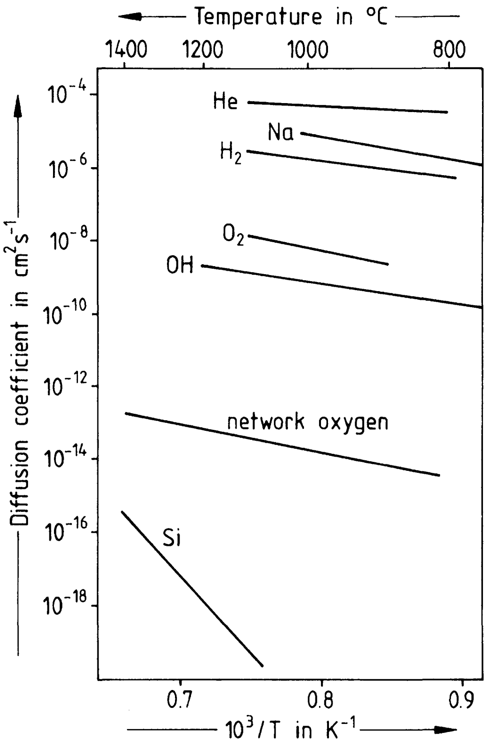

FIG. 10. Diffusion coefficient of various network-forming and network-modifying species in vitreous silica versus reciprocal absolute temperature (from Schaeffer, 1984).

ous silica. The diffusion coefficients and the activation energies of gases depend on the diameter of their ions, atoms, and molecules. Increasing activation energies are obtained in the series He → H_2 → Ne → O_2 → N_2 → Ar and decreasing values for the permeabilities. Structurally incorporated or interacting species usually have low diffusivities depending on bonding forces by which they are attached to the glass network. Network oxygen (not O_2) and silicon have the smallest diffusivites. Typical examples for interacting diffusion processes are for H_2O, which diffuses via an interaction with the silica network by forming SiOH groups, hydrogen, and network oxygen diffusion. Diffusion of O_2 plays an important role in the oxidation processes of silicon-containing materials producing noncrystalline silica layers (see Sec. 2.5.4). Schaeffer (1984) showed that the parabolic growth of SiO_2 layers ($d \propto t^{1/2}$) is connected to a parabolic rate constant that is, by a factor of 2, larger than the diffusion coefficient of network oxygen (see also Lamkin *et al.*, 1992).

In addition to the complex diffusivity of oxygen, the interaction diffusion process of hydrogen is also complex. To understand the behavior of the OH^- groups in vitreous silica preparation, conditions of SiO_2 glasses (see Table 1) must be known since they affect diffusion. In the case of type-Va vitreous silica, it may be noted that chemically bonded hydrogen can be set free at elevated temperatures from the so-called unstable OH^- groups. This reaction was originally written

$$2SiO_2 + H_2 \rightleftarrows \equiv SiOH + SiH \qquad (9)$$

or, better formulated,

$$\equiv Si{-}O{-}Si\equiv + H_2 \rightleftarrows \equiv SiOH + \equiv SiH. \qquad (10)$$

In this reaction diffusion model, the amount of easily and totally extractable hydrogen below T_g increases, together with physically dissolved H_2 (Stone, 1987; Witzke, 1990). It explains the so-called diffusion anomaly of hydrogen (Lee, 1964; Ogawa *et al.*, 1977) in vitreous silica. On the other hand, the interacting diffusion process of Si–OH groups results in, originally written (Stone, 1987; Witzke, 1990),

$$2SiO_2 + H_2O \rightleftarrows 2 \equiv SiOH \tag{11}$$

or, better formulated,

$$\equiv Si{-}O{-}Si \equiv \; + H_2O \rightleftarrows 2 \equiv Si{-}OH, \tag{12}$$

the so-called stable OH groups, which cannot be easily and totally removed by thermal treatment below T_g.

A typical example for a diffusion process of structurally incorporated species is sodium ions, which have the highest diffusivity with the exception of helium (Fig. 10). The diffusivity depends not only on ionic radius and bonding force but also on the network structure. This is obvious from Fig. 11, which shows that there are large differences particularly parallel and perpendicular to the *c*-axis directions in quartz as well as in the hydroxyl-free (e.g., type IV) and hydroxyl-rich (e.g., type III) silica glasses. While the difference between the two orientations in quartz is a pure steric effect, the difference in the two silica-glass types is mainly a consequence of the different structure of the two silica-glass versions and of the interaction between Na^+ and OH^- in type III.

The diffusivity of cations in vitreous silica decreases usually with increasing charge. The ratio of the different diffusion coefficients in type-I silica glasses at 1000 °C is $D_{Na^+}:D_{Ca^{2+}}:D_{Al^{3+}}:D_{O^{2-}} = 1:2.5 \times 10^{-3}:1.3 \times 10^{-8}:1.1 \times 10^{-9}$ on the basis of an exchange of these ions with the impurities of the type-I silica glass.

A very special effect was observed with liquid and gaseous helium in contact with vitreous silica. While ^4He can be absorbed by vitreous silica at temperatures below 4.2 K, ^3He is not absorbed below 4.2 K (see Brückner, 1970/71). Thus, vitreous silica may be useful in separating ^3He from ^4He by selective absorption (isotope sieve).

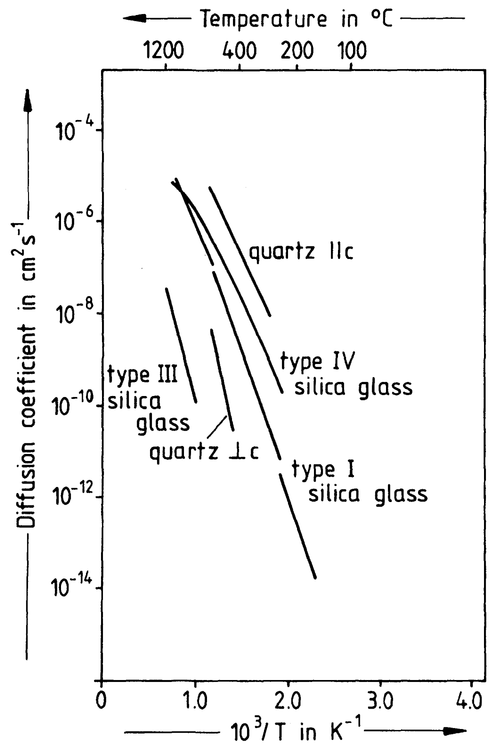

FIG. 11. Diffusivity of sodium in quartz and in various types of vitreous silica versus reciprocal absolute temperature (from Frischat, 1969, in Brückner, 1970/71).

2.3.4 Viscosity The stable and undercooled melt of silica is a high-viscosity fluid with the largest temperature range for a large viscosity range ($\Delta\eta \sim 10^{12}$–10^3 Pa s) and the highest glass transition temperature, $T_g \approx 1200$ °C, among all oxide glass compositions at all. Even at the melting temperature of cristobalite ($T_g = 1713$ °C), the viscosity is still above 10^6 Pa s. This range of temperatures is one of the reasons for the good glass formability of SiO_2 and for the long duration to reach the metastable structural equilibrium for viscosities above 10^9 Pa s (below about 1300 °C) and for various other properties such as density, refractive index, and dielectric properties. On the other hand, this viscosity-temperature range is also the reason for the fact that nearly all properties depend on thermal history, as shown in Sec. 2.1. The hydroxyl content plays an important part, too.

The influence of thermal history on viscosity is seen in Fig. 12 where the low-tem-

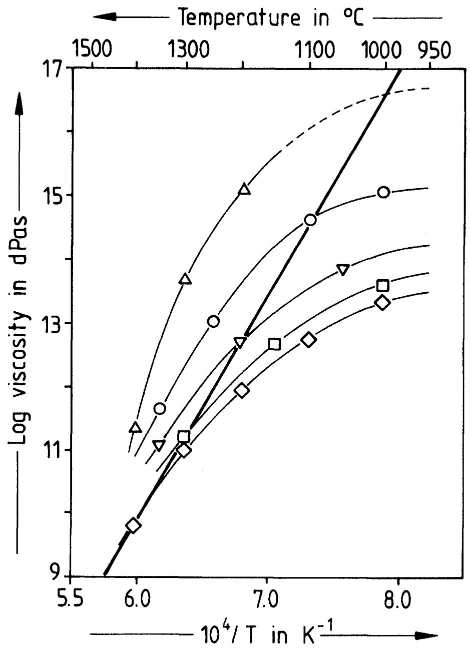

FIG. 12. Viscosity of vitreous silica, type I, versus reciprocal of absolute temperature with different thermal histories (reproduced with kind permission of The Society of Glass Technology from Hetherington *et al.*, 1964). Solid straight line is the metastable equilibrium curve; fictive temperatures: △ 1000 °C, O 1100 °C, ▽ 1200 °C, □ 1250 °C, ◇ 1400 °C.

perature viscosities of type-I silica-glass samples were measured with five different fictive temperatures between 1000 and 1400 °C. Those viscosities that have been determined from samples with fictive temperatures below the metastable equilibrium curve (the straight line in Fig. 12) are larger than the equilibrium viscosities, those that have been determined from samples with fictive temperatures beyond the metastable equilibrium curve are smaller than the equilibrium viscosities. Finally, fictive temperatures that agree with those of the metastable equilibrium curve have equal viscosities, because larger fictive temperatures are connected to a glass network structure of lower viscosities than those of the corresponding (lower) temperatures of the equilibrium curve and vice versa.

The influence of hydroxyl content is clearly seen from Fig. 13. Increasing (OH^-) content decreases viscosities by several orders of magnitude by the creation of open

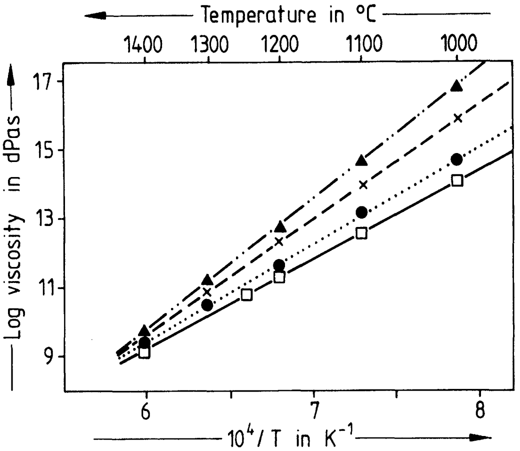

FIG. 13. Equilibrium viscosities of silica glasses with various water content in the high-viscosity temperature range (reproduced with kind permission of The Society of Glass Technology from Hetherington *et al.*, 1964). Curves from above: ▲ = type I, X = type II with 0.027 wt % OH, ● = type II with 0.04 wt % OH, □ = type III with 0.12 wt % OH.

bonds or of nonbridging oxygens in the SiO_2 network: ≡Si–OH HO–Si≡. Similar to that is the creation of open bonds by Cl^-, e.g., in type-IV silica glass: ≡Si–Cl Cl–Si≡.

The whole temperature range for the hydroxyl-poor silica glasses may be represented by the Vogel-Fulcher-Tammann equation (Brückner, 1970/71):

$$\begin{aligned}\log \eta &= A + B/(T - T_0)\\ &= -2.49 + 15\,004/(T - 253). \qquad (13)\end{aligned}$$

The activation energy of viscous flow, Q, obtained from the Boltzmann equation

$$\ln \eta = \ln \eta_0 + Q/kT, \qquad (14)$$

is 670 J/mol at high and 511 J/mol at low viscosities; thus, Q is altered only by a factor of 1.3.

2.4 Optical Properties

Silica in the form of glass and quartz is very useful as optical material because of its high transmittance of visible, ultraviolet, and near-infrared light and, in addition, for quartz because of the birefringence and rotary power for plane polarized light.

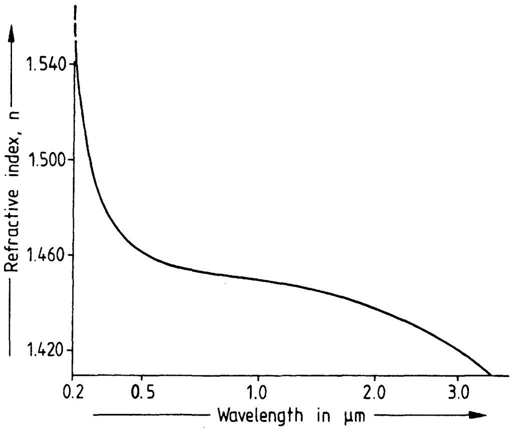

FIG. 14. Refractive index of type-II silica glass versus wavelength (Heraeus prospect).

2.4.1 Refractive Index, Reflectance, Transmittance, Luminescence, Light Scattering Silica glass is optically isotropic with a refractive index of 1.4589 at 789.3 nm wavelength. The dispersion curve is shown in Fig. 14. Like the density, the refractive index depends on thermal history and shows a maximum at a fictive temperature of about 1550 °C (Brückner, 1970/71). Increasing hydroxyl content decreases the refractive index and density. In contrast to the glass, the crystalline silica modifications are birefringent; thus, low-quartz has refractive indices of 1.5442 for the ordinary and 1.5533 for the extraordinary ray at 589.3 nm. The optical activity of quartz, due to its helical structure, corresponds to a rotary power of 21.7° at 589.3 nm for a 1-mm-thick sample. Vitreous silica becomes also birefringent under stress and has a stress-optical coefficient of 3.4×10^{-12} Pa^{-1}.

The optical transmittance of electromagnetic waves is limited at high frequencies by the ultraviolet band edge on account of interactions with the electrons of silica (electron intertransfer between oxygen and silicon or impurity cations). Increasing the electron bonding energy shifts the UV edge to higher frequencies. Therefore, the most "perfect" silica glasses, regarded as especially "pure," such as types III–VI, have the smallest absorbance between the infrared bands and the valence-band to conduction-band transitions (Fig. 15). Because the absorption edge in the UV is affected by impurities (for example, Fe^{3+} absorbs at 230 nm), and by some intrinsic defects of silica (SiO_{2-x}, $x \sim 10^{-4.5}$, type I, absorbs at 247 nm) and chemical disorder in the silica structure (see Sec. 2.5.3), absorption in the UV region may not correspond to the band-to-band transition. While nonstoichiometry causes an absorption band at 247 nm, impurities or doping of vitreous silica, e.g., with titanium or cerium oxide, can cause tailor-made shifts of the UV edge to longer wavelengths. Such shifts are important for applications, among others for special UV lamps (Witzke *et al.*, 1994).

The transmittance in the visible range is usually >99% and reduced only by reflectivity, which is relatively low because the refractive index is low, and it can be reduced by optical interference (λ/4 –) layers.

By incorporation of transition-metal or rare-earth metal oxides, colored silica can be produced (Shultz, 1974; Clasen, 1993), simi-

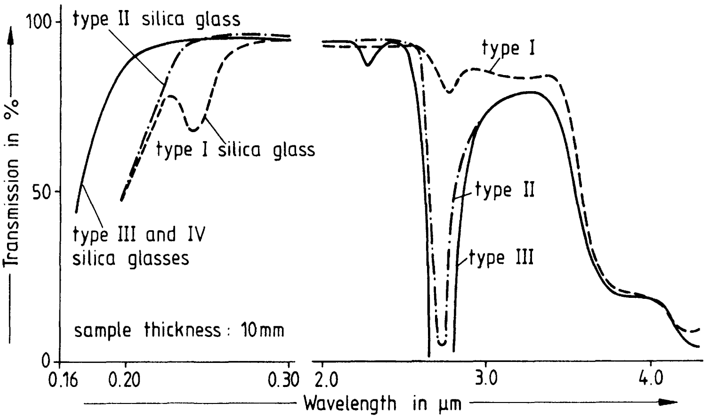

FIG. 15. Spectral transmittance of various types of silica glasses.

lar to silicate glasses, as a result of electron intratransfer within the 3*d* or 4*f* orbitals of the cations.

When natural rock quartz and silica glasses of types I and II are irradiated by x rays, γ rays, or neutrons, they become brown, smoky quartz, and their spectra have three absorption bands near 220, 300, and 550 nm due to the metal impurities (Al, Na; see Table 1) (see Sec. 2.5.2). Irradiated silica glasses of types III and IV as well as types VI and VII do not have color centers absorbing in the visible range but only a shift of the (weak) 220-nm absorption to 215 nm, which is possibly due to a radiation-induced electron transition of an oxygen vacancy into the conduction band (Weeks and Sonder, 1963). The corresponding electron transition of a nonradiated bridging oxygen is due to an absorption below 180 nm. The electronic states of the defects induced by radiation can be removed at temperatures below and around 300 °C by emission of photons of various energies (thermoluminescence bands at about 300, 500, and 600 nm) where traps are involved of which the deepest lie at 1.1 eV (Lippingcott *et al.*, 1958).

The optical transmission in the near infrared also depends on impurities; e.g., Fe^{2+} absorbs at 1.1 μm and the hydroxyl groups at 2.7 μm (Si–O–H) with overtones and combination vibrations at 2.4, 2.2, and 1.4 μm. In the far infrared the Si–O oscillations produce strong and broad absorption bands at 9.1 μm (asymmetric valency vibration), at 21 μm (transversal vibration), and at 12.5 μm in the Raman spectrum (symmetric pumping vibration). The force constant was calculated as $K_{Si-O} = 4.0$ N cm^{-1} (Zarzycki and Naudin, 1960).

Natural quartz and silica glasses of types I and II have luminescence bands at 280 nm (4.4 eV) and 396 nm (3.1 eV) when excited by UV light of wavelengths smaller than 280 nm—e.g., by Hg resonance lamps (254 nm). These luminescence bands are connected to the UV-absorption bands at about 205 nm (6 eV) and 247 nm (5.1 eV), respectively. The origin of the luminescence is not yet adequately understood, but it is assumed that both reduction of SiO_2 to SiO_{2-x} ($x \sim 10^{-4.5}$) and impurities, such as aluminum, germanium, and sodium, produce trapping sites for electrons and holes. Particularly, the luminescence band at 396 nm, connected to the absorption band at 205 nm and to alkali impurities, is assumed to be produced by charge-trapping centers of nonbridging oxygen atoms, while the luminescence band at 280 nm, connected to the absorption band at 247 nm, is produced by vacant oxygen sites (Lell *et al.*, 1966; Stroud *et al.*, 1965).

Very pure silica glasses of types III and IV do not show luminescence and probably also not types VI and VII.

The situation is changed if photoluminescence is considered. Photoluminescence is the result of photoexcitation of electrons in a sample that was previously treated by defect-center–producing or ionizing radiation. Again, the excitation (absorption) energy has to be larger than the energy of emission of the defect center because of nonradiative transitions. The emission spectra of irradiated silica glasses are complex and show bands (Friebele, 1991) at 650 nm (1.9 eV), 440 nm (2.8 eV), 290 nm (4.3 eV), and 185 nm (6.7 eV). The exact position, intensity, and half-width of the bands depend on the type of irradiation, the temperature, and the type of silica glass. Although all bands are influenced by impurities, inclusively OH^- (Wang *et al.*, 1988), they are all related to intrinsic defects. Roughly, the following attributions of these four bands may be given: the 1.9-eV luminescence has an excitation band centered near 4.8 eV and is probably associated with a nonbridging oxygen center of OH^- in high-OH^--containing silica glass, but it seems possible that it also may be several bands arising from different defect centers. The 2.8-eV emission band is common to all quartz and silica-glass types, with larger intensity for quartz than for the glasses. The 4.3-eV emission band is present only in oxygen-deficient and in neutron-irradiated silica samples. The decay of this band correlates well with the time derivative of the decay of the transient E' absorption center at 230 nm, indicating that the emission results from hole–electron pair recombination at the sites of oxygen vacancies. The 6.7-eV emission band was observed in high-purity silica glasses and quartz during 2.5-MeV electron bombardment (radioluminescence), but its origin has not been investigated intensively enough (Treadaway *et al.*, 1975).

Besides the normal luminescence, photoluminescence, and radioluminescence, thermal luminescence is of interest in order to get indications about the origin of the various emission bands observed during heating samples, during or after irradiation, or from annealing at various temperatures. Those investigations are related to optical absorption bands due to radiation and to paramagnetic centers (see Sec. 2.5.3, particularly Fig. 23). These can also be identified by annealing (thermal bleaching or annihilation) at different temperatures between 100 and 800 K (Friebele, 1991; Griscom, 1986). Thus, as an example, the intense blue luminescence during thermal bleaching results from the annihilation rate of the E' centers near 215 nm, unless connected with the restoring of bridging oxygen vacancies with absorption band near 215 nm. It is also obtained during fracture of silica glass under ultrahigh-vacuum condition (triboluminescence). Interesting is the observation that pulsed electron irradiation produces transient absorption centers near 215 nm accompanied by a blue emission with a similar time decay characteristic (see Sigel, 1973/74).

Silica glass for optical purposes has a relatively low light scattering. If, however, optical fibers are taken into consideration, the light scattering as well as relatively low absorption play an important role. This problem will be discussed in more detail in the following section.

2.4.2 Optical Waveguides Optical fibers as waveguides are widely and commercially used as transmittance medium in modern communication technologies. Several hundred times as many channels can be transmitted through one glass fiber with a diameter of 120 μm as can be transmitted through a copper coaxial tube with a diameter of 6 mm.

Silica-glass fibers have a dominant role, particularly in long-distance communication links. The fibers are either a step-index monomode or gradient-index fiber. For short-distance communication links, usually the less expensive multicomponent silicate-glass fibers are used. The production of a step-index monomode fiber starts with a thick-walled silica-glass tube whose interior surface is covered by a chemically or physically vapor-deposited layer of additives giving a higher refractive index than that of the silica glass, e.g., GeO_2 or TiO_2. This prepreg or preform is drawn and collapsed to a fiber with a core diameter of about 1–4 μm and a coating wall thickness of about 60 μm. The criteria that only one mode of a monochromatic wave can be transported is given by the relation between the difference of the two refractive indices, n_i = index of the core glass and n_0 = index of the coating glass, and the maximum diameter $D_{\max}$ of the core:

$$D_{\max} = 2.403\lambda/\pi(n_i^2 - n_0^2)^{1/2}; \qquad (15)$$

e.g., if the difference $n_i - n_0$ is 1.0%, $D_{\max}$ has to be 4 μm at a wavelength of the light, $\lambda = 850$ nm. For more details, see FIBER OPTICS.

The production of a gradient-index fiber, for which the refractive index has a parabolic profile, starts usually with a silica-glass tube with a diameter of 10–15 mm. The interior surface is covered with about 100 or more layers produced by chemical or physical vapor decomposition, each layer changing in chemical composition in such a way that a gradually increased refractive index is produced layer by layer. This prepreg is collapsed by drawing out the composite in a reduced pressure to a rodlike preform with a diameter of about 4–6 mm. Through this collapse the parabolic profile of the refractive index is preformed. The preform is then drawn to the final optical fiber with an outer diameter of about 120 μm.

The vapor deposition process is based on the deposition of the chlorides of Si, together with other chlorides such as $GeCl_4$, BCl_3, $TiCl_4$, etc., giving successively increasing refractive indices as oxides or better as silicates with increasing concentrations of the added components. The modes of the monochromatic optical wave from a laser and the refractive indices of core and coating are shown in Fig. 16 for the different types of optical fibers. In the monomode step index and in the gradient-index fibers, broadening of the light signal over long distances is small, while in the multimode step-index fiber, large broadening of the signal takes place because of different time delays of the various modes.

The most important advantages of the optical fibers based on silica glass are the high transmission or the low optical losses (less than 1 dB/km at 850 and 1300 nm) due to the low impurity concentration of the syn-

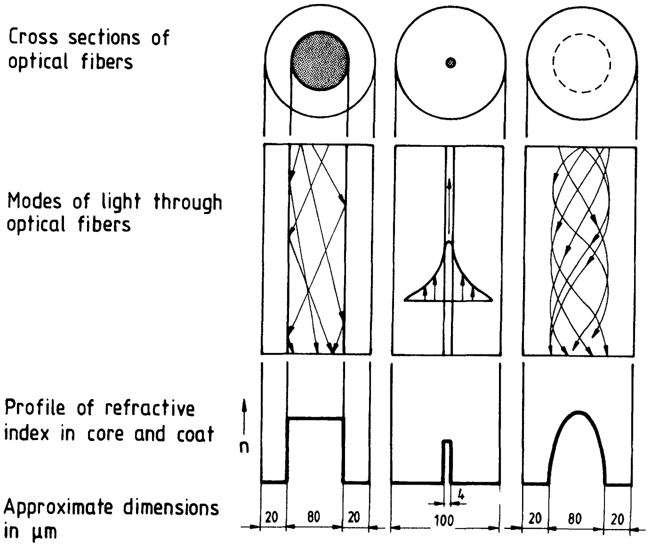
Cross sections of optical fibers
Modes of light through optical fibers
Profile of refractive index in core and coat
n
Approximate dimensions in µm
20
80
20
4
100
20
80
20

FIG. 16. Three types of optical glass fibers (schematic): above, cross section; middle, longitudinal section; and below, profile of refractive index. Step-index or multimode (left), single-mode (middle), and gradient-index fibers (right). n = refractive index; the numbers at the bottom are approximate dimensions in μm. Arrows indicate the propagation of the modes. The single-mode fiber carries along a relative large amount of intensity of the wave within the slad.

thetically produced silica glasses (types IV–VI). The limit of losses has been achieved. It is determined by the Rayleigh light scattering and phonon absorption (Fig. 17). Although this limit is more favorable for halogenide and chalcogenide glasses, a chemical purity approaching that of vitreous silica is not reached yet.

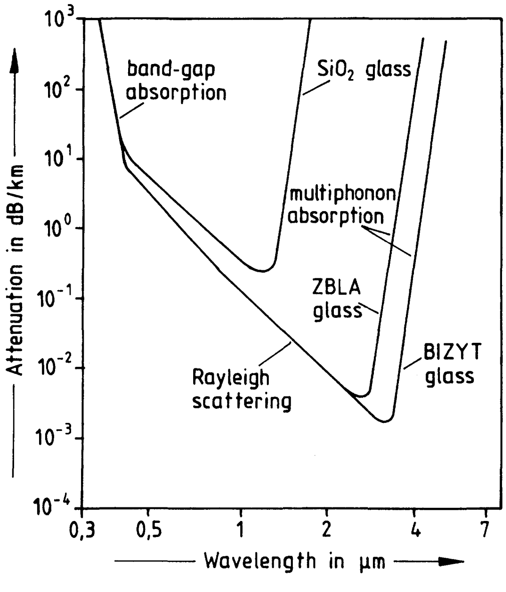

FIG. 17. The attenuation–wavelength window of minimum light losses due to Rayleigh scattering and phonon absorption of silica-glass fibers as compared with zirconium fluoride glasses (from Lucas and Adam, 1989). ZBLA = Zr–Ba–La–Al–fluoride, BIZYT = Ba–In–Zn–Y–Th–fluoride.

Another great advantage of silica-glass fibers is their large tensile strengths and large Young's modulus (see Sec. 2.2.1) as compared to multicomponent, halide, and chalcogenide glass fibers. If fibers are drawn under very clean conditions (CO_2-laser mirror furnace) and coated on line with polymers in order to prevent surface microcracks, average strengths of 5100 MPa for 4-cm proof length and about 4690 MPa for 20-m proof length (Kalish *et al.*, 1979) are measured.

A further advantage is the high chemical resistivity of the silica-glass fibers in contrast to fibers from other systems, particularly of halide and chalcogenide fibers.

2.4.3 Nonlinear Properties of Optical Silica-Glass Fibers; Optical Fiber Sensors

Optical glass fibers are not only passive or linear media for which an increase of power output increases proportionally with the input but are also active ones. In some materials, nonlinear selective power-dependent effects may occur with large optical intensities such as optical gain and frequency conversion (Stolen, 1979; Hill *et al.*, 1979). The main point is that such nonlinear effects depend on intensity and on the interaction length. On account of the low losses in small-core optical fibers and therefore of the high intensities transmitted over lengths of more than 1 km, large intensity amplification of 10^5–10^8 is possible. This enhancement decreases the threshold power for nonlinear processes that makes the effects useful in fiber-optical amplifiers, oscillators, and modulators. Thus, many optical sources used for the measurement of fiber losses and pulse dispersion may be based in future on fibers themselves.

Stimulated Raman scattering (SRS) and stimulated Brillouin scattering (SBS), which are the lowest threshold nonlinear processes in glass fibers, are observed as a result of the long interaction length available in optical glass fibers. Figure 18 shows an example of the Stokes and anti-Stokes scattered intensities for fused silica at 0 and 300 K. Also, optically induced birefringence, multiphoton effects, and two-step absorption from small impurity concentrations have been observed. Phase-matched four-photon mixing, self-phase modulation, and nonlinear absorption have been detected in silica-glass fibers. While these nonlinear effects are of interest as devices with optical glass fibers for tunable Raman oscillators and optical Kerr effect, they have to be also of particular concern for communication fibers because these processes can be detrimental.

Various fiber parameters can serve as sensors for the detection and localization of insertion losses by faults, defects, and connectors using the optical time-domain reflectometer (OTDR). These parameters are attenuation, bandwidth, local losses, mode mixing length, and, under certain conditions, resolution of scattering losses and absorption by attenuation measurements (Rourke *et al.*, 1979). The OTDR method is a nondestructive testing technique and is based on a measurement of the length dependence of optical fiber attenuation. The backscattered power from a probe pulse at a certain position in the fiber is detected by the OTDR. A pulse of light from the source is coupled into the fiber. If the propagating pulse encounters one or several absorption and backscattering sites, which may be dielectric discontinuities that result in Fresnel reflection, point defect sites, and Rayleigh scattering sites within the fiber, each site converts a portion of the forward traveling pulse into a reverse traveling wave that is directed toward the photodetector. Local alterations produced by external physical or chemical influences on the fiber properties can be detected and localized.

Many devices of optical fiber sensors are based on coherently illuminated pairs of fibers, a so-called probe fiber as the active sensor for measuring the parameters of interest such as temperature, pressure, etc., and a shielded reference fiber. Any external influence can change the phase of light in the probe by alteration of the fiber length, the core diameter, or the refractive index. This change of phase can be measured very sensitively in a two-arm fiber interferometer. A silica-glass fiber, for instance, as a pressure sensor has a pressure sensitivity of -4.1×10^{-5} rad $\mathrm{Pa}^{-1}\ \mathrm{m}^{-1}$. An interesting

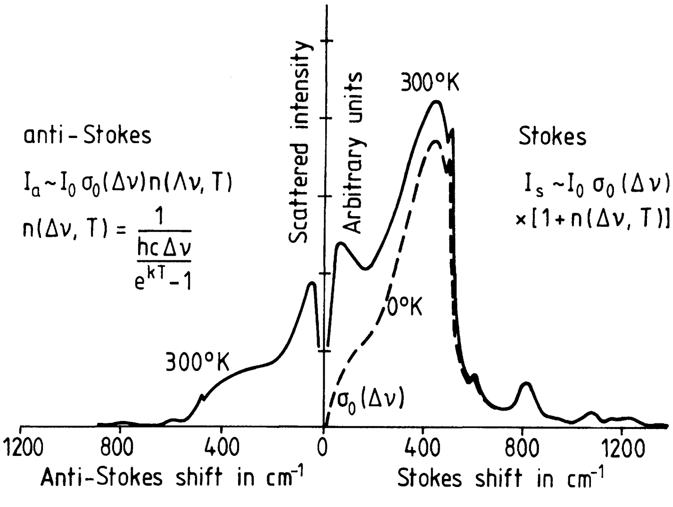

FIG. 18. Stimulated Raman scattering. Stokes and anti-Stokes scattered intensities for vitreous silica at 0 and 300 K; $n(\Delta\nu)$ is the phonon population factor, h is Planck's contant, k is Boltzmann's constant, T is absolute temperature, c is velocity of light, and $\Delta\nu$ is frequency shift in cm^{-1} (from Stolen, 1979, copyright American Telephone and Telegraph Company; used by permission).

application of a pressure-sensing fiber is the hydrophone. The gyroscope is another type of sensor based on fiber optics and on the classical Sanac "ring" interferometer in which the conventional air paths are replaced by single-mode fibers.

2.5 Electrical and Electronic Properties, Structural Defects

2.5.1 Conduction and Losses Quartz and silica glasses are among the best electrically insulating substances even at relatively high temperatures. Also, their surface electrical conductivity is low since hydrated film layers do not contain alkali ions as in the case of conventional silicate glasses.

The electrical properties of vitreous silica are usually attributed to the presence of impurities, especially of sodium ions. This fact holds for dc fields and for low-frequency ac fields. At high frequencies, other relaxation losses occur. In the case of type-III silica glass with reduced OH^- concentration (400 ppm), conductivities of the order of 10^{-12} Ω^{-1} cm^{-1} at 350 °C were measured.

The dielectric constant ϵ and dielectric loss, tan δ, show a typical dispersion and absorption in the temperature range 450–760 °C (Fig. 19; Owen and Douglas, 1959). The larger the hydroxyl content of the silica glasses, the larger are ϵ and the maximum of tan δ. The maximum shifts to higher temperatures with increasing frequency, indicating that a relaxation process is responsible for the shift. Therefore, the results in Fig. 19 are plotted in reduced scales $(\epsilon - \epsilon_\infty)/(\epsilon_S - \epsilon_\infty)$ or tan δ_{ac}/tan δ_{ac}^{max} versus the ratio of frequencies f/f_{max}, where tan δ_{ac}^{max} is the maximum of tan δ_{ac} and f_{max} is the frequency at which tan δ_{ac}^{max} occurs. Figure 19 indicates that the distribution of relaxation times of these migration losses, caused by the migration of the mobile impurity ions (Na^+, H^+, OH^-, etc.) in silica glasses, is less than in alkali silicate glasses.

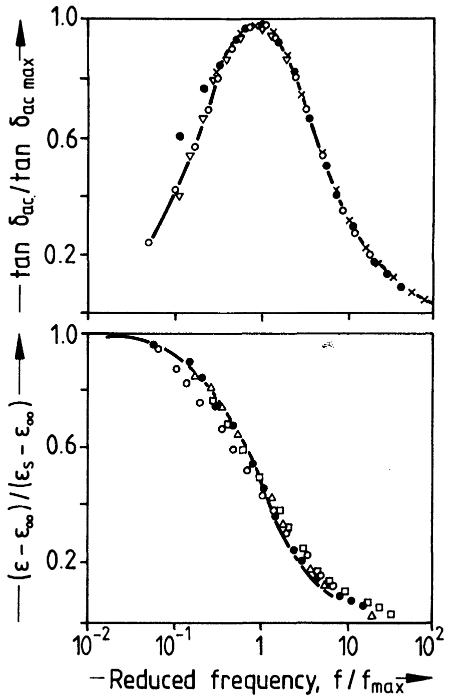

FIG. 19. Dispersion and absorption curves of silica glass versus frequency (reduced variable) in soft electrical fields due to relaxation of mobile impurities (reproduced with kind permission of The Society of Glass Technology from Owen and Douglas, 1959).

Another type of dielectric loss in vitreous silica occurs at frequencies 10^4–10^8 Hz (depending on temperature). The origin of this loss is attributed to bendings and distortions of nonbridging oxygen bonds and to deformations of the positions of bridging oxygens (see Brückner, 1970/71). Therefore, they are called deformation losses and are based on the relaxation of these deformations with a much lower relaxation time and a much lower activation energy than those of the migration losses in the low-frequency range.

The transport of electrons in vitreous silica may be measured particularly at high field strengths (e.g., Ferry, 1978). This transport is of special interest because of the properties of vitreous silica as a dielectric medium for use in metal-oxide-silicon (MOS) semiconductor devices. With respect to the transport properties, the primary attention has usually centered on the breakdown phenomena of vitreous SiO_2 at high electric fields. The drift velocity of hot electrons as a function of the applied electric field begins to saturate at fields of the order of 8 × 10^5 V/cm. Scattering due to polar optical modes of 0.153 and 0.063 eV is found and a "saturated" velocity exists at about 1.9 × 10^7 cm/s. An impact ionization is found to occur at about 10^7 V/cm.

In contrast to the high velocities of the (hot) electrons, the velocity of the holes is found to be much lower when hole–electron pairs are produced by photo- or radiation-induced transient hole conductivity. Electrons that escape initial recombination depending on the applied electric field across the vitreous silica sample are rapidly swept out, and most of the holes remain behind near their points of generation and cause negative threshold voltage shifts in a MOS device. During a period of time the holes are transported relatively slowly, depending on temperature and electric field (10^{-6} s at room temperature, 10^{-3} s at cryogenic temperature), through the wet-grown, 96-nm-thick SiO_2 layer on ⟨100⟩ Si and are removed at the interface (Si under positive gate bias). This hole transport is highly dispersive and takes place over many decades of time (Fig. 20). It is assumed that polaron hopping between randomly distributed localized sites is the microscopic charge-transfer mechanism. The average hopping distance in the direction of the electric field is found to be 0.9 ± 0.2 nm for fields in the range of 1–6 MV/cm (McLean *et al.*, 1978).

2.5.2 Electronic Structure and Properties; Density of States Another approach to properties and structure of a material is the density of states, $N(E)$, of the electronic energy of a material or system, which can be calculated for a band from the band structure. The total density of states is the sum of the states of all bands. The band structure is the energy spectrum of an electron in an ideal crystal. The description of the energy states $E_v(\mathbf{k})$ for nonmetals usually consists of filled states (the valence band) separated from empty states (the conduction band) by an energy (the band-gap energy). These band gaps are large for insulators, as is the case for SiO_2 for which it is about 10 eV. The energy bands have the symmetry of the point group of a crystal. Most important for the properties of a solid are those bands that are near the Fermi energy. In the case of large band-gap (nonmetallic) crystals, the Fermi level is usually assumed to be approximately at the middle of the gap (Kittel, 1949). Figure 21 shows the results of calculations of band structures based on the extended tight-binding method for low-temperature quartz and cristobalite. Their broad energy gaps are very similar. From these band structures the densities of state are calculated and are shown in Fig. 22 for the two crystalline SiO_2 modifications quartz and cristobalite. A handicap exists for the noncrystalline vitreous silica because of its lack of long-range-order structure. Therefore, an approximation of the Zachariasen–Warren network model has been developed that has no rings of bonds but an infinite surface: the Bethe lattice, from which the density of states has been calculated for vitreous silica [Fig. 22(c)] that is very similar to those of the crystalline modifications and relatively slightly averaged in the energy distribution (Chadi *et al.*, 1978).

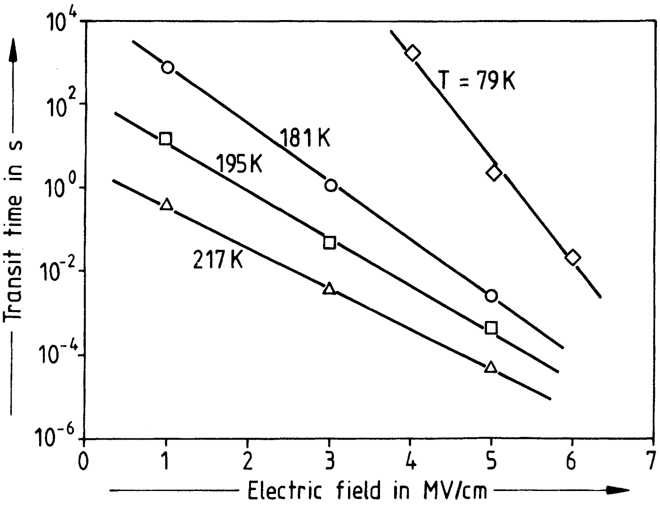

FIG. 20. Hole transit time, $t_{1/2}$, as a measure of the average inverse mobility versus electrical field in the oxide. $t_{1/2}$ is the time at which half recovery occurs (from McLean *et al.*, 1978).

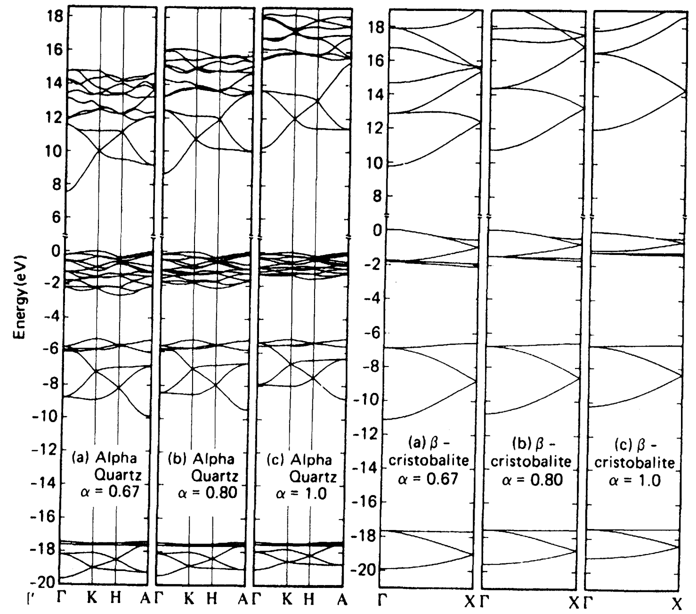

FIG. 21. Energy-band structure of (a) quartz and (b) crystobalite for different values of the exchange parameter α (from Batra, 1978).

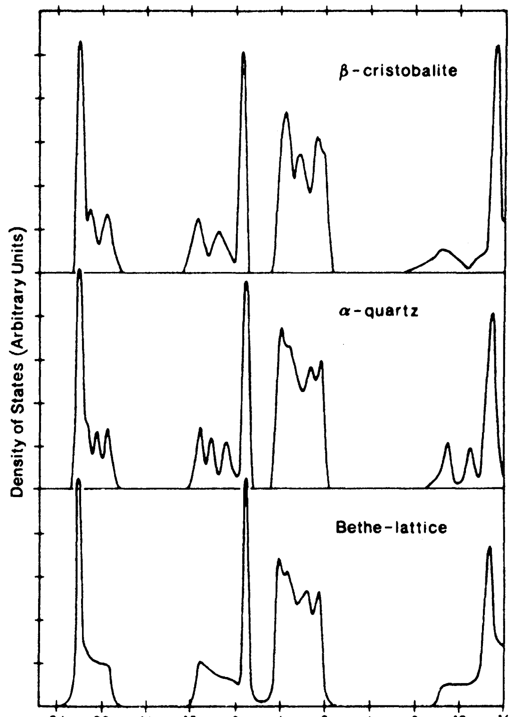

FIG. 22. Density of states for crystobalite, quartz, and vitreous silica (Bethe lattice). The valence-band maximum of quartz has been chosen for zero of energy (from Chadi *et al.*, 1978).

2.5.3 Structural Defects, Paramagnetic and Color Centers Numerous properties of quartz and silica glass are strongly influenced by the presence of defects. There are various defects localized in the structure of SiO_2. A "perfect" structure may be a continuous crystalline lattice (quartz) or a continuous network (silica glass) of $SiO_{4/2}$ tetrahedra joined together at the corners. All kinds of local deviations from a "perfect" structure will be called structural defects, including impurities, dopings, vacancies, radiation damage, and the so-called point defects. These defects may be charged or uncharged, diamagnetic or paramagnetic, and many of them may be associated with optical absorption bands. While the interpretation of optical data is connected with uncertainties, electron paramagnetic resonance (EPR) spectroscopy gives the most detailed data for interpretation. Since the application of EPR implies the presence of unpaired electrons, it is usually necessary to produce such paramagnetic centers by radiation in order to activate precursor sites prior to trapping of electrons and holes at the defect centers of interest. Various impurities such as Fe^{3+}, Cu^{2+}, and Ti^{3+} are paramagnetic.

In natural quartz there are frequently aluminum centers as impurities, which are responsible for the absorption and brown/black color when irradiated (smoky quartz). Due to the two different Si–O distances (0.160 and 0.162 nm) in quartz at exact crystallographic positions that determine the dextro- and levo-oriented arrangement of the SiO_4-tetrahedra, the exact local position of the paramagnetic center within the unit cell can be determined if angle-dependent studies of single crystals relative to the magnetic field of the EPR spectrometer are made and compared with the known point-group symmetries of the crystal. In such a way and by x-ray or γ-ray irradiation at room temperature and cryogenic temperatures, it could be shown that Al is coordinated by four oxygens forming an AlO_4-tetrahedron and that the paramagnetic center is $[\mathrm{Al},e^+/\mathrm{M}^+]^{1+}$ with the precursor $[\mathrm{Al}^-/\mathrm{M}^+]^\circ$, where the hole is trapped at the one oxygen with the small distance and the alkali ion (M^+) between the two oxygens with the longer distance to Si. Many other impurities in quartz have been studied in this way, such as Ge^{4+}, Ti^{3+}, H^0, and Fe^{3+}, the last one being partly responsible for the amethyst color (Griscom, 1978).

If pure quartz is irradiated by neutrons, paramagnetic centers of a type other than those produced by impurities occur that are defect centers termed E_1', E_2', and E_4'. They are correlated with optical bands at 5.8 eV. These paramagnetic states are due to unpaired electrons at oxygen vacancies that belong to two differently and asymmetrically relaxed silicon atoms (Griscom, 1978; Weeks, 1994). In neutron, electron, γ-ray, and x-ray irradiated silica glass, only the E_1' paramagnetic center has been observed (Weeks, 1994).

Intrinsic and extrinsic defect centers play an important role also in pure silica glass (types III and IV) and in silica-glass fibers. Three general intrinsic defect types are known to date: the E' center at a bridging oxygen vacancy ($\equiv\mathrm{Si}\cdot$, where the point denotes an unpaired spin). Thus, starting from the neutral situation of an oxygen vacancy $\equiv\mathrm{Si{-}Si}\equiv$, the creation of an $E_1' = E_\gamma'$ center by hole trapping can be expressed:

$$\equiv\mathrm{Si{-}Si}\equiv \xrightarrow{h\nu} \equiv\mathrm{Si}\cdot\ \mathrm{Si}\equiv + e^-, \tag{16}$$

with the spin at the site of the oxygen vacancy (Weeks, 1956; Griscom, 1986; Weeks and Nelson, 1960). The second paramagnetic center is nearly the same one with the only difference being that the unpaired spin is at the site away from the vacancy toward an interstitial position and a radiolytically produced hydrogen atom at the second silicon atom with a preexisting oxygen atom (Weeks, 1994):

$$\equiv\mathrm{i{-}Si}\equiv + \mathrm{H}^\circ \xrightarrow{h\nu} \equiv\mathrm{Si}\cdot\ \mathrm{H{-}Si}\equiv. \tag{17}$$

In another model the strained bond $\equiv\mathrm{Si{-}Si}\equiv$ is not attacked by a hydrogen, and the radiolytically displaced oxygen is separated far away, thus, the following reaction is assumed: $\equiv\mathrm{Si{-}Si}\equiv \xrightarrow{h\nu} \equiv\mathrm{Si}\cdot\equiv\mathrm{Si} + e^-$, where the unpaired spin is at the site away from the oxygen vacancy. In both models for E_β', the threefold-bonded silicons with no unpaired electrons, $\equiv\mathrm{Si}$, are electrically charged with e^+; the sites of the trapped electrons, e^-, are still not known. The third defect center is connected to the formation of peroxide radicals by hole trapping on peroxide linkages. The creation of it is assumed by a reaction such as (Griscom, 1984, 1986)

$$\equiv\mathrm{Si{-}O{-}O{-}Si}\equiv \xrightarrow{h\nu} \equiv\mathrm{Si{-}O{-}O}\cdot\mathrm{Si}\equiv + e^- \tag{18}$$

being a hypothetical one (Weeks, 1994).

All these kinds of defects were analyzed by EPR-hyperfine interactions of the unpaired spin with $^{29}\mathrm{Si}$ and $^{17}\mathrm{O}$ isotopes, respectively, and additionally by isochronical radiation and anneal kinetics. In this way Griscom (1986, 1984) has found that subtle but definite differences in the g values seem to correlate with the defect production and/or annealing kinetics, and he has found on this basis an additional very short-living E' center (E_α') that anneals at 100 K in the order of minutes to hours. With respect to the high OH^--containing type-III silica glasses, there exists a second oxygen-associated hole center that is related not to a bridging but to a nonbridging oxygen, and therefore it is not an intrinsic but an extrinsic defect center produced by the radiolysis of the hydroxyl group:

$$\equiv\mathrm{Si{-}OH} \xrightarrow{h\nu} \equiv\mathrm{Si{-}O}\cdot + \mathrm{H}^\circ. \tag{19}$$

Its lifetime is relatively short at 300 K as compared with those of the E_β' and E_γ' and with that of the peroxide radical (Fig. 23). The annealing mechanisms are connected to the diffusion of various species and observed in silica-glass types III and IV.

Other extrinsic defects than those based on OH^- groups are associated with Cl^-, particularly in type-IV silica glass. There are those based on dopants such as Ge and P in the core of optical fibers (Griscom, 1988). In type I and II silica glasses, the Al and Na centers as natural impurities from rock crystals are centers that are very similar to those in quartz.

2.5.4 SiO_2 Films in Microelectronic Devices This section will consider the key developments and novel aspects of technology associated with the use of silica films in solid-state silicon-based devices and circuits. Investigations of the physical and electrical properties in thin dielectric SiO_2 glass films are based on a perspective other than those in studying bulk SiO_2 glass. This other perspective is connected to the various demands on dielectric properties of the thin SiO_2 layers on Si substrates in contrast to bulk silica glass. Two examples may be given for a first impression:

1. Electrical fields across the layers are in the MV/cm range, but those across bulk samples are in the kV/cm range. This difference is due to the relatively small thickness of SiO_2 in films (10–100 nm) compared with typical bulk samples that are about a millimeter thick.
2. Dielectric losses in bulk glasses are measured at sodium concentrations in the tenth-percent range and larger, while instabilities in device properties are observed at concentrations of one part per million and less.

Therefore, any kinds of impurities, defects, or charges have consequences in the quality and reliability of the numerous MOS-based systems. With respect to the principal functions, to the large variety, and to the technology of integrated solid-state circuits, the reader is referred to the articles SI-SIO₂ INTERFACE, ELECTRONIC PROPERTIES OF and HETEROSTRUCTURES, SEMICONDUCTOR in this encyclopedia and to Sze (1983, 1991), Widmann *et al.*, (1988), and Aitken and Irene (1985). Only the preparation techniques, functions, and properties of the pure and doped SiO_2 films in contact with the heavily doped Si substrates will be described here.

The most usual preparation techniques are the direct oxidation of silicon and chemical vapor deposition (CVD). Variations are physical vapor deposition (PVC), high-pressure oxidation, and plasma oxidation.

The process of thermal oxidation is of the diffusion-reaction type and leads to a linear-parabolic equation for the thickness growth of the oxide layer with linear and parabolic rate constants, k_l and k_p, respectively:

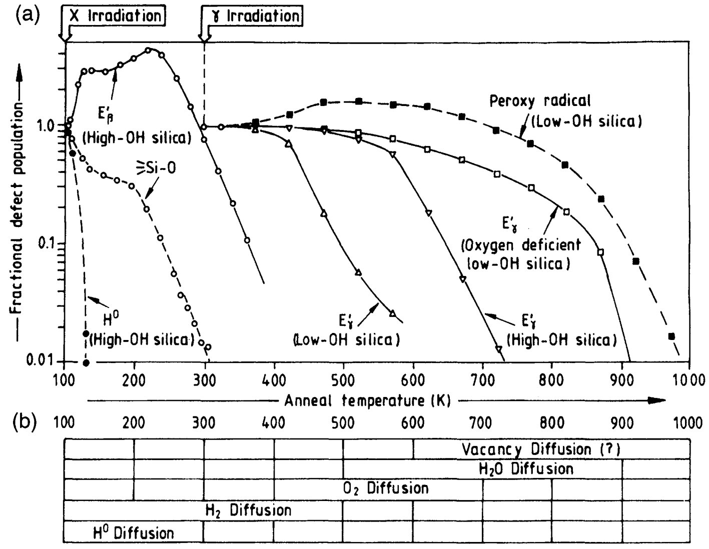

FIG. 23. (a) Isothermal anneal curves for radiation-induced defect centers in high-purity type-III (1200 ppm OH) and type-IV (<5 ppm OH) silica glasses. (b) Effective temperature ranges for various diffusion-limited anneal mechanisms, predicated on the 5–10-min anneal times employed in the experiment (with kind permission of Materials Research Society, from Griscom, 1986).

$$k_l^{-1}(l - l_0) + k_p^{-1}(l^2 - l_0^2) = t - t_0, \qquad (20)$$

where l and t are the oxide thickness and time of oxidation, respectively, and l_0 and t_0 are the offsets on the l and t axes. For oxidation in pure H_2O, l_0 and t_0 are both zero, whereas for the oxidation in dry O_2 there exists an initial oxidation regime that is not described by the linear-parabolic law but for which l_0 and t_0 are valid (see Aitken and Irene, 1985). The silicon interface is currently removed during oxidation, and the oxide layer is fully oxidized because the oxidant is moving through the oxide. The electrically active region, the Si–SiO_2 interface, is affected by the oxidation process, and any impurity present in the oxidation ambient or defects caused by oxidation will be located at this important interface. Usual temperatures for oxidation are between 780 and 1200 °C. Lower-temperature oxidation results in a higher density of SiO_2. High-pressure oxidation accelerates the oxidation rate; e.g., for 10 atm steam, a temperature of about 900 °C is comparable in oxidation rate with 1 atm steam at 1200 °C.

Both CVD and PVD are usual techniques for masking and insolation. However, CVD-SiO_2 films do not possess the same electrical quality in terms of fixed and mobile charges, charge trapping, interface states, and dielectric integrity as compared with thermally oxidized SiO_2 layers. However, there are several advantages to the CVD process. The CVD leaves the Si–SiO_2 interface intact, and high deposition rates at lower temperatures, as compared with thermal oxidation of Si, are usual, e.g., 10 nm/min SiO_2 at 750 °C with CVD for which in O_2 1200 °C is needed.

In the field of microelectronics it is well known that differently prepared films may yield considerably different electrical behavior. Therefore, a current physicochemical characterization besides that of the electrical properties is necessary. Quantitative compositional analysis is available from electron microprobe, nuclear backscattering, secondary ion mass spectroscopy (SIMS), and Auger and x-ray photoelectron spectroscopy (XPS). Scanning and transmission electron microscopy (SEM and TEM) techniques are tools for determining the morphology of SiO_2 films and Si–SiO_2 interfaces. EPR has also been used to determine interface states.

The knowledge and preparation of the SiO_2 films produced by growth, by etching, or by secondary passivation with Si_3N_4 films is very important, particularly that at the gate (base) of MOSFETs (FET: field effect transistor). The most common mechanical method uses a sensitive stylus, and optical measurements are usually made by visible-light interference or by ellipsometry.

The principal functions of SiO_2 thin films in MOSFET devices are as follows:

1. primary passivation to reduce the number of electrically active surface defects of the various doped semiconductors (Si) (defects influence the dielectric properties);
2. insulation (secondary passivation) of conducting layers from each other or to insulate individual devices on a common silicon substrate—in this case, phosphosilicate glass films act as an impurity getter, and Si_3N_4 layers act as protection against corrosion, further oxidation, and mechanical damage;
3. masking against dopants in the silicon surface;
4. diffusion source for doping the silicon interface; and
5. increase in the breakdown voltage.

Although the dielectric phenomena and explorations of dielectric constant and dielectric breakdown in thin SiO_2 glass films are similar to those for bulk glass (see Sec. 2.5.1), the passivation of the silicon surface by SiO_2 and the Si–SiO_2 interface are unique phenomena. The MOS-based devices therefore are defined and described by the dielectric and passivating properties of the thin glass films—that means, particularly, by the capacitance-voltage (*C-V*) characteristics. A MOS capacitor and its equivalent circuit [Fig. 24(a)] has usually a *C-V* curve, shown in Fig. 24(b) for low and high frequencies. In this diagram a small-signal ac capacitance is measured while a dc bias is applied to the gate. The capacitance of this device depends on the properties of the glass layer and the semiconductor. The *C-V* data contain information about the dielectric constant, film thickness, interfacial quality, and doping type and concentration of the semiconductor. For more information, see Grove (1967) and Sze (1991). In principle the capacitance of the semiconductor arises from the space-charge layers that are formed by different potentials at the Si/SiO_2 interface due to dif-

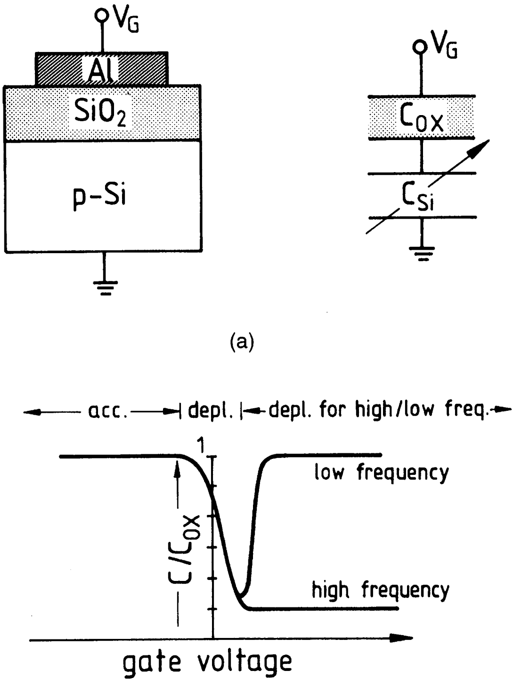

(a)

(b)

FIG. 24. *C-V* characteristic of an Al–SiO_2–Si device: (a) physical setup of a MOS capacitor and its equivalent circuit, (b) the capacitance-voltage curve measured at high and low frequencies (reproduced with kind permission of Academic Press Ltd., after Aitken and Irene, 1985).

ferent dc bias conditions. Three space-charge conditions in the *p*-type Si, marked along the voltage axis, are produced: At negatively biased voltage the majority of the holes are drawn to the interface, forming a thin dense layer that makes the interface look metallic (accumulation of the mobile majority carriers at the interface, and only the capacitance of the SiO_2 is measured). At smaller negative and positive voltages, the majority carriers are driven away from the interface (depletion range) and the finite capacitance is inversely proportional to the depletion width. This depletion width is increased, and the capacitance is decreased, as voltage increases until any additional positive bias at the gate does not increase the charge in the depletion layer but rather supplies additional minority carriers to the inversion layer. Depletion and inversion layers have quite different frequency responses because the depletion charge is a majority-carrier effect and responds rapidly to the changes in the applied field, and the inversion layer is a minority-carrier effect with a slow response to the field due to the fact that time is needed for the thermal generation of these carriers. Thus, at high frequencies (beyond 10 Hz) only the depletion layer and the steplike high-frequency curve are obtained.

From the preceding description of the MOS *C-V* properties, it becomes clear that not only the quality of the doped semiconductor but also that of the SiO_2 film and Si–SiO_2 interface are of significant importance. The quality of passivation of the thin SiO_2 films determines the quality of the intrinsic electrical properties of the Si–SiO_2 interface and its stability. Thus, any kind of electrically active defects at the interface and in the SiO_2 layer influences the dielectrical properties of the device negatively. These defects fall into two categories: interface traps (also called surface states) and charges, the latter being fixed or mobile.

With respect to the SiO_2 layers, different types of defects have to be considered and their numbers reduced in order to produce high-quality SiO_2 films and Si–SiO_2 interfaces. Since electrically active defects are traps, able to capture carriers, one has to know the various kinds of traps in SiO_2 and their concentrations. DiMario (1978) has given a review on electron and hole traps in thermal SiO_2. The most important traps are sodium-connected traps. There is first the 2.4-eV trap, which is a Coulombic attractive site for electrons in the SiO_2 band gap that reduces the well-known total band gap of about 9 eV. This trapping site is possibly related to sodium in an immobile configuration, particularly produced in steam-grown SiO_2. In contrast to that there are usually mobile traps that are connected to Na^+ ions that are immobilized at 77 K near the Si–SiO_2 interface. In as-grown SiO_2, traps were observed caused by water incorporation into the film during deposition. Trapped hole sites near the Si–SiO_2 interface that trap electrons into energetically deep sites can be injected into SiO_2 films or may be generated during processing conditions (growth, temperature, and annealing). There are hole traps in an as-grown silica film that are introduced into the SiO_2 valence band through excitation by far UV light, x rays, or high electric fields and injections from the Si substrate across the SiO_2 band gap. Deep hole traps near the Si–SiO_2 interface capture holes permanently at room temperature. Hole traps are sensitive to processing conditions and can be increased by annealing above 950 °C. It is believed that they are caused by impurities introduced into the SiO_2 glass and/or by strained or broken bonds due to stresses at the interface. Even the flow of current across the Si–SiO_2 interface can create new traps (interface traps). This creation is observed when capacitors have been stressed by hot-electron injections, by exposure to ionizing radiation, and by high electric fields.

Besides the just-described defects, there is a series of intentionally produced traps with differing properties produced by introduction of tungsten (W traps), by x-ray radiation, and by ion implantation of Al, As, and P, for example. Usually the last ones produce such extensive damage that any carrier injected into the SiO_2 layer is trapped. For more information about these special kinds of traps, see DiMario (1978).

2.6 Applications of Vitreous Silica

The fields of application of vitreous silica are determined by the properties of the various types, which again depend on quality, i.e., on the kinds and concentrations of impurities (see Table 1). Therefore, it will be suitable to list the fields of application of the seven types of vitreous silica, which naturally have relatively large regions of overlap.

Type I is used for IR optics in optical applications. The priority of applications is tubing for lamps of different kinds, particularly for IR lamps that are used for drying applications in the field of coatings. Generally, it is used for nonoptical applications, e.g., for semiconductor-processing equipment, and as a material for structural elements.

Type II is suitable for general and precision optics, for tubing elements of optical fiber preforms, for tubing elements of high-quality discharge lamps, and for semiconductor-processing equipment.

Types III and IIIa were developed for a series of modern applications such as precision (UV) optics, high-power laser optics, mask substrates of optical microlithography, substrates for optoelectronics, and, last but not least, core material for speciality fiber preforms, e.g., PCS.

Type IIIb is useful for optical telecommunication fibers as preforms with doped and undoped core and as material for cladding and jacketing.

Type IV is a nearly hydroxyl-free vitreous silica with highest purity and serves as standard material for scientific research and is used in precision optics and in fluorine-doped cladding for speciality fiber preforms.

Type Va is a precursor for type Vb silica glass.

Type Vb has applications in semiconductor-processing equipment and as special lamp tubing, e.g., metal-halide, mercury, and other discharge lamps as well as metal-halide incandescent lamps.

Type VI is fabricated on a pilot scale so far and not yet on an industrial scale. It appears useful for experimental studies on purification and doping of the porous precursor material, on sintering, and on preparation techniques on the basis of colloidal SiO_2 particles as raw material.

Type VII is a material prepared on a laboratory scale and is a typical laboratory material. It is based on the sol-gel technique (silicon alkoxide); it has mainly experimental applications in production of coatings and thin films on substrates.

Additional applications due to typical properties of vitreous silica and quartz are given in Secs. 1.3.1 and 1.3.2. Special applications are less connected to the typical properties but rather to the material SiO_2 as a part of a system, e.g., as optical waveguides, sensors, etc., as pointed out in Secs. 2.4.2 and 2.4.3.

Not included in the seven types of vitreous silica are the thermally produced thin films of SiO_2 on metallic Si surfaces by oxidation, particularly as a passivation element in semiconducting devices (Sec. 2.5.4). The reason is that the quality of these thin layers depends strongly on preparation conditions such as furnace atmosphere, impurities, and dopings of the Si substrate. In this way this group of silica glasses is not a class in the sense of the other types, and, additionally, it is strongly connected to the metallic Si substrate with special preparation-dependent properties of the interface Si/SiO_2.

3. SILICA GEL, AEROGELS

During the last decade, a pure SiO_2 product with remarkable properties due to its high porosity has been given attention. Silica gels with densities between 0.3 and 0.005 g cm^{-3}, called aerogels because they contain more air than amorphous SiO_2, have been prepared. There are usually two routes of production. In the first route, water glass (sodium silicate) plus HCl is mixed. After gelation the sodium is washed several times with water until a pure SiO_2 skeleton is obtained. The water in this aqua gel has to be washed out by methanol and is dried under supercritical conditions (above 240 °C and above a pressure of 80 bars) in order to prevent shrinking and the collapse of the porous gel by capillary forces. The second route starts from a mixture of tetramethoxysilane, $Si(OCH_3)_4$, and water plus an acid as catalyzer. After formation of hydroxyl, $Si(OH)_4$, and release of CH_3OH and separation of water, Si–O–Si bondings are formed into a very fine wet network of SiO_2 particles (gel). This product has to be dried supercritically.

By virtue of the high porosity and small pore sizes of 1–100 nm (comparable with the mean-free-path length of the air molecules of 70 nm!), these gels have excellent heat isolation for windows and solar energy collectors (Fricke, 1988). Heat conductivities are obtained between 0.005 and 0.010 W m^{-1} K^{-1} (for comparison, styropore has conductivity of 0.026). Aerogels are used as Cherenkov sensors in high-energy physics in order to record relativistic protons, muons, and pions. On account of their nanometer structure, they have very low ultrasonic velocities (about 100 m s^{-1}, which is 50 times smaller than that for bulk silica glass). At a density of 0.005 g cm^{-3} the velocity is 20 m s^{-1}. Such velocities provide material for acoustic antireflection layers because the product of the small densities and sonic velocities has a low acoustic impedance. Therefore, a high acoustic intensity (by a factor of 1000) is obtained in air if a thin layer of aerogel is glued on to a usual ultrasonic piezo source because the impedance in air is also very low. Applications are seen for measuring distances to the rear of cars or in roboter systems, for instance.

4. SUMMARY

In this article silicon dioxide is shown to be a chemically simple composition, but a very complicated one with respect to its crystallographic modifications and properties, especially quartz, cristobalite, and fused silica. Particularly the noncrystalline version of it, the vitreous (glassy) silica or silica glass, is of great interest in science and of great importance in practical applications due to its excellent physical, chemical, and thermal properties and due to its good workability. It has applications in numerous devices, in electronics, in microelectronics, in optics, in sensors, in the modern communication field (optical fibers), for the industrial fabrication of special lamps, and in semiconductors as melting crucibles with highest purity. The survey includes the crystalline, vitreous, and amorphous states of silicon dioxide; its mechanical, thermal, optical, electrical, and electronic properties; and also the various anomalies of vitreous silica as compared with the properties of the large family of the multicomponent silicate glasses. Additionally, some discussion is given to silica gels and aerogels.

ACKNOWLEDGMENTS

The author is grateful to Dr. Englisch and Dr. Witzke, Heraues Quarzschmelze, for fruitful discussions and supplementations and to Prof. Dr. R. Weeks for linguistic and complementary revision of the manuscript.

GLOSSARY

Aerogel: High-porous silica gel, amorphous SiO_2.

Cristobalite: Modification of SiO_2, stable above 1470 °C.

Extrinsic Defects: Defects produced by impurities or nonstoichiometric local composition.

Intrinsic Defects: Structural defects.

Keatite, Coesite, Stishovite: High-pressure, high-temperature modifications of SiO_2.

Quartz: Most frequent and thermodynamically stable modification of SiO_2 on Earth.

Silica Gel: Amorphous SiO_2.

Silica W: Fibrous SiO_2, stable above 1000 °C in water-free atmosphere.

Silicon Dioxide, Silica: Chemical compound SiO_2.

Thermal SiO_2: Glassy SiO_2 films grown on Si surfaces at elevated temperatures (passivation films).

Trapping Sites: Electron or hole traps.

Tridymite: Modification of SiO_2, stabilized by alkali oxide.

Vitreous Silica, Fused Silica, Silica Glass: Glassy modification of SiO_2.

Works Cited

Aitken, J. M., Irene, E. A. (1985), in: M. Tomozawa, R. H. Doremus (Eds.), *Treatise on Material Science and Technology,* Vol. 26, Orlando: Academic, pp. 1–56.

Batra, J. P. (1978), in: S. T. Pantelides (Ed.) *The Physics of SiO_2 and Its Interfaces,* New York: Pergamon, pp. 65–69.

Brill, R. W., Hermann, G., Peters, C. (1939), *Naturwissenschaften* **40,** 676–677.

Brill, R. W., Herman, G., Peters, C. (1942), *Ann. Phys. (Leipzig)* **41,** 233–244.

Brückner, R. (1970/71), *J. Non-Cryst. Solids* **5,** 123–216.

Chadi, D. J., Laughlin, R. B., Joannopoulos, J. D. (1978), in: S. T. Pantelides (Ed.), *The Physics of SiO_2 and Its Interfaces,* New York: Pergamon, pp. 55–59.

Clasen, R. (1987), *Glastechn. Ber.* **60,** 125–132.

Clasen, R. (1993), *Glastechn. Ber.* **66,** 299–304.

Cohen, A. F. (1958), *J. Appl. Phys.* **29,** 591–598.

DiMario, J. G. (1978), in: S. T. Pantelides (Ed.), *The Physics of SiO_2 and Its Interfaces,* New York: Pergamon, pp. 160–178.

Evans, K. M., Gaskell, P. H., Nex, C. M. M. (1982), in: P. H. Gaskell, I. M. Parker, E. A. Davis (Eds.), *The Structure of Non-Crystalline Materials,* Proceedings of the Second International Conference, London: Taylor & Francis.

Fanderlik, J. (1991), *Silica Glass and Its Application,* Amsterdam: Elsevier.

Ferry, D. K. (1978), in: S. T. Pantelides (Ed.), *The Physics of SiO_2 and Its Interfaces,* New York: Pergamon, pp. 29–34.

Flörke, O. W. (1956a), *Naturwissenschaften* **43,** 419–420.

Flörke, O. W. (1956b), *Ber. Dtsch. Keram. Ges.* **33,** 319–321.

Fricke, J. (1988), *Sci. Am.* **258** (5), 92–97.

Friebele, E. J. (1991) in: D. R. Uhlmann, N. J. Kreidl (Eds.), *Optical Properties of Glass,* Westerville, OH: American Ceramic Society, pp. 205–262.

Frischat, G. H. (1969), *Phys. Status Solidi* **35,** K47–K49; *Glastechn. Ber.* **42,** 351–359.

Griscom, D. L. (1978), in: S. T. Pantelides (Ed.) *The Physics of SiO_2 and Its Interfaces,* New York: Pergamon, pp. 232–252.

Griscom, D. L. (1984), *Nucl. Instrum. Methods* **B1,** 481–488.

Griscom, D. L. (1986), in: F. L. Galeener, D. L. Griscom, M. J. Weber (Eds.), *Defects in Glasses,* Pittsburgh: Materials Research Society, pp. 213–221.

Griscom, D. L. (1988), in: R.A.B. Devine (Ed.), *The Physics and Technology of Amorphous SiO_2,* New York: Plenum Press, pp. 125–134.

Grove, A. S. (1967), *Physics and Technology of Semiconductor Devices,* New York: Wiley.

Hench, L. L. (1993), in: A. K. Varshneya, D. F. Bickford, P. P. Bihuniak (Eds.), *Advances in Fusion and Processing of Glass, Ceram. Trans. Vol. 29,* Westerville, OH: American Ceramic Society, pp. 591–597.

Hetherington, G., Jack, K. H., Kennedy, J. C. (1964), *Phys. Chem. Glasses* **5,** 130–136.

Hill, K. O., Kawasaki, B. S., Johnson, D. C., Fujii, Y. (1979), in: B. Bendow, S. S. Mitra (Eds.), *Fibre Optics,* New York: Plenum Press, pp. 211–240.

Hosemann, R., Hentschel, M. P., Schmeißer, U., Brückner, R. (1986), *J. Non-Cryst. Solids* **83,** 223–234.

Hunklinger, S. (1982), *J. Phys. (Paris)* **C9,** suppl. au no. 12, Tome 43, 461–474.

Hunklinger, S., Schickfus, M. von (1981), in: W. A. Phillips (Ed.), *Amorphous Solids,* Topics in Current Physics Vol. 24, Berlin: Springer-Verlag, pp. 81–104.

Jebsen-Marwedel, H., Brückner, R. (1980), *Glastechnische Fabrikationsfehler,* Berlin: Springer-Verlag, pp. 358–384.

Kalish, D., Key, L. L., Kurkjian, C. R., Tariyal, B. K., Wang, T. T. (1979), in: S. E. Miller, A. G. Chynoweth (Eds.), *Optical Fibre Telecommunications,* New York: Academic, pp. 401–433.

Kittel, C. (1949), *Introduction to Solid State Physics,* New York: Wiley.

Konnert, J. H., D'Antonio, P., Karle, J. (1982), *J. Non-Cryst. Solids* **53,** 135–141.

Lamkin, M. A., Riley, F. L., Fordham, R. J. (1992), *J. Europ. Ceram. Soc.* **10,** 347–367.

Lee, R. W. (1964), *Phys. Chem. Glasses* **5,** 35–42.

Lebedev, A. A. (1921), *Arb. Staatl. Opt. Inst., Leningrad* **2** (10).

Lebedev, A. A. (1940), *Bull. Acad. Sci. USSR, Phys. Ser.* **4,** 584–587.

Lell, E., Kreidl, N. J., Hensler, J. R. (1966), *Prog. Ceram. Sci.* **4,** 1–38.

Lippingcott, E. R., van Valkenburg, A., Weir, Ch. E., Bunting, E. N. (1958), *J. Res. Natl. Bur. Std.,* 61–72.

McLean, F. B., Boesch, H. E., McGarrity, J. M. (1978), in: S.T. Pantelides (Ed.), *The Physics of SiO_2 and Its Interfaces,* New York: Pergamon, pp. 19–23.

Lucas, J., Adam, J.-L. (1989), *Glastechn. Ber.* **62,** 422–440.

Ogawa, K., Morikawa, T., Imamura, H., Akai, Y. (1977), *J. Light Vis. Env.* **1,** 16–21.

Owen, A. E., Douglas, R. W. (1959), *J. Soc. Glass Technol.* **43,** 159–165.

Primak, W. (1958), *Phys. Rev.* **110,** 1240–1251.

Porai-Koshits, E. A. (1977), *J. Non-Cryst. Solids* **25,** 87–128.

Rourke, M. D., Jensen, S. M., Barnoski, M. K. (1979), in: B. Bendow, S.S. Mitra (Eds.), *Fibre Optics*, New York: Plenum Press, pp. 255–268.

Schaeffer, H. (1984), *J. Non-Cryst. Solids* **67,** 19–33.

Schultz, P. C. (1974), *J. Am. Ceram. Soc.* **57,** 309–313.

Sigel, G. H. (1973/74), *J. Non-Cryst. Solids* **13,** 372–398.

Sosman, R. B. (1965), *The Phases of Silica*, New Brunswick, NJ: Rutgers Univ. Press.

Stolen, R. H. (1979), in: S. E. Miller, A. G. Chynoweth (Eds.), *Optical Fibre Telecommunications*, New York: Academic, pp. 125–150.

Stone, J. (1987), *J. Lightwave Technol.* **5,** 712–733.

Stroud, J. S., Schreures, W. H., Tucker, R. F. (1965), *Proceedings of the 7th International Congress on Glass*, Bruxelles, 42.1–42.18.

Tool, A. Q. (1946), *J. Am. Ceram. Soc.* **29,** 240–253.

Treadaway, M. J., Passenheim, B. C., Kitterer, B. D. (1975), *IEEE Trans. Nucl. Sci.* **NS-22,** 2253–2258.

Valenkov, N., Porai-Koshits, E. A. (1936), *Nature* **137,** 273–274.

Wang, P. W., Haglund, R. F., *et al.* (1988), *J. Non-Cryst. Solids* **102,** 288–294.

Warren, B. E., Krutter, H., Morningstar, O. (1936), *J. Am. Ceram. Soc.* **19,** 202–206.

Warren, B. E. (1941), *J. Am. Ceram. Soc.* **24,** 256–261.

Weeks, R. A. (1956), *J. Appl. Phys.* **27,** 1376–1381.

Weeks, R. A. (1994), *J. Non-Cryst. Solids* **179,** 1–9.

Weeks, R. A., Nelson, C. (1960), *J. Am. Ceram. Soc.* **43,** 399–402.

Weeks, R. A., Sonder, E. (1963), in: *LOW Symposium on Paramagnetic Resonance*, Vol. 2, New York: Academic, pp. 869–879.

Witzke, H.-D. (1990), *Glastechn. Ber.* **63K,** 333–341.

Witzke, H.-D., Takke, R. St. Thomas, Englisch, W. (1994), *Glastechn. Ber., Glass Sci. Technol.* **67c,** 346–349.

Wong, J., Angell, C. A. (1976), *Glass Structure by Spectroscopy*, New York: Marcel Dekker.

Wright, A. C., Leadbetter, A. J. (1976), *Phys. Chem. Glasses* **17,** 122–145.

Wright, A. C., *et al.* (1982), *J. Non-Cryst. Solids* **49,** 63–102.

Zachariasen, W. H. (1932), *J. Am. Chem. Soc.* **54,** 3841–3851.

Zarzycki, J. (1982), *J. Non-Cryst. Solids* **52,** 31–43.

Zarzycki, J., Naudin, F. (1960), *Verres Réfractaires* **14,** 113–123.

Further Reading

Fricke, J., Emmerling, A. (1992), in: R. Reisfeld, C. K. Jörgensen (Eds.), *Structure and Bonding*, Vol. 77, Berlin: Springer-Verlag, pp. 37–87.

Phillips, W. A. (Ed.) (1981), *Amorphous Solids, Low Temperature Properties*, Topics in Current Physics Vol. 24, Berlin: Springer-Verlag.

Sze, S. M. (Ed.) (1983), *VLSJ Technology*, New York: McGraw-Hill.

Sze, S. M. (Ed.) (1991), *Physics of Semiconductor Devices*, New York: Wiley.

Walrafen, G. E., Revesz, A. G. (Eds.) (1986), *Structure and Bonding in Noncrystalline Solids*, New York: Plenum Press.

Widmann, D., Mader, H., Friedrich, H. (1988), *Technologie hochintegrierter Schaltungen, Halbleiterelektronik*, Vol. 19, Berlin: Springer-Verlag.

SILICON NITRIDE

F. L. RILEY, *School of Materials, University of Leeds, Leeds, United Kingdom*

Introduction 133
1. **Crystal Chemistry** 134
1.1 Chemical Bonding 134
1.2 Crystal Structures 135
1.3 Amorphous Silicon Nitride 135
2. **Production Methods** 137
2.1 General 137
2.2 Powders 137
2.2.1 Silicon Nitridation 138
2.2.2 Carbothermal Nitridation of Silicon Dioxide 138
2.2.3 Diimide Decomposition 138
2.3 Bulk Forms 138
2.3.1 Reaction-Bonded Silicon Nitride 138
2.3.2 Sintered Silicon Nitride 139
2.3.3 Hot-Pressed Silicon Nitride 140
2.4 Thin Films 140
3. **Properties of Materials** 142
3.1 Physical 142
3.1.1 Density 142
3.1.2 Diffusivity 142
3.1.3 Electrical 142
3.1.4 Optical 143
3.1.5 Thermal 143
3.2 Mechanical 143
3.2.1 Creep 143
3.2.2 Fracture Toughness 143
3.2.3 Hardness 144
3.2.4 Strength 144
3.2.5 Young's Modulus 144
3.3 Chemical 144
3.3.1 High-Temperature Oxidation and Corrosion 145
3.3.2 Molten-Metal Attack 145
4. **Applications** 145
4.1 Thin Film 145
4.1.1 Masking Layers for Silicon Local Oxidation 146
4.1.2 Diffusion Barriers 146
4.1.3 Silicon Nitride Dielectrics 146
4.2 Bulk Ceramic 147
4.2.1 Cutting-Tool Inserts 147
4.2.2 Internal-Combustion Engines 147
4.2.3 Bearings 148
4.2.4 Industrial 148
5. **Summary** 148
Works Cited 149
Further Reading 149

INTRODUCTION

The element silicon forms only one nitride that is stable at normal temperatures. This is a solid and has two crystalline phases, designated α and β. The stoichiometric composition is Si_3N_4, but departures from this stoichiometry can occur in the amorphous, noncrystalline form. Silicon nitride does not occur naturally in the oxidizing environment of the earth's crust (though trace occurrence in minerals of meteoritic origin has been reported); it can thus be regarded as a man-made material.

Although silicon nitride has been known to chemists for over a century, its potential as a polycrystalline ceramic material only became apparent in the late 1950s to early 1960s, when considerable development work was initiated on the production and characterization of nonoxide compounds of low molecular weight and strong interatomic bonding, and thus likely to have useful electrical and structural properties. The possibility of using silicon nitride as a component to withstand temperatures of 1300 °C and higher in a new range of high-efficiency internal-combustion engines was quickly appreciated and was to lead to intensive efforts over the next 25 years to develop high-strength silicon nitride ceramics for a wide range of applications at room and high temperature. Much of this work has been reported in conference proceedings, of which Carlsson *et al.* (1992) and Hoffmann *et al.* (1994) are good examples.

This development work in the ceramic area, involving the production of increasingly

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

complex chemical and crystallographic systems, has led to the realization that silicon nitride, strictly, should be regarded as a very large class of materials. Some members of the class are relatively simple [pure chemical vapor deposition (CVD) silicon nitride, for example]; others (those prepared with oxide-sintering additives) can be complex solid solutions involving other metallic elements and oxygen. Well-recognized examples of this type are the "sialon" group of materials, based on two- or more-component subsystems. Full discussion of these complex systems is beyond the scope of a short overview article on silicon nitride, which will be restricted to single-component materials and, with only brief excursions, where essential, into the field of silicon nitride–metal oxide systems. A detailed overview of the sialon materials has been provided by Ekstrøm and Nygren (1992).

Simultaneously, though on a smaller scale and less visibly, the uses of silicon nitride have been explored in a range of applications in electronics. This work led in particular to improved understanding of the production and electrical and electronic properties of amorphous silicon nitride, usually in thin (~100 nm)-film form. The use of silicon nitride in microelectronic devices was first reported in 1964, when it was discovered that thin films, because of their dense structures, had better masking properties than silicon dioxide (silica) against metals and, in particular, against the diffusion of sodium. It was also established that the films had the capability of remembering and storing for long times electrons and holes, and thus acting as "read only" memory (ROM) devices. Subsequently the effectiveness of silicon nitride as a mask against oxygen and water vapor at temperatures up to ~1000 °C allowed the development of the local silicon oxidation technique; silicon nitride thin films now play a key role in silicon- and gallium-arsenide–based microelectronic technology. Because of their mechanical hardness and radiation resistance, silicon nitride thin films are also used as final protection layers to finished devices. Silicon nitride is now being considered as an alternative or additional gate or node dielectric in advanced metal-oxide-semiconductor (MOS) devices such as random-access memory arrays. A full treatment of the application of silicon nitride in the electronics area is given by Belyi *et al.* (1988).

The purpose of this article is to introduce silicon nitride and its properties and to indicate the nature of the large volume of work dedicated to realizing its considerable potential as a structural material on a very wide range of scale. At the present time there is no single-authored textbook dedicated to silicon nitride; important volumes of other types dealing wholly or in part with silicon nitride are listed under the "Works Cited" and "Further Reading" headings. The Further Reading section gives selected volumes the contributions to which are also well referenced; both sets provide an introductory guide to the very extensive literature (over 25 000 publications) dealing wholly or in part with silicon nitride materials. Numerical data quoted in this review are taken from these volumes. A comprehensive survey and compilation of factual and numerical data on silicon nitride is in the process of being constructed (Gmelin Institut für Anorganische Chemie, 1991, 1995, 1996).

1. CRYSTAL CHEMISTRY

This section deals with the structural arrangement of atoms within the silicon nitride crystal, and the differences between the two crystalline phases, α- and β-Si_3N_4.

1.1 Chemical Bonding

The silicon–nitrogen single bond is strongly covalent; estimates of ionicity are in the range 11–30%. This entails considerable localization of the electrons and directed bonding. Silicon forms four sp^3 hybrid bonds directed to the corners of a tetrahedron with 109.5° angles, and nitrogen, three trigonal sp^2 hybrids at 120° angles in a plane. The basic structural units can thus be regarded as SiN_4 tetrahedra, joined together in a three-dimensional network by sharing (N) corners; each N is thus common to three tetrahedra. Figure 1 shows schematically the arrangement of the SiN_4 tetrahedra in the puckered sheets that form the basis for the α and β structures of Si_3N_4. The marked tetrahedra are at height $\frac{1}{2}$; the basis for the unit cell is also shaded. Using normal covalent radii for Si and N, the Si–N bond length can be cal-

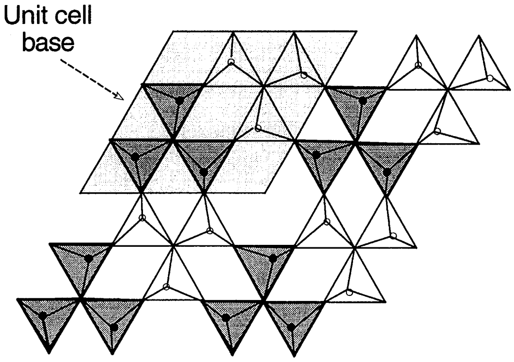

FIG. 1. Schematic illustration of the arrangement of the SiN_4 tetrahedra in the puckered sheets making up the Si_3N_4 structure. The shaded tetrahedra are at a height of $\frac{1}{2}$; the basis for the unit cell is also shaded (Courtesy D. P. Thompson).

culated to be 187 pm, which is much larger than those seen experimentally in silicon nitride (170–177 pm). It is assumed that the Si 3d orbitals make a significant contribution to p^{π}-d^{π} interactions. The mean Si–N bond energy is calculated to be 330 kJ mol^{-1}.

1.2 Crystal Structures

The SiN_4 tetrahedra link to form puckered six-membered rings, which in turn delineate large (and normally empty) voids; these linked rings can be regarded as "AB" and "CD" layers, in which the CD layer is an inverted AB, with protruding Si and N bonds. In the idealized silicon nitride structures, the layers can be stacked in two ways: AB layers stacked vertically on top of each other give the β structure, . . .ABAB. . . (space group $P6_3$); stacking of an AB layer alternately with an offset CD layer obtained by the operation of a c-glide plane gives the α structure, . . .ABCDABCD. . . (designated space group $P31_c$). In α-Si_3N_4 the bottom half of the unit cell is the same as in β-Si_3N_4, and the unit-cell dimension in the c axis is approximately twice that of the β-Si_3N_4 unit cell. The unit cells of α-Si_3N_4 and β-Si_3N_4 are shown schematically in Figs. 2 and 3. These arrangements lead to two rather open, and low-density, crystal structures that, while they contain marked differences, are also in many respects closely similar. The hexagonal β unit cell containing Si_6N_8 has continuous channel-like voids, aligned parallel to the c axis: in the α structure these are replaced by large closed cages. The β structure is almost strain free and has an average Si–N bond length of 173.6 pm. In contrast, the α structure in departing from the idealized structure has a marked degree of strain and a slightly lower single-crystal density (α-Si_3N_4 ~ 3.167–3.171, β-Si_3N_4 ~ 3.192 Mg m^{-3}). For this reason it is generally assumed that the β structure is thermodynamically the more stable. However, practical experience is that under clean conditions and particularly at relatively low temperatures (<1300 °C), the α phase is kinetically the more readily formed, both as a result of gas-phase reactions and by the lower-temperature recrystallization of amorphous silicon nitride.

These observations would be in accordance with the view that the α phase is the low-temperature modification, and the β phase is the high-temperature modification. The α- to β-phase reconstructive transformation of secondary coordination occurs fairly readily at high temperature (>1600 °C), particularly in the presence of a liquid solvent; the reverse process has, however, never been reliably observed.

There has been considerable controversy over the nature of the α-Si_3N_4 phase, particularly with regard to whether it should be considered to be the nitride low-temperature phase or, alternatively, as an oxynitride of low oxygen content and fairly precisely defined stoichiometry. This view translates to that of the β phase being formed under "low" oxygen potentials ($\leq\sim10^{-20}$ bar) and the α phase under "high" oxygen potentials. Uncertainty was sustained by observations that α-Si_3N_4 could be shown to contain significant amounts of structural oxygen and had a wider range of experimental unit-cell dimensions (a = 774.91–775.72 pm; c = 561.64–562.21 pm), and thus densities, than the β phase (a = 760.3 pm; c = 290.6 pm). The generally accepted current view is that both are true nitrides but that the strained α structure is more readily able to accommodate the small amounts of oxygen inevitably present in formation environments; oxygen seems not to be necessary for the stability of the structure.

1.3 Amorphous Silicon Nitride

The use of chemical vapor deposition techniques at relatively low temperatures (<1000 °C) to form thin films of silicon ni-

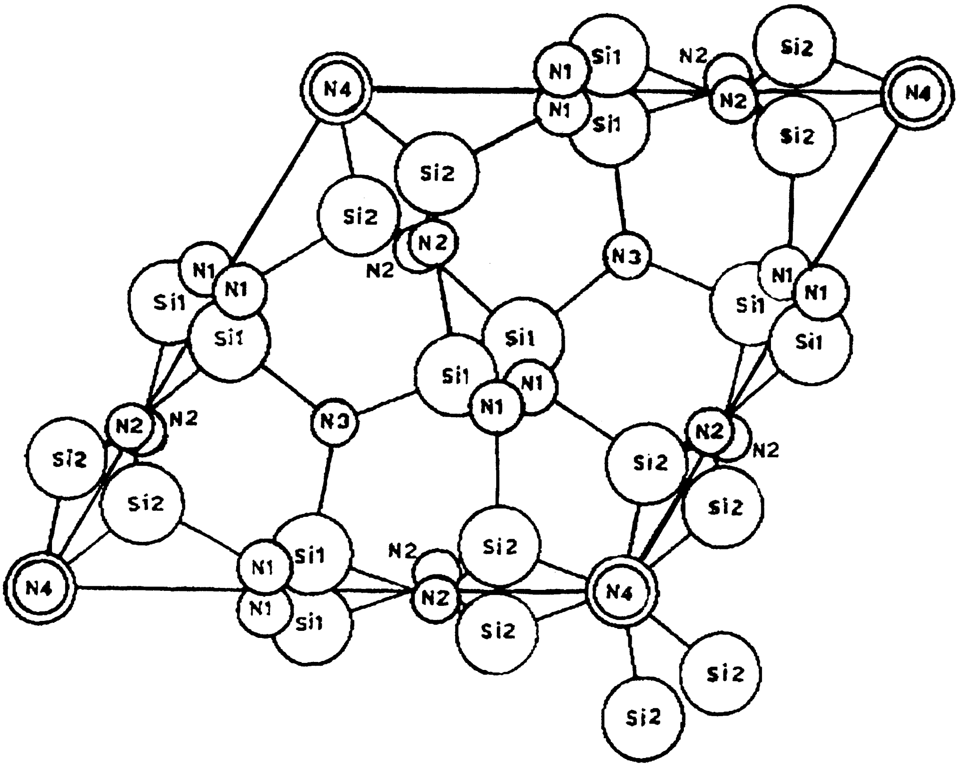
N4
N1
Si1
N2
Si2
N4
N1
Si1
N2
Si2
Si2
Si2
N2
N2
N3
N1
N1
Si1
Si1
N1
N1
Si1
Si1
Si1
Si1
N1
N1
N2
N2
N3
N2
N2
Si2
Si2
Si2
Si2
N4
N1
Si1
N2
Si2
N1
Si1
N2
Si2
N4
Si2
Si2

FIG. 2. The unit cell of α-Si_3N_4 projected along the *c* axis (courtesy S. Wild).

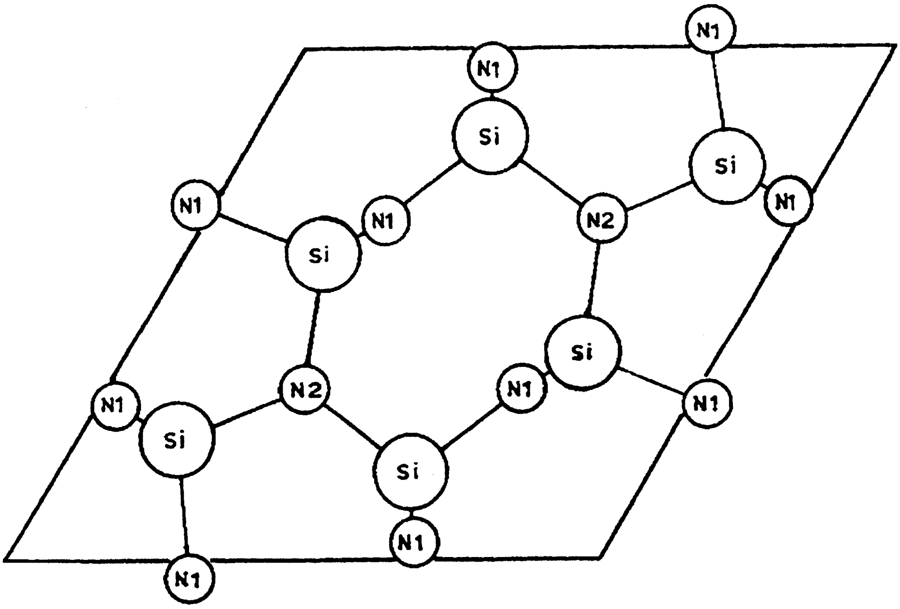
N1
N1
Si
Si
N1
N1
N1
N2
Si
Si
Si
N2
N1
N1
N1
Si
Si
N1
N1

FIG. 3. The unit cell of β-Si_3N_4 projected along the *c* axis (courtesy S. Wild).

tride invariably gives a noncrystalline, amorphous material. Because of the presence of hydrogen as a carrier gas, or in the ammonia (NH_3) and silane (SiH_4) reactant gases, hydrogen is always incorporated into this form of silicon nitride. Strictly its composition should be represented by $Si_xN_yH_z$, in which the N/Si ratio ranges from 0.5 to 1.6 (stoichiometric = 1.33), depending on the synthesis conditions. The amount of hydrogen incorporated is proportional to the hydrogen concentration in the reaction environment and inversely proportional to the deposition temperature: its concentration can range from <1 to >30 at.%. Hydrogen forms chemical bonds with the Si and N (Si–H and N–H groups), with a distribution between the two types depending on deposition conditions. Annealing in vacuum at temperatures above ~900 °C readily leads to complete loss of the hydrogen. The presence of hydrogen in amorphous silicon nitride is of fundamental importance for the chemical, physical, and electrophysical properties of the thin silicon nitride films essential for applications in microelectronics technology, and it significantly affects the properties of the film, such as etch rate, film density, and diffusion-barrier characteristics.

Oxygen is also often present in amorphous silicon nitride as an impurity element, as it is in α-Si_3N_4, and this too can have a very great effect on film properties. However, amorphous CVD films normally contain <0.5% of oxygen.

2. PRODUCTION METHODS

This section provides an introduction to the methods used to produce silicon nitride in its forms of technical importance: powders (the precursors to silicon nitride ceramics), amorphous thin film, and polycrystalline (and often multiphase) ceramic.

2.1 General

Because most ceramic production processes use powders as the starting point, much attention has been directed to the preparation of high-purity and fine-particle-size silicon nitride powders. In order to appreciate better the objectives in this respect, it has to be realized that it is impossible to sinter even submicrometer particle-size silicon nitride powder to full density using conventional ceramic conditions. There are two major reasons for this: the very low intrinsic diffusivities of Si and N in the Si_3N_4 lattice, even at temperatures above 1700 °C, and the volatility of silicon nitride provided by the decomposition

$$Si_3N_4 = 3Si_{(v)} + 2N_{2(g)}. \qquad (1)$$

The nitrogen molecule is a very stable species, and silicon nitride undergoes appreciable decomposition to silicon and nitrogen gas at temperatures above 1200 °C; this process provides an alternative surface-energy lowering mechanism to that required (the movement of material from particle–particle contact points into the surrounding void space) to permit the approach of particle–particle centers and densification. Silicon nitride powder must therefore be sintered in the presence of small amounts of a wetting liquid solvent of carefully controlled composition, to provide the low-conductivity pathway necessary for silicon nitride transport and particle fusion. Suitable liquids have been found to be the metal silicates, normally used at the 5- to 10-vol.% level and giving liquids at temperatures (depending on composition and quantity) from ~1200 °C upward. These silicates are formed *in situ* by reaction of added metal oxide with the silicon dioxide film naturally present on the silicon nitride particles, and providing 3–5% by weight of SiO_2; for this reason the "oxygen" content of a silicon nitride powder is an important aspect of its specification.

Silicon nitride powder suitable for sintering to a high-density and high-strength component will thus have high chemical purity (>99.5%), low oxygen content (<2%), and a specific surface area of ~10–20 $m^2\ g^{-1}$ (corresponding to mean particle sizes in the region of 200–100 nm). It will also consist of >95% α phase, which generally gives a higher mechanical strength sintered product, for reasons connected with the α-phase to β-phase transformation that takes place during sintering, with associated important microstructural developments (see Sec. 2.3.2).

2.2 Powders

Four routes are used for the commercial production of high-grade powders. These rely on the ready availability of materials such as high-purity silicon, silicon dioxide, and silicon tetrachloride. High-quality commercial silicon nitride powders are widely available at prices currently in the region of ($50 to $100)/kg.

2.2.1 Silicon Nitridation The earliest commercial synthesis of silicon nitride used the direct nitridation of micron-dimension, loosely packed, silicon powder by reaction with nitrogen or ammonia gas at temperatures in the range of 1250–1450 °C. This reaction can be summarized by the simple equation

$$3Si_{(s)} + 2N_{2(g)} = Si_3N_{4(s)}, \quad \Delta H_{1700\,K} = -724\ kJ\,mol^{-1}, \tag{2}$$

though this disguises a complex reality. Because the process is strongly exothermic, conditions must be carefully controlled to avoid a runaway reaction and premature fusion of the silicon powder particles (melting point, 1409 °C), which cannot then be fully nitrided. For this reason the process must be very slow (of the order of 20–50 h depending on the furnace loading), even with computer interlinking of reaction rate and furnace temperature control. The product powder tends to be strongly agglomerated and requires extensive wet milling to give the required submicrometer particles, coupled with subsequent chemical purification to remove metallic contamination and excess surface silicon dioxide.

2.2.2 Carbothermal Nitridation of Silicon Dioxide Silicon dioxide is not reactive toward nitrogen but can be converted to silicon nitride in the presence of carbon at temperatures using nitrogen or nitrogen/hydrogen mixtures, in the temperature range 1200–1450 °C:

$$3SiO_{2(s)} + 2N_{2(g)} + 6C = Si_3N_{4(s)} + 6CO_{(g)}. \tag{3}$$

This is also a slow reaction, and several hours can be required for completion, depending on temperature and gas flow rate: too high a temperature (>1460 °C with 1 bar pressure of nitrogen) favors the competing reaction of C with SiO_2 to form silicon carbide. Very fine amorphous silicon dioxide powder is used; the silicon nitride powder product is also fine, with a high α-phase content and of high chemical purity, with the exception of some contamination by residual carbon (which can with care be reduced to <0.5%).

2.2.3 Diimide Decomposition Silicon tetrachloride and liquid ammonia at −40 °C react rapidly to form the intermediate silicon diimide, which after removal of the ammonium chloride by sublimation or solution at 900 °C, or by washing with liquid ammonia, is decomposed by heating at ~1300 °C to yield crystalline silicon nitride. This complex series of steps, which may be summarized as

$$SiCl_{4(l)} + 6NH_{3(l)} = Si(NH)_{2(s)} + 4NH_4Cl_{(s)}, \tag{4}$$

$$3Si(NH)_{2(s)} = Si_3N_{4(s)} + 2NH_{3(g)}, \tag{5}$$

leads to a product of very high-purity α-Si_3N_4, with well-controlled submicrometer particles and morphology; there may be some chlorine contamination. The route is, however, expensive.

2.3 Bulk Forms

Silicon nitride ceramic is widely produced with two distinctive types, porous reaction-bonded silicon nitride and dense sintered or hot-pressed silicon nitride.

2.3.1 Reaction-Bonded Silicon Nitride Work in the 1950s on the nitridation of micron-size silicon powders led to the discovery that strong components of dimensions exactly those of the compacted silicon starting powder could be produced on heating at 1250–1450 °C under nitrogen or ammonia gas. This porous form of ceramic was termed "reaction bonded," or "reaction sintered," and was, historically, the first ceramic form of the material to be produced. The microstructure consisted of well-bonded silicon nitride microcrystals, with interconnected voids filled with material of a fine, whisker morphology.

Continuing detailed investigations showed that the silicon was nitrided through vapor-phase reactions involving $Si_{(g)}$, and probably also the gaseous silicon monoxide ($SiO_{(g)}$) generated from contaminant oxygen:

$$2Si_{(s)} + O_{2(g)} = 2SiO_{(g)} \tag{6}$$

$$3SiO_{(g)} + 2N_{2(g)} = Si_3N_{4(c)} + \tfrac{3}{2}O_{2(g)}. \qquad (7)$$

Nucleation and growth of silicon nitride takes place in the interparticle void spaces on silicon and silicon nitride surfaces, with the result that the increase in molar volume of ~22% is accommodated entirely within the interparticle void spaces of the compacted silicon powder; at the low temperatures used, there can be no sintering shrinkage of the product. Because compacted powders normally contain ~40–50% of void space, the reaction-bonded silicon nitride product contains 18–28% of porosity, which then has a dominant influence on the properties of the material. Figure 4 illustrates the characteristic appearance of a polished section of reaction-bonded silicon nitride, in which porosity shows as the dark phase.

2.3.2 Sintered Silicon Nitride A fully dense form of polycrystalline silicon nitride was first produced in 1961 by hot-pressing silicon nitride powder in the presence of magnesium oxide (generating liquid magnesium silicate at temperatures of ~1550 °C). This method is cumbersome, slow, and restrictive in terms of the shapes that can readily be produced by it. In 1975 a technique for the more versatile production of pressureless sintered silicon nitride was developed, based on the recognition that the main problem with obtaining dense sintered material was the volatility and dissociation of silicon nitride. This could be suppressed by immersing the component in a bed of powder of the same composition, or by using high pressures (>10 bar) of nitrogen gas. Both techniques have as objective the attainment of a thermodynamically equilibrated environment for the sintering silicon nitride powder, suppressing the tendency for dissociation into $N_{2(g)}$ and $Si_{(v)}$ [reaction (1)].

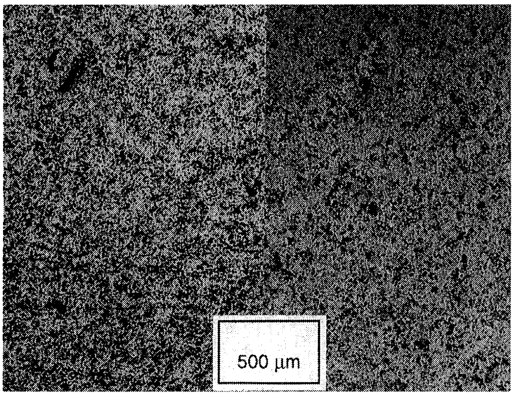

FIG. 4. Typical microstructures in two grades of porous reaction-bonded silicon nitride, shown in polished section. The dark regions are interconnected porosity. (Courtesy R. G. Stephen.)

Liquid-forming systems developed for hot pressing prove to be equally effective as sintering aids, though increased quantities are needed. Oxides such as MgO and Y_2O_3 at the 3–15 mol% level have been widely used as liquid-forming densification aids, and many other liquid-forming systems have since been investigated. Al_2O_3 is also a widely used additive, often in conjunction with Y_2O_3 although because a significant proportion of the Al and O now enter the Si_3N_4 crystal lattice, the product phase should strictly be regarded as a sialon. The additive oxides initially react with the nanometer-thickness oxide film naturally present on the Si_3N_4 particle surfaces at the ~3% level to form viscous silicate liquids in which silicon nitride has at high temperature (>1700 °C) significant solubility and mobility. Normal liquid-phase sintering occurs by dissolution and recrystallization of the silicon nitride, and fully dense materials can be obtained in 1–2 h at temperatures in the range 1700–1850 °C, provided effective measures are taken to control Si_3N_4 evaporation.

During the sintering process the ~100-nm-dimension α-Si_3N_4 particles convert to the more stable β phase. A major objective of the sintering process is to ensure the development of a microstructure of fine (1–10 μm) β-Si_3N_4 grains with fibrous, high-aspect-ratio, morphology shown in Fig. 5, which is of importance for the mechanical properties of the material. This microstructure is normally achieved by the use of a high-α-phase starting powder containing a low concentration of β-Si_3N_4 nuclei, and by prolonging sintering times to allow full development of the β-Si_3N_4 grains. A standard powder-bed sintering schedule for a 10–15 $m^2\,g^{-1}$ silicon nitride powder containing 5% Al_2O_3 and 5% Y_2O_3 by weight would consist of heating to 1000 °C under vacuum with a 1-h hold, followed by heating under 1 bar nitrogen at 1700 °C for 1 h, with a final anneal at 1800 °C for 2 h. Alternatively, the nitrogen pressure could be increased to 10 bar and

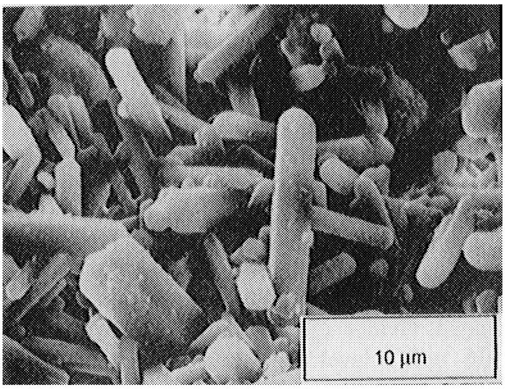

FIG. 5. The characteristic high–aspect-ratio β-Si_3N_4 grains in a high strength-sintered silicon nitride, delineated through removal by chemical etching of the inter-grain glass phase (courtesy P. Andrews).

heating continued at 1700 °C for a further 2 h. The powder in which the component is immersed is composed of a coarser grade of silicon nitride powder and contains ~50% of inert boron nitride powder to prevent excessive sintering of the bed.

During cooling the intergranular liquid forms secondary intergranular phases, which may be glassy or crystalline. Generally the silicate glass is the continuous phase, and it thus controls to a very large extent the high-temperature properties of the material. There is also a very thin (~1–2 nm) film of amorphous silicate material at all β-Si_3N_4 grain faces, thus blocking the continuity of the crystalline lattice. Crystallization of the intergranular glass would be highly desirable, but it is difficult to achieve fully at the grain boundaries. Densities of materials of this type are normally in the range 3.20–3.35 Mg m^{-3}, depending on the additive composition and quantity.

2.3.3 Hot-Pressed Silicon Nitride The uniaxial hot pressing of silicon nitride powder using graphite dies and pressures up to 30 MPa was the first method by which fully dense silicon nitride was obtained, but it has now generally been replaced by the more versatile pressureless sintering process. However, even with careful control of the sintering environment, it remains difficult consistently to produce 100% dense silicon nitride, completely free from voids. For this reason a high-pressure treatment may be carried out, in order to eliminate completely the last traces of porosity.

A well-established procedure for the production of small components is hot-isostatic pressing. This consists of the encapsulation of the component in a glass sheath (or presintering under normal conditions to the closed porosity stage, corresponding to >90% density) and heating in a cold-wall furnace at temperatures up to 1800 °C using nitrogen at pressures up to 400 MPa as the pressurizing fluid. Unlike uniaxial die pressing, there are no restrictions on component shape. A further major advantage of this process is that, because of the much higher pressures available, smaller volumes of liquid-forming additives can be used, with consequent benefits for the high-temperature mechanical properties of the material.

2.4 Thin Films

With the early realization that silicon nitride powder was very difficult to sinter to a fully dense material (although porous silicon nitride components could be readily produced by the nitridation of silicon powder), a natural step was to investigate the possibility of using CVD methods to produce dense silicon nitride in thin layer or sheet form. The CVD process utilizes the vapor-phase reaction at high temperature of volatile or gaseous precursor components, with nucleation and growth of the solid product phase occurring on any convenient surface or substrate. In the case of silicon nitride, silicon tetrachloride and ammonia are cheap and readily available materials, reacting at temperatures >700 °C:

$$3SiCl_{4(g)} + 4NH_{3(g)} = Si_3N_{4(s)} + 12HCl_{(g)}. \tag{8}$$

Silane, SiH_4, is an alternative and more reactive gaseous silicon compound:

$$3SiH_{4(g)} + 4NH_{3(g)} = Si_3N_{4(s)} + 12H_{2(g)}. \tag{9}$$

Carrier nitrogen gas is used. The reaction temperature and pressure determine the type and morphology of the product, with reactant pressures normally being of the order of ~10^{-3}–10^{-1} bar. At temperatures below ~1200 °C, the product is dense amorphous silicon nitride. This may be crystallized by heating at higher temperatures, or by prolonged heating at 1000 °C, though shrinkage and cracking is likely to occur. At higher deposition temperatures, films of dense, high-purity, polycrystalline α-silicon nitride are formed. Reaction rates under these conditions can be of the order of many μm h^{-1}, and while films of up to 100-μm thickness are easily developed, it is difficult to build up thicker stress- and crack-free layers by this method, and the technique has limited applicability in the context of silicon nitride component production. Prolonged annealing of α-silicon nitride films at temperatures >1500 °C permits conversion to the β form; this is a process useful for the production of very high-purity β-Si_3N_4, but otherwise, at present, it is only of academic interest.

Very thin, amorphous, silicon nitride films can be obtained by the direct nitridation of silicon surfaces by ammonia at temperatures up to 950 °C. This reaction is, however, self-limiting to thicknesses of the order of 10 nm because of the very low diffusivity of nitrogen in silicon nitride.

Chemical vapor deposition is the most common deposition method used in the electronics industry; this can be carried out at atmospheric pressures (APCVD) and at pressures between 10^{-4} and 10^{-2} bar (LPCVD). Until 1976 the standard deposition reaction used silane and ammonia at 1 bar pressure in a cold-wall reactor, and with temperatures between 700 and 900 °C. The SiH_4/NH_3 ratio was >150, and deposition rates were >10 nm min^{-1}. A typical, now commonly used, hot-wall low-pressure LPCVD reactor uses dichlorosilane ($SiCl_2H_2$) and ammonia at temperatures between 700 and 800 °C, and with pressures between 2×10^{-4} and 1×10^{-3} bar:

$$3SiCl_2H_{2(g)} + 4NH_{3(g)} = Si_3N_{4(s)} + 6HCl_{(g)} + 6H_{2(g)}. \qquad (10)$$

Growth rates are ~10 nm min^{-1}, and a hundred and more silicon wafers can be coated simultaneously and uniformly. Figure 6 shows schematically a horizontal hot-wall LPCVD reactor, in which silicon slices are mounted vertically and with a spacing corresponding to the molecular mean free path, to improve film quality.

This technique is now routinely coupled with complex molecular excitation schemes such as plasma enhancement (PECVD), laser enhancement (LECVD), photo (UV) enhancement (PHCVD), and microwave and electron cyclotron resonance (ECR). When silicon nitride coatings are used with devices containing aluminum or gold metallization, lower deposition temperatures, between 200 and 350 °C, are needed; to achieve this, plasma enhancement is used, in which a low-pressure (10^{-4} to 4×10^{-3} bar) glow discharge provides the reaction energy, with power supplies in the 30-kHz to 13-MHz range, and effective temperatures as high as 10^5 K. Growth rates are usually ~30 nm min^{-1}.

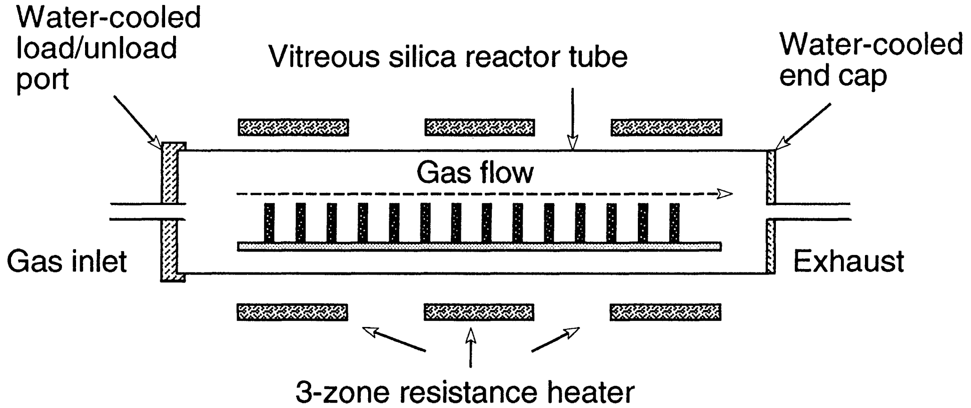

FIG. 6. Schematic diagram of a hot-wall LPCVD reactor.

3. PROPERTIES OF MATERIALS

The materials requirements for any one successful application are generally several in number. Silicon nitride has a very wide range of applications, and the range of properties of interest is therefore correspondingly broad.

3.1 Physical

3.1.1 Density Because of the open rather than "close-packed" crystal structures of α- and β-Si_3N_4, the density of pure crystalline silicon nitride is relatively low at ~3.17–3.19 Mg m^{-3}. Polycrystalline ceramics densified with the aid of heavier metal oxide systems (Y_2O_3, for example) can have densities higher than 3.2 Mg m^{-3}, because of the presence of the metal silicate intergranular phases. Amorphous silicon nitride has densities in the lower range from 2.4 to 3.1 Mg m^{-3} depending on production conditions: the low-pressure (and low-temperature) plasma technique yields the lowest densities and highest extents of hydrogen incorporation.

3.1.2 Diffusivity The intrinsic mobility of both silicon and nitrogen atoms in the Si_3N_4 lattice is very low and provides an additional impediment to the sintering of silicon nitride powder. Values for Si and N self-diffusion coefficients in Si_3N_4 have been difficult to determine, partly because of the difficulty of producing large enough high-purity specimens, and also because of the absence of convenient isotopes of Si and N. The most reliable data are for the tracer diffusion of ^{30}Si in α-Si_3N_4, for which $D = 10^{-15}$ m^2 s^{-1} at 1500 °C. The extrapolated value for the ^{15}N tracer-diffusion coefficient at 1500 °C is 10^{-19} m^2 s^{-1}, indicating that in this temperature region Si is the more mobile species.

3.1.3 Electrical The electrical properties of silicon nitride depend strongly on its purity and, in the case of sintered materials, particularly on the presence of intergranular glass. Pure silicon nitride is an electrical insulator with a band gap of between 3.85 eV (experimentally for amorphous silicon nitride) and 6.49 eV (theoretically for β-Si_3N_4). The dielectric constant of pure CVD α-Si_3N_4 depends on its density, ranging at 35 GHz from 6.77 (2.64 Mg m^{-3}) to 7.68 (3.15 Mg m^{-3}). This is in close agreement with room-temperature values measured for dense hot-pressed polycrystalline silicon nitrides containing 1% MgO and 8% Y_2O_3, which gave values of 8.32 and 8.25, suggesting that the behavior is intrinsic in silicon nitride. The value rises by a factor of ~1.1 over the temperature range 25–1400 °C. The loss tangent, in contrast, is almost entirely controlled by the intergranular glass, which contains mobile or polarizable metallic ions. For pure CVD α-Si_3N_4, it is $\sim 2 \times 10^{-4}$ at room temperature, rising to 3×10^{-2} at 1400 °C. Values for commercial dense silicon nitrides containing glass are ~10× larger. Amorphous silicon nitride films have dielectric constants in the range 6–7 for LPCVD material and 6–9 for PECVD materials.

The room-temperature dc resistivity of pure silicon nitride is $\sim 10^{14}$ Ω m, though PECVD films of variable composition have a wide spread of 10^4 to 10^{11} Ω m. In the case of impure materials with intergranular glass, this value drops rapidly with increasing temperature. At 400 °C the reported dc resistivity of silicon nitride hot-pressed with 5% MgO is $\sim 4 \times 10^7$ Ω m; at 1000 °C the value is $\sim 10^3$ Ω m. With larger amounts of additive oxide, the values fall even further. The ac resistivity becomes less frequency dependent with increasing temperature and approaches dc values at ~1000 °C. The dielectric strength of CVD silicon nitride is of the order of 10^9 V m^{-1}.

The electrophysical properties of amorphous CVD silicon nitride have been studied in detail. The properties of nonstoichiometric silicon nitride vary over wide ranges, probably because of the important role of trapped hydrogen in controlling the transport properties of silicon nitride. Charges in near-electrode regions of silicon nitride are caused by the electrons and holes injected, and the electrophysical properties of silicon nitride can only be studied effectively together with the properties of the contacts in metal-insulator-semiconductor structures. The trap depth varies from ~2.5 eV (the middle of the forbidden zone) to ~1.2 eV, depending on the fabrication method and hydrogen content. It is considered that electrons and holes have essentially the same importance for charge transport.

3.1.4 Optical Pure silicon nitride is colorless, though polycrystalline ceramic forms normally range from light gray to almost black, presumably because of the presence of traces of metallic impurity. Absorptions in the infrared occur at 830–835 cm^{-1} and 870–880 cm^{-1} (the strongest two bands), 920 cm^{-1}, and 1015–1020 cm^{-1} (weak bands). The refractive index of amorphous silicon nitride is in the range 1.8–2.5 and is sensitive to purity: low values indicate oxygen contamination; values higher than 2.0 indicate excess silicon.

3.1.5 Thermal The thermal expansivity of crystalline silicon nitride is very low, even for ceramic materials. The thermal expansion coefficient is a function of temperature, but a mean value for pure β-Si_3N_4 of 3.4 MK^{-1} over the temperature range 40–1200 °C is reported. Commercial polycrystalline materials containing intergranular glass, and possibly substitutional aluminum and oxygen, often include slightly lower values with a reported range of 2–4 MK^{-1}. The low expansivity is a major reason for the exceptionally good resistance to thermal shock of silicon-nitride-based materials, and the basis for many of their high-temperature structural applications.

The thermal diffusivity of dense sintered silicon nitride is mainly controlled by the intergranular phases, and by the glass in particular. Occurrence of the α-Si_3N_4 to β-Si_3N_4 phase transformation is seen to increase thermal diffusivity, possibly because of the higher intrinsic diffusivity of the β phase. Room-temperature values typically are in the range 10–25 $m^2\ s^{-1}$. Measured thermal conductivities at room temperature are 15–50 $W\ m^{-1}\ K^{-1}$ for dense material and 4–30 $W\ m^{-1}\ K^{-1}$ for porous reaction-bonded materials. Quoted specific heat values are in the range 700–1000 $J\ kg^{-1}\ K^{-1}$ with a mean of 780 $J\ kg^{-1}\ K^{-1}$ and independent of material type.

3.2 Mechanical

Silicon nitride is intrinsically hard and strong, and ceramic forms have been developed largely because of their good mechanical properties, particularly at high (>1000 °C) temperature. The strength of a good-quality sintered silicon nitride is almost constant up to 1200 °C. At higher temperatures strength begins to fall because of grain-boundary sliding aided by the intergranular glass film. This feature is enhanced by a good resistance to thermal shock, allowing its use for components that may be subjected to rapid temperature changes. Quenching from temperatures in the range 400–700 °C is needed to cause significant loss of strength during a standard water quench test.

3.2.1 Creep The slow deformation under load at high temperature is an important aspect controlling the high-temperature applications of silicon nitride materials. Most creep measurements are made at temperatures above 1100 °C and under loads typically of ~100 MPa. The creep of dense silicon nitride is largely controlled by grain-boundary sliding through the intergranular glass. This leads to grain-boundary separation, cavity formation, and cracking. Cavity nucleation occurs mainly at multiple grain edges. The process is complicated in air by simultaneous oxidation, during which intergranular cations diffuse to the outer surface, leading to a hardening effect and reduced creep rates. The volume fraction and composition of the intergranular phase have an important bearing on creep rates, and crystallization of the intergranular glass gives significant reductions in creep rates. The creep behavior of porous reaction-bonded silicon nitride is adversely affected by internal oxidation, which generates silicate glass with incorporation of trace metallic impurities in the silicon powder.

3.2.2 Fracture Toughness Most measurements of fracture toughness made on sintered polyphase silicon nitride using the diamond-indentation or notched-beam techniques show that fracture toughness is strongly influenced by the microstructure, and by grain aspect ratio in particular. In the case of dense β-silicon nitride materials with a fibrous grain structure, K_{1c} values as high as 12 $MPa\ m^{1/2}$ have been reported. Normal sintered materials will have values in the region of 5–7 $MPa\ m^{1/2}$. Measurements made using the Vickers indentation method on isotropic single-crystal α-silicon nitride give much smaller values of 1.9–2.9 $MPa\ m^{1/2}$, indicating that crack bridging and grain pullout are the primary mechanisms for toughening in sintered silicon nitride. The porous reaction-bonded form has a lower fracture toughness; values in the range 1.5–2.8 MPa $m^{1/2}$ have been reported.

3.2.3 Hardness Silicon nitride has an indentation hardness similar to that of silicon carbide (carborundum). Measurements made on single crystals of CVD α-silicon nitride show an average Knoop hardness value of 23 GPa, and the Vickers values are in the range ~28 to ~30 GPa for the 1010, 0001, and 1120 faces. These values are significantly higher than those measured on sintered polyphase silicon nitride materials, which contain a softer intergranular glass.

3.2.4 Strength Like all ceramics, silicon nitride fails at low temperature by brittle fracture: because there is no plastic flow (except at very high temperature), the stress-strain curve is virtually linear to the fracture point, and stress-concentrating defects are significant factors determining the strength of a component. The strength (σ) of a brittle material is often expressed in terms of the Griffith relationship

$$\sigma = (1/Y)[2E\gamma_s/c]^{1/2}, \qquad (11)$$

where E is Young's modulus; γ_s is the fracture surface energy; c is the dimension of the critical defect and often equated to the size of a microstructural defect such as a pore, large grain, or surface scratch; and Y is a constant related to flaw geometry. The term $[2E\gamma_s]^{1/2}$ is generally condensed to a critical fracture-toughness term, K_{1c}. Strength is normally measured by breaking bars in bend; tensile strength is experimentally much more difficult to measure, and values are rarely quoted (as a rough rule of thumb, the corresponding tensile strength would be 0.6 times a bend strength value).

The strength of the porous but otherwise single-phase reaction-bonded silicon nitride is controlled by the void fraction (indicated by density) and the size of the largest void. Most commercial materials have densities in the range 2.3–2.5 Mg m^{-3}, corresponding to void fractions of 0.72–0.80, and bend strengths in the range 150–250 MPa. Maximum void sizes are normally in the range 100–200 μm, though with care in the preparation and packing of the silicon powder before nitriding, they can be reduced to 50–20 μm, allowing strengths of 400–500 MPa to be obtained.

The strengths of dense, sintered silicon nitrides are much higher; room-temperature bend strength values are typically around 700 MPa, though values above 1000 MPa are routinely obtained. This is the result of the 1- to 10-μm-dimension β-Si_3N_4 grains and absence of large voids, coupled with the high fracture toughness given by the fibrous microstructure. Quoted strength values must, however, be treated with caution because of their sensitivity to sample size and to the quality of the surface finish.

The high-temperature strength of these materials is largely controlled by the intergranular glass. There is very little glass in reaction-bonded silicon nitride, and the strength remains almost unchanged up to 1400 °C. Sintered silicon nitride materials, in contrast, may contain 15% or more by volume of glass, and although they may retain good strength to 1200 °C, above this point their strength falls rapidly as a result of grain-boundary sliding and cavitation. The thickness of the intergranular amorphous film at the grain faces is ~1 nm and independent of liquid volume though controlled by the liquid composition. The bulk of the glass phase is concentrated at grain edges and grain corners. An important feature, therefore, is the use of liquid-forming sintering additives that either give very high-viscosity glasses, or can be crystallized by a post-densification annealing process. For this reason much development work has been carried out into silicon nitride materials sintered with additions of Al_2O_3 and Y_2O_3, which subsequently allow the crystallization by a high-temperature anneal of the intergranular yttrium aluminosilicate glass to α'- and β'-sialon solid-solution phases.

3.2.5 Young's Modulus This is a function of density (or void fraction), and to a lesser extent of composition. Porous reaction-bonded materials (of variable void fraction) usually have Young's-modulus values in the range 120–220 GPa. Dense sintered silicon nitride has a value in the region of 300–330 GPa, depending on the volume of lower-modulus glass present.

3.3 Chemical

Because most of the applications foreseen for the ceramic forms of silicon nitride involve high temperatures, oxidation and corrosion behavior have been studied in detail. High-temperature reactions with metals are less well studied. The properties of thin films of amorphous films of silicon nitride have been examined from the point of view of their effectiveness as barriers to the diffusion of metals, and oxygen, at temperatures in the region of 1000 °C.

3.3.1 High-Temperature Oxidation and Corrosion Silicon nitride is thermodynamically unstable with respect to oxidation to silicon oxynitride and silicon dioxide:

$$2Si_3N_{4(s)} + \tfrac{3}{2}O_{2(g)} = 3Si_2N_2O_{(s)} + N_{2(g)}, \qquad (12)$$

$$Si_3N_{4(s)} + 3O_{2(g)} = 3SiO_{2(s)} + 2N_{2(g)}. \qquad (13)$$

Oxide films of molecular thickness are always present on silicon nitride surfaces, and because of the low mobility of oxygen in them, they provide a barrier against continuing oxidation. Within thin films the film stoichiometry is continuously variable, from Si_3N_4 at the interior to SiO_2 at the surface. At temperatures above ~1000 °C, oxygen mobility becomes significant, and there is continuing oxidation, normally following approximately parabolic kinetics. However, oxidation rates are still relatively slow, and components can be used for many thousands of hours at this temperature. The high-temperature oxidative and corrosive degradation of silicon nitride materials has been studied extensively: the reviews in Nickel (1994) provide a good introduction to this subject.

The key factor under these conditions is the very effective diffusion barrier provided by both silicon oxynitride and pure silicon dioxide toward oxygen transport to the underlying silicon nitride. However, silicon dioxide reacts readily at high temperature with most metal oxides to form silicates of much higher conductivity toward oxygen; the protectiveness of the silicon dioxide film is thereby lost. Metal oxides, such as sodium oxide, potassium oxide, or vanadium pentoxide, are present in significant concentrations in many combustion environments and will accelerate oxidation by reacting with the protective silicon dioxide film. In sintered silicon nitrides, a process that can be regarded as autocorrosion occurs, through the diffusion of metal ions from the intergranular glass to the surface silicon dioxide, with the same effect. The composition and quantity of the continuous intergranular glass phase thus become of significance for the high-temperature oxidation resistance of sintered silicon nitride.

The oxidation behavior of reaction-bonded silicon nitride is influenced by similar factors, and also by the accessible void space, which permits internal oxidation to occur. Oxidation rates become slower at temperatures >~1200 °C, because of the development of a sealing outer barrier layer of silicon dioxide.

High-purity CVD films of silicon nitride are also effective barriers to the diffusion of oxygen and water vapor at temperatures of interest for the production of microelectronic devices (700–900 °C). This is likely to be the result of the formation of a very thin outer layer of silicon oxynitride or silicon dioxide; the nitride film is thus to this extent sacrificial.

The ability of thin, CVD- or plasma-deposited amorphous films of silicon nitride to act as diffusion barriers to dopants or impurities such as B, Al, Ga, In, and As is of great importance in the production of Si- and GaAs-based devices. Sodium penetrates less than 5 nm in 22 h at 600 °C, for example. Film stoichiometry is of importance for its effectiveness; small amounts of oxygen allow readier diffusion of Ga.

3.3.2 Molten-Metal Attack At high temperatures silicon nitride has a good resistance to attack by many molten metals, such as iron. It is particularly effective in the presence of molten aluminum, which reacts readily with many oxide refractory materials by abstraction of oxygen. In the case of the nitride it is likely that a thin barrier film of AlN is developed.

4. APPLICATIONS

This section deals with the applications of silicon nitride, treated in terms of the two important types of material—amorphous thin film and the bulk polycrystalline ceramic.

4.1 Thin Film

Extensive use is made of 100- to 200-nm thin films of amorphous silicon nitride in the semiconductor device industries. This is because the production of an integrated circuit uses multiple stages to give precisely controlled doped layers making up the individual components, as well as the deposition of other dielectric or metallic layers. Because of the high stress levels attained in CVD silicon nitride films applied to a silicon substrate, they are normally only deposited on top of a 5- to 50-nm pad or buffer layer of silicon dioxide. This serves to prevent peeling and cracking of the silicon nitride layer.

4.1.1 Masking Layers for Silicon Local Oxidation Because the oxidation rate of silicon nitride is <0.1 that of silicon, silicon nitride films are extensively used in most silicon device processes to allow the local oxidation of silicon. The surface of a silicon wafer is selectively protected by a 100- to 400-nm LPCVD or PECVD layer of silicon nitride, and the oxide layer is thermally grown in the exposed areas. At the end of the oxidation cycle, the nitride film is selectively removed by hot phosphoric acid. Figure 7 shows schematically the steps in the production of a simple recessed-oxide (ROX) structure, prepared on a silicon surface using a silicon nitride film as an oxidation mask.

4.1.2 Diffusion Barriers Amorphous silicon nitride has a higher density than amorphous silicon dioxide (~2.2 Mg m^{-3}) and is a better barrier to impurity-metal diffusion. It is therefore often used in very large-scale integration (VLSI) production lines to prevent the cross penetration of dopants during thermal processing stage—for example, where a lightly doped circuit component layer must be grown over a more heavily doped substrate. To prevent dopant escaping from the back surface of the substrate to the front, the back of the substrate is sealed by protective CVD silicon nitride film. The CVD silicon nitride films are also used to prevent the diffusion of elements such as boron from heavily doped devices already grown in a silicon substrate to neighboring regions. In this situation the film can have an additional function of providing a dielectric layer to form a capacitor.

Silicon nitride overcoating films are used to seal devices during fabrication against alkali-ion (sodium, for example) contamination arising from some encapsulants and to improve the resistance of the chips to mechanical damage before final bonding or encapsulation operations. The PECVD silicon nitride is often used between metallizing wiring levels as protection against mechanical damage.

Silicon nitride films are equally important in GaAs device fabrication, because of the tendency of GaAs to lose As by evaporation at temperatures >650 °C. A stoichiometric PECVD silicon nitride film deposited at 250–350 °C is an effective barrier against both As and Ga diffusion and maintains GaAs stoichiometry during thermal treatment (as well as improving the mechanical strength of the device). Films are readily removed after completion of processing by dilute hydrofluoric acid or a plasma etch.

4.1.3 Silicon Nitride Dielectrics Silicon nitride is a component of a useful gate dielectric in specific metal–insulator–semiconductor (MIS) memory device applications. Individually erasable cells are pro-

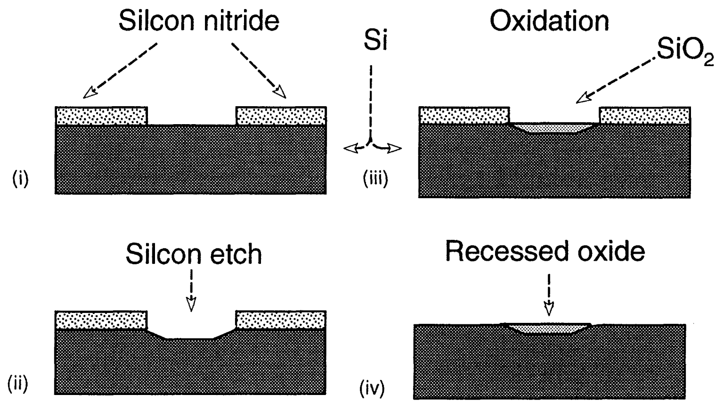

FIG. 7. Steps in the production of a recessed oxide pattern on a silicon surface, by localized oxidation using a silicon nitride mask.

duced with a dual dielectric control gate consisting of a silicon dioxide (~2 nm) and a silicon nitride (~50 nm) film, in metal nitride oxide silicon (MNOS) devices. Charge stored in traps at the oxide–nitride interface is retained when power is removed. The write/erase properties depend on the properties of the interfaces, and on those of the CVD silicon nitride, and are thus very sensitive to processing conditions.

The dielectric constant of silicon nitride is ~2× that of silicon dioxide and allows the production of a capacitor of the same value, but of half the thickness, as one made from silicon dioxide. It is thus an important material when very thin (<10 nm) dielectric films are required. Silicon nitride is used in very thin (~10 nm) gate dielectrics as a dual dielectric to provide increased breakdown field strength in field-effect transistors, and in dual-dielectric storage capacitors in multimegabit random-access memory arrays. Composite, (NO) LPCVD silicon nitride (4 to 8 nm)–silicon dioxide films with low leakage and a high resistance to boron penetration are also successfully used as gate insulators in insulated-gate field-effect transistors (IGFETs).

4.2 Bulk Ceramic

Dense silicon nitride has properties that distinguish it from other technical ceramics: its major features are very good resistance to thermal shock (because of a low thermal expansion coefficient), high strength and wear resistance, and a high fracture toughness giving a relatively low defect sensitivity (arising from a fine grain and fibrous microstructure). It also has the hardness of silicon carbide and is thermally stable under nitrogen or air to very high temperatures. Porous reaction-bonded silicon nitride has the same good thermal properties and a complex shape fabricability. As a structural material, silicon nitride has a number of established applications, most of which are generally on a small scale. Popper (1994) has provided a summary overview. Despite the many demonstrated viable applications, it has still to realize its full potential as a mass-produced high-temperature structural ceramic material.

4.2.1 Cutting-Tool Inserts Cermets such as cobalt-bonded tungsten carbide perform well at cutting speeds of <100 surface m min^{-1} but are thermally degraded at higher speeds. Dense sintered silicon nitride materials (including sialon-based materials) will cut cast iron, hard steel, and nickel-based alloys with surface speeds up to 25× those obtainable with tungsten carbide. Extensive use is made of silicon-nitride–based cutting tool inserts for machining cast iron in the automotive industry and for machining nickel-based superalloys for the aero industries.

4.2.2 Internal-Combustion Engines Much of the early work carried out on silicon nitride in the 1970s and 1980s was directed at the introduction of dense and porous silicon nitride hot-zone components into a new range of advanced gas-turbine and reciprocating engines, with projected significantly higher operating temperatures and therefore efficiencies. As a result of many intensive research programs, a large number of test engines were developed and satisfactorily evaluated. However, major stumbling blocks to the widespread incorporation of silicon nitride components into standard engines have been the cost of the raw materials and, probably more importantly, the technological problems associated with the mass production of high-grade ceramic engine components with a very high degree of reproducibility and long-term reliability. Reproducibility and reliability are not problems unique to silicon nitride ceramics: they have their roots in the most fundamental aspects of ceramic powder blending and sintering. Nonetheless, a number of engines make routine use of dense sintered silicon nitride components such as diesel glow plugs and turbocharger rotors, where stresses and temperatures are more modest and the consequences of failure less catastrophic. Current development work concerns the introduction of silicon nitride components where improved hot wear resistance is required, such as exhaust valves, bucket tappets, and rocker-arm pads. Figure 8 shows examples of engine components made from sintered silicon nitride. A useful feature of silicon nitride is its very low density, low inertia, and high modulus compared with metallic materials, which make possible the design of valve

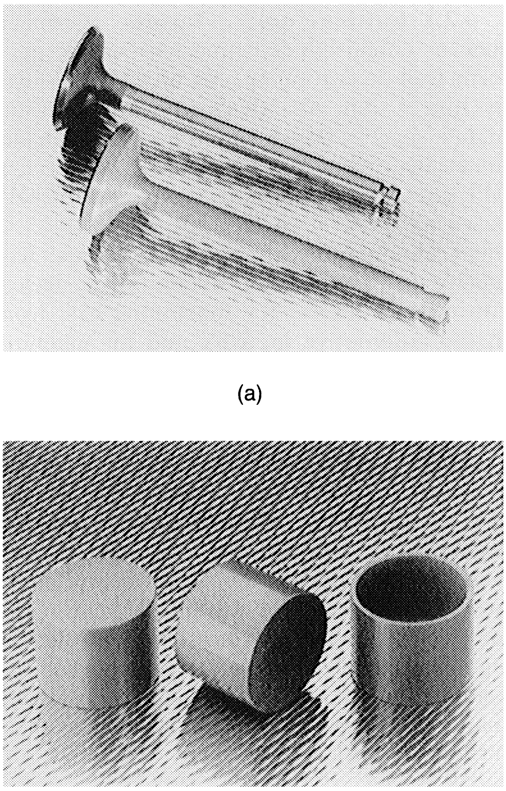

(a)

(b)

FIG. 8. Examples of sintered silicon nitride components: (a) silicon nitride (bottom) and metal (top) engine valves; (b) automobile bucket tappets. (Courtesy T&N Technology Ltd.)

springs for use in very high-frequency applications without resonance being set up.

4.2.3 Bearings The wear resistance, high stiffness, and low density of dense silicon nitride have led it to be considered for a high-temperature roller and ball bearing material. These features are well suited to combat the high Hertzian and centrifugal stresses induced in high-speed systems working at well above room temperature. Rolling contact fatigue tests have shown that silicon nitride can have a much longer life than conventional higher-density steel and cermet bearings (Horton, 1991).

4.2.4 Industrial Lower-density silicon nitride has been found to be useful under lower stress conditions, where refractoriness, chemical stability, and a good resistance to thermal shock are the main requirements. There is a steady demand for thermocouple sheath materials for use in the hot-metal processing areas and for components involved in the handling of molten reactive metals such as aluminum. Reaction-bonded silicon nitride is ideal for such components, in that fairly complex shapes can readily be fabricated in compacted silicon powder, and then nitrided. Threaded inert-gas–welding torch nozzles also provide a consistent market for this material. On a smaller scale, specialist kiln furniture of low thermal mass and good thermal shock resistance has been satisfactorily produced in reaction-bonded silicon nitride. Figure 9 shows examples of reaction-bonded silicon nitride components.

Dense silicon nitride has been evaluated as a coil and disc spring material for use in chemical plant in oxidizing or corrosive environments, at temperatures (>500 °C) where normal metallic springs would not be viable.

5. SUMMARY

This article has attempted to introduce the reader to some of the fundamental aspects of silicon nitride and to show its interest and challenge. The material has many usable properties in both the electrical and the structural areas. For this reason it has established a very wide spectrum of applications on the micro and macro scales, as thin films of amorphous material, and as large polycrystalline and multiphase components of carefully tailored microstructure. Many prob-

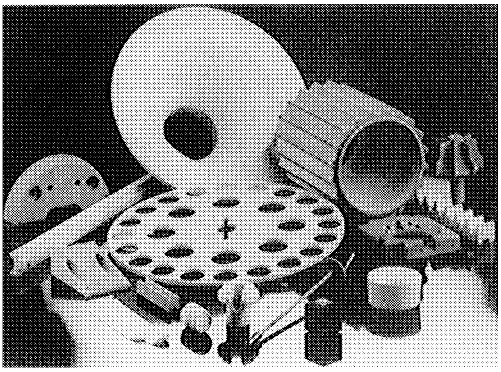

FIG. 9. Examples of reaction-bonded silicon nitride components (courtesy T&N Technology Ltd.).

lems have been overcome; there are many areas still requiring exploration in order to develop a full understanding of the structure-property relationships in these materials. The volume of published literature concerning silicon nitride is vast and generally not easy to plot a course through. The short list below provides no more than a brief introduction to the subject.

Works Cited

Belyi, V. I., Vasilyeva, L. L., Ginovker, A. S., Gritsenko, V. A., Repinsky, S. M., Sinitsa, S. P., Smirnova, T. P., Edelman, F. L. (1988), *Silicon Nitride in Electronics,* Materials Science Monographs, Vol. 34, Amsterdam: Elsevier.

Carlsson, R., Johansson, T., Kahlman, L., (Eds.) (1992), *4th International Symposium on Ceramic Materials and Components for Engines,* London: Elsevier Applied Science.

Ekstrøm, T., Nygren, M. (1992), "SiAlON Ceramics" *J. Am. Ceram. Soc.* **75** (2), 259–276.

Gmelin Institut für Anorganische Chemie (1991), *Gmelin Handbook of Inorganic and Organometallic Chemistry,* 8th ed., Silicon Supplement Vol. B5c: *Silicon Nitride in Microelectronics and Solar Cells,* Berlin: Springer Verlag.

Gmelin Institut für Anorganische Chemie (1995), *Gmelin Handbook of Inorganic and Organometallic Chemistry,* 8th ed., Silicon Supplement Vol. B5d1 and B5d2: *Silicon Nitride: Electrochemical Behavior, Colloidal Chemistry and Chemical Reactions,* Berlin: Springer Verlag.

Gmelin Institut für Anorganische Chemie (1996), *Gmelin Handbook of Inorganic and Organometallic Chemistry,* 8th ed., Silicon Supplement Vol. 5b1, *Silicon Nitride: Mechanical and Thermal Properties,* Berlin: Springer Verlag.

Hoffmann, M. J., Becher, P. F., Petzow, G. (Eds.) (1994), *Silicon Nitride '93,* Aedermannsdorf, Switzerland: Trans Tech Publications.

Horton, S. A. (1991), in: F. L. Riley (Ed.), *3rd European Symposium on Engineering Ceramics,* London: Elsevier Applied Science, pp. 35–50.

Nickel, K. G. (1994), *Corrosion of Advanced Ceramics: Measurement and Modelling,* Dordrecht: Kluwer Academic Publishers.

Popper, P. (1994), in: M. J. Hoffmann, P. F. Becher, G. Petzor (Eds.), *Silicon Nitride '93,* Aedermannsdorf, Switzerland: Trans Tech Publications, pp. 719–723.

Further Reading

Bonnell, D. A., Tien, T. Y. (Eds.) (1989), *Preparation and Properties of Silicon Nitride Based Materials,* Materials Science Forum Vol. 47, Aedermannsdorf: Trans Tech Publications.

Chen I.-W., Becher P. F., Mitomo M., Petzow G., Yen T.-S. (Eds.) (1992), *Silicon Nitride Ceramics,* Materials Research Society Symposium Proceedings Vol. 287, Pittsburgh: Materials Research Society.

Hoffmann, M. J., Petzow, G. (Eds.) (1994), *Tailoring of Mechanical Properties of Si_3N_4 Ceramics,* NATO Advanced Study Institutes, Series E: Applied Sciences, No. 276, Dordrecht: Kluwer Academic Publishers.

Riley, F. L. (Ed.) (1977), *Nitrogen Ceramics,* NATO Advanced Study Institutes, Series E: Applied Sciences, No. 23, Leyden: Noordhoff.

Riley, F. L. (Ed.) (1983), *Progress in Nitrogen Ceramics,* NATO Advanced Study Institutes, Series E: Applied Sciences, No. 65, The Hague: Martinus Nijhoff.

SILICIDES

See TRANSITION METALS, SILICIDES AND NITRIDES

SILICON, AMORPHOUS

MARTIN STUTZMANN AND CHRISTOPH E. NEBEL, *Walter Schottky Institut, Technische Universität München, Garching, Germany*

Introduction 151
1. **Preparation and Structural Properties** 152
1.1 Deposition Methods 152
1.2 Atomic Structure and Structural Stability 152
1.3 Vibrational Modes 154
2. **Electronic Properties** 155
2.1 Optical Absorption 156
2.2 Transport and Doping 157
2.3 Defects 162
2.4 Recombination 165
2.5 Metastability 170
3. **Alloys and Multilayers** 176
4. **Devices** 179
4.1 Overview 179
4.2 Solar Cells and Photodetectors 180
4.3 Field-Effect Transistors 182
5. **Summary and Outlook** 183
Glossary 184
Works Cited 185
Further Reading 185

INTRODUCTION

Amorphous silicon (*a*-Si), the disordered analog of crystalline silicon, is a prominent member of the large family of glassy or amorphous semiconductors that have been studied extensively during the last 50 years. Amorphous semiconductors are interesting materials in the field of applied physics because they can be deposited on many different substrate materials at relatively low temperatures while maintaining their basic semiconducting behavior. In addition, some specific properties of disordered semiconductors (bistable switching, radiation hardness, lack of wave-vector conservation, large-area and thin-film capability, chemical metastability, etc.) are particular to the amorphous state and have no direct analogy in crystalline semiconductors.

Particularly, amorphous silicon has captured most of the attention in the field of amorphous semiconductors since it was discovered around 1970 that *a*-Si deposited by low-temperature plasma-enhanced chemical vapor deposition (PECVD) of silane (SiH_4) has a very low defect density (approximately 10^{15} cm^{-3}) and can be doped both *n* and *p* type, qualitatively similar to crystalline Si (Spear, 1977). As a direct consequence of these discoveries, hydrogen-containing ("hydrogenated") amorphous Si (*a*-Si:H) has been employed in a number of electronic devices: thin-film solar cells and photodetectors using Schottky diodes or *p-i-n* junctions, thin-film field-effect transistors for display or printing applications, *a*-Si:H xerographic drums, particle detectors, and many others.

Whereas all of these devices worked in principle, it was discovered early on that their ultimate performance was seriously limited by the tendency of *a*-Si:H to create additional deep defects spontaneously under conditions of strong electronic excitation. The best known example is the so-called "Staebler–Wronski effect," which manifests itself via the creation of metastable Si-dangling-bond defects under strong illumination (Staebler and Wronski, 1980). Thus, while prior to 1980 most of the investigations dealing with amorphous Si were aimed at a better understanding of the structural, dynamical, and electronic properties of *a*-Si or intrinsic *a*-Si:H, since 1980 more applied issues and questions have dominated the field: modeling and optimization of specific devices and deposition processes, the microscopic nature and energy distribution of structural defects, doping, impurities, or structural and electronic stability.

During the last decade, amorphous silicon has also made a continuously growing contribution to the market of commercial elec-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

tronic devices. Large-area thin-film solar cells with stable conversion efficiencies of 10% or more now enter the mass production stage and will replace the limited applications of low-efficiency cells widely used in solar-powered calculators during the 1980s. Amorphous-silicon–based active-matrix flat displays are currently commercialized in high-end portable computers and present a strongly growing market for the near future. Amorphous silicon technology today is mainly attractive whenever large-area devices have to be combined with standard micrometer-lithography techniques. Besides flat-panel displays and thin-film solar cells, amorphous silicon therefore provides a promising technology for future applications, such as large-area image scanners, high-connectivity neural networks, or three-dimensional integration.

1. PREPARATION AND STRUCTURAL PROPERTIES

1.1 Deposition Methods

Amorphous silicon can be prepared by a number of different techniques, including thermal evaporation from a solid source, sputtering of crystalline Si targets (dc, rf), and decomposition of Si–H compound gases (silane, SiH_4, or disilane, Si_2H_6, etc.) via thermal, optical, or glow-discharge excitation. State-of-the-art *a*-Si:H samples today are mainly prepared by using a plasma-enhanced chemical vapor deposition (PECVD) system schematically outlined in Fig. 1. Reactive and carrier gases (SiH_4, H_2, PH_3, B_2H_6, etc.) enter the reactor via a gas manifold equipped with mass-flow controllers. The gases are dissociated in a parallel-plate capacitor that is part of a critically tuned rf-resonance circuit. One of the two capacitor plates is heated (≅200–400 °C) and acts as the substrate holder. The reactor is operated under constant pressure and flow rate using a throttled pumping system. Most systems use the industrial rf at 13.56 MHz for plasma excitation, but good results have also been obtained for reactors operating from dc up to 2.45 GHz. Thermal decomposition in hot-wall reactors or using hot filaments has also been applied successfully to *a*-Si:H deposition, especially for films with lower hydrogen concentration. Attempts to deposit *a*-Si:H via optical excitation of reactive gases (photo-CVD using CO_2^- or excimer lasers) have not met wide acceptance mainly because of window contamination problems and low deposition rates.

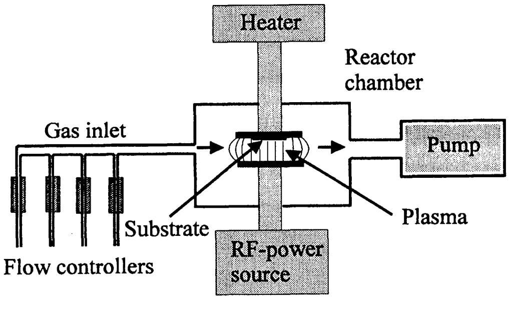

FIG. 1. Schematic diagram of a plasma-enhanced chemical vapor deposition reactor used for the preparation of amorphous-silicon thin films.

1.2 Atomic Structure and Structural Stability

Amorphous Si or hydrogenated amorphous Si prepared at temperatures below 500 °C is amorphous in the sense that neither x-ray scattering, Raman scattering, nor electron microscopy is capable of detecting significant fractions of ordered regions with diameters exceeding 20–30 Å. Instead, the amorphous structure is well described as a continuous random network with fourfold tetrahedral coordination of most atoms, but with no structural order extending beyond second-nearest-neighbor distances. Pure amorphous silicon contains a large number of vacancies and small voids, causing a density deficit of several percent with respect to crystalline Si, as well as a large number of electronic defects that make the material unsuitable for semiconductor devices. Moreover, the reactivity of elemental Si with oxygen leads to a rapid oxidation of internal and external surfaces. The chemical stability of *a*-Si is much improved if reactive atomic hydrogen is present during deposition or diffused into the sample afterwards. H atoms cause a passivation of internal surfaces and electronic defects by the formation of Si–H bonds. This reduces the density of electronic defects (Si dangling bonds) down to about 10^{16} cm^{-3} or less. Device-quality *a*-Si:H films prepared by PECVD of silane typically contain 5–10 at.% hydrogen.

Figure 2 presents a schematic view of the *a*-Si:H network, indicating several characteristic hydrogen-bonding units. In addition to the tetrahedrally coordinated disordered Si lattice, various hydrogen-related local structures can be identified: isolated Si–H bonds or Si–H_2 bridges, short fragments of polysilane $(Si–H_2)_n$, interstitial H_2 molecules, hydrogenated vacancies and microvoids, and other more hypothetical bonding configurations, such as the mobile H_2^* defect thought to be responsible for metastability phenomena of *a*-Si:H networks (see Sec. 2.5).

More quantitative information concerning the structural properties of *a*-Si and *a*-Si:H can be obtained from the radial distribution function, $D(r)$, which is determined from x-ray or electron-scattering experiments. $D(r)$ describes the probability of finding a network atom (only Si, since H does not contribute to the scattering efficiency to first order) at distance r from a given origin atom. In crystalline semiconductors, $D(r)$ consists of a series of discrete peaks with decreasing spacing corresponding to the various shells of atoms around the origin. In amorphous silicon, because of the long-range disorder, only the first two or three coordination shells can be identified (see Fig. 3). For large values of r, $D(r)$ approaches quickly the homogeneous-medium limit, $D(r) = 4\pi r^2\rho$, where ρ is the average macroscopic density. As shown in Fig. 3, the strength, position, and width of the radial distribution peaks contain information about the number of atoms, the average values, and the variation of bond lengths and bond angles. The first peak in *a*-Si is basically identical to that of *c*-Si, except for a small broadening due to bond-length variations ($\Delta a/a \approx 1\%$) and a lower integrated intensity reflecting the smaller density of *a*-Si. Similar arguments hold for the second peak (broadened by bond-angle disorder, $\Delta\varphi \approx 10°$) and the third peak (broadened by variation of the dihedral angle, $\Delta\psi$).

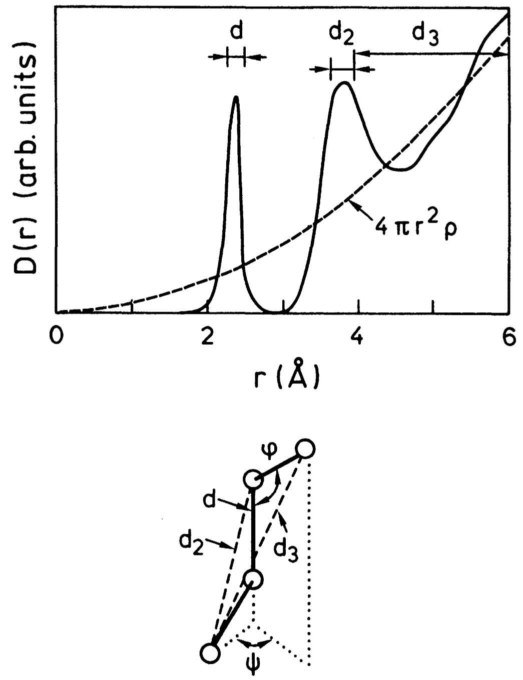

FIG. 3. Experimental radial distribution function $D(r)$ of amorphous silicon. The dashed line shows the homogeneous limit $4\pi r^2\rho$, where ρ is the macroscopic density. d, d_2, and d_3 indicate nearest-, second-nearest-, and third-nearest-neighbor distances according to the schematic model in the lower part of the figure. φ and ψ are the bonding and the dihedral angles, respectively.

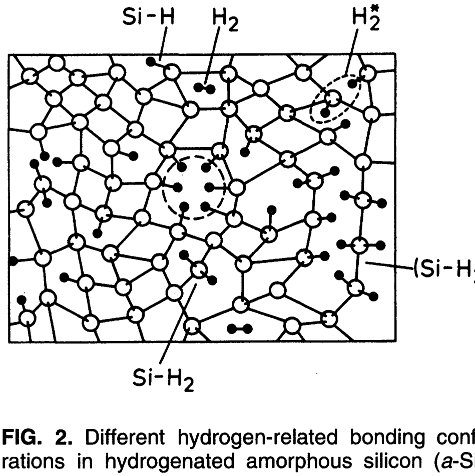

FIG. 2. Different hydrogen-related bonding configurations in hydrogenated amorphous silicon (*a*-Si:H). See text for details.

Hydrogenated amorphous Si is structurally stable over more than 20 years when stored at room temperature. In contrast, amorphous Si without hydrogen quickly oxidizes in air similarly to the crystalline Si surface. All forms of amorphous Si crystallize in an exothermic reaction upon heating to about 650 °C. Highly strained thin amorphous Si films sometimes exhibit so-called

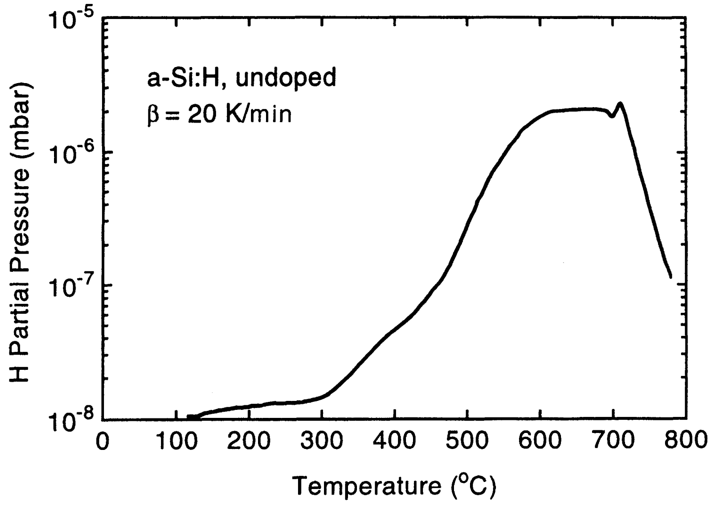

FIG. 4. Hydrogen effusion spectrum of undoped *a*-SiH. The sample was heated with a linear temperature ramp of 20 K/min in a constantly pumped vacuum system. The small peak at ≈700 °C is due to hydrogen release during crystallization.

explosive recrystallization even at lower temperatures when suitably activated. Prior to crystallization, *a*-Si:H looses the bonded hydrogen at temperatures between 300 and 550 °C. Weakly bound hydrogen (SiH_2 bonds, clustered H) is desorbed between 300 and 400 °C, whereas strongly bound H (isolated Si–H bonds) leaves the amorphous network between 500 and 600 °C (Fig. 4) (Beyer, 1985).

1.3 Vibrational Modes

The vibrational modes of amorphous silicon can be separated into network vibrations with energies below 520 cm^{-1} and hydrogen vibrations with energies above 600 cm^{-1}. Si-network vibrations fall into four bands corresponding to the transverse (T) and longitudinal (L) acoustic (A) or optical (O) phonons of the crystalline Si lattice (Lannin, 1984). Since the wave vector **k** of the phonons is no longer a good quantum number in an amorphous semiconductor (missing translational symmetry), only the density of phonon states can be used to describe the vibrational properties of the disordered network. The relation between the phonon density of states (DOS) and the phonon dispersion curves of crystalline Si is shown in Fig. 5. The DOS of lattice vibrations in *a*-Si can then be viewed as a disorder-broadened version of the crystalline DOS. Experimentally, the phonon DOS can be measured by inelastic neutron scattering (upper curve of Fig. 6),

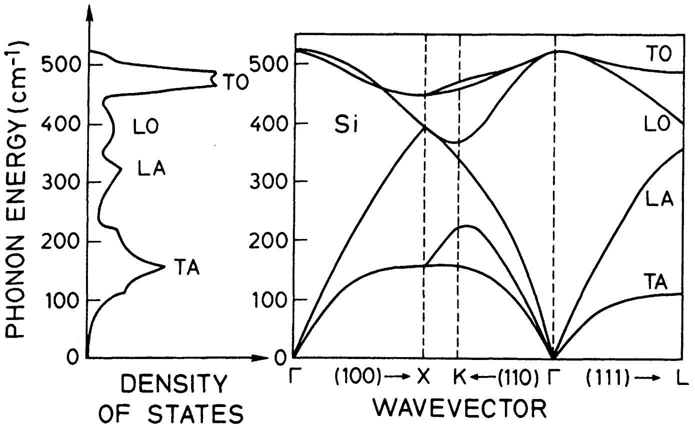

FIG. 5. Phonon dispersion curves of crystalline silicon and the corresponding phonon density of states. TO, LO, LA, and TA denote transverse and longitudinal optical and acoustic phonons.

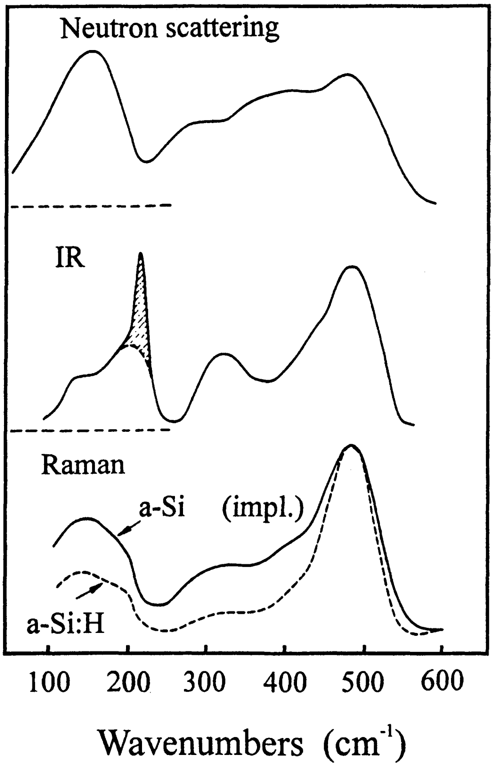

FIG. 6. Neutron scattering, infrared absorption (IR), and Raman spectra of amorphous Si. The shaded peak in the IR spectrum is due to a hydrogen-related mode (Si–H resonant mode). The two different Raman spectra correspond to device quality *a*-Si:H and amorphous Si produced by implantation.

by infrared (IR) absorption (middle curve), or by Raman scattering (lower trace). In the latter two techniques, corrections for the oscillator strengths and the scattering cross section of the different modes, respectively, have to be taken into account. Of practical importance is the use of Raman scattering as a convenient way to analyze the amount of bonding disorder and nanocrystalline admixture. Empirically, the mean deviation $\Delta\varphi$ of the bond angle from the tetrahedral value $\varphi = 109.5°$ is related to the full width at half maximum (FWHM), Γ, of the TO-Raman peak at 480 cm^{-1} via the relation $\Gamma^2 = \Gamma_0^2 + (A\Delta\varphi)^2$, with $\Gamma_0 = 33$ cm^{-1} and $A = 7$ cm^{-1}/deg. The presence of nanocrystallites, on the other hand, can be easily detected via a sharp Raman peak at approximately 520 cm^{-1}, corresponding to the zone-center TO/LO-phonon energy (cf. Fig. 5).

The Si–H bonds in *a*-Si:H give rise to various local vibrational modes (LVM), which can be divided into bond-bending and bond-stretching vibrations. For the assignment of the various modes, see Table 1.

IR absorption spectra of hydrogen LVM are of considerable importance because they allow a fast assessment of the total hydrogen content and the quality of *a*-Si:H films (Cardona, 1983). Dense, homogeneous *a*-Si:H films only show the bending and stretching vibrations of isolated Si–H bonds at 630 and 2000 cm^{-1}, respectively. In contrast, *a*-Si:H samples with inferior electronic quality commonly exhibit Si–H_2- or Si–H_3-related IR features as well as a significant fraction of bond-stretching vibrations at 2100 cm^{-1}. The total hydrogen content can be estimated from the integral of the Si–H bending absorption at 630 cm^{-1} according to the relation

$$[H] \approx (1.6 \times 10^{19}\,\mathrm{cm}^{-2}) \int \frac{\alpha(\omega)}{\omega}\, d\omega,$$

where α is the absorption coefficient in units cm^{-1}.

2. ELECTRONIC PROPERTIES

The discussion of electronic properties of amorphous Si has to take into account three principal features introduced by the characteristic disorder of the amorphous network. A first feature is the preservation of the short-range order. Thus, a similar overall electronic density of states as in crystalline Si is to be expected [Fig. 7(a)].

Second, the long-range disorder introduced by bond-length and bond-angle variations causes localization of electronic states at the band edges of the conduction and valence bands in which the electron wave func-

Table 1.

Mode energy	Bonding configuration	Mode
630	Si–H, Si–H_2	Bend, wag
875	Si–H_2	Bend ("scissors")
2000	Si–H	Stretch
2090	Si–H_2	Stretch
2100	Si–H + void	Stretch

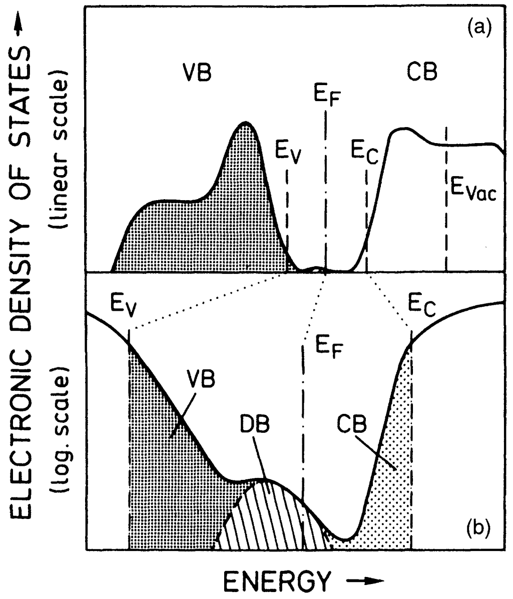

FIG. 7. (a) Main features of the electronic density of states of amorphous silicon. (b) Details of the localized states in the band gap on a logarithmic scale. E_V, E_C, and E_F indicate the valence- and conduction-band mobility edges and the Fermi level. VB, DB, and CB are the valance-band tail, the dangling-bond defect band, and the conduction-band tail.

tion is confined to a small volume. The density of localized states at the conduction- and valence-band edge decreases exponentially toward midgap. These "band-tail states" [VB and CB, Fig. 7(b)] control the transport properties because conduction takes place predominantly at the band edges. The transition energy separating extended from localized states is called the "mobility edge" (E_V and E_C in Fig. 7) because here the carrier mobility drastically decreases at $T = 0$ (Mott, 1987).

Third, deeper in the gap, electronic states due to defects control the recombination of electrons and holes. In *a*-Si:H, the main deep defect is the silicon dangling bond (DB, a threefold-coordinated Si atom), which can be charged positively (D^+) or negatively (D^-), or can be neutral (D^0).

In crystalline semiconductors the wave functions of the electronic states in the bands are solutions of the Schrödinger equation compatible with the periodic lattice potential (Bloch waves). In amorphous Si, however, the lattice potential is not periodic, so that the wave vector **k** is no longer a good quantum number. Instead, inelastic scattering of electrons occurs after a mean free path of only 10–15 Å, strongly decreasing the carrier mobility compared to crystalline semiconductors. A schematic view of delocalized versus localized electronic wave functions in *a*-Si:H is shown in Fig. 8. The detailed effects of disorder on optical absorption and electronic transport are discussed in the following two subsections.

2.1 Optical Absorption

Optical transitions in crystalline semiconductors are subject to the momentum conservation rule $\Delta k = 0$ because of the negligible wave vector of a low-energy photon. Well-known consequences are the strong optical transitions (absorption and luminescence) at the band edge of direct semiconductors and the weak, phonon-assisted transitions of indirect semiconductors, such as silicon. In amorphous Si the missing long-range periodicity destroys the concept of indirect or direct optical transitions. Instead, the dependence of the absorption coefficient α on photon energy $\hbar\omega$ is determined mainly by the combined density of occupied states in the valence band and empty states in the conduction band (Ley, 1984):

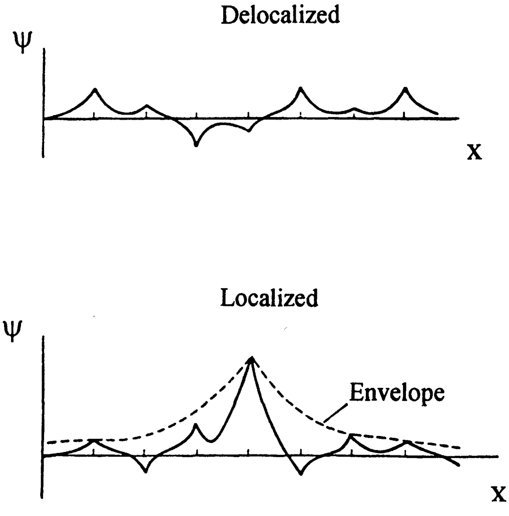

FIG. 8. Schematic view of delocalized and localized electronic wave functions in an amorphous network.

$$\alpha(h\omega) = \left(\frac{2\pi e}{m\omega}\right)^2 \frac{\omega}{hcV} \int M^2(E) N_{\mathrm{VB}}(E) \times N_{\mathrm{CB}}(E + h\omega) dE,$$

where V is the absorbing sample volume, $N(E)$ denotes the electronic density of states at energy E, and M is the optical transition-matrix element. In Fig. 9, experimental absorption spectra for *a*-Si and *a*-Si:H are compared with those of crystalline and microcrystalline Si. All forms of Si exhibit a strong absorption coefficient of $\alpha \approx 10^6$ cm^{-1} (corresponding to an absorption length of ≈100 Å) for photon energies above 3.2 eV. Significant differences only exist below the first direct optical transition of *c*-Si at 3.2 eV. In this energy range, the absorption of *c*-Si decreases rapidly by two to three orders as a result of the indirect, phonon-assisted nature of the transitions. The absorption finally vanishes almost completely below the indirect band gap (≈1.1 eV at 300 K) in high-quality crystals.

In contrast, device-quality *a*-Si:H and pure *a*-Si maintain a high level of absorption ($\alpha \approx 10^5$ cm^{-1}) in the visible spectral range because of the structural disorder (loss of **k**-vector conservation). As a direct consequence, amorphous silicon photoreceptors (solar cells, photodiodes) usually can be built with film thicknesses of 1 μm or less for visible light, whereas crystalline Si devices have to be at least one order of magnitude thicker to ensure a high quantum efficiency. In the near-infrared energy region below 1.5 eV, hydrogenated amorphous silicon according to Fig. 9 has a lower absorption coefficient than *c*-Si, because of the opening of the optical band gap by the incorporation of hydrogen. As a rule of thumb, one finds for the optical band gap E_{04} (energy at which $\alpha = 10^4$ cm^{-1}), as a function of hydrogen content,

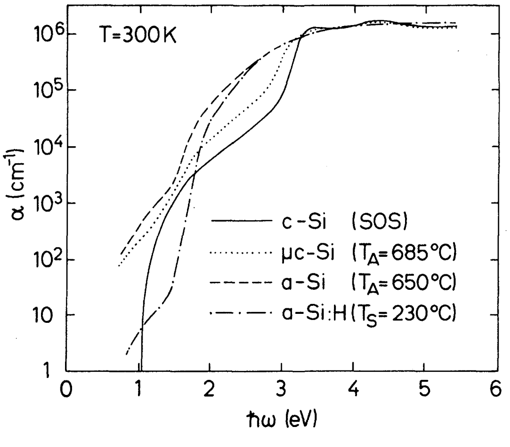

FIG. 9. Linear absorption coefficient, α, of crystalline silicon (*c*-Si), microcrystalline silicon (*μc*-Si), amorphous silicon (*a*-Si), and hydrogenated amorphous silicon (*a*-Si:H).

$$E_{04}(H) \approx 1.67 \text{ eV} + (0.02 \text{ eV})[H],$$

where [H] is the hydrogen concentration in at.%. Another quantity commonly used for the description of the optical band gap in disordered semiconductors is the Tauc gap, which is obtained by plotting $\sqrt{\alpha h\omega}$ as a function of the photon energy $\hbar\omega$ (Tauc, 1974). In *a*-Si:H, one finds $E_{\mathrm{Tauc}} \approx E_{04} - 0.2$ eV. The difference between *a*-Si and *a*-Si:H, as far as the optical gap E_{04} is concerned, can be easily seen in Fig. 9.

A characteristic feature of optical absorption in all amorphous semiconductors in general and amorphous silicon in particular is the existence of an exponential absorption tail ("Urbach tail") below E_{04}, in which the absorption coefficient, α, follows the empirical relation $\alpha \propto \exp[\hbar\omega/E_0]$, where E_0 (the Urbach slope parameter) is a material parameter. Device-quality *a*-Si:H has a rather steep Urbach tail with $E_0 \approx 50$ meV, whereas unhydrogenated *a*-Si has a much shallower tail ($E_0 \approx 100$ meV, see Fig. 9).

At energies around $\hbar\omega \approx 1$ eV, the absorption spectra of amorphous Si are dominated by optical excitation related to defect levels ("dangling bonds") in the gap (see also Fig. 7). This defect absorption appears as a low-energy shoulder in the spectra of Fig. 8. The defect density N and the absorption coefficient at $\hbar\omega = 1$ eV are found to be proportional to each other (see also Sec. 2.3).

2.2 Transport and Doping

The most prominent parameter to describe transport in *a*-Si:H is the electrical conductivity σ_D in the dark. σ_D is a macroscopic quantity that represents an average behavior of the propagating carriers. The interpretation of the experimental data therefore has to take into account all phenomena related to transport in disordered semiconductors like carrier scattering, tunneling rates, and trapping and thermal re-emission processes as well as the distribution of and the thermal occupation of states. In general, the conductivity σ_D is given by the product of carrier density n, carrier mobility μ, and the elementary charge:

$$\sigma_D = ne\mu.$$

In systems like amorphous Si with a distribution of possible conduction paths, it is useful to write the macroscopic conductivity as a summation over different conduction channels (Overhof and Thomas, 1989):

$$\sigma_D = \int N(E)e\mu(E)f(E,T)\,dE,$$

where $f(E,T)$ is the Fermi occupation function and $N(E)$ is the density-of-states distribution. The integral in general contains contributions from electrons and holes. When conductivity takes place far from the Fermi energy E_F and is dominated by a single type of carrier, then activated conduction is observed:

$$\sigma_D = \int N(E)e\mu(E)\exp\left[-\frac{E - E_F}{kT}\right]dE.$$

The dominant transport path is determined by the density of states, the carrier mobility, and the Boltzmann factor. For samples with a high density of localized defect states or at very low temperatures, conduction occurs by hopping in the distribution of localized states close to E_F. However, in state-of-the-art *a*-Si:H, conduction takes place predominantly in extended states above the mobility edge E_C of the conduction band, and in boron-doped *a*-Si:H below the mobility edge E_V of the valence band. Conduction in extended states is by thermal activation of electrons and holes from the Fermi level to the mobility edge. This leads for *n*-type conductivity to

$$\sigma_D = \sigma_{\min}\exp\left[-\frac{E_C - E_F}{kT}\right],$$

where $\sigma_{\min}$ is the minimum metallic conductivity:

$$\sigma_{\min} = N_C e\mu_C kT,$$

with μ_C the carrier mobility at the mobility edge.

For transport in delocalized band states above the mobility edge, the most important parameter is this conductivity prefactor, $\sigma_{\min}$. In the limit of degenerate doping, $\sigma_{\min}$ would be approximated by

$$\sigma_{\min} = en\mu = e(kTN_C)\left(\frac{1}{6}\frac{e}{kT}r_i^2\nu\right) \approx 200\ (\Omega\ \text{cm})^{-1},$$

where $N_C \approx 3 \times 10^{21}$ cm^{-3} eV^{-1} is the density of states at the conduction-band mobility edge, $r_i \approx 5$ Å is the inelastic scattering length, and $\nu \approx 10^{14}$ s^{-1} is the hopping frequency of carriers. This value is also corroborated by experiments performed on highly phosphorus-doped *a*-Si:H samples, where the Fermi level is close to the conduction-band mobility edge, E_C. A different relation for the minimum metallic conductivity was first proposed by Mott (1987) as $\sigma_{\min} = 0.03\ e^2/\hbar r_i$, which for $r_i \approx 5$ Å leads to the same value of 200 (Δ cm)$^{-1}$ determined above.

For transport in the nondegenerate case, where carriers are thermally or optically excited above the mobility edges, the most obvious mechanisms by which the prefactor $\sigma_{\min}$ could be different from the expected value of 200 (Ω cm)$^{-1}$ are thermally induced shifts of the mobility edge E_C, of the optical gap E_{gap}, and of the Fermi level (statistical shift). The statistical shift of the Fermi level and the shift of the mobility edge are dependent on the specific density-of-states distribution and cannot be predicted in the general case, whereas the shift of the optical gap is known experimentally to be

$$\gamma_G = -4.4 \times 10^{-4}\ \text{eV/K}.$$

There is still a discussion going on whether the concept of a sharp mobility edge is a reasonable assumption. However, most of the existing experimental data are analyzed on the basis of this concept. From a theoretical and experimental point of view, hopping conduction becomes significant whenever the density of states at the Fermi level exceeds about 10^{18} cm^{-3} and the temperature falls below a critical value.

Small amounts of phosphine or diborane added to silane in the glow-discharge deposition process cause variations in conductivity ranging from 10^{-13} to 10^{-2} S/cm (Spear, 1977). As an example, Fig. 10 shows the variation of the dark conductivity of *a*-Si:H upon doping with phosphorus. Around room temperature, σ_D is thermally activated with an activation energy between 0.8 eV (undoped) and 0.15 eV (highest doping level). The variations in the slope at high temperatures are discussed in more detail below. The discovery that amorphous silicon can be doped at all was surprising because of the theoretical prediction that substitutional doping of an amorphous semiconductor should be impossible. This prediction was based on the 8 − *N* rule (where *N* is the number of valence electrons per atom), saying that all atoms in a random amorphous network are able to achieve bonding configurations with their optimal chemical valency. Accordingly, both boron and phosphorus are expected to be threefold coordinated and electronically inactive. In crystalline semiconductors, on the other hand, the periodic lattice forces the impurity to have the same coordination as the network atoms, so that the phosphorus and boron atoms are incorporated with fourfold coordination and become electronically active.

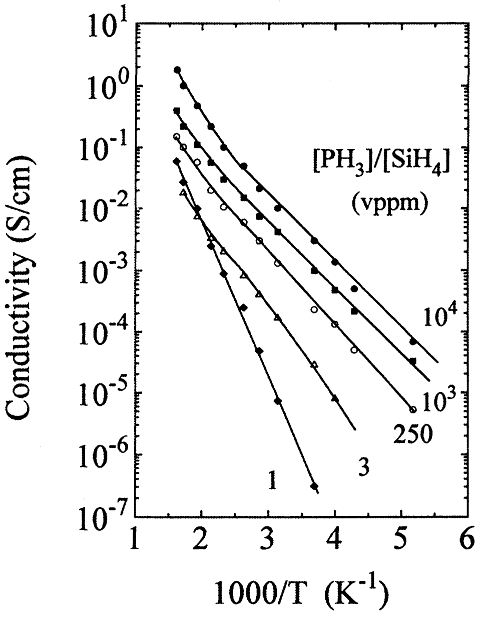

FIG. 10. Dark conductivity of *n*-type amorphous silicon as a function of temperature for different gas-phase doping levels with phosphorus.

Figure 11 summarizes further experimental results related to doping in *a*-Si:H. The variation of the Fermi-level position relative to the mobility edges E_V and E_C as a function of dopant gas concentration is shown in 11(d). As the Fermi level passes through the middle of the gap, the type of conduction changes from *p* type (E_F close to E_V) to *n* type (E_F close to E_C), as evidenced by the sign reversal of the thermopower in 11(c). The movement of the Fermi level through various regions of the localized states distribution (cf. Fig. 7) upon doping can also be traced via the changes of the spin-resonance *g* value in 11(b): for heavy boron doping, the

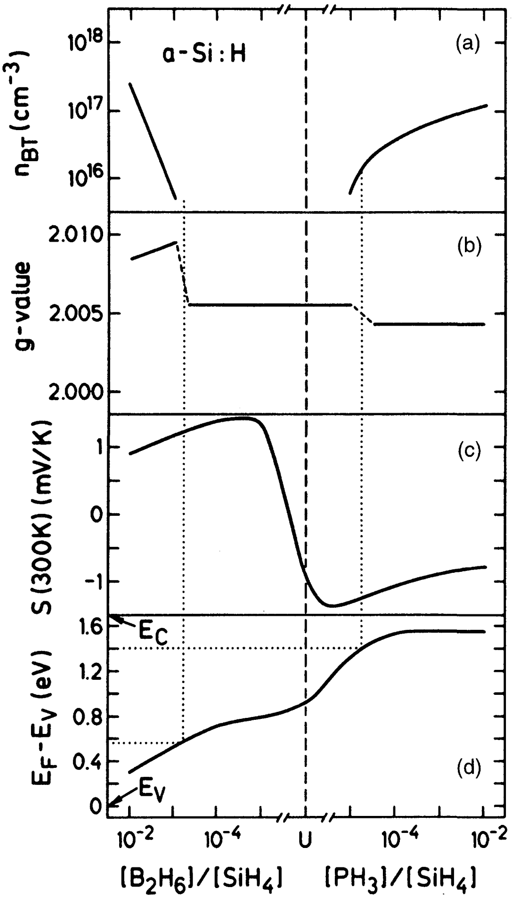

FIG. 11. Dependence of (a) the shallow carrier density in the band tails, (b) the electron-spin resonance *g* value, (c) the thermoelectric power, and (d) the Fermi-level position relative to the valence-band mobility edge in amorphous silicon as a function of doping with diborane and phosphine.

Fermi level lies in the valence-band tail (VB), and spin resonance is due to trapped holes (g = 2.008–2.01). In weakly doped and undoped *a*-Si:H, E_F is situated in the dangling-bond defect band (DB, g = 2.0055). Finally, in heavily phosphorus-doped samples, E_F shifts into the conduction-band tail (CB), and the spin-resonance signal is due to trapped electrons (g = 2.0044). In the limit of strong doping, electron spin resonance can also be used to measure the density n_{BT} of shallow carriers trapped in band-tail states, which increases exponentially as the conductivity activation energy decreases [11(a)].

The mobility μ of carriers is a further important transport parameter, especially for many device applications. It contains information about the transport mechanism and the density-of-states distribution. For insulating thin films such as *a*-Si:H, the time-of-flight technique has been the most successful way to determine μ (Tiedje, 1984), mainly because Hall experiments in amorphous materials are not very well understood. To perform a time-of-flight experiment, the *a*-Si:H layer is deposited between two blocking electrodes consisting of either *n*- and *p*-type layers or two Schottky barriers. Carriers are generated near one of the electrodes by a short laser flash and separated in an applied electric field. One type of charge is collected immediately by the nearby electrode, and the carriers with opposite sign cross the sample to the far electrode. The average motion of carriers is recorded via a fast digitizing oscilloscope. As a result of the fast trapping of carriers in the band tails (within femtoseconds or picoseconds), the deduced drift mobility μ_D represents an average over free and trapped carriers. The free mobility in transport states, μ_C, and the experimentally determined drift mobility are related by

$$\mu_D = \mu_C/(1 + f_{\text{trap}}),$$

where the trapping-time fraction f_{trap} for a single trap level is given by

$$f_{\text{trap}} = \frac{N_T}{N_C kT}\exp\left(\frac{E_T}{kT}\right).$$

Here N_T is the volume trap density, E_T the energy gap between the mobility edge and the trap level, and N_C the density of states at the mobility edge. Thus, the drift mobility in *a*-Si:H strongly depends on the density-of-states distribution close to the mobility edges and on the temperature. For high temperatures the drift mobility approaches the free-carrier mobility because the density of states in the band is higher than in the band tail. It is also important to note that in *a*-Si:H the tail states are distributed in energy (exponential tail), so that a simple one-trap model cannot be applied. More realistically, time-of-flight experiments have to be discussed on the basis of the multiple-trapping model and are commonly interpreted to evaluate the distribution of localized states at the conduction- and valence-band edges.

Experimental results for the electron and hole mobilities as functions of temperature are shown in Fig. 12. The general features of electron and hole mobilities are comparable, but electrons have significantly higher mobilities than holes. The extrapolations of the high-temperature mobilities indicate that the extended state mobility μ_C of electrons is 10–20 cm²/V·s, and that of holes is 0.5–5 cm²/V·s. Toward lower temperatures the drift mobility decreases strongly. Below T = 250 K for electrons and 410 K for holes, μ_D exhibits thermally activated behavior with activation energies dependent on the applied electric field. This is the dispersive transport regime where the transit of the carriers across a thin film is electric-field and thickness dependent (Nebel and Street, 1993). The activation energies of electron mobilities vary from 200 to 90 meV and of hole mobil-

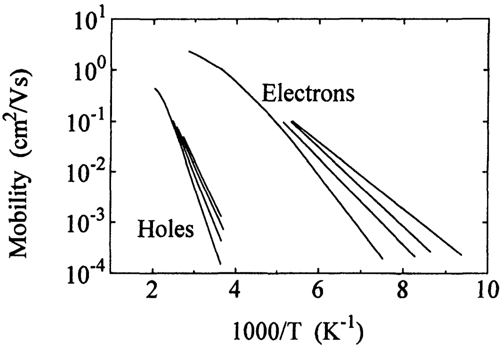

FIG. 12. Temperature dependence of the electron and hole drift mobility for different electric fields. (For a given temperature, the mobility increases with increasing electric field.)

ities from 500 to 350 meV, indicating that on the average holes are more deeply trapped than electrons. This is due to the broader valence-band tail, which decreases exponentially with a characteristic energy of about 45–50 meV (Urbach slope) compared with about 22–25 meV for the conduction-band tail.

The dispersive nature of transport is caused by the distribution of localized states over a broad energy range. Carrier interaction with these states is dominated by successive trapping and release events, which cause the center of the charge distribution to thermalize deeper and deeper in energy during the transit from the front to the back contact. The average trap depth of carriers, as they cross the sample, is given by the demarcation energy:

$$E_D = kT \ln(\nu_0 t),$$

where ν_0 is the attempt frequency for thermal emission from the traps ($\cong 10^{12}\ \mathrm{s}^{-1}$) and t is the time measured relative to the moment of carrier generation. Calculations show that the activation energy of the drift mobility represents reasonably well the demarcation energy at the moment when carriers reach the back contact.

As a third fundamental transport phenomenon, we mention the Hall effect in *a*-Si:H. The Hall effect has not contributed to the understanding of transport and doping because of its "double-sign anomaly," meaning that the Hall field has the wrong sign compared to the case of electron or hole transport in crystalline semiconductors. The sign anomaly is characteristic for amorphous semiconductors and seems to be linked to the small scattering distance in these materials. Figure 13 summarizes the Hall mobilities measured in intrinsic and phosphorus- and boron-doped *a*-Si:H using a conventional Van de Pauw contact configuration. Note that the Hall mobilities are significantly different from the drift mobilities deduced from time-of-flight experiments (Fig. 12).

Although amorphous Si can be doped substitutionally with group-V or group-III atoms just like crystalline Si, considerable differences exist as far as the doping efficiency, and effects of doping on the electronic density of states are concerned. As shown in Fig. 14, the total doping efficiency (number of charges added to the material per incorporated dopant atom) decreases like the square root of the dopant gas concentration in the deposition plasma, from unit doping efficiency in the parts per million range down to a doping efficiency of less than 1% for the highest dopant concentrations. At the same time, the density of deep dangling-bond defects increases, from less than $10^{16}\ \mathrm{cm}^{-3}$ in undoped *a*-Si:H to more than $10^{18}\ \mathrm{cm}^{-3}$ in heavily doped samples (see Fig. 15). Both observations can be qualitatively explained by

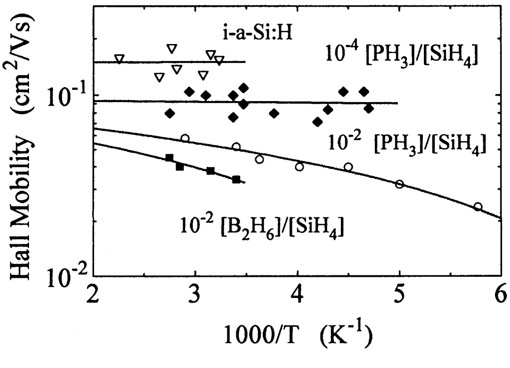

FIG. 13. Temperature dependence of the Hall mobility in intrinsic, *n*-type, and *p*-type *a*-Si:H.

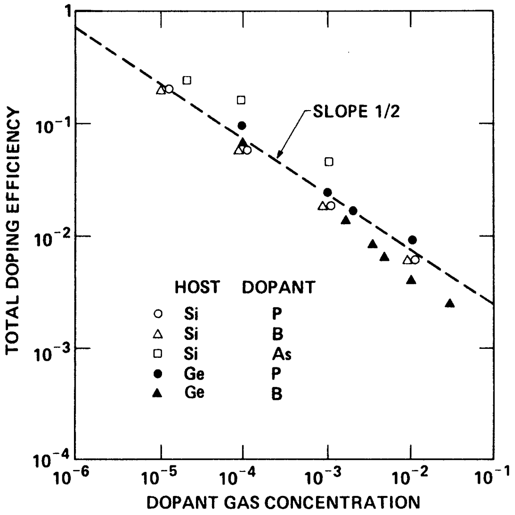

FIG. 14. Total doping efficiency versus gas-phase dopant concentration for different dopant-host systems. The dashed line indicates the expected behavior according to the autocompensation model.

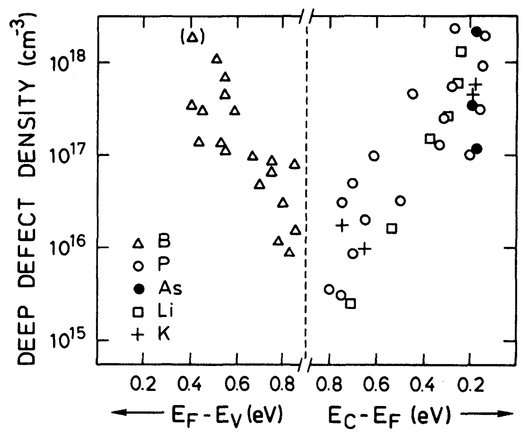

FIG. 15. Dependence of the deep defect density in amorphous silicon on the Fermi-level position for different dopants.

an autocompensation reaction of the form (Street, 1985)

$$\mathrm{Si}_4^0 + \mathrm{P}_3^0 \Leftrightarrow \mathrm{Si}_3^- + \mathrm{P}_4^+.$$

Here Si_4^0 denotes a fourfold-coordinated network atom, P_3^0 an inactive, threefold-coordinated phosphorus atom, Si_3^- a negatively charged, doubly occupied dangling-bond defect below the Fermi level, and P_4^+ an ionized substitutional donor state. Such an autocompensation reaction is believed to occur during and even after deposition. Assuming complete thermal equilibration and applying the law of mass action to the autocompensation reaction, one obtains

$$[\mathrm{P}_4^+] = [\mathrm{Si}_3^-] \propto [\mathrm{P}_3^0]^{0.5},$$

where [ ] denotes a solid phase concentration, and we have used the approximations $[\mathrm{P}_3^0] \gg [\mathrm{P}_4^+]$ and $[\mathrm{Si}_4^0] \approx$ const. Thus, the dangling-bond density should increase as the square root of the phosphorus concentration, whereas the doping efficiency, $[\mathrm{P}_4^+]/([\mathrm{P}_3^0] + [\mathrm{P}_4^+])$, should decrease as $1/[\mathrm{P}_3^0]^{0.5}$, in accordance with Figs. 14 and 15.

The autocompensation effectively keeps the Fermi level in doped *a*-Si:H between the deep defect band and the shallow donors, so that *a*-Si:H cannot be doped degenerately. Indeed, the number of shallow charge carriers, n_{BT} in Fig. 11(a), is always much smaller than the deep defect density in Fig. 15 for the same doping level. Further details of this phenomenon are discussed in Sec. 2.5.

As in crystalline Si, un-ionized neutral donors in *a*-Si:H are paramagnetic and can be observed in electron spin resonance at low temperatures. Figure 16 shows the corresponding ESR spectra of P_4^0 and As_4^0 levels in *n*-type *a*-Si:H. The multiple lines are due to $^{31}\mathrm{P}$ and $^{75}\mathrm{As}$ hyperfine interaction. Analyzing the hyperfine splitting and the occupancy (i.e., the resonance intensity) as a function of temperature allows one to estimate the localization radius (10 Å) and the position of the donor level relative to the conduction-band mobility edge (≈100–150 meV). Compared to crystalline Si, donor states in *a*-Si:H are thus more localized and energetically deeper (Stutzmann *et al.*, 1987).

2.3 Defects

We have already mentioned in conjunction with Fig. 7 that localized states in *a*-Si:H can be divided into three different

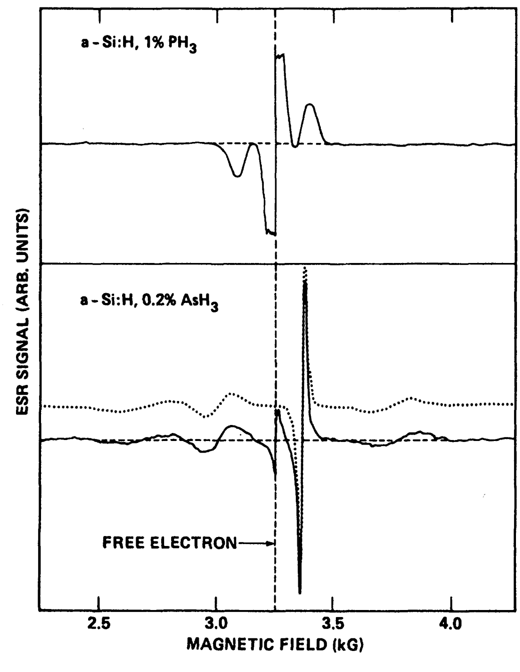

FIG. 16. Electron-spin resonance signal of neutral donors (phosphorus and arsenic) in amorphous silicon. The spin-resonance signal close to the free-electron position (dashed line) is due to electrons in conduction-band tail states present in all *n*-type samples.

types, according to their energetic position in the gap and also via the different g values in spin resonance [cf. Fig. 11(b)]. Microscopically, these three different types of localized states are commonly assigned to bonding and antibonding orbitals of weak or distorted Si–Si bonds (valence-band and conduction-band tail states, respectively) and to threefold-coordinated Si atoms, where one of the four sp^3-hybrid orbitals remains without a suitable bonding partner (deep dangling-bond defects).

Experimental evidence for this assignment only exists for the deep defects. In their neutral charge state, dangling bonds are occupied by a single electron and are, therefore, paramagnetic. The corresponding electron-spin-resonance signal has a characteristic g value of 2.0055 and is shown as the sharp central resonance in Fig. 17. The assignment of this signal to dangling-bond defects is corroborated by the existence of two weak hyperfine lines due to defect orbitals localized on ^{29}Si atoms (natural abundance 4.7%, nuclear spin $I = \frac{1}{2}$) (Stutzmann and Biegelsen, 1988). This hyperfine structure can be seen more clearly in ^{29}Si-enriched samples, where almost all Si atoms carry a nuclear spin and thus produce a hyperfine splitting (lower part of Fig. 17). A detailed analysis of this hyperfine interaction allows one to estimate the degree of localization of the defect orbital and the projected 3s and 3p contributions of the Si atomic orbitals. These estimates (60% localization on the central atom, 10% s, and 90% p character) agree very well with what is to be expected for the relaxed configuration of threefold-coordinated Si. An alternative model for the deep defect in a-Si:H, namely, a fivefold-coordinated Si state ("floating bond," Si_5), has not been widely accepted.

An important property of localized states in general is their effective correlation energy U_{eff}, which determines the relative energetic position of the different charge states of a given defect. In Fig. 18 we describe the case of an amphoteric defect with three different charge states in the gap and a positive correlation energy. The occupation factor $g(E)$ versus energy is shown in 18(a) for $T = 0$ and $T > 0$. At $T = 0$, states above the Fermi energy E_{F} are empty ($g = 0$), states between E_{F} and $E_{\mathrm{F}} - U$ are occupied by one electron ($g = 1$), and states below $E_{\mathrm{F}} - U$ are doubly occupied ($g = 2$). In Figs. 18(b)–18(d) we show the resulting energy distributions of empty, singly, and doubly occupied states for a band of defect levels centered at an energy E_D and having a width W. Note that this defect band describes the energy distribution for the $g = 0$ to $g = 1$ transition of individual defects. As indicated in 18(d), the electron–electron repulsion, which causes a positive correlation energy, shifts the doubly occupied levels up in energy by the amount U. It is commonly accepted that dangling-bond defects in amorphous Si have a positive correlation energy of the order of 0.2–0.4 eV. Negative correlation energies occur when the change in charge state is accompanied by a strong exothermic structural relaxation, which pulls the 1 → 2 transition energetically below the 0 → 1 transition. In this case, the singly occupied states are not stable in thermal equilibrium but decay spontaneously into pairs of empty and dou-

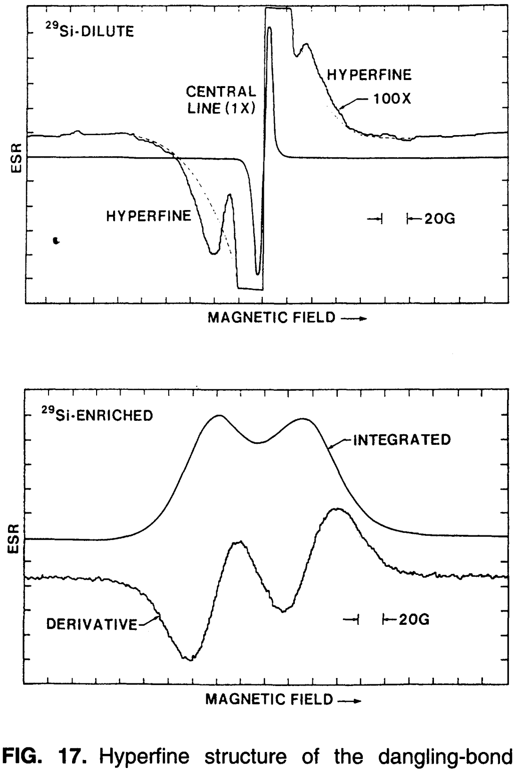

FIG. 17. Hyperfine structure of the dangling-bond spin-resonance signal in intrinsic amorphous silicon with the natural isotope concentration of ^{29}Si (upper part) and in ^{29}Si-enriched material.

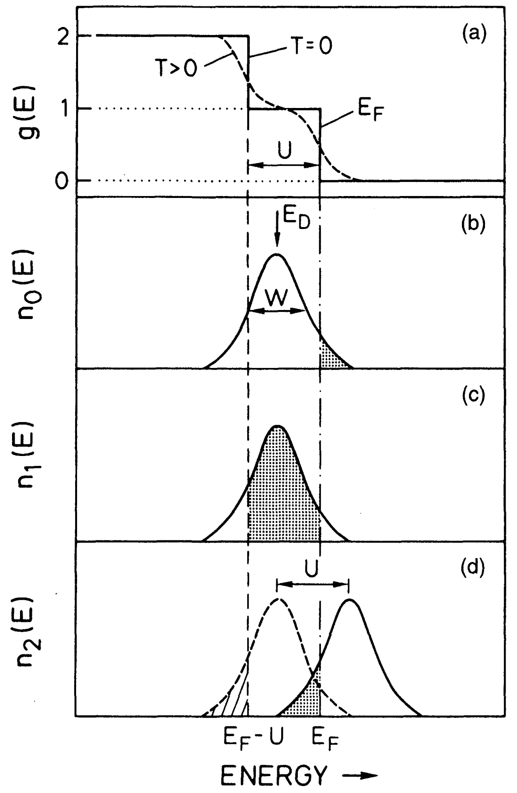

FIG. 18. Electron occupation of a defect band with a positive correlation energy *U*.

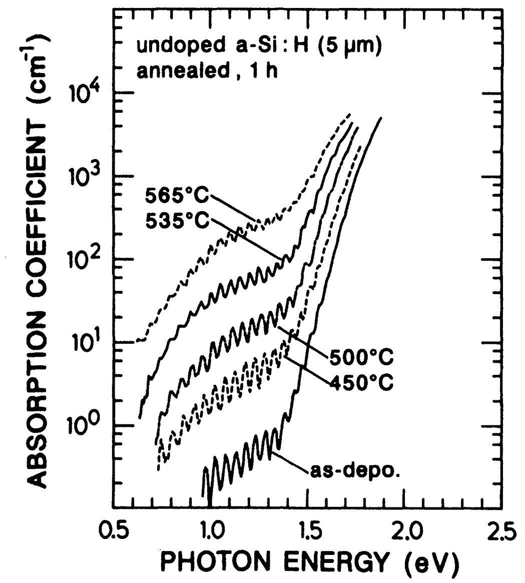

FIG. 19. Sub-band-gap absorption in amorphous silicon with different defect densities produced by thermal annealing.

bly occupied defects. This situation occurs, for example, in chalcogenide glasses.

For all applications of amorphous Si, the density and energy distribution of deep dangling-bond defects are important material parameters. Probably the most reliable experimental method to address this problem is the measurement of the subgap absorption by photothermal deflection spectroscopy (PDS) or the constant photocurrent method (CPM), which can be used for amorphous Si irrespective of doping concentration, film thickness, etc. Typical subgap absorption spectra for undoped *a*-Si:H with different thermal histories are summarized in Fig. 19. As-deposited material exhibits an exponentially decaying Urbach tail (cf. Fig. 8) due to excitation of carriers from the valence-band tail into the conduction band. This absorption tail is followed by the dangling-bond–related defect absorption for photon energies $\hbar\omega \leq 1.4$ eV. Shape and energy position of this absorption band are compatible with a broad (FWHM ≈ 0.3 eV) distribution of defect levels close to the middle of the mobility gap. The absorption coefficient α at a fixed energy position within the defect absorption is proportional to the density of defects, which in the case of Fig. 19 is altered by thermal dissociation of Si–H bonds via high-temperature annealing. As a rule of thumb one has

$$N\,(\mathrm{cm}^{-3}) \approx 10^{16}\alpha(1.3\ \mathrm{eV})(\mathrm{cm}^{-1})$$

as the relation between the defect density in units cm^{-3} and the absorption coefficient α at 1.3 eV in units cm^{-1}. The strict linearity between subgap absorption and deep defect density is demonstrated by comparing subgap absorption with other experimental techniques, such as ESR. Figure 20 shows a plot of $\alpha(1.3\ \mathrm{eV})$ versus the dangling-bond–related ESR spin density N_S for undoped *a*-Si:H in which the defect density has been changed by light soaking (cf. Sec. 2.5) or by thermal annealing, as in Fig. 19. Data such as those in Fig. 20 suggest that in undoped *a*-Si:H indeed the majority of defects are in their neutral, paramagnetic charge state and have a correlation energy *U* much larger than *kT* at room temperature (see Fig. 18).

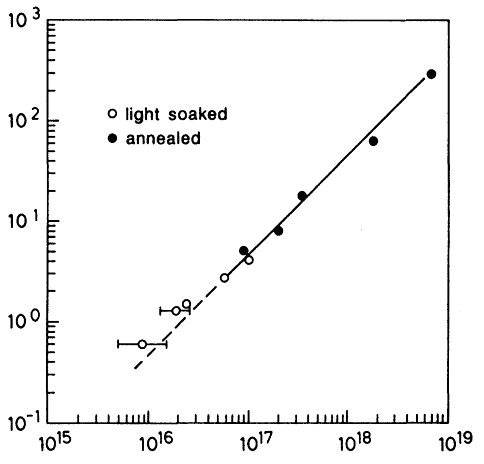

FIG. 20. Sub-band-gap absorption coefficient at a photon energy of 1.3 eV versus spin-resonance defect density in undoped samples. Different defect densities were produced by light-soaking (open symbols) or thermal annealing (closed symbols).

2.4 Recombination

The fate of charge carriers in *a*-Si:H following the creation—e.g., by optical absorption or injection—is another central question for applications of this material. The recombination of excess carriers in *a*-Si:H is schematically displayed in Fig. 21, which summarizes the basic features of a very complex phenomenon. The model has been developed combining data from a variety of experiments, most importantly photoconductivity, photoluminescence (PL), electron spin resonance, spin-dependent recombination, charge-collection measurements, etc. Electron–hole recombination takes place via three basic steps, each of which is sensitively dependent on the density and distribution of gap states, the temperature, the carrier density, and the applied electric field:

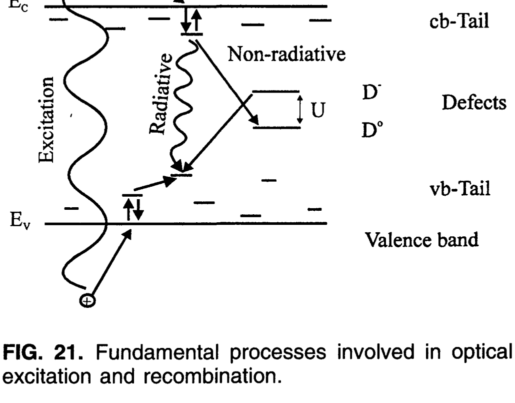

FIG. 21. Fundamental processes involved in optical excitation and recombination.

1. After photogeneration, electrons and holes undergo rapid inelastic scattering (thermalization) in the conduction and valence bands on a time scale typically shorter than 10^{-12} s. Following thermalization, carriers are trapped in the shallow tail states of the respective bands.
2. Further energy dissipation is dominated by thermalization within the band tails. Tunneling transitions (hopping down) are dominant at temperatures lower than 100 K. At higher temperatures multiple thermal excitation and retrapping governs the thermalization process. Typical time scales for this process are between 10^{-10} and 10^{-6} s.
3. Finally, carriers recombine either radiatively or nonradiatively. Radiative recombination proceeds by transitions of electrons from the conduction-band tail to trapped holes in the valence-band tail, causing a broad luminescence emission. The *a*-Si:H luminescence band is featureless, with a peak at 1.3–1.4 eV and a half-width of 0.25–0.35 eV. The luminescence quantum efficiency is practically 100% at low temperatures but decreases rapidly at higher temperatures and also as a function of defect density above 10^{17} cm^{-3}. The radiative recombination is mainly geminate (electron and hole stay together as a pair during their lifetime) when there is insufficient thermal energy to dissociate an electron–hole pair, which is bound by Coulomb interaction. The pair may separate by about 50 Å during thermalization, but when the Coulomb attraction is strong enough, the particles eventually diffuse back together and recombine. The transition from geminate to nongeminate recombination occurs when different electron–hole pairs start to overlap at sufficiently high generation rates or when the pairs diffuse apart at high temperatures or in high electric fields.

With increasing temperature, nonradiative recombination becomes significant because of thermal excitation of trapped carriers, which enhances the mobility. The enhanced mobility gives rise to faster trapping at defects, which is the primary reason for the thermal quenching of the luminescence. In high-quality *a*-Si:H with a low defect density ($<10^{16}$ cm^{-3}), the nonradiative recombination via defects begins to influence the lifetime at temperatures higher than 100 K. At room temperature the luminescence intensity is reduced by about four orders of magnitude, and the recombination is dominated by direct capture of electrons and holes into defects. Whether this capture of carriers by defects occurs via transport in the extended band states or, in contrast, as a diffusive hopping transition depends on the density and distribution of gap states.

Nonradiative recombination of photoexcited carriers in *a*-Si:H above 100 K via dangling-bond defects as the main recombination center has been investigated in great detail by spin-dependent photoconductivity (SDPC). In this technique one makes use of the fact that spin resonance of paramagnetic levels generally enhances electronic transition rates via these levels. SDPC spectra of device-quality, undoped *a*-Si:H taken at 150 K are shown in Fig. 22. The complex SDPC response can be separated into two underlying basic processes with slightly different characteristic time constants. In the first step an excess electron localized in the conduction-band tail (spin g value, $g = 2.0044$, see Fig. 11) is trapped by a neutral dangling-bond defect ($g = 2.0055$), charging the latter negatively. This transition is seen in SDPC as a combined resonance at the mean g value, $g = 0.5(2.0055 + 2.0044) = 2.0050$. In order to complete the recombination, a hole trapped in the valence-band tail ($g = 2.01$) has to recombine with the negatively charged defect. At 150 K, transport of holes in *a*-Si:H occurs by sequential hopping between tail states. This hopping motion is again spin dependent and gives rise to the second constituent resonance in Fig. 22 at $g = 2.01$.

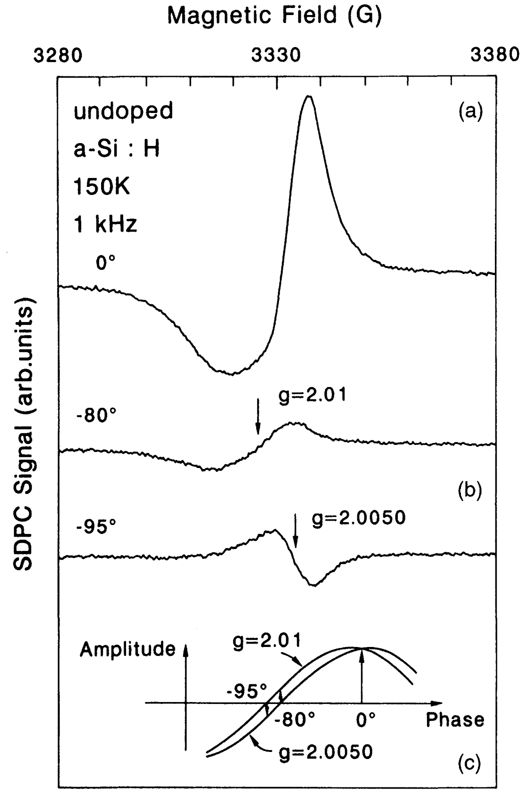

FIG. 22. Spin-dependent photoconductivity response of undoped amorphous silicon at 150 K. (b), (c) Decomposition of the signal in (a) into two constituents with a different phase shift relative to the magnetic-field modulation used for detection.

Whereas spin-dependent photoconductivity provides a qualitative picture of the recombination process, other methods can provide more quantitative parameters for specific recombination steps. For example, the capture cross section of dangling bonds in *a*-Si:H has been measured by a combination of ESR and charge-collection experiments in the temperature regime where the interaction between carriers and defects is dominated by direct capture (ballistic capture). Charge-collection experiments evaluate the $\mu\tau$ products that are obtained by measuring the collected charge as function of applied electric field in a time-of-flight experiment. The collected charge Q versus the applied electric field F is given by the Hecht formula as

$$\frac{Q(F)}{Q_0} = \frac{\mu_D \tau F}{d}\left[1 - \exp\left(-\frac{d}{\mu_D \tau F}\right)\right],$$

where τ is the deep trapping lifetime, Q_0 the total generated charge, and d the layer thickness. Figure 23 shows some results for in-

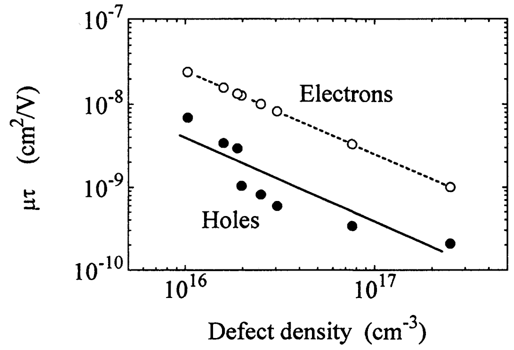

FIG. 23. Mobility-lifetime products of electrons and holes in undoped *a*-Si:H as a function of the defect density.

trinsic *a*-Si:H with different defect densities determined by ESR. The $\mu_D\tau$ products of electrons and holes are inversely proportional to the defect density. The hole $\mu_D\tau$ product is approximately a factor of 7 smaller than the electron $\mu_D\tau$ product, reflecting the smaller drift mobility of holes in *a*-Si:H. The capture cross section Σ_C of the defects for ballistic capture can be calculated as

$$\Sigma_C = \frac{el}{6kT}\frac{l}{\mu_D\tau N_D},$$

where e is the elementary charge, l the mean free path (assumed to be ≅10 Å for electrons and ≅5 Å for holes), k the Boltzmann factor, T the temperature, and N_D the defect density. In intrinsic *a*-Si:H, the dominant defect (Si dangling bond) is neutral and can capture either an electron or a hole. This is the reason for the linear decrease of both electron and hole $\mu_D\tau$ products as the defect density is increased (see Fig. 23). In doped *a*-Si:H, the defects are charged and can only trap carriers of opposite charge. This reduces the minority-carrier $\mu_D\tau$ product significantly for electrons in *p*-type material and holes in *n*-type *a*-Si:H. Charged defects have larger cross sections than neutral defects on account of the long-range attractive Coulomb interaction with excess carriers. The trapping parameters for the four transitions and the related capture cross sections for ballistic capture are summarized in Table 2 (Street, 1991).

Table 2.

Transitions	$\mu_D\tau N_0$ (V cm)$^{-1}$	Σ_C (cm^2)
$e \to D^0$	2.5×10^8	2.7×10^{-15}
$e \to D^+$	5×10^7	1.3×10^{-14}
$h \to D^0$	4×10^7	8×10^{-15}
$h \to D^-$	1.5×10^7	2×10^{-14}

The photogeneration of carriers enhances the conductivity of *a*-Si:H by many orders of magnitude, dependent on illumination power and defect density. Examples are given in Fig. 24, where the illumination power has been kept constant at 50 mW/cm^2 while the defect density has been varied over three orders of magnitude by electron bombardment. The almost linear dependence of σ_{ph} on the inverse defect density $1/N_D$ is in agreement with the role of dangling bonds as the main recombination center.

The photoconductivity in intrinsic *a*-Si:H is governed over the entire temperature range by electron transport. At low temperatures, $T < 150$ K, the weaker localization of electrons and the steeper conduction-band tail prevents freeze-out of electrons in deep traps. At temperatures higher than 150 K, electrons dominate transport as a conse-

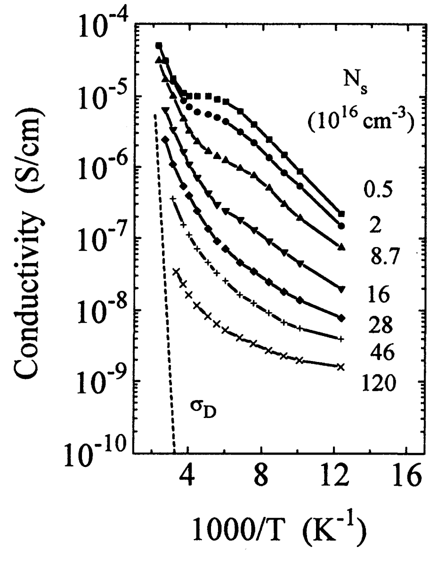

FIG. 24. Dark conductivity (σ_D) and photoconductivity in undoped *a*-Si:H with different defect concentrations produced by electron bombardement. (After Dersch *et al.*, 1983.)

quence of their propagation in delocalized states near or in the conduction band. The mobility of electrons is significantly larger than that of holes (see Fig. 12), which incidentally is also the reason for the *n*-type conduction in intrinsic and slightly compensated *a*-Si:H. In contrast, the broader valence-band tail causes holes to be more immobile, especially at low temperatures where they get trapped in the deep valence-band tail states ("save traps"). Here the density of traps is too small for rapid tunneling.

Of particular interest in Fig. 24 is also the specific temperature dependence of the photoconductivity at a given defect density. In the low-defect material, σ_{ph} is singly activated at low temperatures ($T < 150$ K, regime I), shows a shoulder or even a maximum at ≈170 K (regime II), and finally increases steeply toward higher temperatures (regime III). In regime I all holes are frozen out in the deep valence-band tail states (save hole traps). Recombination is limited by tunneling transitions of holes to the defects. The thermal activation is due to the increase in electron mobility (see Fig. 12). In regime II (170 K $< T <$ 250 K) holes become mobile via thermal excitation out of their deep traps and start to interact with the defects more efficiently. This causes a decrease in the effective electron lifetime and a corresponding decrease of σ_{ph}. Finally, in regime III ($T >$ 250 K) the increase in σ_{ph} is due to thermal excitation of electrons into shallow traps close to the conduction-band mobility edge. Electrons become very mobile and cause the drastic increase in σ_{ph}. With increasing defect density, σ_{ph} decreases, and the specific minimum associated with the hole freeze-out vanishes. For $N_D \geq 10^{17}$ cm^{-3}, σ_{ph} is governed by defect recombination via tunneling capture.

Figure 25 shows the absorption, photoluminescence, and photoluminescence excitation spectra of intrinsic *a*-Si:H. The luminescence is most effectively excited with homogeneously absorbed light ($\hbar\omega \approx E_{04}$). The observed radiative recombination occurs between electrons in the CB tail and holes in the VB tail. The photoluminescence intensity distribution is shifted from the Urbach edge by about 0.4 eV and is centered around 1.3–1.4 eV. The FWHM of the photoluminescence band is 0.25–0.3 eV (Street, 1984). The origin of the luminescence bandwidth and of the energy shift is attributed to

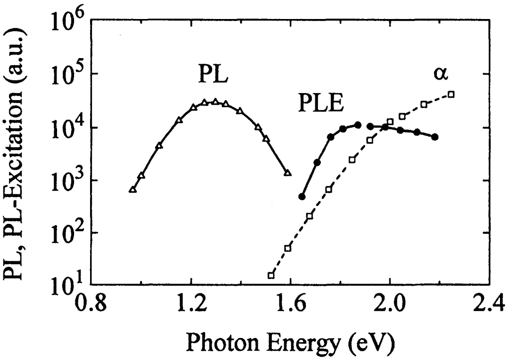

FIG. 25. Photoluminescence (PL), photoluminescence excitation (PLE), and absorption spectra (α) of amorphous silicon at low temperature.

1. disorder broadening due to the spread in energy of band-tail states and
2. dynamic broadening by phonons.

The identification of the relative contributions of disorder and phonon broadening in the luminescence spectrum is rather complex and not clear up to now. The phonon model explains the photoluminescence data reasonably well, indicating that disorder broadening contributes less. On the other hand, the disorder model also gives reasonable agreement, although the interpretation of the low-energy luminescence tail would result in an unrealistically broad valence-band tail.

The effect of defects on the luminescence intensity is shown in Fig. 26. The intensity drops drastically for defect densities larger than 10^{17} cm^{-3}, indicating a change of the recombination mechanism to a nonradiative process. At 10 K it is reasonable to assume that both radiative and nonradiative recombination are governed by tunneling transitions rather than by thermal excitation. The nonradiative tunneling rate is given by

$$R_{nr} = \nu_0 \exp[-2r_c/\alpha],$$

where r_c is the critical tunneling distance for the onset of nonradiative recombination, ν_0 the attempt-to-tunnel frequency (≈10^{12} s^{-1}), and α the localization length (≈10–12 Å). If an electron is closer to a defect than to the nearest hole, nonradiative recombination via defects is more likely. Equating R_{nr} with the observed radiative recombination rate, $1/\tau_r \approx 10^3$ s^{-1}, gives a critical distance $r_c \cong 10\alpha$,

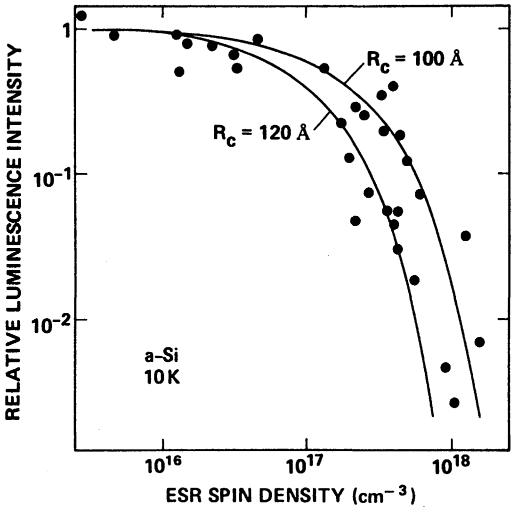

FIG. 26. Defect-density dependence of the luminescence intensity at 10 K. R_C is the critical radius for capture by defects discussed in the text. (After Street *et al.*, 1978.)

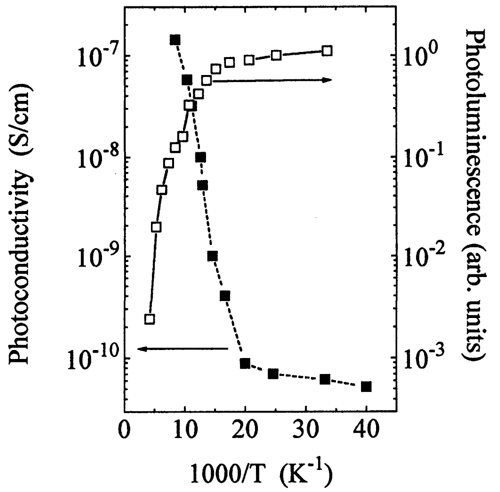

FIG. 27. Photoconductivity (left) and photoluminescence intensity (right) as a function of temperature.

which is 100–120 Å. The functional dependence of the luminescence efficiency η_{PL} on defect density N can be written as

$$\eta_{PL} = \exp[-4\pi r_c^3 N/3].$$

According to Fig. 26, this relation also gives critical radii r_c between 100 and 120 Å. In *a*-Si:H with a high defect density, a second PL band at ≈0.9 eV is observed, which is ascribed to radiative recombination between an electron in the CB tail and a dangling-bond defect.

Figure 27 shows the thermal quenching of the luminescence. Above T = 50 K the luminescence intensity drops significantly, and the photoconductivity rises over orders of magnitude. Obviously, the thermally increased carrier mobility (predominantly of electrons) quenches the luminescence and results in an effective separation of electron–hole pairs, thus giving rise to the increase in photoconductivity. At high temperatures thermal excitation is much faster than tunneling transitions, and, therefore, the interaction with defects and nonradiative recombination increases.

The competition between recombination via close electron–hole pairs (radiative tail-to-tail transitions of excitonic pairs) and via defects (nonradiative recombination) can also be seen in spin-dependent recombination (optically detected magnetic resonance or spin-dependent photoconductivity). As an example, Fig. 28 shows the changes of the SDPC spectra of undoped *a*-Si:H at an intermediate temperature of 175 K as a function of defect density, which in this case is increased by extended exposure to intensive light (light-induced degradation; see next section). At 175 K, radiative recombination and electron–hole pair dissociation followed by nonradiative defect recombination according

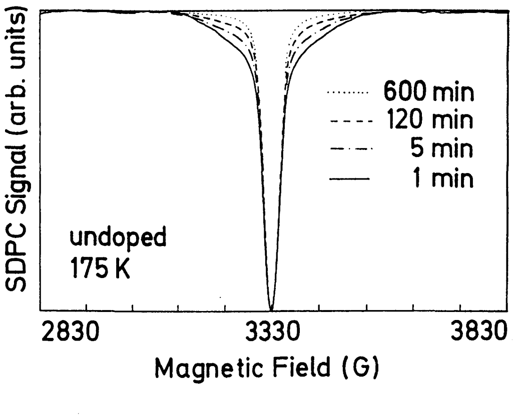

FIG. 28. SDPC signal of *a*-Si:H at 175 K after exposure to intense light for different durations. Note the gradual disappearance of the broad background line due to close electron–hole pairs.

to Fig. 27 both occur with similar probability in device-quality *a*-Si:H. In this case the SDPC spectrum consists of a broad and a narrow quenching line. The narrow line is the same as discussed in Fig. 22, except for the fact that in Fig. 28 we show the integrated SDPC spectrum and not the first derivative spectrum obtained by magnetic-field modulation. The broad quenching line, which can be seen as a pronounced tail to the narrow line, on the other hand, is due to excitonic pairs. The spin–spin interaction between the spatially correlated electrons and holes causes a broadening of the spin resonance to a FWHM of about 200 G, compared to ≈10 G for distant spin pairs. The main observation in Fig. 28 is that with increasing defect density (i.e., increasing illumination time), the relative contribution of the broad signal to the overall SDPC response decreases, showing directly the increase of the defect-related recombination path (Brandt and Stutzmann, 1991).

Another interesting phenomenon is the influence of high electric fields on carrier propagation and recombination. The photoluminescence, for example, is significantly quenched at fields larger than 5×10^4 V/cm. Field effects have also been detected in dark-conductivity experiments on intrinsic and doped *a*-Si:H. Time-of-flight measurements show that carrier mobilities are strongly enhanced in the presence of high fields. On the basis of Monte Carlo calculations, Marianer and Shklovskii (1992) have introduced an effective temperature, T_{eff}, which takes into account electric field effects on transport and recombination in *a*-Si:H:

$$\frac{T_{\mathrm{eff}}^2}{T^2} = 1 + \left\{\frac{0.67e\alpha F}{kT}\right\}^2,$$

where e is the elementary charge, α the localization length, F the electric field, k the Boltzmann factor, and T the temperature. Applying this equation to experimental data results in reasonably good fits with correct values for the localization length α of electrons (≈7–8 Å) and of holes (≈4 Å) in the respective band-tail states.

2.5 Metastability

Of great importance for applications of amorphous silicon is the stability of this material under strong optical or electronic excitation (exposure to intense light or high electric fields, current stressing, thermal cycling). Unfortunately, both undoped and doped *a*-Si:H show a variety of instability and metastability phenomena, which, at present, impose severe limitations on the performance of *a*-Si:H–based electronic devices.

The best known example of such a metastability is the light-induced degradation (metastable defect formation) of *a*-Si:H, also known as the Staebler–Wronski effect (Staebler and Wronski, 1980). The main experimental observations associated with this effect are summarized in Fig. 29. Upon illumination of undoped *a*-Si:H with intense, strongly absorbed light ($\hbar\omega \geq 2$ eV, $I \geq 100$ mW/cm^2), both the photoconductivity σ_{ph} and the dark conductivity σ_D decrease by at least one order of magnitude within several hours or days [Fig. 29(a)]. This change in conductivity is metastable in the sense that the effect of prolonged illumination can be completely reversed by annealing at ≈200 °C in the dark for 1 h. This defines two distinct states of an undoped *a*-Si:H sample, namely, the annealed, stable state *A* and the metastable state *B* obtained after long light exposure. According to Fig. 29, these two states differ not only by their photoconductivities and dark conductivities but also by the conductivity activation energy and the recombination kinetics. As shown in Fig. 29(b), the decrease in the dark conductivity of *a*-Si:H upon light soaking is due to a change in the Fermi-level position from $E_C - E_{\mathrm{F}} \approx 0.7$ eV in state *A* to $E_C - E_{\mathrm{F}} \approx 0.85$ eV in state *B*. Moreover, the dependence of σ_{ph} on the light intensity I in Fig. 29(c) indicates a transition from a bimolecular recombination in state *A* ($\sigma_{ph} \propto I^{1/2}$, recombination via direct electron–hole annihilation) to monomolecular statistics in the light-soaked state, *B* ($\sigma_{ph} \propto I$, recombination via defects acting as recombination centers).

Direct evidence for the creation of metastable defects (Si dangling bonds) upon light soaking of undoped *a*-Si:H comes from spin-resonance measurements. As an example, Fig. 30 shows the increase of the number of dangling bonds as a function of illumination time for an *a*-Si:H sample exposed to 300 mW/cm^2 of monochromatic light ($\hbar\omega \approx 2$ eV) at 40 °C.

The total sample volume in this case is about 3 μm × 0.5 cm^2, so that 10^{12} dangling bonds correspond to a defect density of about 10^{16} cm^{-3}. The main features in Fig. 30, following the initial stage *I*, are a sublinear increase of the number of defects with a characteristic $t^{1/3}$ dependence on illumination time (II) and a subsequent saturation of the effect at a defect density of about 2×10^{17} cm^{-3} (III).

The dependence of spin density on illumination time as observed in Fig. 30 can be explained via the following rate equation:

$$\frac{dN}{dt} = c_{sw}\frac{G^2}{N^2} - (N - N_0)\nu_0 \exp\left[-\frac{E_A}{kT}\right] - (N - N_0)B\frac{G}{N}.$$

In this equation, *G* is the generation rate of photoexcited electrons and holes; *N* is the total defect density; N_0 is the initial (stable) defect density; and c_{sw}, ν_0, E_A, and *B* are material parameters. The first term on the right-hand side describes the creation of new metastable defects by nonradiative tail-to-tail recombination of electrons and holes, where both the electron and hole densities in intrinsic *a*-Si:H at 300 K are proportional to *G*/*N* (monomolecular limit). c_{sw} is a susceptibility parameter of *a*-Si:H for metastable defect formation and is about 10^{-15} cm^3 s^{-1}. Taking into account only this defect creation term, one obtains from the above equation

$$N^3(t) - N_0^3 = 3c_{sw}tG^2,$$

which quantitatively accounts for parts I and II of Fig. 30 (Stutzmann, 1992).

The negative terms in the rate equation for *dN*/*dt* describe the thermal and the light-induced annealing of metastable defects, respectively. Thermal annealing occurs with an activation energy $E_A \approx 1.1$ eV. The prefactor ν_0 is about 10^{10}–10^{12} s^{-1}. Light-induced annealing is only important for high optical generation rates, *G*, and low temperatures ($T \leq 300$ K). The steady state in Sec. III of Fig. 30 is due to a balance between defect creation and annealing. The value of $\approx 2 \times 10^{17}$ cm^{-3} for the saturated defect density occurs under standard illumination conditions. Saturation values in excess of 10^{18} cm^{-3} can be obtained by illumination with high-intensity short laser pulses and active sample-temperature stabilization at 300 K.

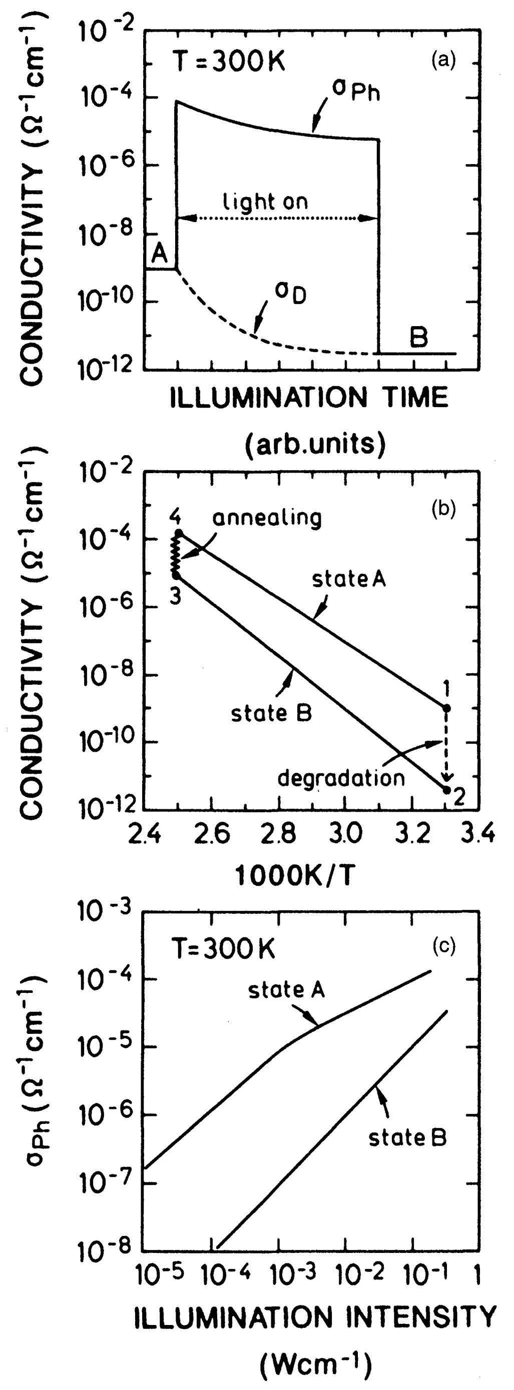

FIG. 29. Main effects of light-induced degradation on dark conductivity and photoconductivity in undoped amorphous silicon (Staebler–Wronski effect).

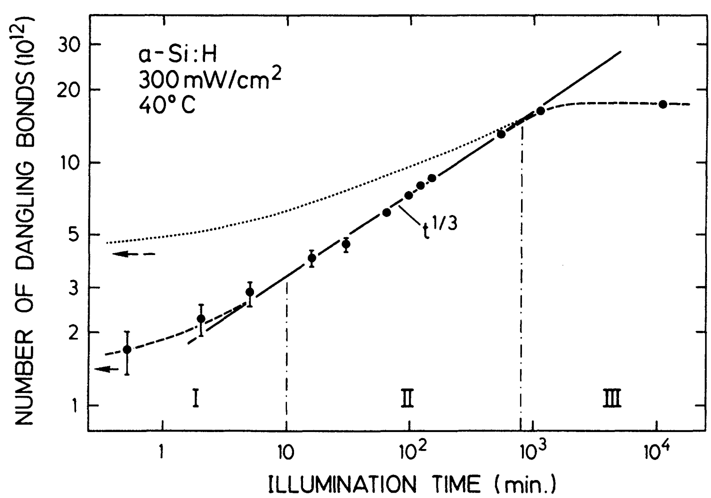

FIG. 30. Evolution of the dangling-bond defect density (areal density) in *a*-Si:H during illumination with white light of intensity 300mW/cm^2. Note the $t^{1/3}$ dependence for intermediate times and the subsequent saturation.

Light-induced metastable defect creation in *a*-Si:H is the major problem for applications of this material in photosensors and solar cells. The metastable defects act as additional recombination centers that decrease the photoconductivity (cf. Fig. 24) and, as a consequence, e.g., the efficiency of a solar cell using *a*-Si:H as the absorber material (see Sec. 4.1).

The microscopic origin of light-induced defect creation in *a*-Si:H is not yet clear. Both intrinsic (involving Si and H atoms only) and extrinsic mechanisms have been discussed. Figure 31 summarizes the most basic intrinsic relaxation processes that could give rise to metastable dangling-bond formation:

1. dissociation of a weak Si–Si bond due to electron–hole recombination;
2. disproportionation of tetrahedral Si atoms into a dangling bond and fivefold-coordinated atom ("floating bond");
3. metastable creation of neutral dangling bonds, D^0, by electronic charge exchange between charged defect pairs, D^+ and D^-;
4. Si–Si bond breaking stabilized by hydrogen bond switching;
5. creation of two metastable dangling bonds by photodissociation of a H_2^* complex;
6. photoinduced Si–H bond dissociation followed by migration and retrapping at an internal surface (e.g., in hydrogen-terminated voids).

An extrinsic origin of light-induced metastability has been suggested in connection with N, C, or O atoms, since only these impurities are present in device-quality *a*-Si:H with sufficient concentrations (10^{17}–10^{18} cm^{-3}, 10–100 ppm). Considerable efforts have been made to minimize light-induced degradation by producing ultrapure *a*-Si:H samples, but below a critical impurity content of typically 10^{20} cm^{-3}, no positive effect has been observed. As an example, Fig. 32 summarizes the dangling-bond densities in the annealed state, *A*, and after light soaking, *B*, as a function of controlled O or N doping of ultrapure *a*-Si:H. Below 10^{20} cm^{-3} neither O nor N contamination affects the degree of light-induced metastability. A similar observation holds for contamination with C. On the other hand, above a critical concentration of about 10^{20} cm^{-3}, all impurities tend to increase both the stable and the metastable defect density in *a*-Si:H, probably through an overall effect on the bonding properties of Si and H network atoms.

A second type of metastability with practical importance is observed in *a*-Si:H intentionally doped with electronically active shallow donors and acceptors (P, B, As, . . .). As already discussed in Sec. 2.2 in some detail, substitutional dopants are mostly incorporated in the growing amorphous film with threefold coordination (P_3,B_3). Partial thermodynamic equilibrium within the film during growth determines the fraction of fourfold-coordinated, electrically active dopants (P_4,B_4), which varies with doping level and is

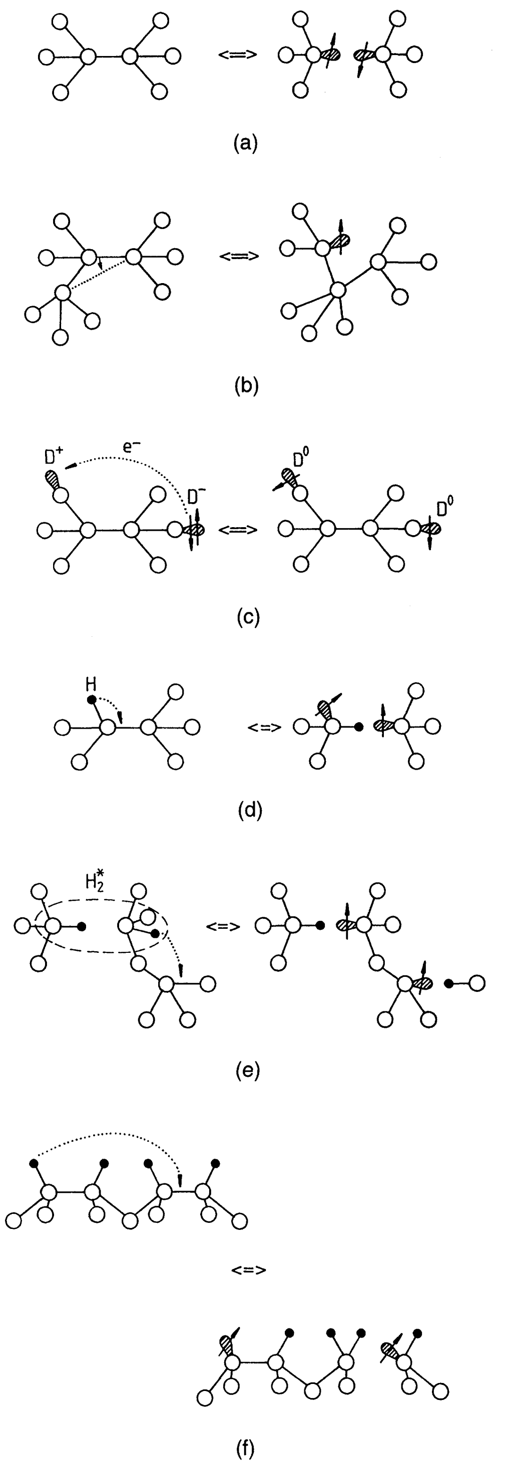

FIG. 31. Atomic models for possible metastability processes: (a) bond breaking, (b) valence alternation, (c) charge exchange, (d) hydrogen bond switching, (e) dissociation of H_2^* complexes, and (f) hydrogen migration on internal surfaces.

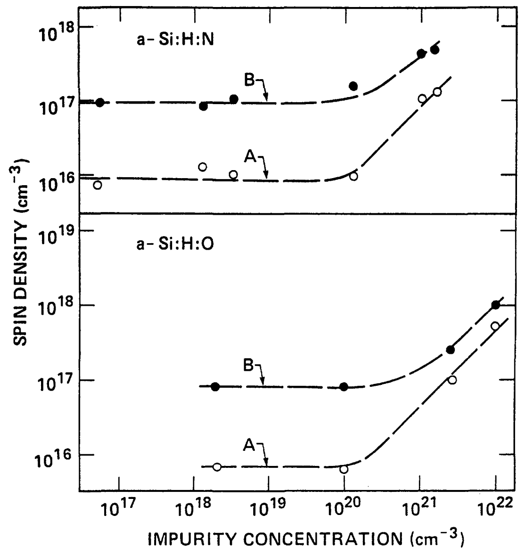

FIG. 32. Dependence of the defect density in the annealed state, curves A, and the light-soaked state, curves B, on the concentration of (a) nitrogen and (b) oxygen impurities.

typically $[P_4]/[P_3] \leq 10^{-1}$ (see Fig. 14). The fourfold dopants are most stable when ionized, and hence an approximately equal density of compensating charged dangling bonds are created (D^- in the case of P, D^+ in the case of B doping).

It is now observed that the densities of gap states and active dopants are only partially defined during the growth of the sample but are also controlled by postdeposition structural relaxation processes, which occur above a certain equilibration temperature T_E (Kakalios and Jackson, 1989). The dopant-defect system is in thermodynamic equilibrium above T_E, which is ≈130 °C for P-doped and $T_E \approx 90$ °C for B-doped *a*-Si:H. Rapid cooling to temperatures below T_E freezes in the atomic and electronic structure at T_E. The effect of fast versus slow cooling on the conductivity of *n*- and *p*-type *a*-Si:H is shown in Fig. 33. The temperature dependence of the dark conductivity for P- and B-doped *a*-Si:H below T_E is strongly dependent on the cooling rate, which has been varied in the case of Fig. 33 from 0.02 K/s (full lines) to 1 K/s (dashed lines). Rapid thermal quenching from temperatures above T_E (arrows in Fig. 33) results in a higher dark conductivity (i.e., a larger fraction of

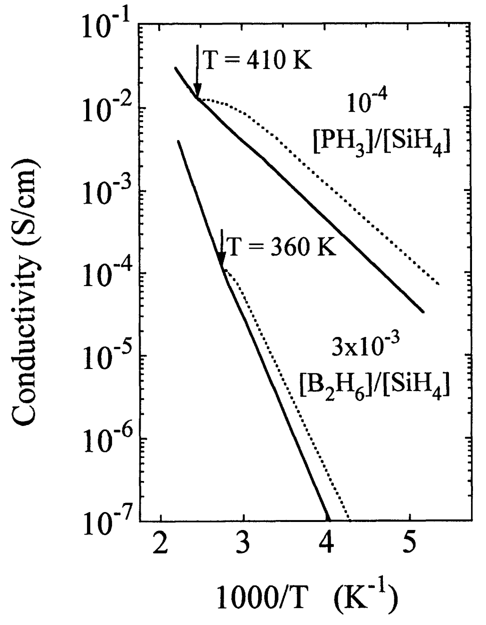

FIG. 33. Dark conductivity of *n*-type and *p*-type amorphous silicon versus temperature after slow cooling (solid curves) and rapid quenching (dashed curves). Arrows at 410 and 360 K indicate equilibration temperatures.

active dopants) than slow cooling, as expected for a thermally enhanced (endothermic) dopant activation reaction:

$$P_3^0 + Si_4^0 + \Delta E \Leftrightarrow P_4^+ + Si_3^-.$$

The frozen-in state obtained by rapid thermal quenching of phosphorus-doped *a*-Si:H below the equilibration temperature T_E is characterized by a higher active dopant concentration $[P_4^+]$. From thermodynamic arguments it is expected that the defect density Si_3^- also increases. However, up to now no significant changes of the defect density could be detected experimentally. Instead, the increase in active dopants leads to an increase of the band-tail carrier density $\Delta n_{BT} \sim \Delta P_4^+$. Analogous arguments are valid for $[B_4^-]$ in *p*-type *a*-Si:H doped with boron.

Experimentally, the density n_{BT} of shallow carriers in band-tail states at a constant temperature $T < T_E$ following rapid thermal quenching is found to decay according to a stretched-exponential relation

$$n_{BT}(t) = n_{BT}(0) \exp[-(t/\tau)^{\beta}].$$

Corresponding decay curves for phosphorus-doped *a*-Si:H ($T_E \approx 130$ °C) are shown in Fig. 34(a). The two parameters characterizing a stretched-exponential relaxation process are the dispersion parameter β and the time constant τ. The dispersion parameter β is given by $\beta = T/T_0$ with $T_0 \approx 580$ K for *n*- and *p*-type *a*-Si:H [Fig. 34(b)]. Physically, stretched-exponential relaxation occurs for a system with an exponential distribution of energy barriers, with kT_0 representing the characteristic decay parameter ($\approx$ spectral width) of the barrier distribution. The time constant τ of the relaxation process is related to the effective barrier height E_{eff} via a Boltzmann term:

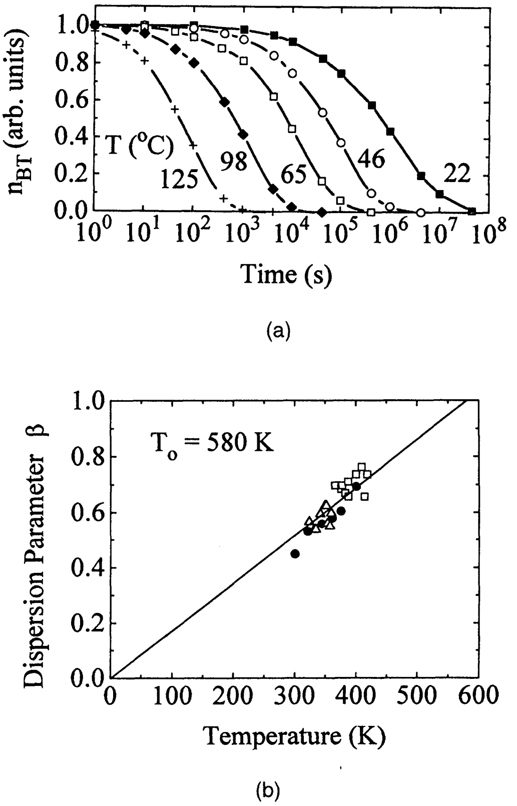

FIG. 34. (a) Stretched-exponential relaxation of shallow band-tail carriers in *n*-type samples at different annealing temperatures. (b) Temperature dependence of the dispersion parameter β obtained from the relaxation of conductivity in *n*-type and *p*-type samples (open squares and triangles, respectively) and from sweep-out measurements in *n*-type material (solid dots). The solid line indicates the expected behavior for a characteristic temperature T_0 of 580 K.

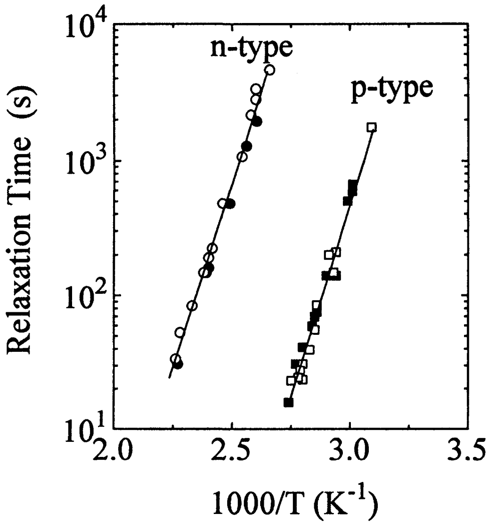

FIG. 35. Arrhenius plot of the relaxation time in doped amorphous silicon samples following rapid quenching.

$$\tau = \tau_0 \exp[E_{\text{eff}}/kT].$$

According to Fig. 35, both *n*- and *p*-type *a*-Si:H exhibit a relaxation behavior following thermal quenching, which can be described by an effective barrier of $E_{\text{eff}} \approx 1.1$ eV and a prefactor τ_0 between 10^{-12} and 10^{-10} s. Note that these values are very similar to those observed for the annealing of light-induced metastable dangling bonds in undoped *a*-Si:H described above. It is thus tempting to propose a common structural origin for metastability in both undoped and doped *a*-Si:H, namely, a reaction between lattice bonds, dangling-bond defects, and dopants that is mediated via the diffusion of hydrogen. Specific processes for doped *a*-Si:H, which are based on corresponding dopant-hydrogen interactions in crystalline Si, are shown in Fig. 36 (Chang and Chadi, 1988). Thus, substitutional boron is rendered electronically inactive by insertion of a hydrogen atom into one of the four Si–B bonds, while P_4^+ is transformed into the passive state by insertion of H into an antisite of Si ("back bond"). Upon heating, the hydrogen atom can migrate to a nearby weak Si–Si bond, where it becomes trapped by the formation of a Si–H bond and a charged dangling-bond defect. In the same process, the dopant atom becomes activated.

The involvement of hydrogen in *a*-Si:H metastability phenomena is a recurring theme in the recent literature. As a matter of fact, the participation of hydrogen in structural relaxation of *a*-Si:H at relatively low temperatures is quite reasonable in view of experimental results concerning hydrogen diffusion in *a*-Si:H. Hydrogen is found to diffuse in *a*-Si:H via a dispersive multiple-trapping process with an activation energy of about 1.4 eV and a diffusivity of about 10^{-20}–10^{-18} cm^2 s^{-1} at 400 K. With this diffusivity, hydrogen can move in the amorphous Si network over a distance of up to several tens of angstroms within one hour's time, a distance that is comparable to or less than the average distance between H atoms in the lattice. Thus, it is possible to view the hydrogen atoms as constituting a distinct sublattice in the *a*-Si:H structure, whose atoms become mobile above a certain temperature T_g. In analogy to network glasses, this temperature T_g has certain characteristics of a typical glass transition temperature: above T_g, the hydrogen atoms are sufficiently mobile to mediate a partial thermal equilibration among particular structural units such

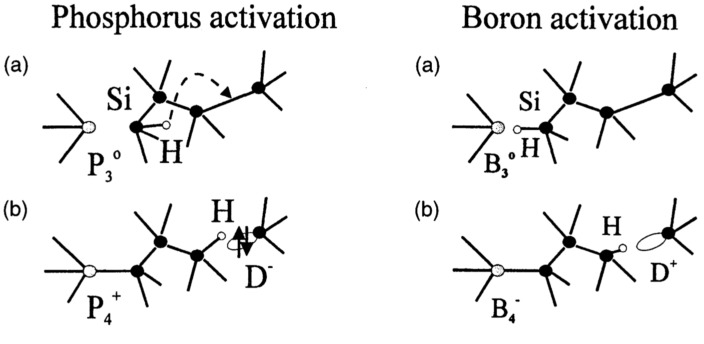

FIG. 36. Microscopic models for thermal activation of dopants via hydrogen migration in phosphorus-doped samples and boron-doped material: (a) describes the dopant passivation by H and (b) the active state of the dopant.

as dopants and defects; below T_g, even the H atoms as the most mobile structural species are frozen into their respective positions, and all relaxation processes come to a halt. In the literature this model is known as the hydrogen glass model (Kakalios and Jackson, 1989). The glass transition temperature T_g is assumed to be equal to the equilibration temperature T_E discussed in connection with Figs. 33 and 34.

3. ALLOYS AND MULTILAYERS

For many applications of a semiconductor it is desirable to have a certain control over the band-gap energy E_g. In the case of *a*-Si:H this is easily achieved by alloying with Ge, C, O, or N. In practice one adds reactive gases, such as GeH_4, CH_4, N_2O, CO_2, NH_3, etc., to the SiH_4 used for preparation of amorphous Si. As already shown in Fig. 32, alloying with nitrogen and oxygen quickly increases the density of structural defects in the material, so that substoichiometric SiO_x and SiN_x alloys play a negligible role for applications. Only wide-gap *a*-SiO_2:H and *a*-Si_3N_4:H close to stoichiometry are used as the dielectric gate insulator in *a*-Si:H thin-film transistors. For smaller variations of the band gap, better results have been obtained with isoelectronic *a*-$SiGe_x$:H and *a*-SiC_x:H alloys. As compiled in Fig. 37, alloying with Ge or C allows a variation of the optical band gap over the range $1 \text{ eV} \leq E_g \leq 2.5 \text{ eV}$, while maintaining relatively good transport properties, at least in the region $1.5 \text{ eV} < E_g < 2 \text{ eV}$. The rapid decrease of the photoresponse, σ_{ph}/σ_D, which is evident from Fig. 37 for increasing Ge or C content, is mainly caused by a corresponding increase of deep defects acting as recombination centers. Thus, the dangling-bond density increases from less than 10^{16} cm^{-3} to more than 10^{17} cm^{-3} when going from *a*-Si:H to *a*-Ge:H, and a similar trend holds for alloying with carbon. In addition, the electron mobility in *a*-$SiGe_x$:H and *a*-SiC_x:H is reduced from its value in *a*-Si:H because of additional carrier scattering by chemical disorder [cluster formation, preferential hydrogen bonding to Si or C, etc. (Wang and Schiff, 1993)]. According to Fig. 38, at the same time the hole mobility remains essentially unchanged, indicating that isoelectronic alloying with Ge and C mainly affects the density-of-states distribution close to the conduction-band mobility edge (alloy broadening).

Whereas the optoelectronic properties of pure hydrogenated amorphous silicon have been more or less optimized in the course of the last 20 years, the same is obviously not the case for all possible *a*-Si:H–related alloys. Mixing of reactive gases in a glow-discharge reactor under otherwise unchanged deposition conditions is only the zero-order approach to a very complex problem. In particular, the work on *a*-Ge:H and *a*-C:H has shown that optimized deposition conditions for these amorphous semiconductors are quite different from those used for *a*-Si:H deposition, leaving a lot of room for the improvement of *a*-Si:H–based alloy materials. The need for variable–band-gap material in amorphous silicon solar cells will certainly motivate future work in this interesting area.

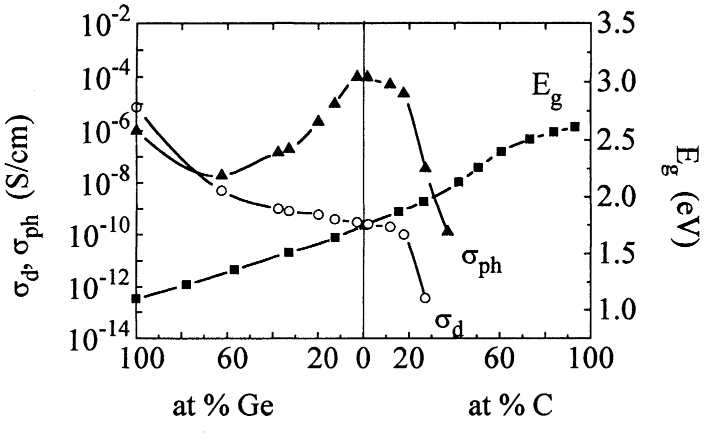

FIG. 37. Dark conductivity (circles), photoconductivity (triangles), and optical band gap (squares, E_g), in silicon–germanium and silicon–carbon alloys as a function of alloy composition.

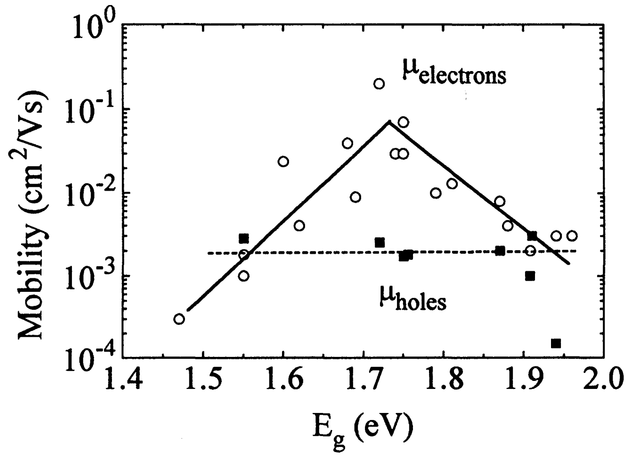

FIG. 38. Dependence of electron and hole mobilities in silicon-related amorphous alloys on the optical band gap, E_g.

In addition to alloying, attempts have also been made to influence the optical and electronic properties of *a*-Si:H–based thin films via the preparation of compositional or doping multilayers. In the field of amorphous semiconductors, the term "multilayer" is preferred over the term "superlattice," since there are no periodic lattices to begin with. Because the deposition of *a*-Si:H and related materials by plasma-enhanced CVD proceeds with a growth rate of typically 1 Å/s, it is fairly easy to produce atomically abrupt interfaces by a corresponding temporal modulation of the reactive-gas partial pressures in the deposition reactor. Indeed, multilayers with periods between 5 and 500 Å and satisfactory long-range periodicity have been routinely produced in this fashion.

The quality of the multilayer structure can be monitored either by direct high-resolution microscopy or by scattering experiments. In the case of compositional superlattices (*a*-Si:H alternating with *a*-$SiGe_x$:H, *a*-SiN_x:H, *a*-SiC_x:H, etc.), combined microscopy and x-ray scattering investigations have shown a high multilayer quality in terms of atomically abrupt interfaces, good periodicity in growth direction, and lateral homogeneity.

Direct evidence for superlattice effects in periodic compositional *a*-Si:H/*a*-SiN_x:H multilayers also comes from Raman-scattering experiments. As a result of the artificial periodicity of the multilayer, the phonon dispersion of long-wavelength acoustic phonons with k vectors parallel to the growth direction is altered by folding back of the approximately linear phonon branches at the superlattice Brillouin-zone boundary $k_z = \pi/L$, where L is the period length of the multilayer. As a consequence, k conservation ($\Delta k \approx 0$) in Raman backscattering is fulfilled not only by the long-wavelength optical phonons near the Brillouin-zone center but also by a ladder of back-folded acoustic-phonon states, ω_n and ω_{-n}, where n indicates the order of folding and the sign describes branches with positive or negative dispersion, respectively ($\partial\omega/\partial k > 0$ or < 0). In the limit of a linear dispersion relation for long-wavelength acoustic phonons, one has $\omega_n \approx \omega_{-n} \approx 2n\nu\pi/L$ (ν is sound velocity), so that the Raman-active $k \approx 0$ acoustic-phonon frequencies should be proportional to $1/L$. The experimental data in Fig. 39 demonstrate that this is indeed the case for $n = 1$.

In contrast to the obvious effects of multilayer formation on the phonon properties of *a*-Si:H–based systems, there is little convincing evidence for pronounced quantization phenomena in the electronic properties of the same kind of multilayers. For example, confinement of electrons and holes in the *a*-Si:H layer of *a*-Si:H/*a*-Si_3N_4:H superlattices should lead to the same quantization condition,

$$E_n = \frac{h^2}{2m^*}\left(\frac{n\pi}{L}\right)^2,$$

for the electron or hole levels as in crystalline quantum wells. Corresponding experimental results in Fig. 40 indeed exhibit a

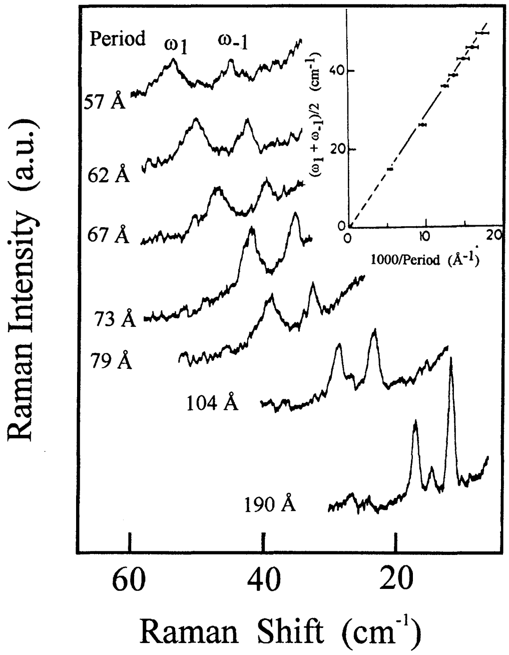

FIG. 39. Raman spectra of *a*-Si/Si$_3$N$_4$:H multilayers with different periods, showing the first and second phonon peaks that become Raman active through back-folding of the acoustic-phonon branches. Inset: The expected linear dependence on inverse layer period. (After Santos *et al.*, 1986.)

trend toward higher band gaps with decreasing *L*.

On the other hand, the absorption spectra in Fig. 40 remain essentially featureless even for very small well widths, in contrast to the expected steplike density-of-states distribution for two-dimensional systems. It has thus also been proposed that a considerable contribution to the band-gap opening in short-period amorphous multilayers is due to alloying effects and other changes in the *a*-Si:H–layer composition.

Evidence for quantum confinement effects in short-period *a*-Si:H–based multilayers and heterostructures has also been claimed on the basis of other experiments, such as resonant tunneling in double-barrier structures or thermally modulated absorption. So far, many of the results claiming evidence for quantum confinement effects lead to conflicting results concerning the effective mass m^* of confined carriers [$m^* \approx (0.3\text{–}0.6)m_0$ for electrons, $m^* \approx m_0$ for holes]. Also, conceptually it is difficult to reconcile quantum confinement effects in 50-Å-period multilayers with the mean free path for carriers in *a*-Si:H of ≈10 Å. Quantum effects require a complete phase coherence of the confined wave function over the entire well width. In *a*-Si:H this coherence is quickly lost by the strong disorder scattering. Nevertheless, it may be possible that a small fraction of confined carriers experiences less disorder scattering and can actually feel the influence of confining potential barriers.

The second type of multilayers comprises the doping superlattices (*n-i-n-i*, *n-i-p-i*, etc.), which are obtained by a periodic modulation of the degree and/or the type of doping during growth. The energy-band diagram of such an *n-i-p-i* structure in its equilibrium

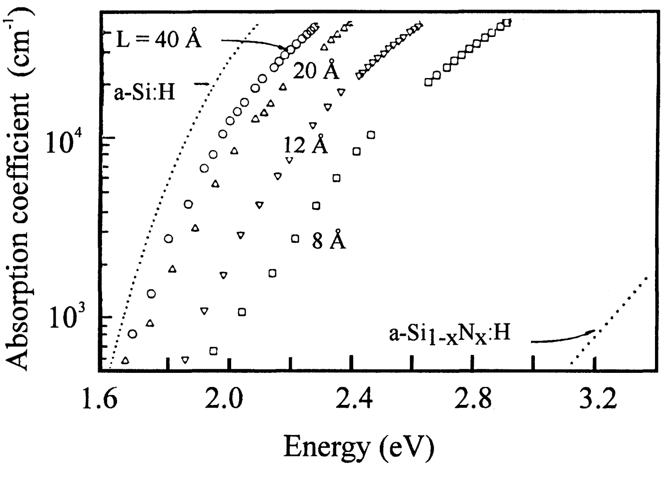

FIG. 40. Absorption coefficients of *a*-Si/Si$_3$N$_4$:H multilayers with small periods *L* between 40 and 8 Å (after Abeles and Tiedje, 1983).

state is shown in the top part of Fig. 41 by the solid lines. The periodic alternation between *n*- and *p*-type layers leads to a corresponding modulation of the band structure in growth direction due to the transfer of electrons from *n* to *p* layers. Upon illumination of such a *nipi* structure, photoexcited carriers are separated by the internal fields. Electrons and holes then accumulate in their corresponding majority-carrier regions (electron in *n* layers, holes in *p* layers), causing a flattening of the potential modulation as indicated by the dashed curves in the top part of Fig. 41. Recombination between photoexcited carriers is thus hindered by the spatial separation between electrons and holes, leading to a phenomenon known as persistent photoconductivity: because of the increased recombination time of spatially separated electron and hole populations, the photoconductivity perpendicular to the modulation direction is strongly enhanced and may even persist for very long times in long-period multilayers at low temperatures. An example of such a persistent photoconductivity in an *a*-Si:H doping multilayer is given in the lower part of Fig. 41. Following optical excitation at $t = 0$, the photoconductivity quickly decays at short times but then reaches an almost constant level, far above the dark conductivity, that persists for minutes to hours, depending on temperature and sample structure. In fact, the extremely long persistence of elevated conductivities in *a*-Si:H doping superlattices has caused a lot of doubt on the applicability of the standard persistent-photoconductivity model outlined above. Even in the absence of thermal excitation, carriers in adjacent *n* and *p* layers should recombine rather rapidly via tunneling. An alternative model for the extremely persisting conductivity in *a*-Si:H *nipi* structures is based on the pronounced structural metastability phenomena in this material (cf. Sec. 2.5). It is known from investigations of bulk doped *a*-Si:H that illumination leads to a light-induced metastable activation of dopant atoms that, however, has different efficiencies in *n*- and *p*-type samples. As a consequence, illumination of *a*-Si:H doping multilayer structures will cause a different activation of dopants in *n* and *p* layers, a situation that may resemble a persistent photoconductivity. In addition, light-induced creation of metastable dangling bonds is known to be differently efficient in *n*- versus *p*-type *a*-Si:H, an effect that may also contribute to a persistent conductivity increase after illumination.

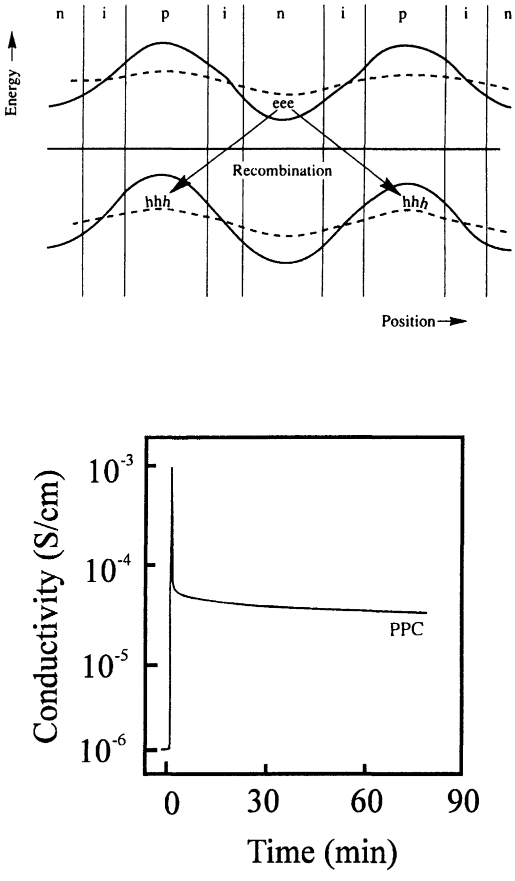

FIG. 41. Schematic band diagram of a *nipi* structure in thermal equilibrium (solid lines) and under illumination (dashed lines). The lower part shows the persistent photoconductivity in such a structure following a short light pulse. (After Kakalios, 1986.)

4. DEVICES

4.1 Overview

The importance of state-of-the-art hydrogenated amorphous silicon for device applications is based on a number of advantages compared to other semiconductors. The most relevant advantages are the following:

- low-temperature processing (≤400 °C)

- easy and uniform preparation over large areas
- electronic quality independent of substrate material
- compatibility with crystalline Si technology
- low defect density (comparable to bulk crystalline GaAs)
- high absorption coefficient
- easy transformation in crystalline state with local selectivity
- easy alloying with other atoms for bandgap variation

Of course there are also a number of deficiencies, such as

- low structural stability under electronic excitation
- low doping efficiency
- low carrier mobility

Nevertheless, amorphous Si has been used successfully in a number of electronic devices listed below and partly described in more detail in the following sections. Among *a*-Si:H–based devices one notably finds

- solar cells, photodetectors
- thin-film transistor arrays
- spatial light modulators
- photocopiers, image sensors, scanners
- large-area light-emitting diodes
- sensors
- memory circuits

and others. A good review of these applications can be found in the books edited by Pankove (1984) and Kanicki (1992). Here, we only discuss briefly the first two kinds of devices, namely, solar cells and thin-film transistors, which at present dominate commercially available *a*-Si:H products.

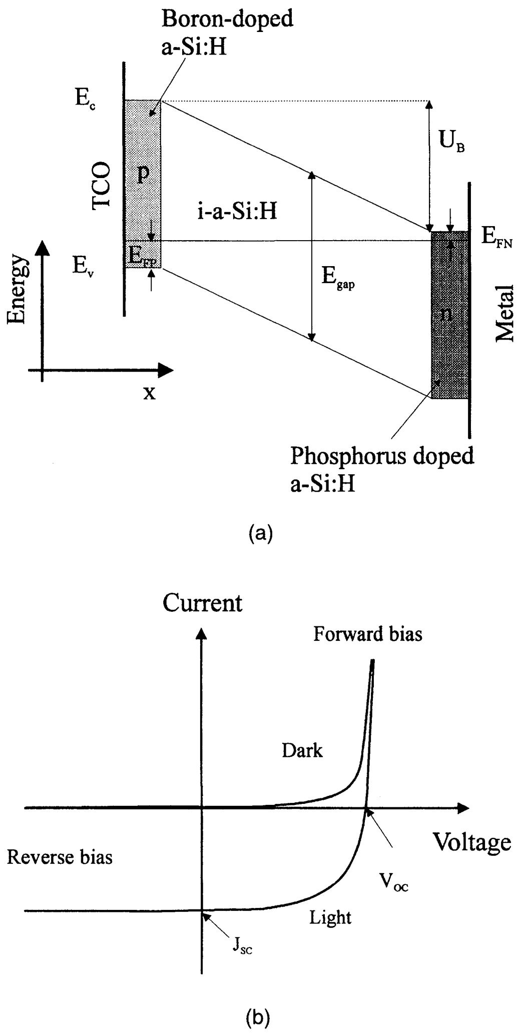

FIG. 42. (a) Energy band diagram of an *a*-Si:H *p-i-n* photosensor (E_{gap}: optical gap; E_{FP} and E_{FN}: Fermi level of holes and electrons in the *p*-type and *n*-type contact layers; U_B: built-in voltage). (b) Current-voltage characteristics of the *p-i-n* diode in (a) under dark and light conditions (I_{SC}: short circuit current; V_{OC}: open circuit voltage).

4.2 Solar Cells and Photodetectors

The main advantages of amorphous silicon as a solar-cell material are the high optical absorption coefficients in the visible spectral range, allowing the use of thin-film technology (cf. Fig. 8), and the possibility to prepare such films on large-area, low-cost substrates and at relatively low temperatures (see PHOTOVOLTAIC DEVICES; SOLAR ENERGY). A typical solar-cell structure used today is shown in Fig. 42(a). Light is absorbed in an intrinsic (undoped) *a*-Si:H layer of thickness 200–800 nm. The field necessary to separate photoexcited electrons and holes is generated by thin, highly doped *n*- and *p*-type layers (≈10 nm). The whole structure is referred to as a *p-i-n* diode. Electrical contact to the diode is made, for example, by a transparent conducting oxide (TCO) on the *p* side and by a metal such as Al or Ag on the *n* side, which also serves as a back-reflector to increase the effective optical path lengths for weakly absorbed light. In addition, surface texture of the substrate material or the TCO on the micron scale is used for light trapping in the solar-cell device. As substrate material, low-cost lime glass is generally employed, but cells deposited on stainless steel foil or plastic substrates have also been developed.

Because holes in *a*-Si:H have a much lower room-temperature mobility than electrons (see Fig. 12), it is necessary to introduce the light via the *p*-type contact, in order to minimize the distance holes have to drift in the built-in field toward the *p* contact. In more advanced cell designs, one also replaces the *p*-type *a*-Si:H layer by *p*-type *a*-SiC:H alloy layers with a wider optical gap. This decreases the absorption of blue light in the *p* contact. At the same time, the conduction-band offset between *a*-SiC:H and *a*-Si:H serves as a back-diffusion barrier, preventing electrons from recombining in the *p* contact. Further improvement of carrier collection can be achieved by using microcrystalline Si with a higher conductivity than *a*-Si:H as the *n*-type contact.

The electrical characteristics of a *p-i-n* solar-cell structure can be easily understood from the *I-U* curves shown in Fig. 42(b). In the dark, the solar cell behaves like a diode, with a low thermal or leakage current in reverse bias, and a high, exponentially increasing injection current in forward bias. Under illumination, the reverse-bias current increases by the amount of photoexcited and collected carriers. For an incident solar light intensity of 100 mW/cm^2, good *a*-Si:H cells obtain short-circuit currents of $I_{SC} \approx 17$ mA/cm^2. Typical values for the open-circuit voltage V_{OC} in *a*-Si:H solar cells fall between 0.85 and 0.95 V.

For a single *p-i-n* device as shown in Fig. 42(a), the theoretical limit for the conversion efficiency has been estimated to be about 14–15%. In state-of-the-art cells, values of $\eta \approx 12\%$ have been reached, at least in small prototype cells with a total area of about 1 cm^2.

For commercially produced *a*-Si:H solar cells, the simple *p-i-n* structure of Fig. 42(a) has been abandoned in favor of a more complex tandem or triple stack design, where two or three *p-i-n* devices are deposited on top of each other. This has mainly two reasons: For one, stacked cells allow the combination of two or three absorbing layers with different band gaps. Variation of the band gap is achieved by alloying with carbon or germanium (see Sec. 3). The advantage of variable–band-gap stacked cells is a much larger maximum conversion efficiency than for single cells. Estimates indicate that ultimately as much as 25% efficiency could be achieved. An example of such a tandem cell is shown (inverted) in Fig. 43. The device consists of a thin top cell made out of *a*-Si:H and a thicker bottom cell made out of a suitable amorphous SiGe alloy. High-energy photons are absorbed in the top cell, whereas lower-energy photons are absorbed in the bottom cell, which has a smaller optical band gap. The main design parameters of such tandem arrangements are the band gaps and the thicknesses of the two absorber layers. An optimized two-terminal structure has to fulfill the condition of current matching; i.e., the current generated by the top cell has to be equal to that generated by the bottom cell. This is generally not possible for all operation situations encountered in real applications. Another possibility is to isolate top and bottom cells from each other electrically (four-terminal device), but this is too expensive for large-area power generation. On the other hand, four-terminal tandem structures have been used as sensitive color sensors.

A second reason for the preference given to tandem structures compared to single *p-i-n* diodes is the light-induced degradation

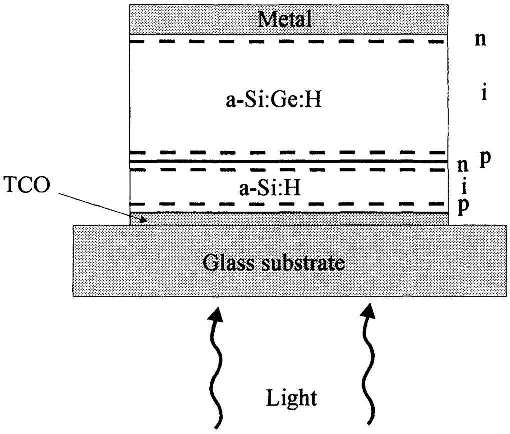

FIG. 43. Layer structure of an *a*-Si:H/*a*-SiGe:H tandem solar cell.

of *a*-Si:H solar cells, which can be linked directly to the light-induced metastable defect creation discussed in Sec. 2.5. The creation of additional defects in the absorber layer causes an increased recombination of photogenerated carriers and a change of internal electric fields. As a result, all three device parameters (U_{OC}, I_{SC}, and *FF*) are decreased during degradation (see Fig. 44), thus strongly affecting the conversion efficiency of the cell. The light-induced degradation of *a*-Si:H solar cells is at present the main hindrance for large-scale applications of such devices, since it decreases the cell efficiency from about 12% to about 8% after long-term exposure to intense sunlight. Since according to Sec. 2.5 the main reason for metastable defect creation in undoped *a*-Si:H is direct nonradiative recombination of photoexcited carriers, photoinduced degradation is particularly severe in devices with a bad collection efficiency. The efficiency of carrier collection is low in regions of the absorber layer with a low electric field or further away from the nearest contact layer than the typical carrier drift length, $\mu\tau F$. As a consequence, cell degradation is much more pronounced in thick cells (*i*-layer thickness > 500 nm) than in thin cells, since in the latter the field $F \approx U_B/d$ (U_B: built-in voltage; *d*: layer thickness) can be maintained at a sufficiently high value to prevent carrier recombination. Thus, a practical way to produce relatively stable solar cells is to divide the total absorber thickness into two separate layers of a *p-i-n-p-i-n* tandem structure. In this way, a compromise between good overall absorbance and good stability can be achieved. Such tandem cells have produced a conversion efficiency around 10% even in their degraded state. Similar efficiencies have also been demonstrated for large-area *a*-Si:H cells (0.1–0.5 m^2), whereas present mass-production modules have lower efficiencies of about 6–8%.

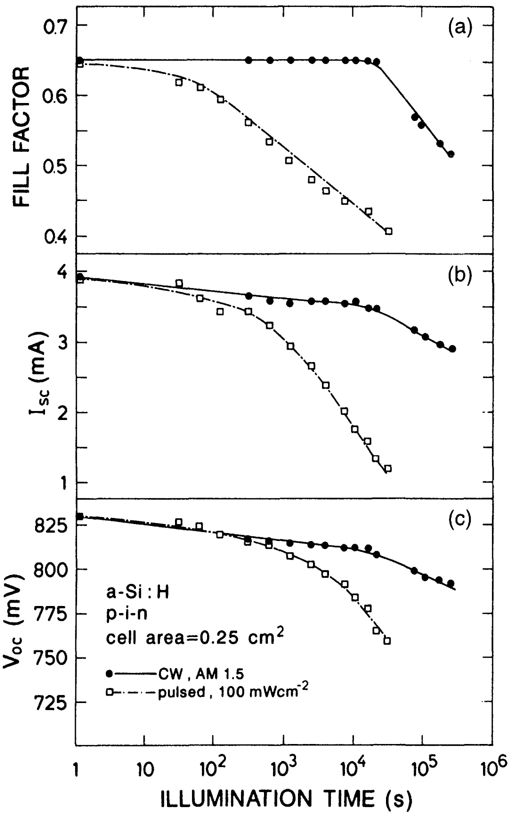

FIG. 44. Light-induced degradation of the main solar-cell parameters for a *p-i-n* cell exposed to 100 mW/cm^2 white light irradiation. Open symbols show accelerated degradation of the same cell using pulsed light from a flash lamp.

The same kind of *p-i-n* structure employed in *a*-Si:H solar cells is also used in various kinds of photodetectors or radiation detectors. Depending on the application desired, the *p-i-n* diodes are either laterally structured as small pixels with dimensions of about 50 × 50 μm^2 for high-resolution scanners or imaging arrays or produced with much larger *i*-layer thicknesses ($\approx$50 μm) for direct detection of ionizing radiation.

4.3 Field-Effect Transistors

In addition to applications as a photodetector, *a*-Si:H is widely used for active thin-film transistor (TFT) devices (see TRANSISTORS, FIELD EFFECT), in particular in the form of active-matrix liquid-crystal displays for portable computers. The basic structure of an *a*-Si:H TFT is shown in Fig. 45. The active layer consists of a thin (<100 nm) undoped *a*-Si:H layer, which is contacted via Ohmic metal/n^+-doped *a*-Si:H layers to form the source and drain contacts. The active layer is isolated from the metal gate by a suitable *a*-Si_3N_4:H or *a*-SiO_2:H dielectric layer produced by plasma-CVD. TFT structures used today have a channel length (source–drain distance) between 10 and 100

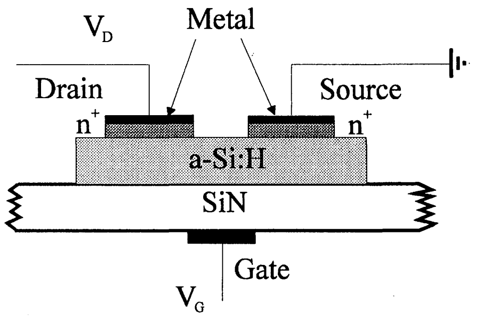

FIG. 45. Schematic design of an *a*-Si:H thin-film field-effect transistor. n^+ indicates heavily phosphorus-doped contact layers; SiN is the silicon–nitride dielectric layer.

μm, depending on the switching speed and maximum source–drain current desired. To a good approximation, the source–drain current can be written as

$$I_{SD} = \mu C_G \frac{W}{L} U_{DS}(U_G - U_{thr} - U_{DS}),$$

where μ is the carrier mobility in the channel, C_G the gate capacitance per unit area, W the gate width, and U_{thr} a threshold voltage, which for *a*-Si:H has a value of approximately 1 V. Typical source–drain currents of *a*-Si:H TFTs are 10^{-11} Å in the "off" state ($U_G = 0$) and 10^{-5} Å in the "on" state ($U_G = 10$ V), both for a source–drain voltage $U_{DS} = 10$ V.

The combination of many such small TFT structures to form a sensor or display array is shown in Fig. 46. Each sensor element or display pixel is addressed by its own TFT. The gate and data (drain) contacts of the TFTs are patterned as indicated schematically in Fig. 46(a) in the form of a gate- and data-line matrix. For displays, the TFTs are turned on to apply a voltage bias to a liquid-crystal element acting as a capacitor in the electric circuit. The TFT has to be able to supply sufficient charge to the electrodes of the liquid-crystal pixel during one frame time to turn the pixel on or off. In sensor arrays, the sensor element (e.g., an *a*-Si:H *p-i-n* or Schottky photodiode) acts as a variable light-dependent resistor that is connected to an external current sensor by turning the TFT on. A micrograph of such a sensor array is shown in Fig. 46(b). Modern TFT array structures for applications as TV screens or computer displays contain typically a million TFTs on an area of several hundred square centimeters and are presently produced at a level of about 100 000 displays per year. Both the size and the production volume of such displays are expected to experience considerable growth in the near future.

5. SUMMARY AND OUTLOOK

In addition to monocrystalline, multicrystalline, microcrystalline, and more recently porous Si, amorphous silicon is a modifica-

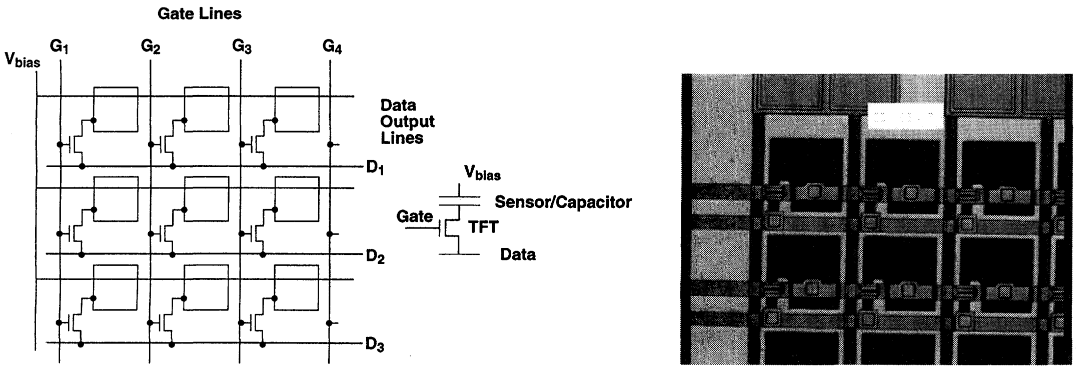

(a)

(b)

FIG. 46. (a) TFT matrix array used for liquid-crystal displays and for two-dimensional high-resolution image sensors. (b) Electron micrograph of a real device using the basic design of (a). The pixel size is 100 × 100 μm^2.

tion of solid silicon with interesting properties for application as an electronic material. The complete absence of long-range order in amorphous silicon gives rise to a number of specific properties that are qualitatively different from those of any other form of Si. This is particularly noticeable in the optical absorption of visible light, which is stronger than in all other forms of silicon and justifies the interest in *a*-Si:H thin-film photosensors and solar cells. Amorphous Si can be prepared with low defect densities over large areas and on many different types of substrates, which is a prerequisite for most present applications. Despite its large-area capability, amorphous Si can be structured by conventional lithographic techniques, thus providing a unique possibility to combine macroelectronics and microelectronics. Of similar importance for applications is the fact that *a*-Si:H can be doped substitutionally both *n* and *p* type and that the band gap of *a*-Si:H can be tuned over a large range by alloying with Ge, C, O, or N. There are, however, also a number of negative features that very often reduce the potential of *a*-Si:H for high-performance devices. In particular, the low carrier mobility, the low doping efficiency, and the various forms of metastability pose severe limitations on present applications. In many cases it would therefore be desirable to combine the good properties of amorphous Si with those of crystalline Si. Activities in this direction are presently pursued by a number of groups, e.g., using laser recrystallization, preferential etching, or plasma deposition of mixed amorphous–microcrystalline Si thin films.

GLOSSARY

***a*-Si:** Amorphous silicon.

***a*-Si:H:** Hydrogenated amorphous silicon.

***a*-SiGe:H:** Hydrogenated amorphous silicon alloyed with Ge.

***a*-SiC:H:** Hydrogenated amorphous silicon alloyed with C.

***a*-SiN:H:** Hydrogenated amorphous silicon alloyed with N.

***a*-SiO:H:** Hydrogenated amorphous silicon alloyed with O.

***a*-Ge:H:** Hydrogenated amorphous germanium.

***a*-C:H:** Hydrogenated amorphous carbon.

Band-Tail States: Localized states at the band edge of the conduction and valence band.

CB: Conduction band.

Correlation Energy: Relative energy position of the singly and doubly occupied charge states of a defect.

CPM: Constant photocurrent method.

DB: Dangling bond.

Demarcation Energy: Energy that separates states in thermal equilibrium with band states from states that are not yet equilibrated.

Dispersive Transport: Decreasing mobility of carriers with increasing time due to interaction with localized states.

DOS: Density-of-states distribution.

ESR: Electron spin resonance.

Geminate Recombination: Recombination of an electron–hole pair generated by the same photon.

Hyperfine Structure: Splitting of the ESR spectrum due to interaction between electronic spins and nuclear magnetic moments.

IR Absorption: Infrared absorption.

Light Soaking: Prolonged exposure to intense light.

LO(A): Longitudinal optical (acoustic) phonon.

LVM: Local vibrational mode.

Metastability: Spontaneous reversible creation of additional defects and/or dopants under strong electronic excitation.

PDS: Photothermal deflection spectroscopy.

PECVD: Plasma-enhanced chemical vapor deposition.

PL: Photoluminescence.

Photo-CVD: Optical excitation of reactive gas.

***p-i-n*:** Boron-doped (*p*), intrinsic (*i*), and phosphorus (*n*) layer sequence.

SDPC: Spin-dependent photoconductivity.

Staebler–Wronski Effect: Light-induced degradation.

Statistical Shift: Temperature-induced shift of the Fermi level.

TCO: Transparent conductive oxide.

TFT: Thin-film transistor.

Thermalization: Energy relaxation of excited carriers.

Trapping: Capture of electrons or holes into localized band-tail states.

TO(A): Transverse optical (acoustic) phonon.

Urbach Tail: Exponential absorption tail below the band-gap energy.

Urbach Slope: Characteristic energy of the exponential absorption.

Van de Pauw Contact Geometry: Ohmic contacts on the perimeter of a flat sample of arbitrary shape.

VB: Valence band.

Works Cited

Abeles, B., Tiedje, T. (1983), *Phys. Rev. Lett.* **51,** 2003–2007.

Beyer, W. (1985), in: D. Adler, H. Fritzsche (Eds.), *Tetrahedrally Bonded Amorphous Semiconductors,* New York: Plenum, p. 129.

Brandt, M. S., Stutzmann, M. (1991), *Phys. Rev. B* **43,** 5184–5187.

Cardona, M. (1983), *Phys. Status Solidi B* **118,** 463–481.

Chang, K. J., Chadi, D. J. (1988), *Phys. Rev. Lett.* **60,** 1422–1425.

Dersch, H., Schweitzer, L., Stuke, J. (1983), *Phys. Rev. B* **28,** 4678–4687.

Kakalios, J. (1986), *Phil. Mag. B* **54,** 199–218.

Kakalios, J., Jackson, W. B. (1989), in: H. Fritzsche (Ed.), *Amorphous Silicon and Related Materials,* Singapore: World Scientific, p. 207.

Kanicki, J. (Ed.) (1992), *Amorphous and Microcrystalline Semiconductor Devices,* Boston: Artech.

Lannin, J. S. (1984), *Semicond. Semimetals* **21B,** 159–195.

Ley, L. (1984), in: J. D. Joannopoulos, G. Lucovsky (Eds.), *The Physics of Hydrogenated Amorphous Silicon II,* Berlin: Springer, p. 61.

Marianer, S., Shklovskii, B. I. (1992), *Phys. Rev. B* **46,** 13100–13103.

Mott, N. F. (1987), *J. Phys. C* **20,** 3075–3102.

Nebel, C. E., Street, R. A. (1993), *Int. J. Mod. Phys. B* **7,** 1207–1258.

Overhof, H., Thomas, P. (1989), *Electronic Transport in Hydrogenated Amorphous Semiconductors,* Berlin: Springer.

Pankove, J. I. (Ed.) (1984), *Hydrogenated Amorphous Silicon, Part D: Device Applications,* Semicond. & Semimetals 21D, New York: Academic Press.

Santos, P., Hundhausen, M., Ley, L. (1986), *Phys. Rev. B* **33,** 1516–1518.

Spear, W. E. (1977), *Adv. Phys.* **26,** 811–845.

Staebler, D. L., Wronski, C. R. (1980), *J. Appl. Phys.* **51,** 3262–3268.

Street, R. A. (1984), *Semicond. Semimetals* **21B,** 197–244.

Street, R. A. (1985), *J. Non-Cryst. Solids* **77 & 78,** 1–16.

Street, R. A., Knights, J. C., Biegelsen, D. K. (1978), *Phys. Rev. B* **18,** 1880.

Stutzmann, M., Biegelsen, D. K., Street, R. A. (1987), *Phys. Rev. B* **35,** 5666–5701.

Stutzmann, M. (1992), in: J. Kanicki (Ed.), *Amorphous & Microcrystalline Semiconductor Devices* Vol. II, Boston: Artech House, p. 129.

Stutzmann, M., Biegelsen, D. K. (1988), in: H. Fritzsche (Ed.), *Amorphous Silicon and Related Materials,* Singapore: World Scientific, p. 557.

Tauc, J. (1974), in: J. Tauc (Ed.), *Amorphous and Liquid Semiconductors,* Berlin: Springer, p. 225.

Tiedje, T. (1984), in: J. D. Joannopoulos, G. Lucovsky (Eds.), *The Physics of Hydrogenated Amorphous Silicon II,* Berlin: Springer, p. 261.

Wang, Q., Schiff, W. A. (1993), *Mater. Res. Soc. Symp. Proc.* **297,** 419–425.

Further Reading

Mott, N. F., Davis, E. A. (1979), *Electronic Processes in Noncrystalline Materials,* Oxford: University Press.

Street, R. A. (1991), *Hydrogenated Amorphous Silicon,* Cambridge: University Press.

SILICON, CRYSTALLINE

SOKRATES T. PANTELIDES, *Department of Physics and Astronomy, Vanderbilt University, Nashville, Tennessee, U.S.A.*

Introduction 187
1. **Atomic Arrangements** 188
1.1 The Bulk Perfect Crystal 188
1.2 Crystalline Surfaces, Interfaces 189
1.3 Plane and Line Defects 190
1.4 Point Defects 192
1.5 Polycrystalline, Amorphous, and Porous Si 192
2. **Vibrational Properties—Phonons** 193
3. **Electronic Structure and Properties** 194
3.1 The Perfect-Crystal Energy Bands 194
3.2 Optical and Dielectric Response 196
3.3 Dopant Impurities 198
3.4 Heavy Doping—The Degenerate Limit 199
3.5 Lattice Point Defects and Deep Impurities 199
3.6 Surfaces 205
4. **Transport Properties** 206
5. **Thermal Properties** 208
6. **Atomic Diffusion** 208
7. **Mechanical Properties** 209
8. **Crystal Growth** 210
8.1 Bulk Crystal 210
8.2 Epitaxial Thin Films 210
8.3 Silicon-On-Insulator 211
9. **Processing** 211
9.1 Oxidation 212
9.2 Etching 212
9.3 Process Modelling 212
10. **Silicon-Based Microelectronics** 213
Glossary 213
Works Cited 213
Further Reading 215

INTRODUCTION

Silicon is one of the most abundant elements in nature. After oxygen, it is the most abundant element making up the earth's crust (28% by weight) in the form of various compounds. Sand is essentially pure silicon dioxide. Lava from volcanoes consists primarily of molten silicon dioxide. Silicon is also found in many animals and plants. The element silicon (Si, atomic number 14, atomic weight 28.0855) was first isolated by the Swedish chemist Baron Jöns Jacob Berzelius in 1824. In its pure form it is a crystalline solid with a metallic sheen. It also exists in an amorphous form. The melting point of crystalline Si is 1410 °C, and its boiling point is 2355 °C. At 20 °C crystalline Si has a density of 2.33 g/cm^3.

Silicon compounds have long been used in many glasses and ceramics and as ingredients in steel and other alloys. Si is also the main element in synthetic compounds known as silicones, which are a cross between organic and inorganic materials and are used as insulators, lubricants, water repellents, and implants in the human body. Above all, however, *crystalline silicon* is the heart and soul of modern microelectronics that, in turn, is the heart and soul of modern computers, the revolution in communications, information processing, and automation in factories, offices, and homes. This Encyclopedia contains articles on many aspects of silicon-based technology. It also contains articles on amorphous silicon and on several silicon-based compounds and structures. The present article focuses exclusively on crystalline silicon and its physical properties.

Crystalline silicon is a semiconductor, a class of materials that falls somewhere between metals, which have high electrical conductivity, and insulators, which do not conduct electricity. Semiconductors have

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

unique electrical properties that can be customized by the controlled incorporation of certain kinds of impurities.

The electrical properties of crystals were explained by quantum mechanics in the early 1930s. Electrons occupy energy levels that form wide bands. In a metal, the highest-energy band containing electrons is only partially filled, up to an energy called the Fermi energy. Externally applied electric fields can set up electrical currents by pushing electrons near the Fermi energy to adjacent empty energy levels. In insulators, on the other hand, the highest-energy band containing electrons, known as the valence band, is completely filled at $T = 0$ K (absolute zero of the temperature scale) and separated from the next band, known as the conduction band, by a forbidden gap. No conduction is possible when all bands are either completely full or completely empty simply because no energy can be imparted to the electronic system by applying moderate voltages. At finite temperatures some electrons are excited from the valence band into the conduction band so that some conduction is possible. For true insulators at room temperature, such conduction is negligible.

Semiconductors are essentially insulators with fairly small forbidden energy gaps. In pure crystalline forms, semiconductors would be fairly useless materials because they are poor insulators (a significant number of electrons are excited in the conduction band at room temperature) and poor conductors (the number of electrons in the conduction band at room temperature is not large enough to yield conductivity comparable to that of metals).

The technological significance of semiconductors arises primarily from two reasons:

1. Unlike metals, semiconductors exhibit two distinct types of conductivity, namely, by electrons in the conduction band (generally known as *n* type for the fact that electrons carry a negative charge) and by "holes" in the valence band (generally known as *p* type because the holes effectively carry a positive charge).
2. The concentration of electrons in the conduction band and holes in the valence band can be controlled via the introduction of certain types of impurities known as dopants. In this manner, the conductivity can be dominantly *n* type or dominantly *p* type and can be varied over more than ten orders of magnitude.

More importantly, by juxtaposing *n*-type and *p*-type material, one can form a *pn* junction with unique electrical properties. Such junctions act as rectifiers for ac current and were indeed the first solid-state electrical devices, before World War II. The first transistor, invented in 1948 by John Bardeen, Walter Brattain, and William Shockley (Bardeen and Brattain, 1948; Shockley, 1949), consisted of two such junctions. Similarly, semiconductor–metal and semiconductor–insulator interfaces have unique electrical properties. Virtually all microelectronic devices that have been developed since the original transistor exploit the unique electrical properties of *pn* junctions and of semiconductor–metal and semiconductor–insulator interfaces.

The first transistor was fabricated from germanium. Soon after that, however, silicon replaced Ge as the semiconductor of choice. The primary reasons are that Si can be grown in large quantities with much better control than any other semiconductor and that Si has a native oxide (SiO_2) that is stable and has other unique properties, e.g., the quality of the Si–SiO_2 interface. Currently the global sales of Si for use in microelectronics approach $\$10 \times 10^9$ annually, fueling microelectronics, the world's largest industry, whose global revenues are approaching the 10^{12}-dollar level annually. About 5000 metric tons of single-crystal Si is currently consumed annually.

This article reviews the properties and uses of crystalline Si, with emphasis on its electronic properties and its use in microelectronics. In the first part the article covers atomic arrangements in perfect bulk crystalline Si, at its free surfaces, and at bulk defects. It then describes the vibrational, electronic, transport, thermal, and mechanical properties of Si and finishes with the methods of bulk-crystal and thin-film growth and basic processing steps used in microelectronics.

1. ATOMIC ARRANGEMENTS

1.1 The Bulk Perfect Crystal

In crystalline Si the atoms are arranged in a diamond lattice. This is a face-centered cubic (fcc) lattice, one of the 14 Bravais lattices that are allowed by symmetry considerations in three dimensions, with a basis of two atoms (Burns, 1985). If the origin of coordinates is taken at one of the Si atoms occupying one of the fcc lattice sites, the second atom in the primitive unit cell occupies a site at (1,1,1)*a*/4, where *a* is the fcc cube edge. For Si, $a = 0.543$ nm. There are eight atoms per primitive cube. Each atom has four nearest neighbors in tetrahedral directions (i.e., the four nearest neighbors are at the apexes of a regular tetrahedron), as seen in Fig. 1. The angle subtended by two neighbors is approximately 109.4°. Each atom has 12 second-nearest neighbors. If one envisions "bonds" between nearest neighbors, these bonds form six-membered rings.

The diamond lattice is a relatively open lattice. If spheres are drawn around atoms so that they just touch, the volume of these spheres is only 34% of the total volume. In contrast, in a monatomic fcc lattice, touching spheres fill 74% of the available volume.

The space group of the diamond lattice is the nonsymmorphic group designated as O_h^7 in the Schoenflies notation and $Fd3m$ in the international crystallographic notation. The point group at an atomic site is T_d. In the open spaces of the diamond lattice, there are interstitial sites that also have full tetrahedral (T_d) symmetry because they are also the centers of regular tetrahedra whose four apexes are occupied by atoms. Another high-symmetry interstitial site is the center of a six-membered ring of atoms. It has full hexagonal symmetry.

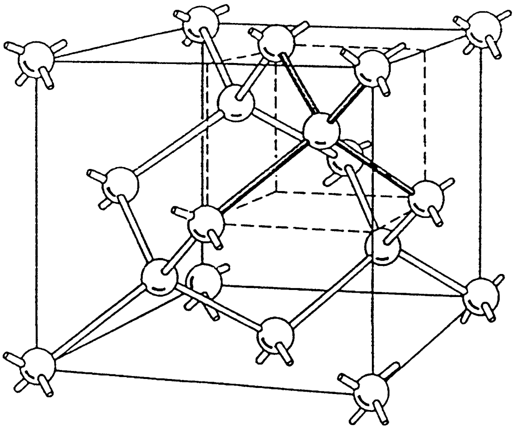

FIG. 1. Schematic showing the tetrahedral coordination of atoms in the diamond lattice.

1.2 Crystalline Surfaces, Interfaces

Since Si is a cubic crystal, the three principal surfaces are the (100), the (110), and the (111). These are shown schematically in Fig. 2 in the case where all atoms are kept at the normal sites they occupy in the perfect crystal. In these "ideal" cases, the surface layers are two-dimensional lattices with primitive translation vectors that we may denote as τ_1 and τ_2. In reality, the surfaces undergo "reconstruction," which means that the symmetry of the surface structure is reduced, leading to a lower-energy state. Equivalently, the surface primitive unit cell is larger, with the primitive translation vectors now being $m\tau_1$ and $n\tau_2$, where m and n are two different integers. One then says that the surface has undergone an $m \times n$ reconstruction.

Low-energy electron diffraction (LEED) is an experimental technique that probes the symmetry of surfaces and can establish the indices $m \times n$. Schlier and Farnsworth (1959) were the first to report a LEED determination of a 2 × 1 reconstruction of the Si(100) surface. Sometimes at low temperatures a 4 × 2 reconstruction was also seen. Later, the Si(111) surface was observed to

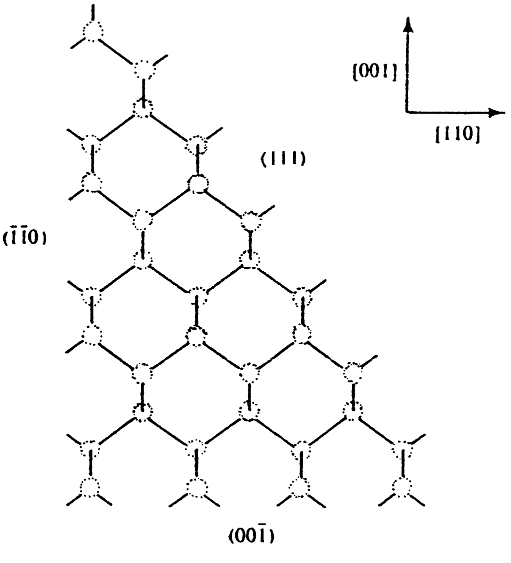

FIG. 2. Schematic showing the atomic arrangements of the principal surfaces of Si in the absence of reconstruction. The crystal is viewed along a [1$\bar{1}$1] direction. From Harrison (1980).

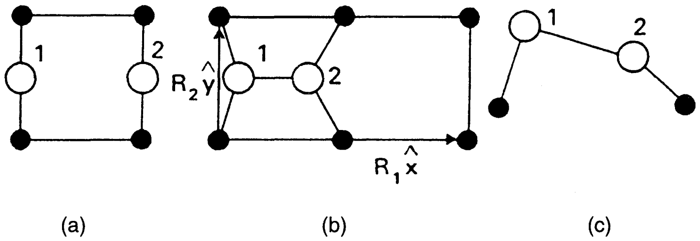

FIG. 3. (a) Top view of the "ideal" Si (100) surface. The open circles depict atoms in the topmost layer, and dark circles depict atoms in the second layer. (b) Top view of the same surface after dimer formation (2 × 1 reconstruction). (c) Side view of the same surface after dimer formation. From Chadi (1979).

undergo several reconstructions as a function of temperature. When cleaved in vacuum at room temperature, the (111) surface exhibits a 2 × 1 reconstruction. If cleaved at very low temperatures (<20 K), a 1 × 1 pattern appears. After annealing to high temperatures (>400 °C), a 7 × 7 reconstruction emerges.

The determination of the actual positions of the surface atoms in the new primitive cell is not an easy task. Intense experimental and theoretical investigations took more than ten years (1970s and early 1980s) to pin down the actual atomic arrangements on Si surfaces. Many experimental and theoretical techniques were developed and contributed to gradual progress. The current consensus is that the surface atoms on the (100) 2 × 1 surface form asymmetric dimers, as illustrated in Fig. 3; the surface atoms on the Si(111) 2 × 1 surface form pi-bonded chains, as illustrated in Fig. 4; and the surface atoms at the (111) 7 × 7 surface form the complex unit cell shown in Fig. 5. The determination of these structures involved the simultaneous elucidation of the surface electronic structure. We therefore postpone further discussion to Sec. 3.6.

The atomic arrangements at Si interfaces depend very much on the second material. For example, it has been determined that a monolayer of hydrogen atoms brings back the surface Si atoms to their "ideal" positions. At the other extreme, the atomic arrangements at the Si–SiO_2 interface are not very regular because SiO_2 is amorphous when grown thermally on Si. In fact it is likely that there exists a thin SiO_x interlayer. Furthermore, it is known that there exist point defects at the interface. The so-called P_b center is believed to be a Si atom that is bonded to only three nearest neighbors, a structure that is commonly referred to as a dangling bond. Many crystalline solids such as transition-metal silicides, CaF_2, etc. form well-ordered interfaces.

1.3 Plane and Line Defects

Stacking faults are the principal plane defects in bulk crystals. An fcc crystal can be thought of as a stacking of planes of close-packed spheres. The planes are stacked in a staggered fashion so that spheres of the second plane sit on top of interstitial sites of the first plane. In this arrangement only half of the interstitial sites have atoms placed on top of them. The third plane of spheres can then go in one of two ways: with its spheres directly over the spheres of the first plane (*ABABA* . . . stacking) or with its spheres over the unused interstitial sites of the first plane (*ABCABC* . . . stacking). The former gives rise to a hexagonal close-packed (hcp)

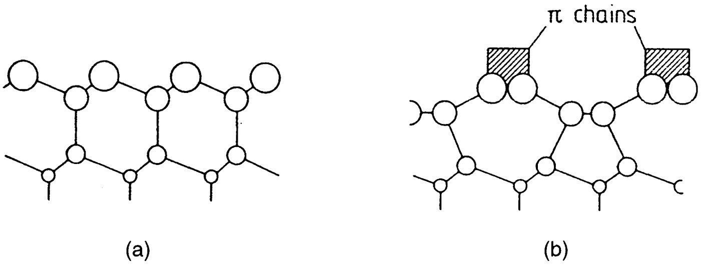

FIG. 4. (a) Side view of a Si(111) "ideal" surface. (b) Side view of the same surface after the formation of pi-bonded chains (2 × 1 reconstruction). From Pandey (1981).

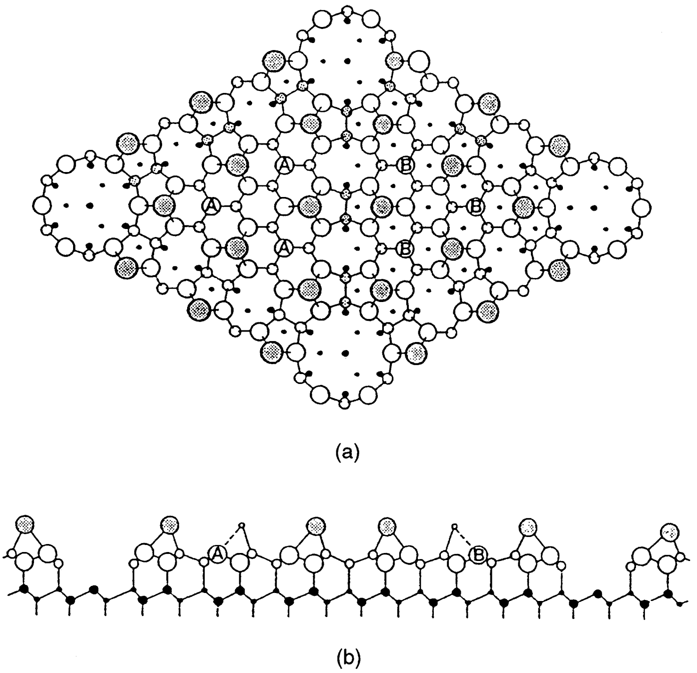

FIG. 5. The complex atomic arrangement of the Si(111) 7 × 7 surface: (a) top view, (b) side view. From Takanayagi *et al.* (1985).

structure while the latter gives rise to an fcc structure. In the fcc case the close-packed planes are {111}. An intrinsic stacking fault corresponds to a missing plane so that the stacking proceeds as *ABCABCBCA*. . . . An extrinsic stacking fault corresponds to an extra plane added in the sequence so that the stacking proceeds as *ABCABCBABC*. . . .

In a diamond lattice, because there are two atoms per unit cell, each fcc {111} plane is actually two planes of atoms, so the stacking sequence should be designated as *AaBbCcAaBb*. . . . In Si, stacking faults correspond to missing or extra *pairs* of planes of atoms. In such faults, the tetrahedral coordination of the atoms is not disturbed so that the total energy of the fault is quite small. Stacking faults where only one plane of Si atoms are missing would disturb the tetrahedral coordination and thus would have a much higher energy.

Dislocations (Hirth and Lothe, 1992) are intrinsic line defects. Edge dislocations correspond to an extra half plane of atoms being inserted in the crystal. The edge of that half plane is the dislocation, and it lies in what is called the slip plane. Screw dislocations correspond to cutting a crystal and shearing it parallel to the plane of the cut by one atom spacing. An important quantity associated with a dislocation is the Burgers vector, defined as follows. One first draws a closed circuit by connecting atoms in a perfect region of the crystal. If the same circuit is drawn around a dislocation, the Burgers vector is defined to be the vector needed to complete the circuit. For edge dislocations the Burgers vector is perpendicular to the dislocation line, whereas for screw dislocations the Burgers vector is parallel to the dislocation line. In the general case of a mixed dislocation, the Burgers vector makes some other angle with the dislocation line. For example, when this angle is 60°, the dislocation is known as a 60° dislocation. For "perfect" dislocations the Burgers vector is exactly equal to a lattice vector. "Imperfect" or partial dislocations are also possible with the Burgers vector being something other than a lattice vector.

As in all crystals, Si may contain edge, screw, or mixed dislocations. Since crystalline Si is fundamentally an fcc crystal, the preferred slip planes for dislocations are the {111} planes, because they have the closest packing of atoms and the maximum interplane separation. This configuration leads to the lowest stress, as calculated using the continuum theory of elasticity. The smallest Burgers vector in the {111} plane, which again minimizes the stress, is in the ⟨110⟩ direction.

In Si it has been found that the prominent dislocations in the early stages of glide are either pure screw or 60° dislocations. There are actually two distinct types of dislocations in Si because of the double-layer atomic arrangement of the diamond lattice. The two types are called dislocations of the glide set and dislocations of the shuffle set. The distinction can be found in any comprehensive book on dislocations.

In Si the perfect dislocations described above dissociate into partial dislocations. Typically, a perfect 60° dislocation dissociates into 30° and 90° partial dislocations. The two partials bound an intrinsic stacking fault.

In dislocations in Si the tetrahedral bonding is disturbed, with some atoms being threefold coordinated. As in surfaces, reconstruction follows to lower the energy of the system. The detailed arrangement of Si atoms in the core of dislocations is still a subject of active research.

1.4 Point Defects

The primitive native point defects in any crystal are the vacancy (missing atom at nominal atomic site) and the self-interstitial (atom at a site that is not a nominal atomic site). Together, they are often called a Frenkel pair. Point defects such as the vacancy occur in different charge states because they have localized energy states within the forbidden band gap. The atomic rearrangements around the point defect may then depend on the charge state. In the case of the vacancy, V^{++} undergoes no lattice reconstruction, i.e., full tetrahedral symmetry is retained (the neighboring atoms only relax in a symmetry-preserving mode that is known as the breathing mode), whereas all other charge states undergo symmetry-reducing reconstructions. These reduced symmetries have been deduced primarily from electron paramagnetic resonance data (Watkins, 1994). The origins of these phenomena are discussed in Sec. 3.6 in the context of the electronic structure of point defects.

The self-interstitial has been more difficult to pin down. Experimentally it has not been observed, apparently because it is highly mobile and reactive. Theoretical calculations that seek to establish its equilibrium configuration by comparing the formation energy of different configurations have found that the lowest-energy configuration depends on the charge state of the defect (see Sec. 3.6). The latest work on the subject finds that the neutral interstitial has a relatively complex structure: it is a form of a split interstitial, namely, a dumbbell of two atoms occupying what is normally an atomic site.

Impurities may occupy substitutional sites, in which case the surrounding lattice usually relaxes a very small amount without lowering the point symmetry. Examples are the dopant impurities (group-III and group-V elements). Other impurities—e.g., Li—occupy the high-symmetry tetrahedral interstitial site. Transition-metal impurities may be either substitutional or interstitial. Oxygen is a rather unique impurity. Its lowest-energy configuration is a low-symmetry interstitial site: oxygen occupies a bridge position between neighboring Si atoms with the Si–O–Si angle being about 145°, mimicking the Si–O–Si bonds in SiO_2. By high-energy electron irradiation, oxygen can be induced into another configuration that is nominally substitutional, but with the oxygen atom shifted off-center, bridging two of its Si neighbors. This center is often referred to as the oxygen-vacancy complex (Watkins and Corbett, 1961). Other impurities—e.g., phosphorus—form true vacancy–impurity pairs; namely, the impurity occupies a substitutional site next to a vacant site. Finally, other complexes are possible—e.g., a boron-hydrogen pair, where the boron is substitutional and the hydrogen occupies a nearby interstitial site.

1.5 Polycrystalline, Amorphous, and Porous Si

For completeness, we mention here briefly forms of Si that are not crystalline.

Polycrystalline Si (see Kamins, 1988; SILICON, POLYCRYSTALLINE), like all polycrystals, consists of crystallites (grains) that meet at grain boundaries (*q.v.*). We do not attempt here a review of grain-boundary structures, which are general for any material. Of interest here would be the particular atomic arrangements in Si grain boundaries. Just like free surfaces, the planes of atoms at a grain boundary undergo relaxations and reconstructions from their "ideal" positions. It is believed that all atoms at Si grain boundaries are fourfold coordinated. Establishing the detailed atomic positions is a demanding task. High-resolution microscopy provides detailed images (Bourret and Rouvier, 1989). Typically one must assume some structure and simulate the image to compare with experiment. Total-energy minimizations also help guide the search for possible structures. Usually, however, more than one structure can be found with similar total energy and similar simulated electron-microscope image.

Real grain boundaries in polycrystalline Si do have point defects that are electrically active and can be passivated by hydrogen. In addition, grain boundaries are low-energy sites for dopant impurities. Since in grain boundaries such impurities are electrically inactive, they are most likely threefold coordinated, which is the low-energy coordination they prefer. It was recently determined that arsenic impurities segregate in Si grain boundaries threefold-coordinated as dimers or ordered chains (Maiti *et al.*, 1996).

Amorphous Si (see SILICON, AMORPHOUS) is a network of fourfold-coordinated Si atoms. For each Si atom the four neighbors are roughly in tetrahedral directions. In contrast to crystalline Si, there is no long-range periodicity. In addition, whereas in crystalline Si the atoms are arranged in buckled six-member rings (where a "member" is a Si–Si bond), in amorphous Si there is a distribution of five-, six-, and seven-member rings.

By viewing "perfect" amorphous Si as a network of fourfold-coordinated Si atoms, one immediately deduces that the native point defects would be threefold-coordinated atoms (generally known as dangling bonds) and fivefold-coordinated atoms. The latter have been labeled floating bonds (Pantelides, 1986). Traditionally, dangling bonds were viewed to be the dominant defect that gives rise to electron-paramagnetic-resonance (EPR) and luminescence signals. The evidence for such an identification, however, is only circumstantial. The concept of floating bonds is of more recent import. They have been found in theoretical simulations of amorphous Si structures. Definitive experimental proof as to the nature of the observed defects and the relative concentrations of dangling and floating bonds is lacking.

Porous Si is a form of disordered Si with a high density of pores. Its claim to fame is that it is the only form of Si known to produce strong luminescence lines in the red region of the optical spectrum. The detailed atomic arrangements that produce luminescence have not been established, though a number of suggestions have been put forth.

2. VIBRATIONAL PROPERTIES—PHONONS

In the previous section, we described the atomic positions in a perfect Si crystal, at surfaces, and at line and point defects. The atoms actually vibrate about these nominal sites. For small vibrations, the allowed frequencies are obtained by a harmonic approximation: the total energy of the system, which is a function of the actual atomic positions, is expanded to second order in the atomic displacements. One then can introduce a set of generalized coordinates (the classical normal vibrational modes) in terms of which the total energy becomes a simple sum of terms, each of which is the Hamiltonian of a simple harmonic oscillator. In the case of a perfect crystal, the normal modes are propagating waves and can be labeled by a wave vector $\mathbf{q}$ and a "branch index" ν. The number of branches is equal to the number of degrees of freedom in the primitive cell, i.e., three for a monatomic solid and six for a diatomic solid like Si. The corresponding frequencies are also labeled by the wave vector $\mathbf{q}$ and the branch index ν. Excitations of these decoupled harmonic-oscillator states are known as phonons.

The phonon spectra of crystalline Si are shown for $\mathbf{q}$ along high-symmetry directions in Fig. 6. The three lower branches are the acoustical branches, with the slope near $\mathbf{q} = 0$ corresponding to the speed of sound in the crystal. In the acoustical branches, the two Si atoms at each lattice site move in tandem. The three upper branches are the optical branches, in which the two Si atoms at each lattice site move in opposite directions. Each set of three branches contains a longitudinal branch—with the atoms oscillating along the direction of wave propagation—and two

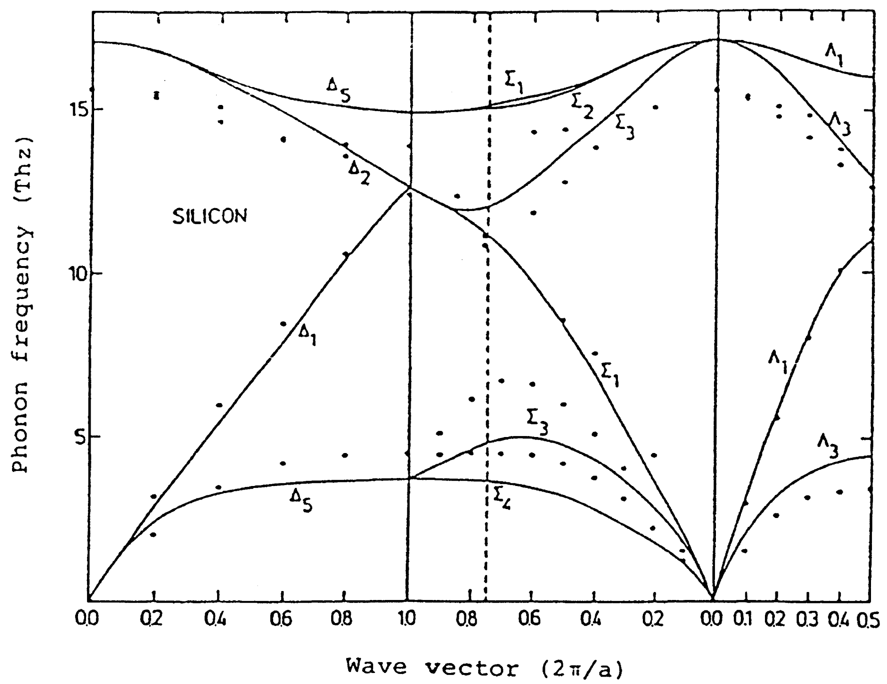

FIG. 6. The phonon spectra of silicon.

transverse branches—with the atoms oscillating perpendicular to the direction of wave propagation.

3. ELECTRONIC STRUCTURE AND PROPERTIES

3.1 The Perfect-Crystal Energy Bands

Electron states in Si are adequately described by the self-consistent-field approximation according to which every electron is acted upon by the same effective periodic potential, arising from all the nuclei and the rest of the electrons. The resulting electron states, characterized by a propagation vector **k** and known as Bloch states, form bands of allowed energies separated by band gaps. The electron energy bands of Si have been calculated by a multitude of techniques, and by now they are thoroughly understood. A plot of the energy bands as a function of the wave vector **k** along high-symmetry directions in the Brillouin zone (i.e., the primitive unit cell in reciprocal space) is shown in Fig. 7. The lowest four bands are completely occupied by the four electrons per unit cell in the Si lattice (two electrons—spin up and spin down—per band at each **k** point) and are known as valence bands. The bands above the forbidden gap are empty, making Si a semiconductor with an energy gap of 1.16 eV at 0 K (1.12 eV at 300 K).

The lowest conduction band has six equivalent minima at $(\pm k_c, 0, 0)$, $(0, \pm k_c, 0)$, and $(0, 0, \pm k_c)$, where $k_c \approx 0.82(2\pi/a)$ and a is the primitive cube edge. In the vicinity of each of these minima, the band is parabolic but with different transverse and longitudinal effective masses:

$$E(k) \cong \frac{\hbar^2}{2m_\perp} k_x^2 + \frac{\hbar^2}{2m_\parallel}(k_y^2 + k_z^2) \qquad (1)$$

for the minima along the k_x axis. The values of the masses are $m_\perp = 0.19 m_e$ and $m_\parallel = 0.97 m_e$, where m_e is the mass of a free electron.

The valence bands have a maximum at **k** = 0. At that point, the bands are degenerate. If spin–orbit interactions are ignored, the degeneracy is threefold (times two for spin) with Bloch functions that transform like the

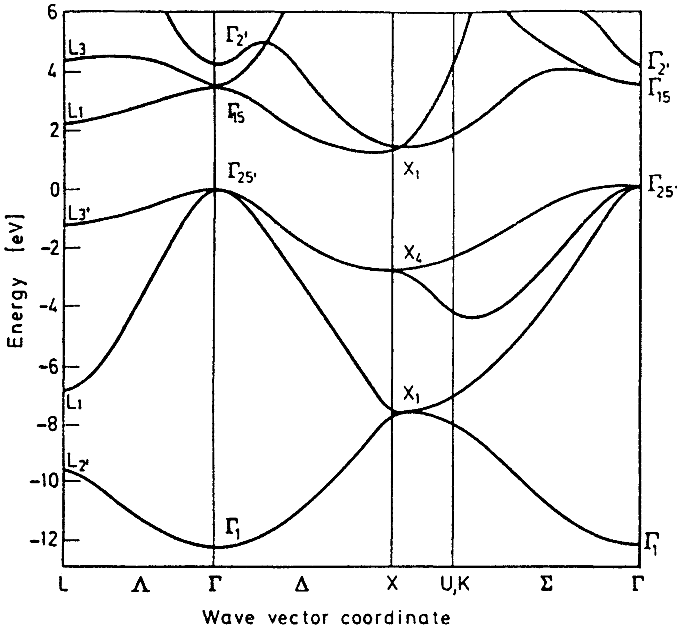

FIG. 7. Energy bands of Si along high-symmetry directions in **k** space. The designations such as L_1, X_4, etc. refer to symmetry properties of the corresponding wave functions. From Chelikowski and Cohen (1976).

functions xy, yz, and zx ($\Gamma_{25'}$ representation of the T_d group). The three bands are the eigenvalues of the following 3 × 3 matrix developed by $\mathbf{k}\cdot\mathbf{p}$ theory (Kohn, 1957):

$$\begin{pmatrix} Ak_x^2 + B(k_y^2 + k_z^2) & Ck_xk_y & Ck_xk_z \\ Ck_yk_x & Ak_y^2 + B(k_z^2 + k_x^2) & Ck_yk_z \\ Ck_zk_x & Ck_zk_y & Ak_z^2 + (k_x^2 + k_y^2) \end{pmatrix}. \tag{2}$$

When spin–orbit interaction is included, the bands at $\mathbf{k} = 0$ split into a fourfold-degenerate set that transforms like Γ_8 and a twofold-degenerate set that transforms like Γ_7. In the region near $\mathbf{k} = 0$, the top four bands split into the top "heavy-hole" band and the lower "light-hole" band. One can approximate these bands with parabolic bands and assign effective heavy-hole and light-hole effective masses. The values of these effective masses are $m_l = 0.16m_e$ and $m_h = 0.5m_e$. The lowest band is known as the split-off band. At $\mathbf{k} = 0$ the spin–orbit splitting is 44 meV. On the basis of these experiments and $\mathbf{k}\cdot\mathbf{p}$ theory, the structure of the energy bands near the energy gap was established in the 1950s, long before it became possible to calculate the full energy bands from first principles.

Energy-band theory evolved since the 1950s when the first computers became available. The energy bands of Si have probably been calculated with just about every conceivable method. Currently, the most rigorous way to formulate the "one-electron" problem for a crystal is the density-functional theory of Hohenberg and Kohn (1964). In this theory, the total energy E is shown to be a universal functional of the electron density $n(r)$, which in turn is written in terms of one-electron wave functions $\psi_i(r)$:

$$n(r) = \sum_i |\psi_i(r)|^2, \tag{3}$$

where the sum is over occupied states. In the corresponding effective one-electron potential, the exchange-correlation term is not known exactly. The most common approximation is the local-density approximation, where one assumes that the known form of exchange and correlation for the homogeneous electron gas can be used at each r for the local value of $n(r)$. This approximation is often coupled with a pseudopotential approximation, where the core electrons and the strong potential that binds them are replaced by an effective potential whose lowest eigenstates are the valence states. As a result, the oscillatory part of the valence wave functions in the core regions is replaced by a smooth part. It is then possible to expand the crystal wave functions in terms of a practically small set of plane waves.

In density-functional theory, the one-electron wave functions and corresponding one-electron energies are not rigorously wave functions and energy levels. They are merely intermediate quantities that enter the theory, in terms of which the total energy is calculated. Nevertheless the calculated one-electron energies give a good representation of the one-electron energy bands that are measured experimentally. The biggest shortcoming is the fact that the theoretical band gap is only about half of what is measured. In part, the problem is a consequence of the local-density approximation for exchange and correlation. Theories that go beyond these approximations have been developed and yield energy bands that are in full accord with experimental data (Hybertsen and Louie, 1985). Silicon has in fact been the testing ground for such theories.

The structure of the energy bands of Si can be understood in terms of simple arguments based on atomic theory. Si atoms have 1*s*, 2*s*, and 2*p* core electrons. They also have two 3*s* and two 3*p* electrons. One can combine the four 3*s* and 3*p* wave functions into four sp^3 "hybrid orbitals" that point in the four tetrahedral directions. For example, the orbital h_{111} defined by

$$h_{111} = \tfrac{1}{2}(\phi_s + \phi_x + \phi_y + \phi_z) \tag{4}$$

points in the (111) direction. Here ϕ_s is the 3*s* orbital of Si and ϕ_x, ϕ_y, and ϕ_z are the three 3*p* orbitals. The energies of the four hybrid orbitals can then be easily calculated to lie at $\epsilon_h = (\epsilon_s + 3\epsilon_p)/4$, where ϵ_s and ϵ_p are the 3*s* and 3*p* atomic energies, respectively. One can then form bonding and antibonding orbitals by combining hybrids pointing toward each other. If we denote two such hybrids on neighboring atoms by h_1 and h_2, the bonding and antibonding orbitals are defined by

$$b = \tfrac{1}{2}(h_1 + h_2), \tag{5}$$

$$a = \tfrac{1}{2}(h_1 - h_2). \tag{6}$$

If the interaction matrix element $\langle h_1|H|h_2\rangle$, where H is the crystal Hamiltonian, is denoted by $-V_2$, the corresponding energies are $\epsilon_b = \epsilon_h - V_2$ and $\epsilon_a = \epsilon_h + V_2$, and the bonding–antibonding splitting is $2V_2$. The notation used here is that introduced by Walter Harrison of Stanford University, who in the 1970s constructed an elementary theory of semiconductor properties based on these concepts (see Harrison, 1980).

One can construct Bloch functions out of the bonding orbitals and, by using the Hamiltonian matrix elements between neighboring bonding orbitals as parameters, reproduce the known valence bands of Si quite well. The conduction bands arise from the antibonding orbitals.

The densities of one-electron states of Si have been calculated and measured by photoemission techniques. They are shown in Fig. 8. Points of high symmetry are reflected in features of the density of states.

3.2 Optical and Dielectric Response

Optical absorption experiments cause electrons to be excited from valence states to conduction states. The optical reflectivity, which is directly related to the optical absorption, of Si is shown in Fig. 9. The main peak around 4 eV arises primarily from transitions between the highest valence bands and the lowest conduction bands that, as can be seen in Fig. 7, are approximately "parallel." In addition to the band-to-band transitions, electron–hole attraction between the electron and the hole gives rise to excitons (*q.v.*) that lie at energies just below the band gap. The excitons are not resolved in Fig. 9.

X-ray absorption experiments excite electrons from the core states to the conduction bands. Because the core states are flat in **k** space (because of negligible interactions between neighboring core wave functions), the x-ray absorption spectrum should reflect the conduction-band density of states. The observed spectra actually rise much more steeply at the threshold than the corresponding density of states (Fig. 10). The effect has been attributed to electron–hole interactions and the formation of excitons just below the

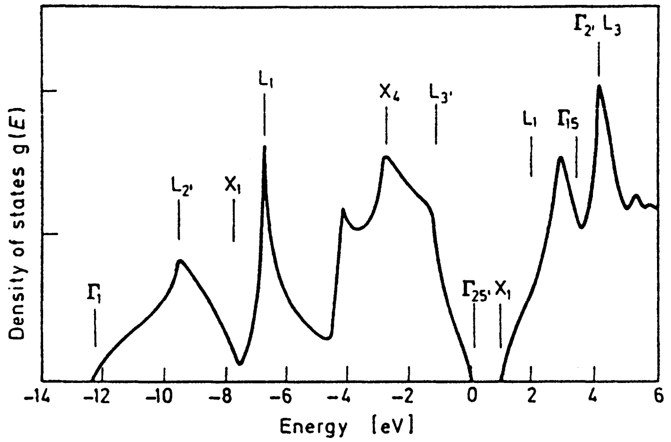

FIG. 8. The density of one-electron states in Si. From Chelikowski *et al.* (1973).

band edge. These excitons have very small binding energies (~30 meV; see Sec. 3.3 for this estimate) and are not resolved.

When an external electrostatic potential $V_{\text{ext}}(r)$ is applied to a crystal, the electrons are polarized and give rise to a screening potential. In linear response theory, the total (screened) potential $V(r)$ is given by

$$V(r) = \int \frac{V_{\text{ext}}(q)}{\epsilon(q)} e^{-i\mathbf{q}\cdot\mathbf{r}} d^3q, \tag{7}$$

where $V_{\text{ext}}(q)$ is the Fourier transform of $V_{\text{ext}}(r)$. The dielectric function $\epsilon(q)$ can be calculated in terms of virtual excitations of the core and valence electrons. Results of numerical calculations have been fitted to the following analytic formula (Nara, 1965):

$$\frac{1}{\epsilon(q)} = \frac{Aq^2}{q^2 + \alpha^2} + \frac{Bq^2}{q^2 + \beta^2} + \frac{1}{\epsilon_0}\frac{\gamma^2}{q^2 + \gamma^2}, \tag{8}$$

where $A = 1.175$, $B = 1 - A = -0.175$, $\alpha = 0.7572$, $\beta = 0.3123$, and $\gamma = 2.044$. The constant ϵ_0 is the static dielectric constant $\epsilon(0)$, which for Si is 11.4.

In the simplest model of a dielectric constant, one assumes that all the virtual excitations can be lumped into a single effective excitation at an energy E_0. A well-known sum rule for the oscillator strengths then yields the following expression for ϵ_0:

$$\epsilon_0 = 1 + (\hbar\omega_p)^2/E_0^2, \tag{9}$$

where $\omega_p^2 = 4\pi ne^2/m$ is the so-called plasma frequency for the electron gas whose density is n. Here e is the electron charge and m is the electron mass. Since ϵ_0 and ω_p are known from experiment, this formula allows one to extract a value of the effective dielectric gap E_0, sometimes referred to as the Penn gap (Penn, 1962). This model of the dielectric constant formed the basis of a theory of ionicity for semiconductors by James C. Phillips of Bell Laboratories (see Phillips, 1973). The theory's main thesis is that the dielectric gap E_0 can be written in terms of a homopolar part E_h and an ionic part C in the form

$$E_0^2 = E_h^2 + C^2. \tag{10}$$

In an alternative approach, Harrison (1980) has suggested that the dielectric constant is

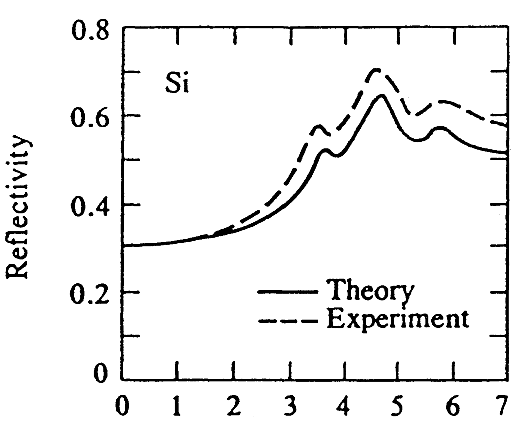

FIG. 9. The optical reflectivity spectrum of Si.

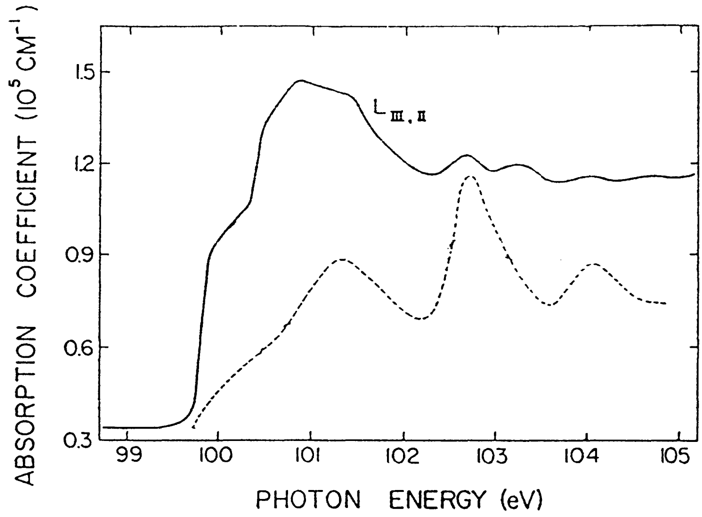

FIG. 10. The x-ray absorption spectrum of Si arising from excitation of 2*p* core electrons. From Brown and Rustgi (1972).

primarily determined by virtual excitations that on average correspond to the bonding–antibonding energy. The solid is then in effect represented by a system of polarizable bonds with a single excitation energy of $2V_2$ in the notation introduced earlier. A formula for the dielectric constant is then obtained that has a different form from Eq. (9) simply because Harrison did not impose the requirement that the oscillator strength of the single average transition be equal to 1. Harrison's theory yields a different ionicity (or polarity) scale for the compound semiconductors.

3.3 Dopant Impurities

When a Si atom is replaced by a group-V element (P,As,Sb), an extra electron is introduced in the crystal and the Si^{+4} core is replaced by a $+5$ core. There is a net increase in the core charge of $+e$. The resulting Coulomb potential is screened, in the simplest approximation, by the dielectric constant ϵ_0:

$$V(r) = -e^2/\epsilon_0 r. \tag{11}$$

The extra electron would normally go to the bottom of the conduction band. If the conduction band had a single minimum with effective mass m^*, the problem would be completely analogous to the free hydrogen atom. A series of hydrogenic bound states appear below the conduction-band edge, but the effective Rydberg is reduced by a factor m^*/ϵ_0^2. If we take round numbers such as $m^* = 0.3$ and $\epsilon_0 = 12$, we find that the effective Rydberg is of order 30 meV, i.e., a small fraction of the Si band gap. At $T = 0$ K, the donor electron would be in the ground state of the hydrogenic series, but at room temperature most donor electrons would be excited into the conduction band where they can contribute to conductivity.

A very analogous situation exists when a Si atom is replaced by a group-III element (B,Al,Ga,In). An electron is now missing, which is equivalent to a hole at the top of the valence band. The net change in the core charge is now $-e$. Hydrogenic states appear above the valence band, where the hole can be bound. Again, at room temperature, most holes are excited into the valence bands.

The simple analysis given above was formulated rigorously by several authors in the 1950s (Kohn and Luttinger, 1955). The resulting theory is known as effective-mass theory, and its range of validity is established. It generally applies when the binding energy is a small fraction of the band gap, and the corresponding hydrogenic wave functions are spread out to encompass several shells of neighbors around the impurity.

In Si, however, the conduction-band minima are anisotropic, as given by Eq. (1). As a result, the *p* and higher-order states are split. (E.g., the 2*p*, $m = 0$ state has a different energy from the 2*p* states with $m = \pm 1$; here m is the azimuthal quantum number. These states are denoted by $2p^0$ and $2p^{\pm}$.) These states were observed by infrared absorption in the 1950s and 1960s. The experimental Rydberg series were in excellent agreement with theoretical effective-mass values.

The Si conduction bands present yet another complication. There are six equivalent minima. Symmetry requires that the bound-state wave functions be constructed as linear combinations of states at all six minima. The coefficients are determined by group theory. The bound-state energies then are determined by intravalley and intervalley terms. For $2p$ and higher-order states, the intervalley terms are negligible. For the $1s$ ground state, however, they are not. As a result, what would have been a sixfold-degenerate $1s$ state splits into a single of A_1 symmetry, a doublet of E symmetry, and a triplet of T_2 symmetry (Kohn and Luttinger, 1955). Here the chemical notation for the irreducible representations of the T_d group are used.

If the impurity potential were simply a screened Coulomb potential, all group-V impurities would have the same hydrogenic energy levels. Experimental data show that the $1s$ ground state actually differs from impurity to impurity (Table 1). The reason is because the $1s$ hydrogenic wave function (scaled by m^* and ϵ_0) is sufficiently localized to sample the so-called central-cell region where the actual potential looks more like the difference between the impurity-atom and host-atom pseudopotentials. The deviations of the ground states from the simple hydrogenic result are known as central-cell corrections or chemical shifts.

The valence bands of Si present altogether different complications. The corresponding kinetic energy operator in the hydrogenic equation is now a matrix corresponding to Eq. (2). The resulting hydrogenic spectrum reflects this multiplicity of bands and is quite complex. The spectra have been measured, and detailed calculations have been carried out. Again, however, the ground states of the various impurities exhibit chemical shifts.

Table 1. Donor energy levels in meV, measured from the conduction-band edge (from Aggarwal and Ramdas, 1965).

Impurity	$1s(A_1)$	$1s(T_2)$	$1s(E)$
Phosphorus	45.5	33.9	32.6
Arsenic	53.7	32.5	31.2
Antimony	42.7	32.9	30.6

3.4 Heavy Doping—The Degenerate Limit

When the concentration of shallow impurities exceeds a certain limit, the impurity ground-state wave functions start to overlap, and the discrete energy levels broaden into a band. This is a classic example of the Mott metal-insulator transition. Let us consider a simple hydrogenic $1s$-like ground state that can take a single electron (a second electron can be bound to form a negatively charged center, but the binding of the second electron, as in the free H atom, is roughly 1/20 of the effective Rydberg). As such levels broaden into a band, there are enough electrons to fill up to half the band. Thus the system is in principle metallic at all impurity concentrations. The system, however, remains insulating up to a critical bandwidth, when delocalization wins. Because of the disorder in the impurity distribution, there are always localized tail states (see ELECTRON STATES, LOCALIZED) at the edge of the band.

Beyond the Mott transition (at about 5×10^{18} cm^{-3} impurities), the impurity band merges with the regular conduction band so that in effect the conduction-band edge is now lower, resulting in a smaller band gap. There are other factors contributing to the band-gap shrinkage in heavily doped Si (Pantelides *et al.*, 1985). The electrons are no longer in discrete states but rather fill a regular band up to the Fermi level. Electron–electron interactions, which are not present in the case of the perfect crystal, have a net effect of lowering the band. The effect is analogous to the usual treatment of electron–electron interactions in a simple metal when one goes beyond the independent electron approximation. The band-gap shrinkage in heavily doped Si is an important effect in Si devices.

3.5 Lattice Point Defects and Deep Impurities

The structure of the electronic states that are localized at a vacancy can be inferred correctly from simple considerations in terms of the sp^3 hybrids that form bonds in the bulk. When an atom is removed and the neighboring atoms are kept frozen at their nominal bulk positions—known as the "ideal" vacancy—we are left with four dangling bonds, i.e., dangling sp^3 hybrids pointing tetrahedrally toward the vacant site. The true localized states need to have the full symmetry of the tetrahedral group and can be represented by the following combinations of the hybrids:

$$\begin{aligned} |A_1\rangle &= \tfrac{1}{2}(h_1 + h_2 + h_3 + h_4), \\ |T_2\rangle_1 &= \tfrac{1}{2}(h_1 - h_2 + h_2 - h_3), \\ |T_2\rangle_2 &= \tfrac{1}{2}(h_1 + h_2 - h_3 - h_4), \\ |T_2\rangle_3 &= \tfrac{1}{2}(h_1 - h_2 - h_3 + h_4). \end{aligned} \tag{12}$$

Here, the h subscripts refer to the four Si atoms neighboring the vacant site and the designations A_1 and T_2 refer to symmetry under the operations of the T_d group, which is the point group of the ideal vacancy. States of A_1 symmetry transform like s orbitals, whereas the T_2 representation is triply degenerate and transforms like p orbitals.

The true localized states are expected to have the same transformation properties as those given above, but the detailed wave functions would of course be more complex. The construction of the states given by Eqs. (12) allows for a simple electron counting. There are only four electrons (one from each of the neighboring Si atoms) to occupy the localized states so that only half of the eight states (4 times 2 for spin) are occupied for the neutral vacancy.

George Watkins at the General Electric Laboratory studied the Si vacancy extensively in the 1960s (see Watkins, 1994). He was able to create and isolate vacancies by high-energy electron bombardment at very low temperatures (4 K). With electron paramagnetic resonance experiments, he was able to determine that the A_1 state is lower in energy and contains two electrons, whereas the T_2 state lies in the gap. The neutral vacancy would thus have two electrons in a sixfold-degenerate state. Such an arrangement triggers the Jahn–Teller effect: any symmetry-lowering distortion splits the T_2 state with a gain in energy that is linear in the distortion amplitude. This linear gain in energy is ultimately balanced by the normal quadratic term in the total energy so that the stable configuration has reduced symmetry. A combination of experiments (EPR, uniaxial stress, optical excitation) led Watkins to conclude that V^+, V^0, and V^- undergo a tetragonal distortion.

The Si vacancy was also the prototype point defect that was studied by the first parameter-free first-principles calculations of the properties of point defects in solids. These techniques were developed in the late 1970s. Baraff and Schluter (1978) at AT&T Bell Laboratories and Bernholc *et al.* (1978) at IBM Research reported the first detailed calculations. The T_2 level is at about the middle of the gap, and the A_1 state is a resonance below the valence-band edge. Electron-density plots for the ideal vacancy are shown in Fig. 11. Note that the wave function of the T_2 state in the gap looks very much like a dangling bond (dangling hybrid) pointing toward the vacant site, with a tail that extends over several neighbors. Note also that, in contrast, the total change in electron density caused by the removal of an atom is very highly localized within roughly an atomic volume.

Further theoretical study of the vacancy by Baraff *et al.* (1979) led to the conclusion that the vacancy is very likely to exhibit so-called "negative-U" behavior. The concept had been introduced earlier by Anderson of Bell Laboratories (1975). Under normal circumstances, if a localized state contains an electron, there is a positive energy cost U to add a second electron to the same level. This energy is often referred to as the Hubbard U. Anderson noted, however, that lattice relaxation around a point defect could in principle lead to enough energy gain so as to overcome the Hubbard U and result in a net negative U. In such a case, the point defect would have either no electrons or two electrons, with the one-electron state being unstable and unobservable. Anderson made the suggestion in the context of chalcogenide glasses to account for the lack of EPR signal from coordination defects. Following the theoretical suggestion for the Si vacancy, Watkins and Troxell (see Watkins, 1994) designed a set of clever experiments that confirmed that the vacancy is indeed a negative-U center.

The self-interstitial in Si has never been isolated. In Watkins's experiments, the electron bombardment creates an equal number of vacancies and interstitials. The signature

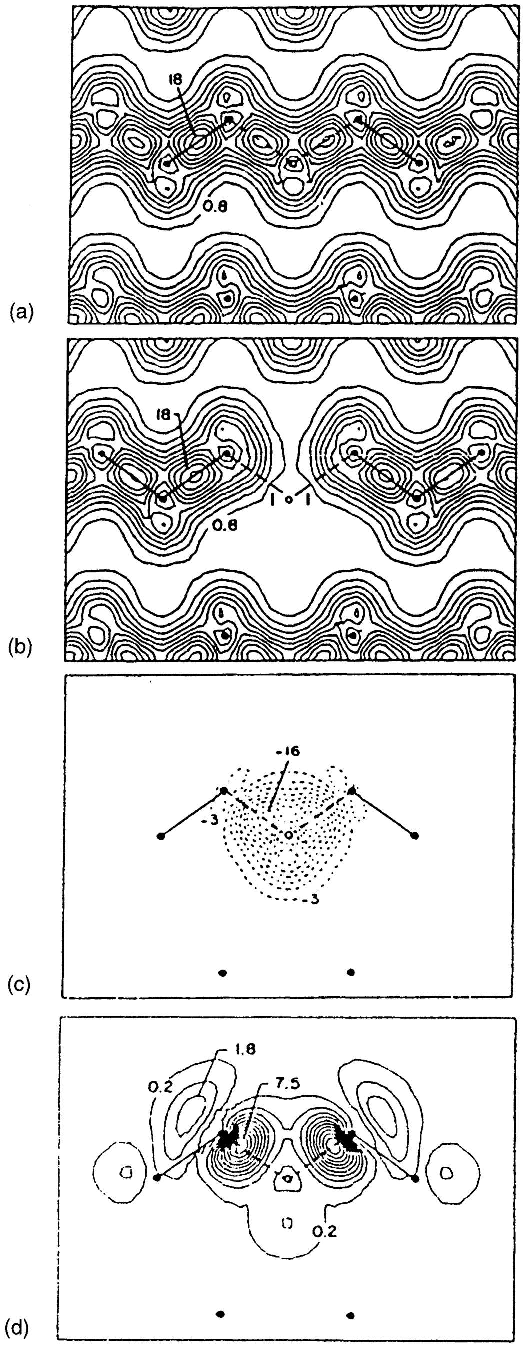

FIG. 11. Electron-density plots in the (110) plane for the "ideal" vacancy in Si. (a) The electron density in a perfect Si crystal; (b) the electron density in a crystal containing a vacancy; (c) the difference between the previous two panels; (d) the electron-density plot for the localized state in the energy gap. From Bernholc *et al.* (1980).

of the interstitial has been indirect. In Al-doped Si, after electron bombardment, the number of vacancies was matched by an equal number of Al interstitials. Since initially all Al atoms are substitutional, the obvious inference is that Si interstitials are extremely mobile even at 4 K, find the substitutional Al atoms, and kick them out. A mechanism for such athermal migration was proposed by Bourgoin and Corbett (1972): Consider an interstitial atom at a site *A* in a positive charge state with a barrier for migration that can be as large as the band gap. Denote the saddle point for this migration by *B*. Now assume that the equilibrium site for the interstitial in the neutral state happens to be the site *B*. The situation is very analogous to what happens at a negative-*U* center where a change of charge state is accompanied by drastic atomic rearrangements. Now consider the interstitial in the presence of free electron–hole pairs. When the interstitial is positively charged at site *A*, it can capture an electron and become neutral. It naturally then rolls over to the nearest *B* site, where it is stable until it captures a free hole to become positive again. Then it rolls over to the nearest *A* site (with some probability that it returns to its original *A* site), and the process continues. In effect, the barrier is overcome by using the energy given up by electron–hole recombination so that a barrier of up to the value of the band gap can be overcome. Detailed first-principles calculations have confirmed that migration along a path involving the tetrahedral and hexagonal sites can proceed athermally. The relevant total energy curves are shown in Fig. 12. The lowest-energy form of the neutral interstitial has since been determined to be a configuration of lower symmetry with two Si atoms forming a dumbbell at a nominal Si site (Blöchl *et al.*, 1993). The athermal path is likely to involve this site and a tetrahedral site.

The simplest deep impurities in Si are the substitutional group-VI elements sulfur, selenium, and tellurium (double donors) and the substitutional group-II elements zinc and cadmium (double acceptors). It is instructive to start with a vacancy and consider the modifications incurred when an impurity atom is added at the vacant site. If a group-I element is added (say copper), one may expect a small interaction between the copper

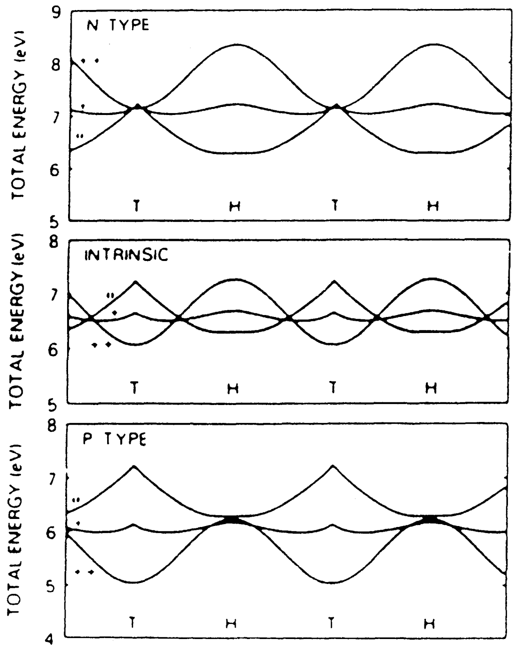

FIG. 12. The total-energy curves for different charge states of the Si interstitial along a path that includes the tetrahedral (T) and the hexagonal (H) sites (from Pantelides, 1994).

3*s* electron and the vacancy A_1 resonance and an even smaller interaction between the copper 3*p* state (which is empty) and the vacancy T_2 state. The net effect should be a T_2 level in the gap containing a total of three electrons in the neutral state. Extending this picture one more step, a group-II impurity should have a T_2 level in the gap with four electrons in the neutral state. Viewed another way, this level in the gap contains two bound holes, which are the two holes one focuses on when treating a Zn impurity as an effective-mass double acceptor. In fact if we extend the theoretical model to a group-III impurity being inserted in a vacant site, we end up with the T_2 state containing five electrons, i.e., only one bound hole, which is the well-known effective-mass hole.

In the model just described, as we proceed from a group-I to a group-II to a group-III impurity, the vacancy T_2 level is successively lowered in energy because the impurity pseudopotential reduces the net repulsion arising from the removal of a Si pseudopotential. Detailed calculations have shown that indeed the level in the gap of group-I and group-II acceptors is virtually identical to that of the vacancy (see Pantelides, 1994).

Group-VI donors have states that are more complex. Four out of the six electrons play the role of Si electrons so that the valence states of the crystal are not seriously perturbed. The two extra electrons again occupy localized states that again have A_1 and T_2 symmetry. Now the A_1 level is in the gap and can take one or two electrons, corresponding to the one or two electrons of effective-mass donors. The T_2 level is a resonance in the conduction bands. The double acceptors have been studied experimentally in great detail by Hermann Grimmeiss's group at the University of Lund in Sweden (see Grimmeiss and Janzen, 1994). It was demonstrated that these deep double donors have full Rydberg excited states just like the single donors. These Rydberg states arise from the Coulombic tail of the impurity potential and should in principle be present in the ionization spectra of any neutral or positively charged double donor. The first and second ionization energies of the group-VI impurities are listed in Table 2.

Substitutional transition-metal impurities can also be understood by starting with a vacancy that has a T_2 level in the energy gap and adding an impurity atom. The relevant atomic energy levels are now the *d* states. The local tetrahedral symmetry splits the five *d* states into a threefold T_2 and a twofold *E* level. Clearly the *E* level does not interact with any of the vacancy localized levels. In contrast, the T_2 level interacts with the vacancy T_2 level. Detailed first-principles calculations have confirmed that the trend of energy levels is consistent with this simple model and provided additional insights. An extensive review of the subject has been given by Zunger (1986). In Fig. 13 we show the results for the series of 3*d* series of ele-

Table 2. First and second ionization energies of group-VI donors in Si (from Grimmeiss and Janzen, 1994).

Impurity	First ionization energy (eV)	Second ionization energy (eV)
Sulfur	0.32	0.61
Selenium	0.31	0.59
Tellurium	0.20	0.41

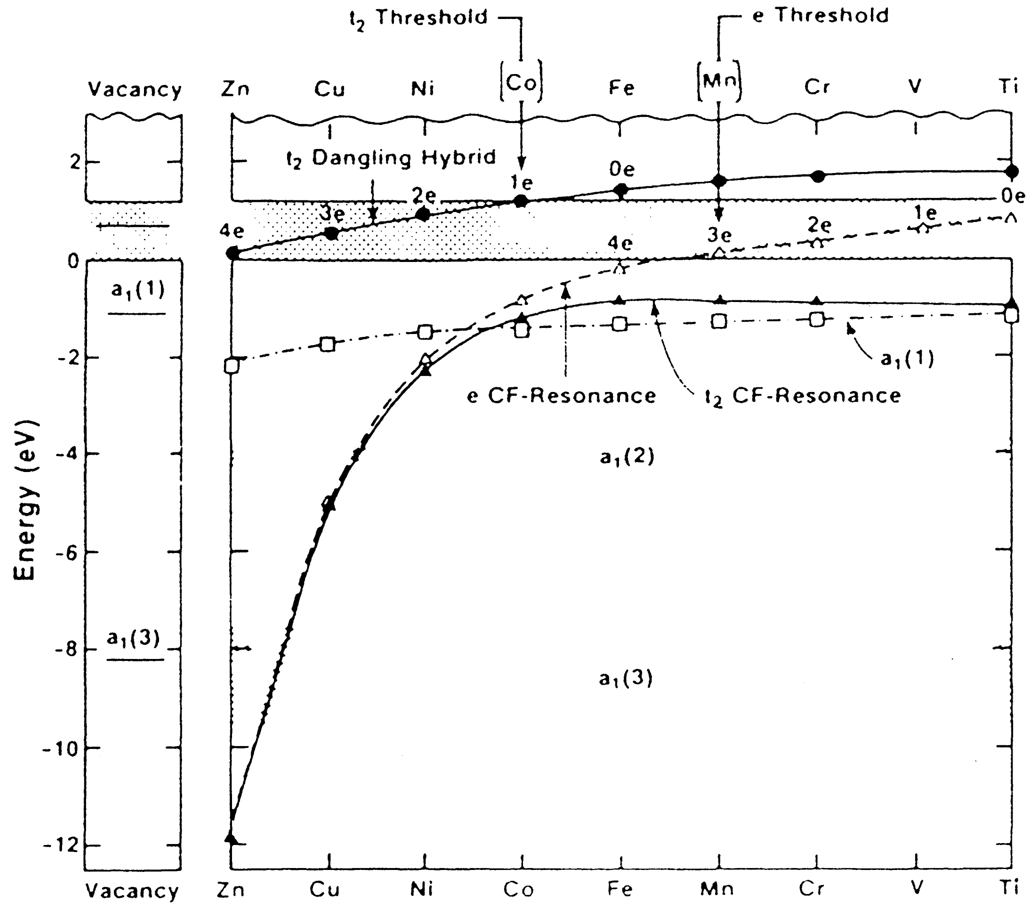

FIG. 13. The localized states in the gap region for transition-metal substitutional impurities in Si (from Zunger, 1986).

ments starting with Zn. As we already noted earlier, Zn, a group-II element, has a T_2 level in the gap that is basically the vacancy level shifted somewhat by the Zn potential. The Zn *d* electrons are near the bottom of the valence bands. Note in Fig. 13 that as we move to Cu, Ni, and so on, the *d*-impurity *d* levels move up in energy and start interacting with the vacancy levels. The T_2 part of the *d* levels interacts with the vacancy T_2 level. They mix and push apart, whereas the *E* level continues its smooth ascent in energy. It is clear that one needs a detailed calculation to establish the actual ordering and position of the levels relative to the gap. For a complete understanding of transition-metal impurities, one needs to consider further the arrangement of the electron spins and form multiplets. The interested reader may refer to the article by Zunger (1986) and references therein. Finally, transition-metal impurities may also occupy interstitial sites. The same articles contain relevant information.

Another important impurity in Si is oxygen. It is present in bulk Si that is used in the form of wafers in microelectronics and contributes to the mechanical stability of these wafers (oxygen-free Si wafers tend to warp). The lowest-energy configuration of oxygen is a bridge position between two neighboring Si atoms. The resulting Si–O–Si chain is analogous to the Si–O–Si chains that comprise SiO_2. The oxygen $2p$ orbitals form bonds and antibonds with Si sp^3 hybrid orbitals. All the available electrons occupy the bonding orbitals or the nonbonding "pi" oxygen $2p$ orbitals, all of which are low in energy, below the valence-band edge (compare with the fact that the band gap in SiO_2 is of order 10 eV). Thus the center has no electrically active levels in the Si band gap.

After annealing at about 400°C, oxygen atoms form clusters of three to four atoms with electrically active shallow levels. These centers have traditionally been known as "thermal donors," and the belief that they are small oxygen clusters comes from the fact that their concentration scales with some power n of the oxygen concentration, with n being 3–4. These centers have been studied extensively both experimentally and theoretically, but the atomic arrangements that give rise to the donor behavior were a mystery until recently (see Chadi, 1996).

Oxygen also segregates into small SiO_2 particles at the back of wafers. As these particles grow by the capture of additional migrating oxygen atoms, Si interstitials are emitted occasionally for volumetric reasons: as oxygen atoms are inserted in Si–Si bonds, the Si–Si distance stretches from 2.35 Å to about 3 Å (Ourmazd *et al.*, 1984). These SiO_2 particles are effective at capturing unwanted transition-metal impurities such as iron, cleaning the near-surface region of the wafer where devices are fabricated. This phenomenon is called intrinsic gettering.

Under electron bombardment—e.g., the experiments by Watkins described earlier in the context of vacancies—oxygen atoms in the bridge position capture a vacancy and form a vacancy-oxygen pair. The center actually consists of a substitutional oxygen atom that is displaced from the nominal site. The oxygen simply strives to form a bridge configuration with two of the neighboring Si atoms. The electronic structure of this center is simple to understand. The oxygen $2p$ orbitals bond with the two Si dangling hybrids and form Si–O–Si chains that have no levels in the gap. The other two Si dangling hybrids give rise to a state in the gap that takes up the two electrons, making this center a deep double donor.

Hydrogen is a unique impurity that has intriguing behavior in Si. It goes in interstitially and is a fast diffuser so that it has not been observed as an isolated impurity. It has been found to pair with impurities. In particular, it binds strongly with shallow acceptors and deactivates them, i.e., it forms a H–impurity pair that is no longer electrically active as a shallow acceptor (see review by Van de Walle, 1994). Theory has established that the lowest-energy site for the H atom in the case of boron impurities is along one of the B–Si bonds (see Van de Walle, 1994). There is actually a low-energy ring around the boron atom with the barrier for H motion in this ring only about 0.2 eV (see Fig. 14). Hy-

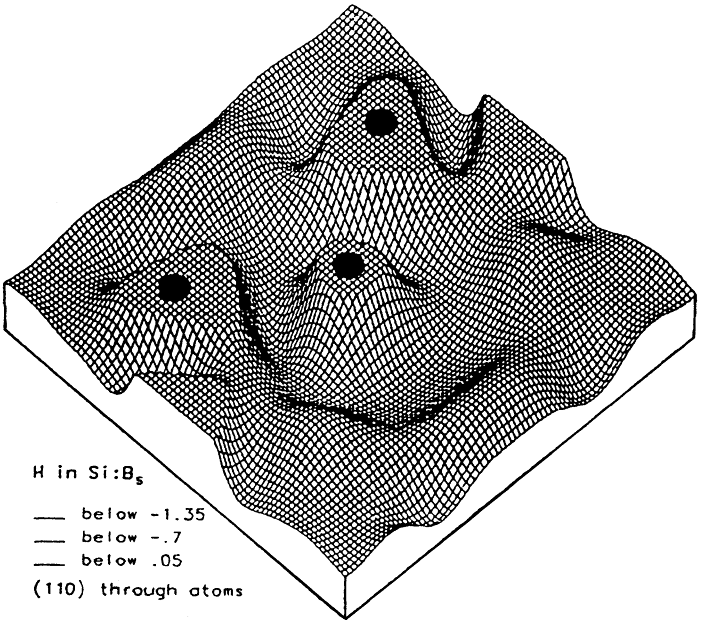

FIG. 14. Total energy for a hydrogen atom in the vicinity of a boron substitutional impurity in Si (from Denteneer *et al.*, 1989).

drogen also forms pairs with donor impurities, but the bonding is much weaker. Finally, H bonds strongly to dangling bonds at vacancies and surfaces forming Si–H bonds.

3.6 Surfaces

In Sec. 1.2 we described how atoms are arranged at the "ideal" primary Si surfaces. If we pursue the simple picture of sp^3 Si orbitals that form bonding and antibonding orbitals in the bulk perfect crystal, we recognize immediately that the creation of an "ideal" surface breaks an array of bonds, leaving the so-called dangling bonds or dangling hybrids. Every such dangling hybrid at a surface contains a single electron whose energy is now at the level of the sp^3 hybrid, much higher than the level of a bonding orbital. Reconstruction occurs because rearrangements of the surface atoms can give rise to some kind of bonding between the dangling sp^3 hybrids. Since the final geometries are not fully tetrahedral, the sp^3 hybridization of course changes, and a more complex description of the electronic states is necessary.

For a given set of atomic positions at a free surface, the electronic structure can be calculated by a variety of techniques. First-principles calculations were pioneered in the early 1970s by Joel Appelbaum and Don Hamann of Bell Laboratories (see Appelbaum and Hamann, 1976) and later by Marvin Cohen's group at Berkeley (see, e.g., Schluter *et al.*, 1975). Semiempirical tight-binding calculations, though less accurate, have the advantage of fast exploration of many structures (see, e.g., Chadi, 1979; Pandey, 1981). Photoemission spectroscopy, pioneered by Dean Eastman at IBM and William Spicer at Stanford (see Eastman and Grobman, 1972, and Wagner and Spicer, 1972, for seminal applications to Si surfaces), was developed as a powerful experimental technique to map out the dispersions of the electron states. Comparison of calculated and measured dispersions is an indirect way to explore surface reconstructions. Similarly, LEED calculations (see, e.g., Duke and Laramore, 1971, 1972) and comparisons with experimental spectra can do the same thing. Scanning tunneling microscopy, developed in the early 1980s by Gerd Binnig and Heine Rohrer of IBM's Zurich Laboratory, is a more direct experimental probe, as it produces atomic-resolution images of surfaces (see Binnig *et al.*, 1983). Finally, first-principles total-energy calculations and dynamical simulations, developed in 1985 by Roberto Car and Michele Parrinello in Trieste, allow direct exploration of reconstructions by computer.

All the above techniques plus many more played a role in unraveling the structure of Si surfaces. In their original LEED article, Schlier and Farnsworth (1959) proposed that the surface atoms of the (100) 2 × 1 surface form dimers. Other competing models were proposed over the years, but Appelbaum and Hamann's calculations showed that the dimer model yielded densities of surface states that were in best agreement with photoemission data. Angle-resolved photoemission data by Himpsel and Eastman (1979), however, found nonmetallic dispersions whereas the theoretical calculations had metallic dispersions. The puzzle was resolved by Chadi (1979) of Xerox with tight-binding calculations. He found that the total energy is lowered and the dispersions become nonmetallic when the dimers are slanted (Fig. 15). These asymmetric dimers account for the 4 × 2

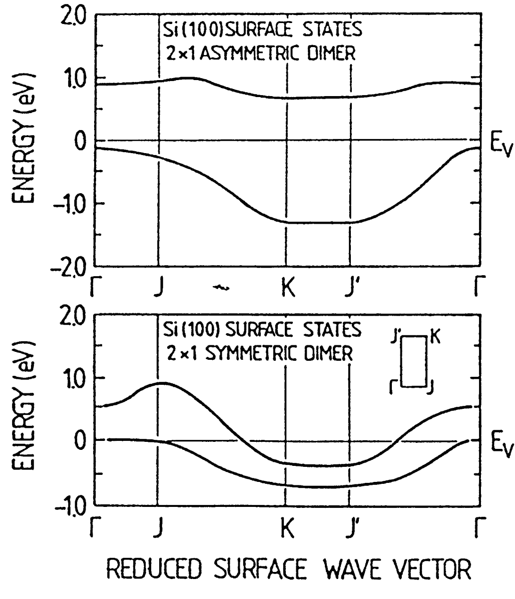

FIG. 15. The surface-state dispersions of a Si(100) 2 × 1 surface for (a) asymmetric dimers and (b) symmetric dimers. The horizontal line in each case marks Fermi energy (all levels below it are occupied by electrons). From Chadi (1979).

structure as well: the dimer slant alternates along a chain of dimers, doubling the size of the unit cell. In the last 15 years, the symmetry of the dimer has been debated, as a number of experiments seemed to find symmetric dimers. It is now well established, however, that the dimers are asymmetric and that some experiments, e.g., scanning tunneling microscopy, may in fact "see" symmetric dimers because of the way they interact with the surface.

The (111) 2 × 1 surface is another interesting story. The ideal surface has an array of dangling bonds sticking out in the ⟨111⟩ direction. Each dangling bond has one electron. In early work by Haneman (1961) it was proposed that a buckling reconstruction would set in: half of the atoms would move up and half would move down. The dehybridization of the sp^3 hybrids would make the down energy level have a lower energy so that both electrons would occupy it. First-principles calculations, however, found that the Coulombic repulsion between the two electrons worked against this picture. Furthermore, calculated dispersions did not agree with photoemission data. The puzzle was resolved by Pandey (1981). Exploring various structures with tight-binding Hamiltonians and comparing the calculated dispersions with experiment, he came upon a reconstruction of pi-bonded chains that worked best. Subsequent total-energy calculations provided confirmation. The atomic rearrangements of the top two layers are significant, but the gain from the pi bonding is sufficient to win over distortion energies.

Finally, the Si(111) 7 × 7 surface has an equally long story. There, experiment pretty much did it alone. First the STM images of Binnig *et al.* (1983) established the existence of adatoms. Subsequent electron-microscopy studies by Takayanagi *et al.* (1985) established the existence of a stacking fault underneath half of the unit cell.

4. TRANSPORT PROPERTIES

Silicon is the prototypical semiconductor and is characterized by n-type or electron conduction and p-type or hole conduction. The conductivity σ is given by

$$\sigma = ne\mu_n + pe\mu_p, \tag{13}$$

where n and p are the electron and hole concentrations, respectively, and the μ's are the corresponding mobilities. The mobility determines the effective average carrier velocity for an externally applied electric field E:

$$\bar{v} = \mu E. \tag{14}$$

In a simple theory, mobilities are given by

$$\mu = e\tau/m^*, \tag{15}$$

where m^* is the carrier's effective mass and τ is a relaxation time, i.e., the average time between collisions by the carrier with whatever scattering centers are present (phonons, impurities, etc.). In the case of electrons in Si, all six equivalent valleys contain on average the same number of carriers. The m^* that appears in Eq. (13) is the so-called conductivity effective mass that can be shown to be

$$\frac{1}{m^*} = \frac{1}{3}\left(\frac{1}{m_\parallel} + \frac{2}{m_\perp}\right). \tag{16}$$

The carrier velocity would increase linearly with the electric field if the mobility were independent of the field. That is true only for weak electric fields, which corresponds to Ohm's law. For such fields, one is interested in the dependence of τ and hence μ on the carrier effective mass and on temperature. In Si the dominant scattering processes are by acoustical phonons and by ionized impurities. The two mean collision times are then combined in the form

$$1/\tau = 1/\tau_1 + 1/\tau_2. \tag{17}$$

The mobility due to scattering by acoustic phonons is given by

$$\mu \sim m^{*-5/2}T^{-3/2}, \tag{18}$$

whereas the mobility due to scattering by ionized impurities is given by

$$\mu \sim m^{*-1/2}N_I^{-1}T^{3/2}, \tag{19}$$

where N_I is the ionized impurity density.

At low electric fields, the carriers maintain a quasi-equilibrium state with the lattice and a Maxwellian velocity distribution. Basically, as the field imparts energy to the carriers, they emit more acoustical phonons than they absorb so that they maintain the same temperature as the lattice. At fairly high fields, the carriers acquire excess energy that they do not share with the lattice very efficiently. They become "hot" with an effective temperature Θ that is higher than the lattice temperature T. The average drift velocity drops below the low-field value $\mu_0 E$ and is given by

$$\bar{v} = \mu_0 E \sqrt{T/\Theta}. \qquad (20)$$

At sufficiently high fields, the average drift velocity saturates at about 10^7 cm/s (Fig. 16).

The lifetimes of carriers are determined by the recombination processes that are active. Typically, deep impurities and other defects with localized energy levels in the band gap provide an intermediate step by first capturing a carrier at the deep level and then releasing it for recombination with a complementary carrier in the other band. Such processes can occur in principle by either the emission of light or multiphonon emission.

In the absence of deep impurities and defects, the lifetimes of carriers are primarily limited by Auger recombination; namely, an electron recombines with a hole with the band-gap energy taken up by another carrier that gets excited to a higher energy. The latter carrier subsequently loses the extra energy to the lattice by phonon emission. Experiments to measure these Auger rates are usually carried out in heavily doped but otherwise very pure, dislocation-free single crystals. Theoretical calculations (Laks *et al.*, 1990) have established that in Si, Auger recombination of electrons is a pure electronic process, whereas for holes the process is assisted by phonons. This difference is caused by the details of the energy bands. Both energy and crystal momentum (the wave vector **k**) need to be conserved during the Auger process. For electrons, it is possible to conserve both with pure electronic transitions, but for holes some of the momentum—and hence some of the energy—must be transferred to the lattice by phonon absorption.

Light emission during a recombination process is not generally efficient in Si because it has an indirect band gap. As a result, crystal momentum cannot be conserved during photon emission. In the last several years, visible luminescence has been observed from porous Si, a form of Si with large voids. It appears that this system has an effective direct gap of about 1.7 eV with large cross section for radiative recombination. A number of suggestions have been made about the origins of the effect, but detailed understanding is still incomplete.

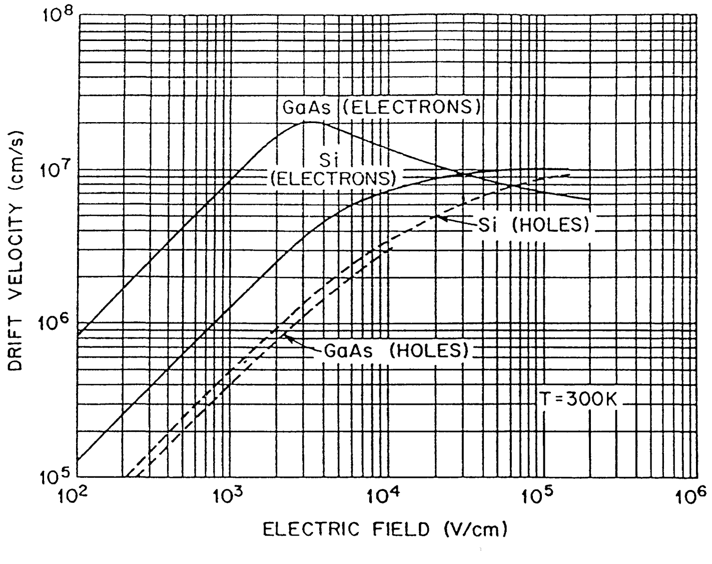

FIG. 16. The drift velocity of electrons and holes in Si and GaS as a function of electric field (from Sze, 1981).

5. THERMAL PROPERTIES

The thermal conductivity of Si has a contribution from the lattice, a contribution from electronic conduction, and a cross term from mixed electron/hole conduction. The direct contribution of conduction carriers to the thermal conductivity is quite small, but the mixed term can be large. The thermal conductivity of Si is shown as a function of temperature in Fig. 17. For comparison, the thermal conductivity of copper is also shown.

A temperature gradient can induce an electric current, and the corresponding linear coefficient P is known as the differential thermoelectric power ($J = -P\partial T/\partial x$). The thermoelectric power is negative for n-type Si and positive for p-type Si.

6. ATOMIC DIFFUSION

Self-diffusion in crystalline Si has been measured by a variety of techniques using radioactive tracers. The diffusion constant has the form

$$D = D_0 \exp(-Q/kT). \tag{21}$$

Measured values of the activation energy Q range from 4 to 5 eV, depending on the range of temperatures that was accessible to the particular experiment. There is no systematic variation of Q, however, so that the range is best viewed as an error bar.

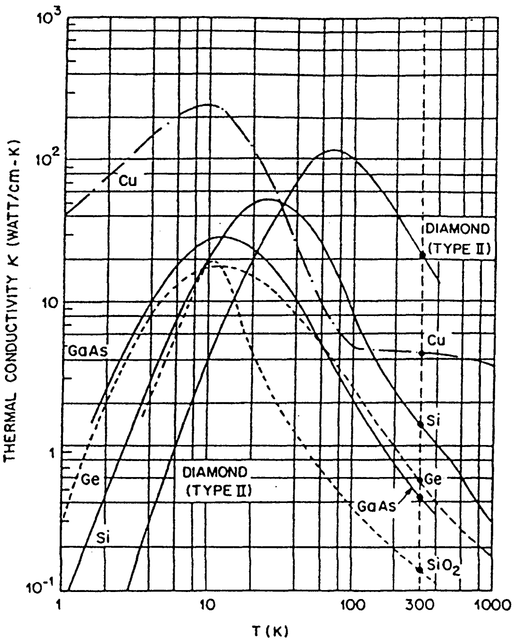

FIG. 17. The thermal conductivity of Si and a few other solids as a function of temperature (from Sze, 1981).

Self-diffusion can be mediated by vacancies or self-interstitials. When a vacancy makes a jump to a neighboring site, a Si atom makes a jump in the opposite direction. A self-interstitial can migrate in the interstitial channels and occasionally kick out a Si atom from a normal atomic site and take its place while the newly created interstitial moves on. In both these mechanisms, the activation energy for self-diffusion is equal to the sum of the defect's formation and migration energies.

Early experiments in the late 1960s and the 1970s led to serious controversies over the role of vacancies and interstitials in Si diffusion. There were strong advocates of one or the other defect as being the dominant or sole mediator of diffusion. First-principles calculations of formation and migration energies were first reported in 1984 by Car *et al.* (see Pantelides, 1994), and the surprise result was that the two defects had comparable formation and migration energies, leading to virtually identical Q's in the range of experimental values. A few years later, Pandey (1986) demonstrated with first-principles calculations that direct exchange between neighboring Si atoms without any defect is possible with an activation energy of about 4.5 eV. In order to determine the relative contributions of the three mechanisms, one would need calculations of the pre-exponentials D_0 or the full diffusion constants. Such calculations were recently carried out (Bloechl *et al.*, 1993), and the results favor the interstitial (Fig. 18).

Dopant diffusion in Si has been measured to have activation energies that are typically about 1 eV smaller than the activation energy of self-diffusion (see Fahey *et al.*, 1989). Dopant impurity diffusion can occur via precisely the same mechanisms as self-diffusion, but at least the defect-mediated mechanisms would have the same activation energy as self-diffusion. Additional mechanisms are possible, however. A vacancy and a substitutional impurity may form a pair, and then

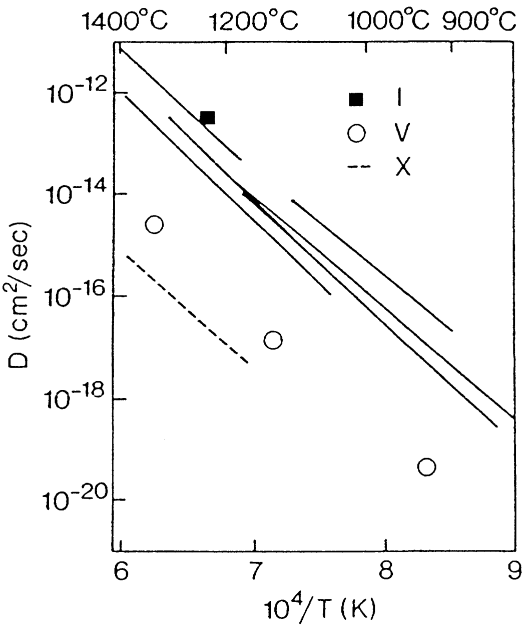

FIG. 18. The self-diffusion constant of Si for the interstitial mechanism (I), the vacancy mechanism (V), and the exchange mechanism (X) compared with experimental data (from Bloechl *et al.*, 1993).

the pair migrates. The process is rather intriguing: The vacancy migrates around a hexagonal ring of atoms and positions itself in "front" of the impurity; the impurity jumps into the vacant site; the vacancy is now at the "back" of the impurity and again migrates to the front via a hexagonal ring, and the process goes on. Calculations have demonstrated that for some impurities the formation plus migration energy of the pair is about 1 eV smaller than the self-diffusion activation energy, in agreement with experiments (Nichols *et al.*, 1989).

An analogous process is possible with a self-interstitial. The self-interstitial can take up the impurity site by pushing the impurity to the next atomic site, where the impurity in turn pushes the resident Si atom out into an interstitial site. The new interstitial can turn around and do the same thing to the impurity and so on. The self-interstitial has yet another possibility: it may simply kick an impurity out into the interstitial region and take its place at the atomic site. The impurity then can migrate in the interstitial region until it kicks back into a lattice site.

Neither experiment nor theory has been able to determine unambiguously the relative roles of vacancies and interstitials. There is sufficient evidence, however, to believe that antimony diffuses predominantly via vacancies, whereas the other impurities have strong interstitial components.

7. MECHANICAL PROPERTIES

The bulk modulus B of a solid is defined by

$$V = -B\frac{dV}{dp}, \qquad (22)$$

where V is the volume and p is the pressure. It is an elastic constant that measures the rigidity of the solid against volume changes. The compressibility is defined as the reciprocal of the bulk modulus. For crystalline Si, the bulk modulus is 0.988×10^{12} dyn/cm^2.

At absolute zero, one has $dp/dV = -d^2U/dV^2$, where U is the internal energy, so that

$$B = V\frac{d^2U}{dV^2}. \qquad (23)$$

This expression allows the direct evaluation of the bulk modulus from first-principles calculations of the total energy as a function of volume. Theoretical values are within 1% of the experimental value.

For elastic deformations that reduce the symmetry of the solid, one in general needs to introduce 21 independent elastic constants that connect the six independent strain components to the corresponding six stress components:

$$\sigma_i = \sum_{i=1}^{6} c_{ij}\epsilon_j. \qquad (24)$$

There are only 21 independent elastic constants because c_{ij} is symmetric. For a cubic solid such as Si, symmetry further reduces the number of independent elastic constants to three. In addition to the bulk modulus discussed earlier, the other two independent elastic constants are c_{12} and c_{44}. Their values for Si are 6.39×10^{11} dyn/cm^2 and 7.95×10^{11} dyn/cm^2, respectively.

Plastic deformation of Si via the generation of dislocations occurs when the material is compressed beyond the elastic limit. The external stress induces dislocations to glide. When dislocations are pinned at the two ends by some crystal defects or impurities, the external stress bows them to the point where new dislocation loops are emitted (Frank-Read sources). In addition to dislocations, large external stresses induce clusters of point defects with a deep electrical level below midgap (about 0.45 eV from the valence-band edge) that pin the Fermi level in lightly doped material. These defects have also been seen by EPR measurements. Annealing data indicate that the level is not directly associated with the dislocations (Kimerling *et al.,* 1981).

8. CRYSTAL GROWTH

8.1 Bulk Crystal

The first step in the production of high-purity single-crystal Si is the reduction of SiO_2 (quartz, common sand) with carbon at high temperatures (in excess of 2000 °C) in a submerged-electrode arc furnace. The oxygen is removed in the form of CO, and the metallurgical grade Si at this point contains 1–2% Fe, Al, and other impurities, including P and B. In the next step, metallurgical grade Si is mixed with anhydrous hydrogen chloride. A number of reactions occur that remove most of the impurities as chloride salts. Silicon is distilled in the form of trichlorosilane ($SiHCl_3$). In a subsequent step, trichlorosilane is decomposed thermally into Si and HCl in the presence of H_2. During the decomposition, Si is deposited in polycrystalline form on a high-purity thin rod of Si known as "slim rod."

The two most important industrial techniques to produce high-grade single-crystal Si are the Czochralski and the float-zone methods (see CRYSTAL GROWTH). In the Czochralski method (Czochralski, 1917), polycrystalline Si is melted in a quartz crucible in argon atmosphere. A single-crystal seed of the desired orientation is dipped into the molten Si and withdrawn from the melt at a controlled rate. The seed and the crucible are usually rotated in opposite directions for uniformity. The diameter of the pulled cylindrical material and hence the resulting wafers was initially 75 mm. Current technology uses 200-mm wafers. Today 300-mm wafers are produced and are expected to be used in microelectronics by the end of the decade.

In the float-zone method no crucible is used. Instead, a rod of polycrystalline Si is converted to single crystal by traversing a molten zone from one end in contact with a single-crystal seed to the other end. The zone melting is done with rf heating. The main difference in the quality of float-zone Si is its low oxygen content. Czochralski Si gets the oxygen and potentially other contaminant impurities from the quartz crucible. By employing vacuum in the growth chamber, float-zone Si can be grown extremely pure.

Czochralski-grown Si is doped as desired by adding appropriate elements or alloys in the crucible so that, when the polysilicon melts, the dopants are distributed uniformly. In the float-zone method, dopants are introduced in the molten zone in the form of gases. During device fabrication, selective doping is done through masks either by in-diffusion or by ion implantation (see ION-BEAM MODIFICATION OF MATERIALS) and subsequent annealing.

8.2 Epitaxial Thin Films

For device fabrication, it is advantageous sometimes to grow epitaxial thin Si films. Two processes are particularly effective. Chemical vapor deposition (CVD) [see CVD (CHEMICAL VAPOR DEPOSITION)] employs a furnace in which many wafers are exposed to a gas of silane. At the Si surface the silane decomposes, and crystalline Si is deposited. In this conventional CVD, used by the microelectronics industry, the temperature is usually high (~1050 °C) in order to keep the surface free of adsorbates. At lower temperatures polycrystalline or amorphous Si is deposited. In recent years Bernard Meyerson of IBM Research (Meyerson, 1986; Meyerson *et al.,* 1988) developed an ultrahigh-vacuum (UHV) version of CVD in which the partial pressures of water vapor and other gases are kept low enough to avoid oxidation and contamination. High-quality crystalline Si films are deposited at ~500 °C. Hydrogen plays an important role in this process by keeping the exposed surface passivated. If the temperature is increased by even 100 °C, hydrogen is removed, and the surface cannot be maintained as clean.

Another method for thin-film deposition is molecular beam epitaxy (*q.v.*) (MBE), where a beam of Si atoms is directed at a Si surface. Low-temperature (~700 °C) crystalline film deposition is again possible. Dopants can be incorporated by simultaneous beam fluxes of atoms or molecules. The drawback of this method is that a film can be deposited on a single wafer at a time by scanning the beam.

Both the UHV-CVD and MBE methods are capable of admixing dopants and of fabricating $Si_{1-x}Ge_x$ films by admixing Ge. For small x the SiGe film can be grown epitaxially on Si with Ge under lateral compression. After a certain thickness, misfit dislocations are induced to relieve the strain. The SiGe system has been studied extensively and can be used to fabricate devices that are much faster than pure-Si devices.

8.3 Silicon-On-Insulator

In microelectronics, Si devices are typically fabricated at the surface of a Si wafer. The Si material underneath the devices is not a good insulator so that devices are not thoroughly electrically isolated from one another. Parasitic transistors through the Si substrate cause a phenomenon called latchup. Technology designers have found ways to suppress the effect. Ideally, however, one would prefer to have an insulating substrate. Beginning in the early 1960s, the idea of growing thin crystalline Si films on insulating substrates has been investigated extensively. Fabrication of devices and circuits on such Si films would avoid all complications one encounters with regular Si wafers. In addition, such chips are more radiation resistant. For example, circuit failures known as soft errors can be caused by electron–hole pairs generated in the bulk Si substrate by impinging alpha particles. If the substrate is insulating, the effect is minimized. Finally, successful fabrication of circuits on thin Si films grown on insulators would allow three-dimensional stacking of devices.

The technology of silicon-on-insulator (SOI) is quite advanced today and is used for niche applications. The cost of SOI wafers is still significantly higher than regular Si wafers, but new advances may eliminate this differential soon. As the level of integration is approaching the 10^9-devices-per-chip level by the end of the century, SOI technology may play a very important role.

Epitaxial SOI has been grown successfully on sapphire (see, e.g., Vasudev, 1986) even though the thermal properties and crystal structures of Si and sapphire are quite different. The Si films are deposited by a variety of methods, including vacuum evaporation, sputtering (*q.v.*), molecular beam epitaxy, or chemical vapor deposition. The last is the most successful and is used almost exclusively in commercial applications.

Growth of Si on sapphire proceeds initially by the growth of islands that then coalesce to form a film. The nuclei are typically (100)- and (110)-oriented crystallites. The (100) crystallites, however, grow faster and eventually trap and cover the (110) nuclei. As the (100) islands coalesce, stacking faults occur at the island boundaries. As the film grows thicker, the density of defects is reduced significantly, and device-grade films can be grown.

Another approach altogether to SOI is the implantation of oxygen or nitrogen in a regular Si wafer to form a buried oxide or nitride (see Lam, 1986). The process is known as SIMOX (separation by implanted oxide). The implantation is done at about 500 °C, which leads to a good dielectric, but the Si layer at the top is quite defective and thin. After annealing, the Si film improves significantly. Additional layers of Si are then grown epitaxially to achieve the desired thickness for device fabrication.

9. PROCESSING

About ten years after the invention of the transistor, another major breakthrough led to the modern microelectronic revolution: the development of the Si planar technology that allowed the fabrication of entire circuits on the surface of a thin Si wafer. The pioneers of these developments were John Kirtley of Texas Instruments and Robert Noyce of Fairchild Semiconductor. Nowadays entire microprocessors are fabricated on a single piece of a Si wafer (chip). The fabrication of integrated circuits entails a series of processing steps, involving masking for selective diffusion, oxidation, deposition, etching, etc.

The use of photoresists and lithography is a key step for selective processing. The description of all the various processing steps used in Si technology is beyond the scope of this article (see VLSI—INTEGRATED CIRCUITS). We already discussed the deposition of epitaxial Si films in the previous section. In this section we touch briefly on oxidation, etching, and process modeling.

9.1 Oxidation

Oxidation of Si for device fabrication is done by exposing Si to O_2 gas at a high temperature (thermal oxidation; see Katz, 1983). After the initial oxidation, oxygen diffuses through the oxide to the interface where it gets incorporated. As oxygen atoms insert themselves between Si–Si bonds to form Si–O–Si bridges, the Si–Si distance increases from 2.35 Å to about 3 Å. As a result, oxidation leads to the emission of Si self-interstitials (Fahey *et al.*, 1989). The resulting interface is generally of high quality and very flat. There is a small concentration of threefold-coordinated Si atoms (dangling bonds) that are known as P_b centers.

The oxide films used for gate dielectrics in field-effect transistors are very thin and by the end of the decade are expected to be about 40 Å. This requirement is putting new constraints on the quality of the oxide and of the interface. This is a highly active area of research today. It has been found that the introduction of small amounts of nitrogen enhances the quality of the oxide and blocks hydrogen diffusion through the interface, which has detrimental effects.

9.2 Etching

The removal of Si material during device fabrication used to be done with wet chemicals. For increased control, dry etching is now the predominant method (see PLASMA ETCHING). The primary gas in the chamber is fluorine. The atomic-scale processes that result in etching have been studied extensively, both experimentally and theoretically. Initially F ions tie off dangling bonds at the Si surface. Then they penetrate into the subsurface region where they can insert themselves between neighboring Si atoms. As a result, surface Si atoms are gradually surrounded by two, three, and ultimately four F atoms. SiF_2, SiF_3, and SiF_4 molecules are ejected, partly helped by the energy of the impinging ions. F ions may also diffuse interstitially further into the bulk Si, but they can no longer insert themselves between neighboring Si atoms because of the stiffness of the lattice. The near-surface etching is assisted by the presence of holes because the F^- ion needs to get rid of its electron when it inserts itself between neighboring Si atoms (Van de Walle *et al.*, 1988). For this reason, *n*-type material etches at a faster rate because at the surface an inversion layer is formed that traps holes.

9.3 Process Modeling

The equations that describe diffusion in a crystal are well known (Fick's laws). If one assumes a particular mechanism for dopant diffusion (e.g., vacancy–impurity pairs), one can systematically write down the equations that describe the evolution of a dopant profile at high temperatures. These equations need to be solved for the particular geometries of devices. Over the last 20 years, powerful computer programs have been developed that do such modeling in one, two, and three dimensions. As the mechanisms of impurity diffusion have been elucidated over the years, these computer programs have become quite sophisticated. They are used extensively in the design and development of electronic devices and circuits.

Modeling of oxidation has also evolved significantly over the years. The modeling is based mostly on phenomenological models of growth and viscous deformation. The models are used in industrial development. Models for deposition and etching are, however, still at an early stage and are not used extensively by industry. The models are again phenomenological based on simple rate equations but are quite successful in predicting observed profiles.

Integrated circuits are very complex mechanical systems. Stresses can develop as a result of the mismatch of thermal expansion coefficients between the different materials (Si, silicon dioxide, metals). Continuum mechanics successfully models these stresses if the constitutive relations of the relevant materials are known. The latter are usually difficult to establish for thin films and complex structures.

10. SILICON-BASED MICROELECTRONICS

In the last 40 years, Si has been the semiconductor of choice for electronic devices. Though faster devices have been demonstrated using GaAs and other compound semiconductors, they are only used in niche applications. The main advantage of Si is the development of the relevant material growth and processing techniques. Though one can grow dislocation-free single-crystal Si ingots 200 mm in diameter, GaAs material is typically 75 mm in diameter. Process control in Si allows a very high level of integration, whereas GaAs lags behind by one to two orders of magnitude. Thus, even if GaAs devices are faster, that advantage is lost if a single Si chip would be replaced by 20 or more GaAs chips. The interconnects will effectively kill any device advantage.

Process-control improvements proceed at an unrelenting pace as Si technology introduces a new generation of microelectronics every three years or so. The pace is expected to continue for at least three to four more generations. Then, the feature dimensions will be so small that quantum phenomena will set in. In anticipation of this, research in the field of quantum electronics has demonstrated many quantum devices, usually fabricated with III–V compounds. The most likely scenario, however, is that Si technology will gradually adapt to and adopt quantum phenomena, and quantum microelectronics will again be Si based. It is hard to imagine that it will be possible to achieve the necessary manufacturing control with any other material. One might project that even when we get to the point when devices will be fabricated atom by atom using a version of scanning tunneling microscopy, the material is likely to be silicon.

GLOSSARY

Absolute Zero: The lowest achievable temperature where all classical motion ceases.

Acceptor: An impurity that introduces a deficiency of one electron (a hole) in a semiconductor.

Atomic Diffusion: In a solid, the ability of atoms to migrate.

Conduction Band: In a semiconductor, the lowest-energy energy band that, at the absolute zero of temperature, contains no electrons.

Donor: An impurity that incorporates an extra electron in a semiconductor.

Doping: The addition of impurity atoms to a semiconductor to alter the concentration of electrons or holes available for conduction of electricity.

Energy Bands: The bands formed by the allowed energy levels of electrons in crystals.

Epitaxy: The process of growing a crystal layer whose atomic structure is in register with that of the substrate, hence, epitaxial growth.

Hole: An empty energy level in the valence band of a semiconductor; it behaves as a particle carrying a positive charge.

Ionization Energy: In a semiconductor, the energy needed to remove a bound electron or hole to the respective energy-band edge.

Optical: Corresponding to the range of frequencies of visible electromagnetic radiation (light).

Photoemission: An experiment in which a photon is absorbed by a solid and an electron is emitted.

Space Group: A group of mathematical operations describing the translational, rotational, and mixed symmetries of a crystal.

Transistor: A device made by heterogeneous doping of a semiconductor—it acts as a solid-state amplifier or a switch. It is the key device used in microelectronics for logic operations and memory. There are two main types of transistors, known as bipolar and field-effect transistors.

Uniaxial Stress: Stress applied along one axis.

Valence Band: In a semiconductor, the highest-energy energy band that, at the absolute zero of temperature, is fully occupied by electrons.

Works Cited

Aggarwal, R. L., Ramdas, A. K. (1965), *Phys. Rev.* **140,** A1246–A1253.

Anderson, P. W. (1975), *Phys. Rev. Lett.* **34,** 953–955.

Appelbaum, J., Hamann, D. R. (1976), *Rev. Mod. Phys.* **48,** 479–496.

Baraff, G. A., Schluter, M. (1978), *Phys. Rev. Lett.* **41,** 892–895.

Baraff, G. A., Kane, E. O., Schluter, M. (1979), *Phys. Rev. Lett.* **43,** 956–959.

Bardeen, J., Brattain, W. (1948), *Phys. Rev.* **74,** 230–231.

Bernholc, J., Lipari, N. O., Pantelides, S. T. (1978), *Phys. Rev. Lett.* **41,** 895–899.

Binnig, G., Rohrer, H., Gerber, C., Weibel, E. (1983), *Phys. Rev. Lett.* **50,** 120–123.

Blöchl, P., Smargiassi, E., Car, R., Laks, D. B., Andreoni, W., Pantelides, S. T. (1993), *Phys. Rev. Lett.* **70,** 2435–2438.

Bourgoin, J., Corbett, J. W. (1972), *Phys. Lett.* **38A,** 135–137.

Bourret, A., Rouviere, J. L. (1987), in: *Proceedings of the International Symposium on Polycrystalline Semiconductors, Grain Boundaries, and Interfaces,* Berlin: Springer-Verlag, p. 8.

Brown, F. C., Rustgi, O. P. (1972), *Phys. Rev. Lett.* **28,** 497–500.

Burns, G. (1985), *Solid State Physics,* Orlando, FL: Academic.

Car, R., Parrinello, M. (1985), *Phys. Rev. Lett.* **55,** 2471–2474.

Chadi, D. J. (1979), *Phys. Rev. Lett.* **43,** 43–47.

Chadi, D. J. (1996), *Phys. Rev. Lett.* **77,** 861–864.

Chelikowsky, J. R., Chadi, J. D., Cohen, M. L. (1973), *Phys. Rev. B* **8,** 2786–2794.

Chelikowsky, J. R., Cohen, M. L. (1976), *Phys. Rev. B* **14,** 556–582.

Czochralski, J. (1917), *Z. Phys. Chem. Stöchiom. Verwandtschafts.* **92,** 219–221.

Denteneer, P. J. H., Van de Walle, C. G., Pantelides, S. T. (1989), *Phys. Rev. B* **39,** 10809–10824.

Duke, C. B., Laramore, G. E. (1971), *Phys. Rev. B* **2,** 4765–4782.

Duke, C. B., Laramore, G. E. (1972), *Phys. Rev. B* **3,** 3183–3197.

Eastman, D. E., Grobman, W. D. (1972), *Phys. Rev. Lett.* **28,** 1378–1381.

Fahey, P., Griffith, P. B., Plummer, J. D. (1989), *Rev. Mod. Phys.* **61,** 289–384.

Grimmeiss, H. G., Janzen, E. (1994), in: S. T. Pantelides (Ed.), *Deep Centers in Semiconductors,* New York: Gordon & Breach.

Haneman, D. (1961), *Phys. Rev.* **121,** 1093–1100.

Harrison, W. A. (1980), *Electronic Structure and the Properties of Solids,* San Francisco: W. H. Freeman and Co.

Himpsel, F., Eastman, D. E. (1979), *J. Vac. Sci. Technol.* **16,** 1297–1299.

Hirth, J. P., Lothe, J. (1992), *Theory of Dislocations,* Malabar, FL: Krieger Publ.

Hohenberg, P., Kohn, W. (1964), *Phys. Rev.* **136,** B864–B871.

Hybertsen, M., Louie, S. G. (1985), *Phys. Rev. Lett.* **55,** 1418–1421.

Kamins, T. (1988), *Polycrystalline Silicon for Integrated Circuit Applications,* Boston: Kluwer Academic Publ.

Katz, L. E. (1983), in: S. M. Sze (Ed.), *VLSI Technology,* New York: McGraw-Hill.

Kimerling, L. C., Patel, J. R., Benton, J. L., Freeland, P. E. (1981), in: R. R. Hasiguti (Ed.), *Defects and Radiation Effects in Semiconductors 1980,* London: Institute of Physics.

Kohn, W. (1957), in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics,* Vol. 5, New York: Academic, p. 257.

Kohn, W., Luttinger, J. M. (1955), *Phys. Rev.* **98,** 915–922.

Laks, D. B., Neumark, G. F., Pantelides, S. T. (1990), *Phys. Rev. B* **42,** 5176–5185.

Lam, H. W. (1986), in: B. J. Baliga (Ed.), *Epitaxial Silicon Technology,* Orlando, FL: Academic.

Maiti, A., Chisholm, M., Pennycook, S. J., Pantelides, S. T. (1996), *Phys. Rev. Lett* **77,** 1306–1309.

Meyerson, B. (1986), *Appl. Phys. Lett.* **48,** 797–799.

Meyerson, B. S., Uram, K. J., LeGues, F. K. (1988), *Appl. Phys. Lett.* **53,** 2555–2557.

Nara, H. (1965), *J. Phys. Soc. Jpn.* **20,** 778–784.

Nichols, C. S., Van de Walle, C. G., Pantelides, S. T. (1989), *Phys. Rev. Lett.* **62,** 1049–1052.

Ourmazd, A., Schröter, W., Bourret, A. (1984), *J. Appl. Phys.* **56,** 1670–1681.

Pandey, K. C. (1981), *Phys. Rev. Lett.* **47,** 1913–1917.

Pandey, K. C. (1986), *Phys. Rev. Lett.* **57,** 2287–2290.

Pantelides, S. T. (1986), *Phys. Rev. Lett.* **57,** 2979–2982.

Pantelides, S. T., Selloni, A., Car, R. (1985), *Solid State Electron.* **28,** 17–24.

Pantelides, S. T. (1994), in: S. T. Pantelides (Ed.), *Deep Centers in Semiconductors,* New York: Gordon & Breach.

Penn, D. R. (1962), *Phys. Rev.* **128,** 2093–2097.

Phillips, J. C. (1973), *Bonds and Bands in Semiconductors,* New York: Academic.

Schlier, R. E., Farnsworth, H. E. (1959), *J. Chem. Phys.* **30,** 917–926.

Schlüter, M., Chelikowsky, J. R., Louie, S. G., Cohen, M. L. (1975), *Phys. Rev. B* **12,** 4200–4214.

Shockley, W. (1949), *Bell Syst. Tech. J.* **28,** 435–489.

Sze, S. M. (1981), *Physics of Semiconductor Devices,* 2nd ed. New York: Wiley.

Takanayagi, K., Tanishiro, Y., Takahashi, M., Takahashi, S. (1985), *J. Vac. Sci. Technol. A* **3,** 1502–1506.

Van de Walle, C. G. (1994), in: S. T. Pantelides (Ed.), *Deep Centers in Semiconductors,* New York: Gordon & Breach.

Van de Walle, C. G., McFeeley, F. R., Pantelides, S. T. (1988), *Phys. Rev. Lett.* **61,** 1867–1870.

Vasudev, P. K. (1986), in: B. J. Baliga (Ed.), *Epitaxial Silicon Technology,* Orlando, FL: Academic.

Wagner, L. F., Spicer, W. E. (1972), *Phys. Rev. Lett.* **28,** 1381–1384.

Watkins, G. D. (1994), in: S. T. Pantelides (Ed.), *Deep Centers in Semiconductors,* New York: Gordon & Breach.

Watkins, G. D., Corbett, J. W. (1961), *Phys. Rev.* **121,** 1001–1014.

Zunger, A. (1986), in: H. Ehrenreich, D. Turnbull (Eds.), *Solid State Physics,* Vol. 39, New York: Academic, p. 276.

Further Reading

Böer, K. W. (1990), *Survey of Semiconductor Physics,* New York: Van Nostrand Reinhold.

Harrison, W. A. (1980), *Electronic Structure and the Properties of Solids: The Physics of Chemical Bonds,* San Francisco: Freeman.

Lannoo, M., Bourgoin, J. (1981), *Point Defects in Semiconductors,* Berlin: Springer-Verlag.

Moss, T. S. (1980), *Handbook on Semiconductors,* Amsterdam: North Holland.

Muller, R. S., Kamins, T. I. (1986), *Device Electronics for Integrated Circuits,* 2nd ed. New York: Wiley.

SILICON, POLYCRYSTALLINE

MAURIZIO ARIENZO, *IBM Research Division, T.J. Watson Research Center, Yorktown Heights, New York, U.S.A.*,

WENDY A. ORR-ARIENZO, *IBM Microelectronics Division, East Fishkill, New York, U.S.A.*

Introduction 217
1. **Polysilicon Deposition and Microstructure** 218
1.1 Polysilicon Deposition 218
1.2 Production Reactors 219
1.3 Safety Considerations 220
1.4 Material Characterization 220
2. **Impurity Doping, Diffusion, and Electrical Properties** 222
2.1 Dopant Incorporation 223
2.2 Dopant Diffusion 223
2.3 Dopant Segregation and Activation 225
2.4 Electrical Properties—Resistivity 225
2.5 Electrical Properties—Carrier Mobility 228
2.6 Diffusion through Polysilicon/Single-Crystal Silicon Interfaces 228
3. **Polycrystalline Silicon Applications** 229
3.1 Silicon-Gate Technology 229
3.2 Polysilicon-Emitter Bipolar Transistors 230
3.3 Other Polysilicon Devices and Applications 231
3.3.1 Trench Isolation 231
3.3.2 Trench Capacitors 231
3.3.3 Resistors and Interconnections 232
4. **Summary** 232
Glossary 232
Works Cited 232
Further Reading 233

INTRODUCTION

Polycrystalline silicon, also referred to as polysilicon, is defined as the state of aggregation of silicon in which a number of single crystals (grains) of various sizes and shapes are intimately networked via grain boundaries. Polysilicon is differentiated from amorphous silicon, since long-range order is present. It is differentiated from single-crystal silicon by the fact that the various grains have different orientations, generating grain boundaries at the crystalline discontinuities. Polysilicon plays a key role in microelectronics. In fact, much of the success of silicon integrated circuits is due to the properties of polysilicon coupled with those of silicon dioxide and silicon nitride. Since it is fully compatible with the high-temperature silicon integrated-circuit processing, it can be deposited during microfabrication on either single-crystal silicon or its conventional dielectrics, silicon dioxide or silicon nitride, creating semiconducting or, by heavily doping, conducting layers. Consequently, it is used as an electrical "contact" to silicon layers, as a capacitor plate on top of oxide [the "gate" of metal-oxide-semiconductor (MOS) transistors], and by running polysilicon lines on top of isolation, either as a local "interconnect" between devices in integrated circuits when doped to low resistivity or as a "resistor" when doped to the required resistivity. Polysilicon is now widely used in CMOS (complementary metal-oxide-semiconductor) technology as the gate, as the contact to both the source and drain, and as the local interconnect between transistors. It is also used to contact both the base and the emitter of bipolar transistors, and as a dopant diffusion source for the extrinsic base and for the single-crystalline emitter. Other applications in microelectronics include its application as a refill and spacer material. In addition to integrated-circuit applications, polysilicon is being employed in solar cells and in thin-film transistors used in flat-panel

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

displays. While most of the considerations in this article apply to the material used in both these applications, the reader is referred to the articles Photovoltaic Devices, Solar Energy, and Display Technology for more detailed descriptions.

1. POLYSILICON DEPOSITION AND MICROSTRUCTURE

Although polycrystalline silicon (polysilicon) can be deposited by sputtering and evaporation, chemical vapor deposition (CVD) is the most common deposition method used in the electronic industry. CVD is the process in which the solid products of a chemical reaction are deposited on a substrate as a thin film. See CVD (Chemical Vapor Deposition). Silicon CVD is commonly obtained by the pyrolyzation of silane (SiH_4) according to the reaction

$$SiH_4(s) \rightarrow SiO_2(s) + 2H_2(g) \qquad (1)$$

or of the silicon halides, primarily dichlorosilane ($SiHCl_2$), trichlorosilane ($SiHCl_3$), and silicon tetrachloride ($SiCl_4$). Pyrolyzation of the halides has been used for years in the production of hyperpure silicon feedstock in the single-crystal silicon wafer manufacturing industry. This process has been particularly successful because of the purity control and its low-cost economics. By pyrolyzing either silane or any of the halides, single-crystal silicon, polysilicon, or amorphous silicon can be obtained, depending on the deposition conditions and on the substrate on which the deposition occurs. Uniform films of high quality and purity can be obtained by CVD, a clear advantage with respect to the films obtained by physical vapor deposition (PVD) methods, such as vacuum evaporation, sputtering, and molecular-beam epitaxy (MBE). PVD is the process in which the material to be deposited is passed into the vapor transport phase by a physical mechanism, like evaporation, sublimation, or ion bombardment. Also, in contrast to the directionality of all PVD techniques, CVD exhibits "conformality," or the ability to deposit a film of uniform thickness over topology or steps. Conformality and reproducibility, key requirements in today's application to device structures, have been the driving forces for the increased application of polysilicon CVD in the microelectronics industry.

1.1 Polysilicon Deposition

Uniquely in CVD, the composition of the transport vapors and of the deposited film are different. Since the chemical reactions involving the transport vapors usually occur at or near the surface, the opportunity for incorporation of contaminants from the gas phase is greatly reduced. In addition, the purity of the gases used for CVD has been steadily increasing over the past decade as a result of technological advances and user demand. The growth rate of the deposited film in a CVD reactor is limited by several kinetic barriers: the gas-phase mass transport of the species to the surface; the adsorption of at least one reactant onto the surface; the reaction on the surface, which may involve several steps each with its own rate; and finally the flux of the reaction by-products away from the surface. Polysilicon is deposited at temperatures between 600 and 650 °C in a low-pressure reactor, in a regime limited by surface reactions rather than adsorption, to obtain uniform nucleation and thus uniform thickness and conformality independently of the material on which it is grown. In this temperature and pressure regime, during the initial phases of the polysilicon growth on any surface, small clusters of adsorbed atoms diffuse on the surface until they reach nucleation sites, where they grow to form polycrystalline films. The amount of surface diffusion is influenced by many parameters, including temperature, pressure, and flow rates. Figure 1 shows the deposition rate as a function of temperature, also called Arrhenius plot, for polysilicon growth for different silane partial pressures (Adams, 1983). As can be observed, the deposition rate increases rapidly as the temperature increases and, for a given temperature, increases, as expected, with increasing silane partial pressure. At lower temperatures, the deposition rate is too slow to be practical, and a film containing microcrystalline silicon embedded in an amorphous network can be obtained. At higher temperatures, gas-phase reactions and silane depletion become significant, and nonuniform films are obtained.

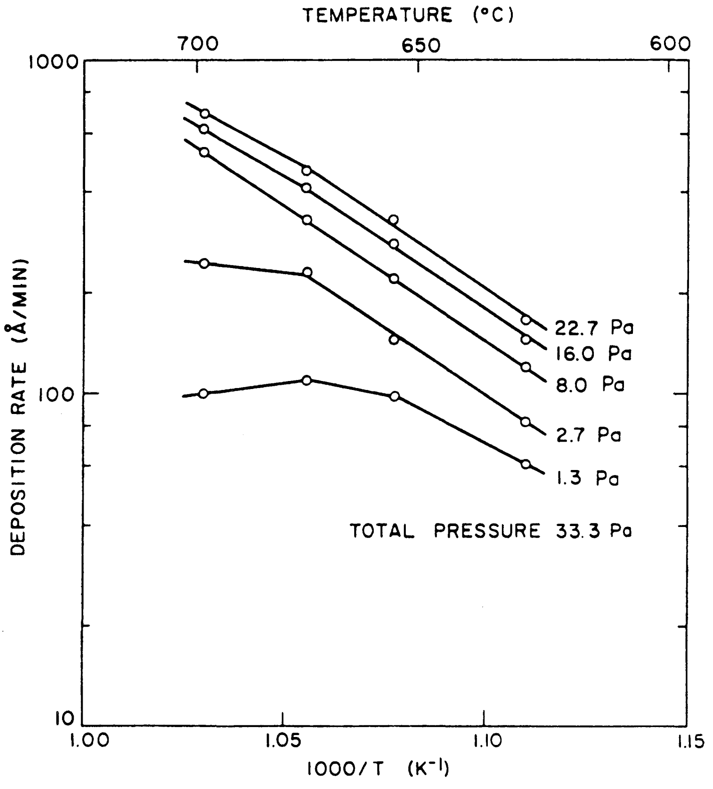

FIG. 1. Arrhenius plot for polysilicon deposition for different silane partial pressures (Adams, 1983).

1.2 Production Reactors

Chemical vapor deposition is performed at low pressures, between 0.1 and 1.0 Torr. Low-pressure CVD (LPCVD) reactors use standard diffusion furnace equipment modified to operate at lower pressures and are described in Sec. 2.2 of the article CVD (CHEMICAL VAPOR DEPOSITION) in this Encyclopedia. The excellent temperature controls present in diffusion and oxidation equipment are used to guarantee good temperature uniformity both across the wafers and from wafer to wafer. These reactors are hot-wall reactors, since the tube walls are at the same temperature as the wafers. As a result, they require periodic cleaning because peeling and particle generation become a concern.

LPCVD is performed at temperatures between 600 and 650 °C, using silane at high inlet reactant-gas mole fraction (little or no carrier gases). These two conditions, together with the lower pressure, cause the rate of the chemical reaction at the wafer surface to be the dominant kinetic mechanism for the growth (Adams, 1983; Rosler, 1977; Kern and Schnable, 1979). It has been demonstrated that virtually no decomposition of the silane source occurs in the gas phase because these reactions are usually surface nucleated (or heterogeneous) given the temperature uniformity inherent in resistance-heated furnaces. The deposited films have superior uniformity and stoichiometry, and very low defect densities. Furthermore, particulates are strongly reduced because of the lack of gas-phase nucleation.

Uniformities in the ±3% range (2σ) are possible both across each wafer and wafer to wafer for a standard LPCVD polysilicon deposition system. Since the interwafer distance is of the same order of magnitude as the molecular mean free path in the gas, while mass transfer is greatly enhanced at reduced pressure, more than 100 wafers of 200-mm diameter can be deposited using standard diffusion quartzware with a growth rate in the 10- to 20-nm/min range.

A typical LPCVD reactor is depicted in Fig. 2. The gases are injected at the system door and exhausted at the rear through a trap using either a mechanical pump or a

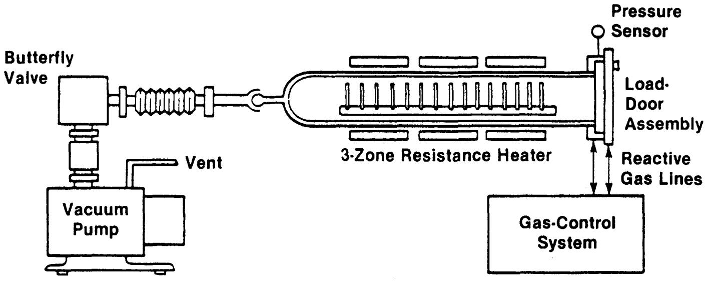

FIG. 2. Schematic of a low-pressure CVD (LPCVD) reactor used for batch deposition of polysilicon in VLSI factories.

booster pump/mechanical pump combination. Pressure control, usually between 0.2 and 1 Torr, requires control of the pumping speed. This can be achieved by introducing a butterfly valve before the mechanical pump, by varying the speed of the booster, or simply by injecting some inert gas (e.g., N_2) before the mechanical pump. However, the latter approach is discouraged because the deposition rate is a strong function of the silane partial pressure, and it is quite difficult to vary the flows and keep the partial pressures constant, especially in the presence of carrier or dopant gases. As discussed earlier, it is the partial pressure of the reactant gases (product of the mole fraction by the total pressure) that is important in determining the deposition rate. The growth rate can be significant at low pressure because of the lack of a carrier gas.

To compensate for silane depletion along the tube caused by the consumption of the reacted gas, most of the processes use a "tilted" temperature profile, with the temperature increasing toward the rear of the tube. This may cause some undesirable effects, such as different grain sizes for wafers at different tube positions, due to the high temperature sensitivity of the growth rate in the temperature range used. Polysilicon deposition using a "flat" temperature profile is now available either by diluting silane, usually 20–30% in nitrogen, increasing the total flow in the system, or by using gas manifolds inside the system. The latter are quartz tubes with appropriately spaced holes to supply reactant gases at different positions inside the tube.

1.3 Safety Considerations

It is well known that silane is explosive in concentrations in excess of 2% in air. To obviate this problem, processes using 2% diluted silane have been developed. If a flow-limiting device is included at the gas cylinder, the risk of explosion in case of a line rupture is reduced. Some cylinder manufacturers now provide such standardized devices on their cylinders. Additional other precautions, needed not only to improve safety but also to reduce the contamination caused by leakage and other sources, have been summarized by Hammond (1980).

The dopant gases used for *in situ* doping (phosphine, arsine, and diborane) are also very toxic, and in LPCVD additional safety problems are associated with the dissolution of the gases in the pump oil. A useful approach is to include a post reactor either in the back portion of the system or as a separate unit in the source cabinet before the pumping stages. Some LPCVD system manufacturers are offering such systems in their units.

1.4 Material Characterization

There are two types of characterization of polysilicon:

1. routine tests that are performed on one or more test wafers on each run (lot monitors) and
2. periodic tests that are performed either in process development or in quality control/process validation.

A standard method to characterize a polysilicon deposition process will include the analysis of the thickness and the resistivity and their uniformity both across the wafer and wafer to wafer across the system load on lot-monitor wafers. This is usually done by using oxide monitors (wafers with approximately 100 nm of silicon dioxide) on which the thickness can be easily measured with either an ellipsometric or an interferometric technique. If the process is *in situ* doped, dopant uniformity is measured by sheet resistance measurement after an appropriate anneal to activate the dopant. This anneal should be as close as possible to the thermal processing that the doped polysilicon will see in the device process since, as will be discussed later, sheet resistance is a complex function of microstructure, dopant distribution, and dopant electrical activation.

Microstructure, dopant profile, and carrier profile are monitored only periodically. By far, the most common characterization technique for polysilicon microstructure is transmission electron microscopy (TEM). A good review of the technique is contained in the book by Thomas and Goringe (1979). TEM is used for grain-size and texture analysis. In a conventional TEM sample prepared by thinning the polysilicon, the various orientations of the grains are made visible through light- and dark-field imaging, highlighting the grain boundaries. By analyzing the diffraction spots, the orientations of the grains can be obtained, and preferential alignment of the grains can be evaluated. In Fig. 3, TEM pictures of a polysilicon film before [Fig. 3(a)] and after [Fig. 3(b)] a thermal treatment are shown. As can be observed, the grains have grown considerably during the thermal treatment (from an average grain size of 30–100 nm for a 90-nm-thick film).

If the TEM sample is prepared in cross section, the presence of grain boundaries in the direction perpendicular to the surface can be verified as well as the presence or absence of voids. Figure 3(c) shows a cross section of the as-deposited film in 3(a). When a sample is viewed in cross section, the thickness of extremely thin films can be accurately determined by use of the directly resolved lattice fringe spacings of the substrate. Using x-ray analysis, more information on the texture and orientation can be obtained, together with, for example, the chemical analysis of inclusions possibly present.

With the introduction of high-resolution TEM and lattice imaging, polysilicon/single-crystal interfaces have been studied (Bravman *et al.*, 1985) to determine the chemical composition of the interface, the presence or absence of an interfacial oxide, the continuity of the interface, and the epitaxial realign-

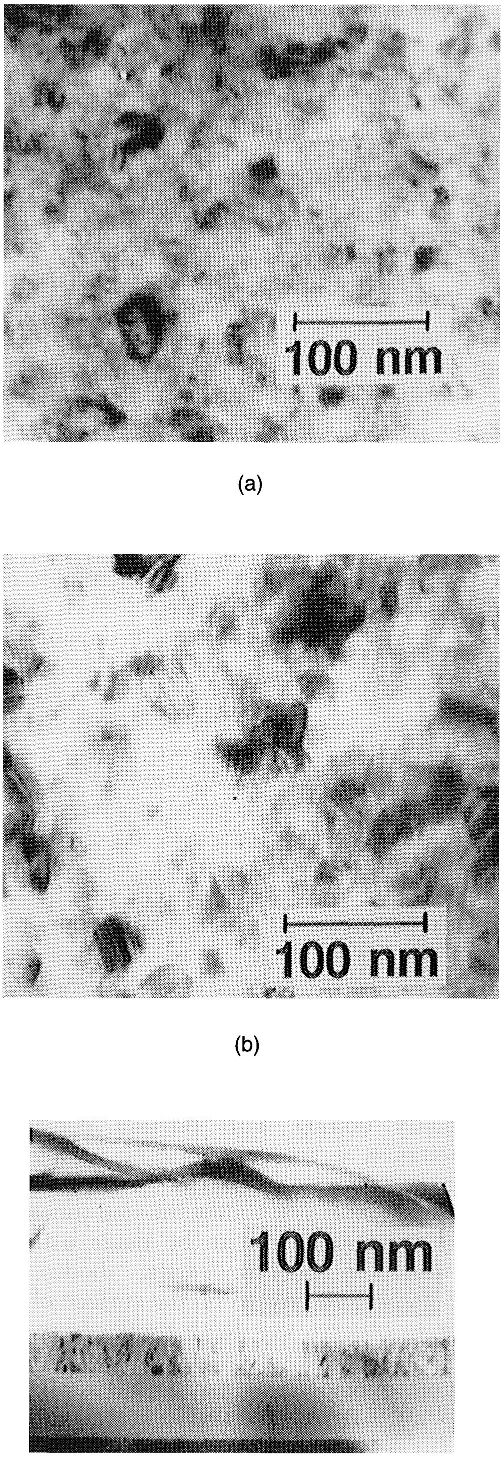

(c)

FIG. 3. Transmission electron micrograph of polysilicon grown on oxide (a) as deposited and (b) after thermal treatment, and (c) cross section as deposited.

ment of the grains at the interface, as we discuss in Sec. 2.6. Finally, combining scanning TEM (STEM) observations with energy-dispersive chemical analysis (EDX) on cross sections, an analysis of the interfacial region is feasible, collecting information on the amount of dopant incorporated there.

Two main dopant profiling techniques are used: Rutherford back-scattering (RBS) and secondary-ion mass spectroscopy (SIMS) (Marcus, 1983). RBS has advantages for dose determination, since it is an absolute measure of the chemical quantity. It cannot be used for boron and lacks sensitivity for arsenic. For doping profiling, RBS lacks resolution where the doping changes by more than one order of magnitude within 20 nm or less. SIMS is the preferred technique for B, As, or P determination, and, since resolutions of 5–10 nm per order of magnitude of dopant change can be achieved, it is also used to study the distribution of dopants in devices where the impurity profiles are very steep.

For carrier profiling, several techniques are used: spreading resistance, junction depletion capacitance, and differential Hall together with the sheet-resistance measurement. Spreading resistance is a well-known method to obtain an electrical depth profile. This technique uses a two-point probe running along a beveled sample edge to identify from the resistivity and type the profile of the active carrier concentration. This method can be used accurately for layers greater than 0.5 μm in depth. However, doping density cannot be accurately inferred in polysilicon because of electrically inactive grain-boundary doping. For junction depletion capacitance, a junction must be formed to determine the active carrier profile on the low-doped side of a unilateral step junction. These measurements can be made using a *p-n* junction, Schottky-barrier diodes, or MOS capacitors formed on the surface of the wafer to observe the depth profile from the surface into the substrate. This method is strongly influenced by the doping level and has limited application because of the junction breakdown.

Hall measurement is the most precise method to determine the mobility and the average active carrier concentration in polysilicon. The two previous methods, in fact, lead to questionable results as a result of the fields present at the grain boundaries and the difficulty of probing single grains. If polysilicon is deposited on thermal oxide, a Hall pattern can be defined by photomasking. A bulk measurement will give information on the bulk resistivity, bulk Hall mobility, and bulk carrier concentration. To find the carrier concentration and mobility profiles within a polysilicon (or single-crystal) deposited film, two methods can be used. In one, various thicknesses can be deposited and the above measurement repeated each time, obtaining the profiles from the data by a single transformation (Kamins, 1971). Another method used (Natsuaki *et al.*, 1983) is to etch layers of the polysilicon film with a dry or wet isotropic etchant and repeat the measurements each time. Again, the profiles can be simply obtained from the measurements. Some of the problems with these two methods are the grain-size variations between films of different thicknesses subjected to the same heat treatment and the difficulty in etching polysilicon isotropically without preferential grain-boundary attack.

2. IMPURITY DOPING, DIFFUSION AND ELECTRICAL PROPERTIES

The first applications of polysilicon in very large-scale integration (VLSI) have been for silicon-gate CMOS transistors and for high-value resistors. In both cases, uniform heavy doping concentration is necessary throughout the polysilicon: heavy in the first case, controlled in the second. More recently, polysilicon has also found a large number of applications as a diffusion source for both bipolar and CMOS processes. By depositing and etching a pattern in a doped polysilicon layer, a self-aligned diffusion can be obtained only in the silicon underneath the patterned polysilicon. The polysilicon itself can also act as the contact to the diffusion and can be used as an interconnection level. Doping of and diffusion in polysilicon have been extensively investigated.

Polysilicon can be doped by diffusion from a solid, liquid, or gaseous source, by ion implantation into the polysilicon, and by the addition of dopant gases during the deposition process (*in situ* doping). The structure of the polysilicon will be influenced by the presence of the dopant, and this, in turn, will influence the dopant diffusion characteristics in the polysilicon, the dopant activation, the dopant segregation, and the carrier trapping at the grain boundary.

None of the three commonly used dopants (B, P, and As) behave in quite the same way. These differences are further enhanced by varying grain-boundary composition and impurity content that are obtained by depositing in various systems and by various techniques. In this section the methods used to incorporate dopants in polysilicon, the thermal diffusion in polysilicon, and the segregation of dopants are discussed, with emphasis placed on the grain-boundary effects. Diffusion throughout the polysilicon/silicon interface is deferred to Sec. 2.6, after the discussion of the interface properties.

2.1 Dopant Incorporation

Gas and solid diffusion sources have been used to incorporate dopants in polysilicon. $POCl_3$ has been used routinely to dope the polysilicon to be used as a gate for CMOS transistors with phosphorus (Kamins, 1971, 1979). Solid sources (such as boron nitride wafers) have been used to simplify the processing steps. However, the most precise method to incorporate dopants in polysilicon is ion implantation. The advantages of ion implantation include the capability to measure accurately the number of ions incorporated (dose) and to adjust the depth at which the ions are implanted by regulating the ion energy (range). For each ion there will be a projected range of penetration, R_P, for every energy used. The depth profile of the implanted ions is approximately the same as in single-crystal silicon, the only difference being the possibility of channeling in aligned grains. However, significant differences in the profiles arise during the subsequent heat treatments. By implantation of arsenic and phosphorus (heavier ions), an amorphous layer is formed within the projected implantation range R_P for doses exceeding the mid-10^{14} cm^{-3} regime. During the subsequent anneal, this amorphous layer will rapidly recrystallize leading, in the case of heavy doping, to large grain growth or even defective crystalline layers.

Variations in polysilicon thickness and the possibility of channeling can affect the dose of dopants incorporated into the polysilicon. Preamorphization of the polysilicon layer or implantation with a tilt angle are also used in polysilicon. Contamination from the implanters (usually heavy metals) can be avoided by implanting through an oxide.

In situ doping can be achieved by the addition of arsine (AsH_3), diborane (B_2H_6), or phosphine (PH_3) during the deposition process. The main advantage with respect to ion implantation is process simplification. In some applications—for example, when polysilicon is used as a plate of the trench capacitors of advanced CMOS circuits (Goto and Inayoshi, 1987) (see Sec. 3.3.2)—*in situ* doping has the distinct advantage of providing the doping inside the trenches. For deep trenches, the only alternatives for doping incorporation are lengthier processes, deep implantation, and long anneals (Arienzo *et al.*, 1986; White *et al.*, 1985).

The addition of dopant sources in the reactors during growth introduces more species that compete for the available adsorption sites. Diborane increases the deposition rate, while the addition of either arsine or phosphine decreases it. For example, in the presence of phosphine, a self-limiting surface layer composed of a layer of adsorbed phosphorus or phosphine has been observed with a variety of UHV probes, reducing the adsorption and decomposition of silane at the surface (Meyerson and Yu, 1984). Diborane, on the other hand, showed no detectable adsorption on the silicon surface.

2.2 Dopant Diffusion

In single-crystal silicon, dopant diffusion proceeds by interaction of the dopant atoms with point defects: vacancies and interstitials. In the case of polysilicon, the grain boundaries play a very significant role in dopant diffusion. This is because grain boundaries provide paths along which dopant atoms can diffuse with less activation energy than in the grains. Dopant atoms can move from within the grains to the grain boundaries, proceed along the grain boundaries, and then diffuse back into the grains. Diffusion kinetics in polycrystalline thin films are discussed at length in Sec. 2 of the article DIFFUSION IN THIN FILMS. Since the electrical activation of the dopant atoms is different in the grains as compared to the boundaries, as discussed in Sec. 2.3, it is im-

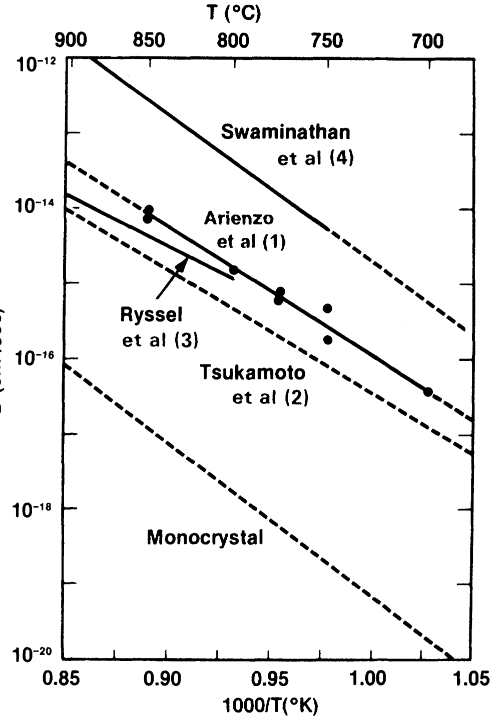

FIG. 4. Diffusion of arsenic in single-crystal and polycrystalline silicon. Experimental data and extrapolations from (curve 1) Arienzo *et al.* (1984), (curve 2) Tsukamoto *et al.* (1978), (curve 3) Ryssel *et al.* (1981), (curve 4) Swaminathan *et al.* (1982).

portant to understand the diffusion coefficients in polysilicon. In the case where polysilicon is used as a diffusion source into single-crystal silicon, the diffusion profile in the single crystal will also be determined by the properties of the polysilicon/silicon interface, as is discussed in Sec. 2.6. Various workers have determined experimentally the diffusion coefficients in polysilicon for arsenic, phosphorus, and boron. The formula governing diffusion is

$$D = D_0 e^{-E_a/kT}, \qquad (2)$$

where D_0 is the pre-exponential constant term, E_a is the activation energy of the diffusion process, k is the Boltzmann constant in eV/K, and T is the absolute temperature. The values obtained for both the pre-exponential constant and the activation energy of the diffusion coefficient of a particular dopant species cover a considerable range. Figure 4 is a semilogarithmic plot of the experimental results and extrapolated data to low temperature obtained in the case of arsenic, and Table 1 is a compilation of the data found in the literature (Tsukamoto *et al.*, 1978; Ryssel *et al.*, 1981; Swaminathan *et al.*, 1982) for As-doped and B-doped polysilicon as compared to single-crystalline silicon diffusion values. It can be observed that diffusion in polysilicon is at least a few orders of magnitude faster than in single-crystal or bulk diffusion at any temperature. In the particular case of arsenic, the large scatter in the reported data is partially due to the fact that in most of the experiments, the dopants are introduced by implantation, which amorphizes the surface region. During the subsequent heat treatments, the amorphous layer caused by the implantation damage will recrystallize, resulting in a larger grain size in the implanted region (Michel *et al.*, 1982; Arienzo *et al.*, 1984). These morphology

Table 1. Diffusivity (cm²/s) of arsenic and boron in single-crystal silicon and polysilicon

Temp (°C)	As crystal	As poly (a)	As poly (b)	As poly (c)	B crystal	B poly
650	1.2×10^{-21}	1.6×10^{-18}	4.4×10^{-17}	9.3×10^{-18}	1.3×10^{-19}	3.5×10^{-18}
700	1.7×10^{-20}	1.3×10^{-17}	5.4×10^{-16}	5.5×10^{-17}	1.2×10^{-18}	2.9×10^{-17}
750	1.9×10^{-19}	8.6×10^{-17}	5.2×10^{-15}	2.7×10^{-16}	8.1×10^{-18}	1.9×10^{-16}
800	1.6×10^{-18}	4.7×10^{-16}	4.1×10^{-14}	1.1×10^{-15}	4.8×10^{-17}	1.1×10^{-15}
850	1.2×10^{-17}	2.2×10^{-15}	2.7×10^{-13}	4.3×10^{-15}	2.4×10^{-16}	5.2×10^{-15}
900	7.0×10^{-17}	9.2×10^{-15}	1.5×10^{-12}	1.4×10^{-14}	1.1×10^{-15}	2.2×10^{-14}
950	3.7×10^{-16}	3.4×10^{-14}	7.3×10^{-12}	4.3×10^{-14}	4.1×10^{-15}	8.1×10^{-14}
1000	1.7×10^{-15}	1.1×10^{-13}	3.1×10^{-11}	1.2×10^{-13}	1.4×10^{-14}	2.7×10^{-13}
1050	6.8×10^{-15}	3.4×10^{-13}	1.2×10^{-10}	3.1×10^{-13}	4.5×10^{-14}	8.4×10^{-13}
1100	2.5×10^{-14}	9.5×10^{-13}	4.2×10^{-10}	7.4×10^{-13}	1.3×10^{-13}	2.4×10^{-12}

Data and extrapolations from (a) Tsukamoto et al., (b) Swaminathan et al., and (c) Ryssel et al. (1981).

changes, as well as the effect produced by residual lattice defects generated by implantation, complicate the dopant redistribution.

Another interesting effect observed during dopant diffusion in polysilicon is grain growth, particularly observed at high doping levels. It has been found experimentally that both phosphorus and arsenic enhance grain growth, while boron does not (Wada and Nishimatsu, 1978; Mei *et al.*, 1982; Smith *et al.*, 1984; Angelucci *et al.*, 1983). The explanation involves the Fermi-level dependence of dislocation motion and the concentration-enhanced diffusivity of arsenic and phosphorus.

2.3 Dopant Segregation and Activation

Unlike single-crystal silicon, the relationship between impurity doping and electrical conductivity is complex. The dopant incorporated in the polysilicon film will either reside interstitially or substitutionally in the crystallites where it will be mostly ionized (active) or segregate in the grain boundary where it will be only partially ionized or not ionized at all (inactive). Since impurity atoms move preferentially in grain boundaries, a substantial amount of impurities are always present in the grain boundaries of polysilicon. The fraction of the dopant atoms segregating to the grain boundaries is proportional to the volume density of grain-boundary segregation sites and, hence, inversely proportional to the size of the grains (Mandurah *et al.*, 1980). The ratio between the electrical concentration and the dopant concentration is called the activation ratio. Generally, the electrical carrier concentration is always lower than the dopant concentration because of the grain-boundary segregation and carrier trapping at states present in the grain boundaries (Solmi *et al.*, 1982). At high dopant concentrations, the trapping sites, generally of the order of 3×10^{12} sites/cm^2 (Baccarani *et al.*, 1978), are saturated with a negligible fraction of the carriers so that only dopant segregation can influence the electrical properties of polysilicon. For arsenic and phosphorus a substantial amount of segregation has been observed by STEM (Rose and Gronski, 1982; Wong *et al.*, 1985). Boron, on the other hand, shows no appreciable segregation: this has been explained (Mandurah *et al.*, 1980) with a model based on the heat of segregation that yields a negative heat of segregation for boron, implying that boron would be depleted from a surface or interface, and a positive heat of segregation for arsenic and phosphorus. In the case of boron, the activation is very high. High concentrations of boron have been observed at the interface (surface) on which polysilicon is deposited. The amount of pile-up at the interface is a function of the interface preparation and has been attributed to the presence of interfacial oxides or contaminants (Michel *et al.*, 1982; Garben *et al.*, 1986).

In order to obtain a film of the desired resistivity, the synergistic effects between grain size, grain growth, dopant diffusion, and segregation must be considered during all thermal processes that the film will undergo. First, as discussed in Sec. 1.4, grain size and dopants and/or contaminants are functions of the deposition system and will strongly affect the film structure both as deposited and during subsequent heat treatments. If an ion implantation is performed, amorphization and defects introduced will further modify the material properties. As discussed earlier, during each heat treatment, dopant will diffuse through and segregate to grain boundaries, and for higher temperatures, the grain size may also increase as a function of the dopant type and amount. Finally, after the last heat treatment, some of the dopant will be ionized, depending on its bonding coordination, and some of the ionized carriers will be trapped. Only the carriers that are not trapped will be participating in electrical conduction.

2.4 Electrical Properties—Resistivity

Dopant segregation, grain size, and grain orientation can all influence the electrical properties of the films. The resistivity of undoped and lightly doped polysilicon is on the order of 10^5 to 10^6 Ω cm, which is several orders of magnitude higher than that of single-crystal silicon, as shown in Fig. 5. In this range the resistivity is nearly constant, and changes in dopant concentration will result in only small resistivity variations. At medium doping levels, the resistivity becomes a strong function of dopant concentration, and a slight increase will cause it to drop sharply. At high doping levels, the resistivity

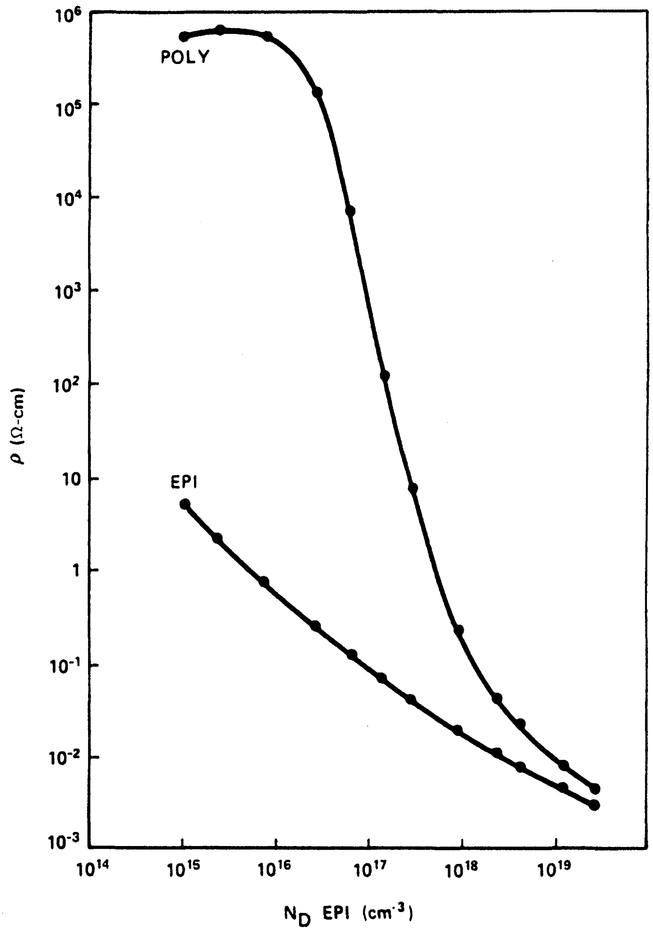

FIG. 5. Resistivity of polysilicon as a function of doping level as compared to that of single-crystal epitaxial silicon (Kamins, 1988).

approaches that of single-crystal silicon, although it will always remain slightly higher (Seto, 1975; Baccarani *et al.*, 1978).

Polysilicon is modeled as an ensemble of small crystallites joined by grain boundaries. Inside each crystallite, the atoms are arranged in a periodic manner forming small single crystals, while the grain boundaries are composed of disordered atoms with incomplete bonding. The high concentration of defects and dangling bonds at the grain boundaries causes trapping states that can immobilize both dopant atoms and charge carriers. Thus, the number of free carriers available for electrical conduction in polysilicon films is always lower than the dopant concentration because of both carrier trapping (Kamins, 1971; Seto, 1975; Baccarani *et al.*, 1978) and dopant-atom segregation at the grain boundaries (Swaminathan *et al.*, 1980; Mandurah *et al.*, 1980).

At low doping levels, the carrier concentration in polysilicon is very close to its intrinsic value in single-crystal silicon. As the doping is increased, the carrier concentration increases slowly and remains very small compared to the average dopant concentration. This is explained by the trapping of most of the carriers at states at or near the grain boundaries, leaving few free to contribute to the conduction; see Fig. 6 (Mandurah *et al.*, 1979). As the doping concentration is increased, the number of trapped carriers will increase in a manner determined by the energy distribution of the traps. Eventually, the traps will approach saturation at a generally accepted concentration in the low 10^{17}-cm^{-3} range. Upon further increase in the doping concentration, the number of trapped carriers will not increase appreciably; however, the potential barrier will decrease and the space-charge region will nar-

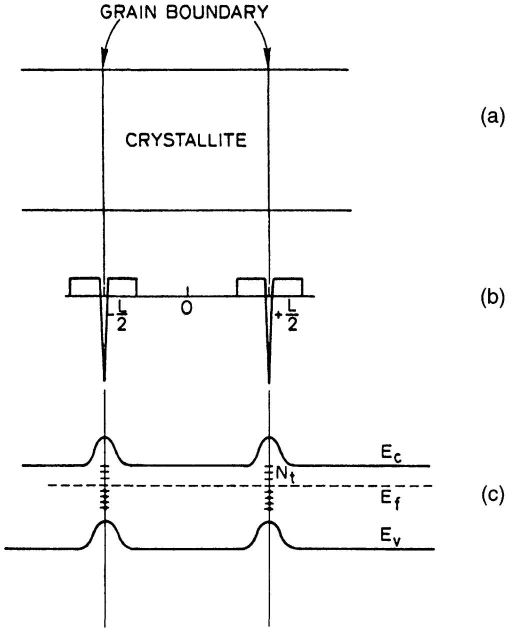

FIG. 6. Model of *n*-type polycrystalline material showing (a) small crystallites surrounded by grain boundaries containing large numbers of traps, (b) resulting space charge, and (c) corresponding energy-band diagram with potential barriers surrounding the grain boundaries (Mandurah *et al.*, 1979).

row. The resistivity will thus decrease sharply with small increases in the dopant concentration as the traps become saturated (Mandurah *et al.*, 1979). Finally, at high dopant concentrations, the region near the grain boundary will no longer limit the conductivity of the samples, and the properties of the material will approach those of the crystallites.

The total dopant concentration is equal to the sum of the active dopant in the crystallites and the dopant segregated at the grain boundaries. Although the dopant atoms in the crystallites have a high probability of being completely ionized, there are several hypotheses on the behavior of the dopant atoms at the grain boundaries. The fraction of the dopant atoms segregated at the grain boundaries is proportional to the volume density of grain-boundary segregation sites and, hence, inversely proportional to the size of the grains. Of the total carriers generated, some ($\sim 10^{17}$) will be trapped at grain-boundary sites, while the remaining n will be finally available for conduction:

$$\sigma = ne\mu, \tag{3}$$

where σ is the conductivity, e is the electron charge, and μ is the mobility (see Sec. 2.5). Figure 7 shows the carrier concentration versus concentration for P- and As-doped films after heat treatment at 1050 °C for 1 h (Solmi *et al.*, 1982). The solid solubilities at the same temperature for P and As in single-crystal silicon are also shown. It can be observed that the electrical carrier concentration is always lower than the dopant concentration, the segregation is more marked for the As-doped films, and as dop-

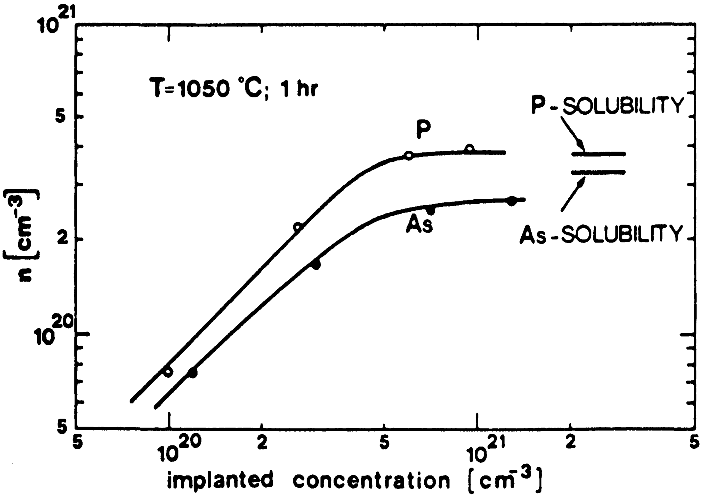

FIG. 7. Carrier concentration vs implanted concentration for P- and As-doped polysilicon films after thermal annealing at 1050 °C for 1 h. Solid solubilities in single-crystal films for P and As are also shown (Solmi *et al.*, 1982).

ing increases, both curves saturate to a value close to the solid solubility at that temperature.

For boron, no measurable changes in carrier concentration were observed, again indicating no appreciable tendency for the boron dopant atoms to segregate at grain boundaries. The boron-doped films are usually activated in the as-deposited form and do not require thermal annealing for activation.

For heavily doped films, the resistivity of the polysilicon is limited by the solid solubility of the dopants in silicon, since we can supersaturate the film with dopant either during deposition or with ion implantation. The resistivity can be reduced only by increasing the mobility, and the dependence of the mobility on the grain size and other polysilicon properties is shown in the next section.

2.5 Electrical Properties—Carrier Mobility

In Eq. (3), the mobility μ describes the ease with which a carrier can move through the polysilicon film. This mobility is influenced by the potential barrier at the grain boundaries and therefore by doping concentration (Seto, 1975). At low doping levels, carrier mobility is on the order of 100 $\mathrm{cm^2/V\ s}$, approximately one order of magnitude lower than in single-crystal silicon, and is relatively insensitive to changes in dopant concentration. At moderate doping levels, it has been observed to drop to a minimum value, and this occurs over the same range of dopant concentrations where the resistivity is most sensitive to changes in dopant (Seto, 1975). At higher levels, the potential barrier at the boundaries decreases and the mobility increases, approaching that for the single crystal.

The mobility in heavily doped polysilicon thin film is affected by ionized-impurity scattering (μ_i), surface scattering (μ_s), and grain boundary scattering (μ_{gb}), as described by (Solmi *et al.*, 1982)

$$\frac{1}{\mu} = \frac{1}{\mu_i} + \frac{1}{\mu_s} + \frac{1}{\mu_{gb}}. \tag{4}$$

The ionized-impurity scattering contribution is the same as found from analysis of carrier mobility versus electron concentration in heavily doped single-crystal silicon. In single crystals, it was verified that the mobility depends only on carrier concentration and is not affected by electrically inactive dopant. Regarding the surface scattering term μ_s, if the thickness of the film d is comparable to the carrier mean free path λ, the mobility μ in the film differs from the bulk mobility μ_b because of surface scattering. The empirical relationship found is

$$\mu = \frac{\mu_b}{1 + 2\lambda/d}. \tag{5}$$

In the heavy doping range, λ is in the 10–15-nm range, and if the film thickness is in the 100–200-nm range, the mobility will then be about 5–10% lower than μ_b.

The behavior of μ_{gb} in the case of phosphorus doping seems to be well characterized empirically by the grain size only and can be described by the relationship

$$\mu_{gb} = KD^{\alpha}, \tag{6}$$

where $K = 4.5 \times 10^4\ \mathrm{cm}^{2-\alpha}/\mathrm{V\ s}$ and $\alpha = 0.59$ are constants and D is the average grain size in centimeters. The direct role of the grain boundary in determining the conductivity in polysilicon films is taken into account in a model where a potential barrier, dependent on the dopant and on processing conditions, is associated with the grain boundary (Mandurah *et al.*, 1981a, 1981b). In the case of arsenic, the mobility in the grain boundary depends strongly on process conditions. The increase of arsenic concentration in the film causes a degradation of the bulk grain crystalline quality (twins and lattice defects) and subsequently a degradation in mobility due to scattering at these defects.

2.6 Diffusion through Polysilicon/Single-Crystal Silicon Interfaces

As discussed earlier, doped polysilicon is often used as a diffusion source for single-crystal silicon during device processing. Through the characterization of the resulting impurity profiles in the single-crystal substrate, it has been found that the interface can profoundly affect the resulting diffusion profiles. More specifically, it has been shown that impurities such as oxygen or carbon that can easily be trapped at the interface during polysilicon deposition can block the diffusion of dopant across the interface. The degree of this effect varies with the dopant and the amount of contaminant. In the case of boron, interface contamination strongly influences the outdiffusion profile in the single crystal. When the interface is free from contaminants, the boron concentration at the interface is equal to the solid solubility at the diffusion temperature. The concentration at the contaminated interface is less than that of the solid solubility limit at the annealing temperature. As the diffusion time increases, the interface concentration increases, gradually approaching the solid solubility limit.

Chemically grown oxides can also substantially retard the diffusion of dopants and have been purposely used to obtain very shallow junctions in silicon (Stork *et al.*, 1985). Studying the oxygen Auger peaks associated with the interface, it has been observed that arsenic diffusion decreases the amount of oxide at the interface. This decrease has been found to be inversely proportional to the initial thickness of the interfacial oxide and the drive-in temperature. These results demonstrate the importance of interface control in achieving reproducible junction profiles.

3. POLYCRYSTALLINE SILICON APPLICATIONS

Polycrystalline silicon has been a key enabling technology in many diverse VLSI applications. It is used extensively in CMOS circuits as a gate contact and for local wiring. Its thermal and chemical compatibility with single-crystal silicon as well as the capability for a wide range of conductivity have made possible the invention of trench (vertical) capacitors and isolation structures. These same properties are utilized to significant advantage in bipolar circuits, increasing switching speeds and decreasing transistor size. Finally, polycrystalline silicon is also used as resistors and resistor contacts.

3.1 Silicon-Gate Technology

Metal-oxide-semiconductor transistors have been used extensively since the 1960s and form the prime building blocks of the well-known *n*-type MOS (NMOS) and CMOS circuits used in memory technology. Through the 1970s, gate electrodes of MOS devices were formed using evaporated aluminum, which was deposited after the source and drain were formed. As circuit densities increased to 1 Mb (1 million transistors/chip), the tight lithographic tolerance required in forming the aluminum gate while ensuring sufficient overlap of source and drain became problematic. This tolerance also drove the need for a large overlap, increasing the parasitic gate-to-drain capacitance, known as the Miller feedback capacitance.

The use of polysilicon as an aluminum-gate replacement is a significant breakthrough. By using polysilicon deposited on the gate electrode as a mask to define both source and drain, a process known as self-alignment, the need for a large overlap of the gate electrode or tight lithographic tolerance is eliminated. Not only does gate polysilicon self-alignment provide miniaturization leverage, it also improves circuit performance by decreasing parametric variability from transistor to transistor. Finally, the precise orientation of gate to drain achieved across all transistors in a circuit by self-alignment considerably reduces and equalizes the Miller feedback capacitance.

Polysilicon gates also provide more suitable threshold voltages for integrated MOS circuits than do aluminum gates. The threshold voltage is the voltage at which a MOS transistor begins to form a conducting channel between source and drain. This voltage is in part determined by the difference in work function between the gate-electrode material and the silicon. While the work function of aluminum is fixed at about 4.1 eV, that of polysilicon can be tailored to the circuit requirement on the basis of doping level and type. For highest electrode conductivity, the polysilicon is normally highly doped by ion implantation. The electron affinity for highly doped *n*-type material, where the Fermi level lies near the conduction band, is close to the work function of aluminum. Therefore, threshold voltages of *n*-gate transistors are similar to those of aluminum-gate transistors. With high *p*-type doping, the Fermi level lies close to the valence band, increasing the work function by nearly 1.1 eV. This makes the threshold voltage of *p*-gate transistors less negative than that of aluminum-gate transistors. The lower threshold voltage is particularly advantageous in CMOS circuits.

3.2 Polysilicon-Emitter Bipolar Transistors

Polysilicon has also provided a number of significant advantages in bipolar circuits. Specifically, its use as an emitter contact has enhanced circuit performance by reducing hole injection from the base even in ultrathin submicron junctions. As a base contact, it has provided a simple means, through the process of self-alignment, to make denser circuits by creating a vertical *npn* bipolar transistor. The process of self-alignment has also enabled miniaturization, eliminating certain critical alignment tolerances.

The most significant early industrial application of polysilicon in bipolar circuits came as an extrinsic base contact in about 1982. The extrinsic base is the conductive region between the lightly doped intrinsic base and the external contact. For high-performance applications, the resistance of the extrinsic base and the capacitance of the doped region must be minimized. The extrinsic base must be separated from the *n*-emitter region to avoid low emitter–base breakdown voltage. In a conventional transistor, this spacing must be at least an alignment tolerance, and this lithographic requirement limited miniaturization.

By using doped polysilicon, the extrinsic base region can be removed from the silicon and placed above an oxide. The polysilicon is highly boron doped and, therefore, very conductive without adding to the capacitance. The intrinsic base contact is achieved by boron out-diffusion from the polysilicon to the single crystal. Furthermore, the emitter can be made to self-align to the extrinsic base so that the two are separated by an oxide sidewall formed on the polysilicon. By using the self-alignment technique, the emitter–base separation can be reduced to far less than an alignment tolerance. Typical sidewall thickness of 250 nm can be formed routinely in production. Therefore, the use of doped polysilicon for the extrinsic base has enabled the reduction of the base resistance and capacitance.

The significant effect of these features on bipolar transistor size is clearly demonstrated in Fig. 8, in which the layouts of two generations of bipolar transistors are compared. The polycrystalline-silicon extrinsic base, coupled with the self-alignment process and advances in photolithography, have enabled the transistor size to be reduced nearly ten times. The speed of polysilicon-base bipolar digital circuits in production has increased from about 4 GHz for conventional transistors to 9 GHz.

A further technical advance was made by using *n*-type polysilicon to contact and form the emitter region. This simplifies the emitter process in two main areas. First, it estab-

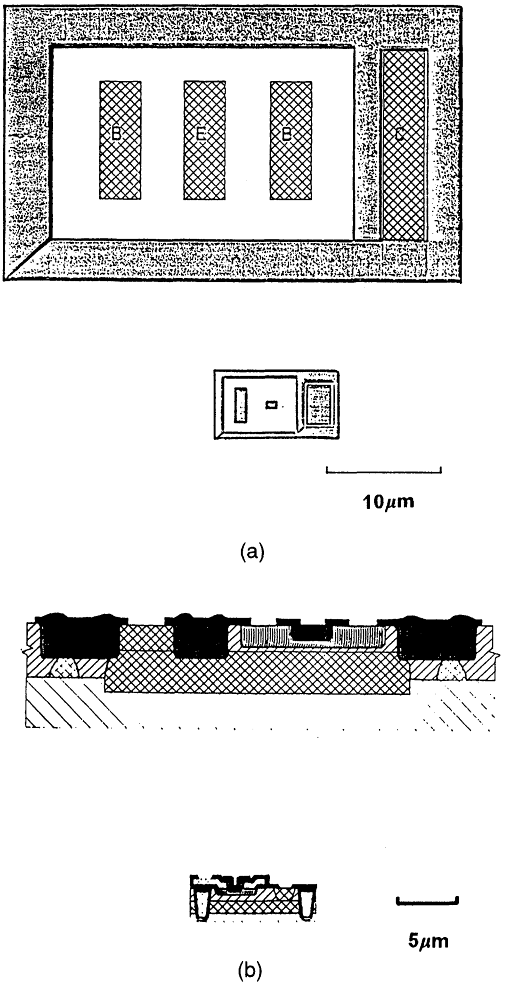

FIG. 8. Layout and cross-section comparison of two generations of IBM bipolar transistors. The transistor (a) is formed with conventional implantation/diffusion, while the transistor (b) is formed using polysilicon.

lishes a low-resistance contact to the single-crystal emitter region without the stringent requirements for metal alloying while avoiding junction penetration. Second, it serves as a source for arsenic dopant for the single crystal, mitigating the requirement for ion implantation. By using polysilicon as an arsenic diffusion source, emitter depths of less than 50 nm can be realized in production without the concerns of implantation damage or contamination. A typical junction profile showing the emitter polysilicon contact, single-crystal emitter, and boron-doped intrinsic base is shown in Fig. 9. With these advances, the circuit speed of digital bipolar circuits in production has exceeded 19 GHz.

3.3 Other Polysilicon Devices and Applications

3.3.1 Trench Isolation Polysilicon has also provided key leverage in integrated-circuit miniaturization through its use in vertical isolation. Conventional isolation structures formed from thermal oxidation consume considerable area and normally require high-temperature processing. A typical polysilicon-filled trench used for isolation is shown in Fig. 10, and the dimensional advantages of such a structure are clear. Two key technological breakthroughs have enabled the vertical isolation structure to be used successfully in manufacturing. First, reactive ion etching has made possible the formation of deep, vertical grooves in the silicon substrate. Many techniques have been attempted to fill these grooves with suitable dielectrics. Materials such as oxides and polyimide have been employed as fillers, with limited success due to the mismatch in the coefficient of thermal expansion and, in the polyimide case, to the lack of chemical stability. The advent of CVD polysilicon was the second breakthrough in achieving vertical isolation: the deposition conditions are controlled so as to grow uniformly along the walls of the groove. The groove will be filled when a polysilicon layer half as thick as the trench is wide has been deposited. During the trench fill process, polysilicon is also deposited on the surface of the silicon wafer. This layer can be removed using dry etching or by chemical mechanical planarization.

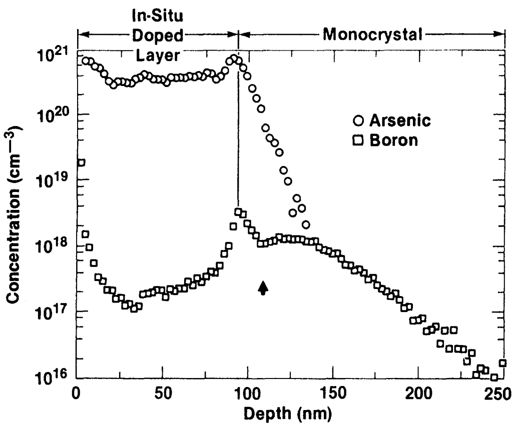

FIG. 9. Typical junction profile of a bipolar transistor formed using *in situ* doped polysilicon.

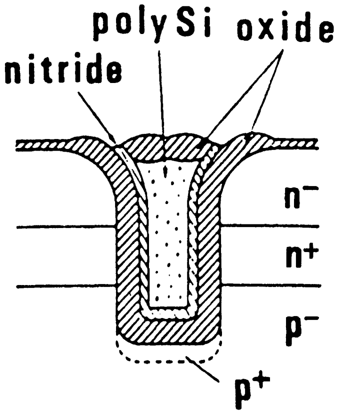

FIG. 10. Schematic of a trench capacitor.

3.3.2 Trench Capacitors The vertical polysilicon structure of Fig. 10 has been further modified and improved to become a capacitor, providing considerable area reduction in specific MOS circuit applications, especially dynamic random-access memories (DRAMs). In DRAM cells, this new feature permitted the planar storage capacitor to be replaced with the trench capacitor, shrinking the cell size while maintaining the required capacitor value. To form the capacitor, a groove is etched using reactive ion etching into the silicon. The groove is oxidized and additional dielectric deposited as required. Doped polysilicon is deposited to fill the trench forming the counter electrode of the capacitor. The excess polysilicon is removed, and finally contact is made to the trench electrode through a conductive layer of polysilicon deposited later in the process. Charge is stored in the inversion layer formed along the vertical walls of the trench, thus increasing the capacitor's effective storage area without increasing the surface area.

3.3.3 Resistors and Interconnections Given the widespread use of polysilicon in circuit applications and its wide range of resistivity, it is frequently used for local interconnections and for high-value resistors. For example, using polysilicon as a local interconnect may eliminate additional masking layers and metal deposition processing, depending on the specific process sequence. However, since the resistance of polysilicon is higher than that of aluminum, it is generally used only over short distances to avoid signal delay. Polysilicon can be used as a high-value resistor when tight electrical tolerances are not required. A typical application is the load resistor in a static RAM cell where resistors are in the range of 10^9 Ω. Polysilicon can be doped using ion implantation and defined lithographically to achieve the required resistor value. The contact areas require additional dopant to ensure good Ohmic contact. The resistivity is difficult to control with great accuracy since as the film's grains grow during subsequent thermal processing, the resistivity of the polysilicon decreases. The resistivity of lightly doped polysilicon varies with temperature, which also degrades resistor operating tolerances.

4. SUMMARY

In this article the salient features of polycrystalline silicon have been discussed ranging from deposition methods to properties and finally applications. Polycrystalline silicon has been used extensively in the VLSI industry because of its ease of deposition and great range of electrical and physical properties. Polycrystalline silicon has also been extensively studied to understand better the influence of crystalline defects and impurities on semiconductor electrical properties.

GLOSSARY

Bipolar Transistor: Junction transistor that provides current gain due to the proximity of forward- and reverse-biased junctions. These two junctions are formed by doping three regions: the *emitter*, the *base*, and the *collector*.

Carrier Trapping: Capture of a free electron or hole by a dangling bond (or recombination center).

CMOS: Complementary metal-oxide-semiconductor circuit family, utilizing both *p*-type and *n*-type MOS transistors.

CVD: Chemical vapor deposition.

Dopant: A chemical impurity that is introduced in a semiconductor material to alter that material's electrical conductivity.

Fermi Level: Chemical potential of electrons in a solid.

Grain Boundaries: The disordered regions between adjacent crystallites in a polycrystalline solid material.

Interferometric Technique: Optical measurement method that relies on the interference of light at film interfaces to determine thickness.

Ion Implantation: Technique to introduce dopants, in which atoms are separated by mass and are highly accelerated to penetrate the semiconductor.

Ionization: Loss of an electron by an atom.

LPCVD: Low-pressure chemical vapor deposition.

MBE: Molecular-beam epitaxy.

Mobility: Ease by which an electrical carrier may move through a material.

MOS: Metal-oxide-semiconductor transistor utilizing a conductor dielectric gate.

Photolithography: Technique by which circuit images are printed on semiconductor wafers.

Polycrystalline: A solid composed of many crystallites of different crystal orientation.

PVD: Physical vapor deposition.

RBS: Rutherford back-scattering.

SIMS: Secondary-ion mass spectroscopy.

Solar Cell: Semiconductor device in which solar energy is converted into electrical energy.

STEM: Scanning transmission electron microscopy.

TEM: Transmission electron microscopy.

VLSI: Very large-scale integration.

Works Cited

Adams, A. C. (1983), in: S. M. Sze (Ed.), *Dielectric and Polysilicon Film Deposition*, New York: McGraw-Hill, pp. 93–129.

Angelucci, R., Severi, M., Solmi, S. (1983), *Mater. Chem. Phys.* **9,** 235–245.

Arienzo, M., Komem, Y., Michel, A. E. (1984), *J. Appl. Phys.* **55,** 365–369.

Arienzo, M., Megdanis, A. C., Sackles, P. E., Michel, A. E. (1986), *IEEE Trans. Electron Dev.* **ED-33,** 1535–1538.

Baccarani, G., Ricco, B., Spadini, G. (1978), *J. Appl. Phys.* **49,** 5565–5570.

Bravman, J. C., Patton, G. L., Plummer, J. D. (1985), *J. Appl. Phys.* **57,** 2779–2782.

Garben, B., Orr-Arienzo, W., Lever, R. (1986), *J. Electrochem. Soc.* **133,** 2152–2156.

Goto, H., Inayoshi, K. (1987), in: G. Soncini, P. U. Calzolari (Eds.), *ESSDERC '87: 17th European Solid State Device Research Conference,* Bologna: Technoprint, p. 369.

Hammond, M. L. (1980), *Solid State Technol.* **23,** 104–109.

Kamins, T. I. (1971), *J. Appl. Phys.* **42,** 4357–4366.

Kamins, T. I. (1979), *J. Electrochem. Soc.* **126,** 833–837.

Kamins, T. (1988), *Polycrystalline Silicon for Integrated Circuit Applications,* Boston: Kluwer Academic Publishers.

Kern, W., Schnable, G. L. (1979), *IEEE Trans. Electron Devices* **ED-26,** 647–657.

Mandurah, M. M., Saraswat, K. C., Helms, C. R., Kamins, T. I. (1980), *J. Appl. Phys.* **51,** 5755–5763.

Mandurah, M. M., Saraswat, K. C., Kamins, T. I. (1979), *J. Electrochem. Soc.* **126,** 1019–1023.

Mandurah, M. M., Saraswat, K. C., Kamins, T. I. (1981a), *IEEE Trans. Electron Devices* **ED-28,** 1163–1171.

Mandurah, M. M., Saraswat, K. C., Kamins, T. I. (1981b), *IEEE Trans. Electron Devices* **ED-28,** 1171–1176.

Marcus, R. B. (1983), in: S. M. Sze (Ed.), *Diagnostic Techniques,* New York: McGraw-Hill, 524–527.

Mei, L., Rivier, M., Kwart, Y., Dutton, R. W. (1982), *J. Electrochem. Soc.* **129,** 1791–1795.

Meyerson, B. S., Yu, M. L. (1984), *J. Electrochem. Soc.* **131,** 2366–2371.

Michel, A. E., Kastl, R. H., Mader, S. R. (1982), *Nucl. Instrum. Methods Phys. Res.* **209,** 719–724.

Natsuaki, N., Tamura, M., Miyazak, T., Yanagi, Y. (1983), in: *IEDM Technical Digest: International Electron Devices Meeting,* New York: IEEE, pp. 662–665.

Rose, J. H., Gronski, R. (1982), *Appl. Phys. Lett.* **41,** 993–995.

Rosler, R. S. (1977), *Solid State Technol.* **20,** 63–70.

Ryssel, H., Iberl, H., Bleier, M., Prinke, G., Haberger, K., Kranz, H. (1981), *Appl. Phys.* **24,** 197–200.

Seto, J. Y. W. (1975), *J. Appl. Phys.* **46,** 5247–5254.

Smith, D. A., Tan, T. Y., Fontaine, C. (1984), in: *IBM Technical Report RC9361.*

Solmi, S., Severi, M., Angelucci, R., Baldi, L., Bilenchi, R. (1982), *J. Electrochem. Soc.* **129,** 1811–1818.

Stork, J. M. C., Arienzo, M., Wong, C. Y. (1985), *IEEE Trans. Electron Devices* **ED-32,** 1766–1770.

Swaminathan, B., Demoulin, E., Sigmon, T. W., Dutton, R. W., Reif, R. (1980), *J. Electrochem. Soc.* **127,** 2227–2229.

Swaminathan, B., Saraswat, K. C., Dutton, R. W., Kamins, T. I. (1982), *Appl. Phys. Lett.* **40,** 795–798.

Thomas, G., Goringe, M. J. (1979), *Transmission Electron Microscopy of Materials,* New York: Wiley.

Tsukamoto, K., Akasaka, Y., Watari, Y., Kusano, Y., Hirose, Y., Nakamura, G. (1978), *Jpn. J. Appl. Phys. (Suppl.)* **17,** 187–192.

Wada, Y., Nishimatsu, S. (1978), *J. Electrochem. Soc.* **125,** 1499–1504.

White, F., Koburger, C., Abernathey, J., Nesbit, L., Arienzo, M. (1985), in: *IBM Technical Report RC11953.*

Wong, C. Y., Grosvenor, C. R. M., Batson, P. E., Isaac, R. D. (1985), in: *IBM Technical Report RC10802.*

Further Reading

Cullen, G. W. (Ed.) (1989), *Proceedings of the Tenth International Conference on Chemical Vapor Deposition 1987,* Pennington, NJ: The Electrochemical Society, Inc.

Ghandi, S. K., *VLSI Fabrication Principles,* New York: Wiley.

Grove, A. S., *Physics and Technology of Semiconductor Devices,* New York: Wiley.

Kamins, T., *Polycrystalline Silicon for Integrated Circuit Applications,* Boston: Kluwer Academic Publishers.

Robinson, M. D. (Ed.) (1984), *Proceedings of the Ninth International Conference on Chemical Vapor Deposition 1984,* Pennington, NJ: The Electrochemical Society, Inc.

Sze, S. M., *VLSI Technology,* New York: McGraw-Hill.

SILVER, GOLD, AND OTHER NOBLE METALS

MONIKA FIEBERG, *Chimet SpA, Badia al Pino (Arezzo), Italy,*
HERMANN RENNER, *Degussa A.G. (retired), Hanau, Germany,* and
GUENTHER SCHLAMP, *Demetron G.m.b.H. (retired), Hanau, Germany*

Introduction 236
1. **General Characteristics of the Noble Metals** 236
1.1 Electronic Structure 236
1.2 Physical Properties 237
1.3 Alloying Behavior 239
1.3.1 Phase Formation 239
1.3.2 Strengthening 240
1.3.3 Oxygen 241
1.3.4 Hydrogen 241
2. **Silver** 241
2.1 Electronic Structure 241
2.2 Special Properties 241
2.2.1 Mechanical and Thermal Properties 241
2.2.2 Electrical Conductivity 242
2.2.3 Magnetic Properties 242
2.2.4 Optical Properties 242
2.2.5 Chemical Properties 242
2.3 Alloying Behavior 243
2.3.1 Phase Formation 243
2.3.2 Influence of Alloying Elements 244
2.3.2.1 Strengthening 244
2.3.2.2 Grain Refining 244
2.3.2.3 Electrical Conductivity 244
2.3.2.4 Oxygen 244
2.4 Abundance, Resources, and Raw Materials 245
2.4.1 Abundance and Resources 245
2.4.2 Raw Materials 245
2.5 Production 245
2.5.1 Primary Production 245
2.5.2 Secondary Production (Recycling) 246
2.6 Manufacturing 246
2.7 Alloys 247
2.7.1 Special Compositions 247
2.7.2 Solders and Brazes 248
2.8 Applications 249
2.8.1 Photography 249
2.8.2 Jewelry and Silverware 249
2.8.3 Dentistry and Biomedical Uses 249
2.8.4 Electrical Technology 249
2.8.5 Electronics 250
2.8.6 Heat Mirrors 250
2.8.7 Chemical Technology 251
2.8.8 Catalysis 251
2.8.9 Batteries 251
3. **Gold** 252
3.1 Electronic Structure 252
3.2 Special Properties 252
3.2.1 Mechanical and Thermal Properties 252
3.2.2 Electrical Conductivity 252
3.2.3 Magnetic Properties 252
3.2.4 Optical Properties 253
3.2.5 Chemical Properties 253
3.3 Alloying Behavior 253
3.3.1 Phase Formation 253
3.3.2 Influence of Alloying Elements 253
3.3.2.1 Strengthening 253
3.3.2.2 Grain Refining 254
3.3.2.3 Electrical Resistivity 254
3.4 Abundance, Resources, and Raw Materials 254
3.4.1 Abundance and Resources 254
3.4.2 Raw Materials 255
3.5 Production 255
3.5.1 Primary Production 255
3.5.2 Secondary Production (Recycling) 255
3.6 Manufacturing 255
3.7 Alloys 255
3.7.1 Special Compositions 255
3.7.2 Solders and Brazes 256
3.8 Applications 256
3.8.1 Jewelry 256
3.8.2 Dentistry 257
3.8.3 Electrical Technology 257
3.8.4 Electronics 257
4. **Platinum and the Platinum-Group Metals** 257

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

4.1 Common Characteristics 257
4.1.1 General Description 257
4.1.2 Electronic Structure 257
4.1.3 Physical Properties 258
4.1.3.1 Mechanical and Thermal Properties 258
4.1.3.2 Electrical Resistivities 259
4.1.3.3 Thermoelectric Behavior 259
4.1.3.4 Magnetic Behavior 260
4.1.4 Alloying Behavior 261
4.1.4.1 Metals 261
4.1.4.2 Carbon 261
4.1.4.3 Oxygen 262
4.1.5 Catalytic Actions 262
4.1.6 Abundance, Resources, and Raw Materials 264
4.1.6.1 Abundance and Resources 264
4.1.6.2 Raw Materials 265
4.1.7 Production 265
4.1.7.1 Primary Production 265
4.1.7.2 Secondary Production (Recycling) 265
4.2 The Elements 266
4.2.1 Platinum 266
4.2.1.1 Influence of Alloying Elements 266
4.2.1.2 Alloys 268
4.2.1.3 Applications 269
4.2.2 Palladium 272
4.2.2.1 Influence of Alloying Elements 272
4.2.2.2 Alloys 273
4.2.2.3 Applications 273
4.2.3 Rhodium 274
4.2.3.1 Applications 274
4.2.4 Iridium 274
4.2.4.1 Applications 274
4.2.5 Ruthenium 275
4.2.5.1 Applications 275
4.2.6 Osmium 275
4.2.6.1 Applications 275
5. Summary 275
Acronyms 276
Works Cited 276
Further Reading 278

INTRODUCTION

The term "noble metals" encompasses the eight elements silver, gold, and the six platinum-group metals (PGM) ruthenium, rhodium, palladium, osmium, iridium, and platinum. According to their electronic structures, the precious metals are classified as the end members of the transition metals placed in the periodic table of the elements in groups VIII and IB of the fifth and sixth periods. They are characterized by dense close-packed crystal structures, outstanding resistance against oxidation, high melting temperatures, good to high electrical conductivities, and mechanical and thermal stability. Silver has the highest electrical conductivity of all elements, gold has the highest oxidation resistance of all metals, and osmium and iridium are the elements with the highest densities. According to their individual electronic structures, they show, unique to each, distinctive physical and chemical properties.

Silver and gold are the classical precious metals, used since ancient times for jewelry, tableware, and coins and in dentistry. Besides these still-continuing applications, they are in elemental and in alloyed form of high technical importance as conductor, connector, solder, and braze materials in the fields of electrical engineering and electronics, batteries, and optical and heat-reflecting coatings on architectural glasses. Silver is the most important constituent in photographic emulsions. The main application fields of the platinum metals are catalysis in organic and anorganic chemical reactions, automotive catalysts, electrodes of galvanic and of fuel cells, highly coercive magnetic alloys for recording and memory devices, resistor elements in electronic devices and in the sensor technology, construction materials of corrosion-resistant chemical equipment, and, as chemical compounds, constituents in medical drugs.

1. GENERAL CHARACTERISTICS OF THE NOBLE METALS

1.1 Electronic Structure

The electronic structure of the noble-metal atoms is characterized by the completion of their $4d$ or $5d$ electron levels with increasing nuclear charge, extending for the "light" noble metals (atomic numbers 44–47) from $4d^7 5s^1$ for ruthenium to $4d^{10} 5s^1$ for silver and for the "heavy" noble metals (atomic numbers 76–79) from $5d^6s^2$ for osmium to $5d^{10}6s^1$ for gold. In the condensed metallic state, the electron orbitals overlap, forming energy bands. The electronic structure is described by the density of states (DOS = number of energy levels in a unit energy range) and the Fermi level, which represents the state occupied with maximum electron energy. The state-density curve is almost the same for all noble metals (Yonemitsu, 1991). Figure 1 shows a schematic DOS curve of the noble metals with the positions of their Fermi levels. The different locations of the Fermi levels of each pair of the group homologs relative to the peaks or valleys of the DOS curve account for the differences in their electric conductivities and in their thermoelectric, magnetic, and optical behavior. Because of the very small energy differences between their outer *s*- and *d*-electron levels, *d* electrons can be easily transferred into the valence band. This enables all noble metals to react in several oxidation states, forming numerous anorganic and organic compounds and—with the exception of gold—to perform manifold catalytic activities by intermediate participation in chemical reactions. Structure-related similarities exist between the group homolog elements of the fourth period (iron, cobalt, nickel, copper) in the alloying behavior, in the formation of magnetic alloys, in the formation of complex anorganic and organic compounds, and in catalytic activities.

The strong tendency to complete the 4*d* and 5*d* levels has the effect that for ruthenium, rhodium, silver, platinum, and gold the outer *s* levels are only singly occupied. Palladium has the basic electron configuration $d^{10}s^0$.

The "noble" character of the precious metals is expressed by their positive reduction potential against hydrogen. They share this characteristic with a few other elements such as antimony, copper, mercury, and rhenium. The reduction potential increases along the periods and inside the groups with increasing nuclear charges.

1.2 Physical Properties

The atomic diameters of the transition metals of the second long period are systematically larger than those of their group homologs of the first long period. The differences between the atomic diameters of the homologs of the second and the third long periods are, as a result of the lanthanide contraction, very small. The elements of each of the pairs ruthenium/osmium, rhodium/

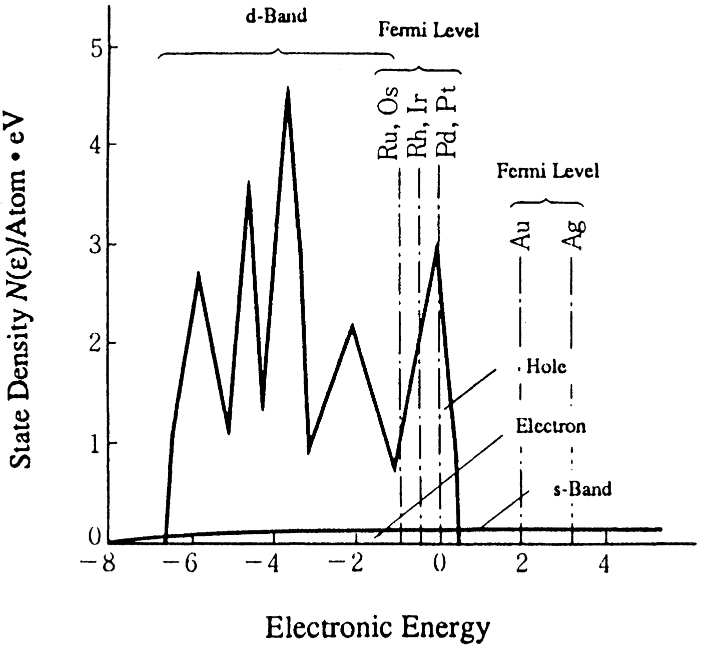

FIG. 1. Schematic DOS curve of the fcc transition metals with location of the Fermi levels of the noble metals relative to the upper state of density (Yonemitsu, 1991, p. 30).

iridium, palladium/platinum, and silver/gold have practically the same atomic diameters.

The crystal structures of the transition metals change with increasing atomic numbers from the more open body-centered-cubic structures to the close-packed structures of highest symmetry and highest coordination numbers. Ruthenium and osmium crystallize hexagonal close packed; rhodium, palladium, silver, iridium, platinum, and gold have face-centered-cubic crystal structure. There are octahedral and tetrahedral holes in the fcc and the hcp structures, which can be occupied by interstitial atoms having maximum atomic radii of $0.41r$ or $0.225r$, respectively (r = atomic radius of the noble metals, if these are considered as close-packed spheres in these lattices).

The densities of all noble metals are above 10 g/cm^3, and those of the the heavy noble metals (osmium, iridium, platinum, gold), above 19 g/cm^3.

The ductility increases in each period with increasing atomic number. Silver, gold, palladium, and platinum exhibit a wide range of ductility; rhodium, iridium, ruthenium, and osmium can only be deformed at elevated temperatures, accompanied by strong work hardening.

The elastic moduli of the single crystals of the fcc noble metals show considerably different values in the ⟨100⟩, ⟨110⟩, and ⟨111⟩ direction, causing anisotropic elastic properties of single crystals and in polycrystalline materials with textures (Table 1).

The high melting temperatures, between 960 °C (silver) and 3045 °C (osmium), are in accordance with the dense packed structures and indicate strong interatomic bonds. The melting points and the recrystallization temperatures decrease in each period with increasing nuclear charge. Likewise, the hardness and the brittleness decrease, while the thermal expansion increases with increasing nuclear charge.

Table 1. *E* moduli in different crystal directions (Jönsson, 1995, p. 204).

Noble metal	Direction	*E* (GPa)
Ag	[100]	44
	[110]	82
	[111]	115
Au	[100]	42
	[110]	81
	[111]	114
Pd	[100]	65
	[110]	129
	[111]	187
Ir	[110]	480
	[111]	660

The individual electrical conductivities differ considerably. Silver has the lowest electrical resistivity of all metals. Palladium and platinum have values as large as six times those of silver, and rhodium, iridium, ruthenium, and osmium have intermediate values (Table 2). The higher resistivities compared with those of silver and gold are mainly due to the fact that these elements have smaller numbers of carriers and that the mobility of the current-carrying *s* electrons is decreased by an increased probability of scattering due to phonons, caused by the higher densities of *d* states at the Fermi level. The carrier density for silver and gold is 1/atom, while for palladium and platinum, because of *d-s* interband transitions, it is reduced to 0.6/atom. The total state densities $N(e_F)$ at the Fermi levels are proportional to the electron specific heat coefficients γ:

$$N(e_{\mathrm{F}}) \alpha \gamma \quad (\mathrm{mJ/mol\ K^2}). \tag{1}$$

The difference of the state densities can be derived from the measured values of γ, which amount for silver and gold to 0.65 and 0.7, respectively, while for palladium and platinum γ is as high as 9.9 and 6.7 (Tanuma, 1991, p. 46). The rule of Wiedemann and Franz of the proportionality of the electrical conductivity and the thermal conductivity is valid for all precious metals. The Lorenz numbers, calculated from the ratios of the measured values of the thermal and their electrical conductivity,

Table 2. Electrical resistivities of the precious metals (Diehl, 1995, p. 155).

Metal	Resistivity	Measurement error ±
Ru	6.7	—
Rh	4.34	0.01
Pd	9.725	0.015
Ag	1.465	0.005
Os	8.2	0.1
Ir	4.7	0.05
Pt	9.825	0.025
Au	2.03	0.01

Table 3. Superconducting precious-metal alloys (Benner *et al.,* 1991, p. 636).

Alloy	Transition temperature (K)
Nb_3Au	10.9
Nb_3Pt	10.6
ZrRuP	12.3
$Tb_3Os_4Ge_{13}$	14.1
ZRuAs	11.9
YOs_4Ge_{10}	8.6
$ScIr_4Si_{10}$	8.4
$ErRhB_4$	8.5

$$L = \lambda\kappa/T \quad (\mathrm{W}\,\Omega/\mathrm{deg}^2), \tag{2}$$

correspond with values between 2.27 for silver and 2.96 for ruthenium to typical metallic behavior in accordance with the theory of the contribution of the free electrons to the thermal conductivity.

Superconductivity has been found for ruthenium (T_c = 0.49 K), rhodium (T_c = 0.9 K), osmium (T_c = 0.71 K), and iridium (T_c = 0.14 K). Higher transition temperatures are shown by some binary and ternary noble-metal alloys (Table 3) (Khan, 1984; Raub, 1984; Benner *et al.,* 1991, p. 636). The thermoelectric properties of the precious metals are of great practical importance for thermocouples used for measurements at high temperatures. The magnetism of the precious metals is very weak. Silver and gold behave (like copper) diamagnetically, while all platinum-group metals behave paramagnetically.

The optical properties are closely related to the interactions of photons with *s* and *d* electrons. Optical constants give information about electronic band structures, the participation of free and bound electrons, phase formation, ordering processes, and lattice defects. Colors are due to selective light absorption by interband transitions of electrons caused by absorption of photons. Photons can be absorbed by exciting an electron of a filled band just below the Fermi surface. The energy difference between this band and the Fermi surface determines the minimum energy of the photon that can be absorbed by interband transition. For the monovalent metals copper, silver, and gold, these energy differences are 2.1, 4.0, and 2.17 eV, respectively, causing the characteristic brilliant colors of the copper red, silver white, and gold yellow (see also COPPER, p. 269). The reflectivity and the color of the noble metals are of special importance for jewelry and dentistry alloys and for optical and heat-reflecting mirrors. Figure 2 shows the reflectivity of the bulk noble metals in the wavelength range of visible light.

Emissivity values are needed in pyrometrical measurements to calibrate deviations against the emission values of a black body. Table 4 gives some selected values.

1.3 Alloying Behavior

1.3.1 Phase Formation The precious metals form numerous alloys with each other and with many subgroup A and B metals. The alloy systems include types of continuous solid solutions, superlattice structures, miscibility gaps, eutectic systems, and intermetallic phases. They have predominantly fcc, hcp, or bcc crystal structures. Pri-

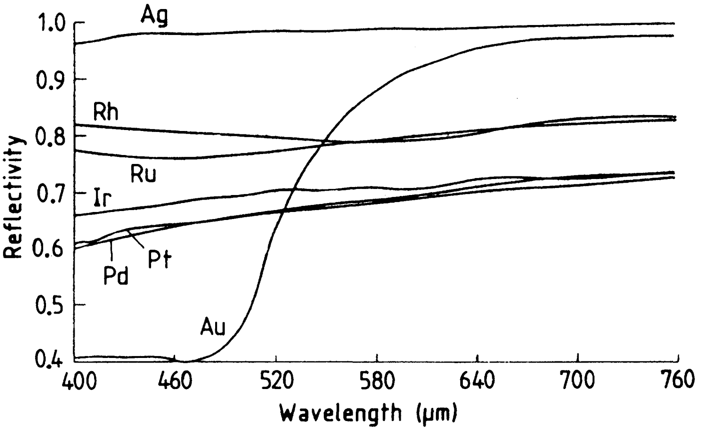

FIG. 2. Optical reflection of the precious metals in the visible light (Völcker, 1995, p. 173).

Table 4. Optical emissivity of precious metals at 1000 °C[a] (Smithells and Brandes, 1977, p. 1019).

Metal	Emissivity
Ag	0.055[b]
Au	0.16
Rh	0.22
Pt	0.29
Ir	0.36
Pd	0.37
Ru	0.42
Os	0.52

[a] At wavelength 0.65 μm.
[b] At 800 °C.

mary solid solutions in precious-metal alloy systems with metals are generally of the substitutional type, where the solute atoms replace some of the solvent atoms on their lattice positions. Continuous solid solutions are formed between most noble metals, and in the alloy systems of gold, palladium, and platinum with iron, cobalt, and nickel. Interstitial-type solid solutions are formed by palladium and platinum with small atoms like hydrogen and carbon.

Some precious-metal alloys undergo ordering reactions forming superlattices. Their generation is bound to differences of the electrochemical properties and to differences of the atomic sizes of the alloying components. The most frequently formed structures are of the composition types AB or AB_3, which derive from the fcc lattices in the disordered state (Table 5). The transformations are accompanied by changes of physical, mechanical, or chemical properties. These order–disorder transformations are of practical importance for the production of mechanically hard alloys and of high-coercivity magnetic alloys.

1.3.2 Strengthening Strengthening of precious metals and alloys is effected either by solid-solution hardening, by phase trans-

Table 5. Compositions and structures of superlattices in precious-metal alloy systems (Hirabayashi, 1991, p. 105).

Atomic ratio	Composition			Superlattice structure	Fundamental structure
3:1 or 1:3	Pd_3Fe	$PdCu_3$	Au_3Cu $AuCu_3$		
	Pt_3Fe	$PtCu_3$	Au_3Pd		
	Pt_3Ti	$PtNi_3$	Ag_3Pt	$L1_2$	fcc
		$PtMn_3$			
	Rh_3Mo		Ag_3In		
	Rh_3W			$D0_{19}$	
	Ir_3Mo				
	Ir_3W		(Au_3Mn)	$D0_{22}$	fcc
	Pd_3V				
	Pd_3Nb				
	Pt_3V				
			Au_3Cd	$D0_{23}$	fcc
			Au_3Zn		
			Ag_3Mg		
2:1	Pd_2V			Ni_2Cr	fcc
	Pt_2V				
	Pt_2Mo				
1:1	PdFe		AuCuI		
	PtFe			$L1_0$	fcc
	PtV				
	PtNi				
	PtCo				
	PtCu				fcc
	PdCu		AuMn	$L1_1$	bcc
			AgCd	$L2_0$ or B2	
			AgZn		
	RhMo				hcp
	IrMo			B19	
	PtMo				
	IrW				
	PtNb				

formations, by precipitation hardening, by dispersion hardening, or by grain-size refinement.

Dispersion hardening is effected by incorporating finely dispersed, small amounts (0.1–2 vol %) of very small particles of high–melting-point compounds (oxides, carbides), which are insoluble or become soluble only near the melting temperature of the matrix. Special properties are available by composite materials made from noble-metal powders and metallic or nonmetallic additives. They are used for electrical contacts.

Fine grain structures of noble metals and noble-metal alloys improve the working behavior and surface quality. Grain refining and the stabilization of grain sizes to prevent grain growth during annealing operations is performed by alloying with finely dispersed insoluble particles, by segregation of insoluble atoms, or by solute atoms with large dimensional misfit to the host metal. They are applied in gold jewelry and dental alloys, and silver and platinum-metal alloys for spinnerets and seal rings.

1.3.3 Oxygen Noble metals, except gold, form solid and gaseous oxides. Dissolved oxygen affects hardness and working behavior. Silver oxides and platinum-metal oxides have practical importance as catalysts in organic chemical reactions.

1.3.4 Hydrogen All precious metals, except palladium, take only very small amounts of hydrogen in solution. Palladium and certain palladium alloys are unique in their capability to dissolve large amounts of hydrogen.

2. SILVER

Silver, chemical symbol Ag (argentum), atomic number 47, relative atomic mass 107.8682, is in its natural abundance composed of two stable isotopes in the composition of 51.84% ^{107}Ag and 48.16% ^{109}Ag. It is a bright, white metal with the characteristic silver luster. In pure condition it is very soft and malleable. The most important properties are the high electrical conductivity, the oxidation resistance, the optical reflectivity, the alloying behavior, and the light sensitivity of silver halides. The main application areas are film/photography (39%), jewelry and silverware (28%), electronics/electrical technology (14%), and solders and brazes (5%) (quantities in weight percent of the total yearly demand in the western world 1994; Degussa, 1994).

2.1 Electronic Structure

The electronic structure is [Kr] $4d^{10}5s^1$. The band diagrams, describing the electron energy levels, show five lower bands, originating from the fully occupied d levels, while one band near the Fermi level corresponds to the s-electron level. Figure 3(a) shows the band diagram of silver, and Figure 3(b), the density of states for the d and s states. Interactions between the $4d$ and the $5s$ states cause the Fermi surface, which is basically spherical for monovalent metals, to intercept the hexagonal faces of the Brillouin zone, forming "necks" in the Fermi surface with ring-shaped intersections around the $\langle 111 \rangle$ direction in **k** space [Fig. 3(c)].

The fully occupied $4d$ level at an energy about 4 eV below the singly occupied s-like states accounts for the preferred monovalent chemical behavior, for the high electrical conductivity, and for the brilliant reflectivity over the total range of the visible light.

2.2 Special Properties

Selected values of basic physical and chemical properties of silver are given in Table 6.

2.2.1 Mechanical and Thermal Properties The fcc crystal structure with its high symmetry provides many slip planes and is the reason for the good mechanical formability of silver. It has low strength and can easily be cold worked. The hardness of pure, annealed silver is ~25 HV; the work hardening is not very marked. It can be rolled down to foil thickness of ~0.2 μm.

The value of the elastic modulus of single crystals of silver differs in the $\langle 100 \rangle$, $\langle 110 \rangle$, and $\langle 111 \rangle$ directions (Table 1), causing anisotropic elastic properties of single crystals and in polycrystalline material with textures. The recrystallization temperature is $0.28T_m$ (T_m = melting temperature in kelvins) (Jönsson, 1995).

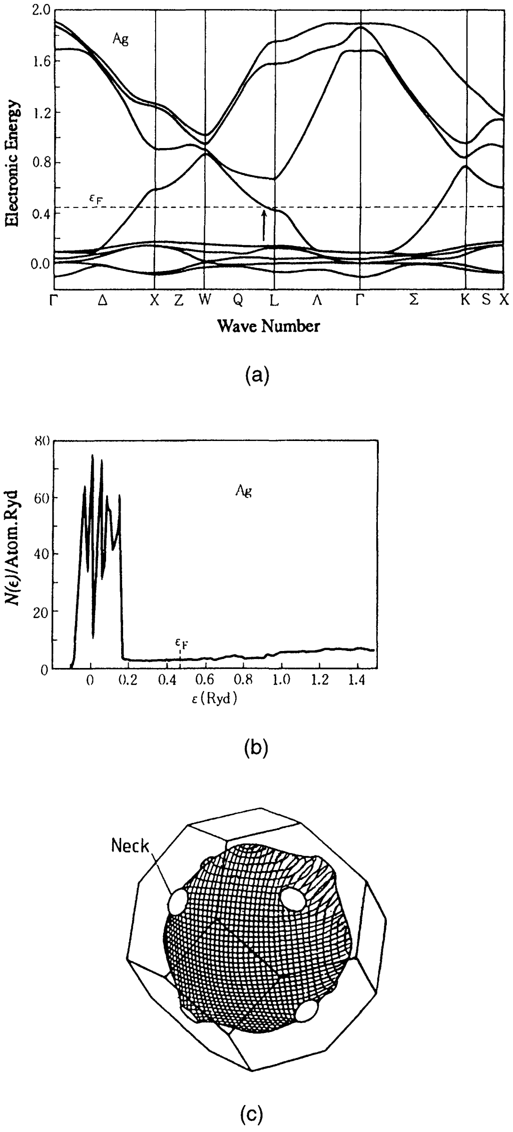

(a)

(b)

(c)

FIG. 3. Electronic structure of silver (Yonemitsu, 1991, pp. 25–28): (a) band diagram, (b) DOS curve (from Christensen, 1972), (c) Fermi surface.

2.2.2 Electrical Conductivity The electrical conductivity $K = en\mu$ is determined by the number of carriers en and their mobility μ, understood as the mean velocity of the carriers at a given field strength and electrical charge. The number of electrons involved in the Fermi surface of silver is 1 electron/atom, which can be converted by calculation with 1 × crystal density × atomic weight × Avogadro number to 5.8×10^{22} cm^{-3}. The high electrical conductivity of silver arises from its free-electron-like monovalent behavior and the small amplitudes of thermal vibrations due to its dense crystal structure and its low compressibility, which lead to only small electron-phonon scattering.

2.2.3 Magnetic Properties Silver behaves diamagnetically. The diamagnetic susceptibility is independent of the temperature between 0 °C and the melting point. Table 7 gives the susceptibilities at zero and 20 °C. Measurement of magnetic moments is used to determine the oxidation state of elements in alloys and compounds.

2.2.4 Optical Properties In the range of the wavelengths of visible light, silver shows the highest optical reflection of all noble metals. Absorption of light is related to interactions with electrons in filled states below the Fermi level. Photons can be absorbed by excitation of electrons only if the photon energy is higher than the energy difference between the lower filled band and the Fermi level. The difference of about 4 eV corresponds to a wavelength of 3000 Å, which is in the ultraviolet range (Hummel, 1971, p. 161). No visible light is therefore absorbed by electron excitation. Silver coatings are therefore best suited for optical mirrors, reflectors, and transparent heat-reflecting architectural glasses. The high reflectivity of silver is in accordance with the Hagen–Rubens relation (Hagen and Rubens, 1903),

$$R_{\mathrm{opt}} = 1 - \sqrt{\nu/\kappa} \tag{3}$$

where ν = frequency of light and κ = electrical conductivity. This states that at small frequencies, metals with high electrical conductivities are also good optical reflectors.

2.2.5 Chemical Properties Silver is resistant against water and shows excellent resistance to cold and hot concentrated organic acids such as acetic, formic, citric, lactic, fumaric, phthalic, and benzoic acids, fatty acids, and phenol. It is soluble in nitric and sulfuric acid. It is to a much higher degree than most other metals resistant to high-temperature caustic alkalies.

Silver is attacked by low-melting-point metals, such as mercury, sodium, potassium, lead, tin, indium, and bismuth. Reactions with sulfur-containing compounds form

Table 6. Selected values of basic physical and chemical properties of silver and gold (Degussa, 1995).

	Ag	Au
Atomic properties		
Atomic number	47	79
Atomic weight	107.868	196.966
Electron configuration	$[\mathrm{Kr}]4d^{10}5s$	$[\mathrm{Xe}]4f^{14}5d^{10}6s$
Crystal structure	fcc	fcc
Lattice constant (Å) (10^{-1} nm)	4.086	4.078
Atomic radius (Å) (10^{-1} nm)	1.445	1.442
Mechanical properties		
Density (20 °C) (g/cm^3)	10.49	19.32
Density, liquid at melting point (g/cm)^3	9.35	17.19
Hardness (HV, kg/mm^2)	26	25
Young's modulus (GPa)	82	79
Modulus of rigidity (MPa)	27	28
Poisson's ratio	0.37	0.42
Coefficient of compressibility	0.95	0.58
Thermal properties		
Melting point (mp) (°C)	961.9	1064.4
Vapor pressure at mp (mbar) (1×10^5 N m^{-2})	1.5	0.02
Boiling point (°C)	2212	3080
Thermal conductivity (20 °C) (W·m^{-1} K^{-1})	418	310
Coeff. of linear expansion (0–100 °C) (10^{-6} K)	19.0	14.3
Electrical properties		
Specific electrical resistance (0 °C) ($\mu\Omega$ cm)	1.465	2.0
Temperature coefficient of electrical resistance (0–100 °C) (10^{-3} K)	4.1	4.03
Thermoelectric voltage against Pt (0–100 °C) (mV)	0.740	0.770
Work function (eV)	4.73	5.22
Chemical properties		
Oxidation states in compounds	2, 1	3, 1
Ionization energy (eV)	7.576	9.225
Covalent radius for single bonds (Pauling) (pm)	134	134
Ionic radius (pm) (oxidation number, coordination number)	79 (2+,4) 67 (1+,2)	68 (3+,4) 137 (1+,6)
Electronegativity (Allred and Rochow)	1.4	1.4
Reduction potential (V) with (*n*) electrons	0.800 (1)	1.498 (3)

black silver-sulfide layers. Very thin surface layers of Ag_2O are formed in oxygen or in air below 180 °C. Above this temperature the oxygen partial pressure of the Ag_2O exceeds 760 Torr and the oxide decomposes. Up to temperatures of 800 °C, no visible attack can be seen, though small losses in weight can be measured.

Table 7. Magnetic susceptibilities of precious metals (Diehl, 1995).

Metal	χ(0 K) (10^{-9} m^3 kg^{-1})	χ(20 °C) (10^{-9} m^3 kg^{-1})
Ru	4.85	5.15
Rh	11.56	12.44
Pd	82.68	63.89
Os	0.60	0.69
Ir	1.29	1.51
Pt	13.32	12.19
Ag	−2.5	−2.5
Au	−1.7	−1.76

2.3 Alloying Behavior

2.3.1 Phase Formation Solid solutions are formed with A and B metals in different compositions. Maximum solubilities of B metals with favorable size factor (ratio of atomic diameter of solute to solvent <15%; Hume-Rothery and Raynor, 1956, p. 127) are determined by the electron concentration *e*/*a*, defined by

$$\frac{e}{a} = \frac{v_1(100 - x) + v_2 x}{100},$$

where v_1 = valency of the solvent, v_2 = valency of the solute, and x = atomic percentage of the solute, at the value of $e/a = 1.4$, these values decreasing with increasing valency of the solute (Raynor, 1976, p. 13).

Intermetallic compounds with B metals with identical or similar crystal structures are formed at compositions corresponding to definite ratios of the number of valence electrons per atom (electron concentration e/a) with values of 3/2, 21/13, and 7/4, ("Hume-Rothery phases"; Hume-Rothery and Raynor, 1956, pp. 201–202). Most of these phases extend over a certain homogeneity range. The observed values of e/a are near the electron concentration at which the Fermi sphere meets the boundary of the first Brillouin zone of the corresponding crystal structures (Kittel, 1980, p. 586). Besides these e/a-determined structures, there exists a huge number of other intermetallic phases, most of them of metallic character with predominantly cubic, hexagonal, or tetrahedric symmetry.

2.3.2 Influence of Alloying Elements

2.3.2.1 Strengthening. Figure 4 shows some examples of the effect of alloy additions on hardening. Silver and gold alloys show the characteristic behavior of solid-solution hardening. Yield stresses depend on the kind and the concentration of solutes. They decrease with increasing temperature to final values ("plateau stress"; Haasen, 1994, p. 325). These are supposed to be caused by remaining interactions of the dislocations with solute-atom groups, which cannot be released by thermal activation. The critical resolved shear stress of the solid solutions is characteristic for each kind of solute.

Precipitation hardening is performed with silver–copper alloys. It is of practical importance for jewelry, dentistry, or industrial apparatus parts and construction components.

Dispersion-strengthened silver and silver alloys have high mechanical strength, corrosion resistance, and electrical conductivity. Figure 5 shows the strength of dispersion-hardened silver compared with pure silver.

2.3.2.2 Grain Refining. The most common used grain-refining additive for silver is nickel, alloyed in concentration of 0.15 wt % ("fine-grain silver"). Gold, zinc, aluminum, antimony, and cadmium increase the recrystallization temperature.

2.3.2.3 Electrical Conductivity. Figure 6 shows some examples of effects of alloy additions on the electrical conductivity. In the low concentration range, the electrical conductivity increases proportionally to the atomic concentration of the solute atom. The absolute values of the resistivity increase for B-metal solutes roughly with the square of the distance of the position of the solute from the solvent in the periodic table (Linde, 1932).

2.3.2.4 Oxygen. Silver forms two oxides, Ag_2O and AgO. Both decompose at higher temperatures (Ag_2O at >150 °C, AgO

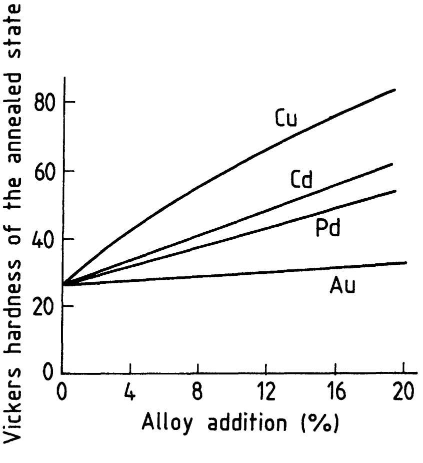

FIG. 4. Influence of alloying elements on the hardness of silver (Butts and Coxe, 1967, p. 111).

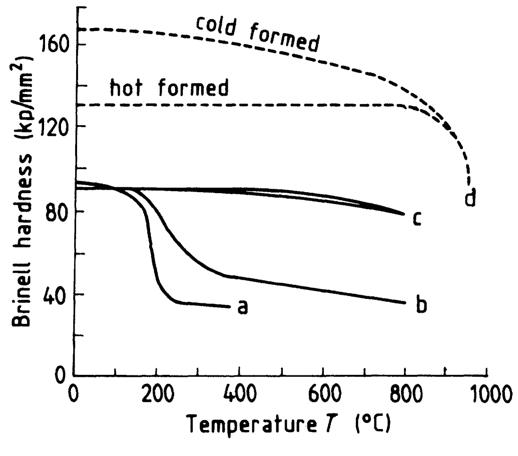

FIG. 5. Brinell hardness of (curve a) fine silver, (curve b) fine-grain silver (99.7%), (curve c) silver (99%) dispersion hardened by oxide-metal powder, and (curve d) silver (99.5%) dispersion hardened by internal oxidation (1 kp/mm^2 ~450 kPa) (Degussa, 1967, p. 36).

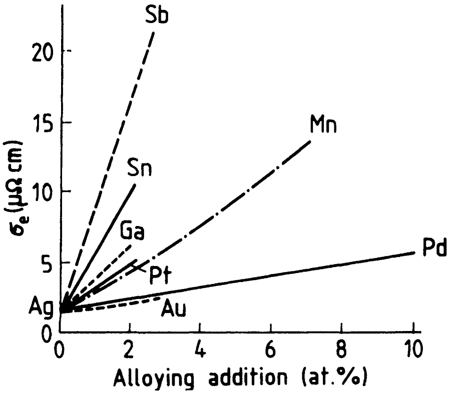

FIG. 6. Influence of alloying elements on the electrical conductivity of silver (Doduco, 1977, p. 20).

at >85 °C). Silver oxides have practical importance as catalysts and are used in considerable quantities in batteries.

Oxygen dissolves in the silver lattice, occupying interstitial lattice sites. The solubility increases with increasing temperature, rises very strongly at the melting temperature by a factor of ~80, and decreases then with further rising temperature (Table 8). The solution of oxygen in liquid silver is favored by the formation of Ag_2O, which dissolves in the liquid silver. The melting temperature of silver decreases with rising oxygen pressure of the surrounding gas atmosphere.

2.4 Abundance, Resources, and Raw Materials

2.4.1 Abundance and Resources The abundance of silver in the region accessible by mining (to a depth of 5 km) amounts to an average of 0.1 to 0.5 ppm, which is at least 100 times higher than that of gold and platinum. The abundance of silver in the oceans is estimated to be ~0.001 ppm (~10^9 t). This enormous reserve cannot, however, be extracted economically. The present known exploitable silver resources are estimated to be ~250 000 t.

Table 8. Oxygen uptake by silver (Degussa, 1967, p. 57).

T (°C)	Solubility (ppm)
200	0.03
400	1.4
600	10.6
800	38.1
973	3050
1000	3000
1100	2700
1200	2500

2.4.2 Raw Materials Native silver occurs in the form of nuggets, dendrites, or lumps. Main sources are silver sulfides in complex ores with lead or lead–zinc found in the United States, Mexico, South America, Africa, and Australia. They account for ~50% of the world silver production. The next most important sources are argentiferous copper ores, with largest reserves in Europe, which contribute ~30% of the world silver demand. Further amounts result from extraction from gold ores (~10%) and tin ores (~2%).

2.5 Production

2.5.1 Primary Production True silver ores are treated by the aqueous cyanide process. The ore is crushed, wet milled, and leached with an aqueous solution of NaCN. Solids are removed by filtering. Silver is then precipitated either by cementation with zinc dust or by electrolytic reduction. By far, most of the silver comes from lead and zinc ores. It is concentrated together with gold and the platinum metals by stirring the lead melt with zinc, precipitating, and separating the formed AgZn crystals. The zinc is distilled, leaving a 50% silver concentrate with predominantly lead (Parkes process). Silver is extracted by blowing air onto the liquid melt. The formed liquid PbO layer dissolves oxides of the base metals and up to 5% silver. It is continuously withdrawn for renewed extraction, leaving highly concentrated silver of about 99% purity (cupellation process). Silver from silver-containing copper ores is collected in the anode slimes of electrolytic copper refining and extracted by hydrometallurgical or pyrometallurgical processes.

Refining is done by electrolysis. Silver is deposited on the cathodes and continuously removed. The remaining slime contains gold and PGMs. Purity comes up with a single run to 99.99, with a double run to 99.999%.

2.5.2 Secondary Production (Recycling)

Recycling processes include chemical, pyrometallurgical, hydrometallurgical, and electrochemical processes. The extraction of low-grade scrap is performed by melting with collector metals. These dissolve the major portion of the precious metals. The slags, oxides, and base-metal alloys with lesser amounts of precious metals are separately treated or fed back to the preceding operation. The collector metal for silver and gold is lead; the collector metal for essentially PGM-containing materials is copper. Figure 7 shows schematically the recycling process of silver scrap by lead extraction in the shaft furnace and the concentration by removing base-metal oxides together with PbO in the converter furnace.

2.6 Manufacturing

Fine silver and silver alloys are melted and cast into bars, rods, and granules either discontinuously or in continuous processing under a protecting-gas or reducing-gas atmosphere. Processing to foils and wires is done by extrusion, rolling, or drawing. Pure silver wires can be drawn to very fine diameters without intermediate annealing. Smaller items are produced by investment casting; complex-shaped pieces are made by the "lost-wax method" (see Gold, page 255). Silver powders of different grain sizes and shapes are produced from the melt by water- or gas-spray processes, or by electrolytical precipitation. They serve as base material for powder-metallurgical–produced composites and as components in metallizing preparations. Silver granules are formed by pouring a thin stream of molten silver into water.

Dispersion-hardened silver and silver alloys are produced either by powder-metallurgical methods or by internal oxidation of alloyed base-metal components—e.g., aluminum. In the concentration range of 0.1–0.2 vol % of solute, spherolithic oxide particles of sizes less than 20 nm are formed (Poniatowski and Clasing, 1968). Silver-containing composites are made from mixtures of silver powder with powders of oxides, refractory-metal carbides, or nickel by pressing, sintering, and extrusion. Extruded powder com-

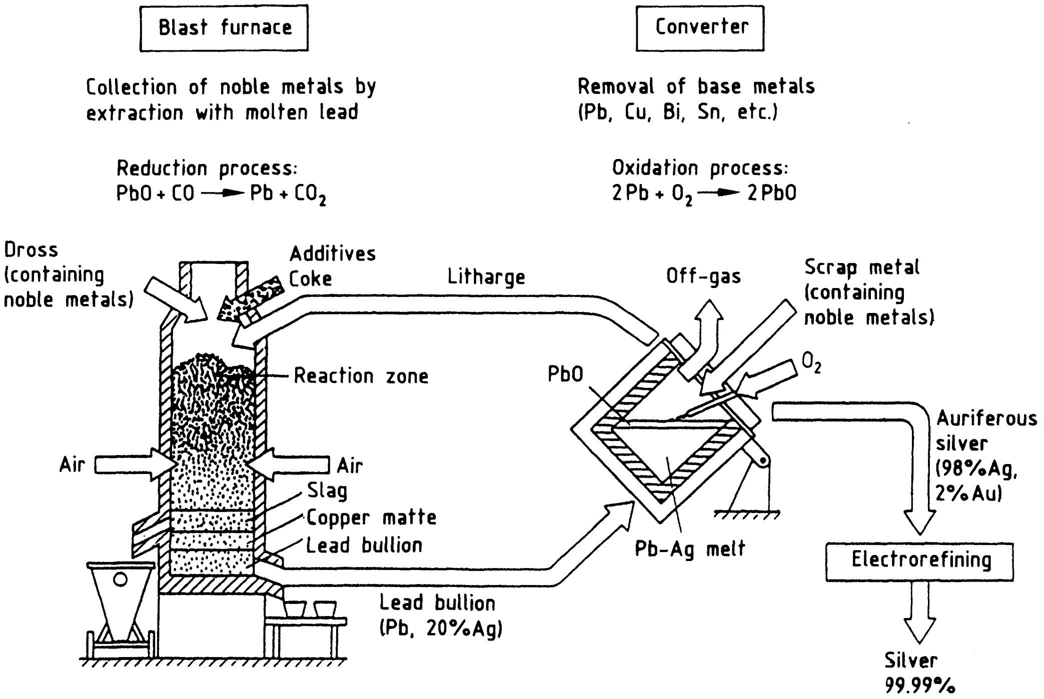
Blast furnace
Collection of noble metals by extraction with molten lead
Reduction process:
$PbO + CO \rightarrow Pb + CO_2$
Converter
Removal of base metals (Pb, Cu, Bi, Sn, etc.)
Oxidation process:
$2Pb + O_2 \rightarrow 2PbO$
Dross (containing noble metals)
Additives Coke
Litharge
Off-gas
Scrap metal (containing noble metals)
O_2
Reaction zone
PbO
Auriferous silver (98%Ag, 2%Au)
Air
Air
Slag
Copper matte
Lead bullion
Pb-Ag melt
Electrorefining
Lead bullion (Pb, 20%Ag)
Silver 99.99%

FIG. 7. Schematic view of silver recycling: extraction with lead in the shaft furnace, concentration in the converter (Renner, 1993, p. 125).

pacts show preferred orientation of the added powder grains along the rod axes. Silver composites with continuous length-oriented fibers can be produced starting with thin nickel wires placed inside silver tubes, which are then drawn together down to small diameters. Silver-nickel fiber composites are magnetic. High coercivity results if the nickel-fiber diameters become so small that no Bloch wall can be generated, and therefore only one single domain exists (Stöckel, 1979). Figure 8 shows the dependence of the coercive force on fiber diameter.

Infiltrated composite materials are produced from sintered skeletons of tungsten, molybdenum, or tungsten carbide, into which liquid silver penetrates by capillary forces.

Surface layers are produced

1. from bulk material components by mechanical cladding (rolling, pressing), thermal cladding (soldering, welding), and physical vapor deposition (PVD: evaporation, sputtering), or
2. from chemical compounds or solutions by galvanic deposition, chemical reduction, powder-based preparations (lacquers, pastes), thick-film pastes (screen printing, firing), thermal decomposition (e.g., from resinates), and chemical vapor deposition (CVD).

2.7 Alloys

2.7.1 Special Compositions The alloys AgNi0.15 ("fine-grain silver") (all compositions are in weight percent), AgSi1.5 ("hard silver"), and AgCd5 exhibit excellent corrosion resistance and high electrical conductivity. They are used as construction material for equipment and apparatus in the chemical industry.

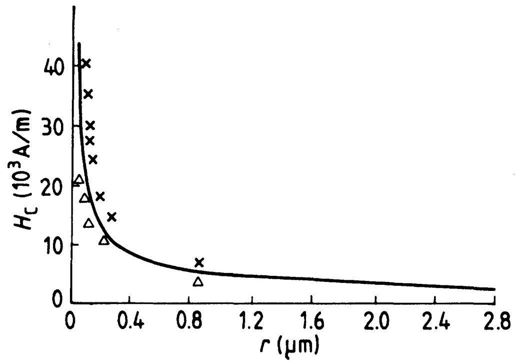

FIG. 8. Coercivity of Ag/Ni wire fiber (AgNi20) composites with small fiber diameters as a function of the fiber diameter. Crosses, 77 K; triangles, 295 K (Stöckel, 1979, p. 238).

Ag–Cu alloys. Copper effects hardening by three different mechanisms: solid solution, improved rate of work hardening, and, as a result of the miscibility gap in the alloy system, hardening by precipitation. The compositions of the majority of industrial important alloys are on the silver-rich side of the eutectic point. Ag–Cu alloys with 3–20 wt % Cu ("hard silver" alloys) are used for contacts, switches, circuit breakers, and motor contactors.

AgCu7.5 ("sterling silver"), AgCu4.2 ("Britannia silver"), and AgCu10 ("coin silver") are used for flat and hollow tableware, jewelry items, coinage, and electrical contacts.

AgCu28 (eutectic) and ternary or multicomponent alloys with Zn, Cd, Sn, Li, and/or Ni find wide application in braze alloys. Ag–Cu–P alloys are braze fillers for flux-free brazing.

AgMg0.25Ni0.2 is used for electrical contacts, and for high–thermal-conductivity spring clips for vacuum tubes, relays, and similar parts. Ag–Mg–Ni alloys harden by heating in air by internal oxidation. Nickel prevents grain growth during annealing treatments. The electrical conductivity is 75% IACS at 20 °C.

Ag–Pd alloys serve as electrical contacts and as conductor material in electronic devices. Alloys with 50–70 wt % silver are applied as electrode layers in capacitors. They are further used for brazing stainless steel, nickel-based, and other heat-resistant alloys. Small tubes of Ag23Pd77 are used in diffusion cells for the production of high-purity hydrogen.

Ag–Hg alloys play an important role in dentistry as material for fillings of cavities.

Silver alloys with tin or lead as binary, ternary, or multicomponent alloys with indium, tin, copper, and/or zinc are used as soft solders for general purpose and in special compositions as solders for electronic devices. The silver additions in the soft-solder matrix increase the mechanical strength and suppress the leaching of silver surfaces during the soldering process.

Ag–Ca alloys are the basis for the production of skeletal silver catalysts. Ag–Cu–Au alloys are used in a great variety of compositions in jewelry and dentistry. AgCd5In15, because of its high thermal-neutron capture cross section and good corrosion resistance against hot water, is used as cladding material of control rods in nuclear fuel reactors.

2.7.2 Solders and Brazes A variety of silver alloy compositions with a wide range of different melting temperatures are used as solders and brazes—*soft solders* (liquidus temperatures below 450 °C), *hard solders or brazes* (liquidus temperatures between 450 °C and 950 °C), and *high-temperature solders* (liquidus temperatures above 950 °C). Table 9 gives some typical compositions. Silver-containing solders based on tin are applied in food canning, in the installation of copper tubes, and in cryotechnology. Silver-containing eutectic binary and ternary alloys of lead and tin are used for the assembling of semiconductor devices.

Silver-based brazes are the biggest group. They are preferably used in industrial applications, electronics, and silverware production. Silver–manganese alloys are especially suited for soldering hard-metal and refractory-metal surfaces (molybdenum, tungsten,

Table 9. Noble-metal–containing solder and braze alloys (Beuers and Krappitz, 1995).

a) Soft solders

Type	Composition	Solidus (°C)	Liquidus (°C)
Sn based	Sn62Pb36Ag2	179 *E*[a]	
	SnAg3.5	221 *E*[a]	
	SnAg25Sb10	228	395
Pb based	Pb92.5Ag2.5Sn5	280 *E*[a]	
	PbAg3	305 *E*[a]	
	Pb92.5Ag2.5In5	307 *E*[a]	
Au based	AuSn20	280 *E*[a]	
	AuGe12	356 *E*[a]	
	AuSi2	370	740

[a] *E* = Eutectic composition.

b) Hard solders (brazes)

Type	Composition	Solidus (°C)	Liquidus (°C)
Ag–Cu	Ag40Cu19Zn21Cd20	595	630
	Ag60Cu27In13	605	710
	Ag44Cu30Zn26	675	735
	Ag72Cu28	779 *E*[a]	
Ag–Mn	Ag85Mn15	950	950
Ag–Pd	Ag68.4Pd5Cu26.6	807	810
	Ag54Pd25Cu21	901	950
	Ag75Pd20Mn5	100	1120
Pd–Cu	Pd18Cu72	1080	1090
Pd–Ni	Pd60Ni40	1237	1237
Au–Cu	Au80Cu20	890	890
	Au33Cu65	990	1010
Au–Ni	Au82Ni18	950	950

[a] *E* = Eutectic composition.

c) Active solders

Alloy	Composition (wt %)						Melting range (°C)	Brazing temperature (°C)
	Ag	Cu	Sn	In	Ti	Pb		
CB 1	72.5	19.5		5	3		730–760	850–950
CB 2	96				4		970	1000–1050
CB 4	70.5	26.5			3		780–805	850–950
CB 5	64	34.5			1.5		770–810	850–950
CB 6	98			1	1		950–960	1030
CS 1	86	10			4		221–300	850–950
CS 2				4	4	92	320–325	850–950

tantalum). Additions of nickel or cobalt increase the strength of the bond. Ag–Pd–Cu alloys are used to join components for high-temperature applications (gas turbines, supersonic aircraft, nuclear technology).

High-purity silver-solder alloys are needed for hard-solder processes under vacuum or in protective-gas or reducing-gas atmospheres. The minimum purity of the components is 99.99%. They must be free of metallic and nonmetallic constituents that have high vapor pressures at the soldering temperature (e.g., zinc, cadmium, arsenic). Palladium additions increase the corrosion resistance and the mechanical strength. These alloys are predominantly applied for solder connections of metallized ceramics with metals, of vacuum electron tubes, and of molybdenum, tungsten, zirconium, and beryllium.

Special compositions for jewelry and dentistry account for requirements of caratage, color, mechanical stability, and corrosion resistance.

Phosphorus-containing copper–silver-solder alloys are self-fluxing on copper, silver, and tin–bronze surfaces. These alloys find application in electrical engineering and in cryotechnology, where solder connections have to withstand temperature differences extending over wide ranges down to −50 °C.

Direct soldering of ceramic is possible with "active solders," containing ~2–4 wt % titanium or zirconium. Bonding is performed by reaction of the refractory metal with the ceramic, forming an adherent bond layer.

2.8 Applications

2.8.1 Photography Silver is used in large quantities as halide compounds, with the main starting material high-purity silver nitrate (see RECORDING, PHOTOGRAPHIC).

2.8.2 Jewelry and Silverware Silver alloys for jewelry and silverware contain copper up to 20%. Alloys with fineness (noble metal content in parts per thousand) of 925/1000 ("sterling silver"), 835/1000, and 800/1000 are the main types. Their low hardness in the soft (quenched) condition (HV 70–85) and the high ductility of sterling silver makes them best suited for sledge, bend, and filigran work. Hardening is performed by annealing at 290 °C, causing segregation of Cu atoms in the grain boundaries of the silver-rich phase. The harder 800/1000 alloy is preferred for the production of cast items with higher resistance against scratching. For enamel work the composition 970/1000 is used to avoid local melting at heating temperature exceeding 780 °C. Silver jewelry has to be protected against corrosion by coating with clear lacquers or by galvanic plating with rhodium.

2.8.3 Dentistry and Biomedical Uses Tooth cavity fillings are made with silver amalgams. They are formed from pasteous mixtures of Ag–Cu–Sn alloy powders and mercury, which harden at room temperature forming a complex matrix of silver amalgam and base-metal compounds. Typical compositions are in the range of 40 wt % silver, 32 wt % tin, 30 wt % copper, 2 wt % zinc, and ~3 wt % mercury.

Colloidal silver sols and organic silver compounds have bactericidal properties. They are used for disinfection. Implants of silver wire, gauzes, and plates have been used in brain surgery. Another area of medical application is acupuncture needles made from high-carat silver and gold alloys.

2.8.4 Electrical Technology Silver, silver alloys, and silver-bearing composites make up the largest amount of *contact materials* as make-break (= switching) contacts, especially in the heavy-current technology. Typical compositions are listed in Table 10.

Fine silver is used as a galvanic coating in thicknesses between 2 and 50 μm to lower the contact resistance and to improve the solderability of base-metal contact pieces. Values of hardness are between 60 and 160 HV, and the electrical conductivity amounts from ~50 to ~80% of that of cast metal.

Laminated composite materials are made by hot pressing, diffusion bonding, hot rolling, cold rolling, or explosive welding. Ag–Ni and Ag–Cu alloys ("fine grain" and "hard-silver" grades), soldered or clad on copper-alloy backings, and Ag–Ni composites are applied for ac and dc devices that work in the low-voltage range and current below 25 A.

Silver–metal-oxide (CdO, SnO_2, ZnO) composites cover the range of 50 to 5000 A in ac switches, especially with high current peaks on makes like low-voltage motor switches, power circuit breakers, and relays in automobiles. Contact faces that have been cut vertical to the axis of the fibers have in-

Table 10. Silver-bearing contact materials (Renner, 1993, p. 156).

Type	Composition (at. %)	Hardness (HV)	Electrical conductivity $(m/\Omega \cdot mm^2)$
Alloys	AgNi(0.15)	100	58
	AgCu(3)	120	52
	↓		
	AgCu(20)	150	49
Composites with:			
Metals	Ag–Ni(10)	90	54
	↓		
	Ag–Ni(40)	115	37
Oxides	Ag–CdO(10)	80	48
	↓		
	Ag–CdO(15)	115	45.5
	Ag–ZnO(88)	95	49
	$Ag–SnO_2(8)$	92	51
	$Ag–SnO_2(12)$	100	42
Carbon	Ag–C(2)	40	48
	Ag–C(5)	40	43.5
Refractory metals/compounds	Ag–W(20)[a]	240	26–28
	↓		
	Ag–W(80)[a]	80	42
	Ag–WC(40)	130	24–30
	↓		
	Ag–WC(80)	470	—

[a] Composite made by infiltration of liquid silver into a tungsten skeleton.

creased erosion resistance in make-break contacts compared to composites with random orientation of composite particles. Silver-graphite composites with 2–5 wt % C offer high protection against contact welding under short-circuit currents. Ag–Mo, Ag–W, and Ag–WC composites are applied for power switches that work at very high temperatures. Ag–W composites with tungsten contents above 60% are made by infiltration. They withstand arc erosion and prevent softening of rocket nozzles by the heat that develops by friction during the re-entry into the atmosphere.

Fuses, made from silver or silver alloys in bulk form or clad on copper, are designed to melt, thereby breaking the current flow, if an overload or a short circuit develops.

Silver coatings on copper or aluminum supports are the current-bearing medium of conductor devices in UHF television, radar technology, coaxial cables, and space technology. Surface protection is provided by thin coatings of either gold, palladium, or rhodium.

2.8.5 Electronics Silver is used in bulk material form and as a surface layer, as conductor material, as electrodes, as contact pads in electronic devices, in assembling parts, in thick-film and thin-film circuits, and in printed circuits. Silver-containing soft solders and brazes are widely used for electrical and mechanical connections of electron tubes and semiconductor devices.

Preparations made from silver or silver-palladium powder and organic binders serve for the production of electrode layers and contact faces of capacitors (Fig. 9). Conductor pastes for metallizing of ceramic thick-film circuits (Fig. 10) are made from powders of silver, silver-palladium, and silver-platinum mixtures together with glass frits and organic binders. They are applied by screen printing and firing.

2.8.6 Heat Mirrors Thin layers of silver with thicknesses below ~0.012 μm are transparent for optical light waves in the visible range (380–780 nm) but reflect in the long-wavelength region. These layers, deposited on architectural window glasses by sputtering, reduce energy losses caused by outgoing heat radiation. Additional thin, transparent, and color-neutral oxide coatings serve as antireflex layers and to improve adhesion, corrosion resistance, and mechanical stability. Special sputtering equipment and

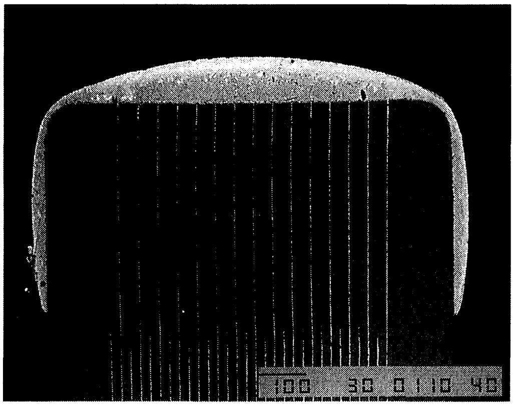
100
30 0 1 10 40

FIG. 9. Multilayer ceramic capacitor: cross section, with silver-palladium electrode layers and silver contact edge (Demetron).

target material conditions have been developed for large-scale batch type and continuous processing of coherent layers of homogeneous thickness (Dachselt *et al.*, 1982; Gläser, 1989).

2.8.7 Chemical Technology In chemical equipment, silver is used in bulk form and as surface layers for corrosion-resistant apparatus parts—e.g., crucibles for alkaline fusions or evaporations. For high-temperature applications, dispersion-hardened parts are used. Examples are fittings, bursting disks to protect equipment against damage by overpressure, and wires and rings as sealing materials for oxygen compressors and for ultrahigh-vacuum equipment. Filter elements of spheroidal silver powder are used in fruit presses. Silver-lined reactors find extended use in the food industry as vacuum pans, condensors, and evaporators. Hormones and vitamins generally are prepared in silver equipment.

2.8.8 Catalysis Metallic silver is an excellent catalyst for oxidation processes of organic compounds. It is applied in the form of wire gauzes or as precipitate on γ-Al_2O_3 or silicate carrier. Technically important processes are the production of ethylene oxide and the formation of formaldehyde from methanol.

2.8.9 Batteries The AgO/Zn button battery type is of great practical importance as current source for cameras, watches, hearing aids, and similar small devices. The worldwide yearly demand exceeds 10^9 pieces. Silver oxide forms the cathode; zinc, cadmium, or magnesium forms the anode. AgO batteries have high energy densities (80–250 W·h/dm^2), high discharge current, and high long-life stability. The cell voltage is 1.55 V. Rechargeable units with cadmium anodes reach lifetimes exceeding 1000 cycles (Benner *et al.*, 1991, p. 541).

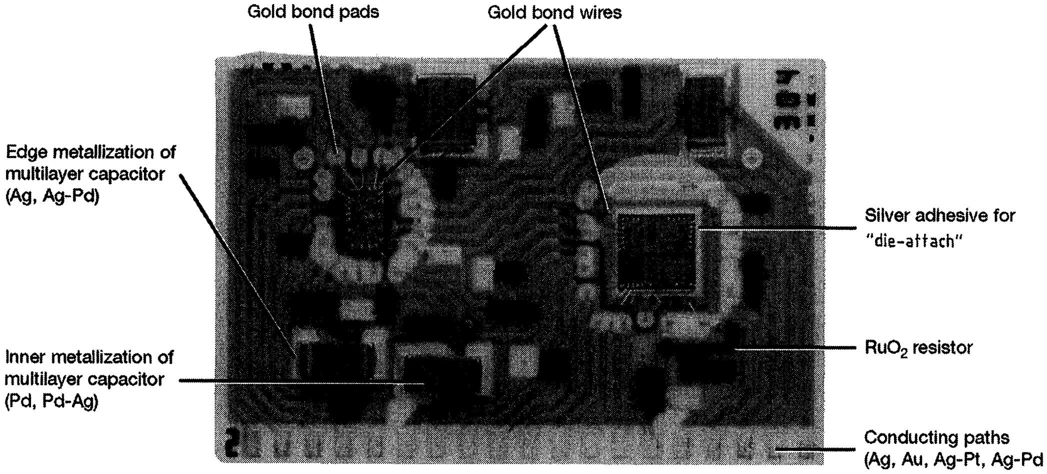
Gold bond pads
Gold bond wires
Edge metallization of multilayer capacitor (Ag, Ag-Pd)
Inner metallization of multilayer capacitor (Pd, Pd-Ag)
Silver adhesive for "die-attach"
RuO_2 resistor
Conducting paths (Ag, Au, Ag-Pt, Ag-Pd

FIG. 10. Thick-film hybrid circuit on ceramic substrate with conductor and resistor paths, bond pads, and assembled semiconductor devices (Dorbath *et al.*, 1991, p. 47).

3. GOLD

Gold, chemical symbol Au (aurum), atomic number 79, relative atomic mass 196.96654 (only one natural isotope), is a soft, brightly shining metal of the characteristic gold-yellow color. In pure condition it is very ductile. The most important properties are its oxidation resistance, its easy workability, its mechanical and thermal stability, and its high electrical conductivity. The main application areas are jewelry (86%), electrical technology/electronics (7%), and dentistry (3%) (quantities in weight percent of the total demand in the western world 1994; Degussa, 1994).

3.1 Electronic Structure

The electronic structure is [Xe] $4f^{14}5d^{10}6s$. Band diagrams for the $5d$ and $6s$ states, the density-of-states curve, and the shape of the Fermi level are very similar to those of silver (Fig. 11). The energy difference between the upper filled d band and the Fermi level is about 2.17 eV.

3.2 Special Properties

Selected values of basic physical and chemical properties of gold are given in Table 6.

3.2.1 Mechanical and Thermal Properties Pure gold is very soft and can easily be cold worked without considerable work hardening. It can be cold deformed more than 99% of its cross section without intermediate annealing. High-purity gold wires can be cold drawn without envelope down to diameters of <10 μm. Very thin beaten gold leaf is hammered in packages to foils as thin as 0.1 μm. The hardness of pure, annealed gold is ~25 HV; the recrystallization temperature is $0.28T_m$ (T_m = melting temperature in kelvins).

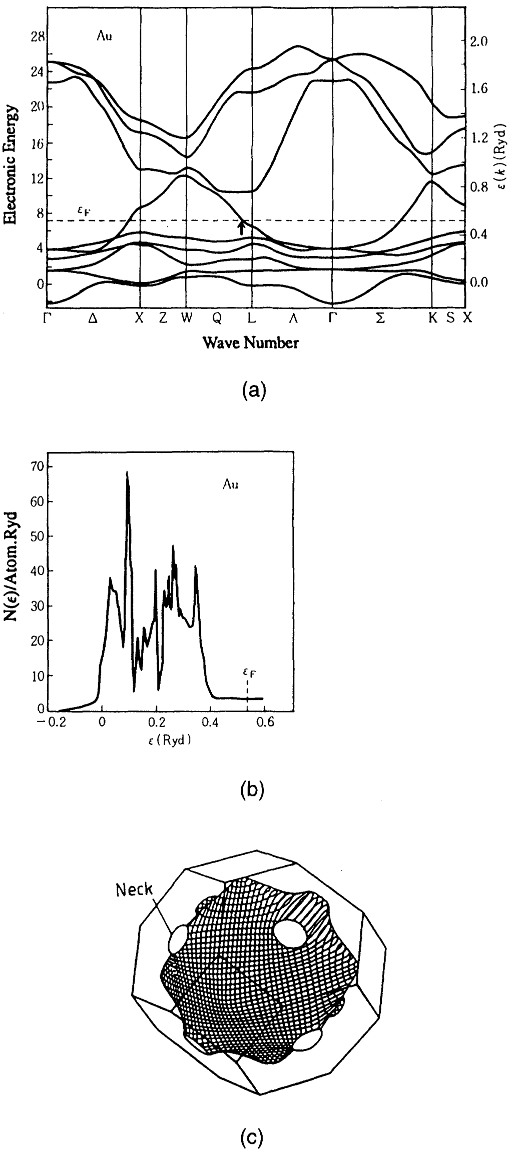

FIG. 11. Electronic structure of gold (Yonemitsu, 1991, pp. 25–28): (a) band diagram, (b) DOS curve (from Christensen and Saraphin, 1971), (c) Fermi surface.

3.2.2 Electrical Conductivity The electrical conductivity of pure gold is ~73% of that of silver. The number of electrons involved in the Fermi surface is 5.7×10^{22} cm^{-3}. Minute quantities of magnetic impurities, particularly iron, down to the part-per-billion level, influence the low-temperature electrical transport properties of gold. They cause a minimum in the electrical resistivity, accompanied by anomalies in thermoelectric power, magnetoresistance, and magnetic susceptibility (Kondo effect: Kondo, 1964, 1965). Measurements of the thermoelectric power can be used to determine the content of iron in gold down to ~0.01 ppm (Kopp, 1976).

3.2.3 Magnetic Properties Gold behaves diamagnetically. Its diamagnetic susceptibility is independent of the temperature between 0 °C and the melting point. Table 7 gives susceptibilities at zero and 20 °C.

3.2.4 Optical Properties The reflectivity of gold shows a marked fall at ~550 nm in the visible range, with a minimum of *r* ~0.25 in the near ultraviolet. The *d* bands are, compared with silver, positioned at higher energy levels. Interband transitions occur already at smaller energies (~2.17 eV) in the visible branch. The reflected light contains all wavelengths above 560 nm, which gives the typical gold-yellow color. Because of its high reflectivity in the infrared range, gold films are used in radiant heating and drying devices, for thermal-barrier windows of large buildings, and as reflecting coatings to protect space vehicles and space suits against excessive solar radiation. Ultrafine particles of gold ("gold-black") show selective light absorption. Reflectivity is less than 0.01 in the visible region and less than 0.1 in the infrared. Gold-black forms, therefore, an excellent light-receptive surface coating for radiation detectors and is also suited as selective absorption film for the capture of solar energy (McKenzie, 1978).

Thin films show optical transmission dependent on the wavelength and film thickness. Thin gold films appear green in transmission. Penetration of photons in the metal surfaces is in the range of 10^{-6} cm. Above this thickness, transmission decreases very rapidly.

3.2.5 Chemical Properties Gold does not react with water, dry or humid air, oxygen, ozone, nitrogen, hydrogen, fluorine, iodine, sulfur, or hydrogen sulfide under normal conditions. Gold withstands practically all anorganic and organic acids and aqueous solutions of alkali-metal hydroxides, alkali salts of mineral acids, and alkali-metal oxides. Also, fused caustic alkalies do not attack gold, provided air and other oxidizing agents are excluded. If a hydrohalic acid is combined with an oxidizing agent, such as nitric acid ("aqua regia"), a halogen, hydrogen peroxide, or chromic acid, gold will dissolve in the mixture. Gold reacts readily with dry chlorine. Gold can be easily recovered from solutions of gold-complex compounds by anorganic or organic reducing agents (e.g., iron, tin, or zinc salts, sulfur dioxide, hydrazine, oxalic acid) and by anion exchangers. Colloidal gold forms stable hydrosols of an intensive red or violet color. Colloidal gold is the coloring component of rubin glass.

3.3 Alloying Behavior

3.3.1 Phase Formation Gold forms alloys with most metals. Alloying behavior and phase formation follow principally the same rules as those valid for silver. Gold alloys readily with mercury at room temperature to form an amalgam.

The maximum solubilities of B metals with favorable size factor are reached at *e/a* values of ~1.3. The eutectic alloys of gold with silicon (AuSi3.1), germanium (AuGe12) and tin (AuSn20) show very low melting temperatures (Table 9). In the system gold–copper, superlattice structures are formed at compositions Cu_3Au, CuAu, and $AuCu_3$ by disorder–order transformation from the cubic A1 crystal structure to orthorhombic ($L1_0$ type for CuAu) and cubic ($L1_2$ type for Cu_3Au and $AuCu_3$). Their formation is accompanied by changes of the electrical conductivity and hardness. The superlattice phases form sublattice structures, showing gold and copper on inverted lattice sites. They form coherent domains of nine or five single cells, respectively, separated by the so-called "antiphase boundaries" (APB; Fig. 12).

3.3.2 Influence of Alloying Elements

3.3.2.1 Strengthening. The effect of alloy additions on hardening corresponds principally to the behavior of silver. Figure 13 shows some examples of the effect of alloy additions on hardening; Fig. 14 shows the dependence of the yield stress on the kind

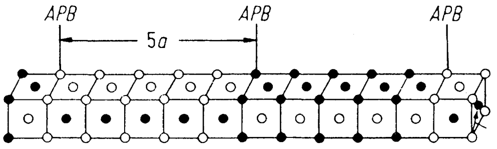

FIG. 12. Antiphase boundaries (APB) in the ordered CuAu superlattice ("CuAuII") (Haasen, 1994, p. 40, copyright Springer Verlag, Heidelberg).

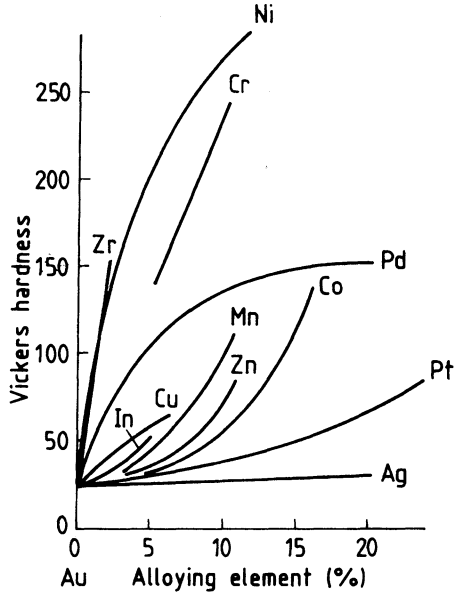

FIG. 13. Influence of alloying elements on the hardening of gold (Wise, 1964, p. 72).

and the concentration of solutes. A miscibility gap in the ternary system Au–Ag–Cu, which broadens with decreasing temperature, enables precipitation hardening in these composition ranges. Additionally, strengthening of the alloys occurs in the composition range >75 wt % Au by the formation of the ordered AuCu phase.

*3.3.2.2 **Grain Refining.*** Effective grain-refining additions for castings of gold and gold alloys (Au–Cu, Au–Ag–Cu in 14 and 18 carat) are small amounts of iridium, ruthenium, rhenium, zirconium, barium, tantalum, niobium (0.01%–0.1%), and combinations of Co/Ba, Co/Mo, and Ni/Mo. Zirconium, barium, and cobalt increase the recrystallization temperature by 150–200 °C and suppress grain growth at further annealing. Tin (0.3 wt %) and antimony (0.1 wt %) decrease the recrystallization temperature of 18-ct gold–copper alloys by ~100 °C (Ott and Raub, 1980, 1981, 1982).

*3.3.2.3 **Electrical Resistivity.*** Figure 15 shows some examples of the increase of the electrical resistivity by alloying elements. The increase of the resistivity in the low concentration range is proportional to the atomic concentration of the solute. Like in the case of silver, the absolute values of the resistivity for B-metal solutes increase roughly with the square of the distance of the position of the solute from the solvent in the periodic table (Linde, 1932).

3.4 Abundance, Resources, and Raw Materials

3.4.1 Abundance and Resources Gold is distributed very unevenly in the earth's crust. The average abundance is estimated as

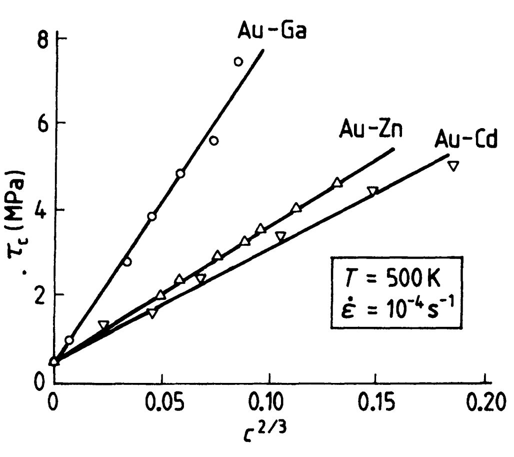

FIG. 14. Critical shear stress of solid-solution hardened gold alloys (Haasen, 1994, p. 326, copyright Springer Verlag, Heidelberg).

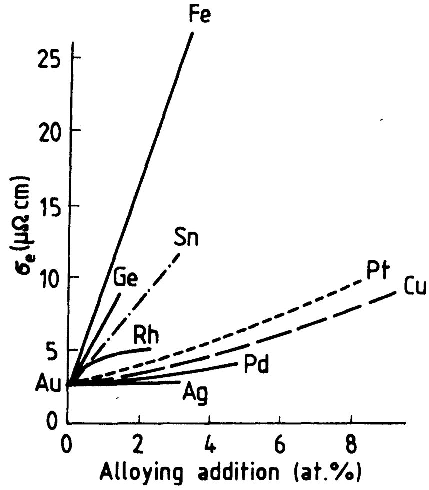

FIG. 15. Influence of alloying elements on the electrical conductivity of gold (Doduco, 1977, p. 50).

~0.005 ppm. The gold content in the oceans varies greatly depending on location. Gold contents of 0.008–4 mg/m^3 have been reported.

World gold reserves for which mining is economically viable are assessed at ~70 000 t.

3.4.2 Raw Materials The main sources contain gold as the metal—generally alloyed with small amounts of silver, copper, or platinum—as particles in quartz veins, in river sands ("placer" deposits), or included in massive rock formations. The average gold contents vary from ~0.5 ppm in placer deposits to ~12 ppm in the rock formations, when separated from the gangue. Most important deposits are in South Africa (Witwatersrand), Russia, the United States, Papua New Guinea (Ok Tedi), and Brazil.

3.5 Production

3.5.1 Primary Production Gold particles are concentrated and separated from the gangue by milling, flotation, or panning. The most important process for extraction is cyanide leaching in the presence of particles of activated carbon ("carbon in pulp" process). The gold-cyanide solution is adsorbed on the surface of the carbon particles. The process of solution and adsorption runs in several stages in countercurrent flow. The loaded carbon particles are removed from the pulp by sieving, and the adsorbed gold cyanide is eluted with hot NaCN solution. Gold is recovered either by precipitation with zinc or by electrolysis with stainless-steel electrodes. New developments are directed to the exchange of carbon against organic resins as adsorbing agent for the AuCN solution ("resin in pulp"). In metallurgical extraction processes, gold and silver follow the same route. Zinc, lead, and copper act as collecting agents for gold through the formation of alloys. Refining is done by leading chlorine gas over the surface of molten gold at about 1100 °C (Miller process), removing silver and base metals as slags and volatile components, and leaving gold of ~99.5% purity. Electrolysis in HCl and $H[AuCl_4]$ in aqueous solution refines to purities of 99.9% to 99.99%. Tetrachloroauric(III) acid can be also extracted from aqueous solution by solvent extraction with organic solvents that form complex gold compounds (Fieberg and Edwards, 1978).

3.5.2 Secondary Production (Recycling) Recovery processes from secondary materials like alloys, sweeps, or surface-coated materials is performed by principally the same kind of processes as applied for the primary production. Gold can be extracted from solutions with very low gold content (<0.1 g/L) by anion exchange resins or activated charcoal.

3.6 Manufacturing

Bars, rods, foils, and wires are manufactured by classical metallurgical processes. For the production of gold leaf, fine gold is rolled down to 20-μm thickness, then placed between sheets of vellum and beaten out with a spring hammer to leaves of thicknesses of ~0.2 μm. Alternatively, those leaves can be produced by cathode sputtering.

Casting of small jewelry items and dental pieces is done by the *lost-wax method.* Wax models of the desired shape are made in a split mold, provided with a joint gate, and embedded in a hardenable paste. After hardening, the wax is melted out and the hollow mold fired at red heat. The molten gold alloy is then poured in. After cooling, the formed pieces are removed, smoothed, and polished.

Gold powders of different grain sizes and shapes are made by precipitation from solutions by means of reducing agents. Thin surface films are produced by galvanic processes or by sputtering; thicker surface layers are produced by applying metallizing pastes made from special gold or gold alloy-powder grades.

3.7 Alloys

3.7.1 Special Compositions AuAg20 (all compositions are in weight percent) is applied for low-voltage electrical contacts, e.g., rivets. AuCo5, AgNi5, and AuAg26Ni3 are used as contact material that is resistant against Ag migration. Au–Ag–Cu alloys ("colored gold alloys") in different compositions are predominantly used in jewelry and dentistry. Melting temperatures are adjusted by additions of zinc or nickel for jewelry alloys and palladium or platinum for dental alloys.

Au–Ni–Cu and Au–Ag–Pd alloys ("white-gold alloys") are jewelry alloys, applied also in different compositions corresponding to requirements for mechanical strength and oxidation resistance. Palladium is used as an alloy component instead of nickel, if nickel may cause allergic reactions. They have higher melting temperatures but are relatively soft and cannot be age hardened by precipitation. They may be made susceptible to precipitation hardening by additions of copper.

Au–Pt alloys are, because of their high corrosion resistance, used for spinnerets in rayon production. AuPt30 serves as a high-melting precious-metal braze with liquidus temperature of 1450 °C and solidus temperature of 1228 °C. The alloys with 40%–65% Au can be effectively age hardened by quenching from 1100 °C and annealing at ~500 °C to tensile strengths of ~1400 N/mm^2, using the existing miscibility in this composition range. Small amounts (~0.5 wt %) of Rh, Rn, or Ir prevent segregation effects during the ingot production. They broaden the miscibility gap and turn the system into the peritectic type by extending the two-phase region up to the solidus.

AuPt10 and AuAg25Pt5 are alloys for electrical contacts working under heavily corrosive conditions. AuCu14Pt9Ag4 is applied for oxide-free slip contacts of high reliability at very low voltage and current—for instance, as a test value transmitter in the form of bent wire pieces or cut foils.

Au–Ni alloys form a continuous series of solid solutions at high temperature. These alloys are used as brazing alloys. The most common composition is AuNi18. Au alloys with small amounts (up to 1–3 wt %) of iron, cobalt, or nickel are electrolytically deposited as hard, wear-resistant surface coatings of electrical contacts.

3.7.2 Solders and Brazes

The low-melting-point eutectic alloys with silicon, germanium, and tin have high mechanical strength and are resistant against thermocycling. They are of special interest in the manufacturing of electronic devices.

Gold-based brazing alloys are important for jewelry, dentistry, and high-temperature-resistant solder connections. The compositions of jewelry and dentistry solders must correspond in color, corrosion resistance, and mechanical strength to the materials to be joined. Melting temperatures are adjusted by alloying with zinc and nickel to lower, and with palladium and platinum to raise, the melting temperature. Au–Ni and Au–Cu alloys have high oxidation resistance and good thermal stability up to 800 °C. The production of turbines, jet engines, and components and devices in nuclear and space technology are fields of application (Table 9).

3.8 Applications

3.8.1 Jewelry

Gold jewelry alloys are usually classified as colored and as white gold alloys (Table 11). Colored gold alloys are based on the ternary Au–Ag–Cu alloy system. It offers, in different compositions, a variety of colors and the possibility to adjust the mechanical properties according to requirements of production and wear resistance. The colors are described by special parameters (T = tone, S = saturation, D = darkness). Values for different color grades and methods for their determination have been appointed by standards (DIN 5033, 6164, 8238). The binary alloy of 99% gold and 1% Ti shows the bright color of pure gold. Its hardness of 70 HV in soft condition can be raised by work hardening and by subsequent age hardening up to 170 HV, with a fine distribution of the intermetallic $TiAu_4$ phase in the gold matrix (Gafner, 1989). The high caratage and its strength are of interest both for jewelry and for technical applications—e.g., bond wire for semiconductor devices.

White gold alloys contain either nickel or palladium. They are typically used with gold

Table 11. Main components (wt %) of gold-based jewelry alloys (Renner, 1990, p. 526).

	Gold content, wt %	Content of alloy components, wt %
Colored gold		
20 carat	88.3	Ag 0–16.7, Cu 0–16.7
18 carat	75	Ag 0–20, Cu 5–25
14 carat	58.5	Ag 8–34, Cu 7.5–33.5
8 carat	33.3	Ag 8–35, Cu 30–55, Zn 0–20
White gold		
18 carat	76	Cu 4–8, Ni 10–18, Zn 3–6
18 carat	75	Pd 10–20, Cu + Zn 5, rest Ag
14 carat	59	Cu 15–25, Ni 10–16, Zn 5–8

contents of 14 and 18 ct. Nickel white gold may cause allergic reactions. The trend is therefore toward palladium white-gold alloys, which are also superior in their ductility and corrosion resistance.

Very thin (<0.1 μm) layers of gold, silver, platinum, and palladium on ceramic, porcelain, glass, quartz, and mica, predominantly for decoration, but also for technical purposes, are deposited by thermal decomposition of liquid noble-metal preparations. They contain organic noble-metal compounds ("resinates," organic thio compounds, and carboxylates). They are applied by painting, brushing, or screen or offset printing and firing. The adherence on ceramic and porcelain is effected by additions of base-metal oxides (Bi_2O_3, SnO_2) or oxide-forming base-metal compounds. Rhodium-oxide prevents grain growth and agglomeration of the deposited noble-metal pigments (Hunt, 1979). Addition of fine precious-metal powders to the preparations leads to the formation of matte surfaces ("burnished gold," "burnished silver").

3.8.2 Dentistry Pure gold foil and matte gold are used for direct filling of cavities. Gold alloys form the major part of inlays, crowns, multiple-unit bridges, and partial dentures. Special gold alloys are provided for successive surface sealing with ceramic veneers (PFM alloys—porcelain fused to metal). The available hardness of dental alloys ranges from 50 to ~275 HV 5. Alloy additions of silver, palladium, or platinum and copper, tin, indium, zinc, or gallium affect hardness and melting range and improve the adhesion to ceramic (Kempf and Hausselt, 1992).

3.8.3 Electrical Technology Gold surfaces enable the formation of electrical connections with lowest Ohmic resistances. Gold and gold alloys are indispensible for conductor material and low-voltage contacts for communication and information, where even thin oxide layers would cause interruptions or failures in the signal transfer. AuCr2 serves as resistor material in corrosive environments with stable temperature coefficient of resistance (TCR) between -20 and $+40$ °C. Wires of multicomponent alloys based on gold or palladium serve as potentiometer windings. Their specific resistance values range from 15 to 150 $\mu\Omega\cdot$cm with TCR extending from $+10^{-4}$ to -10^{-5} (deg^{-1}). Sliding contacts are made from Pd–Ag or Ag–Au–Cu alloy.

3.8.4 Electronics Gold-plated surfaces serve for connections by pressure, bond, or solder contacts on semiconductor chips and for the joining of assembling parts and package components. The surface coats are deposited by galvanotechnic, by sputtering, or by screen printing of thick-film pastes. Gold powders of different grain sizes and spherical or platelike grain shapes are used in metallizing and thick-film pastes to produce bond pads and conductor layers. The eutectic alloys of gold with silicon, germanium, and tin (Table 9) are used for low-temperature solder connections with high mechanical strength. Gold and gold–silicon alloys are used to connect the semiconductor crystals with the base contact supports ("die-attach") on lead frames, packages, or hybrid circuits. Thin wires ("bond wire", standard diameter 25 μm) made of pure gold, alloyed with small amounts (~7 ppm) of beryllium, serve for the connection of the majority of monolithic integrated semiconductor chips ("IC") with the inner leads of the device (thermocompression bonding). Gold-plated metallic lids and solder preforms of the alloy AuSn20 are used for the hermetic (vacuum tight) sealing of devices.

4. PLATINUM AND THE PLATINUM-GROUP METALS

4.1 Common Characteristics

4.1.1 General Description The six platinum-group elements ruthenium, rhodium, palladium, osmium, iridium, and platinum are high–melting-point, silver-white metals of high density. The most important properties are the high resistance against oxidation, high strength and thermal stability, and their capability to act as catalysts in numerous inorganic and organic chemical reactions.

4.1.2 Electronic Structure Palladium, platinum, rhodium, and iridium have fcc structures. The electronic structures of the isolated atoms are for rhodium $[Kr]4d^85s^1$, for palladium $[Kr]4d^{10}5s^0$, for iridium $[Xe]4f^{14}5d^76s^2$, and for platinum $[Xe]4f^{14}5d^96s^1$ (see Table 14). They have similar densities of state (DOS) with more than five high *d*-band peaks that overlap with the broad *s* band. The Fermi levels of palladium and platinum are only slightly above the *d*-band peak. The corresponding Fermi surfaces have complicated shapes, forming Fermi hole sheets and Fermi electron sheets (Figs. 16 and 17). The numbers of *d* electrons of rhodium and iridium are smaller than those of palladium and platinum; the Fermi levels are located at the midslope of the peak of the DOS curve. The band diagrams show an *s*-like electron sheet around the same symmetry point as for palladium and platinum. The Fermi surfaces have more complicated shapes and consist of several sheets (Figs. 18 and 19).

The crystal structures of ruthenium and osmium are hexagonal closed packed. The electronic structures of the isolated atoms are for ruthenium [Kr]$4d^7 5s^1$ and for osmium [Xe]$4f^{14} 5d^6 6s^2$. The band diagrams and the Fermi surfaces (Figs. 20 and 21) are much more complicated than those of the elements with the fcc structure and show a variety of hole and electron sheets (Yonemitsu, 1991).

4.1.3 Physical Properties

4.1.3.1 Mechanical and Thermal Properties. Palladium and platinum have low strength and can easily be cold worked. The work hardening is not very marked. Rhodium and iridium have much higher strength and have only limited workability at room temperature. They can be easily deformed at higher temperatures (rhodium >200 °C, iridium >600 °C), accompanied by strong work hardening. Ruthenium and osmium exhibit high strength. They can only be deformed at high temperatures and only to a limited degree. Work hardening for these metals is stronger than that for rhodium and iridium. Ruthenium that has been deoxidized by melting with rare-earth metals can be hot rolled.

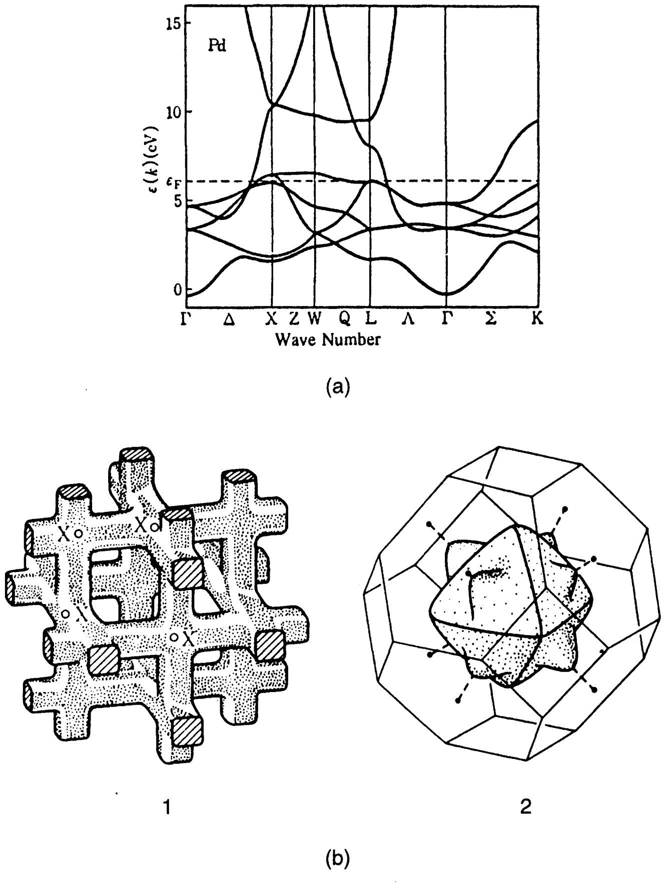

FIG. 16. Electronic structure of palladium: (a) band diagram (from Fong, 1974), (b) Fermi surface: (1) hole sheet, (2) electron sheet (from Mueller *et al.*, 1970).

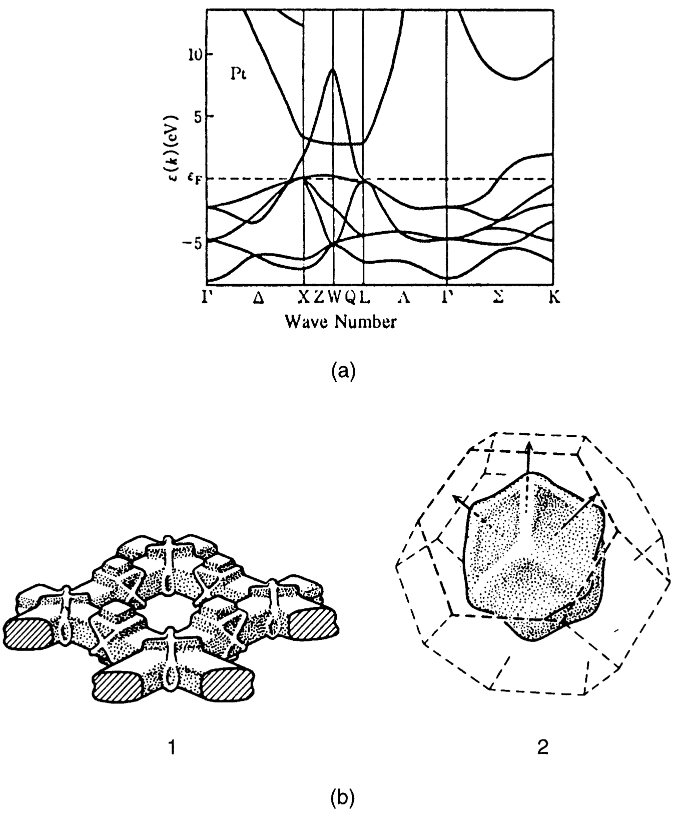

FIG. 17. Electronic structure of platinum (Yonemitsu, 1991, pp. 31–32): (a) band diagram (from Rogers, 1974), (b) Fermi surface: (1) hole sheet, (2) electron sheet (from Ketterson and Windmiller, 1970).

Figure 22 shows the increase of hardness caused by cold deformation. The rate of work hardening increases with increasing modulus of elasticity. There are two separate branches for cubic and hexagonal structures (Fig. 23; Darling, 1973, p. 93). The deformation behavior can be related to the ratio of the bulk modulus to the rigidity modulus (Table 12). Lower values correspond to the harder and brittle metals (Pugh, 1954).

The melting temperatures of the PGMs are between 1554 °C for palladium and 3045 °C for osmium. The recrystallization temperatures in relation to the melting temperatures decrease with increasing nuclear charge (ruthenium $\sim 0.57 T_m$, rhodium $\sim 0.43 T_m$, palladium $\sim 0.41\ T_m$, iridium $\sim 0.55\ T_m$, platinum $\sim 0.31\ T_m$). The melting points, the hardness, and the brittleness likewise decrease. The thermal expansion and the ductility increase with increasing nuclear charge.

4.1.3.2 Electrical Resistivities. The electrical resistivities are between $10.55 \times 10^{-8}\ \Omega \cdot \mathrm{m}$ for palladium and $4.78 \times 10^{-8}\ \Omega \cdot \mathrm{m}$ for rhodium. They decrease in the order palladium > platinum > osmium > ruthenium > iridium > rhodium. (Table 2).

4.1.3.3 Thermoelectric Behavior. Thermocouples made from platinum-group alloys are used to measure high temperatures because of their refractoriness and oxidation resistance. The thermoelectric voltage is defined as the voltage $E_{A,\mathrm{Pt}}$ in mV that a wire made from material A develops in conjunction with a wire made from physically pure platinum. The device forms with two junctions a closed electric circuit, with the two junctions held at different temperatures. One junction is supposed to have a temperature of 0 °C. For metals A and B, one has

$$E_{\mathrm{AB}} = E_{\mathrm{APt}} - E_{\mathrm{BPt}}. \tag{5}$$

The thermoelectric voltage is positive if on the hot junction the current flows from the platinum to material A. The thermoelectric power ϵ,

$$\epsilon_{\mathrm{APt}} = \frac{dE_{\mathrm{APt}}}{dt} \quad (\mu\mathrm{V/K}), \tag{6}$$

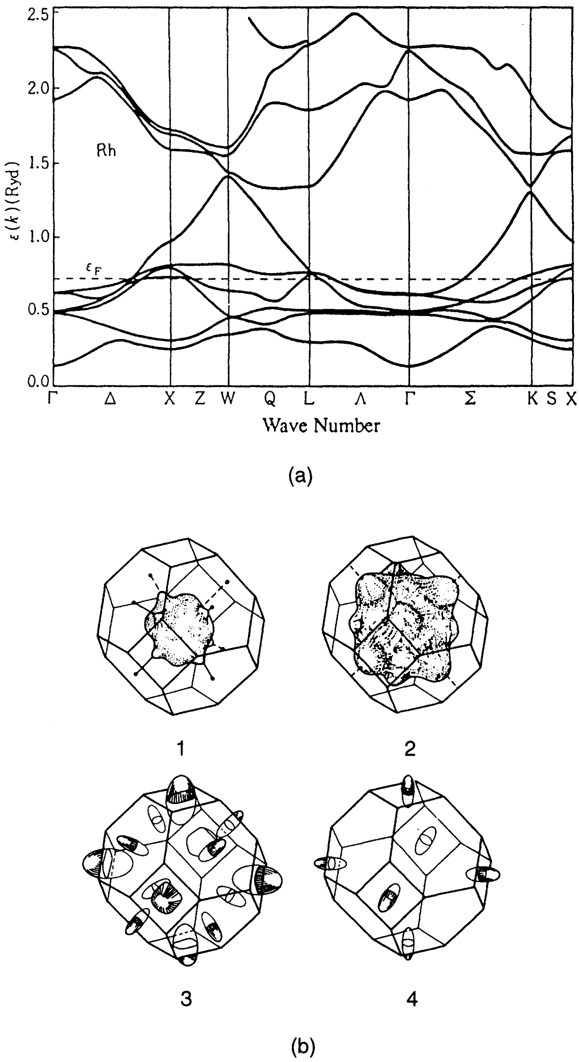

FIG. 18. Electronic structure of rhodium (Yonemitsu, 1991, pp. 33–34): (a) band diagram (from Christensen, 1973), (b) Fermi surface: (1) and (2) hole sheets in band 1 and 2, (3) and (4) electron sheets in band 3 and 4 (from Coleridge, 1966a, 1996b).

is derived from the thermoelectric voltage as a function of the temperature t by differentiation or by measuring over a small temperature difference. It is a material characteristic and defines the basic thermoelectric behavior of different materials. Figure 24 illustrates the temperature dependence of the thermoelectric power of the precious metals. Metal pairs belonging to the same group in the periodic table show similar values and temperature dependences. The values of the thermoelectric power of palladium and platinum are about five times higher than those of ruthenium and osmium. Copper and silver have positive values and show linear temperature dependence (Tanuma, 1991, p. 53).

For practical purposes, the thermoelectric voltage is the most important parameter. The thermoelectric voltage is changed by alloying elements. Table 13 gives basic values of the thermoelectric voltage of some common thermocouples in millivolts.

4.1.3.4 Magnetic Behavior. The magnetic properties are described by the mass susceptibility χ_m, which is composed of the diamagnetic portion χ_{Dia} and the paramagnetic portion χ_{para} according to

$$\chi_m = \chi_{\mathrm{Dia}} + \chi_{\mathrm{para}} \tag{7}$$

with

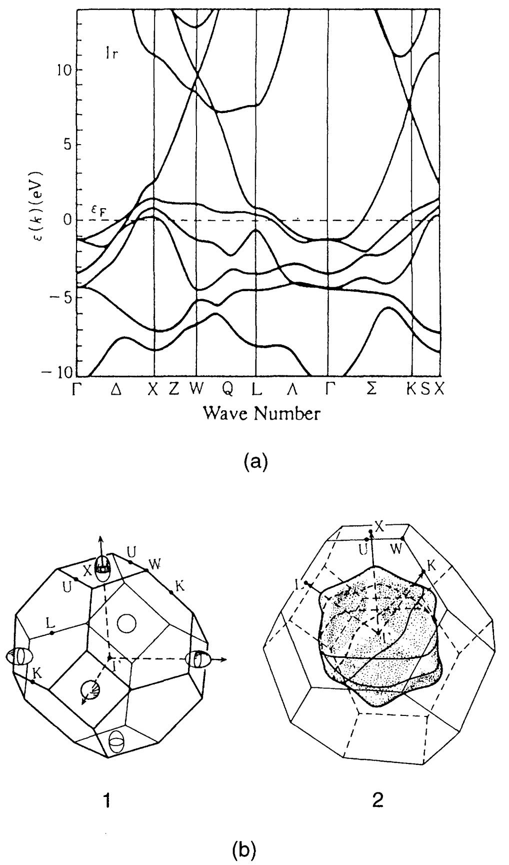

FIG. 19. Electronic structure of iridium (Yonemitsu, 1991, pp. 33, 34): (a) band diagram (from Smith, 1974), (b) Fermi surface: (1) hole sheet (from Grodski and Dixon, 1969), (2) electron sheet (from Hornfeldt *et al.*, 1973).

$$\chi_{\text{para}} = c/T \qquad (8)$$

(Curie's law) and

$$C = M^2/3R, \qquad (9)$$

where M is the magnetic moment per mole and R is the gas constant. The platinum-group metals behave paramagnetically. Palladium and platinum exhibit fairly strong paramagnetism ($\chi > 0$), since the Fermi level is located at a peak of the state-density curve in the d band, corresponding to the electronic structure of nickel. The paramagnetic susceptibilities of palladium and platinum decrease with increasing temperature, while in contrast the susceptibilities of rhodium, iridium, ruthenium, and osmium increase with increasing temperature. The change of the susceptibility with the temperature is related to the excitation of the electrons below the Fermi level. The temperature dependence of the magnetic susceptibility of the PGMs (Fig. 25) gives the experimental evidence for the different location of the Fermi levels (Chikazumi, 1991, p. 98). Palladium and platinum tend to become ferromagnetic when they are alloyed with ferromagnetic elements.

PGMs show magnetostriction under the action of a magnetic field:

$$\Delta l/l = s_l H^2, \qquad (10)$$

where s_l = specific magnetostriction and l = length of the sample. Based on this and in combination with ordering processes, alloys in the systems Pt–Fe and Pd–Fe show around the Fe_3Pt and Fe_3Pd stoichiometries in the disordered state, like Fe–Ni alloys (e.g., NiFe36, "Invar"), zero or negative coefficients of thermal expansion for similar magnetic reasons (Kussman and von Rittberg, 1950; Kussman and Jessen, 1962).

4.1.4 Alloying Behavior

4.1.4.1 Metals. Compared to silver and gold, the behavior of the PGMs is rather complex both in the range of solid solubilities and in intermetallic-phase formation. Ruthenium, iridium, palladium, and platinum form very stable compounds with the refractory transition elements, especially zirconium and hafnium. Platinum takes up to 18 at. % zirconium into solid solution and reacts even with ZrC and ZrO_2 with formation of the hexagonal compound Pt_3Zr. Relatively high solid solubilities exist for the PGMs in molybdenum and tungsten (Darling, 1973, p. 116).

A high affinity of the PGMs exists for the metals of the rare-earth group. They form numerous intermetallic compounds of different compositions and different structures. In view of the exceptionally high thermal-neutron absorption cross sections of samarium, europium, and gadolinium, these PGM–rare-earth-metal alloys are of special interest for the atomic industry (Funston and McGurty, 1961; Loebich, 1971).

4.1.4.2 Carbon. The platinum-group metals dissolve large quantities of carbon in the molten state, but the solubility in the

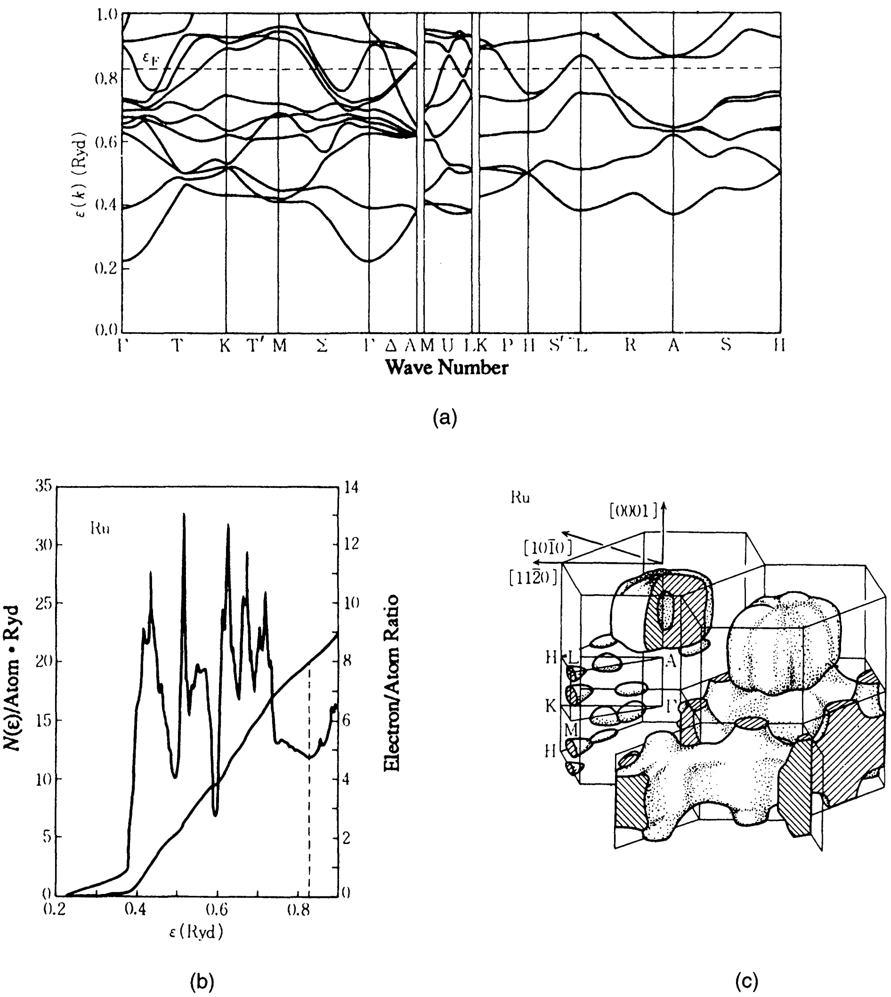

FIG. 20. Electronic structure of ruthenium (Yonemitsu, 1991, pp. 35–37): (a) band diagram, (b) DOS curve (from Jepsen *et al.*, 1975), (c) Fermi surface: connected hole sheets in the seventh and eighth bands, electron sheets in the ninth band (from Alekseev *et al.*, 1979).

solid state is very small. Carbon dissolves interstitially and effects a considerable increase of hardness, accompanied by slightly depressed melting temperatures.

4.1.4.3 Oxygen. The platinum metals form a variety of solid and gaseous oxides in different oxidation states. The solid oxides have tetragonal and hexagonal crystal structures. Rhodium and palladium form stable oxide skins when heated in air. Both also take oxygen into solid solution. The solubility of oxygen increases in the order Rh < Pd < Ag. PdO has a tetragonal crystal structure and is semiconducting; RhO_2 is an electrical conductor; RuO_2 is a semiconductor.

4.1.5 Catalytic Actions The platinum-group metals show pronounced catalytic actions in numerous chemical reactions (e.g., reforming processes in petrochemistry, oxidation of NH_3 to NO, and hydrogenation of hydrocarbon compounds). The catalytic activities are essentially determined by the heat of adsorption of the reactant molecules on the catalyst surface and the bond strength of the adsorbed molecules. The platinum metals provide moderate values of heat of adsorption, which correspond to the dissociation energy of the reactant molecules. The heat of adsorption for each element differs also for different crystal planes; it increases

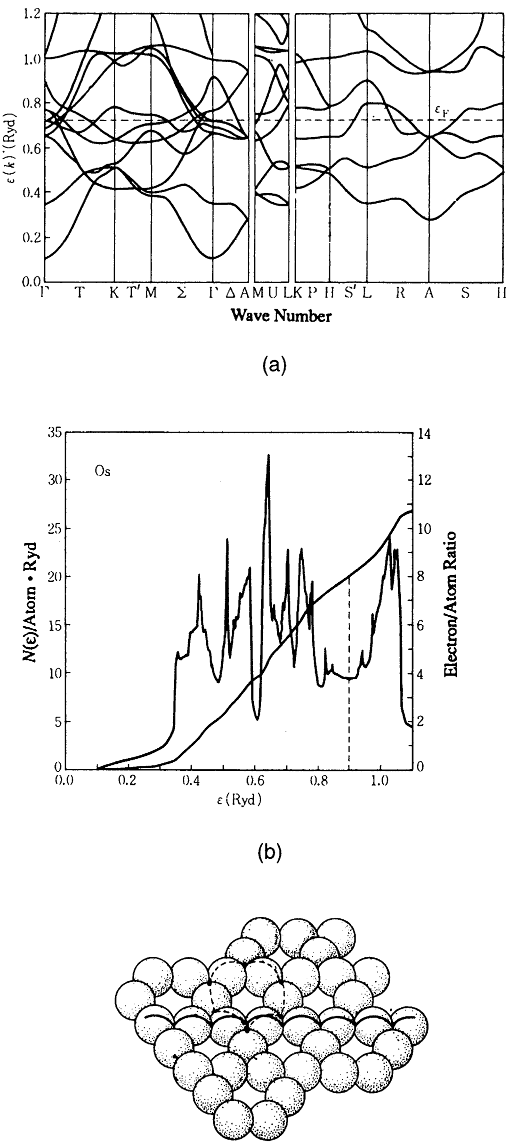

(a)

(b)

(c)

FIG. 21. Electronic structure of osmium (Yonemitsu, 1991, pp. 35–37): (a) band diagram, (b) DOS curve (from Jepsen *et al.*, 1975), (c) Fermi surface: connected hole sheets of the eighth band (from Plachy and Dixon, 1974).

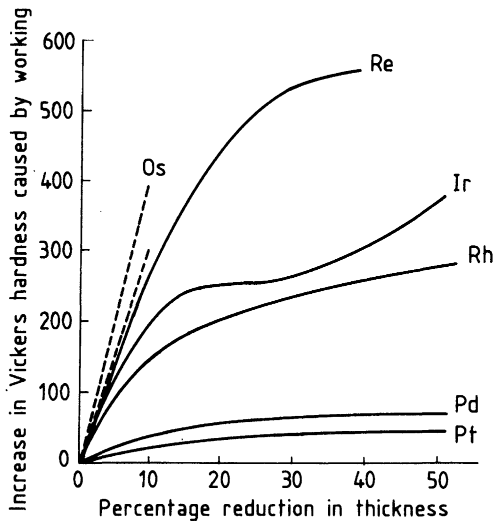

FIG. 22. Work hardening of the PGMs: increase by reduction (Darling, 1973, p. 93).

for fcc crystals with decreasing density of the surface atoms in the order [111] < [100] < [110] (Toyoshiura and Somorjai, 1979; Tamaru and Naito, 1991, p. 266). Catalytic actions are structure sensitive. Correlations exist between the lattice dimensions of catalysts and the dimensions of adsorbed molecules. As a consequence of these influences, the effectiveness of the catalytic action of the individual PGMs differs for different reaction types and follows changing order ranges. They decrease, for example, for the hydrogenation of C = C double bonds in the order Rh > Ru > Pd > Pt, for the oxidation of paraffines Pd > Pt ≫ PdO (Benner

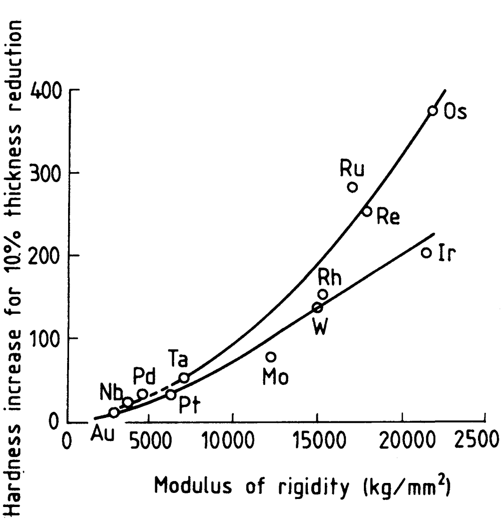

FIG. 23. Work hardening of cubic and of hexagonal metals: hardness increase at different values of the rigidity (Darling, 1973, p. 93).

Table 12. Mechanical properties of PGMs (Darling, 1973, p. 94).

Metal	Crystal structure	Young's modulus (GPa)	Modulus of rigidity, *G* (GPa)	Bulk modulus, *K* (GPa)	Poisson's ratio	*K*/*G*
Os	hcp	560	220	380	0.25	1.73
Ir	fcc	538	214	378	0.26	1.76
Re	hcp	472	180	340	0.26	1.89
Ru	hcp	430	172	292	0.25	1.71
W	bcc	396	151.4	318.6	0.29	2.11
Rh	fcc	386.4	153	280.1	0.26	1.83
Pd	fcc	128.3	46.1	190.9	0.39	4.13
Pt	fcc	174	62.2	280.9	0.39	4.52
Au	fcc	80.2	28.2	174.6	0.42	6.18

et al., 1991, p. 765), and for the synthesis of methane from CO and H_2O in the order Ru > Rh > Pd > Pt > Ir (Schuit and van Reijen, 1958; Vannice, 1975; Broden *et al.*, 1976). The selectivity of the catalytic actions and the long-time stability of the catalysts can be improved by alloying.

In heterogeneous catalysis the PGMs are applied in wire gauze form or as highly dispersed metal coatings ("platinum black," "palladium black") onto supports such as carbon, γ-Al_2O_3, silica gel, aluminum silicate (zeolite), or $CaCO_3$. The deposition is performed by thermal decomposition or chemical reaction of precious-metal salts. The supporting material can be in the form of powder or granules; the concentration of noble metals varies between 0.5% and 5%. Because of the surface-induced reaction, high specific surfaces (usually in ranges between 100 and 1000 m^2/g) are favorable for higher yields and reaction rates.

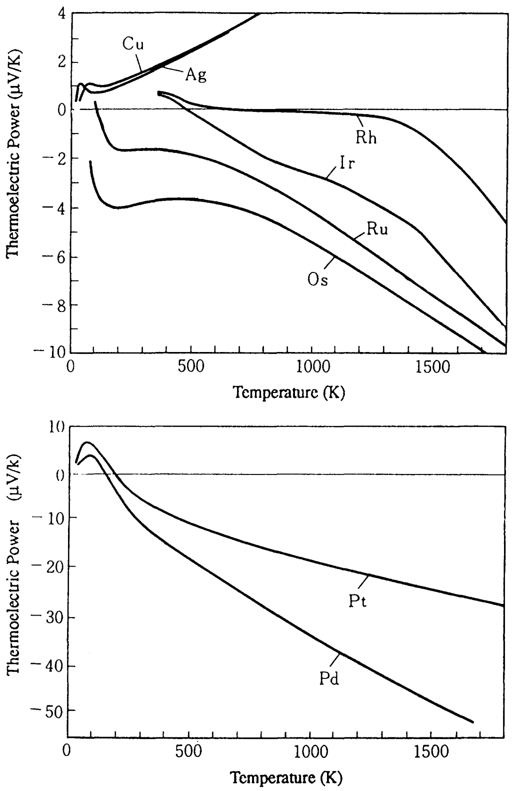

FIG. 24. Thermoelectric power of precious metals (from Vedernikov, 1969).

In homogeneous catalysis the reactants and the catalyst are both present in liquid, preferably organic, solution. Halide compounds of palladium and complex organic compounds of rhodium, iridium, and ruthenium are important for the stereospecific synthesis of emantiomeric compounds.

4.1.6 Abundance, Resources, and Raw Materials

4.1.6.1 Abundance and Resources.

The platinum-group metals are among the scarcest of metallic elements. The abundance of platinum and the associated platinum-group metals is estimated as on the order of 10^{-4} ppm, concentrated in planetary regions reaching up to ~30 ppm.

The largest deposits are found in South Africa (Bushveld), in the United States (Stillwater, Oklahoma), in Canada, and in Russia. The latter two are the most important sources for palladium. PGM contents in exploitable minerals and deposits vary from 0.05 ppm (secondary deposits in washable sands) to 3–20 ppm for the most important deposits and up to platinum concentrations of 200 ppm in special ores.

Table 13. Basic values of the thermoelectric voltage[a] of common themocouples (Degussa, 1967, p. 100).

T_1 (°C)	T_2 (°C)	Pt-Rh10/Pt	Pt-Rh20/Pt-Rh5	Rh-Ir60	Pt-el[b]	Pd-or[c]
0	100	0.643	0.074	0.371	3.31	4.6
	500	4.221	1.447	2.562	20.20	27.9
	1000	9.570	4.921	5.495	41.65	59.6
T_1 (K)	T_2 (K)	Au-Co2.1/Cu[d]	Au-Fe0.02/Cu[d]			
4.2	10	0.044	0.093			
	20	0.173	0.208			
	40	0.590	0.423			

[a] In mV.
[b] Pt-el = Platinel, Pd83Pt14Au3/AuPd35.
[c] Pd-or = Pallador, PtIr10/AuPd40.
[d] At. %.

For the platinum-group metals together, the current estimate of workable deposits is ~70 000 tons.

4.1.6.2 Raw Materials. The PGMs occur as native alloys or in mineral compounds in placer deposits, sometimes together with gold, as well as in lode deposits in basic rocks together with nickel and copper. Deposits contain varying compositions in regard to the PGM mix with average distribution of the proportion in percent for platinum, 45 (19–59); palladium, 30 (25–66.5); ruthenium, 5 (1–12); rhodium, 4 (2–7.6); iridium, 1 (<1–2.4); and osmium, <1–1.2.

4.1.7 Production

4.1.7.1 Primary Production. Production steps for the PGMs are schematically shown in Fig. 26. After flotation, the material is melted under alternating reducing and oxidizing conditions, removing iron and SO_2, followed by grinding and then magnetic separation of the enriched PGM-containing Ni–Cu–Fe phase. Solution with HCl/HNO_3 or HCl/Cl_2 removes most of the base metals, leaving a PGM concentrate of above 30%. Osmium and ruthenium are separated by distillation of their volatile oxides from concentrated hydrochloric solutions. The remaining PGMs are separated in aqueous solution by subsequent oxidation, reduction, and precipitation of complex PGM chlorides in the sequence platinum, iridium, palladium, rhodium.

PGMs can also be recovered from solutions by solvent extraction. It works with quantitative yield and is highly selective.

Refining is done by repeated solution, recrystallization, and precipitation of compounds. Conversion of metals to sponge or powder is carried out for platinum, palladium, iridium, and rhodium by thermal decomposition of the ammonium-chloro complexes ("calcination").

Ruthenium is prepared by precipitating RuO_2 from aqueous solution and reduction with hydrogen; osmium is prepared by precipitating potassium osmate (IV) and heating in hydrogen atmosphere or by reduction with $NaBH_4$ in aqueous solution.

4.1.7.2 Secondary Production (Recycling). Recycling of PGM scrap has become of special importance for platinum, palladium, and rhodium with the growing

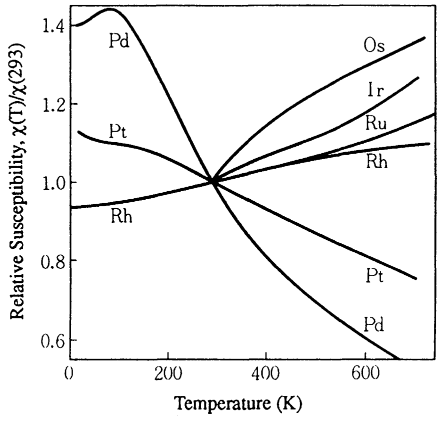

FIG. 25. Temperature dependence of the magnetic susceptibility of the PGMs (from Berkowitz and Kneller, 1969).

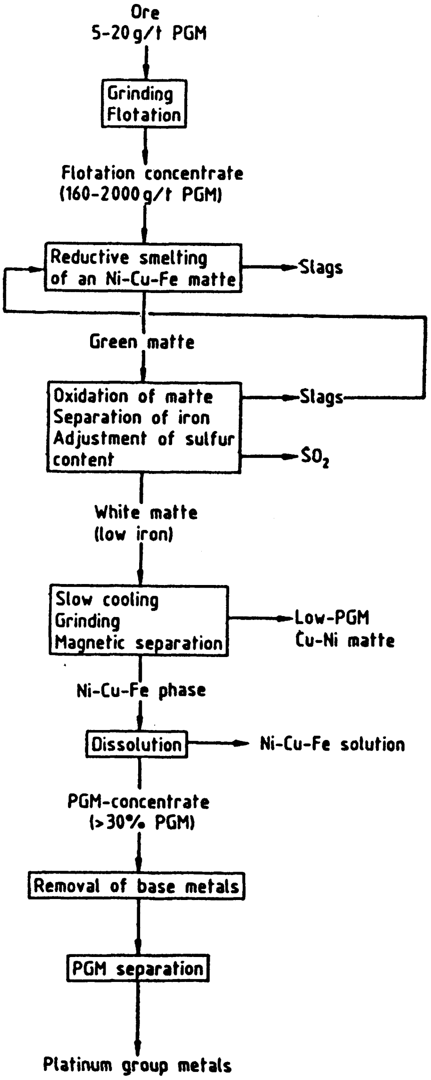
Ore
5-20g/t PGM
Grinding
Flotation
Flotation concentrate
(160-2000g/t PGM)
Reductive smelting
of an Ni-Cu-Fe matte
Slags
Green matte
Oxidation of matte
Separation of iron
Adjustment of sulfur
content
Slags
SO₂
White matte
(low iron)
Slow cooling
Grinding
Magnetic separation
Low-PGM
Cu-Ni matte
Ni-Cu-Fe phase
Dissolution
Ni-Cu-Fe solution
PGM-concentrate
(>30% PGM)
Removal of base metals
PGM separation
Platinum group metals

FIG. 26. Production steps for PGMs (Renner, 1992, p. 88).

demand for autocatalysts. Metallic waste is dissolved or treated with chlorine with subsequent reduction and precipitation of PGM powders. Low-grade wastes are recovered by pyrometallurgical processes using lead as collector metal. Supported catalysts and automobile exhaust catalysts are melted in electric furnaces with iron or copper as collector metals, forming alloys with ~20 wt % of PGM (predominantly for platinum, palladium, and rhodium).

4.2 The Elements

4.2.1 Platinum Platinum, atomic number 78, relative atomic mass 195.08, six natural isotopes (between 190 and 198), is a silvery white, bright metal, ductile and malleable, resistant against air and water, soluble only in HNO_3–HCl and molten alkali. The most important properties are oxidation resistance; mechanical, thermal, and electrical stability; and the ability to act as catalyst in numerous different chemical reactions. The main application areas are jewelry (39%), automotive catalytic converters (42%), chemistry (4%), electronics/electrical technology (5%), glass industry (4%), petroleum (6%) (quantities in weight percent of the total yearly demand in the western world 1994; Degussa, 1994).

Selected values of basic physical and chemical properties are given in Table 14.

4.2.1.1 Influence of Alloying Elements. *Metals: Strengthening.* Platinum can be hardened by alloying with other precious metals and base-metal elements. The most effective hardeners are precious metals in the order Ru, Os, Ir, Au, Rh and the base metals in the order Co, Ni, Cu, Fe. Figure 27 shows the increase of hardness by increasing amounts of alloy additions.

Dispersion-strengthened platinum and platinum alloys are produced by coprecipitation of the metals and refractory oxides—for instance, with ZrO_2 (Selman *et al.*, 1974; zirconium-oxide grain-stabilized platinum, "ZGS"®, Johnson Matthey, 1974). Platinum can also be effectively dispersion hardened by incorporating very small amounts (0.04–0.08 wt %) of refractory carbides (TiC, ZrC). Additive strengthening is effected by partial decomposition and reaction of the carbide with the matrix. The low carbide concentration offers the advantage that ductility, electrical conductivity, and the recrystallization behavior are not seriously impaired.

Metals: Electrical resistance. Figure 28 shows the influence of some alloying elements on the electrical conductivity. In the Pt–Mo system, at composition with 30 at. % Pt, a superconducting ordered phase is formed. The partially ordered state shows the maximum effect with the highest critical transition temperature (7.5 K) (Flükiger *et al.*, 1972).

Metals: Magnetic properties. In the Cu–Pt

Table 14. Selected values of basic physical and chemical properties of platinum metals (Degussa, 1995).

	Ru	Rh	Pd	Os	Ir	Pt
Atomic properties						
Atomic number	44	45	46	76	77	78
Atomic weight	101.07	102.905	106.42	190.2	192.22	195.08
Electron configuration	$[Kr]4d^7 5s$	$[Kr]4d^8 5s$	$[Kr]4d^{10}$	$[Xe]4f^{14}5d^6 6s^2$	$[Xe]4f^{14}5d^7 6s^2$	$[Xe]4f^{14}5d^9 6s$
Crystal structure	hcp	fcc	fcc	hcp	fcc	fcc
Lattice constant (Å) (10^{-1} nm)	$a = 2.701$ $c = 4.275$ $c/a = 1.583$	3.803	3.887	$a = 2.735$ $c = 4.319$ $c/a = 1.579$	3.839	3.924
Atomic radius (Å) (10^{-1} nm)	1.325	1.345	1.376	1.377	1.357	1.377
Mechanical properties						
Density (20 °C) (g/cm^3)	12.45	12.41	12.02	22.61	22.65	21.45
Density, liquid at melting point (g/cm)^3	10.90	10.70	10.49	20.1	19.39	18.91
Hardness (HV, kg/mm^2)	250–500[a]	130	50	300–680[a]	200	48
Young's modulus (GPa)	485	386	124	570	538	173
Modulus of rigidity (MPa)	172	153	51	220	214	67
Poisson's ratio	0.29	0.26	0.39	0.25	0.26	0.39
Coefficient of compressibility	0.31	0.36	0.52	0.28	0.28	0.36
Thermal properties						
Melting point (mp) (°C)	2310	1966	1554	3045	2410	1772
Vapor pressure at mp (mbar) (1×10^5 Nm^{-2})	5.5	3.5	20	20	4	0.1
Boiling point (°C)	3900	3700	2970	5000	4130	3827
Thermal conductivity (20 °C) ($W \cdot m^{-1} K^{-1}$)	117	150	75	87	147	73
Coeff. of linear expansion (0–100 °C) (10^{-6} K)	9.5	8.5	11.1	6.5	6.8	9.0
Electrical properties						
Specific electrical resistance (0 °C) ($\mu\Omega$ cm)	6.7	4.34	9.725	8.2	4.7	9.825
Temperature coefficient of electrical resistance (0–100 °C) (10^{-3} K)	4.0	4.6	3.8	4.2	4.3	3.92
Transition temperature of superconductivity	0.47	0.9		0.71	0.11	
Thermoelectric voltage against Pt (0–100 °C) (mV)	0.684	0.70	−0.570		0.660	0
Work function (eV)	4.52	4.8	4.97	4.55	5.40	5.27
Chemical properties						
Oxidation states in compounds	8, 6, 4, 3, 2, 0, −2	5, 4, 3, 1, 2, 0	4, 2, 0	8, 6, 4, 3, 2, 0, −2	6, 4, 3, 2, 1, 0, −1	4, 2, 0
Ionization energy (eV)	7.37	7.46	8.34	8.7	9.1	9.0
Covalent radius for single bonds (Pauling) (pm)	125	125	128	126	127	130
Ionic radius (pm) (oxidation number, coordination number)	62 (4+,6) 68 (3+,6)	55 (5+,6) 67 (3+,6)	64 (2+,4) 86 (2+,6)	39 (8+,4) 63 (4+,6)	63 (4+,6) 68 (3+,6)	63 (4+,6) 60 (2+,4)
Electronegativity (Allred and Rochow)	1.4	1.5	1.4	1.5	1.6	1.4
Reduction potential (V) with (n) electrons	0.455 (2)	0.758 (3)	0.951 (2)	0.85 (8)	1.156 (3)	1.118 (2)

---

[a] Depending on crystal orientation.

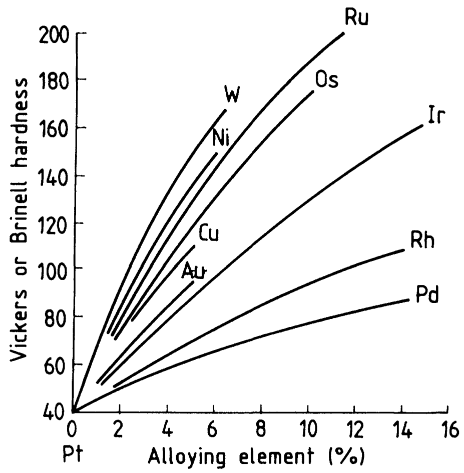

FIG. 27. Influence of alloying elements on the hardening of platinum (Zysk, 1979, p. 689).

system a long-range–ordered rhombohedrical lattice is formed below 812 °C. The equiatomic PtCo alloy is of great importance for permanent magnets. Ordering occurs in such a manner that cobalt and platinum atoms arrange on alternate planes. The strongest increases of hardness and magnetic properties are found at intermediate states where maximum strains exist by the combination of the two different lattice structures. Considerable stresses are caused during the ordering process by coherence of tetragonal platelets on the [110] planes of the cubic lattice. The highest magnetic characteristics are developed if the ordered and disordered phases are present in comparable proportions.

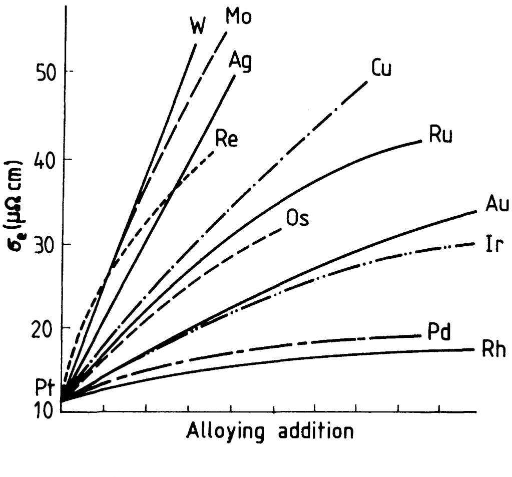

FIG. 28. Influence of alloying elements on the electrical conductivity of platinum (Doduco, 1977, p. 67).

Superlattices, formed in Pt–Cr alloys in the composition ranges of 17–65 wt % Cr, are ferromagnetic with the maximum magnetic saturation for the phase PtCr (extending from 43 to 84 wt % Cr) at ~30 at. % Cr. FePt and CoPt with tetragonal crystal symmetries have large magnetic anisotropies. In the alloy systems Fe–Pt and Fe–Pd the saturation moment increases with the addition of noble metals. Such alloys can be used as thin-film head materials for high-density recording devices.

Oxygen. The solubility of oxygen in platinum is very low. Very thin coatings of platinum on reactive metals are therefore an effective protection against oxidation. Platinum forms three solid oxide compounds, which are stable at different temperatures and exhibit different oxygen partial pressures. Solid PtO_2 in the hexagonal modification is used as a catalyst. Volatile oxides form if platinum or platinum–rhodium alloys are heated in oxygen-containing atmospheres. They cause remarkable weight losses in long-term runs. The oxides can be recovered by leading the outcoming gas stream through a gauze of gold-palladium alloy, on which the platinum and rhodium oxides are deposited.

4.2.1.2 Alloys. Platinum–palladium alloys are used in jewelry, for electrical contacts, and for equipment parts for glass production and construction elements in highly corrosive environments.

Platinum–iridium alloys find various applications in electrochemistry, chemical technology, medicine, and jewelry. Iridium contents (in weight percent) are for laboratory ware 0.4%–0.6%, jewelry 5%–15%, medical 10%, electrical contacts 10%–25%, electrodes 10%, and pens, needles, and springs 25%–30%. Maximum hardness (~360 HB) is reached with 30% iridium.

Platinum–rhodium alloys are applied in composition with 10% rhodium as catalyst for the oxidation of NH_3, as a standard element in thermocouples (Pt–Rh of different compositions against Pt), and as heater elements in electrically heated high-temperature furnaces. Pt–Rh alloys with addition of Au are especially resistant against wetting by molten glass.

Ruthenium strengthens platinum very effectively. Ten percent ruthenium raises the hardness to 168 HV. The alloy PtRu5 has the same strength, hardness, and corrosion resistance as the alloy PtIr10 ("hard platinum"). Applications are jewelry items, electrodes for electrochemistry, spark-plug electrodes, electrical contacts, and wear-resistant wires in potentiometers. Higher alloy compositions up to 14% ruthenium and ternary types with 15% rhodium and 6% ruthenium are used for electrical contacts and potentiometer wires intended for heavy-duty applications.

Platinum–nickel alloys have been proven as electrical switches and as high-strength and high-temperature–resistant construction materials in the chemical industry. Compositions reach from trace amounts up to 20% nickel with maximum tensile strength of ~1725 MPa after 90% cold reduction.

Platinum–tungsten alloys, commonly used in compositions with 5–8 wt % Co, have excellent wear resistance. They are used for potentiometers, glow wires in copiers, diode switches, heater elements, and claddings on rods for fuel elements in nuclear reactors.

Platinum–cobalt alloys with compositions near 50 at. % Co form strong permanent magnets with coercivities up to 540 kA/m at room temperature. Typical uses are focusing magnets, hearing aids, miniature motors, medical instruments and implants, and films for magneto-optical recording.

Platinum–gold and platinum–gold–rhodium alloys have higher strength than pure platinum at 1000 °C, finer grain structure, and better resistance against wetting by molten glass.

4.2.1.3 Applications. *Jewelry.* The reflectivity of platinum in the visible-light spectrum of 69% (for comparison: silver 91%, palladium 57%) gives it its characteristic appearance, similar to white gold. This, together with the higher hardness of platinum and platinum alloys compared with gold, has led to a fast-growing platinum demand for jewelry items. Platinum resinates are employed for decoration on porcelain surfaces.

Catalysis. The largest demand for platinum catalysts in the chemical industry comes from the petroleum refinery. Catalysts are Pt–Al_2O_3, Pt–Ir, and Pt–Re. Further uses are hydrorefining, hydrocracking, and hydrogenation processes and synthesis of medical drugs.

Automobile exhaust-gas catalysts take the biggest portion of the total demand of platinum (~42%) and rhodium (~89%). They reduce and remove the gaseous and particulate pollution constituents from exhaust-gas streams by catalytically activated reduction and oxidation reactions. The three-way catalyst for gas engines contains as active part a highly dispersed metallic coat of a mixture of platinum and 20 wt % rhodium and palladium in changing proportions. The metal coat is precipitated on a honeycomb-shaped ceramic support ("monolith"), made from cordierite (Fig. 29). The monolith is coated with a ceria-doped "washcoat" of Al_2O_3, which forms on firing a porous structure with a high specific surface area. The oxidation of hydrocarbons (CH_x) is activated by platinum and palladium, the reduction of nitrogen oxides (NO_x) by rhodium (Amirnazmi and Boudart, 1975; Obuchi *et al.*, 1982, 1983). The reaction has its best yield at an air-to-fuel ratio of 14.7:1. The ratio is measured and maintained by means of two platinum sensors. One is installed in the inlet manifold to monitor the proportion of air in the air-fuel mixture as it enters the combustion chamber; the second forms the tip of the so-called lambda probe (see p. 271), inserted in the exhaust stream ahead of the catalyst to measure the amount of oxygen in the exhaust. Each catalytic converter contains ~2 g of a PGM mixture, either platinum/rhodium or platinum/palladium/rhodium mixture in varying proportions.

Three-way catalysts are also used in thermal power stations.

Electrodes in fuel cells. Fuel cells convert chemical energy directly into electrical energy by catalytic oxidation of hydrogen. The battery-type devices consist of two catalyst-bearing electrodes, separated by an ion-conducting electrolyte. Hydrogen is oxidized at the anode, and oxygen is reduced at the cathode, generating electric current together with some reaction heat. Two fuel-cell types have been developed to practical use: the phosphoric acid fuel cell (PAFC) and the proton-exchange membrane (PEM) fuel cell. The electrodes on the anode side of the PAFC devices consist on the gas side of platinum on coal fiber and on the electrolyte side of a porous structured layer of platinum/carbon and PTFE. Cathode catalysts are Pt/Co/Cr, Pt/Co/Fe, or Pt/Ni alloys. In PEM-

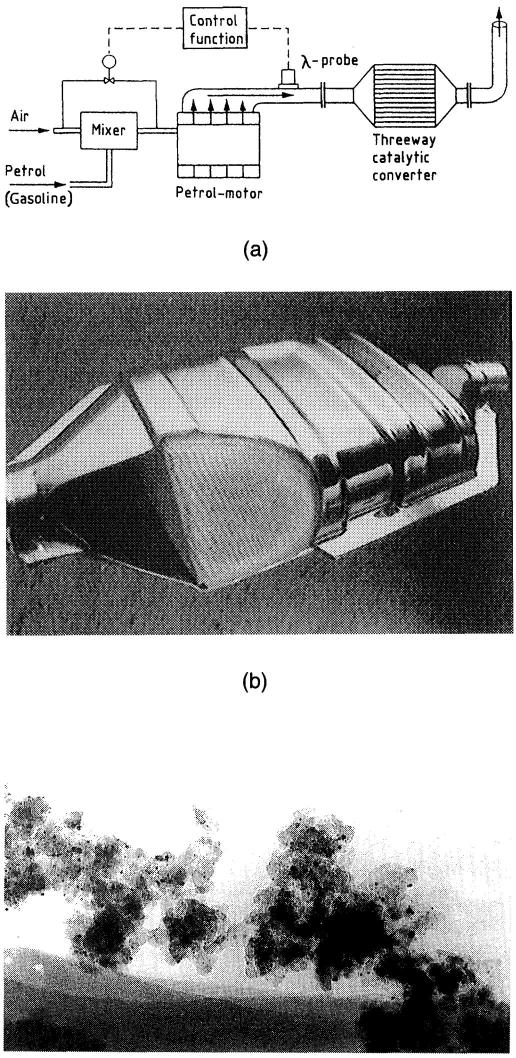

(c)

FIG. 29. Three-way catalytic converter: (a) schematic view (Prescher, 1995, p. 369), (b) cutaway view (Coombes, 1992), (c) Pt/Rh catalyst surface on the washcoat (Prescher, 1995, p. 366).

type fuel cells the electrodes are directly coated on both sides with porous layers of PTFE loaded with platinum/carbon on the cathode and platinum (+ ruthenium and palladium)/carbon on the anode side (Seibold, 1995).

Sensor devices: Thermocouples. Platinum and Pt–Rh alloys are used for thermocouples to be applied at high temperatures. Standard combinations are the following:

1. "S,"* PtRh10/Pt, up to 1600 °C;
2. "R,"* PtRh13/Pt or
3. "B,"* PtRh30/PtRh6, up to 1800 °C; and
4. IrRh40/Ir up to 2000 °C.

Measurement of very low temperatures can be made with the systems

5. AuFe(0.03 at. % Fe)-Chromel from 4.2 to 273 K,
6. AuCo(2.11 at. % Co)-AuAg(0.37 at. % Ag) or Cu from −240 to 0 °C, and
7. AuFe(0.02 at. % Fe)-Cu from −270 to −230 °C.

Thermocouples for heavily corrosive atmospheres are

8. AuPd40-PdPt14Au3, corresponding to NiCr–Ni;
9. AuPd35-PtPd12,5 corresponding to NiCr–Ni; and
10. AuPd46-PtIr10 corresponding to Fe-Constantan.

Sensors: Resistance thermometers. Platinum resistance thermometers are based on the temperature dependence of the electrical resistivity. The temperature coefficient of electrical resistance of platinum in high-purity condition is 3.927×10^{-3}/K. Basic values (= demanded values for the *R-T* relation) of platinum resistance thermometers are specified in international standards (DIN IEC 751). At temperatures below about 20 K the TCR of pure metals becomes too low in comparison to the absolute value of the resistance and cannot be measured. Alloys of rhodium with 0.5 at. % iron or alloys of gold with 1.15 at. % manganese show by magnetic effects an increasing TCR with decreasing temperature (Kondo effect; see p. 252) to sufficiently high, measurable values.

Film-type resistance thermometers contain thin platinum or iridium deposits of 1–2 μm thickness, deposited on flat ceramic substrates. The required resistance value is adjusted by laser-beam cutting. Iridium layers have values of thermal expansion that are comparable to those of the ceramic. They are therefore applied on devices for temperature measurements at lower temperatures down to ~200 °C.

---

*The types "S", "R", and "B" correspond to the international standard IEC 584. Chromel is Ni80Cr20. Constantan is Cu40Ni40.

Sensors: Special sensor devices. Platinum-equipped sensor devices are used to detect combustible gases in air or explosive gas mixtures. In the so-called pellistor devices, a platinum wire is placed in a small sphere of Al_2O_3 (diameter ~1.5 mm). The sphere is plated with platinum-black deposit that heats up by catalytically activated reaction of the gas composition in case of the presence of combustible constituents. The increase of temperature is measured by the electrical resistance of the platinum wire.

Platinum contact electrodes are constituents of sensor devices made from semiconducting materials (e.g., SnO_2, ZnO, piezoceramic, NTC resistors) or special ceramics. They serve for the determination of gas concentration, humidity, or pressure load. Oxygen-determining sensors ("lambda probes") use the high electrical conductivity imparted to solid electrolytes (e.g., ZrO_2) by the presence of oxygen ions, compared to their normal conductivity due to electrons. If this ionic conductor acts as a separator between two gas spaces, and if it is provided with platinum electrodes on both sides, a difference in the oxygen partial pressures causes the formation of an electromotive force ("Nernst emf") between the electrodes. Lambda probes are used to control the fuel-to-air ratio in conjunction with automobile exhaust-gas catalysts in Otto engines.

Magnetic storage. Thin-film cobalt-platinum alloys and thin-film cobalt-platinum multilayers play an important role as magnetic media on hard disks in magnetic and magneto-optic storage and recording devices. (see MAGNETIC STORAGE OF INFORMATION).

Electrical technology/electronics. Resistor-heated furnaces working in air at temperature >1200 °C are equipped with heaters made from platinum wires (up to 1600 °C), PtRh40 (up to 1700 °C), or rhodium bands (up to 1800 °C). Pt-Ni, Pt–Rh, and Pt–Ru alloys are used in spark-plug electrodes. Platinum finds applications in electronic devices as barrier layers, as contact pads on semiconductor chips, and as a stabilizing component against silver-ion migration in silver-based metallizing preparations and thick-film conductor pastes for screen printing of hybrid circuits. Platinum resinates are used for the production of conductive coatings on ceramic substrates onto which further layers can be deposited electrolytically.

Chemical technology. Platinum crucibles, dishes, boats, and electrodes for electrogravimetry are basic tools in the laboratory technology. If dimensional stability is required, components are made from the alloys PtIr3, PtAu5, and PtIr0.3.

Platinum coatings serve as corrosion protection on reaction equipment. Porous filters of platinum, palladium, or silver for corrosive media are produced by sintering of spherolithic powders. High-purity optical glass is melted in pure platinum crucibles. Platinum components (skimmers, pouring funnels, protection tubes for thermocouples, level-measuring devices) and dispersion-hardened platinum are used for equipment parts in the glass-working technology. Dispersion-strengthened platinum–rhodium alloys serve as heater elements in high-temperature furnaces. Platinum centrifuges are used for the production of rock wool and slag wool. Spinnerets for the manufacturing of glass fiber and glass wool are made of Pt–Rh alloys. Spinnerets for textile fibers like rayon or acrylic are made from platinum–gold and platinum–rhodium alloys. Platinum–rhodium tubes protect high-temperature thermocouples against oxidation. Platinum-plated wires of iron–nickel or molybdenum, or gold-plated iron–nickel wires, serve as vacuum-tight leads through glass tubes. Platinum and iridium are indispensible crucible materials for the production of high-purity single crystals for optical and laser technology.

Biomedical uses. The stereoisomer dichloro-diammino Pt^{II}-complex compound ("*cis*-platinum") has antineoplastic effects by suppressing the DNA synthesis but not the synthesis of RNA or proteins. This compound and less toxic successor products ("Carboplatin," "Paraplatin") have been proven as highly effective agents against cancer (Rosenberg *et al.*, 1965, 1969). Platinum and platinum–iridium alloys find use for electrodes in pacemakers, in catheters, as marker bands, as guide wires, and as temporary and permanent implants. Irradiated Ir wire, sheated in Pt, can be implanted in the body to deliver controlled doses of radiation for cancer therapy. The radio-opacity of the platinum shields the healthy tissues from radiation, while the tumor is irradiated by the exposed iridium tip. Platinum coils made from thin wire are applied for treatment of aneurysm in the brain, where traditional recourse of surgery is most difficult (Coombes, 1993).

4.2.2 Palladium Palladium, atomic number 46, relative atomic mass 106.42, six natural isotopes (between 102 and 110), is a silvery white, bright metal, malleable and ductile. It is soluble in oxidizing acids and molten alkali. The main application areas are electronics (46%), dentistry (26%), automotive catalytic converters (17%), jewelry and others (10%), and chemistry (7%) (quantities in weight percent of the total yearly demand in the western world 1995; Johnson Matthey, 1995).

Selected values of basic physical and chemical properties are given in Table 14.

4.2.2.1 Influence of Alloying Elements. *Metals: Strengthening.* Figure 30 shows the increase of hardness with increasing amounts of alloy additions. Grain refining of palladium alloys is performed with additions of ruthenium and germanium. Palladium and copper form superlattices in the composition ranges between 10 and 25 wt % palladium (tetragonal $L1_0$ structure) and between 30 and 50 wt % palladium (B2 structure).

Metals: Electrical resistance and magnetic behavior. Figure 31 shows the influence of alloying elements on the electrical resistance.

Palladium alloys with iron and with cobalt show ferromagnetism. On the alloying of palladium with 0.1 wt % cobalt, the palladium atoms surrounding a cobalt atom are strongly magnetized to magnetic moments on the order of 10 Bohr magnetons (M_B), which is ~6 times that of a single cobalt moment.

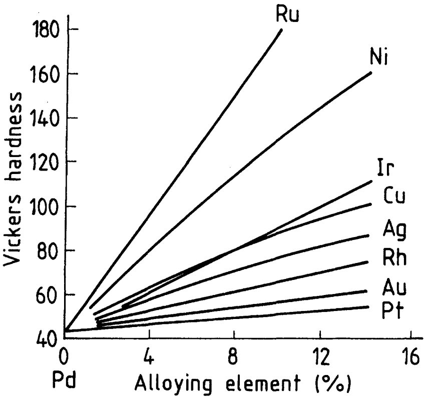

FIG. 30. Influence of alloying elements on the hardening of palladium (Wise and Vinnes, 1979, p. 700).

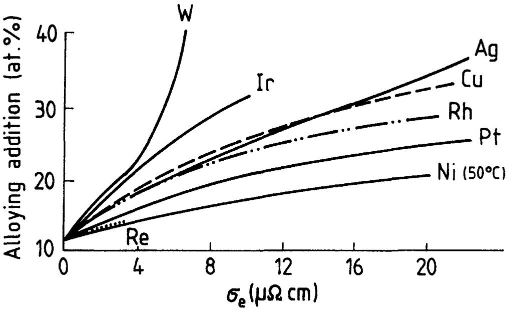

FIG. 31. Influence of alloying elements on the electrical conductivity of palladium (Doduco, 1977, p. 67).

Palladium dissolves in diamagnetic copper, forming diamagnetic alloy phases in the composition ranges up to 50 at. % palladium. The partial dissociation of the metallic palladium according to

$$\mathrm{Pd_{Dia}} \leftarrow \mathrm{Pd^{+}_{Para}} + e^{-}, \tag{11}$$

which is responsible for its paramagnetism, is suppressed by the high electron-gas concentration of the metallic copper. The ordered phases in this system have higher diamagnetic susceptibilities than the disordered solid-solution phase. Only disordered Cu–Pd alloys with palladium contents above 50 at. % palladium are paramagnetic. Similar behavior is found when hydrogen is dissolved in palladium, where the paramagnetism of palladium decreases to reach zero at the composition $PdH_{0.66}$ and above, leaving a pure diamagnetic palladium–hydrogen alloy.

Hydrogen. Palladium and certain palladium alloys are unique in their capability to dissolve large amounts of hydrogen, up to ~2800 times their own volume (Moore, 1939; Lewis, 1967, 1994). Hydrogen enters palladium interstitially. The hydrogen atoms are located in the octahedral holes of the fcc lattice. Above 295 °C, palladium and hydrogen form a continuous series of solid solutions with fcc structure. Below that temperature, the phase splits into a fcc palladium-rich α phase and also a fcc hydrogen-rich β phase, forming a miscibility gap that broadens with decreasing temperature. The equilibrium hydrogen pressure at 295 °C amounts to 19.87 atm.; the composition on the critical point corresponds to palladium with 21 at. % hydrogen. The β phase takes up hydrogen up to 1300 times the volume of the palladium, corresponding to a composition of palladium with 50 at. % hydrogen. Further quantities of hydrogen up to a total of 2800 times the volume of palladium can be loaded by cathodic deposition. The two-phase region shows, according to the phase rule, constant vapor pressures at constant temperatures, which amounts to 1 atmosphere at ~140 °C. The occupation of interstitial sites by hydrogen atoms causes expansion of the palladium lattice. The lattice parameters increase with increasing hydrogen content from 3.891 Å for the pure palladium up to 4.06 Å at 75 at. % hydrogen (Nelin, 1971). Thermal cycling of Pd–H alloys in the duplex phase region causes high mechanical stresses due to changes of the lattice dimensions with changing amounts of dissolved hydrogen. Pure palladium, if repeatedly heated and cooled in hydrogen-bearing atmosphere, becomes brittle and cracks. Palladium–silver alloys with 20–25 wt % silver dissolve higher amounts of hydrogen than pure palladium and resist thermal cycling (Lewis, 1982). Ordered A_3B phases of Pd–Mn and Pd–Fe alloys show considerably higher solubility for hydrogen than the corresponding disordered phases. In Pd–Mn alloys, hydrogen supports the ordering process to proceed at lower temperatures (Flanagan and Sakamoto, 1993).

4.2.2.2 Alloys. PdAg23 exhibits extremely high hydrogen permeability and is used for semipermeable membranes for the extraction and purification of hydrogen. The Pd–Ag alloy with 40 at. % Ag has the highest electrical resistivity at room temperature. It finds use for precision resistors in the electrical technology. The specific resistance between 20 and 100 °C is nearly constant and amounts to 42.5 Ω·cm; the specific resistance between ~20 and 800 °C changes from 42.5 to 44 Ω·cm.

Palladium–copper (15–40 wt %) alloys serve as surface material for electrical contacts in the milliampere range and as hard, wear-resistant material for slip rings with high conductivity. Pd–Ag–Cu alloys containing ~45% Pd can be age hardened. These alloys find application in dentistry. Addition of 1% Zn accelerates age hardening. Ternary alloys containing 10 to 25% Pd are used as hard solders for the formation of high-strength brazed joints.

Palladium–gold alloys are the basis of precision resistance alloys. Addition of iron accentuates the formation of long-range ordering. The alloy Au50Pd45Mo5 has the specific resistance of 100 $\mu\Omega\cdot$cm with a temperature coefficient of resistance (TCR) of 0.00012/°C between 0 and 100 °C. The tensile strength reaches values of ~1060 N/mm^2.

Palladium–ruthenium with 4.5 wt % ruthenium is a standard jewelry alloy in the United States. Alloys with ruthenium contents up to 12% serve as material for electrical contacts and as catalysts for the reduction of nitric oxide.

4.2.2.3 Applications. *Electronics:* Palladium is, pure or in combination with silver, the most important component in capacitor electrode layers, in silver conductor paths, and in resistor elements in thick-film circuits.

Dentistry. Palladium-based alloys (palladium content from about 60–80 wt % Pd) contain copper and gallium as hardening elements and tin to lower the melting temperature. The best corrosion resistance is attained with a single-phase structure in the composition range of Cu 4–6 wt %, Ga 5–7 wt %, and Sn 6 wt %.

Chemistry. Hydrogen purification equipment based on the alloy of the composition palladium with 23 wt % silver is used to produce ultrapure hydrogen up to 99.999 99% purity. Palladium-silver permeation membranes are provided in the form of bundled tubes with wall thickness of ~75 μm, sealed on one end. Hydrogen is pressed from outside through the tubes, dissociating and diffusing through the alloy to the inner side of the tubes, leaving most impurities outside the membrane (Hunter, 1960; Knapton, 1977).

The solubilities of the hydrogen isotopes in palladium and palladium–silver alloys differ to an appreciable degree. Diffusion membranes of palladium or palladium–silver alloys can therefore be used for separation and concentration processes of the different hydrogen isotopes. The equilibrium hydrogen/deuterium separation factors vary from ~2.4 at 25 °C to 3.7 at −80 °C (Wicke and Nernst, 1964). For the hydrogen/tritium ratio, a value of 2.8 has been measured at 25 °C (Sicking, 1970).

4.2.3 Rhodium Rhodium, atomic number 45, relative atomic mass 102.905 50, no second natural isotope, is very rare. It is a bright, silvery, and hard metal, oxidation resistant up to 875 K, inert against all acids but attacked by molten alkali. The most important properties are catalytic action in heterogeneous and homogeneous catalysis. The main application areas are automotive catalytic converters (89%), chemistry (3%), electronics/electrical technology (1.8%), and glass (3.3%) (quantities in weight percent of the total yearly demand in the western world 1995; Johnson Matthey, 1995).

Selected values of basic physical and chemical properties are given in Table 14. The characteristic fall of the optical reflectivity in the visible range (Fig. 2) is smallest for rhodium. It has a neutral color; the reflectivity is ~20% lower than that of silver.

4.2.3.1 Applications. *Catalysts.* Rhodium is an important constituent in automotive catalysts and in numerous heterogeneous and homogeneous catalytic reactions. In automotive catalysts, rhodium effects the reduction of nitrogen oxides to nitrogen by intermediate adsorption of NO and CO on the surface followed by dissociation of the NO molecule and reaction of the desorbed oxygen atom with CO to form CO_2 (Madix, 1986; Friend, 1993; Cho, 1994).

Chemical technology. Because of their ability to withstand highly corrosive operating environments, platinum–rhodium alloys find extended use for equipment and apparatus parts in the glass industry, especially for the production of high-quality glass for optical devices, liquid-crystal–display glasses, and screens for television sets and desk-top computers.

Platinum-rhodium alloys find application in net catalysts for the oxidation reaction of ammonia to nitrogen oxide. Complex organic rhodium compounds are of great importance as catalysts for hydroformylation (Osborne *et al.*, 1965; Coffey, 1966) and for the production of enantiomerically pure compounds by stereospecific chemical reactions. Rhodium and iridium are suited as susceptors for medium- and high-frequency fields for inductive heating on air. Iridium crucibles used for induction melting of glasses or ceramics are protected against oxidation by rhodium coatings.

Electrical technology/electronics. Rhodium serves as a hard, wear-resistant, electrically conductive surface coating for slide and pressure contacts. It is a constituent in the most commonly used thermocouples in various compositions (e.g., PtRh10/Pt, PtRh13/Pt, PtRh30/PtRh6) applied for temperature measurements in the medium- and high-temperature range.

Jewelry, mirrors. Silver mirrors and jewelry items made from silver and white-gold alloys are protected against corrosion and wear by thin coatings of rhodium.

4.2.4 Iridium Iridium, atomic number 77, relative atomic mass 192.22, only two natural isotopes (191 and 193), is a bright silvery, hard, high–melting-point metal. Together with rhodium, it belongs to the rarest members of the platinum metals. This is in accordance with the general rule that isotopes with odd atomic numbers are rarer than those with even atomic numbers that are divisible by 4. Iridium is inert against all acids, air, and water, but attacked by molten alkali. The main application areas are chemistry and catalysis (58%), crucibles (5%), and others (37%) (quantities in weight percent of the total yearly demand in the western world 1995; Johnson Matthey, 1995).

Selective values of basic physical and chemical properties are given in Table 14. Iridium has the highest density of all metals. It is an effective hardener and grain-refining agent in PGM alloys.

4.2.4.1 Applications. *Chemistry.* In chlor-alkali plants and in electrolytic cells for the production of sodium chlorate, iridium and iridium–ruthenium are applied as surface coatings on anodes. Iridium crucibles serve for the production of high-purity single crystals for optical and laser technology—e.g., gadolinium–gallium garnets and monocrystalline Al_2O_3 substrates. Special catalytic actions of iridium find application for reforming catalysts in the petroleum refining industry and in the synthesis of pharmaceuticals. Quartz glass fibers are drawn through iridium spinnerets.

Electrical technology. Iridium coatings are used as sensors in layer resistor devices. Thermocouples made from combinations of iridium with iridium–rhodium 10 alloy are suited for temperature measurement at very high temperatures up to 2300 °C. Spark-plug electrodes tipped with iridium-containing alloys are used on helicopters and in heavy-duty automobile spark plugs.

4.2.5 Ruthenium Ruthenium, atomic number 44, relative atomic mass 101.07, seven natural isotopes (between 96 and 104), is a bright, silvery metal, resistant against air, water, and acids, but soluble in molten alkali. The main application areas are electronics/electrical technology (45%), chemistry (45%), and catalysis (10%) (quantities in weight percent of the total yearly demand in the western world 1995; Johnson Matthey, 1995).

Selective basic physical and chemical properties are given in Table 14. Ruthenium is, like iridium, an effective hardener and grain-refining agent for PGM alloys.

4.2.5.1 Applications. *Electronics.* Ruthenium is the most important base element of resistor devices and resistor layers in thick-film circuits. The starting material is RuO_2. Resistor layers are printed with paste preparations made from RuO_2 with additions of PbO and Bi_2O_3. They form by firing complex Ru compounds of high thermal stability and very low temperature coefficients of electrical resistivity. The measurement unit is the so-called "sheet resistivity," defined as the electrical resistance of a square area, measured between opposite edges. Exact resistance values are adjusted by applying small cuts by means of laser beam into the resistor layer ("laser trimming").

Electrochemistry. Ruthenium-coated titanium electrodes are used in the electrolysis of brine to produce chlorine and caustic soda. They lower the overpotential against chlorine, saving energy and giving controlled composition of the chlorine gas with low oxygen concentration.

Catalysis. RuO_2 and complex organic ruthenium compounds are specific catalysts in numerous heterogeneous and homogeneous catalytic reactions. Metal-organic ruthenium compounds are proven as catalysts for the photodissociation of water (Watkins, 1978; Harriman, 1983, 1991).

4.2.6 Osmium Osmium, atomic number 76, relative atomic mass 190.23, seven natural isotopes (between 184 and 192), is a bright, silvery metal, resistant against water, oxygen, and acids, but attacked by molten alkali. Application areas are catalysis, alloy components in hard precious-metal alloys, and electronics.

Selected values of basic physical and chemical properties are given in Table 14. Osmium and iridium are efficient reflectors in the wavelength range below 100 nm (Hass and Hunter, 1974).

4.2.6.1 Applications. Osmium is the least practical used noble metal. There are only a few applications.

Chemistry. Osmium tetroxide is used as a catalyst in special oxidation reactions. Complex organic osmium compounds are used in homogeneous catalytic reactions, especially for drug synthesis.

Optics. Osmium and iridium have been used as mirror coatings for vacuum-UV spectroscopy to study celestial phenomena that cannot be found with Al-based mirrors. Ir-coated mirrors have been used in large-aperture telescopes of space rockets for exploration tasks in enhanced space astronomy (Herzig and Spencer, 1982; Herzig, 1983).

Electronics. Surface layers of osmium and ruthenium on dispenser cathodes lower the electronic work function (Zalm and Van Stratum, 1966). Thin osmium layers are applied on tungsten and molybdenum wires for *thermionic sources.*

5. SUMMARY

Silver, gold, and the platinum metals have the highest oxidation resistance of all metals. They are construction materials in the chemical technology and used as conductor, resistor, and contact materials. The platinum metals and platinum-metal compounds are effective and highly selective catalysts in numerous chemical reactions and in automotive exhaust-gas converters. Platinum and platinum alloys are used in sensor elements and thermocouples. The noble metals are also key materials in the fields of magnetic recording; as energy-saving coatings on architectural glass panes; and as electrodes in fuel cells, electrolytic production equipment, and batteries. New technologies open promising application areas. Progress in the automobile engine technology and the growing demand for multilayer capacitors in electronic devices have caused 35% growth of the palladium demand from 1993 to 1994. The demand for ruthenium grew in the same year by 76%, mainly on account of applications in catalysis, electrodes in the chlor-alkali electrolysis, and alloy improvement. The demand for iridium grew by 36% by new catalytic activities, electrodes, and the need for hard, corrosion-resistant alloys—e.g., for spark plugs in aircraft engines. Because of the limited reserves, the production from lower-grade raw material and the recycling of secondary materials have become of increasing importance.

ACRONYMS

APB:	Antiphase boundary
CVD:	Chemical vapor deposition
DNA:	Deoxyribonucleic acid
DOS:	Density of states
fcc:	Face centered cubic
hcp:	Hexagonal close packed
HV:	Vickers hardness
NTC:	Negative temperature coefficient
PFM:	Porcelain fused to metal
PGM:	Platinum-group metal
PTFE:	Polytetrafluorethylene
PVD:	Physical vapor deposition
RNA:	Ribonucleic acid
ZGS:	Zirconium-oxide grain stabilized

Works Cited

Alekseev, E. S., Ventsel, V. A., Voronov, O. A., Likhter, A. I., Magnitskaya, N. V. (1979), *Zh. Eksp. Teor. Fiz.* **76,** 215–222 [*Sov. Phys. JETP* **49,** 110–113].

Amirnazmi, A., Boudart, M. (1975), *J. Catal.* **39,** 383–394.

Benner, L. S., Suzuki, T., Meguro, K., Tanaka, S. (Eds.) (1991), *Precious Metals, Science and Technology,* Allentown, PA: International Precious Metals Institute.

Berkowitz, A., Kneller, E. (1969), *Magnetism and Metallurgy,* New York: Academic, p. 259.

Beuers, J., Krappitz, H. (1995), in: Degussa (Ed.), *Edelmetalltaschenbuch,* Heidelberg: Hüthig.

Broden, G., Rhodin, T. N., Brucker, C. (1976), *Surf. Sci.* **59,** 593–611.

Butts, A., Coxe, C. D. (1967), *Silver,* Princeton, NJ: Van Nostrand, p. 111.

Chikazumi, S. (1991), in: L. S. Benner, T. Suzuki, K. Meguro, S. Tanaka (Eds.), *Precious Metals, Science and Technology,* Allentown, PA: International Precious Metals Institute, p. 98.

Cho, B. K. (1994), *J. Catal.* **148,** 697–706.

Christensen, N. E. (1972), *Phys. Status Solidi B* **54,** 551–563.

Christensen, N. E. (1973), *Phys. Status Solidi B* **55,** 117–127.

Christensen, N. E., Seraphin, B. O. (1971), *Phys. Rev. B* **4,** 3321–3344.

Coffey, R. S. (1966), British Patent No. 1,121,642.

Coleridge, P. T. (1966a), *Proc. R. Soc. A* **295,** 458–475.

Coleridge, P. T. (1966b), *Phys. Lett.* **22,** 367–369.

Coombes, J. S. (1992), in: Johnson Matthey Co., *Platinum 1992 (Annual Report),* London: Johnson Matthey Publishers, pp. 37–43.

Coombes, J. S. (1993), in: Johnson Matthey Co., *Platinum 1993 (Annual Report),* London: Johnson Matthey Publishers, pp. 35–38.

Dachselt, W. D., Münz, W. D., Scherer, M. (1982), in: Carl M. Lampert (Ed.), *Optical Coatings for Energy Efficiency,* SPIE Proceedings No. 324, Bellingham, WA: SPIE, pp. 37–43.

Darling, A. S. (1973), *Inst. Met. Rev.* **175,** 91–122.

Degussa (1967), *Edelmetalltaschenbuch 1973,* Frankfurt/Main: Degussa.

Degussa (1994), *Edelmetallmärkte 1994 (Annual Report),* Frankfurt/Main: Degussa A. G.

Degussa (1995), *Edelmetalltaschenbuch,* Heidelberg: Hüthig, cover pages.

Diehl, W. (1995), in: Degussa (Ed.), *Edelmetalltaschenbuch 1995,* Heidelberg: Alfred Hüthig, pp. 155–170.

Doduco (Ed.) (1977), *Doduco Datenbuch,* Pforzheim: Doduco, pp. 19–80.

Dorbath, B., Frischkorn, H. J., Jeske, G., Starz, K. A. (1991), in: Degussa Ressort Metall (Ed.), *Metall Forschung und Entwicklung,* Frankfurt/Main: Degussa A. G., pp. 47–58.

Fieberg, M., Edwards, I. R. (1978), Report No. 1996, National Institut van Metall, Randburg, South Africa.

Flanagan, T. B., Sakamoto, Y. (1993), *Plat. Met. Rev.* **37,** 26–37.

Flükiger, R., Paoli, A., Roggen, R., Yvou, K., Muller, K. J. (1972), *Solid State Commun.* **11,** 61–63.

Fong, C. Y. (1974), *J. Phys. F: Metal Phys.* **4,** 775–782.

Friend, C. M. (1993), *Sci. Am.* **268**(4), 42–47.

Funston, E. S., McGurty, J. A. (1961), *Plat. Met. Rev.* **5,** 56.

Gafner, G. (1989), *Gold Bull.* **22,** 112–122.

Gläser, H. H. (1989), *Glastech. Ber.* **62,** (3), 93–99.

Grodski, J. J., Dixon, A. E. (1969), *Solid State Commun.* **7,** 735–737.

Haasen, P. (1994), *Physikalische Metallkunde,* 3. Aufl., Berlin: Springer.

Hagen, E., Rubens, H. (1903), *Ann. Phys. (Leipzig)* **11,** 873–901.

Harriman, A. (1983), *Plat. Met. Rev.* **27,** 102–107.

Harriman, A. (1991), *Plat. Met. Rev.* **35,** 22–23.

Hass, G., Hunter, W. R. (1974), in: B. J. Thompson, R. R. Shannon (Eds.), *Proceedings of the Ninth International Congress of the International Committee for Optics, Santa Monica, California, October 1972,* Washington, D.C.: National Academy of Sciences, pp. 525–553.

Herzig, H. (1983), *Plat. Met. Rev.* **27,** 108–109.

Herzig, H., Spencer, R. S. (1982), *Appl. Opt.* **21,** 15–16.

Hirabayashi, M. (1991), in: L. S. Benner, T. Suzuki, K. Meguro, S. Tanaka (Eds.), *Precious Metals, Science and Technology,* Allentown, PA: International Precious Metals Institute, pp. 103–115.

Hornfeldt, S. P., Windmiller, L. R., Ketterson, J. B. (1973), *Phys. Rev. B* **7,** 4349–4357.

Hume-Rothery, W., Raynor, G. V. (1956), *The Structure of Metals and Alloys,* London: The Institute of Metals.

Hummel, R. E. (1971), *Optische Eigenschaften von Metallen und Legierungen,* Berlin: Springer, p. 161.

Hunt, L. B. (1979), *Gold Bull.* **12,** 116–127.

Hunter, J. B. (1960), *Plat. Met. Rev.* **4,** 130–131.

Jepsen, O., Andersen, O. K., Mackintosh, A. R. (1975), *Phys. Rev. B* **12,** 3084–3103.

Johnson Matthey Co. (1992), *Platinum 1992 (Annual Report),* London: Johnson Matthey Publishing.

Johnson Matthey Co. (1995), *Platinum 1995 (Annual Report),* London: Johnson Matthey Publishing.

Jönsson, S. (1995), in: Degussa (Ed.), *Edelmetalltaschenbuch,* Heidelberg: Hüthig, pp. 202–223.

Kempf, B., Hausselt, J. (1992), *Interdiscip. Sci. Rev.* **17,** 251–259.

Ketterson, J. B., Windmiller, L. R. (1970), *Phys. Rev. B* **2,** 4813–4838.

Khan, H. R. (1984), *Gold Bull.* **17,** 94–100.

Kittel, C. (1980), *Einführung in die Festkörperphysik,* München: R. Oldenbourg.

Knapton, A. G. (1977), *Plat. Met. Rev.* **21,** 44–50.

Kondo, J. (1964), *Prog. Theor. Phys.* **32,** 37–49.

Kondo, J. (1965), *Prog. Theor. Phys.* **34,** 372–382.

Kopp, J. (1976), *Gold Bull.* **9,** 55–57.

Kussman, A., Jessen, X. (1962), *J. Phys. Soc. Jpn.* **17,** (B1), 136–139.

Kussman, A., von Rittberg, G. (1950), *Ann. Phys. (Leipzig)* **7,** 173–181.

Lewis, F. A. (1967), *The Palladium-Hydrogen System,* New York: Academic.

Lewis, F. A. (1982), *Plat. Met. Rev.* **26,** Pt. I, 70–78; Pt. II, 20–27; Pt. III, 121–128.

Lewis, F. A. (1994), *Plat. Met. Rev.* **38,** 112–118.

Linde, J. O. (1932), *Ann. Phys. (Leipzig)* **4,** 219–248, 353–366.

Loebich, O. (1971), Doctoral Thesis, Technische Hochschule Aachen.

Madix, R. J. (1986), *Science* **233,** 1159–1166.

McKenzie, D. R. (1978), *Gold Bull.* **11,** 49–53.

Moore, C. A. (1939), *Trans. Electrochem. Soc.* **75,** 237–269.

Mueller, F. M., Freeman, A. J., Dimmock, J. O., Fardgna, A. M. (1970), *Phys. Rev. B* **1,** 4617–4635.

Nelin, G. (1971), *Phys. Status Solidi* **45,** 527–536.

Obuchi, A., Naito, S., Onishi, T., Tamaru, K. (1982), *Surf. Sci.* **122,** 235–255.

Obuchi, A., Naito, S., Onishi, T., Tamaru, K. (1983), *Surf. Sci.* **130,** 29–40.

Osborne, J. A., Wilkinson, G., Young, J. F. (1965), *J. Chem. Soc. Chem. Commun.* 131–132.

Ott, D., Raub, Ch. J. (1980), *Metall.* **34,** 629–633.

Ott, D., Raub, Ch. J. (1981), *Metall.* **35,** 543–548, 1005–1011.

Ott, D., Raub, Ch. J. (1982), *Metall.* **36,** 150–156.

Plachy, A. L., Dixon, A. E. (1974), *Can. J. Phys.* **52,** 1295–1303.

Poniatowski, M., Clasing, M. (1968), *Z. Metallk.* **59,** 165–170.

Prescher, N. N. (1995), in: Degussa (Ed.), *Edelmetalltaschenbuch 1995,* Heidelberg: Alfred Hüthig, pp. 351–401.

Pugh, S. F. (1954), *Philos. Mag.* **45,** 823–843.

Raub, Ch. J. (1984), *Plat. Met. Rev.* **28,** 63–75.

Renner, E. (1990), in: W. Gerhartz (Ed.), *Ullmann's Encyclopedia of Industrial Chemistry,* Weinheim: VCH, Vol. A12, pp. 499–533.

Renner, E. (1992), in: W. Gerhartz (Ed.), *Ullmann's Encyclopedia of Industrial Chemistry,* Weinheim: VCH, Vol. A21, pp. 75–127.

Renner, E. (1993), in: W. Gerhartz (Ed.), *Ullmann's Encyclopedia of Industrial Chemistry,* Weinheim: VCH, Vol. A24, pp. 107–163.

Rogers, D. L. (1974), *Phys. Status Solidi B* **66,** K53–K57.

Rosenberg, B., Van Kamp, L., Krigius, T. (1965), *Nature* **203,** 698–699.

Rosenberg, B., Trosko, J. E., Mansour, V. H. (1969), *Nature* **222,** 385–386.

Schuit, G. C. A., van Reijen, L. L. (1958), *Adv. Catal.* **10,** 242–317.

Seibold, K. (1995), in: Degussa (Ed.), *Edelmetalltaschenbuch,* Heidelberg: Hüthig, pp. 414–421.

Selman, G. L., Day, J. G., Bourne, A. A. (1974), *Plat. Met. Rev.* **18,** 46–57.

Sicking, G. (1970), Ph.D. Thesis, Univ. Münster, Germany.

Smith, N. V. (1974), *Phys. Rev. B* **9,** 1365–1376.

Smithells, C. J., Brandes, E. A. (1977), *Metals Reference Book,* 5th ed., London: Butterworth, p. 1019.

Stöckel, D. (1979), *Z. Werkstofftech.* **10,** 230–242.

Tamaru, K., Naito, S. (1991), in: L. S. Benner, T. Suzuki, K. Meguro, S. Tanaka (Eds.), *Precious Metals, Science and Technology,* Allentown, PA: International Precious Metals Institute, pp. 262–307.

Tanuma, S. (1991), in: L. S. Benner, T. Suzuki, K. Meguro, S. Tanaka (Eds.), *Precious Metals, Science and Technology,* Allentown, PA: International Precious Metals Institute, pp. 35–59.

Toyoshima, I., Somorjai, G. A. (1979), *Cat. Rev. Sci. Eng.* **19,** 105–159.

Vannice, M. A. (1975), *J. Catal.* **37,** 449–473.

Vedernikov, M. V. (1969), *Adv. Phys.* **18,** 337–370.

Völcker, A. (1995), in: Degussa (Ed.), *Edelmetalltaschenbuch 1995,* Heidelberg: Alfred Hüthig, pp. 170–173.

Watkins, D. M. (1978), *Plat. Met. Rev.* **22,** 118–125.

Wicke, E., Nernst, G. H. (1964), *Ber. Bunsenges.* **1964,** 224–235.

Wise, E. M. (1964), *Gold,* Princeton, NJ: Van Nostrand, p. 72.

Wise, E. M., Vines, R. F. (1979), in: W. H. Cubberly, H. Baker, D. Benjamin (Eds.), *Metals Handbook,* Vol. 2, *Properties and Selection: Nonferrous Alloys and Pure Metals,* 9th ed., Metals Park, OH: American Society for Metals, pp. 699–705.

Yonemitsu, K. (1991), in: L. S. Benner, T. Suzuki, K. Meguro, S. Tanaka (Eds.), *Precious Metals, Science and Technology,* Allentown, PA: International Precious Metals Institute, pp. 13–35.

Zalm, P., Van Stratum, J. A. (1966), *Philips Tech. Rev.* **27** (3/4), 69–75.

Zysk, E. D. (1979), in: W. H. Cubberly, H. Baker, D. Benjamin (Eds.), *Metals Handbook,* Vol. 2, *Properties and Selection: Nonferrous Alloys and Pure Metals,* 9th ed., Metals Park, OH: American Society for Metals, pp. 688–698.

Further Reading

Bard, J. A. (1979), "Properties of Gold and Gold Alloys," in: W. H. Cubberly, H. Baker, D. Benjamin (Eds.), *Metals Handbook,* Vol. 2, *Properties and Selection: Nonferrous Alloys and Pure Metals,* 9th ed., Metals Park, OH: American Society for Metals, pp. 679–683.

Benner, L. S., Suzuki, T., Meguro, K., Tanaka, S. (Eds.) (1991), *Precious Metals, Science and Technology,* Allentown, PA: International Precious Metals Institute.

Butts, A., Coxe, Ch. D. (1967), *Silver, Economics, Metallurgy, and Use,* Princeton, NJ: Van Nostrand.

Coxe, C. D., McDonald, A. W. (1979), "Properties of Silver and Silver Alloys," in: W. H. Cubberly, H. Baker, D. Benjamin (Eds.), *Metals Handbook,* Vol. 2, *Properties and Selection: Nonferrous Alloys and Pure Metals,* 9th ed., Metals Park, OH: American Society for Metals, pp. 671–678.

Darling, A. S. (1973), *Some Properties and Applications of the Platinum Group Metals,* London: The Institute of Metals, pp. 91–122.

Degussa (1967), *Edelmetalltaschenbuch,* Frankfurt/Main: Degussa.

Degussa (1995), *Edelmetalltaschenbuch,* 2nd ed., Heidelberg: Alfred Hüthig.

Degussa (Ed.) (1991), *Metall Forschung und Entwicklung,* Frankfurt/Main: Degussa Ressort Metall.

Degussa (annual reports), *Edelmetallmärkte,* Frankfurt/Main: Degussa.

Emsley, J. (1991), *The Elements,* 2nd ed., Oxford: Clarendon.

Haasen, P. (1994), *Physikalische Metallkunde,* 3. Aufl., Berlin: Springer.

Hume-Rothery, W., Raynor, G. V. (1956), *The Structure of Metals and Alloys,* London: The Institute of Metals.

Hummel, R. E. (1971), *Optische Eigenschaften von Metallen und Legierungen,* Berlin: Springer.

Johnson Matthey (annual reports), *Platinum,* London: Johnson Matthey Publishing.

Kittel, C. (1980), *Einführung in die Festkörperphysik,* authorized translation of *Introduction to Solid State Physics,* 5th ed., München: R. Oldenbourg.

Kondo, J. (1969), "Theory of dilute magnetic alloys," in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics,* Vol. 23, New York: Academic, pp. 184–281.

Raub, E. (1940), *Die Edelmetalle und ihre Legierungen,* Berlin: Springer.

Raynor, G. V. (1976), "The Alloying Behavior of Gold," *Gold Bull.* **9,** 12–18, 50–55.

Renner, H. (1992), "Platinum Group Metals and Compounds," in: W. Gerhartz (Ed.), *Ullmann's Encyclopedia of Industrial Chemistry,* Vol. A21, Weinheim: VCH, pp. 75–131.

Renner, H. (1993), "Silver, Silver Compounds, and Silver Alloys," in: W. Gerhartz (Ed.), *Ullmann's Encyclopedia of Industrial Chemistry,* Vol. A24, Weinheim: VCH, pp. 107–163.

Renner, H., Johns, M. W. (1990), "Gold, Gold Alloys, and Gold Compounds," in: W. Gerhartz (Ed.), *Ullmann's Encyclopedia of Industrial Chemistry,* Vol. A12, Weinheim: VCH, pp. 499–533.

Savitskii, E. M., Prince, A. (1989), *Handbook of Precious Metals,* New York: Hemisphere.

Schlamp, G. (1995), "Noble Metals and Their Alloys," in: R. Cahn (Ed.), *Materials Science and Technology,* Vol. 8/9 (K. Matucha, Ed.), *Structure and Properties of Nonferrous Metals,* Weinheim: VCH, pp. 467–587.

Wise, E. M., Vines, R. F. (1979), "Properties of Palladium and Palladium Alloys," in: W. H. Cubberly, H. Baker, D. Benjamin (Eds.), *Metals Handbook,* Vol. 2, *Properties and Selection: Nonferrous Alloys and Pure Metals,* 9th ed., Metals Park, OH: American Society for Metals, pp. 699–705.

Zysk, E. D. (1979), "Precious Metals and Their Uses," in: W. H. Cubberly, H. Baker, D. Benjamin (Eds.), *Metals Handbook,* Vol. 2, *Properties and Selection: Nonferrous Alloys and Pure Metals,* 9th ed., Metals Park, OH: American Society for Metals, p. 659–657.

Zysk, E. D. (1979), "Properties of Platinum and Platinum Alloys," in: W. H. Cubberly, H. Baker, D. Benjamin (Eds.), *Metals Handbook,* Vol. 2, *Properties and Selection: Nonferrous Alloys and Pure Metals,* 9th ed., Metals Park, OH: American Society for Metals, pp. 688–698.

SIMULATION BY MOLECULAR DYNAMICS

Berni J. Alder and Anthony J. Ladd, *University of California, Lawrence Livermore National Laboratory, Livermore, California, U.S.A.*

Introduction 281
1. **Time Considerations** 282
1.1 The Fermi–Pasta–Ulam Problem 282
1.2 Hard Particles 282
1.3 The *H* Theorem 282
1.4 Irreversibility 282
1.5 Numerical Noise 283
1.6 Computational Limitations 284
1.7 Length of Trajectory 284
1.8 Rare Events 285
2. **Equilibrium Properties** 285
2.1 Phase Transitions 286
2.2 Ensembles 286
2.3 van der Waals Approximation 287
3. **Complex Molecules** 288
3.1 Water 288
3.2 Constraint Dynamics 288
3.3 Biological Applications 289
3.3.1 Protein Folding 289
3.3.2 Membranes 289
3.3.3 Electron Transfer 290
3.3.4 Crystal Structure 290
4. **Transport Coefficients** 290
4.1 Autocorrelation Functions 290
4.2 Long-Time Tails 291
4.3 Generalized Hydrodynamics 292
4.4 Microscopic Boundaries 293
4.5 Nonequilibrium Molecular Dynamics 293
4.5.1 Boundary-Driven Flows 294
4.5.2 Modified Equations of Motion 294
4.5.3 Nearby Trajectories 294
5. **Particle Hydrodynamics** 294
5.1 Lattice Gas 295
5.2 Lattice Boltzmann 296
5.3 Direct-Simulation Monte Carlo 296
6. **Macroscopic Particle Dynamics** 296
6.1 Polymers and Brownian Dynamics 297
6.2 Hydrodynamic Interactions 297
6.3 Granular-Particle Dynamics 298
7. ***Ab Initio* Molecular Dynamics** 298
7.1 The Local-Density Approximation (LDA) 299
7.2 Car–Parrinello Method 299
8. **Prospects** 300
Glossary 300
Works Cited 301
Further Reading 303

INTRODUCTION

Molecular dynamics has come to mean the numerical solution of the classical Newtonian equations of motion for a representative sample of interacting particles. In effect one is solving the 6*N*-dimensional Liouville equation (*N* is the number of particles), from which, without invoking any ergodic hypothesis regarding the accessibility of regions of phase space, one can obtain long-time (equilibrium) average properties (Alder and Wainwright, 1959a). Equilibrium properties can also be obtained by a competitively powerful numerical method called Monte Carlo (Metropolis *et al.,* 1953), which does assume that the ergodic hypothesis is valid, but molecular dynamics is uniquely useful in solving time-dependent many-body problems.

In fact, the computer-based Monte Carlo and molecular-dynamics methods are the only available techniques by which to solve many-particle problems “exactly.” It is due to this circumstance that these two techniques have found widespread use in almost all branches of science. Some of the more interesting time-dependent calculations from both the fundamental and applied viewpoints will be illustrated in this article. Before that, what is meant by an exact solution must be

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

clarified. Previously, exact solutions of many-particle systems have only been possible through analytic treatment of idealized systems, such as perfect gases or harmonic oscillators. For more realistic interactions, molecular dynamics can give the exact solution for a finite number of particles by following the equations of motion for a finite time. In most applications, these equations of motion are integrated by a finite-difference scheme, which introduces numerical error (Verlet, 1967). However, for rigid elastic spheres, the trajectories are followed collision by collision, so that the only numerical errors involved are roundoff errors in calculating the time to the next collision (Alder and Wainwright, 1963). The effect of each of these limitations is discussed next.

1. TIME CONSIDERATIONS

The very first attempt at integrating the equations of motion numerically led to the misleading conclusion that statistical mechanical systems do not relax to equilibrium rapidly or that systems very much larger than 100 particles are required to reach equilibrium, that is, to observe ergodic behavior.

1.1 The Fermi–Pasta–Ulam Problem

This first investigation (Fermi *et al.*, 1965; Tuck and Menzel, 1972) involved a linear chain of either 32 or 64 nearly harmonic oscillators with an additional fixed particle at each end. This isolated system of slightly anharmonic particles was started from an initial condition in which all the energy was localized in one of the long-wavelength modes. As the classical equations of motion are integrated in time, one expects that the energy will flow from the long-wavelength modes to the shorter-wavelength modes, until an equipartition of energy among all the modes is reached. However, instead of thermalization, the energy was observed to recur periodically (to a very close approximation) in the original mode after being redistributed in only a few other modes. This lack of ergodicity could not be ascribed to the initial preparation of the system nor to the restricted number of degrees of freedom in the finite system but ultimately was attributed to the weak nonlinearity. As the particles interacted more anharmonically (are more strongly coupled), the system behaved as expected.

1.2 Hard Particles

The validity of the ergodic hypothesis was in fact first demonstrated for hard spheres, the extreme anharmonic interaction. For as few as ten particles it was shown that within statistical errors identical equilibrium results were obtained by Monte Carlo and molecular dynamics. Moreover, a rigorous proof of the ergodic hypothesis has been achieved for four hard spheres (Sinai, 1973). The practical numerical demonstration is, moreover, valid for almost all initial conditions one might randomly choose. Some highly selected initial choices, such as all spheres moving in the same direction with the same speed, would not be ergodic.

1.3 The *H* Theorem

A demonstration that the velocity distribution of a hard-sphere system rapidly and monotonically approaches the Maxwell–Boltzmann equilibrium distribution, from almost any randomly chosen initial velocity distribution $f(\mathbf{v})$, has been carried out by calculating the Boltzmann H function (Alder and Wainwright, 1958),

$$H(t) = \int f(\mathbf{v},t) \ln f(\mathbf{v},t)\, d^3v. \tag{1}$$

For an initial distribution where all the particles had the same speed, but random velocity directions, the H function reached its equilibrium value in less than 200 collisions for a 100-particle system, or less than 4 collisions per particle (see Fig. 1). H-function calculations for particles interacting with an attractive potential, in addition to the hard-sphere repulsive potential, show that thermalization now depends on the rearrangement of particle positions, which can be quite time-consuming at high density, but nevertheless, the H function still approaches the final equilibrium value monotonically.

1.4 Irreversibility

The H theorem attributes a direction in time to a system of particles that obey time-reversible equations of motion. This has led

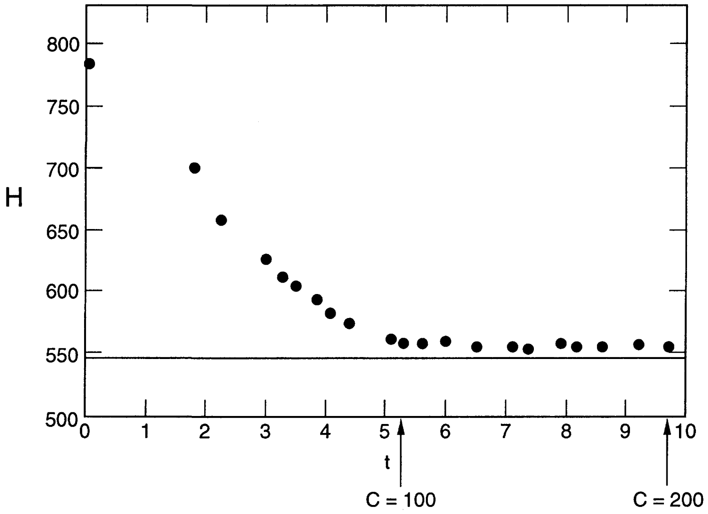

FIG. 1. The behavior of the Boltzmann *H* function with time *t*, or collision number *C*, for 100 hard spheres at low density. The horizontal line represents the equilibrium *H* value for an infinite system.

to a great deal of discussion in the literature, but the statistical nature of this theorem (Lebowitz, 1993) is perhaps best illustrated by the following computer simulation. Suppose you put four spheres in the left half of a box and integrate the equations of motion forward for some time. Subsequently, you reverse the time, that is, reverse the directions of all velocities; nobody is particularly surprised if, after you integrate for an equal time in the reverse direction, you find the same four particles in their original positions on the left-hand side. On the other hand, if the experiment is repeated for 100 particles, everybody recognizes that upon finding all 100 particles again on the left-hand side, a highly unusual event has occurred. This illustrates not only the irreversible nature of systems as small as 100 particles but also the incredibly long (larger than the age of the universe) Poincaré cycle or the time it takes for all 100 particles to return, without intervention, arbitrarily close to their initial position in phase space. The essential point is that coarse-grained distributions, such as $f(\mathbf{r},\mathbf{v},t)$, evolve monotonically from a low-probability initial state prepared by the experimenter to overwhelmingly more probable equilibrium states.

1.5 Numerical Noise

In the above numerical experiments, the 100 particles only return to their original positions if the time to reversal is less than about ten collisions per particle. For longer times the accumulation of roundoff error in the last of the customary 12 digits with which the positions and velocities are specified causes the reversed trajectory to lead to a different configuration. The cause for this rapid divergence of paths is the dynamical fact that closely spaced trajectories in phase space separate exponentially in time. Roundoff error can be considered as the addition of some random noise to the otherwise completely deterministic dynamics. It has never been observed to cause any problem in calculating any transport property, even if the correlation time involved in determining that property exceeds ten collision times per particle. For equilibrium properties, the noise can be considered equivalent to choosing different members of the microcanonical ensemble. In the process of determining averages over initial states, one has the choice of either running a single long trajectory and chopping it up to establish a series of initial equilibrium states or setting up stochastically a set of initial equilibrium states by, for example, the Monte Carlo method.

Roundoff error can be avoided (Levesque and Verlet, 1993) entirely by algorithmic design as well as eliminated in certain simplified models—for example, lattice-gas models (Frisch *et al.*, 1986), which are discussed later. Experiments with such models show that numerical precision is not a limiting factor. In lattice gases only exact logical operations are required, since they are cellular automata algorithms involving only Boolean arithmetic. These models are exactly reversible microscopically, yet for large systems macroscopic observables evolve monotonically, as expected from the second law of thermodynamics.

1.6 Computational Limitations

Molecular-dynamics calculations for hard spheres make it particularly evident that one is dealing with logically sequential calculations, which cannot take advantage of parallel computers to increase the length of time for which the trajectory can be followed. It is easier to study larger systems on parallel computers than it is to lengthen the simulation time. For particles with continuous interaction potentials, where the parallel aspect of computers can be taken advantage of, for example, in the force-field part of the calculation, up to 100 million particles have been simulated for a few picoseconds of real time (Plimpton, 1995).

Molecular-dynamics calculations for particles with short-range interaction forces typically use periodic boundary conditions so as to minimize boundary effects and thus mimic an infinite system as accurately as possible by surrounding a finite system by replicas of itself. The system can also be viewed as residing on a toroidal surface. Long-range Coulombic interactions can be dealt with by a summation method developed by Ewald (Madelung, 1918) for calculating the total energy of an ionic crystal, since interactions between periodically repeated cells must be taken into account explicitly (Brush *et al.*, 1966). However, more computationally efficient methods have been developed recently (de Leeuw *et al.*, 1980; Greengard, 1994). For the more common short-range–force situation, the periodic cell is further divided into regions of size somewhat larger than the range of the forces, typically three particle diameters, to take advantage of the fact that distant regions do not contribute significantly to the force calculation. This domain decomposition, each cell of which typically contains 100 particles, significantly reduces the calculation time in systems of more than 1000 particles, so that it is then proportional to the number of particles rather than the number of particles squared. The neighbor list in each compartment does not have to be updated until the maximum particle displacement exceeds the difference between the compartment size and the range of the forces. Strips of that size difference occasionally have to be interchanged between nodes of a parallel computer.

1.7 Length of Trajectory

The time-consuming process in a molecular-dynamics simulation is the force calculation necessary to advance all the particles one time step. The step size is usually taken to be the same for all particles and is limited by the requirement that the force does not change too much in the time interval. For simple pairwise interacting particles at liquid densities, that time step typically corresponds to 10^{-14} s in real time. There are many algorithmic choices for updating the particle positions and velocities, but important considerations are simplicity, efficiency, accuracy, stability, and reliability. Another consideration is that only one force calculation per time step is usually practical. Because force can change rapidly with distance, it does not pay to go much beyond third-order accurate integration schemes; this has led to the Verlet leapfrog algorithm as the method of choice (Berendsen and Van Gunsteren, 1986). The algorithm is a simple finite-difference scheme to solve second-order differential equations for the particle coordinates; particle velocities are obtained by differencing coordinates.

In a sample simulation of 10^3 particles, the force calculation requires about 10^4 machine clock cycles per particle per time step or of the order of 10^7 cycles per time step. In a long calculation of 30 h or 10^5 s that runs on a processor with a clock speed of 10^8 cycles/s, a total of 10^{13} cycles can be executed; this means 10^6 of the above time steps or 10^{-8} s of real time. Massively parallel computers allow simulations of much larger numbers of particles, where each processor typically handles a domain consisting of 10^3 particles, but not for longer real times since each processor is subject to the above limitations. Unless clock speeds improve significantly, this limitation will persist. Lattice gases overcome this limitation by reducing both the compartment size and the number of neighbors to the order of ten particles each, as well as significantly reducing the number of cycles per time step. This allows systems to be followed for many orders of magnitude longer in time.

1.8 Rare Events

Because trajectories cannot be followed for macroscopic times, events that occur very infrequently—for example, as a result of a rare fluctuation—cannot be studied without some special preparation of the system that focuses on the unusual happening. In general, dealing with phenomena that occur at vastly different time scales in a given system presents formidable numerical problems unless one can ignore the coupling between the different time scales (Brackbill and Cohen, 1985). Then one can either average over the short time scale or ignore the long time scale.

As an example, consider vacancy diffusion in which a hard sphere occasionally jumps into a neighboring vacancy (Bennett and Alder, 1971). The bottleneck can readily be identified by physical insight, although this is not always the case. One then resorts to the transition-state theory, developed long ago for chemical reactions, which gives the rate of a process as the probability of the system being in the activated state multiplied by the transmission coefficient, which measures whether the reaction will successfully proceed from one side of the barrier to the other, once the particle is at the barrier. The only assumption made in this theory is that of local equilibrium. As a consequence, the jumping sphere is successively confined between two closely spaced walls along the rare-event trajectory, and the system is allowed to equilibrate at each successive position of the two walls (Connick and Alder, 1983). From the pressure difference between the two walls, derived from the collision rate with the walls, the work necessary to drag the sphere to the saddle point can be evaluated (see Fig. 2). Having thus obtained the probability of being at the top of the barrier by an adiabatic process, the two walls are removed, and the transmission coefficient is obtained by finding how often the forward trajectory ends in the hole and the backward trajectory at the original site of the jumping sphere. The transmission coefficient will automatically correct for placing the dividing surface somewhere other than at the exact saddle point. There are many variations of this theme, but the key ingredient is the choice of the transition state to be near the bottleneck.

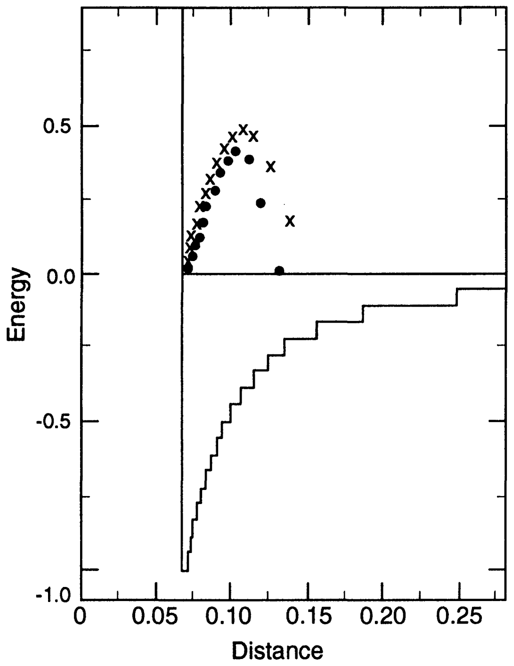

FIG. 2. Energy (dots) and free energy (crosses) of activation for exchanging a sphere in contact with an attracting particle located at the origin with another neighboring sphere, determined by confining the exchanging sphere between successively more distant pairs of walls located at the indicated steps in the attractive potential.

Another example of rare events is the mechanism by which a crystal melts; in this case the bottleneck was not known in advance, and so brute-force long runs were required. The rare events are schematically indicated in Fig. 3. It was shown (Barker *et al.*, 1981) later that this mechanism represents an instability that occurs at densities slightly below melting. However, the only melting criterion that has been rigorously established is the thermodynamic one, namely, that the free energy of the solid equals that of the liquid.

2. EQUILIBRIUM PROPERTIES

Although we do not want to dwell on equilibrium properties, a few key points are worth enumerating insofar as they relate to

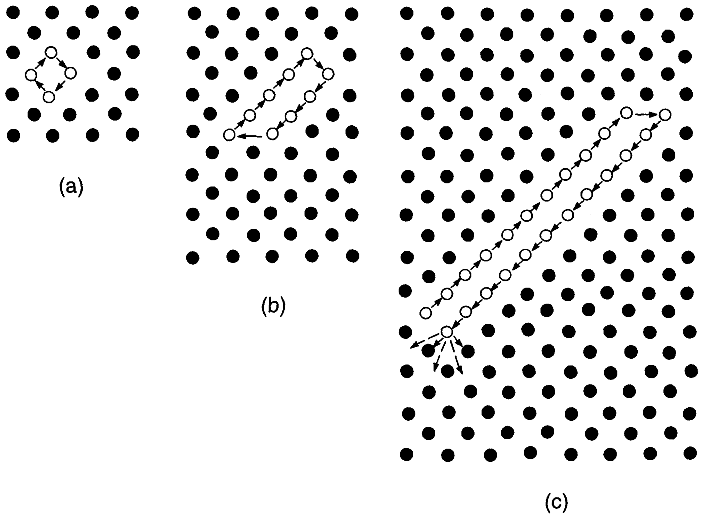

FIG. 3. A schematic diagram of the particle motion (indicated by arrows) in a two-dimensional solid, shown at three successively lower densities near the melting density. (a) First phase of closed-loop motion showing four particles changing position in lattice. (b) As the density is lowered, the loop elongates and contains more particles. (c) Breakup results when the loop becomes so large that the jump of one atom at one end of the loop is no longer correlated with the jump at the other end.

nonequilibrium properties. The first point to consider is the relative efficiency between Monte Carlo and molecular dynamics in determining equilibrium properties. For simple systems such as hard spheres, it is difficult to make a choice (Jacucci and Rahman, 1984). As the properties of more complex molecular systems are studied, Monte Carlo may be most efficient if "smart" or biased Monte Carlo moves can be employed (Rao and Berne, 1979; Siepmann and Frenkel, 1992).

2.1 Phase Transition

The first surprising result that came out of a comparison of Monte Carlo and molecular-dynamics calculations was that, except in regions near a phase transition, a 100-particle system could mimic an infinite system, in that quantitative differences in thermodynamic properties were only of the order $1/N$. Near phase transitions, long-range fluctuations are large, and thus those regions of the phase diagram are greatly distorted in small systems. At a second-order phase transition, such as the critical point between a liquid and a gas, the additional problem of critical slowing down has to be faced, that is, the infrequency with which such collective long-range fluctuations occur. There are special Monte Carlo techniques to deal with that (Swendsen *et al.*, 1992). In first-order phase transitions, such as the liquid–solid transition, the difficulties are the rarity of fluctuations that lead to nucleation of the solid (see Fig. 3) and the large number of particles that are required for the liquid and solid phases to coexist (Alder and Wainwright, 1959a). Free-energy matching of different phases is the most important technique to determine phase boundaries (Frenkel and Ladd, 1984).

In spite of the above difficulties, it proved possible, in very long runs, to demonstrate the coexistence of a solid and fluid phase for two-dimensional systems of hard disks with about 1000 particles (Alder and Wainright, 1959b, 1962) (see Fig. 4). This and complementary simulations of three-dimensional hard-sphere systems settled an important issue, namely, whether purely repulsive forces alone could lead to solid formation. No rigorous analytic proof of this transition is available as yet.

2.2 Ensembles

A further problem in simulating freezing from a liquid is that for small systems the kind of crystal that can form is determined by the boundary conditions, namely, that an integer multiple of unit cells must fit in each dimension in order to form a perfect crystal of a specified structure. This is because most molecular-dynamics calculations use a microcanonical ensemble, where the energy and, in particular, the volume are fixed. However, other ensembles can be formulated

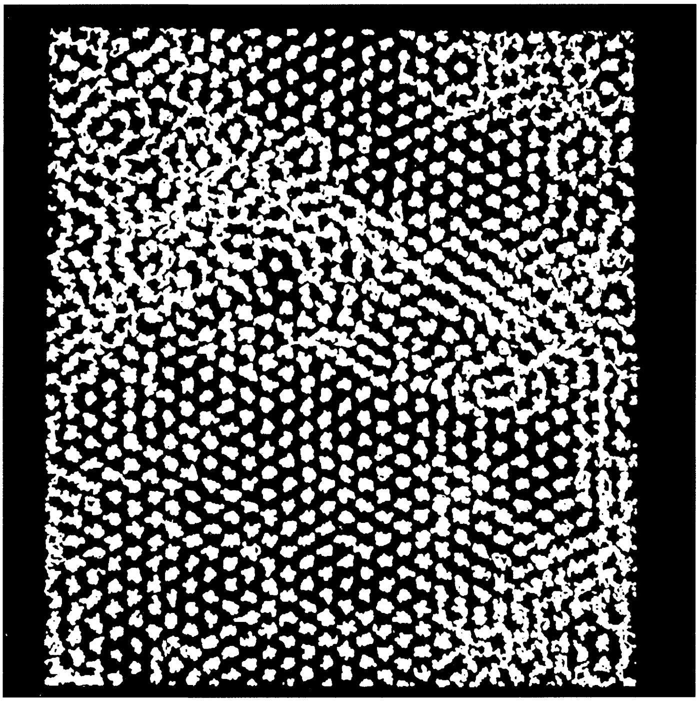

FIG. 4. The trajectories of the centers of particles for a few thousand collisions in the two-phase region. Ordered regions, where particles oscillate about their solid lattice positions, can be clearly seen to coexist with liquid regions where the particle centers trace out short line trajectories as the particles diffuse between lattice positions.

(Anderson, 1980)—in particular, constant-pressure and -temperature ones that correspond more naturally to the freezing process and that allow the system to accept a variety of crystal structures. Such molecular-dynamics ensembles can provide correct equilibrium properties, including the crystal structure (Nosé and Klein, 1983). Correct dynamical properties for such ensembles can only be attained under conditions such that perturbations vary sufficiently slowly that the system is in effect microcanonical for times comparable to a typical relaxation for the process under consideration. A constant-pressure ensemble can also be formulated for Monte Carlo simulations and shows a bimodal density distribution between fluid and solid in the two-phase region (Wood, 1970).

2.3 van der Waals Approximation

The physical significance of the hard-sphere system is that it provides a basis from which to determine the properties of real fluids by treating the attractive potential as a perturbation. The van der Waals equation is an important prototype of this idea; it relates the pressure p, volume V, and temperature T by

$$(p + a/V^2)(V - b) = RT. \qquad (2)$$

The pressure is modified by the attractive potential contribution given as a constant a divided by V^2, and the effective volume is reduced by the excluded volume occupied by the spheres b (R is the gas constant). Rewriting this equation as

$$pV/RT = V/(V - b) - a/VRT \qquad (3)$$

and replacing the hard-sphere term $V/(V - b)$ by the more accurate computer-generated hard-sphere pressure allows the a/VRT term to be identified as a perturbation due to the attractive forces, namely, the ratio of the potential energy a/V to the kinetic energy RT. This view can be made rigorous if the attractive potential is sufficiently weak and long range (Uhlenbeck *et al.*, 1963), so that no matter how the particles are arranged they are always within each other's range of interaction. Then the mean-field theory represented by the van der Waals equation is rigorously valid; that is, the higher-order attractive perturbation terms $O(T^{-2})$ are negligible. It is also noteworthy that with the establishment of the solid phase for hard spheres, this simple theory contains all three states of matter, that is, solid, liquid, and gas.

The van der Waals theory is the foundation of the modern theory of simple liquids, not only for equilibrium but also for nonequilibrium properties. For transport properties, the van der Waals theory permits the attractive potential to be ignored, since in the mean-field approximation the constant attractive potential never exerts any force on a particle. Indeed, to a good approximation the transport coefficients of hard spheres reproduce those of real monatomic fluids (Dymond and Alder, 1966).

3. COMPLEX MOLECULES

It is fortunate that most gross macroscopic properties are primarily determined by the repulsive forces, since we do not yet know the intermolecular forces very accurately. Furthermore, even for simple systems such as argon, the potential is not, as is often assumed, pairwise additive (Barker, 1986). It is known that the effective pair potential at liquid densities differs by 20% in well depth from the true pair potential (Maitland *et al.*, 1981). Besides repulsion, the next most important force to take into account in more complex molecules is the Coulomb charge distribution. This is often done by empirically assigning partial charges to various atoms in the molecule. Such partial charges can be obtained from quantum mechanical calculations, from which the correct potential energy of interaction must ultimately be obtained. However, this task is at present much too complex to be carried out with sufficient precision. Nevertheless, it is amazing that by merely assigning a charge and size to an atom, one can obtain semiquantitative results, even ignoring the important polarization effects. More delicate questions, such as the conformation of a polymer, require more accurate knowledge of the intermolecular potential. This is because differences in conformation involve tiny energy differences. At present one must resort to effective empirical potentials for complex molecules, obtained (in a building block fashion) by reproducing properties of simpler molecules by molecular dynamics (Lifson *et al.*, 1979).

3.1 Water

It is remarkable how well a distributed-charge model can reproduce the properties of water, even though the hydrogen bond can only be accurately represented quantum mechanically (Rahman and Stillinger, 1971). Studies of water and ionic solutions are, of course, important precursors to the study of biological systems. From molecular-dynamics studies employing simple model potentials for water, dynamic properties such as the residence time of a water molecule near a positively or negatively charged ion can be obtained. Structural properties such as the difference in orientation of a water molecule in the neighborhood of an ion can also be ascertained. Such properties are not very sensitive to the details of the interaction potential.

3.2 Constraint Dynamics

The most commonly used algorithm for molecules, even for water, is constrained molecular dynamics (Ciccotti *et al.*, 1982). The constraint for the water calculation, for example, would be to keep the two oxygen–hydrogen bond lengths fixed at their experimental value. A completely rigid model of water would also keep the angle (the bending angle) between these two bonds fixed. This makes physical sense since the vibrational and bending frequencies are not very different. In general, the motivation for introducing constraints is to eliminate fast modes, such as vibrational modes, which do not significantly affect the macroscopic properties but which do affect the efficiency of the molecular-dynamics calculation. If one had to take into account these fast modes, a considerably smaller integration time step would be necessary. Since vibrational periods are typically at least 100 times less than correlation times for translational or rotational velocities, they can safely be ignored.

In general, complex molecules are modeled by chains of point particles interacting by a potential energy that is split into an intermolecular and an intramolecular part. The latter part is usually written as a sum of various bond-bending or torsional energies, and pair interactions between nonbonded atoms on the same molecule. Most commonly only the direct intermolecular bonds are constrained since they differ significantly in frequency from the others. The simplest and most common implementation of the constrained procedure is to take unconstrained steps in the atomic coordinates, utilizing the leapfrog algorithm, and then apply a resetting mechanism to the positions to satisfy the constraints. The constraints are introduced as a set of Lagrange multipliers to be evaluated by an auxiliary calculation at each time step. This involves solutions to a set of nonlinear algebraic equations that can be obtained iteratively, treating each individual constraint in succession. The Lagrange multipliers represent the forces directed along the bonds required to keep the bond lengths constant. The method works well, even for a very complex network of bond constraints, as demonstrated by simulation of small proteins (Brooks *et al.*, 1988).

3.3 Biological Applications

The principal difficulties in applying molecular dynamics to biological macromolecules are the computational problem of the long time scales on which interesting phenomena occur and the lack of an accurate knowledge of the interatomic forces.

3.3.1 Protein Folding The physical time scales involved in the very important problem of protein folding are of the order of seconds or minutes and impossible to reach computationally (McCammon *et al.*, 1977). The main reason for this slow relaxation rate is the vast dimensionality of configuration space that has to be sampled. The only reasonable approach to the protein-folding problem is to decompose the calculation by investigating marginally stable intermediate secondary structures that can be observed experimentally. Simulations have so far been carried out for folding and unfolding of small peptides and proteins (Bishop and Schulten, 1994) (see Plate 1). The suggestion has been made that such a process might be speeded up by funneling into a singular deep energy minimum (Shakhnovich, 1994). To investigate various folding mechanisms, it is possible to speed up the process by artificially altering the force field along relevant bond angles.

A particular example of the use of molecular dynamics in structural biology is the study of the interaction of proteins with nucleic acids (DNA or RNA). These interactions are the basis of a variety of essential cellular functions, including DNA replication and transcription. The hormone receptor proteins represent a class of proteins that interact with DNA to regulate gene expression. The cocrystal x-ray structure of the DNA binding domain of one member of this class, the glucocorticoid receptor, interacting with a specific sequence of DNA has been determined. However, the observed features in the cocrystal structure are not biologically active. With the use of molecular-dynamics techniques, a complex that represents the biologically active element in this regulatory mechanism was constructed that is more closely related to the physiologic environment [see Plate 1(a)] by placing the system in a drop of water. The simulations indicate that when the protein interacts with DNA, the conformation of the DNA assumes a bent and unwound conformation as compared with uncomplexed DNA [see Plate 1(b)]. These observations are not borne out in the crystallographic environment; however, they are observed on a macroscopic scale by the use of other experimental techniques that are conducted under physiological conditions. The conformational changes that the protein induces in the DNA are essential to a proposed mechanism of gene regulation for this class of proteins by making a particular gene accessible for transcription.

3.3.2 Membranes Fortunately, there are a number of interesting phenomena that occur on much shorter time scales. An example is binding of a small molecule to a macromolecular site, a subject intensively pursued in drug design (Berendsen, 1986). If the rate is too slow, the rare-event algorithm, discussed earlier, can be invoked. Other important studies involve the transport of small molecules across membranes. Bilayer membranes self-assemble by using a molecule that has a polar group at one end that likes to interact with the water solvent (hydrophilic) and a nonpolar group at the other end that does not (hydrophobic). These systems are similar to soap solutions, which can assume a variety of micellar phases, depending on concentration (Smit *et al.*, 1990). In certain concentration ranges a lamellar structure is obtained consisting of bilayers that are permeable to various smaller molecules. Simulations have also been used to study the insertion of lipid molecules into the bilayer.

3.3.3 Electron Transfer Another interesting example where molecular dynamics is making an impact is in the mechanism of electron-transfer rates in photosynthetic reactions (Gehlen *et al.*, 1994). The free electron generated by photon-induced excitations appears to couple electrostatically to the nuclear charges that fluctuate in position in the macromolecule. From molecular dynamics the thermal motion of the charges for a given molecular structure yields the spectral density of the charge fluctuations, which apparently controls the electron transfer rate.

3.3.4 Crystal Structure A final example where molecular dynamics has had an impact is in the determination of crystal structures of proteins from x-ray diffraction. Initial crystal structures can be considerably refined by utilizing additional data that are sensitive to different distance scales and features of the molecular configuration. For example, nuclear magnetic resonance is sensitive to the proximity of different hydrogen atoms within the molecules. Utilizing that information, in addition to x rays, helps define the global structure of the molecules more accurately. Similarly, molecular dynamics, which is sensitive to short-range spatial interactions, can help refine the local protein structure, leading to a considerably more accurate location of the atoms in a crystal structure (Karplus, 1986).

4. TRANSPORT COEFFICIENTS

Molecular-dynamics simulations of single-particle diffusion led to a fundamental disagreement with kinetic theory, namely, that successive collisions in a fluid would be uncorrelated. This "molecular chaos" assumption corresponds mathematically to a Markov relaxation process; it necessarily leads to an exponentially decaying velocity correlation function, contradicting the computer-observed power-law decay (Alder and Wainright, 1970). This unexpected discovery of long-term memory effects through computer experiments is particularly noteworthy since it preceded laboratory experiments (Paul and Pusey, 1981). It led to a profound change in the mathematical structure of kinetic theory (Dorfman and Cohen, 1975). In retrospect this phenomenon could have been predicted, if hydrodynamics had been thought to be applicable on molecular time and distance scales. In fact, quite the contrary was believed, so that the coherent hydrodynamic flows that lead to correlated events were ignored. Surprisingly, the observed correlation effects are quantitatively accounted for hydrodynamically, as is shown in Sec. 4.2. This in turn leads to the intriguing practical consequence that the relatively small systems studied by molecular dynamics for short times can be used as an alternative to continuum mechanics to explain hydrodynamic phenomena. The advantage of this approach is discussed in Sec. 5.

4.1 Autocorrelation Functions

There are two methods to calculate transport coefficients; both have analogies to real experiments. In one case, one studies the decay of spontaneous fluctuations by means of an autocorrelation function (Kubo, 1957). In the alternative method, one sets up a steady nonequilibrium state through an external force (Ashurst and Hoover, 1975); this is discussed below.

Many relaxation processes have been studied by molecular dynamics via correlation functions. For example, dielectric relaxation monitors the average decay in time of fluctuations in the total dipole moment of the system. Other examples involve the decay of orientational components of a molecular system to determine various spectroscopic properties. We concentrate here on the velocity autocorrelation function $\rho(t)$, by which the self-diffusion coefficient D can be determined (Rahman, 1964):

$$\rho(t) = \langle v(0)v(t)\rangle, \tag{4}$$

where $v(0)$ is the velocity of a particle at the origin of time and $v(t)$ is its velocity at some later time t. The diffusion coefficient is the integral of the autocorrelation function:

$$D = \int_0^\infty \rho(t)\,dt. \tag{5}$$

This expression is equivalent to the more familiar Einstein expression

$$D = \lim_{t\to\infty} \langle \Delta r^2(t) \rangle / 6t, \tag{6}$$

where Δr is the displacement of a particle position in time t. The angular brackets in these expressions refer to averages over many initial equilibrium states. The shear viscosity can be obtained in like manner from an integral of the stress autocorrelation function (Levesque *et al.*, 1973).

4.2 Long-Time Tails

The decay of the velocity autocorrelation function of a fluid of hard disks shows a long positive persistence of velocity that implies that a particle remembers its initial velocity after it has undergone in excess of 100 collisions (Alder and Wainwright, 1969). This long-time tail can be quantitatively accounted for by a hydrodynamic model in which a solid disk is surrounded by a two-dimensional continuum fluid with a shear viscosity equal to that of the hard-disk system. The solid particle is given an initial velocity, which leads to a positive pressure wave in the fluid ahead of the disk and a corresponding rarefaction wave behind. The positive pressure in front and the negative pressure behind equalize by creating a double vortex flow around the particle (see Fig. 5). By comparing the velocity field around the disk calculated by molecular dynamics with the continuum hydrodynamic one under identical conditions, the quantitative validity of hydrodynamics at length scales of three molecular diameters (10 Å) and time scales of 10 mean collision times (10^{-13} s) was established. A straightforward dimensional analysis of the hydrodynamic model leads to the conclusion that the tail decays as $t^{-d/2}$, where d is the dimensionality of the system (see Fig. 6). This power-law decay is to be contrasted with the exponential decay predicted by Markov processes. It is also to be noted that in strictly two-dimensional systems ($d = 2$), the diffusion constant diverges; in fact, in two-dimensional systems none of the transport coefficients exist, requiring a reformulation of continuum hydrodynamics that has not as yet been carried out.

The theoretical prediction of divergent transport coefficients in two-dimensional systems has been confirmed by lattice-gas simulations of the stress autocorrelation function, which shows a long-time t^{-1} tail. The shear viscosity, which is proportional to the integral of the stress autocorrelation function, therefore diverges logarithmically. At liquid densities, near solidification, molecu-

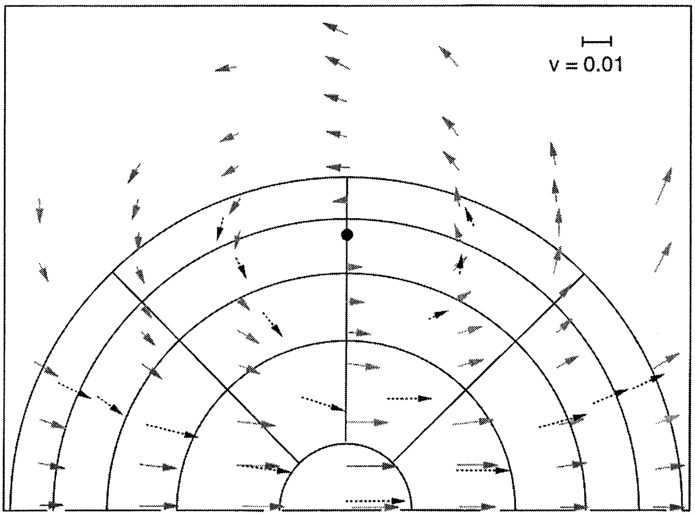

FIG. 5. The velocity correlation between a central particle and its neighboring particles after the central particle has undergone about ten collisions. (Because of symmetry, only the upper half of the diagram is shown; the vectors shown are the average of those of the upper and lower halves.) The size of the central particle is shown by the central half-circle and its velocity by the size of the arrow with a dotted tail near its center. The other such arrows are of a size to indicate the velocity average (as obtained by molecular dynamics) of whatever neighboring particles are found in that section of space. These arrows are to be compared with the light arrows, with line tails, which are obtained from a hydrodynamic solution of the velocity field on a rectangular grid for the identical physical situation. The agreement is striking, confirming the vortex motion.

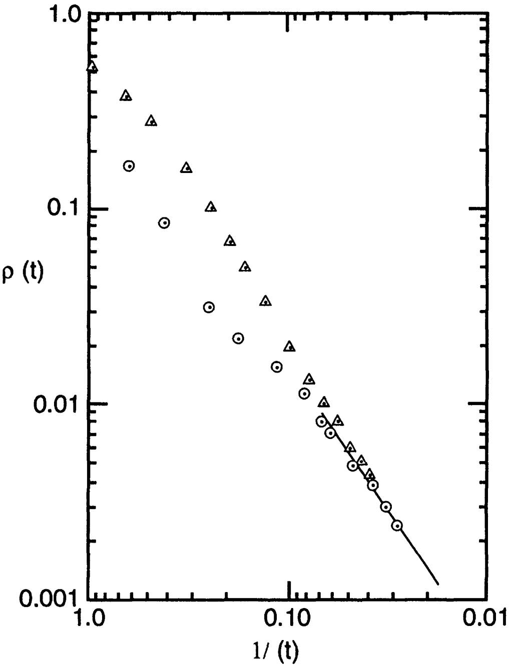

FIG. 6. Comparison of the velocity autocorrelation function $\rho(t)$ as a function of time (in terms of mean collision times) between the hydrodynamic model (circles) and a 500-hard-sphere molecular-dynamics calculation (triangles) at a volume relative to close packing of 3 on a log-log plot. The straight line is drawn with a slope corresponding to $t^{-3/2}$.

lar dynamics shows another type of slow decay for the stress autocorrelation function, characterized by a "stretched" exponential, $e^{-t^{\beta}}$ (β is a constant) (Palmer *et al.*, 1984), similar to experimental observations of viscous and dielectric relaxation. This tail, often called the "molasses" tail, appears to be connected with a slow structural adjustment of the dense fluid in the presence of a shear (Ladd and Alder, 1989). This process is characterized by a series of different exponential relaxations, the totality of which can be expressed as a stretched exponential. The detailed structural relaxation mechanisms in dense liquids have yet to be identified and very likely would lead to a dynamic theory of glass formation.

4.3 Generalized Hydrodynamics

Fluctuations in the hydrodynamic variables (mass, momentum, and energy) also depend on length scale (or wavelength), as well as on amplitude. We forego discussing dependence on amplitude, which leads to nonlinear transport coefficients, because that subject is not well understood. However, we briefly discuss the dependence of the linear transport coefficients on wavelength; this is the subject of generalized hydrodynamics (Alley and Alder, 1983). Introducing a molecular length scale into hydrodynamics permits a quantitative discussion of the corrections to macroscopic hydrodynamics as one approaches atomic dimensions. The only wavelength-dependent fluctuations of atomic dimensions accessible experimentally are density fluctuations, measured by neutron-diffraction studies (see NEUTRON DIFFRACTION). Molecular-dynamics calculation of the neutron-scattering function (Alley and Alder, 1983) have been made, and the experiments compare quantitatively to the prediction of generalized hydrodynamics (see Fig. 7).

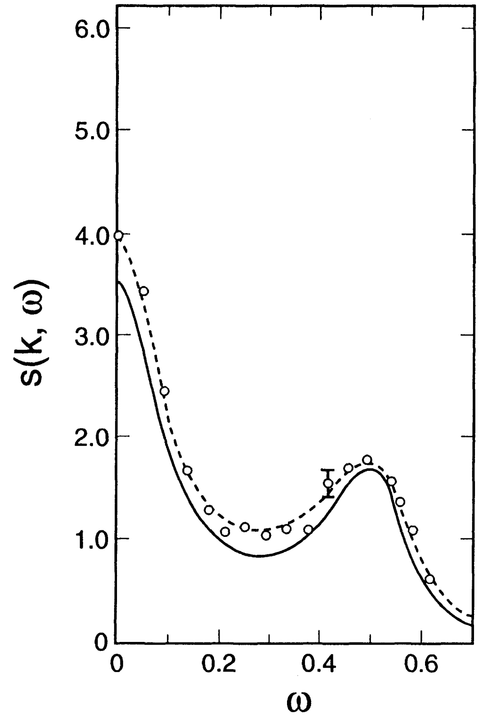

FIG. 7. The neutron-scattering function $S(\kappa, \omega)$ for hard spheres as a function of frequency ω, at a volume three times larger than close packing and a wavelength ($2\pi/k$) 2.63 times the hard-sphere diameter. The molecular-dynamics results are represented by circles, and the hydrodynamic results, by the dashed curve. The solid curve represents the results of kinetic theory, valid only at lower densities.

As an illustration of generalized hydrodynamics, consider the problem of an infinitely massive sphere of atomic dimensions falling in a fluid, in order to find out what the correction to Stokes's law would be (Alley and Alder, 1980). By analysis of the wavelength dependence of the stress fluctuations, a wavelength-dependent viscosity $\eta(k)$ was deduced by molecular dynamics, with the result that

$$\eta(k)/\eta(0) = (1 + a^2k^2)^{-1}, \tag{7}$$

leading to a correction to Stokes's law of $1 - a/\sigma_{12}$, where σ_{12} is the mean diameter of the falling sphere and the solvent molecules. The crucial distance scale a that is introduced through the generalized viscosity is of the order of molecular size, that is, $a/\sigma_{12} \sim 0.3$ for a dense hard-sphere solvent. This Stokes prediction agrees well with a separate molecular-dynamics calculation of the Stokes friction coefficient and shows that even for atomic dimensions the correction to Stokes's law is only about 30% (see Fig. 8).

4.4 Microscopic Boundaries

The above Stokes problem requires that boundary conditions be imposed on the surface of the sphere to determine the hydrodynamic friction coefficient. Two common boundary conditions correspond to slip or stick; they determine how much of the tangential momentum of a colliding solvent molecule is randomized. From an atomistic view, slip boundary conditions correspond to elastic collisions, whereas stick corresponds to diffusive collisions, whereby the post-collision velocity direction is randomly distributed. The real situation lies somewhere between the above two extremes, depending on the roughness of the surface, and can be ascertained only by a molecular-dynamics simulation. Thus, by forming the falling sphere from smaller spheres representing the size and density of the atoms of the material of which it is constituted, molecular dynamics yields the correct boundary conditions when this aggregate object is suspended in the solvent spheres that collide elastically with the falling sphere (see Fig. 8). Such realistic boundary conditions generated microscopically are important when one studies, for example, fluid flow through microscopic pores (Koplic and Banavar, 1995).

4.5 Nonequilibrium Molecular Dynamics

The linear transport coefficients (diffusion, viscosity, and thermal conduction) can be calculated by monitoring the decay of

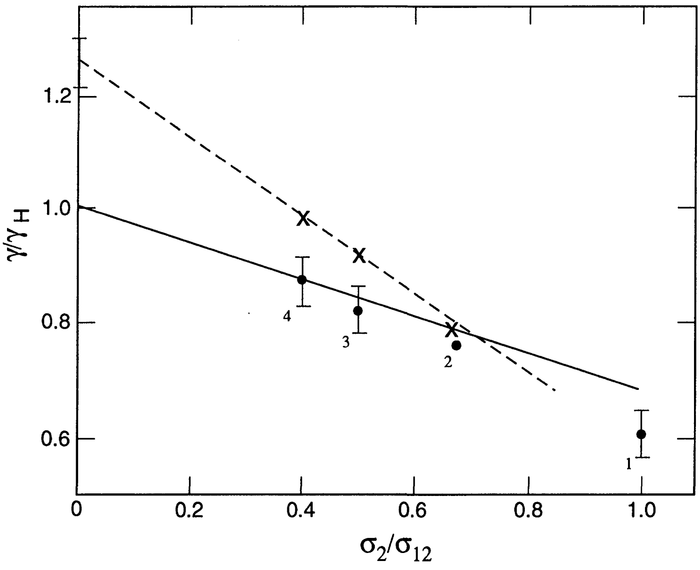

FIG. 8. The ratio of the molecular-dynamics determined friction coefficient γ to that of the hydrodynamic one, γ_H, at various ratios of the diameter of the solvent, σ_2, to that of the average diameter of the sphere and the solvent in which the sphere is immersed, σ_{12}. The four circles with their indicated error bars correspond to the molecular-dynamics results for spheres, respectively, 1, 2, 3, and 4 times the size of the solvent; the solid line represents the generalized hydrodynamic result, extrapolating to the hydrodynamic result for slip boundary conditions, in the large-ratio size limit ($\sigma_2/\sigma_{12} = 0$, $\gamma/\gamma_H = 1$). The crosses correspond to molecular-dynamics results for a big sphere having a rough surface, made up of little spheres of the same size as the solvent. The dashed line shows that in the hydrodynamic limit the results extrapolate to a value in between that for slip boundary conditions (1) and stick boundary conditions (1.5).

fluctuations in an equilibrium ensemble, as discussed above. However, it is also possible to drive a nonequilibrium flow of mass, momentum, or energy and measure transport coefficients directly (Hoover *et al.*, 1980). Moreover, nonequilibrium molecular dynamics (NEMD) is a unique probe of nonlinear transport phenomena, such as occur during the shearing of a polymer solution or the plastic flow of a crystalline solid (Hoover *et al.*, 1982). In such situations, rearrangements of the microstructure in the externally imposed flow field lead to nonlinear responses. Such flows are of considerable practical importance in understanding the rheological properties of complex fluids and solids. NEMD simulations fall into two categories: boundary-driven flows and modifications of the equations of motion.

4.5.1 Boundary-Driven Flows In such simulations flows of momentum or energy are set up by interactions between the particles and some external reservoir, for example, a collection of particles with a preassigned velocity distribution, or a container wall. Two such reservoirs act as a source and sink of momentum and energy and allow nonequilibrium flows to be set up in the system.

Boundary-driven flows are easy to visualize and are closely related to experimental methods. However, in systems of microscopic size, the reservoirs introduce substantial spatial inhomogeneities of heat in NEMD simulations, so that the thermodynamic state of the system is spatially varying. Hence, these methods are not suitable for accurate calculations of transport coefficients.

4.5.2 Modified Equations of Motion For an accurate calculation of transport coefficients, it is necessary to set up a homogeneous perturbation in the distribution function of particle coordinates and velocities, so that the system approximates a small volume of a macroscopic system at constant mass, momentum, and energy but with fluxes of mass, momentum, or energy flowing through it (Hoover *et al.*, 1980; Evans, 1982). To illustrate the principle, consider the simplest case of the diffusion coefficient D of a tagged particle in a system with periodic boundary conditions. The NEMD method consists of applying an external force F to the tagged particle [a compensating force $-F/(N-1)$ must be applied to the $N-1$ other particles] and measuring its mean velocity u. For small enough driving forces, the ratio u/F is equal to D/kT, the self-diffusion coefficient divided by the Boltzmann constant k times the temperature T. The other transport coefficients can also be measured from homogeneous modifications of the equations of motion. Furthermore, it is possible to establish a rigorous connection between the NEMD simulations and the autocorrelation formulas for the linear transport coefficients. NEMD is about as efficient numerically as calculating autocorrelation functions.

Because in NEMD simulations very large gradients are necessary to establish a significant signal-to-noise ratio, large amounts of energy are dissipated, and the systems consequently heat rapidly. Thus, methods of uniformly thermostatting the system had to be developed, and again, suitable modifications to the equations of motion allowed for simulation of a steady state (Nosé, 1984). It should be noted that viscous heating is a second-order effect (in the external driving force), so that the linear transport coefficients are not sensitive to the details of the thermostatting process.

4.5.3 Nearby Trajectories It has been suggested that the signal-to-noise problem in nonequilibrium simulations could be overcome by a "subtraction" technique (Ciccotti *et al.*, 1979). The idea is to start two trajectories from the same initial phase point: an equilibrium trajectory and a weakly perturbed nonequilibrium one. These trajectories might be expected to be highly correlated. If so, the nonequilibrium response could be determined with small statistical noise, by subtracting the fluctuating equilibrium one, which has zero mean, from the nonequilibrium response. This technique also allows the use of small perturbations, resulting in negligible viscous heating. Unfortunately, the chaotic nature of the underlying dynamics, and the resulting exponential divergence of neighboring trajectories, makes this scheme of limited practical value.

5. PARTICLE HYDRODYNAMICS

Encouraged by the hydrodynamic explanation of the long-time tails, one might think of doing other fluid-flow problems by molecular dynamics, instead of solving the continuum Navier–Stokes equations (see Fig. 9) (Meiburg, 1986). There are distinct advantages as well as disadvantages. For the Navier–Stokes equations there is no proof of the existence, convergence, or uniqueness of the numerical solutions, while molecular dynamics, being a particle method, is numerically stable, although computationally very expensive. Navier–Stokes ignores fluctuations while molecular dynamics includes them within the limitations of a finite system. Navier–Stokes is only strictly valid when gradients are small, even though it is often used to calculate shock-wave propagation, while molecular dynamics is valid even when the linear-gradient approximation breaks down. Finally, the continuum approximation of Navier–Stokes is not applicable when the mean free path is comparable to the characteristic length scale of the flow, such as in rarefied gas flows; in such cases, one needs to solve the Boltzmann equation, as discussed in Sec. 5.3.

5.1 Lattice Gas

Although hydrodynamic simulations have been carried out by molecular dynamics (Rapaport, 1988), they are computationally intensive and cannot be pursued for very long times. Thus an alternative, simpler, particle method is highly desirable, particularly since molecular dynamics often produces much more detailed information than is either necessary or digestible. This is the motivation behind lattice-gas models where the particles' positions are restricted to a lattice and the particles have a very limited set of velocities. Collisions between particles still must conserve mass, momentum, and energy. The only other restriction is that of detailed or semidetailed balance, which means that irreversibility is not introduced at the microscopic level or, in other words, into the equations of motion. Detailed balance is satisfied if the probability of a collision event is equal to that of the reverse one, while for semidetailed balance it is true on the average. It can be shown (Frisch *et al.*, 1987) that the conservation laws and semidetailed balance are sufficient to derive the equations of hydrodynamics. Some lattice-gas models violate semidetailed balance to reach a particular goal—for example, to lower the viscosity of the gas so as to achieve higher Reynolds-number flow; such models are nonphysical.

Lattice gases involve only integer algebra and are orders of magnitude faster than molecular dynamics, especially on special-purpose computers. The algorithms are numerically stable and well suited to massively parallel computers. However, because of the limited set of velocities, a number of problems arise that have limited the application of lattice gases. One such problem arises because, at a given lattice point, only distinguishable particles are allowed; that is, all particles must have different combinations of position, speed, and direction. As a consequence, an exclusion effect arises, which leads to a Fermi-like particle velocity distribution rather than the Boltzmann distribution. As a result, special measures must be taken to ensure that translational invariance is obeyed for high-velocity flows. Moreover,

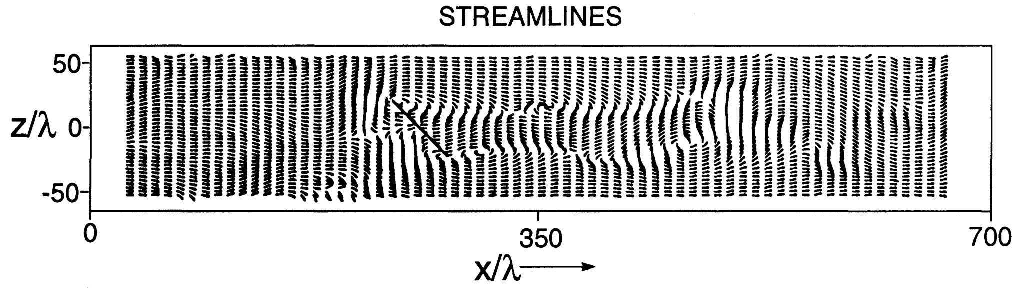

FIG. 9. The velocity of volume elements of a fluid streaming past a plate (shown by the solid line) set at an angle to the flow, in a container 700 mean free paths long, 100 mean free paths wide, and 1 mean free path deep. The molecular-dynamics simulation of some 100 000 spheres shows the alternate vortex-shedding phenomenon behind the plate leading to an observable von Kármán wake.

the lattice must be of sufficiently high symmetry that the viscous stresses are isotropic. Furthermore, lattice gases, being only point particles, are restricted to a perfect-gas equation of state. There is also a computational drawback that in practice long runs are needed to average out the naturally occurring fluctuations. Nevertheless, lattice gases have been used to advantage to gain more detailed knowledge of the behavior of the long-time tails (van der Hoef and Frenkel, 1991). They can also deal with flows in complex geometries, such as porous media and particulate suspensions (Ladd *et al.,* 1988).

5.2 Lattice Boltzmann

Since lattice gases are inherently dilute, one can also solve the corresponding Boltzmann equation and avoid the time-consuming average over the fluctuations (McNamara and Zanetti, 1988). The idea is to solve the Boltzmann equation on a lattice with a discrete set of velocities. In this case one calculates the probability densities of particles that move along lattice links rather than tracking the individual particles. Their interactions at lattice sites are subject to the usual conservation laws. Dealing with the continuous probability density variables avoids the "exclusion" problem of lattice gases as well as their translational invariance difficulties. Furthermore, it proved possible to reduce the viscosity to arbitrarily low values in the lattice Boltzmann scheme, whereas the constraints of lattice gases made that impossible for models consisting of particles with only a few different velocities (Higuera *et al.,* 1989). Thus, lattice versions of the Boltzmann equation can reach high Reynolds-number flows, but at the cost of losing the fast integer-algebra aspect of lattice gases and the absolute numerical stability, since this is no longer a particle method. The method is numerically equivalent to a discretized solution of the Navier–Stokes equations, with similar limits on its numerical stability, although one might have thought that the existence of an H theorem for the Boltzmann equation would have significantly improved stability. These calculations require roughly the same amount of computer time as Navier–Stokes solutions.

5.3 Direct-Simulation Monte Carlo

A particle method has also been developed to solve the Boltzmann equation (Bird, 1976). It has found wide use in rarefied-gas dynamics and has supplemented wind tunnel tests, for example, in airplane design. In this method the particle positions and velocities are described by continuous variables as in molecular dynamics. Collisions are executed stochastically by having particles assigned to cells of approximately a mean free path in size. Collisions among the particles within that cell are selected with the Boltzmann predicted probability, according to the relative velocity of randomly selected pairs and their collision cross section. After the collision, the pair remains in place but with stochastically determined post-collision velocities. The number of collisions is set by the collision rate and the selected time step. After that, all particles are advanced by about one mean free path, after which the collision process within a cell is repeated. In effect this algorithm ignores spatial correlations over a distance of one mean free path, consistent with the assumption of the Boltzmann approach. The procedure is about three orders of magnitude faster than a corresponding molecular-dynamics calculation. By having a full spectrum of velocities, the scheme has none of the limitations of lattice gases or lattice Boltzmann, and even the assumption of an ideal-gas equation of state can be removed by a small modification of this procedure (Alexander *et al.,* 1995). In practice it has not been possible to reach high Reynolds-number flows with this method.

6. MACROSCOPIC PARTICLE DYNAMICS

A wide variety of systems containing macromolecular particles can be studied by molecular dynamics and related methods. Examples of such systems include polymer solutions, suspensions, microemulsions, fluidized beds, grain silos, powder hoppers, and slurries. Particle sizes vary considerably, from less than 0.1 μm in fine-particle suspensions to granular particles of sizes up to a few centimeters. What distinguishes these calculations from molecular dynamics is that collisions between macromolecular particles are inelastic. Energy is absorbed, either by the background fluid or by internal degrees of freedom within the particle itself; in either case the energy dissipated has little effect on the subsequent dynamics, but the loss of kinetic energy leads to complex phenomena. Inelastic collisions can easily promote the formation of transient, but large-scale, clusters of particles. It is this clustering that is the primary cause of the complex dynamics. Three important examples of macroscopic particle dynamics are discussed here: dilute solutions of polymers, dilute to dense suspensions of hydrodynamically interacting spheres, and dense systems of granular particles.

6.1 Polymers and Brownian Dynamics

The simplest model of a polymer chain in dilute solution is the Zimm model, in which the polymer backbone is approximated by a chain of beads connected by harmonic springs. The beads interact with one another via a short-range excluded-volume potential. The model also takes into account the effects of the solvent by a mean-field or "preaveraged" approximation of the hydrodynamic forces acting between beads. The Zimm model successfully accounts for many of the experimentally observed scaling laws for dilute polymer solutions. More realistic models require numerical simulation.

Polymer solutions can be numerically modeled by "Brownian dynamics" in which the beads representing the monomer units of the polymer chain are represented by Brownian particles subject to random forces as well as viscous and intermolecular forces. Brownian dynamics is used to describe the behavior of macromolecules so large and massive that their position and velocity can be considered as frozen during collisional momentum transfer with solvent particles. In other words, the Brownian particle experiences many collisions with solvent particles before it changes its velocity significantly. The dynamics of a Brownian particle are stochastic, as described by the Langevin equation (Ermak and McCammon, 1978). For a single sphere

$$m\dot{u} = -\xi u + R(t) + F, \qquad (8)$$

where $\xi = 6\pi\eta a$ is the friction coefficient describing the drag on a sphere of radius a in a solvent of viscosity η, R is the random force representing the effect of collisions between solvent molecules and the Brownian particle, and F is force due to interparticle interactions. The friction constant used is that appropriate for "sticky" collisions, since fluid in contact with a macroscopic solid-particle surface is found to move with the local velocity of the contact point. In a suspension containing many Brownian particles or in a single polymer, the presence of neighboring spheres introduces a many-body hydrodynamic interaction, so that the force on each sphere includes significant contributions from every other sphere (Ladd, 1988).

6.2 Hydrodynamic Interactions

Hydrodynamic interactions are the forces between macroscopic particles arising from their motion; they are transmitted by the fluid in which they are suspended (Happel and Brenner, 1986). If the macroscopic particles have sufficiently small velocities, the fluid flow can be described by low Reynolds-number or "creeping-flow" hydrodynamics. There is then a linear relationship between particle forces and particle velocities,

$$u_i = \sum_{j=i}^{N} \mu_{ij} F_j. \qquad (9)$$

The mobility matrix μ_{ij} depends on the configuration of all the spheres in the system. An extension of Eq. (9) to include particle torques and particle rotations can readily be made. Given the complete set of mobilities for translation and rotation, it is then possible to solve for the particle dynamics. However, the mobilities are difficult to calculate in general, so that in many instances hydrodynamic interactions are neglected altogether: this leads to the isolated-sphere results represented by Eq. (8). These are strictly valid only at infinite dilution. Dilute suspensions can be treated fairly simply by using the mobility tensor for an isolated pair of spheres, but even then analytic expressions are only valid at large separations of the pair. When the pair of particles is near contact, the character of the interaction changes; the critical length is then the gap between the neighboring surfaces rather than the interparticle separation. This region is described by lubrication theory.

For denser suspensions it is not possible to enumerate all the many-body hydrodynamic interactions analytically, but direct numerical calculations can determine the mobility matrix for arbitrary configurations of spheres (Brady and Bossis, 1988). This has made it possible to compute the hydrodynamic properties (permeability, sedimentation velocity, self-diffusion coefficient, and viscosity) of random dispersions of rigid spheres, at any concentration, with a high degree of accuracy and reliability (Ladd, 1990). Much less work has been done on the computationally more demanding task of determining the changes in suspension microstructure under imposed external flows. Semiquantitative results have been obtained for small numbers of spheres in Couette flow and channel flow and for bulk sedimentation. Because the above simulations are so time-consuming, it is natural to try the lattice-gas/lattice-Boltzmann models discussed earlier. The simplicity and robustness of the algorithms has made it possible to handle the complex boundary conditions that occur in particle suspensions with relative ease (Ladd, 1994).

6.3 Granular Particle Dynamics

Despite extensive empirical experience, the dynamics of granular flow is poorly understood. There is both a scientific challenge and a technological need to develop an understanding of the mechanisms underlying the flow of grains. In a typical granular flow, such as the emptying of a hopper, there are regions of kinetically excited flows, where the grains have a random "thermal" component to their velocities, and stagnant regions, where the grains are quiescent and in constant contact with one another. Numerical simulation requires a formulation of grain–grain and grain–wall interactions in order to predict these macroscopic flow properties. The interactions between grains or between a grain and a container wall depend on the specific contact mechanics of the two surfaces. The important qualitative feature of granular particle dynamics is that, at each collision between grains, some of the kinetic energy of the colliding pair is lost. This energy is absorbed as heat, by sound waves inside the grains, and by irreversible deformation (plastic flow). Such systems eventually fall into a stagnant rest state with no particle motion, unless energy is supplied from an outside source, for instance, a shearing flow or a gravity flow. More realistic models of grain mechanics allow for distortion of the grains, for rolling and sliding friction between the contact points, and for adhesive surface forces (Walton and Braun, 1986) (see Fig. 10).

7. AB INITIO MOLECULAR DYNAMICS

It is apparent that the fundamental limitation of molecular dynamics is the parametrization of the force field in which the atoms move. In condensed phases many-body forces are not negligible, even for the simplest materials, and must be absorbed into an effective few-body potential. There are two drawbacks to this procedure: first, an effective potential depends on the thermodynamic state, so that different potentials are necessary for dilute and condensed phases; second, this effective potential must be determined by extensive fitting to experimental

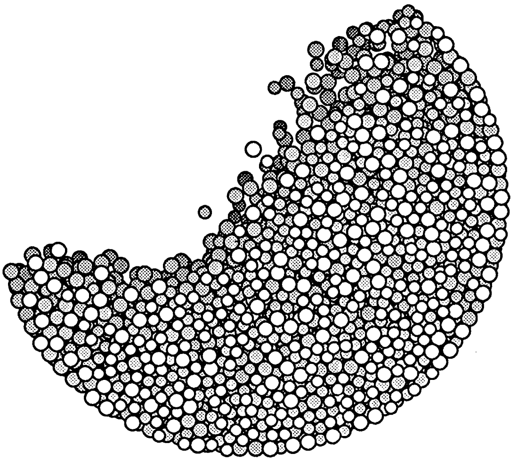

FIG. 10. Simulation of inelastic, frictional spheres in a rapidly rotating horizontal drum (not shown). Coefficient of friction, 0.2; coefficient of restitution for two-body normal impacts, 0.85. The rotation rate is such that at approximately six particle diameters from the outer wall the centrifugal acceleration is equal to the gravitational acceleration. The slice is approximately 5 diameters thick with periodic boundaries along the horizontal axis.

data (Vitek and Srolovitz, 1989). This is a cumbersome, ill-defined, and imprecise prescription. Instead, it would be desirable to introduce an accurate quantum mechanical calculation of the forces, particularly for metals, where pairwise additivity is even qualitatively inadequate. This requires that, for a given instantaneous configuration of the nuclei, the Schrödinger equation be solved to determine the spatial distribution of the electron density. Once the electron density is known, the interatomic forces can be calculated and the (classical) equations of motion of the nuclei solved for a small time step. This approach makes it possible to determine the properties of materials *ab initio,* requiring only the atomic number of the nuclei of the various elements making up the substance.

7.1 The Local-Density Approximation (LDA)

Despite decades of effort, neither the traditional variational methods of molecular quantum mechanics nor the more recent quantum Monte Carlo techniques (Ceperley and Kalos, 1986) have succeeded in developing a practical scheme for solving the Schrödinger equation with sufficient accuracy, for more than a few electrons. For systems with many electrons, the local-density approximation is the most widely used method in *ab initio* molecular dynamics (Car and Parrinello, 1985). It is chosen for its conceptual simplicity and predictive accuracy. Although density-functional theory was originally developed for metals, it can be used to describe covalent bonding as well.

Density-functional theory rests on the principle that the exact ground-state energy is a minimum of the total energy with respect to all possible variations in the electron density. However, the scheme is only computationally viable in the independent-electron or mean-field approximation. In practice this means that the total charge density is expressed in terms of noninteracting wave functions (orbitals). Missing from this description are the effects of electron exchange and electron–electron correlations. However, this exchange-correlation energy has been accurately calculated for a uniform electron gas by quantum Monte Carlo (Ceperley and Alder, 1980). In the LDA approach it is assumed that the exchange-correlation energy at each space point in the inhomogeneous electron system around a nucleus can be replaced by the exchange correlation of the uniform electron gas at the appropriate local electron density. Further corrections caused by the exchange and correlation energy due to the local density gradient are being attempted (Perdew and Wang, 1992). With this approach, systems of 100–1000 atoms can be accurately simulated (Brommer *et al.,* 1992).

7.2 Car–Parrinello Method

The Car–Parrinello method allows explicit treatment of the electronic degrees of freedom. This is accomplished by treating the electronic degrees of freedom as classical particles by incorporating their motion into the Lagrangian. The orbitals of the electronic wave function are assigned a fictitious mass, which serves to make the time scale of the variations in electron density slower than in reality, but still fast compared with nuclear motion (Payne *et al.,* 1992). The wave function thus follows the Born–Oppenheimer (adiabatic) surface staying near the true electronic ground state at all times. Nevertheless, a rather small time step ($\sim 10^{-15}$ s) is still required. More recently the alternative idea of fully relaxing the electronic degrees of freedom to the ground state at every time step has been implemented. The extra cycles necessary to achieve that are compensated for by a less stringent time-step requirement.

The other important ingredient of the Car–Parrinello method is the expansion of the orbitals in terms of a plane-wave set, which requires the use of accurate pseudopotentials to represent the nuclei with the tightly bound core electrons. Otherwise, many basis functions are required to represent the rapid variation of the wave function near the ion core. These core states usually play only a minor role in determining the interatomic forces, and thus most simulations deal only with valence electrons. These interact with one another and with the atomic pseudopotential, which represents the effective interaction due to the nucleus and core electrons. Most importantly, the Car–Parrinello method provides, for the first time, a practical way of coupling rearrangements of the atomic structure and the electronic structure. It is now possible to follow changes in electronic structure with changes in the thermodynamic state as, for example, when a substance melts (Sugino and Car, 1995). Recent applications include properties of amorphous solids, the motion of defects in crystals, and surface reconstruction. Materials studied include covalently bonded structures, semiconductors, and metals.

8. PROSPECTS

Applications of molecular dynamics to condensed systems have mushroomed from its inception, when simple systems were studied, to the investigation of increasingly more complex systems. Simple systems allowed theoretical concepts to be tested quantitatively and, in analogous fashion to real experiments, led either to verification or to modification of theories. Molecular dynamics and closely related methods are still being applied to relatively idealized problems, but in an effort to reach larger distance and time scales. Examples of such problems are dense polymeric systems, glasses, and particle suspensions. Although computer power is increasing rapidly, such calculations usually require algorithmic innovations—for instance, the use of lattice-gas and lattice-Boltzmann models to describe the fluid in two-phase solid–fluid flows. In some cases it may by that there is no substitute for brute-force computing; here parallel computation is often essential.

Another area where large distance and time scales are necessary is in hydrodynamic flow problems, such as rarefied-gas flows, which cannot be described by the Navier–Stokes equations. Here, another variant of molecular dynamics, the Direct Simulation Monte Carlo (DSMC) or Bird method, is useful. It can simulate a wide range of gas flows, on much larger length and time scales than is possible with conventional molecular dynamics. The range of applicability of DSMC and related methods can be enhanced by embedding them in continuum finite-difference or finite-element codes. Thus a Direct Simulation Monte Carlo algorithm (or perhaps even molecular dynamics) could be used to simulate "critical" regions of the problem, such as narrow gaps, less than or comparable to the collisional mean free path, where Navier–Stokes breaks down. The DSMC code can be directly coupled to a continuum code that would be used to simulate the large-scale motions, at enormous computational savings compared with a fully molecular simulation. Such "adaptive" algorithms, where the computational method varies with the local physics, could be very important in simulating a much wider class of fluid flows—for example, flows with very steep gradients or where realistic representations of boundaries are required.

Realistic simulations of complex molecules are even more difficult. In addition to the problems associated with the multiple time scales, there is very little reliable information about the intermolecular forces in real materials. Even if intermolecular forces can be calculated, there is the further difficulty of devising a suitable functional form for the potential; this is especially problematic in systems with large many-body forces. As a result, it seems clear that *ab initio* methods, which calculate the intermolecular interactions *in situ*, will become increasingly important. However, although density-functional methods have been used to calculate some dynamical properties, like diffusion coefficients in liquid metals, much more needs to be done to include electron correlations accurately, and beyond that, to allow electron transport properties, like thermal and electrical conduction, to be determined.

GLOSSARY

Adiabatic: A process where the system is in equilibrium at each stage; thus a reversible process with no change of entropy.

Boltzmann Equation: A kinetic theory of transport that accounts for collisions by the molecular chaos (or Markov) approximation, assumed to be valid for low-density gases.

Cellular Automata: A computational scheme in which neighboring cells of a regular lattice exchange only binary bits of information.

Couette Flow: A flow between a stationary plate and a slowly moving one.

Ensemble: A collection of macroscopically equivalent equilibrium states in which a given set of thermodynamic variables are held fixed while the others are allowed to fluctuate.

Ergodicity: A condition in which time averages are equivalent to ensemble averages.

Markov Process: A process where the probability of an event is independent of preceding events.

Mean Field: An approximation where the effect of the neighbors on a given particle depends only on the average (mean) force they exert.

Mean Free Path: The average distance a particle travels between successive collisions.

Monte Carlo Method: A stochastic method to evaluate ensemble averages.

Navier–Stokes Equation: The equation of macroscopic hydrodynamics, consisting of the conservation laws and the linear dissipation assumption.

Rheology: The study of nonequilibrium properties of highly viscous media where the transport is not linear in the applied gradients.

Reynolds Number: The dimensionless number describing the flow rate, uL/v, where u is the velocity, L is the characteristic dimension of the flow, and v is the kinematic viscosity.

Stokes Drag: The frictional or drag force on a macroscopically large sphere falling freely in a fluid.

Works Cited

Alder, B. J., Wainwright, T. E. (1958), in: I. Prigogine (Ed.), *Transport Processes in Statistical Mechanics,* New York: Interscience, pp. 97–131.

Alder, B. J., Wainwright, T. E. (1959a), *J. Chem. Phys.* **31,** 459–466.

Alder, B. J., Wainwright, T. E. (1959b), *Sci. Am.* **201** (4), 113–123.

Alder, B. J., Wainwright, T. E. (1962), *Phys. Rev.* **127,** 359–361.

Alder, B. J., Wainwright, T. E. (1963), in: *Conference on the Many-Body Problem,* Interscience Publishers, New York, p. 97.

Alder, B. J., Wainwright, T. E. (1969), *J. Phys. Soc. Jpn.* **26,** 267–269.

Alder, B. J., Wainwright, T. E. (1970), *Phys. Rev. A* **1,** 18–21.

Alexander, F. J., Garcia, A. L., Alder, B. J. (1995), *Phys. Rev. Lett.* **74,** 5212–5215.

Alley, W. E., Alder, B. J. (1980), in: M. Balaban (Ed.), *Molecular Structure and Dynamics,* Philadelphia: International Science Services.

Alley, W. E., Alder, B. J. (1983), *Phys. Rev. A* **27,** 3158–3173.

Alley, W. E., Alder, B. J. (1984), *Phys. Today* **37** (1), 56–63.

Andersen, H. C. (1980), *J. Chem. Phys.* **72,** 2384–2393.

Ashurst, W. T., Hoover, W. G. (1975), *Phys. Rev. A* **11,** 658–678.

Barker, J. A. (1986), *Mol. Phys.* **57,** 755–760.

Barker, J. A., Henderson, D., Abraham, F. F. (1981), *Physica* **106A,** 226–238.

Bennet, C., Alder, B. J. (1971), *J. Phys. Chem. Solids* **32,** 2111–2122.

Berendsen, H. J. C. (1986), "Biological Molecules and Membranes," in: G. Ciccotti and W. G. Hoover (Ed.), *Molecular Dynamics Simulation of Statistical Mechanical Systems,* Proceedings of the Enrico Fermi International School of Physics Summer School, Varenna, Bologna: Società Italiana de Fisica, pp. 496–519.

Berendsen, H. J. C., Van Gunsteren, W. F. (1986), in: G. Ciccotti and W. G. Hoover (Ed.), *Molecular Dynamics Simulation of Statistical Mechanical Systems,* Proceedings of the Enrico Fermi Summer School, Varenna, Bologna: Società Italiana de Fisica, pp. 43–65.

Bird, G. A. (1976), *Molecular Gas Dynamics,* London: Oxford University Press.

Bishop, T., Schulten, K. (1994), in: G. Wipff (Ed.), *Computational Approaches in Supramolecular Chemistry,* Boston: Kluwer Academic Publishers.

Brackbill, J. U., Cohen, B. I. (Eds.) (1985), *Multiple Time Scales,* New York: Academic.

Brady, J. F., Bossis, G. (1988), *Annu. Rev. Fluid Mech.* **20,** 111–155.

Brommer, K. D., Needels, M., Larson, B. E., Joannopoulos, J. D. (1992), *Phys. Rev. Lett.* **68,** 1355–1358.

Brooks, C. L., III, Karplus, M., Pettitt, B. M. (1988), *Proteins: A theoretical perspective of dynamics, structure, and thermodynamics,* Advances in Chemical Physics No. 71, New York: Wiley.

Brush, S. G., Sahlin, H. L., Teller, E. (1966), *J. Chem. Phys.* **45,** 2102–2118.

Car, R., Parrinello, M. (1985), *Phys. Rev. Lett.* **55,** 2471–2474.

Ceperley, D., Alder, B. J. (1980), *Phys. Rev. Lett.* **45,** 566–569.

Ceperley, D. M., Kalos, M. H. (1986), in: K. Binder (Ed.), *Monte Carlo Methods in Statistical Physics,* 2nd ed., Topics in Current Physics, Vol. 7, Springer: Berlin, pp. 145–194.

Ciccotti, G., Jacucci, G., McDonald, I. R. (1979), *J. Stat. Phys.* **21,** 1–22.

Ciccotti, G., Ferrario, M., Ryckaert, J.-P. (1982), *Mol. Phys.* **47,** 1253–1264.

Connick, R. E., Alder, B. J. (1983), *J. Phys. Chem.* **87,** 2764–2771.

de Leeuw, S. W., Perram, J. W., Smith, E. R. (1980), *Proc. R. Soc. London, Ser. A* **373,** 57–66.

Dorfman, J. R., Cohen, E. G. D. (1975), *Phys. Rev. A* **12,** 292–316.

Dymond, J., Alder, B. J. (1966), *J. Chem. Phys.* **45,** 2061–2068.

Ermak, D. L., McCammon, J. A. (1978), *J. Chem. Phys.* **69,** 1352–1360.

Evans, D. J. (1982), *Phys. Lett.* **91A,** 457–460.

Fermi, E., Pasta, J. G., Ulam, S. M. (1965), in: E. Amaldi, H. L. Anderson, E. Persico, F. Rasetti, C. S. Smith, A. Wattenburg, and E. Sergré (Eds.), *Collected Works of Enrico Fermi,* Vol. II, Chicago: University of Chicago Press.

Frenkel, D., Ladd, A. J. C. (1984), *J. Chem. Phys.* **81,** 3188–3193.

Frisch, U., Hasslacher, B., Pomeau, Y. (1986), *Phys. Rev. Lett.* **56,** 1505–1508.

Frisch, U., d'Humières, D., Hasslacher, B., Lallemand, P., Pomeau, Y., Rivet, J.-P. (1987), *Complex Syst.* **1,** 649–707.

Gehlen, J. N., Mardin, M., Chandler, D. (1994), *Science* **263,** 499–502.

Greengard, L. (1994), *Science* **265,** 909–914.

Happel, J., Brenner, H. (1986), *Low-Reynolds Number Hydrodynamics,* Dordrecht: Martinus Nijhoff.

Higuera, F., Succi, S., Benzi, R. (1989), *Europhys. Lett.* **9,** 345–349.

Hoover, W. G., Evans, D. J., Hickman, R. B., Ladd, A. J. C., Ashurst, W. T., Moran, B. (1980), *Phys. Rev. A* **22,** 1690–1697.

Hoover, W. G., Ladd, A. J. C., Moran, B. (1982), *Phys. Rev. Lett.* **48,** 1818–1820.

Jacucci, G., Rahman, A. (1984), *Nuovo Cimento* **D4,** 341–356.

Karplus, M. (1986), in: E. Clementi, S. Chin (Eds.), *Structure and Dynamics of Nucleic Acids, Proteins, and Membranes,* New York: Plenum, pp. 113–126.

Koplik, J., Banavar, J. R. (1995), *Annu. Rev. Fluid Mech.* **27,** 257–292.

Kubo, R. (1957), *J. Phys. Soc. Jpn.* **12,** 570–586.

Ladd, A. J. C. (1988), *J. Chem. Phys.* **88,** 5051–5063.

Ladd, A. J. C. (1990), *J. Chem. Phys.* **93,** 3484–3494.

Ladd, A. J. C. (1994), *J. Fluid Mech.* **271,** 311–339.

Ladd, A. J. C., Alder, B. J. (1989), *J. Stat. Phys.* **57,** 473–482.

Ladd, A. J. C., Colvin, M. E., Frenkel, D. (1988), *Phys. Rev. Lett.* **60,** 975–978.

Lebowitz, J. L. (1993), *Physica A* **194,** 1–27.

Levesque, D., Verlet, L. (1993), *J. Stat. Phys.* **72,** 519–537.

Levesque, D., Verlet, L., Kürkijarvi, J. (1973), *Phys. Rev. A* **7,** 1690–1700.

Lifson, S., Hagler, A. T., Dauber, P. (1979), *J. Am. Chem. Soc.* **101,** 5111–5121.

Madelung, E. (1918), *Z. Phys.* **19,** 524–532.

Maitland, G. C., Rigby, M., Smith, E. B., Wakeham, W. A. (1981), *Intermolecular Forces: Their Origin and Determination,* Oxford: Claredon Press.

McCammon, J. A., Gelin, B. R., Karplus, M. (1977), *Nature* **267,** 585–590.

McNamara, G. R., Zanetti, G. (1988), *Phys. Rev. Lett.* **61,** 2332–2335.

Meiburg, E. (1986), *Phys. Fluids* **29,** 3107–3113.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. (1953), *J. Chem. Phys.* **21,** 1087–1092.

Nosé, S. (1984), *J. Chem. Phys.* **81,** 511–519.

Nosé, S., Klein, M. L. (1983), *Mol. Phys.* **50,** 1055–1076.

Palmer, R. G., Stein, D. L., Abrahams, E., Anderson, P. W. (1984), *Phys. Rev. Lett.* **53,** 958–961.

Paul, G. L., Pusey, P. N. (1981), *J. Phys. A* **14,** 3301–3327.

Payne, M. C., Teter, M. P., Allan, D. C., Arias, T. A., Joannopoulos, J. D. (1992), *Rev. Mod. Phys.* **64,** 1045–1097.

Perdew, J. P., Wang, Y. (1992), *Phys. Rev. B* **46,** 12947–12954.

Plimpton, S. (1995), *J. Comp. Phys.* **117,** 1–19.

Rahman, A. (1964), *Phys. Rev. A* **136,** 405–411.

Rahman, A., Stillinger, F. H. (1971), *J. Chem. Phys.* **55,** 3336–3359.

Rao, M., Berne, B. J. (1979), *J. Chem. Phys.* **71,** 129–132.

Rapaport, D. C. (1988), *Phys. Rev. Lett.* **60,** 2480–2483.

Shakhnovich, E. I. (1994), *Phys. Rev. Lett.* **72,** 3907–3910.

Siepmann, J. I., Frenkel, D. (1992), *Mol. Phys.* **75,** 59–70.

Sinai, Ya. G. (1973), in: E. G. D. Cohen, W. Thirring (Eds.), *The Boltzmann Equation,* Vienna: Springer-Verlag.

Smit, B., Hilbers, P. A. J., Esselink, K., Rupert, L. A. M., van Os, N. M., Schlijper, A. G. (1990), *Nature* **348,** 624–625.

Sugino, O., Car, R. (1995), *Phys. Rev. Lett.* **74,** 1823–1826.

Swendsen, R. H., Wang, J. S., Ferrenberg, A. M. (1992), in: K. Binder (Ed.), *The Monte Carlo Method in Condensed Matter Physics,* Berlin: Springer.

Tuck, J. L., Menzel, M. T. (1972), *Adv. Math.* **9,** 399–407.

Uhlenbeck, G. E., Hemmer, P. C., Kac, M. (1963), *J. Math. Phys.* **4,** 229–247.

Vitek, V., Srolovitz, D. J. (Eds.) (1989), *Atomistic Simulations of Materials: Beyond Pair Potentials,* New York: Plenum Press.

van der Hoef, M. A., Frenkel, D. (1991), *Phys. Rev. Lett.* **66,** 1591–1594.

Verlet, L. (1967), *Phys. Rev.* **159,** 98–103.

Walton, O. R., Braun, R. L. (1986), *J. Rheol.* **30,** 949–980.

Wood, W. W. (1970), *J. Chem. Phys.* **52,** 729–741.

Further Reading

Allen, M. P., Tildesley, D. J. (1987), *Computer Simulation of Liquids,* Oxford: Clarendon. A more advanced textbook devoted to the foundation and application of molecular dynamics to liquids.

Brooks, C. L., III, Karplus, M., Pettit, B. M. (1988), *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics,* Advances in Chemical Physics No. 71, New York: Wiley. A more advanced textbook devoted to proteins.

Cicotti, G., Hoover, W. G. (Eds.) (1986), *Molecular Dynamics Simulation of Statistical Mechanical Systems,* Proceedings of the International School of Physics "Enrico Fermi," Course 97, Amsterdam: North-Holland. Advanced topics in molecular dynamics by some of the principal contributors to the field.

Cicotti, G., Frankel, D., McDonald, I. R. (Eds.) (1987), *Simulation of Liquids and Solids,* Amsterdam: Elsevier. A good collection of pioneering papers in the field.

Gould, H., Tobochnik, J. (1988), *An Introduction to Computer Simulation Methods/Application to Physical Systems,* Pts. 1 and 2, Reading, MA: Addison-Wesley. A general book intended to be used in a course.

Hansen, J. P., McDonald, I. R. (1990), *Theory of Simple Liquids,* 2nd ed., London: Academic. Another more advanced textbook devoted to the foundation and application of molecular dynamics to liquids.

Heermann, D. W. (1990), *Computer Simulation Methods in Theoretical Physics,* 2nd ed., Berlin: Springer. An elementary textbook introduction to molecular dynamics.

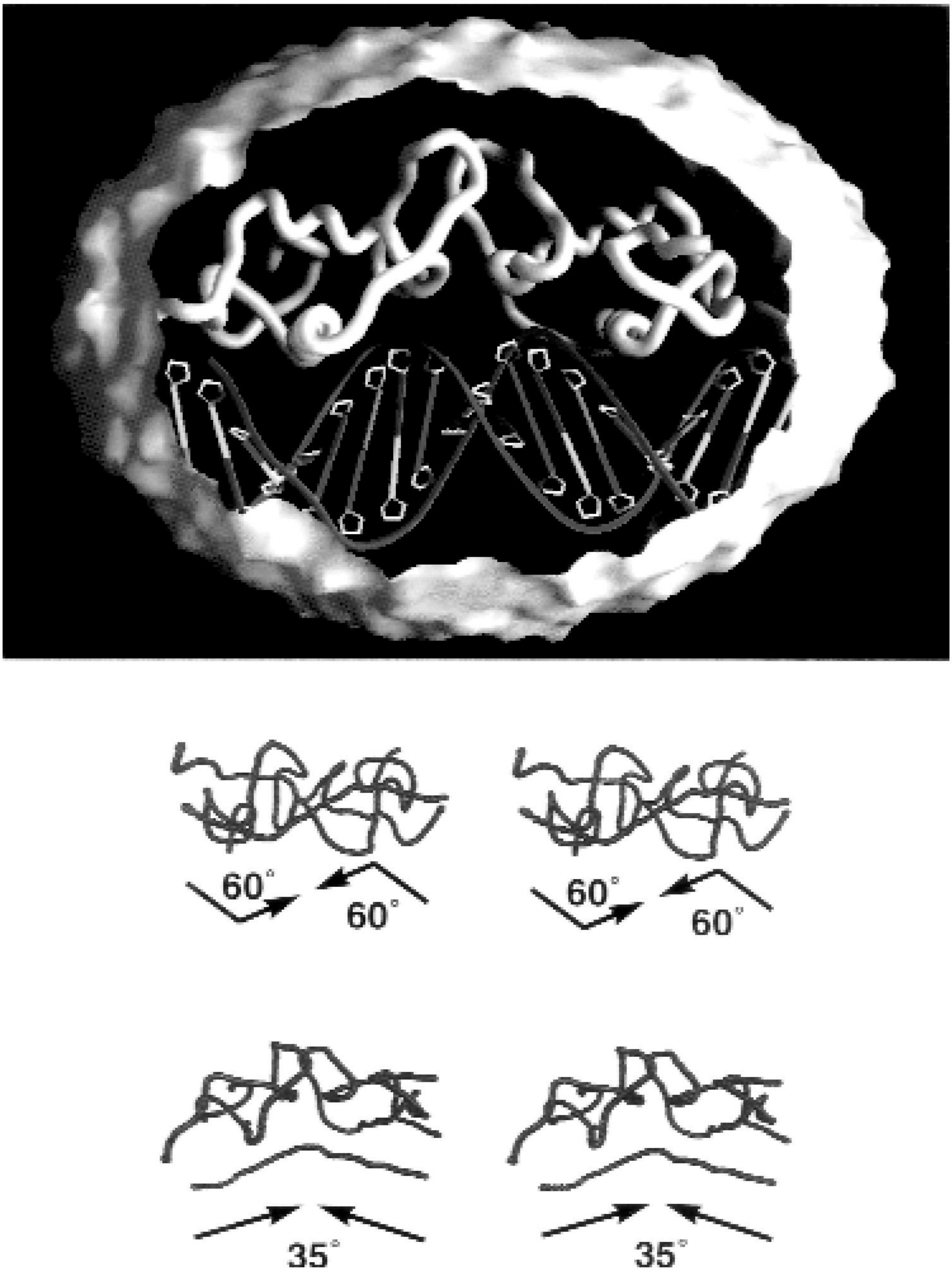

Vol. 18 – PLATE 1. (a) Cutaway view of the solvent (water) surface of the protein-DNA complex. The white "worm" represents two identical proteins (glucocorticoid receptor) facing each other, which bind to a specific sequence of the DNA (color) structure. (b) Line representations, in cross-eyed stereo, of the DNA-protein complex after 70 ps of molecular-dynamics simulation. The system is viewed with the DNA located behind the protein (top) and with the DNA below the protein (bottom). The DNA and the two monomers are represented schematically through a single line, each of which traces the axis of the DNA and the backbone of the protein, respectively. The views illustrate how the DNA is bent through the interaction with the protein dimer; the solid black lines indicate the deviation of the DNA strand from its ideal linear geometry.

SOFTWARE ENGINEERING

HERBERT J. BERNSTEIN, *Bernstein+Sons, Bellport, New York, U.S.A.*

	Introduction	305
1.	**Background**	308
1.1	The Problem	308
1.2	The Context of Design and Development	309
1.3	Tools Used in Solving the Problem	312
1.3.1	Conceptual Disciplines	312
1.3.2	Selection of Appropriate Tools	312
1.3.3	Human Factors	313
1.3.4	Standards and Controls, QA/QC	313
2.	**Basic Concepts**	313
2.1	What Is Software?	314
2.2	Types of Software	314
2.2.1	Systems Software	314
2.2.2	Application Development Support	314
2.2.3	Applications	315
3.	**Software Engineering**	315
3.1	Software Design Cycles	315
3.2	Software Design Paradigms	315
3.2.1	Bottom-Up, Top-Down, Outside-In	316
3.2.2	Functional and Applicative Programming	318
3.2.3	Structured Programming	319
3.2.4	Proving Correctness	320
3.2.5	Software Engineering and ADA	320
3.2.6	CASE Tools	320
3.2.7	Object-Oriented Software	321
3.2.8	Usability—Testing and Prototyping	322
3.2.9	Requirements Engineering	323
3.2.10	Re-engineering	323
4.	**Security Issues**	324
5.	**Other Issues**	325
5.1	Issues in Maintenance	325
5.2	Reuse of Software	325
5.3	Decision Support	326
6.	**Support for Groups**	327
6.1	Organizational Issues	327
6.2	Including the Client in the Design Process	328
7.	**Quality Assurance and Control**	329
8.	**Examples of Tools and Techniques**	330
8.1	Communications Networks	330
8.2	Graphical User Interfaces	333
9.	**Future Directions**	337
	Acknowledgment	338
	Glossary	338
	Works Cited	339
	Further Reading	340

INTRODUCTION

Software engineering is the disciplined use of well-understood techniques to analyze, design, build, repair, or use software. The main focus of software engineering is on building new software systems by analyzing system requirements and then by designing and developing software to satisfy those requirements. As is true of all the engineering professions, software engineering provides the interface between technology and its uses within society.

Much of this article will be devoted to organizational issues and human factors. This reflects the state of the software engineering profession at the time of writing this article. Only a few years before, there had been a widespread belief that automata would rapidly be able to replace people in most of the detailed work of software construction, and the focus of software engineering had been on suites of automated tools that would write software without human intervention. Those approaches failed to provide reproducible results, and most of the effort in software engineering has now turned back toward understanding how people work effectively.

As a young field, software engineering is

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

still defining itself, still learning how to build the tools it needs, learning how to build reliable software. In design and development, the goal of software engineering is to produce software that performs as intended, performs reliably, is delivered on schedule, and is delivered at predictable and reasonable cost. There have been serious problems in achieving these goals. There is an ongoing effort to improve the approaches we now use, but the basic concepts of top-down analysis, bottom-up synthesis, object-oriented representation, and the use of software to process software are, and are likely to remain, the core of software engineering. Approaches to analysis, repair, and operation of existing software systems are strongly influenced by the major approaches used in design and development.

In this article we concentrate on the problems and solutions of interest to physicists. We look at the engineering approaches that are important for producing large scientific production codes and for producing reliable data-acquisition and control systems for experiments. This causes us to look at a much wider range of issues than might be the case in discussing software engineering in the context of business applications. We give less emphasis to any single paradigm, such as object orientation, than to basic engineering approaches.

In this context, the line between engineering approaches needed for software and those needed for hardware blurs. It is, however, important to understand what software is and when it is appropriate. Computers, communications media, detector arrays, and even people are the hardware, the tangible components of information-processing systems. Software is the intangible part, the sets of instructions and lists of data that determine what the tangible components will do. Software is the means by which we organize and adapt appropriate hardware to specific tasks.

Especially in data-acquisition and control systems, it is common to be faced with multiple viable alternative solutions using complex combinations of hardware and software, some relying primarily on hardware, some relying primarily on software, and most somewhere in the middle. How do we choose among such alternatives?

Software is intangible. Hardware is tangible. If we process information with an assembly of physical components, we are using a tangible system for information processing. We can touch the parts, examine them, and determine what the system will do with information. In a laboratory, we might take a signal from a detector, amplify it by a set gain, send it through a low-pass filter, and derive new information from the signal. We use well-understood techniques to select the cables, match the amplifier to the signal, and choose component values for the filter. We apply those techniques with sufficient care and sufficient discipline to have a reasonable assurance that what we build will perform in a predictable manner. Assembling electronic components to process data is an example of electrical engineering.

Now suppose we need a more complex filtration of the signal. Perhaps we need a Fourier transform to help detect certain interesting frequency components. We could feed the signal to a computer via an analog-to-digital converter and write a program to transform the signal. When it is running, the program does not have an obvious tangible expression in the computer. Software is information (a concept with powerful implications, since we are working with information-processing systems). Software changes the computer, making it behave as if it were a different machine than we would deduce from simple examination of the raw hardware. By loading the program into a computer, we have made a virtual machine. Instead of writing a program, we could have built hardware to perform the same information-processing tasks. Though either approach could have been used to solve the problem, using software is often more cost effective. Even though planning the logical sequence of actions to be taken may require the same amount of work for both the real machine and the virtual machine, loading instructions and data into a computer is faster and simpler than wiring a circuit, and it is much easier to modify a program than it is to rewire a circuit. Therefore we choose to build the virtual machine. There is a much better chance that the virtual machine will function reliably if we follow engineering approaches, and discipline ourselves to use well-understood techniques to design the code, instead of compounding the experiment we are performing on the signal with an experiment in programming. For example, we might use a well-tested Fourier-transform code and run it under a proven operating system. We might take care to plan and document the program in a clearly organized and structured manner before writing code. Such disciplined use of well-understood programming techniques is an example of software engineering.

Not all problems can be solved this way. The first time that problems in software design or analysis are faced there may be no well-understood techniques to apply, or the problem may prove to us that techniques we thought we understood we did not understand at all. Then artistry, creativity, and the courage to try and to fail may be needed to chart new territory. Many mistakes may be made, but, usually, with time and patience, the parameters of the problem and the tools needed to solve the problem become well understood. Creating software for similar problems becomes less of a creative exercise, becomes more routine, more predictable, more mechanized, becomes engineering rather than artistry. When we are in the experimental, creative, artistic mode of programming, we may be doing interesting research of great value to future efforts, a necessary prerequisite to future software engineering, but we are not doing software engineering. When we can approach a software problem and know very early in the process how to solve the problem, how long it will take to solve the problem, and how reliable our solution is likely to be, then we are doing software engineering.

In making any estimates of how much effort it may take to solve a particular problem, it helps to step back from the problem itself and first answer some questions. In general, we need to identify precisely what is to be produced. Are we producing a product, a process, or a document? Who will use what we produce? What are the constraints on how to produce it and, once produced, on how it is to be used? Are there standards and codes we should follow? Who is the customer, and when is delivery to be made? Who pays (an often ignored question, but clearly those who pay must be satisfied, or the project has failed), and what are the financial constraints on the project? How do we tell if we have a success, and how do we fix things if there are problems? What happens when we are done?

Such questions sound fussy and mundane, much as if saying we had to go to an architect and make fancy blueprints every time we wished to make a small change in our house. Clearly there are times when just going ahead and remodeling a house by the seat of our pants will work and is the right thing to do, but there are many more times when it might be a disastrous mistake to do so. If it matters that the software we produce work, it is best to produce it in an organized, well-documented manner, to define clearly what we are trying to do and for whom we are trying to do it. Asking such questions encourages us to impose structure on the problem, to reduce the context we have to deal with at any given time to manageable scope, and, most importantly, to keep things simple.

If we plan first and act second, we have a chance to select the right tools to do the job; to select appropriate hardware platforms, appropriate paradigms, languages, and libraries; to consider the human factors involved so that we can match the skills of the developers to the tasks to be done and match the product produced to the needs of its users; to conform to the ever-growing maze of rules and regulations that surround what we do; and to produce a quality product and have some assurance that it actually works.

Among the most powerful tools that have been developed for software engineering are structured programming, which focuses on the top-down analysis—the hierarchical decomposition of complex actions into simpler actions with more limited context—and object-oriented representation, which focuses on interactions among the information structures (objects) involved and state-changing communication among objects. The structured programming approach has particular application to large systems performing numerical calculations, while object-oriented representation lends itself well to handling the kinds of multiple simultaneous interactions encountered in real-time data acquisition and experiment control, as well as in dealing with human control and display interfaces.

Despite the importance of these two approaches, it is not the choice of a particular technique that defines software engineering. We can no more say that software engineering is object-oriented programming or structured programming or any other popular technique than we can say that the only engineering approach to crossing a river is to design a suspension bridge. Just as some river-crossing problems are better solved with ferries than with bridges, some software design problems are better solved with procedural designs than with object-oriented ones. Converting a procedural matrix inversion on a single-processor SIMD (single-instruction, multiple-data, i.e., vector-processor) machine into object-oriented form may be entertaining, but it is not likely to be an optimal choice for solving the problem of inverting a matrix. The wider the range of tools in the software engineer's kit and the better our awareness of the appropriate domain of applicability of each tool, the more likely that the software engineer will find sound solutions to the problems we face. Software engineering is not defined by particular methods but by being methodical.

Because of the limitations of space and the need to focus on issues of methodology, we cannot address a full range of tools and techniques in detail. The tools of software engineering cover a wide range of arts and sciences. In this article we look at examples of two important types of tools: those drawn from the design of communications networks and those drawn from interactive graphics. Applying the mathematical models of queues and graphs that have long been used in the design of telephone systems, highways, and railroads to flows of information helps to provide the means to allocate necessary resources to tasks with minimal waste. Allowing designers and users to work with graphical representations of subsystems of information systems allows for a more natural design paradigm in many cases than does working with commands and lists of numbers.

1. BACKGROUND

A brief history of software engineering can be found under COMPUTER PROGRAMMING LANGUAGES. Here we examine the present state of the field. Why is there a need for a distinct field of software engineering? What problems need to be addressed for software design that differ from those in electrical engineering, digital engineering, or systems engineering? Software engineering has a unique perspective because of the remarkable flexibility of computers. We can change what a computer does much more quickly than we can build the equivalent physical systems. New programs can be loaded into a computer in seconds. This allows the same device to attack a wide range of problems or, more importantly, to attack immense problems one small portion at a time. In comparison to the problems they can solve, computers are simple devices. The problems computers can be used to solve are complex. Software, of necessity, carries the increased level of complexity of the problems being solved. The most challenging issues in software engineering arise because of the need to deal with complexity.

1.1 The Problem

Software engineering is, or at least aspires to be, one of the engineering professions, the one that brings the reliability of engineering approaches to problems that involve computer software. Engineers work at the so-called "socio-technological interface," the interface between science and technology on one hand and the practical needs of society on the other. In simple terms, engineers build things for other people to use. As Herbert Hoover said, "To the engineer falls the job of clothing the bare bones of science with life. . . ." (Wright, 1989, p. 49).

The goal of software engineering is to produce software that performs as intended, performs reliably, is delivered on schedule, and is delivered at predictable and reasonable cost. In order to be able to do that, the software engineer must understand computers and the applications to which they are put. The complexity of the tasks we are able to address with computers makes each of the goals of software engineering difficult to achieve. It is hard to build a reliable system when you do not understand what the system does and when many of the components you will use are themselves systems that are complex and poorly understood. This gives a high priority in software engineering to techniques that help in the management of complexity. The most important issues are psychological. Many people have a difficult time dealing with too much new or unrelated information at one time. They need to have the information context kept to something small—a page or two of text or a handful of distinct concepts—to understand what they are dealing with. Unless we are working on something very familiar, if what we intend to produce goes beyond that context we well may forget parts of the project from the beginning when we are working in the middle and the end, and produce something that, though it works, does not do what we had originally intended. Worse, because of this tendency to lose track of things, we may leave out critical cases and create errors. These conceptual errors and the errors in operation then have to get fixed, causing delays in the project, and making any original estimates of time and cost wrong. These are not abstract concerns. Many more large software projects have failed than have succeeded. The failures are rarely because of a lack of skill in programming, nor because of a lack of powerful tools. The failures are almost universally due to a failure to manage complexity, because the problem took on a life of its own and got away from the people trying to solve it.

However, abstract managerial skills are not sufficient. The engineer must have a solid understanding both of the tools to be used and of the technical details of the problem to be solved. Even more important than understanding what can be done with the resources at hand is to understand what cannot be done with the resources at hand, to have a quantitative understanding of rates and flows and design limitations.

Let us look in more detail at what software is and the environment in which software is produced.

1.2 The Context of Design and Development

Software is information that affects the processing of information. It may be data, instructions, images, etc. It may exist in a computer, on a piece of paper, or in someone's head. It may be a static product or a dynamic process, which is a system able to interact with other systems and change state. It may be a small piece of code created quickly and used by one person for a brief period or a system of millions of lines of code and documentation created by years of effort by teams of people and computers and used by millions of people over decades. It has a life, sometimes the brief life of a butterfly, sometimes the well-ordered life of a child born to great expectations, sometimes the tumultuous vitality of a prolific family, and sometimes a life that never leads anywhere.

The well-ordered life is rare, but easy to understand and illuminating. One begins with a need, a requirement for a system to do something useful with information, say, to perform a certain experiment. Just as the well-ordered personal life tends to arise in the context of generations of similar lives, well-ordered program development tends to arise within well-understood contexts. The need involved may never before have been seen, or never before have been automated, but it has strong family ties to earlier generations of problems or to similar problems that have been automated. The need exists within the context of the systems within which the need was perceived and of similar and related systems for which analysis has been done. We analyze the need within that context, trying to define precisely what it is we are asking for. Perhaps it is to control experimental apparatus, collect data, and try to fit that data by appropriate models. Having clearly defined the need, we carefully design information data structures, flows of information, and flows of control that we believe will meet those needs, and once that design is complete, we translate the abstractions of the design into a concrete implementation as computer code and instruction manuals. We validate and test what we have produced to convince ourselves that we have indeed met the need, and then, if all is well, we put the new system into use. After some period of use, we make some changes to the system, whether because of problems, changing needs, or a changing environment. This happens some number of times until the system somehow feels too old and patched, and, after a new system is built, the old system is laid to rest.

Such a software development life cycle itself defines an interesting information-processing system—one that accepts ideas, produces software, and then applies it. Therefore, it is challenging to attempt to produce software to support this system. As we discuss in Sec. 3.2.6, major portions of this system have been automated using "Computer Assisted Software Engineering" (CASE) tools.

This well-ordered flow from concept to design to construction to use is called the "waterfall model" (see, for example, Pressman, 1992, p. 24), in which analysis and design of the external system to which the software belongs, analysis of the software system, design of the software system, programming the software system, testing of the system, and maintenance of the system are shown as cascading flows from higher to lower levels, on the assumption that it requires considerable effort to make the process flow uphill. If this is the life cycle inherent in the system and in the organizations using it, then the reverse flows do require special care. This is especially true in large bureaucratic organizations with complex hierarchical decision making. The stately procession of steps in the waterfall model (and in the necessary reverse flows to deal with changes) can require a great deal of time.

Many applications, however, require a higher rate of progress or at least a higher rate of change, more in line with the image of a tumultuous family. In this environment new requirements and design concepts are continually being derived from experience with systems and rapidly changing external demands. As with the well-ordered life, the tumultuous family has many generations, but instead of coming neatly, one after the other, they are jumbled up with new family members coming on the scene and usurping the rights of the old before they think themselves old. This is common when there have been several cycles of earlier systems, many of which are still in use, and control over the development process no longer rests with a small group of decision makers but is driven by a democratic market process. Seeing the overall flow of development may be impossible in such chaos, and, in the large, the development environment becomes one in which psychology, group dynamics, and organizational issues dominate the technical issues of software development. Interacting design and implementation decisions may be made independently by many groups of players. We look at some of those issues in Sec. 6. However, in the small, the major fragments of the basic flows seen in the waterfall model remain important, and the related issues are worth considering.

The issues are much the same as in writing a good newspaper article: what, why, who, how, when, and where. At each step in the flow, we need to consider what to produce, why it should be produced, who will produce it for whom, and how, when, and where to produce it. In addition, in the real world, costs and who will pay them always enter the equation, and it is important to define clearly criteria and mechanisms for the customer to decide if we have done what was intended and to bring the software into actual use.

What to produce is a basic issue. In the waterfall model we produce an analysis, then a design, then a program, then a system in use. But what are we actually producing? Are we just producing more and more concrete versions of code or data for a computer? This is rarely the case. Especially when a process is being created, the real demands on the system are likely to be from people rather than computers. The software produced is not just the internal state of a computer. At the conceptual stage of analyzing needs, no code at all may be produced, just documents and conversations about the desired functioning of systems able to process information. Documentation is needed not just of the system in isolation but also as part of the broader systems it supports and with which it interacts.

In order to design software that will meet expectations and to decide if the software is functioning correctly, we need to know why the software is being produced and for whom (or for what). The consideration of why a system is being produced extends beyond understanding the context of systems within which it is to fit, to understanding the goals and purposes of those systems. That requires a clear understanding of who the customers and users are. If output of the system is input to equipment, we only have to consider the interface requirements of the equipment. If the results are intended for people, we have to consider the interface requirements of people, which can be very demanding. The study of people's requirements for use of machines is called ergonomics, and involves consideration of how people work effectively. For example, if we design a system requiring manual entry of certain data, it is important to consider the demands being made on the human operator in order to understand the chances of errors in that data entry process. The choices of keyboards, displays, and even chairs can be software engineering issues.

The user of the system is rarely the only entity imposing constraints on what is to be produced. Most projects exist within a complex context of management, rules, laws, and regulations. Independent of the technical requirements of the project at hand, external standards may be imposed on how code is to be produced and how it is to be used. Depending on the discipline, there may be legally enforceable rules, laws, and regulations that impose the use of certain programming languages and require special features in the system. For example, a government agency might require use of the programming language ADA and, for code controlling certain experiments, require doubly redundant control over safety shutters.

In addition, many projects are paid for by a customer who is not the user. The project is only partially successful if the user is satisfied, but the customer is not happy. The customer may not have the necessary background to decide if the project has done what it is supposed to do, and it often becomes part of the design process to produce necessary information. This layer of management information can be more complex than the project upon which it is providing reports.

Considering for whom the software is being produced constrains how it is to be produced but does not necessarily provide a viable mechanism that will meet those constraints. The constraints may leave no feasible means for production or may be so loose as to leave us without guidance on how to go forward. In the well-ordered environment we can look to the past and similar projects to find mechanisms to resolve this, but, in general, making the compromises on how to go forward and getting them accepted is critical to the success of the project.

Once it is clear what is to be delivered to whom and how it is to be done, the question of schedules arises. Because software has such a reputation for untimely delivery, it becomes important for large projects to track detailed timelines and subproducts and subprocesses. The dependences among tasks have to be considered as part of the project design. If *C* cannot be delivered until both *A* and *B* are done, then, in tracking the project, knowing that either *A* or *B* is late gives an early warning that *C* and everything dependent on *C* will also be late.

If we can understand the time lines and dependences, then often we can make a crude estimate of costs. Information-processing system costs have hardware and software components. The software costs usually dominate in large systems, and so tracking how many workers will be needed for what steps and organizing for a steady use of people through the project with appropriate contingencies for the inevitable delays is critical to estimating costs.

If the project does succeed, we need some way of convincing ourselves that it is indeed a success. If it fails, it is even more important to be aware of that fact. Testing and validation are important parts of the design process. Deciding who will make the decisions about testing and validation and who will have the authority to approve the product are important decisions in the design process. When the locus of authority to make these decisions is far removed from the locus of the decisions about the actual production of the software, it may become impossible to produce a product that is acceptable.

If the product is found to be flawed, either because of mistakes in design or, more commonly, because of changes in requirements while the design was in progress, then, if the project is not to be scrapped, a mechanism must be found to repair or rebuild the software, which may involve restarting the entire process.

If at the end of all this the product is working and acceptable, we enter the most important and generally least considered part of the life of a software product, the period of its actual operation. Then the user will have access to the software and perhaps do things that the designer never considered. At the very least, the users push systems and discover the limits experimentally. As an abstract pattern of information, software does not degrade with time. Yet it is a common observation that most software does seem to degrade and eventually stop working. This is a result of two phenomena: the environment in which any given software is used keeps changing, and "slight" changes to the software accumulate and corrode it like a layer of rust.

Unlike software, computers and related hardware fail, stop working, and have to be replaced. When they are replaced, it is usually with newer, different machines, which provide at best a slightly different environment for execution. Even when instruction sets are identical, newer machines tend to be faster, which may upset timing assumptions. It is common practice to replace and upgrade operating systems and language compilers during a hardware change. For example, the era of physics on computers has overlapped more than six distinctly different versions of FORTRAN (see COMPUTER PROGRAMMING LANGUAGES), and the transition has in each case disabled some number of programs.

Even when great care is taken to preserve the hardware and operating-system environment, it is a rare software system that does not accumulate maintenance changes, patches, and other improvements. Code that once provided printed lists of numbers for output might have seen a conversion to black and white graphics and then to color graphics. Ensuring that all the changes work with the original software and with one another as intended can be a complex combinatorial problem, eventually making a total redesign simpler than making the next "small" change.

Thus we have extremely complex environments in which we are trying to produce complex systems. We need tools to bring order out of such chaos.

1.3 Tools Used in Solving the Problem

The major tools used in software engineering rest on abstraction and the creation of hierarchies of abstractions to reduce the level of complexity to a manageable level. We save ourselves from having to deal with all issues at once by looking for some common characteristics by which to label groups of issues, hoping to reduce our confusion with new, broader categories. If we choose well, the new labels help. If we choose poorly, the new labels become additional distracting details we must track.

1.3.1 Conceptual Disciplines In the broadest terms, the major simplifying steps we can take are to define clearly the tasks to be performed, to define clearly the client for whom the tasks are being performed, to impose a structure on the tasks and their relationship to the client, to take that structure and reduce the context if it is not sufficiently simple, and, while we are doing this simplification, to try to keep what we are doing simple. This last concept, which in classical engineering is called KISS for "Keep It Simple, Stupid," applies not just to subproblems within the design of a system but to the system as a whole and at all levels in between. The fact that a system is large and may require the efforts of tens, hundreds, or even thousands of people to build is not a reason to give up on KISS. The basic task of software engineering is to take what at first appears to be impossibly complex and beyond the comprehension of any one person and find the tools and techniques to recast the problem in terms one person can understand. At the very least we must make the external appearance of what is produced sufficiently simple for a person to be able use it.

Most large problems require two modes of thought: analysis and synthesis. In analysis, we discover the inner workings of a system. In synthesis, we build a system from components. When we have an existing, working concrete system, analysis is a process of examination. When all we have is a problem requiring us to design a new concrete system, we may have to analyze an abstract system in terms of internal structures that do not yet exist but that we infer from the logical necessities of the abstract system and from our experience. Analysis of components is an essential step in synthesis, since it is difficult to combine components that are not understood, and synthesis can create powerful tools for analysis. We return to this interaction when we discuss top-down analysis and bottom-up synthesis below.

There are a large number of ways to analyze a system—i.e., different, often arbitrary, ways to slice something large into smaller pieces. We can follow the flows of control or the flows of information, or we can try to follow both together. There are always new tools to do this that can be found or synthesized. For each tool, we need to ask if there is sufficient benefit to justify the costs, delays, and complications involved.

1.3.2 Selection of Appropriate Tools In the early days of computing, the designer of a system had a limited range of hardware from which to choose. Picking another platform was a very expensive decision. It is important to get past that mind-set today and understand that the cost of hardware is no longer the dominant factor in system design. Trade-offs between choices of hardware platforms and the time and effort to be invested in using different platforms must be evaluated. The ripple effects of delays caused by changes requiring new documentation and training for experienced programmers and users must also be considered. Similarly, choices have to be made among design paradigms, languages, libraries, software tools, etc., with concern for the impact of delays and costs of any change from that which is familiar. For each tool choice we should attempt to quantify the benefits (estimated savings in time, reduced probability of errors implying a quantitative estimate of reduced rework effort) and costs (estimated expenditures in time and money to acquire or build the tools, train people, etc.). Some tools will show clearer gains than others and will be worth considering further. We look at the tools of software engineering in Sec. 3. There are many among which to choose. It is rare for one tool to be the right choice. In the end what matters is that the tools used be capable of doing the job and be well understood by the people using them. Delaying a project while people are retrained in the latest technology may be worth while for a large project with long time scales, but rarely for a small project. However, one must be aware of the not-invented-here syndrome, in which designers stay with familiar approaches and blind themselves to tools or even complete solutions they might simply pick off the shelf from elsewhere. Minimizing the impact of such biases is one reason why it is advisable to do a quantitative cost/benefit analysis of the impact of the choice of tools, rather than relying solely on qualitative impressions.

1.3.3 Human Factors The element of familiarity is one of a wide range of critical human factors in the design process. Unlike hardware, people are a very expensive resource and need to be used effectively. This is a difficult challenge. People have their own lives and interests, and their interests and skills do not necessarily fit well with the needs of projects. Matching people to tasks and motivating them to work is a subtle and demanding task. Some people work well in highly structured environments; some do not. Some people are natural team players; some work better alone. Encouraging people to work in ways that produce results of consistently high quality is critical to organizations that produce software. The process of making decisions, of selecting one approach over another, depends on human value judgments that are difficult to quantify.

As difficult as it may be to get people to work together, dealing with the clients for software and getting them to state their requirements clearly can be even more difficult. Organized approaches have been developed to deal realistically with people and try to produce a system that actually meets their needs and to some extent even does what they want. We look at these issues in more detail in Sec. 6.2.

1.3.4 Standards and Controls, QA/QC The minimum that is required to obtain quality results is a set of standards against which to measure processes and products. In small groups with innate motivation, these standards may be sufficient, but, in general, full systems are needed for quality assurance and quality control (QA/QC). Quality assurance is the term for the entire set of mechanisms used to assure those who need to be assured (which hopefully includes the clients and users of systems) that the system has an acceptable chance of doing what is needed of it. Quality control is the term to describe the detailed control mechanisms used to ensure that every step of the way in production the quality of the product is being maintained. In software systems, which rarely can be tested for every eventuality, quality control is an essential means of quality assurance. We return to issues of quality in Sec. 7.

2. BASIC CONCEPTS

In the Introduction to this article, we explored the distinction between software and hardware. See COMPUTERS, COMPUTER HARDWARE, and COMPUTER PROGRAMMING LANGUAGES for some of the terminology we will now use.

2.1 What Is Software?

Software is information that affects information-processing systems. From the time of Babbage (see Morrison and Morrison, 1961, pp. 245–295; Hollingdale and Toothill, 1965), the basic elements of software have been operations (actions to be performed), variables (information that may change during a calculation), and constants (numbers that do not change). These concepts spring from the design of computer hardware. A computer typically has a memory, a processor for arithmetic and control, and input-output capabilities. The details of hardware, instruction sets, memory sizes, and device controls now change very rapidly. Software built on such shifting sands must be designed with a certain independence of the particulars of the hardware, or any hardware dependence must be parametrized. Multiple computers may be joined together to form one larger machine. Software must then be able to present and distinguish among various models of machines: multiprocessors based on message passing among processors with independent memories, multiprocessors based on sharing one large memory among many processors in some coordinated manner, or collections of specialized processors working in bucket-brigade fashion to handle large arrays of data efficiently. Systems that provide the ability to manipulate images require software abstractions for graphics and visualization.

From the elementary concepts of operations, variables, and constants, we need to build abstractions that allow us efficiently to map real-world problems onto a wide variety of hardware designs. This is commonly done by working in hardware- and software-independent pseudocode, using graphical representations of systems, and deferring coding to near the end of analysis and design.

2.2 Types of Software

It helps clarify our thinking about software to distinguish some major types of software. Systems software is used to make hardware more manageable. Application development support software provides tools used to build software. Finally, application software is the software solution to a particular problem.

2.2.1 Systems Software Systems software is used to present us with well-controlled reliable interfaces, which isolate us from the specifics of particular hardware and which provide interfaces to the hardware on a level of abstraction more appropriate for use by other software. When a machine will be used by a single task at a time, a simple monitor may be used to provide higher-level interfaces for the execution of jobs. Systems that will be shared by more than one user need operating systems to control the allocation of shared resources. When the sharing will involve multiple threads of independent execution, it can be very useful for the operating system to provide virtual machines, which give each task the illusion of a private computer. Commonly used accesses to system functions can be gathered into collections of routines called application programming interfaces (APIs). The evolving need for graphical user interfaces (GUIs), which allow users to point and click instead of type, and real-time demands such as data acquisition can call for specialized APIs. Services can be provided for process-to-process communications within systems, on local area networks, and on internetworks. Operating system accounting and security, predefined interfaces for the creation of client/server applications, and access locks to allow users to reliably share their own resources without confusion are all useful. Software engineering for systems software and for all software dependent on systems software has been greatly simplified by the widespread acceptance of variants of the UNIX operating system (see Martin *et al.*, 1984). While there are many variants of UNIX and many competing systems, UNIX now provides a common background against which a large portion of software engineering is done, especially in the academic world.

2.2.2 Application Development Support In some cases all that is needed is an operating system and the appropriate application, but, especially for software engineering itself, specific tools are needed to support application development. Assemblers, compilers, interpreters, and revision control systems are an essential minimum. See COMPUTER PROGRAMMING LANGUAGES for a discussion of the first three classes of tools. A revision control system allows us to keep track of multiple versions of documents. As changes are made to a program or design document and it is time to replace an older version, instead of giving the new version of the document a distinct name, or discarding the old document completely, a revision control system allows us to hold both versions and get at either one. In addition, a good revision control system includes access locks, so that in a team development environment, one member of the team can check out a document for changes and be sure that the system will not allow someone to slip in another change out of order.

Perhaps the most important development support tool is a mechanism to maintain commonly accessible libraries of previously developed routines, so that earlier efforts can be reused. A permanent shared file system with library mechanisms is now a given in all software engineering efforts.

2.2.3 Applications In most cases, the ultimate objective of all we have described thus far is the support of applications. There are many types of applications, as many types as there are useful ways to process information. For example, we could have numerical, administrative, laboratory automation, real-time, embedded, GUI, graphical, distributed, client/server, or database applications. Each makes different demands. A numerical application may require extremely efficient access to the processor and libraries of highly optimized linear algebra routines. Administrative applications may require ultrareliable security and development support tools suited to large groups. Real-time applications may require extremely responsive support of multiple threads of execution. Embedded applications may need the same features as real-time, plus specialized device control and the ability to clone applications to diskless environments. Software engineering needs to look at what is common in all of these and make the most of these commonalities while preserving the unique flavor of each application.

3. SOFTWARE ENGINEERING

As we noted in the Introduction, the main focus of software engineering is on building new software systems. However, it is important to see the building of systems as a part of the complex lives of systems and to ensure that the approaches used in construction integrate well with the rest of the design cycle.

3.1 Software Design Cycles

The waterfall model discussed in Sec. 1.2 is one of many possible patterns of life for software. No matter what the actual order of events, however, the software engineer must have the tools to deal with all stages in the life of a project. It is never safe to assume that the process flows one way. The same or similar problems tend to recur, and involvement in earlier or later stages can help to make sense of the current stage. While it may seem desirable to "build it right the first time," this is rarely realistic. Most software products come back for reconstruction, redesign or even reanalysis many times. Nor should we restrict communications to be only between adjacent stages, with analysts communicating only with designers, designers only with programmers, etc. This may lead to serious miscommunications and a failure ever to close on a working product. The safest assumption is that software will flow in a series of nested loops, with all paths possible. Further, as pressures for rapid development become more common, it becomes desirable to skip steps, to make analysis feed directly to code generation, or to allow users to program without intermediaries. This is neither good nor bad, just a fact of the world in which software engineers work. When well-understood techniques are available to achieve such bypasses, an engineer may well use them, but it is important that they be well understood, or the result may be a nonworking system or, worse, a system that appears to work but produces wrong answers. Therefore, let us look at the major design paradigms, understanding that some of them may be applied by automata, rather than by people.

3.2 Software Design Paradigms

There are many ways to look at a system analysis and design problem. None of them is perfect. When faced with a totally new problem, it may be necessary first to invent a design paradigm and then to return to the specific problem. However, most problems of a modest scale or a great deal of regularity (such as many major scientific calculations) are handled more than adequately by some combination of bottom-up synthesis and top-down analysis used together in what is called "outside-in" design.

3.2.1 Bottom-Up, Top-Down, Outside-In Most software analysis and design projects start at one of two places: at the top, considering the highest level of abstraction and working down to understand the necessary details; or at the bottom, starting with existing tools, resources, and capabilities and working up to create more general, abstract tools closer to the level of abstraction of the problem. In most cases, working top down is a process of analysis, discovering the inner structure of the problem, and working bottom up is a process of synthesis, creating the solution to the problem. However, as we shall see, a top-down process can be used as a means of synthesis, and a bottom-up process can be used as a means of analysis.

Consider the flow in Fig. 1. We start with the original problem—for example, the requirement to "do" the experiment mentioned in Sec. 1.2. The overall task is to do the experiment. We can analyze that requirement as consisting of two major subtasks: data acquisition and data processing. We can then further decompose data acquisition into two subtasks: experiment control and data logging. We can also decompose data processing into two subtasks: data reduction and modeling. If we continue this approach to a fine enough level of detail, that detail will map directly onto the necessary code for implementation. Indeed, if we are working in almost any suitable higher-level language, we can make this into a direct code synthesis technique by making each transition from higher-level tasks into lower-level tasks into a sequence of calls to subroutines, which get written after their use has been defined by those calls. That is top-down analysis used as a means of code synthesis.

Now consider the opposite flow in bottom-up synthesis shown in Fig. 2. We start with a collection of simple tools, the software equivalent of knives and screws and levers, and put them together into increasingly more powerful and complex tools until we have created a tool adequate to solve the problem. Just as combining two knives with the fulcrum of a lever produces a pair of scissors, which can do jobs neither knives nor levers can do, combining simple software tools can produce systems of remarkable power and complexity. Returning to the example we used above, we might first build a library of basic digital and analog device-control routines, and mathematical routines for linear algebra, then combine the basic device-control routines into motor-control routines and pulse-acquisition routines, and

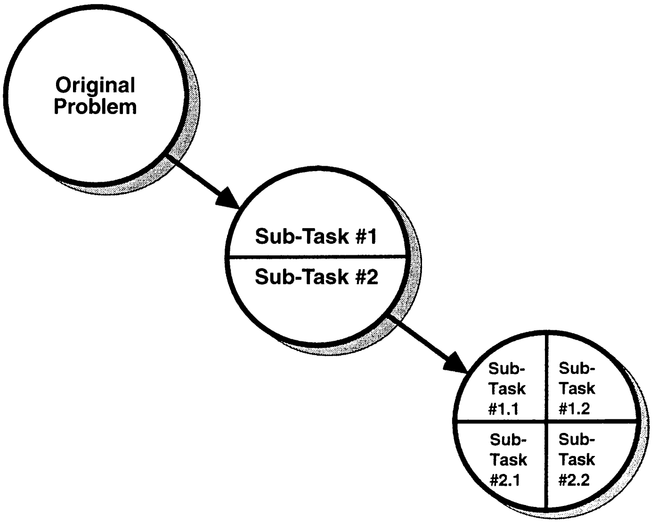

FIG. 1. Top-down analysis.

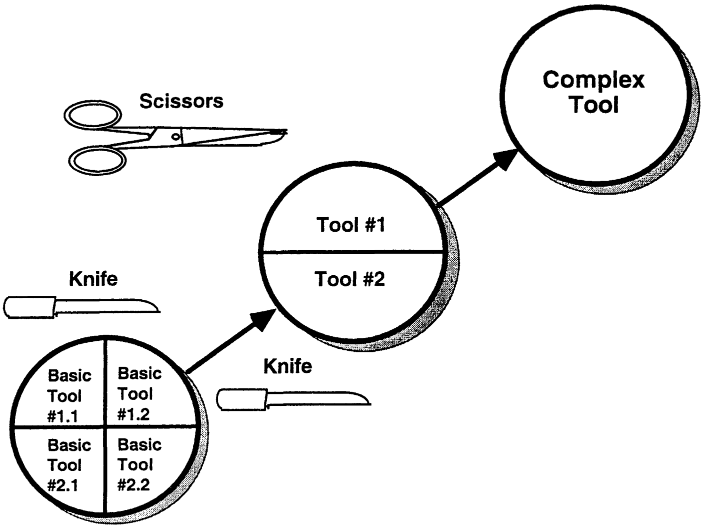

FIG. 2. Bottom-up synthesis.

the linear algebra routines into a least-squares program. That this is a synthesis technique is obvious. It is also useful in analysis, especially when finding errors in existing systems or trying to change them. Following the structure of a system up from the bottom helps to ensure an awareness of the fine details of the actions the system will take.

There are benefits and risks in both of these approaches. One can misuse either technique. In top-down analysis, one could try to skip all intermediate decompositions and go directly to the most elementary subtasks. In bottom-up synthesis, one could try to build the entire system in one pass. In either case, the result is usually a single large tangle of code. Such code is almost impossible to understand and is very prone to errors. However, there is no assurance that hierarchical decompositions used in a top-down analysis will lead to a feasible implementation. When we finally get to the bottom level we may be demanding resources that cannot be delivered. Worse yet, by decomposing each branch independently, we may waste a great deal of effort creating slightly different solutions to similar problems from other branches, making the code unnecessarily complex. Pure bottom-up synthesis suffers from complementary problems. There is no assurance that the structures we build by assembling simple substructures will be combinable into a solution to the overall problem, and we may waste a great deal of effort creating or gathering libraries of substructures that are irrelevant to the problem at hand.

It might appear, however, that both techniques at least reduce complexity by creating hierarchical structure. This is desirable, but it is not an unmixed blessing. Consider the example in Fig. 3. If we did no hierarchical analysis or synthesis, we might build one very large program out of many elementary operations. At any point in that one big program, we have a tremendous width of context, i.e., the number of lines of code, in this case 100 000 lines, all of which must be considered at the same time to understand the operation of the code.

As we recast the large program hierarchically, the width of context is reduced, but the depth increases, and the number of interfaces rises sharply. By the time we have gone four levels deep, no routine demands a width of context of more than 20 lines, but we must deal with over 10 000 routines and over 10 000 interfaces. We have traded one form of complexity for another. Finding the optimum balance point is difficult. The trade-off can be made more economical if there is some significant commonality and reuse among the routines we generate. Bottom-up synthesis provides a bias toward commonality and reuse of code. Therefore, for large projects in which optimizing the depth and width of context and minimizing

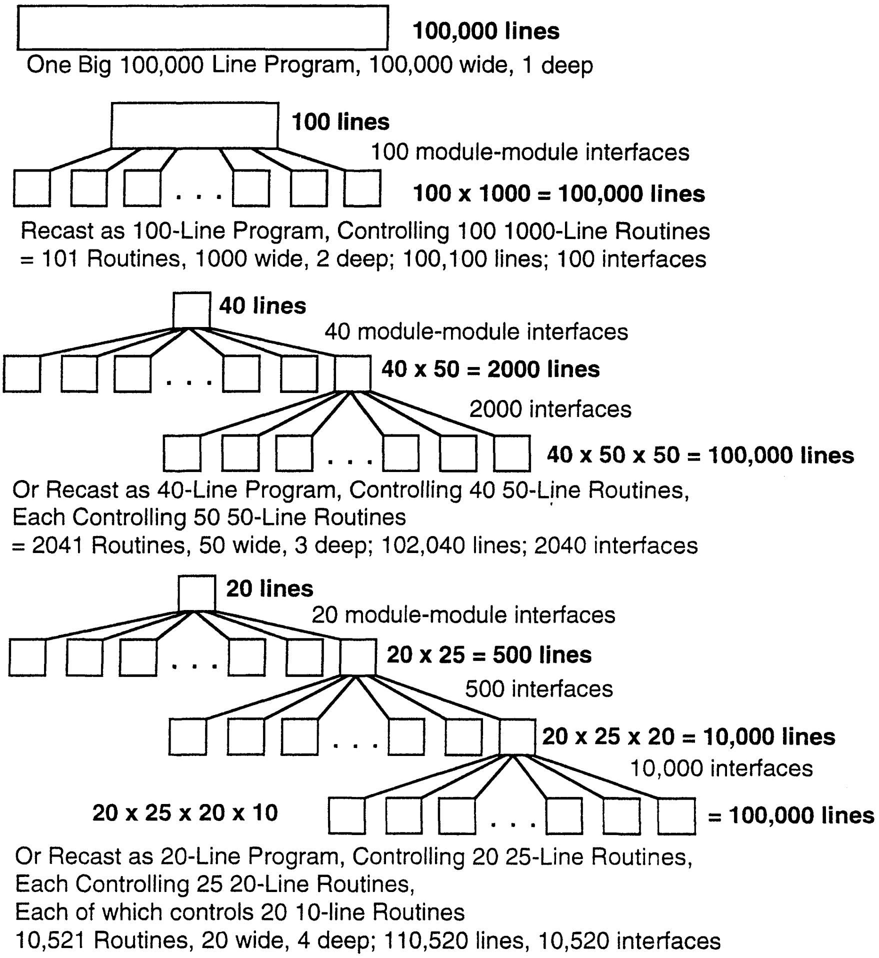
100,000 lines
One Big 100,000 Line Program, 100,000 wide, 1 deep
100 lines
100 module-module interfaces
100 x 1000 = 100,000 lines
Recast as 100-Line Program, Controlling 100 1000-Line Routines
= 101 Routines, 1000 wide, 2 deep; 100,100 lines; 100 interfaces
40 lines
40 module-module interfaces
40 x 50 = 2000 lines
2000 interfaces
40 x 50 x 50 = 100,000 lines
Or Recast as 40-Line Program, Controlling 40 50-Line Routines,
Each Controlling 50 50-Line Routines
= 2041 Routines, 50 wide, 3 deep; 102,040 lines; 2040 interfaces
20 lines
20 module-module interfaces
20 x 25 = 500 lines
500 interfaces
20 x 25 x 20 = 10,000 lines
10,000 interfaces
20 x 25 x 20 x 10
= 100,000 lines
Or Recast as 20-Line Program, Controlling 20 25-Line Routines,
Each Controlling 25 20-Line Routines,
Each of which controls 20 10-line Routines
10,521 Routines, 20 wide, 4 deep; 110,520 lines, 10,520 interfaces

FIG. 3. Width and depth of context.

the number of distinct interfaces and routines can be critical to success, it is common to mix aspects of top-down analysis and bottom-up synthesis, working in on the problem simultaneously from both requirements and resources, i.e., working outside-in. Consider Fig. 4. In dealing with a large problem, it is common to find several pools of existing tools and several areas of broadly specified abstract subtasks. We try to extend the domain of applicability of each pool of tools by building libraries and appropriate application interfaces to serve as virtual tools oriented toward the known subtasks. At the same time, we bring each of the major abstract subtasks closer to the level of the available tools, by further decomposition into simpler and more concrete sub-subtasks, keeping in mind the capabilities of the libraries and interfaces we have. The abstractions and the virtual tools are brought together in a common ground at an intermediate level of abstraction and complexity. It may take many years and many false starts to find the right common ground for a domain of applications, but once found, that common ground can provide the most productive level at which to work.

3.2.2 Functional and Applicative Programming When execution of a piece of code creates a "side effect," a nonobvious change to the state of data structures, the context of that piece of code has been broadened, making it harder to keep track of what it is doing. Functional (or "applicative") pro-

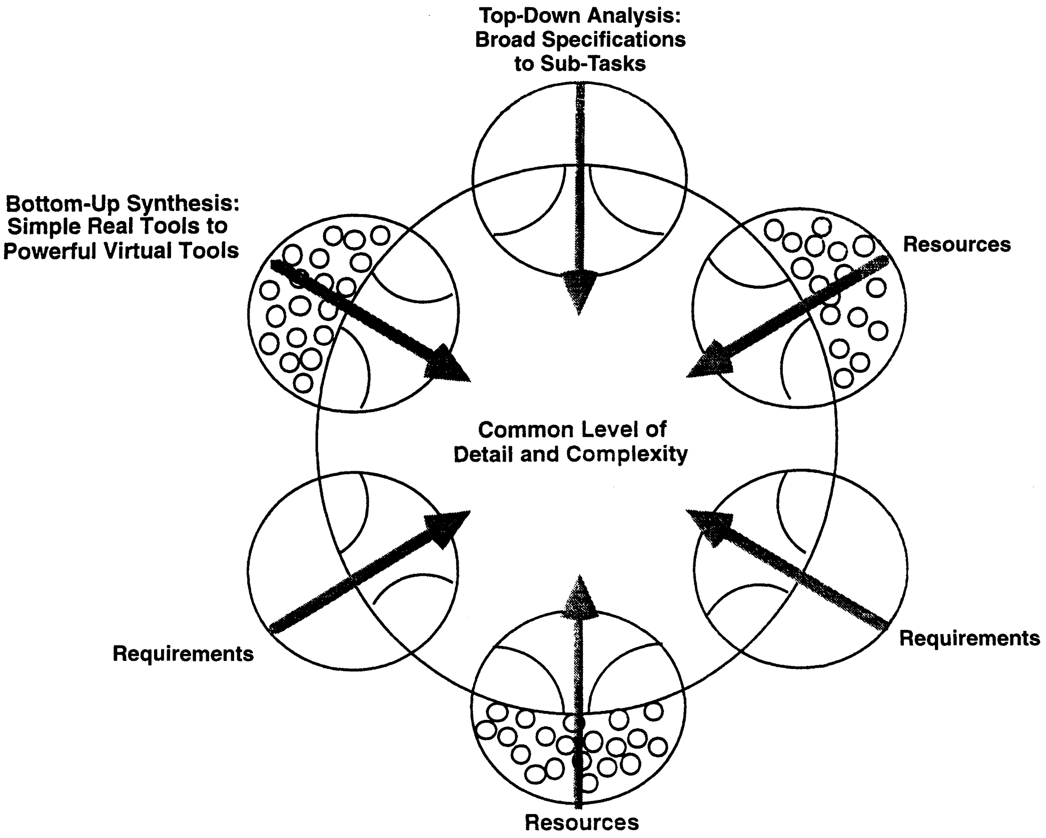
Top-Down Analysis: Broad Specifications to Sub-Tasks
Bottom-Up Synthesis: Simple Real Tools to Powerful Virtual Tools
Resources
Common Level of Detail and Complexity
Requirements
Requirements
Resources

FIG. 4. Working outside-in.

gramming is an approach used to avoid that problem, making it easier to deal with the routine–routine interface problems discussed above. In functional programming, problems are solved by pure function evaluations. Routines have no memory of prior state; they take exactly the same action every time they are called. Thus, no matter how deep a hierarchy we create, at each level and in each routine along that level we need only validate execution locally. The concepts began with work of Church on the "Lambda Calculus" in the 1930s and found a viable programming language expression in McCarthy's LISP. This approach to software design has seen extensive use in natural language processing and symbol manipulation, but the burden of working without side effects can be unacceptable in applications for which side effects are useful, such as device control, operating systems design, and other real-time applications that need to remember where they are as they are interrupted. In such cases other paradigms are better suited. For more background on functional programming, see Hudak (1989).

3.2.3 Structured Programming Structured programming is the application of top-down analysis to the flows of control in an information-processing system and has been used very effectively in systems dominated by considerations of control flow independent of data. Dijkstra (1968) used the technique to design an operating system three levels deep with each routine sufficiently small and well understood that he had the confidence to assert that the system would be "flawless" before having done final testing, and he was essentially correct. To make the technique work, it is important that each module be entered and exited cleanly and used only through its "official" interfaces (a doctrine to which all major software engineering paradigms now adhere). As the technique matured, practitioners tried to ensure clean entries and exits by simply prohibiting the use of direct jumps to particular lines of code, creating so-called "goto-free" code. This extreme rigidity is no longer considered necessary, but certainly unusual changes of flow of control have to be well justified. See Dijkstra (1976).

3.2.4 Proving Correctness Dijkstra attempted to prove the "correctness" of his implementation. Hoare (1969) made this approach rigorous by providing formal proofs of programs. There is a fundamental limitation to this technique. We can only prove an internal consistency between a program and a sufficiently detailed specification. We cannot prove that the program does what was intended by the user. However, removing the possibility of a gap between specification and implementation is highly desirable. While a proof can be attempted after the fact, it is most useful as an intrinsic part of the design and creation of code. By bringing formal methods up into the specifications, the chance of the final system failing to be consistent with those specifications is very small. Such approaches have even been extended to verifying compliance with timing constraints in real-time systems (Felder and Morzenti, 1994).

3.2.5 Software Engineering and ADA Real-time systems are among the most challenging to software designers. Such code is often used for control of devices, and errors can have disastrous consequences. This provides a strong incentive for the application of engineering approaches. ADA (Booch, 1983) was designed to establish a sound engineering environment for the creation of embedded real-time systems, which are information-processing systems embedded into and controlling devices. ADA drew on earlier work in simulation languages and uses "abstract data types" or "classes" that combine data-structure definitions with definitions of the actions that can be taken on those structures. ADA makes use of "strong typing," forcing actions to be consistent with the abstract data-type declarations. Instances of abstract data types coupled with the necessary constants and implementations of the actions make what we now call objects. Only the portions of the specification of an object necessary to determine its interactions with other objects are visible. Private, internal data structures and actions are hidden. This was an important step in assuring that a system could be built reliably, since one could know with certainty which portions of an object could be changed without having to consider interactions with other objects. This "encapsulation," a bottom-up creation of new abstractions, is a powerful design-control mechanism. ADA is still used, but much current work with objects is now done with languages that support additional capabilities such as "inheritance," to which we return in discussing object-oriented programming. ADA provides a complete environment for software development and was an important step in the development of CASE tools.

3.2.6 CASE Tools Computer-aided software engineering (CASE) is a term used to define the use of computer-based tools to automate steps in software development. The original CASE tools are "ordinary" assemblers, compilers, and interpreters (see COMPUTER PROGRAMMING LANGUAGES), which allow software to be defined in human-readable terms. Auxiliary programs that support compilers, such as loaders, linkers, and librarians, allow independent compilation of routines, allow creation and maintenance of libraries of routines for multiple use, and encourage modular development. Interactive context-sensitive editors facilitate the rapid production of both code and documents, and revision-control systems allow the multiple versions created to be tracked. Profilers, monitors, and analyzers allow the software performance to be monitored and, hopefully, improved.

Such tools are often combined into smoothly functioning "seamless" environments, in which the flow from code creation in an editor to compilation to test execution is facilitated. This tool-oriented approach also is used in the administration and control of the development process by means of project management tools. These tools allow time and resources for tasks to be estimated, tasks to be linked showing their dependences, and critical path dependences to be noted. Used interactively, such tools can flag trouble while there is still time to correct the situation. Everything that helps in communication and record keeping becomes a valuable tool in development—databases are used to track the complex dictionaries of entities in object-oriented analysis, electronic mail is used to provide internal project communications among participants, etc. See Fuggetta (1993) for a "Classification of CASE Technology."

Since most people work better with graphical representations than with purely verbal descriptions, there are CASE tools that allow object-oriented software (see below) to be defined graphically. See Armbruster (1995) for a review of such tools. We explore a simple Hypercard® example of partial graphic design of an application in Sec. 8.

3.2.7 Object-Oriented Software As we noted in the discussion of ADA above, at the time this article was being written the most popular software engineering paradigms related to the representation of software systems as collections of communicating objects. The messages that objects send one another may then cause the objects to take some appropriate actions, but the actions are hidden internal mechanisms within the objects, known to other objects only by the outwardly defined communications mechanism. This is precisely how basic computer hardware is designed and presented to software and has long been the realistic approach to design of demanding real-time applications such as operating systems. As general applications are expected to display more real-time interactive capabilities, object-oriented approaches become essential. Object-oriented paradigms have been applied to all stages and aspects of software development, including to software development organizations. The focus on semiautonomous interdependent teams is itself an expression of human software development processes in object-oriented terms.

What is an object? An object is an information-processing system or subsystem or component. In contrast with functional or applicative programming, in object-oriented representations side effects are not only allowed, they are actively encouraged, since an object is allowed to process that same information in different ways at different times by having hidden internal mechanisms; i.e., objects may have a private internal memory of their state that is not accessible to any other object. An object may do anything with the messages it receives that the designer wishes, as long as the object deals with other objects in the ways it is expected to, i.e., that it honors its "contracts" with other objects. This may seem to be a very vague definition. It is. It permits almost anything to be called object oriented. Indeed, we could define an object called "global memory" that accepts the messages "STORE value in location" and "FETCH data from location" and that communicates back the values fetched. We can design all other objects to use global memory to store their internal state and thereby make what had been hidden completely visible. (Such techniques, which evolved in hardware design to permit analysis of internal states of complex logic, have become essential to debugging complex object-oriented systems.)

The main focus of object-oriented representations is on object-oriented programming (OOP), which is done in several languages, the most popular of which is C++ (see, for example, Halladay and Wiebel, 1993). C++, an extension of C, allows abstract data types, encapsulation, inheritance, polymorphism, and dynamic binding.

Let us look at how an experiment control problem might be viewed in terms of objects. Looking at the bottom, we see we have some devices to control, and so we will certainly need some software objects for device control. We are not yet certain of the details of the control we will provide in each case, but it certainly is worthwhile to define an abstract "class" of such objects, each of which would have "methods" to initialize its associated device(s) and its own internal data structures, to send commands and data to its devices, to accept status information and data from its devices, and to shut down the devices and release resources when they are no longer needed. Let us call the class of objects "DeviceController." Now, in looking at an example, we find that we have a four-axis positioner, so we define a particular object "FourAxis," which is a kind of device controller. Therefore we use the general class definition of DeviceController to create the framework we will need and add the specific mechanisms needed for the methods FourAxis will use. When we are building the system, we have incremental encoders for position, but we expect to acquire absolute encoders. We now are facing a design decision whose impact can be carefully controlled through the definitions in FourAxis, with the internals such as encoder type that have limited impact on the processes using FourAxis hidden from view and the internals that do have serious impact, such as the positioning accuracy and timing, made visible through well-defined object interfaces.

In designing the interface to FourAxis, it would improve clarity to allow client objects to work in terms of laboratory coordinate systems for positioning. We find reason to permit polar coordinates in both degrees and radians and Euclidean coordinates in centimeters and inches. So we "overload" the positioning methods of FourAxis to allow the same operation name to work reliably and unambiguously in any of the coordinate systems driven by the type of the position data. FourAxis has become "polymorphic."

That is a small example of bottom-up object-oriented programming. Looking at the same problem top-down, we might start with the view of the entire experiment control system as one object, then decompose it into user-interface, device-control, and data-management subobjects. Clearly the preliminary work we just did on FourAxis fits well with this structure, and the thought about coordinate systems interacts both with the user-interface objects and the data management objects. Hopefully, we will work toward a common meeting ground.

For more on object-oriented software, see Booch (1991).

3.2.8 Usability—Testing and Prototyping A system may work according to its specifications but be difficult to use. It may demand too much of its users or demand things of them that conflict with their training, abilities, interests, or tastes. A system may work for one user and not another. For example, if a manual control input is required and requires use of the right hand, left-handed people may find the system very frustrating. If critical messages are displayed in pure red, those messages may be invisible to the 8% of males who are color blind. Ensuring that a system is usable requires attention at all stages of the design process. Testing and prototyping can be effective techniques to keep a design usable.

We must clearly understand who and/or what will use the system and early in the design process present to them as realistic a simulation as possible of what the system will be like, to see if our understanding of what the system is to do is actually correct. We may have to start with verbal descriptions, but people do not usually absorb the full impact of a system from a verbal description. It is important at least to provide accurate screen images for review and, where possible, to create an adequately functional mockup or prototype for the user to try. Once the system has gotten to the point of having completed modules, testing by using the real modules in place of prototypes is important, both to check for errors in the modules and to see if they meet user expectations. If this is to have any meaning as a test, allowance must be made for the possibility that the software may fail and have to be repaired or redesigned. It is an excellent moral challenge to "build it right the first time" and a goal to which all software engineers aspire, but it is unrealistic in general. It is better to have allowed for extra design cycles and not to have to use them than to fail to allow for those cycles and be left with a useless product.

In many organizations the transition from development to actual use is formalized into a process called a "handoff" or release. The development team is intentionally kept distinct from the operations team so that there will be someone ready to test the product who has not been prejudiced by being involved in developing it. When development thinks it has a product ready for actual use, it prepares a complete handoff package or release kit with all the software and documentation, supposedly ready to stand on its own legs. Operations takes the package, tries to work from it without side communications, and prepares its own tests of usability. If the product has problems, development is given all the details so that they can reproduce the problem, but they are not allowed to tinker with the production version. They have to prepare fully documented patch releases and upgrades, which go through the same handoff procedure. This may seem a cumbersome process, but it has proven so useful that most serious professional developers actively seek parties willing to accept preliminary trial handoffs (so-called "alpha" and "beta" releases) to find remaining problems with their product.

For more on usability, see Nielsen (1992).

3.2.9 Requirements Engineering A system may be usable, yet not usable for the intended purpose, because the intended purpose was not clearly understood or not clearly communicated. Determination of requirements, creating clear, complete, and consistent specifications, is one of the most difficult aspects of software engineering. There are many strongly held opinions on how to address this task, but only partly understood methods with limited records of success. We look at the issues and approaches in requirements engineering involving group efforts, where the designer and the customer (or "stakeholder") are distinct, in Sec. 6. Here let us just consider the abstract issue of defining the question.

A software system exists as part of a larger system. If the system as a whole is not well understood, there is a significant risk that the software will fail to meet the needs of the system. In some areas of software design, the person analyzing the system is untrained and has at best a rudimentary familiarity with the larger systems involved. Fortunately, in the physical sciences, it is not unusual for software designers to be trained and active in the application area. However, even then it can be desirable to put aside any presumptions we may have and come to a new specification question with a completely open mind. We first try to establish a consistent framework within which to pose the questions about the project. We might focus on procedural aspects or information flows or object-oriented representation. Within that framework we ask the question, "what?" That is, we might ask, "what will this data acquisition system do," or "what information do we wish to get out of this system, and what information might we have available to put in," or "what are the major elements of this system, and what services will they provide to one another?" All are valid possible starting points, and in most cases it is worthwhile to pursue all of them and see where they lead, but it is usually less confusing to pursue them one at a time. The answer to the first-level question usually involves sentences about actions taken on information. If we are pursuing a procedural thread, we would then inquire into the detailed meaning of each of the actions, if a data flow thread, into the detailed meaning of each item of information, and if an object-oriented thread, into the detailed structure of each clause. After some amount of iteration and comparison among threads, we are likely to have a trial-balloon document describing requirements.

There is considerable disagreement over the desirability of keeping a wall between the creation of such a requirements document and the creation of code. If we let them merge, then we must impose a strong formality on the specification so that it can be analyzed by software, but we gain the ability to have the software check for consistency and do most of the programming for us. Changes in documentation and changes in code become one and the same thing. On the other hand, the formality we impose may preclude us from clearly stating what we want the system to do, since we can only be as expressive as the formal specification language allows. Arguments that a specification language allows any idea to be expressed, provided we elaborate enough, are unrealistic, since clarity can be lost in the complexity of elaboration.

As of this writing there is an ongoing effort to clarify the choices to be made in requirements engineering and establish the reliable range of utility for the major tools. For a review of the questions, see Siddiqi (1994) and the other articles in the same issue.

3.2.10 Re-engineering Suppose you have a software system that is doing something useful, handling, for example, a diffraction-data acquisition and display for a single detector data stream with four-circle positioning. You want a new system that shows in some ways a strong family resemblance, say an area detector system with an array of data streams and detector roll, pitch, and yaw as additional positioning parameters. Even if the existing system is poorly documented, it is a working prototype for important aspects of the new system and may be worth using as a partial basis for building the new one. The overall process of evolving a system with one functionality into a system with different functionality is called re-engineering. As the value in this reuse of system structure as well as components is recognized, it is becoming the practice of both good designers and good suites of design tools to retain a richly detailed trace of the design flow of a system so that future re-engineering can be facilitated. For older "legacy" systems, an essential first step to permit re-engineering is to reverse engineer the system to discover its inner workings.

Reverse engineering of software is a complex process. Software tools can be used to look for relevant patterns in the code and to attempt to recover the original design information, but often a software system is a complex combination of pieces added by random people at random times in random ways, sometimes without a clear design goal in mind, and often incorrectly. It is rare to reverse engineer a large system and not find previously undiscovered bugs in the process. It is sometimes cheaper and safer to discard the old system completely and start fresh, rather than risk provoking buried bugs in poorly understood sections of code.

If we have (or create) sound documentation of the design of the existing system, it is tempting just to glue on new features. This is a high-risk approach, for the same reasons that starting to extend a house by just knocking down walls and framing new rooms is risky. The whole thing may collapse. For example, our area detector may force a higher interrupt rate on the processor than it can sustain, causing fatal stack overflows. The new system has to be designed before it is built. The old system is there as a resource and helpful guide, not as a straitjacket or a foundation.

In addition to all the standard tools and techniques of software engineering, the use of "wrappers" and "middleware" has proven very effective in re-engineering efforts. A wrapper is a layer of software that presents itself to an existing subsystem as a client with which the subsystem is known to work and that presents itself to outside clients as a server meeting new interface requirements. For example, the positioning subsystem for the single-detector data stream might be a highly efficient work of art with which we have neither the desire nor the need to tinker. However, assume it had messy handling of data structures inconsistent with a good object-oriented design. Rather than modify it, we could create a general positioning layer as a wrapper, providing the necessary data structures and hiding them from the rest of the software, as well as adding the hooks for the area detector roll, pitch, and yaw. If we intended to come back later for a full rewrite, this approach would allow prototyping of the rest of the system quickly and efficiently.

See Winsberg (1995) and Sneed (1995) for more on re-engineering.

4. SECURITY ISSUES

No creation of man is ever completely secure. Imperfections in design, the wear and tear of normal use, environmental insults, deliberate assaults by persons with evil intent, accidental damage by persons who are careless, and simple entropy all conspire against what we create. Information systems are no exception, and, as information systems become increasingly interconnected and interdependent and failures in one system can propagate to many others without any chance for human recognition that anything is wrong, security issues become critical topics in software engineering.

While changes to ensure security are often added to a system after it is built, as with any changes, it would have been better to consider them in the original design. In general, we need to extend the models we use to include all the failure modes the system is likely to encounter. Where possible we have to ensure that failures are detected and reported or corrected. This can increase the complexity of the design, and the tradeoffs of costs and benefits have to be examined.

There are two major categories of failures to consider: those that are due to random events, and those that are due to intentional attacks on the system. For independent random failure following a random probability density distribution p over a domain of events x, with costs c, then the expected cost (usually called the "risk") is

$$\int p(x)c(x)\,dx.$$

We then try to adjust the design to minimize the risk. This can be done by reducing p where c is high or reducing c where p is high. For example, if the cost of losing certain data is high, we can reduce the probability of loss by check sums and replication for that data or reduce the cost of loss by making modules dependent on that particular data able to function from other information.

Caution must be exercised in mechanically using probabilistic risk assessment for securing a system against random failures. Averages can be deceptive. If there is a high-cost, high-probability failure mode, the efforts we put into dealing with that failure mode may blind us to a large number of lower-cost and lower-probability failure modes that make the system performance so annoying as to be unacceptable. This can be viewed either as a misestimation of costs or as a poor choice of measures in the integration.

The more difficult category of failures to model and deal with is that of deliberate attacks on the system. We now must assume an active malign intelligence introducing failures. We can no longer assume a probability distribution as given, but must enter the realm of game theory (*q.v.;* von Neumann and Morgenstern, 1944) in which our opponents will adjust the distribution to our disadvantage. We then have to find a design or operation strategy that will minimize our maximum losses. Thus, even if the compromise of a critical password is estimated by us to be a low-probability event, prudence would argue for building in safeguards to ensure that the damage that can be done by loss of that password be limited to some acceptable level.

In planning methods for dealing with failures—random or intentional—we need accurate models of the behavior of the system. Complex systems built using very abstract models to hide the complexity of the system may also hide critical details of potential failure modes. It does no good to have a system designed with multiple redundant independent data communications paths if the physical implementation puts the data through virtual channels on a common medium. Security modeling has to be done on all levels of abstraction, from the highest to the lowest. For more information on security design approaches, see Baskerville (1993). For considerations in the compartmentalizing of data flows to help protect against intentional system failures and loss of confidential information, see Sandhu (1993).

5. OTHER ISSUES

5.1 Issues in Maintenance

Software should not need maintenance, but it does. While the logical abstractions of code do not wear out, the environment in which software is used changes and necessitates changes in the software. In addition, no matter how carefully software is designed and tested, there is some chance of errors. A person may make a mistake, or a failure in some software or hardware component used to generate or transcribe the software may be reflected in the software. Problems can lie dormant for months or years and then be discovered by users or client systems doing something different or by a change in the environment.

Once we accept the need for change, we must address mechanisms to effect change. Especially when coding is a separate process from analysis, changes can have a corrosive effect, making the system in use drift further and further from its specifications. Suppose we have an experiment with 21-bit encoders for position that we upgrade to 24-bit encoders and just make the minimal changes in the positioner driver software to handle the new encoders. Since we are approaching and sometimes exceeding the accuracy of real arithmetic, failing to reflect this change in the system documentation invites mistakes at a later date. We need mechanisms to hold code and its documentation together. This is perhaps the strongest argument for specifications and other documentation to be made a part of code against which the code is automatically verified or from which the code is generated. Without such automatic tools, tremendous discipline is required of programmers to ensure that they reflect all changes back up to the highest levels of the documentation that might in any way be affected.

5.2 Reuse of Software

A very effective tool for reducing the chance for errors and the need for changes is to reuse existing, appropriate, well-understood, well-tested software. Subject to the constraints stated, there is much to gain from this. The more the same pieces are used in different projects, the better understood they become, and the more certain we can be that all errors have been found and either removed or clearly defined. The time and effort that goes into the initial development gets amortized over a much longer life than would be the case for code used only once. Major libraries of routines are now a staple of software engineering.

There are, however, risks and limitations to this approach. Intellectual property law may give other people rights in software you wish to reuse. The fees to them then become part of the cost equations in doing the project, and potential future fees become a continuing budget item. More importantly, the owners of software sometimes are reluctant to release it in its entirety, providing only external documentation and machine-executable code (known as "binaries") rather than detailed internal documentation and human-readable source code. This limits the degree of understanding of the code that can be achieved and weakens the value of reuse of software as a sound engineering tool.

Even without the intellectual property limitations, and even with every line of source code in front of you, you may fail to understand a software subsystem completely. Subtle errors can lie dormant in the most heavily used routines for decades.

5.3 Decision Support

All of software engineering involves decisions: what design approach to take, which software and hardware to select, what resources to apply to various parts of the problems. The software that is produced may have to make decisions. Making a decision is a complex process, especially because it involves value judgments. Each decision made may create a cascade of further decisions that have to be made. Let us examine some of the issues in making decisions. When decision making is automated, the resulting software provides decision support tools.

In its simplest form, a decision is a choice between two alternatives. Let us call the alternatives *A* and *B*. If we know nothing more about them than their names, there is no way to decide between them. We must know some attributes of *A* and *B* in order to choose between them. Each alternative may have a cost, and each may have a value. The cost is the price we must pay for choosing that alternative, and the value is the benefit we derive. If the costs and values are given in measurable units, the decision would appear to be simple: Choose the alternative for which the value less the cost is greatest. It is not so simple. Suppose alternative *A* costs \$100 and returns a benefit of \$1000. Suppose alternative *B* costs \$1 000 000 and returns a benefit of \$2 000 000. Clearly we have more to gain from alternative *B*. In decision support terms, we say that alternative *B* "dominates" alternative *A*. This leads us to a simple mechanical step in making decisions: Compare values and costs of decision alternatives, and discard the alternatives that are clearly dominated by other alternatives.

However, if we do not have \$1 000 000, we must, with regret, choose alternative *A*. Perhaps we could pursue alternative *B* by borrowing the money and then paying it back when we receive the value of our investment. This sounds attractive but leads up to wonder if we have stated the problem correctly. Are we certain of the value of *B*? Perhaps there is a risk that we will not ever see the \$2 000 000 or that we will see it at a later time.

Let us state the problem more realistically. We need to assign probabilities to the costs and values. Suppose we know that, for each of the two alternatives, there is a 50% chance of the cost being the stated amount and a 50% chance of it doubling. Suppose we know that there is a 90% chance of getting the stated value and a 10% chance of getting nothing. The expected costs and values are then

Alternative	Cost	Value
A	\$150	\$900
B	\$1 500 000	\$1 800 000

Alternative *B* still looks more attractive, as long as we borrow \$1 500 000. We have nine chances out of ten of getting \$2 000 000 and one chance in two of only spending \$1 000 000 of the \$1 500 000 we borrow. But what if we hit the one chance in ten and have to pay back \$1 500 000 with no value in our hands? Alternative *B* could cause ruin. In decision support, the rules of "ruin" and "quasiruin" are applied. If an alternative admits the possibility of a disastrously high net cost, i.e., of ruin, it is rejected; if it admits a net cost that could be absorbed, but that crosses a subjectively set limit of risk, it is rejected as quasiruinous.

Having pruned the alternatives by removing those that are dominated by others and those that may cause ruin or quasiruin, we can compare the remaining quantifiable alternatives by computing appropriately weighted sums of the possible costs and possible values of each alternative. It is tempting simply to apply the probabilities of the various outcomes and use those as the weights, as we did above, but that is rarely sufficient for a sound decision. Each probability must also be multiplied by appropriate subjective measures of the desirability of each outcome, creating what is called the "subjective expected utility" or "SEU." Comparisons of SEUs are at the heart of most decision support tools. However, a word of caution is appropriate. Such methods can only be as good as the accuracy of the values and weights assigned. In many cases it is difficult to assign numeric values accurately, and sometimes it is impossible (see Salvendy, 1987).

6. SUPPORT FOR GROUPS

As Pressman (1995) noted, "People—not technology or process—build software." People working in groups raise some of the most difficult issues in software engineering.

A person developing a completely new program for his own use and on his own initiative has very effective internal lines of communication for the project. He need only consult with himself to define his requirements, approve his specifications, decide on changes, and gain approval from the customer for job completion. Things are not so simple when more than one person is involved as a developer, or if the users, customers, or other decision makers are different people than the developers. Issues of effective communications, team formation, and group interaction become critical to the software development process.

In this respect, software engineering has much in common with any other group engineering activity, and the lessons of classical industrial management apply. Once the job is no longer to be accomplished by an individual working alone, some form of allocation of tasks among people must be made, and some mechanism must be set up to make those allocations of tasks, to monitor progress, and to ensure integration of subtasks into a coherent whole. Except for the use of people instead of computers, this has much in common with the problems presented in organizing a complex program into cooperating modules. The software engineer has the advantage of bringing concepts and tools he routinely uses for software design to the task of designing the process of designing software. There are limits to this, however. People are not computers.

Coordination of tasks being done by different people is a major stumbling block in group development activities. According to Kraut and Streeler (1995), the most valuable and most used coordination technique in software engineering is discussions among peers, especially in the same project. The challenge is to integrate this powerful communication mechanism into the development process in ways that help it fit project goals, rather than in ways that work at cross purposes.

6.1 Organizational Issues

Creating an environment in which people can work effectively on software is difficult, especially when an existing external organizational structure has to be considered. If the organization primarily rewards managerial career paths and not lower-level line activities, then there is a strong disincentive for programmers to stay on the line and code. This has led to the creation of programming management structures that are not related to the technical issues of creating software but that must be considered in the process. If organizational lines of communication are vertical and centralized, the essential interpersonal communications needed for software engineering by groups may be blocked. In some organizational cultures, there is a sharp distinction between analysts who define what has to be created and implementors who create what the analysts have designed. This can be a valuable division of labor, equivalent to that between architects and builders in construction or, as is so often the case in software design, one that breaks down because of blocked or garbled communications. It can work well when the analyst has a deep understanding of both what the client says and what the implementor does, and when the implementor is working on common ground with the analyst, and when implementors have sufficient cross communication with one another and with the analyst to stay on track with one another. In some organizational cultures it is common for those who will use or operate a system to be a distinct group from those who designed it, and perhaps even from those who requested it. A formal process of handoff of a system from development to production happens. Then, when further changes to the system are needed, if smooth paths back into development are not available, conflicts may develop and production may start doing its own programming. As we will see in discussing approaches to ensuring quality in Sec. 7, these traditional data-center structures are under strong competition from more team-oriented, flat hierarchy approaches. For more detail on the classic data-center organization, see Withington (1972).

6.2 Including the Client in the Design Process

People who ask for software from others rarely know exactly what they want when they first ask. Even if they are well versed in both the application domain and the computer technology involved, decisions about very fine details of an information system can seriously affect their perception of that system. As a system develops, it takes on a character of its own and, for good reasons, may drift from the original specifications. Once built, a system may demand too much familiarization time and effort from the client. The obvious way to minimize these problems is to include the client in a significant portion of the design process. As details are refined, the client can react before those details are so deeply embedded in the system design that change is horrendously expensive. As the system evolves, client involvement ensures that mutations are acceptable or at least understood. Being involved in the design gives the client a head start on familiarization. There are many informal and formal approaches to involving clients in software design, ranging from one extreme of presenting the client with *faits accomplis* to take or leave when the project is done to the other, so-called "Scandinavian approach" of having the entire user community be part of the design process throughout the life of the project. Let us look at two approaches. One, called joint application design (JAD), is in the middle of the range, and the other, called participatory design (PD), is an example of full involvement.

JAD began as a way for clients to refine their requirements before presentation to software designers but has evolved into a joint effort on both requirements and design involving clients and software designers. JAD is a very formal method based on meetings. Each meeting is attended by the clients and system designers. The meetings are "facilitated" by a meeting leader. The facilitator must be a neutral party adequately trained to lead the meeting and well versed in system development methods, but not seen as an advocate for either the users or the developers. The facilitator controls the agenda and the discussion and ensures full documentation of the meeting results, usually using a person designated as "scribe" to capture and preserve all the design work done at the meeting. Visual aids, and sometimes highly automated design systems, are used to help present ideas for discussion with maximum clarity.

PD is a less rigid approach with more intimate and continuous user involvement. The focus is brought out of the meeting room and into the workplace where the system will be used. Users and software designers try to teach one another about their working environment and the range of technical options in what is called "mutual reciprocal learning" and do the "design by doing," using hands-on approaches that let the users work with simulations of systems to get a feel for what is being proposed and show the designers what they are doing or want to do with the system. The approach may be very low-tech, with cardboard models, or medium-tech, with simulations on modest personal computers. The demands of such detailed interactions are beyond the capabilities of most existing higher-technology systems.

These approaches both have had some success, but JAD, with its meeting orientation and formality, has had its greatest impact in helping users to clarify their requirements, and its impact tends to fade in the later design stages until prototype systems are available for review. PD maintains a deeper involvement further into the design cycle and, by forcing designers to mock up their ideas before the prototype stage, keeps the development closer to evolving user needs. JAD tends to be a middle-manager methodology, while PD provides more involvement from lower-level line people. Because of that, use of both approaches could be complementary, provided means were found to ensure a commonality of goals between the two groups. For more information on JAD and PD, see Muller and Kuhn (1993) and Carmel *et al.* (1993). For a discussion of a JAD-derived approach for the implementation stages, called the joint implementation process (JIP) involving implementation coordinators (ICs) working in dynamic unstructured environments, see Fallon (1995).

7. QUALITY ASSURANCE AND CONTROL

Quality assurance and quality control begin with a simple goal: to assure ourselves that what we produce will function for its intended customer in its intended manner. In manufacturing industries this effort has been formalized into two related disciplines: quality assurance (QA), which is the overall effort by management to assure itself of the quality of a product, and quality control (QC), which is a control system integrated with the manufacturing process as a mechanism to assist in QA.

Some aspects of QA and QC can be mechanized. One can have inspection and test procedures that are imposed on intermediate and final products. One can use languages and processes that include cross checks for internal consistency and for conformance with machine-readable specifications. But in most respects, software engineering remains a human endeavor and quality can only result when all the people in all the critical paths to creation of a sound product are committed and have the skills and authority to turn that commitment into action. As Demming (1991) notes in commenting on American industrial ills, "measures of productivity do not lead to improvement in productivity [and] . . . new machinery and gadgets are not the answer." When an institutional structure isolates line people from the decision making about their activities with many layers of bureaucracy, tries to treat people as interchangeable parts in a machine, actively encourages them to close their eyes to defects in what they are producing, then quality suffers. There is a growing realization that the quality of production and service efforts can be improved by what amounts to an object-oriented structuring of the organizations that create products and services. This consists of bringing the actions together with the structures upon which they act by creating small functional teams with as much decision-making autonomy as possible and by keeping the cross communications among teams as simple and effective as possible. See the discussion of total quality management by Creech (1994).

There have been serious efforts to bring a sound understanding of QA/QC into software engineering. One very popular effort in the United States has been the capability maturity model (CMM) developed by the Software Engineering Institute (SEI). In Europe efforts have focused on the International Standards Organization (ISO) standards in the 9000 series, particularly ISO 9001.

The SEI CMM focuses on five levels of organizational maturity in the processes of software production. At the lowest level there is no formal structure. On the next level, the organization has built the essential tools to be able to repeat past successes in producing software. It has mechanisms supporting the waterfall model to understand requirements, to plan and track projects, and to verify the quality of the software produced, and it has a software configuration management system. On the third level, the organization-wide scope of software development processes is recognized, creating one organization-wide approach to documentation, standardization, and integration of software. On the fourth level, management of software processes, quantitative measures are used for control of the process. On the fifth and highest level, the organization enters into a phase of continuous improvement using quantitative feedback and anticipating future directions. The view is one of bottom-up growth of quality, with quality spreading into an existing organization and changing the internal culture of the organization.

The ISO 9000 series approach is more

top-down, starting with a top-level organizational commitment to deliver what the organization's customers have asked for. An ISO 9001–compliant organization will have a fully documented quality system integrated into the entire production process, mechanisms to ensure a match between customer requirements and software planning, mechanisms to plan and track projects, mechanisms to fit in externally acquired software whether from third parties or from the customer, mechanisms to trace the life of all software products from conception to creation to installation, mechanisms to plan and control the process of creating software, etc. In most respects, ISO 9000 corresponds to SEI CMM level 4 and does not explicitly address the issue of continuous improvement. However, a lack of commitment to improve and a failure to create mechanisms for continuous improvement would indicate a failure of the top-level organizational commitment on which ISO 9000 is based. See Paulk (1995) for a discussion of the SEI CMM and the ISO 9000 series.

It is interesting to consider the interaction of attempts to achieve quality software with the evolution of organizations into dynamic unstructured environments. Especially in view of the organization-wide demands of the SEI CMM approach and the ISO 9000 approach, one might assume that quality requires imposition of more formal organizational structures. This is generally untrue. Indeed the move to "empower" small teams by giving them autonomy and responsibility has taken place in order to improve quality. Quality is improved by tightened lines of communication for decision making and the emotional commitment of people to a process about which they feel a personal sense of control. But there is also a more subtle reason. Products built by small groups have a better chance of being simple enough to understand. Where it is not possible to handle an entire project within a small cohesive group, quality can often be maintained by having a single facilitator or implementation coordinator who, by having to understand the entire project him/herself, implicitly forces simplification onto the process. KISS is an essential element of systems that produce quality software.

8. EXAMPLES OF TOOLS AND TECHNIQUES

Software engineering draws on tools and techniques from computer science, mathematics, and many other disciplines. A good software engineer will be able to use a variety of programming languages and software libraries. He or she will be proficient in establishing appropriate data structures, be adept at sorting and searching, linear algebra, calculus, the theory of approximation, etc. At this writing, however, two areas of expertise are essential in almost all software engineering problems: an understanding of the basics of communications networks and the use of graphical user interfaces (GUI).

8.1 Communications Networks

The design of information-processing systems requires careful consideration of the interactions among different flows of information. Information from two sources may have to be combined before a critical step can be taken. Portions of an information-processing system may be separated by relatively slow and unreliable communications links. The tools and techniques of network design are of increasing importance in software engineering. We will look at two of the most important concepts from network design, which apply to almost all information-processing systems: that queuing-system response times rise sharply as capacity limits are approached, which teaches us how to allocate sufficient resources to produce a system with good response, and that the flow in a network is limited to the capacity of the narrowest bottleneck, which explains the need to focus design effort in removing the narrowest bottlenecks in a system.

In most cases, each subcomponent of an information-processing system may be viewed as a server of processing requests, acting very much like a teller at a bank. Customers for service arrive in some pattern of arrivals, are placed in some order for eventual processing, and then are served with service times that depend on the nature of their particular requests and the present capabilities of the server.

The study of such systems is called queuing theory, which may be viewed as a subfield of the study of "stochastic processes" (see, for example, Karlin, 1969). Stochastic processes are those governed by the behavior of some family of random variables, $R(t)$, where t is an index variable, often representing time. For a particular value t_0 of t, we do not necessarily have a unique value $R(t_0)$. Rather, we know the probabilities that $R(t_0)$ will assume particular values. By probability, we mean the limiting value over infinitely many trials of the ratio of the number of times $R(t_0)$ assumes the specified values to the total number of trials.

The model we use in dealing with queues considers customers and servers. We have a supply of customers requiring service. We have servers to provide that service. We do not know when any particular customer may arrive, nor do we know how long a server will take to service a particular customer. We only know the general pattern of arrival of customers and provision of service. For example, we might know that customers arrive at an overall average rate r_c, with no dependence between the arrival of any two customers and that servers serve at an overall average rate r_s, with no dependence between the service time for any two customers.

Such customer and server random variables are said to obey Poisson distributions, with a mean interarrival time of customers of $1/r_c$ and a mean service time of $1/r_s$. If there are m servers, such a model is called an M/M/m system, where M stands for Markov, since Markov chains are an effective tool to study such systems. More generally, if the customer random variables obey a rule X, and the server random variables obey a rule Y, and there are m servers, then the queue model is called an $X/Y/m$ system. In Fig. 5, we see two $X/Y/1$ queues and one $X/Y/3$ queue. For most networks encountered in software engineering problems, using M/M/1 queues provides a sufficiently accurate approximation to the behavior encountered, and we use M/M/1 queues as the primary model.

Analysis of M/M/1 queues shows that the average queue length L is given by

$$L = \rho/(1 - \rho), \quad \rho = r_c/r_s,$$

where ρ, the ratio of the average rate of arrival of customers to the average service rate, is known as the utilization. The average waiting time W can be derived by "Little's Formula" (Little, 1961),

$$L = r_c W,$$

from which we conclude that both the average queue length and the average waiting time rise rapidly toward infinity as the average rate of arrival of customers approaches the average service rate. For example, while we have a queue length of 1 for 50% utilization, the length rises to 4 for 80% utilization

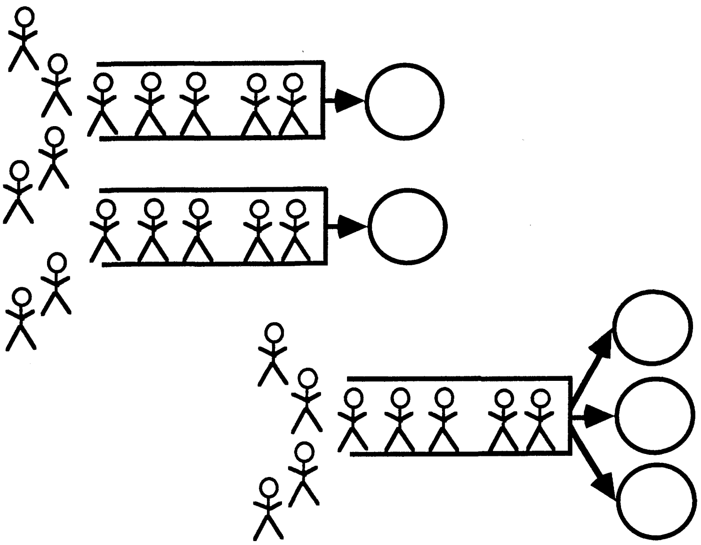

FIG. 5. Two queues with one server each and one queue with three servers.

and to 100 for 99% utilization. Behavior similar to this "soft knee" in an M/M/1 queue is seen in most information-processing systems. For that reason, it is a common rule of thumb to keep at least 20% of the capacity of a system in reserve for almost all systems and a full 50% for systems that must be particularly responsive. For more on the behavior of queues in information-processing systems, see Schwartz (1987).

A full system consists of more than one queue. It will consist of networks of queues, in which the requests serviced by one or more queues become the customers for other queues. We can draw a picture of an information-processing system as a set of processing nodes connected by links along which information flows. The nodes and links joining them form what is called a "graph," which, when we assign a specific direction to each link and forbid links from a node to itself, is called a "network." The study of graphs and networks is called graph theory. For a basic grounding in the subject, see Busacker and Saaty (1965).

When we draw a picture of an information-processing system, we can either draw the precise physical characteristics or abstract only the logical connectivity. If we record the physical locations of the nodes as points in three-space and the paths of the lines as curves in three-space, the sets V of points and E of curves constitute a "geometric graph". While there are design problems that require exact knowledge of node locations and line paths, as in cabling a room full of equipment, much can be learned from the abstract topology of the graph of a system.

Let v and w be nodes, A v-w cut is a set of links, the removal of which disconnects all paths from v to w. A cut is called minimal if restoring any of the links in the cut will reconnect the vertices in question. The capacity of a link is the limit on the rate of flow of information it can carry (e.g., 1000 characters per second). The capacity of a cut is the sum of the capacities of the links in the cut. In Fig. 6, the capacity of each link is given by the number beside it. We have marked a v-w cut of links with total capacity 14, which is less than the capacity of any other cut that could have been used to disconnect v from w.

We can use judiciously chosen minimal cuts to analyze feasible flows between nodes. It is important to realize that we are only looking at the flow between pairs of vertices, not at some generalized total flow in the network.

If we have a feasible flow in which v is the only source and w is the only sink, we say the flow is from v to w. The magnitude of the flow cannot exceed the capacity of a v-w cut (*a fortiori* of a cut of minimal capacity), so the maximal flow from v to w is bounded above by the capacity of a v-w of minimal capacity.

Further, if the maximal flow could not reach the capacity of a cut of minimal ca-

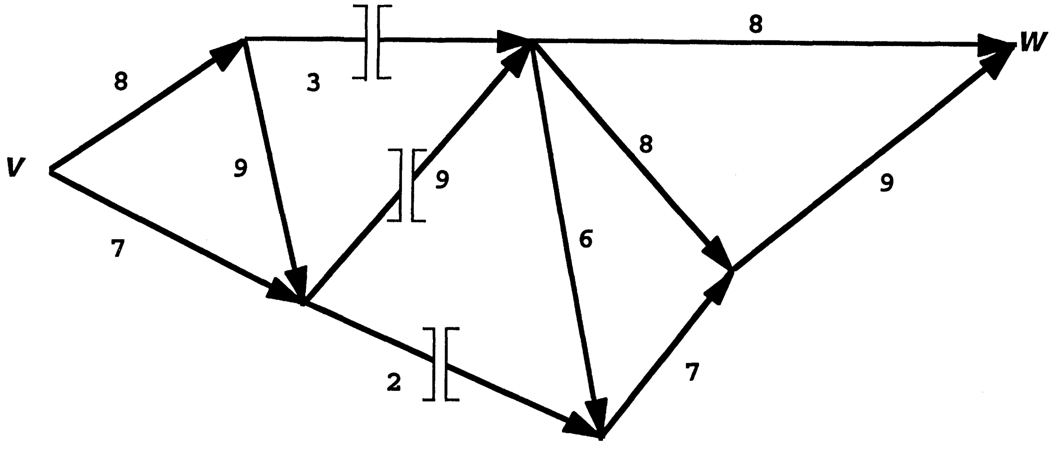

FIG. 6. A network with a v-w cut of minimal capacity.

pacity, then the limitation on flow would be due to a cut of still smaller capacity, a contradiction. Thus we have the Ford-Fulkerson "Max-Flow-Min-Cut" theorem:

> The maximal feasible flow from *v* to *w* is equal to the capacity of a *v*-*w* cut of minimal capacity.

This result is critical to efficient system design. It is saying that we cannot make the flow through an information-processing system be any better than the flow permitted by the capacity of the narrowest bottleneck, that it simply does not pay to put great effort into improving the performance of arbitrary portions of a system. If we want users to see a payback from the effort, we must focus our efforts on those portions of the system where the information flows are constricted. If we combine this with what we have learned about M/M/1 queues, above, we can see that much of the work in software engineering has to be concerned with identifying information flows and capacities and ensuring that we are providing more than adequate capacity at the critical bottlenecks of a system.

8.2 Graphical User Interfaces

A control system for an experiment from the nineteenth century and the first half of the twentieth century would provide its users with a control panel consisting of buttons, switches, dials, and meters. The earliest computers were controlled in a similar manner, but by the 1960s computers and apparatus controlled by computers were managed by typing in commands and looking at lines of characters on paper or on a display screen to see results. For example, to position an axis of a positioner and view related detector output, the user might engage in the following dialog (shown with user responses in bold face):

```
COMMAND: POSITION
WHICH AXIS? THETA
THETA = 25.08, NEW THETA? 30.105
POSITIONING, PLEASE WAIT
POSITIONING, PLEASE WAIT
THETA = 30.105
COMMAND: DETECTOR
DECTECTOR = 1.287V
COMMAND:
```

A lot of good science was done this way, but it is a lot more cumbersome than holding down a switch and watching a position indicator. Graphical user interfaces (GUIs) have allowed for the creation of the modern equivalent of the nineteenth-century control panel on a computer display screen. One can create the necessary software from scratch, but there are now many sound packages that allow rapid and reliable creation of GUI interfaces to information-processing systems. We will look at two of the most popular, Apple Computer's Hypercard® and the so-called "web browsers" based on the hypertext markup language (HTML). Both are examples of hypertext systems, systems that allow documents to be constructed from fragments of text, sound, images, and applications in order to assemble complex assemblies of information with active links among the network of fragments.

Hypercard is an example of a system in which we work to the "WYSIWYG" ("what you see is what you get") paradigm. Hypercard is not the most sophisticated of the GUI-oriented application-building packages, but it was one of the first and is still heavily used, and it illustrates the essential features of such WYSIWYG packages. HTML makes it feasible to build network-based GUI-oriented applications with a high degree of platform independence. In order to provide such independence, there is no assurance that the interface presented on one system will be the same as that presented on another, just that it will work in well-understood and predictable ways. Web browers based on HTML have made distributed graphical access to applications and data commonplace.

If we were to design a control panel for a four-axis positioner, it might look like Fig. 7. The four axes have been labeled with Greek letters, and for each axis, there is an accurate numerical display of the position, buttons to decrease or increase the angular position, and a panel-meter display of the angular position. The detector output is shown as a strip chart. In a full system, we would use a strip chart with more detail, logging the angles in different colors or line styles along with the detector output, but the amount of detail in Fig. 7 should be sufficient for discussion of the graphical user interface. In Fig. 8 we see a partially built con-

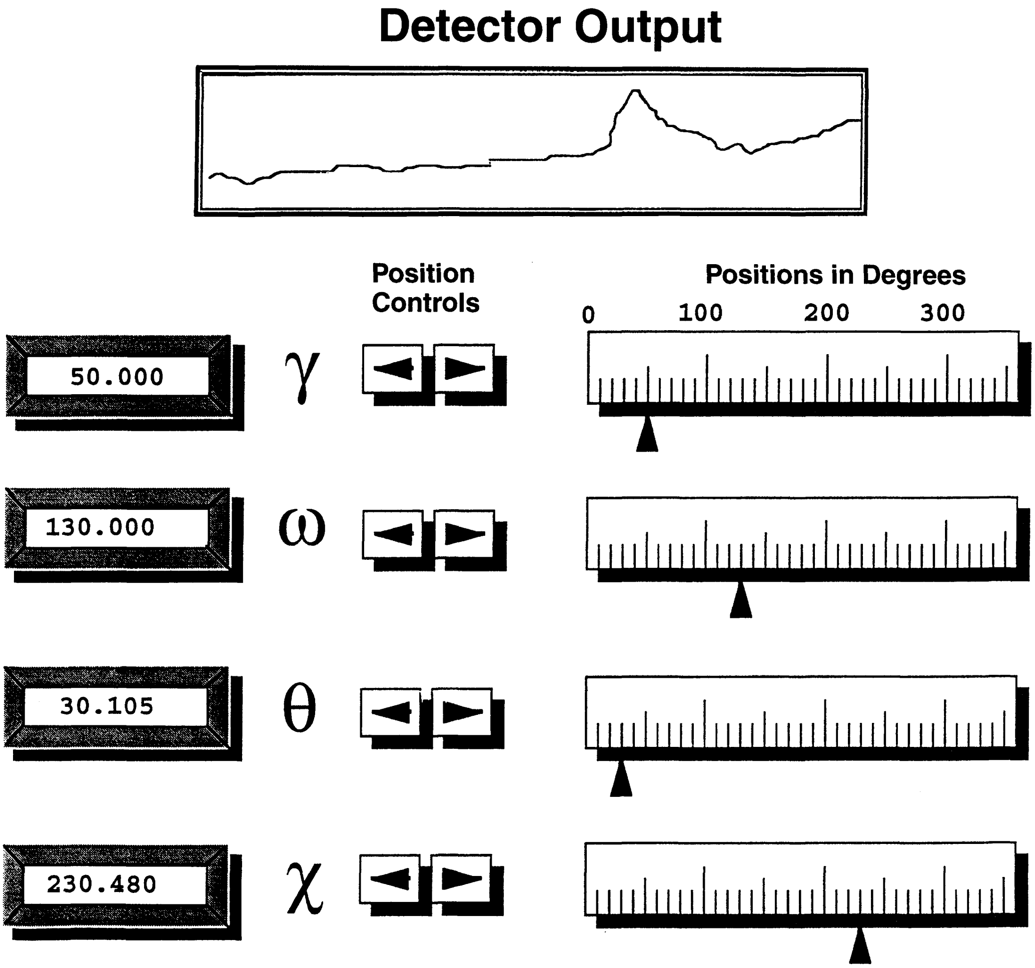

FIG. 7. Design drawing of control panel for four-axis controller.

version of the design drawing into a Hypercard stack. The position controls for the gamma axis have been overlaid with active buttons that can be "pushed" by clicking or holding down the button on a pointing device such as a mouse or trackball. The graphic position indicator on the scale for the gamma axis has been replaced with another button, which can be dragged to the desired angle by placing the pointer into the position indicator and then holding down the button on the pointing device while moving the pointer left or right. The numeric position indicator has been overlaid with a field that displays the current position as the control buttons are pushed or the graphic position indicator is dragged. The user also has the option of typing an exact position into the field. Later in the building of the interface, we would take a section of the display screen for the detector output strip chart and periodically pick up a strip-chart image being maintained by an independent data-logging task.

What speeds the process of creating such useful active images is that they are themselves created by the same sorts of click and drag operations as we just described. For each button, a template of a button was dragged with the mouse to the desired location and then shaped by pushing or pulling on the corners. This saves much tedious measuring with a ruler or the use of graph paper.

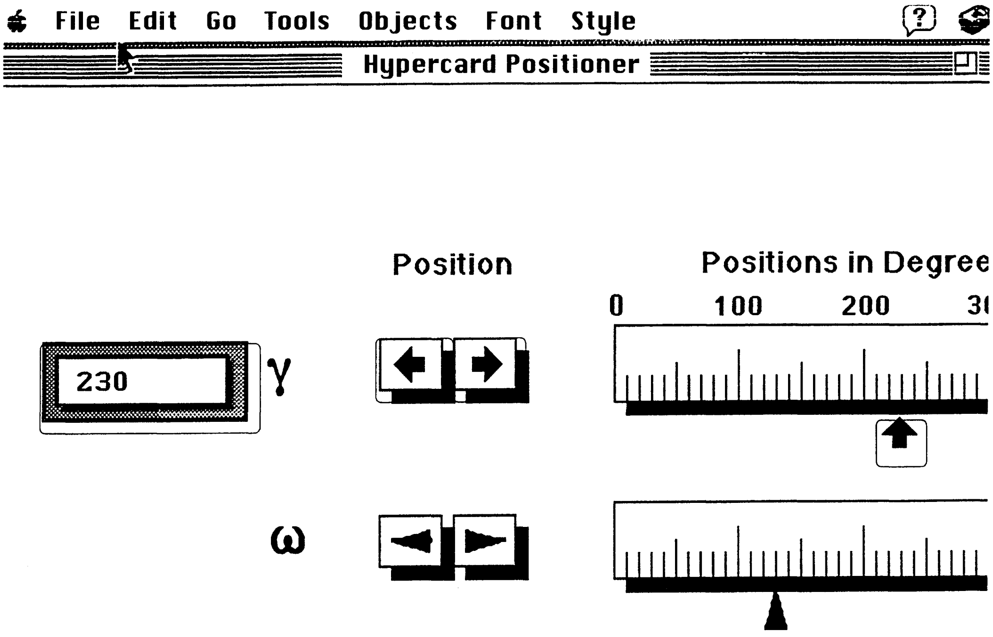

FIG. 8. Screen dump of Hypercard four-axis positioner partially built.

There is much more possible with the use of packages such as Hypercard than just convenient ways to draw images of information-processing systems. Each of the objects in the picture, as well as the entire picture itself, can have an active program, called a "script" in Hypercard, associated with it. Within the scripts are handlers for the messages that may be sent to the object, for example, by pressing the button on a pointing device. Here is part of the script for the positioning button for the gamma axis that decrements the position:

```
on mouseUp
global gammaPos, gammaRate, gammaStart
put .001 into gammaRate
end mouseUp

on mouseDown
    global gammaPos, gammaRate, gammaStart
    put gammaPos into gammaStart
    put .001 into gammaRate
    put gammaPos - gammaRate into xx
    if xx < 0 then put xx+360 into xx
    put xx into gammaPos
    pass mouseDown
end mouseDown

on mouseStillDown
    global gammaPos, gammaRate, gammaStart
    put gammaStart-gammaPos into gd
    if gd < 0 then add 360 to gd
    if gammaRate < .01 then
        if (gd > .01) then
            put trunc (gammaPos*100) into gp
            put gammPos*100-gp into gr
            add .001 to gammaRate
            if gr/100-gammaRate < 0 then
                put gp/100 into gammaStart
                put gp/100 + .01 into gammaPos
                put .01 into gammaRate
            end if
        end if
    else
            .
            . [portion of script omitted]
            .
        end if
    put gammaPos - gammaRate into xx
    if xx < 0 then put xx + 360 into xx
    put xx into gammaPos
    pass mouseStillDown
end mouseStillDown
```

We have provided handlers for three messages in this script, one for the case that the mouse button has just been released, one for the case that it has just been pressed, and one for the case that it is still being held down. Notice that all these handlers do is manage the conversion of information from the state of button presses to two numbers, one held in the global variable gammaPos and one held in the global variable gammaRate. A third global variable gammaStart is used to keep track of when the user started holding down the button. The actual work of displaying the position and managing the actual axis controller is done elsewhere, in the scripts of other objects, which peek at the state of these variables and are also notified of the need for an update by the "pass" commands that cause the messages that were received by these handlers to be passed on to other objects by the rules of inheritance. This incomplete style of programming, where we see only a small part of the effects of an action, this programming by "side effects," is typical of object-oriented designs and appropriate for the design of systems that must handle multiple cooperating autonomous processes. Notice also that the bulk of the code in the script for the case of the button still being held down derives its complexity from human engineering concerns. When we first press the button, we wish the axis to take its highest resolution steps. As we hold it down, we wish the axis to move faster and faster. Therefore we keep increasing gammaRate first in unit steps within each decade and then bump it to the next decade. However, in order to produce a display that causes minimal confusion, it is important to make the transition on rate decades only when the position is on a decade boundary, but also to postpone the transitions long enough to allow for realistic reaction times by human operators. For more on Hypercard and hypertext in general, see Nielsen (1990) and the references therein.

Now suppose we want to allow a similar four-axis control to be operated from a variety of terminal types. We want to design our software just once to allow position control from high-performance workstations, medium-scale graphics-oriented personal computers, and text-oriented terminals. The display station might be directly coupled to the system controlling the position or connected via a wide-area network with long delays. For experimental facilities in the physical sciences with visitors from other institutions, this is a common requirement.

The most popular means of solving such problems is to design the interface for the control application using HTML, the hypertext markup language. This language allows us to specify a hypertext document in broad functional terms, leaving it to specific browser applications on the particular display systems to decide how best to present them. Thus, although HTML documents can be edited with WYSIWYG editors, it is important to realize that the result is a simple text document marked up with minimal instructions for display. For example, the portion of an HTML document to provide control of the gamma position for a four-axis positioner might be as shown in the HTML fragment in Fig. 9. The portion of the image produced on a simple browser by this sec-

```
<!doctype html public "-//IETF//DTD HTML 2.0//EN">
<HTML>
<HEAD><TITLE>Positioner Control Page</TITLE></HEAD>
<BODY> ...

<FORM ACTION="...[positioner server goes here]/cgi-bin/script">
<p> Position Controls:
<H2> Gamma </H2>
Position:<INPUT TYPE="text" NAME="Gamma_Position" VALUE="000.00" SIZE="7">
<menu>
<li><INPUT TYPE="radio" NAME="Gamma_Step" VALUE="up"> up
<li><INPUT TYPE="radio" NAME="Gamma_Step" VALUE="down"> down
<li><INPUT TYPE="checkbox" NAME="Gamma_Rate" VALUE="slow"> slow (.01)
<li><INPUT TYPE="checkbox" NAME="Gamma_Rate" VALUE="medium"> medium (.1)
<li><INPUT TYPE="checkbox" NAME="Gamma_Rate" VALUE="fast"> fast (1)
<li><INPUT TYPE="checkbox" NAME="Gamma_Rate" VALUE="gallop"> gallop (10)
    </menu> ... </FORM>
```

FIG. 9. Fragment of HTML document for one axis of four-axis positioner.

tion of the HTML document is shown in Fig. 10. Note that we have checked the box for "slow" positioning and clicked the "up" positioning button. We could just as well have typed in a new position within the position box. These actions are sent back to the server by the browser, and it will reply with an updated form with an updated position. Different browsers may use different graphics for the check box or the position or other elements, or even do them as straight text, but they all would look the same to the server updating the position. Pictures as good as and better than the examples we used for Hypercard, with portions made active for detection of the position of a pointing device, can be used, as well, but it is a good practice to provide the sort of simple alternative shown here so that even the most primitive display terminals can be used.

HTML and related approaches are evolving rapidly. They have drastically changed perceptions and expectations about what can be done with computers. Coupled with an understanding of the data communications networks upon which they are based, they are becoming the most visible, if not necessarily the most important, of the tools of software engineering. For more on HTML, see Graham (1995).

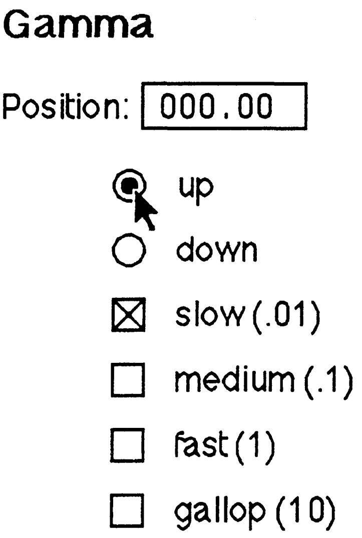

Gamma

Position: 000.00

● up

○ down

☒ slow(.01)

☐ medium(.1)

☐ fast(1)

☐ gallop(10)

FIG. 10. Browser display for gamma axis.

9. FUTURE DIRECTIONS

Software engineering is a far-from-mature discipline. There is no general agreement on the practical details of how to produce quality software reliably and consistently. Methods that work in one organization may fail completely when transplanted to another. No single method has been a complete success. Top-down analysis in structured programming, side-effect–free applicative programming, using ADA as an environment for software engineering, side-effect–rich object-oriented programming, and bottom-up synthesis to do encapsulation all have advantages and limitations. Where do we go from here?

Software engineering exists within a context of rapidly changing hardware. In the near future we can expect increasing uses of networks of computers as tightly coupled symmetric multiprocessor configurations, as loosely coupled groups intentionally joined to solve cooperatively pieces of a larger problem, and as randomly interacting communities of computers dealing with one another as autonomous sources, transformers, and sinks of information. Software engineering in such environments will have to merge with a more general concept of information systems engineering in which hardware, software, and human-based systems are all components of larger abstract systems for the processing of information and in which the system has to transcend changes in any of these areas. Integrated approaches to hardware, software, and human interactions are needed. See White *et al.* (1993) for a discussion of "Systems Engineering of Computer-Based Systems."

If one were to predict the major impact of such pressures, it would be for an increasing emphasis on the basic tools of top-down analysis, bottom-up synthesis, object-oriented representation, and use of software to process software with an emphasis on simplifying and delayering the processes, as we have already seen in industrial processes. Rather than relying on separate teams specializing in discrete steps in the waterfall model, there will have to be more emphasis on broadly trained people able to follow a project through from beginning to end, able to understand all aspects from the applications to systems software and hardware, and able to follow such processes more efficiently with fewer handoffs.

In the longer term, as distributed processing expands and self-defining processing systems such as neural nets are used, we can expect the need for probabilistic approaches to increase. When one cannot be certain where the "latest" version of information exists, or if it has been lost, delayed, or garbled, when information becomes soft and fuzzy, the best we can do is estimate probabilities that a system is doing what we want and work to make the risk acceptable (see THEORY AND APPLICATIONS OF FUZZY LOGIC, and Graf and Reyneri, 1995). We will learn to build in redundancy, having multiple systems handle the same tasks with as much independence as possible and compare results. Complete uniformity of process will become a source of risk rather than of reliability, since such uniformity would remove the possibility of truly independent checks. The standard accounting practice of one person doing and another checking will have to extend to software systems design.

As always, there will be panaceas presented, perfect methods and systems that will produce perfect software directly from vague statements of requirements without the need for messy, uncooperative human efforts. Some of these systems will work for some problems, but the truly challenging problems in software engineering, the ones that require vision and commitment, will remain a human endeavor.

ACKNOWLEDGMENT

We wish to thank Frances C. Bernstein and Michael E. Bernstein for their helpful comments and suggestions.

GLOSSARY

Abstract Data Type (ADT): Abstract definition of a class of objects giving the data structures, actions, or methods that may be applied and giving the relationships to other ADTs.

Action: Anything an object may do, e.g., by changing its internal state or by sending a message to another object.

Bottom-Up Synthesis: Building a system or product by hierarchically assembling simple tools and resources into more complex and abstract ones.

CASE: Computer-aided software engineering, the use of computer-based tools to automate steps in software development.

Class: An **Abstract Data Type.**

CMM: SEI capability maturity model, an approach to quality assurance based on increasing levels of control over the production process.

Encapsulation: Creation of more abstract classes from more concrete ones, hiding internal methods and data structures not needed for external actions, a bottom-up technique.

GUI: Grapical user interface.

HTML: Hypertext markup language, a language for the creation of hypertext documents that supports text fragments, images, sounds, movies, and application fragments distributed over wide-area networks.

Hypercard®: A registered trademark of Apple Computer Inc., the name of a GUI-based application building system with hypertext capabilities.

Hypertext: A network-oriented document structure providing active links among distributed document fragments.

Inheritance: Definition of a class in terms of another class, so that the new class inherits the data structures and/or methods of the old. Often considered an essential characteristic of object-oriented programming. Inheritance also is used to denote the implied passing of messages up a hierarchy of actual instances of objects created by inheritance among classes.

ISO 9000: An approach to quality assurance promulgated as a suite of standards by the International Standards Organization (ISO). The approach is based on top-level managerial commitment to quality.

Message: Information passed from one entity to another, possibly to itself. For objects, often used interchangeably with **Action.**

Method: Action a class may take.

Multiple Inheritance: Inheritance by a new class from more than one old class.

Object: A specific instance of an **Abstract Data Type,** i.e., an entity in an information-processing system stated in terms of the data it can manipulate, the actions it can take on those data, and its relationships to other objects.

Object-Oriented Analysis (OOA): Analysis of requirements in terms of objects and their relationships. Not necessarily done top-down.

Object-Oriented Programming (OOP): Programming in terms of objects that take actions on the basis of messages passed to them.

Outside-In Analysis and Design: The alternate or simultaneous application of bottom-up synthesis and top-down analysis to a problem, possibly from more than one "top" starting point and from more than one "bottom" set of resources.

Operation: Same as **Action.**

Overloading: Using the same name for different actions in different contexts, e.g., "+" for integer addition, floating-point addition, and logical union.

Polymorphism: The ability to use a common name for multiple classes or methods. See **Overloading.**

Procedure: A **Routine.**

Quality Assurance (QA): Overall management effort to assure the quality of a product, i.e., that the product conforms to the expectations of the customer.

Quality Control (QC): Control system integrated with the manufacturing process as a mechanism to assist in **QA.**

Re-engineering: Using an existing system as a guide in building a new system.

Reverse Engineering: Recreating design documents by analyzing code and external behavior.

Risk: The expected cost of a failure.

Routine: A self-contained portion of code used by other code. It may invoke other routines.

Side Effect: A nonobvious change to the state of data structures resulting from the invocation of a routine.

Software: Information in the form of data and instructions that affects information-processing systems.

Software Engineering: The disciplined use of well-understood techniques to analyze, design, build, repair, or use software.

Structured Programming: Top-down analysis applied to the flow of control of a system.

Top-Down Analysis: Understanding a system or concept by hierarchically decomposing it starting with the most general level of abstraction and working downward toward increasing specificity.

TQM: Total Quality Management, an approach to assuring quality by restructuring organizations into decentralized interdependent semiautonomous teams.

Works Cited

(Some citations appear under Further Reading, below)

Armbruster, J. L. (1995), *Dr. Dobb's J.* **20,** 76–86.

Baskerville, R. (1993), *ACM Comput. Surv.* **25,** 375–414.

Booch, G. (1983), *Software Engineering with Ada,* Menlo Park, CA: Benjamin/Cummings.

Booch, G. (1991), *Object Oriented Design with Applications,* Redwood City, CA: Benjamin/Cummings.

Busacker, R. G., Saaty, T. L. (1965), *Finite Graphs and Networks: An Introduction with Applications,* New York: McGraw-Hill.

Carmel, E., Whitaker, R. D., George, J. F. (1993), *Commun. ACM* **36** (4), 40–48.

Creech, B. (1994), *The Five Pillars of TQM: How to Make Total Quality Management Work For You,* New York: Truman Talley Books/Dutton.

Demming, W. E. (1991), *Out of the Crisis,* Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Dijkstra, E. W. (1968), *Commun. ACM* **11,** 341–346.

Dijkstra, E. W. (1976), *A Discipline of Programming,* Englewood Cliffs, NJ: Prentice-Hall.

Fallon, H. (1995), *How to Implement Information Systems and Live to Tell About It,* New York: Wiley.

Felder, M., Morzenti, A. (1994), *ACM Trans. Software Eng.* **3,** 308–339.

Fuggetta, A. (1993), *Computer* **26** (12), 25–38.

Graf, H. P., Reyneri, L. M. (Eds.) (1995), *IEEE Micro* **15,** 10–11.

Graham, I. S. (1995), *The HTML Sourcebook.* New York: Wiley.

Halladay, S., Wiebel, M. (1993), *Object-Oriented Software Engineering.* Lawrence, KS: R&D Publications, Inc.

Hoare, C. A. R. (1969), *Commun. ACM* **12,** 576–580.

Hollingdale, S. H., Toothill, G. C. (1965), *Electronic Computers,* Harmondsworth, U.K.: Penguin.

Hudak, P. (1989), *ACM Comput. Surv.* **21,** 359–411.

Karlin, S. (1969), *A First Course in Stochastic Processes,* New York: Academic.

Kraut, R. E., Streeler, L. A. (1995), *Commun. ACM* **38** (3), 69–81.

Little, D. (1961), *Operations Res.* **9,** 383–387.

Martin, R. L. *et al.* (1984), *AT&T Bell Laboratories Tech. J.* **63,** 1571–1910.

Morrison, P., Morrison, E. (Eds.) (1961), *Charles Babbage and His Calculating Engines,* New York: Dover.

Muller, M. J., Kuhn, S. (Eds.) (1993), *Communications of the ACM: Special Issue on Participatory Design* Vol. 36, New York: Association for Computing Machinery.

Nielsen, J. (1990), *Hypertext & Hypermedia,* San Diego, CA: Academic.

Nielsen, J. (1992), *Computer* **25** (3), 12–22.

Paulk, M. C. (1995), *IEEE Software* **12,** 74–83.

Pressman, R. (1992), *Software Engineering, A Practitioner's Approach,* 3rd ed., New York: McGraw-Hill.

Pressman, R. (Ed.) (1995), *IEEE Software* **12,** 101–102.

Salvendy, G. (Ed.) (1987), *Handbook of Human Factors,* New York: Wiley.

Sandhu, R. S. (1993), *Computer* **26** (11), 9–19.

Schwartz, M. (1987), *Telecommunications Networks,* Reading, MA: Addison-Wesley.

Siddiqi, J. (1994), *IEEE Software* **11** (2), 18–19.

Sneed, H. M. (1995), *IEEE Software* **12** (1), 25–34.

von Neumann, J., Morgenstern, O. (1944), *Theory of Games and Economic Behavior,* Princeton, NJ: Princeton University Press.

White, S., Alford, M., Holtzman, J., Kuehl, S., McCay, B., Oliver, D., Owens, D., Tully, C., Willey, A. (1993), *Computer* **26,** 54–65.

Winsberg, P. (1995), *Datamation* **41** (9), 36–41.

Withington, F. G. (1972), *The Organization of the Data Processing Function,* New York: Wiley-Interscience.

Wright, P. H. (1989), *Introduction to Engineering,* New York: Wiley.

Further Reading

Blum, B. I. (1992), *Software Engineering, A Holistic View,* New York: Oxford University Press.

Booch, G. (1991), *Object Oriented Design with Applications,* Redwood City, CA: Benjamin/Cummings.

Creech, B. (1994), *The Five Pillars of TQM: How to Make Total Quality Management Work for You,* New York: Truman Talley Books/Dutton.

Dijkstra, E. W. (1976), *A Discipline of Programming,* Englewood Cliffs, NJ: Prentice-Hall.

Muller, M. J., Kuhn, S. (Eds.) (1993), *Communications of the ACM: Special Issue on Participatory Design* Vol. 36, New York: Association for Computing Machinery.

Salvendy, G. (Ed.) (1987), *Handbook of Human Factors,* New York: Wiley.

SOIL POLLUTION

Janick F. Artiola and Mark L. Brusseau, *Soil, Water, and Environmental Science Department, University of Arizona, Tucson, Arizona, U.S.A.*

	Introduction	341
1.	**Nature of Soils**	341
2.	**Sources of Pollution**	345
2.1	Controlled Waste Disposal	345
2.1.1	Landfills	345
2.1.2	Municipal Sewage Disposal	345
2.1.3	Residential Septic Systems	345
2.2	Leaking Storage Tanks	345
2.3	Industrial Manufacturing Plants	346
2.4	Commercial Activities	346
2.5	Energy and Material Production	346
2.6	Agricultural Activities	346
2.6.1	Agrochemicals	346
2.6.2	Feedlots	346
3.	**Measuring and Monitoring Pollution**	346
4.	**Transport and Fate of Pollutants**	349
4.1	Quantifying Transport and Fate	349
4.2	Dispersion	350
4.3	Sorption	351
4.4	Transformation Reactions	352
4.5	Multiphase Pollution	353
4.6	Soil Heterogeneity	353
5.	**Cleaning Up Soil and Water Pollution**	353
5.1	Legislative Programs for Pollution Cleanup	353
5.2	Cleanup Technologies	354
5.2.1	Containment	355
5.2.2	Soil Flushing	355
5.2.3	Soil Vapor Extraction	356
5.2.4	*In Situ* Remediation	356
6.	**Soil Reclamation**	356
7.	**Summary**	360
	Glossary	360
	Works Cited	361
	Further Reading	361

INTRODUCTION

Soil, although often overlooked, is a critical component of the earth and especially for human existence. Soil, directly or indirectly, serves as the major source of food production for humans. All of the structures and dwellings used by humans are built on soil. And, central to the topic of this article, the majority of wastes produced by human activity are purposefully or accidentally released to soil. In this article we define the nature and properties of soils and discuss some of the important interactions between wastes and soil as well as methods for detecting, remediating, and restoring contaminated soil. It should be stressed that the behavior and pollution of water in the subsurface (soil moisture, groundwater) is intimately related to that of soil. Thus, our discussion will, in effect, cover pollution of subsurface water and soil.

1. NATURE OF SOILS

The soil environment is a mass of unconsolidated material that lies between geologic deposits and the atmosphere. Thus, soils cover much of the exposed earth's surface and support many forms of life, including humankind. The soil environment is a highly dynamic system in which a multitude of physical, chemical, and biological processes are constantly operating. Examples of these processes include chemical oxidation/reduction, hydrolysis, physical breakdown of minerals due to temperature changes and mechanical action, and elemental biocycling in organisms that use soils as a growth medium. These energetically diverse processes act on any exposed geologic deposits and transform them into "soil" material.

Soils are formed by the combined contributions of five factors, namely, parent material, climate, topography, biological activity,

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

and time. The geologic deposits provide the parent material from which soils are formed. The composition of the soil will in large part be determined by the parent material. Through their combined actions, climate, topography, and biological factors produce the soil by weathering the parent material. The soil-forming processes operate through the course of time, the fifth factor.

Soil weathering processes decrease rapidly with depth, leading to the formation of soil horizons. These layers, typically parallel to the surface, are studied by soil scientists and provide descriptions of soil-forming processes acting on the horizons of soils. A typical soil profile will have no fewer than three horizons named A, B, and C (see Fig. 1). The A horizon is the layer of soil directly exposed to atmospheric, physical, chemical, and biological processes. This horizon is manipulated to sustain agricultural food production. Horizon B lies directly below A and is considered a zone of accumulation as well as a transitional zone between the exposed A horizon and consolidated geologic material. Horizon C signifies the beginning of unweathered geologic material that may or may not have been the origin of the overlying soil.

Soil is a heterogeneous medium consisting of four major components. The solids component consists of the inorganic materials (e.g., sand, silt, clay) that serve as the support structure of soil. This component is probably what most people think of as soil. However, soil also contains liquids, gases, and biological matter. The liquid component is a solution of water and many dissolved chemicals and is referred to as the soil solution. The gas components that form the soil atmosphere consist of air and fractions of inorganic and sometimes organic gases. The biological component consists of both living (e.g., bacteria) and dead (soil organic matter) material. These components will be discussed in greater detail in forthcoming paragraphs.

Soils are classified first by their physical properties and second by their chemical properties. The commonly accepted triangular textural classification of soil based on percentages of sand, silt, and clay size particles is presented in Fig. 2. It should be noted that clay particles, because of their small diameter, have up to 10 000 times more surface area per unit weight than do sand particles. Thus, clay minerals have unique surface properties that affect many soil physical and chemical processes and will be briefly discussed.

Initial soil classification methods assumed that all clay-size particles are composed of

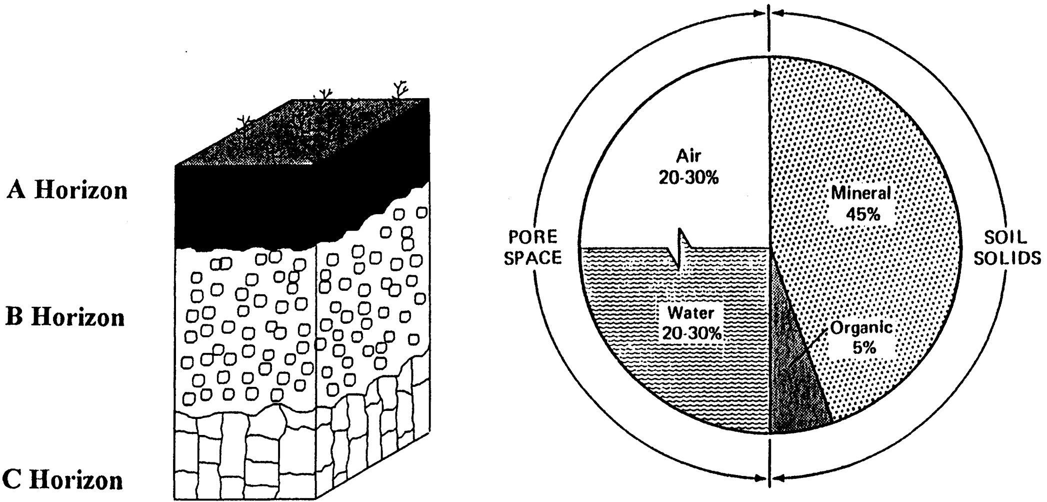

FIG. 1. Soil profile with major soil horizons (A, B, and C) and typical composition (air, water, minerals, and organic matter) of the A horizon (~plow layer). Modified from Buol *et al.* (1973) and Brady (1974).

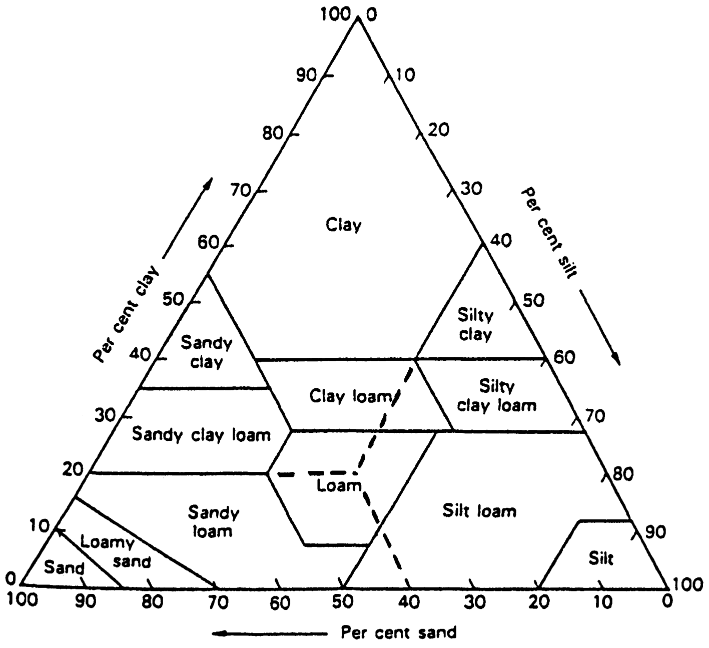

FIG. 2. Textural classes of soils, United States Soil Survey classification. Sand 2.0–0.05 mm, silt 0.05–0.002 mm, and clay less than 0.002 mm in diameter. A typical loam may have about 40% sand and 40% silt particles and about 20% clay. (Buol *et al.*, 1973.)

weathered primary minerals that are called clays, secondary minerals, or phyllosilicates (discussed below). However, it is not uncommon to have in soils significant amount of clay-size particles composed of minerals based on iron, aluminum, and/or manganese oxides, calcium and magnesium carbonates and sulfates, and even organic matter. When these components are found in small quantities in soils (<5%), they act as binding agents among clay, silt, and sand particles. But, when found in larger concentrations, these components can impart irregular and unpredictable physical and chemical properties to soils. Nonetheless, most clay soils have properties distinctly different from sandy soils. Of particular importance is the ability of clay particles to adsorb and to retain water via capillary suction within the micropores of clay particle aggregates. Thus, the water retention capacity of soil decreases rapidly with increasing percentage of sand. Sand particles form large pores that drain easily in response to gravitational forces and do not retain much water.

Clay minerals form during the weathering of primary silicate (silicon based) minerals such as micas and feldspars (Birkeland, 1974). The process of weathering includes selective removal or replacement of elements such as potassium, magnesium, silicon, and calcium from the crystal structures of the primary minerals (aluminosilicates). Clays such as montmorillonite may also be formed by the recrystallization of some minerals such as vermiculite. Substitution of elements such as aluminum(III) for silicon(IV) or iron(II) or magnesium(II) for aluminum(III) results in a gain in net negative crystal charge and gives rise to the negative surface charge of clay minerals. Thus, these minerals have significant cation exchange properties and readily exchange soluble cations with the soil solution. The interactions between negatively charged surfaces of clay and soluble cations have profound implications in soil–plant fertility relationships, as well as the movement and attenuation of inorganic cationic pollutants such as cadmium(II) and lead(II) in the soil environment.

Soils with shrink-swell properties contain dominant amounts of expanding clay minerals, such as montmorillonite (smectite) and vermiculite. Soils with visible surface cracks usually contain significant amounts of these clays. These clay minerals have an expanding interlayer that can gain (absorb) or lose water depending on the amount of water available and thus change crystal dimensions. Kaolinite, a common clay mineral used in ceramics, does not have expanding crystal interlayer spaces and does not exhibit shrink-swell properties during wet-dry cycles.

Organic matter affects the nature and properties of soils in several important ways. The bulk of the organic matter found in soils has its origins in the cycles of plant growth and decay. Dead plant tissue decomposes and undergoes a series of microbially and chemically catalyzed reactions that lead to the formation of humic substances in the soil environment. The mechanisms for the formation of humic substances are not completely understood. But they may include polymerization of small fractions of oxidized organic matter and the restructuring of partially oxidized natural macromolecules such as lignin, the major component of wood. Humic substances have very complex and variable three-dimensional macromolecular structures of polymerized aliphatic and aromatic hydrocarbons (MacCarthy *et al.*, 1990). In general, humic materials have an outer structure with ionizable groups that make them stable in aqueous soil environments. However, the inner structure of these substances appears to be composed of less polar (more hydrophobic) groups. These combined external and internal properties impart to soil humic substances a "reactivity" for organic and inorganic substances. Thus, humic substances play a key role in the fate and transport of pollutants such as metals and of organic chemicals such as pesticides, industrial solvents, and petroleum hydrocarbons.

Humic substances accumulate in the soil environment in response to several environmental factors associated with climate, landscape, and plant species. For example, forested soils in cold climates accumulate several times more organic matter than prairie soils in hot semidesert environments. Humic substances act as binding agents for soil mineral particles and can promote the formation of soil aggregates. These aggregates improve water holding capacity, water movement, nutrient availability, and oxygen-gas transfer properties of soils.

Soils have distinct colorations that extend from black to white and from green to red. The two dominant soil fractions that control color are organic matter, which darkens soils, and iron oxides, which impart reddish-brown tones to otherwise gray silicate minerals. Desert soils that have dominant fractions of calcite and gypsum minerals can have light beige to near-white colorations that may be accentuated by the accumulation of sodic salts. Less common green-colored soils may be found in buried waterlogged soil environments and are due to the formation of reduced iron and manganese minerals.

The soil environment is very rich and diverse in organisms. A list of types and numbers of organisms that inhabit a typical fertile soil is presented in Table 1. Bacteria are by far the most common and diverse group of microorganisms found in soils. In fertile soils, the number of bacteria per gram of soil ranges in the millions. Bacteria are involved in major nutrient recycling processes such as nitrogen fixation, as well as organic-matter decomposition and chemical transformation processes. Soil bacteria are the most active organisms involved in the biotransformation and detoxification of organic wastes and organic pollutants disposed of and/or de-

Table 1. Organisms in the soil environment (from Brown, 1978).

Microorganisms in fertile soil[a]	(10^6 g^{-1})	Animals in mull soil under beech	(10^6 ha^{-1})	% of total animal mass[b]
Bacteria	1–100	Earthworms	1.8	75.1
Actinomycetes	0.1–1	Enchytraeid worms	5.3	1.5
Fungi	0.1–1	Gastropods	1.0	7.0
Algae	0.01–0.1	Millipedes	1.8	10.6
Protozoa	0.01–0.1	Centipedes	0.8	1.8
		Mites and springtails	44.1	0.4
		Other	7.2	3.6

---

[a] The carbon in soil microorganisms accounts for about 3% of the soil organic carbon; that is, they contain about 1 ton of carbon per hectare in a soil with 1.5% carbon to 15 cm depth.

[b] Total mass of animals was 286 kg/ton of carbon.

posited on land surfaces. Higher organisms such as earthworms are very active not only in organic-matter transformations but also in physical mixing and translocation of minerals in the soil environment.

The nature of soils is very complex and dynamic. Soil scientists must apply fundamental concepts in physics, chemistry, and biology to study and understand the interrelated processes ongoing in the soil environment. Soil systems have a unique capacity to adapt to external stresses. Humankind depends on a healthy soil environment for survival. Thus, humankind must not take the nature of soils for granted by exceeding their ability to cope with external stresses that may damage the soil environment irreversibly.

2. SOURCES OF POLLUTION

There are numerous sources of soil pollution. Some of the sources are the result of planned waste-disposal activities, whereas other sources are the result of accidental releases. Just about all conceivable human activities result directly or indirectly in the production of waste materials. The management, treatment, and disposal of this waste is an important aspect of environmental science and engineering. This aspect is, however, outside the scope of the present article. Brief discussions of the major sources of soil pollution are presented below.

2.1 Controlled Waste Disposal

2.1.1 Landfills Through the course of everyday activity, humans produce enormous quantities of wastes. These wastes consist primarily of paper, plastic, glass, aluminum, and food and plant (e.g., yard) materials. It is estimated that each person in the United States produces approximately 1.8 kg of solid waste each day (Pye *et al.*, 1983). Despite increasing efforts to reduce and recycle this material, much of the waste ends up in landfills. The basic premise of a landfill is to bury the waste in a controlled subsurface environment. The primary means by which landfills can serve as a source of soil pollution is through the generation and escape of landfill leachate.

Landfill leachate consists of water in which are dissolved chemicals that, as a result either of the concentrations involved or of their inherent toxicity, pose a risk to human health or the environment. Landfill leachate is typically high in salts and often contains synthetic organic compounds and heavy metals. Leachate is produced by the infiltration of water through the waste material. As the water moves through the waste, components of the waste dissolve into the water. Current regulations require landfills to be constructed so that leachate production and escape are minimized. However, landfills constructed prior to the regulations were usually not so constructed, and, therefore, many older landfills have problems associated with leachates. Compounding this is the fact that many landfills were located in lowlands near streams overlying porous sediments.

2.1.2 Municipal Sewage Disposal In most larger municipalities, sanitary and storm sewage is processed through wastewater treatment plants. The liquid (treated waste-water effluent) and solid (sewage sludge) wastes produced during treatment are often disposed of by land spreading. This involves placing the wastes on or incorporating the wastes into the soil. Sewage waste can contain synthetic organic compounds, heavy metals, and microbial contaminants (pathogens). Once the waste has been disposed to the soil, rain or irrigation water can dissolve contaminants as it moves through the waste, thereby moving the contaminants deeper into the soil. Some waste components can enter the food chain by plant and animal uptake. Pollutant transport by surface-water runoff and by wind is also a concern.

2.1.3 Residential Septic Systems In smaller municipalities and in most rural areas, sewage waste is disposed through the use of septic systems. A septic system consists of a waste storage tank and distribution piping buried in the soil near a home. Sewage eventually makes its way from the home to the soil. If these systems are improperly designed, they introduce contaminants such as microbial pathogens into the soil.

2.2 Leaking Storage Tanks

Liquids are routinely stored in tanks located above or below the ground. These tanks are used to hold, for example, gasoline, aviation fuel, chemical stocks for manufacturing, and cleaning solvents. An estimated 3 to 5 million underground storage tanks exist in the United States, and it is thought that approximately 10% of these may be leaking (EPA, 1987). Liquids leaking from ruptured tanks or from spills near the tank enter the soil, thereby serving as a source of pollution. Fuel components (e.g., benzene, toluene, xylene) and chlorinated solvents (e.g., tetrachloroethene, trichloroethene) are major subsurface contaminants because of their ubiquitous use. Leaks and spills associated with storage tanks are a major mode of entry into the environment for these compounds. Recently, requirements for emplacing and monitoring storage tanks have been enacted to help reduce such problems.

2.3 Industrial Manufacturing Plants

Waste materials are a byproduct of manufacturing processes, and disposal of these wastes serves as a source of soil pollution. The type of pollution will depend on the item being produced and the method of manufacture. Recently there has been a concerted effort to develop manufacturing processes that minimize waste production and dependence on hazardous materials.

2.4 Commercial Activities

Most commercial activities produce waste materials that find their way into soil. For example, chlorinated organic compounds (e.g., tetrachloroethene, trichloroethene) are used for dry cleaning clothes and for degreasing automobile and airplane engine parts. On-site disposal of cleaning wastes can be a source of soil pollution. Heavy-metal pollution is also associated with commercial activities, such as electroplating and automobile repair.

2.5 Energy and Material Production

Energy-production activities such as electricity generation can be a source of soil pollution. For example, when coal is used for electricity generation, devices are often placed in the plant smoke stack to remove sulfur, metals, and particulates so that they are not released into the atmosphere. The waste materials so collected are often disposed of on site by application to soil. Precipitation infiltrating into the waste may dissolve and transport contaminants deeper into the soil. Mining activities can also serve as a source of pollution, such as surface-water leaching of mining spoils.

2.6 Agricultural Activities

2.6.1 Agrochemicals Fertilizers and pesticides (agrochemicals) are routinely used in current crop production activities in the United States. The use of these agrochemicals has helped to increase crop production capacity greatly. However, the application of these agrochemicals to soil can serve as a source of pollution, especially if their use is not properly managed.

2.6.2 Feedlots Current animal production activities (beef, swine, poultry) make use of feedlots. These feedlots can produce large quantities of animal wastes (manure). The flow of precipitation runoff over the feedlots can wash nutrients such as nitrogen and phosphorus off the site. If the nutrient-laden water enters a surface-water body, it can contribute to the process of eutrophication (see Glossary).

3. MEASURING AND MONITORING POLLUTION

The soil environment is a large repository of anthropogenic and natural pollutants. The soil environment by its nature is able, to a certain extent, to sorb, neutralize, degrade, and retain pollutants and thus protect humankind and other living organisms from deleterious exposure to many kinds of toxic chemicals and pathogens. However, excessive disposal of pollutants into the soil environment usually leads to an imbalance in the processes responsible for this protective effect. The soil then becomes contaminated (overloaded) and is no longer able to cope with the pollutants, therefore releasing them into other environments.

It is worth stressing that soil pollution is a relative term that should be qualified with some examples. For example, a soil that has high concentrations of nitrates, as a result of excessive applications of wastes or fertilizers, may not be damaged nor considered polluted. But this condition usually leads to excessive uptake of N by plants, with negative physiological effects. Also, high levels of nitrates in soils can result in nitrate contamination of surface and groundwater resources. Another example is a soil in which a toxic organic or inorganic chemical has been spilled (such as waste oils, pesticides, solvents, metal plating wastes). Such a spill is likely to influence many chemical and biological processes and therefore induce changes in the soil. This soil is likely to release some of these pollutants into other environments and will need extensive remediation and reclamation. This soil is considered to be severely contaminated.

Soil pollution monitoring has two major objectives. The first is to determine the "health" of the soil itself. The other is to determine the location, amount, and character of any pollutants that may be present in the soil.

Soil pollution can be indirectly estimated by measuring soil properties that are affected by pollutants. These properties include water infiltration, organic matter, biological-microbial activity, plant growth, and biological-species diversity. A good example of this approach to soil organic pollutant monitoring involves changes in the biological activity of some microorganisms when exposed to polluted soils or soil-water extracts. In this case, photoluminescent microorganisms exposed to active soil pollutants die or cease to function, resulting in a decrease in light emission. Since these organisms are sensitive to many common pollutants, their responses do not provide identification of the specific pollutant. New developments in this field have enhanced our ability to identify and quantify soil microbial pathogens (bacteria and viruses). These new soil-monitoring techniques make use of soil microbial population DNA extractions combined with DNA amplification by polymerase chain reaction (PCR).

Other important aspects of soil pollution monitoring involve methods of sample collection and sample preparation, which can be difficult to implement. Typical problems encountered during the collection of soil samples include sample contamination and cross-contamination and pollutant losses such as volatilization in the case of volatile organic chemicals. Soil sample preparation often biases results by contamination, and/or by the exclusion of soil components that may or may not be contaminated. Thus, with all of these aspects considered, the data collected should be sufficient in quality and quantity to be able to quantify with a known degree of precision and accuracy the pollutant(s) status in the soil environment.

Soil pollution is usually measured in the context of a three-dimensional space called the soil environment. Some of the important characteristics of this environment were discussed in Sec. 1 and can be drastically affected by pollutants. Soil pollution measurements require that the extent of the contaminated soil volume or mass be delineated. An example of this approach is finding the extent of pollutant contamination of a soil following a chemical spill. To do this, data on the distribution of pollutant concentrations must be collected. The data can then be plotted in three dimensions using statistically derived interpolation functions that help define a contour of the pollutant in the soil environment. Although this approach may not define the exact extent of the contaminated soil, it is the best tool available to soil scientists to measure distributions of pollutants, short of excavating and analyzing all of the soil.

Temporal measurements of soil pollution are often referred to as soil pollution monitoring. These involve repetitive measurements of soil pollution status over time. The accumulation of pollutants in the soil environment must be monitored over time to ensure that they do not exceed the maximum allowable limits. This practice is becoming more common as more wastes are being disposed of in the soil environment. For instance, federal regulations require that agricultural fields to which municipal solid wastes are applied must be monitored for pollutants such as metals and toxic organic chemicals, as well as for microbial pathogens. In this case a soil-sampling schedule is defined and implemented to help monitor the changes in key pollutants in the soil environment.

There are many methods by which to collect samples that can be analyzed for target pollutants. Solid-phase soil samples are collected using devices as simple as a hand shovel and as sophisticated as a truck-mounted hollow-stem auger. Environmental soil sampling requires that soil samples be collected free of cross-contamination and with minimal physical disturbance. This is usually accomplished by collecting soil core samples. For this purpose a variety of hand hollow-stem and mechanical augers have been devised. The sealed soil cores can be stored and transported to the laboratory for analysis, thereby minimizing field contamination and physical disturbance. Plastic sleeves are used to collect samples for routine soil inorganic parameters, whereas metal sleeves are used when organic contaminants are expected to be present in the soil.

Nondestructive or *in situ* soil monitoring is performed primarily for two soil parameters, these being soil moisture and salinity. Infiltrometers are field devices that are placed on the surface of the soil and allow for the measurement of water infiltration rates and hydraulic conductivity in soils. These two parameters are linked to changes in physical and chemical soil properties, sometimes induced by the presence of pollutants such as soluble salts. There are two devices commonly used for this purpose: the double-ring infiltrometer and the tension-disc infiltrometer. Both devices allow for careful metering of water as it infiltrates into the soil profile.

Neutron-activation probes, time-domain reflectometry (TDR), and electrical-conductivity (EC) soil probes are also used for nondestructive measurements of soil-water content and can be used to measure salinity in soils. However, these devices require some disturbance of the soil environment as an access tube or metal probes must be inserted into the profile. The limitations of these devices include the high cost, the use of a low-energy radioactive source (for neutron probe) or of sophisticated electromagnetic radiation sources and measuring devices (for TDR and neutron probe), and frequent calibrations.

Soil-pore water sampling devices are used to collect water-soluble pollutants from the soil environment. These devices do not directly measure pollutant parameters but allow the *in situ* collection of samples of the soil-pore water, which may then be analyzed for compounds of concern. These devices also require some soil disturbance as they must be inserted into the soil profile.

There are two major types of soil-pore water collection devices: passive or gravity fed and active or suction devices. Passive devices such as glass brick lysimeters are inserted into the soil, and gravitational water is allowed to percolate into them. This water is subsequently withdrawn using an access tube and preserved for analysis. Numerous types of porous cups are used as suction devices to collect soil-pore water. These include porous cups made out of ceramic, Teflon, or stainless steel. These cylindrical devices are inserted into the soil, and vacuum is applied to create suction across the porous membrane in contact with the soil solution. This process draws soil-pore liquid into the cup; the water is subsequently withdrawn via an access tube and subject to chemical analysis.

Methods are also available for collecting samples of the soil gas. This is of special interest for monitoring levels of important gases such as oxygen and carbon dioxide and for determining gas-phase concentrations of volatile organic pollutants (e.g., gasoline components, cleaning solvents). Active gas-sampling devices are conceptually similar to pore-water samples in that volumes of the fluid (in this case gas) are somehow extracted from the soil and then removed to the subsurface for subsequent analysis for the target pollutants. Passive gas-sampling devices are based on emplacing a device in the soil that contains a material to which the gas-phase compound of interest can sorb. After a given time, the device is recovered and the pollutants thermally desorbed and subjected to appropriate analysis.

We have seen that the most common approach to measuring soil pollution is done by measuring the concentrations of pollutants in soil samples or soil-water extracts. This in many instances requires destructive sampling or disruptive sample collection. This activity can potentially expose site workers and sampling personnel to the pollutants. Another concern is that many soil sampling and monitoring programs primarily involve collection and analysis of single-phase samples (e.g., soil-pore water or soil cores). However, many pollutants can reside in several phases associated with the soil environment. In these cases, the determination of total pollutant masses is constrained by several sources of uncertainty when based on single-phase samples. For example, if only soil-pore water samples are available, concentrations and masses of pollutants associated with the solid and gas phases of the soil will have to be estimated by use of distribution coefficients. As is discussed in Sec. 4, the use of these coefficients is subject to and limited by several factors. A complete soil sampling and monitoring program requires collection and analysis of all pertinent phases of the soil.

A further complication to soil environmental sampling is the very limited knowledge about the heterogeneity of the soil environment. Since each soil sample collected is part of a larger population of samples present in the soil, a valid and defensible estimate of soil pollution levels requires

1. *a priori* knowledge of the soil heterogeneity and pollutant variance in the soil system;
2. careful use and implementation of soil sampling equipment, sampling protocols, and analytical techniques; and
3. a rigorous statistical approach to determine the appropriate soil-sampling scheme and to analyze the data.

4. TRANSPORT AND FATE OF POLLUTANTS

An understanding of how contaminants move in the subsurface is required to address environmental problems. For example, such knowledge is needed to evaluate the probability of contaminants associated with a chemical spill reaching an aquifer and contaminating groundwater. Such knowledge is also required to develop and evaluate methods for cleaning up contaminated soils and aquifers. Just as importantly, knowledge of contaminant transport and fate is necessary to design "pollution-prevention" strategies.

Four major processes that control the movement of contaminants in porous media are advection, dispersion, interphase mass transfer, and transformation. Advection, also referred to as convection, is the transport of dissolved matter (solute) by the movement of a fluid responding to a gradient of fluid potential. Water is the primary fluid of concern for the systems in which we are interested. Advection, then, is the resulting movement of a solute or pollutant as it is carried along by the flowing water. It is a primary factor for pollutant transport. Dispersion represents spreading of solute about a mean position, such as the center of mass. Phase transfers, such as sorption, liquid–liquid partitioning, and volatilization, involve the transfer of matter in response to gradients of chemical potential. Transformation reactions include any process by which the physicochemical nature of a contaminant is altered. Examples include biotransformation, radioactive decay, and hydrolysis.

4.1 Quantifying Transport and Fate

The equation that is most often used to describe transport of dissolved contaminants in porous media is (Fetter, 1993; Brusseau, 1994)

$$R\frac{\partial C}{\partial t} = -v\frac{\partial C}{\partial x} + D^*\frac{\partial^2 C}{\partial x^2} - \mu C, \qquad (1)$$

where x is a spatial coordinate representing one or more directions, t is time, C is the concentration of contaminant in the fluid (e.g., water), v is the average linear velocity of the fluid in the pores of the medium, D^* is the dispersion coefficient with the asterisk signifying the existence of multiple terms for multiple-dimension systems, μ is a first-order transformation coefficient, and R is the retardation factor, defined as

$$R = 1 + (\rho/\theta)K_d, \qquad (2)$$

where ρ is the bulk density of the porous medium, θ is the fractional volumetric fluid content (equal to porosity when pores are completely saturated with the fluid), and K_d is a coefficient representing distribution of the contaminant between the solid and liquid phases (e.g., sorption).

Equation (1), known as the advection–dispersion equation, is by far the most widely used equation for describing the transport of dissolved matter in porous media. It is used in chemical engineering, petroleum engineering, and chromatography, in addition to earth sciences. The term on the left-hand side of Eq. (1) represents the change in contaminant mass occurring at a specified location in response to transport and fate processes. The retardation factor represents the influence of phase transfer (sorption) on transport. The first term on the right-hand side represents advective transport, while the second term represents dispersive transport. The third term represents a loss of contaminant from the solution due to transformation reactions. To develop this equation it must be assumed that the porous medium is homogeneous (e.g., v, D^*, and R are spatially invariant) and that the transfer of contaminant between the solid and liquid phases is instantaneous.

The movement of a contaminant from a point where it is first introduced to a point at which it is being monitored can be described by a breakthrough curve. The measurement and analysis of breakthrough curves is a widely used means of investigating contaminant transport in porous media. An example of a breakthrough curve, generated by solving (1) with $\mu = 0$, is shown in Fig. 3. Note the sharp, symmetrical curve obtained for the transport of a contaminant under ideal conditions (e.g., homogeneous porous medium, instantaneous sorption). However, transport often deviates from this ideal case. A breakthrough curve exhibiting nonideal transport is also shown in Fig. 3. This curve is clearly asymmetrical, exhibiting early breakthrough and tailing. Nonideal transport is often observed in real conditions.

4.2 Dispersion

The dispersion coefficient is usually defined as (Fetter, 1993; Brusseau, 1994)

$$D = \alpha v + D_0/\tau, \tag{3}$$

where α is the dispersivity, D_0 is fluid-phase diffusion coefficient, and τ is a factor accounting for the tortuosity of the porous medium. The first and second terms on the right-hand side of (3) represent the contribution of mechanical mixing and axial diffusion, respectively, to total dispersion. Mechanical mixing results primarily from nonuniform flow, which is induced by nonuniform pore-size distributions. Axial diffu-

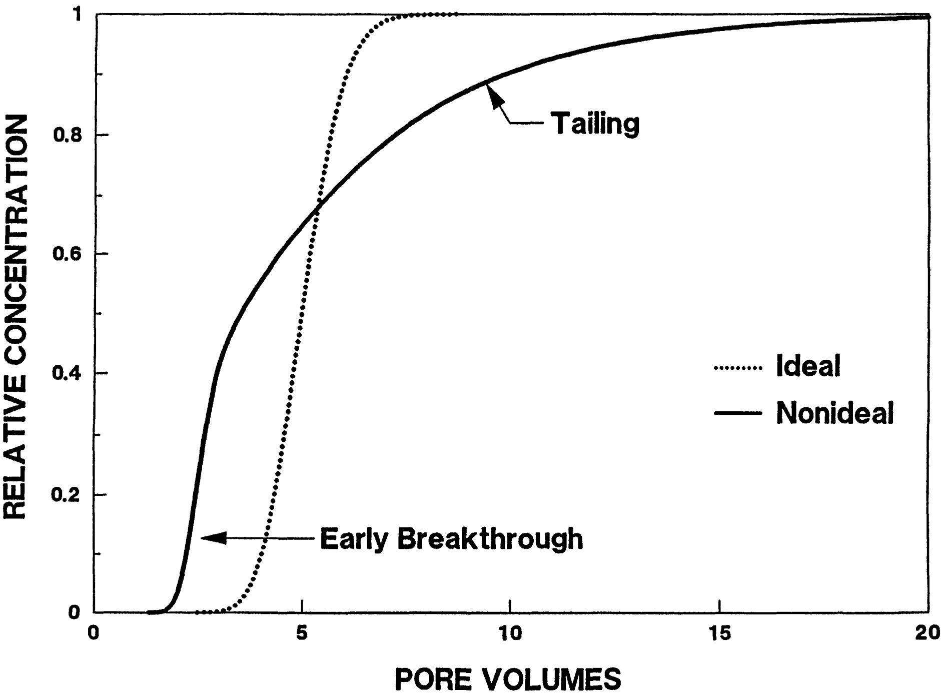

FIG. 3. Ideal and nonideal breakthrough curves. Relative concentration is defined as measured concentration divided by input concentration; pore volume is equivalent to the volume of pumped solution divided by the volume of solution the column can retain.

sion results from natural thermal activity of molecules and usually is described in terms of Fick's law. Generally, axial diffusion is an important component of dispersion only when pore-water velocities are low (e.g., <0.1 cm/h). Such velocities are typical of flow in the vadose zone (the portion of the subsurface between the soil surface and surficial aquifers), in clay layers, and in some consolidated aquifers.

Under ideal conditions, the dispersivity is a function only of the porous medium. A porous medium that has a wide range of particle sizes will also tend to have a wide range of pore-space sizes. Since the flow rate of a fluid is proportional to the size of the pore, a wide range of pore sizes will increase the degree of spreading that occurs during transport. Conversely, a porous medium that is very homogeneous—for example, one that consists of uniform particles—will induce very little spreading.

4.3 Sorption

Sorption or retention is a major process influencing the transport of pollutants in soils. The broadest definition of sorption or retention is the association of pollutant molecules with the solid phase of the soil (soil particles). Properties of the pollutant (e.g., is it nonpolar or polar, organic or inorganic) are key determinants of the degree to which the pollutant will be sorbed or retained by the soil. The specific mechanism by which a particular pollutant is sorbed will depend, in part, on the nature of the pollutant. The physical/chemical properties of the soil are also important.

Sorption is quantified by measuring a sorption isotherm, which is simply a description of the relationship between the concentration of pollutant sorbed by the soil and the concentration in water. Many different forms of isotherms have been proposed and used to describe sorption. The simplest is the linear isotherm (the same idea as Henry's law), which is given by

$$S = KC_w, \tag{4}$$

where S is the concentration of pollutant sorbed by the soil (M/L^3) and K is the sorption coefficient (L^3/M). The larger the value of K, the greater the degree to which a pollutant is sorbed by the soil. A linear isotherm signifies that for all concentrations of pollutant in water, there will always be proportionally the same sorbed concentration. The sorption of many nonpolar organic pollutants is linear. Some organic and many inorganic pollutants exhibit nonlinear sorption. An example of a widely used nonlinear isotherm is the Freundlich isotherm, given by

$$S = K_f C_w^n, \tag{5}$$

where K_f is the Freundlich sorption coefficient and n is, in essence, the power function.

The use of sorption isotherms is based on the existence of equilibrium between the solid and water phases. Research has shown that the rate of sorption of many nonpolar organic pollutants is very slow, taking anywhere from several hours to several months or even years to reach equilibrium. This slow rate of sorption is often due to the slow diffusion of pollutant molecules into spaces in the soil that have very small openings or pores. In addition, pollutants that have been in contact with soil for long times are often difficult to remove from the soil. These factors make it difficult to clean up contaminated soil.

Sorption is such an important process because it reduces the velocity of a pollutant with respect to the movement of water. As long as a pollutant molecule is sorbed by soil, it is prevented from moving with the water. Thus, it will take longer for the pollutant to travel a given distance than it would for water. The effect of sorption on pollutant transport is accounted for by the retardation factor. When the pollutant is not sorbed by the soil (i.e., $K = 0$), the retardation factor R is equal to 1. This means that the pollutant will move at the same velocity as water. When $R = 2$, the pollutant moves at an effective velocity that is half that of water. In other words, the pollutant moves only half as fast as water. When $R = 10$, the pollutant moves at one-tenth the velocity of water. Pollutants that have small retardation factors (<10) are considered to be relatively mobile. They can move rapidly from a spill site and thus quickly contaminate a large area. Conversely, pollutants that have very large retardation factors (>1000) move very slowly with respect to water. Thus, they probably will not create a contaminated zone as extensive as that created by mobile pollutants.

4.4 Transformation Reactions

The transformation of pollutants by microorganisms is referred to as biodegradation. Many organic compounds of interest can be biodegraded by soil microorganisms. Many factors influence the degree to which a compound will be biodegraded in the subsurface. These factors include

1. existence of microbial populations capable of degrading the compound,
2. the availability of the compound to the microbial population (bioavailability), and
3. environmental factors (e.g., *p*H, temperature, nutrient or electron-acceptor availability, etc.).

The bioavailability of a compound is influenced by processes such as mass transfer, rate-limited sorption/desorption, and transport in heterogeneous porous media. Hence, the kinetics of biodegradation can in many cases be controlled by processes other than the physiological activity of the microorganisms.

Some pollutants can be transformed by abiotic (physical/chemical) processes as well. Two major abiotic transformation processes are hydrolysis and radioactive decay. Water is a ubiquitous component of soil systems. As a result, most pollutants will come into contact with water to some extent. It is important, therefore, to understand if and when a pollutant will react with water when they are in contact. Reaction of a pollutant with water is termed hydrolysis. A generalized example of this reaction is given by

$$\mathrm{R{-}}X + \mathrm{H_2O} \rightarrow \mathrm{R{-}OH} + X^- + \mathrm{H}^+,$$

where R–X is an organic compound with X representing a functional group such as a halide (e.g., Cl). By reacting with water, the original compound (R–X) has been transformed to another compound (R–OH).

The two key factors in hydrolysis reactions are the charge properties of the pollutant molecules and the *p*H. Hydrolysis is essentially an interaction between nucleophiles (substances with excess electrons, such as OH^-) and electrophiles (substances deficient in electrons, such as H^+). Thus, the charge properties of the molecule will govern its reactivity with water. For many compounds, hydrolysis may be catalyzed or enhanced under acidic or basic conditions. This means that the occurrence and rate of hydrolysis is often *p*H dependent. For example, a hydrolysis reaction catalyzed by OH^- would occur more rapidly at higher *p*H values because of larger OH^- concentrations. Hydrolysis can also be influenced by sorption interactions. For example, the *p*H at the surface of many soils is lower than the *p*H of the water surrounding the soil particles. Thus, an acid-catalyzed hydrolysis reaction could be enhanced when the pollutant is associated with the soil.

Radioactive decay is an important transformation reaction for a special class of pollutants: radioactive elements. Radioactive decay is caused by an instability of the nucleus in the atom, whereby either protons and neutrons or electrons are emitted in the form of radiation. This transformation is spontaneous, but the rate at which it occurs varies widely depending on the element concerned.

The rate of radioactive decay and of hydrolysis (for a fixed *p*H) can be described with a first-order equation:

$$\frac{\partial C}{\partial t} = -kC, \tag{6}$$

where k is the rate constant $[1/T]$. A first-order equation can also be used to represent biodegradation under certain conditions (e.g., when the growth of the microbial population is minimal). For transformation reactions it is useful to define a half-life, which is the time required for half of the original pollutant mass to be transformed. For reactions that follow first-order kinetics, the half-life ($T_{1/2}$) is defined as

$$T_{1/2} = 0.693/k_h. \tag{7}$$

This equation can be used to estimate the time required for a pollutant to be transformed, which is related to its persistence in the environment. For example, less than 1% of the original pollutant mass remains after a time period equal to seven half-lives. After ten half-lives, less than 0.1% remains. For a pollutant that has a half-life of 1 day, this would mean that only 0.1% would remain untransformed after 10 days.

4.5 Multiphase Pollution

Some of the most important soil pollutants are immiscible organic liquids. Examples include gasoline, diesel fuel, and aviation fuel associated with various types of fuel stations, chlorinated solvents associated with cleaning facilities, and creosote associated with wood-treatment operations. Separate phases of immiscible liquid may exist in portions of the subsurface at hazardous waste sites contaminated by these liquids. It is well known that when an immiscible liquid moves through a porous medium, a fraction of the liquid can become entrapped, thereby forming a residual saturation. Very large pore-water velocities (i.e., hydraulic gradients) are usually required to displace residual quantities of immiscible liquids. Hence, the primary means of removal will be dissolution into water and volatilization into the soil atmosphere. The residual liquid, therefore, serves as a long-term source of contaminant.

Many organic contaminants of interest are volatile (e.g., gasoline components, chlorinated solvents). For systems that are fully saturated with water (e.g., aquifers), this is of no special concern. However, it is of great significance for the vadose zone, which, because it is usually not completely saturated with water, contains a gas phase. Consequently, a description of the transport of organic contaminants in the vadose zone will be incomplete without consideration of transport in the gas phase. Research in this area has increased greatly with the advent of "soil venting" or "vacuum extraction" as a popular means of remediating vadose-zone contamination.

4.6 Soil Heterogeneity

Soils and geologic materials are inherently nonuniform. This is a direct result of the formation of these materials and their exposure to various events during past geologic times. Thus, properties that affect solute transport can be expected to vary spatially. One of the main reasons for deviations from ideal transport behavior in the field are heterogeneities in the physical, chemical, and biological properties of subsurface material. Any analysis of the transport and fate of pollutants in soil is incomplete without consideration of the nature, magnitude, and impact of heterogeneity.

5. CLEANING UP SOIL AND WATER POLLUTION

The existence of pollutants in soil can pose a risk to human health and to the "health" of the environment. As such, the cleanup of polluted soil and groundwater has become a major component of environmental activity. In the following sections the legislative and technical framework of subsurface cleanup are discussed.

5.1 Legislative Programs for Pollution Cleanup

The first major legislation that dealt with cleanup of contaminated environments was the Water Pollution Control Act (1972). A provision of this act provided authority to fund the cleanup of hazardous substances released into navigable waters. However, there are no provisions for cleanup of contaminated land. The first legislation to directly address cleanup of contaminated land was the Resource Conservation and Recovery Act (1976). This act provided authority to order an operator to cleanup hazardous waste emitted at the site of operation. This act did not cover the many contaminated sites that have been abandoned.

The Superfund program, also known by its official title, CERCLA (Comprehensive Environmental Response, Compensation, and Liability Act—1980) was designed explicitly to address the cleanup of hazardous waste sites. It, along with later amendments (SARA, Superfund Amendments and Reauthorization Act—1986), is the major legislation governing activities associated with the cleanup of environmental pollution. Some states have smaller versions of the federal Superfund program. In addition, the Department of Defense has a separate cleanup program (Installation Restoration Program) for military sites.

Because of its importance, we briefly discuss the major components of Superfund. Its purpose is twofold:

1. to respond to releases of hazardous substances on land and in navigable waters, and
2. to clean up contaminated sites.

The first purpose deals with future releases, whereas the second one deals with sites of existing contamination. There are two types of responses available within Superfund:

1. removal actions: response to immediate threats—e.g., remove leaking drums; and
2. remedial actions: cleanup of hazardous sites.

The Superfund provisions can be used when a hazardous substance is released or when there is a substantial threat of release. They may also be used when there is a release or substantial threat of release of a contaminant that poses imminent and substantial endangerment to public health/welfare.

The first step is to place the potential site in the Superfund Site Inventory. After a preliminary assessment and site inspection, the decision is made as to whether or not the site will be placed on the National Priorities List (NPL). Sites placed on this list are those sites deemed to require a remedial action. There are currently more than 1200 sites on the NPL. The remedial investigation/feasibility study (RI/FS) is the next component. The purpose of the RI/FS is to characterize the nature and extent of risk posed by contamination and to evaluate potential remedial options. The investigation and feasibility study components are normally conducted concurrently and with a "phased" approach. This allows feedback between the two components. The selection of the specific remedial action to be used at a particular site is a very complex process. The goal of the remedial action is that it be protective of human health and the environment, that it maintain protection over time, and that it minimize untreated waste. Specifically, CERCLA mandates that the remedial action must

1. protect human health and the environment,
2. comply with applicable or relevant requirements (ARARs),
3. be cost-effective,
4. use permanent solutions or resource recovery technologies to the maximum extent practicable, and
5. satisfy preference for treatment as a principal element.

After a remedial action has been selected, it is designed and put into action.

An important component of Superfund and all other cleanup programs is the cleanup target. Related to this is the question: How clean is clean? As can be imagined, the stricter the cleanup provision (e.g., lower contaminant concentration), the greater are the attendant cleanup costs. It may require tens to hundreds of millions of dollars to clean large, complex hazardous waste sites completely. In fact, it may be impossible to clean many sites completely. It is very important, therefore, that the physical feasibility of cleanup and the degree of potential risk posed by the contamination be weighed against the economic impact *and* the future use of the site. The consideration of the future use of the site will help focus our scarce resources on those sites that pose the greatest current and future risk.

5.2 Cleanup Technologies

There are three major categories or types of remedial actions:

1. containment—contamination is restricted to a specified domain to prevent further spreading;
2. removal—contaminant is removed from the open environment and transferred to a place of control; and
3. treatment—the contaminant undergoes transformation to a nonhazardous substance.

Since the inherent hazardousness of a contaminant is eliminated only by treatment, the last is the preferred approach of the three. Containment and removal techniques are very important, however, in cases where it is not feasible to treat the contaminant. At many hazardous waste sites, the remedial action in use consists of a combination of containment, removal, and treatment. We focus on each of the three types of remedial actions in turn. Although numerous techniques have been proposed for soil cleanup, only a very few have been tested to any significant extent. Our discussion focuses on those that have been used in the field.

5.2.1 Containment Containment can be accomplished by controlling the flow of the fluid that contains the contaminant, such as controlling the flow of groundwater. This can be done with the use of wells, drains, or physical barriers. The principle of a physical barrier is to install it in front (downgradient) of the contaminated zone, which prevents further spread of the contaminant. An important component of employing physical barriers is the presence of a zone of low permeability into which the physical barrier can be "keyed." Without such a key, the contaminated water could flow underneath the barrier. Usually, physical barriers can be emplaced to depths of about 50 m.

There are three major types of physical barriers: slurry trenches, grout curtains, and sheet piling. Slurry trenches, the most widely used type of barrier, are constructed by filling a trench with clay or a mixture of clay and soil. Grout curtains are formed by injecting into the ground chemicals that are somewhat like cement. In some cases, large sheets of iron (sheet piling) are driven into the ground to form a barrier.

The principle of containment by wells and drains is to manipulate and control water flow by extraction/injection of water. Since contaminant movement is governed largely by water flow, controlling water flow results in contaminant movement being controlled. The key performance factor for this approach is the capture of the contaminant plume. An advantage associated with the use of wells to control contaminant movement is that this is the only containment technique that can be used for deep systems (i.e., >50 m). For this and other reasons, wells are the most widely used method for containment. Some disadvantages associated with wells are the cost of long-term operation and maintenance of the pumps, and the need to store, treat, and dispose of large quantities of contaminated water that is pumped to the surface.

Containment can also be accomplished by immobilizing the contaminant in place. This can be accomplished by cementation, vitrification, freezing, or fixation. Cementation involves solidifying the porous media by injecting cement. A key performance criterion is the long-term stability of the "concrete" block, including the leachability of the contaminant. Vitrification involves "melting" the porous medium to form a "glass" block. This is done by use of an electric current to generate heat (2000–3000 °C). As for cementation, a key performance criterion is the long-term stability of the block, including the leachability of the contaminant. The concept of fixation is to immobilize the contaminant by promoting binding to the solid phase (covalent bond). Key performance factors are delivery of catalyst to target areas and the long-term stability and leachability of the fixated contaminant. Freezing is normally limited to use as a temporary containment technique.

5.2.2 Soil Flushing Soil flushing or "pump and treat" is currently one of the most widely used remedial action techniques. The basic principle of soil flushing is to remove contaminated water from the subsurface by pumping it with one or more wells. Furthermore, clean water brought into the contaminated region by the pumping action will remove additional contamination by inducing desorption from the solid phase (soil particles). The contaminated water pumped from the subsurface is directed to some type of treatment operation. This treatment may consist of air stripping, carbon adsorption, or perhaps an above-ground biological treatment system. There is one major performance criterion when using wells for containment, that being capture of the contaminant. When using wells for contaminant removal (i.e., pump and treat), the effectiveness of contaminant removal is a second major performance criterion in addition to contaminant capture. Recent studies of operating pump-and-treat systems have shown that contaminant removal by pumping is often not very effective. There are many possible factors that can limit the effectiveness of contaminant removal. Some major ones are presence of low-permeability zones, rate-limited desorption, and presence of immiscible liquids.

Because the use of pumping to remove contaminants is a major remedial action technique, methods are being tested to enhance the removal effectiveness. One such approach is to inject into the aquifer a chemical, such as a surfactant, that will promote dissolution and desorption of the contaminant, thus enhancing the removal effectiveness. A key factor controlling the success of this approach is the ability to deliver the chemical to the places that contain the contaminant.

5.2.3 Soil Vapor Extraction The basic principle of soil vapor extraction, or soil venting, is very similar to that of pump and treat: A fluid is pumped through a contaminated domain to enhance removal. In the case of soil venting, however, the fluid is air rather than water. There are two key conditions for using soil venting. First, there must be a gas phase present in the soil, through which the contaminated air can travel. This condition limits the use of soil venting to the vadose zone. Second, soil venting is based on transfer of contaminant from other phases (solid, water, immiscible liquid) to the gas phase. This limits soil venting to volatile contaminants.

Because air is much less viscous than water, much less energy is required to pump air. Thus, it is usually cheaper to use soil venting for removing volatile contaminants from the vadose zone. Once the contaminant is removed from the vadose zone, it is either released to the atmosphere or placed into a treatment system. The major performance factors for soil venting are the same as for water pumping, the effectiveness of capturing and removing the contaminant. The effectiveness of contaminant removal by soil venting can be limited by the same factors that limit removal by water flushing.

5.2.4 *In Situ* Remediation Removing pollutants from soil is often difficult. Hence, techniques have been developed that are meant to treat the pollutant in place. The primary *in situ* treatment technology used to date is *in situ* bioremediation. This usually involves stimulating naturally occurring microorganisms to degrade the pollutant. This stimulation can be effected, for example, by the addition of nutrients or oxygen. The factors previously discussed that control biodegradation (Sec. 4.4) also control the degree to which *in situ* bioremediation will be successful. An additional key factor controlling the efficacy of this technique is the delivery of the nutrients or oxygen to the contaminated domain.

In situ chemical remediation is similar to *in situ* bioremediation in that the contaminant is degraded by promoting a transformation reaction. For chemical remediation techniques, however, this is done with a chemical reaction such as hydrolysis or oxidation/reduction. This approach has been used much less frequently than *in situ* bioremediation.

6. SOIL RECLAMATION

The nature of soils is being constantly challenged by human's activities on the surface of the planet. The reclamation of disturbed and/or contaminated soils is very important to secure a healthy natural soil environment for the future. Soil reclamation implies that the soil must be returned to a state such that natural soil-forming processes (described in Sec. 1) can continue indefinitely. Thus, after soil reclamation activities have been terminated, natural soil physical, chemical, and biological processes and conditions will approximate their original natural state. It is now recognized that soil restoration, that is, returning to exact natural conditions, is often not practically achievable.

Soil systems that often must be reclaimed include

1. agricultural lands that have been excessively cropped, abandoned and mismanaged, resulting in severe physical (erosion), chemical (salt accumulation), and biological (loss of biological diversity) deterioration;
2. lands that have been subjected to extraction of underground resources (mining);
3. lands that have been used to dispose of or stockpile excessive amounts of wastes;
4. lands that have been damaged by natural catastrophes such as flood, fire, and wind erosion; and
5. lands that have been damaged by excessive use or changes in use such as overgrazing, deforestation, and varying urban growth patterns.

Traditional soil reclamation techniques include the physical manipulation of the plow layer (~30 cm) with the addition of a soil amendment. In general, soil amendments re-

Table 2. Bulk density and porosity of soils. Extreme low bulk densities are usually associated with coarse (sandy) soils, whereas extreme high bulk densities are associated with fine textured (clay) soils.

Bulk Density (g/cm^3)	Porosity (%)	Ranking
1–1.3	60–50	Low bulk density, high porosity
1.4–1.7	50–35	Medium bulk density, medium porosity
>1.8	>33	High bulk density, very low porosity

fer to natural materials or chemicals that help restore the natural chemical and physical properties of soils. Early soil reclamation efforts were focused on agricultural lands that had been mismanaged, usually by a combination of poor water quality and poor irrigation practices. These two factors often result in accumulations of salts in the soil profile that severely restrict crop production. Thus, there are several criteria that are used to determine the suitability of soils to sustain plant growth. The two most important physical criteria are permeability and bulk density, and are listed in Tables 2 and 3.

Compacted soils (high bulk density) do not allow proper plant root penetration and severely limit soil–atmospheric gas exchanges. When oxygen diffusion into the soil profile is limited, the soil system can become anaerobic and undergo chemical and biological changes that are detrimental for the sustained growth of most plants. Another way of measuring soil compaction is the soil porosity (the ratio of the total soil volume to the volume of the solids). These two parameters are inversely related (see Table 2).

Soil-water permeability is very important to insure proper water flow into the soil profile. If water is slow to infiltrate into the soil, it tends to run across the land and create erosion. At the same time, soils with low infiltration rates tend to accumulate salts and may not provide sufficient water for sustained plant growth (see Table 3). The bulk density and infiltration of soils can in general be improved via physical manipulation such as deep plowing. However, such practices offer a temporary remedy at best, as they do not address the more fundamental problems responsible for these deleterious physical changes in the soil.

Changes in soil salinity have a progressive and profound impact on the overall physical and chemical status of soils. Soil salinity is a measure of the soluble salt content of a soil and is measured by electrical conductivity of a soil-water extract. Excessive soil salinity limits plant growth and affects soil microbial activity. Most food production plants are sensitive to salt, and yields are therefore limited when plants are exposed to high salinity. The criteria to evaluate soil salinity are given in Table 4.

Sodium chloride is one of the most deleterious minerals to affect the soil environment. Excessive additions of NaCl from brine spills and/or accumulations due to poor irrigation practices result in several important changes in the chemical and physical structure of clay minerals in soils. Clay minerals have a cation exchange capacity (CEC) or ability to adsorb/exchange solution cations from the soil solution (see Sec. 4). Thus, since NaCl is very soluble in water, the Na^+ ion exchanges onto the clay, causing a change in the positive charge densities of the clay particles. Clay particles that have predominantly monovalent cations such as Na^+ in the exchange sites have a diffuse charge density that makes them repulse other similarly charged clay particles. This change results in the physical dispersion of clay particles in the soil environment. Dispersed clay particles form very small soil pores that severely limit water movement through the soil profile. On the other hand, Ca^{2+} ions do not have this effect on soils. Because clay particles saturated with divalent cations have

Table 3. Soil infiltration rankings.

Permeability (cm/h)	Ranking	Texture[a]
<0.15	Very slow	Usually heavy clay soils
0.15–0.5	Slow	Clay soils
0.6–1.5	Moderately slow	Clay loam soils
1.6–5	Moderate, adequate	Loam soils
5–15	Moderately rapid	Sandy loamy soils
15–50	Rapid	Sandy soils
>50	Very rapid	Coarse sandy soils

---

[a] Other factors besides texture may also affect soil permeability.

Table 4. Soil salinity rankings.

Parameter (mS/cm)	Nonsaline	Slighty saline	Moderately saline	Saline
Electrical conductivity[a]	<4	4–8	8–16	>16

[a] Measured on a water-saturated soil paste extract.

more localized charge densities, they tend to form aggregates with other clay particles. Soil aggregates are of vital importance for a healthy soil environment as they improve porosity, facilitate plant root penetration, improve oxygen and water movement and availability, and provide excellent microenvironments for soil microbes.

The ratio of sodium ions to common divalent ions such as calcium and magnesium provides a good measure of the sodium content in clays. The criterion for evaluating sodium content in clays is called exchangeable sodium percent (ESP) and is defined in Table 5. Alkaline waters used in irrigation of soils can also produce soil conditions similar to those described above. Excessive concentrations of soluble carbonate (CO_3^{2-}) and bicarbonate (HCO_3^-) ions in irrigation water limits availability of soluble calcium ions in the soil solution. This can result in an increase of exchangeable sodium in clay particles, thereby increasing the soil SAR and creating sodic soils (see Table 5).

Saline and sodic soils can be reclaimed by a combination of physical soil manipulations and chemical soil amendments that include the following:

Table 5. Sodium exchangeable percent rankings. These ESP values are generally considered guidelines. Thus, these rankings may vary from soil to soil.

ESP = (Na/CEC) × 100[a]	Ranking	Comments
<10	Adequate	Sandy soils can tolerate ESP values up to 15 without problems
10–15	Borderline	Problems likely to occur in fine texture soils
>15	Inadequate	Infiltration problems

[a] Sodium and cation exchange capacity of clay soils are in milliequivalents per 100 g of soil.

1. Incorporation into the soil plow layer of liberal (>5 tons per acre) amounts of gypsum ($CaSO_4 \cdot 2H_2O$), followed by repeated irrigations.
2. Additions of acidifying materials such as elemental sulfur into the soil, which ultimately produces sulfuric acid, or directly adding sulfuric acid to the irrigation water. This acid reacts with carbonate ions liberating CO_2 gas and forming a more soluble form of calcium, as shown in the following summary reaction:

$$H_2SO_4\ \text{(in water)} + CaCO_3\ \text{(in soil)} \rightarrow CaSO_4 \cdot 2H_2O\text{(gypsum)} + CO_2\ \text{(gas)}.$$

After repeated irrigations the soluble sodium salts are usually leached below the root zone. These treatments restore a more natural chemical and physical balance in the soil environment.

Acid soils develop as a result of continuous fertilization with acid-forming fertilizers. The most common reduced forms of nitrogen fertilizers such as ammonia and urea form nitric acid upon oxidation in the soil environment, thereby lowering the soil *p*H. Soils in the eastern half of the United States tend to become acidic with time, as these soils do not have large reserves of acid-neutralizing minerals such as calcite commonly found in western soils. Acid soils (*p*H < 5.5) quickly lose macro- and micronutrients via leaching and develop aluminum toxicity problems with plants, through the following reaction:

$$Al(OH)_3\ \text{(soil minerals)} + 3H^+\ \text{(in soil solution)} \rightarrow Al^{3+}\ \text{(in soil solution, toxic!)} + 3H_2O.$$

Lime and calcite additions to these soils are a common reclamation practice that provide a temporary remedy for the deleterious effects that low *p*H has on both plants and the soil itself. In this case, lime (CaO) or calcium carbonate ($CaCO_3$) additions neutralize

Table 6. Major soil macronutrients.

Element	Desired soil concentration ranges	Desired plant-available concentration and forms
Carbon	0.5–3% in organic matter (O.M.)	C:N:P = 100:10:1
Nitrogen	>0.1% in O.M. optimum C:N = 10:1	>20 mg/Kg as NO_3–N
Phosphorus	>0.01% inorganic form	>10–20 mg/kg as PO_4–P extractable. Variable depending on soil and extraction method
K, Ca	>0.1%; supplies vary depending on mineral form	>100 mg/Kg as water-soluble Ca^{2+} and K^+
Sulfur	>0.1% as gypsum	>100 mg/Kg as water-soluble SO_4^{2-}

excess H^+ ions in the soil, thereby promoting the formation of very insoluble aluminum oxides minerals such as gibbsite.

Soil reclamation may also include the reestablishment of a nutrient balance in the affected soil system. Fertility levels of a soil are usually assessed by measuring the total and plant-available concentrations of macro- and micronutrients. These are listed in Tables 6 and 7. Chemical fertilizers are commonly added to agricultural soils and disturbed lands as a means to restore nutrient balance to the soil system. However, excessive amounts of some fertilizers can produce *p*H and salt problems, as described above.

Organic wastes, such as composted plant residues and municipal solid sludges, are common soil amendments used in the reclamation of soils deficient in organic matter and plant nutrients. Many studies to date indicate that repeated applications to soils in need of organic matter and nutrients have beneficial consequences. However, this practice may not be sustainable (indefinitely) as repeated applications of wastes may lead to increases in toxic metals in the soil environment. Thus, waste applications to the soils require a more rigorous monitoring program to control possible deleterious effects that these materials can have in the soil environment. Nonetheless, recycling organic wastes by using them as soil amendments is very cost-effective and beneficial to the soil environment. This practice is becoming a more acceptable form of waste disposal, as the natural cycle of the major plant and animal nutrients listed in Table 6 is completed in the soil environment.

Following successful soil reclamation activities, plants are usually seeded or planted to provide physical stability to the soil surface and help re-establish the natural biological and nutrient cycle to this environment. To this end, many types of grasses, shrubs, and trees are available (Munshower, 1993). These plant species have been found to be tolerant to the less-than-optimum conditions that recently reclaimed soils have.

Soil reclamation practices are very beneficial and necessary to restore the natural physical, chemical, and biological balance of soil systems. These practices are becoming increasingly more efficient as our understanding of soil science continues to increase. Soil reclamation activities will likely

Table 7. Major soil micronutrients.

Element	Desired soil concentration ranges	Desired plant-available concentration and forms
Boron	>10 mg/kg, variable	>0.5 mg/kg as H_3BO_3–B hot-water extractable
Copper, zinc, iron	>10 mg/kg, variable	>0.5 mg/kg as Cu^{2+} EDTA[a] extractable
Manganese	>100 mg/kg, variable	>5 mg/kg as Mn^{2+} EDTA[a] extractable
Molybdenum	>1 mg/kg, variable	>0.01 mg/kg as MoO_4^{2-}–Mo, but varies depending on presence of other micronutrients and extracting solution used

---

[a] EDTA is ethylene diamine tetraacetic acid, a metal chelating agent.

increase in the future as more and more soils are being affected by human's activities.

7. SUMMARY

We have briefly discussed the nature and importance of soils and the movement and impact of pollutants in soil and groundwater environments. We have also discussed methods to monitor, remediate, and restore polluted soils and groundwaters. It must be stressed that it is very difficult, expensive, and often impossible fully to remediate and restore polluted subsurface systems. It is clear that preventing pollution from occurring is a much more effective strategy. It is also clear that pollution prevention programs can be fully successful only when the public, industry, academia, and government work together.

GLOSSARY

Advection: The transport of solute by the movement of a fluid; also called **Convection.**

Advection–Dispersion Equation: The most widely used equation for simulating solute transport in porous media.

Aluminosilicates: Minerals composed primarily of aluminum (Al) and silicon (Si) atoms.

Aquifer: A water-bearing formation in the subsurface.

Axial Diffusion: Movement of solute in response to a concentration gradient in the direction of flow.

Breakthrough Curve: A representation of the concentration of solute in a fluid as a function of time at a selected point.

CEC: Cation exchange capacity, a surface property of clay minerals enabling them to exchange cations to and from water solutions containing dissolved ions.

Clays: Secondary minerals. Examples: montmorillonite, vermiculite, and illite.

Convection: A term used interchangeably with **Advection.**

Dispersion: The spreading of solute about a mean position.

Dispersivity: A parameter representing the spreading potential of a solute-porous medium system.

ESP: Exchangeable sodium percent, the percentage of sodium ions that cover the surface of clay particles.

Eutrophication: A process whereby the trophic status of surface-water bodies changes with time, and the type and quantity of biota change as well. This process is mediated by the availability of nutrients and can be deleteriously accelerated by runoff containing high levels of fertilizers.

Facilitated Transport: Enhanced rate of solute transport caused by an increase in apparent solubility and a concomitant decrease in sorption and retardation.

Humic Substances: Complex organic (carbon-based) materials that form from decaying plant and animal tissue in the soil environment.

Hydrodynamic Residence Time: The average time required for fluid particles to traverse the flow domain.

Hydrolysis: Chemical process that breaks a molecule in two by insertion of a water molecule, the hydrogen attaching to one fragment and the hydroxyl to the other.

Ideal Transport: Transport that can be predicted with models based on assumptions of subsurface homogeneity and instantaneous mass transfer.

Local-Scale Dispersion: Spreading caused by axial diffusion and mechanical mixing.

Mechanical Mixing: Pore-scale spreading caused by flow through a macroscopically tortuous and nonuniform porous medium (i.e., one with nonuniform pore-size distribution).

Nonideal Transport: Transport that does not coincide with that predicted from models based on assumptions of subsurface homogeneity and instantaneous mass transfer.

Oxidation and Reduction: Chemical processes that result in the loss or gain of electrons from elements and molecules.

PCR: Polymerase chain reaction, laboratory biochemical process whereby DNA molecules are replicated enabling their detection and characterization.

Plume: A spatial distribution of solute.

Pump and Treat: A method for removing contaminants from porous media based on inducing water flow through the contaminated domain.

Retardation Factor: A parameter in the advection–dispersion equation that accounts for association of solute with the solid phase.

Soil Venting: A method for removing contaminants from porous media, based on inducing gas flow through the contaminated domain.

Solute: Matter dissolved in a fluid.

Sorption: The association of solute with a solid phase.

Sorption Isotherm: A regression of sorbed-phase concentrations against solution-phase concentrations.

TDR: Time-domain reflectometry, electrophysical process that characterizes and measures the propagation of electromagnetic radiation through the soil and relates it to water content.

Vadose Zone: The portion of the subsurface located between the soil surface and surficial aquifers and characterized by conditions of variable water saturation.

Weathering: Chemical and physical soil-forming process that transforms primary minerals into secondary minerals (silt and clay).

Works Cited

Birkeland, P. W. (1974), *Pedology, Weathering, and Geomorphological Research,* New York: Oxford University Press.

Brady, N. C. (1974), *The Nature and Properties of Soils,* 8th ed., New York: MacMillan.

Brown, A. L. (1978), *Ecology of Soil Organisms,* London: Heinemann.

Brusseau, M. L. (1994), *Rev. Geophys.* **32,** 285–314.

Buol, S. W., Hole, F. D., McCracken, R. J. (1973), *Soil Genesis and Classification,* Ames, IA: The Iowa State University Press.

Environmental Protection Agency (1987), Underground Storage Tank Corrective Action Technologies, EPA/625/6-87-01115, Cincinnati, OH.

Fetter, C. W. (1993), *Contaminant Hydrogeology,* New York: Macmillan Press.

MacCarthy, P., Clapp, C. E., Malcom, R. L., Bloom, P. R. (1990), *Humic Substances and Crop Sciences: Selected Readings,* American Society of Agronomy, Madison, WI: Soil Science Society of America.

Munshower, F. K. (1993), *Practical Handbook of Disturbed Land Revegetation,* Boca Raton, FL: Lewis Publishers.

Pye, V. I., Patrick, R., Quarles, J. (1983), *Groundwater Contamination in the United States,* Philadelphia: Univ. of Pennsylvania Press.

Schaller, F. W., Sutton, P. (1978), *Reclamation of Drastically Disturbed Lands,* American Society of Agronomy, Crop Science Society of America, Madison, WI: Soil Science Society of America.

Further Reading

Boulding, J. R. (1994), *Description and Sampling of Contaminated Soils: A Field Guide,* 2nd ed., Boca Raton, FL: Lewis Publishers.

Brady, N. C. (1974), *The Nature and Properties of Soils,* 8th ed. New York: MacMillan.

Brusseau, M. L. (1994), *Rev. Geophys.* **32,** 285–314.

Dixon, J. B., Weed, S. B. (1989), *Minerals in the Soil Environment,* 2nd ed., SSSA Book Series, Madison, WI: Soil Science Society of America.

Klute, A. (1986), *Methods of Soil Analysis—Part 1, Physical and Mineralogical Properties,* 2nd ed. Agronomy No. 9, ASA, Inc., Madison, WI: SSSA, Inc. Publishers, 1188 pp.

Munshower, F. F. (1994), *Practical Handbook of Disturbed Land Revegetation,* Boca Raton, FL: Lewis Publishers.

Page, A. L., Miller, R. H., Keeney, D. R. (1982), *Methods of Soil Analysis—Part 2, Chemical and Microbiological Properties,* 2nd ed. Agronomy No. 9, ASA, Inc., Madison, WI: SSSA, Inc. Publishers, 1159 pp.

Pye, V. I., Patrick, R., Quarles, J. (1983), *Groundwater Contamination in the United States,* Philadelphia: Univ. of Pennsylvania Press.

Wild, A. (1993), *Soils and the Environment,* New York: Cambridge Univ. Press.

Wilson, L. G., Everett, L. G., Cullen, S. J. (1995). *Handbook of Vadoze Zone Characterization & Monitoring,* Boca Raton, FL: Lewis Publishers.

SOLAR CELLS

See PHOTOVOLTAIC DEVICES

SOLAR ENERGY

I. Dostrovsky, R. Tenne, and A. Yogev, *Weizmann Institute of Science, Rehovot, Israel*

Introduction 363
1. **Thermal Processes** 364
1.1 Low-Temperature Applications 365
1.1.1 Domestic Hot Water 365
1.1.2 Solar Ponds 365
1.2 Medium Temperatures 366
1.2.1 Process Heat 367
1.3 High Temperatures 368
1.3.1 Trough Systems 368
1.3.2 Dish Systems 369
1.3.2.1 Dish Farms 370
1.3.3 Central Receivers 371
1.3.4 Thermal Receivers 372
1.3.4.1 Molten-Salt and Sodium Receivers 372
1.3.4.2 Volumetric Receivers 373
1.4 Very High Temperatures 373
1.4.1 Solar Gas Turbines 375
1.4.2 Chemical Reactions 375
1.5 Storage 376
1.5.1 Chemical Storage 377
1.6 Economic Considerations 378
2. **Photovoltaic Systems** 379
2.1 Introduction 379
2.1.1 Single-Crystal Silicon 380
2.1.2 Gallium Arsenide Cells 381
2.2 Thin-Film Devices 381
2.2.1 Polycrystalline Silicon 382
2.2.2 Amorphous Silicon 382
2.2.3 Cadmium Telluride 383
2.2.4 Chalcopyrites 383
2.2.5 Tandem and Multijunction Cells 384
2.3 Concentrators 384
2.4 System Integration 385
2.4.1 Modules and Arrays 385
2.4.2 Encapsulation and Support 385
2.4.3 Power Conditioning 386
2.5 Storage 386
3. **Photochemistry and Solar Lasers** 387
3.1 Applied Solar Photochemistry 387
3.2 Solar-Driven Lasers 388
3.2.1 Materials for Solar Lasers 388
3.2.2 Nonimaging Concentrators 388
3.2.3 Heat Removal and Efficiency 390
3.2.4 The Wavelength of Solar Lasers 390
3.2.5 Applications of Solar Lasers ... 391
Works Cited 391
Further Reading 392

INTRODUCTION

In this article we concentrate on technologies for harvesting solar radiation that reaches the surface of the earth for the benefit of mankind. We do not deal here with those indirect forms of solar energy such as wind energy, hydropower, biomass, or fossil fuels.

The earth intercepts about 5.6×10^6 exajoules (EJ) per year of solar energy (1 EJ = 10^{18} J), about 1.7×10^6 EJ of which are backscattered by the upper atmosphere into space, allowing 3.9×10^6 EJ to reach the surface of the earth or to be absorbed by the lower atmosphere. The bulk of this energy is ultimately reradiated to space in the form of infrared radiation, but a small fraction is converted to the kinetic energy of winds, flowing water, and ocean currents; the potential energy of precipitation; and chemical energy produced by photosynthesis. The latter, biomass, is, of course, the basis of all life on the planet.

Of the total solar radiation reaching the surface of the earth, about 7.9×10^5 EJ/yr are absorbed by the land masses, of which some 9.3×10^4 EJ/yr are in the sunny deserts. If we consider only the latter quantity as available for harvesting, it still amounts to about 300 times the present world demand for energy. At the present time, on a global scale, the proportion of the solar energy harvested directly by mankind is negligible—in

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

spite of the enormous potential of this renewable source and its relatively benign environmental nature. In the present article we review the problems involved in the large-scale utilization of solar energy and the various efforts under way to overcome them.

Solar radiation is not distributed uniformly around the globe, as the map in Plate 1 shows. We note the expected relatively high insolation values for the world's major deserts. Tracts of land that could be devoted to solar energy harvesting without competing unduly with other essential needs (food production, for example) are available in the desert areas. In fact it may be shown that a small fraction of the world's deserts, if devoted to solar energy harvesting, could cover mankind's need "forever." These considerations point to the world's deserts as the future sources of energy.

But Plate 1 also shows the lack of overlap between these potential sources of "unlimited" energy and the major centers of population and industry where this energy is needed. This, perhaps, is one of the major handicaps to the large-scale use of solar energy, for it requires developing technologies for the large-scale storage and transportation of the harvested solar energy. There is, of course, also the inherent problem arising from the intermittent nature of solar radiation (day and night), and its somewhat stochastic behavior (weather dependence). Possible solutions to all these problems are discussed in the following pages.

Solar radiation reaching the earth must be converted to some other forms of energy before it can be put to use by mankind. There are three possible routes for such conversion (see also ENERGY CONVERSION):

1. the thermal route, where the radiation is first absorbed and converted to heat and then to other forms of energy, if needed. (Sec. 1);
2. the direct conversion to electricity (Sec. 2); and
3. the photochemical route, where the radiation induces a chemical reaction generating products of higher-energy content than the starting material (Sec. 3).

Although in nature the last route dominates, it has not so far been possible to emulate it artificially.

1. THERMAL PROCESSES

Solar radiation reaching the earth corresponds to that emitted by a blackbody at 5780 K after some modification by absorption and scattering in the atmosphere. For the thermal processes the exact spectrum of the radiation is unimportant. On reaching the surface a fraction α is absorbed and is converted to heat. Some of this heat, in turn, is lost by various processes, and only the residual amount is available for harvesting. The efficiency of the collection is defined as the ratio of useful heat recovered from the collector to the solar irradiation impinging on an equal area if set to face the sun. Its maximum value, for a given insolation, depends clearly on the magnitude of the losses. In analyzing the various thermal losses of a solar collecting system, one must distinguish between those due to convection and conduction and those due to radiation from the heated surface. The former can, in principle, be reduced to very small values by proper design. The radiative losses, however, cannot be reduced at will. At the very least, the surfaces that are exposed to sunlight (and there may be additional ones) radiate energy back into space. The rate of reradiation is determined by a property of the surface material known as its emissivity, ϵ, and its temperature according to the Stefan-Boltzmann relationship:

$$E = \epsilon\sigma T^4. \tag{1}$$

The value of the constant $\sigma = 5.6705 \times 10^{-8}$ W m^{-2} K^{-4}. The value of ϵ depends on the spectrum (and hence also on the temperature) and may vary from nearly 1 to less than 0.1 in different materials.

Because of this, the thermal efficiency of solar collectors is a somewhat complicated function of the temperature at which it is operated. A very simplified form of this function (where heat losses by conduction and convection and the effect of covers or windows are neglected) will be used to illustrate the relationship between the thermal efficiency of collectors and the temperature:

$$\eta = \alpha - \epsilon\sigma T^4/SI, \tag{2}$$

where ϵ and σ have been defined before, I is the solar flux (in units of W/m^2), and S is degree of concentration of the solar radiation. In Fig. 1 is shown the relation between the efficiency of heat recovery and collector temperature assuming a perfect black surface ($\alpha = 1$ and $\epsilon = 1$, curve *a*). We note the rapid drop in efficiency as the temperature rises toward 100 °C, with useful operational range of 60–70 °C.

A good deal of control of the collector's operating temperature and efficiency is possible by manipulating the parameters of the energy balance equation [Eq. (2)] by proper design and appropriate choice of materials. In this simplified equation there are three parameters that can be manipulated: α, ϵ, and *S*. In general, each application has its own combination of parameters that optimize the system, and we shall discuss them further for each class. We shall proceed from those involving low temperatures to the highest ones and in each case outline the principles of the design.

1.1 Low-Temperature Applications

These are characterized by temperatures below 100 °C. As is shown in Fig. 2, curve *a*, at this temperature level no special adjustment of the parameters is necessary, and the system will operate with reasonable efficiency of conversion to useful heat with α and ϵ close to 1 (i.e., ordinary black surface) and with no concentration of the solar radiation ($S = 1$). However, a great improvement is possible with relatively little extra cost, through the use of selective surface coatings of the collector. These are surfaces that show very high absorption of the solar spectrum (the absorption coefficient α) while possessing a low emissivity in the infrared region (the coefficient ϵ). Several commercial processes are available that give surfaces with α greater than 0.9 and ϵ of about 0.1. Figure 2, curve *b*, shows the theoretical maximum efficiency curve for a collector fitted with such a selective surface.

1.1.1 Domestic Hot Water The most widespread application of low-temperature solar energy is probably in domestic hot-water supplies. Here the aim is to heat water to about 70 °C and store it for use for about a day. Very simple systems, with no moving parts, have been developed requiring no special maintenance efforts. The heart of such systems is a collector surface of a few square meters area covered by a transparent window (usually ordinary window glass). Many designs are available commercially. Most include a selective coating of the absorber surface. The collector is connected to a storage vessel, and water is circulated through both by the thermo-syphon principle. Such a design is sufficient for locations free from frost. In places where freezing of the water may occur, a somewhat more complicated design is necessary, where a separate loop of an antifreeze medium is used to couple the collector and storage tank.

1.1.2 Solar Ponds Solar ponds are also designed to operate at temperatures below 100 °C. These are aimed at the large-scale

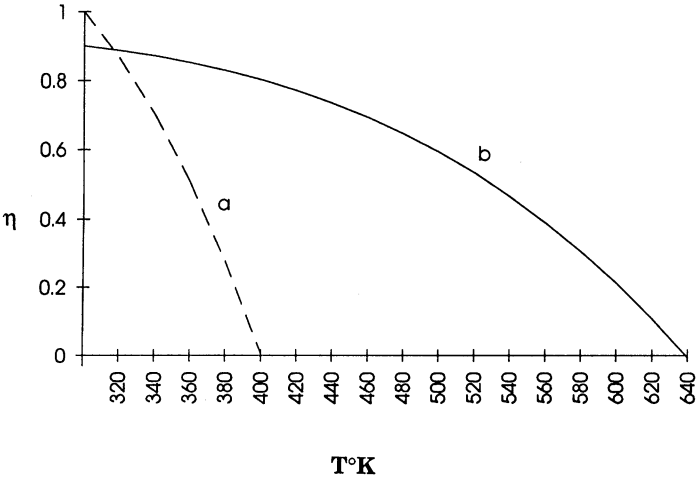

FIG. 1. Efficiency (η) of heat recovery as function of collector temperature (K). Curve *a*, $\alpha = 1$, $\epsilon = 1$, $c = 1$; curve *b*, $\alpha = 0.9$, $\epsilon = 0.1$, $c = 1$.

production and long-term storage of solar hot water for further use either for the generation of electricity or as a source of energy for desalination.

The natural pond or lake does not get heated up much by the sun because convection soon brings any warm water from the lower layers to the surface where it cools. If such convection could be prevented, water heated up by absorption of solar radiation near the bottom would stay there, and its temperature would rise (and could even reach the boiling point). The upper water layers then act as transparent thermal insulation. Convection can be stopped by establishing and maintaining a vertical density gradient in the pond, and this is achieved by creating a salinity gradient. Under such conditions the hot, saline, water near the bottom of the pond will still be heavier than the upper layers, and so no convection will occur (Fig. 2). Systems based on this technology have been under development for several decades in several countries, the largest demonstration plant being in Israel, where this work originated. A 2.5 × 10^5-m^2 pond, constructed on the north shore of the Dead Sea, supplied hot water to an organic-fluid turbine to generate 2.5 MW of electricity. This plant operated for several years and proved the principle. However, no larger-scale facility has been built anywhere in the world so far. The main handicap arises probably from the low efficiency of the energy conversion to electricity (about 1% overall), which is not completely compensated for by the inherent simplicity of the collection system and its large storage capacity. It is quite possible that as a source of heat for multiple-effect desalination plants specially designed for low–source-temperature operation, solar ponds may become an ideal energy source.

1.2 Medium Temperature

As soon as temperatures above about 200 °C are required, at reasonable recovery efficiency, controlling heat losses by the use of selective surfaces alone becomes insufficient (see Fig. 1, curve *b*). However, the last term of Eq. (2) can be further reduced by increasing the factor *S* in the denominator. This means that the solar flux must be concentrated before impinging on the absorbing surface. The effect of doing this is equivalent to reducing the area of the surface from which heat can be radiated, without at the same time reducing the solar energy input.

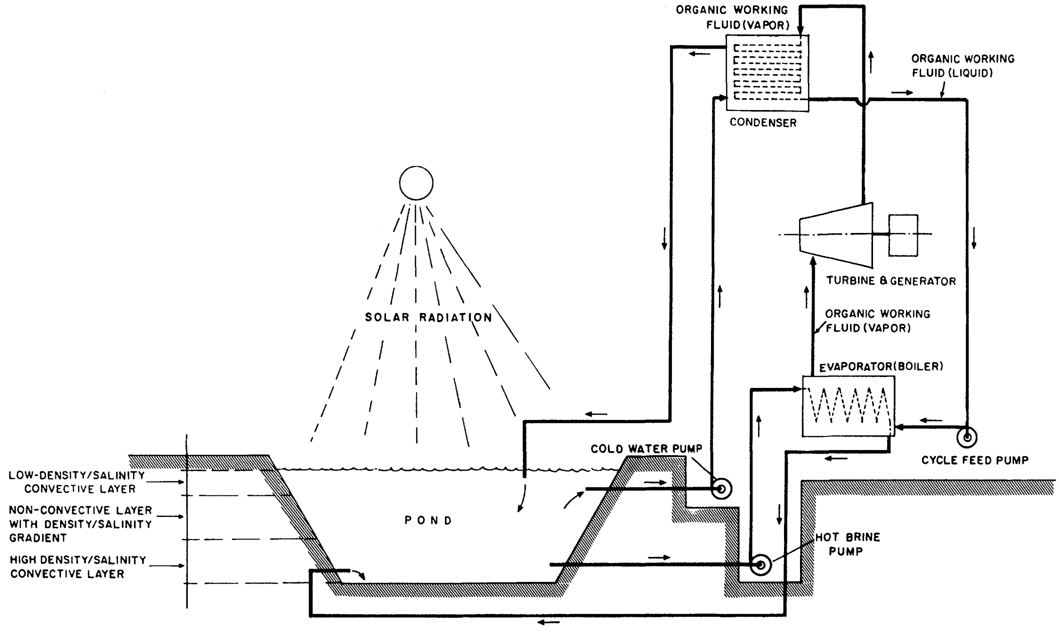
ORGANIC WORKING FLUID (VAPOR)
ORGANIC WORKING FLUID (LIQUID)
CONDENSER
TURBINE & GENERATOR
ORGANIC WORKING FLUID (VAPOR)
EVAPORATOR (BOILER)
SOLAR RADIATION
COLD WATER PUMP
CYCLE FEED PUMP
LOW-DENSITY/SALINITY CONVECTIVE LAYER
NON-CONVECTIVE LAYER WITH DENSITY/SALINITY GRADIENT
HIGH DENSITY/SALINITY CONVECTIVE LAYER
POND
HOT BRINE PUMP

FIG. 2. Schematic diagram of a solar pond power plant.

To achieve this in practice requires a sharp change in the technologies involved and in the economics of the systems.

Once the need for concentration has arisen, the emphasis shifts from the absorber to the concentrating collector. In the low-temperature range, the absorber (which is also the collector) is the extended element, and the cost of the plant is proportional to its size. With concentration, the size of the absorber shrinks drastically, and the extended area is now the collector and not the absorber, which now largely determines the economics of the plant. Ironically, at the very high temperatures—at the other end of the scale—the emphasis shifts back to the design of the absorber (see Sec. 1.4).

The only practical method of concentrating solar radiation over large areas is by reflection, and the cost of the reflecting surface becomes a dominant factor. The degree of concentration required depends very much on the details of the processes that follow. It will be recalled, of course, that in general the higher the source temperature, the higher will be the efficiency of any thermal conversion process (Carnot principle), but so will be the losses. Therefore, for applications where a conversion of solar thermal energy to mechanical energy is involved, there is a conflict in the temperature requirements. For the thermal collection step, the lower the working temperature, the higher will be the efficiency, while for the conversion to mechanical energy, the higher the temperature, the greater will be the efficiency. Obviously an optimum value for the process as whole must exist somewhere. Figure 3 shows the result of the folding of the two efficiency curves. The values shown are for the theoretical ideal efficiency. In reality, of course, the actual values will be much smaller (could be as small as half the theoretical maximum).

1.2.1 Process Heat The temperature range 100–300 °C covers industrial and commercial applications known generically as "process heat." This temperature range is too low for efficient primary electric power generation. At the lower end of this temperature range, fairly simple concentrating devices may be sufficient, and tracking the sun is not essential. At the upper end, shaped mirrors and solar tracking are necessary.

The applications in this temperature range are increasingly moving into the domain of cogeneration (electricity and heat). Because the additional effort of achieving higher solar concentration, and consequently higher temperatures, is not very great, it may become economically advantageous under certain conditions to combine electricity gen-

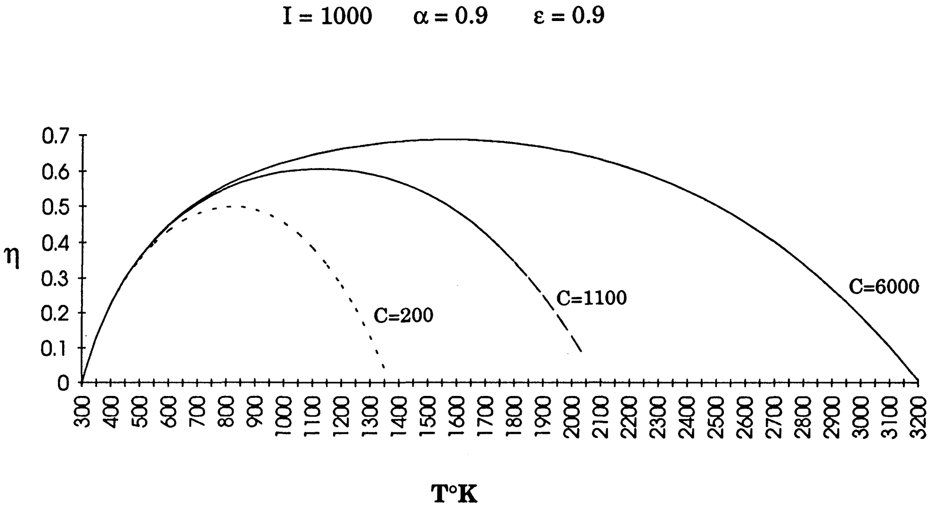

FIG. 3. Efficiency of conversion (η) of solar radiation to mechanical energy as a function of working temperature for various degrees of concentration *C*. For solar normal irradiation of 1000 W/m^2, ambient temperature of 300 K, α = 0.9, and ϵ = 0.9.

eration and process-heat production into one cycle. The high temperature is used to generate electricity, and the rejected heat from this stage, at a lower temperature, serves as process heat. It is likely that in the future the needs for the latter will be covered mostly through such cogeneration plants whenever both electricity and heat are needed in the same site.

1.3 High Temperatures

The temperature range of 300–700 °C is applied mainly to the generation of electric power, using the Rankine cycle. To achieve these temperatures, highly concentrated solar energy is needed (see Fig. 3), and there are several technologies available: parabolic troughs, reflecting dishes, and heliostat fields, all involving tracking the apparent motion of the sun.

1.3.1 Trough Systems This technology is the basis of the largest solar commercial electricity-generating plants erected to date. In southern California, a total of some 350 MW_e (megawatts electric) were in operation as of 1994 using the trough technology. As the name implies, these concentrators consist of long parabolic reflecting troughs, usually of silvered glass, giving a line-focussed solar radiation. Along the focal line a pipe is located that acts as the radiation absorber. The heat generated in it is transferred to a flowing oil stream, which in turn transfers it to a water boiler to generate steam and drive a conventional Rankine-cycle turbine and generator. To reduce heat losses to the minimum, the receiver pipe is coated with a selective surface and enclosed in an evacuated glass jacket. The parabolic trough assembly, usually aligned in a N–S orientation, tracks the sun along one axis during the day using a servo system driven by a sun sensor (Fig. 4). The combination yields a concentration ratio (defined here as the ratio of the area of the aperture collecting the solar radiation to the area from which heat is lost from the absorber tube) of about 26.

The oil exits the solar field at a temperature of 390 °C, and the steam generated by the circulating oil is delivered to the turbine at 371 °C and 100 bar. Overall efficiency (solar radiation to electricity) is reported as

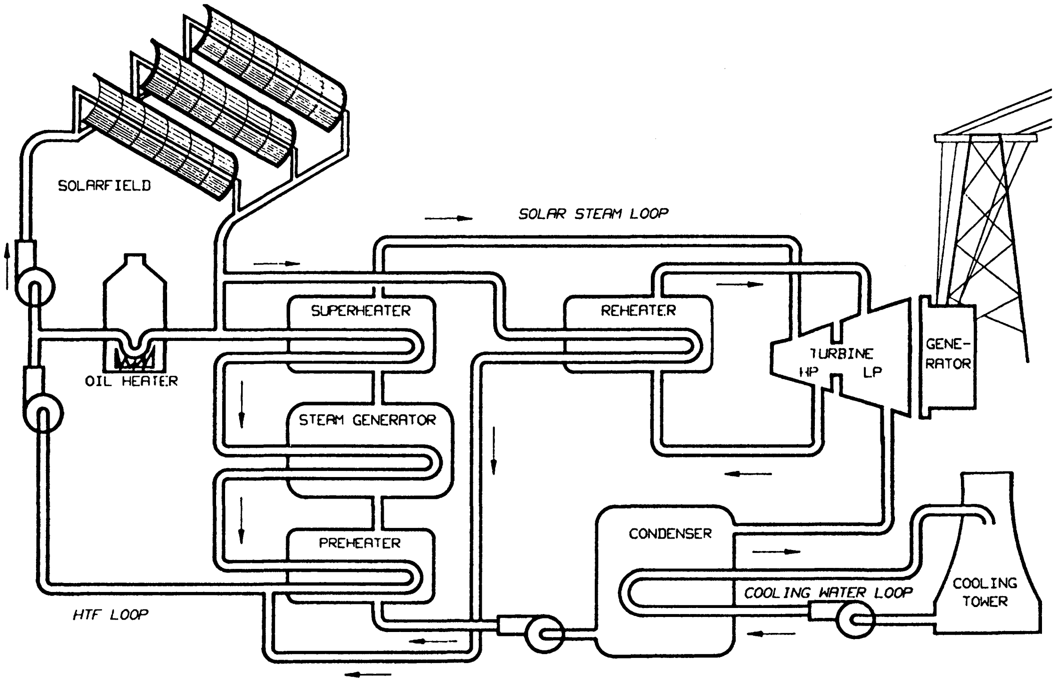

FIG. 4. Schematic diagram of a trough solar power plant (LUZ SEGS VIII).

13.9%. Because of its essentially modular design, there are no special problems in scaling the plant size to meet any demand. This provides the first example of a distributed system, where the heat output of many individual collectors arranged in a "farm" are combined to feed a power block to generate electricity.

In these plants no provision is made for solar energy storage, and instead, natural-gas–burning oil heaters are included in the design as standby. These not only serve to stabilize the operation of the plant against fluctuation in the solar energy flux but also enable it to operate in a fossil-fuel-only mode. They are therefore an example of "hybrid" plants. In California, the maximum permitted fraction of fossil fuel for these plants was 25%, and so the operators have to plan the use of this resource very carefully to maximize their profit. These plants are connected to a large grid and try to feed electricity to it mainly at times of high demand. At their specific locations these match roughly the high-insolation periods.

The design described above is the state of the art of this technology at present. Research and development efforts are under way to eliminate the oil heat-transfer circuit and to raise the steam directly in the absorber tube. If this is successful, it is estimated that considerable economic advantage will result.

1.3.2 Dish Systems The parabolic trough described above delivers a line-focus image with a modest concentration of the solar flux. Higher concentration ratios and, therefore, higher conversion efficiencies can be obtained, in principle, by optical systems that form a "point" image. The simplest of these is the parabolic dish. A single glass mirror, like the ones used in the largest telescopes, would be too heavy and far too expensive to be of practical use in solar energy applications. So other ways were developed by which a reflecting parabolic dish surface could be approximated of as large a size as is consistent with adequate mechanical strength and rigidity and of sufficient optical quality to reach high solar concentration.

It must be remembered that the sun is not a point source, but rather a disc with an angular radius of 4.653 mradians. So even in the ideal case the image will be a disc and not a point. Furthermore, the sun's disc is not uniformly radiant. Scattering processes both in the solar photosphere and in the earth's upper atmosphere cause a spreading of the theoretical "flat top" intensity distribution to a more "bell-like" shape with certain "limb darkening" and spreading beyond the angular radius (the "sunshape"). The actual sunshape has been determined by measurements, and various approximation formulas are available. It should be noted that the sunshape is not independent of atmospheric conditions and for certain very precise applications must be determined locally. These effects will further degrade the image. In general, however, for most practical and quick calculations the idealized disc with uniform distribution may serve as a sufficiently accurate model of the sun.

Many approximations to the ideal configuration of a parabolic dish concentrator have been tried (Fig. 5). They can be classified into the following three types:

1. A reflecting surface assembled from a large number of suitably shaped pieces of a matrix carrying a reflective surface of silvered glass or a metallized plastic film. The resulting reflecting shape can give a reasonable approximation to a paraboloidal surface. Several variants of this technology have been constructed and de-

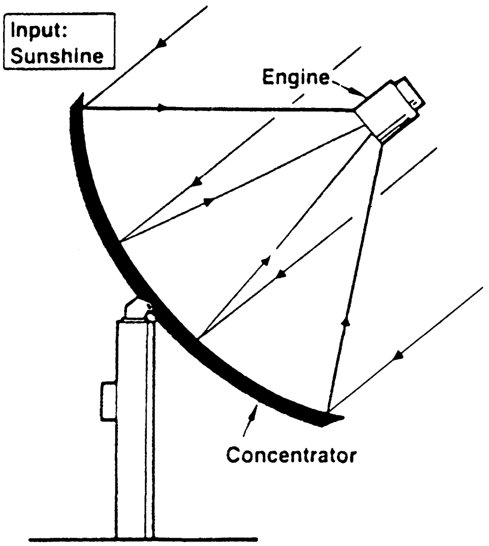

FIG. 5. Parabolic dish collector system.

ployed with reflecting area of up to 400 m^2. Concentrations of a few thousand suns have been obtained.
2. A single stretched reflecting membrane. The best example of this design is that of SBP (Germany). A thin sheet-metal membrane is stretched across a drumlike structure and deformed to the desired shape. For better reflectivity thin second-surface silver glass mirrors are glued onto the metal membrane. The largest dish constructed according to this design has a diameter of 17.5 m and a reflecting area of 227 m^2 (Lampoldshausen, Germany). Smaller units, of 7.5-m diameter, have also been built. They provide a concentration ratio of about 600 and are used to drive Stirling engines (see below) in Almeria, Spain.
3. A multifaceted assembly of reflecting membrane elements. The difficulty of manufacturing large single stretched-membrane dishes has led to the design whereby the reflecting surface is broken up to a number of smaller stretched-membrane units. An example of this approach is the LaJet design, where 24 facets (made of metallized PMMA foil), 1.8 m in diameter each, are assembled on a suitable structure to simulate a large dish of 9.4-m diameter. The curved shape of each facet is achieved by maintaining a slight vacuum behind the membrane. Solar concentration ratio of 225 has been reported.

The difficulties resulting from the need to locate complex energy converting devices at a movable focal point "floating in mid-air" has prompted attempts at designing a fixed-focus dish. A model of such a solar collector is now under development in Germany (Bomin Solar).

1.3.2.1 Dish Farms. The maximum practical size of an individual dish of any of the above designs is limited by mechanical considerations of weight, rigidity, wind load, and other. A size of around 400 m^2 seems to be the limit at present. This means that under good insolation conditions (say 900 W/m^2) there will be some 340 kW of thermal energy at peak time delivered to the absorber. Clearly, therefore, many such units have to be connected together, to provide a meaningful commercial output.

There are two ways of connecting together many dishes into a solar "farm," each requiring a different technology:

1. Each dish can carry only the heat absorbing unit at the (moving) focus, and the hot heat-transfer fluid is then pumped through an extensive piping network to a central generating unit. The reduced aperture of the cavity receivers will reduce the heat losses and thereby improve the efficiency, but on the other hand, the extensive piping network will entail extra heat losses and pumping work. Several dish farms were constructed and operated in the last decade according to this scheme. As examples one can quote the STEP cogeneration plant at Shenandoah, Georgia, U.S.A., where 114 dishes with a combined aperture of 4386 m^2 heated oil to 400 °C to raise steam and produce electricity. A somewhat more recent example is the Solar Plant 1 at Warner Springs, California, U.S.A., where 700 dishes with a combined aperture of 30 261 m^2 generated steam at 371 °C and 41 bar to generate 4.8 MW_e of electricity at peak. In Australia a dish farm uses steam as the heat-transfer medium.
2. In another approach each dish carries a complete electricity-generating unit at its focus, and the output from all of them is combined to give the required plant capacity. For this approach it is necessary to develop a relatively small conversion unit that will accept the highly concentrated radiation at the focus of the dish and convert it to electricity. The unit, together with its heat rejection system, has to be located at the focus and move with the dish as it tracks the sun. This requirement imposes severe constraints on possible designs.

In comparing the two approaches one must remember that in choosing oil as a heat-transfer medium or even steam at "conventional" characteristics, one is back at the relatively poor thermodynamic conditions and must accept the low efficiencies that they entail with little benefit from the higher-temperature capabilities of the dish. It is such considerations that are encouraging the efforts to develop a compact heat engine that will take full advantage of the high solar concentrations available at the focal area of a dish.

One generation system that has been receiving much attention in recent years for this application is the Sterling engine (*q.v.*). This is a compact and self-contained heat engine that can operate at elevated temperatures (>700 °C) with hydrogen or helium as working fluid. The design was conceived in Sweden many decades ago and has been developing slowly. Under the impetus of possible solar energy application, this development has received a boost in recent years. Development programs are under way now in several countries to marry the dish and Sterling engine technologies to produce a compact and reliable solar-power unit. A number of such units are undergoing solar testing at present.

A key component in these units is the absorber that is exposed to the highly concentrated radiation and transfers the high-temperature heat to the operating gas in the engine. Two approaches were demonstrated. In one, the concentrated solar flux impinges directly onto thin tubes through which the gas is circulating. In the second approach sodium is boiled (either in a heat-pipe configuration or in a boiler) and then transfers the heat to the working gas. Efficiencies close to 30% net conversion to electricity have been achieved. The modular nature of these units would permit the combination of any number of them to meet local needs, with the flexibility of a distributed cluster system.

At the present time, apart from the few experimental dish–Sterling rigs, the dish collectors with their high concentration ratios have not achieved much more than the troughs with their very modest concentrations.

1.3.3 Central Receivers A basic issue in all large-scale solar energy harvesting is the need to combine the output from a large number of collecting devices. So far we have described concepts based on combining the outputs of heat-transfer medium from many receivers in a central power-generating unit or combining the electrical outputs from many independent generators. There is, however, another way of achieving the final aim, and it is to combine the solar flux from many concentrating reflectors into one receiver to drive a compact central power-generating unit. This is known as the central receiver (or "Solar Power Tower") concept (Fig. 6).

According to this concept the solar energy is collected by a large number of curved mirrors that track the sun on two axes and direct the image onto a common target, the receiver. These mirrors are known as "heliostats" and are themselves constructed from a number of curved facets, with focal lengths corresponding to their distance from the target. They are mounted on a common tracking structure and each canted in such a way as to approximate a section of a paraboloidal surface.

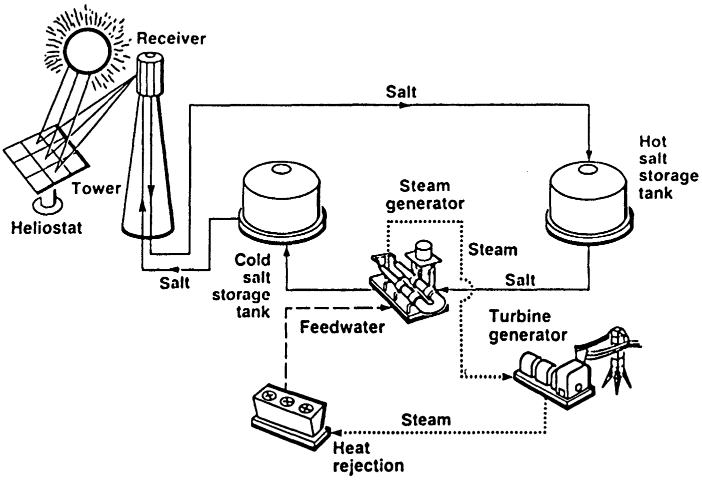

FIG. 6. Schematic diagram of a central receiver plant using molten salt as heat-transfer medium.

For fairly obvious reasons the receiver has to be mounted high enough above the plane of the mirror field to minimize blocking and shadowing of the reflected light from one heliostat by those in front of it. The tower that carries the receiver, which may be over 100 m in height, gives the alternative name to this concept, "solar tower."

Clearly the total amount of solar energy reflected into the receiver is the sum of the light reflected from the individual heliostats. But this varies greatly according to the location of the heliostat in the field and the time of day. The optimization of the layout of the field of heliostats, taking into account the blocking and shadowing by neighbors and the height of the tower, can be made using computer codes that are available for this purpose.

It is easy to see that large amounts of concentrated radiation can be delivered to the receiver to be converted to other forms. The largest field built so far, which collects some 45 MW_t (megawatts thermal) of solar energy, is in Barstow, California, U.S.A. The optimal commercial units would probably be at least ten times larger. Because of the eminent suitability of this concept to large-scale harvesting of concentrated solar radiation, much work has been devoted to the development of the various components.

The key component, in the sense that it determines the economics of the whole system, is formed by the heliostats. They provide the primary energy-collecting and -concentrating function, and their number is large. Many designs have been tried in the last two decades, and work is still continuing. The aim is to reduce cost per square meter and ensure adequate lifetime (some 30 years). At present, the designs seem to converge to heliostats of between 50 and 150 m^2 in area, made of second-surface curved glass facets. Under development are heliostats using metallized plastic films and stretched membranes (similar to those described above for the dishes).

Each heliostat in a field is individually computer controlled. The position of the sun is calculated every few seconds, and for each heliostat the positions of the azimuth and elevation drive motors needed to reflect the beam to its target are determined, and the proper commands are issued. It will be appreciated that the required algorithm here is somewhat more complicated than in the dish case, since the heliostat is not aimed at the sun but along a line bisecting the angle between the sun and the target, as seen from the heliostat. Furthermore, this direction is different for each of the thousands of heliostats in the field and is varying all the time. This may seem a formidable task, and indeed would have been impossible (or prohibitively expensive) a couple of decades ago, but is easily within the capabilities of modern computing technology.

1.3.4 Thermal Receivers The availability of large amounts of concentrated (over 1000 suns) solar energy at the receiver has opened up a new development need—how to handle this intense energy without losing its advantages. These are not at all trivial, and the emphasis of the research and development in recent years has therefore shifted from the problems of collecting and concentrating solar energy to the problem of receiver design at these high flux levels. Fortunately, in the central receiver concept there are fewer physical restraints on the design of the receiver as compared with dishes or troughs.

The mere fact that much energy is absorbed in a small area means that the fraction of the heat lost by all modes (including radiation) is drastically reduced, and the efficiency of heat production will be high. The problem is then how to transfer these high fluxes of energy to the working fluid of the plant. In most of the experimental plants built so far, the working fluid was water, as is customary in conventional power plants. The high-temperature and high-pressure steam was then used to drive a Rankine-cycle power block. Difficulties were encountered with this approach due to the limited heat-transfer capability to water boiling in tubes and the very demanding conditions imposed by the high and rather unstable solar fluxes.

1.3.4.1 Molten-Salt and Sodium Receivers. Because of the limitations of water, much attention was devoted to the search for other heat-transfer media. Two such systems were tried: liquid metal (sodium) and molten salt (a mixture of alkali nitrates). With both of these materials very high heat-transfer rates are possible, leading to efficient solar absorbers.

Both types were tested in Almeria, Spain, in the 1980s, and the expected efficiencies of the receivers were confirmed. Further work on the liquid-sodium receiver was abandoned there because of the perceived hazards of this application.

Molten-salt solar receivers were also developed and tested in France and the United States. An experimental power station designed for 2.4 MW_e (THEMIS) was built and operated for 3 years in Targasonne, France, in the 1980s. Components for molten-salt systems were studied further at Sandia laboratory in the United States. On the basis of the results obtained there, it was decided to modify the Solar One plant at Barstow (California, U.S.A.) (initially designed with a water-in-tube receiver) and change it to a molten-salt system. This modified plant has now (1996) started operation. In all these plants the power block is still based on the conventional Rankine cycle at more or less conventional thermodynamic conditions.

A number of test facilities were built and operated over the last 20 years in several countries ranging in power from about 0.5 to 10 MW_e. They all performed below expectation and showed that none of the technologies tested then was ready for commercial exploitation. Among the problems that surfaced was the great thermal inertia of the systems, which severely reduced the amount of energy supplied.

1.3.4.2 Volumetric Receivers. All the receivers described so far suffer from the disadvantage that large fluxes of heat have to be forced through tube walls and into the working fluid. This constrains the design and limits the possible efficiency of the receivers and the utility of solar energy. Development projects were initiated in several countries aimed at overcoming this difficulty by causing the concentrated solar energy to be absorbed directly in the bulk of the working fluid, usually a gas. Since air and all other working gases are transparent to solar radiation, some means of absorbing it must be devised. Two general approaches were studied. In the first, fine particles of a material suitable for absorbing the solar radiation are suspended in the gas stream, and in the second, some sort of structure is provided that will stop the solar flux and transfer the heat from its surface to the gas stream flowing past it. These approaches led to what are known as the "volumetric receivers." They all involve a cavity, with or without a window depending on the application, illuminated with very concentrated solar radiation. Inside the cavity there is the absorbing structure. Various designs of this structure have been tried: wire-mesh screens, ceramic-fiber screens, specially shaped ceramic forms, and ceramic foams. All these are undergoing laboratory and sun tests. Very encouraging absorption rates (over 2 MW_t/m^2) have been obtained with some of them, leading to physically small receivers of high efficiency and much higher gas temperature (see Fig. 7). For applications such as gas-turbine drives or chemical reactors (Secs. 1.4.1 and 1.4.2), the receivers need a window, and this is possibly the most demanding component. It is required to transmit very high fluxes of solar radiation and to operate hot and under pressure. The size of a module will probably be limited by the maximum size of the window that could be installed. In this connection the application of secondary concentrators (see Sec. 3.2.2) will be crucial, since for a given power rating they will reduce the size of window needed.

Another, and simpler, application is to use a windowless volumetric receiver to heat air (at atmospheric pressure of course) and then use the hot air to raise steam and drive a conventional power block. A 30 MW_e plant using this approach has been designed in Europe but has not been constructed yet.

1.4 Very High Temperatures

There is considerable interest in temperatures above 1000 °C for higher efficiency electricity generation and for a whole range of chemical and materials applications. As can be seen in Fig. 4, the requirement of temperatures above 1000 °C immediately translates to the need for much higher concentrations of the solar energy. And to achieve these, secondary concentrators are required. These devices are discussed fully in connection with solar lasers (see Sec. 3.2.2 and Fig. 18). Basically they are specially configured reflectors, designed according to the principles of nonimaging optics, and known as compound parabolic concentrators (CPCs) and Trumpet-type concentrators. They are placed ahead of the receiver and further con-

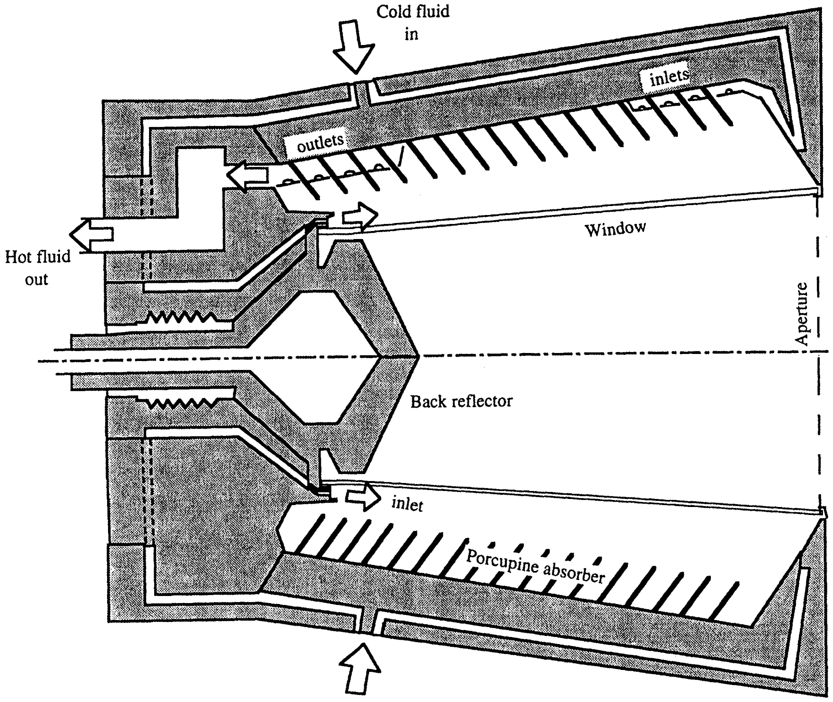
Cold fluid in
inlets
outlets
Hot fluid out
Window
Aperture
Back reflector
inlet
Porcupine absorber

(a)

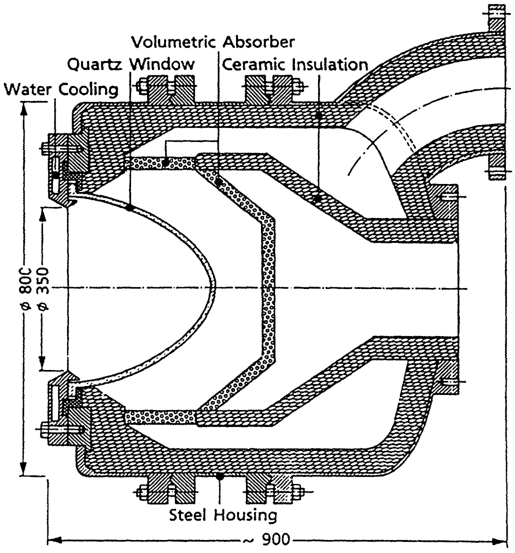
Volumetric Absorber
Quartz Window
Ceramic Insulation
Water Cooling
Ø 800
Ø 350
Steel Housing
~ 900

(b)

FIG. 7. Volumetric receivers. (a) DIAPR 50-kW receiver (WIS). Aperture diameter 115 mm. (b) VORBREC receiver (DLR). Dimensions in mm.

centrate the incoming concentrated radiation from a dish or heliostats.

1.4.1 Solar Gas Turbines Although the use of highly concentrated solar radiation is the key to high efficiency of collection in the dish and central receiver technologies, no advantage has been taken so far in these devices (apart from the dish–Sterling combination) of the availability of high temperatures to improve the thermodynamic efficiency of the conversion to electricity. To do so would require a change in both the power block and the receiver that feeds it.

Developments under way today in several countries attempt to adapt the Brayton cycle (gas turbine; see TURBINES, GAS) to solar use (Fig. 8). According to this concept, the working fluid (usually air) is compressed and heated to a high temperature (in a volumetric receiver, for example) and is then allowed to expand through a gas turbine that drives both the compressor and the electricity generator. Commercial gas turbines are available that can operate at 1300 °C, and their conversion efficiency is correspondingly high (over 50% in a combined-cycle mode). The adaptation to solar use requires the development of receivers that can deliver air at the temperature and pressure needed as input to a modern gas turbine (1300 °C and 20 bars).

Early experiments with solar gas turbines, without secondary concentration, were based on a cavity receiver lined with silicon carbide tubes through which air at the required pressure was passed. Temperatures of about 1000 °C were reached and used to drive a hot-air turbine coupled to an electricity generator. The limited heat-transfer capability of the tubes led to a bulky and rather inefficient design. Experiments are now in progress (Karni, 1994) with a volumetric receiver replacing the ceramic-tube one.

1.4.2 Chemical Reactions The very high temperatures that can be reached readily by means of concentrated solar radiation have attracted the attention of material processors and chemists. There are a large number of high-temperature endothermic processes of possible industrial interest, and some of them have been examined as candidate systems for solar energy application. Of particular relevance to the energy field is the search for fuels and for storage systems.

Hydrogen, as an alternative and environmentally benign fuel, has been the center of attention for decades (Fletcher, 1977). It can

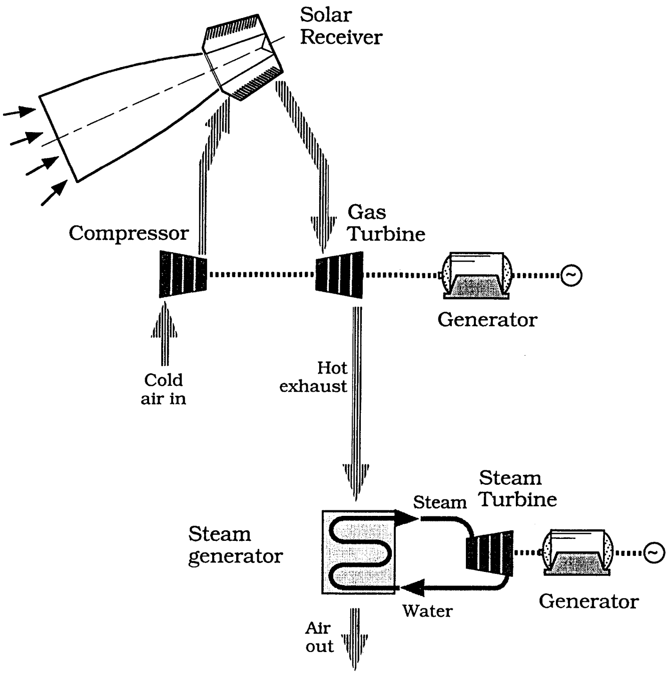

FIG. 8. Solar gas turbine (in combined-cycle mode).

be used not only directly in various engines but also to generate electricity via fuel cells and as a raw material for the synthesis of methanol and other liquid fuels. At temperatures above 2000 °C water is appreciably dissociated to its elements. If a method of separating hydrogen and oxygen, while still at the high temperature, can be developed, there will be a direct route to hydrogen production not involving electricity (Winter, 1988). Research to this end is under way, and some progress has been made.

A way of sidestepping the problem of separating the two elements is to carry out the decomposition of water in two or more steps. Much work was done on such an approach in connection with nuclear energy, but the upper temperature available from the reactors is around 900 °C. This limited the choice to fairly complex chemical systems. The higher temperature available from the concentrated sunlight opens additional, and simpler, reactions. One such reaction scheme involves the decomposition, by intense concentrated sunlight, of magnetite (Fe_3O_4) particles suspended in a gas stream to ferrous oxide and oxygen. The ferrous oxide is then reacted with water vapor to give hydrogen and re-form the magnetite (Kuhn *et al.*, 1994). The thermal decomposition of certain oxides (e.g., zinc) can also make it possible to design similar cycles, and in addition provide a means of energy storage (Fletcher, 1984).

Another direction in the search for fuels is the gasification of biomass (wood, forestry, and agricultural residues) and other carbonaceous matter. Under high temperatures (around 900 °C) such materials (with the addition of water if necessary) form a gaseous mixture containing hydrogen, carbon monoxide, and other components depending on the nature of the feedstock. This mixture can be used as such, or purified to give hydrogen or synthesis gas (a mixture of hydrogen and carbon monoxide), and they, in turn, can be used as starting points for liquid fuels. It should be pointed out here that when the feedstock is biomass, the fuels produced in this way are environmentally benign because their use and production do not add any carbon dioxide to the atmosphere (Fletcher, 1991; Epstein, 1994).

Much of the research described in this section is carried out with the help of a solar furnace. Essentially it is a fixed-dish concentrator illuminated by one or more heliostats. The principal advantage of such a facility is that the experimental zone, at the focus of the dish, is stationary. The largest solar furnace, 1000 kW (thermal), is in Odeillo, France. Many other much smaller solar furnaces, in the tens of kilowatts rating, are operating in research facilities around the world.

Another possible application of solar energy has surfaced in recent years, and it is the detoxification of wastes and the purification of contaminated water and soil. These problems are assuming a major proportion, and considerable efforts are being devoted to solve them. Solar energy can play a part in two ways:

1. by a thermal decomposition of the contaminants at high temperature, and
2. by photochemical decomposition of the contaminants.

The second, and more extensively researched photochemical route, is discussed in Sec. 3.1.

1.5 Storage

No mention has been made so far of a crucial element of solar energy plants, and it is that of energy storage. Energy storage is an important, but not essential, feature of any conventional power system and is used primarily as a means of load leveling and improving the economics.

In the case of solar energy the problem of storage becomes crucial. The sun's radiation is not a steady source of energy either on the short term or on longer terms, diurnal or seasonal. The short-term fluctuations are stochastic in nature and cannot be programmed into the operational procedure, yet may have a profound effect on the performance of the plant. The longer-term storage is particularly important where the solar plant is expected to be the sole, or the major, source of power ("stand-alone" operation).

The technical solutions to these problems vary greatly according to the specific requirements of the system. For the short-term storage, whose purpose is mainly to stabilize the operation of the power-generation equipment, some partial uncoupling (buffering) between the heat-generation part (the solar component) and the power block is the common solution. In systems using a liquid heat-transfer medium—oil, molten salt, or molten metal—a common solution is to provide storage tanks in the heat-transfer loop. In some systems only one tank is provided where mixing of the hot and cold liquid (usually oil) is avoided. This is known as the "thermocline" design. A stratified regime is maintained in the tank with the interface between the hot and cold layers moving up and down according to the charge/discharge status.

More common are the two-tank systems, one cold and one hot. Liquid from the "cold" tank flows to the solar receiver, where it is heated to the working temperature, and from there flows to the hot storage tank. The power-generation part of the plant is fed from the hot tank and is therefore completely decoupled from the solar receiver (Fig. 6). When the heat-transfer fluid is a gas, the same result may be achieved by placing a sensible heat-storage device in the circuit between the receiver and the power block. It may take the form of towers or vessels filled with solids, such as ceramic bricks, metal forms, boulders, gravel, and similar materials.

The various systems have been tried in the experimental plants operated so far and ranged in storage capacity from tens of minutes to a few hours. The shorter–storage-time devices were used mainly as stabilizers to deal with short-term fluctuations of insolation. The longer-duration storage devices were used to improve the economics of the plant by extending the operating periods by a few hours so as to match closer the demand curve. In this case the solar-collecting field must be larger than that needed to provide the nominal output at the design point, with the excess energy collected going to the storage device. The ratio of the size of the actual field to the nominal one is known as the "solar multiple" of the plant.

Another approach to extending the storage capability is to use phase-change systems. In these the heat is stored in the form of latent heat of melting of suitable solids and is released upon solidification. Such devices, therefore, operate by cycling around the melting point of the working substance. The early development of such systems was aimed at the long-term (diurnal) storage of low-temperature heat (for domestic use), and the materials used were various salt hydrates of rather low melting point. More recent advances, aimed at the power industry, developed composite ceramic-salt materials loaded with high-melting salts (Gluck *et al.*, 1992).

The really long-term storage of solar energy is a much more difficult problem. So far only low-temperature seasonal heat-storage capability has been demonstrated, and it is in the solar pond (see Sec. 1.1.2). A more universal system is under development at present and involves the conversion of solar energy into chemical energy, i.e., into suitable chemicals that can be stored indefinitely and release the energy on demand. These systems are discussed in the following section.

The problems associated with storage can be avoided if one is prepared to sacrifice the total dependence on solar energy and adopt the hybrid mode of operation. In this mode the use of fossil fuel together with, or instead of, the sun is possible and permitted. In this case fossil burners are placed in the heat-transfer circuit and supplement or replace the solar radiation. In fact most commercial plants operating today are of this nature (see Fig. 4).

Where cheap natural gas is available, the decisions whether to adopt the hybrid approach, and to what extent, is a matter of economics and environmental considerations.

1.5.1 Chemical Storage The importance of long-term storage of solar energy cannot be overstated. It is the key to the large-scale exploitation of the largest potential source of renewable energy on earth. Without long-range storage ability, solar energy can only supplement existing power networks to a limited degree. A related factor was mentioned early in this article, and it is the mismatch between the regions with ample insolation and regions of intense energy use (see Plate 1). The solution to this problem requires the ability to collect solar energy in the sunny areas, store it, and then deliver it to the centers of demand.

Research and development have been under way for over a decade on a possible solution to both the above challenges. It involves the conversion of solar energy to chemical energy, that is, to the energy stored in energy-rich substances from which they can be recovered. In principle any endothermic, reversible, reacting system could meet such a need, but careful analysis of various possibilities has limited the choice so far to a very few systems. The most widely studied is the reversible reaction between methane and water (or carbon dioxide) to give a mixture of hydrogen and carbon monoxide. This is a well-known reaction in the petrochemical industry where it is known as the "reforming reaction." It proceeds in the presence of suitable catalysts at around 900 °C with the absorption of large amounts of heat. The mixture of the products (hydrogen and carbon monoxide) can be stored at ambient temperature indefinitely. This mixture of gases, also known as "synthesis gas," is a very useful starting material for many chemical compounds (hence the name), including hydrogen and fuels. It can be used in fuel cells to generate electricity, and it can be burnt very effectively in gas turbines and to provide process heat. But in all these applications, most of the energy originates in the methane used in the reforming reaction, and the contribution of solar energy is only 27%.

Nevertheless, this reaction is of great interest to solar energy because of its reversibility. With the help of suitable catalysts, it is possible to reverse the reactions, this time with the evolution of all the solar heat absorbed in the reforming reaction and the regeneration of the starting materials. This makes it possible to use these reactions in a closed-cycle mode (see Fig. 9) to store and transport solar energy without using up any materials and without any emissions to the atmosphere. This cycle has initially been proposed in Germany as a means of transporting nuclear-generated heat over long distances (Harth, 1981). The complete cycle has been demonstrated on about 10 kW (thermal) of solar energy at the Weizmann Institute in Israel, and a scaled-up version (to 500 kW thermal) is now operating there (Levy, 1993). The reversible decomposition of certain metal oxides, mentioned above, being studied at present, is also a possible route to long-term storage of solar energy.

1.6 Economic Considerations

Solar energy is by far the largest renewable energy resource on earth, and moreover it is almost entirely benign from the environmental point of view. And yet it has not been exploited directly on a large scale by humankind. The reason, obviously, is that to date it has not competed successfully with inexpensive fossil sources. The lack of commercial interest, in turn, suppressed research and development efforts. Over the past 30 years, with the increasing awareness of the finite nature of the fossil resources and the adverse environmental effects of continued dependence on them, the attitude has changed somewhat. More effort is being exerted on the international level to develop the commercialization of solar energy, although by no means commensurate with the effort devoted for many decades to the development of nuclear energy. The following paragraphs discuss some of the economic hurdles confronting solar energy.

Since the incoming solar radiation costs nothing, the economics of solar energy is dominated by the capital investment in the plant. Solar radiation is a rather dilute source of energy, in the sense that the flux reaching a unit area on earth is rather low, seldom reaching over 1000 W/m^2. Potential sites for exploitation of solar energy are con-

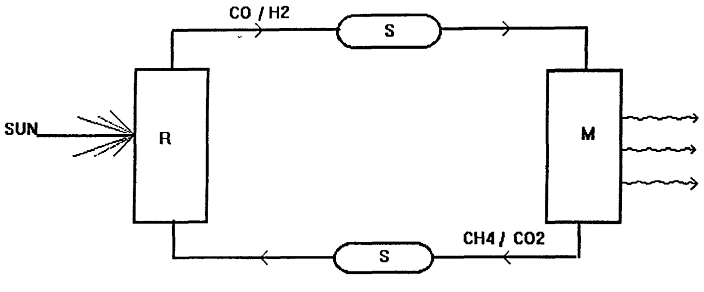

FIG. 9. Solar chemical storage and transport system. R, reformer; M, methanator; S, storage vessels.

sidered very good if the annual irradiation is greater than 2300 kW h/m^2 and good if it is over 2000 kW·h/m^2. This immediately translates to the need for very large areas to collect a meaningful amount of energy. Even if one neglects the cost of the "real estate," the cost of the collecting surface becomes a large fraction of the total cost of the plant.

There are several economic implications associated with the intermittent nature of solar energy. First, of course, is the fact that the solar plant is utilized for a smaller fraction of the day than it could be in a conventional plant. Furthermore, intermittent operation results in additional loss of production due to the thermal inertia of the plant. The need for storage, even for short terms to ensure stability of operation, which is absent in nuclear and conventional power stations, also puts solar energy at an economic disadvantage.

The economics of solar energy depends very much on how it fits into national energy policies. These, in turn, vary greatly from country to country. The parameters involved relate to the level of insolation available, to the availability and economics of other energy resources, to the sensitivity of the nation to environmental considerations, and to the international policies in this regard. Also of importance are questions relating to how solar energy will fit into the power economy. Are the solar stations to be part of a large network, or are they to serve isolated communities or small networks? How large will they be? What are the characteristic load curves?

We have seen already that where fossil fuels are combined with a solar plant, hybrid operation has great advantage. The most economical arrangement is where the solar plant supplies power to a network only at peak load time. This is the time when a utility is prepared to pay most for additional generated electricity. This only applies to locations where the electricity load curve overlaps the insolation curve significantly, a factor that severely limits the possible areas of applicability of solar energy. And it is not an accident that the only example of large commercial solar plants is from such a location (southern California). When the match is not perfect, some storage capacity is required to gain a few working hours, or, alternatively, hybrid operation is adopted. Careful economic analysis is necessary to determine the optimal proportion of the additional storage. In general, small solar plants that feed into a large electric network are relatively free from problems of storage (beyond what is necessary for the smooth operation of the plant).

Modern electricity-generation plants are shifting to a combined-cycle mode of operation, where the efficiency of conversion to electricity is much higher and can reach 60% as compared with about 40% for the "conventional" power station. In the combined-cycle mode, both gas turbines and steam turbines are used. The former use natural gas, operate at high inlet temperature, and exhaust the combustion product gases at temperatures of around 600 °C. These in turn are used to raise steam of a quality suitable to drive a conventional Rankine cycle. The combination leads to the high overall conversion efficiency. A solar component may be integrated into such a combined-cycle system in several ways. At present most such systems assign the solar energy to the "bottoming," i.e., the Rankine steam cycle, and use the fossil fuel to drive the gas turbine. However, with the development of solar-driven gas turbines, it will become possible to use them also in the "topping" cycle (Fig. 10).

Difficult problems arise when the solar plant is required to be of the "stand-alone" type, that is, to be the sole, or major, source of power to the network. This situation may arise when the network being served is small and remote, or when the use of fossil fuels is precluded, by regulation or availability. In such situations the need for large and long-term storage becomes unavoidable. As was mentioned above, the technology required for this purpose is still under development (see Sec. 1.5.1).

2. PHOTOVOLTAIC SYSTEMS

2.1 Introduction

Photovoltaic (PV) cells are solid-state electronic devices that convert electromagnetic radiation to electricity. Full description of the physics of these devices is given in PHOTOVOLTAIC DEVICES, and is not repeated here. Instead we concentrate on the application of

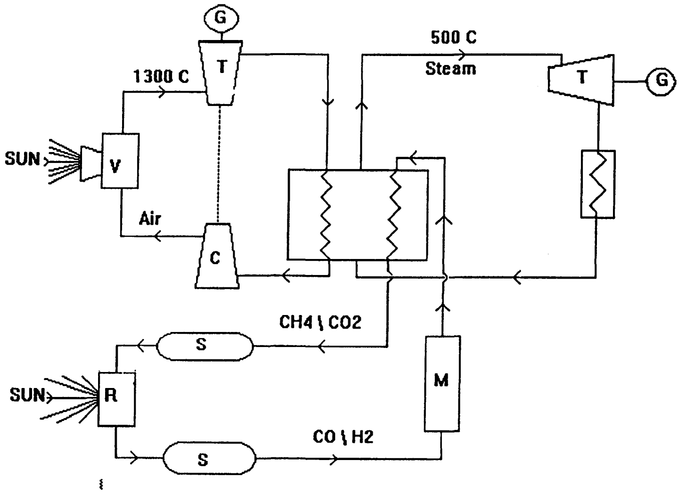

FIG. 10. A combined-cycle solar stand-alone system. V, volumetric receiver; R, reformer; T, turbine; C, compressor; M, methanator; S, storage; G, generator.

such devices to the harvesting of the major renewable energy resource—solar energy.

PV cells are quantum devices in the sense that only photons above a certain energy (the forbidden gap energy) will promote an electron from the valence band to the conduction band of a semiconductor and, if conditions are right, ultimately to an outside circuit to provide useful electric current. PV devices are therefore spectrum sensitive and can only use part of the solar spectrum. In Table 1 (Emery, 1996) we list the major photovoltaic materials relevant to solar energy along with the parameters important for PV operation.

Photons with energy lower than the band gap are not absorbed, while the excess energy of photons with energy higher than the band gap is ultimately converted to heat. Therefore the maximum efficiency is to first order determined by the band gap of the semiconductor from which the PV cell is made. Figure 11 shows these theoretical efficiencies and their dependence on gap energy. In practice the efficiencies are lower than the ideal maximum for a variety of reasons, and much of the research and development in PV is directed to improving it; and in fact the efficiencies have been rising steadily over the last few years.

2.1.1 Single-Crystal Silicon The first Si photovoltaic cell was demonstrated in 1954, a year that marked the birth of the modern photovoltaic cells. This is still the most mature technology today with some

Table 1.

Semiconductor	Band gap (eV)	Type, dopant	PV technology	Efficiency (%)	Type of junction
InP	1.35	*n*: Zn, Cd *p*: Se, Te, Si	Single crystal	21.9	Heterojunction
GaAs	1.45	*n*: Zn, Cd *p*: Se, Te, Si	Single crystal	24.3	Heterojunction
$CuInSe_2$	1.0	*n*: In_{Cu} *p*: Cu_{In}	Thin film	17.7	Heterojunction
CdTe	1.45	*n*: In, Cl *p*: P	Thin film	15.8	Heterojunction
a-Si:H	1.7	*n*: P, Sn *p*: B, Al	Thin film	10	Homo/heterojunction
Si	1.1	*n*: P, Sn *p*: B, Al	Single crystal; polycrystalline thick film	24 18.6	Homojunction

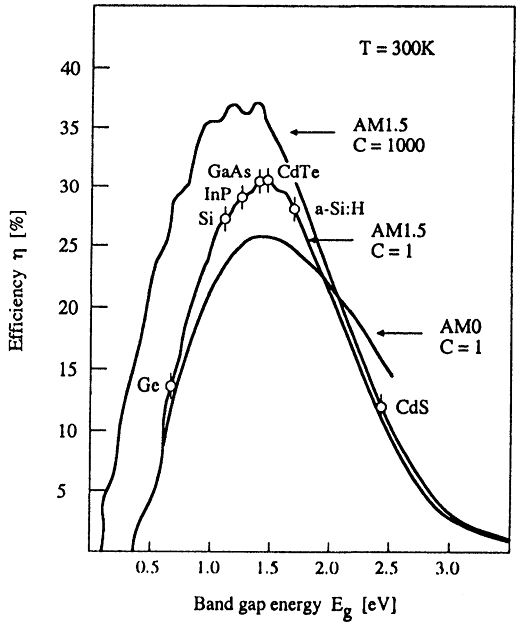

FIG. 11. Ideal solar-cell efficiency as a function of band-gap energy when illuminated under spectral conditions of outer space (AM0) and under terrestrial conditions with the sun at zenith angle of 48° (AM1.5). *C* = degree of concentration of the solar radiation.

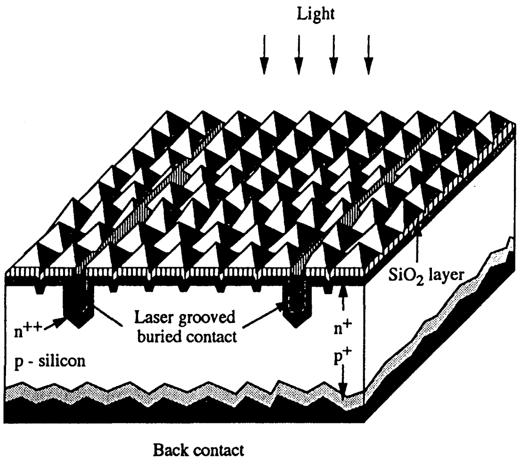

FIG. 12. Schematic representation of high-efficiency design for silicon solar cells.

40% share of the sales of all PV cells. A concise review of this technology has been published by Green (1987). A diagram of one of the most advanced versions of such a cell is shown in Fig. 12. Typical efficiency of laboratory-sized (4 cm^2) advanced PV Si cells is 24% for unconcentrated sunlight, and 28% with concentrated (200 times) sunlight. Commercially available silicon cell modules typically have efficiencies of about 12%. Modules manufactured in a semi-industrial environment have a narrow distribution of efficiencies (19.5% to 21.5%).

The high-performance (>23% efficient) Si cells are made of high-purity silicon prepared by the float-zone or Czochralski technique. They are doped with boron to make them *p* type. The back surface is made p^+ by Al alloying, which also provides the Ohmic contact and a field to drive away holes from the back surface and thus reduce recombination losses (back surface field). The front (100) surface is texturized by alkali treatment to reduce reflection losses. A further reduction in reflection and front-surface recombination losses is achieved by depositing a layer of SiO_2. Sometimes a second antireflective film of Si_3N_4 or MgF_2 is added.

In some advanced designs the Ohmic contacts are produced by grooving the front and back surfaces by a laser technique and then depositing a series of metals (Ni, Cu, and Ag) in them by an electroless process. These contacts are called "buried contacts" and minimize shadowing without increasing series resistance.

2.1.2 Gallium Arsenide Cells A number of compounds of group-III and -V elements, like GaAs, InP, GaP, etc., are semiconductors. Some of them have useful photovoltaic properties and have been investigated in the hope of producing more efficient or more economic PV cells. Gallium arsenide PV cells are particularly attractive. Single-crystal GaAs has a band gap of 1.41 eV, close to the optimum value on the efficiency–band-gap curve (see Fig. 11). GaAs cells have exhibited efficiencies of 25% with no concentration and 28% efficiency with concentrated sunlight (200 suns).

2.2 Thin-Film Devices

Until the early 1970s, the main applications of photovoltaic devices were in space with a few specialized uses in military and civilian applications. In all these applications, the cost of the PV components was of secondary importance, and emphasis was given mainly to efficiency and reliability. In the last two decades, however, with the increasing importance given to renewable energy sources and environmental issues, attention has been directed to the possibility of large-scale (utility) use of PV. While the currently available single-crystal Si (and GaAs) PV cells show high efficiency and reliability, they suffer from one major disadvantage: they are too expensive, mainly because of the need for very high-purity materials and complex and expensive production methods. Much of the research effort in recent years was directed, therefore, to cost reduction.

The strategies being pursued for cost reduction are aimed at increasing efficiency, reducing the amount of material required, obtaining cheaper materials, and finding simpler, less expensive, production methods. On reflection it will become evident that of the various options available, reducing the amount of material presents the greatest opportunity for cost saving because 100-fold savings are possible, while modifying technologies offers a more modest opportunity for savings. Economic analyses of solar PV plants lead to the conclusion that the efficiency of the less expensive cells must still be high because of the area-proportional value of system costs (balance of system).

2.2.1 Polycrystalline Silicon Polycrystalline silicon is less expensive to produce than single-crystal silicon, and although polycrystalline cells are somewhat less efficient than single-crystal cells, they continue to be a possibly viable PV technology. A number of methods have been used to make polycrystalline silicon for PV use. Large ingots (of up to 100 kg) have been cast from the melt in crucibles of high-purity graphite or silica or silicon nitride. A directional solidification was established by maintaining a temperature gradient in the crucible. The ingot was then sliced into wafers and further processed according to procedures established for the monocrystalline material. The best polycrystalline PV cells are made by this method.

Self-supporting sheets of silicon can be produced by pulling out of a melt or by extrusion from graphite die. Thin films of Si can also be grown on suitable substrates (ceramic, coated steel, and industrial silicon) by one of several methods. These include chemical deposition from solution, vacuum vapor deposition, chemical vapor deposition, and others, all potentially cheaper than high-temperature melt procedures. The thickness of the polycrystalline layers produced by these techniques is of the order of tens of microns, i.e., a factor of 10 lower than in the single-crystal cells.

Since the thickness of Si needed to absorb 90% of the solar photons incident on it is about 200 μm and the minority-carrier diffusion length in polycrystalline Si seldom exceeds 100 μm, thin polycrystalline Si PV cells require design changes. It is necessary to ensure that the light entering the cell is trapped within the active layer and that its effective path within it is much greater than the thickness of the material. A textured front surface (which can be obtained by electrochemical etching—so-called porous silicon), coupled with a mirrorlike back surface, creates a light-trapping configuration that results in a large increase in the absorption coefficient. Efficiency of 17% in 4-cm^2 polycrystalline cells has been achieved.

Even when electronic-grade silicon is used, many undesirable impurities (oxygen, carbon, transition metals), which are minority lifetime killers, may be picked up during processing and must be removed. This is done by heating between 600 and 1000 °C after depositing, by evaporation, a layer of Al to provide gettering action. After the treatment, this film, now partially oxidized, is removed by etching.

A novel approach to the fabrication of Si PV cells has been reported recently (Levine, 1991). Silicon spheres of 0.75 mm diameter are produced by a spray technique. During this process dopants are introduced, and a superficial *p-n* junction is formed. The spheres are embedded in glass matrix and contacts established by metal evaporation. An efficiency of 11.5% has been reported for arrays of this type.

2.2.2 Amorphous Silicon A further major step in the reduction of costs of materials and production has been achieved with the introduction of PV materials of high optical absorption that can be applied in very thin layers, of the order of 1 μm thick, by relatively simple techniques (Zweibel, 1993). We describe here some of the most promising examples.

Amorphous silicon, *a*-Si, is usually prepared by the glow-discharge decomposition of silane gas (SiH_4) in a radio-frequency–generated plasma and the deposition of the resulting silicon on a suitable substrate. The high density of dangling bonds that are formed in the material prepared in this way introduces states within the forbidden gap, which have a deleterious effect on the electronic properties of *a*-Si. An excess of hydrogen (up to about 10%) is introduced to saturate most of the dangling bonds and improve the characteristics of the material (which is then designated as *a*-Si:H).

a-Si:H is unique in that its band gap can be controlled, in the range 1.5 to 1.8 eV, by adjusting the hydrogen content, and even further by alloying with Ge or C. In this way materials may be obtained with band gaps between 1.3 eV (for *a*-Si,Ge:H) and 2.0 eV (for *a*-Si,C:H). This flexibility permits the design of multijunction cells that can be "tuned" for maximum efficiency. Figure 13 shows a diagram of a double-junction *a*-Si:H/*a*-Si,Ge:H cell. *a*-Si:H cells suffer from light-induced degradation. This is particularly severe in single-junction cells. Thus, a 1-cm^2 cell of *a*-Si:H with a 12.5% efficiency (compared with the theoretical maximum of 28%) degrades with time to a stable 6% efficiency. A similar three-junction cell starting with 13.7% efficiency degrades to a stable 10% value. Nevertheless, this technology is still very attractive and developing, with modules of 1 m^2 being produced for utility applications. However, unless the efficiency of *a*-Si:H PV panels increases to about 15%, they will not satisfy the requirements for large-scale PV power plants.

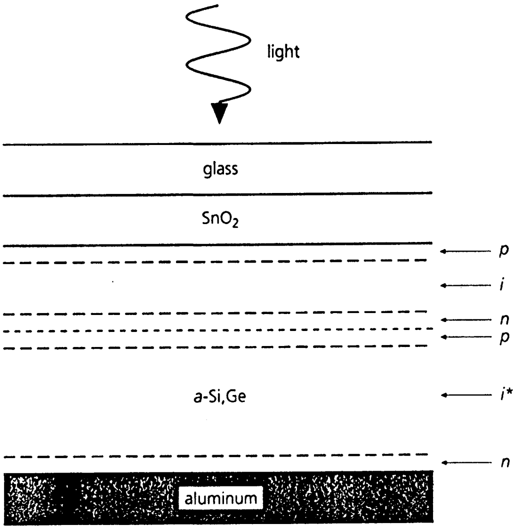

FIG. 13. A schematic diagram of an *a*-Si multijunction cell. The *i** layer of the bottom cell is an amorphous silicon–germanium alloy with a narrow band gap.

2.2.3 Cadmium Telluride Some of the semiconductors of the II–VI family also have properties suitable for thin-film photovoltaic applications. Thus, CdTe has a band gap of 1.45 eV (close to optimum) and a high optical absorption coefficient permitting the use of very thin films (a few microns). In addition it is less sensitive to impurities and lattice defects.

PV cells of this material are formed by vacuum evaporation of CdTe, followed by dipping in $CdCl_2$ solution and annealing at 450 °C. The $CdCl_2$ serves as a flux, which leads to recrystallization of the compound to give large crystallites (>1 μm). The heterojunction is completed by depositing thin CdS film. Somewhat inferior results were obtained with films formed by electrodeposition and screen-printing techniques. HgTe can be used with advantage for the Ohmic back contacts instead of gold. Air annealing of the whole cell at 200 °C for 30 min brings about substantial improvements in these cells. Small-area cells with a solar conversion efficiency of about 16% have been reported (Chu *et al.*, 1992). Large modules (1 m^2) with 8% efficiency are on the market today. As with other technologies where large amounts of Cd are involved, CdTe cells would be subject to strict environmental control, which may impede their utilization on a large scale.

2.2.4 Chalcopyrites Other promising semiconductors for solar applications belong to the group of chalcopyrites, compounds with the formula $CuMX_2$ where M = In,Ga and X = S,Se. The most developed member of this group is the compound copper indium diselenide (CISe) with a band gap of 1.0 eV. Its high absorption coefficient permits the use of very thin films (<0.5 μm). A diagram of a CISe cell is shown in Figure 14. It possesses the highest conversion efficiency of any thin-film cell, 17.7% for a 1-cm^2 cell, 14.6% for a 3.5-cm^2 unit, and 9.7% for a 0.37-m^2 module (Hedstrom *et al.*, 1993). The

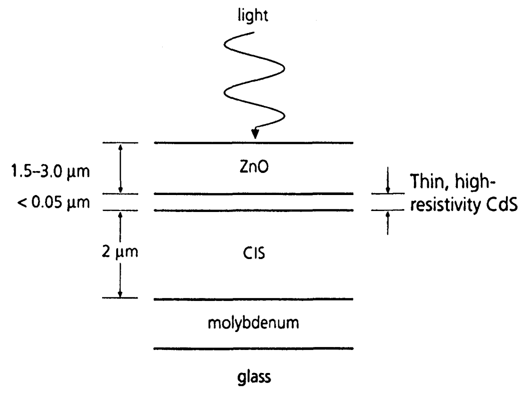

FIG. 14. A copper-indium-diselenide cell.

chalcopyrites can be produced either by vapor deposition or by sputtering of the metals followed by annealing in a selenium-laden gas. The latter mode is the one preferred now. The heterojunction is formed by depositing a 0.05-μm insulating CdS layer followed by a sputtered ZnO transparent conductor a few microns thick.

An intriguing recent discovery is the Cu-deficient ordered vacancy phase (OVP) of *n*-$CuIn_3Se_5$, which forms spontaneously on the last few monolayers of CIS during vapor deposition (Schmid *et al.,* 1993). This new phase has a band gap of 1.5 eV and consequently may serve as the *n*-type layer of an internal heterojunction. Another chalcopyrite, copper indium disulfide, gained attention in recent years. It has the advantage of somewhat higher band gap (1.5 eV) and the replacement of Se by the more benign sulfur. Efficiencies of about 11% and later 12% have been reported for 1-cm^2 cells (Scheer *et al.,* 1993). Dissolution of excess Cu*X*S on the surface of $CuInS_2$ in KCN solution provides the best interface with the overlayer CdS window film (Klenck *et al.,* 1993).

2.2.5 Tandem and Multijunction Cells

In the single-junction PVC, all photons of energy lower than the band gap cannot be absorbed by the semiconductors and are lost. The conversion efficiency of the cells can be increased if layers of two or more photoactive materials are arranged in series in such a way that the photons that were not absorbed in the first layer are absorbed by the following one. Model calculations show (Bennet, 1978) that an AM1 efficiency of 36% can be attained by two junctions and of 40% by three properly tuned junctions. Further increase in the number of junctions leads to only a marginal gain in efficiency. Figure 15 shows a diagram of a typical two-contact arrangement.

The preferred fabrication technique for III–V semiconductor tandem cells is metal-organic chemical vapor deposition. One of the major challenges in fabricating monolithic cells is the Ohmic contact between the two junctions. It is required to be transparent to light, and its lattice must match as far as possible that of the substrate. The preferred arrangement is a layer that acts as a tunnel diode operating under forward bias.

2.3 Concentrators

Up to now we have considered only PV cells that operate under unconcentrated insolation, i.e., not exceeding about 1000 W/m^2. It is clear that if more concentrated sunlight could be used, the area of the PV cells

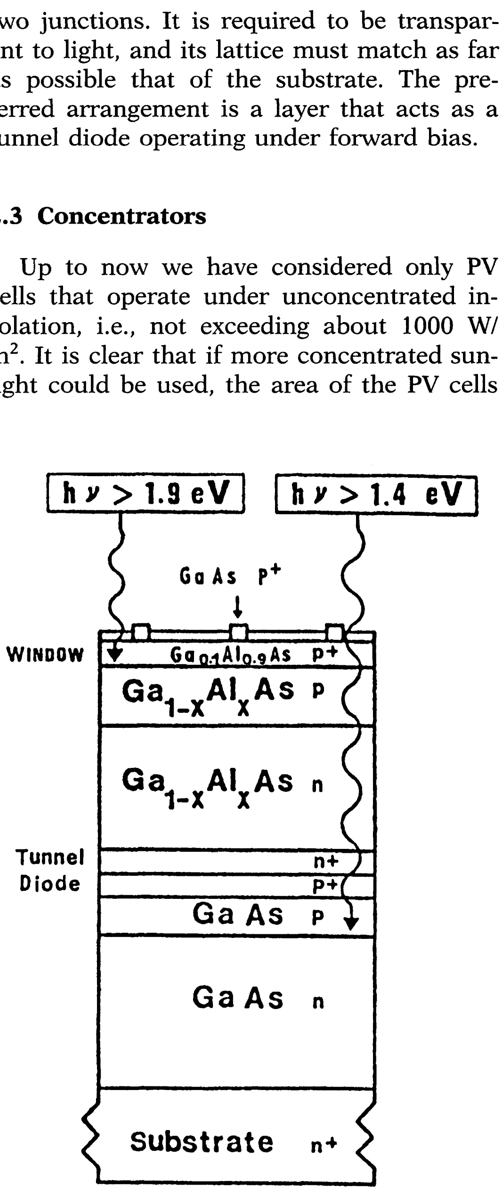

FIG. 15. Schematic diagram of an Al_xGa_{1-x}/GaAs monolithic tandem solar cell with a tunnel diode as intercell Ohmic contact.

required for the same power output would be reduced in inverse proportion. Thus, even a concentration of 100 times, which is considered very modest in solar thermal technologies, will reduce the area of PV cells by a factor of 100 and reduce the cost of cells substantially. However, it will be the cost of the solar collecting and concentrating components that will play a significant role in the economic viability of the plant. Concentrator systems require PV cells that can operate efficiently under intense illumination. Such cells are, indeed, available (Cuevas *et al.*, 1989; Kuryla *et al.*, 1990) at a cost not much greater than cells for unconcentrated sunlight. The economic advantage of operating under concentrated sunlight depends, therefore, on the relative savings on cells as compared to the cost of the solar collector and concentrator components.

A large variety of solar concentrators is available, some designed specifically for PV operation and others through the efforts of the solar thermal development programs. Those designed for PV use are based on Fresnel lenses, while those designed primarily for thermal applications are based on reflectors. The latter are described extensively in Secs. 1 and 3 of this article. With concentrations above about 100, the economic considerations become very similar to those involved in the thermal technology with the replacement of the power-block components used there by photovoltaic power units.

A particularly attractive feature of the use of concentrated sunlight is the possibility of splitting the solar spectrum and illuminating different PV cells with those portions of the spectrum that each cell can use most effectively. A more ambitious proposal involves powering by the same source of concentrated sunlight, at the same time, different conversion technologies, e.g., PV units, lasers, and thermal converters (Krupkin, 1993). Such an arrangement is discussed and illustrated in Sec. 3.

The selection of the various parts of the solar spectrum may be done by means of dichroic filters. These can withstand high light intensities and are relatively inexpensive.

2.4 System Integration

2.4.1 Modules and Arrays The output voltage of an individual cell is low, of the order of 1 V, and the current delivered by it is a few tens of mA/cm^2. A large number of cells have to be connected together in a series and parallel arrangement to obtain the desired output. The hierarchy of these connections starts with the modules that are then connected into arrays that are finally assembled into a system. Because of the large number of units differing slightly in characteristics and operating conditions connected together, and the lack of uniformity in cells of larger area, the efficiency of the assemblies decreases appreciably as we progress along this hierarchy. Thus, the efficiency of a module is considerably lower than that of a laboratory-sized cell, and that of the array smaller still. For this reason we have been careful when quoting efficiencies to mention the size of the unit involved.

Great care has to be exercised in the design and assembly of the final system. In such complex systems provisions have to be made to deal with conditions of off-specification operation of some of the components that may affect seriously the operation of the whole. For example, if one cell of a module loses illumination for some reason, all the string in series with it will be put out of action, but will remain under the voltage of the other modules in parallel with it. This may induce current to flow in the wrong direction through the shadowed cell, heating it ("hot spot") and possibly damaging it permanently. Analogous situations may arise in parallel-connected units. These calamities are avoided by incorporating diodes in strategic points to shunt or prevent the flow of unwanted currents.

2.4.2 Encapsulation and Support A major difference between the laboratory cell and the actual field module is its protection from the weather. The components of an actual photovoltaic power unit are expected to operate, exposed to all weather conditions, for many years, and the economics of the system depends to a large measure on such survival. This is achieved usually through careful encapsulation of the modules.

The assemblies are usually supported on conventional structures, designed, of course, for long outdoor exposures. There is, however, a specific feature of solar assembly support—whether it is stationary or tracking. The simplest design is, of course, the simple fixed support adjusted during construction for maximal solar exposure. Usually this will call for a south-facing (in northern latitudes) or north-facing (in southern latitudes) structure with its normal set at an angle to the vertical equal to the latitude of the location. Then at the equinoxes the panels will receive the maximum insolation, while at other times of the year there will be a reduction of intensity by a factor equal to the cosine of the angle between the sun and the normal to the panel. Such loss can be minimized or avoided entirely if the panel is made to track the apparent motion of the sun. Such double-axis tracking is common in the field of thermal applications of solar energy, and fairly simple and reliable equipment is available. It should be noted that for PV use a far less accurate tracking is sufficient than that needed for the thermal or laser applications (see Secs. 1 and 3). The economic considerations are simple: Is the gain in the annual solar electrical energy produced by a tracking assembly greater than the extra costs of the tracking mechanisms? Concentrating PV systems, of course, always track the sun, and this partly compensates for the fact that they use only the direct normal fraction of the solar radiation.

Obviously the exact relative advantage of one design over the other will be site specific. Comparative studies in various parts of the world showed that, for example, the annual solar radiation per unit area in Rome, Italy, and Denver, Colorado, U.S.A., were the same for the fixed-bed and the concentrating systems, while in Boston, Massachusetts, and Seattle, Washington, U.S.A., the direct normal radiation captured by the concentrating cells was about 80% of the total. Two-axis tracking flat-bed PV assembly will, of course, always receive more total irradiation than the direct normal component.

Another function of the mounting structure is to provide cooling for the PV cells. The efficiency of all PV cells drops rapidly as the operating temperature rises. In fact all the values of efficiencies quoted in this article were those measured at "room" temperature (25 °C). Actual modules operating in the field operate at temperatures slightly higher than ambient, which may often be higher than the nominal value quoted above. This is particularly so with cells operating under concentrated sunlight, and cooling is essential or the cell may even be destroyed. With nonconcentrating cells it is usually sufficient to ensure good thermal contact with the backing structure that may be designed for enhanced radiation. With concentrating cells, forced cooling of the backing must be included in the design, and this, of course, adds to the cost of construction and operation. In advanced designs the hot water obtained from the cooling cycle is used for heating or other household purposes.

2.4.3 Power Conditioning The electrical power produced by PV cells and therefore PV panels and systems is low-voltage direct current. Apart from certain very specific local applications, such power does not match the rest of the local power network and the equipment connected to it. In addition, the raw power produced by the PV assembly is not steady because the solar illumination is not steady even during an apparently clear and cloudless day. So some power conditioning is an essential feature of a PV power supply. The details will, again, depend on the specific site and conditions. The simplest arrangement, for small and independent units, will call for some storage batteries to stabilize the output and an inverter to deliver ac power at the voltage and frequency to match local usage.

For larger units, particularly those connected to networks, more sophisticated controls are necessary. These may include synchronization, various safety interlocks, and power regulators. The last of these are necessary to ensure that the PV units are operated at their maximum power points even when the illumination and load are changing.

2.5 Storage

The problem of energy storage in solar plants is a general one and not specific to photovoltaic systems (see Sec. 1.5). Since the sun does not shine continuously, and in addition disturbances due to weather conditions are unavoidable, solar energy plants suffer a severe handicap. The magnitude of this problem and its possible solutions depend very much on local conditions and policies. This issue is discussed at some length in Sec. 1.5 of this article, and therefore only those aspects more specific to PV systems will be described here.

For small stand-alone units (in the kilowatt range), storage batteries are an adequate solution. For larger systems (in the megawatt range), this solution is not practical. If, however, the solar unit is part of a utility network, then the question of storage depends on the match between the load curve of the network and the insolation curve for the location of the solar plant. If the match is reasonable, then the solar plant can be used as a peaking plant only, supplying power only at times of maximum demand when it is worth most, and storage can be dispensed with. However, if larger-scale utility use or stand-alone operation is contemplated, the problem of storage becomes particularly severe.

The options open to PV systems are more limited than those for thermal systems. The electricity supply industry has been facing this problem for many decades and has evolved some solutions, and these of course are open also to the PV industry and will be discussed further here. However, there is one channel that is particularly suitable to the PV industry. It can take advantage of the fact that the primary product is dc electricity and is thus eminently suitable for electrochemical processes. In particular, the electrolysis of water to produce hydrogen is an attractive route for storage (and use). Hydrogen is a valuable material and is seriously being considered as an environmentally benign replacement of fossil fuel for transportation and other uses. At the same time hydrogen, in a fuel cell, is an excellent source of electricity. Thus, solar energy from PV plants can be stored as hydrogen indefinitely, can be transported in pipe lines, and can be converted back to electricity at will via fuel cells. A joint German–Saudi Arabia 300-kW solar PV-hydrogen experiment (Hysolar) has been running for a number of years (see also Sec. 1.4.2).

As was already discussed in connection with the thermal applications (see Sec. 1.6), the optimal choice of a process and its detailed design depends on many factors in addition to simple cost, most of which are equally valid for photovoltaic systems, and will not be discussed here. They are site specific (in the general sense) and include questions of government policies and environmental regulation.

3. PHOTOCHEMISTRY AND SOLAR LASERS

In this section we deal with the conversion of sunlight to chemical energy without converting it first to heat or electricity. Although solar photochemistry is the basis of all life on earth, and has been studied from the early days of chemistry, up to the present time there have been no major artificial (i.e., nonbiological) applications to industry or the energy economy.

3.1 Applied Solar Photochemistry

Some recent attempts to use solar radiation to detoxify wastes and treat contaminated water may be the first examples of a new trend. The driving force behind this new field is the increasing awareness of the magnitude of environmental damage caused by continuing pollution of the water resources and atmosphere by human activity. The permissible concentrations of most pollutants are so low that the classical technologies are inapplicable or too expensive. The hope that photochemical processes, driven by the plentiful solar radiation, can be used for the degradation of these pollutants has led to a large number of experiments and tests in various parts of the world. Most of these are still in the laboratories (using solar simulators) and are too numerous to be detailed here. A recent review appears in a book by Ollis (1993). Reports of experiments in progress can be found in almost all the international conferences on solar energy, some of which are quoted in the bibliography at the end of this article. In particular we would like to mention the reviews by Larson (1991) and Mills (1993), as well as the field experiment reported by Mehos (1993).

Two different methods are studied. The first method is based on the activation of semiconductors like TiO_2 using the UV fraction of solar radiation and subsequent interaction with organic contaminants leading to destruction of the latter. The main drawback of this method is that only a very small fraction of solar radiation can be utilized.

The other method is based on the excitation of oxygen molecules dissolved in water by energy transfer from an organic dye that absorbs visible solar radiation (Acher, 1990). A more significant fraction of the solar spectrum is involved in this process. The oxygen molecules are excited to their singlet state and can oxidize various organic contaminants present in the solution. For complete detoxification, solar radiation at concentration of more than 1000 suns may be necessary.

The most efficient photochemical processes are those that are induced by monochromatic radiation tuned to the specific molecular species involved. The classical way of obtaining approximately monochromatic light from solar radiation is either by means of absorption filters or by diffraction. A more recently developed method is through the use of solar-driven lasers.

3.2 Solar-Driven Lasers

A detailed description of the physics of lasers is found in LASER PHYSICS and will not be repeated here. A common way of exciting solid-state lasers today is by illumination with a powerful beam of light, usually obtained from an electrically powered source. It is but a conceptually short step to replace the electrically driven light source by intense solar radiation and thus obtain a solar-driven laser. In practice it is somewhat more complicated, and in this section we describe the special features of solar lasers and report on the state of the art today. The motivation for developing such devices is the elimination of the need for a power supply (particularly important in space applications) and the possibility of high overall efficiency of conversion of sunlight to monochromatic and coherent radiation.

3.2.1 Materials for Solar Lasers In Table 2 we list some of the laser materials suitable for solar driving together with their thresholds, spectral absorption bands, and other parameters. In Fig. 16 we show the solar spectrum and note the overlap with the absorption bands of some of the materials listed in Table 2. We note that all of them are garnets possessing high refractive index and high thermal conductivity. They all have absorption bands in the solar spectrum, a high refractive index, and good heat conductance. Yttrium aluminum garnet (YAG) doped with neodymium is a common combination for many laser applications. However, the overlap of the absorption spectrum of the doped crystal with the solar spectrum is rather poor, leading to low efficiency even with chromium as codopant. Gadolinium gallium garnets (GGG) and gadolinium scandium gallium garnets (GSGG) give very high energy-transfer efficiency when codoped with chromium but suffer from poor thermal properties and tendency to fracture. These difficulties may be minimized by spectrally selected illumination and special geometric design of the crystal. It is then possible to double the efficiency as compared with Nd:YAG (Noter *et al.*, 1988).

Table 2. Properties of some laser materials for solar pumping.

Laser material	Center of absorption band (nm)	Lasing threshold (w/cm^2)	Laser wavelength (nm)
NdYAG[a]	540 580 740 800 880	30	1064
NdCrGSGG[b]	460 640 740 810 880	50	1054
Alexandrite	410	200	750–850

[a]YAG stands for yttrium aluminum garnet.
[b]GSGG stands for gadolinium scandium gallium garnet.

Sapphire doped with Ti^{+3} has an excellent thermal conductivity, but the threshold of brightness for laser pumping is so high that it is practically impossible to remove all the heat that is associated with continuous pumping. It cannot, therefore, be considered a good candidate for a cw solar laser (Krupkin, 1993).

3.2.2 Nonimaging Concentrators As can be seen in Table 2, some of the thresholds for lasing action are within the light intensities easily achievable by the concentration technologies developed for thermal application and are described in Sec. 1. But for some of the more interesting materials, the thresholds are very high and require light intensities that cannot be obtained by these techniques. However, as Winston has shown (Welford, 1989), if in the design of the concentrators the requirement of faithful imag-

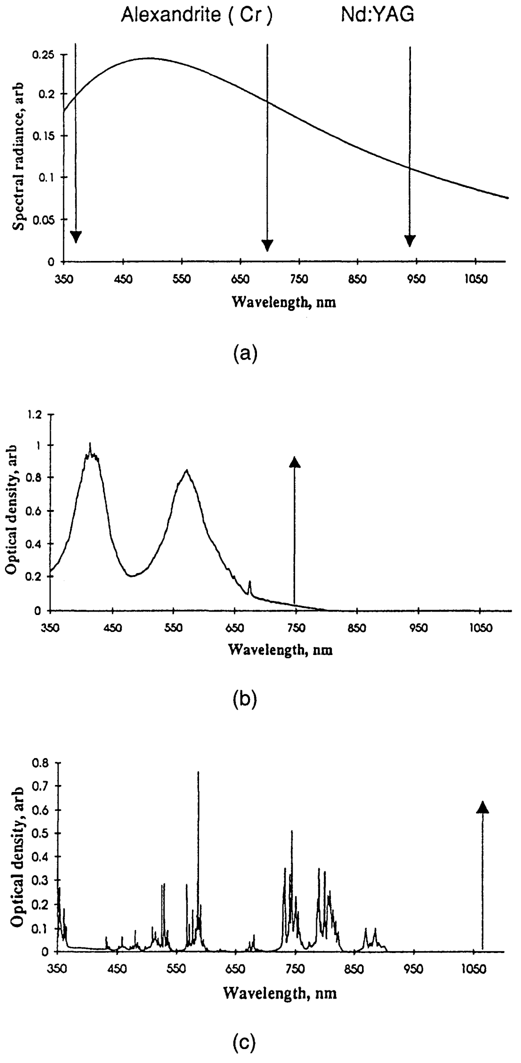

(a)

(b)

(c)

FIG. 16. Solar spectrum and absorption bands of various laser materials. (a) Solar spectrum outside the atmosphere; (b) absorption spectrum of alexandrite; (c) absorption spectrum of Nd:YAG.

ing is waived, it is possible to attain much higher light intensities.

A very common design of nonimaging concentrators is the compound parabolic concentrator (CPC) shown in Figs. 17 and 18. The three-dimensional shape is obtained by rotating a section of a parabola with its axis tilted an angle θ to the axis of revolution. The degree of concentration provided by this device depends on the ratio of the entrance and exit angular apertures according to the following relations: $C = \sin \alpha'/\sin \alpha$ for the two-dimensional case (trough), and $C = \sin^2\alpha'/\sin^2\alpha$ for the three-dimensional case, where α is the entrance aperture half-angle and α' is the exit aperture half-angle. One can see that for the value of the angle $\alpha' = 90°$, the concentration is maximal and equals $1/\sin \alpha$ and $1/\sin^2\alpha$ for the two- and three-dimensional cases, respectively. The rays then leave the exit into a 2π solid angle.

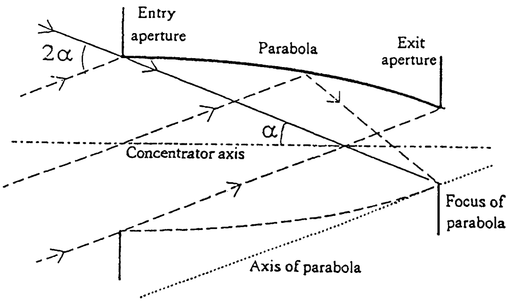

FIG. 17. A compound parabolic concentrator.

For very small entrance angles such as, for example, that of the solar disc (0.266°), the concentration ratio is $46\,200n^2$ (for a 3D device), but the concentrator will be very long and unwieldy. Here n is the refractive index of the medium through which the light propagates within the concentrator. The main use of such CPCs is, however, as secondary concentrators receiving the output of dishes, troughs, or heliostats. In these cases the entrance half-angle will be large and the concentration ratio correspondingly smaller. Thus, for a 45° entrance half-angle the ideal maximum concentration of the CPC will be only 2, but the overall concentration of the assembly can still be several thousand.

When the object to be illuminated is of cylindrical shape (rod or tube), as in most solar laser applications, the configuration shown in Fig. 18 is appropriate (it is a version of 2D concentrators).

When this technology is used with solid-state lasers, the brightness of the pump light can be further increased by refraction. The upper limit can then be increased by n^2 (for a 3D system), where n is the refractive index of the CPC material. In this case, the concentrator itself must be made from transparent dielectric material, relying on total internal reflections for almost lossless operation. Advantage of this circumstance can only be

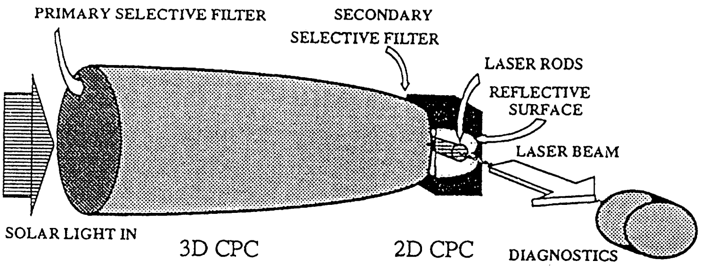

FIG. 18. A solar pumped laser (WIS).

taken if the laser itself is placed in close optical contact with the concentrator and its refractive index is not too different from that of the latter.

3.2.3 Heat Removal and Efficiency

Having achieved the high solar fluxes into a solid-state laser, the next major problem is to remove the excess heat generated within it. While about 50% of the light absorbed within the spectral band of the laser material is converted to radiation and leaves the crystal, the remaining light that is absorbed by the solid is converted to heat and has to be removed. Apart from the obvious danger to the crystal by fracture, there is also a very deleterious optical effect. Temperature gradients within the crystal, by causing refractive index variations, can lead to birefrigence and lensing, which interfere with the optics of the laser cavity. Effective cooling is therefore a prerequisite of high-powered solar lasers, with the added complication that the cooling system is part of the optical cavity. In many cases it is the heat-removal capacity that sets the limit to the laser power output. A diagram of a complete solar laser is shown in Fig. 18.

High-power solar lasers can be obtained by bundling together several rod lasers in one solar concentrator (see Fig. 19) in the same optical cavity. Power output of 500 W cw has been obtained from such a device with four 6.3-mm Nd-YAG rods (Thompson, 1992, 1993).

Another way of increasing the power output of lasers, when the rate of heat removal is the limiting factor, is to remove from the illuminating light part of the short-wavelength portion of the spectrum, which contributes only slightly to the lasing. In this way the thermal load is reduced to the minimum. This concept of splitting the solar spectrum and then using each spectral band separately can be generalized to other systems where solar radiation is converted directly to other energy forms. For example, one may consider a more complex system where part of the spectrum is used for lasers, part for photovoltaic conversion, and the rest for thermal applications (Krupkin, 1993). Spectral splitting can also be applied to simultaneous pumping of several lasers with different absorption bands and thus obtain a much higher conversion efficiency. An example of such a pair may be alexandrite and Nd-YAG. See Fig. 20.

The spectral splitting may be obtained with the aid of dielectric filters. These can operate with highly concentrated light without cooling and can be integrated in the optical design of the collection and concentration devices (Krupkin, 1993). Such schemes may be very attractive economically, since 1 m^2 can provide 10 MW of solar light.

3.2.4 The Wavelength of Solar Lasers

All available optically pumped lasers are active in the near IR. The holmium laser emits

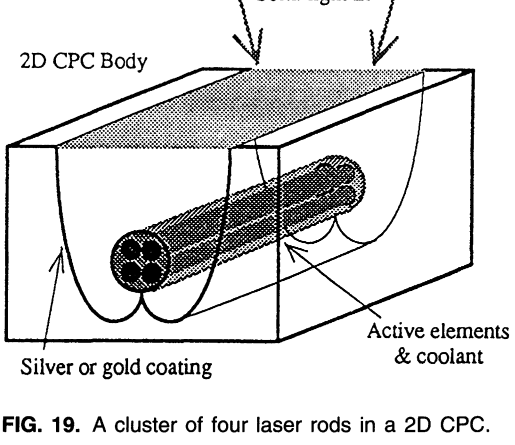

FIG. 19. A cluster of four laser rods in a 2D CPC.

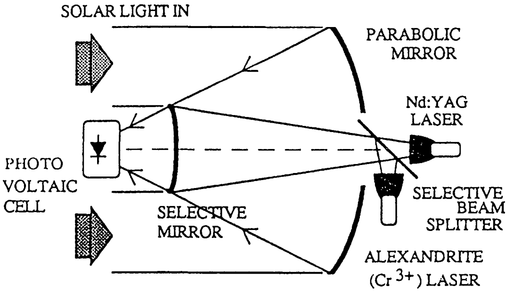

FIG. 20. An example of the use of split solar spectrum to drive two lasers and a photovoltaic cell assembly.

at 2 μm and the neodymium laser at 1.06 μm, and alexandrite is tunable over the range 0.75–0.9 μm. Many applications require shorter wavelength and higher pulse power. Thus, most photochemical reactions require UV light and tunability. Solar lasers are not different from any other continuous-wave lasers, and standard laser techniques for frequency manipulation may be applied to them. Thus, *Q* switching to obtain giant pulses, and the use of nonlinear optical materials to get tunability and frequency up-conversion, have been demonstrated with solar lasers (Arashi, 1993). But since they are not solar-specific techniques, they will not be discussed further here.

3.2.5 Applications of Solar Lasers Solar lasers are still in the early stages of development, and no commercial application has been realized so far. For space applications solar lasers have an obvious advantage due to their independence from other power sources. Even the presently available lasers have enough power for missions such as remote sensing, weather prediction, and communications. Probably the most attractive terrestrial application of powerful tunable solar lasers is in industrial photochemistry, but these have not been developed yet.

Works Cited

Acher, A., Fischer, E., Zelinger, R., Manor, Y. (1990), *Water Res.* **24,** 837–843.

Arashi, H., Kameda, Y. (1993), *Solar Energy* **50,** 447–451.

Bennet, A., Olsen, L. C. (1978), in: *Conference Record, 13th IEEE Photovoltaic Specialists Conference,* New York: IEEE, pp. 860–868.

Buck, R., Abele, M., Bauer, H., Seitz, A., Tamme, R. (1994), in: D. E. Klett, R. E. Hogan, T. Tanaka (Eds.), *Solar Engineering, 1994,* New York: ASME, pp. 73–78.

Cuevas, A., Sinton, R. A., Midkiff, N. E., Swanson, R. M. (1990), *IEEE Electron Devices Lett.* **11,** 6–8.

Chu, T. L., Chu, S. S., Britt, J., Ferekides, C., Wang, C., Wu, C. Q. (1992), *IEEE Electron Devices Lett.* **13,** 303–304.

Emery, K. (1996), in: D. Faiman (Ed.), *Proceedings of the 7th Sede Boqer Symposium on Solar Electricity Production,* Beer Sheva, Israel: Jacob Blaustein Institute for Desert Research, Ben Gurion University of the Negev, pp. 51–63.

Epstein, M., Spiewak, I. (1994), in: *Proceedings of the 7th International Symposium on Solar Thermal Concentrating Technologies, Moscow, September 26–30, 1994,* Moscow: Institute for High Temperature of the Russian Academy of Sciences (IVTAN), pp. 958–971.

Epstein, M., Spiewak, I., Funken, K.-H., Ortner, J. (1994), in: D. E. Klett, R. E. Hogan, T. Tanaka (Eds.), *Solar Engineering, 1994,* New York: ASME, pp. 79–91.

Fletcher, E. A., Moen, R. (1977), *Science* **197,** 1050.

Fletcher, E. A. (1984), *J. Minn. Acad. Sci.* **49,** 30.

Fletcher, E. A. (1991), in: M. E. Arden, S. M. A. Burley, M. Coleman (Eds.), *1991 Solar World Congress,* Tarrytown, NY: Pergamon, p. 790.

Gluck, A., Tamme, R., Streuber, C. (1993), in: M. Macias (Ed.), *Proceedings of the 6th International Symposium on Solar Thermal Concentrating Technologies, Mojacar (1992),* Madrid: Editorial CIEMAT, pp. 445–457.

Harth, R. E., Boltendahl, U. (1981), *Interdisciplinary Sci. Rev.* **6,** 221–228.

Hedström, J., Ohlsen, H., Bodegård, M., Kylmer, A., Stolt, L., Hariskos, D., Ruckh, M., Schock, H. W. (1993), in: R. J. Schwartz (Ed.), *Proceedings of the IEEE 23rd Photovoltaic Specialists Conference, Louisville, Kentucky,* New York: IEEE, p. 364.

Karni, J. (1994), in: *Proceedings of the 7th International Symposium on Solar Thermal Concentrating Technologies, Moscow, September 26–30, 1994,* Moscow: Institute for High Temperature of the Russian Academy of Sciences (IVTAN), pp. 838–848.

Klenck, R., Walther, T., Schock, H. W., Cahen, D. (1993), *Adv. Mater.* **5,** 114.

Kreith, F., Krieder, J. (1978), *Principles of Solar Engineering,* New York: McGraw-Hill.

Krupkin, V., Kagan, Y., Yogev, A. (1993), in: R. L. Holman, R. Winston (Eds.), *Non-Imaging Optics: Maximum Efficiency Light Transfer II,* SPIE Proceedings Vol. 2016, Bellingham, WA: SPIE, pp. 50–60.

Kuhn, P., Enrensberger, K., Frei, A., Oswald,

H. P. (1994), in: *Proceedings of the 7th International Symposium on Solar Thermal Concentrating Technologies, Moscow, September 26–30, 1994,* Moscow: Institute for High Temperature of the Russian Academy of Sciences (IVTAN), pp. 614–626.

Kuryla, M. S., Kaminar, N. R., MacMillan, H. F., Ristow, M. L., Virshup, G. F., Klausmeier-Brown, M. R., Partian, L. D. (1990), in: *Conference Record of the 21st IEEE Photovoltaic Specialists Conference,* Vol 2, New York: IEEE, pp. 1142–1146.

Larson, R. A., Marley, K. A., Schlauch, M. B. (1991), in: D. W. Tedder, F. G. Pohland (Eds.), *Emerging Technologies in Hazardous Waste Management,* ACS Symposium Series No. 468, Washington, DC: American Chemical Society, pp. 66–82.

Levine, J. D., Hotchkiss, G. B., Hammerbacher, M. D. (1991), in: *Conference Record, 22nd IEEE Photovoltaic Specialists Conference,* New York: IEEE, pp. 1045–1048.

Levy, M., Levitan, R., Rosin, H., Rubin, R. (1993), *Solar Energy* **50,** 179–189.

Mehos, M., Turchi, C. (1993), in: M. Macias (Ed.), *Proceedings of the 6th International Symposium on Solar Thermal Concentrating Technologies, Mojacar (1992),* Madrid: Editorial CIEMAT, pp. 1107–1122.

Mills, A., Davies, R. H., Worsley, D. (1993), *Chem. Soc. Rev.* **22,** 417–425.

Noter, Y., Oron, M., Shwartz, J., Weksler, M., Yogev, A. (1990), in: B. P. Gupta, W. H. Traugott (Eds.), *Proceedings of the 4th International Workshop on Solar Thermal Central Receiver Systems,* Bristol, PA: Hemisphere Publishing Corp., pp. 587–596.

Ollis, D. F., Al-Ekabi, H. (1993), *Photocatalytic Purification and Treatment of Water and Air,* Amsterdam: Elsevier.

Scheer, R., Walther, T., Schock, H. W., Fearheiley, M. L., Lewerenz, H. J. (1993), *Appl. Phys. Lett.* **63,** 3294–3296.

Schmid, D., Ruckh, M., Grunwald, F., Schock, H. W. (1993), *J. Appl. Phys.* **73,** 2902–2909.

Thompson, G., Krupkin, V., Yogev, A., Oron, M. (1992), in: M. Macias (Ed.), *Proceedings of the 6th International Symposium on Solar Thermal Concentrating Technologies, Mojacar (1992),* Madrid: Editorial CIEMAT, pp. 553–563.

Thompson, G., Krupkin, V., Oron, M., Yogev, A. (1993), in: M. Oron, I. Shladov, Y. Weissman (Eds.), *8th Meeting on Optical Engineering in Israel: Optical Engineering and Remote Sensing,* SPIE Proceedings Vol. 1971, pp. 400–407.

Welford, W. T., Winston, R. (1989), *High Collection Nonimaging Optics,* New York: Academic.

Zweibel, K. (1993), *Am. Sci.* **81,** 362–369.

Further Reading

Winter, C.-J., Sizmann, R. L., Vant-Hull, L. L. (Eds.) (1991), *Solar Power Plants,* New York: Springer-Verlag.

Möllers, J. (1993), *Semiconductors for Solar Cells,* Boston: Artech House.

Becker, M., Klimas, P. C. (Eds.) (1993), *Second Generation Central Receiver Technologies,* Karlsruhe: Verlag C. F. Muller.

Green, M. A. (1987), *High Efficiency Solar Cells,* Switzerland: Trans Tech Publications.

Winter, C.-J., Nitsch, J. (Eds.) (1988), *Hydrogen as an Energy Carrier,* New York: Springer-Verlag.

At the present time there is an intense international activity in research and development in all areas related to the applications of solar energy. Progress is reported in regular international meetings and symposia (annual or biannual) and published in their proceedings. These publications present a valuable source of updated information on the status of the various topics. Prominent among the organizations sponsoring such conferences are the IEA, IEEE, OSA, and ASME.

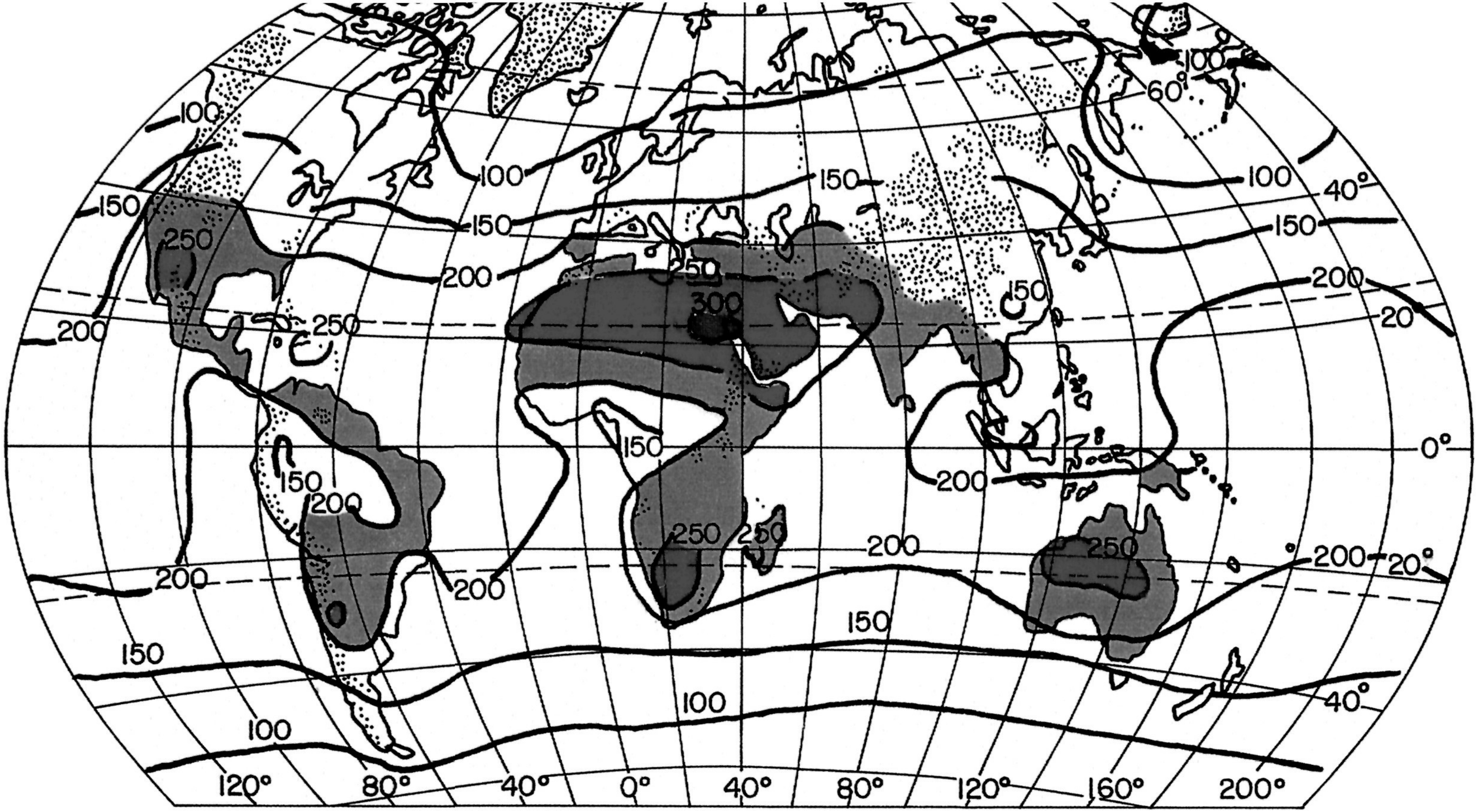
100
100
100
60°
150
250
40°
150
150
150
200
200
200
250
300
150
250
200
20°
200
150
150
0°
150
200
200
250
250
200
200
250
200
20°
200
200
150
150
100
100
40°
120°
80°
40°
0°
40°
80°
120°
160°
200°

Vol. 18 – PLATE 1. World solar insulation map (W/m² over 24 h, annual average) (source: Kreith and Krieder, 1978).

SOLAR RADIATION

D. F. Smart and M. A. Shea, *Space Physics Division, Geophysics Directorate, Phillips Laboratory, Hanscom Air Force Base, Bedford, Massachusetts, U.S.A.*

Introduction 394
1. **The Solar Structure** 394
2. **Solar Neutrino Emission** 396
3. **Solar Electromagnetic Radiation** 397
3.1 Solar Irradiance from the Quiet Sun 397
3.1.1 The Solar Radio Spectrum 398
3.1.2 The Solar Far-Infrared Spectrum 399
3.1.3 The Solar Optical and Infrared Spectrum 400
3.1.4 The Solar Ultraviolet Spectrum 400
3.1.4.1 The Lyman-Alpha Flux 401
3.1.5 The Solar Extreme-Ultraviolet Spectrum 401
3.1.6 The Solar X-Ray Flux 401
4. **Solar Particle Emission from the Quiet Sun** 401
4.1 Solar Plasma Emissions 401
4.1.1 The Solar Wind 402
4.1.1.1 Solar-Wind Composition 402
4.1.2 The Interplanetary Magnetic Field 403
5. **Solar Electromagnetic Radiation from the Active Sun** 405
5.1 Solar Activity and Solar Cycles 405
5.1.1 Sunspot Cycles 405
5.2 Solar Flares 406
5.2.1 Solar-Flare Reporting 408
5.2.1.1 Optical Solar Flares 408
5.2.1.2 Solar X-Ray Flares 408
5.2.2 Solar-Flare Emissions 409
5.2.2.1 Solar-Flare Radio Emissions .. 409
5.2.2.2 Solar-Flare Gamma-Ray Emission 411
5.2.2.3 Solar-Flare Hard X-Ray Emission 411
5.2.2.4 Solar-Flare EUV and Soft X-Ray Emissions 412
6. **Solar Particle Radiation from the Active Sun** 412
6.1 Solar Plasma Emissions 412
6.1.1 Coronal Mass Ejections 412
6.2 Solar Energetic Particles 413
6.2.1 Propagation Effects in the Interplanetary Medium 414
6.2.2 Solar-Electron Events 416
6.2.3 Solar-Proton Events 416
6.2.4 Solar-Particle Event Composition 419
6.2.4.1 Large Solar-Particle Events 419
6.2.4.2 Small and Iron-Rich Solar-Particle Events 419
6.2.4.3 Charge State of Solar Particles 419
7. **Terrestrial Effects from Transient Solar Emissions** ... 421
7.1 Communication Effects 421
7.1.1 Polar Communication Effects 421
7.1.2 Aurora 422
7.2 Geomagnetic-Field Variations 422
7.3 Atmospheric-Density Effects on Satellite Operations 422
7.4 Solar-Particle Radiation Effects 423
7.4.1 Radiation-Dose Effects on Solar-Cell Power Systems 423
7.4.2 Soft Errors in Electronics 423
7.4.3 Solar-Particle Radiation Effects on Human Space Flight 423
8. **Summary** 424
Glossary 424
Works Cited 428
Further Reading 429

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

This article is a general summary of the photonic and particulate emissions from the Sun and describes the effects that these emissions have on our technology. The solar emissions define the character of the entire heliosphere and control the environment that makes life on Earth possible. The general organization of this report starts with the solar structure and electromagnetic radiation, continues with the particulate radiation, and concludes with terrestrial effects resulting from transients in solar emissions.

The Sun is a dynamic star currently having an approximate 11-yr activity cycle. The discussion of the various solar emissions has been separated into emissions from the "quiet Sun" and those from the "active Sun." Quantities that are "well measured" have the briefest description. Quantities that are not well understood nor well measured and are variable have a more detailed discussion in an attempt to define our current understanding.

The effect of solar emissions on the terrestrial environment is worldwide in scope. Research results and technological advances since the advent of the space era in 1957 clearly emphasize the interrelationships of what were once considered separate and independent phenomena. This report discusses both the various interrelationships where they are known and those areas where much is yet to be learned.

1. THE SOLAR STRUCTURE

The Sun is a relatively small, faint, cool star at an average distance of 1.496×10^8 km from the Earth. The composition of the outer layers by mass is approximately 73% H, 25% He, and 2% other elements. This rotating mass ($\sim 1.971 \times 10^{30}$ kg) of hot plasma is in dynamic equilibrium between the gravitational energy of its mass and the energy of its internal fusion heat source. The key word is *dynamic*. The Sun is a very complex and structured body. There are fluctuations in the energy output of the Sun on time scales ranging from seconds to billions of years. The radiative output of the Sun has increased approximately 30% since the Sun began its hydrogen burning sequence over three billion years ago. The Earth's climatic record suggests changes in the radiative output on the time scale of hundreds or thousands of years. The radiative output of the Sun has been observed to vary about 0.3% during a solar cycle since space measurements of the solar irradiance began in 1978.

Models of the Sun are generally organized into various regions where specific processes occur. Each of these regions is often subdivided into smaller specific regions where one particular process dominates. The "standard solar model" (see Cox *et al.*, 1991, for a detailed description) is constantly being updated and improved. For this article the Sun's structure has been simplified and described (see Fig. 1) from the core outward as follows:

1. The core of the Sun has several definitions. The heart, or inner core, where the temperature and density are sufficient to support fusion reactions that are the source of solar energy and the source of the solar neutrino flux, extends to ~0.15 of the solar radius (R_0). The core of the Sun, which contains 96% of the solar mass, extends to $\frac{1}{4}$ of the solar radius.
2. The radiative zone (sometimes called intermediate interior) is the region where radiative transport is the dominant heat-transfer process. This region extends from the core boundary to a poorly defined and model-dependent boundary between about (0.7 and 0.86)R_0. The radiation that begins in the core as high-energy gamma rays is degraded by suppressive absorptions and re-emissions until it finally emerges as visible radiation. Away from the core, the temperature, density, and pressure decrease until free electrons can be trapped by atoms in bound states. This transition in material structure causes an increase in opacity, and so radiation becomes a less efficient method of transporting energy, resulting in a large temperature gradient.
3. The convective zone, where convection becomes the dominant process for transporting heat and magnetic flux, extends from the radiative zone to the photosphere. In the convective zone, there are at least three or four different scales of convection.

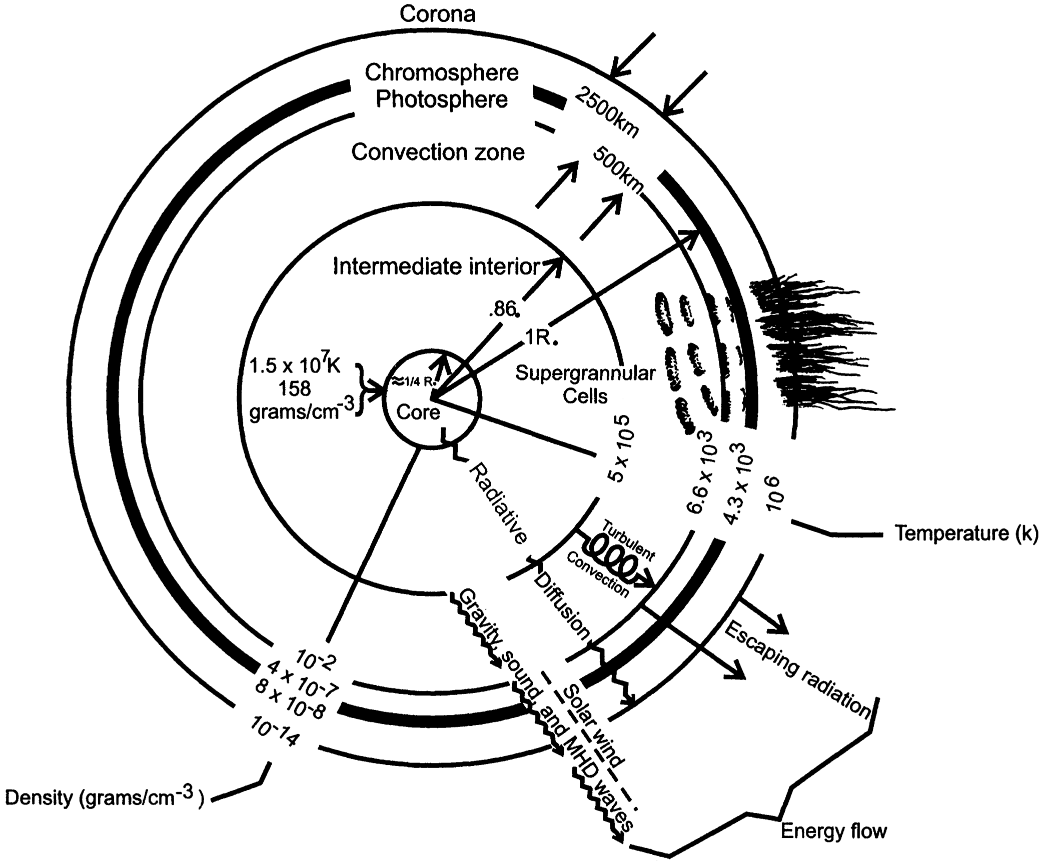
Corona
Chromosphere
Photosphere
Convection zone
2500km
500km
Intermediate interior
.86
1R.
1.5 x 10^{7}K
158
grams/cm^{-3}
≈1/4 R.
Core
Supergrannular
Cells
5 x 10^5
6.6 x 10^3
4.3 x 10^3
10^6
Temperature (k)
Radiative
Diffusion
Turbulent
Convection
Gravity, sound, and MHD waves
Solar wind
Escaping radiation
Energy flow
10^{-2}
4 x 10^{-7}
8 x 10^{-8}
10^{-14}
Density (grams/cm^{-3})

FIG. 1. Idealized general solar structure and modes of energy outflow. The features shown are not to scale.

4. The photosphere, which we observe as the "surface" of the Sun, is where most of the solar energy radiates into space. The photospheric diameter is well defined at the limb of the Sun and is considered to be the diameter of the Sun, having a usually accepted value of 1.391 × 10^6 km. This photosphere is not a "surface" but actually has a thickness (depending on the wavelength and position) of several hundred kilometers. A formal definition of the extent of the photosphere is the depth in the solar atmosphere from which the probability of a photon escaping is 1/*e*. The photosphere "surface" is not uniform but is highly structured, with granulations of the order of 1000 km and supergranulations, which have vertical and horizontal motion apparent in solar photographs. (See Zirin, 1988, for illustrative examples.) Granules have a mean lifetime of about 8 min, while supergranules may have a lifetime of 1–2 days. There are also a number of observable convective motions and oscillation modes. The most prominent oscillation mode has a period of 300 s.
5. The chromosphere is a complex layer of the solar atmosphere above the photosphere and beneath the transition region and the corona. The chromospheric network has a pattern similar to the granulation pattern observable in the photosphere and is an important source of extreme-ultraviolet (EUV) emission. The chromospheric network is also the source of spicules, which appear as brilliant flames of gas that project upward and have lifetimes of 2–5 min. The chromosphere is the source of the strongest lines in the solar spectrum, including the Balmer alpha line of hydrogen (Hα) and the

H and *K* lines of calcium, and is the source of the red (chromium) color often seen during a total solar eclipse. The principal absorption lines in the solar spectrum, called the Fraunhofer lines, originate primarily in the chromosphere. The temperature of the solar atmosphere reaches a minimum in the lowest part of the chromosphere and then increases approximately monotonically with height. The upper boundary of the chromosphere is about 2500 km above the photosphere and is defined as the region where the temperature becomes too high ($\sim 10^4$ K) for hydrogen to give observable line emission. The rise in chromospheric temperature with height is usually explained as being the result of nonradiative energy input via mechanical energy generated in the convection zone. It is presumed that acoustical waves propagate outward and form shocks in the low-density corona, generating collisions between the atoms in the chromosphere. The low densities present also permit the magnetic fields to carry energy in the form of waves.

6. The transition region (an extremely complex region, where the physical processes are poorly understood) is a relatively narrow region (~ 100 km) between the chromosphere and the corona where the temperature rapidly increases to about 10^6 K.
7. The corona is a complex, structured region of very hot tenuous gas above the chromosphere extending into space. It has been further subdivided into three components by its emission characteristics: the *E* corona, where most of the line emission occurs; the *K* corona, where the emission is essentially continuum; and the *F* corona, where lightly broadened Fraunhofer lines are the identifying signatures. The corona is dominated by magnetically induced motions, sudden releases of energy, and explosive expansion into space. The corona is heated from below by mechanical, electrical, and magnetic dissipation. Favorite candidates are shock waves produced by the mechanical motion of upward surging granules, propagating waves induced by shocks or associated with the spicules, and Alfvén waves propagating upward from the turbulent photosphere along magnetic lines of force. The average particle's total thermal energy ($\frac{5}{2}kT$) is 0.8 keV, considering fully ionized hydrogen at 2×10^6 K. The poorly understood nonthermal energy input into the corona results in some of the particles acquiring sufficient energy to overcome the gravitational potential energy (2.0 keV per particle at $1R_0$). These particles escape into interplanetary space as the solar wind. The solar corona viewed from space in soft x-ray wavelengths reveals an inner corona whose structure is dominated by the solar magnetic fields.

The solar emissions observable at the Earth reveal many surprises and define the limits of our knowledge and ignorance. The solar neutrino flux measured at the Earth is about a factor of 3 less than expected from our knowledge of fusion reactions and our estimates of the interior of the Sun. The tenuous outer layers of the Sun are much hotter (coronal temperatures range between one and two million degrees) than expected from the photospheric temperature. The majority of the photonic emission of the Sun comes primarily from the photosphere, with absorption and additional emission in the chromosphere and corona. The majority of the radio and x-ray emission comes from the solar corona.

2. SOLAR NEUTRINO EMISSION

Neutrinos are produced by the fusion reactions deduced to occur in the core of the Sun. The standard solar models postulate a proton–proton (*p-p*) chain of reactions, as tabulated in Table 1. The most important neutrino-production processes are the chains noted by (*a*), (b_1), and (b_2).

In the "standard solar model," the product of these *p-p* reactions and the additional neutrino flux from the carbon–nitrogen cycle reactions should result in a neutrino flux at the Earth's orbit of 7.9 ± 0.9 SNU (solar neutrino unit). The SNU definition is a result of using the reaction $^{37}\mathrm{Cl} + \nu \rightarrow {}^{37}\mathrm{Ar} + e^-$ for neutrino detection where one SNU = 10^{-36} solar neutrino captures per second per target atom. The 20-yr average flux observed by the ^{37}Cl neutrino capture experiment is 2.1 ± 0.2 SNU. The more recent solar neutrino experiments in Europe and Japan designed to measure solar neutrinos at differ-

Table 1. The standard solar model: most important *p-p* reaction chains.

Reaction	Energy (MeV)	
$4{}^1\mathrm{H} \Rightarrow {}^4\mathrm{He} + 2e^+ + 2\nu$	26.731	
${}^1\mathrm{H} + e^- + {}^1\mathrm{H} \Rightarrow {}^2\mathrm{D} + \nu$	1.442	$E_\nu = 1.442$ MeV
${}^1\mathrm{H} + {}^1\mathrm{H} \Rightarrow {}^2\mathrm{D} + e^+ + \nu$	1.442	$E_{\nu(\max)} = 0.420$ MeV
${}^2\mathrm{D} + {}^1\mathrm{H} \Rightarrow {}^3\mathrm{He} + \gamma$	5.493	
(a) ${}^3\mathrm{He} + {}^3\mathrm{He} \Rightarrow {}^4\mathrm{He} + 2{}^1\mathrm{H}$	12.859	
(b) ${}^3\mathrm{He} + {}^4\mathrm{He} \Rightarrow {}^7\mathrm{Be} + \gamma$	1.587	
(b_1) ${}^7\mathrm{Be} + e^- \Rightarrow {}^7\mathrm{Li} + \nu$	0.862	$E_\nu = 0.862$ MeV (89.5%) $E_\nu = 0.384$ MeV (10.5%)
${}^7\mathrm{Li} + {}^1\mathrm{H} \Rightarrow 2{}^4\mathrm{He}$	17.347	
(b_2) ${}^7\mathrm{Be} + {}^1\mathrm{H} \Rightarrow {}^8\mathrm{B} + \gamma$	0.135	
${}^8\mathrm{B} \Rightarrow {}^8\mathrm{Be}^* + e^+ + \gamma$	15.075	$E_{\nu(\max)} = 14.02$ MeV
${}^8\mathrm{Be}^* \Rightarrow 2{}^4\mathrm{He}$	2.995	

ent energies are confirming this apparent neutrino deficit. A further detailed description of the solar neutrino measurement techniques can be found in Bahcall (1990) or in references on the solar interior such as Cox *et al.* (1991).

3. SOLAR ELECTROMAGNETIC RADIATION

A crude approximation of the solar visible light emissions is that of a 5770-K "blackbody," varying between 6200 and 5400 K depending on the photospheric depth. The relatively thin photosphere is the primary radiative region above which there are an extended solar atmosphere, the chromosphere, and the corona. These layers of gas above the photosphere complicate the solar emission spectrum by introducing both absorption bands and additional emission from the hot corona. The photonic spectrum received from the Sun is not simple but consists of a very large number of emission and absorption lines. Detailed solar spectral atlases used at solar observatories are very large data files filling computer tapes. A very simplified solar electromagnetic emission spectrum is illustrated in Fig. 2.

3.1 Solar Irradiance from the Quiet Sun

The total irradiance of the Sun received at the Earth is historically called the "solar constant." The value of the total solar irradiance at 1.0 AU adopted for the 1986 solar minimum was 1367 ± 3 W m^{-2} (Lean, 1991). The solar irradiance received at the Earth has an annual modulation of about ±3% due to the elliptical orbit of the Earth about the Sun. Analysis of the solar irradiance variations has identified 11-, 22-, and 80-yr components (Lee *et al.*, 1995). The mean irradiance value between 1984 and 1993 determined from measurements by the Earth Radiation Budget Satellite (ERBS) was 1365.3 ± 0.7 W m^{-2} with an estimated accuracy of 0.2%. Unfortunately, the mean numerical value of equivalent measurements from 1978 to 1993 by the Nimbus 7 satellite was 1372.1. The value from the Solar Maximum Mission Satellite (SMM) was 1367.4; the Upper Atmosphere Research Satellite value was 1365.9; and the Space Shuttle Atmospheric Laboratory for Applications and Science (ATLAS1) Solar Constant (SOLCON) active-cavity radiometer value was 1366.9. These discrepancies are a cause of serious concern since the "solar constant" is a fundamental value used in geophysical models of the Earth's atmosphere and climate (see Foukal, 1990, for a more detailed discussion). The observations are being evaluated and refined by technical specialists with regard to the sensor response over the entire solar spectrum, the sensor calibration, sensor degradation, and measurement precision (see Hudson, 1988, or Donnelly, 1992, for a detailed discussion). There is not yet an ideal device for measuring the solar irradiance, and different space and ground-based experiments yield slightly different values. Figure 3 shows 11 years of unadjusted satellite measurements of the "total solar irradiance," illustrating the solar-cycle trend, the variability of the solar flux, and the discrepancies between different instruments.

The solar electromagnetic emission spec-

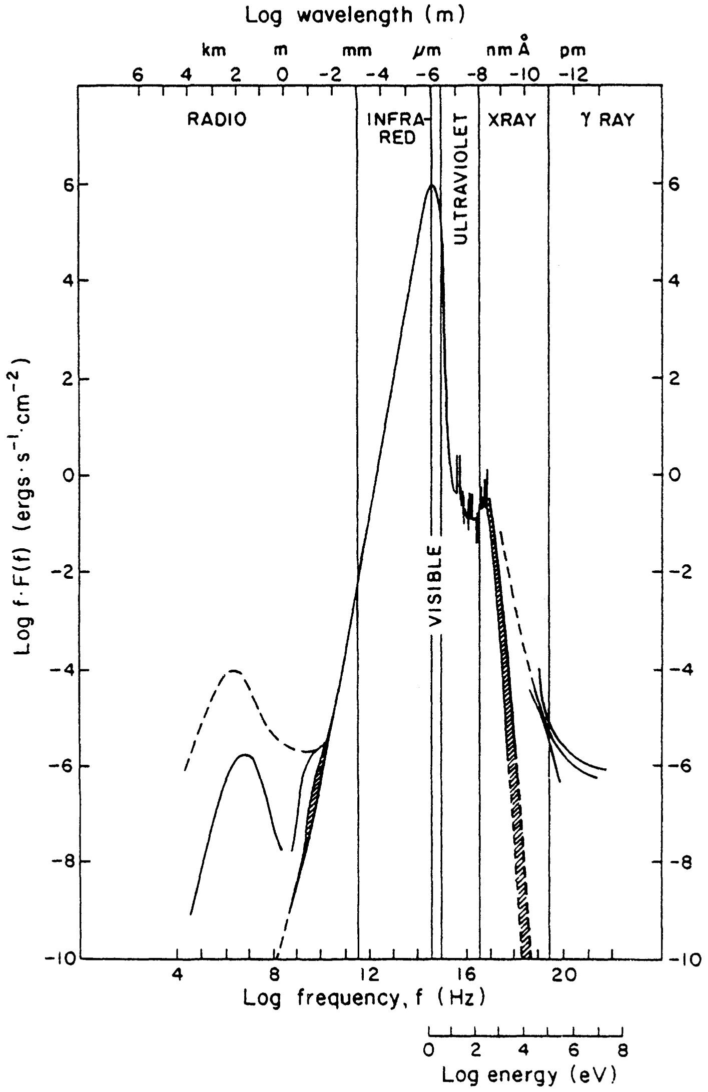

FIG. 2. Overview of the solar radiation spectrum and its variations (adapted from Sawyer *et al.*, 1986). The ordinate is the logarithm of the product of the frequency in hertz by the radiative flux in units of ergs·s^{-1}·cm^{-2}·Hz^{-1}. In the radio region, the solid and dashed lines represent the envelopes expected from "average" and large radio bursts. The shaded areas indicate the solar-cycle variability. The dashed line in the x-ray region indicates the envelope expected for "large" x-ray bursts. The solid lines at the far right indicate gamma-ray bursts.

trum can be divided into regions by wavelength. These regions (see Fig. 2) are detailed in Table 2, adapted from White (1977), and discussed individually in the following sections. Approximately 99% of the total solar radiation is contained in the region between 0.3 and 10 µm. About 2% of the solar irradiance appears at ultraviolet and x-ray wavelengths shorter than 0.32 µm. This radiation is strongly absorbed in the upper atmosphere. More than 70% of the solar irradiance is concentrated in the near-ultraviolet, visible, and near-infrared portions of the spectrum lying between the atmospheric transmission cutoffs near 0.32 and 1.0 µm. Most of this radiation reaches the Earth's surface, and a large portion of this energy enters the lower atmosphere through the evaporation–precipitation cycle of water (see also ATMOSPHERIC PHYSICS; ATMOSPHERIC STRUCTURE; and Figs. 3 and 4 of IONOSPHERE).

3.1.1 The Solar Radio Spectrum The solar radio spectrum ($\lambda > 1$ mm) contains only a very small part of the total solar irradiance; however, the emission in this part of the spectrum is very dynamic. The solar radio emission can be classified as a basic sun component, a slowly varying component, and

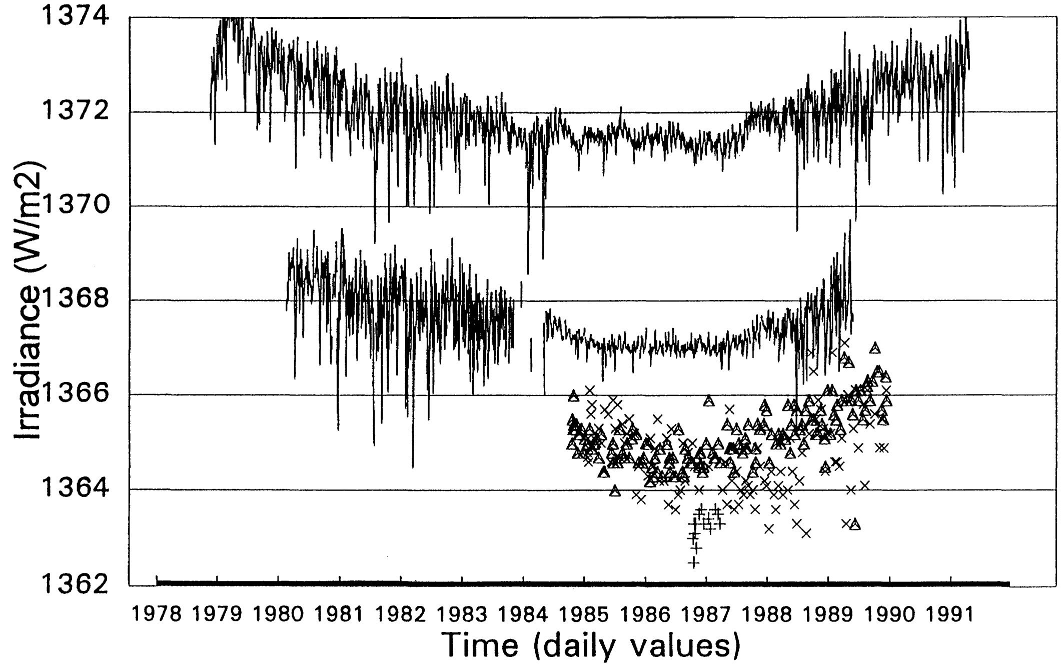

FIG. 3. Unadjusted and uncorrected measurements of the "total solar irradiance" from different instruments on various spacecraft. Data from the NIMBUS 7 spacecraft are at the top of the figure. Data from the Solar Maximum Mission Satellite (SMM) are in the center. Clustered near the bottom are data from NOAA9 (×), NOAA10 (+), and ERBS (Δ), the Earth Radiation Budget Satellite. (Figure courtesy of National Geophysical Data Center.)

a burst component (discussed later in Sec. 5.2.2.1). The solar radio emissions are of interest to radio communicators and radar operators and are diagnostics of solar activity processes. The 10-cm flux is used as a proxy of solar extreme-ultraviolet (EUV) flux.

If the Sun radiated only as a thermal source, the emitted energy density would vary with frequency and temperature according to Planck's radiation law. The following equation gives the equivalent blackbody temperature or brightness temperature:

$$B_f = 2\kappa T f^2 c^{-2} = 2\kappa T \lambda^{-2}. \qquad (1)$$

In this equation, B_f is the "brightness" at frequency f in W m^{-2} Hz^{-1} sr^{-1}; the frequency f is in Hz, or cycles per second; the wavelength λ in meters; the temperature T in kelvins; the velocity of light c in meters per second; and Boltzmann's constant κ in joules per kelvin. Figure 4 illustrates the solar radio spectrum and the spectra of a blackbody at various temperatures. In this figure, the solar power-flux density (power per unit area per bandwidth) is plotted against wavelength.

Table 2. Regions of the solar electromagnetic emission spectrum

Spectral region	Wavelength range
Radio	$\lambda > 1$ mm
Far infrared	1 mm $> \lambda >$ 10 μm
Infrared	10 μm $> \lambda >$ 0.75 μm
Visible	0.75 μm $> \lambda >$ 0.3 μm
Ultraviolet (UV)	300 nm $> \lambda >$ 120 nm
Extreme ultraviolet (EUV)	120 nm $> \lambda >$ 10 nm
Soft x rays	10 nm $> \lambda >$ 0.1 nm
Hard x rays	0.1 nm $> \lambda$

3.1.2 The Solar Far-Infrared Spectrum

The far-infrared portion of the solar disk spectrum from 1000 to 10 μm contains an integrated irradiance of about 0.0802 ± 0.026 W m^{-2}, which is 0.057% of the solar constant. This part of the solar emission spectrum is essentially continuous, and there is relatively little solar-cycle change.

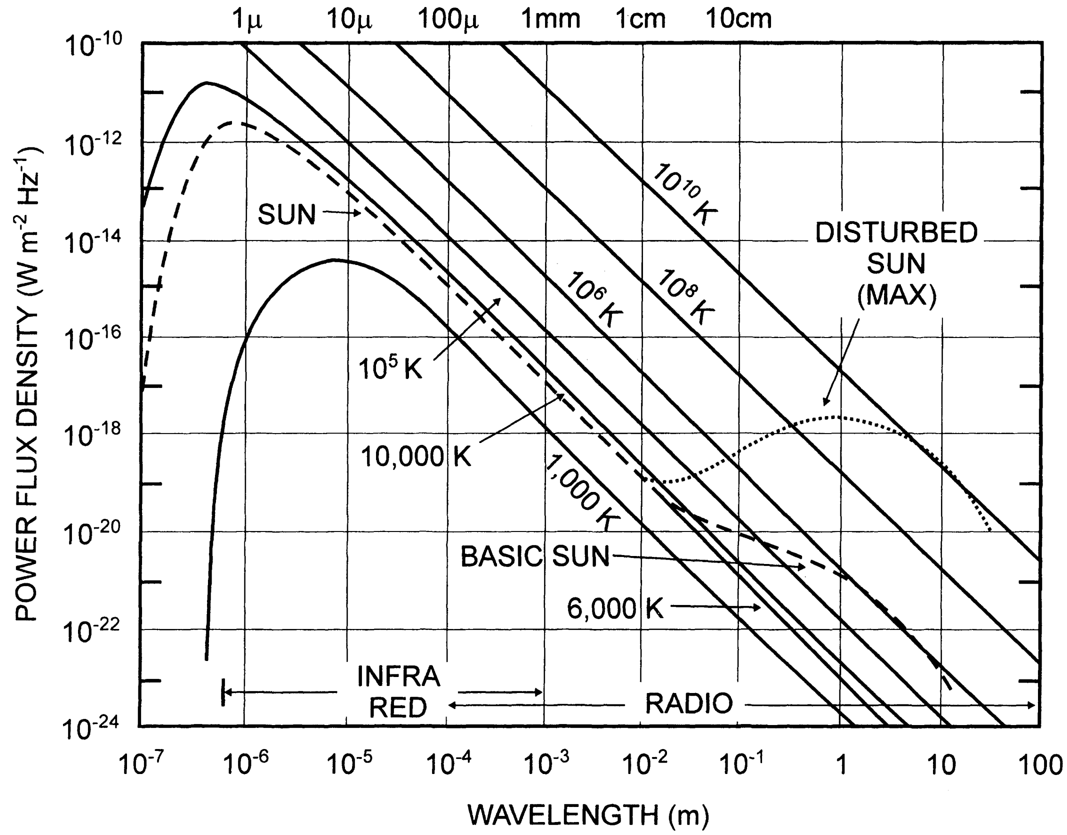

FIG. 4. The solar radio spectrum and the spectra of a blackbody at various temperatures.

3.1.3 The Solar Optical and Infrared Spectrum The optical and infrared portion of the solar spectrum in the wavelength range from 0.3 to 10 nm contains about 99% of the total solar irradiance and is the most important part of the solar constant with respect to the energy input to the Earth. Optical and infrared radiation are usually considered to be "quiet Sun" emissions. There are changes in the emission in the Fraunhofer lines from active regions. The accuracy of the observations remains a point of controversy because of systematic differences between the flux determination found by different measurement methods (see Fig. 3). The absolute accuracy should be in the 1 to 2% range; however, calibration errors of up to 8% have been reported that affect recent measurements of solar irradiance.

3.1.4 The Solar Ultraviolet Spectrum The ultraviolet region of the solar spectrum in the wavelength range from 300 to 120 nm contains only about 1% of the total solar irradiance. This radiation is important because it heats the stratosphere and photodissociates the major molecular constituents of the upper atmosphere, being almost totally absorbed by the ozone and diatomic oxygen molecules (see also ATMOSPHERIC PHYSICS; ATMOSPHERIC STRUCTURES; IONOSPHERE). The spectrum of this region is complex; at about 140 nm, there is a transition from a strong continuum broken by Fraunhofer absorption lines to a weak continuum dominated by emission line spectra. The ultraviolet wavelength region is where there are the greatest technical difficulties in making accurate measurements. The measurement accuracies in this part of the solar spectrum are often about ±10%, sometimes achieving a relative accuracy of 2%. Measurement accuracies of 1% are needed for global modeling of the upper atmosphere, particularly for ozone time variations. Because this portion of the solar spectrum is so complex, reference spectra (see Hinteregger *et al.*, 1981, or Tobiska, 1993) are compiled and updated as more precise measurements become available.

The solar ultraviolet (UV) flux variations contribute between one-fifth and one-third of the solar-cycle enhancement in the total solar irradiance. There are significant changes in the solar ultraviolet due to solar activity. There are 27-day (the average solar rotational period) variations that can be as large at 25% at 120 nm, these variations decreasing approximately exponentially with wavelength to about 1% at 300 nm. The largest changes in the emission in this entire wavelength region occur during major solar flares and can be as great as 16%. Because of the solar-activity dependence, the recent UV and EUV observations are now being published monthly in *Solar-Geophysical Data* by the NOAA National Geophysical Data Center.

3.1.4.1 The Lyman-Alpha Flux. The Lyman-alpha line centered at 121.6 nm contains more radiative energy than all the other shorter wavelengths in the solar spectrum. This "line" is not simple and is actually a complex profile. The Lyman-alpha emission is a principal source of ionization in the upper *D* region of the Earth's atmosphere at times of low solar activity (see also IONOSPHERE). The integrated Lyman-alpha flux depends on solar activity and can vary between 3.3 and 6.5 erg cm^{-2} s^{-1}. The Lyman-alpha flux variation during a solar rotation can be ±15% and about a factor of 2 between solar maximum and solar minimum.

3.1.5 The Solar Extreme-Ultraviolet Spectrum

The solar flux in the extreme ultraviolet region of the solar spectrum is the main ionization source for the Earth's ionosphere (see also IONOSPHERE). It contains a very complex series of resonance lines and continua from the major constituents of the solar atmosphere. The energy flux at wavelengths below 120 nm is small, but it is important because it determines the ionization state of the Earth's ionosphere. At wavelengths greater than 17 nm, the variation during a solar rotation is typically ±15% and can be a factor of 5. The variation over a solar cycle can be two orders of magnitude.

3.1.6 The Solar X-Ray Flux

Solar x rays are an important source of ionization in the Earth's ionosphere, with absorption wavelength being a function of depth in the terrestrial atmosphere (see also ATMOSPHERIC PHYSICS; IONOSPHERE). The Lyman-alpha flux and the cosmic radiation are the main sources of ion production in the Earth's ionosphere; however, when the 0.05- to 1-nm flux exceeds 10^{-6} W m^{-2}, the solar x rays become an important, and for large transient events the dominant, source of ionization in the Earth's atmosphere. The variations in the 0.1–1.0-nm flux are so large and unpredictable that the soft x-ray flux is published for general use in *Solar-Geophysical Data.* The 0.1–0.8-nm (1–8-Å) flux varies by orders of magnitude over a solar cycle. Large variations occur during a solar rotation with marked differences from rotation to rotation depending on the number, size, growth rate, and spatial distribution of solar-active regions. During solar-flare events the 0.1–0.8-nm flux can increase by more than three orders of magnitude, and the solar x-ray and gamma-ray bursts are the cause of the "sudden ionospheric disturbance" (SID) troublesome to terrestrial radio communications. For a very large X10-class solar x-ray event, the peak radiative power received at the Earth in the 0.1–0.8-nm region will exceed 1.0×10^{-3} W m^{-2} and will have a duration of several hours; the total energy received in this wavelength band may exceed 1 J. (See Sec. 5.1.1.2 for an explanation of classification of solar x-ray events.)

4. SOLAR PARTICLE EMISSION FROM THE QUIET SUN

4.1 Solar Plasma Emissions

The Sun is constantly emitting plasma, called the solar wind. The existence of the solar wind was hypothesized, and many of its characteristics were theoretically predicted, before it was actually measured by spacecraft. The solar corona is very hot, and as a result of poorly understood nonthermal energy input, some of the particles acquire sufficient energy to overcome the gravitational potential energy (2.0 keV per particle at $1R_0$) and escape into interplanetary space. These particles form a highly ionized plasma in which is embedded the coronal magnetic field. This becomes the interplanetary magnetic field, which is carried out from the Sun by the solar wind. The plasma kinetic energy is generally much larger than the magnetic energy, and the field is said to be "frozen in" the solar-wind flow.

4.1.1 The Solar Wind In the general solar-wind theory devised by Parker (1963) and modified by later researchers, there is a continuous expansive motion of the solar wind, reaching supersonic flow and continuing to accelerate slowly until it approaches an asymptotic value at some great distance from the Sun. In its supersonic flow condition, the solar-wind bulk-motion speed is faster than the speed of sound in the interplanetary medium and is faster than the Alfvén speed at which magnetic wave phenomena propagate. In its radial expansion, the solar-wind density is predicted to decrease as $1/r^2$. Many of the Parker (1963) predictions of the general characteristics have been verified; however, spacecraft measurements show a solar wind that is faster, cooler, and more turbulent, variable, and structured than the predictions of the simple Parker theory. The goal of modern solar-wind theorists is to explain the space observations. However, there is no satisfactory general model of the solar wind. Currently, there are a number of individual theories that attempt to model several aspects of the solar-wind behavior.

Various space probes have sampled the solar wind from distances as close to the Sun as 0.3 AU to as far from the Sun as 70 AU (in 1995), from latitudes in the ecliptic plane to 80° solar latitude. In addition to the limited *in situ* measurement of the solar wind by spacecraft, the interplanetary scintillation radio observation technique can determine the pattern speed of irregularities in the electron density as the plasma moves through the line of sight to various stellar point sources, thus providing information on the structure of the interplanetary plasma. The solar wind is a polytropic plasma with a proton component having an average particle kinetic energy of approximately 1 keV and an electron component having average particle kinetic energy of approximately 100 eV. The solar wind is very structured, and it is customary to refer to solar-wind streams. These solar-wind streams corotate with the Sun and retain their individuality to very large distances from the Sun (they are still observable near the ecliptic plane by our most distant interplanetary spacecraft). Low-speed streams are generally at low solar latitudes and seem to be associated with "magnetically closed" coronal features. The high-speed streams generally originate from "magnetically open" coronal structures called coronal holes. High-speed streams "override" slow-speed streams generating compression density structures and corotating shocks. The result of the stream structure of the solar wind is that the behavior of the individual plasma elements can be approximated as motion radially outward with an essentially constant velocity (except when there are stream–stream interactions) with a slight slowing down of the fast solar-wind streams at large distances. Transient structures propagating in the solar wind are often the result of coronal mass ejections. It is common to consider two different classes of solar wind: the steady or normal solar wind and the transient solar wind. Other classifications often used for the solar wind in the ecliptic plane are slow (<400 km s^{-1}), fast (>600 km s^{-1}), or average.

The general characteristics of the solar wind near the ecliptic plane (normalized to 1 AU) are summarized in Table 3. There is a great variability in the solar wind. Many of the properties of the solar wind are dependent on the flow speed, as shown in this table. However, the momentum flux and the total energy are remarkably independent of solar-wind speed. The HELIOS spacecraft data set covering the period from 1974 to 1982 indicates that the density of the solar wind between 0.3 and 1.0 AU can be modeled as a power law $n(r) = 6.1r^{-2.1}$, where r is the radial distance from the center of the Sun.

4.1.1.1 Solar-Wind Composition. The solar wind is an electrically neutral plasma. The average ionic component consists principally of protons, about 4% alpha particles, and a small number of other positively charged ions. The solar-wind composition and charge state are variable (stream structure and temperature dependent) and different from either the solar photospheric or the solar coronal composition, probably reflecting the ionization and acceleration processes. The measurements of the elemental

Table 3. Solar-wind parameters at 1 AU in the ecliptic plane.

Parameter[a]	Units	Slow solar wind	Fast solar wind	Average solar wind
Proton bulk flow speed	km s^{-1}	335	680	475
Proton density	cm^{-3}	9.5	3.0	6.5
Proton flux density	10^8 cm^{-2} s^{-1}	3.2	1.9	2.7
Proton momentum flux density	10^8 dyn cm^{-2}	2.12	2.16	2.1
Proton temperature	10^5 K	0.5	2.6	1.2
Electron temperature	10^5 K	1.6	1.5	1.4
Proton heat flux	10^{-3} erg cm^{-2} s^{-1}	0.05	0.23	0.13
Electron heat flux	10^{-3} erg cm^{-2} s^{-1}	4.3	5.1	4.3
Alpha temperature	10^5 K	1.4	10.	5.8
Alpha/proton ratio		0.031	0.042	0.04
Total energy flux	erg cm^2 s^{-1}	1.37	1.49	1.45

[a] Parameters averaged from several sources.

abundances of the solar wind (tabulated in Table 4) are still being improved. Some of the most detailed measurements have been made in the magnetosheath of the Earth's magnetosphere.

4.1.2 The Interplanetary Magnetic Field The interplanetary magnetic field (IMF) is convected out of the Sun via the solar wind (a highly ionized conductive plasma). Its basic properties were successfully predicted by Parker (1963); however, spacecraft measurements reveal an IMF that is more turbulent, variable, and structured than the predictions of the simple Parker theory. It is customary to organize IMF data according to the basic Parker theory and to organize variations in the behavior of the IMF by their deviation from the Parker theory predictions.

Table 4. Solar-wind abundances relative to oxygen (Gloeckler and Geiss, 1989).

Element	Relative abundance: Interplanetary	Relative abundance: Magnetosheath
H	1900 ± 400	1125 ±100
He	75 ± 20	45 ± 5
C		0.529 ± 0.013
N		0.129 ± 0.008
O	1.000	1.000
Ne	0.17 ± 0.02	0.114 ± 0.008
Mg		0.106 ± 0.001
Si	0.19 ± 0.04	0.103 ± 0.011
S		0.038 ± 0.009
Ar	0.004 ± 0.001	
Fe	0.19 ± 0.1	0.124 ± 0.004

In the time-stationary spherically symmetric Parker theory, assuming an expanding solar wind at constant velocity at a distance a few solar radii from the Sun, the three components of the interplanetary magnetic field are given by

$$B_r(r,\theta,\varphi) = B(\phi,\theta_0)(b/r)^2, \tag{2}$$

$$B_\theta(r,\theta,\varphi) = 0, \tag{3}$$

$$B_\phi(r,\theta,\varphi) = B(\theta,\varphi_0)(\Omega b^2/vr^2)(r_0 - b)\sin\theta, \tag{4}$$

where b and φ_0 are the distance and longitude at which the inner-boundary radial field $B(\theta,\varphi_0)$ is assumed to exist, v is the solar-wind velocity, and Ω is the solar angular rotation rate. The distances r and b are connected by the equation

$$r/b - 1 - \ln(r/b) = (v/b\Omega)(\varphi - \varphi_0), \tag{5}$$

which for $r \gg b$ reduces to the Archimedean spiral equation. This formulation works well for an average description of the interplanetary magnetic field to the distances traveled by man-made spacecraft.

The basic topology of the interplanetary magnetic field is very important in the understanding of solar energetic-particle events. The outward flow of the solar wind with its "frozen in" magnetic field from a rotating sun leads to an Archimedean spiral geometry. This spiral geometry is easiest to visualize in the ecliptic plane, as illustrated in Fig. 5. The spiral geometry is predicted to be present at all latitudes, and the latest Ulysses solar polar mission results confirm these expectations of the Parker theory.

There are variations in the magnitude of the interplanetary magnetic field observed at all wavelengths. The magnitude fluctuation (field variance σ_B^2) can be approximated by a power law with a slope of -3. Like the solar wind, the interplanetary magnetic field is highly structured. The gross characteristics of the solar magnetic field can be approximated by an inclined dipole (the inclination varying as the solar cycle) with well-defined northern and southern polarities separated by a heliospheric current sheet, as illustrated in Fig. 6. (A useful concept for visualizing this is the ballerina skirt.) A spacecraft in the ecliptic plane would sense changes in the polarity of the interplanetary magnetic field during a solar rotation, as it is sequentially connected to the polarity of the "northern" and then "southern" solar latitudes. This regular variation in the sign of the interplanetary magnetic field was originally called the "sector structure." This structured interplanetary magnetic field exists primarily in the ecliptic plane, and the latitudinal extent of this "sector structure" is called the tilt of the interplanetary current sheet. The inclination angle of the interplanetary current sheet is solar-cycle dependent, minimum at solar minimum and maximum during the solar magnetic polar reversal at solar maximum. Interplanetary spacecraft at heliolatitudes "above" or "below" the current sheet observe a unipolarity interplanetary magnetic field reflecting the solar magnetic cycle polarity of the solar magnetic field. Solar source surface maps of the solar magnetic field modeled and extrapolated to 2.5 solar radii and then projected to 1 AU at the average solar-wind speed have been shown to have a 0.8 correlation with the sign of the interplanetary magnetic field observed at the Earth (Hoeksema and Scherrer, 1986).

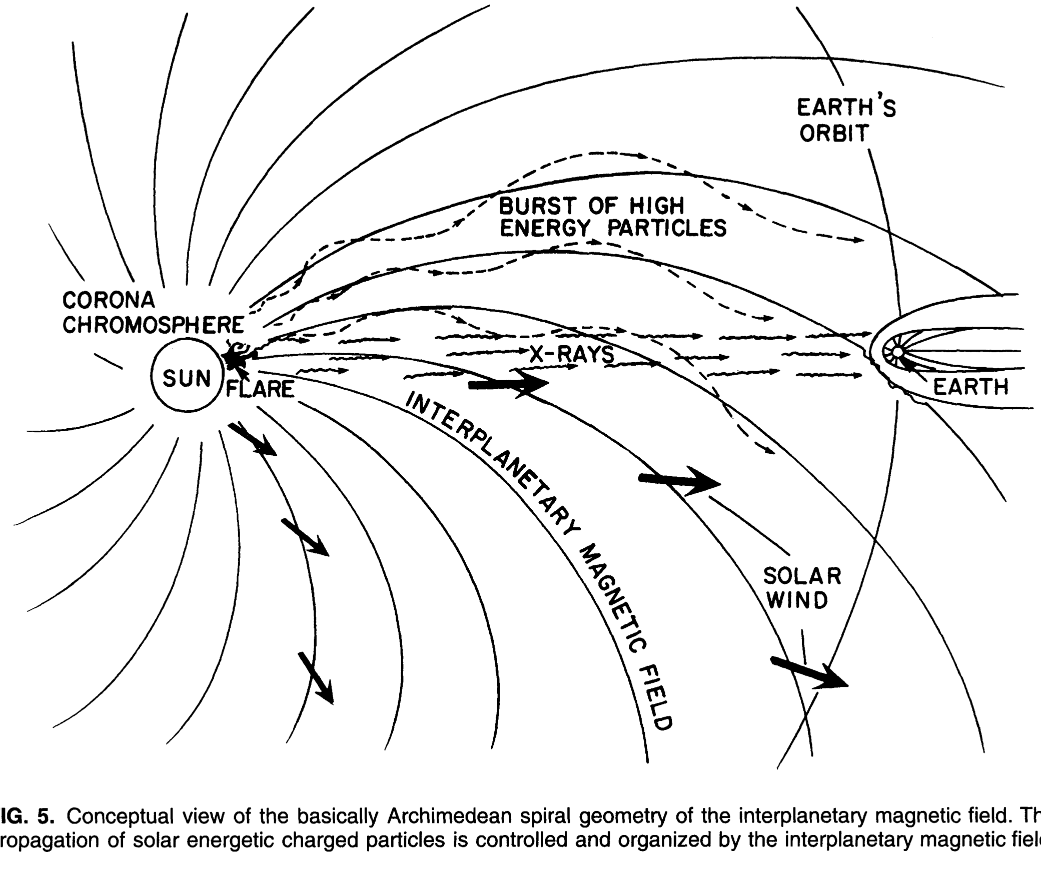

FIG. 5. Conceptual view of the basically Archimedean spiral geometry of the interplanetary magnetic field. The propagation of solar energetic charged particles is controlled and organized by the interplanetary magnetic field.

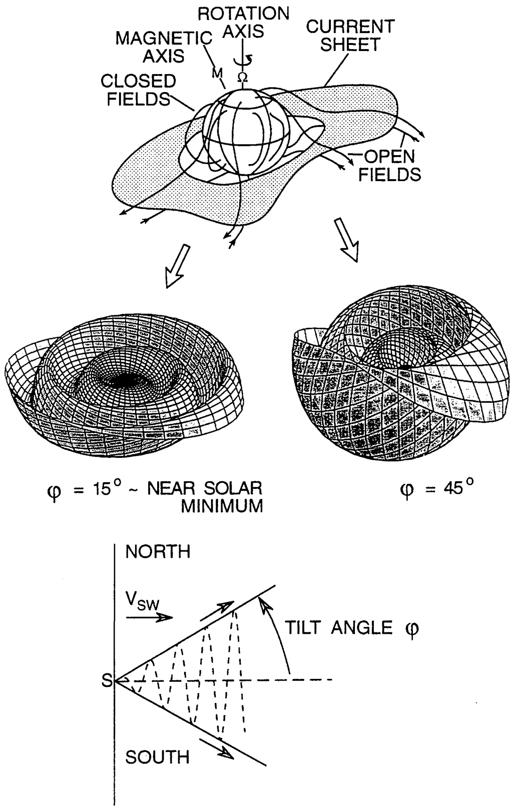

FIG. 6. Conceptual illustration of the heliospheric current sheet during solar minimum and solar maximum. Top: The concept of the heliospheric current sheet. Center: Conceptual view of the amplitude and structure of the heliospheric current sheet near solar minimum (left) and during more active periods of the solar cycle (right). Bottom: Conceptual view of the "wavy" heliospheric current sheet being convected out by the solar wind.

5. SOLAR ELECTROMAGNETIC RADIATION FROM THE ACTIVE SUN

Features visible in the photosphere and chromosphere (and, with space technology, the solar corona) are observed to wax and wane in a cyclic manner with an approximate 11-yr period. The concept of solar activity has been broadened from the classical narrow concept of sunspots and solar flares to embrace an entire range of solar processes, time scales, and energy-release processes. The cyclic behavior affects the electromagnetic output of the Sun; however, the effect is very dependent on the wavelength. The solar-cycle effect is smallest in the visible wavelengths and has a small observable effect on the solar constant (see Fig. 3). The solar-cycle effect is largest at the extreme ends of the electromagnetic spectrum (the long radio wavelengths and the short x-ray wavelengths).

5.1 Solar Activity and Solar Cycles

The cyclic solar activity is the result of variations in the energy in the form of magnetic flux, convected to the solar surface by the action of the solar dynamo operating in the convection zone. The sun has a magnetic cycle with a period of about 22 years, with the average magnetic field observed at the solar poles reversing about every 11 years during the maximum of the solar-activity cycle. This same solar magnetic polarity phase cycle is observed in sunspot pairs where the leader spot and the follower spot have opposite polarities. The sunspots observed on the solar photosphere are manifestations of locally strong magnetic fields. The visible solar flare is the result of the explosive release of energy, stored as magnetic stress, into the solar atmosphere. The number of "flares" on the Sun and the number of magnetic disturbances observed on the Earth are roughly correlated with the solar-activity cycle.

5.1.1 Sunspot Cycles Sunspots and solar flares are the most obvious manifestations of solar activity. The average period of the sunspot cycles for the last 280 years has been 11.1 years with an average rising phase of 4.8 years and an average declining phase of 6.2 years. Figure 7 illustrates the sunspot cyclic behavior since 1900. It is traditional to use the International sunspot number as a proxy to indicate solar activity, and there are many methods of predicting the "next solar cycle." These different methods generate diverse predictions, and there is no universally accepted method of sunspot-number prediction. Some contemporary theories consider the solar cycle to be chaotic and not amenable to precise statistical prediction.

At the beginning of a solar-activity cycle, new sunspots are observed at midlatitudes on the sun, the average sunspot position gradually migrating toward the solar equator as the solar cycle progresses, resulting in a "butterfly diagram" in a map of sunspot heliolatitudes over a solar cycle. Large sunspots

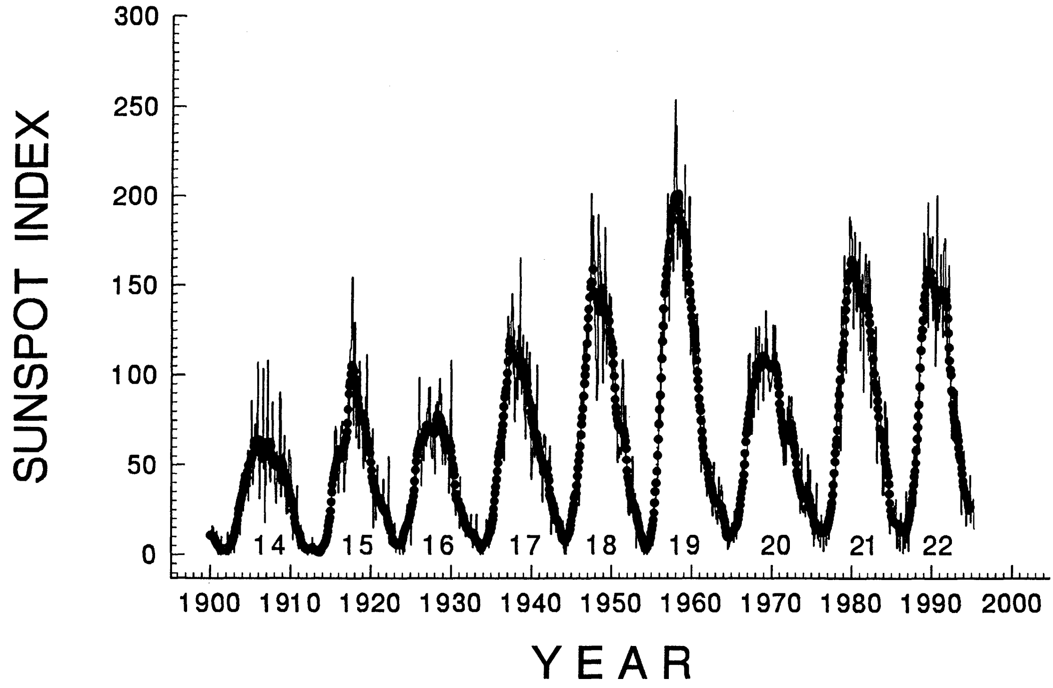

FIG. 7. Sunspot cycles since 1900. The light line is the monthly sunspot index. The large solid dots denote the smoothed sunspot index (a 13-month running mean).

can be observed on the solar surface with an unaided human eye, and the historical record of sunspot observations can be extended for approximately 2000 years. The current cyclic behavior has not always been apparent; during the Maunder minimum and the Spöre minimum, very few sunspots were visible (or reported) for long periods of time (see review by Eddy, 1978). Sunspot numbers extend from 1610 to the present, and the sunspot cycles have been numbered since 1700. Sunspot numbers are considered "reliable" for statistical purposes only from 1848 when the Swiss astronomer Wolf defined rules for counting sunspots and introduced the sunspot index. These numbers are currently called the International sunspot numbers; historically they were called the Zürich sunspot numbers [see McKinnon (1987) and data files maintained by the National Geophysical Data Center]. There are several different types of sunspot numbers, such as the number as determined by the Wolf sunspot index and smoothed sunspot numbers that are a 13-month running mean of the Wolf sunspot index. A less-well-known and little-used sunspot count is the AVSO number (American visual sunspot observations).

5.2 Solar Flares

The English term "solar flare" is the name given to the sudden release of energy (occasionally $>10^{32}$ ergs) in a relatively small volume of the solar atmosphere. The original term referred to a *chromospheric brightening* that is more descriptive of the observations, occasionally seen in white light but usually observed in the hydrogen-alpha line. The term has been expanded to include all of the electromagnetic and particle emissions associated with the energy release. There is a general correlation between the optical importance of a solar flare and its radio and x-ray output, but there are very large variations in the data. In extreme instances, optical subflares are associated with large x-ray emissions, and in other extreme instances very large optical flares have been almost "radio quiet."

The concept of what is a solar flare has undergone considerable evolution. Contemporary concepts based on results from the NASA Solar Maximum Mission and the Japanese YOHKOH satellite favor energy release in loops. As seen from space in ultraviolet and x-ray wavelengths, characteristic looplike structures outline the magnetic fields from active regions. The loop sizes vary considerably, from a few thousand kilometers to more than 100 000 km. The disruption of magnetic stress in these loops seems to result in the solar flare. In a large solar flare, many loops, including entire arcades, are involved. The mechanism that triggers the flare processes is not understood. The impulsive phase of the solar flare is the sudden release of energy either directly or indirectly into electromagnetic and particle radiation. The flaring loop model developed by NASA as a result of the Solar Maximum Mission is shown in Fig. 8.

While there are a number of theories involving processes occurring during a solar flare, there is no general solar flare theory. Current concepts refer to an impulsive phase and a gradual phase. The theoretical concept of how the impulsive phase of the flare proceeds is referred to as the thick-target model. In this model the primary outcome of the magnetic reconnection process is the production of large quantities of energetic electrons and protons. These charged particles will move away from the reconnection site (believed to be in the corona) along the magnetic field lines toward the denser lower portions of the solar atmosphere. As they encounter dense plasma in the chromosphere, they will interact and deposit their energy and further heat the plasma. Energetic electrons will generate gyrosynchrotron radiation as they gyrate along the field lines; they also generate bremsstrahlung (German for braking radiation) as they interact with the denser material. The more energetic ions can cause nuclear interactions, generating gamma rays and knocking neutrons out of the target nuclei. As the tube fills with hot

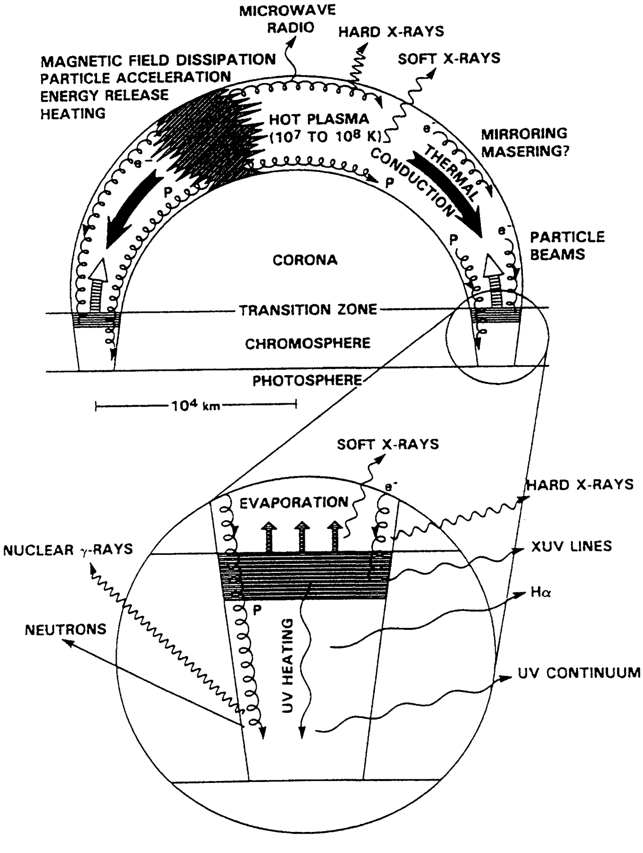
MICROWAVE RADIO
HARD X-RAYS
MAGNETIC FIELD DISSIPATION
PARTICLE ACCELERATION
ENERGY RELEASE
HEATING
SOFT X-RAYS
HOT PLASMA (10⁷ TO 10⁸ K)
e⁻
P
THERMAL CONDUCTION
MIRRORING MASERING?
PARTICLE BEAMS
CORONA
TRANSITION ZONE
CHROMOSPHERE
PHOTOSPHERE
10⁴ km
SOFT X-RAYS
HARD X-RAYS
EVAPORATION
NUCLEAR γ-RAYS
XUV LINES
Hα
NEUTRONS
UV HEATING
UV CONTINUUM

FIG. 8. An idealized flaring loop, illustrating probable sources of the radiation involved (courtesy of NASA).

plasma, it emits ultraviolet and soft x-ray radiation, initiating what is called the gradual phase. Once the heating begins to decrease, the plasma cools by radiation. Figure 9 is a sketch of an idealized solar-flare radiative-emission curve.

The gradual phase is a more time-extended portion of the solar flare. Essentially, explosive evaporation carries large quantities of high-temperature plasma into the magnetic flux tubes where it is observed as ultraviolet and soft x-ray emission. Much of the energy put into the corona by the flare is gradually moved by thermal conduction to the chromosphere where it is radiated away by the strong emission lines of hydrogen and helium.

5.2.1 Solar-Flare Reporting

5.2.1.1 Optical Solar Flares. Optical solar-flare reporting is done in accordance with specifications of the International Astronomical Union. The optical classification of a solar flare is based on its measured optical area (in millionths of the solar disk) corrected for foreshortening and brightness at the time of maximum brightness in the Hα line (centered at 656.3 nm). Optical solar flares are currently (since 1965) ranked in importance as 0 (or *s*), 1, 2, 3, to a maximum of 4. Solar flares are subjectively ranked in brightness as *f* (faint), *n* (normal), or *b* (bright). Prior to 1965, optical solar flares were ranked as *s* (sub) or 1, 2, 3, and 3+. The *Quarterly Bulletin on Solar Activity* is a publication of the International Astronomical Union in which the solar-flare reports have been evaluated and grouped. (Unfortunately, this is a time-consuming process, and the publications are several years "late.") In the United States, the NOAA National Geophysical Data Center collects and maintains files of unevaluated solar-flare reports that are published monthly in *Solar-Geophysical Data.* The statistical occurrence of optical solar flares for the last four solar cycles is given in Table 5.

5.2.1.2 Solar X-Ray Flares. Since about 1969, there have been routine space measurements of the solar x-ray flux by the NOAA Space Environment Laboratory. The NOAA classification of solar-flare soft x-ray emission is generally accepted worldwide. This is a letter–number combination denoting the magnitude of the peak x-ray burst intensity measured at the Earth in the 0.1–0.8-nm band, detailed in Table 6. The letter denotes the class (order of magnitude) of the x-ray emission, and the number describes the magnitude in each class. For example, a

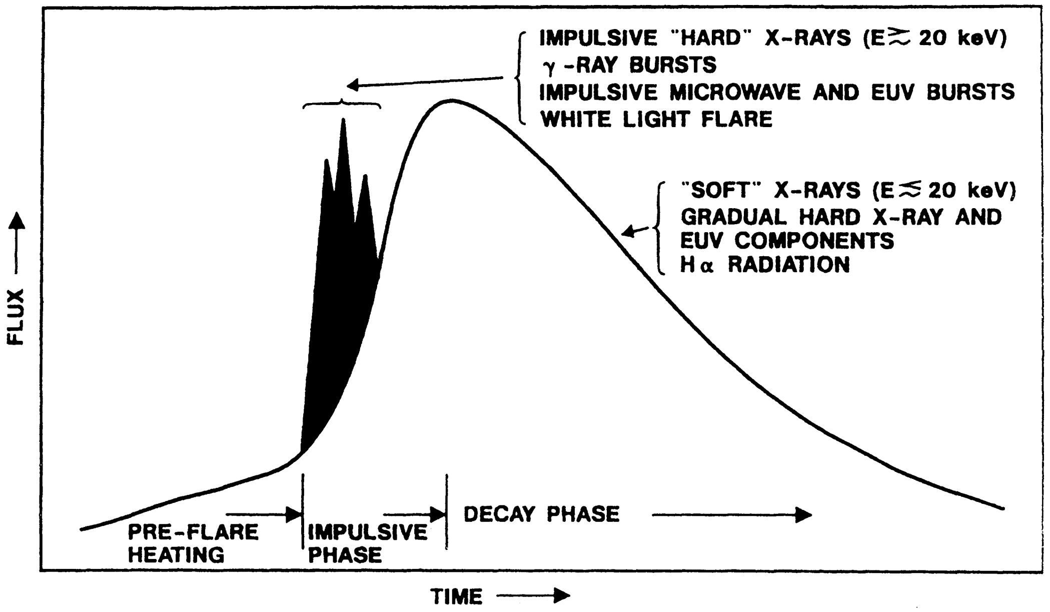

FIG. 9. Idealized solar-flare radiative energy intensity-time curves.

Table 5. Optical solar-flare occurrence frequency.

Optical importance	Approximate flares per year Solar minimum	Approximate flares per year Solar maximum
s[a]	~800	>10 000
1	37	1 100
2	3	120
3	0	10
4	0	2

---

[a] A curious fact is that there are no rules for the reporting of subflares (individual observatory option), which leads to a skewing of the statistics when all reported optical brightenings are plotted.

Table 6. The NOAA SEL solar x-ray flare classification.

Class	Power in the 0.1–0.8-nm band
B	$<10^{-6}$ W m^{-2}
C	$\geq 10^{-6}$ to $<10^{-5}$ W m^{-2}
M	$\geq 10^{-5}$ to $<10^{-4}$ W m^{-2}
X	$\geq 10^{-4}$ W m^{-2}

peak soft x-ray flux emission of 6.2×10^{-5} W m^{-2} would be classified as an *M* 6.2 event. Figure 10 illustrates the solar flare x-ray frequency observed since 1969.

5.2.2 Solar-Flare Emissions

5.2.2.1 Solar-Flare Radio Emissions. Radio burst emission occurs during the solar-flare process; the burst characteristics vary with wavelength. The duration and intensity are extremely variable; the duration may last from seconds to hours, and the intensity may range from small to extremely large. Accelerated electrons are the source of the radio emission, and the principal mechanisms are gyrosynchrotron radiation from electron motion around magnetic field lines or plasma radiation stimulated by electrons penetrating into the preexisting plasma. Another radiation mechanism hypothesized as a result of asymmetric electron distributions in magnetic flux tubes is the cyclotron maser radiation. The electron gyrofrequency is given by

$$\nu_B = eB/2\pi m_e c \approx 2.8 \times 10^6 B. \quad (7)$$

The plasma frequency is given by

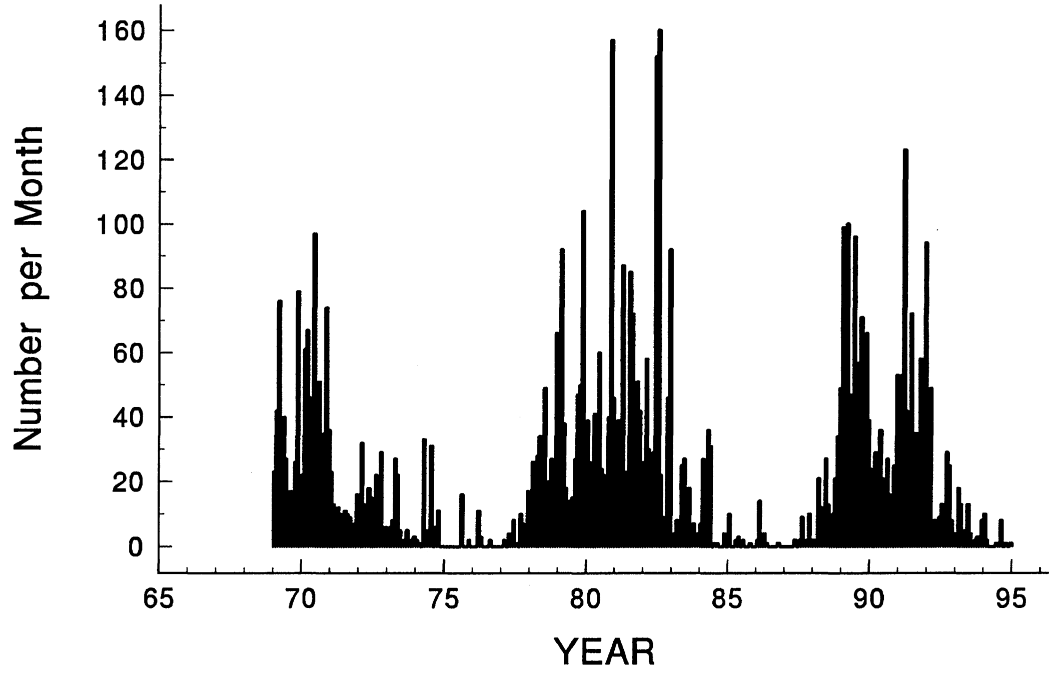

FIG. 10. The frequency of solar flares with x-ray emission $\geq 10^{-5}$ W m^{-2} (Class M1) since 1969 (data courtesy of NOAA Space Environment Laboratory).

$$\nu_p = (n_e e^2/\pi m_e)^{1/2} \approx 9000\sqrt{n_e}. \qquad (8)$$

In these equations e and m_e are the charge and the mass of the electron, B is the magnetic field strength in gauss, n_e is the plasma electron number density, and c is the speed of light.

In the low corona near active regions and/or near sunspots, the magnetic field may be so strong that the electron gyrofrequency is higher than the plasma frequency. Energetic electrons with energy >100 keV that are confined to move around the magnetic field lines emit radiation at frequencies up to 10–100 times their gyrofrequency. In other locations in the solar atmosphere where the magnetic field is weaker and the gyrofrequency is less than the plasma frequency, electrostatic waves (called Langmuir waves) stimulated by anisotropic electron distributions generate plasma radiation.

The intensity of the solar-flare radio burst is generally reported in solar flux units (SFU). One solar flux unit is 10^{-22} W m^{-2} Hz^{-1} or 10^4 Jansky. Moderate solar radio bursts have amplitudes in the hundreds of SFU. Large solar radio events have amplitudes exceeding 1000 SFU. Occasionally a very powerful solar flare may generate a solar radio burst with a peak flux exceeding 10 000 SFU for several minutes. In the United States, the NOAA National Geophysical Data Center (which also functions as World Data Center A for Solar–Terrestrial Physics) collects and maintains files of solar radio bursts, which they publish monthly in *Solar-Geophysical Data.* There is a detailed classification system by spectral type for bursts in the metric wavelength range illustrated in Fig. 11 and described below.

- Type I solar radio emission: A noise storm composed of many short, narrow-band bursts in the meter wavelength range (300–500 MHz), of extremely variable intensity. The storm may last from several hours to several days.
- Type II solar radio emission: Narrow-band emission that begins in the meter range (300 MHz) and sweeps slowly (tens of minutes) toward decameter wavelengths (10 MHz). Harmonics at twice the frequency are often observed. Type II emissions often occur in association with solar flares and fast coronal mass ejections. The metric type II is indicative of a shock wave moving through the solar atmosphere from 1.2 to 2 solar radii.
- Type III solar radio emission: Narrow-band bursts that sweep rapidly (seconds) from decimeter to dekameter wavelengths (500–0.5 MHz). They often occur in groups and are an occasional feature of complex solar-active regions. The fast fre-

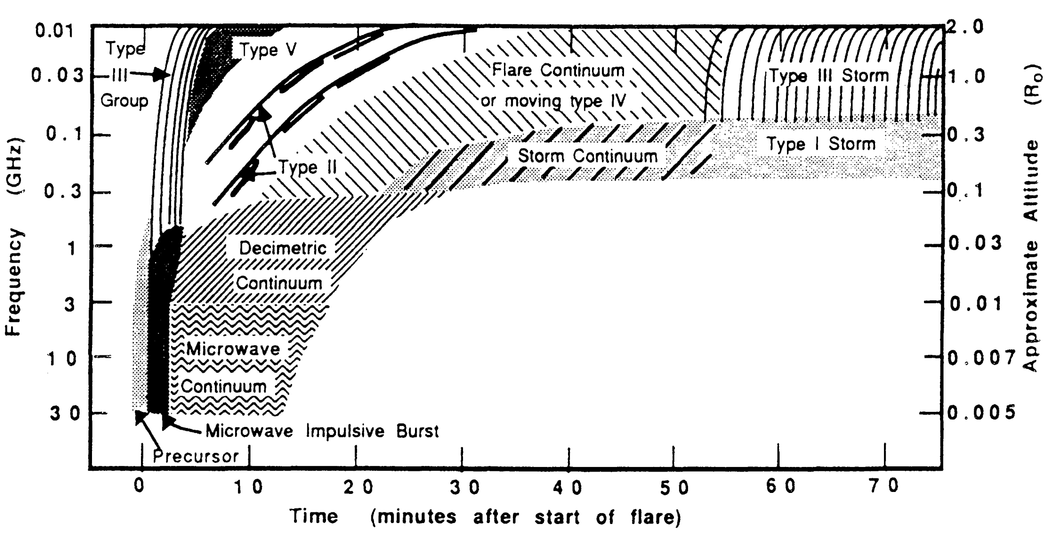

FIG. 11. Schematic diagram illustrating the types of solar radio emission associated with solar flares as a function of dynamic spectra, emission frequencies, and time. The emission altitude is indicated on the right.

quency drift is believed to be caused by fast electrons (electrons traveling about 10% of the speed of light) streaming outward. Type III emission is often associated with solar electron events (see Sec. 6.2.2).

- Type IV solar radio emission: A smooth continuum of broad-band bursts primarily in the meter range (300–30 MHz). These bursts occur with some major flare events; they begin 10–20 min after the flare maximum and can last for hours. Type IV bursts are commonly associated with CMEs (see Sec. 6.1.1).
- Type V solar radio emission: Short-duration (a few minutes) broad-band continuum noise in the dekameter range usually associated with type III bursts.

5.2.2.2 Solar-Flare Gamma-Ray Emission. Ions are simultaneously accelerated by the solar-flare process into the MeV energy range and, in exceptional cases, to the GeV energy range. As these ions travel down the magnetic flux tubes in a coronal loop to the chromosphere or even to the photosphere, there will be nuclear interactions with the ambient material producing nuclear line emission. The fragments of the nuclear interactions are highly radioactive, generating positrons and pi mesons. There is extremely strong emission at 0.511 MeV resulting from electron–positron annihilation. Strong lines are emitted from carbon and oxygen nuclei at 4.44 and 6.13 MeV. Another prominent nuclear line at 2.223 MeV is the result of neutron capture by hydrogen. Theoretical models of the solar-flare high-energy emission are still being developed (Ramaty and Murphy, 1987; Ramaty and Mandzhavidze, 1993). Table 7 lists high-energy–photon production mechanisms.

The gamma-ray emission was initially thought to be limited to the impulsive phase of the solar flare; however, the results from the more sensitive gamma-ray experiments on both the U.S. Gamma Ray Observatory and the Russian MIR space station show that the very powerful solar flares in June 1991 had extended gamma-ray emission for several hours.

The gamma-ray continuum emission is slightly directional, reflecting its high-energy ion-interaction initiation. The high-energy ions traveling down the coronal loops will be magnetically focused and have a high probability of having a velocity component parallel to the solar surface when they interact with the ambient nuclei. The Solar Maximum Mission satellite results noted that gamma-ray emission from the solar flare on the limb of the Sun (as viewed from the Earth) was about an order of magnitude "brighter" than those from flares near the solar central meridian. There is also great variability between the power in the continuum emission and the line emission. The line emission can be observed only if it is stronger than the continuum emission; for some solar flares the electron bremsstrahlung is the dominant emission.

5.2.2.3 Solar-Flare Hard X-Ray Emission. Hard x-ray (ten to hundreds of keV) emissions are most intense during the initial part of the solar flare. Hard x-ray emission is

Table 7. High-energy–photon production mechanisms.

Emissions	Process	Observed photons
Continuum	Electron bremsstrahlung	keV to MeV
Nuclear deexicitation lines	^{4}He (α,n) ^{7}Be*	0.429 MeV
	^{4}He (α,p) ^{7}Li*	0.478 MeV
	^{20}Ne (p,p) ^{20}Ne*	1.634 MeV
	^{12}C (p,p) ^{12}C*	4.438 MeV
	^{16}O (p,p) ^{16}O*	6.129 MeV
Neutron capture line	^{1}H (n,γ) ^{2}H	2.223 MeV
Positron annihilation radiation	$e^{+} + e^{-} \rightarrow 2\gamma$	0.511 MeV
Pion (π^{0} and π^{+}) decay radiation	$\pi^{0} \rightarrow 2\gamma$ $\pi^{\pm} \rightarrow \mu^{\pm} \rightarrow e^{\pm}$	10 MeV to 3 GeV

extremely impulsive, composed of bursts that range from milliseconds to seconds in duration. The hard x-ray bursts are accompanied by microwave radio and ultraviolet bursts with similar time profiles, suggesting that a common mechanism stimulates the emissions.

The hard x-rays are thought to be the result of interactions between accelerated electrons and ambient ions (bremsstrahlung) as the accelerated electrons travel down the magnetic flux tubes in a coronal loop. The hard x-ray emission energy spectra are often approximated either by a multithermal source with a peak temperature in the several hundred million degree range or by a nonthermal power-law distribution of energies with a spectral index having typical values ranging from -2 to sometimes -5. At low energies the spectrum merges with the low-energy thermal spectrum. There are large variations in the amount of hard x-ray flux emitted from different solar flares. These variations are probably the results of variations in the amount of energy involved in the various acceleration mechanisms. There is not a good correlation between solar-flare optical importance and hard x-ray emission. The correlation between solar-flare hard x-ray emission and the number of energetic particles observed in space is also poor.

5.2.2.4 Solar-Flare EUV and Soft X-Ray Emissions. The hot (multimillion degree) plasma resulting from the solar flare radiates at wavelengths corresponding to the plasma temperature. The measurement technology for soft x-ray measurements is better developed than the EUV measurement techniques, and most of the data reflect measurements in the soft x-ray wavelengths. The x-ray emission at a few kilovolts (soft x rays) increases rapidly to a maximum intensity and then decays on a longer time scale. A general classification of x-ray flares as impulsive or gradual has developed. The x-ray emission from "impulsive" solar flares is generally less than an hour in duration, having characteristic short rise times and rapid decay times. The extended x-ray emission from "gradual" solar flares may persist for several hours. These extended solar x-ray flares are often called long-duration events because of a characteristic long decay time of many hours that is suggestive of extended thermal heating of the plasma, perhaps due to magnetic reconnection as the active regions involved relax to simpler magnetic configurations. The long-duration soft x-ray–emitting flares have a good association with the observations of coronal mass ejections (CMEs) and are considered a good proxy indicator of a probable CME.

6. SOLAR PARTICLE RADIATION FROM THE ACTIVE SUN

The particle emission for the active Sun ranges from enhanced plasma emission to copious fluxes of energetic ions. Both types of emissions have effects on the Earth's atmosphere and human activities. When the solar plasma interacts with the Earth's magnetosphere, the additional energy input causes perturbations that affect communications, navigation, and even our satellites and power grids. The energetic ions affect communications in the Earth's polar regions and are operational constraints on human activities in space.

6.1 Solar Plasma Emissions

It was initially assumed that transient, fast, dense plasma emissions would be associated with the solar-flare activity; however, these associations proved to be tentative at best. Most solar flares (brightenings observable in the Hα emission) have no observable effect in the solar wind. Only the "big flares," those that are conspicuous by emissions in all electromagnetic wavelengths and associated with energetic particles, can be related to transients in the solar wind. Beginning with the Skylab mission in 1972, a new solar phenomenon was identified, the coronal mass ejection. Rather than association with "explosive" energy releases on the Sun, these emissions were found to be related to large-scale changes in solar features, such as the "disappearing filament" or erupting filament and streamer "blow outs." The total energy involved in a very large mass ejection is equivalent to and may exceed the energy released in a large solar flare.

6.1.1 Coronal Mass Ejections There is a complete spectrum of mass motions on the Sun. The small-scale mass motions may contribute to coronal heating and provide additional mass to the solar wind. The large-scale CMEs result in significant restructuring of the solar corona and may involve up to 10^{16} g of material. The large-scale CMEs are self-contained structures of plasma and embedded magnetic field ejected into space, which perturb the structure of the interplanetary medium. Approximately 30% of all CMEs have an internal field topology characteristic of twisted flux ropes (Gosling, 1990, 1993) and are commonly called "magnetic clouds." There is a continuum in the size and speed of coronal mass ejections ranging over several orders of magnitude. The speed of the leading edges of CMEs observed by the Solar Maximum Mission satellite ranged from 10 to 2100 km s^{-1}. Most coronal mass ejections have speeds comparable with the solar-wind speeds. Only about one-third have a speed sufficiently in excess of the solar wind that they can drive an interplanetary shock. Fast coronal mass ejections generate magnetohydrodynamic (MHD) shock waves as they move through the interplanetary medium at speeds faster than the local Alfvén speed. This shock wave imparts energy into the interplanetary medium and is capable of accelerating electrons and ions to higher energies.

Most coronal mass ejections arise from initially large-scale closed structures, with many from preexisting coronal streamers. However, more than one-third of the observed CMEs cannot be associated with any activity observable on the solar surface. The latitudinal distributions of CMEs is similar to that of the streamers and prominences, principally at low latitudes near the "streamer belt" near the solar-activity cycle minimum and becoming very broad near the solar-activity cycle maximum. The occurrence frequency of coronal mass ejections tracks the solar-activity cycle in both phase and amplitude. During solar minimum the average CME rate is about 0.25 per day; during solar maximum the average CME rate is about 2.5 per day (Webb and Howard, 1994; Webb, 1995). It is estimated that during the maximum of the solar-activity cycle, the coronal mass ejections may contribute about 15% of the mass flux in the near-ecliptic solar wind.

Large coronal mass ejections may be observed as density enhancements against the plane of the sky by space-borne instruments that observe the "Thomson scattered" light from density structures. State-of-the-art ground-based coronagraphs are also capable of observing coronal mass ejections but over a much smaller density range than the space-borne instruments. Unfortunately, these space-borne instruments are technically demanding and extremely expensive, and only a few have been orbited. Figure 12 shows a large coronal mass ejection observed on 24 October 1989. This very fast coronal mass ejection was associated with a large solar flare and high-energy solar-particle event.

6.2 Solar Energetic Particles

There has been a dramatic change in perspective regarding the source of energetic-particle events. Since the large transient increases in particle flux could often be associated with solar flares, it was assumed that the solar-flare process was the source of energetic particles observed in space. The names "solar cosmic rays" and "solar energetic particles" were commonly used to describe transient increases in particle flux. Papers written in the 1950s through the 1980s

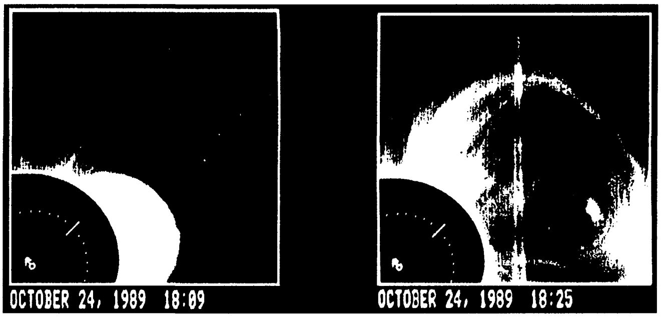

FIG. 12. The coronal mass ejection observed on 24 October 1989 by the SMM coronograph. The dotted circle indicates the visible Sun; the arrow indicates the direction of the solar rotation axis (courtesy of A. Hunthausen, HAO).

reflect the assumption that solar energetic particles were accelerated from the available material in or above the solar-active region in which the flare occurred. As measurement techniques improved, the elemental and isotopic composition of solar particles observed at the various interplanetary spacecraft was found to be consistent with the particles having passed through less than 30 mg cm^{-2} of matter from the acceleration site to their detection location. Thus, the solar photosphere and chromosphere were eliminated as the source region of the particles observed in space since the elemental and isotopic composition of these solar energetic ions did not appear to have undergone fragmentation due to interaction with the significant mass of the solar atmosphere.

A new paradigm suggests that a source of the energetic ions observed in space is the result of acceleration associated with interplanetary shocks generated by fast coronal mass ejections (see Reames, 1995, for a review). It is supposed that a shock wave has accelerated some fraction of the ions in the local solar wind. The diffusive shock acceleration scenario (Lee, 1983), a concept where particles near the shock are scattered back and forth across the shock by waves, gaining energy each time, is favored as a mechanism for continuous acceleration of particles as the shock travels out from the Sun. The accelerated particles stream away from the shock along the interplanetary magnetic field lines. The particle intensity–time profile will depend on how the observer is connected to the shock via the interplanetary magnetic field. If the shock approaches the observer, the maximum low-energy flux will be measured as the shock passes the observer.

Both the solar-flare scenario and the shock-acceleration scenario have enthusiastic advocates. Many of the "large" solar-particle events are associated with "large" solar flares. However, only about $\frac{1}{2}$ of the solar-proton events observed at the Earth can be confidently time associated with specific solar flares. The intensity–time profiles of solar energetic particles leave no doubt that interplanetary shocks also accelerate ions.

Solar-particle events are described by their flux, fluence, and spectra. Solar-particle flux is a measure of particle intensity, which can be either an integral flux above a specified energy level, in units of particles $(\mathrm{cm}^2 \cdot \mathrm{s} \cdot \mathrm{sr})^{-1}$, or a differential measurement, which specifies the flux at a specific energy in units of particles $(\mathrm{cm}^2 \cdot \mathrm{s} \cdot \mathrm{sr} \cdot \mathrm{MeV})^{-1}$. Omnidirectional particle fluxes are given in units of $(\mathrm{cm}^2 \cdot \mathrm{s})^{-1}$. Solar-particle fluence is the total number of particles experienced throughout an individual solar-particle event or episode. Fluence may be given in either directional units of particles $(\mathrm{cm}^2 \cdot \mathrm{sr})^{-1}$ or omnidirectional units of particles (cm^{-2}). The fluence is generally of concern for the total radiation exposure. The solar-particle flux (or fluence) as a function of energy is described by the particle spectrum. Many different spectral forms are used. The most commonly used spectral forms are a power law in energy or rigidity (rigidity is momentum per unit charge), an exponential form, or more recently the use of Bessel functions.

Energetic solar-particle events occur in many sizes and on many time scales. Several different classification systems can be used to group different types of events. A relatively recent grouping is to classify them as "impulsive" or "gradual" (a classification of the associated solar flare x-ray activity), as shown in Table 8. The "impulsive" events are generally small particle-flux and fluence events. Some of the "impulsive" events, particularly the ^{3}He- and Fe-rich events, seem to be directly associated with solar flares. Gradual events tend to be larger flux and fluence events with "normal" composition and are associated with coronal mass ejections and/or large, gradual, long-duration x-ray–emitting solar flares.

6.2.1 Propagation Effects in the Interplanetary Medium The transport of solar protons in interplanetary space is controlled by the topology and characteristics of the interplanetary magnetic field. The topology of the magnetic field lines in interplanetary space is controlled by the flow speed of the ionized plasma and the rotation rate of the sun, resulting in the so-called Archimedean-spiral configuration (see Fig. 6). Solar-particle events observed in space resulting from solar activity at the probable "foot point" of an idealized Archimedean-spiral interplanetary magnetic field line leading from the observer back to the Sun are considered to be "well-connected" events. This means that the probable propagation path of the solar par-

Table 8. Properties of impulsive and gradual solar-proton events (adapted from Reames, 1995).

	Impulsive particle event characteristics	Gradual particle event characteristics
Particles	Electron rich	Proton rich
^{3}He/^{4}He	~1	~0.0005
Fe/O	~1	~0.1
H/He	~10	~100
Charge state of iron	~20	~14
Duration	Hours	Days
Heliolongitude range	<30° (generally "well connected")	~180°
Associated solar radio type	III, V, (II)	II, IV
Associated solar x-ray emission	Impulsive	Gradual
CME association	. . .	96%
Associated interplanetary shock	. . .	Yes
Events per year (at solar maximum)	~1000	~10

ticles was probably along the interplanetary magnetic field lines with minimal scattering due to irregularities in the interplanetary magnetic field. This concept is often referred to as the "favorable propagation path." It is often observed that the highest particle-flux intensities are along this favorable propagation path.

The particle-flux longitudinal gradients in the inner heliosphere are variable, and local interplanetary conditions and structures greatly influence the observed time–intensity profile. The average heliolongitudinal gradient of particle flux is difficult to determine; however, the longitudinal proton-flux gradient used in the U.S. Air Force proton-prediction system is one order of magnitude decrease in flux per radian of heliolongitudinal distance away from the most favorable propagation path (Smart and Shea, 1985).

There is a radial gradient in the spacecraft observation of energetic-particle flux at different distances from the Sun, which is most pronounced in the inner heliosphere. These gradients are the result of an initial near-Sun particle injection and particle propagation away from the Sun along the interplanetary field lines. The most extensive measurements are comparisons between the flux measured by Earth-orbiting spacecraft and interplanetary spacecraft with orbits closer to the Sun than 1 AU (such as the HELIOS spacecraft). There are fewer data available between 1 and 3 AU. The form and magnitude of these radial gradients are the subject of controversy, reflecting the arguments between a solar-flare acceleration source and a CME/shock acceleration source. In the outer heliosphere, at distances of 5 AU and beyond, the particle-flux profiles are dominated by the major interplanetary shock structures.

If it is assumed

1. that the maximum possible prompt solar-proton flux is at the position that is "well connected" to the particle source acceleration region,
2. that diffusion across the interplanetary magnetic field lines is negligible, and
3. that the volume of the magnetic flux tube expands in the manner expected from classical geometry as the distance from the Sun increases,

then a power-law function of the form R^{-3} can be used to extrapolate proton flux at 1 AU to other distances. In the energy range of 10–70 MeV and for radial distances from 1 to 5 AU, the average solar-proton radial gradient derived by Hamilton (1977), considering the probable effects of diffusion, has a functional form of $R^{-3.3}$, where R is the radial distance from the Sun. The limited measurements available suggest solar-proton radial gradient variations ranging from R^{-3} to R^{-4}. For distances less than 1 AU, the proton flux at 1 AU seems to extrapolate with a functional form of R^{-3}. Again, the limited measurements available suggest that radial gradient variations ranging from R^{-3} to R^{-2} should be expected. When extrapolating proton fluence from 1 AU to other distances in the heliosphere, a functional form of $R^{-2.5}$ has been recommended with variations expected that may range from R^{-3} to R^{-2} (Feynman and Gabriel, 1988).

6.2.2 Solar-Electron Events Solar-electron events in space are now recognized to be a common result of solar activity. Solar-electron events were not identified until the beginning of the space era. Systematic identification of solar-electron events began in 1965 when space instrumentation was developed to measure specifically these events. The electron fluxes in space do not constitute a significant source of radiation dose in comparison with the energetic-ion events. Solar-electron events are often used as diagnostics of solar activity or "tracers" of the interplanetary magnetic field topology since their primary motion is essentially confined along the interplanetary magnetic field lines.

Solar-electron events were initially classified into two groups: nonrelativistic solar-electron events, which are the most common, and relativistic solar-electron events. Some events were called "pure" electron events since they did not have a significant ionic component. A more recent grouping is to classify the solar-electron events by the associated soft x-ray emission into impulsive and long-duration (sometimes called gradual) events.

Many of the solar-electron events observed in space do not have an association with visible solar-flare activity; however, there is an excellent correlation between the observation of solar-electron events in space and the occurrence of a type III solar radio burst. There is a very good correlation between large solar x-ray events and the observation of energetic electrons in space at positions that are "well connected" via the interplanetary magnetic field to the solar-flare position on the solar disk.

A comprehensive systematic survey of electron data acquired by the ISEE-3 and HELIOS 1 and 2 spacecraft (Moses *et al.*, 1989) found that the spectral characteristics of solar-electron events could be organized by the type of associated solar-flare soft x-ray event (impulsive or long duration). Solar-electron events associated with long-duration x-ray events (LDE) had a characteristic particle spectrum typified by a power law in momentum with a characteristic spectral index ranging in magnitude between 3 and 4.5. (For electrons, except at low energies, the energy spectrum and the rigidity spectrum are essentially the same.) The solar-electron events associated with impulsive solar x-ray events have a characteristic double power-law spectrum with a spectral index that hardens (the magnitude of the absolute value of the spectral index decreases) with rigidity (or energy). The transition range in these broken power-law spectra generally occurs between 100 and 200 keV, with a spectral index ranging in magnitude from 0.6 to 2 in the low-energy region of the spectra and a spectral index ranging in magnitude between 2.4 and 4.3 in the high-energy range of the spectra.

The characteristics of the electron spectra observed on widely separated spacecraft were found to be similar. After adjustment for flux amplitude, which is a function of relative connection to the solar activity source, the same characteristics in spectral shape are found. The long-duration events have the spectral form of a power law in momentum, and the impulsive events have the double power law even at widely separated heliolongitudinal distances.

The number of solar-electron events in a specific energy range can be described by a power-law distribution. Cliver *et al.* (1991) found that in the interval from 1973 to 1989, the number of solar-electron events in the 3.6–18-MeV energy range and differential flux range of 10^{-2}–10^{2} $(\text{cm}^2 \cdot \text{s} \cdot \text{sr} \cdot \text{MeV})^{-1}$ exceeding a specific threshold could be approximated by a power law with a slope of about -1.3.

6.2.3 Solar-Proton Events It is common to classify energetic solar-proton flux by the observed proton flux and fluence at a specific energy. Proton events in the energy range between 10 and 30 MeV are often used, both for traditional historical reasons and because there are enough statistics to be meaningful. There is a general association of proton-event frequencies observed at the Earth with the solar-activity cycle, but there is no discernible pattern in the proton event occurrence. (See Shea and Smart, 1990, for solar cycles 19–21.) However, the expected solar-proton-event flux and fluence frequency distribution is important for planning space missions, and so there is considerable interest in this type of data. Table 9 gives a summary of the solar-proton fluences observed at the Earth for solar cycles 19–21. Table 10 presents a summary of the largest events that have occurred during solar cycle 22 (not

Table 9. Summary of solar-proton fluence for solar cycles 19–21.

Cycle No.	Start	End	No. of months in cycle	No. of discrete proton[a] events	No. of discrete proton[a] producing regions	Solar cycle integrated solar-proton fluence	
						>10 MeV (cm^{-2})	>30 MeV (cm^{-2})
19	May 54	Oct 64	126	65	47	7.2×10^{10}	1.8×10^{10}
20	Nov 64	Jun 76	140	72	56	2.2×10^{10}	6.9×10^{9}
21	July 76	Sept 86	123	81	57	1.8×10^{10}	2.8×10^{9}

[a] Proton events with $\geq$10 MeV peak flux $\geq 10\ (\text{cm}^2 \cdot \text{s} \cdot \text{sr})^{-1}$.

yet completed as of the preparation of this article).

It had been previously suggested, as a result of the analysis of the solar-proton events observed in solar cycle 20, that there were "ordinary" and "anomalously large" events. However, analysis of a larger data sample of all the solar-proton events in solar cycles 19–21 shows that the proton-event fluence distribution observed at the Earth during the "active years" of the solar cycle is well approximated by a log-normal function (Feynman *et al.*, 1990). Including the solar-proton events of solar cycle 22 improves the statistics and reinforces the utility of the log-normal approximation (Feynman *et al.*, 1993). These data can be used to estimate the probable free-space exposure to solar protons for various space missions beyond the Earth's geomagnetic shield (including geosynchronous orbit). The probable exposure of a space vehicle at 1.0 AU to solar protons with energy >30 MeV during the solar-active years for various mission times is given in Fig. 13.

The number of solar-proton events in a specific energy range can be described by a power-law distribution. The distribution of the most commonly occurring solar-proton events exceeding a specific threshold in the 24–40-MeV energy range can be approximated by a power law with a slope of about -1.1 over the differential flux range between 10^{-3} and $10^{2}\ (\text{cm}^2 \cdot \text{s} \cdot \text{sr} \cdot \text{MeV})^{-1}$ (van Hollebeke *et al.*, 1975; Cliver *et al.*, 1991). The occurrence of very large solar-proton events appears to be less frequent than would be expected from an extension of the most commonly occurring event frequency. When large data samples are analyzed, there appears to be a broken power-law distribution. The distribution of the number of solar-proton events with energies $\geq$10 MeV expected during a solar cycle is given in Fig. 14. These

Table 10. Summary of the largest solar-proton fluence occurrences for solar cycle 22.

Cycle No.	Yr	Begin		End		Event-integrated solar-proton fluence	
		Mo	Dy	Mo	Dy	>10 MeV (cm^2)	>30 MeV (cm^2)
22	89	03	09	03	24	0.12×10^{10}	0.03×10^{9}
	89	08	12	08	18	0.76×10^{10}	1.4×10^{9}
	89	09	29	10	02	0.38×10^{10}	1.4×10^{9}
	89	10	19	10	30	1.9×10^{10}	4.2×10^{9}
	89	11	27	12	03	0.51×10^{10}	0.16×10^{9}
	89	12	30	01	02 (90)	0.21×10^{10}	0.13×10^{9}
	90	05	21	05	31	0.035×10^{10}	0.14×10^{9}
	91	03	22	03	26	0.96×10^{10}	1.8×10^{9}
	91	06	04	06	21	0.32×10^{10}	0.79×10^{9}
	91	07	07	07	12	0.11×10^{10}	0.01×10^{9}
	92	05	09	05	11	0.066×10^{10}	0.02×10^{9}
	92	10	30	11	05	0.35×10^{10}	0.43×10^{9}
	94	02	20	02	22	0.099×10^{10}	0.02×10^{9}
Totals:						5.82×10^{10}	1.0×10^{10}

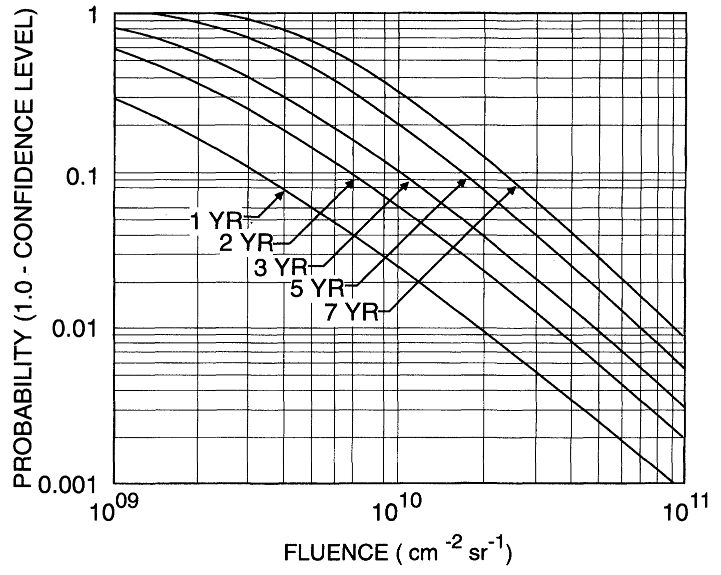

FIG. 13. The probable exposure to protons with energy >30 MeV during the solar-active years for various exposure times in interplanetary space at 1 AU. (Original figure from Feynmann *et al.*, 1993, reproduced by permission of the American Geophysical Union.)

data represent the average of ≥10-MeV proton events with peak flux ≥10 $(\mathrm{cm}^2 \cdot \mathrm{s} \cdot \mathrm{sr})^{-1}$ measured by Earth-orbiting satellite instruments during solar cycles 20–22. The increasing slope of the curve at very large flux events suggests that there may be limits to the acceleration processes, and extraordinarily large events are very rare. The analysis of total solar-proton fluence data also noted the reduced number of very large events, $>10^{10}$

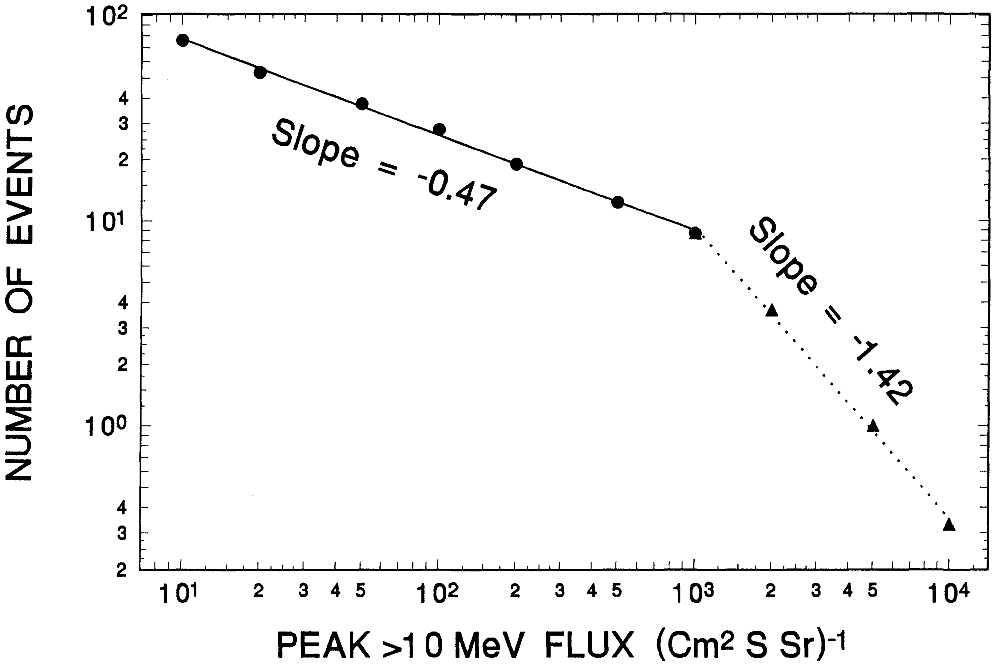

FIG. 14. The expected number of proton events during a solar cycle with protons ≥10 MeV exceeding a specified peak flux threshold. Data from solar cycles 20–22.

cm^{-2}, first noted by Lingenfelter and Hudson (1980) and subsequently confirmed by later investigators.

There is no universally accepted solar-proton classification system. The solar-proton event classification system introduced by Smart and Shea (1971) separated the >10-MeV peak flux in decade groups. A statistical treatment of the ≥30-MeV fluence data by Nymmik (1993) suggests a fluence classification based on groups separated by one standard deviation in the log-normal distribution of fluence (see Table 11) where the average solar-proton-event fluence occurrence is "normal." Stassinopoulos *et al.* (1996) adopted an order-of-magnitude classification system for convenience.

Table 11. The Nymmik solar-proton fluence classification.

Class	$n\sigma^a$	Omnidirectional ≥30-MeV fluence (cm^{-2})
Small	Average − 1σ	~2 × 10^6
Normal	Average	~9 × 10^6
Large	Average + 1σ	~3 × 10^7
Very large	Average + 2σ	~2 × 10^8
Extremely large	Average + 3σ	~5 × 10^9

[a] σ is one standard deviation of the log-normal frequency distribution.

6.2.4 Solar-Particle Event Composition

The composition of "small" solar-particle events is variable and has been separated into distinct composition groups. From an astrophysical point of view, the "iron-rich" and ^{3}He-rich events are extremely interesting, perhaps reflecting processes in the solar corona, but the particle flux in these events is small. The composition of "large" solar-particle events is relatively consistent with an ion selection process based on the first ionization potential of the elements in the solar corona or solar wind.

6.2.4.1 Large Solar-Particle Events. Large solar-particle events tend to have a "normal" composition that is relatively consistent, although there are event-to-event variations, particularly in the helium/hydrogen ratios. The ratio of the abundances of the elemental composition observed in large solar energetic-particle events to the abundances of the elemental composition observed in the photosphere is a function of the first ionization potential. Some researchers contend that large solar energetic-particle events are a reasonable sample of the solar corona (Breneman and Stone, 1985; Mewaldt and Stone, 1989); however, there are relatively large uncertainties in coronal abundance measurements. The elemental composition of large solar-particle events is given in Table 12.

6.2.4.2 Small and Iron-Rich Solar-Particle Events. When space particle-detector technology developed the capability to determine the composition of "small events," the initial results were regarded as "composition anomalies" since there were distinct differences from "large" solar-particle events. It became common to classify these as ^{3}He-rich or iron-rich events. A more recent classification is to consider the class of particle emission associated with the particle source: "impulsive-flare" events and "gradual events." In general, the fluxes in the impulsive events are orders of magnitude smaller than large or ordinary solar-particle events. The elemental composition of impulsive-flare–associated solar-particle events is detailed in Table 13.

6.2.4.3 Charge State of Solar Particles. The charge state of solar energetic ions provides information on their origin and also affects their propagation through interplanetary space since the charge state determines the magnetic rigidity (momentum per charge) and the particle scattering from magnetic irregularities in space as a function of the particle rigidity. Partially ionized solar energetic ions are more "rigid" (less susceptible to geomagnetic bending) and capable of greater penetration into the Earth's magnetic field, an important effect in radiation-dose calculations. In the early decades of solar particle measurements, it was assumed that the energetic ions would be fully stripped of electrons. As space instrument technology improved and direct ionic-charge measurements became possible, it was found that the elemental nuclei with $Z > 6$ were not fully stripped of electrons. Table 14, adapted from Luhn *et al.* (1985), gives the mean charge states for 12 solar-particle events observed with the ISEE-3 spacecraft between September 1978 and September 1979. These events were specifically selected to exclude fluxes

Table 12. Composition of "normal" solar-particle events (adapted from Reames, 1992).

Element	Z	Mean solar particle (O = 1000)	Mean solar corona (Si = 1000)	Accepted solar photosphere (Si = 1000)
H	1	1 170 000 ± 89 000	7 550 000 ± 570 000	27 900 000
He	2	55 000 ± 3 000	355 000 ± 19 000	2 270 000
C	6	471 ± 14	3 040 ± 90	10 100
N	7	128 ± 4	826 ± 26	3 130
O	8	1 000 ± 30	6 450 ± 190	23 800
Ne	10	151 ± 5	974 ± 32	3 440
Na	11	11.8 ± 1.1	76 ± 7	57.4
Mg	12	203 ± 8	1 310 ± 52	1 074
Al	13	14.0 ± 0.7	90.0 ± 5	84.9
Si	14	155 ± 7	1 000 ± 45	1 000
P	15	0.74 ± 0.11	4.7 ± 0.7	10.4
S	16	35.6 ± 1.1	230 ± 7	515
Cl	17	0.33 ± 0.11	2.1 ± 0.7	5.24
Ar	18	3.3 ± 0.5	21 ± 3	101
K	19	0.53 ± 0.24	3.3 ± 1.5	3.77
Ca	20	12 ± 1.6	77 ± 10	61.1
Ti	22	0.61 ± 0.16	3.8 ± 1.1	2.4
Cr	24	2.3 ± 0.4	15 ± 3	13.5
Fe	26	155 ± 15	1 000 ± 100	900

associated with the passage of interplanetary shocks by the spacecraft.

The charge state of energetic ions is a particularly difficult measurement, and the charge states are generally determined at relatively low energies, typically of the order of 1 MeV or less. The initial measurements have been verified by subsequent independent measurements, and the charge states have been measured at energies of tens of MeV (Oetliker *et al.*, 1995). These higher-energy charge states are generally consistent with the low-energy charge states, except for some unresolved ambiguity in the charge state of iron. The general view is that the charge state of the elements in large events is consistent with coronal temperatures ($\sim 10^6$-K plasma). A recent complex analysis found an ionic charge state of 12 for iron nuclei at ~200 MeV per nucleon. These results were from the analysis of the solid-state nuclear track detectors on the Long Duration Exposure Facility (LDEF) in 28° low altitude Earth orbit during the September and October 1989 large energetic solar-particle events.

The charge state of iron is particularly sensitive to the temperature of the plasma from which the ion is accelerated. In the "gradual" events, the typical charge state of iron seems to be $Q = 14$ and is consistent with the charge state of iron in the solar corona and in the solar wind. In the particle events associated with small impulsive solar flares, the mean charge state of iron is reported to be $Q = 20$, which is consistent with the ions having a source in a hot ($\sim 2 \times 10^7$ K) plasma.

Table 13. Averaged elemental abundances in impulsive-flare associated solar-particle events (adapted from Reames, 1992).

Element	Z	Averaged impulsive-flare (O = 1000)
^{4}He	2	46 000 ± 4 000
C	6	406 ± 18
N	7	172 ± 11
O	8	1000 ± 27
Ne	10	453 ± 18
Mg	12	478 ± 18
Si	14	425 ± 17
S	16	171 ± 11
Ca	20	78 ± 14
Fe	26	1025 ± 30

Table 14. Charge state of "normal" solar-particle events.

Element	Atomic number	Measured mean charge state (Q)
C	6	5.7
N	7	6.3
O	8	7.0
Ne	10	9
Mg	11	10.7
Si	14	11.0
S	16	10.9
Fe	26	14.9

7. TERRESTRIAL EFFECTS FROM TRANSIENT SOLAR EMISSIONS

The consequences of the solar emissions on the Earth have been noted for centuries, although it has not always been recognized that the Sun was the initial source of the energy manifestation. Auroras have been observed in northern Europe throughout recorded history. The effects of magnetic storms on navigation have been noted since the 18th century. As technology grew more complex, the influence of the Sun on this technology became more apparent. In the 19th century, the effects on long telegraph lines were noted. In the 20th century, effects on radio communications, satellite operations, and power grids have been experienced. Figure 15 gives the time scales of solar effects on the Earth's environment.

7.1 Communication Effects

One of the most immediate effects of solar flares is the radio emission that may cause interference throughout the electromagnetic spectrum, effectively jamming radios and radars. The solar x-ray event energy deposition increases the electron density in the ionosphere, which for major perturbations can suddenly disrupt high-frequency (HF) propagation paths (see also IONOSPHERE). These sudden ionospheric disturbances (SIDs) affect HF radio point-to-point communication over the entire sunlit hemisphere, often inconveniencing HF communication users. In some cases, such as aircraft operations, the loss of communication may present control difficulties.

7.1.1 Polar Communication Effects

As the solar energetic charged particles initially accelerated as a result of solar activity travel along the interplanetary magnetic field lines and intercept the Earth, additional energy is deposited in the polar atmosphere, increasing the electron density in the polar ionosphere. In the case of major solar-particle events, the dramatic enhancement in the electron density in the *D* region of the

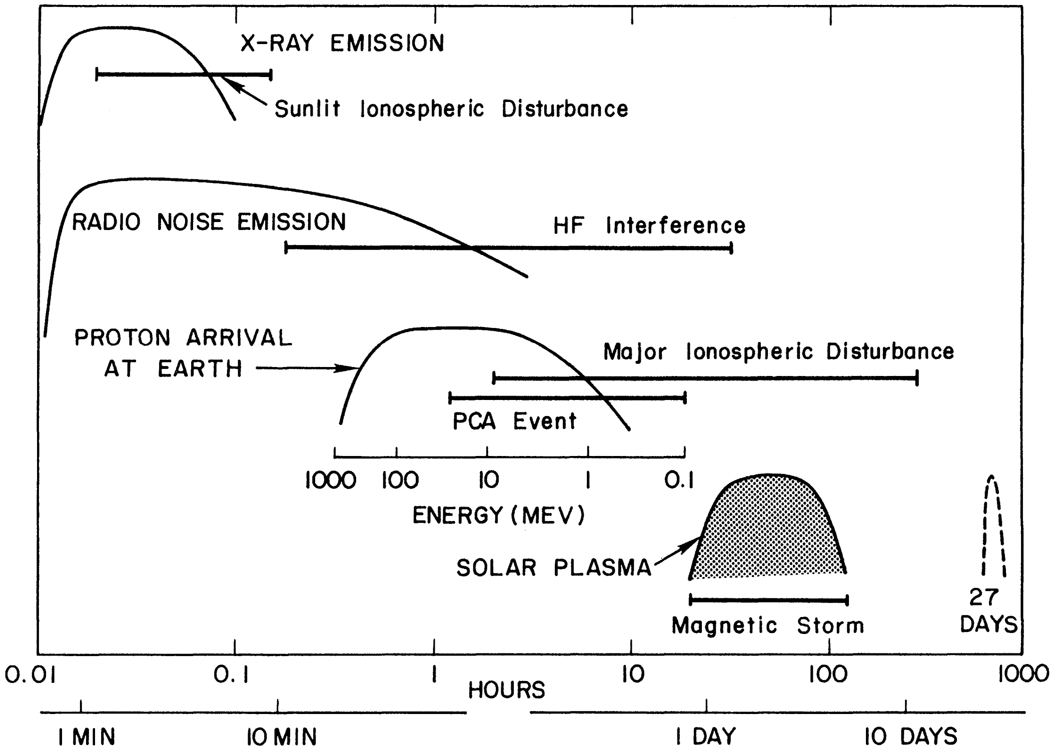
X-RAY EMISSION
Sunlit Ionospheric Disturbance
RADIO NOISE EMISSION
HF Interference
PROTON ARRIVAL AT EARTH
Major Ionospheric Disturbance
PCA Event
1000 100 10 1 0.1
ENERGY (MEV)
SOLAR PLASMA
Magnetic Storm
27 DAYS
0.01
0.1
1
HOURS
10
100
1000
1 MIN
10 MIN
1 DAY
10 DAYS

FIG. 15. The time scale of various effects on the Earth. The effects are offset vertically for clarity. (See text for definitions.) The dashed feature at 27 days is indicative of the 27-day recurrence pattern in many solar phenomena affecting the Earth.

Earth's polar ionosphere attenuates radio signals traversing the ionosphere, leading to a phenomena known as polar cap absorption (PCA) events. Analysis of the PCA events recorded in the 1950s, 1960s, and 1970s aided in the determination of the solar-particle radiation levels in these time periods. On rare occasions, the solar-particle flux was so intense that signals from ionosondes used to determine the structure of the Earth's ionosphere were completely absorbed, causing a phenomenon historically known as polar cap blackout. In extreme cases, the polar cap blackout continued for several days during which time point-to-point HF radio communications were essentially completely disrupted. Even with more modern VLF long-range navigation systems, such as the LORAN system, during a solar-particle event (PCA event) the increased ionization in the polar ionosphere modifies the characteristic of the Earth–ionosphere waveguide, resulting in changes in the phase and amplitude of received signals that translate into navigation errors.

7.1.2 Aurora The polar aurora is a consequence of energy initially emitted from the Sun via the solar wind (including the coronal mass ejection) entering the Earth's magnetosphere and, through complex magnetospheric processes, eventually resulting in the local acceleration of electrons that travel down geomagnetic field lines, depositing sufficient energy into the atmosphere so as to stimulate emission from atomic oxygen atoms. (The common auroral emissions are at 557.7 and 630.0 nm, which appear as green and red when the intensity exceeds the color threshold of the human eye.) This locally intense energy deposition again increases the local electron density in the Earth's ionosphere and can interfere with HF point-to-point communication in the high-latitude auroral regions. During large magnetic storms, the aurora moves toward the equator, on occasions reaching the Earth's midlatitudes, transporting what are normally polar-region effects to these latitudes.

7.2 Geomagnetic-Field Variations

The geomagnetic storm, historically named because of variations induced in the Earth's magnetic field, is both a direct effect and an indicator of complex processes in the Earth's magnetosphere resulting from solar activity, such as the coronal mass ejection. The variations induced in the Earth's magnetic field are maximum in the polar regions (up to several thousand nanotesla variation in an average 60 000-nT surface polar magnetic field) and minimum at the Earth's equator (extreme variations to perhaps 400 nT in an average 30 000-nT equatorial surface magnetic field). These variations affect the direction indicated by magnetic compasses, a fact that has been a technological driver leading to the development of other types of navigation systems that might be immune to these effects like the LORAN system and more recent space navigation systems such as the Transit satellites and the Global Positioning satellites. A more recently noticed effect on human activity is the action of induced currents, resulting from extremely intense and rapid magnetic field variations, exceeding the design capacity of power grids, thereby damaging equipment or, in more spectacular cases, causing the collapse of a major power grid such as occurred in Quebec, Canada, in March 1989.

7.3 Atmospheric-Density Effects on Satellite Operations

The density at a specific height in the upper atmosphere changes during a solar cycle as the atmosphere expands and contracts because of variations in the solar EUV flux that heats the upper atmosphere. The upper-atmosphere density controls the lifetime of orbiting satellites, since even these very low densities result in slight drag forces that limit the lifetimes of orbiting objects. Predictions of satellite orbits such as the U.S. space shuttle or the MIR space station are extremely dependent on our ability to specify the atmospheric density. Low-altitude Earth-orbiting satellites must be either periodically reboosted, such as the Russian MIR space station or the U.S. Hubble Space Telescope, or replaced. Failure to reboost a decaying satellite into a higher orbit will result in reentry of the satellite in the Earth's atmosphere, such as happened to the U.S. Skylab spacecraft in July 1979.

Besides the long-term solar-cycle variations, the coronal mass ejection plasma emissions may couple a tremendous amount of energy (in human terms) into the Earth's magnetosphere. The geomagnetic storm is a result of this energy input, which initiates and intensifies current systems flowing in the magnetosphere and ionosphere. In the polar regions, the phenomenon known as Joule heating greatly increases the atmospheric density, thereby changing satellite orbits. During the major geomagnetic storms in March 1989, several hundred orbiting objects were temporarily "lost" to the NORAD catalog of Earth-orbiting objects, because of unanticipated changes in their orbits.

7.4 Solar-Particle Radiation Effects

Energetic-particle ionizing radiation has many effects on today's technology. Besides degrading polar communications as described in Sec. 7.1.1, energetic solar-particle radiation also presents operational constraints on human activities in space.

7.4.1 Radiation-Dose Effects on Solar-Cell Power Systems Many space vehicles are powered by solar cells, which convert a fraction of the Sun's electromagnetic energy into electrical power. Modern spacecraft are designed to last for many years or even decades. The cumulative radiation-dose effect on solar cells results in their becoming gradually less efficient, and eventually their power output drops below the minimum necessary to operate all of the vehicle subsystems. The total fluence from one very large solar proton event can double the cumulative exposure for a solar cycle (see Tables 9 and 10), and the output of solar-cell power systems can be significantly degraded, reducing the operational lifetime of a space vehicle by years (Marvin and Gorney, 1991).

7.4.2 Soft Errors in Electronics Modern semiconductor elements are so small (feature sizes less than a micron) and the currents in an individual element so low that the number of electrons involved in a change of state (from off to on) is of the same order of magnitude as the number of electrons generated by the passage of an energetic ion through semiconductor material. This phenomenon is called "single-event upset" or "soft error" since the device will continue to operate correctly even though the information it contains is now incorrect. Semiconductor devices that are vulnerable to these types of errors are called "soft" devices, and semiconductor elements that are immune to soft errors are called "hard" devices.

There are two primary mechanisms for causing soft errors in space-borne semiconductor electronics. Energetic heavy ions, such as ^{26}Fe, passing through the sensitive volume of a device liberate enough electrons to cause a change of state in a logic element. (Engineers use the term "linear energy transfer" to compute the probability of soft errors.) The other mechanism is that an energetic nucleus has a direct collision with an atom, and in its recoil, the energy imparted to the atom by the collision is converted to enough electrons to cause a change of state in a logic element. Selection of space-qualified parts to minimize the soft-error effect is a major challenge in space-vehicle design.

7.4.3 Solar-Particle Radiation Effects on Human Space Flight The free-space dose that might be experienced by an unshielded astronaut during the largest recorded solar-proton events would be debilitating (exceeding the threshold for inducing nausea) and perhaps potentially lethal. Studies of the probable effects of solar-particle radiation of space mission scenarios to the Moon or Mars recommend that a location with a minimum of 20 $\mathrm{g \cdot cm^{-2}}$ shielding be included in the spacecraft to reduce the effects of large energetic solar-particle events to tolerable levels.

The orbits of the Russian MIR space station and the projected International Space Station have been selected partly to minimize the solar-particle radiation risk by taking advantage of the geomagnetic shielding effect offered by the Earth's magnetic field and still permit maximum coverage of the Earth's surface. The Russian MIR space station has the longest actual history of radiation measurements on manned spacecraft during large energetic solar-particle events. The very energetic solar-particle event of 29 September 1989 (evaluated as the third largest in the history of cosmic radiation observations) resulted in a measured radiation exposure to the Cosmonauts in the MIR space station of 0.005 Gy (0.5 REM). Following the extremely intense solar-particle–interplanetary-shock event that occurred on 20 October 1989, during the intense geomagnetic storm that reduced the normal geomagnetic shielding, the Cosmonauts in the MIR space station experienced an additional exposure of 0.015 Gy (1.5 REM).

These large energetic solar-particle events pose no significant hazard to ordinary commercial aviation at altitudes below 40 000 ft. At very high altitudes (in the 50 000–70 000-ft range) of the Concorde SST, there is a possibility that the radiation exposure could exceed the recommended levels for the general public, and these aircraft carry radiation warning devices in case of such an occurrence. Only twice in the operational history of the SST has this occurred, but radiation intensity never approached the hazard level. (On 29 September 1989, and again on 20 October 1989, low-level warnings were registered.)

8. SUMMARY

The Sun is the source of energy that permits life on Earth. It has only been during the past century or so that we have been able to identify some of the more elusive "causes and effects" of solar activity and solar emissions on our lives. Much of our knowledge emanates from our increased technology and our dependence on this technology. With progressive miniaturization of electronic components used in space and on earth, and the increased susceptibility of these components to charged particles and current systems, it is imperative that we understand solar emissions and interplanetary phenomena that perturb our spatial and terrestrial environment.

In the infrared and visible wavelengths, the solar output is relatively constant with small identifiable changes occurring during the solar-activity cycle. However, our current technology is not adequate to provide the precision we desire about the solar constant. The solar electromagnetic radiations are more variable at both ends of the solar electromagnetic emission spectrum. At long wavelengths (radio frequencies) sporadic large solar outbursts occur that are associated with solar-flare activity, the frequency of which is directly related to the solar-activity cycle. At short wavelengths, the frequency of UV and x-ray outbursts that affect the electron density in the Earth's ionosphere and hence its electromagnetic propagation properties are again related to the solar-activity cycle. Unfortunately, the solar-activity cycles have not proven to be very predictable.

It is in the solar plasma and solar-particle emission that the most significant concept changes have occurred. Solar energetic-particle emission constitutes an operational constraint on human space activities. This includes limiting the lifetime of space-vehicle solar-cell power systems and interference with semiconductor logic operations. The occurrence of these solar plasma and particle emissions is sporadic and currently not very predictable.

Research on the solar-wind plasma and embedded magnetic field continues in order to understand better this method of energy transport to the Earth's environment. The energy involved in a coronal mass ejection is of the same order of magnitude as a large solar flare, and those CMEs that are directed toward the Earth impart a large amount of energy into the Earth's magnetosphere, resulting in the terrestrial disturbances once thought to be the result of solar-flare activity. These magnetospheric perturbations, which often are manifested as geomagnetic storms, interfere with navigation systems (particularly those employing magnetic direction-finding systems for orientation) and can temporarily disrupt communication links. The energy input to the Earth's upper atmosphere can change the atmospheric density and affect satellite operations. In extreme cases, a major geomagnetic storm can interfere with and even disrupt electrical power grids.

The very nature of the effect of solar emissions on humans and their environment is worldwide in scope. In reviewing the available information for this article, it was obvious that the research results and technological advances over the past quarter century clearly emphasize the interrelationships of what were once considered separate and independent phenomena. As our technology advances we will continue to learn more about the complexities of solar emissions on the solar–terrestrial system.

GLOSSARY

Active Region (AR): A localized, transient volume of the solar atmosphere in which plages, sunspots, faculae, flares, etc. may be observed. Active regions are the result of enhanced magnetic fields; they are at least bipolar and may be complex if the region contains two or more bipolar groups.

Alfvén Wave: A transverse wave in magnetized plasma characterized by a change of direction of the magnetic field (rather than a change of intensity).

Anisotropy: The ratio of the maximum to the average particle-flux distribution as a function of angle.

Archimedean Spiral: A mathematical curve resulting from a linear angular rotation with increasing distance.

Astronomical Unit (AU): The average distance from the Sun to the Earth, 149.6×10^6 km.

Chromosphere: The layer of the solar atmosphere above the photosphere and beneath the transition region and the corona. It is the source of the strongest lines in the solar spectrum, including the Balmer alpha line of hydrogen ($H\alpha$) and the H and K lines of calcium. It is the source of the red (chromium) color often seen around the rim of the Moon at total solar eclipses.

Chromospheric Brightening: A literal description of the "solar flare" that is optically visible as a brightening of a small area of the solar chromosphere.

Corona: The outermost layer of the solar atmosphere, characterized by low densities ($<10^{-9}$ cm^{-3} or 10^{-15} m^{-3}) and high temperatures ($>10^6$ K).

Coronagraph: An optical device that makes it possible to observe the corona at times other than during an eclipse. An occulting disk blocks light from the solar disk, effectively providing an artificial eclipse.

Coronal Hole: An extended region of the corona, exceptionally low in density and associated with unipolar photospheric regions having "open" magnetic field topology. Coronal holes are largest and most stable at or near the solar poles and are a source of high-speed solar wind. Coronal holes are visible in several wavelengths. They are most notable in solar soft x rays and can be found in the He 1083-nm emission.

Coronal Loops: A typical structure of enhanced corona observed in EUV lines and soft x rays. They are sometimes related to $H\alpha$ loops. Coronal loops represent "closed" magnetic topology.

Coronal Mass Ejection (CME): A transient outflow of plasma from or through the solar corona. The CMEs are often, but not always, associated with erupting prominences, disappearing solar filaments, and flares.

Differential Particle Flux: The energetic particle flux at energy E per unit energy interval, per unit area, per unit time, per unit solid angle of observation, passing through an area perpendicular to the viewing direction. The directional differential flux of solar particles is usually specified in units of $(\text{cm}^2 \cdot \text{s} \cdot \text{sr} \cdot \text{MeV})^{-1}$. The omnidirectional differential flux is usually specified in units of $(\text{cm}^2 \cdot \text{s} \cdot \text{MeV})^{-1}$.

Differential Rotation: The change in solar rotation rate with latitude. Low latitudes rotate at a faster angular rate (~14°/day) than do high latitudes (~12°/day).

Disappearing Solar Filament (DSF): An English term for the sudden disappearance of a solar filament (prominence) on a time scale of minutes to hours. The DSFs are a possible indicator of coronal mass ejections. The proper name of the phenomenon is *disparition brusque* given by French astronomers.

Extreme Ultraviolet (EUV): The portion of the electromagnetic spectrum from approximately 10 to 100 nm.

Facula: White-light plage. A bright region of the photosphere seen in white light, seldom visible, except near the solar limb. Corresponds with concentrated magnetic fields that may precede sunspot formation.

Favorable Propagation Path: A concept suggesting that the Archimedean spiral path from the Earth to the Sun would connect to a specific solar longitude. It is based on the principle that charged particles travel along the interplanetary magnetic field lines that are transported out from the Sun. For an idealized constant-speed solar-wind flow, if the interplanetary magnetic field is "frozen" in the plasma, then the result would form an Archimedean spiral.

Fibril: A linear feature in the $H\alpha$ chromosphere of the Sun, occurring near strong sunspots and plage or in filament channels. Fibrils parallel strong magnetic fields, as if mapping the field direction.

Filament: A mass of gas suspended over the chromosphere by magnetic fields and seen as a dark ribbon threaded over the solar disk. A filament on the limb of the Sun seen in emission against the dark sky is called a prominence. Filaments occur directly over magnetic-polarity inversion lines, unless they are active.

First Ionization Potential: The energy required to remove the first electron from an electrically neutral atom. (The ionization potential is usually given in electron volts.)

Flare: See Solar Flare.

Fluence: The time-integrated flux throughout a solar-particle event or a specific time period.

Flux: The rate of particle flow through a reference surface. The directional flux is usually specified in units of $(\mathrm{cm}^2 \cdot \mathrm{s} \cdot \mathrm{sr})^{-1}$. The omnidirectional flux is usually specified in units of $(\mathrm{cm}^2 \cdot \mathrm{s})^{-1}$.

Granulation: Cellular structure of the photosphere visible at high spatial resolution. Individual granules, which represent the tops of small convection cells, are 200–2000 km in diameter and have lifetimes of 8–10 min.

Gray (Gy): The new international system unit (SI unit) of absorbed dose of radiation. $1\ \mathrm{Gy} = 1\ \mathrm{J \cdot kg^{-1}} = 100$ rad.

Hα: The first atomic transition in the hydrogen Balmer series, with wavelength 656.3 nm. This absorption line of neutral hydrogen falls in the red part of the visible spectrum and is convenient for solar observations. The Hα line is universally used for patrol observations of solar flares, filaments, prominences, and the fine structure of active regions.

Heavy Ions: Nuclei of elements such as nitrogen, carbon, oxygen, or iron that are positively charged because some or all of the planetary electrons have been stripped from them.

Heliocentric: A measurement system with its origin at the center of the Sun.

Heliocentric Angular Distance: A measurement of angular distance from the center of the Sun.

Heliographic: Referring to coordinates on the solar surface referenced to the solar rotational axis. (See **Solar Coordinates.**)

Heliosphere: The plasma domain surrounding the Sun and dominated by the solar-wind outflow. The heliospheric boundary, where the solar-wind flow is hypothesized to change from supersonic flow to subsonic flow, is estimated to be at >100 AU.

Integral Particle Flux: The number of particles of energy equal to or greater than E, per unit area, per unit solid angle, per unit time, passing through an area perpendicular to the viewing direction. The directional integral flux is specified in units of $(\mathrm{cm}^2 \cdot \mathrm{s} \cdot \mathrm{sr})^{-1}$. The omnidirectional integral flux is specified in units of $(\mathrm{cm}^2 \cdot \mathrm{s})^{-1}$.

Interplanetary Magnetic Field (IMF): The magnetic field in interplanetary space transported out from the Sun via the solar wind. The average magnitude of the IMF at 1 AU is ~5 nT.

Interplanetary Shock: An abrupt change in velocity or density that is moving faster than the wave-propagation speed in interplanetary space.

Mean Free Path: The distance between scattering events in interplanetary particle propagation.

Photosphere: The lowest visible layer of the solar atmosphere; corresponds to the solar surface viewed in white light. Sunspots and faculae are observed in the photosphere.

Plage: On the Sun, an extended emission feature of an active region that is seen from the time of emergence of the first magnetic flux until the widely scattered remnant magnetic fields merge with the background. Magnetic fields are more intense in plages, and temperatures are higher than in the surrounding, quiescent regions.

Pore: A feature of the photosphere, 1–3 arc sec in extent, usually not much darker than the dark spaces between photospheric granules. It is distinguished from a sunspot by its short lifetime, 10–100 min.

Proton Event: A generic term referring to an enhancement in the energetic-particle flux observed in space, usually as a result of solar activity.

Rigidity: The momentum per unit charge of a charged particle. Rigidity is usually specified in units of MV (10^6 V) or GV (10^9 V).

Sievert (Sv): The new international system unit (SI unit) of biologically effective radiation absorbed dose. 1 Sv = 1 Gy times dose equivalent (Q). 1 Sv = 100 REM.

Shock: A discontinuity in pressure, density, and particle velocity, propagating through a compressible fluid or plasma.

Smoothed Sunspot Number: A statistical 13-month running mean of the sunspot numbers, centered on the month of concern. The 1st and 13th months are given weights of 0.5.

Solar Coordinates: Specifications for a location on the solar surface. The location of a specific feature on the Sun (for example, a sunspot) is complicated by the fact that there is a tilt of 7.25° between the ecliptic plane and the solar equatorial plane as well as a true wobble of the solar rotational axis.

Solar Cycle: The solar cycle refers to the solar-activity cyclic behavior, usually represented by the number of sunspots visible on the solar photosphere. The average length of solar cycles since 1900 is 11.4 yr.

Solar Flare: The name given to the sudden release of energy (often more than 10^{32} ergs) in a relatively small volume of the solar atmosphere. Historically, an optical brightening in the chromosphere, now expanded to cover almost all impulsive radiation from the Sun.

Solar Flare Importance: The classification of a solar flare based on its measured optical area (measured in millionths of the solar disk) and brightness at the time of maximum brightness in the Hα line, by a number–letter combination. Solar flares are ranked in area as importance 0 (or *s*), 1, 2, 3, to a maximum of 4. Solar flares are subjectively ranked in brightness as *f* (faint), *n* (normal), or *b* (bright).

Solar Maximum: The month(s) during the sunspot cycle when the smoothed sunspot number reaches a maximum. The maximum for solar cycle 22 occurred in July 1989.

Solar Minimum: The month(s) during the sunspot cycle when the smoothed sunspot number reaches a minimum. The recent solar minimum between the 21st and 22nd solar cycles occurred in September 1986.

Solar Rotation Rate: Relative to the Earth (synodic), $(13.39° - 2.7° \sin^2\theta)$/day. Relative to an inertial system (sidereal), $(14.38° - 2.7° \sin^2\theta)$/day ($\theta$ = solar latitude). The difference between sidereal and synodic rates is the Earth orbital motion of 0.985°/day.

Solar Wind: The plasma flowing out into space from the solar corona. Typically at 1 AU, the average solar-wind velocity is ~ 375 $\text{km}\cdot\text{s}^{-1}$, and the average density is $\sim 5\ \text{cm}^{-3}$.

Spicules: Rapidly changing, predominantly vertical, spikelike structures in the solar chromosphere observed about the limb. Spicules appear to be ejected from the low chromosphere at velocities of 20–30 $\text{km}\cdot\text{s}^{-1}$, reaching a height of about 9000 km and then falling back or fading. The total lifetime is 5–10 min.

Spray: Luminous material ejected from a solar flare with sufficient velocity to escape the Sun (675 $\text{km}\cdot\text{s}^{-1}$). There is little evidence that sprays are focused by magnetic fields. Compare with **Surge.**

Subsonic Flow: Plasma flowing slower than the local speed of sound in the medium.

Sunspot: An area seen as a dark spot, in contrast with its surroundings, on the photosphere of the Sun. Sunspots are concentrations of magnetic flux, typically occurring in bipolar clusters or groups. They appear dark because they are cooler than the surrounding photosphere. Larger and darker sunspots sometimes are surrounded (completely or partially) by penumbrae. The dark centers are umbrae. The smallest, immature spots are sometimes called pores.

Sunspot Cycle: The approximately 11 yr quasiperiodic variation in the sunspot number. Other solar phenomena, such as the 10.7-cm solar radio emission, exhibit similar cyclical behavior. The polarity of the solar magnetic field reverses during the maximum phase of each cycle.

Sunspot Number: A quantification of the number of sunspots and sunspot groups visible on the Sun according to the International Astronomical Union (IAU) definitions of sunspot measurement. This number is often used as a proxy of solar activity. The daily sunspot index is identified as *A*. The sunspot index *R* is defined as $R = k(10g + s)$, where s = number of individual spots meeting the IAU criterion, g = number of sunspot groups, and k is an observatory and/or observer factor. The standard sunspot number was for many years derived at Zürich and was referred to as R_z. It is now being derived at Brussels and is called the International sunspot number R_I. Often, the term "sunspot number" is used in reference to the widely distributed smoothed sunspot number.

Supergranulation: A system of large-scale velocity cells that does not vary significantly over the quiet solar surface or with phases of the solar cycle. The cells are presumably convective in origin with weak upward motions in the center, downward motions at the borders, and horizontal motions of typically 0.3 to 0.4 km s^{-1}. Magnetic flux is more intense along the borders of the cells.

Supersonic Flow: Plasma flowing faster than the local speed of sound in the medium.

Surge: A jet of material from active regions that reaches coronal heights and then either fades or returns into the chromosphere along the trajectory of ascent. Surges typically last 10–20 min and tend to recur at a rate of approximately 1/h. Surges are linear and collimated in form, as if highly directed by magnetic fields. Compare with **Spray.**

Ultraviolet (UV): That part of the electromagnetic spectrum between 5 and 400 nm.

Well Connected: A concept suggesting that the Archimedean spiral path from the specified position in space would connect to a specific solar longitude. It is based on the concept that charged particles travel along the interplanetary magnetic field lines that are transported out from the Sun.

X-ray Flare Class: Rank of a solar flare based on its x-ray energy output. Solar x-ray flares are classified by the NOAA Space Environment Services Center by a letter–number combination denoting the magnitude of the peak x-ray burst intensity measured at the Earth in the 0.1–0.8-nm band. See Table 6.

Works Cited

Bahcall, J. N. (1990), *Sci. Am.* **262** (5), 54–61.

Breneman, H. H., Stone, E. C. (1985), *Astrophys. J.* **299,** 959–963.

Cliver, E., Reames, D., Kahler, S., Cane, H. (1991), in: *22nd International Cosmic Ray Conference,* Vol. 3, Dublin: Dublin Inst. for Adv. Studies, pp. 25–28.

Cox, A. N., Livingston, W. C., Matthews, M. S. (Eds.) (1991), *Solar Interior and Atmosphere,* Tucson, AZ: University of Arizona Press.

Donnelly, R. P. (Ed.) (1992), *Proceedings of the Workshop on the Solar Electromagnetic Radiation Study for Solar Cycle 22,* Boulder, CO: ERL, NOAA, DOC.

Eddy, J. A. (Ed.) (1978), *The New Solar Physics,* Boulder, CO: Westview Press for the AAAS.

Feynman, J., Armstrong, T. P., Dao-Gibner, L., Silverman, S. (1990), *J. Spacecraft Rockets* **27** (4), 403–410.

Feynman, J., Gabriel, S. (Eds.) (1988), *Interplanetary Particle Environment,* JPL Publication 88-28, Pasadena, CA: Jet Propulsion Laboratory, California Institute of Technology.

Feynman, J., Spitale, G., Wang, J., Gabriel, S. (1993), *J. Geophys. Res.* **98,** 13281–13294.

Foukal, P. V. (1990), *Sci. Am.* **262** (2), 34–41.

Gloeckler, G., Geiss, J. (1989), in: C. J. Waddington (Ed.), *Cosmic Abundances of Matter,* AIP Conference Proceedings No. 183, New York: American Institute of Physics, p. 49.

Gosling, J. T. (1990), in: C. T. Russell, E. R. Priest, L. C. Lee (Eds.), *Physics of Magnetic Flux Ropes,* Geophysical Monographs No. 58, Washington, DC: American Geophysical Union, p. 343.

Gosling, J. T. (1993), *J. Geophys. Res.* **98,** 18937–18949.

Hamilton, D. C. (1977), *J. Geophys. Res.* **82,** 2159–2169.

Hinteregger, H. E., Fukui, K., Gilson, B. R. (1981), *Geophys. Res. Lett.* **8,** 1147–1150.

Hoeksema, J. T., Scherrer, P. H. (1986), *The Solar Magnetic Field—1976 through 1985,* UAG-94, Boulder, CO: WDC-A, NGDC, NOAA, DOC.

Hudson, H. S. (1988), *Annu. Rev. Astron. Astrophys.* **26,** 473–507.

Lean, J. (1991), *Rev. Geophys.* **29,** 505–535.

Lee, M. A. (1983), *J. Geophys. Res.* **88,** 6109–6120.

Lee, B. L., Gibson, A. M., Wilson, R. S., Thomas, S. (1995), *J. Geophys. Res.* **100,** 1667–1675.

Lingenfelter, R. E., Hudson, H. S. (1980), in: R. O. Pepin, J. A. Eddy, R. B. Merrill (Eds.), *Proceedings of the Conference on the Ancient Sun,* New York: Pergamon Press, pp. 69–79.

Luhn, A., Hovestadt, D., Kleckler, B., Scholer, M., Gloeckler, G., Ipavich, F. M., Galvin, A. B., Fan, C. Y., Fisk, L. A. (1985), in: *19th International Cosmic Ray Conference,* Vol. 4, Greenbelt, MD: NASA Conference Publication 2376, pp. 241–244.

Marvin, D. C., Gorney, D. J. (1991), *J. Spacecraft Rockets* **29,** 713–719.

McKinnon, J. A. (1987), *Sunspot Numbers: 1610–1985,* UAG-95, Boulder, CO: WDC-A, NGDC, NOAA, DOC.

Mewaldt, R. A., Stone, E. C. (1989), *Astrophys. J.* **377,** 959–963.

Moses, D., Dröge, W., Meyer, P., Evenson, P. (1989), *Astrophys. J.* **346,** 523–530.

Nymmik, R. A. (1993), in: *23rd International Cosmic Ray Conference,* Vol. 3, Calgary: Univ. of Calgary, pp. 29–32.

Oetliker, M., Klecker, B., Hovestadt, D., Scholer, M., Blake, J. B., Looper, M., Mewaldt, R. A. (1995), in: *24th International Cosmic Ray Conference,* Vol. 4, Rome: Universitá La Sapienza, pp. 470–474.

Parker, E. N. (1963), *Interplanetary Dynamical Processes,* New York: Wiley-Interscience.

Quarterly Bulletin on Solar Activity, Zurich: Eidgen, Strenwarte.

Ramaty, R., Mandzhavidze, N. (1993), in: J. T.

Ryan, W. T. Vestrand (Eds.), *High Energy Solar Phenomena—A New Era of Spacecraft Measurements,* AIP Conference Proceedings No. 294, New York: Institute of Physics, p. 26.

Ramaty, R., Murphy, R. J. (1987), *Space Sci. Rev.* **45,** 213–268.

Reames, D. V. (1992), in: *Proceedings of the First SOHO Workshop,* ESA Special Publication No. 348, Paris: European Space Agency, pp. 315–323.

Reames, D. V. (1995), *Rev. Geophys. Suppl.* **33,** 585–589.

Sawyer, C., Warwick, J. W., Dennett, J. T. (1986), *Solar Flare Prediction,* Boulder, CO: Colorado Associated Press.

Shea, M. A., Smart, D. F. (1990), *Sol. Phys.* **127,** 297–320.

Smart, D. F., Shea, M. A. (1971), *Sol. Phys.* **16,** 484–487.

Smart, D. F., Shea, M. A. (1985), in: A. S. Jursa (Ed.), *Handbook of Geophysics and the Space Environment,* Bedford, MA: Air Force Geophysics Laboratory, Hanscom AFB, Chap. 6.

Solar-Geophysical Data, Boulder, CO: NGDC, NOAA, DOC.

Stassinopoulos, E. G., Brucker, G. J., Nakamura, D. W., Stauffer, C. A., Gee, G. B., Barth, J. L. (1996), *IEEE Trans. Nucl. Sci.* **43,** 369–382.

Tobiska, W. K. (1993), *J. Geophys. Res.* **98,** 18879–18893.

van Hollebeke, M. A., Ma Sung, L. S., McDonald, F. B. (1975), *Sol. Phys.* **41,** 189–223.

Webb, D. F. (1995), *Rev. Geophys., Suppl.* **33,** 557–583.

Webb, D. F., Howard, R. W. (1994), *J. Geophys. Res.* **99,** 4201–4220.

White, O. R. (Ed.) (1977), *The Solar Output and Its Variations,* Boulder, CO: Colorado Associated University Press.

Zirin, H. (1988), *Astrophysics of the Sun,* London: Cambridge University Press.

Further Reading

Akasofu, S.-I., Chapman, S. (1972), *Solar-Terrestrial Physics,* London: Oxford University Press.

Bruzek, A., Durrant, C. J. (1977), *Illustrated Glossary for Solar and Solar-Terrestrial Physics,* Dordrecht/Boston: D. Reidel.

Durrant, C. J. (1988), *The Atmosphere of the Sun,* Bristol: Adam Hilger.

Foukal, P. (1990), *Solar Astrophysics,* New York: Wiley Interscience.

Hruska, J., Shea, M. A., Smart, D. F., Heckman, G. (Eds.) (1993), *Solar-Terrestrial Predictions-IV,* Boulder, CO: SEL, NOAA, DOC.

Hundhausen, A. J. (1972), *Coronal Expansion and Solar Wind,* New York: Springer-Verlag.

Kundu, M. R., Woodgate, B., Schmahl, E. J. (Eds.)(1989), *Energetic Phenomena on the Sun,* Dordrecht: Kluwer Academic Publishers.

Jursa, A. S. (Ed.) (1985), *Handbook of Geophysics and the Space Environment,* Bedford, MA: Air Force Geophysics Laboratory, Hanscom AFB.

Noyes, R. W. (1982), *The Sun, Our Star,* Cambridge MA: Harvard University Press.

Reviews of Geophysics (1987), *Contributions in Solar-Planetary Relationships,* Vol. 25 Supplement, Washington, DC: American Geophysical Union.

Reviews of Geophysics (1991), *Contributions in Solar-Planetary Relationships,* Vol. 29 Supplement, Washington, DC: American Geophysical Union.

Reviews of Geophysics (1995), *Contributions in Solar-Planetary Relationships,* Vol. 33 Supplement, Washington, DC: American Geophysical Union.

Schwenn, R., Marsch, E. (Eds.) (1990), *Physics of the Inner Heliosphere I,* New York: Springer-Verlag.

SESC Glossary of Solar-Terrestrial Terms (1990), Boulder, CO: SEL, NOAA, DOC.

Solar Cosmic Rays; Proton Fluxes Model (1986), State Standard 25645. 13-86, Moscow.

Stenflo, J. O. (Ed.) (1990), *Solar Photosphere: Structure, Convection and Magnetic Fields,* Dordrecht: Kluwer Academic Publishers.

Sturrock, P. A., Holzer, T. E., Mihalas, D. M., Ulrich, R. K. (1986), *Physics of the Sun,* Dordrecht/Boston: D. Reidel Publishing Co.

Svestka, S. (1976), *Solar Flares,* Dordrecht/Boston: D. Reidel.

Tandberg-Hanssen, E., Emslie, G. (1988), *The Physics of Solar Flares,* London: Cambridge University Press.

Wentzel, D. G. (1985), *The Restless Sun,* Washington, DC: Smithsonian Institution Press.

Wilson, J. W., Townsend, L. W., Schimmerling, W., Khandelwal, G. S., Khan, F., Nealy, J. E., Cucinotta, F. A., Simonsen, L. C., Shinn, J. L., Norbury, J. W. (1991), *Transport Methods and Interactions for Space Radiations,* NASA Reference Publication 1257, Washington: U.S. GPO.

SOLID, LIQUID, AND GASEOUS ELECTRICAL INSULATION

MARKUS ZAHN, *Laboratory for Electromagnetic and Electronic Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*

Introduction 432
1. **Important Electrical Properties of Insulation** 432
2. **Representative Analytical Descriptions of Electrical Insulation** 433
2.1 Linear Lossy Dielectrics 433
2.1.1 Governing Equations 433
2.1.2 Ohmic Material 434
2.1.3 Transient Charge Relaxation 434
2.1.4 Interfacial Charge 434
2.1.4.1 Step Transient 434
2.1.4.2 Sinusoidal Steady State 435
2.2 Charge-Injection Models 436
2.2.1 Schottky Emission 436
2.2.2 Fowler–Nordheim Field Emission 436
2.3 Avalanche Modeling 436
2.4 Statistical Methods 437
2.5 Life-Prediction Techniques 438
2.6 Space-Charge Effects 438
2.6.1 Space-Charge Distortion of the Electric Field 438
2.6.2 Governing Equations for Drift-Dominated Unipolar Conduction 439
2.6.3 dc Steady-State Distributions 440
2.6.4 Space-Charge Effects in Pulse-Power Technology 441
2.7 Drift-Diffusion Conduction 443
2.7.1 Physical Modeling 443
2.7.2 Governing Equations 444
2.7.3 Equilibrium Solutions 444
3. **Applications of Electrical Insulation** 445
3.1 Gaseous Insulation 445
3.1.1 Corona 445
3.1.2 Streamer Mechanisms 445
3.1.3 Electronegative Gas—SF_6 446
3.1.4 Arc Discharges 446
3.1.5 Vacuum Insulation 446
3.2 Liquid Insulation 446
3.2.1 Characteristics 446
3.2.2 Insulation Properties and Standards 447
3.2.2.1 Electrical Properties 448
3.2.2.2 Physical Properties 448
3.2.2.3 Chemical Properties 449
3.2.3 Analysis of Gas in Transformer Oils 450
3.2.3.1 Background 450
3.2.3.2 Gas Cause and Origin 450
3.2.3.3 Detection Methods 451
3.2.3.4 Interpretation Theories 451
3.3 Solid Insulation 451
3.3.1 Characteristics 451
3.3.2 Treeing 452
4. **Measurement Methodologies of Electrical Insulation** 453
4.1 Dielectrometry Measurements 453
4.1.1 Sinusoidal Steady State 453
4.1.2 Parallel-Plate Sensor 453
4.1.3 Oil-Impregnated Pressboard 454
4.1.4 Imposed ω-k Dielectrometry 455
4.2 Flow Electrification 456
4.2.1 Background to the Problem 456
4.2.2 The Absolute Charge Sensor 456
4.2.3 Flow-Loop Measurements 457
4.2.4 Couette Charger 458
4.3 Kerr Electro-optic Field-Mapping Measurements 459
4.3.1 Governing Equations 459
4.3.2 Representative Measurements in Highly Purified Water 459
4.3.3 Electron-Beam–Irradiated Polymethylmethacrylate 461
4.4 Acoustic Measurements of Space-Charge Distributions 462
4.4.1 Pressure-Wave Method 464
4.4.2 Electroacoustic Method 464
4.5 Thermally Stimulated Measurements 464
Glossary 465
Works Cited 465
Further Reading 466

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

The essential purpose of electrical insulation is to prevent the flow of electricity. An ideal insulator has zero electrical conductivity and is only characterized by its dielectric permittivity ϵ. However, "real" insulators have finite electrical conduction that leads to power dissipation, and to chemical degradation that causes the insulation to age and ultimately to fail. If the electric field exceeds the electrical breakdown strength of the insulator, spark discharges will occur with large values of current often leading to failure.

The presence of matter between voltage-stressed metallic electrodes often modifies the electric field because even though the material is initially uncharged, the electric field causes charge motion, called conduction, and small charge displacements, called polarization. Because of the large number of atoms present—an Avogadro's number of 6.02×10^{23} per gram molecular weight—slight imbalances in the distribution have large effects on the electric field inside and outside the material. We must then self-consistently account for the electric-field effect on charge motion and redistribution, together with the local net charge as a source of electric field.

If the material polarization and conduction constitutive laws are known, the electric-field distribution throughout the insulation can be calculated. However, often materials contain structural defects and impurities. Charge carriers are also generated from electron, light, or nuclear radiation, from high electric-field strengths that cause charge emission from electrodes, by dielectric ionization, or by contact charging of impurities. With imprecise control or knowledge of such physical mechanisms and material constituents, it often becomes impossible to calculate insulation performance, and many tests are necessary to develop statistics of electric breakdown strength and insulation life.

1. IMPORTANT ELECTRICAL PROPERTIES OF INSULATION

To characterize the electrical insulation properties of an insulator, it is necessary to specify the dielectric permittivity ϵ (or relative dielectric constant $\epsilon_r = \epsilon/\epsilon_0$ with $\epsilon_0 \approx 8.854 \times 10^{-12}$ F/m the permittivity of free space), the electrical resistivity ρ (or its reciprocal, the conductivity $\sigma = 1/\rho$), and the dielectric breakdown strength under dc, ac, and impulse voltages. Table 1 lists represen-

Table 1. Typical values of dielectric strength, resistivity, and relative permittivity of common electrically insulating materials.

	Dielectric strength (MV/m)	Resistivity (Ω m)	Relative permittivity
Gases (STP)			
Air	~3		1.0
Nitrogen	~3		1.0
SF_6	~7.5		1.0
Liquids			
Transformer oil	35	$>10^{12}$	2.2
Pure water (25 °C)	~13	$\sim 1.8 \times 10^5$	78
Pure ethylene glycol (25 °C)	~27	$\sim 9 \times 10^6$	38
Solids			
Bakelite	~24	$>10^{10}$	4.9
Cellulose paper	~10	$>10^{13}$	3.8
Ethylene propylene rubber	30–40	$\sim 10^{13}$	2.5–3.0
Mica	~100	$>10^{16}$	5.4
Polycarbonate	~250	$\sim 10^{14}$	3.0
Polymethylmethacrylate	~350	$>10^{13}$	3.5
Polystyrene	~200	$\sim 10^{14}$	2.5
Polyethylene (25 °C)	~570	$>10^{14}$	2.2
Polypropylene	~300	$\sim 10^{16}$	2.2
Porcelain	~10	$>2 \times 10^{11}$	5–7
Glass (Pyrex)	~17	$>2 \times 10^{11}$	5
Barium titanate	~5	$>10^9$	1200

tative values for typical insulating gases, liquids, and solids.

The study of electrical insulation is interdisciplinary, as the insulation properties depend strongly on electrical, thermal, physical, and chemical parameters. The fundamental limit for most high-voltage systems is the insulation strength—that is, the highest voltage that can be safely sustained without spark discharges.

Gaseous and liquid insulation offer the advantage of filling all available space and of often being self-healing after a spark discharge as the debris is dispersed or carried away by a flow. Liquids also increase the rate of heat transfer by their high thermal conductivity and by convection. Transformer oil is about 20–30 times more effective than air in removing the heat generated in the windings and magnetic core of a transformer. A solid offers the advantage of mechanical strength but often has voids that are not self-healing after electrical breakdown, because the insulation is punctured, leaving a carbonized discharge channel that is conducting.

Even vacuum is not a perfect insulator and, in a real system, is really a gas at very low pressure. Electron emission from the cathode due to microscopic particles or protrusions releases dissolved and absorbed gas on the electrodes and causes heating that slightly melts the metal electrode. Eventually, sufficient vapor is produced so that a low-pressure gas discharge can result if sufficient voltage is applied. For a typical ~1-mm gap at a pressure of $\sim 10^{-3}$ Torr, the breakdown field strength is of order E_{bd} ~375 kV/cm, whereas for typical liquids and solids the working breakdown strength is of order 100–200 kV/cm, although thin-film solids in laboratory measurements can have breakdown field strengths that exceed 1 MV/cm.

The initiation of spark discharges is generally by charge injection from the electrodes. Such discharges contribute to thermal, oxidative, radiative, mechanical, and chemical degradation of insulation that lead to aging and loss of life. Thermal failure occurs when the heat generated by Joule heating and by other dielectric losses cannot be easily lost to the surroundings. The temperature increases, which further increases dielectric loss, leading to thermal runaway. For most practical applications it is important to estimate insulation life from sets of measurements at elevated voltage stresses to measure the time to failure under accelerated aging conditions. Various statistical models are used to estimate the probability of in-service insulation failure from laboratory tests.

2. REPRESENTATIVE ANALYTICAL DESCRIPTIONS OF ELECTRICAL INSULATION

2.1 Linear Lossy Dielectrics

2.1.1 Governing Equations In most electrical insulation applications, the frequency or time rates of change of electric field are sufficiently slow that Maxwell's equations are well approximated in their electroquasistatic limit. Then, electrically insulating materials are essentially capacitive with the material characterized by its dielectric permittivity ϵ. Maxwell's equations can then be written as

$$\nabla \times \mathbf{E} = 0 \Rightarrow \mathbf{E} = -\nabla \Phi, \tag{1}$$

$$\nabla \cdot \mathbf{E} = \rho/\epsilon \Rightarrow \nabla^2 \Phi = -\rho/\epsilon, \tag{2}$$

$$\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} = 0, \tag{3}$$

where for simplicity we take ϵ to be a constant in the region of interest. The irrotational electric field $\mathbf{E}$ in (1) comes from Faraday's law with negligible magnetic field, which allows the definition of the electric scalar potential Φ. In (2), Gauss's law relates the electric field to the source volume charge density ρ. Together with (1), (2) shows that the potential Φ obeys Poisson's equation. Conservation of charge in (3) relates the current density $\mathbf{J}$ to the time rate of change of charge density.

If a pair of electrodes at voltage v enclose a dielectric medium, even with ϵ as a function of position, the total energy stored in the capacitor is

$$W = \int_{\text{all space}} \tfrac{1}{2} \epsilon |\mathbf{E}|^2 \, dV = \tfrac{1}{2} C v^2, \tag{4}$$

where C is the capacitance between electrodes.

2.1.2 Ohmic Material For electrical insulation applications, it is usually desired that the current density $\mathbf{J}$ be as small as possible. Many materials are accurately described by an Ohmic constitutive law where σ is the Ohmic conductivity:

$$\mathbf{J} = \sigma\mathbf{E} \Rightarrow v = iR, \tag{5}$$

which is the field form of the familiar circuit Ohm's-law proportionality relating voltage v to current i via resistance R.

A homogeneous lossy dielectric between a pair of electrodes at voltage difference v has a circuit model of resistance R in parallel with capacitance C such that

$$RC = \epsilon/\sigma, \tag{6}$$

where the RC product is independent of voltage, electrode shape, area, or spacing. The power dissipated in the system is

$$P = \int_{\text{all space}} \sigma|\mathbf{E}|^2\,dV = i^2R = v^2/R. \tag{7}$$

2.1.3 Transient Charge Relaxation Substituting Ohm's law of (5) into (3) for constant conductivity σ yields

$$\sigma\nabla\cdot\mathbf{E} + \frac{\partial\rho}{\partial t} = 0 \Rightarrow \frac{\partial\rho}{\partial t} + \frac{\sigma}{\epsilon}\rho = 0, \tag{8}$$

where the right-hand equation eliminates $\nabla\cdot\mathbf{E}$ using (2). The solution to (8) is

$$\rho(x,y,z,t) = \rho(x,y,z,t=0)e^{-t/\tau}, \tag{9}$$

where $\tau = \epsilon/\sigma$ is called the dielectric relaxation time and equals from (6) the circuit RC time constant. This solution shows that if any region of space has an initial volume charge distribution $\rho(x, y, z, t = 0)$, it will exponentially decay with time constant τ. If an Ohmic region has no initial volume charge so that $\rho(x, y, z, t = 0) = 0$, then the volume charge density remains zero in that region thereafter.

2.1.4 Interfacial Charge

2.1.4.1 Step Transient. Surface charge can accumulate on the interface between two dissimilar lossy dielectric materials. This is illustrated for the series lossy capacitor shown in Fig. 1, which is modeled by a pair of parallel RC circuits connected in series. We consider a step voltage applied at $t = 0$ to an initially unexcited system. Neglecting end effects, the electric field can only be in the x direction and at best vary with that coordinate. In the absence of any volume charge ($\rho = 0$), Eq. (2) for the electric field requires $dE_x/dx = 0$ so that the electric fields in each region, E_1 and E_2, are uniform in space within each lossy dielectric but can vary with time.

The interfacial form of (3) gives the required conservation of charge boundary condition at the interface as

$$J_2 - J_1 + \frac{\partial\sigma_s}{\partial t} = 0; \quad \sigma_s = \epsilon_2E_2 - \epsilon_1E_1, \tag{10}$$

where σ_s is the interfacial surface charge density, which equals the difference in perpendicular displacement fields ($\mathbf{D} = \epsilon\mathbf{E}$) across the interface.

At $t = 0_-$, the interfacial surface charge density is zero and remains continuous at $t = 0_+$ so that $D = \epsilon_1E_1 = \epsilon_2E_2$, while in the dc steady state as $t \to \infty$, Eq. (10) requires that the current density $J = \sigma_1E_1 = \sigma_2E_2$ be continuous across the interface.

The solutions for the electric fields in each region are then

$$\begin{aligned} E_1 &= \frac{\sigma_2 v}{\sigma_2 a + \sigma_1 b}(1 - e^{-t/\tau}) + \frac{\epsilon_2 v}{\epsilon_2 a + \epsilon_1 b}e^{-t/\tau}, \\ E_2 &= \frac{\sigma_1 v}{\sigma_2 a + \sigma_1 b}(1 - e^{-t/\tau}) + \frac{\epsilon_1 v}{\epsilon_2 a + \epsilon_1 b}e^{-t/\tau}, \end{aligned} \tag{11}$$

where the time constant

$$\tau = \frac{\epsilon_1 b + \epsilon_2 a}{\sigma_1 b + \sigma_2 a} \tag{12}$$

is a thickness-weighted average of relaxation times in each material. The interfacial surface charge density is

$$\begin{aligned} \sigma_s &= \epsilon_2E_2 - \epsilon_1E_1 \\ &= \frac{(\epsilon_2\sigma_1 - \epsilon_1\sigma_2)v}{\sigma_2 a + \sigma_1 b}(1 - e^{-t/\tau}), \end{aligned} \tag{13}$$

which is zero at $t = 0$ and increases to a

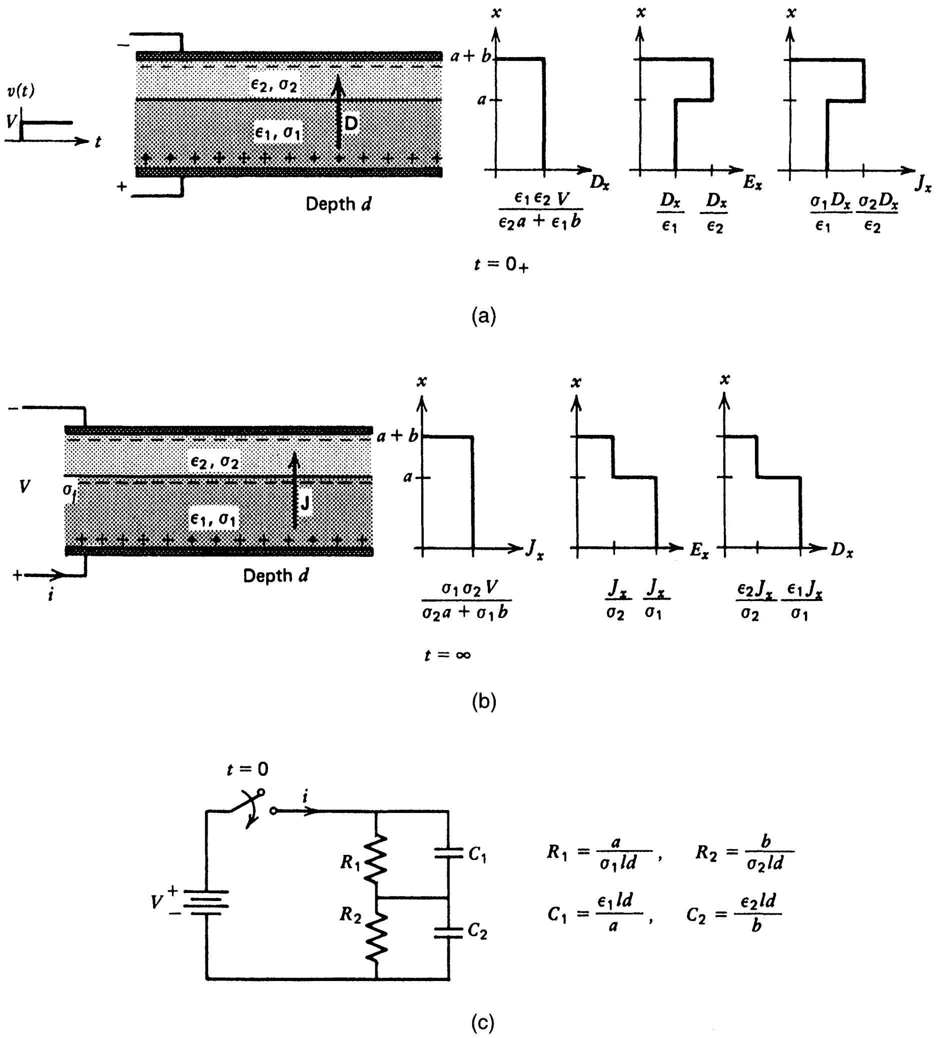
v(t)
V
t
ϵ_2, σ_2
ϵ_1, σ_1
D
Depth d
$a+b$
a
x
D_x
E_x
J_x
$\frac{\epsilon_1 \epsilon_2 V}{\epsilon_2 a + \epsilon_1 b}$
$\frac{D_x}{\epsilon_1}$ $\frac{D_x}{\epsilon_2}$
$\frac{\sigma_1 D_x}{\epsilon_1}$ $\frac{\sigma_2 D_x}{\epsilon_2}$
$t = 0_+$
(a)
V
σ_f
J
i
$\frac{\sigma_1 \sigma_2 V}{\sigma_2 a + \sigma_1 b}$
$\frac{J_x}{\sigma_2}$ $\frac{J_x}{\sigma_1}$
$\frac{\epsilon_2 J_x}{\sigma_2}$ $\frac{\epsilon_1 J_x}{\sigma_1}$
$t = \infty$
(b)
$t = 0$
i
R_1
R_2
C_1
C_2
$R_1 = \frac{a}{\sigma_1 ld}$, $R_2 = \frac{b}{\sigma_2 ld}$
$C_1 = \frac{\epsilon_1 ld}{a}$, $C_2 = \frac{\epsilon_2 ld}{b}$
(c)

FIG. 1. Two different lossy dielectric materials in series between parallel plate electrodes have permittivities and Ohmic conductivities that change abruptly across the interface. (a) At $t = 0_+$, right after a step voltage is applied, the interface is uncharged so that the displacement field $D = \epsilon_1 E_2 = \epsilon_2 E_2$ is continuous with the solution the same as for two lossless dielectrics in series. (b) Since the current is discontinuous across the boundary between the materials, the interface will charge up. In the dc steady state, the current density $J = \sigma_1 E_1 = \sigma_2 E_2$ is continuous. (c) Each region is circuit equivalent to a resistor and capacitor in parallel.

positive steady-state value if $\epsilon_2/\sigma_2 > \epsilon_1/\sigma_1$, decreases to a negative steady-state value if $\epsilon_2/\sigma_2 < \epsilon_1/\sigma_1$, or remains zero for all time if $\epsilon_2/\sigma_2 = \epsilon_1/\sigma_1$.

2.1.4.2 Sinusoidal Steady State. If, rather than a step voltage, the applied voltage is sinusoidal, the electric fields in each region will also vary sinusoidally with time:

$$v(t) = Re\{\hat{v}e^{j\omega t}\} \Rightarrow E_1(t) = Re\{\hat{E}_1 e^{j\omega t}\}, E_2(t) = Re\{\hat{E}_2 e^{j\omega t}\}. \qquad (14)$$

In (14) it is convenient to use complex notation with careted symbols representing complex amplitudes. Then (10) requires that the sum of conduction and displacement current densities be continuous across the interface:

$$(\sigma_2 + j\omega\epsilon_2)\hat{E}_2 = (\sigma_1 + j\omega\epsilon_1)\hat{E}_1. \qquad (15)$$

The electric field complex amplitudes are then related as

$$\frac{\hat{E}_2}{(j\omega\epsilon_1 + \sigma_1)} = \frac{\hat{E}_1}{(j\omega\epsilon_2 + \sigma_2)} = \frac{\hat{v}}{[b(\sigma_1 + j\omega\epsilon_1) + a(\sigma_2 + j\omega\epsilon_2)]}, \tag{16}$$

and the interfacial surface charge-density complex amplitude is

$$\hat{\sigma}_s = \epsilon_2\hat{E}_2 - \epsilon_1\hat{E}_1 = \frac{(\epsilon_2\sigma_1 - \epsilon_1\sigma_2)\hat{v}}{[b(\sigma_1 + j\omega\epsilon_1) + a(\sigma_2 + j\omega\epsilon_2)]}. \tag{17}$$

2.2 Charge-Injection Models

2.2.1 Schottky Emission Schottky emission is essentially thermionic emission from a metal electrode into the conduction band of a dielectric taking into account the lowering of the work-function potential barrier by the image force on the injected electron of charge $e = 1.6 \times 10^{-19}$ Coulombs. The injected current density at an absolute temperature T with an electric field E into a dielectric of permittivity ϵ is

$$J = AT^2 \exp\{-[\phi - (e^3E/4\pi\epsilon)^{1/2}]/kT\}, \tag{18}$$

where ϕ is the work function, m is the effective electron mass, $k = 1.38 \times 10^{-23}$ J/K is Boltzmann's constant, $h = 6.6256 \times 10^{-34}$ J s is Planck's constant, and

$$A = 4\pi emk^2/h^3 \sim 120 \text{ A cm}^{-2} \text{ K}^{-2} \tag{19}$$

is the Richardson–Dushman constant of thermionic emission.

2.2.2 Fowler–Nordheim Field Emission Quantum mechanical tunneling of electrons from a metal surface into a dielectric under a strong electric field results in the current density

$$J = \frac{e^3E^2}{8\pi h\phi} \exp[-8\pi\sqrt{2m}\, \phi^{3/2}/3heE]. \tag{20}$$

2.3 Avalanche Modeling

The fundamental model of electric breakdown is the avalanche model, where an injected electron collides with a molecule to release an additional electron. Each of these electrons then continues with more collisions, multiplying the number of electrons and leading to current growth. With α, known as Townsend's first ionization coefficient, denoting the number of ionizing collisions per electron per unit distance, the increase in the number density n of electrons in a distance dx is

$$dn = n\alpha\, dx \Rightarrow n = n_0e^{\alpha x}; \quad \alpha = e^{-\nu/eE\lambda}/\lambda, \tag{21}$$

resulting in the number density of electrons exponentially increasing with distance, with n_0 the original number of electrons at the cathode, $x = 0$. The Townsend coefficient α increases with increasing electric field E, depends strongly on the mean free path λ between collisions, and is characterized by the energy ν required to ionize gas molecules. The conduction current flow through the anode a distance d from the cathode is then

$$I(x = d) = I(x = 0)e^{\alpha d}, \tag{22}$$

where $I(x = 0)$ is the cathode emission current. Such impact ionization by electrons alone is not sufficient to cause electrical breakdown.

As the fast-moving electrons enter the anode, they leave behind the relatively slower positive ions. As these positive ions drift to the cathode, they also collide with neutral molecules and liberate additional electrons with probability γ, known as Townsend's second ionization coefficient. The parameter γ also includes the effects of additional free electrons generated by photoionization processes.

The anode current then becomes

$$I(x = d) = I(x = 0)\frac{e^{\alpha d}}{1 - \gamma(e^{\alpha d} - 1)}. \tag{23}$$

The current grows without bound when the denominator in (23) becomes zero, so that the discharge becomes self-sustaining when

$$\gamma(e^{\alpha d} - 1) = 1. \tag{24}$$

Electronegative gases, such as SF_6, have an electron attachment η, so that Townsend's breakdown criterion of (24) is modified to

$$\frac{\gamma\alpha}{\alpha - \eta}[e^{(\alpha-\eta)d} - 1] = 1, \tag{25}$$

which requires a larger value of α and thus a larger electric field before the equality is satisfied.

With the mean free path λ inversely proportional to pressure, the Townsend coefficients α and γ are functions of E/p, where E is the electric field and p is the gas pressure. This model is appropriate for gases, but it is also used to describe electrical breakdown in liquids with formation of a gaseous bubble (perhaps due to heating) and in solids that have voids. In a uniform electric field, the curve of gas breakdown voltage versus the pressure–gap (pd) product is known as the Paschen curve, as shown in Fig. 2. For a gap of fixed spacing d, as the pressure decreases, the gas density decreases and the electron mean free path between collisions increases. Thus, each electron makes fewer collisions with gas molecules as it travels to the anode and thus avoids loss of energy due to collisions. A lower electric field can then accelerate electrons to sufficient kinetic energy for ionizing collisions. As the pressure further decreases, the breakdown voltage decreases to a minimum value, v_p, known as the Paschen minimum. As the density is decreased below the Paschen minimum, there are even fewer collisions, so that self-sustaining ionization is maintained only with a corresponding increase in voltage. To the right of the Paschen minimum, there are so many collisions that much of the electron kinetic energy is dissipated so that a large voltage is necessary for ionizing collisions.

2.4 Statistical Methods

To model insulation behavior, three statistical distributions are usually used. The normal distribution, also known as the Gaussian distribution or bell-shaped curve, has the probability $F(x)$ of a value of a measured property x given by

$$F(x) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{26}$$

where μ is the mean of x and σ is the standard deviation. The value of μ gives the center of the distribution peak, and σ determines the width of the distribution and describes the scatter of data. This is the simplest statistical method and is typically used to describe such insulation properties as dielectric constant, dissipation factor, insulation resistance, tensile strength, and elongation.

The log normal distribution is described as

$$F(x) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left[-\frac{(\log x-\mu)^2}{2\sigma^2}\right], \tag{27}$$

where μ is now the logarithmic mean and σ is the logarithmic standard deviation. The logarithm can be base e or base 10. This distribution is typically used for such insulation

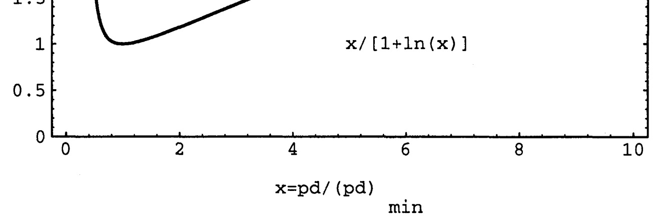

FIG. 2. The Paschen curve describes the voltage breakdown strength in gases as a function of the pressure-gap (*pd*) product. The nondimensional function given in the plot with voltage and pressure–gap product normalized to the Paschen minimum approximates measured curves, as described in von Hippel (1959).

properties as corrosion, gaseous diffusion, chemical reaction rates, and time to failure.

The Weibull distribution is usually used to describe life and failure distributions, time to failure, and dielectric strength data,

$$F(x) = 1 - \exp\left[-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right], \tag{28}$$

where α is a scale parameter analogous to the mean, β is a shape parameter analogous to the standard deviation, and γ is a location parameter that indicates a threshold below which there is no failure.

2.5 Life-Prediction Techniques

Thermal aging relates aging time to temperature and often assumes a single chemical reaction whose rate depends on an activation energy H and absolute temperature T. The rate of reaction R is usually described by the Arrhenius model,

$$R = A\exp[-H/kT], \tag{29}$$

where A is a frequency factor. The rate of insulation aging approximately doubles for every 10 °C rise.

To estimate insulation life with accelerated aging by elevating testing voltage v_t above normal service voltage v_s, an inverse-power model is often assumed:

$$t_t/t_s = (v_s/v_t)^n, \tag{30}$$

where t_t is the testing time, t_s is the time in service, and n is called the life exponent. For many cases, $n \approx 7$. Tests usually take the lowest voltage for v_t to yield failures with a value of $t_t \sim 6$ months. By combining Weibull statistics with (30), an estimate of the probability of in-service insulation failure can be determined from laboratory measurements. Life-prediction techniques are greatly complicated when there are multiple simultaneous stresses, typically from electrical, thermal, mechanical, and environmental factors.

2.6 Space-Charge Effects

2.6.1 Space-Charge Distortion of the Electric Field Most analytical models assume that the electric field is spatially uniform. This motivates many tests to use parallel plane electrodes, but the uniform-field assumption is only true in the central region between electrodes in the absence of net volume charge.

To help understand space-charge effects, often due to charge injection from electrodes as described in Sec. 2.2, we consider the simplest case of planar electrodes with an x-directed electric field E that is distorted by net charge density $\rho(x)$ only dependent on the x coordinate. In such a one-dimensional electrode geometry, Gauss's law of Eq. (2) requires that the slope of the electric-field distribution be proportional to the local charge density:

$$\nabla\cdot\mathbf{E} = \frac{\rho}{\epsilon} \rightarrow \frac{\partial E}{\partial x} = \frac{\rho}{\epsilon}. \tag{31}$$

For the case of no volume charge, shown in Fig. 3(a), the electric field is uniform, given by $E_0 = v/d$ for a voltage v across a gap d. The electric field drops at a charge-injecting electrode, but its average value E_0 remains constant at v/d. Charge injection thus causes the electric field to increase above the average value at the non-charge-injecting electrode. In Fig. 3(b), unipolar injection has the electric field maximum at the non-charge-injecting electrode, thus possibly leading to electrical breakdown at lower voltages. For bipolar homocharge injection in Fig. 3(c), positive charge is injected at the anode and negative charge is injected at the cathode, so that the electric field is lowered at both electrodes and is largest in the central region. Since electrical breakdown often initiates at the electrode-dielectric interface, this case can allow higher-voltage operation without breakdown, shown by Zahn *et al.* (1986) to be up to ~40% higher in highly purified water. If the voltage suddenly reverses, the electrical field must also instantaneously reverse, but the charge distribution cannot immediately change because it takes some time for the volume charge to migrate. Thus, for early times with voltage reversal, positive charge is near the cathode and negative charge is near the anode, enhancing the fields near the electrodes, as shown in Fig. 3(d). Such a bipolar heterocharge configuration leads to breakdown at lower voltages, as has been found in HVDC polyethylene cables when the voltage is instantaneously reversed

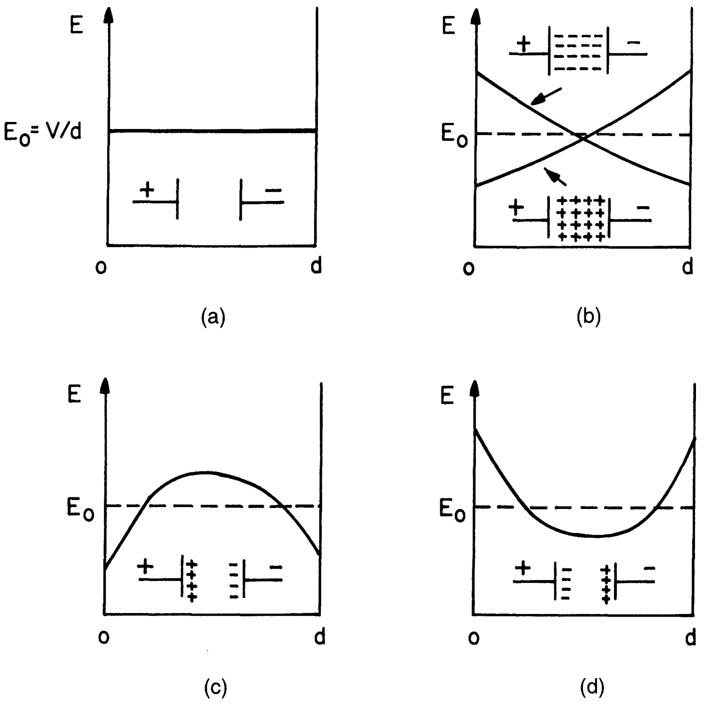

FIG. 3. Space-charge distortion of the electric-field distribution between parallel-plate electrodes with spacing d at voltage v so that the average electric field is $E_0 = v/d$. (a) No space charge so that the electric field is uniform at E_0; (b) unipolar positive or negative charge injection so that the electric field is reduced at the charge-injecting electrode and enhanced at the non-charge-injecting electrode; (c) bipolar homocharge injection so that the electric field is reduced at both electrodes and enhanced in the central region; (d) bipolar heterocharge distribution where electric field is enhanced at both electrodes and depressed in the central region.

to reverse the direction of power flow. A similar configuration to Fig. 3(d) also occurs when the dielectric is ionized so that free ions are attracted to their image charges on the electrodes.

2.6.2 Governing Equations for Drift-Dominated Unipolar Conduction Because Ohmic conduction involves the presence of equal amounts of positive and negative charge, there can be a net current flow with zero volume charge. However, a volume charge distribution within a dielectric can arise from spatially varying conductivity, irradiation, charge injection from high-field–stressed electrodes as described in Sec. 2.2, dielectric ionization, or contact charging of impurities.

To model net charge transport, we use a drift-dominated unipolar mobility model where the velocity $d\mathbf{r}/dt$ of a positive charge carrier is proportional to the local electric field through the mobility μ, $d\mathbf{r}/dt = \mu\mathbf{E}$, and in turn the electric field is related to the charge density through Gauss's law of Eq. (2). Such a conduction model is often used in the analysis of high-field conduction in solid-state semiconductors when diffusion is negligible.

We consider a one-dimensional model in a medium of thickness d and permittivity ϵ with x-directed electric field E. Without loss of generality, we assume that positive mobile charge with density ρ has a mobility μ. The current-density constitutive law is then

$$\mathbf{J} = \rho\frac{d\mathbf{r}}{dt} = \rho\mu\mathbf{E}. \tag{32}$$

Substituting Eq. (32) into Eq. (3) and integrating with respect to x together with one-dimensional forms of Eqs. (1) and (2) yields the governing equations

$$\epsilon\frac{\partial E}{\partial t} + \rho\mu E = J(t), \tag{33}$$

$$\frac{\partial E}{\partial x} = \frac{\rho}{\epsilon}, \tag{34}$$

$$v = \int_0^d E\,dx, \tag{35}$$

where $J(t)$ is the terminal current per unit electrode area and v is the voltage difference across the electrodes at $x = 0$ and $x = d$. Differentiating (33) with respect to x and using (34) gives an equation for the charge density,

$$\frac{\partial \rho}{\partial t} + \frac{\rho^2 \mu}{\epsilon} + \mu E \frac{\partial \rho}{\partial x} = 0. \tag{36}$$

Using (34) in (33), integrating between the electrodes, and then using (35) yields the terminal voltage–current relationship

$$\frac{dv}{dt} + \frac{\mu}{2}[E^2(x = d) - E^2(x = 0)] = \frac{J(t)d}{\epsilon}. \tag{37}$$

The first term is proportional to the familiar capacitive displacement current, while the second term is due to migrating charge.

Equations (33) and (36) are a pair of quasilinear partial differential equations, which can be converted to a set of ordinary differential equations by jumping into the frame of reference of the migrating charge moving at the charge velocity

$$\frac{dx}{dt} = \mu E. \tag{38}$$

In this moving frame of reference, the total time derivatives of electric field and charge density are

$$\frac{dE}{dt} = \frac{\partial E}{\partial t} + \frac{dx}{dt}\frac{\partial E}{\partial x} = \frac{J(t)}{\epsilon}, \tag{39}$$

$$\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t} + \frac{dx}{dt}\frac{\partial \rho}{\partial x} = -\frac{\rho^2 \mu}{\epsilon}. \tag{40}$$

The solution to (40) is

$$\rho(t) = \frac{\rho_0(t = t_0)}{[1 + \rho_0(t = t_0)\mu(t - t_0)/\epsilon]} \text{ on } \frac{dx}{dt} = \mu E, \tag{41}$$

where $\rho_0(t = t_0)$ is the initial charge density on a trajectory that starts at $t = t_0$. Physically, this solution of decreasing charge density with time is due to the self-field Coulombic repulsion of charge so that any initial packet of charge expands with time. Even though the total charge within a packet is constant, the density decreases because the volume of charge is increasing. Note that if the initial charge density $\rho_0(t = t_0)$ is zero, the charge density remains zero along the charge trajectory.

If the terminal current density $J(t)$ is known, the electric field along a charge trajectory is obtained by direct time integration of (39), and the position of the charge is found by time integration of the electric field using (38). Often the terminal voltage is constrained so that the terminal current must be found using (37). Because the electric field at both electrodes is generally not yet known, usually a numerical method is needed to solve self-consistently for all variables.

2.6.3 dc Steady-State Distributions

The dc steady-state distributions for electric field and charge density are found by setting the time derivatives in (33) and (36) equal to zero:

$$E_{ss} = \left[\frac{2Jx}{\epsilon\mu} + E_0^2\right]^{1/2}, \tag{42}$$

$$\rho_{ss} = \frac{J}{\mu E_{ss}} = \frac{\epsilon dE_{ss}}{dx} = \frac{J}{\mu}\left[\frac{2Jx}{\epsilon\mu} + E_0^2\right]^{-1/2}, \tag{43}$$

where J is the dc steady-state current density and E_0 is the emitter electric field at $x = 0$, which must be specified as a charge-injection boundary condition, such as from (18) or (20).

The current–voltage relation is found from (35) by integrating (42) between electrodes stressed by dc voltage v:

$$v = \int_0^d E_{ss}\, dx = \frac{\epsilon\mu}{3J}\left\{\left[\frac{2Jd}{\epsilon\mu} + E_0^2\right]^{3/2} - E_0^3\right\}. \tag{44}$$

Simplifying (44) gives solution

$$\frac{16Jd^3}{\epsilon\mu v^2} = \left(9 - 12\left(\frac{E_0 d}{v}\right)^2 + \left\{\left[9 - 12\left(\frac{E_0 d}{v}\right)^2\right]^2 + 192\left(\frac{E_0 d}{v}\right)^3\left[1 - \frac{E_0 d}{v}\right]\right\}^{1/2}\right). \tag{45}$$

For positive charge injection the emitter electric field E_0 must obey the inequality $0 \le E_0 \le v/d$. For negative E_0 there is no charge injection as the electric field would push the positive charge back into the electrode. For $E_0 > v/d$, the spatial derivative of E_{ss} could not be positive as required by Gauss's law of

(34) for positive charge injection and simultaneously satisfy (35) of having an average field strength between electrodes of v/d. The positive space-charge distribution distorts the electric field so that the maximum electric field strength occurs at the negative electrode at $x = d$. This field enhancement is important in electric breakdown studies as breakdown initiates where the electric field is largest.

The largest electric field at the non-charge-emitting electrode at $x = d$ is $E(x = d) = 1.5v/d$, which occurs for space-charge limited conditions when $E_0 = 0$ and $\rho_{ss}(x = 0) = \infty$, so that $Jd^3/\epsilon\mu v^2 = 9/8$. As $E_0 d/v$ approaches unity, the electric field distribution approaches a uniform field as the injected charge density and current density become small.

We expect that the injected charge density at $x = 0$, $\rho_{ss}(x = 0) = \rho_0$, would increase as the electric field E_0 at the charge-injecting electrode increases. The simplest relationship that approximates (20) in the high-field limit when the exponential is approximately unity is

$$\rho_0 = AE_0; \quad A \approx \frac{e^3}{8\pi h\phi\mu}, \tag{46}$$

where A is an injection constant. The steady-state current density, which is independent of x, can be particularly evaluated at $x = 0$ as

$$J = \rho\mu E = \rho_0\mu E_0 = \mu A E_0^2. \tag{47}$$

Using (47) in (44) lets us solve for E_0 as

$$\frac{E_0 d}{v} = \frac{3Ad/\epsilon}{(2Ad/\epsilon + 1)^{3/2} - 1}. \tag{48}$$

2.6.4 Space-Charge Effects in Pulsed-Power Technology For inertial-confinement fusion machines, as shown in Fig. 4, and directed-energy devices, high peak power at the terawatt level is needed for short times of the order of 100 ns. Pulsed-power technology collects and stores electric energy at a low input power, typically of order 1 kW, for a long time of order seconds in a capacitor bank, and then delivers this ~1-kJ energy in a much shorter time, ~100 ns, at a much higher power level, ~10^{10} W.

Highly purified water and water/ethylene glycol mixtures have been studied for use as the dielectric in pulse-forming transmission lines because the high relative dielectric constant, $\epsilon_r \approx 40$–88, and high resistivity, $\rho \approx 18$–8500 MΩ cm, allow short and efficient low-impedance transmission lines, especially at ~ −35 °C with a mixture of 60% glycol/40% water, as described in Zahn *et al.* (1986).

In testing such pulse-forming lines it was found that at high electric stress, the open-circuit voltage across immersed parallel-plate electrodes decayed more quickly as the initial voltage increased and that the decay rate depended on electrode spacing and was not exponential, as shown in Figs. 5(a) and 5(b)

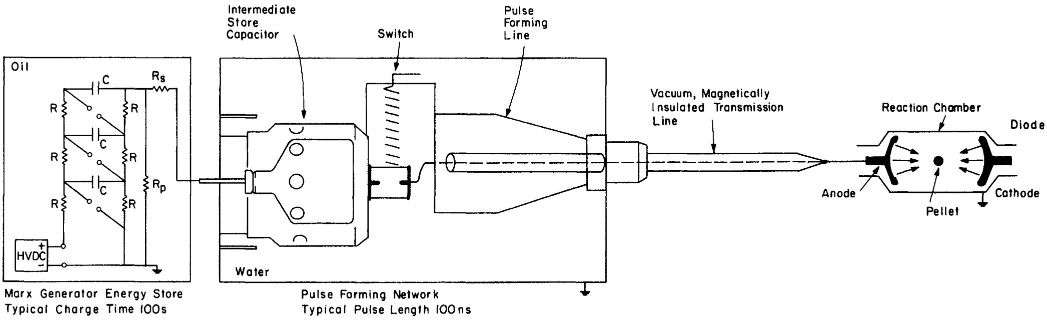

FIG. 4. A representative pulsed-power machine used for inertial-confinement fusion experiments consists of a slowly charged Marx capacitor bank under insulating oil as primary energy storage; a water dielectric capacitor as intermediate energy storage; a switched pulse-forming line that compresses the voltage in time; and a magnetically and vacuum insulated transmission line that delivers the power pulse to a vacuum diode to produce energetic electrons or light ions, which are then accelerated and focused by electric and magnetic fields onto a target, typically a deuterium-tritium pellet.

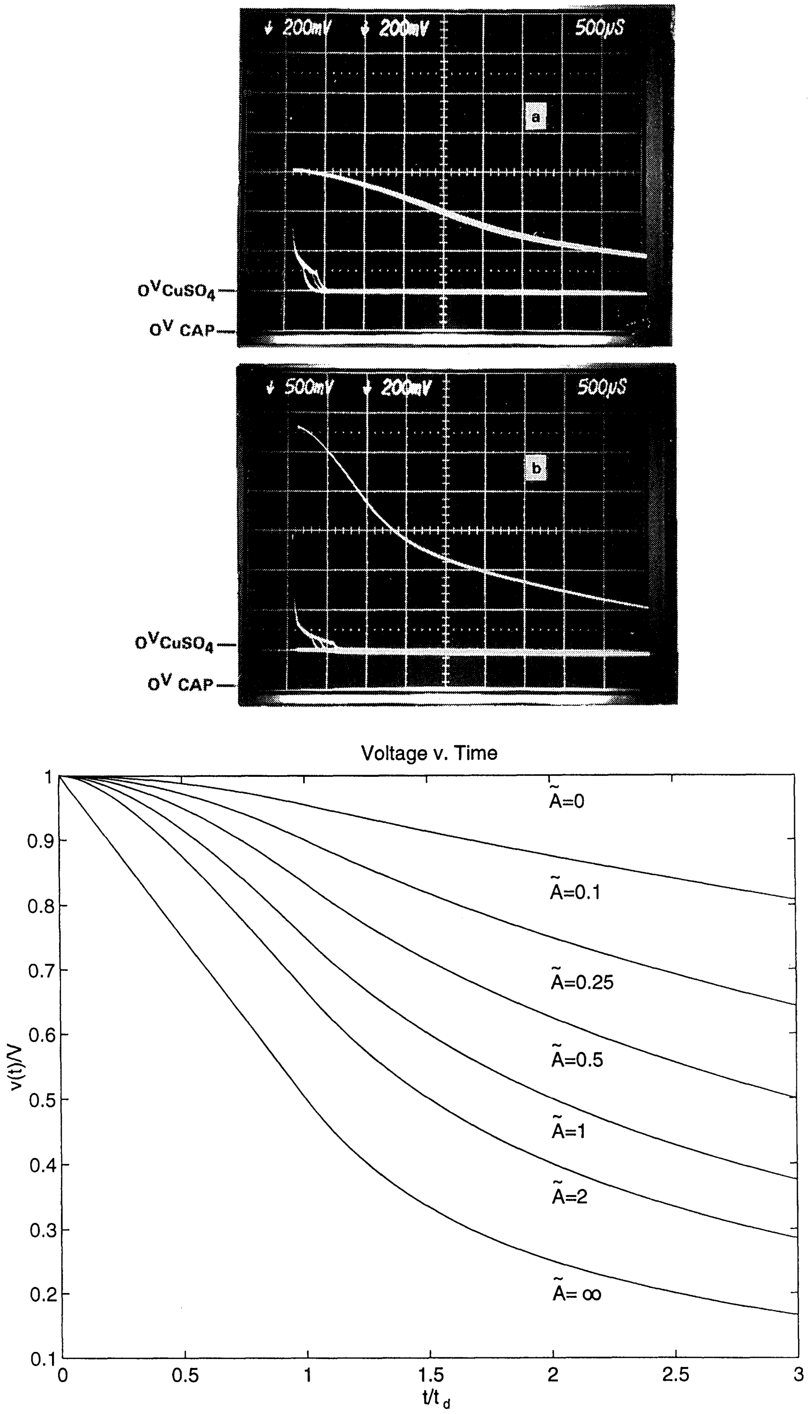

(c)

FIG. 5. (a) Measured open-circuit voltage decay vs. time for an 80% ethylene glycol/20% water mixture at -10 °C with a dielectric relaxation time ~24 ms for parallel-plate stainless-steel electrodes with 5.1-mm gap with initial voltages of (a) 55 kV and (b) 90 kV. Note the change in sign of the curvature of both traces at early time and that by $t \approx 4.5$ ms, both voltages are about the same despite the large difference in initial voltage. (c) Theoretical curves of Eq. (55) using a drift-dominated unipolar conduction model for various values of positive charge injection parameter $\tilde{A} = Ad/\epsilon$ showing nonexponential open-circuit voltage decay.

for two different initial voltages. This disagrees with the simple Ohmic model of a lossy homogeneous capacitor where the open-circuit decay is exponential with decay rate $\tau = RC = \epsilon/\sigma$ independent of capacitor geometry and voltage as given by (6).

To explain the nonlinear behavior, we use the unipolar ion-conduction law of (32) where in a qualitative sense, the quantity $\rho\mu$ is an effective conductivity that varies with time and position. In addition to the dielectric relaxation time of (9), there is a time-of-flight time constant, $t_d = d^2/\mu v$, for injected charge to reach the opposite electrode, which depends on electrode spacing and voltage.

To emphasize space-charge effects, we assume that the Ohmic conductivity is zero and take the capacitor to be initially uncharged in the dielectric volume but with an initial applied voltage v at $t = 0$ and then open circuited thereafter. We find the injection electric field at $x = 0$, $E_0(t)$, using (33) with $J(t) = 0$ and the linear charge-injection law of (46),

$$\frac{dE_0}{dt} + \frac{A\mu}{\epsilon}E_0^2 = 0 \Rightarrow E_0(t) = \frac{v/d}{1 + A\mu vt/\epsilon d}, \tag{49}$$

where the initial condition is $E_0(t = 0) = v/d$.

For those trajectories of injected charge that start at $x = 0$ at time $t = t_0$, Eq. (39) with $J(t) = 0$ shows that the electric field along the trajectory remains constant at the $x = 0$ emitter electric field, $E(t) = E_0(t = t_0)$. Then the trajectory is given from (38) as

$$\begin{aligned}\frac{dx}{dt} &= \mu E_0(t = t_0) \Rightarrow x(t) = \mu E_0(t = t_0)(t - t_0)\\ &= \frac{(\mu v/d)(t - t_0)}{[1 + A\mu v t_0/\epsilon d]}.\end{aligned} \tag{50}$$

In particular, the trajectory that starts at $t_0 = 0$ defines the migrating charge front that reaches $x = d$ at time t_d

$$x_d(t) = \frac{\mu vt}{d};\ t_d = \frac{d^2}{\mu v}. \tag{51}$$

Solving (50) for t_0 gives

$$t_0 = \frac{\mu vt/d - x}{(\mu v/d)(1 + Ax/\epsilon)}, \tag{52}$$

so that the electric field and charge density at time t and position x are

$$E(x,t) = \begin{cases}\dfrac{v(1 + Ax/\epsilon)}{d(1 + Adt/\epsilon t_d)}, & 0 \le x \le x_d(t),\\ v/d, & x_d(t) < x < d;\end{cases} \tag{53}$$

$$\begin{aligned}\rho(x,t) &= \epsilon\frac{\partial E}{\partial x}\\ &= \begin{cases}\dfrac{Av/d}{1 + Adt/\epsilon t_d}, & 0 < x < x_d(t),\\ 0, & x_d(t) < x < d.\end{cases}\end{aligned} \tag{54}$$

The terminal voltage is then computed as

$$\begin{aligned}v(t) &= \int_0^d E dx = \int_0^{x_d(t)} E dx + \int_{x_d(t)}^d E dx;\\ &= \begin{cases} v\left[1 - \dfrac{Adt^2}{2\epsilon t_d^2(1 + Adt/\epsilon t_d)}\right], & 0 \le t \le t_d,\\ \dfrac{v(1 + Ad/2\epsilon)}{1 + Adt/\epsilon t_d}, & t \ge t_d.\end{cases}\end{aligned} \tag{55}$$

Figure 5(c) shows the nonexponential decay of (55) for various values of charge injection constant A, which has similar shape to the measurements shown in Figs. 5(a) and 5(b).

2.7 Drift-Diffusion Conduction

2.7.1 Physical Modeling In such materials as metallic conductors, ionized gases, electrolytic solutions, and semiconductors, the charge carriers can be classically modeled as an ideal-gas plasma. We assume that we have two carriers of equal magnitudes but opposite signs $\pm q$, with respective masses $m_\pm$ and number densities $n_\pm$. These charges can be holes and electrons in a semiconductor, oppositely charged ions in an electrolytic solution, or electrons and essentially stationary nuclei in metals. When an electric field is applied, the positive mobile charges move in the direction of the field, while the negative mobile charges move in the opposite direction. These charges collide with the host medium at respective collision frequencies ν_+ and ν_-. This process then acts as a viscous or frictional dissipation opposing the motion. In addition to electrical and frictional forces, the particles exert a force on their neighbors through a pressure term due to thermal agitation that would be present even if the particles were uncharged. For an ideal gas, the partial pressure is $p = nkT$, where n is the number density of charges, T is the absolute temperature, and $k = 1.38 \times 10^{-23}$ J/K is Boltzmann's constant.

2.7.2 Governing Equations Because the collision frequencies $\nu_\pm$ are typically of the order 10^{13} Hz, the particle inertia is often negligible. Then force equilibrium on each carrier moving at velocities $\mathbf{v}_+$ and $\mathbf{v}_-$ requires

$$\begin{aligned} 0 &= q\mathbf{E} - m_+\nu_+\mathbf{v}_+ - \frac{1}{n_+}\nabla(n_+kT), \\ 0 &= -q\mathbf{E} - m_-\nu_-\mathbf{v}_- - \frac{1}{n_-}\nabla(n_-kT), \end{aligned} \tag{56}$$

where we recognize in the last terms that the force per particle due to pressure is given by $(1/n)\nabla p$.

We solve (56) at constant temperature for the velocity of each carrier as

$$\begin{aligned} \mathbf{v}_+ &= \frac{q}{m_+\nu_+}\mathbf{E} - \frac{kT}{m_+\nu_+n_+}\nabla n_+, \\ \mathbf{v}_- &= -\frac{q}{m_-\nu_-}\mathbf{E} - \frac{kT}{m_-\nu_-n_-}\nabla n_-. \end{aligned} \tag{57}$$

The charge density for each carrier is $\rho_+ = qn_+$ and $\rho_- = -qn_-$. Then the current density for each carrier can be written as

$$\begin{aligned} \mathbf{J}_+ &= \rho_+\mathbf{v}_+ = qn_+\mathbf{v}_+ = \rho_+\mu_+\mathbf{E} - D_+\nabla\rho_+, \\ \mathbf{J}_- &= \rho_-\mathbf{v}_- = -qn_-\mathbf{v}_- = -\rho_-\mu_-\mathbf{E} - D_-\nabla\rho_-, \end{aligned} \tag{58}$$

where $\mu_\pm$ are the particle mobilities and $D_\pm$ are the particle diffusion coefficients:

$$\mu_\pm = \frac{q}{m_\pm\nu_\pm}, \quad D_\pm = \frac{kT}{m_\pm\nu_\pm}. \tag{59}$$

We note that the ratio $D_\pm/\mu_\pm$ for each carrier equals the thermal voltage, kT/q (~25 mV at room temperature), the Einstein relation:

$$\frac{D_+}{\mu_+} = \frac{D_-}{\mu_-} = \frac{kT}{q}. \tag{60}$$

2.7.3 Equilibrium Solutions In equilibrium when the net current density of each carrier is zero, with $\mathbf{E} = -\nabla\Phi$, Eq. (58) shows that the charge densities are related to the potential by Boltzmann distributions:

$$\begin{aligned} \mathbf{J}_+ &= 0 \Rightarrow \rho_+ = \rho_0 e^{-q\Phi/kT}, \\ \mathbf{J}_- &= 0 \Rightarrow \rho_- = -\rho_0 e^{+q\Phi/kT}, \end{aligned} \tag{61}$$

where we use Eq. (60) and $\pm\rho_0$ is the equilibrium charge density of each carrier when $\Phi = 0$.

To find the spatial dependence of ρ and Φ, we use Eq. (61) in the Poisson's-equation form of (2),

$$\begin{aligned} \nabla^2\Phi &= -\frac{\rho_+ + \rho_-}{\epsilon} = -\frac{\rho_0}{\epsilon}\left(e^{-q\Phi/kt} - e^{q\Phi/kT}\right) \\ &= \frac{2\rho_0}{\epsilon}\sinh\frac{q\Phi}{kT}, \end{aligned} \tag{62}$$

which is known as the Poisson–Boltzmann equation.

To examine the form of simple solutions to (62), we take the small-potential limit so that $q\Phi/kT \ll 1$:

$$\nabla^2\Phi - \frac{\Phi}{\lambda^2} = 0, \quad \lambda = \left[\frac{\epsilon kT}{2q\rho_0}\right]^{1/2}, \tag{63}$$

where λ is called the Debye length.

If an electrode is placed at $x = 0$ with potential $\Phi(x = 0) = v_0$ with respect to the potential at $x = \pm\infty$, the potential then decreases exponentially with distance as

$$\Phi(x) = \begin{cases} v_0 e^{-x/\lambda}, & x > 0, \\ v_0 e^{+x/\lambda}, & x < 0. \end{cases} \tag{64}$$

Similarly, if a point charge Q is inserted into the medium, the solution to (63) is

$$\Phi(r) = (Q/4\pi\epsilon r)e^{-r/\lambda}. \tag{65}$$

In Eqs. (64) and (65), the potential decays away from the source with characteristic length λ because counter charges of opposite sign are attracted to the source charge, tending to shield the surrounding region.

3. APPLICATIONS OF ELECTRICAL INSULATION

3.1 Gaseous Insulation

3.1.1 Corona An understanding of electrical breakdown in gases can also be applied to a bubble model in liquids or a gaseous cavity model in solids. The fundamental phenomenon is described by the Paschen curve of Fig. 2. Representative voltage (v_P) and pressure-distance products at the Paschen minimum are listed for various gases in Table 2. When the applied electric field exceeds v_P/d, corona onsets, resulting in light, UV radiation, audible noise, radio-frequency interference, nitric acid in the presence of moisture and air, heat generation and power loss, and mechanical erosion of surfaces by ion impacts. Corona partial discharges are localized ionizations of gas forming a plasma of electrons, ions, excited molecules, and free radicals. These form chemically reactive species that often degrade nearby liquid and solid interfaces. Corona is often pulsating, as the buildup of charge near the electrode causes a decrease in the electric field below that necessary to sustain charge emission or dielectric ionization, as illustrated in Fig. 3(b). Corona can be reduced by rounding sharp edges to lower the local electric field. Above the Paschen minimum, an increasing pressure reduces the mean free path between collisions, and higher voltages are necessary for ionization, thereby preventing corona. At a pressure 10^{-4}–10^{-5} Torr, the mean free path is a few meters. Decreasing the pressure to 10^{-6}–10^{-8} Torr increases the mean free path to the order of 50 m.

Table 2. Typical Paschen minimum breakdown voltages and pressure–distance products for various gases [adapted from Meek and Craggs (1978) and Khalifa (1990)].

Gas	v_P (V)	pd at v_P (Pa m)
Air	327	0.754
Argon	137	1.197
Hydrogen (H_2)	273	1.530
Helium (He)	156	5.320
Neon	244	3.987
Carbon dioxide (CO_2)	420	0.678
Nitrogen (N_2)	251	0.891
Nitrous oxide (N_2O)	418	0.665
Oxygen (O_2)	450	0.931
Sulfur dioxide (SO_2)	457	0.439
Hydrogen sulfide (H_2S)	414	0.798
Mercury vapor (Hg)	~425	2.392

In air-insulated overhead power lines, a 350-kV line has corona losses ~4 kW/m in dry weather and ~70 kW/m in wet weather. Raising the voltage to 400 kV increases the corona losses to ~10 kW/m in dry weather and ~180 kW/m in wet weather.

The effects of corona can be reduced by increasing the conductor diameter to lower the electric field. This is most efficiently done using four parallel conductors spaced as a 6-in. square, which approximates the corona characteristics of a 6-in.-diam cable. Two parallel conductors 6–8 in. apart can operate up to ~500 kV.

Some corona characteristics can be applied in useful ways:

1. Xerography—The first step in copying an image is to lay down a uniform layer of static charge from a corona source over the surface of a light-sensitive plate.
2. Electrostatic separation of ores—Corona-charged conducting and nonconducting mineral particles can be separated using the Coulomb force.
3. Electrostatic paint spraying—Corona-charged paint articles will be directed by the electric field onto a grounded surface.
4. Printing on polymer surfaces—Thin films such as polyethylene cannot be wetted by printing inks. Chemical action from rf corona oxidizes the polymer to form surfaces that inks can wet.
5. Ozone generators—Ozone is generated by corona ionization of oxygen.
6. Geiger counter—With a voltage applied to a wire–cylinder electrode geometry to just below corona initiation, a passing cosmic ray causes an avalanche and a measurable current pulse.

Failures in electrically stressed gas-insulated systems usually occur at the interfaces of solid insulators used for mechanical support. Such surface flashovers depend on the materials, surface condition, particle contamination, and the electric fields at the triple junction of electrode, solid insulator, and gaseous insulation.

3.1.2 Streamer Mechanisms With high over-voltages above the Paschen curve in a uniform electric field gap, avalanches grow into irregular filamentary luminous branches known as streamers. According to streamer theory, ionization from an avalanche increases the conductivity, and the emitted light increases photoionization of gas molecules ahead of the streamer. This causes a space-charge distribution that increases the electric field at the tip of the streamer.

The time to breakdown has a statistical lag, t_s, of the average time after voltage is applied before an electron that initiates an avalanche appears in a gap, and a formative time lag, t_f, which is the time the discharge takes to completely develop from the time the initiating electron appears. The statistical time, $t_s \sim 10^{-2}$ s, can be made shorter by ultraviolet radiation of the cathode or large over-voltage of the gap. The formative time is of order $t_f \sim 10^{-4}$–10^{-9} s.

3.1.3 Electronegative Gas—SF_6 Gas-insulated systems (GISs) use sulfur hexafluoride (SF_6), an electronegative gas, as an arc-quenching medium in such high-voltage equipment as circuit breakers, switchgear, and underground cables above 123 kV, electrostatic accelerators, and x-ray equipment. A representative *E*/*p* value at the Paschen minimum is 86.2 V/m·Pa for SF_6, versus 23.6 V/m·Pa for air. Mixing a small amount of SF_6 with other less expensive gases such as nitrogen, hydrogen, carbon dioxide, or air increases the breakdown strength of those gases. For example, 20% by partial pressure SF_6 to nitrogen raises the breakdown strength to about 80% of that of pure SF_6. SF_6 is nontoxic but denser than air (so it can cause suffocation in a closed tank), nonflammable, noncorrosive, inert, and nonreactive. It decomposes above 500 °C. Its high molecular weight and low viscosity make it more effective with heat transfer than other gases. Its thermal time constant is about 1000 times shorter than that of air.

3.1.4 Arc Discharges Whenever a circuit breaker or load-break switch is opened while carrying a current, an arc forms between separating contacts. As the contacts move apart, the contact area decreases and the contact resistance goes up. The current becomes highly concentrated and melts metal. The metal vapor between contacts sustains the arc.

3.1.5 Vacuum Insulation Vacuum insulation is limited by Fowler–Nordheim electron emission [Eq. (20)] at microprotrusions on the electrode surface that can typically magnify the local electric field by a factor of 200–500. Electrical breakdown results when the Joule heating from the injected current leads to an explosive growth of plasma that initiates the formation of an arc. Microparticles on the electrode also detach and are accelerated by the electric field to impact explosively on the opposite electrode. The need for vacuum insulation is especially important in designing power systems for space.

3.2 Liquid Insulation

3.2.1 Characteristics Liquids are usually better insulators than solids and gases. Their high density from Paschen's law indicates a higher breakdown strength than gas. Liquid fills the space to be insulated and by thermal conduction and convection helps dissipate thermal losses.

Most insulating liquids used in transformers, cables, capacitors, switches, and circuit breakers are complex mixtures of paraffinic, naphthenic, and aromatic hydrocarbons. The paraffinic base is characterized by the chemical formula $C_{2n}H_{2n+2}$, the naphthenes by $C_{2n}H_{2n}$, and the aromatics by C_nH_n. Naphthenic oils are generally preferred because they are less viscous and have good gas-absorbing properties. Representative physical characteristics of mineral insulating oils for use in transformers, switches, and circuit breakers, with typical ASTM tests, are listed in Table 3.

These oils have relative dielectric constants of about 2, a dissipation factor of about 0.001 at 60 Hz, working electrical breakdown strength of about 120 kV/cm, and a resistivity greater than 10^{12} Ω cm. Their flammability is high, but the hazard from fire is minimal when operating at a temperature far below the flash point.

In transformers, the oil serves the dual function of high-voltage insulation and cooling with oil flow. Analysis of the insulating oil in an operating transformer provides information on the state of health of the system. This monitoring has become more important in recent years because of the trend toward higher voltages and smaller size,

Table 3. Representative physical characteristics and ASTM standards of mineral insulating oils [adapted from Clark (1962) and the ASTM Standards].

Property	
Condition	Clear [D1524]
Viscosity	12 cSt (40 °C) [D445, D88]
Specific gravity	0.91 [D1298]
Color (max)	0.5 [D1500]
Neutralization number	0.03 mg KOH/g [D974, D2440, D1533]
Flash point (min)	145 °C [D92]
Pour point (max)	−40 °C [D97]
Free sulfur	nil [D1275]
Total (fixed) sulfur	0.1% or less [D1275]
Evaporation (8 h/100 °C)	8%
Dielectric strength	30 kV/2.5 mm [D877] 28 kV/1.02 mm [D1816]
Specific heat (30–35 °C)	0.4252
Power factor (100 °C)	0.003 [D924]
Chlorides and sulfates	nil [D878]
Resistivity (100 °C)	(1–10) × 10^{12} Ω cm [D1169]
Coefficient of expansion	0.00070 [D1903]
Specific optical dispersion	110–115 [D1807]
Thermal conductivity	0.39 cal cm^{-1} sec^{-1}/°C
Refractive index (25 °C)	1.4828
Aniline point	63–78 °C [D611]

thereby increasing the design stress. This has led to the development of on-line hydrocarbon gas monitors, indicators of the degradation of cellulosic insulation, computerized methods to predict temperatures throughout transformer windings by calculating precise oil flow and heat flux, and development of a microprocessor-based transformer temperature monitor to improve monitoring and control of in-service transformers.

Moisture is the most common cause of deterioration, either as tiny droplets mixed with the oil as an emulsion or as a pool at the bottom of the transformer when the droplets coalesce to form larger drops that sink to the bottom. A slower but more damaging deterioration is the formation of acids and sludge, caused by oxidation. Fault conditions such as overheating (hot spots), partial discharges (corona), and continuous arcing lead to accumulation of combustible gases and to chemical degradation of the oil, resulting in changes in the molecular-weight distribution.

In the recent past, mineral-oil flammability was reduced by adding polychlorinated biphenyls (PCBs). However, later research has shown that the nonbiodegradability of PCBs is a health hazard so that current standards do not allow the use of PCB oil. This has resulted in a search for new natural or synthetic oils that are biologically and environmentally safe, yet still have high electric strength, low dielectric dissipation, good chemical stability, low volatility and high flash point, and good arc-quenching properties, and are nonflammable, nontoxic, and inexpensive. Silicone oil is one candidate for replacement for PCB-contaminated oil.

To decrease power dissipation in transformers, oil and pressboard insulation are dried to remove moisture. However, as drying processes have been improved, a new problem has arisen, flow-induced electrification. Here the mobile part of the electrical double layer that forms at the interface of dissimilar materials is swept away by oil flow, leaving the countercharge behind on the interface. The net accumulation of charge on highly insulating surfaces leads to strong electric fields and ultimately to spark discharges that can cause a transformer to fail.

Electrical breakdown in liquids is thought to be primarily due to the following mechanisms:

1. Electronic avalanches as in gases.
2. Bubbles that form near surface irregularities on electrodes that concentrate the electric field and cause Joule heating that boils the liquids to form bubbles. Once a gas region is formed, avalanche theory describes the breakdown phenomena in the gaseous bubble. Figure 6 shows a hydrocarbon liquid stressed by a high-voltage pulse. The electron avalanches within the vapor region grow into treelike structures at the cathode, which can initiate the breakdown streamer shown.
3. Suspended particles that form chains along the electric field to partially short circuit the gap.

3.2.2 Insulation Properties and Standards From the current state of knowledge of oil aging and degradation, the power industry has developed measurement proce-

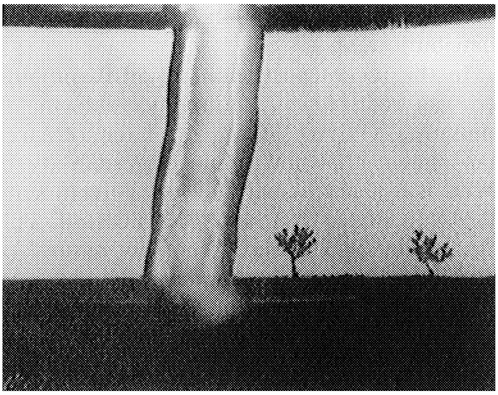

FIG. 6. Parallel-plate electrodes with a hydrocarbon dielectric (toluene) stressed by a 250-kV pulse show electrical trees emanating from the cathode, one of which led to the breakdown streamer that short circuits the 5-mm gap. The photograph was taken about 200 ns after electrical breakdown.

dures, interpretation, and corrective actions for monitoring the electrical, physical, chemical, and thermal properties of oil. The important properties are listed and defined below together with the appropriate ASTM tests.

3.2.2.1 Electrical Properties *Dielectric Breakdown Voltage* [D877, D1816, D3300] is important at 60-Hz power frequency as a measure of the liquid's ability to withstand electric stress at operating and test voltages without failure. Low values indicate the presence of contaminants such as water, dirt, or conducting particles. Impulse testing is used to simulate transient voltage stresses arising from such causes as lightning strokes and high-voltage switching.

Gas Evolution [D2298, D2300] determines the rate of gas evolution or gas absorption under ASTM-specified test conditions of electrical discharge.

Power Factor [D924], the ratio of power dissipation in watts to the product of voltage and current in volt-amperes, equivalent to cos ϕ where ϕ is the phase angle between voltage and current, is a measure of the oil dielectric losses and thus the amount of energy dissipated as heat.

Resistivity [D1169]—High resistivity reflects low content of free ions and ion-forming particles and normally indicates a low concentration of conductive contaminants. Resistivity of less than 15 MΩ cm generally indicates problems.

3.2.2.2 Physical Properties *Aniline Point* [D611]—Testing provides an estimate of the aromatic hydrocarbon content of mixtures and indicates the solvency of the oil for some materials.

Coefficient of Thermal Expansion [D1903] is used to compute the size of the transformer structure to accommodate oil volume over the full temperature range of operation.

Color [D1500]—Oil gets darker with age because of formation of acids, resins, and sludge from insulating materials in the transformer. The color should not progress beyond a light brown, remaining clear and free from turbidity or cloudiness. Black coloring or floating carbon particles indicates severe arcing.

Visual and Qualitative Infrared Absorption Examination [D1524, D2144]—Visual examination of color and homogeneity of the oil permits a determination of whether the oil should be sent to the laboratory for testing. The infrared spectrum indicates the general chemical composition of the oil.

Flash and Fire Point [D92]—The flash and fire points indicate the flammability of an oil. The flash point is the temperature at which vapors above the oil surface first ignite when a small test flame is passed across the surface. The fire point is the temperature at which the oil first ignites and burns for at least 5 s when a small test flame is passed across the surface.

Interfacial Tension [D971, D2285] is a reliable indication of the presence of hydrophilic compounds, which are soluble polar contaminant products of oxidation. A high interfacial tension indicates the absence of undesirable polar contaminants. The test is frequently applied to service-aged oils as an indication of the degree of deterioration.

Molecular Weight [D2224] is a basic quantity required in analysis to convert weight fractions to mole fractions.

Pour Point [D97] is the lowest temperature expressed as a multiple of 5 °F (or 3 °C) at which the oil is observed to flow when cooled without seriously limiting the degree of circulation.

Refractive Index and Specific Optical Dispersion [D1218, D1807]—Changes in refractive index with service indicate changes in composition or the degree of contamination.

Specific optical dispersion is the difference between the refractive indices of light of two different wavelengths at the same temperature, divided by the specific gravity. It serves as a quick index to the amount of unsaturated compounds present in the oil.

Specific Gravity [D287, D1298, D1481]—Insulating oils are usually sold on the basis of volume delivered at 15.6 °C (60 °F), while delivery is usually based on weight. The specific gravity influences the heat-transfer rates. In cold climates, ice may be formed in de-energized transformers below 0 °C, and the maximum specific gravity of the oil should always be less than that of ice, so that ice will settle out at the bottom. When new oil is mixed into apparatus in service, differences in specific gravity may indicate a tendency of the two oils not to mix, which can cause overheating of self-cooled apparatus.

Steam Emulsion Number [D1935] provides an approximate evaluation of the ability of mineral insulating oils to separate from water. The test indicates the presence of contaminants such as metallic soaps, asphaltic materials, etc.

Viscosity [D88, D445, D2261] influences heat-transfer properties and consequently their temperature rise. At low temperatures, the higher viscosity influences the speed of moving parts such as load tap-changer mechanisms and pumps. Viscosity controls mineral insulating oil processing conditions such as dehydration, degasification and filtration, and oil impregnation rates. Viscosity affects pressure drop, oil flow, and cooling rates in circulating oil systems.

3.2.2.3 Chemical Properties *Acidity* [D1534, D1902] in used insulating oils can be compared to that of new oil in detecting contaminants or chemical changes and can be used as a measure for determining when an oil should be replaced to prevent further decomposition and consequent sludging. For a power transformer, 500 kV A and larger, operating at 60 °C maximum, the transformer oil should reach the critical acid number of 0.25 mg KOH/g in about 15 years.

Carbon-Type Composition [D2140]—Percentages of aromatic carbons, naphthenic carbons, and paraffinic carbons relate chemical nature of an oil to other properties related to oil composition.

Characteristic Groups [D2007]—A rapid procedure for classifying mineral insulating oil into hydrocarbon types of polar compounds, aromatics, and saturates.

Compatibility with Construction Material [D3455]—Screening for compatibility of oil with varnishes, dip coatings, core steel, gaskets, and wire enamels.

Copper Content [D2608, D2675]—The presence of copper accelerates the deterioration of oil.

Gas Analysis [D3612] covers the extraction and measurement of gases dissolved in electrical insulating oil having a viscosity of 20 centistokes (100 SUS) or less at 40 °C, and the identification and determination of the individual component gases extracted, including hydrogen (H_2), nitrogen (N_2), carbon monoxide (CO), carbon dioxide (CO_2), oxygen (O_2), methane (CH_4), ethane (C_2H_6), ethylene (C_2H_4), acetylene (C_2H_2), propane (C_3H_8), and propylene (C_3H_6).

Gas Content [D831, D1827, D2945]—Gas bubbles can lead to gaseous ionization in high fields. A low gas content reduces foaming and reduces available oxygen in sealed equipment, increasing service life of the oil.

Inorganic Chlorides and Sulfates [D878]—Presence of inorganic chlorides and sulfates indicates improper refining or contamination affecting corrosivity and oil dielectric properties.

Neutralization Number [D664, D974] determines acidic or basic constituents in oil. The total acid number is the number of milligrams of KOH required to neutralize all acidic constituents in a 1-g sample. A low acid number is necessary to minimize electrical conduction and metal corrosion to maximize the insulation life. In used oils, an increase in acid number from the value for new oil indicates contamination or chemical changes such as oxidation. The acid number should not increase beyond 0.10 mg KOH/g under normal operating conditions.

Oxidation Inhibitor Content, Phenolic [D1473, D2668]—New electrical insulating oil may contain phenolic oxidation inhibitors, 2,6-ditertiary-butyl para-cresal (DBPC) and 2,6-ditertiary-butyl phenol. DBPC reduces oil color, acidity, and sludge in operating transformers.

Oxidation Stability [D1313, D2112, D2440]—Minimizing the development of oil sludge and acidity resulting from oxidation during storage, processing, and long service life reduces electrical conduction and metal corrosion and increases insulation system life, electrical breakdown strength, and heat transfer. Open breathing transformers have faster oxidation deterioration than sealed transformers. The rate of oxidation generally doubles with each 10 °C increase.

Peroxide Number [D1563] magnitude indicates the quantity of oxidizing constituents present. Deterioration of oil results in the formation of peroxides and other oxygen-carrying compounds. The peroxide number measures those compounds that will oxidize potassium iodide.

Saponification Number [D94], the number of milligrams of KOH consumed by 1 g of oil under prescribed conditions. Components that react with KOH may consist of oil oxidation products, added compounds, or contaminants.

Sediment and Soluble Sludge [D1698] determines the amount of sediment and soluble sludge in service-aged oils, including organic and inorganic content. Sediment may deposit on transformer parts and interfere with heat transfer, and may choke oil ducts and thus hinder oil circulation and heat dissipation. Inorganic sediment indicates contamination, and organic sediment indicates deterioration or contamination.

Sulfur, Corrosive [D1275]—Sulfur compounds that severely discolor a copper surface are termed corrosive.

Water Content [D1315, D1533]—In ASTM test D1315, "Test for Water in Insulating Oils by Extraction," moisture is removed from oil by heat, vacuum, and shaking. The moisture is absorbed in a P_2O_5 trap and is determined by the weight gain. ASTM test D1533, "Water in Insulating Liquids," uses the Karl Fischer method, usually in the range of 0–75 ppm, in insulating liquids having a viscosity less than 19 centistokes (100 SUS) at 40 °C. The method is based essentially upon the reduction of iodine by sulfur dioxide in the presence of water. A low water content is necessary to achieve adequate electric strength and low dielectric losses, maximize insulation life, and minimize metal corrosion. A normal power transformer will have a water content under 2% in the paper and 10–20 ppm in the oil.

3.2.3 Analysis of Gas in Transformer Oils

3.2.3.1 Background. A study of gas formation in transformer oil stressed by high fields frequently gives an early indication of abnormal behavior in a transformer and can allow corrective action before the equipment suffers greater damage. Improvements in available instrumentation allow utilities to adopt oil-analysis programs as failure prevention programs rather than just failure analysis programs. The distribution of gases is related to the type of fault, and the rate of gas formation is indicative of the severity of the fault.

3.2.3.2 Gas Cause and Origin. Fault gases are generally due to corona or partial discharge, pyrolysis or thermal heating, and arcing. The hydrocarbon gases methane (CH_4), ethane (C_2H_6), ethylene (C_2H_4), and acetylene (C_2H_2) are produced by mineral oil decomposition. Carbon dioxide (CO_2) and carbon monoxide (CO) arise from thermal decomposition of the cellulosic insulating material. Hydrogen (H_2) is a component of both cellulosic material and any water present in the oil. Oxygen (O_2) and nitrogen (N_2) are not considered fault gases but are usually present.

Arcing causes large amounts of hydrogen and acetylene to be produced, with minor quantities of methane and ethylene. Carbon dioxide and carbon monoxide may also be formed if the fault involves cellulose. The oil may be carbonized.

Low-energy electrical discharges from corona produce hydrogen and methane, with small quantities of ethane and ethylene. Comparable amounts of carbon monoxide and dioxide may result from discharges in cellulose.

Decomposition products from overheated oil include ethylene and methane, together with smaller quantities of hydrogen and ethane. Traces of acetylene may be formed if the fault is severe or involves electrical contacts.

Large amounts of carbon dioxide and carbon monoxide are evolved from overheated cellulose. Hydrocarbon gases, such as methane and ethylene, will be formed if the fault involves an oil-impregnated structure.

All these gases are dissolved in the oil as well as in the gas blanket above the oil. Because of differences in solubilities of these gases in the oil, their distribution in the oil and gas blanket will differ. Equal quantities of hydrogen and acetylene in the gas blanket will result in concentrations in the oil that differ by almost two orders of magnitude because hydrogen is less soluble in oil, while acetylene is more soluble in the oil.

The energy released by any fault, including heat, fragments the oil into smaller hydrocarbon chains and hydrogen. Corona leads to large amounts of hydrogen, sparking gives rise to methane and ethane, severe local heating causes the formation of ethylene, and arcing forms acetylene. These are the key gases for each fault type, but other gases can also be formed. Faults near cellulosic materials give rise to carbon dioxide and carbon monoxide with the ratio of carbon monoxide to carbon dioxide increasing as the severity of the faults increases.

3.2.3.3 Detection Methods. The most widely used method is the determination of the total combustible gases present above the oil. It is rapid and can be continuously monitored, but it does not determine which fault gases are present and therefore cannot indicate either the type or severity of a fault.

Normal oil aging has up to 500 ppm fault gases. Decomposition may be in excess of normal aging with gas levels of 500–2500 ppm, while severe problems are developing for gas levels above 2500 ppm.

A rate of increase of combustible gas generation of 100 ppm or more for a 24-h period, on a continuing basis, with a relatively constant load, indicates a deteriorating condition and requires an assessment of continued operation, especially if acetylene is in excess of 20 ppm.

Dissolved-gas analysis (DGA) takes a sample of oil, and the dissolved gases are extracted and then separated, identified, and quantitatively determined with a gas chromatograph to detect fault gases at the earliest point in time and can ascertain the type and severity of the fault.

Because dissolved hydrogen in oil is of the same order of concentration as the other fault gases, hydrogen monitors using hydrogen diffusion through a Teflon or polyimide membrane have been developed for cheaper and easier measurements that allow continuous monitoring.

3.2.3.4 Interpretation Theories. Faults of thermal or electrical origin can be differentiated by comparing pairs of gases with approximately equal solubilities and diffusion coefficients, as described by Rogers (1978). Thus, an increase in the ratio of ethylene to acetylene above unity indicates an electrical fault, while the ratio of methane to hydrogen exceeding 0.1 suggests a thermal fault and less than 0.1 suggests a corona discharge.

A thermodynamic assessment of the formation of the simple decomposition gaseous hydrocarbons based on equilibrium pressures at various temperatures suggests that the proportion of each gaseous hydrocarbon in comparison with each of the other hydrocarbon gases varies with the temperature of decomposition. This leads to the assumption that the rate of evolution of any particular gaseous hydrocarbon varies with temperature and that at a particular temperature there would be a maximum rate of evolution where each gas attains the maximum rate at a different temperature. This thermodynamic equilibrium suggests that with increasing temperature, the maxima would be in turn methane, ethane, ethylene, and acetylene.

3.3 Solid Insulation

3.3.1 Characteristics Solid insulation is generally used in the form of porcelain, glass, and epoxy resin high-voltage bushings; in oil-impregnated cellulose in transformers; in polymer power cables; in thin-film polymer capacitors; and in mica insulation where high temperature and surface discharges are experienced. The primary advantage of solid insulation is that it also provides mechanical strength, but its primary disadvantage is that solids are not self-healing after electrical discharges, as a conducting channel remains. In carefully controlled laboratory conditions, the breakdown strength in thin-film solids can be as high as 1–15 MV/cm, but in practical working systems the working stress drops to ~200 kV/cm. The primary failure mechanism is electrical discharge across voids. In a free space void with relative dielectric constant of 1 as compared with the typical relative dielectric constant in solids of 2–3, the ac electric field in the void is about 2–3 times that in the solid, often leading to spark discharges.

Thin films usually have many defects of pin holes and irregular thickness, but with thin metallic electrodes they can be self-healing after electrical breakdown as the heat released by the breakdown evaporates the metal, reducing the electric stress across the breakdown channel. The remaining part of the film can be high-voltage stressed.

Glass and ceramic solids can typically work up to ~250 °C. Thermoplastic polymers can be molded and extruded at about 100–120 °C. Heat-curing solids, such as epoxy, have substantial mechanical strength and hardness and prevent the entry of moisture. Epoxy insulators reinforced with glass fibers weigh about one-fourth to one-half as much as corresponding porcelain insulators for the same mechanical and electrical strength.

Many polymers of extremely large molecular size such as polyethylene (C_2H_4), polypropylene (C_3H_6), polyvinylchloride (C_2H_3Cl), polystyrene (C_8H_9), polymethylmethacrylate ($C_5H_9O_2$), and polytetrafluoroethylene (C_2F_4) are highly insulating as a consequence of their covalent bonding. Conduction is usually due to polar or ionic impurities. Cross-linking by chemical or radiation means joins molecular chains at many points to form an extensive, insoluble network. Cross-linked polyethylene (XLPE) has been a replacement for thermoplastic polyethylene in power cables because it can operate at higher temperatures (~90 °C versus ~70 °C for polyethylene) and has improved chemical resistance. It has a higher breakdown strength and longer life, primarily due to its resistance to treeing from acetophenone, which results from a peroxide chemical cross-linking reaction that acts as a tree-retardant additive by absorbing energetic electrons. In the past XLPE was produced by steam curing at 200–220 °C at a pressure of 1.6–2 MPa, but partly in order to minimize absorption of moisture, present processes often use dry curing with hot nitrogen.

Surface failures can occur as a result of surface contamination, cracks, imperfections, and roughness. Insulators should have adequate electrical and mechanical strength with temperature variation, rain, and pollution. When pollution accumulates on insulator surfaces and its soluble ingredients dissolve in moisture from dew or fog, a conducting path is established from the high-voltage conductor to ground. The leakage currents form hot spots that dry up the moisture, forming dry bands. The dry band takes on a major fraction of the voltage drop, causing partial arcs leading to flashover.

3.3.2 Treeing The major degradation process in solids is the formation of electrical and water trees. Treeing is an electrical pre-breakdown phenomena that often initiates at a point of high electric stress and progresses by periodic partial discharges. Treeing lowers the electric breakdown strength and often leads to electric breakdown, and it is the most likely aging mechanism in solid extruded polymers. Electrical trees in solids have a similar appearance to the “trees” shown in Fig. 6 for an electrically stressed hydrocarbon liquid.

Electrical trees are dark, hollow channels resulting from the decomposition of material and are characterized by carbonized branch or bush-type dendrites. Electrical trees initiate from electron injection and mechanical fatigue due to periodic electrical stresses, Joule heating and thermal decomposition, and the electric stress enhancement due to contaminants and voids. Water trees are very fine filamentary paths between small cavities often caused by cracks from pressure through which moisture penetrates driven by an electric field. They are more bushlike or fanlike than electrical trees. Vented electrical or water trees start at an insulation surface, while nonvented (bow-tie) trees grow from contaminants or cavities within the insulation and progress symmetrically outward parallel to the electric field toward both electrodes. Water trees can be made visible by means of methylene blue dye.

Polymer cables are typically used in HVDC power transmission, space and airborne applications, and x-ray machines. HVDC can compete economically with HVAC and offers advantages in long underground, submarine, and long overhead lines, and as interconnections between large ac systems that differ in frequency such as in Japan, which has power networks that operate at 50 or 60 Hz.

The main form of cable deterioration under ac conditions is discharges. At the discharge inception voltage there are four discharges per cycle, which gradually causes erosion and electrical trees that rapidly propagate through the insulation to complete the breakdown. Under dc, the repetition rate of partial discharges is several orders of magnitude less than for similar ac voltages.

4. MEASUREMENT METHODOLOGIES OF ELECTRICAL INSULATION

Electrical insulation applications and research have a venerable history and continue to play an essential role in modern technology. As new materials are developed and as insulation system requirements are increased, the need for improved measurement methods continues to use the most modern tools. In this section current measurement methodologies are briefly described.

4.1 Dielectrometry Measurements

4.1.1 Sinusoidal Steady State The most direct measurement of the electrical properties of insulators is from terminal voltage–current measurements of electrical impedance. Under ac steady-state operation, every electrical quantity $F(t)$ may be expressed as the real part of a complex quantity varying at radian frequency ω:

$$F(x,y,z,t) = \mathcal{R}\{\hat{F}(x,y,z)^{j\omega t}\}. \tag{66}$$

If permittivity ϵ and conductivity σ are constant with time, we may write the total current density as a sum of displacement and conduction current densities. In terms of complex amplitudes,

$$\begin{aligned}\hat{J} &= \hat{J}_d + \hat{J}_c = j\omega\epsilon\hat{E} + \sigma\hat{E} \\ &= j\omega\hat{E}(\epsilon - j\sigma/\omega) = j\omega\epsilon^*\hat{E},\end{aligned} \tag{67}$$

where $\epsilon^* = (\epsilon - j\sigma/\omega)$ is the complex permittivity.

The dielectric spectrum of a material is a representation of its complex permittivity $\epsilon^* = \epsilon' - j\epsilon''$ as a function of frequency. The real component, $\epsilon' = \epsilon$, gives the dielectric constant while the imaginary component, $\epsilon'' = \sigma/\omega$, determines the power dissipation (loss) in the material.

In an Ohmic material, ϵ and σ are independent of the frequency or amplitude of the applied electric field, and a plot of log (ϵ''/ϵ_0) versus log ω has a slope -1. In a dispersive material when ϵ'' is plotted against frequency on a log-log scale, it can be characterized by one or more loss peaks.

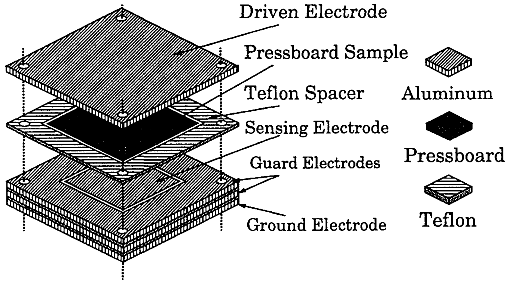

FIG. 7. Structure of the parallel-plate dielectrometry sensor used by Sheiretov and Zahn (1995).

4.1.2 Parallel-Plate Sensor The simplest measurement is to place the material between a pair of parallel-plate electrodes of known area and separation as described by Sheiretov and Zahn (1995). This test cell can be modeled as a resistor in parallel with a capacitor. The complex admittance of the structure can be measured, and from that its permittivity and conductivity can be calculated.

The parallel-plate sensor shown in Fig. 7 is used to measure the dielectric spectrum of typical transformer insulation, oil-impregnated pressboard, as a function of moisture and temperature. The capacitive structure is composed of the driven electrode, the sensing electrode, and guard electrodes. The input admittance, with which the sensing electrode is loaded, is that of a known parallel RC pair, modeled as shown in Fig. 8 with

$$R_T = d/\sigma A, \quad C_T = \epsilon A/d, \tag{68}$$

where d is the plate separation distance, A is the sensing electrode area, and σ and ϵ are the material's conductivity and permittivity, respectively.

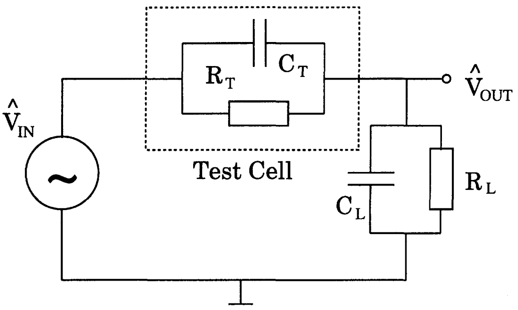

FIG. 8. Equivalent circuit of the test structure for dielectrometry measurements.

For linear time-invariant (LTI) systems we take the standard complex sinusoidal form:

$$v_{\mathrm{IN}} = \Re\{\hat{v}_{\mathrm{IN}} e^{j\omega t}\} \quad v_{\mathrm{OUT}} = \Re\{\hat{v}_{\mathrm{OUT}} e^{j\omega t}\} \tag{69}$$

with complex voltage amplitudes $\hat{v}_{\mathrm{IN}}$ and $\hat{v}_{\mathrm{OUT}}$ defined in Fig. 8. A microprocessor-driven controller is responsible for generating the driving voltage and measuring the output voltage. The data that it produces are expressed in terms of a magnitude 20 log(M) [dB] and phase $180\varphi/\pi$ [deg], which are related to the complex amplitudes defined in Fig. 8 as

$$\frac{\hat{v}_{\mathrm{OUT}}}{\hat{v}_{\mathrm{IN}}} = Me^{j\varphi} \Rightarrow M = \left|\frac{\hat{v}_{\mathrm{OUT}}}{\hat{v}_{\mathrm{IN}}}\right|,$$
$$\varphi = \angle\left(\frac{\hat{v}_{\mathrm{OUT}}}{\hat{v}_{\mathrm{IN}}}\right) = \arctan\frac{\Im\{\hat{v}_{\mathrm{OUT}}/\hat{v}_{\mathrm{IN}}\}}{\Re\{\hat{v}_{\mathrm{OUT}}/\hat{v}_{\mathrm{IN}}\}}. \tag{70}$$

The next step in calculating σ and ϵ is to calculate R_T and C_T from measurements of M and φ, and the known values of R_L and C_L. We define the admittances of the test and load branches as $Y_T = 1/R_T + j\omega C_T$ and $Y_L = 1/R_L + j\omega C_L$, respectively. Then from the voltage-divider relationship we obtain

$$\frac{\hat{v}_{\mathrm{OUT}}}{\hat{v}_{\mathrm{IN}}} = \frac{Y_T}{Y_T + Y_L} = \frac{1/R_T + j\omega C_T}{1/R_T + 1/R_L + j\omega(C_T + C_L)} = Me^{j\varphi}. \tag{71}$$

Solving (71) for R_T and C_T in terms of known values for R_L and C_L and measured values for M and φ yields

$$R_T = \frac{1 + M^2 - 2M\cos\varphi}{M[\cos\varphi - M - \omega R_L C_L \sin\varphi]} R_L,$$
$$C_T = \frac{M[\cos\varphi - M + (1/\omega R_L C_L)\sin\varphi]}{1 + M^2 - 2M\cos\varphi} C_L. \tag{72}$$

Representative measurements in Fig. 9 show that Shell Diala A transformer oil is Ohmic. On a log-log scale, the plot of ϵ'' versus frequency is a straight line of slope -1, which means that σ is independent of the frequency. For linear dielectric materials, ϵ' should also be constant with frequency. The observed rise of ϵ' at low frequencies can be attributed to electrical double-layer formation at the aluminum-oil interfaces.

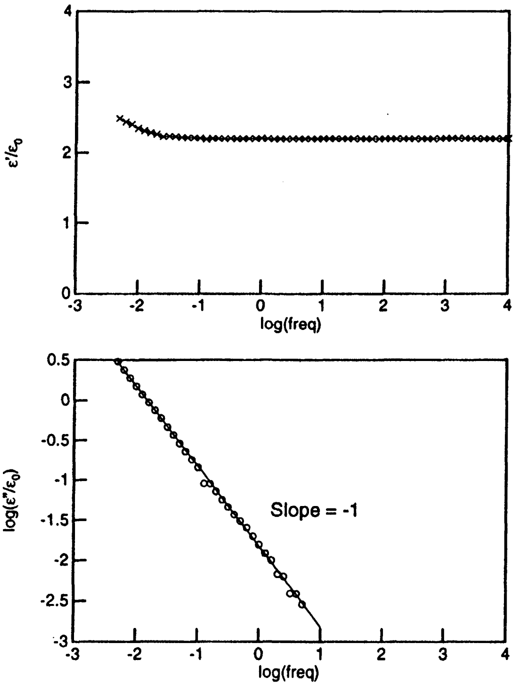

FIG. 9. Complex permittivity of transformer oil measured with the parallel-plate sensor. The -1 slope for the loss measurement indicates that transformer oil is Ohmic. The electrical double layers at the electrode/oil interfaces cause the low-frequency value of ϵ' to increase.

The plot of ϵ'' is not shown for frequencies higher than $10^{0.7}$ Hz. This is because at that high-frequency range the response is dominated by the capacitive element, and the gain/phase measurement is insensitive to the value of σ.

4.1.3 Oil-Impregnated Pressboard

The parallel-plate sensor was also used to study how the dielectric spectrum of oil-impregnated pressboard changes with variations in temperature and moisture content in order to model temperature and moisture dynamics in an operating transformer. First, many samples of pressboard were prepared, each with a different content of water. Then the moisture content of a small piece of each sample was measured, and the remaining sample was placed in the sensor structure, its dielectric spectrum was scanned at five different temperatures, and finally its mois-

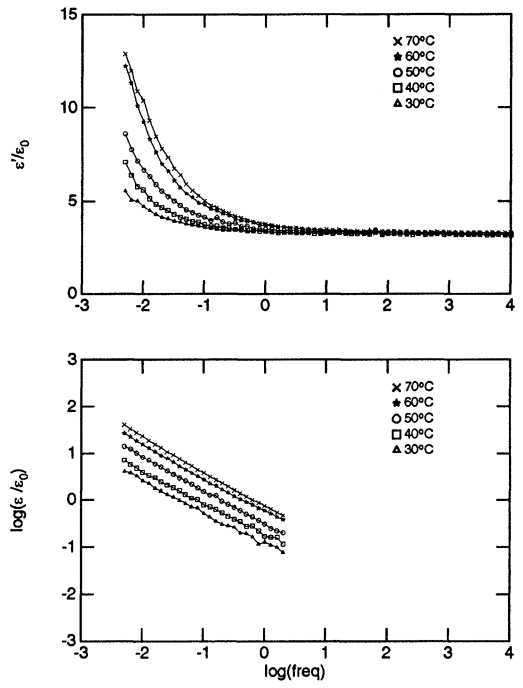

FIG. 10. Dielectric spectra of an oil-impregnated pressboard sample of 2.3% moisture at five temperatures. The slope of the loss characteristic is ~ −0.7, indicating that pressboard is dispersive.

ture content was measured again. Figure 10 shows representative dielectric spectrum measurements as a function of temperature for pressboard with 2.3% moisture. The slope of log ϵ'' versus log ω is -0.7, indicating that pressboard is dispersive, i.e., the conductivity is a function of frequency.

4.1.4 Imposed *ω-k* Dielectrometry If electrodes are placed side by side in a repeating pattern on one surface of the material, the electric fields will decrease away from the electrodes, and the complex impedance between the electrodes will be most sensitive to the material adjacent to them. Periodic electrode structures with a multitude of interdigitated fingers, as shown in Fig. 11, have been designed. The electric fields are uniform in the z direction and periodic in the y direction with a spatial wavelength of λ, equal to the distance between repeating electrodes. Thus at every surface of constant x, the electric potential is periodic in y and can be expanded as an infinite series of sinusoidal Fourier modes of spatial wavelengths $\lambda_n = \lambda/n$. This is very convenient, because the two-dimensional solutions to Laplace's equation $\nabla^2\Phi = 0$ in Cartesian geometry are of the form

$$\Phi(x,y) = \sum_{n=1}^{\infty} \Phi_n \,\mathrm{hyp}(k_n x)\,\mathrm{trig}(k_n y), \tag{73}$$

where $k_n = 2\pi/\lambda_n$ is the wave number of the nth mode; hyp(x) stands for any one of the hyperbolic exponential functions sinh(x), cosh(x), e^x, or e^{-x}; and trig(y) stands for one

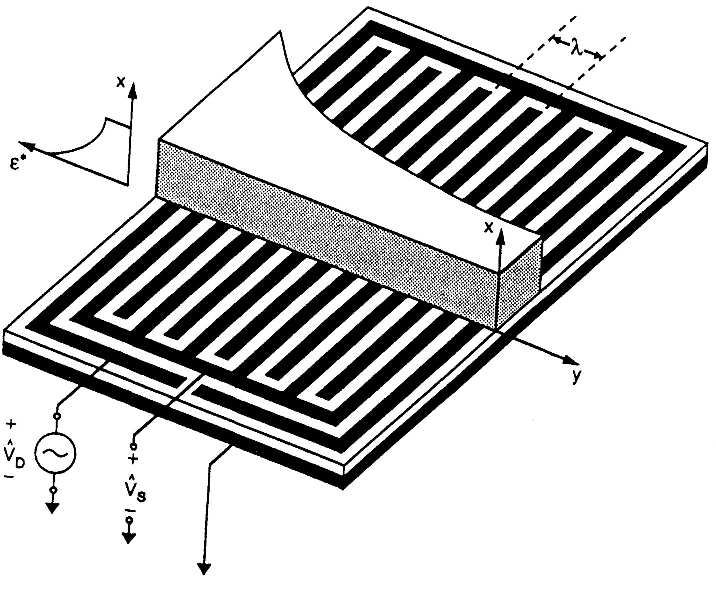

FIG. 11. Interdigital electrode structure for imposed *ω-k* dielectrometry measurements to measure the complex permittivity ϵ^* of the dielectric material ($x > 0$) above the interdigitated electrodes.

of the trigonometric functions $\sin(y)$ or $\cos(y)$. For every Fourier mode n, the electric fields decrease with x as $\exp(-2\pi nx/\lambda)$ with the fundamental mode $n = 1$ penetrating farthest into the material.

By designing sensors with various fundamental spatial wavelengths λ, it is possible to measure the dielectric properties of materials at different depths. Combining the results from several such sensors makes it possible to determine the x-dependent spatial profiles of the complex permittivity. The three-wavelength sensor structure shown in Fig. 12 consists of three sets of interdigitated electrodes, deposited on a common flexible Kapton (a polyimide) substrate. Current research is using this sensor to measure the moisture diffusion profiles in transformer pressboard insulation as the temperature changes.

4.2 Flow Electrification

4.2.1 Background to the Problem Electrification has been long recognized as a problem in the flow of petroleum liquids in pipes, through filters, and in charge accumulation in storage tanks. With changes in the chemical makeup of insulating liquids used both to transfer heat and/or to withstand high electric stress, the use of new dielectric materials such as polymers and cellulosic materials, and an increase in the flow speeds for greater cooling, the electrification problem has recently arisen in electric power apparatus and in automotive fuel systems. Polymeric replacements for metal parts in automotive fuel lines, filters, and pumps have caused flow-electrification–initiated discharges that are hazardous.

A number of failures of large oil-cooled power transformers and in a trichlorotrifluorethane-cooled HVDC valve have been attributed to flow-induced electrification. Oil is generally used to cool the windings of a power transformer with the flow by either forced or natural convection through a heat exchanger to dissipate the generated heat. The moving oil provides the transport of entrained charge separated at liquid/solid interfaces by physical, chemical, and electrical phenomena. Moving transformer oil generally entrains positive charge in the charged oil through the transformer structure, with the oil playing the role of the belt of a Van de Graaff generator. Charge separation at interfaces between moving fluid and boundaries with the accumulation of charge on insulators or isolated conductors can lead to high field strengths and electrical discharges. The potential builds up until the rate of charge accumulation equals the rate of charge leakage, or until spark discharges occur.

4.2.2 The Absolute Charge Sensor To overcome measurement uncertainties from charge separation at surfaces of charge-measuring devices, an absolute charge sensor (ACS) using Faraday-cage principles has been developed that provides an absolute measurement of the charge density that is independent of the fluid's electrical properties, the velocity of the fluid, and any electrification process within the instrument. More information about the ACS can be found in the reference by Morin *et al.* (1991a).

A schematic of the ACS is illustrated in Fig. 13. Fluid such as transformer oil or hydrocarbon fuel is drawn into an electrically conductive container as the metal bellows expands. As the ACS is sampling fluid, either the voltage or the current is measured by a high- or low-impedance electrometer, respec-

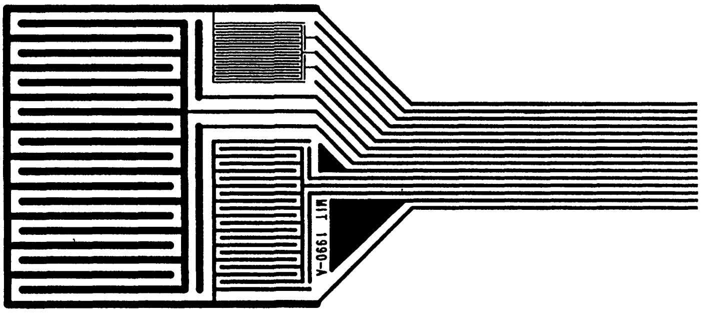

FIG. 12. Structure of a three-wavelength interdigitated dielectrometry sensor to measure spatial distributions of complex permittivity.

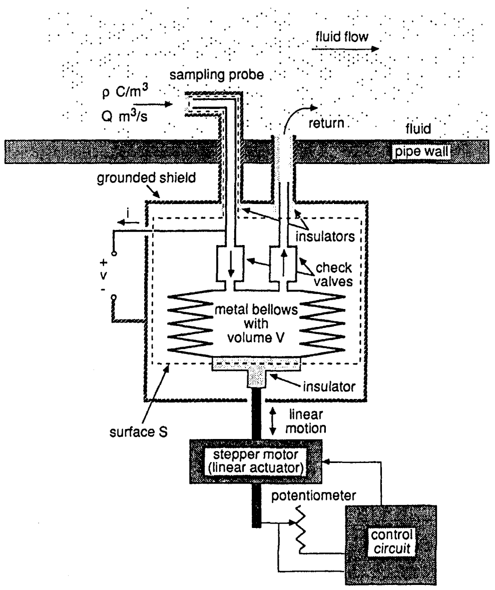

FIG. 13. Electromechanical schematic of the absolute charge sensor (ACS).

tively, connected across the electrical terminals. Once the measurement is complete, the fluid is expelled through the exit check valve and conduit by reversing the direction of the actuator to compress the bellows volume.

A charged body introduced into a perfectly insulating volume surrounded by a conductive container will induce an opposite charge on the interior wall of the container. Thus, without actually being transported to the wall, the net charge induced onto the conductive container inner wall would be equal to but of opposite polarity to that in the fluid. Because the rate of accumulation of the charge so induced on the inner wall of the container constitutes a capacitive electrical current that can be measured externally, it is then possible to measure this net charge by means of an external circuit having a sufficiently low impedance that the potential of the container would remain essentially zero relative to that of the surrounding shield. Because the conductive container forms a Faraday cage, the current measured during the filling cycle is independent of charging processes that occur inside the Faraday cage whether they are due to imaging, charge relaxation, or electrification. The total charge within the ACS with entrained volume charge is balanced by image surface charge on the wall. Thus, as soon as charge enters the ACS inner probe conduit, it is immediately recorded by the electrometer, whether the charge remains entrained in the fluid or is conducted to the wall. The charge density ρ is related to the short-circuit current i as $\rho = i/Q$, where Q is the fluid flow rate into the ACS.

4.2.3 Flow-Loop Measurements A typical measurement flow loop that models oil circulation for cooling in a transformer is shown in Fig. 14. Similar flow loops are also used to model fuel flows in vehicles. In Fig. 14, a pump drives transformer oil through piping connected to a large-volume reservoir (55-gal. oil drum), electrically isolated from the grounded piping by insulated tubing. The system includes a parallel flow loop that is electrically isolated by nylon fittings and contains a paper oil filter that can be valved in or out through valves 2 and 3, a gate valve (valve 1) to control the flow rate through the main flow path, and a flow meter to monitor the total flow rate. The flow speed could also be independently controlled by varying the pump impeller speed from the ac motor controller. The paper oil filter is used as a source of electrification charge that can be controlled by appropriate setting of valves 2 and 3. If all the flow is directed through the filter by fully closing valve 1 and fully opening valves 2 and 3, the maximum flow rate is 18 gal/min and the volume charge density is of order 100–500 μC/m^3 depending on oil temperature over the range of 65–100 °F. If none of the flow passed through the filter as a result of fully opening valve 1 and fully closing valves 2 and 3, the maximum flow rate is 26 gal/min and the volume charge density is of order 4–10 μC/m^3.

The electrometer current from the oil drum equals the difference in convection currents entering (measured by ACS 1) and leaving (measured by ACS 2) the oil drum. The ACS-measured charge densities at the inlet, ρ_{in}, and outlet, ρ_{out}, of the oil drum with flow rate Q give the difference in convection currents, i_{acs}, as

$$i_{acs} = (\rho_{in} - \rho_{out})Q. \tag{74}$$

From conservation of charge, the difference in convection currents should equal the total circuit current flowing from the drum

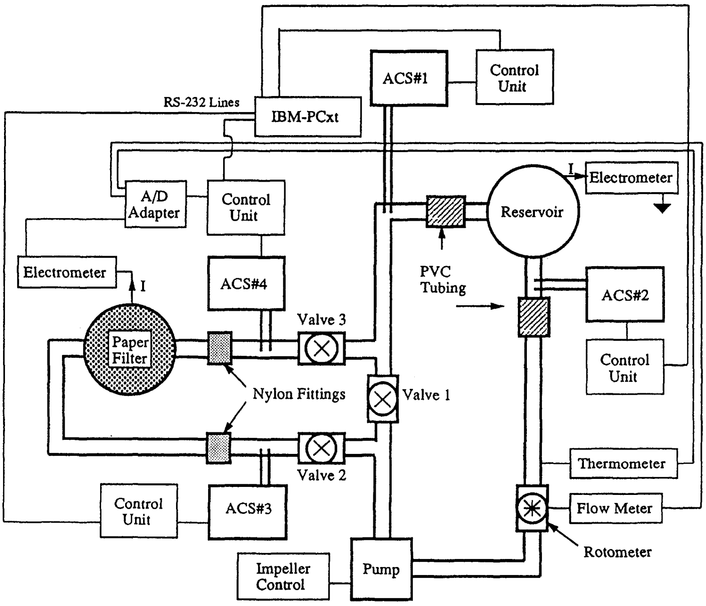
RS-232 Lines
IBM-PCxt
ACS#1
Control Unit
A/D Adapter
Control Unit
Electrometer
I
ACS#4
Valve 3
Paper Filter
Nylon Fittings
Valve 1
Valve 2
Control Unit
ACS#3
Impeller Control
Pump
PVC Tubing
I
Electrometer
Reservoir
ACS#2
Control Unit
Thermometer
Flow Meter
Rotometer

FIG. 14. A pump drives transformer oil through a closed flow loop connected through a large-volume reservoir (55-gal. drum). Two ACSs monitor the charge densities at the inlet (1) and outlet (2) to the electrically isolated reservoir, which is virtually grounded through a current-measuring electrometer. The oil charge density is increased by closing valve 1 and opening valves 2 and 3 to direct the flow through the electrostatically charging paper filter. The charge densities at the inlet and outlet of the paper filter are measured by ACSs 3 and 4. Measured electrometer currents at the paper filter and reservoir should equal the difference in convection currents in and out.

through the electrometer to ground, $i_{acs} = i_E$.

4.2.4 Couette Charger A Couette charger (CC), shown in Fig. 15, simulates flow electrification processes in transformers, where transformer oil fills the annulus between coaxial cylindrical electrodes that can be bare metal or transformer pressboard covered, as described by Morin *et al.* (1991b). The inner cylinder can rotate at speeds giving controlled turbulent flow that brings electric charge to the volume from the electrical double layer at the liquid/solid interfaces. Figure 16 shows the approximate linear charge-density profiles in each hydrodynamic sublayer of thickness δ_1 and δ_2, joining the wall charge-density values ρ_{w1} and ρ_{w2}, to the turbulent core charge density ρ_0. This compact apparatus allows for flexibility in testing various oils and transformer pressboards at controlled temperatures and moisture levels in the oil and pressboard. Flow electrification measurements as a func-

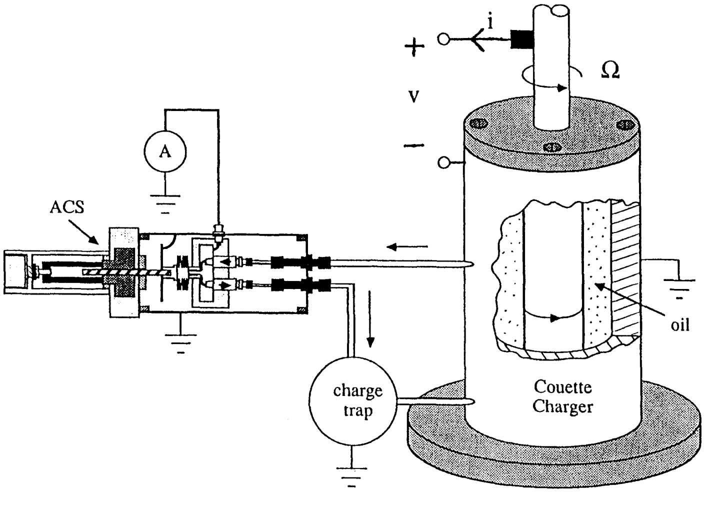
+
i
v
−
Ω
A
ACS
oil
charge trap
Couette Charger

FIG. 15. Couette charger with electrical terminals connecting inner and outer cylinders for measuring open-circuit terminal voltage or short-circuit current or for applying a voltage between the cylinders. The absolute charge sensor is used to measure the charge density in the turbulent core. The charge trap is a large volume where the fluid residence time is much longer than the fluid dielectric relaxation time. It is used so that fluid exiting the ACS into the Couette charger is uncharged.

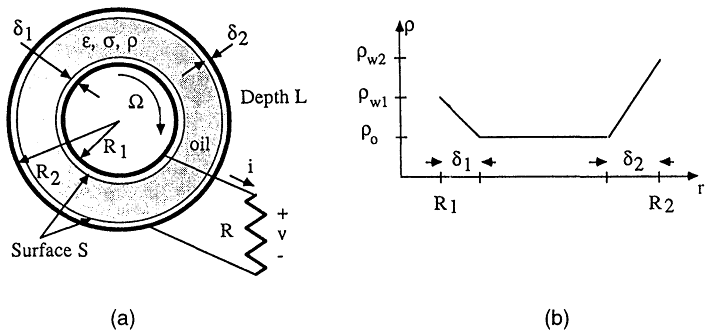

(a) (b)

FIG. 16. (a) Cross section of the Couette charger with load resistance R connected to the terminals. (b) Charge-density distribution between the inner and outer cylinders showing the wall charge densities ρ_{w1} and ρ_{w2} that describe the governing electrokinetics at the oil/cylinder interfaces. Also shown is the approximately linear charge-density profile in each sublayer joining the wall charge-density values to the turbulent core charge density ρ_0 at the edge of each sublayer with thicknesses δ_1 and δ_2.

tion of inner cylinder speed are also performed from open-circuit voltage and short-circuit current terminal measurements, as shown in Fig. 17. Transient measurements with a step change in temperature have shown the charge density to change from an initial value to a new steady-state value, including cases of polarity reversal, both values dependent on the equilibrium moisture levels in oil and pressboard, as shown in Fig. 18.

4.3 Kerr Electro-optic Field-Mapping Measurements

4.3.1 Governing Equations High-voltage–stressed materials are usually birefringent, in which case the refractive indices for light (of free-space wavelength λ) polarized parallel, $n_{\parallel}$, and perpendicular, $n_{\perp}$, to the local electric field are related by

$$n_{\parallel} - n_{\perp} = \lambda B E^2, \tag{75}$$

where B is the Kerr constant and E is the magnitude of the applied electric field. The phase shift ϕ between light-field components propagating in the direction perpendicular to the plane of the applied electric field along an electrode length L is

$$\phi = 2\pi B E^2 L = \pi(E/E_m)^2, \quad E_m = (2BL)^{-1/2}. \tag{76}$$

If the birefringent dielectric is placed within a circular polariscope with aligned or crossed polarizers, the transmitted light intensity is

$$\frac{I}{I_0} = \begin{cases} \sin^2 \frac{1}{2}\pi(E/E_m)^2, & \text{crossed polarisers,} \\ \cos^2 \frac{1}{2}\pi(E/E_m)^2, & \text{aligned polarisers.} \end{cases} \tag{77}$$

Field minima and maxima occur when

$$\frac{E}{E_m} = \sqrt{n} \begin{cases} n \text{ odd} \begin{pmatrix} \text{minima} & \text{AP} \\ \text{maxima} & \text{CP} \end{pmatrix}, \\ n \text{ even} \begin{pmatrix} \text{minima} & \text{CP} \\ \text{maxima} & \text{AP} \end{pmatrix}. \end{cases} \tag{78}$$

E_m is thus the field magnitude for first maximum with crossed polarizers and first minimum with aligned polarizers. For test cells of length $L = 1.1$ m with highly purified water at $\lambda = 590$ nm, $B \approx 2.8 \times 10^{-14}$ m/V^2 and $E_m \approx$ 35–36 kV/cm. As an example of Kerr measurements of electric field distributions, Fig. 19 shows a circular polariscope experimental apparatus with representative measurement using coaxial cylindrical electrodes.

When a high-voltage pulse is applied, as in Fig. 20, a photomultiplier-tube optical detector goes through the series of maxima and minima related by $\sqrt{n}$ as in (78), as described in Zahn (1988).

4.3.2 Representative Measurements in Highly Purified Water Representative pulsed high-voltage measurements of highly purified water are shown in Fig. 21 with parallel-plate electrodes for different metallic electrodes showing differences in the magnitude and sign of the injected charge as presented by Zahn *et al.* (1986). Stainless-steel electrodes generally inject positive charge; aluminum injects negative charge, while brass can inject either positive or negative

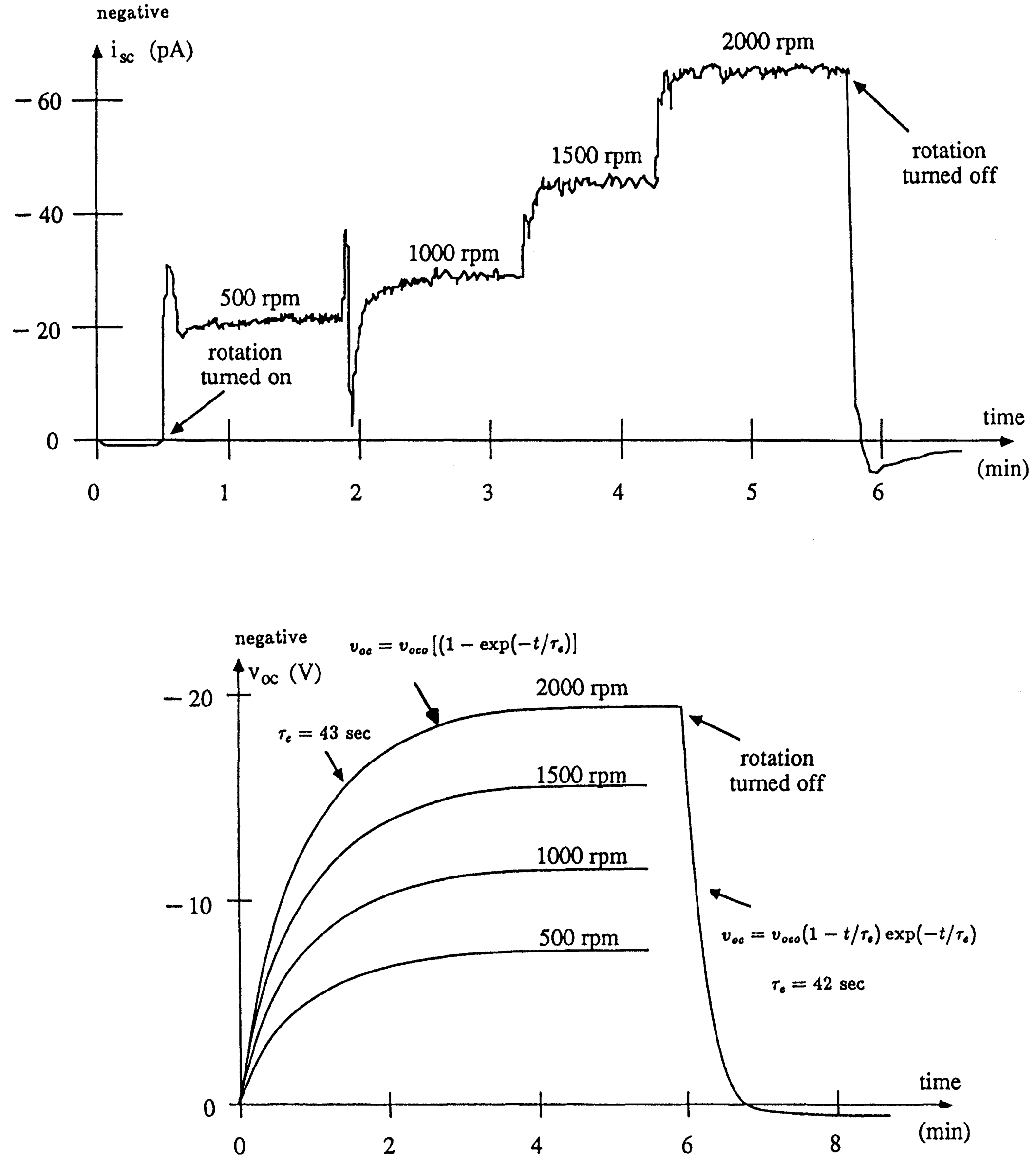

FIG. 17. Measured transients of short-circuit current and open-circuit voltage as functions of time for various inner-cylinder rotational speeds.

charge. Thus by appropriate choice of electrode-material combinations and voltage polarity, it is possible to have uncharged, unipolar charged negative or positive, or bipolar homocharge charged liquid.

Breakdown strengths are generally higher with bipolar injection. For example, with brass–aluminum electrodes, the polarity for bipolar injection had a breakdown strength of ~125–135 kV/cm, while the reverse polarity had negative charge injection with breakdown strength ~90–95 kV/cm. Similarly, stainless steel–aluminum electrodes had a breakdown strength with bipolar injection of ~125–140 kV/cm, while the reverse polarity had no charge injection with a breakdown strength of ~105 kV/cm. This increase in breakdown strength for homocharge distri-

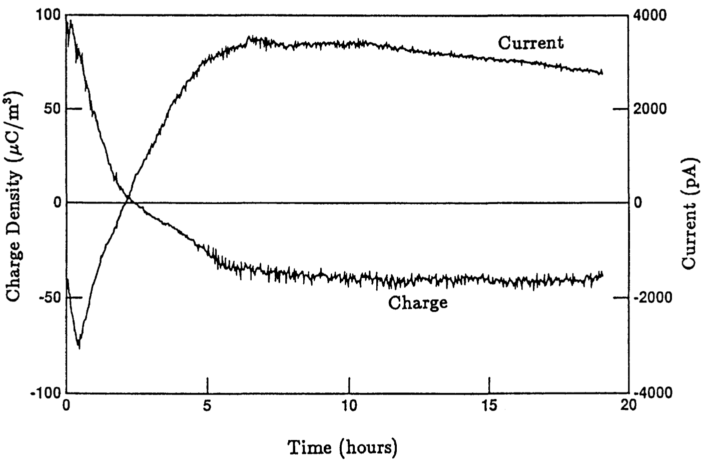

FIG. 18. Representative charge density and short-circuit terminal current responses to a step change in temperature from 40 to 70 °C with pressboard of 1.7% moisture covering inner- and outer-cylinder surfaces. The inner cylinder was rotating at 1000 rpm. Oil moisture started at 21 ppm and ended at a value of 35 ppm, which was approaching equilibrium at 70 °C.

butions is due to the decrease in electric field at both electrodes resulting from the space-charge shielding, as shown in Fig. 3(c). The electric field is increased in the center of the gap, but breakdown does not occur because the intrinsic strength of the dielectric in the volume is larger than at an interface.

4.3.3 Electron-Beam–Irradiated Polymethylmethacrylate Polymethylmethacry-

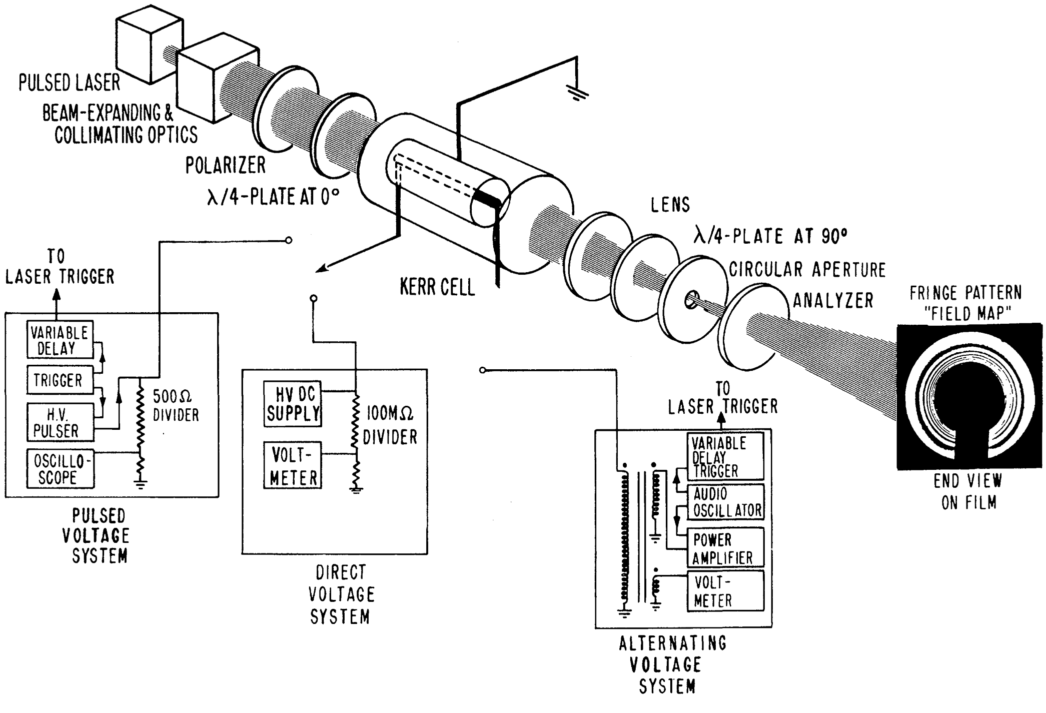
PULSED LASER
BEAM-EXPANDING & COLLIMATING OPTICS
POLARIZER
λ/4-PLATE AT 0°
KERR CELL
LENS
λ/4-PLATE AT 90°
CIRCULAR APERTURE
ANALYZER
FRINGE PATTERN "FIELD MAP"
END VIEW ON FILM
TO LASER TRIGGER
VARIABLE DELAY
TRIGGER
H.V. PULSER
OSCILLO-SCOPE
500Ω DIVIDER
PULSED VOLTAGE SYSTEM
HV DC SUPPLY
VOLT-METER
100MΩ DIVIDER
DIRECT VOLTAGE SYSTEM
TO LASER TRIGGER
VARIABLE DELAY TRIGGER
AUDIO OSCILLATOR
POWER AMPLIFIER
VOLT-METER
ALTERNATING VOLTAGE SYSTEM

FIG. 19. Circular polariscope experimental configuration shown for Kerr electro-optic field-mapping measurements using coaxial cylindrical electrodes for pulsed, dc, and ac high voltages. The photographic film shows circular light maxima and minima as the nonuniform electric field between cylindrical electrodes satisfies the conditions of Eq. (78) for multiple values of *n*.

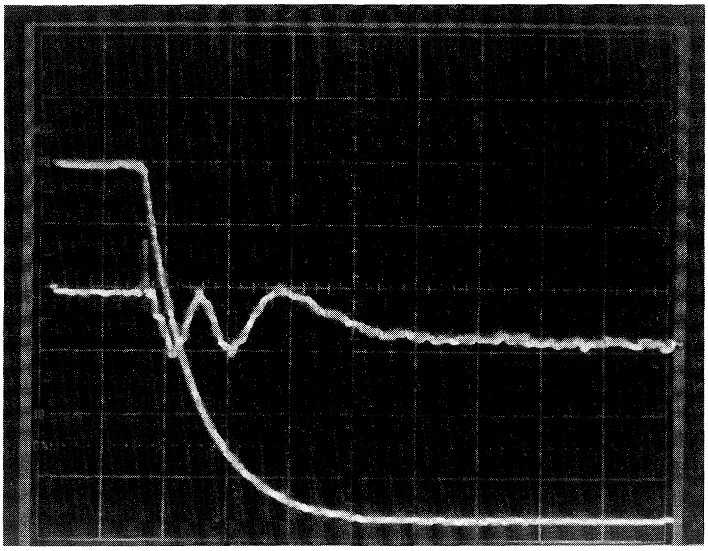

FIG. 20. The output of an optical photomultiplier tube goes through a series of maxima and minima [here $n = 1$–4 in Eq. (78)] when a 100-kV negative high-voltage pulse is applied, here shown for 77% water, 23% ethylene glycol by weight at 2.7 °C using crossed polarizers (time base 20 μs cm^{-1}). From this measurement, the Kerr constant of the mixture at $\lambda = 633$ nm wavelength is $B \approx 2.2 \times 10^{-14}$ m/V^2 with $E_m \approx 45.6$ kV/cm.

late (PMMA) is a convenient highly insulating polymer to work with because it is transparent at room temperature. When an electron beam is used to irradiate PMMA samples, the accumulated trapped charge leads to high internal self electric fields up to 3.5 MV/cm so that for a 10-cm-long PMMA sample, numerous light minima and maxima arise, making photographic and videotape measurements possible, as demonstrated by Zahn *et al.* (1987) and Hikita *et al.* (1988).

With $B \approx 2 \times 10^{-15}$ m/V^2 for PMMA, $E_m \approx 0.5$ MV/cm for $L = 10.2$ cm and $E_m \approx 0.44$ MV/cm for $L = 12.7$ cm, for which representative images are shown in the linear polariscope configuration in Fig. 22. Incident light is in the z direction, polarized at ±45° to the direction of the electric field E with an analyzing polarizer placed after the sample either crossed or aligned to the incident polarization. The He–Ne laser at 633 nm wavelength has its beam expanded to ≈7.5 cm to allow measurements of the light intensity distribution over the entire sample cross section. Beam splitters allow simultaneous measurements for aligned and crossed polarizers using Polaroid cameras as well as a videotape recording system.

The electron beam is generated by a Van de Graaff generator, exits from the accelerator tube through a thin (76 μm) aluminum window, and passes through ≈50 cm air to the PMMA sample that is short-circuited through current monitors at the top and bottom surfaces. The energy loss in the window and intervening air is about 160 keV. Figure 22 also shows representative data near the cameras for an accelerator beam energy of 2.6 MeV and a current density of 20 nA/cm^2 after ≈60 s of beam irradiation for aligned and crossed polarizers for a $d \approx 1.27$-cm-thick PMMA sample uniformly irradiated over its width, 5.1 cm, and length, 10.2 cm. The sample had $E_m \approx 0.5$ MV/cm. The sample is shown before irradiation to have some mechanical birefringence. Also shown in Fig. 22 is a representative frame of light intensity distribution from a computerized image digitizer for a larger current density of 110 nA/cm^2 after 11 s of irradiation with $E_m \approx 0.44$ MV/cm and $L = 12.7$ cm.

4.4 Acoustic Measurements of Space-Charge Distributions

In nontransparent materials, acoustic measurements offer an alternative to optical measurements. The two usual approaches for measuring space-charge distributions are a pressure-wave method and a pulsed electroacoustic method. Both methods result in a measured time signal that is a direct image of the spatial profiles of charge density or electric field. In addition to fundamental measurements of conduction properties, trapped charges in solids are useful for microphones, sensors, actuators, and optoelectronics.

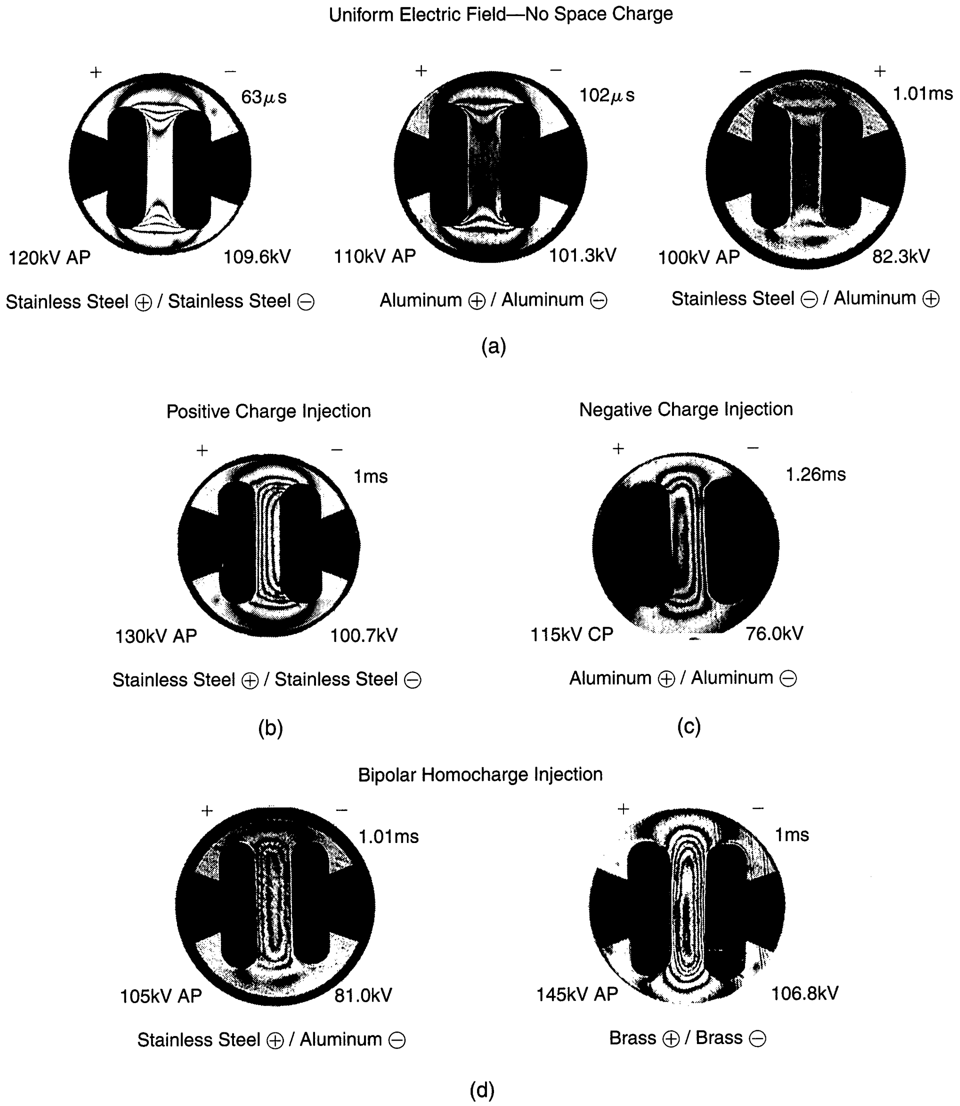
Uniform Electric Field—No Space Charge
+
−
63μs
120kV AP
109.6kV
Stainless Steel ⊕ / Stainless Steel ⊖
+
−
102μs
110kV AP
101.3kV
Aluminum ⊕ / Aluminum ⊖
−
+
1.01ms
100kV AP
82.3kV
Stainless Steel ⊖ / Aluminum ⊕
(a)
Positive Charge Injection
+
−
1ms
130kV AP
100.7kV
Stainless Steel ⊕ / Stainless Steel ⊖
(b)
Negative Charge Injection
+
−
1.26ms
115kV CP
76.0kV
Aluminum ⊕ / Aluminum ⊖
(c)
Bipolar Homocharge Injection
+
−
1.01ms
105kV AP
81.0kV
Stainless Steel ⊕ / Aluminum ⊖
+
−
1ms
145kV AP
106.8kV
Brass ⊕ / Brass ⊖
(d)

FIG. 21. Representative Kerr-effect measurements with a 1-cm gap between parallel-plate electrodes stressed by a high-voltage step in highly purified water with $E_m \approx$ 35–36 kV/cm. (a) The space-charge–free uniform electric-field distribution so that the Kerr effect shows uniform light transmission in the interelectrode gap region. At times less than about 100 μs after the step was applied, the dielectric is essentially space-charge free for all electrode metal combinations. With stainless steel negative and aluminum positive, the dielectric is space-charge free for all time. (b) Positive charge injection with stainless-steel electrodes so that the electric-field distribution has positive slope with decreased electric field near the positive electrode and increased field at the negative electrode. (c) Negative charge injection with aluminum electrodes so that the electric-field distribution has negative slope with decreased electric field near the negative electrode and increased field at the positive electrode. (d) Bipolar homocharge injection, where the electric field is decreased at both charge-injecting electrodes, with the peak field in the central gap region. In each photograph, the electrode metals and polarity are listed at the bottom; the initial charging voltage and whether aligned (AP) or crossed (CP) polarizers are used are given at the lower left; the instantaneous voltage is given at the lower right; and the time after high voltage is applied is given in the upper right.

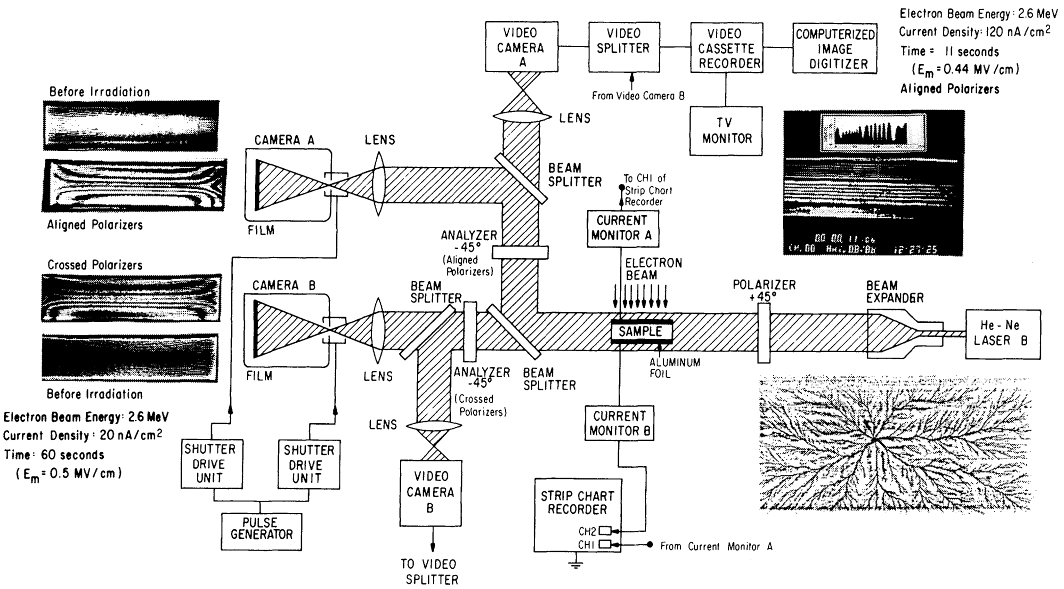
Before Irradiation
Aligned Polarizers
Crossed Polarizers
Before Irradiation
Electron Beam Energy: 2.6 MeV
Current Density: 20 nA/cm²
Time: 60 seconds
(Em = 0.5 MV/cm)
CAMERA A
LENS
FILM
CAMERA B
FILM
LENS
LENS
SHUTTER DRIVE UNIT
SHUTTER DRIVE UNIT
PULSE GENERATOR
VIDEO CAMERA A
VIDEO SPLITTER
VIDEO CASSETTE RECORDER
COMPUTERIZED IMAGE DIGITIZER
From Video Camera B
LENS
TV MONITOR
BEAM SPLITTER
To CH1 of Strip Chart Recorder
CURRENT MONITOR A
ANALYZER -45° (Aligned Polarizers)
BEAM SPLITTER
ELECTRON BEAM
SAMPLE
ALUMINUM FOIL
ANALYZER -45° (Crossed Polarizers)
BEAM SPLITTER
CURRENT MONITOR B
VIDEO CAMERA B
TO VIDEO SPLITTER
STRIP CHART RECORDER
CH2
CH1
From Current Monitor A
POLARIZER +45°
BEAM EXPANDER
He-Ne LASER B
Electron Beam Energy: 2.6 MeV
Current Density: 120 nA/cm²
Time = 11 seconds
(Em = 0.44 MV/cm)
Aligned Polarizers

FIG. 22. Apparatus and representative data for Kerr electro-optic field-mapping measurements with simultaneous aligned and crossed polarizers in electron-beam–irradiated samples using photographic film and a computer-interfaced videotape recording system as detectors. In the lower right is a representative electric tree in PMMA after electric breakdown.

4.4.1 Pressure-Wave Method A pressure wave is produced at one end of a sample using a piezoelectric transducer or by the heating from an intense laser pulse, as developed by Lewiner (1986) and Maeno *et al.* (1988). As the pulse propagates through the sample at the acoustic wave speed c, space charge will be displaced with the material causing a displacement current in the sample. The resulting open-circuit voltage is proportional to the electric field, $v(t) = \alpha E(x = ct)$, while the short-circuit current density is proportional to the volume charge density, $j(t) = \alpha c \rho(x = ct)/s$ where s is the sample thickness and α is the proportionality constant related to sample compressibility, pulse pressure, and pulse duration. The shape of measured voltage or current in time is identical to the spatial distribution of electric field E or charge density ρ. Measurements have typically been applied to polymers and ceramics but have also been applied to water trees in polymers and oil-impregnated pressboard. With a typical acoustic speed of $c = 2000$ m/s, a 1-ns pressure pulse has a spatial resolution of 2 μm.

4.4.2 Electroacoustic Method A high voltage is applied, and the Coulomb force on the space charge produces a propagating pressure wave that is detected by a piezoelectric transducer. The transducer signal is proportional to the convolution of the charge density and applied electric field, so that the Fourier transform of the transducer signal in time is proportional to the product of spatial Fourier transforms of charge density and electric field. Recent work has used this method to measure space-charge distributions in XLPE power cables.

4.5 Thermally Stimulated Measurements

A solid sample that is metallized on both surfaces is poled with a high electric field at high temperature, allowing permanent dipoles to orient with the electric field and charge to build up at interfaces and at structural defects. With the electric field maintained, the sample is cooled to trap the charge and dipoles. Then with the sample short-circuited, the sample is reheated at a constant rate of the order of 1 °C/min, and the discharge current is measured with an electrometer, of order 10^{-12}–10^{-10} A. A number of peaks in current as a function of temperature occur whose shape and location are different for each material and characteristic of permanent dipole density, relaxation time, activation energy, and other trapping parameters.

GLOSSARY

Arrhenius Aging Model: An exponential thermal aging model that assumes a single chemical reaction rate dependent on an activation energy and absolute temperature.

Avalanche: A cumulative multiplication of charge carriers due to electric-field–induced impact ionizations.

Corona: Luminous partial discharges in gases.

Debye Length: Approximate thickness of space-charge layer near interfaces for drift/diffusion conduction.

Einstein Relation: Statement that the ratio of molecular diffusivity to charge mobility equals the thermal voltage, $D/\mu = kT/q$.

Electroacoustic Measurements: Electric-field and space-charge measurements by propagating acoustic waves through the dielectric.

Flow Electrification: Electrostatic hazard that arises from entrainment of electrical double-layer charge at interfaces that accumulates on isolated surfaces to cause spark discharges.

Fowler–Nordheim Field Emission: Quantum mechanical tunneling of electrons from a metal into vacuum or a dielectric under a strong electric field.

Kerr Electro-optic Effect: An electric-field–induced birefringence that can be used to optically measure electric-field distributions.

Paschen's Law: In a uniform electric field, the breakdown voltage in a gas is a function only of the pressure-gap product. The minimum voltage for breakdown is called the Paschen minimum.

Poisson–Boltzmann Equation: Poisson's equation for the scalar electric potential under equilibrium drift/diffusion conduction with the charge carriers obeying Boltzmann distributions.

Schottky Emission: The increase in thermionic emission of electrons from a metal due to the electric field.

Streamers: Growth of avalanches into irregular filamentary luminous discharges in gas and liquid dielectrics.

Thermally Stimulated Measurements: Flow of charge as poled dielectric is reheated at constant rate.

Townsend Ionization: Ionizing collisions that can lead to electrical breakdown if the increase in electrons is unbounded.

Treeing: Partial discharge dendritic patterns of hollow or water channels in solids that can lead to electrical breakdown.

Weibull Distribution: A statistical distribution to typically model life, failure, and dielectric strength of dielectrics.

Works Cited

Clark, F. M. (1962), *Insulating Materials for Design and Engineering Practice,* New York: Wiley.

Hikita, M., Zahn, M., Wright, K. A., Cooke, C. M., Brennan, J. (1988), "Kerr Electro-optic Field Mapping Measurements in Electron Beam Irradiated Polymethylmethacrylate," *IEEE Trans. Electrical Insulation,* **EI-23,** 861–880.

Khalifa, M. (1990), *High-Voltage Engineering, Theory and Practice,* New York: Marcel Dekker.

Klinkenberg, A., Van der Minne, J. L. (1958), *Electrostatics in the Petroleum Industry,* Amsterdam: Elsevier.

Lewiner, J. (1986), "Evolution of Experimental Techniques for the Study of the Electrical Properties of Insulating Materials," *IEEE Trans. Electrical Insulation,* **EI-21,** 351–360.

Maeno, T., Futami, T., Kushibe, H., Takada, T., Cooke, C. M. (1988), "Measurement of Spatial Charge Distributions in Thick Dielectrics Using the Pulsed Electroacoustic Method," *IEEE Trans. Electrical Insulation,* **EI-23,** 433–439.

Meek, J. M., Craggs, J. D. (1978), *Electrical Breakdown of Gases,* New York: Wiley.

Morin II, A. J., Zahn, M., Melcher, J. R., Otten, D. M. (1991a), "An Absolute Charge Sensor for Fluid Electrification Measurements," *IEEE Trans. Electrical Insulation,* **26,** 181–199.

Morin II, A. J., Zahn, M., Melcher, J. R. (1991b), "Fluid Electrification Measurements of Transformer Pressboard/Oil Insulation in a Couette Charger," *IEEE Trans. Electrical Insulation,* **26,** 870–901.

Rogers, R. R. (1978), "IEEE and IEC Codes to Interpret Incipient Faults in Transformers, Using Gas in Oil Analysis," *IEEE Trans. Electrical Insulation,* **EI-13,** 349–353.

Sheiretov, Y., Zahn, M. (1995), "Dielectrometry Measurements of Moisture Dynamics and Oil-Impregnated Pressboard," *IEEE Trans. Dielectrics Electrical Insulation,* **2,** 329–351.

von Hippel, A. R. (1959), *Molecular Science and Molecular Engineering,* New York and London: Technology Press of MIT and Wiley.

Zahn, M., Ohki, Y., Fenneman, D. B., Grip-shover, R. J., Gehman, Jr., V. J. (1986), "Dielectric Properties of Water and Water/Ethylene Glycol Mixtures for Use in Pulsed Power System Design," *Proc. IEEE,* **74,** 1182–1221.

Zahn , M., Hikita, M., Wright, K. A., Cooke, C. M., Brennan, J. (1987), "Kerr Electro-optic Field Mapping Measurements in Electron Beam Irradiated Polymethylmethacrylate," *IEEE Trans. Electrical Insulation,* **EI-22,** 181–185.

Zahn, M. (1988), "Space Charge Effects in Dielectric Liquids," in: E. E. Kuhnhardt, L. G. Christophorou, L. H. Luessen (Eds.), *The Liquid State and Its Electrical Properties,* New York and London: Plenum Publishing Corp., pp. 367–430.

Further Reading

Insulation Theory, Practice, and Measurement Methods

Annual Book of ASTM Standards; Vol. 10.01, *Electrical Insulation-Solids, Composites and Coatings;* Vol. 10.02, *Electrical Insulation-Wiring and Cable, Heating and Electrical Tests;* Vol. 10.03, *Electrical Insulating Liquids and Gases; Electrical Protective Equipment,* Philadelphia: ASTM.

Cygan, P., Laghari, J. R. (1990), "Models for Insulation Aging under Electrical and Thermal Multistress," *IEEE Trans. Electrical Insulation,* **25,** 923–934.

Gallagher, T. J., Pearmain, A. J. (1983), *High Voltage Measurement, Testing and Design,* New York: Wiley.

Haus, H. A., Melcher, J. R. (1989), *Electromagnetic Fields and Energy,* Englewood Cliffs: Prentice Hall.

Zahn, M. (1979), *Electromagnetic Field Theory, A Problem Solving Approach,* New York: Wiley; also (1987), Malabar: Krieger.

Vacuum

Latham, R. V. (1995), *High Voltage Vacuum Insulation. Basic Concepts and Technological Practice,* London: Academic.

Miller, H. C. (1990), "Electrical Discharges in Vacuum (1877–1979)," *IEEE Trans. Electrical Insulation,* **25,** 765–860.

Miller, H. C. (1991), "Electrical Discharges in Vacuum, 1980–1990," *IEEE Trans. Electrical Insulation,* **26,** 949–1043.

Gases

Bartnikas, R., McMahon, E. J. (1979), *Engineering Dielectrics, Volume I, Corona Measurement and Interpretation,* Philadelphia: ASTM Special Technical Publication 669.

Dakin, T. W. (1962), "Theory of Gas Breakdown," in: *Progress in Dielectrics,* Vol. 4, London: Heywood & Company, pp. 152–198.

Sudarshan, T. S., Dougal, R. A. (1986), "Mechanisms of Surface Flashover along Solid Dielectrics in Compressed Gases: A Review," *IEEE Trans. Electrical Insulation,* **EI-21,** 727–746.

Van Brunt, R. J. (1994), "Physics and Chemistry of Partial Discharge and Corona. Recent Advances and Future Challenges," *IEEE Trans. Dielectrics and Electrical Insulation,* **1,** 761–784.

Liquids

Bartnikas, R. (1994), *Engineering Dielectrics, Volume III, Electrical Insulating Liquids,* Philadelphia: ASTM, Monograph 2.

Erdman, H. G. (1992), *Electrical Insulating Oils,* Philadelphia: ASTM, Special Technical Publication 998.

Gallagher, T. J. (1975), *Simple Dielectric Liquids,* Oxford: Clarendon Press.

Wilson, A. C. M. (1980), *Insulating Liquids: Their Uses, Manufacture and Properties,* Stevenage and New York: Peter Peregrinus, Ltd.

Zahn, M. (guest editor) (1988), "Flow Electrification in Electric Power Apparatus," *IEEE Trans. Electrical Insulation,* **23,** 101–176.

Solids

Bartnikas, R., Eichhorn, R. M. (1983), *Engineering Dielectrics, Volume IIA, Electrical Properties of Solid Insulating Materials: Molecular Structure and Electrical Behavior,* Philadelphia: ASTM, Special Technical Publication 783.

Bartnikas, R. (1987), *Engineering Dielectrics, Volume IIB, Electrical Properties of Solid Insulating Materials: Measurement Techniques,* Philadelphia: ASTM, Special Technical Publication 926.

Blythe, A. R. (1979), *Electrical Properties of Polymers,* Cambridge, England: Cambridge Univ. Press.

Devins, J. C. (1984), "The Physics of Partial Discharge in Solid Dielectrics," *IEEE Trans. Electrical Insulation,* **EI-19,** 475–495.

Gerhard-Multhaupt, R. (1987), "Electrets: Dielectrics with Quasi-Permanent Charge or Polarization," *IEEE Trans. Electrical Insulation,* **EI-22,** 531–554.

Jonscher, A. K. (1983), *Dielectric Relaxation in Solids,* London: Chelsea Dielectrics Press.

Mason, J. H. (1959), "Dielectric Breakdown in Solid Insulation," in: *Progress in Dielectrics,* Vol. 1, New York: Wiley, pp. 3–58.

O'Dwyer, J. J. (1973), *The Theory of Electrical Conduction and Breakdown in Solid Dielectrics,* Oxford: Clarendon Press.

Sessler, G. M. (Ed.) (1980), *Electrets,* Berlin: Springer-Verlag.

Shaw, M. T., Shaw, S. H. (1984), "Water Treeing in Solid Dielectrics," *IEEE Trans. Electrical Insulation,* **EI-19,** 419–452.

Steennis, E. F., Kreuger, F. H. (1990), "Water Treeing in Polyethylene Cables," *IEEE Trans. Electrical Insulation,* **EI-25,** 989–1028.

van Turnhout, J. (1975), *Thermally Stimulated Discharge of Polymer Electrets,* Amsterdam: Elsevier.

SOLIDS, ACOUSTICAL PROPERTIES OF

A. V. GRANATO, *Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois, U.S.A.*

Introduction 467
1. **Perfect Crystals—Elastic Constants** 468
1.1 Second-Order Elastic Constants of Perfect Crystals 468
1.2 Third-Order Elastic Constants of Perfect Crystals 469
1.3 Interatomic Potentials 469
1.4 Thermal Properties of Crystals 469
2. **Real Solids—Acoustic Response; Electrons, Dislocations, and Point Defects as Sources of Loss** 469
2.1 Electrons 470
2.2 Dislocations 471
2.2.1 Dislocation Relaxation Effects 471
2.2.2 Dislocation Resonance Effects 471
2.2.2.1 Low-Amplitude Dislocation Resonance 473
2.2.2.2 Amplitude-Dependent Damping Theory 475
2.3 Point Defects 478
2.3.1 Tunneling of Trapped Hydrogen in Niobium 479
2.3.2 Theory of Two-Level Systems. Relaxation and Resonance Effects 481
2.3.3 Acoustic Measurements of Tunneling Hydrogen in Niobium 483
2.3.4 Classical Point-Defect Resonance Effect 484
Acknowledgment 485
Glossary 485
Works Cited 486
Further Reading 487

INTRODUCTION

This is a short survey of a very large topic, using selected examples illustrating some of the principal effects found by acoustical methods over the past 40 years. It starts with perfect crystals, or at least acoustically perfect crystals. These are defined as crystals for which the defects are either static, such as dislocations uniformly pinned by impurities, or in low enough concentrations so as not to make their presence known. In perfect crystals, acoustic waves travel without attenuation or dispersion. Measurements of the sound velocity v are most often used to determine elastic constants c through relations of the form $\rho v^2 = c$, where ρ is the density. The literature on elastic constants is old and large, and many excellent reviews and compilations are available (Huntington, 1958; Simmons and Wang, 1971; Fedorov, 1968; Hearmon, 1979). The elastic constants are thermodynamic properties that are functions of applied stress and temperature. Measurements of the change of elastic constants with applied stress have been made in the past few decades to obtain third-order elastic constants, giving the first deviations from linearity in the interatomic force–displacement relations.

Real crystals contain imperfections, which give rise to attenuation and velocity changes. The acoustic response of solids is of two types, relaxational and resonance. These are analogous to the paramagnetic and diamagnetic response of magnetic dipoles to external magnetic fields. Paraelasticity and relaxation are the same. The defect strain results from the competition between the ordering effects produced by an external field on an existing elastic dipole and the disordering effects due to thermal fluctuations. It requires thermal activation over or tunneling through an energy barrier and is normally therefore strongly temperature dependent. Diaelasticity

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

and resonance are also the same. They are the direct response corresponding to an induced elastic dipole under an applied force even where there is no motion over or through a barrier.

Defects break up into two major classes, dislocations and point defects. Little will be said about the Bordoni relaxation peaks arising from dislocation–lattice interactions, but the resonance described by the string model that has been used to explain the dislocation behavior that is seen in a wide variety of situations will be discussed, including the amplitude dependence with its inertial and viscosity effects, as well as the Brownian-motion theory that has been developed for dislocation theory that finds application in many other areas of physics.

Point defects are another major category of defects. Acoustic measurements have been used to determine the configuration and dynamics of self-interstitials in metals and to resolve a long-lasting discussion about what the structure of such defects is. A particularly simple and illustrative example is hydrogen in niobium, of special interest because of the quantum tunneling behavior of the hydrogen. Electrons are another important source of acoustic attenuation. This attenuation proved to be one of the major sources of information for testing and establishing the Bardeen–Cooper–Schreiffer (BCS) (Bardeen *et al.*, 1957) mechanism of superconductivity.

In solid studies, the types of techniques that have been used vary a great deal because the techniques depend on the frequency range, although the basic concepts and ideas are pretty much the same at all frequencies. Measurements have been made at frequencies as low as 10^{-2} Hz and occasionally even lower using stress–strain curves. In the 1-Hz range pendula are used, and in the 100-Hz range bending-beam techniques are often employed. In the kilohertz region, usually a resonant-bar technique is used. The pulse-echo technique is mostly used for the megahertz region from about 10 to 1000 MHz. The available range continues on up into the gigahertz range with phonon spectroscopy techniques available there, and more recently into the terahertz region with again basically pulse techniques, but now the sound is stimulated by laser excitation and usually detected by laser techniques as well. Many of the various measurement techniques available are described by Krautkrämer and Krautkrämer (1983).

Most of these techniques are characterized by a very strong sensitivity, which proves to be a decisive advantage over nonacoustic techniques for many applications. Typically, sensitivity to defects in the parts-per-million range is available. For the pulse-echo technique, one can measure the attenuation of the amplitude and the velocity of the signal. By using interferometer techniques, one can measure velocity changes, for example, at 10 MHz up to about 1 Hz or about one part in 10^7. This is especially important if one wants to know the properties of isolated defects, because typically defects interact strongly through their shear stress fields, and at concentrations of even a thousand parts per million, the interactions become so strong that it is not possible really to study the properties of isolated defects. These interactions severely limit direct techniques such as x-ray, neutron, and phonon spectroscopies for the study of isolated defects.

For the pulse-echo technique, a second property has proved to be particularly important and useful. This is the fact that one can quite easily vary the polarization, for example, by looking at longitudinal and shear waves in different directions. This leads directly to complete symmetry information, useful in determining configurations of defects.

This large field has many international conferences, reports, reviews, and monographs (Nowick and Berry, 1972; De Batist, 1972; Hiki, 1972; Truell *et al.*, 1969; Mason, 1958, 1964; Zener, 1948; Mason, 1964–1968; Mason and Thurston, 1970–1984) available for detailed discussions. An extensive reading list with references to many conference reports has been compiled by Magalas (1996).

1. PERFECT CRYSTALS—ELASTIC CONSTANTS

1.1 Second-Order Elastic Constants of Perfect Crystals

Elastic constants have provided a means for testing various cohesion theories. One approach has been a phenomenological one in which the free energy is expanded in a Taylor series in the strains in a crystal. The first nonzero terms are quadratic in the strains and represented by a fourth-order elastic-constant tensor. The next terms in such a series would be given by a sixth-rank tensor, called the third-order elastic constants (Thurston and Brugger, 1964). The elastic constants, defined as the second derivatives of the free energy against the strains, give the second-order elastic constants, but in general they depend on the strain. These tensors can be written in a reduced notation because of the symmetries involved, and the three familiar elastic constants for the cubic case, C_{11}, C_{12}, and C_{44}, are more usually expressed in a different combination: the Zener (1948) combination of bulk modulus $B = (C_{11} + 2C_{12})/3$ and the two shear constants, C_{44} and $C' = (C_{11} - C_{12})/2$.

1.2 Third-Order Elastic Constants of Perfect Crystals

There are six third-order elastic constants for cubic symmetry, obtainable by measuring the change of the second-order elastic constants with applied stresses. With the sensitivities now available, it is not necessary to apply the kind of stresses (kilobars of pressure) that occurred in most of the older literature. The fact that one can work with extremely low loads means that plastic-deformation–induced defects can be avoided, and this permits measurements under uniaxial strain conditions so that one can get many different kinds of measurements and determine the full set of third-order elastic constants. We mention two important applications for these constants.

1.3 Interatomic Potentials

Third-order elastic constants provide information about interatomic potentials for crystals, as well as useful data for calculations of the properties of defects. Generally speaking, one knows that, at least in the simpler crystals, there is normally a short-range, fast-changing part to the potential and a long-range slow part. For example, in the alkali halides the Coulomb attraction and the closed-core repulsion are the dominant terms. The binding energy is usually determined mostly by the long-range part, typically about 90% from the Coulomb term and perhaps 10% from the short-range repulsive part. When one goes to the lattice constant, this is obtained by a derivative of the potential against distance, so that the short-range part becomes more important, giving more of a 50-50 kind of composition. The second-order elastic constants require another derivative, giving more weight to short-range terms again. The relative values of the third-order elastic constants depend almost completely on the structure and no longer in detail on the potential, and so one can predict relative values fairly well knowing only the geometrical structure (Holder and Granato, 1971).

1.4 Thermal Properties of Crystals

The third-order constants can also be used for descriptions of thermal properties. This was an idea first introduced by Einstein to connect the mechanical and thermal properties of crystals. Using a harmonic oscillator of frequency ω as a model of the solid, he was able to calculate the Einstein temperature ϑ_E (with $\hbar\omega = k\vartheta_E$) by expressing the specific heat in terms of ϑ_E and ϑ_E in terms of the elastic constants. This idea was developed further by Debye, who made a more correct calculation and expressed the specific heat in terms of the normal modes of the crystal, finding the Debye theta Θ_D in terms of the known elastic constants. This works exactly at low temperatures, where in the long-wavelength limit, one can justify use of the continuum elastic theory. This process can be continued. With third-order elastic constants, one can obtain values for the thermal expansion. The same applies for quantities like the temperature dependence of the specific heat and the elastic constants, which, however, require fourth-order elastic constants as well.

2. REAL SOLIDS—ACOUSTIC RESPONSE; ELECTRONS, DISLOCATIONS, AND POINT DEFECTS AS SOURCES OF LOSS

The principal sources of attenuation in metals come from dislocations, point defects, and electrons. Metals not containing defects may be regarded as perfect crystals, but we choose to include them here since the electrons are, like the defects, a source of loss. The losses can be identified by their temperature dependence and their dependence on the point-defect density ϕ, among other things, as illustrated in Fig. 1. The electron loss has a temperature dependence that depends on the electrical conductivity. It is relatively large and constant at low temperatures and falls off quickly at high temperatures. In a superconductor it drops suddenly at the transition temperature T_c [Fig. 1(a)]. The characteristic shape is given very well by the BCS theory. Also, the velocity has a temperature dependence that is T^4 at low temperatures according to theory similar to that for the Debye specific heat (Leibfried and Ludwig, 1961; Garber and Granato, 1975). In a metal there is a contribution to the velocity from the electrons that is T^2, which is normally overwhelmed by the lattice contribution, except at the very lowest temperatures.

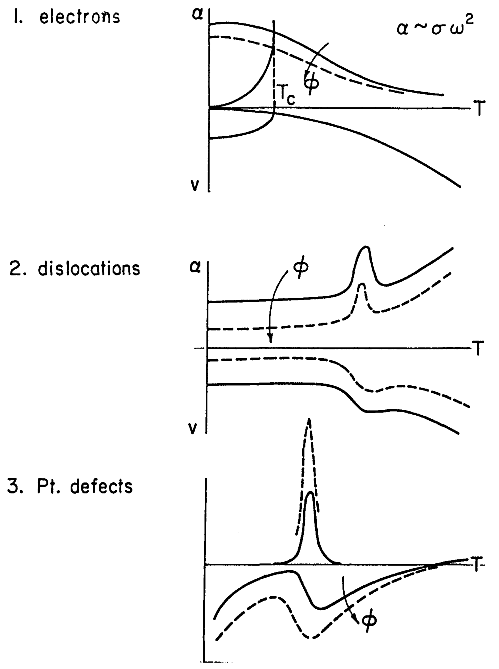

FIG. 1. Schematic temperature and point-defect-density (ϕ) dependences of the principal sources of attenuation. The curves above and below the horizontal line are attenuation and velocity, respectively, as functions of temperature. With increasing point-defect concentration ϕ, the solid curves change to the dashed curves.

In a superconductor there is also a small change in the velocity at the transition temperature. If one irradiates or puts in point defects that change the conductivity, the electronic loss decreases [dashed line, Fig. 1 (a)].

Dislocations provide a second very efficient source of loss, which is felt over the whole available temperature range. For pure and lightly deformed crystals, there is a resonance loss (Granato and Lučke, 1956) that varies approximately linearly with temperature. Point defects immobilize the dislocations and reduce the loss [dashed line, Fig. 1(b)]. For heavily deformed crystals, a relaxation effect known as the Bordoni peak (Bordoni, 1949) occurs, which is less sensitive than the resonance effect to point defects.

Point defects also have a characteristic relaxation loss called the Snoek effect (Snoek, 1941) that has been known for a long time. This gives rise to a loss peak when the frequency matches the rate of motion of the point defect between positions equivalent by symmetry. At the same time, a dispersion in the velocity occurs. More recently appreciated is the fact that there is also a resonance effect for point defects in general (Jäckle *et al.*, 1976; Hunklinger and Arnold, 1976), which gives a negative contribution to the modulus that increases at low temperatures and has a temperature-independent component as well (Granato, 1994a). These are the main mechanisms, and their characteristics are easily identified so that they can be used to sort out multiple sources of loss.

2.1 Electrons

A simple model for the effect of electrons was given by Mason (1955), who supposed that the viscosity of the electron gas produced attenuation in the same way as in the case of acoustic propagation in an ordinary gas or liquid. Mason then applied the standard expressions for sound attenuation due to viscosity, which give, for the amplitude attenuation of transverse waves,

$$\alpha_t = 2\pi^2 f^2 \eta / \rho \vartheta_t^3, \tag{1}$$

where f is the frequency, η is the viscosity, ρ is the density, and ϑ_t is the transverse-wave velocity. For longitudinal waves, a similar expression is obtained with $\alpha_l = (\frac{4}{3})(\vartheta_t/\vartheta_l)^3\alpha_t$, where ϑ_l is the longitudinal-wave velocity. The viscosity η can be expressed in terms of the conductivity from kinetic theory as $\eta = (3\pi^2N)^{2/3}(\hbar/5e^2)\sigma$, where N is the number of particles per unit volume, $\hbar$ is Planck's constant, e is the electronic charge, and σ is the conductivity. An equivalent result based on a relaxation mechanism was given by Morse (1955), and both are subject to the restrictions $ql_e < 1$, where q is the propagation constant and l_e is the electronic mean free path. Pippard (1955) extended these treatments to arbitrary l_e. The present understanding of ultrasonic attenuation in normal metals has been reviewed by Rayne (1975). An application of the Bardeen–Cooper–Schrieffer (Bardeen *et al.*, 1957) theory of superconductivity by Morse and Bohm (1957) provided a successful explanation for the temperature dependence of the ultrasonic attenuation in superconductors.

2.2 Dislocations

2.2.1 Dislocation Relaxation Effects

The explanation of the Bordoni peak came about mainly through the work of Seeger (1956) and his colleagues at Stuttgart, in terms of kinks on dislocations. That is, dislocations that are nearly aligned along crystallographic directions feel an interaction with the lattice, which provides a periodic potential. For the dislocation to advance, it is necessary to stimulate thermally the creation of a nucleus, a place where the dislocation crosses over into the next valley. That is done by creating a kink pair. There are many models for these kinks, starting with the Frenkel–Kontorova (1939) model. They have a very interesting mathematical structure with particle-type properties. They have a mass, can be accelerated, and have a limiting relativistic velocity given by the sound velocity. The particles can oscillate, radiate, and interact with themselves, with phonons and electrons, and so on. Nowadays the literature is pretty much similar to the older literature, but the kinks are usually referred to as solitons.

The use of the kink-pair model for the explanation of the Bordoni peak has been reviewed recently by Ritchie and Fantozzi (1992). Some questions about the interpretation are still under discussion (Kosugi and Kino, 1993). The relationship of the kink model to the vibrating-string model and to dislocation resonance effects is also reviewed by Ritchie and Fantozzi, and earlier as well by Granato and Lücke (1966).

2.2.2 Dislocation Resonance Effects

For dislocations that lie at a finite angle to the crystallographic directions, the kink density becomes so high that the kinks become irrelevant. That is, although the motion still occurs by kink motion sideways, the displacement actually is very much like that of a simple string. There is an effective tension because a dislocation under stress increases its length or increases its energy. That provides a very strong restoring force. Also phonons and electrons are scattered by a moving dislocation, providing a resistance or viscosity.

One of the characteristic features of the resonant dislocation component to the loss is that it is very amplitude dependent. For example, measurements by Read (1940), shown in Fig. 2, may already be the first strong experimental evidence for the existence of the dislocations in crystals that had been postulated just a few years earlier. The convincing part of the proof is the orientation dependence of the measurements made in single crystals of zinc where the slip planes were known. It takes a shear stress to move a dis-

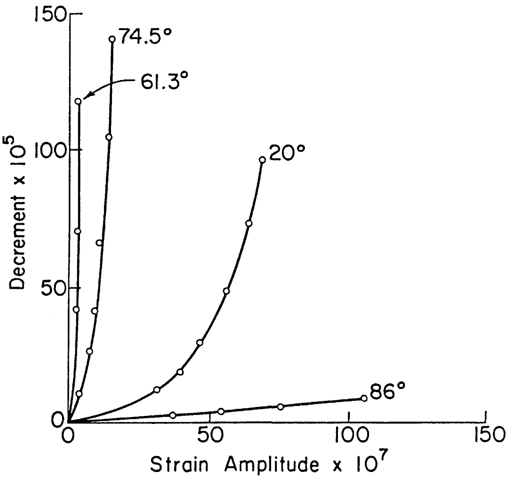

FIG. 2. Variation of decrement with strain amplitude in zinc crystals whose cylinder axes make various angles θ with the hexagonal c axis (after Read, 1940).

location, so that when the slip plane is at 0° or 90° to the tensile axis, there should be no shear stress and no effect. In fact, for crystals where the slip planes were near 0° or 90°, there is very little effect. The maximum effect should and does occur near 45°, providing very strong evidence for the existence of dislocations.

One rather convincing demonstration of the radiation (Eshelby, 1949; Garber and Granato, 1970) that occurs from moving dislocations comes from a simple experiment (Schwenker and Granato, 1969) in which one side of a sample of lithium fluoride is permanently displaced relative to the other so that there is a wall of dislocations on one slip plane. Now, if the dislocations cannot move, then one should simply get the usual echo from the back face of the sample for an incident sound wave. But if the shear stress of the sound wave is polarized so that it can move the dislocations, an additional echo is generated by the wall. Each dislocation radiates a cylindrical wave. By Huyghens's principle, these add up to a macroscopic plane wave that can also be detected. One should expect an echo for the polarization that moves the dislocations and no echo for the polarization at right angles. That is just what happens, as shown in Fig. 3. Also if these

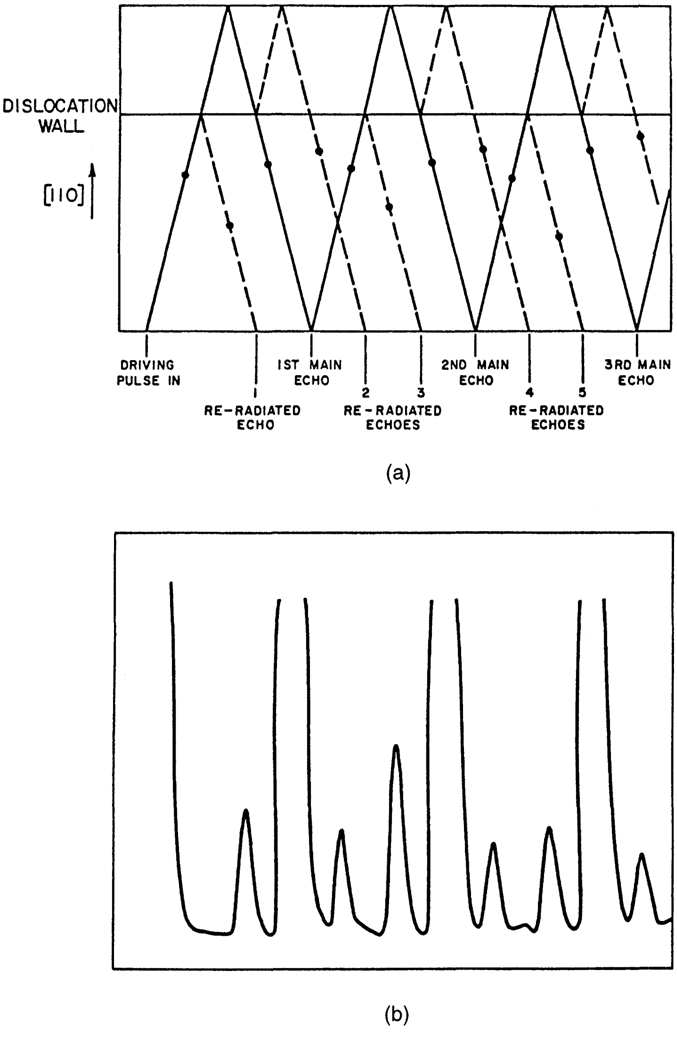

FIG. 3. (a) Schematic representation of wave positions along the crystal axis vs time for a deformed crystal. (b) The reradiated dislocation wall echoes for a deformed LiF crystal vs display time (after Schwenker and Granato, 1969).

crystals are lightly irradiated so as to pin down the dislocations, the wall echoes go away while the attenuation of the main echoes decreases.

2.2.2.1 Low-Amplitude Dislocation Resonance. For low stress amplitudes, the theory is analogous to dielectric theory, with dislocation displacement taking the place of charge displacement. The elastic shear-constant change $\Delta G/G$ is given by

$$\Delta G/G = \epsilon_d/\epsilon_{\mathrm{el}}, \tag{2}$$

where ϵ_d is the dislocation strain and ϵ_{el} is the elastic strain. The dislocation strain is given by the Orowan (1940) relation

$$\epsilon_d = \Lambda b\bar{y}, \tag{3}$$

where Λ is the dislocation density, b is the Burgers vector, and $\bar{y}$ is the average dislocation displacement, obtained from Koehler's (1952) vibrating-string model,

$$A\ddot{y} + B\dot{y} - Cy_{xx} = b\sigma, \tag{4}$$

where A is the dislocation mass per unit length, B is the viscous force per unit velocity, C is the dislocation line tension, and $b\sigma$ is the applied force per unit length. Using Eqs. (4) and (3) in (2), one obtains a damped harmonic oscillator response of magnitude

$$\frac{\Delta G}{G} = \frac{\Omega \Lambda L^2}{6} \frac{1}{(1 - \omega^2/\omega_0^2) + id\omega/\omega_0^2}, \tag{5}$$

where $d = B/A$ and $\omega_0^2 = \pi^2 C/AL^2$, where L is the length of a pinned dislocation segment. In Eq. (5), Ω is an orientation factor taking account of the fact that only the resolved shear-stress component in the slip systems produces dislocation motion.

The real part of Eq. (5) gives the modulus change $\Delta c/c$, and the imaginary part gives the decrement Δ of the vibration, defined in the Glossary. A distribution of dislocation loop lengths L needs to be taken into account for comparisons with experiment.

The first striking evidence for the string model from low-amplitude measurements was that of Thompson and Holmes (1956), who irradiated pure single crystals of copper with neutrons. It is reasonable to assume, with Thompson and Holmes, that the dislocation density does not change in this process, so that observed changes in the damping modulus should be caused only by changes in the average loop length. Thompson and Holmes found that the damping and modulus changes depended upon the fourth and second powers of the average loop length, respectively, as predicted by Eq. (5). In addition, they found that the size of the effect varied considerably from specimen to specimen, all of nominally the same purity. The specimens could be restored to their original condition by a suitable high-temperature annealing treatment. Also, it was found that there was a considerable background damping that was not removed by the radiation. The importance of these results for the further development of the field cannot be overemphasized. They demonstrated that the low-amplitude damping effects follow definite characteristic laws but that it is practically impossible to prepare two different specimens in exactly the same state. Radiation techniques provide a means for continuously varying the "purity" and studying the dependence on loop length in detail. Also the results showed that in order to attempt quantitative analysis of damping measurements, it is necessary to have some means of separating the effects from any unknown background effects. Once these conditions are realized, the method proves a means of extraordinary sensitivity in studying defect interactions.

When pinning occurs without changing the dislocation structure, Eq. (5) can be written in terms of the number of pinners added to a dislocation segment in an irradiation experiment so that a "direct count" scheme for the number of pinners on a dislocation segment is available. Thus, if n_0 is the original number of pinners per unit length of dislocation line, L_0 is the original segment length, n is the number of defect pinners added, and L is the average segment length with the added pinners, then

$$L = \frac{L_0}{1 + n/n_0}. \tag{6}$$

If we define the parameters Y and Z as Thompson and Pare (1960) have done, as the fraction of the modulus defect and decrement remaining after n pinners have been added, it follows that

$$n/n_0 = Z^{-1/4} - 1 = Y^{-1/2} - 1. \qquad (7)$$

The ratio n/n_0 may be interpreted directly as the number of pinners added per dislocation segment since, by definition, the number n_0 is the same as the number of dislocation segments in a unit length of dislocation.

Since Y is proportional to L^2 and Z is proportional to L^4, a line of slope unity must result upon plotting Y^2 vs Z on a logarithmic plot if the loop-length expectation of Eq. (5) is realized. Alternatively, plots of $Y^{-1/2}$ and $Z^{-1/4}$ vs radiation time should coincide, as shown in Fig. 4 from the data of Thompson and Pare (1960). The experimental check of the loop-length dependence extends through three orders of magnitude. This is about the normal range of variation found for dislocation damping in high-purity copper crystals. It should be noted that the extent of experimental agreement with theory in this test depends to some extent upon the specimen itself. Deviations are sometimes found. However, these have been interpreted as being caused by the presence of two or more dislocation components that possess different line tensions, and apparently they do not represent a failure of the string model.

The frequency dependence of the damping predicted in Eq. (5) is illustrated in Fig. 5. These are measurements by Stern and Granato (1962) showing the effect of cobalt gamma irradiation on the damping of high-purity copper. Before irradiation, the damping has a maximum at a few megahertz. After 50 h of irradiation in a 6000-Ci cobalt source, the height of the maximum decreased, and the location is increased to about 100 MHz. The damping at low frequencies is much more sensitive to the increased number of pinning points than that at high frequencies, as expected from Eq. (5). More recently, the underdamped dislocation resonance has been observed by Beamish and Franck (1982), in bcc ^{3}He.

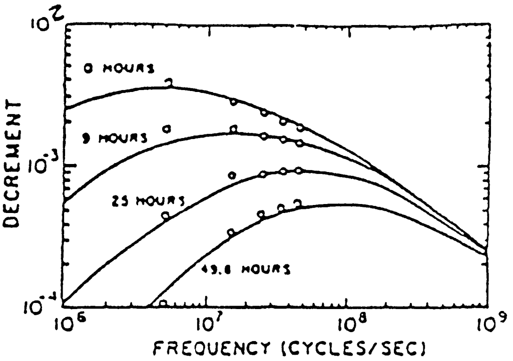

FIG. 5. The dislocation decrement in copper as a function of frequency for several times during cobalt gamma irradiation. The solid curves are theoretical (after Stern and Granato, 1962).

According to Eq. (5), the drag constant B may be measured if the dislocation density is known. Drag constants were determined by Fanti *et al.* (1969) from measurements of ultrasonic attenuation and velocity changes as functions of frequency in specimens of LiF and NaCl for which the dislocation density Λ was determined by etch-pit counts. The values obtained are in good agreement with those to be expected for phonon drag, according to Leibfried (1950). When these values are used to calculate the limiting dislo-

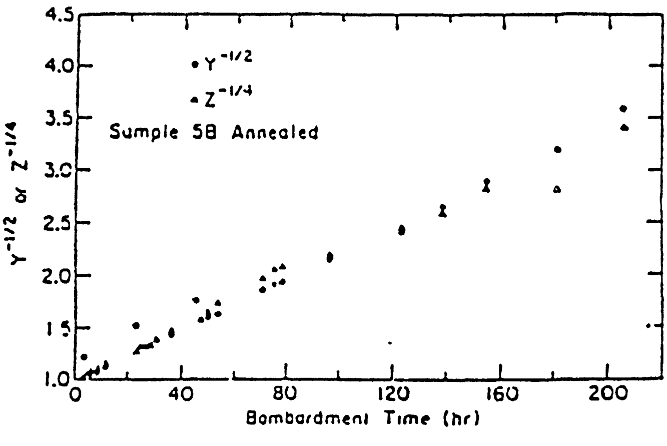

FIG. 4. $Y^{-1/2}$ and $Z^{-1/4}$ vs time at room temperature in copper sample 5*B* (after Thompson and Pare, 1960).

cation velocities expected at high stress at room temperature, according to

$$B_p v = b\tau, \tag{8}$$

the results in Fig. 6 are obtained for NaCl. This figure also shows the velocities at high stresses determined by the etch-pit technique by Gutmanas *et al.* (1963). It may be seen that the etch-pit velocities do not exceed, and appear to be limited by, the velocities predicted by extrapolation of the ultrasonic results from lower stress levels ($\sim 10^2$ cm/s). This is strong evidence that the same drag mechanism is operative on the instantaneous velocities in both stress ranges. As was pointed out by Baker (1962), the fact that velocities determined from ultrasonic measurements are many orders of magnitude higher than those measured by etch-pit techniques at the yield stress and the fact that dislocation displacements as large as 1000*b* are deduced at stress amplitude levels an order of magnitude below the yield stress demonstrate that the Peierls force is ineffective in limiting dislocation motion, at least at room temperature. This result is an example of ways by which acoustic measurements can help to distinguish between various postulated deformation mechanisms.

These results, with others, show that under controlled conditions the predictions of the string model for low strain amplitudes have been verified, and, where contact can be made with results of other measurements, the comparison is favorable. The predictions can therefore be used with some confidence in the study of dislocation–point-defect interactions. The controlled conditions include

1. well-defined experimental arrangements;
2. discrimination of dislocation losses from other losses, for example, as obtained by irradiation; and
3. self-checking arrangements.

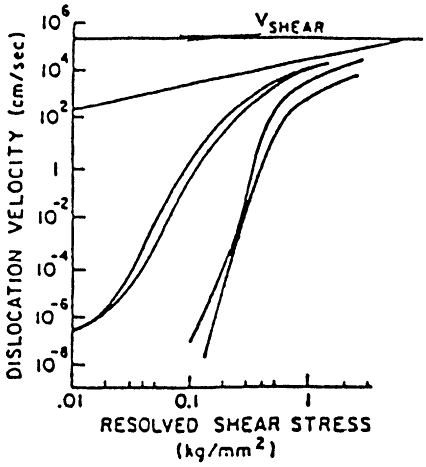

FIG. 6. Dislocation velocities as a function of applied stress in NaCl. The straight line is a linear extrapolation of ultrasonic velocities to higher stresses. The curved lines are the etch-pit measurements of Gutmanas *et al.* (1963) (after Fanti *et al.*, 1969).

For the kilohertz region, the latter includes making both modulus and damping measurements, and for the megahertz region, both modulus and damping measurements as a function of frequency. In experiments, the main complicating feature found over the simple model is the possibility of a distribution of line tensions. This must be assessed by studies of **3** above. The conditions mentioned favor irradiation techniques over plastic-deformation techniques. In the former, low dislocation densities may be used, reducing interaction possibilities, and better estimates of defect concentrations and distributions may be obtained.

2.2.2.2 Amplitude-Dependent Damping Theory. For high enough stress amplitudes, the dislocations may be pulled away from the weak pinning points, as illustrated in Fig. 7. A large increase in strain for no increase in stress then occurs, and for a cyclic stress, the area in the stress–strain curve represents the energy loss per half cycle. This loss is amplitude dependent. It is zero until the stress amplitude equals the mechanical breakaway stress $\sigma_M = U_0/b^2L$, where U_0 is the pinning energy (Cottrell, 1948), and L is the distance between weak pinning points. Beyond this, the decrement, which by definition is the loss per half cycle divided by the maximum stored elastic energy $\sigma^2/2G$, decreases as $1/\sigma_0^2$ and is as shown in Fig. 7(c). In a real crystal, however, there is a distribution of dislocation segments L, which changes the shape of the decrement vs amplitude from that shown in Fig. 7(c) to that in Fig. 7(d).

It might be anticipated that the effect of thermal fluctuations would be to change the breakaway stress from σ_M at $T = 0$ to a lower value σ_T at finite T, as indicated in Fig.

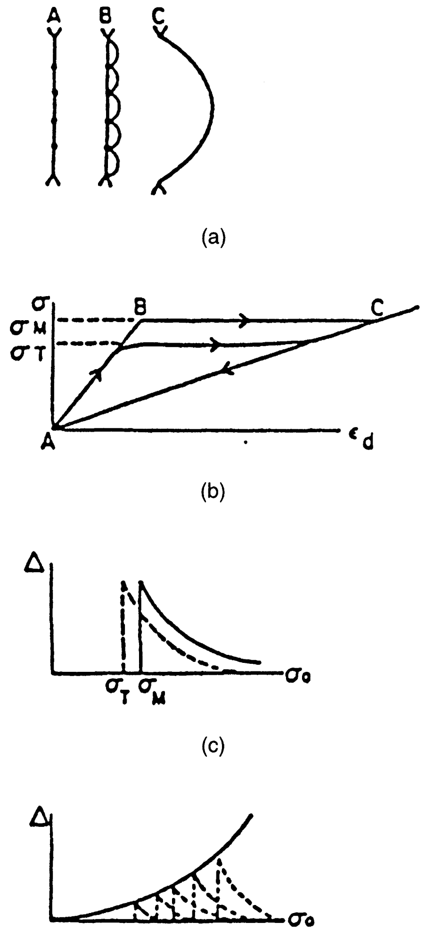

(a) (b) (c) (d)

FIG. 7. (a) Dislocation displacement as a function of increasing stress for a dislocation segment with weak interior and strong exterior pinning points. (b) Stress–dislocation-strain curve corresponding to (a) for a cyclic stress. Breakaway occurs at a lower stress σ_T at finite temperatures than the mechanical breakaway stress σ_M. (c) Decrement vs stress amplitude for segments of same length. (d) Decrement vs stress amplitude for a distribution of weak pinning points.

7(b). A summary and physical discussion of the theory is given by Granato and Lücke (1981) and Teutonico *et al.* (1964). For the low-temperature region, the principal results are as follows:

1. The stress amplitude $\sigma_\Delta(T)$ required to produce a constant Δ, at a temperature T, corresponds to a constant fraction of dislocation segments being broken away from point defects and is directly related to calculations for the yield stress of macroscopic flow measurements.
2. For high stress amplitudes, low temperatures, and relatively pure materials, the stress should have the temperature dependence

$$\sigma = \sigma_M - AT^{2/3}, \tag{9}$$

   where σ_M is the stress required to overcome the pinning barrier in the absence of thermal fluctuations and A is a constant for constant frequency regardless of the form of the dislocation force–distance law, provided only that the force has a smooth maximum.
3. For low stresses, the activation process no longer depends only on the details of the force–distance law. For such low stresses, breakaway from a single pinning point becomes ineffective because the broken-away state is a relatively high-energy state, and the repinning rate is higher than the unpinning rate. Breakaway to a lower-energy state can now occur only by simultaneous unpinning from more than one pinning point, but the activation energy for this is higher, and the process occurs only at higher temperatures.

Measurements of the critical unpinning stress (stress amplitude for given decrement value) are closely related to macroscopic flow-stress measurements for dilute alloys. An observation of the $T^{2/3}$ law of Eq. (9) is of special significance since this form is required theoretically on very general grounds, and measurements that purport to measure the force–distance profile between a dislocation and a single point defect should meet this requirement as a necessary and crucial test.

Superconducting materials entering the superconducting state suffer a loss of mechanical strength. Since the initial reports by Kojima and Suzuki (1968) and Pustovalov *et al.* (1967), many studies have been made, which have been reviewed by Startsev (1983). This effect was unexpected and unexplainable on the basis of traditional theories of plastic flow, which take no account of viscous or inertial effects.

Internal-friction measurements by Schwarz *et al.* (1977) on pure and dilute-alloy crystals of copper and lead are shown in

Fig. 8. The maximum in the stress of the dilute alloy of copper is similar to a maximum in the flow stress of a similar specimen seen by Kamada and Yoshizawa (1971). For $T <$ 6 K, the data for pure copper follow the strict $T^{2/3}$ dependence of Eq. (9), as had been found earlier for pure aluminum (Schwarz and Granato, 1975) and which was explained in terms of a classical thermally activated process. The measurements in lead are similar to those in copper for the two purities, except that there are additional changes during the *N*–*S* transition. This change can be used as a key to determine whether the dislocations are overdamped or underdamped in the normal state. In the pure lead, the curve for the *S* state joins smoothly onto the curve for the *N* state below the transition temperature where the drag is still increasing. This shows that the dislocations in pure lead are overdamped. In contrast, dislocations in the dilute alloy are underdamped in the *N* state so that they can become overdamped when the damping is increased when the temperature is raised sufficiently that phonons contribute significantly to the viscous damping. This explains the existence of the anomalous maximum observed for the dilute lead alloy at about 20 K. The fact that the temperature dependence of $\sigma_0(\Delta)$ in the *N* state of the dilute lead alloy is essentially the same as in the dilute copper alloy provides strong experimental evidence that the anomalous maximum in the dilute copper alloy is of inertial origin. Figure 8 also shows that below $T = 3.5$ K, the stress amplitude $\sigma_0(\Delta)$ for the *S* state resumes a linear $T^{2/3}$ dependence with approximately the same slope as in the *N* state. Thus the change in stress vs temperature is predicted well by a purely mechanical model. However, the temperature dependence of the absolute stress is characteristic of a rate process. This suggested to Schwarz *et al.* (1977) that in the presence of inertial effects, dislocation motion is initiated by a thermally activated process and continues by the overcoming of additional obstacles inertially. The amount of plastic deformation in the latter process is determined by the degree of underdamping.

A theory of this process was given by Isaac and Granato (1988), who developed a Brownian-motion rate theory that incorporates viscous and inertial overshoot effects. The theory is an extension to the many degrees of freedom corresponding to dislocation motion of some earlier results found for Brownian motion of a particle in a force field by Kramers (1940), Chandrasekhar (1943), and Wang and Uhlenbeck (1945).

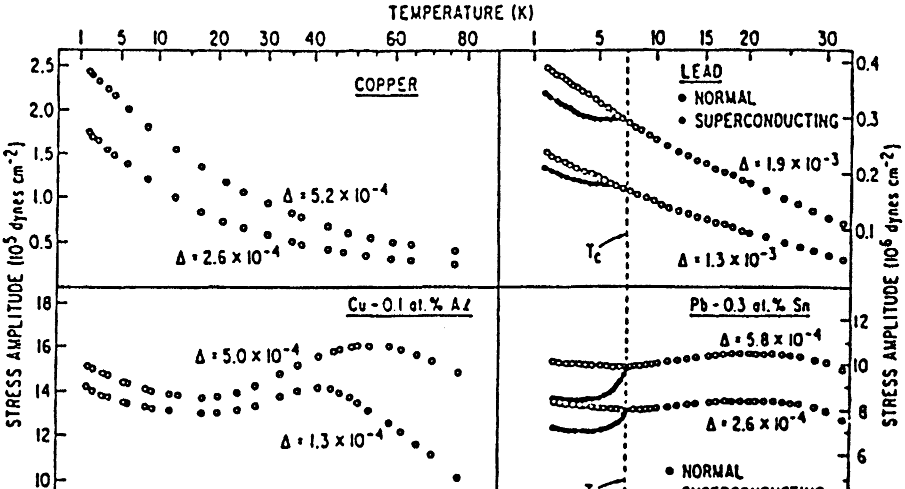

FIG. 8. Temperature dependence of the stress amplitude for constant decrement in internal friction measurements on pure and dilute alloys of copper (left column) and lead (right column) (after Schwarz *et al.*, 1977).

A Langevin equation for a dislocation in a stress and impurity field is solved for a distribution function $p(x,v,t)$, which gives the probability of finding a dislocation at position x with velocity v at time t. The form of this distribution at once explains one of the principal puzzles about many observations of the temperature dependence of the flow stress and its change at the superconducting transition. This is that a purely mechanical inertial theory (Granato, 1971a, 1971b) gives a satisfactory account of the changes in flow stress at given temperature, including the predictions, subsequently verified (Schwarz *et al.*, 1977), that the flow stress change will go to zero below the transition temperature for sufficiently pure materials, and that an anomaly in the temperature dependence of the flow stress above the transition temperature will appear otherwise. On the other hand, this mechanical theory cannot account for the fact that the absolute value of the flow stress has a normal temperature dependence expected from thermal overcoming of barriers in either the normal or superconducting state. This mechanical–thermal duality is explained by the form of the distribution function, which tends to be a cloud centered about the purely mechanical path, which differs for two different viscosities. The path depends on the viscosity while the width of the cloud depends upon the temperature. If one holds the temperature constant and changes the viscosity, the cloud moves, but overall the change is mechanical-like. On the other hand, if one holds the viscosity fixed, and varies the temperature, the cloud position stays fixed, but the cloud width changes, thus emphasizing the fluctuation aspect of the thermal bath. The superconducting transition thus provides a unique opportunity to emphasize either the systematic part of the Brownian-motion effect leading to viscosities or the random part leading to thermal fluctuations.

In fact, these considerations are general, applying for any diffusing "particle" in an external force field. The quantities needed in the theory are the viscosity B and the resonant frequencies of the "particle." The advantages of plastic flow measurements are that dislocations are Brownian particles for which

1. these have been calculated, and mechanisms
2. have been determined,
3. can be measured acoustically, and
4. can be controlled (as in the normal to superconducting transition).

2.3 Point Defects

The diffusion of point defects over obstacles under the influence of an external force is a process that plays a role in many different kinds of physical effects. These have been well studied classically and are normally described by an Arrhenius temperature dependence. More recently, there has been increased interest in quantum mechanical tunneling effects and the effect of viscosity on tunneling rates.

Many different techniques have been applied to such measurements, but ultrasonic measurements have particular advantages. Dipolar point defects interact strongly with ultrasonic shear stress. Measurements of ultrasonic attenuation and velocity as functions of polarization, frequency, temperature, and defect density display both resonance and relaxation behavior. Ultrasonic measurements in the 10–100-MHz range enjoy benefits generally attributable to internal-friction techniques (Robrock *et al.*, 1977). Most notably, they provide information about the symmetry of defects, can often be used at low defect concentration where defect interactions can be neglected, and permit the determination of properties of individual defect species even when several different types of defects are present simultaneously in the sample. In addition ultrasonic measurements can be used to obtain all independent elastic constants and the bulk resistivity from the attenuation in the same specimen. They are particularly well suited for the study of quantum systems. Information about energy gaps and stress coupling strengths is readily obtained from ultrasonic attenuation peaks and velocity changes at low temperature for defects in concentrations of a few parts per million. These peaks would normally occur, if at all, below the readily accessible temperature range for internal-friction measurements at lower frequencies.

Textbooks are available (Nowick and Berry, 1972; De Batist, 1972) for detailed discussions of classical point-defect relaxation effects. Resonant effects for $\omega \ll \omega_0$, where ω_0 is the resonance frequency, are discussed by Granato (1994a). Quantum effects for both relaxation and resonance are reviewed by Hunklinger and Arnold (1976). We outline here the quantum theory, which contains the classical theory as a high-temperature limit.

To illustrate the types of effects that are found, we focus on the system of hydrogen trapped at an impurity in niobium. This is a rather rich system, which has been well studied, with many aspects arising from isotope and multiple-trapping possibilities. We then outline the theory for the simplest system, the two-level tunneling system (TLS), which already contains most of the physical effects of interest, and is easily extended to more complicated systems.

One of the most active areas of condensed matter physics over the past two decades has been the question of the properties of noncrystalline materials, or the structure, dynamics, and physical properties of amorphous and liquid materials. It is known from low-temperature studies of specific heat, thermal conductivity, and ultrasonic attenuation that the properties can be understood in terms of a TLS model, with an unspecified tunneling entity. The same parameters appear in the theories of both the thermal properties and the ultrasonic properties. The system of hydrogen in niobium is a model system for such studies because it is simple, with no need to average over a distribution of barrier heights, as is necessary in the TLS model for amorphous or glassy systems. An additional question of much current interest is the question of the effect of viscosity on tunneling rates. A theory by Caldeira and Leggett (1981) predicts that tunneling rates are reduced by the viscosity provided by interactions of the tunneling system with the environment. There are unique possibilities of studying this effect provided by the possibility of observing the ultrasonic response of tunneling systems in both the normal and superconducting states of superconductors.

The general program for ultrasonic measurements is to measure the temperature, frequency, polarization, and isotope dependence of the attenuation β and velocity v of ultrasonic waves. These represent the out-of-phase and in-phase response of the system to a periodic external force. The attenuation and velocity changes can alternatively be expressed as the quality factor Q^{-1} and modulus change $\Delta C/C$ using $Q^{-1} = \alpha\lambda/\pi$ and $v = \sqrt{C/\rho}$, where λ is the sound wavelength, C is the elastic modulus, and ρ is the density. Q^{-1} and $\Delta C/C$ are given as

$$Q^{-1} = \Delta_R \frac{\omega\tau}{1 + \omega^2\tau^2}$$

and

$$\frac{\Delta C}{C} = \Delta_R \frac{1}{1 + \omega^2\tau^2} + \Delta_s, \tag{10}$$

where Δ_R is the relaxation strength, Δ_s is the resonance strength, τ is the relaxation time, ω is the frequency, and $\omega \ll \omega_0$. For a classical system

$$\Delta_R(T) = \frac{\alpha^2 c}{CkT},$$
$$\Delta_s = \beta c,$$

and

$$\tau(T) = \tau_0 \exp(U/kT), \tag{11}$$

where α is the coupling constant for the defect interaction with the sound wave, β is the diaelastic susceptibility, τ_0 is the prefactor, U is the barrier height for the relaxing point defect, and c is the defect concentration. The same form [Eqs. (10)] holds for quantum systems, but the temperature dependence is different. The relaxation height goes to zero for low enough temperature, and the relaxation time is often weaker than exponential. Quantum effects are generally recognized, though, by analysis of the temperature dependence of Δ and τ. It can happen that the relaxation time is Arrhenius in form but still be quantum in nature. But then the preexponential factor τ_0 is generally much smaller than a characteristic atomic period. By analysis of $\Delta_R(T)$ and $\tau(T)$, the basic parameters of the tunneling system, α and Δ (tunneling gap), are deduced.

2.3.1 Tunneling of Trapped Hydrogen in Niobium Measurements by Sellers *et al.* (1974) of a large isotope effect in the excess specific heat of niobium containing hydrogen or deuterium showed convincingly that hydrogen in niobium formed a quantum tunneling system. More detailed measurements of specific heat by Morkel *et al.* (1978) showed that if there is no oxygen or nitrogen impurity, then there is no excess specific heat. This implies that the tunneling system consists of a complex made up of hydrogen trapped at an oxygen or nitrogen impurity. More specific information has since been obtained from neutron-scattering and ultrasonic measurements. A brief review of these results with references is given by Drescher-Krasicka and Granato (1985).

An example of the type of measurement obtained for the Nb–O/H system is shown in Fig. 9. The specimen contains 700 ppm of oxygen and 100 ppm of hydrogen. In the superconducting state, a small relaxation peak is found for the *C′* mode at 2.5 K at 10 MHz. The effect is small but can be measured with great accuracy. The relaxation strength behaves quantum mechanically, going to zero at low temperatures. The velocities in these samples show a dispersion at 2.5 K, but they also decrease further at lower temperatures. The latter decrease is a diaelastic effect. In the normal state there is no evidence of a relaxation. Since the hydrogen effect occurs in the *C′* mode, but not in the C_{44} mode, we obtain immediately important symmetry information useful for discriminating between candidate models for the configuration, which is information not available from specific-heat or neutron-scattering experiments. The relaxation nature of the peak in the superconducting state is established from the shift in the peak position with frequency in the usual way. The remarkable feature of this peak is that it disappears in the normal state, showing a strong and unusual dependence of the relaxation state on the normal electron density. In the normal state, even with the absence of a relaxation, the modulus decreases with decreasing temperature. This is a quantum resonance effect, about which more will be said later.

Measurements as a function of frequency between 10 and 150 MHz permit a determination of the temperature dependence of the relaxation strength and relaxation time. The relaxation time and strength found by Poker *et al.* (1979) are shown in Figs. 10 and 11. The relaxation strength deviates from the classical 1/*T* dependence. The relaxation rate is Arrhenius, but the intercept is too low for a classical process. A classical 1/*T* dependence is shown for reference.

If a specimen containing 700 ppm H and 100 ppm O is rapidly quenched from 240 K, a second much larger peak appears (Huang

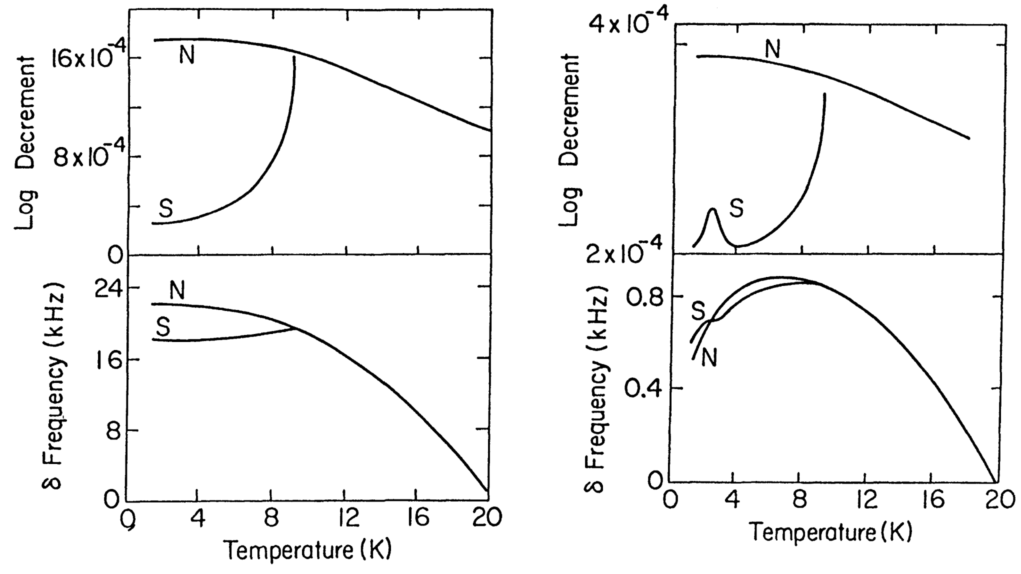

(a) (b)

FIG. 9. Decrement (Δ) and velocity change ($\Delta v/v = \Delta f/f$) vs temperature at 10 MHz for Nb containing 100 ppm of oxygen and 700 ppm of hydrogen for the (a) C_{44} mode and (b) *C′* mode (after Drescher-Krasicka and Granato, 1985).

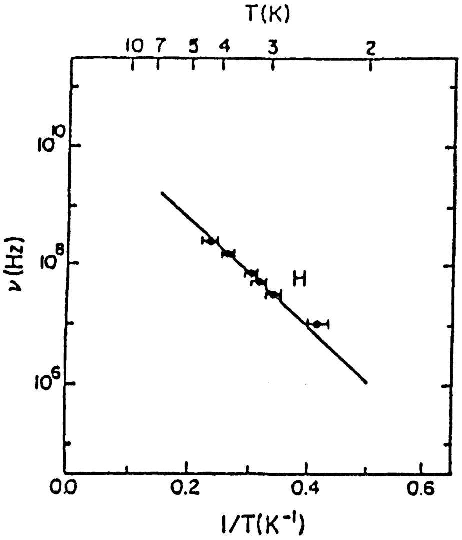

FIG. 10. Logarithm of relaxation frequency as a function of inverse temperature. Straight line is a fit to the data with an activation energy of 1.8 meV.

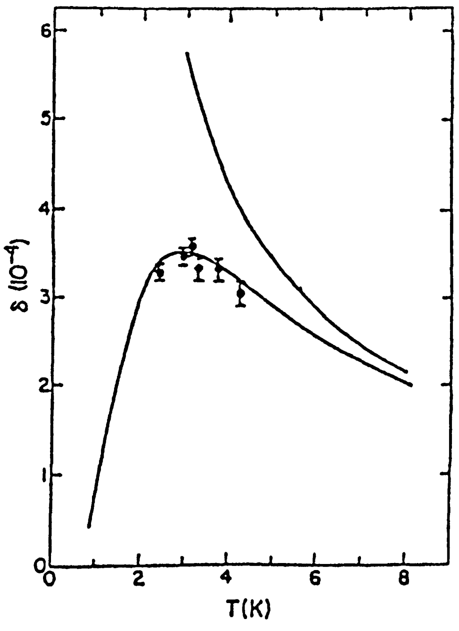

FIG. 11. Relaxation strength of the *C′* mode vs temperature for hydrogen.

et al., 1985) near 6.5 K. This peak is unstable against annealing near 90 K, as shown in Fig. 12.

2.3.2 Theory of Two-Level Systems. Relaxation and Resonance Effects The theory of attenuation and elastic-constant changes for TLS in crystals is basically the same as that for amorphous systems, but simpler. For amorphous systems there is not only a distribution of strains but also a distribution of tunneling gaps, extending down to zero. For TLS in crystals, there is a minimum gap, Δ_0, and for ultrasonic measurements $\hbar\omega < \Delta_0$, there is no direct resonant absorption. The theory was given already by Jäckle (1972) and applied immediately to amorphous systems.

The elastic constant C is given by definition as

$$C = \frac{\partial^2 F(\epsilon)}{\partial \epsilon^2}, \tag{12}$$

where F is the free energy per unit volume, ϵ

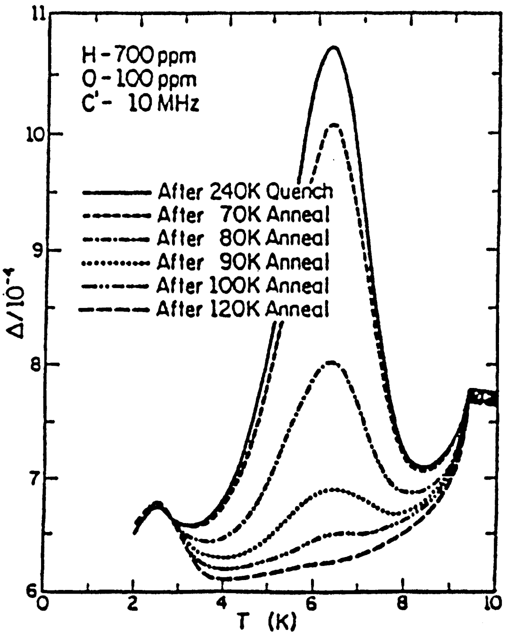

FIG. 12. Decrement vs temperature for a 10-MHz *C′* mode in a niobium crystal containing about 100 ppm O and 700 ppm H after a rapid cooldown from 240 K. The rapid change of attenuation near 9.2 K is due to the transition from the normal to the superconducting state.

is an elastic strain, and C is the elastic constant belonging to that strain. The free energy is given by

$$F = -kT \ln Z, \tag{13}$$

where Z is the partition function given by

$$Z = \sum_i \exp(-E_i/kT), \tag{14}$$

where $E_i(\epsilon)$ are the states of the system. The relaxation and resonance effects can be illustrated with the simple two-level system shown schematically in Fig. 13. For $\epsilon = 0$, the dipole is supposed to have two equivalent orientations. The ground state is tunnel-split with a gap of 2Δ. If a stress or strain ϵ is applied, one orientation becomes energetically favored. If time is available, transitions will occur, bringing the system back into thermal equilibrium.

For $T \ll \Delta$, only the ground state is populated. For small ultrasonic stress amplitudes, the elastic constant is given by the curvature of the ground state. The dipole moment is induced, proportional to the strain ϵ, and the change is called diaelastic. This response has a resonance-like character. The resonance frequency is normally far above the megahertz range. As a result, no attenuation is seen, but only the low-frequency change of elastic constant is found.

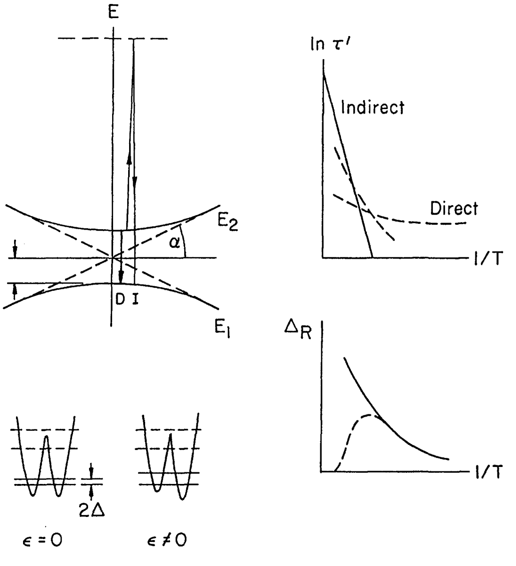

FIG. 13. Schematic illustration of potential, energy levels, transition processes, transition rates, and relaxation strength Δ_R for a two-level system, or a four-level system including dashed lines for E. Classically $\Delta_R = 0$, τ^{-1} is Arrhenius, and $\Delta_R \propto 1/T$.

At finite temperature, higher levels become populated, and phonon-induced transitions become possible. For the two-level system shown, however, no relaxation takes place for small-amplitude ultrasonic waves about $\epsilon = 0$ because the energy of the states is flat for small ϵ, maintaining equilibrium population of the states. For $\Delta = 0$, the states are given by the dashed lines. This can represent a classical Snoek system—for example, Fe–C (where one of the levels is doubly degenerate). Now if a strain is applied, equilibrium is disturbed, transitions take place, and the effect is that of a permanent dipole reorienting in a field. This is a paraelastic effect, represented by a relaxation process. The low-frequency (high-temperature) elastic constant is given by Eqs. (12)–(14), and the response for all frequencies, by Eq. (10).

For classical systems Δ_R is inverse with temperature as a result of the competing thermal disordering and external strain-aligning effects. Also Δ_R is proportional to the defect concentration and the square of the strain coupling constant α, given by the slope of $E(\epsilon)$. The relaxation time is Arrhenius with an activation energy U given by the height of the barrier. Transitions may occur by several processes, which act in parallel. Direct transition rates are independent of temperature at the lowest temperature and occur by tunneling, increasing linearly with temperature at higher temperature where they become phonon assisted (Jäckle, 1962). Indirect transitions through other states may also occur, but with a higher temperature dependence. The fastest rate dominates the temperature dependence. At higher temperatures, in the classical regime, transitions through states near the top of the barrier are most effective.

For a somewhat more complicated four-state system, the four levels indicated in Fig. 13 by both the solid and the dashed lines occur. In this system, relaxation paraelastic effects are obtained from the (degenerate at $\epsilon = 0$) first excited states and resonance diaelastic effects from the others. The relaxation disappears, however, for $T < \Delta$ since the excited states become depopulated. The relaxation strength Δ_R then falls below the classical $1/T$ curve at a temperature determined by Δ.

For TLS made asymmetric by a strain ϵ, the gap 2Δ is given by the three parameters Δ_0, α, and ϵ in

$$\Delta_{1,2} = \mp(\Delta_0^2 + \alpha^2\epsilon^2)^{1/2}, \tag{15}$$

where $2\Delta_0$ is the tunnel splitting, $2\alpha\epsilon$ is the strain-induced asymmetry energy, α is a strain coupling coefficient, and ϵ is the external elastic strain. Then

$$Z = 2\cosh(\Delta/kT). \tag{16}$$

Using Eqs. (13)–(16) in (12), one obtains

$$\Delta C/C = \Delta_R + \Delta_S, \tag{17}$$

where

$$\Delta_R = \left(\frac{c\alpha^2}{CkT}\right)\left[\frac{\alpha\epsilon}{\Delta}\right]^2 \operatorname{sech}^2\left(\frac{\Delta}{kT}\right) \tag{18}$$

and

$$\Delta_S = \left[\frac{c\alpha^2}{C\Delta_0}\left(\frac{\Delta_0}{\Delta}\right)^3\right]\tanh\left(\frac{\Delta}{kT}\right) + \gamma c. \tag{19}$$

The factor in parentheses in Eq. (18) is the familiar result for the classical Snoek effect for $E_{1,2} = \mp\alpha\epsilon$. The factor $\operatorname{sech}^2(\Delta/kT)$ cuts off the relaxation at temperatures low enough for the upper state to become depopulated. What is new for a quantum system is the bracketed factor in Eq. (18) for Δ_R and Eq. (19) for Δ_S. The factor $(\alpha\epsilon/\Delta)^2$ in Eq. (18) shows explicitly that for a symmetric potential, there will be no relaxation because there is no energy shift for a small perturbing ultrasonic strain. For large strains $\alpha\epsilon/\Delta \to 1$, the new term Δ_S is the resonance response. Even though there is no resonant absorption for $\hbar\omega \ll \Delta_0$, there is still a response in phase with an applied strain. At zero temperature, where only the ground state is populated, Δ_S is given by the square bracket in Eq. (19), as is readily calculated using $F = E_1$. From Eq. (12), Δ_S is the curvature of this state as a function of strain. This curvature goes to zero for large strains where E_1 becomes linear in strain. At finite temperature, there is some population of E_2 with opposite curvature, reducing the size of the effect by the factor $\tanh(\Delta/kT)$.

For high temperatures,

$$\begin{aligned}\Delta_R &= \frac{\alpha^2 c}{CkT}\frac{\alpha^2\epsilon^2}{\Delta_0^2 + \alpha^2\epsilon^2},\\ \Delta_S &= \frac{\alpha^2 c}{CkT}\frac{\Delta_0^2}{\Delta_0^2 + \alpha^2\epsilon^2} + \gamma c,\end{aligned} \tag{20}$$

and $\Delta C/C$ achieves the classical value

$$\frac{\Delta C}{C} = \Delta_R + \Delta_S = \left(\frac{\alpha^2}{CkT} + \gamma\right)c. \tag{21}$$

For a random distribution of strain, these results must be averaged over a strain distribution function, with the results expressed now in terms of ϵ_0, the average absolute value of the strain. Below the temperature for which $\omega\tau = 1$, the elastic-constant change is given by the resonant response Δ_S. In what follows, we focus mainly on the velocity measurements. For small strain $\beta = \alpha\epsilon_0/\Delta \ll 1$, the resonance is large and relaxation is a small fraction of the total elastic-constant change at high temperatures. For large strain, the resonance is greatly reduced, and the relaxation is a larger fraction of the total change at high temperatures. The total change is the same for both strains at high temperature. For $kT < \Delta_0$, $\Delta C/C$ deviates from the $1/T$ classical high-temperature response and becomes constant for $kT \ll \Delta_0$.

These features provide a means for analyzing data in terms of a TLS formalism. In the superconducting state $\Delta C/C$ changes from $\Delta_R + \Delta_S$ at high temperatures to Δ_S below $\omega\tau = 1$. In the normal state $\Delta C/C$ is given by $\Delta_R + \Delta_S$ over the whole measured temperature range. A deviation from $1/T$ behavior signals the approach of kT to Δ_0. A relaxation that is a small fraction of the total elastic constant change means β is small.

2.3.3 Acoustic Measurements of Tunneling Hydrogen in Niobium The velocity data for a rapidly cooled specimen (Drescher-Krasicka and Granato, 1985) are plotted in Fig. 14. One sees that the resonance is smaller than the relaxation in this case. Independent of the nature of the tunneling system and the strain distribution, the temperature-independent velocity change at low temperature must be a resonance effect.

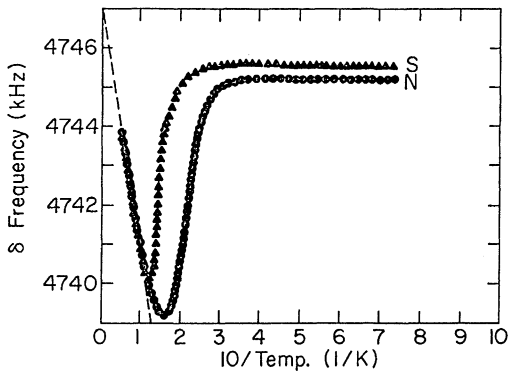

FIG. 14. Frequency change vs 10/T for rapidly cooled specimen. The relaxation strength is large compared to the resonance strength. The velocity change observed at low temperature is larger in the normal state than in the superconducting state (after Drescher-Krasicka and Granato, 1985).

The fact that this is different in the N and S states implies that the tunneling gap must also have changed. This is strong evidence for a correlation of the energy gap with the viscosity of the system.

In Fig. 15, the elastic-constant data, by Maschoff *et al.* (1986), for C' at 10.8 MHz for both the N and S states are plotted vs $1/T$. The two curves, after correction for the electron–phonon interaction, do not come together at the lowest temperatures, as should be expected if Δ_0 does not change. This curve is obtained by subtracting the host and electron–phonon terms from the data. At high temperatures, both the N- and S-state curves vary as α^2/T. This remains true for a strain distribution. Below 2.4 K, in the S state, the relaxation component Δ_R is lost, and only the resonance component Δ_S remains. The deviation from a $1/T$ dependence in the N state occurs for $kT = \Delta_0$, showing directly that $\Delta_0/k \approx 1$ K. The relative size of Δ_R/Δ_S is a measure of $\alpha\epsilon/\Delta_0$, and one sees directly that $\alpha\epsilon < \Delta_0$. Quantitative values for the three parameters of the model, Δ_0, α, and the Lorentzian strain distribution parameter ϵ_0, are obtained by fitting the data as shown by the solid lines in Fig. 15. These are $\Delta_0(S) = 0.094$ meV, $\alpha = 45$ meV, and $\epsilon_0 = 1.1 \times 10^{-3}$. The elastic-constant change, $\delta C(S\text{–}N) = 5.1 \times 10^{-5}$, is 21% and corresponds to a tunneling-parameter change of $\{\Delta_0(S) - \Delta_0(N)\}/\Delta_0(S) = 36\%$. The middle curved line in Fig. 15 is the calculated N-state curve one would expect if Δ_0 does not change, using the parameters Δ_0, α, and ϵ_0 obtained in fitting the S-state data.

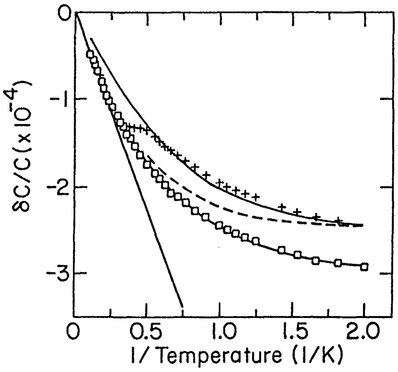

FIG. 15. Relative modulus change vs 1/T after background subtraction for N(□) and S(+) Nb–O–H and fits to the two-level system model (solid lines). The middle curve (dashed) is the computed N-state modulus for no change in Δ_0 (after Maschoff *et al.*, 1986).

2.3.4 Classical Point-Defect Resonance Effect The effect is normally difficult to detect because $\Delta s = \beta c$ in Eq. (20) is both frequency and temperature independent. Although relative changes in $\Delta c/c$ can be measured to great accuracy, absolute changes cannot. Since Δs is normally small, it is only detected when the defect concentration is changed during a measurement, as in a radiation-damage experiment where measurements can be made as a function of radiation flux, as shown in Fig. 16.

This is an example of an important application of acoustic measurements to the determination of the structure of self-interstitials in copper from measurements of the elastic-constant changes measured in thermally radiated copper single crystals (Rehn *et al.*, 1974), where about three parts per million of interstitials were introduced into a sample. One gets changes of the elastic constant that are not very big, in the largest case about 100 ppm, but one can measure that acoustically with very great accuracy. The important features of the results are, first of all, that the effect is linear in defect density and large, about 30 ppm per parts per million of defects for the C_{44} shear constant. The effect is very anisotropic, and this is an important clue for the identification of the structure of the interstitial. Thirty is a very

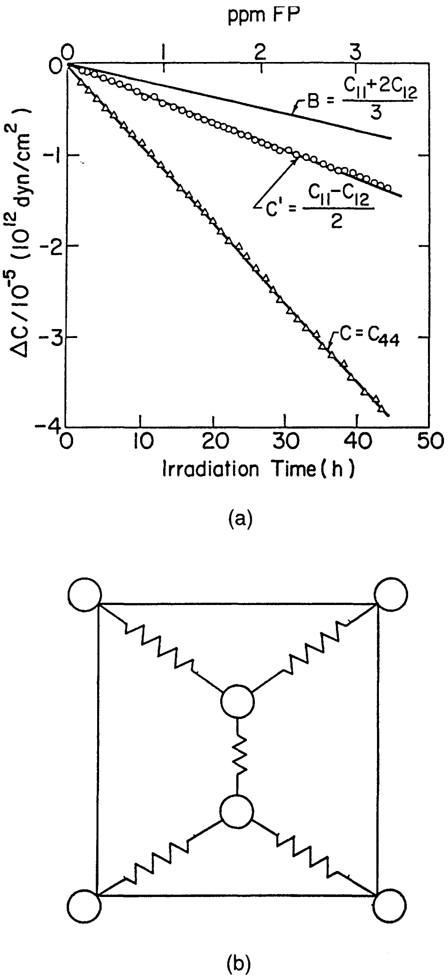

FIG. 16. (a) Change of the elastic constants of a Cu single crystal after irradiation with thermal neutrons at 4 K (data after Rehn *et al.*, 1974). (b) Split ⟨100⟩ configuration of an interstitial in Cu.

large effect. Normally, property changes such as the volume change or the bulk modulus change per point defect are of order unity. This is more than one order of magnitude larger and that shows that the interstitial produces a very large elastic polarizability in the sample. These clues permit one to identify the structure of the interstitial among several possibilities as one for which two atoms are trying to sit on a normal lattice site, spread out into a dumbbell configuration oriented along ⟨100⟩, as shown in Fig. 16(b), interacting strongly with its neighbors in such a way that it is very susceptible to a shear stress of one type, as compared to the other shear stress, giving a soft spot in the crystal with a resonance mode of the order of $\frac{1}{7}$ of the Debye frequency.

These results form the basis for a new model for liquids and amorphous or glassy solids. According to the interstitialcy model (Granato, 1992), liquids are crystals containing a few percent of interstitials in thermal equilibrium, and glasses are frozen liquids. The strong shear diaelastic polarizability of the interstitial reduces the interstitial formation energy with concentration so that substantial numbers of interstitials exist in thermal equilibrium at high temperature. The properties established for interstitials in fcc metals of

1. a strong shear-modulus softening,
2. a degeneracy in orientation, and
3. low-frequency resonance modes

are those known experimentally to be universal in glasses. Good quantitative agreement has been found for the predictions of the model with the known structure, dynamics, and thermodynamic properties of liquids and glasses (Granato, 1994b).

ACKNOWLEDGMENT

This work was supported by NSF grant DMR 93-19773.

GLOSSARY

Decrement Δ: The ratio of the amplitudes of two successive oscillations of an underdamped harmonic motion. $\Delta = \alpha\lambda = \pi Q^{-1}$, where α is the attenuation, λ is the wavelength, and Q is the quality factor.

Diaelastic Effect: Diaelasticity and resonance are the same. They are the direct response corresponding to an induced elastic dipole under an applied force even when there is no motion over or through a barrier.

Internal Friction: Conversion of mechanical strain energy to heat within a material subjected to fluctuating stress.

Paraelastic Effect: Paraelasticity and relaxation are the same. The defect strain results from the competition between the ordering effects produced by an external field on an existing elastic dipole and the disordering effects due to thermal fluctuations. It requires thermal activation over or tunneling through an energy barrier and is normally therefore strongly temperature dependent.

Relaxation Effect: Same as **Paraelastic Effect.**

Resonance Effect: Same as **Diaelastic Effect.**

Second- (Third-) Order Elastic Constants: Coefficients of the quadratic (cubic) terms in a Taylor series expansion of the free energy in terms of strains.

Works Cited

Baker, G. S. (1962), *J. Appl. Phys.* **33,** 1730–1732.

Bardeen, J., Cooper, L. N., Schrieffer, J. R. (1957), *Phys. Rev.* **108,** 1175–1204.

Beamish, J. R., Franck, J. P. (1982), *Phys. Rev. B* **26,** 6104–6175.

Bordoni, P. G. (1949), *Ric. Sci.* **19,** 851–862.

Caldeira, A. V., Leggett, A. J. (1981), *Phys. Rev. Lett.* **46,** 211–214.

Chandrasekhar, S. (1943), *Rev. Mod. Phys.* **15,** 1–89.

Cottrell, A. H. (1948), in: *Report of a Conference on Strength of Solids, H. H. Wills Physical Lab., University of Bristol, 7–9 July 1947,* London: The Physical Society of London, p. 30.

De Batist, R. (1972), *Internal Friction of Structural Defects in Crystalline Solids,* Amsterdam, London: North Holland Publ. Co., New York: American Elsevier Publ. Co., Inc.

Drescher-Krasicka, E., Granato, A. V. (1985), *J. Phys. (Paris),* Colloq. **C10,** 73–80.

Eshelby, J. D. (1949), *Proc. R. Soc. London, Ser. A* **197,** 396–416.

Fanti, F., Holder, J., Granato, A. V. (1969), *J. Acoust. Soc. Am.* **45,** 1356–1366.

Fedorov, F. I. (1968), *Theory of Elastic Waves in Crystals,* New York: Plenum Press.

Frenkel, J., Kontorova, T. (1939), *J. Phys. Acad. Sci. USSR* **1,** 137–149.

Garber, S. A., Granato, A. V. (1970), *J. Phys. Chem. Solids* **31,** 1863–1867.

Garber, J. A., Granato, A. V. (1975), *Phys. Rev. B* **11,** 3990–3997.

Granato, A. V. (1971a), *Phys. Rev. B* **4,** 2196–2201.

Granato, A. V. (1971b), *Phys. Rev. Lett.* **27,** 660–664.

Granato, A. V. (1992), *Phys. Rev. Lett.* **68,** 974–977.

Granato, A. (1994a), *J. Alloys Compounds* **211/212,** 503–508.

Granato, A. V. (1994b), *J. Phys. Chem. Solids* **55,** 931–939.

Granato, A. V., Lücke, K. (1956), *J. Appl. Phys.* **27,** 789–805.

Granato, A. V., Lücke, K. (1966), in: W. P. Mason, R. N. Thurston (Eds.), *Physical Acoustics,* New York: Academic Press, Vol. IVA, pp. 225–276.

Granato, A. V., Lücke, K. (1981), *J. Appl. Phys.* **52,** 7136–7142.

Gutmanas, E., Nadgornyi, Y., Stephanov, A. V. (1963), *Fiz. Tverd. Tela* **5,** 1021 [*Sov. Phys. Solid State* **5,** 743–747].

Hearmon, R. F. S. (1979), (1) The Elastic Constants of Crystals and Other Anisotropic Materials, p. 1; (2) The Third and Higher Order Elastic Constants, p. 245, in: K.-H. Hellwege, A. M. Hellwege (Eds.), *Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology,* Group III, *Crystal and Solid State Physics,* Vol. 11, *Elastic, Piezoelectric, Pyroelectric, Piezooptic, Electrooptic Constants, and Nonlinear Dielectric Susceptibilities of Crystals,* Berlin-Heidelberg-New York: Springer-Verlag.

Hiki, Y. (1972), *Elasticity and Anelasticity,* Tokyo: Kyoritsu (in Japanese).

Holder, J., Granato, A. V. (1971), in: W. P. Mason, R. N. Thurston (Eds.), *Physical Acoustics,* Vol. VIII, New York: Academic, pp. 237–277.

Huang, K. F., Granato, A. V., Birnbaum, A. G. (1985), *Phys. Rev. B* **32,** 2178–2183.

Hunklinger, S., Arnold, W. (1976), in: W. P. Mason, R. N. Thurston (Eds.), *Physical Acoustics,* Vol. XII, New York: Academic, pp. 155–215.

Huntington, H. B. (1958), in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics,* New York: Academic, Vol. VII, pp. 214–351.

Isaac, R. D., Granato, A. V. (1988), *Phys. Rev. B* **37,** 9278–9285.

Jäckle, J. (1972), *Z. Phys.* **257,** 212–223.

Jäckle, J., Piehé, L., Arnold, W., Hunklinger, S. (1976), *J. Non-Cryst. Sol.* **20,** 365–391.

Kamada, K., Yoshizawa, I. (1971), *J. Phys. Soc. Jpn.* **31,** 1056–1068.

Koehler, J. S. (1952), in: W. Shockley *et al.* (Eds.), *Imperfections in Nearly Perfect Crystals,* New York: Wiley, pp. 197–216.

Kojima, H., Suzuki, T. (1968), *Phys. Rev. Lett.* **21,** 896–898.

Kosugi, T., Kino, T. (1993), *Mater. Sci. Eng. A* **164,** 368–372.

Kramers, H. A. (1940), *Physica* **7,** 284–304.

Krautkrämer, J., Krautkrämer, A. (1983), *Ultrasonic Testing of Materials,* 3rd ed., Berlin, Heidelberg, New York: Springer-Verlag.

Leibfried, G. (1950), *Z. Phys.* **127,** 344–356.

Leibfried, G., Ludwig, W. (1961), in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics,* Vol. 12, New York: Academic, pp. 275–444.

Magalas, L. (Ed.) (1996), *Mechanical Spectroscopy,* in press.

Maschoff, K. R., Drescher-Krasicka, E., Granato, A. V. (1986), in: A. C. Anderson, J. P. Wolfe, (Eds.), *Phonon Scattering in Condensed Matter V*, New York: Springer Verlag, pp. 64–66.

Mason, W. P. (1955), *Phys. Rev.* **97,** 557–558.

Mason, W. P. (1958), *Physical Acoustics and the Properties of Solids*, Princeton, NJ: Van Nostrand.

Mason, W. P. (Ed.) (1964–1968), *Physical Acoustics*, Vols. I–V, London, New York: Academic.

Mason, W. P. (Ed.) (1964), *Physical Acoustics, Principles and Methods*, New York: Academic.

Mason, W. P., R. N. Thurston (Eds.) (1970–1984), *Physical Acoustics*, Vols. VI–XVIII, London, New York: Academic.

Morkel, C., Wipf, H., Neumaier, K. (1978), *Phys. Rev. Lett.* **40,** 947–950.

Morse, R. W. (1955), *Phys. Rev.* **97,** 1716–1717.

Morse, R. W., Bohm, H. V. (1957), *Phys. Rev.* **108,** 1094–1096.

Nowick, A. S., Berry, B. S. (1972), *Anelastic Relaxation in Crystalline Solids*, New York, London: Academic.

Orowan, E. (1940), *Proc. Phys. Soc. (London)* **52,** 8–22.

Pippard, A. B. (1955), *Philos. Mag.* **46,** 1104–1122.

Poker, D. B., Setser, G. G., Granato, A. V., Birnbaum, H. K. (1979), *Z. Phys. Chem. N.F.* **116,** 39–45.

Pustovalov, V. V., Startsev, V. I., Didenko, D. A., Fomenko, V. S. (1967), *Fiz. Met. Metalloved.* **23,** 312–318.

Rayne, J. A. (1975), in: D. Lenz, K. Lenz (Eds.), *Proceedings of the Fifth International Conference on Internal Friction and Ultrasonic Attenuation in Crystalline Solids, Aachen, Germany, August 1973*, Berlin: Springer, pp. 13–51.

Read, T. A. (1940), *Phys. Rev.* **58,** 371–380.

Rehn, L. E., Holder, J., Granato, A. V., Coltman, R. R., Young, F. W. Jr., (1974), *Phys. Rev. B* **10,** 349–362.

Ritchie, I. G., Fantozzi, G. (1992), in: F. Nabarro (Ed.), *Dislocations in Solids*, Vol. 9, Amsterdam: North Holland, pp. 57–133.

Robrock, K.-H., Rehn, L. E., Spiric, V., Schilling, W. (1977), *Phys. Rev. B* **15,** 680–688.

Schwarz, R. B., Granato, A. V. (1975), *Phys. Rev. Lett.* **34,** 1174–1177.

Schwarz, R. B., Isaac, R. D., Granato, A. V. (1977), *Phys. Rev. Lett.* **38,** 554–557.

Schwenker, R. O., Granato, A. V. (1969), *Phys. Rev. Lett.* **23,** 918–921.

Seeger, A. (1956), *Philos. Mag.* **651;** (1981), *J. Phys. C* **5,** Suppl. 10, 201–228.

Sellers, G. H., Anderson, A. C., Birnbaum, H. K. (1974), *Phys. Rev. B* **10,** 2771–2776.

Simmons, G., Wang, H. (1971), *Single Crystal Elastic Constants and Calculated Aggregate Properties: A Handbook*, 2nd ed., Cambridge, MA: MIT Press.

Snoek, J. L. (1941), *Physica* **8,** 711–733.

Startsev, V. I. (1983), in: F. R. N. Nabarro (Ed.), *Dislocations in Solids*, Vol. 6, New York: North Holland, Ch. 28, pp. 143–233.

Stern, R. M., Granato, A. V. (1962), *Acta Metall.* **10,** 359–381.

Teutonico, L. J., Granato, A. V., Lücke, K. (1964), *J. Appl. Phys.* **35,** 220–234.

Thompson, D. O., Holmes, D. K. (1956), *J. Appl. Phys.* **27,** 713–723.

Thompson, D. O., Pare, V. K. (1960), *J. Appl. Phys.* **31,** 528–535.

Thurston, R. N., Brugger, K. (1964), *Phys. Rev.* **133,** A1604–A1610.

Truell, R., Elbaum, C., Chick, B. B. (1969), *Ultrasonic Methods in Solid State Physics*, New York, London: Academic Press.

Wang, M. C., Uhlenbeck, G. E. (1945), *Rev. Mod. Phys.* **17,** 323–342.

Zener, C. (1948), *Elasticity and Anelasticity of Metals*, Chicago: Univ. Chicago Press.

Further Reading

Bhatia, A. (1967), *Ultrasonic Absorption*, Oxford: Oxford Clarendon Press.

De Batist, R. (1972), *Internal Friction of Structural Defects in Crystalline Solids*, Amsterdam, London: North Holland Publ. Co., New York: American Elsevier Publ. Co., Inc.

Hunklinger, S., Arnold, W. (1976), in: W. P. Mason, R. N. Thurston (Eds.), *Physical Acoustics*, Vol. XII, New York: Academic, p. 155.

Krautkrämer, J., Krautkrämer, A. (1983), *Ultrasonic Testing of Materials*, 3rd ed., Berlin, Heidelberg, New York: Springer-Verlag.

Mason, W. P. (1958), *Physical Acoustics and the Properties of Solids*, Princeton, NJ: Van Nostrand.

Nowick, A. S., Berry, B. S. (1972), *Anelastic Relaxation in Crystalline Solids*, New York, London: Academic.

Schreiber, E., Anderson, O., Soga, N. (1973), *Elastic Constants and Their Measurements*, New York: McGraw-Hill.

Truell, R., Elbaum, C. (1962), "High Frequency Ultrasonic Stress Waves in Solids," in: S. Flügge (Ed.), *Handbuch der Physik*, Vol. X1/2, Berlin: Springer, p. 153.

Truell, R., Elbaum, C., Chick, B. B. (1969), *Ultrasonic Methods in Solid State Physics*, New York, London: Academic.

Zener, C. (1948), *Elasticity and Anelasticity of Metals*, Chicago: Univ. Chicago Press.

SOLIDS, CRYSTALLINE—MECHANICAL PROPERTIES

ROB PHILLIPS, *Division of Engineering, Brown University, Providence, Rhode Island, U.S.A.*

Introduction 489
1. **Elastic Response of Materials** 495
1.1 Phenomenology of Elastic Deformation 495
1.2 Generalized Hooke's Law 495
1.3 Microscopic Origins of Elastic Moduli 497
1.4 Beyond the Elastic Regime 497
2. **Strength of Crystalline Solids** 498
2.1 Ideal Tensile Strength 498
2.2 Ideal Shear Strength 499
2.3 Defect-Mediated Deformation 500
3. **Permanent Deformation of Materials** 501
3.1 Phenomenology of Permanent Deformation 501
3.2 Time-Independent Plastic Deformation 501
3.3 Creep Phenomena 505
4. **Fracture** 507
4.1 Phenomenology of Fracture 507
4.2 Stress Concentration 507
4.3 The Griffith Concept 509
4.4 Fatigue 512
5. **Influence of Material Properties** 513
5.1 Grain-Size Effects 513
5.2 Work Hardening 514
5.3 Solid-Solution Hardening 515
5.4 Second-Phase Particles 517
6. **Concluding Remarks** 518
Acknowledgments 518
Glossary 518
Works Cited 519
Further Reading 519

INTRODUCTION

Much of what makes the properties of materials so intriguing is a reflection of their response to external stimuli. In addition to their rich electrical and optical properties, materials reveal a host of remarkable responses when subjected to mechanical loads. One abiding concern is that materials fail in dramatic ways. For example, metals have always been of special interest because of their mechanical properties. During World War II, the *Liberty* ships repeatedly broke in half, sometimes under the presumed shelter of safe harbor. Earthquakes continue to fascinate and horrify us if for no other reason than they show us how illusory is the solidity of our surroundings. By way of contrast, whole industries depend upon the ability to shape and form materials in ways that are almost unbelievable. Metal ingots that begin with dimensions of order meters can, after deformation processing, reach dimensions in excess of kilometers. Thus both failure and intentional deformation form the backdrop for the study of the mechanical properties of materials.

The mechanical properties of materials pose challenges on experimental and theoretical fronts alike. The variety of observed mechanical phenomena is perhaps best typified by a specimen loaded in tension. Under these circumstances, a cylindrical sample, for example, is loaded parallel to the cylinder axis, and the resulting elongation is considered as a function of the applied load. Though the province of the mechanical properties of crystalline solids is wider than evidenced in a tension test, this simple experiment serves to fix some of the key ideas that characterize the type of behavior requiring explanation and that the engineer seeks to control—ideas such as elastic deformation, plastic flow, and eventual catastrophic failure via fracture. These same features are exhibited whether the test is carried out in tension, torsion, or bending.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

An example of these features is illustrated schematically via the stress–strain curves shown in Fig. 1. Here it is imagined that the tensile specimen has been subjected to increasing loading, which engenders deformation and ultimate failure. Figure 1(a) shows the response of a ductile material to increasing loads and exhibits plastic deformation while Fig. 1(b) illustrates the stress–strain curve for a brittle ceramic, which exhibits no plastic deformation prior to fracture. Elastic deformation is characterized by its reversibility—once the load is removed, the solid returns to some nascent reference configuration that is free of deformation. *Linear* elastic materials are characterized by a linear relation between stress and strain, just the behavior canonized in the law of Hooke

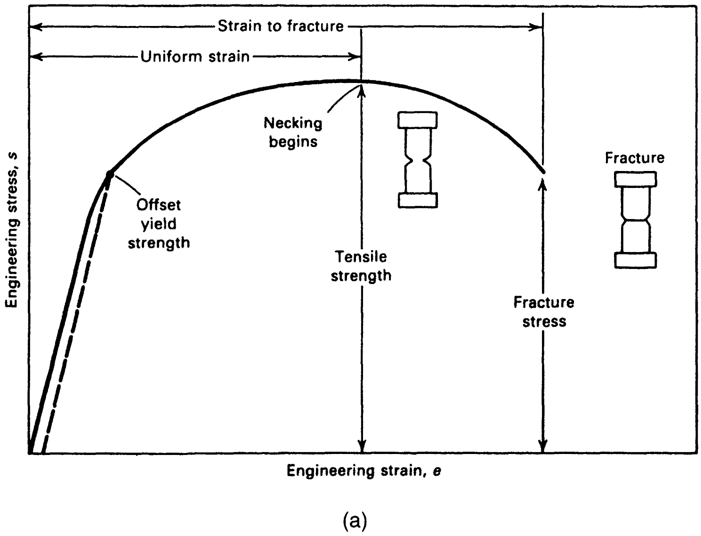
Strain to fracture
Uniform strain
Necking begins
Offset yield strength
Tensile strength
Fracture stress
Fracture
Engineering stress, s
Engineering strain, e
(a)

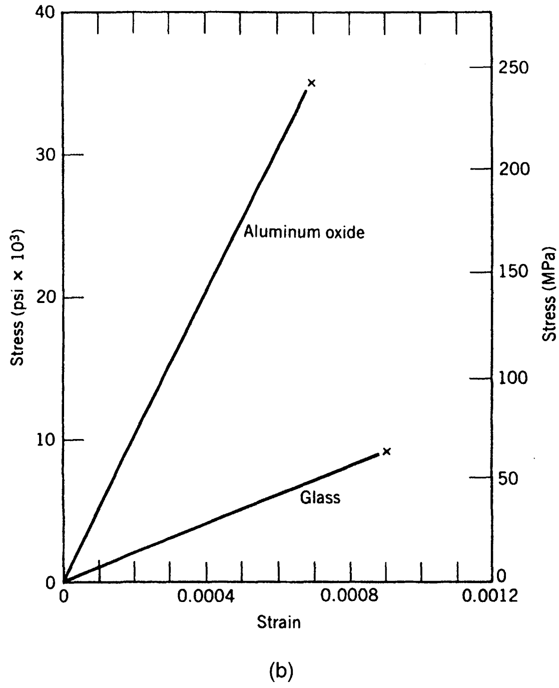

FIG. 1. Generic stress–strain curves for samples loaded in tension. (a) Stress–strain behavior for ductile material that exhibits plastic deformation. After Boyer (1987). (b) Stress–strain curve for brittle ceramics. After Callister (1994).

that bears his name. This behavior is exhibited in the small-strain region of the figure. At a stress known as the yield stress σ_y, ductile materials undergo permanent deformation. Beyond this point, if the load is removed, the sample does not return to its undeformed state; rather, the specimen exhibits a permanent change of shape. This behavior is familiar to anyone who has bent a clothes hanger and is the basis of technologies such as the sheet forming used in shaping car bodies. For a generic ductile solid, after some range of plastic deformation, the solid fails catastrophically via a fracture process. At this point, the loading part of the experiment has ended.

Our discussion of the mechanical response of a generic crystalline solid has thus far been highly idealized, implying that the mechanical response of a particular material is determined by nothing more than the state of applied stress. However, the observed range of mechanical behavior is determined by a number of competing factors (both external and internal) that we have thus far ignored; factors such as temperature, strain rate, and thermal and mechanical loading history are especially important. Ashby has pioneered a diagrammatic representation of the different observed behaviors known as deformation mechanism maps (Ashby and Jones, 1980), an example of which is shown in Fig. 2. This figure depicts the deformation mechanisms (i.e., elastic response, dislocation-mediated plastic deformation, and creep) that can be realized by a material as simple as elemental copper under different loads and temperatures. For example, at high enough temperatures even for modest loads, it is seen that the solid suffers deformation in the form of creep. Creep deformation refers to the permanent plastic deformation of a material at stresses below the yield stress, usually as a result of heightened temperatures. An additional level of complexity is introduced as a result of the dependence of the dominant mode of deformation on strain rate. Contours of constant strain rate are seen as the thin lines that rib the figure and allow for the data at different strain rates to be collapsed onto a single plot.

Not only are there a variety of mechanisms by which a solid may undergo deformation, but their eventual failure via fracture occurs by a number of mechanisms as well. A fracture mechanism map is illus-

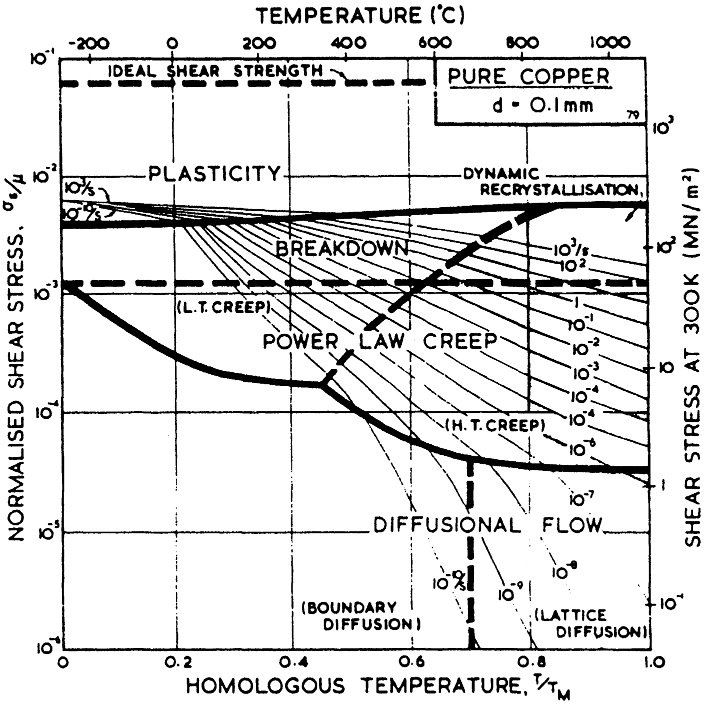

FIG. 2. Deformation mechanism map for elemental Cu with a grain size of 100 μm illustrating various permanent deformation mechanisms available to the solid under thermomechanical loads. Temperature normalized by melting temperature T_m, and stress normalized by shear modulus μ. After Frost and Ashby (1982).

trated in Fig. 3, which demonstrates that at elevated temperatures materials may suffer ductile fracture (fracture preceded by plastic deformation), while at lower temperatures they will fail in a brittle manner via cleavage. One measure of the ductility that is clearly evident in the fracture mechanism map is the reduction in the cross-sectional area of the sample prior to failure.

A second feature that broadens the range of observed mechanical properties is that of various microstructural influences. For example, trace amounts of Au or Al in elemental Cu change the yield stress (stress to plastic deformation) dramatically. This is illustrated in Fig. 4, which depicts the concentration dependence of the yield stress for various impurities in Cu and reveals that the hardening effect is an appreciable fraction of the yield stress itself for a soft metal such as Cu. Similarly, by tuning the grain size of a given material via application of both heat treatment and deformation, the yield strength can again be varied substantially. Thus, while Fig. 1 serves to provide a schematic idea of the types of mechanical behavior, both testing conditions and the material's history should be considered as well.

The geometric description of shape changes in materials and the applied forces needed to maintain them serve as the language of deformation for describing the mechanical behavior of solids. It is most convenient to idealize a solid body as a continuum, with the underlying discrete lattice properties blurred into a set of field variables such as displacement, velocity, strain, and temperature. The deformation of a body is characterized by imagining an initially undeformed state (dubbed the reference configuration) in which the material particles that make up the body are labeled by their Cartesian coordinates **X**. After undergoing some shape change, the material particles find themselves in new positions given by

$$\mathbf{x} = \mathbf{X} + \mathbf{u}(\mathbf{X}), \tag{1}$$

where $\mathbf{u}(\mathbf{X})$ is known as the displacement vector and represents the displacement suffered by the material particle that was initially at position **X**. Such deformations are schematized in Fig. 5, which illustrates that in general the deformation may consist not only of relative motion of material particles but of rotations as well. Though the kinematics of finite deformation as illustrated in Fig. 5 is intriguing in its own right, in this article we emphasize a geometric description of the deformations of interest in terms of the infinitesimal or small-strain tensor, defined by

$$\epsilon_{ij} = \tfrac{1}{2}(u_{i,j} + u_{j,i}). \tag{2}$$

The displacement gradients are given via indicial notation as

$$u_{i,j} = \frac{\partial u_i}{\partial X_j}, \tag{3}$$

and this notation will be used throughout the article. It is important to note that this deformation measure is really only appropriate in the limit of small displacement gradients. In particular, the small-strain tensor, while blind to rigid-body translations, is nonzero in the presence of *finite* rotations. On the other hand, for most purposes, the interesting deformations from the standpoint of the mechanical properties of materials are those involving changes in *relative* position, such as are picked up by the symmetric part of the displacement gradient, more familiar as the strain tensor. The components of the small-strain tensor admit a simple interpretation if we restrict our analysis to the Cartesian directions. The diagonal components of the small-strain tensor determine the stretching of these vectors, while the off-diagonal elements provide a measure of the change in angles between them as a result of the deformation.

While strain serves as the kinematic basis of our discussion of the mechanical properties of materials, kinematics must be supplemented by a description of the forces that result in deformation. In this case, the solid is idealized as a continuum in which forces are transmitted locally and are described in terms of the stress tensor. For example, if we reconsider the tensile specimen described above and imagine a loading in which the tensile axis points along the z direction, the stress tensor is given by

$$\sigma = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_0 \end{pmatrix}. \tag{4}$$

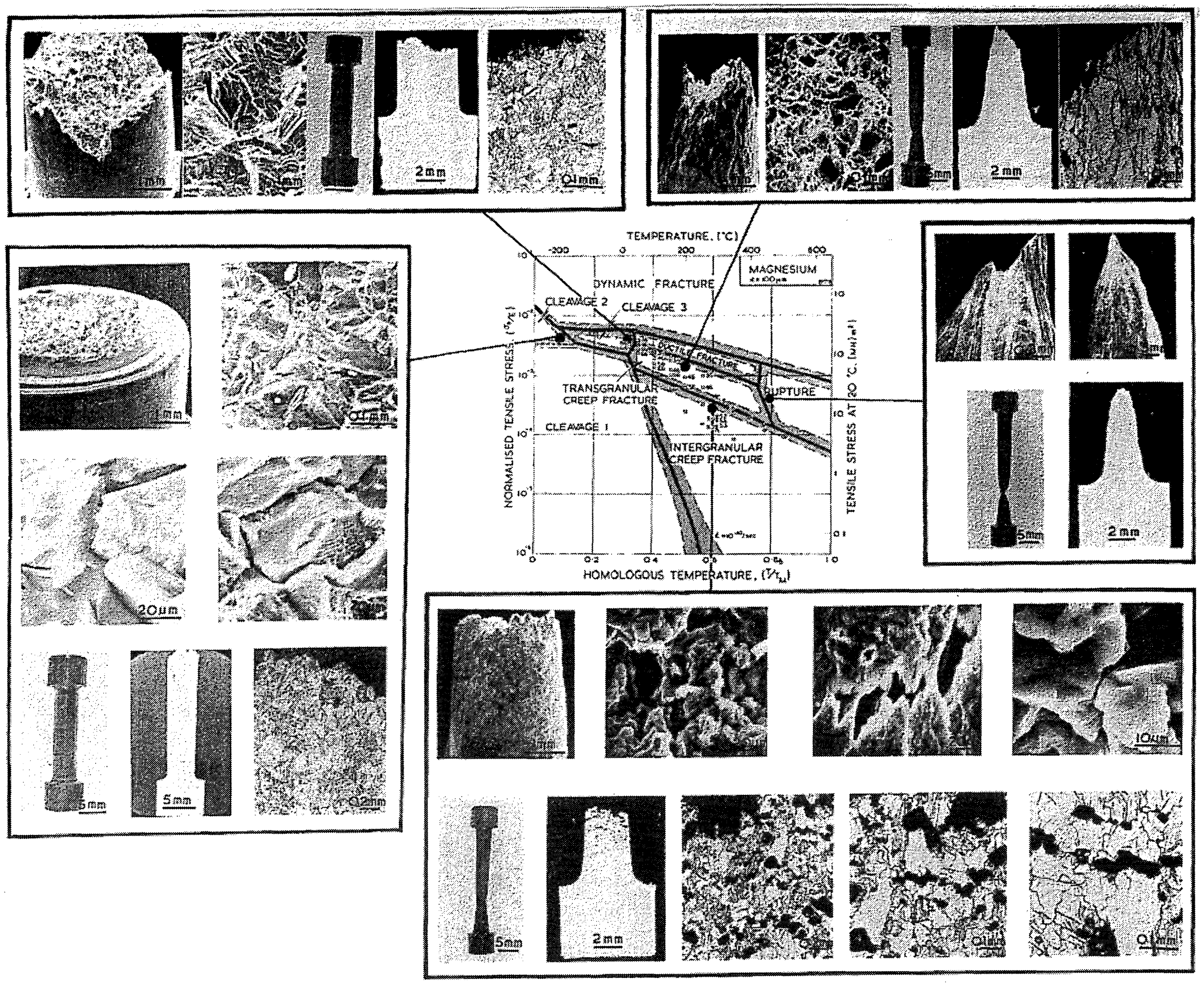

FIG. 3. Fracture mechanism map illustrating various fracture mechanisms available to solid Mg upon fracture. Central map shows mechanisms for different loads and temperatures, while surrounding figures show fracture surfaces and deformation of sample prior to fracture. After Gandhi and Ashby (1979).

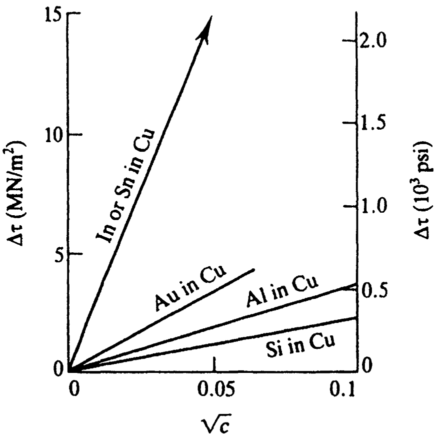

FIG. 4. Dependence of the yield stress of copper on impurity concentration. After Courtney (1990).

Given the stress tensor at a point, the force per unit area acting on a plane with normal **n** passing through that point is determined by

$$t_i^{(\mathbf{n})} = \sigma_{ij} n_j, \tag{5}$$

where $\mathbf{t}^{(\mathbf{n})}$ is known as the stress (or traction) vector. For more complicated states of stress, the off-diagonal components of the stress tensor determine the shear stresses on the planes orthogonal to the Cartesian directions. This can be seen by recourse to Eq. (5) if we consider a plane with normal in the z direction. In this case, σ_{xz} gives the component of the traction vector pointing along the x direction. Hence, these components determine the shearing forces.

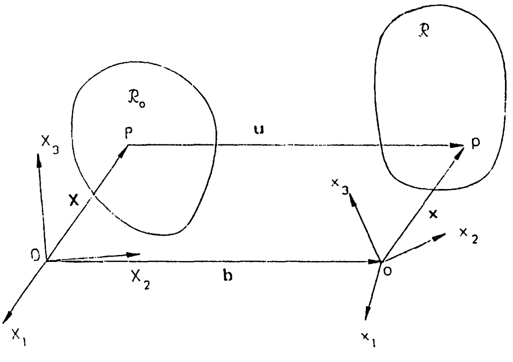

FIG. 5. Schematic of the mapping that carries a reference configuration with material particles labeled by **X** into the deformed configuration, with positions $\mathbf{x} = \mathbf{X} + \mathbf{u}(\mathbf{X})$. After Khan and Huang (1995).

The equilibrium of a solid body in the absence of any body forces follows from Newton's first law as

$$\sigma_{ij,j} = 0, \tag{6}$$

where the summation convention, which instructs for the summation over repeated indices, is implied. The symmetry of the stress tensor ($\sigma_{ij} = \sigma_{ji}$) follows from the absence of body moments. The point of our excursion into rudimentary continuum mechanics is that the notions of stress and strain not only have been found as the most powerful tool for describing the experimental basis of the mechanical properties of materials, but they provide the foundation upon which models of elastic and plastic deformation as well as fracture are built.

The plan of the remainder of the article is as follows. The main body of the article begins in Sec. 1 with a discussion of the elastic behavior of crystalline solids and an emphasis on linear elastic response. Section 2 follows by describing the notion of the ideal strength of crystalline solids. The ideal strength concept illustrates the importance of crystalline defects as the agents of mechanical behavior, with dislocations as the primary players in plastic deformation and cracks as the source of fracture at stresses below the ideal tensile strength. Sections 3 and 4 are the centerpiece of the article and consider permanent deformation in the form of plasticity, creep, fracture, and fatigue. Here it is shown how crystal lattice defects mediate the various mechanisms responsible for observed macroscopic mechanical behavior. Section 5 illustrates the subtleties introduced as a consequence of microstructural influences and demonstrates the refinements needed to understand the dependence of mechanical behavior on a material's thermal and mechanical history. At best, our treatment of this vast subject can only be cursory. In particular, the emphasis here is purely on crystalline solids, with metals serving as the basis of the majority of the discussion. The suggestions for further reading gives a list of references that are particularly enlightening.

1. ELASTIC RESPONSE OF MATERIALS

1.1 Phenomenology of Elastic Deformation

Prior to permanent deformation, crystalline solids typically undergo elastic deformation. Such deformation is reversible with the result that the undeformed state may be instantaneously recovered once any applied stresses are removed. The physics of such deformation may be neatly captured via a few simple material parameters known as the elastic moduli. That is, the elastic response of a material subjected to some arbitrarily complicated state of stress is ultimately determined by a set of numbers (up to 21 for crystals of low symmetry) known as the elastic constants. Experimentally, it is often found that there is a linear relation between these strains and the stresses needed to impose them.

The simplest sample for exploring the linear elastic behavior of materials is the tensile specimen. In this case, it is found that there is a linear relation between the dimensionless tensile elongation $\epsilon = (l - l_0)/l_0$ and the stress needed to produce it. This behavior is parameterized via a material constant known as the Young modulus, E, with the result that

$$\sigma_n = E\epsilon, \tag{7}$$

where σ_n is the normal stress. This is Hooke's law in its simplest form and accounts for the initial linear regime depicted in Fig. 1. Hooke's law is perhaps the simplest version of a material constitutive law, a necessary ingredient to supplement the kinematics and balance laws of conventional continuum formulations. Such constitutive laws serve to relate kinematic variables such as the strain or strain rate to the stresses. An accompanying effect of such tensile deformation is a lateral contraction of the sample perpendicular to the applied load that is parameterized via a second elastic material constant known as the Poisson ratio, ν. The strains in this simple case, then, are

$$\epsilon_{xx} = -\nu\epsilon_{zz} = -\frac{\nu\sigma_n}{E}$$

$$\epsilon_{yy} = -\nu\epsilon_{zz} = -\frac{\nu\sigma_n}{E}$$

$$\epsilon_{zz} = \frac{\sigma_n}{E},$$

where we have assumed the tensile axis is in the z direction. For isotropic materials, these two material parameters (E and ν) serve to characterize the elastic material response for loads as complicated as desired as long as the strains are sufficiently small. For larger elastic deformations, this description must be extended to the consideration of higher-order moduli. For many applications, the elastic regime is all that a material is ever hoped to sample—the bending of an airplane's wing while it is being bumped around in turbulence is intended to be reversible.

1.2 Generalized Hooke's Law

Thus far, we have ignored any intrinsic anisotropy induced by the underlying crystal structure of which the sample is composed. However, if we pull on a sample of zinc, the moduli describing both the elongation and the lateral contraction will depend strongly upon the crystal orientation. For example, if we are to pull along the c axis in a zinc crystal with a particular stress, it will lead to one tensile elongation. On the other hand, if we pull along one of the close-packed directions in the hexagonal basal plane with the same stress, the resulting elongation will be different. These anisotropy effects suggest a generalized form of Hooke's law that posits a relation between stress and strain of the form

$$\sigma_{ij} = C_{ijkl}\epsilon_{kl}, \tag{8}$$

where C_{ijkl} is known as the elastic modulus tensor, and the summation convention is in effect. The number of independent elastic moduli is determined by the symmetry of the crystal, with cubic crystals having three independent moduli and hexagonal crystals, five such moduli, and for isotropic solids, the modulus tensor reduces to the two independent material constants (Young's modulus and Poisson ratio) discussed above. For example, in cubic crystals like Al and Cu, the three independent moduli are denoted C_{1111}, C_{1122}, and C_{1212}. A more compact notation is often adopted in which one defines $C_{1111} = C_{11}$, $C_{1122} = C_{12}$, and $C_{1212} = C_{44}$. For the cubic crystal, the use of these constants leads to the conclusion that the stress needed to induce a 1% strain is the same for the (100) and (001) directions, but not for the (111) direction. In cubic crystals, a useful measure of the elastic anisotropy is given by

$$A = 2C_{44}/(C_{11} - C_{12}), \tag{9}$$

a quantity known as the anisotropy ratio. For an isotropic material, this constant is unity.

The observed range of elastic moduli for different materials spans some six orders of magnitude, with diamond at the pinnacle with a Young's modulus of a thousand gigapascals, typical structural materials such as steel coming in with moduli of a few hundred gigapascals, and rubbers and other polymers reaching values of less than 0.1 GPa. These numbers can be put in perspective by considering the stresses needed to induce a 1% strain in a few different materials. To put these stress measures into everyday terms, we first consider the pressure exerted on the floor by one's feet when standing. If we estimate a typical person's weight as 700 N and the contact area between their feet and the floor as 0.04 m^2, this corresponds to a stress of roughly 2×10^4 N/m^2 or 2×10^{-5} GPa. These numbers should be contrasted with the stresses of 10 GPa needed to induce a 1% strain in a cylindrical tensile specimen of diamond and even the stress of 10^{-3} GPa needed to strain a generic rubber specimen by 1%.

A second view of an elastic solid is provided by the notion of a stored energy function. By virtue of its elastic deformation, a solid stores strain energy, which can be recovered upon release of the deformation. The work done in bringing a solid to some state of strain is indicated schematically in Fig. 6. For a linear elastic solid, the strain energy density is given by

$$W(\epsilon) = \tfrac{1}{2}C_{ijkl}\epsilon_{ij}\epsilon_{kl}. \tag{10}$$

As a result, the total strain energy stored in the solid by virtue of its elastic strain is computed by integrating the strain energy density over the entire solid,

$$E_{\text{strain}} = \int W(\epsilon)\, dV. \tag{11}$$

The strain energy provides another window on the significance of the elastic moduli. For example, in the case of a material with cubic symmetry, the strain energy density is given by

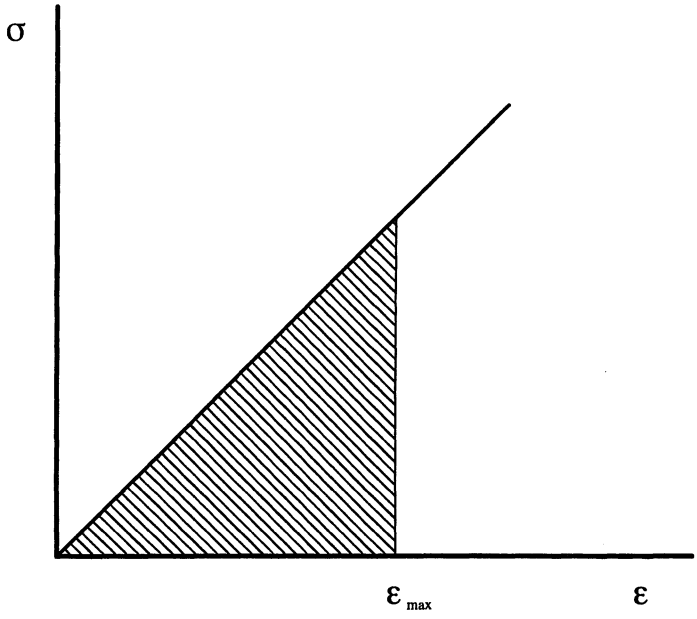

FIG. 6. Stress–strain curve illustrating energy stored in a solid by virtue of elastic deformation. Upon reaching a strain of $\epsilon_{\max}$, the strain energy is given by the area of the shaded region.

$$W = \tfrac{1}{2}C_{11}(\epsilon_{xx}^2 + \epsilon_{yy}^2 + \epsilon_{zz}^2) + C_{12}(\epsilon_{xx}\epsilon_{yy} + \epsilon_{xx}\epsilon_{zz} + \epsilon_{yy}\epsilon_{zz}) + 2C_{44}(\epsilon_{xy}^2 + \epsilon_{xz}^2 + \epsilon_{yz}^2). \quad (12)$$

Thus if we consider pure shear for which only the off-diagonal components of the strain tensor are nonzero, the energy cost of such deformations is governed by C_{44}. Alternatively, if we consider purely dilatational strains for which $\epsilon_{xx} = \epsilon_{yy} = \epsilon_{zz}$ and for which $\epsilon_{ij} = 0$ for $i \neq j$, then only C_{11} and C_{12} determine the energy cost. The linear elastic description of such strains breaks down once the strains become sufficiently large. In particular, the energy cost described above, which is quadratic in the strains, is no longer an appropriate description of the strain energy and requires the inclusion of nonlinear effects.

1.3 Microscopic Origins of Elastic Moduli

The use of the type of linear elastic models described above is intriguing for a number of reasons. One important feature of such models is that they replace the underlying lattice degrees of freedom with a continuum whose properties are specified by a few independent material parameters. However, ultimately the values adopted by these material parameters are a direct consequence of the underlying interactions between the atoms that make up the solid. Recent advances in the solution of the Schrödinger equation for electrons in a solid have made the first-principles calculation of material parameters such as the elastic moduli routine. As is shown below, the moduli can be interpreted as the curvature of the potential energy surface for the deformation of interest.

A more transparent microscopic model of the origins of the moduli in terms of bonds between atoms is afforded by considering a pair-potential description of cohesion. In this case, the total energy of the solid may be written as

$$E_{\text{tot}} = \tfrac{1}{2}\sum_{ij}^{N} V(R_{ij}), \quad (13)$$

where the sum is over all of the atoms in the solid and R_{ij} is the distance between the ith and jth atoms. The moduli are computed by seeking an analogy between the microscopic expression for the total energy and that arising from the continuum picture. For a crystal subjected to a homogeneous deformation described by the strain tensor $\boldsymbol{\epsilon}$, the microscopic expression for the total energy may be written as

$$E_{\text{tot}} = \tfrac{1}{2}\sum_{ij}^{N} V((\mathbf{I} + \boldsymbol{\epsilon})R_{ij}). \quad (14)$$

In this expression, $\boldsymbol{\epsilon}$ is the strain tensor and is a measure of the extent to which the crystal is deformed, and hence interatomic distances are altered, and $\mathbf{I}$ is the identity tensor. By expanding the energy to second order in the relevant strains and equating this expression for the energy to that arising from the continuum model, the moduli may be read off as

$$C_{ijkl} = \frac{1}{2\Omega}\sum_{n}\left(V_n'' - \frac{V_n'}{R^n}\right)R_i^n R_j^n R_k^n R_l^n/(R^n)^2. \quad (15)$$

In this expression, the sum is over the neighbors of the atom at the origin, R^n is the distance to the nth neighbor, R_i^n is the ith component of the vector to the nth neighbor, and Ω is the volume per atom in the crystal. The critical insight provided by this model calculation is the relation between the curvature of the interatomic potential (a microscopic quantity) and the elastic moduli (macroscopic observables). Though this model provides useful insights into the microscopic origins of the moduli, for a covalent material such as Si this approach will not capture the bond-bending mechanisms relevant to its moduli, and more sophisticated treatments are required.

1.4 Beyond the Elastic Regime

Earlier we noted that much of the permanent deformation that accounts for the broad variety of observed mechanical properties is mediated by extended defects such as dislocations and cracks. Interestingly, the primary conceptual basis for understanding such defects is the type of linear elastic models elucidated in the preceding section. As we discussed earlier, the phenomenological basis for the continuum description of phenomena such as elastic and plastic response is constitutive models. Thus, despite the fact that the presence of defects such as cracks and dislocations results in macroscopic behavior that cannot be described by linear elastic constitutive models (i.e., Hooke's law no longer applies), the fields surrounding the defects themselves are amenable to a linear elastic treatment.

In subsequent sections our understanding of crack-tip phenomena, for example, will be founded upon an elastic analysis of the crack problem that allows for a characterization of the crack-tip stress fields in terms of an $r^{-1/2}$ singularity, where r denotes the distance from the crack tip. Similarly, elastic analyses of dislocations are characterized by stress fields that decay with a generic r^{-1} singularity and logarithmic interaction energies.

Reference to the stress–strain curve exhibited in Fig. 1 makes it clear that for most materials, the elastic regime constitutes but a small part of the observed mechanical response. A foreshadowing of the qualitative changes that arise beyond the elastic regime is first noted at the so-called proportional limit, the point on the stress–strain curve where linearity is lost. The proportional limit is followed by plastic yield at a stress often referred to as the offset yield strength. The offset yield strength is a stress that will lead to some prespecified permanent strain upon unloading. Once the yield strength is reached, upon unloading there remain permanent strains, and it is the understanding of these permanent strains that occupy us in coming sections.

2. STRENGTH OF CRYSTALLINE SOLIDS

2.1 Ideal Tensile Strength

Early attempts to explain the permanent deformation of crystalline solids centered on the notion of the ideal strength. The ideal strength concept envisions failure as a homogeneous process in which all atoms across a given plane are displaced uniformly, either in shear or in tension. For example, failure in tension was thought to occur as indicated schematically in Fig. 7, with the result that once sufficiently large strains were reached, the atomic bonds could no longer resist the applied tension. This simple model of fracture allows for an atomistically based estimate of both the stress and strain to failure, an estimate found to be in gross disagreement with observed tensile strengths. The ideal strength calculation can overestimate the strength of a solid by as much as a factor of 10^3. This failure of solids to reach their ideal strengths served as an impetus for the development of defect-based models of permanent deformation.

One particularly transparent model of ideal tensile failure approximates the normal stress σ_n that develops in response to the separation of adjacent planes by an opening displacement Δ in excess of their equilibrium spacing as a sinusoidal function

$$\sigma_n = A\sin(2\pi\Delta/a), \tag{16}$$

where a is a length that sets the scale over which cohesion between adjacent planes occurs. The prefactor A is determined by insist-

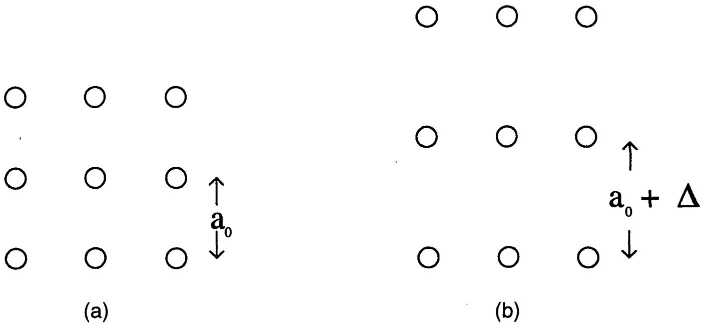

FIG. 7. Fracture mechanism envisioned in calculation of ideal tensile strength. (a) Perfect crystal. (b) Crystal subjected to homogeneous elongation.

ing that in the limit of small opening displacements, the normal stress is linear in the opening displacement, with the proportionality constant given by Young's modulus, E. In particular, it is found that the prefactor A is given by $E/2\pi$. In turn, this model implies that the maximum tensile stress that can be supported by the crystal is $E/2\pi$. While providing a physically simple model of the fracture process, the predictions implied by this model are generally found to be at odds with measured values of both the stress and strain to failure. Our simple model of the normal stresses that develop in response to the normal strains preceding fracture may be challenged as overly facile with the hope that a more realistic law based upon a correct treatment of the atomistic underpinnings of this process would reduce the estimate of the ideal strength. However, such refinements do not suffice to bring this simplified law into correspondence with the experimental values, except perhaps in the case of whiskers that have stress–strain behavior like that shown in Fig. 8, where it is seen that stresses much closer to the ideal stress may be reached. In particular, for Cu with a Young's modulus slightly larger than 100 GPa, the whisker reaches a value of roughly $E/50$. The difference between whiskers and the tensile samples considered until now is that whiskers are nearly defect free, thus lacking the stress-concentration features thought to operate in more macroscopic samples.

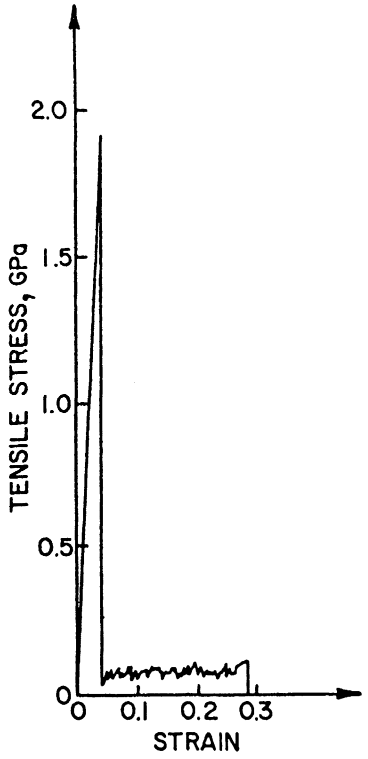

FIG. 8. Stress–strain curve for Cu whisker. Note the stress scale, which is in units of GPa, typical of ideal yield strength, rather than MPa units typical of conventional yield. After Boyer (1987).

2.2 Ideal Shear Strength

Analogous arguments have been advanced for evaluating the strength of solids in shear. In this case, plastic deformation is treated as arising from the shearing of adjacent crystal lattice planes. As in the model of the ideal tensile strength elucidated above, a relation between the shear stresses and slip of adjacent planes is assumed, the simplest of which takes the form

$$\tau = \tau_{\max} \sin(2\pi\delta/a). \tag{17}$$

The prefactor in the stress-displacement relation is again determined by insisting upon a fit to the linear elastic value, which, upon following an argument like that given above, yields the result $\tau_{\max} = \mu/2\pi$, where μ is the shear modulus for the plane of interest. This result provides an estimate of the ideal shear strength of crystalline materials. As might have been expected in light of our experience with the ideal tensile strength, the ideal shear strength is an overestimate of the observed values by as much as a factor of 10^5. The wide discrepancy between the observed and ideal shear strengths of crystalline solids can be gleaned from Fig. 9, which reports the yield stresses observed in wide classes of materials. The scale associated with these strengths is in the megapascal range, while the shear modulus of typical materials is in the gigapascal range. It is noted that the ideal strength is often orders of magnitude higher than the observed strength, with structural materials such as metals only reaching as much as 10^{-5} of their ideal values. Clearly, these simple models of mechanical yielding are missing something crucial.

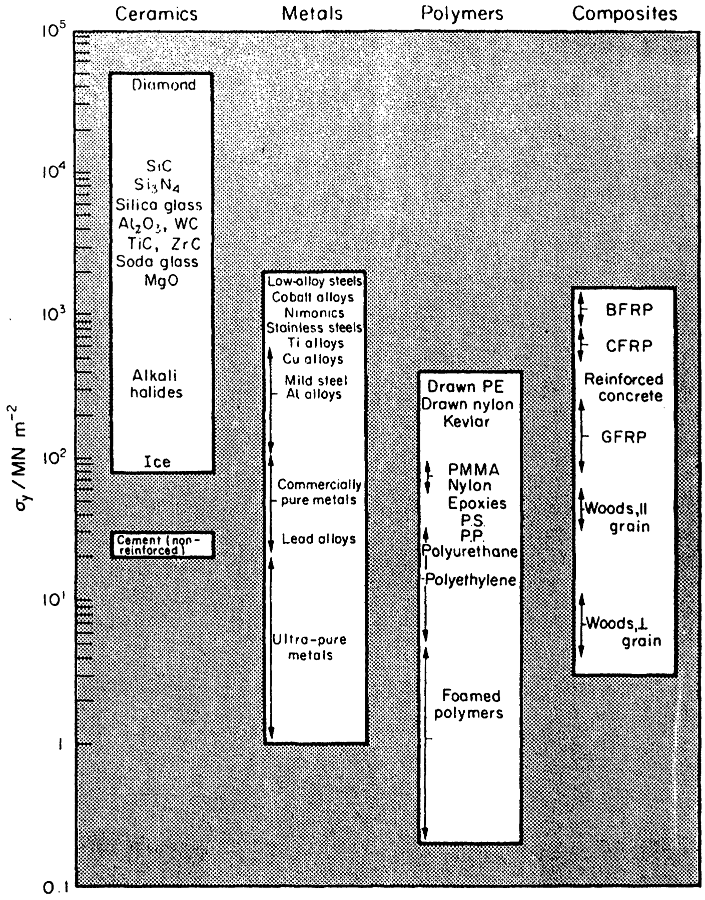

FIG. 9. Illustration depicting range of observed yield stresses for wide classes of materials. Note that the vast majority of materials yield at stresses orders of magnitude lower than estimates based on ideal strength. After Ashby and Jones (1980).

2.3 Defect-Mediated Deformation

The paradox left in the wake of the calculation of both the ideal shear and tensile strengths led to some of the most powerful theoretical insights concerning the mechanical behavior of solids. In particular, these failed models led to the conviction that extended defects such as dislocations and cracks are the carriers of permanent deformation in materials. As will be shown subsequently, cracks can be thought of as stress concentrators, which can in the vicinity of the crack tip ramp up the local stresses to values as high as the ideal tensile strength, while the remote loads remain much lower. It is this ability of cracks to magnify applied stresses locally that resolves the ideal-strength paradox. Similarly, the presence of dislocations reduces the stress needed to initiate plastic deformation. In this case, the presence of these line defects allows for a net translation of one half of a crystal with respect to the other at stresses well below those demanded by the ideal shear-strength calculation. In large measure, then, the interesting deformation of materials is primarily a consequence of the presence of extended defects, most notably dislocations and cracks.

The following sections focus on two central issues in the mechanical behavior of materials, namely, plasticity and fracture. Our treatment of plasticity first considers plastic flow as a rate-independent process that results once a certain critical stress is attained. Subsequently, we consider time-dependent plastic deformation in the form of creep with an emphasis on the mass-transport and dislocation-based mechanisms that make time-dependent permanent deformation possible, even though the stresses are lower than the yield stress. Fracture is then considered both from the standpoint of the stress concentration concept that holds that fracture occurs because of a magnification of stresses due to the presence of flaws such as cracks, and from the perspective of the energetic arguments first introduced by Griffith that assert that fracture is the result of a competition between the release of elastic energy and the creation of new surface. Fatigue is treated briefly, illustrating that fracture can occur (like plastic flow before) at stresses lower than the critical stress when the material is subjected to cyclic loading.

3. PERMANENT DEFORMATION OF MATERIALS

3.1 Phenomenology of Permanent Deformation

Some of the primary criteria for the selection of materials to be used in applications where the material is subjected to mechanical loads center on the competition between a few key material properties, namely, yield strength, weight, and toughness. The yield strength is a measure of the difficulty to induce plastic deformation. For example, among metals, steel is a high–yield-strength material, whereas lead is not. On the other hand, toughness refers to a material's resistance to fracture. Interestingly, much of the emphasis on the permanent deformation of materials focuses on the detrimental consequences of such deformation. On the other hand, deformation processing has existed for as long as man has used metals for the construction of various objects. The use of tungsten filaments in light bulbs relies on the ability to form these metals to extreme plastic strains. Thus, while plastic deformation can have detrimental consequences, in some cases the failure of a material to exhibit sufficient plastic deformation can mean the difference between success and failure in its processing.

3.2 Time-Independent Plastic Deformation

Plastic deformation refers to the irreversible shape change of materials that have been subjected to stresses in excess of their yield strength. Figure 10 shows the stress–strain curve of a generic tensile specimen loaded into the permanent deformation regime. As long as the stress is less than σ_y, upon release of the load the material will traverse the same curve upon unloading that it did upon loading, with the result that the initially undeformed state is recovered. On the other hand, if the sample is loaded to

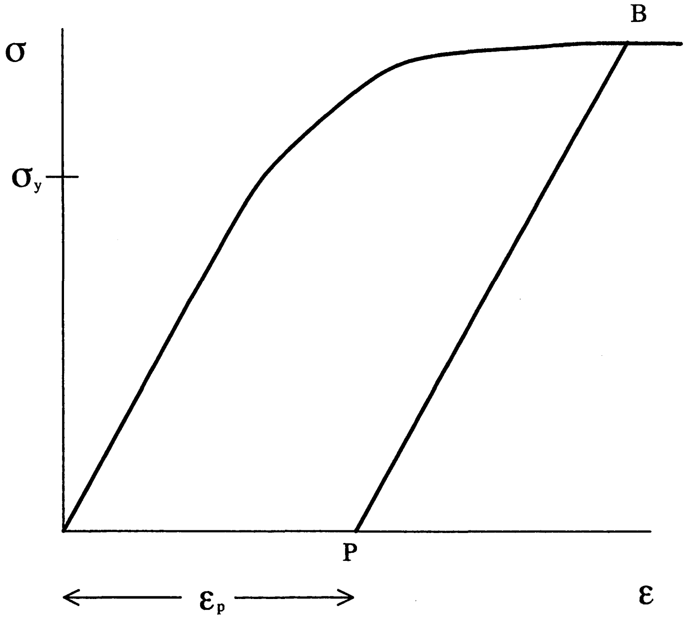

FIG. 10. Stress–strain curve illustrating the characteristics of permanent deformation. Plastic strain ϵ_p is strain that remains after load has been removed. Hardening refers to the increase in flow stress needed to maintain continued plastic deformation once yield has been attained.

point B in the figure, and the loads are then released, the unloading curve will correspond to that labeled BP in the figure. The plastic strain is defined as the strain that remains after unloading and is labeled ϵ_p in the figure. Further, if the sample is now loaded again, to a first approximation, it will load elastically up the curve PB with a new and larger yield stress than was found in the original experiment. The material has hardened, a consequence of the interactions between dislocations.

The simplest microscopic picture of plastic deformation centers upon the motion of dislocations. One immediate task we face is to demonstrate how dislocations can be the carriers of macroscopic plastic strains. In Fig. 11, a schematic of a dislocated crystal shows how a dislocation can be thought of as a one-dimensional boundary between slipped and unslipped regions of the crystal. The planes along which dislocations move are known as slip planes. The unit of slip carried by a particular dislocation is known as the Burgers vector, **b**. In fcc metals, slip typically occurs on {111} planes with Burgers vectors along the ⟨110⟩-type direction. hcp crystals allow for slip on basal planes and the perpendicular pyramidal planes, while in bcc metals, slip typically occurs on {110}- and {112}-type planes with Burgers vector along the ⟨111⟩-type directions.

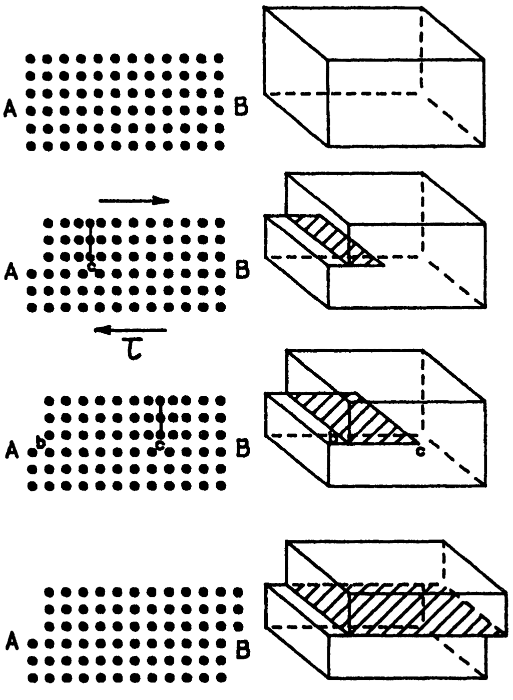

FIG. 11. Dislocated solid, illustrating the way in which dislocation is a boundary between slipped and unslipped regions of crystal. After Khan and Huang (1995).

An argument for the claim that dislocations are the agents of plastic change can be made by imagining the dislocation to move all the way to the opposite boundary of the crystal, as shown in Fig. 11. In this simple two-dimensional model, we note that the average plastic strain carried by such dislocations is given by

$$\Delta\epsilon_p = nb/h, \tag{18}$$

where b is the magnitude of the Burgers vector of the relevant dislocations, h is the height of the region of interest, and n is the number of dislocations. If the dislocations travel with a mean velocity $\bar{v}$, then the time to traverse the distance L is given by

$$\Delta t = L/\bar{v}. \tag{19}$$

In the limit that the number of dislocations is sufficiently large to allow for this average description, the net plastic strain rate can be approximated as

$$\dot{\epsilon}_p \approx \Delta\epsilon_p/\Delta t = \rho b\bar{v}, \tag{20}$$

where $\rho = n/hL$ is the density of mobile dislocations. This result is known as Orowan's equation and is an example of a constitutive law that links a macroscopic quantity (the plastic strain rate) to microscopic quantities (the average dislocation velocity and the density of mobile dislocations). In treating the dislocation density as time independent, we have ignored the effects of dislocation sources, which can enhance the plastic strain rate.

What is the evidence that a dislocation-based mechanism, while plausible, is indeed responsible for plastic deformation in metals? One piece of evidence is that of well-defined slip traces that can be observed on the surfaces of deformed materials. Such slip traces are the points of exit of dislocations at the surface of a crystal, just as indicated schematically in Fig. 11, though for traces to be observed optically, large numbers of dislocations must have exited the crystal. A second crucial outcome of the dislocation model of plastic deformation is that the presence of dislocations must resolve the paradox associated with the failure of materials to reach their ideal shear strength. By virtue of the fact that dislocations respond to an applied stress, it is possible to perform a model calculation (Hirth and Lothe, 1982) of the stress needed to move a dislocation (this is the so-called Peierls stress), which yields a value for the critical shear stress for dislocation motion of

$$\tau_{\text{Peierls}} = \frac{2\mu}{(1-\nu)} \exp\left(\frac{-4\pi\zeta}{b}\right), \tag{21}$$

where ζ is the width of the dislocation and is the length scale over which the slip goes from zero to a full Burgers vector; typical widths range from one to ten lattice spacings. This expression should be contrasted with the ideal shear strength; the Peierls stress is reduced by an exponential factor relative to the ideal shear strength.

A more immediate treatment of the coupling of dislocations to stress fields is provided by the elastic theory of dislocations. This analysis is important not only because it yields insights into the stress needed to induce plastic deformation but also because it is the basis of understanding various hardening mechanisms. For example, solid-solution hardening is a reflection of the interaction between dislocations and the stress fields induced by point defects. Similarly, strain hardening arises because dislocations interact with each other with the result that as the density of dislocations on various competing slip systems increases, dislocation–dislocation interactions inhibit further dislocation motion. In general, for a dislocation with line direction given by the vector $\boldsymbol{\xi}$ and with Burgers vector $\mathbf{b}$ the force on that dislocation per unit length due to a stress $\boldsymbol{\sigma}$ is given by

$$\mathbf{F} = -\boldsymbol{\xi} \times (\boldsymbol{\sigma}\mathbf{b}), \tag{22}$$

a result known as the Peach–Koehler force. As seen below, the Peach–Koehler force presides over phenomena as diverse as conventional plastic yielding and strain hardening arising from dislocation–dislocation interactions.

A critical testing ground for the dislocation-based model of plastic deformation is that of flow in single crystals. It is imagined that the sample is loaded in tension, as shown in Fig. 12. The pioneering observations of Taylor and Schmid (Honeycombe, 1968), among others, led to a number of key insights concerning plastic flow in single crystals. One such insight is the observation that the stress to induce plastic flow is dependent upon the orientation of the underlying crystal lattice with respect to the loading axis. Careful analysis led Schmid to elucidate the law that bears his name, namely, that the stress to induce plastic flow is given by the component of the shear stress in the slip plane that is directed along the slip direction, the so-called resolved shear stress. Plastic flow will initiate when this stress reaches a critical value known as the critical resolved shear stress. For a slip plane with normal $\mathbf{n}$ and slip direction $\mathbf{s}$ as shown in Fig. 12, the resolved shear stress is given by

$$\tau_r = \mathbf{s}\cdot\boldsymbol{\sigma}\mathbf{n}. \tag{23}$$

In terms of the angles ϕ and λ defined in the figure, this leads to the equivalent expression

$$\tau_r = \sigma_0 \cos\phi \cos\lambda, \tag{24}$$

where the geometric term is known as the

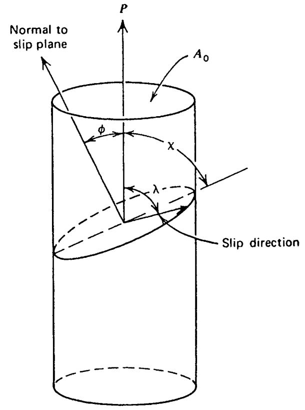

FIG. 12. Cylindrical test specimen that exhibits geometric origins of Schmid factor. Tensile loading occurs along cylinder axis while slip plane is oriented in a direction characterized by angle ϕ, and slip direction is characterized by angle λ. After Hertzberg (1976).

Schmid factor and σ_0 is the remote tensile load (P/A_0 in the figure). Hence there is a characteristic geometric signature for plastic flow in single crystals. In particular, depending upon the orientation of the tensile axis with respect to the crystal axes, the critical resolved shear stress will be reached at different tensile loads as determined by the Schmid factor and shown in Fig. 13. Though the dislocation picture of plastic deformation provides a mechanistic underpinning for such deformation, the phenomenology of permanent deformation via plastic flow requires the motion of dislocations in such large numbers that for many purposes it is preferable to invoke an average description in terms of macroscopic plasticity considerations. In these models, the dislocations are never considered explicitly. Rather, the emphasis is on determining the macroscopic criteria for plastic flow.

We now revisit the stress–strain curve of Fig. 10 with an eye to a phenomenological description of plastic flow. The simplest model of plastic flow holds that a solid undergoes strictly elastic deformation until a critical stress (known as the yield stress) is reached. Of course, so far our analysis has ignored the tensor character of the stress state since we have only considered tensile specimens. In principle, there are six components of the stress tensor that can be specified independently. These more general stress states lead to plastic flow as well. The locus of all points in stress space that result in plastic flow is known as the yield surface. To refine our insights into yield it is necessary to revisit the idea of the stress tensor in order to characterize the conditions under which yield will occur. The yield criterion is most transparent if we choose a coordinate system in which the stress tensor is diagonal. This coordinate system is characterized by the fact that the traction and normal vectors are parallel. This condition can be stated quantitatively in the form of an eigenvalue problem,

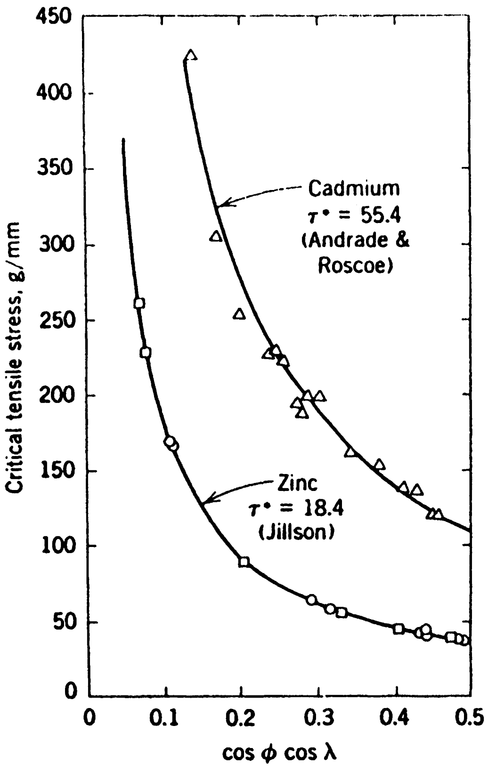

FIG. 13. Experimental data showing correlation between critical tensile load and Schmid factor. After Gilman (1969).

$$\mathbf{t}^{(\mathbf{n})} = \boldsymbol{\sigma}\mathbf{n} = \lambda\mathbf{n}, \qquad (25)$$

where $\mathbf{n}$ is known as a principal axis of the stress tensor and λ is a principal stress. A yield condition for plastic flow can be constructed on the basis of the insight that the maximum shear stresses occur on (110) planes with respect to these principal axes, with the value of such shear stresses given by

$$\tau_{\text{shear}} = \tfrac{1}{2}(\sigma_i - \sigma_j), \qquad (26)$$

where σ_i and σ_j are the ith and jth principal stresses. The Tresca yield condition holds that plastic yield will occur when the difference between the maximum and minimum principal stresses reaches some critical value. This criterion may be written

$$\tau_{\text{critical}} = \tfrac{1}{2}(\sigma_{\max} - \sigma_{\min}) = k, \qquad (27)$$

where k is the critical shear stress for plastic flow. Though this yield condition is consistent with the lessons of the Schmid law since it emphasizes the role of critical shear stresses in inducing plastic flow, it is important to note that this yield criterion is isotropic. In particular, it does not acknowledge the discreteness of the crystal lattice that allows for slip only on certain specific slip planes. Within this criterion, by rotating the state of stress continuously, the plane on which the Tresca condition envisions slip to occur can be rotated continuously as well. On the other hand, isotropic yield is a reasonable representation for polycrystals with their slip systems oriented over a range of directions, and it is this class of behavior that such yield criteria are intended to describe. A second yield condition, known as the von Mises yield criterion, argues that plastic flow will proceed when an averaged version of the shear stresses reaches a critical value. In particular, if the *deviatoric stress* is defined as

$$S'_{ij} = \sigma_{ij} - \delta_{ij}\tfrac{1}{3}tr(\boldsymbol{\sigma}), \tag{28}$$

which amounts to subtracting off the hydrostatic part of the stress, then the *second invariant of the deviatoric stress* is defined as

$$J_2 = \tfrac{1}{2}S'_{ij}S'_{ij}. \tag{29}$$

The von Mises yield criterion contends that plastic flow will initiate when

$$J_2 = k^2. \tag{30}$$

Note that for both of the yield criteria described above, under a state of hydrostatic stress in which $\sigma_1 = \sigma_2 = \sigma_3$, there will be no plastic flow, in keeping with the observations of Schmid and others. Experimental results contrasting these two yield criteria are shown in Fig. 14. These experiments were performed on thin-walled tubes subjected to both torque and tension and show that the initiation of plastic flow occurs for states of stress that are bounded by the Tresca and von Mises values.

Models such as those we have described above are known as models of perfectly plastic solids. These models hold that below some critical stress the deformation will be strictly elastic and that once the critical stress is reached, plastic deformation will proceed. The highly idealized models of plastic flow make no provision for hardening. On the other hand, as illustrated in Fig. 10, plastic deformation is characterized by *hardening* in which it is necessary to increase the stress in order to continue plastic deformation. This hardening is a reflection of the interaction between dislocations, which makes it increasingly difficult to move such dislocations in the presence of the stress fields induced by neighboring dislocations. This topic is taken up in more detail in Sec. 5.

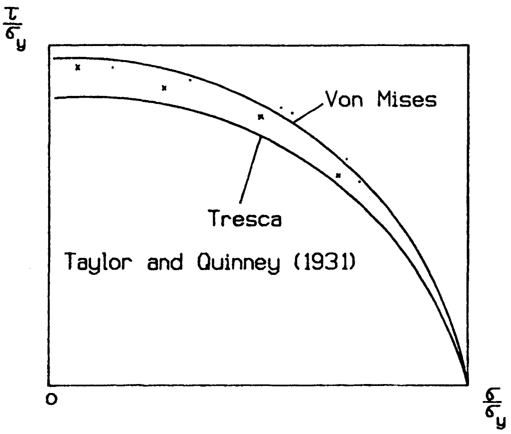

FIG. 14. Experimental comparison of yield stress for different values of shear and normal stress components to those predicted by von Mises and Tresca criteria. After Khan and Huang (1995).

3.3 Creep Phenomena

Our description of permanent deformation thus far implies that for stresses lower than that of the yield strength, a material's deformation will be strictly elastic. This is an idealization of observed behavior. However, once temperatures reach some fraction of the melting temperature, materials exhibit permanent deformation even for stresses below the yield stress. Such processes are generally denoted as creep phenomena, though, as will be seen in the case of fracture, a variety of creep mechanisms are operative there as well.

Besides being of fundamental interest, creep mechanisms are of particular concern in high-temperature material applications, since they pose threats to performance over the life of a component. The two fundamental creep tests that set ideas for these phenomena are those carried out at constant load and those occurring at constant strain. In constant-load creep tests, the sample is loaded, and the subsequent strain is monitored as a function of time, leading to strain vs time profiles of the type shown in Fig. 15.

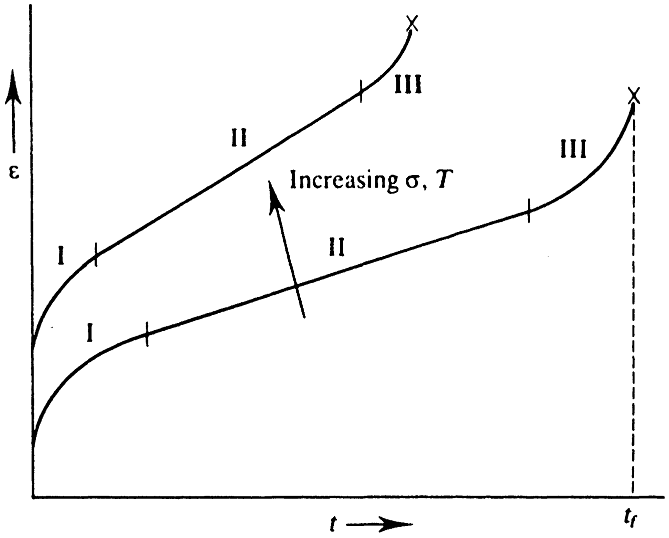

FIG. 15. Generic creep curve exhibiting transient regime (I) followed by steady-state creep regime (II) and tertiary creep regime (III). After Courtney (1990).

Alternatively, the sample can be loaded to a fixed strain with the resulting stress relaxation monitored as a function of time. The underlying mechanisms are defect based and can result from thermally activated dislocation motion, stress-assisted mass-transport mechanisms, and motion of or along grain boundaries.

It is conventional in the analysis of creep curves such as that shown in Fig. 15 to divide the response into three regimes (I, II, and III) known respectively as primary (or transient), secondary, and tertiary creep. The primary regime is characterized by a decreasing creep rate, while the secondary regime exhibits steady-state behavior, as is evidenced in the figure. The third stage reveals an increasing strain rate and precedes failure of the material. Experiments have suggested a number of empirical relations between the strain rate, the applied stress, and the temperature, a typical example of which is

$$\dot{\epsilon} = \dot{\epsilon}_0 \sigma^n \exp(-Q/RT), \qquad (31)$$

where Q is the activation energy for creep and n is a phenomenological parameter relating the strain rate to the stress. An interesting correlation between the creep activation energy and that for lattice self-diffusion may be noted, leading to a conjectured relation between creep and mass transport via diffusion. One of the challenges posed by the type of expression just discussed is the construction of microscopically based models that result in such behavior. The two types of thermally activated and stress-assisted creep mechanisms considered here are dislocation-based models and diffusion-based models.

The key idea behind the diffusion-based creep mechanisms is that the presence of a stress field can bias mass transport, leading to a net shape change. Perhaps the simplest such picture is that of Nabarro–Herring creep. Here it is imagined that a sample is loaded in tension in one direction and in compression in an orthogonal direction. Further, it is imagined that the vacancies that mediate bulk diffusion have as their primary source the crystal surface or the boundaries surrounding a grain of size d. The presence of the applied stress state inhibits vacancy formation on the crystal faces in compression while it enhances the vacancy concentration on the faces in tension, leading to a net concentration gradient. As a result of this concentration gradient, a diffusive flux will be set up between these faces, having the net effect of transporting matter to the faces in tension. A simple computation of this diffusive flux leads to the Nabarro–Herring creep rate given by

$$\dot{\epsilon} = A(D_L/d^2)(\sigma\Omega/kT), \qquad (32)$$

where D_L is the self-diffusion coefficient for *bulk* diffusion, Ω is the atomic volume, and d

is the average grain size over which such diffusion occurs. A mechanistic variant of this creep mechanism known as Coble creep considers the dominant diffusion to take place along the grain boundaries themselves rather than through the bulk as postulated in the Nabarro–Herring case. In this case, the resulting strain rate is given by

$$\dot{\epsilon} = A(D_{GB}\delta/d^3)(\sigma\Omega/kT), \tag{33}$$

where here the relevant diffusion coefficient is that for grain boundary diffusion (D_{GB}) and δ refers to the grain-boundary thickness.

Reflection on the deformation mechanism map shown in Fig. 2 reveals that in addition to the Nabarro–Herring and Coble creep regimes, there is an additional region of parameter space that is dominated by so-called power-law creep. In this case, it is thought that it is the diffusion-assisted motion of dislocations that carries the plastic deformation. In contrast with the linear stress dependence found for both the Nabarro–Herring and Coble mechanisms, the strain rate in this regime is characterized by a power-law stress dependence ($\dot{\epsilon} \propto \sigma^n$). A mechanistic picture of the processes operating in this regime centers on the idea that pinned dislocation segments are able to surpass the barriers that pin them via a thermally activated climb mechanism. Dislocation climb refers to the motion of a dislocation perpendicular to its slip plane via mass transport, either to or away from the dislocation core.

4. FRACTURE

4.1 Phenomenology of Fracture

As indicated in Fig. 1, at sufficiently large stresses, solids fail via fracture. Taken literally, fracture refers to the breaking of interatomic bonds across some entire cross section of the solid. Though all fractures share certain generic features, an elucidation of fracture at the microscopic level reveals a host of processes that can ultimately result in catastrophic failure.

Fracture that is not accompanied by some prior plastic deformation is known as brittle fracture. Alternatively, fracture can be preceded by massive plastic flow, a phenomenon known as ductile fracture. The variety of fracture mechanisms are captured in the fracture mechanism map shown in Fig. 3. Here it is seen that, depending upon the temperature and the loading, the generic tensile specimen can exhibit a variety of interesting behaviors. Though broadly speaking the division into ductile and brittle fracture is convenient, a more precise characterization of the fracture process must allow for a number of different processes within each of these classes of failure. These mechanisms are summarized pictorially in Fig. 16, which gives an interpretation of the dominant fracture mode in various regimes. Among other things, this figure makes it clear that microstructural influences, particularly grain boundaries, are crucial in determining the dominant fracture mode.

Brittle fracture can occur via either a transgranular or an intergranular mechanism. Intergranular failure points to the important role played by the types of microstructural details normally ignored in a preliminary analysis. Here, the grain boundaries serve as the mechanical weak link, opening up the attendant possibility of toughening mechanisms that center on grain-boundary chemistry. In the ductile regime, it is seen that again a number of different mechanisms are available, with necking and progressive thinning of the cross section as in pulling taffy as the most extreme variant of ductile failure. At elevated stresses, ductile fracture arises from the growth and coalescence of voids under the influence of the concentrated stresses ahead of the crack tip. However, even at loads much lower than the putative critical stress for fracture, creep fracture becomes possible. This type of failure depends upon the operation of mass transport mechanisms that are biased by the presence of the applied and crack-tip stresses. The wide range of mechanisms displayed in the fracture mechanism map makes it clear that a detailed understanding of fracture is a challenging task. On the other hand, certain generic features of fracture can be enumerated by considering both the energetics associated with crack propagation and the stress enhancements caused by the presence of the crack. It is this route that we follow here.

4.2 Stress Concentration

As noted earlier, the conceptual basis for the resolution of the paradox posed by the failure of materials in general to reach their

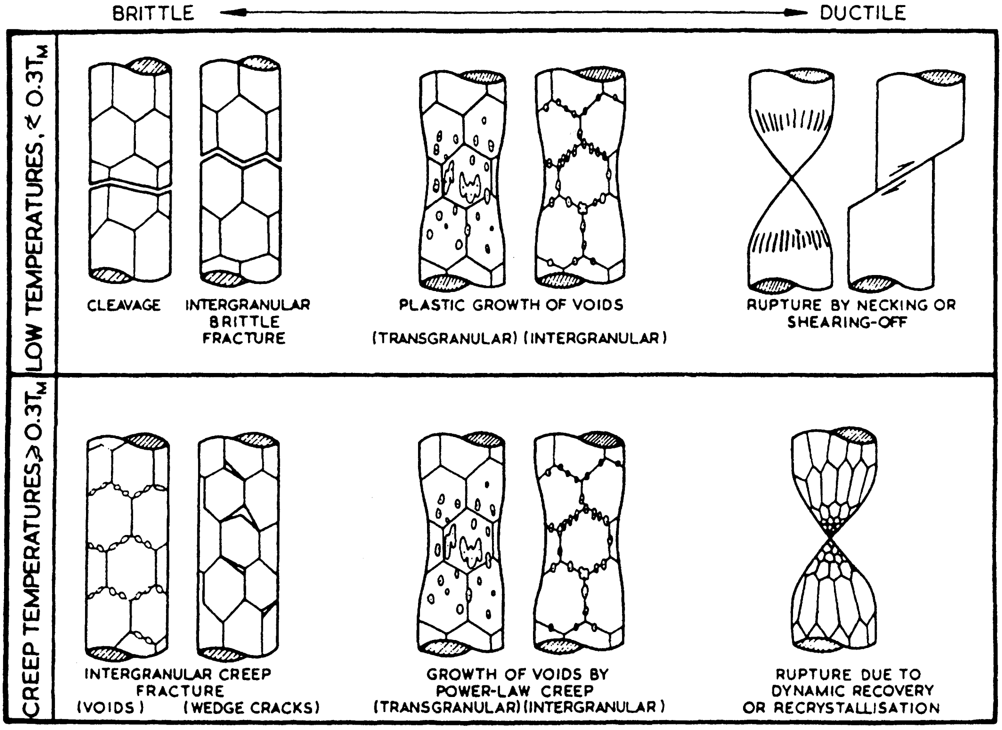
BRITTLE
DUCTILE
LOW TEMPERATURES, ≮ 0.3T_M
CLEAVAGE
INTERGRANULAR BRITTLE FRACTURE
PLASTIC GROWTH OF VOIDS
(TRANSGRANULAR) (INTERGRANULAR)
RUPTURE BY NECKING OR SHEARING-OFF
CREEP TEMPERATURES, ≥ 0.3T_M
INTERGRANULAR CREEP FRACTURE
(VOIDS) (WEDGE CRACKS)
GROWTH OF VOIDS BY POWER-LAW CREEP
(TRANSGRANULAR) (INTERGRANULAR)
RUPTURE DUE TO DYNAMIC RECOVERY OR RECRYSTALLISATION

FIG. 16. Pictorial representation of key observed fracture mechanisms for different temperature regimes. After Ashby *et al.* (1979).

ideal tensile strength is the notion of stress enhancement due to the presence of flaws such as cracks, steps, and notches. What is the experimental basis for considering cracks as the agents of fracture? Advances in optical techniques now allow for the direct observation of moving cracks, as is shown in Fig. 17. The idea is to use the nonuniform deformation fields in the near crack-tip region to produce fringe patterns in real time. For the case shown in the figure, the crack is found to travel at somewhat less than half of the shear-wave speed. Post mortem analysis of fracture surfaces also plays a key role in elucidating the underlying fracture mechanism. In Fig. 18, the fracture surfaces from a ductile fracture arising from void coalescence and a brittle intergranular fracture are contrasted. These probes, and many others, provide definitive insights into the role of cracks as the agents of failure via fracture.

With the insight that cracks are the key players in mediating fracture, it is important to systematize the way in which such defects concentrate stresses well in excess of the applied loads. One particularly convenient avenue for examining this is to think of cracks as the singular limit of elliptical holes. A linear elastic analysis of the elliptical hole was first provided by Inglis, who demonstrated that for a geometry such as that shown in Fig. 19, the stress along the continuation of the semimajor axis of the ellipse is given as shown in the figure. This analysis is suggestive in that as the radius of curvature goes to zero (the ellipse becomes increasingly eccentric) the expression for maximum normal stress σ_{yy} that occurs at the termination of the semimajor axis may be expanded to give

$$\sigma_{\max} = \sigma_0(1 + 2\sqrt{a/\rho}). \qquad (34)$$

For a given elliptical hole, the stress concentration factor may be defined in terms of the remote load σ_0 as the ratio $\sigma_{\max}/\sigma_0$, which reveals the way in which such geometrical inhomogeneities can magnify stress. Indeed, even a circular hole results in a ramping up of the remote stress by a factor of 3.

The sharp-crack limit can be thought of

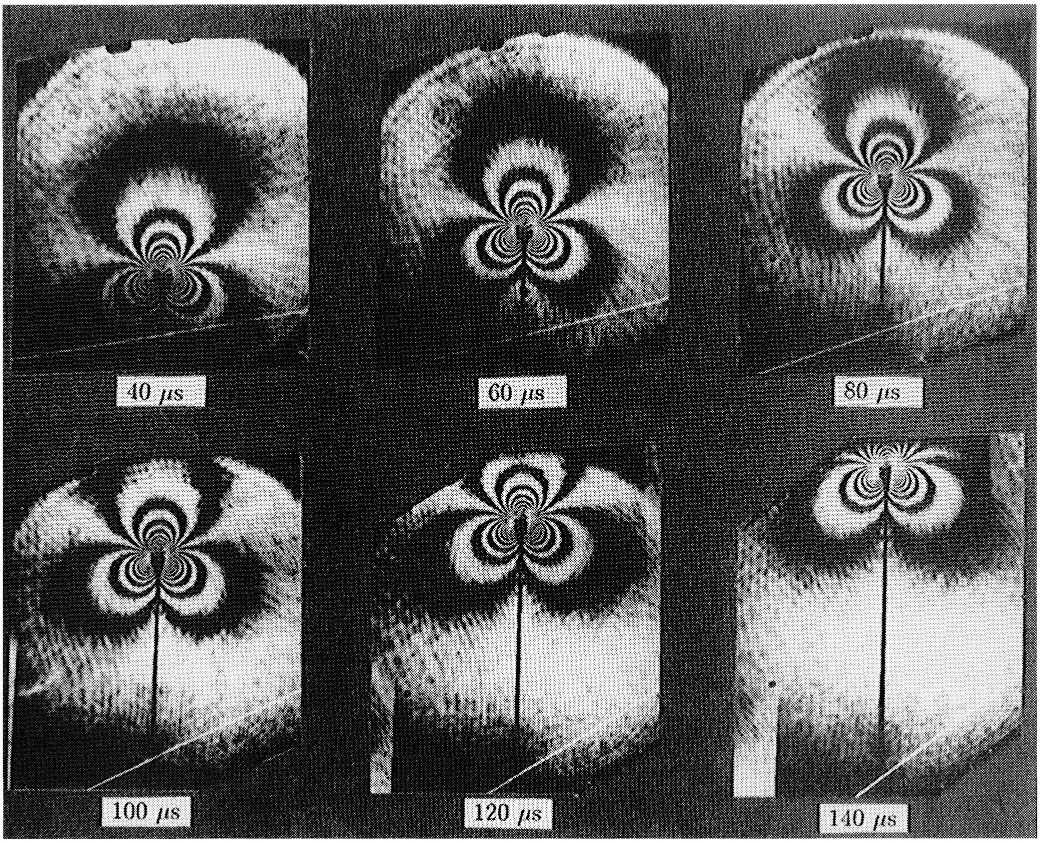

FIG. 17. Time evolution of crack tip observed experimentally. Courtesy of A. Rosakis.

in the context of the elliptical hole discussed above, or it can be solved explicitly. Because the underlying theoretical framework in our analysis is linear elasticity, superposition allows for the consideration of three independent modes of loading that "span" the space of possible loading conditions of a sharp crack, as illustrated in Fig. 20. Recall that the equilibrium equations may be written as $\sigma_{ij,j} = 0$. A solution to the boundary-value problem of the sharp crack in the absence of body forces yields a solution for the nonvanishing components of the stress tensor under mode I loading of the form

$$\sigma_{ij} = \frac{K_I}{\sqrt{2\pi r}} f_{ij}(\theta), \tag{35}$$

where $f_{ij}(\theta)$ is an angular factor [for example, $f_{\theta\theta} = \cos^3(\theta/2)$], K_I is known as the mode I stress intensity factor, and r is the distance from the crack tip.

One of the more beautiful results of the theory of linear elastic fracture mechanics is the way in which this singular solution for the crack problem points the way to a material parameter that characterizes the resistance of a given material to crack propagation. In particular, by performing experiments on specimens with known geometry and with a preexisting flaw, the stress to fracture can be measured. This experiment then defines the critical stress intensity K_{IC}, which serves as a material parameter known as the fracture toughness. Once the fracture toughness is known, the critical load for different geometries can be determined.

4.3 The Griffith Concept

The preceding discussion has highlighted the way in which stresses can be locally enhanced by the presence of stress-concentrating inhomogeneities. A complementary view-

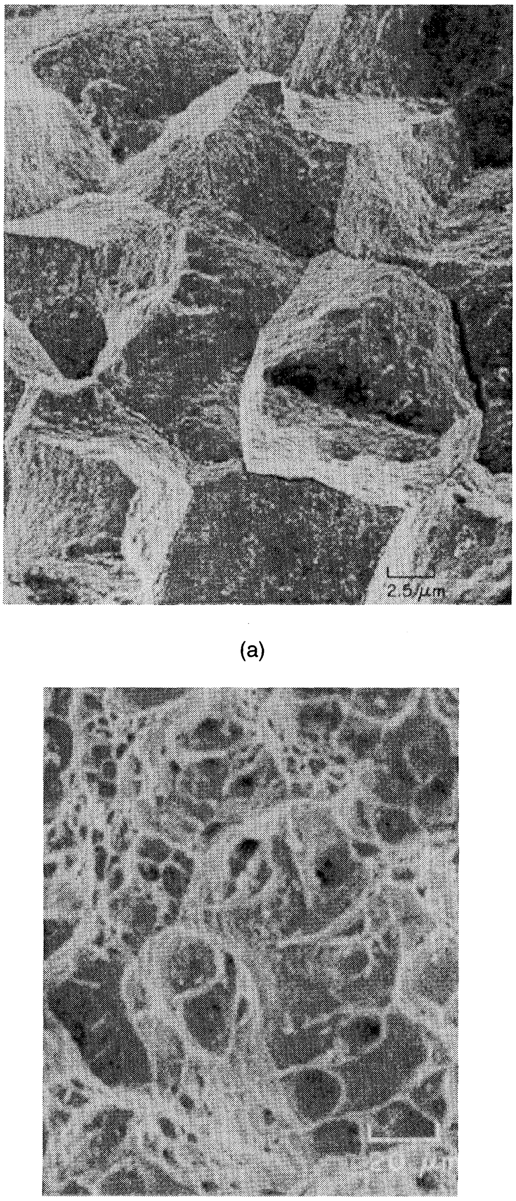

FIG. 18. Fracture surfaces arising from (a) brittle intergranular fracture and (b) ductile fracture via void coalescence. The intergranular fracture surface clearly exhibits facets from grain structure, while ductile fracture surface shows dimpled features. After Boyer (1974).

point, pioneered by Griffith (1920), evaluates the energetics of crack extension via much the same type of competition between energies familiar from nucleation theory. In particular, the Griffith energy-balance concept argues that the driving force for crack extension is the result of a competition between the elastic energy released upon crack extension and the increasing surface energy attending such extension. If we consider a centered crack of length $2a$ (the sharp limit of the elliptical hole discussed above), on the basis of the elastic stress fields derived above, it can be shown (Lawn, 1993) that the elastic energy release upon extending the crack by amount da is given by

$$dU_{\text{elastic}} = -[\pi\sigma^2 a(1 - \nu^2)/E]\,da, \tag{36}$$

where σ is the applied load, a is the crack half length, E is the Young modulus, and ν is the Poisson ratio. The negative sign indicates that crack extension leads to a reduction in the elastic energy of the system. On the other hand, the energy cost associated with such crack extension in the form of surface energy is

$$dU_{\text{surface}} = 2\gamma\,da, \tag{37}$$

where γ is the surface energy. If we seek the critical condition for which the driving force vanishes, we find that the critical load for crack extension is given by

$$\sigma_c = \sqrt{\frac{2\gamma E}{\pi a(1 - \nu^2)}}. \tag{38}$$

This condition is known as the Griffith criterion and implies a fracture toughness of

$$K_{IC} = \sqrt{2\gamma E/(1 - \nu^2)}. \tag{39}$$

For a typical surface energy of 1 J/m^2 and a typical Young's modulus of 100 GPa, this implies a fracture toughness of roughly 0.5 MPa m$^{1/2}$. It is interesting to contrast the Griffith value for the fracture toughness with typical observed values. Only the most brittle materials such as glasses and concrete, for example, have fracture toughnesses as low as those computed from the simple Griffith criterion. On the other hand, the fracture toughness exhibited by materials such as me-

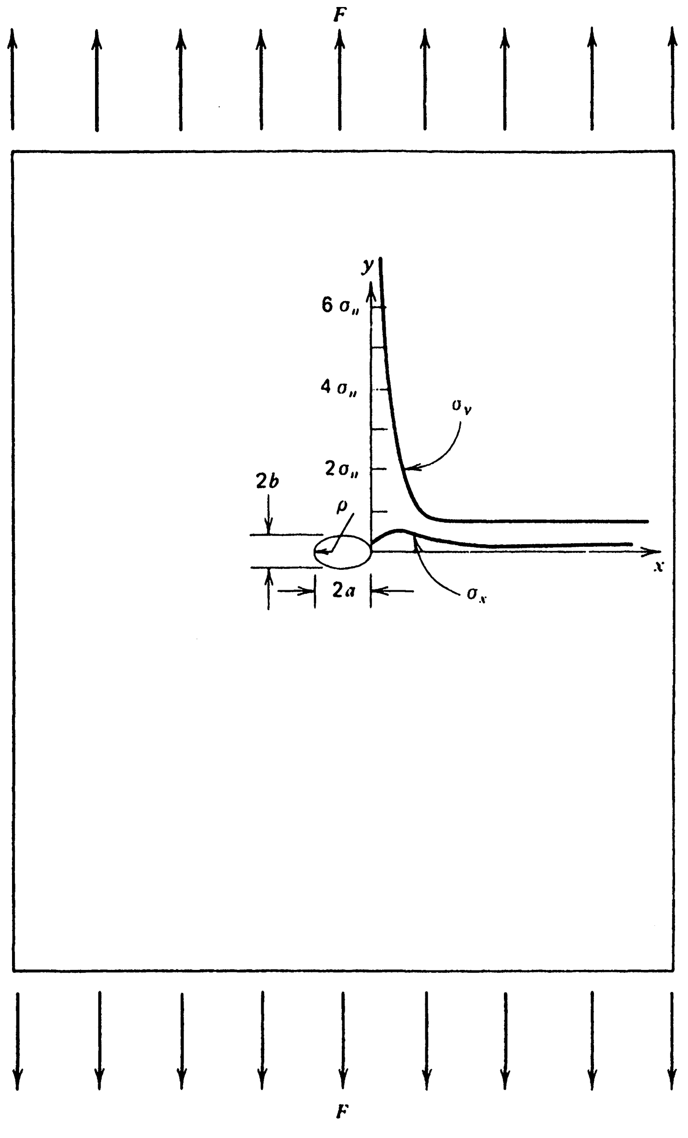

FIG. 19. Geometry considered by Inglis in evaluating stress concentration due to an elliptical hole. Figure shows enhancement of normal stresses (σ_y) relative to remote values (σ_n). Note that notation in this figure is different than in remainder of article. After Collins (1993).

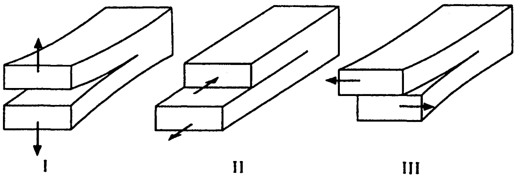

FIG. 20. Three modes of loading that characterize crack-tip response. In mode I, loading is perpendicular to crack faces, while modes II and III involve shears. After Lawn (1993).

tallic alloys easily exceed 10 MPa $m^{1/2}$ and can climb in excess of 100 MPa $m^{1/2}$.

Evidently, the model of fracture toughness described above misses a key feature of most real fracture processes. Irwin (1948) and Orowan (1955) realized that the simple nucleation-theory–type argument formulated above must be augmented by an additional term representing the energy that goes into plastic deformation in the vicinity of the crack tip. At the mechanistic level, one mechanism that engenders such deformation is the emission of dislocations from the crack itself as well as the activation of dislocation sources in the vicinity of the crack tip. We may estimate the size of the region over which plastic deformation is important by recourse to the linear elastic crack-tip stress fields described earlier. The plastic zone is defined as that region over which the near-tip stresses have induced plastic deformation. The size of the plastic zone may be estimated by considering the region over which the elastic stresses have reached the yield strength of the material. That is,

$$\sigma_y = K_I/\sqrt{2\pi r_p}, \tag{40}$$

where r_p is the plastic zone size and is given by

$$r_p = K_I^2/2\pi\sigma_y^2. \tag{41}$$

In its simplest form, the plastic work can be thought of as leading to an effective surface energy of the form

$$\gamma_{\text{eff}} = \gamma_s + \gamma_p, \tag{42}$$

where γ_s is the surface energy and γ_p is the contribution to the effective surface energy arising from the plastic work. Use of this effective surface energy in Eq. (39) rather than its bare value demonstrates qualitatively the way in which plastic dissipation raises the fracture toughness.

A final interesting point concerning the fracture toughness is the notion of the brittle-to-ductile transition. As the temperature of many materials is raised above some threshold value, the fracture mechanism can pass from that of brittle fracture to ductile fracture over a narrow temperature range. Accompanying this change is a significant increase in the material's toughness. An example of this behavior is shown in Fig. 21.

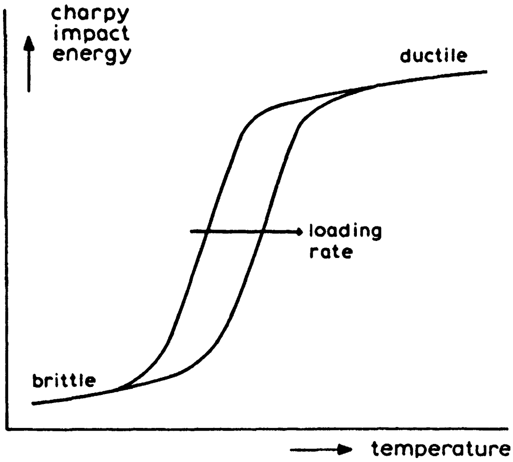

FIG. 21. Brittle–ductile transition as evidenced by energy absorbed in fracture test. High temperatures require more energy to achieve fracture. After Broek (1991).

4.4 Fatigue

A more insidious form of failure is fracture via fatigue. In this case, the material can fail at stresses lower than the stress required for conventional fracture as a result of being subjected to cyclic loading. Hence, even if the service loads have been selected so as to be much lower than the critical stress for fracture, cracks may first be initiated and then slowly grow in response to a cyclic stress field. Many headline structural disasters are ultimately traced to fatigue failure. The detective work of determining the mechanism by which some material failed is aided by consideration of the morphology of the fracture surface. As noted in the previous section, different fracture mechanisms can lead to widely different morphologies in the resulting fracture surfaces. In the case of fatigue failure, the fracture surface reveals one part that reflects the slow–crack-growth regime of fatigue failure and a second part that is a result of the catastrophic fast crack growth that preceded ultimate failure.

One of the major successes in the study of fatigue is the understanding of fatigue crack growth. Here the point is that although the stress amplitude falls short of the stress needed for failure under static loading, subcritical cracks can grow as a result of the cyclic variation in the stresses. Earlier it was shown that the stress intensity factor serves as a governing parameter that presides over fracture processes under constant loads. Remarkably, the variation in the stress intensity has been determined to be a governing parameter in characterizing the slow crack growth found in fatigue. In particular, the rate of crack extension per cycle is given by

$$\frac{da}{dN} = C(\Delta K)^n, \tag{43}$$

where C and n are material constants and $\Delta K = K_{\max} - K_{\min}$ is the difference between the maximum and minimum stress intensities. An example of this dependence is illustrated in Fig. 22.

This section of the article has, for the most part, treated the various deformation mechanisms as distinct phenomena. However, crack-tip phenomena, for example, should really be seen as the site of many of the mechanisms described above taking place concurrently. Stress concentrations at the crack tip can lead to dislocation nucleation at the tip and activation of sources in the tip neighborhood, emphasizing the coupling between plasticity and fracture.

5. INFLUENCE OF MATERIAL PROPERTIES

So far, our treatment has emphasized the mechanical features common to wide classes of materials without noting the ways in which details such as grain size, chemical composition, and heat treatment dictate these properties. However, from the standpoint of the engineering of materials for specific uses, it is just these details that are manipulated, sometimes with dramatic effect, in order to obtain some desired property. A key tool in the hands of the materials scientist is the ability to control microstructures. Often, such control is carried out via programs of both deformation and annealing that yield variations in the resulting mechanical properties. For example, schoolchildren used to be taught of the bronze age in which the strengthening of copper by additions of tin made possible a new generation of materials. These developments were a consequence of the use of solid-solution hardening to alter dramatically the mechanical properties of a particular class of metals. Chemical effects are important not only because impurities impede the motion of dislocations but also because suitable tuning of the chemistry at grain boundaries can have dramatic consequences as well, resulting in either embrittlement or enhanced ductility.

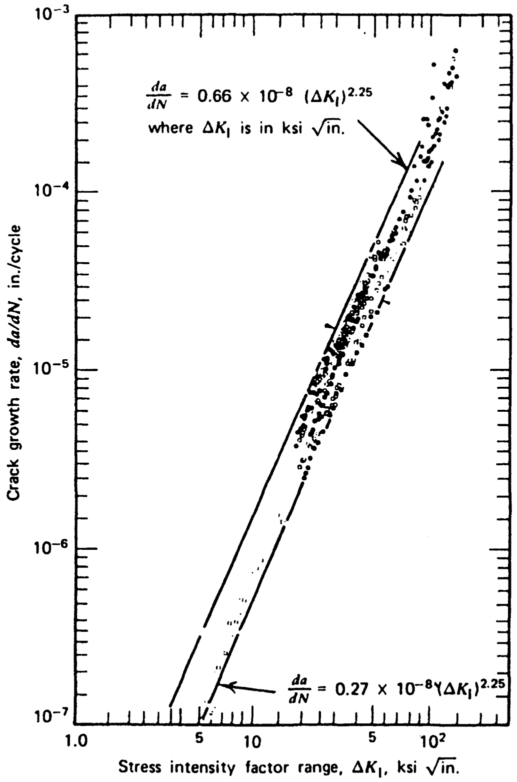

FIG. 22. Crack growth rate as a function of stress intensity difference for steels. Stresses measured in units of kilopounds/in^2 (ksi). After Fuchs and Stephens (1980).

An additional way in which mechanical behavior can be altered beyond that of chemical effects is via programs of heat treatment, which, among other things, have the effect of altering the grain size. It is the purpose of this section to highlight the ways in which the materials designer can tune the mechanical properties of a given material. Our treatment has emphasized the ways in which materials can be altered intentionally. However, anyone who has broken rusted metal knows that nature alters materials as well. Consideration of these effects requires an analysis of the ways in which exposure to various environmental factors can lead to embrittlement and is outside of the scope of this article.

5.1 Grain-Size Effects

Most engineering materials are polycrystalline. That is to say that the microstructure of such materials is built up of a collection of grains, each oriented differently, and bounded by grain boundaries. Interestingly, the geometric character of the grains is surprisingly well described to first approximation in terms of a single parameter—the grain size. This implies that under typical circumstances, the microstructure is composed of a number of grains of similar size. Some examples of microstructures with different grain sizes are shown in Fig. 23. Hall (1951) and Petch (1953) found that the yield strength depends upon this mean grain size according to the simple law

$$\sigma_y = \sigma_0 + k_y d^{-1/2}, \tag{44}$$

where d is the grain size and k_y is a material-specific parameter. This relation implies that larger grain sizes tend to reduce the yield strength of a particular material. An example of this effect is shown in Fig. 24.

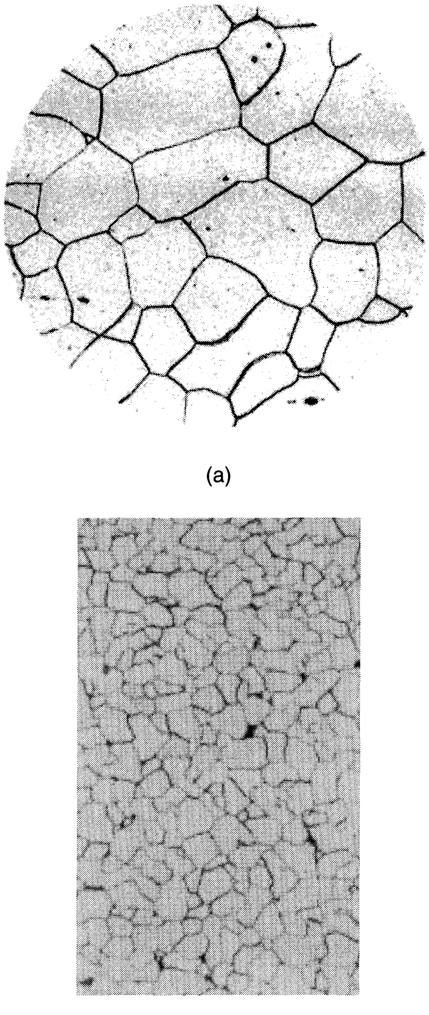

FIG. 23. Typical microstructures as found in steel. Figures show two different microstructures at same magnification. After Lyman (1972).

As with many of the empirical laws that characterize such mechanical behavior, there are qualitative explanations that serve to indicate the underlying physics of these phenomena. In the case of the Hall–Petch relation described above, the key point is thought to be the role of grain boundaries in inhibiting dislocation transmission from grain to grain. Dislocations are often found to glide until they reach a boundary, which then impedes their motion. As further dislocations accumulate on a given slip plane in the vicinity of the boundary, a dislocation pileup results, which has a stress field of its own. The presence of the dislocation pileup in a given grain enhances the applied stresses that are acting on dislocations in adjacent grains. Macroscopic plastic deformation requires slip to occur in multiple grains, not just those optimally oriented for slip. The stress enhancement due to the pileup in a favorably oriented grain assists the slip in adjacent, less favorably oriented grains. The dependence on grain size enters through the fact that the strength of the dislocation pileup is determined by the grain size, with larger pileups leading to larger stress concentrations in adjacent grains and hence lower yield stresses.

5.2 Work Hardening

The yield stress does not only depend on the grain size but also can be determined by the extent to which a material has been previously deformed. Examination of Fig. 10 reveals that the onset of plastic flow is followed by a regime in which it requires ever higher stresses to maintain plastic deformation. This effect is known as work (or strain) hardening and is attributed to the internal stresses that arise as a result of the dislocation–dislocation interaction. Recall our discussion of the Peach–Koehler force in Sec. 3.2, in which it was shown that a dislocation has an interaction energy with the stress fields that surround it. One source of such stress fields is other dislocations. In this case, if we imagine two parallel edge dislocations, both taken along the x axis for convenience, the x component of the Peach–Koehler force on the dislocation at the origin is given by

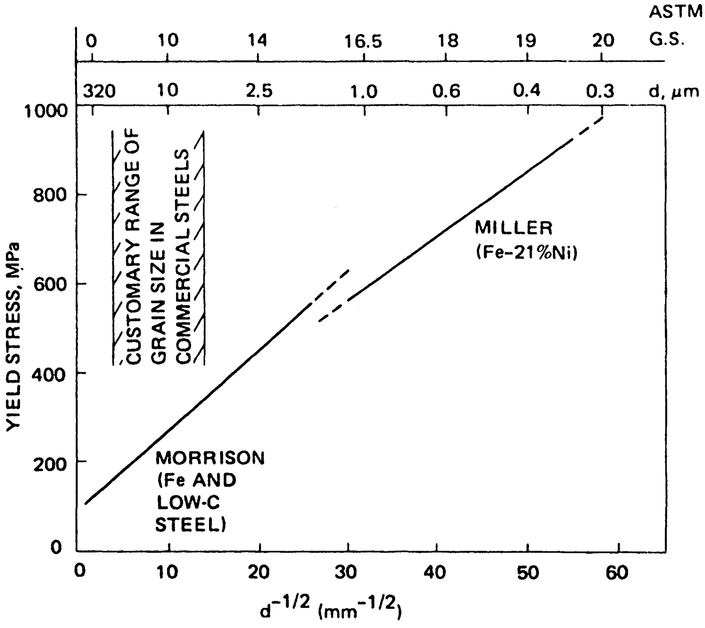

FIG. 24. Relation of yield strength to grain size in steels illustrating the Hall–Petch grain size dependence. After Leslie (1981).

$$F = -\mu b^2/2\pi(1 - \nu)d, \qquad (45)$$

where d is the separation between the two dislocations and μ is the shear modulus. The important insight afforded by this result is that the presence of neighboring dislocations can impede the motion of the dislocation of interest.

Taylor formulated a simple but enlightening model of strain hardening that argues that the increase in the flow stress as a consequence of work hardening should scale as $\rho^{1/2}$, where ρ is the dislocation density. Taylor's argument considers an ordered array of edge dislocations with uniform spacing a and interplanar distance d. By appealing to the type of interaction force between neighboring dislocations given above, but by summing the contributions due to all the neighboring dislocations, it can be shown that the critical resolved shear stress to move dislocations scales with the $\rho^{1/2}$ law mentioned.

Though the arguments advanced by Taylor are pleasing, they cannot account for the various stages of hardening revealed in Fig. 25. The Taylor model concentrates on the interactions between dislocations associated with the same slip system. However, once a particular slip system is sufficiently hardened, slip can commence on alternative slip systems. Stage 2 hardening reflects this interaction between dislocations associated with different slip systems. Detailed understanding of the various stages of hardening requires careful analysis of the underlying dislocation processes and resulting dislocation–dislocation interactions.

5.3 Solid-Solution Hardening

As alluded to above, another mechanism leading to an increased yield stress is solid-solution hardening, in which the presence of foreign impurities in the form of point defects can increase the stress needed to move dislocations. As seen in Fig. 4, there is a characteristic scaling of the yield stress with the impurity concentration. However, despite a number of competing models for this behavior, quantitative understanding of this effect remains elusive. One of the difficulties in constructing models of hardening is the role of a number of simultaneous competing mechanisms—mechanisms such as grain-size effects, an initial density of dislocations leading to strain hardening, and the pinning of dislocations by impurities or second-phase particles. Though it is convenient to assume that these affects conspire to alter the flow stress in an additive way, such an assumption may not be warranted.

A simple qualitative model for this hard-

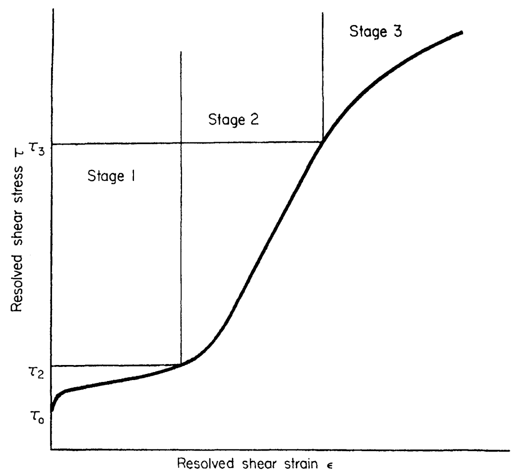

FIG. 25. Schematic of various stages of hardening observed in an fcc metal. After Honeycombe (1968).

ening may be constructed as follows. First, it is necessary to account explicitly for the way in which the stress field of the point defects interacts with the dislocation. One way in which the point defect alters the behavior of the dislocation is through the volume change accompanying the point defect, which is coupled to the hydrostatic component of the stress produced by the dislocation at the position of the point defect. If we consider the geometry depicted in Fig. 26, the hydrostatic pressure at the site of the point defect is given by

$$p(r,\theta) = \frac{(1 + \nu)}{(1 - \nu)} \frac{\mu b \sin\theta}{3\pi r}, \tag{46}$$

where μ is the shear modulus and, as earlier, ν is the Poisson ratio. For the coupling between the volume change and the hydrostatic stress considered above, there results an interaction energy of the form

$$E_{\text{int}} = A \sin\theta/r, \tag{47}$$

where the constant A depends upon the elastic properties of the medium, the Burgers vector of the dislocation, and the volume change resulting from the presence of the point defect. Differentiation of this expression with respect to the dislocation coordinate x along the slip plane results in an interaction force of the form

$$F = -2Kxd/(x^2 + d^2)^2, \tag{48}$$

where d is the perpendicular distance between the slip plane and the point defect. The interaction force between the point defect and the dislocation is shown in Fig. 27. Note that this result is strictly built around a linear elastic description of the fields due to both the dislocation and the point defect. This implies that for close contact distances between the point defect and the dislocation, this description will break down.

Once the interaction energy between the dislocation and the point defect has been determined, the next step in pushing the model toward a quantitative insight is to obtain an

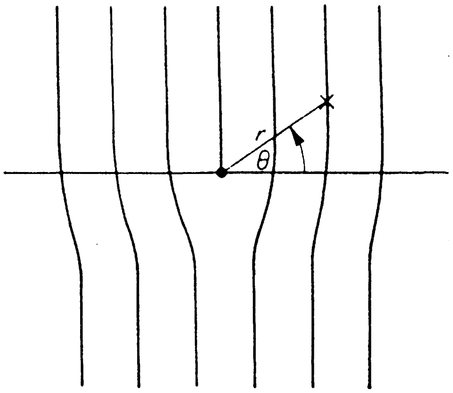

FIG. 26. Schematic of interaction between point defect at (r,θ) and edge dislocation at the origin. After Honeycombe (1968).

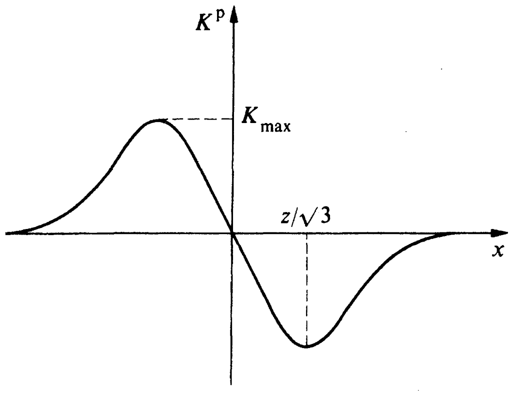

FIG. 27. Force exerted on dislocation due to presence of point defect. After Haasen (1978).

estimate of how a random distribution of such impurities superposes to produce some mean increase in the yield stress. Note that such models presuppose that the distribution of the impurities themselves is not affected by the presence of the dislocation itself, an assumption that breaks down if temperatures are high enough to allow for solute diffusion. In this regime, it is also necessary to consider the breaking away of dislocations from defect "atmospheres."

Theories designed to treat the very dilute limit as well as the higher-concentration regime lead to different predictions as to how the yield strength will scale with impurity concentration. In the Fleischer model (Fleischer, 1963), the impurities are sufficiently dilute so that it can be imagined that the operative process is the breaking away from single pinning points with a force equal to the maximum strength of these impurities. This model leads to a scaling of the yield strength of the form $\sigma_y \propto c^{1/2}$. In the Labusch model (Labusch, 1970), on the other hand, the dislocation is thought to interact with pinning centers of a number of strengths that are spread along the dislocation line with some continuous density. Here the scaling of the yield strength occurs as $\sigma_y \propto c^{2/3}$. On the other hand, in some cases it is experimentally difficult to settle the question of how the yield stress scales with impurity concentration. The key point remains, however, that the point-defect–dislocation interaction is a crucial part of the hardening that is known to occur in real alloy systems.

5.4 Second-Phase Particles

Under appropriate conditions, a particular material can be littered with particles of a second phase, which can significantly alter the mechanical properties of that material. A schematic example of the types of variations in mechanical properties that result from annealing programs is shown in Fig. 28, where the properties of Al–Cu alloys are plotted as a function of aging time. The varying mechanical properties reflect a concomitant change in the disposition of the Cu, which is initially quenched in the Al at concentrations in excess of those allowed by the equilibrium phase diagram at low temperatures. If the alloy is aged at intermediate temperatures, diffusion allows the Cu atoms to cluster into small particles that hinder dislocation motion.

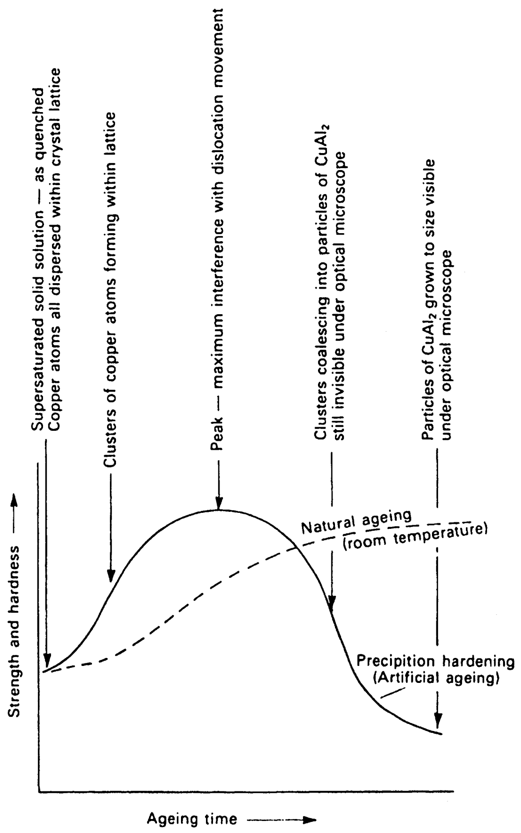

FIG. 28. Diagram illustrating the role of aging in altering the mechanical properties in the Al–Cu system. After Bradbury (1985).

From the standpoint of mechanical properties, the interaction with dislocations is one of the features that makes second-phase particles important. The action of such particles on dislocations has been postulated to lead to two primary mechanisms. In the one extreme, the particles do not allow for the passing of an impinging dislocation. In this case, the dislocation segment between two such particles will bow out until the adjacent bowing segments intersect and pinch off. This mechanism is known as the Orowan process. Alternatively, it is possible that the dislocation can pass through the particle via a cutting mechanism. In this case, the second-phase particles transmit the dislocation with the result that they too suffer slip. These two processes are contrasted in Fig. 29.

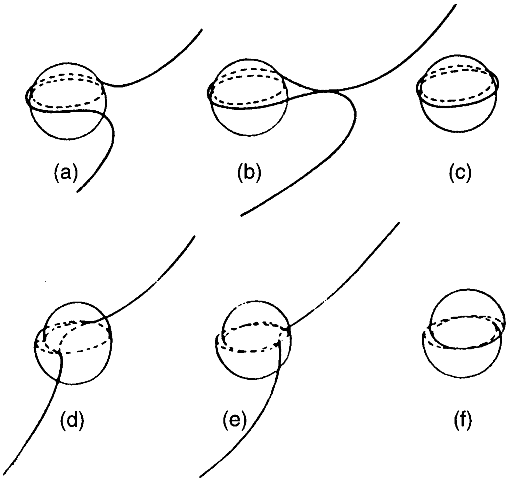

FIG. 29. Illustration of two dislocation mechanisms when second-phase particles are encountered. (a)–(c) Case in which dislocation bows out around particle; (d)–(f) particle cutting mechanism. After Gerold (1979).

Second-phase particles not only are important in dictating the strengthening of some materials but can play a critical role in governing a material's fracture behavior as well. One example of this role is provided by steels, in which it is found that second-phase particles such as carbides and sulfides serve as the sites at which voids are initiated via a mechanism of interfacial debonding (see STEEL). The spacing of these particles, then, in part determines the rate at which these voids coalesce leading in the end to ductile fracture.

6. CONCLUDING REMARKS

The mechanical properties of crystalline solids continue to pose challenges of vital interest to both engineers and physicists. Fundamental questions remain concerning the mechanistic origins of phenomena ranging from ductile fracture via void coalescence to power-law creep. In addition, the ability to optimize materials for both strength and toughness has led to new materials such as composites and lightweight intermetallics. The crucial role of defects in mediating mechanical behavior has been highlighted in this article and will continue to provide problems in applied physics for a long time to come.

ACKNOWLEDGMENTS

I am grateful to N. Bhate, A. Bower, A. Carlsson, R. Miller, V. Shenoy, and E. Tadmor, all of whom made valuable suggestions concerning this article. Thanks also to A. Rosakis for the use of his sequence of figures illustrating dynamic fracture.

GLOSSARY

Brittle Fracture: Fracture in which there is no accompanying plastic deformation.

Brittle–Ductile Transition: Temperature-dependent change in character of fracture mechanism. At low temperatures material fails via brittle fracture, while at high temperatures it fails via ductile fracture mechanisms.

Burgers Vector: The unit of slip carried by a dislocation.

Coble Creep: Diffusion-based creep mechanism in which diffusion takes place along grains.

Creep: Time-dependent plastic deformation of material, usually thought of as occurring at high temperature and small loads.

Ductile Fracture: Fracture preceded by plastic deformation.

Elastic Moduli: Material constants relating stresses to elastic strains.

Fatigue: Failure of material subjected to cyclic loading.

Fracture Toughness: A measure of a material's resistance to fracture.

Hardening: Increase in the stress needed for plastic flow. There are a variety of mechanisms (chemical impurities, second-phase particles, dislocation–dislocation interactions) responsible for such hardening.

Nabarro–Herring Creep: Diffusion-based creep mechanism in which diffusion takes place across grains.

Peierls Stress: The stress needed to move a dislocation against intrinsic lattice resistance.

Resolved Shear Stress: Component of traction vector within slip plane resolved along slip direction.

Slip Plane: Crystal lattice plane along which dislocations move.

Solid-Solution Hardening: Increase in yield stress resulting from chemical impurities.

Strain Energy Density: Elastic energy per unit volume stored in body by virtue of its elastic deformation.

Strain Tensor: Geometric measure of deformation suffered by body. Strain characterizes changes in *relative* position.

Stress Intensity Factor: Numerical constant characterizing the strength of the elastic $r^{-1/2}$ singularity associated with a crack.

Stress Tensor: Tensor carrying information about state of internal forces within continuous body. Components of stress tensor give components of force on given planes.

Yield Stress: Stress above which ductile materials suffer permanent deformation.

Works Cited

Ashby, M. F., Gandhi, C., Taplin, D. M. R. (1979), *Acta Metall.* **27,** 699–729.

Ashby, M. F., Jones, D. R. H. (1980), *Engineering Materials,* Oxford: Pergamon.

Boyer, H. E. (Ed.) (1974), *Metals Handbook,* Vol. 9, *Fractography and Atlas of Fractographs,* Metals Park, OH: American Society for Metals, 8th ed.

Boyer, H. E. (Ed.) (1987), *Atlas of Stress-Strain Curves,* Metals Park, OH: ASM International.

Bradbury, E. J. (Ed.) (1985), *Essential Metallurgy for Engineers,* New York: Van Nostrand Reinhold.

Broek, D. (1991), *Elementary Engineering Fracture Mechanics,* Dordrecht: Kluwer Academic Publishers.

Callister, W. D. (1994), *Materials Science and Engineering,* New York: Wiley.

Collins, J. A. (1993), *Failure of Materials in Mechanical Design,* New York: Wiley.

Courtney, T. H. (1990), *Mechanical Behavior of Materials,* New York: McGraw Hill.

Fleischer, R. L. (1963), *Acta Metall.* **11,** 203–209.

Frost, H. J., Ashby, M. F. (1982), *Deformation Mechanism Maps,* Oxford: Pergamon.

Fuchs, H. O., Stephens, R. I. (1980), *Metal Fatigue in Engineering,* New York: Wiley.

Gandhi, C., Ashby, M. F. (1979), *Acta Metall.* **27,** 1565–1602.

Gerold, V. (1979), in: F. R. N. Nabarro (Ed.), *Dislocations in Solids,* Vol. 4, Amsterdam: North Holland.

Gilman, J. J. (1969), *Micromechanics of Flow in Solids,* New York: McGraw Hill.

Griffith, A. A. (1920), *Philos. Trans. R. Soc.* **221,** 163–198.

Haasen, P. (1978), *Physical Metallurgy,* Cambridge, U.K.: Cambridge University Press.

Hall, E. O. (1951), *Proc. Phys. Soc. London, Ser. B* **64,** 747–753.

Hertzberg, R. W. (1976), *Deformation and Fracture Mechanics of Engineering Materials,* New York: Wiley.

Honeycombe, R. W. K. (1968), *The Plastic Deformation of Metals,* New York: St. Martins Press.

Irwin, G. R. (1948), *Fracturing of Metals,* p. 147, Cleveland, OH: American Society for Metals.

Khan, A. S., Huang, S. (1995), *Continuum Theory of Plasticity,* New York: Wiley.

Labusch, R. (1970), *Phys. Status Solidi* **41,** 659–669.

Lawn, B. (1993), *Fracture of Brittle Solids,* Cambridge, U.K.: Cambridge University Press.

Leslie, W. C. (1981), *The Physical Metallurgy of Steels,* New York: McGraw Hill.

Lyman, T. (Ed.) (1972), *Metals Handbook,* Vol. 7, *Atlas of Microstructures of Industrial Alloys,* Metals Park, OH: American Society for Metals, 7th ed.

Orowan, E. (1955), *Welding J.* **34,** 157–160.

Petch, N. J. (1953), *J. Iron Steel Inst.* **174,** 25–28.

Further Reading

Ashby, M. F., Jones, D. R. H. (1980), *Engineering Materials,* Oxford: Pergamon. A gold mine of insights and phenomenology concerning the mechanical properties of materials; both the best place for a first exposure to topics and a reference for many of the types of figures shown in this article.

Batchelor, G. K. (Ed.) (1958), *The Scientific Papers of Sir Geoffrey Ingram Taylor,* Vol. 1, Cambridge, U.K.: Cambridge University Press. Interesting historical insights into the role played by one of the 20th century's great practitioners of classical physics in the understanding of mechanical behavior.

Boyer, H. E. (Ed.) (1974), *Metals Handbook,* Vol. 9, *Fractography and Atlas of Fractographs,* Metals Park, OH: American Society for Metals, 8th ed. A source of a great variety of pictures illustrating fracture surfaces. An enjoyable afternoon's browsing.

Boyer, H. E. (1987), *Atlas of Stress-Strain Curves,* Metals Park, OH: American Society for Metals. Compendium with wealth of stress–strain behavior for variety of metals. Perusal of this volume leads to insights into elastic behavior, plasticity, and the role of crystal orientation and a number of other topics.

Courtney, T. H. (1990), *Mechanical Behavior of Materials,* New York: McGraw Hill. General textbook with interesting coverage of wide range of topics.

Gleiter, H. (1983), "Fundamentals of Strengthening Mechanisms" in: R. C. Gifkins (Ed.), *Strength of Metals and Alloys,* Oxford: Pergamon. An overview of current understanding of many of the strengthening mechanisms discussed in the article.

Haasen, P. (1979), "Solution Hardening in f.c.c. Metals" in: F. R. N. Nabarro (Ed.), *Dislocations in Solids,* Vol. 4, Amsterdam: North Holland. Review article describing efforts to understand solid solution hardening in metals.

Hirth, J. P., Lothe, J. (1982), *Theory of Dislocations,* New York: Wiley. The classic treatise on the theory of dislocations. Covers elastic theory of dislocations as well as the role played by crystal lattice effects.

Kelly, A., Macmillan, N. H. (1986), *Strong Solids,* Oxford: Clarendon. Excellent account of both ideal strengths of solids in addition to role of defects in mediating mechanical behavior.

Khan, A. S., Huang, S. (1995), *Continuum Theory of Plasticity,* New York: Wiley. Discusses phenomenology of plastic deformation and macroscopic ideas used in explaining the phenomena.

Knott, J. F. (1973), *Fundamentals of Fracture Mechanics,* London: Butterworths. Excellent introductory treatment of fracture mechanics with balanced treatment of theoretical insights and experimental foundations.

Kocks, U. F. (1985), *Metall. Trans.* **16A,** 2109–2129. Discussion of difficulties associated with simplified models of solid solution hardening.

Kocks, U. F., Argon, A. S., Ashby, M. F. (1975), *Thermodynamics and Kinetics of Slip,* Oxford: Pergamon. Intriguing book that shows how dislocation considerations may be used to understand macroscopic constitutive ideas used to describe plastic flow.

Kuhlmann-Wilsdorf, D. (1985), *Metall. Trans.* **16A,** 2091–2108. Discussion of experimental and theoretical background on study of work hardening.

Landau, L. D., Lifshitz, E. M. (1959), *Theory of Elasticity,* Oxford: Pergamon. A terse description of many of the key results and ideas of the theory of elasticity.

Lawn, B. (1993), *Fracture of Brittle Solids,* Cambridge, U.K.: Cambridge University Press. A very readable introduction to fracture mechanics.

Lemaitre, J., Chaboche, J.-L. (1990), *Mechanics of Solid Materials,* Cambridge, U.K.: Cambridge University Press. Encyclopedic view of much of theoretical apparatus used in analysis of solid mechanics.

Lyman, T. (Ed.) (1972), *Metals Handbook,* Vol. 7, *Atlas of Microstructures of Industrial Alloys,* Metals Park, OH: American Society for Metals, 8th ed. This entire series is full of interesting figures and data. The microstructures handbook shows many beautiful and intriguing microstructures.

Martin, J. W. (1980), *Micromechanisms in Particle-Hardened Alloys,* Cambridge, U.K.: Cambridge University Press. An undergraduate text with wide coverage on microstructural hardening mechanisms, especially those arising from second-phase particles within the microstructure.

McClintock, F. A., Argon, A. S. (1966), *Mechanical Behavior of Materials,* Reading, MA: Addison-Wesley. A complete appraisal of the field of mechanical properties as of the mid 1960s. Besides its wide coverage, the book suggests a number of interesting "desk-top experiments" that demonstrate various mechanical effects.

Poirier, J.-P. (1985), *Creep of Crystals,* Cambridge, U.K.: Cambridge University Press. A description of both the phenomenology and mechanisms that govern high-temperature deformation.

Suresh, S. (1991), *Fatigue of Materials,* Cambridge, U.K.: Cambridge University Press. A comprehensive research level monograph on the field of fatigue.

Sutton, A. P., Balluffi, R. W. (1995), *Interfaces in Crystalline Materials,* Oxford: Oxford University Press. This book treats interfaces with the same wealth of information and insights that Hirth and Lothe bring to the subject of dislocations.

SOLIDS, DIFFUSION IN

D. J. Young, *School of Materials Science and Engineering, The University of New South Wales, Sydney, Australia*

	Introduction	521
1.	**Point Defects**	521
2.	**Driving Forces and Mobilities**	523
3.	**Vacancy and Tracer (Self-Diffusion) Coefficients**	524
3.1	Random-Walk Description	524
3.2	Absolute-Rate–Theory Description	525
3.3	Self-Diffusion Coefficient Values	527
4.	**Use of Phenomenological Equations**	528
4.1	Relationship with Correlation Coefficients	529
5.	**Diffusion in Alloys**	530
5.1	Origins of Cross Effects	530
5.2	Interstitial and Substitutional Diffusion	531
5.3	Kirkendall Effect	532
6.	**Diffusion in Ionic Solids**	534
6.1	Relative Building Units	534
6.2	Reduction of Transport Equations	535
7.	**Experimental Methods**	536
7.1	Calculation of Diffusion Coefficients	537
	Glossary	538
	Works Cited	538
	Further Reading	539

INTRODUCTION

Diffusion is the sole mechanism of mass transport in isothermal solids free of external fields (magnetic, mechanical stress, etc.). It is consequently of great practical importance in a number of branches of materials science and engineering. These include semiconductor device fabrication, corrosion and scaling of metals, sintering of ceramics and metals, creep, surface engineering, and a number of phase transformations. The phenomenon can be described without reference to the nature of the solid, or its constituent particles, in the classical approach represented by Fick's (1855) first law

$$\mathbf{J} = -D\nabla C.$$

Here **J** is the flux in moles per unit area per second, ∇C is the gradient in molar concentration, and D is the diffusion coefficient, approximated as a constant characteristic of the material. Such an approach lacks any predictive capacity. However, the diffusion coefficient can ultimately be related to the frequency of particle jumps and the lattice constant of a solid, permitting at least qualitative prediction of material behavior.

Section 1 considers the defect species in solids responsible for diffusion, and Sec. 2 uses the thermodynamics of irreversible processes to identify chemical-potential gradients as the driving forces for their diffusion. In Sec. 3, the kinetics of vacancy diffusion is discussed, and expressions are arrived at for self-diffusion, i.e., diffusion in the absence of any gradient other than that of an isotope. This section concludes with a brief examination of experimental data.

In the practical situations of interest where diffusion occurs in a chemical-potential gradient, the diffusion coefficient defined in Fick's law becomes a function of concentration. Irreversible thermodynamics is used in Sec. 4 to produce a more general description, appropriate to this situation. This approach is then applied to alloys in Sec. 5 and to ionic solids in Sec. 6. Finally, the principal methods of measuring diffusion coefficients are described in Sec. 7.

1. POINT DEFECTS

Solid-state diffusion involves the movement of individual particles (molecules, atoms, or ions) that constitute the material.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

These particles are capable of movement because they vibrate around their mean positions and because the existence of defects in the solid crystal permits an occasional vibration to extend into a translation to an adjacent lattice site. Two common defects are illustrated in Fig. 1 for the case of a pure monatomic solid: a vacancy, or unoccupied lattice site, and an interstitial atom, *i.e.*, one located between normal lattice sites.

A lattice atom can move into an adjacent vacancy, exchanging sites with the defect. Movement via this vacancy mechanism is the most common way in which diffusion occurs. Clearly, the concentration of vacancies present is important in determining the probability of atomic translation occurring. The interstitial species can contribute to diffusion simply by moving into an adjacent interstitial site. This is improbable in pure metals, because the atoms are large, but operates for interstitial impurities such as C, H, N, and O dissolved in metals. The interstitial species can also move onto an adjacent normal lattice site, displacing its occupant into one of the adjacent interstitial sites, as shown in Fig. 2. This is the "interstitialcy" mechanism and is known to be important in the diffusion of silver in AgBr (Friauf, 1957). Whichever the mechanism, the concentration of defects is an important factor in the particle movement rates. The question of defect concentrations is now considered.

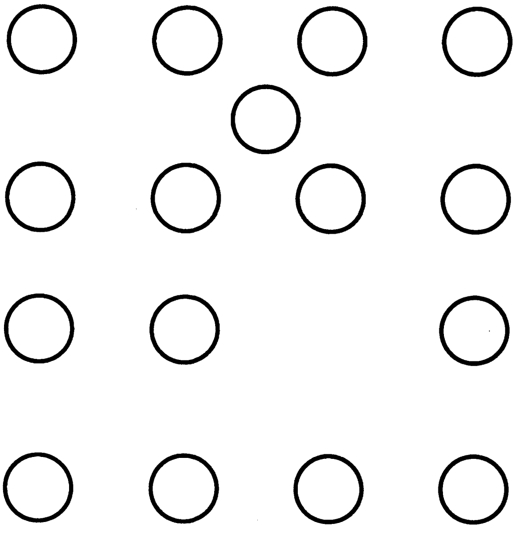

FIG. 1. Schematic view of a solid monatomic crystal containing a vacancy and an interstitial.

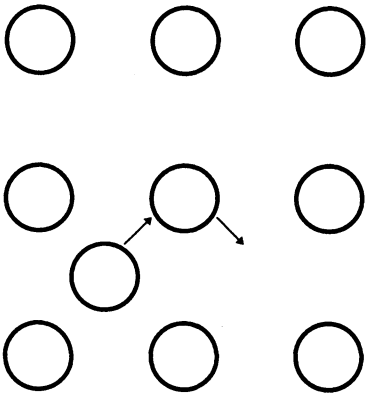

FIG. 2. Interstitialcy diffusion involves simultaneous movement of interstitial and lattice species.

Equilibrium concentrations of point defects in crystals are calculated by the methods of statistical thermodynamics. The application of these methods to crystals has been reviewed in detail by Schottky (1958), and their use in diffusion calculations has been explored by several authors, notably Howard and Lidiard (1964). The Gibbs free energy G for a monatomic crystal containing n_v vacancies and N atoms is

$$G = G_0 + n_v g_v - kT \ln \frac{(N + n_v)!}{N! n_v!}, \qquad (1)$$

where g_v is the free energy of formation of a vacancy, k is Boltzmann's constant, and the logarithmic term is recognized as the configurational entropy resulting from the presence of defects. It is this term that makes vacancy formation inevitable at all temperatures above absolute zero. The free-energy minimum representing the equilibrium state of the crystal defines the chemical potential of the vacancies as zero:

$$\left(\frac{\partial G}{\partial n_v}\right)_{T,P,N} = \mu_v = 0. \qquad (2)$$

Application of this to Eq. (1), making use of Stirling's approximation ($\ln N! = N \ln N - N$), then yields

$$X_v = n_v/(N + n_v) = \exp(-g_v/kT), \qquad (3)$$

where X indicates mole fraction. Thus the equilibrium concentration of vacancies rises rapidly with temperature. A similar expression is obtained for interstitials. These expressions neglect interactions between defects, and are thus suitable only for low defect concentrations.

The same treatment is successful in the case of ionic solids. Vacancy formation in an insulating compound must involve simultaneously both the cation and anion sublattices in order to preserve charge neutrality. In a compound MX of 1:1 stoichiometry, "Schottky pairs" of vacancies result. Similarly, interstitial formation involves the simultaneous formation of a lattice vacancy and an interstitial species, thus again preserving charge neutrality. The same statistical approach applies, and the overall configurational entropy for Schottky defect formation is given by

$$S = kT \ln\left[\frac{(N_m + n_{vm})!}{N_m! n_{vm}!} \frac{(N_x + n_{vx})!}{N_x! n_{vx}!}\right], \tag{4}$$

where n_{vm} is the number of cation vacancies and n_{vx} is the number of anion vacancies. Hence, following the same free-energy minimization procedure, it is found that

$$\mu_{vm} + \mu_{vx} = 0, \tag{5}$$

and

$$X_{vm} X_{vx} = \exp[-(g_{vm} + g_{vx})/kT]. \tag{6}$$

Equations (3) and (6) are recognized as forms of the mass-action law.

2. DRIVING FORCES AND MOBILITIES

Having identified the principal mechanisms of diffusion in crystalline solids, we now seek to calculate the rate of the process. In addition to a knowledge of defect concentrations, evaluation of their mobilities and the driving forces that induce their movement is required. The most general approach to the latter question is provided by the methods of irreversible thermodynamics. The principal concepts were developed by Onsager (1931). More complete accounts are given by deGroot and Mazur (1962) and Prigogine (1967), while the application to solid-state diffusion has also been reviewed (e.g., Howard and Lidiard, 1964; Kirkaldy and Young, 1987).

The procedures of irreversible thermodynamics enable us to calculate the rate of entropy production per unit volume, σ, in terms of the various fluxes flowing within the system. The result is a bilinear expression involving the fluxes themselves and a set of thermodynamic forces. These forces are thereby identified as those responsible for the fluxes, each flux being linearly dependent on all the forces. The description is applicable only to systems that are not far removed from equilibrium and is therefore appropriate to diffusion in a solid within which the local equilibrium state is closely approached.

For isothermal diffusion in a closed, isobaric, field-free system, it is found for a temperature T that

$$T\sigma = -\sum_i \mathbf{J}_i \cdot \nabla \eta_i \tag{7}$$

and hence that

$$\mathbf{J}_i = -\sum_j L_{ij} \nabla \eta_j, \tag{8}$$

where $\mathbf{J}_i$ is the flux of component i, and the coefficients L_{ij} will be discussed shortly. The driving forces are seen to be the gradients $\nabla \eta_j$, known as electrochemical potential gradients. They are defined by

$$\eta = \mu + zF\psi, \tag{9}$$

where z is the effective charge on the particle, F is the Faraday, and ψ is the local electrostatic potential. Gradients in potential constitute fields, but these are internal to the solid, and the conditions for the validity of Eq. (7) are maintained. In the case of a metallic alloy, the constituent atoms have no effective charge, and the driving force is the chemical potential gradient, $\nabla\mu$. This result is intuitively satisfactory in the sense that diffusion is perceived (under the conditions specified earlier) as a process that eliminates differences in chemical potential, thereby achieving equilibration.

Now that the driving forces have been identified, attention is directed to the phenomenological coefficients L, so called because they do not rely on any detailed description of the solid structure or the diffusion mechanism. The method of dealing with the set of equations (8) will be explained and its value illustrated later. For the moment, one-dimensional diffusion in a binary system is examined, and the approximation $L_{12} \simeq 0$ leads to the simple result

$$J_1 = -L_{11}\frac{d\mu_1}{dx}. \tag{10}$$

Methods of evaluating L_{ij} are discussed later.

If we recall that

$$\mu_1 = \mu_1^0 + RT\ln a_1, \tag{11}$$

with μ_1^0 the standard chemical potential and a_1 the activity of component 1, then it is seen that for an *ideal solution* ($a_1 = X_1$),

$$J_1 = -\frac{L_{11}RT}{C_1}\frac{dC_1}{dx}, \tag{12}$$

where C is the molar concentration. This is equivalent to Fick's first law of diffusion,

$$J_1 = -D_1\frac{dC_1}{dx} \tag{13}$$

with

$$D_1 = L_{11}RT/C_1, \tag{14}$$

which is the diffusion coefficient. This coefficient is commonly regarded as a material property. It must be remembered, however, that D_1 varies with the concentration of component 1, and also with the concentrations of all other components. This point is explored further after evaluation of L_{11}.

3. VACANCY AND TRACER (SELF-DIFFUSION) COEFFICIENTS

The measurement of D can be laborious, and considerable effort has been devoted to theoretical calculation of the quantity. Theory does not yet permit *a priori* calculation, but it does provide a number of predictive tools that can be used to minimize experimental effort. The most widely used approach is based on random-walk theories originally due to Einstein (1905, 1906) and Smoluchowski (1908).

3.1 Random-Walk Description

Atom–vacancy exchanges occur about 10^8 times per second in metals just below their melting points. Because the direction of interchange is random and because the number of events is very large, statistical methods work well. In the random progress of a single vacancy over n successive interchanges or jumps, the squared displacement is

$$\mathbf{x}^2 = (\mathbf{x}_1 + \mathbf{x}_2 + \ldots \mathbf{x}_n)^2. \tag{15}$$

Rearrangement to collect diagonal terms yields

$$\mathbf{x}^2 = \sum_{i=1}^{n}\mathbf{x}_i^2 + 2\sum_{j-1}^{n-1}\sum_{k=j+1}^{n}\mathbf{x}_j\cdot\mathbf{x}_k. \tag{16}$$

In a purely random process the terms in the second summation have equal probabilities of being positive or negative. Averaging over the trajectories of many vacancies then yields the mean displacement

$$\langle\mathbf{x}^2\rangle = \left\langle\sum_{i=1}^{n}\mathbf{x}_i^2\right\rangle = \sum_{i=1}^{n}\langle\mathbf{x}_i\rangle^2. \tag{17}$$

For cubic lattices all jumps are of the same length, λ. Then

$$\langle\mathbf{x}^2\rangle = n\lambda^2. \tag{18}$$

The number of jumps in the x direction occurring in time t is related to the isotropic jump frequency ν through

$$n/t = \nu/3 \tag{19}$$

and

$$\langle\mathbf{x}^2\rangle = \nu\lambda^2 t/3. \tag{20}$$

Since (Manning, 1968)

$$\langle\mathbf{x}^2\rangle = 2Dt, \tag{21}$$

it follows that

$$D = \nu\lambda^2/6, \tag{22}$$

the Einstein–Smoluchowski relation. However, this diffusion coefficient applies to the vacancy, whereas knowledge of atomic (or ionic) diffusion is usually the objective.

Although the atoms of a single-component solid are continually moving, their indistinguishibility means that the motion is not observable. This difficulty is overcome by observing the diffusional intermixing of different isotopes of the same atom (or ion). In the ideal case, the enthalpy of mixing is zero, and a purely entropic process is observed. The relationship between the diffusion coefficient of the tracer D_{A^*} and that of the vacancy D_v is approached by recognizing that the jump frequency of a moving species is proportional to the probability of finding the appropriate species on a nearest-neighbor site. Thus a tracer atom can move only when a vacancy is adjacent (probability X_v), and a vacancy can move when an atom is adjacent (probability $X_A \approx 1$). It follows then that

$$D_{A^*} = fD_vX_v, \tag{23}$$

where f is a constant of proportionality. If the tracer movement were completely random, then we would have $f = 1$. However, this cannot be so for diffusion via a vacancy mechanism, and successive jumps are said to be correlated.

The analytic calculation of correlation effects is complicated (Manning, 1968; Kidson, 1978; Murch, 1984) but amenable to Monte-Carlo methods (*q.v.;* see also Murch, 1984). The physical origin of the effects can be easily understood from a consideration of the possibilities confronting a tracer atom immediately after it has interchanged with a vacancy. The most likely event for the tracer is prompt reinterchange with the vacancy. Thus the tracer will travel a shorter distance in a given number of jumps than will a vacancy. This means that the nondiagonal terms in Eq. (16) no longer sum to zero. An approximate estimate of the size of the effect is arrived at (Shewmon, 1963) by noting that for a crystal coordination number of z, the probability that a vacancy will jump back to the tracer position is $1/z$. This will cancel out the first jump, making two tracer jumps ineffective. Thus

$$f \approx 1 - 2/z. \tag{24}$$

Estimates of the correlation factor f made in this way are compared with precise calculations in Table 1. It is to be noted that $f = 1$ for dilute interstitial diffusion because randomness is assured.

The random-walk description informs us that diffusion coefficients can be calculated from a knowledge of defect concentrations and jump frequencies and that the result will vary with direction in anisotropic solids. It also predicts that the correlation factor varies with defect type (values are also available for interstitialcy and divacancy diffusion: Murch, 1991; Philibert, 1991). This fact is made use of in distinguishing diffusion mechanisms in ionic solids, for which the correlation factor can be experimentally accessible (Murch, 1991). In addition, the study of correlation effects has at least the potential to help calculate the diffusion properties of highly disordered solids (Murch, 1984; Sato, 1989). Finally, the methodology has been applied to binary alloys, but with mixed results (Bocquet *et al.*, 1983; Kirkaldy and Young, 1987).

3.2 Absolute-Rate–Theory Description

We turn now to the evaluation of individual jump frequencies, using absolute-rate theory. The first applications to solid-state diffusion were reported by Wert and Zener (1950) and Seitz (1950), and subsequent extensions for various cases have been provided by others (e.g., Lidiard, 1955; Dignam *et al.*, 1973; Young and Kirkaldy, 1984). Alternative approaches exist, notably linear-response theory (Kubo, 1957; Allnatt and Okamura, 1984) and the path-probability method (Kikuchi and Sato, 1969; Sato and Zhang, 1993), but will not be considered here.

When a particle moves from one lattice position to another, it passes through an intermediate state that has a higher energy because adjacent particles must be perturbed

Table 1. Comparison of approximate and exact values of f for vacancies.

Structure	$1 - 2/z$	f
Simple cubic	0.67	0.6555
Body-centered cubic	0.75	0.7215
Face-centered cubic	0.83	0.7814
Diamond	0.5	0.5

from their mean lattice positions in order to accommodate the passage of the moving particle. During this lattice distortion, an activated complex involving the two interchanging species (e.g., particle plus vacancy) is formed. The activity a_{iv} of the complex is described via the equilibrium constant K_{iv} for its formation:

$$\frac{a_{iv}}{a_i a_v} = K_{iv} = \exp\left(-\frac{\Delta H_{iv}}{RT}\right)\exp\left(\frac{\Delta S_{iv}}{R}\right), \qquad (25)$$

where ΔH_{iv} is the enthalpy and ΔS_{iv} is the entropy of complex formation.

A profile of the periodic internal-energy surface in a direction parallel to that of diffusion is shown in Fig. 3. An electrostatic field can be externally imposed, or can arise through the movement of the charged species themselves, and will in this case be aligned with the diffusion direction. The height of the energy barrier to diffusion of a charged species is modified by the field, being lower for downfield movement than for upfield movement of an appropriately charged species. It will be assumed that the field does not affect ΔS_{iv}. We may write for the interchange of species i and a vacancy between planes (1) and (2) separated by a distance λ

$$\begin{aligned} J = {} & m\lambda\nu_{iv}\exp\left(\frac{\Delta S_{iv}}{R}\right) \\ & \times \left\{a_i^{(1)}a_v^{(2)}\exp\left[-\frac{\Delta U_{iv} - zF(\psi^{(0)} - \psi^{(1)})}{RT}\right]\right. \\ & \left. - a_v^{(1)}a_i^{(2)}\exp\left[-\frac{\Delta U_{iv} - zF(\psi^{(2)} - \psi^{(0)})}{RT}\right]\right\} \\ & \times \gamma_{iv}^{-1}, \end{aligned} \qquad (26)$$

where m is the volume concentration of lattice sites, ν_{iv} is a kinetic frequency term, and γ_{iv} is the activity coefficient for the transition-state complex. Here z is the effective charge of the vacancy, that of the cation being zero. In the more general case of interchange between any two differently charged species, z is replaced by the difference in effective charges of the participating species. Superscripts in parentheses represent the location at which the quantity in question is evaluated.

Equation (26) is cleared of common terms and subjected to Taylor series expansion of the terms $a_v^{(2)}$ and $a_i^{(2)}\exp(-zF\psi^{(2)}/RT)$. Retention of linear terms, in the case where the field is not inordinately high, leads to

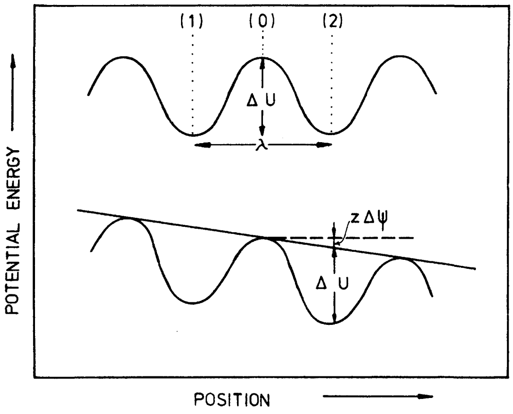

FIG. 3. Potential-energy profile in diffusion direction: upper curve, no electrostatic field; lower curve, showing effect of electrostatic field.

$$J = -\frac{m\lambda^2}{RT}\nu_{iv}K_{iv}a_i a_v\{\nabla\mu_i - \nabla\mu_v - zF\nabla\psi\}, \qquad (27)$$

which, upon substituting from Eq. (9), becomes

$$J = -\frac{m\lambda^2}{RT}\nu_{iv}K_{iv}a_i a_v\{\nabla\eta_i - \nabla\eta_v\}. \qquad (28)$$

Expressions of this sort always apply to pairwise site exchanges.

For diffusion of noncharged species in a metal or an alloy, $z = 0$, and we obtain

$$J_i = -\frac{m\lambda^2}{RT}\nu_{iv}K_{iv}a_i a_v\{\nabla\mu_i - \nabla\mu_v\}. \qquad (29)$$

If the equilibrium condition of Eq. (2) is realized and the off-diagonal terms in Eq. (8) are ignored, this result simplifies to Eq. (10), whence

$$L_{11} = (m\lambda^2/RT)\nu_{iv}K_{iv}a_i a_v,$$

and therefore, in the dilute (ideal) solution approximation, it is found from Eq. (14) that

$$D_1 = \lambda^2\nu_{iv}\exp\left(\frac{-\Delta H_{iv}}{RT}\right)\exp\left(\frac{\Delta S_{iv}}{R}\right)X_v, \qquad (30)$$

which is seen to correspond to Eq. (23), except that the correlation coefficient cannot be evaluated from this simple procedure, which omits the off-diagonal L_{ij} terms.

We now combine Eqs. (3) and (30) to determine the temperature dependence of the diffusion coefficient:

$$D = D_0 \exp[-(\Delta H_{iv} + \Delta H_v)/RT], \qquad (31)$$

where ΔH_v is the enthalpy (per mole) for vacancy formation, and the remaining constants have been collected in D_0. The expected Arrhenius form is arrived at and is commonly used to interpolate or extrapolate sparse experimental data.

3.3 Self-Diffusion Coefficient Values

The discussion thus far has been concerned with diffusion in a pseudo one-component system ($A + A^*$) and is applicable to both a pure metal and a single sublattice of a pure compound. The measured tracer diffusion coefficient D_{A^*} is also frequently known as the self-diffusion coefficient, because the tracer and solvent species are chemically identical. However, it should be noted that some writers reserve the term self-diffusion coefficient for the quantity D_{A^*}/f. Extensive tabulations of self-diffusion are available (Askill, 1970; Brandes, 1983; Freer, 1980; Mehrer, 1990), and a good general discussion is provided by Philibert (1991). Brown and Ashby (1980) have examined a wide range of data and reconfirmed a number of general correlations.

The self-diffusion coefficient at the melting point, D_{T_m}, is very roughly constant for a given class of materials, as shown in Fig. 4. Using Eq. (31), it is seen that

$$D_{T_m} = D_0 \exp[-(\Delta H_{iv} + \Delta H_v)/RT_m]. \qquad (32)$$

The factor D_0 contains the atomic vibration frequency, the lattice constant λ, a lattice-specific factor including the correlation coefficient, and the entropies of vacancy formation and movement. For solids of the same crystal structure and bond type, these quantities are all approximately equal, provided

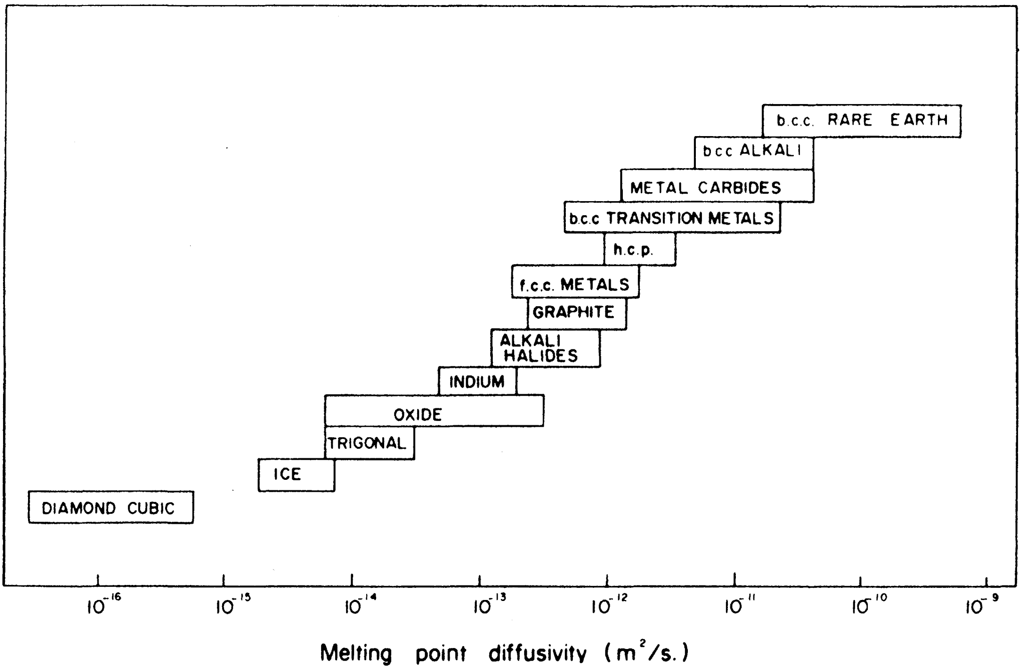

FIG. 4. Comparison of self-diffusion coefficients extrapolated to melting points. Reprinted with permission from *Acta Metall.* **28,** Brown and Ashby, "Correlations for Diffusion Coefficients," pp. 1085–1101, Copyright 1980, with kind permission from Elsevier Science Ltd., The Boulevard, Langford Lane, Kidlington OX5 1GB, UK.

that particle masses (which affect ν) do not vary greatly.

Following classical arguments (e.g., Mott and Gurney, 1964), it is noted that the formation of a vacancy in a constant-mass system involves movement of a particle from the interior of a solid to its surface in the usual event that interstitial formation is impossible. This is energetically equivalent to breaking half the particle's bonds, and the enthalpy of the process would be expected to be proportional to the heat of sublimation, ΔH_s. The movement of a particle into the vacancy during diffusion involves temporary local disordering and is therefore like local melting (Sherby *et al.*, 1973). The enthalpy of movement is expected to be proportional to the heat of fusion, ΔH_m. Thus the activation energy Q appearing in Eq. (32) can be written

$$Q = \alpha \Delta H_s + \beta \Delta H_m, \qquad (33)$$

with α and β as constants. At the melting point, phase equilibrium is achieved, and

$$\Delta S_m = \Delta H_m / T_m, \qquad \Delta S_s = \Delta H_s / T_m, \qquad (34)$$

with ΔS_m the entropy of fusion and ΔS_s the entropy of sublimation at the melting point. Then we may write

$$Q/RT_m = (\alpha \Delta S_s + \beta \Delta S_m)/R. \qquad (35)$$

If these entropy changes are approximately constant for a particular crystal structure and bond type, then so is Q/RT_m, and the constancy of D_{T_m} is explained. Data collected by Brown and Ashby (1980) show the approximate constancy of Q/RT_m for a number of crystal types, as seen in Fig. 5. The data in Figs. 4 and 5 can be used to arrive at order-of-magnitude estimates for self-diffusion at lower temperatures. Other formulas for Q have recently been reviewed by Cahoon and Sherby (1992).

4. USE OF PHENOMENOLOGICAL EQUATIONS

The application of Eq. (8) to a solid is now explored. For an n-component substitutional solid in which diffusion occurs

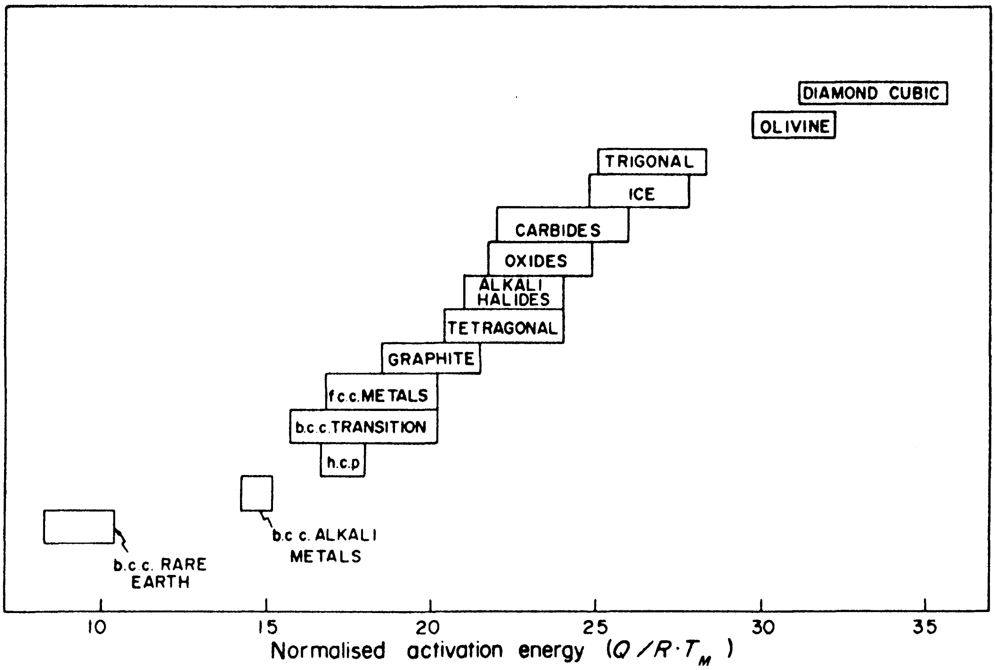

FIG. 5. Normalized activation energies for diffusion in different crystal types. Reprinted with permission from *Acta Metall.* **28**, Brown and Ashby, "Correlations for Diffusion Coefficients," pp. 1085–1101, Copyright 1980, with kind permission from Elsevier Science Ltd., The Boulevard, Langford Lane, Kidlington OX5 1GB, UK.

through pairwise exchanges of lattice-site occupants with vacancies, the transport equations are

$$J_i = -\sum_{j=1}^{n} L_{ij}\nabla\mu_j - L_{iv}\nabla\mu_v, \tag{36}$$

where charge-neutral species are concerned. The equations contain redundancies as a result of site and mass conservation in a lattice-fixed frame of reference:

$$\sum_{i=1}^{n} J_i + J_v = 0, \tag{37}$$

$$\sum_{i=1}^{n} X_i + X_v = 0. \tag{38}$$

The constraints (37) and (38) can be applied to Eqs. (36), reducing them to a linearly independent set. A simpler method is to commence with the entropy-source expression (7), rewritten now as

$$T\sigma = -\sum_{i} J_i\nabla\mu_i - J_v\nabla\mu_v. \tag{39}$$

Application of (37) to (39) then leads to

$$T\sigma = -\sum_{i} J_i\nabla(\mu_i - \mu_v) \tag{40}$$

and its consequence

$$J_i = -\sum_{j=1}^{n} L_{ij}\nabla(\mu_i - \mu_v). \tag{41}$$

Comparison with Eq. (29) shows how the elementary kinetic theory will provide an evaluation of the L_{ij}.

The most important finding of the irreversible-thermodynamics approach is the Onsager reciprocal relationship between off-diagonal terms,

$$L_{ij} = L_{ji}, \tag{42}$$

which applies when Eqs. (41) constitute an independent set. This makes the description of multicomponent diffusion much more efficient and reduces the number of measurements necessary. It is important, wherever possible, to reduce still further the experimental effort required, by arriving at approximations to the L_{ij}.

4.1 Relationship with Correlation Coefficients

The off-diagonal terms describe diffusion driven by gradients other than that of the diffusing species itself. Estimation of these "cross effects" is important. One approach to the problem consists of relating the L_{ij} to correlation effects. A large and difficult literature has accumulated on this subject (see, e.g., Howard and Lidiard, 1964; Manning, 1968; LeClaire, 1975; Allnatt and Allnatt, 1984; Heumann, 1979; Kirkaldy *et al.*, 1987). We consider here only the application to tracer diffusion in the alloy *A*–*A**–*V*, i.e., an alloy of *A* with its tracer isotope *A**, containing vacancies, *V*. The appropriate independent flux equations are

$$J_A = -L_{AA}\nabla(\mu_A - \mu_v) - L_{AA^*}\nabla(\mu_{A^*} - \mu_v), \tag{43}$$

$$J_{A^*} = -L_{A^*A}\nabla(\mu_A - \mu_v) - L_{A^*A^*}\nabla(\mu_{A^*} - \mu_v), \tag{44}$$

and the correlation factor is related to these coefficients by

$$\begin{aligned} L_{AA} &= \frac{D_vN}{kT}\frac{X_AX_v}{X_A + X_{A^*}}[1 - X_v - X_{A^*}(1 - f)], \\ L_{A^*A^*} &= \frac{D_vN}{kT}\frac{X_{A^*}X_v}{X_A + X_{A^*}}[1 - X_v - X_A(1 - f)], \end{aligned} \tag{45}$$

and

$$L_{AA^*} = L_{A^*A} = \frac{D_vN}{kT}\frac{X_AX_{A^*}X_v}{X_A + X_{A^*}}(1 - f). \tag{46}$$

It is seen that only if tracer jumps are completely uncorrelated ($f = 1$) do the cross terms vanish. Calculations for systems containing more than one chemical component are much more complex. Since the magnitudes of the effects predicted are generally small compared with other factors, they will not be considered further.

In discussing diffusion in real solutions, it becomes impossible to retain the ideal-solution approximation leading from Eq. (10) to

Eqs. (12) and (14). Instead, the activity coefficient γ is introduced,

$$a_i = \gamma_i X_i, \tag{47}$$

and, if off-diagonal terms are neglected, it is found that

$$D_1 = \frac{L_{11}RT}{X_1}\left(1 + \frac{\partial \ln \gamma_1}{\partial \ln X_1}\right), \tag{48}$$

where use has been made of the definition $X_1 = C_1/C$. The magnitude of the thermodynamic effect can be quite significant, as shown in Sec. 5.2.

5. DIFFUSION IN ALLOYS

5.1 Origins of Cross Effects

Equation (8) may be rewritten for atomic diffusion as

$$J_i = -\sum_j L_{ij} \nabla \mu_j, \tag{49}$$

whence it is clear that cross effects can arise through either kinetic interactions, as represented by the Onsager coefficients, or thermodynamic interactions, represented by the variation of chemical potential with composition. For simplicity, we assume equal molar volumes for all constituents and write the local thermodynamic equilibrium statement for the chemical potential of a component in an isothermal, isobaric n-component system:

$$\mu_i = \mu_i(C_1, C_2, \ldots . C_{n-1}). \tag{50}$$

In this case

$$\nabla(\mu_i - \mu_n) = \sum_j^{n-1} \mu_{ij} \nabla C_j, \tag{51}$$

where

$$\mu_{ij} = \frac{\partial(\mu_i - \mu_n)}{\partial C_j}. \tag{52}$$

Substitution into Eq. (41), with μ_v replaced by μ_n, yields

$$J_i = -\sum_j^{n-1} D_{ij} \nabla C_j \tag{53}$$

with

$$D_{ij} = \sum_k^{n-1} L_{ik} \mu_{kj}. \tag{54}$$

Equation (53) is recognized as a generalized Fick equation for the description of diffusion driven by multiple concentration gradients. It is now explicit that the cross effects are due to thermodynamic interactions (μ_{ij}) and kinetic effects (L_{ij}).

The need for this degree of complexity is illustrated by Darken's (1948) famous experiment on carbon diffusion in Fe–Si–C alloys. A bar of homogeneous Fe–0.44%C was joined to one of Fe–0.48%C – 4%Si and diffused at 1050 °C. At this temperature the diffusion couple is fcc austenite, and carbon diffuses rapidly, whereas silicon diffusion is extremely slow. The resulting carbon concentration profile is shown in Fig. 6, where the carbon is seen to have "unmixed," i.e., to have spontaneously adopted a nonhomogeneous concentration distribution. This cannot possibly be predicted or analyzed by use of Fick's law [Eq. (13)]. The reason for the unmixing lies in the thermodynamic effect of silicon on carbon: the addition of silicon to Fe–C alloys increases μ_C. The initial state of the diffusion couple contained equal carbon *concentrations* but a higher *activity* in the half containing silicon. In fact, the system shown in Fig. 6 is moving toward the equilibrium state in which μ_C would be everywhere the same.

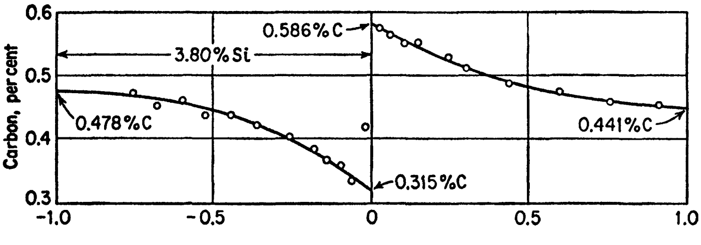

FIG. 6. Distribution of carbon in Fe–Si–C infinite diffusion couple, after 13 days at 1050 °C. Abscissa is fractional position. After Darken. Reprinted with permission from *Trans. Metall. Soc.* **180,** p. 43 (1949).

Solid-solution thermodynamic models are usually expressed in terms of mole fractions, and Eq. (52) is therefore transformed into

$$\bar{\mu}_{ij} = \frac{\partial(\mu_i - \mu_j)}{\partial X_j} = C\mu_{ij}. \qquad (55)$$

In the case of equal partial molar volumes, it can be shown (Miller *et al.*, 1986; Kirkaldy and Purdy, 1969) that the D matrix, as determined on a mole fraction scale, is given by the matrix product

$$C[D] = [L][\bar{\mu}] \qquad (56)$$

where the $\bar{\mu}$ matrix is symmetric. A useful model for dilute solution (solvent component n) was provided by Wagner (1952):

$$\ln \gamma_i = \ln \gamma_i^0 + \sum_{j=1}^{n-1} \epsilon_{ij} X_j, \qquad (57)$$

where γ_i^0 is the value of the activity coefficient at infinite dilution and ϵ_{ij} are the "interaction parameters." Combination of Eqs. (47), (55), and (57) leads to

$$\bar{\mu}_{ij} = RT\epsilon_{ij} \quad (i \neq j), \quad \epsilon_{ij} = \epsilon_{ji}; \qquad (58)$$

$$\bar{\mu}_{ii} = RT[1/X_i + \epsilon_{ii}]. \qquad (59)$$

For a ternary system, the two necessary flux equations are found in Eq. (53), the coefficients of which are evaluated from Eqs. (56) and (59) as

$$\begin{aligned} D_{11} &= RT\{L_{11}(\epsilon_{11} + 1/X_1) + L_{12}\epsilon_{21}\}, \\ D_{12} &= RT\{L_{11}\epsilon_{12} + L_{12}(\epsilon_{22} + 1/X_2)\}, \\ D_{21} &= RT\{L_{22}\epsilon_{21} + L_{21}(\epsilon_{11} + 1/X_1)\}, \\ D_{22} &= RT\{L_{22}(\epsilon_{22} + 1/X_2) + L_{21}\epsilon_{12}\}. \end{aligned} \qquad (60)$$

In an ideal solution all $\epsilon_{ij} = 0$, and the D_{ij} reduce to the purely kinetic form

$$D_{ij} = RTL_{ij}/X_j. \qquad (61)$$

For real solutions, if no kinetic cross effects occur, i.e., $L_{ij}(i \neq j) = 0$, it is clear that the diffusional cross terms are nonzero. In this case the dilute-solution limit (X_1, $X_2 \to 0$) may be described by

$$D_{12}/D_{11} = X_1\epsilon_{12}/(1 + \epsilon_{11}X_1), \qquad (62)$$

and

$$D_{21}/D_{22} = X_2\epsilon_{21}/(1 + \epsilon_{22}X_2). \qquad (63)$$

Thus the ternary coefficients are determined uniquely by the binary ones in dilute solutions.

5.2 Interstitial and Substitutional Diffusion

For interstitial diffusion there are negligible correlations between sublattices so that the approximation $L_{ij}(i \neq j) = 0$ is valid. Practical examples are steels Fe–C–*M* in which *M* is a substitutional metal (Si, Mn, Ni, Cr, Mo, Co) and the carbon is interstitial. The variation of D_{CM}/D_{CC} with carbon concentration is shown in Fig. 7, compared with the predictions of Eq. (62) using independently measured interaction parameters (Brown and Kirkaldy, 1964). Agreement is quite good, and the error of up to about 10% could be due to inaccurate values of the interaction parameters.

In considering diffusion in substitutional alloys, it is normal to use a volume-fixed frame of reference. If the partial molar volumes can be adequately approximated as

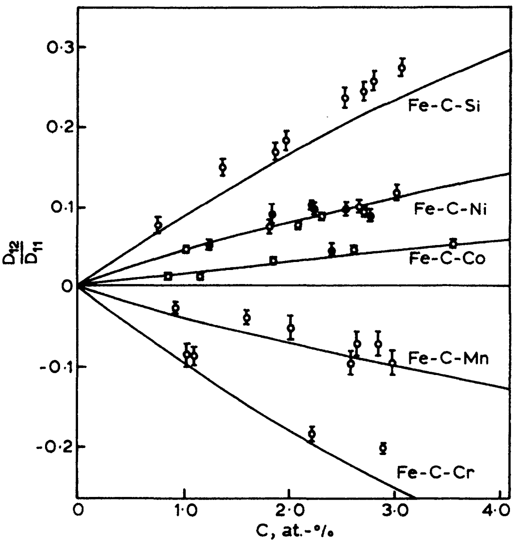

FIG. 7. Variation of D_{12}/D_{11} in Fe–C–M alloys. The solid lines show the thermodynamic prediction. After Brown and Kirkaldy. Reprinted with permission from *Trans. Metall. Soc.* **230**, p. 223 (1964), a publication of the Minerals, Metals & Materials Society, Warrendale, PA.

equal, then the frame of reference is simply defined by

$$\sum_{i=1}^{n} J_i = 0. \tag{64}$$

Physically, this corresponds to recognition of the fact that diffusion of a particular species from a solid region does not leave void space behind: rather another species will diffuse into it. It is recognized that this conservation law leads directly to a diffusional cross effect: movement of one component leads necessarily to the movement of others. The process can be visualized as shown in Fig. 8, where it is seen that while vacancies are necessarily involved, they function simply as intermediaries. The kinetics of this process is described using absolute-rate theory applied to the interchange of any two species i and j, thereby arriving at estimates of L_{ij}.

The applicability of this description to experimental data for ternary alloys has been discussed by Kirkaldy and Young (1987), Purdy (1990), and Dayananda (1991, 1992). However, to understand measured data, it is necessary to relate the experimental (laboratory) reference frame with the lattice frame used in theoretical discussions. The physical manifestation of the difference between the two frames is known as the Kirkendall effect.

5.3 Kirkendall Effect

The discussion commences with a substitutional binary alloy AB in which diffusion occurs via atom-vacancy exchanges, and Eq. (29) applies to both A and B. It is seen that these individual fluxes are in general not equal and opposite. If, for example, $D_A > D_B$ in a sample initially A rich on the left, there will be an excess flux of A from left to right over B atoms moving to the left. Thus the diffusion zone as a whole drifts to the left, compensating for the accumulation of matter and hydrostatic pressure that would otherwise occur on the right. Since the diffusion zone is generally a small part of a larger sample, measurements of position that are referred to the end of the sample (the laboratory reference frame) are affected by this drift, and so, in consequence, is the estimate of diffusion rate. This is dealt with by relating the two frames of reference.

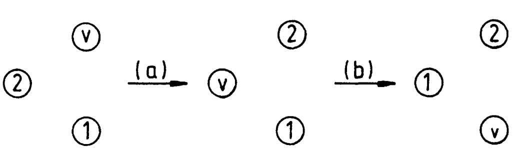

FIG. 8. Site-exchange model for diffusion in (A,B)O with fixed anion sublattice. Species shown are the two different cations and a vacancy.

In the laboratory frame, Eq. (64) applies. However, in the lattice frame [to which Eq. (29) refers] it does not. We therefore denote this latter frame by primes, observing that

$$J'_A + J'_B = -J'_v. \tag{65}$$

If the lattice frame moves with respect to the laboratory frame with a velocity v, then

$$J_i = J'_i + C_i v, \qquad i = A,B, \tag{66}$$

where the nonprimed fluxes refer to the laboratory frame. These equations are solved using (64) to obtain

$$v = -\frac{J'_A + J'_B}{C_A + C_B}, \tag{67}$$

or, upon resubstitution,

$$J_A = -J_B = X_B J'_A - X_A J'_B. \tag{68}$$

In the simple situation in which the off-diagonal Onsager coefficients are set equal to zero, and local equilibrium applies ($\nabla\mu_v = 0$), Eq. (13) applies. Since, moreover, for an isobaric system in which partial molar volumes $\bar{V}$ are equal,

$$\bar{V}(C_A + C_B) = 1, \tag{69}$$

combination of Eqs. (67), (68), (69), and (13) yields

$$v = \bar{V}(D_A - D_B)\nabla C_A, \tag{70}$$

$$J_A = -(X_B D_A + X_A D_B)\nabla C_A. \tag{71}$$

The value of v can be measured using inert markers.

The first demonstration of lattice drift was performed by Smigelskas and Kirkendall (1947) using the diffusion arrangement shown schematically in Fig. 9. Annealing caused rapid outward diffusion of the zinc and inward drift of the molybdenum markers. The effect is quite general and is widely used in diffusion measurements.

For an infinite diffusion couple (sample much larger than the diffusion zone), it can be shown that

$$C_A = C_A(\lambda), \qquad \lambda = x/t^{1/2}, \tag{72}$$

and hence

$$v = \frac{D_A - D_B}{t^{1/2}} \frac{dC_A}{d\lambda}. \tag{73}$$

Because the markers are located at a point of fixed composition and therefore at a fixed value of $dC_A/d\lambda$, Eq. (73) integrates immediately to yield

$$x_m = 2(D_A - D_B)\frac{dC_A}{d\lambda}t^{1/2} \tag{74}$$

for the marker displacement.

The quantities D_A, D_B are known as the *intrinsic diffusion coefficients* because they refer to diffusion with respect to the lattice planes in the presence of an activity gradient. It is necessary now to relate these to the measured *tracer* coefficients, D_{A*}, D_{B*}. Recognizing that the tracer coefficient refers to an ideal solution for which Eq. (14) applies and that the intrinsic coefficient refers to a solid solution for which Eq. (48) applies, it is found that

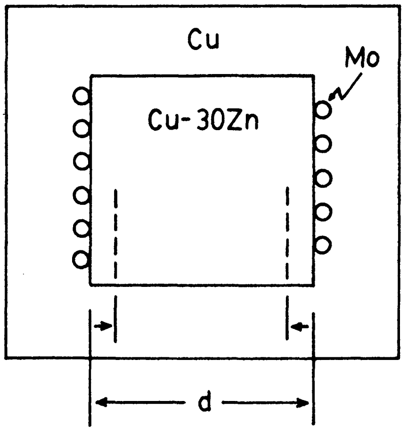

FIG. 9. Schematic representation of diffusion couple used to demonstrate lattice drift.

$$D_A = D_{A*}\left(1 + \frac{d\ln\gamma_A}{d\ln X_A}\right). \tag{75}$$

Since from the Gibbs–Duhem equation for equilibrium in a solution

$$\sum X_i d\mu_i = 0, \tag{76}$$

we may write

$$1 + \frac{d\ln\gamma_A}{d\ln X_A} = 1 + \frac{d\ln\gamma_B}{d\ln X_B}; \tag{77}$$

then Eq. (71) becomes

$$J_A = -\tilde{D}\nabla X_A, \tag{78}$$

with

$$\tilde{D} = (X_B D_{A*} + X_A D_{B*})\left(1 + \frac{d\ln\gamma}{d\ln X}\right). \tag{79}$$

This is the Darken–Hartley–Crank equation, and $\tilde{D}$ is the *chemical diffusion coefficient* (Darken, 1948; Hartley and Crank, 1949). The quantity $\tilde{D}$ is obtained from a diffusion-couple measurement, and if markers are used values of D_A, D_B may also be obtained. If the thermodynamic factor in Eq. (77) is known, then the validity of Eq. (79) can be verified. It has been found to be rigorously true for dilute solution but slightly inaccurate for rich solutions in which the correlation effects suggested by Manning (1968) may have some importance.

The above analysis has been extended to multicomponent systems (see, e.g., Kirkaldy and Young, 1987). The lack of balance among the intrinsic diffusive flows always leads to a compensating mass flow of material. That is to say, diffusional cross effects arise even in the absence of kinetic or thermodynamic correlations. Thus even a component with a negligible intrinsic mobility will move. The simple form of Fick's law fails, and the generalized form (53) must be used.

6. DIFFUSION IN IONIC SOLIDS

Self-diffusion in ionic solids is discussed elsewhere in DIFFUSION AND IONIC CONDUCTION IN SOLIDS. Attention will be confined here to chemical diffusion in the presence of activity gradients. This is technologically important, e.g., in the manufacture of ceramic materials and in the growth of oxide scales on alloys. Discussion is facilitated by the use of relative building units (Kroger *et al.*, 1959). These are used to provide a link between macroscopic thermodynamic/kinetic properties and the point-defect structure, thus allowing access to the phenomenological equations of irreversible thermodynamics.

6.1 Relative Building Units

We employ the defect notation of Kroger and Vink (1956) in which a lattice species is represented by the symbol S_m^*. Here the subscript represents normal occupancy in a perfect crystal, the principal symbol represents the species actually present, and the superscript represents the charge of the species relative to normal site occupancy: a prime for negative, a dot for positive, and a cross for zero charge. In a ternary substitutional oxide $(A,B)\mathrm{O}$ containing fully ionized cation vacancies and positive holes as the only defects, the species constituting the solid are

$$A_m^{\times},\ B_m^{\times},\ V_m'',\ \mathrm{O}_0^{\times},\ h^{\bullet}, \tag{80}$$

and thus outnumber the thermodynamic compositional variables (two independent variables for a three-component compound).

The electrochemical potential of each species is defined through Eq. (9). However, this is not operationally useful as neither η nor ψ can be measured locally. This is overcome by grouping species in charge-neutral sets. If these sets are chosen so as to preserve the cation/anion ratio characteristic of the solid, they are building units. Obvious units for $(A,B)\mathrm{O}$ are $\{A_m^{\times} + \mathrm{O}_0^{\times}\}$, $\{\mathrm{O}_0^{\times} + V_m'' + 2h^{\bullet}\}$, etc. A subset of building units is made up of relative building units that consist of the difference between two lattice species. Thus they represent compositional change arising from the replacement of one species with another and may therefore be used to describe diffusion (Young and Kirkaldy, 1984). A flux of units $\{A_m^{\times} + V_m^{\times}\}$ corresponds to diffusion of A via a vacancy mechanism; a flux of $\{A_m^{\times} - B_m^{\times}\}$ corresponds to interdiffusion.

Since relative building units conserve sites and charge, their movement necessarily satisfies the flux constraints appropriate to diffusion in the absence of an external field: Eq. (37) together with charge conservation

$$2J_{V''} = J_{h^{\bullet}}, \tag{81}$$

where the fluxes are measured in a solvent-fixed reference frame provided by an immobile anion sublattice. The constraints (37) and (81) serve both to reduce the transport equation set (8) and to eliminate dependences among the species set (80). Thus, for example, movement of a vacancy must be accompanied by movement of positive holes and an opposing cation flux. The relative building units that describe these exchanges are

$$\{A_m^{\times} - V_m'' - 2h^{\bullet}\}, \{B_m^{\times} - V_m'' - 2h^{\bullet}\}, \{A_m^{\times} - B_m^{\times}\},$$

and the third unit is merely a linear combination of the first two. The oxygen sublattice is assumed to be immobile but can still play a chemical role. The appropriate unit is $\{\mathrm{O}_0^{\times} + V_m'' + 2h^{\bullet}\}$. It is seen that combination in the appropriate proportions of the above units yields a solid $(A,B)_{1-\delta}\mathrm{O}$ of any desired degree of substitution or nonstoichiometry.

Individual lattice species lack thermodynamic significance, but a precise meaning attaches to the building units that they constitute. This is seen from the reactions leading to the introduction of defects:

$$\tfrac{1}{2}\mathrm{O}_2(g) = \mathrm{O}_0^{\times} + V_m'' + 2h^{\bullet}, \tag{82}$$

$$A(g) + V_m'' + 2h^{\bullet} = A_m^{\times}, \tag{83}$$

$$B(g) + V_m'' + 2h^{\bullet} = B_m^{\times}, \tag{84}$$

the Gibbs equilibrium equations for which yield

$$\tfrac{1}{2}\mu_{\mathrm{O}_2} = \eta(\mathrm{O}_0^{\times}) + \eta(V_m'') + 2\eta(h^{\bullet}), \tag{85}$$

$$\mu_A = \eta(A_m^{\times}) - \eta(V_m'') - 2\eta(h^{\bullet}), \tag{86}$$

$$\mu_B = \eta(B_m^{\times}) - \eta(V_m'') - 2\eta(h^{\bullet}). \tag{87}$$

Thus the relative building units have potentials corresponding to the chemical potentials of the constituent elements. Appropriate grouping of the individual species (80)—for which kinetic models can be developed—has dealt with the redundancy of the description while preserving the detailed microscopic description.

6.2 Reduction of Transport Equations

Reduced sets of phenomenological equations are found from the entropy-source expression

$$-T\sigma = J_A\nabla\eta_A + J_B\nabla\eta_B + J_v\nabla\eta_v + J_h\nabla\eta_h \tag{88}$$

by applying the constraints (37) and (81). Inspection of the latter reveals that five distinguishable pairs of species can be chosen for elimination, leading to the results

$$\begin{aligned} -T\sigma &= J_A\nabla\{\eta_A - \eta_v - 2\eta_h\} \\ &\quad + J_B\nabla\{\eta_B - \eta_v - 2\eta_h\}; \\ &= J_i\nabla\{\eta_i - \eta_j\} + J_h\nabla\{\tfrac{1}{2}\eta_v + \eta_h - \tfrac{1}{2}\eta_j\}, \\ &\qquad i,j = A,B; \\ &= J_i\nabla\{\eta_i - \eta_j\} + J_v\nabla\{\eta_v + 2\eta_h - \eta_j\}, \\ &\qquad i,j = A,B. \end{aligned} \tag{89}$$

The entropy source term has a unique value, and these expressions are equivalent. Using the defect equilibria (86), (87), and the flux constraint equations, it is readily demonstrated that each of Eqs. (89) is equivalent to

$$-T\sigma = J_A\nabla\mu_A + J_B\nabla\mu_B, \tag{90}$$

where now the chemical potentials are those of the elements.

The above procedure removes the redundancy in the species set (80). The flux equations arising from Eq. (90) are

$$\begin{aligned} J_A &= -L_{AA}\nabla\mu_A - L_{AB}\nabla\mu_B, \\ J_B &= -L_{BA}\nabla\mu_A - L_{BB}\nabla\mu_B. \end{aligned} \tag{91}$$

Because isothermal diffusion experiments conducted in the absence of an external field disclose nothing but mass transfer rates, the description is complete. It is seen that the use of relative building units is consistent with the fact that the thermodynamics and diffusion kinetics of ionic crystals can always be described in terms of elemental chemical potentials. Information on correlations among the diffusing species can be recovered from further consideration of (89).

The relative building units involving A revealed in Eqs. (89) are

$$\{A_m^\times - B_m^\times\}, \qquad \{A_m^\times + V_m'' - 2h^\bullet\}.$$

The former unit corresponds to diffusion via an exchange mechanism involving vacancies as intermediaries, but with $J_v = 0$, as illustrated in Fig. 8. A strict correlation arises through the fact that the ions share the use of the same vacancies. The second unit describes counter-current flow of material and vacancies and expresses the strict correlation with free-carrier movement. This mechanism can operate only when a gradient of vacancies is available, the effect upon which of an oxygen potential is explicit through Eq. (82). A linear combination of the fluxes arising from the movement of these units represents a complete accounting for diffusion of A, and similarly for B:

$$\begin{aligned} J_A &= -l_1\nabla(\mu_A - \mu_B) - l_2\nabla\mu_A, \\ J_B &= -\lambda_1\nabla(\mu_B - \mu_A) - \lambda_2\nabla\mu_B. \end{aligned} \tag{92}$$

Rearrangement and comparison with (91) yields

$$\begin{aligned} &L_{AA} = (l_1 + l_2), \qquad L_{BB} = \lambda_1 + \lambda_2, \\ &L_{AB} = l_1, \qquad L_{BA} = \lambda_1. \end{aligned} \tag{93}$$

Because l_1 and λ_1 refer to the exchanges $\{A^*–B^*\}$ and $\{B^*–A^*\}$, they must be identical, and symmetry of $[L]$ is assured. Similar treatments of more complex examples have been presented by Young *et al.* (1979, 1990) and Kirkaldy and Young (1987).

A practical example of the application of this methodology is the description of solid-solution oxide scale growth on alloys (Young and Gesmundo, 1988). The rate of scale growth supported by cation diffusion is expressed in terms of scale thickness, x_s:

$$\frac{dx_s}{dt} = V\sum_i J_i, \tag{94}$$

where V is the volume of oxide formed per mole of metal. Oxide diffusion data have been deduced from a number of binary-alloy oxidation experiments (Narita *et al.*, 1982). The type of data and the degree of success are illustrated in Fig. 10.

The principal value of this approach lies in equations of the type shown in (93), whereby microscopic kinetic evaluations of the l_i lead to estimates of L_{ij} and hence the D_{ij}. In this way it can be shown that for substitutional cation interdiffusion in (A,B)O when $\nabla\mu_O = 0$,

$$\tilde{D} = \frac{D_{A^*}D_{B^*}}{C_A D_{A^*} + C_B D_{B^*}}(C_A + C_B) \tag{95}$$

for the case of ideal or Henrian (see Glossary) solution. The description is easily extended to the quaternary compound (A,B,C)O. An analogous expression (79) was found for binary alloys to result from the differing intrinsic mobilities, which led to compensating bulk material flow. No such effect is possible in ionic solids, in which the anion lattice is immobile. It is instead the electrostatic potential or field (developed by charge separation within the solid) that brings into balance the fluxes of charged species having different mobilities.

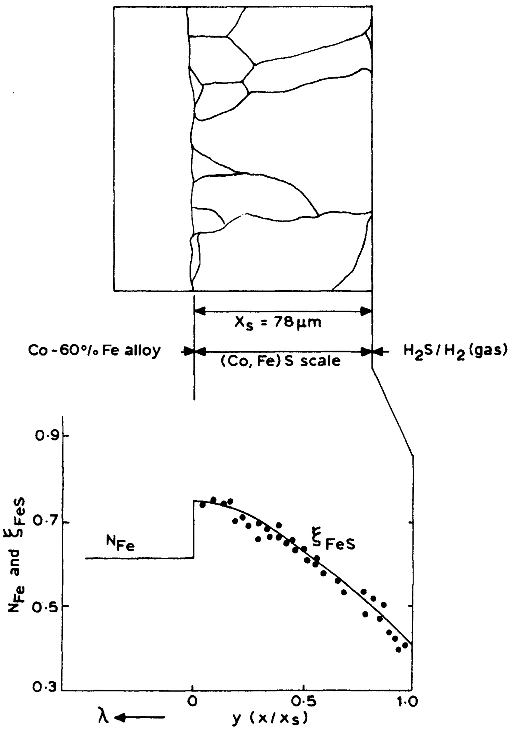

FIG. 10. (Co,Fe)S scale formed on Co–60Fe alloy at 600 °C and p_{S2} = 2 × 10^{-6} atm in 1 h. Compositions measured by electron probe microanalysis (ξ_{FeS} denotes mole fraction of FeS), and continuous line shows best fit to diffusion model. After Young *et al.* (1980). Reprinted with permission from the Electrochemical Society.

7. EXPERIMENTAL METHODS

In the most common diffusion measurements, the movement of a system toward homogeneity is observed and compared with the predictions of the diffusion equations. These equations, together with appropriate boundary conditions, yield solutions for the one-dimensional case of the general form

$$C_i = C_i(x,t,D). \tag{96}$$

Thus D is evaluated by fitting the expressions to experimental data $C_i = C_i(x,t)$. In these diffusion couple experiments, two different mixtures are brought into contact at a planar interface, and diffusion is observed along the direction normal to that interface. The method is appropriate to diffusion in the presence of a concentration gradient (chemical diffusion) and is also used for self-diffusion studies where the only gradient is that of a radiotracer.

Measurements can also be carried out on homogeneous materials by observing the interaction of the solid's constituent particles with a transient perturbation, or their subsequent relaxation to a lower energy state. Interaction or relaxations that occur via translational motion (or are thereby modified) display a time dependence determined by particle mobilities, and their measurement leads to a determination of D. Because individual particle events occur over very short ranges, these techniques are valuable for measuring very slow self-diffusion. However, because they are not sensitive to position within a sample, they cannot reveal gradient effects and will not be described here. A general review of these techniques is provided by Philibert (1991). For a more extensive review of the important nuclear magnetic resonance technique applied to solids, see Stokes (1984).

Two types of diffusion couple are important. If sample dimensions and the period of diffusion are such that concentrations at the ends of the sample do not change, then the experiment is described as an *infinite* diffusion couple, which is used to measure chemical diffusion. In a tracer diffusion measurement, the couple consists of a thin film of isotopically labeled material on the end of a long bar of unlabeled but compositionally identical material. In both cases the experiment involves preparation of the couple (while avoiding diffusion), an annealing period at constant temperature in a controlled environment, subsequent analysis of the concentration profiles, and calculation of the diffusion coefficients. Experimental aspects have been reviewed in some detail by Dunlop *et al.* (1992).

7.1 Calculation of Diffusion Coefficients

Predicted profiles of the form (96) are obtained from Fick's law [Eq. (13)], which is subject to the continuity condition

$$\frac{\partial C_i}{\partial t} = -\frac{\partial J_i}{\partial x}, \tag{97}$$

leading to Fick's second law

$$\frac{\partial C_i}{\partial t} = D\frac{\partial^2 C_i}{\partial x^2}, \tag{98}$$

where D has been approximated as constant. The solution of (98) is required for the appropriate boundary and initial conditions. Methods and a number of solutions are available from Carslaw and Jaeger (1959) and Crank (1970).

The thin-film solution applies to the tracer diffusion experiment:

$$C(x,t) = \frac{M\exp(-x^2/4D_{A*}t)}{2(\pi D_{A*}t)^{1/2}}, \tag{99}$$

where M is the amount of labeled material. With knowledge of the annealing time t, D_{A*} is evaluated from a logarithmic plot according to this equation.

For an infinite diffusion couple consisting initially of one half containing a uniform concentration C_0 and the other a concentration C_1, after diffusion time t we have

$$\frac{C(x,t) - C_0}{C_1 - C_0} = \tfrac{1}{2}[1 - \operatorname{erf}(x/2\sqrt{Dt})], \tag{100}$$

where erf is the Gaussian error function,

$$\operatorname{erf} z = \frac{2}{\sqrt{\pi}}\int_0^z \exp(-u^2)\,du. \tag{101}$$

Corresponding solutions are available for ternary systems (Kirkaldy 1958a,b). These solutions rely on D being constant, an improbable event in the presence of a concentration gradient. Instead, the analysis of Boltzmann (1894) and Matano (1933) must be used. Here the new variable $\lambda = x/\sqrt{t}$ is introduced, and the final solution is

$$\tilde{D}(c') = -\tfrac{1}{2}\left(\frac{d\lambda}{dc}\right)\int_0^{c'} \lambda\,dc, \tag{102}$$

with

$$\int_{C_0}^{C_1} x\,dc = 0 \tag{103}$$

defining the origin of coordinates. Graphical or numerical evaluations of the differential and the integral in (102) are used to evaluate $\tilde{D}(c')$. Errors are particularly large as $C \to C_0$ or C_1. For this reason, the method is best suited to diffusion zones of substantial compositional width. For narrow compositional widths, it is often better instead to approximate $\tilde{D}$ as constant and use error function solutions (see, e.g., Jan *et al.*, 1991).

For multicomponent systems where the diffusion equations (53) hold, the Matano analysis is extended (Kirkaldy, 1957) for the infinite diffusion couple:

$$\tfrac{1}{2}\int_{C_i^0}^{C_i} \lambda\,dC_i = -\sum_j^{n-1} D_{ij}\left.\frac{dC_j}{d\lambda}\right|_{C'}, \tag{104}$$

with the origin defined by

$$\int_{C_i^0}^{C_i^1} x\,dC_i = 0, \tag{105}$$

the latter condition being simultaneously satisfied for all components. Data on diffusion in ternary metallic solutions have been collected and reviewed by Kirkaldy and Young (1987) and Dayananda (1991). The difficulty of constructing couples that will satisfy Eq. (105) has prevented extension to higher-order systems.

This discussion has been based on the assumption that diffusion occurs by a random process. In real materials, certain paths such as dislocations or grain boundaries may be favored, and this will be apparent in distortions of the concentration profile. A good discussion is provided by Kaur and Gust (1989).

GLOSSARY

Building Unit: Group of lattice species that, when added to or taken from a crystal, preserve its cation/anion site ratio and its electroneutrality.

Chemical Potential: Partial molar free energy.

Correlation Coefficient: A number reflecting the degree of correlation (or departure from the purely random) between successive diffusive jumps of a tracer atom or ion.

Cross Effect: Effect of the potential gradient of one component on the diffusive flux of another.

Henrian Solution: Dilute solution in which the activity coefficient of solute is independent of concentration.

Interaction Parameter: The numerical value of the differential of the logarithm of an activity coefficient with respect to the concentration of a different component.

Mobility: Velocity of a particle moving under the influence of a field having unit magnitude.

Nonstoichiometry: Deviation in the composition of a compound from the ideal stoichiometric ratio of components, the deviation being accommodated by the presence of vacancies or interstitials. For example, $Fe_{1-\delta}O$ contains high concentrations of cation vacancies.

Off-Diagonal Terms: Terms in phenomenological diffusion equations expressing dependence of a given component flux on the potential gradients of other components. Also known as *cross terms.*

Phenomenological Equation: Equation stating linear dependence of each component flux on the chemical potential gradients of *all* components.

Point Defects: Lattice sites that differ in their occupancy or charge level from the situation in a perfect crystal.

Tracer: Radioactive isotope used to enable observation of particle movement and hence diffusion.

Works Cited

Allnatt, A. R., Allnatt, E. L. (1984), *Philos. Mag.* **A 49,** 625.

Allnatt, A. R., Okamura, Y. (1984), in: G. E. Murch, H. K. Birnbaum, J. R. Cost (Eds.), *Nontraditional Methods in Diffusion,* Warrendale, PA: AIME.

Askill, J. (1970), *Tracer Diffusion Data for Metals, Alloys and Simple Oxides,* New York: IFI/Plenum.

Bocquet, J. L., Brebec, G., Limoge, Y. (1983), in: R. W. Cahn, P. Haasen (Eds.), *Physical Metallurgy,* Part 1, Amsterdam: North Holland.

Boltzmann, L. (1894), *Ann. Phys. (Leipzig)* **53,** 90.

Brandes, E. A. (Ed.) (1983), *Smithells Metal Reference Book,* 6th ed., London: Butterworths.

Brown, A. M., Ashby, M. F. (1980), *Acta Metall.* **28,** 1085.

Brown, L. C., Kirkaldy, J. S. (1964), *Trans. AIME* **230,** 223.

Cahoon, J. R., Sherby, O. D. (1992), *Metall. Trans. A,* **23A,** 2491.

Carslaw, H. S., Jaeger, J. C. (1959), *Conduction of Heat in Solids,* Oxford: Clarendon Press.

Crank, J. (1970), *The Mathematics of Diffusion,* Oxford: University Press.

Darken, L. S. (1948), *Trans. AIME* **175,** 184.

Darken, L. S. (1949), *Trans. AIME* **180,** 430.

Dayananda, M. A. (1991), *Diffusion in Metals and Alloys,* Landolt-Bornstein, Berlin: Springer.

Dayananda, M. A. (1992), *Defect Diffus. Forum* **83,** 73.

de Groot, S. R., Mazur, P. (1962), *Non-Equilibrium Thermodynamics,* Amsterdam: North Holland.

Dignam, M. J., Young, D. J., Goad, D. W. G. (1973), *J. Phys. Chem. Solids* **34,** 1227.

Dunlop, P. J., Harris, K. R., Young, D. J. (1992), in: B. W. Rossiter, R. C. Baetzold (Eds.), *Physical Methods of Chemistry,* 2nd ed., Vol. 6, New York: Wiley.

Einstein, A. (1905), *Ann. Phys. (Leipzig)* **17,** 549.

Einstein, A. (1906), *Ann. Phys. (Leipzig)* **19,** 371.

Fick, A. E. (1855), *Ann. Phys. Chem.* **94,** 59.

Freer, R. (1980), *J. Mater. Sci.* **15,** 803.

Friauf, R. J. (1957), *Phys. Rev.* **105,** 843.

Hartley, G. S., Crank, J. (1949), *Trans. Faraday Soc.* **45,** 801.

Heumann, T. (1979), *J. Phys. F* **9,** 1997.

Howard, R. E., Lidiard, A. B. (1964), *Rep. Prog. Phys.* **27,** 161.

Jan, C.-H., Svenson, D., Zheng, X.-Y., Lin, J.-C., Chang, Y. A. (1991), *Acta Metall.* **39,** 303.

Kaur, I., Gust, W. (1989), *Fundamentals of Grain and Interphase Boundary Diffusion,* 2nd ed., Stuttgart: Ziegler Press.

Kidson, G. V. (1978), *Philos. Mag.* **37,** 305.

Kikuchi, R., Sato, H. (1969), *J. Chem. Phys.* **51,** 161.

Kirkaldy, J. S. (1957), *Can. J. Phys.* **35,** 435.

Kirkaldy, J. S. (1958a), *Can. J. Phys.* **36,** 899.

Kirkaldy, J. S. (1958b), *Can. J. Phys.* **36,** 907.

Kirkaldy, J. S., Purdy, G. R. (1969), *Can. J. Phys.* **47,** 865.

Kirkaldy, J. S., Young, D. J. (1987), *Diffusion in the Condensed State,* London: Institute of Metals.

Kirkaldy, J. S., Young, D. J., Lane, J. E. (1987), *Acta Metall.* **35,** 1273.

Kroger, F. A., Stieltjes, F. H., Vink, H. J. (1959), *Philips Res. Rep.* **14,** 557.

Kroger, F. A., Vink, H. J. (1956), *Solid State Phys.* **3,** 307.

Kubo, R. (1957), *J. Phys. Soc. Jpn.* **12,** 570.

LeClaire, A. D. (1975), in: N. B. Hannay (Ed.), *Treatise in Solid-State Chemistry,* Vol. 4, *Reactivity of Solids,* New York: Plenum.

Lidiard, A. B. (1955), *Philos. Mag.* **46,** 1218.

Manning, J. R. (1968), *Diffusion Kinetics for Atoms in Crystals,* Princeton: Van Nostrand.

Matano, C. (1933), *Jpn. J. Phys.* **8,** 109.

Mehrer, H. (1990), *Diffusion in Metals and Alloys,* Landolt-Bornstein New Series, Berlin: Springer.

Miller, D. G., Vitagliano, V., Sartorio, R. (1986), *J. Phys. Chem.* **90,** 1509.

Mott, N. F., Gurney, R. W. (1964), *Electronic Processes in Ionic Crystals,* New York: Dover.

Murch, G. E. (1984), in: G. E. Murch, A. S. Nowick (Eds.), *Diffusion in Crystalline Solids,* Orlando, FL: Academic Press.

Murch, G. E. (1991), in: R. W. Cahn, P. Haasen, E. J. Kramer (Eds.), *Materials Science and Technology,* Vol. 5, *Phase Transformations in Materials,* New York: VCH.

Narita, T., Nishida, K., Smeltzer, W. W. (1982), *J. Electrochem. Soc.* **129,** 209.

Onsager, L. (1931), *Phys. Rev.* **37,** 405; **38,** 2265.

Philibert, J. (1991), *Atom Movements Diffusion and Mass Transport in Solids,* Courtaboeuf, France: Les Editions de Physique.

Prigogine, I. (1967), *Introduction to the Thermodynamics of Irreversible Processes,* New York: Wiley-Interscience.

Purdy, G. (Ed.) (1990), *Fundamentals and Applications of Ternary Diffusion,* New York: Pergamon.

Sato, H. (1989), in: A. R. Laskar, S. Chandra (Eds.), *Superionic Conductors,* New York: Academic Press.

Sato, H., Zhang, H. (1993), in: B. Fultz, R. W. Cahn, D. Gupta (Eds.), *Diffusion in Ordered Alloys,* Warrendale, PA: TMS.

Schottky, W. (1958), *Halbleiterprobleme* **4,** 235.

Seitz, F. (1950), *Acta Crystallogr.* **3,** 355.

Sherby, O. D., Robbins, J. L., Goldberg, A. (1973), *J. Phys. Chem. Solids* **34,** 1025.

Shewmon, P. G. (1963), *Diffusion in Solids,* New York: McGraw-Hill.

Smigelskas, A., Kirkendall, E. (1947), *Trans. AIME* **171,** 130.

Smoluchowski, M. von (1908), *Ann. Phys. (Leipzig)* **25,** 205.

Stokes, H. T. (1984) in: G. E. Murch, H. K. Birnbaum, J. R. Cost (Eds.), *Nontraditional Methods in Diffusion,* Warrendale, PA: AIME.

Wagner, C. (1952), *Thermodynamics of Alloys,* Reading, MA: Addison-Wesley.

Wert, C., Zener, C. (1950), *J. Appl. Phys.* **21,** 5.

Young, D. J., Gesmundo, F. (1988), *Oxid. Met.* **29,** 169.

Young, D. J., Kirkaldy, J. S. (1984), *J. Phys. Chem. Solids* **45,** 781.

Young, D. J., Delamotte, E., Kirkaldy, J. S. (1979), *Can. J. Phys.* **57,** 722.

Young, D. J., Narita, T., Smeltzer, W. W. (1980), *J. Electrochem. Soc.* **127,** 679.

Young, D. J., Dorward, R. C., Kirkaldy, J. S. (1990), in: G. Purdy (Ed.), *Fundamentals and Applications of Ternary Diffusion,* New York: Pergamon.

Further Reading

Barrer, R. M. (1951), *Diffusion in and through Solids,* Cambridge, U.K.: Cambridge Univ. Press.

Crank, J. (1970), *The Mathematics of Diffusion,* Oxford: Oxford Univ. Press.

Denbigh, K. G. (1951), *The Thermodynamics of the Steady State,* London: Methuen.

Flynn, C. P. (1972), *Point Defects and Diffusion,* Oxford: Clarendon Press.

Kirkaldy, J. S., Young, D. J. (1987), *Diffusion in the Condensed State,* London: Institute of Metals.

Kofstad, P. (1972), *Nonstoichiometry, Diffusion and Electrical Conductivity in Binary Metal Oxides,* New York: Wiley-Interscience.

Kroger, F. A. (1973), *The Chemistry of Imperfect Crystals,* 2nd ed., Amsterdam: North Holland.

Manning, J. R. (1968), *Diffusion Kinetics for Atoms in Crystals,* Princeton, NJ: Van Nostrand.

Mrowec, S. (1980), *Defects and Diffusion in Solids,* Amsterdam: Elsevier.

Philibert, J. (1991), *Atom Movements Diffusion and Mass Transport in Solids,* Courtaboeuf, France: Les Editions de Physique.

Porter, D. A., Easterling, K. E. (1981), *Phase Transformations in Metals and Alloys,* Wokingham, Berks: Van Nostrand Reinhold.

Prigogine, I. (1955), *Introduction to the Thermodynamics of Irreversible Processes,* Springfield, IL: Charles C. Thomas.

Shewmon, P. G. (1963), *Diffusion in Solids,* New York: McGraw-Hill.

SOLIDS, DYNAMIC HIGH-PRESSURE EFFECTS IN

W. J. Nellis, *Condensed Matter Physics Division and Institute of Geophysics and Planetary Physics, Lawrence Livermore National Laboratory, University of California, Livermore, California, U.S.A.*

	Introduction	541
1.	**Scope and Background**	542
1.1	General	542
1.2	History	542
1.3	Relation to Static High Pressures	542
1.4	Scalability	543
2.	**Methods to Generate Shock Pressures**	543
3.	**Shock Compression**	544
3.1	Simple Shock Waves	544
3.2	Rankine–Hugoniot Relations and Equations of State	545
3.3	Off-Hugoniot Equations of State: Thermal Equation of State	546
3.4	Pressure Scales and Standards	547
3.5	Shock Temperatures and Optical Emission Spectra	547
3.6	Shock-Wave Profiles	547
3.6.1	Elastic–Plastic Flow and Material Strength	547
3.6.2	Shock-Induced Phase Transitions	548
3.7	Shock Release	548
3.7.1	Speed-of-Sound Measurements	549
3.7.2	Grüneisen Parameter and Phase Transitions	549
3.8	Electrical Resistivity	549
3.9	Raman Spectroscopy	549
3.10	Flash X-Ray Diffraction	549
4.	**Computational Simulations**	550
5.	**Materials**	550
5.1	Experimental Methods	550
5.1.1	Explosive Systems to Synthesize Diamond	550
5.1.2	Two-Stage Light-Gas Gun	550
5.2	Specimens	551
5.2.1	Films	551
5.2.2	Hard Materials	551
5.2.3	Bond Strengths between Film and Substrate	551
5.2.4	Single Crystals	551
5.2.5	Highly Defected Superconducting $YBa_2Cu_3O_{7-x}$	551
5.2.6	Powder Consolidation	551
5.2.7	Shock-Induced Chemical Reactions	552
6.	**Summary**	552
	Acknowledgment	552
	Appendix A: Units and Conversion Factors	552
	Glossary	552
	Works Cited	553
	Further Reading	554

INTRODUCTION

Dynamic high pressures are applied to materials to increase density and temperature, to alter crystal structure and microstructure, and to change physical and chemical properties. These effects are achieved at high pressures, and many are retained on release from high pressures. This article treats high pressures achieved dynamically by shock compression. In fact, the terms dynamic and shock are used interchangeably to describe high pressures. Shock compression is achieved by high-velocity impact, laser pulses, and explosives, for examples. Typically, shock pressure pulses have a duration of ~1 μs, although other pulse lengths can be achieved depending on how the dynamic pressure is achieved. Because of their fast adiabatic nature, dynamic high pressures cause heating of the specimen, both homogeneous and heterogeneous depending on

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

the nature of the specimen and the magnitude of the pressure. The rate of application of pressure is very high, which causes very high strain rates and the generation of very high densities of lattice defects. Shock pressures up to several hundred GPa (1 Mbar) and temperatures up to several thousand K are common. On release of pressure, very high quench rates up to 10^{12} bar/s and 10^9 K/s are achieved, which can cause retention of high-pressure phases at ambient conditions. These quench rates are the maxima in macroscopic specimens because they occur at the speed of sound. Materials made using dynamic high pressures are fine grained because of the fast time scale. Dynamic high pressures can be applied to virtually arbitrary volume, depending on the method employed.

Thus, dynamic methods

1. achieve very high pressures, densities, temperatures, strain rates, and quench rates;
2. affect crystal structure, microstructure, and properties, often uniquely;
3. treat specimen volumes of arbitrary size;
4. are used both to investigate states of matter at high pressures and to produce unusual materials after recovery from high pressures; and
5. are used to make materials that are produced commercially.

A well-known example of the last is the synthesis of fine-grained diamond powder using explosives (Cowan *et al.,* 1968).

1. SCOPE AND BACKGROUND

1.1 General

The purpose of this article is to give the reader an overview of dynamic high-pressure techniques and examples of materials that are made using dynamic high pressures. Both are important in their own right, and they are interrelated. For example, determining what occurs at dynamic high pressures can guide experiments to recover materials from dynamic high pressures.

1.2 History

The outbreak of World War II stimulated the development of shock-compression research. The primary purpose was to obtain equation-of-state data to generate material models to design devices for national defense. These same equations of state and computer codes are now used to design and interpret a wide variety of interdisciplinary high-pressure experiments, which extend well beyond the initial goals. Examples include general effects of shock waves, the nature of condensed matter at extreme pressures and temperatures, and the structure and properties of materials recovered from high shock pressures. Computations are essential for interpreting and guiding experiments. As one example of multidisciplinary activity, dynamic high-pressure experiments provide laboratory simulations to understand the deep interiors of the Earth and the outer giant planets, which are at high internal pressures and temperatures, as well as the cratered surfaces of bodies of the solar system, which were impacted at high velocities (Syono and Manghnani, 1992).

1.3 Relation to Static High Pressures

Static high pressures are achieved with presses of various types and sizes. Static compression takes seconds or longer to achieve and lasts for longer times. The difference in time duration between static and dynamic methods causes substantial differences in temperature and microstructure. For example, static high pressures typically cause larger grain size and fewer lattice defects than dynamic pressures. Static high-pressure methods, including methods to heat specimens, are described in HIGH-PRESSURE TECHNIQUES and SOLIDS, STATIC HIGH-PRESSURE EFFECTS IN; descriptions are also given by Jayaraman (1986) and by Bundy (1988).

Briefly, for small-scale scientific research at extremely high pressures, a pair of diamond anvils forced against each other is used with suitable gasketing to confine the specimen. Sample dimensions are typically 50 to a few hundred μm in diameter and up to a few tens of μm thick. Pressures range from 1 to a few hundred GPa at room temperature. Specimens are heated either by raising the temperature of the entire cell or by laser heating the specimen locally in a spot about 10 μm in diameter. Much higher pressures are achieved with laser heating that ranges up to 5000 K. Large-volume multianvil presses use much larger samples for scientific research and for commercial applications. Sample dimensions are typically a millimeter up to a centimeter or more on edge. Maximum pressures range up to about 20 GPa at room temperature and decrease as the temperature of the apparatus is increased because strength decreases with increasing temperature. If the specimens are in bulk thermodynamic equilibrium at both dynamic and static high pressures, data from the two types of experiments are related by a thermal model. Because of this, primary pressure scales at static high pressures above about 7 GPa are derived from shock-wave data.

1.4 Scalability

Results obtained using small specimens in the laboratory can be scaled up in size using different dynamic high-pressure systems. If rate effects are not important, pressures and temperatures determined with a diamond cell could also be obtained in shock experiments with larger samples. Scaling to larger specimens using dynamic high pressures can be done using existing computational and experimental shock-wave technology.

2. METHODS TO GENERATE SHOCK PRESSURES

A shock is essentially a discontinuity in flow and thermodynamic variables that propagates in a specimen at supersonic velocity (see SHOCK WAVES). For a strong shock in dense matter, the discontinuous rise in pressure can occur on a subnanosecond time scale. For a shock amplitude comparable to material strength or in a porous specimen, the rise in pressure can be tens of nanoseconds or longer and have a multiwave structure, depending on material strength, phase transitions, and porosity. Since the shock pulse is applied for short times, dynamic pressures can only cause structural changes that occur during the time of the pressure pulse.

Dynamic high pressures are generated typically by hypervelocity impact, by rapid energy release from high explosives, or by rapid energy deposition by lasers. The impact method is illustrated in Fig. 1, in which a projectile is accelerated to velocities in the range 0.1 to 8 km/s with a two-stage light-gas gun. The corresponding pressures range from 0.1 GPa to a few 100 GPa. For velocities below about 2 km/s, a single-stage gun is sufficient. The pressure pulse is generated on impact of the projectile onto the target. Much higher pressures and temperatures in the plasma range can be generated with lasers than with the other methods. Explosive and laser-generated shock generators are described in SHOCK WAVES.

Typical shock samples are tens of millimeters in diameter and a few millimeters thick for gun and explosive-driven experiments. Specimens have been as small as 100 μm on a side and 20 μm thick in gas-gun and laser experiments, as well as up to a few meters in length in explosively driven production experiments. Typical durations are

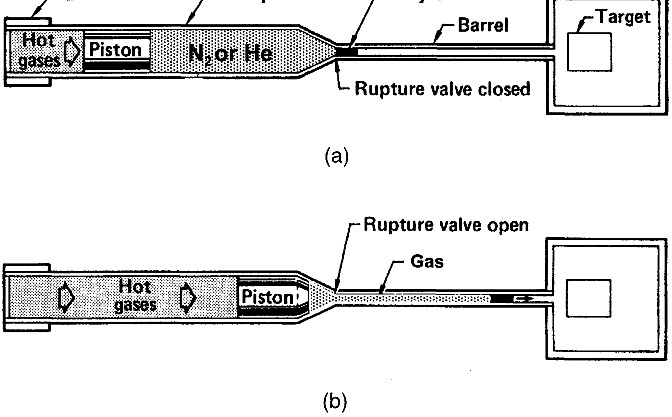

(a)

(b)

FIG. 1. Schematic of a two-stage light-gas gun. The first stage is the pair consisting of the driving gases produced by the burned gunpowder and the piston; the second stage is the pair consisting of the light gas and the projectile. The first stage compresses the light gas, which breaks the rupture valve at a predetermined pressure. Light gases have a high sound speed, which causes high projectile velocity, which in turn causes high pressure on impact of the projectile with the target. Light gases are N_2, He, and H_2, which produce maximum impact velocities of about 2, 5, and 8 km/s, respectively.

50 to 500 ns for gas-gun experiments, a few microseconds for explosives experiments, and a few nanoseconds for lasers. Small dimensions are often used for investigating effects, while large dimensions are often used for applications. Explosives are used for large-volume samples; lasers are used for surface layers.

The temperatures achieved in shock compression are determined by choice of shock pressure, specimen, and geometry. For example, shock temperatures are influenced by choice of specimen density and morphology, initial temperature, and shock-wave structure. Porous samples have lower initial and shock densities and higher shock temperatures than samples initially at crystal density. Lower initial temperatures cause lower shock temperatures. The highest temperature at a given pressure is achieved by a single shock to that pressure; a pulse with several step rises in pressure causes a lower final shock temperature than a single shock to the same final pressure.

3. SHOCK COMPRESSION

Dynamic high-pressure experiments have short time durations. Thus, fast electronic, optical, and x-ray techniques have been developed to diagnose specimen response. Since the shock process is so fast, the location of the shock wave must be synchronized with the turn-on of the recording system. Modern instrumentation makes the triggering and diagnostics possible with remarkable accuracy. Shock experiments are often performed together with theoretical calculations and have stimulated the development of condensed matter theory at high pressures (see EQUATIONS OF STATE), molecular-dynamics and Monte Carlo techniques, simulations of the dynamic response of powder particles, and hydrodynamic computer calculations. Articles in these areas are presented elsewhere (Davison and Graham, 1979; Schmidt *et al.*, 1992, 1994).

3.1 Simple Shock Waves

A simple, one-dimensional shock front typically involves an effective sharp step rise in shock pressure and in other variables in going from the unshocked to the shocked state. A simple shock front is illustrated in Fig. 2, where P_H is shock pressure, V_H is specific shock volume, E_H is specific internal shock energy, u_s is shock velocity, u_p is mass velocity behind the shock front, and x is distance. The zero-subscripted variables are initial values ahead of the shock front. The spatial width of the discontinuity of a strong shock in dense solids is several lattice spacings, which corresponds to a rise time $\ll 1$ ns. Material is shocked directly along the Rayleigh line from (P_0,V_0) to (P_H,V_H), as shown in Fig. 3. A plot of the locus of any two shock parameters is called the Hugoniot, shock adiabat, or shock-compression curve. Typical representations are u_s vs u_p, which are determined in an experiment, and P_H vs V_H, which are derived from u_s vs u_p.

The example in Fig. 2 represents a simple wave in which all effects except the equation of state are said to be overdriven and so rep-

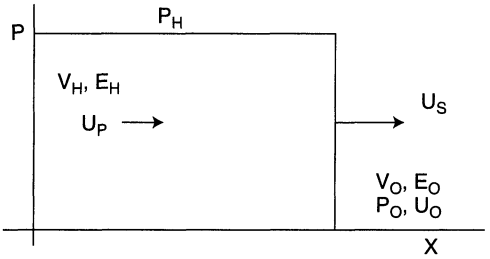

FIG. 2. Schematic of a simple shock wave traveling at velocity u_s in the x direction. This figure illustrates that the shock is essentially a traveling discontinuity. (Note: u_p and u_s are velocity vectors.)

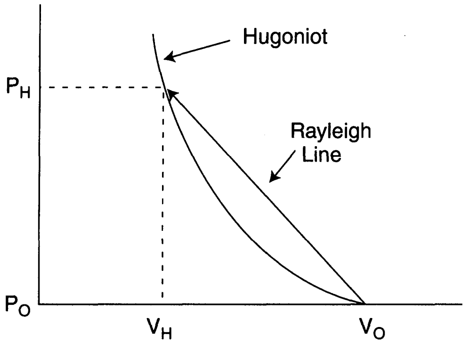

FIG. 3. Illustration in P-V space of the shock compression of a material initially at V_0 to a P-V point on the Hugoniot curve. The loading is along the Rayleigh line.

resents a relatively high-pressure wave. Figure 2 could also represent a purely elastic wave below the Hugoniot elastic limit (HEL). P_{HEL} is the maximum shock pressure at which the material acts as an elastic solid in shock compression. In Fig. 4(a) P_B is the minimum pressure required to overdrive the HEL, that is, to have a single plastic wave with no sign of the HEL in the wave profile. At intermediate pressures between the HEL and P_B, a two-wave structure exists, as shown in Fig. 4(b). In this case, a shock wave at intermediate pressure P_C splits into two waves. The leading wave at P_{HEL} is called the longitudinal elastic precursor because its wave speed is higher than that of the plastic wave at a pressure above P_{HEL}.

3.2 Rankine–Hugoniot Relations and Equations of State

The experimental design of shock-wave equation-of-state experiments is based on the Rankine–Hugoniot equations, which are derived from conservation of momentum, mass, and energy across the front of a shock wave (Fig. 2). These equations are given by

$$P_H - P_0 = \rho_0(u_s - u_0)(u_p - u_0), \tag{1a}$$

$$V_H = V_0[1 - (u_p - u_0)/(u_s - u_0)], \tag{1b}$$

$$E_H - E_0 = \tfrac{1}{2}(P_H + P_0)(V_0 - V_H), \tag{1c}$$

where $\rho_H = 1/V_H$ is mass density and the other variables are defined in Sec. 3.1. For a single wave, u_s is often measured; u_p is measured or is calculated using the shock-impedance-match principle, which states that P_H and u_p are continuous across an interface. Shock impedance is defined to be $\rho_0 u_s$, the slope of P_H vs u_p from Eq. (1a), where $P_0 = u_0 = 0$. For the case of a symmetric impact in which an impactor of a given material is accelerated to velocity u_I and strikes a target of the same material, $u_p = u_I/2$ by symmetry. Equations 1 define all primary pressure scales above about 7 GPa, both static and dynamic.

For many solids in the absence of phase transitions, experimental Hugoniot data follow the relation

$$u_s = C + Su_p, \tag{2}$$

where C is bulk sound speed at $P = 0$ and S is the slope of the linear relation between u_s and u_p. The constant S is a function of the pressure derivative of the isentropic bulk modulus (Ruoff, 1967). This result can be used to calculate P_H, V_H, and E_H from Eqs. (1). For comparison, the experimental Hugoniot curve and the theoretical 0-K isotherm and isentrope of Al are plotted in Fig. 5 (McMahan, 1976). For strong shocks, the total specific internal energy of Eq. (1c) becomes $E = \frac{1}{2}P_H(V_0 - V_H)$. As shown in Fig. 5, the total specific internal energy is represented by the area of the triangle under the Rayleigh line. Internal energy caused by isentropic compression to the same volume is reversible and is represented by the area under the isentrope. Reversible internal energy is removed from the sample when pressure is released. Irreversible internal energy is, thus, represented by the hatched area in Fig. 5. This irreversible internal energy is the reason for shock-induced heating and is the reason why shock pressures cause higher thermal and total pressures than isothermal,

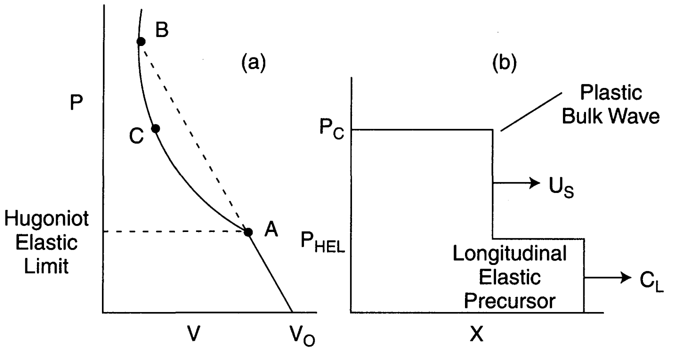

FIG. 4. (a) Hugoniot of a solid with a Hugoniot elastic limit, $P_A = P_{\mathrm{HEL}}$. P_B is the minimum shock pressure required to overdrive the HEL, that is, to achieve a single shock wave. (b) At intermediate pressure P_C, a two-wave structure propagates. The elastic precursor travels at the longitudinal sound speed C_L, followed by a plastic shock traveling at shock velocity u_s.

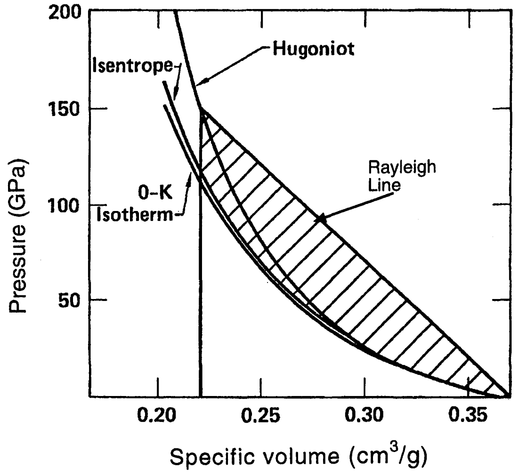

FIG. 5. Illustration of irreversible internal energy (shaded region) achieved in shock compression. This example is for Al shocked to 150 GPa and shows the measured Hugoniot and calculated 0-K isotherm and the isentrope from crystal volume (after McMahan, 1976).

static compression to the same volume. In addition to heating materials, irreversible shock energy is also available for inducing phase transitions, changing microstructures, causing chemical reactions, and absorption by other internal degrees of freedom.

Single-shock equation-of-state experiments are illustrated in Fig. 6(a), in which a Ta impactor strikes a stepped Al target. Impactor velocity u_I and shock velocity u_s in the target are obtained by measuring the time between flash x rays of the impactor and the shock-transit time across the measured step height. u_I and u_s and the shock-impedance-match principle determine (P_H, V_H, E_H) in the target with $u_0 = 0$ in Eqs. (1). Double-shock equation-of-state experiments are illustrated in Fig. 6(b), in which a Ta impactor strikes a soft Al target backed by a stiff Ta anvil. After the first shock traverses the Al, it reflects backward off the stiffer Ta anvil, double shocking the Al and sending a shock forward into the Ta anvil. By taking the first-shock state as the zero-subscripted variables in Eqs. (1), measuring u_I, and measuring u_s in the anvil, the second-shock state can be obtained from Eqs. (1). Double shocking to a given volume produces lower pressures and temperatures than a single shock to the same volume.

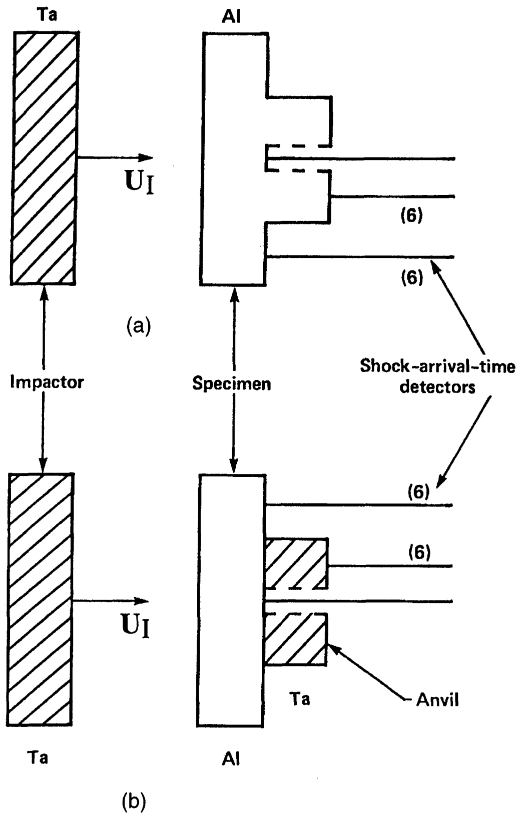

FIG. 6. Illustration of (a) single-shock and (b) double-shock equation-of-state experiments.

Hugoniot equation-of-state data for Al, Cu, Mo, Ta, and Pt, for examples, have been measured up to a few hundred GPa using a two-stage light-gas gun at higher pressures and a planar explosive system at lower pressures (Marsh, 1979; van Thiel, 1977). Maximum pressures in these experiments are 170, 330, 470, 560, and 660 GPa, respectively (Mitchell and Nellis, 1981; Holmes *et al.*, 1989).

3.3 Off-Hugoniot Equations of State: Thermal Equation of State

Shock-compression equation-of-state data provide a reference curve for calculating off-Hugoniot states, that is, states that are not on the Hugoniot. Below about 100 GPa, thermal pressure at a given volume can be calculated from the Grüneisen model, which assumes that $(\partial P/\partial E)_V = \gamma/V$ is a constant, where γ is the Grüneisen parameter, dependent only on V. The Grüneisen model assumes that thermal energy is absorbed by the lattice. γ is obtained by measuring various discrete (P,E) states at the same V and approximating differentials with differences. The latter can be obtained by single- and double-shock data to the same volume using both solid and porous specimens. Porous specimens, with lower initial densities than solid ones, are more compressible and thus have larger internal energies and thermal pressures at a given volume than do shock-compressed crystal-density specimens. Equation-of-state data for porous and solid specimens of Al, Cu, and Fe suggested that γ/V is constant (Carter *et al.*, 1971), which is a common form used for γ for most materials. Also, γ has been determined from sound speed measurements, described in Sec. 3.7. Heat capacities are estimated to calculate temperatures from internal energies.

The latest approach to theoretical calculations of the equation of state of metals at high pressures and temperatures has been developed by Moriarty (Mitchell *et al.*, 1991). Electron band theory is used to calculate the 0-K isotherm in P-V space, to which are added electron-thermal and ion-thermal contributions to the pressure as functions of V and T.

3.4 Pressure Scales and Standards

Shock data are the primary data for static pressure scales above about 7 GPa. At higher static pressures the optical shift of ruby fluorescence and x-ray diffraction densities are calibrated vs shock-wave–reduced P-V isotherms. A shock-wave–reduced isotherm is that in which internal shock energies are taken into account to derive a 300-K isotherm from the Hugoniot curve, as described in Sec. 3.3.

Since static-pressure diamond-anvil-cell experiments are now up to the 500-GPa range (Ruoff *et al.*, 1990), accurate 300-K and higher-temperature P-V isotherms are needed for static pressure scales to ultrahigh pressures. Common x-ray standards for static high-pressure experiments are Pt and Mo because of their high atomic numbers and strong x-ray scattering. Powders of these metals are x rayed at very high static pressures to obtain sample volume. The corresponding pressure is derived from shock-reduced isotherms by the methods in Sec. 3.3. The optical ruby-fluorescence shift is a pressure scale calibrated by this x-ray diffraction method (see SOLIDS, STATIC HIGH-PRESSURE EFFECTS IN).

3.5 Shock Temperatures and Optical Emission Spectra

Shock temperatures have been measured in shocked transparent solid specimens by fitting the spectral dependence of radiation emitted from the shock front to a blackbody or graybody emission spectrum. If the specimen is transparent, the optical emission is observed for several times 100 ns while the shock front transits the specimen. The effective emission temperature can be taken in many cases as the shock temperature of the specimen. Results for transparent CsI are an example (Radousky *et al.*, 1985). Because of their thin optical depths of only about 100 Å and $<10^{-12}$ s, which precludes accurate measurements, shock temperatures of opaque solids are calculated as in Sec. 3.3. In one case, shock temperatures of metallic Fe have been measured by observing optical radiation from partially released Fe through a transparent diamond window, taking into account shock-wave reflections and thermal transport at the Fe–window interface (Yoo *et al.*, 1993). Shock temperatures of powders have also been measured (Boslough, 1992).

3.6 Shock-Wave Profiles

In shocked solids at pressures just above the HEL or, similarly, just above phase-transition pressures, a shock can have a multiple-wave structure (Graham and Asay, 1978). By measuring time-resolved shock-wave profiles, equations of state, mechanical constitutive properties and phase transition pressures under dynamic loading are derived.

3.6.1 Elastic–Plastic Flow and Material Strength Ideal elastic–plastic flow is illustrated in Fig. 4. For $P_A = P_{\mathrm{HEL}} < P_C < P_B$, a two-wave structure propagates. Because elastic wave speeds are greater than plastic ones, the elastic wave propagates as a precursor ahead of the plastic wave. The HEL is typically about a factor of 2 greater than the yield strength. The HELs of Cu, hardened steel, alumina, and B_4C are about 0.3, 2, 10, and 15 GPa, respectively. HELs are directionally dependent in anisotropic crystals.

A solid material with strength supports shear stress. As a result, the Hugoniot lies above the hydrostat in *P-V* space for a strong solid, stress is a more accurate representation than pressure, and stress is different in different directions. The principal stresses σ_i and the hydrostatic pressure P are given by

$$\sigma_i = -P + s_i, \quad i = 1, 2, 3 \tag{3}$$

and

$$-P = \tfrac{1}{3}(\sigma_1 + \sigma_2 + \sigma_3), \tag{4}$$

where the s_i are the stress deviators in the principle directions i and $\Sigma s_i = 0$. The yield strength Y^0 is such that

$$s_1^2 + s_2^2 + s_3^2 \leq (Y^0)^2. \tag{5}$$

For a perfectly plastic material, Y^0 is a constant independent of pressure. For many materials, called elastic plastic, work hardening increases yield strength Y for shock stresses above the HEL. That is, as shock pressure increases the amount of plastic work and associated shock temperature, Y increases to a maximum and then decreases until $Y = 0$ at melting, where individual stress deviators $s_i = 0$ in the fluid state. For Cu, Y^0 is about 0.1 GPa, and $Y_{\max}$ is about 2.5 GPa at 70-GPa shock stress and 1000-K shock temperature (Morris and Fritz, 1980). Cu melts at a shock pressure of about 230 GPa (Moriarty, 1986).

3.6.2 Shock-Induced Phase Transitions

Phase transitions can be induced by dynamic high pressures and temperatures. If crystal volume changes in the transition, then pressure and material velocity also change. By measuring time-resolved pressure or mass-velocity histories, phase changes can be detected. The graphite-to-diamond transition is an example. The *P-V* curve and wave structure are similar to those in Fig. 4, except that in this case P_A is the onset phase transition pressure, rather than the HEL. The measured velocity of an interface between shocked graphite and a LiF window is shown in Fig. 7. The first wave, or velocity jump, is graphite in the shock-compressed state at 21 GPa. The second wave is caused by the transformation to diamond, which occurs in 10 ns in the highly ordered pyrolytic graphite specimen, oriented with the *c* axis of the hexagonal-graphite crystal structure parallel to the shock direction. The fast speed of the phase transition is the signature of a martensitic transformation from hexagonal graphite probably to hexagonal diamond (Erskine and Nellis, 1992). The kinetics of this transformation are quite sensitive to the degree of crystalline order and to the relative orientation of the specimen with respect to the direction of shock propagation. The velocity profile of Fig. 7 was measured using a fast optical interferometer called a VISAR, or velocity interferometer for a surface of any reflector (Barker and Hollenbach, 1972). Shock-wave profiles can also be measured with a Fabry–Pérot interferometer or with piezoelectric gauges.

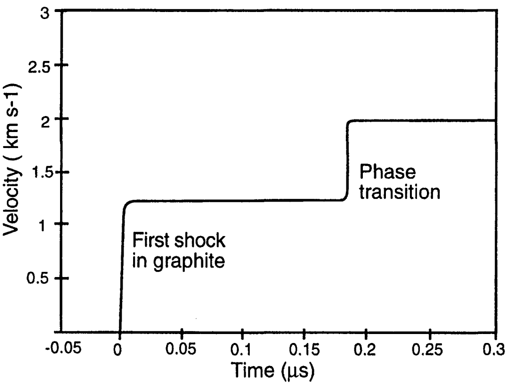

FIG. 7. Measured wave profile in highly oriented pyrolytic graphite shocked into the region of the phase transition to diamond. In this case, the first knee in the curve corresponds to the onset of the phase transition, rather than to the HEL illustrated in Fig. 4 (after Erskine and Nellis, 1992).

Melting is also induced by high shock pressures and temperatures. Melting temperatures increase with pressure. However, shock temperatures also increase with pressure and eventually exceed the melting temperature at sufficiently high pressures (Moriarty, 1986).

3.7 Shock Release

A calculated temporal wave profile of a strong planar shock in Cu is shown in Fig. 8 (Neumeier *et al.*, 1989). This shock is in-

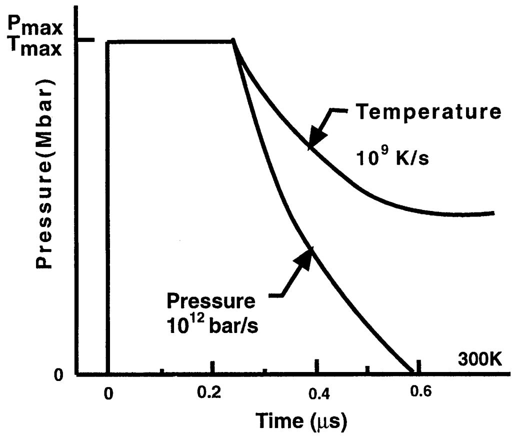

FIG. 8. Shock and release profiles calculated in planar geometry for Cu shocked to high pressure. The pressure releases to zero, but the temperature does not because of the irreversible internal shock energy shown in Fig. 5 (after Neumeuier *et al.*, 1989).

duced by impact; a shock is also induced in the impactor and travels back toward the rear surface of the impactor (Fig. 6). On reaching the rear surface of the impactor, pressure drops, and a release wave is transmitted forward into the impactor, eventually overtaking and releasing the original shock wave as it travels in the target. At overtake or release, both pressure and temperature decrease. The release rates are about 10^{12} bar/s and 10^9 K/s. While pressure releases to zero quickly, temperature releases to a residual value, which is higher than the initial value because of irreversible shock heating (Fig. 5). The residual temperature approaches the ambient value by thermal conduction.

3.7.1 Speed-of-Sound Measurements The high-pressure release wave described in Sec. 3.7 travels longitudinally at the speed of sound c in the material, which is traveling at mass velocity u_p in the laboratory frame. Thus, the release wave travels at velocity $u_p + c$, which is greater than u_s. The longitudinal speed of sound C_L of a high-pressure shock state is determined by measuring the depth in the target at which the release wave just overtakes the shock, provided u_s and u_p are known (Brown and McQueen, 1986).

3.7.2. Grüneisen Parameter and Phase Transitions For the case in which shear strength is negligible, the bulk thermodynamic speed of sound C_B is given by

$$C_B = V_H\left\{\left(\frac{\partial P}{\partial V}\right)_H\left[(V_0 - V_H)\left(\frac{\gamma}{2V_H}\right) - 1\right] + P_H\left(\frac{\gamma}{2V_H}\right)\right\}^{1/2}, \qquad (6)$$

where the partial derivative is taken along the Hugoniot. For this case, C_B is assumed equal to C_L, the measured longitudinal sound speed. Thus, sound speeds measured in the shock state in the fluid can be used to calculate $\gamma(V)$, which can be used to calculate off-Hugoniot states, as in Sec. 3.3.

For the case in which a solid has shear strength, we have

$$C_B^2 = C_L^2 - \tfrac{4}{3}C_T^2, \qquad (7)$$

where C_T is the sound speed of the shear wave. Melting is often taken to occur at the shock pressure at which $C_L = C_B$, where C_L is measured and C_B is calculated from Eq. (6). Because the speed of sound is related to a derivative of pressure with respect to volume, sound speed is usually more sensitive to phase transitions than Hugoniot P-V relations.

3.8 Electrical Resistivity

Electrical resistivities are measured up to the 100-GPa range. The resistivity of FeO at 100 GPa shock pressure is about 100 μΩ-cm (Knittle *et al.*, 1986), which is in the range for metallic Fe at 100 GPa.

3.9 Raman Spectroscopy

Raman spectroscopy has been used to observe shifts in the vibrational frequency of diamond with increasing shock pressure (Gupta *et al.*, 1989).

3.10 Flash X-Ray Diffraction

Flash x-ray diffraction has been performed on graphite, BN, and LiF single crystals at shock pressures up to 100 GPa. The observation of diffraction shows that the material has long-range crystalline order, despite high densities of shock-induced lattice defects. The experimental x-ray density is observed to be equal to the Hugoniot density [Eq. (1b)] up to 100 GPa shock pressures for LiF (Johnson and Mitchell, 1980).

4. COMPUTATIONAL SIMULATIONS

The experimental data described above are often used to develop material databases, such as equations of state and constitutive properties, for implementation into computer codes. These codes are often called hydrodynamic, even for solids that have strength. Hydrodynamic codes are used to simulate, design, and interpret a variety of dynamic high-pressure experiments. In addition, molecular-dynamics, Monte Carlo, and electron-energy-band techniques are now used to understand dynamic high-pressure phenomena on an atomic or molecular scale (Schmidt *et al.*, 1992).

5. MATERIALS

Dynamic pressures discussed above are diagnosed and simulated computationally in real time. In this section, we discuss how dynamic high pressures are used to produce novel crystal structures, microstructures, and associated properties. That is, specimens are subjected to dynamic pressures and recovered intact for characterization. Pressure is applied at strain rates up to 10^8/s and higher, and materials are shock compressed to pressures as high as 100 GPa and temperatures as high as a few times 1000 K, held at pressure for times on the order of microseconds, quenched at rates up to 10^{12} bar/s and 10^9 K/s (Fig. 8), and recovered for characterizations of structure and properties. These quench rates occur at the physical limit for macroscopic samples because they occur in bulk at the speed of sound. Even higher quench rates are achieved by thermal diffusion in micron-thin layers at interfaces. These high quench rates mean that phases and structures synthesized at high shock pressures and temperatures might be quenched at ambient conditions. Materials treated this way often have unique structures and tailored properties.

Effects can be investigated in small samples using, for example, a small gas gun or a laser-heated diamond anvil cell. Larger specimens can be produced by using larger dynamic systems, such as explosives or large lasers (see SHOCK WAVES; Cowan *et al.*, 1968).

5.1 Experimental Methods

Materials are recovered from dynamic high pressures using a variety of methods. As examples, we discuss diamond powders made with explosives and a method to investigate effects in small specimens using a gas gun.

5.1.1 Explosive Systems to Synthesize Diamond Probably the best known example of materials produced commercially with dynamic pressure is the synthesis of diamond powders for abrasives (Cowan *et al.*, 1968). This system is about 4 m long and about 1 m in diameter. A mixture of carbon and metal powders is placed inside a cylindrical explosive system, which drives the carbon into the pressure–temperature stability field of diamond. Polycrystalline diamond powders are synthesized from the carbon, which is quenched thermally by the metal powders before the pressure is released. Because the process is fast, polycrystalline diamond is produced with a range of particle sizes of about 0.1 to 40 μm.

Nanocrystalline diamond powders are produced by the nucleation and growth of diamond within the reaction products of detonated explosives. Carbon is a common component, or carbon additives are mixed with explosives. Detonation takes the reaction products to the pressure-temperature stability field of diamond, which forms on phase separation of carbon from the other reaction products. The fast pressure-release path quenches the diamond formed at high pressure. Typical grain size is about 4 nm, small because of the microsecond time scale (Lyamkin *et al.*, 1988). These grains agglomerate into particles about 0.1 μm across. This work was based on the theoretical analysis of the Hugoniot data of hydrocarbons, which assumed that hydrocarbons decompose into diamond-like carbon and molecular hydrogen at high shock pressures and temperatures (Ree, 1979).

5.1.2 Two-Stage Light-Gas Gun Effects of dynamic high pressures in small specimens are investigated using a two-stage light-gas gun (Fig. 1). A shock is generated on impact of a projectile with a target capsule in which a specimen is embedded. Typical dimensions of a gun for this purpose are a length of 6.5 m and a bore diameter of 20 mm. Maximum impactor velocity is 4 km/s with He driving gas, which produces 120 GPa (1.2 Mbar) for a Cu plate impacting a Cu target. Velocities up to 7 km/s could be achieved with H_2 driving gas. Specimens are typically 10 mm in diameter and 0.001 to 1 mm thick. Initial temperatures can be varied in the range 100–1300 K. A representative configuration is illustrated in Fig. 9.

5.2 Specimens

Various types of specimens are subjected to dynamic high pressures and characterized. A few are described below, most of which are shocked with a two-stage gun.

5.2.1 Films Films are used because a wide variety of materials can be made with this method, mixing is on a fine scale for fast reactions during fast pressure pulses, and maximum quench rates are achieved in thin layers that can retain shock-induced structures and properties. To illustrate that films can be recovered successfully from high dynamic pressures, Nb films 1–10 μm thick were recovered from 100 GPa shock pressure (Koch *et al.*, 1990).

5.2.2 Hard Materials Potentially new hard materials can be synthesized and recovered from dynamic high pressures. For example, shock-compressed mixtures of C_{60} fullerenes and Cu powders produced fine-grained diamond (Sekine, 1992). Nanocrystalline metastable diamond films have been produced by shock compression of 2-μm-thick C_{60} fullerene films (Yoo *et al.*, 1992).

5.2.3 Bond Strengths between Film and Substrate The strength of bonding between a film made by conventional means and its substrate is measured by launching a short-pulse planar shock into a substrate, which then transits the film. The shock reflects as a tensile wave from the free surface of the film and travels back in the direction from which the original shock came. The tensile stress required to delaminate the film from the substrate is determined from the shape of the time-resolved free-surface velocity history of the film, which is measured with a fast optical velocimeter similar to the design of Barker and Hollenbach (1972). Short-pulse incident shock waves are generated by impact of thin metal foils (Nutt, 1992) or by a laser (Gupta *et al.*, 1993).

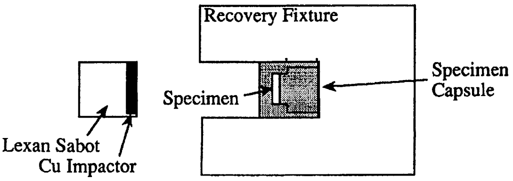

FIG. 9. Illustration of an experiment to recover a specimen shocked to high pressure. The projectile is launched by the gun in Fig. 1.

5.2.4 Single Crystals Three-phase SiO_2 has been synthesized above 20 GPa by shocking single-crystal quartz, which fractures into grains. The three phases are lamellar glass and quartz within grains and melt glass between grains (Gratz *et al.*, 1992). Above about 40 GPa, only amorphous SiO_2 is found.

5.2.5 Highly Defected Superconducting $YBa_2Cu_3O_{7-x}$ Shock deformation in ceramics induces dislocations and stacking faults, which enhance magnetic hysteresis and magnetic levitation force in superconducting $YBa_2Cu_3O_{7-x}$. By using highly ordered melt-textured specimens and orienting them with respect to the shock direction, these brittle oxide specimens can be shocked without macroscopic fracture. After annealing in oxygen, the magnetic hysteresis at 1 kOe and 70 K is about 20% greater than the value before shock at 7 GPa (Nellis *et al.*, 1994).

5.2.6 Powder Consolidation Dynamic compaction rapidly consolidates powders by depositing compressive energy on particle surfaces, which are heated heterogeneously, often bond together, and then quench thermally to the interiors of the particles, forming a dense compact. Advantages include

1. focussing energy on particle surfaces so that whole systems do not need to be heated for compaction,
2. consolidating nanocrystalline powders so quickly that grains do not have time to grow, and
3. bonding powders of composite materials.

A common problem with this method is cracking that occurs on release of pressure. Single-piece disks of both conventional and rapidly solidified alumina–zirconia ceramics have recently been consolidated by a reverberating shock wave (Freim *et al.*, 1994). Diamond particles have been shock consolidated. For example, good results have been obtained by choice of diamond particle size and a relatively small admixture of graphite particles (Potter and Ahrens, 1988). A computational model is being developed for dynamic compaction that could be applied to a wide variety of materials, pressures, and consolidation rates and, thus, address key issues computationally (Benson and Nellis, 1994).

5.2.7 Shock-Induced Chemical Reactions Dynamic high pressures and temperatures cause chemical reactions between powder particles and, thus, offer a way to synthesize novel materials. Recent results are described in Schmidt *et al.* (1994). Shock-induced defects also enhance chemical reactivity and catalysis (Graham *et al.*, 1986; Thadhani, 1993).

6. SUMMARY

High dynamic pressures achieve unusual states of matter through changes in density and temperature, which cause changes in properties at high pressures. In recovered specimens, high dynamic pressures, densities, temperatures, and strain rates can cause changes in microstructures and phases, which in turn can change properties of materials. An extensive array of fast diagnostics have been developed to probe matter at high dynamic pressures. Computational and theoretical methods interpret results and are used to design dynamic high-pressure experiments. The fast real-time experiments are important for validating computational models and for determining materials that can be recovered from dynamic high pressures. A wide variety of materials have been recovered intact for characterization of structure and properties. Material produced with dynamic high pressures are produced commercially, such as synthetic diamond powders.

ACKNOWLEDGMENT

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

APPENDIX A: UNITS AND CONVERSION FACTORS

The SI system provides five basic units used in high-pressure research: meter (m), kilogram (kg), second (s), ampere (A), kelvin (K), and mole (mol). Pressure is the derived quantity of force per unit area. Many kinds of units of pressure are used in engineering and science according to the particular traditions. In this article we use the units GPa (Gigapascal) and kilobar (kbar); 1 GPa = 10 kbar, where 1 GPa = 1 N $m^{-2} \times 10^9$ and 1 kbar is the pressure of about 1000 atmospheres. Units of density are 1 g cm^{-3} = 1 kg m^{-3}.

GLOSSARY

Bond Strength: Strength of the bond between a film and its substrate. One technique to measure this is to measure the strength of the tensile wave, induced by a shock pulse, required to break this bond.

Dynamic Powder Compaction: Compaction of a powder specimen by a dynamic pressure pulse.

Dynamic Strength: The strength of a material at dynamic high pressure. Strength is generally higher for short pressure pulses than for static pressures.

Elastic–Plastic Flow: Flow in which a material first compresses elastically to the Hugoniot elastic limit and plastically at higher shock pressures.

Fabry–Pérot Interferometer: A velocity interferometer that measures the time dependence of the velocity of a surface.

Flash X-Ray Diffraction: Pulsed x-ray diffraction from a material at high shock pressure.

High Strain Rate: The strain rate $\dot{V}V^{-1}$ achieved by shock compression ranges up to $>10^8$ s^{-1}, the physical limit in macroscopic bodies.

Hugoniot: Locus of states reached by shock compression from a given initial condition. Common representations include pressure vs volume, shock velocity vs mass velocity, and pressure vs mass velocity.

Nanocrystalline Diamond: Grains of diamond 1–20 nm across, produced by high shock pressures. The brief time duration (of order microseconds) restricts particle growth to small dimensions.

Pulsed Raman Spectroscopy: Light scattering by vibrational modes at high shock pressures.

Quench Rates: Rates at which pressure and temperature decrease when shock pressure releases. These rates, about 10^{12} bar/s and 10^9 K/s, are the physical limit in macroscopic bodies.

Shock-Induced Chemical Reactions: Chemical reactions caused by high shock pressures, densities, temperatures, and defect concentrations. Synthesis of new phases from elemental powders is an example.

Shock-Induced Chemical Reactivity: Chemical reactivity induced by shock-induced defects. Enhanced catalytic activity is an example.

Shock-Induced Defects: Lattice defects, usually dislocations and stacking faults, induced in a lattice by the high–strain-rate deformation produced by shock compression.

Shock Wave: A discontinuity in flow and thermodynamic variables that propagates at supersonic velocity.

Single-Stage Gun: A gun consisting of a projectile and a breech in which the projectile is accelerated by compressed gas or by burned gunpowder in the breech.

Two-Stage Light-Gas Gun: A gun consisting of two stages of driving gas and piston. The first stage consists of the products of burned gunpowder and a heavy piston; the second stage consists of light gas and a light projectile. The light gas and the projectile are separated by a nozzle that amplifies the gas velocity and by a valve that opens when the gas pressure exceeds a given value. The light gas has a high sound speed and, thus, produces a high velocity when it accelerates the projectile.

VISAR: A velocity interferometer for a surface of any reflector, which measures the time dependence of the velocity of the surface.

Works Cited

Barker, L. M., Hollenbach, R. E. (1972), *J. Appl. Phys.* **43,** 4669–4675.

Benson, D. J., Nellis, W. J. (1994), *Appl. Phys. Lett.* **65,** 418–420.

Boslough, M. B. (1992), in: S. C. Schmidt, R. D. Dick, J. W. Forbes, D. G. Tasker, (Eds.), *Shock Waves in Condensed Matter 1991,* Amsterdam: Elsevier Press, pp. 617–620.

Brown, J. M., McQueen, R. G. (1986), *J.Geophys. Res.* **91,** 7485–7494.

Bundy, F. P. (1988), *Phys. Rep.* **167,** 618–630.

Carter, W. J., Marsh, S. P., Fritz, J. N., McQueen, R. G. (1971), *Accurate Characterization of the High-Pressure Environment,* Washington, DC: National Bureau of Standards, pp. 147–158.

Cowan, G. R., Dunnington, B. W., Holtzman, H. A. (1968), U.S. Patent No. 3,401,019.

Davison, L., Graham, R. A. (1979), *Phys. Rep.* **55,** 255–379.

Erskine, D. J., Nellis, W. J. (1992), *J. Appl. Phys.* **71,** 4882–4886.

Freim, J., McKittrick, J., Nellis, W. J. (1994), in: S. C. Schmidt, J. W. Shaner, G. A. Samara, M. Ross, (Eds.), *High-Pressure Science and Technology—1993,* New York: American Institute of Physics Press, pp. 1263–1266.

Graham, R. A., Asay, J. R. (1978), *High Temperatures—High Pressures* **10,** 355–390.

Graham, R. A., Morosin, B., Venturini, E. L., Carr, M. J. (1986), *Annu. Rev. Mater. Sci.* **16,** 315–341.

Gratz, A. J., Nellis, W. J., Christie, J. M., Brocious, W., Swegle, J., Cordier, P. (1992), *Phys. Chem. Minerals* **19,** 267–288.

Gupta, Y. M., Horn, P. D., Yoo, C. S. (1989), *Appl. Phys. Lett.* **55,** 33–35.

Gupta, V., Yuan, J., Martinez, D. (1993), *J. Am. Ceram. Soc.* **76,** 305–315.

Holmes, N. C., Moriarty, J. A., Gathers, R. G., Nellis, W. J. (1989), *J. Appl. Phys.* **66,** 2962–2967.

Jayaraman, A. (1986), *Rev. Sci. Instrum.* **57,** 1013–1031.

Johnson, Q., Mitchell, A. C. (1980), *High Pressure Science and Technology,* Oxford: Pergamon, pp. 977–978.

Knittle, E., Jeanloz, R., Mitchell, A. C., Nellis, W. J. (1986), *Solid State Comm.* **59,** 513–515.

Koch, R., Nellis, W. J., Hunter, J. W., Davidson, H., Geballe, T. H. (1990), *Pract. Met.* **27,** 391–405.

Lyamkin, A. I., Petrov, E. A., Ershov, A. P., Sakovich, G. V., Staver, A. M., Titov, V. M. (1988), *Dokl. Akad. Nauk SSSR* **302,** 611–620 (in Russian).

Marsh, S. P. (1979), *Los Alamos Shock Hugoniot Data,* Berkeley: University of California Press.

McMahan, A. K. (1976), *Bull. Am. Phys. Soc.* **21,** 1303.

Mitchell, A. C., Nellis, W. J. (1981), *J. Appl. Phys.* **52,** 3363–3374.

Mitchell, A. C., Nellis, W. J., Moriarty, J. A., Heinle, R. A., Holmes, N. C., Tipton, R. E. (1991), *J. Appl. Phys.* **69,** 2981–2986.

Moriarty, J. A. (1986), *Shock Waves in Condensed Matter,* New York: Plenum, pp. 101–106.

Morris, C. E., Fritz, J. N. (1980), *J. Appl. Phys.* **51,** 1244–1246.

Nellis, W. J., Weir, S. T., Hinsey, N. A., Balachandran, U., Kramer, M. J., Raman R. (1994), in:

S. C. Schmidt, J. W. Shaner, G. A. Samara, M. Ross, (Eds.) *High-Pressure Science and Technology—1993,* New York: American Institute of Physics Press, pp. 695–697.

Neumeier, J. J., Nellis, W. J., Maple, M. B., Torikachvili, M. S., Yang, K. N., Ferreira, J. M., Summers, L. T., Miller, J. I., Sales, B. C. (1989), *High Pressure Res.* **1,** 267–289.

Nutt, G. L. (1992), *J. Mater. Res.* **7,** 203–213.

Potter, D. K., Ahrens, T. J. (1988), *J. Appl. Phys.* **63,** 910–914.

Radousky, H. B., Ross, M., Mitchell, A. C., Nellis, W. J. (1985), *Phys. Rev. B* **31,** 1457–1462.

Ree, F. H. (1979), *J. Chem. Phys.* **70,** 974–983.

Ruoff, A. L. (1967), *J. Appl. Phys.* **38,** 4976–4980.

Ruoff, A. L., Xia, H., Luo, H., Vohra, Y. K. (1990), *Rev. Sci. Instrum.* **61,** 3830–3833.

Schmidt, S. C., Dick, R. D., Forbes, J. W., Tasker, D. G. (1992), *Shock Compression of Condensed Matter—1991,* Amsterdam: Elsevier Press.

Schmidt, S. C., Shaner, J. W., Samara, G. A., Ross, M. (1994), *High-Pressure Science and Technology—1993,* New York: American Institute of Physics Press.

Sekine, T. (1992), *Proc. Jpn. Acad. Ser. B.* **68,** 95–99.

Syono, Y., Manghnani, M. H. (1992), *High-Pressure Research: Applications to Earth and Planetary Sciences,* Tokyo: Terra Scientific Publishing.

Thadhani, N. N. (1993), *Prog. Mater. Sci.* **37,** 117–226.

van Thiel, M. (1977), *Compendium of Shock Wave Data,* Livermore: Lawrence Livermore Laboratory Report UCRL-50108.

Yoo, C. S., Nellis, W. J., Sattler, M. L., Musket, R. G. (1992), *Appl. Phys. Lett.* **61,** 273–275.

Yoo, C. S., Holmes, N. C., Ross, M., Webb, D. J., Pike, C. (1993), *Phys. Rev. Lett.* **70,** 3931–3934.

Further Reading

Asay, J. R., Shahinpoor, M. (Eds.) (1993), *High Pressure Shock Compression of Solids,* New York: Springer-Verlag.

Bushman, A. V., Fortov, V. E. (1987), in: A. E. Scheindlin, V. E. Fortov (Eds.), *Thermal Physics Reviews,* London: Harwood Academic Publishers, pp. 220–336.

Courant, R., Friedrichs, K. O. (1976), *Supersonic Flow and Shock Waves,* New York: Springer-Verlag.

Duvall, G. E., Graham, R. A. (1977), *Rev. Mod. Phys.* **49,** 523–579.

Eliezer, S., Ghatak, A., Hora, H. (1986), *An Introduction to Equation of State: Theory and Applications,* Cambridge: Cambridge Univ. Press.

Graham, R. A. (1993), *Solids Under High-Pressure Shock Compression,* New York: Springer-Verlag.

McQueen, R. G., Marsh, S. P., Taylor, J. W., Fritz, J. N. (1970), in: R. Kinslow (Ed.), *High-Velocity Impact Phenomena,* New York: Academic Press.

Ross, M., Young, D. A. (1993), *Annu. Rev. Phys. Chem.* **44,** 61–87.

Schmidt, S. C., Shaner, J. W., Samara, G. A., Ross, M. (Eds.) (1994), *High-Pressure Science and Technology—1993,* New York: American Institute of Physics Press.

Zel'dovich, Ya. B., Raizer, Yu. P. (1966), *Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena, Volumes I and II,* New York: Academic Press.

SOLIDS, STATIC HIGH-PRESSURE EFFECTS IN

RUSSELL J. HEMLEY AND HO-KWANG MAO, *Geophysical Laboratory and Center for High-Pressure Research, Carnegie Institution of Washington, Washington, D.C., U.S.A.*

Introduction 555
1. **High-Pressure Techniques** 556
1.1 Large-Volume Devices 556
1.2 Diamond-Anvil Cells 557
1.3 Pressure Calibration 557
1.4 Variable Temperature 559
1.4.1 Cryogenic Methods 559
1.4.2 Resistive Heating 559
1.4.3 Laser Heating 559
2. **High-Pressure Measurements** 559
2.1 Diffraction 559
2.2 Spectroscopy 561
2.3 Transport 561
2.4 Macroscopic 561
3. **Examples of High-Pressure Studies** 562
3.1 Fundamental Physics 562
3.1.1 Insulator–Metal Transitions 562
3.1.2 Hydrogen 562
3.1.3 Superconductivity 563
3.1.4 Semiconductors 563
3.1.5 Melting 564
3.1.6 Pressure Amorphization 564
3.1.7 Phonon Localization Transitions 565
3.2 Chemistry and Materials Science 565
3.2.1 Diamond and Superhard Materials 566
3.2.2 New Low-*Z* Compounds 566
3.2.3 New Amorphous Materials 566
3.2.4 Biological Materials 567
3.3 Earth and Planetary Science 567
3.3.1 Earth's Crust 567
3.3.2 Earth's Mantle 567
3.3.3 Earth's Core 568
3.3.4 Other Planets 568
4. **Outlook** 568
4.1 Higher Pressures 568
4.2 New Analytical Techniques 569
5. **Summary** 569
Works Cited 569
Further Reading 571

INTRODUCTION

Pressure is firmly established as a key variable in many areas of the natural sciences. When pressure is increased on solids, interactions among atoms increase, in some cases radically altering the physical and chemical properties of the material. In contrast, changing interatomic distances by varying composition is restricted by the discrete nature of chemical elements, and tuning by temperature is complicated by thermal excitations (i.e., vibrations) and ultimately the melting and boiling points of the material. Measurements of physical properties tuned over extended pressure ranges now provide crucial experimental bases for testing and developing theories of condensed matter. In addition to its fundamental significance, the effects of pressure on solids have important implications for the syntheses of materials. New materials, with novel physical and chemical properties, have been discovered, and the newly accessible pressure domain provides the prospect of greatly enlarging the number of known materials. High-pressure experiments on geological and planetary materials are providing a unique window on the interiors of the Earth and other planets.

In the past decade, high-pressure research has been revolutionized, principally as a result of breakthroughs in diamond-cell technology. Static pressures of several hundred GPa (gigapascal, 1 GPa = 10 kbar = 10^4 bar) can be reached in the laboratory (Fig. 1). Even more significantly, physical properties of materials can be characterized *in situ* under these conditions, and the accuracy of many high-pressure measurements is now approaching those on samples under ambi-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

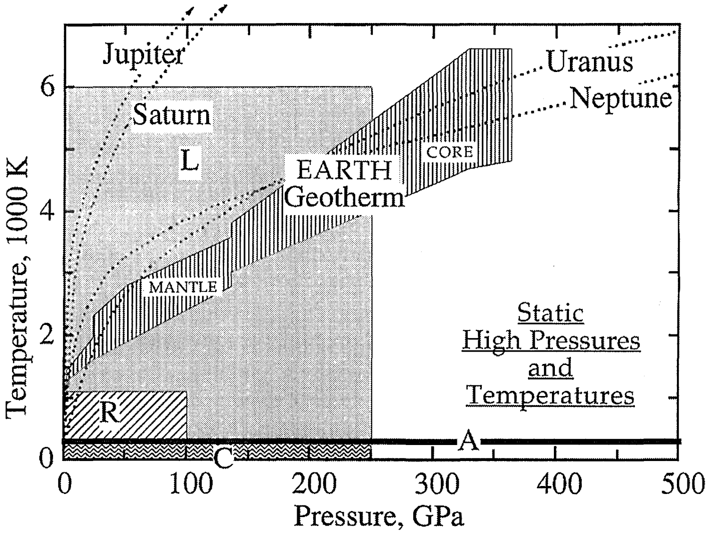

FIG. 1. Pressure–temperature range for static high-pressure experiments superimposed upon the pressure–temperature profiles (geotherms) of the interiors of the Earth and other major planets. Most experiments, including the highest-pressure records, have been performed at the ambient temperature (trajectory "A"). Region "C" denotes the cryogenic range, which has been particularly useful for fundamental studies of physics and chemistry of the materials. The range of simultaneous high temperatures and high pressures achieved by external-resistance heating and/or laser heating are denoted by *R* and *L*, respectively. These ranges represent the reported maxima rather than the intrinsic limitation of the diamond cell. The regions of the Earth's mantle (to 135 GPa) and core (135–364 GPa) are also indicated in the figure.

ent conditions. Typically, a number of high-pressure transitions are observed in each solid in this wide range of pressures. In addition, theoretical work has been growing at a steady rate in all aspects of high-pressure work. Also, high-pressure materials theory, including large-scale computations, approximate methods, and analytical theory, has advanced significantly and plays an essential role in the interpretation of experimental results and in making useful predictions.

1. HIGH-PRESSURE TECHNIQUES

Nearly all static pressure techniques are based on a few conceptually simple principles in which pressure (defined as force/area) is maximized, generally along with tradeoffs in one or more other variables, such as sample size, temperature range, or ability to measure specific physical or chemical properties. High-pressure techniques are reviewed in the article by that title. Bridgman articulated many of the basic principles of high-pressure techniques in use to this day. This included the discovery of the "principle of unsupported area" and the invention of the Bridgman anvil, based on the "principle of massive support." Although the pressure range of laboratory measurements has increased by nearly two orders of magnitude since this early work, devices in use today may be considered descendants of many designs outlined in the early work of Bridgman and others. This advance has come about as a result of continued optimization of many of these early designs, and the fabrication and/or implementation of high-strength materials, such as metal carbides, sintered diamond, and single-crystal diamond.

1.1 Large-Volume Devices

Developments until the 1960s consisted mainly of what are now considered "large-volume" devices. These included the belt or girdle apparatus, the solid-media piston-cylinder apparatus, the split-sphere multianvil (or multipiston) device, the tetrahedral press, and the uniaxial Bridgman (or Drickamer) apparatus (see Sec. 3.2.1 and HIGH-PRESSURE TECHNIQUES). Such large-volume devices create a relatively uniform sample environment at high pressure and temperature, while maintaining the chemical and physical integrity of the samples. Experiments with such devices, while limited to lower pressures (<40 GPa), can produce well-characterized samples for phase equilibria, spectroscopy, calorimetry, x-ray diffraction, and ultrasonic investigations on recovered samples. Studies to temperatures of order 3000 K have been performed. Other modern devices permit a number of *in situ* measurements, including x-ray diffraction. Such experiments have sample dimensions of order several millimeters. New developments include the use of sintered diamond for anvils, instead of tungsten carbide, for reaching higher pressures. Diamond synthesis continues to be an active area of research with these devices. Particularly significant is the growth of chemically and isotopically pure large diamond single crystals.

1.2 Diamond-Anvil Cells

Much of today's knowledge of the behavior of solids at very high pressures can be attributed to the use of diamonds to compress materials in a diamond-anvil cell. The cell is based upon the original Bridgman anvil design, in which the sample is placed between the polished culets of two diamonds and is contained on the sides by a metal gasket. Since pressure is force divided by area, enormous pressures can be generated by opposing anvils with a very small tip (culet), providing that the anvil material is strong enough to withstand the pressure without yielding (Jayaraman, 1983). Flawless single-crystal diamond is chosen for its strength. The anvils exert pressure on a thin metallic gasket within which a circular sample chamber is confined (Fig. 2). Samples immersed in pressure-transmitting media are sealed in the chamber. The gasket also functions as a supporting belt without which the anvils will fail at a maximum pressure of only 35 GPa.

Diamond is transparent to much of the electromagnetic spectrum—including high-energy x and γ radiation above 10 keV and low-energy ultraviolet–visible–infrared radiation below 5 eV. Numerous diffraction, scattering, and absorption probes can be employed for determination of electronic, dynamic, and structural properties of samples under pressure. One of the main constraints on diamond-cell technology is the small sample volume, which in turn is determined by the limited anvil size ($\sim$0.1–0.4 carat; 1 ct = 0.2 g). At a constant anvil size, higher pressures are reached with a reduction in sample volume: there is a tradeoff between the maximum pressure and the maximum volume of the compressed sample. For instance, diamonds with 1-mm large culet diameter are generally limited to pressures below 5 GPa; diamonds with 20-μm small culet diameter are needed for pressures above 300–400 GPa. The diameter of the sample chamber is approximately one-half of the culet diameter, and the thickness, one-tenth. Studying such minute samples requires development of microsampling techniques.

1.3 Pressure Calibration

In lower-pressure work, the pressure may be calculated from the applied force and the geometry of the apparatus. Such calculations become increasingly inaccurate at higher pressures because of the deformation of materials of the apparatus. The limit of the accurate range of pressures determined by force per unit area is about 7 GPa. Because of this, an internal standard is required. A fair amount of work has gone into the problem of pressure calibration at very high pressures. If a physical property of a material is pressure dependent, it can be calibrated to provide a pressure scale. One type of calibrant is a fixed point, which is a discontinuous change in a sample property that has independently been determined to occur at a specific pressure and temperature. An example is a phase transformation, which can be observed following quenching to ambient conditions or observed *in situ* at high pressures (and variable temperature), for example, as a change in crystal structure or electrical resistance. Another class of calibrants exploits a systematically varying pressure–property relation; for this, highest precision obtains when the property exhibits a large pressure effect and can be readily measured. Examples are unit-cell volume of simple metals (*P-V* equations of state) determined by x-ray diffraction. Pb and Pt equations of state have been determined by shock-wave compression to 600 GPa for use as x-ray standards for static pressure experiments (Holmes *et al.*, 1989) (see SOLIDS, DYNAMIC HIGH-PRESSURE EFFECTS IN). Another class comprises spectroscopic techniques, the most popular being ruby fluorescence, which has been calibrated to 180 GPa against equations of state of simple metals (Bell *et al.*, 1986).

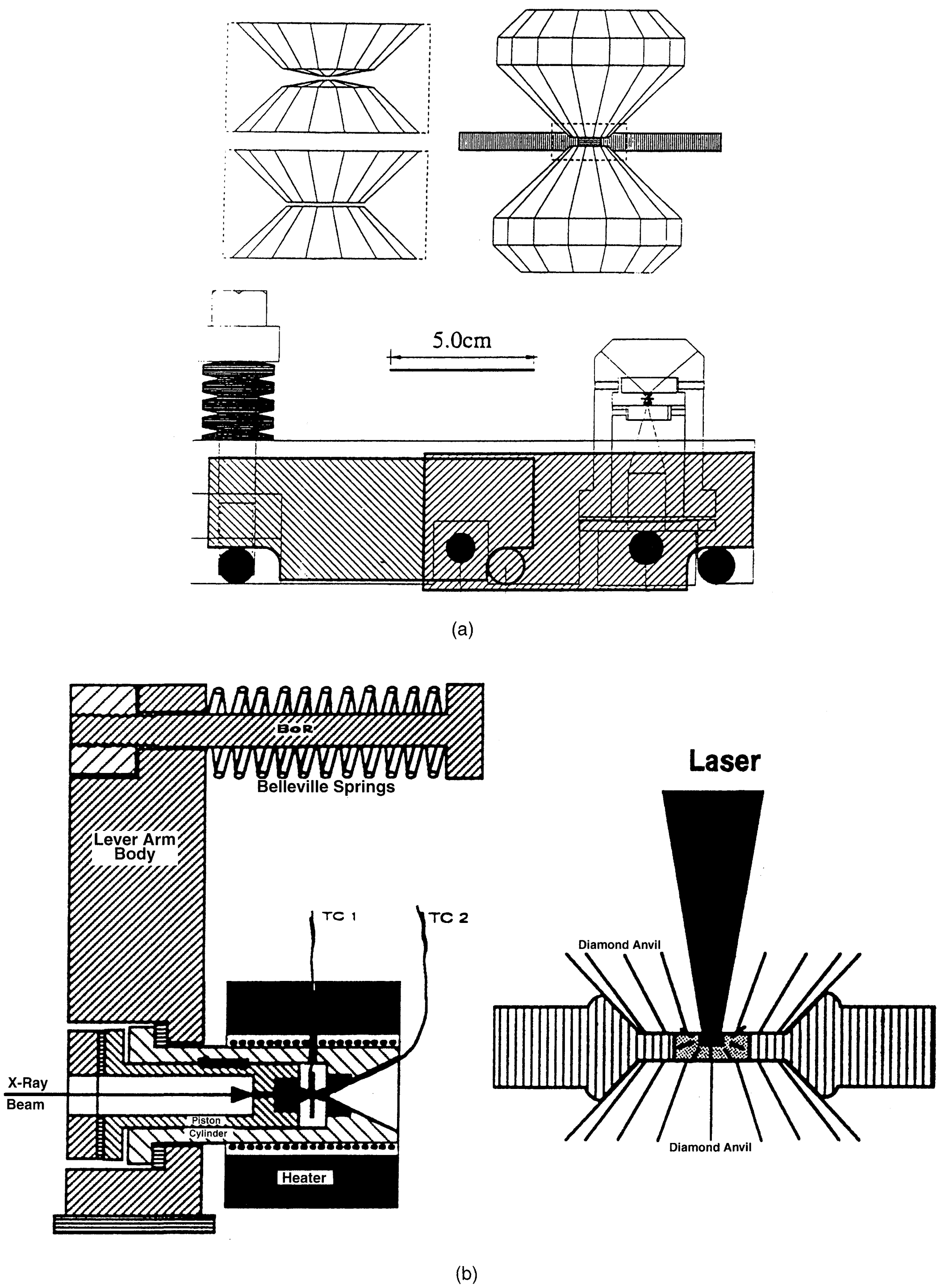
5.0cm
(a)
BoR
Belleville Springs
Lever Arm Body
TC 1
TC 2
X-Ray Beam
Piston
Cylinder
Heater
Laser
Diamond Anvil
Diamond Anvil
(b)

FIG. 2. (a) Diamond-anvil cell design. Top left: Beveled anvils used for reaching multimegabar pressures, and flat anvils used for lower-pressure studies. Top right: Opposed brilliant-cut diamond anvils and gasket. Bottom: Multimegabar diamond-anvil cell with piston-cylinder and two-stage lever arm. (b) High-temperature/high-pressure methods with the diamond-anvil cell. Left: Resistively (externally) heating technique. Right: Laser (internal) heating technique.

1.4 Variable Temperature

1.4.1 Cryogenic Methods In principle, there are no limits on the lowest temperatures for high-pressure studies. A variety of lower-pressure devices (defined here as those achieving maximum pressures of less than 7 GPa), including piston cylinders, have been used to liquid helium temperatures. For practical purposes, however, again the small size and therefore low thermal load of anvil devices are ideally suited for low-temperature studies with flow and bath cryostats as well as refrigerators (Jayaraman, 1983). Diamond-anvil cell experiments into the megabar range (>100 GPa) have been performed with cryostats at liquid helium temperatures. There have also been a few studies at lower pressures (~1 GPa) to temperatures below 0.1 K.

1.4.2 Resistive Heating A variety of resistance techniques can be exploited for high-*P-T* studies, including both external and internal methods [Fig. 2(b)]. External furnaces provide accurate and controlled temperatures to ~1300 K, the primary limitation of which is the graphitization of diamond (i.e., it is unstable relative to graphite at low pressures). These heaters include large coils surrounding the entire assembly (which provide more uniform temperatures but typically lower maximum pressures) or smaller heating elements located in the immediate vicinity of the cell (providing higher temperatures but requiring accurate characterization of axial temperature gradients). Recently, the advantages of both types of furnaces have been realized by implementing both in a two-stage heating technique. Alternatively, the sample can be heated within the sample chamber with a miniature heater, which circumvents the problem of breakdown of diamond because the sample is heated locally and the hottest portion of the diamond is in its stability field.

1.4.3 Laser Heating Higher temperatures (>3000 K) can be reached in diamond-anvil cell experiments by heating samples within the cell with lasers. Crucial to this is absorbance of the sample at the appropriate wavelengths [Fig. 2(b)]. For this purpose, far-infrared CO_2 lasers operating at 10.6 μm and Nd-YAG (yttrium aluminum garnet) and Nd-YLF (yttrium lithium fluoride) lasers at 1.06 μm are used (Boehler, 1996; Jeanloz, 1990; Ming and Bassett, 1974). New developments include the use of laser-heating techniques in combination with synchrotron x-ray methods for simultaneous high-*P-T* diffraction measurements (Fiquet *et al.*, 1996; Yoo *et al.*, 1995).

2. HIGH-PRESSURE MEASUREMENTS

One of the significant features of current high-pressure research is the development of an increasingly wide range of *in situ* techniques that are being applied to the study of static pressure effects in solids. Table 1 lists some of the major categories of analytical techniques used in these studies. Because different techniques require different sample sizes, there is a wide variation in the pressure range possible for different classes of experiments with current technology. In general, great progress in microsampling techniques for probing materials at the highest pressures (>100 GPa) has resulted from recent advances on two fronts: radiation sources and the detector systems. Lasers and synchrotrons provide collimated beams of electromagnetic radiation many orders of magnitude more brilliant than those from conventional light sources. In addition, new spallation sources provide an unparalleled source of neutrons. Multichannel detector technology—such as energy-dispersive x-ray detectors, two-dimensional position-sensitive x-ray detectors, storage-phosphor x-ray imaging plates, image intensifiers, intensified diode-array detectors, charge-coupled detectors, and new Fourier-transform interferometer techniques—allows simultaneous detection in thousands of channels, thus providing sufficient measurement sensitivity for studying ultrasmall samples at ultrahigh pressures. These developments have also resulted in marked improvement in measurement accuracy and sensitivity at high pressures.

2.1 Diffraction

A variety of diffraction techniques are typically used to probe the structures of materials at high pressure (Jayaraman, 1983). These include x rays, neutrons, and optical diffraction experiments. The use of these re-

Table 1. Examples of various measurement techniques and materials studied at high pressures.

Technique	Materials (P_{max})
Diffraction	
X-ray	
Powder	CsI, Fe, W, Re, Pt, Pb (>300 GPa)[a,b]
Single crystal	SiO_2 (70 GPa),[c] $Ar(H_2)_2$ (150 GPa)[d]
Amorphous	SiO_2 glass (42 GPa)[e]
Neutron	
Crystalline	D_2 (30 GPa),[f] Fe (35 GPa)[g]
Amorphous	Liquid D_2 (1 GPa)[h]
Spectroscopy	
X-ray (EXAFS, XANES)	Br_2 (58 GPa)[i]
Optical	
Absorption	H_2 (~300 GPa),[j] Diamond (420 GPa)[k]
Luminescence	Ruby (>300 GPa)[l]
Reflectivity	W, Pt, Pb (>300 GPa)[k]
Multiphoton	KI, RbI (1 GPa)[m,n]
Raman	
Spontaneous Raman	Diamond (400 GPa)[k]
Coherent anti-Stokes Raman	H_2 (37 GPa)[n]
Hyper-Raman	$SrTiO_3$ (5 GPa)[o]
Infrared	H_2 (250 GPa)[j,p]
Acoustic	
Brillouin scattering	SiO_2 (58 GPa)[q]
Impulsive stimulated scattering	$(Mg,Fe)_2SiO_4$ (12 GPa)[r]
Ultrasonic interferometry	Mg_2SiO_4 (12 GPa)[s]
Acoustic emission	Si, Ge (70 GPa)[t]
Inelastic neuron scattering	H_2 (1 GPa)[u]
Mössbauer effect	FeO (100 GPa)[v]
Nuclear magnetic resonance	$FeSiF_6 \cdot H_2O$ (7 GPa),[w] H_2 (7 GPa)[x]
Nuclear quadrupole resonance	Cu_2O (2 GPa)[y]
Electron paramagnetic resonance	$CdCr_2Se_4$ (8 GPa)[z]
Transport	
Electrical	
Resistivity	$(Mg,Fe)SiO_3$ (100 GPa)[aa,bb]
Superconductivity	$HgBa_2Ca_{n-1}Cu_nO_{2n+2+d}$ (30 GPa),[cc] Se (60 GPa)[dd]
Thermoelectric power	LiH (50 GPa)[ee]
Magnetic	
Susceptibility	Sr (50 GPa)[ff]
Thermal conductivity	$(Mg,Fe)_2SiO_4$ (12 GPa)[s]
Macroscopic techniques	
Deformation and strength	Diamond (170 GPa),[gg,hh,ii] SiO_2 glass (80 GPa)[jj]
Macroscopic length	SiO_2 glass (10 GPa)[kk]

---

[a] Mao *et al.* (1989). [b] Ruoff *et al.* (1992). [c] Mao *et al.* (1994). [d] Loubeyre, (1996). [e] Meade *et al.* (1992). [f] Glazkov *et al.* (1988). [g] Shil'shtein *et al.* (1984). [h] Ishmaev *et al.* (1988). [i] Itie *et al.* (1986). [j] Mao and Hemley (1994). [k] Vohra (1992). [l] Xu *et al.* (1986). [m] Lipp *et al.* (1993). [n] Daniels *et al.* (1994). [o] Yamanaka and Inoue (1989). [p] Hanfland *et al.* (1993). [q] Zha *et al.* (1994). [r] Zaug *et al.* (1992). [s] Li *et al.* (1995). [t] Meade and Jeanloz (1989). [u] Kobelev *et al.* (1989). [v] Pasternak *et al.* (1993). [w] Vaughan *et al.* (1971). [x] Lee *et al.* (1989). [y] Reyes *et al.* (1992). [z] Sakai and Pifer (1985). [aa] Peyronneau and Poirier (1989). [bb] Li and Jeanloz (1991). [cc] Gao *et al.* (1994). [dd] Akahama *et al.* (1993). [ee] Babushkin *et al.* (1993). [ff] Bireckoven and Wittig (1988). [gg] Mao *et al.* (1979). [hh] Mao and Hemley (1991). [ii] Weidner *et al.* (1994). [jj] Meade and Jeanloz (1988). [kk] Meade and Jeanloz (1987).

quires a sufficiently transparent material either as the gasket or as the anvil. The transparency of diamond to hard x rays, neutrons, and optical wavelengths makes the diamond-anvil cell an ideal device for a wide range of diffraction studies over the broadest pressure regime. The development of techniques utilizing synchrotron radiation to probe materials has revolutionized many areas of science. Synchrotron radiation is intense, polarized, and highly collimated light emitted over a wide continuum of the spectrum by electrons accelerated in circular orbits to the velocities approaching that of light. In addition, the radiation is produced with a precise time structure that can be exploited for time-resolved and kinetic studies of samples under pressure (Skelton *et al.,* 1991). By far, the most commonly used radiation is the hard and soft x-ray region for diffraction, scattering, and spectroscopic applications.

2.2 Spectroscopy

High-pressure optical spectroscopic methods provide crucial information on behavior of solids at very high pressures (Hemley *et al.,* 1987; Jayaraman, 1983). Vibrational spectroscopy provides detailed information on bonding properties, phase transitions, and thermodynamic properties. Electronic spectroscopy includes absorption, reflectivity, and various luminescence techniques. Again, the development and application of intense synchrotron sources for spectroscopy has been highly significant for high-pressure studies—for example, in the lower energy range, from near- to far-infrared wavelengths. Laser techniques using visible radiation are particularly useful for studies at very high pressures because the laser beam can in principle be focused down to 1 μm, or near the diffraction limit of the radiation, to provide high spatial resolution within a sample [e.g., at ultrahigh pressures above 300 GPa (Hemley *et al.,* 1987)]. In recent years, there have been growing applications of nonlinear optical techniques to study electronic transitions, acoustic velocities, and thermal conductivity at high pressure (Daniels *et al.,* 1994; Zaug *et al.,* 1992). High-pressure microspectroscopic methods have been extended to infrared wavelengths with the development of synchrotron infrared techniques (Hanfland *et al.,* 1993). Other spectroscopic methods include Mössbauer spectroscopy, which has measured to pressures near 100 GPa (Pasternak *et al.,* 1993), and EXAFS, measured to as high as 58 GPa (Itie, 1988).

2.3 Transport

The effect of pressure on the electrical resistance of solids was first studied systematically by Bridgman using a variety of large-volume apparatus. Important subsequent measurements with large-volume apparatus include the resistivity studies of the graphite-to-diamond transition and the phase diagram of carbon (Bundy, 1963; Togaya *et al.,* 1994). Investigations at the highest pressures typically use the four-lead technique with the diamond-anvil cell. The technique was first used in the 10-GPa range by Mao and Bell (1981) in a study of the electrical conductivity of (Mg,Fe)O to 30 GPa. More recently, measurements have been carried out at high temperature and pressure using simultaneous resistance heating or laser-heating techniques (Knittle and Jeanloz, 1986; Li and Jeanloz, 1991; Peyronneau and Poirier, 1989). Direct resistance techniques for determining the pressure dependence of superconductivity have been developed at low temperatures (Akahama *et al.,* 1993; Erskine *et al.,* 1986). Recently, the technique has been used to investigate the pressure dependence of superconductivity in high-temperature superconductors (Gao *et al.,* 1994). At the highest pressures, the materials in the sample chamber deform, causing the leads and contacts to the sample to break or insulation to be damaged. To overcome these problems, and to reach still higher pressures, microlithographic techniques have been used to construct microcircuits on the tips of diamond anvils (Hemmes *et al.,* 1989).

A number of techniques have been developed for the study of the effect of pressure on magnetic properties—for example, variations in magnetic moment with pressure, paramagnetic–diamagnetic transitions, and superconductivity. Magnetic methods can thus be used to study changes in electrical properties (e.g., superconductivity), thereby obviating the need for introducing electrical leads into the sample chamber (Timofeev, 1992). The use of inductive techniques to detect superconductivity through the Meissner effect is considered the most reliable. Such experiments can utilize superconducting flux transformers, multiple modulation techniques, and SQUIDs (Bireckoven and Wittig, 1988; Wijngaarden *et al.,* 1991).

2.4 Macroscopic

Another class of techniques is that involving measurements and observations of the entire sample at very high pressures (or stresses). Griggs first showed that solid materials deform and undergo diffusive creep under pressure (see Griggs and Handin, 1960). More recently, such measurements of deformation and material strength have been extended to higher pressures with diamond-anvil cell techniques (e.g., Meade and Jeanloz, 1988; Sung *et al.*, 1977). The capability of such cells for *in situ* observations also permits direct observations of compressibility from changes in macroscopic sample dimensions (Meade and Jeanloz, 1987). In addition, the pressure dependence of phase transitions such as melting and crystallization can be directly observed (e.g., Lu *et al.*, 1991; Williams *et al.*, 1987).

3. EXAMPLES OF HIGH-PRESSURE STUDIES

Below we give a few examples of the application of new high-pressure techniques to scientific problems. These applications typically involve the use of a number of the techniques described above.

3.1 Fundamental Physics

Pressure of several hundred gigapascals causes the density in a solid to increase as much as 50% for the most incompressible materials (e.g., ceramics and many metals), and 1000% for compressible ones (e.g., solidified gases). A variety of pressure-induced transitions typically occur over this range of compressions. These transitions include crystallographic phase transitions (both discontinuous and continuous), electronic and magnetic transitions, and novel transitions involving localization of excitations (both vibrational and electronic).

3.1.1 Insulator–Metal Transitions With increasing density in a material, electrons in general become less stable localized in bonds and near atomic cores. From the early days of quantum theory it was predicted that above a critical density electrons in all materials must delocalize into conduction states, thus forming a metal (Wigner and Huntington, 1935). The metallization process could involve, for example, a change in molecular or crystallographic structure, or an overlap of electronic valence and conduction bands. The insulator–metal transition in iodine at 16 GPa is well established (Balchan and Drickamer, 1961) and has become a classic example of such a pressure-induced insulator–metal transition. Recently, phase transitions, crystal structures, and electronic properties of iodine have been characterized well into the metallic state at multimegabar pressures (276 GPa) by optical and x-ray diffraction measurements and band-structure calculations (Reichlin *et al.*, 1994). Another example of interesting high-pressure behavior is provided by the pressure-induced transitions in Xe, a rare-gas solid with van der Waals bonding, and CsI, an alkali halide with ionic bonding. Despite their very different bonding properties at ambient pressure, both materials transform to metals above approximately 150 GPa (Goettel *et al.*, 1989; Reichlin *et al.*, 1989, 1987). Moreover, in addition to the convergence in bonding properties, both materials transform to an hcp-like structure and have identical equations of state at these pressures (Jephcoat *et al.*, 1987; Mao *et al.*, 1989).

3.1.2 Hydrogen The behavior of hydrogen at very high pressure is a fundamental problem in condensed matter physics. At low pressures and temperatures, hydrogen is an insulating molecular solid. However, it has been predicted theoretically that at extreme pressures, the molecules will dissociate to form a monatomic metal (Wigner and Huntington, 1935). Moreover, this hypothetical monatomic phase of dense hydrogen was subsequently predicted to exhibit several unusual properties, including very high-temperature superconductivity (Ashcroft, 1968) and a liquidlike ground state (Brovman *et al.*, 1972). The characterization of hydrogen at very high densities has become a major goal of high-pressure research. There has been marked progress in the effort in recent years (Mao and Hemley, 1994). Experiments demonstrate that the molecular bond is stable to at least ~250 GPa (Hemley and Mao, 1988), but several intriguing changes in bonding and physical properties are observed at intermediate pressures (Fig. 3). The changes can be understood in terms of a series of symmetry-breaking transitions with increasing pressure and decreasing temperature (to phases II and III). At higher pressures, there is a drop in the intramolecular stretching mode frequency and a large increase in its absorption strength (phase III) (Hanfland *et al.*, 1993; Hemley and Mao, 1988). Notably,

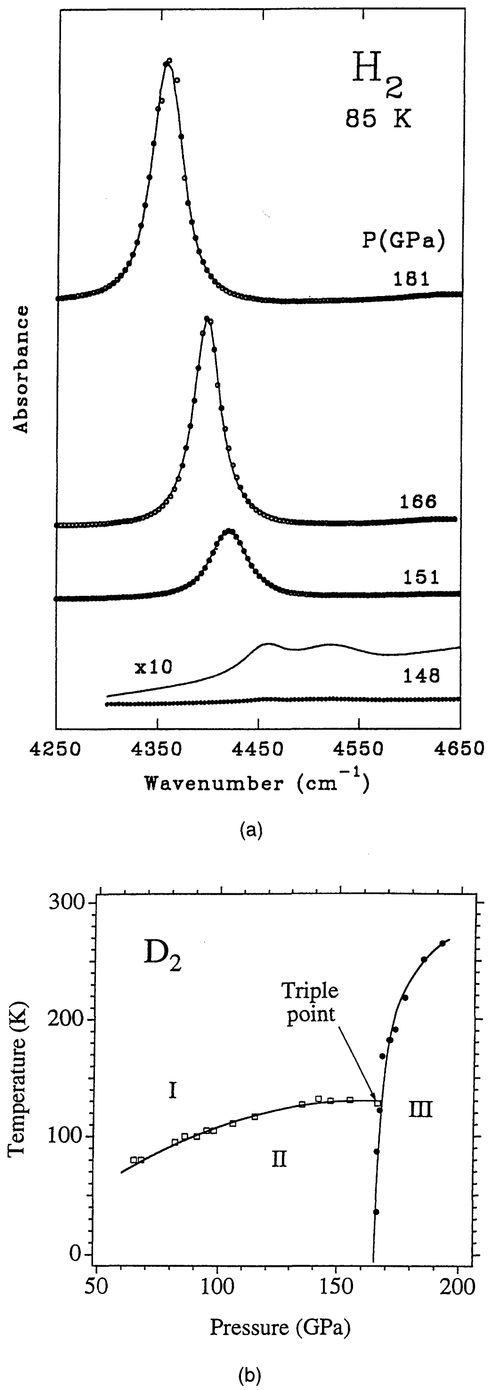

(a)

(b)

FIG. 3. High-pressure studies of solid hydrogen isotopes (Mao and Hemley, 1994). (a) Synchrotron infrared absorption spectroscopy of hydrogen showing the intensity enhancement at the transition at 150 GPa to phase III. (b) Phase diagram of deuterium at 100–200 GPa. The boundary between I and III terminates at a higher-temperature invariant point, above which the transition becomes continuous.

the material remains transparent in the visible and midinfrared spectrum (Eggert *et al.*, 1991; Hemley *et al.*, 1996), indicating that metallization occurs at higher pressures.

Despite its apparent simplicity, accurate theoretical calculations for dense hydrogen are particularly difficult and pose a difficult challenge to modern condensed-matter theory. This difficulty arises from the significant quantum character of the atomic nuclei (zero-point energy) associated with their low mass, which requires the proton and electron dynamics to be treated on the same theoretical footing (Ashcroft, 1993). This has been done in recent quantum Monte Carlo calculations for both the atomic and molecular phases at high pressure (Natoli *et al.*, 1993). The large zero-point motion can play a significant role in stabilizing structures [e.g., the diamond structure for the atomic phase (Natoli *et al.*, 1993)]. In addition, the unusual dynamics of the dense solid may give rise to new phenomena such as new types of pairing interactions and high-T_c superconductivity in the theoretically predicted molecular metallic state (Moulopoulos and Ashcroft, 1991).

3.1.3 Superconductivity Physical properties of fundamental and technological interest, such as superconductivity, can also be favorably tuned or enhanced by pressure. Understanding pressure effects on T_c played a major role in the discovery of the first 90-K cuprate superconductors (Wu *et al.*, 1987). The recently discovered mercury-based oxides $HgBa_2Ca_{n-1}Cu_nO_{2n+2+d}$ (n = 1, 2, 3) have superconducting onset temperatures (T_c) of 94, 128, and 135 K, respectively (Gao *et al.*, 1994). Their T_c's were found to increase with pressure. When studied in a diamond cell with the four-lead electrical conductance technique (maximum pressure of 45 GPa), record high T_c's of 118, 154, and 164 K were reached (Fig. 4).

3.1.4 Semiconductors There are also interesting high-pressure effects in semiconductors, in part due to processes associated with changing the band gap, including closure of the gap at moderate pressures and crystallographic transitions to metallic states (see Yu *et al.*, 1995). The prototype example of the latter is the transformation in Si from the semiconducting diamond structure to the metallic β-Sn structure at 12 GPa (Jayara-

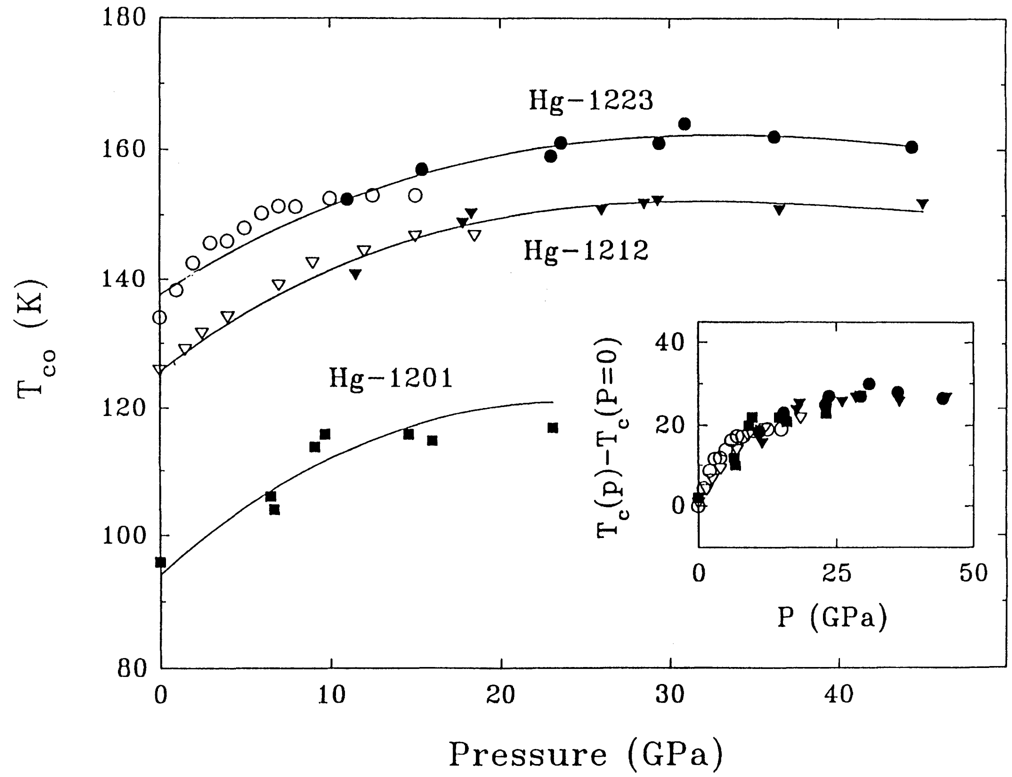

FIG. 4. Pressure dependence of the superconducting onset temperature T_{co} for $HgBa_2Ca_{n-1}Cu_nO_{2n+2+d}$ (Gao *et al.*, 1994).

man, 1983). Recent diffraction studies reveal that a number of the high-pressure phases in semiconductors have structures that are more complex than previously thought (McMahon *et al.*, 1994). Recent studies have also revealed unusual pressure effects on deep-level impurity centers and electronic levels in multiple quantum well and superlattice structures (Paul *et al.*, 1994).

3.1.5 Melting After centuries of study, a quantitative and complete microscopic understanding of melting is still not in hand. Pressure studies provide important tests of theories of melting. The form (slope) of the melting curve under pressure provides information on the structural and thermodynamic changes in the solid and liquid as a function of pressure. There have been numerous debates about the melting behavior of specific materials at very high pressure, including melting of materials that constitute interiors of the planets (Boehler, 1996). Whether or not a melting line will in general terminate in a critical point like the vaporization curve has been a recurring question in high-pressure physics. Experimental work to this day has failed to find evidence for such solid–liquid critical behavior. However, if sufficiently extreme conditions are reached, new phenomena associated with atomic breakdown are expected. The plasma phase transition in fluid hydrogen has been predicted to occur at 8000 K and 190 GPa and has a negative *P-T* slope (Saumon and Chabrier, 1989). The nature of such transitions, including the associated critical phenomena, is an important question addressed in current research.

3.1.6 Pressure Amorphization When a crystalline material is compressed to pressures well outside of its stability field, it may convert to an amorphous material if the temperature is too low for it to recrystallize to the equilibrium high-pressure crystalline phase. This unusual transformation was first documented in ice I; instead of transforming to the equilibrium crystalline phase, ice I transforms to a dense glass when compressed above 1 GPa at 77 K (Mishima *et al.*, 1984). Such pressure-induced amorphization transformations have been observed in a growing number of materials during the past ten years (Hemley *et al.*, 1994). These transformations may access unusual metastable states with varying degrees of disorder and possibly unique properties.

3.1.7 Phonon Localization Transitions

Pressure can also be used to study unusual transitions associated with the localization and delocalization of excitations. Most solids can be described by the harmonic approximation; in some others, anharmonicity prevails. In the harmonic case, noninteracting phonons couple strongly and form a broad continuum band (Kimball *et al.*, 1981). In the anharmonic case, strong interaction of phonons leads to the formation of bound states only weakly coupled in the crystal. For example, the bound state in the vibrational spectrum forms a sharp peak separated from the origin of the broad two-phonon continuum band by the quantity of its anharmonicity. If the relative magnitudes of the continuum band width and the anharmonicity could be varied so that the former would exceed the latter, it was postulated, the bound two-phonon bound state would disappear into the continuum band. Such a transition was observed for the first time by varying pressure in solid deuterium (Fig. 5) (Eggert *et al.*, 1993). At ambient condition, the two-phonon bound state is well separated from the continuum, but at high pressures the continuum bandwidth, which is a measure of the intermolecular coupling, increases sharply. The two become equal at 36 GPa, and indeed, the two-phonon band vanishes into the continuum.

3.2 Chemistry and Materials Science

High-pressure methods up to several gigapascals have been used for years for synthesis of materials—for example, in the growth of high-quality single crystals using hydrothermal and piston–cylinder techniques. However, at higher pressures, bonding properties in materials can be altered considerably, in some cases drastically, as the relative energies of the valence electrons change in response to compression. Thus, chemical af-

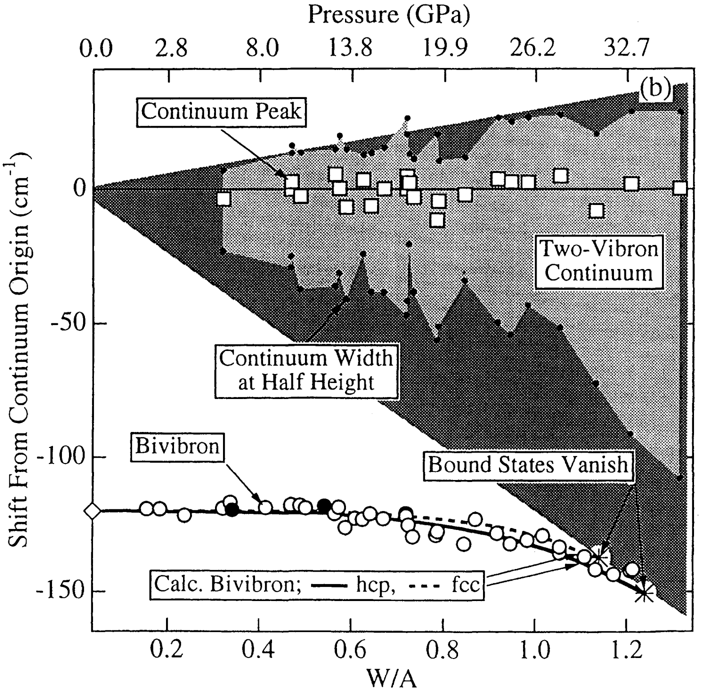

FIG. 5. Pressure-induced two-phonon ("bivibron") bound–unbound transition in solid deuterium. The shaded area shows the continuum bandwidth, *W*. *A* is anharmonicity (Eggert *et al.*, 1993).

finities of the elements can become very different from those at ambient pressure. Thus, new metastable materials with novel or otherwise useful physical and chemical properties can be tailored by appropriate high-pressure synthesis. In this regard, it is important to examine the mechanical stability on decompression of solid phases that are metastable at ambient pressure.

3.2.1 Diamond and Superhard Materials Probably the most celebrated example of high-pressure synthesis was the production of diamond by General Electric Co. in 1955 (see Bundy *et al.*, 1996). This advance was important for demonstrating the economic and technological utility of high-pressure research as well as for initiating new and innovative static high-pressure techniques in general. [Subsequently, diamond was synthesized by shock-wave methods in 1961 (DeCarli and Jamieson, 1961).] Together with the development of chemical vapor deposition techniques for diamond growth, high-pressure diamond synthesis remains a major industry today. The primary product of this industry is diamond powders for abrasives. A highly significant advance in diamond synthesis in the 1990s is the growth of large, perfect single-crystal (gem quality) diamonds, including those that are close to being isotopically pure and having unique physical properties, such as exceptionally high thermal conductivity (Wei *et al.*, 1993) and strength (Vohra *et al.*, 1994). This breakthrough is turning out to be very important for high–heat-load optics applications (e.g., synchrotron radiation research) as well as for anvils in high-pressure experiments. Diamond is predicted to transform to a metallic phase at pressures above 1.2 TPa (12 Mbar) (Biswas *et al.*, 1984; Yin and Cohen, 1983). So far, these pressures have not been reached in static experiments; diamond-cell experiments using natural diamonds do reveal signs of transformations and plastic flow under the extreme nonhydrostatic stresses produced at the tips of the anvils (Mao and Hemley, 1991).

Unusual transformations in graphite to a diamond-like phase have been documented under pressure [pressure-induced metal-to-insulator transition (Bundy *et al.*, 1996)]. The high-pressure phase has the characteristics of a superhard material like diamond. New superhard phases have also been produced from subjecting fullerenes to high pressures (Nunez-Regueiro *et al.*, 1995). The theoretical prediction of superhard carbon nitrides, in particular β-C_3N_4 (Liu and Cohen, 1989), has inspired a large number of experimental and theoretical investigations. Recently, a new class of such materials having a bulk modulus (incompressibility) greater than that of diamond has been predicted theoretically (Teter and Hemley, 1996).

3.2.2 New Low-Z Compounds Inert-gas molecules interact weakly at low density through van der Waals forces. They also do not readily form stoichiometric compounds at ordinary pressure. However, such solid compounds can be made at high pressures, as was first demonstrated with the synthesis of $He(N_2)_{11}$ (Vos *et al.*, 1992). Since then, a number of such compounds of this novel type, $NeHe_2$, $Ar(H_2)_2$, and several in the CH_4–H_2 system, have been observed at high pressures (Loubeyre, 1996). Their crystal structures and stoichiometry can be understood in terms of molecular packing efficiency. Clathrate hydrates consist of hydrogen-bonded networks of cages containing larger guest molecules typically bound by van der Waals forces. They are unstable at moderate pressures, as the open networks break down under compression. In a high-pressure study of the H_2–H_2O binary system, a novel type of clathrate with 1:1 ratio was discovered (Vos *et al.*, 1993) (Fig. 6). In this high-pressure clathrate, H_2O and H_2 form two interlocking networks, both with the diamond structure. With the efficient packing of molecules afforded by the structure, the clathrate is stable to at least 30 GPa.

3.2.3 New Amorphous Materials A number of amorphous materials, including high-silica glasses, can be irreversibly compacted by pressure; that is, a densified form can be retrieved on quenching to ambient conditions (Hemley *et al.*, 1994). This provides materials with a range of optical properties, densities, and chemical properties that can be continuously tuned by changing synthesis conditions. New quenchable nitride glasses with high refractive indices have been produced in recent high-pressure studies (Grande *et al.*, 1994).

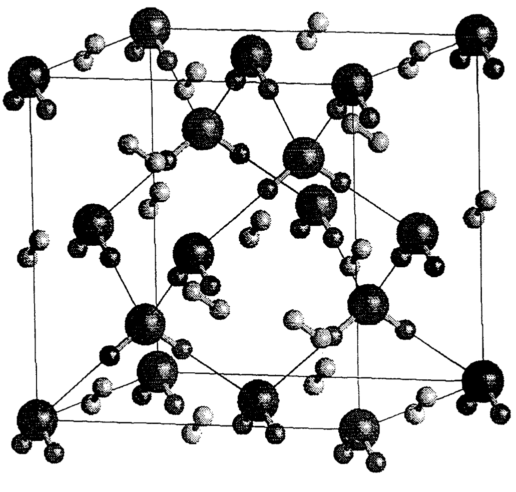

FIG. 6. Crystal structure of the high-density clathrate. The dumbbells are idealized representations of the rotationally disordered hydrogen molecules; the large spheres denote the oxygens of the H_2O molecules (Vos *et al.*, 1993).

3.2.4 Biological Materials There are interesting effects of pressure on biological materials, and most of these changes occur at comparatively low pressures (e.g., below 1 GPa). These include pressure-induced denaturation of proteins, an effect first observed by Bridgman in 1914 in egg albumen. Pressure is also used for food purification (similar to heat treatment or pasteurization) (Hayashi, 1991). The stability of micro-organisms at high temperatures and moderate pressures (<0.1 GPa) is important for understanding the extraordinary biological activity associated with volcanic seafloor vents (e.g., "black smokers").

3.3 Earth and Planetary Science

Laboratory experiments utilizing high pressure and/or high temperature offer the only means to examine directly the chemistry and physical properties of materials that constitute deep planetary interiors.

3.3.1 Earth's Crust Applications of the new techniques described above are providing new insights into a number of high-pressure phenomena associated with the properties and processes of deeper portions of the Earth's crust. Recently, large enhancements of solubility of minerals, such as quartz in aqueous solutions, have been observed under lower crustal pressures (i.e., 0.5–2 GPa) (Manning, 1994). Pressure- and temperature-induced dehydration reactions of low-pressure hydrous minerals have been observed directly, including observations of the kinetics of such transformations using synchrotron x-ray diffraction. Diamond-anvil studies have also been used to determine the equations of state of aqueous fluids to deep crustal conditions (Shen *et al.*, 1992). In addition, the methods provide direct observation of crystal growth of crustal minerals.

3.3.2 Earth's Mantle Studies of pressure-induced polymorphism in minerals have been particularly important for developing progressively more refined models of the mineralogy and composition of the Earth's deep mantle (Liu and Bassett, 1986). The general role of phase transitions in the Earth's deep interior, and how these may give rise to the well-known seismic wave discontinuities, was articulated by Birch (1952). The olivine–spinel transformation was discovered in silicates and investigated in detail in the 1960s, and it remains the leading hypothesis for explaining the 400-km discontinuity. Magnesium metasilicate (enstatite, $MgSiO_3$) transforms to the perovskite-type phase (Fig. 7) (Liu, 1975) at pressures corresponding to the second major seismic discontinuity in the Earth's mantle at 660 km (or 23 GPa). Moreover, silicate minerals in general transform to perovskite-dominated assemblages at lower mantle pressures. This has given rise to the view that (Mg, Fe)SiO_3 perovskite is the dominant mineral in the Earth's lower mantle, and as a result of the large volume of this region, it is likely to be the most abundant mineral in the planet as a whole.

Pressure-induced polymorphism of SiO_2—in particular, the transition from quartz to the high-pressure phases coesite (Coes, 1953) and stishovite (Stishov and Popova, 1961)—had a significant impact on understanding of the deep solid Earth. The higher-pressure form is a tetragonal structure with the rutile-type structure and involves a change in coordination of Si by O from fourfold (the coordination found in crustal silicates) to sixfold (the typical coordination found in the deep mantle). Recent diamond-cell experiments demonstrate that stishovite transforms

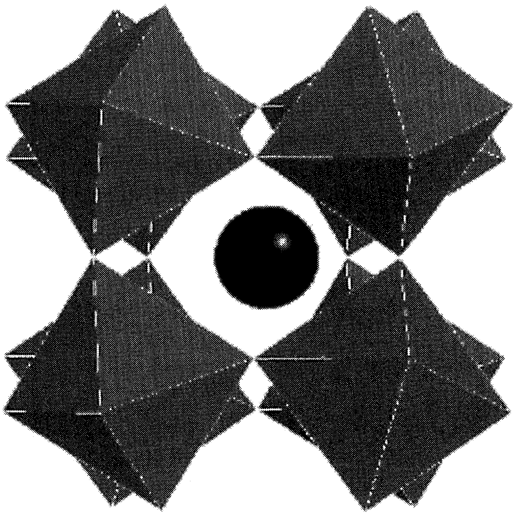

FIG. 7. Crystal structure of orthorhombic (Mg, Fe)SiO_3 perovskite, the dominant phase of the lower mantle of the Earth and most abundant mineral in the planet. The structure consists of corner-sharing SiO_6 groups (octahedra) with Mg and Fe ions (spheres) in the larger crystallographic sites.

to an orthorhombic form near 50 GPa (Kingma *et al.*, 1995). In all of this work, theory has played an important role by providing both fundamental understanding of observed mineral behavior and predictions to be tested subsequently by experiment (Bukowinski, 1994; Cohen, 1994).

3.3.3 Earth's Core *In situ* diamond-anvil cell studies have been instrumental in the characterization of polymorphism and melting of iron to core conditions (Boehler, 1996; Mao *et al.*, 1990; Williams *et al.*, 1987). Although theories for the dominant role of metallic hydrogen in the Earth's core were eventually discredited on the basis of the low density predicted for metallic hydrogen relative to that of the core, the effect of alloying hydrogen with iron (an effect that does not occur at low pressures) was not ruled out. The latter idea gained some favor beginning in the 1980s when iron hydride was shown to be stable under pressure (Badding *et al.*, 1991).

3.3.4 Other Planets High-pressure studies are also providing new insights into the nature of the interiors of other planets in the solar system. Experiments on silicates, oxides, and sulfides yield information relevant to other terrestrial planets such as Mars. The large gaseous planets are believed to consist of approximately 90% hydrogen, with most of this material existing at ultrahigh pressures in a fluid metallic state. Together with shock-wave studies, static high-pressure investigations of hydrogen and helium, together with other low-Z materials (e.g., the C–H–O–N system), thus provide an experimental basis for understanding the structure, composition, dynamics, and evolution of the large gaseous planets. In addition, studies of low-Z molecular materials (e.g., CH_4, NH_3, CO_2, H_2O) are relevant to the moons that orbit these bodies (Johnson and Nicol, 1986). Of particular interest is the observation of new solid compounds in mixtures of simple molecular materials (Vos *et al.*, 1993).

4. OUTLOOK

4.1 Higher Pressures

The upper limit of static pressure techniques is currently unknown. Efforts to reach the highest static pressures continue (Mao *et al.*, 1989; Ruoff *et al.*, 1992; Xu *et al.*, 1986). In principle, the diamond cell is limited by possible ultrahigh-pressure transformations in diamond. Studies to >300 GPa reveal no evidence for metallization of diamond; as discussed above, theory predicts that this should occur at significantly higher hydrostatic pressure. However, even this may not represent a practical limit since the lower-pressure insulating diamond may be metastable above the transition (just as diamond is metastable relative to graphite at ambient pressure). Stress-induced transition and plastic flow has been documented in regions of high stress concentration localized at the anvil tips, but this may be circumvented by further refinements of cell designs. Further advances may result from the discovery and synthesis of new superhard materials for anvils. These advances may also be useful for the related goal of enlarging sample volumes in order to improve the accuracy and precision of measurements as well as permitting a still broader spectrum of techniques to be used to probe materials under extreme pressure conditions.

4.2 New Analytical Techniques

High-pressure research is still in the early stages of utilizing a number of new analytical techniques. These include, for example, the construction of ultrabright synchrotron sources (such as the European Synchrotron Radiation Facility in France, the Advanced Photon Source in the United States, and SPring-8 in Japan), which will make new classes of diffraction, spectroscopic, and imaging experiments possible with high-pressure devices. Also, new spectroscopic measurements will become possible with the continued development of infrared and ultraviolet synchrotron sources, including those utilizing coherent radiation (free-electron lasers). There are efforts to develop lithographic techniques for producing microcircuits on anvil surfaces [perhaps in conjunction with CVD diamond (McCauley and Vohra, 1995)] for electrical and magnetic measurements at multimegabar pressures. There is also the prospect of observing new phenomena at ultralow (millikelvin) temperatures and megabar pressures, currently uncharted territory.

5. SUMMARY

With recent breakthroughs in high-pressure technology, a wide range of measurements can now be conducted *in situ* on compressed solids over an unprecedented range of pressures. In the small number of systems investigated to date, a variety of pressure-induced transitions have been observed in solids under ultrahigh pressures. It is conceivable that novel physical phenomena may emerge with detailed study of matter in the new pressure dimension made possible by the development of new techniques. Analogy could be drawn to the discoveries of superfluidity and superconductivity. These properties were unpredicted and unexpected and could only be found with the application of extreme temperature conditions and careful experimental observations. Searching for novel physical phenomena that are unique to extreme pressure conditions requires similar efforts.

Works Cited

Akahama, Y., Kobayashi, M., Kawamura, H., Endo, S. (1993), *Rev. Sci. Instrum.* **64,** 1979–1983.

Ashcroft, N. W. (1968), *Phys. Rev. Lett.* **21,** 1748–1750.

Ashcroft, N. W. (1993), *J. Non-Cryst. Solids* **156–158,** 621–630.

Babushkin, A. N., Pilipenko, G. I., Gavrilov, F. F. (1993), *J. Phys. F: Condensed Matter* **5,** 8659–8664.

Badding, J. V., Hemley, R. J., Mao, H. K. (1991), *Science* **253,** 421–424.

Balchan, A. S., Drickamer, H. G. (1961), *J. Chem. Phys.* **34,** 1948–1949.

Bell, P. M., Xu, J., Mao, H. K. (1986), in: Y. Gupta (Ed.), *Shock Waves in Condensed Matter,* New York: Plenum, pp. 125–130.

Birch, F. (1952), *J. Geophys. Res.* **57,** 227–286.

Bireckoven, B., Wittig, J. (1988), *J. Phys. E.: Sci. Instrum.* **21,** 841–848.

Biswas, R., Martin, R. M., Needs, R. J., Nielsen, O. H. (1984), *Phys. Rev. B* **30,** 3210–3213.

Boehler, R. (1996), *Philos. Trans. R. Soc. London A* **354,** 1265–1278.

Brovman, E. G., Kagan, Y., Kholas, A. (1972), *Sov. Phys. JETP* **34,** 1300–1315.

Bukowinski, M. S. T. (1994), *Ann. Rev. Earth Planet. Sci.* **22,** 167–205.

Bundy, F. P. (1963), *J. Chem. Phys.* **38,** 618–630.

Bundy, F. P., Weathers, M. A., Bassett, W. A., Hemley, R. J., Mao, H. K., Goncharov, A. F. (1996), *Carbon* **34,** 141–153.

Coes, L. (1953), *Science* **118,** 131–132.

Cohen, R. E. (1994), in: P. Heaney, G. V. Gibbs, C. T. Prewitt (Eds.), *Reviews of Mineralogy,* Washington, D.C.: Mineralogical Society of America, pp. 369–402.

Daniels, W. B., Lipp, M., Strachan, D., Yoo, C., Zhang, H. M., Yu, Z. (1994), in: S. C. Schmidt, J. W. Shaner, G. Samara, M. Ross (Eds.), *High-Pressure Science and Technology—1993,* New York: American Institute of Physics, pp. 1499–1502.

DeCarli, P. S., Jamieson, J. C. (1961), *Science* **133,** 1821–1822.

Eggert, J. H., Moshary, F., Evans, W. J., Lorenzana, H. E., Goettel, K. A., Silvera, I. F., Moss, W. C. (1991), *Phys. Rev. Lett.* **66,** 193–196.

Eggert, J. H., Mao, H. K., Hemley, R. J. (1993), *Phys. Rev. Lett.* **70,** 2301–2304.

Erskine, D., Yu, P. Y., Chang, K. J., Cohen, M. L. (1986), *Phys. Rev. Lett.* **57,** 2741–2744.

Fiquet, G., Andrault, D., Itie, J. P., Gillet, P., Richet, P. (1996), *Phys. Earth Planet. Sci.,* in press.

Gao, L., Xue, Y. Y., Chen, F., Xiong, Q., Meng, R. L., Ramirez, D., Chu, C. W., Eggert, J. H., Mao, H. K. (1994), *Phys. Rev. B* **50,** 4260–4263.

Glazkov, V. P., Besedin, S. P., Goncharenko, I. N., Irodova, A. V., Makarenko, I. N., Somenkov, V. A., Stishov, S. M., Shil'shteyn, S. S. (1988), *JETP Lett.* **47,** 661–664.

Goettel, K. A., Eggert, J. H., Silvera, I. F., Moss, W. C. (1989), *Phys. Rev. Lett.* **62,** 665–668.

Grande, T., Holloway, J. R., McMillan, P. F., Angell, C. A. (1994), *Nature* **369,** 43–45.

Griggs, D., Handin, J. (1960), in: D. Griggs, J. Handin (Eds.), *Rock Deformation,* Memoir 79, New York: Geological Society of America, pp. 347–364.

Hanfland, M., Hemley, R. J., Mao, H. K. (1993), *Phys. Rev. Lett.* **70,** 3760–3763.

Hayashi, R. (1991), *High Press Res.* **7,** 15–21.

Hemley, R. J., Mao, H. K. (1988), *Phys. Rev. Lett.* **61,** 857–860.

Hemley, R. J., Bell, P. M., Mao, H. K. (1987), *Science* **237,** 605–612.

Hemley, R. J., Prewitt, C. T., Kingma, K. J. (1994), in: P. Heaney, G. V. Gibbs, C. T. Prewitt (Eds.), *Reviews of Mineralogy,* Washington, D.C.: Mineralogical Society of America, pp. 41–81.

Hemley, R. J., Mao, H. K., Goncharov, A. F., Hanfland, M., Struzhkin, V. V. (1996), *Phys. Rev. Lett.* **76,** 1667–1670.

Hemmes, H., Driessen, A., Kos, J., Mul, F. A., Griessen, R., Caro, J., Radelaar, S. (1989), *Rev. Sci. Instrum.* **60,** 474–480.

Holmes, N. C., Moriarty, J. A., Gathers, G. R., Nellis, W. (1989), *J. Appl. Phys.* **66,** 2962–2967.

Ishmaev, S. I., Sadikov, I. P., Chernoyshov, A. A., Isakov, S. L., Vindryaeskii, B. A., Kobelev, G. V., Sukhoparov, V. A., Telepnev, A. S. (1988), *Sov. Phys. JETP* **68,** 1403–1406.

Itie, J. P., Jean-Louis, M., Dartyge, E., Fontaine, A., Jucha, A. (1986), *J. Phys. (Paris)* **47,** 897–900.

Jayaraman, A. (1983), *Rev. Mod. Phys.* **55,** 65–108.

Jeanloz, R. (1990), *Annu. Rev. Phys. Chem.* **40,** 237–259.

Jephcoat, A. P., Mao, H. K., Finger, L. W., Cox, D. E., Hemley, R. J., Zha, C. S. (1987), *Phys. Rev. Lett.* **59,** 2670–2673.

Johnson, M. L., Nicol, M. (1986), *J. Geophys. Res.* **92,** 6339–6349.

Kimball, J. C., Fong, C. Y., Shen, Y. R. (1981), *Phys. Rev. B* **23,** 4946–4959.

Kingma, K. J., Cohen, R. E., Hemley, R. J., Mao, H. K. (1995), *Nature* **374,** 243–245.

Knittle, E., Jeanloz, R. (1986), *Solid State Comm.* **13,** 1541–1545.

Kobelev, G. V., Sukhoparov, A. A., Telepnev, A. S., Vindryaevskyi, B. A., Ishmaev, S. N., Sadikov, I. P., Chernychov, A. A. (1989), in: N. V. Novikov, Y. M. Chistyakov (Eds.), *High Pressure Science and Technology, Proceedings of the XIth AIRAPT Conference,* Kiev: Naukova Dumka, pp. 294–298.

Lee, S. H., Conradi, M. S., Norberg, R. E. (1989), *Phys. Rev. B* **40,** 12492–12498.

Li, B., Liebermann, R. C., Gwanmesia, G. D. (1995), *Eos Trans. Am. Geophys. Union (Fall Meeting)* **76,** F619.

Li, X., Jeanloz, R. (1991), *J. Geophys. Res.* **96,** 6113–6120.

Lipp, M. J., Blechschmidt, J., Daniels, W. B. (1993), *Phys. Rev. B* **48,** 687–694.

Liu, A., Cohen, M. L. (1989), *Science* **245,** 841.

Liu, L.-g. (1975), *Geophys. Res. Lett.* **2,** 417–419.

Loubeyre, P. (1996), *Proceedings of the XV AIRAPT Conference,* in press.

Lu, G. Q., Nygren, E., Aziz, M. J. (1991), *J. Appl. Phys.* **70,** 5323–5345.

Manning, C. E. (1994), *Geochem. Cosmochem. Acta* **58,** 4831–4839.

Mao, H. K., Bell, P. M. (1981), *Rev. Sci. Instrum.* **52,** 615.

Mao, H. K., Hemley, R. J. (1991), *Nature* **351,** 721–724.

Mao, H. K., Hemley, R. J. (1994), *Rev. Mod. Phys.* **66,** 671–692.

Mao, H. K., Bell, P. M., Dunn, K. J., Chrenko, R. M., DeVries, R. C. (1979), *Rev. Sci. Instrum.* **50,** 1002–1009.

Mao, H. K., Wu, Y., Hemley, R. J., Chen, L. C., Shu, J. F., Finger, L. W. (1989), *Science* **246,** 649–651.

Mao, H. K., Wu, Y., Chen, L. C., Shu, J. F., Jephcoat, A. P. (1990), *J. Geophys. Res.* **95,** 21737–21742.

Mao, H. K., Shu, J., Hu, J., Hemley, R. J. (1994), *Eos Trans. Am. Geophys. Union (Fall Meeting)* **75,** 662.

McCauley, T. S., Vohra, Y. K. (1995), *Appl. Phys. Lett.* **66,** 1486–1488.

McMahon, M. I., Nelmes, R. J., Wright, N. G., Allan, D. R. (1994), in: S. C. Schmidt, J. W. Shaner, G. A. Samara, M. Ross (Eds.), *High-Pressure Science and Technology—1993,* New York: American Institute of Physics, pp. 629–636.

Meade, C., Jeanloz, R. (1987), *Phys. Rev. B* **35,** 236–244.

Meade, C., Jeanloz, R. (1988), *Science* **241,** 1072–1074.

Meade, C., Jeanloz, R. (1989), *Nature* **339,** 616–618.

Meade, C., Hemley, R. J., Mao, H. K. (1992), *Phys. Rev. Lett.* **69,** 1387–1390.

Ming, L. C., Bassett, W. A. (1974), *Rev. Sci. Instrum.* **45,** 1115–1118.

Mishima, O., Calvert, L. D., Whalley, E. (1984), *Nature* **310,** 393–395.

Moulopoulos, K., Ashcroft, N. W. (1991), *Phys. Rev. Lett.* **66,** 2915–2918.

Natoli, V., Martin, R. M., Ceperley, D. M. (1993), *Phys. Rev. Lett.* **70,** 1952–1955.

Nunez-Regueiro, M., Marques, L., Hodeau, J. L., Bethoux, O., Perroux, M. (1995), *Phys. Rev. Lett.* **74,** 3483–3486.

Pasternak, M., Taylor, R. D., Dillon, T., Jeanloz, R. (1993), *Eos Trans. Am. Geophys. Union (Fall Meeting Suppl.)* **74,** 416.

Paul, W., Burnett, J. H., Cheong, H. M. (1994), in: S. C. Schmidt, J. W. Shaner, G. Samara, M. Ross (Eds.), *High-Pressure Science and Technol-*

ogy—1993, New York: American Institute of Physics, pp. 545–548.

Peyronneau, J., Poirier, J. P. (1989), *Nature* **342,** 537–539.

Reichlin, R., Ross, M., Martin, S., Goettel, K. A. (1987), *Phys. Rev. Lett.* **56,** 2858–2861.

Reichlin, R., Brister, K. E., McMahan, A. K., Ross, M., Martin, S., Vohra, Y. K., Ruoff, A. L. (1989), *Phys. Rev. Lett.* **62,** 669–672.

Reichlin, R., McMahan, A. K., Ross, M., Martin, S., Hu, J., Hemley, R. J., Mao, H.-K., Wu, Y. (1994), *Phys. Rev. B* **49,** 3725–3733.

Reyes, A. P., Ahrens, E. T., Heffner, R. H., Hammel, P. C., Thompson, J. D. (1992), *Rev. Sci. Instrum.* **63,** 3120–3122.

Ruoff, A. L., Xia, H., Xia, Q. (1992), *Rev. Sci. Instrum.* **63,** 4342–4348.

Sakai, N., Pifer, J. M. (1985), *Rev. Sci. Instrum.* **56,** 726–731.

Saumon, D., Chabrier, G. (1989), *Phys. Rev. Lett.* **62,** 2397–2400.

Shen, A. H., Bassett, W. A., Chou, I.-M. (1992), in: Y. Syono, M. H. Manghnani (Eds.), *High-Pressure Research: Application to Earth and Planetary Sciences,* Tokyo: Terra Scientific; Washington, D.C.: American Geophysical Union, pp. 61–68.

Shil'shtein, S. S., Glazkov, V. P., Makarenko, I. N., Somenkov, V. A., Stishov, S. M. (1984), *Sov. Phys. Solid State* **25,** 1907–1909.

Skelton, E. F., Ayers, J. D., Qadri, S. B., Moulton, N. E., Cooper, K. P., Finger, L. W., Mao, H. K., Hu, Z. (1991), *Science* **253,** 1123–1125.

Stishov, S. M., Popova, S. V. (1961), *Geochemistry (USSR)* **10,** 923–926.

Sung, C.-M., Goetze, C., Mao, H. K. (1977), *Rev. Sci. Instrum.* **48,** 1386–1391.

Teter, D. M., Hemley, R. J. (1996), *Science* **271,** 53–55.

Timofeev, Y. A. (1992), *Prib. Tekh. Eksper.* **5,** 186–189.

Togaya, M., Sugiyama, S., Mizuhara, E. (1994), in: S. C. Schmidt, J. W. Shaner, G. Samara, M. Ross (Eds.), *High-Pressure Science and Technology—1993,* New York: American Institute of Physics, pp. 255–258.

Vaughan, R. W., Lai, C. F., Elleman, D. D. (1971), *Rev. Sci. Instrum.* **42,** 626–629.

Vohra, Y. K. (1992), in: A. K. Singh (Ed.), *Recent Trends in High Pressure Research,* New York: International Science, pp. 349–358.

Vohra, Y. K., McCauley, T. S., Gu, G., Vagarali, S. S. (1994), in: S. C. Schmidt, J. W. Shaner, G. A. Samara, M. Ross (Eds.), *High-Pressure Science and Technology—1993,* New York: American Institute of Physics, pp. 515–518.

Vos, W., Finger, L. W., Hemley, R. J., Hu, J. Z., Mao, H. K., Schouten, J. A. (1992), *Nature* **358,** 46–48.

Vos, W. L., Finger, L. W., Hemley, R. J., Mao, H. K. (1993), *Phys. Rev. Lett.* **71,** 3150–3153.

Wei, L., Kuo, P. K., Thomas, R. L., Anthony, T. R., Banholzer, W. F. (1993), *Phys. Rev. Lett.* **70,** 3764–3767.

Weidner, D. J., Wang, Y., Vaughan, M. T. (1994), *Science* **266,** 419–422.

Wigner, E., Huntington, H. B. (1935), *J. Chem. Phys.* **3,** 764–770.

Wijngaarden, R. J., Ennige, E. N., Scholtz, J. J., Hemmes, H. K., Griessen, R. (1991), in: R. Pucci, G. Piccitto (Eds.), *Molecular Systems under High Pressure,* Amsterdam: Elsevier, pp. 157–180.

Williams, Q., Jeanloz, R., Bass, J., Svendsen, B., Ahrens, T. J. (1987), *Science* **236,** 181–182.

Wu, M. K., Ashburn, J. R., Torng, C. J., Hor, P. H., Meng, R. L., Gao, L., Huang, Z. J., Wang, Y. Q., Chu, C. W. (1987), *Phys. Rev. Lett.* **58,** 908–910.

Xu, J., Mao, H. K., Bell, P. M. (1986), *Science* **232,** 1404–1406.

Yamanaka, A., Inoue, K. (1989), *Rev. Sci. Instrum.* **60,** 3253–3257.

Yin, M. T., Cohen, M. L. (1983), *Phys. Rev. Lett.* **50,** 2006–2009.

Yoo, C. S., Akella, J., Campbell, A. J., Mao, H. K., Hemley, R. J. (1995), *Science* 1473–1475.

Yu, P. Y., Paul, W., Weinstein, B. A. (Eds.) (1995), "Proceedings of the 6th International Conference on High Pressure Semiconductor Physics," *J. Phys. Chem. Solids* **56,** 311–672.

Zaug, J., Abramson, E. H., Brown, J. M., Slutsky, L. J. (1992), in: Y. Syono, M. H. Manghnani (Eds.), *High-Pressure Research: Application to Earth and Planetary Sciences,* Tokyo: Terra Scientific; Washington, D.C.: American Geophysical Union, pp. 157–166.

Zha, C. S., Hemley, R. J., Mao, H. K., Duffy, T. S., Meade, C. (1994), *Phys. Rev. B* **50,** 13105–13112.

Further Reading

Bridgman, P. W. (1949), *The Physics of High Pressure,* London: G. Bell and Sons.

Hazen, R. M. (1993), *The New Alchemists,* New York: Random House.

Ichimaru, S., Ogata, S. (Eds.) (1995), *Elementary Processes in Dense Plasmas,* New York: Addison-Welsey.

Liu, L.-g., Bassett, W. A. (1986), *Elements, Oxides, Silicates,* New York: Oxford Univ. Press.

Schmidt, S. C., Shaner, J. W., Samara, G. A., Ross, M. (Eds.) (1994), *High-Pressure Science and Technology—1993,* New York: American Institute of Physics.

Wentorf, R. H. (Ed.) (1962), *Modern Very High Pressure Techniques,* Washington, D.C.: Butterworths.

SOLIDS

See CRYSTALLINE STATE; ELECTRON STRUCTURE OF SOLIDS; ELECTRONIC TRANSPORT IN SOLIDS; HEAT-PULSE PROPAGATION IN SOLIDS; INTERACTION OF SOLIDS WITH PARTICLES AND RADIATION; IONIC CONDUCTION AND DIFFUSION IN SOLIDS; MAGNETIC ORDERING IN SOLIDS; MATERIALS PREPARATION, SOLIDS; MECHANICAL AND ELASTIC WAVES IN SOLIDS; MECHANICAL PROPERTIES OF SOLIDS; PHASE TRANSFORMATIONS, STRUCTURAL; POINT DEFECTS AND EXTENDED DEFECTS IN CRYSTALS; RADIATION DAMAGE IN CRYSTALS; SOLIDS, ACOUSTIC PROPERTIES OF; SOLIDS, CRYSTALLINE, MECHANICAL PROPERTIES OF; SOLIDS, DIFFUSION IN; SOLIDS, DYNAMIC HIGH-PRESSURE EFFECTS IN; SOLIDS, STATIC HIGH-PRESSURE EFFECTS IN; SURFACES AND INTERFACES OF SOLIDS, STRUCTURE OF; THERMAL CONDUCTION IN SOLIDS

SOLUBILITY AND MIXING IN FLUIDS

ROBERT HOŁYST, *Institute of Physical Chemistry, Polish Academy of Sciences, Warsaw, Poland*

Introduction 573
1. **Thermodynamics** 574
1.1 Colligative Properties 574
1.2 Binary Mixtures: Liquid–Vapor Coexistence 575
1.3 Partially Miscible Liquids 576
2. **Statistical Mechanical Theories of Mixtures** 578
2.1 Homogeneous Systems and Ideal Mixtures 579
2.2 Homogeneous Systems and Excess Mixing Functions 579
2.3 Conformal Fluid Theories and the One-Fluid Approximation 580
3. **Interaction Potential and Mixing** 580
3.1 Van der Waals Interactions 581
3.2 Steric Interactions 581
3.3 Hydrogen Bonding 582
4. **Polymer Blends and Polymers in Solutions** 582
4.1 Polymer Blends 583
4.2 Polymers in Solutions 584
5. **Ordering and Demixing** 585
5.1 Mixtures of Liquid Crystals 585
5.2 Diblock Copolymers 586
5.3 Amphiphilic Systems 586
6. **Kinetics of Demixing** 588
Glossary 589
Works Cited 589

INTRODUCTION

The standard definition (Mc-Graw Hill Editors, 1984) of solubility is as follows: It is the ability of the substance to form a solution with another substance. The solution is a single homogeneous liquid, solid, or gas phase that is a mixture in which the components (liquid, gas, solid, or the combination thereof) are uniformly distributed throughout the mixture. Finally, miscibility denotes the tendency or capacity of two or more liquids to form a uniform blend, that is, to dissolve in each other. For historical reasons the substance less abundant in the mixture is called a solute, while the more abundant one is called a solvent. Here we shall concentrate on nonelectrolytes, i.e., solutions in which none of the components is in the form of free ions.

Traditionally, solubility and mixing belong to the realm of chemistry and material sciences, and many standard textbooks on physical chemistry (Nernst, 1904; Atkins, 1994) or chemistry (Grant and Higuchi, 1990; James, 1986) treat this problem, usually within the scope of thermodynamics. Here, apart from thermodynamics, we discuss other issues such as the statistical mechanical theories of mixtures, the relation between intermolecular interactions and demixing, the coupling between ordering and demixing, and the kinetics of demixing, which includes spinodal decomposition in binary liquid mixtures. The following special examples will be discussed: polymer blends, diblock copolymers, liquid crystals, and ternary mixtures including surfactants (amphiphiles). Within the scope of the physics of solubility and mixing, one can study such diverse phenomena as mixing of two simple liquids, collapse of the polymer chain in a solvent, flocculation of colloidal particles upon the addition of polymer chains, mixing of two polymer components in a liquid state, ordering of copolymers, and ternary mixtures of oil, water, and surfactant or formation of micelles in aqueous solutions. In reality, very rarely do we have completely pure substances; thus the properties of mixtures and the phenomena of mixing and demixing are of prime importance for science and technology.

An old alchemist's maxim, "similia similibus solvuntur" ("like dissolves like"), is the oldest rule of solubility. This rule can be a

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

very good guide in the study of mixing and solubility, provided one can precisely define what in a given case is the degree of likeness. For example, two simple liquids of low molecular mass can easily mix, say, at room temperature, while after polymerization they demix well above the room temperature.

1. THERMODYNAMICS

Solubility also means the maximum amount of the solute that can be solubilized in the solvent under given thermodynamic conditions. Within thermodynamics one formulates simple rules for solubility of gases and solids in liquids (Hildebrand and Scott, 1950):

1. The solubility of a gas is proportional to its partial vapor pressure (Henry's law).
2. The gas with the higher critical temperature and boiling point is more soluble than one with a lower critical temperature.
3. The solubility of a gas diminishes with increasing temperature.
4. The solubility of a solid increases with increasing temperature.
5. A solid having a higher melting point is less soluble at a given temperature than one having a lower melting point, provided that the enthalpies of melting are comparable.
6. A solid with large enthalpy of melting is less soluble at a given temperature than one with small enthalpy of melting, provided that the melting temperatures are comparable.
7. If rule **1** is satisfied for a solute in two immiscible solvents in contact, then the ratio of its concentration in these solvents is, for a given pressure and temperature, a constant (Nernst distribution law).

These rules are not always satisfied, sometimes only for very dilute solutions. Nonetheless, in many cases they provide very valuable information about solubilities of different substances in different solvents based on a few thermodynamic properties of pure substances. All these rules can be expressed in simple mathematical forms (Atkins, 1994; Hildebrant and Scott, 1950); e.g., rules **4–6** can be expressed as follows:

$$\ln x = \frac{-\Delta H^s_{\text{solute}}}{R}\left(\frac{1}{T} - \frac{1}{T_m}\right), \tag{1}$$

where x is the mole fraction of the solute in the saturated solution (solution in equilibrium with the solute solid), $\Delta H^s_{\text{solute}}$ is the enthalpy of melting for the solute, T_m is its melting temperature, R is the gas constant, and T is the temperature.

1.1 Colligative Properties

The elevation of the boiling point, the depression of the freezing point in a solution, and osmosis are properties that in the first approximation do not depend on the specific nature of the solute, but only on its amount in the solvent. Such properties are called colligative properties. For a dilute solution we find that the addition of the solute into the pure solvent raises its boiling temperature by

$$\Delta T = \left(\frac{RT_b^2}{\Delta H^v_{\text{solvent}}}\right)x, \tag{2}$$

where T_b is the boiling temperature of the pure solvent and $\Delta H^v_{\text{solvent}}$ is the enthalpy of vaporization for the pure solvent. Similarly the addition of a small amount of solute into the solvent lowers its freezing temperature by

$$\Delta T = -\left(\frac{RT_m^2}{\Delta H^s_{\text{solvent}}}\right)x, \tag{3}$$

where T_m is the freezing temperature of the pure solvent.

The phenomenon of osmosis (*q.v.*) takes place when a solution is separated from the pure solvent by a semipermeable membrane, which allows the flow of solvent from the pure phase to the solution. The pressure that has to be applied to stop this flow is called the osmotic pressure Π. For a dilute solution one gets the van't Hoff equation relating this pressure, the volume of the solution, V, and the number of moles of the solute in the solution n:

$$\Pi V = nRT. \tag{4}$$

Since n/V is given by the total mass of the solute divided by the molar mass M and the volume, combination of a measurement of the osmotic pressure and the van't Hoff formula allows one to determine the molar mass of the solute. This method is particularly useful for determining the mass of macromolecules; however, in this case it is often necessary to include in Eq. (4) the next term in the virial expansion of the osmotic pressure.

1.2 Binary Mixtures: Liquid–Vapor Coexistence

In an ideal mixture the components mix in all proportions without any change of volume or enthalpy. In the ideal binary mixture, the partial vapor pressure of component *A* (or *B*) coexisting with the ideal liquid mixture is given by $P_A = x_A(g)P$ $[P_B = x_B(g)P]$. The total pressure P is related to the (Raoult law) vapor pressures above pure *A* (P_A^*) and pure *B* (P_B^*) liquids by

$$P = P_A + P_B = x_A(l)P_A^* + x_B(l)P_B^*. \tag{5}$$

These two equations define the vapor $[x_A(g) = 1 - x_B(g)]$ and the liquid $[x_A(l) = 1 - x_B(l)]$ compositions at coexistence. In Fig. 1 the composition–pressure diagram is shown; the region between the two curves (given by the above equations) is the two-phase region, where the relative amounts of vapor and liquid are given by the lever rule (Atkins, 1994).

This type of diagram (now temperature and composition, Fig. 2) is the basis of fractional distillation. We start with the *A*, *B* mixture at point 1 on the diagram and evaporate it (1–2 dashed line). Next we condense it again until we reach point 3. The resulting liquid mixture is now richer in component *B*. Repeating these steps (evaporate it along the 3–4 dashed line, etc.), we can obtain almost pure *B* phase. The analogous process is used to purify solids and is called zone refining. In many mixtures the process of distillation is stopped at a certain point where the composition of the vapor is the same as the composition of the coexisting liquid (Figs. 3 and 4). This point is called the azeotrope. Since at this point the liquid boils without changing its composition, the fractional distillation proceeds only until the positive azeotrope is met (Fig. 4). In the example shown in Fig. 3, we reach a negative azeotrope by straight distillation, i.e., by continuously removing the vapor from the vessel of the boiling liquid mixture. For example, the mixture of ethanol and water has the azeotrope at $T = 78$ °C and at 4% of water (by mass); the mixture of water and nitric acid has the azeotrope at $T = 122.4$ °C and 60% of HNO_3 (mole fraction) (at 1 atm); that of

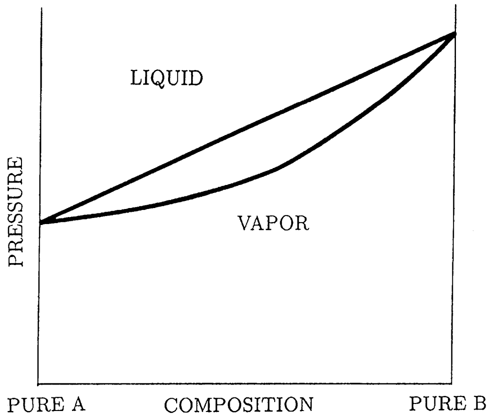

FIG. 1. The pressure–composition phase diagram for an ideal binary mixture (*A*,*B* components) [see Eq. (5)]. The region between the solid lines is the two-phase region where vapor and liquid coexist. Along the upper curve, evaporation of the liquid starts, and along the lower curve, condensation of the vapor starts.

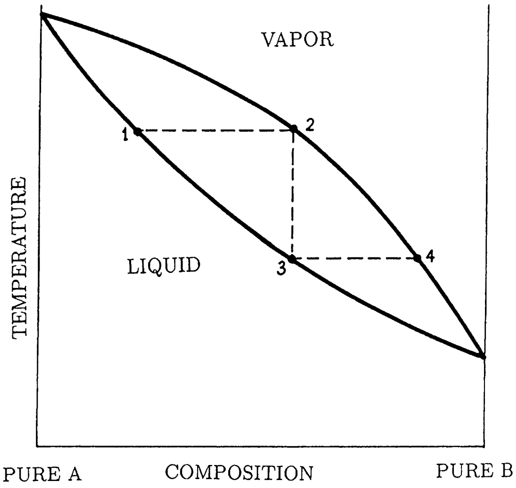

FIG. 2. The temperature–composition phase diagram for the binary mixture. The dashed line shows how the process of fractional distillation proceeds. In this way we can obtain almost pure (*B*-rich) liquid.

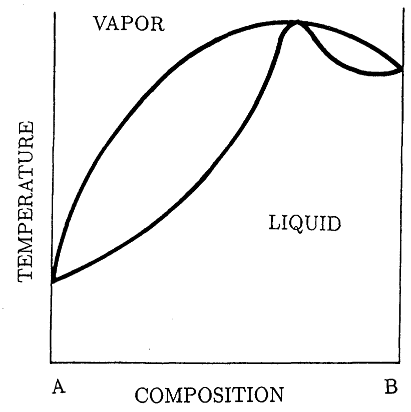

FIG. 3. A phase diagram with a high-boiling or negative azeotrope (maximum on the temperature–composition diagram). At this point liquid boils without changing its composition. This type of diagram occurs, for example, for water/nitric acid, chloroform/acetone, and hydrochloric acid/water.

hydrogen chloride and water has the azeotrope at $T = 108$ °C and 20% (by mass) of HCl (at 1 atm). Of course, the location of the azeotrope changes with pressure. The following rules can help us in deciding when we can expect an azeotrope in a liquid mixture (Rowlinson and Swinton, 1982):

1. The closer the vapor pressures of the pure components of the mixture, the more likely is azeotropy. It is inevitable at any temperature at which they are the same. It also means that when the vapor pressures are very close, then even small departures from ideality of the mixture can produce an azeotrope.
2. The closer the vapor pressures of the pure components, the more rapidly does the azeotropic composition change with temperature.
3. An increase of temperature and of vapor pressure in a positive azeotrope (maximum on the *p-x* diagram and minimum on the *T-x* diagram, Fig. 4) increases the mole fraction of the component whose vapor pressure increases more rapidly with temperature. The converse holds for a negative azeotrope (Fig. 3).

These rules are not without exceptions, but usually can be a very useful guide.

1.3 Partially Miscible Liquids

Mixtures are often far from being ideal, and consequently under suitable conditions we can expect demixing in the liquid state. In some systems demixing can take place as we lower the temperature (Fig. 5) or as we raise the temperature (Fig. 6). More complicated cases are shown in Figs. 7–10 (Atkins, 1994; Landau and Lifshitz, 1980). The maximum (minimum) of the curve in Fig. 5 (Fig. 6) is called the upper (lower) critical (or consolute) point. Gases usually mix very well at normal pressures; however, at very high

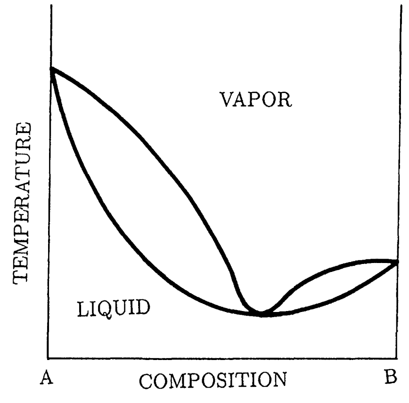

FIG. 4. A phase diagram with a low-boiling or positive azeotrope (minimum on the temperature–composition diagram). This type of diagram can occur for water/ethanol and dioxane/water.

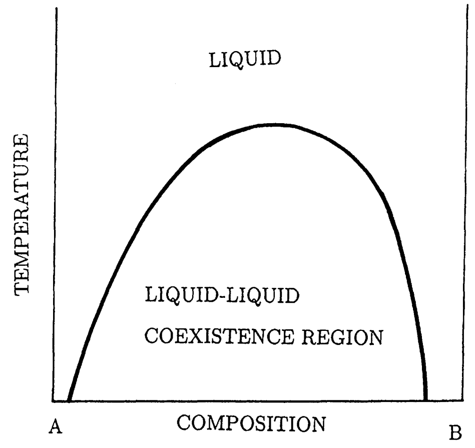

FIG. 5. Typical phase diagram for partially miscible liquids with an upper critical point (maximum on the coexistence curve) (e.g., methanol and hydrocarbon liquids).

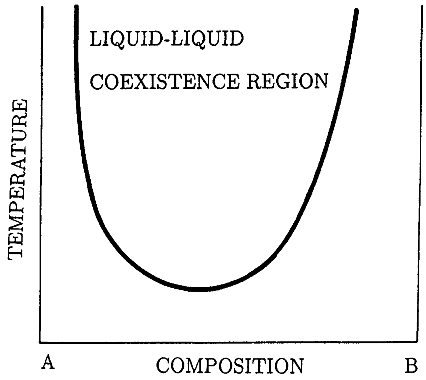
LIQUID-LIQUID
COEXISTENCE REGION
TEMPERATURE
A
COMPOSITION
B

FIG. 6. Typical phase diagram for partially miscible liquids with a lower critical point (minimum on the coexistence curve) (e.g., water and diethylamine or triethylamine).

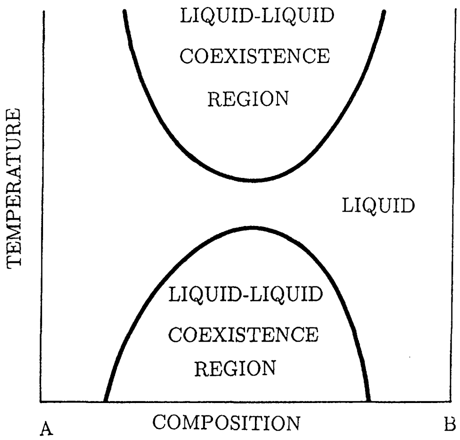
LIQUID-LIQUID
COEXISTENCE
REGION
LIQUID
TEMPERATURE
LIQUID-LIQUID
COEXISTENCE
REGION
A
COMPOSITION
B

FIG. 8. The phase diagram with a lower and an upper critical point and a miscibility gap between these point (e.g., acetone and polystyrene).

pressures they can demix; then sometimes the phrase gas–gas immiscibility is used (Rowlinson and Swinton, 1982). The densities of these fluid phases at high pressures are comparable to the density of the liquid phases at "ordinary" low-pressure critical points.

Figure 10 is very similar to the liquid–solid phase diagram for a binary mixture (instead of vapor we have liquid, and instead of liquid we have solid). The lowest temperature of the liquid mixture at freezing in this case is obtained for the eutectic composition (point "E" on the diagram).

For ternary mixtures the phase diagram is often represented in the form of the Gibbs triangle. The Gibbs phase rule states that in a system of r components and M coexistent phases it is possible arbitrarily to preassign $r - M + 2$ variables from the set T, P, x_j^i, $i = 1, \ldots, M$, $j = 1, \ldots, r - 1$ (Callen, 1960). In particular, it means that the maximum number of coexistent phases in the system is $r + 2$. One should note here that if the system spontaneously orders in an arbitrarily small external field (magnetic or electric), then the number of coexistent phases can be larger if we include these fields in the set of thermodynamic variables.

So far we have considered liquid mixtures

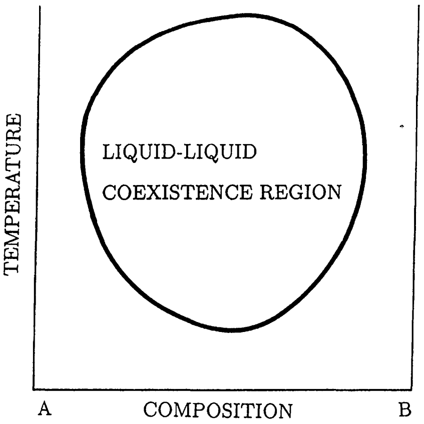
TEMPERATURE
LIQUID-LIQUID
COEXISTENCE REGION
A
COMPOSITION
B

FIG. 7. The phase diagram with upper and lower critical points (e.g., nicotine and water).

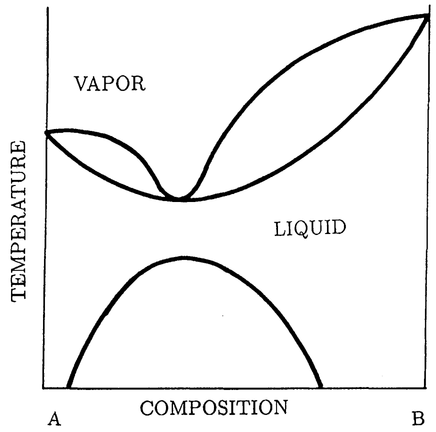
VAPOR
TEMPERATURE
LIQUID
A
COMPOSITION
B

FIG. 9. The phase diagram for a mixture where the upper critical point is below the boiling curve.

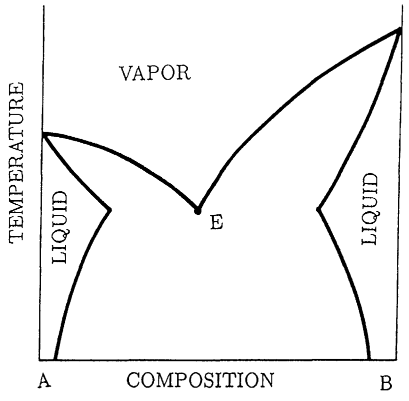

FIG. 10. The phase diagram for a mixture not fully miscible in the liquid state (e.g., water/diethyl ether, water/chloroform, and high hydrocarbons and methanol).

in a bulk system. The confinement of the fluid in a capillary also affects mixing and solubility. The detailed discussion of the thermodynamics of confined mixtures can be found in Evans and Marconi (1987). However, the theoretical problem of whether confinement increases or decreases solubility and mixing is still not resolved. The problem is especially important for narrow pores of the size of few molecular diameters.

2. STATISTICAL MECHANICAL THEORIES OF MIXTURES

The modern statistical mechanical approach to inhomogeneous and/or homogeneous fluids and fluid mixtures is based on density-functional theory (DFT) (Evans, 1979). The central quantity of interest in DFT is the Helmholtz free energy $F[\rho_1 \cdots \rho_m]$, which is a unique functional of the local densities $\rho_i(\mathbf{r})$ $(i = 1 \cdots m)$ in an m-component simple atomic mixture. In a homogeneous system, $\rho_i = N_i/V$, where N_i is the number of molecules of type i and V is the volume. The Helmholtz free energy as a functional of the densities completely specifies its statistical and thermodynamical properties. The equilibrium state of the system corresponds to the global minimum of the grand thermodynamic potential, obtained from F by a Legendre transform.

In the case of anisotropic liquids (e.g., liquid crystals) consisting of elongated rigid molecules, the one-particle distribution function $\rho(\mathbf{r},\omega)$, which depends on the position of the center of mass $\mathbf{r}$ of the molecule and its orientation in space ω, represented by the three Euler angles, describes the structure of the system. Consequently, the free energy of liquid crystals is a functional of $\rho(\mathbf{r},\omega)$. More complicated cases are also possible, depending on the structure and flexibility of molecules.

For simplicity we confine further discussion to simple atomic fluids. The free energy can be split into an ideal part and an excess part as follows:

$$F[\rho_1 \cdots \rho_m] = F_{id}[\rho_1 \cdots \rho_m] + F_{ex}[\rho_1 \cdots \rho_m]. \tag{6}$$

For the noninteracting system the excess part is zero, and the ideal part can be easily calculated from the partition function:

$$F_{id} = k_{\mathrm{B}}T \sum_{i=1}^{m} \int d^3r\rho_i(\mathbf{r})\{\ln[\rho_i(\mathbf{r})\lambda_1^3] - 1\}, \tag{7}$$

where λ_i is the thermal de Broglie wavelength of the ith component in the mixture. The integrals are performed over the whole volume of the system. The excess part, arising from the interactions, can be written in the following general form:

$$\begin{aligned} F_{ex}[\rho_1 \cdots \rho_m] = \sum_{i,j=1}^{m} & \int d^3r_1 d^3r_2 \rho(\mathbf{r}_1)\rho(\mathbf{r}_2) \\ & \times \int_0^1 dl(l - 1) \\ & \times c_{ij}(\mathbf{r}_1\mathbf{r}_2; [l\rho_1, \cdots, l\rho_m]). \end{aligned} \tag{8}$$

The functions c_{ij} are the Ornstein–Zernicke (OZ) direct correlation functions (Hansen and McDonald, 1986). Although this relation is formally exact, the form of the direct correlation functions is in general unknown for inhomogeneous and even for homogeneous liquids. Various approximations to the exact DFT functional have been proposed: the weighted density approximation (WDA) (Tarazona, 1985; Curtin and Ashcroft, 1985) and the modified weighted density approximation (MWDA) (Denton and Ashcroft, 1989) for pure systems, later applied to binary mixtures of hard spheres (Denton and Ashcroft, 1990, 1991) (for a different approach to mixtures, see Xu and Baus, 1987, 1992;

for a review, Baus, 1990). Density functional theories have been also applied to the description of structural phase transitions in liquid crystals (Poniewierski and Holyst, 1988, 1990; Allen *et al.*, 1993). Recent reviews on this subject are given by Löwen (1994) and Allen *et al.* (1993). In general, the theory has been aimed at explaining the properties of inhomogeneous systems from the known properties of homogeneous systems. Usually, the form of the direct correlation functions for homogeneous systems are required as the input for the theory. For the homogeneous mixture, these functions are defined via the Ornstein–Zernicke equation:

$$h_{ij}(|\mathbf{r}_1 - \mathbf{r}_2|) = c_{ij}(|\mathbf{r}_1 - \mathbf{r}_2|) + \sum_{k=1}^{m} x_k \rho \int d^3 r_3 c_{ik}(|\mathbf{r}_1 - \mathbf{r}_3|) \times h_{kj}(|\mathbf{r}_3 - \mathbf{r}_2|), \tag{9}$$

where $\rho = \Sigma_{i=1}^{m} N_i/V$ is the total density and h_{ij} is the two-point correlation function. This set of equations with suitable closure [e.g., Percus–Yevick (PY) for hard spheres; Henderson and Leonard, 1971] can be solved, and the functions c_{ij} can be calculated. Now combining these results together with MWDA completely specifies the properties of the inhomogeneous and/or homogeneous mixture.

In general, solving the OZ equation for mixtures interacting through a complicated potential is very difficult. One can then resort to some old methods known in the theory of mixtures as the one-fluid approximation in the more general scheme known as conformal theories of mixtures (Sec. 2.3) (Rowlinson and Swinton, 1982; Hansen and McDonald, 1986).

2.1 Homogeneous Systems and Ideal Mixtures

In homogeneous mixtures, one defines mixing functions; i.e., for any thermodynamic function $A(N_i)$, the mixing function is defined as

$$A^{\text{mix}} = A(N_1 \cdots N_m) - \sum_{i=1}^{m} A_i(N), \tag{10}$$

where

$$N = \sum_{i=1}^{m} N_i$$

is the total number of particles and A_i is the thermodynamic function of interest for a pure system of the ith component. The mixing function vanishes for a pure system. If the interactions between particles are the same, then the form of the thermodynamic potentials is very simple. The Helmholtz free energy (same for Gibbs free energy) is then [Eqs. (6–8)]

$$F^{\text{mix}} = Nk_BT \sum_{i=1}^{m} x_i \ln x_i. \tag{11}$$

Strictly speaking, real mixtures are nonideal, except in very special cases when we have mixtures of isotopes of low molecular mass and when quantum effects are negligible. However, often the ideal solution is a very good starting point in the description of many properties of mixtures—e.g., gas–liquid coexistence.

2.2 Homogeneous Systems and Excess Mixing Functions

The difference between the actual value of the thermodynamic mixing function and its ideal value [Eq. (11)], for the same temperature, composition, and volume (or pressure), is called the excess function {the same name is used in DFT [Eq. (8)], but its meaning there is different}. The simplest form of the mixing function for a binary mixture is given by the Guggenheim quadratic form (Henderson and Leonard, 1971; Rowlinson and Swinton, 1982),

$$F^E = N\chi x_1 x_2, \tag{12}$$

obtained in the simplest lattice approximation. In the lattice approximation both the excess Gibbs free energy and the excess Helmholtz free energy are the same, and χ/κ_BT is the interaction parameter independent of temperature and pressure (volume). Lattice models are often used for the description of certain properties of mixtures (Furman *et al.* 1977; Walker and Vause, 1983; Carneiro and Schick, 1988). By comparing Eqs. (10)–(12) and Eqs. (6)–(8), it follows that in general χ is a complicated integral of the direct correlation function. Only in the limit of low density is χ independent of x_i. In this limit $c_{ij} = f_{ij} = \exp(-v_{ij}/k_BT) - 1$, where $v_{ij}(|\mathbf{r}|)$ is the two-body interaction potential between molecules of components i and j. In this approximation χ can be explicitly calculated from Eq. (8).

Usually in thermodynamics one defines by Eq. (12) the Gibbs free energy G^E, assuming that χ is a function of temperature and pressure. Then the heat of mixing (enthalpy) is

$$H^E = N\left(\chi - T\frac{\partial\chi}{\partial T}\right)x_1x_2, \tag{13}$$

and the change of volume in the system upon mixing is

$$V^E = Nx_1x_2\frac{\partial\chi}{\partial P}. \tag{14}$$

Most mixtures of simple liquids (e.g., argon, krypton) have positive heat of mixing; i.e., heat is absorbed, but the excess volume V^E can be of either sign. The excess quantities are usually small; e.g., V^E per mole in a 50% solution (in mole fraction) of tetrachloroethene in cyclopentane at 25 °C is -8 mm^3 (the mixture contracts) (total volume is ~10 cm^3) and the excess heat of mixing H^E is 800 J/mol.

2.3 Conformal Fluid Theories and the One-Fluid Approximation

One of the oldest approaches to the theory of homogeneous mixtures dates back to van der Waals and is known as the one-fluid approximation. In this approach the excess properties of the mixture are expressed in terms of the quantities of a hypothetical pure fluid. It applies to a mixture in which the interaction potential $v_{ij}(r)$ satisfies the scaling relation

$$v_{ij}(r) = \epsilon_{ij}v(r/\sigma_{ij}) \tag{15}$$

for all i, j. Often σ_{ij} and ϵ_{ij} (for $i \neq j$) are related to pure components. The most frequently used combination rules are as follows: $\sigma_{ij} = (\sigma_{ii} + \sigma_{jj})/2$ and $\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}}$, known as the Lorentz and Berthelot rules, respectively. The hypothetical fluid is characterized by the potential given by Eq. (15) with some energy parameter ϵ and length parameter σ. In order to specify these parameters, one makes the approximation

$$h_{ij}(r) = h(r/\sigma), \tag{16}$$

and sets

$$\sigma^3 = \sum_{i,j=1}^{m} x_ix_j\sigma_{ij}^3. \tag{17}$$

The interaction parameter for this hypothetical one-component fluid reads

$$\epsilon = \sum_{i,j=1}^{m} x_ix_j\epsilon_{ij}\frac{\sigma_{ij}^3}{\sigma^3}, \tag{18}$$

and the interaction potential is $\epsilon v(r/\sigma)$. Now the excess mixing properties of the mixture are related to the excess properties of this fluid. This approximation is known as the van der Waals one-fluid approximation (Henderson and Leonard, 1971). The comparison with computer simulations for the model binary mixture of Lennard–Jones fluids shows a very good agreement with the theory (Hansen and McDonald, 1986). The approximation breaks down when the molecules of different components are very disparate in sizes and have very different energy parameters. Another of its drawbacks is that it always leads to a negative volume change V^E upon mixing. More complicated approximations are also possible but usually lead to worse agreement with computer simulations. For real systems, the simple Lennard–Jones mixture constitutes a rather poor approximation (Hansen and McDonald, 1986).

One has to remember that excess mixing quantities are rather small and that small deviations of the model interaction potential from the real potential between the molecules can change even the sign of the excess quantity (e.g., V^E).

3. INTERACTION POTENTIAL AND MIXING

The intermolecular forces have their origin in quantum mechanics and classical electrostatics. According to the Hellman–Feynman theorem, the interactions between the molecules can be calculated according to the laws of electrostatics, once the distribution of electrons has been determined from the Schrödinger equation. From the simple Coulomb potential originate various important interactions at the molecular level. The dispersion van der Waals attractive interactions are responsible for boiling (gas–liquid transition). It is now known that the reduction of the range of attraction in the system results in the complete disappearance of the liquid phase even if the attraction is strong at short distances (Tejero *et al.,* 1994). It implies, for example, that C_{60}, where interactions are of the short range, is a substance that should not have a liquid phase (Hagen *et al.,* 1993). Molecules, when brought together, strongly repel each other, preventing the overlap of electronic orbitals. This interaction gives the molecule its size and shape and is known as hard-core repulsion or steric interaction. Steric interactions are responsible for the freezing transition (Alder and Wainwright, 1957), the structure of liquids (Hansen and McDonald, 1986), and structural transitions in liquid crystals (Onsager, 1949; Frenkel, 1991). Van der Waals attraction and steric repulsion are very important since they are present in every molecular system. Apart from them there are many specific interactions. One interesting example considered here is hydrogen bonding occurring between, e.g., water molecules. This interaction is responsible for the facts that ice has lower density than liquid water and that the highest density of water occurs at 4 °C. The extremely high boiling and freezing points of water are also due to this interaction and the tetrahedral structure formed by water molecules. Here we try to answer partially the question of how these various interactions affect mixing.

3.1 Van der Waals Interactions

If we apply the Berthelot rule, which holds reasonably well for van der Waals forces (Israelachvili, 1985), in the Guggenheim model [Eq. (12)] for the binary mixture, we find

$$\chi \sim (\sqrt{\epsilon_{11}} - \sqrt{\epsilon_{22}})^2 > 0. \qquad (19)$$

Since χ is always greater than zero, we may conclude that van der Waals interactions promote demixing. At sufficiently low temperatures, the excess part of the Gibbs free energy outweighs the ideal part [Eq. (11)], and demixing occurs. With this simple model we obtain the type of phase diagram shown in Fig. 5. In the first approximation the square roots of the energies ϵ_{ii} are proportional to the polarizability of the molecules of ith component.

The van der Waals interactions for any two bodies in vacuum are always attractive. Also they are attractive for identical bodies in a solvent. However, one may have repulsive van der Waals interactions between two different solute molecules in a solvent. This happens whenever the index of refraction (related to polarizabilities) of the solvent is intermediate between those of two different solute molecules (Israelachvili, 1985). It happens in many mixtures of organic solvents and different organic polymers (Van Oss *et al.,* 1980).

3.2 Steric Interactions

It has been believed until recently that steric interactions alone cannot induce demixing. This belief has been based on the fact that the ideal entropy of mixing decreases if the demixing transition takes place. This decrease of the entropy S can be compensated by the decrease of the energy U so that the free energy $F = U - TS$ decreases upon mixing. Since in the hard-core system the interaction energy is zero, one has not expected the demixing transition in such a system. This conclusion is supported also by the explicit calculation of the parameter χ [Eq. (12)] for a binary fluid mixture of hard spheres in the low-density approximation. One finds $\chi < 0$ in this limit and concludes that mixing is favored in the system.

Recent computer simulations have shown that this is not the case and that in a binary mixture of a model system of hard-core molecules differing in size only, the demixing transition driven by the increase of entropy occurs. Although the ideal entropy is the largest in a homogeneous mixture, the entropy associated with the free volume is larger in the demixed fluid (Biben and Hansen, 1991; Frenkel and Louis, 1992; Van Duijneveldt and Lekkerkerker, 1993; Dijkstra and Frenkel, 1994; Dijkstra *et al.,* 1994; Frenkel, 1994). At a certain concentration of solute molecules, the latter outweighs the former, and demixing occurs. The role of steric interactions in real mixtures has not been resolved yet. Clearly, it is a many-body effect.

3.3 Hydrogen Bonding

The unique properties of water follow from its ability to form a tetrahedral structure in the liquid, induced by hydrogen bonding (Israelachvili, 1985). The interaction is electrostatic, but its strength (10–40 kJ/mol) is one order of magnitude larger than the strength of the van der Waals energy (~1 kJ/mol). This interaction depends on the orientation of molecules and in general is not pairwise additive.

The solubilization of molecules that do not form hydrogen bonds (e.g., alkanes and other hydrocarbons) in a hydrogen-bonding solvent (methane, water, acetic acid) involves forming a clathrate cage around the solute molecule (Fig. 11). The solvent molecules adapt such an orientation close to the nonbonding solute as to saturate its hydrogen bonds with other solvent molecules. It means that solvent molecules around the solute molecules are more ordered than the molecules in the bulk solvent. It leads to the decrease of entropy in the process of solubilization, which for this reason becomes unfavorable. For *n*-butane in water the heat of mixing is $\Delta H = -4.2$ kJ/mol, while the change of the entropy $-T\Delta S$ (at $T = 25$ °C) is 28.7 kJ/mol. Thus, the change of entropy contributes 85% to the change of the Gibbs free energy $\Delta G = \Delta H - T\Delta S$. For longer hydrocarbons the contribution of entropy to the total change of the Gibbs free energy is even larger (Israelachvili, 1985). The low solubility of non-hydrogen-bonding molecules in water is entropic in nature and is known as a hydrophobic effect. The solubility of hydrocarbons in a hydrogen-bonding solvent (e.g., methane) decreases with the length of the hydrocarbon; e.g., the critical temperature for *n*-pentane is $T_c = 287$ °C, while for *n*-hexane it becomes $T_c = 307$ °C (at normal pressure) (Rowlinson and Swinton, 1982).

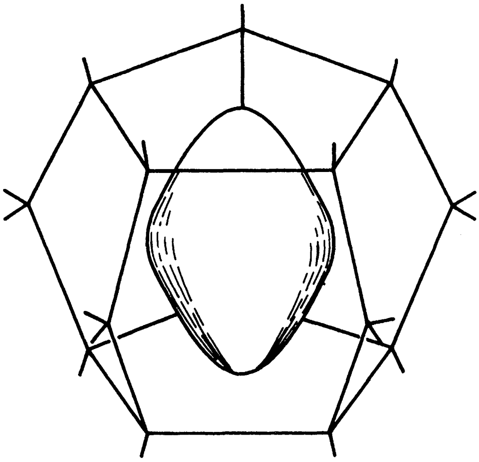

FIG. 11. The clathrate cage around a non-hydrogen-bonding solute molecule. The lines indicate the hydrogen bonds between the solvent molecules.

Methane and water mix very well in all proportions at room temperature, and this is due to the fact that both form hydrogen bonds. We expect that molecules of the right geometry, containing electronegative atoms (oxygen atoms in alcohols, nitrogen atoms in amines), are capable of forming hydrogen bonds. At low temperatures the hydrogen bonds forming between the solute and solvent molecules enhance mixture ($\chi < 0$). If we raise the temperature, the bonds are broken and the liquids demix above the lower critical temperature as a result of van der Waals forces ($\chi > 0$). Thus, hydrogen bonds are responsible for enhanced mixing at low temperatures and consequently for the existence of the lower critical point (Fig. 6) (Walker and Vause, 1983). We expect that this phenomenon should strongly depend on the molecule geometry.

4. POLYMER BLENDS AND POLYMERS IN SOLUTIONS

The simple lattice model presented in Sec. 2 [Eq. (12)] has been applied to polymer blends and polymers in solutions by Flory and Huggins (Flory, 1953). The free energy for the binary mixture of two homopolymers *A* and *B* consisting of N_A and N_B monomers, respectively, is given by the Flory–Huggins expression:

$$\frac{F^{\mathrm{mix}}}{Mk_{\mathrm{B}}T} = \frac{x_A}{N_A}\ln x_A + \frac{x_B}{N_B}\ln x_B + \frac{\chi}{k_{\mathrm{B}}T}x_Ax_B. \tag{20}$$

Here $x_A = n_AN_A/(n_AN_A + n_BN_B)$ is the mole fraction of A monomers in the mixture; n_i is the number of polymer molecules of component i (A or B); $M = n_AN_A + n_BN_B$ is the total number of monomers in the system, and χ, in polymer physics and chemistry, is called the Flory–Huggins interaction parameter. In the Flory–Huggins theory, incompressibility is assumed, just as in the Guggenheim model of simple mixtures. The critical temperature is

$$T_c = \frac{2N_AN_B}{k_B\chi(\sqrt{N_A} + \sqrt{N_B})^2}, \tag{21}$$

and at the critical composition

$$x_A^c = \frac{\sqrt{N_B}}{\sqrt{N_A} + \sqrt{N_B}}. \tag{22}$$

The ideal mixing entropy is reduced by a factor N_A and/or N_B, and, thus, the critical temperature is very large for polymer mixtures (N_A, $N_B \gg 1$). Indeed, polymers separate very easily. Even small differences in the interaction potential get magnified by a factor of N_A, N_B, and, thus, the critical temperatures are often much higher than the temperature at which the polymer molecule breaks through the breakage of chemical bonds linking the monomers. Even a binary mixture of isotope polymers (one deuterated) can undergo a demixing transition at room temperature if N_A, N_B are sufficiently large (Gehlsen *et al.*, 1992).

A mesoscopic theory of polymer chains has been formulated by Edwards (1966) and applied to polymer blends, polymers in solutions, membranes in solutions, etc.

4.1 Polymer Blends

Despite the fact that the Flory–Huggins theory has been used for more than four decades, only recently has it been carefully checked experimentally (Bates *et al.*, 1988; Gehlsen *et al.*, 1992) and in computer simulations (Deutsch and Binder, 1992). It follows from computer simulations that χ parameter has the following form:

$$\chi/k_BT = \alpha/T + \beta, \tag{23}$$

where α and β are constants independent of temperature or composition. In the first approximation α is related to the attractive van der Waals forces, while β is related to steric interactions (called sometimes excluded-volume interactions). Sometimes, in experimental works, it is assumed that α depends linearly on the concentration x (Roe and Zin, 1984). Typically, we find $\alpha \sim 1$ and $\beta \sim 10^{-3}$–10^{-4}. From the previous section we know that the effective interaction χ is related to the integrals of the direct correlation functions (Schweizer and Curro, 1988), and thus χ can be, in principle, a complicated function of temperature, density, and concentrations. In general, χ should not depend on the global architecture of the polymer chain.

In polymer systems the quantity of interest is the radius of gyration, i.e., the linear size of the region occupied by a single polymer molecule consisting of N monomers of length l. It is defined as follows (Doi and Edwards, 1986):

$$R^2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\langle(\mathbf{r}_i - \mathbf{r}_j)^2\rangle. \tag{24}$$

Here $\mathbf{r}_i$ gives the location of the ith monomer in the polymer chain and $\langle\cdots\rangle$ denotes the statistical average. For the polymer blend the dependence of R on N is (the leading term in N) as follows:

$$R = \sqrt{Nl^2/6}, \tag{25}$$

characteristic for the noninteracting chain. It means that interactions between monomers in the same polymer chain are screened by the presence of other chains (Doi and Edwards, 1986; De Gennes, 1979) or a solvent (Edwards, 1975). As the temperature is lowered, the scaling of R with N does not change. However, the chains shrink progressively (R decreases) as the critical temperature is approached (Hołyst and Vilgis, 1994).

One can deduce the structure of a polymer blend from scattering experiments. The scattering intensity $S(q)$ has the following form (De Gennes, 1979):

$$S^{-1}(q) = \frac{1}{x_Ag(N_A,q)} + \frac{1}{x_Bg(N_B,q)} - 2\frac{\chi}{k_BT}, \tag{26}$$

where $g(N,q) = 2N(y + e^{-y} - 1)/y^2$ and $y =$

$Nq^2l^2/6$. The contrast for neutron scattering is achieved by deuteration of one of the components. This approximation is known as the random-phase approximation (RPA). The characteristic length in the problem is given by the radius of gyration R [Eq. (25)].

From the theoretical point of view, it is interesting to note that although the critical point in a binary mixture of polymers belongs to the universality class of the Ising model (see UNIVERSALITY; PHASE TRANSITIONS: RENORMALIZATION AND SCALING), in the limit $N \to \infty$ the mean-field theory becomes exact. In particular, the critical exponents measured for the system have the mean-field values for $|T - T_c|/T_c \geq 1/N$ (De Gennes, 1977; Hołyst and Vilgis, 1993). This behavior has been confirmed in experiments (Schwahm *et al.,* 1987).

Polymer blends are compressible mixtures (Floudas *et al.,* 1993), but the role of this factor in the Flory–Huggins theory is still under theoretical study (Lifschitz *et al.,* 1994).

4.2 Polymers in Solutions

In the theory and experiment one often discusses three types of solvents: good, bad, and theta solvents. Since for $N_A \gg N_B = 1$ the critical concentration is very low ($x_A^c \sim \sqrt{1/N_A}$), we can expand F^{mix} in x_A and find for this dilute solution

$$\frac{F^{\text{mix}}}{Mk_BT} = \frac{x_A}{N_A} \ln x_A + Bx_A^2 + \tfrac{1}{6}x_A^3 \cdots, \tag{27}$$

where $B = \frac{1}{2} - \chi/k_BT$ is the osmotic virial coefficient (Atkins, 1994; De Gennes, 1979). For $B > 0$ the solution is classified as bad, for $B < 0$ it is classified as good, and for $B = 0$ it is classified as a theta or Flory solvent. The nature of the solvent depends on temperature. For each solvent there is a unique temperature (called the theta temperature) when $B = 0$ and the solution becomes nearly ideal. In a theta solvent the chains behave, in the first approximation, according to Eq. (25). In a good solvent the polymer molecule swells and

$$R \sim N_A^\nu \tag{28}$$

with the Flory exponent $\nu \approx \frac{3}{5}$, which should be compared with $\nu = \frac{1}{2}$ [Eq. (22)] for a polymer chain in a blend and a polymer chain in a theta solution. In a bad solvent the chains shrink and $\nu \leq \frac{1}{2}$. In fact, close enough to the critical temperature each solvent becomes a bad solvent. The Flory–Huggins theory predicts correctly the location of the critical point (De Gennes, 1979; Hołyst and Vilgis, 1993) but not the shape of the coexistence curve close to the critical point. More information on polymers in solutions can be obtained in a recent monograph (Des Cloizeaux and Jannik, 1990).

It is often assumed that the theta temperature must reflect the competition between the attractive van der Waals forces and the short-range steric interactions. However, it does not have to be the case. As shown explicitly by Frenkel and Louis (1992), in a simple lattice model of a polymer chain consisting of monomers interacting through steric interactions only and solvent consisting of small molecules interacting through the same steric interaction, the Flory–Huggins parameter is

$$\chi/k_BT = \tfrac{1}{2}c \ln(1 + z_s) > 0. \tag{29}$$

Here c is the coordination number of the lattice and z_s is the fugacity of the solvent molecules. Thus, in this system with purely steric interactions, we may have a bad, good, or theta solvent by changing the chemical potential of the solvent (Dijkstra *et al.,* 1994). The mechanism of demixing in such a mixture is clear. The fraction of the volume accessible to small solvent particles increases when the large particles cluster. Of course, in this model system, demixing is a purely entropic effect. Also the coagulation (clustering) of colloidal particles in a polymer solution is an entropic effect (Meijer and Frenkel, 1994). The clustering of colloidal particles increases the total number of accessible polymer configurations. The effective attraction between the colloidal particles, responsible for coagulation, is not pairwise additive (Shaw and Thirumalai, 1991), and that is why its quantitative theoretical description is so difficult. The role of steric interactions in polymer solutions is not fully understood.

We note that in the process of preparation of a polymer blend, one first prepares a solution of A and B polymers in a solvent and then evaporates the solvent. The comparison of the theory and experiment is often plagued by the inevitable polydispersity of the polymer masses. And finally, we point out that often instead of using mole fraction one uses either volume fraction or simply the density of monomers in the description of the free energy of mixing in polymer solutions (Atkins, 1994; De Gennes, 1979).

As we see, polymers are not easily mixed even at high temperatures. Usually chemical means are used to induce mixing. One of them is a chemical modification of one of the components in such a way as to induce hydrogen bonds between the different components. For example, PSD (deuterated polystyrene) and PBMA [poly(butyl methacrylate)] are not miscible. However, modifying PSD by attaching at random OH groups along the chain makes the PSD(OH)/ PBMA blend miscible. This is because OH groups from PSD(OH) form hydrogen bonds with CO groups on any segment of PBMA (Hobbie *et al.*, 1994). Another way to induce miscibility is even more direct. One can simply join one polymer chain with the other by chemical methods. In this way diblock and multiblock copolymers are formed.

5. ORDERING AND DEMIXING

There are many examples of demixing induced by ordering and vice versa. These effects occur in complex fluids such as diblock copolymers (Bates, 1991; Bates and Fredrickson, 1990), amphiphilic systems (Gompper and Schick, 1994; Laughlin, 1994), and mixtures of liquid crystals (De Gennes and Prost, 1993).

5.1 Mixtures of Liquid Crystals

For commercial applications in the display industry, only mixtures of liquid crystals are used. By mixing two nematic liquid crystals, one can lower the freezing point (eutectic composition; see Sec 1) without strongly affecting the isotropic–nematic phase transition point. Such a mixture exhibits anisotropic properties in the liquid state in a much wider temperature range than the pure substances. Usually the point at which the nematic–isotropic transition occurs is called the clearing point.

Miscibility is used as a first test for new phases in liquid crystals. If the two materials give the same texture and are miscible in all proportions maintaining this texture, they must have the same symmetry. However, we have learned that demixing does not have to reflect differences in the symmetry of phases. Differences in the shape and size of molecules are sufficient to induce demixing. Nonetheless, the miscibility test is still very useful, since it is quick and simple.

NMR allows us to measure internuclear distances of solute molecules if the solvent is nematic. Elongated isomers are more readily solubilized by nematic solvents, and this effect is used in chromatography.

Doping the nematic solvent with a chiral solute induces a cholesteric (twisted) ordering in the system, with a very large pitch (the wavelength of the twist). Changing the composition of this mixture allows us to vary the pitch continuously. The pitch, in a dilute solution, is inversely proportional to the concentration of the solute molecules. In chiral, ferroelectric smectic C^*, one can increase the pitch by dissolving in the system a nonchiral solute.

In a homogeneous solution of small monomeric units and nematic solvent, the fast polymerization reaction between the monomers results in demixing and leads to the formation of polymer-dispersed liquid crystals (PDLC), which are used as switchable windows, light shutters, and displays. Demixing is induced here since the length of the polymeric component of the mixture grows in the polymerization reaction, and according to Eq. (21) the critical temperature gets higher for longer chains. If the reaction is fast, droplets of the nematic solvent, which form in the process of demixing, are homogeneously distributed in the solid polymeric matrix. Usually the reactions require more components such as curing agents, catalysts, *etc.* (Doane, 1990).

The extremely rich phase behavior of liquid crystals and liquid-crystalline polymers has been described theoretically in several models: the lattice spin model (Sivardiére, 1980), the Onsager model for elongated molecules (Deblieck and Lekkerkerker, 1980; Van Roij and Mulder, 1994), and a combination of the Flory–Huggins model for mixtures and the Maier–Saupe model for liquid crystals (Brochard *et al.*, 1984; Maïssa and Sixou, 1989; Hołyst and Schick, 1992a). The following results follow from these theoretical studies. The nematic–isotropic phase transition in mixtures is always accompanied by partial demixing. The nematic phase is always richer in the component that orders more easily. The azeotrope at the nematic–isotropic phase transition is found in binary mixtures where both components have very similar isotropic–nematic transition temperatures in pure systems. The azeotrope concentration is very close to the critical concentration for demixing inside the nematic phase. All the diagrams shown in Figs. 1–10 are found in binary mixtures of liquid crystals if instead of vapor we have the isotropic phase and instead of liquid the nematic phase. The topology of many diagrams is even more complicated (Brochard *et al.*, 1984). The nematic order parameter (De Gennes and Prost, 1993) always couples to the concentration. The demixing at the isotropic–nematic phase transition is driven by ordering; however, inside the nematic phase we can also observe isotropic demixing (Figs. 5 and 9), which changes the nematic order parameter. The demixing inside the ordered nematic mixture has the same origin (isotropic van der Waals or steric interactions) as the demixing shown in Fig. 5 for a simple isotropic liquid mixture. The roles of the anisotropic steric (shape) interactions and attractive interactions in the process are not fully understood. Demixing in an ordered nematic liquid mixture (Casagrande *et al.*, 1982; Dorgan and Soane, 1990) or ordered smectic liquid mixture (Sigaud *et al.*, 1990) has been observed in experiments, confirming many of the above-stated predictions.

5.2 Diblock Copolymers

Polymers are usually not miscible (Sec. 3), but at the same time for practical applications we need homogeneous mixtures. One way to prevent a demixing transition is the chemical bonding of the *A* polymer to the *B* polymer. In this way diblock and multiblock copolymers are formed (Bates and Fredrickson, 1990). Although phase separation is now prohibited on the macroscopic scale, the system can undergo so-called microphase separation. The "demixing" in the *AB* diblock copolymer system takes place at the scale given by the size of the radius of gyration and at much lower temperatures than the demixing in the polymer *AB* blend. The system forms many ordered phases in a liquid state: lamellar, gyroid (Matsen and Schick, 1994), diamond, cubic, and hexagonal. The microscopic interactions responsible for the formation of these phases are the same as those responsible for the demixing transition in the *AB* polymer blend (Leibler, 1980), and the same Flory–Huggins parameter χ is used in the description of both systems. The *AB* diblock copolymer molecules form *A*-rich and *B*-rich domains, and chemical bonds joining *A* and *B* molecules reside at the interface between the domains (Fig. 12). The molecules are stretched in the ordered phase at low temperature, and the characteristic period of the ordered structures scales with the number of monomers in a molecule, N, as $N^{2/3}$. The mixing of the *A* and *B* homopolymers can be enhanced by the addition of the *AB* diblock copolymer. The latter has a similar effect on the *AB* polymer blend as an amphiphilic molecule on the mixture of oil and water, although amphiphiles can be more effective in enhancing mixing between immiscible liquids (Hołyst and Schick, 1992b).

5.3 Amphiphilic Systems

Amphiphilic molecules consist of two parts: a polar head that can form hydrogen bonds and a hydrocarbon tail. Such mole-

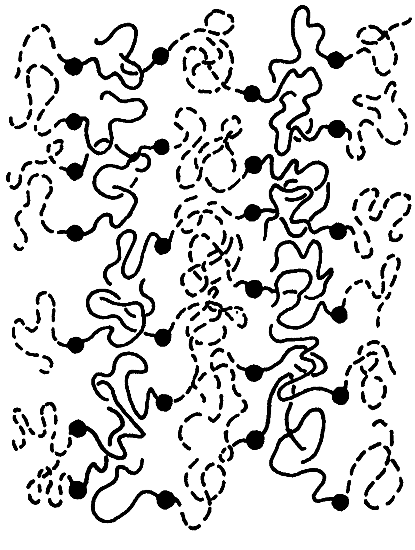

FIG. 12. Lamellar phase (periodic in one direction) of the *AB* diblock copolymer. The dots denote the position of the points where an *A* block (dashed line) is chemically joined to a *B* block (solid line).

cules are not easily solubilized in water because of their long tails. At a certain, usually low, concentration called critical micelle concentration, the amphiphilic molecules aggregate and form closed structures called micelles. The interior of the micelle consists of the hydrocarbon tails and has the structure and density of the hydrocarbon liquid. The polar heads reside on the surface of the micelle, shielding solvent molecules (water) from the interior hydrocarbon liquid. The geometrical shape of the micelles depends on the maximum length of the hydrocarbon chain, l_c, the cross-sectional area occupied by the polar head, a_0, and the hydrocarbon volume, v (Israelashvili, 1985; Degiorgio, 1985). For example, spherical micelles can form for $v/a_0l_c < \frac{1}{3}$, nonspherical micelles (in particular cylindrical) for $\frac{1}{3} < v/a_0l_c < \frac{1}{2}$, and flat bilayers or vesicles for $\frac{1}{2} < v/a_0l_c < 1$. They are usually polydisperse in size. As we see, the hydrogen bonding and hydrophobic effect are responsible for the formation of these structures. When the concentration of amphiphilic molecules becomes comparable to the concentration of the water solvent, certain ordered, bicontinuous phases are formed such as simple cubic (Fig. 13), diamond (Fig. 14), and gyroid structures (Fig. 15). The internal surfaces of these structures are formed by the polar heads. The structures are periodic in three dimensions and the internal surfaces may assume the configuration of minimal surfaces. The minimal surface is characterized by zero mean curvature at every point; each point on this surface is a saddle point.

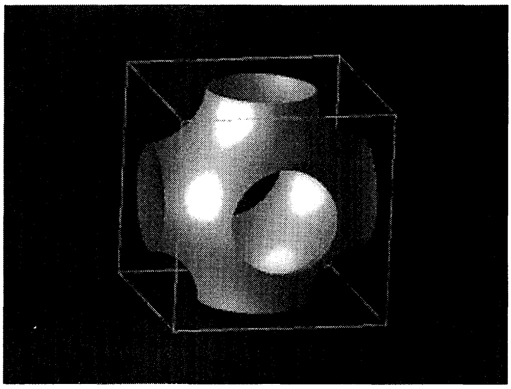

FIG. 13. The cubic bicontinuous phase in the amphiphilic system. Symmetry $Pm\bar{3}m$. Known as the Schwartz *P* triply periodic minimal surface. Here only the unit cell is shown. The internal interfaces are shown only. In the binary mixture of water and amphiphiles, water is on both sides of the surface. In the ternary mixture of oil, water, and amphiphile, we have oil on one side of the surface and water on the other. The interface is formed by the amphiphilic molecules. Water (oil) forms interconnected channels that span the whole volume.

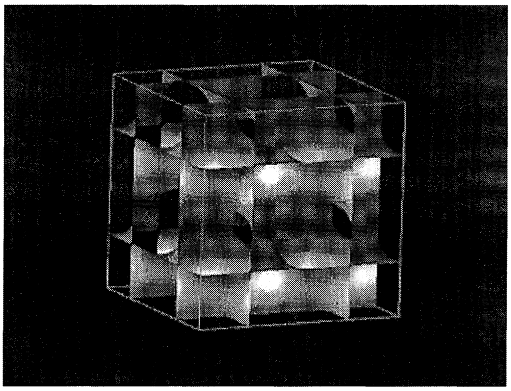

FIG. 14. The cubic bicontinuous phase in the amphiphilic system. Symmetry $F\bar{4}3m$. Known as the Schwartz *D* triply periodic minimal surface. Legend as in Fig. 13.

Amphiphilic molecules strongly reduce the surface tension between hydrocarbon liquid and water and enhance mixing between these two immiscible liquids. One characteristic structure formed in the ternary mixture of oil, amphiphiles, and water is the microemulsion. This homogeneous phase is characterized experimentally by enhanced scat-

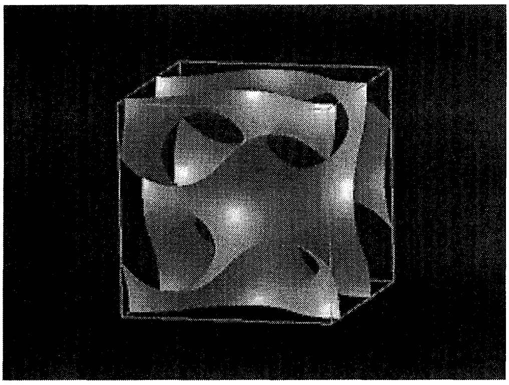

FIG. 15. The gyroid structure. Known as the Schoen *G* surface. Symmetry $Ia\bar{3}d$. Legend as in Fig. 13.

tering at a certain nonzero $\mathbf{q}$ vector. The form of the water–water scattering intensity

$$S(\mathbf{q}) = (a + g\mathbf{q}^2 + c\mathbf{q}^4)^{-1} \tag{30}$$

with negative g fits the data from scattering experiments very well (Teubner and Strey, 1987). It reveals the internal structure of the microemulsion with a characteristic length of order of 400 Å (the size of the amphiphile molecule is about 30 Å). Inside the microemulsion, water-rich regions of the size of the characteristic length are separated by the amphiphiles from oil-rich regions of a similar size. On the macroscopic scale, the system can be characterized by the large amount of these internal water–oil interfaces. Various surface and bulk properties of microemulsions are described by Ciach (1992) and Gompper and Schick (1994). The simplest Landau model that can be used for the description of bulk and surface properties of microemulsion is given by the following functional:

$$F[x(\mathbf{r})] = \int d^3r[(\Delta x)^2 + g(x)(\nabla x)^2 + f(x)], \tag{31}$$

where Δ denotes the Laplacian and ∇, the gradient,

$$g(x) = g_2x^2 - g_0, \tag{32}$$

and

$$f(x) = x^2(x - 1)^2(x + 1)^2. \tag{33}$$

Here x is the local concentration difference between oil and water, $f(x)$ is the simplest Landau bulk free energy for the three phases (pure water, pure oil, and microemulsion) at coexistence, and g_2 and g_0 are constants. Interestingly, the bicontinuous phases shown in Figs. 13–15 correspond to the local minima of this functional.

6. KINETICS OF DEMIXING

When a homogeneous binary AB mixture above its critical point is suddenly cooled (quenched) below its critical temperature, it ceases to be in thermodynamical equilibrium [for a review on kinetics, see Langer (1992)]. The homogeneous state can be now either a metastable or an unstable state. In the case of a metastable state, the process of demixing requires, in the first place, nucleation of droplets of the minority phase, say the A-rich phase. Then the droplets start to grow. At first they grow independently, and their radius $L(t)$ changes with time t according to the formula

$$L(t) \sim \sqrt{t}. \tag{34}$$

This behavior has been observed in binary polymer blends (Cumming *et al.*, 1990). At a later stage a significant fraction of A molecules disappear from the homogeneous mixture (most of them form droplets), and competitive growth starts. Small droplets decrease in size, and the A atoms diffuse toward large droplets, which grow. This mechanism is known as the "evaporation–condensation" mechanism and is described by the Lifshitz–Slyozov–Wagner (LSW) law for the growth rate of large droplets:

$$L(t) \sim t^{1/3}. \tag{35}$$

When the system is quenched into the thermodynamically unstable region, the demixing proceeds via the spinodal decomposition mechanism. Early stages of this process are described by the Cahn–Hilliard theory. According to the theory the system becomes unstable with respect to small fluctuations, of wave vector $\mathbf{q}$ smaller than some value $\mathbf{q}_0$. The key prediction of the theory is the exponential growth of the scattering intensity $S(\mathbf{q},t)$ in time, with a well-defined maximum at $\mathbf{q}_{\max} = \mathbf{q}_0/\sqrt{2}$. Interpenetrating A-rich and B-rich domains of the size of $L \sim 1/\mathbf{q}_{\max}$ are formed. The mixtures, of low viscosity, undergoing spinodal decomposition do not remain for a long time in their unstable early configuration, in contrast to the high–molecular-weight polymer blends. The latter are very viscous liquids, and the whole kinetics of phase separation is very slow, allowing a detailed experimental observation of the process (Bates and Wiltzius, 1989). In the late stages of spinodal decomposition, the scattering intensity can be represented by the following scaling form:

$$S(\mathbf{q},t) \approx L^d(t)Y(\mathbf{q}L(t)) \times \text{const}, \tag{36}$$

where Y is a scaling function, d is the dimension of space, and $L(t)$ is the time-dependent length characterizing the size of the interpenetrating A-rich and B-rich regions when the volume fractions are equal. $L(t)$ grows in time, the pattern coarsens, and the interfacial area, whose energy drives the coarsening, decreases. In the case of unequal volume fractions, $L(t)$ is the size of A (minority phase) droplets [Eq. (35)]. The late-stage configuration depends on the initial volume fractions, but not on the early-stage mechanism of demixing. In the late stage we have either a dilute gas of droplets of the minority phase in the sea of the majority phase, a dense system of the droplets of one of the phases in a sea of the another, or an interpenetrating network of A-rich and B-rich domains, which coarsen in time. The growth rate and the growth mechanism for the first case is described by the LSW law [Eq. (35)]. In the second case a Binder–Stauffer mechanism of collisions and coalescence of droplets is valid. It gives $L(t) \sim t^{1/3}$, similarly as in the first case. Finally, in the last case, if the coarsening proceeds via the flow induced by the surface tension of interfaces, the scaling law $L(t) \sim t$ holds. The scaling function Y and the late stage of spinodal decomposition have been studied in computer simulations (Koga and Kawasaki, 1993). Although the field of kinetics of demixing or, in general, of first-order phase transitions is rather old, new experimental results suggest that the problem is far from being understood. First of all, a new mechanism of coarsening called "collision-induced collision" has been observed (Tanaka, 1994). It has been also shown that surface effects and confinement play an important role in spinodal decomposition (Tanaka, 1993; Jones *et al.*, 1991; Wiltzius and Cumming, 1991). Finally, the problem of heat release during demixing has been addressed in experimental studies (Bailey and Cannell, 1993). Concluding: the kinetics of demixing is still an active field of research.

GLOSSARY

Azeotrope: The point on the phase diagram where two different phases of same composition coexist. Usually, in simple liquids the two phases are vapor and liquid.

Eutectic Composition: The liquid mixture composition with the lowest freezing point.

Excess Thermodynamic Quantities: The difference between actual mixing functions and ideal mixing functions.

Ideal Mixture: Mixture in which all the components mix in all proportions without change of volume or enthalpy. Interactions between various components in the mixture are the same [see also Eq. (11)].

Miscibility: The tendency or capacity of two or more liquids to form a uniform blend, that is, to dissolve in each other.

Mixing Function: The difference between a function defined for the mixture and the sum of these functions for pure systems. The mixing function vanishes for the pure system [Eq. (10)].

Solubility: The ability of the substance to form a solution with another substance. It also denotes the maximum amount of the solute that can be solubilized in the solvent at the given thermodynamic conditions.

Solute: The substance less abundant in the mixture.

Solution: A single homogeneous liquid, solid, or gas phase that is a mixture in which the components (liquid, gas, solid, or the combination thereof) are uniformly distributed throughout the mixture.

Solvent: The substance most abundant in a mixture.

Spinodal Decomposition: Process of phase separation of the mixture in the region (of the phase diagram) of thermodynamic instability of the mixture.

Upper (Lower) Consolute (or Critical) Temperature for a Liquid Mixture: Temperature, for the binary mixture of A and B components, above (below) which A and B components mix in all proportions; i.e., they exhibit full miscibility.

Works Cited

Alder, B. J., Wainwright, T. E. (1957), *J. Chem. Phys.* **27,** 1208–1209.

Allen, M. P., Evans, G. T., Frenkel, D., Mulder, B. M. (1993), in: I. Prigogine, S. A. Rice (Eds.), *Advances in Chemical Physics,* Vol. 36, New York: Wiley, pp. 1–166.

Atkins, P. W. (1994), *Physical Chemistry,* Oxford: Oxford Univ. Press.

Bailey, A. E., Cannell, D. S. (1993), *Phys. Rev. Lett.* **70,** 2110–2113.

Bates, F. S. (1991), *Science* **251,** 898–905.

Bates, F. S., Fredrickson, G. H. (1990), *Annu. Rev. Phys. Chem.* **41,** 525–557.

Bates, F. S., Wiltzius, P. (1989), *J. Chem. Phys.* **91,** 3258–3274.

Bates, F. S., Muthukumar, M., Wignall, G. D., Fetters, L. J. (1988), *J. Chem. Phys.* **89,** 535–544.

Baus, M. (1990), *J. Phys. Cond. Matter* **2,** 2111–2126.

Biben, T., Hansen, J. P. (1991), *Phys. Rev. Lett.* **66,** 2215–2218.

Brochard, F., Jouffroy, J., Levinson, P. (1984), *J. Phys. (Paris)* **45,** 1125–1136.

Callen, H. B. (1960), *Thermodynamics,* New York: Wiley.

Carneiro, G. M., Schick, M. (1988), *J. Chem. Phys.* **89,** 4368–4373.

Casagrande, C., Veyssié, M., Finkelmann, H. (1982), *J. Phys. (Paris) Lett.* **43,** L671–L675.

Ciach, A. (1992), *Polish J. Chem.* **66,** 1347–1381.

Cumming, A., Wiltzius, P., Bates, F. S. (1990), *Phys. Rev. Lett.* **65,** 863–866.

Curtin, A., Ashcroft, N. W. (1985), *Phys. Rev. A* **32,** 2909–2919.

Deblieck, R., Lekkerkerker, H. N. W. (1980), *J. Phys. (Paris) Lett.* **41,** L351–L355.

De Gennes, P. G. (1977), *J. Phys. Lett.* **38,** L441–L443.

De Gennes, P. G. (1979), *Scaling Concept in Polymer Physics,* Ithaca, New York: Cornell Univ. Press.

De Gennes, P. G., Prost, J. (1993), *The Physics of Liquid Crystals,* Oxford: Clarendon Press.

Degiorgio, V. (Ed.) (1985), *Physics of Amphiphiles: Micelles, Vesicles and Microemulsions,* Proceedings of the International School of Physics "Enrico Fermi," Course XC, Amsterdam: North Holland.

Denton, A. R., Ashcroft, N. W. (1989), *Phys. Rev. A* **39,** 4701–4708.

Denton, A. R., Ashcroft, N. W. (1990), *Phys. Rev. A* **42,** 7312–7329.

Denton, A. R., Ashcroft, N. W. (1991), *Phys. Rev. A* **44,** 8242–8248.

Des Cloizeaux, J., Jannik, G. (1990), *Polymers in Solutions: Their Modelling and Structure,* Oxford: Clarendon Press.

Deutsch, H.-P., Binder, K. (1992), *Macromolecules* **25,** 6214–6230.

Dijkstra, M., Frenkel, D. (1994), *Phys. Rev. Lett.* **72,** 298–300.

Dijkstra, M., Frenkel, D., Hansen, J. P. (1994), *J. Chem. Phys.* **101,** 3179–3189.

Doane, J. W. (1990), in: B. Bahadur (Ed.), *Liquid Crystals: Applications and Uses,* Singapore: World Scientific, pp. 362–396.

Doi, M., Edwards, S. F. (1986), *The Theory of Polymer Dynamics,* Oxford: Clarendon Press.

Dorgan, J. R., Soane, D. S. (1990), *Mol. Cryst. Liq. Cryst.* **188,** 129–146.

Edwards, S. F. (1966), *Proc. Phys. Soc.* **88,** 265–280.

Edwards, S. F. (1975), *J. Phys. A* **8,** 1670–1680.

Evans, R. (1979), *Adv. Phys.* **28,** 143–200.

Evans, R., Marconi, M. B. (1987), *J. Chem. Phys.* **86,** 7138–7148.

Flory, P. (1953), *Principles of Polymer Chemistry,* Ithaca, NY: Cornell Univ. Press.

Floudas, G., Pakula, T., Stamm, M., Fischer, E. W. (1993), *Macromolecules* **26,** 1671–1675.

Frenkel, D. (1991), in: J. P. Hansen, D. Levesque, J. Zinn-Justin (Eds.), *Liquids, Freezing and Glass Transition,* Ecole d'été de physique théorique, Session LI, Amsterdam: North Holland, pp. 689–762.

Frenkel, D. (1994), *J. Phys. Cond. Matter* **6,** A71–A78.

Frenkel, D., Louis, A. A. (1992), *Phys. Rev. Lett.* **68,** 3363–3365.

Furman, D., Dattagupta, S., Griffiths, R. B. (1977), *Phys. Rev. B* **15,** 441–464.

Gehlsen, M., Rosendale, J. H., Bates, F. S., Wignall, G. D., Lotte, H., Almdal, K. (1992), *Phys. Rev. Lett.* **68,** 2452–2455.

Gompper, G., Schick, M. (1994), in: C. Domb, J. L. Lebowitz (Eds.), *Self-Assembling Amphiphilic Systems,* Vol. 16, *Phase Transitions and Critical Phenomena,* New York: Academic, pp. 1–181.

Grant, D. J. W., Higuchi, T. (1990), *Solubility Behavior of Organic Compounds,* New York: Wiley.

Hagen, M. H. J., Meijer, E. J., Mooij, G. C. A. M., Frenkel, D., Lekkerkerker, H. N. W. (1993), *Nature* **365,** 425–426.

Hansen, J. P., McDonald, I. R. (1986), *Theory of Simple Liquids,* New York: Academic.

Henderson, D., Leonard, P. J. (1971), in: D. Henderson (Ed.), *Physical Chemistry (an Advanced Treatise),* Vol. VIIIB, New York: Academic, pp. 414–510.

Hildebrand, J. H., Scott, R. L. (1950), *Solubility of Nonelectrolytes,* 3rd ed., New York: Reinhold Publishing Corp.

Hobbie, E. K., Bauer, B. J., Han, C. C. (1994), *Phys. Rev. Lett.* **72,** 1830–1833.

Hołyst, R., Schick, M. (1992a), *J. Chem. Phys.* **96,** 721–729.

Hołyst, R., Schick, M. (1992b), *J. Chem. Phys.* **96,** 7728–7737.

Hołyst, R., Vilgis, T. A. (1993), *J. Chem. Phys.* **99,** 4835–4844.

Hołyst, R., Vilgis, T. A. (1994), *Phys. Rev. E* **50,** 2087–2092.

Israelachvili, J. N. (1985), *Intermolecular Interactions and Surface Forces,* New York: Academic.

James, K. C. (1986), *Solubility and Related Properties,* New York: Dekker.

Jones, R. A. L., Norton, L. J., Kramer, E. J., Bates, F. S., Wiltzius, P. (1991), *Phys. Rev. Lett.* **66,** 1326–1329.

Koga, T., Kawasaki, K. (1993), *Physica A* **196,** 389–415.

Landau, L. D., Lifshitz, E. M. (1980), *Statistical Physics,* 3rd ed., Pt. 1, Oxford: Pergamon Press.

Langer, J. S. (1992), in: C. Godrèche (Ed.), *Solids Far from Equilibrium,* Cambridge, U.K.: Cambridge Univ. Press, pp. 298–362.

Laughlin, R. G. (1994), *The Aqueous Phase Behavior of Surfactants,* New York: Academic.

Leibler, L. (1980), *Macromolecules* **13,** 1602–1617.

Lifschitz, M., Dudowicz, J., Freed, K. F. (1994), *J. Chem. Phys.* 3957–3978.

Löwen, H. (1994), *Phys. Rep.* **237,** 251–324.

Maïssa, P., Sixou, P. (1989), *Liq. Cryst.* **5,** 1861–1873.

Matsen, M. W., Schick, M. (1994), *Phys. Rev. Lett.* **72,** 2660–2663.

Mc-Graw Hill Editors (Eds.) (1984), Dictionary of Chemical Terms, 3rd ed., New York: McGraw-Hill.

Meijer, E. J., Frenkel, D. (1994), *J. Chem. Phys.* **100,** 6873–6887.

Nernst, W. (1904), *Theoretical Chemistry,* New York: MacMillan.

Onsager, L. (1949), *Proc. NY Acad. Sci.* **51,** 627–655.

Poniewierski, A., Hołyst, R. (1988), *Phys. Rev. Lett.* **61,** 2461–2464.

Poniewierski, A., and Hołyst, R. (1990), *Phys. Rev. A* **41,** 6871–6880.

Roe, R.-J., Zin, W.-C. (1984), *Macromolecules* **17,** 189–194.

Rowlinson, J. S., Swinton, F. L. (1982), *Liquids and Liquid Mixtures,* Guilford, UK: Butterworth Scientific.

Schwahm, D., Mortensen, K., Yee-Madeira, H. (1987), *Phys. Rev. Lett.* **58,** 1544–1546.

Schweizer, K. S., Curro, J. G. (1988), *Phys. Rev. Lett.* **60,** 809–812.

Shaw, M. R., Thirumalai, D. (1991), *Phys. Rev. A* **44,** R4797–R4800.

Sigaud, G., Nguyen, H. T., Achard, M. F., Twieg, R. J. (1990), *Phys. Rev. Lett.* **65,** 2796–2799.

Sivardiére, J. (1980), *J. Phys.* **41,** 1081–1089.

Tanaka, H. (1993), *Phys. Rev. Lett.* **70,** 2770–2773, 3524.

Tanaka, H. (1994), *Phys. Rev. Lett.* **72,** 1702–1705.

Tarazona, P. (1985), *Phys. Rev. A* **31,** 2672–2679.

Tejero, C. F., Daanoum, A., Lekkerkerker, H. N. W., Baus, M. (1994), *Phys. Rev. Lett.* **73,** 752–755.

Teubner, M., Strey, R. (1987), *J. Chem. Phys.* **87,** 3195–3200.

Van Duijneveldt, J. S., Lekkerkerker, H. N. W. (1993), *Phys. Rev. Lett.* **71,** 4264–4266.

Van Oss, C. J., Absolom, D. R., Neumann, A. W. (1980), *Colloid. Surf.* **1,** 45–56.

Van Roij, R., Mulder, B. (1994), *J. Phys. II France* **4,** 1763–1769.

Walker, J. S., Vause, C. A. (1983), *J. Chem. Phys.* **79,** 2660–2676.

Wiltzius, P., Cumming, A. (1991), *Phys. Rev. Lett.* **66,** 3000–3003.

Xu, H., Baus, M. (1987), *J. Phys. C* **20,** L373–L380.

Xu, H., Baus, M. (1992), *J. Phys. Cond. Matter* **4,** L663–L668.

SOLUBILITY AND SEGREGATION IN ALLOYS

CHARLES WERT, *Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, U.S.A.*

Introduction 593
1. **Thermodynamics of Alloys** 594
2. **Concentrated Solutions** 598
3. **Gas–Metal Equilibrium** 600
4. **Compounds** 602
5. **Ordering** 603
6. **Precipitation of Second Phases** 603
7. **Rapidly Solidified Alloys** 607
8. **Mechanical and Dielectric Spectroscopy** 608
9. **Conclusions** 613
Glossary 613
Works Cited 614
Further Reading 615

INTRODUCTION

The recorded use of materials dates back more than 200 centuries, and so the sciences of physics and chemistry are only recently on the scene. Thus, it is instructive to begin the discussion by noting how ancient craftsmen made useful materials, especially of metals. We can see how both solutions of elements and preparation of segregated phases produced desirable products. To be sure, pottery and glass were also important, but the inability of craftsmen to alter their chemistry and internal structure appreciably made their form and surface design more important. Thus, major eras have been named the Bronze Age and the Iron Age, not the pottery or glass age. Three examples of metallic alloys illustrate how more complex products were developed as understanding increased of the chemistry and the internal arrangement of atoms: gold, copper, and iron.

Gold and copper (along with silver) were first acquired in nearly pure form. At first, they could be altered only by pounding (to harden them) or by heating (to soften them). Gold in nearly pure form was hammered into thin sheets to stretch its usefulness. It was later alloyed with mercury to form an amalgam that could be spread in thin layers on copper, brass, or bronze. A thin coating of gold was formed after evaporation of the mercury, a coating whose thinness and quality rivals modern electroplated coatings. Silver and copper are often found associated with native gold, and further, both form complete solid solutions with gold; craftsmen around the world developed alloys comparable to the 10- to 22-carat gold alloys presently used for jewelry and for coins.

Copper was also used in pure form, hammered and annealed to produce a desired strength and hardness. It was alloyed with tin—found also in pure form, but also widely present in cassiterite, which could be easily reduced by a wood fire—to provide a bronze with properties comparable to modern bronzes. Brass was also produced by heating copper in the presence of zinc carbonate, whose dissociation provided zinc. Finally, middle eastern craftsmen and much later, Peruvian Indians, made copper–arsenic alloys with good casting properties.

Iron has an even more complex history. It was first obtained in metallic form from iron-rich meteorites, in which it is always alloyed with nickel. Iron oxides can be reduced in a wood fire only with a forced air stream, and only gradually was a means found to provide a spongy iron-slag product from iron oxides. It could then be transformed into more useful materials by heating—particularly to cast irons and later to steels. Craft practice far predated an understanding of the role of carbon in strengthen-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ing steels—the tortuous path of this understanding is well described in the books of Smith (1962, 1981) and of Aitchison (1960). To note how recently modern steels have developed, recall that the beams of the Eiffel Tower are made from a rolled iron containing strings of slag—wrought iron.

These examples—and many others that might be cited—describe ancient alloys that are either solid solutions or mixtures of solutions and segregated phases. That they resulted in useful products is a remarkable testimony to the acumen of early craftsmen.

Materials development still provides a vast frontier of opportunity. Within the past 50 years, countless new materials with novel combinations of solutions and segregated phases have entered the marketplace. Tiny "shear pins" hold a filling of my tooth in place; a light-cured epoxy fills in a chipped spot on an incisor, color-matched to the adjacent enamel. Small additions of impurities to silicon make the computer possible. Silica glass fibers of extraordinary purity transmit voice or data signals with low loss using infrared radiation—the small loss inherently present can be made up by laser amplification every 50 km in short lengths of fiber containing Er in solution. Super-bright light-emitting diodes (LEDs) are bright enough to serve as warning lights in the back windows of automobiles; they are highly efficient and have a lifetime of perhaps 100 years. The bearings and bearing races in my 50-year-old bicycle seem unworn—even after some 60 000 miles. My father could expect to get 5000 miles from a set of tires in 1925; now I am unhappy if I don't get 50 000. A modern jet aircraft engine could make about 200 trips around the world at the equator between major overhauls.

These and a thousand other possible examples describe the enormous gains that have been made in materials science and engineering in the past 50 years. Steady gains were made through the centuries, of course, but a chief factor in the recent extraordinary progress has been the avalanche of new techniques of examination of materials—especially the microscopic, spectroscopic, and diffraction techniques that have come from physics. This section describes the nature of solutions and of materials containing segregated phases and shows how applications from science have aided in their specification.

Part of our understanding of solutions and compounds is based on chemistry, part on physics. Yet practitioners of both of these sciences have great difficulty in predicting a precise combination of elements that will produce a useful product. Practical alloy development is still largely the domain of the technologist, and most useful alloys are an exquisite combination of elements heated in a particular way to satisfy a specific need. Reading the vast literature on materials demonstrates forcefully, though, that the science and technology are merging into a continuum of practitioners.

1. THERMODYNAMICS OF ALLOYS

Most materials used technologically for products are in a metastable state, not in a thermodynamically stable configuration. Nevertheless, it is useful to start with the *thermodynamics* of the stable state.

A metal like copper can dissolve other elements by substituting them on sites in the host crystal; the extent of solubility may be calculated by application of the laws of thermodynamics. Let the incremental change in Gibbs free energy, ΔG, in substitution of atoms of B for the host atoms A be expressed by two terms, an incremental change in enthalpy, ΔH, and temperature times a change in entropy, $T\Delta S$:

$$\Delta G = \Delta H - T\Delta S. \tag{1a}$$

The change in entropy may be divided into two terms, an entropy of mixing and an excess entropy, ΔS_{excess}, occasioned by the possible alteration of the vibrational states of atoms in the crystal by the alloying addition. The first may be calculated from statistics of atom placement, and so we may write

$$\Delta G = \Delta H + RT(X_a \ln X_a + X_b \ln X_b) - T\Delta S_{\text{excess}}. \tag{1b}$$

Here X_a is the atom fraction of element A; X_b is that of element B. Several possibilities exist in specification of the minimum in ΔG, depending on the relative values of the several terms. A few special cases are described in the following paragraphs.

The enthalpy change ΔH may be zero, or it may be so small that its effect on ΔG is outweighed by the entropy term for the temperature range of interest. Then the components of Eq. (1b) may yield a monotonic negative value of ΔG for all compositions, and complete solubility exists. An example of this condition for the terms of Eq. (1b) (with ΔS_{excess} set equal to zero) is shown in Fig. 1(a), and the corresponding phase diagram for the binary alloy AuAg (which satisfies this condition) is shown in Fig. 1(b). In both the liquid and the solid alloys, complete miscibility exists for all compositions. Upon melting, the two-phase region demanded by thermodynamics is remarkably narrow, as can be seen; this results from similarity of electronic structure and nearly equal atomic radii of the two elements. This complete solid miscibility results in the presence of naturally appearing solutions of various composition valuable to early craftsmen—historians lump them together under the term *electrum.*

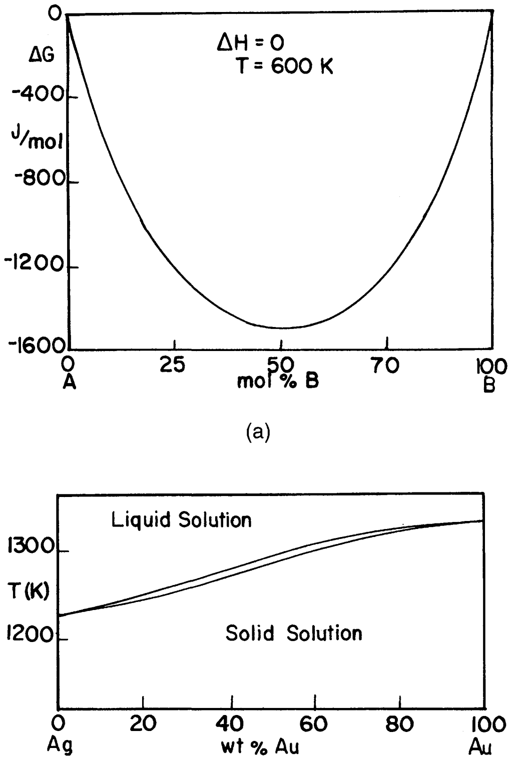

(b)

FIG. 1. (a) Free energy vs concentration for $\Delta H = 0$ and $T = 600$ K. (b) Phase diagram for the Au–Ag system.

Not many alloy systems have such a simple phase diagram. A more common case is one in which ΔH is large enough for concentrated solutions so that a mixed-phase region exists over a large range of the solid alloys. The thermodynamic parameters for that case are sketched in the upper part of Fig. 2. The enthalpy increase on addition of the alloying element at either end of the concentration range is large enough that ΔG goes through a minimum at specific concentrations and then increases. For alloys of intermediate concentration, a solid solution is energetically unfavorable with respect to separation into two phases whose concentrations are given by the common tangent line to the curve of ΔG vs concentration. Those concentrations c_1 and c_2 are functions of temperature—this sketch is appropriate for a temperature of about 340 K for the values of ΔH assumed. The large entropy contribution to ΔG for liquid alloys causes a depression of the freezing point for alloys of intermediate concentration. A typical phase diagram for such a system is that drawn for Pb–Sn, shown in the lower half of Fig. 2. This alloy system includes the technologically impor-

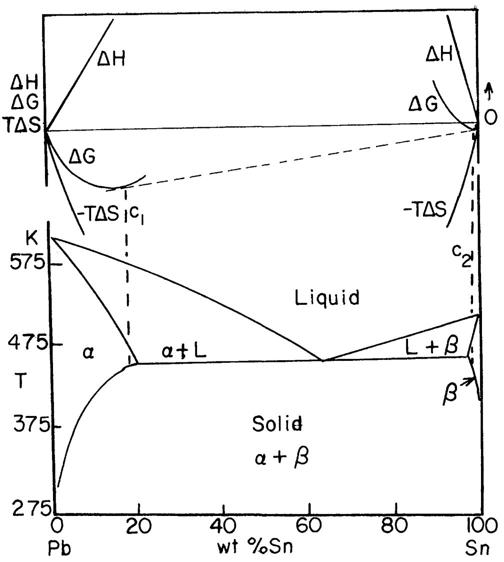

FIG. 2. Phase diagram for Pb–Sn. A eutectic. The upper part of the figure shows the components of Eq. (1) at the extremes of the concentration range. For dilute enough alloys, solid solutions exist. But for concentrated alloys, a mixed-phase region is energetically favorable.

tant solders. The lowest-melting alloy at the eutectic composition 60% Sn, 40% Pb is the common soft solder; compositions away from that point remain mushy during a large part of their solidification. Such phase diagrams also exist for ceramic combinations; complete solubility exists for the system NiO–MgO, Fig. 3(a), and a eutectic for the system MgO–CaO, Fig. 3(b). Observe that such portrayals for ceramic combinations are usually expressed as relationships between compounds, not between the constituent elements.

The "eutectic alloy system," which we (in cold climates) perhaps use more widely than any other, though, is that of ice and salt (NaCl). At one atmosphere of pressure, ice melts at 32 °F, NaCl at 1474 °F. When salt is put in contact with ice, a solution results, which lowers the melting point as a salt-water solution forms. Finally, when the water solution reaches 23 wt % salt, the temperature has fallen to −6 °F. This is the eutectic point. Adding more salt will not lower the water temperature further: the minimum is −6 °F (at a concentration of 23% salt). We all know that, at least in general, for we are aware that salting our sidewalks and streets does not help if the temperature is "too cold."

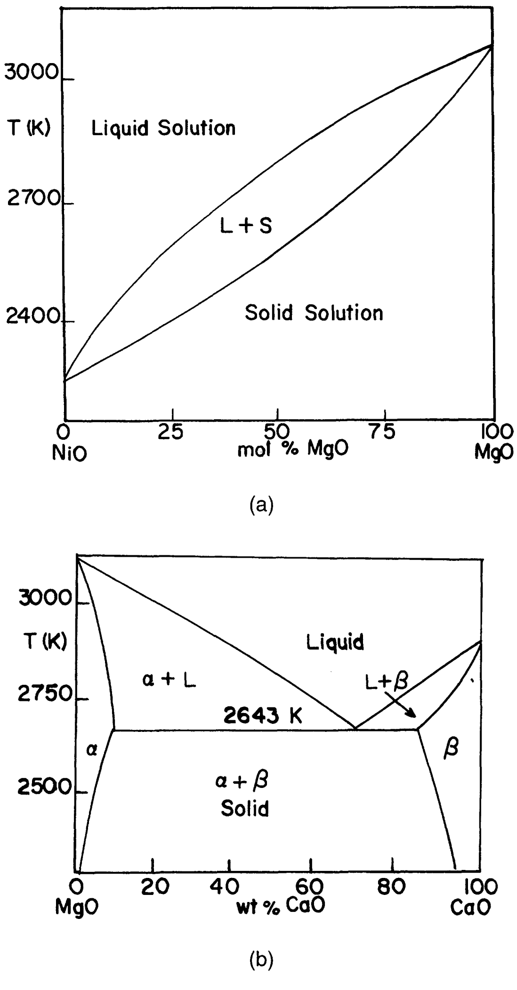

FIG. 3. Oxide phase diagrams: (a) MgO–NiO, (b) MgO–CaO.

Countless examples of binary and multicomponent phase diagrams (and others with complex sets of intermediate compounds) have been determined and published, mostly by metallurgists and ceramists. First-principles calculations have correctly predicted a few phase diagrams, but their number is so small compared with the vast number of possible combinations of alloying elements that such successes are not much more than demonstrations of the correctness of the physical models. In addition, virtually all commercial materials are multicomponent combinations not easily amenable to modeling, and so lines in binary diagrams (and surfaces in multiconstituent systems) have usually been determined experimentally. Again, though, microscopy and the diffraction and spectroscopic techniques of physics are of inestimable value in determination of the character of phases present in alloy systems.

The extremes of such phase diagrams are often technologically important. For dilute solution of B in A, X_a is about unity, and X_b may be expressed as the ratio of the number of B atoms to the total number of atoms in the solid, N. If the enthalpy required to introduce an atom B is a constant δh, then, at equilibrium,

$$X_b = n/N = \exp(-\delta h/RT) \qquad (2a)$$

or in logarithmic form,

$$\ln(n/N) = -\delta h/RT. \qquad (2b)$$

Thus, the limiting solubility of B in A is exponential in $1/T$ and δh is the unit enthalpy of solution. Countless important phase diagrams have limiting solid solutions that fit that equation. Not only do the phase boundaries of Pb in Sn and Sn in Pb fit that expression, but so also do such alloys such as Cu, Zn, and many other elements dissolved in Al, alloys that provide the important hardened duralumins. Beryllium dissolved in Cu has a limiting solubility line obeying this equation; proper heat treatments provide a strong spring alloy, as well as nonmagnetic tools, useful around large magnets. They are also nonsparking tools, so that they are useful in potentially explosive gas atmospheres. Carbon, nitrogen, and oxygen dissolved in such metals as Fe, Ta, Nb, and V obey this equation. Examples of alloys whose free energy obey Eq. (2) are nearly endless.

Equation (2) also describes the solubility of vacancies and other point defects dissolved in metals, insulators, and semiconductors. Investigations of δh for these point defects—usually in the range 0.5–3 eV per defect—have been an important part of diffusion studies in solids. See the data for the concentration of single vacancies in Cu, plotted according to Eq. (2b) in Fig. 4; δh is about 1.3 eV/vacancy.

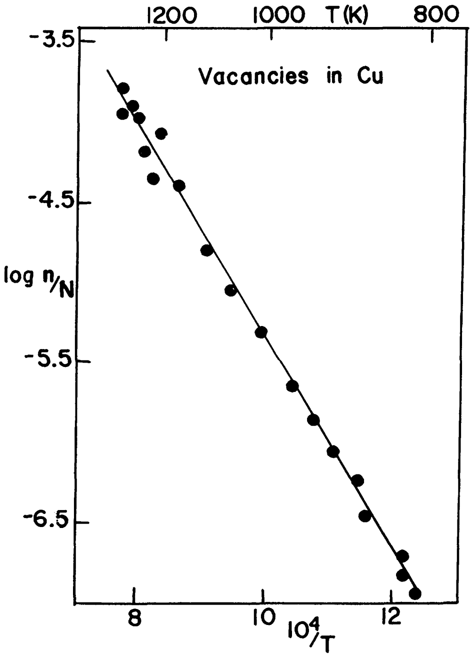

FIG. 4. Equilibrium concentration of vacancies in Cu. The data points are a small fraction of reported values. The slope of the line is the energy of formation, 1.3 eV per vacancy.

An important factor of Eq. (1) is the way in which ΔG approaches zero as the concentration of element B approaches zero. We can see the key features with the aid of the upper curves of Fig. 2. Since ΔH approaches zero linearly with X_b and the entropy term approaches zero with infinite negative slope, a minimum must always exist in ΔG for some finite concentration, no matter how large is the term ΔH. Thus, impurities always exist in thermodynamic equilibrium in solids, and costly purification schemes must often be employed to purify them—as for silicon, whose use in electronic devices is made possible only by zone refining.

The semiconductors, silicon and germanium, present a special case of purification and subsequent alloying. The enthalpy of solution of elements in these semiconductors may be very large; under these circumstances the solubility line above the eutectic may not show the usual linear behavior as is seen in Fig. 1 for Pb in Sn. Rather the line may loop around in a behavior called *retrograde* (moving backward); see an example for Sb in Ge, Fig. 5. Such retrograde behavior for a wide range of alloying elements in Ge is shown in Fig. 17 of the article GERMANIUM. Such lines for all impurities are significant in the purification of Si and Ge by zone refining.

Measurement of the components of the Gibbs free energy of multicomponent systems is an immense task, but it continues methodically. Also difficult is extrapolation of fragmentary values to other temperatures and determining the "best" set of values when data conflict. Finding minima in multicomponent alloys is not always easy, and making sure that these are not metastable minima is difficult. Many calculational and smoothing computer programs are now in wide use. For binary and ternary systems, programs as simple as macro-driven spreadsheet files may be appropriate; for more complex systems, more sophisticated programs are required. The journal *CALPHAD* and the annual CALPHAD conferences provide perhaps the best source of such information. Data bases of thermochemical values are being assembled, some on line, others on diskettes or CD-ROMS, still others in analytical form. Handbooks of tables of data, once common, are becoming passe. A consortium of organizations worldwide is engaged in

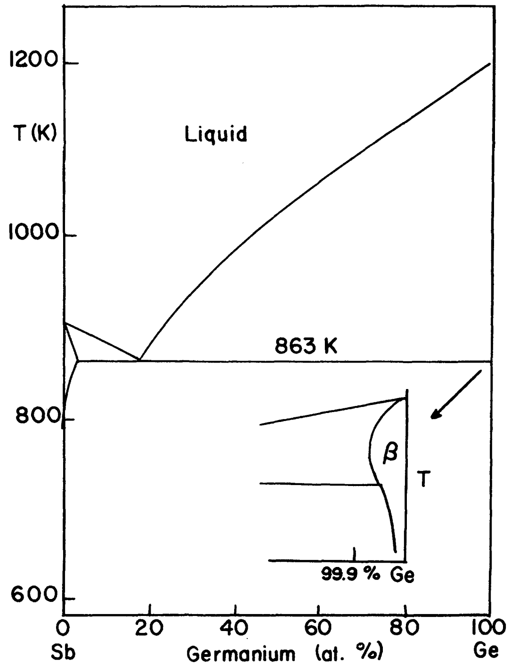

FIG. 5. The Ge–Sb eutectic. The solubility of Sb above the eutectic temperature shows retrograde behavior. The narrow region labeled β is the region of doping. After Wert and Thomson (1970).

collecting and publishing both phase diagrams and supporting thermochemical data on both binary and multicomponent systems (that consortium was assembled under the leadership of ASM International).

2. CONCENTRATED SOLUTIONS

Some alloys of technological importance display great solubility, among them the brasses. The reason for this large solubility was examined by the metallurgist Hume-Rothery (1948), not only for the alloys based on copper but also for those based on the companion metals gold and silver. He and his colleagues (using straightforward metallurgical techniques) assembled a huge amount of data on phase equilibria of alloys of these noble metals, analysis of which led them to propose that the extent of solubility depended on a number of rather simple "rules." The most significant are the following:

1. Large solubility is favored if the elements are not far apart in electronegativity.
2. It is also favored if the atom sizes are not too different (the difference in atomic radii should be less than 15%).
3. These two factors being favorable, the solubility (for these face-centered-cubic alloys) still does not exceed that value corresponding to an average electron to atom ratio (*e*/*a*) of about 1.4.

The results for copper alloyed with several multivalent elements are shown in Fig. 6. One sees that the limit of solubility occurs for a value of *e*/*a* in the range 1.3 to 1.4. A two-phase mixture exists for larger values of *e*/*a*, but other phases (having different crystal structures) exist at higher *e*/*a* ratios: 1.5, 1.6, and 1.7.

Hume-Rothery and his colleagues assembled their many experimental observations in the 1920s and 1930s, just when the band theory of solids was being developed by physicists. Using ideas from that theory, the group proposed that the limitations placed on the *e*/*a* ratio were the result of the reduction in electronic density of states when the expanding Fermi surface impinged on spe-

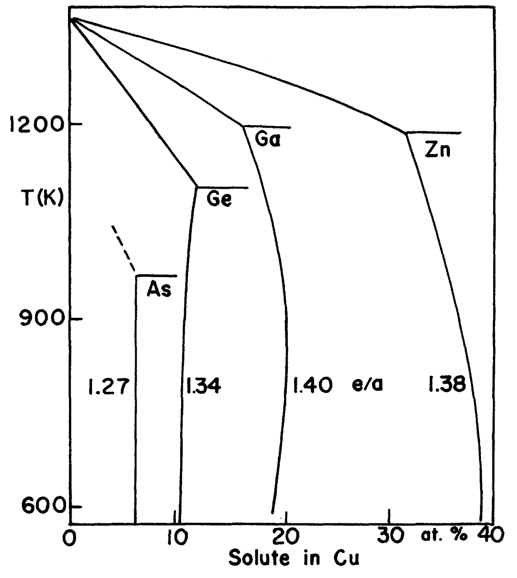

FIG. 6. Multivalent impurities in Cu have decreasing solubility as their valence difference with Cu becomes larger. But the *e*/*a* ratio at maximum solubility is about the same—between 1.3 and 1.4. After Barrett and Massalski (1966).

cific boundaries of the Brillouin zones as the concentration of multivalent elements was increased. For the face-centered-cubic copper alloys, this was assumed to be impingement of a spherical Fermi surface on octahedral faces of the first zone boundary. The critical *e/a* ratios for stable phases at higher concentrations result from impingement of the Fermi surface on boundaries of zones that become successively more nearly spherical.

The experimental evidence of the Hume-Rothery group is not to be doubted; an excellent summary is found in the book of Barrett and Massalski (1966). The simple interpretation initially proposed provoked much discussion among physicists, in particular of the "magic" *e/a* ratios. That the initial explanations were too simple is certain; experiments in the late 1950s showed that the Fermi surface already is in contact with the zone boundary for pure copper. Thus the limiting *e/a* ratio of 1.4 could not be explained by the initial ideas of Hume-Rothery and his colleagues. An alternative model based on more refined principles was proposed shortly thereafter in a series of articles by Blandin and colleagues at Orsay (see, for example, Blandin *et al.*, 1966). Other solutions were later offered by Heine and Weaire (1970) and by several authors in the more recent books *The Physics of New Materials* (Fujita, 1994) and *Electron Theory in Alloy Design* (Pettifor and Cottrell, 1992).

One can appreciate the difficulty of making accurate enough calculations by noting that the enthalpy difference associated with changes in phase is usually very small compared with the total enthalpy of either phase. Consider a really simple case—the enthalpy change for pure iron when it transforms at 1183 K from the low-temperature bcc phase to the high-temperature fcc phase is only about 6% of the enthalpy of the bcc phase just before the transformation. Even in this simple one-element case, one must calculate a relatively small difference between two large numbers. How much more difficult is the task for an alloy of two (or more) elements! These difficulties do not prevent physicists and materials scientists from attempting calculations of many types. Such attempts are described in many of the references given in the list of Further Reading.

Many combinations of elements other than the noble metals form technologically important solid solutions, of course—perhaps none more important than the stainless steels, combinations principally of Fe, Ni, and Cr. These elements have about the same atomic radii—about 0.125 nm. They are in the 3*d* series, so that their electronic similarities are favorable for large solubility. Iron (below 1183 K) and chromium (at all temperatures up to its melting point) have the same crystal structure (body-centered cubic), and so they are miscible in any ratio below 1183 K (except for the intrusion of a troublesome low-temperature σ phase at a 50:50 Fe–Cr ratio). Nickel has a different crystal structure (face-centered cubic), and so a change in crystal type must occur at some ratio for either Fe–Ni or Cr–Ni alloys. That limitation is not insurmountable, in technological practice.

Thousands of combinations of the three have been produced, of which the following are of primary importance. A ternary solid solution of 74% Fe, 18% Cr, and 8% Ni is of special value. It has good strength, yet can be rolled, forged, or pressed into complex shapes. It is resistant to corrosion and oxidation. Its commercial designation is type 304—but it is known in the trade as "18-8," and that symbol often appears inconspicuously on products such as flatware or serving dishes. Its corrosion resistance comes from a thin surface layer of the oxide CrO_2, which forms instantly on the surface—and which re-forms if the surface is abraded or scratched. Because of the high Cr content, 18-8 is not ferromagnetic. This alloy and other ternary combinations can be strengthened further by addition of a small amount of carbon, which forms a segregated carbide with appropriate heat treatment. Two binary higher-nickel alloys—55% Fe, 45% Ni and 22% Fe, 78% Ni—are both ferromagnetic with excellent permeability. Indeed, substitution of a little Mo in the latter alloy produces Supermalloy, an alloy with the highest permeability known—a rod becomes fully saturated, magnetically, when it is placed parallel to the Earth's field.

An enormous number of other alloys of Fe are also important. At one extreme are the "plain-carbon" steels, the cheapest steel, used for a myriad of products. At the other extreme are the complex multicomponent steels alloyed with small amounts of Ni, Cr, Mo, W, Nb, . . . along with carbon. After a hard landing of an aircraft on a runway, I always give a private word of thanks for "4340," a multicomponent high-strength steel with good ductility, used for such crucial applications as aircraft landing gear. Compilations of compositions, properties, and applications are found in many handbooks and commercial catalogs.

One interesting group of stainless steels are the iron meteorites. The iron always is alloyed with nickel—present in the range 7% to more than 15%. This is the concentration range for which phase separation occurs into an Fe-rich bcc phase (called kamacite) and a higher-Ni fcc phase (called taenite). Because of the billions of years of their existence, the iron–nickel alloy of most iron-based meteorites has separated into bands of kamacite and taenite of massive size easily visible with the unaided eye. The pattern is known as the "Widmannstätten structure" after the Viennese scientist who first identified it in the early 1800s. Many pictures of such bands are to be found in the handbook authored by Buchwald (1975)—that work is both an excellent compilation of the astronomical nature of such meteorites and the metallurgy of their constituents. Past technological use of "iron" of meteoritic origin has been discussed by Aitchison (1960) and by Smith (1960, 1962, 1968, 1981). The reader may find amusing the historical uncertainty that has arisen over "first rights of discovery" of the Widmannstätten pattern, for another observer, Thomsen, may have observed it somewhat earlier (see Smith, 1962). The varied paths of melting, freezing, segregation, and redistribution of minerals in the 50 or so parent bodies of meteorites are well described by Dodd (1986).

3. GAS–METAL EQUILIBRIUM

Equilibrium between an atmosphere of a gas and atoms of that gas dissolved in a metal has been studied extensively. Such characterization is important because of the gradual dissolution of compressed gases in the walls of metal containment cylinders and because of reactions of atmospheric gases with hot metals and alloys. An excellent review of the thermodynamics and kinetics of such interactions and a huge compilation of data are found in the book *Gase und Kohlenstoff in Metallen* written by the group at the Max-Planck-Institute in Stuttgart under the leadership of E. Gebhardt (Fromm and Gebhardt, 1976). They also carried on much work themselves, especially for the diatomic gases dissolved in the bcc metals. One example is given here from the work of Cost (1963) to show the principles involved, the pressure-temperature-composition equilibrium in the Nb–N system.

First, the solution of a diatomic gas dissolved atomically in a metal should obey Sievert's law—the dissolved concentration should vary as the square root of the gas pressure. Measurements for such equilibrium, Fig. 7, show that law to be obeyed; the lines through the data points are drawn with a slope of 2.

Next, the partial molar heat of solution of nitrogen in Nb [for dilute solutions, this is ΔH of Eq. (1)] should be a constant with both temperature and composition. That this

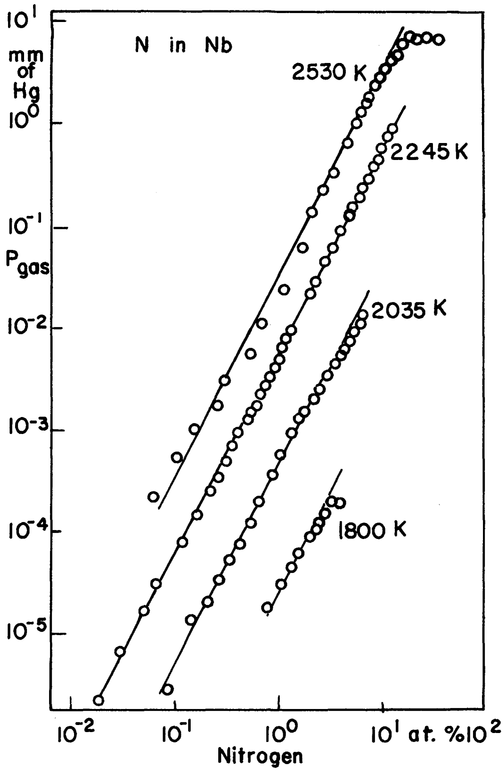

FIG. 7. The concentration of N atoms in Nb as a function of the gas pressure of nitrogen at constant temperature. After Cost and Wert (1963).

law is obeyed can be shown from the data plotted in Fig. 8. The straight lines on the plot of log(p) vs $1/T$ yield a value for ΔH of about -190 kJ/g-atom of nitrogen. The limiting line of Fig. 8 (delineated by the filled circles) yields a value of about -280 kJ/g-atom of nitrogen for the standard heat of formation of the equilibrium solid compound Nb_2N. From these data, the standard molar entropy of solution of nitrogen in Nb may be calculated to be -62 J/K·g-atom of nitrogen, a number just about that expected from a random distribution, with little excess entropy, i.e., change in vibrational character of the metal crystal. Finally, the limiting terminal solubility of N in Nb with respect to the nitride Nb_2N is displayed in Fig. 9; the standard heat of solution of the nitride in Nb at high temperatures is about 88 kJ/g-atom of nitrogen.

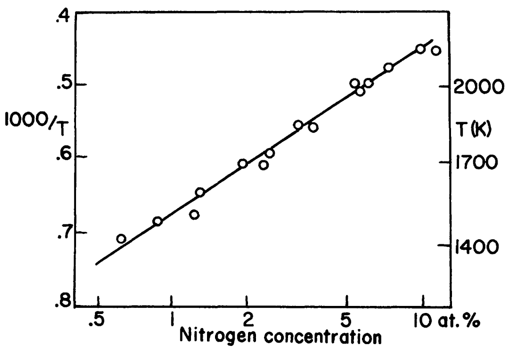

FIG. 9. The concentration of N atoms in solid solution in equilibrium with the stable nitride, Nb_2N, as a function of temperature. After Cost and Wert (1963).

These measurements specify completely both the equilibrium between the gas N_2 and N atoms dissolved in the metal and the equilibrium between the nitride Nb_2N and N atoms dissolved in the metal. From the derived parameters, the pressure-temperature-composition phase diagram may be drawn. That construct is shown in Fig. 10 (in coordinates log p, $1/T$, and log c). The phase diagram is a eutectic with eutectic composition at point G. At these low pressures, the composition of the nitride is invariant with either pressure or temperature. The limits of the solubility of the terminal solid-solution phase, labeled α, are planes in these coordinates whose slopes and lines of intersection are given by

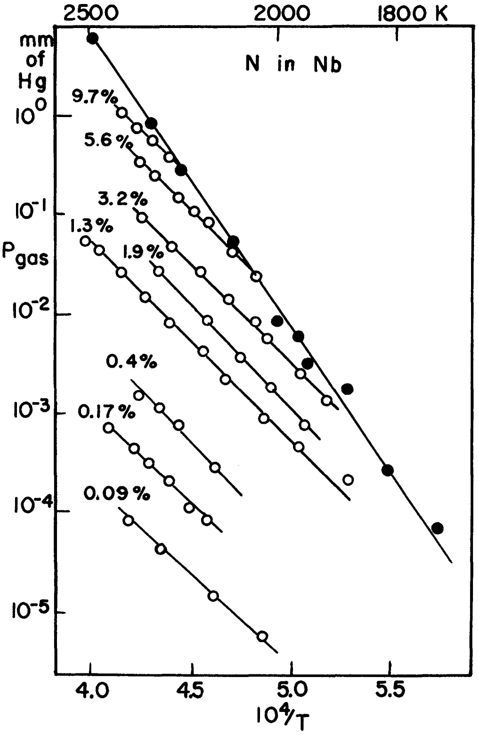

FIG. 8. The pressure-temperature equilibrium at constant concentration of N atoms in Nb. After Cost and Wert (1963).

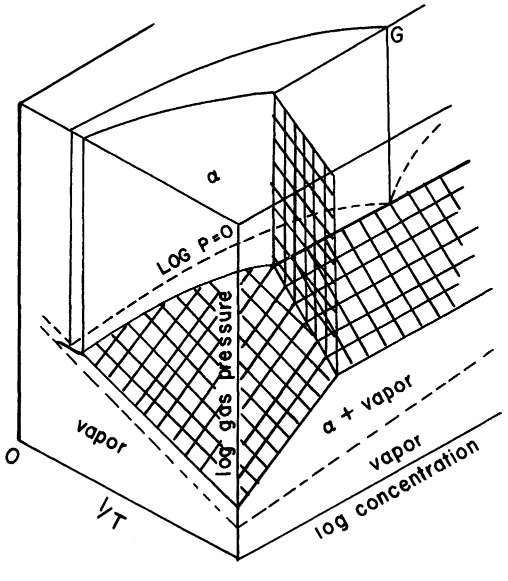

FIG. 10. A part of the phase diagram of N in Nb as a function of log pressure, $1/T$, and log concentration. After Cost and Wert (1963).

the values of the constants derived from the data embodied in Figs. 7–9. In addition, the concentration of dissolved nitrogen may be expressed in terms of temperature and gas pressure p (in pascals) as

$$c(\text{atom }\%) = 7 \times 10^{-3}\{p\}^{1/2} \exp(190\,000/RT).$$

From data in the book of Fromm and Gebhardt (1976), similar evaluations can be made for many gas–metal systems.

4. COMPOUNDS

Few alloy systems display large solubility. Most are dominated by compounds that exist as second phases. An example is shown in Fig. 11 for the system Cu–Zn. Although Zn is dissolved in the fcc Cu to about 40%, above that concentration several phases exist at higher compositions. Each of these phases exists over a range of several atom %; their compositions are not sharply set. These phases are called the Hume-Rothery phases; the pattern also exists for other binary alloys, such as Ag and Au with these alloying elements (see Chap. 11 of Barrett and Massalski, 1966). Their stability results from conduction-electron interactions—the phases themselves are still metallic conductors. An immense number of important alloy systems have their stability controlled by such phases.

The second set of important phases are the ionic crystals—formed from elements with large differences in electronegativity, such as the metal oxides. A typical example is the double eutectic system SiO_2–Al_2O_3 dominated by an oxide of Al and Si, mullite, Fig. 12. This oxide has a large energy of formation; the large value of ΔH overwhelms the entropy term of Eq. (1), and the solubility of mullite in both SiO_2 and Al_2O_3 in equilibrium with the oxide is small. All of these compounds, SiO_2, Al_2O_3, and mullite, are of large technological value. Many metallic and ceramic alloy systems are controlled by the high stability of such ionic compounds.

A large number of metallic and ceramic systems fall between these two extremes, especially for multicomponent combinations of elements. The Superalloy metallic phases based on Ni have been exquisitely combined to give a mixture of solid solutions and second phases to provide properties that make possible the turbine blades in jet engines. Iron alloys range from the solid-solution stainless steels to the cast irons (with their segregated graphite) to multiphase steels used for aircraft landing gear. The ionic oxides of copper combined with other metallic additions are under intense scrutiny for high-temperature superconductivity (see SUPERCONDUCTIVITY, HIGH-TEMPERATURE).

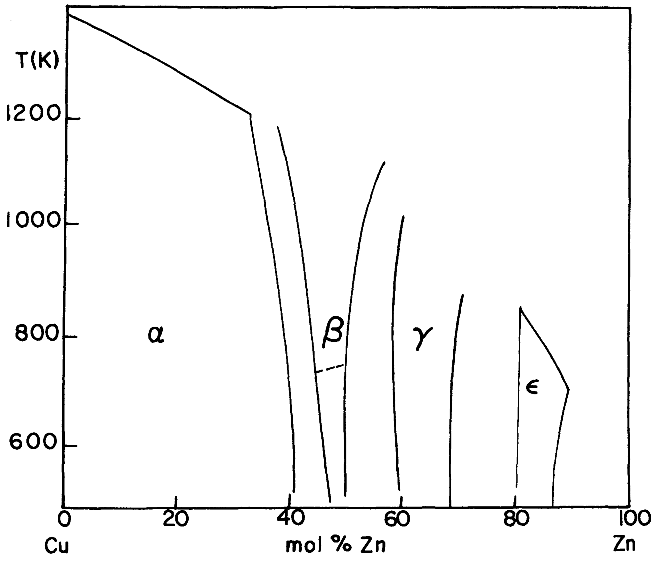

FIG. 11. The Hume-Rothery phases in Cu–Zn.

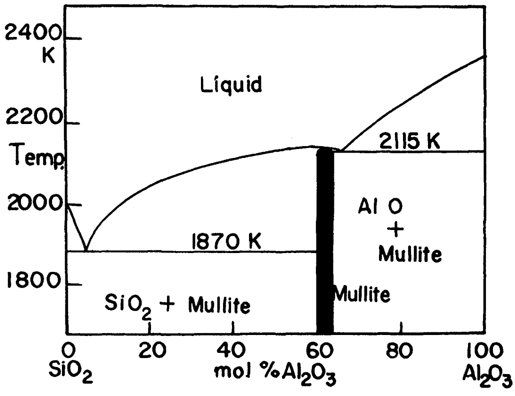

FIG. 12. Phase diagram dominated by an intermediate compound with large stability.

5. ORDERING

The thermodynamics of solid solutions described above is predicated on randomness of atom placement in the crystal. For many alloys, randomness is nearly achieved, but slight departures occur in most instances. One such departure is short-range order, in which atoms of one type have a preference of atoms of the second type for nearest neighbors. Two types of measurements of physical origin are commonly used to investigate the changes in this ordering—changes in electrical resistivity accompanying changes in temperature and x-ray scattering at small angles.

In some instances, the changes may be of long range with predictable atom placement over large distances. An example occurs for the phase 50% Cu–50% Zn; see the location of this phase in Fig. 11. The phase, called β brass, has the bcc crystal structure above 540 °C with nearly random placement of Cu and Zn in the crystal. This randomness gradually disappears with decreasing temperature, and below 540 °C (the short-dashed line in Fig. 11) both the Cu and Zn have only atoms of the other as nearest neighbors.

6. PRECIPITATION OF SECOND PHASES

Precipitation of second phases may occur when a temperature change makes a solid solution unstable. Several distinct reactions may occur. Chief among them is a process called *nucleation and growth.* Nuclei of the second phase (initially unstable) form within the solution by thermal fluctuations. These nuclei then grow by diffusion to form massive quantities of the second phase. An example of such a process may be described for nitrides forming in bcc iron (a choice based not so much on the technological importance of this alloy system as for the opportunity it affords to see how increasingly sophisticated techniques from physics aid in understanding the precipitation process). A thermodynamically stable solid solution of nitrogen in iron may be prepared at some high temperature (900 K is appropriate). Upon rapid cooling to some lower temperature (500 K is appropriate), a stable phase, Fe_4N, nucleates and grows as plates on the (210) planes of the iron. This phase is in equilibrium with a solid solution of lower concentration.

But this is not all. In this instance another nitride exists, $Fe_{16}N_2$, whose free energy in bcc iron is higher than that of Fe_4N, but which forms first at temperatures around 500 K, perhaps because of easier nucleation. The compound forms first as platelets on the (100) planes of the bcc iron and reaches a temporary equilibrium with a solid solution of N in the bcc iron, an equilibrium higher in N in the bcc iron than is true for N in equilibrium with the stable Fe_4N phase. The relationship of these phases and the thermodynamic parameters of this edge of the phase diagram may be seen in Fig. 13. The limiting solubility lines are straight lines in these coordinates, log c vs $1/T$ [in accordance with Eq. 2(b)]. The equations of the lines are put in as labels in the figure. The two sets of measurements, made by independent observations of Dijkstra (1949) and of Fast and Verrijp (1955), agree well. (Both measurements were made using mechanical relaxation techniques described in the last section.) The insets to this figure show optical micrographs of both the metastable phase $Fe_{16}N_2$ and the stable phase Fe_4N, which form in sequence. Observe that the solubility line in equilibrium with the metastable phase $Fe_{16}N_2$ occurs for all temperatures at a higher N concentration than that for the stable phase Fe_4N. This thermodynamic relationship always must exist in phase equilibria where metastable phases occur.

Optical micrographs such as those inset

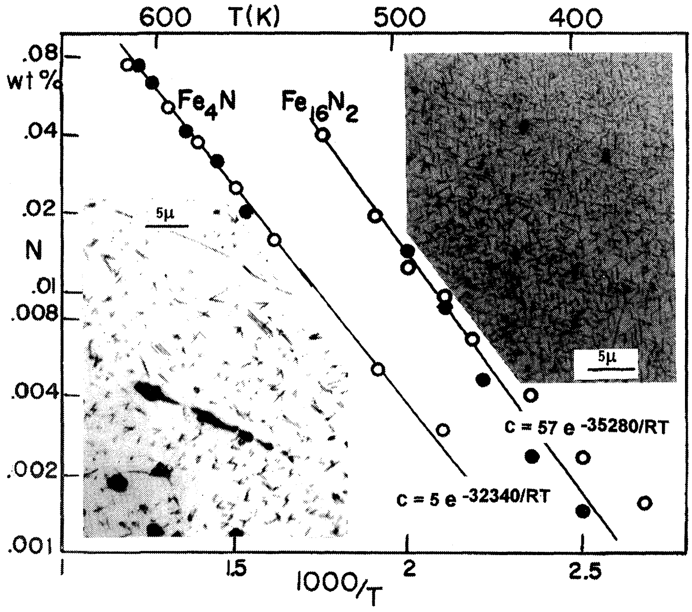

FIG. 13. A part of the phase diagram for Fe–N showing equilibrium for both the stable phase Fe_4N and the metastable phase $Fe_{16}N_2$. The open circles are taken from Dijkstra (1949), the filled circles from Fast and Verrijp (1955). The adjacent insets show precipitates: $Fe_{16}N_2$, upper right; Fe_4N, lower left. The large black blobs are iron oxide, initially present in the iron.

into Fig. 13 clearly show that $Fe_{16}N_2$ forms on the (100) habit plane of the host iron. Later calculations of elastic energy indicated that the preferred habit plane of such a tetragonal precipitate in a cubic host might not be precisely the (100) plane. See Wen *et al.* (1981). This possibility was found to be the case in this instance; electron microscope imaging by Ferguson *et al.* (1984) showed that the "plates" in the inset of Fig. 13 are, in fact, puckered rosettes of platelets, segments that are, indeed, tilted from the (100) plane by about 10°. The average habit plane, though, is (100).

No unusual magnetic characteristic of the compound $Fe_{16}N_2$ seemed evident when it was first observed as a metastable nitride in iron. However, crystals of this phase produce an unusual magnetic effect when grown epitaxially on the surface of an In–Ga–As wafer, whose chemistry was selected to produce a lattice constant close to that of $Fe_{16}N_2$. It possesses a "giant" magnetic moment per iron atom—more than 3 Bohr magnetons, a value much higher than should be expected.

Almost all commercially valuable alloys are heat treated to produce second phases to enhance some desirable combination of properties: strength, ductility, corrosion resistance, magnetic or electrical characteristics, ... In nearly all instances the appropriate final product is not a mixture of thermodynamically stable phases but a combination of metastable phases that experience shows give the best product. The steels, aluminum alloys, brasses, permanent magnets, bronze, semiconductor chips, . . . , all are metastable combinations. Processing is carefully formulated to enhance the desirable features; many failures of products occur because of improperly designed processes or because of errors in carrying out specified processes. In addition, the necessity of making appropriate metastable combinations is a major reason that first-principles calculations of physics are of small importance in formulating multicomponent alloys with useful properties. Again, though, the experimental techniques of physics—especially the electron microscope, together with spectroscopic and diffraction techniques—are of inestimable value in characterization of useful combinations.

Another mode of phase separation exists when the concentration of the alloy lies within the *spinodal* lines of the free energy plot. Understanding this reaction may be aided by the plots of Fig. 14. This is an extension of the plot of ΔG vs concentration (of Fig. 2) over the entire range. At high temperatures, the free-energy curve may be negative and monotonic in concentration; see

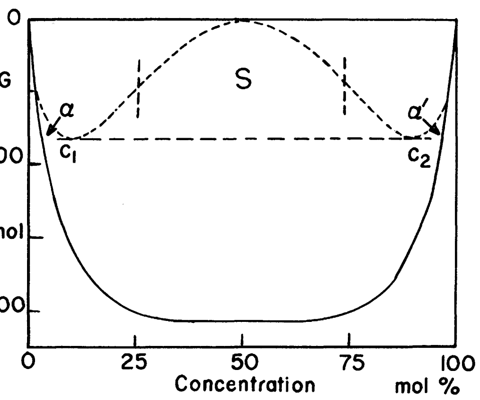

FIG. 14. Free-energy plot for an alloy system showing the influence of the value of ΔH on the extent of solubility. The solid curve shows the case where ΔH is small enough that ΔG varies monotonically with concentration over the entire range. With ΔH large enough, the curve takes the form drawn with the short-dashed line. At either extreme, regions of solid solubility exist. Between the points of common tangency, segregated phases of composition C_1 and C_2 coexist. However, within a central region, labeled S, segregated phases can form with no requirement for nucleation.

the solid line. A complete solid solution exists for this condition. At lower temperatures, where the entropy term is reduced, a curve like the short-dashed line may exist. For alloy compositions in the regions marked α and α', solid solutions exist. For alloys with intermediate composition, two equilibrium phases exist whose compositions C_1 and C_2 are defined by the common tangent line. But between these concentrations, another region S exists, within which $d^2\Delta G/dc^2$ is negative. For alloys with composition within this region, the solution is unstable with respect to small compositional fluctuations within itself. Such fluctuations are commonly bands that grow in spacing and in concentration differences as time goes on. Ultimately, the equilibrium phases C_1 and C_2 should result. This mode of phase separation, called *spinodal decomposition,* is attractive because no nucleation barrier exists.

An excellent example of this effect is decomposition of an oxide of Co and Fe, $(Co,Fe)_3O_4$, which had been deposited on a Si(111) surface. After being rapidly cooled from a high temperature where it exists as a solid solution (Hirano *et al.,* 1993), two phases formed spinodally upon aging at 975 K, one phase rich in Co and the other rich in Fe. The progress of the band formation with time is clearly shown in Fig. 15 by the splitting of the [310] diffraction line of Cu *Kα* into two adjacent peaks. A photograph by the authors shows that the alternate bands of Co- and Fe-rich phases are some 5 to 10 nm broad after some 45 minutes of aging. A great many similar examples of this continuous precipitation have been reported in which such continuous banding occurs: solid metal alloys, solid and liquid ceramic combinations (see GLASSES), and liquid polymeric mixtures.

A striking example of this type of decomposition are bands that have been seen in the olivine-rich achondrite Divnoe (Petaev and Brearly, 1994). Using electron-microprobe microanalysis, the authors observed alternate bands of iron-poor and iron-rich phases on a scale of 1 or 2 μm. They propose that the bands in the olivine phase of this meteorite formed at relatively low temperatures (around 500 K) during many millions of years.

The spinodal concept goes back to Gibbs (1928), but many others have contributed to

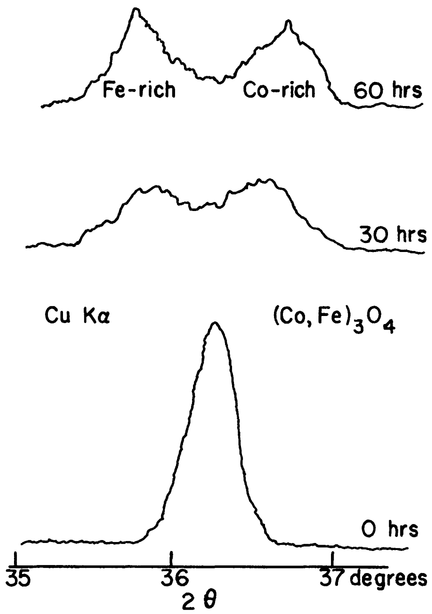

FIG. 15. Formation of the spinodal phases in decomposition of a mixed oxide of Co and Fe. Successive x-ray patterns show formation of alternating Co-rich and Fe-rich bands. After Hirano *et al.* (1993).

its development. Quantitative details of the initial stages of this process have been described more recently by Cahn and Hilliard (1958) in terms of a nonlinear diffusion equation; later stages are not well described. Whether the many reported examples of such phase separation are all spinodal phase changes fitting the Cahn–Hilliard formulation (or one of the many modifications proposed) is not clear, since competing nucleation-and-growth modes could also occur. Both experimental and theoretical investigations provide a fertile area for both materials science and applied physics research.

The strengthening of aluminum alloys containing several percent of such elements as Cu, Li, and Zn displays a sequence of several such metastable phases before the stable phase $CuAl_2$ forms. The commercially important alloys are treated to produce a banded structure resembling the spinodally produced phases. These first phases are thin plates, only a few atoms thick, which are enriched in the alloying constituent. They strengthen the alloy by greatly impeding the motion of dislocations. With time, other metastable phases form, and finally the phase $CuAl_2$; at this point the alloy is much softer than that containing the earlier metastable phases. First found in a laboratory in the German city Düren, the phases have been given the name Guinier–Preston zones in honor of the two French and English scientists who first described the details of their structure. The "overaging" of the aluminum alloys, with consequent softening, requires use of other alloys for high-performance, supersonic aircraft.

The historical account of the development of hardened aluminum alloys is fascinating. In part it shows initial applications of a craftlike nature, with improvements that come later with basic understanding of the geometry of the metastable precipitated phases. The Wright brothers, for example, fabricated a light-weight, but strong, aluminum–copper alloy for the crankcase of their early aircraft engine in 1903. Recent examination by Gayle and Goodway (1994) of the microstructure of the microstituents from specimens cut from the original crankcase shows the presence of thin plates, Fig. 16. Some 8 years after the flight of their plane, the *American Flyer,* Wilm (1911) showed the hardening that takes place upon a certain heat treatment of such alloys. Gradually, the deduction was made that the hardening came from formation of a precipitate during this heat treatment, a precipitate postulated by some to be $CuAl_2$; see Merica *et al.* (1920). Some years later, Guinier and Preston independently showed, by x-ray analysis, that the precipitate was not $CuAl_2$, but consisted of thin segregated platelets of some metastable, Cu-rich preprecipitate phase, now called Guinier–Preston zones [see Guinier (1938); Preston (1938)]. Still later, electron-microscope examination gave direct visual evidence of the geometry of the segregated phases, such as shown in the example of Fig. 16.

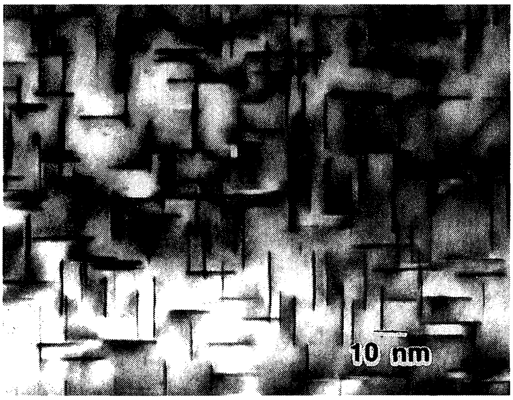

FIG. 16. Thin plates of Cu–Al zones in the aluminum alloy from the crankcase of the aircraft engine from the first Wright brothers' aircraft. Reprinted with permission from Gayle and Goodway (1994). Copyright 1994 American Association for the Advancement of Science.

Other modes of formation of segregated phases may occur. One, of importance, is decomposition of a single supersaturated phase along a reaction front to produce two new phases. The most famous example is the decomposition of austenite (a face-centered-cubic solid solution of carbon in iron) into two new phases arranged in a lamellar pattern—the new phases being a nearly pure iron (a body-centered-cubic phase) and the compound Fe_3C. The product phases are called *pearlite;* a typical example is shown in Fig. 17. This steel is hard, but not brittle—hence it is useful for such applications as a hammer, which is hard enough not to be dented by a nail, yet not so brittle that chips of steel would break off during the impact.

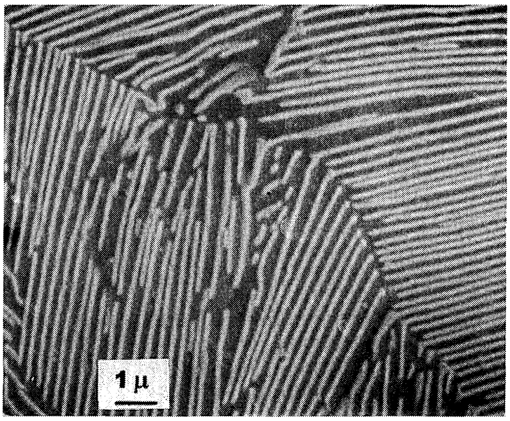

FIG. 17. Micrograph of a typical pearlite in a eutectoid steel. This layered structure shows of alternate bands of ferrite (nearly pure iron) and the carbide Fe_3C. Courtesy of R. Bohl.

7. RAPIDLY SOLIDIFIED ALLOYS

The metals used by humans for at least 8000 years were always aggregations of crystals. Although inorganic glasses were also used during much of that time, no way had been found to make their metallic counterpart; solidification of molten alloys always resulted in a crystalline product. Then in 1958, Prof. Pol Duwez and his colleagues in California produced, by rapid quenching, samples of a series of *amorphous* solid *alloys* in the eutectic system Cu–Ag (Duwez, 1960; Klement *et al.* 1960). Soon (as always happens) anyone could make such alloys—hundreds of alloy systems have been found to yield amorphous material with fast enough quenching. The first samples were measured in grams, but now many thousands of tons/year of amorphous magnetic materials alone are being sold.

Amorphous film production by rapid quenching has been supplemented by other methods. Milling of fine powders can produce amorphous material. Fragmentation of the structure of certain crystalline alloys by hydrogen absorption also may produce an amorphous product. Amorphous surface layers may be produced by laser heating or by electroplating. Irradiation may induce amorphization. Also interdiffusion of ultrathin layers of certain combinations can produce amorphous materials. See Greer (1995).

Rapidly quenched alloys are generally solid solutions; rapid quenching inhibits phase separation and compound formation. Their composition, though, may be well outside the solid-solution range of slowly cooled alloys. The full array of diffraction techniques using x rays, electrons, and neutrons has been employed to determine atom configurations. They show that the amorphous state is not a unique state; processing variables and alloy differences make the product somewhat variable. However, a geometrical short-range order of atom positioning is observed that is highly similar to that in the liquid state. A chemical short-range order is often observed—traces of the chemical affinity characteristic of the crystalline state may remain.

Magnetically soft iron-based amorphous ribbon has remarkably low core loss, about $\frac{1}{3}$ that of grain-oriented silicon iron. It is therefore attractive for power transformers from an engineering perspective. Its higher material cost offsets part of that economic gain, but some 10 000 transformer units are being installed in the United States each year, many additional units in Japan.

Not all rapid quenching of a liquid alloy produces an amorphous product—crystalline phases sometimes occur or can be formed by devitrification. One such phase is the magnetically hard compound $Nd_2Fe_{14}B$. Rapid quenching from the liquid state prevents formation of other equilibrium phases, which would form for that composition upon slow cooling. This rapidly cooled compound (and other rare-earth combinations) has a magnetic energy product in excess of 200 kJ/m^3, a value far exceeding that of any other material. In just a few years it is rapidly replacing the manufacture of Alnico and $SmCo_5$ magnets, the previously common high-energy-product materials. Magnets may be produced from the rapidly solidified ribbons by stacking them and pressing or by use of a binder material. Sintering of appropriately prepared powders may also provide a suitable fabrication technique for this material.

Yet other novel materials—the quasicrystals—may be produced by rapid solidification of particular alloys. These alloys have very large atom ratios compared to the relatively small ratios for most compounds; a particular quasicrystal of aluminum, copper, and iron has the atom ratio $Al_{65}Cu_{23}Fe_{12}$. X-ray Bragg reflection patterns show sharp spots.

However, the patterns show icosahedral (fivefold) symmetry, so that the usual crystallographic space-filling mode by linear translation of the unit cell does not exist. Further, the quasicrystal "cell" is so large—in the range 500–2500 nm—that it is indeterminate, and detailed placement of specific atoms cannot be made. Although the diffraction spots are sharp, careful examination shows a slight irregularity that is related to small improper atom placement, a defect called a phason.

Many quasicrystals are rapidly solidified (ternary, or higher) aluminum alloys, for which the rapid cooling extends the normal solubility range. The materials are generally poor conductors, a condition assumed to result from an extremely low density of states at the Fermi level. The *e*/*a* ratio is about 2.2 electrons per atom, a number generally considered to be the location of a "Hume-Rothery pseudogap", a designation somewhat of a turnaround from the skepticism of the earlier Hume-Rothery *e*/*a* concept for crystalline alloys.

Quasicrystals have been observed as a stable feature of a few phase diagrams at high temperature. See, for example, a cut through the Al–Fe–Cu ternary at 680 °C in Fig. 18 (Bradley and Goldschmidt, 1939). Other such phases have apparently been observed in the Al–Cr and Al–Mn binary systems. The reader may see an excellent review of the subject in a paper by Kelton (1993). No practical application has been found for these materials; they are an interesting laboratory curiosity. Undoubtedly, though, use will be found for them (see CRYSTALLINE STATE; QUASICRYSTALS).

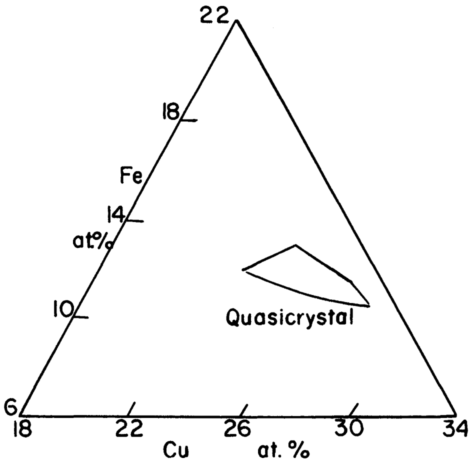

FIG. 18. A cut at 950 K through the ternary AlFeCu phase diagram showing the range of compositions of the quasicrystal phase. For simplicity, the equilibrium relationships with the other crystalline binary and ternary phases are not shown.

8. MECHANICAL AND DIELECTRIC SPECTROSCOPY

Previous paragraphs have emphasized the important role played by the spectroscopic techniques in specification of properties of solutions and segregated phases. In this final section, examples of two spectroscopies, mechanical and dielectric loss, are made to demonstrate their importance for specifying atomic configurations and kinetics of rate processes for metals, ceramics, and polymers. The principles can be best introduced using a formalism defined by Zener in which Hooke's law is modified to take account of both the usual interatomic force constants and a contribution from additional time-dependent relaxations that may produce additional strain. In the simplest case the relationship between stress and strain is modified by addition of only their first time derivatives:

$$\sigma + \tau\dot{\sigma} = M_R\epsilon + M_U\tau\dot{\epsilon}. \qquad (3)$$

Here, the parameter M_R is the usual static elastic constant, and M_U is the elastic constant at infinite frequency. The time constant τ is called the relaxation time; it is that time required for the relaxation to achieve $1/e$ of its completion value. Strictly, three relaxation times have been defined: one at constant stress, τ_s, one at constant strain, τ_e, and a third, τ, defined as the square root of the product of these two, $\tau = (\tau_s\tau_e)^{1/2}$. For most cases, there is little numerical difference between the three, so I use, for these examples, the designation τ.

This expression describes cases for constant applied stress, constant strain, or a time-dependent stress and strain. An enormous number of solids have been shown to possess recoverable atomic rearrangements described by this expression (see Zener, 1948; Nowick and Berry, 1972; McCrum *et* *al.*, 1967). The usual case is that for which a sinusoidal stress is applied at some angular frequency ω; the strain, of course, possesses the same frequency ω, but it is not in phase with the stress. The elastic modulus has a real and an imaginary part, therefore. These components may be expressed in the following way:

$$M_{\text{real}} = M_U - (M_U - M_R)/(1 + \omega^2\tau^2), \qquad (4a)$$

$$M_{\text{imag}} = (M_U - M_R)\omega\tau/(1 + \omega^2\tau^2). \qquad (4b)$$

These expressions are called the Debye equations, since they were first derived by Debye (1928, 1945) for the dielectric case, described later. The first equation describes the change in the real part of the elastic constant as the frequency decreases. The second gives the energy loss as a function of frequency, a loss termed the internal friction. In principle, either of these equations can be used for interpretation of measurements; but the second is customarily used. In succeeding paragraphs, examples are used to illustrate the applications to solutions and compounds. These examples depend on changes of atom positions upon application of a stress, i.e., on diffusion. Since thermally activated motion varies with temperature according to an Arrhenius expression $e^{-H/RT}$, τ is assumed to be given by an equation $\tau = \tau_0 e^{H/RT}$, where τ_0 is the time constant at infinite temperature. Then, one has the choice of making measurements as a function of temperature at constant angular frequency ω, or as a function of frequency at constant temperature. For experimental simplicity, the first is commonly used. For simplicity of nomenclature, Eq. (4b) is usually rewritten in terms of the phase angle δ between stress and strain as

$$\tan\delta = A\omega\tau_0 e^{H/RT}/(1 + \omega^2\tau_0^2 e^{2H/RT}). \qquad (5)$$

All spectroscopic techniques have two significant parameters; this type is no exception. The first is a *rate constant,* here the material parameter τ. The second is the *amplitude factor A*; it describes the strength of the relaxation. Although both τ and A have been used extensively to characterize solutions and compounds, most of the examples described here refer to the first, the rate constant.

The first example is for one of the oldest metallic solutions used by mankind, a Cu–Zn brass. Zener showed that a plot of tanδ vs T has the form shown in Fig. 19 for a measuring frequency of 620 Hz. Zener proposed that this effect is caused by reorientation under stress of Zn–Zn dipoles in the brass, the rate being determined by local diffusion of the zinc atoms. At low temperatures, the rate of jumping is slow, i.e., τ is large, so that $\omega\tau$ in Eq. (4b) in both the numerator and denominator is large compared with 1, and the energy loss given by Eq. (4b) is low. At high temperature, the value of τ is small, so that $\omega\tau$ is small, and the energy loss is again small. At that temperature for which $\omega\tau$ is about unity, the energy loss is at a maximum.

This measurement has been repeated for many such substitutional alloys (see Nowick and Berry, 1972), and the following features have been observed:

1. The effect is indeed a reorientation of strain dipoles under stress. The strength of the effect, i.e., the magnitude of tan δ, for dilute alloys is proportional to the square of the atomic concentration, as would be expected for random placement of alloying atoms among the crystal sites.

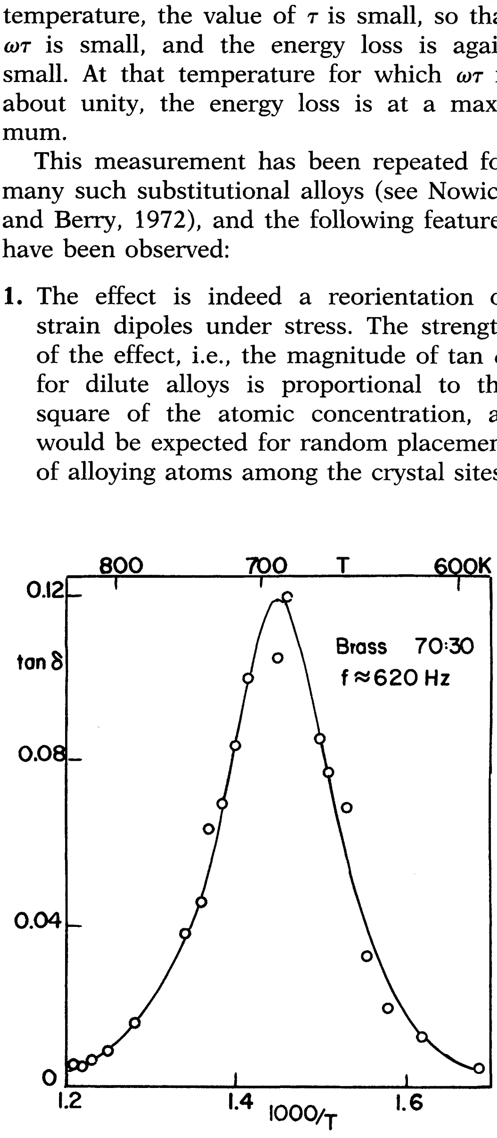

FIG. 19. Zener's first measurement of internal friction caused by spatial reordering under a mechanical stress of Zn–Zn dipoles in a 30% brass. After Zener (1948).

2. Measurements on single crystals give the proper orientation effects for nearest-neighbor pairs.
3. The activation enthalpy [H in Eq. (5)] is closely the same as that measured by other means for diffusion of the alloying elements.

A second example, first examined by the physicist Snoek and his colleagues (Snoek, 1941) and later developed in detail by Zener and his colleagues, is for interstitial alloys. It is typified by the solutions described earlier: nitrogen in bcc iron and nitrogen in niobium (also bcc); but it also is characteristic of other small atoms, carbon, oxygen, and hydrogen, dissolved in the bcc metals. These elements do not replace atoms of the host metal upon alloying; rather, they fit into interstices in the host crystal. If the host is fcc, the dissolved atoms fit into sites also with cubic symmetry, and the local distortion is merely a volumetric change. But if the host crystal is bcc, the distortion has a tetragonal component, and strain dipoles thus exist that can reorient under stress with a rate constant characteristic of local diffusion.

One of the most complete sets of measurements of diffusion in such is for carbon atoms in bcc iron. A set of measurements at several frequencies near 1 Hz is shown in Fig. 20 (Wert and Zener, 1949). Gradually, as experimental techniques developed to allow measurement over a much broader range of frequency, determination of the diffusion coefficient was extended over a range of some 10^{15} [see Fig. 21, adapted from an article of Lord and Beshers (1966)]. The wide range of the measured values allows the activation energy H to be determined with high reliability—the value is 81 470 J/mol. Extrapolation of the data to high temperatures shows a nonlinearity in this semilogarithmic plot, implying some additional enhancement of the diffusion rate at high temperatures, a phenomenon not yet explained.

Interstitials in the bcc metals reorient under the application of an applied stress because the tetragonal symmetry of such an interstitial impurity is lower than the cubic symmetry of the cubic host—the interstitial is thus a strain dipole. Isolated interstitials in fcc hosts (such as the stainless steels) do not produce a mechanical energy loss because they fit into interstices also of cubic symmetry, and thus they produce no strain dipole. But a strain dipole could exist for close H–H pairs or close clusters of H with a substitutional alloying atom, such as Ni or Cr. In such a case, an internal friction peak would be observed if the dipole strength were large enough and the clusters were present in suf-

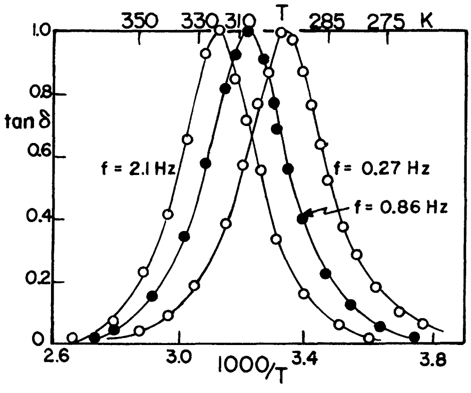

FIG. 20. Measurements of the internal friction Snoek peak of carbon in iron at several frequencies of measurement. The peak shifts in temperature as the frequency of the applied stress is changed. After Wert and Zener (1949).

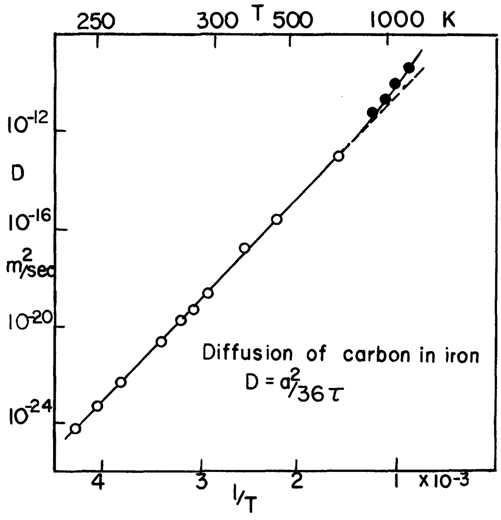

FIG. 21. Plot of diffusion coefficients for carbon in iron as a function of $1/T$. Only a minute fraction of the many reported values are shown. This line represents measurements of a physical parameter over as wide a range of values as for nearly any physical parameter ever measured. After Lord and Beshers (1966).

ficiently large concentration. Many such measurements have been reported for hydrogen in stainless steels of many compositions. Loss peaks have been observed in the temperature range 200–300 K for frequencies from 200 to 2000 Hz. See a measurement for hydrogen in three Fe–Ni binary alloys plotted in Fig. 22 (Nishino *et al.*, 1987); a damping peak is seen at 240 K for a frequency of 700 Hz. These authors supposed that the loss mechanism is indeed the reorientation of H–H pairs in the alternating stress field. Such an interpretation would require that the magnitude of the peak, i.e., the constant A in Eq. (5), should increase as the square of the dissolved H concentration. A later study by Gavriljuk *et al.* (1993) showed that A increases linearly as the H concentration increases, not as the square, and so it appears that a Ni–H cluster is favored. Variation of the rate constant τ with temperature shows that the enthalpy of motion, about 36 kJ/mol, is about right for hydrogen-ion diffusion in the stainless steels.

The measurements described in the preceding paragraphs relate to local diffusion of atoms. However, energy loss may occur by relaxation effects associated with long-range diffusion of atoms for proper specimen geometry. Let a rod be flexed so that a linear stress gradient exists across the rod. Atoms in the compression region possess a higher potential energy than those in the dilated region and will diffuse to reduce their energy. If the stress is periodic with angular frequency ω, this diffusion may result in an alternating strain relaxation that is out of phase with the stress—energy loss occurs. The relaxation time has been shown by Gorsky (1935) to be given in terms of the diameter of the rod, d, and the diffusion coefficient, D, by the expression $\tau = d^2/\pi^2 D$. The energy loss is a maximum, as before, when $\tau = 1/\omega$. This technique has been especially useful for measurement of hydrogen and deuterium in the bcc metals; see the example in Fig. 23 for diffusion of these isotopes in Nb, Ta, and V. Perhaps no measurement of diffusion has ever been made more frequently than this one, since the measurement has been for many years part of a laboratory exercise in the physics program of the Technical University of Munich, in Garching.

The data plotted in Fig. 23 allow us to examine the isotope effect on diffusion. The classical ratio of the diffusion coefficients for hydrogen and deterium, D_H/D_D, at any temperature should be 1.4 (i.e., $2^{1/2}$). This is so since the vibrational frequency of an atom in a crystal should vary as the reciprocal of the square root of the mass of the atom, $1/m^{1/2}$. The data of the figure show a much larger ratio than that for all three metals V, Nb,

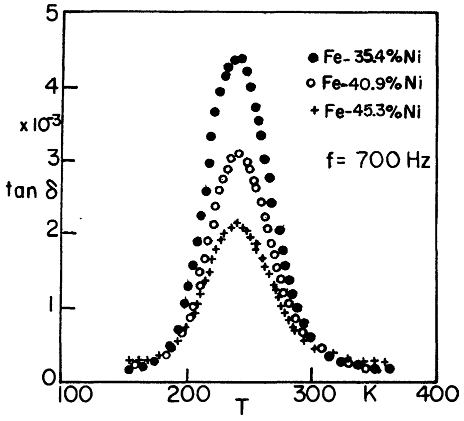

FIG. 22. Internal-friction peak for binary Fe–Ni alloys containing dissolved hydrogen. For these three alloys, the peak stays constant in temperature, i.e., τ is a constant, although the magnitude of the peak changes, i.e., the constant A changes. Nishino *et al.* propose that the peak is caused by relaxation of H–H clusters, but later investigations suggest that the relevant cluster is Ni–H. Curves after Nishino *et al.* (1987).

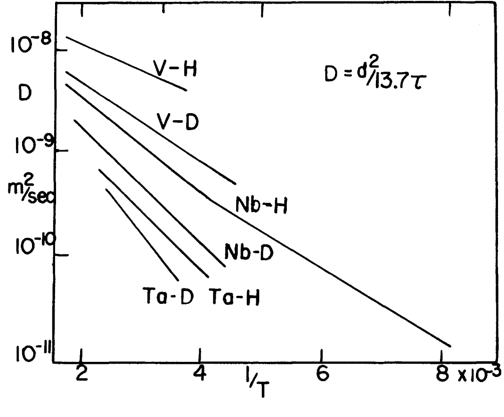

FIG. 23. Measurements of the diffusion coefficient for isotopes of Hydrogen in the bcc metals using the Gorsky effect. After Schaumann *et al.* (1970).

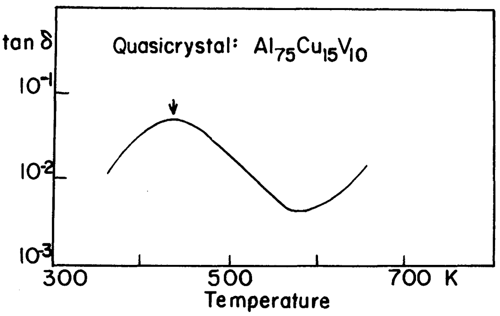

FIG. 24. Internal friction peak resulting from atom reorientation under a mechanical stress for a quasicrystal of Al, Cu, and Fe. After Okumura *et al.* (1994).

and Ta, and so the classical model is not correct in this instance.

Similar loss measurements exist for rapidly solidified alloys—both for amorphous alloys and for quasicrystals. One such measurement is shown in Fig. 24 for the quasicrystal FeAlCu (Okumura *et al.*, 1994). The origin of the energy dissipation is thought by the authors possibly to be caused by an atomic rearrangement associated with phasons.

The "plastic" nature of polymers provides ample opportunity for atomic rearrangements to produce internal friction loss. Energy loss occurs both for polymers formed from a single monomer and for polymer blends. In general, three distinct peaks are seen, as is seen in Fig. 25 for nylon. That at the lowest temperature is called the γ peak. It is caused by reorientation under stress of segments of the main chain. The next peak, the β peak, is caused by reorientation of side chains. That at the highest temperature, termed the α peak, is caused by a structural change at the glass transition temperature. The applications of internal friction to polymer structure occurred during the same time frame as was true for the metals, mainly after 1950. Even so, the two paths of investigation have remained almost separate; few investigators have worked in both. The atomic arrangements that produce these mechanical relaxations in polymers may also be characterized by neutron scattering (Frick and Richter, 1995).

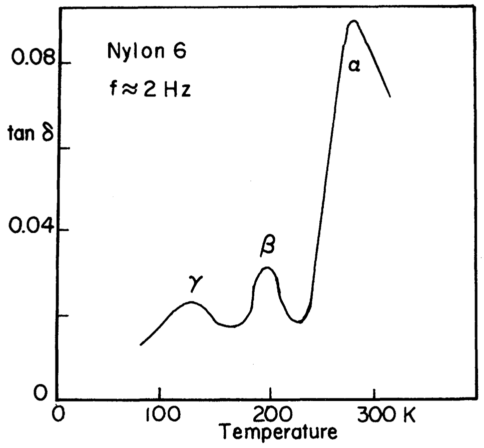

FIG. 25. Internal friction loss peaks for nylon-6.

Ceramic materials have a special attribute that has provided an additional technique of investigation. Many are good insulators, and they thus may show dielectric loss. That application depends on a set of principles first enunciated by Debye (1928, 1945). For a lossy material, the electric polarization **P** may not be in time phase with the field **E**; the controlling expression has the same functional relationship as for the mechanical case,

$$\mathbf{P} + \tau\dot{\mathbf{P}} = \epsilon_0(\chi_R\mathbf{E} + \tau\chi_U\dot{\mathbf{E}}). \qquad (6)$$

Here, ϵ_0 is the permittivity of a vacuum and χ_R and χ_U are the susceptibilities in the relaxed and unrelaxed states, respectively. A real and an imaginary part of the dielectric capacitivity exist, and the dielectric loss in an ac field **E** has the form

$$\tan\phi = A'^{*}\omega\tau_0 e^{H/RT}/(1 + \omega^2\tau_0^2 e^{2H/RT}).$$

Defects in insulators that have both a mechanical strain dipole and an electric moment may show both mechanical loss and dielectric loss. This joint behavior may allow additional features of the defect to be determined. An example of such a determination is shown in Fig. 26 for both $\tan\delta$ and $\tan\phi$ using tetragonal ZrO_2 containing 3 mol% Y_2O_3. A loss peak appears in the same temperature region for both (Weller, 1994), and so the authors conclude that the same defect is responsible for both the mechanical and the dielectric loss. They supposed that this defect is an associated yttrium-vacancy complex that reorients under the influence of

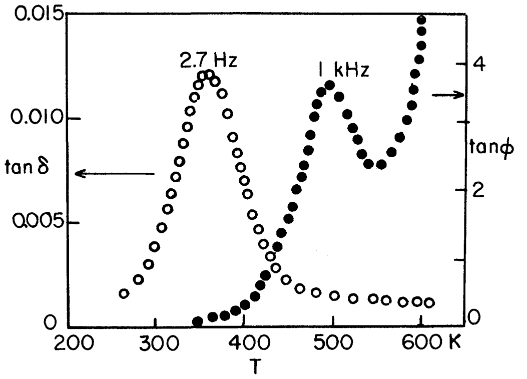

FIG. 26. Mechanical and dielectric loss spectra for ZrO_2 containing 3 mol% of Y_2O_3. The peak is caused by a spatial reorientation, under either an electric or a mechanical stress, of an yttrium-vacancy pair. It is displaced between the two measurements because of the frequency difference. This difference, in fact, allows the activation energy for the reorientation to be determined: 0.9 eV. After Weller (1994).

both a mechanical stress and an electric field. Measurements at several frequencies show that both have the same activation enthalpy, H, of about 90 kJ/mol. Weller and his colleagues have made many measurements on other ionic crystals, as have Wachtmann (1963) and Nowick and Heller (1963). A huge number of observations have been reported by a myriad of investigators for the many high-T_c superconductors.

9. CONCLUSIONS

Materials have been used by mankind for tools and weapons for hundreds of thousands of years. Modification of metals and ceramic materials by heat treatment and alloying began perhaps 200 centuries ago. The first metallic alloys were fabricated about 7000 B.C., gold and copper alloys in the beginning and later alloys of iron. The role of physics came, therefore, rather late—only in the last two centuries. Development of materials for mechanical applications has profited in the past century more by the application of physical measurement techniques than by theory; theory has been more after-the-fact explanatory rather than predictive. Development of materials for electrical applications—especially for modern electronics—though, has been aided immensely by both theory and measurement techniques of physics. Polymers, whose production in synthetic form has occurred only in the past century, have been characterized almost from the beginning by the physical and chemical sciences.

GLOSSARY

Alnico: A permanent-magnet alloy developed in the 1930s. It contained aluminum, nickel, and cobalt (and small quantities of other elements) as well as iron. It was a big improvement over the carbide-hardened steels previously used.

Bohr Magneton: The magnetic moment associated with the electron spin. Its value is 9.27×10^{-24} A-m^2.

Brillouin Zones: Boundaries in momentum space at which the energy–momentum relation suffers a discontinuity.

Eutectic: Point of lowest melting temperature as the composition of an alloy is changed.

Habit Plane: The plane of the host crystal along which a precipitate forms.

High-Strength Alloy Steels: Steels alloyed with small amounts of metallic elements (commonly totaling less than 5%). These improve the ability of the alloy to be hardened by influencing the nature of the carbide precipitate. A widely used alloy is "4340." Its composition is about the following: Ni: 1.8%; Cr: 0.8%; Mo: 0.25%; Mn: 0.7%; Si: 0.3%; C: 0.4%. When properly heat treated, its tensile strength at room temperature is about 250 000 psi (about 1750 MPa). It will deform in tension about 10% to fracture.

Kamacite: The low-Ni phase of the Fe–Ni constituent of the iron meteorites.

Phase: A physically homogeneous part of a material's system.

Phase Diagram: A graphical construct in which the phases present are plotted as a function of composition, temperature, and pressure.

Sigma (σ) Phase: An intermediate Fe–Cr phase with composition nearly 50:50. It forms at temperatures below 800 K. It is a brittle phase, usually not desired in alloy steels.

Spinode: Dictionary definition—cusp. In our context, it refers to a maximum in the slope of the line of G vs composition. At that point, $d\Delta G/dc$ goes through zero with increasing composition. Between the spinodal points $d^2\Delta G/dt^2$ is negative.

Taenite: The high-Ni phase of the Fe–Ni constituent of the iron meteorites.

Vacancy: A vacant lattice site in a crystal.

Widmannstätten Pattern: A pattern of precipitates of two phases of iron–nickel solid solutions that form in the iron meteorites over billions of years. The most striking pattern occurs when plates of an iron-rich phase (kamacite) precipitate on (111) planes of the remaining iron–nickel phase (taenite). The pertinent figure shows that pattern for one of the most famous iron meteorites (Gibeon) that fell in the early 1800s in Southwest Africa (now Namibia). The white bands are the kamacite—they have formed on planes of the taenite having octahedral symmetry. Meteorites that show this pattern are called, therefore, *octahedrites*.

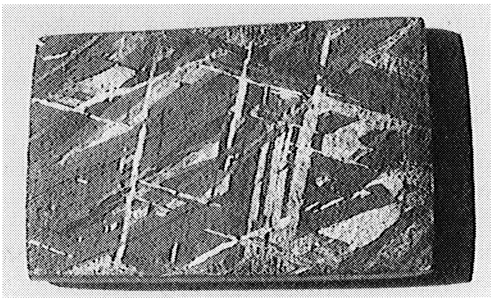

Works Cited

Aitchison, Leslie (1960), *A History of Metals*, New York: Interscience Publishers.

Barrett, Charles S., Massalski, T. B. (1966), *Structure of Metals*, 3rd ed., New York: McGraw-Hill.

Blandin, A., Friedel, J., Saada, G. (1966), *J. Phys. (Paris)* Colloque **27,** C3, 128.

Bradley, A. J., Goldschmidt, H. S. (1939), *J. Inst. Met.* **65,** 389–418.

Buchwald, Vagn (1975), *Handbook of Iron Meteorites*, Vols. 1, 2, and 3, Berkeley: University of California Press.

Cahn, J. W., Hilliard, J. E. (1958), *J. Chem. Phys.* **28,** 258–267.

Cost, J. R., Wert, C. A. (1963), *Acta Metall.* **11,** 231–242.

Debye, P. (1928, 1945), *Polar Molecules*, New York: Dover.

Dijkstra, L. (1949), *Trans. AIME* **185,** 252–260.

Dodd, Robert T. (1986), *Thunderstones and Shooting Stars, The Meaning of Meteorites*, Cambridge, MA: Harvard Univ. Press.

Duwez, Pol, Willens, R. H., Klement, W., Jr. (1960), *J. Appl. Phys.* **31,** 1136–1137.

Fast, J. D., Verrijp, M. B. (1955), *J. Iron Steel Inst.* **180,** 337–343.

Ferguson, P., Dahmen, U., Westmacott, K. H. (1984), *Scr. Metall.* **18,** 57–60.

Frick, B., Richter, D. (1995), *Science* **267,** 1939–1945.

Fromm, E., Gebhardt, E. (1976), *Gase und Kohlenstoffe im Metallen*, Berlin: Springer.

Fujita, F. (1994), *The Physics of New Materials*, Berlin: Springer.

Gavriljuk, V. G., Hänninen, Yu, Tarasenko, A. V., Tähtinen, S., Ullako, K. (1993), *Scr. Met. Mater.* **28,** 900–906.

Gayle, Frank W., Goodway, Martha (1994), *Science* **266,** 1015–1017.

Gibbs, J. Willard (1928), *The Collected Works of J. Willard Gibbs*, New York: Longsmans Green.

Gorsky, W. S. (1935), *Phys. Z. Sowjetunion* **8,** 457–465.

Greer, A. Lindsay (1995), *Science* **267,** 1947–1953.

Guinier, A. (1938), *C. R. Acad. Sci.* **206,** 1641–1643.

Heine, Volker, Weaire, D. (1970), in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, vol. 24, New York: Academic Press, 249–462.

Hirano, S., Yogo, T., Kikuta, K., Asai, E., Sugiyama, K., Yamamoto, H. (1993), *J. Am. Cer. Soc.* **76,** 1788–1792.

Hume-Rothery, W. (1948), *Atomic Theory for Students of Metallurgy*, London: Institute of Metals.

Kelton, K. F. (1993), *Int. Mater. Rev.* **38,** 105–137.

Klement, K., Willens, R. H., Duwez, P. (1960), *Nature* **187,** 869.

Lord, A. E., Beshers, D. N. (1966), *Acta Metall.* **14,** 1659–1672.

McCrum, N. G., Read, B. E., Williams, G. (1967), *Anelastic and Dielectric Effects in Polymeric Solids*, New York: Wiley.

Merica, P. P., Waltenberg, H. G., Freeman, J. R. (1920), *Trans. AIME* **64,** 3–25.

Nishino, Y., Kato, T., Tamaoka, S., Asano, S. (1987), *Scr. Metall.* **21,** 1235–1239.

Nowick, A., Berry, B. (1972), *Anelastic Relaxation in Crystalline Solids*, New York: Academic Press.

Nowick, A. S., Heller, W. R. (1963), *Adv. Phys.* **12,** 251–298.

Okumura, H., Tsai, A. P., Inoue, A., Masumoto, T. (1994), *Mat. Sci. Eng. A* **182,** 781–1792.

Petaev, Michail I., Brearley, Adrian J. (1994), *Science* **266,** 1545–1547.

Pettifor, D., Cottrell, A. H. (1992), *Electron Theory in Alloy Design,* London: Inst. of Metals.

Preston, G. D. (1938), *Proc. R. Soc.* **A167,** 526–538.

Schaumann, G., Völkl, J., Alefeld, G. (1970), *Phys. Stat. Sol.* **42,** 401–413.

Smith, C. S. (1960), *A History of Metallography,* Chicago: University of Chicago Press.

Smith, C. S. (1962), *Geochim. Cosmochim. Acta* **26,** 971–972.

Smith, C. S. (1968), *Sources for the History of the Science of Steel,* Jointly Published by the Society for the History of Technology and the MIT Press. Cambridge, MA and London, England.

Smith, C. S. (1981), *A Search for Structure,* Cambridge, MA: The MIT Press.

Snoek, J. L. (1941), *Physica* **8,** 711–733.

Wachtman, J. B., Jr. (1963), *Phys. Rev.* **131,** 517–527.

Weller, M. (1994), *Abschlussbericht DFG—Kolloquium Keramische Hochleitungswerkstoffe* (to be published by VCH, Weinheim).

Wen, S. H., Kostlan, E., Hong, M., Khachaturyan, A. G., Morris, J. W. (1981), *Acta Metall.* **29,** 1247–1254.

Wert, C. A., Thomson, R. M. (1970), *Physics of Solids,* New York: McGraw-Hill.

Wert, C. A., Zener, C. (1949), *Phys. Rev.* **76,** 1169–1175.

Wilm, A. (1911), *Metallurgie* **8,** 125–130.

Zener, C. (1948), *Elasticity and Anelasticity of Metals,* Chicago: University of Chicago Press.

Further Reading

Alper, Allen M. (Ed.) (1970–1978), *Phase Diagrams: Materials Science and Technology,* Vols. I–VII, San Diego: Academic Press. Excellent articles by competent writers.

Alper, Allen M. (Ed.) (1995), *Phase Diagrams in Advanced Ceramics,* San Diego: Academic Press.

Baker, H. (Ed.) (1992), *Handbook of Phase Diagrams,* ASM International, OH: Materials Park.

Bennett, L. H. (Ed.) (1980), *Theory of Alloy Formation,* Pittsburgh: TMS-AIME.

Bergeron, C. G., Risbud, S. H. (1984), *Introduction to Phase Equilibria in Ceramics,* Columbus, OH: Amer. Cer. Soc.

Blythe, A. R. (1979), *Electrical Properties of Polymers,* Cambridge, U.K.: Cambridge Univ. Press.

CALPHAD: Computer Coupling of Phase Diagrams and Thermochemistry, Oxford: Pergamon Press. Includes publication of a summary of the Proceedings of the annual CALPHAD conferences.

Gaskell, David R. (1981), *Introduction to Metallurgical Thermodynamics,* Washington, DC: Hemisphere.

Grimwald, Mark (1992), The Metallurgy of Gold, *Interdisciplinary Sci. Rev.* **17,** 371–381.

Guinier, A. (1959), "Heterogeneities in Solid Solutions," in: H. Ehrenreich, D. Turnbull (Eds.), *Solid State Physics,* Vol. 9, New York: Academic, pp. 293–398.

Haasen, Peter (1978), *Physical Metallurgy,* Cambridge, U.K.: Cambridge Univ. Press.

Massalski, T. B., Okamota, H., Subramanian, P. R., Kacprzk, L. (Eds.) (1990), *Binary Alloy Phase Diagrams,* 2nd ed., Materials Park, OH: ASM International.

Matsumiya, Tooru, Nogami, Atsushi, Sawada, Hideaki (1995), "Monte Carlo Simulations of Intermetallic Compounds," in: *Advanced Materials and Processes,* Materials Park, OH: ASM International, 51–53.

Nambu, Shinji, Sato, Akini, Sagala, D. A. (1992), "Computer Simulation of Spinodal Decomposition in the Tetragonal TiO_2–SnO_2 System," *J. Am. Ceram. Soc.* **79,** 1906–1913.

Paneth, F. A. (1960), "The Discovery and Earliest Reproductions of the Widmanstätten Figures," *Geochim. Cosmochim. Acta* **18,** 176–182.

Quasicrystals: a. (1991), *Reports on Progress in Physics,* Vol. **54**, London: Institute of Physics, 1373–1425. b. (1993), *J. Non-Cryst. Solids* **153–154,** Elsevier, Amsterdam.

Sakurai, Yoshifumi (Ed.) (1993), *Current Topics in Amorphous Materials,* Amsterdam: Elsevier.

Seitz, Fred (1992), *The Science Matrix,* Berlin: Springer. A review of the development of physical thought from ancient time to the present.

Stosks, G. M., Turchi, P. E. A. (1994), *Alloy Modeling and Design,* Pittsburgh: TMS-AIME.

Vickers, Charles (1923), *Metals and Their Alloys,* New York: Henry Cary Baird and Co. An excellent account of the state of knowledge of metallic alloys in the pre-X-ray diffraction, pre-electron microscopy, pre-Band Theory era.

Villers, Pierre, Prince, Alan, Okamoto, Hiraoki (Eds.) (1995), *Handbook of Ternary Alloy Phase Diagrams,* Vols. 1–10, Materials Park, OH: ASM International. The largest compilation of ternary phase diagrams yet published.

SONIC NOISE

George C. Maling, Jr., *Noise Control Foundation, Poughkeepsie, New York, U.S.A.*

Introduction 617
1. **Measures of Noise Immission and Noise Emission** 618
1.1 Sound Pressure and Sound-Pressure Level 618
1.2 Sound Power, Sound-Power Level, and Sound Intensity 618
1.3 Determination of Sound Power and Sound-Power Levels 619
1.3.1 Determination of Sound Power in Hemi-Anechoic Rooms 619
1.3.2 Determination of Sound Power in a Reverberant Room 620
2. **Sound Spectra** 620
2.1 Broad-Band Sound-Pressure Levels 620
2.2 Octave and One-Third Octave Band Levels 620
2.3 Narrow-Band Levels 621
3. **Instrumentation for Noise Measurement** 621
3.1 The Sound-Level Meter 621
3.2 Spectrum Analyzers 622
4. **Mechanisms of Noise Generation** 622
5. **Stationary Noise Sources** 623
6. **Moving Noise Sources** 623
6.1 Aircraft Noise 623
6.2 Road-Vehicle Noise 624
7. **Sound Propagation** 624
7.1 Sound Propagation Indoors 625
7.2 Sound Propagation Outdoors 626
8. **Sound-Attenuating Elements** 627
8.1 Outdoor Noise Barriers 627
8.2 Transmission Loss of Walls and Other Structures 628
8.3 Sound-Absorptive Materials 628
9. **Active Noise Control** 628
10. **Effects of Noise on People** 629
10.1 Physiological Effects 630
10.2 Psychological Effects 630
11. **Criteria for Noise Control** 631
11.1 Indoors 631
11.2 Outdoors 632
12. **Regulatory Aspects of Noise: Standards** 633
12.1 Regulations 633
12.2 Standards 633
Acknowledgments 634
Glossary 634
Works Cited 636
Further Reading 637

INTRODUCTION

Noise is defined as *undesired sound*. The classical *system problem* approach (Bolt and Ingard, 1957) to any problem related to the control of noise is to treat the problem in three parts, the *source*, the *path*, and the *receiver*. In this approach, each component of the problem can be examined separately, and appropriate action can be taken. Even when noise *control* is not the objective, the system approach is useful for describing sonic noise.

From a physical viewpoint, noise is equivalent to sound, and the physics of noise generation and transmission over a path to a receiver have little to do with the term *noise*. When the sound reaches a receiver, a human being or an animal, however, the extent to which the sound is undesired has to be considered. The important effects are primarily psychological (i.e., loudness or annoyance) and physiological (i.e., damage to the hearing mechanism). It is the combination of the physical aspects of generation and propagation of sound, its effects on receivers, and the need to find engineering techniques for noise abatement that makes noise and its control an important professional discipline for noise-control engineers.

There is a wide variety of noise sources in modern society. Sources in and around the

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

home include gardening equipment, lawn-maintenance equipment, kitchen and home-cleaning appliances, and furnaces. Sources at work include industrial machinery, ventilating-system noise, and office equipment. Sources that intrude upon the community include surface-transportation noise—generated by automobiles, trucks, trains, and outdoor recreational vehicles—and aircraft—commercial, general aviation, military, and helicopters.

It is not possible to detail all of the characteristics of all of these sources in this article. The term *emission* is used when the acoustical characteristics of a source are being described. The sound-power level of the source or the sound-pressure level measured under carefully prescribed environmental conditions are common descriptors of noise emission.

The path over which the source reaches the receiver generally provides some attenuation of noise, although, in some cases, the noise may be amplified. Indoor and outdoor sound propagation must both be considered. The characteristics of silencers, mufflers, and similar devices placed between a source and a receiver are considered to be part of the path. Characteristics of sound-absorptive materials used to change the attenuation over the path are an important part of noise control.

Finally, the effects of noise on the receiver must be considered. Important topics include methods for rating the effects of noise, both psychological and physiological, on humans and animals; criteria for control of noise; and regulatory aspects of noise—the setting and enforcement of limits on noise. The sound-pressure level is the commonly used measure of noise incident on a receiver, and the term *noise immission* is sometimes used to describe noise that reaches a receiver.

1. MEASURES OF NOISE IMMISSION AND NOISE EMISSION

1.1 Sound Pressure and Sound-Pressure Level

The immission of noise upon a receiver is usually described in terms of sound-pressure level. Sound consists of the very small variations in pressure p above and below atmospheric pressure, which propagate with a speed c, the speed of sound. A propagating plane wave has the form

$$p(x,t) = p(x - ct), \tag{1}$$

where x is distance and t is time. Although air is fundamentally a nonlinear medium, the pressure fluctuations are, in most situations, so small relative to the atmospheric pressure that a linear approximation may be used. Exceptions include propagation of sonic booms and high-intensity sound generation from engines.

Sound pressure at some location is conventionally defined as a root mean square (rms) pressure averaged over a stated time duration. The level of the sound pressure is determined from the mean square pressure $\overline{p^2}$ and a reference sound pressure p_0, defined to be 20 μPa. Thus, the sound-pressure level may be expressed as

$$L_p = 10 \log(\overline{p^2}/p_0^2). \tag{2}$$

As an example, a sound-pressure level of 60 dB corresponds to an rms sound pressure of 0.02 Pa. This is small compared with atmospheric pressure, which is approximately 10^5 Pa.

The particle velocity u associated with a sound wave is obtained from the linear (one-dimensional) momentum equation given by

$$\rho \frac{\partial u}{\partial t} = -\frac{\partial p}{\partial x}, \tag{3}$$

where p and u are instantaneous values of the pressure and particle velocity at some location and ρ is the density of the medium (air).

1.2 Sound Power, Sound-Power Level, and Sound Intensity

The sound-pressure level defined above may, in some cases, be used as a measure of the noise *emission* of a sound source. An example is specification of the noise emissions of road vehicles. The sound pressure measured under well-defined environmental conditions is also sometimes used to specify the noise emission of machinery and equipment (*emission sound-pressure level*).

The specification of the noise emissions of sources is important because governmental bodies are beginning to require the use of national and international standards to determine noise emissions in terms of sound-power level and emission sound-pressure level, and to require the publication of noise-emission values in specifications for industrial machinery and consumer products.

For stationary noise sources, the sound power and sound-power level produced by the source are most satisfactory as descriptors of noise emission. The sound power W emitted by a source is expressed (Beranek and Vér, 1992, Chap. 4) in terms of the surface integral

$$W = \int_S \mathbf{I} \cdot d\mathbf{S}, \tag{4}$$

where W = sound power (W), $\mathbf{I}$ = sound intensity (W/m^2), $d\mathbf{S}$ = element of surface area (m), and S = surface that surrounds the source. The sound intensity $\mathbf{I}$ is the average rate of sound energy transmitted in the specified direction per unit area at a given position. The integral of Eq. (4) is over a surface that surrounds the source (the *measurement surface*), and the dot product of the two vectors indicates that the component of sound intensity normal to the surface is taken. The integral may be taken over any closed surface surrounding the source (hemispherical, spherical, or irregular), but in practice, rectangular or hemispherical measurement surfaces are most often used.

Most noise sources of practical interest are *directive,* which means that the sound intensity varies over the measurement surface. In practice, the surface integral of Eq. (4) is approximated by dividing the measurement surface into segments, denoted by S_i, and by replacing the integral by a sum over the segments:

$$W = \sum_{i=1}^{N} I_i S_i. \tag{5}$$

Here, I_i is the magnitude of the sound intensity normal to the ith surface segment, and S_i is the area of the ith surface segment.

When the sound power W has been determined, the sound-power level L_W, expressed here in bels, is determined relative to a reference sound-power level W_0 of 10^{-12} W by

$$L_W = \log(W/W_0). \tag{6}$$

The sound-power level may also be expressed in decibels. One decibel is one-tenth of a bel. When the decibel is used as the unit of both sound-power level and sound-pressure level, confusion between the quantities may arise.

1.3 Determination of Sound Power and Sound-Power Levels

Sound intensity and therefore sound power can be measured directly with modern acoustical instruments. If the sound pressure at each of two closely spaced microphones is measured, the difference between these two pressures divided by the microphone spacing is an approximation to the pressure gradient on the right-hand side of Eq. (3), and the particle velocity in the sound wave can be determined by time integration. Standard methods have been developed (ISO, 1993, 1996) for selecting the number of surface segments used in Eq. (5) and determining the accuracy of the calculated sound-power level.

Determination of sound power by means of sound-pressure level measurements is a more widely used technique. Two test environments are in common use: anechoic (or hemi-anechoic) rooms and reverberation rooms.

1.3.1 Determination of Sound Power in Hemi-Anechoic Rooms The sound field produced by a source in a hemi-anechoic room approximates a free field over a reflecting plane. The sound intensity normal to a surface that surrounds the source is approximated by

$$\mathbf{I} = \overline{p^2}/\rho c, \tag{7}$$

where $\overline{p^2}$ is the mean square sound pressure at a point, ρ is the density of air, and c is the speed of sound. Using nominal values for ρ and c and standard reference quantities for sound pressure and sound power, the sound-power level (in bels) produced by the source is given approximately by

$$L_W = 0.1[L_p + 10\log(S/S_0)], \tag{8}$$

where L_p is the average sound-pressure level in decibels over the measurement surface, S is the area of the measurement surface in square meters, and S_0 is 1 m^2.

1.3.2 Determination of Sound Power in a Reverberant Room A reverberation room is a room in which all surfaces are highly reflective for sound waves. The sound-pressure level produced by a source is approximately independent of the field position within the room, and the space-average sound-pressure level can be determined from measurements at a few microphone positions or by means of a moving microphone. The sound-power level of a source can be calculated from the space-time–averaged sound-pressure level and a knowledge of the reverberation time (see ACOUSTICS, ARCHITECTURAL) in the room. A more commonly used method is to compare the average mean square sound pressures in the room produced by the source under test and by a source of known sound-power output (a *reference sound source*). In this *comparison method,* the sound-power level of the source under test, L_{WE} (in bels), is given by

$$L_{WE} = L_{WR} + 0.1(L_{pE} - L_{pR}), \tag{9}$$

where L_{WR} = sound-power level of the reference sound source, L_{pR} = average sound-pressure level in the room produced by the reference sound source, and L_{pE} = space-time–averaged sound-pressure level produced by the sound source under test.

An overview of standard methods of determining sound-power level from sound-pressure level measurements is given in ISO (1980).

2. SOUND SPECTRA

The sound pressure at a point is a function of time and, therefore, may also be described in the frequency domain. Three descriptions of noise spectra are in common use. The frequency-domain descriptions below may also be used for sound-power levels.

2.1 Broad-Band Sound-Pressure Levels

A single-number rating for a noise spectrum is obtained by determination of the level of the mean square pressure over a wide frequency range such as 20 Hz to 20 kHz—depending on the characteristics of the spectrum and the frequency response of the microphone used for the measurements. For sound-level meters, the standard frequency weighting C provides a good approximation to a broad-band sound level. This frequency weighting C is not widely used, except for determination of peak levels of noise of impulsive character.

Another standard frequency weighting is A-weighting. Figure 1 illustrates the design goal for the response of frequency weighting A. The response for frequency weighting A decreases especially at frequencies less than 1 kHz and approximates the frequency response of the human ear at low sound-pressure levels. Sound-pressure levels measured with use of this weighting are called *A-weighted sound levels.* Although it is the *level* that is weighted, and not the *unit of level,* the decibel (dB), the designation dB(A) or even dBA is sometimes used to indicate the A-weighted sound-pressure level in decibels. When A-weighting is applied to a sound-power spectrum, the result is defined to be the *noise-power emission level,* i.e., the A-weighted sound-power level in bels.

2.2 Octave and One-Third Octave Band Levels

While the A-weighted sound-pressure level frequently is adequate for evaluating human response to noise, a more detailed description of the spectrum is required for the specification of the noise emission of sources, for selection of noise-control measures, for diagnosis of noise problems, for determination of sound transmission through structures, and often for calculation of single-number ratings related to human response to noise (such as loudness). A description of noise spectra in terms of *octave* or *one-third octave bands* is common. The transmission characteristics of octave and one-third-octave band filters have been standardized. An octave band or one-third octave band is usually specified by its nominal midband frequency. The lower pass-band frequency of an octave-band filter is one-half the upper pass-band frequency. There are three one-third octave bands in an octave band. For most noise spectra, an adequate description of the spectrum is given by octave-band levels having

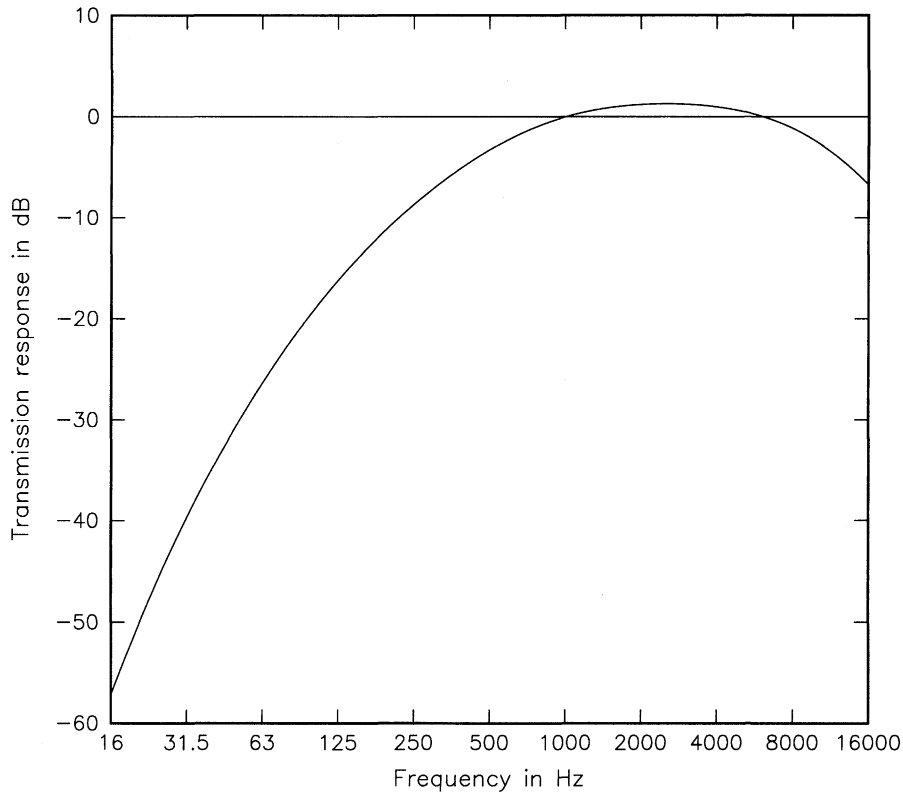

FIG. 1. Relative response vs frequency for A-weighting.

nominal midband frequencies of 125, 250, 500, 1000, 2000, 4000, and 8000 Hz.

2.3 Narrow-Band Levels

There are some situations where a description of a spectrum in narrow bands is required. Examples include the spectra of some axial-flow fans, propellers, and some rotating electrical machinery. Nearly all available narrow-band spectrum analyzers perform narrow-band analysis by using the fast Fourier transform; the characteristics of these instruments are not standardized; however, 4096-line analyses are common. Constant-percentage narrow-band spectrum analyzers are also available and provide the narrow-band levels over a wide range of frequencies in "real time."

3. INSTRUMENTATION FOR NOISE MEASUREMENT

3.1 The Sound-Level Meter

The basic instrument for measurement of noise levels is the *sound-level meter.* The two basic measurements obtained with a sound-level meter are exponential-time–averaged sound levels measured with standardized frequency weightings and time-averaged sound levels determined without exponential time weighting. The exponential time weightings are designated *fast* (or F) and *slow* (or S). The former corresponds to a time constant of 0.125 s, and the latter corresponds to a time constant of 1 s. All conventional sound-level meters are required to have frequency weighting A and time weighting F. Using these time weightings, one can estimate an average level over a specified time interval or determine the maximum sound level with a given time weighting. An example of a time-averaged sound level is an 8-h average sound level that one might measure over the duration of an 8-h work period. The measurement microphone for a sound-level meter is typically a capacitor, electret, or crystal microphone.

A block diagram of a typical sound-level meter is shown in Fig. 2. The signals are amplified, filtered (frequency weighted), and averaged using rms averaging. The output indicator may be either analog or digital, or a combination of digital and quasi-analog. The electroacoustical characteristics of sound-level meters and integrating-averaging sound-level meters have been standardized (IEC, 1979, 1985).

3.2 Spectrum Analyzers

In sound-level meters, the standard frequency weightings are usually implemented by means of analog circuit elements. Design-goal frequency weightings may also be applied to constant-percentage bandwidth sound-pressure levels or to the output of a fast Fourier-transform analyzer.

The sound-level meter may be a simple hand-held analyzer or part of a complicated data acquisition system in which frequency-weighted sound-pressure signals from several microphones are digitized, and the results are stored in computers for later analysis or transmitted directly to a central computer for real-time processing. Some of the most complex systems in use are for monitoring of noise levels around airports (Timmerman, 1993; Nitsche, 1995).

4. MECHANISMS OF NOISE GENERATION

There are four fundamental mechanisms for the generation of sound. These are usually described as *source terms* in the wave equation. The basic fluid flow equations (see ACOUSTICS, LINEAR; ACOUSTICS, NONLINEAR) are usually put in the form of a wave equation for a medium at rest with source terms on the right-hand side of the equation. The sound-pressure levels are assumed to be small enough to avoid nonlinear effects. The mathematical details are in the articles referenced above. For the purposes of this chapter, the acoustic wave equation may be written

$$\nabla^2 p - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = S_m + S_d + S_q + S_h. \qquad (10)$$

The left-hand side of Eq. (10) is a wave equation for a medium at rest. The first of the four source terms on the right-hand side of Eq. (10), S_m, represents generation of noise by fluctuating *volume* (monopole) sources. A classical example is the baffled loudspeaker where the loudspeaker diaphragm "pumps" air and generates sound. The second term, S_d, represents fluctuating forces in the medium. For example, air turbulence can cause fluctuating lift forces on an airfoil, and sound is generated by these forces. This mechanism is responsible for the noise generated by low-speed axial-flow fans and centrifugal blowers. Such sources are often referred to as *dipole* sources because there is no net introduction of mass into the medium. A classical example is two closely spaced monopole sources that operate out of phase. Dipole sources radiate sound less efficiently than monopole sources because of the partial cancellation of the sound waves. This principle is used in many methods for "cancellation of noise by noise," or "active noise control," a subject of intense current research. See Sec. 9 for a discussion on active noise control.

The third source term, S_q, represents a source term for generation of sound by aerodynamic sources (quadrupole sources) and was first understood by Lighthill (1954). The

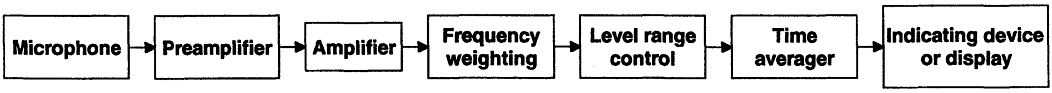

FIG. 2. Block diagram of the principal components of a sound-level meter. The time averager can include exponential weighting for sound-level meters. The frequency response of the microphone is sometimes included in the frequency weighting.

basic mechanism is momentum fluctuations across fixed surfaces in a fluid; sound generation takes place without the influence of solid surfaces. This mechanism is responsible for sound generation by turbojet engines where the major portion of the radiated sound is generated in the air exhaust downstream of the engine itself. A classical description of quadrupole sources is in terms of two closely spaced dipoles where the dipole moments are in opposite directions. Quadrupole sources are less efficient radiators of sound than dipole sources because of the cancellation that takes place between the two dipoles.

The fourth source term, S_h, represents the generation of sound by the introduction of fluctuating heat source into a medium. It is a less common source than those described above. One example is the *Rijke phenomenon,* which occurs when heated filaments are introduced into the lower half of a vertical tube. A strong tone is generated by a feedback mechanism in which the particle velocity in the air induces fluctuating heat flow from the heated filament and acts as a source of sound.

5. STATIONARY NOISE SOURCES

Noise emissions from stationary sources are usually described by the sound-power level emitted by the source. Typical sources include machinery in factories, large air-moving devices used to circulate cooling air in buildings, air-conditioning units, and business equipment used in offices.

The general approach to reduce the noise from stationary sources is to determine the relative strength of the noise emitted from various parts of the source, and to apply appropriate noise-control measures. Sound intensity is one useful tool for identifying noise sources. Another tool is the selective covering or wrapping of portions of a machine with a heavy material such as lead-loaded vinyl in combination with sound-absorptive materials.

Noise-control measures may include replacement of solid surfaces such as protective covers with perforated metal or wire screens, reduction of the size of vibrating surfaces, reduction of the speed at which impacts occur, and reduction of the area of impacting surfaces, damping machine surfaces with materials that absorb energy when flexed, enclosing areas of high noise emission, lowering the speed of rotating equipment, mechanical isolation of equipment so that it does not transmit sound energy into structures, and reduction of the time rate of change of force and pressure.

6. MOVING NOISE SOURCES

Aircraft and road vehicles are the two most common examples of noise generated by moving sources. Although there have been attempts to use sound-power level as the measure of noise emission of such sources, simplified measurement procedures have been adopted for practical reasons. A review of transportation noise in general has been given by Nelson (1987).

6.1 Aircraft Noise

For jet-powered and propeller-driven airplanes, the engines are the principal sources of noise during ground roll, climbout, and landing. Sound produced by vortices shed by the extended landing gear and by wing leading- and trailing-edge devices (flaps) contributes to the noise level under landing-approach flight paths and may be an important source of noise in the future as engine noise is reduced.

The methodology of determining the noise emission of aircraft for certification purposes is well developed and will be covered in this section. Certification noise levels are not intended to represent noise levels that are measured in communities around airports; methods for predicting noise levels around airports are described in Sec. 10.2.

Certification noise limits are specified at three measurement points—known as *lateral* (or sideline), *flyover* (or takeoff), and *landing approach.* The limits are a function of takeoff gross weight and apply to both jet-powered and propeller-driven airplanes. The limits apply to new-design transport-category civil airplanes and to retrofit modifications of old-design transport-category civil airplanes.

Perceived noise levels are determined from 500-ms–average one-third octave band sound-pressure levels, at 500-ms intervals, and are adjusted for the additional annoyance caused by prominent tones, if present. Data measured during a noise-certification compliance test are adjusted to reference atmospheric conditions, reference flight paths, and reference engine-power settings.

Certification limits are set by Part 36 of the Federal Aviation Regulations in the United States (FAR 36) and by Annex 16 to the Convention on International Civil Aviation (ICAO, 1993). The effective perceived noise level (EPNL), a quantity that is regulated by the above documents, is determined from the time integral of tone-corrected perceived noise level.

The first issue of FAR Part 36 became effective in 1969; ICAO Annex 16 was first issued in 1971. The main objective of those regulations was to ensure that noise produced by new aircraft would be less than the noise produced by earlier designs. In the years since the initial issue, the certification noise-level limits were gradually reduced in stages. Stage 1 airplanes are the noisiest and include the original Boeing 707 and McDonnell Douglas DC-8. Stage 2 airplanes have lower noise levels than stage 1 airplanes, and the noise levels of airplanes that comply with stage 3 are even lower. For example, the lateral (sideline) effective perceived noise-level limit at a takeoff gross weight of 200 000 lb ($\approx$91 Mg) is 104.8 dB for a stage 2 airplane with two or three engines, and 97.5 dB for a stage 3 airplane—independent of the number of engines. An airplane that complies with the stage 3 requirements of FAR Part 36 also complies with the Chapter 3 requirements of ICAO Annex 16.

A summary of engine-noise–control designs is available (Hubbard, 1991). A summary of retrofit applications intended to allow stage 2 (Chapter 2) airplanes to meet a phaseout deadline of the year 2000 for stage 3 (Chapter 3) compliance has been published by Marsh (1994). Eldred (1992) has given an overall review of the aircraft noise problem.

6.2 Road-Vehicle Noise

The main sources of road-vehicle noise are engine noise, exhaust noise, and tire–pavement interaction noise. Which source predominates depends on the speed of the vehicle, the condition of the muffler, the design of the tires, and the road surface. Typical noise-reduction measures for vehicles include the use of sound-absorptive materials, optimization of the exhaust-system muffler, optimal design of the cooling fan, and control of gear and transmission noise.

Authorities in most industrialized countries have introduced legislation to control road-vehicle noise by establishing noise-emission limits. New vehicles in many countries are subject to "type approval," and older vehicles may, in some countries, be required to meet noise limits for vehicles in use. The most stringent noise regulations for new road vehicles are currently (1996) in effect in the European Union (Directive 92/97/EEC) where, for example, automobiles must comply with a limit of 74 dB and heavy trucks with a limit of 80 dB. The measurement procedures are described below.

Nationally and internationally standardized noise-measurement procedures for road vehicles differ in some respects, but ISO 362 (ISO, 1994) is widely recognized. This document specifies that the test vehicle is driven over a designated test area with a standardized road surface. Fixed microphones are located 7.5 m from the vehicle path center line. (In the United States, the microphone distance is 15 m, and the relevant standards are SAE J986 and SAE J366.) Vehicles approach the test area at a constant speed, usually 20 to 50 km/h, but the throttle is fully open when the vehicle is within ± 10 m of the microphone positions. The maximum A-weighted fast sound level during the acceleration is measured.

A general review of road-vehicle noise is available (Sandberg, 1995). This review covers the effects of regulations on road-vehicle noise emissions on the noise levels measured in communities.

7. SOUND PROPAGATION

Equation (1) describes the propagation of a plane wave; there is no attenuation of sound with distance. Since sound generally propagates as a spherical wave or into a hemisphere above a reflecting plane, a $1/r$ dependence must be added to Eq. (1) to determine the rms sound pressure at a location at a distance r from the source. The sound-pressure level decreases at a rate of 6 dB per doubling of distance from the source. The sound intensity varies inversely as the square of the distance from the source (inverse-square spreading). This spreading must be taken into account for sound propagation indoors and outdoors. Other effects that influence the variation of sound level with distance from the source are different for sound propagation indoors and propagation outdoors and will be described separately.

7.1 Sound Propagation Indoors

Indoors, the most important effect that modifies the $1/r$ dependence of sound pressure with distance from the source is the reflection of sound from fixed room surfaces. The amount of sound reflected depends on the sound-absorptive characteristics of the surface, usually described in terms of the *absorption coefficient.* There is a variety of methods used to describe absorption coefficients (see ACOUSTICS, ARCHITECTURAL), but for the purposes of this article, the absorption coefficient α is defined as

$$\alpha = 1 - |R|^2, \tag{11}$$

where $|R|^2$ is the squared magnitude of the reflection coefficient, p_r/p_i. Here p_r is the rms amplitude of the reflected sound and p_i is the rms amplitude of the sound incident on the surface.

The classical variation of sound level with distance in a room occurs when reflections from all room surfaces contribute equally to the level of the reflected sound in the room. In this case, the sound field is conveniently described in terms of *direct sound* and *reflected sound.* Near the source, the $1/r$ term predominates, and far from the source, sound reflected from the room surfaces predominates. The total mean square pressure is obtained by adding component mean square sound pressures, since, at least for broad-band sounds, the direct and reflected sounds are essentially uncorrelated.

Under these conditions, the variation of sound level with distance is (see ACOUSTICS, ARCHITECTURAL)

$$L_p(r) = L_W + 10 \log\left(\frac{Q}{4\pi r^2} + \frac{4}{R}\right), \tag{12}$$

where all terms are in decibels and L_W is the sound-power level of the source. Here, r is the distance from the source, Q is a source directivity factor (Beranek and Vér, 1992 Chap. 4), and R is the room constant, which is related to an average Sabine absorption coefficient α_s (neglecting air absorption) by $R = \alpha_s/(1 - \alpha_s)$. At large distances from the source, the level is uniform within the room, and one is in the *reverberant field.* The room constant R is not equal to the reflection coefficient R in Eq. (11).

In most situations of interest in noise control, the classical relationship of Eq. (12) rarely holds because the reflections from all of the room surfaces usually do not contribute equally to the sound-pressure level at a point. Many rooms of practical interest have low ceilings that are partially sound absorptive, and the uniform level in the reverberant field does not occur. In addition, any room of interest contains scattering objects, and the scattering from these objects must be accounted for in any calculation of sound-pressure level vs distance from a source.

Methods for prediction of noise levels in factories have been described by Hodgson (1986). One useful method for describing the variation of sound level with distance indoors has been published by the European Computer Manufacturers Association (ECMA, 1985).

For a "flat room" having a length and width $>3h$ where h is the room height, the variation of sound-pressure level L_p with distance d from a source having a sound-power output L_W may be calculated from

$$L_p = L_W - 20 \log(h/h_0) + \Delta L_f \tag{13}$$

(taken from the ECMA report), where h_0 is 1 m and ΔL_f is given as a function of d/H in Fig. 3. The two parameters in the figure are q, the density factor of the reflecting objects in the room ($q = 0.2$ in the figure), and the average absorption coefficient, $\bar{\alpha}$. The density factor is given by

$$q = (1/4S_f)\Sigma S_e, \tag{14}$$

where S_f is the surface area of the room floor in square meters and ΣS_e is the total area of the scatterers in the room in square meters. The average absorption coefficient is the average of the Sabine absorption coefficients of the ceiling and floor of the room at

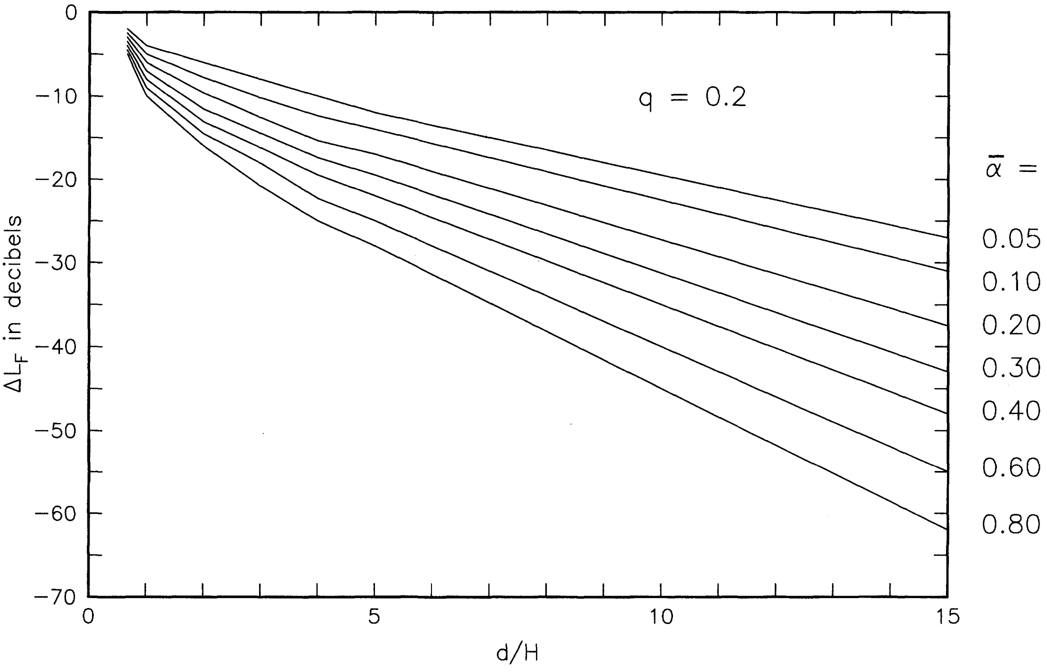

FIG. 3. Correction, ΔL_F, for calculating the decrease with distance of the sound-pressure level from a source for a density factor $q = 0.2$.

frequencies of 250, 500, 1000, and 2000 Hz. For Eq. (13), unlike the classical reverberant-room theory described by Eq. (12), the sound-absorptive properties of the walls in the room are not important.

A second effect that changes the variation of sound-pressure level with distance in rooms is *air absorption.* Sound is absorbed in air because of viscous and heat-conduction effects and (mainly) by oxygen and nitrogen vibrational relaxation phenomena (ANSI, 1995b). This mechanism affects sound levels in large spaces such as concert halls, and is very important in rooms that have very highly reflective room surfaces or for small-scale models, but does not usually play a major role in indoor noise problems.

7.2 Sound Propagation Outdoors

Inverse-square spreading is the most important effect that reduces the sound-pressure level as distance from a source is increased. There are, however, several other factors that influence outdoor sound propagation:

1. Atmospheric absorption. At long distances and at high frequencies, the effects of atmospheric absorption are significant. These effects are usually described by an attenuation coefficient in decibels per meter. Atmospheric absorption is greatly dependent on air temperature and water-vapor content (ANSI, 1995b). As an example, at 4000 Hz, at a temperature of 20 °C and a relative humidity of 50%, it is 29 dB/km (Harris, 1991).
2. Ground effects. When the sound source is located above a ground surface, sound waves that reflect from the ground will constructively and destructively interfere with those propagating directly from the source. In general, the ground is partially reflecting; the reflected wave is modified in amplitude and phase by its interaction with the ground surface. The amount of attenuation attributable to this ground interaction and its variation with frequency depend on the surface characteristics, the source and receiver heights, and their separation. The effects of the ground are largest for intermediate frequencies (~500 Hz) when the source is above the ground

(1 m or more). If the source is close to the ground at all frequencies, all frequencies above about 500 Hz are highly attenuated.

3. Temperature gradients and wind-speed gradients. The speed of sound is proportional to the square root of the absolute temperature of the medium. The normal temperature lapse with height above the ground means that a sound wave front moves more rapidly near the ground surface. This causes the wave front to bend upward, creating a "shadow zone" of low sound-pressure level near the surface. Wind-velocity gradients have a similar effect since the speed of the wave is the wind speed plus the speed of sound. Since the wind speed is usually less close to the ground, propagation upwind tends to bend the wave front upward, creating a shadow zone similar to that produced by a normal temperature lapse. For sound propagation downwind, the opposite effect occurs, and the wave front is bent toward the ground; there is no significant attenuation of the sound wave relative to the attenuation in the absence of wind.
4. Turbulence. Since turbulence always accompanies wind outdoors, the effects of atmospheric turbulence must be included in any analysis of sound propagation outdoors. There are three major effects of turbulence. First, at single frequencies, atmospheric turbulence affects the coherence between waves that propagate directly from the source and waves that are reflected from the ground surface. This results in sound-pressure levels that are higher than would be expected in the absence of turbulence at those points where destructive interference between the waves occurs. Second, atmospheric turbulence produces scattering into the shadow zones caused by air-temperature and wind-speed gradients, tending to raise the sound-pressure level in these areas. Third, in a highly directive sound beam, the presence of turbulence results in an excess attenuation produced by sound scattering out of the beam. Turbulence is an important *source* of sound only in high-speed flows (Lighthill, 1954).

The basic equation for sound propagation outdoors is (Beranek and Vér, 1992, Chap. 5)

$$L_p(r) = L_W - 20\log(r/r_0) + 10\log Q - \log(\Omega/4\pi) + \text{constant} - A, \qquad (15)$$

where all terms are in decibels. Here r_0 is a reference distance of 1 m, Q is a directivity factor (Beranek and Vér, 1992, Chap. 4), the $\Omega/4\pi$ term is related to the solid angle into which the source propagates, the constant depends on the units used, and A is an attenuation that depends on all of the effects described above. Several reviews of sound propagation outdoors are available (Embleton, 1982, 1989; Embleton and Daigle, 1991; Ingard, 1953).

8. SOUND-ATTENUATING ELEMENTS

In the previous section, the reduction of noise over the path from the source to the receiver was discussed for the case where no attempt was made to erect a barrier between the source and the receiver. Such barriers take many forms. Outdoors, noise barriers are being installed along highways to protect residents from the effects of traffic noise. Indoors, walls and other structures form barriers that attenuate sound. In the design of building air-conditioning systems, sound-absorptive ducts are commonly used to reduce the noise level produced by ventilating fans in rooms. Sound-absorptive materials are also used in rooms and in factories to reduce sound levels. Machinery noise reduction is frequently accomplished through partial or full enclosures around machines.

8.1 Outdoor Noise Barriers

The actual performance of barriers designed, for example, to reduce traffic noise depends greatly on traffic flow, since the noise source is basically a line source. It also depends on the height of the source above ground so that the predominant noise sources must be identified. The source heights will be different for slow-moving trucks and for high-speed automobile traffic.

Computational methods for the prediction of barrier noise attenuation are changing rapidly. One set of programs used by noise-control engineers is the STAMINA 2.0/OPTIMA model (Bowlby *et al.*, 1982).

8.2 Transmission Loss of Walls and Other Structures

Indoors, walls and other structures act as barriers to noise. The motion of a wall is controlled by its mass and stiffness, the wall being driven by the pressure of the sound wave incident on one side of a wall. Sound is reradiated on the other side, usually at a much lower level, by the motion of the wall. The sound attenuation of a wall is generally described by a *mass law* (see ACOUSTICS, ARCHITECTURAL) modified by effects such as flanking transmission and coincidence effects. The latter occur when the bending-wave speed in the wall is equal to the speed of sound along the wall.

The sound attenuation of composite structures such as walls with windows is of interest when protecting residents near highways and airports from transportation noise. Sound intensity has proven to be a useful tool for measuring the transmission loss of both solid walls and composite structures. A review of progress in building acoustics and modern methods for the analysis of sound transmission paths is available (Cops and Vermeir, 1995).

8.3 Sound-Absorptive Materials

For a material to absorb sound, it must have a *flow resistance;* that is, the ratio of the pressure drop across the material, Δp, in pascals and the air flow velocity u in meters per second through it must be finite. The flow resistance ζ is usually normalized by the product of air density ρ in kilograms per cubic meter and the speed of sound c at standard sea-level air pressure and an air temperature of 15 °C:

$$\zeta = (\Delta p/u)/\rho c. \tag{16}$$

Materials that have a normalized flow resistance in the range 1–2 are normally good sound absorbers.

A second requirement is that the particle velocity in the sound wave within the material must not be too small in order to achieve high viscous losses from friction. As an example, a thin material with a finite flow resistance is placed close to a solid wall. An incident sound wave is reflected from the wall; the boundary condition is that the particle velocity in the sound wave normal to the (solid) wall is zero, and the incident pressure is doubled at the wall. The material will be ineffective as a sound absorber because of the low (near zero) particle velocity through the material. The same material placed in a region of high particle velocity may be a very effective sound absorber.

Sound-absorptive materials may be bulk materials such as glass-fiber materials or other porous materials. Sound is also absorbed by nonporous materials such as screens and perforated metals placed over partitioned backing materials (i.e., honeycomb structures). In many noise-control situations, a composite structure is used; one example is a bulk material protected by a perforated metal facing. Resonance effects in such structures often enhance the sound absorption at particular frequencies.

A complete description of sound-absorption technology, and computer programs that may be used to estimate sound attenuation by composite structures, sound attenuation in ducts, resonance effects, and sound attenuation in porous materials, is available (Ingard, 1994).

9. ACTIVE NOISE CONTROL

A sound field can be described in space and time by a sound pressure field $p(\mathbf{x},t)$ where $\mathbf{x}$ is a position vector and t is time. To apply active control to this field, one may use a secondary sound source that produces a field closely approximating $-p(\mathbf{x},t)$ within some region of space. Since, in a linear approximation, the two pressure fields add, there is a net reduction of sound pressure in the region. Active noise control is basically an interference phenomenon. Active control can also in theory be applied to the boundaries of an enclosed space as well as to vibrating structures.

The first patent on active noise control was obtained by Lueg (1934), and an early demonstration of how active control can be applied was given by Olson and May (1953). Since that time, improvements in transducers and the development of electronic controllers that use digital signal processing techniques have made active control of sound practical, at least at low frequencies, in many cases (Sommerfeldt and Hamada, 1995; Tichy, 1995; Elliott and Nelson, 1994; Nelson and Elliott, 1992).

Four application areas are of practical interest:

Noise-canceling headsets. Use of a microphone, electronic controller, and loudspeaker mounted in a headset to generate an out-of-phase signal in the ear canal has proved to be effective in reducing noise exposure for persons wearing the headsets (Bartholomae and Stein, 1991).

Global reduction of sound. Placing an out-of-phase secondary source in the vicinity of primary source effectively converts a monopole to a dipole and reduces the radiation efficiency of the primary source. The important parameter is *kd*, where *k* is the wave number and *d* is the distance between the primary and secondary sources. For effective cancellation, $kd \ll 1$. This technique has been demonstrated by Neise and Koopman (1988) in connection with the cancellation of blade-passage frequencies and harmonics in centrifugal fans.

Cancellation of sound in ducts. Use of sound-absorptive materials is an effective way of attenuating sound in ducts such as those used in buildings for heating and ventilating systems. At low frequencies, sound-absorptive materials are less effective, and active techniques may be used to reduce low-frequency sound. A *feed-forward* system is illustrated in Fig. 4. An input microphone detects the sound in the duct, and a loudspeaker is used as a secondary source. The electronic controller calculates the signal wave form to be sent to the loudspeaker in order to minimize, usually in a least-squares sense, the sound pressure at the error microphone. This technique is most useful at low frequencies when only a plane wave can propagate in the duct. See Beranek and Vér (1992, Chap. 15).

Spatial reduction of sound pressure. Effective reduction of the sound-pressure level in a given volume remote from a source is only possible at low frequencies, and it generally requires multiple microphones to sense the field, multiple loudspeakers to act as secondary sources, and a multichannel electronic controller to generate the optimal loudspeaker inputs. Such a system has been described by Peterson *et al.* (1994), and active control of road noise inside vehicles has been described by Sutton *et al.* (1994).

10. EFFECTS OF NOISE ON PEOPLE

As mentioned in the Introduction, receivers of noise are usually human beings; the effects can be broadly divided into two categories, physiological and psychological. A summary of the effects of noise on humans has been published for the World Health Organization (Bergland and Lindvall, 1995).

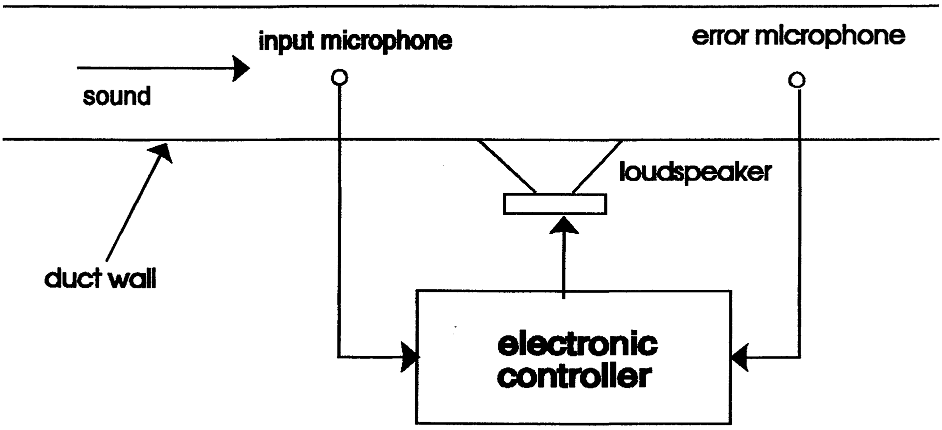

FIG. 4. Concept for a system to cancel sound in a duct.

10.1 Physiological Effects

The most important physiological effect of noise is hearing loss caused by long-term exposure to high levels of noise. This hearing loss may occur through occupational exposure—for example, workers in factories—or through noise exposure through leisure activities or activities around the home. The amount of hearing loss depends on both sound-pressure level and exposure time. It is widely believed (Suter, 1993) that cumulative noise exposure should be calculated on an "equal-energy" basis; that is, a 3-dB increase in sound level accompanied by a decrease by one-half in the exposure time results in the same noise exposure and produces the same amount of hearing loss. This position was adopted by the American Conference of Governmental Industrial Hygienists. The U.S. Department of Labor uses different criteria.

In the 1970s, the U.S. Environmental Protection Agency published a study of the effects of noise (EPA, 1974) and concluded that exposure to an A-weighted sound level of 70 dB averaged over a 24-h period is "safe" for the most noise-sensitive persons, at the frequencies of sound that cause the greatest damage, and when the exposure to noise is for long periods of time. For an exposure of 8 h/d, the "safe" level increases to 75 dB, according to the equal-energy theory.

In industrial situations, it is common to begin monitoring hearing acuity when 8-h exposure levels are 80 dB or greater and to provide hearing protection or engineering controls when 8-h exposure levels exceed 85 dB. The relationship between hearing loss, sound level, and exposure time has been standardized (ISO, 1990), and a description of worldwide requirements on exposure to hazardous noises and recommendations for future action is available (Embleton, 1994).

10.2 Psychological Effects

The primary psychological effect of noise is annoyance. Annoyance depends not only upon the physical characteristics of noise but also on the time of day, background sound levels, personal attitudes, and a number of other considerations. The most serious problems with outdoor noise in today's society are the effects of noise in communities around airports and exposure to road-traffic noise. The former are being addressed by careful control of takeoff and landing patterns of airplanes, by land-use planning around airports, and, in some cases, by sound insulation of homes near airports. A-weighted sound levels in a number of forms are used to describe noise levels around airports. These include the community noise equivalent level (CNEL) and the day–night average sound level (DNL). The day–night average sound level is described in the next paragraph.

Cumulative exposure to noise is often specified through the day–night average sound level L_{dn}, which is a 24-h average A-weighted level determined with the sound exposures during the hours of 24.00–07.00 and 23.00–24.00 increased by a factor of 10. In the 1970s, the U.S. Environmental Protection Agency identified 55 dB as the day–night average sound level "required to protect the public health and welfare with an adequate margin of safety" (EPA, 1974).

The psychological effects of single events are more difficult to quantify. One measure of the noise from a single event, such as an aircraft flyover or automobile passby, is the A-weighted *sound exposure level.* The sound exposure level is given by

$$L_{AE} = 10 \log\left[\frac{\int_0^T p_A^2(t)\,dt}{p_0^2 t_0}\right]. \tag{17}$$

Here, p_A^2 is the squared instantaneous A-weighted sound pressure. The reference sound exposure in the denominator of Eq. (17) is given by the square of the reference pressure, $p_0 = 20\ \mu\text{Pa}$, and the reference time, $t_0 = 1$ s.

As an aid to land-use planning near airports, the Federal Aviation Administration has developed (FAA, 1996) a new version of the Integrated Noise Model (INM, Version 5.1) that enables noise levels in communities around airports to be calculated considering such variables as type of aircraft, number of operations, and flight paths. The main enhancements in Version 5.1 are the addition of military aircraft, the noise-power–distance database from the U.S. Air Force NOISEMAP model, and support for the Microsoft Windows 95® operating system.

As an aid to land-use planning near highways, the U.S. Federal Highway Administration has developed a new (1996) highway noise prediction model, the Traffic Noise Model (Menge, 1996), to replace the STAMINA 2.0/OPTIMA model (Bowlby *et al.*, 1982). The Traffic Noise Model (TNM) incorporates variable ground impedance, reflection from sound-absorptive walls, multiple diffraction, and shielding from barriers, rows of buildings, and trees.

In the United States, near federally funded highways, traffic-noise reduction is being addressed through federal regulations (FHWA, 1976) and, in many cases, the construction of noise barriers along highways. Attempts are also being made to control road-traffic noise through noise-emission standards as described in Sec. 6.2. The relationship between community noise exposure and regulatory limits is being studied. It has been found (Sandberg, 1995) that imposition of limits does not always reduce the noise levels to which communities are exposed. A summary of the effects of noise on people has been given by Eldred and von Gierke (1993).

11. CRITERIA FOR NOISE CONTROL

11.1 Indoors

Criteria for control in the workplace for purposes of hearing conservation have been discussed above. Methods of rating noise for the purpose of providing a satisfactory working environment either use an A-weighted level or use a method that requires an octave-band (or one-third octave band) analysis. In the latter case, a set of *noise criterion* (NC) curves may be used to convert an octave-band spectrum into a single-number rating (Beranek, 1990). Alternatively, *balanced-noise criterion* (NCB) curves or *room criterion* (RC) curves may be used to determine when a given noise specification is met (Beranek, 1992; ANSI, 1995a).

For use of the NC curves, shown in Fig. 5, the octave-band spectrum is plotted on the graph, and the highest point of penetration into the set of NC curves determines an NC value for the spectrum. This is a "curve-tangent" method.

Use of the NCB or RC curves, illustrated in Figs. 6 and 7, is more complex but has

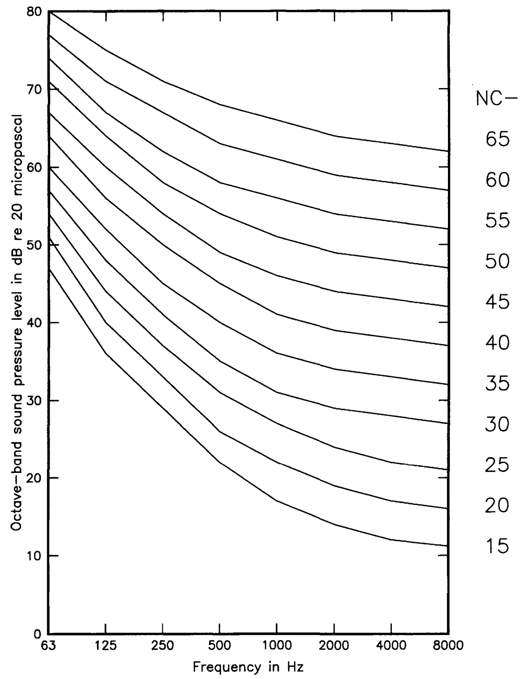

FIG. 5. Noise criterion (NC) curves.

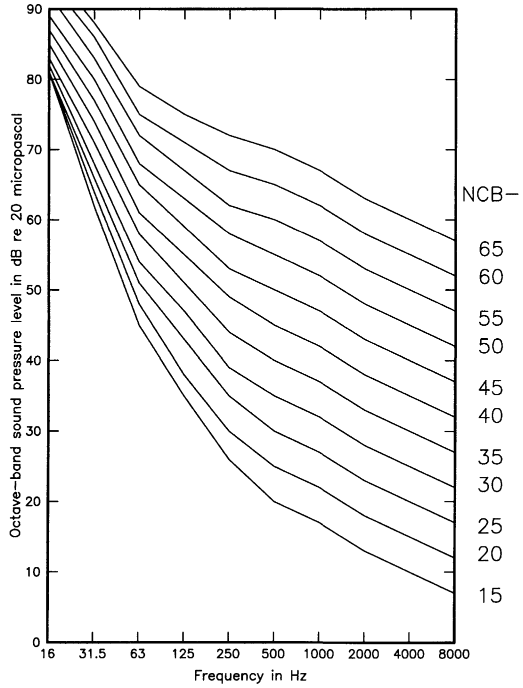

FIG. 6. Balanced-noise criterion (NCB) curves.

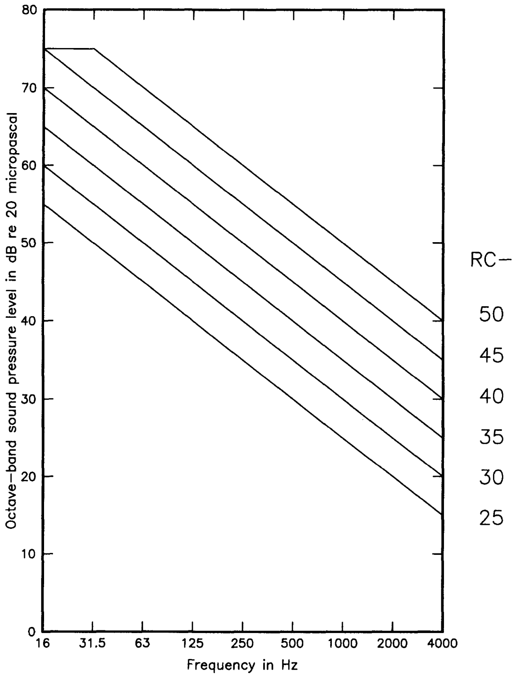

FIG. 7. Room noise criterion (RC) curves developed from measurements of noise from heating, ventilating, and air-conditioning equipment noises in unoccupied offices.

the advantage that the quality of the sound is evaluated and that *rumble* (low-frequency noise) and *hiss* (high-frequency sound) can be explicitly taken into account. If a room noise specification is given in terms of a NCB value (NCB–*NN*, where *NN* is an integer), a given octave-band spectrum complies with this specification if all of the following three criteria are met:

1. The average of the octave-band levels in the 500-, 1000-, 2000-, and 4000-Hz octave bands does not exceed *NN*.
2. The levels in the octave bands from 16 to 500 Hz do not exceed the average calculated in step **1** plus 3 dB.
3. The octave-band levels in the bands from 1000 to 8000 Hz do not exceed an NCB_{avg} curve, where this average curve is calculated by the following procedure (ANSI, 1995a). Find the three NCB curves that have the same levels in the 125-, 250-, and 500-Hz bands as the measured sound-pressure levels in those bands, respectively. Call the three curves so determined $\text{NCB}_h(f)$, and calculate the average, NCB_{avg}.

In the above procedure, step **1** is speech-interference level compliance, step **2** is rumble compliance, and step **3** is hiss compliance. An example of an NCB specification for an occupied activity area is that in a large conference room, an NCB curve in the range 25–30 is suitable (ANSI, 1995a).

If a specification is given in terms of an RC value (RC–*NN*, where *NN* is an integer), a given octave-band spectrum complies with this value if both of the following conditions are met (ANSI, 1995a):

1. The average of the octave-band levels in the 500-, 1000-, and 2000-Hz octave bands does not exceed *NN*.
2. Any level in the octave bands below 500 Hz does not exceed the RC–*NN* curve by more than 5 dB, and any level in the octave bands at 1000 Hz and above does not exceed the RC–*NN* curve by more than 3 dB.

The first criterion relates to average midband levels, and the second relates to rumble and hiss. An example of an RC criterion for an unoccupied space is that for conference rooms. An RC curve in the range 25–30 is a suitable specification (ANSI, 1995a).

Room criteria for noise from heating, ventilating, and air-conditioning systems have been discussed by Blazier (1995) and by Kingsbury (1995).

11.2 Outdoors

There has been a wide variety of noise surveys that attempt to define the relationship between community annoyance and some measure of noise emission. Two such measures are the sound level averaged over some time interval (i.e., the level of the rms sound pressure measured over a given time interval), and the day–night average sound level defined in Sec. 10.2. The results from several studies have been summarized by Eldred and von Gierke (1993) with a graph, shown in Fig. 8, for the best estimate of the percentage of the population highly annoyed vs the day–night average sound level. As an example, it is estimated that more than 30% of the population is highly annoyed by aircraft noise when the day–night average

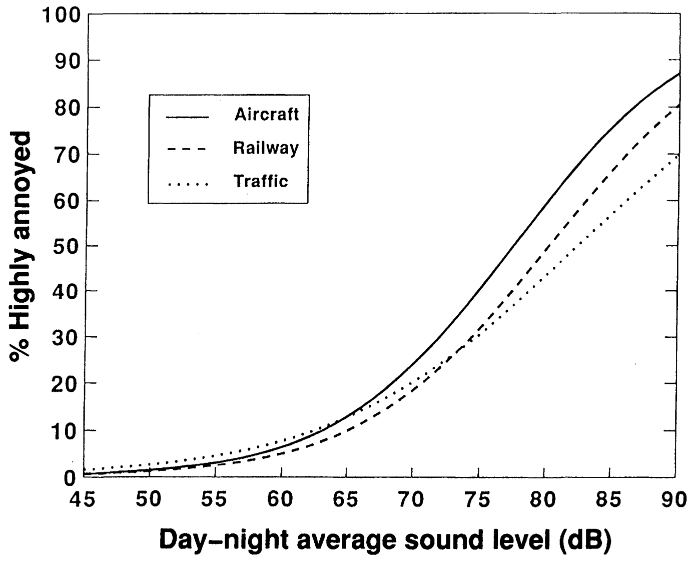

FIG. 8. Percentage highly annoyed vs day–night average sound level from aircraft, road traffic, and railway noise. Based on data from Fidell *et al.* (Eldred and von Gierke, 1993).

sound level exceeds 73 dB. Similar data have been summarized by Finegold *et al.* (1994).

12. REGULATORY ASPECTS OF NOISE: STANDARDS

12.1 Regulations

Noise regulations can be divided into two general areas, regulations on noise *emission* and regulations on noise *immission.* In the United States, major emphasis is placed on regulation of the noise emissions of aircraft through Federal Aviation Administration (FAA) regulations, and these regulations are generally in harmony with international requirements (ICAO). In the 1980s, the European Union began a program to regulate the noise emissions of specific noise sources, but the modern approach is to issue general regulations applicable to a wide variety of equipment and to prepare measurement standards for specific sources. A summary of European noise regulations has been given by Higginson *et al.* (1994). There is a variety of noise-immission regulations, many issued by state and local authorities. A summary of requirements was compiled by Beranek and Vér (1992, Chap. 17).

12.2 Standards

Several organizations produce standards related to noise and its measurement and control. Internationally, standards for measuring instruments are developed by Technical Committee 29 of the International Electrotechnical Commission (IEC). These standards include requirements and specifications for microphones and sound-level meters, sound calibrators, octave- and one-third–octave-band filter sets, and sound-intensity analyzers.

International standards on noise are developed by Technical Committee 43 of the International Organization for Standardization (ISO). Subcommittee 1, Noise, develops standards on determination of the noise emission of sources and other related topics. Subcommittee 2 develops standards on building acoustics, including sound transmission and determination of the properties of acoustical materials.

In the United States, there is a federated standards system coordinated by the American National Standards Institute (ANSI). Several national organizations produce noise standards in their specific areas of interest.

The Acoustical Society of America coordinates the activities of Accredited Standards Committees S1, S2, S3, and S12. Committee S1 develops standards on physical acoustics, including terminology, microphones, sound-level meters, and filter sets. S2 develops standards on mechanical vibration and shock, condition monitoring, and diagnostics. S3 develops standards on bioacoustics—standards that relate to safety, tolerance, and comfort. S12 develops noise standards, including noise-emission standards and standards for the description of environmental sound, for hearing conservation, and for occupational noise exposure.

There are several other national organizations that have significant activities related to noise and its control. The American Society for Testing and Materials (ASTM) develops standards related to building acoustics and acoustical materials, the Society of Automotive Engineers (SAE) develops standards for measurement of the noise of moving sources such as road vehicles and aircraft, and the American Society of Heating, Refrigerating, Airconditioning and Ventilating Engineers (ASHRAE) develops noise standards related to air-conditioning equipment.

There are several other professional and trade associations that develop standards related to specific areas of noise and its control.

ACKNOWLEDGMENTS

The author would like to thank T. F. W. Embleton, G. A. Daigle, A. S. Harris, R. K. Hillquist, K. U. Ingard, and U. Sandberg for reading portions of the manuscript and contributing many helpful suggestions for changes and additions to the text. A. H. Marsh read the entire manuscript and suggested many helpful changes and additions.

GLOSSARY

This glossary has been adapted from American National Standard acoustical terminology (ANSI, 1994) and a noise-control engineering glossary (Anon., 1995).

A-Weighted Sound-Pressure Level: Ten times the logarithm of the ratio of A-weighted squared sound pressure to the squared reference sound pressure of 20 μPa, the squared sound pressure being obtained with fast (F) (125-ms) exponentially weighted time averaging. Alternatively, slow (S) (1000-ms) exponentially weighted time averaging may be specified. Unit, decibel (dB); symbol, L_A.

Anechoic Room: Test room whose surfaces absorb essentially all of the incident sound energy over a frequency range of interest, thereby affording nearly free-field conditions over a measurement surface.

Average Sound-Pressure Level in a Room: Ten times the logarithm of the ratio of the space and time averages of squared sound pressure to the squared reference sound pressure, the space average being taken over the total volume of the room, except for the regions of the room where the direct field of the source and the near field of the boundaries are of significance. Unit, decibel (dB).

Bel: Unit of level when the base of the logarithm is ten, and the quantities concerned are proportional to power. Symbol, B.

Continuous Spectrum: Spectrum of a wave whose frequency components are continuously distributed over a frequency region.

Decibel: Unit of level equal to one-tenth of a bel. Symbol, dB.

Discrete Tone: A discrete tone can be either a sinusoidal variation (see **Pure Tone**), in which case the frequency spectrum would show a single component at the sinusoidal frequency, or, more typically, a nonsinusoidal variation, in which case the spectrum would show a component at the fundamental frequency and other components at harmonics of the fundamental.

Filter Nominal Midband Frequency: For a set of contiguous one-third octave bandpass filters, frequency of a specified series such as the preferred frequency series that includes 100, 125, 160, 200, 250, 315, 400, 500, 630, 800, and 1000 Hz, extended by successive multiplication or division by 10. For a set of contiguous octave bandpass filters, the nominal midband frequency is one of a series such as the preferred frequency series that includes 16, 31.5, 63, 125, 250, 500, 1000, 2000, 4000, and 8000 Hz, extended by successive multiplication or division by 10. Unit, hertz (Hz).

Hemi-anechoic Room: Test room with a hard, reflecting floor whose other surfaces absorb essentially all the incident sound energy over the frequency range of interest, thereby affording nominal free-field conditions above a reflecting plane.

Integrating–Averaging Sound-Level Meter: Device for measuring the level of the ratio of a mean square frequency-weighted sound pressure during a stated time interval to the square of the reference sound pressure.

Level: In acoustics, logarithm of the ratio of a quantity to a reference quantity of the same kind. The base of the logarithm (always 10, unless otherwise specified), the reference quantity, and the *kind* of level shall be explicitly specified. Examples of kinds of levels are sound-pressure level, sound-power level, and sound-intensity level. For common logarithms on the base 10, the symbol for logarithm, for the purposes of this article, is "log."

Noise: Undesired sound.

Noise Emission: Airborne sound radiated by a well-defined noise source (e.g., a machine under test).

Noise Immission: Airborne sound received at the ear of an observer, being a composite of all sounds in the vicinity of the observer.

Noise-Power Emission Level: A-weighted sound-power level. Unit, bel (B).

Particle Velocity: In a sound field, the velocity caused by a sound wave propagating through a given infinitesimal part of the medium, with reference to the medium as a whole. Unit, meter per second (m/s); symbol, u.

Porous Sound Absorber: Material with interconnected voids that presents resistance to airflow through the material.

Pure Tone: (a) Line spectrum consisting of a signal at a single frequency. (b) Sound wave, the instantaneous sound pressure of which is a simple sinusoidal function of time. Also referred to as a **Discrete Tone.**

Reverberation Room: Room having a long reverberation time, specially designed to make the sound field therein as diffuse as possible.

Sabine Absorption: Sound absorption defined by the Sabine reverberation-time equation. Unit, metric sabin. In a reverberant room of volume V, speed of sound c, and decay rate d, Sabine absorption is $A = 0.921Vd/c$.

Sound: (a) Oscillation in pressure, stress, particle displacement, particle velocity, etc., in a medium with internal forces (e.g., elastic or viscous), or the superposition of such propagated oscillations. (b) Auditory sensation evoked by the oscillation described above.

Sound Exposure Level: Ten times the logarithm of the ratio of a given time integral of squared instantaneous A-weighted sound pressure, over a stated time interval or event, to the reference sound pressure determined from the product of the squared reference sound pressure of 20 μPa and reference duration of 1 s. The frequency weighting may be otherwise if stated explicitly. Unit, decibel (dB); symbol, L_{AE}.

Sound Intensity: Average rate of sound energy transmitted in a specified direction at a point through a unit area normal to this direction at the point considered. Unit, watt per square meter (W/m^2); symbol, I.

Sound-Level Meter: Device used to measure sound-pressure level with a standardized frequency weighting and indicated exponential time weighting for measurements of sound-pressure level. Without time weighting, measurement of time-average sound-pressure level is achieved by use of an integrating–averaging sound-level meter. Sound exposure level is determined using an integrating sound-level meter.

Sound Power: Sound energy radiated by a source per unit of time. Unit, watt (W); symbol, W.

Sound-Power Level: Logarithm of the ratio of a given sound power in a stated frequency band or with a stated frequency weighting, to the reference power of one picowatt (1 pW). Unit, bel (B) or decibel (dB); symbol, L_W.

Sound-Pressure Level: Ten times the logarithm of the ratio of the mean square pressure of a sound, in a stated frequency band or with a stated frequency weighting, to the square of the reference sound pressure of 20 μPa. Unit, decibel (dB); symbol, L_p.

Spectrum: (a) Description, for a function of time, of the resolution of a signal into components, each of different frequency and (usually) different amplitude and phase. (b) "Spectrum" is also used to signify a continuous range of components, usually wide in extent, within which waves have some specified common characteristic, e.g., "audio-frequency spectrum."

Spectrum Analyzer: Device used to determine the frequency spectrum of a sound.

Speed of Sound: At a temperature of 15 °C, the speed of sound in air at sea level is approximately 341 m/s and is proportional to the square root of absolute temperature; symbol, c.

Time-Average Equivalent Continuous *A*-Weighted Sound-Pressure Level: Ten times the logarithm of the ratio of the *A*-weighted mean square sound pressure to the square of the reference sound pressure. The duration of the integration time must be explicitly stated. Unit, decibel (dB); symbol, L_{AT} or L_{AeqT}.

Works Cited

Anon. (1995), *Noise/News International* **3,** 161–168.

ANSI (1994), Standard S1.1, New York: Acoustical Society of America.

ANSI (1995a), Standard S12.2, New York: Acoustical Society of America.

ANSI (1995b), Standard S1.26, New York: Acoustical Society of America.

Bartholomae, R. C., Stein, R. R. (1991), in: *Proceedings of NOISE-CON 91,* 239–244, Poughkeepsie, NY: Noise Control Foundation.

Beranek, L. L., Vér, I. (Eds.) (1992), *Noise and Vibration Control Engineering: Principles and Applications,* New York: Wiley.

Berglund, B., Lindvall, T. (Eds.) (1995), *Community Noise,* Stockholm: Stockholm University and Karolinska Institute.

Blazier, W. E. Jr. (1995), *Noise Control Eng. J.* **43,** 53–63.

Bolt, R. H., Ingard, K. U. (1957), in: C. Harris (Ed.), *Handbook of Noise Control,* 1st ed., Chap. 22, New York: McGraw-Hill.

Bowlby, W., Higgins, J., Reagen, J. (Eds.) (1982), Report FHWA-DP-58-1, Washington, DC: Federal Highway Administration.

Cops, A., Vermeir, G. (1995), *Noise/News International* **3,** 10–25.

ECMA (1985), Report No. TR27, Switzerland: European Computer Manufacturers Association (ECMA).

EPA (1974), EPA 550/9-74-004, Washington, DC: U.S. Environmental Protection Agency.

Eldred, K. M. (1992), in: *Proceedings of INTER-NOISE 92,* 3–12, Poughkeepsie, New York: Noise Control Foundation.

Eldred, K. M., von Gierke, H. E. (1993), *Noise/News International* **1,** 67–89.

Elliott, S. J., Nelson, P. A. (1994), *Noise/News International* **2,** 75–98.

Embleton, T. F. W. (1982), *Noise Control Eng. J.* **18,** 30–39.

Embleton, T. F. W. (1989), in: *Proceedings of the Institute of Acoustics,* 11, 5, 1–16: United Kingdom.

Embleton, T. F. W. (1994), *Noise/News International* **2,** 227–237.

Embleton, T. F. W., Daigle, G. A. (1991) in: H. H. Hubbard (Ed.), *Aeroacoustics of Flight Vehicles,* NASA Reference Publication No. 1258, Chap. 12, Hampton, VA: National Aeronautics and Space Administration (republished, 1995, New York: Acoustical Society of America).

FAA (1996), *Integrated Noise Model, Version 5.1,* Washington, DC: Federal Aviation Administration.

FHWA (1976), 23 CFR, Chap. 1, Part 772, Washington, DC: See also *Federal Register,* 23 April 1976.

Finegold, L. S., Harris, C. S., von Gierke, H. E. (1994), *Noise Control Eng. J.* **42,** 25–30.

Harris, C. M. (Ed.) (1991), *Handbook of Acoustical Measurements and Noise Control,* 3rd ed., Chap. 3, New York: McGraw-Hill.

Higginson, R. F., Jacques, J., Lang, W. W. (1994), *Noise/News International* **2,** 156–184.

Hodgson, M. R. (1986), in: *Proceedings of INTER-NOISE 86,* 2, 1319–1322, Poughkeepsie, New York: Noise Control Foundation.

Hubbard, H. H. (Ed.) (1991), *Aeroacoustics of Flight Vehicles,* NASA Reference Publication No. 1258, Vols. 1 and 2, Hampton, VA: National Aeronautics and Space Administration.

ICAO (1993), *International Standards and Recommended Practices, Environmental Protection,* Volume 1: *Aircraft Noise,* Annex 16 to the Convention on International Civil Aviation, 3rd ed., Montréal: International Civil Aviation Organization.

IEC (1979), Standard No. 651, Switzerland: International Electrotechnical Commission.

IEC (1985), Standard No. 804, Switzerland: International Electrotechnical Commission.

Ingard, K. U. (1953), in: *Proceedings of the Fourth Annual National Noise Abatement Symposium,* 4, 1–15.

Ingard, K. U. (1994), *Sound Absorption Technology,* New York: Noise Control Foundation.

ISO (1994), Standard No. 362:1994(E), Geneva, Switzerland: International Organization for Standardization.

ISO (1980), Standard No. 3740, Geneva, Switzerland: International Organization for Standardization.

ISO (1990), Standard No. 1999, Geneva, Switzerland: International Organization for Standardization.

ISO (1993), Standard No. 9614-1, Geneva, Switzerland: International Organization for Standardization.

ISO (1996), Standard No. 9614-2, Geneva, Switzerland: International Organization for Standardization.

Kingsbury, H. F. (1995), *Noise Control Eng. J.* **43,** 64–72.

Lighthill, M. J. (1954), in: *Proceedings of the Royal Society of London, Ser. A* **211A,** 564–587; **222A,** 1–32.

Lueg, P. (1934), U.S. Patent No. 2,043,416.

Marsh, A. H. (1994), in: *Proceedings of INTER-*

NOISE 94, Vol. 1, Poughkeepsie, New York: Noise Control Foundation, pp. 199–204.

Menge, C. (1996), in: *Proceedings of INTER-NOISE 96,* Vol. 6, St. Albans, Herts, U.K.: Institute of Acoustics, pp. 3121–3126.

Neise, W., Koopman, G. H. (1988), in: *Proceedings of INTER-NOISE 88,* Vol. 2, Poughkeepsie, New York: Noise Control Foundation, pp. 801–804.

Nelson, P. A., Elliott, S. J. (1992), *Active Control of Sound,* New York: Academic.

Nelson, P. M. (1987), *Transportation Noise Reference Book,* London: Butterworths.

Nitsche, V. (1995), in: *Proceedings of INTER-NOISE 95,* Vol. 2, Poughkeepsie, New York: Noise Control Foundation, pp. 755–762.

Olson, H. F., May, E. G. (1953), *J. Acoust. Soc. Am.* 1130–1136.

Peterson, D. K., Weeks, W. A., Nowlin, W. C. (1994), in: *Proceedings of NOISE-CON 94,* 315–320, Poughkeepsie, New York: Noise Control Foundation.

Sandberg, U. (1995), *Noise/News International* **3,** 82–113.

Sommerfeldt, S., Hamada, H. (Eds.) (1995), in: *Proceedings of ACTIVE 95, The 1995 International Symposium on Active Control of Sound and Vibration,* Poughkeepsie, New York: Noise Control Foundation.

Suter, A. H. (1993), *Noise/News International* **1,** 131–151.

Sutton, T. J., Elliott, S. J., McDonald, A. M., Saunders, T. J. (1994), *Noise Control Eng. J.* **42,** 137–147.

Tichy, J. (1995), in: *Proceedings of INTER-NOISE 95,* 1, 17–28, Poughkeepsie, New York: Noise Control Foundation.

Timmerman, N. S. (1993), in: *Proceedings of NOISE-CON 93,* 425–428, Poughkeepsie, New York: Noise Control Foundation.

Further Reading

Anon., Home Page of the Institute of Noise Control Engineering, http://users.aol.com/inceusa/ince.html.

Bell, L. H., Bell, D. H. (1994), *Industrial Noise Control,* New York: Marcel Dekker.

Beranek, L. L. (Ed.) (1988), *Noise and Vibration Control,* revised ed., Poughkeepsie, New York: Institute of Noise Control Engineering.

Beranek, L. L., Vér, I. (Eds.) (1992), *Noise and Vibration Control Engineering: Principles and Applications,* New York: Wiley.

Harris, C. M. (Ed.) (1991), *Handbook of Acoustical Measurements and Noise Control,* 3rd ed., New York: McGraw-Hill.

Noise Control Engineering Journal and *Noise/News International,* published by the Institute of Noise Control Engineering, P.O. Box 3206 Arlington Branch, Poughkeepsie, NY 12603, USA.

SONOLUMINESCENCE

Chandra M. Sehgal, *University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.*

	Introduction	1
1.	**Historical Perspective**	1
2.	**Characteristics of Cavitation Bubbles**	2
2.1	Single Isolated Bubble	3
2.1.1	Bubble Motion	3
2.1.2	Hot-Spot Theory	6
2.1.2.1	Adiabatic Bubble Collapse	6
2.1.2.2	Heat Loss by the Collapsing Bubbles	7
2.1.2.3	Hot Spot and Chemiluminescence	7
2.2	Cavitation Field (Multiple Bubbles)	8
3.	**Experiments on Sonoluminescence**	8
3.1	Single-Bubble Sonoluminescence	8
3.2	Multiple-Bubble Sonoluminescence	12
3.2.1	MBSL from Aqueous Media	13
3.2.2	MBSL from Nonaqueous Media	15
3.2.3	MBSL as a Probe of Acoustic Cavitation	16
4.	**New Theoretical Models**	17
5.	**Controversies on the Origin of SL**	19
	Acknowledgments	21
	Glossary	21
	Works Cited	21
	Further Reading	23

INTRODUCTION

Sonoluminescence (SL) is the phenomenon of light emission that occurs when high-intensity ultrasound is transmitted in liquids containing dissolved gases. It is a third-order process mediated by acoustic cavitation (Fig. 1).

In the presence of ultrasound the minute inhomogeneities in the bulk of the material grow into microscopic bubbles (Plesset and Prosperetti, 1977). These bubbles oscillate radially and nonlinearly. In some cases the wall motion becomes unstable and the bubbles collapse. It takes bubbles longer to expand than to collapse. Thus bubbles acquire energy over a relatively long period and release it in a very short interval. Through this process the bubbles are able to concentrate acoustic energy in small local volume elements by as much as 12 orders of magnitude (Barber *et al.*, 1992). The unleashing of intense energy by the bubbles generates short flashes of light, called sonoluminescence.

1. HISTORICAL PERSPECTIVE

The phenomenon of sonoluminescence was first observed in the 1930s (Frenzel and Schultes, 1935; Zimakov, 1934). Since then several researchers have investigated this phenomenon; however, the activity has been sporadic. Early work has been covered extensively in several review articles (El'piner, 1964; Flynn, 1964; Margulis, 1969; Verrall and Sehgal, 1988; Walton and Reynolds, 1984) and is not repeated here.

The periodic nature of SL was first noted in 1950–1960 when it was demonstrated that the light emission occurs during the last phases of bubble collapse (Jarman, 1960; Negishi, 1960). The periodicity of SL was further studied under better controlled experimental conditions by Saksena and Nyborg (1970). The SL was noted to be from stable cavitation with flashes repeating at the frequency of the acoustic wave. Prior to this study the emission was believed to occur only during the implosion of transient cavities. In the 1970s the interest in SL and

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

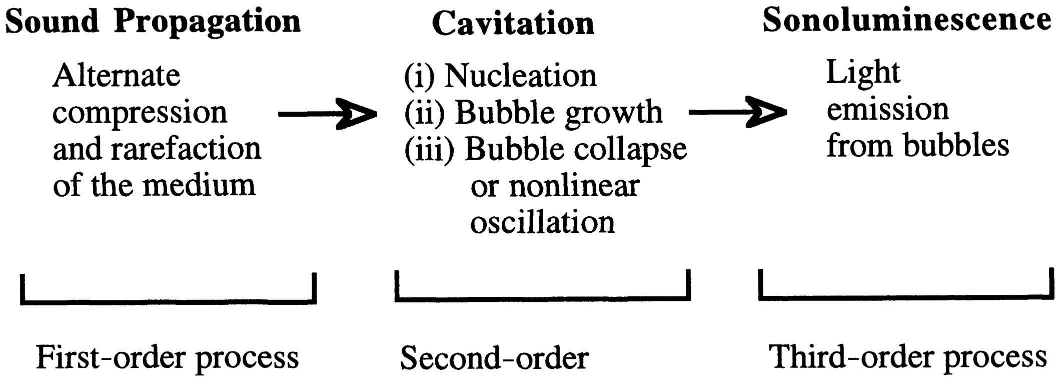

FIG. 1. The relationship between sound propagation, cavitation, and SL.

sonochemistry was revived by Verrall (Mead *et al.*, 1975; Sehgal *et al.*, 1977) in North America and by Margulis (1969, 1978) in Russia. These studies elucidated the free-radical chemistry underlying the phenomenon of cavitation and SL in aqueous media. This generated a widespread interest in the field of sonochemistry. Vigorous and successful research led by Suslick and Boudjouk in the United States, Mason in England, Luche in France, and Heinglein in Germany developed the field of sonochemistry significantly. However, the activity in SL lagged behind. After a hiatus of several years, there is currently an unprecedented activity in this area.

Suslick and Flint (1987) reported detailed spectra of nonaqueous liquids with bands characteristic of the chemical species formed by the breaking of chemical bonds during cavitation. One problem in studying SL has been that the observed light emission is generally from multiple bubbles in a cavitation field. The variable emission from multiple sources makes the analysis of SL difficult. This difficulty was overcome relatively recently by Gaitan *et al.* (1992) in Crum's group by trapping a single bubble in a resonance field. This progress was further exploited in greater detail by Putterman and co-workers (Barber *et al.*, 1992; Barber and Putterman, 1991). They used a fast photomultiplier tube to demonstrate the lifetime of SL pulses to be shorter than 50 ps. These flashes occur once every acoustic cycle with a time jitter of less than 50 ps. These time periods are much smaller than expected on the basis of earlier theories and have generated considerable excitement in the scientific community. The SL spectrum from a single bubble is virtually structureless and is different from the emission from the bubbles in a cavitating field. At this writing, SL from a single bubble and that from multiple bubbles in a cavitation field are evolving as two distinct fields, and it is worth considering them separately.

2. CHARACTERISTICS OF CAVITATION BUBBLES

There are several descriptive models for cavitation based on bubble-wall motion and bubble composition (Flynn, 1964; Neppiras, 1980; Plesset and Prosperetti, 1977). Stable cavitation consists of bubbles that oscillate permanently and often nonlinearly about an equilibrium size over many acoustic cycles. On the other hand, transient cavitation is comprised of bubbles that collapse violently with wall motion approaching or exceeding the speed of sound. They generally last for less than one cycle and release significant energy, often leading to chemical reaction, SL, and the disintegration of the bubbles into smaller cavities. As a result of the violent nature of the bubble collapse, transient cavitation is accompanied by streamer activity, surface waves, and broadband acoustic noise. Stable cavitation, on the other hand, is accompanied by surface waves on bubbles and microstreaming in the surrounding liquid. Either type of bubble may contain dissolved permanent gases like argon or air, vapor of the host liquid, or some combination of the two. Also, the bubbles may exist as single, isolated entities, or coexist with multiple other bubbles and behave as a bubble cloud. In an uncontrolled condition of insonification, a variety of bubbles are formed with variable radial motions and cavity compositions. Under such conditions there is no single, unique model to describe the behavior of all the bubbles.

Furthermore, there is a special category of single bubbles that do not disappear after collapse but maintain their identity and repeat the expansion–contraction cycle indefinitely at the frequency of a continuous sound wave. These bubbles have recently been shown to exhibit sonoluminescence (Barber and Putterman, 1991; Gaitan *et al.*, 1992), an effect associated with transient cavitation, yet they repeat their motion like stable cavitation. In view of this development, the terms "inertial" and "noninertial" have been proposed by the Acoustical Society of America (ASA), the National Council on Radiation Protection and Measurements (NCRP), and the World Federation of Ultrasound in Medicine and Biology (WFUMB) to describe the nature of cavitation. In inertial cavitation the contraction of bubbles is such that the motion approximates that in a Rayleigh cavity. The bubble contracts rapidly to a small fraction of its original volume (equivalent to transient or collapse cavitation), and the contraction is accompanied by the production of shock waves, emission of SL, and the formation of highly reactive chemical species, including free radicals. Noninertial cavitation is "any form of acoustic cavitation that is not inertial. It includes acoustically induced translational motion of bubbles, bubble growth by rectified diffusion or coalescence, continuous heat generation, radiation forces on neighboring particles and microstreaming" (ASA). To date, all the published work has been reported using the terms "stable" and "transient" cavitation. The above discussion provides a perspective on how the terms inertial and noninertial are currently been considered. The language will, undoubtedly, be further developed with use and with the completion of the various reports under consideration by the various organizations.

2.1 Single, Isolated Bubble

2.1.1 Bubble Motion Although SL is most commonly observed from a collection or a cloud of bubbles, the general trend is to explain these observations by the events that occur in single, isolated bubbles modeled as stable or transient cavities. By use of the laws of conservation of mass and momentum, it can be shown that the equation of motion of an isolated cavity of radius R enclosed in an incompressible liquid medium of density ρ is described (Rozenberg, 1971) as

$$R\ddot{R} + \tfrac{3}{2}\dot{R}^2 = (1/\rho)[p(R) - p_\infty], \tag{1}$$

where $p(R)$ and p_∞ are the pressures at the cavity surface and at infinity in the liquid medium, respectively. In a Rayleigh cavity, $p(R) = 0$ and p_∞ is the ambient pressure p_H, and Eq. (1) is mathematically solvable for the cavity-wall velocity U as a function of R, which is described by

$$U^2 = \tfrac{2}{3}\frac{p_{\mathrm{H}}}{\rho}\left(\frac{R_0^3}{R^3} - 1\right), \tag{2}$$

where R_0 is the initial radius. In Eq. (2) there is no constraint on the value of R. As $R \to 0$, $U \to \infty$. This discrepancy arises because the cavity is a void by virtue of the condition $p(R) = 0$. In a cavity containing gas, $p(R)$ is equal to the gas pressure p_g. Taking into consideration the effect of surface tension and viscous damping on a cavity (or a bubble) driven by a sinusoidal acoustic wave of amplitude p_A and angular frequency ω yields

$$p(R) = \left(p_H + \frac{2\sigma}{R_0}\right)\left(\frac{R_0}{R}\right)^{3b} - \frac{2\sigma}{R} - \frac{4\mu\dot{R}}{R} \tag{3}$$

and

$$p_\infty = p_H - p_A \sin \omega t. \tag{4}$$

In the above equations ρ, σ, and μ are the density, surface tension, and viscosity of the medium, respectively; p_H is the hydrostatic pressure; and R_0 is the initial radius of the bubble. The first factor in parentheses of Eq. (3) is the gas pressure p_g for an ideal gas with polytropic index b. For an isothermal process $b = 1$, and it is equal to the specific-heat ratio γ for the adiabatic process. Combining (1), (3), and (4) gives the Rayleigh–Plesset (R-P) equation,

$$R\ddot{R} + \tfrac{3}{2}\dot{R}^2 = \frac{1}{\rho}\left[\left(p_H + \frac{2\sigma}{R_0}\right)\left(\frac{R_0}{R}\right)^{3b} - \frac{2\sigma}{R} - \frac{4\mu\dot{R}}{R} - p_H + p_A \sin \omega t\right]. \qquad (5)$$

Equation (5) was first derived by Noltingk and Neppiras (NN) for isothermal gaseous bubbles (Noltingk, 1950). The radius–time curves computed using (5) show that a bubble increases in size to a maximum and thereafter shrinks with increasing rapidity until the inward motion is stopped by the "cushioning" action of the compressed gas. Since the NN model was first proposed, several elegant mathematical models have been developed that relax one or more assumptions (Flynn, 1975a, 1975b; Plesset and Prosperetti, 1977). However, the NN model continues to provide useful physical insight and is believed to perform reasonably well up to the point at which the radial velocity of the bubble collapse approaches the speed of sound. Figure 2 shows the calculations based on Eq. (5). A bubble of radius near, but below, resonance when driven at increasing p_A/p_H undergoes a change in the nature of cavitation. At low pressure amplitude ($p_A/p_H = 0.5$), the bubble is stable and expands and compresses once every acoustic cycle. The stability in the bubble oscillation is lost as p_A is increased. For example, a change in the driving amplitude from $p_A/p_H = 0.5$ to 2 changes the character of the bubble. The bubble becomes increasingly larger and collapses during the first cycle (see the curves corresponding to $p_A/p_H = 2$ and $p_A/p_H = 4$ in Fig. 2). The bubble no longer vibrates about an equilibrium size but exists only as a transient cavity and collapses violently. A further increase in driving amplitude makes the bubble even larger, and the collapse occurs during the second, third, and subsequent cycles.

One of the limitations of the NN model is that in the final stages of bubble collapse, the radial velocity exceeds that of sound, and the assumption that the liquids are incompressible becomes inadequate. Several studies have addressed the problem of supersonic wall velocity during bubble collapse by treating the liquid as a compressible medium (Flynn, 1975b; Gilmore, 1952; Prosperetti, 1984; Rozenberg, 1971). Gilmore (1952) considered enthalpy,

$$H\left(= \int_{p_\infty}^{p(R)} \frac{dp}{\rho}\right),$$

to be a fundamental quantity to derive the equation of motion. He also used the Kirkwood–Bethe approximation (Kirkwood and Bethe, 1942), which states that pressure disturbances propagate with a velocity equal to the sum of the acoustic velocity and the local fluid velocity. This formalism was developed further by Flynn (1975b) and came to be

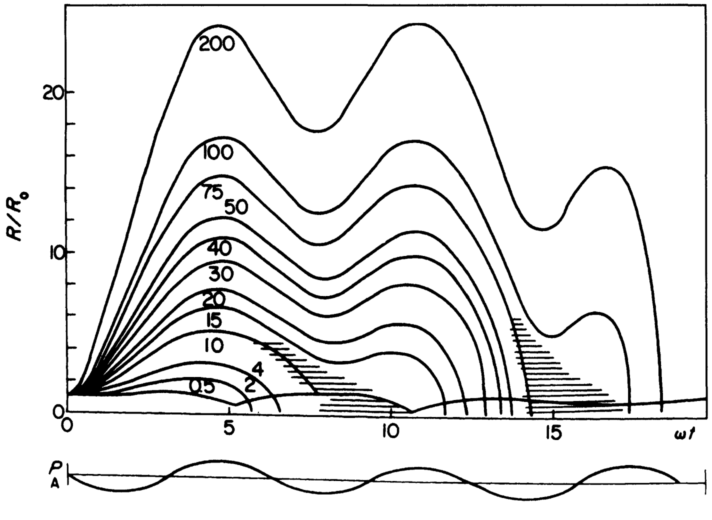

FIG. 2. Radius–time relationships of bubbles in water determined from Eq. (1) for increasing driving acoustic pressure. Initial bubble size $R_0 = 3 \times 10^{-4}$ cm; angular frequency $\omega = 4.2 \times 10^{-6}$ s^{-1}; ambient pressure $P_0 = 1$ bar. The numbers on the figure refer to the ratio P_A/P_0. The hatched area represents an unstable region in the radius–time curves and the generation of subharmonics by the bubbles. (Reproduced with permission from Neppiras, 1980.)

known as the Gilmore–Flynn equation. Prosperetti used first-order compressibility terms and the procedures proposed by Keller and Miksis to derive the equation of motion for the bubbles in compressible fluids,

$$\left(1 + \frac{\dot{R}}{c}\right)R\ddot{R} + \tfrac{3}{2}\left(1 + \frac{\dot{R}}{3c}\right)\dot{R}^2 = \left(1 + \frac{\dot{R}}{c}\right)\frac{1}{\rho} \times [p(R,t) - p_\infty(t + R/c)] + \frac{R}{\rho c}\frac{dp(R,t)}{dt}, \tag{6}$$

where

$$p(R,t) = \left(p_H + \frac{2\sigma}{R_0}\right)\left(\frac{R_0}{R}\right)^{3b} - \frac{2\sigma}{R} - \frac{4\mu\dot{R}}{R} \tag{7}$$

and $p_\infty(t + R/c)$ is the time-delayed pressure. An improvement over this model was obtained by using more accurate expressions for internal pressure than is described in (7) (Prosperetti *et al.*, 1986). Although the formulation is involved, its use provides an insight that is relevant to sonoluminescence from single bubbles (Gaitan *et al.*, 1992). Figure 3 shows the radius–time curve of a 20-μm bubble driven sinusoidally at the pressure amplitude of 1.2 bars. The important feature of this plot is that the bubble does not disappear after rapid compression or "collapse" but repeats the cycle of compression and expansion indefinitely. A similar radius–time curve showing repeated expansion–compression cycles after the bubble collapse was obtained by Lofsted *et al.* (1993) using the van der Waals equation of state with the Rayleigh–Plesset equation.

The models discussed so far assume thermodynamic equilibrium during the entire expansion–compression cycle; isentropic heating of the gases within the bubbles; and a simple equation of state to describe the gases, viz., ideal-gas equation or van der Waals equation. Moss *et al.* (1994) noted that these assumptions are inadequate and do not provide a realistic description of heating during the collapse of a bubble. They proposed the use of numerical techniques to solve the equations of conservation of mass, momentum, and energy for the bubble–liquid region. In the absence of viscosity, surface tension, heat conduction, mass diffusion, and nonmechanical energy loss, these equations are

$$\frac{D\rho}{Dt} + \rho\nabla\cdot\mathbf{v} = 0, \tag{8}$$

$$\rho\frac{D\mathbf{v}}{Dt} = -\nabla P, \tag{9}$$

$$\frac{D\epsilon}{Dt} + P\frac{DV}{Dt} = 0. \tag{10}$$

The operator D/Dt in the above equations is defined as $(\partial/\partial t) + \mathbf{v}\cdot\nabla$. The other variables—$\rho$, $\mathbf{v}$, $P(\rho,\epsilon)$, ϵ, and V—define density, velocity, pressure, specific internal energy, and specific volume, respectively. The equation of state can be represented as

$$p_g = RTF(\rho), \tag{11}$$

where R and T represent the gas constant

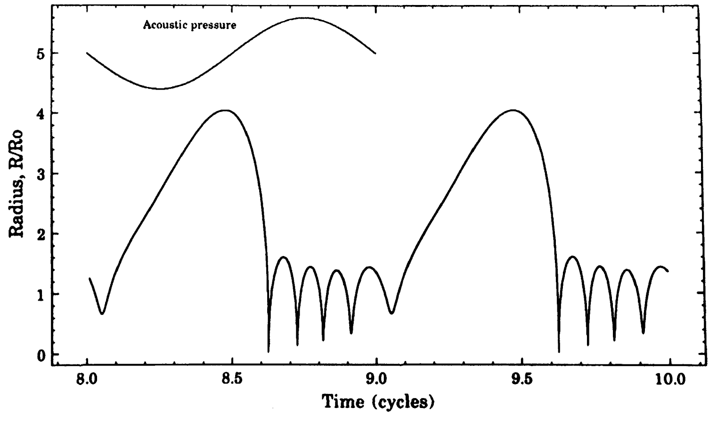

FIG. 3. Calculated radius–time curves illustrating repetitive collapse and rebound of a single spherical bubble in water. (Reproduced with permission from Gaitan *et al.*, 1992.)

and absolute temperature, respectively, and $F(\rho)$ is a function of density. $F(\rho)$ is a simple function for an ideal or a van der Waals gas, but it has a more complex form for a medium undergoing ionization, dissociation, and the like. The equations of motion (7)–(10) and the equation of state (11) constitute the complete set of equations necessary to obtain spatial and temporal information on the field quantities for known boundary conditions. Moss *et al.* (1994) represented the bubble–liquid region by a mesh of concentric shells of fixed mass. The equations of motion were solved numerically using an equation of state that took into consideration the vibrational excitation, dissociation, ionization, and intermolecular repulsion of the air molecules. All these processes become important during the last stages of bubble collapse. The results of this simulation show that the collapse is nearly isentropic until the final few nanoseconds when a shock is generated near the center of the bubble on account of a sudden "stiffening" of the equation of state. The spherical shock converges and rapidly raises the temperature in the center of the bubble. It is estimated that the center (on the order of 0.02 μm) is heated to a temperature greater than 10 eV (1 eV = 11 605 K), which, in turn, radiates light. The pressure–volume cooling occurs subsequent to the compression phase. This rapidly quenches the light emission on a timescale of a few picoseconds, thereby making the SL pulses very short, as has been observed experimentally (Barber and Putterman, 1991).

2.1.2 Hot-Spot Theory The question of how the energy is released by cavitating bubbles has been a topic of debate among researchers. Several theories have been proposed. By far the most widely accepted explanation for sonoluminescence is based on the formation of "hot spots" within the bubbles as they collapse adiabatically.

2.1.2.1 Adiabatic Bubble Collapse. According to the radius–time curves in Fig. 2, the bubble size increases gradually over a relatively long period, whereas the collapse is rapid. That is, a bubble grows isothermally but collapses nearly adiabatically. The local temperature within the bubble during adiabatic collapse increases to a very high value and forms a hot spot from which SL is emitted. The extent of collapse, R_{max}/R_{min}, where R_{max} and R_{min} are the radii at which the collapse begins and ceases, respectively, is related to the ambient pressure P_∞ and the gas pressure p_g inside the cavity (Noltingk, 1950) by

$$(R_{max}/R_{min})^{3(\gamma-1)} \cong P_\infty(\gamma - 1)/p_g, \qquad (12)$$

where γ is the specific-heat ratio. The maximum temperature T_{max} achieved during the adiabatic compression of spherical bubbles is

$$T_{max} = T_0(R_{max}/R_{min})^{3(\gamma-1)}, \qquad (13)$$

where T_0 is the intracavity temperature when the collapse begins. From Eqs. (12) and (13) it follows that

$$T_{max} = T_0 P_\infty(\gamma - 1)/p_g. \qquad (14)$$

Similarly, the maximum pressure P_{max} achieved during a bubble collapse is (Noltingk, 1950)

$$P_{max} = p_g[P_\infty(\gamma - 1)/p_g]^{\gamma/\gamma-1}. \qquad (15)$$

The T_{max} and P_{max} estimated by (14) and (15), although approximate, are useful in providing insight into many experimental observations. The maximum temperature and pressure achieved in a bubble containing air are on the order of 10^4 K and 10^4 atm, respectively. These estimates vary considerably with the nature of the dissolved gas, nature of the host liquid, acoustic pressure amplitude, hydrostatic pressure, and the temperature of the liquid. Nonetheless, these temperatures are large enough to heat the cavity contents to incandescence or to initiate chemiluminescence. Also, the extreme temperature and pressure conditions depend not only on the properties of the medium (viz., p_g and γ) but also on the acoustic pressure amplitude. For example, bubbles containing volatile vapor can yield a higher temperature and pressure (as compared to bubbles containing vapor of a less-volatile liquid), if driven at a higher acoustic pressure. The implication is that the cavitation temperatures are not uniquely dependent on the material properties, but they also vary with the sonification conditions. This aspect has often been overlooked in interpreting the many studies in the literature that have attempted to compare intracavity temperature and pressure from one study with those of another. Small variations in the acoustic conditions are sufficient to explain these differences.

A variant of the adiabatic collapse model is the shock-wave model proposed by Jarman (1960) who postulated that the heating of the cavity contents is due to the shock waves formed during the bubble collapse. A shock wave undergoing multiple reflections leads to the same average behavior as adiabatic heating (Trilling, 1952), and the two theories are in agreement with one another. Recently, there has been a resurgence of interest in the shock-wave theory to explain single-bubble SL (SBSL) (Barber *et al.*, 1994; Greenspan and Nadim, 1993; Lofstedt *et al.*, 1993). This is discussed later in the article.

2.1.2.2 Heat Loss by the Collapsing Bubbles. According to the adiabatic theory, bubbles containing gases with equal specific-heat ratios should yield the same final temperatures, all other factors being equal [see Eqs. (14) and (15)]. It may, therefore, be expected that all the noble gases ($\gamma = 5/3$) should yield similar SL. The measurements show that the emission from water in the presence of xenon is almost 100–500 times stronger than the emission in the presence of helium, even though the two gases have the same specific-heat ratio (Prudhomme and Guiltmart, 1957; Young, 1976). This disagreement between theory and experiment is believed to be due to the heat flow from the collapsing cavities to the surrounding liquid. Hickling (1963) took the heat loss by conduction into consideration in solving the equation of motion of gas bubbles numerically. His analysis showed that the temperatures achieved within collapsing bubbles are indeed lower than those predicted by the adiabatic model. The effect is more pronounced in gases with high thermal conductivity. A closed-form solution for the final intracavity temperature taking heat loss into consideration was obtained later (Young, 1976). Table 1 lists the final temperatures achieved in a collapsing bubble for a fixed sonification condition. Lower temperatures are achieved in gases with higher thermal conductivity, and there is a good correlation between SL intensity and the intracavity temperatures. That is, the bubble collapse is not a purely adiabatic process but is accompanied by considerable heat loss.

2.1.2.3 Hot Spot and Chemiluminescence. The hot-spot models discussed above treat cavitating bubbles as being primarily composed of gases that heat to incandescence. Griffing and Sette (1955) observed a quantitative relationship between intensity of SL and sonochemical yields. Using this relationship as the basis, they postulated SL to be due to the chemical reactions that occur within the hot spots. Since then a considerable amount of spectral data on SL has been obtained by many laboratories using aqueous (Becker *et al.*, 1992; Didenko and Pugach, 1994; Sehgal *et al.*, 1980a; Taylor and Jarman, 1970) and nonaqueous (Flint and Suslick, 1989) media. These studies lead to the view that SL is emission from the reactive chemical species formed, often in excited states, by the intense heating within or around the cavitating bubbles. It is implicit

Table 1. The relationship between specific-heat ratio, thermal conductivity, intracavity temperatures, and sonoluminescence intensity. The calculated intracavity temperatures are for the following conditions: ambient pressure = 1 atm, ambient temperature = 293 K, acoustic pressure amplitude = 6 atm, shell thickness is equal to three times the mean free path, bubble radius at the beginning of collapse = 10^{-4} cm, and radius at the end of collapse = 3.33×10^{-4} cm.

Gas	γ	Thermal conductivity[a] (10^{-6} cal cm^{-1} s^{-1} °C^{-1})	Intracavity temperature[b] (K)	SL[b] (rel. units)
Helium	1.67	109	815	1
Neon	1.67	39	1420	3
Argon	1.67	21	1650	30
Krypton	1.67	12	1890	50
Xenon	1.67		2000	125

[a] Hickling (1963).
[b] Young (1976).

in the hot-spot–chemiluminescence model that one is dealing primarily with the gas–vapor cavities as opposed to purely gaseous cavities. Another consequence of chemiluminescence is that a part of the energy in the collapsing bubbles is used up in breaking chemical bonds. This would imply that the actual cavitational temperature is lower than those predicted by the hot-spot models (Young, 1976).

2.2 Cavitation Field (Multiple Bubbles)

The most common technique for observing SL is by passing continuous ultrasonic waves through a liquid. In general, no attempt is made to control the primary and secondary reflections from the vessel walls or the liquid–air interfaces. Under such conditions the ultrasonic field is complex, the cavitating bubbles are formed at multiple sites, and SL is not a single-bubble event. Instead, the observed SL is a time-space average of the emission from an array or a cloud of bubbles that are distributed in space.

The behavior of multiple bubbles in a cavitation field differs from that of an isolated bubble in many respects. The cavitating bubbles generate pressure waves that affect the acoustic field perceived by the neighboring bubbles and thus influence their motion and behavior. For example, bubbles in close proximity to one another experience secondary Bjerknes forces. When the radial oscillations are in phase, the bubbles are attracted toward one another and often forced to coalesce. Crum (1975) has photographed bubbles in a standing wave and has observed coalescence of the bubbles due to attractive mutual Bjerknes forces.

Indirect evidence that a bubble cloud behaves differently than a single bubble follows from a recent study that shows the resonance frequency of a bubble cloud to be significantly below the resonance frequency of the individual bubbles (Kozhevnikova and Bjorno, 1992). The decrease is related to the gas content and the characteristic dimension of the bubble cloud.

The symmetry of collapse is another aspect in which the behavior of a single, isolated bubble may differ from that of multiple bubbles in a cavitating field. The latter are often formed at rigid boundaries, and the collapse is asymmetric. Such asymmetry dampens the bubble collapse, resulting in lower intracavity temperature. It may also lead to the formation of liquid jets that force the surrounding liquid into the bubble's interior (Crum, 1979). In turn, the change in bubble composition is likely to influence the bubble behavior and the characteristics of SL (Crum, 1994).

3. EXPERIMENTS ON SONOLUMINESCENCE

A wide variety of experimental techniques have been employed by various research groups to observe SL. The ultrasonic frequency and acoustic power used range from 220 Hz to 5 MHz and from 5 to 3000 W/cm^{-2}, respectively (Walton and Reynolds, 1984). The ultrasonic transducers are, generally, in contact with the liquid. There is a significant dissipation of acoustic energy at these power levels, causing liquids to heat. It is therefore necessary to cool the medium constantly during the experiments by using water jackets or cooling coils. A key variable often overlooked in experimental design and data interpretation is the nucleation of the cavitating bubbles. By far the most widely used technique for inducing cavitation has been to insonify a liquid medium containing pre-existing nuclei of entrapped gases or other inhomogeneities. A second approach, used less commonly, is to introduce bubbles of known gas at a specific site in the acoustic field (Barber and Putterman, 1991; Gaitan *et al.*, 1992; Saksena and Nyborg, 1970). The following discussion considers the two cases separately.

3.1 Single-Bubble Sonoluminescence

In the last five years there has been significant progress in studying sonoluminescence from single, isolated bubbles. This is largely due to the extensive work of Putterman and co-workers (Barber and Putterman, 1991) and Gaitan and co-workers (1992). These researchers have developed a technique that allows them to seed a single, isolated bubble with extreme precision in an acoustic resonance field. The bubble levitates in the standing wave and oscillates stably to generate pulses of SL once every acoustic cycle.

The periodicity of SL was first noted by

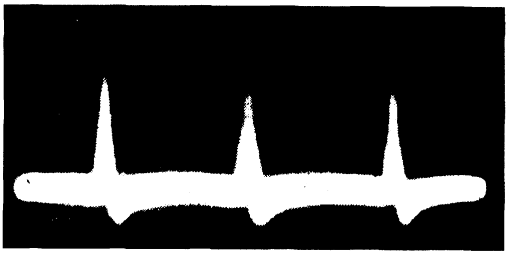

FIG. 4. Periodic SL from bubbles oscillating stably in a glycerol–water mixture. Time between flashes is 33 μs, which is equal to the period of sound wave used for insonification. The width of SL flashes is of the order of 5 μs. (From Saksena and Nyborg, 1970.)

Negishi (1960, 1961). He determined the bubble size by measuring the light scattered by the cavitating region. These measurements revealed the phase of SL to coincide with the minimum volume of the bubble. In this early work no special attempt was made to control the behavior of bubbles and the nature of the acoustic field. It seems that a standing wave was inadvertently set up, and the bubble oscillated stably to emit SL periodically (Walton and Reynolds, 1984). A more controlled study of SL from stable cavitation was reported much later (Saksena and Nyborg, 1970). Bubbles were seeded by forcing a stream of air through a steel capillary into a glycerol and water mixture. The bubbles were entrapped in the center of a vertical tube by a concentric transducer. Sonoluminescence flashes having a duration of 4–5 μs and periodicity of once every acoustic cycle were observed (Fig. 4). There was no broadband acoustic noise and streamer activity with the bubbles. This suggests that the SL was produced by stable pulsation of the bubbles. The light emission was attributed to the heating of the bubble above a certain threshold temperature with each volume compression.

It is important to note that none of the early studies were able to observe SL from isolated single bubbles. This has been possible only recently by seeding bubbles with greater precision and control using high resonance acoustic fields (Gaitan *et al.*, 1992; Putterman, 1995). Figure 5 shows the relationship between SL flashes, bubble radius of the pulsating bubbles, and the ultrasound wave. The SL pulses are much narrower than those reported earlier (Saksena and Nyborg, 1970). Barber *et al.* (1992) used the fastest available microchannel-plate photomultiplier tube to compare SL pulse width with that of a 34-ps laser pulse. Figure 6

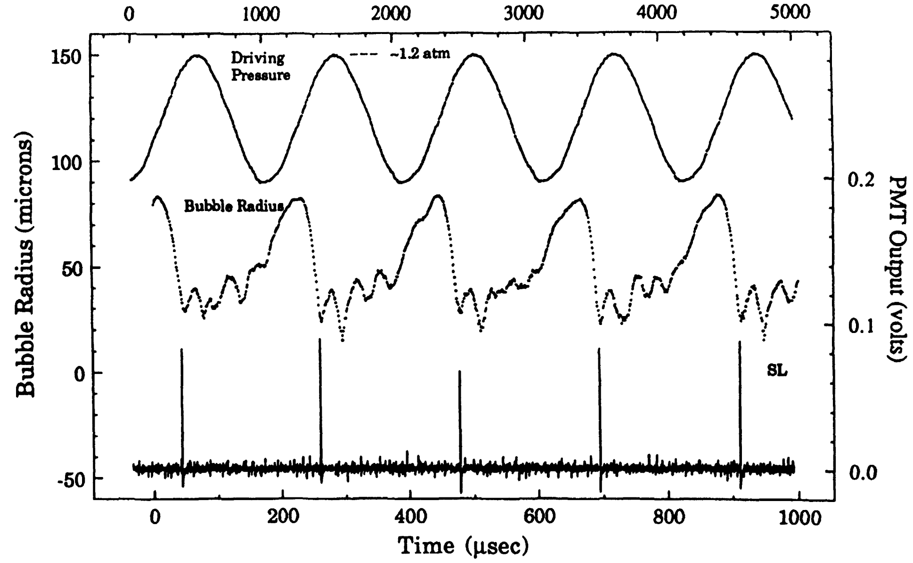

FIG. 5. Correlation between sound wave (top trace), bubble radius determined by light scattering (middle trace), and measured SL (bottom trace) from 21% (by weight) glycerin–water mixture. Ambient pressure = 1.2 atm and insonification frequency = 22.3 kHz. (Reproduced with permission from Gaitan *et al.*, 1992.)

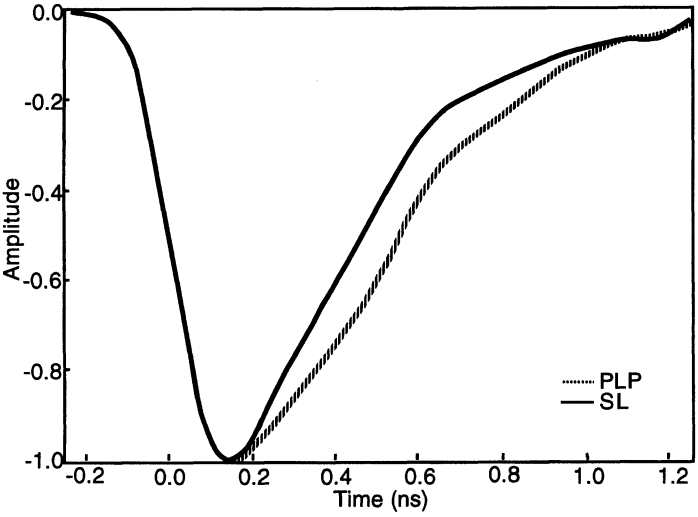

FIG. 6. Comparison of the photomultiplier-tube output for SL and 34-ps laser pulses. The laser operated at 410 nm, matching the blue color of the SL. (Adapted from Barber *et al.*, 1992.)

shows the characteristics of the two pulses. The rise times for the two are identical. The turnoff time of SL is shorter than that of the laser pulse. It is argued that if the rise time of the SL pulse were longer than that of the laser pulse by 4 ps, the SL pulse width would be less than 50 ps. Prior estimates of the lifetime of these flashes are often conflicting and considerably larger, ranging anywhere from 5 ns to 5 μs (Benkovskii *et al.*, 1974; Margulis, 1985; Margulis and Dmitrieva, 1982; Saksena and Nyborg, 1970). The histogram of the time difference between successive SL pulses fits a Gaussian curve with a standard deviation less than 50 ps (Fig. 7). This study for the first time provided direct evidence for the unusual characteristics of SL from stably oscillating bubbles. The observations that SL flashes from single bubbles are shorter than 50 ps and repeat with a jitter less than 50 ps has renewed interest in the phenomenon of SL. The ability to control the behavior of bubbles and measure SL from them with a precision that has not been possible before is likely to open new exciting avenues of research and applications. Nevertheless, there are some caveats that must also be considered.

The aging of water increases jitter and the

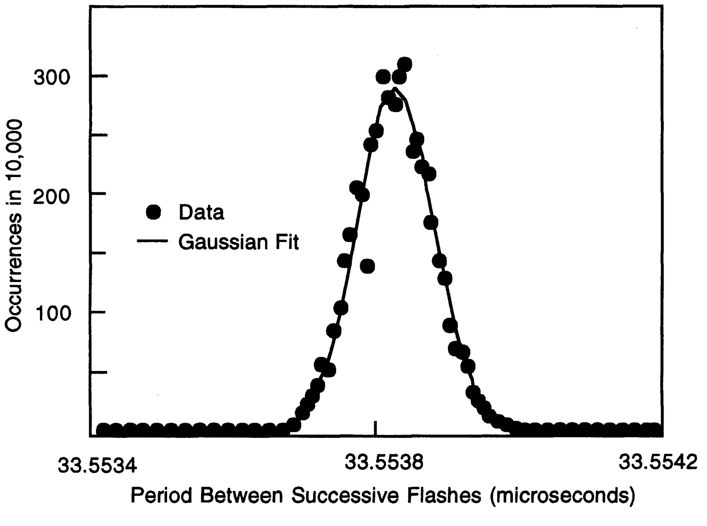

FIG. 7. Histogram of the period between SL flashes from a single, stable, cavitation bubble. Line represents best-fit Gaussian to the experimental data. The Gaussian has a standard deviation of 50 ps. (Adapted from Barber *et al.*, 1992.)

rise time (Barber *et al.,* 1992). The significance of this effect is not known. The acoustic pressure amplitude used to drive bubbles influences synchrony of emission. At high amplitudes the bubbles become unstable, and the precise phase relationship between SL and the instantaneous radius is frequently lost for a short interval before it is restored (Gaitan *et al.,* 1992). In a well-controlled study of single-bubble SL (SBSL), Holt *et al.* (1994) measured the variation in time between pulses by slowly detuning the resonance field at 27 kHz. Figure 8 shows the frequency of time intervals between SL flashes from a single bubble. As the ultrasonic field is gradually detuned from resonance by 0.01 kHz, the histogram bifurcates [Figs. 8(a) and 8(b)], and eventually becomes broadly distributed [Fig. 8(d)]. From a practical point of view, these results have an important implication as they suggest that SBSL is applicable to a small region of pressure–frequency–radius space, viz., the point in space corresponding to resonance. A small deviation from resonance, as may often be encountered in a generalized field, can almost completely eliminate the phenomenon of SBSL. Attempts to measure SL from low-viscosity liquids like silicone were initially unsuccessful (Barber *et al.,* 1994), whereas such emission is readily seen from bubbles in a cavitating field (Suslick *et al.,* 1990). However, when xenon gas was used, SBSL was observed from several nonaqueous liquids, including silicone oil (Weninger *et al.,* 1995). The emission spectra for various liquids were very similar and showed no transitions among the excited states of C_2 that have been observed in multiple-bubble SL (MBSL) from nonaqueous media (Flint and Suslick, 1989).

Simultaneous measurements of light scat-

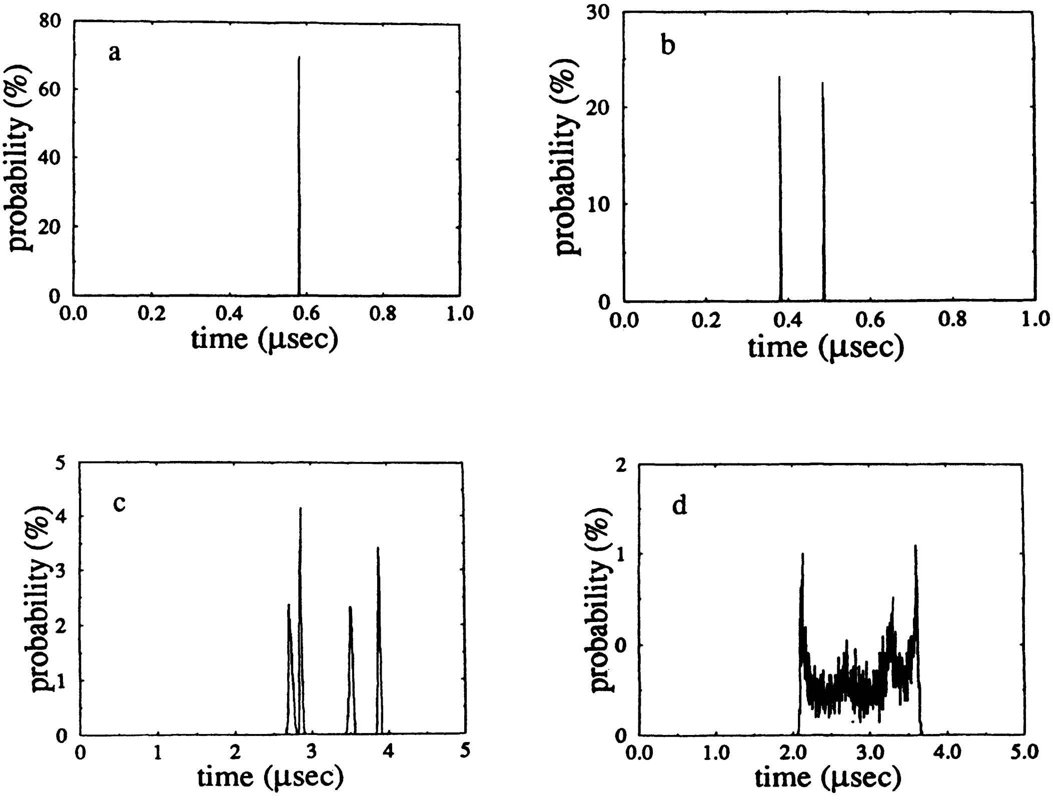

FIG. 8. Probability histogram of the period between SL flashes from a single bubble oscillating stably in an acoustic field. Panels with (a) one, (b) two, (c) four, and (d) broad distribution of peaks demonstrate the change in SL synchrony as the acoustic field is detuned. The frequency was slowly detuned from 27 kHz by 0.01 kHz to initiate bifurcation. (Reproduced with permission from Holt *et al.,* 1994.)

tering by bubbles and SL show the transition of bubbles from nonluminescent to luminescent states to be a well-defined process (Barber *et al.,* 1994). The radius–time relationships vary with the pressure amplitude. At low driving amplitudes, the bubble oscillates but does not emit light. An increase in pressure amplitude makes the bubble smaller but its collapse ratio (maximum to minimum radii) becomes larger, and the bubble starts to luminesce. A further increase in driving pressure increases the bubble radius as well as collapse ratio and leads to a stronger luminescence. The light emission is believed to be due to two-stage heating of the cavity contents: first, by a spherical shock wave converging to the center during implosion, and second, during shock rebound. The increase in the temperature relative to the ambient is proportional to the fourth power of the Mach number M. It is estimated that the Mach number approaches a value of 30 for collapsing bubbles, and the temperatures near the center exceed 3×10^8 K (Barber *et al.,* 1994). The temperatures and densities are high enough for fusion. This raises an interesting speculation that low-level fusion can be produced in a cavitating bubble (Pool, 1994). To date there is no evidence that this is indeed possible.

Single-bubble sonoluminescence also exhibits a strong dependence on ambient temperature (Barber *et al.,* 1994) and the nature of the dissolved gas (Hiller *et al.,* 1994). A decrease in temperature from 35 °C to close to 0 °C increases light emission from 10^2 to 10^6 photons per flash. Doping nitrogen with a small amount of rare gases increases the luminosity by an order of magnitude (Fig. 9). This increase is attributed to the stabilization of bubble motion by the presence of rare gases in the bubble.

The SBSL spectrum extends far into the UV region (Hiller *et al.,* 1992). The decrease in luminescence below 240 nm is believed to be due to absorption by water. The spectrum does not show a peak at 310 nm as observed with MBSL (Didenko and Pugach, 1994; Sehgal *et al.,* 1980a; Taylor and Jarman, 1970). The spectrum is also sensitive to the composition of the dissolved gases. For example, doping nitrogen with a small amount of noble gas gives a stronger emission than either gas alone (Fig. 10).

3.2 Multiple-Bubble Sonoluminescence

As was discussed above, SL from a cavitation field occurs from multiple bubbles. There is often no deliberate attempt to control the size and composition of nuclei and the cavitating bubbles. Therefore, one can expect that within a distribution of bubbles, some will behave as transient and others as stable cavities. In such situations, the relationship between the phase of ultrasound, timing of SL flashes, and bubble oscillations is likely to be much more "chaotic" than that observed from SBSL. Figure 11 shows the phase of SL emitted by multiple-bubble cavitation in the presence of streamer activity (Gaitan *et al.,* 1992). The phase of collapse changes rather irregularly, as one may expect.

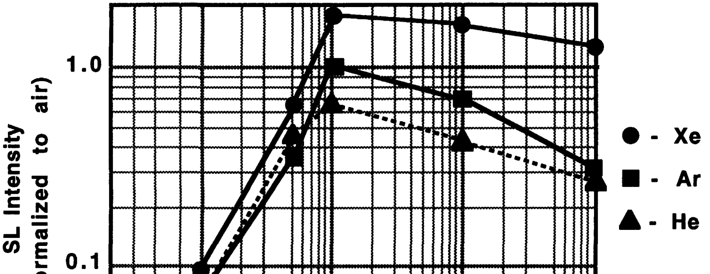

FIG. 9. SBSL intensity from water at wavelengths longer than 380 nm. Percentage composition is in mole fraction. Ambient temperature = 24 °C; ambient pressure = 1 atm; insonification frequency 25 kHz. (Adapted from Hiller *et al.,* 1994.)

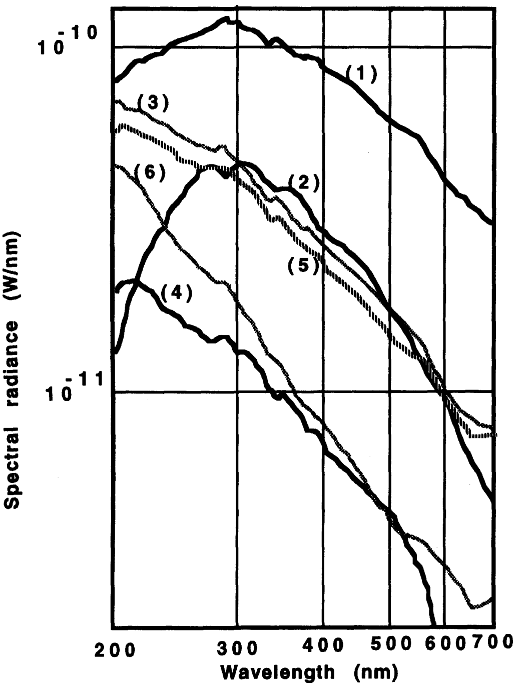

FIG. 10. Effect of dissolved gases on SBSL spectrum of water at 24 °C. Percentage represents mole fraction. Insonification frequency = 33 kHz. (1) 2% Xe in N_2, (2) Xe, (3) 1% Ar in N_2, (4) Ar, (5) 1% He in N_2, and (6) He. (Adapted from Hiller *et al.*, 1994.)

3.2.1 MBSL from Aqueous Media

One of the most widely studied forms of SL is that initiated from pure water. These studies have been reviewed in detail earlier (Verrall and Sehgal, 1988). Studies from several laboratories have shown SL to be a broad-band spectrum that extends from the visible to the ultraviolet with a peak at 310 nm characteristic of emission from hydroxyl radicals (Fig. 12). The emission is attenuated by the addition of small amounts of alcohol, which acts as a radical scavenger (Sehgal *et al.*, 1977). Specific portions of the continuum are attenuated by the addition of nitrates (Sehgal *et al.*, 1980b). The SL from aqueous salt solutions also exhibits atomic emission lines (Fig. 13) (Sehgal *et al.*, 1979). Didenko and Pugach (1994) measured SL from OH* and the simultaneous formation of hydrogen peroxide in the presence of various noble gases. They noted a parallel increase in luminescence intensity and in the yield of H_2O_2, and assigned the emission below 300 nm to be from a noble-gas–water excimer formed by a three-body collision between molecules. The formation of an argon–alkali-metal excimer has also been noted in the study of SL from alkali salt solutions (Sehgal *et al.*, 1979). These studies demonstrate that MBSL in water is due to emission from the excited chemical species.

MBSL also shows a strong dependence on temperature (Sehgal *et al.*, 1980c). Sonoluminescence becomes brighter as the temperature of water is reduced. Sehgal *et al.* explained this change by a thermodynamic model according to which the energy gained by the bubbles during their formation and growth decreases with an increase in vapor pressure (or the liquid temperature). Therefore, when the bubble collapses the energy released is also diminished, and there is a weaker emission. According to this model,

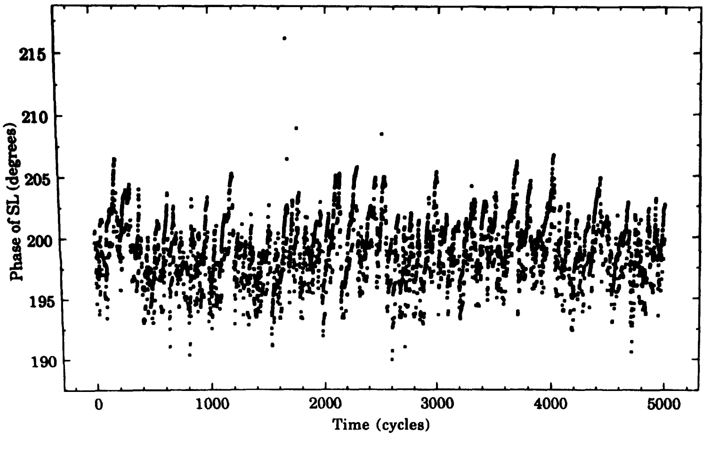

FIG. 11. Phase of SL emitted by multiple-bubble cavitation field. Ambient pressure = 1.3 atm; insonification frequency = 21 kHz. (Reproduced with permission from Gaitan *et al.*, 1992.)

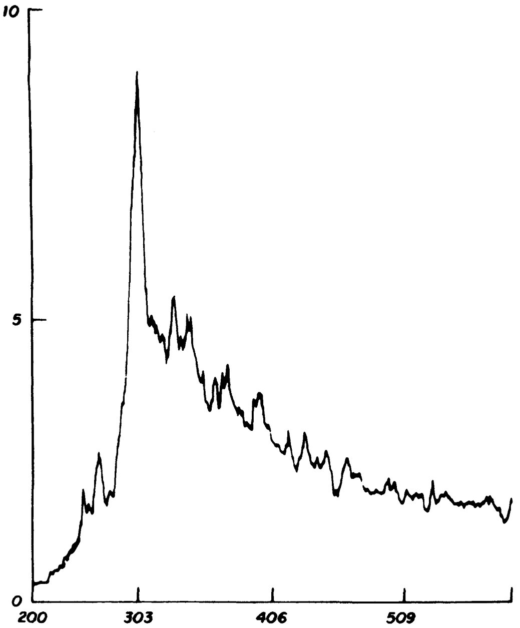

FIG. 12. MBSL spectrum from argon-saturated water at 12 °C. Insonification frequency = 333 kHz. (Adapted from Sehgal *et al.*, 1980a.)

bubbles become relatively inefficient in concentrating acoustic energy as the temperature of the medium is increased. More recently, Didenko *et al.* (1993) have studied the variation in luminosity with temperature in different spectral regions and concluded that the change in SL is also due to the quenching of the emission from H_2O^* and OH^* by the increased number of water molecules in the bubble.

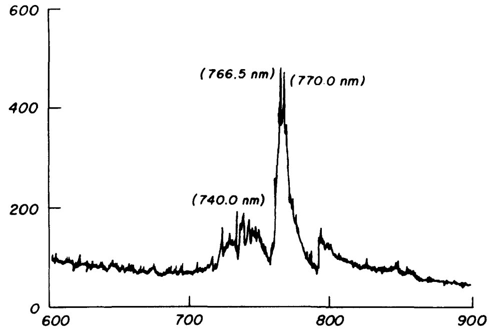

FIG. 13. MBSL spectrum from aqueous solution of potassium iodide saturated with argon at 12 °C. Insonification frequency = 459 kHz. (Adapted from Sehgal *et al.*, 1979.)

Ambient pressure is another parameter that influences MBSL. In general, as the ambient pressure is raised, the luminous intensity increases to a maximum and then rapidly decreases until it becomes undetectable (Fig. 14). The ambient pressure at which SL peaks is also a function of acoustic pressure amplitude. Two alternate approaches have been used to explain the peaking of SL with pressure. The numerical study on bubble motion (Neppiras, 1980) shows that the maximum pressure achieved by an imploding cavity in a liquid at different ambient pressures follows a curve similar to the curve of SL vs. pressure. This suggests that the variation of SL with ambient pressure is due to the change in the intensity of the cavitation event. Chendke and Fogler (1983a, 1983b), on the other hand, have argued that the variation in SL with pressure is due to the change in the number of cavitation events. Henglein and Gutierrez (1993) have used pulsed ultrasound to measure SL at different ambient and acoustic pressures. Their study shows that several acoustic pulses are needed to reach steady-state emission (Fig. 15). The steady-state luminescence at different ambient and acoustic pressures follows a trend similar to that exhibited in Fig. 14. The SL pulses are not rectangular but exhibit a delay in buildup and decay of luminescence. Both the buildup and the decay

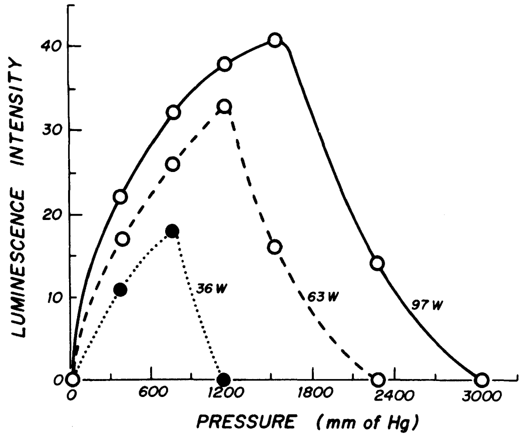

FIG. 14. MBSL intensity as a function of ambient pressure and acoustic power. (Reproduced with permission from Weissler, 1953.)

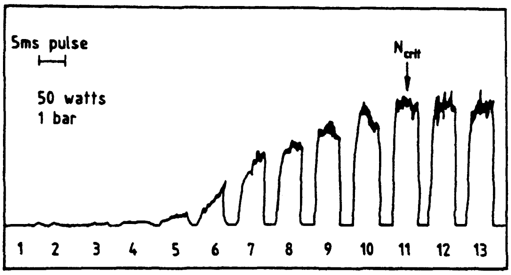

FIG. 15. MBSL from first 13 pulses of ultrasound by insonifying water under an Ar–O_2 atmosphere. Pulse length = 5 ms; duty cycle = 1/11. The time between pulses shown in the figure is compressed. (Reproduced with permission from Henglein and Gutierrez, 1993.)

times vary with applied power and ambient pressure, and the variations are believed to be due to the changes in the growth rates of the bubbles.

3.2.2 MBSL from Nonaqueous Media

Multiple-bubble SL from nonaqueous liquids was first reported by Chambers (1937) soon after SL was discovered. In recent times the most comprehensive studies have been by Suslick and co-workers (Flint and Suslick, 1989). Figure 16 shows SL from dodecane measured in the presence of different gases. In each case the spectrum extends from the visible to the ultraviolet. In the presence of the inert gas argon, the spectrum exhibits chemiluminescence from the excited state of C_2. The same medium when insonified in the presence of a reactive gas like N_2 changes the spectrum significantly. These observations are consistent with observations that the SL spectrum from water changes with the dissolved gas (Sehgal *et al.*, 1980a). Sonoluminescence from silicone oil also exhibits characteristic C_2 bands (Fig. 17) (Flint

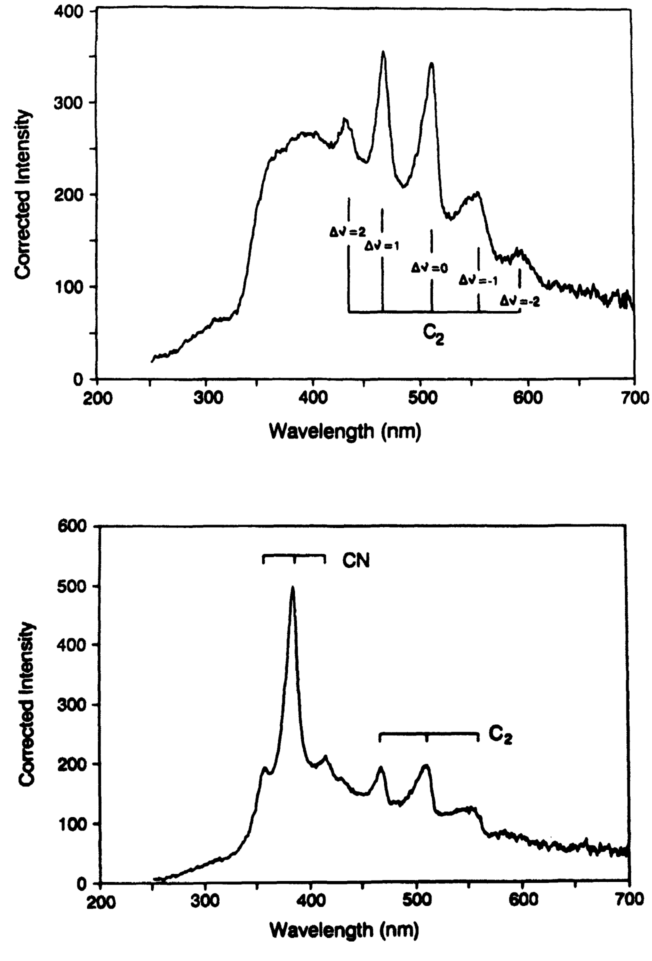

FIG. 16. MBSL from dodecane. (Top) Under Ar at 4 °C; (bottom) under 15% N_2/85% Ar at 4 °C. (Reproduced with permission from Flint and Suslick, 1989.)

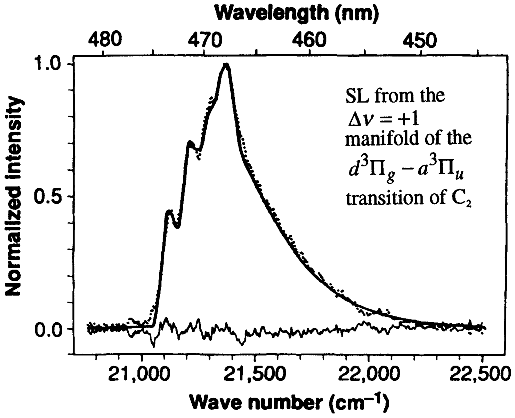

FIG. 17. MBSL from silicone oil (polydimethylsiloxane) under a continuous Ar sparge at 0 °C. Dotted line is the measured spectrum. Solid line is best-fit synthetic spectrum with vibrational and rotational temperatures of 4900 K. (Adapted from Flint and Suslick, 1991a.)

and Suslick, 1991a). Superimposed on the measured spectrum is the spectrum calculated by assuming vibration and rotation temperatures of 4900 K (dotted lines, Fig. 17). The nearly perfect fit between theory and the experiment, the presence of sharp spectral bands characteristic of specific chemical species, and the changing nature of the SL spectrum with the nature of the chemical species formed during cavitation provide strong evidence that the emission is due to chemiluminescence.

3.2.3 MBSL as a Probe of Acoustic Cavitation The concept of using SL as a means to probe cavitation was proposed by Sehgal *et al.* (1979). Previous theoretical studies made a convincing case for the presence of high temperature and pressure within a cavitating bubble (Flynn, 1975b; Noltingk, 1950). It was reasoned that the emission profile of a radiating specimen reflects the physical properties of the medium and, therefore, that the SL could serve as a probe for measuring temperature and pressure within the collapsing cavities. High-resolution SL spectra from sodium and potassium salt solutions were obtained under partial vacuum condition and used to measure the alkali-metal peak broadening and peak shift (Sehgal *et al.*, 1979). Significant peak broadening and peak shift were noted and attributed to the pressure-broadening effect. Using the data on line shift and spectral broadening (Chen and Takeo, 1957), the relative density of the emission phase of the cavity was determined. This analysis indicated the relative density of the emission phase to be on the order of 40. By modeling the emission phase as hot spots formed by collapsing bubbles (Noltingk, 1950), pressures and temperatures within the zone of cavitation bubbles were estimated to be of the order of 300 atm and 3300 K, respectively. An important question that has emerged since these early studies is what does this "emission phase" represent. Does it represent the interior of the bubble, or is it the region surrounding the bubble? Clearly, in view of the high temperatures and pressures involved, the emission is from the region that is in a supercritical state. Under such conditions, it is not feasible to differentiate between the gaseous and liquid states, and the discussion of what physical state constitutes the bubble boundary is somewhat artificial. A recent study suggests that the SL from an alkali metal is from the shell surrounding the bubble and not from within the core of the collapsing cavity (Flint and Suslick, 1991b). The characteristics of the resonance lines normalized to their peak values showed no change when volatile liquids, known to dampen cavitation, were introduced in the aqueous medium. The significant line broadening was ascribed to pressure broadening, and the emission was attributed to the shell of liquid surrounding the cavitating bubble. A theoretical study on SL that takes into consideration the formation of free radicals due to heating of the bubbles during their collapse shows the condition of high temperature to exist only at the center of the bubble and extend to 1% of the bubble radius, or less. On this basis, the size of the light-emitting region should be only a small fraction of the total volume located at the center of the bubble (Kamath *et al.*, 1994). This casts doubt on Suslick's interpretation that SL from alkali metals occurs from the hot, thin shell surrounding the bubbles. Key evidence in support of Kamath *et al.*'s proposal was reported by Matula *et al.* (1995), who compared SBSL and MBSL measured under similar conditions. The characteristic sodium peak commonly seen in the spectrum of MBSL does not occur in the spectrum of SBSL. Matula *et al.* (1995)

interpreted these results to suggest that MBSL is due to the injection of liquid droplets into the hot interior of the bubbles and not from the shell of the liquid surrounding the bubble.

The emission lifetime τ of chromium has also been used as a probe for cavitation. From the measurement of line width at half maximum of the SL lines, τ was determined to be 0.2 ± 0.02 ps (Suslick *et al.*, 1993). This lifetime is shorter than expected under normal conditions of emission. Using the measured τ and assuming the characteristics of collisions, the equations of state, and the cavitation temperatures, the fluid density and the local pressure of the emission phase of the cavitation region were estimated to be 0.15 ± 0.01 g/cm^3 and 1700 ± 110 bars, respectively.

4. NEW THEORETICAL MODELS

It is now widely accepted that SL is caused by intense local heating of a bubble either by adiabatic compression or through the formation of shock waves during its collapse. Electrical theories have been the main contender of the thermal theories. Margulis (1985) has proposed a new electrical model to explain SL. It is based on the experimental observation that stable, large, asymmetric bubbles oscillating in an acoustic field "break off" on one side to form smaller bubbles. It is proposed that there is an accumulation of electrokinetic charge around the "neck" of separation during the detachment of small bubbles from the parent bubble. The discharge and the SL emission occur as the two bubbles separate. Evidence against this model was obtained by measuring SL from $Cr(CO)_6$ in silicone oil in the presence of perfluorcarbon gases CF_4 and C_2F_6 mixed with argon and nitrogen (Suslick and Kemper, 1993). The emission spectra failed to show emission from chemical species (N_2, N^+) expected to be formed during electrical discharge. Moreover, the addition of a small amount of CF_4 or C_2F_6 significantly attenuated the SL even though these gases are capable of electric discharge. The attenuation of SL in the presence of perfluorcarbons is consistent with the thermal theory because fluorocarbons have a lower specific-heat ratio. The SL intensity of chromium was measured by using mixtures of fluorocarbon and argon in varying specific-heat ratios. The luminescence intensity increased in proportion to the specific-heat ratio (Fig. 18) as one may expect on the basis of the thermal theory (Suslick *et al.*, 1993).

Lepoint *et al.* (1994) presented evidence that suggests that sonoluminescence is a microdischarge with properties similar to that of a corona discharge. They propose a model that has attributes of both the thermal and Margulis electrokinetic models. The bubbles heat during the compression phase. During the collapse there is significant bubble deformation and interface disruption, and a fluid jet is forced into the bubble, generating a plasma. The "pinching" of microbubbles, in a manner similar to that proposed by Margulis, generates electrical microdischarge. Lepoint *et al.* (1994) recorded the corona discharge spectrum of an air–water mixture and found its spectral characteristics at a pressure of 10 atm to be very similar to that of SBSL. They noted a significant difference between the SBSL spectrum and the spectrum expected on the basis of shock-wave formation (Lofstedt *et al.*, 1993). Their measurement of corona luminescence intensity in different gases followed the order Xe > Kr >

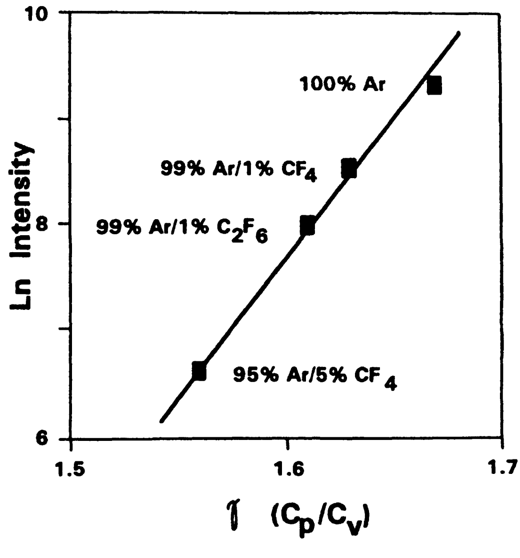

FIG. 18. MBSL intensity of excited Cr atoms as a function of specific-heat ratio. γ was varied by changing the composition of the Ar/perfluorocarbon gas mixture used to sparge the insonified liquid. (Reproduced with permission from Suslick *et al.*, 1993.)

Ar > Ne > He, which is precisely the relative order observed for SL. They observed the SL intensity in the presence of different gases to be on the same order of magnitude as predicted by the product of gas saturation concentration and corona luminescence. All these observations were cited as evidence in support of the similarity between SL and corona discharge.

An elegant thermal model combining complex bubble motion with the chemical kinetics of the reactions that occur within the bubbles has been recently developed to explain SL (Kamath *et al.,* 1994). The model uses numerical methods to determine gas temperature fields and the free radicals formed inside a stably oscillating bubble. It assumes bubble compression to be adiabatic and takes into consideration the heat loss due to conduction to provide useful insight into the phenomenon of SL. It allows for a numerical calculation of the number of free radicals formed at different intracavity temperatures within stably oscillating resonant bubbles driven under varying acoustic conditions. Figure 19 illustrates one such example for a bubble of radius 22.5 μm. The number of free radicals changes with the bubble volume and the intracavity temperature. The number of OH radicals is small below 1000 K but increases rapidly above ~1400 K. At temperatures as low as 1500 to 3000 K the number of free radicals formed within the cavitating bubbles is sufficient to explain the observed SL. These studies do not imply that higher temperatures cannot be achieved within the bubbles but that at these temperatures it is quite conceivable to have significant SL from the bubbles. The number of OH radicals formed is smaller by a factor of 5 when argon is replaced by a 3–4 times more heat-conducting gas such as neon. The largest number of OH radicals is formed in the center of the bubble, and this number decreases rapidly as a function of radial distance from the bubble surface.

The temperature map within and outside the bubble show the highest temperatures to be at the core of the bubble. The temperature gradients are so large that the surface is only a few degrees (~40 K) above the ambient temperature. It is estimated that the light-emitting region is 1% of the bubble radius or less. On this basis SL is unlikely to occur from a thin shell of liquid surrounding the bubble surface (Suslick, 1990). Although this model is able to explain strong emissions of SL (~10^6 photons) as observed experimentally from stable bubbles (Barber and Putterman, 1991), the width of the SL pulse based on this model is ~300 ps, an order of magnitude larger than the experimental estimate of 50 ps (Barber and Putterman, 1991). An interesting possibility has been suggested to explain the difference (Kamath *et al.,* 1994). The primary SL occurs from the free radicals formed in the center of the bubble. Thereafter, the number of ground-state radicals grows in the region surrounding the center, and these radicals block the outgoing radiation from the center by resonance absorption and scattering. In effect, they act as "chemical shutters," allowing the emitted light to be observed only for a fraction of the SL lifetime.

Another proposal used to explain the very

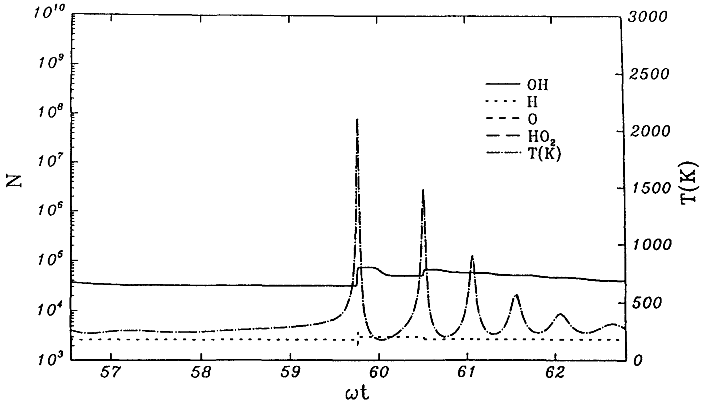

FIG. 19. Number of free radicals and the temperature (right scale) at the center of a 22.5-μm bubble oscillating stably in an acoustic field. (Reproduced with permission from Kamath *et al.,* 1994.)

high cavitational temperatures and very short lifetime of SL flashes is to use shock-wave models (Greenspan and Nadim, 1993; Lofstedt *et al.*, 1993). These models show a spherical shock wave to be formed at regular intervals during bubble oscillations. Initially the shock is weak, but it grows rapidly as the bubble compresses with spherical symmetry. The converging shock heats the core of the bubble to temperatures high enough to luminesce for a time interval an order of magnitude shorter than the bubble oscillation time.

Moss *et al.* (1994), noting that the previous models (Jarman, 1960; Lofstedt *et al.*, 1993; Prosperetti, 1984; Trilling, 1952) did not provide realistic descriptions of the heating of the gas in cavitation, proposed the use of numeric methods to solve equations of motion. They divided the bubble–liquid region by a mesh of concentric shells and solved the equations incrementally in time. The time increment was determined by the ratio of the width of the thinnest shell and the local sound speed. This strategy allowed them not to assume isentropic compression or thermodynamic equilibrium during bubble collapse. Their simulations, using an equation of state that takes into consideration the ionization and dissociation of gases, show that the shock waves are formed in the last few ($\approx$10) nanoseconds of bubble collapse. Prior to that the collapse is near-isentropic. The extreme temperatures on the order of 10 eV occur in the center of the bubble during the formation of the shock wave. Moss *et al.* (1994) noted that the ionization increases the specific heat of the gas in the bubble and lowers the temperature in the bubble. This suggests that temperatures higher than 10 eV can be obtained in the absence of ionization. In a recent study, Moss *et al.* (1996) explored the possibility of achieving temperatures high enough to cause microthermofusion by varying the composition of the bubble and the driving wave form. For this purpose, they performed numerical simulations on bubbles containing deuterium and a mixture of deuterium and water. They also varied the driving sinusoidal pressure by adding a "spike" to the wave form. Their simulations showed the peak temperatures to be on the order of 25 eV in a bubble containing deuterium and deuterated water vapor. When the same bubble was driven with a spiked wave form, the peak temperature increased by two orders of magnitude to the keV range. Moss *et al.* (1996) calculated the number of neutrons formed per hour of SBSL for the thermonuclear reaction

$$\mathrm{D} + \mathrm{D} \rightarrow {}^{3}\mathrm{He} + n + 3.2\ \mathrm{eV}$$

by multiplying the number of SBSL flashes per hour with the reaction volume integrated over the flash time. A conservative estimate, neglecting the innermost two zones, gives a rate of 0.1 neutron/h. The authors note that this rate, although small, is measurable. They believe this rate can be increased by a factor of 50 if the mixture of deuterium is replaced by a mixture of tritium. Further enhancements are also attainable by changing the amplitude, timing, and shape of the spike. They propose the use of SBSL for table-top microthermonuclear fusion and believe that an SBSL system "may produce an experimental crucible that gives the general scientific community easy and inexpensive access to pressures, temperatures, and time scales that have been unattainable previously."

5. CONTROVERSIES ON THE ORIGIN OF SL

Historically, the origin of SL has been a topic of debate almost since the time it was first observed. The early controversies centered around whether the emission is electrical or thermal in nature. Periodically, electrical theories have been postulated but have not yet been accepted as a plausible explanation for SL (Sehgal and Verrall, 1982; Suslick and Kemper, 1993). The early discussions on SL attributed the phenomenon solely to transient cavitation. Now it is widely accepted that SL can also be produced readily by stable oscillating bubbles (Barber and Putterman, 1991; Crum and Falkes, 1986; Gaitan *et al.*, 1992; Greenspan and Nadim, 1993; Margulis and Grundel, 1983; Saksena and Nyborg, 1970; Sehgal *et al.*, 1980a). In general, there is consensus among various researchers that MBSL is a result of chemical species formed in collapsing bubbles. The controversies in this area have centered around whether the intracavity temperatures are of the order of 2000–3000 K in water (Sehgal *et al.*, 1979, 1980d), or higher on the order of 5000–6000 K as measured in nonaqueous media (Flint and Suslick, 1991a; Suslick, 1990). Much has been made of these differences. Several models from the simplest to the most complex (Flynn, 1975a, 1975b; Kamath *et al.*, 1994; Noltingk, 1950) demonstrate tremendous sensitivity of cavitation to small changes in acoustic conditions used to drive the bubbles. For example, it is quite conceivable to obtain higher temperatures in materials with higher vapor pressure as compared to those of lower vapor pressure by driving the bubbles at higher acoustic pressure amplitudes. Another point of debate has been the phase of SL emission, viz., whether it occurs from the core or the shell of liquid surrounding the bubbles. Mathematical models of bubbles oscillating at modest acoustic pressures show the highest temperatures to be at the core of the bubble and the light-emitting region to be on the order of 1% of the bubble radius or less (Kamath *et al.*, 1994). Similar conclusions follow from the shock-wave model. The shock is weak near the bubble surface and grows to significant magnitude only near the core of the bubble (Barber *et al.*, 1994; Greenspan and Nadim, 1993).

Most notable has been the controversy surrounding the differences in SBSL and MBSL. Studies from several laboratories have shown MBSL to be chemiluminescent in nature with peaks corresponding to the chemical structure and composition of the surrounding liquids. Equally convincing is the evidence put forth that the emission from single oscillating bubbles is structureless and characteristic of emission from gases heated to incandescence. To date, the precise reason(s) for these differences is not known.

A possible explanation for the difference between SBSL and MBSL is based on the formation of liquid jets. During asymmetric bubble collapse, a liquid jet is formed at the bubble surface, forcing the liquid into the core of the gas bubble (Crum, 1994). This phenomenon has been photographed by high-speed photography (Crum, 1979) and is likely to occur in the bubbles in cavitation fields because of the asymmetry in collapse due to the interaction with neighboring bubbles and solid boundaries. Because of the presence of liquid inside the bubble, the spectrum of MBSL is likely to be dominated by that of the host liquid. The SBSL is observed by seeding a single gas bubble in a resonant field with very high Q (>1000). The bubble is believed to compress with spherical symmetry with no visible streamer activity. Under such conditions, liquid drops are unlikely to "microjet" into the interior of the bubbles.

Figure 20 shows a direct comparison between SBSL and MBSL from 0.1*M* sodium chloride solution for similar experimental

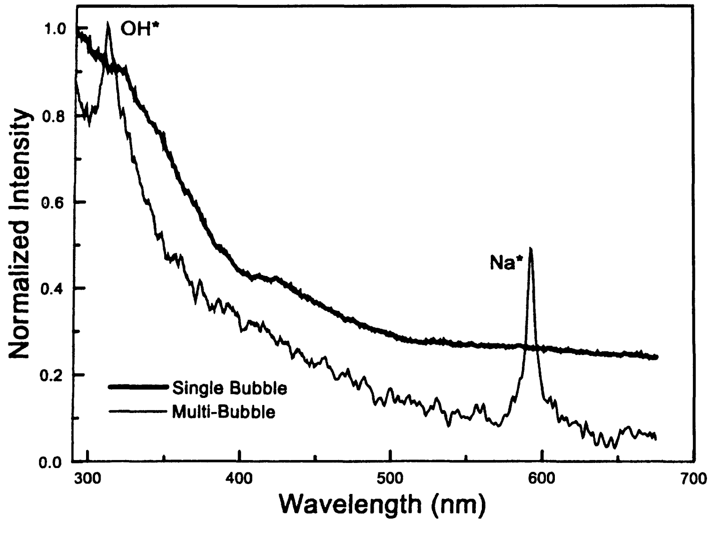

FIG. 20. Comparison of spectra of MBSL and SBSL in 0.1*M* sodium chloride solution. (Adapted from Matula *et al.*, 1995.)

conditions. The SBSL spectrum is structureless and does not show the characteristic sodium emission seen readily in MBSL. This result suggests that sodium emission in MBSL is due to the injection of liquid droplets into the interior of the bubbles in the cavitation field. Also, the entrainment of liquid may further prevent the development of shock waves and thus not allow the temperatures to be as high as those expected in SBSL.

In conclusion, there is an ongoing debate on the nature of SL and the various underlying theories. The recent progress in SBSL has sparked a widespread interest with an increasing coverage of this topic in the press (Amato, 1992; Crum and Roy, 1994; Crum and Suslick, 1995; Levi, 1991; Maddox, 1993; Pool, 1994; Putterman, 1995). These discoveries have undoubtedly opened new areas of research and a myriad of possibilities that were not realized only a few years ago.

ACKNOWLEDGMENTS

Thanks to Professor W. L. Nyborg at The University of Vermont for helpful discussions on the characteristics of cavitation bubbles. Thanks to Professor R. E. Verrall at the University of Saskatchewan for many helpful discussions and suggestions.

GLOSSARY

Cavitation: A three-step process involving nucleation, growth, and subsequent oscillation or collapse of bubbles in a liquid. Cavitation can be induced by acoustic waves, hydrodynamics, lasers, or electrical sparks. The emphasis of this article is on sonoluminescence from acoustic cavitation.

Cavitation Field: A cloud of bubbles with the individual members distributed over the entire acoustic field.

Multiple-Bubble Sonoluminescence: SL from many bubbles distributed in space.

Single-Bubble Sonoluminescence: Emission from an isolated bubble entrapped in a specific region of the acoustic field.

Stable Cavitation: Bubbles oscillating about a mean size.

Transient Cavitation: The violent collapse of bubbles.

Works Cited

Amato, I. (1992), "Making Light of Sound," *Res. News,* 1511.

Barber, B. P., Hiller, R., Arisaka, K., Fetterman, H., Putterman, S. (1992), *J. Acoust. Soc. Am.* **91,** 3061-3063.

Barber, B. P., Putterman, S. J. (1991), "Observation of Synchronous Picosecond Sonoluminescence," *Nature* **352,** 318–320.

Barber, B. P., Wu, C. C., Lofstedt, R., Roberts, P. H., Putterman, S. J. (1994), *Phys. Rev. Lett.* **72,** 1380–1383.

Becker, L., Bada, J. L., Kemper, K., Suslick, K. S. (1992), *Marine Chem.* **40,** 315–320.

Benkovskii, V. G., Golubnichii, P. I., Olzoev, K. F. (1974), *Akust. Zh.* **20,** 126–127 (*Sov. Phys. Acoust.* **20,** 74–75).

Chambers, L. A. (1937), *J. Chem. Phys.* **5,** 290–292.

Ch'en, S. Y., Takeo, M. (1957), *Rev. Mod. Phys.* **29,** 20–73.

Chendke, P. K., Foggler, H. S. (1983a), *J. Phys. Chem.* **87,** 1644–1648.

Chendke, P. K., Fogler, H. S. (1983b), *J. Phys. Chem.* **87,** 1362–1369.

Crum, L. (1975), *J. Acoust. Soc. Am.* **57,** 1363–1370.

Crum, L. A. (1979), *J. Phys. (Paris)* **40** (C8), 285–288.

Crum, L. A. (1994), *J. Acoust. Am.* **95,** 559–562.

Crum, L. A., Falkes, J. B. (1986), *Nature* **319,** 52–54.

Crum, L. A., Roy, R. A. (1994), "Sonoluminescence," *Science* **266,** 233–234.

Crum, L. A., Suslick, K. S. (1995), "Bubbles Hotter Than The Sun," *New Sci.* **146** (1975), 36–40.

Didenko, Y. T., Nastich, D. N., Pugach, S. P., Polovinka, Y. A., Kvochka, V. I. (1994), *Ultrasonics* **32,** 71–76.

Didenko, Y. T., Pugach, S. P. (1994), *J. Phys. Chem.* **98,** 9742–9749.

El'piner, I. E. (1964), *Ultrasound: Physical Chemical and Biological Effects,* New York: Consultants Bureau.

Flint, E. B., Suslick, K. S. (1989), *J. Am. Chem. Soc.* **111,** 6987–6992.

Flint, E. B., Suslick, K. S. (1991a), "The Temperature of Cavitation," *Science* **253,** 1397–1399.

Flint, E. B., Suslick, K. S. (1991b), *J. Phys. Chem.* **95,** 1484–1488.

Flynn, H. G. (1964), in: W. P. Mason (Ed.), *Physical Acoustics,* New York: Academic, p. 57.

Flynn, H. G. (1975a), *J. Acoust. Soc. Am.* **58,** 1160–1170.

Flynn, H. G. (1975b), *J. Acoust. Soc. Am.* **57,** 1379–1396.

Frenzel, H., Schultes, H. (1935), *Z. Phys. Chem.* **27B,** 421–424.

Gaitan, D. F., Crum, L. A., Church, C. C., Roy, R. A. (1992), *J. Acoust. Soc. Am.* **91,** 3166–3183.

Gilmore, F. (1952), *The Collapse and Growth of a Spherical Bubble in a Viscous Compressible Liquid,* Caltech Hydrodynamics Laboratory Report No. 26-4, Pasadena, CA: California Institute of Technology.

Greenspan, H. P., Nadim, A. (1993), *Phys. Fluids A* **5,** 1065–1067.

Griffing, V., Sette, D. J. (1955), *J. Chem. Phys.* **23,** 503–509.

Henglein, A., Gutierrez, M. (1993), *J. Phys. Chem.* **97,** 158–162.

Hickling, R. (1963), *J. Acoust. Soc. Am.* **35,** 967–974.

Hiller, R., Putterman, S. J., Barber, B. P. (1992), *Phys. Rev. Lett.* **69,** 1182–1184.

Hiller, R., Weninger, K., Putterman, S. J., Barber, B. P. (1994), "Effect of Noble Gas Doping in Single-Bubble Sonoluminescence," *Science* **266,** 248–250.

Holt, R. G., Gaitan, D. F., Atchley, A. A., Holzfuss, J. (1994), *Phys. Rev. Lett.* **72,** 1376–1379.

Jarman, P. (1960), *J. Acoust. Soc. Am.* **32,** 1459–1462.

Kamath, V., Prosperetti, A., Egolfopoulos, F. N. (1994), *J. Acoust. Soc. Am.* **94,** 248–260.

Kirkwood, J., Bethe, H. (1942), *The Pressure Wave Produced by an Underwater Explosion,* OSRD Report No. 558, Springfield, VA: National Technical Information Service.

Kozhevnikova, I. N., Bjorno, L. (1992), *Ultrasonics* **30,** 21–25.

Lepoint, T., Voglet, N., Faille, L., Mullie, F. (1994), "Bubble Deformation and Interface Disruption as a Source of Sonochemical and Sonoluminescent Activity," in: J. Blake, J. Boulton-Stone, and N. Thomas (Eds.), *Bubble Dynamics and Interface Phenomena,* Dordrecht:Kluwer Academic Press.

Levi, B. G. (1991), "Light Comes From Ultrasonic Cavitation in Picosecond Pulses," *Phys. Today* **44** (11), 17–18.

Lofstedt, R., Barber, B. P., Putterman, S. J. (1993), *Phys. Fluids A* **5,** 2911–2928.

Maddox, J. (1993), "Sonoluminescence in from the Dark," *Nature* **361,** 397.

Margulis, M. A. (1969), *Akust. Zh.* **15,** 153–173 (*Sov. Phys. Acoust.* **15,** 135–151).

Margulis, M. A. (1978), *Zh. Fiz. Khim.* **52,** 605–609 (*Russ. J. Phys. Chem.* **52,** 342–344).

Margulis, M. A. (1985), *Ultrasonics* **23,** 157–169.

Margulis, M. A., Dmitrieva, A. F. (1982), *Zh. Fiz. Khim.* **56,** 323–327 (*Russ. J. Phys. Chem.* **56,** 198–201).

Margulis, M. A., Grundel, L. M. (1983), *Dokl. Akad. Nauk SSSR* **269,** 405–407 (*Dokl. Phys. Chem. Proc. Acad. Sci. USSR* **269,** 160).

Matula, T., Roy, R., Mourad, P., III, W. M., Suslick, K. (1995), *Phys. Rev. Lett.* **75,** 2602–2605.

Mead, E. L., Verrall, R. E., Sutherland, R. G. (1975), *Can. J. Chem.* **53,** 2394–2399.

Moss, W., Clarke, D., White, J., & Young, D. (1994), *Phys. Fluids* **6,** 2979–2985.

Moss, W., Clarke, D., White, J., Young, D. (1996), *Phys. Lett. A* **211,** 69–74.

Negishi, K. (1960), *Acustica* **10,** 124.

Negishi, K. (1961), *J. Phys. Soc. Jpn.* **16,** 1450–1465.

Neppiras, E. A. (1980), *Ultrasonics* **18,** 201–209.

Noltingk, E. A., Neppiras, B. E. (1950), *Proc. Phys. Soc.* **B63,** 674–685.

Plesset, M. S., Prosperetti, A. (1977), *Annu. Rev. Fluid Mech.* **9,** 145–185.

Pool, R. (1994), "Can Sound Drive Fusion in a Bubble," *Science* **266,** 1804.

Prosperetti, A. (1984), *Ultrasonics* **22,** 69–77.

Prosperetti, A., Crum, L., Commander, K. (1986), *J. Acoust. Soc. Am.* **83,** 502–514.

Prudhomme, R. O., Guiltmart, T. (1957), *J. Chim. Phys.* **54,** 336–340.

Putterman, S. J. (1995), "Sonoluminescence: Sound into Light," *Sci. Am.* **272** (2), 46–51.

Rozenberg, L. (1971), *High Intensity Ultrasonic Fields,* New York: Plenum Press.

Saksena, T. K., Nyborg, W. L. (1970), *J. Chem. Phys.* **53,** 1722–1734.

Sehgal, C., Steer, R. P., Sutherland, R., Verrall, R. E. (1979), *J. Chem. Phys.* **70,** 2242–2248.

Sehgal, C., Sutherland, R. G., Verrall, R. (1980a), *J. Phys. Chem.* **84,** 388–395.

Sehgal, C., Sutherland, R. G., Verrall, R. E. (1980b), *J. Phys. Chem.* **84,** 529–531.

Sehgal, C., Sutherland, R. G., Verrall, R. E. (1980c), *J. Phys. Chem.* **84,** 525–528.

Sehgal, C., Sutherland, R. G., Verrall, R. E. (1980d), *J. Phys. Chem.* **84,** 396–401.

Sehgal, C. M., Steer, R. P., Sutherland, R. G., Verrall, R. E. (1977), *J. Phys. Chem.* **81,** 2618–2620.

Sehgal, C. M., Verrall, R. E. (1982), *Ultrasonics* **20,** 37–39.

Suslick, K. S. (1990), "Sonochemistry," *Science* **247,** 1439–1445.

Suslick, K. S., Doktycz, S. J., Flint, E. B. (1990), *Ultrasonics* **28,** 280–290.

Suslick, K. S., Flint, E. B. (1987), "Sonoluminescence from Non-Aqueous Liquids," *Nature* **330,** 553–555.

Suslick, K. S., Kemper, K. A. (1993), *Ultrasonics* **31,** 463–465.

Suslick, K. S., Kemper, K. A., Flint, E. B. (1993), in: M. Levy, B. R. McAvoy (Eds.), *IEEE 1993 Ultrasonics Symposium,* Vol. 2, New York: IEEE, p. 777.

Taylor, K. J., Jarman, P. D. (1970), *Aust. J. Phys.* **23,** 319–334.

Trilling, L. (1952), *J. Appl. Phys.* **23,** 14–17.

Verrall, R. E., Sehgal, C. M. (1988), "Sonoluminescence," in: K. S. Suslick (Ed.), *Ultrasound: Its Chemical, Physical, and Biological Effects,* New York: VCH Publishers Inc., p. 227.

Walton, A., Reynolds, G. T. (1984), *Adv. Phys.* **33,** 595–660.

Weissler, A. (1953), *J. Acoust. Soc. Am.* **25,** 651–657.

Weninger, K., Hiller, R., Barber, B., Lacoste, D., Putterman, S. (1995), *J. Phys. Chem.* **99,** 14195–14197.

Young, F. R. (1976), *Acoust. Soc. Am.* **60,** 100–104.

Zimakov, P. V. (1934), *Dokl. Akad. Nauk SSSR* **3,** 450–452.

Further Reading

Apfel, R. E. (1981), in: R. J. Emrich (Ed.), *Methods in Experimental Physics,* Vol. 19, New York: Academic, 355–411.

Putterman, S. J. (1995). "Sonoluminescence: Sound Into Light." *Sci. Am.* **272** (2), 46–51.

Suslick, K. S., Doktycz, S. J., Flint, E. B. (1990), "On the Origin of Sonoluminescence and Sonochemistry," *Ultrasonics* **28,** 280–290.

Verrall, R. E., Sehgal, C. M. (1988), "Sonoluminescence," in: K. S. Suslick (Ed.), *Ultrasound: Its Chemical, Physical, and Biological Effects,* New York: VCH Publisher Inc., pp. 227–286.

Walton, A., Reynolds, G. T. (1984), "Sonoluminescence," *Adv. Phys.* **33,** 595–660.

SOUND, PHYSICAL EFFECTS OF

See ULTRASOUND AND OTHER SOUND, PHYSICAL EFFECTS OF

SOUND REPRODUCTION

Manfred Krause, *Institute for Communication, Media, and Music Science, Technical University of Berlin, Berlin, Germany*

Introduction 26
1. **Preliminary Remarks** 26
1.1 Levels 26
1.2 Hearing 27
2. **Emission of Sound** 27
2.1 Elementary Sound Source 27
2.2 Complex Sound Sources 28
2.3 The Piston Source 29
2.4 Loudspeakers 30
2.5 Horns 31
2.6 Driver Principles 32
2.6.1 Electrodynamic Drivers 32
2.6.2 Electrostatic Drivers 33
2.7 Loudspeaker Cabinets 33
2.7.1 Closed-Box Systems 33
2.7.2 Vented-Box Systems 33
2.7.3 Transmission-Line Systems 34
2.7.4 Multiple-Driver Systems 34
2.7.5 Headphones 35
3. **Microphones** 35
3.1 Acoustical Principles 35
3.1.1 Pressure Microphones 36
3.1.2 Velocity Microphones 36
3.1.3 Highly Directional Microphones 37
3.1.4 Pressure-Zone Microphones 38
3.1.5 Sound-Field Microphones 38
3.2 Electrical Working Principles of Microphones 38
3.2.1 Dynamic Microphones 39
3.2.2 Capacitor Microphones 39
4. **Audio Signal Processing** 40
4.1 Amplifiers 40
4.1.1 Amplifier Fundamentals 40
4.1.2 Preamplifiers 41
4.1.3 Operational Amplifiers 41
4.1.4 Processing of Signal Dynamics 42
4.1.4.1 Compressors and Limiters 42
4.1.4.2 Expanders 42
4.1.4.3 Companders 42
4.2 Mixing Consoles 43
4.2.1 Console Layout 43
4.2.2 Automated Consoles 45
4.2.3 Digital Consoles 45
4.2.4 Volume Indicators 45
4.2.4.1 Volume Unit Meter 45
4.2.4.2 Peak-Level Meters 45
4.2.4.3 Stereo Indicators 46
4.3 Line Transmission 46
5. **Digital Audio** 46
5.1 Data Sampling 46
5.2 Oversampling 47
5.3 Analog-to-Digital and Digital-to-Analog Converters 47
5.4 Digital Signal Processing 48
5.5 Audio Data Reduction/ Perceptual Coding 49
6. **Audio Signal Recording** 50
6.1 Analog Disc 50
6.1.1 Disc Recording 50
6.1.2 Disc Playback 51
6.1.2.1 Pickup Styli 51
6.1.2.2 Pickup Cartridges 52
6.1.2.3 Tone Arms 52
6.2 Analog Tape Recording and Playback 52
6.2.1 Tape Recording 52
6.2.2 Tape Playback 53
6.2.3 Recording and Playback Equalization 54
6.2.4 Tape Noise 54
6.3 Analog Recording and Playback Drives 55
6.4 Digital Sound Recording 55
6.4.1 Digital Disc Recording 55
6.4.1.1 Compact Disc 55
6.4.1.2 Magneto-optical Disc (MOD) 56
6.4.1.3 Mini Disc (MD) 57
6.4.2 Digital Tape Recording 57
6.4.2.1 Stationary-Head Tape Recording 57
6.4.2.2 Rotating-Head Tape Recording 58
7. **Room Acoustics and Sound Reinforcement** 58
7.1 Room Acoustics 58
7.2 Sound Reinforcement 59
7.3 Stereophony 59

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

7.3.1 Two-Channel Stereophony 59
7.3.2 Quadrophony and Surround Sound 60
7.3.3 Head-Related Stereophony 61
7.3.4 Multiple-Channel Sound Transmission 61
Works Cited 61
Further Reading 63

INTRODUCTION

There are many aspects to sound reproduction, and everybody is confronted with them in day-to-day life, whether listening to the radio, a Walkman, or the announcements in a department store. Each task has its special demands, but the solutions are based on the same physical principles. Sound reproduction refers primarily to the transmission of speech and music over space as well as time and giving the listener the impression of being present at the place of origin. But there is also music generated with electronic equipment and mixed with new sounds never heard before, combined with voices manipulated with filters and other electronic equipment. Another wide field is sound reinforcement in theaters, concert halls, and open-air stages, designed to improve acoustics or even articulation.

The quality of sound reproduction is determined by the properties of sound sources and microphones as well as the amplifiers, mixers, and sound-recording equipment used.

The technical evolution in sound reproduction is aimed at the transmission of the complete audible frequency band and the full dynamic range of the ear. Besides this, distortion, caused by the nonlinear behavior of the signal-processing components, must be reduced to inaudible values. This evolutionary process started with analog systems and is now, in many cases, replaced by digital systems. However, the ends of the sound reproduction chain, microphones and loudspeakers, are still analog components.

In Sec. 2, after a short discussion of the properties of elementary sound sources, the design of different loudspeaker systems commonly used are described. Section 3 deals with the properties and the design of microphones. Then, in Sec. 4, the audio signal-processing systems in recording studios, such as amplifiers and mixing consoles with their signal-processing units, are treated. Also, aspects of digital signal processing are presented. In Sec. 5, analog and digital audio signal recording are described. This begins with analog disc and tape recording, and continues with digital disc and tape recording. The final section provides an overview of problems of sound reproduction in rooms. Here, aspects of two-channel stereophony, quadraphony, and surround sound are presented.

1. PRELIMINARY REMARKS

1.1 Levels

In acoustics and electroacoustics as well as in telecommunication, power, voltages, amplification, and attenuation are calculated as levels. This means that the logarithm of the ratio of a physical quantity with respect to a reference is used. In order to indicate this mathematical operation, levels are expressed in decibels (dB). With power as the given quantity, levels are defined as

$$L = 10 \log(P/P_0) \quad \text{dB}, \tag{1}$$

where P_0 is the reference power and log is the Briggs logarithm (i.e., to base 10). Since power is proportional to the square of voltage, pressure, etc., voltage levels, e.g., can be expressed as

$$L = 20 \log(U/U_0) \quad \text{dB}, \tag{2}$$

where U is the voltage in question and U_0 is the reference voltage. Sound-pressure level (SPL) is defined as

$$L = 20 \log(p/p_0) \quad \text{dB}, \tag{3}$$

where $p_0 = 2 \times 10^{-5}$ Pa (20 μPa) is the reference sound pressure.

In sound transmission voltage levels are common, but there are different reference voltages in common use. U_0 can be 1 V; voltage level is then denoted as dBV. Or U_0 can be 0.775 V, which is denoted as dBU, or dBm if there is a reference impedance of 600 Ω. Volume indicators will often be adjusted to a system reference voltage, e.g., 1.23 V (equivalent to +4 dBU) or 1.55 V (equivalent to +6 dBU).

1.2 Hearing

Human perception of sound is confined to a frequency range from about 20 to 20 000 Hz. At 1 kHz perception is limited to sound pressures from 20 μPa to about 60 Pa (0 to 130 dB SPL). This is plotted as the auditory sensation area in Fig. 1. Within this area the main ranges of music and speech are inserted.

Curves of equal loudness measured in phons are of interest if music and speech are reproduced at loudness levels other than the original source. They are also important for noise rating.

2. EMISSION OF SOUND

2.1 Elementary Sound Source

If molecules of an elastic medium are excited by oscillating bodies, they will transfer their impact to neighboring molecules, and waves will move away with the speed of sound c. The waves give rise first to pressure deviations with respect to the static air pressure p_s and then to movements of the molecules about their position of rest (ignoring thermal movements), which will be represented by the velocity v of the movement. The speed of sound c is dependent on temperature and air pressure. In air under normal conditions (20 °C and 1023 hPa), c is 334 m/s. Frequencies f between 20 and 20 000 Hz are audible, and the corresponding wavelengths are between 17 m and 17 mm. This has to be taken into account in the construction of audio equipment.

According to the Huygens principle, every point in a wave field originates from an elementary spherical wave. Complex sound fields can be considered as a superposition of elementary waves. This is also the case if the waves meet obstacles distributed in space. The sound field is then influenced by either reflections or diffractions.

Sound waves are characterized by sound pressure p and particle velocity v. These can be computed from the Helmholtz wave equation for sinusoidal spherical waves with the wavelength λ or the frequency f,

$$\frac{\partial \Psi}{\partial t^2} + k^2\Psi = 0, \tag{4}$$

where $\Psi = p/r$ and $k = 2\pi/\lambda$ or $k = 2\pi f/c = \omega/c$. The solution of Eq. (4) is

$$A \exp[j(\omega t - kr)] + B \exp[j(\omega t + kr)], \tag{5}$$

where the negative sign denotes an outgoing and the positive sign an incoming wave. We may consider here only the outgoing wave ($B = 0$).

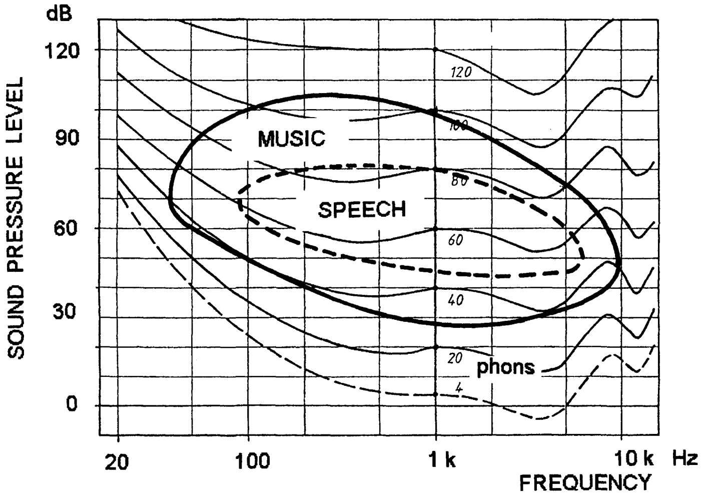

FIG. 1. Auditory sensation area with music and speech ranges marked.

The arbitrarily chosen value A can now be replaced by the so-called volume velocity q. This is the air volume moved by the oscillating surface of a sound source per unit of time. A uniformly sinusoidally oscillating sphere gives the sound pressure with respect to the distance r from the center

$$p(r,t) = \frac{\omega\rho q}{4\pi r}\exp[j(\omega t - kr)], \tag{6}$$

where ρ is the density of the air.

For the velocity of the molecules of the air, we obtain

$$v(r,t) = (q/4\pi r^2)(1 + jkr)\exp[j(\omega t - kr)]. \tag{7}$$

Expressed as magnitude and phase, Eq. (7) gives

$$v(r,t) = (q/4\pi r^2)\sqrt{(1 + k^2r^2)}\exp[j(\omega t - kr + \varphi)] \quad \text{with } \varphi = \tan^{-1}(1/kr). \tag{8}$$

The ratio p/v gives the acoustic impedance Z:

$$Z = \rho c\left\{\frac{k^2r^2}{1 + k^2r^2} + j\frac{kr}{1 + k^2r^2}\right\}. \tag{9}$$

For large r, where the waves can be considered as plane waves, Z becomes ρc. This value is denoted as the characteristic field impedance Z_c. As can be seen from Eq. (6), sound pressure decreases like $1/r$. This is also true for the velocity v at greater distances, as follows from Eq. (7). But near the sound source, velocity v decreases like $1/r^2$. So, in some cases we must discriminate between near and far field (Fig. 2). In the far field, sound pressure and velocity are in phase, while in the near field, a phase shift of 90° occurs.

For a monopole with the surface $S = 4\pi R_0^2$, we get the acoustic impedance

$$Z_s = S\rho c\left\{\frac{k^2R_0^2}{1 + k^2R_0^2} + \frac{jkR_0}{1 + k^2R_0^2}\right\}. \tag{10}$$

The real part Z_r is the acoustic resistance and thus the radiation load due to the emitted sound waves, while the imaginary part Z_i gives the load due to the medium mass. Figure 3 shows the characteristics of Z_r and Z_i with respect to kR_0, relative to the characteristic field impedance Z_c.

At low frequencies Z_r increases in proportion to f^2. At frequencies above the transition frequency $c/\pi R_0 = 2$, Z_r approaches asymptotically Z_c. Then, sound power is emitted with highest efficiency.

The acoustical power radiated by a source with the surface S moving with the velocity v is $P_a = Sv^2Z_r$, and the moving air mass is $m_s = SZ_i/\omega$.

2.2 Complex Sound Sources

Real sound sources are seldom spherical. According to Huygens's principle, they can be constructed as a combination of spherical sources oscillating with different amplitudes and phases. A special case is the dipole source with monopoles separated by a dis-

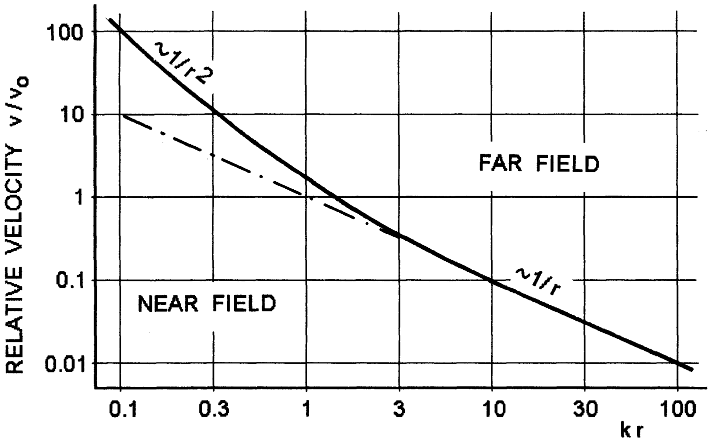

FIG. 2. Attenuation of velocity with respect to distance r from the sound source.

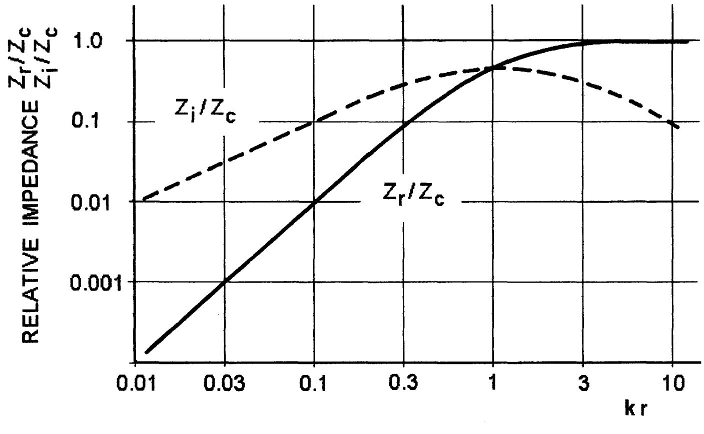

FIG. 3. Real and imaginary parts of acoustic impedance of a spherical sound source.

tance *d* oscillating with the same amplitude but with a phase difference of 180°. Because of the antiphase oscillations, the sound waves cancel partially. This gives a radiation pattern that in polar coordinates is proportional to cos(α), where α is the directional angle from the connecting line of the sources. This results in a figure-of-eight pattern, as in Fig. 4.

Monopoles arranged in a row or column give more complex directional patterns. If a number of monopoles oscillating with the same phase are mounted at the interval *d* as shown in Fig. 5, the three-dimensional radiation pattern is circular around the axis of the column, while the cross section has an elliptical envelope with a number of side lobes. If the number of monopoles is *m*, the axes of the ellipse are proportional to $a = 1$ and $b = 1/m$. The number of side lobes depends strongly on the ratio d/λ and thus on frequency. An example with $m = 4$ and varying d/λ is shown in Fig. 5. The radiation pattern can be rotated with respect to the axis of the column if the signals fed to each of the monopoles have appropriate time delays. By arranging the monopoles in a two-dimensional array, the circular pattern around the axis can be changed accordingly, giving a radiation beam off the plane of the array. This may also rotated by introducing delays in the feeding lines.

Musical instruments have complexly formed bodies oscillating in different modes. Consequently, they have complex directional patterns that depend on the frequencies of the sounds played. The different harmonics may be emitted in different directional patterns and cause colorations of sound with respect to the listening angle. This can also be the case in loudspeakers with large diaphragms or in multipath loudspeaker systems (see Sec. 2.7.4).

2.3 The Piston Source

Most loudspeakers have diaphragms. The model of the diaphragm is a piston mounted on an infinitely large baffle. A piston source

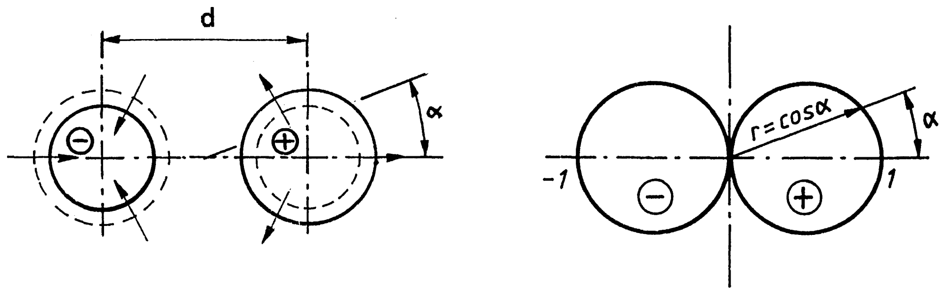

FIG. 4. Directivity pattern of a dipole sound source.

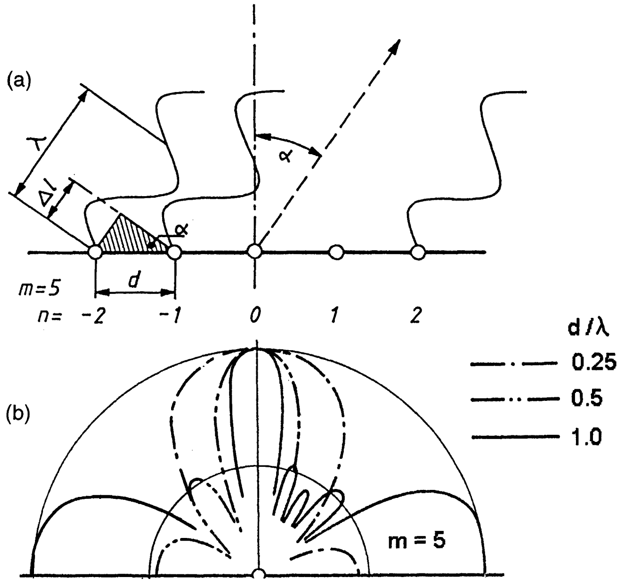

FIG. 5. (a) Interferences caused by column of five monopole sound sources and (b) the resulting directivity pattern.

works as a dipole, and a baffle prevents the antiphase sound of the back of the diaphragm from interfering with the frontal outgoing waves. For low frequencies, when the diameter of the diaphragm is small with respect to the wavelengths of the sound waves, the radiation does not differ from that of a spherical source. At higher frequencies the plane surface acts more like an array of elementary sources with a more complex radiation pattern (Lord Rayleigh, reprint 1945). Therefore, the sound pressure in the near field may become zero. This may become important in near-field listening conditions.

The acoustic impedance Z of the piston shows only slight deviations from that of the monopole, and so for computations the same plots (Fig. 3) may be used.

2.4 Loudspeakers

If a loudspeaker has a stiff diaphragm that can be characterized by one mass m, one spring with the compliance n, and one damping structure with friction r, then according to Newton's law a force F applied to the diaphragm gives

$$F = m\frac{dv}{dt} + \frac{1}{n}\int v dt + rv, \tag{11}$$

where v is the velocity of the diaphragm.

The displacement a and the velocity v of the diaphragm caused by a sinusoidal force $F = F_0 \exp(j\omega t)$ are

$$a = \frac{F}{-\omega^2 m + 1/n + j\omega r} \tag{12}$$

and

$$v = \frac{F}{r + j(\omega m - 1/\omega n)}. \tag{13}$$

The resonance frequency is $\omega_r = 1/\sqrt{mn}$. Figure 6 shows the standardized frequency responses for displacement a and velocity v of the diaphragm with respect to the resonance frequency ω_r.

The acoustical power emitted by a sound source is

$$P_a = v^2 Z_r. \tag{14}$$

To obtain constant power radiation in a given frequency range, this product must be constant. As can be seen from Fig. 3 and Fig. 6(b), this is possible between the resonance frequency f_r and the transition frequency $c/\pi R_0$. Since the displacements of loudspeaker diaphragms are limited to about 10 or 20 mm, high-power loudspeakers for low frequencies require large diaphragms to raise the acoustic resistance. With higher frequen-

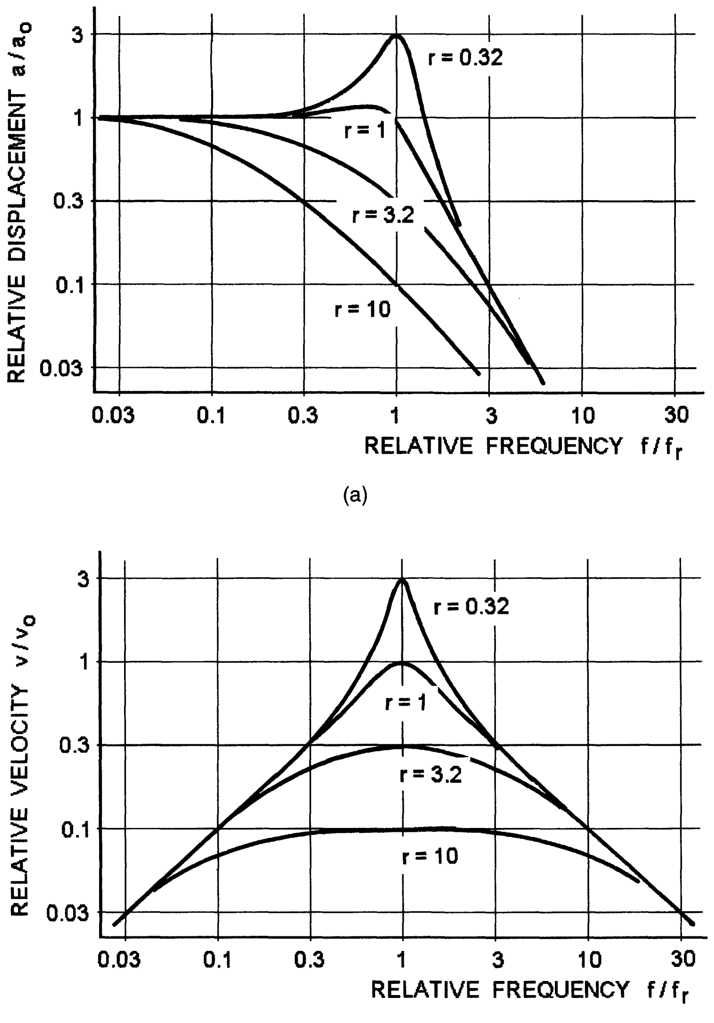

(a)

(b)

FIG. 6. (a) Displacement and (b) velocity of a mechanical system with respect to relative frequency f_r/f_0.

cies the diameter of loudspeakers must be proportionally smaller to achieve a wide-angle directional pattern.

As a consequence of the relatively large dimensions of the diaphragms, the simplified assumptions of infinite stiffness are not true. Therefore, the frequency response of a real loudspeaker exhibits many peaks and dips caused by partial resonances and interferences.

2.5 Horns

In many cases it is not necessary or desirable to emit sound power omnidirectionally. Horns typically concentrate waves in a smaller solid angle and additionally raise the acoustic impedance. The exponential horn is widely used. The cross section S along the axis x increases according to (Webster, 1919)

$$S(x) = S(0)\exp(\gamma x) \text{ with } \gamma = 4\pi f_c/c, \qquad (15)$$

where $S(0)$ is the cross section at the throat of the horn. The factor γ defines a cutoff frequency f_c below which a horn with a circular cross section will radiate no acoustic power.

The acoustic resistance with respect to the characteristic field impedance Z_c is shown in Fig. 7. Compared with the acoustic resistance of the spherical source, it increases more rapidly, and therefore the power emission is more efficient.

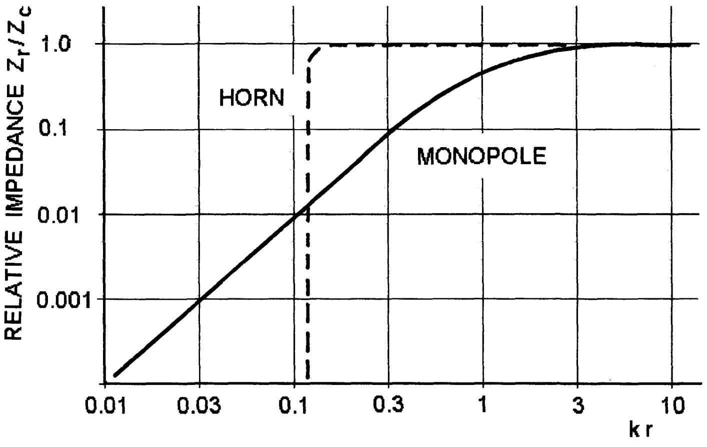

FIG. 7. Real part of the acoustic impedance of an exponential horn (dotted line). For comparison, the real part of the acoustic impedance of a spherical sound source (monopole) is drawn (solid line).

Although the diagram is valid only for infinitely long horns, it may be applied to horns with finite length. In this case the acoustic resistance Z_r shows some peaks and dips, which have influence on the frequency response of the horn, and low frequencies will also be radiated below the cutoff frequency, but with reduced efficiency.

The efficiency of a horn radiator can be further increased by using a compression chamber with a throat having cross section S_t smaller than the surface of the diaphragm S_d. If the velocity of the diaphragm is v_d, the velocity at the throat is then

$$v_t = v_d S_d / S_t. \tag{16}$$

As a result of the increased velocity, the emitted sound power is higher, by a factor S_d/S_t. The dimensions of the compression chamber must be small compared to the shortest wavelengths of the sound to be emitted.

Instead of exponential horns, other forms—e.g., with a hyperbolic shape—are in use. The shape influences the wave fronts leaving the horn and may be designed to achieve constant directivity (Smith *et al.*, 1983). The cross sections of the horns can be made asymmetrical in order to distribute the sound in desired directions.

For low audio frequencies the size of the horn must be very large. This gives rise to constructions with folded horns (Klipsch, 1941). In addition, room boundaries can serve as extensions of the horn. Horns are often used for public address systems and for high-frequency radiation in high-fidelity systems.

2.6 Driver Principles

2.6.1 Electrodynamic Drivers Most drivers work on the electrodynamic principle. A conductor with length l can move freely in a magnetic field with the flux density B. An electric current I in the conductor produces the force $F = BlI$. Apart from so-called ribbon loudspeakers, where the conductor itself forms the diaphragm, the conductor is usually a coil that can move in the ring gap of a magnet. The coil is attached to the diaphragm, which is held by a centering device and the surrounding rim in order to maintain the free movement of the coil in the magnet gap.

The material of the diaphragm is normally made of special paper, although other materials such as aluminum or carbon fiber are sometimes used as sandwich structures. To obtain a stiff structure, diaphragms have a conical form. This form is often modified into a curvilinear one by which the stiffness of the outer part of the diaphragm decreases. This prevents partial oscillations, caused by reflections from the rim of the diaphragm. It also improves the high-frequency response of the loudspeaker, because only the parts near the center of the diaphragm will follow the rapid movements of the coil while they are dampened in the outer parts. For this reason, the center parts are often an aluminum dome. Mid- and high-frequency loudspeakers often have only dome diaphragms or, as in the case of the ribbon loudspeaker, the conductor itself is the diaphragm.

Uniformity of the magnetic flux density in the gap is of great importance. This serves to keep nonlinear distortions low. The decreasing magnetic flux density at the ends of the gap are taken into account by using a coil either somewhat longer or shorter than the gap length. This way the number of turns of the coil is linked always with the same flux within the range of normal displacements. Eddy currents caused by the moving coil can be suppressed with shorting rings, thereby improving the high-frequency characteristic of the driver.

2.6.2 Electrostatic Drivers Electrostatic loudspeakers make use of the effect that a force $F = EQ$ acts on an electrical charge Q in an electrical field with the field strength E. This can be utilized by stretching a diaphragm with a constant electric charge between two perforated narrow-spaced electrodes. To obtain sufficient acoustic output, relatively high signal voltages (1000–2000 V) must be applied, and large diaphragms are necessary. The capacitive load of the electrostatic loudspeaker needs either special driving amplifiers or a coupling network. Because of the large dimensions of the diaphragm, the directivity pattern tends to concentrate acoustic power along the axis. This may be compensated for by dividing the diaphragm into different sections. Of these sections, the center ones will be actuated by mid- and high-frequency signals. The electrostatic loudspeaker has very low distortion on account of the relatively small displacements of the diaphragm and high damping by the air.

2.7 Loudspeaker Cabinets

As discussed in Sec. 2, a piston source needs an infinite baffle, or at least one with dimensions on the order of the wavelengths of the lowest frequencies to be radiated, in order to prevent the cancelling by the antiphase waves coming from the back of the piston. Since the emission of low frequencies needs either large diaphragms or large excursions of them, it is helpful to confine the radiation to smaller solid angles. If the loudspeaker is mounted in a corner of the room, then the acoustic power is radiated only in an eighth of the space. This gives an increase in sound-pressure level of 9 dB with respect to the omnidirectional radiation (Villchur, 1957). For sound waves, walls act as mirrors. Therefore, in enclosed spaces, loudspeakers with omnidirectional radiation patterns will give rise to reflections affecting their frequency response (Allison, 1974). Although loudspeaker enclosures come in a variety of shapes and sizes, the majority of practical designs are variants of the closed box, the vented box, or the transmission line.

2.7.1 Closed-Box Systems The closed box functions, in some respect, as an infinite baffle. However, the sound is (at least in theory) radiated in all directions. The enclosed air volume acts as a spring with a compliance n_b and reduces the compliance of the loudspeaker n_l. The compliance of a box with a volume V with a diaphragm of area S is

$$n_b = V/S^2 p_s \kappa, \tag{17}$$

where p_s is the static air pressure and κ is the adiabatic coefficient of the air. This will increase the resonance frequency of the loudspeaker by the factor $1 + n_l/n_b$. This may be compensated for by increasing the mass and the compliance of the diaphragm.

The construction of boxes has to take into account the fact that standing waves can develop inside. The frequencies f_e of the standing waves in a rectangular enclosure with the dimensions l, w, and h can be computed from

$$f_e = 2\,c\sqrt{[(n_l/l)^2 + (n_w/w)^2 + (n_h/h)^2]}, \quad \text{with } n_l, n_w, n_h = 1, 2, 3, \ldots. \tag{18}$$

Standing waves should be reduced through the use of absorptive material, which should be inserted at places with high particle velocity. Normally, the boxes are stuffed nearly completely with absorptive material [Fig. 8(a)].

With some drivers the suspension of the diaphragm is maintained nearly exclusively by the volume of the box. This is denoted as acoustic suspension.

2.7.2 Vented-Box Systems A box having a volume V and a vent with the length l and the cross section S_v, as schematically

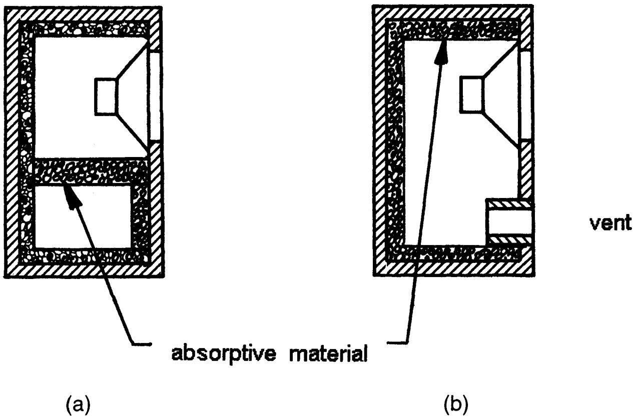

FIG. 8. Construction scheme of (a) a closed and (b) a vented box.

drawn in Fig. 8, is a Helmholtz resonator with the resonance frequency

$$f_r = \frac{c}{2\pi}\sqrt{\frac{S_v}{V(1 + l')}}, \tag{19}$$

where l' is an additive correction length. When this is combined with a loudspeaker, a coupled system results. Depending on the degree of coupling, the electrical resonance curve can split into two resonances, f_{r1} and f_{r2}, and broaden the low-frequency transmission range. Instead of a vent, a properly tuned passive diaphragm (drone cone) can be used.

The design of either a closed box or a vented box can be done by following the so-called Thiele–Small parameters, as developed by Small (1973) and Thiele (1971).

2.7.3 Transmission-Line Systems If a loudspeaker is mounted at one end of an open tube, then sound waves travel through the tube, which may be folded, and leave it at the other end. This causes interferences between the waves emitted from the loudspeaker at certain frequencies. If an odd number of half wavelengths equals the length of the tube, then the sound will be enhanced. If an integral number of wavelengths equals the length of the tube, then the sound will be substantially attenuated. So, a comb-filter effect will occur (see Sec. 3.1.4). The transmission-line enclosure improves the low-frequency sound radiation when higher frequencies are attenuated. This is achieved with absorptive material on the inner walls of the tube.

2.7.4 Multiple-Driver Systems Although a single loudspeaker can be designed to emit the whole audio-frequency band, most loudspeaker designs split the audio band into two to four bands. This means that loudspeakers with optimal properties for each band can be used. The band splitting is done by crossover filters. In many cases there are passive devices directly connected to the loudspeakers. Alternatively, active filters can be used with individual amplifiers for each band. There are many design methods for filters that take into account the frequency and phase response of loudspeakers in boxes. The separation of the bands may be softer or harder according to the choice of the filter ordering. Simple filters comprise a low-pass filter of first order, which is obtained by an inductance, and a high-pass filter with a capacitor, as shown in Fig. 9(a). Filters of second order, as in Fig. 9(b), or even of higher order are often used for three- or four-way systems.

The impedance of a typical electrodynamic loudspeaker is resistive only at two or three frequencies. The simple first-order filter shown in Fig. 9(a) may exhibit unforeseen and undesirable transfer characteristics. Higher-order filters as in Fig. 9(b) interact less strongly with a loudspeaker load but still may require additional components for proper behavior.

At the crossover frequency ranges, two loudspeakers radiate simultaneously. If they are mounted at some separation, there will be interferences in the sound field, as discussed in Sec. 2.2. This is also true if the loudspeaker design has been optimized for a

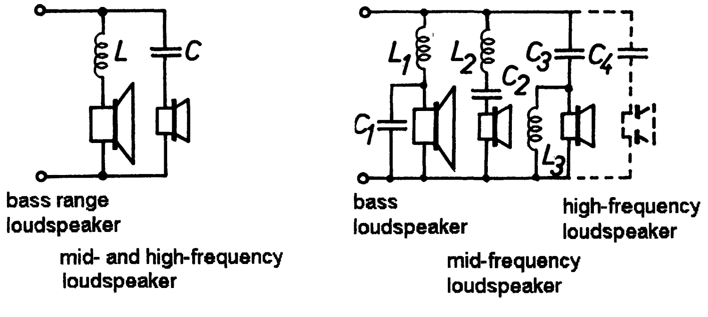
L
C
bass range loudspeaker
mid- and high-frequency loudspeaker
L_1
L_2
C_3
C_4
C_2
C_1
L_3
bass loudspeaker
high-frequency loudspeaker
mid-frequency loudspeaker

(a) (b)

FIG. 9. Circuitry (a) of a two-way loudspeaker system with first-order high-pass and low-pass filtering and (b) of a three-way loudspeaker system with second-order filtering.

flat on-axis frequency response. Interferences can be reduced by arranging the loudspeakers as narrowly as possible; interferences in the horizontal plane are minimized if loudspeakers are vertically mounted. To obtain a good impulse response, the loudspeakers may be arranged in depth. Alternatively, a delay may be introduced electrically so that the sound waves produced by the individual loudspeakers are in phase. A coaxial arrangement is also a possible solution.

Loudspeaker cabinets can cause diffraction effects that will also affect the frequency response. They can be minimized by the layout of the form of the cabinets (Rasmussen, 1995).

2.7.5 Headphones Headphones are widely used in noisy environments or in situations where loudspeakers are unwanted. They are necessary for artificial head stereophony or for virtual-reality systems. They are designed either as pressure or as velocity types (Moller, 1995). Pressure types have a tight coupling to the ear. If correctly fixed to the head, they give a very good frequency response and substantial attenuation of other sounds. Velocity types have an acoustically transparent pad that makes it necessary to enhance low frequencies. Since only small displacements of the diaphragm are needed to generate a high sound pressure at the ear drums, nonlinear distortions in headphones are low even at peak levels.

The frequency responses of more early headphones were tuned for free-field listening, i.e., for frontal incidence of sound; recently, frequency responses are mostly tuned for diffuse incidence of sound (Theile, 1986).

Most headphones have electrodynamic driver systems. However, some more expensive ones have electrostatic drivers with wide bandwidth and improved impulse transmission characteristics.

Except in the case of artificial or dummy-head stereophony, headphones offer less than optimal stereophonic localization. Usually, sound sources seem to be located within the head. This effect can be avoided to some degree if acoustical cross talk, inherent in standard stereophony, is cancelled by appropriate delays and filters. This principle is shown in Fig. 10 (Rubak, 1991), where the cancellation filters are symbolized by their transfer function $H(f)$.

3. MICROPHONES

3.1 Acoustical Principles

Microphones convert the acoustical energy of the sound waves into electrical energy. Depending on the constructive layout,

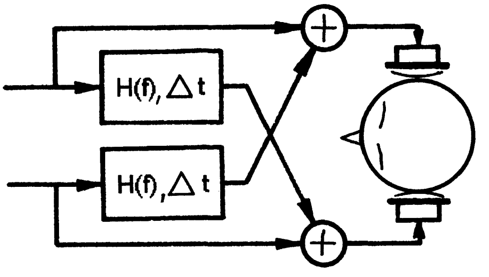

FIG. 10. Principle of cross-talk cancelling for headphones with cross filters and delay to be used with two-channel stereophony.

microphones can be sensitive to sound pressure, velocity, or a mixture of both. Mechanical parameters determine the frequency response of the diaphragm according to Newton's law [see Eq. (12)]. In combination with the electrical transducing principle, this will give the frequency response of the microphone. It should be noted that all frequency responses are referred to constant sound pressure.

3.1.1 Pressure Microphones The diaphragm of a pressure microphone is fixed to a small chamber and thus exposed to sound waves from one side only, as schematically drawn in Fig. 11(a). Variations in "static" air pressure are compensated by a capillary tube. An ideal pressure microphone has an omnidirectional pickup pattern. In a two-dimensional plot, we get a circular pattern, as shown in Fig. 11(b). However, it should be borne in mind that the pickup pattern is three-dimensional. Normally, the dimensions of the microphone will be larger than the wavelengths of the high frequencies of sound. This causes diffractional effects and reflections so that the pickup pattern is no longer omnidirectional.

3.1.2 Velocity Microphones The diaphragm of an ideal velocity microphone is exposed to sound waves on both sides [Fig. 11(a)]. The pickup pattern of an ideal velocity microphone exhibits a bidirectional pickup pattern that in a two-dimensional plot forms a figure-of-eight pattern, as in Fig. 11(b).

Because of the dimensions of real microphones, a frontally incident sound wave must travel around the structure on a path with the length Δs, as schematically shown in Fig. 12(a). This results in a pressure difference Δp. For $\Delta s \rightarrow 0$, we obtain the pressure gradient.

Microphones working with pressure differences are often called gradient microphones. Given a constant sound pressure, the pressure difference increases with frequency or with decreasing wavelength, as is schematically drawn in Fig. 12(b). If the travel path Δs becomes $\lambda/2$, the pressure difference will become zero. The working frequency range as gradient microphone is thus limited.

With suitable design, a mixture of gradient and pressure sensitivity can be achieved. Combining a pressure and a gradient microphone with equal sensitivity to frontal incoming waves gives a so-called cardioid directivity pattern, denoted in Fig. 13 by the factors 0.5 for the omnidirectional and 0.5 for the gradient component. This will give a substantially reduced sensitivity to sound coming from the back. If the ratio of the sensitivities is changed, either a "wide" cardioid or a "supercardioid" pickup pattern can be achieved, as is also shown in Fig. 13.

A more or less unidirectional pattern with one microphone can be obtained with tubes and cavities combined with porous materials. Sound waves arriving from behind the microphone are then appropriately delayed and attenuated.

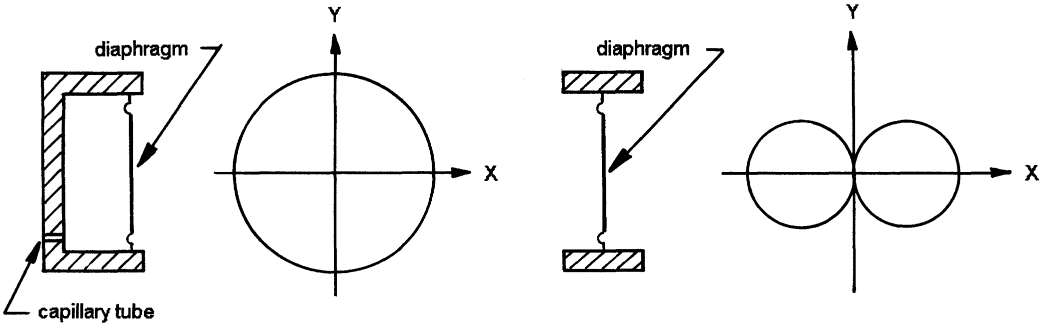

pressure microphone

(a)

gradient microphone

(b)

FIG. 11. Construction scheme and pickup patterns of (a) a pressure microphone and (b) a pressure-gradient microphone.

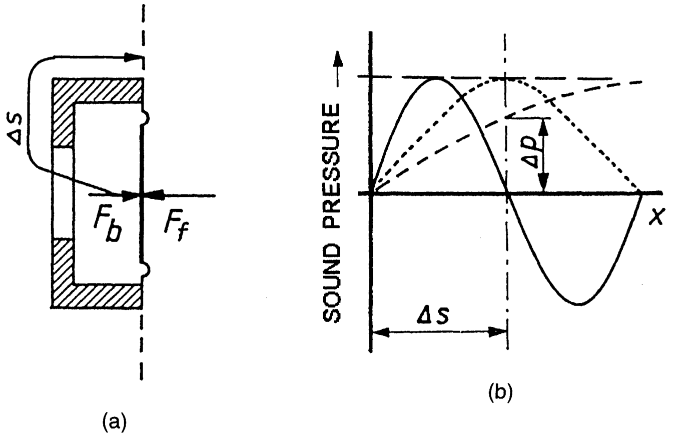

FIG. 12. (a) Construction scheme of a pressure-difference microphone. (b) For sound waves with increasing frequencies = decreasing wavelengths but constant sound pressure, the traveling path Δs from front to back of the diaphragm causes an increasing pressure difference Δp.

3.1.3 Highly Directional Microphones

Arrays of microphones may be used to obtain high directivity. Their pickup patterns depend, as in the case of loudspeakers (see Sec. 2.4), on the ratio of mutual distance and wavelength. The pickup pattern may be rotated by inserting delays in the signal paths.

So-called rifle, or shotgun, microphones are widely used. These have tubes with slits or holes in front of the diaphragm, as shown in Fig. 14. Sound waves hitting the tube cause traveling waves in the tube. Depending on their angle of incidence, the traveling waves are partly cancelled, while no cancellation occurs in the direction of the tube itself.

The response of gradient microphones in the near field of point sources, such as voices, is of special interest. Since the radius of the near field is dependent on the wavelength (see Sec. 2.1), a source with a wide spectrum will be in the near field for low frequencies and, at the same time, in the far field for high frequencies. This results in a low-frequency boost, which in some cases has to be compensated by a high-pass filter. This low-frequency boost in gradient microphones makes them relatively sensitive to streaming air, such as wind or plosive voice sounds. These sounds cause vortices at the edges of the microphone bodies (Wuttke, 1992). This sensitivity may be reduced by windscreens of gauze mesh or special foams. Gradient microphones are also more sensitive to shocks than to pressure microphones.

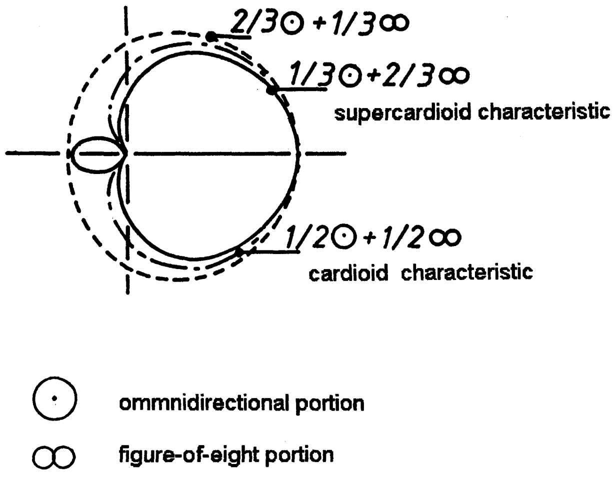
2/3⊙+1/3∞
1/3⊙+2/3∞
supercardioid characteristic
1/2⊙+1/2∞
cardioid characteristic
⊙ ommnidirectional portion
∞ figure-of-eight portion

FIG. 13. The combination of microphones with an omnidirectional pickup pattern and a figure-of-eight pattern, respectively, gives unidirectional pickup patterns. The shapes of the patterns depend on the ratio of the individual sensitivities of the microphones.

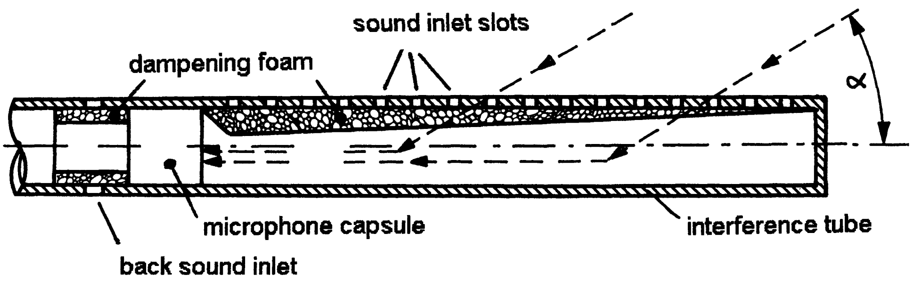

FIG. 14. Longitudinal section of a rifle or shotgun microphone.

Therefore, shock mounts both in the microphone stands as well as within the microphone suspension frames should be provided. Resonance frequencies should be less than 16 Hz and well damped.

3.1.4 Pressure-Zone Microphones If a microphone is mounted near a reflecting surface, its frequency response is affected by the interference of direct and reflected waves. This gives the so-called comb filter effect with a coloration of sound that is normally unpleasant (Fig. 15). This effect can be suppressed with pressure-zone microphones. These are mounted in planes to be laid on tables or floors. This principle is mostly used for pressure microphones but may also be applied to gradient microphones. A special form is the spherical-case microphone. Here, small pressure microphones are mounted diametrically in a sphere about the size of a human head. The diffraction and shadowing of sound waves that arrive from different directions yields characteristic frequency responses similar to that of the human head. This produces stereophonic signals well suited for sound reproduction by loudspeakers (see Sec. 7.3).

3.1.5 Sound-Field Microphones The signals of either three orthogonally directed figure-of-eight microphones and an omnidirectional microphone, or four cardioid microphones at the corners of a tetrahedron, can be weighted and added so that a cardioid pickup pattern can point into any direction of the space. As many directional patterns as wished may be derived simultaneously from these sound-field microphones. Therefore, they can be used not only as coincident stereo microphones (see Sec. 7.3.1) but also for multiple-channel sound transmission as "Periphony" (Gerzon, 1973), "Eidophony" (Scherer, 1977), or "Orthophony" (Hensel *et al.*, 1992).

3.2 Electrical Working Principles of Microphones

Although any effect that converts mechanical movements into electric currents or voltages can be used for microphones, only the electrodynamic and the electrostatic principles are widely used in high-quality sound reproduction.

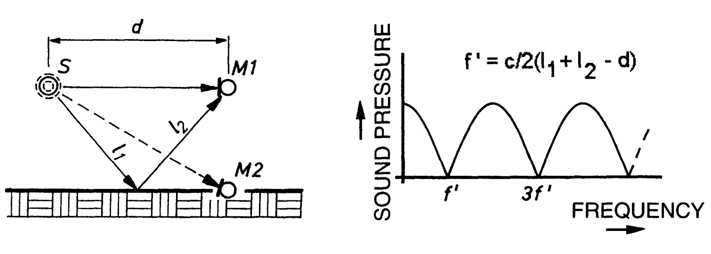

(a) (b)

FIG. 15. (a) Interferences of sound waves caused by reflections from the floor resulting in (b) a so-called comb filter effect at microphone M1. The zeroes of the frequency response are dependent on the difference of the travel paths of the direct and reflected waves. No interferences occur if the microphone M2 is mounted flush with the floor.

3.2.1 Dynamic Microphones In dynamic microphones an electric conductor with length l moves in a magnetic field with the flux density B and the velocity v, giving a voltage $U = Blv$. The frequency response of a dynamic pressure microphone follows the resonance curve of diaphragm velocity v_d (see Sec. 2.4). A flat frequency response can be achieved by damping the diaphragm. An expansion of the frequency range can be realized by additional resonances, e.g., with cavities in front of or behind the diaphragm.

With the surface S of the diaphragm, the pressure difference Δp will give the driving force $F = S\Delta p$. Therefore, a flat frequency response is obtained above the resonance frequency of the mechanical system. The frequency response will, at higher frequencies, be affected by the cutoff conditions with respect to wavelengths, as discussed in Sec. 3.1.2. A sufficient sensitivity at higher frequencies results from the fact that the microphone functions then as a pressure microphone.

There are two types of dynamic microphones. In the ribbon type the diaphragm is simultaneously the conductor. The output voltage is extremely low and has to be enhanced by a built-in transformer to about 0.5 mV/Pa. This is equivalent to 94 dB SPL. In the moving-coil microphone the conductor is wound up to a coil that moves in the gap of a ring magnet. The diaphragm to which the coil is attached is often made of aluminum, but also other materials may be used. The output voltage is about 2 mV/Pa, and the source impedance is about 200 Ω.

3.2.2 Capacitor Microphones The capacitor in this microphone type is formed by a back plate and the diaphragm. The capacity of a capacitor C with the plates in a distance d and surface S is

$$C = \epsilon S/d, \tag{20}$$

where ϵ is the permittivity. If the plates move with excursions s, so that $d = d_0 + s$, this can be written as

$$C = (\epsilon S/d_0)(1 - (s/d_0) + (s/d_0)^2 - + \ldots). \tag{21}$$

This shows that with large displacements of the diaphragm, nonlinear distortions will appear.

There are two operating principles for capacitor microphones: the low-frequency (LF) method and the radio-frequency (RF) method. In the LF system, an electric charge Q is applied to the diaphragm either from a dc source via a large resistor R, as shown in Fig. 16(a), or by the electric field of an electret. The voltage U across the resistor R increases with frequency, but this is compensated by the capacitance C_0 of the microphone capsule. The output voltage is proportional to the voltage of the dc source U_0 and the ratio s/d. In order to obtain a linear frequency response, the cutoff frequency determined by RC_0 has to be at the low end of the audio-frequency band. Thus, the value of R will be about 300 MΩ. As a result, the source resistance is very high, necessitating a high input resistance (about 1 GΩ) of the microphone amplifier. This amplifier is normally integrated in the microphone.

In an RF system the capacitor determines the frequency of an RF oscillator. The frequency is then phase modulated and demod-

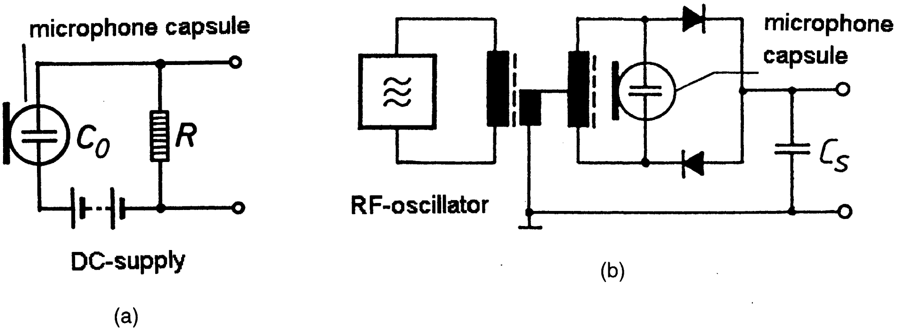

FIG. 16. Circuitry (a) of a LF condenser microphone and (b) of an RF condenser microphone.

ulated by a discriminator circuit, as shown in Fig. 16(b). Another RF system (Hibbing, 1984) uses amplitude modulation, achieved by a push-pull bridge design. The diaphragm is centered between the plates of the condenser and works as the tap of a voltage divider. The high-frequency voltage at the diaphragm is thus proportional to its displacement.

Since the output voltage of condenser microphones is proportional to the displacement of the diaphragm, pressure-sensitive condenser microphones demand a high resonance frequency of the diaphragm. This is achieved by stiff clamping and an appropriate back volume. Velocity condenser microphones must be tuned to a frequency near the middle of the audio-frequency band. A relatively high damping for wide-band operation is needed. One way to achieve this is with a porous back plate.

The pickup patterns of LF capacitor microphones with two diaphragms on both sides of a porous plate can be altered by changing the polarity of one of the diaphragms. If the microphone is designed so that one nonpolarized diaphragm has a cardioid pickup pattern, then the overall pickup pattern is omnidirectional when the other diaphragm has the same polarity. Alternatively, when the other diaphragm is of opposite polarity, the resulting pickup pattern is a figure of eight. The patterns can be altered steadily if the polarization voltage is changed by a potentiometer, as, for example, in Fig. 17.

The output voltage of typical condenser microphones is about 20 mV/Pa, and the output impedance is about 200 Ω. Capacitor microphones always have some electronic circuitry inside and therefore need a power supply. A dc supply can be fed to the microphone via the balanced signal cable and the shielding (phantom feeding, see Sec. 4.3). The feeding, via resistors that are considerably greater than the internal resistance of the signal source, affects the signal level only slightly.

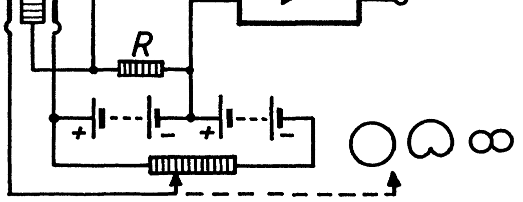

FIG. 17. Circuitry of a dual-diaphragm condenser microphone with variable pickup pattern. The variation is achieved by changing the polarity of the supply voltage of one of the diaphragms.

4. AUDIO SIGNAL PROCESSING

4.1 Amplifiers

4.1.1 Amplifier Fundamentals An amplifier is a two-port device mainly characterized by a transfer function representing frequency and phase response. Ordinarily, the frequency response should be flat within a certain range, e.g., the audio-frequency range. Amplification is achieved by a number of vacuum tubes or transistors. Since these active elements exhibit nonlinear and low-pass behavior, they are the cause of nonlinear and linear distortions. Nonlinear distortions can be reduced with negative feedback. To a certain degree this technique can also extend the frequency range.

Amplification is closely connected with the problem of electronic noise caused by thermal movement of electrons in conducting materials. According to Nyquist, the noise voltage U_n in a circuit with impedance Z is

$$U_n = \sqrt{4kTZ_rB}; \qquad (22)$$

k is Boltzmann's constant (1.38×10^{-23} J/K), T is the temperature in kelvins, Z_r is the real part of the frequency-dependent electrical impedance Z, and B is the bandwidth of transmission in hertz. For a resistor of 200 Ω and a bandwidth of 15 kHz, this yields a noise voltage of 0.23 μV, corresponding to -132.8 dBV. The noise voltage is amplified according to the gain of the amplifier. The circuitry of the latter contributes to the noise of the source, so that the noise voltage at the output of the amplifier is somewhat higher. The difference is denoted as the noise figure of the amplifier, which in the case of microphone amplifiers can be optimized to about

1–2 dB (Netzer, 1981). While the spectral energy density of electronic noise is constant (white noise), transistors and tubes generate a flicker noise at low frequencies, which has a $1/\sqrt{f}$ slope. Since the annoyance of amplifier noise is frequency dependent, with a maximum around 1–3 kHz, audio system specifications have to be weighted according to internationally acknowledged standards. Here, filters with frequency responses of either the *A* curve (after International Electrotechnical Commission IEC 651), which follows the inverted 30-phon curve (see Sec. 1.2), or the CCIR curve (after Comitè Consultatif International des Radio-Communicationes CCIR-468-2) are applied (Fig. 18).

Electroacoustic equipment is, then, characterized by the (unweighted) signal-to-noise ratio (SNR) and the weighted SNR relative either to a reference voltage (e.g., 1 V) or to a maximum output voltage, defined by tolerable nonlinear distortions.

4.1.2 Preamplifiers The output voltage of microphones and other sensing elements is in most cases relatively low. A dynamic microphone with an impedance of 200 Ω may deliver a level of −54 dBU at 94 dB SPL. The noise voltage of the microphone is, according to Eq. (22), −132.8 dBU, which, if converted to a sound-pressure level, is equivalent to 15.2 dB SPL (unweighted). The dynamic range of a microphone channel therefore depends on the maximum sound pressure picked up in a given situation. Microphone amplifiers must handle a wide range of input levels and raise them to the internal reference level of the mixing console, which is normally about 1 V, with an overload margin of about 20 dB. The gain of a microphone amplifier normally can be varied in steps of 6 dB, from 0 to about 80 dB. In some cases there may also be an attenuator for very high-level inputs. The input impedance is typically about 1–2 kΩ, and the output impedance is about 20–50 Ω.

4.1.3 Operational Amplifiers Chip technology is widely used for audio amplifiers. Here, operational amplifiers are used. Operational amplifiers have a balanced input and, normally, an unbalanced output. Operational amplifiers have an extremely high gain, about 100 000, and can do almost any signal-processing task through appropriate application of negative feedback. Because of the high amplification involved, relatively strong feedback is necessary to obtain moderate amplification. Thus, nonlinear distortions are low, and the frequency response is dependent almost exclusively on the frequency characteristic of the components in the feedback loop. With resistors used in the so-called inverting circuitry, gain is determined by the ratio R_g/R_n (Fig. 19).

Operational amplifiers are especially used for summing signals (Fig. 19). Because of high internal gain, the signal voltage at the summing point—the negative input pin of the amplifier while the positive input pin is grounded—is very low. This can be interpreted as if a small resistor is connected between the summing point and the ground pole. The input is thus "virtually grounded"

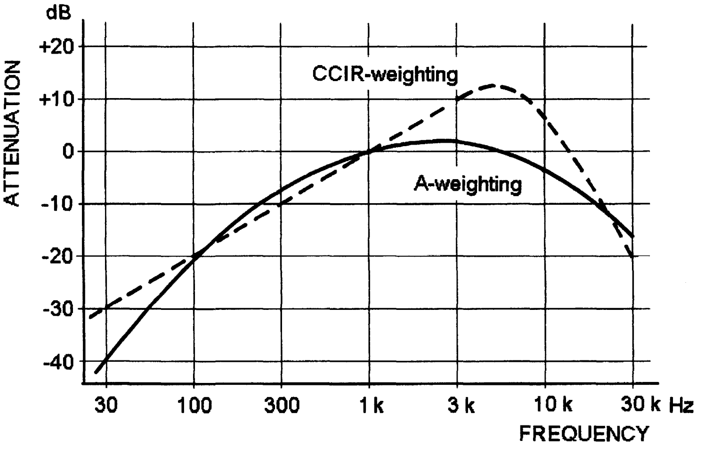

FIG. 18. Weighting curves for noise rating after IEC-268 and CCIR-468-2.

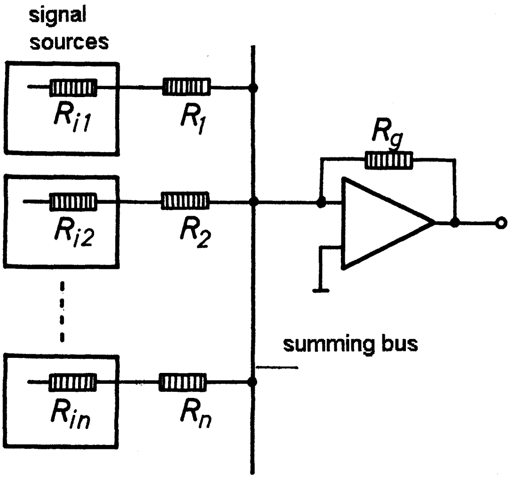

FIG. 19. Circuitry of a summing module for *n* inputs with an operational amplifier.

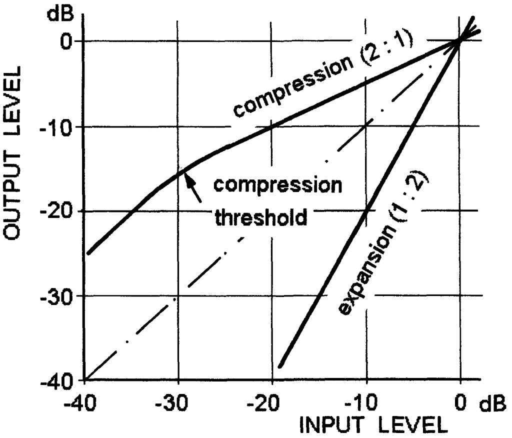

FIG. 20. Level characteristics of compression and expansion. The compression or the expansion may have a threshold of onset. A compression ratio of ∞:1 defines a limiting device.

and therefore substantially not susceptible to cross talk.

4.1.4 Processing of Signal Dynamics

The dynamic range of analog recording and radio transmission systems is limited by inherent noise and maximum allowable signal amplitudes. In many sound reproduction situations, the listener does not want the full dynamic range of the original source, either because of a noisy listening environment or because he does not want to disturb his neighbors. Reducing the dynamics of speech and music is the task of recording engineers and tonmeisters (see Sec. 4.2). It demands a good knowledge of the material to be dynamically processed. The problem can be solved at least to some extent by automatic control.

4.1.4.1 Compressors and Limiters. A compressor reduces a level difference ΔL of the original signal to a lesser value $\Delta L'$, which is determined by the compression ratio, e.g., 2:1. Limiters have a large compression ratio (100:1 or even more). Compressors make use of voltage-controlled amplifiers. The control voltage is derived from either the output or the input of the amplifier by rectification and filtering the signals. The compression as well as the expansion characteristic for steady-state signals may, as shown in Fig. 20, have an onset threshold at low levels and limiting thresholds at high levels. The time constants of filtering may be different for attack and release, and they will influence the envelopes of the signal to some extent. If the onset slope of the control voltage is less steep, then signal overshoot will occur. If the slope of release is smooth, there may be audible "pumping" when background noise is present. If the release time constant is short with respect to the periods, especially at low frequencies, signals will be altered, thus causing audible harmonic distortions.

4.1.4.2 Expanders. Expansion (as the inverse process to compression) functions in a similar manner, also using voltage-controlled amplifiers. Both processes are combined to overcome the noise floor of a transmission system and are widely used in analog tape recording. The dynamics at the output of a mixing console must be reduced to the dynamics of the transmission system by enhancing low levels and, perhaps, reducing high levels. The dynamics must be expanded again with respect to the demands of the listener, either to the original levels or, if desired, somewhat less. Processors may be classified as wide-band or multiband systems. Multiband systems are based on the psychoacoustic effect of masking in order to avoid audible "pumping" effects and nonlinear distortions.

4.1.4.3 Companders. The Dolby A™ noise reduction system, comprising four channels, is widely used in studio tape recording (Dolby, 1967). The compressor enhances levels only below −40 dB by 15–20 dB, while high levels are unaffected. In this manner, overshooting distortions occur only between −40 and −20 dB, a range in which the ear is relatively insensitive. The characteristic is achieved by a limiter with a 15–20-dB higher gain and with a limiting threshold at about −40 parallel to a unity-gain amplifier. The outputs of all bands are summed. For playback, signals are sent again through a limiter and a parallel unity-gain amplifier, but the limiter signal is subtracted in order to give the appropriate attenuation at low levels. The compander characteristic is given in Fig. 21.

Because of the nonlinearity of compression and expansion, the characteristics must be exactly inverse. This requires the adjustment of a reference level and a relatively good linear frequency response of the tape recorders. The adjustment of the transmission level is accomplished by a test tone (Dolby tone). For cassette recorders, simplified versions of the Dolby system are used. These process only the higher-frequency band above about 1 kHz and are denoted as Dolby *B* and *C* (Dolby, 1987b). Improved versions of Dolby noise-reduction systems are denoted as "spectral recording" (Dolby, 1987a).

As mentioned, the original dynamic range of speech and music is often unwanted during reproduction. Since the beginning of radio transmission and high-quality sound recording, engineers and tonmeisters have narrowed the dynamics to 40 or 45 dB. Their knowledge of dynamic contours gave them the opportunity to alter levels both in advance and afterward. Because companders can only react to current signals, this problem can be solved only if there is information regarding upcoming loudness contours. This can be achieved by sending the signals through a delay system and a parallel analyzing processor in order to calculate the control information. This information then must be transmitted to the expander. The look-ahead time may be as long as 20 s (Million, 1985), although for common use this may be much too long (Gilchrist, 1994).

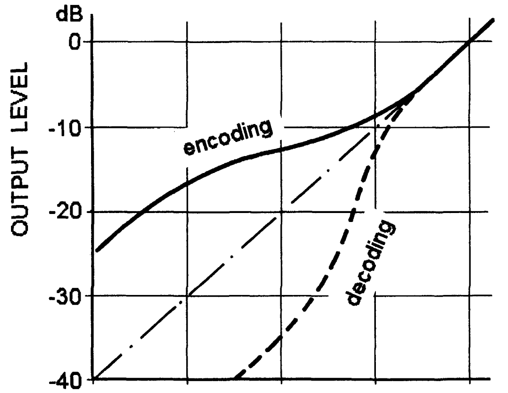

FIG. 21. Level characteristics of Dolby "compander" systems. Encoding boosts low levels, while higher levels are not affected. Decoding must be carried out with an inverse expansion characteristic.

4.2 Mixing Consoles

Work at the mixing console is done normally by recording engineers. In Europe, the tasks sometimes are subdivided between different specialists. Recording engineers are responsible for the technical aspects, while a tonmeister, who has had both musical and technical training, is responsible for the artistic aspects.

4.2.1 Console Layout In mixing consoles, the audio signals of different sources such as microphones, tape recorders, and so on are added, manipulated, and distributed. Although the structure of the mixing consoles is dependent on the application, there are some common features. Level controls (faders) are always associated with each signal path, and, in most cases, filters are present in order to influence frequency response in some manner. Filters are designed to enhance or attenuate low and high frequencies with a moderate slope of ±6 dB/octave. Sometimes there are adjustable filters with bell-shaped frequency responses for the enhancement or attenuation of one-octave frequency bands (Fig. 22).

Many mixing consoles combine *n* input signals to *m* groups, and the groups are then summed to *s* output channels, as schematically drawn in Fig. 23. Besides the components for manipulating the sound signals, there are also provided facilities for intercommunication between the recording studio and the control room. Stereo signals may be

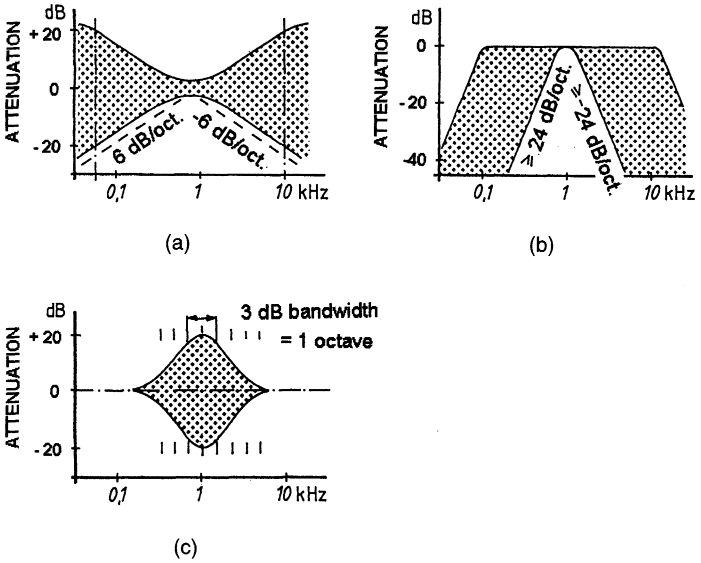

FIG. 22. Schematics of frequency responses of filters in mixing consoles: (a) First-order high and low passes, (b) high and low passes with steep slopes, and (c) band passes.

handled together, and other signals can be fed into the stereo channels by balancing them with panorama potentiometers (panpots). Signal routing can be achieved by switches or by patchboards. Most often group signals are sent to the tape recorder, and then returned to the console. In this case, the summing section serves as control monitor during recording and later as a section for mixing the different signals down to, e.g., two-channel stereo (mixdown). Besides the send and return lines to or from the recorders, there are send and return lines for echo (synthetic reverberation; see Sec. 5.4). Send lines are also provided for soloist headphones. Insert jacks are also provided for ex-

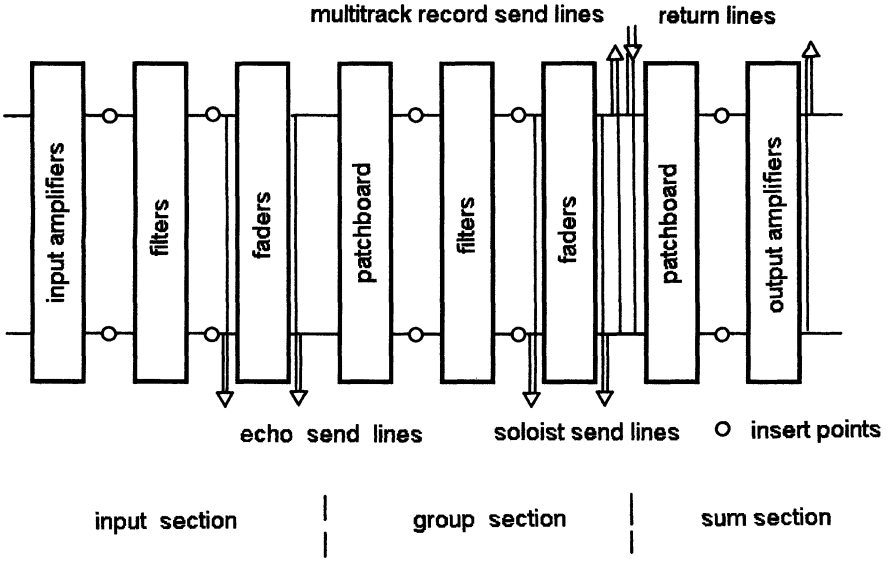
multitrack record send lines
return lines
input amplifiers
filters
faders
patchboard
filters
faders
patchboard
output amplifiers
echo send lines
soloist send lines
O insert points
input section
group section
sum section

FIG. 23. Block diagram of a mixing console with sections for inputs, grouping, and summing.

ternal equipment, such as multiband equalizers, delay lines, and sound-processing devices.

Signals must be monitored acoustically during mixing and recording. Here, switches provide prefade and postfade listening. Signals are further monitored by volume indicators (see Sec. 4.2.4) that may be associated with each signal path.

4.2.2 Automated Consoles A sound engineer has to watch and to manipulate many signals. Although many settings are made in advance, real life often brings unforeseen problems. During mixdown, most of the fader manipulations may have been correct, but one or two may be unsatisfactory. With a nonautomated mixer, mixes have to be done repeatedly until the result satisfies both operator and producer. With an automated mixer, all settings and temporal manipulations are recorded digitally and synchronized with a time code on the tape. If a take has to be repeated, only the signals to be changed are manipulated. The other signals are processed automatically, following the previous mixdown. The faders of those mixing consoles are motor driven, and the automatic control is interrupted only when the operator touches the fader of the signal path to be altered.

Signal-processing components such as amplifiers and filters are, in some cases, separated from the control unit. Processes are often supported by computers with one or more screens, where all settings and routing information are displayed.

4.2.3 Digital Consoles Digital signal processing can take over all the tasks done with analog mixers. Digital consoles do not differ much from analog consoles either in their layout or in their possibilities of manipulation. But recently the computer paradigm has set new trends. The structure of a mixing console can be reduced to a matrix where nodes contain the faders, filters, and other components. Coordination and routing are determined by suitable addressing of the components, or may even be handled automatically. Routing information is then displayed on a computer display (Musialik and Knaeple, 1991). In multitrack and especially in hard-disk editing, signal-processing data such as fader settings or crossfading are performed with the aid of a screen display on which the signals of the tracks are visualized and can be enlarged to any extent. So, for example, hard cuts can be made without audible clicks. Furthermore, the results of manipulations in hard-disc recording can be heard immediately since no rewinding and cueing of a tape recorder is necessary.

Digital signal processing also includes the possibility of delaying signals. This is advantageous in multiple-microphone stereophony (see Sec. 7.3).

4.2.4 Volume Indicators Amplifiers have an upper limit for output voltage, which is often defined by the tolerable signal distortion. Auditive monitoring does not indicate how far from the limit the modulation is. To make best use of the dynamic range of the system, signal strength has to be measured. There are two devices in general use: the volume unit (VU) meter and the peak-level meter.

4.2.4.1 Volume Unit Meter. The VU meter is widely used because of its relatively simple construction. The signal is full-wave rectified with an approximately squaring rectifier (copper oxide), and the instrument is calibrated to the equivalent rms value of the signal. The time constant is about 200 ms so that the 100% point of the scale is reached in 300 ms, and the overshoot is minimal. The time constant of 200 ms simulates to some degree the integrating properties of human hearing. However, the VU meter does not indicate the peak values of the signals. In many cases, these are much higher than the peak value of sine waves. Therefore, many VU meters have a so-called lead of 4 or even 6 dB. This means that a sine signal with a true level of −10 dB will be indicated with −6 or −4 dB. By this, and with some experience of the sound engineer, the modulation of speech and music will not overload the system. Another shortcoming of the VU meter is its limited range with respect to low levels: levels below −20 dB cannot be indicated.

4.2.4.2 Peak-Level Meters. Since it is less necessary to know the volume of signals than their true distance to the upper limit, peak-level meters with some electronics are in use. These are especially important with digital recording, where overloading will produce severe distortions through hard clipping. In Europe, level meters with an approximately logarithmic scale ranging from −50 to +5 dB, an attack time constant of about 10 ms, and a decay time constant of about 0.5 s were commonly used. Although there was no real peak indication, the possible short-time overloads were not audible in analog audio. With digital technology it was felt necessary to shorten the attack time constant to about 1 ms. Peak-level meters are mostly bar graph systems with a large number of light-emitting diodes (LED). They sometimes feature peak hold facilities for comfortable reading. Instead of LEDs, screen displays with different colors for the channels are also used.

4.2.4.3 Stereo Indicators. For stereo signal control it is desirable to monitor the phase relationship of the channels and the correlation of the signals. A simple method is to use an oscilloscope and to deflect the beam by both signals. If the cathode ray tube is oriented appropriately, the light point indicates to some degree direction and intensity of the stereo information.

Correlation of stereo signals may be an important piece of information. In particular, the correct polarity of the signals can be indicated. A simple correlation meter is an "exclusive or" circuit fed with the polarity information of the stereo signals. The integrated output of the circuit is an approximation of the correlation factor.

4.3 Line Transmission

The large dynamic span of audio signals of about 120 dB, corresponding to electrical signals ranging from a few microvolts to about 100 V, makes lines susceptible to cross talk and electromagnetic stray fields. For this reason, proper shielding of individual cables is essential. The most common measures are balanced and floating lines with twisted wires and shielding conductors. Floating signifies, in this context, that there is no direct connection to ground to which shields are connected. In balanced wiring, the induced currents cancel the influence of the stray fields, and the shield additionally also reduces this influence. Since many electronic devices are unbalanced, coupling transformers are used, as shown in Fig. 24 (Whitlock, 1995; Muncy, 1995; Macatee, 1995). With operational amplifiers, balanced circuits can be realized as well, and thus, transformers may sometimes be dispensed with.

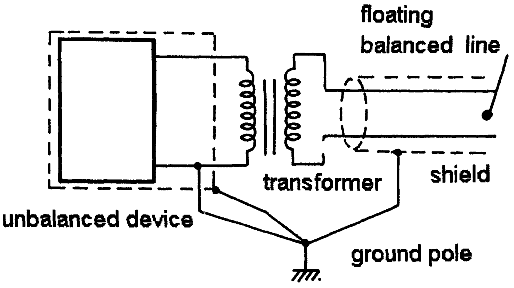

FIG. 24. Transformer outputs or inputs for balanced lines and shielding are used to avoid disturbances from external electromagnetic fields.

Since shields have a certain resistance, currents (either induced by stray fields or potential differences) between different devices must be carefully avoided. For this reason, shields should be grounded only at one end of the line, and all shields should be connected to a central ground pole.

5. DIGITAL AUDIO

The representation of signals by numbers has many advantages. Numbers can be transmitted and stored with fewer errors. If errors do occur, they can be recognized and corrected algorithmically. The development of high-speed signal processors has made it possible to handle vast amounts of multiple-channel audio data.

5.1 Data Sampling

In order to obtain a digital representation of analog signals, these signals must be sampled in equal time intervals T. The sampling process may be modeled as the multiplication of the analog signal $a(t)$ with a sampling signal $s(t)$. The sampling signal is a sequence of small impulses with duration δt and frequency $f_s = 1/T$. With $\beta = \delta t/T$, the spectrum of a sampled sine wave with frequency f is, according to Fourier's theorem,

$$S(\omega) = \beta a_0 \sum \sin[(n\omega f_s + \omega) + \sin(n\omega f_s - \omega)] \quad \text{with } -\infty \leq n \leq +\infty. \tag{23}$$

With a full audio spectrum, this sampling yields a spectrum that is characterized by partial spectra symmetrically to integer multiples of the sampling frequency f_s, as shown in Fig. 25(a).

The envelope of the complete spectrum is determined by the form and duration of the sampling impulse. A rectangular impulse with duration δt gives an envelope that follows the function $y = \sin(x)/x$, where $x = \pi f \delta t$. As can be seen from Fig. 25, a unique representation of the signal exists only if the audio spectrum is in the range $0 \leq f \leq \pm f_s/2$. This is the essence of the Shannon theorem. If signals have higher frequencies, these frequencies must be suppressed by filters; otherwise, so-called aliasing frequencies will occur. Antialiasing filters must have steep slopes above their cutoff frequency because the frequency-cutoff interval is relatively small for the sampling frequencies commonly used (32, 44.1, 48, or 50 kHz).

The reconstruction of a sampled signal gives back a sequence of impulses that, in order to have sufficient signal energy, may be prolonged to a step function. The spectrum is thus of the same form as with sampling and has a null at $1/T$ and an attenuation of 4 dB at the cutoff frequency $f_s/2$. Although the frequencies above the audio range are inaudible, they must be cancelled by analog reconstruction filters. Otherwise, they may modulate analog devices as well as the desired signals.

5.2 Oversampling

Oversampling avoids costly and noise-producing reconstruction filters (Hauser, 1991; Harris, 1993). In n-fold oversampling, the sampling interval is filled with $n - 1$ equidistant zeroes, e.g., three zeroes for a fourfold oversampling [Fig. 25(b)], giving a new sampling rate f_0. Weighting the new sample sequence with a finite impulse response (FIR) filter (see Sec. 5.4) gives an interpolation of the original sequence. Consequently, the spectrum shows high signal values only in the vicinity of multiples of the new sampling rate f_0, while in the intermediate frequency range, the values are low [Fig. 25(c)]. The cutoff slope of the analog reconstruction filter can then be relatively gentle.

5.3 Analog-to-Digital and Digital-to-Analog Converters

The sampling process provides, for the duration of one sampling period T, a voltage by means of a hold circuit. This voltage is

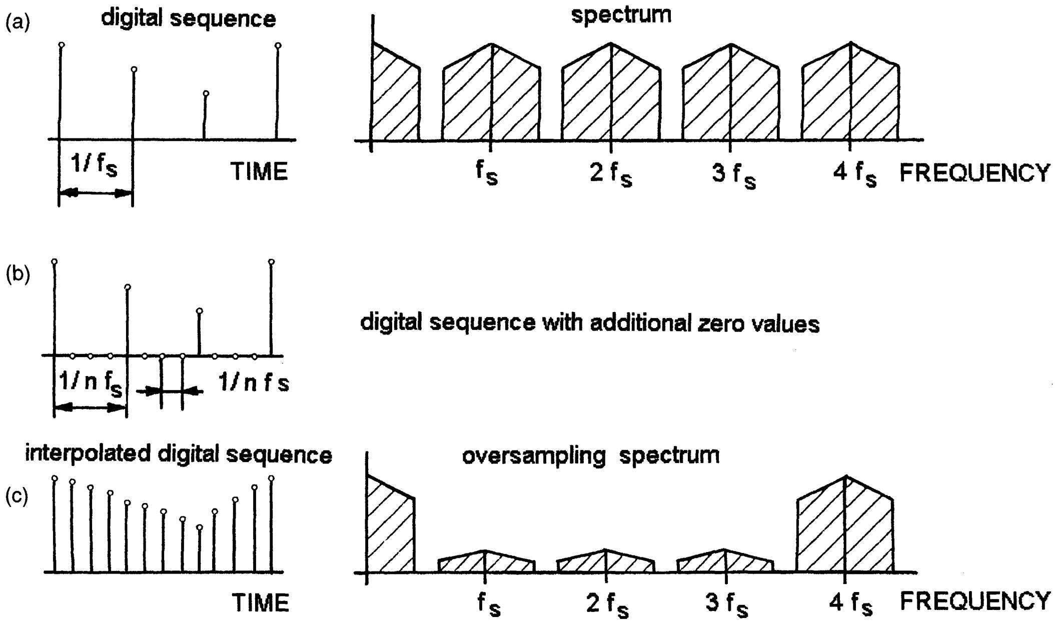

FIG. 25. (a) Spectrum of a sampled signal. (b) With oversampling and digital interpolation, the frequencies above the audio spectrum are suppressed (c).

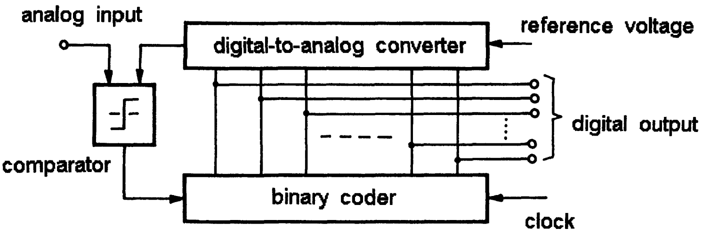

FIG. 26. Structure of an analog-to-digital converter.

compared with a stepped voltage derived from a reference voltage. The steps follow a binary sequence, and the result of the comparison is registered as a chain of N bits "1" or "0," where N determines the resolution of the quantizing process in bits. In digital audio, N is often 16 bits, but for better resolution, it may be extended to 20 or 24 bits. The stepping circuit is by itself a digital-to-analog converter in a feedback loop with the comparator, as shown in Fig. 26. The digital-to-analog converter weighs the currents in a network of resistors that are graduated as powers of two. According to the digital data word, the currents are summed and then transformed into a voltage (Sheingold, 1986).

Analog-to-digital conversion introduces an error that is determined by the quantizing interval q. This is normally a white noise. The error signal voltage U_q is the difference between the analog signal voltage U_s and the reconstructed analog signal voltage U_r. With respect to the maximum sine, voltage swing $U_s = q \times 2^{N-1}$, which, with respect to the maximal sine voltage swing, gives a signal-to-noise ratio of

$$\mathrm{SNR} = 20 \log(U_s/U_q) = 6.02N + 1.78 \text{ dB}. \tag{24}$$

For 16-bit conversion, an SNR from Eq. (24) of about 98 dB results.

The time-consuming stepwise comparison process can be speeded up by a faster 1-bit conversion. This is, in principle, a slope tracking process and may be realized with the circuit shown in Fig. 27. The input signal x is fed into an adder. The output of the comparator is connected with a filter having the gain g and the transfer function $H(f)$, and a 1-bit analog-to-digital converter that produces the digital output y with the quantizing error g. Via a digital-to-analog converter, y is fed back with negative sign to the adder. For y there follows

$$y = \frac{xgH(f) + q}{1 + gH(f)}, \tag{25}$$

giving $y \approx x$, if $gH(f) \gg 1$. By appropriate choice of $H(f)$, the spectrum of the quantizing noise can be shifted to inaudible frequency ranges. This is called *noise shaping* (Lipshitz *et al.*, 1991).

The quantizing noise, although it has a rather low level, may still be audible. Low-frequency signals that normally are below the threshold of audibility will produce wide-band noise by the quantizing steps. Noises of this type are called *granular noises*. The annoying effect of granulation can be diminished by adding low-level dither noise to the analog signal (Schuchman, 1964; Maher, 1992). Dither will also improve low-level signal resolution by reason of the complex properties of the ear (see ACOUSTICS, PHYSIOLOGICAL; also Sec. 5.5). With higher resolutions such as 20 bits per sample or more, the granular noise will no longer be a severe problem.

5.4 Digital Signal Processing

With digital signal processors or high-speed computation, all signal-forming processes can be carried out. In audio technol-

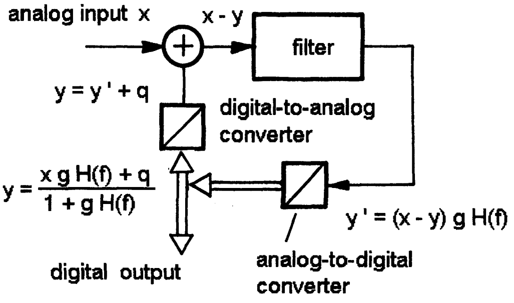
analog input x
x - y
filter
y = y' + q
digital-to-analog converter
y = (x g H(f) + q)/(1 + g H(f))
y' = (x - y) g H(f)
digital output
analog-to-digital converter

FIG. 27. Structure of a 1-bit analog-to-digital converter.

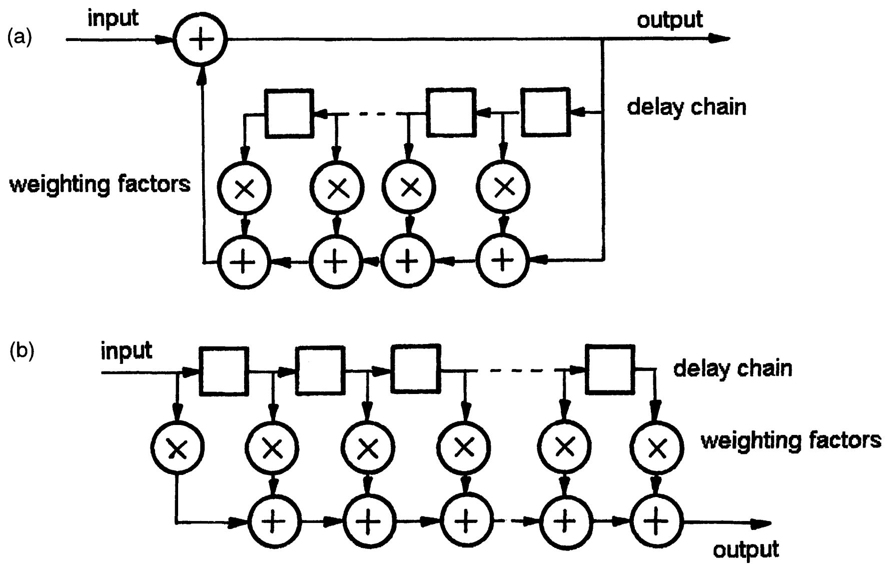
(a)
input
output
delay chain
weighting factors
(b)
input
delay chain
weighting factors
output

FIG. 28. Structure of (a) an IIR filter and (b) a FIR filter.

ogy, filters are particularly important components. There are two types of digital filters. In infinite impulse response (IIR) filters, a number of output values are fed back with delay and processed repeatedly for, at least theoretically, an infinite time. This is shown schematically in Fig. 28(a). The weighting of the output samples with factors a_i determine the frequency response of the filter.

In finite impulse response (FIR) filters, the sampled values pass through a delay chain with taps. Here, the output is the sum of values weighted by factors b_i, as is displayed in Fig. 28(b). The FIR and IIR filters may be combined according to the desired frequency response. Very complex filtering processes, such as the simulation of natural reverberation (see Sec. 7.1), can be realized. A possible scheme is given in Fig. 29. This comprises a FIR filter for early reflections and a multipath IIR filter for diffuse reverberation. With digital signal processing, a multitude of signal forming can be performed, e.g., frequency shifting, comb filtering, and especially signal delay.

5.5 Audio Data Reduction/Perceptual Coding

The perception of sound is characterized by a complex neural coding process, which integrates sound energy in small-frequency bands, denoted as *critical bands*. The com-

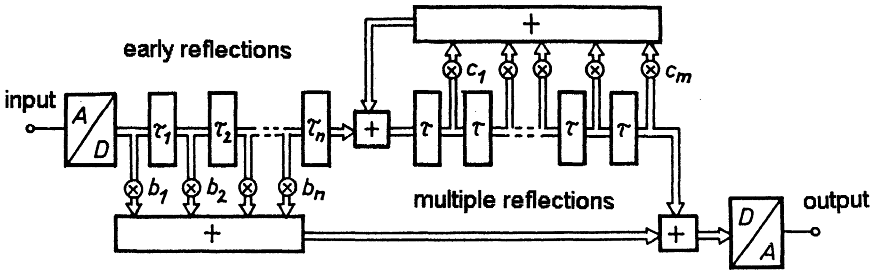
early reflections
input
A/D
τ1
τ2
τn
b1
b2
bn
+
c1
cm
+
τ
τ
τ
τ
multiple reflections
+
D/A
output

FIG. 29. Example of a reverberation filter (synthetic reverberation) with a combination of FIR and IIR filters.

bined neural excitations in the critical bands determine the loudness and timbre of sounds. Of special interest for data reduction is the phenomenon of masking. Masking refers to a shift of the threshold of audibility with respect to intensity and time. Sound components below the masked threshold are inaudible and may be ignored, while sound components whose levels are near the threshold can be coded with fewer bits than the more important components.

Data reduction is carried out in the frequency domain of the signals. The audio signal is split into a number of subbands while a spectral analysis is performed in parallel by a digital Fourier transform (DFT). With the result of the spectral analysis, the threshold of audibility is calculated according to the psychoacoustic model implemented, and data from the filter bank are requantized by bit allocation. Together with the requantized data, coding information is transmitted.

Decoding uses the coding information in order to regain the data, which then, by inverse filtering, gives the sound signal. This signal differs, of course, from the original one, but the coding error is a noise that is masked by the transmitted sound signal. Through this process, the data stream of one audio channel can be reduced from about 700 to 368 kb/s, or even to 64 kb/s with decreased sound quality.

There are many different coding schemes used in audio data recording and transmission, such as MUSICAM, ATRAC, ASPEC, and others. The standardization is specified by MPEG (Motion Picture Expert Group), which defined different so-called layers depending on the complexity of the coding process (Theile and Stoll, 1992).

6. AUDIO SIGNAL RECORDING

6.1 Analog Disc

6.1.1 Disc Recording In analog disc recording, information is stored mechanically by cutting a groove with a lathe into a lacquer surface. In a multistep, electroforming process, copied forms are produced to mold the vinyl discs. The cutting stylus has a triangular cross section with the back angle γ and the front edges inclining at 90°, as shown in Fig. 30. The form of the stylus limits the signal slopes to be recorded. The maximum lateral and vertical cutting velocity v_c depends on the travel speed v_t and is $v_c = v_t \tan \gamma$.

Since the angular speed of analog discs is constant at $33\frac{1}{3}$ or 45 rpm, the travel speed of the inner groove is the most critical case and has to be about 190 mm/s. To obtain a sufficiently long playing time, the separation of the grooves must be as small as possible with respect to a sufficient dynamic range of the record. This depends on the ratio of maximum signal amplitude and the surface roughness of the walls of the groove caused by the cutting stylus and the electroforming process. Peak excursions of the stylus are

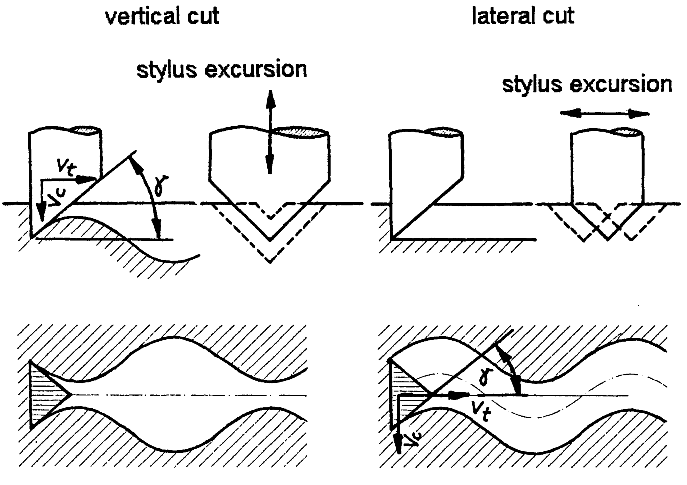

FIG. 30. Geometry of the analog-disc cutting stylus. The form of the groove depends on the movement of the stylus. In stereophony, lateral and vertical movements occur simultaneously.

limited to about 40 μm, and the dynamic range is about 55 to 60 dB. Playing time can be prolonged if the pitch of the tracks is varied with the stylus amplitude.

While in monorecording the stylus moves only laterally, in stereo the signals of each of the two channels are cut separately into the walls of the groove. If both signals are equal and in phase, the excursions of the stylus are lateral. If they are 180° out of phase, the excursions are vertical. To make best use of the recording area on the disc, cutting of the signal amplitude would be optimal. However, since cutting velocity v_c and amplitude a_c are connected by $v_c = \omega a_c$, the amplitude has to be limited at higher frequencies with respect to the maximum stylus velocity. Therefore, characteristics for cutting velocity are internationally standardized. They are characterized by time constants that define cutoff frequencies f_c of first-order filters, as can be seen in Fig. 31. Time constants are computed from $\tau = 1/2\pi f_c$. The boost of high frequencies is necessary because of the increase of noise in this range. This takes into account that most speech and music signals have a roll-off of about −10 dB/octave above 1 kHz. Signals with great high-frequency energy must, however, be handled with care.

Noise in disc recording can be minimized by a careful design and grounding of the cutting stylus. Noise is substantially reduced if the stylus is heated by a coil wound near the tip. The heating current has to be optimized and controlled with respect to the travel speed of the groove.

The electroforming process also adds some noise caused by inner tensions of the plating metals. Direct metal mastering (DMM), in which the signal is cut in a copper master, reduces the number of electroforming processes. This improves the dynamic range by about 6 dB.

Ordinarily, the signals for disc recording come from master tapes that also produce noise. Disc recording noise is normally lower than analog tape noise. Before digital tape recording came into general use, there was some interest in direct-to-disc recording. In this case the cutting process cannot be interrupted, and corrections cannot be made afterward.

6.1.2 Disc Playback

6.1.2.1 Pickup Styli. Whereas the cutting stylus is *V* shaped, pickup tips for playback have rounded-off edges so as not to damage the groove walls. The consequences of this geometric difference are tracking errors. One effect is "clamping" since the walls of the groove incline at 90° only if the groove has tangential direction; otherwise, the angle is smaller, and the tip moves upward at double the frequency recorded. In stereo records this gives records a second-harmonic lateral signal.

Another clamping effect occurs if the tracking radius of the tip is greater than the corresponding radius of the groove. Since each stereo channel is cut into one wall of the groove, the center of the tip does not follow the curvature. Instead, additionally to the signal contour, it moves up and down as

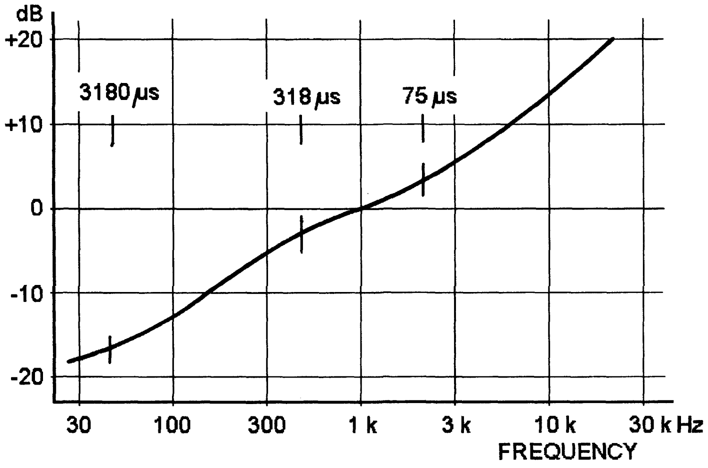

FIG. 31. Standardized frequency response of stylus velocity for disk cutting.

well as back and forth. The playback tip is attached to the end of a cantilever and moves in a circular trajectory. In contrast to this, the recording stylus moves linearly, and as a consequence, the playback signal is different from the recorded signal. If all geometric conditions were known, the distortions could be calculated and taken into account in recording. However, many distortions can be reduced by using an elliptical or biradial stylus because it follows the curvature of the groove much better than a simple rounded tip.

6.1.2.2 Pickup Cartridges. In principle, any physical effect that transforms an excursion into an electric current or voltage may be used for pickups. Magnetic and piezoelectric cartridges are widely used. Other cartridges work like condenser microphones or have semiconducting sensors. Laser scanning has also been proposed.

Magnetic cartridges make use of the induction law. There are moving-magnet or moving-coil systems, and the induced voltage is proportional to the velocity of the stylus. For stereophonic reproduction, the orthogonal movements of the stylus are captured by two orthogonally arranged coils. To separate the stereo information exactly, the angle between the axes of the coils as well as the magnets should be 90°, and the orientation of the cartridge with respect to the groove must be maintained. Even tiny angular deviations generate cross talk. Therefore, stereo channel separation is seldom better than about 30 dB. The output voltage of moving-magnet or moving-coil cartridges is less than 1 mV at the reference velocity and needs amplification. Because of the cutting characteristic (see Sec. 6.1.1), it also requires equalization.

The mass of the moving system must be less than about 1 mg, since together with the stiffness of the disc material a resonance frequency of at least 18–25 kHz is desirable.

6.1.2.3 Tone Arms. There are two categories of tone arms: those with a pivot and those with tangential guiding. In the latter category the pickup stylus moves in the same manner as the cutting head stylus: motion is radial across the disc surface and thus avoids tracking errors in that respect. The stylus of the pivoted tone arm moves on a circular trajectory, and therefore the angle between the directions of the groove and the excursion of the stylus changes. This causes distortions as well. These deviations can be minimized either by long tone arms or for short ones by introducing an offset angle α. For a given short tone arm, there exists an optimal pivot point where the angular deviations are minimized. Introducing the offset angle α develops a so-called skating force. This drives the tone arm toward the center and thus gives an offset excursion of the stylus. This may cause insufficient tracking as the stylus may lose contact with one groove wall. Therefore, antiskating forces can be applied. These are generated, for example, by weights or pulleys. These must be adjusted carefully since the skating force depends on stylus form and the friction between stylus and disc.

Equivalent tone-arm mass, and the compliance of the pickup system, form a resonator with resonance frequencies between 10 and 30 Hz. Oscillations can be excited by shocks from the stand and by warps of the disk. The tone-arm resonance should lie below 16 Hz in order to give the least disturbing effects, and, in addition, it should be well damped.

6.2 Analog Tape Recording and Playback

6.2.1 Tape Recording Tape recording is based on magnetic hysteresis. Magnetic substances are characterized by their permeability μ, which describes the relation between the magnetic flux density B and magnetic field strength H:

$$B = \mu H = \mu_0 \mu_r H, \tag{26}$$

where $\mu_0 = 1.256 \times 10^{-10}\ H/\mathrm{m}$, and μ_r = relative permeability. In ferromagnetic substances μ_r is high, while for most other materials it is almost 1. If ferromagnetic material with the cross section S is exposed to a magnetic field with the field strength H, a magnetic flux Φ with the flux density $B = \Phi/S$ develops. Starting from the magnetic neutral state, the flux density follows the so-called virgin curve, which asymptotically approaches saturation. If the field strength is then reduced to zero, a remanent flux density B_r will remain [Fig. 32(a)]. An alternating saturating magnetic field causes the hysteresis loop.

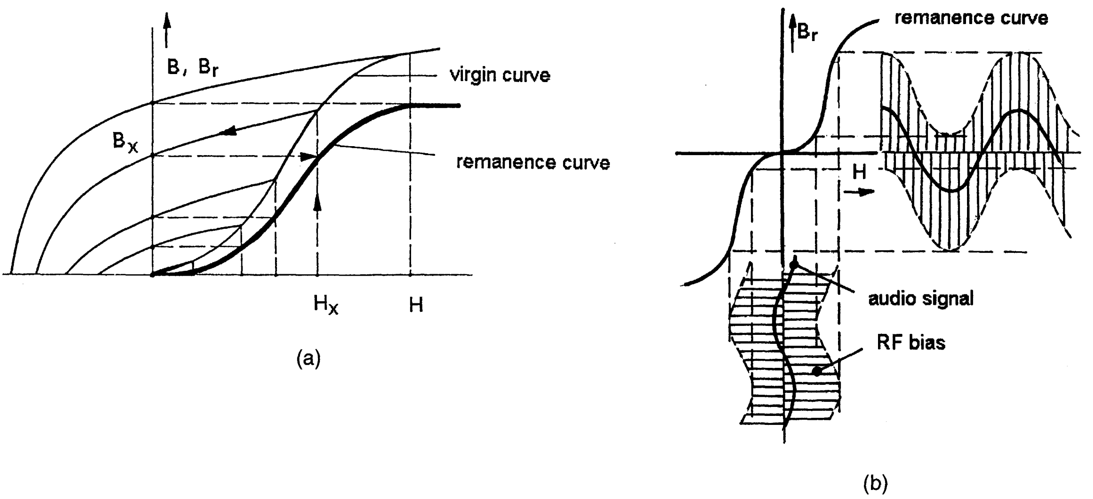

FIG. 32. (a) Hysteresis curve and remanent flux density B_r with respect to field strength H, and (b) the linearization effect of superposed RF (after Camras, 1949).

For recording, the remanent flux density as a function of field strength is important. Its nonlinearity is overcome by a high-frequency bias current that is superposed on the audio signal, as shown in Fig. 32(b) (Camras, 1949).

Magnetic tapes consist of a substrate—for example, polyester—coated with a layer of fine magnetic particles such as ferrooxide, chrome dioxide, or ferric alloy powder dispersed in a binder. The relative permeability of the magnetic layer is about 2–5 and thus much less than that of the record head, which is more than 10 000. The record head has a gap, and its magnetic field penetrates the ferromagnetic layer. The field strength is proportional to the current I in the winding of the head, $H = nI$, where n is the number of turns of the winding. The audio information is laid down at the trailing edge of the gap. Sensitivity, maximum output for prescribed nonharmonic distortions (commonly 3% THD), third-harmonic distortions, and tape noise are dependent on the high-frequency bias current, which is commonly referred to a 1-kHz maximum of sensitivity. The optimal operating bias is somewhat above the reference current and is recommended by tape manufacturers.

6.2.2 Tape Playback For playback a similar or even the same magnetic head is used. In high-quality recorders there are separate heads, with the record head having a wider gap. The gap of the playback head determines a sensing window, and the magnetic flux through the head is a function of the ratio of the recorded wavelength and the gap width w. If the window is presumed to have a rectangular form, the flux is proportional to the function $y = \sin(x)/x$ with $x = \pi w/\lambda$ or, with the tape speed v_t, $x = w\omega/2v_t$.

Each thin partial layer of the magnetic coating contributes to the magnetic flux gathered by the head. However, this is dependent on their distances d from the head and the wavelength. The spacing loss follows the equation

$$y = \exp(-2\pi d/\lambda). \qquad (27)$$

With a tape flux Φ_t, the sum over the thickness d_0 of the coating gives the spacing loss

$$L_s = \Phi_T \frac{[1 - \exp(-2\pi d_0/\lambda)]}{2\pi d_0/\lambda}. \qquad (28)$$

In Fig. 33 spacing loss is drawn as a function of wavelength or frequency, respectively. It can be seen that a thin coating gives a better high-frequency recording but at the cost of less magnetic flux (Krones, 1960). The spacing loss also comes into play when the tape is contaminated with small particles on its surface. Since the effect is dependent on

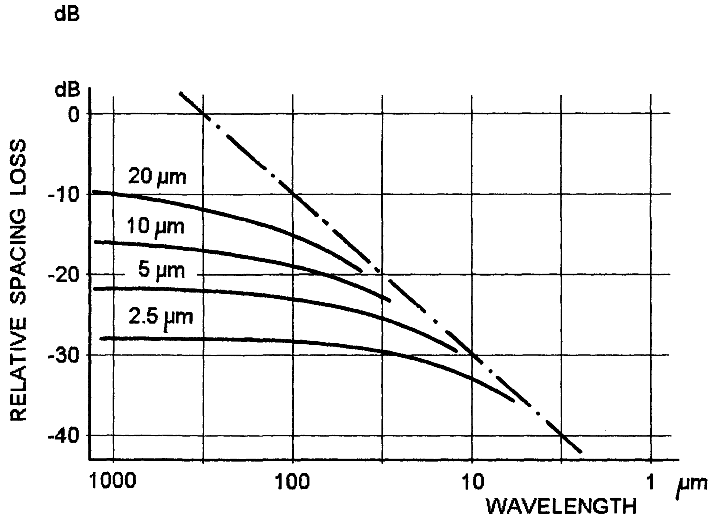

FIG. 33. Spacing loss in magnetic tapes with respect to the recorded wavelength and the thickness of the magnetic layer. The thinner the magnetic layer, the better the high-frequency resolution at the cost of reduced magnetic flux.

the wavelength of the recorded sound, low-speed tape recordings will be more affected by dropouts. Contamination is avoided by the use of cartridges.

Only about 50–70% of the tape flux gathered by the head will permeate the winding, while the rest will pass the gap directly. A sinusoidally alternating flux $\Phi_t = \Phi_t \exp(\omega t)$ on the tape induces a voltage U in the winding with n turns:

$$U = (0.5\text{–}0.7) n \omega \Phi_t \times \frac{\sin(w\omega/2v_t)}{w\omega/2v_t} \frac{1 - \exp(\omega d_0/v_t)}{\omega d_0/v_t}. \qquad (29)$$

Additional losses in recording and playback are hysteresis and eddy currents, which are nonlinearly dependent on frequency and field strength. All frequency-dependent effects must be compensated for during recording and playback.

6.2.3 Recording and Playback Equalization In order to achieve interchangeability, the linearity of playback frequency response is checked with standard tapes. These tapes are recorded with a standardized record head and a high-frequency pre-emphasis characterized by a time constant (see Sec. 6.1) that compensates, to some degree, the spacing losses. For 76.2 and 38.1 cm/s (30 and 15 in./s), the time constant is 35 µs. For compact cassettes (CC), according to IEC I and IEC II, the time constants are 120 and 70 µs, respectively. For low tape speeds there is also a low-frequency boost in order to enhance the induced voltage for a sufficient signal-to-noise ratio. Recording pre-emphasis is allowable because, in most cases, speech and music signal amplitudes in the boosted ranges are low. Standard tapes contain sections for gain adjustment at full level, alignment of the azimuth of the playback head, and equalization of the frequency response. A check of the playback amplifier follows the recorded sequence on the standard tape.

The playback amplifier has to equalize all losses of the playback head, such as gap losses, hysteresis and eddy current losses, and the frequency response of the induction law. The recording amplifier with a bias generator is adjusted to the tape that will be used after the prescribed bias. Gain and frequency response are then corrected in order to obtain an overall linear 1:1 transmission.

6.2.4 Tape Noise Tape noise is caused partially by nonregular dispersion of magnetic particles and tiny distance fluctuations between the tape and heads, which modulate the signals during recording and playback. Since steady magnetic fields raise tape noise, all parts of the tape recorder, including the heads, must be demagnetized carefully, at least before the adjustment procedure. Tape noise may also arise if the bias current has second-order nonlinear distortions.

Signal distortion in tape recording is also produced by print-through. The magnetic field of the recorded signals permeates the layers of the wound-up tape. The thickness of the tape d_t gives a distance effect, and there is a frequency dependence with a maximum at $\lambda = 2\pi d_t$. Print-through is a time-dependent process where the level increases about 0.5 dB with each doubling of time. The effect is enhanced by external alternating magnetic fields and high temperatures. It causes "echoes," which, if they occur in advance of the original signal, are rather annoying. Since trailing echoes normally are not as annoying, tapes may be stored "tail out." That is, they are not wound back after recording. There may also be echo erasers that apply a weak erasing field to the tape before it passes the playback head.

6.3 Analog Recording and Playback Drives

Drive mechanisms for disc and tape recording and playback must be designed carefully. Variations in the number of rotations of the turntable or the capstan may arise from inconstant motor torque, inconstant friction in bearings, and inaccurate centering. These cause frequency modulation perceived as wow and flutter. The ear is most sensitive to frequency modulation at about 4 Hz (wow). Flutter is a modulation with frequencies above 16 Hz. Deviations of about 0.3% from the regular speed are audible. Rotating members of the drive should therefore avoid numbers of rotation at about 4 Hz. Higher frequencies can be suppressed by mechanical filtering such as belt drives. Disks should be level, and turntables should also be carefully isolated against shock and oscillations from the surface on which they stand. Oscillations of this kind will actuate the pickup so much that it will operate within the nonlinear part of its mechanical characteristic, causing amplitude modulations, to which the ear is as sensitive as it is to frequency modulation.

In tape recording, the tape guides must be arranged in such a way that azimuth errors are kept as low as possible. This is achieved by guides that force the tape to travel against one of the guide edges. The friction between tape and tape guide may excite tiny longitudinal oscillations of the tape at about 5 kHz (depending on the mechanical properties of the tape). Modulation by-products, especially difference frequencies, enhance playback noise. At least in some studio tape recorders, oscillations are suppressed by silencer rollers.

6.4 Digital Sound Recording

While in digital signal processing only relatively simple error-detecting and -correcting methods are needed, in digital recording great efforts must be made in order to overcome possible error sequences caused by dropouts. In most cases the errors are burst errors, i.e., the breakdown of the bit stream. Therefore, it is necessary to add a number of parity bits or redundancy to the most significant data bits as is done with the Wyner-Ash-Code, shown in Fig. 34(a). Since burst errors destroy adjacent bits, it makes sense to interleave data bits according to some method such that the original data can be reconstructed with the help of parity bits. This is shown in Fig. 34(b). In cross redundancy check, the data bits are arranged into a two-dimensional array, and parity bits are assigned to lines and columns. The cyclic redundancy check is based on the polynomial representation of data words. N-bit data words to be guarded with P bits are shifted left by P digits, giving an $(N + P)$-digit word. Division by a P-digit prime number yields a remainder R that is added to $N + P$. Then, division by the P-digit prime number leaves no remainder. If transmission errors occur, the division yields a remainder that points at the erroneous bits. Prime numbers can be chosen so that burst errors are corrected effectively (Peterson and Weldon, 1980). There is a multitude of error-correcting schemes combining the methods mentioned.

If error-correcting schemes do not suffice to reconstruct the data words, error concealment will be activated. Here, either the preceding data word is repeated or an interpolation between the preceding and following data words is executed. If a great number of samples drop out, the signal is muted.

6.4.1 Digital Disc Recording

6.4.1.1 Compact Disc. Recording is made with a modulated infrared laser beam that burns tiny holes, called "pits," along a spiral track into a metal-laminated master disc. The track velocity is constant, and one bit is about 0.83-μm long and 0.12-μm deep.

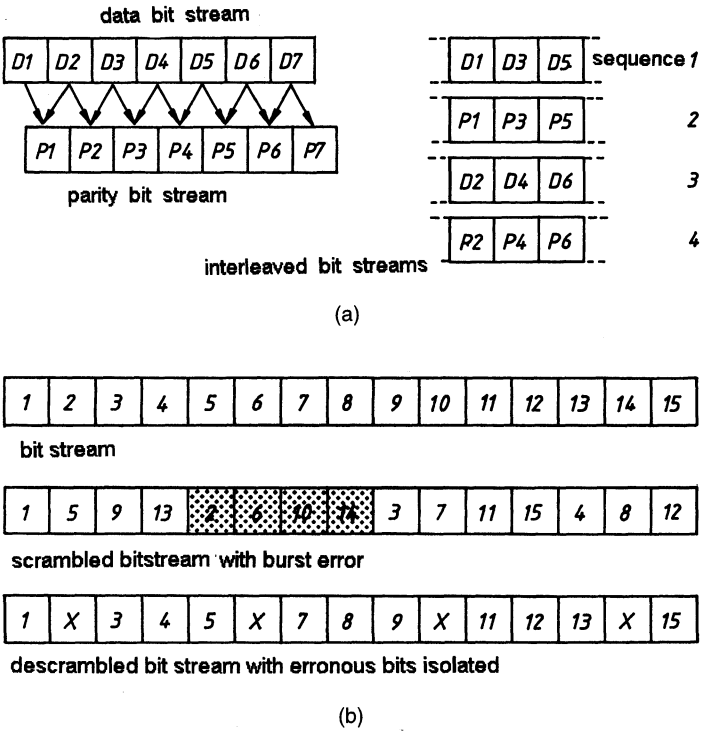
data bit stream
D1 D2 D3 D4 D5 D6 D7
P1 P2 P3 P4 P5 P6 P7
parity bit stream
D1 D3 D5 sequence 1
P1 P3 P5 2
D2 D4 D6 3
P2 P4 P6 4
interleaved bit streams
(a)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
bit stream
1 5 9 13 2 6 10 14 3 7 11 15 4 8 12
scrambled bitstream with burst error
1 X 3 4 5 X 7 8 9 X 11 12 13 X 15
descrambled bit stream with erronous bits isolated
(b)

FIG. 34. Schemes for error-correction coding: (a) With the Wyner-Ash-Code, two adjacent data bits are connected to a parity bit. The distribution of bits on different tracks avoids loss of information by burst errors. (b) The information loss caused by burst errors can also be avoided by separation of adjacent data bits. Erroneous bits are isolated, and with parity bits, the information can be regained.

The track pitch is 1.6 μm. Preparing the discs for molding is nearly the same as in analog disc production. The molded disc is metallized and covered with a guarding layer, as shown in Fig. 35. The information is read by a laser beam that is focused at the surface of the pit layer.

Differences in reflection and diffraction of pits and the "land" between the pits produce modulation of the returning beam, which is sensed by photodiodes. Deviations of focusing are detected by photodiodes, and a correction current is fed into the lens-bearing coil. Tracking is achieved by splitting the laser beam into three beams, so that two of them keep track of the land between adjacent information tracks.

In compact disc (CD) recording, the sampling frequency is 44.1 kHz, and the data are quantized to 16 bits, giving a data stream of about 1.41 Mb/s. Encoding uses the eight-to-fourteen modulation. Here, each data word is split into two groups of 8 bits that are augmented to 14 bits in order to avoid excessively long strings of equal bits. Three additional bits are inserted between each block of 14 bits for the same purpose. Furthermore, there are added bits for synchronization, error correction, time coding, and user information. This gives, finally, an encoded data stream of about 4.32 Mb/s.

Maximum playing time of the standard CD is about 70 min. The recording density can be multiplied by reducing the pitch of the tracks. It may be possible in the future to use blue or ultraviolet lasers. The playing time also can be prolonged by audio data compression (see Sec. 5.5).

6.4.1.2 Magneto-optical Disc (MOD). Magneto-optical recording makes use of the effect that a magnetic field changes the plane of polarization of a reflected light beam. In order to magnetize a tiny spot of the magnetic layer, a laser beam focused on the surface demagnetizes the material by heating it above the Curie temperature. While the spot is cooling down, a magnetic field is applied, impressing the new data bits into the material. In order to read out the data, a polarized laser beam that will not heat the material above the Curie temperature is reflected with a polarization plane according to the

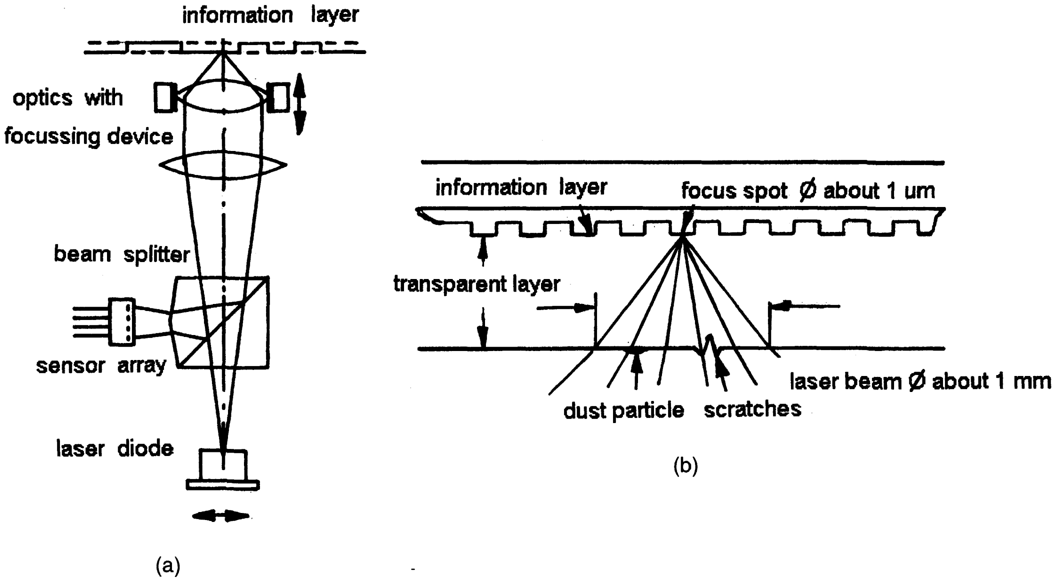

FIG. 35. (a) Main components of the optical system for CD playback. (b) Cross section of CD shows the relatively little effect on laser beam intensity caused by dust and scratches.

direction of the magnetic field. Passing a polarization filter, the beam is intensity modulated and can be evaluated as with the CD. MODs were first used in computer technology and are organized in addressable cylinders and sectors. This permits fast access to data. Since MODs are formatted and faulty sectors are marked, only minor error-correcting measures are necessary.

6.4.1.3 Mini Disc (MD). The mini disc relies, for playback of commercially produced disks, on a method similar to the CD. For home recording, a magneto-optical method is used. Since the recording area is about a quarter of that of a CD, and since the playing time is the same, a data reduction technique (ATRAC = adaptive transform acoustic coding) must be applied (see Sec. 5.5). The audio data rate is reduced to about 0.3 Mb/s. Cartridges are used in order to avoid dropouts caused by contamination. While playback tracking is analogous to the CD, the discs for home recording have tiny grooves for guiding the recording laser. Since the mini disc is designed also for nonstationary use, shocks may cause jumps of the laser system. Up to 3 s of sound are stored in a buffer, together with address information, so that a correct concatenation without clicks can be maintained.

6.4.2 Digital Tape Recording There are two approaches to digital tape recording: either by stationary heads or by rotating heads. Since the gap length of magnetic heads cannot be reduced below about 0.5 μm, the length of one bit is of the same order. On the other hand, the dynamic range to transmit a bit stream can be reduced to about 10 dB. This would be possible with 0.5-μm tracks. A 16-bit stereo data stream at a sampling rate of 50 kHz with a redundancy of 25% for error correction will give a data stream of 2 Mb/s. There are several error-correction schemes in use. They are well adapted to the tape error structures, such as dropouts.

For recording the audio and error-correction data, tape velocity, or the relative speed of the gap of the head with respect to the tape, must be about 1 m/s.

6.4.2.1 Stationary-Head Tape Recording. The tape velocity needed for digital recording of 1 m/s can be handled by stationary heads and the usual tape drives. With 38.1 cm/s (15 in./s), one track per channel is just feasible and is realized in DASH (digital audio stationary head) systems. Each halving of velocity doubles the number of tracks per channel. Multitrack recording is possible if the heads have an appropriate number of gaps. With respect to tape guiding, the track width cannot be reduced below about 0.1 mm. With thin-film heads on a 1" tape, up to 128 different tracks can be realized.

Stationary heads are used also in the dcc™ system, which has the standard tape velocity of music cassettes (MC) of 4.75 cm/s. Digital information is stored on nine parallel tracks. The audio data stream is compressed to 384 kb/s (see Sec. 5.5), which together with track, time, and system information, as well as error-correction data, yields a data stream of 768 kb/s. This can be recorded at this low tape velocity. The playback head, with thin-film technology, is designed for reading of either digital dcc recordings or analog music cassettes.

6.4.2.2 Rotating-Head Tape Recording. With a rotating head, the relative velocity of the heads with respect to the tape is about 5 m/s. This is used in analog video recording and is high enough for the demands of digital recording. Continuous audio signals must be transformed into discontinuous video signals. The commonly used sampling rate of 44.1 kHz is determined by the video system. For every video tape system (U-matic, betamax, S-VHS), there exists an audio recording scheme, some of which can handle up to 12 channels. The recording density in rotating-head systems is rather high since the tracks are small, e.g., 13 μm with the R-DAT system, and are directly adjacent to another. Cross talk between the tracks is avoided by an alternating azimuth, as shown in Fig. 36.

Besides audio error-correction data, there are also data for tracking, synchronizing, time, and titles stored. Multitrack recording is also possible with rotating heads.

7. ROOM ACOUSTICS AND SOUND REINFORCEMENT

7.1 Room Acoustics

Music and speech are usually recorded and reproduced in enclosed rooms. The acoustic properties of these rooms play an important role in defining the characteristics of the sound we hear (Cremer and Müller, 1982; see also ACOUSTICS, ARCHITECTURAL).

A sound impulse emitted by a source *S* arriving at the ears of a listener *L* is a composite of the direct sound and reflections from the walls, the ceiling, and the floor of the room. The reflections constitute the reverberant sound. Its fine structure is given first by a time gap between the direct sound and the first reflection. Second, other first-order reflections arrive according to the different travel paths. Third, higher-order reflections with increasing density appear and fill the room with a diffuse sound field, which dies away according to the degree of absorption at the walls. For different applications there are optimal reverberation times ranging from 0.3 s for speech-recording sound booths, 1.0 to 1.5 s for general-purpose auditoriums, and 1.8 to 2.4 s for concert halls and churches. In some general-purpose auditoriums and recording studios, the room acoustics are variable. This can be achieved by movable panels with a reflective surface on one side and absorbing material on the other, the ceiling may be lifted or lowered in order to change the volume of the room, or movable absorbers may be installed. While these methods can only shorten the reverberation time by about 50%, active systems with microphones and loudspeakers can enhance

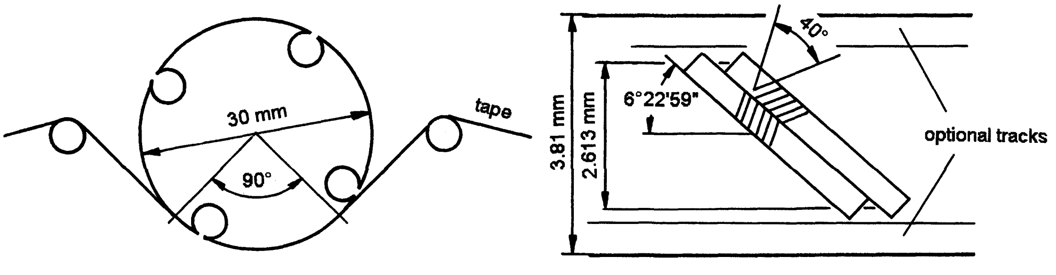

FIG. 36. R-DAT system: (a) Dimensions of the rotating four-head wheel. (b) The tracks for the digital information are densely packed. The alternating azimuth of the gaps of the heads reduces the cross talk between the tracks. On the edges of the tape, there are tracks for special purposes.

reverberation. Because of the danger of acoustic feedback, a large number of individual channels distributed in the walls and the ceiling are necessary, and the gain of each channel must be kept low. The systems can either work broadband, as with the Philips multiple-channel amplification of reverberation (MCR) system (de Koning, 1983), or narrow band, as with the assisted resonance (AR) system (Parkin and Morgan, 1970).

Sounds of longer duration stimulate standing waves. Especially in small rooms, low frequencies give pressure peaks and nodes at different listening locations. Standing waves can be reduced by careful acoustic design.

7.2 Sound Reinforcement

In large rooms or open air, sound amplification is needed to maintain intelligibility of speech. In the free field, as is nearly the case in open-air performances, the sound pressure decreases by 6 dB as the distance is doubled (see Sec. 2.1). In rooms, attenuation near the sound sources is almost the same as in a free field. But at greater distances, the diffuse reverberating sound field dominates, giving a sound level that is nearly independent of distance. Equality of direct and diffuse sound level defines a critical distance d_c in a room, which, with the volume V in cubic meters and the reverberation time T in seconds, can be calculated from

$$d_c = 0.057\sqrt{V/T}. \tag{30}$$

Audible echoes may annoy the listener. Echoes are usually audible if the reflected sound arrives more than 30 ms later than the direct sound, although there is some trade-off between echo level and arrival time (Haas, 1972; Wallach *et al.*, 1973).

Sound reinforcement aims, therefore, to enhance sound level. Under certain circumstances, sound reinforcement may also be used to delay direct sound in order to overcome the level of the diffuse sound and to mask echoes. In sound reinforcement, sound sources are picked up by microphones. Inevitably, some reinforced sound is fed back to the microphones, which, if amplitude and phase meet the proper conditions, gives rise to acoustic feedback or howling. Acoustic feedback can be avoided to a certain degree by the use of microphones and loudspeakers with appropriate directivity patterns. Electronic signal processing—e.g., frequency shifting—may also be used to suppress feedback.

Loudspeakers in sound reinforcement systems may be arranged as central clusters or they may be distributed on the ceilings or the walls. Central clusters are used in rooms with high ceilings, while distributed loudspeakers may be more useful in rooms with low ceilings. Mixtures of both methods are also possible, but if distances are long and the relative arrival times differ by more then 50 ms (equivalent to 17 m), electronic delay is advisable (Dibble, 1991). If the delay is correctly adjusted, the sound waves emitted from the nearest loudspeakers will act to support the attenuated direct sound waves.

The loudspeakers in a cluster must be carefully directed to the listening area in order to obtain relatively uniform coverage with respect to loudness. Of special interest are systems for open-air and stage performances because moving sources must also be amplified. To maintain information about the location of the sound sources, the levels and the delays of the loudspeakers must be balanced carefully, which, especially with wide stage dimensions, will be problematic (Teuber and Voelker, 1991). Although there do exist direction control systems, such as Delta stereophony (Steinke, 1983), in most situations recording engineers tend to do the job themselves.

7.3 Stereophony

7.3.1 Two-Channel Stereophony Sound waves that arrive at the ear drum carry the information of the location of sound sources (Blauert, 1983). This information is determined mainly by the directional dependence of diffraction and shadowing of the human head and the fine structures of the outer ear (pinnae). Head-related stereophony therefore uses microphones mounted in an artificial (dummy) head (Wilkens, 1972). Sounds can be reproduced only by headphones unless compensation for cross talk is provided for loudspeaker presentation. Standard stereophony is based on two-channel sound transmission and makes use of the fact that intensity differences at the two ears, as well as time delays, will give directional sensations. The standard stereophony setup is shown in Fig. 37(a).

The angle of location of speech and music caused by level differences and time delay is shown in Figs. 37(b) and 37(c), respectively. These relations can be produced by appropriate balancing of sound signals with panorama potentiometers in the mixing console (see Sec. 4.2) and/or by the arrangement of stereo and mono microphones in the recording studio.

Microphones for intensity stereophony are mounted as narrowly as possible and separate the sound sources from the left and right side according to pickup patterns, which may diverge between 90° and 150°. Microphones that utilize both intensity and time delay are mounted at a specified distance, also with diverging pickup patterns. In nearly all practical cases, stereo microphones are used as main microphones, while soloists and instrumental groups have additional microphones. Their signals are mixed into the channels by panning (see Sec. 4.2.1), or they may be recorded on separate tracks and later panned and mixed down to a final product. To have a more natural stereo effect, the signals of the additional microphones can be delayed so that the difference of the travel times with respect to the main microphone are compensated for. Stereophonic recordings may also be made by acoustic separation of instrumentalists by sound screens or booths.

7.3.2 Quadrophony and Surround Sound In order to improve sound reproduction and to provide more convincing spatial illusions through sounds arriving from the sides and rear, more channels are necessary. Four channels were originally deemed to be sufficient. For reasons of economy and compatibility, efforts were made to transmit four channels with two channels (Cooper and Shiga, 1972). This was achieved by matrixing the four signals front left (L_l), front right (R_f), rear right (R_r), and rear left (L_l). The rear signals were phase-shifted by 90° and added to the front signals. This resulted in considerable cross talk, which was greater when frequency and phase response of the transmission channel were unsatisfactory.

Many of the localization problems inherent in two-channel stereophony can be overcome by adding a third center channel. In cinemas, especially for large-screen pictures, three front loudspeakers are normally used. Two additional channels for loudspeakers at the sides and the rear produce special effects and reverberation information. The standardized surround scheme is shown in Fig. 38.

The Dolby surround system is widely used. This system encodes the surround signal similar to quadrophony matrixing and 90° phase shifting for two-channel transmission. In decoders, an adjustable delay of 15–30 ms in the surround signals masks the

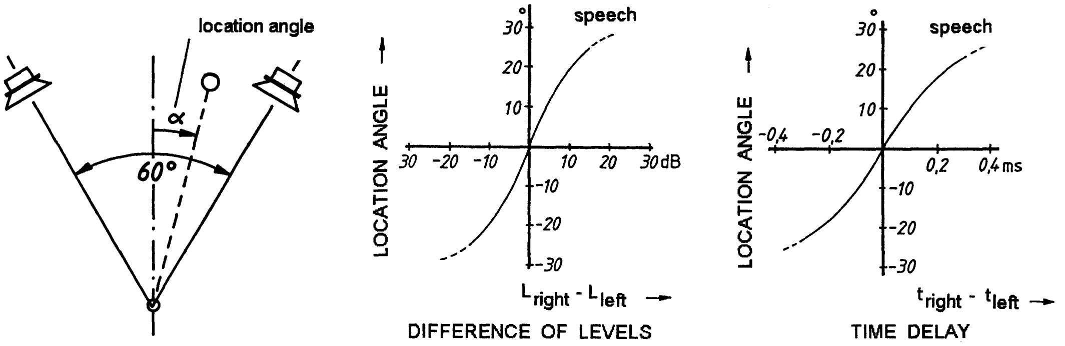

(a) (b) (c)

FIG. 37. Two-channel stereophony: (a) Standard loudspeaker arrangement. The location angle depends either on (b) level differences or on (c) time differences between the two channels. The location angle is also dependent on signal frequency. Here, the location angles for speech are shown.

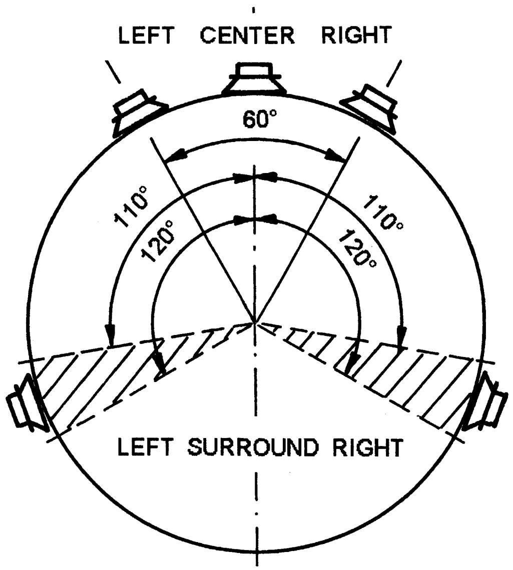

FIG. 38. Standard loudspeaker arrangement for 3/2 surround sound. This is characterized by an additional center loudspeaker and two loudspeakers for (mostly) the reproduction of ambient sound or spatial sound effects.

cross talk of front signals since, according to the Haas effect, the sound is located at the point from which the sound waves first arrive. A level-dependent enhancement of high-frequency components supports the surround localization. The Dolby Prologic™ decoder gives an active enhancement of directivity information. Coherent and noncoherent components can be derived by signal analysis for computing the dominance of one of the five channels. Based on this decoder scheme, different approaches are made in order to enhance further the localization effects.

7.3.3 Head-Related Stereophony Reflection, diffraction, and shadowing by the head, the structure of the pinnae, and the upper part of the body influence the intensities of the sound waves arriving at the ear drums. For each direction, a different frequency response can be measured. The head-related transfer functions (HRTF) are individual, and each person has gained lifelong experience deriving directional information from the intensity and time differences. Head-related stereophony uses artificial heads with pressure microphones at the places of the ear drums. This gives a rather good spatial impression, although the very important structure of the pinnae is not individually adapted. Localization can be improved by a correction of the transfer functions (Moller *et al.*, 1995).

Head-related stereophony may have future applications, such as in the simulation of sound fields and in virtual reality. Here, headphones combined with a motion sensor can produce control information in order to alter the HRTF with respect to the head movements.

7.3.4 Multiple-Channel Sound Transmission For some purposes, such as realistic reproduction of room acoustics without head-related stereophony, proposals have been made for sound reproduction with more than two channels. With sound-field microphones (see Sec. 3.1.5), the waves arriving from different directions can be picked up. Because of the directional patterns, the signals are weighted according to the angles of incidence. The sound field can be approximately reconstructed as it had been in the vicinity of the microphone by appropriate mixing or computing the microphone signals according to the positions of a great number of loudspeakers (Gerzon, 1973; Hensel *et al.*, 1992). At least four transmission channels are necessary, although for better spatial resolution, a greater number of channels must be provided.

Works Cited

Allison, R. F. (1974), "The Influence of Room Boundaries on Loudspeaker Power Output," *J. Audio Eng. Soc.* **22,** 314–320.

Blauert, J. P. (1983), *Spatial Hearing,* Cambridge, MA: MIT Press.

Camras, M. (1949), "Graphical analysis of linear magnetic recording using high frequency excitation," *Proc. IRE* **37,** 569–573.

Cooper, D. H., Shiga, T. (1972), "Discrete-Matrix Multichannel Stereo," *J. Audio Eng. Soc.* **20,** 346–360.

Cremer, L., Müller, H. A. (1982), *Principles and Applications of Room Acoustics,* Vol. 1, London: Applied Science Publishers.

Dibble, K. (1991), "The cathedral church of St. Michael Conventry—A sound system 30 years on," Preprint 3079 (I-3), 90th Convention of the AES, Paris.

Dolby, R. M. (1967), "An Audio Noise Reduction System," *J. Audio Eng. Soc.* **15,** 383–388.

Dolby, R. (1987a), "The Spectral Recording Process," *J. Audio Eng. Soc.* **35,** 99–118.

Dolby, R. M. (1987b), "An Audio Noise-Reduction System," *J. Audio Eng. Soc.* **35,** 99–118.

Gerzon, M. A. (1973), "Periphony: With-Height Sound Reproduction," *J. Audio Eng. Soc.* **21,** 2–10.

Gilchrist, N. (1994), "DRACULA: dynamic range control for broadcasting and other applications," *Report 18th Tonmeistertagung, Karlsruhe,* pp. 36–47.

Haas, H. (1972), "The Influence of a Single Echo on the Audibility of Speech," *J. Audio Eng. Soc.* **20,** 145–159.

Harris, S. (1993), "How to Achieve Optimum Performance from Delta-Sigma A/D and D/A Converters," *J. Audio Eng. Soc.* **41,** 782–790.

Hauser, M. W. (1991), "Principles of Oversampling A/D Conversion," *J. Audio Eng. Soc.* **39,** 3–26.

Hensel, J. *et al.* (1992), "Orthophonie—ein neues Aufnahme- und Wiedergabeverfahren zur Abbildung räumlicher Schallfelder," *Fernseh- und Kinotechnik,* 46. Jg., pp. 165–170.

Hibbing, M. (1984), "Überlegungen zur Entwicklung eines rauscharmen, hochaussteuerbaren Studio-Kondensatormikrofons," *Bericht 13. Tonmeistertagung, München,* pp. 85–94.

Klipsch, P. W. (1941), "A Low Frequency Horn of Small Dimensions," *J. Acoust. Soc. Am.* **13,** 137–144.

de Koning, S. D. (1983/84), "The MCR System—Multiple-Channel Amplification of Reverberation," *Philips Tech. Rev.* **41,** 12–23.

Krones, F. (1960), "Die Theorie des Magnetspeichers," in: F. Winckel (Ed.), *Technik der Magnetspeicher,* Berlin: Springer, pp. 381–473.

Lipshitz, S. P. *et al.* (1991), "Minimally Audible Noise Shaping," *J. Audio Eng. Soc.* **39,** 836–852.

Lord Rayleigh (1945), *The Theory of Sound,* Vols. 1 and 2 (reprint), New York: Dover.

Macatee, S. R. (1995), "Considerations in Grounding and Shielding Audio Devices," *J. Audio Eng. Soc.* **42,** 472–483.

Maher, R. C. (1992), "On the Nature of Granulation Noise in Uniform Quantization Systems," *J. Audio Eng. Soc.* **40,** 12–20.

Million, A., Krause, M. (1985), "Intelligenter automatischer Kompander mit großen Zeitkonstanten," *NTG-Fachberichte* Bd. 91.

Moller, H. *et al.* (1995), "Design Criteria for Headphones," *J. Audio Eng. Soc.* **43,** 218–239.

Moller, H. *et al.* (1995), "Head Related Transfer Functions of Human Subjects," *J. Audio Eng. Soc.* **43,** 300–321.

Muncy, N. A. (1995), "Noise Susceptibility in Analog and Digital Signal Processing Systems," *J. Audio Eng. Soc.* **42,** 435–453.

Musialik, C. M., Knaeple, T. (1991), "A user configurable digital audio mixing console for broadcast applications," Preprint 3043 (G-4), 90th Convention of the AES, Paris.

Netzer, Y. (1981), "The Design of Low-Noise Amplifiers," *Proc. IEEE* **69,** 728–741.

Parkin, P. H., Morgan, K. (1970), " 'Assisted resonance' in the Royal Festival Hall, London," *J. Acoust. Soc. Am.* **48,** 1025.

Peterson, W. W., Weldon, E. J. Jr. (1980), *Error-Correcting Codes,* Cambridge, MA: MIT Press.

Rasmussen, S., Rasmussen, K. B. (1995), "On Loudspeaker Cabinet Diffraction," *J. Audio Eng. Soc.* **42,** 147–150.

Rubak, P. (1991), "Headphone signal processing system for out-of-the head localization," Preprint 3063 (K-5), 90th Convention of the AES, Paris.

Scherer, P. (1977), "Ein neues Verfahren der raumbezogenen Stereophonie mit verbesserter Übertragung der Rauminformation," *Rundfunktechn. Mittlg.* Nr. 5, pp. 196–204.

Schuchman, L. (1964), "Dither Signals and Their Effect on Quantization Noise," *IEEE Trans. Comm. Technol.* **COM-12,** 162–165.

Sheingold, D. H. (Ed.) (1986), *Analog-Digital Conversion Handbook,* Englewood Cliffs, NJ: Prentic Hall.

Small, R. H. (1973), "Vented-Box Loudspeaker Systems, Parts I, II, III and IV," *J. Audio Eng. Soc.* **21,** 363–372, 438–444, 549–554, 635–639; reprint in: R. E. Cookes (Ed.) (1980), *An Anthology of Articles on Loudspeakers from the Pages of the Audio Engineering Society,* 2nd ed., New York: Audio Engineering Society.

Smith, D. et al. (1983), "Improvements in Monitor Loudspeaker Systems," *J. Audio Eng. Soc.* **31,** 408–422.

Steinke, G. (1983), "Delta Stereophony—A Sound System with True Direction and Distance Perception for Large Multi-Purpose Halls," *J. Audio Eng. Soc.* **31,** 500–511.

Teuber, W., Voelker, E. J. (1991), "Acoustical stage support—design and application, Preprint 3052 (I-5)," 90th Convention of the AES, Paris.

Theile, G. (1986), "On the Standardization of the Frequency Response of High-Quality Studio Headphones," *J. Audio Eng. Soc.* **34,** 956–969.

Theile, G., Stoll, G. (1992), "MUSICAM-surround: A Universal Multi-Channel Coding System Compatible with ISO 11172-3," Preprint 3403 (J-5), 93rd Convention of the AES, San Francisco.

Thiele, A. N. (1971), "Loudspeakers in Vented Boxes, Parts I and II," *J. Audio Eng. Soc.* **19,** 382–392, 471–483; reprint in: R. E. Cookes (Ed.) (1980), *An Anthology of Articles on Loudspeakers from the Pages of the Audio Engineering Society,* 2nd ed., New York: Audio Engineering Society.

Villchur, E. M. (1957), "Problems of Bass Reproduction in Loudspeakers," *J. Audio Eng. Soc.* **5,** 122–126.

Wallach, E. B. *et al.* (1973), "The Precedence Effect in Sound Localization," *J. Audio Eng. Soc.* **21,** 817–826.

Webster, A. G. (1919), "Acoustical Impedance, and the Theory of Horns and of the Phonograph," *Proc. Natl. Acad. Sci.* **5,** 275–282.

Wilkens, H. (1972), "Kopfbezügliche Stereophonie—ein Hilfsmittel für Vergleich und Beurteilung verschiedener Raumeindrücke," *Acustica* **26,** 213–221.

Whitlock, B. (1995), "Balanced Lines in Audio Systems: Fact, Fiction, and Transformers," *J. Audio Eng. Soc.* **42,** 454–464.

Wuttke, J. (1992), "Microphones and Wind," *J. Audio Eng. Soc.* **40,** 809–817.

Further Reading

Ahnert, W., Reichert, W. (1981), Grundlagen der Beschallungstechnik, Stuttgart: Hirzel-Verlag.

Ballou, G. (Ed.) (1987), *Handbook for Sound Engineers, the New Audio Cyclopedia,* Indianapolis: Howard W. Sams & Co. Audio Library.

Rossing, T. D., Fletcher, N. H. (1995), *Principles of Vibration,* New York: Springer.

Zwicker, E., Fastl, H. (1990), *Psychoacoustics,* New York: Springer.

SOURCES, PARTICLE—ELECTRON

Richard L. Sheffield, *Advanced Accelerator Concepts Projects, Los Alamos National Laboratory, University of California, Los Alamos, New Mexico, U.S.A.*

Introduction 65
1. **Electron Sources** 65
1.1 Electron Thermal Temperature and Emittance 65
1.2 Thermionic Emission 66
1.2.1 Emission Characteristics 66
1.2.2 Metallic Emitters 66
1.2.3 Alkaline-Earth Oxide Emitters and Dispenser Cathodes 67
1.2.4 Compound Emitters 67
1.3 Photoelectric Emission 67
1.3.1 Low Quantum Efficiency 68
1.3.2 High Quantum Efficiency 68
1.4 Field Emission 69
1.5 Secondary Emission 69
1.6 Ferroelectric Emission 70
1.7 Plasma Emission 70
2. **Electron-Gun Design** 70
2.1 Thermionic and Long-Pulse Guns 70
2.2 rf Gun 72
2.3 Photoinjector 73
2.4 Pulsed Diodes 75
2.5 Virtual Cathodes 75
3. **Applications** 77
Glossary 77
Works Cited 78
Further Reading 79

INTRODUCTION

Electron beams are ubiquitous. Electron beams are used in high-energy physics, radar tubes, lithography, electron microscopes, cancer therapy and radiography, and many industrial processes. This wide range of uses has led to the extensive development of electron-beam sources and gun design.

The generation of low-voltage electron beams for electron tubes is well established and is now described in elementary college textbooks. Higher-power electron sources are used in present research activities and commercial applications. Even so, the physics and engineering of many of the electron-beam designs are closely related to that used in electron tubes.

This article covers the design of the higher-power electron sources. To design an electron source, a supply of electrons is required, and then a structure (an electron gun) must be built that can prepare the electrons to be either accelerated or directly used. A wide range—nanoamperes to megamperes—of current requirements might imply that a wide range of differing electron sources is required. However, many of the present electron sources can cover the wide range in current requirements, and only the gun design is modified for the differing currents. The choices of material, operating conditions, and performance for many sources are covered.

Since this article covers higher-current gun designs, most of the designs entail the use of very high voltages (tens of kilovolts to megavolts) across the cathode–anode gap. All of the gun designs use either a dc or an ac electric field to extract and further accelerate the electrons produced at the source. The key elements of the design, performance, and operational limitations are discussed.

1. ELECTRON SOURCES

1.1 Electron Thermal Temperature and Emittance

To understand the performance of an electron machine, we need to quantify the quality of the electron beam. Charged-particle beam quality is characterized by the ability to focus the beam to a small spot over a given distance. The measure of the quality, emittance, is calculated from the beam angle and radius.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

The beam root mean square (rms) angle x' with respect to the optic axis is defined by

$$x' = \left\langle\frac{dx}{dz}\right\rangle = \left\langle\left(\gamma m \frac{dx}{dt}\right)\left(\gamma m \frac{dz}{dt}\right)^{-1}\right\rangle = \left\langle\frac{p_x}{p_z}\right\rangle, \tag{1.1}$$

where m is the electron mass, γ is the relativistic factor $(1 - v^2/c^2)^{-1/2}$, p_x is the transverse momentum, p_z is the forward momentum, and angular brackets denote an ensemble average over the particle distribution. If the rms spot size at the beam focus is x, then the emittance ϵ is defined to be the product of x and x' at the beam focus:

$$\epsilon = xx'. \tag{2}$$

Thus by determining the emittance and setting the spot size, the optimal convergence of the electron beam is determined. The beam trajectory then determines the required electron-beam focusing system.

The normalized emittance is defined as the emittance times $\beta\gamma$, where β is the particle speed divided by the speed of light. The normalized emittance is an invariant (Lejuene and Aubert, 1980) during acceleration and beam transport if the particle beam experiences only linear forces.

The thermal limit of the normalized emittance of a beam from an emitter of radius r_c at a uniform absolute temperature T is

$$\epsilon_n = 0.5 r_c (kT/mc^2)^{1/2} \quad \text{(units: m·rad)} \tag{3}$$

because $\langle xx'\rangle = 0$ at the cathode, where k is Boltzmann's constant, T is temperature, m is the electron mass, and c is the speed of light (Lawson, 1978). For a uniform current density J, the total current is $I = \pi r_c^2 J$, and thus the lower limit on the emittance is

$$\epsilon_n = 4.0 \times 10^{-6} (kT/J)^{1/2} \text{ mm·mrad}, \tag{4}$$

with J in A/cm^2 and kT in eV.

1.2 Thermionic Emission

Thermionic emission (*q.v.*) occurs when the thermal energy of electrons at a cathode surface exceeds the work function of the surface. The cathode is heated either directly by passing a current through the cathode or indirectly by a filament in proximity to the cathode.

1.2.1 Emission Characteristics The emission current density J_0 follows from the Richardson–Dushman equation (Munro, 1980; Gewartowski and Watson, 1965)

$$\begin{aligned} J_0 &= \left(\frac{4\pi mek^2}{h^3}\right) T^2 \exp\left\{-\frac{e(\phi + \alpha T)}{kT}\right\} \\ &= A(T) T^2 \exp\left[-\frac{e\phi}{kT}\right], \end{aligned} \tag{5}$$

where m is the mass of the electron, e is the electron charge, k is Boltzmann's constant, h is Planck's constant, ϕ is the work function of the surface at absolute zero, T is the temperature, and α is the temperature coefficient of the work function. As defined in the above equation, $A(T)$ is called the emission (or Richardson's) constant and can be found in reference books (Fink and Christiansen, 1982). The theoretical value of A as computed from Eq. (5) is 120.6 A/cm^2·K^2. For most materials the emission constant is significantly lower than the calculated value (see Sec. 1.2.2).

The electric field applied to the surface further reduces the effective work function of the surface by the Schottky effect (Schottky, 1923; Humphries, 1990). This gives an emission current density J of

$$J = J_0 \exp[139(E/T)^{1/2}], \tag{6}$$

where E is the applied surface electric field (kV/cm). The thermionic emission of a material is measured by placing the cathode in a vacuum tube and collecting the emitted electrons on a positive anode. The charge pulled from the cathode reduces (depresses) the electric field at the cathode surface. Therefore, for further emission to occur, the anode–cathode gradient must be greater than the field depression at the cathode surface. If this condition is met, then all the emitted electrons are collected. The emission coefficient is obtained by measuring the current as a function of the anode voltage and extrapolating to zero anode voltage.

1.2.2 Metallic Emitters Metallic emitters are typically refractory metals with a low vapor pressure and high melting temperature. Two metals commonly used are tungsten and tantalum. The work function and emission constant for tungsten are 4.50 eV and 70 A/cm$^2 \cdot$K^2, respectively. The work function and emission constant for tantalum are 4.12 eV and 37 A/cm$^2 \cdot$K^2, respectively. Tungsten heated to 2500 K will produce approximately 0.4 A/cm^2 with a 6000-h lifetime. A higher current can be obtained by raising the cathode temperature [Eq. (5)]. However, the cathode lifetime is a strong function of the cathode temperature. For example, if the temperature is raised to 3000 K, then the lifetime will be reduced to less than a day of operation. Another means to increase the electron emission of metals is by impregnating the metal with thoria and other oxide compounds. This enables operation at lower temperatures, typically 2000 K, and can produce up to 3 A/cm^2.

1.2.3 Alkaline-Earth Oxide Emitters and Dispenser Cathodes Alkaline-earth oxides emit electrons at much lower temperatures than pure metals. However, the lifetime of these cathodes also strongly depends on the operating temperature. As the temperature is raised, the low–work-function material evaporates from the surface. This problem was solved by the development of the dispenser cathode (Gewartowski and Watson, 1965). A dispenser cathode is composed of a porous tungsten matrix impregnated with a compound made of a low–work-function material (such as barium) and a binder. The compound decomposes, and the low–work-function material diffuses through the bulk to the surface at a constant rate. This provides a low–work-function surface at a low operating temperature for long times. A dispenser cathode operates at up to 1400 K and can produce up to 20 A/cm^2. For a dispenser cathode at 1400 K, the average transverse energy of emitted electrons is 0.12 eV. From Eq. (4), a dispenser cathode with a current density of 20 A/cm^2 at 1400 K and a 1-cm^2 area has a 0.3-mm$\cdot$mrad minimum emittance.

1.2.4 Compound Emitters Lanthanum hexaboride (LaB_6) is an example of a compound emitter. LaB_6 requires a high operating temperature (1900 K) but survives in a poorer vacuum than dispenser cathodes (Humphries, 1990). Since some electron accelerator gun systems have poorer vacuums than vacuum tubes, LaB_6 is used as a cathode material for those systems. The maximum current density for LaB_6 cathodes is 20 A/cm^2. Other compound emitters are uranium carbide and zirconium carbide (Lafferty, 1951).

1.3 Photoelectric Emission

Photoemission is the production of electrons from a material by incident electromagnetic radiation. Depending on the choice of photocathode, photoemission can be produced by light wavelengths from the near-infrared to ultraviolet and shorter wavelengths. The quantum efficiency (QE) of a photocathode is the ratio of emitted electrons to incident photons.

The photon source for producing photoemission is a laser. The QE is very important in determining the system design. The QE drives the size, complexity, and cost of the laser system. The laser system is very expensive for high–average-power electron beams.

For semiconductors, the minimum photon energy E_j required to eject an electron from a material is the sum of the electron affinity (Escher, 1981) plus the band-gap energy. The electron affinity is energy difference between the vacuum level and the minimum of the conduction band at the material surface. However, because of phonon and electron scattering in the materials, the emitted electrons do not have an exit energy that is the exact difference between the photon energy and E_j. Typically the photoemitted electrons have a thermalized distribution with an average temperature significantly below the difference between the photon energy and E_j. For example, Cs_3Sb photoemitters have an effective emitted-electron temperature of 0.2 eV with 2.3-eV incident photons (Oettinger and Bursuc, 1987). However, cesium antimonide has an energy band gap of 1.5 eV, thus giving an excess photon energy of 0.8 eV. The extra 0.6 eV of energy goes into heating the material by phonon and electron scattering. Cs_3Sb produces photoelectrons over much of the visible range and has a QE greater than 20% in the ultraviolet. It is widely used in phototubes.

Photoemitters can produce very high charge densities. For instance, Cs_3Sb is capable of delivering over 600 A/cm^2 (Fraser and Sheffield, 1987). From Eq. (4), with an effective electron temperature of 0.2 eV, the minimum emittance for a 1-cm^2 cathode is 0.07 mm·mrad.

The photocathodes discussed in this article are those used for high-current electron beams. The photocathodes are divided into two classes based on QE: low QE and high QE. This class division is chosen because of the significant difference in accelerator design for these two type of photocathodes. Low-QE cathodes are characterized by a relative ease of fabrication and a reduced vacuum requirement. With low-QE cathodes, one cathode could be used for the life of the accelerator. However, the laser power requirements are inversely proportional to the QE. Therefore, low-QE cathodes require much higher peak-power lasers. High-QE cathodes require good vacuums (less than 10^{-8} Torr) and can have extra maintenance requirements, but the laser requirements are less severe.

1.3.1 Low Quantum Efficiency

Many different metals have been used for cathodes. For high-energy electron-beam sources, copper and magnesium (Fischer *et al.*, 1994) are the most common choices. Other metals that have been investigated are Al, Au, stainless steel, Sm, Y, W, Zn, Au, Mo, Ta, Pd, Zr, Ba, Na, and Ca (Suberlucq, 1994; Travier, 1994). Measurements of quantum efficiency vary considerably among individual researchers. This variation can, in part, be attributed to differences in samples, preparation techniques, and contamination before and during measurements. Over all, the measured quantum efficiency of metals varies from less than 2×10^{-6} for copper to 5×10^{-4} for magnesium near a wavelength of 250 nm. The Schottky effect (see Sec. 1.2.1) and the dependence of the absorption of light on angle of incidence have been used to increase the QE of these cathodes. For example, the application of a 100-MV/m surface electric field and of a 60° incidence of the laser increases the QE of a magnesium cathode from 5×10^{-4} to greater than 10^{-3} (Fischer *et al.*, 1994). The application of higher electric fields for still greater QE is limited by electric-field breakdown limits. These limits are much higher for short-pulse and high rf frequencies (see Sec. 2.3) than for dc or long-pulse fields.

The thermionic emitters LaB_6, CeB_6, PrB_6, GdB_6, YB_6, and BaO have also been used as cathodes, both heated and unheated. The heated thermionic emitters are operated just below their emission temperature (typically 1300 K) to keep the surface clean. Again, the measured quantum efficiencies are dependent on many factors and vary among laboratories. The quantum efficiency of LaB_6 and CeB_6 is measured to be 6×10^{-4} (Bamford *et al.*, 1992). The angle of incidence of the laser was 45°.

1.3.2 High Quantum Efficiency

High-QE photocathodes usually require a good vacuum and have a sophisticated fabrication procedure. These types of cathodes can be subdivided into two categories based on application: polarized electron sources and unpolarized electron sources. The first photocathode used in high-energy physics applications was for the production of polarized electrons.

Gallium arsenide (GaAs) has been used for many years as a source for polarized electrons (Chatwell *et al.*, 1994; Alley *et al.*, 1995). Illuminating GaAs with circularly polarized light of a photon energy close to the band gap produces a partially polarized beam of photoelectrons. Single-crystal GaAs can produce a beam with nearly 50% polarization. Recent developments in the production of strained GaAs have led to yields of greater than 80% polarization of the photoelectrons. Strained GaAs is produced by epitaxially growing GaAs on a substrate crystal with a slightly different lattice structure. The substrate is usually $GaAs_{(1-x)}P_x$, where x is a number between 0 and 1 (typically around 0.3). Although the QE of single-crystal GaAs is greater than 10% near 530 nm, the QE of GaAs when used as a polarized electron source is typically close to 1%. To attain the higher QE requires a relatively thick crystal of greater than 1 μm. Because of electron scattering in the GaAs, a beam with a high degree of polarization requires a thin layer of strained GaAs, typically close to 100 nm.

A large number of other multialkali cathodes have been used as electron sources in high-current gun applications—for example, CsK_2Sb, AgO:Cs, CsNaKSb, K_3Sb, and NaK_2Sb. Multialkali cathodes have a significant advantage over metal photocathodes. Multialkali cathodes have QEs (typically 1–15%) that can be many orders of magnitude greater than metal QEs. Since multialkali cathodes rely on a cesium layer to reduce the surface work function to near zero, they tend to be very susceptible to contamination and cesium loss. However, when used in vacuum tubes, the multialkali photocathodes can have a lifetime of years. Vacuum tubes have such a long lifetime because a slight excess of cesium acts as a getter, eliminating contaminants, maintaining the vacuum, and providing a residual vapor pressure of cesium. Unfortunately, in an actively pumped system such as an accelerator, the lifetime can be as short as days. The vacuum requirement for an accelerator system using multialkali cathodes is close to 10^{-10} Torr.

The crystal-like photocathodes, Cs_2Te, CsI, and K_2Te, all require laser wavelengths shorter than 250 nm to obtain quantum efficiencies over 2%. This short wavelength requirement complicates the laser design. However, these cathodes can survive in relatively modest (10^{-8}-Torr) vacuum systems (Chevallay *et al.*, 1994; Kong *et al.*, 1995) Also, initial results indicate that these cathodes can be rejuvenated by heating to 150 °C and then reused.

1.4 Field Emission

The extraction of electrons from a cathode by a high electric field is called field emission. It is a quantum mechanical process relying on the electrons tunneling through a potential barrier, rather than requiring an initial energy higher than the surface barrier (see FIELD EMISSION AND FIELD IONIZATION). Field emission is only weakly temperature dependent and occurs even at a temperature of absolute zero. The current density J is given by the Fowler–Nordheim equation (Feist, 1968),

$$\begin{aligned} J &= 3.2 \times 10^{-6}(E^2/\phi) \\ &\quad \times \exp(-6.8 \times 10^{-7}\phi^{3/2}/E) \quad \text{A/cm}^2, \qquad (7) \end{aligned}$$

where ϕ is the work function and E(V/cm) is the surface electric field.

Field emission is most easily obtained from sharply pointed metal tips. A sharply pointed tip protruding into a high electric field will have a very high surface electric field. Thus, modest anode–cathode voltages suffice for emission. Field emitters are widely used in all electron microscopes as low-emittance point sources of electrons. Tip diameters of 100 nm can be formed. The tips are effectively electron point sources for imaging applications.

Field emission is also used in high-current guns where the cathode structures are large, up to many centimeters in diameter. To obtain sufficiently high electric fields for emission requires very high electric fields with field strengths ranging from 10 to 100 MV/m. However, the electric field at which field emission occurs is typically lower than calculated for a perfectly smooth, clean surface. It is generally believed (Humphries, 1986; Miller, 1982) that this enhancement occurs because the cathode has small microstructures (whiskers) that extend into the cathode–anode gap. The electric-field enhancement at the tip of the whiskers can be several orders of magnitude. After the current flow starts in the whiskers, the whisker tips form a plasma that expands outward covering the surface of the cathode. Since the plasma has a zero-work-function surface, the current is then limited only by the electron charge in front of the cathode. This charge-limited flow can be calculated using the Child–Langmuir emission equation (Humphries, 1990),

$$\begin{aligned} J &= (2e/9\pi m)^{1/2}V^{3/2}/d^2 \\ &= 2.33 \times 10^{-6}V^{3/2}/d^2 \quad \text{A/cm}^2, \qquad (8) \end{aligned}$$

where e is the electron charge, m is the electron mass, V is the cathode–anode voltage, and d is the cathode–anode gap in centimeters.

Field emission depends on the type of surface, and common materials for field emission are metals, velvet, and carbon. Also, surfaces have been deliberately made with uniform patterns of microstructures either to initiate the plasma formation or to have a more uniform pattern of tips on the cathode surface.

1.5 Secondary Emission

Secondary emission (Bruining, 1954; Escher, 1981) is the production of secondary electrons from the surface of a solid into vacuum by energetic incident primary electrons. The number of secondary electrons per incident electron can be greater than 1, and thus this effect is used for electron multiplication. Secondary emission enables the fabrication of high-gain electronic devices such as photomultipliers and image intensifiers. This type of cathode is not typically used for high–energy-density electron beams.

Secondary emission can be measured by first generating a beam of primary electrons with an energy determined by the potential difference between the target and the cathode. These primary electrons pass through a hole in the positive collector and then strike a target. The secondary electrons emitted by the target are then accumulated on the collector. The secondary-emission yield is the ratio of the secondary current (obtained by subtracting the elastically scattered primary current) to the primary current. The secondary-emission yields can vary greatly for differing materials. For instance, lithium has a secondary-emission coefficient of 0.5, and magnesium oxide has a secondary-emission coefficient of 25 (Lide, 1991).

1.6 Ferroelectric Emission

Ferroelectric materials have *E-D* curves analogous to the *B-H* curves for ferromagnetic materials. The remnant polarization of a ferroelectric produces a strong internal electric field that is screened by free charge present on the surface of the ferroelectric. The surface charge density can be many μC/cm^2. By plating a grid pattern on the surface of the ferroelectric and quickly applying a reverse voltage, the polarization of the material can be flipped and the surface charge released. Ferroelectrics, in particular lead-zirconium-titanate and lead-lanthanum-zirconate-titanate (Le Duff, 1990), are presently being investigated for use in high-charge pulsed guns.

1.7 Plasma Emission

As covered in Sec. 1.4, a plasma has no surface work function and can be used to produce very high current densities. Plasma emission can be generated from a pseudospark (Hsu *et al.*, 1993) or a glow discharge (Gavrilov *et al.*, 1994). Complete control over all the electron-emission characteristics, such as making a short pulse duration, is difficult to achieve. Also, the electron temperature in a plasma is several electron volts, which can result in a large-emittance beam. However, since the charge density can be very high, small-radius sources can be considered, thus leading to a low-emittance beam.

2. ELECTRON-GUN DESIGN

The design of electron guns for dc operation is well developed (Pierce, 1954). However, the design of pulsed electron guns is still under development. The design of all high–charge-density electron guns must include the effects of the electric fields present as a result of the electric charge in the beam. These beam self-fields are referred to as the beam space charge. When a pulse is much less than a nanosecond long, the beam fields are never in a steady-state condition for several reasons. The charge of the pulse is increasing and then decreasing in time, leading to a variation of the beam self-fields as a function of time. If a gun with a grid is used, the time to turn the grid on and then off is limited by the capacitance of the grid, causing a time variation in the charge and voltage of the beam. Another effect on the beam quality is the beam fields due to the $\partial I/\partial t$ at the front of the pulse propagating out to the walls of the gun and reflecting back to interact with a following portion of the pulse.

The computer codes (Deaven and Chan, 1990) used for simulating the beam propagation fall into two classes, based on the temporal character of the beam. Examples of codes for dc gun design are EGUN and EBQ. Examples of codes for pulsed beam design are PARMELA and the particle-in-cell (PIC) codes, such as MASK and ISIS.

2.1 Thermionic and Long-Pulse Guns

Thermionic guns produce pulses that range from one-half–nanosecond duration to continuous emission. Even though the beam pulse duration can span a very wide range, the basic gun designs are either gridded or ungridded. At the cathode, the space-charge force of the beam is balanced with a carefully designed focusing electrode, as shown in Fig. 1. By balancing the beam self-repulsion at the beam edge by the field of the focusing electrode, a uniform current density is maintained. The gun current density can be calculated using the Child–Langmuir emission equation [Eq. (8)]. The beam exits

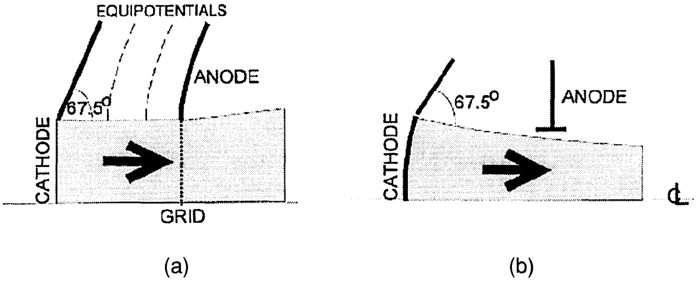

FIG. 1. The Pierce (1949) cathode geometry for a charged-particle beam. The electron beam is produced from the cathode and flows in the direction of the arrow. The gridded gun in (a) has a diverging beam after the anode. The gun in (b) produces a beam that has a focus downstream of the gun.

either through a grid positioned along the anode equipotential line or through an anode aperture. The dc gun can be designed with either a divergent or a focused beam. Figure 2 is a schematic of a dc gun designed for the University of California at Santa Barbara (Herrmannsfeldt, 1988).

The exit energy of the beam from the gun should be as high as possible to minimize further space-charge defocusing downstream from the gun. If the beam maintains a uniform profile up to relativistic energies, then the beam emittance can be near the cathode-limited thermal temperature.

Many applications require a high charge per bunch in short pulses. For instance, a radio-frequency (rf) accelerator requires that the bunch length of the pulse be a small fraction of the rf period, as little as a few picoseconds at the common accelerator frequency of 3 GHz. However, the pulse lengths are limited by grid switching times to a minimum of 0.5 ns. Typically, these guns have rf cavities downstream of the gun that compress the pulse length in time. These cavities are called bunchers. The buncher rf phase is set so that the middle of the electron pulse has nearly no net energy change after exiting from the buncher. The buncher frequency is set so that the electron pulse length is equal to the linear portion of the rf sine wave. The correct phase will give a linear energy slew across the electron pulse with the front of the electron pulse having a relative energy less than the end of the pulse. After the buncher, the beam is allowed to drift so that the end of the pulse catches up to the front of the pulse, giving a reduction in pulse length. Since this bunching technique relies on speed differences along the electron bunch to give compression, this technique can only be used effectively with speeds a fraction of the speed of light. The bunching process can cause the emittance to grow by at least an order of magnitude above the cathode thermal emittance.

Thermionic guns can be pulsed at repetition rates of up to 1 GHz. At very high rep-

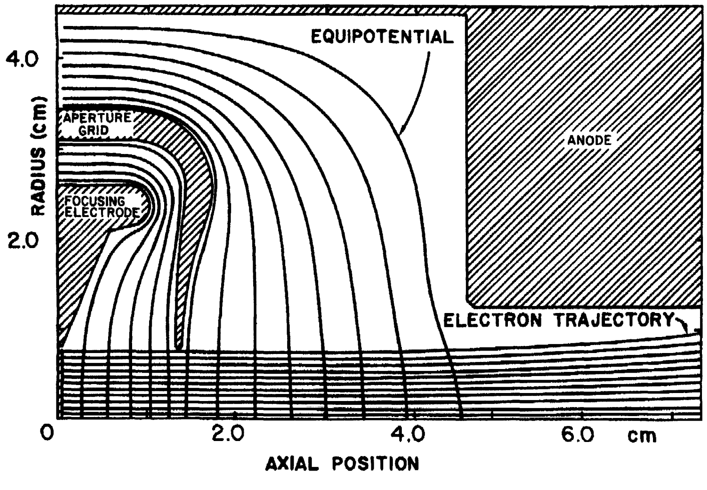

FIG. 2. dc gun designed for University of California, Santa Barbara, in a Pierce geometry. The electron gun was turned on by applying a 10-kV voltage pulse to the aperture grid and turned off by applying −2 kV to the aperture grid. The cathode temperature was 1100 K. The anode voltage was 50 kV. The gun produced 2.2-A pulses that could be varied from 1 ms to dc. (Courtesy of Gerald Ramian, UCSB.)

etition rates, the charge is limited to less than 0.25 nC (van Amersfoort *et al.,* 1992). These high–repetition-rate thermionic guns have a cathode-to-grid spacing of 150 μm. The grid is driven by a 1-GHz rf field, giving a 1-GHz pulse train. For lower pulse repetition rates, higher charges are possible. At Osaka, pulses of 1 nC at repetition rates of 27 MHz are generated (Tsumori *et al.,* 1993). In the Osaka gun, the gun grid is driven by a combined 27- and 81-MHz rf field. The addition of the third harmonic at 81 MHz produces a much faster grid rise and fall time, with a 4-ns effective grid pulse. The longer pulse gives a larger integrated charge.

2.2 rf Gun

Another approach to electron-gun design is to use a thermionic cathode in an rf cavity. Radio-frequency cavities operate in a TEM_{010} mode (Ramo *et al.,* 1965), giving an axial electric field that oscillates at the rf frequency. For a fraction of the rf cycle, electrons are extracted from the cathode by the axial field and are then accelerated out of the rf cavity. Since these electrons cover close to 180° of rf phase, a wide range of energies correlated with the rf phase will be produced by the gun. The Mark III gun system (Benson *et al.,* 1988) shown in Fig. 3 was built for a free-electron laser accelerator. The rf gun has a LaB_6 thermionic electron source inside an rf cavity. The rf gun produces an electron beam for over a quarter of the rf cycle. The beam has a well-correlated with time, but large, range of energies. The peak energy is 1 MeV.

After the gun, the electrons are then transported to a magnetic element called an alpha magnet. The alpha magnet has the lower-energy particles take a shorter path (smaller bend radius) than the higher-energy particles. For a fraction of the rf cycle, the electron pulse from the rf gun has higher-energy electrons that are followed by the lower-energy electrons. The electrons of different energies take different paths through the magnet. Using an energy-selective slit in the magnet region of maximum energy dispersion accomplishes two functions. First, the slit limits the range of electron energy

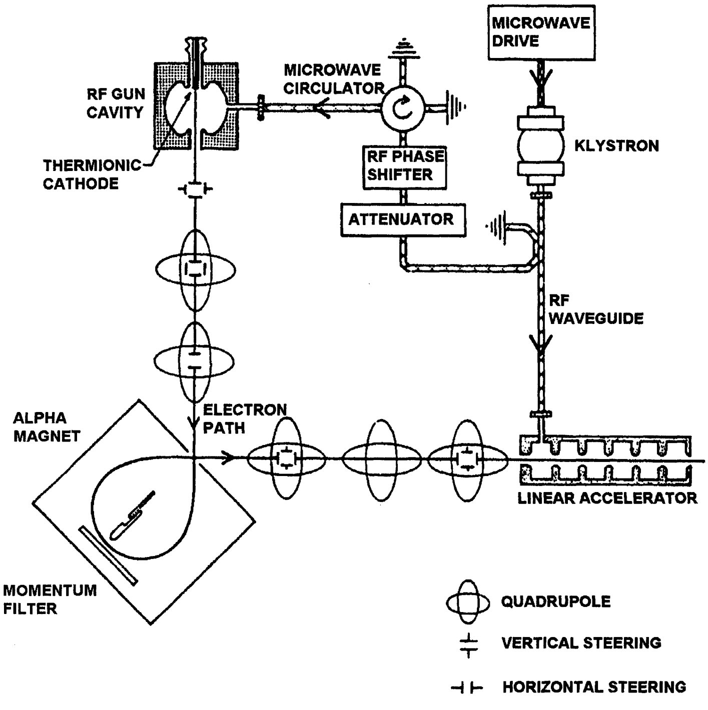

FIG. 3. Schematic of a free-electron laser experiment using an rf gun electron source. Shown are microwave feed system and the path of the electrons from the LaB_6 thermionic gun to the Mark III accelerator. The electron beam passes through beam-line elements for steering (dipoles) and focusing (quadrupoles) and then into an alpha magnet. An alpha magnet (so named because of the electron path) has a momentum slit located at the plane of maximum momentum dispersion. The momentum filter stops all electrons, except for those within a narrow momentum width. The rf gun with alpha magnet (×10 current compression from the alpha magnet) produces a 33-A peak current electron pulse with an energy of 1 MeV. Each rf cycle (rf frequency is 2857 MHz) produces an electron pulse. Each electron pulse has a 1- to 3-ps duration after the alpha magnet. The rf drives the gun for 2 to 5 μs. The electrons are then accelerated to 41 MeV in the accelerator. The gun x by y emittance was approximately 1 by 2 mm·mrad, respectively.

from the gun system. Second, because of the longer path of the higher-energy electrons, the lower-energy electrons catch up to the higher-energy electrons, giving a shorter pulse length. Since the particles are close to the speed of light ($\beta \sim 0.95$), the small difference in relative particle speed for the range of electron energies transmitted through the slit does not appreciably affect the pulse length. In fact, magnetic compressors can be designed that have the lower-energy particles catching up to the higher-energy particles, again giving a bunch compression.

These types of guns are similar to dc guns since the temporal variations are eliminated by the slit. In fact, the beam from an rf gun has a lower emittance than from most pulsed thermionic guns. Since the beam from a thermionic gun is usually not scraped, all of the time-varying beam components are thermalized as the beam propagates down a beam line, resulting in a much higher emittance.

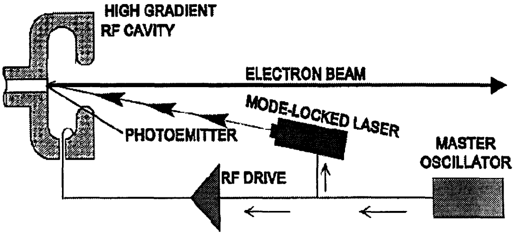

FIG. 4. Basic components of a photoinjector are a laser, a photocathode, an rf source (klystron), and an rf cavity. A single master oscillator provides the rf signal for a laser and a klystron. The laser is typically a solid-state yttrium-lithium-fluoride (YLF) laser. Depending on the choice of photocathode, the laser is either frequency doubled, tripled, or quadrupled to be near the peak of the quantum efficiency. This laser produces a train of short pulses locked to the rf input signal. The laser and klystron thus have a well-defined phase difference with respect to each other. The relative phase of the laser and the rf is adjusted by an rf phase shifter. The relative phase is chosen so that the light pulse hits the photocathode at the best rf phase for the production of a high-quality electron beam.

2.3 Photoinjector

Most of the emittance growth in high–pulse-charge electron accelerators comes from compressing the pulse in time and in space. If the high-charge pulse can be produced with close to the final required pulse shape and then rapidly accelerated to relativistic energies, the emittance growth can be controlled. A design that effectively uses this approach, the photoinjector, combines a photocathode and an rf cavity. This gun is used for the production of subpicosecond- to nanosecond-long pulses of picocoulombs to hundreds of nanocoulombs charge. A schematic of a photoinjector is shown Fig. 4.

Photoinjectors (Fraser and Sheffield, 1987) have several unique characteristics. A high-gradient rf cavity is used to supply the accelerating field. Not only does the high gradient reduce space-charge effects, but the high field also enables laminar flow from the cathode through the accelerator. The high gradient also allows the extraction of high charge for closely spaced pulses, enabling the production of high average currents. A short-pulse laser illuminates the cathode giving short electron pulses. The design of a photoinjector must take into account the large space-charge forces and rapidly changing temporal conditions of the beam. In this regard, the photoinjector is similar to the pulsed diodes described in Sec. 2.4.

Since the electron source is a photocathode illuminated with a laser, the machine designer has extensive control over the spatial and temporal characteristics of the electron emission process. The electron pulse is limited by the surface charge on the cathode produced by the axial electric field present in the rf cavity. The upper limit on extracted charge can be estimated by assuming a surface charge density of $\epsilon_0 E$, where E is the electric field on the cathode during the pulse and ϵ_0 is the dielectric constant. For instance, a cathode with a surface gradient of 30 MV/m has a surface charge density of 27 nC/cm^2. Since photoinjectors routinely use lasers of less than 10-ps pulse lengths, the calculated current density from the gun is greater than 3 kA/cm^2. However, if all of the charge on a surface is extracted, the surface field at the end of the pulse would be zero. The resultant pulse would have a poor emittance and energy spread because of the high degree of field variation at the cathode surface. Also, the electron pulse lengths are limited by space-charge effects. In the first few centimeters in front of the cathode, the space charge causes a longitudinal expansion, thus giving a minimum pulse length. The pulse current density out of the gun is limited to 500 A/cm^2 for 30-MV/m to greater than 1 kA/cm^2 for 100-MV/m surface fields. The measured electron-beam emittance varies, depending on the machine design, approximately as 1 to 2 mm·mrad per nanocouloumb up to 10 nC.

Designs of photoinjectors now integrate the gun rf cavity with several following rf cavities in proximity to the gun. An example of this type of design is the accelerator (Fig. 5) built at Los Alamos National Laboratory for a free-electron laser (Sheffield *et al.*, 1992). This 1300-MHz accelerator has the photoinjector followed by ten accelerating cells, all fed by a single rf port. The photocathode is Cs_2Te. The exit energy of the electrons from the accelerator can be varied between 10 and 20 MeV. The pulse length is between 5 and 20 ps, depending on the pulse charge and accelerator gradient. The beam emittance is 2 mm·mrad for 1 nC.

Research is still proceeding on high–average-current machines at Boeing Defense and Space (Dowell *et al.*, 1993). The first demonstration of a high average current using a photoinjector was on the Boeing accelerator. This 25%-duty-factor, 433-MHz machine has demonstrated an average current of 32 mA at 5 MeV, giving an average beam power of 160 kW. The macropulse average current was 0.13 A. The beam emittance was 5 mm·mrad for 1-nC pulse charge.

Many designs are based on the work done at Brookhaven National Laboratory at 2856 MHz (Batchelor *et al.*, 1992). One of the advantages of operating near 3 GHz is the higher cathode surface electric fields that can be obtained as compared to operation at lower frequencies. The field breakdown level was initially characterized by the Kilpatrick theory of breakdown limits (Kilpatrick, 1957). The experimental data for this initial theory were measured using ions and parallel electrodes. For electrons in cavities fabricated using present-day machining techniques, the breakdown limits in high-frequency rf cavities can be much higher than predicted with the original theory. In general, the breakdown limit in both the aforementioned cases is proportional to the square root of the rf frequency. However, in practice, many other factors are more important than the frequency scaling of field breakdown. In particular, choice of cavity material, surface cleanliness, machining techniques, and vacuum levels are important. Also, the duration of the rf pulse can significantly affect the breakdown limit (the longer the pulse, the lower the breakdown limit). For a 2-μs-long rf pulse at 2.8 GHz, the peak

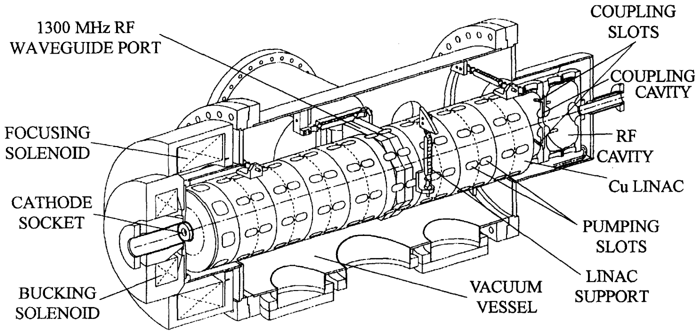

FIG. 5. The FEL accelerator at Los Alamos National Laboratory. The accelerator is supported by titanium struts in an ultrahigh-vacuum container. The focusing solenoid focuses and steers the electron beam out the end of the accelerator. Each accelerating cavity is 180 rf degrees out of phase with the adjacent cavities. The electrons take 180° (traveling at the speed of light, this equals 11.53 cm at 1300 MHz) to cross an accelerator cavity. By the time the electrons arrive at the adjacent cavity, the phase of the next cavity has flipped in sign to one of acceleration for the electrons. Thus, the electrons are accelerated through all the cavities. The rf power is fed in through the waveguide port in the middle of the accelerator. The rf power flows through the structure through slots in the walls of the rf cavities. Between the rf cavities are small cavities called coupling cells. The coupling cells are resonant at 1300 MHz and 90° out of phase with respect to the accelerating cavities. The coupling cells use very little rf power. The coupling cavities force the 180° phase shift between the adjacent accelerator cavities.

surface field measurement is 300 MV/m (Loew and Wang, 1990), approximately eight times the Kilpatrick theory prediction.

2.4 Pulsed Diodes

Pulsed diodes typically use field emission for electron production. However, several examples of thermionic-source systems also exist. For example, the PHERMEX machine has a thermionic cathode (Kauppila *et al.*, 1993). The PHERMEX gun produces a 0.5-MeV, 1-kA beam in a 200-ns-long pulse. Field-emission guns require high cathode surface fields to enable field emission. Two means of producing high surface fields are either sharp points projecting into the cathode–anode gap or very high voltages across that gap. In most diode designs, the high voltages employed are sufficient to generate a high-density plasma in the anode–cathode gap. This plasma would then short out the diode. To generate a beam from these diodes requires a fast (relative to the plasma motion across the gap) voltage rise and fall time across the anode–cathode gap.

Two types of pulsed-diode gun designs are the foil diode and the foilless diode. The foil diode has a conducting surface for the anode plane. The foilless diodes typically use magnetic fields to guide the emitted beam through the anode aperture. Depending on the gun configuration, the electron beam can be either solid or annular, and sheet or cylindrically symmetric. These types of guns typically have a Marx generator to provide a high–energy-density electrical pulse. A Marx generator is a series stack of capacitors connected by spark gaps. The capacitors are charged in parallel. After charging, the spark gaps are closed. So the full voltage across the capacitor stack is the serial sum of the capacitors' voltages. The Marx generator then launches a pulse into a pulse-forming line (PFL). The PFL matches the impedance of the Marx generator to the diode and tailors the pulse shape. Another arrangement for power feed instead of a Marx generator and PFL is to have a pulse-forming network attached to the diode.

For foil diodes, the foil requirements depend on whether the diode is to be used for single-pulse or long-term operation. For long-term operation, a foil must be thin and have good mechanical strength, heat resistance, and radiation resistance. The thinner foils are better for beam propagation since the foil causes transverse scattering of the electrons, increasing the beam emittance. However, a requirement for high average currents or a large single-pulse charge leads to thicker foils, usually of aluminum or titanium. Typical foil thickness for long-term operation ranges from 15 to 100 μm. For low-current beams or single-pulse operation, much thinner foils (0.5 μm thick) can be used, such as Mylar or Kapton with thin metal surface coatings.

Figure 6 shows the foil diode configuration of the Ector (Carlson *et al.*, 1993) machine. This pulsed diode uses a ball cathode. The diode voltage produces a 3.5- to 4.0-MeV, 35-kA beam. The current density is greater than 10 kA/cm^2. The pulse length is approximately 60 ns. The beam is focused onto an x-ray conversion target 4.5 cm from the diode. The beam spot at the target is an approximately uniform 0.7-cm^2 spot.

An example of a foilless diode is the Intense Relativistic Electron Beam (Sheffield *et al.*, 1982) machine shown in Figs. 7 and 8. The field-emission diode produces a 3- to 5-MeV, 100-kA beam. The entire diode region is immersed in a 10-T magnetic field to constrain the radial motion of the beam. The annular beam from the diode has a 1-cm radius and less than 300-μm radial thickness. The current density is greater than 0.5 MA/cm^2. The pulse length is approximately 50 ns. The application of this beam is limited to the high magnetic field region. On exiting the magnetic field, the effective emittance increases substantially.

2.5 Virtual Cathodes

The beam current that can be accelerated through a region is limited by the field depression caused by the beam space charge (see Sec. 1.2.1). A virtual cathode is formed when the accelerating potential is not sufficient to overcome the beam space-charge fields. For this case, the electrons form a plasma cloud between the cathode and anode. Electrons are then reflected back toward the cathode.

High-power microwave generation based on using virtual cathodes (vircators) is presently an area of research. The microwaves can be generated by oscillations of the

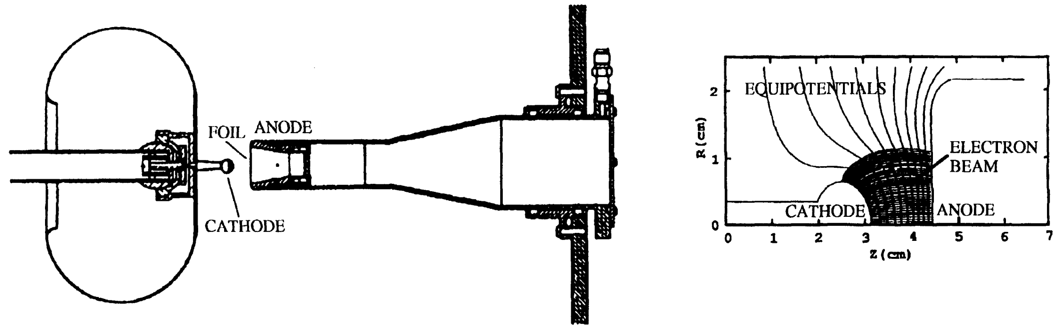
FOIL
ANODE
CATHODE
2
1
0
R (cm)
EQUIPOTENTIALS
ELECTRON BEAM
CATHODE
ANODE
0
1
2
3
4
5
6
7
Z (cm)

(a) (b)

FIG. 6. The Ector pulsed-diode gun. (a) The diode region showing the ball cathode and anode. The field-emission cathode is a 12.7-mm-diameter ball on an 18-mm stem. The polished stainless steel ball is painted with silver to aid the start of emission. The beam produced from the ball is accelerated across a 13-mm cathode–anode gap. The beam then passes through a 25-μm anode foil. The solid emitted beam has a 1-cm radius. (b) A computer simulation showing the electron beam in the cathode–anode region. The electric-field equipotentials are also shown (courtesy of Randy Carlson).

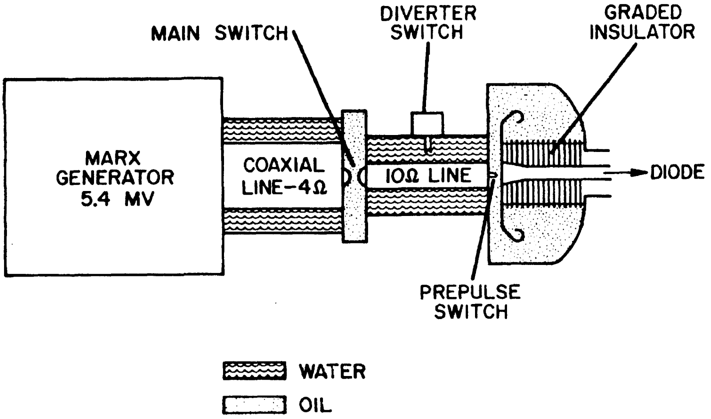
MAIN SWITCH
DIVERTER SWITCH
GRADED INSULATOR
MARX GENERATOR 5.4 MV
COAXIAL LINE–4Ω
10Ω LINE
DIODE
PREPULSE SWITCH
WATER
OIL

FIG. 7. A low-inductance 54-stage Marx generator (216 kJ total energy, 100 kV/stage) charges a 4.5-Ω water transmission line. The main switch breaks down at 95% of peak voltage and a 4-MV pulse is launched into the 10-Ω water transmission line. The prepulse switch is set for breakdown at 1 MV. The low prepulse voltage prevents formation of a plasma in the anode–cathode gap before arrival of the main pulse.

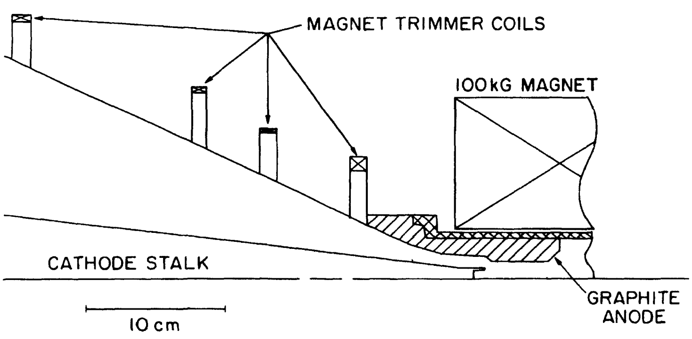
MAGNET TRIMMER COILS
100kG MAGNET
CATHODE STALK
GRAPHITE ANODE
10 cm

FIG. 8. Foilless diode gun that produces an annular electron beam. The diode cathode stalk is attached to the back plate of the diode graded-insulator stack, as shown in Fig. 7 (adjacent to the prepulse switch). The cathode stalk diameter is smoothly decreased over a length of 62 cm to the cathode tip diameter of 2 cm. The cathode tip is a removable stainless steel tube with a wall thickness of 0.076 cm. A 10-T magnet provides a pulsed field that is uniform to 1% over the central 27-cm length of the magnet bore. The trimmer coils contour the fringing fields to be parallel to the surface of the cathode stalk.

plasma cloud. Enhanced frequency selectivity and extraction efficiency are obtained by phase locking of the virtual cathode oscillation in a pre-existing rf field. The microwave generation is enhanced by the buildup of electron bunching at the microwave period. This bunching increases the coupling to the microwave field, providing more microwave power. Gigawatt power levels are possible from 0.5 to 17 GHz (Rix and White, 1994).

3. APPLICATIONS

Electron beams have been in use for many years in electronic device vacuum tubes and then in cathode-ray tubes. Larger electron-beam systems include microwave tubes and high-power radio-frequency devices such as klystrons for radar installations. Electron beams are used in applications as diverse as lithography and the ionization of laser media.

Electron microscopes are widely used in the biological and material sciences. Some of the medical applications of electron beams are cancer therapy and radiography.

High-power electron beams are also used in many industrial processes. The following is a sample list: sterilization of medical products, food containers and products, and sewage; welding, plasma-source ion implantation, material annealing, decomposition and precipitation of chemical pollutants and hazardous materials; cross-linking of polymers in insulation for wires and cables and polymerization; curing of composites, surface coatings, magnetic recording media, and adhesives; radiography for flaw detection; killing insects in grains; fabrication of fiberglass; and radiation hardening.

The electron source requirements for high-energy physics have long provided support for the development of both unpolarized and polarized electron sources. After the advent of synchrotrons for particle physics, the development of electron synchrotrons into x-ray light sources took place. Electron beams are also used for electron cooling of ion beams.

Polarized electrons have many applications. For example, in high-energy physics, the scattering of polarized electrons from polarized nuclei gives information on nucleon form factors. In solid-state physics, the scattering of polarized electrons gives information on the physics of surface magnetism.

High–average-current, low-emittance electron beams find application as sources for free-electron lasers and colliders. High pulse charges are used for wakefield accelerators. A wakefield accelerator is essentially a transformer, using the fields produced by a low-energy, high-charge pulse to drive an accelerator structure designed to accelerate a high-energy, low-charge pulse. Proposed applications include the generation of high–duty-factor, picosecond, high-energy x-ray pulses and picosecond soft x rays by Compton backscattering. Another interesting proposed application requiring both a low-emittance electron beam and a free-electron laser is a gamma–gamma collider (Sessler, 1995).

High-charge pulsed-current electron sources are required for flash x-ray radiography. Other applications include high-power switches, lithography, and plasma processing.

GLOSSARY

Emittance: A parameter used to characterize charged-particle beam quality; a measure of the ability to focus a beam to a small spot over a given distance.

Ferroelectric: A crystalline material with a permanent internal electric polarization.

Field Emission: Production of electrons by electron tunneling through a potential barrier formed by the application of a high surface electric field.

Normalized Emittance: An invariant of electron beam propagation, obtained by multiplying the emittance by relativistic factor $\beta\gamma$.

Photoemission: The production of electrons from a material by incident electromagnetic radiation.

Plasma Emission: The production of electrons by extraction from a plasma.

Polarized Electron: An electron with a preferred spin orientation.

Secondary Emission: The production of secondary electrons from the surface of a solid into vacuum by energetic incident primary electrons.

Space Charge: The net electric charge within a given volume.

Thermionic Emission: The production of electrons by heating a material.

Virtual Cathode: An electron cloud that forms in the anode–cathode gap when the anode voltage is not sufficient to overcome the depression in voltage in the gap due to the field of the emitted electrons.

Works Cited

Alley, R., Aoyagi, H., Clendenin, J., Frisch, J., Garden, C., Hoyt, E., Kirby, R., Klaisner, L., Kulikov, A., Miller, R., Mulhollan, G., Prescott, C., Saez, P., Schultz, D., Tang, H., Turner, J., Witte, K., Woods, M., Yeremian, A. D., Zolotorev, M. (1995), *Nucl. Instrum. Methods Phys. Res.* **A365,** 1–27.

Bamford, D. J., Bakshi, M. H., Deacon, D. A. G. (1992), *Nucl. Instrum. Methods Phys. Res.* **A318,** 377–380.

Batchelor, K., Ben-Zvi, I., Fernow, R. C., Fischer, J., Fisher, A. S., Gallardo, J., Ingold, G., Kirk, H. G., Leung, K. P., Malone, R., Pogorelsky, I., Srinivasan-Rao, T., Rogers, J., Tsang, T., Sheehan, J., Ulc, S., Woodle, M., Xie, J., Zhang, R. S., Lin, L. Y., McDonald, K. T., Russell, D. P., Hung, C. M., Wang, X. J. (1992), *Nucl. Instrum. Methods Phys. Res.* **A318,** 372–376.

Benson, S. V., Schultz, J., Hooper, B. A., Crane, R., Madey, J. M. J. (1988), *Nucl. Instrum. Methods Phys. Res.* **A272,** 22–28.

Bruining, H. (1954), *Secondary Electron Emission,* New York: McGraw-Hill.

Carlson, R. L., George, M. J., Hughes, T. P., Welch, D. R. (1993), in: S. T. Corneliussen (Ed.), *Proceedings of 1993 Particle Accelerator Conference,* Piscataway, NJ: IEEE Service Center, pp. 661–663.

Chatwell, M., Clendenin, J., Maruyama, T., Schultz, D. (Eds.) (1994), *Proceedings of the Workshop on Photocathodes for Polarized Electron Sources for Accelerators,* SLAC Internal Report SLAC-432 Rev.

Chevallay, E., Durand, J., Hutchins, S., Suberlucq, G., Wurgel, M. (1994), *Nucl. Instrum. Methods Phys. Res.* **A340,** p. 146.

Deaven, H. S., Chan, K. C. D. (Eds.) (1990), *Computer Codes for Particle Accelerator Design and Analysis: A Compendium,* Los Alamos National Laboratory Internal Report LA-UR-90-1766.

Dowell, D. H., Davis, K. J., Friddell, K. D., Tyson, E. L., Lancaster, C. A., Milliman, L., Rodenburg, R. E., Aas, T., Bemes, M., Bethel, S. Z., Johnson, P. E., Murphy, K., Whelen, C., Busch, G. E., Remelius, D. K. (1993), *Appl. Phys. Lett.* **63,** 2035–2037.

Escher, J. S. (1981), in: R. K. Willardson, A. C. Beer (Eds.), *Semiconductors and Semimetals,* Vol. 15, New York: Academic, pp. 210 and 265–268.

Feist, W. M. (1968), in: L. Marton, A. B. El-Kareh (Eds.), *Electron Beam and Laser Beam Technology,* Advances in Electronic and Electron Physics, Supplement 4, New York: Academic, pp. 1–57.

Fink, D. G., Christiansen, D. (Eds.) (1982), *Electronic Engineer's Handbook,* 2nd ed., New York: McGraw-Hill, pp. 6–99.

Fischer, J., Srinivasan-Rao, T., Tsang, T. (1994), in: R. Wanzenberg (Ed.), *Sources '94,* Schwerin, Germany, Sept. 29–Oct. 4, Hamburg: DESY, pp. 287–289.

Fraser, J. S., Sheffield, R. L. (1987), *IEEE J. Quant. Elec.* **QE-23,** 1489–1496.

Gavrilov, N. V., Nikulin, S. P., Ponomarev, A. V. (1994), in: W. Rix and R. White (Eds.), *Beams '94,* 10th International Conference on High Power Particle Beams, NTIS PB95-144317, Springfield, VA: NTIS, pp. 507–510.

Gewartowski, J. W., Watson, H. A. (1965), *Principles of Electron Tubes,* Princeton: D. van Nostrand Co., Inc., p. 38.

Herrmannsfeldt, W. B. (1988) in: M. Month, M. Dienes (Eds.), *Physics of Particle Accelerators,* AIP Conference Proceedings 184, Vol. 2, New York: AIP, pp. 1533–1542.

Hsu, T.-Y., Liou, R., Kirkman, G., Gundersen, M. A. (1993), in: S. T. Corneliussen (Ed.), *Proceedings of 1993 Particle Accelerator Conference,* Piscataway, NJ: IEEE Service Center, pp. 3066–3068.

Humphries, S. (1986), *Principles of Charged Particle Acceleration,* New York: Wiley, p. 227.

Humphries, S. (1990), *Charged Particle Beams,* New York: Wiley, p. 199, 283, and 286.

Kilpatrick, W. D. (1957), *Rev. Sci. Instrum.* **28,** 824.

Kauppila, T., Builta, L., Burns, M., Gregory, W., Honaberger, D., Watson, S., Hughes, T. (1993), in: S. T. Corneliussen (Ed.), *Proceedings of 1993 Particle Accelerator Conference,* Piscataway, NJ: IEEE Service Center, pp. 3055–3057.

Kong, S. H., Kinross-Wright, J., Nguyen, D. C., Sheffield, R. L. (1995), *J. Appl. Phys.* **77,** 1–8.

Lafferty, J. M. (1951), *J. Appl. Phys.* **22,** 299.

Lawson, J. D. (1978), *The Physics of Charged Particle Beams,* Oxford: Oxford Univ. Press, p. 199.

Le Duff, J. (Ed.) (1990), *Short Pulse High Current Cathodes,* Gif-Sur-Yvette, France: Editions Frontières.

Lejuene, C., Aubert, J. (1980) in: A. Septier (Ed.), *Applied Charged Particle Optics,* Advances in Electronics and Electron Physics, Supplement 13A, New York: Academic, 159–259.

Lide, D. R. (Ed.) (1991), *CRC Handbook of Chemistry and Physics,* 71st ed., Baca Raton, FL: CRC Press, pp. 12–86.

Loew, G. A., Wang, J. W. (1990), *XIVth International Symposium on Discharges and Electrical Insulation in Vacuum,* Stanford Linear Accelerator Center Report No. SLAC-PUB-5320.

Miller, R. B. (1982), *An Introduction to the Physics of Intense Charged Particle Beams,* New York: Plenum Press, p. 37.

Munro, E. (1980), in: A. Septier (Ed.), *Applied Charged Particle Optics,* Advances in Electronics and Electron Physics, Supplement 13B, New York: Academic, p. 84.

Oettinger, P. E., Bursuc, I. (1987), *Proceedings of 1987 Particle Accelerator Conference,* Piscataway, NJ: IEEE Service Center, pp. 286–288.

Pierce, J. R. (1954), *Theory and Design of Electron Beams,* Canada: D. van Nostrand; Ann Arbor, MI: University Microfilms International.

Ramo, S., Whinnery, J., Duzer, T. (1965), *Fields and Waves in Communication Electronics,* New York: Wiley, 548–551.

Rix, W., White, R. (Eds.) (1994), *Beams '94,* 10th International Conference on High Power Particle Beams, Vol. 1, NTIS PB95-144317, Springfield, VA: NTIS.

Schottky, W. (1923), *Z. Physik.* **14,** 63.

Sessler, A. M. (1995), *Proceedings of 1995 Particle Accelerator Conference,* Piscataway, NJ: IEEE Service Center.

Sheffield, R. L., Browman, M. J., Carlsten, B. E., Young, L. M. (1992), *Nucl. Instrum. Methods Phys. Res.* **A318,** 282–289.

Sheffield, R. L., Montgomery, M. D., Parker, J. V., Riepe, K. B., Singer, S. (1982), *J. Appl. Phys.* **53,** 5408–5413.

Suberlucq, G. (1994), in: R. Wanzenberg (Ed.), *Sources '94,* Schwerin, Germany, Sept. 29–Oct. 4, reprints from DESY, Germany, pp. 557–561.

Travier, C. (1994), in: P. Schoessow (Ed.), *Advanced Accelerator Concepts,* AIP Conference Proceedings 335, Woodbury, NY: AIP Press, pp. 57–74.

Tsumori, K., Okuda, S., Yamamoto, T., Suemine, S., Takamuku, S. (1993), in: S. T. Corneliussen (Ed.), *Proceedings of 1993 Particle Accelerator Conference,* Piscataway, NJ: IEEE Service Center, pp. 650–652.

van Amersfoort, P. W., Bakker, R. J., Bekkers, J. B., Best, R. W. B., van Buuren, R., Delmee, P. F. M., Faatz, B., van der Geer, C. A. J., Jaroszynski, D. A., Manintveld, P., Mastop, W. J., Meddens, B. J. H., van der Meer, A. F. G., Nijman, J. P., Oepts, D., Pluygers, J., van der Wiel, M. J. (1992), *Nucl. Instrum. Methods Phys. Res.* **A318,** 42–46.

Further Reading

Photocathodes

Chatwell, M., Clendenin, J., Maruyama, T., Schutlz, D. (Eds.) (1994), *Proceedings of the Workshop on Photocathodes for Polarized Electron Sources for Accelerators,* SLAC Internal Report SLAC-432 Rev.

Dance, J. B. (1967), *Cold Cathode Tubes,* London: Pittman Press.

Escher, J. S. (1981), in: R. K. Willardson and A. C. Beer (Eds.), *Semiconductors and Semimetals,* Vol. 15, New York: Academic.

Le Duff, J. (Ed.) (1990), *Short Pulse High Current Cathodes,* 91192 Gif-Sur-Yvette, France: Editions Frontières.

Sommer, A. H. (1968), *Photoemissive Materials,* New York: Wiley.

Gun Design

Bakish, R. (Ed.) (1962), *Introduction to Electron Beam Technology,* New York: Wiley.

Miller, R. B. (1982), *An Introduction to the Physics of Intense Charged Particle Beams,* New York: Plenum Press.

Reference Data for Radio Engineers (1982), Indianapolis: Howard W. Sams & Co.

Rix, W., White, R. (Eds.) (1994), *Beams '94,* Proceedings of the 10th International Conference on High Power Particle Beams, Vols. 1 and 2, Springfield, VA: NTIS.

Accelerator Theory and Technology

Corneliussen, S. T. (Ed.), *Proceedings of 1993 Particle Accelerator Conference,* Piscataway, NJ: IEEE Service Center.

Month, M., Dienes, M. (Eds.) (1992), *The Physics of Particle Accelerators,* AIP Conference Proceedings 249, Vols. 1 and 2, New York: AIP.

Septier, A. (Ed.), *Applied Charged Particle Optics,* Advances in Electronics and Electron Physics, Supplements 13A, 13B, and 13C, New York: Academic.

Electron Optics and Beam Lines

Humphries, S. (1986), *Principles of Charged Particle Acceleration,* New York: Wiley.

Lawson, J. D. (1978), *The Physics of Charged Particle Beams,* Oxford: Clarendon Press.

Applications

Abramyan, E. A. (1988), *Industrial Electron Accelerators and Applications,* Washington: Hemisphere Publishing Corp.

Scharf, W. (1986), *Particle Accelerators and Their Uses,* Part 2, Chur: Harwood Academic Pub.

SPACE EXPLORATION

HANS MARK, *Department of Aerospace Engineering and Engineering Mechanics, The University of Texas at Austin, Austin, Texas, U.S.A.*

Introduction 81
1. **The Means to Get There** 83
1.1 The Principles of Rocket Propulsion 86
2. **The Objectives of Space Exploration** 88
3. **The Space Environment** 89
4. **The Sun** 92
5. **The Moon** 93
5.1 The Apollo Program 93
5.2 Scientific Results from the Apollo Lunar Exploration 96
5.3 The Risks of Space Exploration: The Apollo "204" Fire and Apollo 13 97
5.4 The Soviet Lunar Exploration Program 98
6. **The Planets** 98
6.1 Mercury 103
6.2 Venus 103
6.3 Mars 103
6.4 Jupiter 108
6.5 Saturn 110
6.6 Uranus 112
6.7 Neptune 113
6.8 Pluto 113
6.9 Comets and Asteroids 113
7. **Astronomy from Space** 113
7.1 Orbiting Astronomical Observations 114
7.2 Specialized Observatories 114
7.3 High-Energy Astronomical Observatories 115
7.4 The Great Observatories 115
8. **The Space Shuttle** 121
9. **The Space Station** 124
10. **The Future** 126
Works Cited 127
Further Reading 128

INTRODUCTION

Any discussion of the subject of "exploration" must start with the answer to the question: Is there anything to explore? Until the year 1609, the location (in a large sense) of the Earth in the cosmos was a matter of speculation and controversy. It was only when Galileo Galilei first pointed a telescope at the sky in that year that it dawned on people that there was actually something to explore. What Galileo did first was to look at the Moon. In addition to strange new craters, he was startled to see mountains, plains, and valleys similar to those on Earth. The Moon was clearly a place one could visit and explore, so that the original question, in fact, was found to have a positive answer. There was, indeed, something to explore. Galileo even made some drawings of the moon as he saw it through his primitive telescope (Fig. 1) that identified some of the salient features (Galilei, 1610).

Forty-five years later (1654), Christian Huygens had access to a telescope that was good enough to reveal clearly that the planet Mars also had surface features that were also analogous to things on Earth. Specifically, Huygens discovered that Mars had polar snow caps that change with the seasons just the way the snow cover on Earth advances and retreats as seasons change (Fig. 2). In addition to the Moon, then, Mars very probably was also a place to go. By that time, also, Johannes Kepler had developed his laws of planetary motion, and there was no longer any doubt at all that the heliocentric model of the solar system was correct. It was, in short, a theory that had predictive value. Thus, by extrapolation the other planets in the solar system were also places to visit and to explore.

It did not take long for the dreamers and the poets to pick up this new idea. In 1657, the French soldier and adventurer Cyrano de Bergerac was perhaps the first person to speculate on "space" travel using various

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

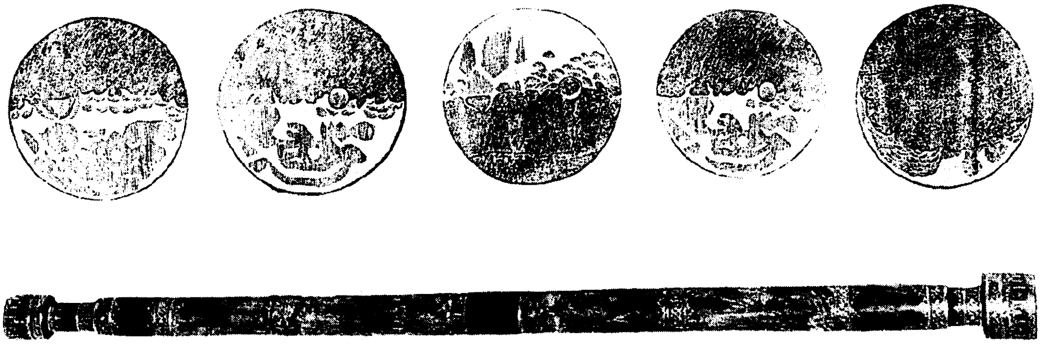

FIG. 1. Drawings of the Moon by Galileo and a picture of the telescope he used to make the drawings (from Bedini *et al.*, 1970).

means of propulsion. (de Bergerac, 1657. This was the real person on which Edmond Rostand's famous play of 1897 is based.) His means were perhaps fanciful (see Fig. 3), but there is no doubt at all that Cyrano was interested in going somewhere. He was also the first to suggest—perhaps with tongue in cheek—that rocket propulsion would be a good way to make the trip. An extensive literature on "space" travel—that is, travel to other worlds—developed even though it was not yet clear what the term "space" really meant.

During the eighteenth century, because of the development of good astronomical telescopes and the development of Newtonian mechanics, this question was cleared up. The "space" environment was understood, and it was realized that the space between the planets was "empty"—that is, there was no air to breathe. Also, it was recognized that to travel in space it would be necessary to escape from the Earth's gravitational field. All of these things came together in Jules Verne's famous novel describing a trip to the Moon that was published in 1865. Jules Verne was a Frenchman who was an attorney by profession, but there is no question that he was also technically literate. His novel *From the Earth to the Moon* describes a circumnavigation of the Moon that was eerily prescient. He talks about a journey to the Moon and back that would begin in Florida, would involve a crew of three people in a spacecraft, would execute a circumlunar trajectory, would end with a landing in the Pacific Ocean, and finally would have an American warship execute the recovery of the crew. More than one hundred years later, we launched our first circumlunar mission on 21 December 1968, Apollo 8. It was also launched in Florida, had a crew of three

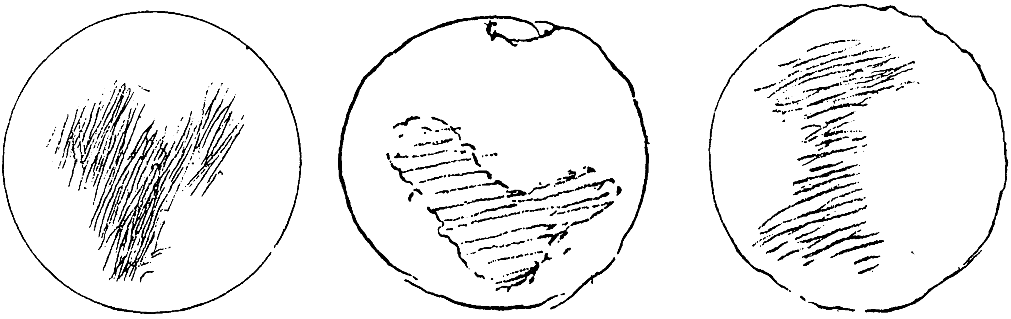

FIG. 2. Drawings of the planet Mars by Christian Huyghens. At left is one dated 28 November 1659, which shows surface markings. The middle picture shows the first observation of the Martian polar cap on 13 August 1672. The picture on the right was made on 17 May 1683. (From Flammarion, 1892.)

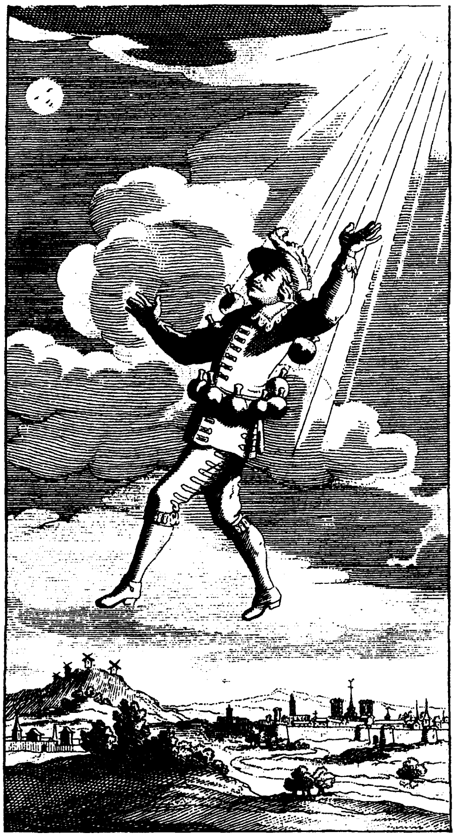

FIG. 3. One way of getting to the Moon according to Cyrano de Bergerac (1657). The method proposed rests on Cyrano's notion that dew is attracted by the Sun. Therefore, tie some bottles full of dew to your belt and, voila, you're off! (From Bedini *et al.*, 1970.)

people, landed in the Pacific Ocean, and was recovered by an American warship. The most interesting prediction that Jules Verne made in his book was that the first trip to the Moon would be made by Americans—and that is exactly how it happened.

All of the stories and events mentioned so far have involved people making these journeys. Exploration has always meant people going to "parts unknown." What has created a spirited and continuing debate in the case of space exploration and planetary exploration is the development of exquisitely sensitive and accurate sensing and data analysis that might be a substitute for people making the exploratory trips. In the past forty years, advances in electronic and computer technology have made it possible to "explore" by remote sensing. Thus, there are people who argue that it is not "necessary" to do "space" exploration with people present. All that is required is the deployment of the appropriate robotic instruments.

What does the term "exploration" really mean? The dictionary definition is simple enough: "To seek for or after; to strive to attain by search; to penetrate or range over for discovery; to examine thoroughly" (Neilson, 1956). Unfortunately, this is not terribly useful in resolving the "people vs. robots" debate in the case of space exploration. The fact of the matter is that there are always people who are true pioneers and who will want to go into the unknown even if very capable robots exist to do a similar job. Furthermore, political considerations are important as well. The political value of putting people in space has been clearly recognized by every president since Dwight Eisenhower. Human space flight has been an element of American foreign policy, the best example being the Apollo–Soyuz mission in 1975 to symbolize the policy of détente with the Soviet Union pursued by President Ford. In 1984, President Reagan initiated the Space Station program as an international effort with our allies in the "Cold War." In 1993, President Clinton invited the Russians to join the Space Station program to symbolize the new collaborative relationship being developed with Russia. These examples illustrate the importance of putting people in space in the minds of successive American Presidents to further foreign policy objectives.

In addition, there is an important domestic policy reason as well. Every President since John Kennedy has rejected advice from various quarters to shut down the effort to put people in space. All of these political leaders recognized a fundamental point: The general public does not recognize robotic missions as "exploration." It is the recognition of this factor that has kept programs involving sending people into space alive. There is, in short, a political imperative that causes political leaders to send people to explore the Moon and eventually the planets. And, as shall be seen, there is scientific value as well.

1. THE MEANS TO GET THERE

The first person to perform detailed and accurate calculations on rocket propulsion was the Russian mathematician and educator Konstantin Tsiolkovsky. It had been recognized for some time that rockets would ultimately be the way to achieve space travel. However, it was Tsiolkovsky who, in the final years of the nineteenth century, performed the quantitative calculations to show that rocket propulsion is the method of choice for space travel. Twenty years later, the German professor Hermann Oberth in his seminal book *Die Rakete zu dem Planetenraum* (*The Rocket to Interplanetary Space*), published in Germany in 1923, repeated Tsiolkovsky's calculations in more detail and also described the infrastructure that would eventually have to be developed to sustain space exploration (Oberth, 1923).

At about the same time, experimental work on rockets began. In the United States, the most important pioneer was Professor Robert H. Goddard of Clark University. In 1919, he published a paper titled "A Method for Reaching Extreme Altitudes" (Goddard, 1919), and on 16 March 1926, he launched the first successful liquid-fueled rocket. Goddard was convinced that liquid-fueled rockets would have to be developed if rockets were to be successfully applied to space travel. He described his ideas in a report to the Smithsonian early in 1920 (Goddard, 1920). During the 1930s, Goddard moved his experimental work to a site near Roswell, New Mexico. It was here that he developed the first large, gyrostabilized liquid-fueled rocket. It was flown successfully on 31 May 1935, and it reached an altitude of about 2286 m with a maximum velocity of about 1130 km/h.

The other active group in the rocket business was a German one based in Berlin and guided by an artillery officer in the German army, Captain Walter Dornberger. This group consisted of a number of more or less amateur enthusiasts who were supported partially by the military and partially by their own resources. Their activities lagged those of Goddard, but they were also convinced that liquid-fueled rockets burning gasoline or kerosene as a fuel and carrying liquid oxygen as an oxidizer would have to be developed. Several of these were successfully tested. The most important contribution of this group to rocket technology was the education and training of Wernher Von Braun, who joined the group at the age of 18 while he was a student at the University of Berlin. Von Braun went on to lead the team that developed the Saturn moon rocket in the 1960s.

The Second World War began in September 1939, and, after some hesitation, the Germans embarked on an all-out effort to develop and then produce large rockets. The same Walter Dornberger (now a colonel and later a major general) who sponsored the Berlin group in the 1930s headed this program, and young Wernher von Braun was his chief engineer. This group picked up Goddard's basic rocket configuration and produced the first successful large gyrostabilized rocket, the A-4. This vehicle could carry a payload of 900 kg out to a range of about 460 km. It was the A-4 that was later turned into the first successful long-range military rocket weapon, the V-2. The Germans fielded these rockets and launched several thousand of them both against Britain and Russia. The V-2 became the prototype for the space launchers that were developed after the Second World War.

In 1946, a detailed study was undertaken by a group at the Douglas Aircraft Company in Santa Monica, California, called "The Design of an Earth Orbiting Spaceship." This group later was separated from Douglas and became the first nucleus of what is now the RAND Corporation (Douglas Aircraft Co., 1946). The group anticipated that large rockets, based on extensions of the V-2 design, would be developed both by the United States and the Soviet Union, the two remaining world powers at the end of World War II. They then calculated what kind of rocket performance would have to be achieved to place significant payloads in Earth orbit. The relevant numbers were, of course, the orbital velocity of the "Earth-orbiting" space craft and also ultimately the velocity that would have to be achieved for a vehicle to escape from the Earth. The orbital velocity is given approximately by

$$V_0 = \sqrt{gR_E}[1 - h/2R_E], \qquad (1)$$

where g is the acceleration of gravity, h is the altitude of the orbiting vehicle, and R_E is the radius of the Earth. Equation (1) is an approximation that is only valid if h is small compared to R_E. Evaluating Eq. (1) yields velocities somewhat larger than 27 400 km/h.

Another important number is the velocity an object must reach to escape from the gravitational influence of the Earth. This velocity is given by

$$V_E = \sqrt{2gR_E}, \tag{2}$$

which, when evaluated, yields about 38 600 km/h. These numbers are the basic thresholds that must be overcome in order to make space exploration possible. The obvious point is that these velocities are very large compared to terrestrial velocities encountered under ordinary circumstances. The Douglas group concluded that reaching space would be difficult but would not be impossible in view of the large rockets that would soon become available because of the military requirements generated by the "cold war."

In the early years of space exploration, all of the launch vehicles used to put spacecraft into Earth orbit and beyond were derivatives of military rockets. Like the V-2, these were originally designed to carry payloads first to intermediate and then to intercontinental ranges. The size of these rockets was determined by the range and by the weights of the nuclear warheads that they were designed to carry. Because the United States had more efficient—and therefore smaller—nuclear weapons than the Soviets in the early years, American military rockets were smaller than those fielded by the Soviet Union. It was this factor, among others, that accounted for the early lead that the Soviets enjoyed in the highly competitive "space race" of the 1960s.

The first artificial Earth-orbiting satellite was launched by the Soviet Union on 4 October 1957. It weighed 83.6 kg, and it was launched by a rocket originally designed to launch intercontinental ballistic missiles (ICBM's), called the R-7 and renamed the SL-1. Sputnik I was placed in an elliptic orbit with a perigee of 227 km and an apogee of 941 km. This was a most significant achievement. Although the satellite carried no scientific instruments, it did have a radio beacon, and the signals from this beacon startled the world. Sputnik I was followed on 3 November 1958 by Sputnik II, which was an entirely different satellite. It weighed 508 kg and it carried a living dog named "Laika." Sputnik II was also launched by an SL-1, and it was the world's first scientific satellite. These achievements were followed by a series of satellites culminating in the first space flight by a human being, Soviet Air Force Lieutenant Yuri Gagarin. Gagarin's "Vostok I" spacecraft was launched on 12 April 1961 by an SL-1 rocket and made a circuit of the world in 108 min before coming back to Earth near the town of Saratov.

The first American attempt to launch an Earth-orbiting satellite, Vanguard I, occurred on 6 December 1957, and it ended in failure. The Vanguard satellite was a scientific research space vehicle designed as part of the "International Geophysical Year" research program slated for 1957. The first successful Vanguard was placed in Earth orbit on 26 September 1958, but this was eight months after the first successful American satellite, Explorer I, was launched on 31 January 1958. Explorer I was put in Earth orbit by a modified Redstone intermediate-range ballistic missile (IRBM) called the Juno I space launcher. The Juno I was created by Wernher Von Braun and his colleagues at the Redstone Arsenal near Huntsville, Alabama. (This was the same group responsible for the development of the V-2 rocket in Germany during the Second World War. Von Braun and many of his senior collaborators were captured and brought to the United States in 1945.) Explorer I weighed 14 kg, and it was placed in an elliptic orbit with a perigee of 360 km and an apogee of 2534 km. It carried a scientific payload designed to measure ionizing radiations above the atmosphere. The Air Force followed with the Thor IRBM, which was also quickly converted to a very capable space-launch vehicle. The highly successful Delta series of space launchers that are still in use today are based on the Thor. Finally, there was the first American intercontinental ballistic missile (ICBM), the Atlas, also developed by the U.S. Air Force. It was a suitably modified Atlas that was used to carry the first American, Lieutenant Colonel John H. Glenn of the U.S. Marine Corps, into orbit around the Earth in a Mercury space capsule on 20 February 1962. The purpose of the Mercury program was to prove that people could survive in the Earth-orbit environment and to bring them safely back to Earth. The Mercury capsule had a blunt heat shield, and it was designed to splash down in water.

All of these rockets used gasoline or kerosene as a fuel and liquid oxygen as an oxidizer. While this provided a high-energy mixture for efficient propulsion, it had the drawback that it was difficult to store liquid oxygen in order to keep these military missiles on instant alert.

To overcome this difficulty, rocket fuels were developed that did not require cryogenic storage systems. The most important of these were hydrazine-fueled rockets that used nitrous oxide as the oxidizer to burn the fuel. The ingredients of this mixture have the advantage that they can be stored indefinitely at room temperature. Also, hydrazine and nitrous oxide will ignite on contact (this is called a hypergolic mixture) so that no ignition system is necessary to start the rocket. The largest American ICBM, the Titan, is based on this principle. The Titan was also converted rapidly into a space-launch vehicle. Titan II rockets were used to put the two-man Gemini spacecraft into Earth orbit for a first flight on 23 March 1965, carrying Astronauts Virgil Grissom and John Young. The Gemini program was a follow-on to Mercury. The purposes of the Gemini flights were to develop orbital maneuvering methods, to test the feasibility of docking, and to practice extravehicular activity (EVA). All of these maneuvers would be employed later on during our trips to the Moon.

The generation of ICBM's following the Atlas and the Titan, the Minuteman I, II, and III missiles, used solid fuels for propulsion. From the military viewpoint, solid-fueled rockets are much preferable to those using liquids for propulsion because solid fuels are much easier to store than liquids, even if the liquids do not require cryogenic cooling. Thus, maintaining solid-fueled rockets in a high state of alert is relatively easy. The price that must be paid for this advantage is that solid-fueled rockets are relatively less efficient and less controllable than those that burn liquid fuels. The solid-fuel technology that was the basis of the Minuteman ICBM system also quickly found a useful application in space-launch vehicles. Solid-fueled "strap-on" rockets are very commonly used on all space launchers from the relatively small Delta rockets to the Space Shuttle.

1.1 The Principles of Rocket Propulsion

It might be worth while here to digress briefly and develop some of the quantitative relationships that govern rocket propulsion. Consider the schematic rocket shown in Fig. 4. The rocket is launched from the surface of the Earth, and it burns fuel at a specified rate that produces exhaust gases that exit from the rocket nozzle with a velocity v_E. The rocket itself has a velocity V_R, which is a function of time as long as the fuel is burning. A force balance can now be written that has the usual inertial term on the left-hand side of the equation. On the right-hand side, there are two terms, one that represents the momentum transferred to the rocket vehicle per unit time and the other that accounts for the force of gravity:

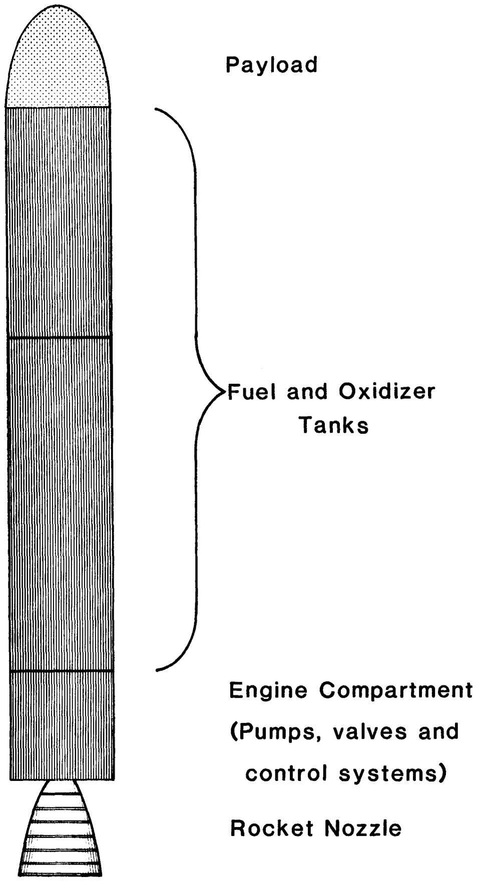

FIG. 4. Schematic diagram of a rocket as it would be launched from the surface of the Earth. The figure shows the fundamental elements of a rocket system: the payload, the fuel, and the rocket, which includes the weight of the fuel tank and the rocket propulsion system.

$$M(t)\frac{dV_R}{dt} = v_E\frac{dM(t)}{dt} - M(t)g. \tag{3}$$

The mass of the entire vehicle is the sum of three terms: the mass of the payload, m_p; the mass of the rocket (which includes the fuel tanks and the rocket engine), m_R; and the mass of the fuel, $m_F(t)$. The mass of the fuel will be a function of time since it is being consumed, and it is for this reason that the mass of the entire vehicle is written explicitly as a function of time in Eq. (3):

$$M(t) = m_P + m_R + m_F(t). \tag{4}$$

Equation (3) can be rewritten as a simple differential equation:

$$\frac{dV_R}{dt} = \frac{dM(t)}{M(t)} - \frac{gdt}{v_E}, \tag{5}$$

where g is taken as constant, which is a good approximation if the rocket is near the surface of the Earth. At the beginning of the flight, before the rocket starts moving, it can be seen from Eq. (5) that the first term on the right-hand side must exceed the second before the rocket can lift off. This says that the rate at which fuel is burned and ejected from the rocket must exceed the force of gravity. The important parameter here is g/v_E, the constant that appears in the second term. This is usually inverted and is called the "specific impulse" of the rocket engine:

$$\tau_s = v_E/g. \tag{6}$$

The specific impulse has the dimensions of time, and it depends only on the exhaust velocity of the rocket. This quantity can be estimated from thermodynamic considerations that depend on the conditions (i.e., temperature) inside the rocket's combustion chamber and on the efficiency (i.e., shape) of the rocket nozzle. It can be shown that approximately,

$$v_E \approx \sqrt{3kT/m}, \tag{7}$$

where T is the temperature in the chamber and m is the molecular weight of the exhaust gases. From Eq. (5), it is obvious that the larger the exhaust velocity v_E, the easier it is to get the rocket off the ground. Thus, high temperatures and low molecular weight of the exhaust gases are important. Good current technology rockets can achieve exhaust velocities of about 3050 m/s. This translates to a specific impulse for the rocket of about 300 s. The advantage of rockets burning liquid hydrogen as a fuel over those burning carbon- or nitrogen-based fuels can also be seen from Eq. (7). The combustion product of a hydrogen-burning rocket is water, which has a molecular weight of 18. In the case of carbon- or nitrogen-based fuels, the molecular weight of the exhaust products is about 40, so that hydrogen-fueled rockets have an exhaust velocity that is about 40% higher.

In order to see what happens at the end of the rocket's flight, trajectory, it is necessary to integrate Eq. (5):

$$\frac{1}{v_E}\int_0^{V_R(\text{final})} dV_R = \int_{m_R+m_P+m_F(0)}^{m_R+m_P} \frac{dM}{M} - \frac{1}{\tau_s}\int_0^{\tau_B} dt. \tag{8}$$

Performing the indicated integrations yields

$$\frac{V_R(\text{final})}{v_E} = \log_e \frac{m_R + m_P + m_F(0)}{m_R + m_P} - \frac{\tau_B}{\tau_s}, \tag{9}$$

where the quantity τ_B is the time to burnout of the rocket.

The performances of an "average" rocket can be determined by looking at Eq. (9). It is not easy to "solve" from first principles, but typical numbers look like this:

1. To get to Earth orbit, the ratio of the final velocity of the rocket, V_R(final), to the exhaust velocity, v_E, is about 2.3 since V_R(final) must be about 27 400 km/h.
2. The time to burnout of a typical orbital rocket, τ_B, is about 10 min so that τ_B/τ_s is about 2.

Therefore, the evaluation of Eq. (9) yields

$$\log_e \frac{m_R + m_P + m_F(0)}{m_R + m_P} \approx 4.3. \tag{10}$$

Inverting the logarithm yields

$$\frac{(m_R + m_P) + m_F(0)}{m_R + m_P} = 1 + \frac{m_F(0)}{m_R + m_P} \approx 100. \tag{11}$$

The ratio of the fuel mass to the sum of the payload and rocket masses is

$$\frac{m_F(0)}{m_R + m_P} \approx 100, \tag{12}$$

since the fuel mass, $m_F(0)$, is much larger than $m_R + m_P$. This result says that it takes about 100 kg of fuel to get 1 kg of material in Earth orbit. This ratio is a little bit larger than typical; but this can be accounted for by the fact that all practical orbital rockets are staged vehicles, which are more efficient than the single-stage rocket considered here.

2. THE OBJECTIVES OF SPACE EXPLORATION

Two important intellectual traditions have dominated the development of the objectives of space exploration. One thread runs from the science fiction writers, Jules Verne, H. G. Wells, Isaac Asimov, Arthur Clarke, and others, that involve people going to other worlds and exploring in the tradition of Christopher Columbus, Vasco da Gama, James Cook, and Robert E. Peary. This type of exploration involves the presence of people and the taking of risks—sometimes great risks. The second thread originates in atmospheric research. This originated with the use of high-altitude balloons to do meteorological research, and in the 1930s the use of small sounding rockets for high-altitude research. Many of the people who originated this research, James Van Allen, Herbert Friedman, and John Simpson, among others, later became prominent in space science research. (See, for example, Van Allen, 1958, which contains many good references to previous sounding-rocket experiments).

In a remarkable series of articles that appeared in *Collier's* magazine between March 1952 and April 1954, a distinguished group of scientists, engineers, and writers laid out the objectives of exploration that would involve sending people into space. (Ryan, 1952–1954) Appropriately enough, the lead article of the series is called "Man Will Conquer Space Soon." Among the authors were Wernher Von Braun (who was building large rockets for the U.S. Army by then); Joseph Kaplan, a professor of physics at UCLA; Fred Whipple, an astronomy professor at Harvard University; and other experts in the medical and legal aspects of space flight, among other things. The spectacular illustrations were provided by Chesley Bonestell, the distinguished artist who specialized in creating the first realistic images of the solar system. These articles presented the program of human space exploration that have been executed ever since the first American spacecraft was launched into orbit around the Earth on 31 January 1958. There are "winged rockets" (today we call these the "space shuttle"), there are space stations, there is the trip to the Moon and the establishing of a permanent colony on the Moon, and finally, there is the trip to Mars. All of the objectives of human space exploration are outlined in these articles in precise detail. Now, some 40 years after these articles were published, about one-third of the program that was outlined has been completed.

In addition to technical matters, the *Collier's* articles also deal with the motivations and the risks of human space flight. One of the articles is completely devoted to describing how astronauts could be rescued by "lifeboats" in space if something went wrong. There are good scientific reasons for incurring the risks of human space flight. First and foremost is the question of how people behave in and react to zero gravity. Second, there are scientific experiments to be conducted either on a space ship or on the surface of the Moon or some asteroid or planet that absolutely require the presence of human judgment and intelligence for their successful execution. There are those who argue that the knowledge accumulated by such work is not worth the risk involved. This is obviously a matter of individual judgment that each person must make for himself. There is such a thing as the imperative to explore the unknown, and there are some who argue that taking great risks in the name of exploration is one of the defining qualities of human beings. Ultimately, it is the exploratory urge together with the political imperative already mentioned that, in addition to the scientific objectives, justifies human space flight.

The first important scientific discovery in space was made in 1958 by the first two spacecraft that were placed in orbit by the United States, Explorer I (31 January 1958) and Explorer III (26 March 1958). Both of these spacecraft were equipped with radiation detectors that were designed and built by Van Allen and his collaborators at the University of Iowa. The discovery was that the Earth was surrounded by a field of energetic charged particles trapped by the Earth's geomagnetic field in spiraling orbits. These were called the "Van Allen radiation belts" after their discoverer. This discovery was very much in the tradition of looking at electromagnetic and auroral phenomena in the Earth's upper atmosphere using balloons and sounding rockets. It has been followed by a long series of important and sometimes startling events in planetary exploration, space-based astronomy, Earth observations, and the physics of the Sun and the space environment. Because of the absence of risk to human life, the NASA Space Science Program has been popular and noncontroversial. The only policy question of importance that has been raised periodically is the level of funding. NASA has traditionally spent more on space science annually than the National Science Foundation has spent on all other branches of science. There are some who claim that this "imbalance" is not healthy in the long run.

The other comparison that needs to be made is the level of spending on space science compared to other NASA programs. Ever since it was founded, NASA has spent somewhere between 15 and 20% of the total NASA annual budget on space science (including earth observations). As a result of this, the highest level of spending on space science was in 1967 during the peak spending year of the Apollo program. There is no specific rationale for this fraction other than historical experience. Nevertheless, the 15–20% figure was strongly endorsed in 1990 by the President's Advisory Committee on the Space Program headed by Norman R. Augustine (Augustine, 1990). The same group also strongly endorsed putting scientific research in space at the highest level of priority in NASA. Thus, scientific research has been done and will continue to be done to expand human knowledge, and it will be well funded as long as NASA can maintain a strong program of human space flight that can maintain a significantly higher spending level for reasons other than the pursuit of scientific knowledge.

3. THE SPACE ENVIRONMENT

The discovery of the Van Allen radiation belts was the first important finding of space science. It has been followed in the past thirty years by a long series of other discoveries that have led to a detailed understanding of the interplanetary medium. Even before the advent of space flight it was known that the Sun emitted charged particles that created a fast-moving plasma called the "solar wind." It is the interaction of the solar wind with the Sun's active and varying electromagnetic field that defines the environment in which the planets execute their own motions around the Sun. The environment was defined by a series of spacecraft sent into orbit around the Sun in the 1960s and 1970s. A typical example of such a spacecraft is shown in Fig. 5. The four Pioneer spacecraft (numbered 6 through 9) were equipped with radiation detectors, plasma probes, and magnetometers that were used map the detailed properties of the solar wind. Because of the Sun's complex and time-varying magnetic field, the solar wind also has a complex and time-dependent structure. By correlating measurements of the plasma and the magnetic fields from several different spacecraft, a dynamic map of the solar wind has been constructed.

The solar wind is the most important source of the charged particles that have been found trapped in the magnetic fields of the Earth, Jupiter, and some of the outer planets. Because of the strong fluctuations of the solar magnetic fields, "holes" can be created in the outer regions of the Earth's magnetic dipole that permit particles to become trapped. The condition that must be met is that when the Earth's field lines close again, the particles with velocities making angles less than a critical angle θ_c with respect to the normal of the field line B where the particle is injected will be trapped:

$$\sin^2\theta_c = B/B_{\max}, \tag{13}$$

where B is the magnetic field value where the particle enters the dipole and $B_{\max}$ is the magnetic field where the particle is "reflected" just before it enters the atmosphere. The approximate flux of the charged particles trapped in the geomagnetic field can be understood by using the solar wind as a

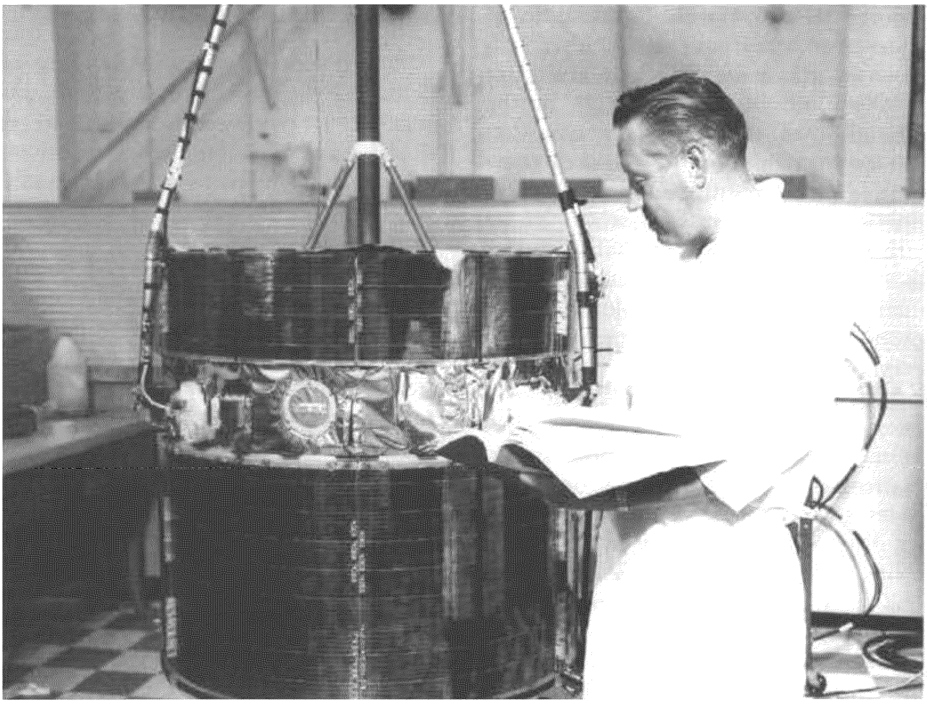

FIG. 5. The Pioneer 6 spacecraft that together with its companions in the Pioneer program provided the first comprehensive map of the solar wind.

source and then applying Eq. (13) to estimate the fraction of particles trapped in the dipole fields assuming certain time-dependent mechanisms to open the outer field lines.

Another mechanism for putting energetic charged particles in the geomagnetic field is the impact of primary cosmic rays on atoms in the Earth's atmosphere. Some of these impacts cause nuclear reactions that produce neutrons. Some of these neutrons—the so-called albedo neutrons—move in directions opposite to the incident cosmic ray and return to regions in the Earth's magnetic field where the atmospheric density is very low. Neutrons are unstable, and they decay into a proton and an electron with a mean life of about 15 min:

$$n \rightarrow p + e^{-} + \nu + 0.275 \text{ MeV}. \qquad (14)$$

If one or both of the charged particles produced by the reaction shown in (14) obey the condition (13), then they will be trapped in the geomagnetic field. Since the reaction is more energetic—on the average—than the charged particles that constitute the solar wind, the high-energy end of the spectrum of Van Allen belt particles is accounted for by the albedo neutrons.

Several spacecraft have been launched to look at the energy and the mass spectrum of the particles that constitute primary cosmic rays. The most important of these spacecraft is the "Einstein" observatory, which is discussed in subsequent sections of this article. Cosmic rays are the result of complex processes that occur in supernova events and also from the varying magnetic field in the Galaxy.

The "edge" of the solar system is loosely defined as the region where the dynamic pressure exerted by the solar wind equals the static pressure of the gas in the interstellar medium. The place where this condition occurs can be detected because a standing "shock" wave would exist where the "supersonic" solar wind strikes the "stationary" interstellar medium. The Pioneer 10 spacecraft, which was launched in March of 1972 and flew past the planet Jupiter in 1973, is likely to be the first manmade object to reach this boundary. The spacecraft is still in contact with the Earth although it is now beyond the orbit of the planet Pluto and is about 7.24 × 10^9 km from Earth. Hopefully, the spacecraft will stay in contact long enough to permit us to see the "edge" of the solar system (Hall *et al.*, 1974, 1976).

The other important aspect of the space environment is its effect on human beings when they fly in space. There are two conditions that need to be considered: Zero gravity is one and the radiation field produced by cosmic rays and by emissions from the Sun is the other.

There is now considerable experience in determining what happens to people when they experience lengthy exposure to the zero-gravity environment. The first really definitive knowledge was gained during the Skylab flights executed by the United States in 1973. The Skylab, the first habitable "space station" placed in Earth orbit, was actually a follow-on to the Apollo lunar landing program that was canceled in 1972. The Skylab spacecraft itself was a modified Saturn IV B stage that was rebuilt so that it could accommodate a crew of three people. (For a complete description of the Saturn launch vehicle and the Apollo spacecraft, see Sec. 5.) It was launched in the unmanned mode on 14 May 1973 and then was visited by three crews starting in July 1973. The longest resident crew was the final one, which stayed in the Skylab for 84 days. The operation of Skylab was terminated in February 1974. Unfortunately, the orbit of Skylab was low enough so that it would decay in a few years. Originally, the hope was that the Space Shuttle would be ready in time to move Skylab into a higher-altitude orbit so that the capability of this facility could be preserved and eventually even improved. However, because of delays in the Shuttle program and an unusually active solar maximum in 1977 and 1978, the Skylab burned up in the Earth's atmosphere on 11 July 1979.

Subsequent to the Skylab activities in 1973 and 1974, the Soviets conducted an extensive series of long-duration space flights. In these, the Soyuz, Salyut, and Mir space vehicles were employed. Both Salyut and Mir are what might be called "mini" space stations that are well suited to the purpose. Soviet (now Russian) cosmonauts have achieved residence times in excess of 1 yr in Earth orbit. A more detailed discussion of these spacecraft is presented in Sec. 5.

Many things have been learned in these flights about the reaction of the human physiology and psychology to the space environment. There are three concerns that seem to be most important: the deconditioning of the cardiovascular system, the alteration of the calcium balance between the bones and the blood, and the long-term effects of ionizing radiations, which include long-term effects of cosmic rays and short-term effects caused by intense bursts of x rays from the Sun. Exercise regimes have been devised to deal with cardiovascular deconditioning and with bone problems. These have been partially successful, but they do not completely overcome the problems that have been encountered. Exposure to radiation is not a serious problem in near Earth orbit since the Earth's magnetic field deflects most of the dangerous radiation. On longer flights outside of the Earth's immediate environment, such as the Apollo missions, radiation becomes an important factor. The Apollo astronauts observed "light flashes" even when their eyes were closed that were caused by Cherenkov radiation due to primary cosmic rays passing through the vitreous humor of the eye. Fortunately, the Apollo missions were conducted at a time when the Sun was not very active. Thus x rays caused by solar prominences were not a serious problem. There are proposals to equip future spacecraft with shielded "shelters" in which astronauts can protect themselves against these x-ray bursts. Nothing can be done about protection against cosmic rays because their energies are so high that no protective shielding can be provided. Fortunately, the cosmic radiation flux is relatively low so that it is not an immediate problem. However, exposure to cosmic rays is the eventual barrier to very long (i.e., many years) human space missions.

The high vacuum of space has also been a useful research tool. Conditions of low pressure can be created by what are called "wake shields" on spacecraft that cannot be reproduced in vacuum chambers on Earth. Experiments have been conducted at the very low pressures created by "wake shields" that involve chemical reactions with atomic oxygen and other atomic species that cannot be created in large quantities on Earth. Also, wake-shields experiments have provided useful information about plasma physics. These experiments are mostly of engineering interest in terms of materials erosion, drag measurements, and other practical results.

What has been discovered in the past four decades about the "empty" space surrounding the Earth is that it is a lively place.

4. THE SUN

The Earth's environment is dominated by the Sun. The Sun is a roughly average star about 333 000 times as massive as the Earth and it has a radius of 700 720 km, which is 109 times that of the Earth. The Sun produces energy from thermonuclear reactions in the core that create a surface temperature of about 6000°C and a total energy flux at the surface of the Earth of about 1 kW/m^2. The Sun rotates on its axis about once every 27 days, and it has a complex magnetic field that rotates with it. The complex motions of the Sun also create turbulence in the upper solar atmosphere that cause sunspots and solar prominences. Since the Sun is hot, some of the plasma "boils off" and becomes the solar wind that has already been mentioned. A complex structure is imposed on the solar wind by the Sun's rotating magnetic field, and this structure has been unraveled by Pioneer and Explorer Spacecraft (Jordan, 1981).

The dynamics of sunspots and solar prominences involve higher-energy processes, and these produce particles and radiations of much higher energies than those that are characteristics of the Sun's surface temperatures. It is these phenomena that give rise to the emission of x rays, high-energy protons and electrons, and even neutrons. These phenomena were actually discovered by Earth-based observations, balloons, and sounding rockets before the advent of orbital space flight. Nevertheless, the Sun was one of the early targets of early Earth-orbiting spacecraft. The first of these was a series of Orbiting Solar Observatories (OSOs), the first of which was launched in 1962. The idea was to look at the Sun with x-ray, ultraviolet, and gamma-ray detectors. This is, of course, the essence of what is meant by "broad-band astronomy." Eight successful flights were made by orbiting solar observatories between 1962 and 1975. The most important results of these flights were detailed examinations of solar flares. The OSOs were placed into orbit around the Earth and were pointed at the Sun to make these observations, using an appropriate pointing system. Two German spacecraft were placed in orbits around the Sun, Helios I and Helios II, to observe surface conditions on the Sun more closely and to make measurements of the solar wind close to the Sun. Helios I was launched on 10 December 1974 and flew to within 48×10^6 km of the Sun's surface. (Note: This is one-third of the distance from the Sun to the Earth, and it is inside the orbit of the planet Mercury.) Helios II, launched on 15 January 1976, came even closer, to within 45×10^6 km. There is no doubt that these spacecraft, as well as many others, launched by Europeans and Russians have provided a detailed picture of how the Sun works.

The Skylab program, which has already been mentioned, also made important contributions to the knowledge of the Sun. The Skylab spacecraft included a large telescope designed to observe the Sun called the Apollo Solar Telescope (AST). This instrument took hundreds of pictures of the Sun, and it was serviced by the astronauts during the Skylab missions.

Perhaps the most ambitious space mission to explore the Sun is "Ulysses." This is a large spacecraft designed to explore the Sun's emissions and magnetic fields in detail over many years. Like "Helios," "Ulysses" is also a European-built spacecraft, but it was launched by the American Space Shuttle "Discovery" on 6 October 1990. Heretofore, spacecraft have been able to explore solar emissions only very close to the ecliptic plane. "Ulysses" flew past the planet Jupiter on 8 February 1992 where, through an orbital-momentum transfer process between the spacecraft and planet (called the "slingshot" effect), the spacecraft was placed in an orbit that makes an angle of about 30° with the ecliptic plane. "Ulysses" will explore solar emissions at much higher latitudes than measurements made in previous flights. Thus "Ulysses" will open a new chapter in the understanding of solar dynamics and particle emissions.

The most important unexplained mystery

about the Sun is that the observed neutrino flux from the star seems to be smaller by a factor of about 3 than what would be expected from the thermonuclear reactions that are going on inside the core. There is some reason to believe that this is actually a very fundamental problem. It is likely that the eventual resolution of this mystery will involve more observations involving the use of spacecraft.

5. THE MOON

On 25 May 1961, President John F. Kennedy made the commitment to place men on the Moon before 1970. Thus began the first major effort to explore another body in the solar system, the program called Apollo. The Earth's satellite, the Moon, is one of the largest satellites relative to its "parent" planet in the solar system. Only Charon, the satellite of Pluto, is larger relative to Pluto. In a real sense then, the Earth–Moon system might be called a "double planet." The Moon has a diameter of about 3218 km (compared to about 12 872 km for the Earth); it has a density of 3.3 g/cm^3, which is somewhat smaller than that of the Earth (5.5 g/cm^3), and a total mass of approximately 1.23% of that of the Earth. The Moon revolves around the Earth once a month and it also revolves around its own axis once a month. Thus, the same face of the Moon always points toward the Earth.

5.1 The Apollo Program

In order to land astronauts on the Moon in a safe manner, a number of robotic missions were conducted as precursors intended to explore the lunar environment. These began with the "Ranger" missions that were "hard landers," essentially spacecraft that were intended to crash on the Moon to determine the mechanical properties of the lunar surface. These were followed by the "Surveyor" missions that were "soft landers," which were intended to make more sophisticated measurements of the composition of the lunar surface and other elements of the lunar environment. Finally, there were orbiters that took very detailed pictures of the lunar surface so that appropriate landing sites could be selected. These pictures were much better than anything that could be obtained from the surface of the Earth.

In order to put people on the Moon, a new space launcher had to be developed. There were three ways of going to the Moon that were considered:

1. a direct ascent and landing;
2. a rendezvous in Earth orbit, from where the lunar landing vehicle would be launched; and
3. a rendezvous in lunar orbit, from which the lunar landing vehicle would be operated.

These alternatives are illustrated in Fig. 6. After some discussion, the lunar-orbit rendezvous method was selected. The launch vehicle built to do this job was called the Saturn V. A picture of the Saturn V is shown in Fig. 7 as it looked ready for take-off. The lunar-orbit rendezvous method was actually a compromise between the direct-ascent method that requires the largest rocket and the Earth-orbit rendezvous method that would have taken the longest time to execute and, therefore, would have been more expensive.

The Saturn V launch vehicle was the largest rocket ever constructed. It was also the first American space launcher that was not a direct descendent of the military rockets discussed in Sec. 1. The Saturn V was a three-stage vehicle that was 111 m long and weighed more than 3400 tons. The Saturn V was capable of placing a payload of approximately 113.4 tons into near-Earth orbit. The Saturn V vehicle consisted of a first stage (the Saturn I) with five F-1 engines that burned gasoline and liquid oxygen. The total thrust developed was nearly 8×10^6 lb. The second stage, the Saturn II C, was powered by five engines, the J-2S, that burned liquid hydrogen as a fuel and carried liquid oxygen as the oxidizer. The third stage, called the Saturn IV B, also was powered by a single J-2S engine that burned liquid hydrogen and liquid oxygen. The purpose of this stage was to move the lunar vehicle from Earth orbit into a trajectory going to the Moon. The last three elements of the Saturn V vehicle consisted of the Command Module, the Service Module, and the Lunar Excursion Module, all of which went to the Moon. They were originally packaged, as shown in Fig. 7, at the very top of the vehicle. On the way to the

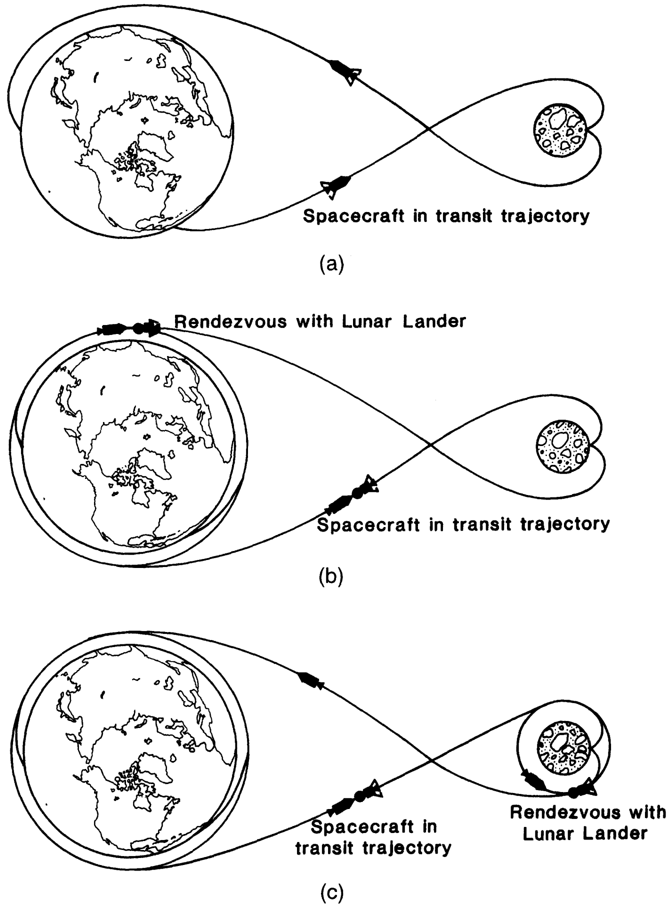
Spacecraft in transit trajectory
(a)
Rendezvous with Lunar Lander
Spacecraft in transit trajectory
(b)
Spacecraft in transit trajectory
Rendezvous with Lunar Lander
(c)

FIG. 6. The three different methods for reaching the Moon: (a) direct ascent and landing, (b) Earth-orbit rendezvous, and (c) lunar-orbit rendezvous.

Moon, a docking maneuver was executed that put the lunar excursion module in the configuration shown in Fig. 8, so that two of the three astronauts in the Command Module could move to the Lunar Excursion Module in preparation for the landing on the Moon. (Some of the expended Saturn IV B stages were set on a trajectory so that these vehicles would eventually crash on the Moon and would cause seismic activity on the Moon that would be picked up by seismometers left on the lunar surface by prior Apollo missions.)

Once the three-piece spacecraft was placed in lunar orbit, the lunar excursion module was flown to the surface of the Moon by the two-man crew. The Apollo crews then spent up to 22 h on the lunar surface exploring and conducting a variety of scientific experiments. Upon conclusion of this work, they returned to the Lunar Excursion Module. The descent stage with its rocket was jettisoned and was left on the Moon. The ascent stage then returned to the Command Module and the attached Service Module in its orbit to perform the return docking maneuver. The two astronauts then rejoined their orbiting colleague in the Command Module. When the docking maneuver was completed, the engine on the Service Module was used to move the spacecraft out of lunar orbit and on to a trajectory back to Earth. The Lunar Module was then dropped off on the way back to the Earth. The final maneuver was to jettison the Service Module and then to put the Command Module back to Earth in the same manner as the Mercury and the Gemini spacecraft.

During the Apollo program, two successful circumlunar missions were flown (Apollo 8 and Apollo 10), and six two-man crews landed on the surface of the Moon. The his-

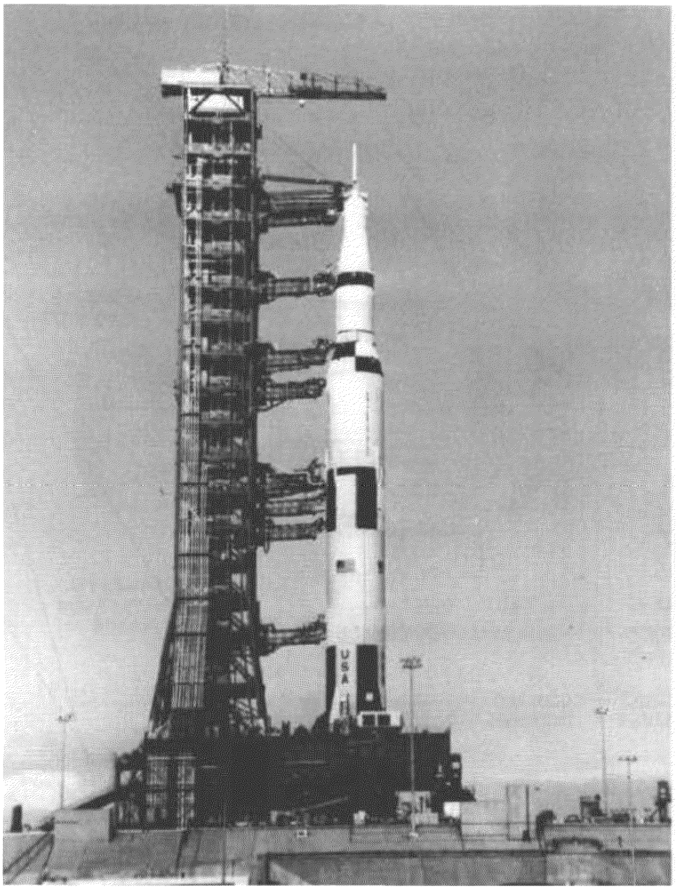

FIG. 7. The Saturn launch vehicle shown on the launch pad at the NASA John F. Kennedy Space Center ready to take off (courtesy NASA).

toric first landing on the moon was carried out by the Apollo 11 mission on 20 July 1969. The mission commander was Neil Armstrong, and he was accompanied by Edwin (Buzz) Aldrin and Michael Collins. A summary of the Apollo missions is shown in Table 1.

There is no question that the Apollo missions had a profound impact both as a great human adventure and in terms of scientific results. Most important from the political viewpoint was that Americans landed on the Moon and the Soviets failed to do so. The United States thus took a firm lead in the "space race," and the objective set by President Kennedy in 1961 was achieved. Unfortunately, by the time that Neil Armstrong and his colleagues became the first human beings to set foot on another world, the political imperatives that drove the program in 1961 were no longer present. In 1970, President Nixon announced the cancellation of the Apollo lunar program. The Skylab program of 1973 and 1974, which has already been mentioned, was a brief follow-on. Finally, the symbolic Apollo–Soyuz Test Program (ASTP) was successfully executed in 1975 as part of the policy of "détente" with the Soviet Union initiated by Secretary of State Henry Kissinger. Although no scientific results were produced, the ASTP mission was a welcome sign of the beginning of international collaboration in space rather than competition.

The exploration of the Moon was an enterprise within the grand tradition of human exploration. It is very likely that a century from now, the only thing that will be remembered about the year 1969 is the landing of Neil Armstrong and Edwin Aldrin on the Moon. This is what happened with 1492, in which case the main thing remembered is the first step Christopher Columbus took on San Salvador Island on October 12 of that year.

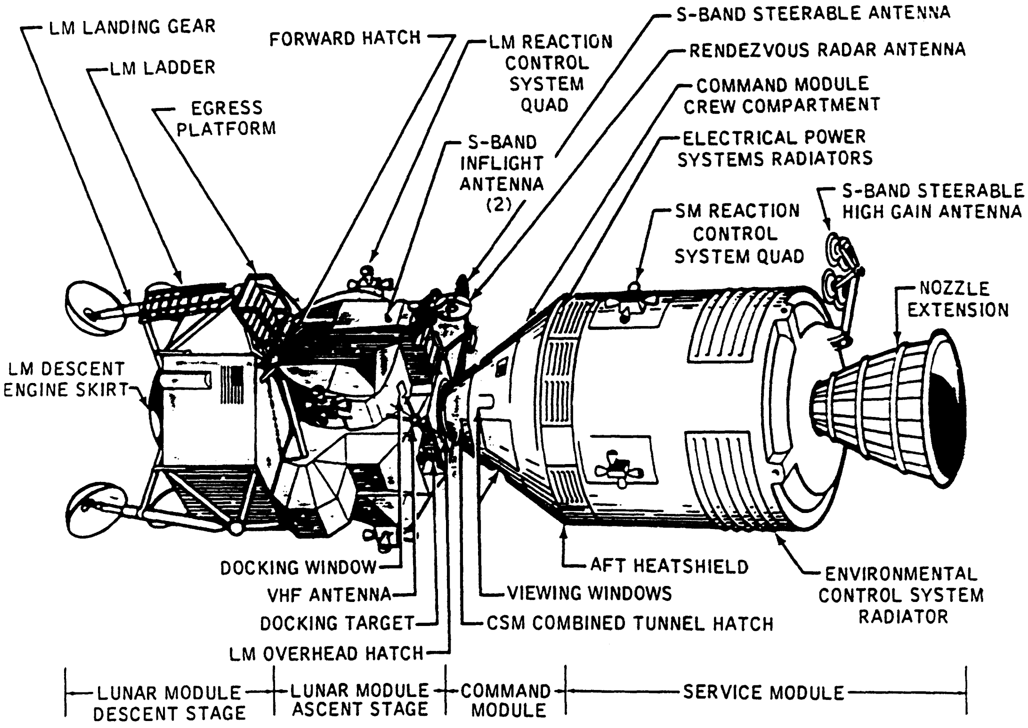
LM LANDING GEAR
LM LADDER
EGRESS PLATFORM
FORWARD HATCH
LM REACTION CONTROL SYSTEM QUAD
S-BAND INFLIGHT ANTENNA (2)
S-BAND STEERABLE ANTENNA
RENDEZVOUS RADAR ANTENNA
COMMAND MODULE CREW COMPARTMENT
ELECTRICAL POWER SYSTEMS RADIATORS
SM REACTION CONTROL SYSTEM QUAD
S-BAND STEERABLE HIGH GAIN ANTENNA
NOZZLE EXTENSION
LM DESCENT ENGINE SKIRT
DOCKING WINDOW
VHF ANTENNA
DOCKING TARGET
LM OVERHEAD HATCH
AFT HEATSHIELD
VIEWING WINDOWS
CSM COMBINED TUNNEL HATCH
ENVIRONMENTAL CONTROL SYSTEM RADIATOR
LUNAR MODULE DESCENT STAGE
LUNAR MODULE ASCENT STAGE
COMMAND MODULE
SERVICE MODULE

FIG. 8. The Apollo spacecraft as configured on its journey to the Moon (courtesy NASA).

5.2 Scientific Results from the Apollo Lunar Exploration

The Apollo program yielded some very important scientific results. The composition of the material found on the lunar surface was accurately established. It was found to be about 75% silicon dioxide (rock) with significant quantities of iron, aluminum, and magnesium and a surprisingly large amount of titanium. In addition to the gross composition, a great deal about crater formation, lunar lava flows, and other surface structures was learned from a very detailed analysis of the lunar material that was returned. A rather detailed picture of the internal structure of the Moon was deduced from seismic and magnetic-field measurements made with instruments placed on the lunar surface by the Apollo astronauts. The Moon probably has no liquid core, which is inferred from the fact that it has no magnetic field. However, the seismic data did reveal a layered structure with a heavy (probably iron) core and a lighter mantle and crust. Finally, the parameters of the lunar orbit around the Earth were very precisely established. This was done by placing corner reflectors at various places on the Moon. These were then illuminated by very short neodymium YAG laser pulses, which were transmitted to the Moon through the Harlan J. Smith reflecting telescope of The University of Texas McDonald Observatory. A lengthy series of measurements permitted the determination of the position of the Moon relative to the Earth within a few centimeters. Using this method, a precise measurement of the rate at which the Moon is receding from the Earth (about 2.5 cm/yr) was made. This, in turn, permitted the development of better methods to estimate the magnitude of the dissipative tidal forces that cause the radius of the lunar orbit to increase slowly as time goes on.

Probably the single most important scientific result of the Apollo lunar exploration was a more precise measurement of the age of the Moon. This was done by comparing the relative content of uranium-235, thorium-232, and potassium-40 in rock samples

Table 1. Summary of the Apollo missions (from Torres, 1986).

Mission	Date	Astronauts	Mission duration (d:hr:min)	Earth orbits	Lunar orbits	Miles traveled	Lunar samples returned
Apollo 7	Oct 11–22, 1968	Walter Schirra Donn Eisele Walter Cunningham	10:20:09	160	0	4 539 959	—
Apollo 8	Dec 21–27, 1968	Frank Borman James Lovell William Anders	06:03:00	1.5	10	578 720	—
Apollo 9	Mar 3–13, 1969	James McDivitt David Scott Russell Schweickart	10:01:00	151	0	4 208 193	—
Apollo 10	May 18–26, 1969	Thomas Stafford John Young Eugene Cernan	08:00:03	1.5	31	828 169	—
Apollo 11	Jul 16–24, 1969	Neil Armstrong Michael Collins Edwin "Buzz" Aldrin	08:03:18	1.5	30	950 598	46 lb
Apollo 12	Nov 14–24, 1969	Charles Conrad Richard Gordon Alan Bean	10:04:36	1.5	45	950 031	75 lb
Apollo 13	Apr 11–17, 1970	James Lovell John Swigert Fred Haise	05:22:54	1.5	Flyby	621 316	—
Apollo 14	Jan 31–Feb 9, 1971	Alan Shepard Stuart Roosa Edgar Mitchell	09:00:01	1.5	34	1 148 560	95 lb
Apollo 15	Jul 26–Aug 7, 1971	David Scott Alfred Worden James Irwin	12:07:11	1.5	74	1 272 187	170 lb
Apollo 16	Apr 16–27, 1972	John Young Thomas Mattingly Charles Duke	11:01:51	1.5	64	1 388 429	209 lb
Apollo 17	Dec 7–19, 1972	Eugene Cernan Ronald Evans Harrison Schmitt	12:13:51	2	75	1 482 721	258 lb

returned from the Moon. These are long-lived isotopes with mean lives of 4.5×10^9, 14×10^9 and 1.2×10^9 yr respectively. Rock samples almost 4×10^9 yr old were returned from the moon. From these results, and theoretical considerations, it was concluded that the age of the Moon is between $(4.5 \text{ and } 4.7) \times 10^9$ yr. This number by inference is also very close to the age of the solar system. Thus these measurements provide an accurate timescale that has so far been obtained to mark geological events on Earth, the length of time that nuclear processes have been going on in the Sun, and the times at which various events in the development of living things have occurred (Heiken *et al.*, 1991).

There are, of course, many open questions that remain. In a real sense, Apollo was no more than a preliminary reconnaissance, and detailed exploration will have to wait until people once again walk on the Moon. One of the recurring questions about the Moon is whether there are traces of water (ice) or perhaps hydrated rocks on the Moon. No evidence of water was found in any of the samples returned by the Apollo astronauts. The importance of finding water has to do with the use of lunar hydrogen (if it exists) as a rocket fuel. It would be very advantageous for future exploration missions in the solar system if water could be mined on the Moon. Since it is very much easier to lift things from the Moon than from the Earth, the use of the Moon as a "cosmic gas station" could greatly facilitate extended trips around the solar system in the future (Arnold, 1985). A mission called "Lunar Prospector" is currently planned to look for water on the moon.

5.3 The Risks of Space Exploration: The Apollo "204" Fire and Apollo 13

Exploration programs such as Apollo are not without risks. This lesson was painfully learned once again when three astronauts, Edward White, Virgil Grissom, and Roger

Chaffee, lost their lives when the command Module in which they were sitting caught fire. The accident occurred on 27 January 1967 during a test with the entire Saturn V launch vehicle on the launch pad. The problem was caused by the fact that the Apollo Command Module was designed to operate with an atmosphere of pure oxygen at normal atmospheric pressure inside the Module. Under these circumstances, apparently a small spark somewhere in the electrical system ignited the hydrocarbon insulation on the wires, which burned vigorously in the oxygen atmosphere of the Apollo Command Module. In addition, the exit hatch of the module was designed in such a way that it could not be easily opened from the inside.

The Apollo-204 fire—it was called that because "204" was the designation number of that particular Command Module—presented NASA with a major crisis. The Accident Investigation Board suggested several design changes to the Command Module, and several important management changes were also made. NASA recovered from the effects of the accident, and the first successful launch of the Apollo system in an Earth orbital flight was Apollo 7 conducted on 11 October 1968. In spite of this setback, NASA was ultimately able to land astronauts on the Moon in time to meet President Kennedy's original schedule.

The second accident of the Apollo program occurred on 13 April 1970 on the Apollo 13 mission when an oxygen fuel tank in the Service Module failed. Although the astronauts had to abort the planned lunar landing, they were able to bring the crippled spacecraft back to Earth. In this case also, a NASA organized Accident Investigation Board quickly found the cause of the problem and recommended some appropriate changes, and the rest of the Apollo program was executed without further incident.

Both the Apollo-204 fire and the Apollo 13 accident illustrate the point that explorers must take great risks in order to accomplish their missions. There is no way that such risks can be avoided, but it is most important to provide an accurate assessment of the risks that must be taken and then to act according to the principle of "calculated risk." This is the time-honored way that explorers have operated, and space exploration is no different from previous experience.

5.4 The Soviet Lunar Exploration Program

It is important also to mention the lunar exploration program mounted by the Soviets in the 1960s. Although they later denied it, there is no question that the Soviets mounted a major program to put men on the Moon. They built a rocket approximately the same size as the Saturn V, called the N-1 booster (G-1), which is shown in Fig. 9. It failed on three launch attempts, the first on July 1969, the second some time in 1970, and the third in May or June of 1972. Since by that time a number of Americans had walked on the Moon, the Soviets canceled their lunar program.

The Soviets did mount a very successful robotic exploration program of the Moon. They were the first to put a "hard lander" on the moon, and they then followed with a series of soft landers that culminated with the automated roving vehicle "Lunakhod." Their most important achievement was that they developed an automated sample-return system that was able to collect and return about half a kilogram of lunar material during three missions (compared with a total of 375.4 kg for the Apollo astronauts). This was an important achievement for them, and it has very interesting future implications. At some point in the future there will probably be an effort to return samples from Mars using robotic techniques, and it is at this point that the automated sample return methods developed by the Soviets will pay off.

6. THE PLANETS

If the centerpiece of human space exploration was the Apollo program—the trip to the Moon—then planetary exploration was the focus for the development of the most sophisticated automated spacecraft. These complex and very capable robots would explore the Solar System without people on board. By 24 August 1989, when Voyager II flew past the planet Neptune, an American robotic spacecraft had flown past every planet in the solar system except Pluto. In addition, the Soviets had also contributed by carrying out a detailed study of the planet Venus with their "Venera" spacecraft and had also sent spacecraft to Mars with mixed

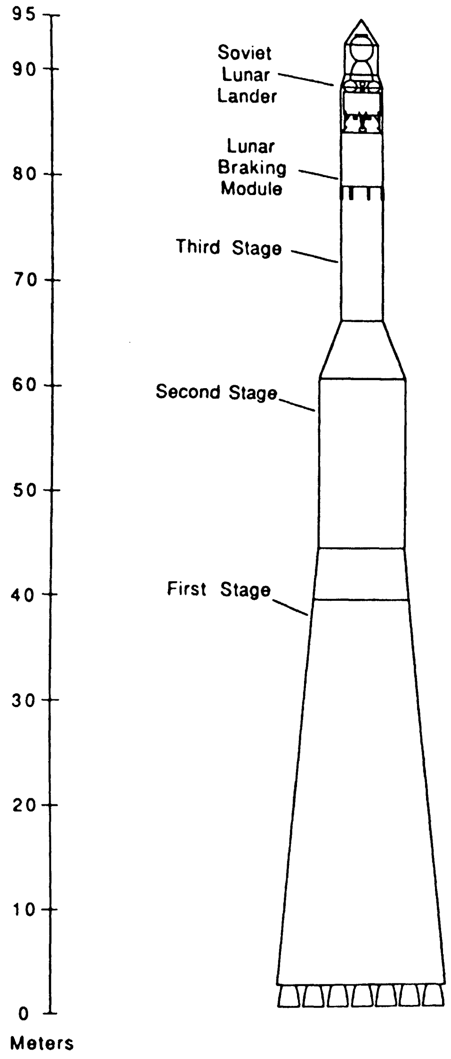

FIG. 9. The Soviet N-1 (G-le) booster that was built to take Soviet cosmonauts to the Moon. This launch vehicle failed three times when launches were attempted: the first exploded shortly after launch on 4 July 1969, the second a few months later failed during the first-stage burn, and the third on 24 November 1972 was destroyed by the range safety officer when the control system failed. (From Newkirk, 1990.)

success. There is no question that this exploratory effort marks a most important scientific achievement.

From the beginning of the planetary exploration effort in 1964, a very clear strategy was followed. The first step was to develop spacecraft that would execute flybys of every planet in the solar system. The second step was to put more capable spacecraft into orbit around each planet, which would make a more detailed study of the planet and, if possible, produce the data so that good maps could be made. The final step was to land automated spacecraft on the surface of the planet where that was feasible or to send a probe into the planet's atmosphere—or both. This very successful strategy has been executed by NASA for the past thirty years.

During the 1960s, the first generation of planetary spacecraft was conceived and developed. These were the "Mariner" class spacecraft that were designed to execute flybys of the inner planets, Mercury, Venus, and Mars. These were relatively small spacecraft, weighing approximately 450 kg, that were powered by solar electric cells and that had an accurate three-axis stabilization system. The use of solar power was feasible because the Mariners stayed relatively close to the Sun, and three-axis stable platforms were necessary in order to achieve the primary objective of the early flybys, which was to take good pictures. Altogether, seven Mariners were successfully launched, and they returned good pictures from the first close flybys of Mercury, Venus, and Mars. A picture of Mercury taken by Mariner 10 is shown in Plate 1. In addition to cameras, the Mariners also carried magnetometers to survey the magnetic fields of the inner planets and spectroscopic instruments to look at planetary atmospheres. It was discovered during these early flights that unlike the Earth, Mercury, Venus, and Mars have no substantial magnetic fields.

A somewhat different approach had to be developed to perform the first flybys of the outer planets. These spacecraft had to be even smaller than the Mariners because they had to go farther away from the Earth. In addition, it was not possible to use solar energy for their power supplies because by the time these spacecraft reached the orbit of Jupiter, the flux of energy from the Sun would be too weak. Nuclear energy in the form of the radioactive isotope plutonium-238 was used to provide electrical power for those spacecraft. These were called Pioneers, and two of them were flown (Pioneer 10 and 11) out to Jupiter and Saturn. Each Pioneer weighed 259 kg. They were spin stabilized, and they did not have an onboard data-storage system. The Pioneers also had cameras that were not as good as those on the Mariners but good enough to get better pictures of Jupiter and Saturn than any that could be obtained from Earth. Pioneer 10 was launched on 2 March 1972 and flew past the Planet Jupiter on 3 December 1973, becoming the first spacecraft to accomplish that objective. As explained in Sec. 3, Pioneer 10 flew past Jupiter in such a way that it picked up enough momentum during the encounter to leave the Solar System. Pioneer 10's sister ship, Pioneer 11, was launched on 5 April 1973 and also flew past Jupiter on 3 December 1974. In this encounter, Pioneer 11 was programmed into a trajectory that would take it to a close encounter with the planet Saturn on 1 September 1979. Pioneer 11 thus became the first spacecraft to fly past both Jupiter and Saturn.

The Pioneers were intended to be precursors for the more complex and sophisticated missions that were to follow. They were simple and robust spacecraft and they accumulated a remarkable number of "firsts" in executing their missions. In addition to being the first spacecraft to fly past Jupiter and Saturn, they were also the first to fly through the asteroid belt between Jupiter and Mars and to measure the micrometeoroid particle density in that region of space and to find that space vehicles could indeed survive passage through that region. And, as has already been noted, Pioneer 10 will become the first manmade object to leave the solar system. What is perhaps most remarkable about these small robots is that both of them are still in communication with the Earth more than two decades after they were launched. A picture of the Pioneer spacecraft is shown in Fig. 10.

After the first generation of the Mariner spacecraft completed their missions to fly past all of the inner planets, a more sophis-

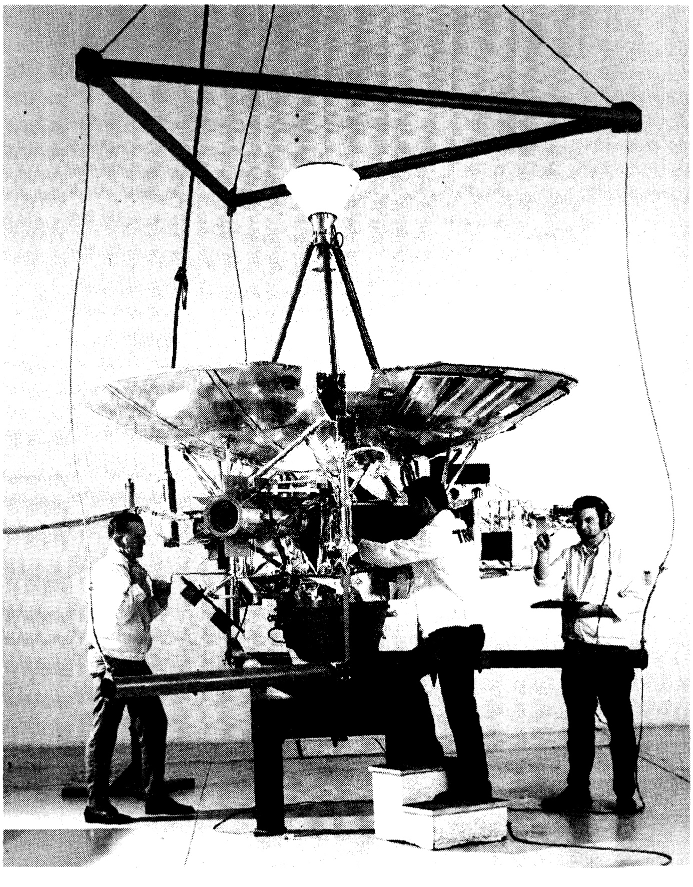

FIG. 10. The Pioneer spacecraft, the first to fly past both Jupiter and Saturn and the first manmade object to leave the solar system (courtesy NASA).

ticated model was built. This was Mariner IX, which was without question one of the most capable spacecraft built up to 1970. Mariner IX was designed as the first space vehicle ever to be put into orbit around another planet in the solar system. Mariner IX was considerably larger than the earlier Mariners, weighing in at over 910 kg. It also had the capability to execute complex maneuvers, which was necessary if it were to fulfill its mission to be placed in an orbit around the planet Mars. Mariner IX was launched on 30 May 1971, and it reached the vicinity of Mars on 13 November 1971. The spacecraft was successfully placed into orbit around the planet and then commenced its mission, which was to begin the creation of a detailed map of the planet. The mission of Mariner IX was an essential precursor to the Viking Mission that would place the first spacecraft on the surface of Mars.

What eventually became the Viking program began as a much larger effort to put a "soft" lander on the planet Mars using a Saturn launch vehicle. With the cancellation of the Apollo program, the program plan was scaled back to use a Titan-Centaur rocket to propel the spacecraft to Mars. The Viking spacecraft consisted of two parts: one that would be placed into orbit around Mars, and a second one that would land on the planet. The Viking lander was definitely the most sophisticated robotic spacecraft built when it was launched in 1975. The scientific objectives of the Viking project were both clear and important. The Viking orbiter was equipped with an improved camera, and it would continue the Mariner IX mission to make an accurate and comprehensive map of the planet. The carrier wave of the orbiter's communication system would be used to study the atmosphere of Mars when the orbiter passed behind the planet and the signal had to pass through the atmosphere immediately before occultation. The Viking lander would also take pictures on the surface of the planet, would analyze soil samples that would be picked up by an especially designed scoop, would look at the atmosphere using a small automated weather station, and would look for evidence of seismic activity. Finally, the single most important scientific objective of the Viking lander was to look for evidence of biological activity.

The Soviet Union also had an active program for the exploration of Mars. They attempted twice to put a "soft" lander on Mars and both attempts failed. The first one apparently missed the planet. The second spacecraft landed on Mars on 2 December 1971, and data transmission failed after 20 s of operation.

It was recognized during the 1960s that starting in the mid-1970s and running on for the next two decades, it would be possible to put a spacecraft on a trajectory such that it could fly past all of the outer planets from Jupiter to Pluto because of the rare alignment of the planets in their orbits. This was called the "Grand Tour," and, as in the case of the exploration of Mars, the early thinking was to put the Grand Tour space vehicles on the correct trajectories using Saturn-type launch vehicles. The most important technical problem that had to be solved to execute the Grand Tour missions was to achieve accurate navigation. The targets of the spacecraft would be very far away but, more important, the trajectory of the spacecraft not only had to hit the right orbit around the planet but also to go into the right trajectory around the planet that would carry the spacecraft to the next target planet. As in the case of Mars, a smaller program was developed called "Voyager" that would come very close to executing all of the objectives of the original Grand Tour.

The Voyager spacecraft was a large (816 kg), very sophisticated spacecraft that was more or less a follow-on to the outer-planet reconnaissance performed by Pioneers 10 and 11. As in the case of the Pioneers, the Voyager spacecraft was powered by three radioisotope thermal generators (RTG's) that used the radioactive decay of plutonium-238 as an energy source. Unlike the Pioneers, the Voyagers had three-axis stabilization systems and very good data-storage systems. They also had high–data-rate telemetry systems so that they could produce high-resolution pictures of the target planets. Two Voyager spacecraft were launched, Voyager I on 5 September 1977 and Voyager II on 20 August 1977. Voyager II is still in contact with the Earth-based ground stations seventeen years after launch. Voyagers I and II both flew past Jupiter and Saturn with encounter dates for Voyager I of 5 March 1979 (Jupiter) and 12 November 1980 (Saturn) and for

Voyager II on 9 July 1979 (Jupiter) and 25 August 1981 (Saturn). Subsequent to these encounters, Voyager II was put on a trajectory that took it to a flyby past Uranus (24 January 1986) and Neptune (24 August 1989). In addition to producing excellent high-resolution pictures of the outer planets, the Voyagers also sent back excellent images of most of the satellites of the outer planets. At the conclusion of the Voyager flybys, Pluto remained the only planet in the solar system that had not yet been visited by a spacecraft launched from Earth.

The third generation (following Pioneer 10 and 11 and Voyager I and II) of outer-planet exploration was initiated in the late 1970s with a program called the Jupiter-Orbiter-Probe (JOP). The first objective of this program was to place an orbiter around the planet Jupiter to take pictures and measure the properties of the space environment in the vicinity of Jupiter. By placing the spacecraft into an orbit around Jupiter, time-dependent and dynamic phenomena both on Jupiter and on Jupiter's satellites could also be observed. The second important objective was to put a probe on the spacecraft that would be dropped into the atmosphere of Jupiter to perform some important *in situ* measurements of the composition and structure of the outer layers of Jupiter's atmosphere. The design and construction of this probe presented some very formidable technical problems since the entry velocity of more than 64 400 km/h is far larger than any that had been encountered in the past. Furthermore, the fact that the outer atmosphere of Jupiter is rich in hydrogen also complicates things because it is more chemically reactive than some other species found in planetary atmospheres.

The spacecraft that is executing the JOP mission is now called "Galileo." It was launched by the Space Shuttle "Atlantis" on 18 October 1989. "Galileo" is a large sophisticated spacecraft in the "Voyager" class weighing 1100 kg with 3-axis stabilization and plutonium-238 powered RTG's. Because of NASA's budget problems in the 1980s, a decision was reached to build only one "Galileo" spacecraft in contrast to the previous planetary exploration program practice that called for two launchers with identical spacecraft. Unfortunately, the high–data-rate antenna failed to deploy properly after the launch. Thus, the absence of a back-up spacecraft caused a major problem. In spite of many attempts to fix the problem, by maneuvering the spacecraft, all of these failed. The absence of the high–data-rate antenna compromises the mission. Much of the high-resolution pictures showing the dynamic behavior of Jupiter and its satellites will be lost. However, the "Galileo" mission managers have been able to develop "work arounds" that will make the "Galileo" mission very valuable in spite of the problem. "Galileo" assumed its orbit around Jupiter on or about 7 December 1995.

The exploration of the planet Venus is unique in the sense that it is the only effort in which the Soviet Union developed a strongly competitive program in planetary exploration. In fact, for a number of years, the Soviet Union led the effort to explore Venus with their "Venera" spacecraft. The first successful "Venera" was Venera 3, and it was launched on 16 November 1965. Venera 3 impacted on the surface of Venus on 1 March 1966. It was the first manmade object to land on another planet. Venera 5 performed the first *in situ* measurements of the composition and structure of the Venus atmosphere. Venera 5 was followed by nine Veneras, and these missions obtained a good picture of the properties of the Venus atmosphere and the properties of the surface. The most spectacular Venera mission was Venera 10, which produced the first picture of the surface of the planet Venus.

When the exploration of Venus was first seriously considered by NASA in 1970, the primary question was whether it would be possible to develop a program that would make a unique contribution to the knowledge of Venus that had not already been done by the Soviet Venera flights. After much discussion, it was decided to focus on two things: The construction of a good map of the planet and the development of some means of measuring winds in the Venus atmosphere to get at least one datum point describing the dynamics of the atmosphere. The result of these considerations was the Pioneer Venus program. Pioneer Venus was a relative of the outer-planet Pioneers in that it was conceived as a small, simple, and rugged spacecraft that could be rapidly developed and launched. There would be two spacecraft; one would carry four probes that would be deployed into the Venus atmosphere. These probes would be followed throughout their descent, and their relative velocities would be determined by careful analysis of the frequency (Doppler) shifts observed in the communications system carrier waves. In addition, the probes would perform measurements of the composition and structure of the Venus atmosphere to follow on the measurements performed by the Venera probes. The Pioneer-Venus orbiter would be placed in a relatively high-inclination orbit around the planet and would be equipped with a radar altimeter. Since the emphemeris of the spacecraft could be accurately determined, a reasonably good map of the planet could be made. The ground resolved distance of this map was about 96 km.

A second Venus mapping effort was initiated in 1982 called the Venus Radar Mapper (VRM). This spacecraft was designed to carry a radar imaging system rather than a simple altimeter and would therefore be larger and more sophisticated than the Pioneer-Venus orbiter. The spacecraft was eventually named "Magellan," and the synthetic-aperture imaging radar eventually produced an excellent map of the planet with a ground resolved distance of about 1 km or about 100 times better than Pioneer Venus.

In the case of comets, a large international effort was mounted to execute a close flyby of Halley's comet during its close approach to the Sun in 1987. Three spacecraft, one European that was built and operated by the European Space Agency (ESA), one Soviet, and one Japanese, succeeded in gathering important information about Halley's comet. It was clearly established by these missions that the comet consisted mostly of water ice in addition to some rocky material as well. The United States elected not to send a spacecraft to Halley's comet. However, NASA provided navigational support for the Halley comet missions through the use of the Deep Space Network, which greatly improved accuracy in tracking and guiding the spacecraft.

A summary of the primary facts learned about each planet by the exploratory efforts that have been described is provided here.

6.1 Mercury

Mercury is a small planet that is a third larger than the Moon in linear dimensions (diameter 4827 km compared to 3475 km). The surface of the planet is hot with a surface temperature of about 500°C on the day side and 210°C on the night side. The surface is rough and cratered with a crater density that is substantially higher than that of the Moon (see Plate 1). Mercury has a substantially different history from that of the Moon and a substantially different composition. It has roughly the same density as the Earth, and, interestingly, it has a small magnetic field, revealed by Mariner 10.

6.2 Venus

Venus has long been regarded as the Earth's "twin" planet because it is in an orbit closest to that of the Earth, and it is about the same size. After that, however, all resemblance ends. Venus has a dense atmosphere that consists mostly of carbon dioxide. The pressure at the surface of the planet is about 95 atm. The temperature at the surface is 550°C, which is hot enough to melt lead. The picture of the surface of Venus taken by Venera 13 is shown in Fig. 11. Even though the atmosphere is very dense, and there is a substantial cloud layer composed of colloidal-size particles, enough sunlight penetrates to the ground to create a permanent kind of "twilight."

The map of Venus obtained from the radar altimeter system on the Pioneer Venus spacecraft is shown in Fig. 12, and a similar map of what the Earth would look like observed by the same system is also shown. In terms of surface features, Venus and the Earth do have some resemblance. Both have elevated regions (i.e., continents on Earth), and both have low-lying regions, which on Earth are ocean bottoms. Both planets have "rift valleys" that are characteristic of volcanic activity. The best evidence of volcanic activity on Venus is the presence of sulfur dioxide and sulfuric acid in the Venus atmosphere. Thus, while the atmospheres and surfaces of Venus and the Earth are very different, it is likely that the internal structures of Venus and the Earth are quite similar.

6.3 Mars

More is now known about Mars than any of the other planets except the Earth itself. Since Mars has a transparent atmosphere, it

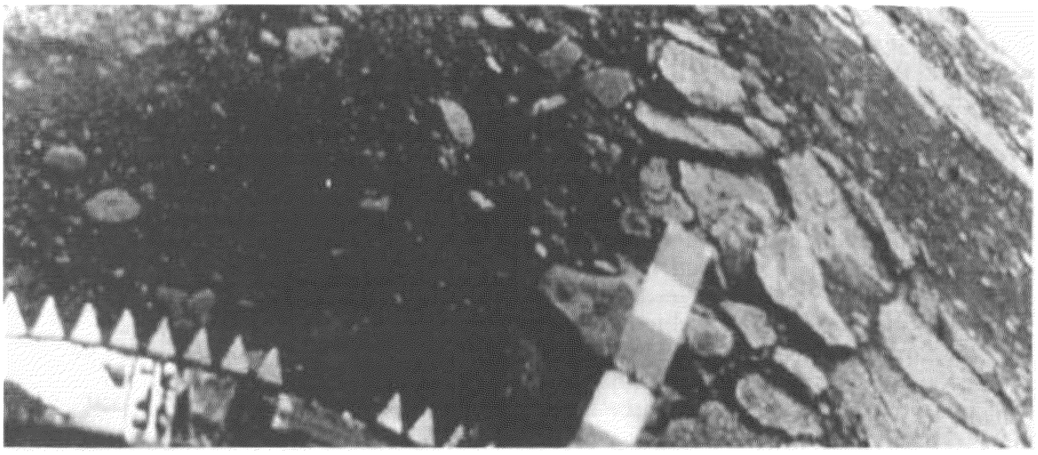

FIG. 11. Picture of the surface of Venus taken with a camera mounted on Venera 13 (courtesy NASA).

is possible to take good pictures both from the Earth and from orbiting spacecraft. It was, in fact, the ability to observe Mars easily through Earth-based telescopes that created the high degree of interest in the planet. By sheer coincidence, Mars has about the same period of rotation as the Earth, and its axis makes about the same angle with the ecliptic plane. Therefore, Mars has about the same diurnal and seasonal variations as the Earth. It was, of course, this circumstance that led to the intense speculations in the nineteenth century about the existence of life—even intelligent life—on Mars.

The first pictures sent back by Mariner IV in 1965 must have been very disappointing to those people who still harbored these thoughts. The pictures showed a barren, cratered surface that was clearly not a hospitable place for life. The results from Mariner IX were more interesting since these indicated that for all its desertlike appearance, Mars was an active place. When Mariner IX arrived in orbit around Mars on 13 November 1971, the surface observed was almost completely featureless. At first, there was fear that, somehow, the camera systems had failed. Actually, this was not the case. Mariner IX arrived at Mars during one of the periodic planet-wide dust storms that completely obscure the planet's surface. Such evidence of storms had already been observed from Earth-based observations. The storm soon ended, and Mariner IX became the most successful mission to Mars that had so far been executed. The Mariner IX pictures began the comprehensive effort to make an accurate map of the planet.

Mariner IX also was the first to take a good picture of Phobos, the larger of the two satellites of Mars. Phobos is an irregular-shaped object about 21 km across; probably it is actually an asteroid captured by Mars at some point in the past.

Two Viking spacecraft were sent to Mars. Viking I was launched on 20 August 1975, went into orbit around Mars on 19 June 1976, and put the Viking I lander on the surface on 20 July 1976. Viking II was launched on 9 September 1976 and orbited Mars on 7 August 1976, and the Viking II lander arrived on the surface of the planet on 3 September 1976. The Viking missions yielded a veritable treasure trove of scientific results. The Viking orbiters completed the map of Mars that is shown in Fig. 13. This map shows the great rift valley on Mars near the equator of the planet (Valles Marineris), and in the upper left of the map, the largest mountain in the solar system, the volcano Olympus Mons, is clearly visible. A polar projection of the planet is shown in Fig. 14 that clearly shows the polar cap. The composition of the polar cap is a mixture of water and carbon dioxide ice. A picture taken by the Viking lander standing on the surface of the planet is shown in Fig. 15. This was the first time that humans have ever had a close look at the surface of another planet from a vehicle standing on that surface. The

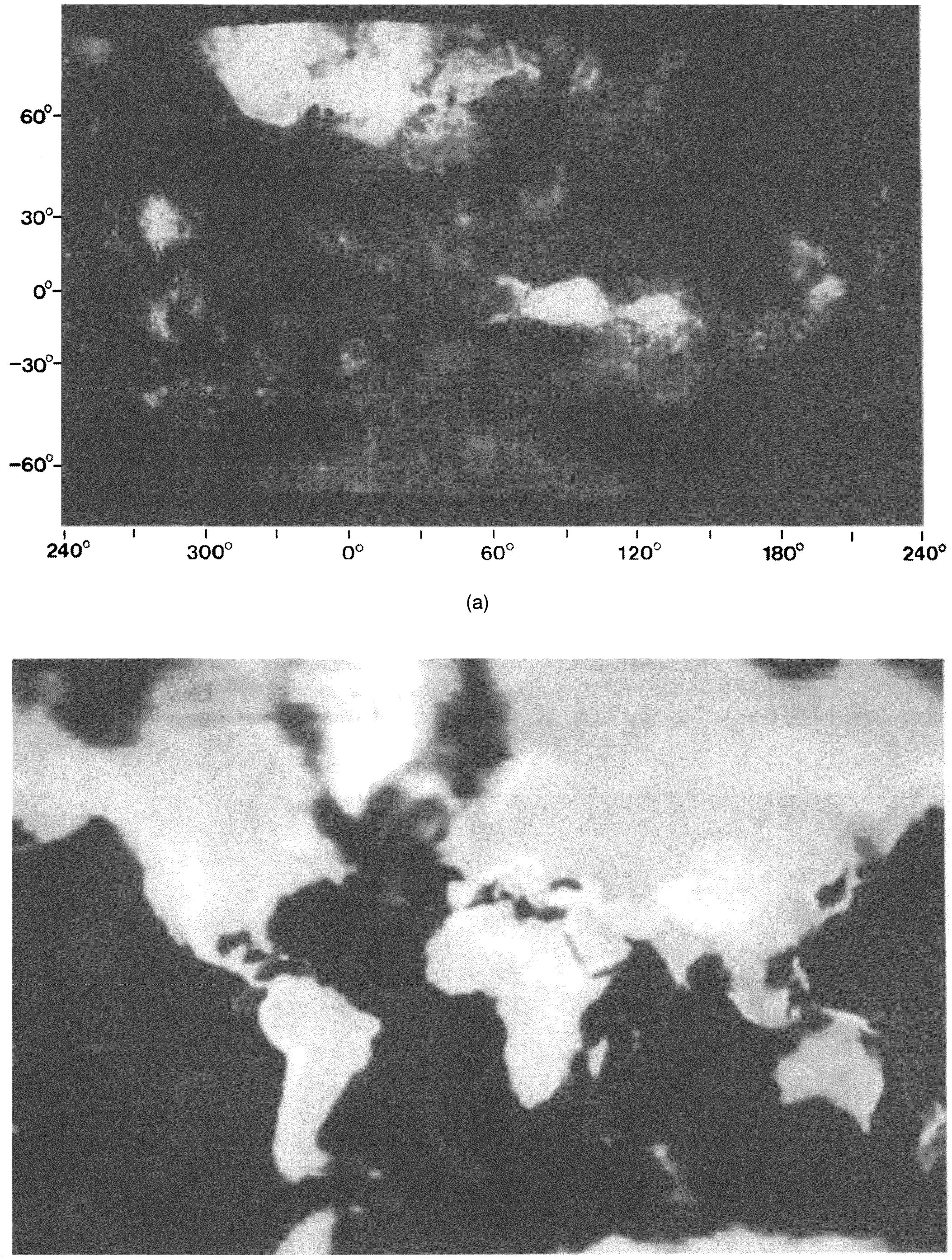
60°
30°
0°
-30°
-60°
240°
300°
0°
60°
120°
180°
240°

(a)

(b)

FIG. 12. (a) Mercator projection of a map of the planet Venus made by the Pioneer Venus Orbiter, along with (b) a map of what the Earth would look like if pictures were taken with a similar instrument (from Fimmel *et al.*, 1983).

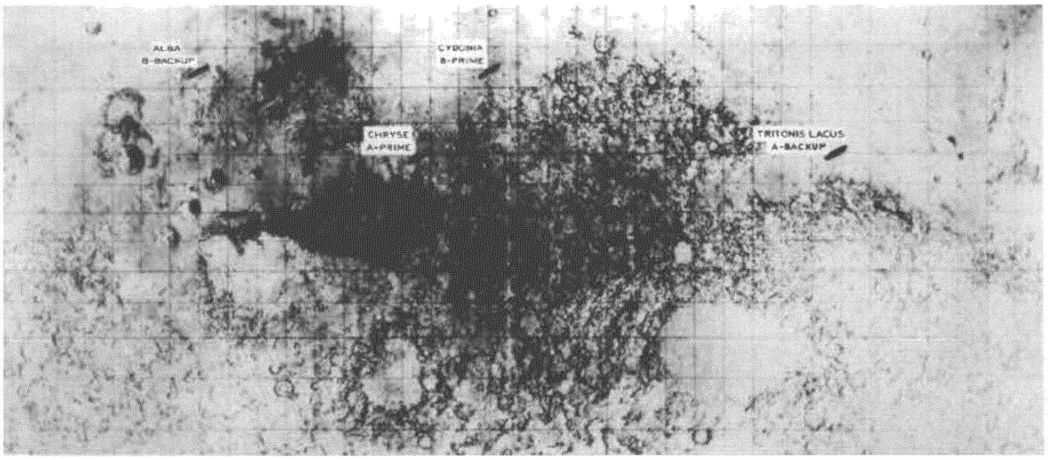

FIG. 13. Mercator projection of a map of the planet Mars. It was produced from pictures taken by the Viking Orbiters. The prime and back-up landing sites of the two Viking landers are also shown. (Courtesy NASA)

lander itself is shown in Fig. 16. There is no doubt that all of these images had a profound effect on the public at the time.

No evidence of biological activity was found by the Viking landers. In fact, the surface chemistry of Mars as it exists today was found to be extremely inhospitable to the carbon-based life that exists on Earth. However, the Viking landers did discover some tantalizing evidence that conditions existing on Mars perhaps some 3×10^9 years ago were similar to those on Earth. It is very likely that life on Earth originated more than 3×10^9 years ago because fossil remains of unicellular organisms have been found in rocks of that age on Earth (Barghoorn and

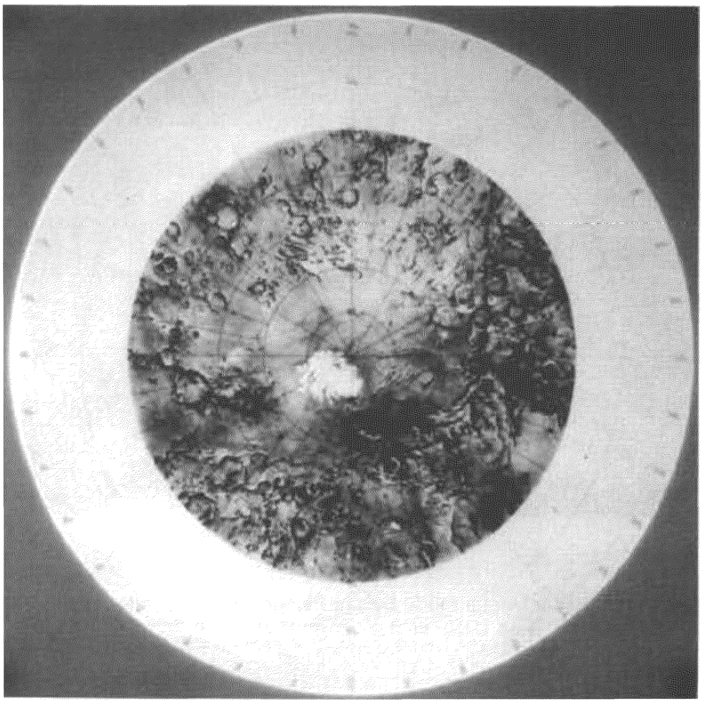

FIG. 14. Polar projection of Mars looking down on the north pole. This map was also made from pictures taken by the Viking Orbiters. (Courtesy NASA)

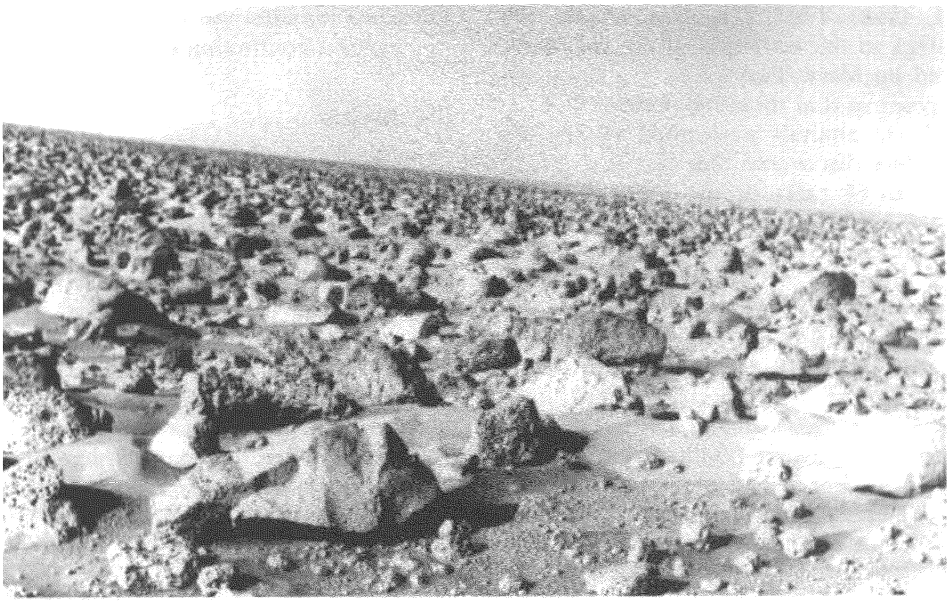

FIG. 15. Picture of the Martian surface as taken by the Viking II lander. This view looks out over Utopia Plain on the northern hemisphere of Mars. It is probably a typical example of Martian flatlands. (Courtesy NASA)

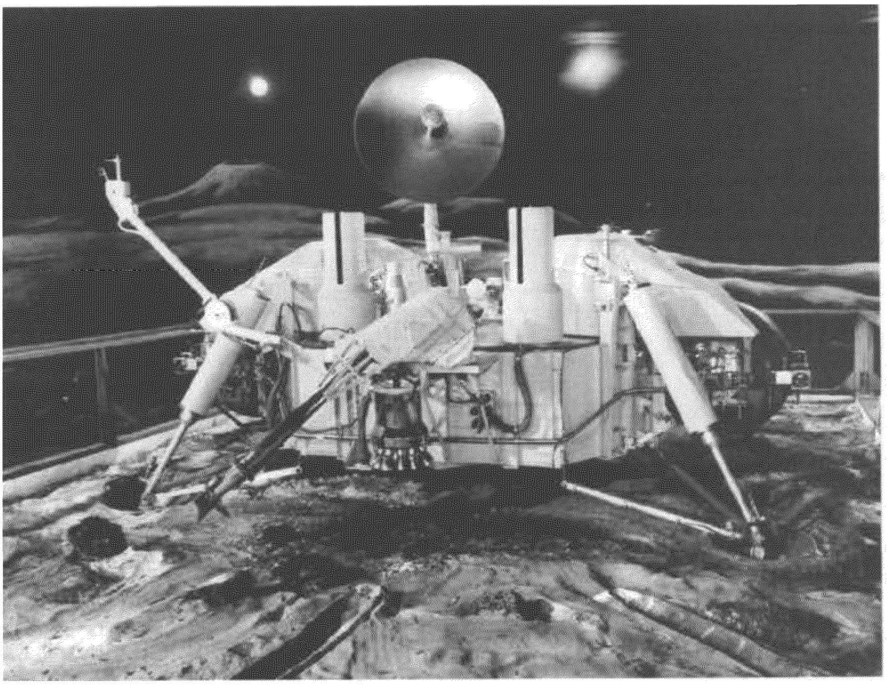

FIG. 16. The full-scale engineering model of the Viking Lander (courtesy NASA).

Schopf, 1966). Thus it is possible that the early steps in the evolution of life may have occurred on Mars. Two major pieces of evidence point in that direction. One is that the atmospheric analysis performed by the Viking landers discovered that the nitrogen-15 to nitrogen-14 ratio in the atmosphere of Mars was twice as large as that in the Earth's atmosphere. By using the differential rates of diffusion of these isotopes of nitrogen out of the Mars atmosphere, the conclusion can be drawn that 3×10^9 years ago, the atmosphere of Mars in terms of pressure, temperature, and composition may have been similar to that existing on Earth at the same time. This means that liquid water may have been present at that time, which is confirmed by many of the pictures taken both by Mariner IX and the Viking orbiters. Some examples of dry lake beds and rivers are shown in Fig. 17. Since conditions on Mars may have been similar to those on Earth at that time, the open and important question is whether what happened here on Earth also happened on Mars. There is no doubt that the origin of life remains one of the major unanswered scientific questions, and this therefore remains the most important motivation for a continuing exploration of Mars.

6.4 Jupiter

Jupiter's properties as determined by the four flybys (Pioneer 10 and 11, Voyager I and II) are now also known in much greater detail than they were before the initiation of planetary exploration. Like Mars, the "surface" features of Jupiter can be seen with good telescopes from the Earth. Among the most interesting and puzzling of these is the "Great Red Spot," which can easily be seen from the Earth. It was observed as early as 1631 using a telescope and it has been visible on the planet ever since. It is now known that the Great Red Spot on Jupiter is a permanent vortex in Jupiter's atmosphere caused by hydrodynamic effects that result from the differential velocities of various regions of Jupiter's surface. This effect also accounts for the familiar band structure that characterizes the planet's surface, which can easily be seen from the Earth. Jupiter rotates around its axis once every ten hours so that enormous shear forces are created between

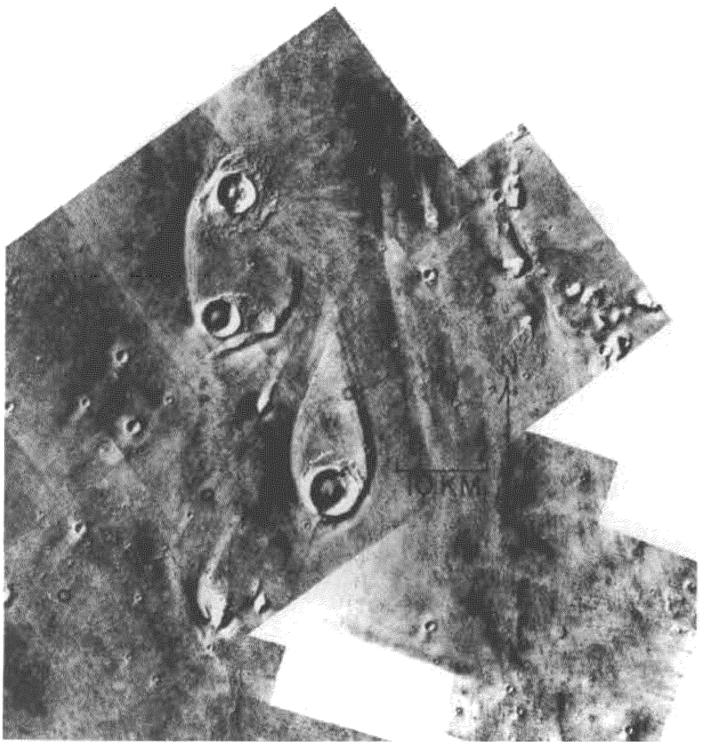

FIG. 17. Some ancient river beds on the planet Mars indicating that water was probably present during the first billion years of the planet's existence. This composite photograph taken by Viking Orbiters shows the apparent outflow from the Ares Vallis into the southern Chryse Planitia. (From Baker, 1982.)

different regions that rotate at different velocities. From the detailed pictures taken by the Voyager Spacecraft, we know that the dark boundaries between the bright bands are turbulent regions.

The chemical composition of the outer layer of Jupiter's atmosphere is mostly hydrogen and helium. There are other species present including ammonia, methane, and some more complex molecules that account for the colors observed in Jupiter's upper atmosphere. There is some mixing caused by the turbulent motion of the gases in the upper atmosphere. The Great Red Spot is the largest vortex of this kind that mixes layers of the atmosphere with those in the upper regions. Complex organic molecules have been observed in the upper atmosphere of Jupiter by both the Pioneer and the Voyager spacecraft.

The average density of Jupiter is 1.33 g/cm^3, which is close to that of normal water. Thus most of the planet is gaseous. Deeper in the planet's atmosphere there may be a layer of liquid water. Below that, there may be a layer of compressed hydrogen that behaves something like a metal with respect to its electrical properties. Finally, there is probably a small (relative to Jupiter's overall size) iron core. The structure is shown in Fig. 18. Jupiter is the largest planet in the solar system with a diameter 11.2 times that of Earth and a mass 318 times as large.

Jupiter has the strongest magnetic field of all the planets. This is not surprising given the planet's rapid rate of rotation and the postulated internal structure of the planet. The same mechanism that causes the Earth's magnetic field (a homopolar generator) probably also causes the magnetic field of Jupiter. Energetic charged particles are trapped in Jupiter's magnetic field just as they are in the magnetic field of the Earth. Very probably the same mechanisms that operate in the Earth's magnetosphere with respect to solar-wind and cosmic-ray interactions are also present on Jupiter. One of the functions of the Galileo orbiter of Jupiter will be to obtain an accurate map of Jupiter's magnetic

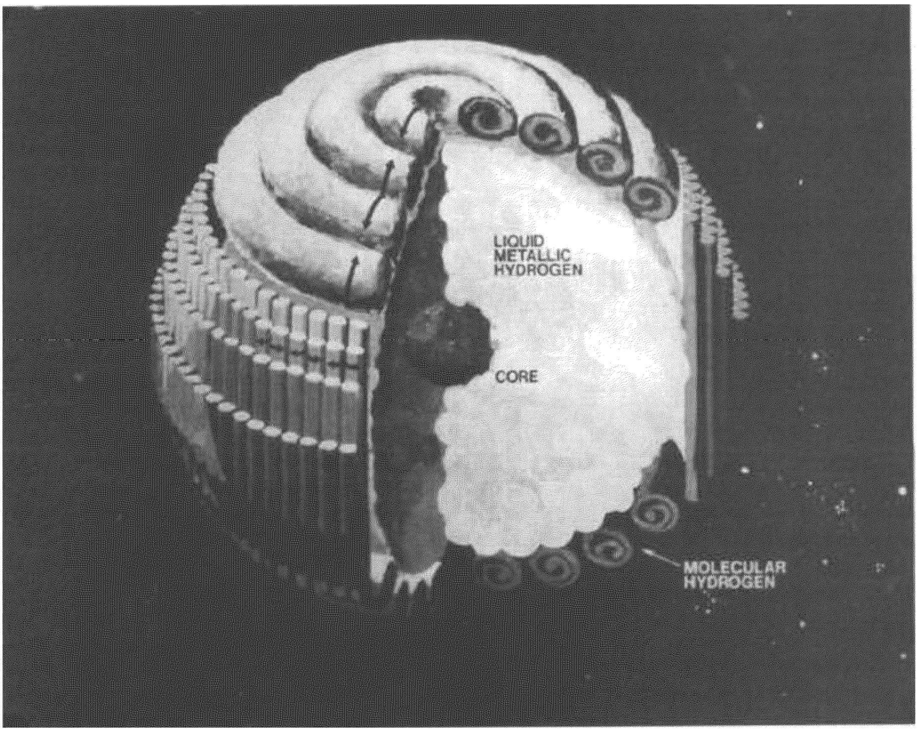

FIG. 18. A model of the structure of Jupiter as deduced from the Pioneer and Voyager flights and other relevant material (courtesy NASA).

field and of the charged particles trapped in the field.

During the Voyager flybys of Jupiter a ring structure was discovered because the telemetry signal was briefly modified as the spacecraft executed its orbit and passed the planet in such a way that the rings were between the spacecraft and the receiving station on Earth. This discovery by such an occultation was one of the unexpected results of the Voyager flybys. The rings are very tenuous when compared to massive ring structures of Saturn.

The most important results other than those from Jupiter itself came from the first detailed examination of the four large satellites of Jupiter—the Galilean moons. It was already known that two inner satellites Io and Europa have densities close to 3 g/cm^3, which characterizes them as more or less rocky (the Earth's density is 5.5 g/cm^3 and Jupiter's is 1.3 g/cm^3). The third moon, Ganymede, has a density is 2.2 g/cm^3 and the fourth, Callisto, has a density very close to that of Jupiter or about 1.3 g/cm^3. The four large moons of Jupiter therefore have very different properties. Io is close enough to the planet so that it experiences considerable tidal forces. The distortions and crustal motions caused by these forces dissipate enough energy to heat Io to the point where volcanic activity has been observed. Figure 19 shows a picture of such an eruption, the first active extraterrestrial volcano ever observed.

At the other extreme, Callisto is very probably a large snowball interspersed with some rocks. Large ice fields with long fault lines running through them have been observed. The Voyager flybys also revealed 12 smaller moons orbiting the planet. This count may increase as more detailed examinations of Jupiter's surroundings are made once Galileo is placed in orbit around the planet.

6.5 Saturn

The most prominent feature of Saturn is the large and complex ring system that surrounds the planet. The three flybys of the

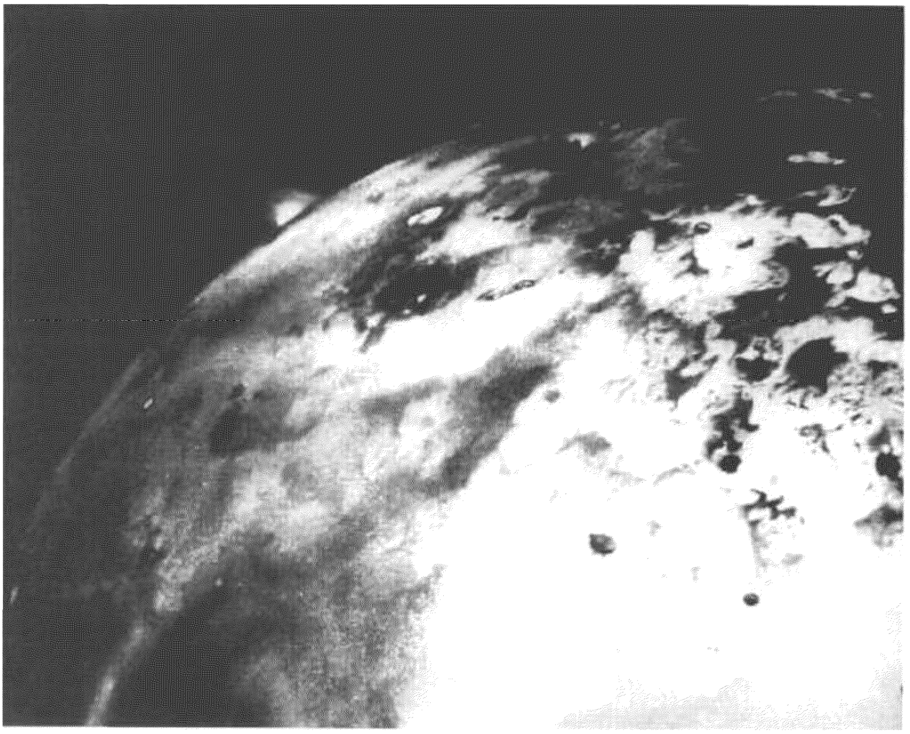

FIG. 19. A volcanic eruption on the Jovian satellite Io. The plume extends 100 miles above the surface of the satellite. This picture was made by Voyager I. (Courtesy NASA.)

planet by Pioneer 11 and Voyagers I and II reveal a surprisingly complex and beautiful structure. Several hundred separate rings were observed separated from each other by sharply defined dark spaces. The rings themselves consist of small bodies ranging probably from objects of the order of a meter in diameter up to some hundreds of meters. These bodies are in very nearly circular orbits around the planet and, in the aggregate, make up the ring system. The bodies in orbit are composed mostly of water ice. The mechanisms that cause the ring structure are well understood. The dark spaces between the rings are regions in which there are no stable orbits for the small particles that constitute the rings because of the influence of larger bodies orbiting Saturn that are not part of the ring system. The largest of these dark spaces, the Cassini Division, is caused by the largest of Saturn's satellites, Titan.

Although none of the pictures from the Voyager missions were actually good enough to resolve individual particles in the rings themselves, both the size distribution and also the composition of the rings can be reliably inferred from nonphotographic information that is available. A photograph of the ring system taken by Voyager I is shown in Fig. 20, which illustrates the great complexity.

Saturn is the second largest planet in the solar system, having a diameter 9.4 times that of the Earth and a mass 95.2 times as large. Saturn has the lowest density of all the planets in the solar system at about 0.7 g/cm^3. It is less dense than water and is about half as dense as Jupiter. It has a relatively rapid rotation rate of one revolution every 10 h. There are surface markings on the planet similar to those on Jupiter, but they are much less prominent. Probably this reflects the less complex chemistry of Saturn compared to that of Jupiter. Saturn has no magnetic field, which may mean that the interior of the planet cannot sustain a homopolar generator such as those that exist on Jupiter and on Earth or that the planet's generator is in a temporary state so that the magnetic field is small. Field reversals of this kind are known to occur on Earth, and the field is small when transitioning between extremes.

Saturn has more satellites known than

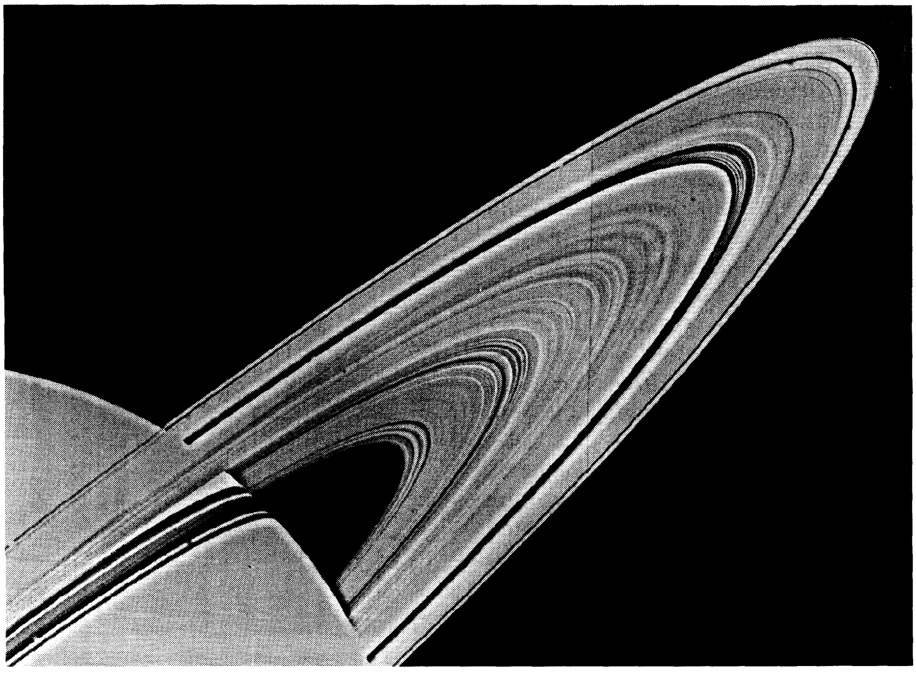

FIG. 20. Picture of Saturn's ring system that illustrates its great complexity. It was taken by Voyager I. (Courtesy NASA.)

any other planet. The number discovered so far is 18. The largest of Saturn's satellites, Titan, is actually the largest satellite in the Solar System. Its diameter is 5139 km, which is roughly the same as that of Mercury. Titan has an atmosphere (mostly of methane) the existence of which was determined by occultation of Voyager II's telemetry signal. Another one of Saturn's satellites, Hyperion, is the only object in the solar system that is definitely known to behave chaotically on a short enough timescale to be directly observable. This Moon tumbles, and measurements of the time history of the tumbling show typical chaotic behavior. This observation may eventually have an important relationship to the ultimate stability of the solar system.

6.6 Uranus

Voyager II, after flying past Saturn, was targeted to fly past Uranus. Uranus is another of the large low-density planets that are the dominant features of the outer solar system. Uranus was discovered by Sir William Herschel in 1781 using one of the first large telescopes built in the eighteenth century. The planet is about 20 times as far away from the Sun and about 14.5 times as massive as the Earth (see Table 2). It was the first planet to be discovered beyond the original ones (Mercury, Venus, Mars, Jupiter, and Saturn) known to the ancients. It is unique in that the axis of rotation of the planet lies approximately in the ecliptic plane rather than more or less perpendicular to it as is the case with the other planets. What is even more interesting is that five Moons that surround Uranus also rotate around an axis that is approximately in the ecliptic plane. The planet looks blue because the atmosphere has less hydrogen and helium than Jupiter or Saturn, which are both a brilliant yellowish in color. The blue color in the case of Uranus is caused by methane. Deeper inside the atmosphere, there are probably layers of ammonia and water. Probably, Uranus has no solid core. An Earth-based telescope mounted on a high-flying airplane first discovered evidence of a ring system on Uranus in 1978. This discovery was confirmed by Voyager II. As in the case of Jupiter, this ring system is very much less well developed than that of Saturn.

Table 2. Summary of the properties of the planets in the solar system (from NASA, 1993).

	Mean distance from Sun		Period of revolution (Earth time)	Period of rotation (Earth time)	Inclination of axis (deg)	Inclination of orbit to ecliptic (deg)	Eccentricity (deg)	Equatorial diameter		Atmosphere	Satellites	Rings
	(10^6 km)	(10^6 mi)						(km)	(mi)			
Mercury	57.9	36	87.97 d	58.65 d	0	7.0	0.206	4 878	3 031	Essentially none	None	None
Venus	108.2	67.2	224.7 d	243.01 d, retrograde	177.3	3.39	0.007	12 104	7 521	Carbon dioxide	None	None
Earth	149.6	93	365.26 d	23 h 56 min	23.5	0	0.017	12 755	7 926	Nitrogen, oxygen	1	None
Mars	227.9	141.6	686.98 d	24 h 37 min	25.2	1.85	0.093	6 790	4 219	Carbon dioxide	2	None
Jupiter	778.3	483.6	1.86 yr	9 h 56 min	3.08	1.31	0.048	142 796	88 729	Hydrogen, helium	16	1
Saturn	1429	888.2	29.46 yr	10 h 40 min	26.7	2.49	0.056	120 660	74 975	Hydrogen, helium	18	Thousands
Uranus	2875	1786	84.07 yr	17 h 14 min	97.9	0.77	0.046	51 118	31 763	Hydrogen, helium	15	11
Neptune	4504	2799	164.82 yr	16 h 6 min	29.6	1.77	0.010	49 528	30 775	Hydrogen, helium	8	5
Pluto	5900	3666	248.6 yr	6.39 d, retrograde	122	17.15	0.248	~2 300	~1 429	Methane	1	Prob. none

6.7 Neptune

This planet is unique because it was the first one whose existence was predicted before it was actually observed. When Uranus was discovered, it was observed that the orbital parameters of the planet could only be explained if there existed another large planet in the solar system outside the orbit of Uranus. The location of this new planet was calculated both by the English mathematician John Adams and the Frenchman U. J. J. LeVerrier. The German astronomer Johann Galle discovered the new planet on 23 September 1846, essentially just at the place predicted. Neptune is very similar to Uranus, just a little bit larger (about 17 Earth masses), and it is about 30 times as far away from the Sun as the Earth. The planet also has a bluish color because of the relative absence of hydrogen and helium compared to Jupiter and Saturn. As in the case of Uranus, methane is the principal component of the upper atmosphere of Neptune. Neptune emits about two and a half times the energy that it receives from the Sun, indicating that the planet is probably still contracting. Voyager II discovered six new satellites of Neptune. The outermost one, Nereid, executes the most eccentric orbit of any satellite in the solar system. Neptune's axis of rotation is tilted about 29° to the ecliptic plane, which is close to the axial tilt of the Earth. Neptune's largest satellite, Triton, is the only satellite in the solar system that rotates around its parent planet in a retrograde direction. It is probable that Triton was captured by Neptune after it was formed somewhere in the outer regions of the solar system. Triton has a tenuous nitrogen atmosphere and has the coldest temperatures of any object observed so far in the solar system. There is a very large polar cap which is apparently dynamic because Voyager II observed liquid nitrogen geysers in that region of Triton.

6.8 Pluto

There are plans to develop a small spacecraft to fly past Pluto in the next decade. No program has yet been approved.

6.9 Comets and Asteroids

In addition to the planets and their satellites, there are other objects of interest in the solar system, the most important of these being the asteroids and comets. No special missions have been mounted by NASA to look at an asteroid in detail. However, the "Galileo" spacecraft passed close enough to a large asteroid to obtain some good pictures. In 1994, a small spacecraft, "Clementine," sponsored by the Department of Defense was flown, first to orbit the Moon for two months and then to execute a close flyby of the asteroid Geographos. Unfortunately, the spacecraft failed before the second portion of the mission was accomplished.

In the case of comets, the 1986–1987 return of Halley's comet to the vicinity of the Sun provided a golden opportunity to take a good look at a large comet. Three spacecraft, one built by the European Space Agency (called "Giotto"), one by the Russians, and one by the Japanese, were launched to fly close to Halley's comet. The most important of these was the "Giotto" spacecraft. Giotto was a large spacecraft—over 900 kg—that was launched by an Ariane IV rocket in July 1985. The spacecraft flew within about 483 km of the nucleus of Halley's comet in March 1986. The spacecraft had a good camera and took some spectacular pictures. There were also sensors onboard that were able to determine the composition of the comet. It was discovered that Halley's comet was not a spherically symmetric object but rather an oblong that had almost a dumbbell shape. Also, the composition was determined to be mostly water ice with some heavy material mixed in. As one observer put it, Halley's comet was essentially a "muddy snowball." The Russian and the Japanese spacecraft also provided important confirming information on the behavior and the structure of Halley's comet.

7. ASTRONOMY FROM SPACE

The topic of this article is "exploration," which, by a simple definition, means going somewhere. However, in the case of space exploration, it is often so profitable to make remote observations from space that this activity as well falls within the definition of the word "exploration." Doing astronomy from space has several advantages. First, there is no atmospheric distortion as the light passes from the object observed to the collection mirror of the telescope. This essentially allows the construction in Earth orbit of large, diffraction-limited telescopes, giving superior angular resolution compared to any that can be built on the ground. Second, there is no atmospheric absorption. This means that it is possible to build telescopes that can be used at all wavelengths of interest, ranging from the far infrared to the ultraviolet. Detectors sensitive to a wide range of wavelengths (i.e., quantum energies) can be put at the focal plane of orbiting telescopes. Finally, in the case of infrared telescopes, the optics and the detectors can be cooled by liquid helium to very low temperatures for long periods of time. Thus, it is possible to see objects that are several orders of magnitude less bright than those that can be seen from the ground. Specialized instruments to look for x rays and gamma rays have also been put aboard Earth-orbiting satellites. In the case of x rays, high-resolution diffraction spectrometers have been flown. Therefore, space-based telescopes can be used to do what might be called "broadband" astronomy for the first time.

The first observations of astronomical objects were made with suborbital sounding rockets. The first target was the Sun, and several investigators developed high-resolution spectrometers for observing radiations emitted by the sun. Solar x rays and ultraviolet emissions were well understood using suborbital techniques. In 1962, the research group led by Ricardo Giacconi discovered the first stellar x-ray object in the constellation Scorpio, which is now called SCO X-1. Several groups organized sounding-rocket experiments to investigate this discovery, and a number of interesting observations were made (Giacconi *et al.*, 1962, 1963; Bowyer *et al.*, 1964, 1965; Chodil *et al.*, 1965). One of these was the discovery that an x-ray source in the constellation Cygnus (Cygnus X-1) seemed to be variable (Bowyer *et al.*, 1965, 1966). It was this source that was later on identified by satellite-born x-ray detectors to almost certainly be the result of a "black hole."

7.1 Orbiting Astronomical Observatories

The first spacecraft designed to look at stars was launched on 8 April 1966. It was called the Orbiting Astronomical Observatory (OAO-1), and unfortunately, it failed. The second Orbiting Astronomical Observatory (OAO-2) was launched in December 1968. It carried 11 different telescopes or radiation detectors. This was really the first example of a "broadband astronomical observatory." It carried infrared, ultraviolet, x-ray, and gamma-ray detectors and clearly demonstrated the power of doing astronomy from space. The third launch of an orbiting astronomical observatory was a failure, but the fourth, which was named "Copernicus," was one of the most successful scientific missions ever attempted. The "Copernicus" spacecraft was launched on 21 August 1972, and it is still working at this writing. The most important discovery of "Copernicus" was the more certain identification of the x-ray source in the constellation Cygnus (Cygnus X-1) as a "black hole." The variability of the Cygnus x-ray source was explained by the occultation of the black hole by the companion star that orbits around the black hole. The x rays emitted by this object are caused by matter that is sucked from the companion star into the black hole by the very intense gravitational field of that object. As the matter accelerates, it gets hotter, and, therefore, energetic x rays are emitted. In addition to the investigations of the Cygnus black hole, the "Copernicus" spacecraft also made a detailed investigation of the Crab Nebula and other objects in the sky that are thought to be remnants of supernova explosions. The structure of these objects was clearly identified by "Copernicus." There is either a neutron star or a black hole at the center, which is surrounded by an expanding gas cloud caused by the energy developed in the supernova explosion. Several of these sources have been identified (Tucker, 1984).

7.2 Specialized Observatories

The Orbiting Astronomical Observatories were followed by a series of more specialized satellites. Some of these were designed to look at specific objects. One of these was the so-called Solar Maximum Mission (SMM), which was launched on 14 February 1980.

This spacecraft was designed as a follow-on to the Orbiting Solar Observatories and was launched in order to observe the Sun during the solar maximum that was expected early on in the 1980s. The SMM spacecraft failed on orbit a few months after launch. It was determined from an analysis that the problem could be repaired on orbit if the satellite were visited by astronauts from a space shuttle mission. This mission was carried out on 10 April 1984 by astronauts flying on the space shuttle "Challenger." The SMM satellite, popularly known as "Solar Max," thus became the first spacecraft to be repaired on orbit by astronauts from the space shuttle. The satellite worked satisfactorily for several years after the repair mission was completed. This operation was a major demonstration of the capabilities of the space shuttle.

Two other specialized astronomical missions deserve mention here. In this case the specialization was with respect to regions of the electromagnetic spectrum that were to be investigated rather than the objects at which the satellites were pointed. The first of these was the International Ultraviolet Explorer (IUE), which was a joint U.S./British mission. The spacecraft was built in the United States, and the ground tracking was done by the British. The IUE was a relatively large spacecraft, weighing about 680 kg and carrying a 44.5-cm-diameter telescope. The IUE spacecraft obtained the first high-resolution spectra of extragalactic stars, and it was the first to do a comprehensive survey of all the ultraviolet sources in the sky. Perhaps the most interesting result from the IUE mission was that the first measurement of the light curve of a supernova was performed in the ultraviolet regions of the spectrum.

A second "specialty" mission of this kind was the Infrared Astronomical Satellite (IRAS). This satellite was designed to look at the infrared region of the spectrum. It carried a telescope with a mirror 57 cm in diameter made of beryllium so that it could be cooled to cryogenic temperatures. The spacecraft had a liquid cryostat that cooled the mirror, as well as the detector array, to a temperature of 2 K. The cryostat carried 473 L of liquid helium. The Infrared Astronomical Satellite was launched on 25 January 1983 and lasted for 1 yr in orbit. In that time the IRAS spacecraft identified thousands of infrared sources in the sky. Probably the most important result was the work done on stellar evolution. The IRAS spacecraft looked at dense gas clouds in which stars were forming, particularly in the constellation Orion. Much was learned about the beginning of the life of a star from this extremely successful mission. Like the International Ultraviolet Explorer, the IRAS was also an international mission. The telescope was built in the United States, the spacecraft was built in the Netherlands, and once again, the British developed the tracking system for the spacecraft. Therefore, in addition to the scientific results of the mission, much was learned about the management of complex international space programs. A picture of the IRAS spacecraft is shown in Fig. 21.

7.3 High-Energy Astronomical Observatories

A third group of "specialty" missions were called the High-Energy Astronomical Observatories (HEAO). The objective of these spacecraft was to put an observatory in Earth orbit that would focus on x rays, gamma rays, and high-energy cosmic rays. Three HEAO observatories were launched. The most important of these was HEAO-2, which was a large spacecraft, weighing about 3275 kg, and was designed to carry a unique imaging x-ray telescope. This was a large grazing-incidence instrument, which served as an x-ray spectrometer and permitted the spacecraft to make significant contributions to cosmology. Probably the most important result of this mission was the identification of an x-ray emitting quasar more than 10×10^9 light years away. The HEAO-2 imaging instrument obtained the beautiful x-ray image of the Crab Nebula shown in Fig. 22. The HEAO-2 spacecraft was launched in November 1978 and was named "Einstein" after it achieved Earth orbit.

7.4 The Great Observatories

The next generation of astronomical observatories are now called the "Great Observatories." These spacecraft were also designed to concentrate on different regions of the electromagnetic spectrum. The Great Ob-

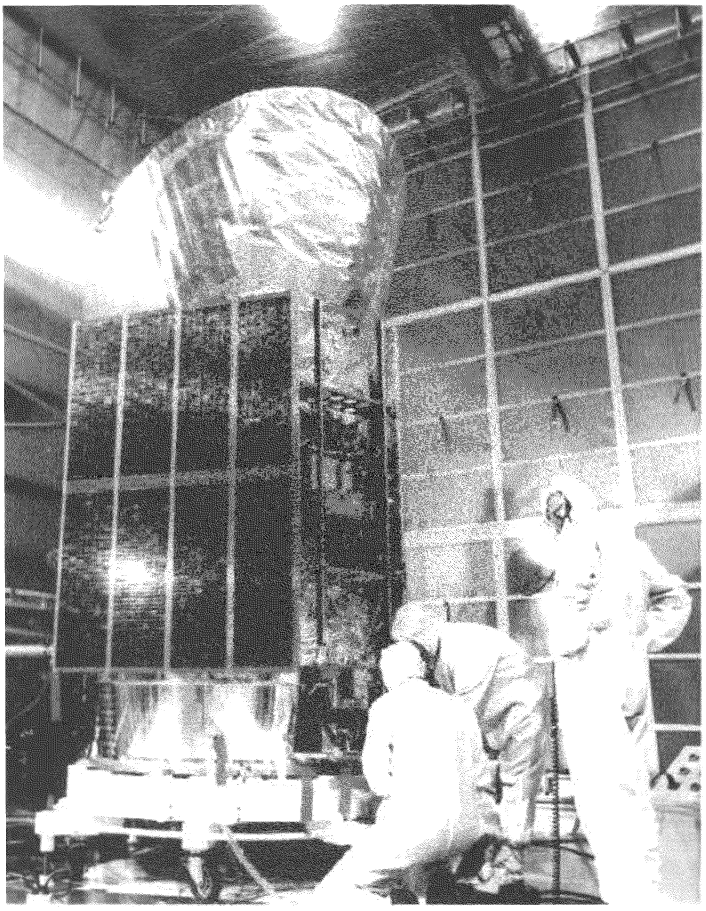

FIG. 21. The Infra-Red Astronomical Spacecraft (IRAS) being prepared for launch. IRAS had a liquid-helium–cooled focal plane and cooled optics. The spacecraft was in orbit for about one year before the liquid helium boiled off and the cooling system failed. (Courtesy NASA.)

servatories are much larger than the previous astronomical spacecraft that have been described. The first one to be launched, on 26 April 1990, was the Hubble Space Telescope, which weighed 10 900 kg. The Hubble Space Telescope (HST) was equipped with a 229-cm-diameter mirror, so that it was a moderately large telescope even by Earth-based standards. The HST was also designed to be launched by the space shuttle, and, more important, it had the capability for revisits. Thus, the HST could be repaired and upgraded by space-shuttle astronauts while it was on Earth orbit. The deployment of the HST from the space shuttle "Discovery" is shown in Fig. 23.

As things turned out, this revisit capability was extremely important. During the construction of the Hubble Space Telescope, a mistake was made in the construction of the optical system. Instead of being a perfect parabola, the mirror was ground with some spherical aberration. After the HST was launched, it was discovered that the telescope could not be properly focused, and after some testing, spherical aberration was found to be the problem. In view of this, a shuttle mission was mounted to fix the difficulty. In December 1993 the space shuttle "Endeavor" was flown to repair the Hubble Space Telescope. A successful rendezvous with the space telescope was made on 5 December 1993. The problem was fixed by inserting small correction lenses into the optics of the telescope. Since then, the Hubble Space Telescope has been working perfectly (Chaisson, 1993).

The Hubble Space Telescope has been used for a remarkably broad series of astronomical observations. In the case of cosmology, excellent pictures of distant galaxies have been obtained (see Figs. 24 and 25).

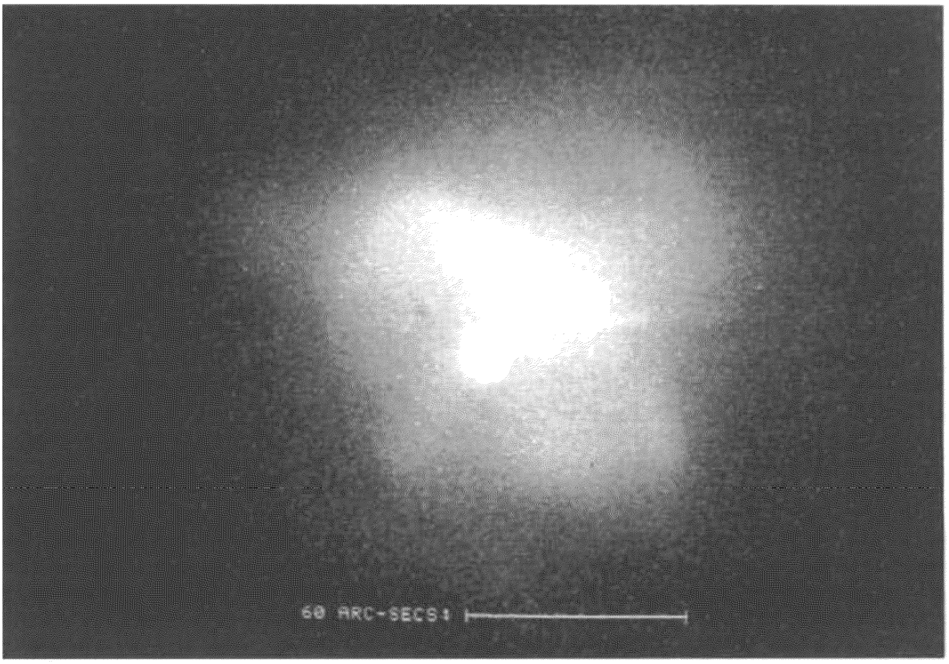

FIG. 22. Picture in the x-ray spectral region of the Crab Nebula produced by the imaging x-ray spectrometer aboard the Einstein (HEAO-2) Observatory. The bright object at the approximate center is the neutron star at the core of the Nebula. (Courtesy NASA.)

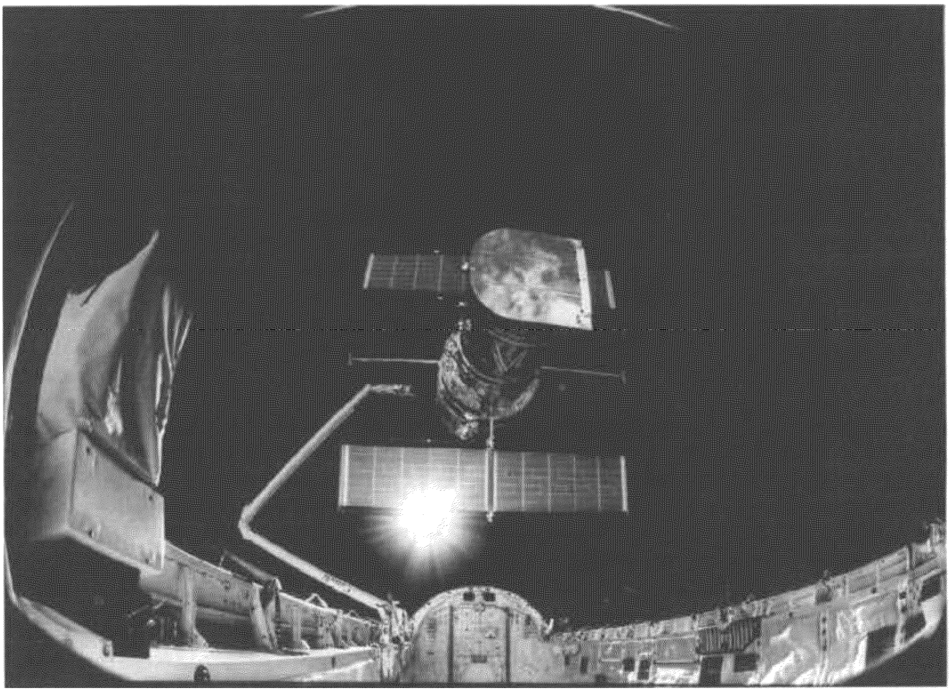

FIG. 23. Deployment of the Hubble Space Telescope from the space shuttle "Discovery" on 25 April 1990 (courtesy NASA).

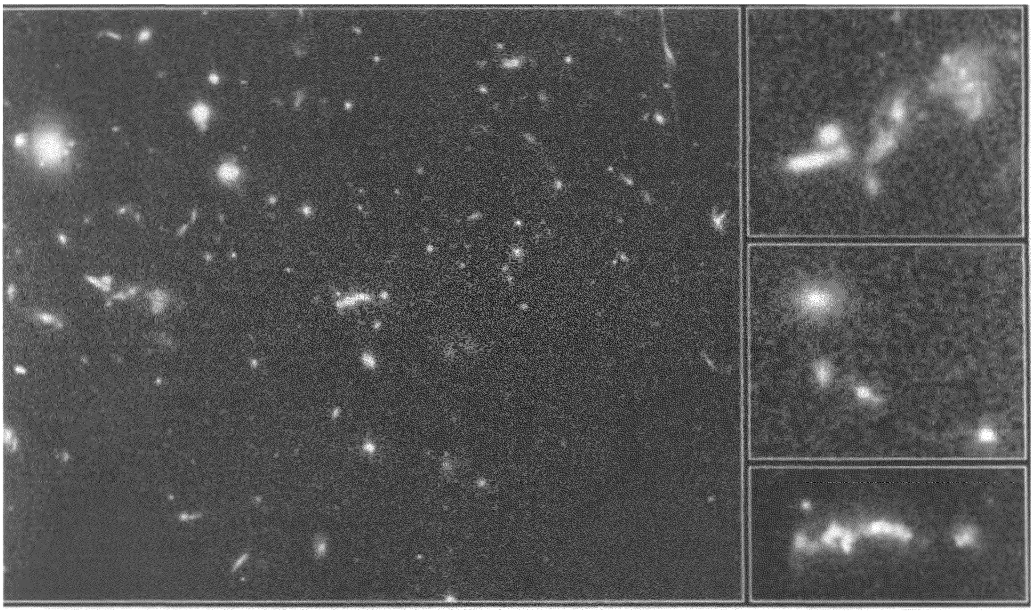

FIG. 24. A distant cluster of galaxies around the object 3D324 taken by the Hubble Space Telescope. The galaxies shown are between (5 and 12) × 10^9 light years away. (Courtesy NASA.)

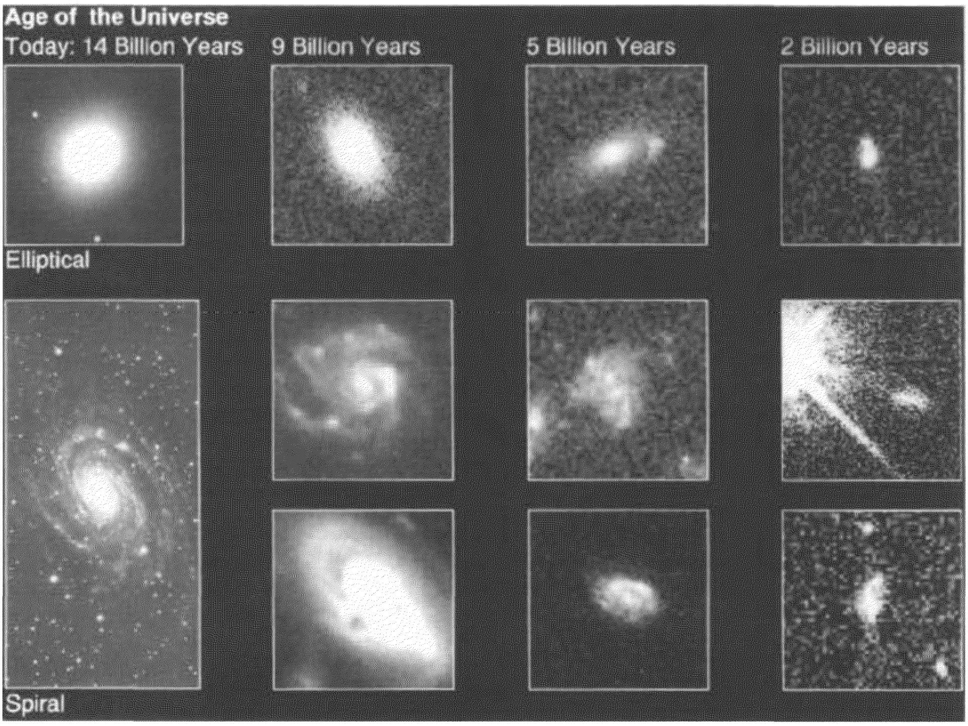
Age of the Universe
Today: 14 Billion Years
9 Billion Years
5 Billion Years
2 Billion Years
Elliptical
Spiral

FIG. 25. A Hubble Space Telescope picture showing how galaxies evolve in time (courtesy NASA).

Perhaps the most important cosmological result so far has been the observation of Cepheid variables in the galaxy M-100, which is 51 × 10^6 light years away (see Fig. 26). This observation has led to a re-evaluation of the "Hubble constant," which measures the rate of expansion of the universe. The new result has reduced the apparent age of the universe from (18 to 20) × 10^9 yr to (12 to 15) × 10^9 yr, which is close to the estimated lifetime of average stars. This has created some questions since the universe may now be too "young" to contain some of the "older" stars. Observations made by the HST have also led to a better understanding of the structure of galaxies. Because of dense dust clouds, it is impossible to see clearly the nucleus of our own galaxy. However, this is not true for other galaxies oriented in other directions, and so good observations of galactic nuclei are now available. From spectrum measurements, it is speculated that galactic nuclei are black holes of great strength.

The Hubble Space Telescope has also been used to make some important observations in our own galaxy. The focus of this work has been on the formation of stars. Figure 27 shows a picture of five young stars in the nebular region near the "belt" of the constellation Orion. Several of these stars are surrounded by disks of dustlike material that may indicate the early stages of the formation of planetary systems.

Finally, the Hubble Space Telescope has been used advantageously to make observations within our own solar system. Some of the most spectacular pictures of the collision of the comet Shoemaker-Levy with the planet Jupiter in July 1994 were taken using the HST (see Fig. 28). Another very important new result from HST observations in the vicinity of our solar system has been the detection of a region of orbiting ice particles

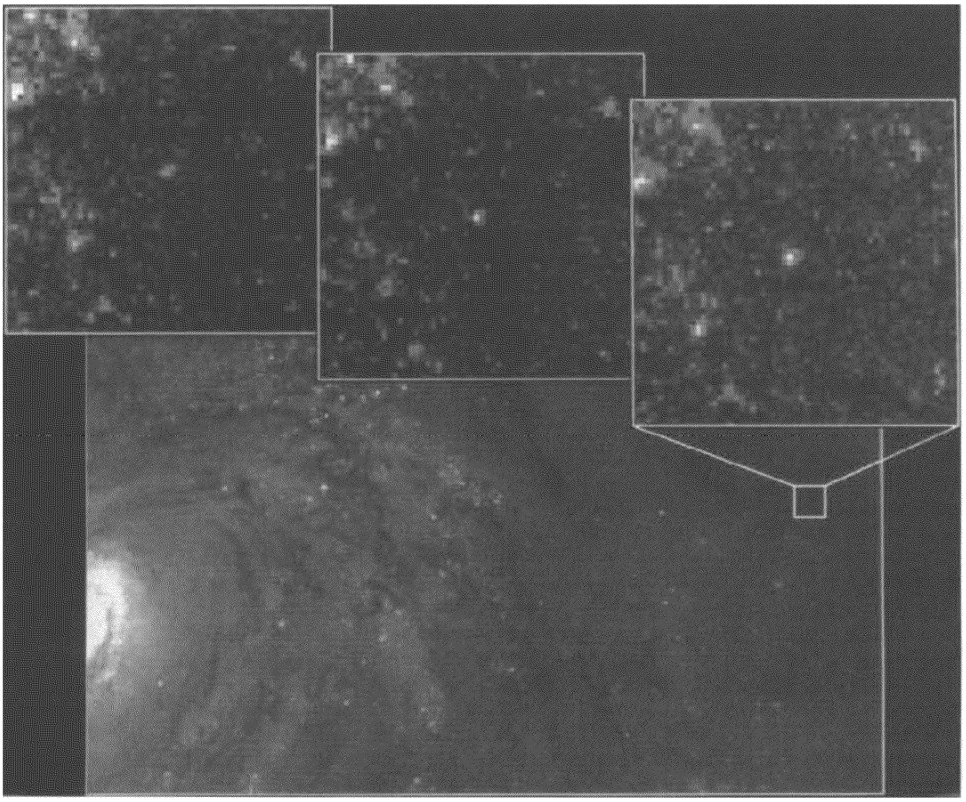

FIG. 26. Cepheid variable stars observed by the Hubble Space Telescope in the galaxy M-100. It was from the analysis of pictures such as this one that astronomers were able to re-evaluate the Hubble Constant. (Courtesy NASA.)

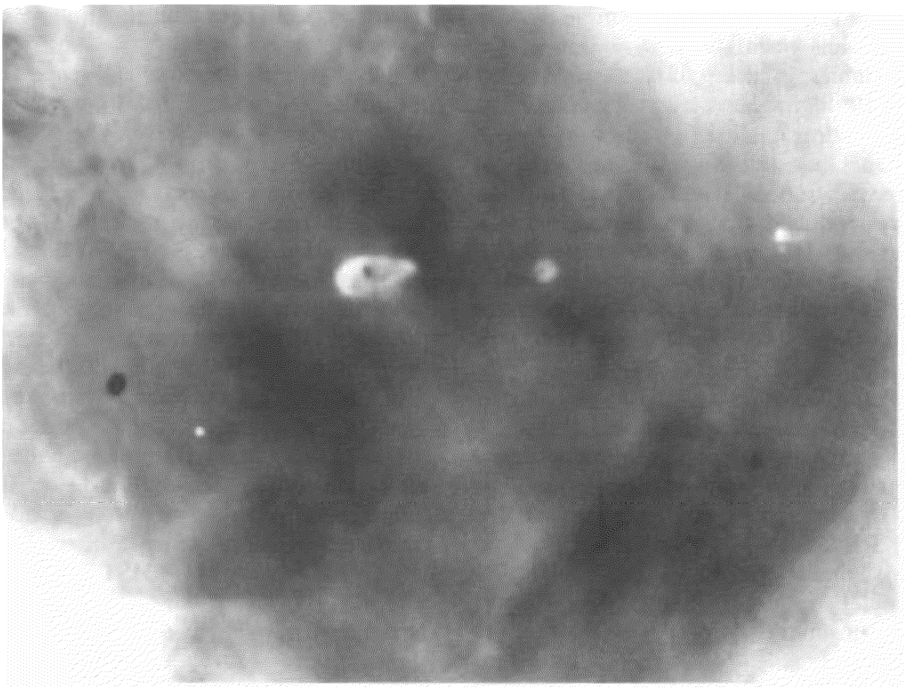

FIG. 27. Five young stars forming in the gaseous region of the constellation Orion (courtesy NASA).

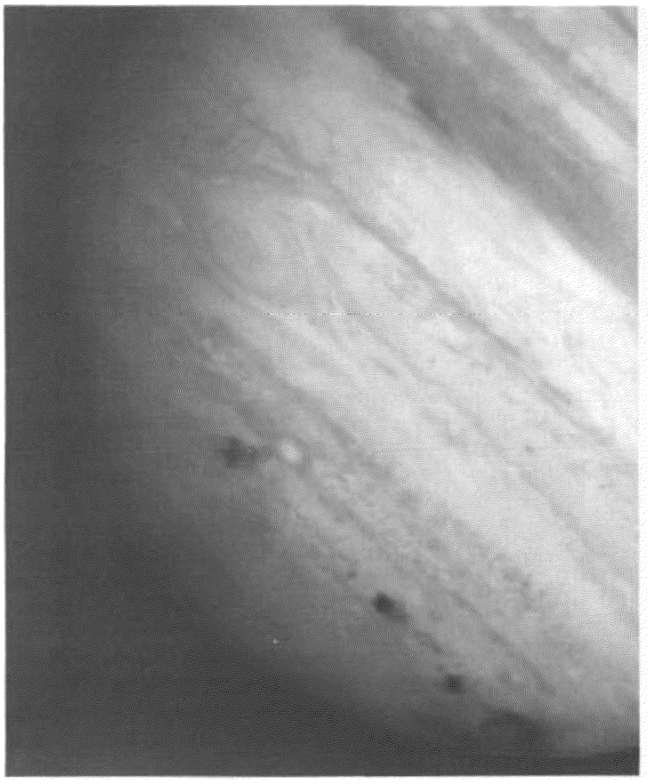

FIG. 28. The pieces of the comet Shoemaker-Levy impacting the planet Jupiter as seen by the Hubble Space Telescope. The black spots on the lower portion of the planet are the impact points. (Courtesy NASA.)

beyond Pluto. This observation confirms the theory of the Dutch astronomer J. H. Oort that comets are formed by the accretion of ice particles in a region beyond the orbit of Pluto. As predicted by Oort, these particles can then be deflected into orbits that are highly eccentric and that permit these objects to come close to the sun.

The Hubble Space Telescope carries a great many other instruments that are designed to be upgraded as detector technology develops. Future shuttle missions are planned in order to implement these upgrades.

The Gamma Ray Observatory (GRO) is another important astronomical observatory that was launched on 8 April 1991. The GRO was renamed "Compton" after it was launched. Because of the high quantum energy of gamma rays, it has been used primarily for cosmological observations. Plate 2 shows the image of a quasar obtained using the Compton GRO. The Compton GRO has also been used extensively to look for evidence of neutron stars and black holes, some of which should be strong sources of gamma rays. Finally, there are the mysterious gamma-ray bursts. The Compton GRO has detected about 1200 of these and has determined that they come from all directions in space—that is, they are isotropic. Whether the gamma-ray bursts are of local (i.e., galactic) origin or whether they come from intergalactic phenomena of some kind is still an open question.

Two other Great Observatories are under development. One is the Advanced X-Ray Astrophysics Facility (AXAF). This spacecraft is meant to be a follow-on to the "Einstein" observatory mentioned earlier. The other spacecraft on the drawing board is a large infrared telescope called the Space Infrared Telescope Facility (SIRTF). It is expected that these Great Observatories will complement and extend the work now being done by the Hubble Space Telescope.

8. THE SPACE SHUTTLE

Even before the Apollo XI astronauts successfully landed on the Moon, NASA was instructed to conduct a series of "Post Apollo Program" studies. In addition, the White House (under President Nixon, by this time) and the National Academy of Sciences also undertook to do "Post Apollo" studies of what should be done. Early in 1969, Vice President Spiro T. Agnew established something called the Space Task Group, which put out an extremely optimistic report arguing that a mission to send people to Mars should be mounted as soon as possible. A study commissioned by the National Academy of Sciences, led by Dr. Louis Branscomb, was more conservative. This study argued that sending people to space was not as productive as using robotic spacecraft to gain scientific information. Their recommendation was to cut back on programs that sent people into space and concentrate more on scientific research with automated spacecraft. Finally, the NASA study, headed by then NASA Associate Administrator for Space Flight, George Mueller, picked up the theme laid down in the articles in *Collier's* magazine some fifteen years earlier. Mueller and his collaborators laid out a program that called for the construction of a space station in Earth orbit, together with the "winged rocket" or reusable spaceship called for in the *Collier's* articles. Eventually, Mueller's plan was adopted, but there was not enough money to do a space station and the "winged rocket," which was named the "space shuttle," at the same time. Thus, a decision had to be made as to which program to execute first. Since it was felt that the space shuttle was technically more difficult and that it would, therefore, be the "pacing item" in the program, the decision was reached to start with the space shuttle.

A great many configurations for a space shuttle were studied. The first designs were intended to be fully reusable. There was a booster, which was designed to fly back to the launch site after taking the orbiter to an altitude from which it could take off and go into Earth orbit. As things turned out, the fully reusable version of the space shuttle was too expensive. Accordingly, another configuration was developed in which only the orbiter was completely reusable. Two solid-fuel rocket boosters designed to get the vehicle into orbit were recoverable, and a large, external fuel tank would be the only item that would be lost in each mission. This so-called "stage and a half" configuration was adopted, and a picture is shown in Fig. 29.

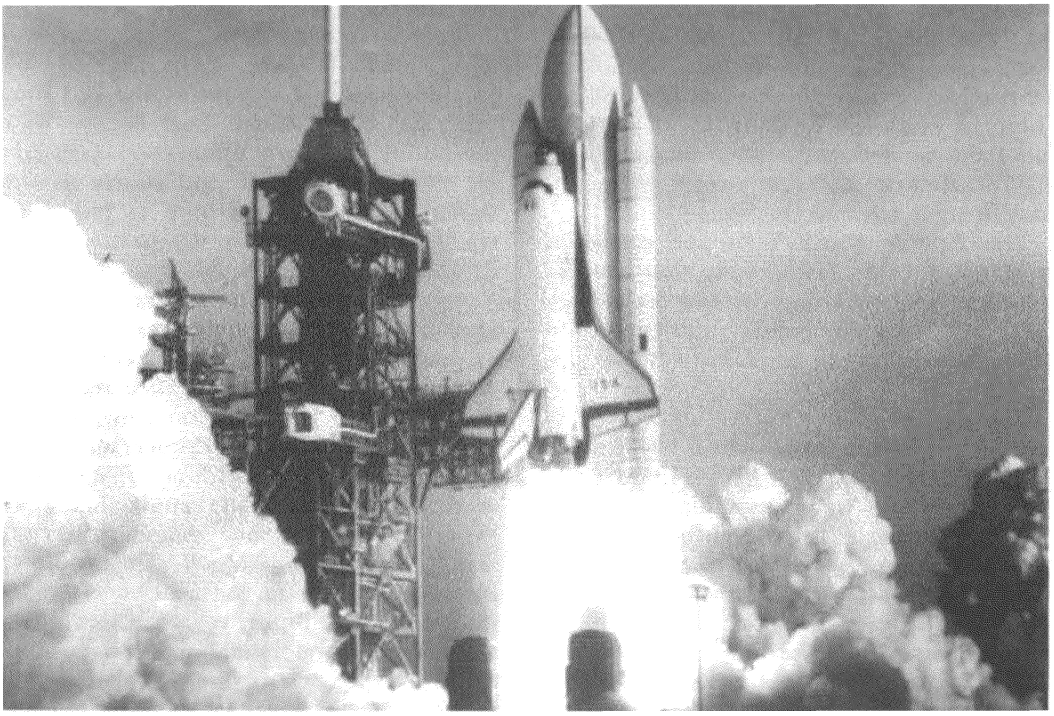

FIG. 29. Space Shuttle "Columbia" taking off from the NASA John F. Kennedy Space Center launch pad on its first flight on 12 April 1981. The orbiter, the solid rocket boosters, and the large external fuel tank are all clearly visible. (Courtesy NASA.)

Another important design consideration in the shuttle was the environment in which the astronauts would be placed. From the very beginning the space shuttle would be different from other people-carrying spacecraft that had been built previously, by the creation of a "shirt sleeves" environment inside the vehicle. This meant that air at atmospheric pressure and atmospheric composition would be the baseline requirement. Furthermore, gravitational forces, which in some cases exceeded $7g$ in the Apollo program, would be held to $2g$ or $3g$. The point here was that the space shuttle should be available for flight by people who were not trained rigorously as astronauts. A category of crew member called "payload specialist" was established, who were trained for a few months to fly on the shuttle to take care of a specific payload. If this were to be done successfully, then the "shirt sleeves" environment described and the relatively low-g force loading had to be built into the vehicle from the very beginning (Mark, 1977).

In 1971, a program was formulated that called for collaboration between NASA and the Air Force to develop the space shuttle. NASA would have the major role to develop the vehicle and the engines, and the Air Force would develop the upper stage, which would be used to boost satellites launched by the shuttle into other orbits. The program as formulated was approved by President Nixon in February 1972. NASA promised to build a reusable orbiter, fly it for the first time in 1978, and do it at a cost of about $6.5 billion in 1972 dollars. The Air Force would develop something called an inertial upper stage (IUS) for something on the order of a billion dollars. As things turned out, the space shuttle flew for the first time in April 1981 and cost about $9 billion in 1972 dollars. While there were cost growths and schedule delays, these were not out of line for a highly sophisticated new technology program such as the space shuttle.

The space-shuttle orbiter was built by Rockwell International; the engines by Rocketdyne, which was a subsidiary of Rockwell; the solid-fuel booster rockets by Morton Thiokol Company; and the external fuel tank by the Martin Corporation. Three NASA centers were principally involved. The overall program management resided at the NASA Lyndon B. Johnson Space Center in Houston, which was also responsible for the development of the space-shuttle orbiter. The propulsion systems were the responsibility of the NASA George C. Marshall Space Flight Center in Huntsville, Alabama. Perhaps one of the most difficult challenges in building the space shuttle was the development of a reusable thermal protection system. The subcontract for the thermal protection system was the responsibility of the Lockheed Missile Systems Division in Sunnyvale, California, and the responsible center was the NASA Ames Research Center in the neighboring city of Mountain View, California.

The creation of a reusable spaceship posed formidable technical challenges. The main engines of the space shuttle operated at pressures that were six times greater than those experienced by the engines that powered the Apollo vehicle. High-speed turbine pumps, for both liquid hydrogen and liquid oxygen, had to be developed. Similar challenges faced the people who were working on the thermal protection system. Ground-based testing facilities were particularly important in this respect, and the high-energy arc jets at the NASA Ames Research Center were critical to the success of this effort. Even so, ground-based testing could not, in the end, simulate the conditions that would be experienced in flight. Therefore, genuine risks had to be taken in the program development effort. Finally, automated control systems had to be developed in order both to launch and to control the space shuttle in flight. The system was sufficiently complicated that no human being could keep track of everything that was going on at the time. This also was the responsibility of the people at the NASA Johnson Space Center and the International Business Machines Corporation (IBM), who developed the computers and software that successfully controlled the space shuttle.

The first flight of the space shuttle "Columbia" occurred on 12 April 1981. Over 70 flights have been conducted in the years since then. In terms of the fulfillment of the original requirements, the program has been very successful. The space shuttle has launched many complex satellites. These were checked out before release, so there have been few failures. Satellites have been retrieved and returned to Earth. Others have been retrieved and repaired on orbit. The retrieval and repair of the Hubble Space Telescope was probably the most spectacular of these missions. Onboard experiments have been carried out using the space shuttle. A project called "Spacelab" is an international element of the space shuttle program. Spacelab is a module system that fits in the payload bay, which was developed and built by the European Space Agency (ESA). Several versions of the Spacelab are available. One is a module designed to be inhabited by astronauts who do experiments while the shuttle is in Earth orbit. Work with the Spacelab module has involved life sciences, materials, and, most important, biological experiments to determine how people can best survive long-duration space flight. In this respect, of course, the space shuttle has not been as important as the work done by the Russians, which is discussed in Sec. 9. Other versions of Spacelab are unmanned palettes mounted in the space-shuttle payload bay, designed to carry instruments that are controlled from the flight deck of the space shuttle. These have been used for Earth observations, atmospheric physics and chemistry, and also astronomical observations. Spacelab modules have collected a respectable body of scientific information over the years.

The major blemish on the space-shuttle program was the loss of "Challenger" on 28 January 1986. Challenger was lost because of the failure of a seal in one of the solid rocket motors that boosts the vehicle into Earth orbit. There were several proximate causes of the accident. One was the cold temperature on the launch day, which made the seal material brittle. A second was that there was some difficulty in assembling the seal, and this may actually have damaged the seal material. Finally, Challenger, on that fateful day, flew through the largest wind shear that had ever been encountered in any space-shuttle flight. Seven astronauts were killed in the accident.

An investigating commission was established under the chairmanship of former Secretary of State, William P. Rogers. The commission recommended various technical and management changes that would serve to prevent future accidents of this kind. As a result of the Rogers Commission recommendations, space-shuttle flights were suspended for 32 months until the successful launch of "Discovery" on 29 September 1988. Ever since then, space-shuttle vehicles have flown successfully on schedule, and there is every reason to believe that this will continue (Rogers, 1986).

The space shuttle has essentially done all of the things that were expected when the program was initiated in 1972. Very probably the space shuttle will be the principal launch vehicle for the next fifteen or twenty years. The space shuttle is essential for the development of a space station and for the things that can be done with that space station. Without the routine access to space that a vehicle like the space shuttle provides, a really capable space station cannot be developed.

The space shuttle has fulfilled the vision that was outlined by the authors of the *Collier's* articles in 1952. There are still cost and management problems with the program, but there is no question that, by any standard, the space shuttle has been a successful exploratory enterprise.

9. THE SPACE STATION

The idea of constructing an Earth-orbiting space station, or "outpost" as it was first called, is not a new idea. It was mentioned for the first time in Hermann Oberth's book (Oberth, 1923). Somewhat later, an Austrian engineer, Count Guido Von Pirquet, published the first calculations of what would have to be done in order to put a space station in Earth orbit (von Pirquet, 1927).

The space station was also a principal item in the *Collier's* magazine articles of 1952–1954. Three important reasons were presented by all the advocates of the space station at the time for constructing such a facility. First, the space station would be a laboratory in Earth orbit in which experiments could be conducted, particularly those related to investigating the effect of long-duration space flight on human beings. Second, the space station would be a maintenance base in Earth orbit from which satellites that had failed could be repaired, or satellites could be upgraded when their instrumentation became obsolete. Third, and perhaps most important, the space station would be a staging base for more ambitious missions to be carried out in the future.

Following the Apollo program, the United States made the decision to develop the space shuttle and to defer construction of a space station until later. The Russians took the opposite view. They have built a series of small space stations, Soyuz, Salyut, and Mir, where each was larger and more sophisticated than its predecessor. The first Salyut was placed into orbit in 1974 and has been followed by several others since then. Perhaps the most unique achievement of the Russian program has been the automated docking and resupply of the Salyut, and later the Mir, space stations.

The focus of the Russian space-station program has been on learning about long-duration space flight. Russian cosmonauts have much more time in space, as of this writing, than American astronauts. Several cosmonauts have spent a year or more in Earth orbit on their Mir space station. Much has been learned from these long-duration space flights about how people can work in the environment of Earth orbit, and how they react to the long confinements in relatively small spaces. All of this is, of course, critically important in terms of learning about future missions to the planets (Newkirk, 1990).

The Russians have done many other things with their space stations. They have conducted extensive Earth-observation programs; they have explored the military value of space stations carrying people; and they have also conducted onboard experiments in materials science and fluid mechanics. All of these things have great value and will be important in the future and larger space stations. A mission to dock an American space-shuttle vehicle with the Russian Mir space station was successfully executed in July 1995. The space shuttle "Atlantis" docked with Mir, and an American astronaut riding on Mir and a Russian cosmonaut on the "Atlantis" exchanged places and returned to Earth with their respective compatriots (see Fig. 30).

The Atlantis–Mir docking is another example of a mission designed to achieve a po-

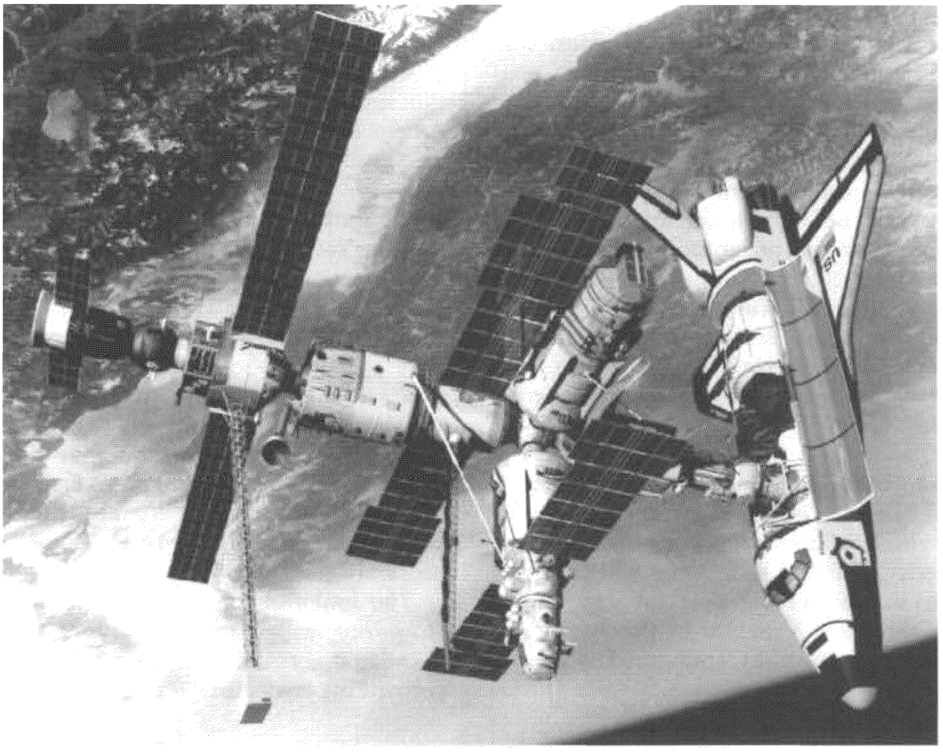

FIG. 30. Space Shuttle "Atlantis" docked to the Russian Mir Space Station as it looked during the joint U.S.–Russian mission executed in July 1995 (courtesy NASA).

litical objective much as the Apollo–Soyuz mission that has already been described. The cost of such missions is sometimes unfavorably compared with the cost of robotic missions that have a much higher scientific payoff. It is important to recognize that the cost of such missions involving international flights with people must be compared to the funds allocated to the execution of U.S. foreign policy (defense, foreign aid, etc.) rather than those devoted to scientific research. In a real way, therefore, the exploration of space with human beings has so far been justified as a "spin-off" resulting from the fulfillment of certain foreign policy objectives.

The United States initiated a Space Station program early in 1984. An International Space Station program was proposed by President Reagan in his State of the Union message delivered on January 25 of that year. The American Space Station program was formulated at the time when the "cold war" was still the dominant feature of the political situation in the world. Following the collapse of the Soviet Union in 1991, Russia was included as an international partner along with the original European partners and the Japanese (Mark, 1993).

Unfortunately, the American Space Station program has had a troubled history. After the Challenger accident there was an escalation of requirements that was, unfortunately, not properly controlled. Therefore, serious increases in the estimated costs were experienced, and there were long schedule delays. As a result of this, the American Space Station program has been on the verge of cancellation a number of times. Were it not for our international commitments, it is quite probable that it would indeed have been halted. When President Clinton was elected in 1992, a complete and thorough review of the Space Station program was conducted (Vest, 1993). The program was cut back so that it would be affordable, and new management techniques were initiated to avoid the mistakes that had been made in the past. Perhaps most important of all, the Space Station program was now reconfigured so that the Russian capabilities were included in the American program.

There is now every reason to believe that a large and capable American space station will be built before the end of the century. The Space Station program remains a truly international effort with European, Japanese, and Russian collaboration. The reasons for building the space station are the same as they were when people first conceived of the idea sixty years ago. The space station will be a laboratory in Earth orbit; it will be a maintenance base with the appropriate orbital transfer vehicles; and finally, it will be a staging base for going to the planets. Most important, it will continue to be an element of American foreign policy that will continue to provide most of the political justification for the program. A picture of the Space Station Alpha as currently configured is shown in Fig. 31.

10. THE FUTURE

On 20 July 1989, the 20th anniversary of the successful landing on the Moon of the Apollo 11 astronauts, President George Bush delivered a speech on the future of the space program. He talked about the shuttle and the space station and the importance of orbiting a space station. He then went on to discuss the longer-term future of the American space program. He made two statements. First he said, "We shall go back to the Moon, this time to stay," and then he went on to say that we will "go on to the next adventure, which is a trip to Mars." These are, of course, the same objectives that were outlined in the *Collier's* magazine articles more than forty years ago. We are still on the same track, and there is no question that these are the right things to do.

The important question is the time scale. How long will it take to achieve the objectives that President Bush outlined in his speech? In answering this question, two im-

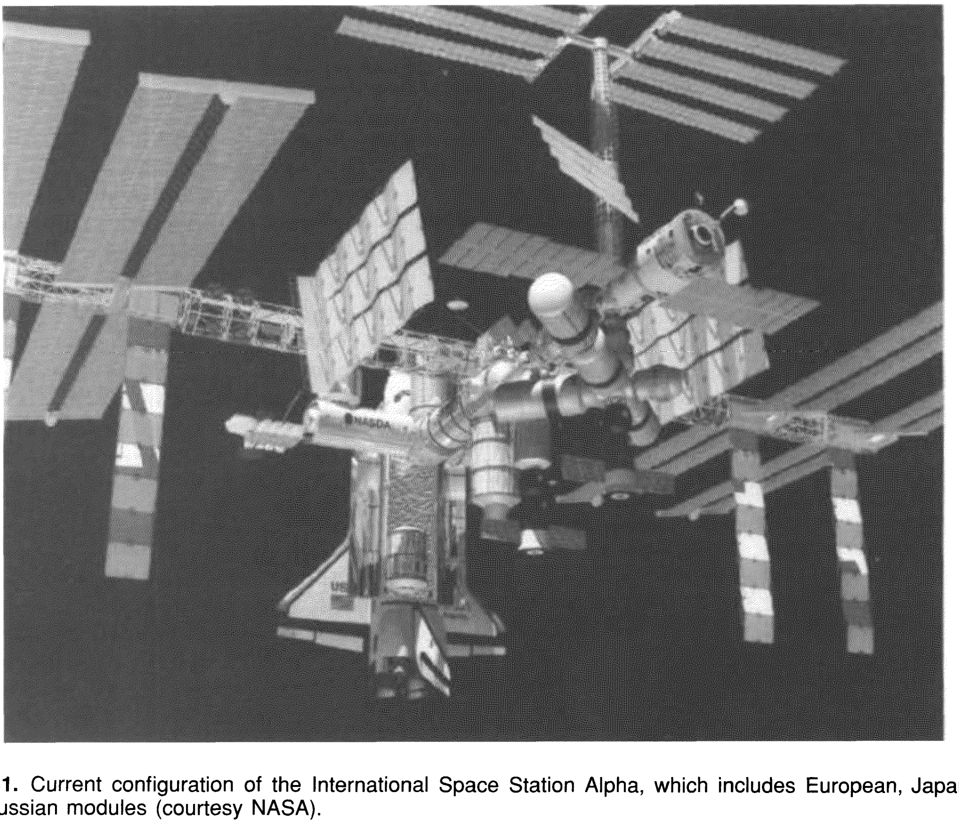

FIG. 31. Current configuration of the International Space Station Alpha, which includes European, Japanese, and Russian modules (courtesy NASA).

portant points need to be made. The first is that the technology to do these things already exists. There is no reason to go to the expense and the delays incurred in the invention of new technology. The second point is that *in situ* resources, both on the Moon and on Mars, should be used to supplement and complement the materials that are brought from Earth (Mark and Smith, 1991).

Given these circumstances, a return to the Moon will probably be possible during the first decade of the next century. The Mars trip will probably have to wait until the year 2020 A.D. This is the most likely timescale given the current state of development.

In addition to the return to the Moon and the trip to Mars, there will also be more sophisticated missions carried out by automated and robotics spacecraft. The most important of these probably is a sample return from the planet Mars. This is a very difficult mission to execute, but it can be done. Sample-return missions would be appropriate precursors for sending people to the planet. There may also be missions to look at asteroids more closely that have already been mentioned. A sample-return mission from an asteroid is probably more easily executed than a sample return from Mars. There is some reason to believe that certain asteroids may actually have some very valuable resources that could be used in the exploration of the solar system.

Finally, improvements will have to be made in launch vehicles and in other elements of the "infrastructure" that we will have to develop to execute the missions that have been described. The space shuttle will fly for another 10 to 15, possibly even 20 years. Nevertheless, we must start now to develop a replacement. The correct approach is to look at vehicles that could fly directly from the Earth to Earth orbit. These so-called "single stage to orbit" vehicles are necessary for really routine space flight. The technology for doing this is not close to being ready. It is probably not as well developed as the technology for the space shuttle was in 1972 when that program was approved. However, it is very likely that the work necessary to bring this technology into existence will be done in the near-term future. It is not too early, therefore, to seriously consider the initiation of a "Single Stage to Orbit" (SSTO) launch vehicle.

Perhaps it is appropriate to end this article by touching, once again, upon the very fundamental question of whether or not the phenomenon that we call "life" exists elsewhere in the universe. NASA has had a small program in the area for over a decade. Recently, this was cut out by the Congress, but private fund raising has kept it going. The fundamental nature of this question is still with us, and we cannot escape that. The importance of the question will cause resurgent interest among many people in continued space exploration. The reason for this was simply stated by Professor Philip Morrison of MIT more than twenty years ago: "Either we find life or we do not. Either answer would be astounding," (McDonough, 1987).

Works Cited

Arnold, James R. (1985), "Ice at the Lunar Poles Revisited," *LPSC* **18,** 29–30.

Augustine, Norman R. (Chmn.) (1990), *Report of the President's Advisory Committee on the Future of the Space Program,* Washington, DC: U.S. GPO.

Baker, Victor R. (1982), *The Channels of Mars,* Austin, TX: Univ. of Texas at Austin Press.

Barghoorn, E. S., Schopf, J. W. (1965), *Science* **150,** 337–339.

Barghoorn, E. S., Schopf, J. W. (1966), *Science* **152,** 775–763.

Bedini, Silvio A., von Braun, Wernher, Whipple, Fred L. (1970), *Moon: Man's Greatest Adventure,* New York: Harry N. Abrams.

Bowyer, S., Byram, E. J., Chubb, T. A., Friedman, H. (1964), *Nature* **201,** 1307–1308.

Bowyer, S., Byram, E. T., Chubb, T. A., Friedman, H. (1965), *Science* **147,** 394–398.

Bowyer, S., Byram, E. T., Chubb, T. A., Friedman, H. (1966), *Science* **152,** 66–71.

Chaisson, Eric (1993), *Astronomy Today,* Englewood Cliffs, NJ: Prentice-Hall.

Chodil, G., Jopson, R. C., Mark, Hans, Seward, F. D., Swift, C. D. (1965), *Phys. Rev. Lett.* **15,** 605–607.

de Bergerac, Cyrano (1657), *Voyages Imaginaires a la Recherche de la Vérité Humaine,* in: R. M. Carré (Ed.) (1970), *Lettres Moderné,* Paris: Minard.

Douglas Aircraft Company (1946), *Preliminary Design of an Experimental World-Circling Spaceship,* Report No. SM-11827, Contract W-33-038 ac-14105, Santa Monica, CA: Douglas Aircraft Company.

Fimmel, Richard O., Colin, Lawrence, Burgess, Eric (1983), *Pioneer Venus,* NAS SP-461, Washington, DC: U.S. GPO.

Flammarion, Camille (1892), *La Planete Mars et ses conditions d'habitabilité,* Paris: Gauthier-Villars et fils.

Galilei, Galileo (1610), *Siderius Nuncius,* Venice: apud Thomam Baglionum.

Giacconi, R., Gursky, H., Paolini, F. R., Rossi, B. B. (1962), *Phys. Rev. Lett.* **9,** 439–443.

Goddard, Robert H. (1919), "A Method for Reaching Extreme Altitudes," Smithsonian Institution, Miscellaneous Collection LXXI, No. 2, p. 69.

Goddard, Robert H. (1920), *Report to the Smithsonian Institution Concerning Further Development of the Rocket Method of Investigating Space,* reprinted in: Esther C. Goddard, G. Edward Pendray (Eds.) (1970), *The Papers of Robert H. Goddard,* New York: McGraw-Hill.

Hall, Charles F., Mark, Hans, Wolfe, John H. (1974), *A Pioneer Odyssey,* NASA SP-349, Washington, DC: U.S. GPO.

Hall, Charles F., Mark, Hans, Wolfe, John H. (1976), "The Journey to Jupiter," *Endeavor* **35** (124), 9–14.

Heiken, Grant H., Yaniman, David T., French, Bevan M. (1991), *Lunar Sourcebook: A User's Guide to the Moon,* Cambridge, U.K.: Cambridge Univ. Press.

Jordan, Stuart (1981), *The Sun as a Star,* NASA Special Publication No. 450, Washington, DC: U.S. GPO.

Mark, Hans (1977), "The Space Shuttle—A Personal View," *J. Vac. Sci. Technol.* **14,** 1234–1249.

Mark, Hans (1993), "The Space Station: In the Beginning," *Planet. Rep.* **13** (6), 8–15.

Mark, Hans, Smith, Harlan J. (1991), "Fast Track to Mars," *Aerospace Am.* **29** (8), 36–41.

McDonough, Thomas R. (1987), *The Search for Extra-Terrestrial Intelligence: Listening for Life in the Cosmos,* New York: Wiley.

NASA (1993), *Pocket Statistics,* Washington, DC: U.S. GPO.

Neilson, W. A. (Ed.) (1956), *Webster's New International Dictionary,* 2nd unabridged ed., Springfield, MA: G. & C. Merriam.

Newkirk, Dennis (1990), *Almanac of Soviet Manned Space Flight,* Houston, TX: Gulf Publishing.

Oberth, Hermann (1923), *Die Rakete zu dem Planetenraum,* Munich: Oldenburg.

Rogers, William P. (Chmn.) (1986), *Report of the Presidential Commission on the Space Shuttle "Challenger" Accident,* Washington, DC: U.S. GPO.

Ryan, C. (Ed.) (1952–1954), "Man Will Conquer Space Soon," *Collier's Magazine,* a series of articles.

Torres, George (1986), *Space Shuttle: A Quantum Leap,* San Francisco: Presidio Press.

Tucker, Wallace H. (1984), *The Star Splitters: The High Energy Astronomy Observatories,* NASA SP-466, Washington, DC: U.S. GPO.

Van Allen, James A. (Ed.) (1958), *The Scientific Uses of Earth Orbiting Satellites,* 2nd ed., Ann Arbor, MI: Univ. of Michigan Press.

Vest, Charles M. (Chmn.) (1993), *Final Report to the President: Advisory Committee on the Redesign of the Space Station,* Washington, DC: U.S. GPO.

von Pirquet, Guido (1927), *Rakete* **2,** 134–140.

Further Reading

Alfvén, Hannes, Arrhenius, Gustav (1976), *Evolution of the Solar System,* NASA Special Publication No. 345, Washington, DC: U.S. GPO.

Collins, Michael, (1988), *Lift Off: The Story of America's Adventure in Space,* New York: Grove Press.

Compton, William David (1989), *Where No Man Has Gone Before: A History of the Apollo Lunar Exploration Missions,* NASA Special Publication No. 4214, Washington, DC: U.S. GPO.

Ezell, Edward Clinton, Ezell, Linda Neuman, (1984), *On Mars: Exploration of the Red Planet 1958–1978,* NASA Special Publication No. 4212, Washington, DC: U.S. GPO.

Gibson, Edward G. (1972), *The Quiet Sun,* NASA Special Publication No. 303, Washington, DC: U.S. GPO.

Katz, Everett J. (1985), *People in Space,* New Brunswick, NJ: Transaction Books.

Harwit, Martin (1981), *Cosmic Discovery: The Search, Scope and Heritage of Astronomy,* New York: Basic Books, Inc.

Hunten, Donald M., Colin, Larry, Donahue, Thomas M., Moroz, V. I. (Eds.) (1983), *Venus,* Tucson: The Univ. of Arizona Press.

Lovell, James A., Kluger, Jeffrey (1994), *Lost Moon,* Boston: Hougton Mifflin Co.

Mark, Hans (1987), *The Space Station: A Personal Journey,* Durham, NC: Duke University Press.

McDougall, Walter A. (1985), *The Heavens and the Earth: A Political History of the Space Age,* New York: Basic Books, Inc.

Nicogossian, Arnauld E., Parker, James F., Jr. (1982), *Space Physiology and Medicine,* NASA Special Publication No. 447, Washington, DC: U.S. GPO.

Ordway, Frederic I., III, Sharpe, Mitchell R. (1982), *The Rocket Team: From the V-2 to the Saturn Moon Rocket,* Cambridge, MA: The MIT Press.

Rossi, Bruno, Olberg, Stanislaw (1970), *Introduction to the Physics of Space,* New York: McGraw-Hill Book Company.

Sagan, Carl (1973), *The Cosmic Connection: An Extraterrestrial Perspective,* Anchor Garden City, NY: Anchor Press, Doubleday.

Trento, Joseph J. (1987), *Prescription for Disaster: From the Glory of Apollo to the Betrayal of the Shuttle,* New Brunswick: Transaction Books.

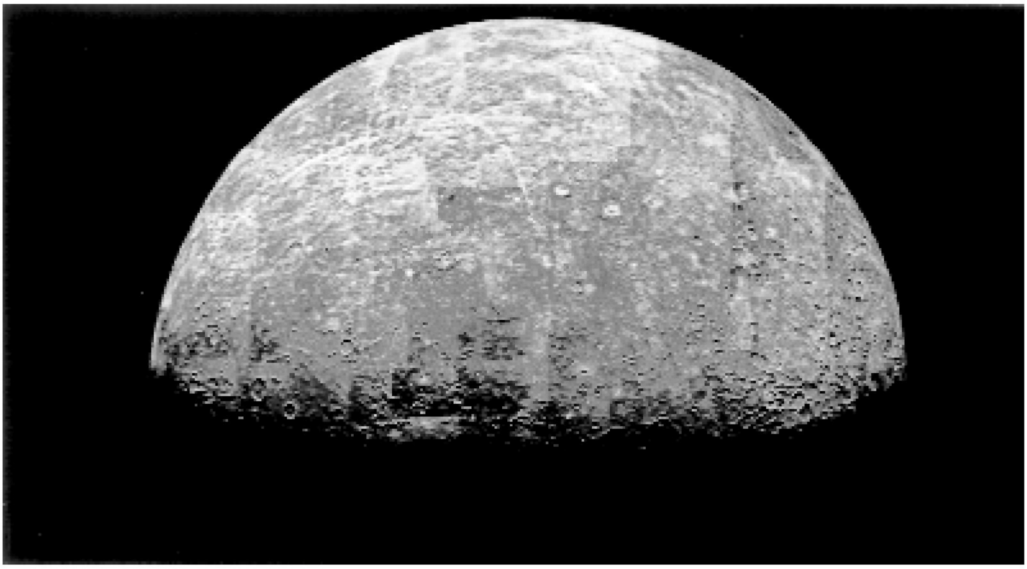

Vol. 19 – PLATE 1. A composite picture of Mercury taken by Mariner 10 on 29 March 1974. The spacecraft approached to within 431 miles of the planet's surface. (Courtesy The Planetary Society).

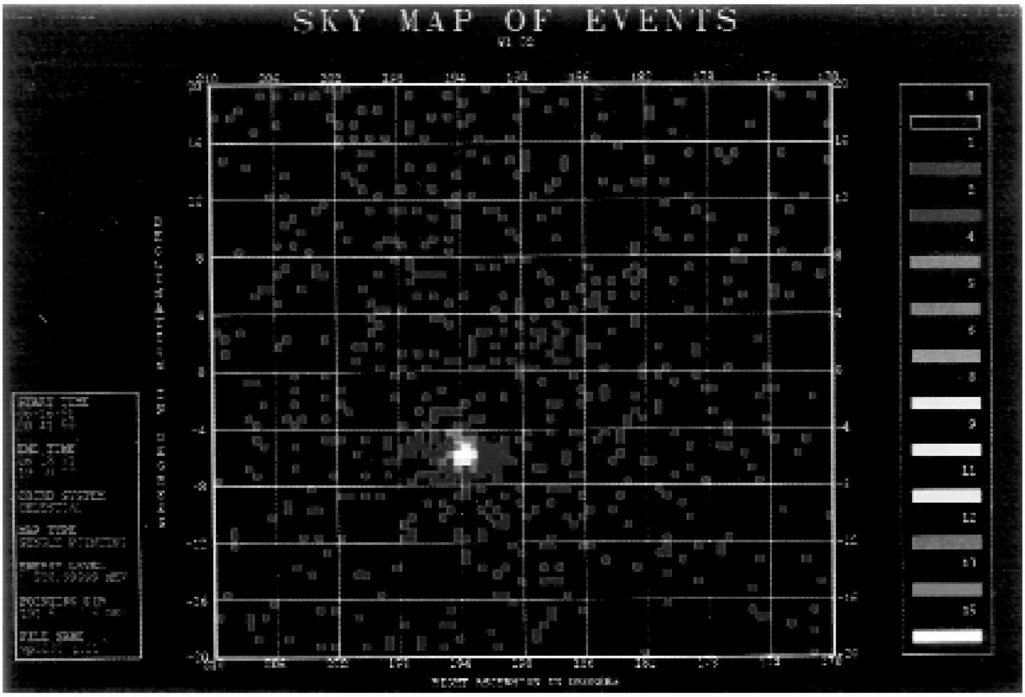
SKY MAP OF EVENTS
START TIME
END TIME
GRID SYSTEM
CELESTIAL
MAP TYPE
ENERGY LEVEL
POINTING DIR
FILE NAME
0
4
8
12
16
20
-4
-8
-12
-16
-20
1
2
4
5
6
8
9
11
12
15

Vol. 19 – PLATE 2. Image of the quasar 3C279 obtained by the Compton Gamma Ray observatory. The quasar is about 6×10^9 light years away, and it emits gamma rays with a total energy a thousand times the energy emitted by the Milky Way galaxy at all wavelengths. (Courtey NASA.)

SPACE-PHYSICS INSTRUMENTATION

Mario H. Acuña and Keith W. Ogilvie, *Laboratory for Extraterrestrial Physics, NASA/ Goddard Space Flight Center, Greenbelt, Maryland, U.S.A.*

Introduction 129
1. **Classification and Types of Instruments** 131
1.1 Fields 131
1.2 Low-Temperature Plasmas 131
1.3 Charged Particles 132
1.4 Energetic Charged Particles and Trapped Radiation 132
1.5 Two-Dimensional Particle Detector Arrays 132
1.6 Imaging and Remote Sensing 132
2. **Instruments Designed for the Measurement of Weak Magnetic, Electric, and Electromagnetic Fields** 132
3. **Instruments for Measurements of the Ambient Thermal Plasma** 138
4. **Instruments for the Measurement of Particles of Medium and Higher Energies** 143
5. **Particle-Imaging Instruments** 145
6. **Visible, UV, and X-ray Imaging Instruments** 146
7. **Special Spacecraft Requirements** 150
8. **Calibration of Space-Physics Instrumentation** 151
9. **Data Handling, Reduction, and Analysis** 151
Glossary 152
Works Cited 154
Further Reading 155

INTRODUCTION

Space physics is that field of science that deals with the study of the medium above a planet's neutral atmosphere and the complex phenomena that take place within it. As such, it includes the ionosphere, magnetosphere, and the interplanetary medium, extending all the way to the Sun's corona. Most of these studies are also relevant to the physics of interstellar space and astrophysical objects.

Early space missions established that this medium is appreciably ionized (i.e., enough to change its properties), is highly conductive, and has a magnetic field imbedded in it, associated either with internal planetary sources or with the supersonic expansion and flow of the extended solar corona (i.e., the solar wind) (see Space Exploration; Magnetospheres). The generic term describing such a medium is a "plasma," an appreciably ionized but electrically neutral gas, and general concepts and principles of plasma physics and magnetohydrodynamics are applicable to its study (see Plasma Physics; Magnetohydrodynamics; Plasma Waves, Oscillations, and Instabilities). Rotation of a planetary body or the supersonic solar wind causes charged particles to flow past the magnetic field lines of a planet, giving rise to complex electric fields with high temporal and spatial variability. The interaction of such flows with different "obstacles" in the space system, such as a planet's magnetic field, atmosphere, or surface, causes the formation of a variety of boundaries, which delineate various regions in space where energy is stored, transported, and dissipated. This system also develops additional and significant temporal variability in response to changing conditions at the plasma source (the Sun), such as solar rotation, sudden energy releases associated with flares and coronal mass ejections (CME's), the 11-year sunspot cycle, and other incompletely understood processes.

The best-known manifestation of this en-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ergy coupling between the Sun's corona and the Earth's ionosphere and atmosphere through the solar wind are the spectacular displays of "northern lights" or aurorae familiar to residents of high-latitude northern and southern regions such as Alaska, Canada, Scandinavia, and New Zealand (see AURORA). In these regions the charged particles coming from the Sun and the solar wind have easy access to the upper atmosphere below by moving along magnetic lines of force that have their footprints near the Earth's magnetic poles and are connected to the distant regions of the magnetosphere where the particles enter the system. The collisions of these particles with atoms in the neutral atmosphere cause them to emit light at characteristic wavelengths, much like a neon sign or electrical gas discharge lamp (see GEOELECTROMAGNETISM; GEOPHYSICS; MAGNETOSPHERES). Complete reviews of these topics and the field of space physics in general are given in Akasofu and Chapman (1972), Hargreaves (1979), Brekke and Egeland (1983), Kennel *et al.* (1979), Al'pert (1990), and Parks (1991).

Space-physics instrumentation refers to the apparatus and instruments carried aboard spacecraft designed to probe these regions of the solar system, generally comprising a variety of detectors that are highly sensitive to magnetic and electric fields, charged particles (electrons and ions), as well as visible, UV, gamma, and x radiation. Neutral-atom imaging instruments that will eventually allow the visualization of the Earth's magnetosphere as an "object" are currently under development. The instruments must be capable of detecting and measuring the very large variety of electrical, magnetic, electromagnetic, and charged-particle phenomena that occur in space; to quantify the flow of energy, mass, and momentum in the system; and to describe fully the populations of charged particles, ions, and electrons at thermal and higher energies, as well as their interaction with the electromagnetic fields. In contrast with the more common remote-sensing functions traditionally carried out from space (telescopes, weather and climate, resource monitoring, infrared radiation, etc.), the majority of the instruments flown on these space missions conduct *in situ* observations. Exceptions to this rule are specialized imagers and electromagnetic radiation detectors designed to conduct remote observations of auroral and other phenomena from space, such as radio waves, gamma and x rays, etc. We have not included in this relatively narrow definition of space-physics instrumentation those instruments associated with remote sensing of solar phenomena (including helioseismology and spectroscopy) because they are more closely related to those designed for astronomical and astrophysical studies. In addition, we do not include ground-based instruments such as radars, ionosondes, and others that are used to infer properties of the medium by probing with electromagnetic waves from below rather than from above.

The increasing dependence of human activities on advanced technology systems, both ground- and space-based, requires that better physical models be developed for the prediction of the response of these highly susceptible systems to solar–terrestrial events. Thus, space-physics instrumentation plays a critical role in obtaining the measurements that will allow a much more detailed and quantitative understanding of the energy flow in the solar–terrestrial system where we live, as well as that of general classes of plasmaphysical phenomena taking place elsewhere in the universe (Acuña *et al.*, 1995).

Two basic factors must be considered in designing space-based measurement systems. They are the *dynamic range* (i.e., "how wide a range") and the *time resolution* of the measurements (i.e., "how fast"). To avoid significant aliasing problems, these must be commensurate with the intensity and the *local* time scales of the phenomena being studied as well as the ratio of the spatial dimension of the regions of interest to the flow and/or spacecraft velocity. The motion of the medium and the spatial boundaries in response to the time variability of the solar inputs affect significantly the interpretation of single-spacecraft observations. The instruments may detect a rapid "change" in the ambient conditions, but the observation cannot be interpreted unambiguously because the crossing of a spatial boundary or a simple temporal variation yield similar signatures in the data. Therefore, simultaneous multi-spacecraft measurements are required to remove these ambiguities from the observations. Simultaneous observations by two spacecraft provide ambiguity resolution only along a straight line (single dimension), while four spacecraft are required to provide full three-dimensional resolution.

Another fundamental fact that distinguishes space physics from other space-based instruments is that the "laboratory" in which they operate (the spacecraft) does not behave just as a simple transportation vehicle to carry them into the regions of interest. In many cases it interacts strongly with the measurements being carried out, and as such it must be carefully designed to minimize interactions that render the observations ambiguous (Harten and Clark, 1995). Often, interactions introduce significant complexity in the analysis and interpretation of the data, making the complete automation of the process a desirable but elusive goal.

An important design consideration for space-physics instrumentation is the "hostility" of certain regions of the space environment, which requires very specialized complex and costly designs. The instruments must operate in vacuum and over a very large temperature range. Spacecraft that operate within planetary magnetospheres may spend significant time within trapped ("Van Allen") radiation belts, which expose the instruments and electronic systems to damaging energetic-particle radiation doses. Ultra-high-reliability designs are mandatory since repairs are generally not possible. High-voltage supplies (up to 30–50 kV) are used in most charged-particle analyzers and detectors and require specialized design and construction to assure their operation and survival. Finally, the limited availability of mass, power, and volume on all spacecraft impose additional constraints on the science instruments that are rarely encountered in ground-based systems.

1. CLASSIFICATION AND TYPES OF INSTRUMENTS

A generic description that has been applied to space-physics instrumentation in the past is "fields and particles instruments." This was based on the typical instrumentation complement flown on the early space missions, which produced discoveries such as the solar wind, the magnetosphere and its boundaries, the Van Allen radiation belts, etc. (see GEOELECTROMAGNETISM; MAGNETOSPHERES; GEOPHYSICS). This very broad characterization is no longer applicable to current instrumentation since imagers and electromagnetic radiation detectors have been added in contemporary missions to provide increased understanding of the coupling among the many phenomena observed locally and their coupling to global processes. Advancing detector and digital technology have also made possible the development of entire new classes of instruments that, for example, can provide multidimensional information about the direction of arrival and velocity of charged particles ("distribution function") and their isotopic composition.

The principles of operation and design concepts for selected instruments from each category are given below. The categories represent an approximate classification scheme useful for descriptive purposes (some of the representative instruments mentioned have multiple measurement capabilities and may appear in more than one category) and are not meant to be exhaustive.

1.1 Fields

Field instruments are those used to detect weak electric, magnetic, and electromagnetic fields. Typical examples include magnetometers (scalar and vector), induction coils, electric field probes (including plasma wave detectors), radio receivers (including appropriate antennas), and particle deflection/reflection instruments that sense indirectly the presence of electric and magnetic fields (see MAGNETIC INSTRUMENTATION; ELECTRICAL INSTRUMENTATION).

1.2 Low-Temperature Plasmas

The low-temperature plasma instruments are used to measure the bulk properties of the local thermal plasma (density, velocity, temperature, etc.). Selected examples include Langmuir probes, Faraday cups [retarding potential analyzers (RPA's)], electron and ion electrostatic analyzers, plasma-wave detectors and sounders, and impedance probes (see SPECTROMETERS, ELECTRON AND ION; SPECTROMETERS, MASS).

1.3 Charged Particles

These instruments are designed to characterize the local population (distribution function, charge state, composition) of charged particles of measurable density. Examples include electrostatic, magnetic, and time-of-flight analyzers (and combinations thereof), and solid-state detectors arranged in a multiplicity of configurations (see SPECTROMETERS, ELECTRON AND ION; DETECTORS, SEMICONDUCTOR; DETECTORS, RADIATION; QUANTUM OPTOELECTRONIC DEVICES).

1.4 Energetic Charged Particles and Trapped Radiation

These instruments are designed to detect and characterize very low densities of medium- and high-energy particles associated with local (e.g., trapped Van Allen belts, auroral particles) or external phenomena (e.g., cosmic rays, solar-particle events, Jovian particles). Examples include high-energy solid-state detector telescopes, scintillation crystals coupled to photodetectors, etc. (see DETECTORS, RADIATION; DETECTORS, SEMICONDUCTOR; SPECTROMETERS, GAMMA; SPECTROMETERS, X-RAY; DETECTORS, SCINTILLATION).

1.5 Two-Dimensional Particle Detector Arrays

The particle-imaging instruments are a new class of instruments that encompass general categories and examples already included in Secs. 1.3 and 1.4. However, the unique characteristics obtained with the use of position-sensitive detectors requires a separate discussion (see DETECTORS, SEMICONDUCTOR).

1.6 Imaging and Remote Sensing

These instruments are designed for imaging and remote sensing of atmospheric or ionospheric emission and absorption phenomena at a variety of wavelengths, including visible, ultraviolet, and x ray. Examples include auroral imagers, x-ray detectors, and imaging and nonimaging photometers and spectrometers at various wavelengths relevant to the phenomena being studied (see OPTICAL INSTRUMENTATION; QUANTUM OPTOELECTRONIC DEVICES).

The foregoing categories are discussed in more detail below.

2. INSTRUMENTS DESIGNED FOR THE MEASUREMENT OF WEAK MAGNETIC, ELECTRIC, AND ELECTROMAGNETIC FIELDS

The accurate measurement of the ambient magnetic-field vector and its orientation in space is recognized as a basic requirement for space-physics studies. The range of field strengths to be measured in exploratory missions may cover up to nine orders of magnitude (5×10^{-3} to 2×10^6 nT; 1 nT $= 10^{-5}$ G $= 10^{-9}$ T), and an angular determination accuracy of the order of one degree is generally sufficient for most studies. In special cases, accuracies of the order of arc seconds are required for detailed mapping surveys of planetary magnetic fields. The time resolution of the measurements ranges from 1 s to several hundred samples per second, depending on objectives (Ness, 1970).

The vast majority of magnetic-field–measuring instruments used to date have been of the *vector* type, which measure three orthogonal components of the local field referenced to an inertial coordinate system. Scalar-type instruments, which measure only the magnitude of the ambient field without providing directional information, have been used for specialized missions where their intrinsic absolute accuracy is required (e.g., geomagnetic-field surveys). The most commonly used vector magnetometers are of the *fluxgate* type. A fluxgate sensor is a magnetic device constructed around a saturable high-permeability core (see Fig. 1). A large-amplitude excitation signal is used to drive the permeable material to saturation at a frequency of 5–20 kHz. The drive current is sufficiently large to cause the permeability of the sensor to switch between a very large value and a much smaller one when the material saturates. Hence, the lines of force associated with an external magnetic field will be "gated" between the two geometries associated with the high- and low-permeability states, as shown in Fig. 1. A balanced winding arrangement is typically used to prevent the large drive signal from appearing at the sense output. The latter is a differential winding, which senses the "gated" external

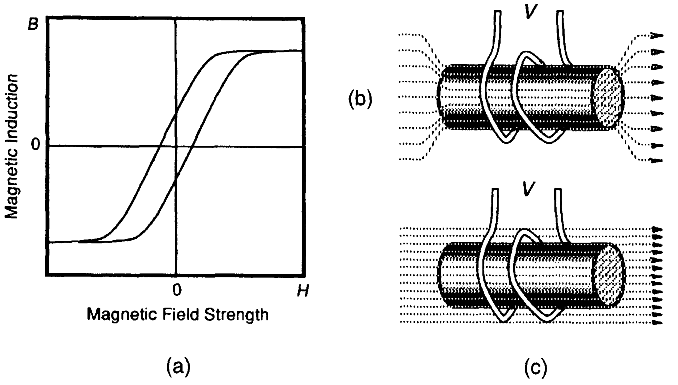

FIG. 1. (a) Magnetization curve of a ferromagnetic core. (b), (c) Basic operation of a fluxgate magnetometer sensor. The ambient magnetic field is switched ("gated") between two geometries (b) and (c) by saturation of a high-permeability material, generating an ac voltage at twice the switching frequency in a pick-up coil.

flux when an external magnetic field is applied. Since the sensor saturates *twice* per drive cycle, the output signal will contain only even harmonics of the drive frequency. This signal is then amplified and synchronously detected to generate a current that is fed back to the sensor through a resistor to cancel out the applied magnetic field. Thus the current amplitude is proportional to the applied field along the axis of the sense/feedback coil. The typical vector magnetometer consists of three fluxgate sensors mounted with their sensitive axes orthogonal to each other.

The ideal transfer function for each individual sensor in this arrangement is given by

$$V_i = |B|k_iR_f\cos\Theta, \qquad (1)$$

where R_f denotes the feedback resistor value, $|B|$ the magnitude of the applied magnetic field, V_i the output voltage corresponding to the ith axis, k_i a proportionality constant determined by the sensor design, and Θ the angle between the external magnetic-field vector direction and the sensitive axis of the fluxgate sensor. By varying R_f, the measurement range can be adjusted over many orders of magnitude. Other less commonly used vector magnetometers include GMR (giant magnetoresistance) based sensors, Hall effect, and optically pumped. The typical frequency response for these instruments is dc to less than 200 Hz (see MAGNETIC INSTRUMENTATION; Ness, 1970).

Scalar magnetometers measure only the magnitude of the ambient magnetic field. The best-known scalar magnetometer is the proton precession type, which takes advantage of the large magnetic moment associated with the proton's spin. In this instrument, a sample of a liquid rich in protons is surrounded by a coil connected to an ac amplifier followed by a frequency (or period) counter. Operation involves a two-step process; the first step consists of applying a dc current to the coil to generate a large magnetic field, which aligns the protons in the sample with the applied field. The second step begins when the current is removed, thus allowing the spinning protons to return to their initial alignment with the ambient magnetic field. This causes the protons to precess about the ambient-field direction, giving rise to a signal in the sense coil at the Larmor frequency given by

$$F[\mathrm{Hz}] = \tau_p|B|/2\pi, \qquad (2)$$

where τ_p is the gyromagnetic ratio of the proton, an absolute quantum mechanical constant, 2.67512×10^8 rad/T·s. This signal is amplified and its frequency or period determined with the electronic counter. Proton precession magnetometers are *absolute* instruments (or primary calibration standards) because their output is related to fundamental constants and does not depend on calibration procedures. However, they only operate in relatively large fields (2×10^4 to 7×10^4 nT) and cannot achieve time resolution better than 1–2 s in their basic configuration. Modified versions have been produced that can generate a continuous precession signal and achieve much higher time resolution (Overhauser effect) at the expense of ultimate accuracy. Another class of scalar magnetometers are of the optically pumped type (rubidium, cesium, helium) where polarized light is used to pump a gas cell into a state where no further absorption is possible and the vapor becomes magnetically polarized and transparent to the pumping light. A weak ac magnetic field is applied transverse to the ambient field to disturb the polarized state of the gas and decrease its transparency to the pumping light, as indicated by a detector at the opposite end of the cell. When the frequency of the transverse applied field coincides with the Larmor frequency given by the gas Zeeman splitting frequency (28 Hz/nT for helium; 6.99 Hz/nT for rubidium), maximum light absorption is obtained. A feedback arrangement is typically used to make the arrangement self-oscillating at the Larmor frequency, thus yielding a continuous signal whose frequency is easily measured. This is a significant advantage over the classical proton-precession magnetometer, but the complexity of these instruments and their significant power consumption has limited their use to very specialized applications (Ness, 1970).

A major problem associated with weak magnetic-field measurements is the magnetic fields generated by the spacecraft subsystems, which can introduce significant errors in the measurements. Of particular concern are high-power circuits like the solar array, batteries, and power distribution systems. To reduce their effect, it has been customary to place the magnetic-field sensors remote from the main spacecraft body, on long deployable booms, which increase the mechanical complexity of the system. This and other accommodation problems are discussed in Sec. 7; see also Harten and Clark (1995).

The outputs from dc and low-frequency magnetic-field sensors are typically digitized with high-resolution analog-to-digital converters to resolve small-amplitude fluctuations superimposed on larger background fields. Spinning spacecraft are preferred for space-physics applications because the rotational motion helps the scanning of ambient particle distribution functions through a large solid angle. However, in the case of magnetometers, the spacecraft spin modulates the ambient magnetic field perpendicular to the spin axis, and additional bandwidth is required to transmit this information to the ground unless onboard de-spin data processing is used. This is a classical example of data-processing and -compression algorithms that are frequently used to reduce the bandwidth required to transmit the information to ground (Lepping *et al.,* 1995; Russell *et al.,* 1995; Behannon *et al.,* 1977).

Another type of ac magnetic-field detectors is classical induction coils, which are also known as "search coils." These are based on the law of induction for a solenoidal coil of n turns wound around a cylindrical core of cross-sectional area A and with permeability μ (Lenz's law),

$$V = \mu n A \cos\Theta \frac{dB}{dt}, \tag{3}$$

where V is the voltage appearing at the coil terminals and Θ the angle between the coil axis and the magnetic-field direction. Like fluxgate magnetometers, an arrangement of three orthogonal sensors is required to carry out vector measurements. Search coils are very sensitive detectors of ac magnetic fields above 10 Hz because their output voltage increases linearly with frequency up to a maximum determined by the self-resonant frequency of the coil. Low-frequency search coils ($f_{\max} < 10$ kHz) use a high-permeability magnetic material as the core to improve the signal-to-noise ratio and minimize mass and size. At higher frequencies, air-core flat rectangular or circular loops outperform magnetic cores but occupy a significant volume if full vector measurements are desired. The outputs of induction-coil sensors are generally coupled to wide-band, large–dynamic-range, low-noise preamplifiers, which feed a signal-processing system similar to those described below (Gurnett *et al.,* 1995).

Electric-field detectors are conceptually extremely simple: Two symmetrical probes or wires are deployed from the spacecraft at a distance d from the center, as shown in Fig. 2. These probes are electrically connected to a very high-impedance differential amplifier whose output is proportional to the difference voltage between the two probes. Thus the measured electric field at dc and frequencies whose wavelength is much longer than the antenna length is simply the ratio of the measured voltage to the distance between the probes. The symmetry of the

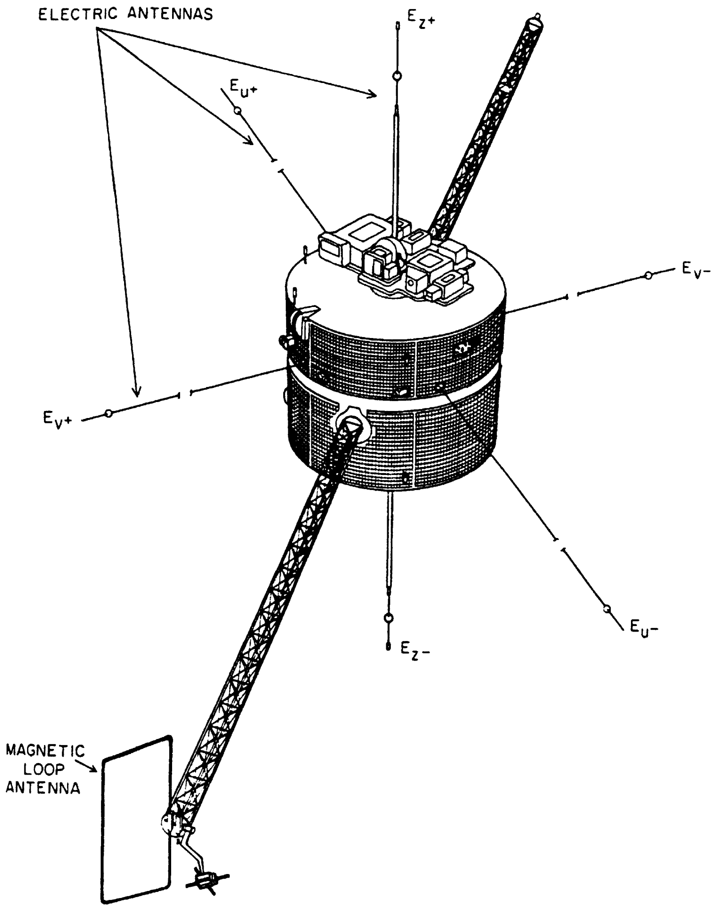

FIG. 2. Electric-field wire antennas and magnetic loop deployed from the POLAR spacecraft. Lengths of tens of meters are required to sense electric fields in very low-density plasmas.

probes and the differential arrangement assure, in principle, that only the external electric field is measured and not that associated with thermoelectric effects and photoemission from probe surfaces, which can significantly affect the accuracy. Measurements can be made along one or more orthogonal axes, each requiring a pair of probes as described above. Although the above may appear to be a straightforward approach, many problems exist with respect to the coupling of the probes to the ambient plasma and the absolute calibration of the measurements, particularly at dc and low frequencies. It is generally accepted that to sense the ambient electric field accurately the distance *d* between the probes must be larger than the shielding distance in the plasma given by the Debye length,

$$D = (kT/4\pi ne^2)^{1/2}, \qquad (4)$$

where *D* is the Debye length in cm, *k* Boltzmann's constant, *T* the temperature in K, *n* the local plasma density, and *e* the electron charge. When the density is small and *T* is high, *D* becomes very large, requiring the deployment of wires or probes separated by tens or even hundreds of meters (Harvey *et al.,* 1995; Maynard *et al.,* 1981).

The electrical signals corresponding to each pair of probes are usually processed on board the spacecraft to extract information regarding wave form, amplitude and phase relationships, spectral content, dynamic range, etc. It is also common to share these signals among several instruments that emphasize different frequency ranges, processing algorithms, and physical phenomena. The most common sharing is between plasma waves and low-frequency electric-field experiments. Electric-field experiments tend to emphasize low-frequency wave-form analysis in the time domain and its correlation to other measurements (like magnetic field and charged particles), while plasma-wave instruments tend to emphasize higher frequencies and spectral analysis with tunable receivers, fast Fourier-transform processors, wide-band wave-form receivers, etc. Typical block diagrams for each class of instruments are given in Figs. 3 and 4. Note that in many cases the receivers and associated spectral and wave-form analysis tools can be connected to either electrical or magnetic antennas, as described above. More recently the data from electric- and magnetic-field detectors have been combined with real-time charged-particle data to study "wave–particle correlations/interactions" in near real time aboard the spacecraft (Gough, 1985). In most cases, the need for advanced onboard processing of multiple-detector signals is driven by the limited data bandwidth available through the telemetry system to the ground or to the onboard data-storage system. A common characteristic of receivers and signal processors associated with electric-field and plasma-wave measurements is their very large dynamic range capability (>100 dB) and the provision of some form of manual or automatic gain control.

A related class of instruments comprises radio-frequency (rf) receivers that are used to monitor the electromagnetic spectrum below approximately 40 MHz. Typical antennas used with these receivers are wire or tubular booms of up to several tens of meters in length. Many physical processes of local and remote origin result in electromagnetic emissions in this range of frequencies. In addition to the Sun, the Earth and Jupiter are powerful sources of electromagnetic emissions associated with plasma-physical processes taking place in their environments. As a representative instrumentation example, the block diagram of the WIND spacecraft WAVES experiment is shown in Fig. 5 (Bougeret *et al.*, 1995). Note the variety of techniques used to sample the electric and magnetic spectrum and process the resulting information:

1. receivers with fixed frequencies and number of channels,
2. receivers tunable by frequency synthesis,
3. time-domain samplers (wave-form receivers), and

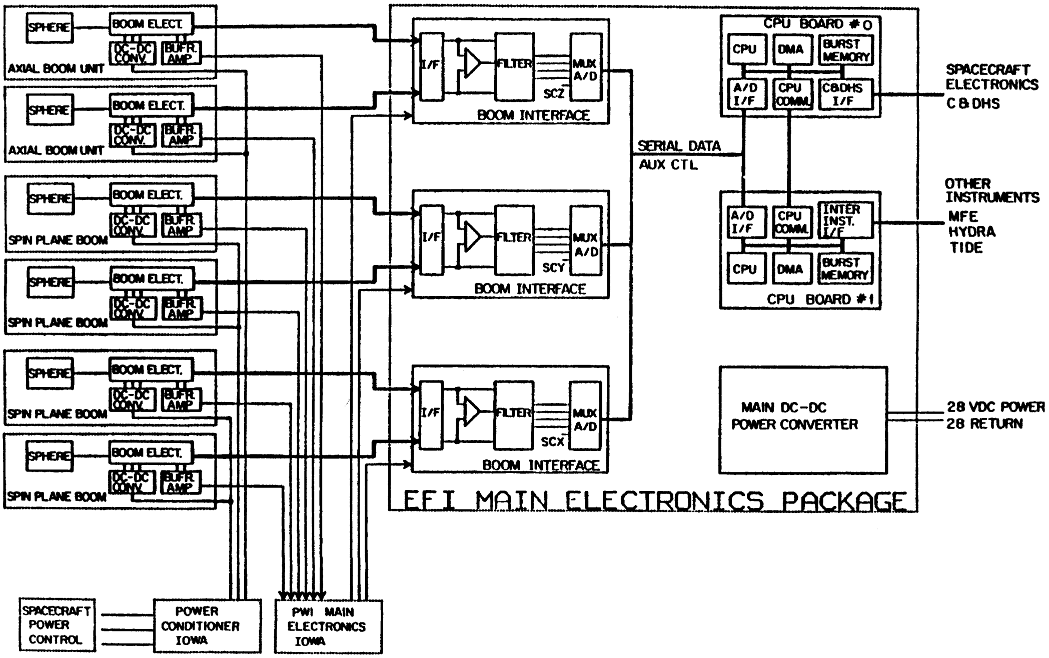
SPHERE
BOOM ELECT.
DC-DC CONV.
BUFR AMP
AXIAL BOOM UNIT
AXIAL BOOM UNIT
SPIN PLANE BOOM
SPIN PLANE BOOM
SPIN PLANE BOOM
SPIN PLANE BOOM
I/F
FILTER
MUX A/D
SCZ
BOOM INTERFACE
SCY
BOOM INTERFACE
SCX
BOOM INTERFACE
CPU BOARD #0
CPU
DMA
BURST MEMORY
A/D I/F
CPU COMM
C&DHS I/F
SPACECRAFT ELECTRONICS C&DHS
SERIAL DATA
AUX CTL
A/D I/F
CPU COMM
INTER INST. I/F
CPU
DMA
BURST MEMORY
CPU BOARD #1
OTHER INSTRUMENTS
MFE
HYDRA
TIDE
MAIN DC-DC POWER CONVERTER
28 VDC POWER
28 RETURN
EFI MAIN ELECTRONICS PACKAGE
SPACECRAFT POWER CONTROL
POWER CONDITIONER IOWA
PWI MAIN ELECTRONICS IOWA

FIG. 3. Block diagram of a representative dc low-frequency electric-field experiment. Wave-form digital signal processing and "burst"-mode data acquisition of fast signals are the main functions implemented in the system.

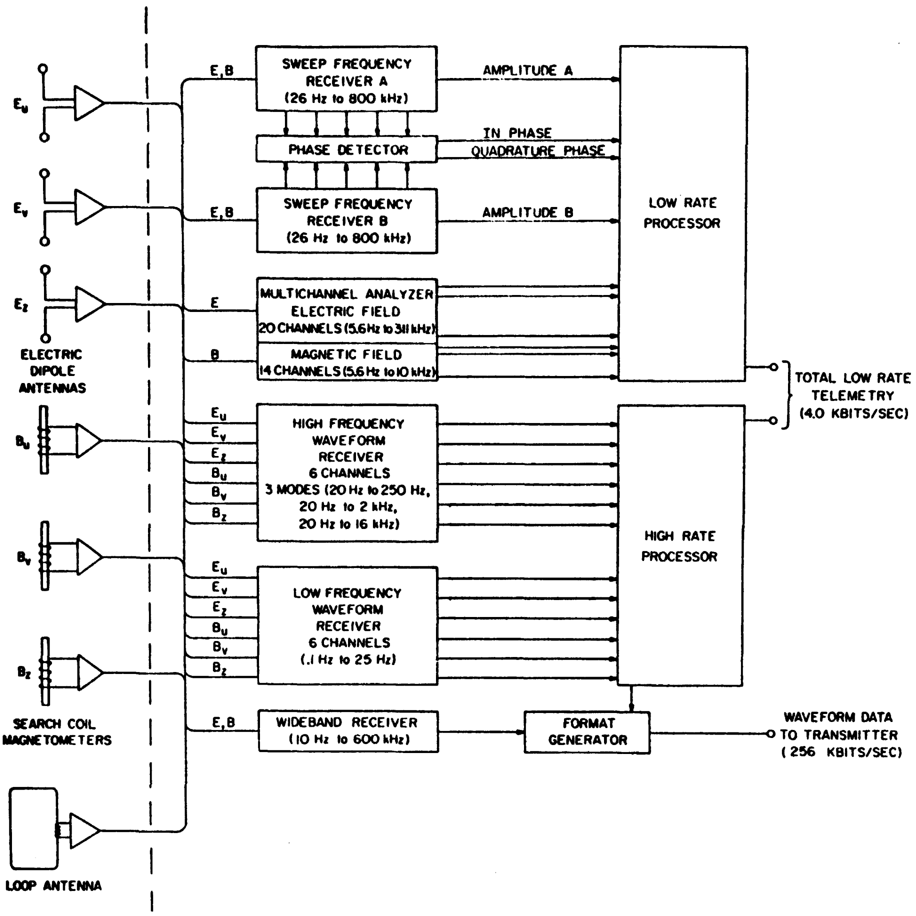
E_u
E_v
E_z
ELECTRIC DIPOLE ANTENNAS
B_u
B_v
B_z
SEARCH COIL MAGNETOMETERS
LOOP ANTENNA
E,B
SWEEP FREQUENCY RECEIVER A (26 Hz to 800 kHz)
PHASE DETECTOR
E,B
SWEEP FREQUENCY RECEIVER B (26 Hz to 800 kHz)
AMPLITUDE A
IN PHASE
QUADRATURE PHASE
AMPLITUDE B
LOW RATE PROCESSOR
E
MULTICHANNEL ANALYZER ELECTRIC FIELD 20 CHANNELS (5.6 Hz to 311 kHz)
B
MAGNETIC FIELD 14 CHANNELS (5.6 Hz to 10 kHz)
TOTAL LOW RATE TELEMETRY (4.0 KBITS/SEC)
E_u E_v E_z B_u B_v B_z
HIGH FREQUENCY WAVEFORM RECEIVER 6 CHANNELS 3 MODES (20 Hz to 250 Hz, 20 Hz to 2 kHz, 20 Hz to 16 kHz)
HIGH RATE PROCESSOR
E_u E_v E_z B_u B_v B_z
LOW FREQUENCY WAVEFORM RECEIVER 6 CHANNELS (.1 Hz to 25 Hz)
E,B
WIDEBAND RECEIVER (10 Hz to 600 kHz)
FORMAT GENERATOR
WAVEFORM DATA TO TRANSMITTER (256 KBITS/SEC)

FIG. 4. Block diagram of a plasma-wave experiment as flown on the POLAR spacecraft. Swept and wide-band receivers, as well as multichannel analyzers, are used to provide information about the ambient ac electric- and magnetic-field frequency spectrum.

4. wide-band receiver and fast Fourier-transform analyzer.

These different receivers represent detailed tradeoffs between time and spectral resolution, dynamic range, and other considerations based on the nature and time variability of the phenomena to be studied. Wide-band receivers (time-domain samplers) are used primarily to obtain time-domain wave-form information, which in many cases is essential to the understanding of the phenomenon being observed. A major problem associated with wide-band data is the very high data rates generated that cannot be handled by the typical onboard telemetry and data-storage systems. Two solutions have been used to circumvent this problem:

1. use the total onboard data-storage capability to store short periods of high-rate data only or
2. transmit directly to the ground for short periods of time when the spacecraft is located in the regions of interest and high data throughputs are possible.

The difficulties associated with the local measurement of weak dc and low-frequency electric fields on dimensional scales much larger than those dictated by probe dimensions and Debye-length limitations has accelerated the development of instruments that inject beams of charged particles to derive the ambient electric (and magnetic) field. The basic principle involves the injection of a narrow beam of electrons of known energy

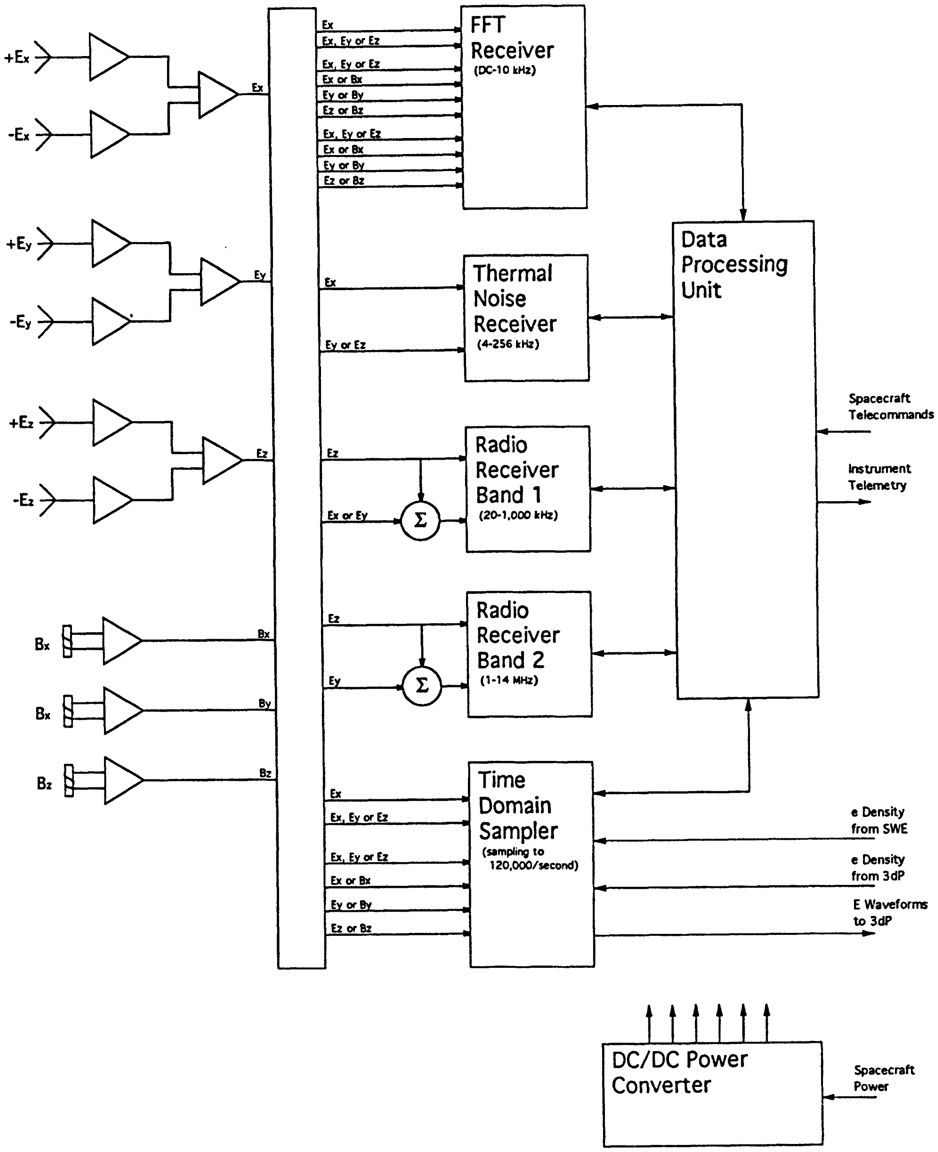
+Ex
-Ex
Ex
+Ey
-Ey
Ey
+Ez
-Ez
Ez
Bx
Bx
Bz
Bx
By
Bz
Ex
Ex, Ey or Ez
Ex, Ey or Ez
Ex or Bx
Ey or By
Ez or Bz
Ex, Ey or Ez
Ex or Bx
Ey or By
Ez or Bz
FFT Receiver
(DC-10 kHz)
Ex
Ey or Ez
Thermal Noise Receiver
(4-256 kHz)
Ez
Ex or Ey
Σ
Radio Receiver Band 1
(20-1,000 kHz)
Ez
Ey
Σ
Radio Receiver Band 2
(1-14 MHz)
Ex
Ex, Ey or Ez
Ex, Ey or Ez
Ex or Bx
Ey or By
Ez or Bz
Time Domain Sampler
(sampling to 120,000/second)
Data Processing Unit
Spacecraft Telecommands
Instrument Telemetry
e Density from SWE
e Density from 3dP
E Waveforms to 3dP
DC/DC Power Converter
Spacecraft Power

FIG. 5. Block diagram of the WIND radio WAVES experiment. A fast Fourier-transform processor (FFT) is used to perform spectral analysis of low-frequency signals. The time-domain sampler provides fast samples (up to 120 000/s) of the signal wave form.

and direction into the ambient medium, and the detection of the return beam, which reflects the effect of the ambient electric and magnetic fields upon the expected direction of arrival. In principle, this technique is free from spacecraft charging and other local contamination effects, but the detection, identification, and tracking of the return beam is a complex process that needs further development (Paschman *et al.*, 1993).

3. INSTRUMENTS FOR MEASUREMENTS OF THE AMBIENT THERMAL PLASMA

As mentioned in the Introduction, these instruments are used to characterize the *bulk* properties of the ambient ionized gas or plasma. In their simplest form, cylindrical, flat, or spherical probes whose potential with respect to the plasma can be varied over a

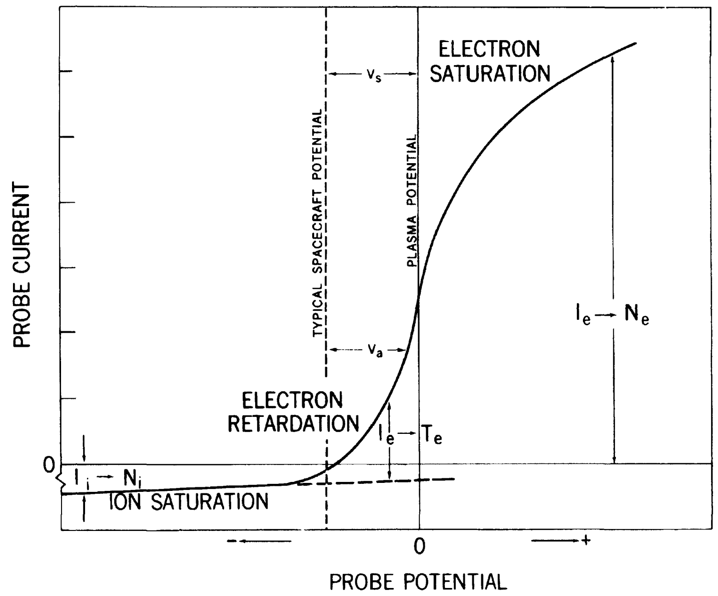

FIG. 6. Current-voltage relationship for a cylindrical Langmuir probe. The regions of the curve from which the ion and electron densities (N_e,N_i) as well as the electron temperature (T_e) are computed are illustrated. The spacecraft potential can also be estimated as shown.

typical range of ±10 to ±30 V with respect to the spacecraft body are used to collect a weak current (typically $<10^{-10}$ A), which is measured as a function of the applied voltage. The magnitude, shape, and slope of the resulting curve of current vs. voltage can be used to infer the electron and ion densities, as well as the temperature and approximate composition of the ambient plasma, as shown in Fig. 6. These probes are generally known as "Langmuir probes" and are often used when the mean free path in the ambient plasma is much greater than the Debye length given in Eq. (4) above (Krehbiel *et al.*, 1981). Their simplicity and ruggedness are attractive features for low-cost missions, but their use is limited to relatively dense plasmas where the collected current is large enough to be measured with conventional techniques. Photoemission from the probe surface and other similar problems (like secondary electrons) can also limit their utility.

Flat-plate collectors with multiple grids or "Faraday cups" are often used for thermal plasma measurements. A simple Faraday-cup arrangement is shown in Fig. 7. This class of plasma sensor belongs to the general category of retarding-potential analyzers or "ion traps." The entrance grid is maintained at a potential that prevents charged particles below a certain energy from reaching the collector. The grid voltage can be varied over a large range to sample the ambient plasma energy distribution. A very successful design based on the application of retarding-potential concepts to solar-wind measurements is the Faraday-cup sensor flown on the WIND spacecraft, a cross section of which is shown in Fig. 8 (Ogilvie *et al.*, 1995). In the supersonic solar wind flow, a simple plate at zero potential will collect a current given by the product of the velocity, charge, and ion density of the solar wind. To determine the energy of the impinging ions and to discriminate between charged particles and the photoelectric current produced by sunlight falling on the collector, a series of grids is used, as shown in the figure. A 200-Hz chopping signal is applied to a modulator grid,

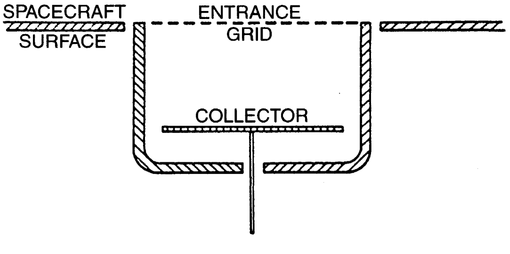

FIG. 7. Simple Faraday-cup collector. The entrance grid shields the collector potential from the exterior of the spacecraft and reduces the errors introduced by photoemission of electrons from the collector surface.

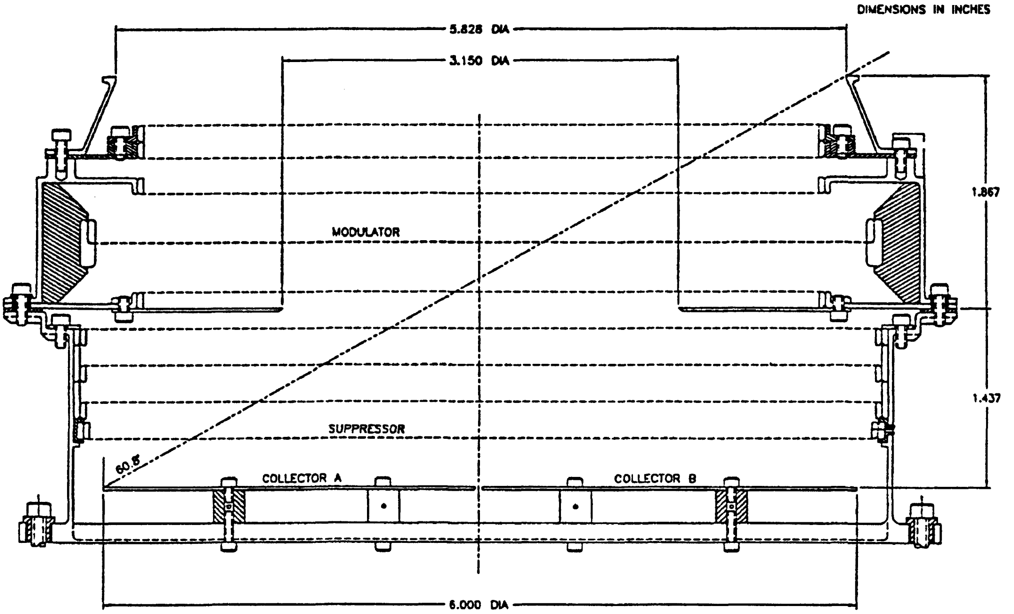

FIG. 8. A cross-sectional view of the WIND Faraday-cup sensor for the solar-wind experiment. All grids are at spacecraft ground potential except for the modulator grid, which has a time-varying voltage applied to it, and the suppressor grid, which is maintained at −120 V to prevent escape of secondary electrons from the collector. Only particles with energy per charge within a narrow range determined by the potentials applied to the modulator grid can make it to the collectors.

which is biased at a potential that can reach as much as 8 kV. The 200-Hz modulating voltage can have amplitudes of up to 1 kV. This modulation results in an ac current in collectors *A* and *B*, which is ac amplified and synchronously detected (see ELECTRICAL MEASUREMENTS). This approach avoids the need for complex, stable, very low-dc-current amplifiers and allows the determination of the direction of arrival of the supersonic flow to better than 1°. In addition, by varying the ratio of bias to chopping voltage amplitudes, the velocity distribution function of the ambient plasma can be determined with good accuracy. The Faraday cup has two great advantages. It only requires approximately five cycles of the modulating voltage to make a measurement, and so it is capable of relatively high time resolution; and the energy per charge or velocity bandwidth $\Delta(E/Q)/(E/Q)$ can be varied, which is not possible in the electrostatic analyzers described below.

Another class of thermal plasma instruments is built around electrostatic deflection analyzers. Charged particles with mass *m* and charge *q* are allowed to enter the space between two parallel cylindrical or spherical plates of radius *b* and *a* (see Fig. 9). A potential difference V_p is applied between the plates to create an electric field that deflects the particles from their straight-line trajectories. At the exit of the deflection assembly, a

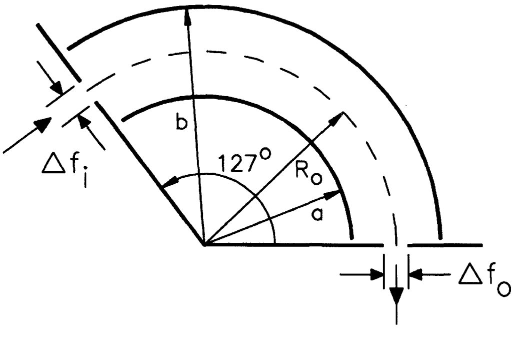

FIG. 9. Parallel-plate cylindrical or hemispherical electrostatic analyzer. The potential applied to the plates deflects the trajectory of the incoming particles such that only those with the correct energy per charge will traverse the analyzer and pass the exit aperture into a suitable detector.

particle detector in the form of an electron multiplier (planar or channel) or collector is used to detect the arrival of the particle and generate a corresponding signal for further processing. It can be shown that for a cylindrical analyzer only particles that satisfy the relationship

$$E_0 = \frac{qV_p}{2\ln(b/a)} \tag{5}$$

will be able to make it through the analyzer along the central ray. By varying the voltage between the plates in steps or by sweeping, the entire distribution function can be sampled. However, this process may involve a long time if the fluxes of particles are small. The *energy resolution* of the analyzer ($\Delta E/E$) is proportional to the ratio $b/(b - a)$, and so high-resolution instruments covering energy ranges up to several keV tend to become large to withstand the high voltages involved. Note that (5) above implies a selection of *energy per charge,* and hence it is possible for more than one species of particles to be allowed through the analyzer. Depending on the application and relative fluxes of particles, this characteristic may introduce errors and limitations in the measurements (see SPECTROMETERS, ION AND ELECTRON).

If the arc covered by the plates of a cylindrical electrostatic analyzer is 127°, it can be shown that particles passing through a narrow slit placed in front of the analyzer aperture will be "focused" onto an equivalent narrow slit at the output, where the detector would be located. This one-dimensional focusing property of the cylindrical analyzer is used very frequently to achieve the desired solid-angle coverage and geometric factor (defined as the product of the sensitive area times the solid angle covered). To obtain full 4π steradian coverage on spinning spacecraft, quadrispherical analyzers have been used. A typical instrument used in the Active Magnetospheric Particle Tracer Explorer (AMPTE), Mars Global Surveyor (MGS), GIOTTO (mission to comet Halley), and WIND spacecraft is the "top-hat" analyzer shown in Fig. 10 (Lin *et al.,* 1995; Acuña *et al.,* 1992; Paschmann *et al.,* 1985). This analyzer consists of two hemispherical plates separated by a small distance. On the outer plate a section is removed at the top to allow entry of the ambient particles into the analyzer over 360° in polar angle but over a narrow range of azimuth (elevation) angles. The particles entering the "top hat" and satisfying (5) are focused onto an annular detector covering almost 360°. A position-sensing detector anode is typically used to determine the polar angle of arrival of the particle. Therefore, if mounted on a spinning spacecraft, a single top-hat analyzer can obtain a sampling of the full distribution function in a single revolution.

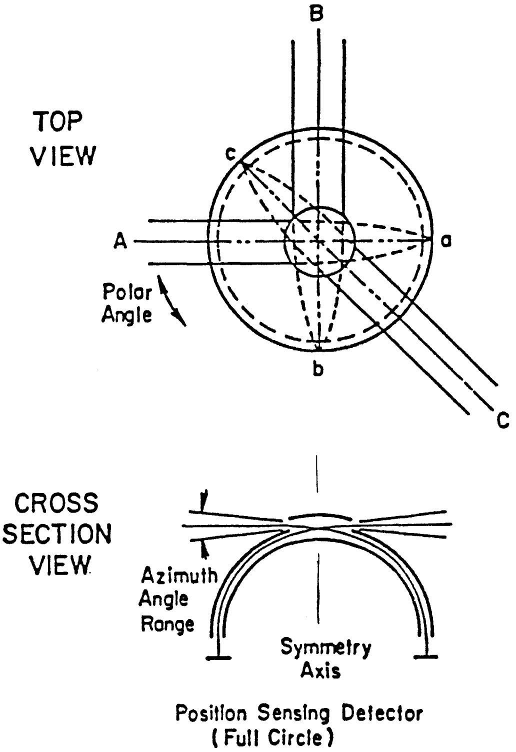

FIG. 10. The "top-hat" hemispherical electrostatic analyzer. The field of view (or angular acceptance region) is a cylindrically symmetrical conical torus, usually of dimensions 10°–20° × 360°. The top view illustrates the focusing properties of the analyzer for particles arriving from representative directions A, B, and C.

Electrostatic analyzers with circular or toroidal geometries similar to the top hat described above have gained popularity in a variety of thermal plasma and ion mass-spectrometry instruments because of their large geometric factor, fast response, and angular resolution (Shelley *et al.,* 1995).

As described above, the fundamental measurement in a plasma is the velocity distribution function, a function of direction and ion species as well as velocity. By taking moments of this function, one can, in principle, learn everything about the plasma. In practice the distribution in three dimensions can be well-enough determined in space to yield the three components of the plasma drift speed, temperatures parallel and perpendicular to the ambient magnetic field, and the heat flux in the plasma for the dominant ion species present. This has been accomplished using both Faraday cups and electrostatic analyzers. The use of spinning spacecraft facilitates the process but limits the time resolution to one spin time. Analogous quantities for electrons can also be obtained by reversing the polarity of the plate or grid potentials.

Because the electrostatic analyzer provides dispersion in energy per charge, a problem arises in assigning a peak in the observed spectrum to a particular ion species. A crossed-field (**E** × **B**) analyzer, usually in the form of a Wien velocity filter, was used in early designs to determine the particle velocity and resolve these ambiguities in ion mass-spectrometer applications (see SPECTROMETERS, MASS). Current practice made possible by technological advances in high-speed logic, carbon foils, channel-plate multipliers, and high-voltage techniques is to obtain individual particle velocity information using time-of-flight (TOF) techniques. In the TOF analyzer, the ions exiting an electrostatic analyzer are post-accelerated by a high voltage (~30 kV) and allowed to strike a thin carbon foil placed in their path. The secondary electrons produced are detected by a microchannel-plate electron multiplier and generate a "start" signal, which is fed to a high-speed logic system. After traversing the length of the analyzer (typically 10–20 cm), the ion strikes a second carbon foil, which, through its associated microchannel-plate electron multiplier, generates a "stop" signal that is used to calculate the time of flight of the ion in question. Many designs add a solid-state detector after the stop foil to measure the total energy of the particle, making possible the determination of its charge state. A typical ion-composition instrument incorporating these principles is the SWICS instrument flown on several interplanetary spacecraft and shown in Fig. 11 (Gloeckler *et al.*, 1995).

Neutral mass spectrometers are used at ionospheric and lower altitudes where the neutral mass density is significant. A typical instrument is the quadrupole mass spectrometer, as illustrated in Fig. 12 (see SPECTROMETERS, MASS). Four cylindrical (or hyperbolic) rods are arranged along the ion path as shown. An electron multiplier or similar particle detector is located at the exit end of

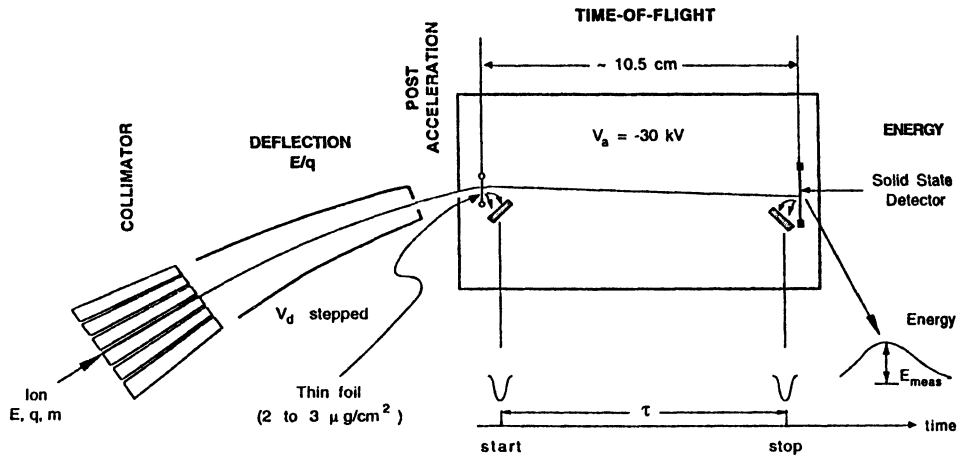

FIG. 11. An overview of the Solar Wind Ion Composition Sensor (SWICS) aboard the WIND spacecraft. Ions with a given energy per charge are selected by the entrance electrostatic analyzer and proceed into a time-of-flight analyzer, which determines their velocity after acceleration by a −30 kV potential. Finally, they stop at a solid-state detector where their total energy is measured.

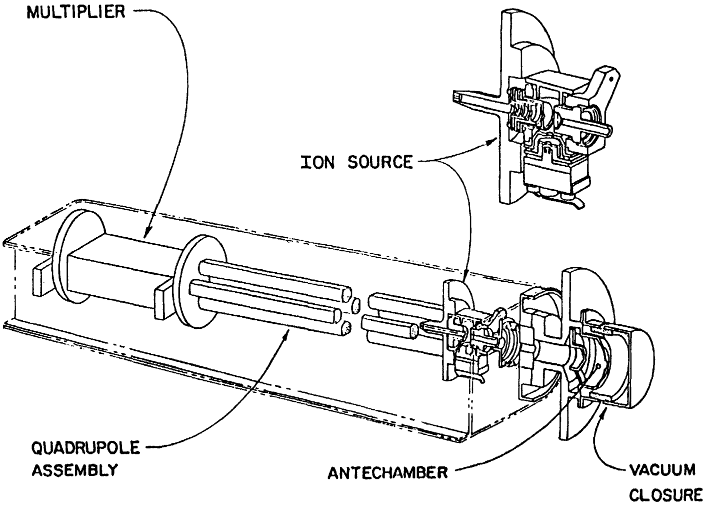

FIG. 12. Quadrupole neutral mass spectrometer. The ambient neutral atoms are first ionized and then admitted into the quadrupole assembly. A combination of dc and ac voltages is applied to the four rods such that only ions within a narrow mass range are able to make it to the exit aperture, where they are detected with an electron multiplier or similar device.

the quadrupole assembly and a device to ionize the neutral gas is located at the entrance end. A combination of dc and rf voltages is applied to the rod pairs and swept over the appropriate range to enable a mass selection between 2 and 50 amu with a resolution of 1 amu. Ions of the selected mass-to-charge ratio are stable in the field and are able to traverse the length of the rods without being deflected out of the central path. These stable ions are therefore able to strike the detector, where each ion produces a pulse of 1×10^6 or more electrons. The number of pulses per unit time is counted and is proportional to the source density of the selected mass (Carignan *et al.*, 1981).

4. INSTRUMENTS FOR THE MEASUREMENT OF PARTICLES OF MEDIUM AND HIGHER ENERGIES

As the energy of the particles to be measured increases, the size, mass, and dimensions of the electrostatic, magnetic, and time-of-flight analyzers described above become impossible to accommodate on typical-size spacecraft. For particles above several tens of keV, the use of thin and thick solid-state detectors, scintillators, and gas-filled counters becomes the preferred detection and measurement technique (see DETECTORS, SCINTILLATION; DETECTORS, RADIATION; Knoll, 1990). Early space missions used Geiger–Mueller counters to detect the existence and characteristics of the Earth's radiation belts. Today, solid-state detector assemblies and scintillators of diverse characteristics and construction are used in most missions because of their versatility and the wealth of information that they provide. Gas-filled detectors (or ionization chambers) are still used for proportional counters, such as those used in the measurement of low-energy x-ray and gamma-ray emissions.

The principles and application of scintillation-based detectors are thoroughly described in Knoll (1990). The basic measurement principle is the detection of light produced in selected materials when an energetic particle traverses the same. This "scintillator" is optically coupled to a photomultiplier tube or avalanche photodiode, which generates a charge pulse in response to the photons arriving at the photodetector. Typical scintillation materials are sodium and cesium iodide crystals as well as many specialized plastics (see DETECTORS, SCINTILLATION). Their primary application is in gamma-ray spectrometry, a field more closely associated with astrophysics and solar physics than space physics. In some instances, interplanetary space-physics missions, because of their orbits and large separation, carry gamma-ray instrumentation to study phenomena such as the mysterious gamma-ray bursts. The KONUS instrument placed aboard the WIND spacecraft is representative of a typical design for this class of instruments, and its detector is shown in

Fig. 13 (Aptekar *et al.*, 1995). A Pb/Sn-shielded sodium iodide crystal scintillator is coupled to a photomultiplier tube and its output processed by a simple multichannel spectrum analyzer, which classifies the detected gamma rays in eight energy bins ranging from 10 to 750 keV and telemeters the information to ground.

The greater density of solid-state detectors implies that particles of much higher energy can be detected for equivalent detector size as compared to gas-filled counters. This is very important in space missions where mass and volume are at a premium. The passage of an energetic particle or gamma ray through a semiconductor material gives rise to the formation of electron–hole pairs in the same. If an electric field is present, such as that associated with a reverse-biased *p-n* junction, the charge carriers will be collected rapidly toward the contacts of the *p-n* junction and can be measured with a charge-sensitive preamplifier (see DETECTORS, SEMICONDUCTOR). The many types of solid-state detectors available today and their characteristics are fully described in Knoll (1990). Silicon diodes have become the detector of choice in the majority of applications. Of particular interest to space-physics applications is the measurement of energy loss of the incident particle in thin solid-state detector assemblies, which allows the identification of the particle. In a thin detector the number of charge carriers created within the detector, Q, is proportional to the energy loss of a particle (dE/dx) as it travels through the detector, times its thickness t:

$$Q = t\frac{dE}{dx}. \tag{6}$$

However, the particle traverses the detector completely, retaining most of its original energy. When a thick detector that completely stops the incident particle is added, its total energy is measured. This arrangement, called a "dE/dx by E telescope," is frequently used in instruments designed to identify simultaneously the mass and energy of each incident particle. An example of a contemporary instrument using a large variety of integrated solid-state detector telescopes is the EPACT instrument flown aboard the WIND spacecraft (Von Rosenvinge *et al.*, 1995). As an example, the alpha-proton-electron (APE) telescope used in this instrument is shown in Fig. 14 and is representative of the complexity and variety of detector types that are in use today. This telescope consists entirely of lithium-drifted (LiD) detectors arranged in the form shown. Some detectors (B1 and B2 in the figure) are curved to minimize path-length variations. Individual charge amplifiers associated with each detector measure the energy loss through the same, and particles stopping in this telescope are identified by coincidence conditions through detectors B1 and B2 and C7. The detailed design of energetic-particle telescopes requires the optimization of a multiplicity of parameters, whose description is beyond the scope of this work. Figure 15 illustrates the isotopic separation that can be achieved with isotopic composition telescopes using solid-state detectors like in the EPACT instrument. Note the clear identification of magnesium isotopes that is achieved. The electronic systems associated with multiple-detector telescopes tend to be fairly complex because of the large number of analog and digital functions required to process, combine, and classify the signals generated by each detector. Extensive use is made of complex microcircuits, microproces-

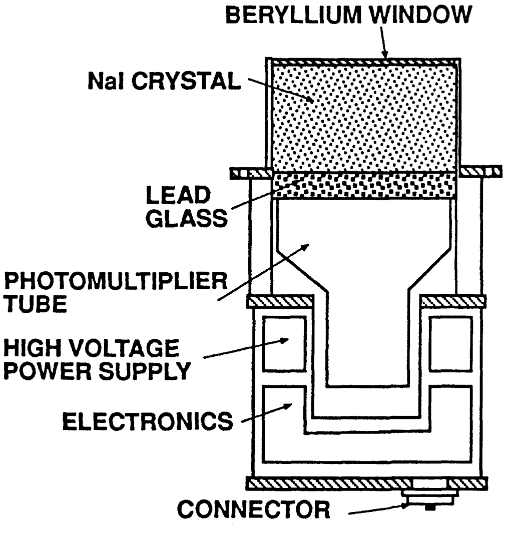

FIG. 13. The KONUS gamma-ray burst-detection instrument. A sodium iodide scintillator crystal is optically coupled to a photomultiplier tube and the assembly shielded to minimize response to energetic particles.

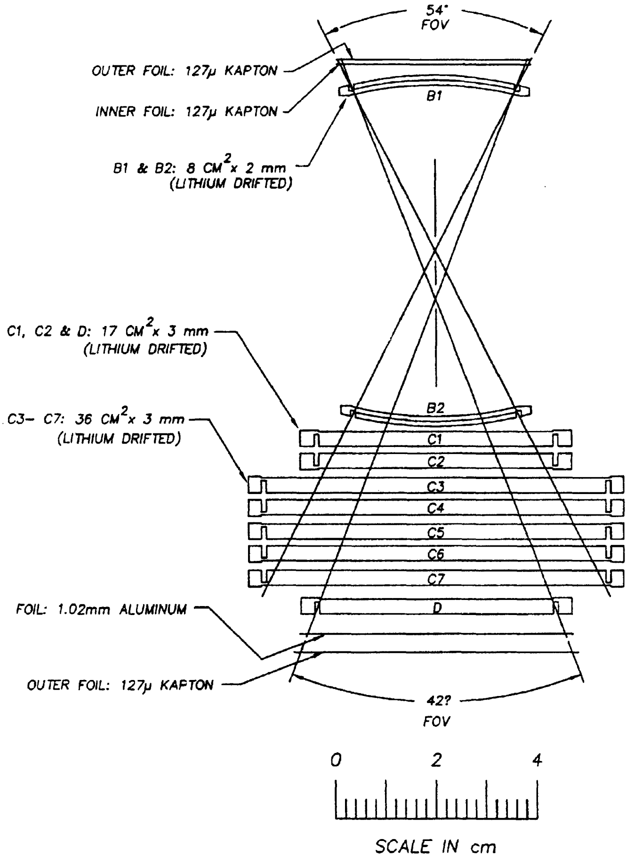
54° FOV
OUTER FOIL: 127μ KAPTON
INNER FOIL: 127μ KAPTON
B1 & B2: 8 CM²x 2 mm (LITHIUM DRIFTED)
B1
C1, C2 & D: 17 CM²x 3 mm (LITHIUM DRIFTED)
C3– C7: 36 CM²x 3 mm (LITHIUM DRIFTED)
B2
C1
C2
C3
C4
C5
C6
C7
D
FOIL: 1.02mm ALUMINUM
OUTER FOIL: 127μ KAPTON
42? FOV
0
2
4
SCALE IN cm

FIG. 14. The WIND spacecraft EPACT alpha-proton-electron (APE) solid-state detector telescope. Multiple solid-state detectors are arranged to measure the energy deposited in each by energetic particles traversing the telescope. Logic functions associated with the signals can determine if the particle has been completely stopped at a given detector.

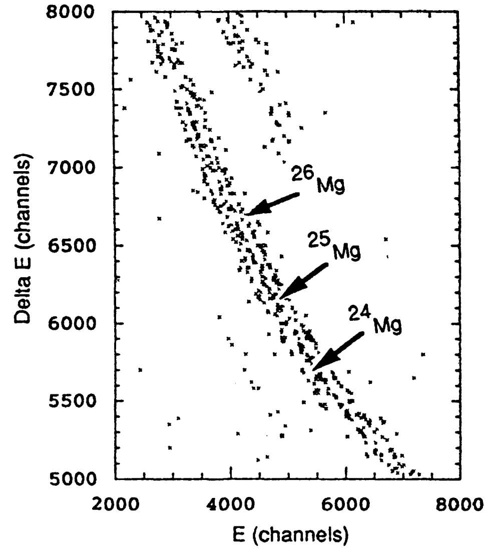

FIG. 15. Isotopic resolution of the EPACT IT solid-state detector telescope. Note the clear separation of Mg isotopes obtained with this arrangement.

sors, and programmable logic devices to reduce mass and power consumption that otherwise would preclude these instruments from being carried on spacecraft.

5. PARTICLE-IMAGING INSTRUMENTS

The development of two-dimensional particle and radiation detectors in the late 1970s and 1980s created a revolution in charged-particle instrumentation, which added to that already started by fast logic and microprocessor-based instrument systems. Up to that point, most particle detectors did not have position-sensing features capable of indicating the arrival point of the photon or particle in a two-dimensional array. The development of microchannel-plate electron multipliers, strip detectors and related technologies, solid-state detector arrays, charge-coupled devices (*q.v.*), etc. made possible a whole new class of instruments with particle-imaging capabilities, analogous to well-known optical instruments (albeit with poorer angular resolution). This feature is extremely useful when narrow spatial distributions of charged particles are being sampled quickly to obtain distribution functions time correlated to electric- and magnetic-field observations. The sequential scanning of narrow fields of view over large solid angles, previously required to obtain complete distribution-function information, introduced significant time aliasing in the plasma and particle measurements. The top-hat electrostatic analyzer described previously is a typical example of particle-imaging instrumentation. The annular microchannel plate placed at the exit of the hemispherical analyzer provides local electron multiplication for the arriving particles; a position-sensitive anode (resistive, segmented, or delay line) is used to determine the angular position of the electron cloud exiting the annular channel plate. The detected signals can then be sorted according to their angular position and energy, making possible a very fast sampling of the distribution function.

Stacks of position-sensitive solid-state detectors have also been used to implement telescopes that are able to produce 2D images of the arriving particles. For example, the SEPS detector of the CEPPAD instrument flown aboard the POLAR spacecraft uses six 16 × 16 telescope arrays coupled to 768 amplifier strings to obtain images of the source/loss cone associated with the trapping of energetic particles in the Earth's magnetosphere (Blake *et al.*, 1995). Each *xy* sensor has an area of 6.25 cm^2 and is divided into 256 individually read out pixels. This type of instrumentation has been made possible not only by advances in solid-state detector technology but also by the availability of very large-scale integration (VLSI) microelectronics devices, which integrate a large number of analog and digital functions into a single chip.

6. VISIBLE, UV, AND X-RAY IMAGING INSTRUMENTS

Imaging of the Earth's auroral regions has become a powerful tool to understand the global plasma processes that result from the excitation of the lower atmosphere by charged particles originating in the solar wind and geomagnetic tail. Imaging photometers at multiple wavelengths were flown initially on spinning spacecraft like Dynamics Explorer (Frank *et al.*, 1981), but the current designs utilize microchannel image intensifiers and two-dimensional charged-coupled-device (CCD) detector arrays to improve greatly the sensitivity and time resolution (see QUANTUM OPTOELECTRONIC DEVICES). Such visible, x-ray, and UV imaging instruments are being flown on the POLAR spacecraft, which is designed to spin at 10 rpm to optimize particle and field measurements. Therefore, to yield maximum sensitivity and resolution, the imaging instrumentation is mounted on a platform that is electronically de-spun to keep it pointed constantly toward the Earth's north pole. The optical design of auroral imagers must take into account the faintness of the emissions as opposed to the intense sunlit portion of the Earth that is visible from the high orbital altitudes (see GEOELECTROMAGNETISM; MAGNETOSPHERES). In the case of ultraviolet imagers, this is achieved with sharp-cutoff filters that block the reflected visible light. Visible imagers, on the other hand, must resort to restrictions of the field of view and offset pointing techniques. Most instruments incorporate multiple sensors and filters at selected wavelengths, which are used to map the electron energy influx into the auroral and polar ionospheres and to determine the precipitation patterns of protons from the geomagnetic tail and ring current, particularly during a geomagnetic storm when the morphology and extent of the auroral emissions change rapidly with time (Akasofu and Chapman, 1972).

An overview of the POLAR visible imager is shown in Fig. 16 (Frank *et al.*, 1995). Three sensor systems are implemented in this design. The low- and medium-resolution cameras share a common optical collimator providing a 20° by 20° field of view as well as primary optics, while the Earth camera has a separate and smaller collimator also providing a 20° × 20° field of view. The Earth camera is equipped with an ultraviolet filter and is used to guide the instantaneous field of view of the low- and medium-resolution cameras so that bright areas of the field of view are avoided, such as that corresponding to the sunlit Earth portion. Table 1 shows

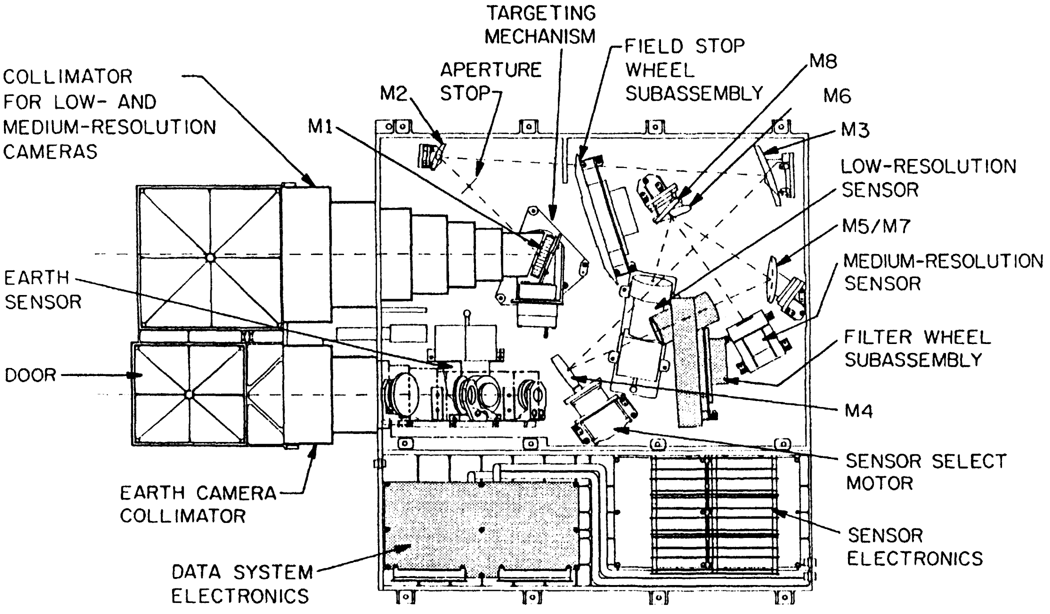
TARGETING MECHANISM
APERTURE STOP
M2
FIELD STOP WHEEL SUBASSEMBLY
M8
M6
M3
COLLIMATOR FOR LOW- AND MEDIUM-RESOLUTION CAMERAS
M1
LOW-RESOLUTION SENSOR
M5/M7
MEDIUM-RESOLUTION SENSOR
EARTH SENSOR
FILTER WHEEL SUBASSEMBLY
DOOR
M4
SENSOR SELECT MOTOR
EARTH CAMERA COLLIMATOR
SENSOR ELECTRONICS
DATA SYSTEM ELECTRONICS

FIG. 16. Top view of the visible imaging instrument for the POLAR spacecraft (VIS). Three sensor systems share two optical paths to provide fields of view of 20° × 20° and varying spatial resolution. Intensified charge-coupled devices operating at −40°C are used as imaging detectors.

the wavelengths of the passband filters used and their functions. They have been carefully selected to optimize observations of the various phenomena, described briefly in the table. Both atmospheric and space-physics objectives are listed (see ATMOSPHERIC PHYSICS; GEOPHYSICS). The image intensifier–charge-coupled-device (see CHARGE-COUPLED DEVICES) combination used in the instrument is shown in Fig. 17. It is heavily shielded against penetrating particles, and the CCD itself is operated at very low temperatures (−40°C) to minimize radiation-induced damage that would significantly degrade its performance and shorten its useful life.

The limited telemetry bandwidth available requires that the large volume of raw data generated by the imaging sensors be com-

Table 1. Passband filters for the low- and medium-resolution cameras.

Filter	λ_p, nm[a]	$\Delta\lambda$, nm	t_f	S, e kR^{-1} pixel^{-1} s^{-1}	Source	Function
1	360.1	2.94	3.9×10^{-5}	7.4×10^{-3}	Surface albedo	Background for ozone measurements
2	317.3	3.22	4.8×10^{-5}	7.3×10^{-3}	O_3	Total ozone
3	396.4	2.17	0.34	73	Continuum	391.4 background
4	531.0	1.13	0.47	59	Continuum	557.7 background
5	391.4	2.02	0.29	63	N_2^+ 1N(0—0)	Aurora
6	557.7	1.02	0.38	41	OI ^1D—^1S	Aurora
7	630.0	1.08	0.48	33	OI ^3P—^1D	Aurora
8	732.0	1.10	0.36	7.8	OII ^2D—^2P	Aurora
9	626.4	1.09	0.46	32	Continuum	630.0 background
10	656.3	1.14	0.40	33	Hα	Aurora (proton)
11	589.0	1.00	0.52	47	NaI ^2S—^2P	Atmospheric nightglow
12	308.5	5.60	0.12	17	OH	Comets, atmosphere

---

[a] Passbands include the decrease of ~0.1 nm for off-normal incidence on the filters.

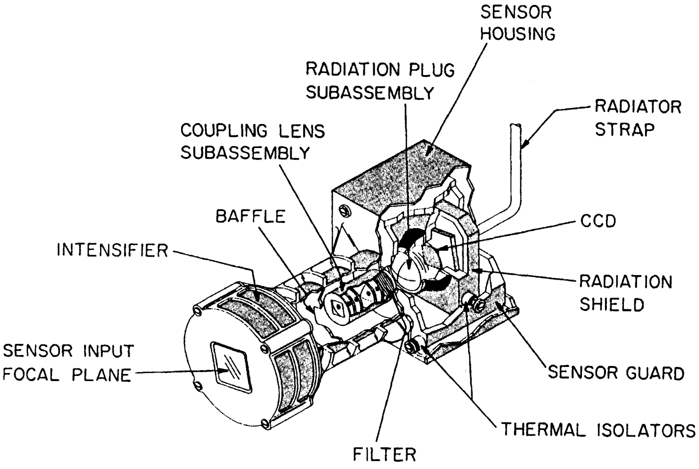
SENSOR HOUSING
RADIATION PLUG SUBASSEMBLY
COUPLING LENS SUBASSEMBLY
RADIATOR STRAP
BAFFLE
CCD
INTENSIFIER
RADIATION SHIELD
SENSOR INPUT FOCAL PLANE
SENSOR GUARD
THERMAL ISOLATORS
FILTER

FIG. 17. Image-intensifier–CCD combination used in visible and UV instruments on the POLAR spacecraft. Radiation shielding is required to operate the devices in the Earth's radiation belts with minimum interference and maximum performance and lifetime.

pressed with efficient algorithms such as discrete cosine transforms or mean–residual-vector quantization codes (see INFORMATION SCIENCE; COMPUTER SOFTWARE). Many classes of "lossless" algorithms are in use today to achieve compression ratios that may reach values as high as 10:1. However, significant problems can exist with data compression unless appropriate communications protocols and error detection and correction techniques are utilized.

The UV-imaging instrument flown on the same spacecraft (POLAR) shares some common design features with the visible imager, such as image-intensified CCD detector assemblies, mirror optics, and passband filters (Torr *et al.,* 1995), and is shown schematically in Fig. 18. However, the UV instrument is "solar blind," which facilitates the task of imaging the auroral emissions in the presence of strong visible light. Figure 19 illustrates the typical response of the UV instrument in the 1400–1600-Å wavelength region.

The bombardment of the atmosphere by energetic electrons also produces bremsstrahlung x rays, which can be sensed from orbital altitudes. The x ray imaging experiment flown on the POLAR spacecraft is shown in Fig. 20 and uses a multiple-pinhole camera to generate nonoverlapping images of the auroral regions (Imhof *et al.,* 1995). These are summed together during ground processing. The x-ray detector itself is a set of two multiwire position-sensing gas proportional counters located at the focal plane and filled with mixtures of argon, xenon, and carbon dioxide and is shown in Fig. 20. The energy range covered is 2–60 keV with a ±21° field of view. The number and size of pinholes used is selectable by means of a movable plate that adjusts the field of view and sensitivity. The multiwire anode is constructed as an *x-y* array, which allows the de-

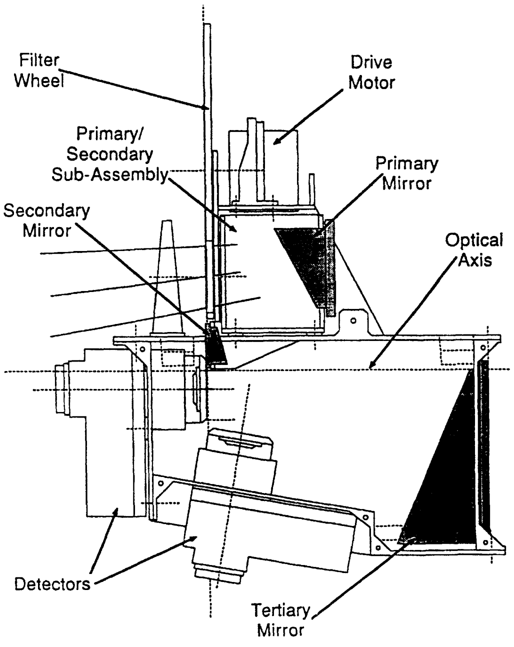
Filter Wheel
Drive Motor
Primary/ Secondary Sub-Assembly
Primary Mirror
Secondary Mirror
Optical Axis
Detectors
Tertiary Mirror

FIG. 18. The ultraviolet imaging instrument for the POLAR spacecraft (UVI). This instrument is "solar blind" and can image the entire auroral oval, including the sunlit portions of the same.

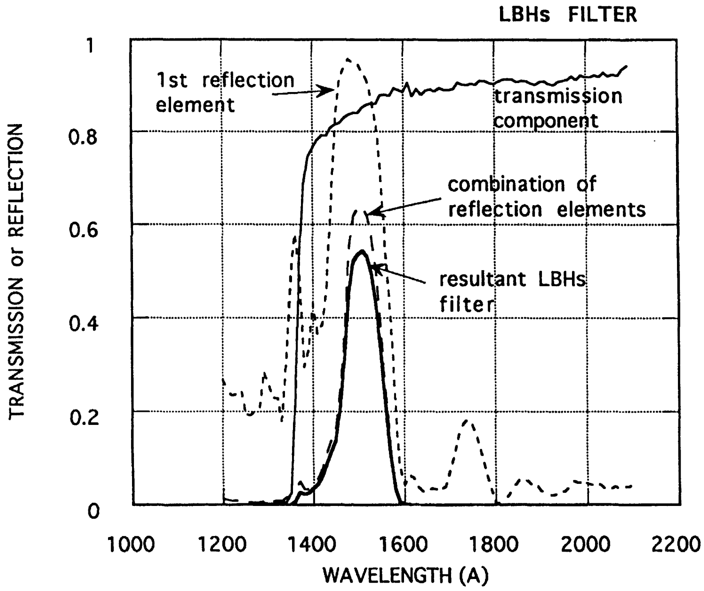

FIG. 19. Combined response of the POLAR UV instrument in the 1400–1600-Å wavelength region. Specially developed optics and multilayer filters were used to achieve this excellent narrow-band response.

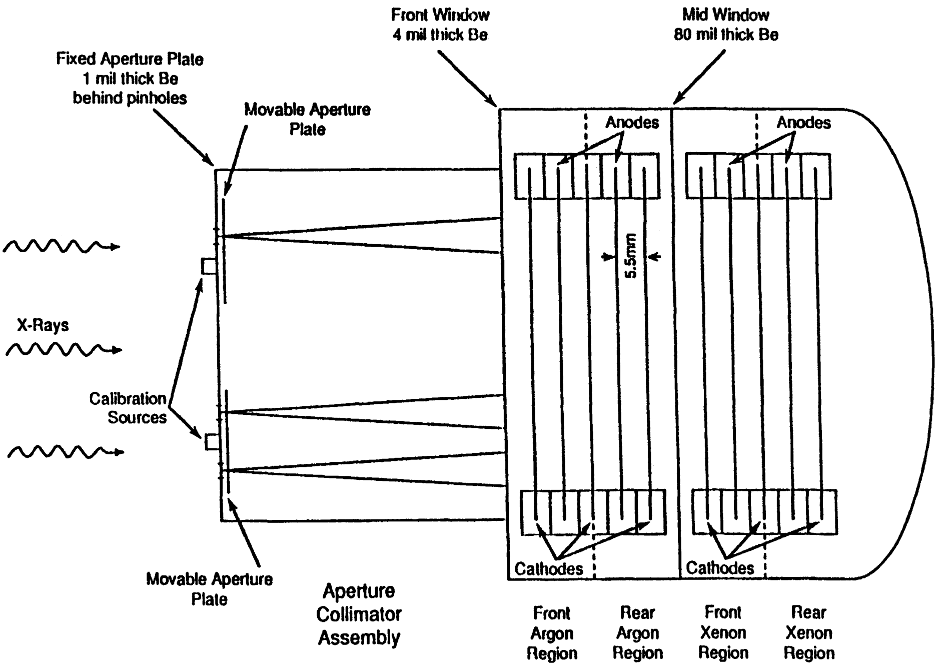
Front Window
4 mil thick Be
Mid Window
80 mil thick Be
Fixed Aperture Plate
1 mil thick Be
behind pinholes
Movable Aperture Plate
Anodes
Anodes
5.5mm
X-Rays
Calibration Sources
Movable Aperture Plate
Aperture Collimator Assembly
Cathodes
Cathodes
Front Argon Region
Rear Argon Region
Front Xenon Region
Rear Xenon Region

FIG. 20. The POLAR x-ray multiple pinhole imager (PIXIE). Two proportional counters with position-sensitive, multiwire readout systems are used, covering the range of 2–60 keV. A movable aperture plate provides “zoom” capabilities.

termination of the position of the local electron avalanche with a precision of 2 mm over the 18 × 18-cm^2 dimension of the detector. The position readout system is based upon measuring the ratio of the charge collected from two distributed cathodes in the gas chambers (Mathieson *et al.,* 1980).

7. SPECIAL SPACECRAFT REQUIREMENTS

The accommodation of space-physics instruments aboard spacecraft represents a significant systems engineering challenge, since many nontraditional requirements are usually invoked. We have already mentioned the problem of spacecraft-generated magnetic fields. In most missions the fields to be measured are extremely weak (<100 nT), and long deployable booms are required to place the sensors away from the main body of the spacecraft where many local sources of magnetic fields are located. In addition, the power generation and distribution system (solar array, batteries, wiring, etc.) must be designed to minimize the generation of magnetic fields by balancing and twisting power-distribution lines, using current-cancelling techniques in the solar array, and selecting batteries with minimum magnetic materials. The use of motors and relays, which incorporate permanent magnets in their design, must be carefully controlled. In sensitive missions, a "magnetics control program" is usually carried out to ensure that the sources of stray magnetism aboard the spacecraft are eliminated or minimized to the required levels (Harten and Clark, 1995).

A similar requirement is associated with the measurement of low-energy plasmas. The energies involved are typically less than a few tens of electron volts, and even minimal charging on the spacecraft surfaces due to ambient particles or photoemission will lead to significant errors in the measurements by deflection of particle trajectories prior to their arrival at instrument apertures. To minimize differential charging between surfaces, the "equipotential spacecraft" requirement was developed. All exposed areas are covered by conductive materials and thermal control paints to ensure that the potential difference between any two points of the spacecraft does not exceed 1 V under worst-case conditions (see Jursa, 1985).

For the electric-field, plasma, and radio-wave experiments, electromagnetic compatibility with spacecraft subsystems is absolutely necessary to minimize interference, particularly since these instruments tend to operate at very high sensitivity. Many spacecraft subsystems generate signals, both conducted and radiated, that can fall within the passbands of these receivers. To minimize these effects it is common practice to synchronize the operating frequency of all power converters on the spacecraft to subharmonics of a master clock. The narrow-band receivers are then designed to avoid these frequencies in their passbands. Wide-band receivers are particularly sensitive to interference since frequency selection and discrimination is not possible. Just like in the magnetics case, an "electromagnetic control plan" is usually implemented to identify, control, and eliminate interfering signals.

Most space-physics instruments must operate in moderate to severe radiation environments such as the belts of trapped energetic particles encountered around magnetized planets. The radiation environment is much more benign, on the average, in the interplanetary medium on account of the absence of trapped radiation. However, solar flares can inject large fluxes of energetic protons, which can interfere with and cause significant damage to spacecraft subsystems and instruments. These are generally known as "solar particle events," and operations may become impossible during a large event. Thus, space-physics instrumentation must be designed to withstand and operate in these environments. High-inclination spacecraft orbits that cut through the Earth's radiation belts may result in a total dose exposure approaching 100 krad or more per year. In addition, energetic particles can create "single-event upsets" (SEU's) and even latch-up in microprocessors, memories, and logic circuits, seriously affecting their operation and eventually causing permanent damage if precautions are not taken (see Jursa, 1985). The design of fault-tolerant systems and software to operate in the space environment with autonomous recovery is an art form in the very early stages of development. Component and system reliability considerations, in addition to the effects already described, add another difficult dimension to the development of space-based instruments. Once the spacecraft are launched, repair or adjustment of the instruments is generally impossible. Finally, most spacecraft today operate in the "store and dump" mode to minimize tracking and data-acquisition costs. This implies that real-time contact with the instruments is infrequent, and sufficient intelligence must be designed into them to perform autonomously a variety of functions, including reconfiguration into a safe condition in the case of unexpected anomalies. The combination of these difficult requirements generally results in very high development and operation costs being associated with space-based instruments.

8. CALIBRATION OF SPACE-PHYSICS INSTRUMENTATION

The complete calibration of this class of instruments is an extensive undertaking that requires highly specialized personnel and institutional facilities such as large magnetic and electromagnetic test environments (where the ambient-field magnitude and direction can be controlled accurately), particle accelerators (as available at national laboratories), radiation sources, large vacuum chambers, and low-energy plasma sources; known composition, charge-state, and energy ion beams; electron sources, etc. The full characterization of an instrument's response may take many months and require several trips to unique facilities located in the United States and around the world. Computer simulations of instrument response functions are becoming widely used since they permit the modeling of instrument responses that may not be easily measured experimentally, or those associated with on-orbit anomalies and recovery (see LABORATORY INSTRUMENTATION; ELECTROMAGNETIC QUANTITIES, MEASUREMENT OF; ACCELERATORS, LINEAR; PARTICLE SOURCES—ION; PHYSICS AND TECHNOLOGY OF ION AND ELECTRON SOURCES; SOURCES, PARTICLE—ELECTRON; TARGETS FOR PARTICLE ACCELERATORS; INTENSE PARTICLE BEAMS; PLASMA DEVICES; SPARKS, ARCS, AND OTHER ELECTRIC DISCHARGES).

Many instruments include in their design limited internal calibration devices, which can be periodically activated to verify performance. For example, magnetometers include precision current sources that can be used to generate magnetic fields of known direction and magnitude at the sensor. Similar devices are used with electric field detectors, while small radioactive sources or UV sources may be incorporated in charged-particle and thermal-plasma detectors. Intercalibration of instruments is usually accomplished during the data analysis phase by verifying the consistency of the results. Imaging instruments may use known objects (e.g., the Moon) for photometric calibrations and sensitivity verification.

At the spacecraft system level, extensive instrument compatibility testing is carried out to verify that instruments and spacecraft do not interact or interfere with each other or that undesirable emissions that could interfere with normal operations are not generated (Harten and Clark, 1995).

9. DATA HANDLING, REDUCTION, AND ANALYSIS

As pointed out in the preceding sections, the limitations on the bandwidth available for the transmission of data to the ground through the spacecraft telemetry system impose strong design constraints on the onboard and ground data-processing algorithms used. The handling of errors in the communication channel becomes a significant issue with highly compressed data streams, since errors tend to destroy large blocks of data. The ground system must decode the telemetry stream for each instrument without introducing additional errors and format the data structures for conversion to physical parameters using appropriate, user-generated algorithms. In contrast to *event-driven* data acquisition, such as is typical of remote-sensing instruments, space-physics instruments collect data continuously and generate extensive time series of physical parameters associated with the observations. Since the spatial and temporal scales used in scientific studies can range over several orders of magnitude, a multiplicity of specialized data sets needs to be produced, accessed, and maintained over long periods of time. Real-time data are usually not required for these studies, except in those cases where monitoring of the space environment for predictive purposes and cause–effect determinations are primary objectives (e.g., "space weather" and related activities). Browsing through "key parameter" summary data sets has become an efficient way of selecting time intervals or events for further study (Mish *et al.,* 1995). These low–time-resolution key parameter data sets are generated by a central facility using approximate algorithms and predictive ancillary data (such as spacecraft orbits) for speed and efficiency. The primary emphasis is on the rapid development of an intelligent index or catalog to the detailed data that can be easily accessed and will allow the identification of significant events or periods. The use of electronic networks and distributed data bases has become commonplace for instrument operation as well as data-reduction and -analysis activities (see INFORMATION SCIENCE; COMPUTERS; COMPUTER NETWORKS; PHYSICS APPLICATIONS OF COMPUTERS—EXPERIMENTAL; COMPUTER SOFTWARE).

The maturity of the space-physics field presents a significant challenge as compared to related disciplines that are just starting their "discovery" phase. Space-physics instrumentation tends to be extremely sophisticated, addressing highly specialized problems that require significant resources for their understanding. The accurate separation of true observations from signals generated by the spacecraft itself can become a serious obstacle and require a major effort for its resolution, particularly for weak-field and low-energy plasma observations. Just a few days of high-resolution data during events related to magnetospheric boundary crossings or solar activity may require several years of analysis for its full understanding. Advanced data visualization techniques including complex analytical simulations are being used to expedite the process and complement the experimental results.

GLOSSARY

Aliasing: Incorrect identification of the temporal variation characteristics of a phenomenon by insufficient time sampling of observations.

Aurora: Lights in the upper atmosphere (90 to 300 km altitude) produced by the excitation of atmospheric gases by energetic particles; the localized areas in the two polar regions where aurorae are typically observed are determined by the detailed topology of the magnetosphere.

Auroral Oval: As seen from space, the auroral emissions occupy a relatively narrow oval extending around the magnetic poles, mapping the magnetic field lines along which particles have access to the atmosphere. During magnetic storms the oval widens significantly and extends to lower latitudes, developing bright emission enhancements.

Bremsstrahlung: Radiation emitted by accelerated or decelerated charged particle, for example, x rays.

Carbon Foil: A foil of carbon with a mass-to-area ratio of a few micrograms per square centimeter.

Coronal Mass Ejection (CME): Emission of a large plasma (gas) cloud by the Sun, usually accompanied by x rays and energetic particles.

Data-Compression Algorithm: A digital algorithm that reduces redundancy in data to make possible the use of lower data rates for transmission.

Data Rate: The speed at which information is transmitted or processed by a digital data telemetry system.

Distribution Function and Moments: A mathematical function describing the number, velocity, and angular direction of charged particles in a plasma. Successive integrations of this function allow the computation of the plasma density, momentum, temperature, heat flux, etc.

Electromagnetic Compatibility: The techniques used to ensure that different components and instruments aboard a vehicle or spacecraft do not interfere with each other in an electromagnetic sense.

Geomagnetic Storm: Interval of 1–3 h of auroral, geomagnetic, and magnetospheric activity, usually followed by a several-hour period of relative quiesence. Full recovery of the geomagnetic field to prestorm values may take several days.

Geomagnetic Tail: A region of the magnetosphere in the antisunward direction; if visible, the magnetotail would have characteristics like the tail of a comet.

Helioseismology: The study of solar interior oscillation modes whose periods may range from seconds to years.

Ionosphere: Ionized region of the upper atmosphere, generally from about 90 km to about 1000 km in altitude; the area beyond this is defined as the magnetosphere.

Ionosonde: A pulse transmitter-receiver system used to "sound" the ionosphere, usually from below, through reflected echoes that can be used to determine the density of ionization, altitude, and profile of the different regions.

Key Parameter Data Set: A rapidly produced, low–time-resolution (one to several minutes) data set of key geophysical parameters or measurements, used to provide a qualitative overview and summary of observations. Usually, these sets are used as catalogs or indices to select intervals or events for further study.

Latch-up: The term applied to describe an uncontrollable avalanche condition in semiconductor devices caused by interaction with energetic particles, leading to high localized currents that can destroy the device.

Loss Cone: A narrow conical angle defined around a magnetic field line through which trapped particles in the radiation belts can reach the atmosphere and be lost.

Magnetosphere: Region of space around the Earth and other magnetized planets in which the planetary magnetic and electric fields usually dominate the transport and motion of charged particles.

Microchannel Plate: A two-dimensional arrangement of thousands of microscopic cylindrical electron multipliers bonded together to form a plate. Particles striking the entrance plate surface produce an amplified electron cloud at the same location on the exit surface. Multiple anode configurations make possible the determination of the location of arrival of each particle.

Overhauser Effect: Magnetic polarization of a liquid sample rich in protons to which free radicals have been added.

Plasma: Gas (atomic and/or molecular species) that is ionized. Naturally occurring plasmas usually contain magnetic fields.

Plasma Temperature and Heat Flux: Plasma temperature is defined as the average energy per particle. The heat flux is the energy flux carried by a flowing plasma.

Plasma Waves: Electric, magnetic, and electrostatic waves that can exist in an ionized medium.

Power Converter: Electronic device used to transform the dc current from the spacecraft primary power system (solar array and/or batteries; distribution bus) into a high-frequency ac voltage that is then stepped up or down with transformers, rectified, filtered, and regulated to obtain multiple dc voltages at its output.

Quadrispherical Analyzer: An electrostatic analyzer with deflectors that are concentric quarter spheres.

Radiation Belts: Localized regions in the magnetosphere that contain energetic charged particles whose motions are primarily controlled by the magnetic field of the planet.

Radiation Dose: The energy deposited into a medium by penetrating energetic particles.

Single-Event Upset: An isolated, nondestructive change in the state of semiconductor memories and other logic devices caused by the interaction of an energetic charged particle with the semiconductor, usually in the form of local ionization and charge trapping.

Solar Flare: Sudden brightening, for several minutes to several hours, of a small area of the solar photosphere that contains a group of sunspots.

Solar-Particle Events: Transient events associated with solar phenomena (such as flares and coronal mass ejections), which can inject very high-energy charged particles into the interplanetary medium representing a significant hazard to spacecraft and human spaceflight.

Solar Wind: Supersonic expansion of the solar corona, primarily hydrogen ions, into the interplanetary medium.

Space Weather: A contemporary term, analogous to atmospheric weather, used to describe the state and temporal variations of the Sun–Earth interactions through the interplanetary medium.

Store-and-Dump Mode: A method to acquire data from spacecraft using onboard storage rather than continuous transmission. The data are collected, stored on board for extended periods of time on tape recorders or solid-state memories, and then "dumped" to the ground at high speed when stations are available or in view of the spacecraft.

Telemetry System: Data-acquisition and -processing systems, usually digital, used by vehicles and spacecraft to transmit measured data to the ground (i.e., "measurements at a distance") for processing and analysis.

Thermal Plasma: A plasma that is approximately in thermal equilibrium.

Time-Domain Sampler: A fast sampling system for electrical and magnetic signals used to transmit wave-form information rather than spectral content.

Time-of-Flight Analyzer: An instrument that measures the velocity of particles by measuring the time they take to cover a given small linear distance.

Wave-Particle Correlator: Instrument designed to perform correlation functions in real time between charged-particle and electromagnetic-field data.

Wien or Velocity Filter: An arrangement of orthogonal electric and magnetic fields used to select particles with a given mass-to-charge ratio.

Works Cited

Acuña, M. H., Connerney, J. E. P., Wasilewski, P., Lin, R. P., Anderson, K. A., Carlson, C. W., McFadden, J., Curtis, D. W., Rème, H., Cros, A., Médale, J. L., Sauvaud, J. A., d'Uston, C., Bauer, S. J., Coutier, P., Mayhew, M., Ness, N. F. (1992), "Mars Observer Magnetic Fields Investigation," *J. Geophys. Res.* **97** (E5), 7799–7814.

Acuña, M. H., Ogilvie, K. W., Baker, D. N., Curtis, S. A., Fairfield, D. H., Mish, W. H. (1995), "The Global Geospace Science Program and its Investigations," *Space Sci. Rev.* **71,** 5–21.

Akasofu, S. I., Chapman, S. (1972), *Solar Terrestrial Physics,* Oxford: Oxford University Press.

Al'pert, Ya. L. (1990), *Space Plasma,* 2 Vols., Cambridge, U.K.: Cambridge University Press.

Aptekar, R. L., Frederiks, D. D., Golenetskii, S. V., Ilynskii, V. N., Mazets, E. P., Panov, V. N., Sokolova, Z. J., Terekhov, M. M., Sheshin, L. O., Cline, T. L., Stilwell, D. E. (1995), "KONUS-W: Gamma Ray Burst Experiment for the GGS WIND Spacecraft," *Space Sci. Rev.* **71,** 265–272.

Behannon, K., Acuña, M. H., Burlaga, L. F., Lepping, R. P., Ness, N. F., Neubauer, F. M. (1977), "Magnetic Field Experiment for Voyagers 1 and 2," *Space Sci. Rev.* **21,** 235–257.

Blake, J. B., Fennell, J. F., Friesen, L. M., Johnson, B. M., Kolasinski, W. A., Mabry, D. J., Osborn, J. V., Penzin, S. H., Schnauss, E. R., Spence, H. C., Baker, D. N., Belian, R., Fritz, T. A., Ford, W., Laubscher, B., Stiglich, R., Baraze, R. A., Hilsenrath, M. F., Imhof, W. L., Kilner, J. R., Mobilia, J., Voss, D. H., Korth, A., Güll, M., Fisher, K., Grande, M., Hall, D. (1995), "CEPPAD," *Space Sci. Rev.* **71,** 531–562.

Bougeret, J. L., Kaiser, M. L., Kellogg, P. J., Manning, R., Goetz, K., Monson, S. J., Monge, E., Friel, L., Meetre, C. A., Perche, C., Sitruk, L., Hoang, S. (1995), "WAVES: The Radio and Plasma Wave Investigations on the WIND Spacecraft," *Space Sci. Rev.* **71,** 231–263.

Brekke, A., Egeland, A. (1983), *The Northern Lights,* Heidelberg: Springer-Verlag.

Carignan, G. R., Block, B. P., Maurer, J. C., Hedin, A. E., Reber, C. A., Spencer, N. A. (1981), "The Neutral Mass Spectrometer on Dynamics Explorer B," *Space Sci. Instrumentation* **5,** 429–441.

Frank, L. A., Craven, J. D., Ackerson, K. L., English, M. R., Eather, R. H., Carovillano, R. L. (1981), "Global Auroral Instrumentation for the Dynamics Explorer Mission," *Space Sci. Instrumentation* **5,** 369–393.

Frank, L. A., Sigwarth, J. B., Craven, J. D., Cravens, J. P., Dolan, J. S., Dvorsky, M. R., Hardebeck, P. K., Harvey, J. D., Muller, D. W. (1995), "The Visible Imaging System (VIS) for the POLAR Spacecraft," *Space Sci. Rev.* **71,** 297–328.

Gloeckler, G., Balsiger, H., Bürgi, A., Bochsler, P., Fisk, L. A., Galvin, A. B., Geiss, J., Gliem, F., Hamilton, D. C., Holtzer, T. E., Hovestadt, D., Ipavich, F. M., Kirsch, E., Lundgren, R. A., Ogilvie, K. W., Sheldon, R. B., Wilken, B. (1995), "The Solar Wind and Suprathermal Ion Composition Investigation on the WIND Spacecraft," *Space Sci. Rev.* **71,** 79–124.

Gough, M. P. (1985), "Spacecraft Particle Correlator Experiment (SPACE) for AMPTE UKS," *IEEE Trans. Geosci. Remote Sens.,* **GE-23** (3), 305–310.

Gurnett, D. A., Persoon, A. M., Randall, R. F., Odem, D. L., Remington, S. L., Averkamp, T. F., Debower, M. M., Hospodarsky, G. B., Huff, R. L., Kirchner, D. L., Mitchell, M. A., Pham, B. T., Phillips, J. R., Schintler, W. J., Sheyko, P., Tomash, D. R. (1995), "The POLAR Plasma Wave Instrument," *Space Sci. Rev.* **71,** 597–622.

Hargreaves, J. K. (1979), *The Upper Atmosphere and Solar-Terrestrial Relations,* New York: Van Nostrand Reinhold.

Harten, R., Clark, K. (1995), "The Design Features of the GGS WIND and POLAR Spacecraft," *Space Sci. Rev.* **71,** 23–40.

Harvey, P., Mozer, F. S., Pankow, D., Wygant, J., Maynard, N. C., Singer, H., Sullivan, W., Anderson, P. B., Pfaff, R., Aggson, T., Pedersen, A., Fälthammar, G., Tanskannen, P. (1995), "The Electric Field Instrument in the POLAR Spacecraft," *Space Sci. Rev.* **71,** 583–596.

Imhof, W. L., Spear, K. A., Hamilton, J. W., Higgins, B. R., Murphy, M. J., Pronko, J. G., Vondrak, B. R., McKenzie, D. L., Rice, C. J., Gorney, D. J., Roux, D. A., Williams, R. L., Berg, J. A., Lutz, L., Detrick, D. (1995), "The POLAR Ionospheric X-Ray Imaging Experiment (PIXIE)," *Space Sci. Rev.* **71,** 385–408.

Jursa, A. S. (Ed.) (1985), *Handbook of Geophysics and the Space Environment,* Air Force Geophysics Laboratory, Air Force Systems Command, ADA 167000.

Kennel, C. F., Lanzerotti, L. J., Parker, E. N. (Eds.) (1979), *Solar System Plasma Physics,* 3 Vols., Amsterdam: North Holland.

Knoll, G. K. (1990), *Radiation Detection and Measurement,* 2nd ed., New York: Wiley.

Krehbiel, J. P., Brace, L. H., Theis, R. F., Pinkus, W. H., Kaplan, R. B. (1981), "The Dynamics Explorer Langmuir Probe Instrument," *Space Sci. Instrumentation* **5,** 493-502.

Lepping, R. P., Acuña, M. H., Burlaga, L. F., Farrell, W. M., Slavin, J. A., Schatten, K. H., Mariani, F., Ness, N. F., Neubauer, F. M., Whang, Y. C., Byrnes, J. B., Kennon, R. S., Panetta, P. V., Scheifele, J., Worley, E. M. (1995), "The WIND Magnetic Field Investigation," *Space Sci. Rev.* **71,** 207–229.

Lin, R. P., Anderson, K. A., Ashford, S., Carlson, C., Curtis, D., Ergun, R., Larson, D., McFadden, J., McCarthy, M., Parks, G. K., Rême, H., Bosqued, J. M., Coutelier, J., Cotin, F., D'Uston, C., Wenzel, K.-P., Sanderson, T. R., Henrion, J., Ronnet, J. C., Paschmann, G. (1995), "A Three-Dimensional Plasma and Energetic Particle Investigation for the WIND Spacecraft," *Space Sci. Rev.* **71,** 125–153.

Mathieson, E., Smith, G. C., Gilvin, P. G. (1980), *Nucl. Instrum. Meth.* **174,** 221.

Maynard, N. C., Bielecki, E. A., Burdick, H. F. (1981), "Instrumentation for Vector Electric Field Measurement from DE-B," *Space Sci. Instrumentation* **5,** 523–534.

Mish, W. H., Green, J. L., Reph, M. G., Peredo, M. (1995), "ISTP Science Data Systems and Products," *Space Sci. Rev.* **71,** 815–878.

Ness, N. F. (1970), "Magnetometers for Space Research," *Space Sci. Rev.* **11,** 459–554.

Ogilvie, K. W., Chornay, D. J., Fitzenreiter, R. J., Hunsaker, F., Keller, J., Lobell, J., Miller, G., Scudder, J. D., Sittler Jr., E. C., Torbert, R. B., Bodet, D., Needell, G., Lazarus, A. J., Steinberg, J. T., Tappan, J. H., Mavretic, A., Gergin, E. (1995), "SWE: A Comprehensive Plasma Instrument for the WIND Spacecraft," *Space Sci. Rev.* **71,** 55–77.

Parks, G. K. (1991), *Physics of Space Plasmas,* Reading, MA: Addison-Wesley.

Paschmann, G., Loidl, H., Obermeyer, P., Ertl, M., Laborenz, R., Sckokpe, N., Baumjohann, W., Carlson, C. W., Curtis, D. W., (1985), "The Plasma Instrument for the AMPTE IRM," *IEEE Trans. Geosci. Remote Sens.* **GE-23** (3), 262–266.

Paschmann, G., Melzner, F., Haerendel, G., Bauer, O. H., Baumjohann, Sckopke, N., Treumann, R., McIlwain, C. E., Fillius, W., Whipple, E. C., Torbert, R. C., Quinn, J. M. (1993), "The Electron Drift Instrument for CLUSTER," European Space Agency Publication ESA SP-1159, 115–132.

Russell, C. T., Snare, R. C., Means, J. D., Pierce, D., Dearborn, D., Larson, M., Barr, G., Le, G. (1995), "The GGS/Polar Magnetic Fields Investigation," *Space Sci. Rev.* **71,** 563–582.

Shelley, E. G., Ghielmetti, A. G., Balsiger, H., Black, R. K., Bowles, J. A., Bowman, R. P., Bratschi, O., Burch, J. L., Carlson, C. W., Coker, A. J., Drake, J. F., Fischer, J., Geiss, J., Johnstone, A., Kloza, D. L., Lennartsson, O. W., Magoncelli, A. L., Paschmann, G., Peterson, W. K., Rosenbauer, H., Sanders, T. C., Steinacher, M., Walton, D. M., Whalen, B. A., Young, D. T. (1995), "The Toroidal Imaging Mass-Angle Spectrograph (TIMAS) for the POLAR Mission," *Space Sci. Rev.* **71,** 497–530.

Torr, M. R., Torr, D. G., Zukic, M., Johnson, R. B., Ajello, J., Banks, P., Clark, K., Cole, K., Keffer, C., Parks, G., Tsurutani, B., Spann, J. (1995), "A Far Ultraviolet Imager for the International Solar-Terrestrial Physics Mission," *Space Sci. Rev.* **71,** 329–383.

Von Rosenvinge, T., Barbier, L. M., Karsch, J., Liberman, R., Madden, M. P., Nolan, T., Reames, D. V., Ryan, L., Singh, S., Trexel, H., Winkert, G., Mason, G. M., Hamilton, D. C., Walpole, P. (1995), "The Energetic Particles, Acceleration, Composition and Transport (EPACT) Investigation on the WIND Spacecraft," *Space Sci. Rev.* **71,** 155–206.

Further Reading

Acuña, M. H., Ousley, G. W., McEntire, R. W., Bryant, D., Paschmann, G. (1985), "AMPTE: Mission Overview," *IEEE Trans. Geosci. Remote Sens.* **GE-23** (3).

Baker, D. N., Mason, G. M., Figueroa, O., Colón, G., Watzin, J. G., Alemán, R. M. (1993), "An Overview of the Solar, Anomalous, and Magnetospheric Particle Explorer," *IEEE Trans. Geosci. Remote Sens.* **31** (3).

Carovillano, R. L., Forbes, J. M. (1983), *Solar-Terrestrial Physics,* Dordrecht, The Netherlands: D. Reidel Publishing Co.

Coplan, M. A., Davis, C. C., Moore, J. H. (1989), *Building Scientific Apparatus,* Reading, MA: Addison-Wesley.

Hoffman, R. A. (Ed.) (1981), "Dynamics Explorer," *Space Sci. Instrumentation* **5** (4).

Kivelson, M. G., Russell, C. T. (1995), *Introduction to Space Physics,* Cambridge: Cambridge University Press.

Ogilvie, K. W., Durney, A., Von Rosenvinge, T. (1978), "Description of Experimental Investigations for the ISEE Spacecraft," *IEEE Trans. Geosci. Electronics* **GE-16** (3).

Russell, C. T. (Ed.) (1995), "The Global Geospace Mission," *Space Sci. Rev.* **71** (1–4).

SPARKS, ARCS, AND OTHER ELECTRIC DISCHARGES

Uwe H. Bauder and Hans Peter Schmidt, *Department of Electrical Engineering and Information Science, Technical University of Munich, Germany*

	Introduction	157
1.	**Introduction to Elementary Collision Theory**	158
2.	**Thermodynamic Equilibrium**	159
2.1	Criteria for the Establishment of LTE	159
2.2	Selected Properties of LTE and PLTE Plasmas	160
3.	**The Electrical Arc Discharge**	161
3.1	High-Pressure Arcs	162
3.1.1	Wall-Stabilized Arcs	162
3.1.2	Flow-Stabilized Arcs	165
3.1.3	Arc Jets	171
3.2	Low-Pressure and Vacuum Arcs	172
4.	**Sparks**	173
5.	**Glow Discharges**	174
5.1	Mechanism and Electrical Characteristics	174
5.2	Hollow-Cathode Discharges	176
5.3	Pseudosparks	177
6.	**High-Frequency Discharges**	178
6.1	Microwave Discharges	179
6.1.1	Electron-Cyclotron Resonance Discharges	179
6.2	Radio-Frequency Discharges	180
6.3	Dielectric-Barrier Discharges	182
7.	**Corona Discharges**	183
8.	**Lightning Phenomena**	183
	Glossary	185
	Works Cited	185
	Further Reading	186

INTRODUCTION

By the term "discharge," usually the physical phenomena associated with the interaction of an electric or electromagnetic field with a volume of gas are described. Since the conditions under which such an interaction takes place vary over a wide range of gas pressures and compositions, temperatures, frequencies of the electromagnetic field, and other factors, such as the presence or absence of electrodes, a generally accepted classification is missing. Electrical discharges may be of natural origin, such as lightning or Saint Elmo's fire, or manmade discharges may be tailored for specific applications. In each case the discharge contains charged particles, such as electrons and ions; frequently, also neutral particles are present. The composition of the particle mixture is often such that it contains equal numbers of different charges. In this case one speaks of "quasineutrality." In general, gas mixtures containing charged particles are called "plasma." In recent years the field of applications has increased dramatically. While the use of electrical arc discharges for the generation of light was one of the earliest applications, today's discharges may still be used for the generation of radiation, but now the range of applications is best described by the term "plasma chemistry and technology." Some of the most important potential applications of discharges such as controlled fusion power generation are described elsewhere in the *Encyclopedia of Applied Physics* (see Fusion, Inertial Confinement; Fusion, Magnetic Confinement; Plasma Devices; Plasma Etching). In the present article applications of discharges are covered after a discussion of the physical phenomena describing the type of discharge in question. Considering the technical applications of the discharges leads to a classification either by the temperature range realized in the discharge plasma or by the gas pressure in which the discharge is operated. Using the temperature as

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

the classifying parameter, one may speak of "high-temperature" and "low-temperature" plasma, the borderline separating the two areas not being well defined. We shall use a classification by pressure: It is usual to call discharges in gases at atmospheric pressure or above "high-pressure discharges."

1. INTRODUCTION TO ELEMENTARY COLLISION THEORY

Let us consider an elastic two-body collision. Particle A with mass m_A and the initial velocity $\mathbf{v}_A$ collides with a particle $B(m_B$ and $\mathbf{v}_B)$; after the collision the velocities are $\mathbf{v}'_A$ and $\mathbf{v}'_B$.

The conservation of momentum requires

$$m_A\mathbf{v}_A + m_B\mathbf{v}_B = m_A\mathbf{v}'_A + m_B\mathbf{v}'_B = m_0\mathbf{v}_S = \text{const} \tag{1}$$

with $m_0 = m_A + m_B$, $\mathbf{v}_S$ being the velocity of the center of gravity of the two particles. The conservation of energy reads:

$$\tfrac{1}{2}m_A\mathbf{v}_A^2 + \tfrac{1}{2}m_B\mathbf{v}_B^2 = \tfrac{1}{2}m_A\mathbf{v}_A'^2 + \tfrac{1}{2}m_B\mathbf{v}_B'^2 + Q = \text{const}; \tag{2}$$

$Q = 0$ characterizes the elastic collision. By introducing the relative velocities before and after the collision, $\mathbf{g} = \mathbf{v}_A - \mathbf{v}_B$ and $\mathbf{g}' = \mathbf{v}'_A - \mathbf{v}'_B$ (χ is the angle between $\mathbf{g}$ and $\mathbf{g}'$), it can be shown easily that $|\mathbf{g}| = |\mathbf{g}'|$.

By estimating the relative kinetic energy (W) loss of a particle (A) with mass m_A colliding elastically with a second particle (B), a first conclusion regarding the establishment of thermal equilibrium between electrons and heavy particles may be drawn from these elementary considerations:

$$\frac{W_A - W'_A}{W_A} = \frac{\Delta W_A}{W_A} = 2\,\frac{m_A m_B}{(m_A + m_B)^2}(1 - \cos\chi). \tag{3}$$

Considering the difference in mass between electrons and heavy particles in the plasma, $m_e \approx 10^{-30}$ kg and $m_{\text{heavy}} > 10^{-27}$ kg, it is evident that it takes a very large number of collisions before thermal equilibrium between these two species can be established.

The probability of a collision is best described by the collision cross section. The differential collision cross section is defined by

$$S_{12}(g) = \frac{dN_1(\theta,\phi)/dt\,d\Omega}{\Phi_1}, \tag{4}$$

where $dN_1(\sigma,\varphi)/dt\,d\Omega$ is the number of test particles ("1") traveling into the solid angle $d\Omega$ in the direction (σ,φ) after interacting with one field particle ("2") at rest, given that the incident flux density of test particles with velocity $\mathbf{g}$ is Φ. Then the total cross section is $\sigma_{12} = \int dS_{12} = \int (dS_{12}/d\Omega)d\Omega$, and the momentum transfer cross section needed for kinetic theory is given by

$$\sigma_{12}^{(1)}(g) = \int_{4\pi} (1 - \cos\chi)dS_{12}. \tag{5}$$

Leaving the microscopic particle model, it is advantageous to work with averages. Let $f_j(\mathbf{r},\mathbf{v},t)$ be the distribution function of particles of type j in the six-dimensional phase space, the μ space. The number of j particles present in the volume element of μ space at a given time is then given by

$$d^6N_j = f_j(\mathbf{r},\mathbf{v},t)d^3\mathbf{v}d^3r. \tag{6}$$

For binary collisions the kinetic equation describing the variation of the distribution function with time is called the Boltzmann equation:

$$\frac{\partial f_j}{\partial t} + \mathbf{v}\cdot\nabla_r f_j + \frac{\mathbf{F}_j}{m_j}\cdot\nabla_v f_j = \left(\frac{\partial f_j}{\partial t}\right)_{\text{coll}}. \tag{7}$$

The $\mathbf{F}_j$ in Eq. (7) represents external forces acting on the particles, while the term on the right-hand side is the so-called collision term. By multiplying the Boltzmann equation (7) with particle properties depending on $\mathbf{r}$, $\mathbf{v}_j$, t such as mass, momentum, and energy and integrating over the velocity space, the so-called Enskog equation follows. It is a general balance equation from which the specific conservation equations for mass, momentum (Navier–Stokes equation), and energy may be deduced. Evaluating the collision terms in Eq. (7) then allows the determination of the transport properties of a plasma: electrical and thermal conductivities, diffusion coefficients, and viscosity.

2. THERMODYNAMIC EQUILIBRIUM

Total thermodynamic equilibrium (TTE) is defined as a state in which $T_e = T_h = T_{\text{radiation}}$ holds, where T_e and T_h are the temperatures of electrons and heavy particles (molecules, atoms, and ions), respectively, and $T_{\text{radiation}}$ is the temperature to be used for the computation of the intensity of the blackbody radiation field according to the Planck function. Because of the poor energy exchange in elastic collisions of electrons with heavy particles [Eq. (3)], TTE is not realized in electrical discharges. If certain conditions are met in the plasma, a state more or less close to TTE may be present. In particular, the concept of local thermal equilibrium (LTE) is a useful means to describe the high-pressure arc column. The underlying idea is that in such an arc column the particle number densities and therefore the number of collisions are sufficient to ensure one common temperature of any species constituting the plasma. The properties of the plasma volume under consideration may then be described by a single temperature and the ambient pressure.

Exempted from this thermal equilibrium is the radiation field only. Since most plasmas radiate a considerable portion of the spectrum without appreciable self-absorption, the radiation field will, in general, not be governed by the Planck radiation law, as required by TTE. A closer look at the implications of LTE leads to the criteria for estimating the parameter range where this concept can be applied successfully.

2.1 Criteria for the Establishment of LTE

The basis for any thermal descriptions is the nearly equal temperatures of the Maxwell–Boltzmann velocity distributions of the heavy species and the electrons.

The amount of energy that is picked up by the electrons from the electric field in between collisions (i.e., $\lambda_e eE$) will not be completely redistributed to the heavy particles by collisions. λ_e denotes the mean free path length of the electrons, given by $\lambda_e^{-1} = \Sigma n_k \sigma_{ek}$, where σ_{ek} is the collision cross section of electrons with the species k of particle number density n_k. The resulting excess temperature of the electrons may be estimated by the Maecker–Finkelnburg (Maecker and Finkelnburg, 1957) criterion, which compares $\lambda_e eE$ to the energy transferred by elastic collisions to the heavy particles:

$$\frac{T_e - T_h}{T_e} = \frac{m_{\text{heavy}}}{4m_e}\left(\frac{\lambda_e eE}{\frac{3}{2}kT_e}\right)^2. \tag{8}$$

This criterion gives a fair insight into the requirements for the application of LTE and yields a good approximation for a homogenous and weakly ionized plasma. The higher the temperature and the larger the particle number densities for a prevailing electrical field, the better is the basis for the establishment of an LTE plasma.

Generally speaking, high pressures will diminish the temperature gap $T_e - T_h$ in an arc discharge, where an increase in pressure reduces λ_e to a larger extent than it strengthens the electric field E. Equally favorable are high current densities, which usually result in high temperatures and high electrical conductivities. The Joule heat input $\mathbf{j} \cdot \mathbf{E}$ necessary to retain a certain temperature requires less field strength when the electrical conductivity is high. This criterion is usually fulfilled to a good degree in a high-pressure and high-current arc column.

A more stringent requirement is the validity of the Boltzmann relation for the population density of any energy level of the heavy species. As mentioned above, radiation may play an important role as a loss term, and the population density of an energy level, especially the lower-lying ones, may deviate from the Boltzmann relation as a result of radiative depopulation. To ensure thermal population densities to within 10%, collisional excitation should exceed the radiative population rate by a factor of approximately 10. Griem (1964) among others gives a criterion for the minimum electron number density required:

$$n_{e\,\min} \geq 10^{18}\sqrt{T_e}(E_{21})^3\ \text{m}^{-3}. \tag{9}$$

The energy gap E_{21} between the ground state and the first excited state, which is usually the largest, has to be inserted in eV and T_e in K. This minimum electron number density might be relaxed by a factor of up to 10 if the resonance lines are reabsorbed. At least a partial reabsorption of resonance line radiation is quite often found in arc columns, and

$n_{e\,\min}$ may be lowered by an appropriate factor.

Another condition might be easily violated in the column of some arc devices. The electron temperature must not vary considerably within a distance that a heavy particle traverses until it is in ionization or excitation equilibrium according to the local electron temperature. An estimate of this diffusion length is given by Griem (1964) and reads

$$d \approx 5 \times 10^{17}\left(\frac{1}{A}\right)^{1/4}\sqrt{\frac{kTE_{21}}{f_{21}}} \times \frac{1}{\sqrt{(n_a + 2n_i)(n_a + 11n_i)}} \exp\left(\frac{E_{21}}{2kT}\right) \text{ m.} \quad (10)$$

Here A is the atomic mass, f_{21} denotes the absorption oscillator strength of the resonance line, i.e., the number of classical absorbers of unit amplitude that would be equivalent to one quantum absorber (see Richter, 1968; Fowler, 1956); and n_a and n_i are the particle number densities of atoms and ions, respectively.

The more restrictive criteria of Eqs. (9) and (10) depend only on the particle number densities for a given temperature, which may be approximated by the ambient pressure for a single ionized gas. Because of quasineutrality, the square root of the particle number densities given in Eq. (10) is larger than the total number density. This may be expressed in terms of pressure according to Dalton's law, as long as the electron temperature does not severely exceed the temperature of the heavy particles; i.e., if the Maecker–Finkelnburg criterion of Eq. (8) yields a sufficiently small number, then

$$\sqrt{(n_a + 2n_i)(n_a + 11n_i)} \geq (n_a + 2n_i) \propto p.$$

Therefore, the pressure dependence can be safely estimated by $d \propto p^{-1}$.

The pressure dependence of the electron number density in a plasma where the temperature gap is not too large may be approximated by $\sqrt{p}$, as may be shown from an Eggert–Saha equation (Eggert, 1919; Saha, 1920):

$$\frac{n_e n_z}{n_{z-1}} = 2\frac{U_z(T,\Delta E)}{U_{z-1}(T,\Delta E_{z-1})}\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2} \times \exp\left[-\frac{E_{z-1} - \Delta E_{z-1}}{kT}\right]. \quad (11)$$

Here U_z denotes the partition function of the z-fold ion (a resulting subscript 0 denotes the atom) and ΔE_{z-1} is the lowering of the ionization potential, which is taken from

$$\Delta E_{z-1} = ze^2/4\pi\epsilon_0 r_{\mathrm{D}},$$

with the Debye radius r_{D}, according to the Debye–Hückel theory.

All of the requirements are more likely to be met at higher pressures. The diffusion length decreases most rapidly with increasing pressure, $d \propto p^{-1}$, while the electron density changes only moderately with pressure, $n_e \propto \sqrt{p}$, and the temperature gap $T_e - T_h$ is usually not a very restrictive criterion for high-pressure arcs.

Therefore, one might use the minimum electron number density and the diffusion length to estimate the validity of LTE. At atmospheric pressure, small electron number densities and large diffusion lengths d may cause deviations from LTE at moderate temperatures. Table 1 gives numerical values for an argon plasma. The minimum electron number density of 10^{23} m^{-3} is attained at 13 000 K. This corresponds to an $n_{e\,\min}$ of Eq. (9) relaxed by a factor of ≅0.6, i.e., only partial reabsorption of resonance line radiation.

2.2 Selected Properties of LTE and PLTE Plasmas

If LTE holds, the particle number densities are calculated from quasineutrality, Dalton's law, and the Saha-Eggert equation (11). The population densities of the excited levels are then taken from the Boltzmann relation:

$$n_k = n_z \frac{g_{z,k}}{U_z} \exp\left(-\frac{E_k}{kT}\right). \quad (12)$$

Table 1. Temperature dependence of LTE criteria for argon at atmospheric pressure.

	LTE Criterion	
T (K)	Temperature gap	Diffusion length d (mm)
11 000	<0.04	~1.5
12 000	<0.02	~0.8
13 000	<0.01	~0.5

The subscript k denotes the energy level k, and $g_{z,k}$ is the statistical weight and E_k the energy of this excited state, where n_z is again the number density of the z-fold ion or the atom, respectively.

The transport coefficients of the LTE plasma depend on the temperature, the pressure, and possibly the mass fraction, if mixing of different gases is to be considered.

If the electron number density is not sufficiently high to ensure that the Boltzmann relation holds for every energy level, a partial local thermal equilibrium (PLTE) model may be used. In this case it is assumed that only the higher-lying states are in equilibrium with the electron temperature while the ground state may be severely overpopulated. A generalized Saha equation may be used, where the number densities n_e, n_z are given with respect to an excited level k of the species n_{z-1}:

$$\frac{n_e n_z}{n_{z-1,k}} = 2\frac{U_z}{g_{z-1,k}}\left(\frac{m_e k T_e}{2\pi\hbar^2}\right)^{3/2} \times \exp\left[-\frac{E_{z-1} - \Delta E_{z-1}}{kT_e}\right]. \tag{13}$$

3. THE ELECTRICAL ARC DISCHARGE

The electrical arc owes its name to the curved shape of the discharge, which was operated between horizontal carbon electrodes since the beginning of the 19th century. The buoyancy of the gas heated by the discharge is the origin of the upward curvature of the field of isotherms. Early applications made use of the plasma radiation for illumination purposes. The high temperatures achievable with this discharge make it suitable for materials processing. Arc discharges are self-sustained discharges with currents typically above 1 A and with a cathode fall of less than 20 V. As in all electrical discharges containing electrodes, a cathode fall is present in the discharge. This is a potential drop between the cathode and the arc plasma; its value depends on gas composition, the gas pressure, and the cathode conditions, such as material shape and temperature. In the cathode-fall region, the electrons released at the cathode gain the energy necessary for ionization processes. The low cathode fall together with the high current distinguishes the arc discharge from the glow discharge with its currents in the milliampere region and with cathode falls well above 100 V. The electron emission mechanism at the cathode of an arc discharge is mostly based on a combination of thermionic and field emission. Different types of arc discharges may be characterized by the pressure range (several times 10^2 to 10^9 Pa) in which they are operated; other distinguishing properties are the type of stabilization and the interaction of the discharge with external flow fields. The discharge is often started by high-voltage breakdown. Other starting procedures consist of initially establishing contact between the electrodes before separation occurs, or making use of an exploding wire. Once the arc discharge is operating, three regions can be distinguished: the cathode and the anode regions and the arc column proper.

The governing equations for the basic variables describing the state of the arc column consist of the set of conservation equations for mass, species, momentum, and energy, Ohm's law, and the appropriate Maxwell equations.

The temperature distribution is determined by the energy conservation equation, which reads in its time-dependent form

$$\rho\frac{Dh_k}{Dt} = \mathbf{j}\cdot\mathbf{E} + \frac{\partial p}{\partial t} + \mathrm{div}(\mathbf{v}\cdot\boldsymbol{\tau}) + \mathrm{div}(\lambda\,\mathrm{grad}T) - u, \tag{14}$$

with ρ = mass density, h_k = specific enthalpy, $\mathbf{j}$ = current density, $\mathbf{E}$ = electrical field strength, p = pressure, $\mathbf{v}$ = velocity, $\boldsymbol{\tau}$ = stress tensor, $\boldsymbol{\lambda}$ = thermal conductivity, $u = e - a$ = balance between emitted and absorbed radiation, and $D/Dt = \partial/\partial t + \boldsymbol{\omega}\cdot\mathrm{grad}$. The convective enthalpy flux balances the difference of Ohmic heating and radiation losses together with the heat conduction, and the change in pressure with time. This equation may be extended to include the diffusive and convective enthalpy fluxes due to concentration gradients, if the mixing of two gases is to be considered.

The velocity is given by the solution of the Navier–Stokes equation

$$\rho\frac{D\mathbf{v}}{Dt} = -\mathrm{grad}\,p + \mathrm{div}(\boldsymbol{\tau}) + \rho\mathbf{g} + \mathbf{j}\times\mathbf{B}. \tag{15}$$

The pressure gradient and the viscous forces together with the Lorentz forces $\mathbf{j} \times \mathbf{B}$ ($\mathbf{B}$ denotes the magnetic field) and natural convection $\rho\mathbf{g}$ balance the acceleration of the plasma. The mass continuity completes the fluid dynamic equations:

$$\frac{\partial\rho}{\partial t} + \operatorname{div}(\rho\mathbf{v}) = 0. \tag{16}$$

If turbulence is to be accounted for, one might use the so-called κ-ϵ model, where two additional transport equations, for the turbulent kinetic energy κ and its rate of change ϵ, are to be solved. From these two quantities the eddy (turbulent) part of viscosity, thermal conductivity, and possibly the diffusion coefficients needed for mixing flows are calculated.

The current density is governed by a generalized Ohm's law,

$$\mathbf{j} = \sigma\left(\mathbf{E} + \mathbf{v} \times \mathbf{B} + \frac{1}{en_e}\mathbf{j} \times \mathbf{B} + \frac{1}{en_e}\operatorname{grad} p_e\right), \tag{17}$$

which often simplifies for the arc column to $\mathbf{j} = \sigma\mathbf{E}$. The electrical conductivity is denoted by σ, and $p_e = n_e kT$ is the partial pressure of the electrons. The magnetic field is found via Maxwell's equations, which reduce to

$$\operatorname{curl} H = \mathbf{j} \tag{18}$$

as long as the displacement current densities can be neglected relative to $\mathbf{j}$. This set of closely coupled nonlinear equations can only be solved numerically.

In many applications the stability of the arc plays an important role. As one example: The proper function of arc discharge lamps depends on the arc stability since instabilities not only may lead to a thermal overload of lamp parts but will also affect the focusing properties of the lamp. On the other hand, instabilities may also have a positive effect in applications such as high-voltage circuit breakers where they can enhance the quenching of the arc during current zero. Different types of instabilities have to be considered: Besides electrode instabilities, hydrodynamic and magnetic instabilities exist. The electrode instability is a complex problem involving properties of the electrode material as well as hydrodynamics of the jet at the arc root formed by evaporating material and self magnetic fields (Mentel, 1978). Hydrodynamic instabilities are similar to the ones encountered in aerodynamics. Only magnetic instabilities are specific for the arc discharge and can be treated in a more general manner (Mentel, 1985). Self magnetic fields play a role as well as superimposed fields. A perturbation theory has been applied to the column; it yields predictions of stability regions that are in good agreement with experiments.

3.1 High-Pressure Arcs

3.1.1 Wall-Stabilized Arcs Arcs operating at atmospheric pressure or above are called high-pressure arcs. In this pressure range collision frequencies are high, and LTE can frequently be assumed. The LTE arc column may then be described as a conducting fluid, with this set of basic variables: temperature T, velocity $\boldsymbol{v}$, current density $\mathbf{j}$, and pressure p. In most applications a flow-stabilized discharge is used. Because of the complexity of this discharge type, most basic investigations have been made up to now with the so-called cascade arc (Maecker and Steinberger, 1967). This situation has changed recently, since the "free burning arc," as described below, is now well understood and investigated from both the theoretical and the experimental standpoints. Beginning with a short review of the wall-stabilized cascade arc, some basic properties of the arc discharge can be described. The arc column is stabilized by enclosing it in a cylindrical tube with an inner diameter in the range of typically 2–30 mm. Its simplest uncooled form is shown in Fig. 1 (taken from Poisel *et al.*, 1986). In this case the arc was operated between a tungsten cathode located on the axis of the system and a tungsten ring anode. The tube that confines the column has to be isolating in the axial direction. It should have high thermal conductivity in the radial direction in order to be able to handle the high power flux coming from the plasma without melting. This combination of properties is obtained by using copper plates with a central bore of 2 mm that are isolated from each other by a ceramic

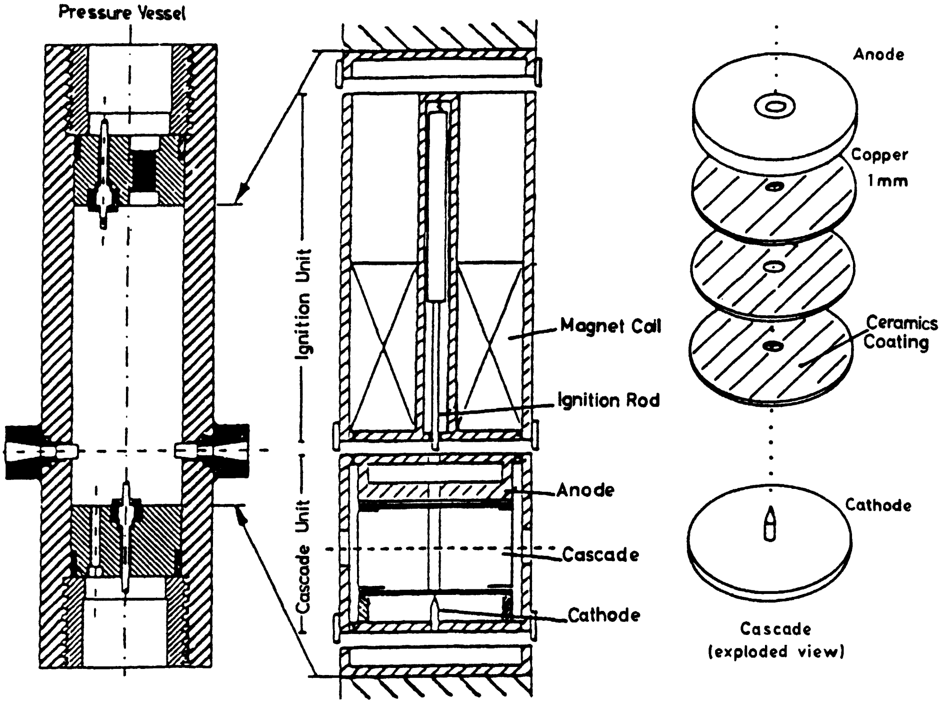

FIG. 1. Cascade arc assembly. A sufficient number of cascade plates fill the space between cathode and anode. (From Poisel *et al.*, 1986.)

coating of 0.2 mm thickness. The arc is operating in argon through the central holes. The cascade assembly consists of 40 plates, each plate is 1 mm thick. Arc currents of 10–100 A are used at atmospheric pressure with a pulse duration of several seconds. A maximum temperature of 15 000 K is obtained, and the power load of the inner channel surface reaches 8 kW/cm^2. With water-cooled cascades of the Maecker type, the corresponding values are 26 000 K in nitrogen, the power load being 20 kW/cm^2. Cascade arcs have been operated at pressures up to 25 MPa. Depending on the pressure, the temperature profiles are more or less flat in the center part. A flat T distribution is found at higher pressures; it is due to the increased importance of radiation transport in the energy balance of the column. Figure 2 (taken from Schade, 1970) gives $T(r)$ profiles measured spectroscopically in an atmospheric-pressure nitrogen arc of 5 mm diameter. At currents below 100 A, a shoulder becomes visible in the profiles. This is typical for molecular gases and is due to the high thermal conductivity in the dissociation region where diffusion processes lead to an enhanced radial transport of dissociation energy. The field-strength/current (E/I) characteristic of high-pressure arc discharges consists of a falling part at low currents and a rising part in the high current range, Fig. 3 (taken from Bauder *et al.*, 1973); see also Fig. 12, part EF, in Sec. 5 (Glow Discharges). If operated in the falling part of the arc characteristic, the discharge requires for stable operation an Ohmic resistance in series with the arc; by this measure a rising characteristic results for the arc/resistance system at a given current.

The energy-balance equation takes a very simple form in the case of a wall-stabilized arc without flow. The general energy equation for an arc column, Eq. (14), simplifies to the modified Elenbaas–Heller equation (Elenbaas, 1936):

$$\sigma E^2 + \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial S}{\partial r}\right) - u = 0, \qquad (19)$$

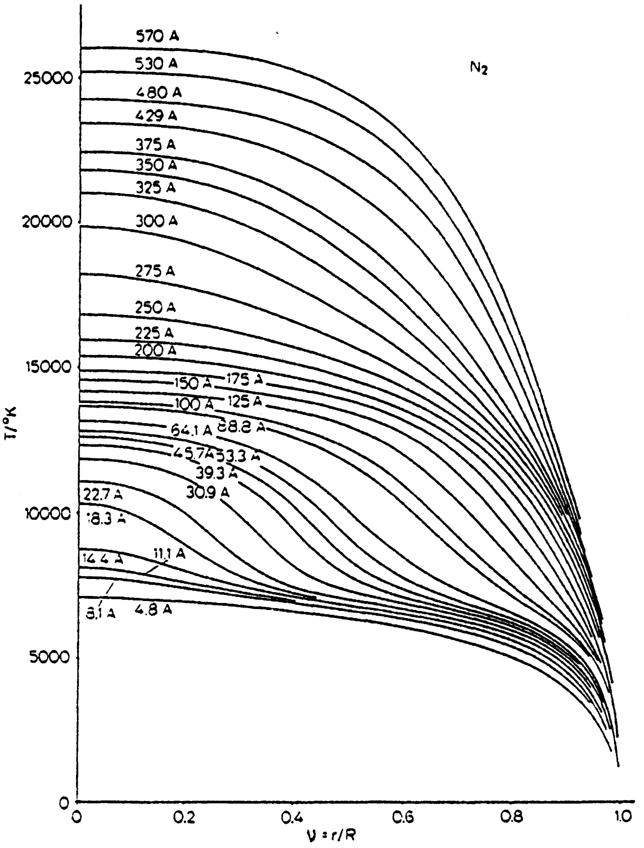

FIG. 2. Temperature distributions of an N_2 arc at different arc currents (from Schade, 1970).

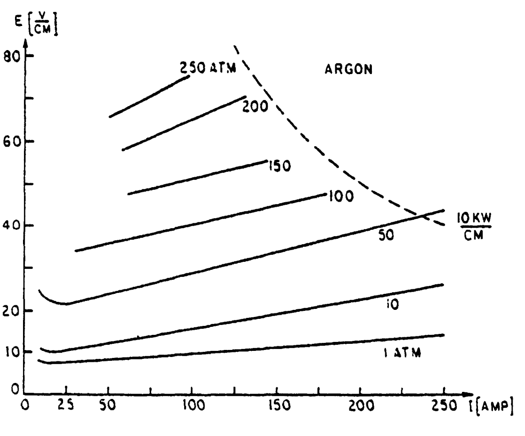

FIG. 3. Voltage–current characteristics of an argon cascade arc at different pressures (from Bauder *et al.*, 1973).

where σ is the electrical conductivity and $S = \int \lambda \, dT$ is the heat-flux potential. If radiation can be neglected, Eq. (19) may be rewritten:

$$\frac{\sigma L}{\pi \int_0^1 \sigma d(\xi^2)} + \frac{1}{\xi}\frac{\partial}{\partial \xi}\left(\xi \frac{\partial S}{\partial \xi}\right) = 0 \qquad (20)$$

with $\xi = r/R$ (R being the radius of the arc column) and $L = IE$ the Ohmic heating per unit arc length. Equation (20) is independent of the total radius R, and therefore equal values of L lead to identical $T(r)$ distributions. This property of the simplified energy balance is used for the prediction of E and I values of columns with different radii (similarity laws).

The cascade arc has been used mostly in scientific applications. As a result of its cylindrical symmetry, it is well suited for the experimental determination of transport properties of plasmas. Values have been published for the electrical conductivity σ, the thermal conductivity λ, the viscosity, and the total emissivity integrated over the whole wavelength range. The pressures used were between 0.1 and 20 MPa; gases like hydrogen, helium, nitrogen, oxygen, argon, and sulfur hexafluoride have been investigated. By performing spectroscopy in the range between the infrared (IR) and vacuum ultraviolet (VUV) and by interferometric measurements on the cascade-arc plasma, atomic data have been determined: transition probabilities and excitation and ionization cross sections as well as data on spectral line broadening and line shift due to the Stark effect and to neutral–neutral and neutral–charged-particle interaction. On the other hand, the technological difficulties with steady-state cascade arcs are considerable. For this reason the free-burning arc is now used frequently for such investigations.

3.1.2 Flow-Stabilized Arcs An arc may generate the stabilizing flow itself, when the arc diameter and therefore the related current density change along its axis. A self-stabilizing arc exhibiting cylindrical symmetry is realized, for example, in the tungsten inert-gas (TIG) welding arc or in high-pressure arc lamps. These arcs are operated between a conical cathode and a more or less flat anode. The electrode spacing is of the order of a few millimeters. dc currents of some ten to a few hundred amperes may be used; the ambient pressure reaches up to 10 MPa.

The current density peaks at the cathode tip and falls off with increasing distance from the cathode. The Lorentz force, which originates from the interaction of the self magnetic field with the current density, is therefore reduced along the axis. This decrease of the Lorentz force results in a pinch-pressure gradient along the axis of the arc, which drives the stabilizing flow. At the anode the plasma impinges on a solid wall and is deflected accordingly.

The voltage–current characteristics of these arcs resemble the E-I curves of wall-stabilized arcs, where the pressure dependence of the arc voltage may be approximated by

$$V = \bar{E}p^{m}l + V_0. \tag{21}$$

The mean value of the electric field $\bar{E}$ in the arc column depends for a given current on the gas used and decreases slightly with increasing arc length l; likewise, the exponent m varies for different gases. The voltage V_0 accounts for the cathode and anode fall and is essentially independent of the ambient pressure.

The temperature distributions of such free-burning argon arcs have been measured at pressures up to 10 MPa by spectroscopic means. The arc is operated in a water-cooled electrode chamber, which removes any excess heat from the high-pressure vessel into which it is fitted. At ambient pressures of several megapascals, more than half of the input power of a few kilowatts leaves the plasma as radiation losses. The water-cooled cathode is a rod of thoriated tungsten with a tip machined to a vertex angle of 60°. The anode consists of a water-cooled copper block lined with tungsten. This design allows for relatively high temperatures in the colder, nonionized gas, which surrounds the arc column. High temperatures are necessary in order to keep the effect of schlieren in the colder gas to a tolerable level. Schlieren are disturbances in the light transmission due to variations in the index of refraction. A Pyrex glass bell fitted with quartz windows ensures that all the gas in the optical path is at elevated temperatures, where the change of the refractive index is less pronounced. Figure 4 illustrates this temperature dependence of the refractive index for argon. The electrode chamber, which is used for temperature measurements, is shown in Fig. 5.

The spectroscopic temperature measurement of the arc column utilizes the known temperature dependence of the particle number and population densities in an LTE plasma for a given ambient pressure. These quantities may be obtained via the intensity of line and continuum radiation. The intensities that are measured consist of the integral of emission and absorption along the line of sight and have to be converted into the local emission quantities. This may be achieved by various methods for solving the Radon integral equation (Radon, 1917), which are routinely employed in computer tomography. In case of cylindrical symmetry and moderate absorption, this simplifies to an Abel integral equation.

The line emission coefficient denotes the radiated power per solid angle, which stems

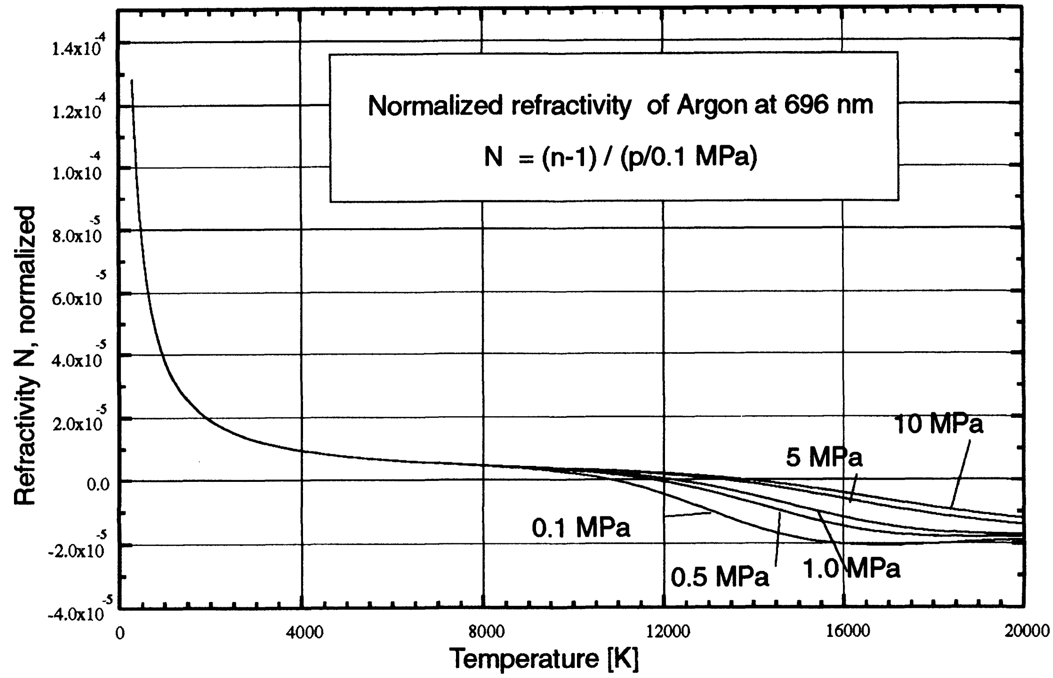

FIG. 4. Index of refraction of argon as function of the temperature.

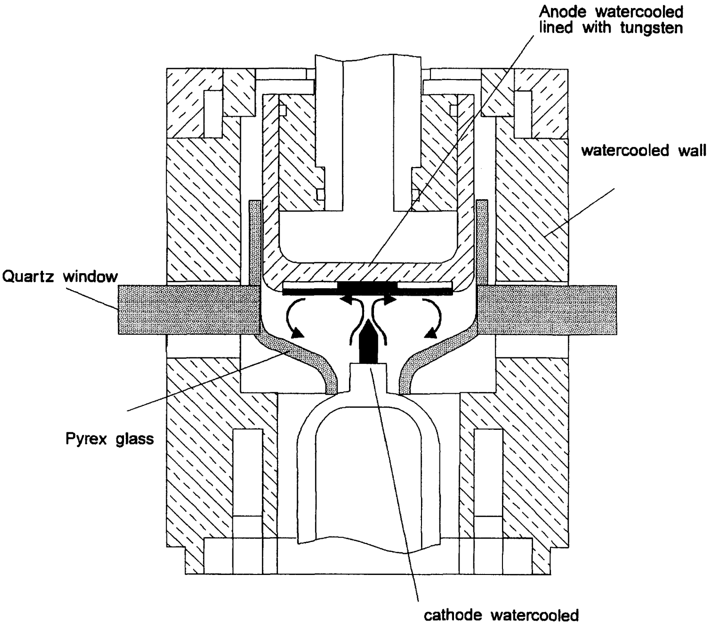
Anode watercooled lined with tungsten
watercooled wall
Quartz window
Pyrex glass
cathode watercooled

FIG. 5. Schematic drawing of the electrode assembly of a high-pressure free-burning arc.

from spontaneous radiative transition between two discrete energy levels. In LTE it depends only on the population density of the upper level E_k, when the transition probability A_{ki} is known, and reads

$$\epsilon_L = \frac{h\nu}{4\pi} A_{ki} n_z(T) \frac{g_k}{U_z(T)} \exp\left[-\frac{E_k}{kT}\right] \quad \frac{\mathrm{W}}{\mathrm{m}^3\mathrm{sr}}, \tag{22}$$

n being the frequency and g_k the statistical weight of level k in the ionization state z. The spectral distribution of the radiated power within this spectral line may further be used to determine particle number densities in the high-pressure arc plasma. In general, the wavelength dependence of the radiated power can be denoted by

$$\begin{aligned} \epsilon_\nu &= P(\nu)\epsilon_L, \\ 1 &= \int_{\mathrm{Line}} P(\nu) d\nu, \end{aligned} \tag{23}$$

with $P(\nu)$ a normalized profile. Stark and possibly van der Waals broadening often dominate the line profile in high-pressure arcs, while Doppler broadening plays a minor role because of the relatively low temperatures. While the electron number density may be taken from the Stark effect, the atom density is taken from van der Waals broadening. Both of these mechanisms lead to a dispersion relation, and the sum of the half widths due to these Lorentzian profiles yields the actual half width of the line.

The continuum radiation arises from recombination (free–bound) and bremsstrahlung (free–free) transitions and is usually given in terms of hydrogen, with appropriate correction factors, the so-called Bibermann factors. For a high-pressure plasma the contribution of electron–atom encounters has also to be taken into account, and the continuum emission coefficient becomes

$$\epsilon_C = \epsilon_{ff}^{ei} + \epsilon_{ff}^{ea} + \epsilon_{fb}. \tag{24}$$

The terms may be calculated from

$$\begin{aligned} \epsilon_{ff}^{ei} &= \frac{1}{(4\pi\epsilon_0)^3} \frac{16\pi e^6}{3c^3(6\pi m_e^3 k)^{1/2}} \sum_z \frac{n_e n_z}{T_e^{1/2}} z^2 \\ &\quad \times \exp\left[-\frac{h\nu}{kT_e}\right] \xi_{ff}(\nu, T_e, z), \end{aligned} \tag{25}$$

$$\begin{aligned} \epsilon_{ff}^{ea} &= \frac{1}{4\pi\epsilon_0} \frac{32e^2}{3c^3} \left(\frac{k}{2\pi m_e}\right)^{3/2} n_a n_e T_e^{3/2} Q_{ea}(T_e) \\ &\quad \times \left\{\left(1 + \frac{h\nu}{kT_e}\right)^2 + 1\right\} \exp\left[-\frac{h\nu}{kT_e}\right], \end{aligned} \tag{26}$$

$$\begin{aligned} \epsilon_{fb} &= \frac{1}{(4\pi\epsilon_0)^3} \frac{16\pi e^6}{3c^3(6\pi m_e^3 k)^{1/2}} \sum_z \frac{n_e n_z}{T_e^{1/2}} \\ &\quad \times z^2 \left\{1 - \exp\left[-\frac{h\nu}{kT_e}\right]\right\} \frac{g_{z,1}}{U_z} \xi_{fb}(\nu, T_e, z). \end{aligned} \tag{27}$$

As an example, for a temperature measurement the two-dimensional intensity distribution of the continuum radiation is recorded, utilizing a CCD camera and an interference filter. The intensities are converted to local continuum emission coefficients by solving the Abel integral equation and are translated into temperatures. Alternatively, line radiation may be recorded using a spectrometer with imaging capabilities and a CCD camera instead of an exit slit. The spectral distribution of the line radiation is obtained for one cross section of the plasma, and the spatially resolved line profiles can be determined from this measurement.

The comparison of two different published temperature measurements is shown in Fig. 6 for an argon arc with 5-mm electrode gap and 200-A dc current at atmospheric pressure. Also shown in this figure is the calculated temperature distribution, which was derived according to the next section.

For a theoretical description of this arc column, the governing equations (14)–(18) simplify for this quite moderate type of complexity: steady state, cylindrical coordinates in two dimensions, no turbulence, and no mixing of different gases. The resulting equations are given below, where the convective parts are expressed in terms of div($\rho\mathbf{v}$ . . .) rather than in $\rho\mathbf{v}\cdot$grad(..), which is due to the numerical scheme used for solving this set of equations.

The energy conservation may be rewritten in terms of enthalpy only, the kinetic energy being small compared to the enthalpy of the gas. Furthermore, viscous effects do not contribute to any significant extent to the energy balance and may be neglected as well. This gives

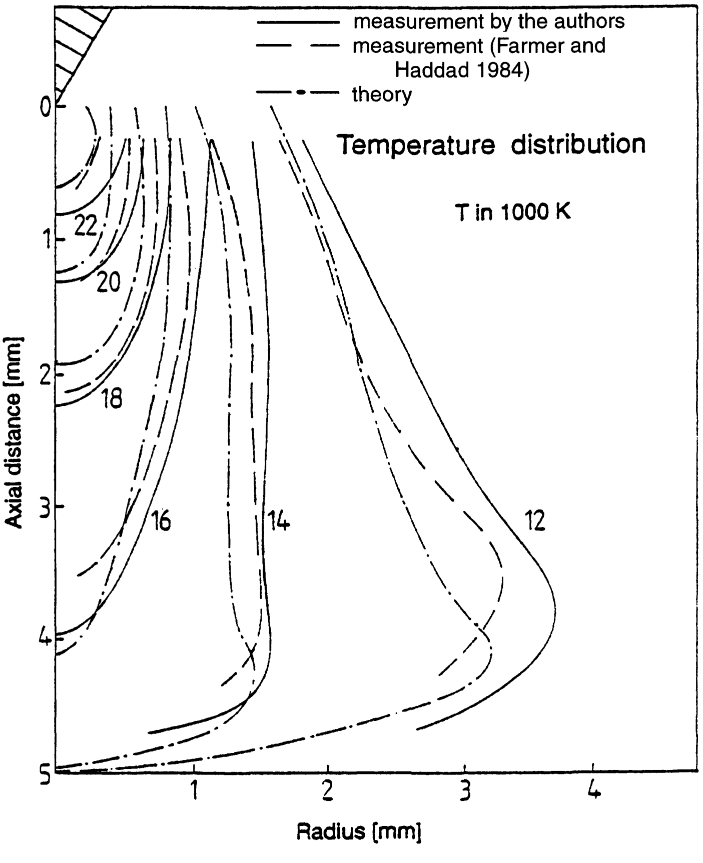

FIG. 6. Temperature distribution $T(r)$ of the free-burning arc at $p = 0.1$ MPa, $I = 200$ A.

$$\frac{1}{r}\frac{\partial}{\partial r}(r\rho v_r h) + \frac{\partial}{\partial z}(\rho v_z h) = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\lambda}{c_p}\frac{\partial h}{\partial r}\right) + \frac{\partial}{\partial z}\left(\frac{\lambda}{c_p}\frac{\partial h}{\partial z}\right) + \frac{j_r^2 + j_z^2}{\sigma} - u. \tag{28}$$

The heat flux is represented in terms of enthalpy with c_p being the specific heat. The radial and axial momentum equations cannot be simplified considerably; the radial momentum balance reads

$$\begin{aligned}\frac{1}{r}\frac{\partial}{\partial r}(r\rho v_r^2) + \frac{\partial}{\partial z}(\rho v_r v_z) &= -\frac{\partial P_\Delta}{\partial r} - j_z B_\theta \\ &+ \frac{1}{r}\frac{\partial}{\partial r}\left(2r\eta\frac{\partial v_r}{\partial r}\right) + \frac{\partial}{\partial z}\left[\eta\left(\frac{\partial v_z}{\partial r} + \frac{\partial v_r}{\partial z}\right)\right] \\ &- 2\eta\frac{v_r}{r^2} - \frac{\partial}{\partial r}\left[\frac{2}{3}\eta\left(\frac{1}{r}\frac{\partial(rv_r)}{\partial r} + \frac{\partial v_z}{\partial z}\right)\right].\end{aligned} \tag{29}$$

The corresponding axial momentum equation is given by

$$\begin{aligned}\frac{1}{r}\frac{\partial}{\partial r}(r\rho v_r v_z) + \frac{\partial}{\partial z}(\rho v_z^2) &= -\frac{\partial P_\Delta}{\partial z} + j_r B_\theta \\ &+ \frac{\partial}{\partial z}\left(2\eta\frac{\partial v_z}{\partial z}\right) + \frac{1}{r}\frac{\partial}{\partial r}\left[r\eta\left(\frac{\partial v_z}{\partial r} + \frac{\partial v_r}{\partial z}\right)\right] \\ &- \frac{\partial}{\partial z}\left[\frac{2}{3}\eta\left(\frac{1}{r}\frac{\partial(rv_r)}{\partial r} + \frac{\partial v_z}{\partial z}\right)\right] + \rho g.\end{aligned} \tag{30}$$

For computational convenience, the pressure is calculated in terms of $P_\Delta = p - p_{\text{ambient}}$. The mass continuity becomes

$$\frac{1}{r}\frac{\partial}{\partial r}(r\rho v_r) + \frac{\partial}{\partial z}(\rho v_z) = 0. \tag{31}$$

The current density is calculated from the charge continuity and Ohm's law, where the electrical field strength is given by the gradient of the potential distribution:

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\sigma\frac{\partial\phi}{\partial r}\right) + \frac{\partial}{\partial z}\left(\sigma\frac{\partial\phi}{\partial z}\right) = 0;$$
$$j_r = -\sigma\frac{\partial\phi}{\partial r}, \quad j_z = -\sigma\frac{\partial\phi}{\partial z}. \tag{32}$$

The magnetic field is given by

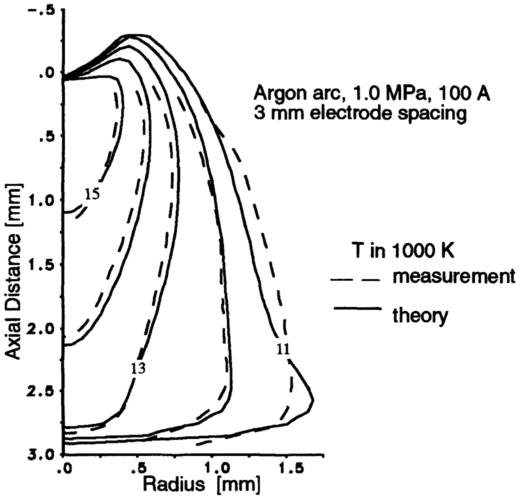

FIG. 7. Comparison of theoretical and experimental $T(r)$ distributions at p = 1 MPa.

$$\frac{1}{r}\frac{\partial}{\partial r}(rB_\varphi) = \mu_0 j_z, \tag{33}$$

where μ_0 is the permeability of free space.

For a numerical solution of the governing equations, the finite-volume approach, also called control volume, is widely used (Patankar, 1980). The computational domain is divided into a sufficiently large number of control volumes, and the conservation equations are integrated over each volume. The diffusive and convective terms yield fluxes across the boundaries of the quantity under consideration. These fluxes are conserved exactly, so that the conservation equations are modeled numerically.

The numerical solution of this set of equations yields the distribution of the basic variables. It has been shown that the arc column can be simulated self-consistently, with the total input current and the electrode geometry as boundary conditions (Lowke *et al.*, 1992), while neglecting the nonthermal sheath layers at the cathode and anode. Figure 7 shows the comparison of calculated and measured temperature distribution of an argon TIG (tungsten–inert-gas)-type welding arc at an elevated pressure of 1 MPa.

An analysis of the energy conservation equation for this arc is shown in Fig. 8,

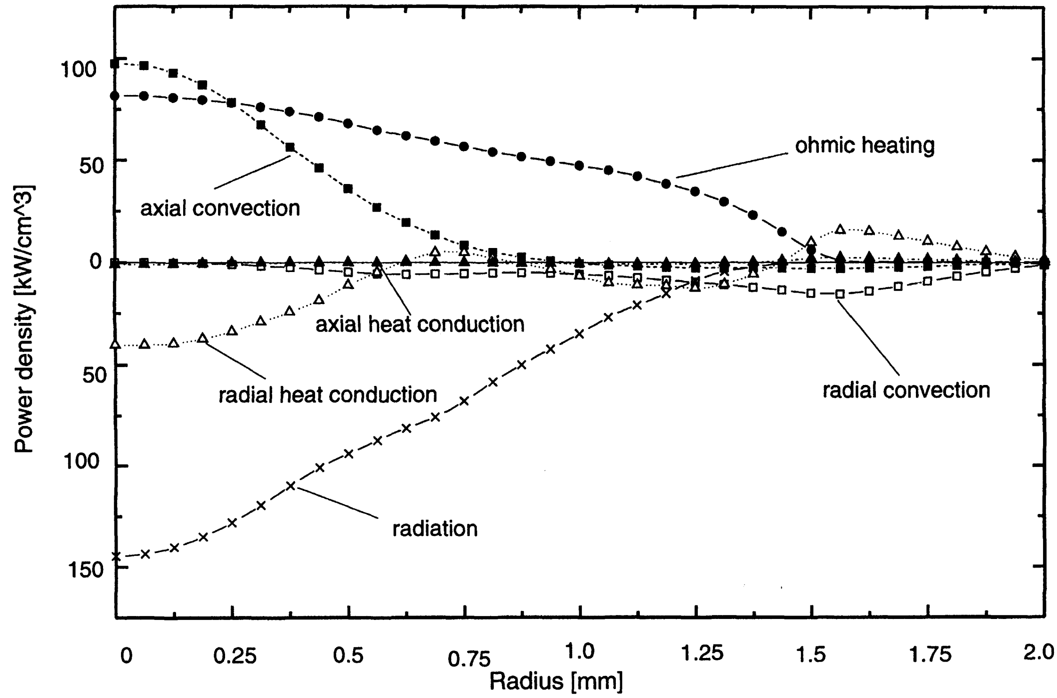

FIG. 8. Radial profile of the energy densities in the midplane between the electrodes of a TIG arc at p = 1 MPa, I = 100 A.

where the radial profile of the energy densities is given for the midplane between the electrodes.

At the center the axial enthalpy flow dominates the energy input and slightly exceeds Ohmic heating, while the radiative losses clearly excel the conduction losses. The arc behavior is strongly influenced by the plasma flow. The cold gas is drawn into the arc at the cathode and heated, and the plasma flows toward the anode. The resulting velocity distribution is shown in Fig. 9. The driving force arises from the gradient of the pinch pressure due to the self magnetic field of the arc. This pressure field is determined by the current density, which thus has a dominating influence on the convective flow of arc plasma and, hence, of temperatures and diameters. The arc properties depend on the arcing current and the arc root at the cathode, where the velocities, temperatures, and diameters are increasing with larger input current.

On account of the strong convective flow, the arc is not influenced largely by the downstream conditions; therefore, a change in arc length has only a minor influence on the properties of the discharge, as long as laminar flow prevails.

The pressure dependence of the discharge is determined by the transport coefficients and the radiation losses. The transport properties, such as electrical conductivity, thermal conductivity, and viscosity, do not depend strongly on the ambient pressure. For example, the conductivities σ and λ of an argon plasma at common arc temperatures vary less than a factor of 2–3 when pressure changes 100-fold from 0.1 to 10 MPa. The mass density scales approximately with p, and hence it dominates the pressure dependence of the transport coefficients. Neglecting the pressure dependence of the reabsorption in the pressure range under consideration, radiation losses are also roughly proportional to the ambient pressure, for singly ionized plasmas.

For the arc at 1 MPa, shown in Fig. 7, radiation is already more important than heat conduction, and therefore the curvature of

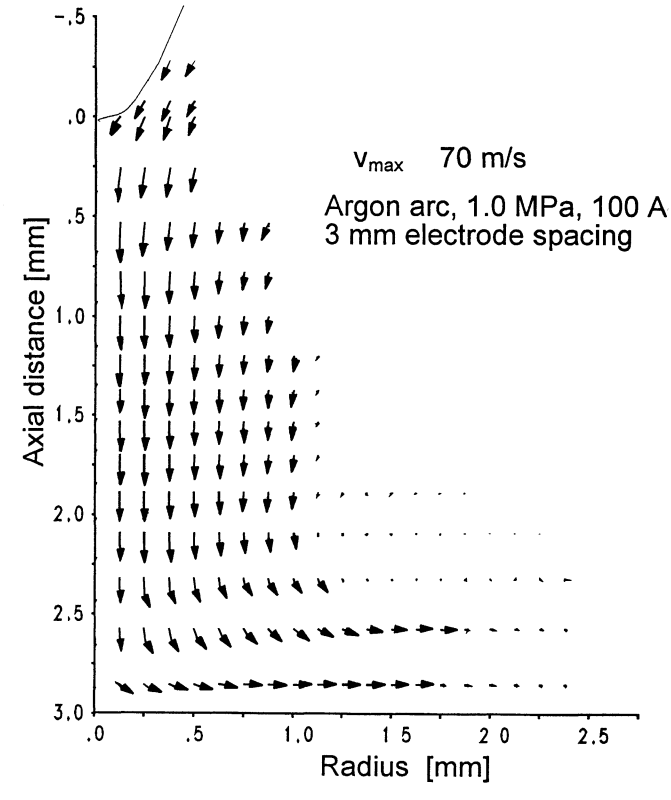

FIG. 9. Velocity distribution of a TIG arc at p = 1 MPa, I = 100 A.

the temperature profiles is only moderate up to a radius where most of the current is carried. With increasing pressure, heat conduction contributes even less to the energy balance, and the radial temperature profiles become more flat. For example, if the pressure is increased to 5 MPa, virtually all the heat is removed by radiation, and hardly any change in the heat flux is required for balancing the heat input near the axis. The temperatures are decreasing with increasing ambient pressure.

The change in mass density results in a stronger convective cooling of the plasma, especially at the cathode at the arc root. The arc constricts, so that the current densities and Lorentz forces are increased, resulting in larger pinch-pressure gradients. This increase is more than compensated for by larger inertia of the plasma; hence the velocities decrease with increasing pressure.

Besides applications in the fields of welding and cutting, the arc discharge influenced by flow effects has important applications in high-power radiation sources, high-power and -voltage circuit breakers (the arc gas being then typically SF_6), and electrothermal launchers.

3.1.3 Arc Jets While in flow-stabilized arcs, convection plays a role in the stabilization of the column, arc discharges making use of gas flow through and around the column have important technological applications. Depending on the design, application, and power range of the device, different names are in use. "Arc jets" can be used for spraying refractory materials for surface treatment or for producing special crystalline structures such as diamonds. The same applications are possible with "plasma torches," which are also used to heat and melt surface material, e.g., for welding and cutting purposes. "Arc heaters," called "Plasmatrons" in the Russian literature, are available with dc, ac, radio-frequency, or microwave power supplies. They can be used for the applications named above or for the production of high-enthalpy gas, e.g., for the test of heat shields of space vehicles returning to Earth (re-entry in an N_2–O_2 atmosphere) or entering the atmosphere of other planets with different gas compositions. Another application is the generation of high-power laser radiation. In space applications "plasma thrusters" are used for propulsion purposes; the high temperatures achievable lead to a more efficient use of the material emitted for propulsion. Additional magnetic effects are used in magneto-plasmadynamic (MPD) thrusters (Auwetter-Kurtz, 1992). In principle, all the named plasma generators use an arc discharge for producing high-enthalpy gas, which flows mainly in the direction of the arc or which is injected radially sometimes as swirl flow. Powdered material that can be heated to the melting point may be injected into the gas before or after it is influenced by the arc discharge (Fig. 10). The high-enthalpy gas or the mixture of gas and refractory material exits through a nozzle. In many applications an axial cathode is used together with a ring anode, which, in turn, may be part of the nozzle. In so-called transferred arcs used for coating purposes and for some arc welding applications, the material to be coated or worked on serves as anode. The arc is operated at high power levels, and it may be additionally confined by a cascaded wall (Buuron, 1993). In industrial torch applications it may be of interest to mix the plasma gas with dissimilar ambient

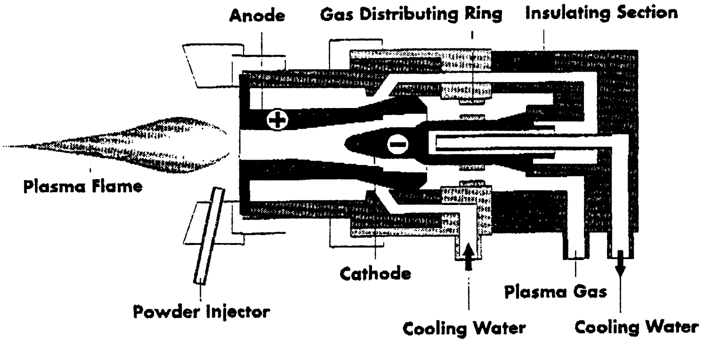

FIG. 10. Arc jet in use for spraying refractory materials (courtesy of Plasma Technik AG, Wohlen, Switzerland; reproduced by permission).

gas. The flow conditions in the arc and in the plume frequently are turbulent; a theoretical model of such torches has been developed (Murphy and Kovitya, 1993), and the comparison of the results with experimental values leads to excellent agreement.

3.2 Low-Pressure and Vacuum Arcs

The main difference between high-pressure and low-pressure arcs ($p < 0.1$ MPa) is concerned with the question of local thermodynamic equilibrium (LTE). While this condition may be nearly realized in the high-pressure arc column, discharges operated at pressures below atmospheric are characterized by a pronounced temperature nonequilibrium. Low-pressure wall-stabilized cascade arcs have been operated in the pressure range down to 1 kPa. They have been used for the determination of electronic excitation coefficients of the O II system (Awakowicz, 1991). This plasma was in a state called "corona equilibrium," where excitation occurs only from the ground level by electron collisions and deexitation is by spontaneous emission.

If a vacuum gap is used for arc initiation, different effects have to be considered. For its existence the electrical arc discharge needs gas that can be ionized and thereby forms the plasma of the arc column. If such gas initially is not present, as in the case of the vacuum-arc discharge, it has to be generated by the evaporation of electrode material. After breakdown of the electrode gap in vacuum has been achieved, in most cases it is the cathode of the discharge that is locally heated to the evaporation temperature of the cathode material. Some types of vacuum arcs exist that use evaporated anode material as arc gas (Ehrich *et al.*, 1990); in either case, vacuum arcs are usually metal vapor arcs. The current passage at the cathode typically occurs in localized zones, which are called cathode "spots." Since the spots are crucial for the vacuum arc, much work has been devoted to the understanding of their mechanism (Mesyats and Proskurovsky, 1989). On account of the small dimensions and of the nonstationary nature of spots, experimental investigations are difficult, and the physical phenomena occurring in the vacuum arc are not all and fully understood as of yet. Depending on the condition of the cathode surface, two different types of cathode spots may be present; they are called spots of type I and II, respectively (Rakhovsky, 1976; Jüttner, 1994). The craters of type I spots are small in dimension (0.1–several μm) and move at high velocities (1 m/s–several hundred m/s). Type II craters are found on clean surfaces. Their characteristic parameters are crater diameters up to several hundred μm, increasing with the temperature of the cathode; slower spot movement with several m/s velocities decreasing with increasing cathode temperature; and slightly higher erosion rates and total arc voltages compared to spots of type I. The type I spot is found on contaminated surfaces, a typical example of the contamination being an oxide layer of the cathode material. If the surface is clean or if it has been cleaned by previous arcing, type II spots are present. In the following the characteristic properties of both spot types are presented together with the known data on the related vacuum-arc discharge columns. Since the detailed spot data vary somewhat with the cathode material, only general properties and ranges of parameters can be given. The spot velocity depends also on the arc current and on the absence or presence of a superimposed magnetic field. On a stainless steel electrode in a cross magnetic field, the spot velocity showed a nearly linear dependence on the current (Fang *et al.*, 1982). Craters similar to the ones formed at the cathode of vacuum arcs are also found in high-pressure arcs. The corresponding arc-spot formation time has been determined experimentally (Nachtigall and Mentel, 1991). It is interesting to note that vacuum arcs exposed to a magnetic field move in a retrograde motion, i.e., a direction opposed to the direction of the Lorentz force $\mathbf{j} \times \mathbf{B}$, j being the current density in the arc. At the present time the question has not been resolved whether this retrograde motion is related only either to the direction of the spot movement or to that of the column, or whether it is a combination of several effects. The current-carrying capability of the cathode spots again depends on the material. Values of 10 A to several hundred amperes have been reported, the highest values being found for tungsten (Harris, 1980). Since vacuum arcs frequently are operated at very high current levels (kA), such as in the case of circuit breakers, a multitude of spots may be present simultaneously. If the current transport in the vacuum arc is mainly through type I cathode craters, the erosion rate is smaller than in the case when type II spots are present. The definition of erosion rate is mass lost per charge transported; values found extend from 10 mg/C to several hundred mg/C, the lower values corresponding to arcs with type I craters. For a discussion of the erosion rate, one has to differentiate between a net erosion of the cathode and the erosion rate of a single crater. Experiments have shown that erosion occurs not only by evaporation but also by the emission of droplets of molten material. These droplets partially fall back on the cathode surface, making the difference between net erosion in the crater and values determined by weighing after arc extinction. The total arc voltage is small as compared with high-pressure arcs. Values below 20 V are typical for arcs with type I craters. Voltage–current arc characteristics are rising, which signifies that vacuum arcs do not require a resistance in series with the arc. By solving the time-dependent conservation equations for mass, momentum, and energy for the crater volume, predictions of the formation time of an individual crater may be obtained (Prock, 1986). Ohmic heating and energy deposited by ion impact are the main energy input mechanisms; the loss terms contain electron emission, thermal conduction, and enthalpy of the material leaving the crater in the form of vapor or droplets. Spot formation times depending on the arc current result. They are in the nanosecond range, and these values compare well with measured formation times (Anders *et al.*, 1992). The plasma of the vacuum arc exhibits some peculiarities: Besides its retrograde motion in magnetic fields, the mechanisms responsible for the very high energies of ions leaving the cathode (velocities of up to 10^4 m/s have been observed) are not fully understood.

Vacuum arcs are of substantial technological interest. High-power circuit breakers for the medium-voltage range ($V < 80$ kV) make use of the fast "resolidification" of the electrode gap with respect to electrical breakdown after current zero when the ac-voltage builds up again. Resolidification signifies that the electrode gap recovers its capability to withstand high breakdown voltages; this effect is due to a fast decrease in the number of charged particles in the electrode gap of vacuum arcs. During the high-current phase, the arc in circuit breakers exists with a multitude of cathode spots, and self magnetic fields are being used to force the spots to move from the center of the electrode toward the edge in a radial direction. Fast-moving spots that disappear when they have reached the electrode edge lead to low electrode erosion, thus preserving the vacuum and diminishing the probability of an unwanted breakdown in the vacuum circuit-breaker gap. Another potentially important application of vacuum-arc technology is related to controlled nuclear fusion. In fusion reactors—e.g., in the JET tokamak operated in Culham, Great Britain—traces of so-called "unipolar arcs" are found on walls that are in contact with the fusion plasma (Behrisch, 1986). Vacuum arcs generating the arc gas by anode erosion are used for surface-treating techniques such as producing coatings.

4. SPARKS

The spark discharge is a nonstationary, high-current discharge at atmospheric pressure or above. In all spark cases there holds for the product of pressure (p) and electrode gap (d) $pd \geq 1$ kPa m; this requires interelectrode distances exceeding several centimeters. The breakdown mechanism for this discharge type may be described by the avalanche–streamer–leader model. If a dc voltage is applied to an electrode configuration, free ions and/or electrons that are either present (e.g., due to natural radioactivity or cosmic radiation) or generated (e.g., by field emission) are accelerated—in the case of negative ions and electrons—toward the anode. On their way they experience elastic and inelastic collisions, thereby creating new electrons through ionization processes. Assuming the electrical field strength to exceed a minimum critical value given by the condition that the rate of production is higher than the rate of electron attachment, a primary and successive avalanches are formed (Fig. 11).

Atoms excited by collisions emit photons, which enhance the formation of secondary avalanches by photoionization processes. The avalanches move toward the anode, and

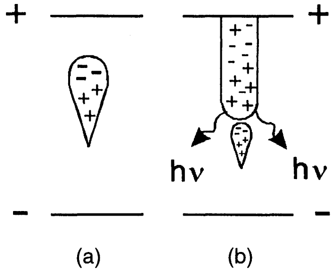

FIG. 11. Breakdown by the avalanche–streamer process between anode and cathode. (a) Avalanche, (b) streamer advancing toward the cathode by interaction with a secondary avalanche, created by photons.

the original field is distorted by the space charge generated in the avalanches. If the amplification of the electron generation process is high and if the total number of electrons contained in the avalanche exceeds $N_{\mathrm{crit}} = 10^8$, a streamer is formed. It consists of a weakly ionized, thin plasma channel and usually grows in the direction of the cathode. The streamer advances probably through the following mechanism: Secondary avalanches that reach the positive space charge at the streamer tip neutralize it and thereby create a new positive space charge located further in the propagation direction of the streamer. Thus, photons play a role in streamers directed toward the cathode. Once the streamer reaches the cathode, free electrons are created at a higher rate, and they are accelerated in the field created by the streamer. A return wave of high ionization propagates with high velocity (10^7 m/s) toward the anode, and the spark channel is formed. Anode-directed streamers are also possible; in this case the propagation is field enhanced and is similar to the propagation of an ionization wave. The ignition process via the described streamer mechanism is much faster than the growth velocity of avalanches or the drift of ions would allow; the streamer velocity exceeds 10^6 m/s. If the electrode gap is very extended, $d \geq$ several meters, the avalanche–streamer process is replaced by the avalanche–streamer–leader process. The leader is a narrow channel containing plasma with a high degree of ionization and is formed in the wake of streamers. Experimental work (Gallimberti, 1977) on streamers has shown that a large amount of the Joule heat deposited by the current is first stored as vibrational energy in molecular gases. Once this energy is relaxed into thermal energy, it increases the electrical conductivity of the plasma. The time constant for this process depends on the nature of the gas, the pressure, and the humidity, the order of magnitude being 10^{-5} s. The velocity of propagation of the leader is similar to the values obtained for streamer propagation. Once the spark channel is formed, current density, field strength, temperatures, and number densities are identical to the ones found in arc discharges. The difference from the arc lies in the fact that usually the spark discharge is too short to allow a heating of the cathode, which would lead to thermionic emission. Applications of the spark are found in radiation sources and for switching purposes (spark gaps).

5. GLOW DISCHARGES

5.1 Mechanism and Electrical Characteristics

Figure 12 shows schematically the voltage–current dependence of a number of different discharges between electrodes. The type of gas may vary as well as the pressure levels.

If the electrical field between the electrodes exceeds values typical for the given gas, Paschen breakdown occurs: $U > U_{\mathrm{P}} = f(pd)$, p being the gas pressure and d the distance between (plane) electrodes. With a field strength high enough, collision processes lead to an avalanche and to the ionization of the interelectrode space. The so-called Townsend breakdown takes place at

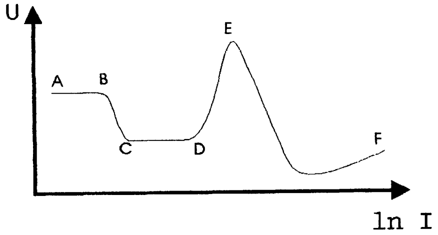

FIG. 12. Voltage–current characteristic of discharges. AB: Townsend dark discharge; CD: normal glow discharge; DE: abnormal glow discharge; EF: arc discharge.

pd values up to approximately $pd = 250$ Pa m. A self-sustained discharge is initiated if

$$\exp[(\alpha - \eta)d] \geq (\alpha - \eta)/\alpha\gamma + 1, \tag{34}$$

where α = first Townsend coefficient for ionization, η = electron attachment coefficient, and γ = coefficient for electron generation by ion impact at the cathode surface.

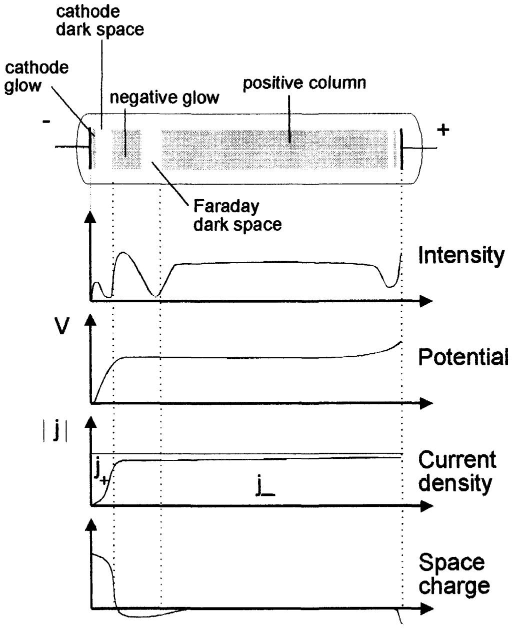

FIG. 13. Schematic drawing of the normal glow discharge.

While the Townsend dark discharge with currents in the range $I < 10^{-5}$ A owes its name to the (near) absence of visible radiation due to the small degree of ionization, the normal glow discharge shows different light-emitting zones. All portions of the *U-I* characteristics in Fig. 12 belong to self-sustaining discharges. While the arc discharge draws its electrons from the cathode via thermal emission or through the combination of thermal and field emission, glow discharges typically operate with cold cathodes. The release of electrons at the cathode is mostly due to secondary emission resulting from the bombardment with positive ions. The pressure range in which such discharges are operated extends from 1 Pa to approximately 1 kPa. In recent years a number of Russian publications (e.g., Raizer, 1991) also described glow discharges at much higher pressures reaching up to several times 10 kPa. Figure 13 shows the most important regions of a normal glow discharge in a simplified and schematic drawing. Let us start with a brief discussion of the properties and phenomena of the normal low-pressure glow discharge. Their voltage–current characteristics correspond to the portion CD of Fig. 12. Such discharges are typically operated with diameters of several centimeters and with a length up to 1 m. Current levels extend up to 0.1 A, and voltages of 10^2–10^3 V result. As shown in Fig. 13, different zones can be distinguished optically along the discharge column; this is because they are either dark or light-emitting spaces. Starting the description at the cathode-fall region, a first narrow dark space named Aston dark space separates the cathode from the cathode glow. Since the electrons leave the cathode with low energies (<1 eV), they have to be accelerated by the electrical field before they can excite atoms (Fig. 14). Emission from atoms that may be in different excited states leads to the cathode glow identified in Fig. 13. These details have been omitted in Fig. 14, which explains only the principal mechanisms in the formation of a glow discharge. Once the electron energy reaches a level where the excitation cross sections decrease significantly, the cathode dark space is formed. In the zone between the cathode and the negative glow, the current density is mainly due to the ionic component, and a positive space charge results. In the negative glow, ionization of atoms is due to inelastic collisions with electrons, and the electric field decreases. The Faraday dark space is due to the fact that most electrons have dissipated their energy. The positive column that follows is a quasineutral plasma with a low degree of ionization and a strong deviation from thermal equilibrium (LTE) be-

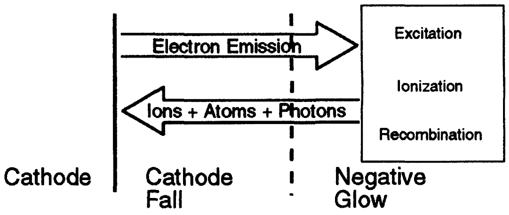

FIG. 14. Cathode region of the glow discharge.

tween electrons and heavy particles, its luminosity being due to new excitation processes of accelerated electrons. The velocity distributions of electrons and heavy particles are Maxwellian but correspond to differing temperatures (several electron volts for the electrons, very low energies for the atoms and ions). The particle number density of electrons extends to values of $n_e < 10^{12}$ cm^{-3}; this is very low compared to the number densities found in arc discharges and is at the origin of the non-LTE condition. At the anode the electrons disappear from the plasma, ions are repelled, and a positive space charge results (anode fall region). The theory of the cathode-fall region was developed very early (Von Engel and Steenbeck, 1934). It allows precise predictions for different gases. Instabilities of the positive column may occur, and they can prevent the use of glow discharges in applications such as gas lasers, e.g., CO_2 lasers. Besides striations of the column, filamentation corresponding to high current densities due to contractions have been observed. Causes of instabilities include thermal effects as well as field and electron temperature disturbances. The corresponding stability theory has made substantial progress over the last years. Moving striations with velocities in the range of 100 m/s have found their explanation by considering ionization waves moving through the plasma.

For laser as well as for pollution-control applications, it is desirable to increase the discharge current to values of at least 1 A. An increase in gas pressure to levels close to atmospheric pressure has also been achieved. At higher pressures the glow discharge becomes unstable; filamentation occurs, leading to a strong nonuniformity of the plasma. It is possible to use high-pressure glow discharges in a number of interesting applications. For this purpose either the current may be pulsed or a gas flow can be used to stabilize the column; in addition it may then be necessary to segment the cathode so as to maintain a constant column diameter. Pulsed discharges have shown great potential for laser applications. In particular, rare-gas halide lasers—e.g., XeCl lasers with preionization by x rays—are under intensive investigation (Bötticher, *et al.*, 1992). Increasing the gas pressure leads to a decrease in length of the different discharge zones described above. The cathode layer, negative glow, Faraday dark space, and anode layer shorten inversely proportionally to the gas pressure. The dependence of current density at the cathode on gas pressure, $j \propto p^2$, is still valid at pressure levels up to several atmospheres, but instabilities of the cathode layer may occur, leading to cathode-spot formation. Experimental investigations performed in nitrogen and air (Napartovich and Akishev, 1993) demonstrated the importance of energy-transfer processes between molecular vibrational levels and electronic levels in such discharges. The applications of high-pressure glow discharges range from the excitation of the active medium of lasers in CO_2, CO, and Ar^+ to the removal of pollutants such as SO_2 and NO_x from air (Akishev *et al.*, 1993). In the technologically interesting case of high-power CO_2 lasers in the multikilowatt range used for welding or cutting, high gas-flow rates of the gas mixture of $CO_2 + N_2 +$ He provide the removal of the excess power. If the current in the glow discharge is increased beyond the value at which the glow covers the whole cathode, the portion of the voltage–current characteristic that is to the right of point D in Fig. 12 applies. In this case an abnormal glow discharge is produced as a transition to the arc discharge. The current density at the cathode rises in this case; values of j in the range 10–10^2 A/cm^2 have been observed. Optically, the abnormal glow discharge exhibits a shorter cathode dark space, and the Aston dark space directly in contact with the cathode disappears. The cathode fall increases to very high values (several kilovolts are possible), and the voltage–current characteristic acquires a positive slope. The electrons leaving the cathode are accelerated in the cathode fall to high energies, but after collision with atoms and ions, low-energy electrons are produced in the negative glow. Ions from the negative glow are accelerated toward the cathode, and charge-exchange processes with neutrals take place in the cathode-fall region. Electron emission results from the impact of ions and neutrals on the cathode surface. Additional energy transfer to the cathode may result from radiation and the release of excitation energy.

5.2 Hollow-Cathode Discharges

Before discussing the so-called pseudospark and its modern applications, it is necessary to consider the hollow-cathode effect and the cold hollow-cathode discharge. A good description of the hollow-cathode effect is given in the recent literature (Gundersen and Schaefer, 1990). By replacing the plane cathode of a glow discharge by a hallow cathode of cylindrical or slit geometry and by operating the discharge in a specific range of parameters, a condition can be obtained where the negative glow is located on the inside of the hollow cathode. Such discharges operate at lower voltages and much higher currents than the normal glow discharge. The pressure range in which the hollow-cathode effect occurs may be expressed by the product *pd*, where *p* is the gas pressure and *d* is the inner diameter of the hollow cathode. In rare gases this range extends between 1 Pa m $< pd <$ 10 Pa m, while molecular gases have lower corresponding values. In a hollow-cathode configuration, the cathode fall increases only weakly with the current density. If the cathode fall is kept at a constant value, the current density increases strongly with decreasing values of *pd*. The cathode fall reaches a minimum for a particular value of *d* at given pressure and current density. The hollow-cathode effect is very sensitive to the geometrical dimensions (length and aperture of the hollow structure) as compared to the mean free path length of electrons. Several mechanisms contributing to the hollow-cathode effect are directly associated with the hollow geometry; they are as follows:

1. There is a higher probability of photons and excited atoms leaving the negative glow on the inside of the structure to reach the cathode surface.
2. Electrons leaving the cathode surface that, after acceleration in the cathode fall, reach the opposite cathode surface are reflected and become pendular electrons, changing the potential distribution in the cathode fall, and increase the ionization rate in the negative glow.
3. Charge-transfer collisions are less likely in the cathode fall under hollow-cathode conditions since the fall length is reduced, as compared to a plane geometry. The results are impacting ions with higher energies and an increase of secondary electrons.
4. The density of sputtered cathode material increases as well as the plasma density inside of the hollow structure. Since the sputtered metal atoms have a lower ionization potential, plasma density is further increased, and multistep processes become more likely.

Different electrode configurations are described in the literature; cylindrical or spherical cathodes have been used as well as several plane structures forming a cavity. By pulsing the current of a hollow-cathode discharge, very high values can be obtained before the cathode heats up, leading to thermal emission and to an arc discharge. Current values of more than 1000 A with current densities up to several kA/cm^2 and with voltages of only several hundred volts have been reported for so-called "superdense cold hollow-cathode glow discharges." Hartmann and Gundersen (1988) described a possible mechanism for this anomalous cathode emission. The applications of hollow-cathode discharges range from high-current closing switches and hollow-anode thyratrons to laser applications and to the production of intense electron beams for the excitation of cw lasers and for material processing. Of particular interest are the metal-ion laser discharges using hollow cathodes, since laser action can be obtained in the UV and VUV spectral region, and even radiation standards for this region have been developed (Warner and Persson, 1979; Gerstenberger *et al.*, 1980; and Danzmann *et al.*, 1988).

5.3 Pseudosparks

A special type of self-sustained hollow-cathode glow discharge has been given the name "pseudospark" because of the high currents and current densities that are used in the discharge for short periods of time but in a repetitive mode for most applications. Four phases in time can be distinguished in a pseudospark discharge:

1. ignition,
2. development of the hollow-cathode discharge,
3. high-current discharge, and
4. decay of the discharge.

Let us consider the basic pseudospark geometry as described in the original paper (Christiansen and Schultheiss, 1979). Two metal plates with a central borehole are separated by an insulator and serve as cathode and anode, respectively (Fig. 15). The diameter of the central hole and the plate separation are small as compared to the diameter of the plates. The insulator ring prevents the discharge from operation in the exterior space (accordingly, d_{eff} is the longest possible electrical field line). For parallel plane electrode geometries, the breakdown curve shows a minimum for pd values around 1 Pa m, p being the gas pressure and d the electrode spacing. If an electrode configuration is used that is similar to the one shown in Fig. 15, a high-current pseudospark discharge can be established within a very short time and at pressures below the minimum of the breakdown curve. The applied electrical field causes values of $E/n \approx 10^{-11}$ V cm^2 in the electrode gap, E being the electrical field strength in V/cm and n the particle number density in cm^{-3}. Electrons coming from the wall of the cathode are accelerated and cause a strong avalanche growth in a very short time (10^{-7}–10^{-8} s). At a discharge current around 10 A, the plasma is formed in the hole region between the electrodes, i.e., in the hollow cathode. A space-charge–neutralized narrow electron beam with very high current densities (10^4–10^5 A/cm^2) leaves the device through the hole on the anode side; this phase lasts 10–30 ns. The high-current discharge that follows is a superdense glow discharge whose duration is determined mainly by the parameters of the external circuit. The internal resistance of the pseudospark is low and decreases with an increase in current. High currents with a short rise time can be carried as a result of a high conductivity. These properties are useful for high-power switches. However, limitations exist regarding the breakdown voltage levels, but for a single-gap system operating with hydrogen, a range up to 40 kV seems possible. Data given for a high-current pseudospark switch with tungsten electrodes are breakdown voltage exceeding 7–18 kV, maximum current 150 kA, lifetime 10^6 discharges, and trigger with a jitter of less than 20 ns. By cascading many single-gap systems in series, it is possible to create an intense, medium-energy electron beam whose small diameter is due to the pinch effect. The applications of pseudosparks range from materials processing to the excitation of laser media.

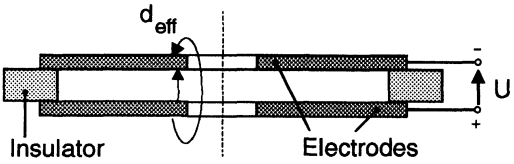

FIG. 15. Schematic drawing of a single-gap pseudospark electrode configuration.

6. HIGH-FREQUENCY DISCHARGES

If breakdown conditions of the gas are met for the frequency range of interest and if power supplies are available for the frequency under consideration, high-frequency discharges can be operated in the whole frequency range from ac to optical frequencies. In the latter case the discharge is called optical discharge, the power of such discharges being supplied, for instance, by a CO_2 laser.

The interaction of an electromagnetic wave with a plasma can be described by solving Maxwell's equations together with the equations of motions of electrons and ions in the E and B field components of the wave. Assuming a plane harmonic wave of frequency ω to travel through a low-pressure plasma, a simple description of the phenomena can be given. Considering the ratio of magnetic and electric forces of a charged particle (charge q) traveling with a velocity v in a direction perpendicular to the B vector of the wave (c being the vacuum velocity of light),

$$\frac{|\mathbf{F}_{\text{magn}}|}{|\mathbf{F}_{el}|} = \frac{qvB}{qE} = \frac{v\mu_0 H}{E} = v\mu_0\sqrt{\frac{\epsilon_0}{\mu_0}} = v\sqrt{\epsilon_0\mu_0} = \frac{v}{c}, \tag{35}$$

it is clear that the magnetic force can be neglected as long as $v \ll c$. Solving the wave equation for the case of negligible friction losses due to collisions leads to the following equation for the index of refraction of the plasma:

$$n^2 = 1 - \omega_{pe}^2/\omega^2, \tag{36}$$

where $\omega_{pe} = (n_e e^2/\epsilon_0 m_e)^{1/2}$ is the plasma frequency of the electrons (n_e and m_e are electron number density and electron mass). The ions may be neglected by reason of their high mass. Only if the plasma frequency exceeds the frequency of the wave in Eq. (36) does wave propagation result. In general, collisions have to be included in the considerations; this leads to a complex index of refraction. Since the plasma is generated by the absorption of electromagnetic energy, the high-frequency discharges have the great advantage of working without electrodes. This allows the use of aggressive gases in the corresponding reactors.

6.1 Microwave Discharges

6.1.1 Electron-Cyclotron Resonance Discharges By virtue of the simplicity of the apparatus and of the wide range of applications, this discharge type has great potential. Its main characteristic is a steady-state magnetic field being superimposed on the discharge that results from the absorption of electromagnetic energy deposited either directly or via capacitive or inductive coupling. Assuming the direction of the static magnetic field to coincide with the direction of wave propagation, two solutions of the wave equation are obtained: a right-hand (R) and a left-hand (L) polarized wave. The complex index of refraction may then be approximated by (Wilhelm, 1993)

$$N_{L/R} \approx \sqrt{1 - \frac{\omega_{pe}^2/\omega^2}{(1 \pm \omega_{ce}/\omega) - i v/\omega}}, \qquad (37)$$

where $\omega_{ce} = eB/m_e$ is the electron gyrofrequency and v is the collision frequency. Without magnetic field and neglecting collisions (i.e., $v/\omega \approx 0$), Eq. (37) becomes equal to Eq. (36). With magnetic field, two characteristic cases may be distinguished for the index of refraction of the right-hand polarized R wave:

1. $N_R \rightarrow \pm\infty$ for $\omega = \omega_{ce}$; i.e., for a certain value of B, this corresponds to the resonance case, phase velocity and wavelength $\Rightarrow 0$
2. $N_R = 0$ for $\omega_{pe}^2/\omega^2 = (1 - \omega_{ce}/\omega)$; i.e., for a certain value of n_e, this corresponds to the "cutoff" case, phase velocity and wavelength $\Rightarrow \infty$

Wave propagation is not possible in the evanescent region where the index of refraction becomes imaginary ($N^2 < 0$). The dispersion conditions for the R wave can be represented in a CMA diagram, Fig. 16 (taken from Wilhelm, 1993). The CMA diagram is a representation of the dispersion relations for an electromagnetic wave in a magnetized plasma; it is named after the initials of its authors: P. C. Clemmow, R. F. Mullaly, and W. P. Allis. The abscissa of this diagram depends on electron number density, whereas the ordinate depends on the magnetic induction. The dashed line in the diagram separates the cutoff region from the evanescent region. In the case of R resonance, i.e., if absorption is to be achieved, curve (a), access is possible from two sides. The "low-field access," curve (b), can be achieved using permanent magnets, but the wave has to tunnel through the evanescent region in front of the resonance. The "high-field access," curve (c), may be difficult in cases where the real part of the index of refraction takes on high values, because of steep magnetic field gradients and high plasma densities.

The power absorbed by the electrons is given by

$$P_{R+L} = \frac{n_e e E_0^2}{2 m_e v}\left[\frac{1}{2}\left(\frac{v^2}{(\omega - \omega_{ce})^2 + v^2}\right) + \frac{1}{2}\left(\frac{v^2}{(\omega + \omega_{ce})^2 + v^2}\right)\right], \qquad (38)$$

where E_0^2 is the amplitude of the wave. The first part in the bracket corresponds to the R

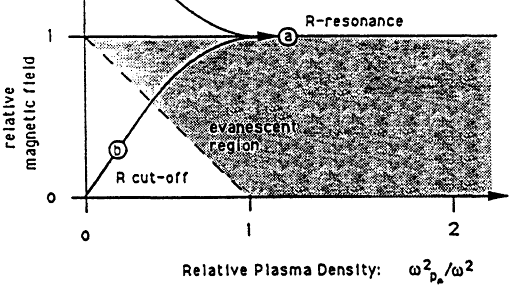

FIG. 16. CMA dispersion diagram (from Wilhelm, 1993).

wave, while the electron cyclotron resonance (ECR) condition is obtained for $\omega = \omega_{ce}$.

The ECR reactors are used for coating, surface treatment, and microstructuring by reactive ion etching. A "standard reactor" is shown together with the magnetic field topology in Fig. 17 (taken from Wilhelm, 1993). This type of reactor cannot be used for large-scale industrial operation, but it is relatively simple and flexible, and therefore it is found in many applications, such as providing carbon coating on the first wall parts (e.g., limiter plates or tiles) of nuclear fusion machines. The cylindrical vessel of the reactor contains the sample holder; it is enclosed by magnetic coils. For better penetration of the ions coming from the plasma, the sample may be subjected to a dc or rf biasing. The *R* wave enters the vessel from the top via a circular horn antenna located behind a quartz window. The pressure range for these discharges lies between fractions of a pascal and several hundred pascals. Frequently, readily available magnetron sources with a frequency of 2.45 GHz and with power levels up to 6 kW are used. If large surfaces have to be treated in mass-production applications, other reactor types, such as the "distributed ECR reactor" (Pomathiod *et al.*, 1984) or a "reactor with elongated mirror geometry (EMG)" (Geisler *et al.*, 1990), are used.

A modification of the ECR reactor is the "electron cyclotron wave resonance reactor" (ECWR). It uses frequencies that are only close to the ECR condition. The increase of the index of refraction in the vicinity of ECR causes a strong reduction of the wavelength in the plasma. This leads to a condition where the wavelength becomes smaller than the plasma dimension. Since the damping of the wave decreases for $\omega < \omega_{ce}$, the wave penetration into the plasma increases, and a standing wave may be obtained in the reactor. The energy introduction into the vessel can be obtained either by capacitive or by inductive coupling. For the case of inductive coupling, one or several ring electrodes are wound around the discharge vessel, and the reactor functions like a transformer whose secondary winding is the plasma itself. Such reactors can be used from the low-pressure range up to atmospheric pressure. In the Russian literature they are called plasmatrons, and the range of frequencies used extends from the microwave frequencies to radio frequencies. Important recent developments make use of microwave plasmas for lasing purposes (CO_2 and excimer lasers) and for excimer lamps.

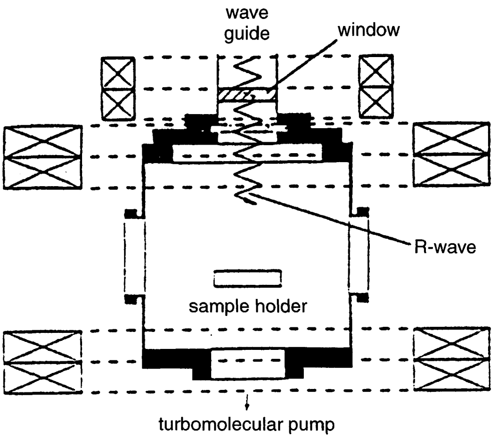

FIG. 17. "Standard ECR reactor" and magnetic field topology (from Wilhelm, 1993).

6.2 Radio-Frequency Discharges

Discharges produced by electromagnetic fields in the radio-frequency (rf) range, i.e., approximately between 100 kHz and 100 MHz, obtain the deposited power either by capacitive or by inductive coupling. While in the inductively coupled discharge as described above no electrodes are used, this may or may not be true for the case of the capacitively coupled discharge. Analytical solutions for the prediction of voltage–current characteristics have been worked out for the electrodeless case as well as for the rf discharge using electrodes (Raizer, 1991). Depending on the gas pressure at which the discharge with electrodes is operated, two different modes of the discharge have been observed, specifically in hydrogen. At pressures below 100 Pa, the light emitted from the discharge is higher in intensity and comes from distinct layers reaching close to the electrodes, whereas rf discharges in H_2 operating at pressure levels above several hundred pascals exhibit a diffuse glow in the center of the gap, the electrode regions being dark. The mechanisms leading to these different discharge modes are not fully understood as yet, but it has been shown that the current densities differ by an order of magnitude for the two cases (Yatsenko, 1981).

Because there is an increased absorption of wave energy in the plasma if a helicon wave is used, discharges of this type have become an interesting and important application of rf discharges. Helicon waves are right-handed circularly polarized waves that propagate in the direction of the superimposed magnetic field and have an electric field component parallel to this B field; see Fig. 18 (taken from Chen, 1991). This type of discharge is operated at a frequency ω that is given by the condition (Komori, *et al.*, 1991) $\omega_{ci} \ll \omega_l \ll \omega \ll \omega_{ce} \ll \omega_{pe}$. The quantity $\omega_{ci} = qB/m_i$ is the ion cyclotron frequency, q being the charge; $\omega_l \approx [\omega_{pi}^2/(1 + \omega_{pe}^2/\omega_{ce}^2)]^{1/2}$ is the lower hybrid frequency, ω_{pi} being the plasma frequency of ions; and $\omega_{pi} = (n_i q^2/\epsilon_0 m_i)^{1/2}$. Depending on the electron number density in the low-pressure (0.1-Pa range) discharge, the role of collisional damping versus Landau damping may be altered, smaller electron number densities ($n_e < 10^{13}$ approximately, in a 7-MHz discharge operated in argon at a pressure of $p < 0.4$ Pa) favoring Landau damping.

Important applications of rf discharges are, among others, the heating of fusion plasmas and the production of high-power

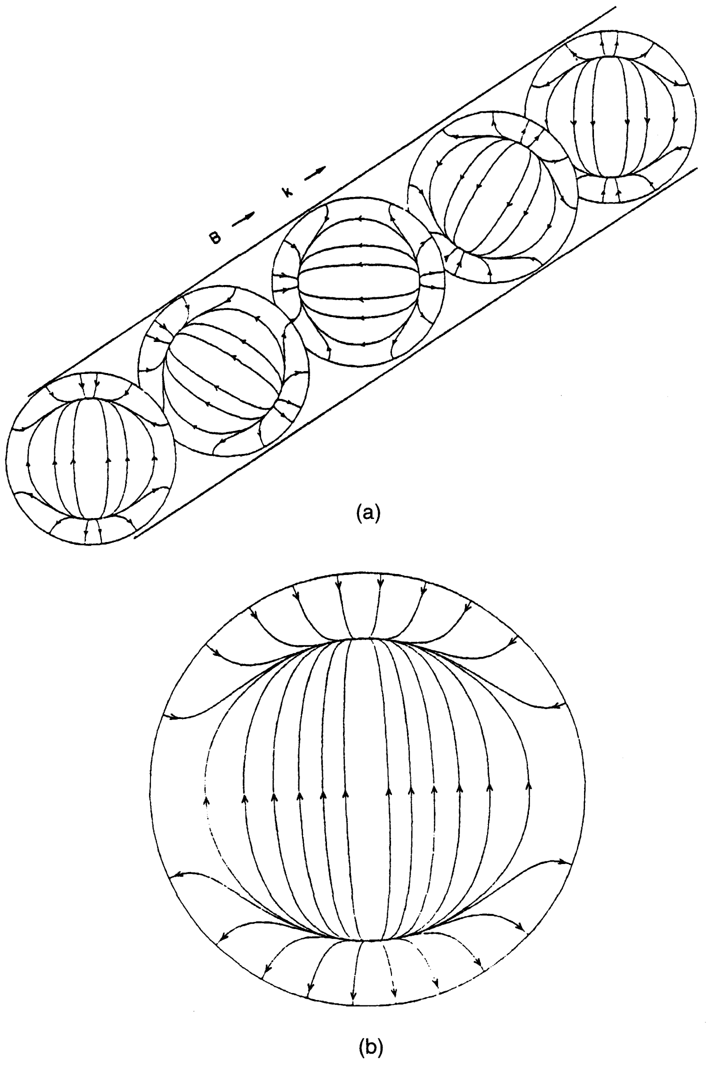

FIG. 18. Electric field line patterns for the $m = 1$ mode of a helicon wave. (a) 3D representation; (b) detailed pattern for $k/a = 1/3$ (from Chen, 1991).

CO_2 laser plasma, of plasmas used for sputtering purposes, and of plasmas used for technological plasma–wall interaction processes.

6.3 Dielectric-Barrier Discharges

The dielectric-barrier discharge (DBD), which is also called "silent discharge," was proposed by Siemens as early as 1857. In recent years it has acquired considerable attention because of its numerous applications. This type of discharge can be operated at different pressure levels (0.01–1 MPa) and frequencies (50 Hz–1 MHz), and a wide range of different gases is used. Typical electrode configurations of DBD's are shown in Fig. 19 (from Kogelschatz, 1992). In all three configurations, which might be used for different applications, a dielectric barrier separates at least one electrode from the discharge space. As a result mainly of the work performed by Eliasson and co-workers (Eliasson and Kogelschatz, 1991; Eliasson *et al.*, 1994), modeling of DBD has made great progress during the last decade. Considering a DBD in a high-pressure gas, the discharge consists of many current filaments existing simultaneously and reforming after extinction. The number of filaments may be high enough to lead to the appearance of a fairly homogeneous discharge in the gap. In a gap with a spacing of the order 1 mm, the radius of the filaments is typically around 0.1 mm, and the current density lies in the range 100–1000 A/cm^2. The electron density in the gap is of the order 10^{14}–10^{15} cm^{-3}, and the mean electron energy is between 1 and 10 eV. Three distinct phases can be distinguished for each filament:

1. the electrical breakdown of the gap, i.e., the formation of the discharge;
2. the passage of the current pulse; and
3. the reaction kinetics in the gas.

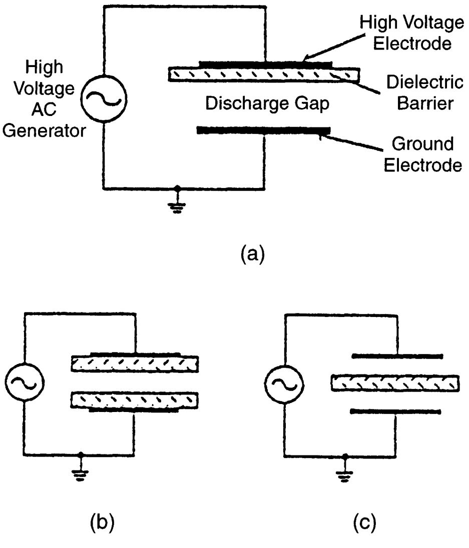

FIG. 19. Typical electrode configurations of dielectric barrier discharges (from Kogelschatz, 1992).

The breakdown occurs if the field inside the gap exceeds the reduced Paschen field, i.e., the so-called Paschen voltage, which corresponds to the minimum of the Paschen curve, divided by the product of number density in the gap and the extent of the gap. An avalanche forms, and at its head a considerable space charge leads to a local self-induced field ("eigenfield") superimposed on the applied field. Thus the field is enhanced in the direction of propagation toward the anode. After the space charge reaches the anode, a high-field wave may be reflected back toward the cathode. This process is completed within nanoseconds; it leaves a discharge channel through which the current pulse occurs. Current transport will take typically 1–100 ns. If there is a dielectric in front of the anode, the charges accumulate on it. This leads to an electric field that is directed against the applied field, and the current flow stops in this filament. If the field is high enough, a second current pulse may be generated; this has been observed in a high-pressure oxygen discharge. During and after the passage of the current, the molecules and atoms go into excited states. Reaction kinetics, i.e., plasma chemistry, takes place, the time scale of which may last up to seconds. Numerical modeling of this process has been successful. By solving the Boltzmann equation (7) together with the continuity equation (for all charged particle species) and Poisson's equation, the relevant features of such discharges can be predicted. Among the most important and oldest applications of such discharges stands the generation of ozone from oxygen or air according to $e + O_2 \rightarrow e + 2\ O$ followed by the reaction $O + O_2 + M \rightarrow O_3 + M$, where M is a collision partner such as O_2, O, O_3, or N_2.

Recently, much work has been devoted toward using DBD for the formation of excimers to be used in UV and VUV radiation sources. Further applications are the pumping of CO_2 lasers, surface treatment (deposition and dry etching), pollution control by the destruction of toxic materials, and, last but not least, CO_2 hydrogenation for the formation of methanol according to the reaction $CO_2 + 3\,H_2 \rightarrow CH_3OH + H_2O$.

7. CORONA DISCHARGES

If the electrical field between electrodes is inhomogeneous, e.g., between a pointed and a plane electrode, a corona discharge may be initiated at one electrode, for example, at the pointed electrode. Depending on the polarity of this electrode, different phenomena are observed. As described in Secs. 4 and 5, a voltage U applied to an electrode gap leads to different types of breakdown mechanisms and discharges, depending on the type of gas, the pressure, the electrode spacing d, and the value of the voltage U. Only if the field strength is high enough to insure multiplication of electron production processes will a self-sustaining discharge start. Let us consider the case of a dc voltage being applied to an electronegative gas like air. The electrode at which the corona discharge should be observed is the negative electrode (negative corona case). Ignition of the corona discharge is only possible if the following ignition criterion is fulfilled:

$$\int_0^P (\alpha - \eta)\,ds = \ln(1 + 1/\epsilon), \qquad (39)$$

where α and η have been defined in Eq. (34) and ϵ is the secondary Townsend coefficient relating the number of electrons coming from the cathode to the corresponding number generated by collision processes in the gas. The integration over the distance from the electrode has to be extended to the point P where the production of electrons ends because $\alpha = \eta$. Equation (39) is similar to the streamer-formation criterion:

$$\int_0^P (\alpha - \eta)\,ds = \ln N_{\mathrm{crit}}, \qquad (40)$$

where N_{crit} is the critical number of electrons in the avalanche, $N_{\mathrm{crit}} = 10^8$; see Sec. 4. Accordingly, the phenomenological description of a negative corona at increasing voltage reaches from a stage (a) in which the space close to the electrode surface shows constantly moving luminous fans, to a glow mode (b), and finally to a prebreakdown streamer (c). The discharge of stage (a) exhibits intermittent short (tens of microseconds) current pulses that have values below a milliampere; they are called Trichel pulses. The processes leading to the Trichel pulses are related to space charge and screening phenomena. Since their frequency is in the range of radio and television signals, they can lead to noise. In stage (b) the glow extends into the gap; the current is more stable and has amplitudes reaching up to 100 mA. If the discharge is of the prebreakdown streamer type (c), a bright streamer of several centimeters in length can be seen in the gap, and the current has pulses in the millisecond range and amplitudes up to amperes. If the dc polarities are reversed, the corona discharge appears at the anode (positive corona). The stages (a)–(c) described for the negative corona exist also; now (a) is called "burst corona," a plasma film that covers the electrode; the glow (b) extending into the gap has higher currents, and the streamer stage (c) exhibits a bright filament with more branching and longer than in the negative corona case. The current pulses have approximately the same duration and amplitude as before. If corona discharges with ac voltages are considered, the effects observed in each half-cycle correspond to the dc effects described above. The only difference between dc and ac corona discharges is that the space charge remaining from the previous cycle leads predominantly to negative-corona Trichel pulses (a) and to positive glow (b) and prebreakdown (c) coronas. Corona discharges have a negative effect on high-voltage power transmission, since the power losses associated with these discharges between power lines may reach the same order of magnitude as the Ohmic losses. On the other hand, corona discharges have applications in Geiger–Müller counters and in electric filters and separators (Goldman and Goldman, 1978).

8. LIGHTNING PHENOMENA

The natural phenomena "lightning" and "Saint Elmo's fire" must always have impressed man, as documented in numerous archeological representations; see, e.g., Poroda (1949). In later times disastrous lighting strokes led to the explosion of stored gun powder, destroying houses and killing people, such as in Brescia in 1769 where 3000 deaths resulted. As described in Golde (1977), early research in different countries was carried out. Once the electrical nature of lightning was understood, man searched for protection. In 1750 B. Franklin proposed the lightning rod. It first proved its usefulness when, in 1760, lightning struck the house of the merchant West in Philadelphia. In recent years lightning recording and research stations have made direct and time-resolved recordings of the current pulse, the electrical and magnetic fields, and the optical phenomena associated with lightning phenomena. From these experimental data and from modeling calculations, the following conclusions may be drawn. At thunderstorm times, the always present natural electrical field with a field strength of 100–400 V/m, which is due to the negative charge of the earth (5×10^5 C) and a corresponding positive charge in an atmospheric layer at a height of approximately 100 km, is disturbed. The disturbance in the lower atmosphere is caused by charge separation in clouds. Different mechanisms, not fully understood as yet, are at the basis of that charge separation: Because of strong convection in the atmosphere, turbulence occurs in clouds, precipitation droplets polarized by the electric field collide, and charge transfer and charge separation may be the consequence. In addition, convective electrification mechanisms may play a role: The cloud may acquire charge by taking in both positive ions from point discharges on the ground and negative ions from the upper atmosphere. Another electrification mechanism that has been proposed is connected with the presence of supercooled rain droplets together with ice crystals. As a result of the electrification, a cloud dipole is formed; maximum field strengths are in excess of 4×10^5 V/m, and charges of many Coulombs are reported. The typical cloud-to-ground stroke starts with a stepped leader, the first one having a step length of 10–200 m, a time interval between steps of typically 50 ms, and a velocity of propagation of approximately 10^6 m/s, the branches having a downward direction. Figure 20 (taken from Berger in Golde, 1977, p. 136) schematically represents the different stages in the formation of the stroke. Once the downward-directed leader of either polarity

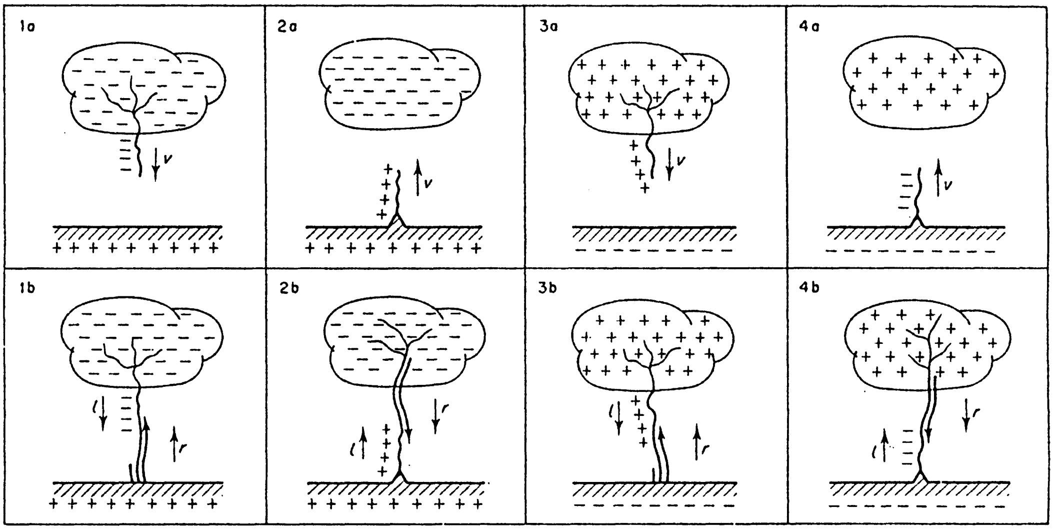

FIG. 20. Schematic representation of the eight types of lightning flash. Types depend on polarities and directions of leader and return strokes: *l*, leader; *r*, return stroke; *v*, direction of propagation (from Golde, 1977, p. 136).

is close to ground, an upward streamer is generated, and a complete discharge channel is formed. From the junction point, an upward-directed ground potential wave travels with a speed on the order of 10^8 m/s. The so-called return stroke carries tens of kiloamperes in an arc channel with the properties described in Sec. 3; its shock wave is at the origin of the thunder. The same formation phases can be identified in an upward lightning stroke. Negative first strokes typically carry a charge of several coulombs. Different return-stroke current models exist—no final theory has been formulated (Uman, 1990). Once the current has stopped, there remain free charges in the extinguished channel for times extending from milliseconds to seconds. If additional charges are available, a subsequent stroke may start, using, approximately, the previous path. Besides strokes of the described type, there exist ground-to-cloud, cloud-to-cloud, and intracloud strokes. Depending on the optical appearance of the stroke channel, different names, such as bead lightning, are given to the phenomenon. Ball-lightning events have also been reported without a final explanation.

GLOSSARY

Abel Equation: Analytical solution of the radiation transport equation for cylindrical nonabsorbing media.

CMA Diagram: Representation of the dispersion relations for an electromagnetic wave in a magnetized plasma. The diagram is named after the initials of its authors: P. C. Clemmow, R. F. Mullaly, and W. P. Allis.

Leader: Propagating channel containing free charges, preceding an electrical breakdown.

Oscillator Strength: Dimensionless quantity that gives the absorption strength of a transition defined as "number of classical oscillators per atom" in the lower state.

Partition Function: Normalization factor for the distribution function, i.e., sum over all energy levels.

Plasma: A mixture of electrons, ions, and neutral particles that is quasineutral, i.e., contains equal numbers of charges of either polarity. Because of the high temperatures necessary for ionization, plasmas emit radiation in a variety of spectral regions.

Radon Equation: Integral that yields the intensity along a line of sight of either an emitting or an absorbing medium. It is the governing equation for computed tomography.

Resonance Line: Line radiation with a transition to the ground state of an atom or ion.

Saint Elmo's Fire: A natural luminous discharge to be seen at the top of pointed objects, e.g., ship masts, at times of high local electrical field strength.

Sputtering: A process by which surface atoms typically from metal electrodes are removed by ion or neutral bombardment. The sputtering process may be used either for a controlled removal of material (e.g., sharpening of razor blades) or for creating deposited layers of sputtered material on a target.

Vacuum Ultraviolet Radiation (VUV): Due to the radiation absorption of molecular oxygen, air is not transparent below 200 nm. Observation of VUV radiation therefore requires evacuated systems.

Works Cited

Akishev, Y. S., Deryugin, A. A., Kochetov, I. V., Napartovich, A. P., Trushkin, N. I. (1993), *J. Phys. D.* **26,** 1630–1637.

Anders, A., Anders, S., Jüttner, B., Bötticher, W., Lück, H., Schröder, G. (1992), *IEEE Trans. Plasma Sci.* **20,** 466–472.

Auwetter-Kurtz, M. (1992), *Lichtbogenantriebe für Weltraumaufgaben,* Teubner: Stuttgart.

Awakowicz, P. (1991), Ph.D. Thesis, Technical University Munich, Germany.

Bauder, U. H., Devoto, R. S., Mukherjee, D. (1973), *Phys. Fluids* **16,** 2143–2148.

Behrisch, R. (1986), in: D. E. Post, R. Behrisch (Eds.), *Physics of Plasma-Wall Interactions in Controlled Fusion,* NATO Advanced Study Institutes Series B, Physics, Vol. 131, New York: Plenum.

Bötticher, W., Lück, H., Niesner, St., Schwabedissen, A. (1992), *Appl. Phys. B* **54,** 295–302.

Buuron, A. J. M. (1993), *Plasma Deposition of Carbon Materials,* Ph.D. Thesis, Technical University, Eindhoven, and (1991) *Fusion Technol.* **19,** 2049.

Chen, F. F. (1991), *Plasma Phys. Controlled Fusion* **33,** 339–364.

Christiansen, J., Schulthesis, Ch. (1979), *Z. Physik A* **290,** 35–41.

Danzmann, K., Günther, M., Fischer, J., Kock, M., Kühne, M. (1988), *Appl. Opt.* **27,** 4947–4951.

Eggert, J. (1919), *Phys. Z.* **20,** 570.

Ehrich, H., Hasse, B., Mausbach, M., Müller, K. G. (1990), *IEEE Trans. Plasma Sci.* **18,** 895–903.

Elenbaas, W. (1936), *Physica* **3,** 947.

Eliasson, B., Kogelschatz, U. (1991), *IEEE Trans. Plasma Sci.* **19,** 309–323.

Eliasson, B., Egli, W., Kogelschatz, U. (1994), *Pure Appl. Chem.* **66,** 1275–1286.

Fang, D. Y., Nürnberg, A., Bauder, U. H., Behrisch, R. (1982), *J. Nucl. Mater.* **111 & 112,** 517–521.

Farmer, A. J., Haddad, G. N. (1984), *Appl. Phys. Lett.* **45,** 24.

Fowler, R. G. (1956), "Radiation from Low Pressure Discharges" in: H. Flügge (Ed.), *Handbuch der Physik,* Vol 22, Berlin: Springer, pp. 209–253.

Gallimberti, I. (1977), *Electra* **76,** 5799.

Geisler, M., Kieser, J., Räuchle, E., Wilhelm, R. (1990), *J. Vac. Sci. Technol.* **A8** (2).

Gerstenberger, D. C., Solanki, R., Collins, G. J. (1980), *IEEE J. Quant. Electron.* **QE-16,** 820–834.

Golde, R. H. (1977), *Lightning,* Vol. 1, London, New York, San Francisco: Academic Press.

Goldman, M., Goldman, N. (1978), "Corona Discharges" in *Gaseous Electronics I, Electrical Discharges,* New York: Academic Press.

Griem, H. (1964), *Plasma Spectroscopy,* New York: Mc Graw-Hill.

Gundersen, M. A., Schaefer, G. (1990), *Physics and Applications of Pseudosparks,* New York, London: Plenum Press.

Harris, L. P. (1980) in: J. M. Lafferty (Ed.), *Vacuum Arcs, Theory and Application,* New York: Wiley.

Hartmann, W., Gundersen, M. A. (1988), *Phys. Rev. Lett.* **60,** 2371.

Jüttner, B. (1994), in: R. L. Boxman, P. Martin, and D. Sanders (Eds.), *Vacuum Arc Science and Technology,* Park Ridge, NJ: Noyes Publications.

Kogelschatz, U. (1992), in: W. T. Williams (Ed.), *Proceedings of the Xth International Conference on Gas Discharges and Their Applications (GD92),* Swansea, U.K.: University College of Swansea, pp. 972–980.

Komori, A., Shoji, T., Miyamoto, K., Kawai, J., Kawai, Y. (1991), *Phys. Fluids B* **3,** 893–898.

Lowke, J. J., Kovitya, P., Schmidt, H.-P. (1992), *J. Phys. D: Appl. Phys.* **25,** 1600–1606.

Maecker, H., Steinberger, S. (1967), *Z. Angew. Phys.* **23,** 456.

Maecker, H., Finkelnburg, W. (1957), *Handbuch der Physik,* Vol. 22, Berlin: Springer.

Mentel, J. (1978), "The Influence of Arc Roots on Current Interruption," in: K. Ragaller (Ed.), *Current Interruption in High-Voltage Networks,* New York: Plenum.

Mentel, J. (1985), in: J. S. Bakos, Z. Soerlei (Eds.), *Proceedings of the XVII International Conference on Phenomena in Ionized Gases,* Budapest: Hungarian Academy of Sciences, pp. 158–171.

Mesyats, G. A., Proskurovsky, D. I. (1989), *Pulsed Electrical Discharges in Vacuum,* Heidelberg: Springer.

Murphy, A. B., Kovitya, P. (1993), *J. Appl. Phys.* **73,** 4759–4769.

Nachtigall, K. P., Mentel, J. (1991), *IEEE Trans. Plasma Sci.* **19,** 942–953.

Napartovich, A. P., Akishev, Y. S. (1993), in: G. Ecker, U. Arendt, J. Boeseler (Eds.), *Proceedings of the XXI International Conference on Phenomena in Ionized Gases,* Vol. 3, Bochum: Arbeitsgemeinschaft Plasmaphysik, Ruhr Universität Bochum, pp. 207–216.

Poisel, H., Landers, F. J., Höß, P., Bauder, U. H. (1986), *IEEE Trans. Plasma Sci.* **PS-14,** 306–318.

Pomathiod, L., Debrie, R., Arnal, Y., Pelletier, J. (1984), *Phys. Lett.* **106A.**

Poroda, E. (1949), Corpus of Ancient Near Eastern Seals, Vol. 1, PI. XXXIV, No. 220. The Collection of the Pierpont Morgan Library, Washington.

Prock, J. (1986), *IEEE Trans. Plasma Sci.* **PS-14,** 482–491.

Radon, J. (1917), *Ber. Verh. Sächs. Akad. Wiss.* **69,** 262–277.

Raizer, Y. P. (1991), *Gas Discharge Physics,* Berlin, Heidelberg: Springer.

Rakhovsky, V. I. (1976), *Proc. 7th Int. Symp. Disch. Electr. Insul. Vacuum,* Novosibirsk, 38–54.

Richter, J. (1968), "Radiation of Hot Gases" in: W. Lochte-Holtgreven (Ed.), *Plasma Diagnostics,* (Ed.), Amsterdam: North-Holland.

Saha, M. N. (1920), *Philos. Mag.* **40,** 472.

Schade, E. (1970), *Z. Phys.* **233,** 53.

Siemens, W. (1857), *Ann. Phys. Chem. (Leipzig)* **102,** 66–122.

Uman, M. A. *et al.* (1990), *J. Geophys. Res.* **95,** 20395–20408.

Von Engel, A., Steenbeck, M. (1934), *Elektrische Gasentladungen,* Vol. II, Berlin: Springer.

Warner, B. E., Persson, K. B. (1979), *J. Appl. Phys.* **50,** 5694–5703.

Wilhelm, R. (1993), "ECR Plasma Sources," in: C. M. Ferreira and M. Moisan (Eds.), *Proc. of a NATO Advanced Research Workshop on Microwave Discharges,* Plenum Press NATO ASI Series 302, pp. 161–179.

Yatsenko, N. A. (1981), *Zh. Tekh. Fiz.* **51,** 1195.

Further Reading

Griem, H. R. (1964), *Plasma Spectroscopy,* New York: McGraw-Hill.

Gundersen, M. A., Schaefer, G. (1990), *Physics and Applications of Pseudosparks,* New York, London: Plenum Press.

Lafferty, J. M. (Ed.) (1980), *Vacuum Arcs, Theory and Application,* New York: Wiley.

Mesyats, G. A., Proskurovsky, D. I. (1989), *Pulsed Electrical Discharges in Vacuum,* Heidelberg: Springer.

Maecker, H., Finkelnburg, W. (1957), in: H. Flügge (Ed.), *Handbuch der Physik,* Vol. 22, Heidelberg: Springer.

Nishikawa, K., Wakatani, M. (1994), *Plasma Physics,* Heidelberg: Springer.

Patankar, S. V. (1980), *Numerical Heat Transfer and Fluid Flow,* New York: McGraw-Hill.

Post, D. E., Behrisch, R. (Eds.) (1986), *Physics of Plasma-Wall Interactions in Controlled Fusion,* New York: Plenum Press.

Raizer, Y. P. (1991), *Gas Discharge Physics,* Berlin, Heidelberg: Springer.

SPATIAL QUANTITIES, MEASUREMENT OF

See Supplement Volume 1

SPECIAL FUNCTIONS

CHARLIE HARPER, *Department of Physics, California State University, Hayward, California, U.S.A.*

Introduction 189
1. **The Sturm–Liouville Theory** 190
1.1 Introduction 190
1.2 Hermitian Operators and Their Eigenvalues 193
1.3 Orthogonality Condition and Completeness of Eigenfunctions 193
1.4 Orthogonal Polynomials and Functions 194
1.4.1 Recurrence Formula for Orthogonal Polynomials 194
1.4.2 Rodrigues Formulas 195
1.4.3 The Generating Function 195
1.5 Factorization of the Sturm–Liouville Equation 195
1.6 Connection with Local Lie-Group Theory 199
2. **The Hypergeometric Function, ${}_2F_1(a,b,c;z)$** 201
2.1 Properties of the Hypergeometric Differential Equation 201
2.2 Properties of $F(a,b,c;z)$ 202
2.3 Helmholtz's Differential Equation in Spherical Coordinates 204
2.4 Associated Legendre Functions and Legendre Polynomials 207
2.5 The Radial Equation 208
3. **The Confluent Hypergeometric Function, ${}_1F_1(a,b,c;x)$** 209
3.1 More on Hermite Polynomials 211
3.2 More on the Laguerre and Associated Laguerre Polynomials 213
4. **Helmholtz's Differential Equation in Cylindrical Coordinates** 216
4.1 Solutions of Bessel's Differential Equation 216
4.2 Bessel Functions of the First Kind 216
4.3 Neumann Functions 216
4.4 Hankel Functions 217
4.5 Modified Bessel Functions 217
4.6 Spherical Bessel Functions 218
5. **Other Special Functions Used in Mathematical Physics** 219
5.1 Some Other Special Functions—Type 1 219
5.2 Some Other Special Functions—Type 2 221
Glossary 222
Works Cited 223
Further Reading 223

INTRODUCTION

This article is devoted to the theory and applications of a set higher transcendental functions that arise naturally in mathematical physics. These higher transcendental functions are referred to as special functions, and they arise

1. when solving, in certain curvilinear coordinate systems, partial differential equations that are defined by physical problems and/or
2. when finding eigenfunctions and eigenvalues of differential operators.

The partial-differential-equations approach to special functions involves use of the separation-of-variables method and either the Frobenius–Fuchs power series solutions of one or more resulting ordinary differential equations or an Infeld–Hull type factorization procedure for finding eigenfunctions and eigenvalues of second-order ordinary linear differential equations.

The focus is on a class of physical prob-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

lems whose differential-equation formulation involves the Laplacian operator, ∇^2. The resulting partial differential equations include the Laplace equation, the heat-conduction (diffusion) equation, the mechanical wave-motion equation, and the Schrödinger wave equation. After application of the separation-of-variables method, the resulting time-independent parts of all these partial differential equations may be written in the form of the Helmholtz differential equation, $\nabla^2 u + k^2 u = 0$. Problems involving the Helmholtz differential equation in spherical coordinates lead to spherical harmonics, Legendre polynomials and associated Legendre functions, Laguerre and associated Laguerre polynomials, and spherical Bessel functions. Problems modeled by use of Helmholtz's differential equation in cylindrical coordinates involve the various types of Bessel functions. Solutions of the Schrödinger wave equation for a linear harmonic oscillator are expressed in terms of Hermite polynomials.

Special functions such as Hermite polynomials, Legendre polynomials and associated Legendre functions, spherical harmonics, Laguerre and associated Laguerre polynomials, and Bessel functions are widely used in mathematical physics and are the main focus of this article; these special functions are special cases of the hypergeometric function ${}_2F_1(a,b,c;z)$ or confluent hypergeometric function ${}_1F_1(a,c;z)$. There exist other useful special functions in mathematical physics that are not expressible in terms of ${}_2F_1$ or ${}_1F_1$. The functions ${}_2F_1$ and ${}_1F_1$ may be developed from the following main viewpoints:

1. ordinary differential equations and the Frobenius–Fuchs power series method,
2. factorization of ordinary differential equations, and
3. representation theory of local Lie groups.

In a lecture course, *The Application of Group Theory to the Special Functions of Mathematical Physics* (unpublished lecture notes, Princeton University, Princeton, NJ, 1955; see Talman, 1968), Wigner pointed out that certain classes of special functions arise as matrix elements of the representations of local Lie groups. Since Wigner's work, many other group-theoretical approaches to special functions have been developed. The purposes of these various group-theoretical approaches are to show unity (or demonstrate a central foundation) among the extremely large number of special functions and to derive their various known basic properties. It is important, however, to note that there exist no single approach to special functions that unites all special functions and illuminates all of their various properties.

Section 1 is devoted to a treatment of the Sturm–Liouville theory and orthogonal polynomials, since these concepts provide insight into properties of solutions of the second-order ordinary linear differential equations in which we are interested. The factorization method of Infeld and Hull and its connection to a group-theoretical foundation of certain special functions are summarized in Secs. 1.5 and 1.6.

A discussion of the properties of the hypergeometric differential equation and the hypergeometric function ${}_2F_1$ is given in Sec. 2. The Legendre polynomials and associated Legendre functions are special examples of ${}_2F_1$, and their important properties are developed in Sec. 2. Also, some discussion of the Chebyshev polynomial, Gegenbauer polynomial, and Jacobi polynomial is given in Sec. 2.

Section 3 is devoted to special functions that are special cases of the confluent hypergeometric function, ${}_1F_1$. These special functions include the various Bessel functions, Laguerre and associated Laguerre polynomials, and the Hermite polynomials. Also, discussions of some properties of the confluent hypergeometric differential equation and properties of confluent hypergeometric function are given in Sec. 3.

In Sec. 4, a comprehensive treatment of the Helmholtz differential equation in cylindrical coordinates and the resulting Bessel functions is given. A summary of certain other special functions used in mathematical physics is given in Sec. 5.

1. THE STURM–LIOUVILLE THEORY

1.1 Introduction

Linear operators are basic to linear differential equations, and the solutions of each of the differential equations we will consider form a vector space. Hence, we begin with the essentials of the theory of linear operators.

A vector space (also known as a linear space or linear manifold) V is a nonempty set of elements $\{\psi_i\}$, called vectors (here vector is used in an abstract mathematical sense), for which the operations of addition and multiplication by a scalar are valid. Addition is a rule that assigns an element for $\psi_1 + \psi_2$ in V for every pair of elements ψ_1 and ψ_2 in V. The operation of multiplication by a scalar is a rule that assigns an element for $a\psi_1$ in V to each complex a and each ψ_1 in V. In addition, a zero element and negative elements exist, and the associative and commutative properties of addition are valid.

A function T that transforms (maps) vectors in V into a vector space W, $T : V \rightarrow W$, is called a linear transformation if $T(\psi_1 + \psi_2) = T(\psi_1) + T(\psi_2)$ and $T(a\psi_1) = aT(\psi_1)$ are valid for all vectors and all real scalars in V. The case $T : V \rightarrow V$ is called a linear operator on V. A real number λ is an eigenvalue (characteristic value) of the linear operator T if there is a nonzero vector ψ (called the eigenvector of T) in V for which $T(\psi) = \lambda\psi$.

In mathematical physics, the linear operator is normally a differential operator, the eigenvalue equation is a differential equation, and the eigenfunctions (solutions) form a vector space and satisfy certain imposed boundary conditions.

For physical problems, the three basic types of boundary conditions are

1. Dirichlet, specification of the solution at each point on the boundary;
2. Neumann, specification of the normal derivative of the solution at each point on the boundary; and
3. Cauchy, specification of both Dirichlet and Neumann conditions.

Typically, the actual physical problem is used as a guide for the formulation of boundary conditions. Sometimes, however, it is difficult to formulate the appropriate boundary conditions for a problem. Hence, it is important to understand what conditions are appropriate for a particular type of differential equation (see, e.g., ANALYTIC METHODS).

The Laplace equation, the time-independent heat-conduction (diffusion) equation, and the time-independent mechanical wave equation may be put into the form of the Helmholtz differential equation. The Helmholtz differential equation, $\nabla^2 u + k^2 u = 0$ for constant k^2, is an eigenvalue equation where $-k^2$ is the eigenvalue, u is the eigenfunction and is subject to boundary conditions, and ∇^2 is the operator. The time-independent Schrödinger wave equation contains the operator ∇^2 and is an eigenvalue equation of the form $\hat{H}\psi = E\psi$ where $\hat{H} = (-\hbar^2/2m)\nabla^2 + V(x,y,z)$.

The separation-of-variables method applied to Helmholtz's differential equation and to the time-independent Schrödinger's wave equation in various coordinate systems leads to ordinary differential equations that may be written in the following general form:

$$\frac{d}{dx}\left\{p(x)\frac{du}{dx}\right\} - q(x)u + \lambda\rho(x)u = 0. \tag{1}$$

The parameter λ is a separation constant (in some cases, more than one separation constant may appear; we will focus on the case of one separation constant). Equation (1) is the well-known Sturm–Liouville equation, and it may be written in the following operator form:

$$\mathcal{L}(u) + \lambda\rho(x)u = 0. \tag{2}$$

The Liouville operator, a linear operator, is defined by

$$\mathcal{L}(u) = \frac{d}{dx}\left\{p(x)\frac{du}{dx}\right\} - q(x)u. \tag{3}$$

For the general differential operator $M(u) = p(x)u'' + r(x)u' + q(x)u$, the operator $\bar{M}(u) = (pu)'' - (ru)' + qu$ is defined as the adjoint of $M(u)$. Note that $M(u) = \bar{M}(u)$ when $p' = r$, and $M(u)$ is said to be a self-adjoint operator in this case. On applying the general definition for the adjoint of an operator to Eq. (3), we find that $\mathcal{L}(u) = \bar{\mathcal{L}}(u)$, which means that the Liouville operator is a self-adjoint operator. In fact, it can be shown that every second-order differential operator can be transformed to the self-adjoint form (see Courant and Hilbert, 1953, p. 279).

In the Sturm–Liouville equation, the function $\rho(x)$ [$w(x)$ and $r(x)$ are also used] is called the density or weight. This name for

$\rho(x)$ is related to the historical origin of the Sturm–Liouville equation, which involved finding the solution for the one-dimensional mechanical wave equation, $[p(x)u_x]_x = \rho(x)u_{tt}$, representing the motion of a nonhomogeneous string. In the mechanical wave equation, $u(x,t)$ is the displacement of the string from its equilibrium position, $p(x)$ is proportional to the modulus of elasticity, and $\rho(x)$ is the mass per unit length of the string. Separation of variables leads to the ordinary differential equations $(pX')' + \lambda\rho X = 0$ and $\ddot{T} + \lambda T = 0$, where λ is the separation constant and $u(x,t)$ is assumed to equal the product $X(x)T(t)$, with typical boundary conditions given by $X(a) = X(b)$ and $p(a)X'(a) = p(b)X'(b)$.

In Eq. (1), the functions $p(x)$, $q(x)$, and $\rho(x)$ are assumed to be real, continuous with continuous derivatives, and nonzero in the region of interest, $[a,b]$. Moreover, it is assumed that $p(x)$ and $\rho(x)$ are always positive in $[a,b]$. The Sturm–Liouville equation is a generalized form of the usual eigenvalue equation since the eigenvalue is multiplied by the density function $\rho(x)$, which may be different from unity. The sign convention for $q(x)$ in Eqs. (1) and (3) conforms to the usage of Courant and Hilbert; some authors use a plus sign for $q(x)$ in these equations. The function $u(x)$ is subject to appropriate boundary conditions. With appropriate substitutions, the following differential equations are among the list of important differential equations in mathematical physics that may be put in the Sturm–Liouville form: Legendre and associated Legendre, Laguerre and associated Laguerre, Schrödinger equation for the linear harmonic oscillator, and Bessel. Hence, a study of the general properties of the Sturm–Liouville equation is extremely useful in mathematical physics. A summary of the relations between the differential equations for many important special functions and the Sturm–Liouville equation is given in Table 1.

The problem of determining the dependence of the eigenfunction $u(x)$ on the eigenvalue λ and of the eigenvalue λ on the boundary conditions imposed on $u(x)$ is often referred to as the Sturm–Liouville problem. The Sturm–Liouville problem is important for problems in both classical and quantum theory. Sturm–Liouville theory unites properties of the solutions of second-order ordinary linear differential equations related to

1. Hermitian and self-adjoint operators;
2. reality of eigenvalues of Hermitian and self-adjoint operators;
3. orthogonality and completeness of eigenfunctions;
4. degeneracy of eigenvalues (if N linearly independent eigenfunctions correspond to the same eigenvalue, then the eigenvalue is said to be N-fold degenerate); and
5. the fact that eigenvalues of the Sturm–Liouville equation form a discrete set of values such that $\ldots \lambda_1 \le \lambda_2 \le \lambda_3 \ldots$

These properties are important in the study

Table 1. Relation to the Sturm–Liouville equation.

$$\frac{d}{dx}\left\{p(x)\frac{du}{dx}\right\} - q(x)u + \lambda\rho(x)u = 0$$

Equation	$p(x)$	$q(x)$	$\rho(x)$	λ
Legendre, $P_n(x)$	$1 - x^2$	0	1	$n(n + 1)$
Associated Legendre, $P_n^m(x)$	$1 - x^2$	$\dfrac{m^2}{1 - x^2}$	1	$n(n + 1)$
Laguerre, $L_n(x)$	xe^{-x}	0	e^{-x}	n
Associated Laguerre, $L_n^k(x)$	$x^{k+1}e^{-x}$	0	$x^k e^{-x}$	$n - k$
Bessel, $J_n(x)$, $Y_n(x)$, $H_n^{(1)}, \ldots$	x	n^2/x	x	1
Hermite, $H_n(x)$	e^{-x^2}	0	e^{-x^2}	$2n$
Quantum oscillator, $\psi_n(x)$	1	x^2	1	λ
Jacobi, $P_n^{(\alpha,\beta)}(x)$; $\alpha, \beta > -1$	$\dfrac{1 - x^2}{(1 - x)^{-\alpha}(1 + x)^{-\beta}}$	0	$\dfrac{(1 - x)^{\alpha}}{(1 + x)^{-\beta}}$	$n(n + \alpha + \beta + 1)$
Chebyshev, $T_n(x)$	$(1 - x^2)^{1/2}$	0	$(1 - x^2)^{-1/2}$	n^2
Gegenbauer, $C_n^{(\alpha)}(x)$; $\alpha > -\frac{1}{2}$	$(1 - x^2)^{\alpha+1/2}$	0	$(1 - x^2)^{\alpha-1/2}$	$n(n + 2\alpha)$

of problems that lead to each of the differential equations we analyze in this article.

1.2 Hermitian Operators and Their Eigenvalues

Consider two twice-differentiable functions u_i and u_j. By use of Eqs. (1) and (2), we obtain

$$u_i^*\mathcal{L}(u_j) - [\mathcal{L}(u_i)]^*u_j = \frac{d}{dx}\left\{p\left(u_i^*\frac{du_j}{dx} - u_j\frac{du_i^*}{dx}\right)\right\}. \tag{4}$$

In Eq. (4), the asterisk is used to denote complex conjugate of the respective functions. Integrating both sides of Eq. (4) over the range of interest yields

$$\int_a^b \{u_i^*\mathcal{L}(u_j) - [\mathcal{L}(u_i)]^*u_j\}dx = \left\{p\left(u_i^*\frac{du_j}{dx} - u_j\frac{du_i^*}{dx}\right)\right\}_{x=b} - \left\{p\left(u_i^*\frac{du_j}{dx} - u_j\frac{du_i^*}{dx}\right)\right\}_{x=a}.$$

Note that the above equation results from the fact that $\mathcal{L}$ is self-adjoint. The operator $\mathcal{L}$ is said to be Hermitian if the following endpoint boundary conditions are imposed on the two functions and their derivatives.

$$\left\{p\left(u_i^*\frac{du_j}{dx} - u_j\frac{du_i^*}{dx}\right)\right\}_{x=b} = \left\{p\left(u_i^*\frac{du_j}{dx} - u_j\frac{du_i^*}{dx}\right)\right\}_{x=a}. \tag{5}$$

By use of the boundary conditions in Eq. (5), the Hermitian relation may be written as

$$\int_a^b u_i^*\mathcal{L}(u_j)dx = \int_a^b [\mathcal{L}(u_i)]^*u_j dx. \tag{6}$$

Thus far, the Liouville operator has been assumed to be real. In quantum mechanics, operators are generally complex (for example, the x component of the linear momentum operator is given by $p_x = -i\hbar\partial/\partial x$), and it is assumed that wave functions satisfy the boundary conditions in Eq. (5). The Hermitian condition in quantum mechanics for linear operator $\hat{A}$ takes the form

$$\int_{\text{all space}} \psi_i^*\hat{A}\psi_j d\tau = \int_{\text{all space}} (\hat{A}\psi_i)^\dagger\psi_j d\tau. \tag{7}$$

An arbitrary linear operator may be put in matrix form, and the notation $\hat{A}^\dagger$ means

1. interchange rows with columns and
2. take the complex conjugate of each element

(this process is called the Hermitian conjugate); in this connection, note that $(\hat{A}\psi_i)^\dagger = \psi_i^*\hat{A}^\dagger$. When an operator satisfies the condition $\hat{A} = \hat{A}^\dagger$, the operator is said to be Hermitian. In the bra and ket vector notation, Eq. (7) becomes $\langle\psi_i|\hat{A}\psi_j\rangle = \langle\hat{A}\psi_i|\psi_j\rangle$.

For solution $u_i = u_j = u$ in Eq. (6) and use of Eq. (2), we obtain

$$\int_a^b [u^*\mathcal{L}(u) - u\mathcal{L}(u^*)]dx = (\lambda - \lambda^*)\int_a^b \rho(x)u^*u dx = 0. \tag{8}$$

The result in Eq. (8) means that eigenvalues of Hermitian operators are real, $\lambda = \lambda^*$.

1.3 Orthogonality Condition and Completeness of Eigenfunctions

By use of Eq. (2) for distinct eigenfunctions u_i and u_j with distinct eigenvalues, the Hermitian relation in Eq. (6) may be written as

$$\int_a^b \{u_i^*\mathcal{L}(u_j) - [\mathcal{L}(u_i)]^*u_j\}dx = \int_a^b [u_i^*(-\lambda_j\rho u_j) + \lambda_i^*\rho u_i^*u_j]dx = (\lambda_i^* - \lambda_j)\int_a^b u_i^*u_j\rho(x)dx = 0.$$

Since $\lambda_i \neq \lambda_j$ and λ_i is real, the above equation implies that

$$\int_a^b u_i^*(x)u_j(x)\rho(x)dx = 0. \tag{9}$$

Equation (9) shows that eigenfunctions corresponding to distinct eigenvalues are orthogonal in the interval $[a,b]$ with respect to the weight function $\rho(x)$.

An orthonormal set of Sturm–Liouville eigenfunctions, $\{u_k(x)\}$, forms a complete set of functions (Courant and Hilbert, 1953, Chap. 6, Sec. 3). This completeness property means that the following equation is valid for any function $f(x)$ that is at least piecewise continuous in the interval $[a,b]$:

$$\lim_{n\to\infty}\int_a^b\left|f(x)-\sum_{k=1}^{n}c_k u_k(x)\right|^2\rho(x)dx=0. \tag{10}$$

The notation $|\ldots|^2$ means the product of the enclosed quantity and its complex conjugate, $|z|^2 = z^*z$. Equation (10) and the orthogonality relation in Eq. (9) lead to

$$f(x)=\sum_{k=1}^{\infty}c_k u_k(x)\quad\text{where } c_k = \int_a^b f(x)u_k^*(x)\rho(x)dx \tag{11}$$

and

$$\sum_{k=1}^{\infty}|c_k|^2=\int_a^b|f(x)|^2\rho(x)dx. \tag{12}$$

Equation (12) is referred to as the completeness relation. By use of Eq. (11) with the appropriate orthogonality relation and weight, one may obtain series expansions for $f(x)$ in terms of any complete set of orthogonal polynomials (or orthogonal functions); for example, the Fourier series, Legendre series, and Hermite series may be written respectively as

$$f(x)=\sum_{n=-\infty}^{\infty}c_n e^{in\pi x/l}$$

where

$$c_n=\frac{1}{2l}\int_{-l}^{l}f(x)e^{-in\pi x/l}dx,\quad -l\le x\le l;$$

$$f(x)=\sum_{n=0}^{\infty}c_n P_n(x)$$

where

$$c_n=\frac{2n+1}{2}\int_{-1}^{1}f(x)P_n(x)dx,\quad -1\le x\le 1;$$

and

$$f(x)=\sum_{n=0}^{\infty}c_n H_n(x)$$

where

$$c_n=\frac{1}{2^n n!\sqrt{\pi}}\int_{-\infty}^{\infty}f(x)H_n(x)e^{-x^2}dx,\quad -\infty\le x\le\infty.$$

1.4 Orthogonal Polynomials and Functions

The power-series solutions of many second-order ordinary linear differential equations in mathematical physics such as the Legendre, Laguerre, and Hermite differential equations may be written as an orthogonal polynomial plus an orthogonal function. The set of real polynomials $\{f_n(x)\}$ is said to be orthogonal with respect to the weight function $\rho(x)$ over the interval $[a,b]$ if the following relations are valid:

$$\int_a^b \rho(x)f_n(x)f_m(x)dx=\begin{cases}0 & m\neq n\\ h_n^2 & m=n\end{cases}. \tag{13}$$

If $h_n^2 = 1$ for all n, then the system of orthogonal polynomials in Eq. (13) is said to be orthonormal. Note that the weight determines the system of polynomials up to a constant factor (the value of h_n) in each polynomial; the specification of this constant factor for each polynomial is referred to a standardization (or standard convention). Summarized in Table 2 are some frequently used orthogonality relations.

1.4.1 Recurrence Formula for Orthogonal Polynomials In many cases, three consecutive orthogonal polynomials f_{n+1}, f_n,

Table 2. Some orthogonality relations.

$$\int_{-1}^{1}P_n(x)P_k(x)dx=\frac{2}{2n+1}\delta_{nk}$$

$$\int_{-1}^{1}P_n^m(x)P_k^m(x)dx=\frac{2}{2n+1}\frac{(n+m)!}{(n-m)!}\delta_{nk}$$

$$\int_0^{\infty}e^{-x}L_n(x)L_m(x)dx=\delta_{nm}$$

$$\int_0^{\infty}e^{-x}x^k L_n^k(x)L_m^k(x)dx=\frac{(n!)^3}{(n-k)!}\delta_{nm}$$

$$\int_{-\infty}^{\infty}e^{-x^2}H_n(x)H_m(x)dx=2^n n!\sqrt{\pi}\delta_{nm}$$

Table 3. Recurrence relations for some polynomials.

$A_n f_{n+1}(x) = (B_n + C_n x) f_n(x) - D_n f_{n-1}(x).$

$f_n(x)$	A_n	B_n	C_n	D_n
$P_n(x)$	$n + 1$	0	$2n + 1$	n
$P_n^m(x)$	$n - m + 1$	0	$2n + 1$	$n + m$
$L_n(x)$	$n + 1$	$2n + 1$	-1	n
$L_n^k(x)$	$n + 1$	$2n + k + 1$	-1	$n + k$
$H_n(x)$	1	0	2	$2n$

and f_{n-1} satisfy a recurrence formula of the form

$$A_n f_{n+1}(x) = (B_n + C_n x) f_n(x) - D_n f_{n-1}(x). \tag{14}$$

The recurrence relations for some of the frequently used orthogonal polynomials are listed in Table 3.

1.4.2 Rodrigues Formulas Formulas involving the nth derivative of an elementary function that can be used to generate orthogonal polynomials f_n are called Rodrigues formulas. It can be shown that the Rodrigues formulas for many orthogonal polynomials may be combined into the following general Rodrigues formula:

$$f_n = \frac{1}{a_n \rho(x)} \frac{d^n\{\rho(x)[g(x)]^n\}}{dx^n}. \tag{15}$$

The quantity $g(x)$ is a polynomial whose coefficients are independent of n, and the factor a_n is determined by the standardization of the orthogonal polynomial system. The Leibnitz formula for the nth derivative of a product should be used to evaluate the right-hand side of Eq. (15); this formula is

$$\frac{d^n}{dx^n}\{A(x)B(x)\} = \sum_{s=0}^{n} \left\{ \frac{n!}{(n-s)!s!} \frac{d^{n-s}}{dx^{n-s}} A(x) \frac{d^s}{dx^s} B(x) \right\}. \tag{16}$$

The Rodrigues formulas for some important orthogonal polynomials are given in Table 4.

1.4.3 The Generating Function The function $\mathscr{G}(x,t)$ is said to be a generating function of the sequence of functions $\{f_n(x)\}$ if the $f_n(x)$ are, up to a constant, the coefficients of t^n in the expansion of $\mathscr{G}(x,t)$ in powers of t. Such an expansion is valid for orthogonal polynomials and most orthogonal functions, and it may be written in the form

$$\mathscr{G}(x,t) = \sum_{n=0}^{\infty} a_n f_n(x) t^n. \tag{17}$$

In Eq. (17), the a_n are independent of x and t. The generating functions for certain orthogonal polynomials that will be used in other parts of this article are given in Table 5.

1.5 Factorization of the Sturm–Liouville Equation

The factorization method for finding eigenfunctions and corresponding eigenvalues of a large class of Schrödinger-type equations (forms of the Sturm–Liouville equation) was introduced by Schrödinginer (1940); in a clearly written paper, Infeld and Hull (1951) further developed the method. By use of the factorization method, a second-order differential equation is factored (transformed) into a product of first-order

Table 4. Rodrigues formula for some polynomials.

$$f_n = \frac{1}{a_n \rho(x)} \frac{d^n\{\rho(x)[g(x)]^n\}}{dx^n}.$$

$f_n(x)$	a_n	$g(x)$	$\rho(x)$
$P_n(x)$	$2^n n!$	$x^2 - 1$	1
$L_n^k(x)$	$n!$	x	$x^k e^{-x}$
$H_n(x)$	$(-1)^n$	1	e^{-x^2}
$P_n^{(\alpha,\beta)}(x)$	$(-1)^n 2^n n!$	$1 - x^2$	$(1 - x)^\alpha (1 + x)^\beta$
$C_n^{(\alpha)}(x)$	$(-1)^n 2^n n! \dfrac{\Gamma(2\alpha)\Gamma(\alpha + n + \frac{1}{2})}{\Gamma(\alpha + \frac{1}{2})\Gamma(n + 2\alpha)}$	$1 - x^2$	$(1 - x^2)^{\alpha - 1/2}$
$T_n(x)$	$(-1)^n 2^{n+1} \dfrac{\Gamma(n + \frac{1}{2})}{\sqrt{\pi}}$	$1 - x^2$	$(1 - x^2)^{-1/2}$

Table 5. Some generating functions.

$$\mathcal{G}(x,t) = \sum_{n=0}^{\infty} a_n f_n(x) t^n; R = \sqrt{1 - 2xt + t^2}.$$

$f_n(x)$	a_n	$\mathcal{G}(x,t)$
$P_n(x)$	1	R^{-1}
$L_n(x)$	1	$(1 - t)^{-1} \exp\{-xt/(1 - t)\}$
$L_n^k(x)$	1	$(1 - t)^{-(k+1)} \exp[xt/(t - 1)]$
$H_n(x)$	$\frac{1}{n!}$	$\exp(2xt - t^2)$
$P_n^{(\alpha,\beta)}(x)$	$2^{-\alpha-\beta}$	$R^{-1}(1 - t + R)^{-\alpha}(1 + t + R)^{-\beta}$
$C_n^{(\alpha)}(x)$	1	$R^{-2\alpha}$
$T_n(x)$	2	$1 + (1 - t^2)/R^2$

differential operators, which results in a pair of first-order differential equations that are equivalent to the original second-order differential equation. The form of the potential function determines if the factorization method will be successful. In terms of the forms of the potential function, Infeld and Hull classified factorizations into six general factorization types. Many examples of the various factorization types are presented in the paper by Infeld and Hull. We give a brief overview of the factorization method because it is

1. a useful and straightforward method for finding eigenfunctions (in many cases, involving special functions) and corresponding eigenvalues for important differential equations in mathematical physics, and
2. equivalent to a local Lie-group theoretical analysis involving differential operators that leads to a group-theoretical approach to special functions.

In this overview of the factorization method, we closely follow the Infeld–Hull approach; using a plus sign for the second term in the Sturm–Liouville equation, we find that it may be transformed (see, e.g., Courant and Hilbert, 1953, p. 292) into the following general differential equation in standard form:

$$\frac{d^2y}{dx^2} + r(x,m)y + \lambda y = 0; m = 0, 1, 2, \ldots. \tag{18}$$

In Eq. (18), the function $r(x,m)$ contains the dynamical information, such as the potential energy, that characterizes the particular problem. The basic idea of the factorization method is that Eq. (18) can be either factored directly as a single differential equation or factored into a pair of differential equations of the following forms:

$$H_{m+1}^{+}H_{m+1}^{-}y(\lambda,m) = [\lambda - L(m + 1)]y(\lambda,m) \tag{19}$$

and

$$H_m^{-}H_m^{+}y(\lambda,m) = [\lambda - L(m)]y(\lambda,m),$$
$$\text{where } H_m^{\pm} = k(x,m) \pm \frac{d}{dx}. \tag{20}$$

The condition that must be satisfied by $k(x,m)$ and $L(m)$ is obtained by carrying out the indicated operations in Eqs. (19) and (20), comparing the result with the original differential equation, and eliminating $r(x,m)$; this condition is

$$k^2(x, m + 1) - k^2(x,m) + \frac{dk(x, m + 1)}{dx}$$
$$+ \frac{dk(x,m)}{dx} = L(m) - L(m + 1). \tag{21}$$

The fundamental idea of the factorization method is established in the five theorems that we now summarize.

Theorem I: If $y(\lambda,m)$ is a solution of the original differential equation, then $y(\lambda, m + 1) = H_{m+1}^{-}y(\lambda,m)$ and $y(\lambda, m - 1) = H_m^{+}y(\lambda,m)$ are also solutions corresponding to the same λ.

Theorem II (mutual adjointness of the operators): For functions g and f satisfying the end-point conditions,

$$\int_a^b g(H_m^{-}f)dx = \int_a^b (H_m^{+}g)f dx.$$

Theorem III (satisfying the boundary conditions): If $y(\lambda,m)$ is quadratically integrable over the entire range of x and $L(m)$ is an increasing function of m $(0 < m)$, then the raising operation, $y(\lambda, m + 1) = H_{m+1}^{-} \times y(\lambda,m)$, produces a function that is also quadratically integrable and vanishes at the end points. If $L(m)$ is a decreasing function of $m (0 < m)$, then the lowering operation, $y(\lambda, m - 1) = H_m^{+}y(\lambda,m)$, produces a function that is also quadratically integrable and vanishes at the end points.

Theorem IV: Class I solutions—$L(m)$ is an increasing function of the integer m. If $0 < m \leq M$ and $\lambda \leq$ the larger of $L(M)$ and $L(M+1)$, then a necessary condition for a square integrable solution is that $\lambda = \lambda_l = L(l+1)$, where l is an integer, and $m = 0, 1, \ldots, l$ and $H_{l+1}^{-}y(\lambda,l) \equiv 0$. Other solutions are given by

$$y_l^{m-1} = [L(l+1) - L(m)]^{-1/2}\left[k(x,m) + \frac{d}{dx}\right]y_l^m.$$

Class II solutions—$L(m)$ is a decreasing function of the integer m. If $0 \leq m \leq M$ and $\lambda \leq L(0)$, then a necessary condition for square integrable solutions is that $\lambda = \lambda_l = L(l)$, where l is an integer, and $m = l, l+1, \ldots$ and $H_l^{+}y(\lambda,l) \equiv 0$. Other solutions are given by

$$y_l^{m+1} = [L(l) - L(m+1)]^{-1/2} \times \left[k(x, m+1) - \frac{d}{dx}\right]y_l^m.$$

Theorem V (normalization of solutions): The $H^{\pm}$ operators preserve the normalization of the eigenfunctions.

The basic task now is to find $k(x,m)$ and $L(m)$ corresponding to a given $r(x,m)$. Infeld and Hull have shown that finding $k(x,m)$ and $L(m)$ is accomplished by use of one of the following six operator types.

Type A solutions are related to spherical harmonic functions and other eigenfunctions that are special cases of the hypergeometric function, ${}_2F_1$; we have

$$r(x,m) = -\frac{a^2(m+c)(m+c+1) + d^2 + 2ad(m+c+\frac{1}{2})\cos a(x+p)}{\sin^2 a(x+p)},$$

$$k(x,m) = a(m+c)\cot a(x+p) + \frac{d}{\sin a(x+p)},$$

and

$$L(m) = a^2(m+c)^2$$

where a, c, d, and p are constants.

Type B solutions are related to associated Laguerre and Laguerre functions and other eigenfunctions that are special cases of the confluent hypergeometric function, ${}_1F_1$; here

$$r(x,m) = -d^2e^{2ax} + 2ad(m+c+\tfrac{1}{2})e^{ax},$$

$$k(x,m) = de^{ax} - m - c,$$

and

$$L(m) = -a^2(m+c)^2.$$

Type C solutions are related to confluent hypergeometric functions, with

$$r(x,m) = -\frac{(m+c)(m+c+1)}{x^2} - \frac{b^2x^2}{4} + b(m-c),$$

$$k(x,m) = \frac{m+c}{x} + \frac{bx}{2},$$

and

$$L(m) = -2bm + \frac{b}{2}.$$

Type D solutions are related to generalization of the Hermite polynomials; here

$$r(x,m) = -(bx+d)^2 + b(2m+1),$$

$$k(x,m) = bx + d,$$

and

$$L(m) = -2bm.$$

Type E solutions are related to hypergeometric functions, with

$$r(x,m) = -\frac{m(m+1)a^2}{\sin^2 a(x+p)} - 2aq\cot a(x+p),$$

$$k(x,m) = ma\cot a(x + p) + \frac{q}{m},$$

and

$$L(m) = a^2m^2 - \frac{q^2}{m^2}.$$

Type F solutions are related to Laguerre polynomials, with

$$r(x,m) = -\frac{2q}{x} - \frac{m(m + 1)}{x^2},$$

$$k(x,m) = \frac{m}{x} + \frac{q}{m},$$

and

$$L(m) = -\frac{q^2}{m^2}.$$

Example 1: Find the eigenvalues and eigenfunctions of the associated Legendre differential equation by use of the factorization method.

Solution: By use of the transformation $u = P \sin^{1/2}\theta$, the associated Legendre differential equation [the theta equation in spherical coordinates, Eq. (49)],

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{dP}{d\theta}\right) + \left[\lambda - \frac{m^2}{\sin^2\theta}\right]P = 0, \tag{22}$$

reduces to

$$\frac{d^2u}{d\theta^2} - \frac{m^2 - \frac{1}{4}}{\sin^2\theta}u + (\lambda + \tfrac{1}{4})u = 0. \tag{23}$$

On comparing Eq. (23) with the general equation, Eq. (18), in standard form, we find that it is a Type A factorization (see Theorem IV) where the parameters are given by $a = 1$, $c = -\frac{1}{2}$, $d = 0$, $p = 0$, $x = \theta$, and $\lambda \Rightarrow \lambda + \frac{1}{4}$; the required quantities for Type A factorization reduce to $r(\theta,m) = -(m^2 - \frac{1}{4})/\sin^2\theta$, $L(m) = (m - \frac{1}{2})$, and $k(\theta,m) = (m - \frac{1}{2})\cot\theta$. Since this is a Class I problem, $\lambda + \frac{1}{4} = L(l + 1)$ or $\lambda = l(l + 1)$ for $l = 0, 1, 2, \ldots$ $l \geq m$. The corresponding eigenfunctions (solutions) of Eq. (23) are obtained by use of Theorem IV for Class I solutions, $H_l^+y(\lambda,l) \equiv 0$; for normalization constants C_l and D_{lm}, the solutions are given by

$$u_l^l = C_l\sin^{(l+1/2)}\theta \text{ and}$$

$$u_l^{m-1} = D_{lm}\left\{(m - \tfrac{1}{2})\cot\theta + \frac{d}{d\theta}\right\}u_l^m.$$

Example 2: By use of a modified factorization method, find the eigenvalues and eigenfunctions of the quantum mechanical linear harmonic oscillator.

Solution: The linear harmonic oscillator problem was used by Schrödinger in his original work on factorization; it is a Type D problem, but $r(x,m)$ does not depend on m. For the linear harmonic oscillator, it is simpler to use the modified version of factorization involving the idea of raising and lowering operators; this modified factorization method is used in many quantum mechanics textbooks and is summarized here. The time-independent Schrödinger equation for the linear harmonic oscillator with potential energy $V(x) = kx^2/2 = m\omega^2x^2/2$ is given by

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + \frac{m\omega^2x^2}{2}\psi = E\psi. \tag{24}$$

The change of variable $\xi = (m\omega/k)^{1/2}x$ reduces Eq. (24) to the dimensionless form

$$\frac{d^2\psi}{d\xi^2} - \xi^2\psi + \lambda\psi = 0, \quad \text{where } \lambda = \frac{2E}{\hbar\omega}. \tag{25}$$

By use of the factorization (of differential operators) method, Eq. (25) may be written as

$$H^+H^-\psi = (\lambda + 1)\psi \text{ and } H^-H^+\psi = (\lambda - 1)\psi,$$
$$\text{where } H^{\pm} = \xi \pm \frac{d}{d\xi}. \tag{26}$$

The above factored equations are equivalent to Eq. (25). The basic idea of this factorization method is that ψ_n and λ_n can be generated from known ψ_0 and λ_0. If ψ_0 satisfies the original equation (factored equations) Eq. (26) with eigenvalue λ_0, then we may write

$$\left(\frac{d}{d\xi} - \xi\right)\left(\frac{d}{d\xi} + \xi\right)\psi_0 = (-\lambda_0 + 1)\psi_0. \tag{27}$$

By use of the identity

$$\left(\frac{d}{d\xi} - \xi\right)\left(\frac{d}{d\xi} + \xi\right)\psi = 2\psi - \left(\frac{d}{d\xi} + \xi\right)\left(\xi - \frac{d}{d\xi}\right)\psi,$$

Eq. (27) reduces to

$$\left(\frac{d}{d\xi} - \xi\right)\left(\frac{d}{d\xi} + \xi\right)\psi_1 = \lambda_1\psi_1. \tag{28}$$

To obtain Eq. (28), we made the following substitutions:

$$\psi_1(\xi) = \left(\xi - \frac{d}{d\xi}\right)\psi_0 \text{ and } \lambda_1 = \lambda_0 + 1. \tag{29}$$

In a similar manner, we obtain

$$\psi_{-1}(\xi) = \left(\xi + \frac{d}{d\xi}\right)\psi_0 \text{ and } \lambda_{-1} = \lambda_0 - 1. \tag{30}$$

The operators in Eqs. (29) and (30) are referred to as raising and lowering operators, respectively. If the above process is repeated n times, we obtain

$$\psi_n(\xi) = \left(\xi - \frac{d}{d\xi}\right)^n \psi_0 \text{ for } \lambda_n = \lambda_0 + n \tag{31}$$

and

$$\psi_{-n}(\xi) = \left(\xi + \frac{d}{d\xi}\right)^n \psi_0 \text{ for } \lambda_{-n} = \lambda_0 - n. \tag{32}$$

As seen from Eqs. (31) and (32), the eigenvalue (eigenenergy) is bounded from below at λ_0, but it is an arbitrarily large negative value, as seen in Eq. (32). It is required that the negative values of λ_{-n} be terminated (Theorem IV) by setting $\psi_{-1} = 0 = (\xi - d/d\xi)\psi_0$; by use of this termination, we find that $\psi_0 = N\exp(-\xi^2/2)$ is a solution of the original differential equation, Eq. (25). The corresponding lower bound for the eigenvalue is $\lambda_0 = 1$ or $E_0 = \hbar\omega/2$. Hence, we find that the knowledge of ψ_0 and λ_0 allows us to find ψ_n and λ_n; we obtain

$$\psi_n(\xi) = N_n H_n(\xi) e^{-\xi^2/2} \text{ for } \lambda_n = 1 + n \text{ or } E_n = \hbar\omega(n + \tfrac{1}{2}). \tag{33}$$

Equation (33) results from the fact that $(\xi - d/d\xi)^n \exp(-\xi^2/2)$ produces $\exp(-\xi^2/2)$ times an nth-order polynomial, $H_n(\xi)$; this polynomial, as is shown in Sec. 2.2, is the Hermite polynomial.

1.6 Connection with Local Lie-Group Theory

Wigner, unpublished Princeton University Lecture Notes in 1955, demonstrated that large classes of special functions arise as matrix elements of the representations of Lie groups such as the groups of rotations in two, three, and four dimensions or the Euclidean groups in two and three dimensions. Talman (1968) extended the work of Wigner. The Wigner group-theoretical approach to special functions shows that

1. addition theorems for special functions result from the group multiplication rule and
2. differential equations giving rise to special functions result from limits of generators.

Other group-theoretical development of special functions exist (see, e.g. Vilenkin, 1968).

The focus in this section is on the approach by Miller (1968), which is based on the recognition of the equivalence between the Infeld–Hull factorization method and the representation theory of complex local Lie groups with the four-dimensional Lie algebras $\mathcal{G}(a,b)$ and the six-dimensional Lie algebra $\mathcal{F}_6$. Local Lie-group theory provides a unifying approach to a large class of special functions and their properties. A comprehensive treatment of group theory is given in GROUP THEORY and in the many excellent books on the subject; we simply summarize essential group concepts needed to outline the Miller group-theoretical approach to special functions.

Local Lie-group theory was developed in the nineteenth century; it is based on the use of local coordinates (needed for a discussion of special functions), and concerned with groups that are analytic only in the neighborhood of the group identity element. A global Lie group involves coordinate-free considerations and is an abstract group.

The Lie algebra (vector space) $\mathcal{G}$ of the local Lie group G is the set of all tangent vectors (generators) at the identity element together with commutation relations $[\alpha,\beta] \in \mathcal{G}$ defined for all tangent vectors α, $\beta \in \mathcal{G}$. Lie's three fundamental theorems and their converses together with the Taylor expansion theorem provide a mechanism for constructing the Lie algebra associated with a Lie group. For local transformation groups, which are important in our analysis, the commutation relations involve Lie derivatives, $[L_\alpha, L_\beta]$, or generalized Lie derivatives, $[D_\alpha, D_\beta]$.

A four-dimensional complex Lie algebra $\mathcal{G}(a,b)$ with basis $\mathcal{J}^\pm$, $\mathcal{J}^3$, and $\mathcal{E}$ is defined by the commutation relations

$$[\mathcal{J}^+,\mathcal{J}^-] = 2a\mathcal{J}^3 - b\mathcal{E};\ [\mathcal{J}^3,\mathcal{J}^\pm] = \pm\mathcal{J}^\pm;\ \text{and}\ [\mathcal{J}^\pm,\mathcal{E}] = [\mathcal{J}^3,\mathcal{E}] = 0. \tag{34}$$

The parameters a and b are a pair of arbitrary complex numbers. For the representation ρ of $\mathcal{G}(a,b)$ on the complex vector space V, we define four operators as follows:

$$J^+ = \rho(\mathcal{J}^+);\ J^- = \rho(\mathcal{J}^-);\ J^3 = \rho(\mathcal{J}^3);\ \text{and}\ E = \rho(\mathcal{E}). \tag{35}$$

The four operators in Eq. (35) obey commutation relations similar to those in Eq. (34); they are

$$[J^+,J^-] = 2aJ^3 - bE;\ [J^3,J^\pm] = \pm J^\pm;\ \text{and}\ [J^\pm,E] = [J^3,E] = 0. \tag{36}$$

The object at this stage is that of finding realizations of the irreducible representations $\rho(\alpha)$ of $\mathcal{G}(a,b)$ [where $\alpha \in \mathcal{G}(a,b)$] that form a Lie algebra of analytic differential operators acting on V. An extremely large number of possible solutions of Eq. (36) exists. By use of generalized Lie derivatives and a classification of all such solutions by differential operators in two variables, Miller (1968, Chap. 8) has shown that only generalized forms of the angular momentum operators are important in special function theory. We therefore write

$$J^3 = \frac{\partial}{\partial y};\ J^\pm = e^{\pm y}\left[\pm\frac{\partial}{\partial x} - k(x)\frac{\partial}{\partial y} + j(x)\right];\ \text{and}\ E = \mu. \tag{37}$$

In Eq. (37), μ is a complex constant; the functions $k(x)$ and $j(x)$ are to be determined. The equations in Eq. (37) satisfy all the commutation conditions for the operators in Eq. (36), except $[J^+,J^-] = 2aJ^3 - bE$, which is satisfied when $k(x)$ and $j(x)$ are solutions of the differential equations

$$\frac{dk(x)}{dx} + k(x)^2 = -a^2\ \text{and}\quad \frac{dj(x)}{dx} + k(x)j(x) = -\frac{d\mu}{2}. \tag{38}$$

The connection of the group-theoretical analysis with the Infeld–Hull factorization method is related to finding the various solutions of the equations in Eq. (38); these various solutions will now be summarized.

$\mathcal{G}(1,0)$ yields

Type A: $k(x) = \cot(x + p)$ and $j(x) = \dfrac{q}{\sin(x + p)}$;

and

Type B: $k(x) = i$ and $j(x) = qe^{-ix}$.

$\mathcal{G}(0,1)$ yields

Type C′: $k(x) = \dfrac{1}{x + p}$ and $j(x) = -\dfrac{\mu(x + p)}{4} + \dfrac{q}{x + p}$;

and

Type D′: $k(x) \equiv 0$ and $j(x) = -\dfrac{\mu x}{2} + q$.

$\mathcal{G}(0,0)$ yields

Type C″: $k(x) = \dfrac{1}{x + p}$ and $j(x) = \dfrac{q}{x + p}$;

and

Type D″: $k(x) \equiv 0$ and $j(x) = q$.

The quantities p and q are complex constants, and $i = \sqrt{-1}$. In the Infeld–Hull notation, Types C′ and C″ are combined and called Type C, and Types D′ and D″ are combined to form Type D. The six-dimensional complex Lie algebra $\mathcal{F}_6$, with generators $\mathcal{P}^\pm$, $\mathcal{P}^3$, $\mathcal{J}^\pm$, and $\mathcal{J}^3$, is defined by appropriate commutation relations for these generators. Because of the difficulty involved in computing matrix elements in the representation theory of $\mathcal{F}_6$, the task of obtaining all of the special-function identities implied by $\mathcal{F}_6$ is incomplete. Some results are obtained by noting that the elements $\mathcal{J}^{\pm}$ and $\mathcal{J}^3$ generate a subalgebra of $\mathcal{F}_6$ that is isomorphic to $sl(2)$, and the elements $\mathcal{P}^{\pm}$ and $\mathcal{P}^3$ generate a three-dimensional abelian subalgebra of $\mathcal{F}_6$. Type A and B operators forming a realization of $sl(2)$ can be extended to Type E and F operators, respectively, forming a realization of $\mathcal{F}_6$ (Miller, 1968, Chap. 6). The connection of these eight types of solutions of Eq. (32) with special functions is now summarized.

$\mathcal{G}(1,0)$: Type A solutions are related to hypergeometric functions, and Type B solutions are related to confluent hypergeometric functions and generalized associated Laguerre functions.

$\mathcal{G}(0,1)$: Type C′ solutions are closely related to confluent hypergeometric functions, Laguerre and associated Laguerre functions. Type D′ solutions are related to parabolic cylinder functions and Hermite polynomials.

$\mathcal{G}(0,0)$: Type C″ solutions are related to Bessel functions, and Type D″ solutions are related to simple transcendental functions, $e^{\pm i\omega x}$.

$\mathcal{F}_6$: Type E solutions are related to hypergeometric functions. Types A and E solutions have different recurrence relations and the same eigenfunctions. Type F solutions are related to confluent hypergeometric functions. Types B and F solutions have different recurrence relations and the same eigenfunctions.

2. THE HYPERGEOMETRIC FUNCTION, $_2F_1(a,b,c;z)$

2.1 Properties of the Hypergeometric Differential Equation

Important members of a large subset of special functions are related to a class of functions called hypergeometric functions, which are solutions of the hypergeometric differential equation (also known as Gauss's differential equation). The hypergeometric differential equation has three regular singular points, and it can be shown that any second-order ordinary linear differential equation with three regular singular points can be transformed (reduced) to the hypergeometric differential equation form. The solutions of many physical problems involve special functions that result from solving second-order ordinary linear differential equations with regular singular points. It is, therefore, natural to expect a connection among hypergeometric functions and certain special functions. The hypergeometric differential equation has the form

$$z(1-z)\frac{d^2w}{dz^2} + [c-(a+b+1)z]\frac{dw}{dz} - abw = 0. \tag{39}$$

Note that Eq. (39) has regular singular points at $z = 0, 1, \infty$. In Eq. (39), parameters a, b, and c are arbitrary complex constants. The hypergeometric differential equation can be solved by use of the Frobenius–Fuchs power series method (see, e.g., ANALYTIC METHODS),

$$w = \sum_{\lambda=0}^{\infty} a_\lambda z^{\lambda+k}, \text{ where } a_0 \neq 0.$$

At the three regular singular points, the respective solutions of the indicial equations are, at $z = 0$: $k = 0$ and $1 - c$; at $z = 1$: $k = 0$ and $c - a - b$; and at $z = \infty$: $k = a$ and b. In general, the solutions of Eq. (39) are the various forms of the Gauss hypergeometric series; they are $[w(z) = {}_2F_1(a,b,c;z) = F(a,b,c;z)]$

$$_2F_1(a,b,c;z) = \sum_{n=0}^{\infty} \frac{(a)_n(b)_n}{n!(c)_n} z_n. \tag{40}$$

The factorial function in Eq. (40) is defined by

$$\begin{aligned}(\lambda)_n &= \prod_{k=1}^{n} (\lambda + k - 1)\\ &= \lambda(\lambda+1)(\lambda+2)\ldots(\lambda+n-1)\\ &= \frac{\Gamma(\lambda+n)}{\Gamma(\lambda)}\end{aligned} \tag{41}$$

for $\lambda = a$, b, and c, respectively. This notation for the factorial function is called a Pochhammer symbol. The notation $_2F_1$ means that there are two factorial functions in the numerator of the series and one factorial function in the denominator of the series in Eq. (40). The gamma function $\Gamma(\lambda)$ is defined in Sec. 5.2.

For convenience, we will write $F(a,b,c;z)$

for ${}_2F_1(a,b,c;z)$ since the three parameters in the parentheses are sufficient to avoid confusion with the notation for the confluent hypergeometric function. Originally, the notation ${}_2F_1(a,b;c;z)$ was used to represent the hypergeometric function. One, however, finds a variety of combinations of commas and/or semicolons used in the literature to represent the hypergeometric function. We will use a comma to separate constants and a semicolon to separate the variable from the constants.

The hypergeometric series $F(-m,b,-n;z)$ is not defined if $n < m$. The hypergeometric series converges if $|z| < 1$ and diverges if $|z| > 1$. For $z = 1$, the series converges if $c > a + b$, and it converges for $z = -1$ if $c > a + b - 1$. The hypergeometric series becomes a polynomial of degree n in z when a or b equals a negative integer.

Note that $F(a,b,c;0) = 1$ and $F(a,b,c;z) = F(b,a,c;z)$. Also, note that many elementary transcendental functions may be expressed in terms of a hypergeometric series; two examples are $\ln(1 + z) = zF(1,1,2;-z)$ and $(1 + z)^a = F(-a,b,b;-z)$. The geometric series is a special case of the hypergeometric series since

$$F(1,1,1;z) = \sum_{n=0}^{\infty} z^n.$$

The numerous properties of $F(a,b,c;z)$ summarized in this article as well as in many other places were developed by Euler and Gauss.

2.2 Properties of $F(a,b,c;z)$

If none of the numbers c, $c - a - b$, or $a - b$ equals an integer, then two linearly independent solutions of the hypergeometric differential equation exist, and each of the six solutions (two at each of three singular points) may be written in four equivalent forms. These 24 forms are referred to as Kummer's 24 solutions of the Gauss hypergeometric differential equation. One of these forms in each case is as follows:

$$w_1(0) = F(a,b,c;z),$$

$$w_2(0) = z^{1-c}F(a - c + 1, b - c + 1, 2 - c; z),$$

$$w_1(1) = F(a, b, a + b + 1 - c; 1 - z),$$

$$w_2(1) = (1 - z)^{c-a-b} \times F(c - b, c - a, c - a - b + 1; 1 - z),$$

$$w_1(\infty) = z^{-a}F(a, a - c + 1, a - b + 1; z^{-1}),$$

and

$$w_2(\infty) = z^{-b}F(b, b - c + 1, b - a + 1; z^{-1}).$$

The functions $F(a \pm 1, b, c; z)$, $F(a, b \pm 1, c; z)$, and $F(a, b, c \pm 1; z)$ are said to be contiguous to $F(a,b,c;z)$. Numerous relations between $F(a,b,c;z)$ and any two contiguous functions, of a form similar to

$$(c - a)F(a - 1, b, c; z) + (2a - c - az + bz) \times F(a,b,c;z) + a(z - 1)F(a + 1, b, c; z) = 0,$$

were developed by Gauss.

When $|z| < 1$, $F(a,b,c;z)$ is analytic, and $F(a,b,c;z)$ has a branch point at $z = 1$ (see Whitaker and Watson, 1947, Sec. 14.53); if a cut is made from 1 to infinity along the real axis, $F(a,b,c;z)$ is analytic throughout the cut plane. Formulas for $F(a,b,c;x)$ for $z = x + iy$ may be obtained from those involving $F(a,b,c;z)$ by use of the following replacements in the cut interval $(-1,1)$: $z - 1$ by $(1 - x)e^{\pm i\pi}$, $z^2 - 1$ by $(1 - x^2)e^{\pm i\pi}$, and $z + 1$ by $x + 1$.

The following differentiation and integration relations are valid for ${}_2F_1$.

A differentiation relation for $F(a,b,c;z)$:

$$\frac{d^nF(a,b,c;z)}{dz^n} = \frac{(a)_n(b)_n}{(c)_n} \times F(a + n, b + n, c + n; z).$$

Fundamental integral relation: The form of the fundamental integral relation for the hypergeometric function is

$$F(a,b,c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c - b)} \times \int_0^1 t^{b-1}(1 - t)^{c-b-1}(1 - tz)^{-a}dt$$

$$\text{for } \mathrm{Re}(c) > \mathrm{Re}(b) > 0.$$

Solutions of the hypergeometric differential equation that are orthogonal polynomials are of particular interest in this article, and polynomial solutions occur when a or b is a negative integer. Examples of the connections of $F(a,b,c;x)$ with some special polynomials of interest are as follows:

$$F(-n,n,\tfrac{1}{2};x) = T_n(1-2x)\text{—Chebyshev;}$$

$$F(-n,n+1,1;x) = P_n(1-2x)\text{—Legendre;}$$

$$F(-n,n+2\alpha,\alpha+\tfrac{1}{2};x) = \frac{n!}{(2\alpha)_n}C_n^{(\alpha)}(1-2x)\text{—Gegenbauer,}$$

and

$$F(-n,\alpha+1+\beta+n,\alpha+1;x) = \frac{n!}{(\alpha+1)_n}P_n^{(\alpha,\beta)}(1-2x)\text{—Jacobi.}$$

The Gegenbauer (also known as ultraspherical), Legendre and associated Legendre, and Chebyshev polynomials are special cases of the Jacobi polynomial (sometimes called hypergeometric polynomial). Chebyshev (Tschebyscheff, Tchebichef, and Tchebicheff are other spellings found in the literature) polynomials involve solutions of separated equations in spherical, parabolic, and prolate and oblate spheroidal coordinates. Chebyshev polynomials converge rapidly and have the special property that $\mathrm{Max}T_n(x) = +1$ and $\mathrm{Min}T_n(x) = -1$; because of this property, Chebyshev polynomials are useful in numerical analysis. Gegenbauer functions result from separated equations in circular cylinder and spherical coordinates with two regular singular points at ± 1 rather than at 0 and 1.

We now summarize some of the basic properties of the solutions of the Jacobi differential equation (see Table 1 in Sec. 1.1.) and express these solutions in terms of ${}_2F_1$. The solutions of the Jacobi differential equation may be written as

$$y = c_1P_n^{(\alpha,\beta)}(x) + c_2Q_n^{(\alpha,\beta)}(x).$$

The quantity $P_n^{(\alpha,\beta)}(x)$ is a polynomial and is called Jacobi polynomial of the first kind. The quantity $Q_n^{(\alpha,\beta)}(x)$ is not a polynomial and is called Jacobi function of the second kind. In terms of hypergeometric functions, we write

$$P_n^{(\alpha,\beta)}(x) = \frac{(\alpha+1)_n}{n!} \times F\left(-n, n+\alpha+\beta+1, \alpha+1; \frac{1-x}{2}\right)$$

and

$$Q_n^{(\alpha,\beta)}(x) = \frac{C(\alpha,\beta)}{(x-1)^{n+\alpha+1}(x+1)^{\beta}} \times F\left(n+1, n+\alpha+1, 2n+\alpha+\beta+2; \frac{2}{1-x}\right).$$

The symbol $C(\alpha,\beta)$ in the above equation represents the quantity

$$C(\alpha,\beta) = \frac{2^{n+\alpha+\beta}\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{\Gamma(2n+\alpha+\beta+2)(x-1)^{n+\alpha+1}}.$$

The standardization for the Jacobi polynomial is given by

$$P_n^{(\alpha,\beta)}(1) = (\alpha+1)_n/n!.$$

Rodrigues's formula and the generating function for the Jacobi polynomial are given in Tables 4 and 5, respectively. The form of the recursion formula for the Jacobi polynomial is

$$\begin{aligned}&2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)P_{n+1}^{(\alpha,\beta)}(x)\\&= (2n+\alpha+\beta+1)\\&\quad\times[(2n+\alpha+\beta)(2n+\alpha+\beta)x+\alpha^2-\beta^2]\\&\quad\times P_n^{(\alpha,\beta)}(x) - 2(n+\alpha)(n+\beta)\\&\quad\times(2n+\alpha+\beta+2)P_{n-1}^{(\alpha,\beta)}(x).\end{aligned}$$

The integral representation of the Jacobi polynomial may be written as

$$P_n^{(\alpha,\beta)}(x) = \frac{1}{2\pi i}\oint_c \frac{1}{2}\left(\frac{t^2-1}{t-x}\right)^n \times\left(\frac{1-t}{1-x}\right)^{\alpha}\left(\frac{1+t}{1+x}\right)^{\beta}dt;\ x \neq \pm 1. \tag{42}$$

The contour in Eq. (42) is a simple closed contour in a positive sense around $t = x$; the points $t = \pm 1$ are outside of the contour. In Eq. (42), the quantities raised to the α and β power are defined to be unity when $t = x$. Graphical illustrations of Jacobi, Chebyshev, and Gegenbauer polynomials are given in Figs. 1–3.

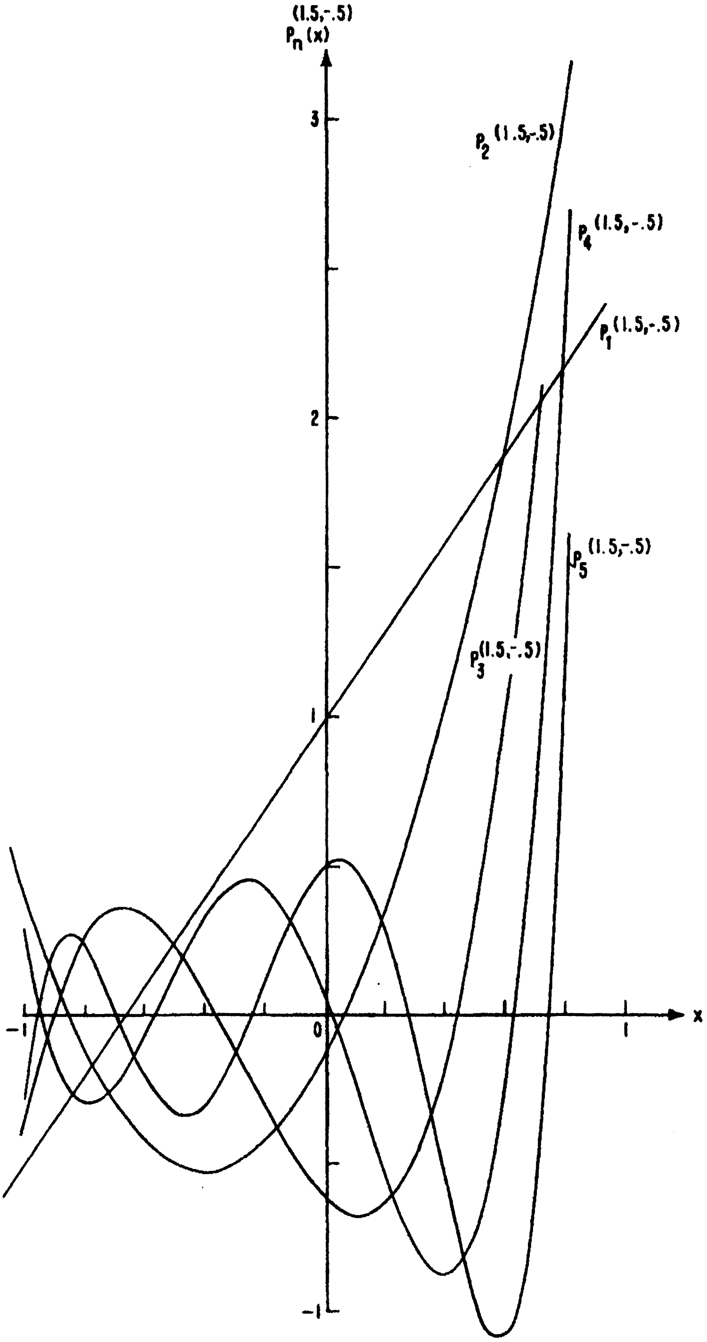

FIG. 1. Jacobi polynomials, $P_n^{(1.5,-0.5)}(x)$ (Abramowitz and Stegun, 1964).

2.3 Helmholtz's Differential Equation in Spherical Coordinates

The Laplacian operator in spherical coordinates (r,θ,ϕ) has the form

$$\nabla^2 = \nabla_r^2 + \frac{1}{r^2}\nabla_{\theta,\phi}^2.$$

The radial and angular parts of the Laplacian are given respectively by

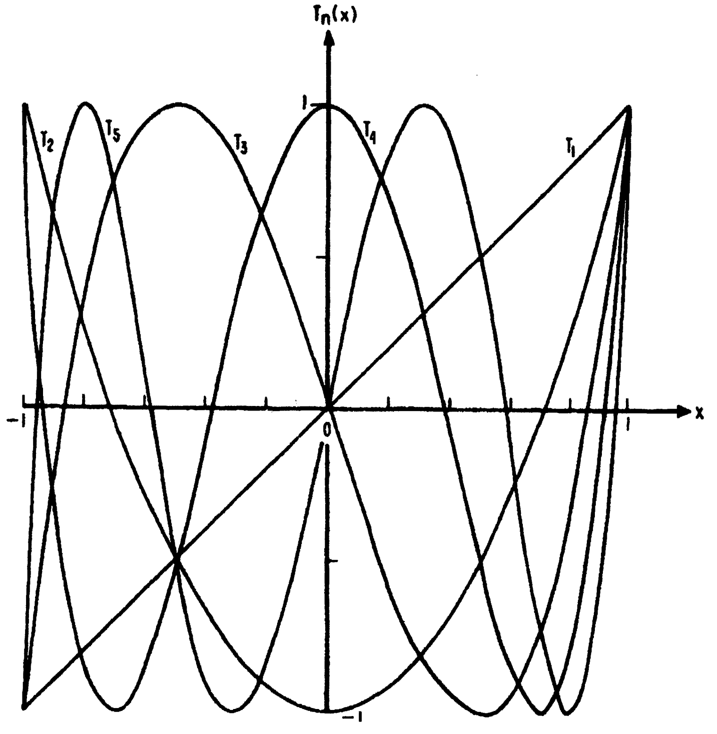

FIG. 2. Chebyshev polynomials, $T_n(x)$ (Abramowitz and Stegun, 1964).

$$\nabla_r^2 = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) \tag{43}$$

and

$$\nabla_{\theta,\phi}^2 = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}. \tag{44}$$

The Helmholtz differential equation in spherical coordinates has the form

$$\left\{\nabla_r^2 + \frac{1}{r^2}\nabla_{\theta,\phi}^2\right\}u + k^2u = 0. \tag{45}$$

The solution of Eq. (45) involves symmetry-type information in terms of the angular parts θ and ϕ, which are common to all problems with spherical symmetry, and dynamical information in terms of the radial part, which characterizes the particular problem under investigation. In this connection, the Helmholtz differential equation reduces to the Laplace differential equation for $k^2 = 0$, to the time-independent heat-conduction (diffusion) or time-independent mechanical wave differential equations for $k^2 =$ constant, and to the time-independent Schrödinger wave equation for $k^2 = (2m/\hbar^2)\{E - V(r)\}$. On separating the angular parts from the radial part in Eq. (45) for $u(r,\theta,\phi) = R(r)Y(\theta,\phi)$, the corresponding differential equations for $R(r)$ and $Y(\theta,\phi)$ with separation constant λ are

$$\nabla_r^2 R(r) + \left\{k^2 - \frac{\lambda}{r^2}\right\}R(r) = 0 \tag{46}$$

and

$$\nabla_{\phi,\phi}^2 Y(\theta,\phi) + \lambda Y(\theta,\phi) = 0. \tag{47}$$

Note that the general solutions of Eq. (47) are independent of the specific problem under investigation but are common to all problems that involve the Laplacian operator in spherical coordinates. The solutions of Eq. (47) with separation constant $-m^2$ are called spherical harmonics (also known as surface harmonics of the first kind), $Y_\lambda^m(\theta,\phi)$. For square integrable solutions, a replacement of the form $\lambda = n(n + 1)$ is required. Tesseral harmonics is the name given to $Y_n^m(\theta,\phi)$ when $m < n$, and the term sectoral harmon-

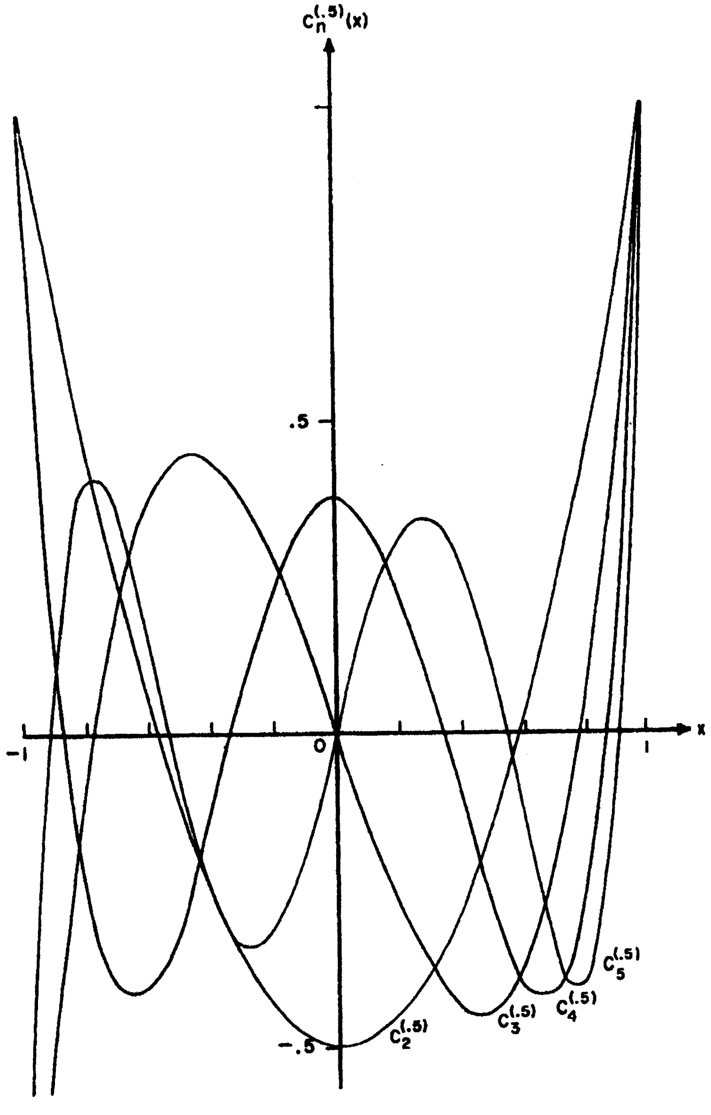

FIG. 3. Gegenbauer polynomials, $C_n^{(0.5)}(x)$ (Abramowitz and Stegun, 1964).

ics is used when $m = n$. Tesseral and sectoral harmonics may be written as $C_n e^{im\phi} P_n^m(\cos\theta)$, where $P_n^m(\cos\theta)$ are associated Legendre functions of the first kind. When $m = 0$, the spherical functions are called Legendre polynomials of the first kind (also known as zonal harmonics and Legendre coefficients).

On substituting Eqs. (43) and (44) into Eqs. (46) and (47), respectively, and separating the variables in Eq. (47) $[Y(\theta,\phi) = \Theta(\theta)\Phi(\phi)$ with separation constant $-m^2]$, we obtain the following three ordinary differential equations: radial equation,

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + \left[k^2 - \frac{n(n+1)}{r^2}\right]R = 0; \qquad (48)$$

theta equation,

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) + \left[n(n+1) - \frac{m^2}{\sin^2\theta}\right]\Theta = 0; \qquad (49)$$

and azimuthal equation,

$$\frac{d^2\Phi}{d\phi^2} + m^2\Phi = 0. \qquad (50)$$

The general solution of the azimuthal equation is $\Phi = c_1 e^{im\phi} + c_2 e^{-im\phi}$. In solving physical problems, the requirement that $\Phi(\phi)$ be a single-valued function is imposed. That is to say, we require that $\Phi(\phi) = \Phi(\phi + 2\pi)$, which leads to the following acceptable values for m: $m = 0, \pm 1, \pm 2, \ldots$; the solution may be written in the form $\Phi(\phi) = Ae^{im\phi}$. In quantum mechanics, the single-valued function requirement is referred to as Born's periodic boundary condition, and m is the magnetic quantum number.

2.4 Associated Legendre Functions and Legendre Polynomials

Solutions of the theta equation, Eq. (49), involve the Legendre polynomials and associated Legendre functions. The traditional treatment of the theta equation involves introducing a new independent variable x by use of the transformation $x = \cos\theta$; here x is not the usual Cartesian coordinate. The associated Legendre differential equation results from this transformation,

$$(1 - x^2)\frac{d^2\Theta}{dx^2} - 2x\frac{d\Theta}{dx} + \left\{n(n + 1) - \frac{m^2}{1 - x^2}\right\}\Theta = 0. \tag{51}$$

The Legendre differential equation is the result when m equals zero in Eq. (51); it is given by

$$(1 - x^2)\frac{d^2\Theta}{dx^2} - 2x\frac{d\Theta}{dx} + n(n + 1)\Theta = 0. \tag{52}$$

The Legendre and associated Legendre differential equations can be solved by use of the Frobenius–Fuchs power series method, factorization method, and hypergeometric functions. The two linearly independent solutions of the associated Legendre differential equation are

$$\Theta_n^m(x) = AP_n^m(x) + BQ_n^m(x).$$

The quantities $P_n^m(x)$ and $Q_n^m(x)$ are associated Legendre functions of the first and second kind, respectively; they are related to the hypergeometric functions as follows:

$$P_n^m(z) = \frac{1}{\Gamma(1 - m)}\left(\frac{z + 1}{z - 1}\right)^{m/2} \times F\left(-n, n + 1, 1 - m; \frac{1 - z}{2}\right), \quad |1 - z| < 2,$$

and

$$Q_n^m(z) = C^{nm} \times F\left(1 + \frac{n}{2} + \frac{m}{2}, \frac{1}{2} + \frac{n}{2} + \frac{m}{2}, n + \frac{3}{2}; \frac{1}{z^2}\right); \quad |z| > 1.$$

The factor C^{nm} in the above equation is given by

$$C^{nm} = e^{im\pi}2^{-n-1}\pi^{1/2} \times \frac{\Gamma(n + m + 1)}{\Gamma(n + \frac{3}{2})} z^{-n-m-1}(z^2 - 1)^{m/2}.$$

Also, the associated Legendre functions $P_n^m(x)$ result when the Legendre polynomials are differentiated m times. That is to say,

$$P_n^m(x) = (1 - x^2)^{m/2}\frac{d^m P_n(x)}{dx^m}.$$

The orthogonality relation, recurrence relation, and Rodrigues formula for the associated Legendre functions are given in Tables 2, 3, and 4, respectively. Graphical illustrations of associated Legendre functions are given in Fig. 4.

The Legendre polynomials of the first kind, $P_n(x)$, and Legendre functions of the second kind, $Q_n(x)$, are related to the Jacobi functions as follows:

$$P_n^{(0,0)}(x) = P_n(x) \text{ and } Q_n^{(0,0)}(x) = Q_n(x).$$

In series form, the Legendre polynomials may be written as

$$\Theta_n(x) \equiv P_n(x) = \sum_{r=0}^{N}\frac{(-1)^r(2n - 2r)!x^{n-2r}}{2^n r!(n - r)!(n - 2r)!}.$$

In the above sum, $N = n/2$ for n even and $N = (n - 1)/2$ for n odd. The general solution of the Legendre differential equation has the form $\Theta_n(x) = C_1P_n(x) + C_2Q_n(x)$; Legendre functions of the second kind, $Q_n(x)$, are not polynomials. The orthogonality relation, recurrence relation, Rodrigues formula, and

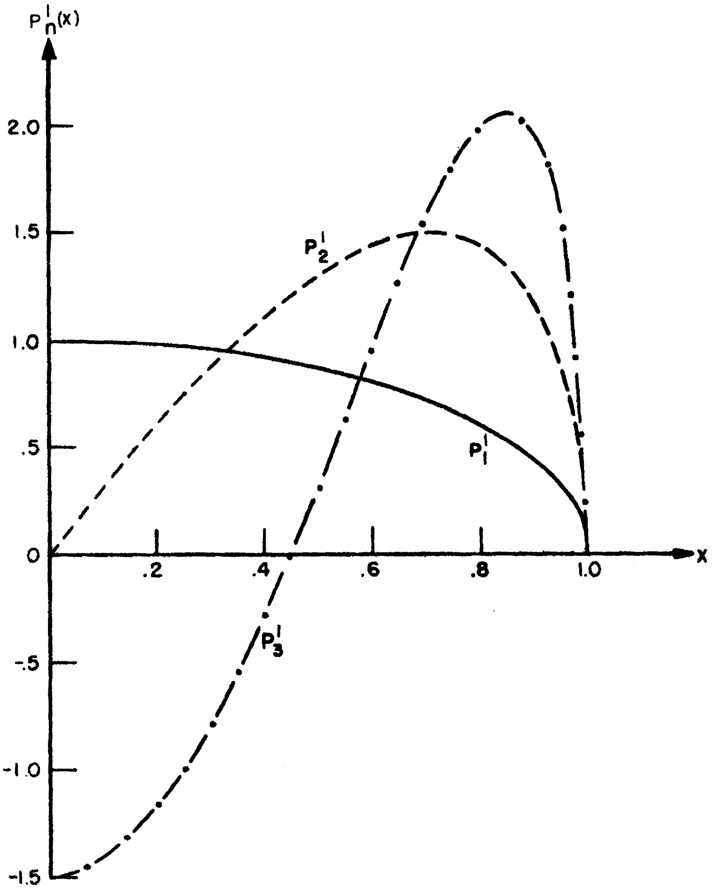

FIG. 4. Associated Legendre functions, $P_n^1(x)$ (Abramowitz and Stegun, 1964).

generating function for the Legendre polynomials are given in Tables 2, 3, 4, and 5, respectively.

The Legendre functions of the second kind, $Q_n(x)$, satisfy a recursion relation of the same form as the one for $P_n(x)$. Graphical illustrations of Legendre polynomials are given in Figs. 5–7, and Legendre functions of the second kind are illustrated in Figs. 8 and 9.

2.5 The Radial Equation

The radial equation, Eq. (48), characterizes the dynamical information of specific problems or classes of problems. The Laplace equation results when $k^2 = 0$; for this case, the solution characterizes such steady-state problems as potentials in electrostatics and temperatures in heat conduction. The time-independent Schrodinger equation results when $k^2 = 2\mu[E - V(r)]/\hbar^2$ for a class of two-body central-force problems; the reduced mass of such a system is given by $\mu = m_1m_2/(m_1 + m_2)$, where m_1 and m_2 are the masses of constituent particles one and two, respectively.

Example 3: Calculate the steady-state temperature distribution $T(r,\theta)$ within a sphere of radius b when the temperature over the surface of the sphere is independent of ϕ. That is to say, $T(b,\theta) = f(\theta)$, where $f(\theta)$ is a known function.

Solution: The general solution of Laplace's differential equation of this problem is independent of ϕ (circular symmetry) and has the form $T(r,\theta) = R(r)\Theta(\theta)$. The radial and theta equations for this problem reduce to

$$r^2R'' + 2rR' - n(n + 1)R = 0$$

and

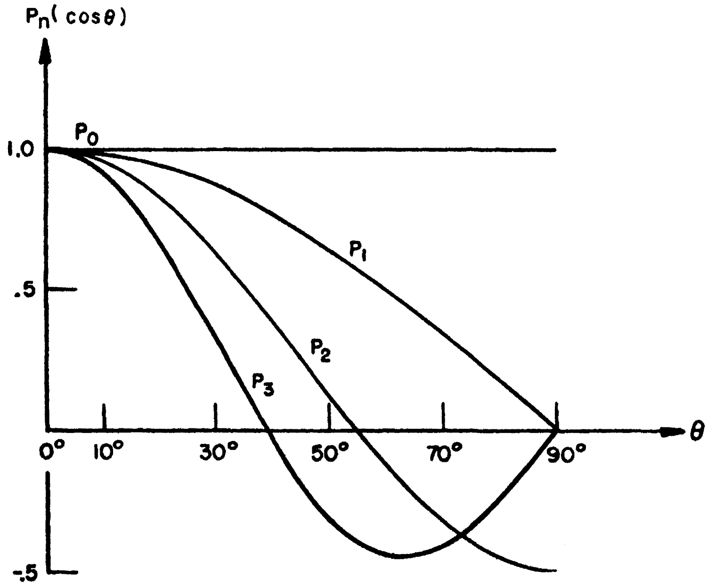

FIG. 5. Legendre polynomials, $P_n(\cos\theta)$ (Abramowitz and Stegun, 1964).

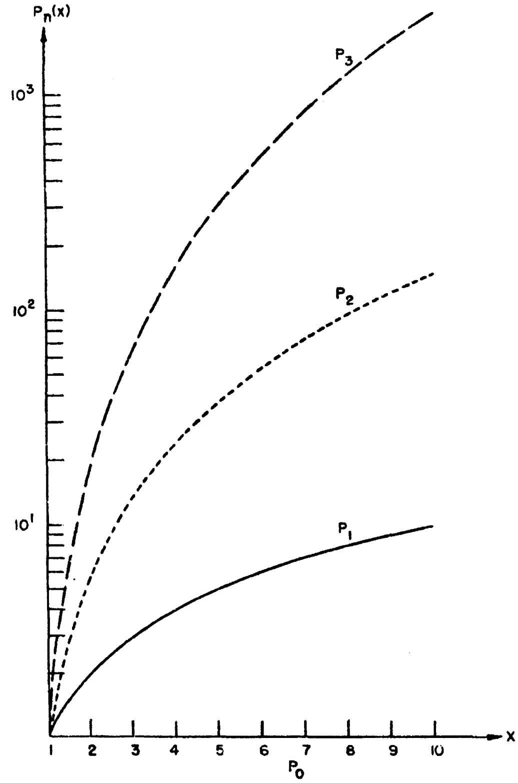

FIG. 6. Legendre polynomials, $P_n(x)$ (Abramowitz and Stegun, 1964).

$$\sin\theta\Theta'' + \cos\theta\Theta' + n(n + 1)\sin\theta\Theta = 0.$$

The general solution of the radial equation is

$$R(r) = Ar^n + B/r^{n+1}.$$

Constants A and B are arbitrary, and the general solution of the theta equation (the Legendre differential equation) is $P_n(\cos\theta)$. We set the coefficient B equal to zero since a solution finite at every point within the sphere is required. The general solution of the problem is a superposition of products of radial and theta solutions; we write

$$T(r,\theta) = \sum_{n=0}^{\infty} A_n r^n P_n(\cos\theta).$$

The coefficients A_n are obtained by use of the boundary condition $T(b,\theta) = f(\theta)$ and the orthogonality relations for the Legendre polynomials. The specific form for $f(\theta)$ must be given if the specific values for A_n are required.

3. THE CONFLUENT HYPERGEOMETRIC FUNCTION, ${}_1F_1(a,c;x)$

The confluent hypergeometric differential equation (also called the Kummer differential equation) has the form

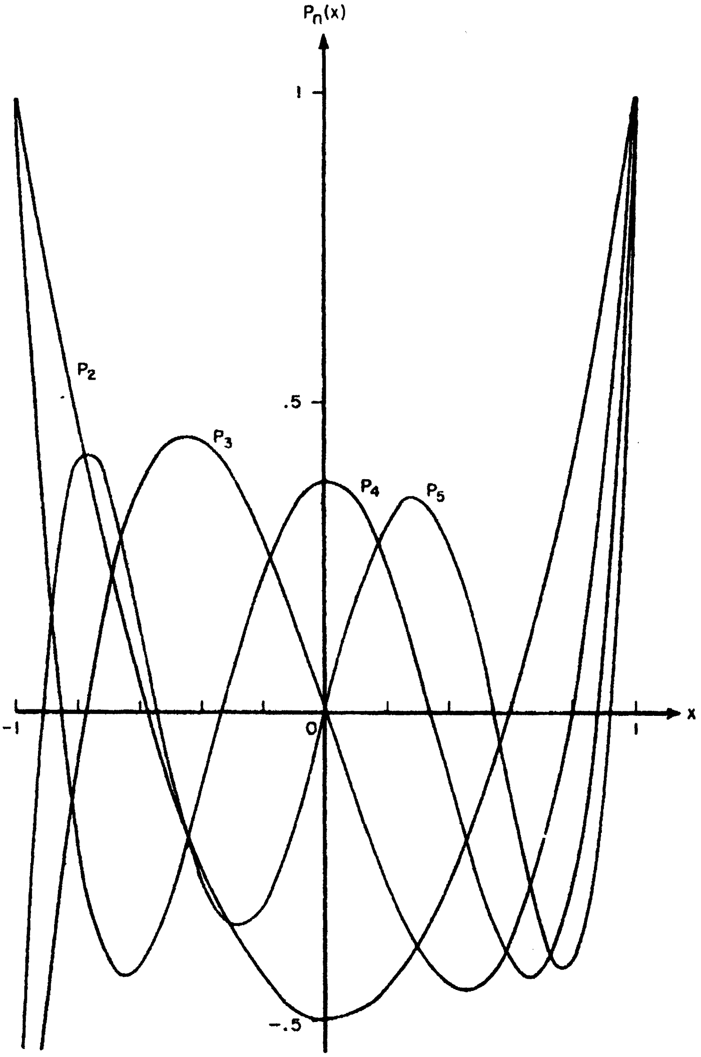

FIG. 7. Legendre polynomials, $P_n(x)$ (Abramowitz and Stegun, 1964).

$$xy'' + (c - x)y' - ay = 0. \tag{53}$$

Equation (53) is obtained from the hypergeometric differential equation by substituting $z = bx$ and letting b approach infinity; this substitution causes a merging or confluence of the two upper singular points. In the confluent hypergeometric differential equation, there is a regular singularity at $x = 0$ and an irregular singularity at $x = \infty$. By use of the power-series method in the neighborhood of $x = 0$, we find that one solution of Eq. (53) has the form (confluent hypergeometric functions or Kummer functions)

$$y(x) \equiv {}_1F_1(a,c;x) = 1 + \frac{ax}{c} + \frac{a(a+1)x}{2!c(c+1)} + \dots$$
$$= \sum_{n=0}^{\infty} \frac{(a)_n x^n}{(c)_n}.$$

The confluent hypergeometric series converges for all values of x.

The Bessel functions and modified Bessel functions in terms of ${}_1F_1$ are respectively given by

$$J_n(x) = \frac{e^{-ix}}{n!}\left(\frac{x}{2}\right)^n {}_1F_1(n + \tfrac{1}{2}, 2n + 1; 2ix)$$

and

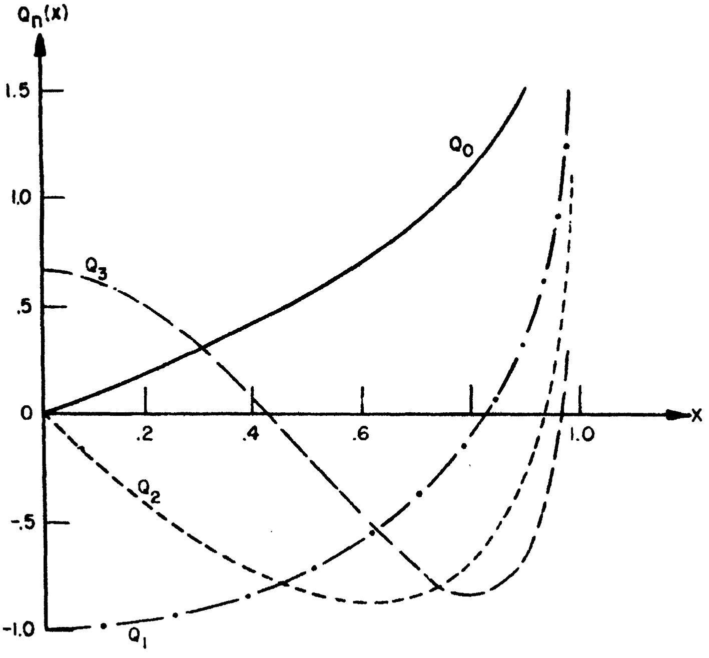

FIG. 8. Legendre functions of the second kind, $Q_n(x)$ (Abramowitz and Stegun, 1964).

$$I_n(x) = \frac{e^{-x}}{n!}\left(\frac{x}{n!}\right)^n {}_1F_1(n + \tfrac{1}{2}, 2n + 1; 2x).$$

The Laguerre and associated Laguerre polynomials in terms of ${}_1F_1$ are respectively given by

$$L_n(x) = {}_1F_1(-n,1;x)$$

and

$$L_n^m(x) = \frac{(n + m)!}{n!m!}\, {}_1F_1(-n, m + 1; x).$$

Hermite polynomials in terms of ${}_1F_1$ have the form

$$H_{2n}(x) = (-1)^n \frac{(2n)!}{n!}\, {}_1F_1(-n, \tfrac{1}{2}; x^2).$$

The error function and complementary error function are respectively defined by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt, \text{ where } \operatorname{erf}(\infty) = 1$$

and

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt.$$

The Hermite polynomials may be obtained from derivatives of the error function as follows:

$$\frac{d^{n+1}\operatorname{erf}(x)}{dx^{n+1}} = (-1)^n \frac{2}{\sqrt{\pi}} H_n(x)e^{-x^2}.$$

In addition, the error function is related to the confluent hypergeometric function by use of the following equation:

$$\operatorname{erf}(x) = \frac{2x}{\sqrt{\pi}}\, {}_1F_1(\tfrac{1}{2},\tfrac{3}{2}; -x^2).$$

3.1 More on Hermite Polynomials

The Hermite differential equation,

$$y'' - 2xy' + 2ny = 0 \quad \text{for } n = \text{constant},$$

is a special case of the Sturm–Liouville differential equation and a special case of the confluent hypergeometric differential equation. The polynomial solution $H_n(x)$ of the Hermite differential equation may be ob-

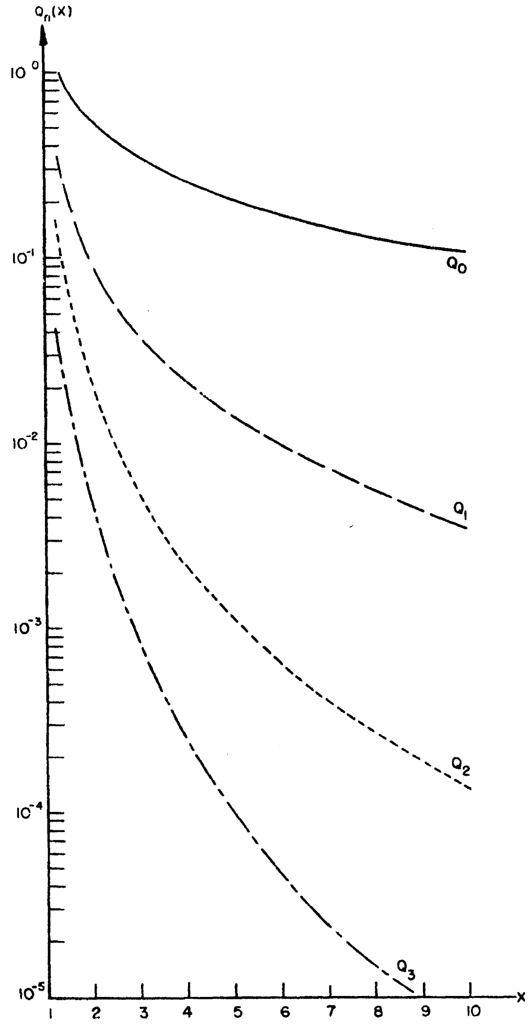

FIG. 9. Legendre functions of the second kind, $Q_n(x)$ (Abramowitz and Stegun, 1964).

tained by use of the Frobenius–Fuchs power-series method,

$$y = \sum_{r=0}^{\infty} a_r x^{k+r}, \text{ where } a_0 \neq 0.$$

For standardization $a_n = 2^n$, the solution has the form

$$y_n \equiv H_n(x) = \sum_{j=0}^{N} \frac{(-1)^j n!(2x)^{n-2j}}{j!(n-2j)!}. \tag{54}$$

Equation (54) is the series form of the Hermite polynomials. In Eq. (54), $N = n/2$ for n even and $N = (n-1)/2$ for n odd. The orthogonality relation, recurrence relation, Rodrigues formula, and generating function for $H_n(x)$ are respectively given in Tables 2, 3, 4, and 5. A representative sketch of $H_n(x)$ is given in Fig. 10.

Example 4: More on the quantum mechanical linear harmonic oscillator.

Solution: The problem of describing the small oscillation of a mass m attached to the end of a spring with force constant k and potential energy $V(x) = kx^2/2$ can be solved exactly in both classical and quantum theory. This system is referred to as a linear harmonic oscillator and is used to represent and analyze more complex physical systems such as vibrations of individual atoms in molecules and in crystals, and classical and quantum theories of radiation. The solution of Schrodinger's wave equation for the linear harmonic oscillator is expressed in terms of Hermite polynomials $H_n(x)$. The equation to be solved is the one-dimensional wave equation for the linear harmonic oscillator, which has the form

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + \frac{kx^2\psi}{2} = E\psi. \tag{55}$$

Note that Eq. (55) is just the one-dimensional Helmholtz equation for which k^2 equals $2m(E - kx^2/2)/\hbar^2$. Solving a problem in quantum mechanics involves finding the wave functions ψ_n and the corresponding eigenenergies E_n. In dimensionless form, Eq. (55) becomes

$$\begin{aligned} &\frac{d^2\psi}{d\xi^2} + (\lambda - \xi^2)\psi = 0, \\ &\quad \text{where } \xi = \left(\frac{m\omega}{\hbar}\right)^{1/2} x,\ \omega^2 = \frac{k}{m}, \\ &\quad \text{and } \lambda = \frac{2E}{\hbar\omega}. \end{aligned} \tag{56}$$

On substituting $\lambda = 1 + 2n$ into Eq. (56), we obtain the Weber differential equation, and the transformation $\psi = \exp(-\xi^2/2)y(\xi)$ reduces the Weber differential equation to the Hermite differential equation. The transformation equation leading to Hermite's differential equation is motivated by use of the Sommerfeld (1949) polynomial method for solving certain differential equations. According to the Sommerfeld method, the so-

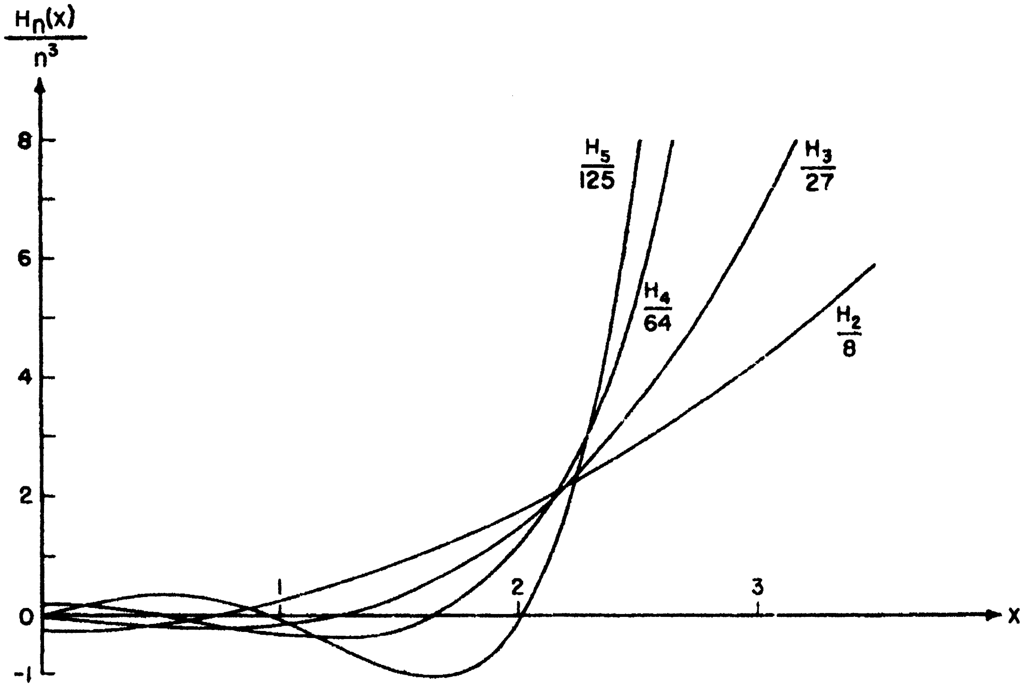

FIG. 10. Hermite polynomials, $H_n(x)/n^3$ (Abramowitz and Stegun, 1964).

lution of Eq. (56) is the product of the physically acceptable asymptotic solution times a polynomial. The resulting asymptotic ($|\xi|$ approaches infinity) differential equation is $\psi'' - \xi^2\psi = 0$, and the physically acceptable solution of this differential equation is given by $\exp(-\xi^2/2)$. In this case, the polynomial part of the solution of Eq. (56) comes from solving Hermite's differential equation. The eigenfunctions and eigenenergy for the linear harmonic oscillator, Eq. (56), are respectively given by

$$\psi_n(\xi) = Ne^{-\xi^2/2}H_n(\xi) \text{ and}$$
$$\lambda_n = 1 + 2n = 2E_n/\hbar\omega \text{ or}$$
$$E_n = \hbar\omega(n + \tfrac{1}{2}).$$

The behavior of $\psi_n(\xi)$ for the first six values of n is illustrated in the sketches in Fig. 11.

3.2 More on the Laguerre and Associated Laguerre Polynomials

The Laguerre differential equation, $xy'' + (1 - x^2)y' + ny = 0$ for n = constant, is a special case of the Sturm–Liouville differential equation as well as a special case of the confluent hypergeometric differential equation; its solution may be obtained by relation to ${}_1F_1$ or by use of the Frobenius–Fuchs power-series method,

$$y(x) = \sum_{\lambda=0}^{\infty} a_\lambda x^{k+\lambda} \text{ for } a_0 \neq 0.$$

The indicial equation in the power-series method has a double root at $k = 0$, and the power-series method yields only one of the two linearly independent solutions of the Laguerre differential equation; this solution is extremely important in mathematical physics. The finite solutions of the Laguerre differential equation, $L_n(x)$, are called Laguerre polynomials. The solution in series form is

$$L_n(x) = \sum_{\lambda=0}^{n} \frac{(-1)^\lambda n! x^\lambda}{(n - \lambda)!(\lambda!)^2};$$
$$a_0 \equiv 1 \text{ (standardization).}$$

The orthogonality and recurrence relations for the Laguerre polynomials are given in Tables 2 and 3, respectively; the generating

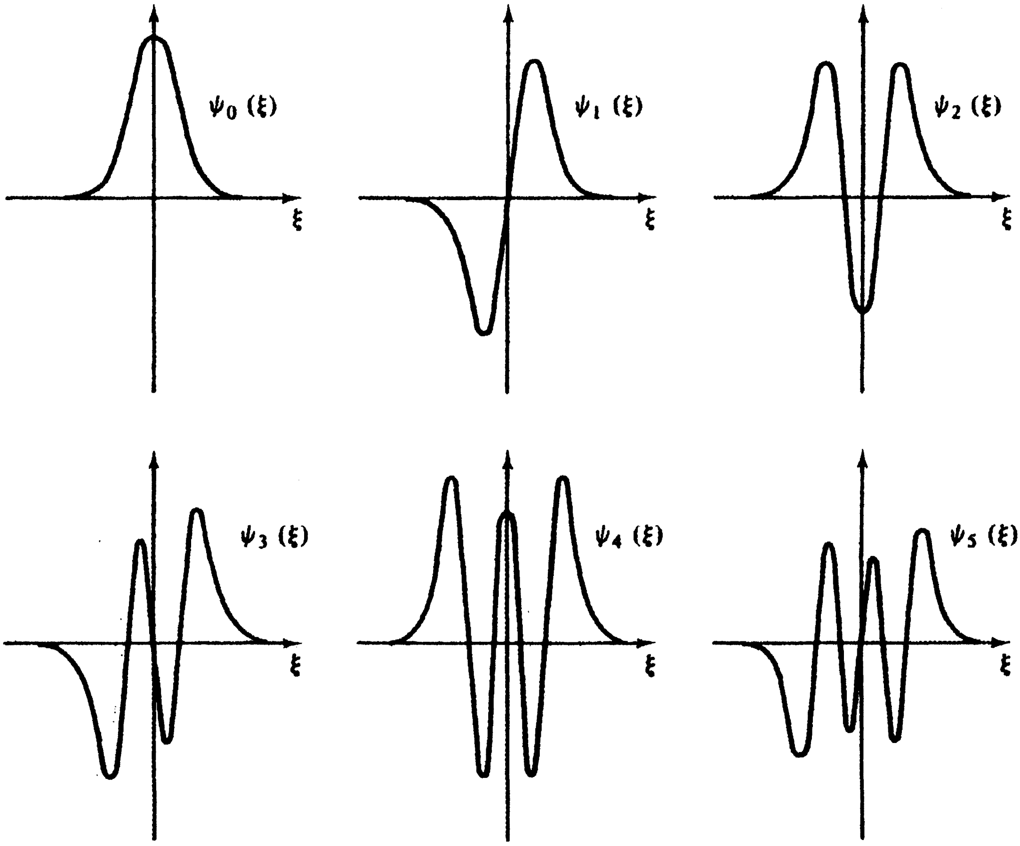

FIG. 11. Linear harmonic oscillator wave functions, $\psi_n(\xi)$.

function and Rodrigues formula are respectively given by

$$\frac{\exp[-xt/(1-t)]}{(1-t)} = \sum_{n=0}^{\infty} \frac{L_n(x)t^n}{n!}$$

and

$$L_n(x) = \frac{1}{n!} e^x \frac{d^n}{dx^n} (x^n e^x).$$

Note that $L_n(0) = 1$, $L_0(x) = 1$, $L_1(x) = 1 - x$, and $L_2(x) = 2 - 4x + x^2$. A sketch of Laguerre polynomials is given in Fig. 12.

Note that the kth derivative of the Laguerre differential equation yields the associated Laguerre differential equation,

$$x \frac{d^2L_n^k(x)}{dx^2} + (k + 1 - x) \frac{dL_n^k(x)}{dx} + (n - k)L_n^k(x) = 0. \qquad (57)$$

In obtaining Eq. (57), the Leibnitz formula, Eq. (16), for finding the kth derivative of a product was used, and we note that the associated Laguerre polynomials $L_n^k(x)$ are related to the Laguerre polynomials by

$$L_n^k(x) = \frac{d^kL_n(x)}{dx^k}.$$

The orthogonality relation, recurrence relation, Rodrigues formula, and generating function for associated Laguerre polynomials are respectively given in Tables 2, 3, 4, and 5.

Example 5, a central-force problem in quantum mechanics.

Solution: The time-independent Schrödinger wave equation is used to study the mechanics of two microscopic particles moving under the influence of a central force (that is to say, the force and potential de-

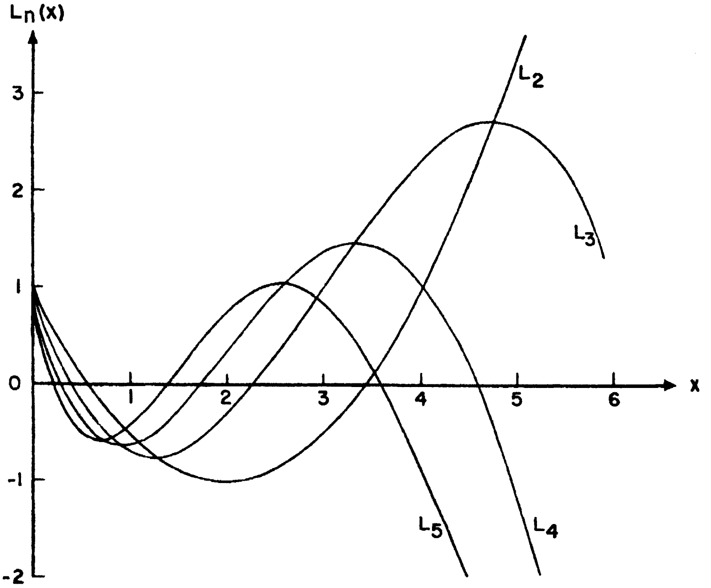

FIG. 12. Laguerre polynomials, $L_n(x)$ (Abramowitz and Stegun, 1964).

pend only on the distance between the two particles). This problem provides the basis for the quantum mechanical treatment of a fundamental class of problems such as the rigid rotator, which is of considerable importance in the study of the spectra of diatomic molecules; the theory of the hydrogen atom; and the nonrelativistic theory of the deuteron. For an attractive inverse-square force law, we substitute $k^2 = 2\mu(E - A/r)/\hbar^2$ into the radial equation; the quantity A is a positive constant. Anticipating use of the well-known solution of the associated Laguerre differential equation, the following substitutions are made in the radial equation for inverse-square force problems:

$$\xi = \beta r,\ \beta^2 \equiv \frac{8\mu|E|}{\hbar^2},\ \text{and}\ \gamma \equiv \frac{2\mu A}{\beta\hbar^2} = \frac{A}{\hbar}\left(\frac{\mu}{2|E|}\right)^{1/2}. \tag{58}$$

By use of the substitutions in Eq. (58), the radial equation, Eq. (48), reduces to

$$\xi\frac{d^2R}{d\xi^2} + 2\frac{dR}{d\xi} + \left\{\gamma - \frac{\xi}{4} - \frac{l(l+1)}{\xi}\right\}R = 0. \tag{59}$$

The symbol l represents the angular momentum quantum number; it is used instead of n in Eq. (48) since n denotes principal quantum number for this problem. Equation (59) is reduced to the associated Laguerre differential equation by use of the transformation

$$R_{nl}(\xi) = \exp(-\xi/2)\xi^l L_{n+1}^{2l+1}(\xi).$$

The desired solution of the equation to be solved, Eq. (59), is a normalization constant times $R_{nl}(\xi)$. A polynomial solution of the associated Laguerre differential equation is obtained when $\gamma = n = l + k + 1$ for $k = 0, 1, 2, \ldots, n - 1$.

Example 6, the hydrogen atom: Obtain the eigenfunction and eigenenergy for the hydrogen atom.

Solution: The hydrogen atom represents a two-body central-force problem in quantum mechanics, where the electron and proton are the two particles under investigation. The Coulomb potential is the central potential for the hydrogen atom. Here, the total energy is negative for bound states, $E < 0$, and the attractive potential energy is given by $V = -e^2/4\pi\epsilon_0 r = -A/r$. The eigenenergy E_n is obtained from Eq. (58); we obtain the usual Bohr result,

$$E_n = -\mu A^2/2\hbar^2 n^2.$$

The corresponding steady-state wave function is the product of solution of the radial part and the angular part, which is $Y_l^m(\sigma,\phi)$ times a normalization constant, C_{nl}; the result is

$$\psi(r,\theta,\phi) = -C_{nl}R_{nl}(\xi)Y_l^m(\theta,\phi). \qquad (60)$$

The eigenfunction solution, Eq. (60), is given in most quantum mechanics books, and the normalization constant is obtained in the usual manner.

4. HELMHOLTZ'S DIFFERENTIAL EQUATION IN CYLINDRICAL COORDINATES

Problems in mathematical physics that involve cylindrical geometry are, in general, simpler to solve in cylindrical coordinates (ρ,ϕ,z) than in Cartesian coordinates. The Helmholtz differential equation in cylindrical coordinates has the form

$$\rho\frac{\partial^2 u}{\partial\rho^2} + \frac{\partial u}{\partial\rho} + \frac{1}{\rho}\frac{\partial^2 u}{\partial\phi^2} + \rho\frac{\partial^2 u}{\partial z^2} + k^2\rho u = 0.$$

By use of the separation-of-variables method for $u(\rho,\phi,z) = P(\rho)\Phi(\phi)Z(z)$ with separation constants $-\lambda^2$ and $-n^2$, the following three ordinary differential equations are obtained:

$$\frac{1}{Z}\frac{d^2 Z}{dz^2} = -\lambda^2$$

with solution $Z(z) = A\cos(\lambda z) + B\sin(\lambda z)$,

$$\frac{1}{\Phi}\frac{d^2\Phi}{d\phi^2} = -n^2$$

with solution $\Phi(\phi) = C\cos(n\phi) + D\sin(n\phi)$, and

$$\xi^2\frac{d^2P}{d\xi^2} + \xi\frac{dP}{d\xi} + (\xi^2 - n^2)P = 0 \text{ for } \xi = \alpha\rho$$
$$\text{and } k^2 - \lambda^2 \equiv \alpha^2. \qquad (61)$$

4.1 Solutions of Bessel's Differential Equation

Equation (61) is Bessel's differential equation, and its solutions are called Bessel (or cylindrical) functions. Bessel's differential equation is solved by use of the power-series method,

$$P(\xi) = \sum_{\lambda=0}^{\infty} a_\lambda \xi^{k+\lambda} \text{ for } a_0 \neq 0.$$

The general solutions of Bessel's differential equation when n is an integer are

$$P(\xi) = AJ_n(\xi) + BN_n(\xi);\ n = \text{integer}.$$

The functions $J_n(\xi)$ and $N_n(\xi)$ are explained in the following sections.

4.2 Bessel Functions of the First Kind

The functions $J_n(\xi)$ are called Bessel functions of the first kind; the series representation, generating function, and recurrence relation for $J_n(\xi)$ are respectively given by

$$J_n(\xi) = \sum_{j=0}^{\infty}\frac{(-1)^j(\xi/2)^{2j+1}}{j!\Gamma(n+j+1)}$$
$$\text{for standardization } a_0 \equiv \frac{1}{2^n\Gamma(n+1)}, \qquad (62)$$

$$\exp\left\{\tfrac{1}{2}\xi\left(t - \frac{1}{t}\right)\right\} = \sum_{n=-\infty}^{\infty} J_n(\xi)t^n,$$

and

$$J_{n-1}(\xi) + J_{n+1}(\xi) = \frac{2n}{\xi}J_n(\xi).$$

Note that $J_{-n}(\xi) = (-1)^n J_n(\xi)$ results from use of Eq. (62). The gamma function Γ is defined in Sec. 5.2. The orthogonality relation for the interval $[0,a]$ may be written in the form

$$\int_0^a J_n\left(\beta_{ni}\frac{\xi}{a}\right)J_n\left(\beta_{nj}\frac{\xi}{a}\right)\xi d\xi = \begin{cases} 0, i \neq j, \\ \dfrac{a^2}{2}[J_{n+1}(\beta_{nj})]^2, \ i = j. \end{cases} \qquad (63)$$

In Eq. (63), $n > -1$, the parameter β_{ni} is the ith zero of J_n, and $0 \leq \xi \leq a$.

The general solutions of Bessel's differential equation when n is not an integer are

$$P(\xi) = CJ_n(\xi) + DJ_{-n}(\xi);\quad n \neq \text{integer}.$$

Sketches of several Bessel functions of the first kind are given in Fig. 13.

4.3 Neumann Functions

The Neumann functions are defined by

$$Y_n(\xi) = N_n(\xi)$$
$$\equiv \frac{J_n(\xi)\cos n\pi - J_{-n}(\xi)}{\sin n\pi};\quad n = \text{integer}.$$

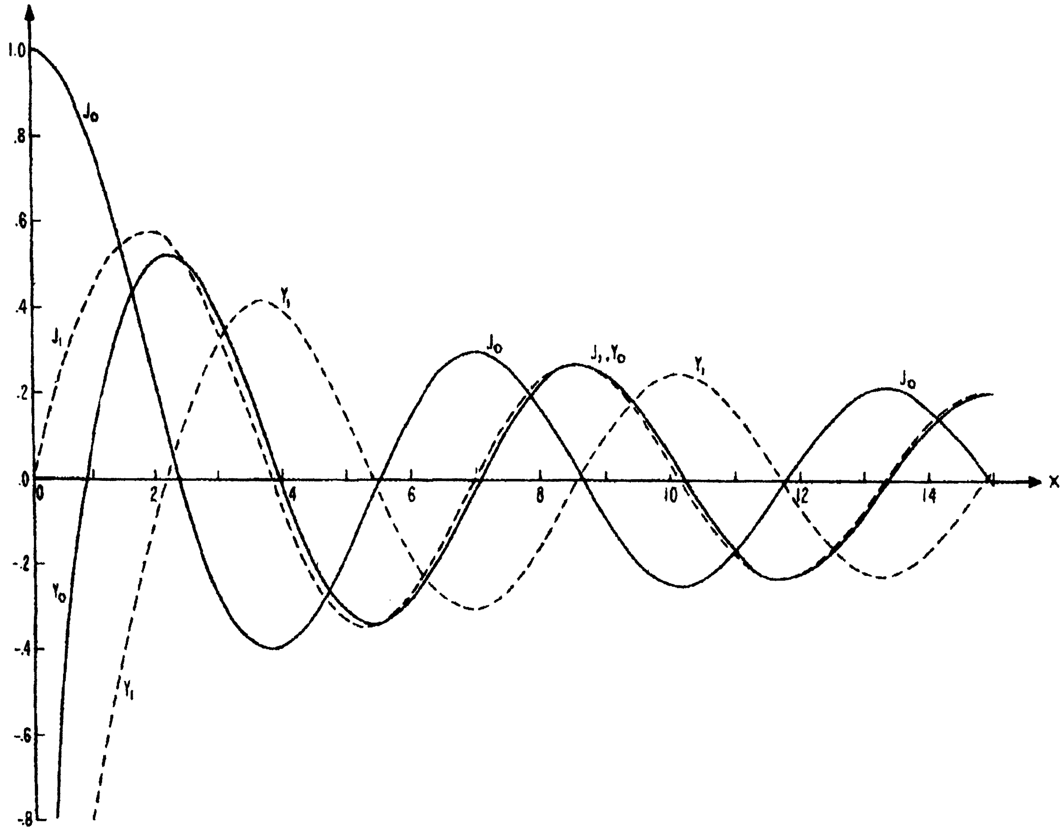

FIG. 13. Bessel functions of the first and second kinds, $J_n(x)$ and $Y_n(x)$ (Abramowitz and Stegun, 1964).

The Neumann functions are called Bessel functions of the second kind. L'Hospital's rule should be used to evaluate $N_n(\xi)$. Sketches of several Neumann functions are given in Fig. 13.

4.4 Hankel Functions

Hankel functions of the first and second kind are respectively defined by

$$\begin{aligned} H_n^{(1)}(\xi) &\equiv \frac{i}{\sin n\pi}\left[e^{-n\pi i}J_n(\xi) - J_{-n}(\xi)\right] \\ &= J_n(\xi) + iN_n(\xi) \end{aligned}$$

and

$$\begin{aligned} H_n^{(2)}(\xi) &\equiv -\frac{i}{\sin n\pi}\left[e^{n\pi i}J_n(\xi) - J_{-n}(\xi)\right] \\ &= J_n(\xi) - iN_n(\xi). \end{aligned}$$

The Hankel functions are independent solutions of the Bessel differential equation, and they are useful in connection with their behavior for large values of ξ since they are infinite at $\xi = 0$. Hankel functions are sometimes referred to as Bessel functions of the third kind.

4.5 Modified Bessel Functions

The modified Bessel differential equation is obtained when ξ is replaced with it in Eq. (61); the result is

$$t^2\frac{d^2P}{dt^2} + t\frac{dP}{dt} - (t^2 + n^2)P = 0. \tag{64}$$

The solutions of Eq. (64) are called modified Bessel functions of the first kind and are denoted by $I_n(\xi)$; they are given by

$$I_n(\xi) = i^{-n}J_n(i\xi) = \sum_{\lambda=0}^{\infty}\frac{(\xi/2)^{2\lambda+n}}{\lambda!(\lambda+n)!}; \quad n = \text{integer}.$$

When n is not an integer, $I_n(\xi)$ and $I_{-n}(\xi)$ are linearly independent solutions of the modified Bessel differential equation, Eq. (64). When n is an integer, $I_n(\xi) = I_{-n}(\xi)$. The modified Bessel functions of the second kind, $K_n(\xi)$, are defined by

$$K_n(\xi) \equiv \frac{\pi}{2}\left\{\frac{I_n(\xi) - I_{-n}(\xi)}{\sin n\pi}\right\}.$$

The modified Bessel functions of the second kind are well behaved for all values of n. Sketches of several modified Bessel functions are given in Fig. 14.

4.6 Spherical Bessel Functions

Solutions of the radial equation, Eq. (48), for k^2 = constant are obtained by comparing the radial equation with the Bessel differential equation; these solutions are called spherical Bessel functions and have the form

$$R(\xi) = Aj_n(\xi) = \left(\frac{\pi}{2\xi}\right)^{1/2} J_{n+1/2}(\xi) \text{ for } \xi = kr.$$

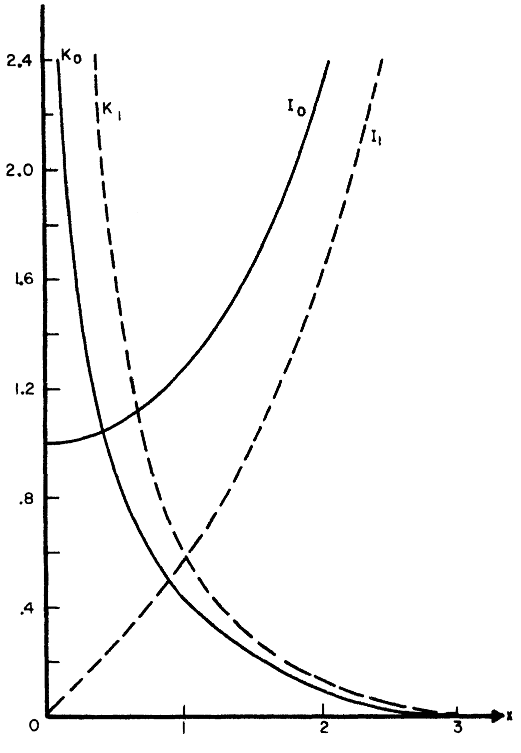

FIG. 14. Modified Bessel functions, $K_n(x)$ and $I_n(x)$ (Abramowitz and Stegun, 1964).

Spherical Bessel functions are often used in quantum mechanics and in other areas of physics. Sketches of several spherical Bessel functions, spherical Neumann functions, and spherical modified Bessel functions are respectively given in Figs. 15, 16, and 17.

Example 7, vibrations of a circular membrane: The displacement $u(r,\theta,t)$ of a stretched circular membrane with mass per unit area μ and under tension T satisfies the two-dimensional mechanical wave equation in plane polar coordinates (r,θ), which may be written in the following form:

$$\frac{1}{r}\left\{\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{\partial}{\partial \theta}\left(\frac{1}{r}\frac{\partial u}{\partial \theta}\right)\right\} = \frac{1}{v^2}\frac{\partial^2 u}{\partial t^2}.$$

The speed of the wave motion is defined by $v = \sqrt{T/\mu}$. Develop the solution of this equation for a vibrating drum head.

Solution: Separating the variables $u(r,\theta,t) = R(r)\Theta(\theta)T(t)$ with separation constants $-\alpha^2$ and $-n^2$ yields the following three ordinary differential equations:

$$\frac{d^2T}{dt^2} + \omega^2 T = 0,$$

with solution $T(t) = A\cos\omega t + B\sin\omega t$, where $\omega^2 \equiv v^2\alpha^2$;

$$\frac{d^2\Theta}{d\theta^2} + n^2\Theta = 0,$$

with solution $\Theta(\theta) = C\cos n\theta + D\sin n\theta$; and

$$\xi^2\frac{d^2R}{d\xi^2} + \xi\frac{dR}{d\xi} + (\xi^2 - n^2)R = 0,$$

where $\xi \equiv \alpha r$. The solution of the last equation is $R(\xi) = EJ_n(\xi) + FN_n(\xi)$ since it is the Bessel differential equation. In addition, it is required that the solution be finite at $\xi = 0$; hence, F is set equal to zero since $N_n(\xi)$ approaches infinity as ξ approaches zero. The general solution for the motion of the drum head is therefore given by

$$u = [A\cos\omega t + B\sin\omega t] \times [C\cos n\phi + D\sin n\phi]EJ_n(\xi).$$

Since the membrane is fixed (no vibration, $u = 0$) around the edge where $r = b$ (radius of

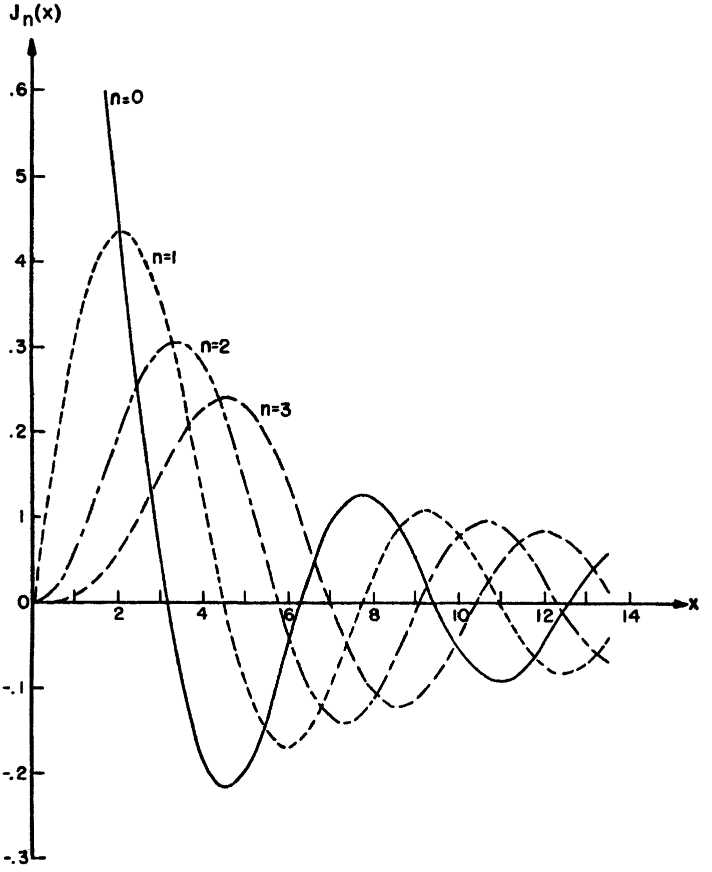

FIG. 15. Spherical Bessel functions of the first kind, $j_n(x)$ (Abramowitz and Stegun, 1964).

the head), the drum head vibrates in circular modes such that $EJ_n(\xi) = 0$. The nodes are located at $\alpha r = \xi_k$, where ξ_k are the values of ξ for which $J_n(\xi)$ has a zero. A single term in the solution corresponds to a standing wave whose modes are concentric circles, and the complete solution is obtained by summing over all such modes of vibration.

5. OTHER SPECIAL FUNCTIONS USED IN MATHEMATICAL PHYSICS

As explained in the Introduction of this article, the list of special functions is extensive. The *Handbook of Mathematical Functions* (Abramowitz and Stegun, 1964) contains a fairly comprehensive list of special functions. The focus in this article has been on special functions that are widely used in mathematical physics to solve classes of problems whose formulations involve special cases of the Helmholtz differential equation. In general, special functions may be classified as Type 1, those special functions that satisfy a differential equation, or Type 2, special functions that do not satisfy a differential equation; for example, the gamma function is a Type 2 special function. In Secs. 5.1 and 5.2, we summarize some other special functions that are used in mathematical physics.

5.1 Some Other Special Functions—Type 1

Some other special functions that satisfy a differential equation are the following:

1. Airy functions are solutions of the Airy differential equation, which has the form $y'' - xy = 0$; the Airy differential equation characterizes constant-force–type problems in quantum mechanical and in elementary particle physics.

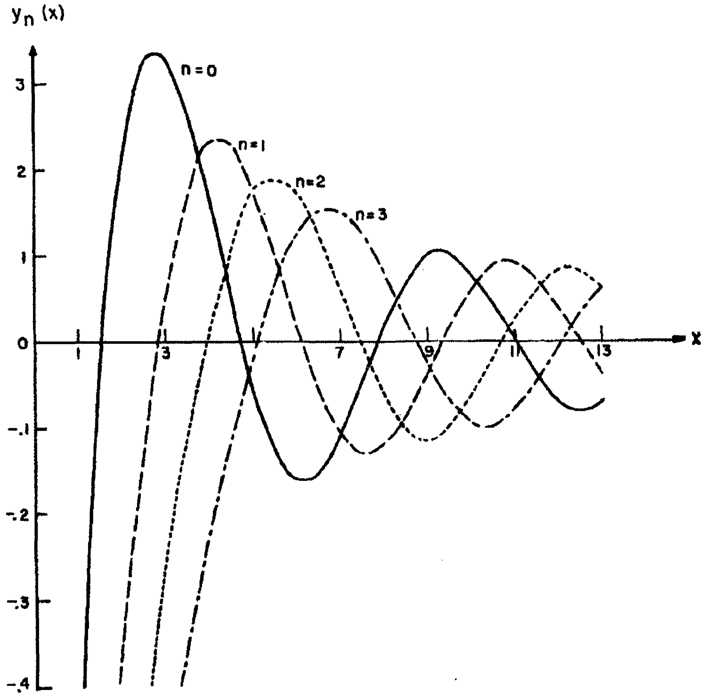

FIG. 16. Spherical Bessel functions of the second kind, $y_n(x)$ (Abramowitz and Stegun, 1964).

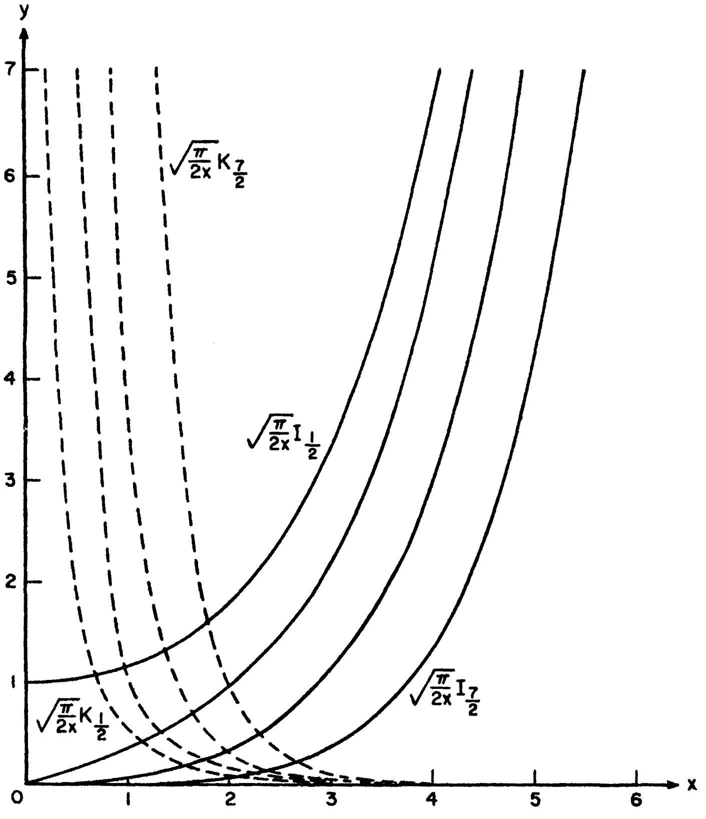

FIG. 17. Modified spherical Bessel functions of the first and second kinds (Abramowitz and Stegun, 1964).

2. Mathieu functions are solutions of the Mathieu differential equation, which has the form $y'' + (a - 2b\cos 2x)y = 0$; the Mathieu differential equation results when a cosine-type potential is substituted into the one-dimensional time-independent Schrödinger wave equation.
3. Parabolic cylinder functions are connected with confluent hypergeometric functions and with Hermite polynomials; they are solutions of differential equations of the form $y'' + (ax^2 + bx + c)y = 0$.

Many bound-state and collision problems in classical and quantum mechanics as well as in other areas of physics involve integrals of the form

$$\int R(x,y)dx. \tag{65}$$

When $R(x,y)$ is a rational function of x and y and $y^2 = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4$, where $a_4 \neq 0$ or $a_4 = 0$ and $a_3 \neq 0$, the integral in Eq. (65) is called an elliptic integral. Complete elliptic integrals of the first and second kind are respectively defined as

$$K(m) = \int_0^{\pi/2} (1 - m\sin^2\theta)^{-1/2}d\theta$$

and

$$E(m) = \int_0^{\pi/2} (1 - m\sin^2\theta)^{1/2}d\theta.$$

Analyzing the motion of a simple pendulum involves an elliptic integral of the first kind. Elliptic integrals can be evaluated directly by use of series expansions or computers. Their importance in mathematical physics, however, is related to their appearance in the solution of physical problems involving certain nonlinear differential equations. Elliptic integrals are special cases of the hypergeometric functions since

$$K(m) = \frac{\pi}{2}F(\tfrac{1}{2},\tfrac{1}{2},1;m) \text{ and}$$

$$E(m) = \frac{\pi}{2}F(-\tfrac{1}{2},\tfrac{1}{2},1;m).$$

The *Handbook of Mathematical Functions* (Abramowitz and Stegun, 1964) is a good reference for additional information on elliptic integrals.

5.2 Some Other Special Functions—Type 2

Some special functions used in mathematical physics that do not satisfy a differential equation are Einstein and Debye functions, which are used in representing the specific heats of solids due to lattice vibrations; error function (defined in Sec. 3.1); gamma function; and beta function. The last two functions are widely used in many areas of mathematical physics and are now summarized.

The factorial, $n!$, is defined as

$$n! = n(n-1)\ldots 2\cdot 1 = \int_0^\infty e^{-t}t^n dt$$

for integer values of n. Note that $0! = 1$, and $n! = \pm\infty$ if n equals a negative integer. The gamma function, Γ, is a generalization of the factorial to cases of noninteger values for n. The Euler definition of the gamma function is

$$\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}dt \quad \text{for } \mathrm{Re}(z) > 0. \tag{66}$$

In Eq. (66), $\mathrm{Re}(z)$ denotes the real part of $z = x + iy$. Note that $\Gamma(\tfrac{1}{2}) = \pi^{1/2}$. Evaluating the integral in Eq. (66) by parts yields the recurrence relation for the gamma function, $\Gamma(z+1) = z\Gamma(z)$. If z is a positive integer n, then $\Gamma(z+1)$ equals $n!$.

The gamma function is used to express, in compact form, solutions of many problems of mathematical physics. The gamma function, however, does not satisfy a differential equation that is related to a physical problem; in fact, the gamma function does not satisfy any differential equation with rational coefficients. A sketch of $\Gamma(x)$ for some positive and negative values is given in Fig. 18.

The beta function, $B(x,y)$, is defined by use of an integral, and it involves a simple and useful combination of gamma functions; it has the form

$$B(p,q) \equiv \int_0^1 t^{p-1}(1-t)^{q-1}dt = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \text{ for } \mathrm{Re}(p) > 0,\ \mathrm{Re}(q) > 0.$$

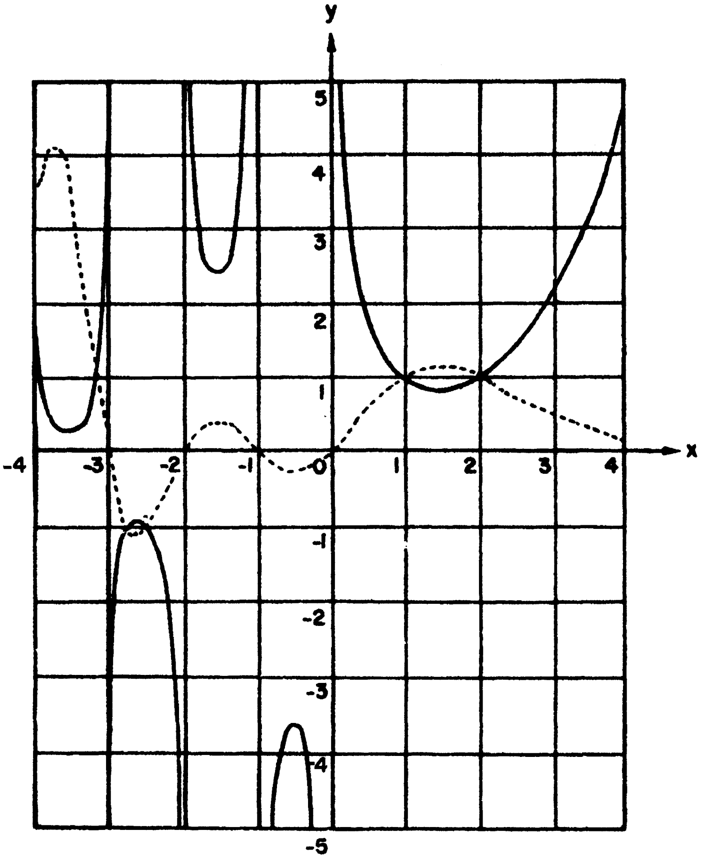

FIG. 18. Gamma function: solid curve, $\Gamma(x)$; dashed curve, $1/\Gamma(x)$ (Abramowitz and Stegun, 1964).

Note that $B(p,q) = B(q,p)$. The beta function is frequently used in high-energy particle physics as well as other areas of mathematical physics.

GLOSSARY

Confluent Hypergeometric Function: A solution of the confluent hypergeometric differential equation, which is a second-order ordinary differential equation with a regular singularity at $x = 0$ and an irregular singularity at $x = \infty$. Laguerre and associated Laguerre polynomials and Hermite polynomials are special cases of the confluent hypergeometric function.

Gamma Function: A generalization, to noninteger values of n, of the factorial $n!$.

Generating Function: A function $\mathcal{G}(x,t)$ that, when expanded in a power series with respect to t, contains the functions (set of polynomials) to be generated as coefficients of the parameter t.

Hypergeometric Function: A generalized geometric series that is the solution of the hypergeometric differential equation, which is a second-order ordinary differential equation with regular singular points at 0,1, and infinity. For example, the Legendre polynomials and associated Legendre functions are special cases of the hypergeometric function.

Rodrigues Formula: A formula used to generate an orthogonal polynomial by taking the nth derivative of an elementary function.

Special Functions: Higher transcendental functions used in mathematical physics.

Sturm–Liouville Theory: The theory devoted to determining the dependence of eigenfunctions on eigenvalues and the dependence of eigenvalues on boundary conditions imposed on eigenfunction solutions of the Sturm–Liouville equation.

Works Cited

Abramowitz, M., Stegun, I. A. (Eds.) (1964), *Handbook of Mathematical Functions,* New York: Dover Publications.

Courant, R., Hilbert, D. (1953), *Methods of Mathematical Physics,* Vol. 1, New York: Interscience Publishers.

Infeld, L., Hull, T. E. (1951), *Rev. Mod. Phys.* **23,** 21–68.

Miller, W. (1968), *Lie Theory and Special Functions,* New York: Academic.

Schrödinger, E. (1940), *Proc. Roy. Irish Acad.* **A46,** 9.

Sommerfeld, A. (1949), *Partial Differential Equations in Physics,* New York: Academic.

Talman, J. D. (1968), *Special Functions: A Group Theoretic Approach,* New York: W. A. Benjamin, Inc.

Vilenkin, N. J. (1968), *Special Functions and the Theory of Group Representations,* Translations of Mathematical Monographs, Vol. 22, Providence, RI: American Mathematical Society.

Whitaker, E. T., Watson, G. N. (1947), *Modern Analysis,* New York: Macmillan Company.

Further Reading

Arfken, G., Weber, H. J. (1995), *Mathematical Methods for Physicists,* New York: Academic.

Beckmann, P. (1973), *Orthogonal Polynomials for Engineers and Physicists,* Boulder, CO: The Golem Press.

Erdelyi, A., Magnus, W., Oberhettinger, F., Tricomi, F. G. (Eds.) (1953), *Higher Transcendental Functions,* Vols. 1, 2, and 3, New York: McGraw-Hill.

Gilmore, R. (1974), *Lie Groups, Lie Algebras, and Some of Their Applications,* New York: Wiley.

Morse, P. M., Feshbach, H. (1953), *Methods of Theoretical Physics,* Vols. 1 and 2, New York: McGraw-Hill.

Sattinger, D. H., Weaver, O. L. (1986), *Lie Groups and Algebras with Applications to Physics, Geometry, and Mechanics,* New York: Springer-Verlag.

Wang, Z. X., Guo, D. R. (1989), *Special Functions,* Singapore: World Scientific.

SPECTROMETERS, ELECTRON AND ION

LÁSZLÓ KÖVÉR, *Institute of Nuclear Research of the Hungarian Academy of Sciences, Debrecen, Hungary*

KAI SIEGBAHN, *University of Uppsala, Uppsala, Sweden*

	Introduction	225
1.	**Basic Concepts, Definitions, and Relations**	225
1.1	Liouville's Theorem	225
1.2	Motion of Charged Particles in Static Fields	226
1.3	General Nomenclature of Parameters Describing the Properties of Electron and Ion Spectrometers	226
2.	**Magnetic Spectrometers**	228
2.1	Momentum-Dispersive Spectrometers	228
2.2	Magnetic Transport Spectrometers	230
2.3	New Applications	233
3.	**Electrostatic Spectrometers**	235
3.1	Dispersive-Type Analyzers	236
3.1.1	Electrostatic Mirror Analyzers	236
3.1.2	Deflectors	240
3.2	Nondispersive-Type Analyzers	241
3.3	Auxiliary Electron and Ion Optics for Transport, Retardation, and Imaging	244
3.4	New Trends in Applications	246
4.	**Spectrometers with Combined Electric and Magnetic Field**	248
5.	**Electron and Ion Detection Systems**	248
	Acknowledgment	251
	Works Cited	251

INTRODUCTION

Charged-particle analyzers are nowadays applied in numerous fields of research to give detailed information about the spectral distributions of electrons and ions emitted from materials under irradiation of various kinds. Such analyzers have long been used extensively to study the radiation from radioactive species. In surface science precise measurements of energies and emission directions of photoelectrons can supply basic and detailed information about electronic structures. Normally such experiments are performed under ultrahigh-vacuum conditions. Similar studies can be performed on free molecules or clusters of atoms or molecules preferably in the form of adiabatically cooled jet beams. Even liquids can be studied if adequate measures are taken for differential pumping of the vapor phase and if arrangements for continuous renewing of the liquid surface exposed to vacuum are made. A range of suitable detectors and electronics has been developed to meet the needs for various experimental situations. Channel-plate detectors are highly efficient tools for these purposes and can be adopted to record particles simultaneously over an extended area. In the following review the basic properties of the various types of analyzers are discussed. Within the scope of this survey, no attempts have been made to include applications in different fields of research.

1. BASIC CONCEPTS, DEFINITIONS, AND RELATIONS

1.1 Liouville's Theorem

It is assumed that the charged particles to be transported, analyzed, and detected are under the action of conservative forces only (due to, e.g., static magnetic and electric fields). Each particle then can be characterized by a point in a six-dimensional phase space, where two three-coordinates represent

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

its geometric position and the components of its momentum vector, respectively. Liouville's theorem, which is fundamental for electron and ion optics, states that in these circumstances the density of points in phase space remains constant throughout the motion of the particles.

As a consequence, the volume of the phase space occupied by the particles does not change, since their number is constant, so that in their phase space these particles behave like an incompressible fluid. For a beamlike particle ensemble, the limits of the area it occupies in the phase subspace containing coordinates perpendicular to the beam direction is the *emittance* of the beam and is an invariant property as a consequence of Liouville's theorem (its shape and orientation, of course, can change) assuming a constant axial momentum component. When the axial momentum is varying along the trajectory, then the emittance varies in inverse proportion. In the case of a paraxial beam having cylindrical symmetry, the emittance is the product of the maximum transverse displacement from the symmetry axis and the angular divergence or aperture angle. The limiting area in the same phase subspace is the *acceptance* of an optical system. For achieving an optimum efficiency for transport or imaging, the matching of the respective emittance and acceptance is necessary.

In the case of small aperture angles, the *Lagrange–Helmholtz equation* holds:

$$\sqrt{E_1}l_1\alpha_1 = \sqrt{E_2}l_2\alpha_2, \tag{1}$$

where E_1 and E_2 are the energy values of the particles in the planes 1 and 2 perpendicular to the beam direction, while l_i and α_i are the respective transverse displacements and aperture angles. This equation is fundamental for focusing of paraxial rays of charged particles.

From Liouville's general theorem one can derive a useful relationship between the source width s, the spectrometer angular acceptance interval α, and the electron energy E:

$$s\alpha\sqrt{E} = \text{const.}$$

Thus, if the electrons are retarded before being analyzed in a given spectrometer, the Lagrange–Helmholtz equation regulates the consequences for the three parameters. This is a frequently occurring case when high resolution is aimed at.

1.2 Motion of Charged Particles in Static Fields

A particle with charge q and velocity v moving in a homogeneous magnetic field B in a plane perpendicular to the lines of force describes a circular path with a radius of curvature ρ. The equation of motion is

$$Bqv = mv^2/\rho, \tag{2}$$

where $m = m_0/\sqrt{1 - v^2/c^2}$. Thus, the momentum is given by

$$p = mv = qB\rho. \tag{3}$$

Because the momentum of the focused particles is proportional to the magnetic field of the spectrometer, it is convenient to express the momentum of a particle (e.g., electron) by means of its $B\rho$ value, which gives information about the magnitude of the magnetic field and spectrometer dimension required in a given experiment. The relationship between the relative energy change $\Delta E/E$ and a respective relative change in the momentum of the particle is

$$\frac{\Delta E}{E} = \left(1 + \frac{\mu}{\mu + E}\right)\frac{\Delta(B\rho)}{B\rho}, \tag{4}$$

where μ denotes the rest-mass energy of the particle. In the case of electrostatic spectrometers, however, it is the energy of the particles that is proportional to the deflecting voltage, and equations similar to (3) and (4) are presented in Secs. 1.3 [Eq. (6)] and 3.1 [Eq. (21)].

1.3 General Nomenclature of Parameters Describing the Properties of Electron and Ion Spectrometers

In this section, the most important parameters specifying and describing electron and ion spectrometers are listed, focusing on the dispersive types. The following definitions contain the energy of the particles, E; similar relations are used in the case of momentum-dispersive devices (magnetic spectrometers), where E should be replaced by the respective momentum values p or $B\rho$.

A charged-particle spectrometer provides a selective transport of the particles emitted from a geometrically well-defined source to the similarly well-defined position detector. The energy interval over which particles are accepted is the *energy range* of the spectrometer, while the *solid angle* Ω of the spectrometer defines the angular region from which particles leaving the source are entering the spectrometer. In the case of sources having a finite *source area,* Ω usually means the solid angle averaged over this area.

The fraction of all monochromatic particles leaving the source and entering the spectrometer that is detected is called the *transmission* of the spectrometer. For a point source, the *point transmission* is the solid angle over which the particles are accepted by the spectrometer multiplied by a factor of less than unity, expressing the intensity loss inside the spectrometer. The *efficiency* of the spectrometer is given by the ratio of the total current detected to the total particle current emitted from the monochromatic source. Characteristic for a specific dispersive spectrometer is the *"central"* or *"main" particle trajectory,* along which a particle entering the spectrometer in a direction tangential to this trajectory moves when the spectrometer is set to the energy of the particle, E. The source is centered on the point where the central trajectory crosses the source plane, and the detector (exit slit) should be positioned (or centered) on the point where this trajectory crosses the plane of the image of the source produced by the spectrometer. On increasing the energy (momentum) of the particle moving in the central trajectory by a small amount dE (dp), the corresponding image point will be shifted by a small displacement dx. The measure of this energy (momentum) dispersive property is the *dispersion* D of the spectrometer, expressed as

$$D = E\frac{dx}{dE}\left(\text{or}\; p\,\frac{dx}{dp}\right). \tag{5}$$

Usually, D is related to the characteristic distance of the spectrometer (focal length or path length along the central trajectory), and this ratio is called *relative dispersion.*

Because of spherical aberrations and other possible imaging errors of the electron or ion optical system of the spectrometers, monochromatic particles emitted from a source (with a constant number of particles emitted per unit time from unit source area within unit solid angle, i.e., with constant *brightness*) and entering the spectrometer will have an apparent energy distribution at the detector in the image plane. The shape and intensity of this distribution depends on different factors (including E, Ω, source size, and type of focusing), and it can be defined as the *spectrometer function* S. For a particle energy E, the energy width (full width at half maximum, FWHM, or base width) of $S(E)$ is the *absolute energy resolution* ΔE of the spectrometer. The *relative energy resolution* $\Delta E/E$ for energy-dispersive (electrostatic) spectrometers can be expressed as a function of the absolute image magnification M, the widths of the inlet and exit slits of the spectrometer (s_1 and s_2), and the dispersion D:

$$\frac{\Delta E}{E} = \frac{Ms_1 + s_2 + \Sigma}{D}, \tag{6}$$

where Σ denotes the contribution coming from the aberrations (here ΔE means base width). Analogous expressions (see later) can be found for the *relative momentum resolution* $\Delta p/p$ in the case of momentum-dispersive spectrometers. A similarly often-used quantity is the *resolving power* $E/\Delta E$ (ΔE usually defined as FWHM, which is about half of the base width, depending on the shape of the spectrometer function), or $p/\Delta p$. For most high-transmission spectrometers, the resolution is 1% or worse, while in the case of high-resolution devices, its value is better than 0.01%.

In the case of extended sources, as an alternative to the transmission used as a measure of the sensitivity for a finite-dimension source, the *luminosity* is applied:

$$L = \Omega\sigma, \tag{7}$$

where σ is the source area. It should be noted here that the luminosity can be increased considerably by increasing the number of detectors in the focal surface (energy or momentum multidetection), or—e.g., in the case of electrostatic spectrometers—by applying pre-retardation of particles before energy analysis.

Spectrometers are often classified on the basis of their focusing properties. When the image width of a point source, expressed as an expansion in powers of the angle, does not contain the first n powers, the spectrometer has nth-order focusing. Focusing in a single direction (nth order) produces a line image of a point, while if the same order of focus exists in the perpendicular direction, the spectrometer is *double focusing*, and for stigmatic focusing (producing images in the same plane in each direction) it provides a point image of a point source.

2. MAGNETIC SPECTROMETERS

Magnetic spectrometers have a long and extended use in the separation and analysis of charged particles. Comprehensive reviews are available (Siegbahn, 1965; Mladjenovic, 1979; Dragoun, 1983) in this field, providing detailed descriptions of the basic principles, theory, design, and applications of these experimental instruments. A classical field where magnetic spectrometers have been of great importance is nuclear structure analysis including the alpha- and beta-ray spectroscopy of radioactive samples, as well as the (in-beam) spectroscopy of particles emitted from the interaction of a target with a high-energy ion beam produced by a particle accelerator. High-resolution spectrometers are used for the spectroscopy of electrons emitted following the decay of excited nuclear states. Applications of devices based on magnetic particle transport or analyzer systems are expanding in the field of analysis of molecular and solid surface structures. Here we should limit ourselves to mentioning only some examples of characteristic types and uses of these devices.

The various magnetic spectrometers can be grouped into four classes (Mladjenovic, 1979):

momentum-dispersive spectrometers for high-performance use or for coincidence measurements;

magnetic transport spectrometers, where the requirement of the good momentum resolution is relaxed and the energy spectra are measured with semiconductor [Si(Li)] detectors;

magnetic guides, for simply transporting particles from the source to the detector, having an input solid angle in some cases nearly 2π (4π) sr and providing a guide for almost the whole spectrum, except usually the low-energy part, in order to avoid the overloading of the detector from the high background produced by the low-energy secondaries (e.g., delta electrons); and

nondispersive focusing magnets for focusing a wider momentum band into a narrow cone at the detector far from the target and stray radiation.

The latest developments in the field of magnetic spectrometers are strongly influenced by the increased ability of performing precise calculation of more sophisticated field distributions, the improved technology in the production of magnetic materials, and the availability of superconducting solenoids (making possible the production of very strong fields while keeping the size of the device relatively small), as well as by the extended application of the modern position-sensitive detectors, making possible the use of the spectrometers as spectrographs, with simultaneous detection of a considerable part of the whole spectrum, and the development of the performance of the semiconductor detectors.

2.1 Momentum-Dispersive Spectrometers

In order to get an impression of the magnitude of the different aberration errors of geometrical nature, it is useful to consider first the *semicircular* focusing principle. Figure 1 shows the paths of monochromatic electrons, emitted from a point source, in a plane perpendicular to the applied magnetic field. Here φ is the pencil angle (defined by a slit) of the electrons emitted, and ρ is the radius of curvature in all cases. The focusing is not perfect; the spherical aberration causes a broadening of the image to width Δx. With a source having a width of x, the full image width is given by

$$\Delta x = s + 2\rho(1 - \cos\varphi), \tag{8}$$

while, assuming small φ and with $\Delta x = \Delta 2\rho$, the base resolution R^0 is

$$R^0 = \frac{\Delta\rho}{\rho} = \frac{s}{2\rho} + \frac{\varphi^2}{2}, \tag{9}$$

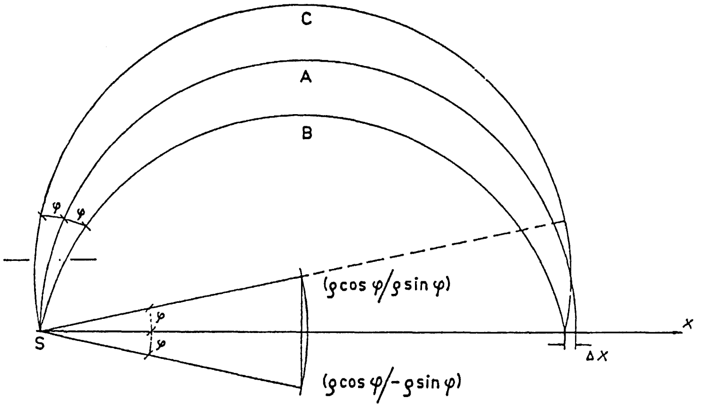

FIG. 1. The semicircular focusing principle (Siegbahn, 1965, by courtesy of Elsevier Science B. V.).

the first term depending upon the width of the source and the radius of curvature and the second upon the angular aperture. For optimum conditions, the two terms should give about equal contributions to the resolving power.

Among the magnetic spectrometers with high momentum dispersion, the best known and most widespread are the double-focusing $\pi\sqrt{2}$ types, combining features of one-dimensional semicircular focusing and two-dimensional helical or lens focusing. These spectrometers have a field of rotational symmetry, decreasing as $1/\sqrt{\rho}$ in the vicinity of the central path ($\rho = \rho_0$). In the symmetry plane of the field, electrons starting from a point on this central circle having angles φ_0 and ψ_0 in the radial and axial directions, respectively, will return to the central circle after focusing angles Φ_ρ and Φ_z. The focusing angles are related as

$$\frac{1}{\Phi_\rho^2} + \frac{1}{\Phi_z^2} = \frac{1}{\pi^2}\,. \tag{10}$$

An anastigmatic focusing is taking place when $\Phi_\rho = \Phi_z = \pi\sqrt{2} = 254°56$. Figure 2 shows the principle of the $\pi\sqrt{2}$ focusing.

Expanding $B_z(\rho,0)$ in a series gives

$$B_z(\rho,0) = B_0(1 + \alpha\eta + \beta\eta^2 + \ldots\ldots), \tag{11}$$

where $\eta = (\rho - \rho_0)/\rho_0$. The first expansion coefficient α is related to the focusing angles by

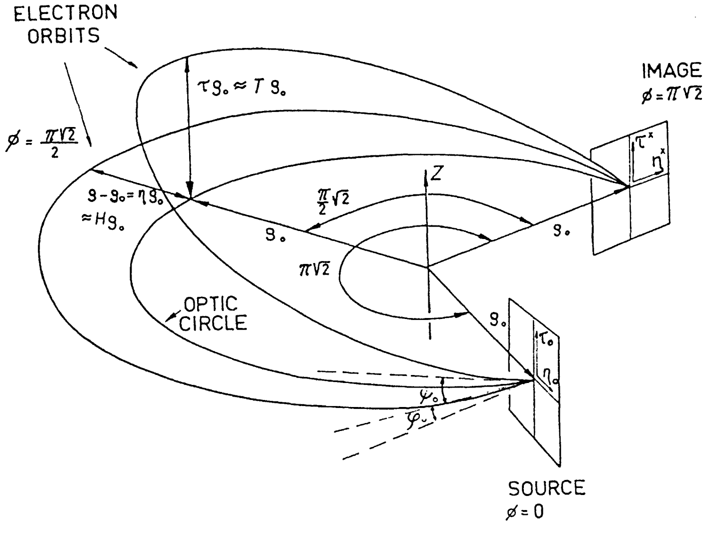

FIG. 2. The principle of $\pi\sqrt{2}$ focusing (Siegbahn, 1965, by courtesy of Elsevier Science B. V.).

$$\Phi_\rho = \pi/(1 - \alpha)^{1/2},\ \Phi_z = \pi/(-\alpha)^{1/2}, \qquad (12)$$

while β is connected with the aberrations influencing the width and height of the image. The focusing is the same for noncentral rays as well.

For the double-focusing case, the relative momentum dispersion of the spectrometer is $D_p = 4$, twice the dispersion compared to that for the homogeneous field. For a source with dimensions of width s and height h, and with a detector width w, the resolution R_0 is (at the base)

$$R_0 = \frac{s}{4\rho} + \frac{w}{4\rho} + \frac{4\beta - 3}{48\rho^2} h^2 + f(\beta, \varphi_0^2, \psi_0^2). \qquad (13)$$

It can be seen that the source height contributes to the resolution with a term of second order, which allows a typical height of 10 mm even for very high resolution. On account of the small applicable axial apertures, $\pi\sqrt{2}$ spectrometers usually work with solid angles smaller than 1%. The luminosity can be considerably improved by applying the method of Bergquist (Siegbahn, 1965), adding two electrostatic correctors, compensating for the source width and for the spherical aberration, respectively. The extended electron source is positioned with its plane at an angle of ~43° relative to the radius of the central circle in the $\pi\sqrt{2}$ magnetic field and by the help of an electric field gradient along the source, using vertical equipotential wires. The difference in ρ coordinates of the source points can be compensated for, enabling use of very extended sources. Compensating for the spherical aberration, a conveniently shaped grid system is placed symmetrically between the spectrometer and the detector, ensuring the achievement of a very high degree of cancellation of the aberration.

The spectrometer magnetic field can be produced either by a coil system only or by using iron for field shaping. The highest attainable precision was obtained by iron-free instruments first designed in Uppsala (Siegbahn, 1965). Present double-focusing systems with position-sensitive semiconductor detectors for multienergy detection can provide a momentum resolution of $\Delta p/p = 0.005$–0.3% in the electron energy range of 300–1000 keV. In coincidence spectroscopy and in the case of in-beam experiments, where a good access is requested to the source and detector, *sector field* devices are often used, in order to place the source and detector outside the field. By using curved pole boundaries, it is possible to achieve simultaneous double and second-order focusing (aberration coefficients depending on φ_0^2 and ψ_0^2 vanish simultaneously) with magnetic sector spectrometers, while the attainable resolution and transmission are comparable to those in the case of the double-focusing $\pi\sqrt{2}$-type spectrometer. An important law for sector-type spectrometers is "Barber's rule," according to which the center of the circle corresponding to the central trajectory in the sector with a homogeneous magnetic field is collinear with the source and focus points, independent of the sector angle. This rule can be generalized for the case of sectors with inhomogeneous field.

2.2 Magnetic Transport Spectrometers

Although some types of magnetic spectrometers (e.g., lens or toroidal types) can provide rather good resolution, they will be mentioned here because of their recent use mostly as particle transporters. These systems have large entrance solid angle and at the same time ensure the elimination of the high background from γ photons or unwanted particles (e.g., positrons), providing optimum conditions for coincidence measurements.

If magnetic sector analyzers are arranged in such a way that the pole pairs (with gaps between them) are situated along a circle, the center of which is collinear with the common source and focal points, a *toroidal* or *"orange"-type* spectrometer is obtained. At a resolution characteristic of sector field types and limited by the fringing fields, much higher transmission ranging between 10% and 20% can be obtained. Toroidal spectrometers with coils containing iron cores usually have 6–8 gaps, while the number of poles in the case of iron-free systems may be one order of magnitude higher. Most of the orange spectrometers are symmetrical, having the same pole profiles at the source and the detector side. Applying different air gaps, this type of spectrometer can be used for angular distribution measurements as well. The background is greatly reduced compared to the lens-type systems. Small, simple, and inexpensive toroidal transport spectrometers can be constructed by using permanent magnets and Si(Li) detectors ("mini oranges"). The characteristic sizes of such a system (source-detector distance, diameter of the magnet configuration) are in the order of 10 cm, and the transmission has a maximum that depends on the field strength and geometry. Figure 3 shows the scheme of a mini orange spectrometer.

The *magnetic lens-type spectrometers* are based on the focusing action of axial fields in long or short solenoids. The simplest case is a long solenoid with homogeneous field parallel to an axis containing the source and the detector. Its focusing properties can be treated completely analytically, which is not possible in the case of more complicated fields. The particle trajectories have a cylindrical symmetry, and the focusing is two dimensional. Particles from the source enter the spectrometer at an angle usually between 10° and 60°, defined by the entrance baffle (no axial rays can reach the detector separated from the source by a long lead absorber). For the case of a long lens with a homogeneous axial field B, the focusing properties are characterized as in the following description (Mladjenovic, 1976).

Dispersion $D = p\,dr/dp$, when measuring the radial displacement of the electron trajectory in the ring focus and when the momentum p is changed by dp, is given by

$$D = r_F/\sin^2\gamma^2, \tag{14}$$

where r_F is the radius of the ring focus and γ is the mean angle of the entering beam, related to the spectrometer axis and defined by the entrance aperture. The resolution (base width) for a point source is given by

$$R = (\Delta\gamma)^2(\beta + \cos^2\gamma\tan^2\beta), \tag{15}$$

where $\Delta\gamma$ is the semiaperture angle and β is defined by

$$\tan\beta/(\pi - \beta) = \tan^2\gamma, \tag{16}$$

while the acceptance solid angle Ω, defined by the entrance solid angle, is given by

$$\Omega = 4\pi\Delta\gamma\sin\gamma. \tag{17}$$

In the case of *short lenses,* the source and the detector are in field-free regions and the magnetic field distribution along the axis has a strong maximum. The solid angle of the lens spectrometers is usually 1–2%.

In order to increase the transmission and reduce the spherical aberrations, Siegbahn and Slätis (Siegbahn, 1965) designed a lens type with magnetic field weaker in the middle, forming an intermediate ring focus after which focusing the particles again at the end to a point focus on the axis. Such an *intermediate-image focusing lens* can be used for electron–electron coincidence measurements. The transmission of this type of spectrometer can reach 8%, and it is most often used for in-beam applications.

When a high-resolution semiconductor spectrometer is used as a particle detector, only the high-efficiency collection of the particles, as well as their transport to the detector, is of importance. To achieve this, *axial magnetic guides* or *solenoid transports* are

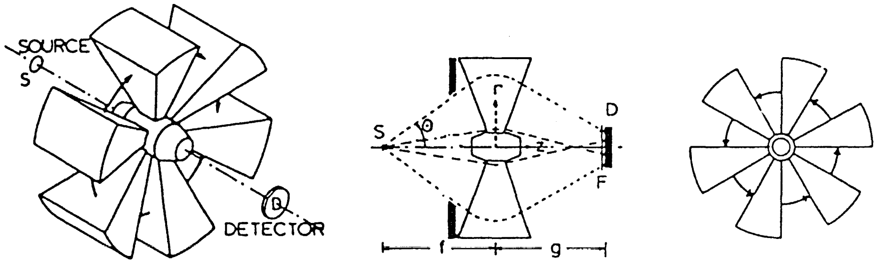

Stereoscopic plane Y plane Z

FIG. 3. The scheme of a permanent magnet "mini-orange" spectrometer (Van Klinken *et al.*, 1975, by courtesy of Elsevier Science B. V.).

used. The homogeneous field of a solenoid represents the guiding channel between the source and the detector on the axis, and when a sufficiently strong field is applied, the particles emitted from the source in 2π solid angle will spiral toward the detector. In order to limit the transport to the particles that can reach the detector surface of a definite size, field gradients are introduced at both the source and detector sides, inducing a magnetic mirror effect for reflecting the particles emitted at angles larger than a maximum angle relative to the axis, while circular baffles can be placed on the axis between the source and the detector to cut off the low-energy part of the spectra, thus saving the detector from the high load. A modulation of the field can improve the efficiency reduced by the baffle or, e.g., reduce the background due to particles with opposite charge (e.g., positrons in the case of electron focusing). If absorbers had been placed on the axis for preventing photons from being detected, the efficiency would have been considerably decreased. The cross section of a superconducting solenoid transport system, used for in-beam experiments (internal conversion spectroscopy), can be seen in Fig. 4 (Árvay *et al.*, 1980). A different concept is realized in the *trochoidal spectrometers and guides,* where the particles, in a cylindrically symmetric field (e.g., having r^{-n} dependence) between two pole pieces, are moving in circles, the centers of which drift along a much

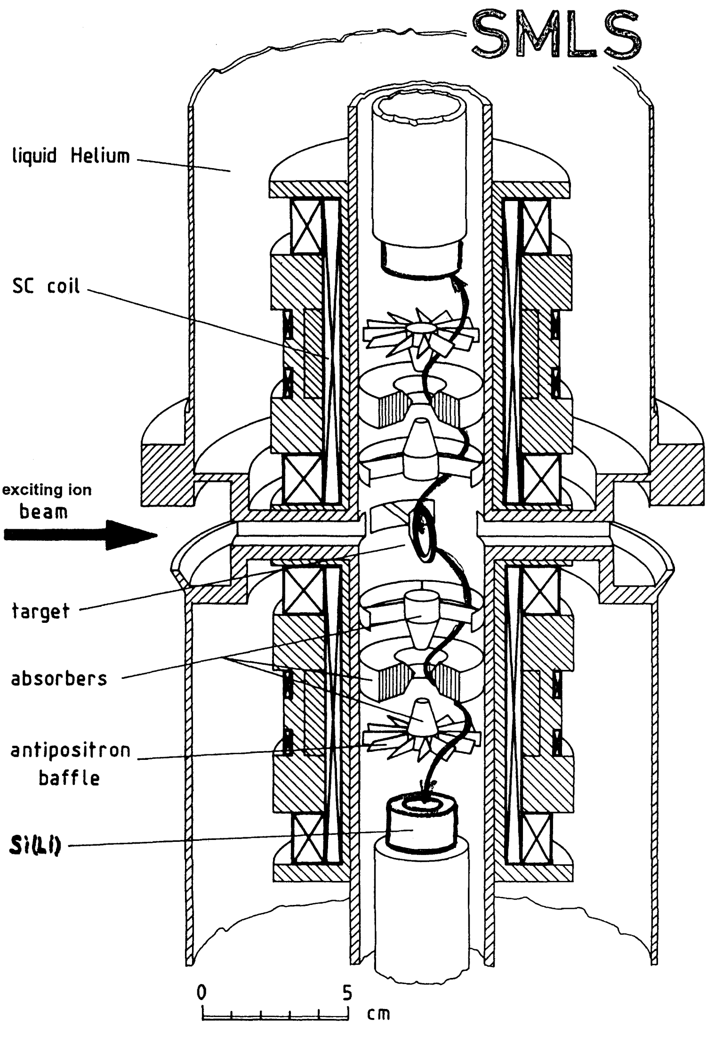

FIG. 4. A superconducting solenoid transport system (modified for lens working mode) for in-beam experiments (Árvay *et al.*, 1980, by courtesy of T. Fényes, 1994).

larger radius circle due to the inhomogeneity of the field. The transmission of the trochoidal systems is independent of the particle energy.

The *nondispersive focusing magnets* collect particles from a source into a converging beam after momentum selection. Figure 5 illustrates the principle of such a system. Here a baffle can be positioned at the intermediate plane of dispersion, while the detector is placed at the converging part of the beam. These instruments ensure a very low background and a good momentum bandwidth. They have, however, a low transmission (usually well below 1%).

2.3 New Applications

Without any attempt for completeness, we mention here some recent applications of magnetic devices, providing an insight into future directions. High-performance time-of-flight spectrometers can be developed using magnetic guides combining high luminosity with high energy resolution. An example is the trochoidal electron spectrometer (Norell *et al.*, 1984), where the electrons are moving in a magnetic field of rotational symmetry with $1/r^2$ dependence. This field ensures very long flight paths within small physical dimensions. The trochoidal spectrometer was successfully applied to studying resonance excitation in He^-, reaching an energy resolution less than 0.2% at an acceptance solid angle of almost 20% of 4π sr (Norell *et al.*, 1984).

A different solution for combining high sensitivity and reasonably high resolution is the *magnetic bottle* device, which can be used as a spectrometer as well as for image magnification and projection. The principle can be understood from Fig. 6. Electrons emitted from a source in a very strong homogeneous magnetic field are forced to follow a spiral around a field line, and they follow the same field line into a region of low and uniform field. Since the total velocity is constant, the axial velocity component increases when go-

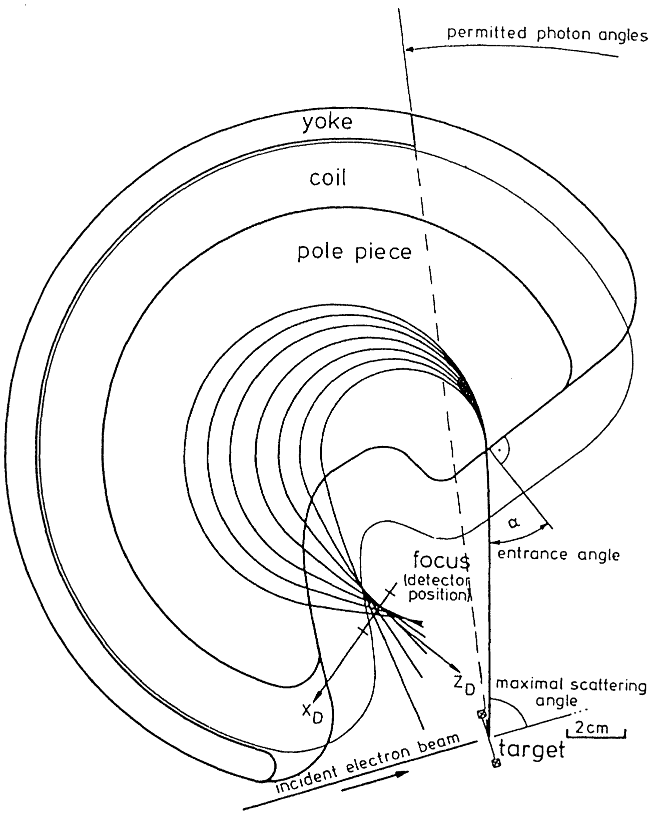

FIG. 5. The principle of operation of a nondispersive focusing magnet (Komma and Ruoff, 1981, by courtesy of Elsevier Science B. V.).

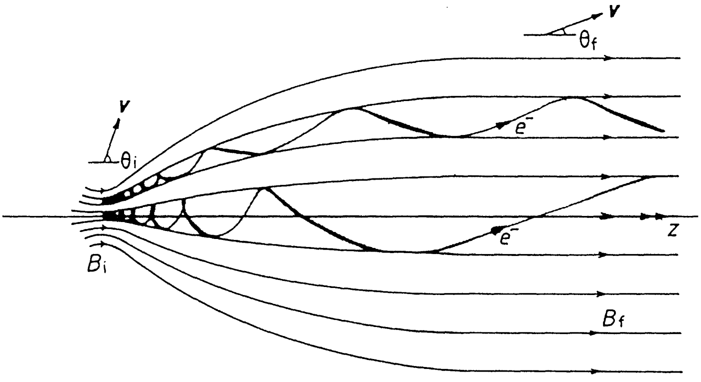

FIG. 6. Magnetic field lines and electron trajectories in a "magnetic bottle." The magnetic field changes gradually from a strong field B_i to a weaker field B_f. (Kruit and Read, 1983, by courtesy of IOP Publishing Ltd.)

ing from the high-field to low-field regions, while the transverse component decreases, leading to the parallelization and unspiralling of the electrons. (Beamson *et al.*, 1980; Kruit and Read, 1983). If the ratio of the field strengths in the high- and low-field region is high enough, electrons follow trajectories parallel to the field lines in the low-field region, even if emitted nearly perpendicularly to the field lines, and can be subjected to the different methods of energy analysis. When the field experienced by an electron changes negligibly in the course of one revolution of the helical motion, the angular momentum is constant, and the angle θ_f of the helical motion in the low-field (B_f) region can be obtained from (Kruit and Read, 1983)

$$\frac{\sin\theta_f}{\sin\theta_i} = \left(\frac{B_f}{B_i}\right)^{1/2}, \tag{18}$$

where the index i refers to the initial parameters at emission. Similarly, the lateral linear magnification M of the orbital motion is given by

$$M = (B_i/B_f)^{1/2}. \tag{19}$$

Assuming that electrons originate from the source with an angular spread 2π sr and energy E, the total possible relative energy spread is

$$\Delta E/E = B_f/B_i, \tag{20}$$

representing an intrinsic limit of the resolution, which is strongly dependent on the initial angular distribution of the emitted electrons. In most cases the magnetic bottle is used as a time-of-flight spectrometer with pulsed sources. Applications are in the low–electron-energy (around 10 eV) region and concern spectroscopy of photoelectrons induced by laser multiphotoionization or UV photoionization processes of atoms, molecules, or cluster ions and also energy-loss spectroscopy of inner-shell excited states. As the diverging magnetic field lines act as guiding tracks for electrons emitted from a solid surface, a good image projection is possible at a high magnification (Beamson *et al.*, 1980), ensuring at the same time high collection efficiency. Combining this magnifier with proper energy selectors [i.e., time-of-flight (TOF) spectrometers, band-pass image filters, or energy analyzers], high-performance photoelectron spectromicroscopes can be developed. Using superconducting solenoids and tunable synchrotron photon sources, submicron lateral resolution can be reached by the help of a special angular selection scheme (Keenlyside and Pianetta, 1993).

Another field of growing interest where magnetic spectrometers can be utilized is the ion-beam analysis of surfaces and interfaces. Here the energy and angular as well as the mass distribution of the ions scattered or recoiling from a solid surface are measured. The design of a magnetic spectrograph connected to a particle accelerator and developed for the analysis of surfaces and thin layers with Rutherford backscattering and recoil spectroscopy (Boerma *et al.*, 1990) is shown in Fig. 7. The main elements of the system are the ultrahigh-vacuum scattering chamber, with a precision three-axis manipulator where the sample is mounted, and the

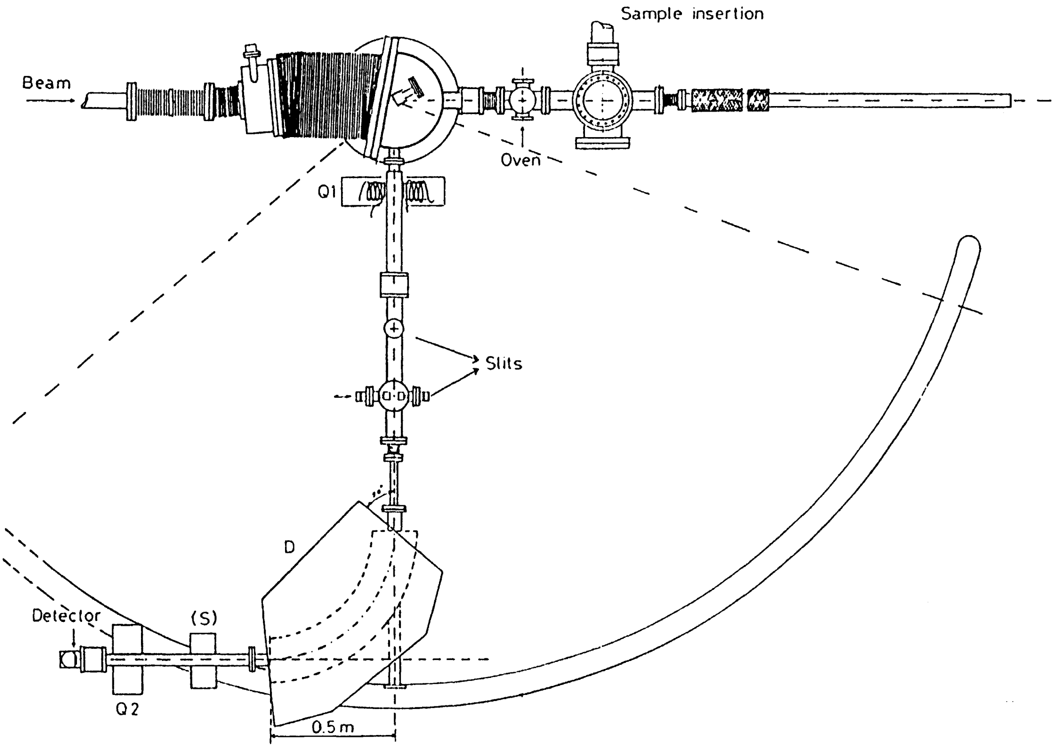

FIG. 7. The design of a magnetic spectrograph for surface and thin-layer analysis using a particle beam from an accelerator in ion-scattering experiments (Boerma *et al.*, 1990, by courtesy of Elsevier Science B. V.).

spectrograph, consisting of two quadrupole magnets and a dipole sector magnet bending the ions in the horizontal plane, equipped with a two-dimensional position-sensitive detector. Together with the chamber, the spectrograph is rotatable around a vertical axis, ensuring a scattering angle variable between 20° and 140°. In the horizontal plane, ions can be measured simultaneously within an energy window of 2% relative width at an energy resolution of 2×10^{-4}, while in the vertical direction their angular distribution can be detected with an angular window of 5° and an angular resolution of 0.1°. The depth resolution is close to one monolayer near the surface, while the mass discrimination of the spectrograph makes possible the background-free detection of light atoms recoiling from a heavier substrate. Providing optimum conditions for surface analysis using the channeling-blocking technique, the magnetic spectrograph provides an important extension of ion-beam analysis through its unique features in applications for recoil spectroscopy and in lattice-site location of light atoms in thin layers.

3. ELECTROSTATIC SPECTROMETERS

In the low–particle-energy region (from a few eV to several keV), the application of the electrostatic spectrometers is the most common. There are many comprehensive reviews available of these systems (Afanas'ev and Yavor, 1976; Granneman and Van der Wiel, 1983). Critical reports on design considerations and recent developments appeared recently (Leckey, 1987; Roy and Tremblay, 1990; Ibach, 1991). Electrostatic spectrometers have been used extensively for a few decades in the fields of analysis of atomic, molecular, and cluster structures, collision processes, and surface and interface analysis through a variety of photoelectron and Auger-electron spectroscopic, electron and ion energy-loss spectroscopic, and electron-diffraction and ion-scattering spectroscopic methods. Further expansion in applications of electrostatic spectrometers is seen in experiments connected with phenomena of molecular and surface dynamics (including processes induced by multiply-charged–particle beams, laser beams, and synchrotron radiation) or with phenomena and structures of microscopic (even molecular) localization. The main advantages of electrostatic spectrometers are their suitability for energy analysis and the simpler way of attaining high stability and precision in forming the field. An electron or ion spectrometer consists of a vacuum (in the case of surface analysis, UHV) system with a sample or scattering chamber, an irradiating unit for exciting the target, and an analyzer including the detector. The analyzer and sample chambers are shielded against even weak external magnetic fields. We restrict ourselves to discussing analyzers and detectors only, while detailed information on component materials, magnetic shielding, working modes, calibration, etc. can be learned from available reports (Granneman and Van der Wiel, 1983; Rivière, 1990). Most electrostatic spectrometers can be grouped into two classes: energy-dispersive and -nondispersive types. Many designs, however (e.g., of display type, TOF spectrometers, or band-pass filters), important for specific applications, fall outside these groups. The latest developments regarding electrostatic spectrometers are motivated by the need for multienergy and multiangle detection in order to increase the speed of information important from the point of view of state and structure analysis, and by the strong requirement of spectromicroscopy for state imaging and the possibility for doing time-resolved spectroscopy at high resolution.

3.1 Dispersive-Type Analyzers

Dispersive analyzers can be of either the mirror or the deflector type. In deflectors the particles follow trajectories close to the equipotentials, while in mirrors the trajectories cross the equipotentials, and the velocity component of the particles perpendicular to these lines decreases to zero and changes to the opposite direction after reflection. In electrostatic dispersive analyzers, the pass energy E is proportional to the potential difference (ΔV) of the electrodes producing the dispersive field:

$$E = qk\Delta V, \tag{21}$$

where q is the charge of the particles with energy E and k is the *spectrometer constant.*

3.1.1 Electrostatic Mirror Analyzers

Among the electrostatic mirror spectrometers, the simplest is the *parallel-plate analyzer,* the scheme of which is shown in Fig. 8. The field between the two parallel plane electrodes is uniform, and the equipotential surfaces are also parallel to the electrodes. In most cases the source (aperture) and the detector are positioned symmetrically relative to the electrodes, in the field-free region. The particles enter into the field at an angle α (Fig. 8) and are focused symmetrically in a plane (in the case of a source at the entrance slit and $\alpha = 45°$, this plane coincides with the plate containing the entrance and exit slits). This property is advantageous from the point of view of simultaneous detection of different energy particles (energy multidetection). At $\alpha = 30°$ beam angle, it is possible to achieve second-order focusing if $h_1 + h_2 = d/(3\sqrt{3})$ (see Fig. 8), where $h_1 + h_2$ is the sum of the distance of the source and detec-

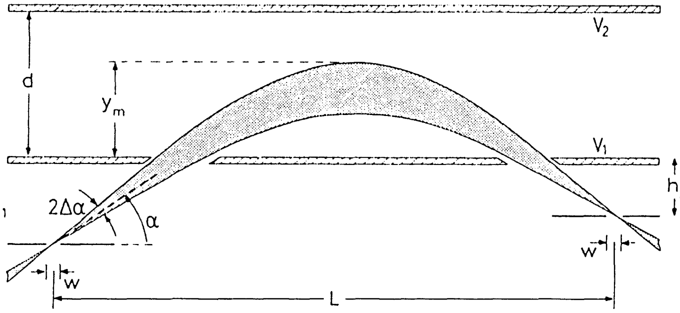

FIG. 8. The scheme of the parallel plate analyzer (Granneman and Van der Wiel, 1983, by courtesy of Elsevier Science B. V.).

tor apertures placed at the focal length d from each other. The energy dispersion in this case is

$$D = \tfrac{2}{3}d. \tag{22}$$

The parallel-plate mirror analyzer has a very simple construction and is often applied in coincidence or in angular distribution measurements. Because of the unidirectional focusing, however, its luminosity is low at high resolution.

One of the most often used electrostatic analyzers is the *cylindrical-mirror analyzer* (CMA). Its scheme is shown in Fig. 9. It consists of two coaxial cylinders with inner radius R_1 and outer radius R_2. The radial logarithmic field between the cylinders reflects the charged particles, which start from the common axis in the field-free region and enter the field through the entrance slit on the inner cylinder, and return to the axis after traveling one focal length (which is dependent on the energy of the particles). Although the focusing in the CMA is unidirectional on account of the cylindrical symmetry, it is equivalent to a two-directional focus (i.e., a point source is imaged to a point). In many applications the point source and the detector are situated on axis (point-to-point focusing); however, either the source or the detector aperture (or both) can be selected to be ring-shaped, providing a considerable flexibility in the design. The focal length L is dependent on the beam entrance angle α, the sum of the radial projection of the central trajectory in the field-free region, $n = h_1 + h_2$ (see Fig. 9), and the spectrometer constant $k = E/q(V_2 - V_1) \times \ln(R_2/R_1)$ (Zashkvara *et al.*, 1966):

$$\frac{L}{R_1} = \left(n + 4pe^{p^2}\int_0^p e^{-x^2}dx\right)\cot\alpha, \tag{23}$$

where $p = \sqrt{k'}\,\sin\alpha$.

When both the source and image are on the axis, first-order focusing can be obtained at practically any beam entrance angle. Second-order focusing can also be achieved at any value of n with properly chosen values for the other parameters. For example, in the case of (second-order) axial point-to-point focusing, $\alpha = 42.3°$ and $k = 1.31$, while $D = 5.6R_1$. The focal surface is a cone centered on the axis with an apex pointing toward the position of the detector. Calculated values of the respective parameters for first- or second-order focusing are available (Risley, 1972) with various combinations of α, n, and k. The point transmission of the CMA is high, and using second-order focusing, the accepted angular range ($2\Delta\alpha$) around the beam angle can be considerably increased at a given resolution. At larger beam angles α, the dispersion is larger, and at a given luminosity, a better resolution can be achieved even for first-order focusing—a practical limit is $\alpha = 60°$. Finite radial dimension of

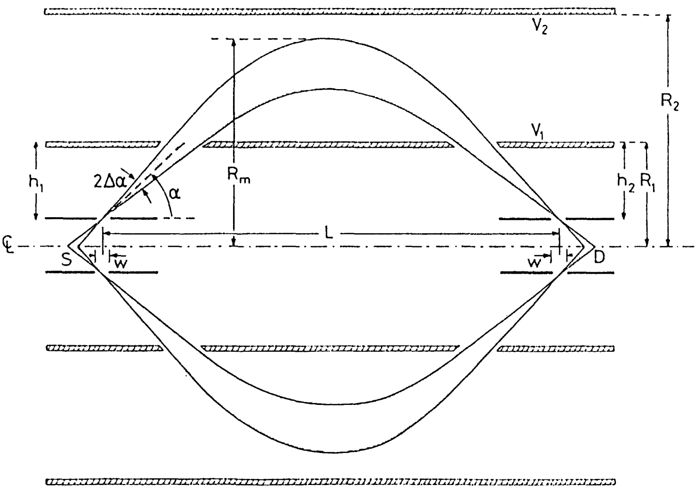

FIG. 9. Cross section of the cylindrical-mirror analyzer (Granneman and Van der Wiel, 1983, by courtesy of Elsevier Science B. V.).

the sample leads to image broadening, and the effective source area is much smaller than the actual one in the case of a radially extended source (Afanas'ev and Yavor, 1976). The "classical" (point-to-point focusing) CMA has the disadvantage of its focusing and spectrometer function (line shape and width) being very sensitive to the axial and radial position of the source (Rivière, 1990). Unfortunately, this property, including the small effective source area, largely undermines the advantage of the high transmission, resulting in low luminosity for extended samples. Therefore, the application of the CMA is most suitable for small-area excitation, e.g., for scanning Auger microscopy, where only point transmission is important. More advantageous are, from the point of view of analysis of particles emitted from extended sources, the double-pass CMA, i.e., CMA applied in series (Palmberg, 1974), with the exit aperture of the first CMA representing the entrance aperture to the second. This arrangement ensures a much better source definition, higher energy dispersion, and a considerable reduction of the spectral background by preventing the greater part of the particles scattered on constructional elements of the analyzer (as well as secondaries) from reaching the detector. Besides this, pre-retardation (e.g., by using spherical grids in front of the entrance slits) makes possible improvement of the energy resolution as well as maintaining a high luminosity by opening the slits of the second stage. For extended sources, the most favorable solution is to position the entrance (source) aperture of the first stage at the entrance slit on the inner cylinder (Varga *et al.*, 1978). A further possible application of the double-pass CMA is for angle-resolved experiments, in combination with a position-sensitive detector for angular multidetection. Figure 10 shows the arrangement of such a double-pass CMA, with the option of pre-retardation (Bosch *et al.*, 1984). By tilting the sample, angular information can be obtained both in azimuthal and in polar directions. One of the disadvantages of the cylindrical mirrors, difficult to improve, is connected with the endings of the cylinders. It is necessary to ensure the correct field distribution at the ends of the analyzer and at the same time to provide a good accessibility of the sample for irradiation. A possible solution, different from the application of equipotential electrodes or conical-shaped endings with a resistive layer (Palmberg, 1974), is the utilization of a distorted field, near the simply closed ends. The scheme of such a second-order focusing *distorted-field* CMA is shown in Fig. 11 (Varga *et al.*, 1985). Here the source and detector are situated well outside the spectrometer, thus providing an increased dispersion and, at the same time, a considerably improved accessibility of the sample and detector (an important advantage when irradiating the sample or in the case of multidetection). This type of analyzer can be easily arranged along a common axis as a double- (or triple-) pass system, with the further advantage of positioning the source between two modules for coincidence experiments or applying different potentials on the inner cylinders of the independent stages facilitating the retardation between two modules. Similar results can be obtained with conical-shaped endings, which further improves the accessibility of the sample, while keeping the second-or-

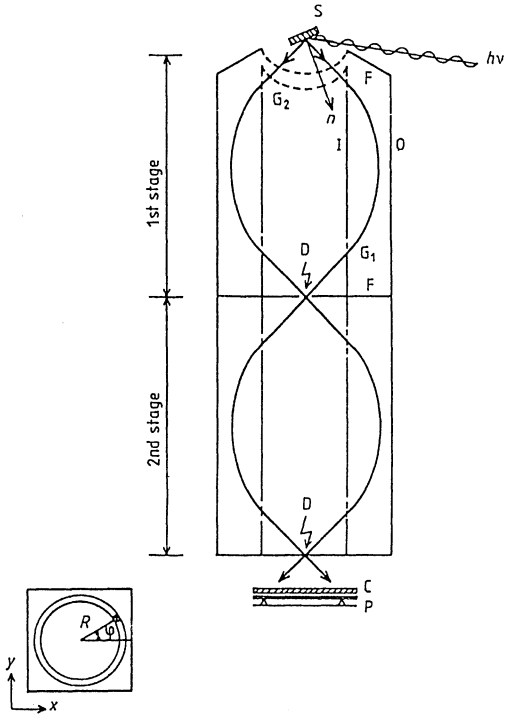

FIG. 10. Double-pass CMA with pre-retardation and angular multidetection (Bosch *et al.*, 1984, by courtesy of IOP Publishing Ltd.)

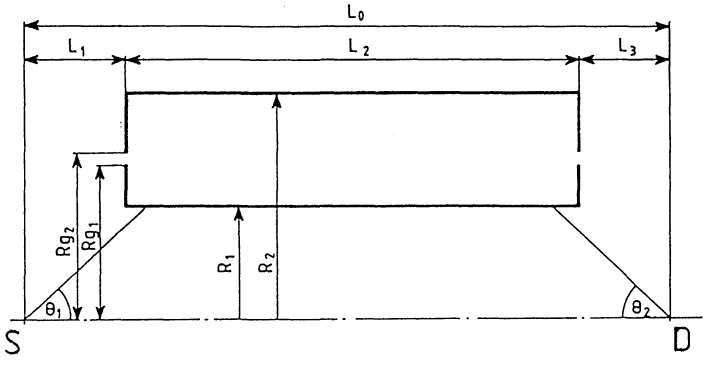

FIG. 11. The scheme of the distorted-field CMA (Varga *et al.*, 1985, by courtesy of Elsevier Science B. V.). S, source; D, detector; θ_1, input beam angle; θ_2, output beam angle.

der focusing properties for the cases of point source, ring focus and ring source, ring focus as well (Tőkési *et al.*, 1994). A considerable improvement of the performance of CMAs can be achieved by applying a similar, axially inhomogeneous field that can be described by an analytical expression, adding a hyperbolic term to the formula giving the potential distribution Φ for the "classical" CMA (Siegbahn *et al.*, 1996):

$$\Phi = \Phi_0[\ln(r/l) - \tfrac{1}{2}(r/l)^2 + (z/l)^2 + \text{const}], \tag{24}$$

where r and z are cylindrical coordinates, l is the linear dimension of the analyzer, and Φ_0 is a constant. The axial potential makes possible the transformation of a large-angle input cone of the beam at the entrance slit into a small-angle output cone at the exit slit of the analyzer. This large input angle ensures a large energy dispersion and a large acceptance solid angle at the same time. A further attractive property of this type of CMA is the applicability as a mass filter with a periodically varying analyzer voltage component, or as a mass reflector with good time dispersion.

Finally, among electrostatic mirrors the *spherical-mirror analyzer* (SMA) should be mentioned most of all because of the increasing interest in its application. First described by Sar-El (1966), the SMA provides an aberration-free image of the source, without linear energy dispersion in the plane of the source and image. In most cases the source and image are located symmetrically on the diameter and inside of the inner sphere (a spherical grid). The particles start from the source in the field-free region, cross this inner spherical surface, and are then reflected by the field between the concentric inner and outer spheres. Energy separation can be made (by using baffles or an exit slit) at an intermediate focus point between the two spherical electrodes (Roy and Tremblay, 1990). Figure 12 shows the application of a hemispherical mirror for energy and angular distribution measurements (Daimon, 1988). The particles, emitted from the sample surface with the same kinetic energy, converge at the small exit aperture *A*. Their angular distribution is detected by a position-sensitive detector. Theoretically, the angular pattern obtained is distortion free, and the solid angle of acceptance is 2π sr (4π sr). The main disadvantages of the SMA are the moderate energy resolution and the difficult accessibility of the sample for irradiation. Recent improvements in the performance

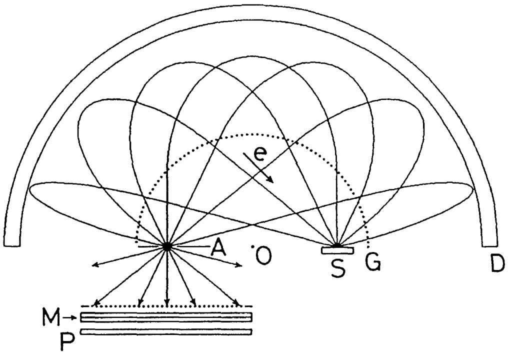

FIG. 12. The principle of angular distribution measurements using the spherical-mirror analyzer (Daimon, 1988, by courtesy of American Institute of Physics). S, sample; A, aperture; M, microchannel plates; *P*, phosphor screen; G, spherical grid; D, spherical deflector.

(application of retarding grids at the source and the detector, and introduction of corrector electrodes) have made possible the successful application of the display-type SMA for studying the band structure of single crystals (Nishimoto *et al.,* 1993).

The combination of different types of electrostatic mirrors can result in a considerable improvement of the overall performance of the spectrometer, compensating the individual disadvantages. An example is shown in Fig. 13 for the combination of a SMA and a double-pass CMA in order to perform simultaneous measurements of high-resolution electron spectra at different angles in ion–atom collision experiments (Kövér *et al.,* 1983; Varga *et al.,* 1992).

3.1.2 Deflectors Electrostatic fields in toroidal capacitors can be used for deflection and imaging of particles entering tangentially to the equipotential surfaces. Energy dispersion occurs at the image in the plane of deflection. The equipotentials of the toroidal field are curved in two perpendicular directions having different radii of curvature. Containing the main path of particles, the main equipotential has a curvature with radius R in the radial direction (in the plane orthogonal to the symmetry axis) and r in the perpendicular (or axial) direction (in planes containing the symmetry axis). When $R = r$, the *spherical-deflector analyzer* (SDA), and in the case of $R = \infty$, the *cylindrical-deflector analyzer* (CDA) are obtained, while for the general case we speak about the *toroidal-deflector analyzer* (TDA). All reviews on electrostatic analyzers mentioned before, as well as the references therein, discuss various aspects of these types in detail, due to their widespread fields of application. The three deflectors can be used as sectors, however, with the condition for zero object and image distances (from the sector) resulting in a sector angle of 127.3° for the CDA (without focusing in the axial direction) and in a sector angle of 180° for the SDA (with the coincidence of the images in both directions). The sector angle for point-to-point focusing, as a function of the ratio of the respective radii of the toroidal deflector, exhibits two asymptotic cases at 127° and 180° (Leckey, 1987).

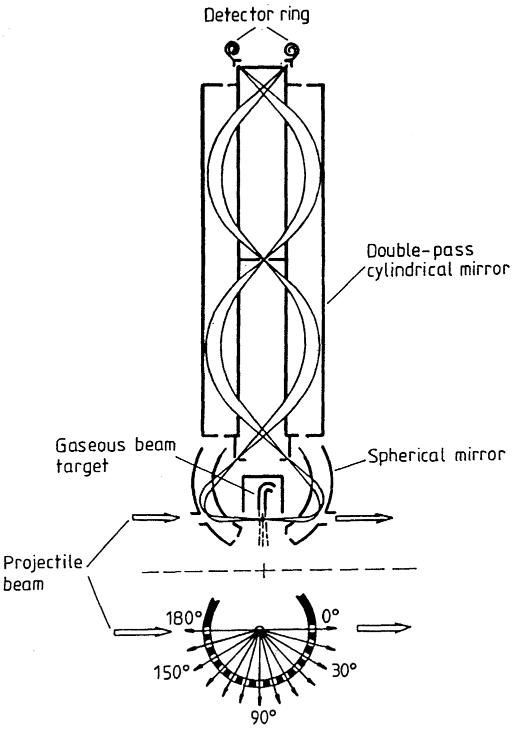

FIG. 13. Combination of spherical-mirror and double-pass CMA analyzers for simultaneous measurements of angular and energy distribution of particles (Kövér *et al.*, 1983; Varga *et al.*, 1992, by courtesy of IOP Publishing Ltd).

The radial dispersion at the focus can be expressed as

$$D = \frac{2r}{2 - r/R}, \tag{25}$$

which gives $D = 2r$ for the SDA and $D = r$ for the CDA. TDAs were recently used for obtaining good angular resolution while keeping the energy resolution and the luminosity high. They take advantage of the possibility of angular multidetection for a ring-shaped focus by using position-sensitive detectors (Roy and Tremblay, 1990; Leckey, 1987), as is shown in Fig. 14. The design, consisting of TDA sectors, is an example of the application of simultaneous energy and angular multidetection (Hellings *et al.,* 1989). The particles, scattered into different angles from the sample, preserve their angular directions and are dispersed by their energy in the radial direction in the plane of the position-sensitive detector.

One of the most popular electrostatic analyzers is the spherical deflector, with its two-directional focusing and its circular cen-

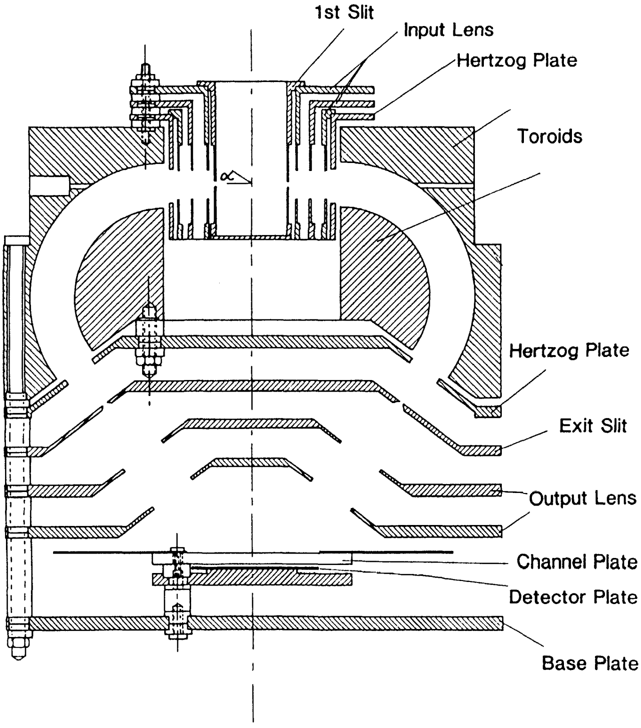

FIG. 14. Toroidal analyzer for simultaneous energy and angular multidetection (Leckey, 1987, by courtesy of Elsevier Science B. V.). The "Hertzog plates" are used for correcting the effects of the fringing fields by introducing electrode-free space between the electrode fringe and the entrance or exit slit (Nishigaki and Kanai, 1986).

tral trajectory. For practical reasons, in order to ensure the necessary space for sample irradiation and the detector, many types of spherical sectors have been developed so far, but to obtain minimal aberrations, the sector angle closest to 180° is preferred. The scheme of the 180° SDA is shown in Fig. 15, illustrating that the object and image points are located in the entrance and exit planes of the analyzer. This arrangement requires the minimization of the fringe fields (Granneman and Van der Wiel, 1983, and references therein). Spherical deflectors are often used for beam definition as particle monochromators, as well as for high-resolution energy analysis. Their space focusing and high luminosity compensates for the low entrance solid angle for point sources and the first-order focus. Because the focal surface is a plane, energy multidetection is facilitated, providing further possibilities for obtaining high-resolution lateral or angular information in the direction perpendicular to the energy dispersion in the image plane. The CDA focuses only in the radial direction and provides a line focus of a point source; therefore, its luminosity is much smaller than that of the SDA; however, it offers the advantage of the simple construction. In the case of a symmetric voltage supply ($V_1 = -V_2$), the radius of the main trajectory in a CDA sector is the geometric mean of the cylinder radii.

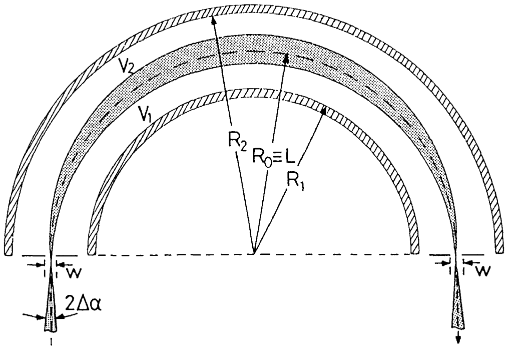

FIG. 15. The scheme of the 180° spherical-deflector analyzer (Granneman and Van der Wiel, 1983, by courtesy of Elsevier Science B. V.).

3.2 Nondispersive-Type Analyzers

Regarding nondispersive electrostatic analyzers, the *retarding-field* types have been used extensively for a long time, mainly in low-energy diffraction experiments (Granneman and van der Wiel, 1983). Usually these systems consist of several high-transmission, concentric spherical grids and a spherical collector. After traveling in a field-free region, retardation of particles from a sample in the center of the spheres takes place between the first two grids. Only particles having kinetic energy exceeding the retarding potential are transmitted through the grids. Therefore, varying the potential difference between the sample and collector leads to an integrated energy spectrum, the differentiation of which with respect to the retarding potential gives the spectrum itself. In order to decrease the influence of lens effects, the retarding grid is constructed of two equipotential grids, while another grid is applied in front of the collector (equipotential with the sample, the first grid and the collector) for avoiding disturbing effects from reflected particles and secondaries. Retarding-field analyzers (RFA) have very large luminosity, while being capable of high energy resolution. Due to the energy-integrating measuring mode as well as to the secondary electrons created on the grids, the level of noise is rather high and the signal-to-noise ratio is low. Improved versions of RFAs contain a cylindrical grid (and a pair of focusing spherical grids) deflecting the particles onto a small collector on the axis, converting the analyzer from a high-pass to a band-pass filter (Huchital and Rigden, 1972; Staib and Dinklage, 1977). In Fig. 16 can be seen the arrangement of Staib. Combining the advantages of retardation with energy-dispersive deflection and completed by a proper input lens, this analyzer is competitive with the dispersive-type devices (no angular information is available in this case). The direct display of the angular distribution can be achieved within a wide cone of acceptance by the help of the ellipsoidal mirror analyzer, shown in Fig. 17. Here the ellipsoidal mirror acts as a low-pass filter and reflects the pre-retarded particles into a variable aperture (preserving the original angular distribution). The energy analysis is then performed by a five-grid high-pass filter, and the image obtained by the position-sensitive detector correlates with the angular distribution of the particles emitted from the sample (Eastman *et al.*, 1980).

With pulsed sources providing a sharp time definition of the emission of (very low-energy) particles from the sample, the dispersion of their velocities can be analyzed by TOF spectrometers. Figure 18 shows the scheme of a paraboloidal electrostatic mirror adapted for this purpose (Trevor *et al.*, 1989), as a typical example. Electrons are created

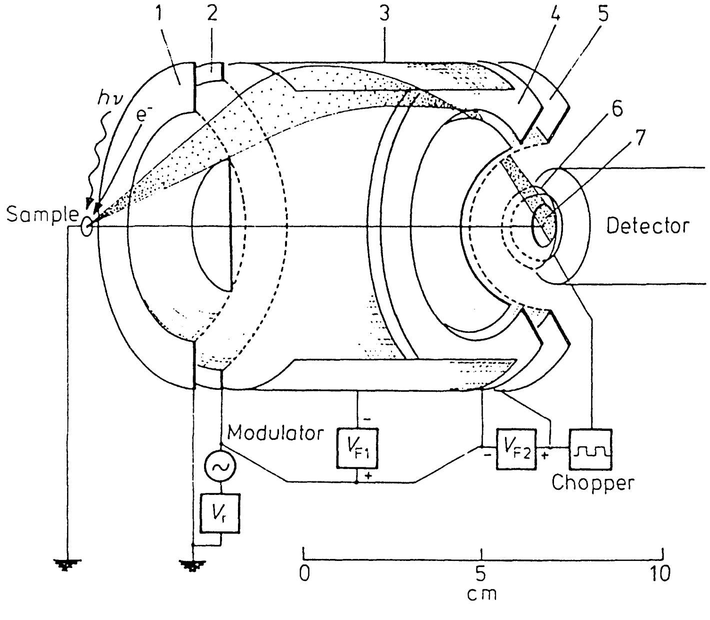

FIG. 16. Band-pass filter retarding-field analyzer (Staib and Dinklage, 1987, by courtesy of IOP Publishing Ltd). 1, entrance grid; 2, retarding grid; 3, cylindrical grid; 4, annular slit (first focus); 5, focusing grid; 6, chopper grid; 7, exit aperture (second focus).

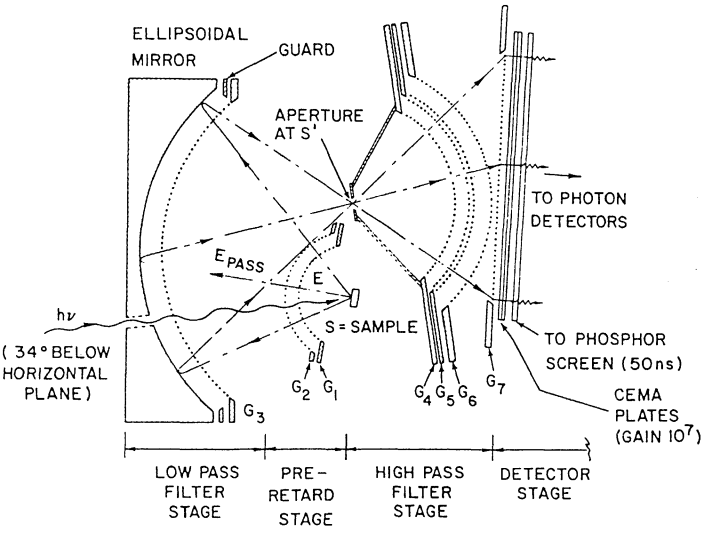
ELLIPSOIDAL MIRROR
GUARD
APERTURE AT S'
TO PHOTON DETECTORS
E_PASS
E
S = SAMPLE
hν
(34° BELOW HORIZONTAL PLANE)
G2
G1
G3
G4
G5
G6
G7
TO PHOSPHOR SCREEN (50ns)
CEMA PLATES (GAIN 10^7)
LOW PASS FILTER STAGE
PRE-RETARD STAGE
HIGH PASS FILTER STAGE
DETECTOR STAGE

FIG. 17. Ellipsoidal-mirror analyzer for wide-range angular multidetection (Eastman *et al.*, 1980, by courtesy of Elsevier Science B. V.).

by a laser pulse in the focal point of a parabolic mirror consisting of two closely spaced grids with a reflection potential on the outer grid. The parabolic shape of the mirror forces all electron trajectories to travel the same total distance to the detector along collimated paths. The flight time is dependent on the energy of the electrons at a given angle of emission, besides geometrical parameters. It is possible to achieve a good energy and angular resolution by using a position-sensitive detector.

Having a tunable, pulsed source for exciting electrons very close to the ionization threshold, the zero–kinetic energy (ZEKE) electrons can be separated from the near-ZEKE ones by the help of a delayed voltage pulse extracting the electrons created in the sample following ionization under field-free conditions. The basic principle of the

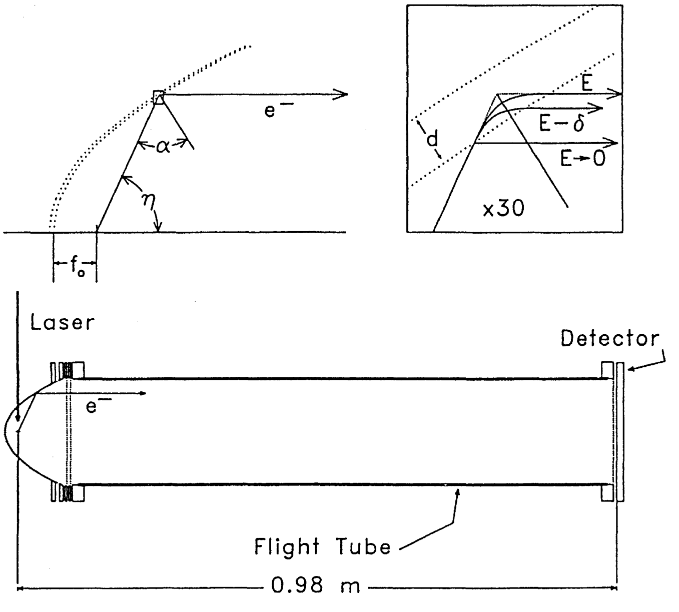
e−
α
η
f0
E
E−δ
E→0
d
x30
Laser
Detector
e−
Flight Tube
0.98 m

FIG. 18. The principle of operation of the paraboloidal-mirror time-of-flight analyzer (Trevor *et al.*, 1989, by courtesy of American Institute of Physics).

method and the scheme of a ZEKE spectrometer is shown in Fig. 19. After a given (~1 μs) delay following the ionization, the near-ZEKE electrons are separated from the electrons with zero velocity, and when the pulsed field is applied, the electrons are extracted into the drift tube and later detected by the TOF detector (Schlag *et al.*, 1993). Small differences are "time amplified" by being converted into a large difference in flight time. Moreover, while the transmission is 100% for the ZEKE electrons, a strong reduction is observed for the near-ZEKE ones.

3.3 Auxiliary Electron and Ion Optics for Transport, Retardation, and Imaging

Important elements of electrostatic spectrometers are the various lenses in front of or after the analyzer, representing an electron- or ion-optical interface between the sample and the analyzer or between the analyzer and the detector. These elements often determine the essential feature of the application (e.g., angular measurements or spectromicroscopy) as well as improving the performance of the spectrometer considerably (i.e., improving resolution and luminosity by retardation or providing optimal conditions for various kinds of sample excitation). Energy or angular multidetection also motivates the design of auxiliary lenses. The principles, types, design, and application of various lens systems are treated in the comprehensive review of Granneman and Van der Wiel (1983). Instead of dealing with details of these devices, here some examples are given for the different functions they can serve.

The arrangement shown in Fig. 20 illustrates how a CMA can scan over polar angles (20°–70°) by the help of a six-element cylindrical angle-selection lens, while simultaneously detecting electrons leaving the sample at different azimuthal angles using a position-sensitive detector (Van Hoof and Van der Wiel, 1980; Van Hoof, 1981).

A good example of the key role for additional electron optics in extending the capability of an electron spectrometer is the case of the lens system of a hemispherical deflection analyzer equipped with a two-dimensional position-sensitive detector. Separating the sample area from the analyzer and ensuring good access for handling and irradiation (as well as the possibility of differential pumping of the sample chamber), the long input lens has to provide the option of retardation of the electrons in order to improve the resolution and the luminosity, while keeping the focus always at the entrance slit of the analyzer (zoom effect) and matching the electron-emitting area on the sample to the area accepted by the analyzer (optimum efficiency). In the direction perpendicular to the direction of the energy dispersion, a wide (limited only by the length of the detector, due to the unit magnification) angular aperture at the entrance of the analyzer leads to a considerable gain in intensity at a given resolution. At the same time, however, in

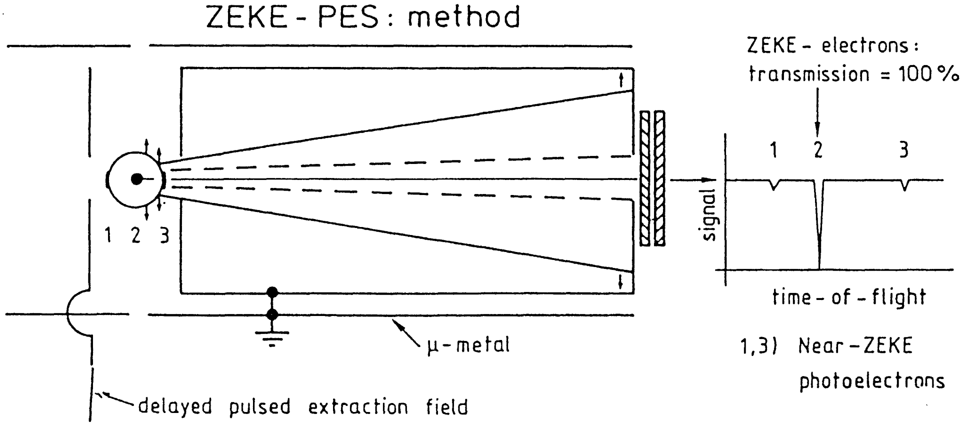
ZEKE - PES : method
ZEKE - electrons : transmission = 100 %
1 2 3
signal
time - of - flight
1,3) Near - ZEKE photoelectrons
1 2 3
μ - metal
delayed pulsed extraction field

FIG. 19. The basic principle of the zero–kinetic-energy (ZEKE) method and spectrometer (Schlag *et al.*, 1993, by courtesy of Elsevier Science B. V.).

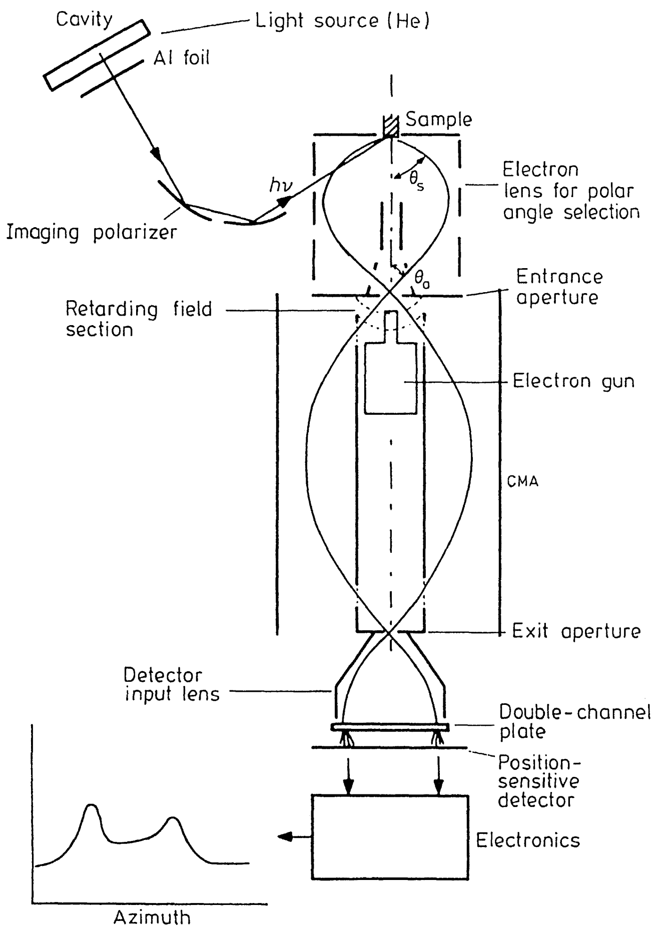
Cavity
Light source (He)
Al foil
Sample
Electron lens for polar angle selection
$h\nu$
θ_s
Imaging polarizer
θ_a
Entrance aperture
Retarding field section
Electron gun
CMA
Exit aperture
Detector input lens
Double-channel plate
Position-sensitive detector
Electronics
Azimuth

FIG. 20. Polar-angle selection lens and azimuthal-angle multidetection for CMA (Van Hoof and Van der Wiel, 1980, by courtesy of IOP Publishing Ltd).

this direction, as a result of the double focusing in the analyzer, the image of the source (entrance slit) can be obtained. By improving the imaging performance of the lens at high (20–25) magnification, very high (several microns) spatial resolution can be achieved (Gelius *et al.,* 1990). The five-element retarding and imaging zoom lens is provided with an exchangeable auxiliary einzel lens close to the lens entrance slit (Gelius *et al.,* 1984; Siegbahn, 1989). In order to scan over the sample surface and direct the beam accurately to a given small entrance slit at the analyzer, an octopole electric deflector is mounted inside the long central section of the lens system, surrounded by two concentric layers of magnetic shielding. While the einzel lens increases the magnification, the five-element part acts as two decoupled three-element lenses.

A completely different idea is followed in the electron optical arrangement of the ESCASCOPE instrument. The diagram of Fig. 21 shows the principle of operation of the lens system (Coxon *et al.,* 1990). There are two additional lens elements (lens 3 and lens 5) in front of the entrance and after the exit of the analyzer compared with a standard transfer lens. Lenses 1 and 2 present a magnified real image of the sample surface in the object plane of lens 3, which, converting the electrons originating from a given point of the surface within a given range of angles into a parallel beam inclined at a particular angle to the object plane of the analyzer, provides the Fourier transform of the surface image at the entrance of the analyzer. Passing through the analyzer, an energy-selected, Fourier-transformed real image is obtained at the image plane of the analyzer, and lens 5 inverts this image into a real, magnified, energy-selected image of the surface, which

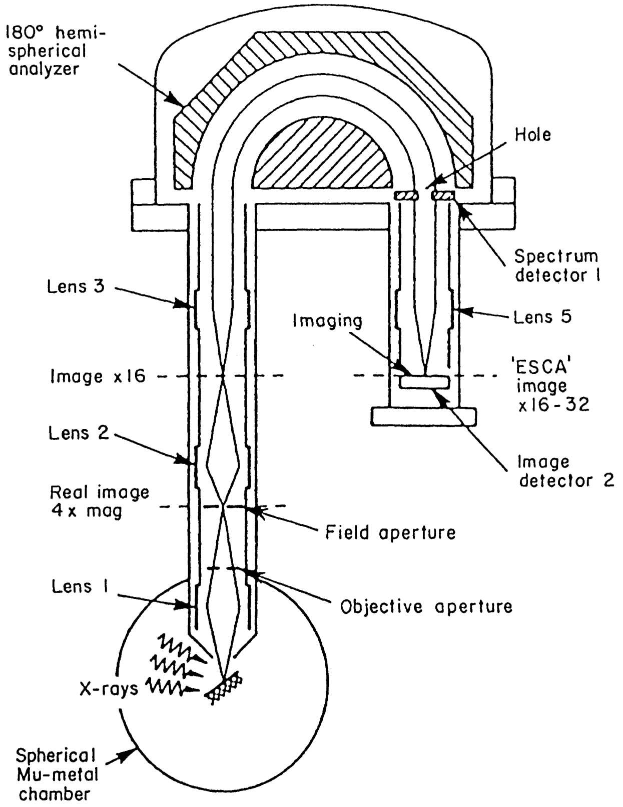

FIG. 21. The principle of operation of the lens system of the ESCASCOPE instrument (Coxon *et al.*, 1990, by courtesy of Elsevier Science B. V.).

is detected by a position-sensitive detector. A spatial resolution approaching 1–2 μm can be achieved in this way.

3.4 New Trends in Applications

Since the first period of the development of high-resolution electron spectroscopy (Siegbahn *et al.*, 1967, 1971), the proliferation of the application of electrostatic spectrometers in various fields continues. Some directions of the evolution carry promises for opening a new era in getting a direct insight into the finest details of the state and dynamic structure of atoms, molecules, and condensed systems.

Tunable laser-induced spectroscopy of excited species has reached an unprecedented level in resolution and sensitivity. In the case of resonant emission multiphoton-ionization photoelectron (or mass) spectroscopy (REMPI-PES), a photoexcited atom or molecule is photoionized, and the respective photoelectron or ion spectra are measured by using mostly TOF or ZEKE techniques, capable, e.g., of resolving the molecular rotation levels. Short-lived interface states in the band gaps between the valence and the conduction bands of semiconductors can be populated optically and photoionized by the help of picosecond or femtosecond pulse-width lasers via the "pump and probe" method, and the energy and angular distribution of the emitted photoelectrons can be measured by a TOF analyzer with angular multidetection (Haight, 1989).

Recent developments in high-resolution electron energy-loss spectroscopy (HREELS, used for obtaining information on surface vibrations and surface phonon dispersion), especially the considerable reduction of the angular aberration of the spherical deflector analyzer by applying deflection plates of optimized shape and voltage, allow attainment of a resolution below 1 meV, making HREELS competitive in this respect with infrared spectroscopy or with inelastic scattering of He atoms (Ibach, 1991, 1993). The scheme of Ibach's spectrometer is shown in Fig. 22. With a careful design, in the case of

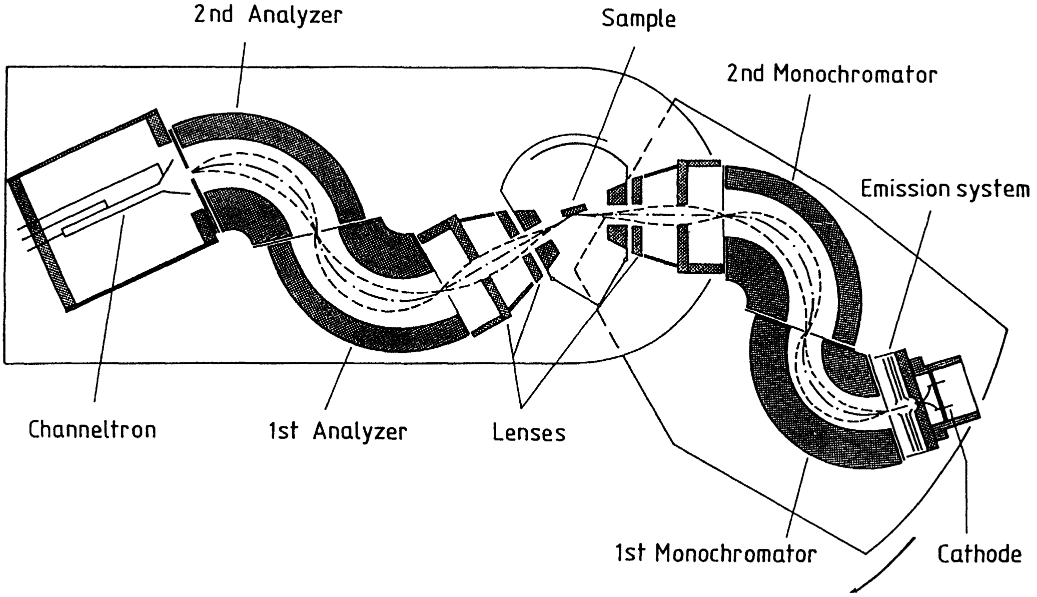

FIG. 22. A typical electron spectrometer for high-resolution electron energy-loss spectroscopy (HREELS) (Ibach, 1991, by courtesy of Springer Verlag). The electron beam emitted from the cathode is monochromatized in the monochromator system and scattered on the sample surface, while the energy spectrum of the scattered electrons is analyzed by the analyzer system with high energy resolution.

this spectrometer (measuring the monochromatic current at the sample as a function of the resolution of the monochromator and comparing to the values obtained from theoretical calculations), no indication of any extra broadening of the energy distribution due to the electron–electron scattering in dense beams known as the Boersch effect (Knauer, 1979) was found (Ibach, 1991). This effect otherwise can be minimized, and as a consequence, the energy resolution can be improved by applying thermal field (Schottky) electron emitters consisting of zirconiated tungsten tips.

The recently elaborated technique of obtaining real-space images around photoemitters by the help of direct Fourier transformation of photoelectron or Auger-electron diffraction data, collected from single-crystal surfaces (*photoelectron holography*), quickly increased the need for devices able to produce full hemispherical angular distribution (electron diffraction) data. One possibility is to scan over the polar and azimuthal angular range by an automated two-axis goniometer, fitted to a commercial hemispherical spectrometer (Naumovic *et al.,* 1993). Utilizing the scanning method and improved correction methods, atomic images with 0.2–0.3 Å (0.02–0.03 nm) accuracy can be obtained even in the case of the determination of adsorbate structures on surfaces (Trevuthasan *et al.,* 1993). Another solution is using a display-type analyzer (Fig. 17) for rapid and straightforward measurements of holograms, permitting a spatial resolution of better than 0.1 Å in transverse planes, due to the large solid angle (Petersen *et al.,* 1993).

As has been discussed in the previous sections, *chemical-state–resolved spectromicroscopy* represents a major trend in the development of electron spectroscopic instruments. There are excellent recent reviews available in this field (e.g., Margaritondo and Cerrina, 1990), reflecting the decisive impact of the development in soft-x-ray optics and of the high-brightness, third-generation synchrotrons. As an illustration, where a different concept using electrostatic lenses is followed, a photoelectron microscope is shown in Fig. 23. Combining the accelerating field of the objective and the small aperture in it and in the focal plane results in a low-pass energy filter reducing the transmission of electrons with more than 1–2 eV energy (Harp *et al.,* 1990). The magnified image is

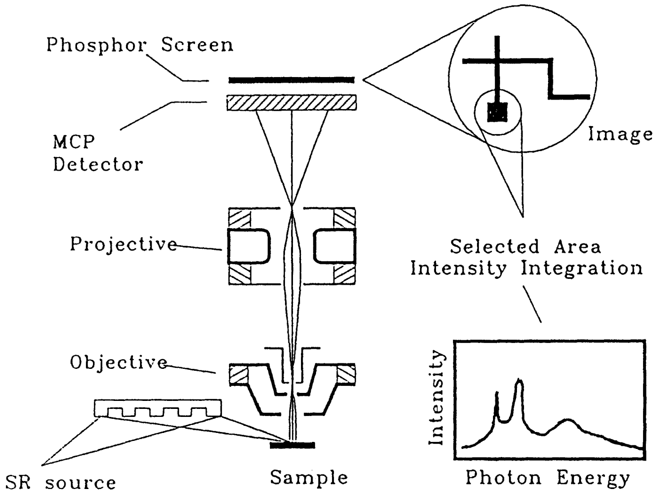

FIG. 23. Photoelectron microscope based on x-ray absorption near edge spectroscopy (XANES) (Harp *et al.*, 1990, by courtesy of The Royal Swedish Academy of Sciences). Here MCP means multichannel plate; SR, synchrotron radiation.

formed primarily by the secondary electrons, whose yield is proportional to the photoabsorption coefficient for the photon energy tuned close to the ionization threshold, applying synchrotron radiation (x-ray absorption near edge spectroscopy, XANES). The achievable spatial resolution is in the micron range.

4. SPECTROMETERS WITH COMBINED ELECTRIC AND MAGNETIC FIELD

With crossed uniform electric and magnetic fields oriented perpendicularly to each other and to the particle trajectory, it is possible to adjust the ratio of these fields to null the Lorentz force on the particle having a particular velocity, while particles with slightly deviating velocities are dispersed and can be filtered out from the beam (*Wien filter*). Minimizing the angular aberration and applying a proper deceleration/acceleration method, the combination of a Wien filter with a scanning transmission electron microscope (STEM) can lead to a high–energy-resolution electron energy-loss spectrometer with a spatial resolution of 1 nm (Batson, 1985). Figure 24 shows the approximate electron-optical ray diagram for the two orthogonal median planes in the spectrometer.

Perpendicularly crossed magnetic and electric fields can be utilized for obtaining energy dispersion of the particles in *trochoidal magnetic guides* of cylindrical symmetry, accepting particles having velocity parallel to the axial magnetic field and being deflected at the end of the guide through a distance dependent on their energy (Granneman and Van der Wiel, 1983).

For operating in a magnetically collimated spectromicroscope (Beamson *et al.*, 1980) and to make use of its axial magnetic field, an electron (positron) energy analyzer has been developed, which acts as a band-pass filter maintaining the spatial integrity of the electron image (Turner and Plummer, 1989). The analyzer design (Fig. 25) is based on a combination of electron image mirrors operating as low- or high-pass filters. In the first stage of the analyzer, the high-energy electrons are filtered out from the electron image through successive reflections by the electron mirrors in this stage, and between reflections the electrons are guided by the crossed field, finally into a high-pass filter, where the electrons having energy lower than a given value are rejected. The second stage of the analyzer, which is similar to the first one except for operating in the opposite direction, corrects the distortion of the image in the first stage. Recent improvements (Turner and Plummer, 1989) achieve a bandwidth of <10 meV for photoelectrons with energy less than 20 eV.

5. ELECTRON AND ION DETECTION SYSTEMS

Except for the direct registration by the help of a Faraday cup, electrons and ions are usually detected through secondary-electron

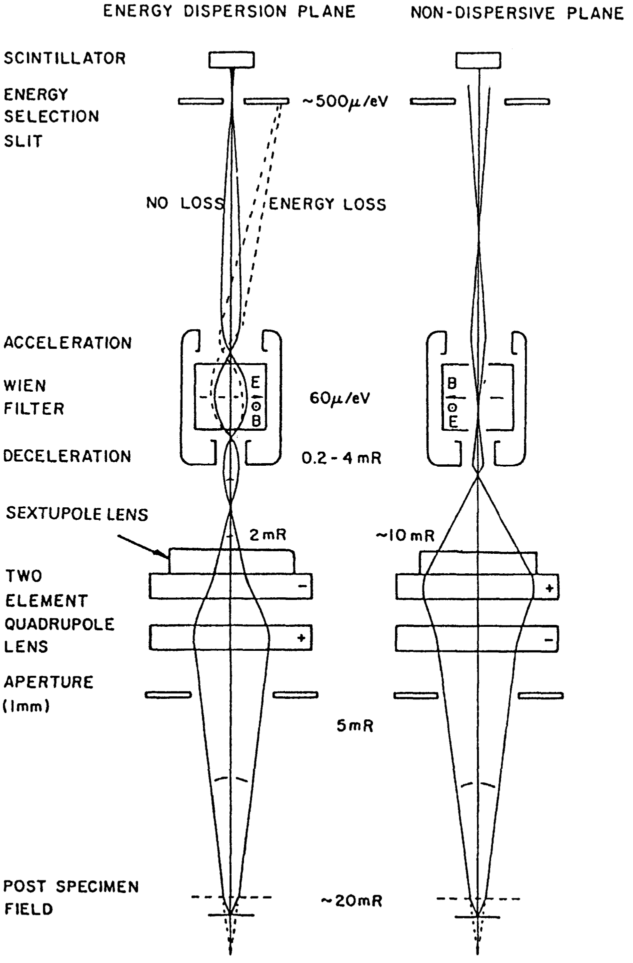
ENERGY DISPERSION PLANE
NON-DISPERSIVE PLANE
SCINTILLATOR
ENERGY SELECTION SLIT
~500μ/eV
NO LOSS
ENERGY LOSS
ACCELERATION
WIEN FILTER
E
B
60μ/eV
DECELERATION
0.2 - 4 mR
SEXTUPOLE LENS
2mR
~10 mR
TWO ELEMENT QUADRUPOLE LENS
-
+
APERTURE (1mm)
5mR
POST SPECIMEN FIELD
~20mR

FIG. 24. Electron-optical ray diagram for the two orthogonal median planes in a Wien-filter spectrometer combined with a scanning transmission electron microscope (Batson, 1985, by courtesy of American Institute of Physics).

multiplication or by electron–hole multiplication in a surface-barrier material or counting diode. For techniques using electron multiplication, recent reviews are available (Granneman and Van der Wiel, 1983; Jede *et al.*, 1992). Detectors based on the principle of electron–hole multiplication are preferred for detecting high-energy (above 10 keV) particles, e.g., in spin analyzers; however, with rapid development of technology, they can provide a promising alternative for high-resolution multidetection.

Electron multipliers can be either *discrete anode* (CuBe or AgMg) or *channeltron* types,

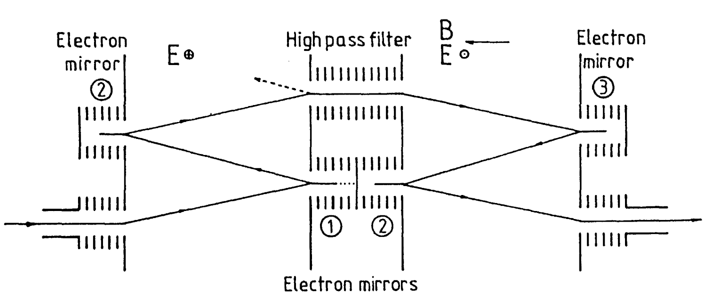
Electron mirror
②
E⊕
High pass filter
B
E⊙
Electron mirror
③
①
②
Electron mirrors

FIG. 25. Imaging band-pass electron analyzer for operation in a uniform magnetic field (Turner *et al.*, 1986, by courtesy of American Institute of Physics). For explanations, see in the text.

providing a gain in the 10^6–10^8 range and being able to work in pulse-counting mode for low particle currents. A channeltron is a curved glass tube with an inner surface layer of a high-resistance, high–secondary-emission–coefficient material representing continuously distributed dynodes. With a high voltage applied between the ends of a channeltron, an incoming particle creates a secondary cascade leading to a gain saturable because of space charge and positive charging of the channel wall. At saturation, a narrow peak occurs in the pulse-height distribution, facilitating the discrimination of the signals from the noise after amplification. The curvature of the channeltron prevents positive feedback from the ions generated from the wall. Because of the recovery time required following a pulse, the maximum count rate is limited to about 10^7 cps. Typical pulse durations are in the range of 10 ns. Although operating only in a good vacuum like discrete anodes, channeltrons are usually less sensitive to venting or to the composition of the residual gas. The efficiency of channeltrons depends on the polarity, charge, and mass of the particles, as well as on the energy (decreasing for higher energies). When higher dynamic range is needed, secondary electrons, induced by ions from a high-voltage metallic electrode, can be accelerated to impinge onto a phosphor screen placed in the window of a photomultiplier (Daly converters: Jede *et al.*, 1992).

Position-sensitive charged-particle detectors are key elements in energy or angular multidetecting systems, as well as in image registration, and are mostly based on the application of *microchannel plates.* Such a device consists of a bundle of short capillary glass tubes, having about a 10-μm inner diameter each, ensuring a large length-to-diameter ratio. Because of the lower (10^3–10^4) gain, at least two plates should be used in series. The maximum count rate is also limited to about 10^6 cps/cm^2. The chevron-type array of the microchannels prevents positive feedback effects. Pulse durations are in the order of 1 ns, due to the small (0.5–2 mm) thickness of the plates, being an important advantage in applications for time-resolved experiments. For position-sensitive detection, various methods can be used. First, *discrete collectors* can be placed behind the channel plates. This solution ensures the highest count rate and signal-to-noise ratio; however, the size of the discrete anodes limits the lateral resolution, and sophisticated electronics are required. Another way is to apply a *fluorescent screen* on top of a fiber-optic faceplate to

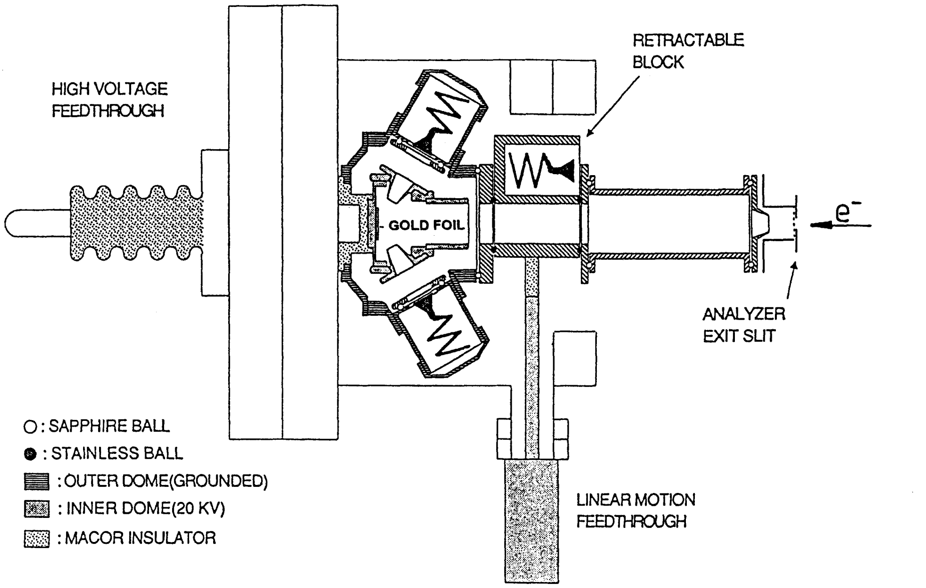

FIG. 26. Cross-sectional view of a micro-Mott spin polarimeter (Huang *et al.*, 1993, by courtesy of American Institute of Physics).

convert the electrons from the channel plate into photons detected by a (one- or two-dimensional) photodiode array or by a sensitive video camera, with digitally controlled readout and image processing. The maximum count rate is limited by the phosphor and the readout time, while the lateral resolution is affected by the broadening from the electron–photon converting and optical elements as well as by the size of the photodiodes. A third possibility is single-particle counting using a resistive anode and a respective anode encoder. In this case the charge of a pulse from the channel plates arriving at the resistive anode is divided according to the resistance ratios between its point of impact and given border points of the resistor. The position of the impact of the particle can be derived from the respective charge (voltage) pulse-height ratios and time differences. Lower count rates are expected with this solution on account of the single-particle counting character and the finite speed of the charge transport.

Detection of *spin-polarized electrons* requires different methods. The most widely used spin analyzer is the Mott polarimeter, based on the asymmetry of the angular distribution of high-energy (10–300 keV), spin-polarized electrons following scattering at large angles from a nucleus of high atomic number. This asymmetry is due to the spin-orbit interaction in the scattering process. The cross section of a micro-Mott polarimeter (Huang *et al.*, 1993) is shown in Fig. 26. It incorporates four channeltrons that detect asymmetric scattering in two orthogonal planes and a fifth retractable one, for cases with no requirement for spin analysis. Incoming electrons from the direction of the analyzer exit slit are accelerated to a gold foil at the end of the detector and biased to +20-kV potential and decelerated by the same field after scattering in front of the channeltrons. In conventional mode, the lens system focuses the electrons into the (fifth) channeltron, while in the spin-analyzing mode, it provides a well-focused beam onto the Au foil. Avoiding the application of high voltage, an alternative option is using electron spin detectors based on the principle of a low-energy (E = 150 eV) diffuse scattering method (Klebanoff *et al.*, 1993).

ACKNOWLEDGMENT

One of the authors (L.K.) is indebted to Dr. D. Varga for the valuable discussions.

Works Cited

Afanas'ev, V. P., Yavor, S. Y. (1976), *Sov. Phys. Tech. Phys.* **20,** 715–733.

Árvay, Z., Fényes, T., Füle, K., Kibédi, T., László, S., Máté, Z., Mórik, Gy., Novák, D., Tárkányi, F. (1980), *Nucl. Instrum. Methods* **178,** 85–93.

Batson, P. E. (1985), *Rev. Sci. Instrum.* **57,** 43–48.

Beamson, G., Porter, H. Q., Turner, D. W. (1980), *J. Phys. E: Sci. Instrum.* **13,** 64–66.

Boerma, D. O., Labohm, F., Reinders, J. A. (1990), *Nucl. Instrum. Methods Phys. Res. B* **50,** 291-299.

Bosch, A., Feil, H., Sawatzky, G. A. (1984), *J. Phys. E* **17,** 1187-1192.

Coxon, P., Krizek, J., Humpherson, M., Wardell, I. R. M. (1990), *J. Electron Spectrosc. Relat. Phenom.* **52,** 821–828.

Daimon, H. (1988), *Rev. Sci. Instrum.* **59,** 545–549.

Dragoun, O. (1983), in: P. W. Hawkes (Ed.), *Advances in Electronics and Electron Physics,* Vol. 60, New York: Academic Press, pp. 1–94.

Eastman, D. E., Donelon, J. J., Hien, N. C., Himpsel, F. J. (1980), *Nucl. Instrum. Methods* **172,** 327–336.

Gelius, U., Wannberg, B., Baltzer, P., Fellner-Feldegg, H., Carlsson, G., Johansson, C.-G., Larsson, J., Münger, P., Vegerfors, G. (1990), *J. Electron Spectrosc. Relat. Phenom.* **52,** 747–785.

Gelius, U., Asplund, L., Basilier, E., Hedman, S., Helenelund, K., Siegbahn, K. (1984), *Nucl. Instr. Methods B* **1,** 85–117.

Granneman, E. H. A., Van der Wiel, M. J. (1983), in: E. E. Koch (Ed.), *Handbook of Synchrotron Radiation,* Vol. 1, Amsterdam: North-Holland, Ch. 6, pp. 367–462.

Haight, R. (1989), *Phys. Rev. Lett.* **62,** 815–818.

Harp, G. R., Han, Zhi-Lan, Tonner, B. P. (1990), *Phys. Scripta T* **31,** 23–27.

Hellings, G. J. A., Ottevanger, H., Knibbeler, C. L. C. M., Van Engelshoven, J., Brongersma, H. H. (1989), *J. Electron Spectrosc. Relat. Phenom.* **49,** 359–382.

Huang, Di-Jing, Lee, Jae-Yong, Suen, Jih-Shih, Mulhollan, G. A., Andrews, A. B., Erskine, J. L. (1993), *Rev. Sci. Instrum.* **64,** 3474–3479.

Huchital, D. A., Rigden, D. J. (1972), *J. Appl. Phys.* **43,** 2291–2302.

Ibach, H. (1991), *Electron Energy Loss Spectrometers,* Springer Series in Optical Sciences, Vol. 63, Berlin and Heidelberg: Springer Verlag.

Ibach, H. (1993), *J. Electron Spectrosc. Relat. Phenom.* **64/65,** 819–823.

Jede, R., Ganschow, O., Kaiser, U. (1992), in: D. Briggs, M. P. Seah (Eds.), *Practical Surface Analysis,* 2nd ed., Vol. 2, New York: Wiley and Salle + Sauerländer, pp. 19–83.

Keenlyside, M., Pianetta, P. (1993), *J. Electron Spectrosc. Relat. Phenom.* **66,** 189–207.

Klebanoff, L. E., Van Campen, D. G., Pouliot, R. J. (1993), *Rev. Sci. Instrum.* **64,** 2863–2871.

Knauer, W. (1979), *J. Vac. Sci. Technol.* **16,** 1676–1679.

Komma, M., Ruoff, H. (1981), *Nucl. Instrum. Methods* **186,** 669–670.

Kővér, Á., Varga, D., Szabó, Gy., Berényi, D., Kádár, I., Ricz, S., Végh, J., Hock, G. (1983), *J. Phys. B: At. Mol. Phys.* **16,** 1017–1027.

Kruit, P., Read, F. H. (1983), *J. Phys. E: Sci. Instrum.* **16,** 313–324.

Leckey, R. C. G. (1987), *J. Electron Spectrosc. Relat. Phenom.* **43,** 183–214.

Margaritondo, G., Cerrina, F. (1990), *Nucl. Instrum. Methods Phys. Res. A* **291,** 26–35.

Mladjenovic, M. S. (1976), *Development of Magnetic β-ray Spectroscopy,* Berlin and New York: Springer Verlag.

Mladjenovic, M. S. (1979), *Nucl. Instrum. Methods* **162,** 193–209.

Naumovic, D., Stuck, A., Greber, T., Osterwalder, J., Schlapbach, L. (1993), *Phys. Rev. B* **47,** 7462–7479.

Nishigaki, S., Kanai, S. (1986), *Rev. Sci. Instrum.* **57,** 225–228.

Nishimoto, H., Daimon, H., Suga, S. (1993), *Rev. Sci. Instrum.* **64,** 2857–2862.

Norell, K. E., Baltzer, P., Wannberg, B., Siegbahn, K. (1984), *Nucl. Instrum. Methods Phys. Res.* **227,** 499–508.

Palmberg, P. W. (1974), *J. Electron Spectrosc. Relat. Phenom.* **5,** 691–703.

Petersen, B. L., Terminello, L. J., Barton, J. J., Shirley, D. A. (1993), *Chem. Phys. Lett.* **213,** 412–421.

Risley, J. S. (1972), *Rev. Sci. Instrum.* **43,** 95–103.

Rivière, J. C. (1990), in: D. Briggs, M. P. Seah (Eds.), *Practical Surface Analysis,* 2nd ed., Vol. 1, New York: Wiley and Salle + Sauerländer, pp. 19–83.

Roy, D., Tremblay, D. (1990), *Rep. Prog. Phys.* **53,** 1621–1674.

Sar-El, H. Z. (1966), *Nucl. Instrum. Methods* **42,** 71–76.

Schlag, E. W., Peatman, W. B., Müller-Dethlefs, K. (1993), *J. Electron Spectrosc. Relat. Phenom.* **66,** 139–149.

Siegbahn, K. (1965), in: K. Siegbahn (Ed.), *Alpha-, Beta- and Gamma-Ray Spectroscopy,* Amsterdam: North-Holland, pp. 79–202.

Siegbahn, K. (1989), *UUIP-1207.*

Siegbahn, K., Nordling, C., Fahlman, A., Nordberg, R., Hamrin, K., Hedman, J., Johansson, G., Bergmark, T., Karlsson, S. E., Lindgren, I., Lindberg, B. (1967), *ESCA, Atomic, Molecular and Solid State Structure Studied by Means of Electron Spectroscopy,* Uppsala: Nova Acta Regiae Societatis Scientarium Upsaliensis.

Siegbahn, K., Nordling, C., Johansson, G., Hedman, J., Hedén, P. F., Hamrin, K., Gelius, U., Bergmark, T., Werme, O., Manne, R., Baer, Y. (1971), *ESCA Applied to Free Molecules,* Amsterdam: North Holland.

Siegbahn, K., Kholine, N., Golikov, G. (1996), UPTEC 96050R, Swedish patent No. 9504642/1.

Staib, P., Dinklage, U. (1977), *J. Phys. E: Sci. Instrum.* **10,** 914–920.

Tőkési, K., Kővér, L., Varga, D. (1994), *Nucl. Instrum. Methods Phys. Res. A* **348,** 173–176.

Trevor, D. J., Van Woerom, L. D., Freeman, R. R. (1989), *Rev. Sci. Instrum.* **60,** 1051–1053.

Trevuthasan, S., Ynzunza, R. X., Tober, E. D., Fadley, C. S., Kaduwela, A. P., Van Hove, M. A. (1993), *Phys. Rev. Lett.* **70,** 595–602.

Turner, D. W., Plummer, I. R., Porter, H. Q. (1986), *Rev. Sci. Instrum.,* **57,** 1494–1500.

Turner, D. W., Plummer, I. R. (1989), *J. Phys. E: Sci. Instrum.* **22,** 593–601.

Van Hoof, H. A., Van der Wiel, M. J. (1980), *J. Phys. E* **13,** 409–414.

Van Hoof, H. A. (1981), *J. Phys. E* **14,** 325–329.

Van Klinken, J., Feenstra, S. J., Wisshak, K., Faust, H. (1975), *Nucl. Instrum. Methods* **130,** 427–441.

Varga, D., Kádár, I., Kővér, L., Mórik, Gy. (1978), *Nucl. Instrum. Methods* **154,** 477–483.

Varga, D., Kővér, Á., Kővér, L., Redler, L. (1985), *Nucl. Instrum. Methods A* **238,** 393–395.

Varga, D., Kádár, I., Ricz, S., Végh, J., Kővér, Á., Sulik, B., Berényi, D. (1992), *Nucl. Instrum. Methods Phys. Res. A* **313,** 163–172.

Zashkvara, V. V., Korsunskii, M. I., Kosmachev, O. S. (1966), *Sov. Phys.—Tech. Phys.* **11,** 96–101.

24

Gamma-Ray Spectrometers

Andrew M. Sandorfi, Christopher J. (Kim) Lister, and Craig Woody

24.1 Introduction 775
24.2 Energy Deposition of γ-Rays in Materials 775
24.3 Solid-State Detectors for Low- and Intermediate-Energy Radiation 776
24.3.1 Solid-State Semiconductors for γ-Ray Detection 777
24.3.2 Requirements of Large Solid-State Detector Arrays 778
24.3.2.1 Efficiency 779
24.3.2.2 Energy Resolution 779
24.3.2.3 Angle Resolution 779
24.3.2.4 Count Rate 779
24.3.2.5 Multiplicity 779
24.3.2.6 Time Intervals 779
24.3.2.7 Spectral Quality 780
24.3.3 Modern Detector Arrays 780
24.4 Shower Development 782
24.4.1 Longitudinal Development 783
24.4.2 Transverse Development 783
24.5 Calorimetry Techniques 785
24.5.1 Calorimeter Types 785
24.5.2 Energy Measurement 786
24.6 Factors Affecting γ-Ray Calorimeter Response 788
24.6.1 Energy Resolution 788
24.6.2 Position Resolution 790
24.6.3 Time Resolution 790
24.7 Readout Devices 792
24.7.1 Optical Detectors 792
24.7.2 Readout Electronics 793
Acknowledgments 793
Glossary 794
References 794
Further Reading 795

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

24.1
Introduction

γ-Rays are the highest-energy photons in the electromagnetic spectrum. Their energies range from tens of kilo electronvolts (10^3 eV, or keV) to multigiga electronvolts (10^9 eV, or GeV) and beyond. Section 24.2 provides an overview of the processes by which γ-rays interact with matter. Below a few mega electronvolts (10^6 eV, or MeV), their detection relies on photoelectric absorption, while at higher energies their interaction results in the creation of electron–positron pairs, which in turn interact with matter by initiating a cascade shower. The detection techniques in these two regimes are distinctly different. The solid-state devices used at low-to-medium energies are discussed in Section 24.3, together with their characteristics and the compromises adopted in modern large segmented detector arrays. In Section 24.4 we discuss the high-energy regime with the physical processes occurring within a cascade shower and the principal shower characteristics. Measurements on the shower allow determination of the γ-ray properties of total energy, arrival position, and arrival time. The various types of calorimeters used for such measurements, along with their characteristics, are discussed in Sections 24.5, 24.6, and 24.7.

24.2
Energy Deposition of γ-Rays in Materials

γ-Rays propagate freely in vacuum at the speed of light leaving no detectable phenomena. It is only through their interaction with matter, and, specifically the generation of charged particles, that they can be detected and studied. Figure 24.1 illustrates the contributions made by the three main interaction processes in the experimentally important sodium iodide scintillation crystal. The figure shows the components of the attenuation coefficient $a(E_\gamma)$, the parameter describing the exponential attenuation of the number N of incident γ-rays as they travel through a crystal to a depth x, as a function of the γ-ray energy E_γ:

$$N = N_0 e^{-xa(E_r)} \tag{1}$$

The dominant process at low energy (≤ 0.2 MeV) is photoelectric absorption of the γ-ray. This results in the emission of a previously bound atomic electron with all the energy of the γ-ray minus the binding energy of the struck electron. As the incident energy is raised, the next phenomenon that becomes important is Compton scattering. In this case, an atomic electron is again struck, but the γ-ray is not absorbed. Instead, the γ-ray scatters off the electron and imparts a fraction of

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

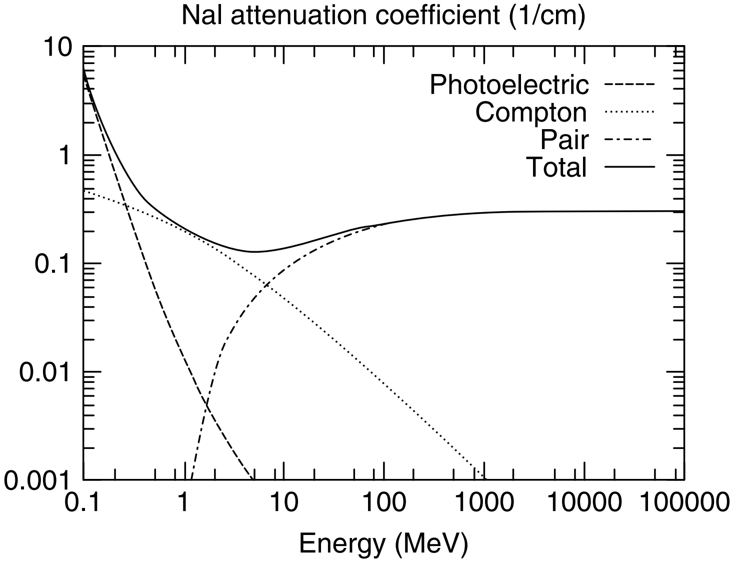

Fig. 24.1 The linear attenuation coefficient of sodium iodide and its decomposition into the contributions from the three interaction processes as a function of γ-ray energy. The results come from the EGS4 Monte Carlo computer program for electron–γ cascade shower modeling (Nelson, Hirayama and Rogers, 1985).

its energy to the struck particle, depending on the angle of scattering. This process is principally important in the transition region of energy before the onset of the third interaction phenomenon.

Pair production is the dominant process by which γ-rays interact with matter in the energy regime above the threshold energy for creating electron–positron pairs ($\geq$1.02 MeV). The γ-ray is transformed into a pair of charged particles, an electron and its antiparticle, a positron. The process requires a virtual photon from the electromagnetic field around a nucleus or electron in order to conserve both energy and momentum. However, essentially, all the γ-ray energy is split between the electron and the positron. The pair emerges nearly symmetrically arranged along the original γ-ray direction with a small average angle between the electron and the γ of $\theta_e \sim m_e c^2/E_\gamma$, where $m_e c^2$ is the rest-mass energy of the electron, 0.511 MeV.

24.3
Solid-State Detectors for Low- and Intermediate-Energy Radiation

Prior to about 1970, γ-rays in the kilo electronvolt to few mega electronvolt range were commonly detected in scintillating materials, which converted the energy deposited into visible light, one of the most common being sodium iodide – see Sections 24.4 and 24.5. A revolution took place with the advent of solid-state devices, particularly germanium counters. These detectors have evolved from tiny crystals with an active volume of a few grams to detector systems containing hundreds of kilograms of active material.

Compared to scintillators, these detectors have remarkable resolution, approaching 1 part in 1000, and are the standard choice for this energy regime.

24.3.1
Solid-State Semiconductors for γ-Ray Detection

The key advantage of using semiconductor materials for detecting electromagnetic radiation lies in the efficient conversion of the photon energy into a large number of charge carriers. The initial photon interaction produces electrons through photoelectric absorption, Compton scattering, and, for energies above 1 MeV, pair production (Section 24.2). The primary electrons thermalize in the semiconductor material, exciting many other electrons from valence states into the conduction band, thus releasing them from lattice sites and allowing them to be swept to the edge of the detector and collected as a charge pulse. In practice, with this process, there are technical issues that must be addressed. First, the material needs be very pure, to keep the number of ambient charge carriers low, and the leakage current small. Second, an electric field needs to be established throughout the detector, to efficiently collect all the charge in the conduction band, before the electrons fall back to valence states through recombination. Third, a rectifying junction needs to be established, to avoid the growth of large leakage currents as the electric field is applied. Reverse biasing this diode junction can allow strong fields to be maintained, more than kilo electronvolts per centimeter, which leads to fast and efficient charge collection. All these issues have been solved and large semiconductor counters are now commonplace for detecting photons with energies below a few mega electronvolts, where a reasonably high fraction of the γ-ray interactions can be contained within the detector material. A comprehensive review of the operating principles of such devices has been made by Knoll (1999).

There are trade-offs to be made in selecting the detector material. In principle, a material with very high density and very small band gap between valence and conduction bands would be ideal: the first to maximize the photon interaction cross section and the second to maximize the number of charge carriers generated. The high-Z aspect of the material is especially important as the γ-ray energy increases, since the probability of interaction changes dramatically both with photon energy and with the detector material (see Section 24.4). In practice, despite many years of effort, silicon and germanium semiconductors still dominate this regime of γ-spectroscopy, though new materials with higher atomic numbers and bigger band gaps, like CdZnTe and HgI_2, and others, are making significant impact in photon imaging and other fields. The band-gap issue is application specific: a small band gap results in more charge carriers, and hence better resolution, but necessitates that the detector be operated at low temperatures, which is usually about 100 K for germanium. A bigger band-gap semiconductor, like in CdZnTe, allows near room-temperature operation, but with loss of resolution. Table 24.1 summarizes the characteristics of common solid-state detector materials.

Large-volume ($>500\,\mathrm{cm}^3$ or $>2.5\,\mathrm{kg}$) germanium detectors, although expensive and requiring cooling, are the counters of choice for research in γ-ray spectroscopy, and usually for isotope identification. They can have good efficiency, excellent energy resolution, and reasonable timing characteristics. For uniform performance,

Tab. 24.1 Characteristics of materials commonly used in semiconductor γ-ray detectors.

Material	Average *Z*	Density (g cm^{-3})	Band gap (eV)	Operating temperature
Si	14	2.33	1.10	$<0\,^\circ$C
Ge	32	5.32	0.67	$<-170\,^\circ$C/100 K
CdZnTe	38	5.50	2.00	0°C
HgI_2	62	6.40	2.15	0°C

detectors are cut from a "boule," a large single crystal of germanium, grown from a molten bath of the material. The first detectors were small and required lithium drifting to keep the leakage currents low, but with improved crystal purification and growing techniques leading to larger volume and purity of material, large intrinsic detectors are now common (up to 10 cm in diameter, and 14 cm in length). These intrinsic materials still retain some impurities. These can be of *p-type*, which can have the largest physical size and highest energy resolution, or *n-type*, which are more radiation hard and can be thermally annealed to remove damage from neutron fluxes encountered in the vicinity of nuclear reactions.

There are limitations on the optimum size of a detector. As the crystals get bigger, more voltage is often needed to deplete the material and allow high internal fields. The physical drift distance for charge also grows, which leads to longer charge collection times and more chance of charge loss in trapping centers. As a result, there has been a move toward mounting multiple crystals in a single cryostat, thus increasing the overall volume of active germanium without pushing the boundaries of crystal growth or operation. The combination of many germanium counters into large arrays has increased the efficiency and sensitivity of experiments by many orders of magnitude.

24.3.2
Requirements of Large Solid-State Detector Arrays

The low and intermediate regimes of γ-rays span the range of energies normally emitted from nuclear bound states. Such emissions arise when the nucleons rearrange themselves to find their most bound configuration. The rearrangements can arise from individual protons or neutrons changing quantum orbits, analogous to atomic X-ray emission, or owing to the more collective bulk motion of nuclei. The resulting radiation has energies from tens of kilo electronvolts to a few mega electronvolts. The study of discrete γ-emissions between nuclear states as the nucleons "fall" from a highly excited configuration to their lowest and most stable arrangement, emitting a burst of γ-rays, has been the most effective tool for investigating nuclear structure. The study of such bursts, which can contain more than 40 γ-rays, requires a range of detector characteristics, sometimes conflicting.

The development of arrays of photon detectors has been progressing for more than 30 years, and is moving into a new generation, the era of γ-ray tracking. Many excellent review articles have been written about this progress, especially for arrays of large germanium counters (Beausang and Simpson, 1996; Lee, Deleplanque, and Vetter, 2003; Gelletly and Eberth, 2006; Eberth and Simpson, 2008).

To develop a spectrometer for investigating modest energy γ-rays, several issues must be considered and optimized. These include the efficiency for detecting photons, the energy resolution, angle resolution, count rate, the number of γ-rays emitted at one time (multiplicity), the time intervals between γ-emissions, and the overall spectral quality (see below). The choice of the detection material and the geometric configuration are largely governed by these issues. All these factors must also be weighed against cost, which can run from thousands to tens of millions of dollars.

24.3.2.1 **Efficiency**

γ-Rays are absorbed in material exponentially, so no detector is ever thick enough to stop all photons. At low-to-medium energies, a compromise thickness corresponding to 80% absorption of the highest-energy photons of interest is often used. The radiation from reactions is emitted in all directions, so a perfect 4π shell of detectors is a goal. In some special applications, for example in decay heat measurements of reactor fuel rods, the total calorimetric sum of γ-energy is more important that individual photons. Maximizing the absolute efficiency, by making a hermetic detector, is very important in such measurements.

24.3.2.2 **Energy Resolution**

Occasionally, just the presence of a γ-ray is important, but most often, high-energy resolution (accurately determining the exact energy of the photon) is critical. At best, with small solid-state detector arrays, energy resolutions of 1 part in 1000 can be approached across this energy regime.

24.3.2.3 **Angle Resolution**

The angle of emission of a γ-ray from a nuclear reaction contains information about the quantum numbers of the parent and daughter states of the decay. For moving nuclei, the Doppler shift of the radiation can be used to infer the lifetime of states. Practically, there is a trade-off between angle resolution and the number of detectors in the system, and how far from the source they are placed.

24.3.2.4 **Count Rate**

The operating mechanism of the detectors determines their count-rate limitation. While for fast scintillators individual counters can detect photons at megahertz rates, large semiconductors are limited to only tens of kilohertz. This has to balance against the number of detectors in the system and their distance from the source to provide a viable operating regime.

24.3.2.5 **Multiplicity**

Some nuclear reactions produce in excess of 40 γ-rays in the cooling process. To retain a good chance of catching individual photons, without any one crystal registering multiple hits, demands high segmentation.

24.3.2.6 **Time Intervals**

The cooling cascade of radiation is usually fast, and is often complete in less than 500 picoseconds (5×10^{-10} seconds). However, occasionally, nuclear states cannot easily decay, and the ‘‘hesitation’’ can be directly measured electronically, especially if the detectors have good time characteristics. Good timing is usually directly associated with high count-rate capability, but it is usually anticorrelated with good energy resolution.

24.3.2.7 **Spectral Quality**

In a perfect detector, a monoenergetic photon would always lead to a monoenergetic signal output. In practice, this never happens, either owing to the primary process (for example, loss of a Compton scattered photon from the counter), or because of detector imperfection (for example, loss of produced light or charge). The spectrum quality is usually defined as the ratio of "full energy," or "photopeak" events, to the total number of interactions. As such, it is essentially the fraction of γ-rays for which a measurement of their total energy is obtained. For low energy radiation (~0.2 MeV) this can be high, >80%, but it falls with energy as secondary photons leak out of the detector.

24.3.3
Modern Detector Arrays

To evaluate the competing criteria, a variety of *figures of merit* have been discussed, which allow arrays to be optimized for specific applications (Beausang and Simpson, 1996; Lee, Deleplanque and Vetter, 2003; Gelletly and Eberth, 2006; Eberth and Simpson, 2008). The most challenging conditions generally involve fusion-evaporation reactions, which produce nuclei at high angular momentum from which large numbers of γ-rays are emitted in the de-excitation process. In such regimes, an effective *resolving power* is frequently used, which reflects the ability to observe the individual γ-ray transitions as the nuclei de-excite. In Figure 24.2, increases in resolving power have accompanied each generation of detector systems. Detector technology is fast approaching a

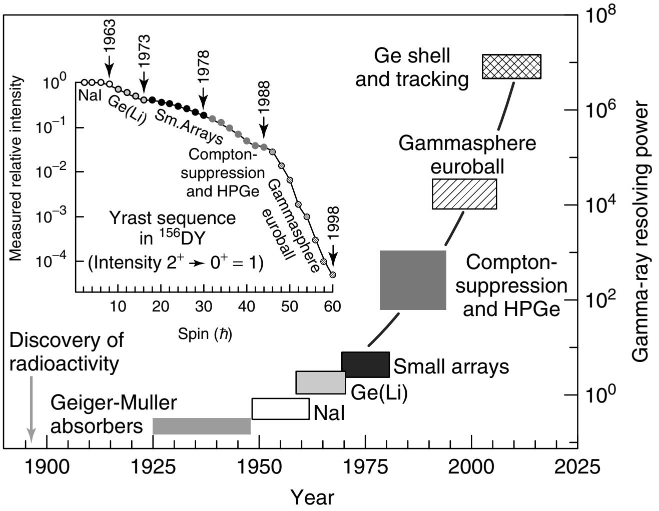

Fig. 24.2 The evolution of γ-ray detection technology from the discovery of radioactivity to the present. The *resolving power* plotted here reflects the ability to observe transitions between rare and exotic nuclear states (LRP, 2002). The insert shows a typical relative intensity for the production of the lowest energy states of a given angular momentum. Excited nuclear states decay by sequentially emitting γ-rays along this *Yrast* sequence. There is a strong inverse relationship between *resolving power* and the experimental limit for the observation of highly excited nuclear states with large angular momentum.

factor of a million in improvement since γ-rays were first observed.

The current state-of-the art detector is the *Gammasphere* array, which was built by a collaboration of US National Laboratories and Universities and has been in operation since 1996 as an international facility for nuclear physics research (Gammasphere, 2008). This is a hybrid system, consisting of 110 separate modules, each with a germanium detector mounted in a bismuth germinate (BGO) shield (Table 24.2) to suppress events in which a Compton scattered photon leaks out, thereby improving the *spectral quality*. The detector can be operated in many modes: as a calorimeter, as a multiplicity filter, or as 110 individual counters. In the high-resolution germanium operation, it has ~10% absolute efficiency for 1.33 MeV γ-rays.

New technologies to progress beyond the Gammasphere generation are currently under study as a project called the *Gamma-Ray Energy Tracking Array* (GRETA, Beausang, 2003). This will involve building a hermetic pure germanium shell. Each detector will have highly segmented electrodes that allow location of the charge deposition. This information allows

Tab. 24.2 Properties of crystal scintillators and Cherenkov materials. The right-most column gives the light output relative to NaI(Tl).

	Density (g cm^{-3})	Refractive index	Radiation length (cm)	Molierè radius (cm)	Emission maximum (nm)	Decay time (ns)	Light output
NaI(Tl)	3.67	1.85	2.59	4.5	410	250	1.00
BGO	7.13	2.20	1.12	2.4	480	300	0.15
LSO	7.40	1.82	1.14	2.1	420	40	0.83
GSO	6.71	1.85	1.38	2.2	430		
Fast						56	0.30
Slow						600	0.03
BaF_2	4.89	1.56	2.05	3.4			
Fast					220	0.7	0.05
Slow					310	620	0.20
CsI(Tl)	4.53	1.80	1.85	3.8	565	1000	1.40
CsI	4.53	1.80	1.85	3.8			
Fast					305	10,35	0.10
Slow					480	1000	0.02
$LaBr_3$(Ce)	5.29	1.90	1.88	2.85	356	20	1.30
$LaCl_3$(Ce)	3.86	1.90	2.81	3.71	335		
Fast						24	0.42
Slow						570	0.13
CeF_3	6.16	1.68	1.70	2.6	340	10,30	0.10
$PbWO_4$	8.28	2.16	0.90	2.2	500	2,15	0.01
PbF_2	7.77	1.82	0.93	2.2	Cher.	Cher.	3.0×10^{-3}
Pb Glass (SF5)	4.08	1.67	2.54	3.7	Cher.	Cher.	1.0×10^{-3}
H_2O	1.00	1.33	36.1	6.8	Cher.	Cher.	0.8×10^{-3}

"tracking" the reconstruction of events. Tracking has many advantages, including recovering interactions where photons scatter out of the shell. The first phase of this array, a 1/4 prototype GRETINA (Lee *et al.*, 2004), is expected to be operational in 2012. For applications involving the observation of multiple γ-rays, GRETA will surpass Gammasphere by a factor of 1000.

Whenever knowledge of the precise energy has been important, solid-state detectors have revolutionized the scientific study of moderate energy γ-rays. Nonetheless, in some investigations, such as in β-decay, the total integrated flux of γ-rays is of paramount importance, as it can reveal the "strength function" distribution of the decay. This has both spectroscopic interests, for example in investigating the distribution of Gammow–Teller decay strength, and practical interest, as it is a factor in determining "decay heat" in a reactor. For these applications, hermiticity is the key factor, and large-volume scintillators still remain the best solution. For low- to medium-energy γ-rays, these are comparatively simple and inexpensive to operate. As the energy grows above the $e^{\pm}$ pair threshold, and a cascade shower begins to develop, hermiticity becomes a complex issue and the detection systems become elaborate. We discuss this in the next section.

24.4
Shower Development

There are two main processes by which energetic electrons (and positrons) are brought to rest:

1. collisions with atomic electrons, resulting in either ionization energy loss for soft collisions or division of the energy between the incident particle and the atomic electron for harder collisions; and
2. emission of a γ-ray in the electromagnetic field of a nucleus or electron, termed *bremsstrahlung*.

(For positrons, there is the additional process of annihilation, but this mainly occurs at rest and produces two γ-rays each with half the pair-creation threshold energy.) Collisional energy losses dominate at low energy, but bremsstrahlung dominates at high energy. The crossover point is at a material-dependent critical energy (Bethe and Ashkin, 1953),

$$E_c \simeq \frac{1600 m_e c^2}{Z_{\text{eff}}} \tag{2}$$

where $m_e c^2$ is the rest-mass energy of the electron, 0.511 MeV, and Z_{eff} is an appropriately weighted average of the nuclear charge of the material constituents. It is interesting to note that the dominant high-energy electron interaction, bremsstrahlung, is related by crossing symmetry to the dominant high-energy photon interaction, pair production. Thus, it is not surprising that the length of material in which a high-energy electron loses $1/e$ of its energy to radiation, the radiation length X_0 (Tsai, 1974, 1977),

$$\frac{1}{X_0} \simeq 4Z\,(Z+1)\,n\alpha r_e^2 \ln\left(183 Z^{-1/3}\right) \tag{3}$$

is approximately the same as the mean distance for a high-energy γ to travel before producing an $e^{\pm}$ λ_{pair}, where

$$\frac{1}{\lambda_{\text{pair}}} \simeq (7/9)4\text{Z}(\text{Z}+1)n\alpha r_e^2 \ln\left(183 Z^{-1/3}\right) \tag{4}$$

where Z is the charge of the nucleus, $n = \rho N_a/A$ is the atomic number density,

ρ is the mass density of the absorber, N_a is Avogadro's number, A is the atomic weight of the nucleus, α is the fine-structure constant 1/137, and r_e is the classical radius of the electron, 2.8×10^{-13} cm. Similarly, the emission angle between the bremsstrahlung γ and the initial electron is characterized by the same angle, $\theta_\gamma \sim m_e c^2/E_e$, as the pair emission.

24.4.1
Longitudinal Development

The effect of combining these two processes, bremsstrahlung and pair production, is the creation of an electron–γ cascade shower, as shown schematically in Figure 24.3. An incident γ-ray produces an electron–positron pair, which then loses energy by radiating lower-energy γ's, which, in turn, generate more pairs, and so on, until the electrons and positrons are below E_c and no longer lose significant energy by radiation. The strong forward peaking of the pair and bremsstrahlung emission means that the shower tends to progress in a straight line along the initial direction. A vastly oversimplified approximation to the showering process would be that the average energy per particle is halved in each radiation length and the number of particles doubles until the average energy is reduced to the critical energy, at which point the shower terminates. Then the linear extent of the shower is $\sim X_0 \ln(E_\gamma/E_c)/\ln 2$, and the maximum number of particles is just $\sim E_\gamma/E_c$, with approximately equal numbers of γ-rays, electrons, and positrons at each step. Finally, the integral track length of the charged particles in this simplified model is $\sim(2/3\ln 2)(E_\gamma/E_c)X_0$. More detailed calculations support the following qualitative features: that the shower depth is logarithmic in incident energy, and the number of shower particles as well as the integral track length are both proportional to the incident energy.

24.4.2
Transverse Development

After the first few radiation lengths, the transverse extent of the shower is determined by the multiple scattering of the electrons and positrons. According to the Molière theory of multiple scattering (Scott, 1963), the lateral displacement of an electron with energy equal to the critical energy traveling a distance of one radiation length defines the so-called Molière radius,

$$R_M = \sqrt{\frac{4\pi}{\alpha}} m_e c^2 \frac{X_0}{E_c} \tag{5}$$

which characterizes the shower distribution perpendicular to the shower axis. Experimentally, it is found that a

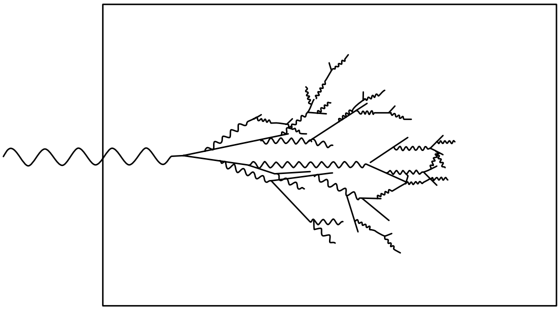

Fig. 24.3 Schematic representation of how an electron–γ cascade shower develops. A high-energy γ-ray enters from the left and produces a pair, which, in turn, emits bremsstrahlung to make more γ-rays, and so on.

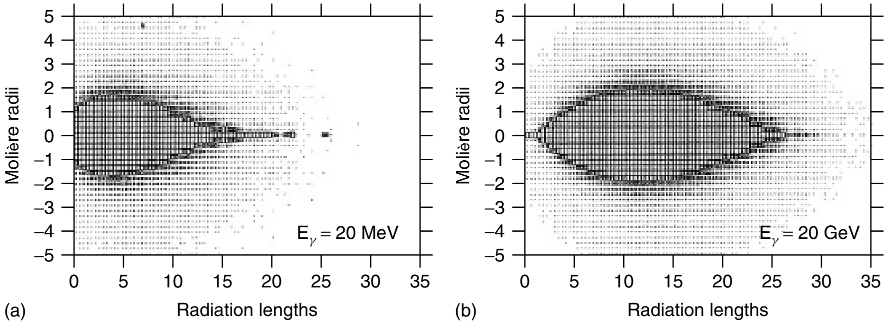

Fig. 24.4 Average relative energy deposition from an electron–γ cascade shower. The size of the box represents the amount of energy deposited in 1-cm cubes of NaI in that position, averaged over many showers. Maximum box size occurs at a level such that the saturated boxes contain 90% of the γ-ray energy. A 20-MeV γ-ray enters from the left in the plot (a) and a 20-GeV one in (b). The radiation length of NaI is 2.6 cm, and the Molière radius is 4.5 cm (see Table 24.2).

cylinder of radius $1R_M$ contains ∼90% of the shower energy deposition; $2R_M$, ∼97%; and $3R_M$, ∼99% (Kleinknecht, 1982).

Figures 24.4 and 24.5 present the results of detailed simulations of such electron–γ cascade showers in NaI with the GEANT Monte Carlo program (Brun and Carminati, 1993). Figure 24.4 shows contours of the energy deposition along and transverse to the initial direction for γ-rays of 20 MeV and 20 GeV. This is the result of averaging over many such showers. As can be seen, the factor-of-1000 increase in energy between the two showers produces a deeper penetration, by about 10 [$\simeq \log_2(1000)$] radiation lengths ($X_0 = 2.6$ cm in NaI, Table 24.2). Both showers exhibit similar radial extent, approximately equal to the Moliere radius ($R_M = 4.5$ cm in NaI, Table 24.2). Any given shower will produce a much more sparse and random energy deposition. Figure 24.5 attempts to give a feel for this by plotting the average number of 1-cm cubes of NaI in which any shower energy

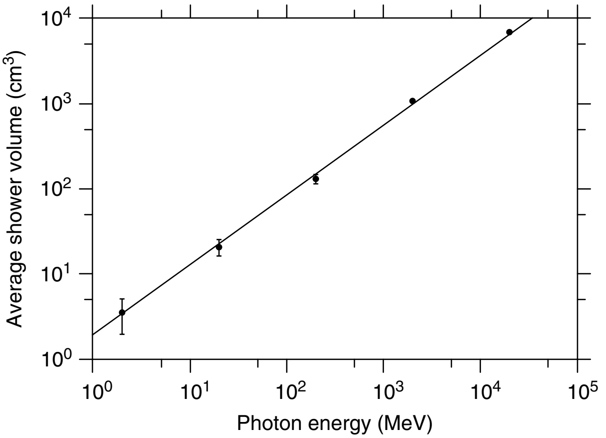

Fig. 24.5 Average number of 1-cm cubes of NaI that receive any energy deposition during an electron–γ-cascade shower as a function of shower energy. The error bars represent the variation in the number of cubes from shower to shower.

deposition has occurred as a function of incident γ-ray energy. For the 20-MeV case, this is only 20 cubes versus the 3400 cm^3 required in Figure 24.4 to contain 90% of the energy. For the 20-GeV case, the comparison is 6820 cubes versus 7000 cm^3.

24.5
Calorimetry Techniques

24.5.1
Calorimeter Types

γ-Ray detection takes place through the fundamental interactions of photons with matter described above. At high energies, the dominant process of pair production leads to electron–positron pairs, which undergo ionization and radiative energy loss in the detector material. The calorimetry technique makes use of the fact that this energy loss is proportional to the total energy deposited in the detector in order to measure the initial photon energy. There are, in general, two classes of calorimeters that are used to measure this energy. The first consists of a *totally active* detector in which all the initial energy is deposited in a material that produces an observable signal. The second is a *sampling* detector in which only a fraction of the initial energy is deposited in an active material, and the rest is deposited in a passive material and is not observed.

Totally active detectors usually consist of a high-density, high-Z material, such as an inorganic crystal or glass, which provides a high conversion efficiency for an incident photon. Table 24.2 lists some of the commonly used inorganic scintillators. The inorganic scintillators have the highest light output and provide the best energy resolution at lower energies. The last (right-most) column of Table 24.2 gives the light output relative to NaI, which yields 38 000 optical photons/MeV of energy deposited (see Table 24.3). The incident photon undergoes an initial interaction on a high-Z atom according to the processes described above, and the resulting electron or positron produces ionization, which leads to the formation of electron–ion pairs. The electrons and ions then transfer their energy to luminescent centers in the material, which can be either intrinsic, such as in a BaF$_2$ crystal, or a dopant or fluor that has been added as an activator, such as thallium (Tl) in the case of NaI or CsI. The excited luminescent centers then radiate their energy in the form of scintillation light, typically in either the visible or the ultraviolet (UV) wavelength range. In this process, each step takes place

Tab. 24.3 Efficiency factors for crystal scintillators.

Scintillator	L (γ's/MeV)	β	S	Q	η
NaI(Tl)	38 000	0.88	0.59	1.00	0.52
CsI(Tl)	65 000	0.97	0.99	1.00	0.96
CsI (pure)	2 000	0.97	0.29	0.10–0.15	0.03
BGO	8 200	0.69	1.00	0.13	0.09
BaF$_2$					
Fast	1 800	<0.82	>0.9	1.00	0.08
Slow	9 950	<0.72	>0.33	1.00	0.24
CeF$_3$	3 200	0.61	0.13	1.00	0.08

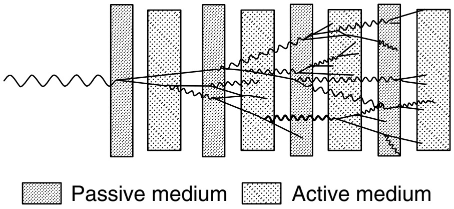

Fig. 24.6 Schematic representation of a sampling calorimeter.

with a certain efficiency that contributes to the overall scintillation efficiency of the material. If we denote the conversion efficiency as β, the efficiency for energy transfer to the luminescent center as S, and the radiative decay efficiency as Q, then the overall scintillation efficiency η is given by $\eta = \beta SQ$. Table 24.3 gives the respective efficiency factors for some commonly used scintillating crystals, taken from Lempicki, Wojtowicz and Berman (1993).

The Cherenkov materials, some of which are also listed in Table 24.2, produce light only by Cherenkov radiation and generally have a much lower light output than inorganic scintillators. Cherenkov radiation is the light emitted by charged particles when their velocity exceeds the phase velocity of light in the medium. It is a very small fraction of the total energy-loss process. These materials are therefore used most often at higher energies where the total amount of energy deposited is much higher. (The columns labeled ''Emission maximum'' and ''Decay time'' in Table 24.2 are not relevant for the Cherenkov materials. The emission is nonresonant and prompt, and so entries have been made with ''Cher.'' to identify these materials.)

A schematic representation of a sampling detector is shown in Figure 24.6. Typically, the passive medium consists of some high-density material, such as lead, which provides the high photon conversion efficiency, while a lower-density material is used as the active medium. Because only a fraction of the total energy is measured, a sampling detector generally has poorer energy resolution than a totally active detector but can be built for a much lower cost on account of the smaller quantity of active detector material required. It is therefore widely used for large-area, high-energy calorimeters.

24.5.2
Energy Measurement

There are several techniques that are used to measure the energy deposited in the active medium in both a totally active and a sampling detector. These are generally based on either the direct detection of the ionization charge or the detection of either scintillation or Cherenkov light produced in the active medium. Direct ionization can be detected in gases, liquids, or solids, and the use of these materials depends largely on the energy range of interest. Because of their low density, gases are rarely used in γ-ray detection, even in sampling calorimeters. Liquids, in particular noble liquids, are often used in sampling detectors because of their relatively high Z and the ability to collect the ionization charge easily using readout electrodes inside the detector. A high voltage is applied across the gap formed by these electrodes, which causes the charge to drift across the gap where it is collected and measured using charge-sensitive detectors. Table 24.4 lists some of the properties of noble liquids that have been used

Tab. 24.4 Properties of noble liquids.

	Liquid Ar	Liquid Kr	Liquid Xe
Density (g cm^{-3})	1.39	2.45	3.06
Radiation length (cm)	14.3	4.76	2.77
Molière radius (cm)	7.3	4.7	4.1
dE/dx (MeV cm^{-1})	2.11	3.45	3.89
Drift velocity (cm μs^{-1})	0.5	0.5	0.3
W-value (eV/ion pair)	23.6	20.5	15.6
Dielectric constant	1.51	1.66	1.95
Triple point (K)	84	116	161

for γ-ray detection. Because of its lower cost, liquid argon has been the preferred choice. All of these liquids operate at cryogenic temperatures and are typically cooled by using liquid nitrogen. Although the implementation of such noble liquid detectors is complex, nonetheless, several have been used in large-area, high-energy physics applications (DØ Collaboration, 1994; SLD Collaboration, 1984).

Plastic scintillators, again because of their low cost, are also widely used in large-area calorimeters. Table 24.5 gives a list of some of the commonly used plastic scintillators along with some of their properties. The light output is given relative to anthracene, an early material now rarely used except as the standard for reporting light output. One aspect of sampling detectors that use scintillating plastic is the need to collect the light produced onto a readout device. This is generally accomplished through the use of light guides, or by allowing the scintillation light to excite another fluorescent material known as a *wavelength shifter* that carries the light to the photodetector. The wavelength-shifting technique has been extensively used because of its ability to provide a wide variety of readout techniques. One such method is illustrated in Figure 24.7 (PHENIX Collaboration, 1993). In this example, the light from scintillation plates, which are interleaved between plates of lead, is used to excite wavelength-shifting fibers that pass longitudinally through the detector. The fibers are then bundled together at the back and read out with photomultiplier tubes (PMT). This provides a convenient way of placing

Tab. 24.5 Properties of Saint-Gobain (Bicron) plastic scintillators (Hurlbut, 1985).

Scintillator	Light output (% anthracene)	Decay time (ns)	Wavelength of maximum emission (nm)	Attenuation length (cm)	Application
BC 400	65	2.4	423	250	General purpose
BC 404	68	1.8	408	120	Fast counting
BC 408	64	2.1	425	380	Large area
BC 412	60	3.3	434	400	Large area
BC 418	67	1.4	391	100	Ultrafast timing
BC 428	50	12.0	490	330	Green emitting

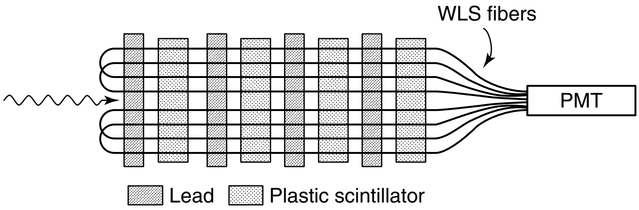

Fig. 24.7 A sampling calorimeter with wavelength-shifting (WLS) fiber-optic readout.

the readout devices at the back of the detector where they will not interfere with any of the incident particles.

Finally, sampling detectors using silicon semiconductor (Table 24.1) as the active medium have also been built and provide good energy resolution because of the large amount of charge produced. However, because of the significant cost of high-purity silicon, these detectors have only been used in rather limited, small-area applications.

24.6
Factors Affecting γ-Ray Calorimeter Response

The overall performance of a γ-ray detector can be characterized by three main categories of resolution that are important experimentally. These are energy resolution, position resolution, and time resolution. There are a number of factors that determine these properties in any calorimeter design.

24.6.1
Energy Resolution

The energy resolution of a calorimeter can be parameterized by the following expression:

$$\left(\frac{\Delta E}{E}\right)^2 = \left(\frac{a}{\sqrt{E_\gamma}}\right)^2 + \left(\frac{b}{E_\gamma}\right)^2 + c^2 \quad (6)$$

The first term is a *statistical* component and contains contributions to the energy resolution due to statistical fluctuations. These can be fluctuations in the amount of primary ionization produced (e.g., the total number of ionizing tracks produced in the electromagnetic shower), or fluctuations in the number of observed particles (either photoelectrons or ion pairs). Usually, for a high-energy shower, the number of ionizing tracks produced is large; hence, the fluctuations are rather small ($\leq 0.5\%/\sqrt{E}$), and the fluctuations in the number of observed particles dominates (Amaldi, 1981). For a sampling detector, this includes a contribution from the so-called *sampling fluctuations,* which are due to the fact that only a fraction of the energy is being detected, and this fraction can vary from one shower to the next. For high-Z inactive radiators of critical energy E_c, radiation length X_0, and total length L_x, the sampling fluctuation contribution to the energy resolution (full width at half maximum of the resulting signal) for γ-rays of energy E_γ can be approximated as (Amaldi, 1981)

$$\frac{\Delta E}{E} = 7.5\% \sqrt{\frac{L_x}{X_0}\,\frac{E_c}{E_\gamma}\,\frac{[\text{MeV}]}{[\text{GeV}]}\,\frac{1}{G(Z, \cos\theta)}} \quad (7)$$

Here, $G(Z, \cos\theta)$ is a complicated function of geometry, the Z of the radiator material, and the angle θ that characterizes the mean angular spread of particles around the shower axis, but it is typically a number between 1.0 and 1.3. In a sampling calorimeter, these fluctuations generally make the largest contribution to the statistical term in the resolution.

The term of Eq. 24.6 with coefficient b contains contributions that are fixed in magnitude relative to the observed signal. This *noise term* contains contributions from various sources of noise. These include any electronic noise in the readout devices or their associated electronics, as well as any constant background noise. Since this term goes as $1/E$, it is most important at lower energies and is often negligible at higher energies.

The term with coefficient c reflects complex functions of the detector that are rather slowly varying with energy and are often approximated by a constant. The two most important are the degree of shower containment and the uniformity of response throughout the active volume. As the fraction of the shower energy escaping a detector volume increases with increasing γ-ray energy, the fluctuations in this signal loss grow until a roughly constant plateau in resolution is reached (Dowell *et al.*, 1987). This contribution can be made small with sufficiently large detector volumes, although the associated costs increase sharply with energy.

Even for large totally active detectors, such as NaI(Tl), inhomogeneities in the crystal can be a fundamental limitation to the ultimate energy resolution one can achieve. The inhomogeneities limit the performance through the variation in light generation and collection for the same amount of energy deposited in different regions of the active volume. This is a significant limitation at moderate energies – for example, the 20-MeV shower shown in Figure 24.4 – where the variations in the shower distributions are large. The effect is much smaller at high energies where each shower fills a large detector volume (see Figure 24.5). Variations are governed by crystal impurities and defects, by deviations in the local concentrations of dopant, by the reflective properties of crystal surfaces, by solid-angle effects and the placement of the light-collection devices, and by nonuniformities in efficiency across the surfaces of the light collectors (Dowell *et al.*, 1990). Controlling, or even predicting, any one of these components is quite difficult. However, their combined effect can be studied with muons and unfolded with the techniques of computer-aided tomography (CAT, or μCAT here) (Dowell *et al.*, 1990).

High-energy muons produce only narrow, well-defined ionization tracks that can completely penetrate even large detectors. Being weakly interacting, they do not undergo nuclear reactions and can be used to deposit a known amount of energy in every element of detector volume along their path. The response to many muons traversing well-defined paths can be combined with Fourier reconstruction algorithms to produce three-dimensional images of the nonuniformities in light generation and collection. As an example, a section of such a μCAT map – a two-dimensional slice through a 24-cm-diameter × 36-cm-long NaI(Tl) crystal viewed at one end by seven 7-cm photomultiplier tubes – is shown in Figure 24.8. The gray-scale image shows the shift in the net collected signal for a nominal muon energy deposition of 5 MeV per cm of track length in each cell. (It is interesting to note that a visual inspection of the interior of this crystal failed to reveal any trace of the complex patterns evident in this μCAT map.) These nonuniformity images can be combined with Monte Carlo simulations of energy deposition in electromagnetic showers to predict

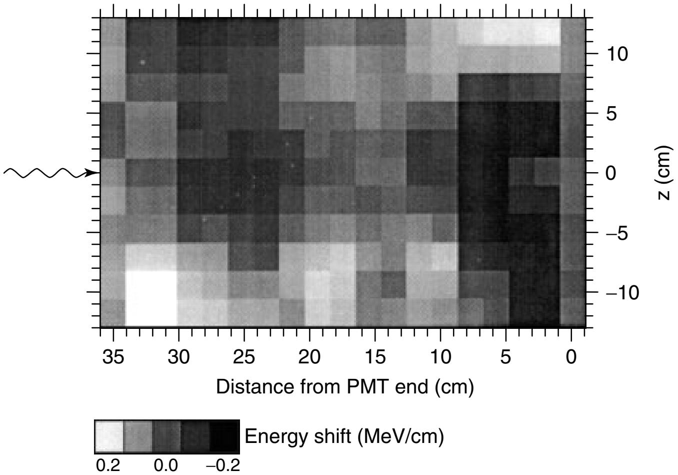

Fig. 24.8 A μCAT scan of an NaI detector, in which cosmic-ray muons deposit the same amount of energy in each cell.

accurately the resolution for γ-rays of any energy (Dowell *et al.*, 1990). On the basis of the μCAT scan in Figure 24.8, the predicted resolution for that detector is shown as the solid line in Figure 24.9, together with measured results. (For comparison, the dashed curve gives the expected behavior of a perfectly uniform detector.) Although the nonuniformities are complex and highly varying, the volume sampled by the γ-ray showers is large (see Figure 24.4) and the net effect is to produce a nearly constant resolution at high energies.

24.6.2
Position Resolution

The spatial resolution of the calorimeter is determined primarily by the density of the detector materials as well as its spatial segmentation. As mentioned earlier, the containment of a shower in the transverse direction is determined by the Molière radius. Therefore, the natural unit for transverse segmentation of a calorimeter is in cells of approximately one Molière radius square. However, the position resolution of the calorimeter will be significantly better than this, since the shower position can be determined from a fit using information from many cells that contain energy from the shower. This can be done as a simple energy-weighted average given by

$$x_{\text{measured}} = \frac{\sum \omega_i x_i}{\sum \omega_i} \tag{8}$$

where ω_i is the energy measured in the ith calorimeter cell and x is the position of that cell for the coordinate being measured. However, more sophisticated techniques are often used where the weighting parameters depend logarithmically on the energy to better account for large transverse energy fluctuations (Awes *et al.*, 1992).

24.6.3
Time Resolution

There are two main aspects of the time resolution of a γ-ray calorimeter. The first is the variation in the measured arrival

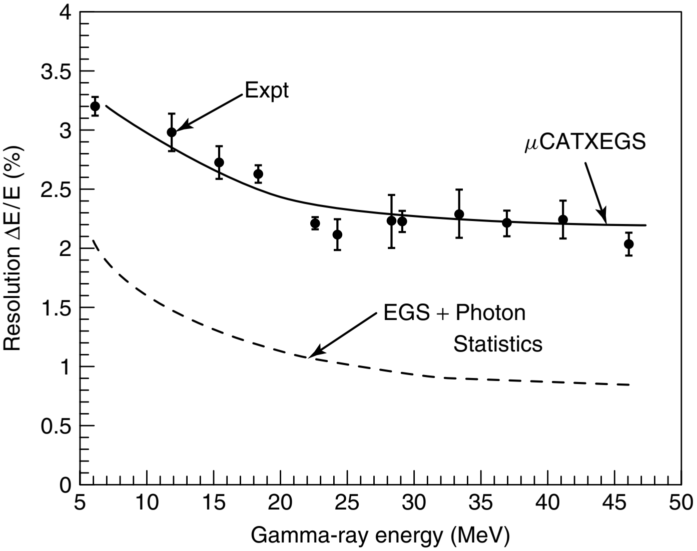

Fig. 24.9 Measured resolution of an NaI detector at several energies, the μCAT predicted resolution (solid line), and the resolution of a perfectly uniform detector (dashed curve).

time of the signal produced by an incident particle, and the second is the ability of the calorimeter to resolve two particles that are close together in time. The first aspect, called the *leading-edge* time resolution, is most important in time-of-flight applications, where the calorimeter is used to measure the transit time of particles arriving from some external source. The principle factors contributing to this resolution are

1. the variation in time in the generation of the detectable signal, which can be either optical photons from a scintillator, Cherenkov photons, or ionization charge;
2. the time variation in the propagation of that signal to the readout device;
3. the variation in time due to changes in the amplitude or shape of the electronic pulse from the readout system, including contributions from electronic noise.

In most cases, the time variation in the production of the primary signal is small compared to the other effects, since the timescale involved for producing either optical photons or ion pairs is rather short ($\leq$1 nanoseconds). The second effect, which is mainly due to light or charge collection, can be significant, especially in a large calorimeter. It is affected by the propagation time in the detection medium (see Table 24.2 for indices of refraction in optical materials and Table 24.4 for drift velocities in noble liquids), as well as the size of the detector. The third effect includes the time variation in the detector response and readout system due to differences in energy deposition. Since the observed energy is subject to statistical fluctuations, this effect can be dominant, particularly at low energies. In general, the larger the prompt signal from the detector and the faster the response time of the readout system, the better will be the time resolution. For optical detectors producing large signals and using fast readout devices, it is possible to achieve time resolutions in the range from a few tenths down to a few hundredths of a nanosecond. For charge-collection devices, depending on the size of the signal, it is possible to obtain time resolutions on the order of a few nanoseconds.

The second aspect of a calorimeter's time resolution is its so-called *pulse-pair* resolution, or resolving time. This is the ability of the calorimeter to resolve two particles closely spaced in time. The main factor determining this resolution is the width in time of the electronic readout pulse. This depends on the response time of the active material (e.g., scintillation decay times, as in Tables 24.2 and 24.5), the propagation time in the detector (e.g., drift velocities in noble liquids, as in Table 24.4), and the response time of the readout system. For a fast device, this can be on the order of a few nanoseconds. For slower detectors, it is possible to resolve two particles close in time by analyzing the pulse shape of the measured signal and comparing it with the known response for a single particle. This technique can provide improved time separation, but may give poorer energy resolution for particles whose arrival times are within the duration of the output pulse.

24.7
Readout Devices

24.7.1
Optical Detectors

γ-ray detectors that produce secondary scintillation photons need a light-sensitive readout device, and a variety of options are available. Which device is used depends strongly on the type of detector (either totally absorbing or sampling), as well as on the type of detector material used. For scintillating materials, the most commonly used readout device is the photomultiplier tube. This device, shown schematically in Figure 24.10, consists of a photocathode that converts either UV or visible photons into electrons (or, more appropriately, photoelectrons), which are then amplified by a series of electron multipliers called *dynodes*. The photoelectrons are produced with an efficiency, known as the *quantum efficiency*, that is typically around 15–25%. The photoelectrons are then accelerated in an applied electric field and they impinge on the dynode, producing several outgoing electrons for each incident electron. This process is repeated many times in subsequent stages of the photomultiplier tube until the desired gain has been achieved and the resulting charge is collected on the anode electrode. Gains of 10^6–10^8 can be readily obtained with modern photomultiplier tubes, providing the possibility of detecting even single photoelectrons and, hence, single incident photons. Photomultiplier tubes are therefore one of the most sensitive types of readout devices used for detecting scintillation light. In addition, they can be designed to have a fast response time (typically a few nanoseconds) and can therefore be used in fast timing applications. Recently, photomultiplier tubes having segmented photocathodes with separate readouts have been developed. These provide an efficient, compact way of processing signals from segmented detector arrays. The disadvantages of photomultiplier tubes are that they require rather high voltages to operate (typically 1–2 kV) and that they do not work well inside a magnetic field.

Another type of optical detector is a light-sensitive photodiode. These can be

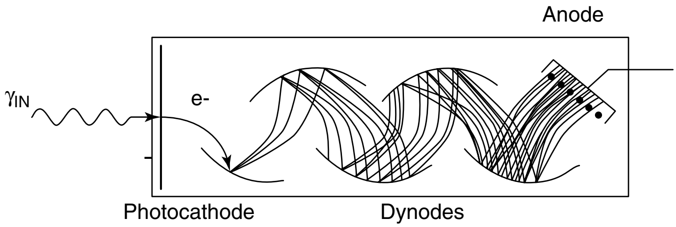

Fig. 24.10 Schematic representation of a photomultiplier tube.

either a silicon photodiode, such as those used in photocell detectors, or a vacuum photodiode, which has a construction similar to that of a photomultiplier tube having only a photocathode and anode electrode. An inherent property of these devices is that they have no intrinsic gain and therefore rely on the incident radiation to produce a sufficiently large signal to be measured directly. However, some devices, such as avalanche photodiodes, can achieve modest gains (10–100) by an amplification process within the silicon detector itself, which greatly improves their sensitivity to low light levels. The active areas of avalanche photodiodes have increased significantly in the last decade (e.g., 25 cm^2 position-sensitive units are currently available), and this has increased their use in many applications. Other types of vacuum devices, such as phototriodes or phototetrodes, are essentially low-gain photomultiplier tubes that can have large photocathode areas and can be used in other applications where modest gain is required.

24.7.2
Readout Electronics

Readout devices that have either no gain or low intrinsic gain usually require the use of a low-noise, charge-sensitive amplifier to detect the signal. The same is true for the detection of the direct ionization charge produced in noble liquids, such as those used in sampling calorimeters. In these cases, the primary signal can be on the order of only a few thousand electrons, and it is generally required to amplify this signal to the level of 10^7 electrons or more in order for it to be used for digitization and readout. This can be achieved with modern electronics through the use of low-noise amplifiers and signal-shaping techniques (Radeka, 1988). These amplifiers are used to integrate the primary charge produced in the detector and deliver a signal that has the proper pulse shape and timing characteristics for the readout system. It is also generally required to carry out this signal processing with a minimum amount of noise introduced by the electronics in order to maximize the signal-to-noise ratio. With a careful design, it is possible to achieve noise levels on the order of a few hundred electrons with signal-shaping times on the order of a few microseconds, or a few thousand electrons with shaping times on the order of a few nanoseconds. However, the choice of the appropriate signal-shaping time and noise considerations depend strongly on the specific experimental application.

Acknowledgments

We gratefully acknowledge additional contributors to the first edition of this chapter: Dr M.M. Lowry (Brookhaven

National Laboratory and Thomas Jefferson National Accelerator facility), Dr S. Hoblit (Brookhaven National Laboratory and the University of Virginia) and Dr M. Lucas (Brookhaven National Laboratory and Ohio University). We are indebted to Dr M. Riley of Florida State University for providing a version of Figure 24.2 suitable for inclusion here.

Glossary

Bremsstrahlung: γ-Ray emitted by a high-energy electron interacting with the electric field of a nucleus or atomic electron.

Calorimeter: Device that measures the total energy it absorbs.

Cascade Shower: The sequential energy-loss process in which γ-rays convert to electron–positron pairs, which then bremsstrahlung, emitting new gamma rays, and these, in turn, pair produce, and so on, until all the energy in the initial γ-ray is dissipated.

Cherenkov Radiation: Photons produced by a charged particle when its velocity in the medium exceeds that of the photons in the medium.

Compton Scattering: Elastic scattering of a γ-ray off an atomic electron.

Hermiticity: The degree to which a detector array is sensitive to all γ-rays, regardless of their directions or angles of emission from the source.

Pair Production: Transformation of a γ-ray interacting with the electric field of a nucleus or atomic electron into an electron–positron pair.

Photoelectric Absorption: Absorption of a γ-ray by an atomic electron.

Photomultiplier Tube: Device that generates a current pulse when struck by an optical photon, via a cascade multiplication of the secondary electron.

Spectral Quality: The fraction of γ–rays for which a measurement of their total energy is obtained.

References

Amaldi, U. (**1981**) *Phys. Scr.*, **23**, 409–24.

Awes, T., Obenshain, F.E., Plasil, F., Saini, S., Sorensen, S.P., and Young, G.R. (**1992**) *Nucl. Instrum. Methods Phys. Res. A*, **311**, 130–38.

Bethe, H.A. and Ashkin, J. (**1953**) *Experimental Nuclear Physics* (eds E. Segrè), Vol. **1**, John Wiley & Sons, Inc., New York, p. 166.

Beausang, C.W. (**2003**) *Nucl. Inst. Meth. Phys. Res.*, **B 204**, 666.

Beausang, C.W. and Simpson, J. (**1996**) *J. Phys.*, **G 22**, 527–28.

Brun, R. and Carminati, F. (**1993**) CERN Program Library Long Writeup W5013.

(DØ Collaboration) Abachi, S. *et al.* (**1994**) *Nucl. Instrum. Methods Phys. Res. A*, **338**, 185–253.

Dowell, D.H., Sandorfi, A.M., Baron, A.Q.R., Fineman, B.J., Kistner, O.C., Matone, G., Thorn, C.E., and Sealock, R.M. (**1990**) *Nucl. Instrum. Methods A*, **286**, 183–201.

Dowell, D.H., Ziegler, W.P., Sandorfi, A.M., Ziegler, B., and Schoch, B. (**1987**) *Nucl. Instrum. Methods A*, **254**, 570–77.

Eberth, J. and Simpson, J. (**2008**) *Prog. Part. Nucl. Phys.*, **60**, 283–337.

(Gammasphere) (**2008**) <*http://nucalf.physics.fsu.edu/~riley/gamma/*>.

Gelletly, W. and Eberth, J. (**2006**) *Lecture Notes in Physics*, Vol. **700**, Springer, pp. 79–117.

Hurlbut, C.R. (**1985**) Paper presented at the American Nuclear Society Winter Meeting, excerpted from the Bicron Catalog, Bicron, Inc, Newbury, OH.

Kleinknecht, K. (**1982**) *Phys. Rep.*, **84**, 85–161.

Knoll, G.F. (**1999**) *Radiation Detection and Measurement*, 3rd edn, John Wiley & Sons, Inc., New York.

Lee, I.Y., Clark, R.M., Cromaz, M., Deleplanque, M.A., Descovich, M., Diamond, R.M., Fallon, P., Macchiavelli, A.O., Stephens, F.S., and Ward, D. (**2004**) *Nucl. Phys.*, **A 746**, 255.

Lee, I.Y., Deleplanque, M.A., and Vetter, K. (**2003**) *Rep. Prog. Phys.*, **66**, 1095–144.

Lempicki, A., Wojtowicz, A.J., and Berman, E. (**1993**) *Nucl. Instrum. Methods A*, **333**, 304–11.

(LRP) (**2002**) Opportunities in Nuclear Science, A Long Range Plan for the Next Decade. *<http://www.sc.doe.gov/np/nsac/docs/LRP_5547_FINAL.pdf>*.

Nelson, W.R., Hirayama, H., and Rogers, D.W.O. (**1985**) *Report No.* 265. SLAC, Stanford, CA.

Kehoe W.L. (ed.) (PHENIX Collaboration, Adare, A., *et al.*) (**1993**) *BNLReport No.* 48922, Brookhaven National Laboratory, Upton, NY.

Radeka, V. (**1988**) *Annu. Rev. Nucl. Part. Sci.*, **38**, 217–77.

Scott, W.T. (**1963**) *Rev. Mod. Phys.*, **35**, 231–313.

(SLD Collaboration, Abe, Kenji *et al.*) (**1984**) *SLAC Report No.* 273, SLAC, Stanford CA; Haller, G.M., Fox, J.D., and Smith, S.R. (**1989**) *IEEE Trans.* **NS 36**, 675.

Tsai, Y.S. (**1974**) *Rev. Mod. Phys.*, **46**, 815–51.

Tsai, Y.S. (**1977**) *Rev. Mod. Phys.*, **49**, 421.

Further Reading

Evans, R.D. (**1955**) *The Atomic Nucleus*, McGraw-Hill, New York.

Fabjan, C.W. and Ludlam, T. (**1982**) *Annu. Rev. Nucl. Part. Sci.*, **32**, 335–89.

Gratta, G., Newman, H., and Zhu, R.Y. (**1994**) *Annu. Rev. Nucl. Part. Sci.*, **44**, 453–500.

Majewski, S. and Zorn, C. (**1992**) *Instrumentation in High Energy Physics*, (ed F. Sauli), World Scientific Publishing, Singapore, p. 157.

25
Spectrometers for Infrared Light

Yukihiro Ozaki and Shigeaki Morita

25.1 Introduction 867
25.2 General Principles of Infrared Spectrometers 869
25.2.1 Principles of Dispersive IR Spectrometers 870
25.2.1.1 Diffraction Grating 870
25.2.1.2 Monochromators 871
25.2.2 Principles of FT-IR Spectrometers 872
25.2.3 Advantages of FT-IR Spectrometers 877
25.2.3.1 Fellgett Advantage (Multiplex Advantage) 877
25.2.3.2 Jacquinot Advantage (Throughput Advantages) 877
25.2.3.3 Connes Advantage 877
25.2.4 Phase Correction 877
25.3 Components of FT-IR Spectrometers 878
25.3.1 Light Sources 878
25.3.2 Interferometers 878
25.3.3 Sample Compartments 879
25.3.4 Detectors 880
25.3.5 Beamsplitters and Mirrors 881
25.3.6 Amplifiers and AD Convertors 881
25.4 Various Types of FT-IR Interferometers 881
25.4.1 Continuous-Scanning Interferometers 881
25.4.2 Step-Scan Interferometers 882
25.4.3 Asynchronous Time-Resolved FT-IR Spectrometer 883
References 885
Further Reading 886

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

25.1
Introduction

Infrared (IR) spectroscopy is spectroscopy in the wavelength region of 0.78 μm–1 mm; the corresponding wavenumber region is 12 800–10 cm^{-1} (Griffiths and de Haseth, 2007; Christy, Ozaki and Gregoriou, 2001). The IR region lies between visible and microwave regions and is subdivided into three: near-infrared (NIR) (0.78–2.5 μm; 12 800–4000 cm^{-1}), mid-IR (2.5–25 μm; 4000–400 cm^{-1}), and far-IR (25 μm–1 mm; 400–10 cm^{-1}) regions. In this chapter, we are concerned only with mid-IR region.

IR spectroscopy provides detailed information about vibrations of a molecule (Griffiths and de Haseth, 2007; Christy, Ozaki and Gregoriou, 2001). Since molecular vibrations reflect the physical and chemical features of a molecule, such as the arrangement of nuclei and chemical bonds within the molecule, IR spectroscopy is very useful not only for identification of the molecule but also for study of the molecular structure. Furthermore, interaction of a molecule with the surrounding environment also causes a change in molecular vibrations, and hence, IR spectroscopy contributes considerably in studying the interaction, too. IR spectroscopy has many uses from basic research to various applications.

The history of IR spectroscopy stretches back to the nineteenth century (Sheppard, 2002). In the 1880s, IR spectra of many basic compounds were measured, and the first commercial IR spectrometer was put on the market as early as 1913. In the beginning of the twentieth century, the idea of group frequencies of IR absorptions for spectral analysis was established. Thus, it can be said that IR spectroscopy is very old, but it has always remained one of the most useful spectroscopic techniques ever in molecular science and analytical science.

When irradiated with IR radiation, a molecule absorbs it under some conditions. The energy $h\nu$ of the absorbed IR radiation is equal to the difference between two energy levels E_m and E_n of vibration of the molecule. In the form of an equation,

$$h\nu = E_n - E_m \qquad (25.1)$$

In other words, absorption of IR radiation occurs principally on the basis of a transition between energy levels of molecular vibration. This is why an IR absorption spectrum is a vibrational spectrum of a molecule.

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

Satisfying Eq. (25.1) is not always the only criterion for IR absorption to occur. There are transitions that are allowed by a selection rule (i.e., allowed transition) and those that are not allowed by the same rule (i.e., forbidden transition). In general, transitions with a change in the vibrational quantum number by ±1 are allowed and others are forbidden. This is known as a *selection rule* with respect to IR absorption. Another selection rule is one that is defined by the symmetry of a molecule. This selection rule is "IR radiation is absorbed when the transition dipole moment of a molecule changes as a whole in accordance with a molecular vibration."

Since most molecules are in the ground vibrational state at room temperature, a transition from the state $v'' = 0$ to the state $v'' = 1$ (first excited state) is possible. Absorption corresponding to this transition is called the *fundamental* band. Although most bands that are observed in IR absorption spectra arise from the fundamental, in some cases, we can find bands that correspond to transitions from the state $v'' = 0$ to the state $v'' = 2,3,4, \cdots\ldots$ (i.e., *overtone* transitions). However, since overtones are forbidden, overtone bands are very weak.

Figure 25.1 shows typical IR spectra in (a) a transmittance unit and (b) an absorbance unit. The abscissa and the ordinate of an IR absorption spectrum must now be explained. Frequency is indicated along the abscissa in units of wavenumber (cm^{-1}) (with higher wavenumbers always on the left-hand side), while *transmittance T* (%) (Figure 25.1a) or an *absorbance E* (Figure 25.1b) is expressed along the ordinate. While IR spectra include reflectance spectra and emission spectra in addition to absorption spectra, reflectance, emission intensity, and so on are expressed along the ordinate in the case of a reflectance spectrum, an emission spectrum, and so on.

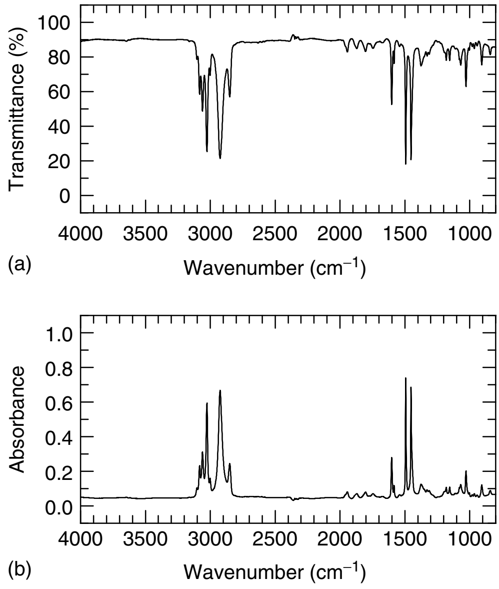

Fig. 25.1 Examples of IR spectra: (a) transmittance spectrum and (b) absorbance spectrum. (Reproduced from Christy, Ozaki and Gregoriou (2001) with permission. Copyright (2001) Elsevier.)

25.2
General Principles of Infrared Spectrometers

IR spectrometers have a long history (Sheppard, 2002). They have made remarkable progress after the Second World War. There were three revolutions in the history of IR spectrometers after the war. The first one, which occurred just after the war because of various technical developments during the war, was concerned with the development of dispersive IR spectrometers using prisms as dispersing elements. During the 1950s, diffraction gratings, detectors, and optical elements developed rapidly, and in 1960s, high-performance dispersive spectrometers with diffraction gratings became widely available commercially. This was the second revolution. The third and the biggest revolution was the advent of Fourier-transform (FT) spectrometers in the 1970s. Indeed, the advances in the field of FT-IR spectroscopy in the past 30 years have been quite remarkable.

In this chapter, we do not describe the detailed history of FT-IR spectrometers, for which many good references are available (Griffiths and de Haseth, 1986, 2007; Ferraro and Basile, 1990; Sheppard, 2002; Kauppinen and Hollberg). However, three important historical events need special mention. In 1952, Fellgett pointed out that, in principle, interferometry also provides a multiplex advantage using broadband sources in that an interferogram records the superimposed signals from all the different wavelengths presented to the detector (Sheppard, 2002). In 1958, Jacquinot emphasized the additional energy throughput advantage of interferometers over dispersive spectrometers on the basis of their circular aperture in comparison with the entrance slits of dispersive spectrometers. Another breakthrough came from the discovery of the *fast Fourier transform* (FFT) by Cooley and Tukey in 1964. The FFT is superior to the discrete FT.

Nowadays, there are two types of spectrometers used to measure IR spectra, namely, dispersive IR spectrometer and FT-IR spectrometer, although the latter is much more popular than the former. Dispersive IR spectrometer is a tool that disperses IR radiation spatially into its constituent wavelengths with a diffraction grating, a prism, or the like. Meanwhile, FT-IR spectrometer utilizes interference of radiation for the purpose of spectral separation of radiation. As each one of the two methods has its own pros and cons, no one can categorically conclude that one is superior to the other. However, FT-IR spectrometer has recently become a principal method for the following reasons:

1. It permits one to measure spectra simultaneously in the entire wavenumber region.
2. Accuracy of the wavenumbers of spectra is extremely high.
3. It secures a constant and high spectral resolution across the whole spectral region.
4. The amount of radiation reaching a detector is high as it utilizes radiation at a very high efficiency.

While FT-IR spectrometer will certainly remain a prominent IR spectroscopic method at least for the time being, dispersive IR spectroscopy could well attract a lot of attention in the future, again if multichannel detectors operable in the IR region advance and improve further. Noting that many researchers today chose FT-IR spectroscopy in almost

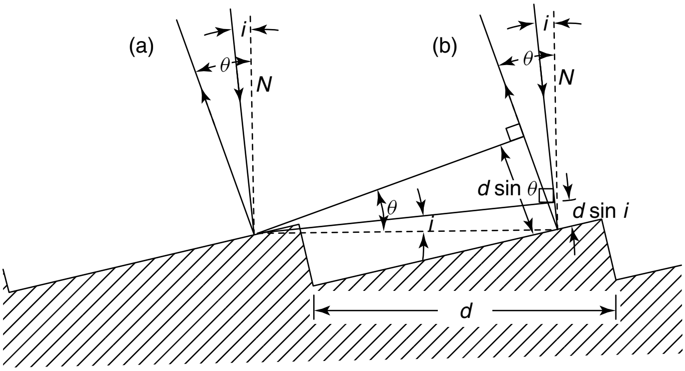

Fig. 25.2 An echelette grating, a typical reflection grating used in a dispersive IR monochromator. (Reproduced from Kauppinen *et al.* (1997) with permission. Copyright (1997) Wiley-VCH.)

all of their research activities that use IR spectroscopy, we discuss FT-IR spectrometer more in detail than dispersive IR spectrometer in this chapter.

25.2.1
Principles of Dispersive IR Spectrometers

A typical IR spectrometer consists of a radiation source, a wavelength selector, a sample compartment, a detection system, and an electronic and computer unit (Kauppinen and Hollberg; Griffiths and de Haseth, 1986, 2007; Christy, Ozaki and Gregoriou, 2001). In dispersive IR spectrometers, prisms and gratings are used as wavelength selectors. They measure a single wavelength at a time, that is, they scan over the wavelength region so that the wavelength through the selector is changed by rotating the prism or grating.

In the following part, we focus on grating spectrometer because prism spectrometer is rarely used in recent years.

25.2.1.1 **Diffraction Grating**

A grating acts like a multislit source when collimated radiation strikes it. Different wavelengths are diffracted and constructively interfaced at different angles. There are two types of diffraction gratings: transmission and reflection gratings. In grating IR monochromators, plane reflection gratings are used because transmission gratings have very poor transmission in the IR region.

Figure 25.2 shows an echelette grating, a typical reflection grating used in a dispersive IR monochromator. It has alternating reflecting and nonreflecting areas called *facets* or *lines*. The constructive interference condition for a reflection diffraction grating with the grating constant d is given by

$$k\lambda = d(\sin i + \sin\theta),$$
$$k = 0, \pm 1, \pm 2, \ldots \qquad (25.2)$$

where i is the angle between the incident radiation and the normal N of the grating (incident angle), θ is the angle between the diffracted radiation and the normal N (diffracted angle), and k is the order of diffraction.

Angular dispersion, D, which represents a variation in the diffracted angle $d\theta$ against a variation in wavelength ($d\lambda$)

is given by

$$D = \frac{d\beta}{d\lambda} = \frac{k}{d\cos\beta} = \frac{\sin i + \sin\beta}{\lambda\cos\beta} \tag{25.3}$$

The resolving power, *R*, is yielded by

$$R = \frac{\nu}{\Delta\nu} = \frac{\lambda}{\Delta\lambda} = \frac{k\pi}{\pi/N} = kN = hD \tag{25.4}$$

where, h is an order of diffraction and N is a total number of facets of a dispersive grating. $\Delta\lambda$ and $\Delta\nu$ are defined by a line width that becomes zero first from the center of a spectral line (Rayleigh's criteria).

In an echelette grating when an incident angle and a diffracted angle satisfy the following relation

$$2\alpha = i + \theta \tag{25.5}$$

where α is an angle between the facet plane and the line plane, called *blaze angle*, the intensity of diffracted radiation becomes strong. For the radiation that illuminates and diffracts with the normal of facet ($i = \theta = \alpha$), one can obtain the following relation:

$$k\lambda = 2d\sin\alpha \tag{25.6}$$

when $k = 1$ and λ_b is called the *blaze wavelength*. This is the wavelength with $k = 1$ where the diffracted radiation becomes the strongest for the grating.

25.2.1.2 **Monochromators**

Figure 25.3 depicts a typical grating monochromator used in an IR spectrometer (Kauppinen and Hollberg). By rotating the grating G or by changing the angle δ, one can scan the wavelength. For a given δ, a certain wavelength is allowed to go through the slit S_2 onto the detector D.

To judge the quality of IR monochromators, optical slit width (ΔV) and mechanical slit width (w) are important. The relation between the two slit widths is given by

$$(\Delta V)^2 = (\Delta V_1)^2 + (\Delta V_0)^2 \tag{25.7}$$

$$\Delta V_1 = \frac{V^2 wd}{kf}\cos i, \quad \Delta V_0 = \frac{V}{Nk}$$

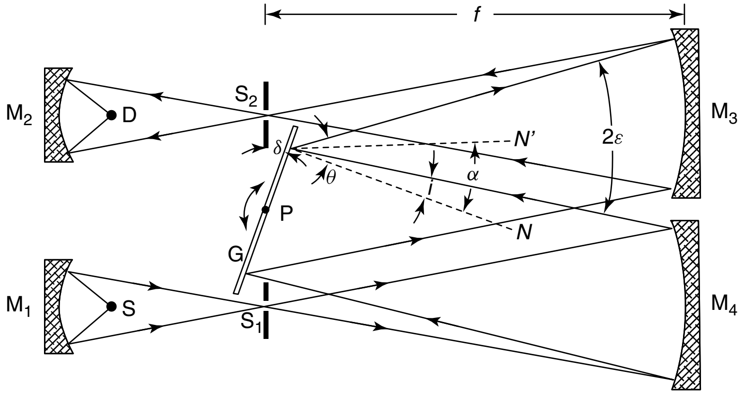

Fig. 25.3 A typical grating monochromator used in an IR spectrometer. (Reproduced from Kauppinen and Hollberg (1997) with permission. Copyright (1997) Wiley-VCH.)

where ΔV_1 and ΔV_0 are determined by a finite slit width and an infinite slit width, respectively, and f is a focal length of a collimator lens.

25.2.2
Principles of FT-IR Spectrometers

The design of many interferometers used for IR spectrometers today is based on that of the two-beam interferometers originally designed by Michelson in 1891 (Griffiths and de Haseth, 1986, 2007). Let us take a look at the principles of FT-IR spectrometers with reference to a Michelson interferometer. It comprises of a half mirror (beamsplitter) and two mutually perpendicular plane mirrors, as shown in Figure 25.4. The beamsplitter is an optical element that separates the incident ray into transmitted ray and reflected ray. Of the two plane mirrors, one is fixed (fixed mirror) and the other is movable parallel to an optical axis (movable mirror). A collimator lens shapes the rays from a light source into parallel rays, which will then impinge upon the beamsplitter. The beamsplitter separates the rays into those heading for the fixed mirror and those heading for the movable mirror (Figure 25.4). As the two rays return to the beamsplitter after being reflected, respectively, by the fixed mirror and the movable mirror, the two rays interfere with each other at the beamsplitter. The intensity of the interference wave is dependent upon how different the distance between the beamsplitter and the fixed mirror is from the distance between the beamsplitter and the movable mirror.

Let us now consider the case when monochromatic radiation is incident upon the beamsplitter. Where the fixed mirror and the movable mirror are equidistant from the beamsplitter, the optical path difference between the two rays is zero, and therefore, the two rays intensify each other. Assuming now that the movable mirror has moved by $\lambda/4$ from its position when the optical path difference is zero, the optical path difference between the two rays becomes $\lambda/2$ and the two rays weaken each other. As the movable mirror further moves by $\lambda/4$, the optical path difference becomes λ, whereby the two rays intensify each other again. The way the two rays

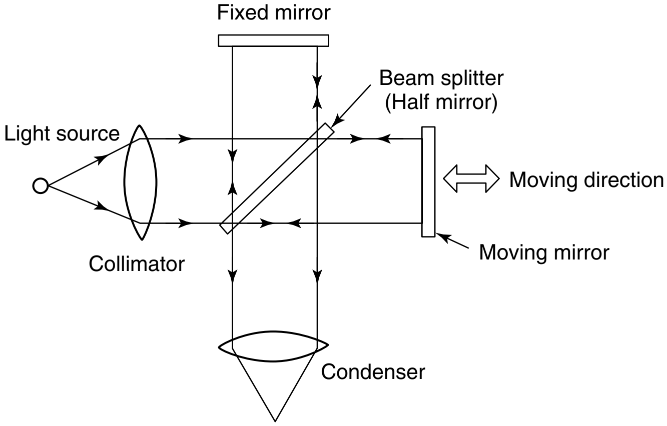

Fig. 25.4 Optical layout of a Michelson interferometer.

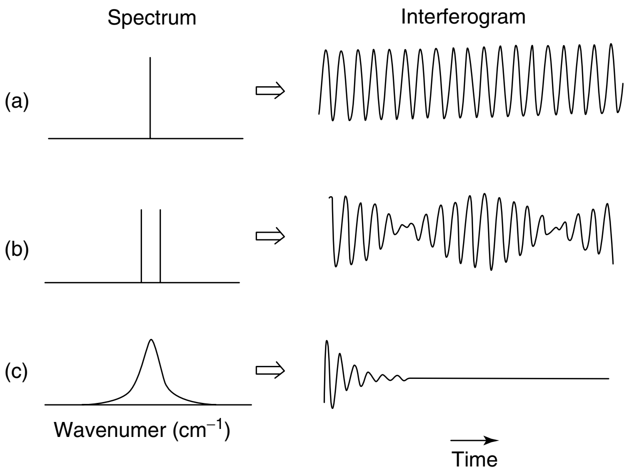

Fig. 25.5 Simple light source spectra and the corresponding interferograms from an interferometer: (a) single infinitesimally narrow line; (b) two infinitesimally narrow lines of equal intensity; (c) continuous light source.

interfere with each other thus changes as the movable mirror moves. In general, two rays intensify each other when their optical path difference is $n\lambda$ (where n is an integer) but weaken each other when their optical path difference is $(n + 1/2)\lambda$. The intensity I of radiation that we observe is expressed as a function of the optical path difference x:

$$I(x) = 0.5P(\lambda)\left(1 + \cos\frac{2\pi x}{\lambda}\right) \quad (25.8)$$

where the $P(\lambda)$ indicates the intensity of the light source. Assuming that the speed of the movable mirror is v and the time is t, x can be expressed as

$$x = 2vt \quad (25.9)$$

Substituting formula (25.9) in formula (25.8) and replacing the wavelength λ with the wavenumber $\tilde{\nu}$, we obtain

$$I(t) = 0.5P(\tilde{\nu})(1 + \cos 4\pi v\tilde{\nu}t) \quad (25.10)$$

Formula (25.10) indicates that the intensity I of radiation consists of a direct current component $0.5P(\tilde{\nu})$ and an alternating current component $0.5P(\tilde{\nu})\cos 4\pi v\tilde{\nu}t$. The alternating current component indicates that radiation of the wavenumber $\tilde{\nu}$ has been amplitude-modulated to a frequency $2v\tilde{\nu}$. If translated into a drawing, the second term in formula (25.10) is as shown in the right-hand side section in Figure 25.5(a).

The component $0.5P(\tilde{\nu})$ in formula (25.10) is what we could obtain if our interferometer is ideal. In reality, it is $0.5P(\tilde{\nu})$ times $H(\tilde{\nu})$, which is the efficiency of the beamsplitter. Let us therefore decide that $0.5P(\tilde{\nu})H(\tilde{\nu}) \equiv B(\tilde{\nu})$.

Let us now consider what happens when two rays of monochromatic radiation that have the wavenumbers $\tilde{\nu}_1$ and $\tilde{\nu}_2$ impinge upon an interferometer at the same time. In this instance, the respective rays are modulated to frequencies $2v\tilde{\nu}_1$ and $2v\tilde{\nu}_2$ and a signal indicative of the sum of them is obtained as an output (Figure 25.5b).

Where continuous ray is incident, the intensity of outgoing ray can be expressed as a function of an optical path difference.

$$I(x) = \int_0^{\infty} B(\tilde{\nu})\ (1 + \cos 2\pi\tilde{\nu}x)\ d\tilde{\nu} \tag{25.11}$$

The second term in this formula, namely, $F(x) = \int_0^{\infty} B(\tilde{\nu}) \cos 2\pi\tilde{\nu}x d\tilde{\nu}$ is an *interferogram*. As an output corresponding to each wavenumber is expressed as a cosine function as shown in Figure 25.5(a), the sum of these outputs $F(x)$ takes a maximum value at the position of $x = 0$. When x shifts from $x = 0$, the outputs corresponding to the respective wavenumbers interfere with each other in various phases, whereby $F(x)$ rapidly becomes weak while winding. Further, an interferogram has a symmetrical profile as $F(x)$ is an even function (i.e., $F(x)$ does not change even when x is replaced with $-x$).

Introducing functions $Be(\tilde{\nu}) = B(\tilde{\nu})/2$ ($\tilde{\nu} \geq 0$) and $Be(-\tilde{\nu}) = Be(\tilde{\nu})$, we obtain

$$F(x) = \int_{-\infty}^{\infty} Be(\tilde{\nu}) \cos 2\pi\tilde{\nu}x d\tilde{\nu} \tag{25.12}$$

Hence,

$$Be(\tilde{\nu}) = \int_{-\infty}^{\infty} F(x) \cos 2\pi\tilde{\nu}x dx \tag{25.13}$$

Formulae (25.12) and (25.13) form a famous FT pair. In short, the interferogram $F(x)$ is obtained through Fourier transformation of $Be(\tilde{\nu})$ and the spectrum $Be(\tilde{\nu})$ is obtained through Fourier transformation of the interferogram $F(x)$.

While FT-IR thus requires measuring an interferogram first and Fourier transformation of the interferogram on a computer thereafter for rendering of a spectrum, an actual spectroscopic system is inclusive of a few more complex processes. First, although integration in formula (25.13) is from $-\infty$ to $+\infty$ of an optical path difference x, this is not realistic. An *apodizing function* is therefore introduced for integration within a finite range after measurement of $I(x)$ while moving the movable mirror within a finite range.

Figure 25.6 shows typical examples of apodizing functions. We introduce an apodizing function $A(x)$ whose profile is shaped as a rectangular wave,

$$A(x) = \begin{cases} 1 & (|x| \leq L) \\ 0 & (|x| > L) \end{cases} \tag{25.14}$$

and we obtain the following definite integration to replace formula (25.13):

$$Be'(\tilde{\nu}) = \int_{-\infty}^{\infty} A(x)F(x) \cos 2\pi\tilde{\nu}x dx \tag{25.15}$$

Since $A(x)$ has a finite value only within the range of $|x| \leq L$, $A(x)$ limits the range of integration. $Be'(\tilde{\nu}) = \widetilde{A(x)F(x)}$ (where the symbol $\sim$ is used to denote Fourier transformation) is a convolution of $\widetilde{A(x)}$ and $\widetilde{F(x)}$ ($= Be'(\tilde{\nu})$) according to the theorem of convolution.

$$Be'(\tilde{\nu}) = \widetilde{A(x)} * \widetilde{F(x)} = \widetilde{A(x)} * Be(\tilde{\nu}) \tag{25.16}$$

Formula (25.16) indicates that the resultant spectrum is multiplication of spectrum $Be'(\tilde{\nu})$ with an instrument function $\widetilde{A(x)}$. It is therefore very important to know the profile of the instrument function $\widetilde{A(x)}$. Calculating $\widetilde{A(x)}$, we obtain

$$\widetilde{A(x)} = \frac{\sin(2\pi L\tilde{\nu})}{\pi\tilde{\nu}} \tag{25.17}$$

This function is similar to the function shown in Figure 25.6(a). The half width

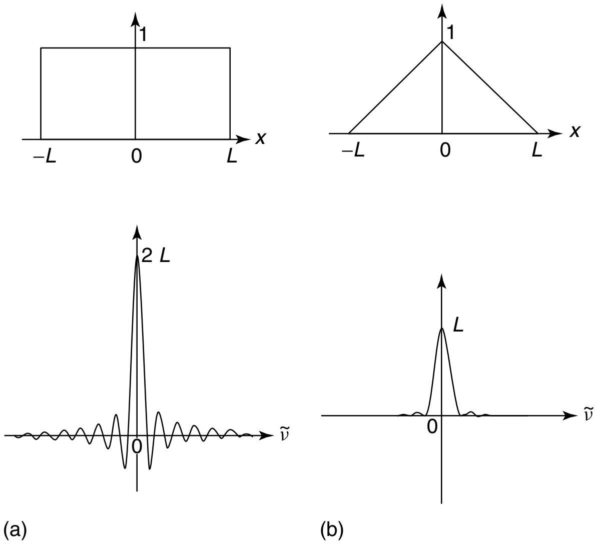

Fig. 25.6 Apodization functions and their corresponding instrument lineshape functions: (a) boxcar truncation and (b) triangular apodization.

of this function has a value $0.6/L$. As a half width determines the spectral resolution, the longer the distance L through which the movable mirror moves, the better the spectral resolution becomes. An instrument function, as that shown in Figure 25.6(a), has a major disadvantage that it includes associated ripples at which it yields positive and negative values. Using the following function $A'(x)$ (Figure 25.6b), we have

$$A'(x) = \begin{cases} 1 - |x/L| & (|x| \le L) \\ 0 & (|x| > L) \end{cases} \tag{25.18}$$

as an apodizing function instead of the function shown in Figure 25.6(a); the instrument function becomes as shown in Figure 25.6(b). In the case of this instrument function, although the half width is $0.88/L$ and the spectral resolution is slightly inferior to what it is in Figure 25.6(a), problems of an associated ripple, a rebound toward the negative side, and the like are suppressed better. Several apodizing functions are implemented in FT-IR spectrometers that are available in the market. One needs to choose a suitable apodizing function in accordance with their purpose.

The second issue on an actual spectrometer is that an interferogram is not an even function but is asymmetric in reality for various reasons. A process called *phase correction* is necessary to solve this.

The flow of an FT-IR signal is described here. Figure 25.7 shows the structure of a spectroscopic unit of an FT-IR spectrometer and the behaviors of the signal within components that form the spectrometer. Radiation from a light source (denoted in [1] in Figure 25.7) impinges upon an interferometer via an incident hole. At the interferometer, the radiation is modulated into frequencies that are

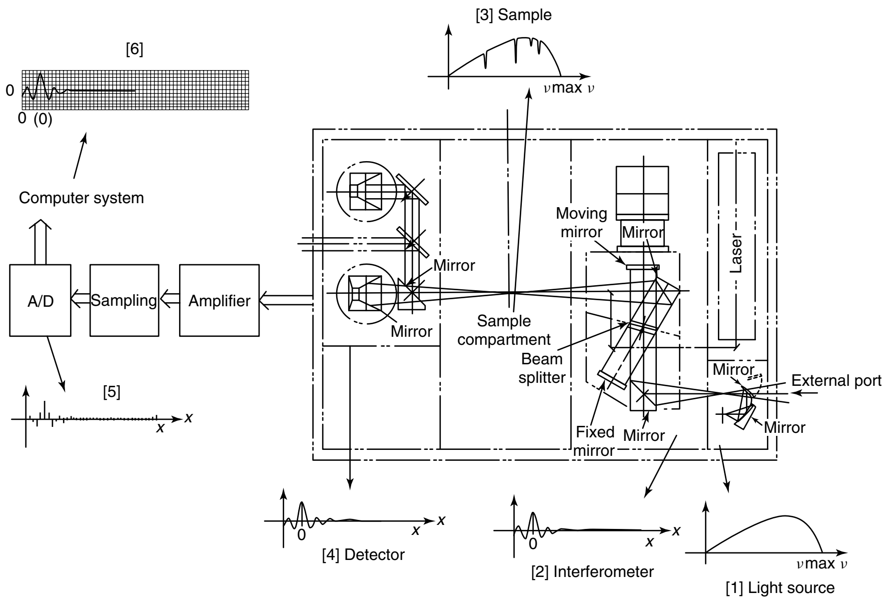
[3] Sample
νmax ν
[6]
0
0 (0)
Computer system
A/D
Sampling
Amplifier
[5]
x
x
Moving mirror
Mirror
Mirror
Mirror
Sample compartment
Beam splitter
Laser
Mirror
External port
Mirror
Fixed mirror
Mirror
0
x
x
[4] Detector
0
x
x
[2] Interferometer
νmax ν
[1] Light source

Fig. 25.7 Optical layout of a typical FT-IR spectrometer and a series of FT-IR signals.

proportional to the respective wavenumbers (denoted in [2] in Figure 25.7).

Leaving the interferometer, the radiation impinges upon a measurement sample that absorbs radiation whose wavenumbers are unique to the sample. Section [3] in Figure 25.7 illustrates absorption of radiation having the characteristic wavenumbers by the sample (Section [3] in Figure 25.7 shows the signal as a spectrum for the simplicity of illustration although an actual signal has the shape of an interferogram). A detector converts the radiation from the sample into an electric signal (the section [4] in Figure 25.7). After being amplified by an amplifier, the electric signal is converted into a digital signal by an A/D converter via a sample hold circuit (which samples the signal) (Section [5] in Figure 25.7). The AD-converted interferogram is fed a set number of times to a computer and accumulated (Section [6] in Figure 25.7). Thus, the accumulated interferogram has the shape of a Fourier-transformed spectrum as shown in Figure 25.8(a).

Fourier spectroscopy is bright because it requires simultaneous capturing of incident radiation and simultaneous measurement of the spectrum in the entire wavenumber region. As the spectrum thus measured is inclusive of the emission spectral characteristic of a light source, the spectroscopic characteristics of an interferometer and a detector and the like, without any processing, it is necessary to measure a background spectrum (Figure 25.8b) and divide the observed spectrum by the background spectrum (Figure 25.8c). The resultant spectrum is a transmission spectrum $T(\tilde{\nu})$. We can calculate an absorbance spectrum $E(\tilde{\nu})$ (Figure 25.8d) using the following formula:

$$E(\tilde{\nu}) = -\log_{10} T(\tilde{\nu}) \tag{25.19}$$

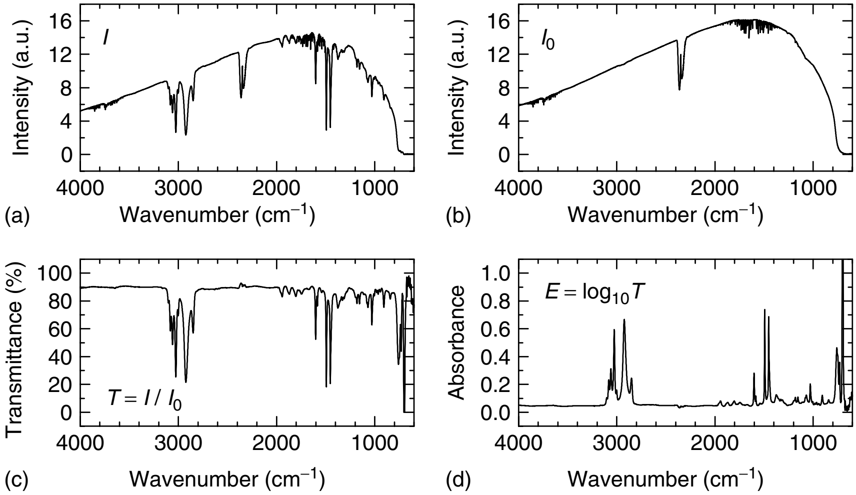

Fig. 25.8 FT-IR spectra: (a) transmission spectrum (before the division by background; (b) background spectrum; (c) transmission spectrum after the division; and (d) absorbance spectrum.

25.2.3
Advantages of FT-IR Spectrometers

It is worth summarizing the three major advantages of FT-IR spectrometers.

25.2.3.1 Fellgett Advantage (Multiplex Advantage)

The biggest advantage of FT-IR spectrometer is the simultaneous detection of the whole spectrum at once. Even though a factor of 2 in signal strength is lost because half the beam is reflected back to the source, the multichannel advantage is nevertheless 10^4 or higher. That is, theoretically, an interferometer can achieve a signal-to-noise (S/N) ratio comparable to that of a dispersive monochrometer 10^4 times faster.

25.2.3.2 Jacquinot Advantage (Throughput Advantages)

Total radiation power on a detector is much higher in a Michelson interferometer than in a grating instrument because at the optimum resolution the area of a circular aperture (the Jacquinot stop) in an FT-IR spectrometer is about 200 times larger than a slit in a grating spectrometer.

25.2.3.3 Connes Advantage

Connes advantage arises from the ability of interferometry using a monochromatic source to accurately and precisely index the retardation, resulting in a superior determination of the retardation sampling position.

25.2.4
Phase Correction

An interferogram measured in the real world, however, has an asymmetric shape because of various reasons although it is supposed to be symmetric with respect to a position of zero optical path difference, and it is therefore impossible to obtain an expected spectrum through mere Fourier cosine transformation of the interferogram (Griffiths and de Haseth, 1986, 2007; Christy, Ozaki and Gregoriou, 2001). The causes of the asymmetric shape include those attributable to a thickness difference between a half mirror and a compensator plate of a beamsplitter, those attributable to the frequency characteristics of an electric circuit system, those attributable to a deviation of the optical path of the main interferometer and that of the associated (laser) interferometer, and so on. Thus, we need what is called *phase correction*. Phase correction is the correction of an asymmetric interferogram into a symmetric interferogram. Equation (25.20) shows the modified relation for the intensity of the interferogram.

$$I(x) = \int_{-\infty}^{\infty} B(\tilde{\nu})\,[1 + \cos(2\pi\tilde{\nu}x + \Phi_{\mathrm{BS}}(\tilde{\nu}))]\,d\tilde{\nu} \qquad (25.20)$$

where $\Phi_{\mathrm{BS}}(\tilde{\nu})$ is the wavelength-dependent phase shift introduced by the beamsplitter.

Phase correction is the mathematical procedure to remove the sine components from the interferogram. There are two popular phase correction algorithms used in single-sided interferograms: the Mertz algorithm and the *Forman* algorithm. In the former, the largest data point in the interferogram is assigned as the zero retardation point and the amplitude spectrum is calculated with respect to this point. A short double-sided interferogram is measured and its corresponding phase array is used to phase correct the entire single-sided spectrum. The *Forman* correction is essentially equivalent to the Mertz routine but it is performed in the retardation space. Besides phase effects, one must consider the effects of beam divergence, of mirror misalignment, and poor mirror drive for the performance of FT spectrometers (Griffiths and de Haseth, 2007).

25.3
Components of FT-IR Spectrometers

The basic components of dispersive spectrometers and FT spectrometers are the same with the exception of the interferometer (Kauppinen and Hollberg; Chalmers and Griffiths, 2002; Griffiths and de Haseth, 2007; Christy, Ozaki and Gregoriou, 2001). Any differences are within the elements and are the result of the different ways in which the source radiation is detected. For example, sensitive thermocouple detectors are commonplace in dispersive instruments, whereas they are not appropriate for rapid-scanning instruments.

25.3.1
Light Sources

Requirements for an IR light source are that the spectrum distribution of the light source is close to that of black body radiation, the light source offers stable emission for long periods of time, the lifetime of the light source is long, and so on. A globar lamp, Nernst heat generator, a nichrome heat generator, and the like have often been used; however, nowadays, ceramic sources have become more popular.

25.3.2
Interferometers

As shown in Figure 25.9, an interferometer of an FT-IR spectrometer is formed by a main interferometer and an associated

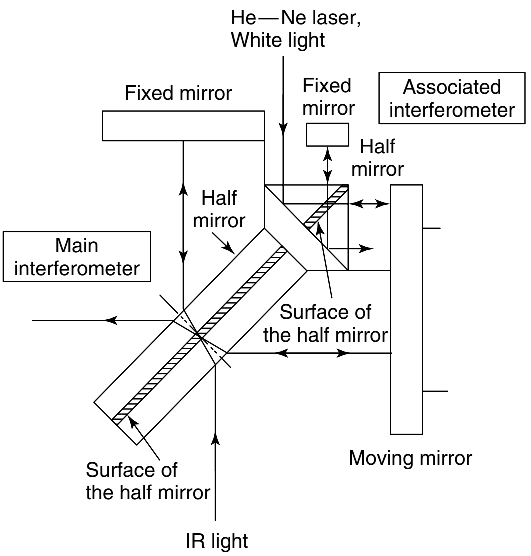

Fig. 25.9 Main and attached interferometers.

interferometer. The associated interferometer is attached for monitoring of an optical path difference x of the main interferometer. Like the main interferometer, the associated interferometer is formed by a half mirror, a fixed mirror, and a movable mirror (which is shared by the main interferometer). The associated interferometer accepts both white light and He–Ne laser radiation and outputs interferometer output signals corresponding to the two types of radiations. The interferometer output signal resulting from the white light is used to denote when to start sampling the interferogram that the main interferometer yields. Meanwhile, the interferometer output signal resulting from the laser radiation is used to monitor the location of the movable mirror and instruct sampling of the interferogram, which is obtained by the main interferometer.

Spectra measured by FT-IR spectrometers have extremely high (on the order of 7 or more digits) wavenumber accuracy. This is because a Michelson interferometer directly reads the frequency of incident radiation while referring to an optical path difference determined by the emission wavelength of the He–Ne laser whose accuracy is on the order of 7 or more digits.

25.3.3
Sample Compartments

Sample compartments of commercial FT-IR spectrometers are quite different from each other depending upon models. We must pay extra attention to the shape of a ray, to where the ray converges, and to a method for measuring the ray. Cross section of the ray is usually circular and converges in the central part of the sample compartment (when the sample compartment is of the center-focus type) or on a wall surface of the sample compartment (when the sample compartment is of the side-focus type) or remains as a parallel beam without converging. We need to adequately utilize the size and the shape of a ray to enjoy the full capability of an FT-IR spectrometer. When setting a cell or accessory to a sample compartment, therefore, we must make sure that the cell or accessory will not block the ray as much as possible.

FT-IR measurement methods include those of a single-beam type and those of a double-beam type. The former is a method for measuring spectra with a measurement sample and a reference sample set alternatively on the same optical path, and is superior in terms of ordinate accuracy because of the use of the same optical path. When we use a single-beam method, we must be careful about any change in measurement environment with time. This is because any change in measurement environment with time will deteriorate the reproducibility of an ordinate. When choosing a double-beam method, we set a measurement sample and a reference sample on different optical paths at the same time and measure the samples alternately for a designated number of integration times. While the measurement environments are considered to be approximately the same, a difference with respect to the optical capabilities between the two optical paths, if any, will give rise to a problem with ordinate accuracy. The two methods thus each have their own pros and cons. This has led to the development of what is known as a *sample switching double-beam method* that enjoys the advantages of the two methods. Where the sample switching double-beam method is used, measurements are taken principally by a single-beam method while a measurement sample and a reference sample are set simultaneously on separate optical paths. After measurements are taken up to a designated number of integration times, an automatic switch mechanism switches the measurement sample and the reference sample and measurements are taken again. Resultant spectra are extremely reliable as the measurement sample and the reference sample are measured on the same optical path in the same environment. The air-tightness of the environment against external air is always critical to any measurement. IR absorption by carbon dioxide, vapor, and the like contained in the atmosphere is strong, and therefore, no matter how strong the spectral correction using a reference sample may be, the influence of the IR absorption will remain persistently. It is important to close the sample compartment airtight, fill the sample compartment with nitrogen gas, dry air, or the like and take the measurement with vapor and carbon dioxide sufficiently removed.

25.3.4
Detectors

Requirements for an FT-IR detector include, among others, that (i) the sensitivity of the detector be high; (ii) the detector remain sensitive in a wide wavenumber region; (iii) a radiation-receiving surface of the detector be large to a certain extent (so as to leverage on the "an optical system is bright" is a characteristic of FT-IR); (iv) the detector be sensitive to a wide frequency region; and (v) the detector exhibit a linear response to the intensity of incident radiation. IR detectors can be divided into two types: thermal detectors and quantum detectors. The former detectors sense the variation in temperature of a detector material due to absorption of the IR radiation. In quantum detectors, incident IR photons result in the production of free charge carriers in the responsive element. No serious temperature change in the element takes place during this process. TGS (tryglycine sulfate) detectors and MCT ($Hg_{1-x}Cd_xTe$; $x \sim 0.2$) detectors are often used in FT-IR spectrometers. The former detectors, utilizing the pyroelectric effect, are characterized by their relative inexpensiveness and sensitivity to a wide IR region. The latter detectors are semiconductor detectors that utilize the photovoltaic power caused by optical absorption by semiconductors or the photoconductivity effect, and are more sensitive than TGS detectors.

25.3.5
Beamsplitters and Mirrors

A typical beamsplitter IR instrument is of the germanium (Ge)-on-potassium bromide (KBr) substrate type. Recently, semiconductor film beamsplitters, which are formed from self-supporting semiconductors, including carbon films, have been described. Preferably, the beamsplitters are formed from silicon, germanium, or diamond films.

25.3.6
Amplifiers and AD Convertors

An interferogram converted into an electric signal by a detector is fed to a computer via a preamplifier, a main amplifier, an electric filter, a sample-hold circuit, and an analog-to-digital (A/D) convertor.

The amplifiers amplify the interferogram up to the input level of the A/D convertor without distortion. The electric filter removes unwanted signals belonging to a high-frequency region and noises that arise from sampling. These electric systems are designed so that their noises are smaller than those caused by the detector, permitting the detector to exhibit its capability adequately.

The A/D convertor is one of the factors that restrict the S/N ratio of FT-IR. Where a high-sensitivity detector such as an MCT is used for measurement, the S/N ratio of an incoming interferogram to the A/D convertor may be significantly high and exceed the dynamic range (the solution power to convert a signal) of the A/D convertor in some instances.

25.4
Various Types of FT-IR Interferometers

25.4.1
Continuous-Scanning Interferometers

Most commercially available FT-IR spectrometers use the continuous-scan mode of operation, where the movable mirror scans at constant velocity. This type of scanning works very well for routine measurements. Also, the fact that continuous-scanning interferometers can measure an interferogram in a time of 1 second or less makes them excellent tools to follow time-dependent processes. In the continuous-scan mode of interferometry, the laser fringe counter is used to sense the accuracy of the scanning velocity. If a deviation is sensed, correction signals are generated that assure proper operation (constant velocity). The consequence of this mode of operation is that each IR wavelength (λ) is modulated at its own particular Fourier frequency, given by Eq. (25.21):

$$f(\lambda) = \frac{2v}{\lambda} \tag{25.21}$$

where v is the mirror velocity.

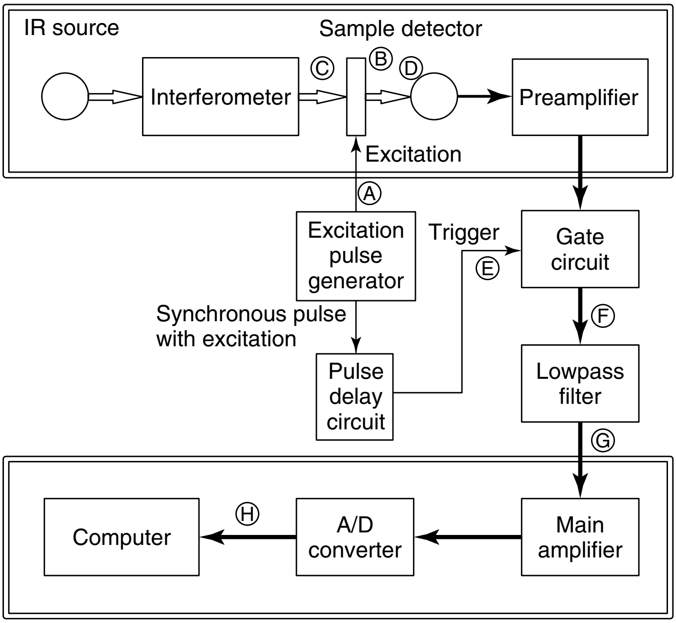
IR source
Sample detector
Interferometer
Ⓒ
Ⓑ
Ⓓ
Preamplifier
Excitation
Ⓐ
Excitation pulse generator
Trigger
Ⓔ
Gate circuit
Synchronous pulse with excitation
Ⓕ
Pulse delay circuit
Lowpass filter
Ⓖ
Computer
Ⓗ
A/D converter
Main amplifier

Fig. 25.10 A schematic diagram for the initial type of the asynchronous time-resolved measurement. The signal-processing assembly for time-resolved measurements is shown in the middle, while the units constituting a conventional FT-IR spectrophotometer are depicted in doubly lined boxes at the top and bottom. (Reproduced from Masutani *et al.* with permission. Copyright (1992) Society for Applied Spectroscopy.)

25.4.2
Step-Scan Interferometers

Step-scan interferometers are powerful for time-resolved FT-IR measurements (Griffiths and de Haseth, 2007; Christy, Ozaki and Gregoriou, 2001). By using a step-scan FT-IR spectrometer, one can monitor much faster reactions if a reaction or sample perturbation is repeatable. In the step-scan interferometer, the optical path difference is held constant over the course of the reaction or perturbation. The optical path difference is then changed, typically by one wavelength of the He–Ne laser, and the process is repeated. This procedure is reiterated at constant intervals of retardation until the optical path difference is sufficiently high that the desired spectroscopic resolution has been reached.

In step-scan FT-IR data are collected while the retardation is held constant or is oscillated about a fixed value. Therefore, to apply the technique to mid-IR and shorter wavelength measurements, a method for controlling the retardation and implementing a special sampling rate comparable to that achieved in modern continuous-scan instruments is required. In recent years, different control methods have been reported. However, all of these basically rely on the use of the He–Ne laser fringe pattern to generate the control signal and to determine the stop size. The biggest advantage of the

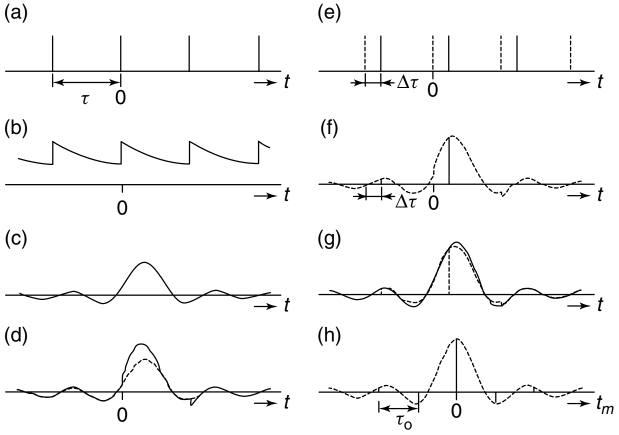

Fig. 25.11 Shape of the signals in processes (a)–(h): (a) Excitation pulses with period τ; (b) responses of the sample to the excitations in (a); (c) ordinary interferogram without excitations; (d) interferogram modulated by the excitations in (a) (full curve) and the unmodulated interferogram (broken curve, same as in (c)); (e) trigger pulses (full line) delayed from the preceding excitation pulses (broken line) by a fixed time interval *Dt*; (f) discrete interferogram time resolved at time delay *Dt* from the excitation (full line) and the modulated analog interferogram (broken curve, the same one as in (d)); (g) time-resolved analog interferogram obtained from the low-pass filter (full curve), the modulated analog interferogram (broken curve, the same one as in (f)), and the discrete interferogram (broken line, the same one as in (f)); (h) discrete interferogram sampled by the A/D converter at its own sampling period t_0 (full time) and the corresponding analog interferogram (broken curve, the same one as in (g)). As for the definitions of t and t_m, see the figure caption for Figure 25.13 of Masutani *et al.* (Reproduced from Masutani *et al.* with permission. Copyright (1992) Society for Applied Spectroscopy.)

step-scan mode is the separation of the time of the experiment from the time of the data collection.

25.4.3
Asynchronous Time-Resolved FT-IR Spectrometer

In 1992, Masutani *et al.* (1992) proposed a time-resolving method for a conventional FT-IR spectrophotometer, which does not require synchronization between the timing for time resolving and that for the sampling of the AD converter. The system was named as an *asynchronous time-resolving system*. This method has the following advantages: (i) the signal-proceeding assembly for time-resolving measurements can be attached to any kinds of commercial FT-IR spectrometers and (ii) in principle, there is no shortest limit in time for transient phenomena to be measured.

Figure 25.10 shows a typical set up for the asynchronous time-resolved

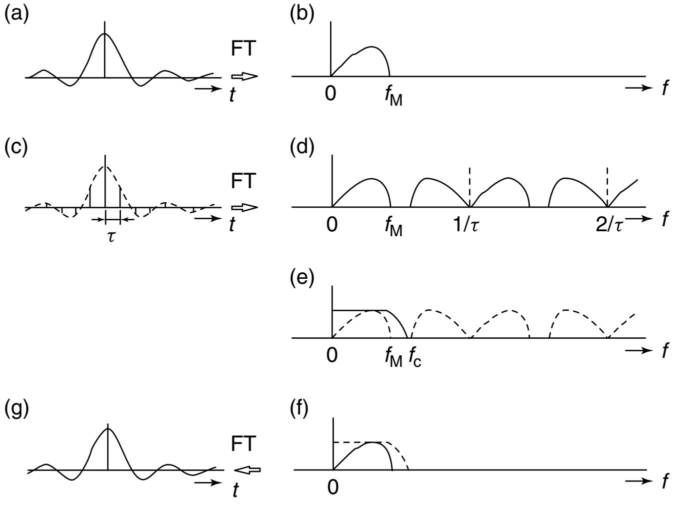

Fig. 25.12 Role of the low-pass filter. (a) An analog interferogram; (b) the spectrum corresponding to the analog interferogram in A with a maximum modulation frequency f_M; (c) a discrete interferogram (full line) with a sampling time interval t and its envelope (broken curve) corresponding to the analog interferogram in (a); (d) the spectrum obtained by inversely Fourier-transforming the discrete interferogram in (c); (e) frequency response of the low-pass filter with a cutoff frequency f_c (full curve) and the same spectrum as in (d) (broken curve); (f) output spectrum from the low-pass filter (full curve) and the same frequency response as in (e) (broken curve); (g) output analog interferogram from the low-pass filter (to be obtained by Fourier-transforming the spectrum in (f)). Horizontal arrows with FT mean the operation of Fourier or inverse Fourier transformation. Variable f in the abscissa is the modulation frequency, equal to 2 us, where u is the velocity of the moving mirror and s is the wavenumber of light. (Reproduced from Masutani *et al.* with permission. Copyright (1992) Society for Applied Spectroscopy.)

measurement (Masutani, 2002). Figure 25.11 depicts shapes of the signals in processes A–H in Figure 25.10 in alphabetical order: The main feature of this method is the use of a low-pass filter placed between the gate circuit and the A/D converter. The role of the low-pass filter is explained in more detail in Figure 25.3.

Figure 25.12 (a) and (b) shows an analog interferogram for radiation from the light source and a corresponding spectrum obtained by inversely Fourier-transforming the interferogram. When a discrete interferogram (c) obtained by sampling this analog interferogram by the gate circuit is inversely Fourier transformed, the spectrum obtained (d) consists of the original component and others that appear at the position determined by multiplying the sampling frequency $1/t$ by integers. Therefore, when this discrete interferogram is passed through a low-pass filter with an appropriate cutoff frequency f_c (e), the high-frequency components

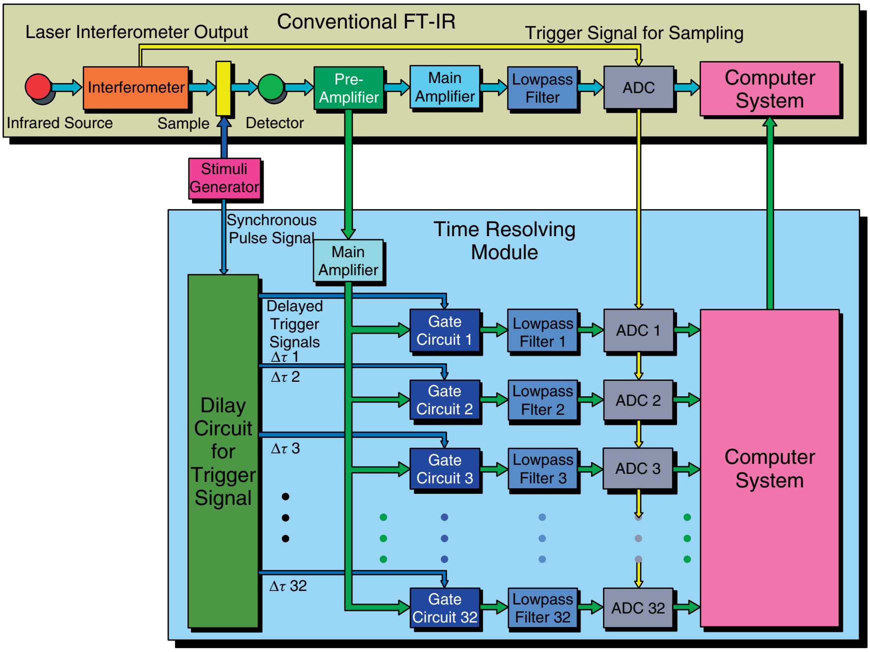
Conventional FT-IR
Laser Interferometer Output
Trigger Signal for Sampling
Interferometer
Pre-Amplifier
Main Amplifier
Lowpass Filter
ADC
Computer System
Infrared Source
Sample
Detector
Stimuli Generator
Synchronous Pulse Signal
Time Resolving Module
Main Amplifier
Dilay Circuit for Trigger Signal
Delayed Trigger Signals
Δτ 1
Δτ 2
Δτ 3
Δτ 32
Gate Circuit 1
Gate Circuit 2
Gate Circuit 3
Gate Circuit 32
Lowpass Filter 1
Lowpass Flter 2
Lowpass Filter 3
Lowpass Filter 32
ADC 1
ADC 2
ADC 3
ADC 32
Computer System

Fig. 25.13 A typical setup of the multichannel asynchronous time-resolving system.

are eliminated (f). The output spectrum (f) is identical to the spectrum (b) for the analog interferogram before sampling. Consequently, the analog interferogram (g) is obtained from the low-pass filter. Although the above Fourier transformations are not actually done by the low-pass filter, it has the function, in effect, as described above.

Figure 25.13 shows a typical setup of the multichannel asynchronous time-resolving system. The time-resolving module of the system has 32 time-resolving units, each of which consists of a gate circuit, a low-pass filter, and an A/D converter. In this system, time resolving is performed by setting each gate circuit at a different time delay. The sample is excited by repetitive stimuli at constant intervals, which do not depend upon the sampling timing of the FT-IR spectrometer. The sample excitations result in transient transmission differences, and thus the analog interferogram is modulated at very high frequency because of the repetitive excitation of the sample. The low-pass filters eliminate high-frequency components from the discrete time-resolved signals generated by the gate circuit and convert them to analog signals. The analog signals thus obtained are processed by the A/D converter at the sampling rate of the FT-IR spectrophotometer itself, as shown by the trigger signal for the sampling in Figure 25.13, which has no relation to the gate timing for time resolution.

References

Chalmers, J.M. and Griffiths, P.R. (2002) *Handbook of Vibrational Spectroscopy*, Vol. 1, John Wiley & Sons, Ltd, Chichester.

Christy, A.A., Ozaki, Y., and Gregoriou, V.G. (2001) *Modern Fourier Transform Infrared Spectroscopy*, Elsevier, Amsterdam.

Ferraro, J.R. and Basile, L.J. (Eds) (1978, 1979, 1982, 1985) *Fourier Transform Spectroscopy*, Vols.1–4, Academic Press, New York.

Griffiths, P.R. and de Haseth, J.A. (1986) *Fourier Transform Infrared Spectroscopy*, John Wiley & Sons, Inc., New York.

Griffiths, P.R. and de Haseth, J.A. (2007) *Fourier Transform Infrared Spectroscopy*, 2nd edn, John Wiley & Sons, Inc., New York.

Kauppinen, Y. and Hollberg, M.R. (1997) Spectrometers, infrared, *Encyclopedia of Applied Spectroscopy*, Wiley-VCH Verlag GmbH, Berlin.

Masutani, K. (2000) Time-resolved mid-infrared spectroscopy using an asynchronous fourier transform infrared spectrometer, in *Handbook of Vibrational Spectroscopy*, Vol. 1 (eds J.M. Chalmers and P.R. Griffiths), John Wiley & Sons, Ltd, Chichester, pp. 655–665.

Masutani, K., Sugisawa, H., Yokota, A., Furukawa, T. and Tasumi, M. (1992) Asynchronous time-resolved fourier transform infrared spectroscopy. *Appl. Spectrosc.*, **46**, 560–567.

Sheppard, N. (2002) The historical development of experimental techniques in vibrational spectroscopy, in *Handbook of Vibrational Spectroscopy*, Vol. 1 (eds J.M. Chalmers and P.R. Griffiths), John Wiley & Sons, Ltd, Chichester, pp. 1–32.

Further Reading

Bell, R.J. (1972) *Introductory Fourier Transform Spectroscopy*, Academic Press, New York.

Bourchareine, P. and Connes, P. (1963) Interferometer with compensated field for fourier transform spectroscopy. *J. Phys. (Paris)*, **24**, 134–138.

Bracewell, R. (1965) *The Fourier Transform and its Applications*, McGraw-Hill, New York.

Brigham, E.O. (1974) *The Fast Fourier Transform*, Prentice-Hall, Inc, New Jersey.

Buijs, H. (1979) in *Multiplex and/or High Throughput Spectroscopy*, SPIE Proceedings, Vol. 191 (ed. G.A. Vanasse), SPIE, Bellingham, p. 116.

Chamberlain, J. (1979) *The Principles of Interferometric Spectroscopy*, John Wiley & Sons, Ltd, Chichester.

Ferraro, J.R. and Krishmam, K. (Eds) (1990) *Practical Fourier Transform Infrared Spectroscopy, Industrial and Laboratory Chemical Analysis*, Academic Press, New York.

Mackenzie, M.W. (Ed.) (1988) *Advances in Applied Fourier Transform Infrared Spectroscopy*, John Wiley & Sons, Inc., New York.

SPECTROMETERS, MASS

R. GRAHAM COOKS, GUODONG CHEN, AND PHILIP WONG, *Department of Chemistry, Purdue University, West Lafayette, Indiana, U.S.A.*

HERMANN WOLLNIK, *II Physikalisches Institut, Universität Giessen, Giessen, Germany*

Introduction 289
1. **Ionization Methods** 292
1.1 Electron-Impact Ionization 292
1.2 Chemical Ionization 294
1.3 Desorption Ionization 295
1.4 Spray Ionization 296
1.5 Glow-Discharge Ionization 299
1.6 Inductively Coupled Plasma Ionization 300
1.7 Thermal Ionization 300
2. **Instrumentation** 301
2.1 Ion Optics 301
2.2 Mass Analyzers 302
2.2.1 Sector Magnets 302
2.2.2 Quadrupole Mass Filters 304
2.2.3 Quadrupole Ion Traps 306
2.2.4 Fourier-Transform Ion-Cyclotron-Resonance Mass Spectrometers 308
2.2.5 Time-of-Flight Instruments 311
2.3 Ion Detection 313
3. **Special Techniques and Their Applications** 314
3.1 Applications of Electron Capture 314
3.2 Chromatography/Mass Spectrometry 315
3.3 Mass Spectrometry/Mass Spectrometry 316
3.4 Mass Spectrometry for Analysis of Biopolymers 319
3.5 Isotope-Dilution Mass Spectrometry 319
3.6 Isotope-Ratio Mass Spectrometry 321
3.7 Mass Measurements of Stable and Unstable Nuclei 324
3.8 Resonance-Ionization Mass Spectrometry 324
3.9 Mass Spectrometry in Surface Analysis 324
3.10 Ion-Mobility Spectrometry 325
3.11 Analysis of Single Particles in Aerosols 327
Acknowledgment 327
Acrynoms 328
Glossary 328
Works Cited 329
Further Reading 330

INTRODUCTION

Mass spectrometry is both a method of physical and chemical analysis and a tool for studying chemical and physical processes (e.g., processes at the solid–vacuum interface, gas-phase ion chemistry, amino-acid sequences in proteins, etc.). It embraces a collection of topics united by common instrumentation. It is simultaneously

1. a form of spectroscopy,
2. a means of carrying out chemical reactions, and
3. a means of separating mixtures of ions on the basis of mass-to-charge ratios.

The principal area of application of mass spectrometry is chemical analysis—molecular analysis, isotopic analysis, isomer analysis, trace analysis, elemental analysis, and surface analysis. Its extremely wide applicability, high sensitivity, and isotopic specificity give mass spectrometry unique applications across the sciences (White, 1986). These applications include investigations of trace contaminants of gases, liquids, or solids as well as the study of the relative abundances of isotopes, and also unimolecular and bimolecular kinetics on time-varying systems including chemical reactors, flames, and the atmosphere. Thermochemical measurements, including measurements of enthalpies, excited-state energies, and bond energies, often rely on mass spectrometry. In

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

addition to these application areas, mass spectrometry is essential to the subject of ion chemistry, which includes ion structures, reaction mechanisms, and reaction dynamics associated with unimolecular and ion/molecule processes in a variety of environments.

In combination with chromatographic separation, principally in the form of the combined gas chromatography/mass spectrometry (GC/MS) instrument, mass spectrometry is the principal method used for organic analysis in environmental science. This application, like many others, is based on its high detection efficiency and specificity of molecular recognition. The combination of mass spectrometry with liquid chromatography (LC/MS) is increasingly important in characterization of biologically derived compounds, and recently much progress has been made in extending the applications of mass spectrometry to include both molecular-weight (MW) determinations and sequence analysis of biopolymers, including proteins, oligonucleotides, and glycopeptides. Often this type of measurement is performed using tandem mass spectrometry (MS/MS), an experiment in which mass analysis is performed twice, once to separate particular (ionized) compounds from a mixture, and then a second time to characterize the fragments generated from the selected ion. Drug metabolite studies in a biological matrix and protein sequencing often depend upon this method, both applications requiring the high sensitivity and chemical specificity of the mass spectrometer. The same characteristics have made mass spectrometry a major method of trace elemental analysis. This latter application is carried out by atomizing and ionizing the sample using a glow discharge or inductively coupled plasma, and the result is a sensitive, precise, and universal method of trace elemental analysis.

Mass spectrometry is also the principal tool in the analysis and separation of stable and of short-lived isotopes. Precise isotopic abundance measurements are possible even when one isotope of an element is many orders of magnitude less abundant than another. One specific method for achieving this uses the accelerator mass spectrometer (AMS). These high-energy (MeV range) instruments, usually based on Van de Graaff accelerators, often employ sputter ionization of solid samples and provide data of value in geology, biology, and other areas of science. For example, age determinations are made in geological as well as in anthropological time.

Mass spectrometry is also used for material modification through ion implantation and in microscale machining. It is an invaluable source of thermochemical information as well as a unique way of preparing and studying elusive neutral molecules and free radicals. It allows interstellar and atmospheric ion chemistry to be studied in the laboratory. It is the principal experimental tool in cluster science and is important in materials science. Finally, it provides unique insights into chemical reaction mechanisms by allowing their study under solvent-free or controlled-solvent conditions.

Mass spectrometry developed from 19th-century physics, starting with the study of electrical discharges in evacuated tubes. J. J. Thomson measured the charge-to-mass ratios of cathode rays, while W. Wien discovered the existence of positively charged ions in 1898 using a velocity filter consisting of crossed magnetic and electric fields (Fig. 1). Thomson built the first mass spectrometer, an instrument to measure the *m/z* values (i.e., mass *m* divided by the positive charge *z*) of the "corpuscules" that make up "positive rays," and this led to his 1912 demonstration of the existence of isotopes. Later in-

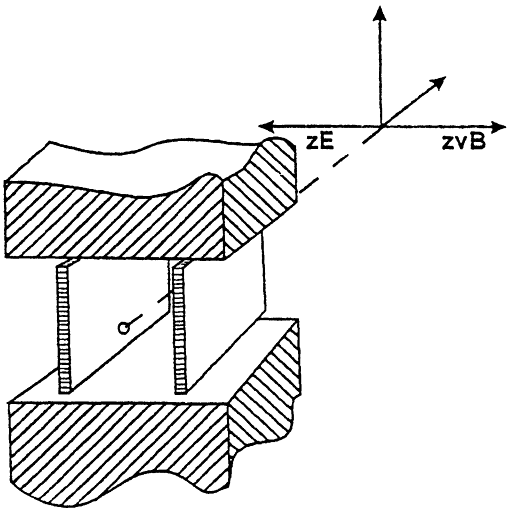

FIG. 1. Crossed magnetic and electric fields as used by Wien as a velocity filter.

struments allowed the determination of the natural abundances of isotopes and their mass defects, which helped unveil the shell structure of protons and neutrons in the atomic nucleus. Early applications in chemistry included bond energies and molecular ionization energies. Around 1940, chemical analysis by mass spectrometry began with the use of mass spectra to characterize petroleum distillates.

The principles of mass spectrometry can be illustrated by describing Thomson's positive-ray apparatus, which used a horizontal magnetic deflection and a vertical electrostatic deflection of ions (Fig. 2). From the patterns recorded on photographic plates, mass-to-charge ratios of the ions and their kinetic energies were determined (Fig. 3). Ions with a fixed mass-to-charge ratio but different kinetic energies appear as a parabola since the magnetic deflection is proportional to z/mv and the electrostatic deflection to z/mv^2. Later, Thomson scanned the magnetic field and used electrical detection to record the first mass spectrum; a 1912 spectrum is shown in Fig. 4.

In these early studies, molecular ions, isotopic ions, multiply charged ions, and negatively charged ions were all observed, as were products of charge exchange and other high-energy collision processes. The first observations of ion–molecule reactions (e.g., the production of H_3^+ from reactions of ion-

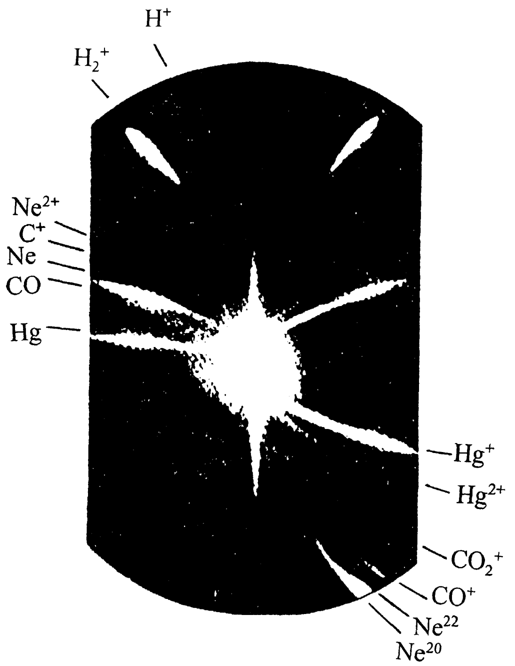

FIG. 3. Photograph of one of Thomson's photographic plates showing the partially resolved parabola due to the heavy isotope of neon, ^{22}Ne, together with parabolas for other ions (Thomson, 1913).

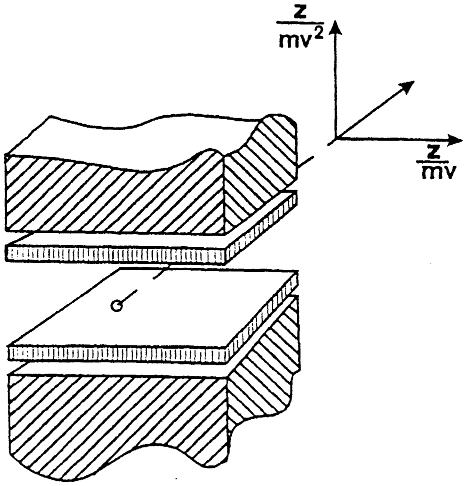

FIG. 2. Parallel magnetic and electric fields used by Thomson and also Kaufmann for mass-to-charge analysis.

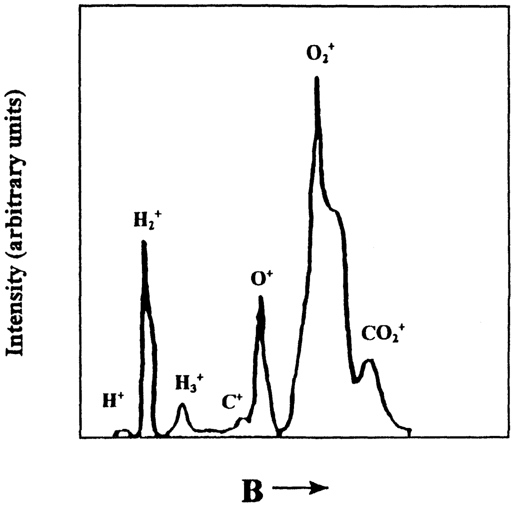

FIG. 4. Early mass spectrum obtained by J. J. Thomson in 1912. The abscissa shows the magnetic field strength (Thomson, 1913).

ized molecular hydrogen) were also made. Incipient in Thomson's work was organic analysis, thermochemistry, and mixture analysis. In 1913 Thomson urged "... especially chemists, to try this method of analysis. I feel sure that there are many problems in Chemistry which could be tackled with far greater ease by this than any other method. The method is surprisingly sensitive—more so than that of Spectrum Analysis, requires infinitesimal amounts of material, and does not require this to be specially purified: the technique is not difficult. ..."

Different physical principles can be used to separate and measure the masses of ions, and this results in a plethora of mass spectrometers of different types. The criteria of greatest importance in a mass spectrometer vary with the type of application. Precise abundance measurements are important in isotope-ratio mass spectrometry, access to ions of very high mass in applications to biopolymers, maximum resolution in identification of new synthetic organic compounds, and maximum sensitivity in trace analysis of environmental toxins. Various types of mass spectrometers best satisfy these different criteria, and this accounts for the variety of mass spectrometers one encounters. For example, widely used instruments are based on physical dispersion of ion beams as in Thomson's magnetic deflection instrument, on measuring the flight time of accelerated ions (time-of-flight mass spectrometers), and on measuring frequencies of motion of trapped ions (ion cyclotron resonance mass spectrometers). If one adds to this variety of mass analyzers the variety of methods of producing ions and detecting them, one understands why there is an enormous diversity of modern instruments.

1. IONIZATION METHODS

An ideal ion source for mass-spectrometric analysis should provide a high ionization efficiency, and high stability with minimum kinetic energy distribution and minimum angular dispersion of the ions produced so as to optimize the performance of the attached mass analyzer. In addition, when molecular analysis is of interest, the ion source should produce intact molecular ions for MW information, fragment ions for structural information, and control over the internal energy transferred to the molecule for control over the degree of fragmentation, and it should be possible to couple it with various types of chromatographs. The various ionization methods suitable for molecular analysis include electron-impact (EI) ionization, chemical ionization (CI), desorption ionization (DI), and spray ionization (SI). Techniques suitable for production of elemental ions include the DI technique of secondary-ion mass spectrometry (SIMS), glow-discharge (GD) ionization, inductively coupled plasma (ICP) ionization, as well as thermal ionization and other less frequently used methods (field ionization, electrohydrodynamic ionization, and spark-source mass spectrometry). A knowledge of the sample type (gas, liquid, solid, single component, complex mixture) and the desired specific application (elemental analysis, MW determination of a biopolymer, isotope ratio measurement, etc.) as well as information on how most compound types behave under different ionization conditions will dictate the most appropriate method. Table 1 summarizes some characteristics of different ionization methods.

1.1 Electron-Impact Ionization

Molecules of vaporized samples are ionized by collisions with energetic electrons (typically of about 70 eV) produced from a heated filament. Figure 5 shows a schematic diagram of a typical EI ion source. Electrons emitted from the filament (typically 100 μA) are accelerated into the ionization region where a small fraction of them collide with the sample molecules. The resulting molecular ions, together with fragment ions produced by their dissociation in the ion source, are then injected into the mass analyzer by the repeller electrode, which is held at a positive voltage relative to the ion-source block.

The mass of the molecular ion yields the nominal MW of the compound (nearest atomic mass unit), while the masses and abundances of the fragment ions can be interpreted to give structural information. Electron-impact sources require volatile, thermally stable samples and generate mainly positive ions.

Because of the physical processes involved in ionization and the limitation of fragmen-

Table 1. Summary of main ionization methods.

Ionization method		Ionization agent	Pressure	Characteristic	Application
EI	Molecular	Electrons (~70 eV)	$\sim 10^{-5}$ Torr	Reproducible spectra, extensive fragmentation	Volatile compounds; structural elucidation
CI	Molecular	Gaseous ions	~1 Torr	Molecular ions and controllable fragmentation	Volatile compounds; MW determination
DI	Molecular	Energetic particles, photons	10^{-6}–10^{-5} Torr	Intact molecular ions from high-mass compounds	Condensed phase, high-mass compounds; MW and structure determination
SI	Molecular	Electrical, thermal, and pneumatic energy	1–760 Torr	Intact molecular ions from high-mass compounds	Solutions of high-MW compounds; used with LC/MS to determine MW's and structures
GD	Elemental	Plasma	0.1–10 Torr	Very stable, reproducible spectra	Elemental analysis of solids
ICP	Elemental	Plasma	Atmospheric pressure	High ionization efficiency	Elemental analysis of solutions

tation to unimolecular processes, EI produces highly reproducible mass spectra. Hence, library searching of EI mass spectra is possible. Algorithms are available that search more than 100 000 mass spectra and record matches with the spectrum of an unknown compound within a few seconds. In order to interpret EI mass spectra and gain structural information, it is necessary to understand the series of competing and consecutive unimolecular fragmentations originating from the molecular ion. Generally, the rate constant for each step can be calculated by applying an appropriate form of reaction-rate theory, and in this way, the mass spectra of simple compounds can be calculated.

Figure 6 shows an EI mass spectrum of acetophenone. The small peak at $m/z = 120$ corresponds to the intact molecular ion. The major fragment ion is at $m/z = 105$ and is due to the loss of a methyl radical, a feature of the structure of the compound. The loss of carbon monoxide (28 atomic mass units) from this fragment ion leads to the fragment ion at $m/z = 77$, and further fragmentation (loss of acetylene, 26 atomic mass units) gives rise to the fragment ion at $m/z = 51$. The peak at $m/z = 43$ is a result of fission of

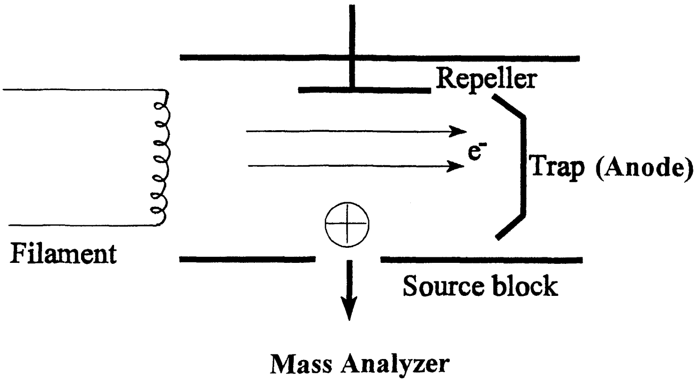

FIG. 5. Electron-impact ionization source.

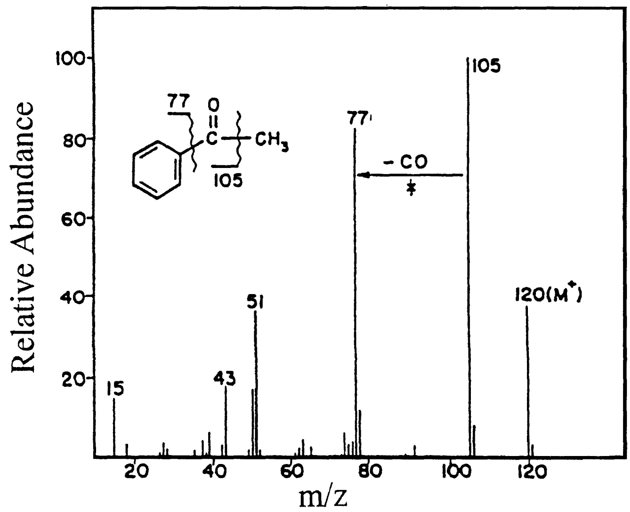

FIG. 6. El mass spectrum of acetophenone (Budzikiewicz *et al.*, 1967).

the phenyl-carbonyl bond with retention of charge on the carbonyl fragment. It can be seen from this example that EI produces structurally characteristic fragmentations. Information of this type can, in turn, be used to facilitate identification of unknown compounds.

1.2 Chemical Ionization

Many molecules do not give recognizable molecular ions in EI because of extensive fragmentation. However, this difficulty can be overcome by using CI. A reagent gas (e.g., CH_4) is ionized in an ion source at a pressure of about 1 Torr, and a characteristic set of product ions is generated in the lightly ionized plasma by ion–molecule and electron–molecule interactions. These reagent ions, in turn, undergo ion–molecule reactions with the analyte, which is present at low partial pressures to avoid its direct EI.

The reaction sequence is as follows:

1. Reagent gas molecules (e.g., CH_4, NH_3, NF_3, N_2O) are ionized by EI,
2. ion–molecule reactions produce reagent ions (e.g., correspondingly, CH_5^+, NH_4^+, F^-, O^-),
3. these reagent ions react with sample molecules (M) to form quasimolecular ions [MH^+, $(M - H)^-$, $(M + O)^-$], and finally
4. unimolecular fragmentation or collisional cooling of these ions produces the ions observed in the mass spectrum.

For example, the reaction sequence in the case of ammonia as the reagent gas is

1. $NH_3 + e^-(70\ eV) \rightarrow NH_3^{+\cdot} + 2e^-$,
2. $NH_3^{+\cdot} + NH_3 \rightarrow NH_4^{+*} + NH_2^{\cdot}$, $NH_4^{+*} + NH_3 \rightarrow NH_4^+ + NH_3^*$,
3. $NH_4^+ + M \rightarrow NH_3 + MH^{+*}$,
4. $MH^{+*} \rightarrow$ fragment ions and neutral fragments, $MH^{+*} + NH_3 \rightarrow MH^+ + NH_3^*$,

where the asterisk indicates internal molecular excitation. Fragmentation patterns for EI and CI give complementary information; often both ion sources are available on the same instrument.

The CI method is versatile because it depends on chemical reactions that yield particular products, and it deposits a controlled amount of energy into them. This leads to selective ionization of particular compounds present in a mixture. The energy deposited into the quasimolecular ion, and hence the extent of its fragmentation, can be changed by choosing a different reagent ion. For a proton-transfer reaction $M + GH^+ \rightarrow MH^+ + G$, the enthalpy change for the reaction is given by the proton affinity difference between the sample (M) and reagent gas (G). The proton affinity (PA) is defined as $PA(M) = \Delta H_{reaction}\ (MH^+ \rightarrow M + H^+) = \Delta H_f(H^+) + \Delta H_f(M) - \Delta H_f(MH^+)$. The maximum internal energy in the protonated molecule produced in CI is therefore given by ΔPA, the difference in proton affinities of M and G, and this energy controls the degree of its fragmentation. For instance, protonated methane (PA = 536 kJ/mol) reacts with acetophenone (PA = 859 kJ/mol), transferring large amounts of energy (ΔPA = 323 kJ/mol) to protonated acetophenone. This ion therefore fragments extensively. Ammonia, by contrast, is a relatively gentle reagent (PA = 857 kJ/mol) in the proton-transfer reaction with acetophenone. The spectrum therefore shows an abundant protonated molecular ion due to the low internal energy deposited in the protonated molecule (ΔPA = 2 kJ/mol). Any compound in the sample with a proton affinity less than that of the reagent gas molecule will not undergo proton transfer, and hence the experiment is transparent to such compounds, which will not interfere with the observation of other compounds in the sample.

1.3 Desorption Ionization

A major limitation of both EI and CI is the requirement that the sample be volatile, and this means that both methods are limited to thermally stable compounds of low molecular weight. Desorption ionization provides one means to ionize nonvolatile samples with high molecular weight. In this family of ionization methods, energy is provided to a condensed-phase sample that is deposited on a surface. The energy can take the form of high-energy particles, or a photon beam. Particle bombardment can be achieved using atoms (e.g., Ar) or ions (e.g., Cs^+) with 5–10-keV energies, as well as multiply charged atomic ions with energies up to 100 MeV, generated in accelerators or from nuclear fission. Laser beams used for DI often have short pulse lengths (10 ns) and photon energies in the UV spectral region. High electrical fields are used in field desorption (FD), where ionization occurs by ion–molecule reactions at a heated surface with transport of ions into the gas phase as a result of a high electrical field. Sample molecules are deposited from solution on fine dendritic needles, which yield field strengths of ca 10^{10} V/m when kilovolt potentials are applied. Table 2 summarizes some of the procedures used in the DI family of ionization methods.

The key feature in DI is the formation of intact quasimolecular ions. When the sample surface is energized, the energy is not transferred directly to the sample molecules. Rather, the transfer of energy is mediated by absorption, excitation, and relaxation processes in the matrix (Fig. 7). This liberates intact molecules of analyte, which are observable by mass spectrometry after ionization, which often occurs in the course of ion/molecule reactions in the energized interface.

Desorption ionization has wide applications both in elemental surface analysis and for the characterization of organic, inorganic, and biological molecules. For example, static secondary-ion mass spectrometry (SIMS) uses a low flux of primary ions and allows the determination of elemental composition, including the spatial distribution of

Table 2. Procedures used in desorption ionization.[a]

Agent	Method	Flux	Matrix	Mass analyzer	Comments
keV ions	Static secondary-ion MS (SIMS)	10^{-10} A cm^{-2} (Ar^+, Cs^+, etc.)	None, solid	Any	Surface sensitive, nondestructive, low signal
keV ions	Dynamic SIMS	10^{-6} A cm^{-2} (Cs^+, etc.)	Liquid	Any	Higher, longer-lasting signal
keV atoms	Fast-atom bombardment (FAB)	10^{13} atoms cm^{-2} s^{-1} (Ar atoms)	Liquid	Any	High long-lasting signal, high background
MeV ions	Plasma desorption (PD)	10^3 particles cm^{-2} s^{-1}	None	Time-of-flight	Best ionization efficiency, low signals
Photons	Laser desorption (LD) and matrix-assisted laser desorption (MALDI)	$\geq 10^6$ W cm^{-2}	Solid matrix absorbs radiation in MALDI	Often time-of-flight	CW lasers can give thermal degradation

---

[a] Adapted from Lambert *et al.* (1987).

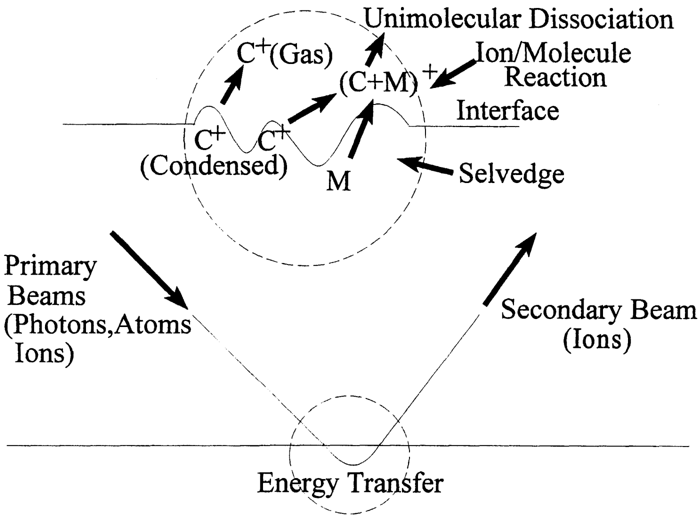

FIG. 7. Ionization in DI, showing the disturbed interfacial region known as the selvedge; M is a neutral molecule, C^+ is a cation (Lambert *et al.*, 1987).

elements in the outermost atomic layers of a material surface. Dynamic SIMS uses a high flux of primary ions, which allows depth profiling by removal of material, layer by layer, via sputtering. Imaging is achieved by rastering a finely focused ion beam across the surface (ion microprobe) or through the use of a direct-imaging instrument (ion microscope). The SIMS process involves a series of binary atomic collisions (a collision cascade) within the near-surface layers of the sample, which serve to transfer momentum and release atoms and ions into vacuum. SIMS has been utilized in material science, metallurgy, catalysis, and polymer science, and it is extremely sensitive. However, a problem associated with SIMS is the difficulty of quantification, since the ionization probability depends strongly on the surface composition. External standards with similar compositions to the sample are essential for quantitation. Traditionally, quadrupole mass filters and sector magnets have been used as SIMS analyzers; more recently, time-of-flight analyzers have been used in such experiments.

An increasingly popular alternative technique for the ionization of nonvolatile compounds, matrix-assisted laser desorption ionization (MALDI), employs a UV-absorbing matrix such as nicotinic acid, present in large excess, to absorb the photon energy. As the energized matrix evaporates, it carries the sample into the gas phase without extensive fragmentation. MALDI mass spectra are dominated by singly charged molecular ions with molecular weights as high as 500 000 atomic mass units. Figure 8 shows a MALDI mass spectrum of the protein glucose isomerase. The highly abundant molecular ion at m/z = 172 450 is evident, as is the relatively low mass resolution.

1.4 Spray Ionization

The DI techniques constitute one means to ionize nonvolatile biological compounds; SI constitutes another group of methods for handling such samples. In the latter case, aqueous solutions are examined directly. In the SI methods, thermal, electrical, or pneu-

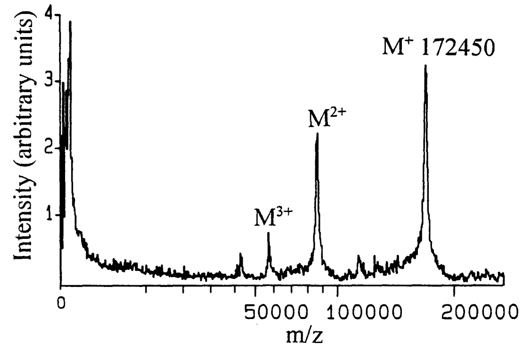

FIG. 8. MALDI mass spectrum of glucose isomerase recorded using nicotinic acid matrix (Karas *et al.*, 1991).

Table 3. Characteristics of the spray-ionization techniques.[a]

Ionization method	Energy source	Operating condition	Ion formation	Comments
Thermospray (TS)	Thermal energy	1–10 Torr, elevated temperature	Gas-phase ion–molecule reactions	Temperature-dependent spectra, interface for LC/MS
Electrospray (ES)	Electric field	Reduced pressure, ambient temperature	Ion–molecule reactions in solution, droplets	Multiply charged ions for high-MW determination; interface for LC/MS and capillary-zone electrophoresis (CZE)/MS
Atmospheric-pressure chemical ionization (APCI)	Corona discharge	Atmospheric pressure, elevated temperature	Gas- and solution-phase ion–molecule reactions	Multiply charged ions; interface for LC/MS and GC/MS

[a] There is considerable overlap in techniques and phenomena between these methods.

matic energy, or a combination of these, is used to accomplish both vaporization and ionization. Small, charged droplets are formed when an aqueous sample solution is sprayed from an electrically charged needle. Evaporation is facilitated by a heated gas, and droplet subdivision occurs by a Coulomb explosion process until, eventually, individual ions remain. These typically have multiple charges, due to attached protons or adventitious alkali metal ions in the case of positive ions. Mass/charge analysis of these ions reveals that the number of charges is not fixed, and so molecular weights can be measured from *m/z* values with some redundancy (see below). Some characteristics associated with different SI methods are summarized in Table 3.

A variant on CI, which is also closely related to electrospray ionization, is atmospheric-pressure chemical ionization (APCI). Figure 9 illustrates a typical APCI source. In APCI, ion–molecule reactions or electron–molecule reactions occur at atmospheric pressure where long residence times in a field-free drift region govern the reaction dynamics. These long residence times, typically 2–5 s, allow sufficient collisions with reagent molecules or electrons to completely ionize trace organic compounds introduced into the source. This method provides high sensitivity (subpicogram detection limits) and is there-

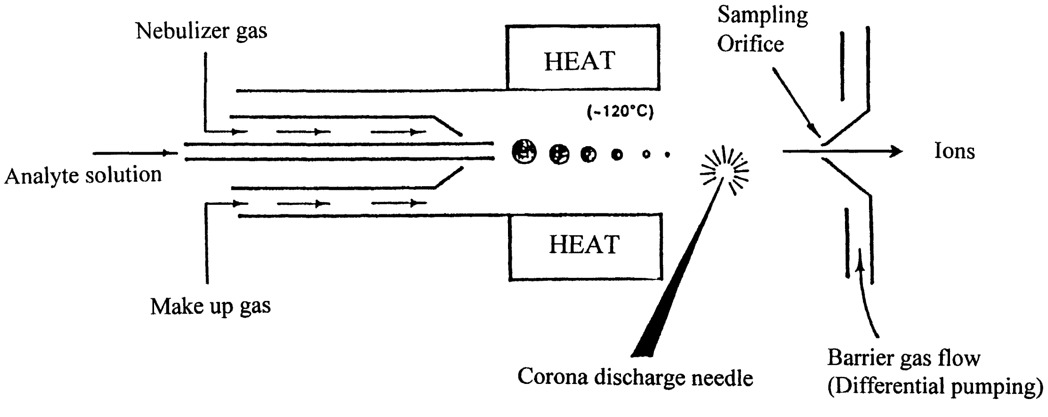

FIG. 9. Typical atmospheric-pressure chemical-ionization (APCI) source (Chapman, 1993).

fore very useful in environmental monitoring and biological-molecule characterization.

A valuable feature of APCI and electrospray (ES) ionization itself is the formation of multiply (as opposed to singly) charged ions from high-MW species. For example, on the order of 100 protons can be attached to a bovine albumin dimer with a MW of about 133 000 atomic mass units to give ions with *m/z* values of only 1330 atomic mass units/charge. These values are easily accessible with conventional instruments. Charge accumulation on the liquid surface and subsequent disruption of the droplet by the applied high electrical field yield, after complete solvent vaporization, intact molecules bearing multiple charges and allow accurate determinations of the molecular weights of large biomolecules. Because multiple charging reduces the *m/z* ratio of an ion, this ionization method has greatly extended the applicability of mass spectrometry to high-MW compounds. For example, simple quadrupole mass analyzers can be used to measure molecular weights of up to 100 000 atomic mass units with a mass accuracy of ±0.01%. Figure 10(a) shows an ES mass spectrum of cytochrome *c*, a compound with a MW of 12 360 atomic mass units. Charge states from 10 up to 19 are observed, i.e., $(M + 10H)^{10+}$ through $(M + 19H)^{19+}$, providing redundancy in establish-

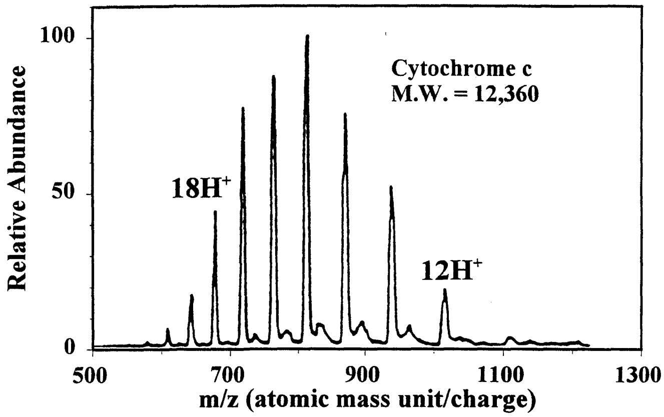

(a)

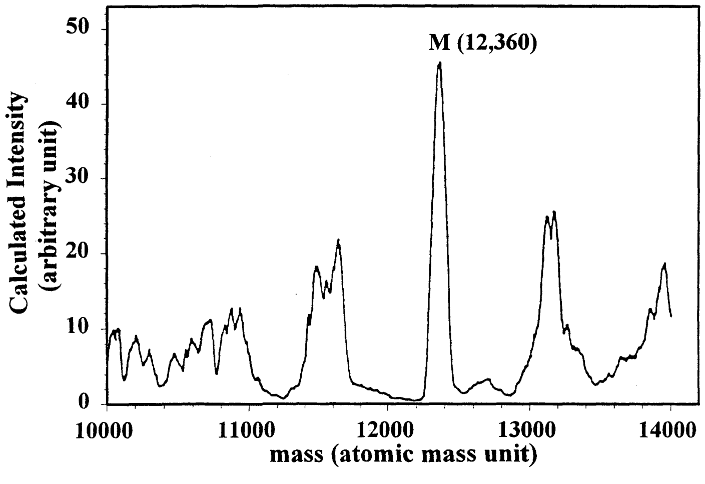

(b)

FIG. 10. (a) Electrospray-ionization mass spectrum of cytochrome *c* plotted conventionally in terms of mass-to-charge ratio. (b) Deconvoluted spectrum from (a), plotted as ion mass vs. abundance. Charge states are shown next to individual peaks in (a). (Mann *et al.*, 1989.)

ing the MW. Figure 10(b) illustrates a deconvoluted spectrum that displays ion abundance vs. mass (not mass–charge ratio). It shows that the major component of the spectrum has a MW of 12 360 atomic mass units. The fact that this ion can acquire a variable number of charges accounts for the appearance of the mass spectrum in Fig. 10(a). The ionization efficiency in ES is extremely high, and this leads to very low detection limits—in exceptional cases down to the attomole range. Electrospray also provides sequence information, for example, for peptides and proteins through the use of tandem mass spectrometry (see Sec. 3.4).

1.5 Glow-Discharge Ionization

Glow discharge is an ionization method that is most often employed in trace elemental analysis of samples in the solid state. A GD results from the passage of electric current (typically in the microampere range) through a gaseous medium (0.1–10 Torr), traditionally using a dc voltage of several hundred volts. Inherent in the operation of the GD source is the cathodic sputtering process that produces a population of atoms directly from the solid cathode. In the steady state, accelerated positive ions of the gaseous species (e.g., Ar^+), present in the discharge, strike the sample surface with sufficient energy to create a collision cascade and to dislodge one or more surface atoms per incoming ion. These sputtered gas-phase atoms are then excited and ionized through collisions in the plasma. Sampling of the plasma with a mass spectrometer produces a mass spectrum from which elemental identities and abundances are determined.

The GD ion source is popular in the elemental analysis of solids on account of its stability, sensitivity, and simplicity of operation. It suffers from minimal matrix effects because atomization and ionization occur as separate steps. This results in a range of relative elemental sensitivities of less than one order of magnitude across the periodic table. Quantification is performed with certified standards. The detection limit for most elements is in the parts-per-billion (ppb) range. The method has been widely applied to the analysis of metals, alloys, high-purity materials, thin films, and coatings, and also for surface studies and in material science applications. Figure 11 shows the results of a thin-film analysis, which includes the detection of two 10-nm-thick iron films separated by EuO layers. The sputter removal rate of material was about 0.035 nm/s.

An older alternative to GDMS, spark-source mass spectrometry, is still used for the direct elemental analysis of solid samples. In this source a pulsed MHz-frequency radio-frequency (rf) voltage of several tens of kilovolts is applied across a small gap between two electrodes, creating a spark plasma. Usually, samples are incorporated directly into the electrodes. The ion beam produced may be erratic and has a wide kinetic energy spread (2–3 keV), which has often resulted in a double-focusing mass spectrometer being employed for analysis (see Sec. 2.2.1). The reproducibility is poor, and long integration times are required because of the instability of the ion signal. However, this method has the advantage of simultaneously monitoring all elements (except hy-

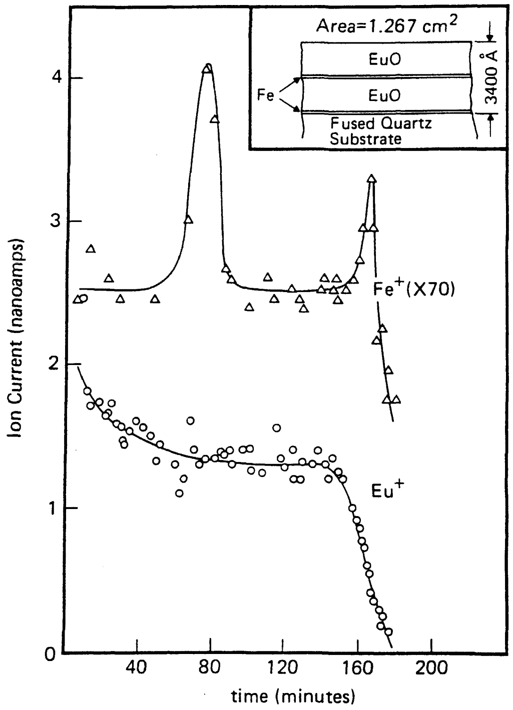

FIG. 11. Glow-discharge mass-spectrometric analysis of a 3400-Å EuO film containing two thin layers of iron, during rf discharge (13.56 MHz) in argon. Iron was detected as m/z = 56 and europium as m/z = 151 and 153. (Coburn and Kay, 1971.)

drogen) with detection limits down to 10 ppb, and it has been used for microsamples and surface contamination studies.

1.6 Inductively Coupled Plasma Ionization

The inductively coupled plasma ion source is used for elemental analysis of aqueous solutions. Figure 12 shows a schematic diagram of the ICP source and its ion-extraction optics. The main body of the ICP torch consists of a quartz tube surrounded by an induction coil that is connected to a rf generator. The sample solution is aspirated into the tube via an argon gas stream. The temperature in the ICP can reach 5 000–10 000 K, resulting in a high ionization efficiency for most elements. The ions leave the ICP together with neutral gas atoms through an extraction aperture in a sampling cone, and they are then analyzed by a mass analyzer, typically of the quadrupole type.

Inductively coupled plasma MS is often used in geological, environmental, clinical, and food analysis. The method has very low detection limits (few pg/mL range), excellent quantification, and a wide linear dynamic range (ca six orders of magnitude). The attainable precision in quantitation in ICP-MS is on the order of 0.1–0.5% relative standard deviation, by using internal standards or isotope dilution techniques. Isobaric interferences and sample matrix interferences cause difficulties in the analysis of a few elements.

In addition to chemical studies, ICP and GD, as well as spark-plasma mass spectrometry, find applications in plasma physics, microelectronics, fusion research, and combustion engineering among other fields. For example, mass-spectrometric studies on phenomena associated with electrical arcs and the thermodynamic and electrical properties of vapors have contributed greatly to the development of electrical discharge lamps. Mass-spectral data relating to chemical intermediates in flames provide valuable diagnostic information in combustion engineering.

1.7 Thermal Ionization

This ionization method produces atomic or molecular ions at the hot surface of a metal filament, the temperature of which can reach 2000 K. High yields of positive and negative ions can be obtained for atoms or molecules with low ionization energy (e.g., metals, such as K with ionization energy $I_P = 4.3$ eV) and with high electron affinity (e.g., nonmetals, such as Cl with electron affinity $E_A = 3.6$ eV). The ion yield β, the ratio of the number of ions to the total number of ions and neutral atoms, is usually represented by the following equations for positive and negative ion formation:

$$\beta^+ = N^+/(N^0 + N^+) = \{1 + (g_0/g_+) \exp[(I_P - W)/kT]\}^{-1},$$

$$\beta^- = N^-/(N^0 + N^-) = \{1 + (g_0/g_-) \exp[(W - E_A)/kT]\}^{-1},$$

where N is the number of ions or atoms; g is a statistical factor; I_P is the first ionization energy of the analyte; k is the Boltzmann constant; T is the temperature of the filament in kelvins; W is the work function of

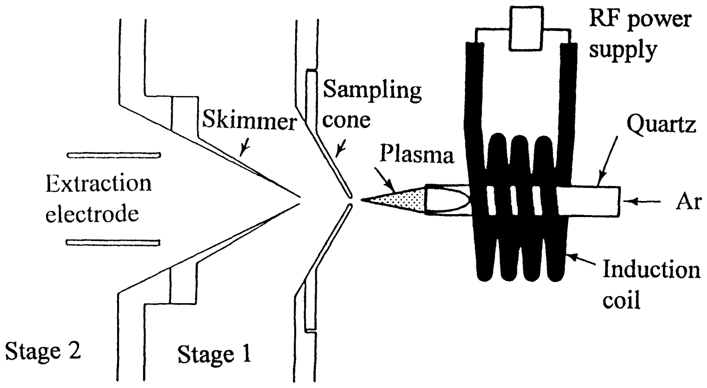

FIG. 12. Schematic diagram of ICP source and its ion-extraction interface. Typical operating pressure: stage 1, 1–3 Torr; stage 2, 10^{-3}–10^{-4} Torr. (Adams *et al.*, 1988; Ingle and Crouch, 1988.)

the filament material; E_A is the electron affinity of the analyte; and the indices 0, +, and − indicate a neutral atom, a positive ion, and a negative ion, respectively. It can be seen that a high work function of the filament material is advantageous in increasing the positive ion yield, whereas for the production of negative ions a material with a low work function should be used.

Thermal ionization generates ions with a very narrow kinetic-energy spread. It is primarily used in conjunction with isotope-dilution (ID) techniques in the nuclear and earth sciences. The detection limits are generally in the order of μg/g, viz., parts per million.

2. INSTRUMENTATION

2.1 Ion Optics

To understand how different mass spectrometers function, it is necessary to understand how ions move in magnetic and electric fields. An ion of mass m, bearing electronic charge z, i.e., of mass-to-charge ratio m/z, is accelerated by a potential V to the kinetic energy $K = mv^2/2 = zV$ and thus to a velocity $v = (2K/m)^{1/2}$. Numerically, this velocity is given in mm/μs, or equivalently in km/s, as

$$v \approx 13.89\sqrt{K/m}, \qquad (1)$$

with K in electron volts (eV) and m in atomic mass units (i.e., 1/12 of the mass of a ^{12}C atom). Assume now that ions of mass-to-charge ratio m/z are formed in processes that leave them moving in all directions with energies $K_i \leq K_{i0}$ or, in other words, velocities $v_i \leq v_{i0}$, where K_{i0} and v_{i0} are the initial energy and velocity of the ions formed in the source. Upon accelerating these ions perpendicularly (in the z direction) at potential V_0 to an energy $K_0 + K_i$, the velocity component v_z is increased while the perpendicular velocity components v_x and v_y are unchanged. Thus one finds

$$v_x \leq 14\sqrt{K_{i0}/m}, \quad v_y \leq 14\sqrt{K_{i0}/m},$$
$$v_z \leq 14\sqrt{(K_{i0} + K_0)/m} \approx 14\sqrt{K_0/m}.$$

One notes also that the ion trajectories are all inclined relative to the ion optic z axis at angles α and β, where

$$\tan \alpha \approx v_x/v_z \leq \sqrt{K_{i0}/K_0},$$
$$\tan \beta \approx v_y/v_z \leq \sqrt{K_{i0}/K_0}.$$

Diverging ions starting from one point can be focused back to a point—for instance, by rotationally symmetric lenses. Such a lens—as indicated in Fig. 13—always has a region in which the ions are forced toward the optic axis (focusing forces) and two regions in which the ions are forced away from the optic axis (defocusing forces). These forces, however, do not cancel and leave the ion trajectories undeflected; rather, over all, the ions are focused toward the optic axis since they move through the focusing regions with a reduced forward velocity component as compared to the defocusing regions. This finding holds true whether the center electrode in the rotationally symmetric lens is at a potential that reduces or one that increases the ion velocity. Note that Fig. 13 shows the case of a reduced intermediate ion velocity. Note also that the second half of the lens in Fig. 13 represents an ion acceleration system, which clearly has an overall focusing action as well.

Alternative focusing devices are quadrupole lenses (Fig. 14) in which the potential increases quadratically with x, so that ions are forced back to the optic axis (in x direction) or away from it (in y direction) with both forces increasing linearly with x or y, respectively. Since quadrupole lenses use their main field components for focusing, they are much stronger than rotationally symmetric lenses whose focusing strength depends on differential effects in their fringing fields. However, quadrupole lenses can only be used in doublets, triplets, quartets, . . . with alternating field configurations, since they always focus the beam in one direction and defocus it in the perpendicular direction and thus achieve overall focusing by appropriate combinations of lenses.

The considerations presented so far assume that all ions start from a pointlike ion source. In reality such a source always has a certain width—for instance, x_0 (Fig. 15)—from each point of which ions leave with divergence angles less than $\pm\alpha_0/2$. These ions then are focused to form an image of width x_1, i.e., each point of the image is reached by

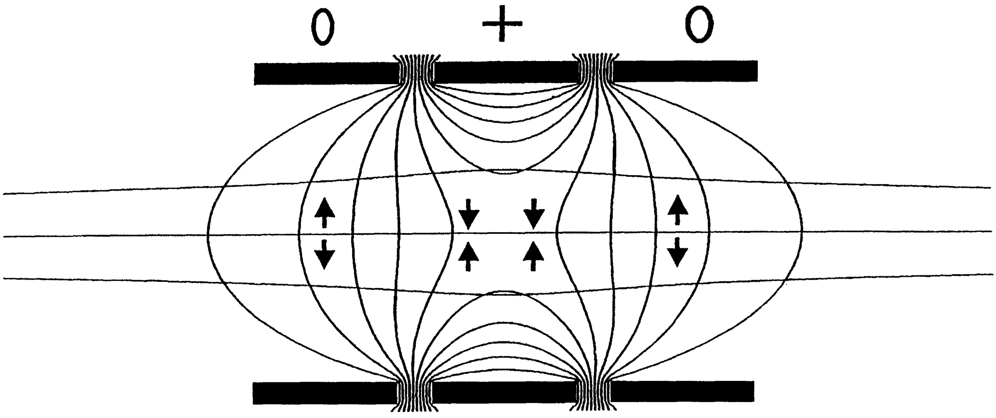

FIG. 13. An electrostatic lens that consists of three coaxially arranged metallic tubes. The two outer tubes are at ground potential, while the middle one is at a potential such that the incoming charged particles are decelerated between the first and second electrode and accelerated again between the second and third electrode with all forces acting perpendicularly to the equipotential lines shown. In the first half of the lens the charged particles are decelerated, and these particles thus are first pushed away from the optic axis (defocusing) and then pushed toward the optic axis (focusing). In the second half of the lens, these forces are reversed so that the charged particles are first pushed away from the optic axis (defocusing) and then toward the optic axis (focusing). Since in the center region the forward component of the particle velocity is reduced, the focusing forces are always larger than the defocusing forces in the initial and final regions.

ions with angles in the range $\pm\alpha_1/2$. If an object–image relationship exists between the ion source and the image, i.e., if ions that leave one point of the source all reach a corresponding point of the image, one finds for ions of given velocity that

$$x_0\alpha_0 = x_1\alpha_1,$$

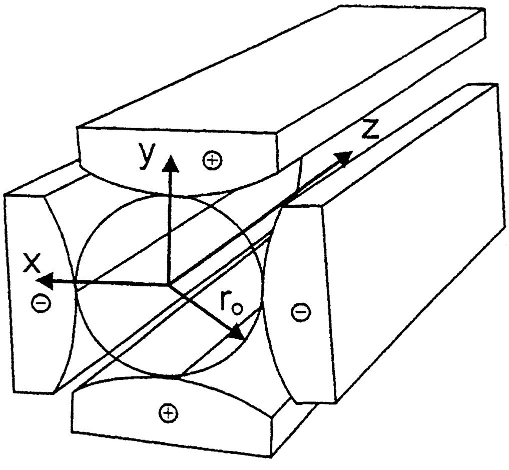

FIG. 14. Quadrupole electrode system that focuses in one direction and defocuses in the perpendicular direction. The inscribed radius is r_0.

which is a consequence of Liouville's theorem derived from thermodynamics. From this relationship it follows that a small image ($x_1 < x_0$) is always accompanied by large angles of divergence ($\alpha_1 > \alpha_0$) of the ion trajectories and a large image by more nearly parallel trajectories.

2.2 Mass Analyzers

The function of a mass analyzer is to measure the mass-to-charge ratios of ions, thus providing a means to identify them. Although there are different types of mass analyzers, their operating principles all depend on the interactions of charged particles with electric or magnetic fields. Five types of mass analyzers in common use are sector magnets, quadrupole mass filters, ion traps, time-of-flight (TOF) instruments, and Fourier-transform ion-cyclotron-resonance (FT-ICR) mass spectrometers. Table 4 summarizes some of their performance characteristics.

2.2.1 Sector Magnets An ion that passes perpendicularly through a magnetic field B moves along a circle of radius r_B along which the centrifugal force mv^2/r_B is balanced by the magnetic force Bzv so that

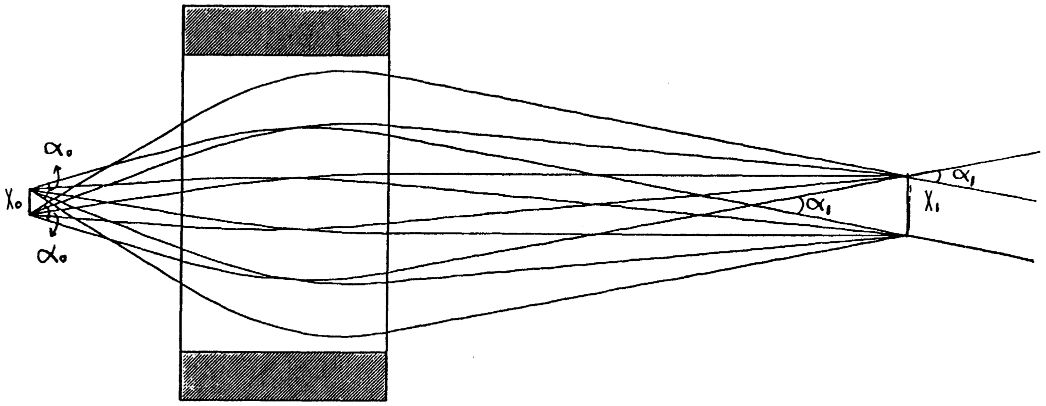

FIG. 15. An ion source of width x_0 from which ions leave with divergence angles $\pm\alpha_0/2$. The corresponding image parameters are x_1 and α_1.

$Br_B = mv/z$. Thus the magnetic field B directs ions along radii r_B that are proportional to their mv/z ratio. Numerically, this value, the magnetic rigidity, is obtained in tesla-millimeters as

$$Br_B \approx 0.144\sqrt{Km/z^2} \qquad (2)$$

where B is in teslas (V m/s^2) and r_B is in millimeters if the mass m is in atomic mass units and the energy K in eV. For ions accelerated through the same potential V one obtains the mass analysis equation $m/z = 48.9B^2r_B^2/V$.

When an ion passes through a cylindrical electrostatic sector field (also called an electrostatic sector analyzer, ESA), it moves along a curve of momentary radius r_E along which the centrifugal force mv^2/r_E is balanced by the electrostatic force zE so that $Er_E = mv^2/z$, where E is the component of the electrostatic field that is perpendicular to the ion trajectory. Thus an ESA directs ions on radii r_E proportional to their translational energy-to-charge (K/z) ratios. Numerically, this so-called electrostatic rigidity is obtained in volts as

$$Er_E = 2K/z = 2V, \qquad (3)$$

where E is in V/mm and r_E in mm.

Equation (2) is the basic equation for magnetic mass analyzers. The geometry of a more realistic system is indicated in Fig. 16, where ions that diverge from a point source are focused to a point on the focal plane (confining attention to the plane of ion deflection). In this type of mass analyzer, a mass-to-charge spectrum can be obtained after accelerating all ions through the same potential (V) and placing an imaging detector along the image plane. While this method collects all ions in the spectrum, one can alternatively collect only those ions that pass through a narrow exit slit placed in the focal plane, while V or, alternatively, B is

Table 4. Performance characteristics of mass analyzers.[a]

Mass analyzer	Quantity directly measured	Ion energy (eV)	Upper mass limit (m/z)	Resolution at m/z = 1000	Dynamic range
Sector magnet	Momentum per charge	10^4	$>10^4$	$>10^5$	$>10^6$
Quadrupole mass filter	Path stability	10	$>10^3$	$>10^3$	$>10^4$
Ion trap	Path stability	1–10^3	$>10^4$	$>10^4$	$>10^3$
Time-of-flight	Time	10^3	$>10^5$	$>10^3$	$>10^3$
FT-ICR	Frequency	1–10^3	$>10^4$	$>10^5$	$>10^4$

---

[a] Adapted from Lambert *et al.* (1987).

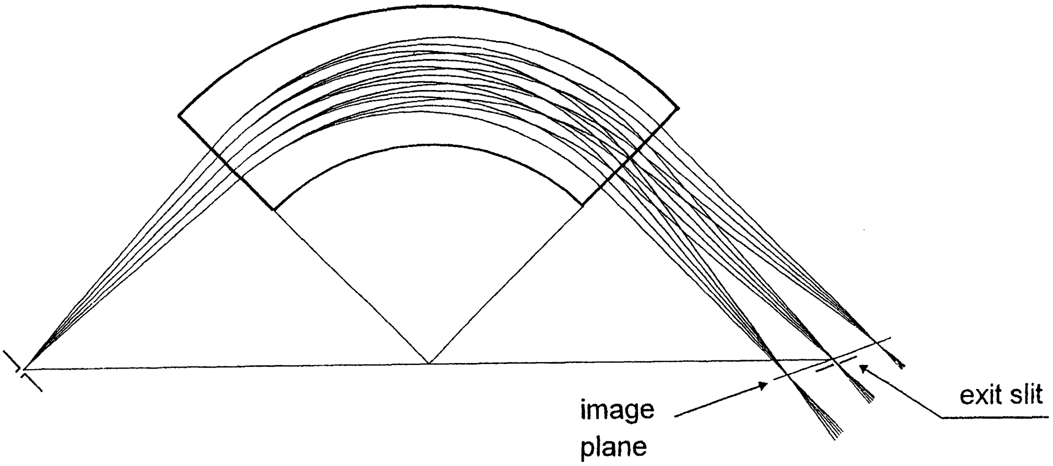

FIG. 16. Monoenergetic ions of three mass-to-charge values that leave a source divergently from a point are focussed by a sector magnet to three points on the image plane.

scanned. This allows ions of different m/z values to successively reach an ion collector placed behind the exit slit.

In a real mass analyzer, as shown in Fig. 16, ions of equal mass-to-charge ratios are perfectly focused only if they all have the same energy and leave the source from a single point. Since a real ion source produces ions with a certain energy distribution, $\Delta K = zV(1 \pm \delta_K)$, the available mass resolving power is always limited. However, a magnetic sector field can be combined with an electrostatic field to produce the high-resolution system shown in Fig. 17. Note that the trajectories for ions of two energies $zV(1 \pm \delta_K)$ and three masses, m_0 and $m_0(1 \pm \delta_m)$ are indicated. Further note that both sector fields are angle focusing and that the energy-to-charge dispersion of the electrostatic sector field is chosen such that it compensates for the energy-to-charge dispersion of the magnetic sector. Such a system is termed double focussing. It produces pure mass-to-charge dispersion independent of the ion energies and gives a mass resolving power $R_s = m/\Delta m$ independent of mass.

2.2.2 Quadrupole Mass Filters If a quadrupole is operated not by applying dc but rather by applying ac potentials, an ion that moves along the system axis experiences the same field distributions as would be produced by a large number of dc quadrupole lenses of alternating sign. If one applies an rf potential $\pm V \cos\omega t$ of angular frequency ω as well as a dc potential U to neighboring pairs of rods, i.e., a potential $U + V \cos\omega t$ to one opposing pair of rods and $-U + V \times \cos\omega t$ to the other, the voltage difference between neighboring electrodes is $2(U + V \times \cos\omega t)$. The ion motion under the influence of this two-dimensional field is described by the Mathieu equation:

$$\frac{d^2u}{d(\omega t/2)^2} + [a_u - 2q_u \cos(\omega t)]u = 0. \tag{4}$$

Here u represents either x or y in Cartesian coordinates, while a_u and q_u are parameters: $a_x = -a_y = -4zU/m\omega^2 r_0^2$ and $q_x = -q_y = 2zV/m\omega^2 r_0^2$. Depending on the magnitude of a and q, this equation has stable or unstable solutions, in which cases the ions either move periodically or diverge to infinity in the x and y directions. Figure 18(a) displays the results of solving Eq. 4 in one dimension in the form of a stability diagram that indicates, by shading, values of the parameters a and q in which stable solutions exist. Figure 18(b) illustrates ion trajectories in one dimension corresponding to two points in the stability diagram, one of which displays motion of a stable ion. In practice, ions will be subjected to forces in two dimensions as they drift down the rod assembly; only ions that undergo stable motion in both directions will remain within the device and be

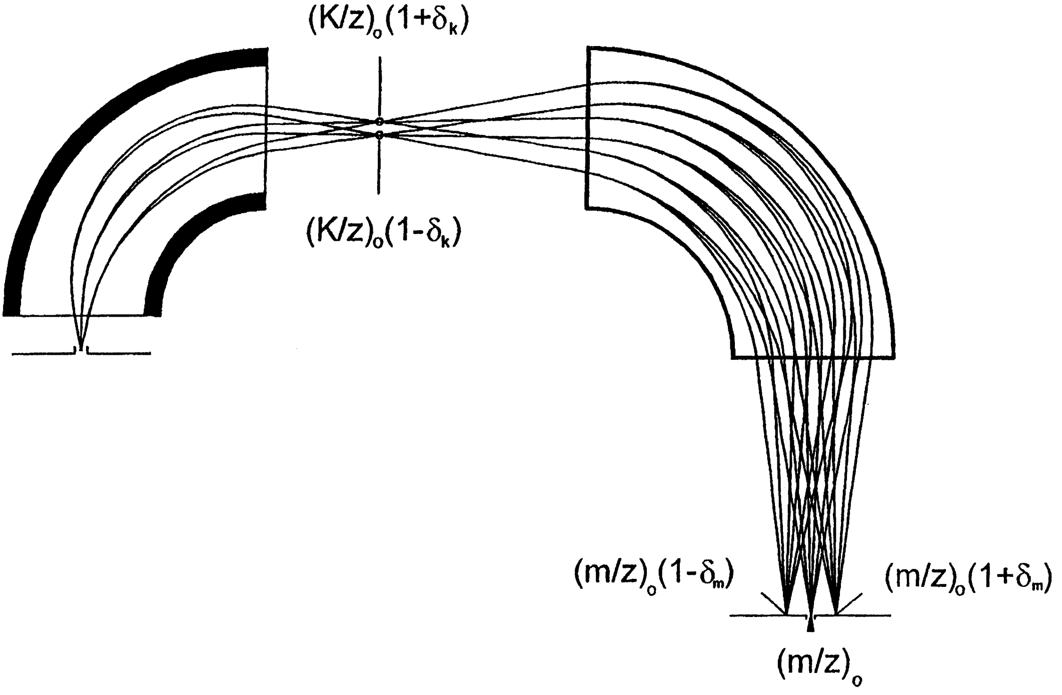

FIG. 17. Schematic diagram of a double-focussing mass spectrometer.

detected by an ion detector as they emerge from the quadrupole. In this experiment, ion motion parallel to the rod assembly is not tightly controlled since the ion separation, made according to mass, occurs only as a result of the field components perpendicular to the rod assembly.

Figure 19 shows an expanded region of part of the stability diagram in which both x and y stability are considered. Points of stability may be satisfied by many combinations of U, V, and m for a constant value of r_0 and ω. However, the shape of the stability region is such that a point near the apex of the diagram is intersected by a line of constant a/q ratio. For a given U-to-V ratio, ions of different mass occur at different points on this line, but only a narrow range of masses (near m_2 in Fig. 19) have stable motion. A mass spectrum may therefore be generated by scanning U and V with this fixed U/V ratio. Thus the mass scan line consists of the locus of values having a constant a/q ratio, and the mass analysis equation for a quadrupole mass filter operated in this way is

$$m/z = 2V/q_z\omega^2r_0^2, \tag{5}$$

showing that V is directly proportional to the m/z ratio of the ejected ions. Differentiating Eq. (5), one finds the mass resolving power R_s to be $m/\Delta m = (m/z)/\Delta V$, showing that for a given precision of V the mass resolving power increases with m/z. An ion transmitted through the array can be detected if its trajectory is characterized by particular values of a and q lying within the stable region, and this condition is established in turn for ions of increasing mass as U and V are scanned together. Typical operating conditions are $U/V = 0.17$, $\omega = 1.0$ MHz, and $r_0 = 1.0$ cm. The devices are typically 10 cm in length, ion energies along the axis are 5–20 eV, and residence times in the quadrupole array are tens of microseconds, corresponding to tens of cycles of ion motion.

Ions heavier than the stable ion, m_2^+, are discriminated against compared with lighter ions because they cannot respond sufficiently fast to the rapidly oscillating field. Ions lower in mass than m_2^+ follow the rapidly changing ac field and leave the device. The combination of high-mass and low-mass filters constitutes a mass analyzer. Resolution depends on the slope of the scan line. If the

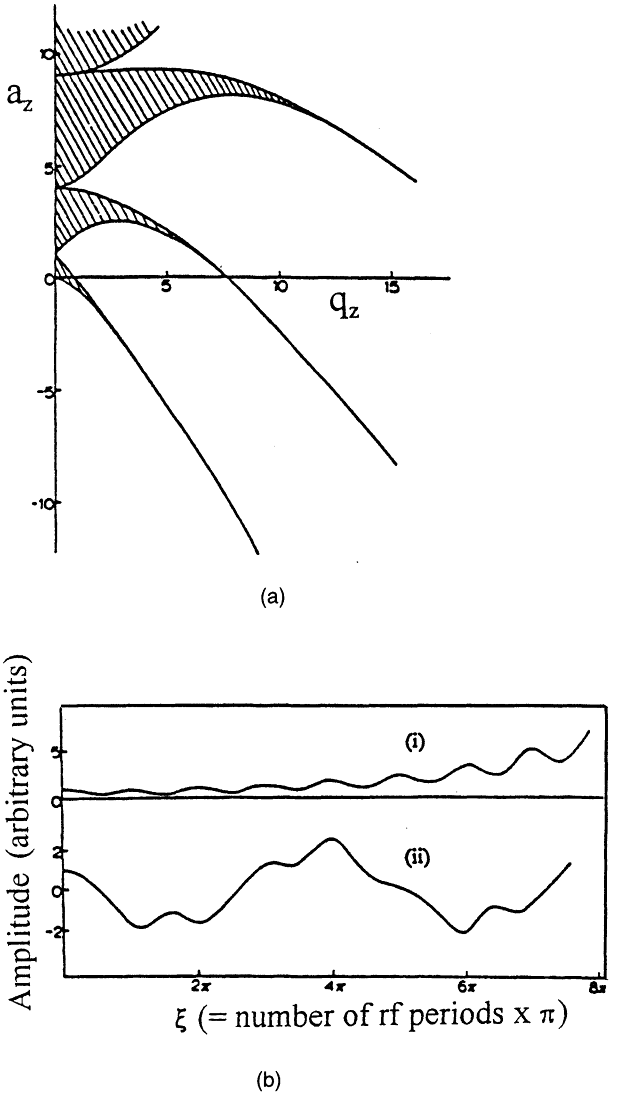

FIG. 18. (a) The *a*-*q* stability diagram displaying stable solutions to the Mathieu equation considering only a single coordinate direction. (b) Ion trajectories for ions with Mathieu parameters (i) $q_z = 0.55$, $a_z = -0.1625$ lying outside the stability region (unbounded motion) and (ii) $q_z = 0.55$, $a_z = 0$ lying within the stability region. (Dawson, 1976.)

scan line intersects the stability diagram very near the apex ($a = 0.23699$, $q = 0.70600$), higher resolution is achieved but at lower sensitivity.

Quadrupole mass spectrometers are small and inexpensive and have found wide application in surface science, leak detection, environmental monitoring, industrial process control, and on-line monitoring. For example, on-line monitoring of fermentations, the trace analysis of chlorinated compounds in drinking water, and the reaction kinetics of the formation of drug metabolites are among many studies reported using quadrupole mass spectrometers.

2.2.3 Quadrupole Ion Traps The quadrupole ion trap is a three-dimensional analog of the quadrupole mass filter. It consists of two end-cap electrodes of hyperbolic cross section that normally are operated at ground potential. Located between the end caps is a ring electrode to which an rf voltage is applied. In such a device a rotationally symmetric electric quadrupole field is generated because of the hyperbolic shape of the electrodes. Because the electric field is rotationally symmetric, it is convenient to consider only radial [$r = (x^2 + y^2)^{1/2}$] and axial (z) displacements. The potential ($\Phi_{r,z}$) at any point in this field is given by

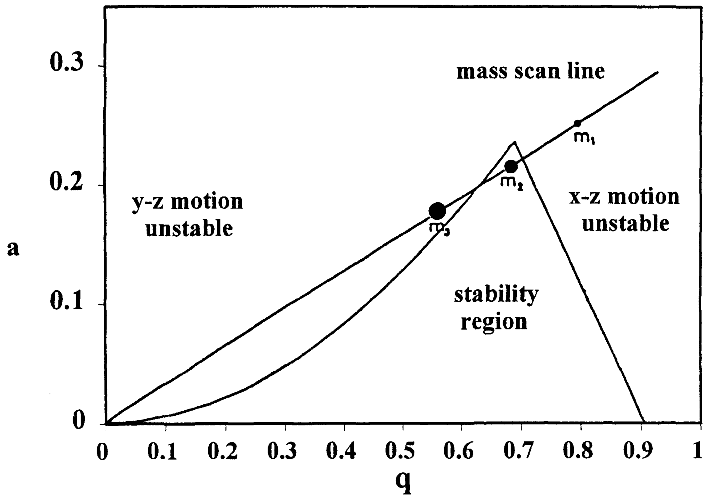

FIG. 19. Part of the Mathieu stability diagram for the linear quadrupole mass filter; m_1, m_2, and m_3 represent three singly charged ions of increasing mass. Only m_2^+ undergoes stable oscillations for the operating voltages represented by the *a* and *q* values shown. (Campana, 1980.)

$$\Phi_{r,z} = (U + V\cos\omega t)\left[\frac{r^2 - 2z^2 + 2z_0^2}{r_0^2 + 2z_0^2}\right], \qquad (6)$$

where r_0 is the internal radius of the ring electrode, z_0 is the closest distance from the center to the end-cap electrode, U is the dc potential and V is the rf potential (zero-to-peak) applied to the ring electrode, ω is its angular frequency, and t is time. In analogy to the two-dimensional quadrupole mass filter, ion motion is governed by the Mathieu equation with the coordinates a_z and q_z defining the operating conditions of the trap [$a_z = -2a_r = -16zU/m(r_0^2 + 2z_0^2)\omega^2$ and $q_z = -2q_r = 8zV/m(r_0^2 + 2z_0^2)\omega^2$]. Depending on the values of a_z and q_z, ions may fall in regions corresponding to simultaneous stability in both the r and z directions, as can be seen in the stability diagram shown in Fig. 20.

Normally, mass spectra are recorded by operating the quadrupole ion trap in a mass-selective instability scan mode in which ions are moved along the q_z axis by raising the rf voltage V until they become unstable at the boundary, where $q_z = 0.908$. They then exit the trap through a hole in the end-cap electrode. Thus ions of increasing m/z are ejected and detected as the rf voltage V is raised. For a quadrupole ion trap operated in this mass-selective instability mode, the mass analysis equation is

$$m/z = 8V/q_z\omega^2(r_0^2 + 2z_0^2), \qquad (7)$$

which is very similar to Eq. (5). Here, too, the mass resolving power increases with m/z. Typical operating conditions are $V_{0-p} = 7500$ V, $r_0 = 1.0$ cm, and $\omega = 1.1$ MHz. The maximum m/z range that can be achieved under these conditions is about 650 atomic mass units/charge. The simplest way to extend the mass range of an ion trap is to cause the trapped ions to become unstable at a lower q_z value by resonant ejection. From Eq. (7) it is clear that means of achieving mass range extension include operation at higher values of the rf voltage V, use of lower angular frequencies (ω), or use of a smaller trap (r_0). A mass/charge range in excess of 10^4 atomic mass units/charge, very high resolution, and very high sensitivity are characteristic features of the ion trap, though one should mention that an ion trap can typically only hold up to about 10^6 ions before Coulombic effects greatly reduce the achievable mass resolving power.

Trapped ions have characteristic frequencies of oscillation, known as secular frequencies. The principal component of these secular frequencies is $(\omega/2)\beta$ radian per second, where β is a parameter that varies with the coordinates a and q. At low values of a and q, β_z is approximately given by $(a_z + q_z^2/2)^{1/2}$. The presence of a helium buffer gas inside the ion trap removes kinetic energy from the ions by collisions, causing them to occupy the central region of the trap. Optimum conditions are found at a helium pressure of about 1 mTorr. Figure 21(a) shows a simu-

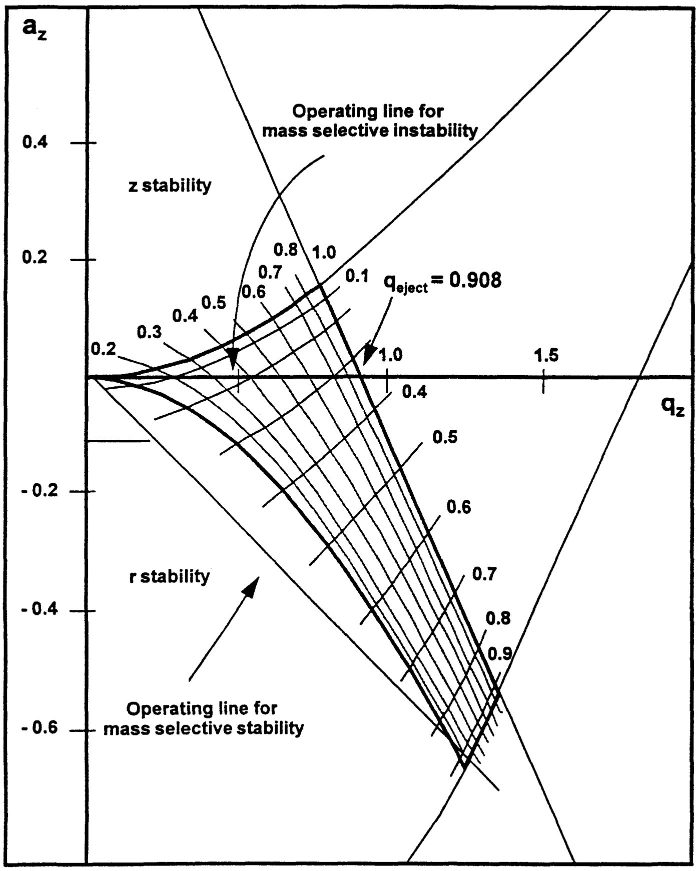

FIG. 20. The Mathieu stability diagram for the quadrupole ion trap. The common operation mode is the mass-selective instability scan in which the rf potential is raised to increase the value of q_z, while $a_z = 0$.

lation of the trajectory of an ion of $m/z = 100$ in an ion trap operated at an rf voltage of 500 V and a frequency of 1.1 MHz. It can be seen that the center of the trap is not occupied by the ion; instead, a "donut" of space is accessed. When a supplementary ac potential, in resonance with the frequency of ion motion in the z direction, is applied across the end-cap electrodes, ions are resonantly ejected from the trap. Figure 21(b) shows the simulation of the ion trajectory in the presence of this ac potential. It can be seen that there is no effect on ion motion in the r direction. However, the excursion in the z direction increases, and the ion is ejected and reaches the end-cap electrode after just a few cycles of application of the ac potential. This method of resonance ejection is used in order to extend the mass-to-charge ratio of the ion trap.

Quadrupole ion traps are small, inexpensive, and highly sensitive instruments. Like quadrupole mass filters, ion traps have found wide applications in environmental, industrial, and on-line monitoring. Their low cost and high sensitivity lead to their wide use as detectors for liquid and gas chromatography, and their small size makes ion traps suitable for field analysis (March, 1989).

2.2.4 Fourier-Transform Ion-Cyclotron-Resonance Mass Spectrometers The operating principle of FT-ICR is that an ion in a strong magnetic field undergoes cyclic motion with an angular frequency ω_c in a plane perpendicular to the direction of the field. This frequency is calculated from the fact that the magnetic force Bzv is balanced by the centrifugal force mv^2/r_B so that

$$\omega_c = \frac{v}{r_B} \approx \frac{13.89\sqrt{K/m}}{(0.144/B)\sqrt{Km/z^2}} \approx \frac{96.485Bz}{m}, \qquad (8)$$

where ω_c is the ion-cyclotron frequency in

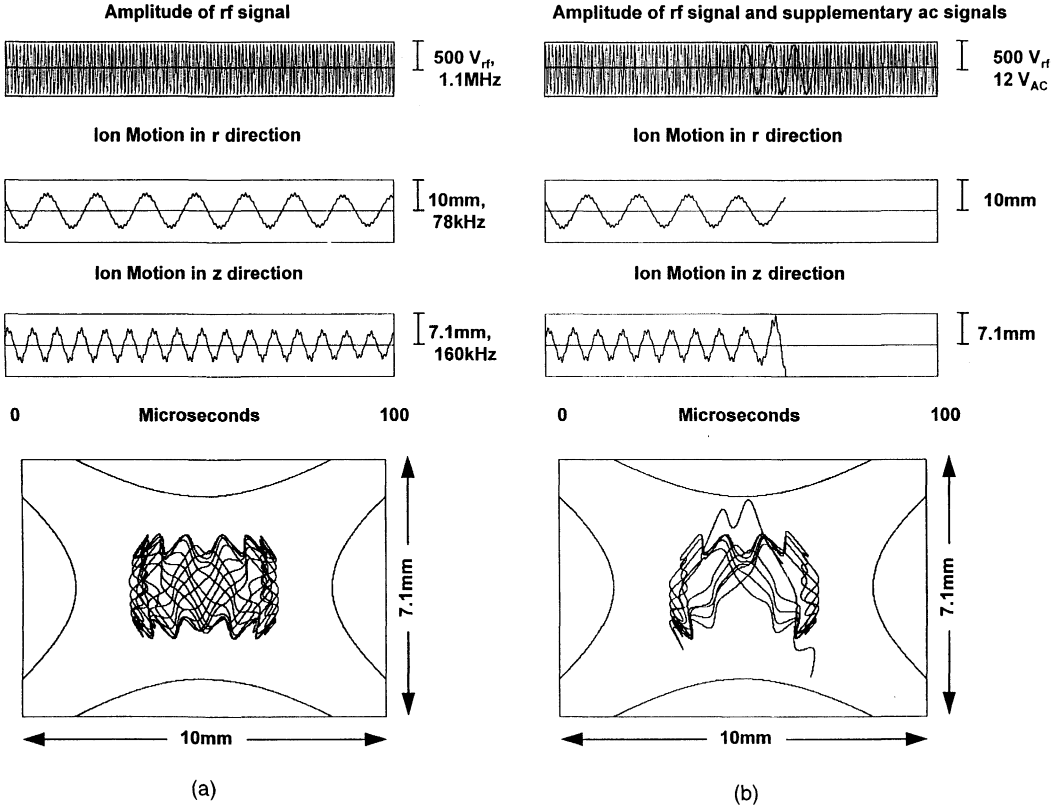

FIG. 21. (a) A simulation of the trajectory of an ion of *m/z* 100 in an ion trap operated at a rf voltage of 500 V and a frequency of 1.1 MHz. (b) The same simulation in which a supplementary ac voltage is used to resonantly excite ion motion. (Cooks *et al.*, 1991.)

MHz, B is in teslas, and m is in atomic mass units with v and r_B taken from Eqs. (1) and (2). The remarkable result is that all ions rotate with frequencies that depend on their mass-to-charge ratios (m/z) but are independent of their energies K, and this feature was put to use in the construction of the cyclotron accelerator by E. O. Lawrence. If the magnetic field strength B is known, measurement of the ion-cyclotron frequency ω_c suffices, in principle, to determine the m/z ratio. Since frequencies can be measured with high precision, mass-to-charge ratios of ions can be determined very accurately from Eq. (8). Besides cyclotron motion of an ion in the ICR cell, the ion also undergoes magnetron motion as a result of a small electric field applied to the two end plates T (Fig. 22). This magnetron motion is a slow—typically a few hundred hertz—circular drift of an ion along a path of constant electrostatic potential. In typical FT-ICR mass spectrometry ex-

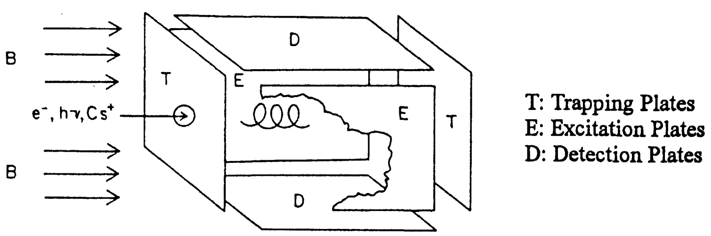

FIG. 22. Schematic diagram of a cubic ICR cell. The method of ionization is indicated by e^-, $h\nu$, Cs^+. (Marshall, 1985.)

periments, the magnetron frequency is much smaller than the cyclotron frequency, and it is often neglected.

Figure 22 shows a schematic of an ion cyclotron cell in which ions are trapped. The magnetic field, *B*, constrains the ions from escaping anywhere in a plane perpendicular to the magnetic field, and a small static trapping voltage (+1.0 V) applied to the two end plates, *T*, prevents the ions from escaping in a direction parallel to the magnetic field. In order to detect the ions, it is necessary to set them in phase-coherent motion by applying to the excitation plates *E* an rf voltage having the cyclotron frequency of the ions. Even after termination of the applied rf voltage, the coherent circular motion of the resonant ions creates in the electrodes an alternating image current, the amplitude of which decreases exponentially with time. The decay of the image current is a result of momentum loss of the ions through collisions with residual gas atoms or molecules. The image current provides a time-domain signal (Fig. 23). If the excitation is performed using a signal containing many frequencies that excite many ions at their respective cyclotron frequencies, the time-domain signal will contain multiple cyclotron frequencies. This signal can be Fourier transformed (FT) into the frequency-domain signal in which each frequency is related to a particular value of *m*/*z* by Eq. (8). Figure 23 displays the time-domain and frequency-domain ion-cyclotron signals for CO^+ and N_2^+. One advantage of simultaneous detection is a significant increase in signal-to-noise ratio.

Differentiating Eq. (8), one finds the mass resolving power R_s of an ICR mass spectrometer to be

$$R_s = \frac{m}{\Delta m} = \frac{96.485Bz}{m\Delta\omega}; \tag{9}$$

$\Delta\omega$ is usually defined as the full spectral peak width at half-maximum peak height. Hence, the mass resolving power decreases with *m*/*z* but increases with *B*, so that commercially available superconducting magnets of field strength of 3–7 T are advantageous. High mass resolving power requires ultrahigh vacuum (below 10^{-8} Torr); otherwise collisions with residual gas atoms and molecules too rapidly degrade the coherent motion of the ions. However, some ionization methods such as electrospray and thermospray require a relatively high pressure. This problem is solved by injecting ions from an external ion source into a differentially pumped ICR cell. Recent methods of injecting ions from external ion sources have led to the coupling of ICR instruments with a variety of chromatographs. In all cases, only a limited number of ions, typically 10^6, can be present in the trapping cell without Coulombic repulsion between the individual ions seriously affecting ion motion and degrading the resolving power and mass measurement accuracy.

The resolving power of FT-ICR is sufficiently high at low mass to allow one to distinguish (Fig. 24) between a positively and a negatively charged ion that differ only by the mass of two electrons. The available high resolving power is particularly valuable in conjunction with electrospray ionization in

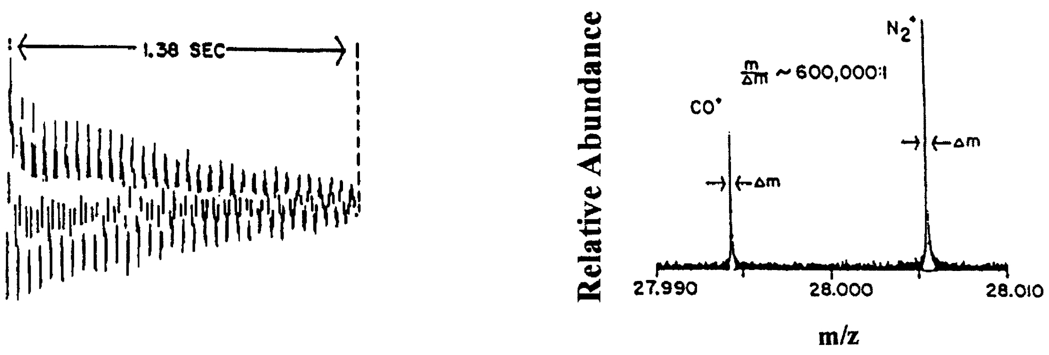

FIG. 23. Time-domain (left) and frequency-domain (right) ion-cyclotron signals for ions of two *m*/*z* ratios (Marshall, 1985.)

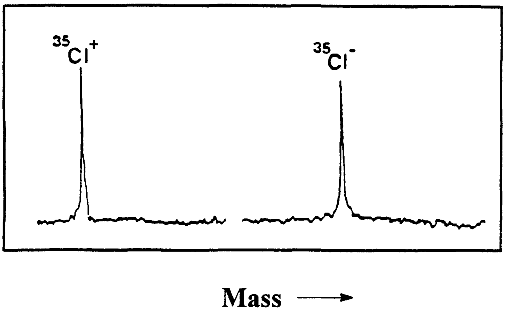

FIG. 24. Mass spectrum of $^{35}Cl^+$ and $^{35}Cl^-$ at a pressure of 10^{-8} Torr and a mass resolution of ~1.5 × 10^6. The Cl^+ and Cl^- ion signals were alternately acquired, stored, and transformed and are presented on a single mass scale. (Allemann *et al.*, 1982.)

which the carbon-isotope peaks in biomolecules are resolved; i.e., mass peaks are separated due to molecules that contain one or several ^{13}C instead of ^{12}C atoms. From the spacing between adjacent isotopic peaks in large molecules, the charge state of multiply charged ions can be determined. Figure 25 shows part of an electrospray ionization spectrum of a mixture of cytochrome c and myoglobin (note that the portion of the spectrum shown occupies a single *m*/*z* unit). The average spacing reveals that the observed ions are due to the 13+ and 18+ charge states of cytochrome c and myoglobin, respectively. That is, they represent intact molecules to which 13 and 18 protons have been added, respectively. The statistical variations in ^{13}C content of large molecules account for the relative abundances of ions with various *m*/*z* values. Multiplying the observed *m*/*z* by the charge state and correcting for the number of added protons yields the molecular weight.

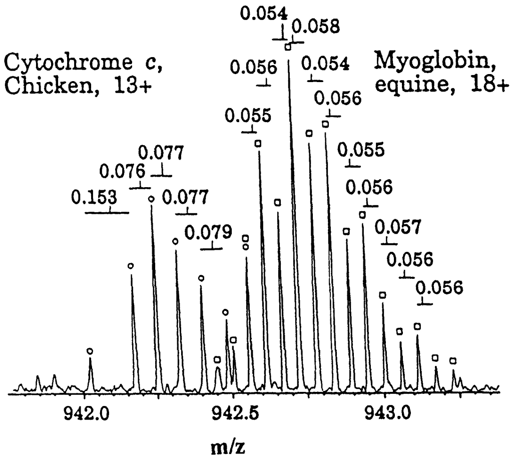

FIG. 25. Single-scan ESI spectrum of the 13+ charge state of cytochrome *c* (circles) and the 18+ charge state of myoglobin (squares). Measured mass-to-charge ratio differences of the isotopic peaks are indicated by values on the bars. (Beu *et al.*, 1993.)

In an FT/ICR, no slits or lenses are required, so that the loss of ions is minimized. Sensitivity is further enhanced by the fact that ion detection is nondestructive and by the fact that a wide range of *m*/*z* values can be measured simultaneously. Because detection does not destroy the ion population of the trap, re-excitation and remeasurement are possible. This step can be repeated hundreds of times with very high efficiency. An ICR cell can trap ions for many seconds, which makes it useful for the study of ion–molecule reactions. The high mass resolving power, R_s, facilitates accurate mass measurement.

2.2.5 Time-of-Flight Instruments In a TOF mass spectrometer, one determines the ion flight time t in μs as it passes through a distance d in mm, where

$$t = \frac{d}{v} = \sqrt{\frac{m}{2zV}}\,d = \frac{d}{13.89}\sqrt{\frac{m}{K}} \tag{10}$$

and v is taken from Eq. (1). This relationship gives the mass-to-charge ratio (m/z) in atomic mass units as

$$m/z = 193t^2V/d^2, \tag{11}$$

where $V = K/z$ is the ion acceleration voltage in volts. Differentiating Eq. (11), one finds the mass resolving power of a TOF mass spectrometer to be

$$R_s = m/\Delta m = t/2\Delta t. \tag{12}$$

The mass resolving power is independent of ion mass but, for a given uncertainty in time Δt, the mass resolving power increases with $(m/z)^{1/2}$. Mass resolving power can be improved by increasing the flight time t, which can be achieved by using lower accelerating voltages or longer flight distances d.

In spite of the simplicity of the method in principle, in practice high performance requires consideration of the effect on the flight time of differences in initial velocity (i.e., kinetic energy spread), initial direction of motion (i.e., angular spread), and initial position at which ions are formed. A well-known method that compensates for velocity differences is to have the ions undergo reflection in a repeller field. The more energetic ions penetrate more deeply into the field so that their flight paths are elongated. By adjusting this field strength appropriately, the overall ion flight time becomes energy independent (Fig. 26). Instead of a homogeneous repeller field, magnetic or electrostatic sector fields can also be used. Additional angular focusing is often achieved by rotationally symmetric lenses (see Fig. 13) or by grid-free ion reflectors (see Fig. 26). The dispersion in ion arrival time due to differences in initial position of ionization is corrected for by generating the ions in an electric field such that those formed further from the detector are accelerated toward it more strongly. Time-focusing of ions of the same mass occurs at the detector focal plane.

In all TOF work, ions must be produced in bursts. Thus, TOF systems are best coupled with ion sources that naturally produce ions in pulses. These include plasma desorption, pulsed SIMS, pulsed laser desorption, and other desorption ionization techniques. However, one can also produce ions continuously, store them for up to 1 ms, and extract them using a voltage pulse of perhaps 1 μs duration and then bunch them in time to perhaps 10 ns. This bunching can be achieved by using the Wiley–McLaren principle by which the leading ions are accelerated to lower energies than the tailing ions. Alternatively, one can produce a continuous low-energy ion beam and apply an appropriate voltage pulse to accelerate these ions at right angles to their initial flight direction.

The predominant method of data collection in TOF mass spectrometry is the integrating transient recording (ITR) in which the arrivals of ions in short time intervals are stored in consecutive recording channels of a multichannel analyzer. Adding several of these spectra achieves relatively noise-free mass spectra. When signals are much weaker, the flight times of individual ions can be determined from their start and stop signals. This method usually achieves better time resolution but requires that the ion intensities be so low that the individual ion ar-

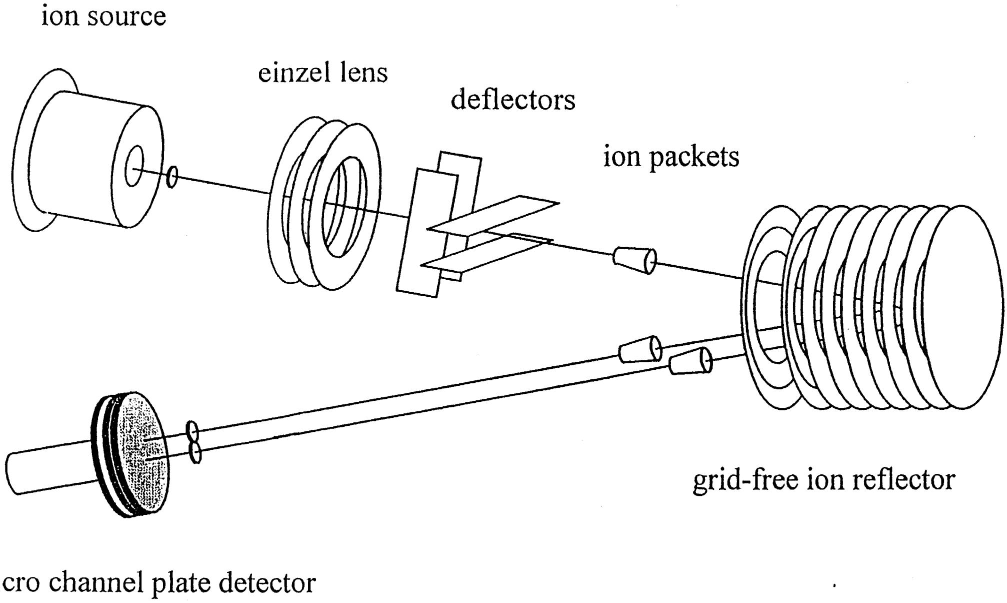

FIG. 26. A reflector is used to compensate for differences in ion kinetic energies and achieve temporal focusing at the detector in TOF mass spectrometry.

rival events not be coincident. In most cases such systems involve a multistop time-to-digital converter (TDC).

The outstanding feature of a TOF/MS is its high mass range ($>10^5$ atomic mass unit/charge). This property has allowed the measurement of molecular weights of proteins in excess of 500 000 atomic mass units using matrix-assisted laser-desorption ionization. However, TOF mass analyzers also have the valuable capability of producing mass spectra of good quality within a short time (a few milliseconds). This makes them ideal for recording mass spectra of time-varying samples as is required in process-control systems and kinetic measurements. Together with their high sensitivity, this feature is also important for their use as mass-sensitive recorders for monitoring the effluent in fast gas chromatography or capillary-zone electrophoresis experiments.

2.3 Ion Detection

When energetic ions impinge on a suitable metal surface, the electron cloud of the ions interacts violently with the electrons of the surface atoms. As a result of this interaction, multiple secondary electrons are released with energies of a few eV. These secondary electrons then can be accelerated and guided from the conversion dynode to an electron-multiplier detector (Fig. 27). In the electron multiplier, N electrons impinge on a surface releasing $\approx 2N$ electrons (the factor 2 being nominal and being chosen only for purposes of illustration). Accelerating these electrons by ≈ 200 V and allowing them to impinge on another surface yields $\approx 4N$ secondary electrons, and by repeating this process $\approx 8N$, $\approx 16N$, etc. Thus the initial number of electrons N is finally multiplied by a large factor. The most common electron multiplier is a continuous-dynode device known as a channeltron across which a potential of several keV is applied. Provided the ion arrival rate is slow ($<10^6$ ions/s), one can record the arrival of each ion on the converter electrode by amplifying the electron multiplier signal. This ion-counting method was introduced by White in the 1950s. In the case of negatively charged ions, a conversion dynode must be used (Fig. 27), but its electrical bias is changed so as to collect positively charged ions generated at its surface by sputtering and other processes. Another detector is the channel plate, which is made up of an array of small (30 μm) individual electron-multiplier elements. The channel plate has a time resolution of about 10^{-10} s, which is much smaller than that of a conventional electron multiplier. Its spatial resolution is an additional advantage, and gains are up to 10^6. The channel plate can be very useful as a beam visualization device when it is operated closely coupled to a phosphor for electron-to-photon conversion. The light from such a phosphor can fall on a vidicon

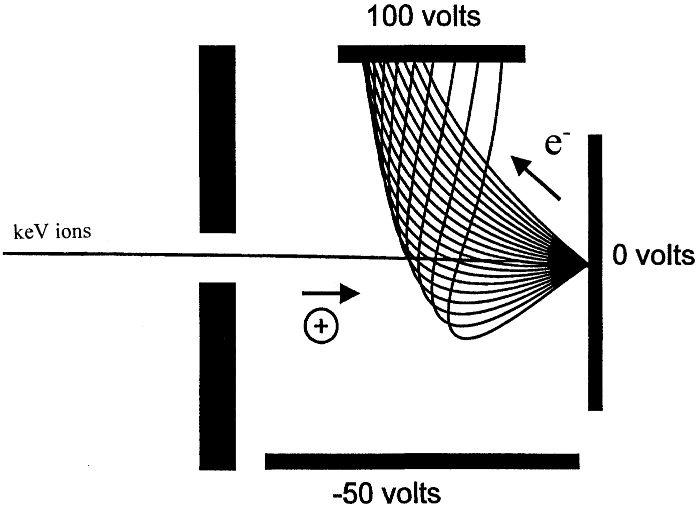

FIG. 27. An ion detector that records incoming ions by the released secondary electrons. The potentials shown give only a rough impression of the required voltages. Analogous procedures are used to record negatively charged ions.

or other photosensitive element to record the ion-beam image. The signal-to-noise ratio for the channel-plate detector is improved over single-channel detectors because of the Felgett advantage.

The release of secondary electrons from a surface is velocity dependent, making it difficult to record very heavy molecular ions. Such ions can be recorded, however, in a two-step process. The ions first impinge on a surface (such as that of the conversion dynode in Fig. 27) where the impact breaks the molecules up into fragments (surface-induced dissociation). The ionic fragments are then accelerated toward a second surface, where these smaller and faster ions form secondary electrons upon impact. In addition to the common electron multipliers, there are several other detectors that are used in mass spectrometry. Photoplates are still occasionally used as imaging detectors. They are sensitive to as few as 50–100 ions. One illustration is shown in Fig. 3. A modern charge-coupled device (CCD) also behaves as an imaging detector. Faraday collectors accept ions quantitatively while preventing loss of any electrons or ions produced. They are used in quantitative work especially in isotopic analysis and can have precisely unit gain.

3. SPECIAL TECHNIQUES AND THEIR APPLICATIONS

3.1 Applications of Electron Capture

The lightly ionized plasma in the CI ion source (Sec. 1.3) contains approximately equal numbers of positive and negative particles, ion–molecule reactions or attachment of thermal energy electrons to molecules yielding negatively charged sample ions. Reagent gases that produce thermal-energy electrons allow electron-attachment chemical ionization to be performed. This method is extremely sensitive (i.e., gives low detection limits) because of the high electron mobility and the large cross section for electron capture by suitable compounds. Often a functional group (e.g., pentafluorophenyl) with high electron affinity is added to a molecule to enhance its sensitivity to electron capture. High selectivity is also obtained since most compounds do not give stable anions upon electron capture. Figure 28 shows a negative chemical ionization mass spectrum of the environmentally important compound 3,3′,4,4′-tetrachlorodiphenyl ether. The most abundant peaks are due to isotopic forms of the intact molecular anion M^- at *m/z* = 306, 308, 310, and 312 containing from zero to three ^{37}Cl atoms. Fragmentation with the loss of an HCl molecule from this anion produces the cluster of peaks centered at *m/z* = 272. The small peak at *m/z* = 161 is due to the product of structurally diagnostic oxygen-phenyl cleavage in the molecular anion. The characteristic isotope ratios of each group of peaks facilitate the identification of the number of chlorine atoms in each fragment (e.g., the natural isotopic abundance ratio for Cl_2 and molecules containing Cl_2 is 9:6:1). The electron-attachment method is used in forensic science, in environmental

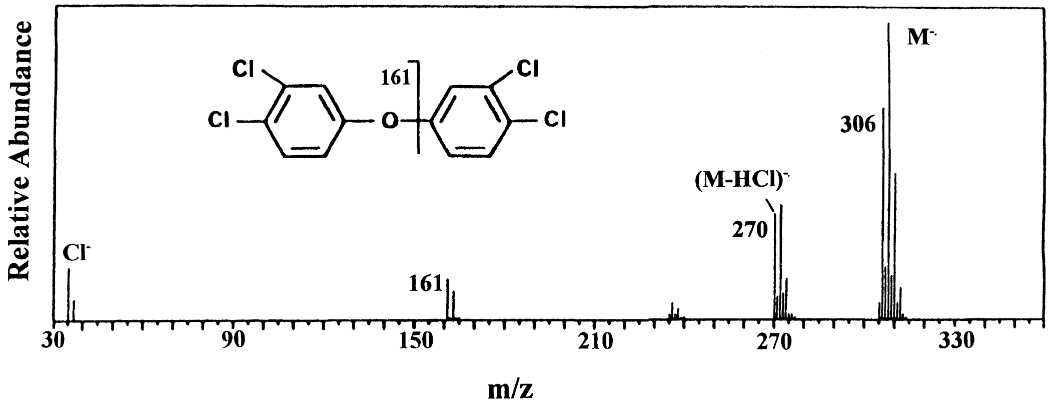

FIG. 28. Electron-attachment chemical-ionization mass spectrum of 3,3′,4,4′-tetrachlorodiphenyl ether using methane to moderate electron energies (ion source temperature = 250°C) (Stemmler and Hites, 1988).

monitoring, and particularly in drug metabolism where selective, ultratrace-level determinations are important and the matrices in which compounds of interest occur are very complex.

3.2 Chromatography/Mass Spectrometry

Chromatography/mass spectrometry is a group of techniques in which the separating power of chromatography is combined with the specificity of mass spectrometry in analyzing particular compounds present in complex mixtures. The chromatography step also serves to preconcentrate analytes and deliver them in a continuous automated fashion for examination by the mass spectrometer. The mass spectrometer can be operated as a universal detector, by recording the total ion abundance in the mass spectrum as each compound elutes from the chromatograph. Alternatively, it can be operated as a selective detector, by providing information on one type of compound through examination of a few characteristic ions, or even just one, in the spectrum.

Selective detection is often performed by recording an ion chromatogram, which displays the abundance of a selected ion(s) as a function of elution time. The total ion chromatogram, on the other hand, is the sum of the individual ion chromatograms (i.e., the total ion current at each point in time) and represents the chromatogram as measured using the total ion abundance, without regard to types of ions. Maximum sensitivity is achieved by limiting the mass-spectrometer scan to one ion, viz., by selected-ion monitoring (SIM), where a single ion of interest is monitored continuously by the mass spectrometer. The signal-to-noise ratio of this experiment is increased greatly compared to an experiment in which much of the analysis time is spent in monitoring ions that provide less or no information on the analyte. SIM trades specificity for sensitivity, and detection limits in SIM can be as low as 100 femtomoles. This is an improvement over full–mass-scan mass spectra by a factor of 100 to 1000, an improvement in sensitivity that is crucial in applications where low levels of specific compounds are to be recognized and quantified.

There are two major types of chromatography/mass-spectrometry instruments: gas chromatography/mass spectrometry (GC/MS) and liquid chromatography/mass spectrometry (LC/MS) instruments. In interfacing a chromatograph to a mass spectrometer, one must maintain chromatographic performance and achieve high transfer efficiency. Three types of GC/MS interfaces are used: direct coupling, splitters, and molecular separators. The interface for the LC/MS combination has similar basic requirements to GC/MS, viz., LC integrity must be maintained without significant loss in efficiency. Table 5 summarizes the main types of LC/MS interfaces.

Both GC/MS and LC/MS play important roles in environment monitoring and life science research. For example, on-line LC/ES/MS is a significant method for the analysis of peptides and other biopolymers. Figure 29 shows the total-ion chromatogram recorded in an LC/ES/MS analysis of a tryptic (enzy-

Table 5. Summary of LC/MS interfaces.

Interface	Ionization mode	Maximum flow rate (μl min^{-1})	Characteristics
Moving belt	EI/CI/FAB	2000	Limited to volatile organic mobile phases, excellent enrichments and efficiency; memory effects
Direct liquid introduction	EI/CI	50	Mechanically simple, no memory effect; inefficienct
MAGIC	EI/CI	500	Limited to relatively volatile compounds; not very sensitive; good for EI
Thermospray	TS	2000	Used for volatile or nonvolatile compounds; is also an independent ionization method in MS
Electrospray	ES	10	Interface for LC/MS and CZE/MS; multiply charged ions; high-MW determination; also independent ionization method

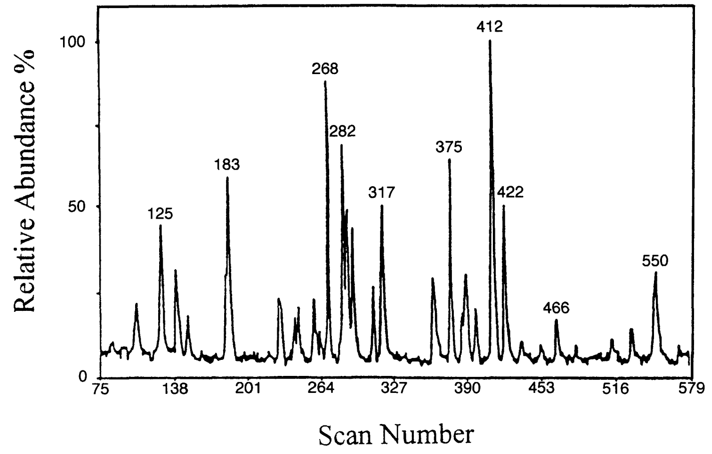

FIG. 29. Total-ion–current chromatogram of an LC/ES/MS analysis of a tryptic digest of reduced and alkylated glycoprotein (Carr *et al.*, 1991).

matic) digest of a glycoprotein. During the acquisition of these data, hundreds of mass spectra were recorded. A representative spectrum is shown in Fig. 30. This type of data is valuable in the recognition of individual glycopeptides; the compound eluting at the time this scan was taken clearly has a molecular weight of 3260 atomic mass units. With the application of tandem mass-spectrometric techniques in conjunction with GC/MS and LC/MS, more structural information on individual compounds can be obtained (see below). For example, sequence information on peptides can be acquired by capillary-zone electrophoresis (CZE) combined with ES/MS/MS.

3.3 Mass Spectrometry/Mass Spectrometry

Mass spectrometry/mass spectrometry (MS/MS), or tandem mass spectrometry, is a procedure for examining individual ions in mixtures of ions. The ions often serve as surrogates for the corresponding neutral molecules, making MS/MS a formidable method of analysis of complex mixtures. The ions of interest are separated by mass and allowed to dissociate upon collision, and the resulting product ions are examined by a second mass measurement. Figure 31 illustrates the concept.

In MS/MS experiments, two mass analyz-

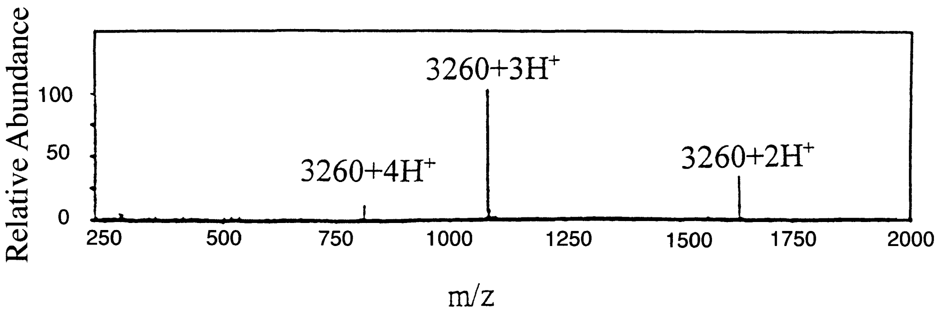

FIG. 30. Representative ES mass spectrum from the LC/ES/MS analysis shown in Fig. 29 (scan 550), showing a compound of molecular weight 3259.7 atomic mass units with two, three, or four charges. The mass assignments are given in parenthesis and are accurate to a single atomic mass unit. (Carr *et al.*, 1991.)

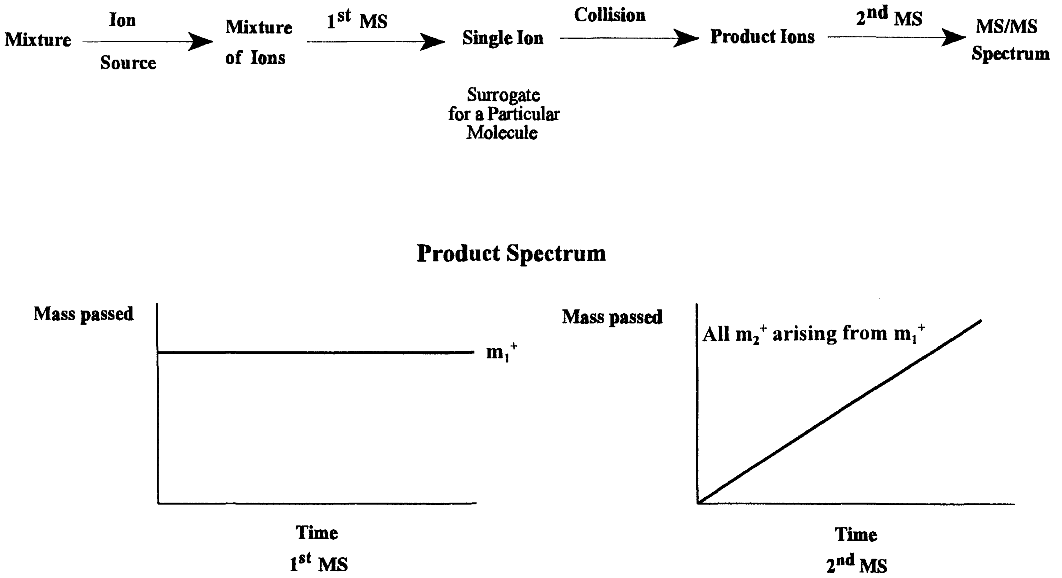

FIG. 31. Concept of tandem mass spectrometry, showing scan (product-ion spectrum) that characterizes particular compounds in mixtures.

ers (or two mass-analysis events) are used to measure the parent and product ions, m_1^+ and m_2^+, explicitly. The first mass analyzer is usually set to pass m_1^+, which then undergoes dissociation to give product ions, m_2^+. Ion dissociation in MS/MS can be achieved by collision with a target gas (collision-induced dissociation, CID), collision with a surface (surface-induced dissociation, SID), or photodissociation (PD), which is a process in which an ion absorbs one or more photons of appropriate wavelength and then dissociates. The cross sections for CID and single-photon photodissociation are about 10–200 and 10^{-2} Å^2, respectively, making collisions with gaseous targets a simple and successful way to excite and dissociate ions. Ion dissociation in MS/MS is represented by the following equation:

$$m_1^+ \to m_2^+ + m_n.$$

Since mass spectrometers naturally can only analyze charged particles, the neutral molecular fragment m_n is not usually detected. But the value of m_n can be inferred from m_1^+ and m_2^+.

Historically, MS/MS arose from the fact that metastable ions, which dissociate in flight in sector instruments prior to mass analysis, give products that appear in a normal mass spectrum at an apparent mass of m^*, where $m^* = m_2^2/m_1$. The positions of the peaks due to these processes provide information on the masses of both the parent ion, m_1^+, and the product ion, m_2^+. For example, benzene shows a metastable peak at an apparent $m/z = 34.7$ due to the following reaction:

$$\begin{array}{lll} C_6H_6^{+\cdot} \to & C_4H_4^{+\cdot} & + C_2H_2 \\ m/z\ 78 & m/z\ 52 & (m^* = 34.7)\,. \end{array}$$

The metastable ion peak usually appears in a sector mass spectrometer as a nonintegral and broad mass peak, and it provides valuable information when interpreting fragmentation patterns.

The product ions formed in MS/MS experiments are usually recorded by scanning the second mass analyzer. This type of scan, a product-ion scan, is very useful in elucidating the structure of the parent ion, which is often an ionized form of the intact analyte molecule. This can be seen for a very simple example in Fig. 32, where a comparison is made between the MS/MS spectrum of an ion of $m/z = 143$, derived from a sample of diesel particulates, and the MS/MS spectrum of authentic protonated 2-methylnaphthalene

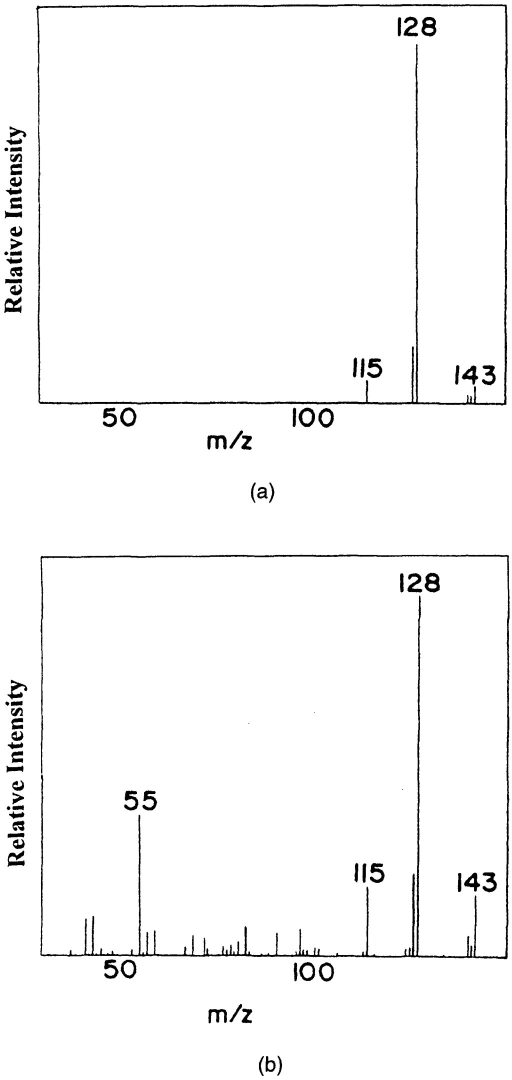

FIG. 32. MS/MS product-ion spectrum of *m/z* = 143 from (a) authentic, protonated 2-methylnaphthalene. (b) A diesel particulate. (Wood *et al.*, 1982.)

(also *m/z* = 143). The agreement between the high-mass fragment ions confirms the presence of methylnaphthalene isomers as a major constituent at *m/z* = 143 in the diesel particulate sample. The ions present at low *m/z* indicate the presence of at least one other component that also gives an ion at *m/z* = 143.

There are other MS/MS scan types. A parent-ion scan is one in which the product ion is selected and a scan is performed to detect all parent ions that fragment to yield this particular product. This scan can be used to identify members of a class of compounds defined by their characteristic product. For example, all alkylnaphthols in a mixture can be identified by setting the second mass analyzer to a fragment ion (*m/z* = 157) characteristic of this class of compounds and scanning the first analyzer so as to record the molecular weights of all compounds that fragment by the following process:

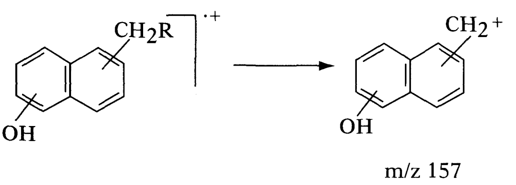

A neutral-loss scan is one that identifies all parent ions formed from a complex mixture that fragment by loss of a particular neutral molecule. Like the parent-ion scan, it is useful in identifying the presence (and molecular weights) of groups of compounds containing a designated functional group. For example, a neutral-loss scan of 44 atomic mass units in a negative ion MS/MS experiment indicates the presence of carboxylic acids, recognized via their characteristic loss of carbon dioxide (mass 44) from the $[M-H]^-$ ions.

MS/MS experiments can be performed with a wide variety of instruments. Multiple analyzer sector, quadrupole, and TOF instruments are all used. Figure 33 shows a schematic diagram of one such instrument, a triple-quadrupole mass spectrometer. The central quadrupole (*Q*2) is operated in the rf-only mode, which allows all ions to pass through, and is used as the collision region. The scanning procedures used for the acquisition of the various types of MS/MS data are straightforward. A product-ion scan is obtained by simply setting the first quadrupole to pass the desired parent ion and scanning the last quadrupole (*Q*3) in the normal manner. Parent-ion scans are performed by setting the last quadrupole (*Q*3) to pass the desired product ion while the first quadrupole *Q*1 is scanned in the normal manner. Neutral-loss spectra are obtained by scanning the first and last quadrupole mass analyzers in concert at the same rate, the offset in the masses of the ions passed by the

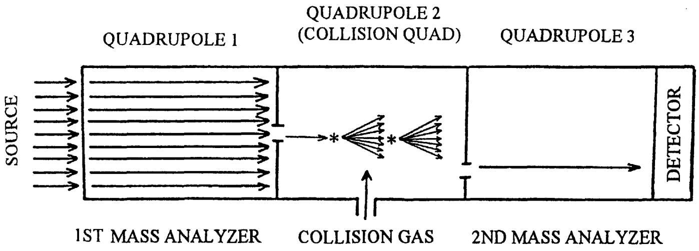

FIG. 33. Operation of the triple-quadrupole mass spectrometer as a MS/MS instrument. The parent scan mode is illustrated.

analyzers corresponding to the desired neutral loss.

Ion-trapping instruments (FT-ICR and quadrupole ion traps) can be used to perform MS/MS experiments, and they require only the use of an additional sequence of operations in the scan function. In the case of the ion trap, the scan function begins with ionization followed by the selection of a parent ion in a step that involves ejecting all other ions from the ion trap. The parent ion is then translationally excited, typically by applying a supplementary rf voltage to the end caps. The product ions resulting from collision-induced dissociation are recorded by scanning the rf voltage to perform a second mass-analysis scan.

The main advantage of MS/MS experiments is their enhanced specificity. This is useful in isomer distinction, biological-molecule sequencing, and the analysis of complex mixtures. For example, a parent ion can be separated from a complex mixture in the first analyzer and then dissociated and its products analyzed by a second analyzer. The MS/MS experiment is often done in conjunction with chromatographic separation.

3.4 Mass Spectrometry for Analysis of Biopolymers

Tandem mass spectrometry can be used to obtain structural information on biopolymers such as proteins and oligonucleotides. Sequence information on peptides is obtained, as shown in Fig. 34. With the development of matrix-assisted laser-desorption ionization (MALDI) and electrospray ionization (see Secs. 1.4 and 1.5), mass spectrometry has begun to make significant contributions to the analysis of higher–molecular-weight biopolymers ($\leq$100 000 atomic mass units). Applications to protein/peptide characterization are well established, and some work on oligonucleotide sequencing has appeared. Conventional methods for DNA sequencing employ sequencing reactions to generate the four nested sets of DNA fragments terminating in dideoxynucleotides A, C, G, or T (where A, C, G, and T represent the four nitrogenous bases adenine, cytosine, guanine, and thymine, respectively) followed by electrophoretic separations in polyacrylamide gel matrices to separate the single-stranded DNA fragments by size and fluorescence detection to characterize the terminal base. In experiments using mass spectrometry, the DNA fragments can be subjected to MALDI or electrospray molecular-weight analysis without intervening electrophoretic separation. These ionization techniques can be coupled with TOF, FT-ICR, or quadrupole mass spectrometers, suggesting that sequences can be read directly from the mass spectra of the mixture. Figures 35(a) and 35(b) show MALDI mass spectra recorded for A, C, G, and T mock-sequence mixtures. The ion signals for each fragment are labeled according to the base length (i.e., the number of synthetic nucleotides). In the A, C, G, and T mixtures, all fragments are terminated with the base A, C, G, and T, respectively. The mass spectrum in Fig. 35(b) was generated by overlaying the four spectra in Fig. 35(a). The order of the peaks corresponds to the sequence, which is GAATTCGAGCTCGGTACCCGGGGA, and each peak differs from its neighbors by one base. DNA sequencing by mass spectrometry potentially has detection limits in the femtomole region.

3.5 Isotope-Dilution Mass Spectrometry

Isotope-dilution mass spectrometry (IDMS) is one of a few methods of quantitation that have high sensitivity, high reproducibility, and excellent accuracy for both elemental and molecular analysis. Provided

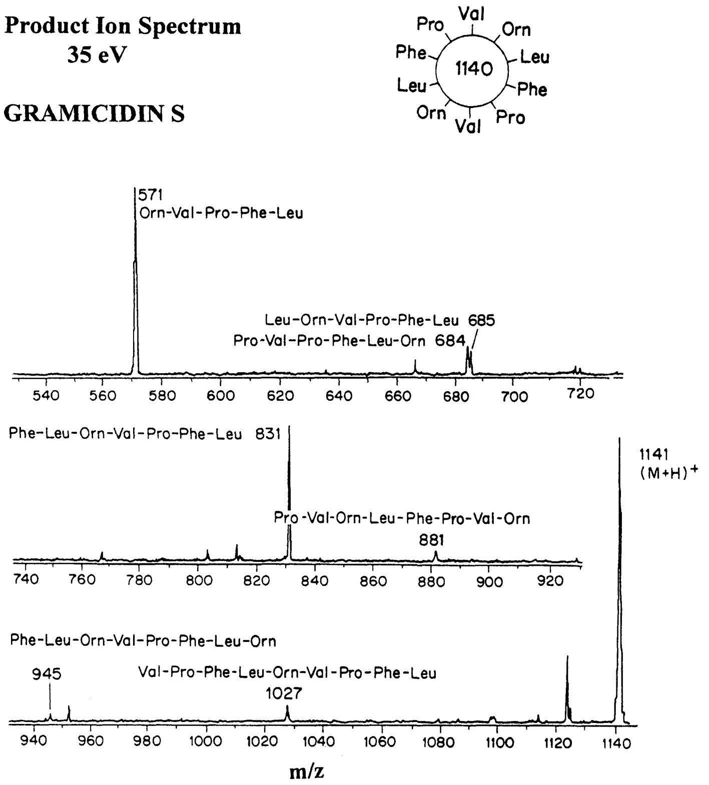

FIG. 34. Use of MS/MS to obtain sequence information on a cyclic peptide (Lambert *et al.*, 1987).

the element of interest has at least two stable or long-lived isotopes, a known quantity of a spike with a different isotopic composition can be added to the sample. Before the determination, the sample and the spike must be completely mixed. Then the ion intensities of two isotopes in the "isotope-diluted" sample are measured using mass spectrometry. From the measured isotope ratio, one can obtain the concentration of the element to be determined in the sample. IDMS can be applied in conjunction with most ionization methods. Table 6 summarizes the most important ionization methods used for elemental analysis by IDMS.

Note that because IDMS is an internal standard method, it provides high precision and accuracy (1% or better), as well as low detection limits (μg/g–pg/g). IDMS is applied in geochronology, radionuclear and isotopic work, certification of standard reference materials, and calibration of other analytical methods. For example, the measured concentrations and isotopic compositions of the noble gases helium, neon, argon, krypton, and xenon in Antarctic meteorites provide information about spallogenic isotopes produced by cosmic rays, trapped primordial or solar gas, and radiogenic ^{4}He and ^{40}Ar from natural decay of unstable potassium (^{40}K), uranium, and thorium. In ^{40}Ar/^{39}Ar age determination for minerals containing potassium, the sample is irradiated in a nuclear reactor. ^{40}Ar is generated by the ^{40}K decay, and ^{39}Ar is produced by the nuclear reaction with ^{39}K. The number of ^{40}Ar atoms generated in the sample can be expressed as $^{40}\mathrm{Ar} = (\lambda_e/\lambda)(e^{\lambda t} - 1)$, where λ and λ_e are the total decay and electron-capture constants for ^{40}K. With careful measurements, the ages of minerals can be determined out to several billion years.

Applications to molecular analysis show similar advantages, and in these cases there are additional choices as to the number and locations of the labeled atoms within the molecular spike. As a matter of convenience, it is customary to use multiple isotopic la-

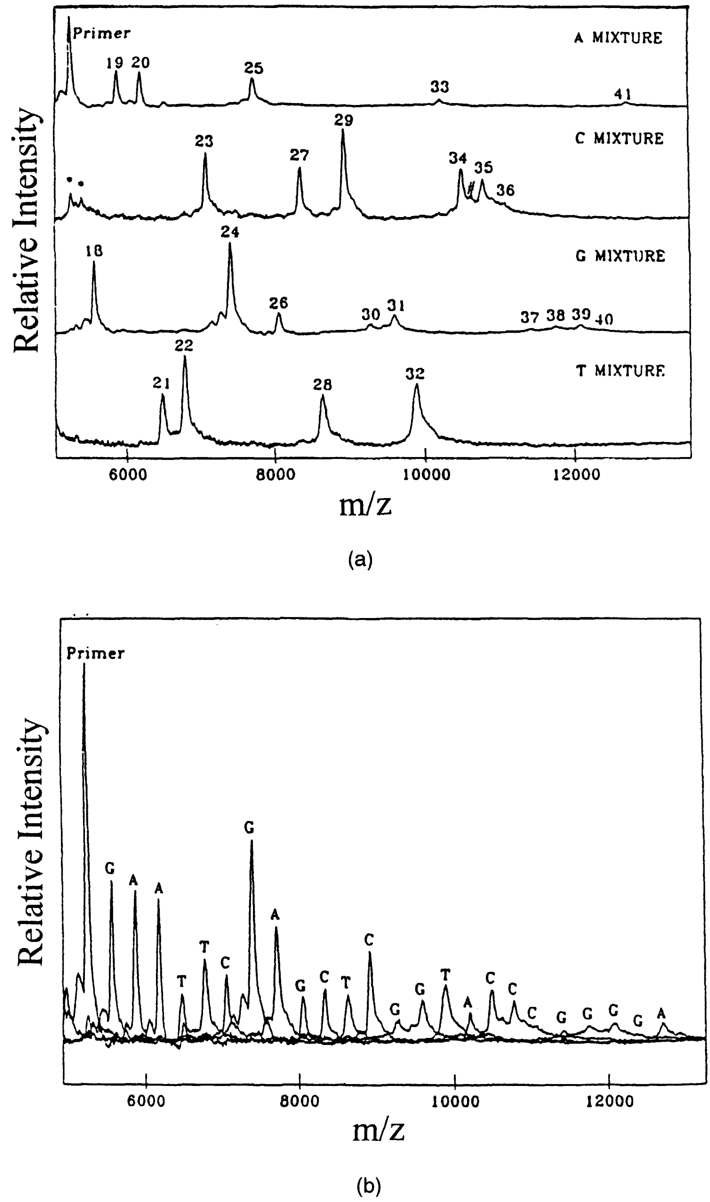

(a)

(b)

FIG. 35. MALDI/TOF negative-ion mass spectra of mock products of sequencing reactions containing mixtures of synthetic oligonucleotides chosen to match the first 24 DNA fragments generated in typical sequencing reactions using a standard template for the primer oligonucleotide of sequence of d(GTAAAACGACGGCCAGT). (a) The four spectra displayed individually. (b) The spectra overlaid and displayed on the same mass scale. Signals due to doubly charged ions and matrix adducts are labeled * and #, respectively. (Fitzgerald *et al.*, 1993.)

bels in isotopically labeled compounds; this gives good mass separation of the ions due to the spike from the naturally labeled isotopes (especially ^{13}C) of the analyte. For example, a primary metabolite of δ-9-tetrahydrocannabinol, 11-nor-δ-9-carboxylic acid, is often determined in clinical laboratories by IDMS using the synthetic trideuterated analog as the spike.

3.6 Isotope-Ratio Mass Spectrometry

Isotope-ratio mass spectrometry (IRMS) is one of the oldest types of mass spectrometry. It finds use in accurate determinations of isotopic compositions of the elements and low-molecular-weight permanent gases such as H_2, CO_2, N_2, and SO_2, as well as more complex compounds. These values provide a

Table 6. Ionization methods for elemental analysis with IDMS.

Ionization method	Preferred analytes
Thermal ionization	
Positive ions	Metals
Negative ions	Nonmetals
Electron ionization	Noble gases
Spark source	All elements
Plasma source (e.g., ICP)	All metals and metalloids
Laser	All elements
SIMS	All elements

rich source of information on the prior physical and chemical history of the sample. The isotope ratios of hydrogen, carbon, nitrogen, oxygen, and sulfur in organic compounds can be determined after combustion. In recent work, separation (by GC) is carried out, and combustion and IRMS on individual compounds present in complex mixtures is performed on line. This experiment can reveal a great deal about the previous history of the compounds; e.g., the biological sources of "marker" compounds such as steranes used in petroleum exploration can be revealed.

In a typical experiment, the ion source is operated in the EI mode with an electron energy of 70 eV and a trap current of 300–600 μA. Gas sample introduction into the ion source is traditionally by means of a dual gas-inlet system. Viscous gas flow from the reservoir (~50 Torr) into the ion source is used to avoid an isotope shift in the reservoir sample over time. The isotope ratio of the sample is measured by direct comparison with a standard gas of known isotopic composition. The resulting molecular ions are usually analyzed by a magnetic sector because of its temporal stability and high precision. The individual ion beams are focused into separate Faraday collectors so that multiple isotopes of a single element may be collected simultaneously (e.g., $^{28}N_2$, $^{29}N_2$, $^{30}N_2$). Modern IRMS systems are operated under automatic control.

IRMS has a high sensitivity to differences in isotopic abundances. A relative standard deviation of 0.001% can be achieved under optimum conditions. The widespread use of stable isotopes has led to increasing development of mass-spectrometric assays of isotopic abundances in fluids, products of combustion, individual organic compounds, and isotopes of mineral elements. For example, stable-isotope–labeled analogs have been used as tracers to follow the metabolism and excretion of biological compounds and drugs. Figure 36 shows different ^{13}C-to-^{12}C ratios of natural carbonaceous materials measured by IRMS and expressed as delta values (differences in parts per thousand

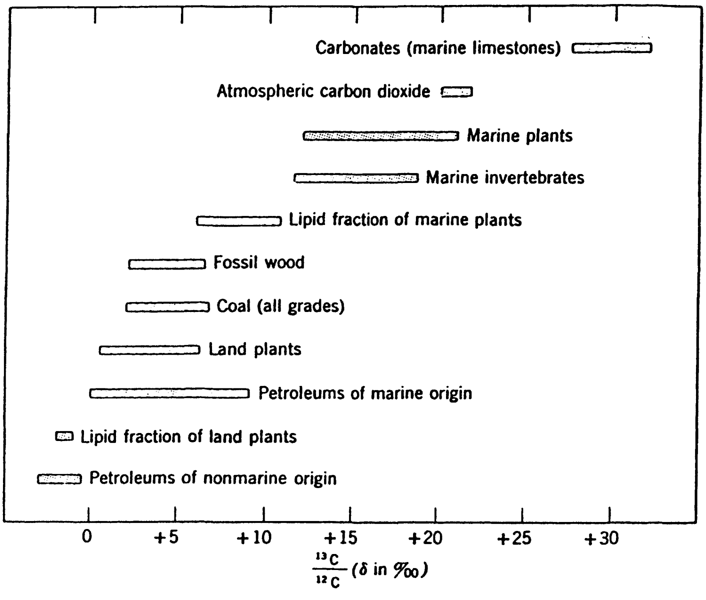

FIG. 36. $^{13}C/^{12}C$ δ-value ranges in natural carbonaceous materials (Craig *et al.*, 1964.)

from a standard material). IRMS continues to play an important role as a highly specialized tool in earth and atmospheric science, biochemistry, and physiology.

Another technique in mass spectrometry that also involves measurement of ratios of abundances of stable or radioisotopes is accelerator mass spectrometry (AMS). An accelerator operated at several million electronvolts is used to produce an ion beam, which is subjected to collisions during which charge-inversion, collision-induced dissociation, and charge-exchange processes may occur. By subjecting the beam to successive mass-to-charge ratio, momentum-to-charge ratio, or energy-to-charge ratio analyses, it is purified of even the smallest traces of molecular-ion contaminants before the isotope ratio of interest is measured. A typical setup is shown schematically in Fig. 37.

The entire instrument provides extremely high abundance sensitivity, viz., the ability to make measurements on specific isotopes even in the presence of other isotopes of the same element whose abundances differ by factors as large as 1:10^{15}. AMS has been applied to measurements of isotopes that are useful as tracers or chronometers including ^{10}Be/^{9}Be, ^{14}C/^{12}C, ^{26}Al/^{27}Al, and ^{36}Cl/^{35}Cl. In the best-known example, the measurement of radiocarbon (^{14}C) is used to determine the age of archeological and geological samples. The procedure only requires a small amount of sample (~1 mg), and with ^{14}C/^{12}C ratios of up to 10^{15}, age determinations up to 100 000 years can be made. Note that the method measures undecayed ^{14}C, *unlike* conventional dating methods that observe the number of decays of ^{14}C per unit time. The detection of tritium (^{3}H) with a half-life of 12.3 yr at a concentration of 1 part in 10^{14} of deuterium has been reported, and in this application the detection efficiency achieved by AMS is two orders of magnitude higher than that achieved by beta counting.

AMS is also used for fundamental studies in atomic and nuclear physics. A combined two-accelerator system at the Lawrence Berkeley Laboratory (U.S.A.) produced the first completely ionized uranium atoms, i.e., $^{238}U^{+92}$, a nucleus with no orbital electrons. Such fully stripped ions exist, for instance, in the interior of many stars and have been produced abundantly for most elements at velocities about 80–90% of the speed of light at the Heavy Ion Accelerator in Darmstadt, Germany. The masses of such nuclei have been investigated there as well as their half-lives. Because there can be no interaction between the atomic nucleus and the electronic shell in the fully stripped ions, their half-lives are different from those ions that possess at least part of their electron cloud.

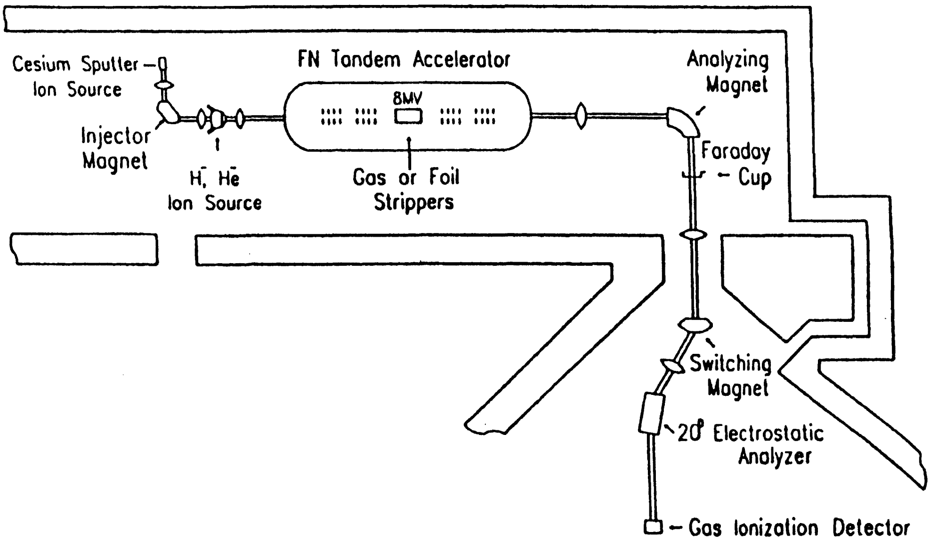

FIG. 37. Schematic diagram of the accelerator mass spectrometer at Purdue University.

3.7 Mass Measurements of Stable and Unstable Nuclei

Precise mass measurements of the isotopes of the elements were first made in experiments begun by Aston in 1919 using a sector-field mass spectrometer. These measurements revealed differences in the binding energies of the different isotopes of the elements, from which the nuclear shell structure was deduced. Even for isotopes of low abundance, such experiments have been conducted with mass resolving powers of about 10^6. For the investigation of difficult-to-produce short-lived neutron-rich or neutron-deficient isotopes, sector-field mass spectrometers have been operated as time-of-flight analyzers. In such systems more energetic and thus faster ions of the same mass-to-charge ratio are guided to the final ion collector so that single-mass ions of all energies arrive simultaneously, provided they started at the same time. Alternatively, the ion energies have all been made equal or been measured for every ion individually by a second sector-field analyzer. With such instruments the binding energies of about 2000 exotic nuclei have been investigated. Some of these exotic nuclei are short-lived but exist abundantly in nova and supernova star explosions, and they have had a major influence on the formation of all heavy elements in the universe.

3.8 Resonance-Ionization Mass Spectrometry

Resonance ionization is a photoionization process where atoms in the gas phase are ionized by successive absorption of photons that energetically match quantum-selected state differences of these atoms. Typically, a laser is tuned precisely to the wavelength required for the excited state, which is unique to the element (and isotope) being measured. Then additional photons of the same or another energy are used to photoionize this excited atom.

Resonance-ionization mass spectrometry (RIMS) is a very selective (even to particular isotopes) and highly sensitive technique since only the atoms of interest are ionized and yet all atoms of the selected element lying in the confines of the laser beam are ionized. The physical basis for the process makes this method yield highly predictable results. The elimination of isobaric interferences is possible in RIMS. The detection and analysis of every chemical element in the periodic table can be realized, and low detection limits in the parts-per-trillion range are normally obtainable. RIMS has found wide applications for elemental analysis in the semiconductor industry and in geological studies.

A technique in which mass spectrometry is used in conjunction with laser spectroscopy to obtain fundamental atomic and molecular information is photodetachment. In this technique, negative ions are generated in a hot-cathode or cold-cathode electric-discharge ion source. Then the mass-selected negative-ion beam is intersected by a laser beam, and the photodetachment spectrum is obtained by recording the number of electrons produced as a function of photon energy. The information obtained can be used to study the electronic structures of molecules. In addition, the electron affinities of atoms and some molecules can be measured precisely by this technique. Fundamental information regarding the electronic, vibrational, and rotational states as well as binding energies can also be provided. A related photoelectron technique is photoionization–photoelectron-coincidence spectroscopy (PIPECO) in which the product ions and photoelectrons are monitored simultaneously. If one performs this measurement while monitoring zero–kinetic-energy electrons, the internal energy distribution of ions can be determined by measurement of the total ion current as a function of the photon energy. Such experiments produce detailed spectroscopic information on gas-phase ions.

3.9 Mass Spectrometry in Surface Analysis

The sputtering process used in DI (Sec. 1.4) releases neutral and ionized species from a surface and can be used for elemental and molecular analysis. In addition, depth profiles are obtained by monitoring the abundance of particular ions as a function of the time of sputtering. The removal of surface layers depends on the current density of the energetic beam and the composition of the material surface. Figure 38 illustrates SIMS depth profiling of boron in

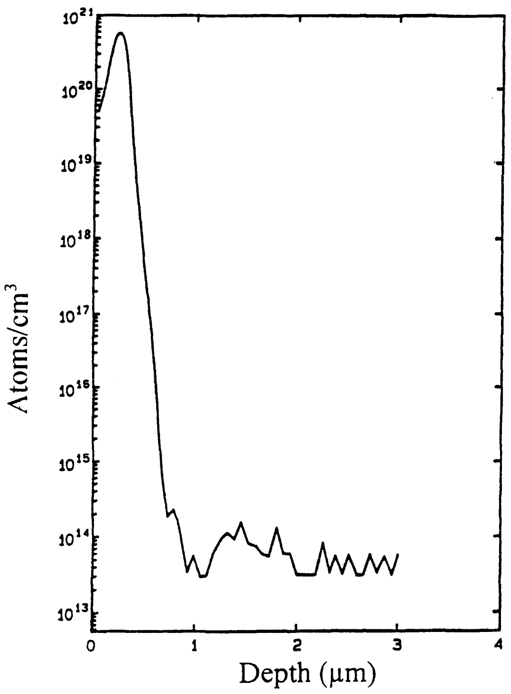

FIG. 38. SIMS depth profile of boron implant in silicon. Primary ions O_2^+. (Adams *et al.*, 1988.)

silicon, where the boron concentrations are recorded quantitatively with a sensitivity of the order of 1 part per billion and a dynamic range of more than seven orders of magnitude, while the depth is followed with a resolution better than ~5 nm. However, this is a destructive method of analysis; and since the percentage of sputtered ions to neutrals varies with compound composition and structure, high quantitative accuracy requires a very good understanding of the underlying processes. A complementary technique to SIMS is Rutherford backscattering spectroscopy (RBS), which employs scattering of high-energy ions, usually $^4He^+$ beams in the MeV energy range, to study the near-surface regions of solids, including implantation profiles, impurity distributions, and diffusion phenomena. Figure 39 shows the energy spectrum of the backscattered products of 2.0-MeV $^4He^+$-ion collisions at 150° from a self-supporting Al_2O_3 film. The result shows two well-resolved peaks corresponding (from the measured He/He$^+$ kinetic energy) to O and Al. RBS has been used to determine impurities in semiconductor devices and the uniformity and stoichiometry of solid-phase epitaxial films in microelectronics research. A lower-energy experiment, ion-scattering spectrometry, uses binary elastic collisions of ions to measure the masses of atoms lying at the surfaces of materials.

Two imaging techniques based on DI, the ion microprobe and ion microscope, have long been used for surface analysis. In the ion microprobe, the desired spatial resolution is achieved by rastering a focused primary beam (≤ 10 μm) across a surface and analyzing the desorbed ions of interest using a mass spectrometer, often a TOF instrument. The lateral resolution in the ion microprobe is essentially given by the beam size. The ion microscope utilizes primary ion beams of relatively large dimension (~5–500 μm). A one-to-one object-to-image transfer is achieved by an ion optical system that transports the desorbed ions from the sample to a detector while maintaining their original spatial relationships.

Another surface analysis technique that employs mass spectrometry is temperature-programmed desorption (TPD), in which the surface is heated at an appropriate rate so that adsorbed species are sequentially desorbed and then ionized and analyzed by mass spectrometry. The dependence of desorbed ion intensities on temperature provides information about the strength of surface binding, while the nature of the adsorbates on the surface at different temperatures is inferred from the ions that appear in the spectrum. Figure 40 shows TPD spectra following Ru (001) oxidation at 1465 K in $\sim 10^{-7}$ Torr O_2. It can be seen from the NO desorption signals that the nature of the adsorption changes between 1110 and 1520 K.

3.10 Ion-Mobility Spectrometry

Mass spectrometry has become a powerful tool in gaining fundamental understanding of atoms and molecules. Ion-mobility measurements in gases have been made for many years, but the recent use of mass-analyzed ion beams has demonstrated unexpected richness in this measurement.

A typical experimental arrangement is shown in Fig. 41. Mass-selected ions generated in the ion source are injected into a high-pressure drift cell (2–5 Torr helium).

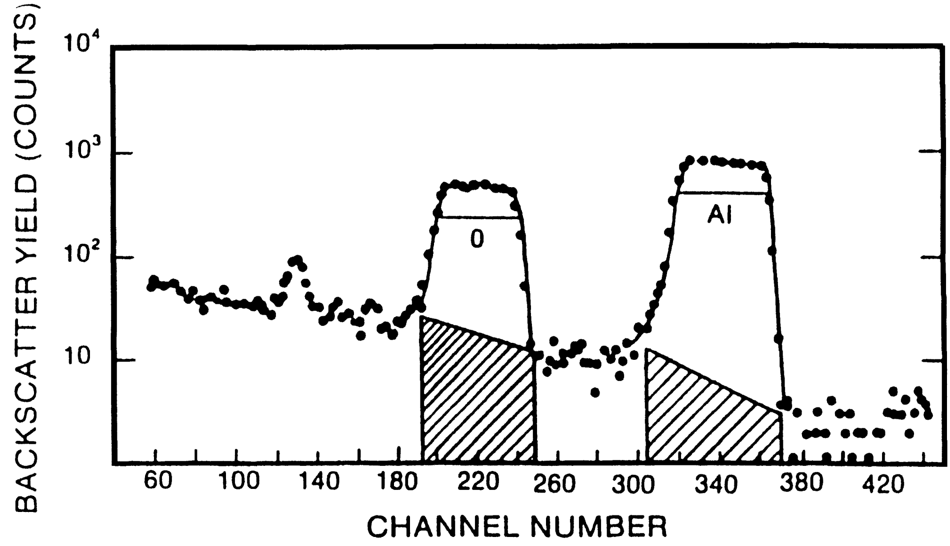

FIG. 39. Energy spectrum of 2-MeV ^{4}He backscattered from a 58 μg/cm^2 self-supporting Al_2O_3 film. The shaded area under each peak indicates the magnitude of the background correction. (MacDonald *et al.*, 1983.)

The ions then drift through the cell under the influence of a weak electric field (1–20 V/cm) and exit the cell through a small aperture. They pass through a quadrupole mass filter and are detected as an arrival-time distribution (ATD). Ions with the same mass but different collision cross sections will have different mobilities and hence drift times, giving rise to different peaks in the ATD. Differences in mobilities have been shown to be associated with differences as subtle as those due to different electronic configurations of transition-metal ions. Figure 42 shows bimodal arrival-time distributions for Co^+ ions generated from different compounds and under different EI conditions. Clearly, Co^+ ions are generated in different electronic states and can be distinguished by their arrival times. This observation offers a good opportunity to select and study reactivities of ions in different electronic states.

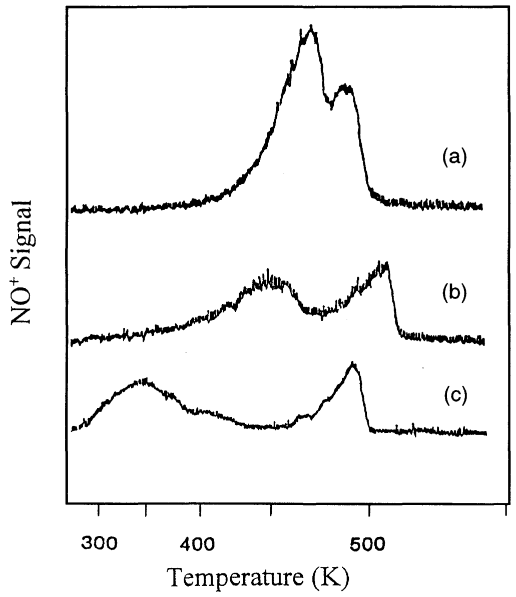

FIG. 40. Temperature-programmed desorption signal for NO released from oxidized Ru. Temperature of annealing (a) 1110 K, (b) 1310 K, and (c) 1520 K. Heating rate 10 K/s. (Vanselow and Howe, 1984.)

Not only can small atomic ions be characterized by their mobilities but large carbon clusters display drift times that depend on their structures. This method has been used to show that carbon atoms first assemble in linear chains, then form monocyclic planar rings at about C_{10}, and finally form families of planar bi-, tri-, and tetra-cyclic rings at C_{20}, C_{30}, and C_{40}, respectively. With advances in ionization methods, macromolecules can be easily ionized, and their structural characterization by mass spectrometry is possible (Sec. 3.4). Ion-mobility measurements hold

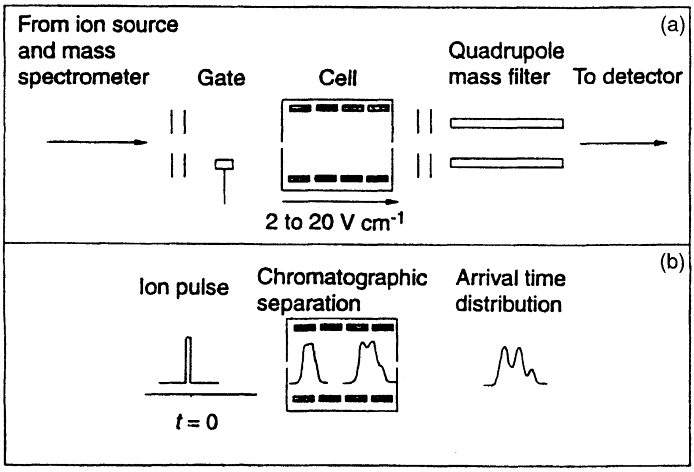
(a)
From ion source and mass spectrometer
Gate
Cell
Quadrupole mass filter
To detector
2 to 20 V cm-1
(b)
Ion pulse
Chromatographic separation
Arrival time distribution
t = 0

FIG. 41. Ion-mobility experiment. (a) Essential elements of the instrument. (b) Spatial and temporal ion-beam profile for a model system. (Bowers *et al.*, 1993.)

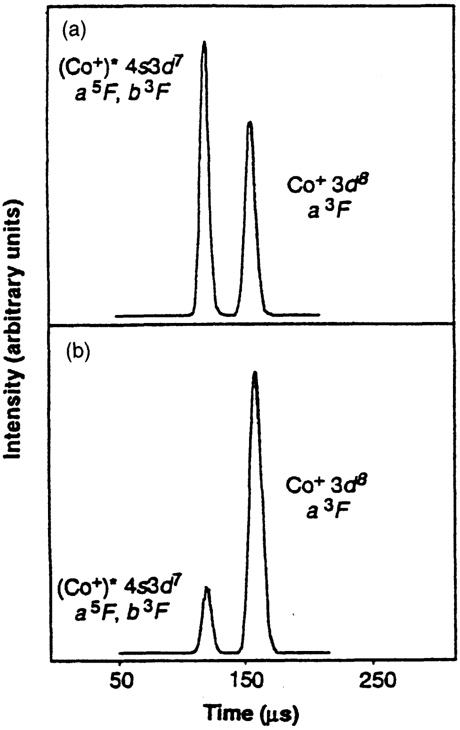

FIG. 42. Arrival-time distributions of Co$^+$ ions passed through a collision cell. The ions are formed (a) from 75-eV electron impact on Co(CO)$_3$NO, yielding 48% ground state and 52% excited state Co$^+$ ions, and (b) by 50-eV electron impact on Co(C$_5$H$_5$)(CO)$_2$, forming 83% ground state and 17% excited state Co$^+$ ions. (Bowers *et al.*, 1993.)

great promise in distinguishing isomeric biological molecules and, more importantly, for identifying the specific conformations associated with the native states of biological compounds.

3.11 Analysis of Single Particles in Aerosols

The high sensitivity of mass spectrometry allows the determination of the chemical composition of single aerosol particles. In this experiment, aerosols are sampled directly using a differentially pumped mass-spectrometer inlet. The particle beam emerging from the inlet is intercepted by a continuous laser beam in the ion source region. Particles passing through the laser beam are detected by light scattering. The scattered radiation from each particle triggers a high-energy laser that ablates the particle in flight. Ions produced from the particle are analyzed by mass spectrometry. The chemical composition of the particle can be inferred from the distribution of ions observed in the mass spectrum. Figure 43 displays mass spectra of single 3.5-μm-diam sodium chloride particles. This technique permits exploration of atmospheric particle size, composition, and morphology, and thus the physical and chemical processes that transform the particles.

ACKNOWLEDGMENT

The authors would like to thank Carsten Weil for his help in preparing figures.

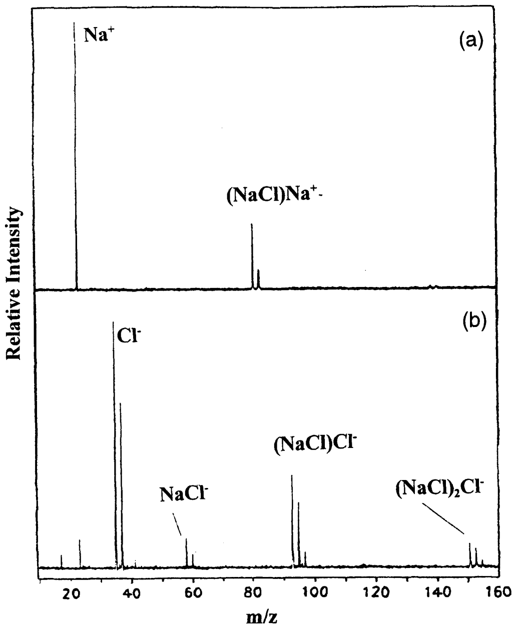

FIG. 43. Mass spectra of single 3.5-μm-diam sodium chloride particles taken with a laser pulse energy of ca. 3 μJ at 248 nm: (a) positive-ion spectrum; (b) negative-ion spectrum (Carson *et al.*, 1995).

ACRONYMS

AMS: Accelerator mass spectrometry.

APCI: Atmospheric-pressure chemical ionization.

ATD: Arrival-time distribution.

CCD: Charge-coupled device.

CI: Chemical ionization.

CID: Collision-induced dissociation.

CZE/MS: Capillary-zone electrophoresis/mass spectrometry.

DI: Desorption ionization.

EI: Electron impact (ionization).

ES: Electrospray.

FAB: Fast-atom bombardment.

FD: Field desorption.

FT-ICR: Fourier-transform ion-cyclotron resonance.

GC/MS: Gas chromatography/mass spectrometry.

GD: Glow discharge.

ICP: Inductively coupled plasma.

IDMS: Isotope-dilution mass spectrometry.

IRMS: Isotope-ratio mass spectrometry.

ITR: Integrating transient recording.

LC/MS: Liquid chromatography/mass spectrometry.

LD: Laser desorption.

MAGIC: Monodispersed aerosol generator for introduction from liquid chromatography.

MALDI: Matrix-assisted laser desorption ionization.

MS/MS: Mass spectrometry/mass spectrometry.

MW: Molecular weight.

NCI: Negative-ion chemical ionization.

PA: Proton affinity.

PD: Plasma desorption.

PIPECO: Photoionization–photoelectron-coincidence spectroscopy.

RBS: Rutherford backscattering spectroscopy.

rf: Radio frequency.

RIMS: Resonance-ionization mass spectrometry.

SI: Spray ionization.

SID: Surface-induced dissociation.

SIM: Selected-ion monitoring.

SIMS: Secondary-ion mass spectrometry.

TDC: Time-to-digital converter.

TOF: Time of flight.

TPD: Temperature-programmed desorption.

TS: Thermospray.

GLOSSARY

Atmospheric-Pressure Chemical Ionization: A variant of chemical ionization performed at atmospheric pressure.

Charge Exchange: Process whereby one particle transfers an electron to another, e.g., $M + A^+ \rightarrow M^+ + A$.

Chemical Ionization: Method in which neutral molecules are ionized by ion–molecule reactions to generate a quasimolecular ion at a pressure of about 1 Torr.

Collision-Induced Dissociation: Process whereby a mass-selected ion is internally excited and caused to fragment by collision with a target gas.

Cyclotron Motion: Cyclic rotation of an ion in a fixed magnetic field.

Desorption Ionization: Method used to ionize nonvolatile solid samples by impact of energetic particles or photon beams.

Double Focussing: A combination of direction and velocity focussing in sector instruments used to achieve high resolution.

Elastic Collision: Collision in which internal energy is conserved.

Electron Affinity: Enthalpy change for the process $M^- \rightarrow M + e^-$.

Electron Capture: Ionization process in which a molecule or atom captures a thermal-energy electron typically in a CI source.

Electron Impact: Ionization method in which molecules are ionized directly by energetic electrons (usually 70 eV) at low pressure ($\sim 10^{-5}$ Torr).

Glow Discharge: Method used to ionize solid samples for elemental analysis by applying an electric field to create an energetic plasma.

Inductively Coupled Plasma: Method used to ionize solution samples for elemental analysis by a plasma.

Inelastic Collision: Collision in which internal energy is not conserved.

Ion Internal Energy: Total electronic, vibrational, and rotational energy referenced to ground state of the ion.

Ion Microprobe: Technique that uses a primary ion beam (≤ 10 μm) rastered across a surface to desorb ions for analysis by mass spectrometry.

Ion Microscope: Technique that uses a large ion beam (5–500 μm) to obtain one-to-one object-to-image transfer by ion-optic lenses that extract and transport the desorbed ions from the sample while maintaining their original spatial relationships.

Ionization Energy: Minimum energy required to remove an electron from a molecule.

Liouville's Theorem: Conservation law in ion optics that states that the product of velocity, angular dispersion, and beam dimension is conserved.

Magnetron Motion: Slow circular drift of an ion along a path of constant electrostatic potential; magnetron motion occurs in ICR as a result of the crossed radial electric field and axial magnetic field.

Mass Resolving Power: $m/\Delta m$, in which m is ionic mass and Δm is the width of the mass peak (typically taken as the full width at half-maximum mass-spectral peak height).

Mathieu Stability Diagram: Diagram showing the solutions to the Mathieu equation that correspond to stable ion trajectories displayed as a function of parameters related to operating voltages and mass and charge of the trapped ions.

Metastable Ion: Ion that fragments slowly and after emergence from the ion source; in sector instruments metastable ions give rise to signals that appear at unique m/z values related to the parent- and product-ion masses.

Molecular Ion: Ion derived from the neutral molecule by loss or gain of an electron.

Neutral-Loss Scan: An MS/MS experiment that records all parent ions that lose a particular neutral fragment.

Parent Ion: Any ion that fragments to a product ion.

Parent-Ion Scan: An MS/MS experiment that records all parent ions that produce a particular product ion.

Photodissociation: Process in which an ion fragments by absorption of one or more photons.

Product Ion: Ion generated by fragmentation of any parent ion.

Product-Ion Scan: An MS/MS experiment that records all product ions derived from a single parent ion.

Proton Affinity: Enthalpy change for the process $MH^+ \rightarrow M + H^+$.

Quasimolecular Ion: Ion that is simply derived from or related to a neutral molecule, e.g., $(M + H)^+$, $(M + Cl)^-$, $(M - H)^-$.

Secondary-Ion Mass Spectrometry: Mass spectrometry based on analysis of particles that are emitted when a surface, usually a solid, although sometimes a liquid, is bombarded by energetic (~keV) primary particles (e.g., Ar^+ and Cs^+).

Spray Ionization: Method used to ionize liquid samples directly by electrical, thermal, or pneumatic energy.

Surface-Induced Dissociation: Process whereby a mass-selected ion is excited and caused to fragment by collision with a target surface.

Tandem Mass Spectrometry: Two-stage mass-analysis experiment, used to study the chemistry of selected ions.

Thermal Ionization: Method used to ionize solid samples on a hot surface of a metal filament.

Works Cited

Adams, F., Gijbels, R., Van Grieken, R. (Eds.) (1988), *Inorganic Mass Spectrometry,* New York: Wiley.

Allemann, M., Kellerhals, Hp., Wanczek, K. P.

(1982), in: Hartmann, H., Wanczek, K.-P. (Eds.), *Ion Cyclotron Resonance Spectrometry II*, Lecture Notes in Chemistry Vol. 31, Berlin: Springer, p. 380.

Beu, S. C., Senko, M. W., Quinn, J. P., McLafferty, F. W. (1993), *J. Am. Soc. Mass Spectrom.* **4,** 190–192.

Bowers, M. T., Kemper, P. R., Helden, G. V., Van Koppen, P. A. M. (1993), *Science* **260,** 1446–1451.

Budzikiewicz, H., Djerassi, C., Williams, D. H. (1967), *Mass Spectrometry of Organic Compounds*, San Francisco: Holden-Day, Inc., p. 163.

Carr, S. A., Hemling, M. E., Bean, M. F., Roberts, G. D. (1991), *Anal. Chem.* **63,** 2802–2824.

Carson, P. G., Neubauer, K. R., Johnston, M. V., Wexler, A. S. (1995), *J. Aerosol Sci.***26,** 535–545.

Coburn, J. W., Kay, E. (1971), *Appl. Phys. Lett.***19,** 350–352.

Cooks, R. G., Glish, G. L., McLuckey, S. A., Kaiser, R. E. (1991), *Chem. Eng. News* **69,** 26–41.

Craig, H., Miller, S. L., Wasserburg, G. J. (Eds.) (1964), *Isotopic and Cosmic Chemistry*, Amsterdam: North Holland.

Dawson, P. H. (1976), *Quadrupole Mass Spectrometry and Its Applications*, New York: Elsevier Scientific Publishing Company.

Fitzgerald, M. C., Zhu, L., Smith, L. M. (1993), *Rapid Commun. Mass Spectrom.* **7,** 895–897.

Ingle, J. D. Jr., Crouch, S. R. (1988), *Spectrochemical Analysis*, Englewood Cliffs: Prentice Hall.

Karas, M., Bahr, U., Giessman, U. (1991), *Mass Spectrom. Rev.* **10,** 335–357.

Lambert, J. B., Shurvell, H. F., Lightner, D. A., Cooks, R. G. (1987), *Introduction to Organic Spectroscopy*, New York: Macmillan Publishing Company.

MacDonald, J. R., Davies, J. A., Jackman, T. E., Feldman, L. C. (1983), *J. Appl. Phys.* **54,** 1800–1803.

Mann, M., Meng, C. K., Fenn, J. B. (1989), *Anal. Chem.* **61,** 1702–1708.

March, R. E., Hughes, R. E. (1989), *Quadrupole Storage Mass Spectrometry*, New York: Wiley.

Marshall, A. G. (1985), *Acc. Chem. Res.* **18,** 316–322.

Stemmler, E. A., Hites, R. A. (1988), *Biomed. Environ. Mass Spectrom.* **17,** 311–328.

Thomson, J. J. (1913), *Rays of Positive Electricity and Their Application to Chemical Analysis*, London: Longmans, Green & Co.

White, F. A., Wood, G. M. (1986), *Mass Spectrometry: Applications in Science and Engineering*, New York: Wiley.

Wood, K. V., Ciupek, J. D., Cooks, R. G., Ferguson, C. R. (1982), *SAE The Engineering Resource for Advancing Mobility*, Serial No. 821217.

Further Reading

Adams, F., Gijbels, R., Grieken, R. Van (Eds.) (1988), *Inorganic Mass Spectrometry*, New York: Wiley.

Aston, F. W. (1924), *Isotopes*, London: Edward Arnold.

Beynon, J. H., Morgan, R. P. (1978), *Int. J. Mass Spectrom. Ion Processes* **27,** 1.

Brunnée, C. (1987), *Int. J. Mass Spectrom. Ion Processes* **76,** 125–237.

Buchanan, M. V. (Ed.) (1987), *Fourier Transform Mass Spectrometry*, Washington, D.C: American Chemical Society.

Busch, K. L., Glish, G. L., McLuckey, S. A. (1988), *Mass Spectrometry/Mass Spectrometry: Techniques and Applications of Tandem Mass Spectrometry*, New York: VCH Publishers, Inc.

Campana, J. E. (1980), *Int. J. Mass Spectrom. Ion Phys.* **33,** 101–117.

Chapman, J. R. (1993), *Practical Organic Mass Spectrometry*, Chichester: Wiley.

Cooks, R. G. (1978), *Collision Spectroscopy*, New York: Plenum Press.

Cotter, R. J. (Ed) (1994), *Time-of-Flight Mass Spectrometry*, Washington, D.C.: American Chemical Society.

Dawson, P. H. (1976), *Quadrupole Mass Spectrometry and Its Applications*, New York: Elsevier Scientific Publishing Company.

Guilhaus, M. (1995), *J. Mass Spectrom.* **30,** 1519–1532.

Howe, I., Williams, D. H., Bowen, R. D. (1981), *Mass Spectrometry: Principles and Applications*, New York: McGraw-Hill.

Kiser, R. W. (1965), *Introduction to Mass Spectrometry and Its Applications*, Englewood Cliffs: Prentice-Hall.

March, R. E., Hughes, R. J. (1989), *Quadrupole Storage Mass Spectrometry*, New York: Wiley.

Marshall, A. G., Verdun, F. R. (1990), *Fourier Transform in NMR, Optical and Mass Spectrometry—A User's Handbook*, New York: Elsevier.

McCloskey, J. A. (Ed.) (1990), *Mass Spectrometry, Methods in Enzymology*, Vol. 193, San Diego: Academic Press.

Vanselow, R., Howe, R. (Eds.) (1984), *Chemistry and Physics of Solid Surfaces V*, New York: Springer-Verlag.

Weber *et al.* (1986), *Rev. Sci. Instrum.***57,** 1251.

White, F. A., Wood, G. M. (1986), *Mass Spectrometry: Applications in Science and Engineering*, New York: Wiley.

Wollnik, H. (1987), *Optics of Charged Particles*, San Diego: Academic Press.

SPECTROMETERS, NEUTRON

John A. Harvey, *Joint Institute for Heavy Ion Research, Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A.*

Robert M. Nicklow, *Oak Ridge National Laboratory (retired), 176 Cedar Point Drive, Kingston, Tennessee, U.S.A.*

Introduction 331
1. **Neutron Spectrometers at Nuclear Reactors** 334
1.1 Neutron Crystal Spectrometers 334
1.1.1 Triple-Axis Spectrometer 336
1.1.2 Polarized-Beam Triple-Axis Spectrometer 336
1.1.3 Spin-Echo Spectrometer 337
1.1.4 Beryllium-Filter Spectrometer 338
1.1.5 Back-Reflection Analyzer Spectrometer 338
1.2 Mechanical Neutron Velocity-Selector Spectrometers 339
1.3 Neutron Time-of-Flight Spectrometers 340
1.3.1 Fast–Neutron-Chopper Spectrometers 340
1.3.2 Slow–Neutron-Chopper Spectrometers 341
2. **Neutron Spectrometers at Accelerators** 342
2.1 Variable-Energy Neutron Source Spectrometers 342
2.1.1 Van de Graaff Accelerator Spectrometer 342
2.1.2 Cyclotron and Proton Linear-Accelerator Spectrometers 343
2.2 Neutron Time-of-Flight Spectrometers 343
2.2.1 Proton Van de Graaff Accelerator Spectrometer 343
2.2.2 Electron-Accelerator Photoneutron Source Spectrometers 343
2.2.3 Deuteron-Cyclotron Source Spectrometer 346
2.2.4 Proton Spallation Source Spectrometers 346
2.2.5 Direct-Geometry Chopper Spectrometer 348
2.2.6 Crystal-Analyzer Time-of-Flight Spectrometer 348
2.2.7 Pulsed Fission- and Fusion-Source Spectrometers 349
2.2.8 Pulsed-Source Slowing-Down Neutron Spectrometer 350
3. **Neutron Detectors** 350
3.1 Neutron Detectors for Spectrometers 350
3.2 Neutron-Detector Spectrometers 350
Glossary 351
Works Cited 351
Further Reading 352

INTRODUCTION

Neutron spectrometers are valuable devices used for condensed-matter studies, for basic nuclear physics research, and for providing cross-section data for fission and fusion reactor programs. Many different kinds of neutron spectrometers have been developed since the discovery of the neutron in 1932 and have been used to explore a very broad energy range, from 10^{-6} eV to 800 MeV. Typically, a spectrometer is used to measure the energy of neutrons incident on and sometimes scattered by a sample. For nuclear physics research or cross-section measurements, this information is usually focused on the neutron–nucleus interaction as a function of the incident neutron energy. For condensed-matter research, the results provide information about the energy and momentum dependence of the neutron–sample interaction. This information is more of-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ten concerned with the internal properties of the sample; the neutron is simply being used as a probe of those properties.

The components of a spectrometer usually depend very strongly on the type of neutron source with which it is used, on the type of research being carried out, and on the neutron energy and energy resolution required by the research. There are two major classes of neutron sources, which differ primarily in the time structure of the neutron beam. Nuclear reactors produce a "steady-state" flow or current of neutrons, whereas accelerator sources usually produce a "pulsed" or chopped flow of neutrons. Both sources provide beams with a broad energy spread, often called a "white" source, and for both classes, the nuclear reactions responsible for producing the neutrons yield neutrons with energies in the MeV range. While neutrons with MeV energies are very useful for certain neutron–nucleus cross-section studies, keV- and eV-energy neutrons are also required. Research in condensed matter requires neutrons in the so-called thermal energy range from about 0.002 to 0.2 eV. Consequently, most neutron sources have moderators that serve to lower the average neutron energy of the beam from the source. For reactors, moderation of the neutron energy is required for sustaining the fission chain reaction even if no neutrons were being extracted for other purposes.

Moderation is achieved by allowing the energetic neutrons to enter a hydrogenous material such as water or heavy water, D_2O. Since the mass of the proton or deuteron is similar to that of the neutron, the neutron loses on the average a large fraction of its energy at each neutron–proton collision. After only about 30 such collisions, neutrons beginning with MeV energy will have an energy close to that of the mean thermal energy of the protons in the moderator material, i.e., $E_{\text{ave}} \propto kT$, which is typically about 0.03 eV. If thermalization is complete, they will also possess an energy distribution that is usually described quite well by a Maxwellian distribution appropriate for the temperature T.

For each source class there is a wide variety of spectrometer designs, each of which is optimized for measurements over only limited energy or momentum ranges. For steady-state neutron sources, it is necessary to have a device to select monoenergetic neutrons of desired energies from a neutron source with a broad energy spread. This energy can be varied systematically in order to make measurements as a function of neutron energy. Monoenergetic neutrons can also be produced directly from monoenergetic ion beams from some accelerators using thin targets, and the neutron energy can be varied by changing the energy of the ion beam. For a pulsed neutron source, the neutron energy is determined by the time-of-flight (TOF) technique by measuring the time taken for a neutron to travel from the source to a detector many meters away. Most neutron spectrometers are located at major neutron sources (see Tables 1 and 2).

Furthermore, there is a significant difference in the type of measurement required for nuclear cross-section research and for condensed-matter research that influences the spectrometer design. For condensed-matter research, one measures the *change* in energy $\Delta E = \hbar\omega$ and momentum $\hbar Q$ of neutrons that have been *scattered* by the sample. That is, we measure a scattering cross section $S(Q,\omega)$ that contains much information about the specimen. For example, one can determine the energy levels of vibrational and magnetic excitations with energies up to a few eV. The diffusive motions of atoms in solids and rotational–tunneling transitions corresponding to molecular reorientations with characteristic energies of only 10^{-6} eV can also be investigated. For all such experiments, the energy of the neutron is measured twice: before and after scattering. See NEUTRON SCATTERING.

For most neutron–nucleus cross-section research, it is the energy of the neutron beam that is *incident* on the sample or detector that is measured. The most striking feature of these measurements is the observation of sharp resonances due to the strong interaction between the target nucleus and the incident neutron at unique energies. These resonances correspond to excited states in the compound nucleus at an excitation energy of many MeV. The widths of these neutron resonances can vary from a few meV with an average spacing between resonances of a few eV for a heavy nuclide such as ^{235}U, to widths as large as many keV with average spacings of hundreds of keV for light nuclides. Detailed properties of these

Table 1. Major reactor neutron sources.

Source name	Location	Spectrum characteristics	Peak thermal flux (n/cm^2 s)
Dhruva	Bhabha Atomic Research Center, Trombay, Bombay, India	Thermal, cold	2×10^{14}
DR 3	Risø National Laboratory, Denmark	Thermal, cold	3.5×10^{13}
FRJ-2	KFA, Jülich, Germany	Thermal, cold	2×10^{14}
HFBR—High Flux Beam Reactor	Brookhaven National Laboratory, Upton, NY, USA	Thermal, cold	8×10^{14}
HFIR—High Flux Isotope Reactor	Oak Ridge National Laboratory, Oak Ridge, TN, USA	Thermal	3×10^{15}
HFR—High Flux Reactor	Institute Laue Langevin, Grenoble, France	Thermal, cold, hot	1×10^{15}
IBR-2—Intense Burst Reactor-2	Joint Institute for Nuclear Research, Dubna, USSR	Pulsed reactor—thermal	1×10^{16} in pulse of 200-μs width and up to 25-Hz rate
JRR-3—Japan Research Reactor 3	Japan Atomic Energy Research Institute, Ibaraki, Japan	Thermal, cold	2×10^{14}
NBSR	National Institute of Standards & Technology	Thermal, cold	4×10^{14}
NRU	Chalk River Laboratory of AECL, Ontario, Canada	Thermal	3×10^{14}
Orphee	Laboratoire Léon Brillion, Saclay, France	Thermal, cold	3×10^{14}
MURR—University of Missouri Research Reactor	Columbia, MO, USA	Thermal	6×10^{14}

Table 2. Major accelerator neutron sources[a]

Source name	Location	Energy, particle	Peak fast neutrons per s	Peak thermal flux[b]
Geel Electron Linear Accelerator—GELINA	Geel, Belgium	120 MeV e^-	4×10^{18} 4 ns	
Intense Pulsed Neutron Source—IPNS	Argonne National Laboratory, Chicago, IL, USA	500 MeV p		1×10^{14}
ISIS	Rutherford Appleton Laboratory, England	800 MeV p		1×10^{16}
KENS	Laboratory for High Energy Physics KEK, Japan	500 MeV p		2×10^{14}
Los Alamos Neutron Scattering Center—LANSCE—plus Pulse Storage Ring—PSR	Los Alamos National Laboratory, Los Alamos, NM, USA	800 MeV p		1×10^{16}
Oak Ridge Electron Linear Accelerator—ORELA	Oak Ridge National Laboratory, Oak Ridge, TN, USA	140 MeV e^-	4×10^{18} 3 ns	
Rensselaer Polytechnic Institute Electron Linac	R.P.I., Troy, NY, USA	80 MeV e^-	2×10^{17} 10 ns	
Weapons Neutron Generator—WNR, no Pulse Storage Ring	Los Alamos National Laboratory, Los Alamos, NM, US	800 MeV p	5×10^{18} 1 ns	

---

[a] Characteristics of many other neutron sources (many no longer in operation) may be found in Rae and Good (1970), Firk (1979), Cierjacks (1983), and Whittemore and West (1983).

[b] Peak thermal neutron flux in moderator, n/cm^2 s.

excited states can be obtained from measurements of the total interaction from neutron transmission through the sample and from interactions producing neutron scattering, capture, fission, etc. as a function of incident neutron energy. Neutron spectrometers for nuclear physics must have very good energy resolution with high intensity to observe many resonances in a nuclide and especially to measure their widths.

Although not all spectrometers at a given source are designed alike, generally they consist of

1. a device to prepare the property of the neutron beam prior to the sample—the energy and its resolution, for example;
2. a device to hold and orient the sample in the beam;
3. a device for measuring the scattering angle and making a "post sample" energy analysis of the beam; and
4. a neutron detector and data-acquisition system.

1. NEUTRON SPECTROMETERS AT NUCLEAR REACTORS

A nuclear reactor is usually fueled with a uranium compound, such as UO_2, enriched in the isotope ^{235}U and placed in an aluminum-clad container. Thermal fission of the ^{235}U produces on the average 2.43 neutrons with a mean energy of about 1 MeV. Typically, the fuel is surrounded by a large tank of water, either H_2O or D_2O, for cooling and for moderating the neutron energy to much lower energies. In the neutron moderation process, an intense source of "resonance" energy neutrons is produced with a $1/E$ energy spectrum down to about 0.2 eV. This reactor source of neutrons is excellent for neutron spectroscopy in the energy range up to about 10 keV. Below 0.2 eV the thermal neutron energy spectrum is described by a modified Maxwellian distribution, and a typical example for a source moderated with D_2O is shown in Fig. 1 (Moon, 1985). This figure shows also that the energy spectrum of the source can be changed significantly by inserting either a cold source (e.g., liquid hydrogen) or a hot source (hot graphite block) in the moderator.

Surrounding the water tank is a thick concrete radiation shield through which are beam tubes that direct the moderated neutrons to the spectrometers. For condensed-matter research, some tubes will be oriented to "view" the cold source, others will view the hot source, and others will view the "room-temperature" moderator. A beam tube is typically a cylinder made of aluminum with an effective diameter of 10 to 15 cm. Because of shielding requirements, reactor beam tubes have a minimum length of several meters. Since the flux of neutrons falls off with the square of the distance from the source, short beam tubes would seem to be preferred for high intensity. However, the number of instruments would then be limited by the floor-space requirements of the instruments. One solution to this problem is the use of neutron guide tubes. For thermal and cold neutrons, guide tubes can transport a beam 30 to 100 m using the phenomenon of total reflection. The refractive index for neutrons in a material such as nickel is less than unity. Consequently, a critical angle θ_c exists below which the neutron beam is totally reflected with no attenuation. With two parallel mirrors, neutrons within the angle $2\theta_c$ will pass by multiple reflection, and the beam divergence will decrease much less with increasing distance than with normal beam tubes. A nuclear reactor source of these thermal- and subthermal-energy neutrons is invaluable for countless condensed-matter investigations.

1.1 Neutron Crystal Spectrometers

As indicated above, one component of a neutron spectrometer, whether it is used for cross-section or condensed-matter research, is a device or technique for determining the neutron energy incident on a sample. One of the best techniques for obtaining an intense beam of low-energy neutrons ($E < 10$ eV) with an energy spread (resolution) of about $\Delta E/E = 1\%$ to 5% is that of a single-crystal monochromator placed in a well-collimated beam of neutrons from a high-flux thermal-fission reactor. When a single crystal (such as beryllium, copper, or pyrolytic graphite) is properly oriented in the collimated neutron beam, neutrons of a discrete energy will be scattered elastically (diffracted) from a particular set of planes of atoms in the crystal through an angle of 2θ given by Bragg's law,

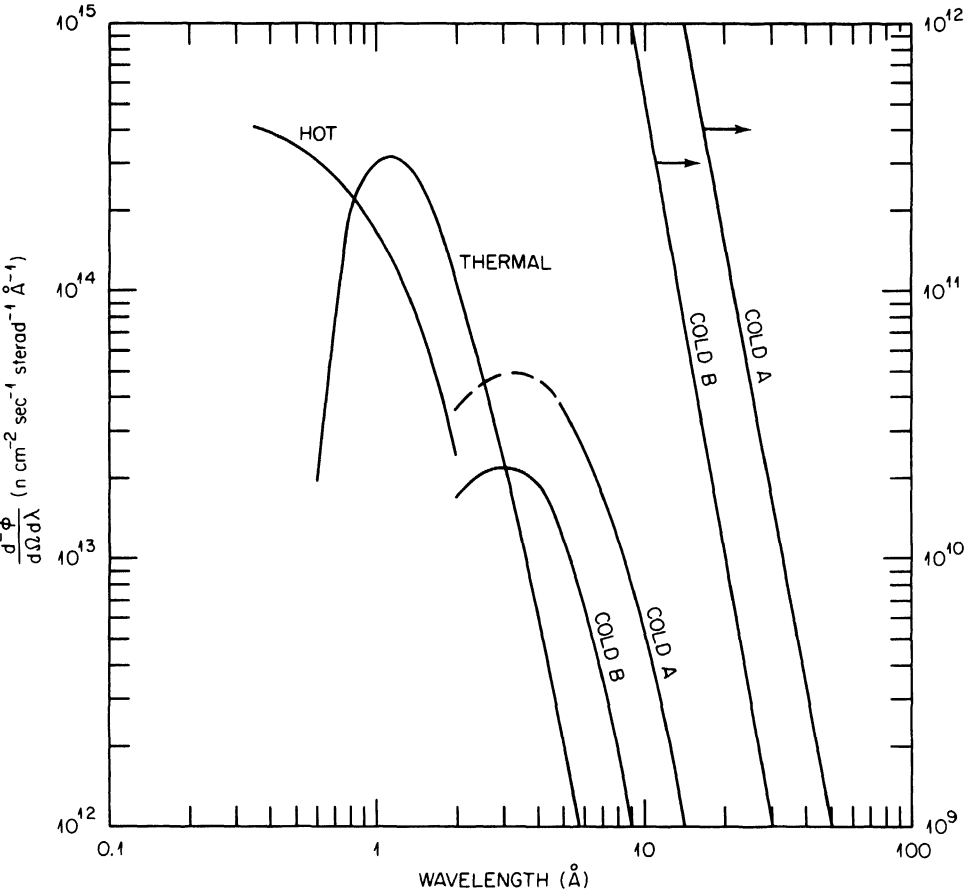

FIG. 1. Thermal-neutron flux distribution from a D_2O-moderated reactor source labeled as thermal and from hot and cold sources.

$n\lambda = 2d\sin\theta$. In this equation, the integer n is the order of the reflection, λ is the wavelength of the neutron, d is the spacing between a particular set of planes of atoms in the crystal, and θ is the angle between the direction of the neutron beam and the set of planes being considered. The neutron wavelength in centimeters equals $0.286 \times 10^{-8}E^{-1/2}$, where E is the neutron energy in eV. Also note that the magnitude of the neutron momentum is $|k| = \hbar/\lambda = \hbar E^{1/2}/0.286 \times 10^{-8}$, i.e., $E \propto |k|^2$. The energy of the diffracted beam can be varied by changing the orientation of the crystal with respect to the neutron beam in order to make measurements as a function of neutron energy.

To increase the intensity of the diffracted neutron beam, multiple parallel slits (called Soller slits) and large single crystals are used to produce a collimated beam with large area. The energy resolution of the beam is determined by the beam divergence and the angular spread (mosaic spread) of the crystal planes, both of which produce an uncertainty in θ. From Bragg's equation above and $n = 1$, one finds that for an angular variation $\Delta\theta$, $\Delta\lambda/\lambda = \Delta\theta \cot\theta$, giving $\Delta E/E = -2\Delta\theta \cot\theta$. The total energy spread from the combined effects of a beam divergence of α_0 and α_1 before and after the monochromator and a monochromator mosaic spread of η_M is

$$\frac{\Delta E}{E} = -2\cot\theta_M\left(\frac{\alpha_0^2\eta_M^2 + \alpha_1^2\eta_M^2 + \alpha_0^2\alpha_1^2}{\alpha_0^2 + \alpha_1^2 + 4\eta_M^2}\right)^{1/2}. \tag{1}$$

Focusing principles that are well known

from optics can also be applied to neutron spectrometer design with a great increase in intensity because samples are often much smaller than neutron beam dimensions. Beam focusing both parallel (usually horizontal) and perpendicular (vertical) to the plane of scattering has been utilized. The case of vertical focusing is easy to illustrate. If one imagines a monochromator crystal shaped in the form of a semicylindrical mirror with the axis horizontal, then neutrons from a point on the neutron source will be focused to a point on the sample if

$$1/L_0 + 1/L_1 = 2\sin\theta/R, \tag{2}$$

where L_0 and L_1 are the distances from the source to the monochromator and from the monochromator to the sample, respectively, R is the radius of curvature of the mirror, and θ is the Bragg angle of the diffracting planes. The intensity gain for a small sample can be expressed as $(h_M/h_s)(L_0 + L_1)/L_1$, where h_M is the height of the monochromator and h_s is the height of the source. This gain can be as large as a factor of 3.

1.1.1 Triple-Axis Spectrometer The monochromator described above is on the first axis of one of the most widely used neutron spectrometers ever designed for condensed-matter research, the triple-axis spectrometer pioneered by Brockhouse (1961), for which he received the 1995 Nobel prize in physics. A schematic diagram of a triple-axis spectrometer located at the High Flux Isotope Reactor of the Oak Ridge National Laboratory, Oak Ridge, Tennessee, is shown in Fig. 2. The neutron beam diffracted by the monochromator crystal is directed onto the specimen located on the second axis, and the energy of the neutrons deflected (scattered) by the specimen through the angle Φ is determined by Bragg diffraction in the analyzer crystal on the third axis. The neutron detector is usually a tube, e.g., 5 cm diam × 15 cm long, oriented with the axis vertical and filled with ^{3}He gas. Neutron detection is by means of the charged particles produced in the $^3\mathrm{He}(n,p)$ reaction.

The great advantage of this spectrometer is the capability it provides to measure, simultaneously and systematically, predetermined values for the momentum change $\mathbf{Q} = \mathbf{k}_i - \mathbf{k}_f$ and the energy change $\hbar\omega = E_i - E_f \propto |\mathbf{k}_i|^2 - |\mathbf{k}_f|^2$ of the neutron due to its interaction with the specimen. Furthermore, the measurements of the neutron intensity corresponding to a wide range of $\hbar\omega$ can be made for a fixed $\mathbf{Q}$. For illustration, if we consider a coordinate system with the x axis parallel to $\mathbf{k}_i$, then $Q_x = |\mathbf{k}_i| - |\mathbf{k}_f| \cos\Phi$ and $Q_y = |\mathbf{k}_f| \sin\Phi$, where Φ is the angle (see Fig. 2) of the scattered neutron relative to the direction of the incident beam. Many pairs of wave vectors $|\mathbf{k}_i|$ and $|\mathbf{k}_f|$ and angles Φ will give the same $\mathbf{Q}$ but different values for $\hbar\omega$. This capability of the spectrometer is particularly desirable for data analysis, because usually the comparison between theory and experiment is most easily carried out for fixed $\mathbf{Q}$. In a typical experiment, either $\mathbf{k}_i$ or $\mathbf{k}_f$ is fixed by selecting the Bragg angle of the monochromator, θ_M, or the analyzer angle θ_A. The other wave vector, and hence the energy transfer, are varied by scanning the appropriate angle. $\mathbf{Q}$ is fixed by simultaneously scanning Φ. If it is necessary to fix $\mathbf{Q}$ with respect to a reference coordinate system of the specimen, then the angle Ψ is scanned also along with the other angles. While such scans are complex, with the computer control of modern spectrometers many such scans are easily carried out automatically. See NEUTRON SCATTERING.

1.1.2 Polarized-Beam Triple-Axis Spectrometer The neutron possesses a spin of $\frac{1}{2}$ and a large magnetic moment. This magnetic moment can interact with and, thereby, provide much information about the magnetic moments of unpaired electrons within the specimen. The neutrons emerging from the reactor beam tube are unpolarized; that is, they can be considered to be equally distributed between up (↑ or +) and down (↓ or −) spin states. It is possible to replace the monochromator and analyzer crystals in the triple-axis spectrometer with crystals that are polarization sensitive. Both are magnetic devices that have zero reflectivity for (−) neutrons, for example. In addition, electromagnetic devices for maintaining or reversing the polarization (flippers) are positioned in the beam before and after the specimen. With both flippers off, the beam incident on the specimen is in the (+) spin state, and only those neutrons that are scattered without change of spin will be reflected by the analyzer. We thus measure a (+ +) scatter-

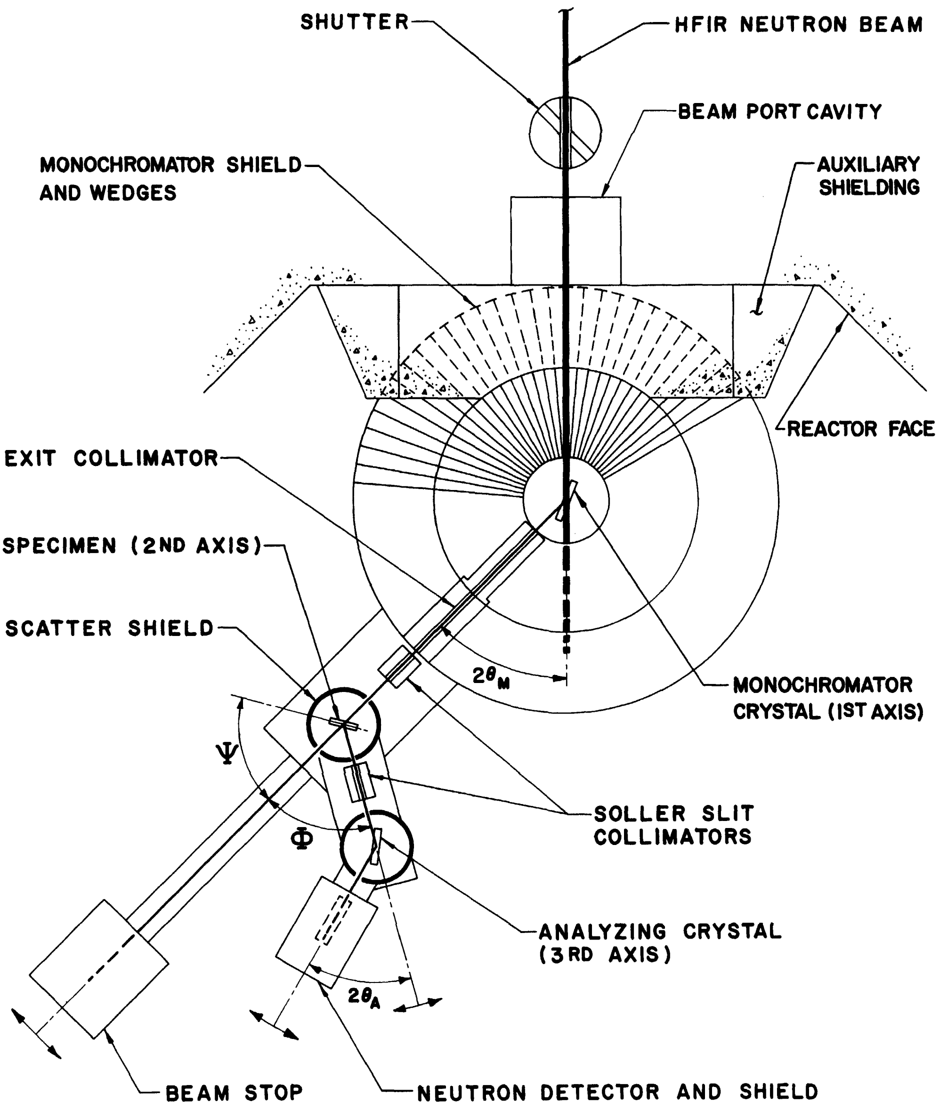
SHUTTER
HFIR NEUTRON BEAM
BEAM PORT CAVITY
MONOCHROMATOR SHIELD AND WEDGES
AUXILIARY SHIELDING
REACTOR FACE
EXIT COLLIMATOR
SPECIMEN (2ND AXIS)
SCATTER SHIELD
2θ_M
MONOCHROMATOR CRYSTAL (1ST AXIS)
Ψ
Φ
SOLLER SLIT COLLIMATORS
ANALYZING CRYSTAL (3RD AXIS)
2θ_A
BEAM STOP
NEUTRON DETECTOR AND SHIELD

FIG. 2. A schematic diagram of the triple-axis spectrometer at the High Flux Isotope Reactor at the Oak Ridge National Laboratory.

ing cross section. By setting the various states of the two flippers, we can measure four partial scattering cross sections: spin flip (+ −), (− +), and nonspin flip (+ +), (− −). We thus have a spectrometer that can be used for energy and spin spectroscopy, providing very detailed information about the magnetism of the specimen at the atomic level (Moon *et al.*, 1969).

1.1.3 Spin-Echo Spectrometer The spin-echo method is utilized in a polarized-beam spectrometer in which the energy transfers are measured directly instead of measuring a difference between specified incident and scattered energies. This method allows a very good resolution, of the order 10^{-9} eV, with good intensity (Mezei, 1979). The neutron spin is utilized for energy measurement in this spectrometer, not for measurement of the magnetic properties of the sample. The method is illustrated in Fig. 3. Neutrons from a cold source are monochromatized and then polarized by reflection from a magnetized neutron mirror. The neutron polarization is first rotated into the hor-

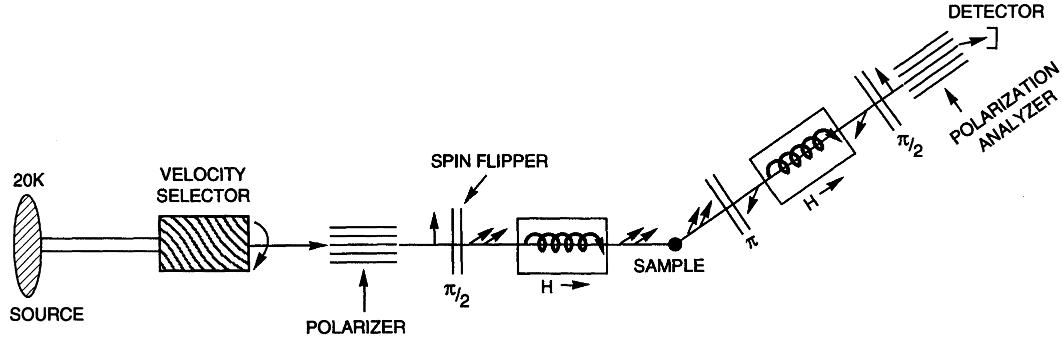

FIG. 3. A schematic illustration of the spin-echo spectrometer using polarized cold neutrons from a reactor.

izontal by a $\pi/2$ flipper. The incident neutrons then precess hundreds of rotations in a magnetic field aligned parallel to the beam path before the sample. The total rotation angle α is inversely proportional to the neutron velocity v_i of the incident neutron.

After the sample, the neutron spins are first rotated by π, and then the scattered neutron spins are passed through another precession field to reverse (unwind), or produce an "echo" of, the precessions that occurred in the incident beam. If the total unwinding angle is β, then the net precession angle θ from monochromator to the detector is $\theta = \alpha - \beta = \gamma HL(1/v_f - 1/v_i)$. Here H is the magnetic field strength and L is the path length of the beam in the field. The polarization P of the emerging neutrons gives the projection or cosine of the precession angle. For $v_i \cong v_f$, we insert $Q = 2mv_i \sin(\Phi/2)$ and $v_i - v_f = \hbar\omega/mv_i$; hence we can write

$$P_f = P_i \int S(Q,\omega) \times \cos[(2 \sin\Phi/2)^3(\gamma HL/\hbar^2 Q^3)m^2\omega]d\omega. \tag{3}$$

Thus the polarization gives directly the Fourier transform of the scattering function $S(Q,\omega)$. The argument of the transform is varied by varying the precession field H.

1.1.4 Beryllium-Filter Spectrometer If a thick polycrystalline block of beryllium is placed in a neutron beam, all the neutrons with energies above about 5 meV are diffracted out of the beam according to Bragg's law, discussed above. However, neutrons below that energy are transmitted with little attenuation because their wavelengths are too long to satisfy the Bragg law for any set of atomic planes in this material and because beryllium has a negligible neutron-absorption cross section. One therefore has a device that serves as a neutron window only for energies below 5 meV.

The beryllium-filter analyzer is an alternative to the analyzer crystal in a triple-axis spectrometer when the energy of the neutrons emerging from the sample is <5 meV. For an energy transfer $\hbar\omega$ much larger than this, for example, >100 meV, the 5-meV window does not deteriorate the resolution appreciably. In such experiments, the beryllium is simply inserted in the scattered-beam path in front of the detector. The high solid angle for this arrangement results in a large increase in intensity, albeit with a correspondingly poor Q resolution. The other limitation of the spectrometer results from the condition that $E_f \ll \hbar\omega$. This condition is equivalent to $k_f \ll k_i$ and, therefore, $Q \approx k_i \propto E_i^{1/2}$. Consequently, there is no flexibility of preselecting Q. Nevertheless, this type of spectrometer has been extremely useful for certain experiments requiring the measurement of high-energy and low-intensity peaks that are nearly Q independent.

1.1.5 Back-Reflection Analyzer Spectrometer In Sec. 1.1, the energy spread of a neutron beam due to the angular variation $\Delta\theta$ of the beam direction or of the diffracting planes in a crystal was shown to be $|\Delta E/E| = (2 \cot\theta)\Delta\theta$. Obviously for $\theta \approx 90°$, i.e., back reflection, the resolution can be extremely good. For example, by employing nearly back reflection in both the monochromator and analyzer that are made from perfect crystals of silicon, a resolution of up to 0.3

μeV is obtained for a neutron energy of about 2.1 meV. A schematic diagram of the IN10 spectrometer at the Institute Laue Langevin (ILL) is shown in Fig. 4. In this type of spectrometer, the energy of the incident or scattered neutron beam cannot be varied by scanning θ, as is done for the triple-axis spectrometer, since we want to keep θ as near to 90° as possible. One solution developed is to scan the incident energy by scanning the neutron velocity changes caused by the Doppler effect and an oscillating-crystal monochromator that moves parallel to the direction of the reactor beam (Birr *et al.*, 1971). The neutrons diffracted by the monochromator are then deflected (by diffraction) toward the sample by a crystal through which the reactor beam was transmitted before it encountered the monochromator. The neutrons scattered by the sample are subsequently backscattered from the array of analyzer crystals, which are focused on the detectors grouped around the sample. The chopper located in front of the sample provides a means, by the TOF technique (see Sec. 1.3 below), to distinguish neutrons that have been scattered by the sample directly to the detectors from those that have backscattered from the analyzer crystals.

An alternative to the Doppler effect is to place the monochromator crystal in a cryofurnace. In this configuration, the incident neutron energy is varied by heating or cooling the crystal, thereby changing the spacing d of the diffracting planes. For example, heating a KCl monochromator from liquid nitrogen to near its melting point provides an energy scan of $-160\ \mu\text{eV} < \hbar\omega < 16\ \mu\text{eV}$ with about 1.0-μeV resolution.

1.2 Mechanical Neutron Velocity-Selector Spectrometers

A mechanical neutron velocity selector is a device that produces a narrow velocity band of neutrons by mechanical means from a steady-state source of neutrons with a broad energy spread. This device is useful for neutron activation and scattering measurements using thermal- and subthermal-energy neutrons, since a crystal spectrometer may have problems in these low-energy regions from higher-order reflections. One technique for selecting velocities consists of several rotating disks of cadmium, containing radial slits, mounted many centimeters apart on a single axis, which is parallel to a collimated neutron beam from a nuclear reactor. The energy of the neutrons passing through the slits is determined by the spacing between the disks, the angular displacement of the slits in adjacent disks, and the speed of rotation. The intensity and energy resolution are determined by the width of the slits, their distance from the axis of rotation, and the rotational speed. For reasonable rotational speeds and with slit widths sufficiently large to give usable intensities, the energy resolution is about 10%.

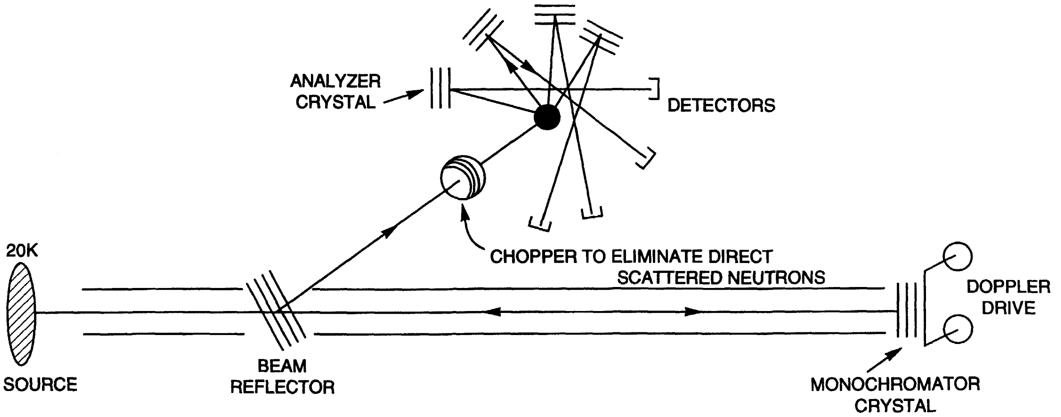

FIG. 4. A schematic diagram of the IN10 back-reflection analyzer spectrometer at ILL reactor, Grenoble, France.

1.3 Neutron Time-of-Flight Spectrometers

Another method to determine the energy of a neutron is to measure its velocity from the time interval that the neutron takes to travel a specified distance, called the time-of-flight (TOF) technique. This method requires the determination of well-defined beginning and ending times for the neutron's transit along the beam path. For a reactor source that is producing a continuous beam of neutrons, it is necessary to "chop" the beam by passing it through an opening in a rotating disk or cylinder to produce microsecond bursts of neutrons and to have an efficient neutron detector located many meters away at the end of a flight path to determine the neutron's flight time. Long flight paths must be evacuated to prevent serious loss of neutrons, and shorter ones often use helium gas, which has a small total cross section for eV and keV neutrons.

The energy resolution of a TOF spectrometer, ΔE, is equal to $2(\Delta t/t)E$, where Δt is the time width of the neutron burst plus the time spread in the neutron detector and electronics and t is the TOF of the neutron. This TOF in microseconds is $72.3L/E^{1/2}$, where L is the path length in meters and E is in eV. The resolution of a TOF spectrometer is often quoted in microseconds per meter. For low neutron energies, the energy resolution in eV is equal to $0.0277E^{3/2}(\Delta t/L)$, where E is in eV and $\Delta t/L$, the resolution, is in microseconds per meter. The time between pulses, the flight path length, and filters in the beam must be selected so that the low-energy neutrons from one pulse do not interfere with high-energy neutrons from the following pulse. By the use of a multichannel storage system utilizing a computer, the complete neutron spectrum consisting of thousands of data points can be obtained with good energy resolution over a broad energy range.

1.3.1 Fast–Neutron-Chopper Spectrometers Fast–neutron-chopper spectrometers are used for spectroscopy in the eV and keV neutron energy regions. A fast chopper consists of a rather massive rapidly rotating shutter in a well-collimated beam of neutrons capable of stopping, or greatly attenuating, neutrons of all energies and also gamma rays from the neutron source. A schematic section of the ORNL fast-chopper TOF spectrometer constructed by Block *et al.* (1960) is shown in Fig. 5. This chopper was capable of producing 1-μs bursts with sufficient intensity from the high flux reactor ORR (neutron flux of 4×10^{14} n/cm^2 s) for transmission measurements using a 180-m flight path. For transmission measurements

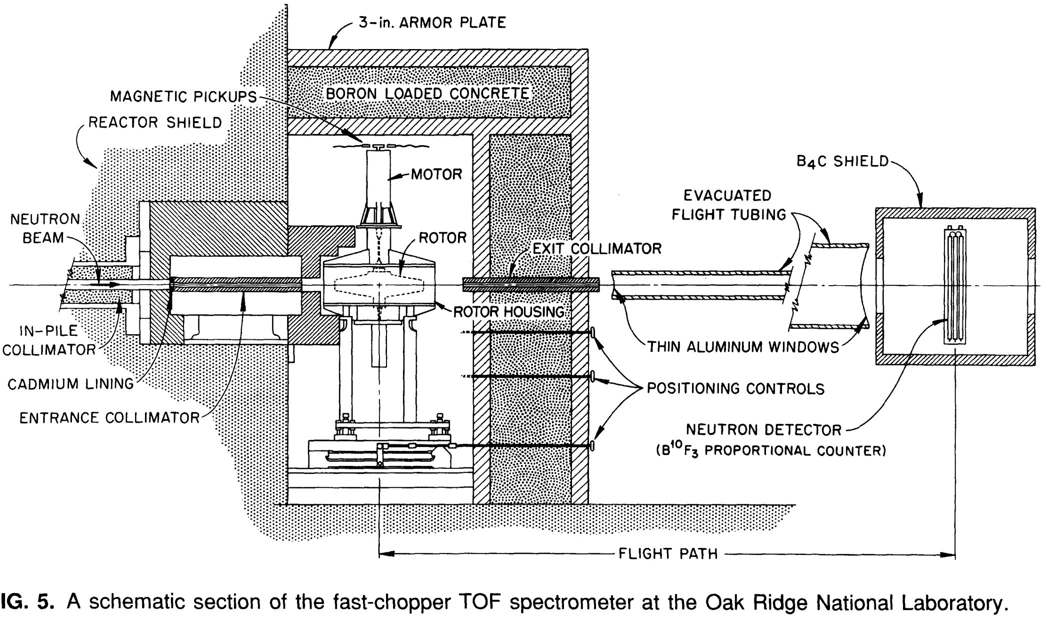
3-in. ARMOR PLATE
MAGNETIC PICKUPS
BORON LOADED CONCRETE
REACTOR SHIELD
MOTOR
B$_4$C SHIELD
EVACUATED FLIGHT TUBING
NEUTRON BEAM
ROTOR
EXIT COLLIMATOR
ROTOR HOUSING
IN-PILE COLLIMATOR
THIN ALUMINUM WINDOWS
CADMIUM LINING
POSITIONING CONTROLS
ENTRANCE COLLIMATOR
NEUTRON DETECTOR
(B^{10}F$_3$ PROPORTIONAL COUNTER)
FLIGHT PATH

FIG. 5. A schematic section of the fast-chopper TOF spectrometer at the Oak Ridge National Laboratory.

an array of 100 $^{10}BF_3$ gas proportional counters 3 cm in diameter and over 1 m long was used as the detector. This resolution of only 0.006 μs/m gives an energy resolution of 0.2 eV at 100 eV (less than the Doppler width of resonances at this energy) and 5 eV at 1000 eV, which is less than the average spacing of resonances for most nuclides and is sufficient for studying some nuclides up to about 10 keV. The low-energy cutoff for a chopper spectrometer depends on the width of the slits, the length of the slit in the rotor, the speed of rotation, and the curvature or shape of the slit along its length, as discussed by Chrien (1983). For partial cross-section measurements, spectrometers with wider slits and shorter flight paths must be used. A spectrometer installed at the BNL High Flux Beam Reactor by Chrien and Reich (1967) produced 5-μs bursts. This spectrometer had sufficient flux and energy resolution using a 22-m flight path for partial cross-section work including capture gamma-ray spectral measurements up to several hundred eV.

1.3.2 Slow-Neutron-Chopper Spectrometers A "slow"-chopper spectrometer (first used by Fermi in 1943) is a much simpler device than a fast chopper since thermal-energy neutrons can easily be stopped by a small thickness of cadmium. Also, detectors can be made that are very efficient for thermal-energy neutrons and inefficient for higher-energy neutrons and gamma rays. Burst widths of 20 μs and flight paths of only 2 m produce an energy resolution ($\Delta E/E$) of 3% at 0.01 eV, which is adequate for many measurements. For neutron measurements below 0.002 eV, filters of beryllium or graphite are often used since they are almost transparent to these low-energy neutrons and attenuate higher-energy neutrons.

Two slow-chopper spectrometers using TOF techniques have been used for condensed-matter research. One utilizes a crystal monochromator to produce a monoenergetic incident beam that is followed by a chopper to define the beginning time for the TOF measurement (Fig. 6). This design is analogous to the triple-axis spectrometer, with the function of the post-sample analyzer crystal replaced with the TOF measure-

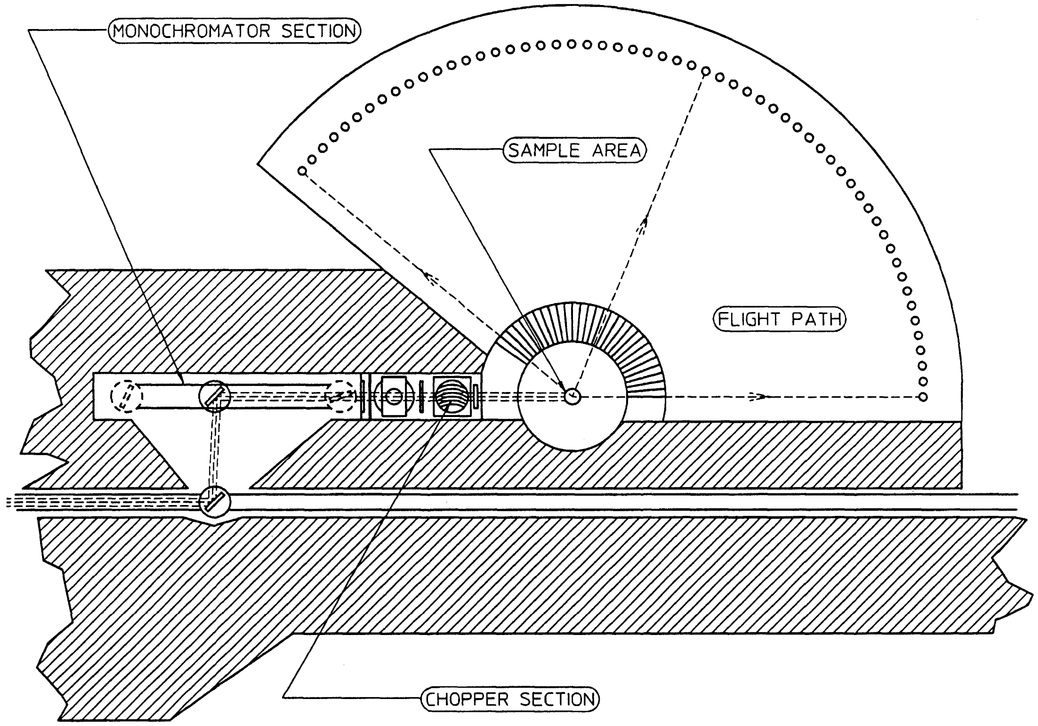

FIG. 6. A schematic of a combination of a crystal monochromator followed by a slow chopper at the National Institute of Standards and Technology reactor research facility.

ment. Such a design is very efficient for collecting spectroscopy data for many scattering angles, and therefore many values of *Q*, simultaneously. The large, space-consuming analyzer-crystal machinery and shielding is replaced with many small detectors located in an evacuated flight path. Each detector measures the time spectrum, and hence the energy spectrum, of the neutrons scattered at the angle corresponding to that detector. Note, however, that constant angle is not constant *Q*.

For the second spectrometer the neutron energy incident on the sample is determined from the preselected transit time between two (or more) choppers separated by a known distance. The first chopper produces a burst of neutrons with a broad distribution of energies. A second chopper is needed to produce a pulsed beam with an acceptable energy resolution and to determine the zero of time for the subsequent TOF spectroscopy of the neutrons scattered by the sample. The neutron energy incident on the sample can be varied by changing the relative phase of the rotations of the choppers, i.e., the time the second chopper is open relative to the time the first chopper is open. If additional choppers are placed before the sample, one can reduce the energy spread incident on the sample, to improve the signal-to-noise ratio and to suppress frame overlap problems (Copley and Udovic, 1993).

2. NEUTRON SPECTROMETERS AT ACCELERATORS

Intense sources of fast neutrons have also been produced from the bombardment of thick targets of heavy elements such as tantalum or uranium with high-current beams of MeV protons, deuterons, and electrons from accelerators. For 50-MeV deuterons on a heavy-element target, neutrons are produced mainly from a nuclear evaporation process and have a Maxwellian energy distribution that extends from about 0.1 MeV to tens of MeV with a peak at about 1 MeV. About 0.3 neutron is produced per deuteron. With 800-MeV protons on a heavy-element target, neutrons are produced by a spallation process, which results in tens of neutrons per proton with energies up to 800 MeV. With an electron linear accelerator with an energy >50 MeV, neutrons are produced by a photoneutron process from bremsstrahlung generated from the electrons striking a heavy-element target. Approximately one neutron is produced for twenty 140-MeV electrons with a thick tantalum target. Most accelerators produce narrow time pulses of charged ions or electrons, and, hence, the fast neutrons are also produced in short time pulses, of the order of nanoseconds, which are needed for TOF spectroscopy for MeV neutron energies. Neutron moderators, which are used to produce intense sources of neutrons below 100 keV and down to thermal energies, increase the pulse width of the moderated neutron pulse.

Useful sources of monoenergetic neutrons down to about 50 keV are produced from the bombardment of thin targets of light elements such as ^{7}Li or ^{3}H with monoenergetic MeV proton or deuteron beams from Van de Graaff or other higher-energy accelerators.

2.1 Variable-Energy Neutron Source Spectrometers

When monoenergetic neutron sources are used for neutron spectrometry for nuclear physics research, measurements are made at a specific neutron energy, then the energy is changed by a small amount (only 0.1% to 1.0%), and a second measurement is made. The process is repeated until the entire energy region of interest is covered. For most of these measurements with monoenergetic MeV neutrons, a pulsed source is not needed. However, for nuclear inelastic neutron-scattering measurements, a technique has been developed that requires nanosecond pulses of monoenergetic MeV neutrons.

2.1.1 Van de Graaff Accelerator Spectrometer Monoenergetic neutrons from Van de Graaff accelerators were the principal source for neutron spectroscopy in the energy range from a few keV to 20 MeV during the 1940s and 1950s. The most useful (*p*,*n*) reactions to cover the neutron energy range from a few keV to a few MeV are those on ^{7}Li and ^{3}H targets, as reviewed by Uttley (1983). The (*d*,*n*) reaction on deuterium is useful from about 1 to 10 MeV, and the (*d*,*n*) on tritium from about 12 to 20 MeV. In the energy range up to 1 MeV, an energy resolution of about 1 keV is possible, which is not adequate for most measurements below about 50 keV neutron energy. The measurement of a complete neutron cross section of a nuclide up to 1 MeV may require up to several thousand sequential energy measurements at slightly different energies, each requiring several minutes. Monoenergetic neutron sources above about 0.1-MeV neutron energy are invaluable for activation measurements that cannot be made using the TOF technique and for nuclear inelastic scattering measurements. See ACCELERATORS, POTENTIAL-DROP.

2.1.2 Cyclotron and Proton Linear-Accelerator Spectrometers For neutron energies above about 21 MeV, truly monoenergetic sources are not possible since additional excited states in the final nucleus are populated for the three most popular targets used for lower neutron energies. Also, multiparticle breakup of the incident particle (if deuterons or tritons are used) produces an intense source of lower-energy neutrons. Hence, even though the intensity of the $^7\mathrm{Li}(p,n)$ monoenergetic neutron source with about 1% resolution up to 300 MeV is comparable to that from the LANL-WNR pulsed neutron source discussed in Sec. 2.2.4, the TOF technique is usually favored for this high-energy region. See CYCLOTRONS and ACCELERATORS, LINEAR.

2.2 Neutron Time-of-Flight Spectrometers

In the late 1930s and early 1940s, accelerators with microsecond pulses and neutron moderators were used for TOF spectroscopy up to only about 100 eV. However, after short-pulse high-intensity cyclotrons, high-current electron linear accelerators, and spallation sources were developed, high-resolution TOF nuclear data could be obtained up to hundreds of MeV. The TOF technique is the only method for high-resolution neutron spectroscopy in the keV energy region. Moderators, which are used to increase the intensity of neutrons below about 100 keV, increase the width of the neutron pulse. For a dense hydrogenous moderator, this time spread of the low-energy leakage neutrons in nanoseconds is about $2000/E^{1/2}$, where E is in eV. This time spread is equivalent to a flight-path distance spread ΔL of about 3 cm for low-energy neutrons and increases to about 7 cm for 100-keV neutrons. Hence, adding this moderation length to the accelerator pulse width (Δt) results in an energy resolution that can be written as $\Delta E/E = 2[(\Delta t/t)^2 + (\Delta L/L)^2]^{1/2}$.

2.2.1 Proton Van de Graaff Accelerator Spectrometer In order to span the neutron energy region from a few keV (the upper limit of a fast-chopper spectrometer) to about 50 keV (the useful lower limit of the variable-energy monoenergetic Van de Graaff source), a pulsed Van de Graaff TOF neutron spectrometer was developed at the ORNL by Good *et al.* (1958). With a proton-bombarding energy about 20 keV above the threshold energy of 1.88 MeV for the $^7\mathrm{Li}(p,n)$ reaction and a moderately thick $^7\mathrm{Li}$ target, 1-ns neutron pulses can be produced with well-defined upper and lower energy limits, from about 3 to 80 keV. Using a flight-path length of only 1 m results in a nominal resolution of only 0.001 μs/m, which corresponds to a percentage energy resolution of 0.4% at 20 keV and 0.8% at 80 keV. However, using a 1-cm-thick $^{10}\mathrm{B}_4\mathrm{C}$ slab with a NaI crystal directly behind as the neutron detector with such a short flight path limits the energy resolution to 2% over the entire energy region. With no low-energy neutrons and short flight paths, high repetition rates up to nearly 10^6 pps can be used without problems from overlap neutrons produced from few-keV neutrons from the previous pulse.

The initial incentive to produce short nanosecond pulses of protons was to obtain monoenergetic MeV neutrons to investigate the excited states of nuclei from inelastic neutron scattering. Various techniques have been developed to produce nanosecond pulses with repetition rates up to 10^6 pps and to measure neutron flight times with nanosecond accuracy up to a few microseconds. The TOF technique to measure the energy of neutrons scattered from a target bombarded with nanosecond pulses of monoenergetic incident neutrons has been the major source of inelastic neutron-scattering data, which are valuable for basic nuclear physics and needed for applied reactor programs.

2.2.2 Electron-Accelerator Photoneutron Source Spectrometers Although electron accelerators were initially built for nuclear physics research such as electron-scattering and photonuclear studies, they have been adapted to be excellent pulsed sources for neutron spectroscopy (Rae and Good, 1970, p. 23; Firk, 1979, p. 550). Electron linear accelerators for pulsed neutron sources have energies from 40 to 180 MeV but have very high peak currents (tens of amperes) in short pulses. Above about 100 MeV electron energy, little is gained in neutron production per kilowatt of electron beam power, and above 200 MeV meson production can cause unwanted backgrounds. Electron linear accelerators have been built at several laboratories in the United States and in other countries mainly for neutron research and have provided the bulk of neutron cross-section data over the entire energy region. See ACCELERATORS, LINEAR.

The latest electron accelerator designed specifically to be used for neutron spectrometry is the Oak Ridge Electron Linear Accelerator (ORELA) [described in review articles by Cierjacks (1983), p. 113, and Bockhoff *et al.* (1990), p. 194]. This accelerator is capable of producing 15-A pulses of 140-MeV electrons, representing a power in the pulse of over 2000 MW. With 24-ns-wide pulses, a beam power of 50 kW is obtained at the maximum repetition rate of 1000 pps. Neutrons are produced from these high-energy electrons by bombardment of a tantalum target (about 4 cm thick) that produces bremsstrahlung, and these high-energy gamma rays then produce neutrons via the (γ,n) reaction. The neutron energy spectrum has a maximum at about 1 MeV and extends from about 10 keV to 80 MeV. The maximum neutron production rate at the ORELA spectrometer in narrow pulses (3–24 ns) is 4×10^{18} n/s. The average production rate is 10^{14} n/s at a repetition rate of 1000 pps and 24-ns-wide pulses.

A schematic of the ORELA facility is shown in Fig. 7. Ten evacuated flight paths with 18 underground stations ranging from 9 to 200 m enable many different experiments to be carried out simultaneously. The target room is also evacuated to eliminate a background due to neutrons scattering from air near the target. A resolution of 0.02 ns/m can be realized using unmoderated neutrons and the 200-m flight station for total and scattering cross-section measurements, which corresponds to a percentage energy resolution of 0.05% at 1 MeV.

The water-cooled tantalum target with a water moderator that is used at ORELA is shown in Fig. 8 (Macklin, 1971). The moderated neutron spectrum has an $E^{-0.78}$ energy distribution down to thermal neutron energies. Below about 300 keV, more usable neutrons are available from the water moderator than from the tantalum target. For most low-energy measurements, "shadow" bars are located in the flight paths or collimation is used to prevent gamma rays (or reduce their intensity) from the tantalum target from reaching the detector directly. The addition of a 5-cm-thick polyethylene moderator on the side of the ORELA target increases the neutron intensity below a few electron volts by up to a factor of 6.

A second target at ORELA, consisting of a thin tantalum converter separated by about 30 cm from a large block of beryllium, is also used for specific MeV neutron measurements. Collimators are designed so that detectors at the ends of flight paths do not "see" the gamma rays from the tantalum converter. Above about 3 MeV, more usable neutrons are available for a neutron transmission measurement from this target than from the thick tantalum target.

Other electron-accelerator spectrometers use somewhat different fast-neutron targets and moderator arrangements. The Geel Electron Linear Accelerator (GELINA), Belgium, has neutron production capabilities very similar to ORELA (Bensussan and Salome, 1978). A unique feature of this facility is a post buncher installed on the accelerator in the mid 1980s that compresses 10-ns, 10-A pulses from the accelerator into 0.6-ns pulses of over 100-A peak current with negligible beam loss. In addition to a stationary uranium target for producing neutrons, a rotating target is used for electron-beam powers above 6 kW. Both are cooled by liquid mercury to avoid neutron moderation. From a thick uranium target, about 0.06 neutron is emitted for each 100-MeV electron. A heterogeneous uranium–beryllium target was also constructed that increased the neutron yield above 10 MeV by a factor of 2 to 3. At GELINA, in addition to polyethylene or water moderators placed close to the uranium target, a liquid-methane moderator has been developed (Bockhoff *et al.*, 1990). This moderator enhances the intensity of neutrons below 0.01 eV by up to a factor of 5 over that

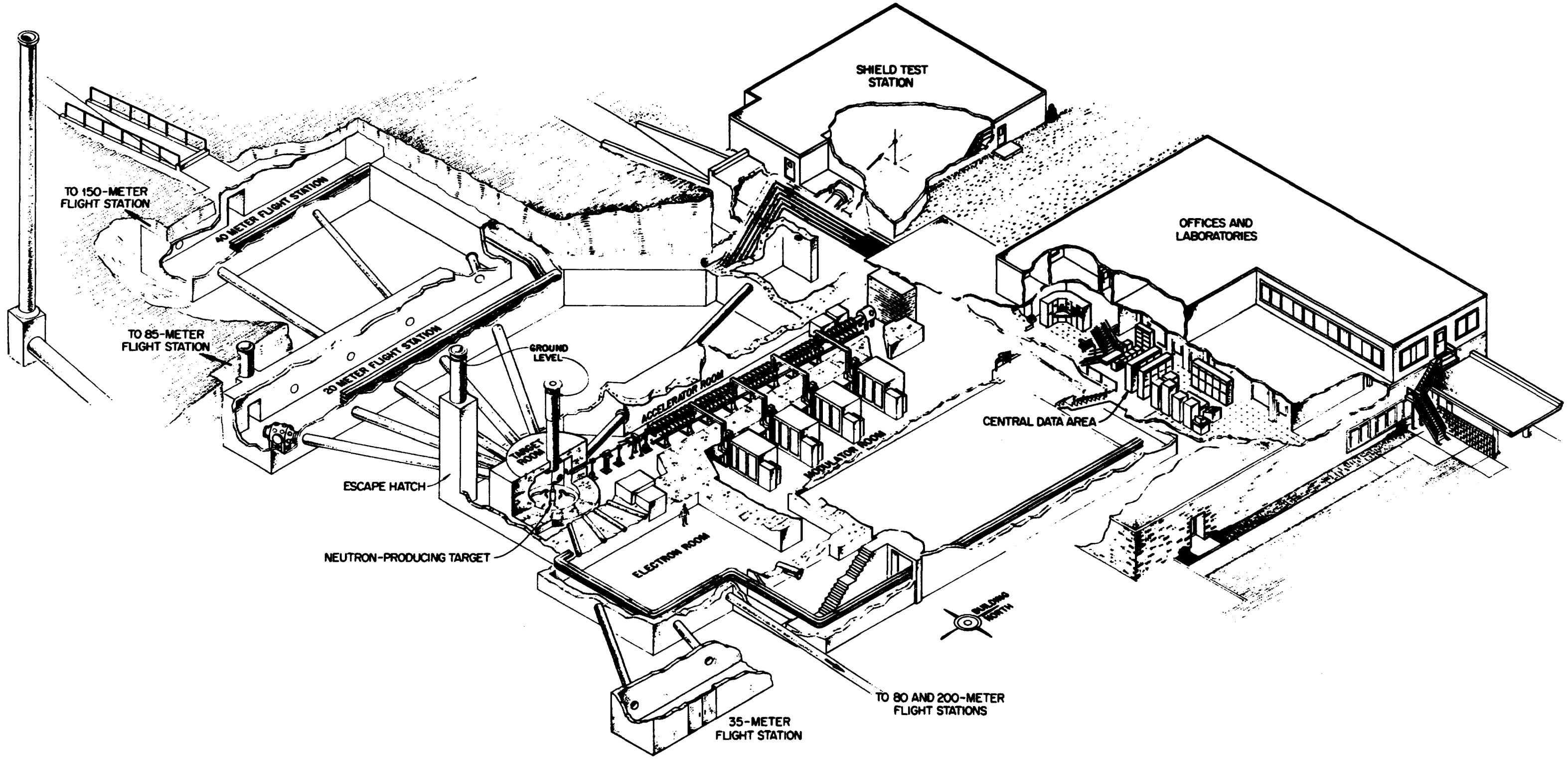
SHIELD TEST STATION
TO 150-METER FLIGHT STATION
40 METER FLIGHT STATION
OFFICES AND LABORATORIES
TO 85-METER FLIGHT STATION
20 METER FLIGHT STATION
GROUND LEVEL
ACCELERATOR ROOM
TARGET ROOM
CENTRAL DATA AREA
MODULATOR ROOM
ESCAPE HATCH
ELECTRON ROOM
NEUTRON-PRODUCING TARGET
BUILDING NORTH
TO 80 AND 200-METER FLIGHT STATIONS
35-METER FLIGHT STATION

FIG. 7. A schematic of the Oak Ridge Electron Linear Accelerator Neutron Spectrometer, ORELA, showing the ten neutron flight paths and the eighteen underground stations.

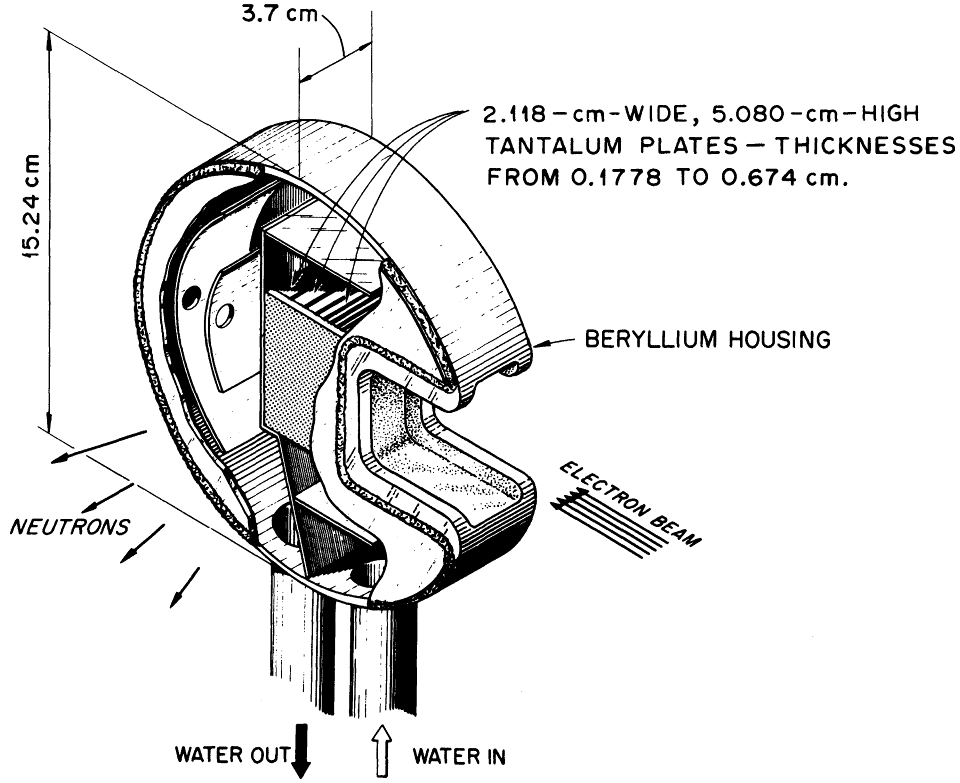
3.7 cm
15.24 cm
2.118-cm-WIDE, 5.080-cm-HIGH TANTALUM PLATES — THICKNESSES FROM 0.1778 TO 0.674 cm.
BERYLLIUM HOUSING
ELECTRON BEAM
NEUTRONS
WATER OUT
WATER IN

FIG. 8. The beryllium-clad tantalum target assembly with H_2O moderator at the ORELA neutron spectrometer.

from a water moderator. The GELINA facility has 31 detector stations and flight-path lengths from 8 to 400 m.

2.2.3 Deuteron-Cyclotron Source Spectrometer The addition of a deflection bunching system to the 50-MeV isochronous cyclotron at Karlsruhe resulted in intense pulses of deuterons only 1 ns in width (Cierjacks, 1983), which is excellent for a neutron spectrometer from 0.3 MeV to about 30 MeV. The peak deuteron current in the pulse of 1.5 A at a repetition rate of 2 × 10^4 pulses/s produces a neutron intensity of 2.2 × 10^{17} n/s sr in the forward direction with a uranium target. Neutrons from the bombardment of a heavy-element target such as tungsten, lead, or uranium with 50-MeV deuterons are produced mainly from (*p,xn*) or (*d,xn*) processes, where *x* specifies the number of neutrons produced in the reaction. The neutrons are produced by an evaporation process and have a Maxwellian energy distribution with a maximum at about 1 MeV and extending from about 0.1 MeV to tens of MeV. Neutrons are also produced up to tens of MeV from direct reactions and from the breakup of the bombarding deuterons. Detector stations at the Karlsruhe facility are located at distances of 12, 56, and 190 m from the neutron source. With a fast efficient detector for neutrons above 300 keV, measurements have been made up to 30 MeV with very good energy resolution. The nominal energy resolution is only 0.04% at 4 MeV using the 190-m flight path. No moderator is used with this Karlsruhe spectrometer since even at 100 keV, the time width of the neutron pulse produced by the moderator would be increased to 5 ns and would require a much lower repetition rate to prevent overlap. See CYCLOTRONS.

2.2.4 Proton Spallation Source Spectrometers The 385-MeV proton synchrocyclotron at Columbia University (built for meson physics research) was the first spallation source used for neutron TOF measurements (Rainwater *et al.*, 1964). A deflected 300-mA proton beam pulse 25 ns wide on a water-cooled lead target produces spallation neutrons with an intensity of 1.6×10^{19} *n*/s at a repetition rate of 60 pulses per second. These properties are best suited for high-resolution neutron spectroscopy below 10 keV and, hence, a 4-cm-thick polyethylene moderator was used to increase the intensity of neutrons in this energy range. The neutron detector was a ^{10}B slab 30 cm high, 120 cm wide, and 4 cm thick and contained 15 kg of ^{10}B located at 200 m with four large NaI crystals either above the slab for most measurements or directly behind for low-energy measurements resulting in a higher efficiency. With a 200-m flight path, the nominal resolution is 0.13 ns/m above about 5 keV, where the moderation time equals the 25-ns pulse width. The resolution at 5 keV is less than the Doppler widths of resonances of even the heaviest nuclides and is sufficient for interesting studies on medium- and lightweight nuclides up to 100 keV.

The 800-MeV proton linear accelerator at the Los Alamos National Laboratory (LANL) has also been used as an intense pulsed neutron source for neutron spectroscopy for neutron energies up to 800 MeV (Lisowski *et al.*, 1990). The Weapons Neutron Research (WNR) facility uses the microstructure of the proton beam from the Los Alamos Meson Physics Facility (LAMPF), i.e., 0.2-ns pulses separated by a few microseconds, to produce neutron pulses having a width of about 1 ns from a tantalum target. Each micropulse contains about 3×10^8 protons and, depending on the experiment and the flight-path length, more than 10 000 micropulses/s may be used. With a thick tungsten target, about 17 neutrons are produced per proton at 800 MeV. Flight-path lengths of the six flight paths are from 9 to 90 m. This unmoderated neutron source is excellent for neutron measurements in the energy region from about 0.3 to 800 MeV. See article on ACCELERATORS, LINEAR.

In the late 1980s, a Proton Storage Ring (PSR) was installed at the 800-MeV proton linear accelerator at LANL (LAMPF) for condensed-matter research. The PSR accumulates an unpolarized H^- beam for about 450 μs and then extracts it in a single 250-ns burst at a repetition rate of 20 pulses per second. This bunched beam of about 80 μA is channeled to a target area to produce an intense source of moderated neutrons. The layout of the Los Alamos Neutron Scattering Center (LANSCE) is shown in Fig. 9 (Lisowski *et al.*, 1990). LANSCE has a split tantalum target with several water moderator and reflector assemblies. One moderator is liquid parahydrogen operated at about 20 K, which results in an increase of 100 in intensity for neutrons below 0.01 eV. The time spread in microseconds of epithermal neutrons is approximately $7/E^{1/2}$, where E is in eV, and the intensity varies as $1/E$. Above 1 keV the resolution for neutron spectroscopy is determined by the pulse width from the PSR. Most of the flight paths at LANSCE are devoted to condensed-matter research, but some are used for basic nuclear physics such as parity violation. The extremely high neutron intensity and low pulse rate make LANSCE very useful for thermal and epithermal neutrons up to 1 keV and even to 100 keV if good energy resolution is not required. In fact, the intensity is so high that for some experiments individual neutron pulses are not counted, but a current as a function of time produced by the detector is recorded.

The most powerful spallation source for condensed-matter studies at the present time is the ISIS facility located at Rutherford Appleton Laboratory in England. At this facility an ion source produces H^- ions, which are accelerated to an energy of 70 MeV prior to injection into a proton synchrotron. The synchrotron then further accelerates the protons to an energy of 800 MeV before they are extracted (50 times a second) and sent to the target station. The target is either tantalum or depleted uranium. For an 800-MeV proton beam, there are typically 30 neutrons per proton from a uranium target.

For condensed-matter research at ISIS, an array of hydrogenous moderators is located around the target. These moderators are at various temperatures to produce different spectral distributions for different types of experiments, much like the situation depicted for reactors in Fig. 1. The moderators are at ambient temperature (316K, H_2O), liquid methane (100K, CH_4), and liquid hydrogen (20K, H_2). Neutron beams are transported through beam tubes from the moderators to various spectrometers and diffractometers as discussed for reactor

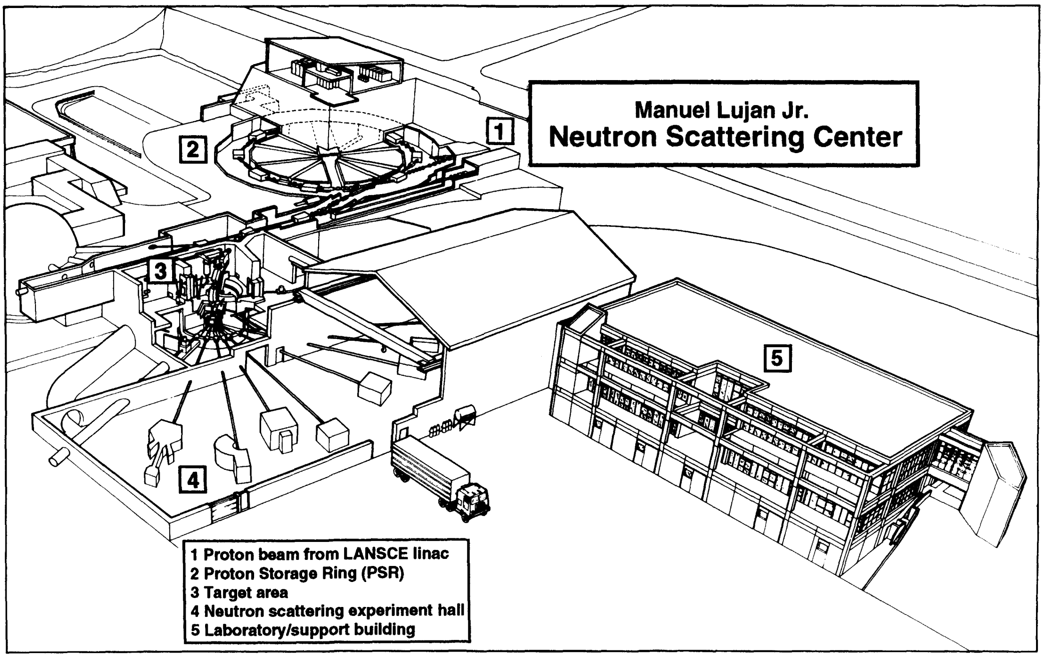
Manuel Lujan Jr.
Neutron Scattering Center
1
2
3
4
5
1 Proton beam from LANSCE linac
2 Proton Storage Ring (PSR)
3 Target area
4 Neutron scattering experiment hall
5 Laboratory/support building

FIG. 9. The layout of the Pulsed Storage Ring and the Los Alamos Neutron Scattering Center, LANSCE.

sources. The pulsed generation of neutrons in a spallation source means that TOF is the natural method to use for spectrometers at such sources. Two general categories of standard spectrometer designs are the most common at pulsed sources: direct geometry and indirect or inverted geometry. Several of the spectrometers located at the ISIS facility are described in Secs. 2.2.5 and 2.2.6. However, similar instruments are also in operation at other spallation sources around the world, such as LANSCE at LANL, IPNS at ANL, and KENS at KEK, Japan (see Table 2).

2.2.5 Direct-Geometry Chopper Spectrometer The direct-geometry chopper spectrometer utilizes the pulse of the source to define the initial time of each neutron's flight. However, the energy distribution of the neutron pulse is as broad as that provided by the moderator. As with many of the reactor-source experiments, it is often necessary to know accurately the neutron energy incident on the sample. Consequently, a slow chopper is located several meters from the source before the sample to monochromatize the incident beam. The High Energy Transfer Spectrometer (HET) at ISIS utilizes a Fermi chopper rotating at frequencies in the range 400–600 Hz for this purpose. Coupling the chopper pulse with the ISIS pulse gives an energy resolution of about 1% in the energy range 0.020–2.0 eV. The energy of the beam that passes through the chopper can be varied at will by altering the phase of the chopper rotation, i.e., altering the time at which the chopper is open relative to the time of the source pulse. Neutrons scattered by the sample are detected by one of many detectors mounted in an evacuated flight path downstream from the sample position, as illustrated previously for the analogous reactor-based spectrometer in Fig. 6. For polycrystalline or liquid samples, a great increase in intensity is achieved on HET by utilizing detectors mounted on annular rings centered on the axis of the incident beam. All detectors on the same ring measure the neutron-scattering spectrum for the same scattering angle and, hence, the same $|\mathbf{Q}|$. The energy of the detected neutron is determined from the total flight time from the Fermi chopper to a detector.

2.2.6 Crystal-Analyzer Time-of-Flight Spectrometer The inverted-geometry spectrometer utilizes an analyzer crystal after the sample to specify accurately the energy of the scattered neutrons. A substantial fraction of the energy spectrum of the white beam emerging from the moderator of a spallation source can be utilized by a spectrometer that has multiple analyzer crystals, each with a given detector, as is illustrated in Fig. 10. This multianalyzer crystal configuration was first used on the KENS spallation source in Japan. A similar version, the PRISMA spectrometer, is located at ISIS. Analogous to the triple-axis spectrometer, each analyzer crystal is set to diffract a preselected energy, but the energy scan is accomplished by measuring the TOF from the source to each detector. The energy transferred to the sample is determined by comparing the total energy from the TOF measurement with that corresponding to the energy of the analyzer setting.

Such a spectrometer has some of the desirable features of a triple-axis spectrometer, while also utilizing the time structure of the pulsed source and multiple analyzers set at many different scattering angles. By appropriate settings of the scattering angles ϕ and analyzer angles θ_A, scans through numerous preselected values of Q,ω can be carried out, thereby achieving a high rate of data collection. The PRISMA spectrometer is described in detail by Steigenberger *et al.* (1991), and a more complete description of indirect-geometry TOF spectrometers is given by Robinson and Pynn (1988).

2.2.7 Pulsed Fission- and Fusion-Source Spectrometers In order to increase the neutron output over that normally obtained from a tantalum or uranium target with an electron linac, a unique target consisting of a small subcritical assembly of ^{235}U was installed on the Harwell accelerator (described by Rae and Good, 1970, p. 34). This "booster" assembly consists of a small cell (about one cm^3) of flowing mercury, which can dissipate up to 5 kW of electron beam power, and several cylinders of ^{235}U. This mercury target (primarily for bremsstrahlung generation) is backed by a solid cylinder of ^{235}U, and both together are surrounded by a set of concentric hollow cylinders of ^{235}U. The total fission power of the booster assembly is about 2.5 kW and gives a factor of 10 enhancement of neutron output over a natural-uranium target. The neutron production rate in the 100-ns burst is 7×10^{18} n/s for 5 kW of 128-MeV electrons. At the maximum repetition rate of 390 pulses per second, the peak neutron yield is 3×10^{14}.

Several "booster" assemblies have been built at Dubna in the USSR and have been used for neutron spectroscopy (described by Whittemore and West, 1983, p. 169; Firk, 1979, p. 554). A neutron multiplication factor of 100 has been accomplished with a plutonium target on an electron accelerator source with a burst width of about 1.7 μs. The Intense Burst Reactor IBR-30 facility operating in a booster mode with a multiplication factor of 200 produced 4-μs pulses 100 times a second. Pulsed reactors (such as the IBR-2) have fast-neutron pulse widths of about 200 μs. Moderators are used with these sources for measurements below about 10 keV.

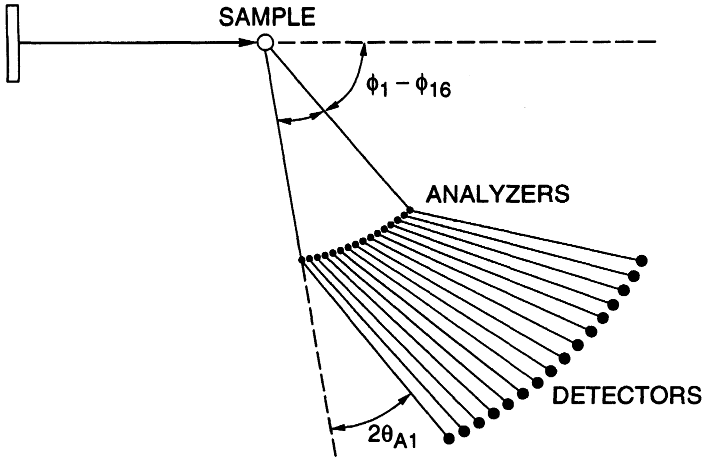

FIG. 10. A schematic of PRISMA, a crystal-analyzer TOF spectrometer at the ISIS facility at the Rutherford Appleton Laboratory, England.

The most intense pulsed neutron source that has been used as a TOF neutron spectrometer is a nuclear fission or fusion explosion. The burst width is about 80 ns, and moderated neutrons extend down to about 20 eV. Flight paths up to 300 m long result in energy resolutions that are poorer in the keV energy region than what is attainable from pulsed accelerators. However, with the extremely high intensity, measurements can be made on highly radioactive materials, on samples with very small cross sections, and on samples only available in very small quantities.

2.2.8 Pulsed-Source Slowing-Down Neutron Spectrometer A unique neutron spectrometer with a broad energy resolution but high intensity was developed in the USSR in the 1960s from a pulsed fast-neutron source located near the center of a large 1-m cube of lead. Fast neutrons are slowed down by scattering from lead nuclei, and the median energy of these moderated neutrons is a function of the slowing-down time. The Rensselaer Intense Neutron Spectrometer (RINS) system (Maguire *et al.*, 1985) consists of a lead slowing-down–time spectrometer coupled to the Rensselaer Polytechnic Institute 80-MeV electron linac. The median neutron energy in electron volts for this RINS system equals $165/(t + 0.3)^2$, where t is the slowing-down time in milliseconds. The percentage energy resolution is about 30% in the electron volt energy region but increases to about 50% at 10 keV. This spectrometer is valuable for measuring fission cross sections of samples available only in very small amounts of nuclides with very small cross sections.

3. NEUTRON DETECTORS

3.1 Neutron Detectors for Spectrometers

Since neutron spectrometers cover a very wide energy span, it is not surprising that a multitude of neutron detectors have been developed, and a variety of techniques have been used for these spectrometers. The important characteristics for these neutron detectors are size, efficiency, timing resolution, discrimination against gamma-ray backgrounds, and discrimination against neutrons whose energies are outside the region of interest. For most spectrometers used in nuclear physics research, the detectors should be quite large (100 cm^2), and efficiencies should be as high as possible (hopefully greater than 10% for the neutrons of interest). For some spectrometers the absolute efficiency of the detector may need to be known to a few percent, but for most spectrometers the relative efficiency as a function of neutron energy is sufficient.

Gas proportional counters using $^{10}BF_3$ or ^3He or thin ^6Li glass scintillators are used for most condensed-matter research spectrometers since they are relatively simple and have high efficiency for thermal-energy neutrons. Inorganic scintillation detectors, which detect charged particles from neutron-induced reactions such as $^{10}B(n,\alpha)$, $^6Li(n,\alpha)$, or $^3He(n,p)$, or the 478-keV gamma ray from the $^{10}B(n,\alpha\gamma)$ reaction, are valuable for neutron spectrometers below about 500 keV.

Since most high-resolution neutron spectrometers for nuclear physics research today use the TOF technique with a pulsed white neutron source, the timing of the detector and electronics should be less than the pulse width of the source in order not to degrade the energy resolution, hence less than a nanosecond for the MeV energy region, but microsecond timing is adequate below 1 eV.

Organic scintillation detectors, which may be a liquid or plastic containing hydrogen, are the most valuable detectors for neutron energies above about 100 keV. Neutrons are detected from the light resulting from recoil protons when neutrons are elastically scattered from hydrogen atoms in the scintillator. Improved low-noise photomultipliers and electronics have made it possible to achieve >10% efficiency for 5-keV energy neutrons and >50% above 20 keV with a scintillator a few centimeters thick (Harvey and Hill, 1979). Very large organic scintillation detectors (called black detectors) have also been constructed that have efficiencies of over 90% up to 10 MeV known to an accuracy of <1%. See DETECTORS, SCINTILLATION.

3.2 Neutron-Detector Spectrometers

In addition to simply detecting neutrons, some detectors are also able to determine the energy of the neutron that induced the event. For example, for an organic scintillator, the pulse height is directly related to the energy of the recoiling proton. The pulse-height distributions for many sizes of organic scintillators have been calculated by Monte Carlo techniques for neutron energies from 0.1 to 70 MeV. For a 4.6-cm × 4.65-cm-diam organic scintillator, the neutron energy resolution is about 10% up to 14 MeV neutron energy. The energy resolution of a thin organic scintillation spectrometer with a transmission of about 50% can range from 10% at 1 MeV to about 3% at 20 MeV. For very large scintillation detectors, the energy resolution is about 30%.

Both ^{3}He gas ionization chambers and ^{3}He gas and liquid scintillation detectors have been used as spectrometers for neutrons whose energies are somewhat greater than the 0.764-MeV Q value for the (n,p) reaction. With a ^{3}He-gas cylindrical gridded ionization chamber of 6.0-cm diam filled with 6 atm of ^{3}He, 3 atm of Ar, and 1 atm of CH_4, an energy resolution of about 2% has been obtained for incident neutrons around 3 MeV. A gas scintillator containing 90% ^{3}He and 10% Xe at a gas pressure of 20 atm has yielded a pulse-height resolution of about 14% in the MeV energy range. A liquid-^{3}He scintillator that has an efficiency of several percent for MeV neutrons has a percentage neutron energy resolution from 0.8 to 8.5 MeV of $21/(E + 0.764)^{1/2}$, where E is the neutron energy in MeV.

GLOSSARY

Cold Neutrons: Neutrons with energies <0.002 eV.

Doppler Broadening: The broadening of a nuclear resonance due to the thermal motion of the target nucleus.

Epithermal-Energy Neutrons: Neutrons with energies >0.2 eV.

Fast Chopper: A massive rotating shutter to produce microsecond bursts of eV and keV neutrons.

ISIS: Spallation neutron source located at the Rutherford Appleton Laboratory, England.

Monochromator: A device for selecting monoenergetic neutrons from a white neutron source with a broad energy spread.

Monoenergetic: $\Delta E/E$ from about 0.001 to 0.1.

Neutron Energy, E: $= 939.553\,[(1 - v^2/c^2)^{-1/2} - 1]$ MeV, where v is the velocity of the neutron and c is the speed of light. For neutron energies below 1 MeV, the approximation $E = mv^2/2$ or time of flight in μs/m $= 72.298/E^{1/2}$ is adequate.

PRISMA: A multiple–analyzer-crystal TOF neutron spectrometer located at the ISIS spallation neutron source.

Slow Chopper: A shutter of cadmium for producing pulses of thermal and subthermal neutrons >10 μs wide.

Spallation: A nuclear process for producing MeV energy neutrons from nuclei bombarded by energetic protons >100 MeV.

Subthermal-Energy Neutrons: Neutrons with energies <0.002 eV.

Thermal-Energy Neutrons: Neutrons with energies from 0.002 to 0.2 eV.

White-Neutron Source: A neutron source with a broad range of energies, often many decades.

Works Cited

Bensussan, A., Salome, J. M. (1978), *Nucl. Instrum. Meth.* **155,** 11–23.

Birr, M., Heidemann, A., Alefeld, B. (1971), *Nucl. Instum. Meth.* **95,** 435–439.

Block, R. C., Slaughter, G. G., Harvey, J. A. (1960), *Nucl. Sci. Eng.* **8,** 112–121.

Bockhoff, K. H., Carlson, A. D., Wasson, O. A., Harvey, J. A., Larson, D. C. (1990). *Nucl. Sci. Eng.* **106,** 192–207.

Brockhouse, B. N. (1961), *Inelastic Scattering of Neutrons in Solids and Liquids,* Vienna: IAEA, p. 113–151.

Chrien, R. E. (1983), "Pulsed Neutron Beams from Steady State Reactors," in: S. Cierjacks (Ed.), *Neutron Sources for Basic Physics and Applications,* New York: Pergamon Press, Chap. IX, 197–216.

Chrien, R. E., Reich, M. (1967), *Nucl. Instrum. Meth.* **53,** 93–107.

Cierjacks, S. (1983), "Accelerator-Based Pulsed White Neutron Sources," in: S. Cierjacks (Ed.), *Neutron Sources for Basic Physics and Applications,* New York: Pergamon Press, Chap. V, 81–132.

Copley, J. R. D., Udovic, T. J. (1993), *J. Res. NIST* **98,** 71–87.

Firk, F. W. K. (1979), "Neutron Time-of-Flight Spectrometers," *Nucl. Instrum. Meth.* **162,** 539–563.

Good, W. M., Neiler, J. H., Gibbons, J. H. (1958), *Phys. Rev.* **109,** 926–933.

Harvey, J. A., Hill, N. W. (1979), "Scintillation

Detectors for Neutron Physics Research," *Nucl. Instrum. Meth.* **162,** 507–529.

Liskowski, P. W., Bowman, C. D., Russell, G. J., Wender, S. A. (1990), *Nucl. Sci. Eng.* **106,** 208–218.

Macklin, R. L. (1971), *Nucl. Instrum. Meth.* **91,** 79–84.

Maguire, H. T. *et al.* (1985), *Nucl. Sci. Eng.* **89,** 293–304.

Mezei, F. (1979), *Nucl. Instrum. Meth.* **164,** 153–156.

Moon, R. M. (1985), in: G. H. Lander, V. J. Emery (Eds.), *Nucl. Instrum. Meth. B* **12,** Appendix B, 556–560.

Moon, R. M., Riste, T., Koehler, W. C. (1969), *Phys. Rev.* **181,** 920–931.

Rae, E. R., Good, W. M. (1970), in: J. A. Harvey (Ed.), *Pulsed Accelerator Time-of-Flight Spectrometers in Experimental Neutron Resonance Spectroscopy,* New York: Academic Press, pp. 1–99.

Rainwater, J., Havens, W. W., Jr., Garg, J. B. (1964), *Rev. Sci. Instrum.* **35,** 263–275.

Robinson, R. A., Pynn, R. (1988), *Nucl. Instrum. Meth. A* **272,** 758–762.

Steigenberger, U., Hagen, M., Cacuiffo, R., Petrillo, C., Cilloco, F., Sachetti, F. (1991), *Nucl. Instrum Meth. B* **53,** 87–96.

Uttley, C. A. (1983), "Sources of Monoenergetic Neutrons," in: S. Cierjacks (Ed.), *Neutron Sources for Basic Physics and Applications,* New York: Pergamon Press, Chap. III, 19–55.

Whittemore, W. L., West, G. B. (1983), "Pulsed Reactor Sources," in: S. Cierjacks (Ed.), *Neutron Sources for Basic Physics and Applications,* New York: Pergamon Press, Chap. VII, 157–175.

Further Reading

Brooks, F. D. (1979), "Development of Organic Scintillators," *Nucl. Instrum. Meth.* **162,** 477–505.

Dickens, J. K. (Ed.) (1994), *Proceedings of the International Conference on Nuclear Data for Science and Technology,* La Grange Park, IL: American Nuclear Society.

Hughes, D. J. (1953), *Pile Neutron Research,* Reading, MA: Addison-Wesley.

Igarasi, S. (Ed.) (1988), *Proceedings of the International Conference on Nuclear Data for Science and Technology,* May 30–June 3, 1988, Mito, Japan: Saikon Publishing Co. Ltd.

Lander, G., Axe, J., Endoh, Y. (Eds.) (1995), *Neutron News,* Vol. 1 (1990) to Vol. 6 (1995), New York: Gordon and Breach.

Price, D. L., Skold, K. (Eds.) (1987), *Methods of Experimental Physics-Neutron Scattering,* Vols. 23A, B, and C, Orlando, San Diego: Academic Press.

Quim, S. M. (Ed.) (1992), *Proceedings of the International Conference on Nuclear Data for Science and Technology, May 13–17, 1991, Jülich, FRG,* Berlin: Springer-Verlag.

SPECTROMETERS, OPTICAL

See UV AND VISIBLE LIGHT SPECTROMETERS

4
X-Ray Spectroscopy

Thomas H. Markert[†] *and Eckhart Förster*

4.1 **Introduction 89**
4.2 **General Concepts 90**
4.3 **Dispersive Spectrometers 91**
4.3.1 Bragg Crystal Devices 91
4.3.1.1 Bragg's Law 91
4.3.1.2 Flat Crystal Bragg Spectrometer 93
4.3.1.3 Bent-Crystal Spectrometer 95
4.3.1.4 Bragg Spectroscopy with Synthetic Crystals (Multilayers) 98
4.3.2 Diffraction Gratings 100
4.4 **Nondispersive (Energy-Dispersive) Spectrometers 103**
4.4.1 Semiconductor Devices 103
4.4.1.1 Lithium-Drifted Silicon [Si(Li)] Detectors 105
4.4.1.2 Germanium Detectors 105
4.4.1.3 Silicon Drift Detectors (SDDs) 106
4.4.1.4 Charge-Coupled Devices (CCDs) 106
4.4.1.5 Room-Temperature Devices 106
4.4.2 Cryogenic Devices 107
4.5 **Applications 109**
Acknowledgments 110
Glossary 110
References 112
Further Reading 113

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

4.1
Introduction

X-ray spectrometers are devices designed to study X-ray beams, with the goals of determining the wavelengths (or equivalently the energies) and the fluxes of the incident X-rays. Although spectroscopy of a sort was done from within a few years after Röntgen's discovery of X-radiation (1895), an enormous breakthrough was achieved when crystal diffraction revealed the first detailed X-ray spectrum (Bragg and Bragg, 1913). The Bragg technique has been refined over the years but is still used extensively and, for many purposes, gives the most useful results (see Section 4.3.1). Since that time X-ray spectroscopy has been applied, with great success, in many branches of pure and applied science (Agarwal, 1991). X-ray spectrometers have improved significantly, and new classes of such instrumentation are evolving rapidly (Tsuji, Injuk and van Grieken, 2004; Janssens, Adams and Rindby, 2000). In this chapter, the tools of X-ray spectroscopy are briefly described and the most common spectrometer devices are discussed. The strengths and limitations of the various spectrometers are indicated, and brief examples of applications and the kinds of devices that might be appropriate for real-world experiments are provided.

To some degree, all X-ray detectors are spectrometers, since all have some ability to determine wavelengths and measure fluxes. However, here we concentrate on those detectors (and spectroscopic devices) that have moderate spectral resolving powers (roughly, the resolving power $E/\Delta E$ must be better than about 20 over at least some of the spectral range). This criterion excludes, in general, proportional counters, microchannel plate devices, and scintillation detectors. Useful discussions of these classes of detectors (including their spectroscopic capabilities) can be found in the books by Fraser (1989), Knoll (1989), and Michette and Buckley (1993).

The survey here is further restricted to the energy (wavelength) range of 100 eV (124 Å) to 100 keV (0.124 Å). Lower and higher energy spectrometers are discussed in articles and books on ultraviolet and γ-ray spectroscopy. Photons with energies above the creation threshold for electron–positron pairs ($E \geq 1.02$ MeV) interact with matter by initiating a cascade shower. Measurements on these hard photons allow their total energy, arrival position, and arrival time to

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

be determined (Sauli, 1992; Ramana Murthy and Wolfendale, 1993). Although this chapter is mainly concerned with the energy range up to 30 keV, we briefly consider, at this point, the range beyond this up to 511 keV, that is, half the electron–positron creation threshold. The resolving power of high-purity germanium (HPGe) detectors (see Section 4.1.2) is typically several hundreds, which is excellent for resolving lines of the highest energy atomic transitions and the lower energy nuclear transitions. The disadvantage of germanium detectors, which progressively worsens as energies increase, is that the percentage of photons that convert via the photoelectric effect becomes smaller with higher energies (about 75% at 100 keV to about 10% at 511 keV). Consequently, the full energy efficiency becomes small.

Sophisticated X-ray spectrometers have been recently designed to study X-ray emission of new types of hot plasma flashes in order to learn about space- and time-dependent ion distributions as well as electron density and temperatures (see Section 3.1.3). Commercially available X-ray spectrometers register X-ray fluorescence of a wide range of samples being irradiated either by an X-ray source or directly by an electron beam in order to obtain their elemental composition. Well-developed X-ray spectrometers with powerful, water-cooled X-ray tubes of many Kilowatts power now compete with small spectrometers equipped by air-cooled X-ray minitubes of a few up to 50-W power and adapted focusing optics (e.g., polycapillaries).

Finally, X-ray spectrometers have been constructed to study X-ray absorption of species in experiments with bright synchrotron and hot plasma sources. The X-ray absorption fine structure provides information about the geometrical distribution of electrons in the vicinity of the absorbing atom. Some recent practical applications are outlined in Section 4.5.

4.2
General Concepts

The primary figure of merit of most spectrometers is the resolving power:

$$R = \frac{\lambda}{\Delta\lambda} = \frac{E}{\Delta E} \tag{4.1}$$

where $\Delta\lambda$ is the full width at half maximum (FWHM) (in wavelength units) of the spectrometer response to a monochromatic X-ray of wavelength λ. The energy E is related to λ by $E = hc/\lambda$, where h is Planck's constant and c is the speed of light. The relationship $E\lambda \sim 1.24$ (E in kilo electronvolt and λ in nanometer units) is often useful in X-ray spectroscopy. Note that X-ray spectroscopists also rely on the older units (angstroms and electronvolt or kilo electronvolt) for wavelength and energy. In this chapter, we typically use these as well, since the interested reader will find them in most of the standard references. To convert to SI units, note that $1\ \mathring{A} = 0.1$ nm and $1\ \text{eV} = 1.60218 \times 10^{-19}$ J.

Besides the (i) resolving power, other relevant performance parameters of X-ray spectrometers, which may, in fact, be decisive in the selection of an instrument for a particular experiment, are as follows: (ii) sensitivity (or instrumental efficiency); (iii) linearity of (or at least information on) spectral response; (iv) maximum counting rate of the detector; (v) breadth of spectral coverage; (vi) field of view; (vii) physical size; (viii) simplicity of operation; and

(ix) cost. The strengths of various spectrometer types are illustrated as they are discussed later.

X-ray spectrometers can be divided into two general categories. The first is the dispersive spectrometers that use X-ray diffraction to spread out the incident spectrum (a prism is an optical analogy). A detector can then read off the dispersed X-rays and associate a wavelength to the position of the detected photons. Dispersive spectrometers are sometimes called *wavelength-dispersive spectrometers* (*WDSs*).

Nondispersive spectrometers, on the other hand, convert the energy of an incident X-ray to a number of "particles" (electron–ion pairs, electron–hole pairs, phonons, superconductor quasiparticles, etc.), and the energy resolution of the spectrometer is determined by the statistical fluctuations of the number of particles (this statement holds only for ideal detectors; no actual device is free of systematic uncertainties). All designs of nondispersive spectrometers suffer from the problem of "dead time," the time required to process an X-ray absorption event. Note that some authors refer to nondispersive spectrometers as *energy-dispersive spectrometers* (*EDSs*), generalizing the word "dispersive" beyond simple spatial separation.

Broadly speaking, dispersive spectrometers work by taking advantage of the wave nature of the X-rays (interference produces the diffraction) and nondispersive (energy-dispersive) spectrometers work by taking advantage of the particle nature of the X-rays (photons ionize or otherwise interact with the detectors). Of course, this comment is overly simplistic, since dispersive spectrometers always use photon detectors to read out the dispersed spectrum. In fact, it is often useful to combine dispersive and nondispersive spectrometers, such as diffraction gratings and solid-state detectors (Sections 4.3.2 and 4.4.1), so as to combine the strengths of the different devices.

In Section 3, the two types of dispersive X-ray spectrometers, Bragg crystal diffractors and diffraction gratings, are discussed. In Section 4.4, the nondispersive detectors (those with reasonably high resolving powers) are considered.

4.3
Dispersive Spectrometers

4.3.1
Bragg Crystal Devices

4.3.1.1 Bragg's Law

Figure 4.1 illustrates the elementary derivation of the fundamental equation of crystal diffraction, that is, Bragg's law,

$$n\lambda = 2d\ \sin\theta \left(1 - \frac{\delta}{\sin^2\theta}\right) \tag{4.2}$$

where λ is the wavelength of the incident X-ray, d is the spacing between the planes of the diffracting atoms, often called *net planes*, θ is the angle of incidence (and reflection) of the X-ray beam onto the crystal surface, and n is a positive integer, namely, the order of diffraction. Bragg's law states that only X-rays with wavelengths that satisfy Eq. (4.2) will be efficiently reflected from the crystal. Figure 4.1 shows how Bragg's law is derived: X-rays of wavelength λ are incident on a crystal surface with a glancing angle θ_{in}. Although each scattering site scatters in all directions, only those rays that satisfy $\theta_{\text{in}} = \theta_{\text{out}}$ will constructively interfere and result in a significant reflected flux (this is,

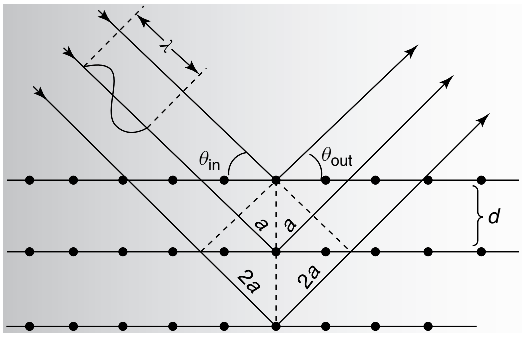

Fig. 4.1 Derivation of Bragg's law.

of course, consistent with the general law of reflection in geometrical optics). Furthermore, a second constraint on the reflected ray arises from the requirement that X-rays diffracted from each crystal plane must also interfere constructively (since the X-rays have relatively high energies, they can typically penetrate through hundreds to thousands of crystal planes). The X-rays penetrate the crystal surfaces and are scattered by the various crystal molecular planes (the planes are separated by d). The differences in the path lengths between the X-ray waves reflected off the various planes are $2a$, $4a$, $6a$, ... Clearly, there will be constructive interference between the various waves only if the path-length differences are integral multiples of the wavelength λ, that is, if $n\lambda = 2a$, where n is a positive integer. Simple trigonometry shows that $a = d \sin\theta$, thereby giving Eq. (4.2).

The crystal lattice parameter of silicon was linked to the SI meter base unit by combined optical and X-ray interferometry experiments with an experimental uncertainty error of about 10^{-8} (reviewed in Becker (2001)). For practical purposes, the lattice parameters (or net plane distances d) of any crystal can be derived at lower accuracy from a high-resolution X-ray diffractometer experiment. Here, the Bragg angles θ of the silicon gauge crystal are compared with those of the crystals being measured (Klöpfel *et al.*, 1997). In such thorough evaluations of crystal parameters, several correction factors have to be considered, for example, refraction effects on the boundary between air and crystal. This leads to a correction of Bragg's law:

$$n\lambda = 2d\ \sin\theta \left(\frac{1-\delta}{\sin^2\ \theta}\right) \tag{4.3}$$

where $1-\delta$ is the real part of the index of refraction of the crystal at the wavelength of interest. Other effects must also be considered for a rigorous treatment of the Bragg diffraction – such as

1. extinction (the primary beam cannot reach deeply into the crystal since it is reflected by intervening crystal layers);
2. absorption (again, the primary beam is absorbed by intervening crystal layers); and
3. multiple reflection (crystal layers reflect both "up" (toward the surface of the crystal) and "down" (away from the surface)).

Thorough discussions of these topics are beyond the scope of this chapter but may be found in some detail in the books by Compton and Allison (1967), Pinsker (1978), James (1982), and Authier (2001).

4.3.1.2 Flat Crystal Bragg Spectrometer

Figure 4.2 is a schematic drawing of a generic X-ray spectrometer based on the Bragg principle. Here, X-rays are collimated so that they strike the diffracting crystal at the Bragg angle for X-rays of wavelength λ. Only those X-rays that satisfy Eq. (4.2) are reflected and registered in the detector. Since (presumably) the spacing d and incident angle θ are known, the wavelength of the diffracted X-ray can be determined to be within a factor of $1/n$ (n is the diffraction order). Removing order ambiguity is an issue for all dispersive spectrometers. Techniques for determining the actual X-ray wavelength (i.e., removing the $1/n$ uncertainty) are discussed later.

To be useful as a spectrometer, it is necessary for a device to cover a range of wavelengths. This is achieved for the generic device in Figure 4.2 by mounting the crystal and the detector on coaxial rotary stages, so that the crystal can present a range of incident angles to the X-ray beam. Any rotation of the crystal by an angle $\Delta\theta$ is accompanied by a rotation of the detector by $2\Delta\theta$, so that the diffracted X-rays fall on the detector. To study a single X-ray emission line, for example, the crystal must be rotated over a range of angles that cover the response of the crystal to a monochromatic line. It should be apparent, from this discussion, that a Bragg device can only measure X-rays with wavelengths in the range $0 \leq \lambda \leq 2d$ (i.e., when $0^\circ \leq \theta \leq 90^\circ$); in practice, however, both the short- and long-wavelength limits are restricted by geometrical constraints. To measure longer wavelengths, one requires crystals (or multilayer diffractors discussed below) with larger $2d$ spacings.

From Bragg's law, the angular dispersion ($d\lambda/d\theta$) can be easily derived: after division by the reflection order n, the wavelength λ is a product of a constant factor ($2d/n$) and $\sin\theta$. If this factor is expressed by $\lambda/\sin\theta$ in the derivative, we obtain

$$\frac{d\lambda}{d\theta} = \frac{\lambda}{\tan\theta} \tag{4.4}$$

Thus, the angular dispersion is inversely proportional to $\tan\theta$, which means that the resolving power R is directly proportional to $\tan\theta$. If we have a set of

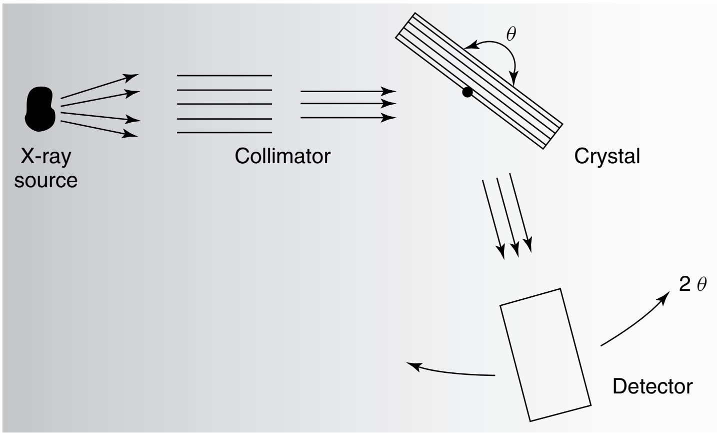

Fig. 4.2 Schematic drawing of a generic crystal spectrometer.

crystals with different net plane distances d, we should choose a plane with $2d$ slightly larger than λ to obtain the highest resolving power R.

The FWHM of the response function of an ideal crystal spectrometer is dominated by the natural resolving power of the crystal (the angular dispersion of the collimated X-rays also contributes, however, and in some cases will dominate). The response function of the crystal alone is called the *rocking curve*, because the complete response of the crystal to a monochromatic X-ray is measured by rotating (rocking) the crystal over a small angular range. The FWHM is the rocking curve width. Very early on, Compton and Allison (Becker, 2001) had derived a general theory of X-ray spectrometers. In the case of Figure 4.2, the relation between the X-ray emissivity J, the rocking curve C, and the reflected intensity P holds with some minor simplifications:

$$P(\lambda) = \int d\lambda' J(\lambda - \lambda') \int_{\alpha_{\min}}^{\alpha_{\max}} d\alpha \times \int_{\phi_{\min}}^{\phi_{\max}} d\phi \cdot C\left(\alpha - \frac{\phi^2}{2} \tan\theta_0 - \frac{\lambda' - \lambda_0}{\lambda_0} \tan\theta_0\right) \tag{4.5}$$

where the divergence angles α and ϕ describe the angular source emission in and out of the reflection plane (Figure 4.2). To minimize the broadening of X-ray emissivity J by convolution with the rocking curve C, the FWHM of C should be much smaller than the divergence $(\alpha_{\max} - \alpha_{\min})$ in order not to lose X-ray photons in the Bragg reflection. The meridional divergence (multiplied by the angular dispersion of Eq. 4.4) should, in addition, be much smaller than the FWHM of X-ray emissivity J. This is normally realized in WDSs, if the resolving power is high and the collimator is well adapted (cf. Figure 4.2). For small sagittal divergences $(\phi_{\max} - \phi_{\min})$, this broadening influence in Eq. (4.5) can be neglected since it depends on the second order of ϕ.

If the absorption in the spectrometer crystal is very small (which holds for hard X-rays), the rocking curve C has the form of a cylinder head (Darwin curve) with a flat top of almost unity reflection (i.e., in the range of interference total reflection). For the general case of structurally perfect crystals, the rocking curves can be computed using the dynamical theory of X-ray interferences (see the book by Authier (2001) and references therein). A narrow FWHM of the rocking curves leads not only to a good spectral resolution but also eventually to a loss of X-ray photons on the detector, whereas a wide FWHM provides many photons at the cost of spectral resolution.

Finally, the sensitivity of the Bragg devices must be known if one wishes to determine the incident flux as a function of wavelength (this is not always the case; sometimes, experimenters may be primarily interested in the locations (i.e., wavelengths) of spectral features or in the relative intensities (as opposed to the absolute fluxes)). To determine the flux in an X-ray line, the crystal must be rocked over the nominal Bragg angle of the line, covering a $\Delta\theta$ range that includes the response of the crystal (most of the width of the rocking curve), as well as contributions due to the dispersion of the incident beam. With the knowledge of the crystal reflectivity as a function of the angle (and other parameters such as the detector quantum efficiency), the incident flux in the X-ray line can be computed. The parameters of interest for computing the flux are not always easy to determine, however, particularly at all energies.

As noted above, Bragg's law permits a residual uncertainty in the X-ray wavelength because of the reflectivity of various diffraction orders of *n*. (This uncertainty is also present for grating spectrometers.) This ambiguity may be minimized or eliminated in a number of ways. The simplest way of eliminating it is to use a detector that has a sufficient inherent spectral resolution that the orders may be distinguished by the detector. Even low-resolution devices such as proportional counters can separate orders sufficiently in a wide energy range. Other techniques are to use filters and/or X-ray optics (which do not reflect efficiently at higher energies) to minimize the contribution from higher orders.

Collection and diffraction efficiency of a flat crystal spectrometer can be improved if parallel beam X-ray optics is incorporated. A diverging beam originating from the X-ray source is parallelized by polycapillary and grazing-incidence optics. The now parallel beam impinges on the spectrometer crystal (Love, 2002). This results in a much larger solid angle of X-rays that can be collected. The wavelength-dispersive unit can, for example, be incorporated in a scanning electron microscope.

4.3.1.3 **Bent-Crystal Spectrometer**

Besides the industrial applications of X-ray spectrometers where a high X-ray flux is often available, there are tasks in which the detector flux is rather limited. This is the case for X-ray diagnostics of high-temperature plasmas and for electron-beam X-ray microanalysis (see Section 4.3.1.4). There is, therefore, a need for additional X-ray optic elements to collect and focus X-rays to the detector. Depending on the required resolving power and luminosity as well as the geometry of spectrometer, different X-ray optics, grazing-incidence mirrors, capillaries, bent gratings, bent crystals, and so on must be selected.

Bent-crystal spectrometers employ, in the simplest form, a cylindrically bent crystal (instead of a flat one) as the dispersing element and a detector that is positioned on the circumference of the Rowland circle (see Figure 4.3(a) and (b)). The curvature of the reflecting net planes is twice that of the radius of the Rowland circle in this (Johann-type) spectrometer. Equation (4.5) is still applicable when the rocking curve is corrected for crystal bending (Uschmann, Malgrange and Förster, 1997) and modified relations are used between the divergence angles α, ϕ, and the wavelength λ'. Since rocking curves are usually wider and bent crystals accept a diverging beam of X-rays, photon numbers on the detector can be increased by 1 or 2 orders of magnitude in comparison to the flat crystal spectrometer. Figure 4.3(c) shows the dominant resonance lines of hydrogen-like and helium-like aluminum ions together with intercombination and satellite lines on a relatively low continuum emission. The electron density and temperature can be obtained from plasma physics models from ratios of selected spectral lines (Attwood, 1999).

The highly excited aluminum ions were produced by focusing high-power Nd : glass laser pulses (200 J, 1 nanosecond) on solid aluminum target (Renner *et al.*, 1994).

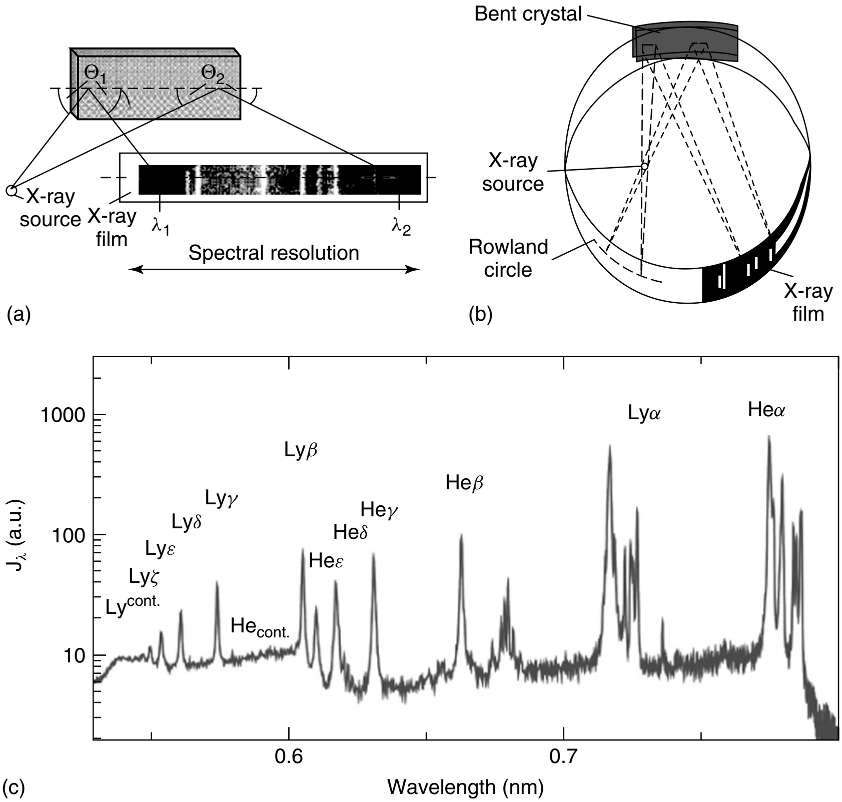
Θ1
Θ2
X-ray source
X-ray film
λ1
λ2
Spectral resolution
(a)
Bent crystal
X-ray source
Rowland circle
X-ray film
(b)
1000
100
10
Jλ (a.u.)
Lyα
Heα
Lyβ
Heβ
Lyγ
Heγ
Lyδ
Heδ
Lyε
Heε
Lyζ
Ly^cont.
He_cont.
0.6
0.7
(c)
Wavelength (nm)

Fig. 4.3 Schematic drawings of (a) a flat crystal spectrometer, (b) a Johann-type bent-crystal spectrometer, and (c) X-ray spectrum of a laser-produced Al plasma.

If the cylindrical crystal is rotated by 90° around its central surface normal, then X-rays are collected in the ϕ angle (instead of the α angle) direction. Figure 4.4(a) shows the von Hamos-type spectrometer in which the distances from the crystal to the source and the detector are equal. By bending a thin crystalline wafer in two dimensions, spherical, toroidal, or ellipsoidal crystals can be produced with curvature radii down to 10 cm. As shown in Figure 4.4(b) and (c), luminous X-ray spectrometry and quasimonochromatic X-ray imaging can be performed. Mißalla *et al.* (1999) have discussed reflection and focusing properties of these schemes when short X-ray pulses in the kilo electronvolts range are considered.

A cylindrically bent analyzer crystal can be combined with a flat position-sensitive detector in the von Hamos geometry (see Section 4.4.1.4). This makes parallel data collection possible for light elements in WDSs (Love, 2002). Toroidal crystals combine focusing in both sagittal and meridional directions if the ratio of the respective radii is $\sin^2\theta$. They are characterized by high resolving power and efficiency but require much careful preparation and testing.

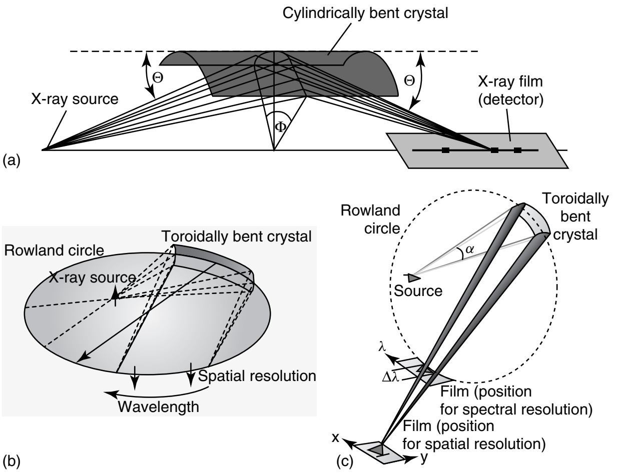
Cylindrically bent crystal
Θ
Θ
X-ray source
X-ray film
(detector)
Φ
(a)
Rowland circle
Toroidally bent crystal
X-ray source
Spatial resolution
Wavelength
(b)
Rowland
circle
Toroidally
bent
crystal
α
Source
λ
Δλ
Film (position
for spectral resolution)
Film (position
for spatial resolution)
x
y
(c)

Fig. 4.4 Schematic drawings of bent-crystal spectrometers: (a) von Hamos-type, (b) toroidally bent crystal spectrometer, and (c) quasimonochromatic X-ray imaging setting.

An even higher spectrometer efficiency can be achieved by applying ellipsoidal highly oriented pyrolithic graphite (HOPG) crystals if a wider reflection curve (a 10th of a degree) is tolerated.

All the spectrometer types discussed so far are based on the Bragg equation and its corresponding angular dispersion ($d\lambda/d\theta = \lambda/\tan\theta$). There is, however, a special variant of a Johann spectrometer in which a point source and film are both located on the Rowland circle (Figure 4.5(b)).

Assuming that the X-ray wavelength λ_0 is reflected by the cylindrical crystal in the Rowland circle plane, X-rays with nonzero sagittal angles ϕ impinge on the crystal under slightly smaller Bragg angles. Geometrical considerations lead to the following relation between

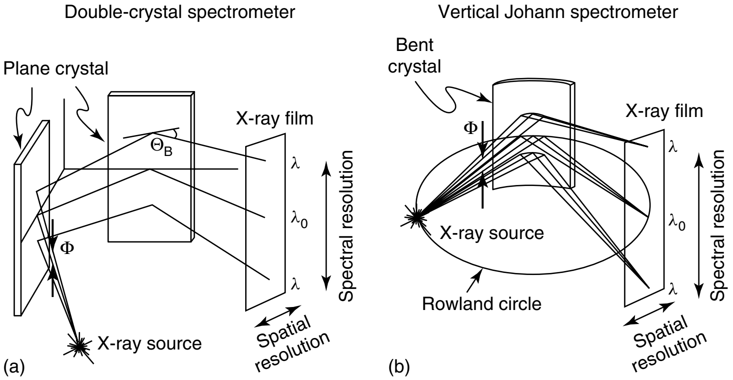
Double-crystal spectrometer
Plane crystal
Θ_B
X-ray film
λ
λ_0
λ
Spectral resolution
Φ
X-ray source
Spatial resolution
(a)
Vertical Johann spectrometer
Bent
crystal
Φ
X-ray film
λ
λ_0
λ
Spectral resolution
X-ray source
Rowland circle
Spatial resolution
(b)

Fig. 4.5 Schematic drawings of high-resolution X-ray spectrometer with vertical (sagittal) dispersion: (a) double flat crystal and (b) Johann-type cylindrically bent-crystal setting.

Tab. 4.1 Key parameters for different X-ray crystal spectrometers.[a]

Type of spectrometer	Luminosity (mrad^2)	Spectral resolution	Range (Å)
Flat single crystal	1.74×10^{-2}	4000	1.14
Johann classical scheme	1.88×10^{-1}	12 800	1.34
Double-crystal vertical dispersion	4.36×10^{-3}	13 000	0.18
Johann vertical dispersion	2.72	10 400	0.18

[a] Quartz crystals of dimension 20 × 30 mm^2 either flat or with curvature radius of 76.8 mm were used. Source-to-detector distance was 130 mm; source of size 1 mm^2 emits a certain number of photons per square milliradian per square millimeter into a solid angle 1 mrad^2 (Renner *et al.*, 1994).

wavelength λ and the divergence angle ϕ:

$$\lambda = \lambda_0 \cos\phi \tag{4.6}$$

where the maximum wavelength λ_0 can be simply changed by moving the X-ray source and the film along the Rowland circle. This leads to a high luminosity and a high spectral resolving power (Renner *et al.*, 1994), which is needed for sophisticated X-ray plasma diagnostics, but is limited to a small spectral coverage. There is also a simple extension of the flat crystal spectrometer to a double-crystal spectrometer in which wavelengths are determined by Eq. (4.6). It suffers, however, from a very low luminosity and is therefore applicable only for strong X-ray sources. Table 4.1 summarizes some key parameters of the X-ray spectrometers of Figures 4.3(a) and (b) and 4.5(a) and (b), respectively.

4.3.1.4 Bragg Spectroscopy with Synthetic Crystals (Multilayers)

Most Bragg diffractors are either natural crystals or crystals grown in the laboratory according to certain chemical and thermal specifications. Crystals of this kind have $2d$ spacings in the range 0.2–2.7 nm and can be used to cover energies between ~500 eV and 35 keV (Attwood, 1999). For lower energies, synthetic multilayer diffractors sometimes called *layered synthetic microstructures* (*LSMs*) or simply multilayers are frequently used, replacing in most cases, the Langmuir–Blodgett soap films used for many years. Multilayer diffractors are made by depositing layers (using the sputtering or evaporation techniques of semiconductor technology) of alternating high-Z and low-Z materials (e.g., carbon and tungsten) with precise thicknesses so as to perform the same functions as one-dimensional crystals. In the multilayer case, the effective spacing d is the sum of the thicknesses of the high-Z and low-Z materials. Up to several hundreds, layers can be laid down in this way (more layers are of little value since the relatively low-energy X-rays cannot penetrate deeper). Multilayers can now be deposited on curved substrates (up to 30-cm lateral size), even with gradients in depth and lateral thickness in order to optimize imaging performance. A comparison of focusing optics for femtosecond X-ray diffraction with Cu Kα line radiation shows differences in the key parameters, focus size, flux, and photons per second (Bargheer *et al.*, 2005). The authors considered a toroidally bent Ge Crystal, multilayer optics, ellipsoidal monocapillary, and polycapillary X-ray optics.

A clear advantage of using multilayer films is that the spacing *d* can be selected to be any desired value (although *d* spacings smaller than about 25 Å are difficult to fabricate). Generally, the resolving power of multilayer diffractors is limited to about 50, which is superior to many other kinds of spectrometers but will not be adequate for some applications. As shown in Figure 4.6, effective methods are available to change the form of the multilayer reflection curve. By using higher reflection orders *n*, the FWHM of the reflection curve can be reduced to obtain higher resolution. In contrast, the FWHM can be broadened by using depth gradients in layer thicknesses to obtain higher spectrometer sensitivity on the expense of peak reflectivity.

In recent years, multilayers have been used as the reflecting surfaces of many different optical devices, providing enhanced reflection at certain soft X-ray wavelengths (i.e., those wavelengths in which the Bragg condition is met). For example, traditional X-ray telescopes utilize grazing incidence ($\theta <$ a few degrees) on polished metallic surfaces (Fraser, 1989). Such telescopes are difficult to fabricate and typically have a relatively small collecting area. More traditional telescope geometries (i.e., those with nearly normal incidence) can be used, however, if the reflecting surfaces are coated with multilayers. The resulting telescope forms sharp images over a narrow wavelength band. Figure 4.7 shows an image of the sun obtained with an X-ray telescope coated with layers of cobalt and carbon with a period of 31.8 Å. At nearly normal incidence, the mirrors reflect efficiently at a wavelength of 63.5 Å. The bandpass of 1.4 Å encompasses the wavelengths of two bright X-ray emission lines from ions of highly ionized Mg and Fe, which are sensitive diagnostics of the plasma temperature in the range $(1-3)\times10^6$ K, in which these ions are abundant. The figure shows the high-temperature material in the solar corona, in contrast to the much cooler photosphere that appears relatively dark

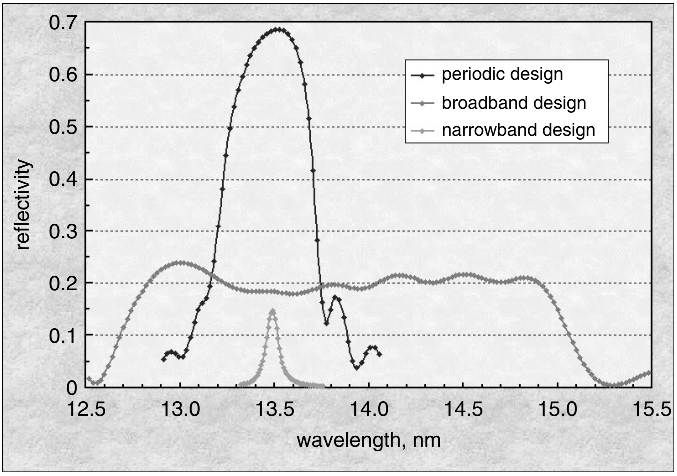

Fig. 4.6 Change in Mo/Si multilayer reflection curves to obtain higher resolution and higher sensitivity, respectively. (courtesy of Torsten Feigl and Sergij Yulin.)

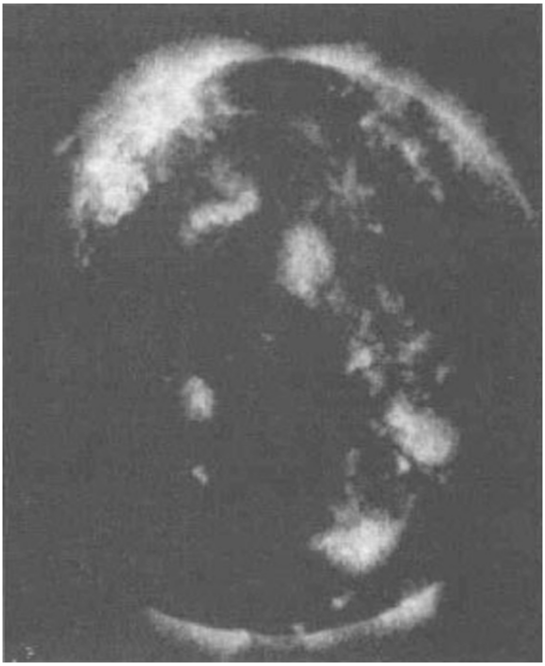

Fig. 4.7 Image of the sun taken on 11 July 1991 at $\lambda = 63.5$ Å with a bandwidth of 1.4 Å. The hot solar corona is seen clearly, while the cooler surface of the sun is relatively dark at this wavelength. (courtesy of Leon Golub.)

at this wavelength (Golub, Zirin and Wang, 1994).

4.3.2
Diffraction Gratings

Diffraction gratings are similar to crystals in that a periodic structure sets up an interference pattern that disperses the X-rays according to wavelength. They differ from crystals in that the structure is periodic only on the surface of the diffractor instead of both on the surface and within the body of the diffractor, as for crystals. As a result, diffraction gratings reflect (or transmit) incident X-rays at all wavelengths regardless of the incidence angle.

Gratings disperse light according to the grating equation:

$$n\lambda = p(\cos\beta - \cos\alpha) \tag{4.7}$$

where the incident and diffracted angles (α and β) are defined in Figure 4.8 for the two types of gratings: reflection and transmission gratings. As was the case for crystal spectrometers, n is the integer order of diffraction (except that for gratings, n can be positive, negative, or zero). For gratings, p is the period of the grating lines or grooves. X-ray reflection gratings operate at small incidence angles ($\alpha < 2°-3°$) because surfaces reflect X-rays efficiently only at grazing angles (unless enhanced with multilayer coats; see Section 3.1.4 and (Spiller, 1994)). Transmission gratings generally operate at nearly normal incidence ($\alpha \sim 90°$) where the transmitted efficiency is greatest, although they can be used with moderate efficiency over a large range of angles.

Ordinary ruled reflection gratings were among the first X-ray spectroscopic

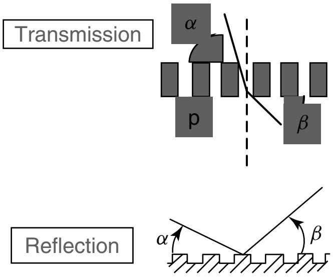

Fig. 4.8 Schematic illustration of the two kinds of X-ray gratings, the transmission and the reflection gratings. Although not indicated in the diagram, both kinds of gratings can be blazed to enhance the efficiency in some of the orders.

instruments, although they have been limited, until recent years to the longer wavelengths ($\lambda > \sim 20$ Å) since it is difficult to rule gratings with sufficiently small periods. In recent years, maturing technology (essentially, replacement of the mechanical ruling method by lithography or holography) has led to reflection gratings that can be used at wavelengths as short as ~5 Å. As is the case for reflection gratings used at optical wavelengths, X-ray gratings can be blazed (i.e., can have asymmetric reflecting surfaces) so as to enhance reflections into particular orders.

X-ray transmission gratings are a relatively recent development. Originally conceived for applications in X-ray astronomy (Gursky and Zehnpfennig, 1966), they have also been applied as spectrometers in synchrotrons and in X-ray lasers, and as interferometers in atomic diffraction experiments. Advantages of transmission gratings are their light weight and that they are typically used at normal incidence so that the spectrometer can be more compact than for a reflection grating. Effective transmission gratings must be thick enough so that the grating bars are at least partially opaque to the X-rays and must have small enough periods so that the dispersion angle is large enough to be useful. Figure 4.9 shows a sketch of a transmission grating developed for X-ray astronomy.

Transmission gratings are made with the techniques of semiconductor technology (microfabrication and nanofabrication). Periodic patterns are generated from holographic interference techniques or from mechanical ruling for larger period gratings. The periodic pattern is transferred to thicker metallic grating bars (gold is most often used). The first-generation transmission gratings were used primarily at lower energies ($E < 3$ keV) because the grating–bar thicknesses were too small (gratings became transparent at higher energies) and/or the bar periods were too large (dispersion angles were small). Recent developments have extended the effective energy range of transmission gratings to ~10 keV (Schattenburg *et al.*, 1991).

Transmission gratings are usually fairly small (a few square centimeters, Figure 4.9), but individual gratings can be combined into a larger assembly. For X-ray astronomy applications, for example, grating areas of over 1000 cm^2 have been used at the objective of X-ray telescopes to provide a large collecting area for spectroscopy of cosmic sources. For laboratory applications, free-standing gold bars with up to 5000 lines mm^{-1} have been produced where the gold bars were fixed on a periodic gold support

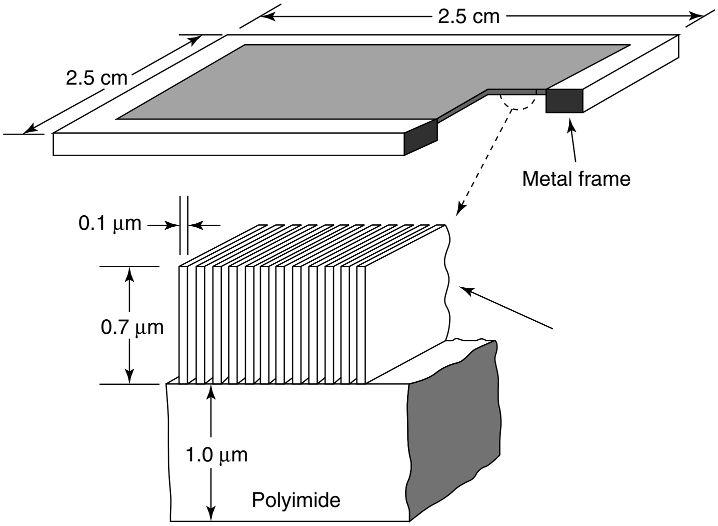

Fig. 4.9 Schematic of a high-energy grating facet being built for the US observatory Advanced X-ray Astrophysics Facility (AXAF). The gold grating bars (period = 2000 Å) are supported by a thin (1-µm) plastic film.

grid with 17 µm × 17 µm active areas. In combination with a toroidal grazing-incidence mirror, these transmission gratings were used in X-ray spectroscopy of both laser-produced plasmas and scattered radiation from a soft X-ray free electron laser (Jasny *et al.*, 1994; Höll *et al.*, 2007).

Like Bragg spectrometers, grating spectrometers can, in theory, achieve extremely large resolving powers. The theoretical maximum resolution for diffraction order n is given as

$$R = \frac{E}{\Delta E} = nN \qquad (4.8)$$

where N is the number of grating grooves or lines illuminated by the X-ray beam. For a grating with 1000 lines mm^{-1} (routine for transmission gratings), a 5-mm-wide X-ray beam can achieve a resolving power of 5000. It is, however, difficult to fabricate perfect periodic gratings, so that the variation in the grating period usually limits the resolving power to $p/\Delta p$, where Δp is the variation in the grating period over the illuminated surface of the grating.

Note that both classes of dispersive spectrometers (crystals and gratings) can, with relatively minor modifications, be used as monochromators as well as spectrometers, that is, they can select a narrow energy band out of a broad spectrum and transfer this band for further processing and analysis. Both grating and crystal monochromators are employed (generally, the crystals at the shorter wavelengths, where the gratings are less effective) in this way.

As was noted in the discussion of crystal spectrometers, there are potential benefits to curving the diffractors. In such a design, the incident X-rays can diverge from a point (parallel or nearly parallel beams have been tacitly assumed for the flat diffractors discussed above), and the curved gratings perform both energy dispersion and focusing. In the Rowland configuration (which is the most commonly used for curved diffractors), a point source at the entrance slit will focus to a (slightly curved) line at the exit slit, each wavelength focusing to a different position as shown in Figure 4.4(b) and (c). More complex curvatures (spheres, ellipsoids, and toroids) can improve the focusing properties at some wavelengths, although it can be more difficult to fabricate the curved diffractor elements (Michette and Buckley, 1993).

4.4
Nondispersive (Energy-Dispersive) Spectrometers

4.4.1
Semiconductor Devices

Semiconductor-based detectors, sometimes simply called *solid-state detectors,* have been used as X-ray spectrometers for more than 30 years. A basic description of how such devices work can be found in Tsuji, Injuk and van Gieken (2004); Fraser (1989); Knoll (1989). The fundamental principle is that an X-ray will liberate electron–hole pairs as it interacts within the semiconductor. The number of such pairs is given on the average by $N = E/w$, where E is the X-ray energy and w the ionization energy. For silicon and germanium, the ionization energies w are approximately 3.62 and 2.96 eV, respectively, and are slightly temperature dependent. An X-ray photon with energy 1 keV, therefore, will liberate on average several hundred electron–hole pairs. This free charge can then be gathered, as in a gas ionization detector, by applying an electric field to the ionized region. The amount of charge gathered is proportional to the energy of the incident X-ray.

Figure 4.10 shows a schematic diagram of a generic semiconductor detector. Two thin surfaces are heavily doped so as to be n-type (where electrons are the majority charge carriers) and p-type (where holes are the charge carriers). A central region, considerably thicker than the n- and p-types, is relatively free of charge carriers and is called the *depletion region*. The key feature of the depletion region is that it has a nonzero electric field, so that electrons and holes that are produced by photon interactions are easily separated. For X-ray detectors, the depletion region can consist of either an extremely pure semiconductor (typically germanium or silicon) or a volume of a semiconductor into which an element of opposite polarity (n- or p-type) has been diffused to compensate for the excess of electrons or holes (lithium drifted into p-type silicon or germanium is the most common). This compensated region (often called the *intrinsic region* since the layer has physical properties similar to pure, intrinsic semiconductors) is where the X-ray interaction takes place. To be effective as an X-ray detector over a broad energy range, the depletion region (whether naturally free of charge carriers or compensated) should be adequately thick (so that high-energy X-rays will be stopped), and any dead or absorbing layers (the n- or p-type regions or other nondetecting layers) should be thin.

When an X-ray enters the depletion layer (assuming that the layer is thick enough), it interacts with the semiconductor to produce electron–hole pairs. A voltage is applied to the surfaces to attract the electrons and holes, which migrate rapidly to the electrodes.

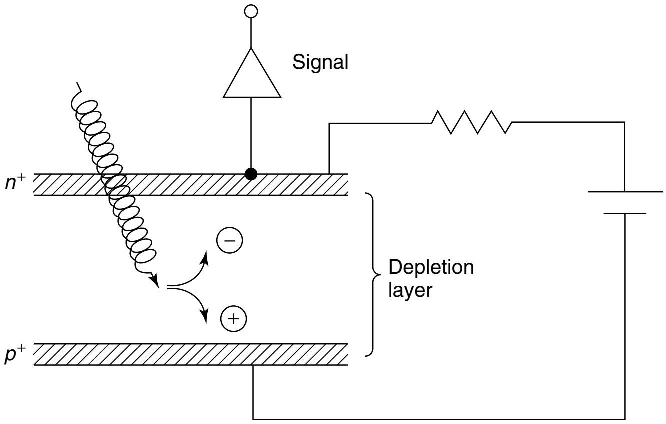

Fig. 4.10 Drawing of a (very) generic semiconductor spectrometer detector.

The heavily doped n- and p-type regions serve as barriers to any leakage current (the current that would flow in an undoped semiconductor as a result of the finite resistivity of the material). The leakage current must be suppressed for any prospect of detecting the small signal arising from an X-ray interaction. The signal goes to a preamplifier as shown in Figure 4.10 and is then processed electronically.

The collected charge following an X-ray photoelectric interaction is $N = E/w$ electrons (and holes). The deviation in the collected charge is related to X-ray energy according to

$$\Delta E = 2.354\, w \left(\frac{FE}{w} + K^2 + D^2 \right)^{0.5} \quad (4.9)$$

where ΔE is the FWHM response to a monochromatic X-ray line (see Eq. (4.1)), w is the ionization energy, E is the energy of the incident X-ray, K arises from the system electronic noise, and D is the dark current (the flow of electrons and holes that is present, owing to various impurities and imperfections in the device, even when there are no X-rays incident on the detector). Both K and D are measured in equivalent free electrons. F is the so-called Fano factor (Fano, 1947), which relates the statistical fluctuations in the number of electrons liberated to the expected number (E/w). (Fano factors for semiconductors have values of about 0.1, although there is an uncertainty in F of at least 50% due to the difficulties in precisely measuring this parameter.) Most semiconductor devices are operated at low temperatures (liquid nitrogen is typically used in the laboratory with $T = 77$ K), which helps to limit the dark current. The resolving power $E/\Delta E$ for a lithium-drifted silicon [Si(Li)] detector with system noise $\sqrt{K^2 + D^2}$ of 10 electrons is about 20 for 2 keV and grows to about 60 for 10 keV. At higher energies, the resolution is limited by the statistical fluctuations in the electron number; at low energies, the system noise dominates. Figure 4.11 shows a spectrum obtained with an Si(Li) device (see below), indicating the usefulness of a semiconductor device. The spectrum is the result of X-ray fluorescence of a sample consisting of thin films of gold,

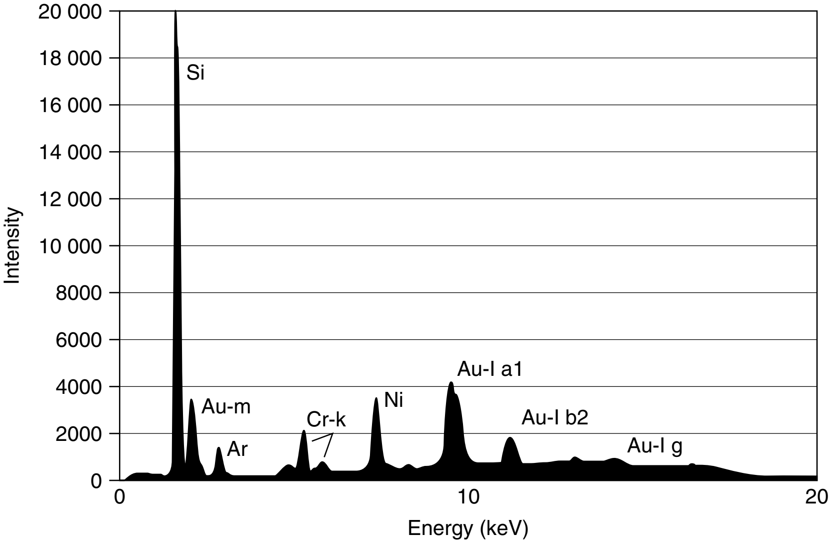

Fig. 4.11 Spectrum obtained with an Si(Li) detector.

nickel, and chromium deposited on a silicon substrate. The intensities of the various fluorescent lines can be used to determine the thicknesses of the films. Note that the resolving power of the semiconductor device is adequate for the quantitative analysis required.

There are a number of semiconductor devices that are used for X-ray spectroscopy. Here, four general types are listed, giving some of the strengths and weaknesses of the various classes. All of these devices are available commercially. See Knoll (1989) and Fraser (1989) for good introductory discussions on these devices and more detailed references.

4.4.1.1 Lithium-Drifted Silicon [Si(Li)] Detectors

Si(Li) detectors have depletion layers of thickness of a few millimeters to ~1 cm, manufactured by introducing ("drifting") a lithium dopant into slightly p-type silicon. The net result of this drifting process (which can take several weeks) is to produce a compensated depletion region that is essentially free of charge carriers. Si(Li) detectors can have significant efficiencies for energies up to about 30 keV. At lower energies, the dead layers limit the performance to energies greater than about 0.1 keV. For many commercial detectors, furthermore, a vacuum-tight window of 1-μm polymer with a cover of 30-nm aluminum is now used.

4.4.1.2 Germanium Detectors

Lithium-drifted germanium detectors [Ge(Li)] were used for many years, particularly at higher energies. In more recent years, extremely pure germanium crystals have been grown (HPGe) that have relatively few free charges and hence can be used directly as a depletion layer without the introduction of a compensating dopant. HPGe detectors complement Si(Li) detectors by having efficiencies at significantly higher energies (up to tens of mega electronvolts; see Section 4.5), but are usually not preferred at lower energies ($E < 20$ keV) since the Si(Li) detectors have better resolution and less prominent escape peaks.

4.4.1.3 Silicon Drift Detectors (SDDs)

A drift chamber detector is a large semiconductor wafer of, for example, high-resistivity n-type silicon; it is fully depleted from a small n^+ ohmic contact, which is positively biased with respect to the p^+ contacts covering both wafer surfaces. By implanting a parallel p^+ strip pattern and applying an electrical field on both wafer sides, electrons are forced to move to the central n^+ readout anode (Strüder *et al.*, 1998). The silicon drift detector (SDD) incorporates an on-chip amplifier and has a very low capacitive loading; energy resolution is 127 eV for 6 kcal at $-20°$. A particular advantage is that commercially available SDDs (Röntec) can be employed for high X-ray counting rates in the range of 400 000 counts per second. The idea for this new detector scheme originated from Gatti and Rehak (1984); the SDDs are fabricated for a wide field of applications, for example, for several synchrotron radiation experiments (Strüder *et al.*, 1998).

4.4.1.4 Charge-Coupled Devices (CCDs)

Charge-coupled devices (CCDs) are silicon-based detectors used extensively in television technology but are also effective (with some modifications) in the X-ray regime. While having spectral resolving powers similar (or even superior) to the best Si(Li) and HPGe devices, they are also excellent imaging detectors, with pixel sizes typically <20 μm and array dimensions 1024 × 1024 pixels or greater. The most common CCD design is a two-dimensional array of Metal-Oxide-Semiconductor (MOS) capacitors on the surface of a semiconductor. (An alternative, first proposed by Gatti and Rehak (1984), is to use an array of pn junctions on the surface of n-type silicon.) The charge from an X-ray interaction in the depletion layer of the semiconductor is stored, momentarily, in the nearest capacitor. By varying the surface voltages, the charge can then be passed onto adjacent capacitors until it reaches the edge of the detector, at which point it can be read out (in practice, the charge packet is read into a perpendicular register once it reaches the edge of the array, and is further transferred in this perpendicular direction; in this way, only a single amplifier circuit is required to read out an entire CCD chip). This charge transfer, or coupling, gives the device its name.

The depletion regions of CCDs are significantly shallower than that of the other semiconductor devices discussed here (of order of tens of microns), primarily because of the lower voltages required for the charge transfer. Consequently, currently available X-ray-sensitive CCDs are used most generally at energies $E < 10$ keV since higher energy X-rays simply pass through the device. The capacitor structure on the front surface, furthermore, acts as a dead layer that can absorb photons with $E < 0.5$ keV.

4.4.1.5 Room-Temperature Devices

The solid-state devices described above operate most efficiently when cooled to temperatures <77 K (liquid-nitrogen temperatures). A class of detectors composed of semiconductor compounds (HgI_2, CdS, CdTe, and GaAs) is actually more effective at room temperatures than when cooled significantly. These materials have larger band gaps (than Si or Ge), so that thermal fluctuations result in a much lower leakage current. Furthermore, because all of these compounds have a much higher effective atomic number, a viable detector can be much thinner (have a thinner depletion layer) and still detect higher energy photons. Unfortunately, the room-temperature devices are difficult to work with, have poor mobility (particularly for holes), and have a generally inferior resolving power. In spite of some clear benefits, the various room-temperature detectors are not as commonly used as the silicon- and germanium-based devices.

4.4.2
Cryogenic Devices

A relatively recent advance in X-ray spectroscopy is the development of what are often called *cryogenic detectors*. These devices convert incident X-rays into thermal-acoustic quanta (phonons) or superconductor quasiparticles (broken Cooper pairs). The energy w required to create such a particle is quite small (10^{-5}–10^{-3} eV) so that extremely high resolving powers are possible in principle (the number of particles is large and thus the relative uncertainty in their number is small). To achieve the optimal resolution of such devices (theoretically ΔE can be as small as $\sim$1 eV), various noise sources must be suppressed by operating at extremely low temperatures ($T < 100$ mK typically) and, hence, the generic name for such detectors – *cryogenic devices*. Although cryogenic detectors are not yet available commercially, they hold significant promise for future applications.

Two basic classes of cryogenic detectors have been developed in recent years: calorimeters and quasiparticle detectors. The calorimeters measure the temperature in a low-temperature substrate due to the absorption of a single X-ray photon (hence, they are phonon detectors). The quasiparticle detectors most often discussed for X-ray spectroscopy are superconducting tunnel junctions (STJs) for which the mean particle energies are 0.001 eV. The quasiparticles are detected by their quantum-mechanical current tunneling through a thin oxide layer. Although the STJ devices are promising, they have not achieved, to date, the resolving power of the calorimeters and thus are not discussed further (see (Labov and Young, 1993) for some recent articles on this topic).

The microcalorimeter has undergone extensive development over the past decade. Resolutions as high as $\Delta E \sim 2$ eV at 1.5 keV (Newbury *et al.*, 2002) have been achieved. Thus, the microcalorimeter has a resolving power that is 10–50 times better than that of the best semiconductor devices. A schematic diagram of a generic calorimeter is shown in Figure 4.12. Basically, a calorimeter absorbs an incoming X-ray and converts the X-ray energy to heat (phonons), which raises the temperature of a cryogenically cooled substrate. The increase in temperature is measured by a thermometer connected to the absorber. After a period of time, the equilibrium temperature is restored through the contact of the substrate with a cold bath. The increase in temperature is proportional to the energy of the X-ray.

The energy resolution of an ideal X-ray calorimeter is given as (Moseley, Mather

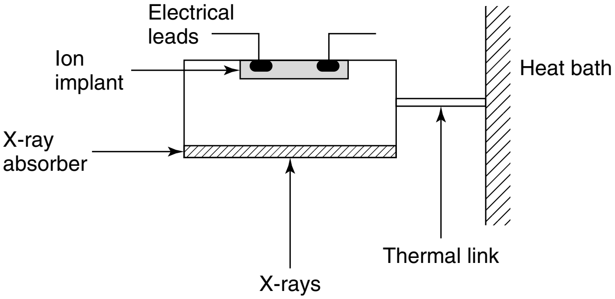

Fig. 4.12 Schematic diagram indicating the operation of an X-ray calorimeter. (Holt, 1987.)

and McCammon, 1984)

$$\Delta E = 2.35\xi\sqrt{kT^2C} \tag{4.10}$$

where T is the equilibrium temperature, C is the heat capacity of the device, k is Boltzmann's constant, and ξ is a nearly constant parameter with magnitude ~1 to 3. The parameter ξ is derived from the temperature variations of the heat capacity and thermal conductivity of the device, and the properties of the thermometer (Kelley *et al.*, 1988). For small detectors ($0.5 \times 0.5 \times 0.1\ \text{mm}^3$ are typical dimensions), cooled to ~0.1 K, Eq. (4.10) yields $\Delta E \sim 1$ eV, but, in practice, additional noise terms in the thermometer limit the current actual performance of calorimeters to ~7 eV. Note that the size of the detector is an important issue since ΔE scales as $C^{1/2}$. Also note that heat capacity often scales as T^3, so that the operating temperature is also of great importance; the best calorimeters built to date operate at about 0.1 K.

The obvious advantages of an X-ray calorimeter are as follows:

1. its grasp of a wide energy range (since it is nondispersive) and
2. its extremely high resolving power.

The resolution ΔE is essentially a constant; thus, the resolving power ($E/\Delta E$) varies as E varies. Figure 4.13 shows the resolution E_{res} as a function of the X-ray energy E_ν for several types of X-ray spectrometers. If digital processing of microcalorimeter signals is performed, energy resolutions comparable to those of the Bragg crystal devices in some spectral ranges (Newbury *et al.*, 2002) can be obtained.

The practical disadvantages of the X-ray calorimeter are as follows:

1. its small size (although mosaic detectors have been designed) and
2. the need to have a complex cryogenic system for the detector to function optimally.

The latter is a particular problem for space applications in which weight and risk are always a matter of concern. For typical designs, a large Dewar flask of liquid helium and an additional refrigeration stage are required.

X-ray microcalorimeters may not be the optimal choice of spectrometer in an environment in which an extremely high counting rate is expected, because of the relatively long time required to process each event (several milliseconds).

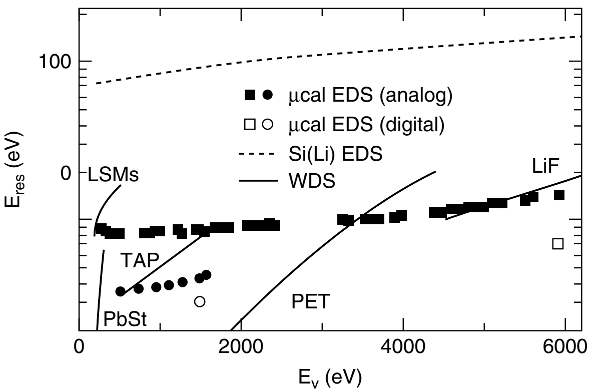

Fig. 4.13 Energy resolution (FWHM) of different types of X-ray spectrometers: microcalorimeter (energy-dispersive spectrometer (µcal EDS)) with analog and digital data processing, lithium-drifted silicon detector (Si(Li) EDS), and wavelength-dispersive spectrometer (WDS). Dispersive elements for WDS are LSMs, thallium acid phthalate (TAP), pentaerythrol (PET), lithium fluoride (LiF), and PbSt (lead stearate). (Newbury *et al.*, 2002.)

4.5
Applications

X-ray spectrometers are used in virtually every area of pure and applied science. Extensive discussion of specific applications can be found in the references following this chapter. In a general sense, however, X-ray spectrometers are employed

- to measure the properties (such as the temperatures, ionization states, and elemental compositions) of hot (>10^6 K) plasmas, both cosmic and laboratory (Silver and Kahn, 1993), where in extreme conditions time evolution and spatial distribution of laboratory plasma parameters have been obtained (Golovkin *et al.*, 2002);
- to determine compositions of various materials (qualitative and quantitative analysis via the monitoring of characteristic lines from X-ray fluorescence or direct bombardment with electrons or other charged particles; see, e.g., (Jenkins, 1988)); and
- to measure and interpret the absorption spectra of various solids at high resolution in order to determine the distribution of electrons. The study of X-ray absorption fine structure (XAFS) has blossomed in recent years with the advent of bright, monochromatic X-ray beams available from synchrotrons (Koningsberger and Prins, 1988). Ultrafast time-resolved X-ray absorption spectroscopy is a nascent technique that permits study of photoinduced processes on many systems down to the subfemtosecond time domain (Bressler and Chergui, 2004).

Depending on the application, some of the various spectrometers discussed earlier may be more appropriate than others. For example, semiconductor devices, with their moderate resolving powers and relatively broadband energy response, are often used for X-ray fluorescence analysis (see Figure 4.11), since the fine structure of the spectral features is not generally an issue. When extremely high resolution is required, at least at the lower energies, crystal spectrometers are still the instruments of choice (the spectrum in Figure 4.3, for example, would be extremely difficult to unravel with a semiconductor spectrometer). When low flux (or short exposure times) is an issue and/or a significant energy band is desired, the narrow bandpass of the crystal spectrometers argues in favor of a nondispersive device (or a grating spectrometer in which the entire dispersed spectrum can be read off simultaneously with an imaging detector). The desire for a broad bandpass and high resolution may lead the user to a calorimeter, in spite of its size, weight, and complexity. For some applications, the expected counting rate may be so high as to rule out some of the energy-dispersive detectors. For example, in many synchrotron applications, the detector of choice may be a photodiode (which measures current and not individual photons), and a dispersive device will be used as the spectrometer. Besides these complex and high-resolution experiments at synchrotron radiation centers, portable equipment for X-ray fluorescence analysis has been developed. The technological progress is based upon on miniature and air-cooled X-ray tubes, often polycapillary optics, thermoelectrically cooled X-ray detectors, multichannel analyzers, and dedicated software (Tsuji, Injuk and van Grieken, 2004). All components have small weight and are incorporated in a pistol-like housing. This portable energy-dispersive X-ray fluorescence equipment has been employed in many applications, such as archaeometry, analysis of lead in paint, environmental analysis, analysis of industrial alloys, soil analysis even on Mars, and so on.

Another factor that affects the choice of spectrometer is the nature of the optical system of which it is a component. For example, the small size of a calorimeter detector requires a tightly focused beam. Flat dispersive spectrometers (as in Figure 4.4) must receive all the radiation at the same incident angle; otherwise, the dispersed image will be blurred (thus effectively reducing the resolving power). As noted in Section 3.1.3, for some applications, crystals and gratings are curved so that the incident radiation will strike the crystal at nearly the same angle.

Acknowledgments

Thomas H. Markert is grateful to Richard Aucoin, Daniel Dewey, Keith Gendreau, George Clark, and Una Hwang for helpful discussions.

Eckhart Förster wishes to thank Ingo Uschmann, Oldrich Renner, Konrad Goetz, and Michael Wendt for their critical remarks and is also grateful to Jana Brusberg and Richard Hutcheon for their help in the preparation of the chapter.

Glossary

Blazed Gratings: Gratings for which the bar, groove, or line shape is structured so as to enhance reflection into a particular order.

Cryogenic Devices: A relatively new class of X-ray spectrometers that operates at very low temperatures ($T < 2$ K typically). Such devices are capable of extremely high resolving powers ($\Delta \mathrm{E} \sim$ 7 eV has been demonstrated at 5.9 keV for at least one such device).

Diffraction: Modification of the intensity and/or phase of an electromagnetic wave by the presence of an object (such as a slit, hole, edge, etc.) in the path of the wave.

Dispersive Spectrometer: An X-ray spectrometer that employs wavelength dispersion. X-rays are dispersed (i.e., spatially separated) by the effects of diffraction. Dispersive spectrometers are sometimes called *wavelength-dispersive spectrometers*. The two classes of dispersive spectrometers are Bragg devices and diffraction gratings.

Escape Peak: For monochromatic incident X-ray photons with an energy E_x smaller than that of the absorption edge of detector atoms, the pulse height output is proportional to E_x. However, when the energy E_x of the monochromatic incident X-ray photons exceeds the absorption edge of the detector atoms, the output may contain two pulse height distributions. The additional or escape peak has a mean pulse height proportional to the difference between the energies of the incident photons E_x and the escaping (from active detector volume) photons $E_{\mathrm{K}\alpha}$.

Monochromators: Devices used for producing radiation at a single wavelength (or, more realistically, in a narrow range of wavelengths). The elements of dispersive spectrometers (crystals and gratings) can also be used as X-ray monochromators.

Multilayers: Synthetic crystals formed by sputtering or evaporating alternating layers of high-Z and low-Z materials onto a substrate. Multilayers can be used as spectrometers or monochromators and can be a part of an X-ray optics system to enhance the performance of the system over a narrow band and is also called *layered synthetic microstructures* or *LSMs*.

Nondispersive Spectrometer: An X-ray spectrometer that does not employ wavelength dispersion. Such devices generally operate by converting the photon energy into some other sort of particle or quasiparticle (e.g., electron–ion pairs or phonons). Nondispersive spectrometers are often called *energy-dispersive spectrometers*.

Resolving Power: The resolving power of a spectrometer (at an energy $E = hc/\lambda$) is usually defined as the ratio of the energy (or wavelength) of interest to the width of the response function of the spectrometer to a monochromatic X-ray line, that is, $R(E) = E/\Delta E = \lambda/\Delta\lambda$. The width of the line is usually chosen to be the FWHM of the spectral response function. Note that some authors invert this definition of resolving power so that it equals $\Delta E/E = \Delta\lambda/\lambda$ and often quote it as a percentage.

Rocking Curve: In crystal spectrometers, the rocking curve is a measure of the spectral resolving power of the crystal, independent of any geometric effects introduced by the geometry of the spectrometer. Specifically, it is the FWHM in degrees of the response of the crystal (as it is rotated or rocked) to a monochromatic X-ray beam.

Rowland Circle Configuration: This configuration is common in focusing X-ray spectrometers. The diverging X-ray source, the diffraction grating, and the detector all lie on a circle (the Rowland circle, after Henry Rowland) that has a diameter equal to the radius of curvature of the curved diffractor. Such a configuration gives one-dimensional focusing and disperses the spectrum along the circle. A variation, using a curved crystal rather than a grating, allows for imaging of small fields along the circle.

References

Agarwal, B.K. (1991) *X-ray Spectroscopy*, Springer, Berlin.

Attwood, D. (1999) *Soft X-rays and Extreme Ultraviolet Radiation: Principles and Applications*, Cambridge University Press, Cambridge.

Authier, A. (2001) *Dynamical Theory of X-ray Diffraction*, Oxford University Press, Oxford.

Bargheer, M., Zhavoronkov, N., Bruch, R., Legall, H., Stiel, H., Woerner, M., and Elsaesser, T. (2005) *Appl. Phys. B*, **80**, 715–719.

Becker, P. (2001) *Rep. Prog. Phys.*, **64**, 1945.

Bragg, W.H. and Bragg, W.L. (1913) *Proc. Phys. Soc., London A*, **88**, 428–438.

Bressler, C. and Chergui, M. (2004) *Chem. Rev.*, **104**, 1781–1812.

Compton, A.H. and Allison, S.K. (1967) *X-rays in Theory and Experiment*, 2nd edn, Van Nostrand, Princeton.

Fano, U. (1947) *Phys. Rev.*, **72**, 26.

Fraser, G.W. (1989) *X-ray Detectors in Astronomy*, Cambridge University Press, Cambridge, UK.

Gatti, E. and Rehak, P. (1984) *Nucl. Instrum. Methods, A*, **225**, 608–614.

Golovkin, I., Mancini, R., Louis, S., Ochi, Y., Fujita, K., Nishimura, H., Shirga, H., Miyanaga, N., Azechi, H., Butzbach, R., Uschmann, I., Förster, E., Delettrez, J., Koch, J., Lee, R.W., and Klein, L. (2002) *Phys. Rev. Lett.*, **88**, 0450021–0450024.

Golub, L., Zirin, H., and Wang, H. (1994) *Solar Phys.*, **153**, 179–198.

Gursky, H. and Zehnpfennig, T. (1966) *Appl. Opt.*, **5**, 875–876.

Höll, A. *et al.* (2007) *High Energy Density Phys.*, **3**, 120–130.

Holt, S.S. (1987) *Astrophys. Lett. Commun.*, **26**, 35.

James, R.W. (1982) *The Optical Principles of the Diffraction of X-rays*, Ox Bow Press, Woodbridge, CT.

Janssens, K., Adams, F., and Rindby, A. (2000) *Microscopic X-ray Fluorescence Analysis*, John Wiley & Sons, Ltd, Chichester.

Jasny, J. *et al.* (1994) *Rev. Sci. Instrum.*, **65**, 1631.

Jenkins, R. (1988) *X-ray Fluorescence Spectrometry*, John Wiley & Sons, Ltd, New York.

Kelley, R.L., Holt, S.S., Madejski, G.M., Moseley, S.H., Schoelkopf, R.J., Szymkowiak, A.E., McCammon, D., Edwards, B., Juda, M., Skinner, M., and Zhang, J. (1988) in *X-ray Instrumentation in Astronomy II, SPIE Proceedings No. 982* (ed. L. Golub), SPIE, Bellingham, WA, p. 219.

Klöpfel, D., Hölzer, G., Förster, E., and Beiersdörfer, P. (1997) *Rev. Sci. Instrum.*, **68**, 3669.

Knoll, G.F. (1989) *Radiation Detection and Measurement*, 2nd edn, John Wiley & Sons, Ltd, New York.

Koningsberger, D.C. and Prins, R. (eds) (1988) *X-ray Absorption: Principles, Applications, Techniques of EXAFS, SEXAFS, and XANES*, John Wiley & Sons, Ltd, New York.

Labov, S.E. and Young, B.A. (1993) *J. Low Temp. Phys.*, **93**, 185–858.

Love, G. (2002) *Microchim. Acta*, **138**, 115.

Michette, A.G. and Buckley, C.J. (eds) (1993) *X-ray Science and Technology*, Institute of Physics Publishing, Bristol, UK.

Mißalla, T., Uschmann, I., Förster, E., Jenke, G., and von der Linde, D. (1999) *Rev. Sci. Instrum.*, **70**, 1288–1299.

Moseley, S.H., Mather, J.C., and McCammon, D. (1984) *J. Appl. Phys.*, **56**, 1257.

Newbury, D., Wollman, D., Nam, S.W., Hilton, G., Irwin, K., Small, J., and Martinis, J. (2002) *Microchim. Acta*, **138**, 265.

Pinsker, Z.G. (1978) *Dynamical Scattering of X-rays in Crystals*, Springer, Berlin.

Ramana Murthy, P.V. and Wolfendale, A.W. (1993) *Gamma–Ray Astronomy*, Cambridge University Press, Cambridge, MA.

Renner, O., Kopecky, M., Krousky, E., Förster, E., Mißalla, T., and Wark, J.S. (1994) *Laser Particle Beams*, **12**(3), 539.

Sauli, F. (ed) (1992) *Instrumentation in High Energy Physics*, World Scientific, Singapore.

Schattenburg, M.L., Canizares, C.R., Dewey, D., Flanagan, K.A., Levine, A.M., Lum, K.S., Manikkalingam, R., and Markert, T.H. (1991) *Opt. Eng.*, **30**, 1590–1600.
Silver, E. and Kahn, S. (eds) (1993) *UV and X-ray Spectroscopy of Laboratory and Astrophysical Plasmas*, Cambridge University Press, Cambridge, UK.
Spiller, E. (1994) *Soft X-ray Optics*, SPIE Optical Engineering Press, Bellingham, WA.
Strüder, L., Fiorini, C., Gatti, E., Hartmann, R., Holl, P., Krause, N., Lechner, P., Longoni, A., Lutz, G., Kemmer, J., Meidinger, M., Popp, M., Soltau, H., Weber, U., and von Zanthier, C. (1998) *J. Synchrotron Radiat.*, **5**, 268.
Tsuji, K., Injuk, J., and van Grieken, R. (2004) *X-ray Spectrometry: Recent Technological Advances*, John Wiley & Sons, Ltd, New York.
Uschmann, I., Malgrange, C., and Förster, E. (1997) *J. Appl. Crystallogr.*, **30**, 4554.

Further Reading

Bertin, E.P. (1975) *Principles and Practice of X-ray Spectrometric Analysis*, 2nd edn, Plenum, New York.
Dyson, N.A. (1990) *X-rays in Atomic and Nuclear Physics*, 2nd edn, Cambridge University Press, Cambridge, UK.
Hows, M.J. and Morgan, D.V. (eds) (1979) *Charge-Coupled Devices and Systems*, John Wiley & Sons, Ltd, Chichester.
Russ, J.C. (1984) *Fundamentals of Energy Dispersive X-ray Analysis*, Butterworths, London.
VanGrieken, R.E. and Markowicz, A.A. (eds) (1993) *Handbook of X-ray Spectroscopy: Methods and Techniques*, Marcel Dekker, New York.

19
Linear Laser Spectroscopies

Stephen H. Ashworth

19.1 Introduction 697
19.2 Direct Techniques 698
19.2.1 Direct Absorption 698
19.2.2 Multipass Techniques 698
19.2.3 Cavity Ring Down 699
19.2.4 Cavity-Enhanced Absorption 699
19.2.5 Intracavity Techniques 700
19.2.6 Laser Magnetic Resonance Spectroscopy 700
19.2.7 Raman Spectroscopy 701
19.2.8 Polarization Spectroscopy 703
19.3 Proxy Techniques 703
19.3.1 Laser-Induced Fluorescence 704
19.3.1.1 Imaging 705
19.3.1.2 Dispersed Fluorescence 705
19.3.1.3 Stimulated Emission Pumping 706
19.3.1.4 Intermodulated Fluorescence 706
19.3.1.5 Saturation Dip Spectroscopy 706
19.3.1.6 Quantum Beat Spectroscopy 707
19.3.1.7 Single-Molecule Spectroscopy 707
19.3.1.8 Light Detection and Ranging (LIDAR) 707
19.3.2 Photoacoustic Spectroscopy 708
19.3.3 Photoelectron Spectroscopy 708
19.3.4 Resonance-Enhanced Multiphoton Ionization 709
19.3.5 Optogalvanic Spectroscopy 710
19.3.6 Frequency Modulation Spectroscopy 711
19.3.7 Thermal Lensing Spectroscopy 712
19.4 Generally Applicable Techniques 712
19.4.1 Velocity Modulation 712
19.4.2 Double Resonance 713

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

19.4.3 Stark Spectroscopy 713
References 713
Further Reading 716

19.1
Introduction

As lasers have developed, they have become a common and important tool for spectroscopy. Applications range from probing samples remotely at great distance to generating transient species. In this chapter, we confine ourselves to linear spectroscopic techniques. In general, therefore, the signal response is proportional to the laser power. The only exceptions to this are when saturation spectroscopies and other two-photon techniques are considered. Here, the overall signal response is no longer directly proportional to the laser power but the absorption of each individual photon is still in the linear regime.

It is the unique properties of laser radiation that have made it such a powerful spectroscopic tool. Continuous wave (CW) lasers can be constructed such that their linewidth is very narrow (below 1 MHz), thus enabling high-resolution measurements. The high spectral brightness of lasers increases the probability that two or more photons may interact with the same absorber, which also makes multiple photon spectroscopies possible. In addition, the very well defined temporal envelopes produced by pulsed laser systems allow time-resolved measurements to be carried out on a range of timescales. Finally, the highly directional nature of laser radiation makes optical manipulation very simple and allows some measurements to be made remotely.

No distinction has been made here between samples. The techniques outlined can be applied to a variety of samples: not exclusively atomic or molecular. Many of the techniques, however, lend themselves to application in the gas phase where the full resolution of laser sources may be applied. Liquid, solid, or solution phases have, however, not been excluded.

This chapter is divided into two main sections. The first deals with applications that detect laser intensity directly. To monitor a small change on what is often very high laser intensity reliably, a detector has to be capable of a very high dynamic range. For this reason, many laser interactions are detected indirectly by means of a side effect. These techniques are here referred to as *proxy techniques* and are described in the second section.

A short concluding section deals with examples of named techniques that straddle these rather artificial boundaries or may be used with a variety of

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

approaches. This is by no means meant to be an exhaustive list and neither is it designed to describe these methods in great depth. The reader can use the citations and review articles to obtain further details.

19.2
Direct Techniques

19.2.1
Direct Absorption

The very high spectral brightness of laser sources means that direct absorption measurements are generally limited by the dynamic range of the detector. In many cases, the signal in which we are interested is a small change in a large detected intensity, hence limiting the overall sensitivity. Methods have been developed to overcome this disadvantage, which either increase the proportion of the sample response or detect the response as a function of time.

One method to increase the proportion of sample response is simply to increase the length over which the sample and laser radiation interact. This is of special importance in dilute samples of low-pressure gases. Pressure broadening studies or measurements of small concentrations of atmospheric absorbers might require such an approach. The length of the sample cell can only be extended if the sample itself is stable and the physical dimensions of the laboratory allow this. Frequency measurements of pure rotational transitions in vibrationally excited H_2O molecules (Matsushima *et al.*, 2006) have used the latter approach with a single pass in a 1.8-m sample cell.

19.2.2
Multipass Techniques

In many cases, it is not possible simply to increase the length of the sample. This may be a result of the physical limitations of a laboratory or the spatial extent of the sample. Few laboratories can accommodate sample cells much longer than 2 m and transient species are often confined to a small region of space. In order to overcome such limitations, or to maximize path length when the sample can only be produced in a limited region, the laser radiation may be passed through the sample many times. In the simplest of these, a pair of plane mirrors may be arranged such that the light is redirected through the sample as many times as possible before reaching the detector. The low divergence of a laser source is a great benefit in these experiments (Qian and Howard, 1997).

More sophisticated approaches use systems of curved mirror cells such as the White (White, 1942) and Herriott (Herriott, Kogelnik and Kompfner, 1964; Herriott, 1963) cells. These optical arrangements are designed to minimize losses in order to produce a much longer effective path length than with plane mirrors. The beam path, however, imposes some geometrical constraints with the result that a given solution is not universally appropriate: a laser beam in a traditional White cell arrangement, for example, tends to probe a single plane. A White or Herriott cell would be a good choice if the sample were sufficiently stable to fill the cell. On the other hand, a multipass arrangement with fewer reflections from plane mirrors may well prove to be a more sensitive solution for a transient sample, which is confined to a small region of space.

19.2.3
Cavity Ring Down

Cavity ring-down spectroscopy (*CRDS*) might be considered close to the ultimate multipass technique, which is reviewed in Berden, Peeters and Meijer (2000). The sample is contained in a high-finesse optical cavity that is designed to minimize losses from diffraction and scattering. Laser light is coupled into the cavity and then quickly switched off. CW lasers are generally coupled to the cavity using an active control to scan or lock the CRDS cavity and often require active switching of the light. On the other hand, the broad bandwidth of a pulsed laser generally overlaps with at least one of the closely spaced cavity modes without the need for active cavity length adjustment. As a result, what the arrangement lacks in resolution it more than makes up for in simplicity.

Once the cavity excitation stops, the stored intensity leaking through one of the mirrors may be monitored as a function of time. A compromise has often to be reached between the wavelength range over which the mirrors are highly reflective and the reflectivity of the mirrors. In general, the more highly reflective the mirrors are, the smaller the wavelength range over which they are highly reflective.

In a CRDS cavity, the number of round trips, n, required for the light to reach $1/e$ of the original intensity is given by $n = -1/[2\ln(R)]$, where R is the mirror reflectivity. The intensity of light confined in a cavity bounded by mirrors of reflectivity $R = 0.9999$ will fall to $1/e$ of the original intensity after 5000 round trips. Thus, a cavity with such mirrors separated by 1 m and filled with sample produces an effective path length of around 10 km.

Active switching of light out of (or into) the cavity is obviously not necessary if pulsed lasers are used; for CW lasers, the switching is usually effected by deflecting the beam with an acousto-optic modulator. The intensity leaking from the cavity decays exponentially with a time constant that is determined by cavity losses and may be as long as tens of microseconds. The difference in time constant between an empty cavity and one containing sample gives a direct measure of the absorption coefficient of the sample. One elegant feature of this technique is that the signal is, at least in principle, immune to power fluctuations in the laser. Part of the absorption spectrum of the iodine monofluoride radical recorded using CRDS is shown in Figure 19.1.

19.2.4
Cavity-Enhanced Absorption

The technique of CRDS has been developed further to avoid switching the laser out of the cavity. This has become known as *cavity-enhanced absorption spectroscopy* (*CEAS*) or *integrated cavity output spectroscopy* (*ICOS*). The laser may be scanned rapidly and sometimes repeatedly over a portion of the sample spectrum and the absorption recorded is the integrated output of the cavity. This has the advantage of simplicity compared to cavity ring-down as the decay of intensity is not recorded as a function of time. It does, however, lose the immunity to power fluctuations of the light source. Off-axis cavity designs have recently become popular as the alignment becomes very robust, which is attractive when designing field instruments (Moyer *et al.*, 2008).

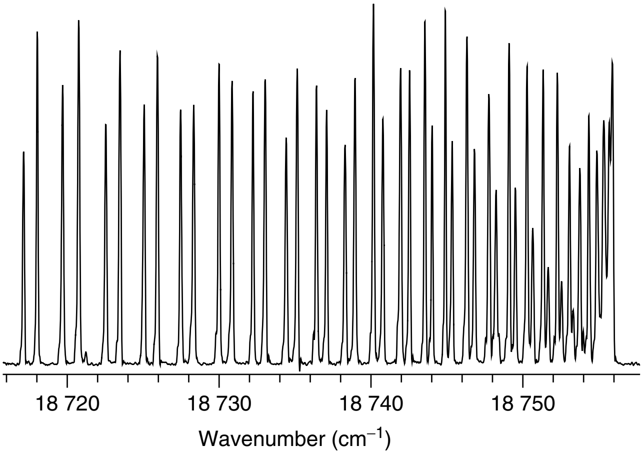

Fig. 19.1 A portion of the absorption spectrum of the iodine monofluoride radical produced by the flash photolysis of CF_3I recorded using cavity ring-down spectroscopy. (Ashworth S.H., Joseph D.M. and Plane J.M.C. *unpublished*.)

The integration of the cavity output and the need for stable lasers means that CEAS requires CW lasers or at least quasi-CW lasers. In common with CRDS, the technique may be applied over a wide wavelength range. It has been applied to make a portable spectrometer to measure ratios of carbon isotopes of CO_2 in the 2 μm region (Wehr *et al.*, 2008) and also to detect atmospherically important species in the ultraviolet region (Mejean, Kassi and Romanini, 2008).

19.2.5
Intracavity Techniques

In some cases, the sample may be usefully placed within a CW laser cavity. The technique is often known as *intracavity laser absorption spectroscopy* (*ICLAS*). These samples may be thought of as being exposed to an infinite number of round trips of the laser light. A small loss due to absorbance has a large effect on the gain of the laser (Peterson *et al.*, 1971). Such a description is, however, deceptively simple. Although the signal is obtained simply by monitoring the laser output, sensitive measurements are only possible with very stable laser sources. The laser itself needs to have been modified. The cavity quality factor, Q, has probably been reduced and the alignment changed by the introduction of a sample cell. Second, the signal itself is a result of a further change in Q as the measurement is taken. Consequently, active stabilization of the laser can become challenging. Nevertheless, the technique is capable of extremely sensitive measurements (Ding *et al.*, 2005).

19.2.6
Laser Magnetic Resonance Spectroscopy

Laser magnetic resonance (*LMR*) is a technique used in spectral regions where laser radiation is not easily tunable. This is the case in the mid- and far infrared. The mid-infrared carbon monoxide laser and a number of far infrared gas lasers operate over a range of discrete frequencies. The gases are lasing on one of many molecular transitions so the output may be one of a range of discrete frequencies. Therefore, these discretely tunable lasers are known as *line tunable lasers*.

The samples under investigation rarely have absorptions that coincide with individual laser transitions, so in the absence of tunable laser light the molecular transitions themselves are tuned. The sample must, therefore, be a radical species in order that the magnetic field may interact with the sample. This selectivity may be used to advantage to detect radical species in a complex mixture. The sample is arranged so that it may be exposed to a magnetic field either between the poles of an electromagnet or within a solenoid. The spectrum is recorded as a function of field. A small, sinusoidal magnetic field is superimposed on the scanned magnetic field and phase-sensitive detection is generally used.

In the far-infrared, gases such as methanol, hydrazine, and CF_2H_2 are used as lasing media. They are generally pumped by CO_2 lasers and give a range of discrete lines with wavelengths in the 50–500-μm range. The gases are lasing on rotational transitions and isotopic variants give access to a range of alternative laser frequencies. The frequencies of known laser transitions are independently measured and periodically tabulated (Inguscio *et al.*, 1986; Zink *et al.*, 2008). New laser transitions are occasionally discovered and have to be precisely measured before they may be used for high-resolution measurements. The development of frequency combs has meant that complicated frequency chains are no longer necessary for accurate frequency determinations (Beverini *et al.*, 2005).

Most LMR spectrometers are designed with the sample within the laser cavity and thus LMR is usually an intracavity technique. The signal is obtained by monitoring the laser output as a function of magnetic field. Figure 19.2 shows part of an LMR spectrum recorded during a study of the HSS radical in the far infrared.

The spectrometer can, however, be designed so that the laser is remote from the sample. The intracavity advantage is lost in this case, so it may be necessary to construct a multipass cell around the magnet (Hinz *et al.*, 1985). On the other hand, this arrangement allows the Faraday (*vide infra*) or Voigt effects to be exploited, which increases the detection sensitivity (Gillett, Cooksy and Brown, 2006).

19.2.7
Raman Spectroscopy

The narrow linewidth and intensity of laser sources has enabled *Raman spectroscopy* to become a mainstream technique. Raman spectroscopy enables vibrational spectra to be recorded using visible radiation albeit with rather different selection rules than apply to infrared absorption. Energy may both be removed from (Stokes) or added to (anti-Stokes) the incident radiation field from ground state and excited vibrational modes, respectively. These inelastic scattering processes may conveniently be pictured as absorption to, and emission from, a virtual state. A single-frequency CW laser is generally used and scattered light is collected and dispersed. The shifts in frequency are generally small as

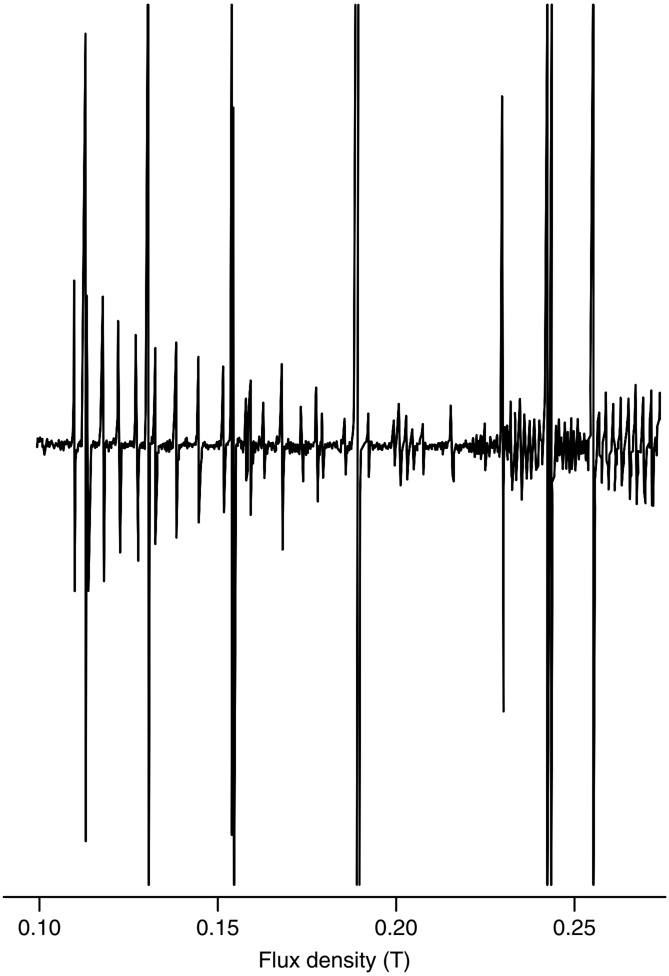

Fig. 19.2 Part of a far-infrared LMR spectrum recorded during a study of the HSS radical using the 203 μm laser line of $^{13}CH_3OH$. The first derivative lineshape is a result of the phase-sensitive detection used. (Ashworth S.H., Brown J.M. and Evenson K.M. *unpublished*.)

they usually correspond to vibrational spacings in the sample.

Rayleigh scattering is far more intense than the Raman shifted components of the scattered light and care must be taken to reduce the intensity of this light reaching the detector. Double and even triple monochromators may be employed for this. Advances in coating technologies, however, mean that filters can now be produced with extremely sharp cutoffs between high and low optical density, which makes experimental arrangements much simpler.

Naturally, the region where the Stokes lines are sought in a Raman spectrum is exactly the region where one might expect intense fluorescence. This is a problem, which often occurs because fluorescence is generally orders of magnitude more intense than the Raman shifted signals. This can be avoided by depositing the sample on a suitable substrate (e.g., graphite), which efficiently quenches fluorescence, or by using a sufficiently long excitation wavelength that the sample is not electronically excited. The latter approach cannot, of course, be applied to more sophisticated variants of the technique such as resonance Raman spectroscopy (Mak and Kincaid, 2008).

A suitable surface, such as rough silver, may increase the signal by orders of magnitude. This constitutes *surface-enhanced resonance Raman spectroscopy* (*SERRS*) (Smith, Faulds and Graham, 2006). Coherent anti-Stokes Raman spectroscopy (CARS) is strictly a nonlinear technique but should at least be mentioned here (Zheltikov, 2000). The CARS signal has the advantage of being highly directional and is therefore relatively straightforward to manipulate optically, to achieve high rejection of unwanted light.

19.2.8
Polarization Spectroscopy

Polarization spectroscopy may be implemented in a number of ways. In the simplest form, the sample is placed in a strong magnetic field, directed along the laser axis, between crossed polarizers. The Faraday effect in an isotropic sample, such as a gas, causes absorptions to rotate the plane of polarization of the laser light. Signals are thus detected against a very low background making this potentially a very sensitive technique.

In a second application, counterpropagating beams of unequal intensity from a narrow bandwidth laser may be used. The higher intensity pump beam is circularly polarized and the weaker probe beam is linearly polarized. The sample is placed between crossed polarizers through which the probe beam is directed. Absorption of the pump beam produces in a nonequilibrium population of magnetic sublevels. This disequilibrium is referred to as an *orientation* and in turn leads to a rotation in the polarization of the probe beam. The detector, behind the crossed polarizer, therefore, only detects signal when both pump and probe interact with the same population. This only occurs when the interaction is with the fraction of the population, which has no component of velocity along the laser axis, and is thus an inherently sub-Doppler technique (Demtröder, 2008).

An extension of this is known as polarization labelling spectroscopy. In this variant, the two beams are no longer derived from the same laser (and as a result the technique is not sub-Doppler). Like other double-resonance techniques (see below), it is usually applied to spectroscopic problems to use secure assignments from one set of bands to gain information about the assignment of a second set (Adohi-Krou *et al.*, 2008). The known transition is pumped with the first, fixed, laser and only transitions in the second set of bands with common electronic states are observed on scanning the second laser.

19.3
Proxy Techniques

The techniques considered up to now have required that a change in laser power be detected in some manner after interaction with a sample. An alternative and sometimes more sensitive approach is to detect a process that results from the absorption of a photon. Care must be exercised with what are here referred to as *proxy techniques*, as they will tend to yield different information from, and have a more limited range of applicability than, direct absorption.

19.3.1
Laser-Induced Fluorescence

The technique of *laser-induced fluorescence* (*LIF*) relies on the fate of the excited state formed on absorption. The word fluorescence itself implies that the upper state is an excited electronic state. As a result, the majority of LIF applications are in the visible and ultraviolet regions of the spectrum.

If the sample has a nonzero quantum yield of fluorescence, the total fluorescence intensity measured as a function of wavelength produces an excitation spectrum, which is a proxy for the absorption spectrum. The spectrum thus formed is a measure of the production of fluorescing, excited state molecules. The fluorescence signal may be shown to be proportional to the number of molecules present in the sample and, in the absence of other wavelength-dependent effects, the signal faithfully reproduces the absorption spectrum.

The absorption spectrum is not faithfully reproduced if other pathways to dispose of the excitation are available. In the gas phase, at low pressure, examples of potential pathways are dissociation and predissociation; the energy of the photon is sufficient to cause bond rupture. Often the fragments are not electronically excited and thus unable to emit, thereby reducing or even completely quenching the overall fluorescence intensity. At higher pressures in the gas phase or solution phase, collisional quenching may become sufficiently important that the efficiency of LIF is reduced.

A great advantage of LIF is that it is essentially a zero background technique. Fluorescence emission is collected off-axis from the excitation source (usually at right angles), so a highly sensitive detector may be used. Unfortunately, a lot of effort and careful experimental design are required to collect more than only a small proportion of the emission. In most cases, the detection sensitivity is sufficient so that it is unnecessary to make this additional effort. The most sophisticated arrangements implement collection optics that not only increase the efficiency of collection but may also provide spatial filtering, which helps discriminate against other sources of light. One of the most efficient systems reported uses spherical mirrors and collects around 30% of the emitted light (Korter *et al.*, 2001).

Scattered light from the excitation source is often the hardest to reject but LIF may also be applied in environments where other sources of light are visible to the detector. LIF is readily applied to furnaces, plasmas, and flames. Such sources may generate a continuum of frequencies where others may generate a structured emission or a combination of the two. A variety of techniques are used to reject unwanted signal at the detector. Some, such as time gating, lock-in detection, and phase-sensitive detection, use a well-defined time dependence of signal – either a modulation or a pulse – to discriminate against both noise and unwanted light. These techniques are, however, of no use if the detector itself is swamped by unwanted illumination.

The sensitivity of LIF, the fact that the signal is proportional to the number of emitters, and above all the selectivity of the technique means that it is easily applied to analytical and kinetic measurements. When observing changes in concentration using LIF, the laser source is not scanned and the signal will often be normalized to take account of variations in the laser power. Once again background light, including laser scatter, may be a problem. In such cases, bandpass filters are effective means to reduce the unwanted light intensity reaching the detector. A monochromator is a very convenient adjustable filter, both in terms of wavelength and bandpass.

When the band of wavelengths detected includes the laser wavelength, this is known as *resonant fluorescence detection*. Nonresonant fluorescence detection is possible if there is a well-defined transition to another energy level. The band of wavelengths required to detect fluorescence emission now no longer includes the laser wavelength. The great advantage of nonresonant fluorescence detection is, therefore, the high rejection of scatter from the excitation source.

19.3.1.1 Imaging

One of the benefits of a laser source is that the radiation need not be in the form of a beam on interaction with the sample. If the laser light is transformed into a sheet, fluorescence may be excited in a plane in the sample. The fluorescence may then be imaged onto a two-dimensional detector allowing dynamic processes, such as combustion, to be probed (Luong *et al.*, 2008). If the sheet of light is arranged so that it may be scanned in a direction perpendicular to the plane of light, a three-dimensional image may be built up.

19.3.1.2 Dispersed Fluorescence

A natural extension to the technique of nonresonant LIF detection mentioned above is to measure the fluorescence intensity as a function of emission wavelength. This is known as *dispersed* or *resolved fluorescence* and gives information on the ground state of the sample, shown schematically in Figure 19.3. Given that the collection efficiency of fluorescence is generally rather low, dispersed fluorescence usually requires quite careful alignment and sensitive equipment.

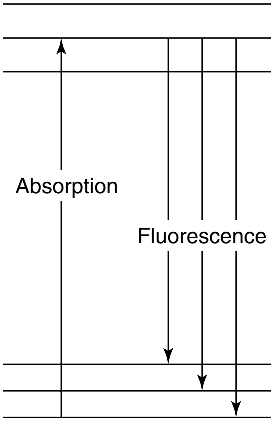

Fig. 19.3 The principle of laser-induced fluorescence. The excitation spectrum is recorded by scanning the absorbed photon and recording the total fluorescence. The wavelength dependence of the emission recorded with a fixed absorption wavelength forms the dispersed (or resolved) fluorescence spectrum.

The resolution is generally determined by the recording instrument. Any systems that are capable of resolving the fluorescence may be used to record dispersed fluorescence. Compact charge-coupled device (CCD) spectrometers offer low resolution rugged devices, whereas the highest resolution is more likely to be afforded by Fourier transform spectrometers (Hodges *et al.*, 2007), but these are much less rugged. The instrument of choice is often dictated by a compromise between the required resolution, fluorescence intensity, and the robustness that is required of the equipment.

19.3.1.3 **Stimulated Emission Pumping**

An alternative to dispersed fluorescence is the technique of *stimulated emission pumping* (*SEP*). A first photon fixed at one wavelength, as in dispersed fluorescence, is used to excite fluorescence. When a second photon incident on the sample is resonant with a downward transition, emission is stimulated or pumped. This leads to a decrease in the excited population and hence a decrease in the fluorescence intensity. The signal level is monitored as a function of the wavelength of a second photon. The signal is usually displayed as the ratio between fluorescence intensity before and after the pump laser as shown by the shaded areas in Figure 19.4. A good example is given in Mukarakate *et al.* (2008) where the normally downward going lines have also been reversed for clarity. The additional complexity of the technique in that it requires two lasers is compensated by the resolution, which is generally limited only by the pump laser.

19.3.1.4 **Intermodulated Fluorescence**

Intermodulated fluorescence is one of the techniques that may be used to reach sub-Doppler resolution. It involves single photon absorption from counterpropagating laser beams. The beams are chopped at different frequencies ω_1 and ω_2. When the beams are probing the same population, a signal is produced at the sum frequency $\omega_1 + \omega_2$. The only population that is probed by both beams simultaneously is that which has no velocity component along the laser beams. Thus the signal is inherently sub-Doppler and is often used to investigate hyperfine structure (Manke *et al.*, 2008).

19.3.1.5 **Saturation Dip Spectroscopy**

Saturation dip spectroscopy has many names and incarnations. In simple terms, the sample is exposed to counterpropagating beams of sufficient intensity that at line center the signal intensity drops because the lower state population has been depleted. This

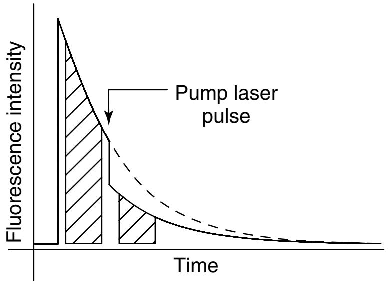

Fig. 19.4 The shaded areas show the areas to be compared in stimulated emission pumping. When the second (pump) laser interacts with the initially excited population, the fluorescence intensity is reduced.

sort of spectroscopy is variously referred to as *Lamb dip, saturation dip,* and *hole burning spectroscopy* in different fields and applications. It is treated in depth in Demtröder's classic book (Demtröder, 2008).

Saturation spectroscopy may be relatively straightforward if the sample is within the cavity of a standing wave CW laser as it is naturally exposed to collinear, counterpropagating radiation.

19.3.1.6 Quantum Beat Spectroscopy

The observation of quantum beats may be thought of as an analog of the observations in a Young's double slit experiment. Closely spaced molecular energy levels may be excited simultaneously by a single, usually pulsed, laser. The laser bandwidth obviously has to be sufficiently broad to excite two or more levels and form a coherent superposition (Ishii, Takeuchi and Tahara, 2008). Modulation at a frequency corresponding to the difference between the excited energy levels is superimposed on the exponential decay of fluorescence. The principle is illustrated in Figure 19.5. A Fourier transform of the beat frequency reveals the energy difference of the levels involved and produces a Doppler-free spectrum even when the energy difference is smaller than the Doppler width of the fluorescence.

19.3.1.7 Single-Molecule Spectroscopy

In many cases, spectroscopic signals are averaged over a large number of molecules, each of which is in a slightly different environment. This averaging effect is known as *heterogeneous broadening* as the signal is made up of contributions from many different individual absorptions. In environments where absorbers are constrained to a small region of space, it is possible to overcome this by recording signal from a single molecule or a single emitter, such as a quantum dot. Obviously, the fluorescence signal will tend to be weak in comparison with ensemble spectroscopy so a single molecule has to be pumped many times to its excited state in a single measurement. The field of view is constrained and the concentration of emitters kept low so that on average only one emitter may be detected at a time. Single-molecule spectroscopy may be used to investigate guest–host interactions and the homogeneity of the local environment. Wustholz *et al.* (2008) and Wirth and Legg (2007) have reviewed this extensive area.

19.3.1.8 Light Detection and Ranging (LIDAR)

Backscattered radiation from a laser directed through the atmosphere may be

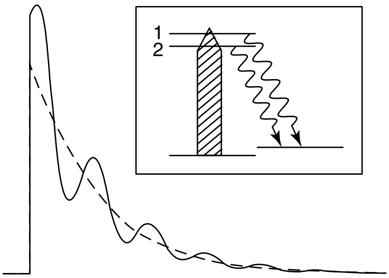

Fig. 19.5 Quantum beats (solid line) superimposed on a simple exponential decay (broken line). The inset shows the associated energy level scheme. The laser excites states 1 and 2 simultaneously and the interference results in quantum beats.

collected and, under favorable circumstances, used to measure the quantity and range of airborne substances. Rayleigh scatter is the strongest signal and is the result of inelastic scattering from all molecules in the atmosphere. This scatter is often carefully screened from detectors in order to allow detection of atomic and molecular signals. On the other hand, the Rayleigh scatter itself may be used to measure important quantities (Schoch, Baumgarten and Fiedler, 2008). The intensity of the scatter is proportional to gas density, the width of the scattered peak is related to gas temperature, and, in addition, one component of bulk velocity results in a shift of the peak wavelength of the scatter.

Resonant atomic fluorescence is perhaps the next strongest source. This allows the concentration of species such as iron and sodium to be measured in the mesosphere, at a range of around 85 km, from ground-based measurements (Gardner *et al.*, 2005). Even a highly collimated laser irradiates a substantial area at this range and very large and efficient collection optics are required to detect the very weak signals involved. Indeed, the signals are often so weak that astronomical telescopes are employed to collect the signal. Closer to home LIDAR is routinely mounted on mobile platforms and may be used to map atmospheric constituents, for example, plumes from volcanoes (Neri *et al.*, 2008) or chimneys.

19.3.2
Photoacoustic Spectroscopy

In an appropriate sample, the absorption of laser radiation results in a pressure wave. This can be detected using the simple expedient of a microphone. Care must be taken, however, that windows and other optical components in the sample cell do not also contribute to the signal. In order to avoid these effects when using a pulsed laser, the cell must be carefully designed (Rabasovic *et al.*, 2008). A CW laser obviously has to be modulated in order to produce a pressure pulse. This may be turned to advantage by ensuring that the sample chamber is resonant at the modulation frequency of the laser. Further isolation may be achieved by using buffer volumes as acoustic filters (Li *et al.*, 2008). This method has been combined with fluorescence detection to enable the quenching mechanisms in solution to be investigated (Kamath, Kartha and Mahato, 2008). *Photoacoustic spectroscopy* appears to be particularly suited as a detection method of absorption in the midlinfrared region (5–12 μm) (Patel, 2008).

19.3.3
Photoelectron Spectroscopy

Photon absorption may transfer enough energy to the sample to liberate an electron. The excess kinetic energy of the electron so liberated may be measured and thus the energy of the resultant ion state inferred. The resolution is limited by the electron energy analyzer and the radiation source but is usually high enough to resolve vibrational structure. The difference between the excitation energy and the kinetic energy of the electron gives information on the electronic structure of the resulting ion as shown in Figure 19.6. *Anion photoelectron spectroscopy*, on the other hand, allows the electronic state of the resulting neutral to be determined.

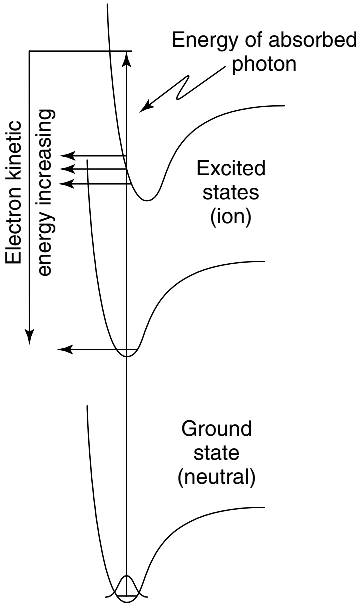

Fig. 19.6 In photoelectron spectroscopy, the ground state of the absorber undergoes a vertical transition to the ionization continuum. The electron carries away the difference in kinetic energy between the absorbed photon and the resultant ion state as indicated by arrows. The pattern of peaks gives information on the geometry change on ionization.

Traditionally, the hard ultraviolet light from a helium discharge has been used to generate photoelectron spectra. The advent of lasers and the availability of synchrotrons have widened the wavelength range over which photoelectrons may be generated and also reduced the linewidth of the exciting light.

The high resolution and tunability of laser sources may be employed to measure the yield of photoelectrons generated just at the ionization threshold. These electrons have a single value of kinetic energy, which is essentially zero, and makes the electron analyzer redundant. This is known as *zero electron kinetic energy* or *ZEKE* spectroscopy (Cockett, 2005). In ZEKE spectroscopy, the electrons produced with nonzero kinetic energy are allowed to disperse before an electric field is applied to the sample region to extract those that remain. The principle of ZEKE was first demonstrated by Chewter and coworkers (Chewter, Müller-Dethlefs and Schlag, 1987). The advantage of this approach is that the resolution is no longer limited by the electron analyzer; instead, it is limited by the laser resolution and the electric fields used to extract the electrons.

19.3.4
Resonance-Enhanced Multiphoton Ionization

Detection of ions is generally much more efficient than the detection of photons. In the simplest implementation of *resonance-enhanced multiphoton ionization* (*REMPI*), the sample is arranged so that it is exposed to a laser whose energy is capable of ionizing the

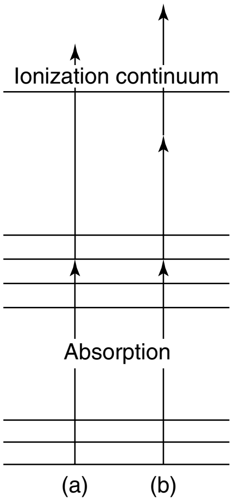

Fig. 19.7 The principle of REMPI. When absorption results in population in an upper state further photons are able to ionize the sample so absorption is detected as an increase in ion current. The second photon may be (a) the same wavelength as the first or (b) a different wavelength.

sample from the upper state of a transition of interest but not directly from its lower state. Another photon, which may result from the same source or a different source, excites the transition of interest results in an increased ion current. Both situations are illustrated in Figure 19.7. The spectrum is recorded as ion current as a function of laser wavelength.

Not only is detection of ions more efficient than the detection of photons, but also time of flight techniques may be used to assist in the identification of ions and fragments. In fact, ionic fragments may be detected by imaging using a two-dimensional detector at the end of a flight tube (Ashfold *et al.*, 2006). These techniques allow the kinetic energy release of the fragments to be reconstructed so that both energetic and dynamic information may be obtained.

19.3.5
Optogalvanic Spectroscopy

A plasma, such as that found in a hollow cathode lamp, is a complex system of excited states and energetic electrons. With a given electrical power input, the system remains in equilibrium. If light whose energy corresponds to an absorption in one of the plasma species, be it an ionic or neutral species is incident on the plasma, the equilibrium is perturbed. This perturbation leads to a change in the power demand from the power supply. The effect was first described in the mid-1920s (Penning, 1928) but not used with lasers until the 1970s Green *et al.* (1976a,b).

Plasmas are generally run at moderately high voltages and a very simple method to obtain the optogalvanic signal is to AC couple the voltage. Thus, the small voltage fluctuations as a result of irradiation are readily detectable. Given that the change in voltage on absorption is often on the order of millivolts, the stability of the power supply is important. Examples of one- and two-photon absorption along with the waveform observed in a commercial hollow cathode lamp are given in Omidyan *et al.* (2008).

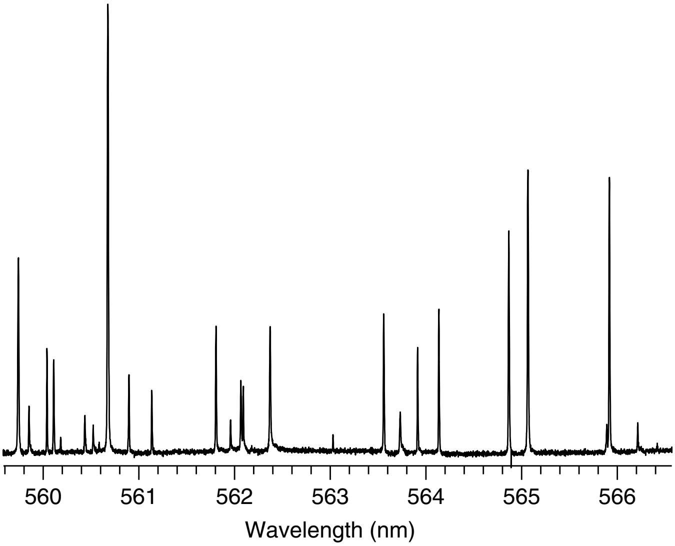

Fig. 19.8 Part of the spectrum of argon atoms recorded using the optogalvanic effect in a commercial hollow cathode lamp. (Greenacre C. and Ashworth S.H. *unpublished*.)

The technique is very straightforward to implement with commercial hollow cathode lamps and is routinely used as a calibration method for pulsed tunable lasers. Part of the spectrum of atomic argon in a commercial hollow cathode lamp is shown in Figure 19.8. Optogalvanic detection can also be applied to plasmas excited by radiofrequency discharges and has been used to detect attomole quantities of $^{14}CO_2$ (Murnick, Dogru and Ilkmen, 2008).

19.3.6
Frequency Modulation Spectroscopy

A laser beam may have sidebands superimposed on it by the use of an electro- or acousto-optic modulator or, in some cases, by direct modulation of the laser itself. The dominant sidebands are 180° out of phase and shifted in frequency from the carrier $\pm\omega$ Hz, where ω is the frequency of the local oscillator. Both sidebands and carrier are directed through the sample and incident on the detector. Off resonance, there is no absorption so the sidebands cancel exactly. This is also the case with the carrier at the absorption line center as both sidebands are absorbed to the same extent – given that the absorption line is symmetric. With the carrier frequency at any other position in the absorption one sideband is preferentially absorbed. The asymmetric absorption of the sidebands results in an AC signal at the detector as shown schematically in Figure 19.9.

One of the first demonstrations of frequency modulation (FM) spectroscopy was on sodium vapor and molecular iodine (Bjorklund, 1980). It is now routinely applied to high-resolution and high-sensitivity measurements in both

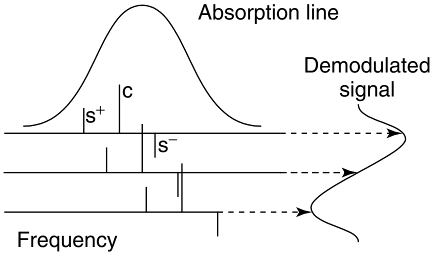

Fig. 19.9 An illustration of FM spectroscopy. The carrier frequency (c) and the sidebands (s+ and s−) which are 180° out of phase are scanned across an absorption line. The resultant demodulated signal is a first derivative lineshape.

spectroscopy and kinetics (Friedrichs, 2008). Combined with cavity enhancement it is known as *noise-immune cavity-enhanced optical heterodyne molecular spectroscopy* or *NICE-OHMS* (Foltynowicz *et al.*, 2008).

19.3.7
Thermal Lensing Spectroscopy

The *thermal lensing* technique relies on the change in density of a sample on interaction with laser radiation. For this reason, it is usually applied to dilute samples in the solid or solution phase. Laser energy is absorbed by the chromophore under investigation, which subsequently undergoes energy transfer. Energy released to the immediate environment results in local heating and thus a change in density. The thermally induced lens causes the laser beam downstream of the target to change its divergence. The irradiance is measured as a function of time by measuring the total energy passing through a small aperture. Thermal lensing may be detected in the excitation beam or in a second probe beam, which is not absorbed by the sample. It has been shown that the influence of solvent (Proskurnin *et al.*, 2008) and the experimental matrix (Georges, 2008) may be important in the interpretation of results. Recent reviews detail the application to detection in flow injection analysis and separation (Franko, 2008) and food analysis and environmental research (Franko, 2001).

19.4
Generally Applicable Techniques

The remaining techniques may find application in situations where either direct or proxy detection schemes are used.

19.4.1
Velocity Modulation

The study of molecular ions either by direct or proxy techniques is often complicated by the presence of neutral species. Such interferences are especially prevalent in the spectroscopy of plasmas and discharges where concentrations of ions may be orders of magnitude lower than that of neutral species. This difficulty was first overcome by Gudeman *et al.* (1983) by the expedient of modulating the electric field responsible for the discharge. The Doppler shift imparted by the sinusoidal variation of the drift velocity acts as a selective frequency modulation. By detecting at the same frequency, by either lock-in or phase-sensitive detection, only the signals resulting from ions are detected.

19.4.2
Double Resonance

SEP *wide infra* is an example of a double-resonance technique. Signal is only observed when two resonant conditions are fulfilled. Such applications may have many different names and involve alternative radiation sources. SEP is one example of *optical–optical double resonance* (*OODR*) where the population of a quantum state is detected by one photon and modulated by a second (Ye, Pang and Cheung, 2007). Population transfer may, however, be achieved by the absorption of a microwave photon that produces either enhancement or reduction of a laser signal: *microwave-optical double resonance* (*MODR*) (Barnett, Ramsay and Zhu, 2005). Similarly, ion dip spectroscopy measures the influence of a second, usually laser, photon on the ion current usually produced by photoionization by a laser (Sakota, Kageura and Sekiya, 2008).

19.4.3
Stark Spectroscopy

Stark spectroscopy is the electric field analog to LMR spectroscopy but tends to be less limited in its range of application. The Stark splitting patterns of transitions may be used to assist in assignments. In combination with tunable lasers, the shift of a single transition may be measured as a function of electric field, which allows the dipole moment in both ground and excited states to be determined (Chen and Steimle, 2008).

References

Adohi-Krou, A., Jastrzebski, W., Kowalczyk, P., Stolyarov, A.V., and Ross, A.J. (2008) Investigation of the D $^1\Pi$ state of NaK by polarisation labelling spectroscopy. *J. Mol. Spectrosc.*, **250**(1), 27.

Ashfold, M.N.R., Nahler, N.H., Orr-Ewing, A.J., Vieuxmaire, O.P.J., Toomes, R.L., Kitsopoulos, T.N., Garcia, I.A., Chestakov, D.A., Wu, S.M., and Parker, D.H. (2006) Imaging the dynamics of gas phase reactions. *Phys. Chem. Chem. Phys.*, **8**(1), 26.

Barnett, M., Ramsay, D.A., and Zhu, Q. (2005) Studies of collisional selection rules in thioformaldehyde (H_2CS) by microwave-optical double resonance. *J. Chem. Phys.*, **123**(15), 154310.

Berden, G., Peeters, R., and Meijer, G. (2000) Cavity ring-down spectroscopy: experimental schemes and applications. *Int. Rev. Phys. Chem.*, **19**(4), 565.

Beverini, N., Carelli, G., De Michele, A., Maccioni, E., Moretti, A., Nyushkov, B., and Sorrentino, F. (2005) High-accuracy frequency measurements in the far infrared. *Laser Phys.*, **15**(7), 1014.

Bjorklund, G.C. (1980) Frequency-modulation spectroscopy – new method for measuring weak absorptions and dispersions. *Opt. Lett.*, **5**(1), 15.

Chen, J.H. and Steimle, T.C. (2008) A molecular beam optical Stark study of nickel monohydride, NiH. *Chem. Phys. Lett.*, **457**(1–3), 23.

Chewter, L.A., Müller-Dethlefs, K., and Schlag, E.W. (1987) Determination of the ionization energy of the benzene-argon complex by zero kinetic energy photoelectron spectroscopy. *Chem. Phys. Lett.*, **135**(3), 219.

Cockett, M.C.R. (2005) Photoelectron spectroscopy without photoelectrons: twenty years of ZEKE spectroscopy. *Chem. Soc. Rev.*, **34**(11), 935.

Demtröder, W. (2008) *Laser Spectroscopy: Experimental Techniques*, 4th revised ed., Springer-Verlag GmbH, Berlin and Heidelberg.

Ding, Y., Campargue, A., Bertseva, E., Tashkun, S., and Perevalov, V.I. (2005) Highly sensitive absorption spectroscopy of carbon dioxide by ICLAS-VeCSEL between 8800 and 9530 cm^{-1}. *J. Mol. Spectrosc.*, **231**(2), 117.

Foltynowicz, A., Schmidt, F.M., Ma, W., and Axner, O. (2008) Noise-immune cavity-enhanced optical heterodyne molecular spectroscopy: current status and future potential. *Appl. Phys. B-Lasers Optics*, **92**(3), 313.

Franko, M. (2001) Recent applications of thermal lens spectrometry in food analysis and environmental research. *Talanta*, **54**(1), 1.

Franko, M. (2008) Thermal lens spectrometric detection in flow injection analysis and separation techniques. *Appl. Spectrosc. Rev.*, **43**(4), 358.

Friedrichs, G. (2008) Sensitive absorption methods for quantitative gas phase kinetic measurements. Part 1: Frequency modulation spectroscopy. *Z. Phys. Chem-Int. J. Res. Phys. Chem. Chem. Phys.*, **222**, 1.

Gardner, C.S., Plane, J.M.C., Pan, W.L., Vondrak, T., Murray, B.J., and Chu, X.Z. (2005) Seasonal variations of the Na and Fe layers at the South Pole and their implications for the chemistry and general circulation of the polar mesosphere. *J. Geophys. Res.-Atmosph.*, **110**(D10), D10302.1–D10302.13.

Georges, J. (2008) Matrix effects in thermal lens spectrometry: Influence of salts, surfactants, polymers and solvent mixtures. *Spectrochim. Acta Part A-Mol. Biomol. Spectrosc.*, **69**(4), 1063.

Gillett, D.A., Cooksy, A.L., and Brown, J.M. (2006) Infrared laser magnetic resonance spectroscopy of the ν_3 fundamental and associated hot bands of the NCO free radical. *J. Mol. Spectrosc.*, **239**(2), 190.

Green, R.B., Keller, R.A., Luther, G.G., Schenck, P.K., and Travis, J.C. (1976a) Galvanic detection of optical absorptions in a gas-discharge. *Appl. Phys. Lett.*, **29**(11), 727.

Green, R.B., Keller, R.A., Schenck, P.K., Travis, J.C., and Luther, G.G. (1976b) Opto-galvanic detection of species in flames. *J. Am. Chem. Soc.*, **98**(26), 8517.

Gudeman, C.S., Begemann, M.H., Pfaff, J., and Saykally, R.J. (1983) Velocity-modulated infrared-laser spectroscopy of molecular-ions - the ν_1 band of HCO^+ . *Phys. Rev. Lett.*, **50**(10), 727.

Herriott, D.R. (1963) Spherical-mirror oscillating interferometer. *Appl. Opt.*, **2**(8), 865.

Herriott, D., Kogelnik, H., and Kompfner, R. (1964) Off-axis paths in spherical mirror interferometers. *Appl. Opt.*, **3**(4), 523.

Hinz, A., Zeitz, D., Bohle, W., and Urban, W. (1985) A Faraday laser magnetic resonance spectrometer for spectroscopy of molecular radical ions. *Appl. Phys. B: Lasers Optics*, **36**(1), 1.

Hodges, P.J., Ross, A.J., Crozet, P., Salami, H., and Brown, J.M. (2007) On the spin-orbit splitting of $CuCl_2$ in its $^2\Pi_g$ ground state. *J. Chem. Phys.*, **127**(2), 024309.

Inguscio, M., Moruzzi, G., Evenson, K.M., and Jennings, D.A. (1986) A review of frequency measurements of optically pumped lasers from 0.1 to 8 THz. *J. Appl. Phys.*, **60**(12), R161.

Ishii, K., Takeuchi, S., and Tahara, T. (2008) Pronounced non Condon effect as the origin of the quantum beat observed in the time-resolved absorption signal from excited-state cis-stilbene. *J. Phys. Chem. A*, **112**(11), 2219.

Kamath, S.D., Kartha, V.B., and Mahato, K.K. (2008) Dynamics of l-tryptophan in aqueous solution by simultaneous laser induced fluorescence (LIF) and photoacoustic spectroscopy (PAS). *Spectrochim. Acta Part A: Mol. Biomol. Spectrosc.*, **70**(1), 187.

Korter, T.M., Borst, D.R., Butler, C.J., and Pratt, D.W. (2001) Stark effects in gas-phase electronic spectra. Dipole moment of aniline in its excited S_1 state. *J. Am. Chem. Soc.*, **123**(1), 96.

Li, J.S., Liu, K., Zhang, W.J., Chen, W.D., and Gao, X.M. (2008) Pressure-induced line broadening for the (30012) ← (00001) band of CO^2 measured with tunable diode laser photoacoustic spectroscopy. *J. Quant. Spectrosc. Radiat. Transfer*, **109**(9), 1575.

Luong, M., Zhang, R., Schulz, C., and Sick, V. (2008) Toluene laser-induced fluorescence for in-cylinder temperature imaging in internal combustion engines. *Appl. Phys. B-Lasers Optics*, **91**(3–4), 669.

Mak, P.J. and Kincaid, J.R. (2008) Resonance Raman spectroscopic studies of hydroperoxo derivatives of cobalt-substituted myoglobin. *J. Inorg. Biochem.*, **102**(10), 1952.

Manke, K.J., Vervoort, T.R., Kuwata, K.T., and Varberg, T.D. (2008) Electronic spectrum of TaO and its hyperfine structure. *J. Chem. Phys.*, **128**(10), 104302.

Matsushima, F., Tomatsu, N., Nagai, T., Moriwaki, Y., and Takagi, K. (2006) Frequency measurement of pure rotational transitions in the $v_2 = 1$ state of H_2O. *J. Mol. Spectrosc.*, **235**(2), 190.

Mejean, G., Kassi, S., and Romanini, D. (2008) Measurement of reactive atmospheric species by ultraviolet cavity-enhanced spectroscopy with a mode-locked femtosecond laser. *Opt. Lett.*, **33**(11), 1231.

Moyer, E.J., Sayres, D.S., Engel, G.S., Clair, J.M.S., Keutsch, F.N., Allen, N.T., Kroll, J.H., and Anderson, J.G. (2008) Design considerations in high-sensitivity off-axis integrated cavity output spectroscopy. *Appl. Phys. B-Lasers Optics*, **92**(3), 467.

Mukarakate, C., Tao, C., Jordan, C.D., Polik, W.F., and Reid, S.A. (2008) Stimulated emission pumping spectroscopy of the X $^1A'$ state of CHF. *J. Phys. Chem. A*, **112**, 466.

Murnick, D.E., Dogru, O., and Ilkmen, E. (2008) Intracavity optogalvanic spectroscopy. An analytical technique for C^{14} analysis with subattomole sensitivity. *Anal. Chem.*, **80**(13), 4820.

Neri, M., Mazzarini, F., Tarquini, S., Bisson, M., Isola, I., Behncke, B., and Pareschi, M.T. (2008) The changing face of Mount Etna's summit area documented with Lidar technology. *Geophys. Res. Lett.*, **35**(9), L09305.

Omidyan, R., Fathi, F., Farrokhpour, H., and Tabrizchi, M. (2008) One- and two-photon laser optogalvanic spectroscopy of neon in the 570-626 nm region. *Opt. Commun.*, **281**(22), 5555.

Patel, C.K.N. (2008) Laser photoacoustic spectroscopy helps fight terrorism: high sensitivity detection of chemical Warfare Agent and explosives. *Eur. Phys. J.-Special Top.*, **153**, 1.

Penning, F.M. (1928) Demonstratie van een nieuw photoelectrisch effect. *Physica*, **8**, 137.

Peterson, N.C., Kurylo, M.J., Braun, W., Bass, A.M., and Keller, R.A. (1971) Enhancement of absorption spectra by dye-laser quenching. *J. Opt. Soc. Am.*, **61**(6), 746.

Proskurnin, M., Bendrysheva, S., Kuznetsova, V., Zhirkov, A., and Zuev, B. (2008) Effect of a solvent on the parameters of the analytical signal, detection limit, and analytical range of the determination in analytical thermal lens spectrometry. *J. Anal. Chem.*, **63**(8), 741.

Qian, H.B. and Howard, B.J. (1997) High resolution infrared spectroscopy and structure of CO-N_2O. *J. Mol. Spectrosc.*, **184**(1), 156.

Rabasovic, M.D., Nikolic, J.D., Markushev, D.D., and Jovanovic-Kurepa, J. (2008) Pulsed photoacoustic gas cell design for low pressure studies. *Opt. Mater.*, **30**(7), 1197.

Sakota, K., Kageura, Y., and Sekiya, H. (2008) Cooperativity of hydrogen-bonded networks in 7-azaindole$(CH_3OH)_n$(n=2,3) clusters evidenced by IR-UV ion-dip spectroscopy and natural bond orbital analysis. *J. Chem. Phys.*, **129**(5), 054303.

Schoch, A., Baumgarten, G., and Fiedler, J. (2008) Polar middle atmosphere temperature climatology from Rayleigh lidar measurements at ALOMAR (69° N). *Ann. Geophys.*, **26**(7), 1681.

Smith, W.E., Faulds, K., and Graham, D. (2006) Quantitative surface-enhanced resonance Raman spectroscopy for analysis. *Surf.-Enhanc. Raman Scatter.: Phys. Appl.*, **103**, 381.

Wehr, R., Kassi, S., Romanini, D., and Gianfrani, L. (2008) Optical feedback cavity-enhanced absorption spectroscopy for in situ measurements of the ratio $^{13}C : {}^{12}C$ in CO_2. *Appl. Phys. B-Lasers Optics*, **92**(3), 459.

White, J.U. (1942) Long optical paths of large aperture. *J. Opt. Soc. Am.*, **32**(5), 285.

Wirth, M.J. and Legg, M.A. (2007) Single-molecule probing of adsorption and diffusion on silica surfaces. *Annu. Rev. Phys. Chem.*, **58**, 489.

Wustholz, K.L., Sluss, D.R.B., Kahr, B., and Reid, P.J. (2008) Applications of single-molecule microscopy to problems in dyed composite materials. *Int. Rev. Phys. Chem.*, **27**(2), 167.

Ye, J.J., Pang, H.F., and Cheung, A.S.C. (2007) Optical-optical double resonance spectroscopy of YBr and YCl. *Chem. Phys. Lett.*, **442**(4–6), 251.

Zheltikov, A.M. (2000) Coherent anti-Stokes Raman scattering: from proof-of-the-principle experiments to femtosecond CARS and higher order wave-mixing generalizations. *J. Raman Spectrosc.*, **31**(8–9), 653.

Zink, L.R., Willcutt, A., Murphy, M., and Jackson, M. (2008) Frequencies of cw FIR laser lines for use in laser magnetic resonance spectroscopy. *Appl. Phys. B-Lasers Optics*, **92**(1), 5.

Further Reading

Andrews, D.L. (1997) *Lasers in Chemistry*, 3rd edn, Springer-Verlag, Berlin.

Andrews, D.L. and Demidov, A.A. (eds) (2002) *An Introduction to Laser Spectroscopy*, 2nd edn, Kluwer Academic/Plenum, New York.

Ellis, A.M., Feher, M., and Wright, T.G. (2005) *Electronic and Photoelectron Spectroscopy: Fundamentals and Case Studies*, Cambridge University Press, Cambridge.

Hollas, J.M. (1998) *High Resolution Spectroscopy*, 2nd, John Wiley & Sons, Ltd, Chichester.

Stephenson, S.K. and Saykally, R.J. (2005) Velocity modulation spectroscopy of ions. *Chem. Rev.*, **105**(9), 3220.

17
Photoacoustic Spectroscopy

Frans J. M. Harren, Stefan Persijn, and Simona M. Cristescu

17.1 Introduction 583
17.2 Devices and Equipment 585
17.2.1 Light Sources 585
17.2.2 Gas-Phase Photoacoustic Cells 587
17.2.2.1 Nonresonant Cells 587
17.2.2.2 Resonant Cells 589
17.2.2.3 Cantilever-Type Optical Microphone 592
17.2.2.4 Quartz Tuning Fork 593
17.2.3 Condensed-Matter Photoacoustic Cells 594
17.2.4 Photothermal Detection 594
17.2.4.1 Photothermal Deflection Spectroscopy 594
17.2.4.2 Thermal Lensing 595
17.2.5 Fourier-Transformed Infrared Photoacoustic Spectroscopy 597
17.3 Application 597
17.3.1 Photoacoustic Spectroscopy in the Condensed Phase 598
17.3.1.1 Solid-State Spectroscopy 599
17.3.1.2 Depth Profiling 600
17.3.1.3 Thermal Diffusivity 600
17.3.1.4 Phase Transitions 601
17.3.1.5 Traces in Solution 601
17.3.1.6 Chemical Reactions 602
17.3.1.7 Biological Processes in Solution 603
17.3.2 Photoacoustic Spectroscopy in the Gas Phase 604
17.3.2.1 Overtone Studies 604
17.3.2.2 Relaxation 604
17.3.3 Trace Gas Detection 607
17.3.3.1 Environmental Applications 609
17.3.3.2 Plant Physiology 610
17.3.3.3 Oxidative Stress 611
17.3.3.4 Microbiology 615
17.3.3.5 Entomology 616

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

17.3.3.6 Human Health Research 618
Acknowledgment 618
Glossary 618
References 620
Further Reading 622

17.1
Introduction

Photoacoustic (PA) spectroscopy is an indirect technique in the sense that an effect of light absorption is measured rather than light absorption itself. Light that is absorbed by a sample is, completely or partly, converted into heat, giving rise to pressure waves that can be detected. Using PA spectroscopy (PAs), regular absorption spectra can be obtained. PA spectroscopy is complementary to regular absorption/transmission spectroscopy, because only absorbed light contributes to the PA signal and drawbacks of regular spectroscopy (e.g., background absorption, scattered light, opaque media) do not play a role in PA spectroscopy. In general, when matter absorbs light the transmitted power follows Beer's law: $P = P_0 e^{-\alpha l}$ in which P_0 the incident laser power, P the transmitted laser power after distance l in centimeters through the sample, and α in cm^{-1} the absorption strength at that frequency. For low absorptions, this exponential expression can be linearized:

$$\Delta P = P_0 - P = P_0 \alpha l \tag{17.1}$$

It shows that the generated PA effect (i.e., decrease in power) is proportional to the incident power. On the other hand, other factors, such as the optical, thermal, and acoustical properties of the sample and media involved, are now more important and can even be the direct objective of study in many applications of PAs.

Like in all spectroscopy, the spectral dependence of absorption permits selectivity and identification of the absorbing components. The PA response itself can be fast, highly sensitive, and in principle, background-free. Thanks to these qualifications, no lengthy accumulation periods (e.g., in trace gas detection) or long absorption pathways are needed. Because transparency and scattering play no role, solids, whether crystalline, powdered, or amorphous, can be studied. Using the dependency of the PA signal on thermal and acoustic properties, surfaces and solids can be tested noninvasively for structural inhomogeneity. For these latter applications, the spectroscopical aspect plays a minor role. PA spectroscopy is especially suited to monitor local effects. Remote sensing, conversely, is alien to PAs.

A medium that absorbs electromagnetic radiation experiences a local temperature increase by relaxation processes that transfer electronic, vibrational, or rotational energy into translation energy (phonon energy in case of solids). This nonradiative relaxation process occurs if its relaxation time can compete with the radiative

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

lifetime of the excited energy levels. If the absorption is modulated, at an acoustic frequency, for example, by modulating the intensity of the radiation source, these temperature changes occur periodically and give rise to acoustic phenomena. Hence the name ''photoacoustic''; light absorption is detected through its accompanying acoustic effect.

A sensitive microphone can detect the effect, especially, when it occurs in the gas phase. In the case of solids or fluids, a piezoelectric transducer is often used. Thus, the minimum requirement is a suitable source of modulated electromagnetic radiation for which the wavelength can be selected, such that the absorption spectrum of the substance investigated can be accessed. In addition, a detection device such as the mentioned microphone or piezo transducer must indicate that absorption takes place (Figure 17.1). This short description already clarifies that PA spectroscopy is based on a two-step process; first, electromagnetic radiation energy must be absorbed, this being the spectroscopically selective step; second, the absorbed energy must be transferred to sound-producing degrees of freedom, by relaxation processes. If this relaxation is too slow with respect to the employed modulation times, PA spectroscopy will not work. Note that the relaxation rate depends on the properties of the participating molecules.

The instrumentation required for PA spectroscopy can be relatively inexpensive and simple; the powerful and tunable laser that may be used forms the cost-determining part, but for some applications a simple mercury lamp and an inexpensive microphone suffice. However, the instrumentation has to be well designed and cleverly chosen, and its operation requires skillful people. Especially, the PA cell has to be designed with care, fitting with the chosen application.

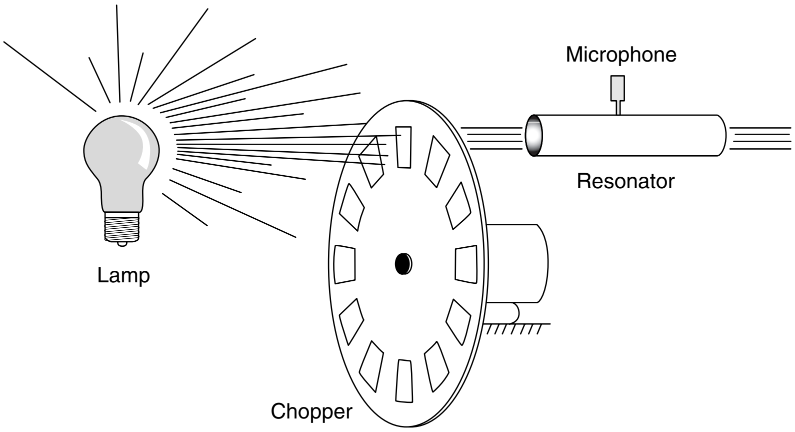

Fig. 17.1 A schematic representation of the photoacoustic setup. The lamp represents the light source, which was the sun in the original experiments by Bell. Nowadays lasers are used because of their tunability and brilliance. When, for example, a gas absorbs the light, the temperature of the gas increases, the gas expands adiabatically, generating a pressure increase. By switching off the light, temperature and pressure drop. The light is modulated, in this case mechanically chopped to generate an AC temperature and/or pressure wave (i.e., sound). If the modulation frequency corresponds to the resonance frequency of the acoustical resonator, the sound is amplified. A microphone is inserted into the resonator at the point where the resulting pressure standing wave has maximum amplitude.

The PA effect has been discovered and clarified almost simultaneously by Bell (1880), Röntgen (1881), and Tyndall (1881). In those days, neither laser sources (which are ideally suited for PA spectroscopy, owing to their brilliance) nor sensitive small microphones were available. Therefore, being well understood, but rather impractical, the effect remained slumbering in the cabinet of curiosities. In 1938, Veingerov refined the PA technique for the first spectroscopic gas analysis (Veingerov, 1938); and in 1943, Luft measured trace gas absorption spectra with an infrared broadband light source down to the part per million level (Luft, 1943). By the end of the 1960s, after the invention of the laser, the scientific interest expanded again. Kerr and Atwood (1968) utilized the laser PA detection to obtain the absorption spectra of small gaseous molecules. Owing to the high spectral brightness of lasers and improved phase-sensitive lock-in techniques to amplify the acoustic signal, they were able to determine low absorption strengths. Since then, a broad field of pure and applied spectroscopic methods has emerged using the PA effect. The last decades have seen an ever-increasing use of PA techniques in a very wide field of applications. These activities range from fundamental research (e.g., the investigation of vibrationally highly excited molecules) to applications (e.g., testing of solids, surfaces, biomedical applications). A number of useful reviews have appeared over this period (see Further Reading), covering most of these areas.

Firstly, the equipment is discussed. Various types of PA cells are introduced, which are resonant and nonresonant with respect to the modulation frequency of the light. In addition to the sound detection by microphones, related techniques such as thermal lensing and photothermal deflection are discussed, which are often applied if liquid- or solid-state samples are investigated. In Section 17.3, applications to solids, liquids, and surfaces and to gas spectroscopy are dealt with. As one of the more recent applications, trace gas detection is also considered with environmental and biological studies. Besides these key references, books and review articles are given for further reading.

17.2
Devices and Equipment

17.2.1
Light Sources

In principle, there is no need for lasers to operate a PA detection system. Light from lamps can be filtered by a monochromatic system to obtain radiation needed for spectroscopic discrimination to distinguish between components. Other highly sophisticated instruments use, for the selection of wavelength, the gaseous species under investigation itself (Figure 17.2); all wavelengths at which absorption takes place are simultaneously active. With the latter instrument (URAS, ABB Germany) detection limits in the sub-ppmv (1 ppmv = 1 part per million volume = $1 : 10^6$) range can be achieved for more than 25 gaseous compounds.

However, more often PA detection goes together with the application of laser light sources, ranging from the infrared to the visible and UV, including both continuous wave (CW) and pulsed lasers. Since the PA signal is proportional to the laser power, strong lasers are advantageously applied.

In the mid-infrared wavelength region, CW CO_2 (wavelength region from 9 to 11 μm) and CO lasers (4.6–8.2 μm) were

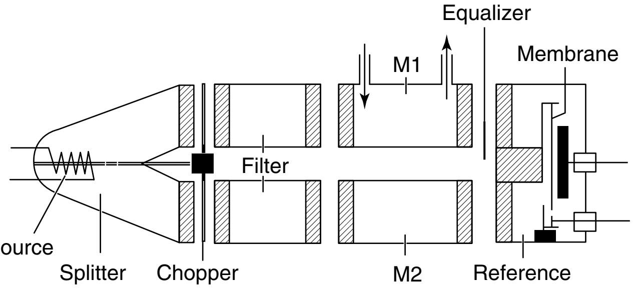

Fig. 17.2 The infrared absorption spectrometer setup (URAS). Light from the thermal IR source is divided into two paths. The chopper modulates the intensity for both paths. The filters serve to filter out light of wavelengths not needed for the detection process; they can be filled with gases whose absorption spectra do not overlap with those of the species under scrutiny. M1 represents the sampling cell through which the sampled gas is flowed. M2 serves as a reference cell with a calibrated gas mixture in which the traversing light is attenuated to equalize the light intensities at the entrance of the last cell at the far right. The light is equalized before the cells M1 and M2 are filled. The last cell consists of two compartments with a membrane in between. Both compartments are filled with the gas under scrutiny so that all wavelength characteristics for this gas contribute to the signal. If in M1 the attenuation differs from that in the reference cell, the membrane starts to oscillate with the frequency of the chopper. This oscillation (typically hertz) is detected capacitively.

among the most used. At present, developments in solid state physics and nonlinear optics have expanded the potential for PA spectroscopy to other wavelength regions. The recent availability of high power, pulsed, and CW quantum cascade lasers (QCL)s, operating at room temperature with power levels up to 1 W, has boosted PA spectroscopy in the mid-infrared wavelength region. Even more, single-mode tuning ranges, approaching 300 wavenumbers, are now achievable from commercially available external cavity QCL products. QCLs operate in the mid-infrared wavelength region from 3.5 to 24 μm (Kosterev and Tittel, 2002). For the 2.5–5 μm region CW periodically poled lithium niobate optical parametric oscillators (PPLN–OPOs) became available at high (several watts) power level and narrow linewidth. Nowadays, OPOs are pumped by high power, near-infrared lasers with excellent spectroscopic properties (narrow linewidth, Gaussian beam profile, fast wavelength scanning), which they can convert into the mid-infrared wavelength region due to their instantaneous nonlinear effects (Ebrahim-Zadeh and Sorokina, 2005).

In the near infrared wavelength region, CW titanium–sapphire ring lasers (650–1000 nm) are intensively used for high-resolution studies of weak overtone absorption spectra of molecular gases. High-overtone spectroscopy profits from sensitive PA detection, since one deals with high vibrational excitation, where fluorescence-based techniques do not work. Electronic spectroscopy with light sources in the visible or in the ultraviolet region does not immediately ask for PA detection because other techniques provide more powerful means (e.g., laser-induced fluorescence (LIF) or resonance-enhanced multiphoton ionization).

In addition to this, near-infrared “telecom” diode lasers have the advantage of small size, reliability, and low costs. Because of their low power the combination with PA spectroscopy is not always very favorable. However, the development of fiber amplifiers in this wavelength region could overcome this disadvantage. Telecom lasers are well suited to be used in gas sensing instruments in which the sensitivity is limited to the ppmv range. These lasers are tunable only over a limited wavelength region (1 cm^{-1}), but larger tuning ranges can be obtained by external cavity setups (external cavity diode laser). The near-infrared diode lasers have the main disadvantage that they operate at the weak overtone absorption bands instead of the strong fundamental vibrational absorption bands in the mid-infrared wavelength region.

By using pulsed lasers and gated detection, the pulse character of the signal allows suppression of a possible window signal by setting the boxcar delay such that the window signal has not yet arrived at the piezo transducer (or microphone) when the trace signal is detected. In general, however, the highest sensitivity of PA detection has been achieved for modulated CW lasers; if no other arguments prevail, the use of CW lasers is recommended.

17.2.2
Gas-Phase Photoacoustic Cells

The proper choice of the PA cell geometry depends on the intended application. For gas-phase measurements, mainly resonant cells are combined with frequency- or amplitude-modulated CW lasers. For condensed-phase spectroscopy (e.g., surfaces, thin surface layers, or depth profiling) with CW lasers, gas-coupled microphones are used within a small nonresonant cell. Utilizing pulsed lasers, the response is monitored with a piezoelectrical crystal directly coupled to the material. In competition with this is the noncontact detection by the photothermal effect, in which a laser beam is deflected or dispersed by the heat gradient from the sample.

17.2.2.1 **Nonresonant Cells**

If the modulation frequency of the incident light is much smaller than the lowest resonance frequency of the cell, it is to be considered a nonresonant cell. The performance of a cell is related to its efficiency in converting absorbed photon energy into acoustic energy by a cell constant *F;* that is, $p_{\text{gas}} = F\Delta P$ with p_{gas} being the generated acoustic pressure in the gas and ΔP derived from Equation 17.1 The cell constant F in (Pa m W^{-1}) can be derived by

$$F_{\text{nonresonant}} = \frac{G(\gamma - 1)l_{\text{nr}}}{2\pi \nu V_{\text{nr}}} \tag{17.2}$$

with l_{nr} and V_{nr} being the length and the volume of the cell, respectively; γ the specific heat constant, ν the modulation frequency, and G a geometrical factor, typically of the order of 1. From the above formula, one can derive that the F is independent of the cell length. The diameter should be as small as possible to generate high signals, just large enough to allow proper alignment of the incident light beam. To minimize the detection of absorption by the cell windows, one has to maximize the cell length; the energy absorbed by the windows is then distributed over a larger volume. The PA signal decreases with higher modulation frequencies; therefore, it is advantageous to use low modulation frequencies.

For the investigation of solids, liquids, and surfaces, another type of nonresonant PA cell has to be designed. The sample has to be sealed in a gas cell having a window

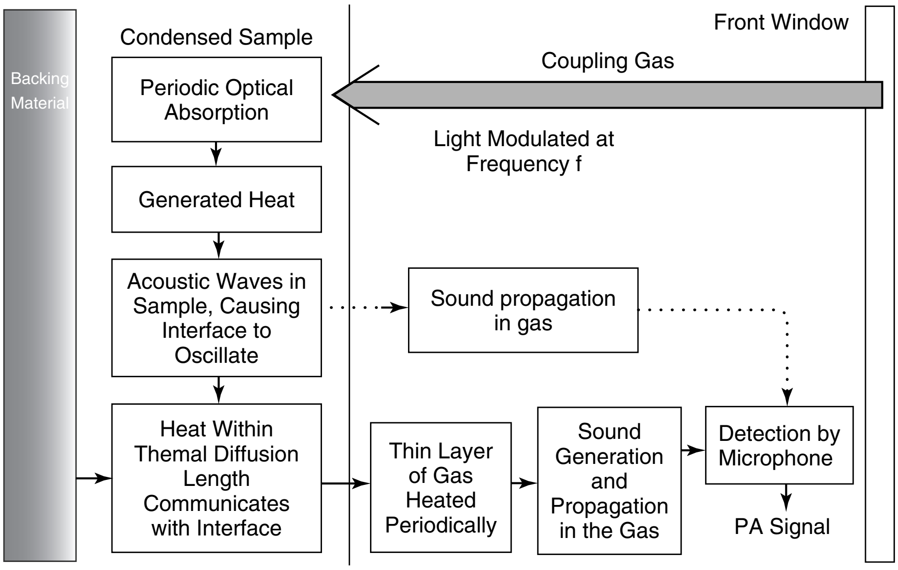
Backing Material
Condensed Sample
Periodic Optical Absorption
Generated Heat
Acoustic Waves in Sample, Causing Interface to Oscillate
Heat Within Themal Diffusion Length Communicates with Interface
Front Window
Coupling Gas
Light Modulated at Frequency f
Sound propagation in gas
Thin Layer of Gas Heated Periodically
Sound Generation and Propagation in the Gas
Detection by Microphone
PA Signal

Fig. 17.3 Block diagram to explain photoacoustic detection of absorption in condensed matter using gas coupling. (Figure adapted from Tam, 1983.)

and a microphone. In the extreme case of an open cell, the wall opposite the window can be the material under investigation. The gas buffer volume in this cell should be designed as small as possible; in some cases, the heat generated by the absorption of (modulated) light diffuses to the surface and into the buffer gas (e.g., helium). The probing depth into the surface is determined by the modulation frequency: at low frequencies deeper layers of the material are investigated as compared to high frequencies. This probing depth, the thermal diffusion length, μ_s is given by

$$\mu_\mathrm{s} = \left\{\frac{D}{\pi \nu}\right\}^{1/2} \tag{17.3}$$

where D is the thermal diffusivity of the sample and ν the modulation frequency of the light. To characterize the complete sample, the optical absorption length μ_α (the reciprocal of the optical absorption coefficient α) and the thickness of the sample L are also important (Figure 17.3). Because the values of all three properties (μ_s, μ_α, and L) can vary relative to each other and within orders of magnitude, six possible limiting cases can be distinguished, as elaborated by Tam (1983).

For pulsed operation, the bandwidth of the microphone is too small (typically 20 kHz) as compared to the fast nanosecond processes in materials. The microphone must then be replaced by a piezoelectric element, either ceramic or a polyvinylidene difluoride (PVDF) foil, which is directly attached to the material. These detectors can show an acoustic response up to tens of megahertz. High-output pulsed lasers combined with gated extraction of the acoustic signals yield a highly sensitive

($\alpha = 1 \times 10^{-7}\,\mathrm{cm}^{-1}$) technique that can discriminate against low-frequency noise and window absorption.

17.2.2.2 **Resonant Cells**

In a resonant cell setting, the modulation frequency is chosen to match one of the resonant frequencies of the cell. This produces a standing, amplified, sound wave in the resonator. For a gas-type resonant cell, two varieties of acoustical resonators have found widespread use: the Helmholtz resonator and the cylindrical resonator excited in a longitudinal, radial, and/or azimuthal mode (Figure 17.4).

A Helmholtz resonator consists of a closed volume (cavity) connected via a long narrow tube to the open air. It is the acoustic equivalent of a mechanical oscillator composed of a mass (long tube) and a spring (cavity). The oscillation frequency (ν_H) of the system depends on the length (l_H) and cross section (A_H) of the tube and the volume (V_H) of the cavity via the following relation:

$$\nu_\mathrm{H} = V_\mathrm{s} \left\{ \frac{A_\mathrm{H}}{V_\mathrm{H} l_\mathrm{H}} \right\}^{1/2} \tag{17.4}$$

with V_s being the velocity of sound (Kapitanov *et al.*, 2001). Although the acoustical response of such a resonator is somewhat lower as compared to the response of the cylindrical resonators described below, the type of resonator has some advantages when low or high temperatures are needed. Microphones have a limited operating temperature range, that is, around room temperature. When gases are investigated at low (e.g., 80 K) or high (350 K) temperatures, it is advantageous to have a homogeneous temperature distribution across the gas volume while the microphone operates at room temperature. The temperature gradient is over the long narrow tube of the resonator at the end of the microphone is placed.

For cylindrical resonant cells, one has to multiply the cell constant F (Equation 17.2) with the quality factor Q of the generated acoustic resonance:

$$F_\mathrm{resonant} = \frac{G(\gamma - 1) l_\mathrm{r} Q}{2\pi \nu V_\mathrm{r}} \tag{17.5}$$

with l_r and V_r being the length and the volume of the acoustic resonator, respectively; Q equals the ratio of the energy stored in the acoustical standing wave to the energy losses per cycle. For resonant cylindrical cells, one can derive for the cell constant: $F_\mathrm{resonant} \propto \sqrt{l_\mathrm{r}}/R$, with R being the radius of the resonator (Equation 17.5.) (Bijnen, Reuss, and Harren, 1996). Cell geometries with large diameter-to-length ratios, to excite the resonance in the radial or azimuthal acoustic mode, possess high Q-values and high resonance frequencies; they have, however, low F-values. PA cells with high Q-values are sensitive to long-term drifts, for example, owing to thermal expansion if the temperature is not carefully controlled. Such cells require active locking of the modulation frequency on the resonance frequency of the cell. In a longitudinally excited resonator, a smaller acoustic gain, as a consequence of a relatively low Q-value, is compensated for by the signal gain due to the smaller diameter.

The resonant acoustical amplification is limited by various dissipation processes, which can be divided into surface and volume effects. Main surface losses are viscous and thermal losses at the resonator surface; in addition there are microphone losses, and acoustic wave scattering losses at obstacles inside the cell. The less-important volumetric losses are radiation losses, free-space viscous and thermal losses, and dissipative relaxation losses within the absorbing gases. For a cylindrical resonator, the resonant frequencies are given by Hess (1983)

$$\nu_{\mathrm{nmp}} = V_{\mathrm{s}}\left[\left\{\frac{\alpha_{\mathrm{mn}}}{d}\right\}^2 + \left\{\frac{p}{2l}\right\}^2\right]^{1/2} \qquad (17.6)$$

with V_{s} being the velocity of sound of the gas inside the cavity, d the cell diameter, l the axial length, and $p = 0, 1, 2, 3 \ldots$ axial mode numbers; α_{mn} is the nth root of the derivative of the Bessel function, $dJ_m/dr = 0$ at $r = R$ (e.g., $\alpha_{00} = 0, \alpha_{01} = 1.2197$, $\alpha_{10} = 0.5861$, and $\alpha_{11} = 1.6970$).

To choose the proper frequency for an optimal signal-to-noise ratio, one has to take into account the noise spectrum. The theoretical minimum acoustical noise results from random pressure fluctuations and their frequency distribution due to Brownian motion. The total power of these pressure fluctuations is constant, but the noise spectrum depends on the Q-value of the acoustical resonance of the PA cell; a high Q-value will significantly amplify noise power at these resonances. However, often this Brownian noise level is below what is produced by other noise sources such as electronic amplifier noise and external acoustical noise from the surroundings; their power decreases with increasing frequency. It is then advantageous to utilize a resonant cell to improve the signal-to-noise ratio. The resonance frequency might be chosen such that contributions from Brownian noise and amplifier/external noise become comparable. At resonance, external acoustical noise within the cell will be amplified at a high Q-value. Here acoustical shielding helps, that is, a proper cell wall construction (massive), material choice (brass), and good design of in- and outlet ports are required. For the latter, $\lambda/4$ notch filters (λ the acoustical wavelength) are positioned in line with the gas in- and outlet (Figure 17.5). In practice, for the B&K 4179 condenser microphone mounted on a well-shielded resonant ($Q = 40$) PA cell, a noise level has been observed that is an order of magnitude above the noise of the complete system (microphone and pre-amplifier) quoted by the manufacturer (Bijnen, Reuss, and Harren, 1996).

By definition, noise components do not possess a fixed phase relation with the periodical modulation of the light intensity. External perturbing sources may have a fixed phase relation, for example, the directly generated acoustical sound caused by chopping the light beam. They must be minimized in the same way as the external acoustical noise. In the case of a mechanical chopper, a good choice of the chopper position and removal of objects in the neighborhood of the chopper wheel are advantageous. A more serious problem

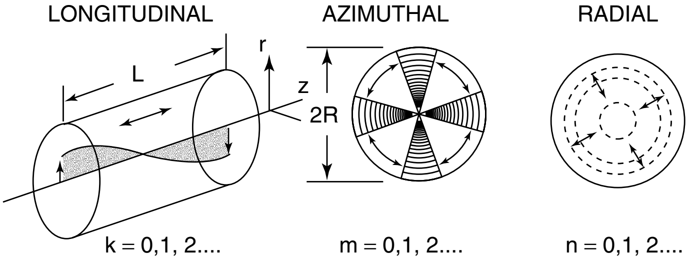

Fig. 17.4 Resonant acoustic modes of a cylindrical closed chamber, the fundamental longitudinal, azimuthal, and radial modes.

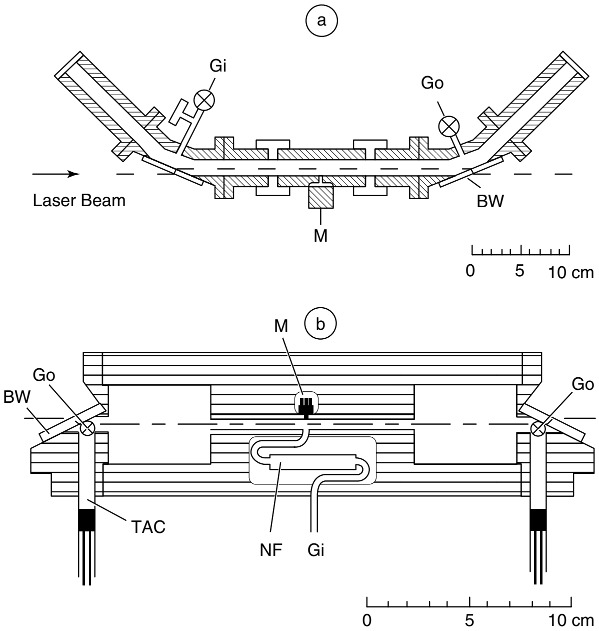

Fig. 17.5 Two photoacoustic cell types successfully used in trace gas detection. (a) The Zurich-banana cell, where the laser beam enters from the left through a Brewster window (BW) to leave the central part ($l = \lambda/2$) of the resonator at the right Brewster window. The position of the microphone M is indicated. The two pieces of length $\lambda/4$ before and after the bends serve to suppress the window signal. The total length of the banana cell amounts to λ. In the bends, the gas inlet (Gi) and gas outlet (Go) are shown. (b) The Harvard–Nijmegen open-organ-pipe cell; the main part is manufactured of a block of massive brass. The central organ pipe acts as an open resonator with length $\lambda/2$; the copper resonator tube has a highly polished, gold-coated inner surface to minimize wall heating by stray light. The central position of the gas inlet (Gi) is essential to obtain short measuring times; only the resonator volume of 2.8 ml must be replenished before an independent concentration measurement can be performed. TACs: $\lambda/4$ tunable air columns to suppress window signal. NF: $\lambda/4$ notch filter to suppress acoustic in coupling of noise via gas inlet.

is formed by the PA background signal; absorption of the light beam in the window material or light scattered/reflected from the resonator wall causes a signal at the frequency of the modulated light beam. In resonant cells, using $\lambda/4$-tubes next to the windows can diminish window signals. These tubes, placed perpendicular to the resonator axis close to the windows, are tuned to the resonator frequency and act as interference filters for the window signals (see Figure 17.5). Influence of scattered light on the PA background signal can be minimized by the use of highly reflecting polished material with a good thermal conductivity. In the case of infrared light a polished gold-coated copper wall of the acoustical resonator has been found to work satisfactorily.

We describe two successfully applied cells, the Zurich-banana-type cell and the Harvard–Nijmegen open-organ-pipe cell (Figure 17.5). In both cases, the modulation frequency ν is chosen in resonance with a longitudinal acoustic mode of the cell (typical length 150 mm and typical diameter

Tab. 17.1 Photoacoustic cell parameters; in the case of the open-organ-pipe cells in which the window signal is so far suppressed as to no longer limit the sensitivity, α_{min} is given for 1 W of laser power, that is, for 10 W of laser power α_{min} becomes smaller by a factor 10.

Cell type	Banana	Organ pipe
Microphone	Bruel and Kjaer	Knowless
Type number	4179	EK 3024
Microphone sensitivity (mV Pa^{-1})	1000	20
Microphone noise (μV $Hz^{-1/2}$)	5	0.5
Number of microphones	1	3
Cell constant F (Pa m W^{-1})	16.4	45.0
Q-value	43	32
Resonance frequency (Hz)	555	1600
Volume diameter ratio (ml mm^{-1})	80/15	3/6
α_{min} (cm^{-1} W^{-1})	5×10^{-8}	2×10^{-8}

10 mm). The banana-type cell has as an important advantage its small volume; it forms an acoustically closed system with a moderately high Q ($Q = 53$ for its utilized second longitudinal mode); the "bent" geometry allows Brewster windows (BWs) to be positioned at pressure nodes, thereby minimizing the window signals (similarly, the gas in- and outlet with their external acoustical noise). The cell is operated at 555 Hz. The end pieces beyond the bends are optimized to suppress the window signal by interference. The open-organ-pipe cell may have a somewhat lower Q-value ($Q = 32$, inner radius $R = 3$ mm) but a higher F-value due to the smaller radius of the resonator tube ($F_{\text{resonant}} \propto \sqrt{l_r}/R$). Table 17.1 shows the properties of the cells.

17.2.2.3 Cantilever-Type Optical Microphone

Microphones, where the flexible membrane is strained over a frame or the pressure sensor is a thinner portion in a metal plate, are not very sensitive and the response is not linear because of the fact that the material has to stretch out radially under the pressure variations. A cantilever-type pressure sensor has been proposed as shown in Figures 17.6 and 17.7 (Uotila, Koskinen, and Kauppinen, 2005). The cantilever size is 2 mm × 4 mm and made out of silicon. Because of its extreme thinness (5 μm) the cantilever moves like a flexible door owing to the pressure variations in the surrounding gas. The frame around the silicon cantilever is thick (380 μm) and there is a narrow (30 μm) gap between the frame and the cantilever on three sides. As the pressure varies, the cantilever only bends but it does not stretch. Therefore, the movement of the free end of the cantilever can be about 2 orders of magnitude greater than the movement of the middle point of the tightened membrane under the same pressure variation. With this setup a detection limit of 0.9 ppmv for ethylene, 0.3 ppmv for methane, and 0.1 ppmv for carbon dioxide was achieved with a mechanically chopped blackbody radiator.

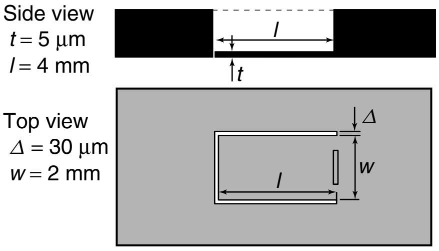

Fig. 17.6 The dimensions of the cantilever and the gap between the cantilever and the frame.

17.2.2.4 **Quartz Tuning Fork**

In the gas-phase spectroscopy, the common approach is to use a resonant PA cell and accumulate the absorbed energy in the gas. However, a recent developed approach is to store the absorbed energy in a sensitive element. A well-suited material for this is a high-*Q* crystal quartz element with piezoelectric properties. These elements are mass produced and inexpensive; every electronic watch or clock is built around a high-*Q* quartz crystal frequency standard. Usually, it is a quartz tuning fork (TF) with a resonant frequency close to 32 768 Hz. The mode at this frequency corresponds to a symmetric vibration (the prongs move in opposite directions). The antisymmetric vibration is piezoelectrically inactive. These quartz TFs have recently become widely used materials for atomic-force and optical near-field microscopy, and therefore their properties have been carefully analyzed (Kosterev *et al.*, 2005). A typical watch TF has a *Q* value of 20 000 in vacuum; at normal atmospheric pressure this is lower (~8000) because of the viscous properties of air. The typical energy accumulation time at atmospheric pressure is 250 ms, which is quite longer than any practical gas-filled resonator (Figure 17.8).

The special features of using a TF are as follows: (i) Ambient acoustic noise is very low above 10 kHz. (ii) External sound will always excite both prongs of the TF and, as a result, will not excite the piezoelectrically active mode, in which the two prongs move

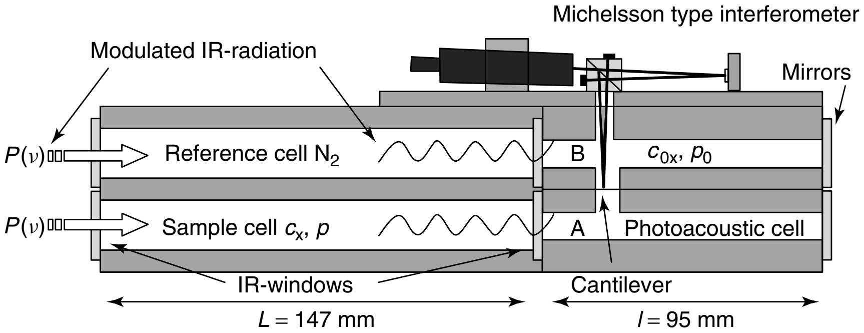

Fig. 17.7 The setup containing the cantilever is in principle the same setup as the nondispersive setup from Figure 17.2, but now with a cantilever detection setup. It consists of three cells: the reference cell, sample cell, and photoacoustic cell. The photoacoustic cell acts as a detector and includes high concentration of the gas to be detected. The cell is divided into two parts A and B which are separated by the cantilever and its frame. The position of the cantilever is proportional to the pressure difference between the cell parts A and B. The displacement of the cantilever end is measured continuously with a Michelson type interferometer.

in opposite directions. (iii) The width of the TF resonance at normal pressure is 4 Hz, and only frequency components in this narrow spectral band can produce efficient excitation of the TF vibration. (iv) The gas sampling volume is extremely small ($0.15\,\text{mm}^3$).

17.2.3
Condensed-Matter Photoacoustic Cells

Most spectroscopy on solid samples is done with gas-coupled microphones. Little sample preparation is needed and PA cells are commercially available. Modulation frequencies are generally low, 2–500 Hz, and coupling volumes are as small as possible. One increasingly popular design is the open photoacoustic cell (OPC) in which the sample makes up one wall of the cell.

The investigation of condensed matter using gas coupling becomes more complex when the sample absorption coefficient is small and the modulation frequency is high (e.g., pulsed lasers). The sample vibration itself will start to contribute in a significant way to the microphone signal. In addition, gas microphone detection of acoustic waves inside a condensed sample is not an efficient way of detection, because the acoustic impedance mismatch at the sample–surface interface causes a low transmission across the interface (Tam, 1983). A better, direct coupling, method involves the insertion or attachment of a transducer (piezoelectric) into or onto the sample.

When CW lasers are used in combination with condensed-matter PA cells, wavelength tuning allows a considerable reduction of the background signal generated by light scattering and window absorption. Factors that limit sensitive detection are as follows: besides light scattering and window absorption, mechanical noise at low frequencies, heating of the sample (self-defocusing), and convection currents in the case of liquid.

These limitations are avoided with pulsed lasers. Pulsed lasers allow time gating to discriminate against window absorption (arrives later) and light scattering (is instantaneous). Heating of the sample is minimized by the low average power (pulses are at nanosecond timescale) in combination with low repetition rates. Mechanical noise and the electrical pickup can be eliminated because the pulsed PA signals are at high acoustic frequencies ($>100\,\text{kHz}$).

Piezoelectric ceramics are (much more) suitable and many types are commercially available, for example, lead zirconate titanate (PZT), lead metaniobate, lithium niobate, crystalline quartz, and so on (Mason and Thurston, 1979). Thin, highly insulating polymeric films poled at elevated temperatures in strong electric fields also show a piezoelectric character: PVDF, Teflon, and Mylar are most commonly used. They combine a fast rise time (2 ns) with good flexibility and good acoustic impedance to liquids such as water (Tam, 1986).

17.2.4
Photothermal Detection

17.2.4.1 Photothermal Deflection Spectroscopy

Schemes different to PA detection are considered by Sigrist (1994). His conclusion is that for many applications, the PA technique utilizing a sensitive microphone appears to yield the simplest and the most sensitive solution. There are instances, however, in which PA fails. For the investigation of condensed matter using a microphone, a gas buffer is needed that results in low signals due to the acoustical mismatch between gas and solids. In addition, for fast (nanoseconds) experiments microphones are too limited due to their bandwidth. Piezoelectric transducers do have wider bandwidths but need to be attached to the condensed sample. When noncontact methods are preferred, techniques such as photothermal deflection and thermal lensing can be used. An additional advantage is that the bandwidth of detection is very wide and only limited by the bandwidth of the photodiodes.

In photothermal deflection spectroscopy, the basic principle rests on the change of refraction index n caused by a local temperature increase and a density decrease upon absorption of radiation. This change Δn results in deflection of a probe laser (in many cases a Helium–Neon laser with a good pointing stability), observed with the help of a position-sensitive detector. The basic idea can be implemented by two different geometries; the two lasers (absorption laser and probe laser) can either be aligned collinearly or perpendicularly (Figure 17.9). The advantages of the perpendicular arrangement are ease of reproducible adjustment and local probing since it defines a probing volume of some cubic millimeters. This small size is important in many applications such as local gas emission or local heating of the solid-state material for (sub)surface imaging. In the latter application, the excitation light is directed perpendicular to the surface, while the probe laser is parallel, and very close, to the surface.

17.2.4.2 **Thermal Lensing**

Like other photothermal phenomena, thermal lensing is based on the temperature increase after illumination of the sample and the nonradiative relaxation of the absorbed energy. In this case, a laser beam is needed because its radially symmetric Gaussian intensity distribution generates more heat at the center of the

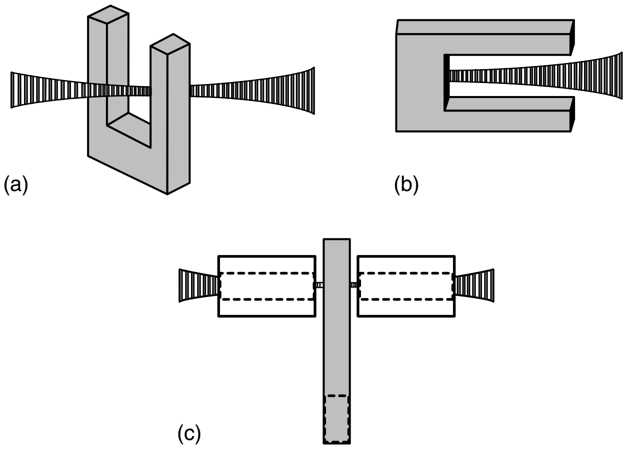

Fig. 17.8 Optical configurations for photoacoustic signal detection with a TF. (a) The laser beam is perpendicular to the TF plane. (b) The laser beam is in the TF plane. (c) An acoustic resonator (sound tube) is added to enhance the signal. The laser beam is directed through the tube. The pressure anti-node is at the center at the location of the tuning fork between the two pieces of the tube.

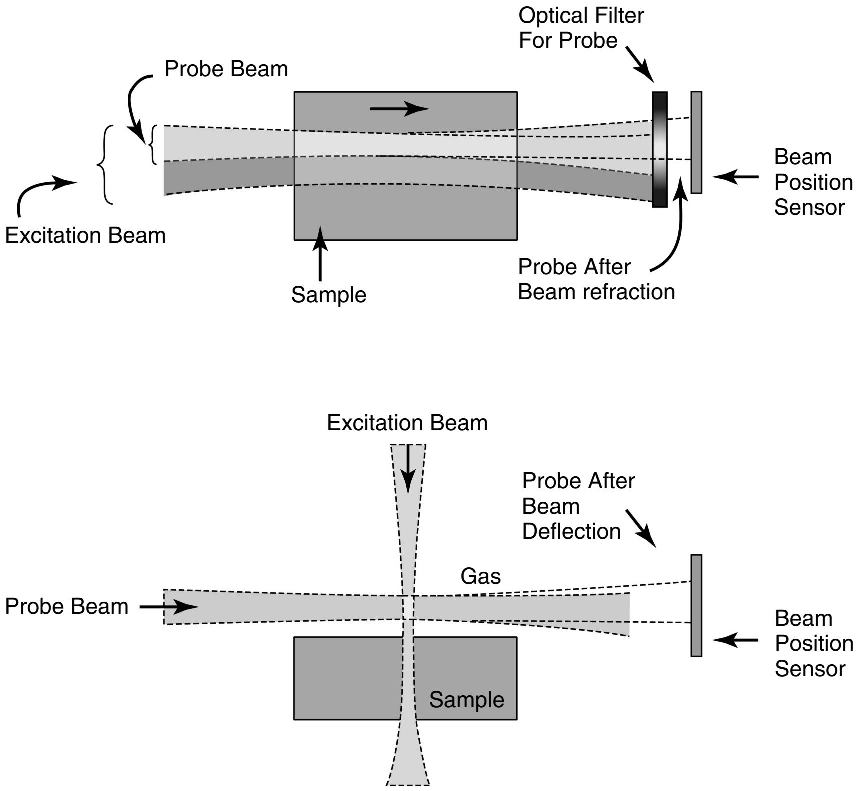
Optical Filter
For Probe
Probe Beam
Beam
Position
Sensor
Excitation Beam
Sample
Probe After
Beam refraction
Excitation Beam
Probe After
Beam
Deflection
Gas
Probe Beam
Beam
Position
Sensor
Sample

Fig. 17.9 Two schematic configurations of the photothermal beam deflection in which the sample absorbs the radiation of the pump laser and the probe laser (in many cases a He–Ne laser) is deflected by the refractive index gradient induced by the generated heat. (a) Parallel setup and (b) transverse setup.

beam than in the wings (Franko and Tran, 1996). Owing to this temperature gradient across the beam, a lens-like optical element is formed in the material. This thermal lens effect can be observed by a second laser beam or by the initial laser beam itself (single-beam thermal lensing). Single-beam instruments are relatively frequently used because of their simplicity and ease of operation compared to dual-beam instruments (see Figure 17.10). In such an instrument, the laser beam is focused with a lens and modulated by a chopper. After passing through the sample, the beam intensity is usually measured in the far field with a photodiode placed behind a pinhole. The thermal lens signal is linearly dependent on the energy of the pulsed laser or the power of the CW laser. The signal is also strongly dependent on the size of the pump beam; it is therefore necessary to focus the laser beam to obtain higher power densities and thus higher sensitivities. For a two-beam setup, a collinear configuration (two parallel laser beams) is more favorable as compared to a transverse configuration (beams crossing at a 90° angle) because of its higher sensitivity, although the second setup is sometimes preferable because of its small sample volume and independence of the sample length. Sampling volumes as small as a few microliters and absorbances as low as 10^{-7} have been detected and characterized by this technique.

17.2.5
Fourier-Transformed Infrared Photoacoustic Spectroscopy

Infrared lasers have mostly limited wavelength coverage. This restriction is circumvented if Fourier-transform techniques are combined with PA detection. As compared to dispersive techniques using filters or monochromators, Fourier-transformed infrared (FTIR) spectroscopy has the advantages of a relatively high spectral throughput (important since the PA signal is proportional to the light intensity), high spectral resolution, wide spectral range, and speed (Jiang, 2002).

The advantage of combining PA detection with a scanning FTIR derives from the signal being background-free; there is no contribution to the FTIR interferogram from IR frequencies where the sample does not absorb. In FTIR, the modulation frequency ν in Hz depends on the wavelength of the light λ and the velocity of the mirror V in the interferometer: $\nu = 2\,V/\lambda$. A drawback in using PAs is that now the generated PA signal depends on the wavelength of the light: it weakens at shorter wavelengths as the modulation frequency increases. An improvement is introduced by stepwise scanning of the interferometer. The PA signal is generated by an external intensity modulator (e.g., chopper) while the interferometer is scanned stepwise rather than continuously. The advantage of stepscan FTIR is that the strength of the PA signal only depends on the frequency of the external modulator. In addition, now relevant phase information can also be collected. In normal FTIR, short wavelengths will generate high Fourier frequencies and thus shorter penetration depth in solid-state material. Assuming equal thermal properties at both wavelengths, phase information will be different, with stepscan FTIR-PA proper phase information making spectral depth profiling possible for, for example, multilayer polymeric materials with a depth resolution of 1 μm. Initially, the stepscan interferometers were limited to only a few laboratories because they were very sensitive to environmental vibrations and had a poor mirror position control. Nowadays, they are commercially available, and use the computational power for interferometer function control, such as dynamic alignment, step and hold servo control, and modulation–demodulation signal generation.

17.3
Application

In general, PA spectroscopy can be an excellent spectroscopic tool, but more factors than just light absorption play a role. This might seem to complicate interpretation, but this makes it an ideal technique to monitor all kinds of other processes and to determine not only concentrations but also various material properties and local distributions of certain spectroscopically well-characterized compounds. In the following section, a number of typical studies will be shortly described with emphasis on those in which spectroscopic selectivity plays a role.

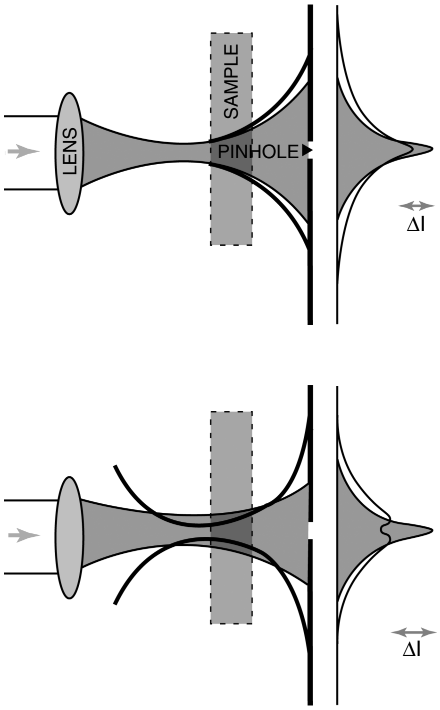

Fig. 17.10 Effect on the laser beam intensity behind a pinhole in a thermal lens effect. The upper panel shows a single-beam thermal lens effect. The lower panel shows a dual beam setup, with the effect of pump beam absorption on the probe beam intensity distribution.

17.3.1
Photoacoustic Spectroscopy in the Condensed Phase

The theory of PA spectroscopy in solids and semisolids was developed in the early seventies by Rosencwaig and Gersho at Bell labs (RG theory) not long after its revival in gaseous samples (Rosencwaig and Gersho, 1976). In those studies, chopped CW light and a gas-coupled microphone were used. In the late 1970s and early 1980s, Patel and Tam (1981) developed PA techniques on liquids and solids using pulsed laser light and a piezoelectric transducer as a contact device. Several aspects play an equally important role in the PA effect in the condensed phase: spectroscopic selectivity, optical, thermal, and acoustic properties of the sample, and the modulation frequency of the light, pulsed light being at the very high frequency end. This gives rise to a multitude of applications focusing on one or the other of these facets.

Despite their simple applicability, still, many of these have not found general use. This may very well be because of the sometimes complex nature of the PA signal in condensed samples. The amplitude and phase of the PA signal depend strongly on the modulation frequency, and both contain important information. There is a depth and absorption dependent phase lag in the signal, and if piezoelectric devices are used, the observed signal is a convolution of the PA signal and response properties of the detector itself.

For CW, light modulation frequencies between 1 and few 100 Hz are used. This can be by amplitude modulation (e.g., using a mechanical chopper), by frequency modulation, or by interferometry. Note that these frequencies are far lower than those applied in (resonance) gas spectroscopy. In

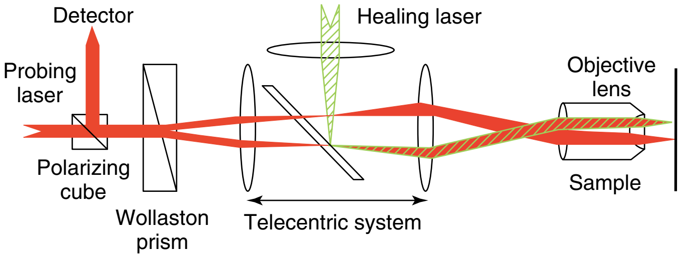

Fig. 17.11 The heating laser beam heats the sample containing the metallic nanoparticles behind the objective lens. The probing laser beam is modulated and split by a Wollaston prism into two probe interferometric arms, which are retro-reflected and again recombined by the Wollaston prism. After demodulation of the interferometer output, the modulated phase-shift induced by the heating can be detected when a nanoparticle is present. (Figure taken from Boyer *et al*., 2002.)

pulsed experiments, the short light pulses (microseconds, nanoseconds) correspond to modulation frequencies several orders of magnitude higher (megahertz, gigahertz) and are comparable to the time domain of absorption and relaxation processes. There, chemists and biologists find important applications, since PA provides a method that is complementary to flash photolysis techniques. All the more, since calorimetric information is also contained in the PA signal, it represents the part of the absorbed light energy that is not used in luminescence or in chemical processes.

The production of a PA signal in a solid (and to some degree also in a liquid) can be characterized by three parameters as follows: sample thickness, optical absorption length (depth), and thermal (heat) diffusion length. The relative values of these determine what kind of information (spectroscopy, depth profile, mechanical) can be obtained from the signal. The theoretical analysis of these cases also shows that there is an optimum coupling gas thickness (Tam, 1983). The thermal diffusion length, usually denoted as μ, depends on the chopping frequency of the incident light: $\mu = (D/\pi\nu)^{1/2}$, with D being the thermal diffusivity of the sample and ν the chopping frequency. It represents the distance traveled by the generated heat wave during one modulation period of the light. For slow modulation, this is longer than that for fast modulation, which shows in the signal phase lag, and the signal originates from deeper inside the sample.

17.3.1.1 **Solid-State Spectroscopy**

Although in gases the PA signal is inherently stronger, application of the technique to solids has important advantages over regular solid-state spectroscopy. Since only absorbed light contributes to the PA signal, spectroscopy on scattering and opaque samples is very well possible. Numerous spectroscopic studies on powders, gels, crystals, multilayer solids, nanomaterials, and surfaces have been done. Applications range from determining spectra of organic/inorganic compounds or complexes, soil samples, powders, surfaces, or thin-layer compositions (e.g., paint layers), catalysis research, nanomaterials, to monitoring living biological tissue. PA cells can be designed for each application but they are commercially available as well. Some commercial FTIR instruments have a PA cell as a standard option. In principle, a black reference material is necessary to correct for the change in spectral intensity over the wavelength range of a scan. Black rubber or carbon is normally used.

As an example, an important application is the spectroscopy of silica and semiconductor ceramics, whether in powdered form or as a crystal or a mixture of both. The band gap within the material is usually observed as an increase of the PA signal together with a change (up to 30°) in the PA phase angle at the band edge. In powdered samples several other bands may be observed, depending on the particle size and caused by quantum effects in small particles. Spectroscopy on powdered or porous samples requires some modifications in the RG theory, becaue of the presence of interstitial gas. There is some debate on the exact role that is played by this.

There is good agreement between spectra obtained with a gas-coupled microphone and with a piezoelectric transducer, but, since the detection with a transducer is directly connected to the sample, and usually at the backside, (thermo-)acoustic properties of the material can be more prominently present in the signal.

17.3.1.2 **Depth Profiling**

Many studies combine PA spectroscopy with depth profiling. As shown above, $\mu = (D/\pi\nu)^{1/2}$ is the depth from which the detected PA signal originates, as determined by the chopping frequency ν and the thermal diffusivity D in m^2 s^{-1} of the sample. A plot of the PA signal or phase versus the chopping frequency can be read as a scan through the layers of material. With thermal diffusivities in the range of $10^{-5}-10^{-9}$ m^2 s^{-1}, penetration depths can be obtained varying from a few micrometers (kHz modulation) to a few hundred micrometers (~2 Hz). It is important to realize that the depth accessible with this technique depends heavily on the absorption and thermal properties of the sample.

An objective can be to study the thermal/mechanical properties of a sample, looking at defects, or to find a distribution of (different) compounds at different depths in the sample. Likewise, the oxidation/etching depths of materials have been followed in many studies. Other examples include multilayer studies, thickness analysis, and material defects. A nonspectroscopic method is to monitor material defects by detecting the acoustic waves traveling through the material originating from surface heating by a modulated light beam.

Recent examples of this are studies in which the penetration of cosmetics or drugs into the skin (or model layers) is monitored. Using modulation frequencies between 13 and 122, or 418 Hz, the diffusion of compounds into depths between 8 and 26 μm was followed over a time course of several hours. Diffusion coefficients could be calculated from these data for different parts of the penetration layer (Hanh *et al.*, 2001).

17.3.1.3 **Thermal Diffusivity**

As a related subject, the thermal diffusivity: $D = \kappa/\rho C_p$ in m^2 s^{-1} with κ in W m^{-1} K^{-1}, the thermal conductivity, C_p in J kg^{-1} K^{-1}, the heat capacity, and ρ in kg m^{-3}, the density of the material, can be determined from the same relationship between the chopping frequency and the PA signal, using differential equations for heat flow and applying the RG theory.

For the study of thermal properties such as thermal diffusivity in solid samples, the OPC is a popular device. The sample is mounted directly on top of the (gas-coupled) microphone and is illuminated from the other side, resulting in a heat transmission configuration. Advantages of this OPC are the minimal gas volume (as small as 9 mm^3) between sample and microphone, and little sample preparation. Of course, the sample has to close off the PA cell gastight. The thermal diffusivity can be determined from the relation between the modulation frequency and the PA signal or the signal phase. A double-log plot or a plot of the PA signal versus the square root of the frequency can be used to derive the diffusivity. The technique usually does not rely on spectroscopic selectivity.

Similar schemes are possible in liquid samples as well, but then a PA cell is used in a transmission mode. Detection takes place using a piezoelectric transducer. The group of Mandelis has presented various treatises on these techniques, proposing improved modulation and referencing methods (Balderas-López and Mandelis, 2001).

Thermal diffusivities are important parameters in the thermal characterization of materials (dental resins, epoxy resins, wood, metals, alloys) and fluids (heat exchange), and the technique has been applied to this wide range of subjects.

17.3.1.4 **Phase Transitions**

Thermal (and acoustic) material properties, such as conductivity, heat capacity, and thus also diffusivity, change abruptly at phase transitions. This implies that it should be simple to determine the exact phase-transition temperatures from changes in the amplitude and phase of the PA signal. This has been used to study solid–liquid and solid–solid phase transitions in solid and liquid crystals. The transition range can vary from several degrees Kelvin to a few tens of milliKelvin.

More often, however, a pyroelectric detector is used as it proves to be more sensitive in this application than a PA sensor. The photo-pyroelectric effect detects the periodic temperature variation in a sample that is illuminated with modulated light, rather than the acoustic signal. In general, the study of phase transitions is mostly used for liquid crystals, but also for the determination of the properties of natural oils and waxes.

17.3.1.5 **Traces in Solution**

Just as in trace gas detection, traces of solutes can be detected photoacoustically with great sensitivity and by a variety of related techniques. And just as in PA spectroscopy on solids, detection in opaque samples is possible. The most straightforward way is to detect photothermal changes directly by a piezo transducer immersed in the solvent or acoustically coupled to the cell wall, or using a gas-coupled microphone. The latter option has the obvious disadvantage of acoustic mismatch between liquid and coupling gas. Another possibility is to use thermal lensing. This technique has been used to follow chemical reactions in solution in real time, and, for example, the disappearance of reactants or the formation of products can be monitored. It was proposed to use PA detection methods as detectors in high performance liquid chromatography (HPLC) systems, which has definite advantages, owing to their high sensitivity.

The sensitivity for weak absorption in liquids was compared for pulsed and CW laser excitation (Faubel *et al.*, 1994). In both cases, a piezoelectric detector has been employed. For pulsed dye lasers, the limiting sensitivity is estimated to be 1×10^{-7} cm^{-1} (the minimum detectable absorption). The corresponding value found for a CW dye laser of 55 mW, chopped at about 80 kHz, was $1 \times 10^{-5}\,\text{cm}^{-1}$ (Manzanares *et al.*, 1995). Note that applying PA to liquids has the distinct advantage of enhanced density of the investigated species, either traces in solution or pure liquids versus their vapors.

17.3.1.6 **Chemical Reactions**

Calorimetry If during light absorption processes the absorbed light energy is ‘‘used,’’ ‘‘stored,’’ or ‘‘lost,’’ PA techniques can be very helpful in resolving timescales and energy balances of processes. One excellent example is time-resolved PA calorimetry; it has developed a distinct use in (bio)-chemical research since the early 1980s. When light is absorbed in the UV–VIS range other de-excitation channels become more important, depending on the chemical structure of the compound under study and of its surroundings (solvent/reactants). Chemical reactions, luminescence, or intersystem crossing to a long-living (e.g., triplet) state take away energy from the direct thermal de-excitation that produces the PA signal. A simple calculation involving excitation energy and PA intensity then gives the amount of energy ‘‘stored’’ in products or ‘‘lost’’ in luminescence:

$$E_\text{a} = \varphi_\text{f} E_\text{f} + \alpha E_\text{a} + \varphi_{st} E_\text{st} \tag{17.7}$$

where E_a is the absorbed light energy ($h\nu$), ϕ_f and E_f are the quantum yield and energy of fluorescence, respectively, α is the fraction of absorbed energy immediately released as heat, and ϕ_st and E_st are the quantum yield and energy involved in energy-storing processes (reaction, intersystem crossing, all on a molar basis). By comparing the PA signals (over a range of concentrations) of the process under study with those of a standard that returns all absorbed light energy immediately as heat, calorimetric information is obtained. Over the years, knowledge on energy balance and PA response has become more detailed and it is now fairly well understood how reaction enthalpies, reaction volumes, and the timescales of processes are incorporated into the PA signal. It was also shown on many occasions that changes in chemical volume, solvent interactions, and enthalpy contributions of phenomena such as hydrogen bonding need to be taken into account (Laarhoven, Mulder, and Wayner, 1999).

Volume Effect The acoustical signal is not generated only by thermal expansion due to heating. Many chemical reactions are accompanied by changes in volume between reactants and products (reaction stoichiometry), or between ground state and transition state (e.g., protein folding). One of the reactants or products may be gaseous (photosynthesis), which produces a significant volume (decrease) increase. Also, differences in solvation between reactants and products can produce a volume effect (and also affect the reaction enthalpy) that can become incorporated in the observed PA signals.

Time Resolution Since most chemical reactions take place in the submicrosecond time domain, usually pulsed techniques are used in combination with piezoelectric transducers such as PZT (lead titanate zirconate crystal) or PVDF/polyvinyl difluoride (PVF_2). The first has a relatively slow time response and the signal obtained is a convolution of the PA response and the response/ringing of the piezo element. Deconvolution methods and a reference compound must be used to obtain time constants of the processes involved. The

PVDF film has a faster time response, and the PA trace obtained from this device is in the nanosecond–microsecond domain. If reaction rates are faster than the time response of the transducer, all energy used/produced in the reaction is integrated in the PA signal. If reactions are slower (factor sometimes used: 5τ, where τ is the instrument response time), the energy is stored. In the intermediate time domain, deconvolution methods must resolve reaction rates.

Because enthalpy, volume change, and time resolution are contained in one signal, a wealth of information can be obtained using PA techniques, but disentangling the signal can be cumbersome and requires an insight into the processes under study.

As a more recent development, the thermal grating method was proposed as a means to disentangle thermal and volume effects more easily (Terazima, Hara, and Hirota, 1995; Nishioku *et al.*, 2002). In this method, the interference of two light waves causes a transient grating in the solution. Photoexcitation of the sample by this light creates a (sinusoidal) modulation in the refractive index that contains contributions from the released reaction heat, the reaction volume change, and the change in absorption spectrum. The grating is probed by a second laser, similar to the thermal lensing system.

17.3.1.7 **Biological Processes in Solution**

Along the same line as the chemical investigations, important research on biologically relevant molecules and *in vivo* processes is in progress (Herbert, Han, and Vogelmann, 2000). For instance, inhibition of photosynthesis systems I and II has been studied photothermally in leaves and in cyanobacteria. Inhibited photosynthesis leads to an increase in heat produced under illumination. Various schemes to attach a PA cell to a leaf have been designed. The main complication is to assure gas exchange to keep the leaf functioning, while closing the cell tight enough to obtain a valid PA signal. The signal from an *in vivo* experiment has to be disentangled both for time resolution and for gas evolution (photobaric signal).

A great deal of research has been done on *in vitro* photosystems. Time resolution and volume changes associated with the different steps in the primary photochemistry of various rhodopsins have been studied by various laser-induced PA spectroscopy experiments (Losi *et al.*, 1999). Photothermal deflection measurements on bacteriorhodopsin have been shown useful to investigate conformational changes in photorefractive compounds (Schulenberg, Gaertner, and Braslavsky, 1994).

Laser-induced PA spectroscopy in biological tissue has developed itself toward three-dimensional tomography, to such an advanced state that animal brain structures can noninvasively be imaged, *in vivo* with skin and skull intact (Wang *et al.*, 2003). The imaging modality combines the advantages of optical contrast and ultrasonic resolution. The intrinsic optical contrast could not only reveal blood vessels but also other detailed brain structures.

Photothermal interference combined with high-frequency modulation and polarization interference provides an efficient, reproducible, and promising contrast method to visualize low amounts of proteins in cells (Boyer *et al.*, 2002). Membrane proteins in cells labeled with 10-nm gold nanoparticles could be visualized (Figure 17.11). The high sensitivity of the method and the stability of the signals allow 3D imaging of individual nanoparticles without the drawbacks of photobleaching and blinking of fluorescent dyes normally used for optically marking molecules. Furthermore, the photothermal image is immune to the effects of scattering background, which limits particle imaging through Rayleigh scattering to diameters larger than 40 nm.

17.3.2
Photoacoustic Spectroscopy in the Gas Phase

The full potential of molecular selectivity of the PA method becomes apparent when it is applied to the gas phase. It is not a high-resolution method (being limited by Doppler and/or pressure broadening); its successful application derives from high sensitivity, easy handling, and fast response.

17.3.2.1 **Overtone Studies**

Intracavity laser PA spectroscopy is one of the most sensitive techniques for studying (highly forbidden) vibrational overtone transitions. First used by Stella, Gelfand, and Smith (1976), this type of spectroscopy, utilizing tunable dye lasers or titanium–sapphire lasers, can determine weak overtone absorptions up to frequencies in the visible region because of its inherent high sensitivity. Overtone excitation deals with vibrational transitions. Since fluorescence lifetime of these energy levels in the infrared is much longer, as compared to collision-induced rotational–vibrational relaxation lifetimes, the sensitive LIF detection is not applicable. Information about highly excited vibrational levels is most useful to study molecular structure, anharmonicity constants, and (non-)localized bonds. Without these types of measurements, the local mode model for R–H bonds would be poorly developed. Typical spectra are obtained with Doppler-limited linewidths (<100 mbar) and at room temperature; the latter to ensure sufficient vapor pressure of larger organic molecules. A typical example is the high-resolution study of the 4 ν_1 band of monofluoracetylene with the Ar^+-pumped titanium–sapphire ring laser (laser bandwidth 20 MHz) (Vaittinen *et al.*, 1993). Within the tunability of this laser (9090–17 800 cm^{-1}), Doppler-limited rotational lines of 58 overtone bands in the wavelength region of 10 750–14 500 cm^{-1} have been observed with a very good signal-to-noise ratio. In this way, a study of Fermi resonance interactions at high overtone excitations has been performed. A resonant PA cell was used, excited in its first longitudinal mode (resonator length 180 mm, cross section 2.1 mm × 5.3 mm, overall length 200 mm) (Lehmann, Scherer, and Klemperer, 1982).

The group of Demtröder (Hornberger *et al.*, 1995) has utilized a 500-mW single-mode F–Center laser (6300–6400 cm^{-1}) and a radially excited resonant PA multipass cell (50 passes) to detect weak combination bands of acetylene. In spite of the relatively modest laser power, a minimum detectable absorption of $\alpha_{\min} = 10^{-9}\ \mathrm{cm}^{-1}$ has been achieved (Figure 17.12). Often, multiple passing is considered to be a nuisance with PA detection because each mirror reflection produces a mirror signal (analogous to the window signal discussed earlier). Here, however, a radial acoustic mode has been excited, to which the mirror signal barely contributes.

17.3.2.2 **Relaxation**

There have been many proposals and attempts to study gaseous relaxation processes by PAs. Phase and amplitude of the PA signal change in a characteristic way when the modulation frequency ν and the relaxation time τ obey $2\pi\nu\tau = 1$. However, before the group of Hess really started

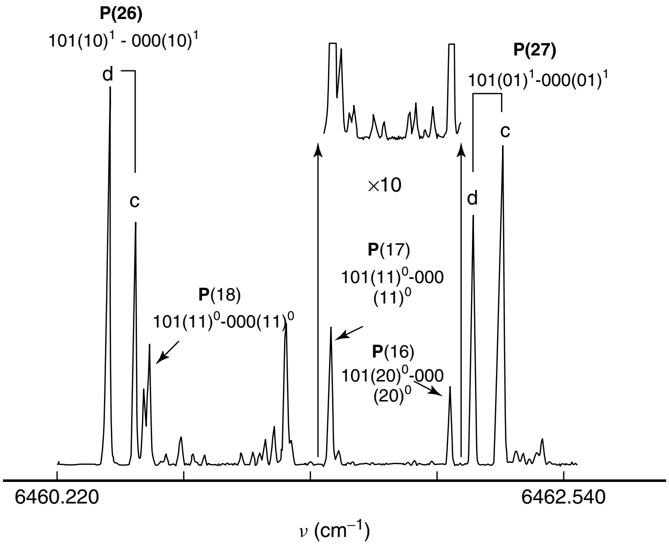

Fig. 17.12 Photoacoustic spectrum of C_2H_2 around 6461 cm^{-1}, the acetylene pressure is 16 mbar. The peaks belong to weak hotband transitions, that is, starting from vibrationally already excited levels with small populations at room temperature. The sensitivity of photoacoustic spectroscopy is illustrated by the inset, which has a 10-fold amplification (Hornberger *et al.*, 1995). (Reproduced with permission from Hornberger, C., Konig, M., Rai, S.B., and Demtroder, W. (1995) *Chem. Phys.*, **190**, 171–177).

an accurate analysis of laser excitation of acoustic modes in cylindrical and spherical high Q-resonators, the results suffered from ambiguities and systematic uncertainties (Karbach, Röper, and Hess, 1983). Essential in the treatment of Hess and coworkers was the theoretical and experimental study of the shift of the resonance frequency and the width of the acoustic resonance curve for various sample pressures. Consistent rate constants have been derived for several reactions, for example, $N_2O_4 \rightarrow 2NO_2$. Similarly, vibrational and rotational energy transfer has been studied for CH_4 excited to the (1,0,0) vibrational state at about 3000 cm^{-1}. At room temperature, rotational relaxation (equilibration) of 1.18 ns bar has been found to be very different from the value for vibrational relaxation (1.74 μs bar).

Vibrational relaxation occurring in the PA cell also has unintentional, practical consequences for other applications. Perhaps the most spectacular is the so-called kinetic cooling effect. PAs relies on (fast) relaxation into thermal heat. If the relaxation is too slow compared to the modulation frequency, the absorbed energy is not represented in the PA energy, that is, the energy is lost (the gas is kinetically cooled). The results are a lower signal and a phase shift. Well known for this is CO_2 in air; traces of CO_2 are excited by the CO_2 laser radiation (e.g., $(02^00) \rightarrow (001)$). Quasi-resonant, slightly endoergic vibrational energy transfer occurs to N_2 molecules: CO_2 (001) $\rightarrow$ N_2 ($v = 1$). Some translational energy from the CO_2 molecules is thus transferred and stalled as internal vibrational energy of the N_2 molecules and, temporarily, cooling occurs. Collisional relaxation of the excited N_2 ($v = 1$) molecules is too slow for PA

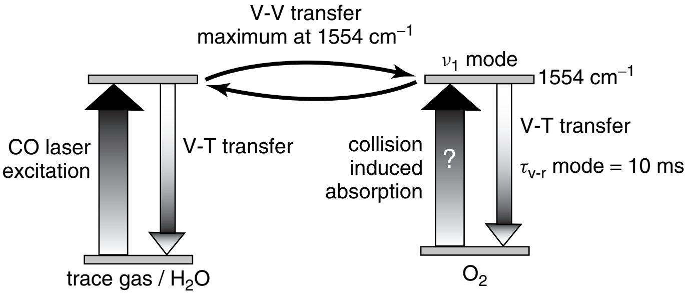

Fig. 17.13 Schematic overview of the excitation and energy transfer process in ethanol in the presence of O_2. Oxygen molecules can become vibrationally excited via vibrational energy transfer from a vibrationally excited trace gas, or directly via collision-induced absorption. Vibrational–translational transfer of H_2O and most trace gases takes place at a nanosecond timescale, while for O_2 this process takes about 10 ms.

measurements at high modulation frequencies and will therefore not contribute to the PA signal. A phase shift of the signal, however, can be observed at typical circumstances as a consequence of the temporary cooling. Addition of H_2O or SF_6 molecules (notorious relaxers) can reverse the situation.

An example is shown in Figure 17.13. Collisional relaxation allows observation of collision-induced absorption phenomena within oxygen. Oxygen molecules, whose ν_1 vibrational mode is centered at 1554 cm^{-1}, can be vibrationally excited via collisional energy transfer. Through collision-induced absorption, that is, a collision between oxygen and a trace molecule or a collision between two oxygen molecules, a dipole moment is induced, which creates an infrared absorption band. Because the induced dipole only exists for times that are on the order of the collision rate, the spectra exhibit a large bandwidth. The induced absorption strength of oxygen is relatively low but still significant due to the high abundance of O_2. The number of collisions required for a vibrationally excited O_2 molecule to relax to the ground state is 8.3×10^7 in pure O_2, while at standard temperature and pressure, 8×10^9 collisions per second take place (Lambert, 1977). Thus, relaxation of vibrationally excited O_2 to the ground state takes about 10 ms, which is long compared to a modulation period of 0.9 ms.

Figure 17.14 presents phase measurements in a mixture of water vapor and ethanol diluted in air and nitrogen. In addition, it shows the collision-induced absorption strength as taken from Thibault and coworkers (Thibault *et al.*, 1997). The observed phase has a pattern similar to the collision-induced absorption strength, however, with a long tail at higher frequencies. If collision-induced absorption would be the main mechanism, such a tail is not expected. Molecules excited at frequencies above 1554 cm^{-1} probably lose some of their energy via rotational–translational energy transfer because in polyatomic molecules the rotational–translational energy transfer is very rapid. As a result, they are closer in resonance to O_2 and vibrational energy transfer is likely to occur. A molecule excited at a frequency below 1554 cm^{-1} can lose some of its energy making vibrational energy transfer to O_2 less probable and hence no tail is expected.

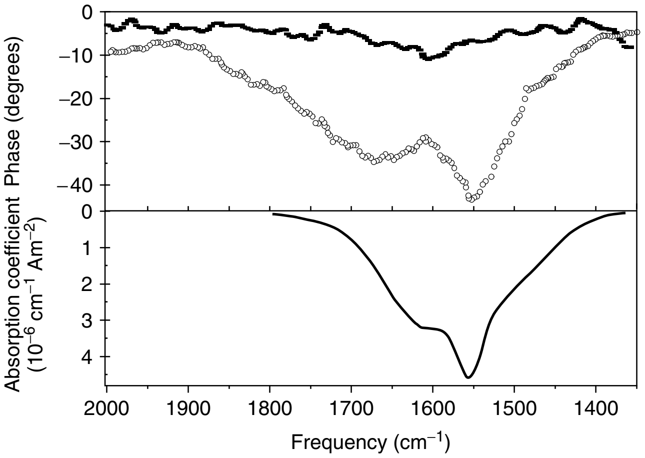

Fig. 17.14 The upper panel presents the phase angle for water vapor and ethanol, both at ppmv level recorded in air (○) or nitrogen (■). In nitrogen, the remaining structure in the phase angle is partly due to out-of-phase contributions of the background signal. The lower panel presents the O_2-air normalized absorption coefficients for collision-induced absorption as taken from Thibault *et al.* (1997)). The main mechanism responsible for the observed phase lag is vibrational energy transfer from excited water/trace gas molecules to the slowly relaxing O_2 molecules.

17.3.3
Trace Gas Detection

PA spectroscopy shows its full flavor in trace gas detection for three main reasons. Firstly, the spectral selectivity becomes exploited; many gases have a clear spectrum of identifiable absorptions in the infrared. Secondly, PA and related techniques are fast and thirdly, they are sensitive. Trace gas detection techniques are important in view of environmental applications, but also for their possibilities in basic science (biology, agriculture, human health) (Cristescu *et al.*, 2008). With their help, it is possible to discover and control mechanisms in plant physiology such as: germination, blossoming, water household, stress reaction, respiration, fermentation, senescence, ripening, injury, and so on. In the following sections, examples of these important applications will be discussed. In the context of human health, an example of analyses of exhaled air and its interpretation toward the physiological status of the human body is given. For PA trace gas detection, a few, usually less sensitive, instruments are commercially available.

Fast and sensitive trace gas detection is not exclusively reserved for PAs. In Table 17.2, various spectroscopic methods are compared, all utilizing modern laser equipment. The first one, direct absorption (DA) spectroscopy, relies on long effective absorption path lengths, since in general the minimum observable laser power difference $\Delta P/P$ is in the order of 10^{-4}. This, combined with Beers' law: $P = P_0 e^{-\alpha l}$, in which P_0 is the incident laser power, P the transmitted laser power after distance l in centimeters through the sample, and α in cm^{-1} the absorption strength at that frequency, shows that sensitivity increases with increasing path length. Even better detection limits can be obtained by combining the multipass cell with a modulation technique, such as wavelength

Tab. 17.2 Feature comparison of the different spectroscopic detection schemes.

	DA	WMS	CRDS	PAS	QEPAS
Calibration required	Yes	No	Yes	No	No
Detection time constant	ms	ms	ms	s	s
Low pressure response (∼10 mbar)	Good	Good	Good	Some loss	Good
Alignment, complexity	Average, multipass alignment	Average, multipass alignment	Difficult. Mode-matching	Easy	Average, tuning fork alignment
Spectral coverage, limiting factor	Wide, photodetector	Wide, photodetector	Limited, mirror coatings	Very wide, windows	Very wide, windows
Cell volume	300 ml	300 ml	500 ml	30 ml	1 μl
Ability to detect broad absorption features	Yes	No, only species till HWHM modulation amplitude	Yes	Yes	No, only species till HWHM modulation amplitude

[a] DA: Direct Absorption, WMS: Wavelength Modulation Spectroscopy, CRDS: Cavity Ring-Down Spectroscopy, PAS: Photoacoustic Spectroscopy, QEPAS: Quartz-Enhanced Photoacoustic Spectroscopy, Half-Width Half Maximum (HWHM) of the absorption profile.

modulation spectroscopy (WMS) (Bomse, Stanton, and Silver, 1992). Originally, WMS was designed for microwave spectroscopy where it is a standard method. However, now it is also commonly used in laser spectroscopy for fast and ultrasensitive trace gas detection applications. The cavity ring down (CRD) method combines accurate decay time measurements of light pulses in an optical cavity filled with weakly absorbing gas ($\Delta\tau/\tau = 10^{-4}$ is observable) with extremely long effective absorption lengths (Berden, Peeters, and Meijer, 2000). The last method, mainly advocated in this contribution, rests upon very sensitive detection of acoustical signals generated by infrared laser absorption in an intracavity-placed resonant PA cell and with quartz enhanced photoacoustic spectroscopy (QEPAS). Each of these spectroscopic detection methods can offer the sub-ppbv (1 ppbv = 1 part per billion volume = 1: 10^9, 1 pptv = part per trillion volume = 1 : 10^{12}) sensitivities required for trace gas detection (Kosterev and Tittel, 2002). However, they also have their own advantages and disadvantages, which should be considered carefully for each specific application. When the experiment, for example, requires determining absolute absorption values, cavity ring-down spectroscopy or DA with multipass enhancement is preferred. These two methods and WMS would also be good choices if highly dynamic processes are studied, since they require a short measurement time per point. However, the required gas volume of these methods is quite large (∼liter) which can be an obstruction for small gas samples. In that case it will be better to use normal PA- or quartz-enhanced PA spectroscopy, which have volumes of

30 ml and 1 μl, respectively. The PA detection schemes also have the advantage of having a very broad spectral usefulness, covering a wide wavelength range.

17.3.3.1 **Environmental Applications**

At low absorption signal levels uncertainties in the measurements are determined by the microphone noise and/or background signal fluctuations. For stronger absorption signals, uncertainties can be either limited by the used calibration mixture, presence of spectral interfering compounds, or uncertainties in the used absorption coefficients. Under good controlled conditions a 1% uncertainty has been achieved (Morgando, 2008).

Measurement of practical samples can pose a real challenge as besides the compound(s) of interest other absorbing gases are present in often much higher levels. For instance, exhaled human breath contains typically 4–5% of both water vapor and carbon dioxide, both of which absorb in large part of the infrared spectrum. When no special precautions are taken, the measurement uncertainties will be considerably higher than for the ideal case of measurement in pure nitrogen or clean air. In most of our measurements, we lower the level of compounds showing spectral interference by using a cold trap and/or chemical scrubbers such as $CaCl_2$ for H_2O removal or KOH for CO_2 removal. Consequently, uncertainties in the measurements are only slightly higher as compared to the ideal case. For molecules like ammonia or nitric oxide these tricks cannot be played since they readily adsorb to most scrubber materials. As these molecules have a permanent electric (ammonia) or magnetic dipole (nitric oxide) moment they can be selectively detected with Stark or Zeeman modulation PA spectroscopy.

Early attempts have been made to measure stack gas emissions from power plants with laser PAs. Owing to their high amounts in the emissions, nitrogen oxide (NO and NO_2) compounds contribute significantly to photochemical smog formation and acidification of the soil. To reduce the total amount of NO_x in the stack gas, NH_3 is added in the exhaust gas toward the chimney. Using a voluminous catalyst, NO_x reduction (within the stack gas) takes place. To check the performance and to avoid an excess of ammonia injection, the ammonia concentration in the chimney is monitored. For this, a CO_2 laser-based PA detection system was successfully applied for *in situ* monitoring of ammonia concentrations (Olafsson *et al.*, 1989). Owing to the difficult, hostile environment halfway up the chimney (vibrations, temperature fluctuations, etc.), concessions have to be made as to the sensitivity of the apparatus. With a nonresonant PA cell at 125 °C a detection limit of 1 ppmv NH_3 was achieved in a multicomponent gas mixture containing 10–15% CO_2.

Another example of a mobile system was developed and used in a field campaign by Sigrist and coworkers (Moeckli, Fierz, and Sigrist, 1996). Installed in a small trailer, the stress on the equipment is less severe. Thanks to this approach, the system has been operational for years and has been applied for several field campaigns in urban and rural environments. It can operate, for example, over a consecutive five-day period continuously to analyze multicomponent gas mixtures during sunny days in the summer, using 12 laser lines; three of them representative for ammonia absorption, three for water vapor, two for ethene, and the others for CO_2, benzene, and toluene. The spectrum of these 12 selected lines could be measured within about 10 minutes, thereby determining the time resolution of the derived pollutant concentrations (Moeckli, Fierz, and Sigrist, 1996). In literature, extremely low detection limits are mentioned on the basis of a signal-to-noise ratio of one, extrapolated from larger quantities of trace gases in a clean buffer gas. In reality, the detection limits are higher due to multicomponent gas mixtures, which induce cross sensitivities in the absorption coefficients. A mathematical analysis of the PA spectra of multicomponent mixtures is based on the weighted least squares fit of the measured spectra with additional iteration steps. By choosing the best set of laser lines with known absorption constants of all the relevant gases, the error in the calculated concentrations can be reduced.

A disadvantage of lab-based systems is that they are still rather bulky. Nowadays, complete PA systems are integrated into durable, man portable, computer size boxes, completely protected from dust, precipitation, and damage, making them applicable in field measurements. Two examples are given, one for measuring ambient ammonia concentration in the lower ppbv (1 ppbv = 1 part per billion volume = $1 : 10^9$) concentration range (Huszar *et al.*, 2008). Within the environment, ammonia is the third most abundant nitrogen compound in the atmosphere and plays an important role in cloud and rainwater chemistry. Typical ammonia sources include livestock, fertilizers, soils, forest fires, humans, animals, oceans, industry, and traffic. The main emission source is the use of animal manure and fertilizers for agricultural soils (responsible for more than 80% of the total ammonia emission). The PA instrument is based on a wavelength-modulated, near-infrared, room temperature-operated diode laser, and a PA cell in its longitudinal resonance mode, with compact electronics for fully automated, long-term operation. The PA cell was made of polyvinylidene fluoride to reduce ammonia adsorption effects. Cross sensitivities (from atmospheric water vapor) were suppressed by using a multiwavelength approach. The system featured highly reliable, automatic operation achieving a detection limit of 50 ppbv for ammonia.

Another laser-based PA instrument was developed and used for aircraft measurements of ethene from industrial sources near Houston. This instrument provided 20-second measurements with a detection limit of ~0.7 ppbv (de Gouw *et al.*, 2009). Data collected from this instrument in flight were compared with gas chromatographic measurements and they agreed within 15% on average. Ethene fluxes from a chemical complex near Houston could be quantified during 10 different flights. The average measured fluxes were an order of magnitude higher than the values regulatory emission inventories indicate.

17.3.3.2 **Plant Physiology**

It is a fortunate coincidence that the ubiquitously active plant hormone ethylene, C_2H_4, is especially easy to detect with CO_2 laser-based PAs. Ethylene is the only known gaseous plant hormone; it is involved in virtually all aspects of the plant life starting from seed germination through growth and flowering, until fruit ripening and plant senescence. The biosynthesis in plants is well understood, leading from the amino acid methionine over two intermediates, S-adenosyl-methionine and amino-cyclo-propane-carboxylic acid, to C_2H_4. Ethylene is produced by the plant tissue as a response to many external factors such as wounding, pathogen attack, light/temperature stress, chilling, drought, and so on. Modern gas chromatographic systems can be used in many instances but still lack the time resolution and extreme sensitivity of PA spectroscopy. With a detection limit of 10 pptv and a short time constant of about 10 seconds, PA detection is ideally suited to obtaining an insight into C_2H_4-triggered plant physiological properties. Examples of recent investigations are listed in Table 17.3 (Ramina *et al.*, 2007).

Numerous studies of the cellular and molecular biology of plants use *Arabidopsis thaliana* as model system. This is an organism about which much is already known, that is, easily manipulated, genetically tractable and offers the ability of fast and efficient testing of hypotheses. By using Arabidopsis as a reference system, we gain knowledge to move forward with molecular biological and genetic analyses on many physiological processes involving ethylene. With PA system dynamics of many processes in plants become easily visible. For example, ethylene released by Arabidopsis displays a circadian rhythm (high emission during the light and low emission during dark) with a peak in the mid-subjective day. When the plants were grown for six days in a succession of light (16 hours) and dark (8 hours) and afterwards were kept in continuous light for several days, they showed an ethylene pattern as would follow the light–dark periods (Figure 17.15).

17.3.3.3 **Oxidative Stress**

In humans, animals, and plants, many diseases are caused by the imbalance between the formation of oxygen radicals and other reactive oxygen species (ROS), which might ultimately result in cell injury. This is a condition of oxidative stress. These reactive molecules can start a chain of reactions known as lipid peroxidation, which leads to degradation of the outer and inner membranes of cells (Halliwell and Gutterridge, 1989).

Ethane (C_2H_6) is a product of lipid peroxidation as a result of oxidative stress and can serve as a marker for membrane degradation. Ethane is detected very sensitively using a laser operating around 3 μm, such as in the OPO-based system. As a practical example, one has investigated the damage of the chilled cucumber leaves under light and dark conditions by monitoring the ethane emission. Under light, the damage of leaf tissues was associated with the presence of free radicals. Additionally, low temperatures slow down the electron transport chains, and inhibit alternative energy dissipation paths. Both factors enhance the formation of free radicals at low temperatures. Furthermore, low temperatures also decrease the activity of antioxidants. In consequence, light-chilling leaves show elevated ethane levels in correlation with a damage leaf area (yellow coloring) of 15–30% of the total area after one-day chilling and up to 65% in a two-day treatment. The dark-chilling leaves did not produce measurable ethane and present no damage.

Plant recognition of pathogen infection leads to the so-called hypersensitive response (HR) indicated by a fast, localized cell death at the site of infection. Upon pathogen invasion an oxidative burst can be observed in plant; however, ROS alone cannot stop the pathogen invasion. A synergistic mechanism has been proposed between ROS and NO during HR in plants. Within the HR, the plant initiates a suicidal strategy, signaled by the production of NO that helps to kill the infected cells and activate the defense mechanism to isolate the infection (Figure 17.16). Using laser-based detection, the first in planta and direct measurements of NO emission from plants as response to *Pseudomonas syringae* challenge has been obtained (Mur *et al.*, 2005).

Tab. 17.3 Recent biological photoacoustic investigations monitoring ethylene emission (Ramina *et al*., 2007).

Studied process	Subject	Remarks
Germination	Pea, bean, rice seeds	Ethylene accompanies starting of germination, growth of shoots and roots
Root development	Arabidopsis seedlings	Ethylene emission during the root development
Flowering	Orchids, Carnations, tulips, Alstromeria, tobacco, transgenic tomatoes	Dynamics of ethylene emission during flower development and wilting
Pollination	Tobacco flower	Ethylene response to pollen tube growth in tobacco flower
Ripening	Fruit	Ethylene and CO_2 emission rates during ripening and post harvest of fruits
Cell death	Algae, tomato cells	Ethylene production during inducing cell death
Circadian rhythms	Arabidopsis plants	Circadian rhythms of ethylene emission in Arabidopsis
Wounding	Tomatoes, endive, cut fresh herbs	Wound-induced ethylene emission in tomatoes, endive, and cut fresh herbs
Water stress	Rumex plants	Ethylene production in relation with shoot elongation during submergence
Drought stress	Wheat, sun flowers	Ethylene rate in sensitive vs. tolerant wheat species and in sunflowers grown under potassium starvation and subjected to drought
Heavy metal stress	Aquatic plants	Ethylene from aquatic plants under heavy metals stress
Plant–insect interaction	Pine-sawfly *Diprion pini*	Plant responses induced by insect egg deposition
Plant–pathogen interaction	Tomato, Arabidopsis – *B. cinerea;* barley-powdery mildew; apples-bacteria	Ethylene production by *B. cinerea in vitro* and in tomato fruit; ethylene response from Powdery mildew infected barley plants; apples infected with *Erwinia amylovora*
Nitrogenase activity	Cyanobacteria	By adding acetylene, ethylene is produced functioning as a sensitive and fast indicator for nitrogen fixation
Hormone interaction	Mutated Arabidopsis plants, transgenic tomato	Cross talk between ethylene, abscisic acid, and auxin in mutated Arabidopsis and transgenic tomatoes

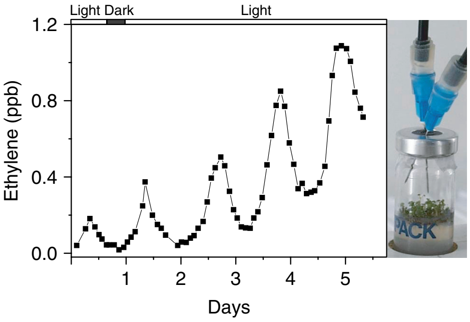

Fig. 17.15 Influence of day length on ethylene production in Arabidopsis. Seedlings were grown for six days in 16 hours light and 8 hours dark and then tested for ethylene emission during one light–dark cycle followed by constant light. Ethylene released by Arabidopsis displays a circadian rhythm with a peak in the mid-subjective day.

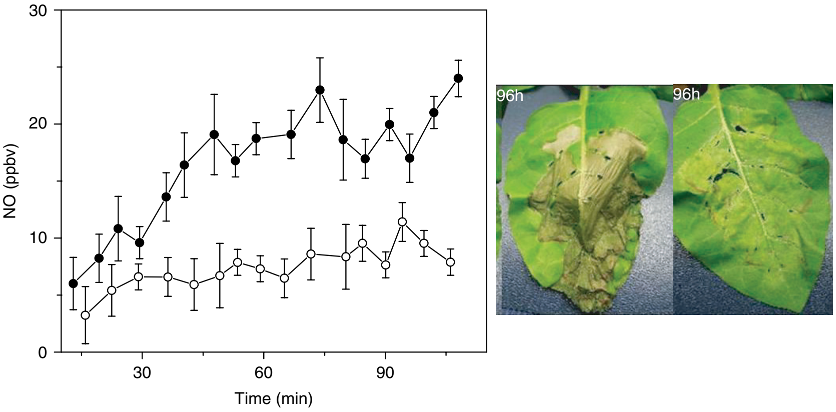

Fig. 17.16 NO formation after infection in tobacco leaves by virulent *P.Syringe pv. tabaci* (dark circles) and avirulent *P.Syringe pv. phaseolicola* (open circles). A suicidal strategy is applied by plant during the avirulent bacteria infection; plant localized fast the lesion due to the HR mechanism (a) compared to the one that did not recognize the bacteria and develops a typical disease response (b).

The quality of agricultural products at the time they arrive at the consumers strongly depends on the developmental stage at harvest, shipping, and storage conditions. Therefore it is commercially advantageous that fruits and vegetables have a long shelf life and do not deteriorate immediately after harvest. Fungal pathogens are known to be responsible for many post-harvest deteriorative processes. Enhanced production of the plant hormone ethylene is one of the earliest responses of plants to the perception of a pathogen.

Using a CO_2 laser-based PA detector the ethylene production by fungi was studied under *in vitro* conditions and then correlated with the ethylene emission from the infected host.

Infection of tomato fruits with *Botrytis cinerea*, a worldwide abundant fungus that attacks more than 200 plant species, resulted in enhanced ethylene release which started to rise before visible decay development (Figure 17.17). This demonstrates that ethylene can be considered a sensitive marker for early infection in harvested fresh products.

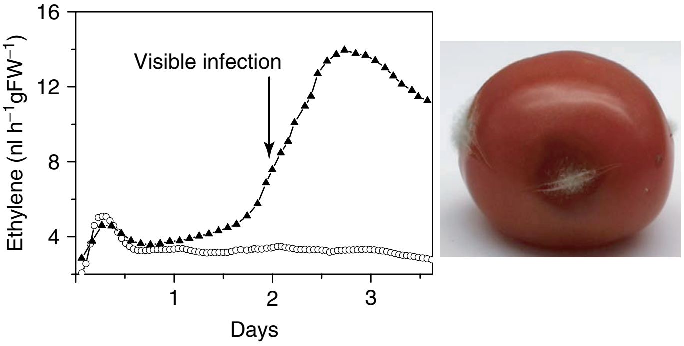

Fig. 17.17 Early detection of *B. cinerea* infection in tomato fruits. Ethylene production from artificially inoculated tomato fruits (triangles) starts to increase one day before visible infection. Production rate is given in nanoliter per hour per gram Fresh Weight (nl h^{-1} gFW^{-1}). For comparison ethylene emission of the uninfected fruit (circles) is shown (Cristescu *et al.*, 2002).

Besides ethylene, other compounds related to the quality control especially during storage can be monitored in real-time with the laser-based PA detectors. Currently, a trend in fruit storage is to reduce the oxygen level in order to slow down ripening and senescence. By lowering the oxygen level, aerobic respiration is gradually replaced by alcoholic fermentation that produces acetaldehyde and ethanol. These two compounds can thus serve as markers for suboptimal storage conditions; a timely observation is required in order to extend the storage period. Since in the storage room the concentration of these (and other trace gases) is normally below the detection limit of conventional gas analyzers, long accumulation periods are needed to obtain detectable concentrations. By avoiding the necessity of accumulation, online detectors such as laser-based PA trace gas detectors have proven to be capable of sensitive detection of ethanol and acetaldehyde at and below the part per billion levels.

Methane is an important greenhouse gas that originates both from anthropogenic and natural sources. In 2006 it has been reported that terrestrial plants may also emit methane under aerobic conditions by an unknown physiological process, and in this way may substantially contribute to the annual global methane budget with 10 Tg up to 260 Tg per year (Keppler *et al.*, 2006). The reported data have generated debates among the scientific community and also the general public. So far these findings were obtained in one single laboratory which was criticized for the experimental setup. Using an OPO-based system we have re-examined whether plants are producing methane in aerobic conditions by monitoring the $^{12}CH_4$ and $^{13}CH_4$ emissions of plants. Toward this aim ^{13}C-labeled plants grown under controlled conditions were used. Since about 99% of the carbon found in these plants was in the form of ^{13}C, one can expect that nearly 99% of the methane emitted by these plants is in the form of ^{13}C-methane. It is preferable to use these ^{13}C-labeled plants since the natural background concentration of ^{13}C-methane is 20 ppbv compared to 1.7 ppmv of ^{12}C-methane. An overall average production of 21 ng g^{-1} h^{-1} ^{13}C-methane was measured online which is 6–18 times lower than the average methane emission rates reported for the same plant species under ''sunlight'' and ''no sun'' conditions, respectively (Dueck *et al.*, 2007). Furthermore, less than 1 ppbv ^{13}C-methane was produced over a six-day period by the ^{13}C-labeled plants while the total plant biomass increased by 30%.

These tests indicated that plants produce insignificant amounts of methane in aerobic conditions, and the overestimated methane previously reported can be partly due to the diffusion of atmospheric methane from the plant tissue. Our results have been later confirmed by several other groups (Nisbet *et al.*, 2009; Beerling *et al.*, 2008).

17.3.3.4 **Microbiology**

The growth of all organisms depends on the availability of mineral nutrients, and nitrogen is one the most required nutrients in many biological reactions. Although the earth's atmosphere contains nearly 80% of nitrogen gas, this is unavailable for use by most organisms, being an inert molecule. In order for nitrogen to be used for growth it must be ''fixed'' (combined) in the form of ammonium or nitrate ions. This conversion process is known as nitrogen fixation and is exclusively performed by prokaryotes (bacteria) using an enzyme complex called nitrogenase. The acetylene reduction is the most widely used method for measuring the nitrogenase activity. It is based on the nitrogenase property to reduce compounds with a triple bond. Thus, by reducing nitrogen to ammonia, the nitrogenase reduces acetylene to ethylene. Both of these gases can be measured with high sensitivity by gas chromatography with flame ionization detection. However, due to its higher sensitivity and fast time response (about 20 seconds), laser PA spectroscopy offers the advantage of monitoring in real time small changes in the nitrogenase activity of low amounts of biomass.

The PA technique was applied for the first time to investigate the nitrogenase activity in a sample of the heterocystous cyanobacterium *Nodularia spumigena*. The

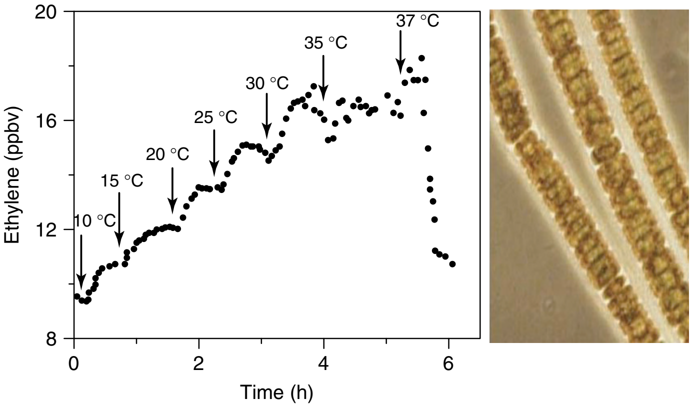

Fig. 17.18 Response of cyanobacteria *N. spumigena* to change in the temperature from 10 to 37 °C under light conditions. Significant instability in the nitrogenase activity is observed at 35 °C, leading to a short inhibition tendency at 37 °C, followed by a drastic decrease.

online monitoring allows measurements under constant conditions, which eliminates artifacts caused by changes in concentrations of oxygen or carbon dioxide due to respiration or photosynthesis during the incubations. Changes in light intensity and temperature, respectively, induce fast response in the nitrogenase activity. Knowing the level of saturation of the enzyme at different light intensities is important when rates of acetylene reduction are converted to rates of nitrogen fixation. At 35 °C and above this temperature, the nitrogenase activity is inhibited, most probably because nitrogenase, or because of the other substrate-generating enzyme systems (ATP and e^-) like the electron transport chain, are breaking down irreversibly (Figure 17.18).

17.3.3.5 **Entomology**

Laser PA trace gas detection is able to investigate the (respiration) behavior of the small insects like the Western Flower thrips, *Frankliniella occidentalis* (the smallest winged insects, size 1–2 mm, weight 50 μg) or single ant (weight of few milligrams) and *Drosophila melanogaster*, fruit fly (weight of 1–2 mg). In Figure 17.19 the respiration behavior of a single ant of 3.7 mg is shown. The ant is free moving in a 2-cm long tube with 4.5 mm internal diameter, and the CO_2 released is directed to the OPO-based system using air as the carrier gas. The respiration pattern periodically shows the release of CO_2 over several minutes that have been associated with the opening and closing, respectively, of the ant spiracles (the respiratory apertures that control the gas exchange of the insect).

Since the capability of the OPO-based system for monitoring CO_2 release of small insects has been demonstrated, one may think of investigating the metabolic changes in a single cell. In the future, this is in reach with the developments of more powerful lasers. For example, with a powerful pump laser which can supply more than 20 W of radiation between 1020 and 1050 nm, 10 times more power at the maximum of the CO_2 absorption can be

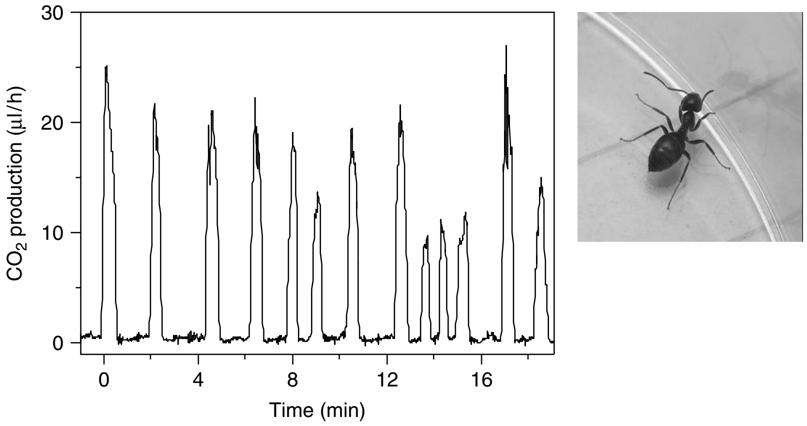

Fig. 17.19 Real-time CO_2 emission of a small ant. The respiration pattern displays periodic peaks of CO_2, which repeat on average once every 1.5 minutes. During the experiment the ant lives inside a small tube (diameter 4.5 mm) through which air is flowed.

Tab. 17.4 Molecules present in exhaled human breath with their physiological pathway and related health condition (Dueck *et al*., 2007; Nisbet *et al*., 2009; Beerling *et al*., 2008).

Compound	Concentration in breath	Physiological origin	Biological-pathology Indication
Acetaldehyde	ppb	Ethanol metabolism	Metabolism of lung cancer cells
Acetone	ppm	Decarboxylation of acetoacetate	Diabetes, fasting response, glucose metabolism
Ammonia	ppb	Protein metabolism	Liver and renal diseases, fasting response
Carbon dioxide	%	Respiration product	$13CO_2/12CO_2$ test for detection of ulcer bacteria *Helicobacter pylori*
Carbon monoxide	ppm	Degradation of hemoglobin by the heme oxigenase enzyme	Blood flow promotion
Carbonyl sulfide	ppb	Gut bacteria	Acute lung rejection in transplantations, liver and bile duct diseases
Ethane	ppb	Oxidative stress Lipid peroxidation	Chronic obstructive pulmonary disease, Rheumatoid arthritis, diabetes mellitus, lung cancer
Ethanol	ppb	Alcoholic fermentation of glucose by gut bacteria and yeast	Glucose metabolism
Ethylene	ppb	Oxidative stress Lipid peroxidation	Free radicals-mediated damage in the human body
Isoprene	ppb	Cholesterol biosynthesis	Hepatocellular injury
Methane	ppm	Gut methanogenic bacteria	Bacterial overgrowth in the small intestine
Methanol	ppb	the degradation of pectin by bacteria in the colon	Carbohydrate malabsorption, pancreatin insufficiency, renal failure
Nitric Oxide	ppb	Oxidative stress, endogenously produced by nitric oxide synthase enzyme	Lung diseases (asthma, COPD, cancer, airways inflammation, etc.), vascular smooth muscle response
Pentane	ppb	Oxidative stress Lipid peroxidation	Inflammatory diseases, transplant rejection, breast and lung cancer

generated. This will lead to an increase in the sensitivity by the same factor and pushes the CO_2 detection limit under 1 ppbv.

17.3.3.6 Human Health Research

Since ancient times, it has been known that the smell of exhaled air can be used as an indicator for several processes taking place in the human body; uncontrolled diabetes produces a sweet, fruity odor; advanced liver diseases entail a musty, fishy reek; failing kidneys bring about a urinelike smell and a lung abscess can be brought to light by its putrid stench. In the past, several attempts to use trace gas detection of exhaled air have been performed with varying success because of the insufficient detection limits of the detectors available. Gases produced in the body are transported to the lungs and are diluted (at rest, healthy persons exhale approximately 1000 l per h) before being exhaled. To measure them, it is usually necessary to accumulate the air samples by adsorbing the gases on a carrier column and releasing the concentrated gas at a certain moment, which decreases accuracy and time resolution.

Under stress conditions (e.g., ionizing radiation, toxic chemical substances, diseases, etc.), the production of free radical in the body is significantly increased. Subsequently, the capacity of the free radical scavengers in the body is overloaded and a chain of chemical reactions is activated. This leads ultimately to cell membrane damage that plays an important role in the aging processes and pathogenesis of some diseases such as cancer, Alzheimer, atherosclerosis, and so on. The cell damage is accompanied by the production of small hydrocarbons such as ethane, pentane, and ethylene that can be easily and quickly measured using PAs.

Other exhaled compounds were found to be by-products of endogenous biological processes. Ethanol, for example, is produced via alcoholic fermentation of glucose by intestinal bacteria and yeast. Acetone, a volatile ketone of fruity aroma, in breath results from the oxidation of free fatty acids.

Laser PA spectroscopy offers several advantages over the GC/MS analysis: (i) allows single breath collection since small sample volumes (few hundred milliliters) are required, (ii) no preconcentration steps are needed, (iii) excellent selectivity, and (iv) the detection and analysis procedure is fast (seconds–minutes) and computer-automated.

A summary of several compounds out of more than 300 present in the breath with their physiological origin and biological–pathological indication is given in Table 17.4 (McCurdy *et al.*, 2007; Kharitonov and Barnes, 2001; Risby and Solga, 2006). Many of these compounds have well-defined biochemical pathway and can be monitored in real time by PAs.

Acknowledgment

The authors wish to thank Frans Bijnen, Huug de Vries, Jos Oomens, Tim Groot, Edi Santosa, Iulia Boamfa, Bas Moeskops, Anthony Ngai, Sacco te Lintel Hekkert, Luc-Jan Laarhoven, Marc Staal and Jörg Reuss for their valuable contributions to this article.

Glossary

Absorption Coefficient: The absorption coefficient α in per centimeter is the attenuation coefficient of the light transmitting through a material. The transmitted light intensity I through a material with thickness l in meters is related to the incident intensity I_0 according to Beers law: $P = P_0 e^{-\alpha l}$, since for small absorptions this equation can be linearized: $\Delta P = P_0 - P = P_0 \alpha l$; the minimum detectable absorption from a device can be denoted as $\alpha_{min}/\text{cm}^{-1}\ \text{W}^{-1}$

Acoustic Resonator: Here, part of the photoacoustic cell allows a standing wave pattern to be formed (resonance). The modulation frequency of the exciting light beam should be tuned to the resonance frequency of the resonator.

Band Gap: Energy difference between the valence band and the conduction band in a semiconductor.

Boxcar Amplifier: Amplifier in which signals that occur within a preset (subsecond) time gate are selectively sampled.

Brewster(s) Angle: The angle of incidence on a window at which plane-polarized light suffers no reflection loss.

Brownian Motion: Random motion of small particles due to statistical variations in molecular pressure around the particles.

Cavity Ring Down (CRD): Sensitive direct absorption method. A laser pulse is stored in a high finesse optical cavity containing the sample. The decay time of the light intensity indicates the absorption strength.

Cell Constant: Parameter that describes the efficiency of conversion of absorbed modulated light energy into acoustical pressure amplitude, in Pa m W^{-1}.

Continuous Wave (CW): Constant, steady state delivery of laser power, as opposed to pulsed.

Depth Profile: Dependence of the photoacoustic signal (phase of amplitude) on sample depth as a consequence of a concentration gradient or changes in the thermal properties of the substrate.

Ethylene: Trivial name of ethene, C_2H_4, mostly used in biology; a gaseous plant hormone, involved in many processes. It can be detected very sensitively by photoacoustic spectroscopy.

HPLC: High performance liquid chromatography. A technique that is widely used to separate compounds in the liquid phase on a packed, (particulate) solid phase column. Often equipped with a UV-VIS detector.

Intracavity Photoacoustic (IPA) Detection: Sensitive method for trace gas detection; the photoacoustic cell is placed inside the laser cavity; due to the higher laser power better sensitivities can be achieved.

Laser-induced Fluorescence (LIF): Sensitive detection of fluorescence – radiative relaxation of electronically excited molecules – is used to detect specific laser-excited molecules. A technique used, for instance, for monitoring molecules in a flame.

Line Tunability: Indicates the restriction for gas lasers to emit light only for a discrete set of wavelengths.

Lock-in Technique: Only signals with a well-defined AC frequency are sampled. A lock-in amplifier locks to and measures the photoacoustic signal using the resonance frequency, ignoring all other signals (noise).

Multiple Pass Setup: A method to increase the absorption pathlength x-fold by reflecting a light beam many times between two mirrors.

Notch Filter: A simple acoustical interference filter that suppresses specific acoustical wavelengths, for example, perturbations from outside entering the acoustical resonator of a photoacoustic detector.

Open "Organ Pipe" Resonator: The central part of an often-employed resonant photoacoustic cell; it consists of an open-ended straight $\lambda/2$ tube, with the microphone mounted at its center where the pressure amplitude has its maximum.

OPC, Open Photoacoustic Cell: Simple photoacoustic cell, in which the sample serves as one wall, usually opposite the (gas-coupled) microphone.

Optical Parametric Oscillator (OPO): A nonlinear optical device where incoming laser radiation of frequency ν_0 is transformed into two light beams of frequencies ν_1 and ν_2 with $\nu_1 + \nu_2 = \nu_0$.

Overtone: Vibrational transitions (due to infrared light absorption) to a higher quantum level than the first excited level. Usually observed as weak absorption lines at shorter wavelengths.

Phase (angle): Position on a periodic wave (sinus) relative to a reference point on that wave (usually a node). In a photoacoustic signal the phase carries temporal information.

Piezo (electric) Element: A material, mostly a ceramic crystal or a polymer foil, that undergoes charge separation when experiencing pressure; piezo = pressure.

Quality Factor: Number that characterizes the quality of a cavity. It is given by the ratio between the stored energy and the energy loss during one cycle under stationary operation.

Quantum Cascade Laser: Laser action based on photon emission by a well-defined cascade of electrons through the layered semiconductor material.

Relaxation: De-excitation of a species after absorption of (light) energy. Relaxation can be radiative (fluorescence, phosphorescence) or nonradiative and within a wide time-window.

Rhodopsin: Protein that efficiently absorbs light in the middle of the visual spectrum with a maximum at 500 nm. Part of the light receptor in the eye and in, for example, halobacterial photoreceptors.

Spectral Coincidence: (Near) resonance between an employed laser transition and one or more transitions of the investigated molecule.

Spectral Selectivity, Spectroscopic Selection: The structural and vibrational properties of molecules allow selective excitation (and detection) of one component in a gas mixture while the other species do not respond to the chosen wavelength of the laser.

Thermal Diffusivity: Thermal diffusivity: $D = \kappa / \rho C_p$ in $m^2\ s^{-1}$, with κ in $W\ m^{-1}\ K^{-1}$, the thermal conductivity, C_p in $J\ kg^{-1}\ K^{-1}$, the heat capacity, and ρ in $kg\ m^{-3}$, the density of the material; the thermal diffusivity determines the response of the material to heat produced by, for example, absorption of electromagnetic radiation.

Thermal Lensing: Thermal nonequilibrium as a consequence of local heating changes the refractive index of a (transparent) material in such a way that light rays become deflected as by the action of a lens.

Trace Gas Detection: Techniques to determine concentrations of species present below about 1 ppmv in the atmosphere.

Window Absorption: Loss of energy of a light beam traversing an optical window, for example, at the entrance or exit of a photoacoustic cell; the resulting heating produces an unwanted photoacoustic signal that must be suppressed.

References

Balderas-López, J.A. and Mandelis, A. (**2001**) *J. Appl. Phys.*, **90**, 2273–2279.

Beerling, D.J., Gardiner, T., Leggett, G., McLeod, A., and Quick, W.P. (**2008**) *Global Change Biol.*, **14**, 1821–1826.

Bell, A.G. (**1880**) *Am. J. Sci.*, **XX**, 305–324.

Berden, G., Peeters, R., and Meijer, G. (**2000**) *Int. Rev. Phys. Chem.*, **19**, 565–607.

Bijnen, F.G.C., Reuss, J., and Harren, F.J.M. (**1996**) *Rev. Sci. Instrum.*, **67**, 2914–2923.

Bomse, D.S., Stanton, A.C., and Silver, J.A. (**1992**) *App. Opt.*, **31**, 718–731.

Boyer, D., Tamarat, P., Maali, A., Lounis, B., and Orrit, M. (**2002**) *Science*, **297**, 1160–1163.

Cristescu, S.M., De Martinis, D., Te Lintel Hekkert, S., Parker, D.H., and Harren, F.J.M. (**2002**) *Appl. Environ. Microbiol.*, **68**, 5342–5350.

Cristescu, S.M., Persijn, S.T., te Lintel Hekkert, S., and Harren, F.J.M. (**2008**) *Appl. Phys. B*, **92**, 343–349.

Dueck, T., De Visser, R., Poorter, H., Persijn, S., Gorissen, A., De Visser, W., Schapendonk, A., Verhagen, J., Snel, J., Harren, F.J.M., Ngai, A.K.Y., Verstappen, F., Bouwmeester, H., Voesenek, L.A.C.J., and Van der Werf, A. (**2007**) *New Phytol.*, **175**, 29–35.

Ebrahim-Zadeh, M. and Sorokina, I.T. (eds) (**2005**) *Mid-infrared Coherent Sources and Applications*, Springer, Dordrecht.

Faubel, W., Schulz, T., Seidel, B.S., Steinle, E., and Ache, H.J. (**1994**) *J. Phys. IV*, **C7**, 531–534.

Franko, M. and Tran, C.D. (**1996**) *Rev. Sci. Instrum.*, **67**, 1–18.

de Gouw, J.A., te Lintel Hekkert, S., Mellqvist, J., Warneke, C., Atlas, E.L., Fehsenfeld, F.C., Fried, A., Frost, G.J., Harren, F.J.M., Holloway, J.S., Lefer, B., Lueb, R., Meagher, J.F., Parrish, D.D., Patel, M., Pope, L., Richter, D., Rivera, C., Ryerson, T.B., Samuelsson, J., Walega, J., Washenfelder, R.A., Weibring, P., and Zhu, X. (**2009**) *Environ. Sci. Technol.*, **3**, 2437–2442.

Halliwell, B. and Gutterridge, J.M.C. (eds) (**1989**) *Free Radicals in Biology and Medicine*, Oxford University Press, Oxford.

Hanh, B.D., Neubert, R.H.H., Wartewig, S., and Lasch, J. (**2001**) *J. Controlled Release*, **70**, 393–398.

Herbert, S.K., Han, T., and Vogelmann, T.C. (**2000**) *Photosynth. Res.*, **66**, 13–31.

Hess, P. (**1983**) *Top. Curr. Chem.*, **111**, 1–32.

Hornberger, C., König, M., Rai, S.B., and Demtröder, W. (**1995**) *Chem. Phys.*, **190**, 171–177.

Huszar, H., Pogany, A., Bozoki, Z., Mohacsi, A., Horvath, L., and Szabo, G. (**2008**) *Sens. Actuators B, Chem.*, **134**, 1027–1033.

Jiang, E.Y. (**2002**) *Spectroscopy*, **17**, 23–34.

Kapitanov, V.A., Ponomarev, Y.N., Song, K., Cha, H.K., and Lee, J. (**2001**) *Appl. Phys. B*, **73**, 745–750.

Karbach, A., Röper, J., and Hess, P. (**1983**) *Chem. Phys.*, **82**, 427–434.

Keppler, F., Hamilton, J.T.G., Brass, M., and Rockmann, T. (**2006**) *Nature*, **439**, 187–191.

Kerr, E.L. and Atwood, J.G. (**1968**) *Appl. Opt.*, **7**, 915–921.

Kharitonov, S.A. and Barnes, P.J. (**2001**) *Am. J. Respir. Crit. Care Med.*, **163**, 1693–1722.

Kosterev, A.A. and Tittel, F.K. (**2002**) *IEEE J. Quant. Electron.*, **38**, 582–591.

Kosterev, A.A., Tittel, F.K., Serebryakov, D.V., Malinovsky, A.L., and Morozov, I.V. (**2005**) *Rev. Sci. Instrum.*, **76**, 043105.

Laarhoven, L.J.J., Mulder, P., and Wayner, D.D.M. (**1999**) *Acc. Chem. Res.*, **32**, 342–349.

Lambert, J.D. (ed.) (**1977**) *Vibrational and Rotational Relaxation in Gases*, Clarendon Press, Oxford.

Lehmann, K.K., Scherer, G.J., and Klemperer, W. (**1982**) *J. Chem. Phys.*, **77**, 2853–2861.

Losi, A., Wegener, A.A., Engelhard, M., Gärtner, W., and Braslavsky, S.E. (**1999**) *Biophys. J.*, **77**, 3277–3286.

Luft, K.F. (**1943**) *Z. Tech. Phys.*, **5**, 97–104.

Manzanares, I.C., Mina-Camilde, N., Brock, A., Peng, J., and Blunt, V.M. (**1995**) *Rev. Sci. Instrum.*, **66**, 2644–2651.

Mason, W.P. and Thurston, R.N. (eds) (**1979**) *Physical Acoustics*, vol. **XIV**, Academic Press, New York.

McCurdy, M.R., Bakhirkin, Y., Wysocki, G., Lewicki, R., and Tittel, F.K. (**2007**) *J. Breath Res.*, **1**, 014001.

Moeckli, M.A., Fierz, M., and Sigrist, M.W. (**1996**) *Environ. Sci. Technol.*, **30**, 2864–2867.

Morgando, I. (**2008**) Design and realization of a primary and secondary leak standards for the measurements of leak flow rates of refrigerants, Thesis, L'Ecole Nationale Superieure des Mines de Paris, Paris.

Mur, L.A.J., Santosa, I.E., Laarhoven, L.J.J., Holton, N.J., Harren, F.J.M., and Smith, A.R. (**2005**) *Plant Physiol.*, **138**, 1247–1258.

Nisbet, R.E.R., Fisher, R., Nimmo, R.H., Bendall, D.S., Crill, P.M., Gallego-Sala, A.V., Hornibrook, E.R.C., López-Juez, E., Lowry, D., Nisbet, P.B.R., Shuckburgh, E.F.,

Sriskantharajah, S., Howe, C.J., and Nisbet, E.G. (**2009**) *Proc. R. Soc. B*, **276**, 1347–1354.
Nishioku, Y., Nakagawa, M., Tsuda, M., and Terazima, M. (**2002**) *Biophys. J.*, **83**, 1136–1146.
Olafsson, A., Hammerich, M., Bülow, J., and Henningsen, J. (**1989**) *Appl. Phys. B*, **49**, 91–97.
Patel, C.K.N. and Tam, A.C. (**1981**) *Rev. Mod. Phys.*, **53**, 517–550.
Ramina, A., Chang, C., Giovannoni, J., Klee, H., Terata, P., and Woltering, E. (eds) (**2007**) *Advances in Plant Ethylene Research*, Springer, Dordrecht.
Risby, T.H. and Solga, S.F. (**2006**) *Appl. Phys. B*, **85**, 421–426.
Röntgen, W.C. (**1881**) *Ann. Phys. Chem.*, **1**, 155–159.
Rosencwaig, A. and Gersho, A. (**1976**) *J. Appl. Phys.*, **47**, 64–69.
Schulenberg, P.J., Gaertner, W., and Braslavsky, S.E. (**1994**) *Biochim. Biophys. Acta*, **1185**, 92–96.
Sigrist, M.W. (ed.) (**1994**) *Air Monitoring by Spectroscopic Techniques*, John Wiley & Sons, Inc., New York.
Stella, G., Gelfand, J., and Smith, W.H. (**1976**) *Chem. Phys. Lett.*, **39**, 146–149.
Tam, A.C. (**1983**) Photoacoustics: spectroscopy and other applications, in *Ultrasensitive Laser Spectroscopy* (ed. D. Kliger), Academic Press, New York, pp. 1–108.
Tam, A.C. (**1986**) *Rev. Mod. Phys.*, **58**, 381–431.
Terazima, M., Hara, T., and Hirota, N. (**1995**) *Chem. Phys. Lett.*, **246**, 577–582.
Thibault, F., Le Doucen, R., Rosenmann, L., Hartmann, J.M., and Boulet, C. (**1997**) *Appl. Opt.*, **36**, 563–567.
Tyndall, J. (**1881**) *Proc. R. Soc. London*, **31**, 307–317.
Uotila, J., Koskinen, V., and Kauppinen, J. (**2005**) *Vib. Spectrosc.*, **38**, 3–9.
Vaittinen, O., Saarinen, M., Halonen, L., and Mills, I.M. (**1993**) *J. Chem. Phys.*, **99**, 3277–3287.
Veingerov, M.L. (**1938**) *Compt. Rend. Acad. Sci. USSR*, **19**, 687–688.
Wang, X., Pang, Y., Ku, G., Stoica, G., and Wang, L.V. (**2003**) *Opt. Lett.*, **28**, 1739–1741.

Further Reading

Kinsley, L.E., Fey, A.R., Coppens, A.B., and Sanders, J.V. (**2000**) *Fundamentals of Acoustics*, 4th edn, John Wiley & Sons, Inc., New York.
Mandelis, A. (ed.) (**1992**) *Progress in Photothermal and Photoacoustic Science and Technology*, vols. **1, 2 and 3**, Elsevier.
Miklos, A., Hess, P., and Bozoki, Z. (**2001**) *Rev. Sci. Instrum.*, **72**, 1937–1955.
Morse, P.M. and Ingard, U.K. (**1986**) *Theoretical Acoustics*, Princeton University Press, Princeton.
Tam, A.C. (**1983**) in *Ultra Sensitive Laser Spectroscopy*, Chapter 2 (ed. D. Kliger), Academic Press, New York.
Zharov, V.P. and Letokhov, V.S. (**1986**) *Laser Optoacoustic Spectroscopy*, Springer Series in Optical Science, Vol. 37, Springer-Verlag, Berlin, Heidelberg.

SPUTTERING

Akira Kinbara, *Kanazawa Institute of Technology, Ishikawa, Japan*

	Introduction	437
1.	**Sputtering Yield**	439
1.1	Definition	439
1.2	Experimental Results	440
1.3	Theory of Sputtering Yield	442
2.	**Fundamentals of Sputtering Apparatus**	443
2.1	Ion-Beam Sputtering	443
2.2	Discharge Sputtering	443
2.3	Magnetron Sputtering	444
2.4	Modulation of Discharge Sputtering	445
2.4.1	rf Sputtering	445
2.4.2	Collimated Sputtering	445
2.4.3	Unbalanced Sputtering	445
2.5	Atom Motion in the Discharge Space	445
2.5.1	Sputtered Atoms	445
2.5.2	Reflected Discharge-Gas Atoms	447
2.5.3	Energy of the Sputtered Atoms	447
3.	**Sputtering Apparatus for Thin-Film Production**	447
4.	**Special Phenomena**	448
4.1	Reactive Sputtering	448
4.1.1	Definition and General View	448
4.1.2	Apparatus	449
4.1.3	Thin-Film Formation	450
4.2	Chemical Sputtering	450
4.3	Selective Sputtering	450
5.	**Special Features of Sputtered Films**	451
5.1	General View	451
5.2	Inclusion of Discharge Noble-Gas Atoms	451
5.3	Columnar Structure	451
5.4	Epitaxial Films by Reactive Sputtering	453
5.5	Internal Stress	453
6.	**Conclusion**	453
	Glossary	454
	Works Cited	454

INTRODUCTION

Sputtering is an atom emission phenomenon from condensed matter. In other words, sputtering is an evaporation process or a gas-phase production from condensed matter using impact of high-speed particles such as accelerated ions without heating the condensed matter. A schematic illustration of the sputtering process is given in Fig. 1. The collision process, which is essentially a nonthermal process, suggests momentum transfer between the atoms of the condensed matter and the high-speed particles. Therefore, it should be possible that we can vaporize condensed matter even at 0 K if sputtering technology is applied to the matter because the momentum transfer is not affected by the temperature of the condensed matter.

In principle, any kind of high-energy (high-speed) particles such as electrons, neutrons, or photons can sputter matter. However, the production rate of the gas phase by use of those particles is extremely low, and their use is limited only for physical interests. As shown below, the momentum transfer between two particles is most efficiently performed by the collision between similar-mass particles. Hence, accelerated ions are usually used as the high-energy particles. The condensed matter (mostly solid) is called the *target* in sputtering technology.

As the gas phase is not in the state of thermal equilibrium with the surroundings, it condenses at the wall and other parts of the vessel in which the gas phase is produced. Many researchers often use the term "sputtering" for the combination of the emission of atoms from their condensed matter and their condensation or deposition on the surroundings. The special receiver of the depositing atoms for some applications is called the *substrate.*

The depositing atoms form thin films on the substrate. Nowadays, thin films produced

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

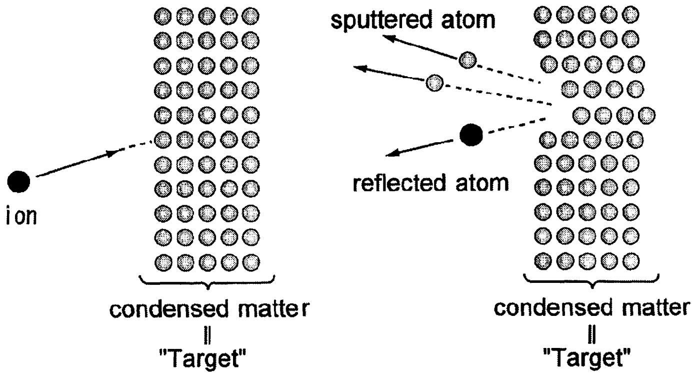

FIG. 1. Schematic illustration of the sputtering process (courtesy of Prof. Kusano, Kanazawa Institute of Technology).

by sputtering technology are applied in a variety of industries such as electronics, optics, mechanics, tribology, etc. There are a number of methods for thin-film production, but the sputtering technology has the highest flexibility for materials. Most materials including refractory metals, alloys, many inorganic compounds, and some polymers can be produced by sputtering technology.

Typical examples of the sputtered films are Al or Ti films for metal electrodes on Si integrated circuits (IC), indium tin oxide films as electrodes for liquid-crystal display (LCD) devices, TiC, TiN, and other hard coatings for machine tools, Pt films for gas sensors, etc.

The sputtering technology is very popular in advanced industries. A serious weak point has been pointed out by workers in these industries that the deposition rate of thin films by sputtering is lower than that by vacuum deposition. But this point has been overcome by introducing a magnetic field parallel to the target. This method is called magnetron sputtering (see Sec. 2.3), and is one of the most useful technologies in material science and industry. In addition, reactive sputtering, which uses reactive gas atoms as the sputtering particles, has produced good results in ceramics (refractory inorganic materials) coatings. Sputtering is no doubt one of the most important and valuable methods in advanced science and technology.

Sputtering is essentially an interaction between high-energy particles and a solid surface. Thus many phenomena in the target are related to sputtering. For example, secondary-electron emission, radiation damage of the target materials, ion implantation, modification of the target surface, etc. occur simultaneously with the sputtering. When we analyze the sputtering phenomenon in the target, we should take into account all phenomena. The sputtering theory should include a self-consistent interpretation for all of them. This makes it difficult to establish the theory and the modeling of the sputtering.

Here we consider a very primitive collision of two particles. Of course the particle–solid surface interaction is different from a particle–particle interaction, but the latter is closely related with the former. Let us consider the complete elastic collision of the two particles, of masses m_1 and m_2. At the initial stage, m_1 moves at the velocity v_1, and m_2 stands still. After the collision, the velocities of two particles change to v_1' and v_2', respectively. From the conservation of momentum and energy, we obtain

$$m_1v_1 = m_1v_1' + m_2v_2', \tag{1}$$

$$m_1v_1^2/2 = m_1v_1'^2/2 + m_2v_2'^2/2. \tag{2}$$

When we define the energy transfer coefficient Γ by

$$\Gamma = (m_2v_2'^2/2)/(m_1v_1^2/2), \tag{3}$$

we obtain from (1) and (2)

$$\Gamma = 4m_1m_2/(m_1 + m_2)^2, \tag{4}$$

$$= 4r/(1 + r)^2, \tag{5}$$

where $r = m_1/m_2$. Γ is plotted against r in Fig. 2.

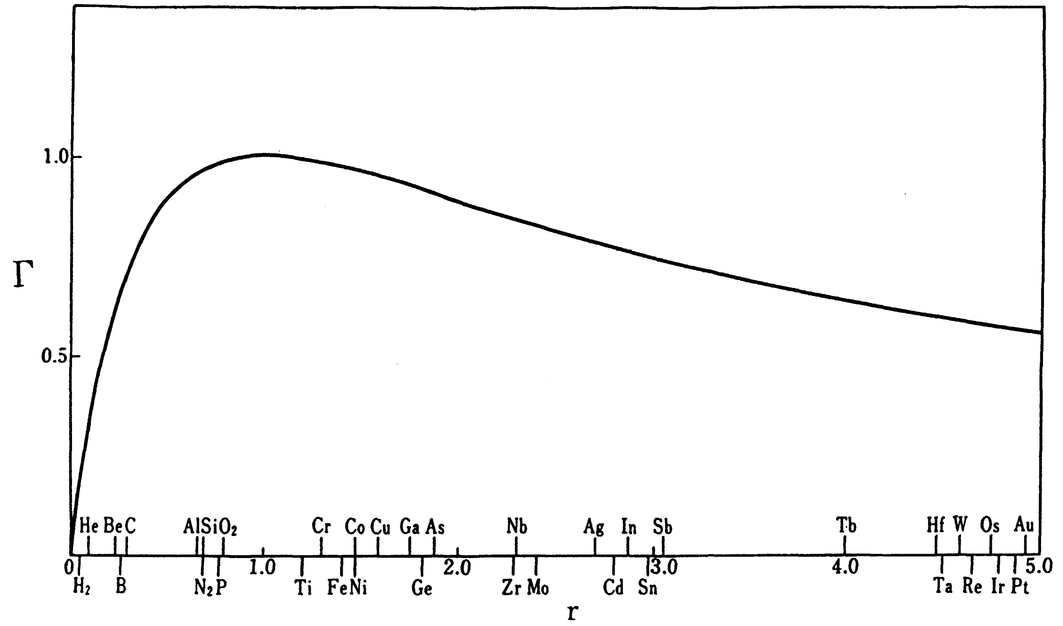

FIG. 2. Energy transfer coefficient Γ as a function of mass ratio r. The chemical symbols on the abscissa indicate the values of r for the respective substances when the bombarding gas is argon (m_1 = 39.948).

Γ is an indication of the energy transfer and often appears in the interpretation of sputtering mechanism. In actual sputtering, Ar ions are mainly used as colliding (sputtering) particles. The names of the elements shown on the horizontal axis in Fig. 2 correspond to m_2 when the mass of Ar is taken to be m_1. The dependence of Γ on r is rather dull, particularly in the range of $r > 1$. According to Fig. 2, we can qualitatively understand that Ar as a sputtering particle is excellent for the energy transfer to a variety of elements. However, electrons, neutrons, protons, and other low-mass atoms or ions are not appropriate as sputtering particles even if they have high energy.

During sputtering, the high-energy particles actually do not collide with isolated particles, but they do collide with a target consisting of many particles. This collision causes not only sputtering but also damage to the target. In addition, the cross section of the collision depends on the energy of the sputtering particles.

Therefore, the dependence of the sputtering phenomenon on energy and the related material, etc. is complicated, and the energy transfer in sputtering is, of course, not completely expressed in terms of (5), but a quantity named sputtering yield, which is one of the most important factors in the sputtering phenomenon and will be mentioned in the next section, is known to be strongly connected with Γ.

1. SPUTTERING YIELD

1.1 Definition

The sputtering yield of a target material is defined as the ratio of the number of sputtered (emitted) atoms of the target material to the number of sputtering (incident) particles in the particle impact process on the material. This quantity is denoted by Y hereafter. When we denote the number of sputtered atoms by N_s and the number of sputtering particles by N_i, Y is given by

$$Y = N_s/N_i \qquad (6)$$

Y is a fundamental quantity expressing the sputtering characteristics of a material. A large value of Y means that a small number of incident particles produces a large number of sputtered atoms. This characteristic is preferable for the production of thin films by sputtering but is not preferable when the material is used as a vessel that should be protected from the attack of fast particles such as nuclear fusion vessel.

It should be noted that *Y* is not inherent in the material, but it depends on many factors of both the sputtering particles and the target materials such as the energy, type, and incident angle of sputtering particles, the crystal surface structure, surface roughness of the target material, etc.

The direct measurement of *Y* is performed through the measurement of the weight loss of the target by sputtering (Kaminsky, 1965), even though a number of indirect methods have been developed (Krauss and Gruen, 1977). For this purpose, a constant and stable particle beam (generally, ion beam) of high energy should be produced. The energy and the intensity of the ion-beam current are varied by a controller.

The weight of the target before and after the ion-beam irradiation is measured by microbalance or another weight monitor. In a typical experimental condition, the irradiation period is an order of an hour or less and the weight loss is of the order of milligrams.

1.2 Experimental Results

Measurements of *Y* have been carried out by many workers. There is a wide variety of results. In Table 1, examples of elements sputtered by Ar ions accelerated in the energy range of 500–600 eV are shown (Maissel, 1970).

Table 1. Sputtering yield (Ar: 500–600 eV) and threshold energy of elements.

	Y	eV		Y	eV
Ag	3.4	4–15	Pb	2.7	
Al	1.05–1.2	13	Pd	2.4	20
Au	2.4	10–20	Pt	1.40–1.6	25
Be	0.51–0.8	15	Rb	1.15	
C	0.12–0.2	7–12	Re	0.9	35
Co	1.22–1.4	7–22	Rh	1.5	24
Cr	1.18–1.3	15–25	Ru	0.7	
Cu	2.3–2.55	12–17	Si	0.5	
Fe	0.84–1.3	14–20	Ta	0.6	13–26
Ge	1.1–1.2	25	Th	0.62–0.7	24
Hf	0.4		Ti	0.51–0.6	20
Ir	1.01–1.2	8	U	0.85–1.0	23
Mo	0.9	24	V	0.65–0.7	23
Nb	0.7	25	W	0.6	13–85
Ni	1.45–1.5	7–21	Y	0.68	
Os	1.0		Zr	0.65–0.8	7–22

One of the most noticeable features of *Y* is its ion-energy dependence. The particle-energy dependence of *Y* has been extensively researched. Ar-ion energy dependences of *Y* of several elements are shown in Fig. 3 as examples (Hosokawa, 1992). *Y* for Cu target has been well investigated and several data have been collected in Fig. 4 (Kinbara, 1985). The values and behavior of *Y* in most metal and semiconductor elements have been studied for a long period of time. The general features of the experimental results using noble-gas ions are as follows (*V* indicates the acceleration voltage of the ions):

1. The *V*-*Y* relation has a threshold voltage.
2. *Y* increases with *V* monotonically until a certain value of *V*.

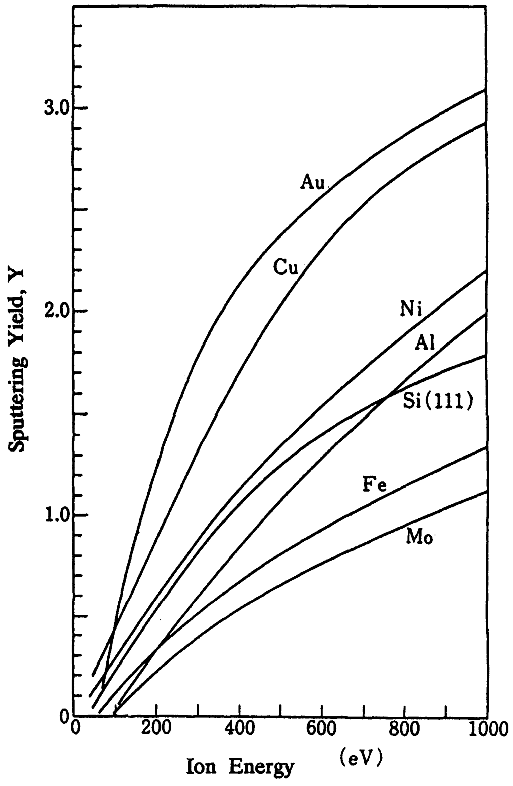

FIG. 3. Sputtering yield of various elements as functions of incident Ar-ion energy.

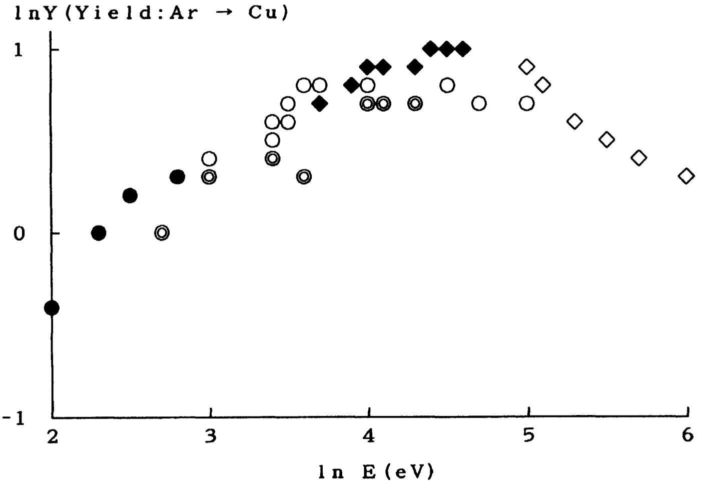

FIG. 4. Experimental results on the measurement of sputtering yield Y of a polycrystalline Cu target irradiated by Ar ions obtained by various workers: ○, Southern *et al.* (1963); ●, Laegreid and Wehner (1961); ◇, Dupp and Scharmann (1966); ◆, Yonts *et al.* (1960); ◎, Rol *et al.* (1960).

3. The *V-Y* relation reaches a maximum value of *Y* at around *V* = 10–100 keV.
4. The maximum value of *Y* does not exceed 100.

In general, there exists a threshold of energy in each target for the initiation of sputtering (Wehner, 1954; Harrison and Magnuson, 1961; Behrisch *et al.*, 1976). The determination of the threshold energy is complicated because it is difficult to generate ion beams of small energy and to observe the complete extinction of the sputtering. In addition, surface modification such as oxidation, contamination, etc. makes it difficult to obtain a precise determination of *Y* for a pure material. Many data have suggested, however, that the threshold energy ranges from a few to about 10 eV.

The measured values are not necessarily the same as the binding energies of the target materials themselves but are the same order, and it is suggested that a certain correlation exists between them. The experimental results on the sputtering threshold energy of elements are listed in Table 1.

There have been many data on polycrystalline copper target irradiated by argon ions, as shown in Fig. 4. *Y* is observed to increase with the ion energy in the low-energy region. The *V-Y* relation is affected by the surface crystal structure. Experiments on single-crystal copper targets have been done, and the dependence of *Y* on the particle energy is indicated in Fig. 5, where the Miller indices of the surface are taken as parameters (Naundorf and Macht, 1980). The curves in the figure are calculated results by Onder-

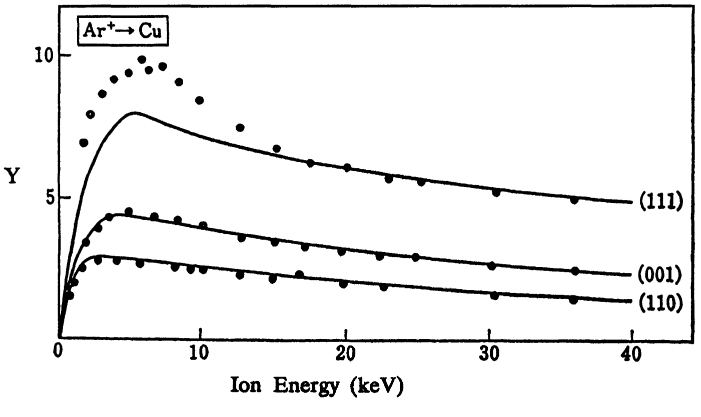

FIG. 5. Sputtering yield of three single-crystal faces of Cu vs Ar-ion energy.

delinden (1966). This figure shows that the larger surface number density of atoms causes the larger sputtering yield. The sputtering yield is related to the particle energy and the cross section for atom collisions at the target surface.

When the energy of the sputtering particles becomes very large—say, 10–100 keV or more—the value of Y is usually observed to decrease. This decrease is believed to be the effect of "channeling" (see ION-BEAM MODIFICATION OF MATERIALS). Namely, if the target is a single crystal, the incident particle can see channels from the front surface to the back surface in some directions, and the incident particles can go through without resistance. This is called the channeling effect. Of course, most targets are not completely single crystal, but the reduction of the cross section can probably cause the reduction of collision of incident particles with target surface atoms, and this is the origin of the decrease of Y.

1.3 Theory of Sputtering Yield

Modeling of sputtering phenomena has been performed by many authors in the process of the sputtering research. Among them, the theory of Sigmund (1968, 1969) is most well known. His model is based on the cascade knock-on atom generation in the target by the impact of the incident particles. Consider an infinite flat plane of finite homogeneous matter. An incident particle arrives at a point on the plane. From that point, a knock-on atom cascade starts, and atoms generated by the cascade diffuse in the target to form a dilute gas. The interaction among the atoms is ignored, and the linearity of the distribution function is considered to be preserved. The distribution function of the atoms is assumed to be determined by Boltzmann's transport equation. The incident atoms are supplied constantly, and a steady-state distribution is formed. Therefore, this theory may be called steady-state linear collision cascade theory. Final results of the calculation on the sputtering yield are related to the incident particle energy and the incident angle of the incident particles to the target surface. The sputtering in this model, in principle, is the diffusion of the gas atoms from the infinite plane to a certain direction.

An enormous amount of the calculation has been done in his model. They are so complicated that the process of the derivation is omitted here, and only the results are described. The main results are the relations of incident particle energy E and incident angle θ vs Y. As long as the elastic collision between the particles is assumed, Y is approximately expressed as follows in usual sputtering conditions:

$$Y(E) = 0.042bA/U_0, \tag{7}$$

where b is a correction factor of the energy transfer in the collision between a particle and a solid body, U_0 is the sublimation energy of the target material expressed in eV, having a value ~10 eV, and A is the nuclear stopping cross section expressed in Å^2. While the experiments showed that threshold energies for the sputtering exist, they have not been taken into account in the theory. In addition, it has been observed that a dependence of Y on the sputtered angle θ exists. These aspects have been considered by Shimizu *et al.* (1981). The relation for the modified sputtering yield Y' obtained by them is

$$Y' = [0.042bA/U_0]Y_n(E,\theta)[1 - (E_0/E)]^{1/2}, \tag{8}$$

where E_0 is the threshold energy and b is a function of θ, the angle of sputtering measured from the normal to the target surface. $Y_n(E,\theta)$ is expressed in an approximation,

$$Y_n(E,\theta)/Y(E,0) = \cos\theta - f; \tag{9}$$

here f is close to 1 in usual sputtering conditions.

Computer simulations of the sputtering process have been carried out in various ways (Ishitani and Shimizu, 1975; Helbig, 1976). These are useful to visualize what happens in the target. The collision process is repeated, a cascade is generated, and trajectories of recoil atoms and incident particles are drawn. As an example, the trajectories given by Ishitani and Shimizu (1975) are shown in Fig. 6. Two-body collision and Monte Carlo simulation are the main bases of the calculation. The Born–Mayer potential is assumed for the interaction between two

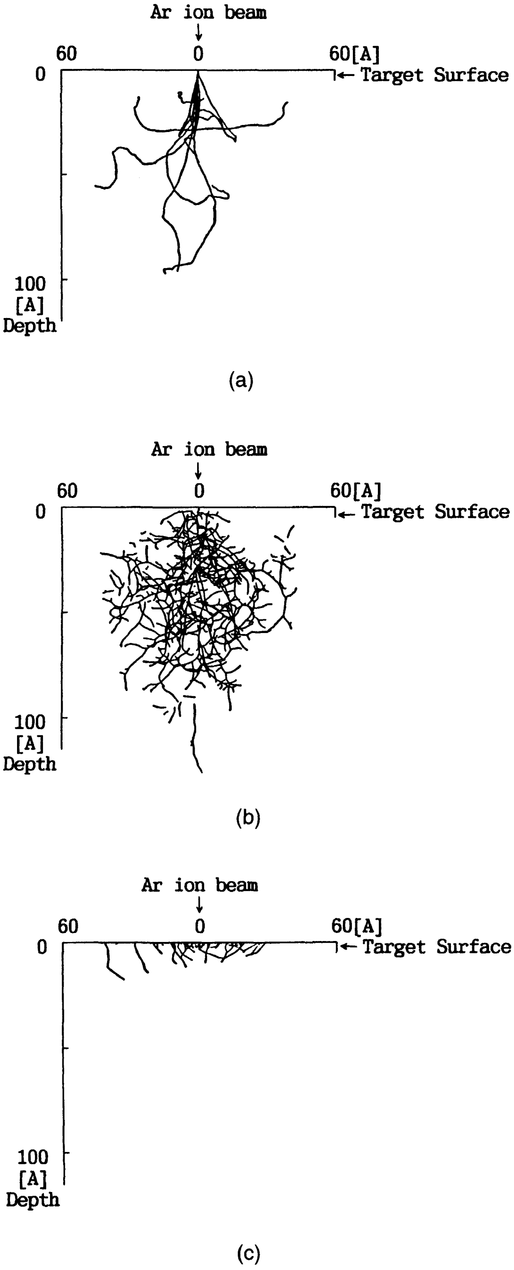

FIG. 6. Monte Carlo simulation on trajectories of (a) incident Ar ions, (b) knocked-on Cu atoms in target, and (c) sputtered Cu atoms (courtesy of Prof. Shimizu, Osaka University).

particles. Two well-known codes of the simulation, MARLOW (Robinson and Torrens, 1974; Hou and Robinson, 1979) and TRIM (Haggmark and Wilson, 1978), have been developed.

2. FUNDAMENTALS OF SPUTTERING APPARATUS

2.1 Ion-Beam Sputtering

Sputtering phenomena have been observed in various ways, but the most primitive method for the observation of sputtering is the ion-beam sputtering method. Ion beams are produced in ion sources. They are accelerated and collimated and arrive at the target. The beam current is directly measured by an ammeter to count the number of ions arriving at the target. For the measurement of sputtering yield, the secondary-electron emission from the target should be taken into account. Sometimes, a Faraday-cup–type structure is used for the target. Atoms emitted from the target move to substrates located around the target to condense and form thin films. The target is etched, and the target weight loss is measured. The number of atoms sputtered from the target is counted by the weight loss, and, hence, the sputtering yield is determined. In this method, the gas pressure around the substrate can be kept at a fairly low pressure. This method is frequently used except in the mass production of thin-film devices, because of the difficulty of manufacturing large-scale ion-beam sources.

2.2 Discharge Sputtering

A more prevailing method is discharge sputtering. A sputtering chamber is essentially almost equivalent to a discharge tube. Most of the discharge sputtering devices are composed of two facing electrodes in the vacuum chamber, and one of them, usually the anode, is grounded. On the anode, substrates are fixed, and high negative voltage is applied to the cathode. On this cathode, a target is tightly fixed.

Electrons emitted by the high voltage from the cathode are accelerated by the electric field between the two electrodes. Those electrons collide with the discharge-gas atoms and generate electrons and positive ions. The generated electrons also generate ions and electrons; thus a cascade is formed. The positive ions collide with the cathode (target) and emit secondary electrons. The cascade generation and the secondary-electron emission are essential to sustain the discharge. During the process of the positive-ion collision with the target, the target emits both electrons and neutral target atoms. This emission of neutral atoms is the sputtering by discharge.

It should be noted that the ratio of ions to discharge-gas atoms under usual glow-discharge conditions is not high. Plasma diagnostic shows that the fraction ranges between 10^{-3} and 10^{-4}.

When we use this sputtering method to produce thin-film devices in industries, a high production rate of the films is strongly required. This requires the production of a high density of ions. To satisfy this requirement, we should use high-pressure (100–10 Pa) discharge gas. The high-pressure gas increases the yield of ions, yet at the same time, it increases the scattering of the sputtered atoms and hinders the arrival of them at the substrate. Hence, the production rate of thin films by sputtering had been considered to have a limit. In addition, the rise in temperature of substrates caused by secondary electrons accelerated by the electric field during the sputtering becomes an important problem when the substrates are made of low melting point materials.

2.3 Magnetron Sputtering

In order to produce a high density of ions in the region close to the target in high vacuum (1–10^{-2} Pa), a magnetic field parallel to the target surface is applied. This field generates a cycloid motion of electrons in the space near the target surface, and the efficiency of the ion generation is drastically increased. This type of sputtering is called magnetron sputtering, and at present, most commercially available sputtering apparatus belongs to this category. Here, we interpret briefly the principle of the magnetron sputtering.

Consider an electron in the space in front of the target to which a magnetic field **B** parallel to the target surface is applied in addition to the electric field **E** perpendicular to the target surface. The equation of motion in this **E** × **B** field is

$$\frac{md^2\mathbf{r}}{dt^2} = -e[\mathbf{E} + (\mathbf{v} \times \mathbf{B})], \qquad (10)$$

where **r** is the position of the electron, t the time, and $-e$ the electron charge. This equation is simplified by taking an appropriate rectangular coordinate system and can be solved under proper boundary conditions. If **E** and **B** are constant and homogeneous, and the directions of **E** and **B** are taken as the z and y axes, the path of an electron is a cycloid and is restricted around the target surface. The trace of the electron motion in the **E** × **B** field is

$$x = x_0 + (v_0/\omega) - (v_0/\omega)\cos(\omega t + \beta), \qquad (11)$$

$$y = y_0 + (E/B)t + (v_0/\omega)\sin(\omega t + \beta), \qquad (12)$$

where $\omega = eB/m$ is called the cyclotron frequency and β is the initial phase angle. The maximum excursion of x is given by $r = v_0/\omega$. This is called the Larmor radius, and the motion of electrons on the target is limited to this radius.

This means that the ion generation occurs in a very limited space in front of the target. The restriction of the space of electron motion causes a high probability of ionization of discharge gas atoms in the restricted area, so that ions are generated even in the 10^{-2}-Pa range and sputtered atoms travel in the discharge space without being scattered by the discharge-gas atoms. The motion of an electron on the target under an **E** × **B** field is shown in Fig. 7.

Magnetron sputtering was one of the most innovative methods for the production of thin films in the 1960s. By this method, high-speed and low-temperature deposition of refractory metals could be realized. Magnetron sputtering is, as it were, an ultimate method, and no methods superior to this based on other principles have appeared yet.

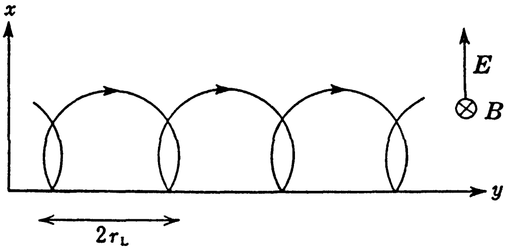

FIG. 7. Trajectory of an electron under an E × B field.

2.4 Modulation of Discharge Sputtering

2.4.1 rf Sputtering When the target (cathode) consists of an insulating material such as SiO_2, dc voltage cannot sustain the discharge because of positive charging of the target surface. Radio-frequency (rf: mainly 13.56 MHz) voltages are applied to the insulating target to maintain the sputtering. This is called rf sputtering. In rf sputtering, both electrodes (target and substrate) are sputtered. But because of a self-bias voltage appearing on the target surface, which is caused by the difference between ion temperature and electron temperature in the discharge plasma, the target is charged negatively, irradiated more heavily by positive ions, and sputtered more preferably.

2.4.2 Collimated Sputtering The direction of the atoms sputtered from the target seems to have some preference, as described in Eq. (9), but it is rather ambiguous, and on a practical level, the direction is nearly random except for a decrease at large sputtering angles. In the case of a single-crystal target, the direction of the sputtering is restricted, and what are called spot patterns appear (Wehner, 1955; Musket and Smith, 1968). They are interesting from the physical viewpoint but have little practical significance because single-crystal targets are never used in thin-film production systems.

In order to arrange the direction of the sputtered atoms, a mesh of a collimator made of stainless steel is placed in front of the target, as shown in Fig. 8. If the mesh is deep enough, the collimation is better but the loss of the film material becomes larger. This method is called "collimated sputtering" and is used for the case where the normal incidence or a certain angle of incidence of sputtered atoms is required in the deposition of the film.

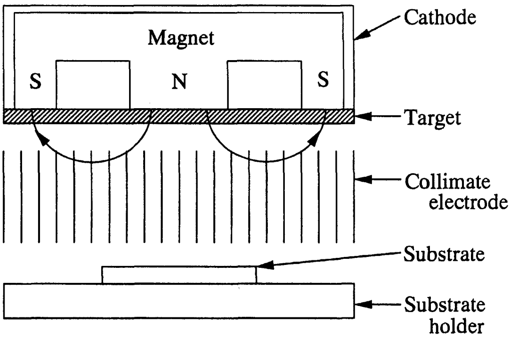

FIG. 8. Schematic illustration of collimated sputtering apparatus (courtesy of Prof. Hata, Kanazawa University).

2.4.3 Unbalanced Sputtering The magnetic field in the magnetron sputtering apparatus limits the discharge space, and the electrons emitted from the cathode cannot collide with the target. Therefore, heating of the substrate and the film by electron irradiation is prevented. The heating changes the film properties, and depending on circumstances, it gives rise to a positive effect such as removal of surface contamination. In order to make the best use of its decontamination effect, the magnetic field is intentionally distorted by adding extra magnets around the regular magnets. This type is called "unbalanced magnetron sputtering." In actual processes, the substrate seems to be immersed in the discharge space. The electrons and neutral reflected atoms from the target, and part of the ions in the discharge space, may collide with the substrate, provided the self-bias potential of the substrate is adequate.

2.5 Atom Motion in the Discharge Space

2.5.1 Sputtered Atoms The sputtered atoms emitted from the target come into the discharge space and move to the substrate. During this movement, sputtered atoms are scattered by discharge-gas atoms. This scattering impedes the direct arrival of the sputtered atoms on the substrate. Hence, some of the sputtered atoms cannot reach the substrate but are evacuated from the discharge chamber. Traces of the scattered atoms are estimated by Monte Carlo simulation (Nakano *et al.*, 1991; Kinbara *et al.*, 1993). Figure 9 shows the relative rate of sputtered atoms of La and B returned on the target. It is clear that heavy atoms such as La do not scatter much, while light atoms such as B scatter abundantly. This should cause a lack of B atoms in sputtered films produced by the direct sputtering of LaB_6, and experiments show the validity of the simulation. Figure 10 shows the composition of sputter-deposited films measured by Auger-electron

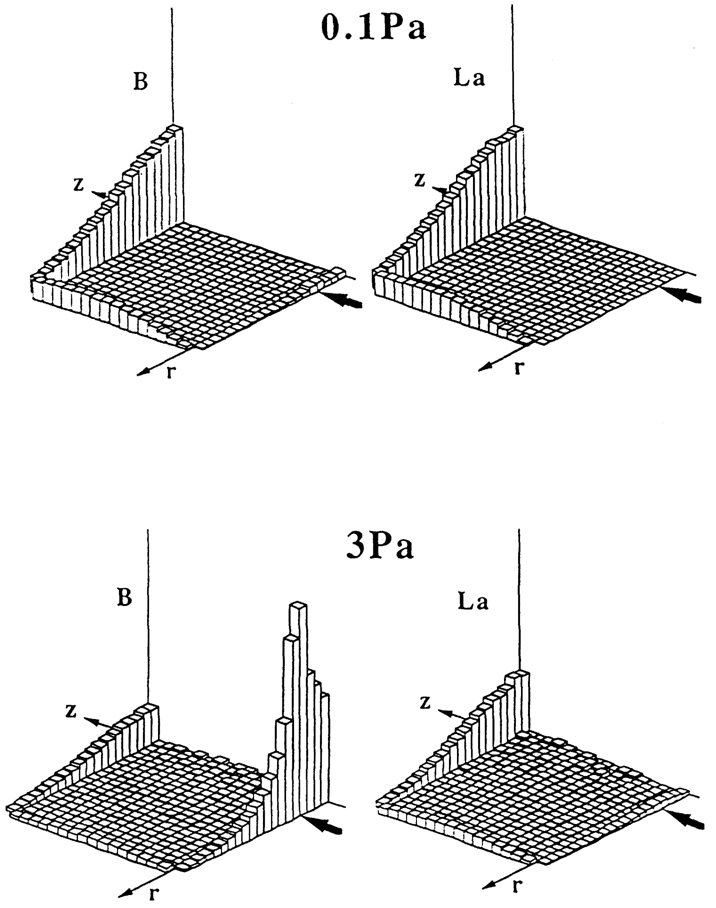

FIG. 9. Computer simulation of arriving rate of La and B atoms on a substrate under different discharge-gas pressures. Thick arrows show the starting point of both atoms at the substrate. It is clear that most La and B atoms arrive at the substrate under a pressure of 0.1 Pa, while most B atoms return to the target under a pressure of 3 Pa.

spectroscopy as a function of Ar discharge-gas pressure during deposition. At higher gas pressures, the rate of B atoms included in the film is observed to be reduced. Hence, we can conclude that the loss of B atoms is larger in thin films. It is considered that sputtering of a composite target causes a lack of the lighter component in deposited

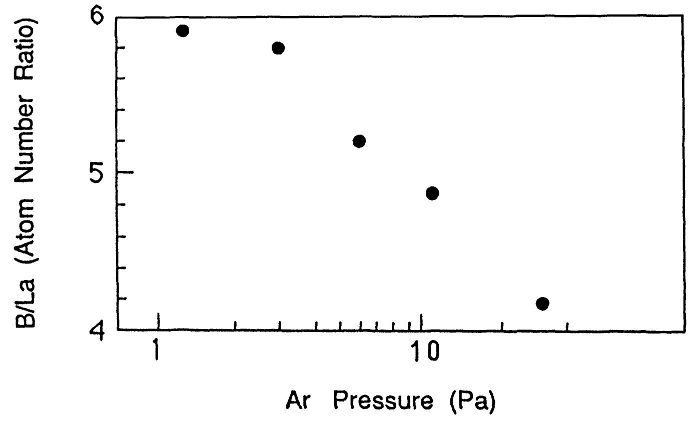

FIG. 10. Atom-number ratio of sputter-deposited La-B thin films measured by Auger-electron spectroscopy at various discharge-gas pressures.

films, and this is confirmed by the experiment.

2.5.2 Reflected Discharge-Gas Atoms

In the process of sputter deposition using a glow discharge, including magnetron sputtering, ions colliding with the target are then partially implanted in the target but mostly are reflected as their charge is lost. The reflected discharge-gas species are believed to be neutral atoms. Part of the energy of the incident ions is given to the target, and the rest remains in the reflected neutral atoms. For example, if the incident ion energy is 1 keV, the energy of reflected atoms is considered to range between several tens and a few hundreds of electron volts. These high-energy atoms cannot be controlled by electromagnetic field; therefore, part of the reflected atoms collide with the growing thin films. This effect hinders the adsorption of residual gas atoms on the film surfaces. Consequently, similar to electron irradiation, the reflected atoms "clean" the films on one hand but on the other hand, they destroy the crystal structure of the film. Some of the reflected atoms hit the atoms of the growing film and hammer them into the film inside. This effect is called peening effect and will be mentioned in Sec. 5.5.

2.5.3 Energy of the Sputtered Atoms

The energy of the sputtered atoms has been measured. There are several methods for its determination, but all the methods reveal that the energy of the emitted atoms sputtered by Ar ions of 10^2–10^3 eV ranges between 1 and 100 eV (Kruger and Scharmann, 1970; Oechsner, 1975; Elbern *et al.*, 1978).

3. SPUTTERING APPARATUS FOR THIN-FILM PRODUCTION

An ion-beam sputtering apparatus consists, in principle, of an ion-beam source and a target. Such a system is shown in Fig. 11. This system works in a high vacuum, but in the operating condition of the ion source, the vacuum is usually 10^{-1}–10^{-2} Pa.

The sputtering apparatus using a discharge is essentially equivalent to a discharge tube. But some ideas are naturally taken into account to generate effective sputtering and a high productivity of thin films.

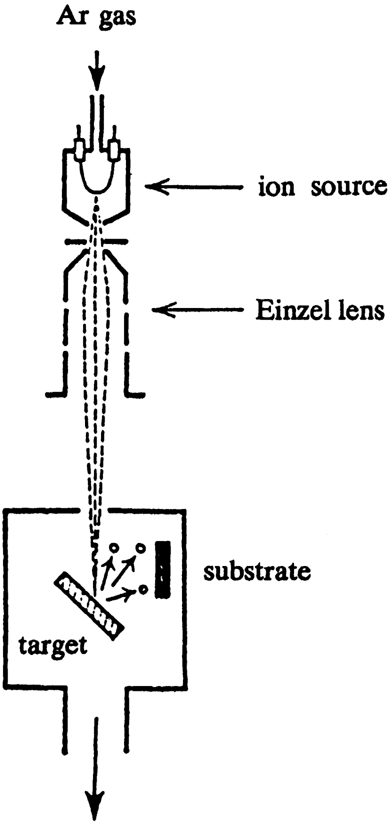

FIG. 11. An example of ion-beam sputtering apparatus.

Most sputtering devices are composed of two facing electrodes in the vacuum chamber, and generally, the anode is grounded, with the substrates fixed on it. High negative voltage is applied to the cathode. A target is tightly fixed on this cathode. In order to produce a high density of ions in the region close to the target, a magnetic field parallel to the target surface is applied as mentioned before. However, in a practical magnetron sputtering apparatus, it is not easy to produce a high parallel magnetic flux (0.1T or more) at the target surface. In reality, permanent magnets are placed on the back side of the target. The actual form of the magnetic flux used in practical equipment is schematically shown in Fig. 12. The top of the arc-shaped flux is parallel to the target, and many ions are produced at the tops.

The cathode/target is cooled by running water on its back side in order to prevent a

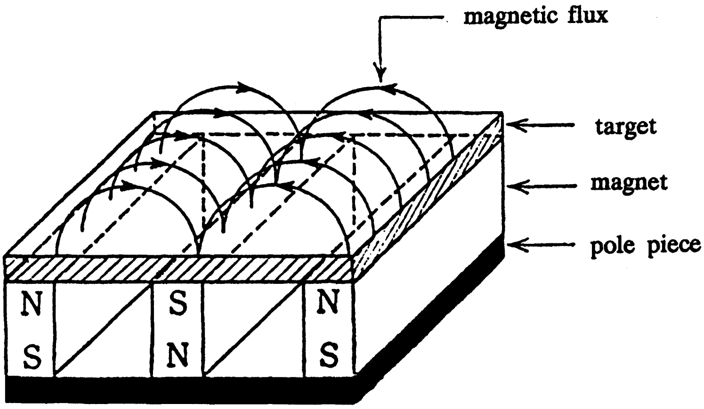
magnetic flux
target
magnet
pole piece
N
S
S
N
N
S

FIG. 12. Magnetic flux in a practical magnetron sputtering apparatus.

temperature rise during sputtering. As long as we use tap water, its electric conductivity is small enough and the electric loss through the water is ignored even if direct contact of water with a conductive cathode exists. The discharge from the cathode occurs in every direction. Namely, the discharge occurs not only between the cathode and the anode but also between the cathode and a chamber wall if the chamber is made of metal. To prevent a discharge in the unnecessary direction, the back side of the cathode is shielded by a metal plate placed close to it. Electrons emitted from the cathode to the plate arrive at the plate without colliding with discharge-gas atoms. This means that the discharge cannot occur in the back direction, and the discharge is limited to the space between the anode and the cathode. Schematic illustrations of a magnetron sputtering apparatus and a cathode/target are shown in Figs. 13 and 14.

4. SPECIAL PHENOMENA

4.1 Reactive Sputtering

4.1.1 Definition and General View Reactive sputtering is a process that forms compound thin films using a conductive (mainly metal) target and reactive gases. Many compounds such as ceramic films produced by reactive sputtering are used in various kinds of industries.

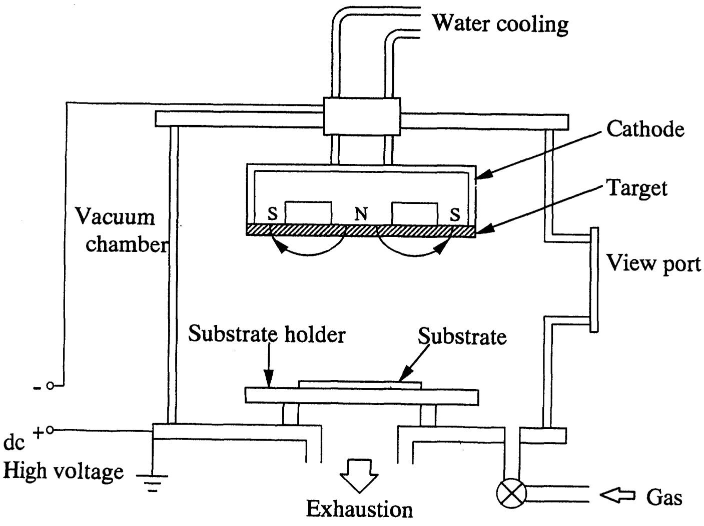
Water cooling
Cathode
Target
S
N
S
Vacuum chamber
View port
Substrate holder
Substrate
- o
dc +o
High voltage
Exhaustion
Gas

FIG. 13. A practical magnetron sputtering apparatus (courtesy of Prof. T. Hata, Kanazawa University).

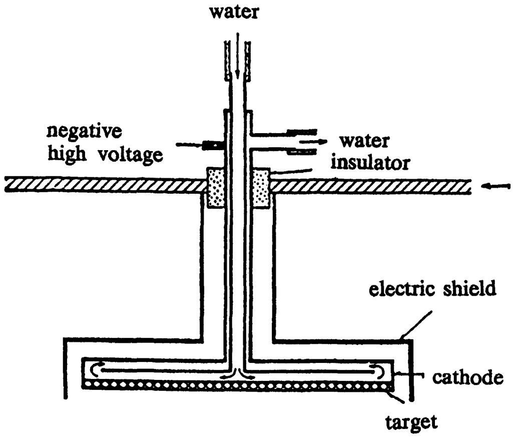

FIG. 14. Structure of a sputtering target.

Table 2. Products of reactive sputtering.

	Gas		
	O_2	N_2	CH_4
Al	Al_2O_3	AlN	
Cd	CdO		
Cr	Cr_2O_3		CrC
Co	CoO_2		
Cu	CuO		
Fe	Fe_2O_3		FeC
Ge		GeN_2	
Hf	HfO_2	HfN	HfC
In	In_2O_3		
Mg	MgO		
Mn	MnO		
Mo	MoO		MoC
Nb		NbN	NbC
Ni	NiO		NiC
Pb	PbO	PbN	
Sb	Sb_2O_3		
Si	SiO_2	Si_3N_4	SiC
Sn	SnO_2		
Ta	Ta_2O_5	TaN	TaC
Ti	TiO_2	TiN	TiC
V	VO_2		VC
W	WO_3		WC
Zn	ZnO		
Zr	ZrO_2	ZrN	

When we use a compound such as a ceramic target in order to produce compound (mainly inorganic) thin films by regular sputtering technology, we often encounter several difficulties:

1. Production of the target is not simple. For example, ceramics are generally brittle and refractory and are not easy to manufacture or to form in an appropriate shape to fit in the sputtering apparatus.
2. Selective sputtering from compound target hinders keeping the stoichiometry of the deposited film composition. A certain component of the target is short in the deposited films.
3. In many cases, the bonding of atoms forming the compound is more tight than that of conductors. This implies that the sputtering yield is small, and, hence, the deposition rate is small.

In order to avoid these difficulties, a pure metal (occasionally semiconductor) target is sputtered in a discharge gas including reactive gases such as oxygen, nitrogen, methane, silane, hydrogen sulfide, etc. Examples of generated materials are listed in Table 2 (Thornton *et al.*, 1981; Ritchie and Harding, 1979). The metal atoms sputtered from the target arrive at the substrate, while reactive gas atoms in the discharge gas also diffuse in the space and arrive at the substrate. Both metal and reactive-gas atoms/molecules migrate on the substrate and collide with each other to form compound molecules (Berg *et al.*, 1987, 1989; Kusano and Goulart, 1990; Kinbara *et al.*, 1992). This process is known as reactive sputtering. Reaction to form compounds occurs on the substrate. The collision probability of metal atoms and reactive gas molecules in the space is small and is negligible under usual sputtering conditions (~1 Pa). The reaction also occurs on the target surface. The reaction at the target surface generates a compound, inorganic, insulating material on the target, and it changes the sputtering characteristics. In particular, it has been well known that by a change in the reactive-gas flow rate, the production rate of thin films will be changed. Reduction and increase of the flow rate are not reversible but cause well-known hysteresis characteristics of the production rate. A schematic plot of the hysteresis is shown in Fig. 15. This hysteresis is caused by compound formation on the target surface. On the target surface, sputtering and compound formation occur simultaneously. The abrupt change of the production rate is related with the compound formation and disappearance (Stirling and Westwood, 1971; Donaghey *et al.*, 1976).

4.1.2 Apparatus Reactive sputtering apparatus is not specially designed. All sputtering apparatus is applicable for this pur-

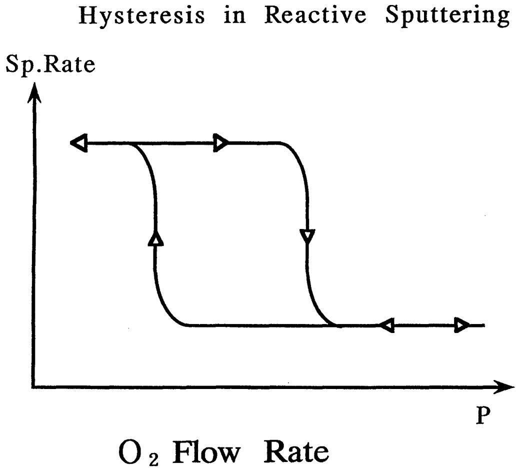

FIG. 15. Schematic form of hysteresis observed for sputtering rate with change of oxygen partial pressure in the reactive sputtering process.

pose. We only need a device to introduce the reactive gas. An accurate gas-flow controller is generally provided.

In some apparatus, the reactive gas is introduced into the chamber through a nozzle, which makes it possible for the gas to blow against the substrate (Vossen *et al.*, 1991; Niwa *et al.*, 1993). A gradient of the reactive-gas density is formed on the substrate. The degree of compound film formation is different from place to place on the substrate. We can clearly recognize and distinguish differences of compound formation on the substrate. The best amount of the reactive-gas supply is evaluated by this nozzle experiment.

4.1.3 Thin-Film Formation Many kinds of compound thin films have been produced by the reactive sputtering method. In general, sputtering is performed by use of Ar and reactive discharge-gas mixture. The most favorable ratio of Ar to reactive gas depends on the apparatus and cannot be precisely specified. However, generally the percentage of the reactive gas plays a rather small part in the total discharge-gas pressure. The role of the Ar is the sputtering of the target. This includes the cleaning of the target surface that is going to be contaminated or combined with the reactive gases.

The ratio of the number of reactive-gas atoms to the number of sputtered metal atoms should be carefully taken into account in order to produce stoichiometric materials. Sputtering and film deposition is not a thermal equilibrium process, and hence, nonstoichiometric compounds are likely to be formed. Therefore, the control of the ratio is very important in forming stoichiometric thin films.

4.2 Chemical Sputtering

Sputtering is essentially the emission process of atoms from a target. However, in the case of reactive sputtering, the reaction occurs on the substrate surface but not on the target. The name reactive sputtering has not been considered to be adequate because the reaction is not concerned with atom emission from the target.

Since nuclear fusion has been developed, the erosion of the reaction chamber wall has been widely investigated (Roth *et al.*, 1976; McCracken and Patridge, 1976; Webb *et al.*, 1980). The wall of the nuclear fusion reactor made of "low-*Z*" material such as carbon is exposed to the high-temperature hydrogen plasma. Chemical reaction of the hydrogen and the wall material (carbon) occurs at the wall and generates methane. This phenomenon has not fully been interpreted, but it is clear that accelerated H^+, D^+, or T^+ ions collide with the carbon on the wall and react to form CH_4 (Erents *et al.*, 1976). The wall is equivalent to the target in the sputtering apparatus, and this phenomenon is a genuine surface reaction on the target. This phenomenon is rather suitable for the name of reactive sputtering, but this name has already been used for the substrate surface reaction. For that reason this target reaction is called chemical sputtering. Of course, when carbon is exposed to H radicals, methane production appears. The efficiency of the methane production is much smaller than that produced by accelerated H^+ ions. The origin of chemical sputtering is still being investigated.

4.3 Selective Sputtering

When an alloy is heated and vaporized, the component of the alloy that has a higher vapor pressure at the alloy temperature evaporates more quickly than the others. The evaporation of an alloy is considered to obey Raoult's law. Hence, the composition of a vacuum-evaporated thin alloy film is not the same as that of the material in the evaporation source.

The sputtering process is close to milling or shaving, so that it is expected that the composition or the stoichiometry of the source (target) material is kept in the thin film during the sputtering process. However, actually, according to many experiments on alloys such as Ag–Cu, Cu–Ni (constantan), Ni–Cr, etc., the stoichiometry does not hold in sputtering, although the vaporization pressure of each component is not directly related with the sputtering. It has been reported, for example, that the ratio of the sputtering yield of Cu to Ni in Cu–Ni alloy ranges between 1.6 and 1.8 (Goto *et al.*, 1978; Coburn, 1979).

There exists a difference of sputtering yields among the components of the target material. The yield is not directly related to the yield of the single component state. In addition, atom diffusion of the component supplies the sputtered material onto the target surface (Pickering, 1976; Ho, 1978). Consequently, the composition of the surface of the target is kept almost constant at the initial stage of sputtering, and the larger-yield material in the alloy is constantly supplied to the target surface for a certain period. This process is called selective sputtering and makes it difficult to use alloy targets for the production of alloy films.

5. SPECIAL FEATURES OF SPUTTERED FILMS

5.1 General View

Sputtered films have characteristic features in structure and properties. They are mainly caused by the discharge-gas atoms in the sputtering chamber. During the sputtering process, the film surfaces are exposed to the discharge-gas atmosphere and adsorption of the gas atoms occurs. In addition, high-energy particles such as reflected sputtering particles at the target and, occasionally, ions irradiate the film surfaces. The adsorption and irradiation cause several effects on the structure and properties of the sputtered films.

5.2 Inclusion of Discharge Noble-Gas Atoms in the Films

In the case of the sputtering process using noble-gas glow discharge, the collision probability of the discharge-gas atoms with the growing thin-film surface is rather large. Most of those atoms desorb very quickly because the adsorption energy of the gas atoms on metal surfaces ranges from 0.05 to 0.35 eV. However, part of them remain on the film surface and are included in the film with the growth of the films. Sputtered films include more or less discharge-gas atoms. It is more probable that the reflected high-energy discharge gas atoms hit the film surface and intrude into the film (Winters and Kay, 1967; Hoffman and Thornton, 1980). The quantity of the atoms is evaluated by heating the film (Fagen *et al.*, 1974; Lee and Oblas, 1975) and x-ray fluorescence (Cuomo and Gambino, 1977; Bayne *et al.*, 1978). The total quantity of included Ar gas atoms are observed to amount to an order of 0.01–1 at. % under usual sputtering conditions. This quantity is affected by the bias voltages applied to the substrate during sputtering. When the bias voltage is −40 V, it was reported to increase to the order of 1–10 at. % (Lee and Oblas, 1970).

5.3 Columnar Structure

The structure of sputtered films is very characteristic. The structure is called "columnar structure." The columnar structure has been observed by a number of workers in the early stage of research on thin-film structure, and it has been investigated by many workers (Dahlgren *et al.*, 1970; Thornton, 1974; Dirks and Leamy, 1977).

TEM and SEM observations show that the cross section of sputtered films is observed to be thick with crystal columns grown perpendicular to the substrate surface. This structure is peculiar not only to sputtered films but to films prepared by CVD or other production methods that use gases rather than vacuum during film formation. An example of the columnar structure of sputtered Ti and Ni films prepared by magnetron sputtering is shown in Fig. 16. On top of the Ni film, a thin Au film is formed by vacuum evaporation.

Thornton (1974) accumulated and ar-

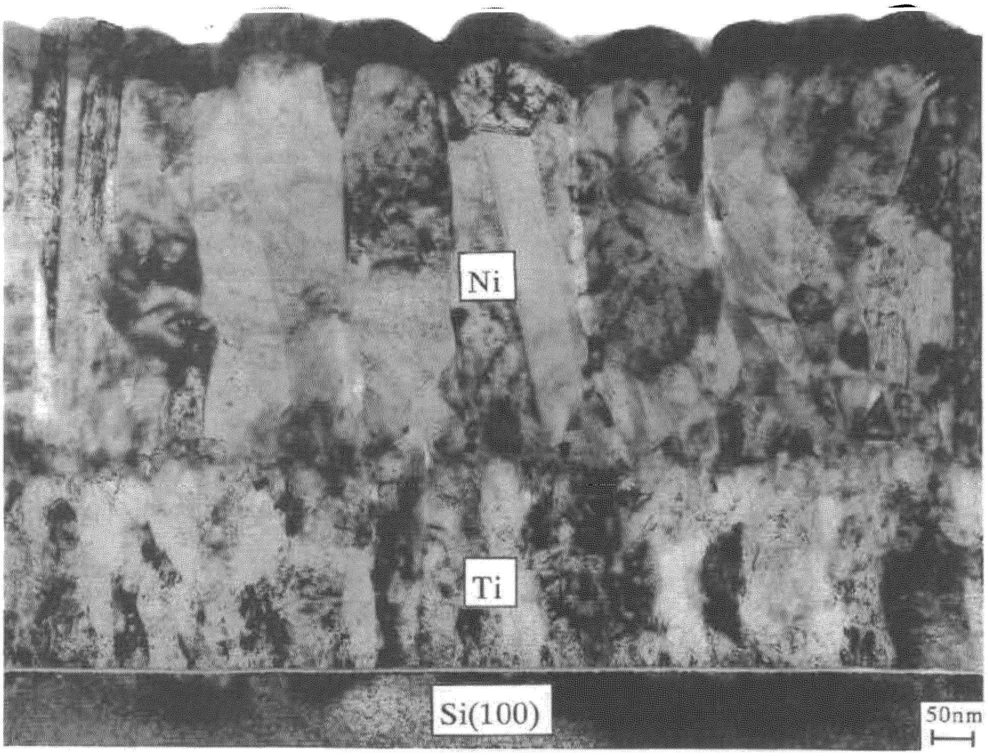

FIG. 16. Columnar structure of sputter-deposited Ti and Ni thin films (courtesy of Dr. I. Kondo, Nippondenso Co. Ltd.).

ranged the experimental data. He proposed a famous illustration of the columnar structure where sputtering-gas pressure and substrate temperature are taken as the parameters. This illustration is idealized more or less, and we occasionally observe discrepancies from this illustration in actual films, but in general, we can estimate the morphological situation of the sputtered films through this illustration. In short, the illustration shows that the columns become rough and thin at low deposition temperature and at high discharge pressure. With increasing temperature and decreasing pressure, the columns become more clear and thick and have better crystallinity. Also, the morphology of the film surface becomes flat and smooth.

One of the very peculiar features concerning the column formation is the effect of oblique incidence (Dirks and Leamy, 1977). Namely, when the incident film atoms arrive at the substrate obliquely to the normal to the substrate surface at an incident angle ϕ from the normal, then the growth direction of the axis of the column becomes ψ. It has been observed that there exists a relation between ϕ and ψ:

$$\tan\phi = 2\tan\psi. \qquad (13)$$

This relation is called the tangent rule and fits the experimental results fairly well.

In order to interpret the formation of the columnar structure and the oblique-incidence effect, computer simulations have been done by Dirks and Leamy (1977). The assumptions used in the simulation are as follows:

1. A self-shadowing effect occurs. Namely, sputtered atoms cannot arrive in the area of the shadows generated by nuclei, and islands of sputtered atoms formed on the substrate.
2. Arriving film atoms at a substrate collide with film atoms already arrived at the substrate to form nuclei. However, the movement of those atoms is hindered by

the adsorbed discharge-gas atoms on sites of the nuclei. Their movement is limited to one atomic interval.

3. The direction of arriving atom movement is random.

The simulation gives a kind of columns and, also, gives the oblique-incidence effect of the column growth, but the diameter of the columns obtained by the simulation seems to be too small. The simulation interprets the tangent rule fairly well.

5.4 Epitaxial Films by Reactive Sputtering

It is rather difficult to generate single-crystal films by the reactive sputtering method because during sputtering, growing films are always exposed to the reactive gas atoms that hinder the movement of film atoms during the film growth. The atom movement or crystallite motion is essential to form epitaxial films. However, recent studies have shown that single-crystal films are produced when the films are produced under a special condition. Horita *et al.* (1993) succeeded in producing oxide single-crystal thin films on Si substrates. They formed YSZ [yttria-stabilized zirconia, $(ZrO_2)_x(Y_2O_3)_{1-x}$] films by sputtering a Zr-Y alloy target. Before the deposition of the film, they formed an ultrathin (<1 nm) oxide layer by sputtering the target in O_2 discharge gas on the Si in order to prevent silicide formation. This metal deposition and oxidation process was repeated, and they observed the growth of fairly good single-crystal oxide thin films.

5.5 Internal Stress

Deposited thin films are well known to be in a stressed state. We call the state that the film has an internal stress. The stress is classified into two categories, namely, tensile and compressive. The internal stress of vacuum-evaporated thin films has been investigated by many workers, and common tendencies and behaviors of the stress have been understood. Most of the metal films prepared by vacuum deposition are in the tensile stress state. The stress values of many metals lie in the 10^6–10^8-Pa region. However, the stress of sputtered films depends strongly on the sputtering conditions. The stress of a given material is changed from tensile to compressive according to the conditions. Several workers investigated the internal stress as a function of the pressure of discharge gases (Thornton *et al.*, 1979; Wu, 1979). They found that the stress was tensile when the pressure of the discharge gas was high, but it changed to compressive at lower pressures. The compressive stress is ascribed to a so-called "peening" effect. This effect is essentially the irradiation of growing films by the reflected atoms at the target. When the pressure is low, the reflected high-energy atoms arrive at the film surface without being scattered by discharge-gas atoms; thus the reflected atoms maintain high energy. Hence, some atoms on the film surface are jammed into the interior of the film. They act as interstitials in the crystal lattice to expand it. The lattice expansion causes the compressive stress. This is the peening effect, and, at present, this model of the stress generation mechanism is accepted by many workers as the origin of the compressive stress of sputtered films.

6. CONCLUSIONS

The fundamentals of sputtering and its application to thin-film formation have been surveyed. Notable matters are indicated below.

Sputtering is regarded as a phenomenon connected with two-body collision. However, other effects peculiar to condensed matter such as cohesion, diffusion, radiation damage, etc. influence the sputtering. At present, simplified models have been proposed, and a formulation for dependence of sputtering yield on the energy of incident ions has been given.

One of the most effective applications of sputtering is thin-film production. In order to form thin films efficiently, a variety of sputtering methods have been developed. Ion-beam sputtering, dc-discharge sputtering, etc. have been well known. Above all, the magnetron sputtering method is now known to be one of the most excellent methods to produce thin-film devices and is widespread in industries, although other methods have their own advantages. No other method superior to this method on the whole has yet been developed in these 30 years.

Recently, reactive sputtering has been used in a wide range of industries. The apparatus is the same as the magnetron sputtering using a metal target, but the discharge gas used is a mixture of Ar and reactive gas such as oxygen, nitrogen, methane, etc. It makes it possible to produce high–melting-point inorganic compounds, ceramics, etc. This method is very useful for the production of oxides, nitrides, carbides, etc., but the process control should be drastically improved because of the instability of the discharge. For that purpose, plasma diagnostics including electromagnetic and optical measurement should be carried out more precisely during the sputtering process.

Sputter-deposited thin films include discharge-gas atoms in them such as Ar. Most of the films have a special structure called columnar structure, and often, they are in highly stressed state, but those features are complicatedly related to the sputtering conditions. These characteristics of thin films are interesting to many workers and are still in the stage of investigation.

Relatively novel technologies such as arc sputtering for high deposition rate, self-sputtering for less contamination, microsputtering for small areas, etc. have been tested. Some of them—for example, arc sputtering—have been already found practical, and some others are still in early stages of development.

In addition to the sputter deposition onto substrates of usual size, large-area sputter deposition such as the coating of window glasses of buildings has gradually been performed in industry. The homogeneity of the film thickness and the cost to produce the film by sputtering are the present points at issue. As future materials, so-called graded materials, quasi-equilibrium materials will possibly be produced by sputtering. The present sputtering technology rather depends on the know-how in individual cases. A more general and physically significant basis of sputtering technology is required.

GLOSSARY

Auger Electron Spectroscopy: An energy analysis of emitted electrons to identify species in matter by an electron bombardment. When an electron in an inner shell (K shell) of an atom is ejected by the bombardment, one electron in the upper shell (e.g., L shell) fills up the electron hole. During this process, another electron in the atom may be emitted from the atom (Auger process). The emitted electron has an energy peculiar to the atom. Hence, we can identify the material by the energy analysis of the electron.

Boltzmann's Transport Equation: A statistical equation to describe the general behavior of a distribution function of particles in phase space assuming the conservation of particle number.

Born–Mayer Potential: An exponentially decreasing potential between two particles.

Cyclotron Frequency: An electron moving perpendicularly to a homogeneous magnetic flux at a constant velocity moves on a circle of constant radius by the Lorentz force. The frequency of the circular motion is the cyclotron frequency.

Ion Implantation: The inclusion of accelerated ions in a solid by the ion bombardment of the solid.

Larmor Radius: The radius of circular electron motion in a homogeneous magnetic flux (see **Cyclotron Frequency**).

Low-Z Material: Material of low atomic number. Radiation loss of these materials in fusion plasma is smaller than that of high–atomic-number materials.

Radical: An atom or molecule that has unpaired electrons.

Raoult's Law: A law stating that the rate of the vapor pressure decrease of a solute in a dilute solution is equal to the mole fraction of the solute.

Secondary Electron Emission: Electron emission from the surface of matter or a particle by particle bombardment of the matter.

Works Cited

Bayne, M. A., Moss, R. W., McClanahan, E. D. (1978), *Thin Solid Films* **54,** 327–336.

Behrisch, R., Bohdansky, J., Oetjen, G. H., Roth, J., Schilling, G., Verbeek, H. (1976), *J. Nucl. Mater.* **60,** 321–329.

Berg, S., Blom, H.-O., Larsson, T., Nender, C. (1987), *J. Vac. Sci. Technol.* **A5,** 202–207.

Berg, S., Blom, H.-O., Moradi, M., Nender, C. (1989), *J. Vac. Sci. Technol.* **A7,** 1225–1229.

Coburn, J. W. (1979), *Thin Solid Films* **64,** 371–382.

Cuomo, J. J., Gambino, R. J. (1977), *J. Vac. Sci. Technol.* **14,** 152–157.

Dahlgren, S. D., McClanahan, E. D., Johnston, J. W., Graybeal, A. G. (1970), *J. Vac. Sci. Technol.* **7,** 398–402.

Dirks, A. G., Leamy, H. J. (1977), *Thin Solid Films* **47,** 219–233.

Donaghey, L. F., Kenneth, G., Geraghty, K. G. (1976), *Thin Solid Films* **38,** 271–280.

Dupp, G., Scharmann, A. (1966), *Z. Phys.* **192,** 284–298.

Elbern, A., Hintz, E., Schweer, B. (1978), *J. Nucl. Mater.* **76/77,** 143–148.

Erents, S. K., Braganza, C. M., McCracken, G. M. (1976), *J. Nucl. Mater.* **63,** 399–404.

Fagen, E. A., Nowicki, R. S., Seguin, R. W. (1974), *J. Appl. Phys.* **45,** 50–59.

Goto, K., Koshikawa, T., Ishikawa, K., Shimizu, R. (1978), *Surf. Sci.* **75,** L373–L375.

Harrison Jr., D. E., Magnuson, G. D. (1961), *Phys. Rev.* **122,** 1421–1430.

Haggmark, L. G., Wilson, W. D. (1978), *J. Nucl. Mater.* **76/77,** 149–155.

Helbig, H. F. (1976), *J. Vac. Sci. Technol.* **13,** 368–373.

Ho, P. S. (1978), *Surf. Sci.* **72,** 253–263.

Hoffman, D. W., Thornton, J. A. (1980), *J. Vac. Sci. Technol.* **17,** 380–383.

Horita, S., Tajima, T., Murakawa, M., Fujiyama, T., Hata, T. (1993), *Thin Solid Films* **229,** 17–23.

Hosokawa, N. (1992), in: Japan Soc. Appl. Phys. (Eds.), *Handbook of Thin Film Production,* Tokyo: Kyoritsu-shuppan.

Hou, M., Robinson, M. T. (1979), *Appl. Phys.* **18,** 381–389.

Ishitani, T., Shimizu, R. (1975), *Appl. Phys.* **6,** 241–248.

Kaminsky, M. (1965), *Atomic and Ionic Impact Phenomena on Metal Surfaces* Berlin: Springer-Verlag, pp. 142–237.

Kinbara, A. (1985), *Sputtering Phenomena,* Tokyo: The University of Tokyo Press.

Kinbara, A., Nakano, T., Kobayashi, A., Baba, S., Kajiwara, T. (1993), *Appl. Surf. Sci.* **70/71,** 742–745.

Kinbara, A., Kusano, E., Baba, S. (1992), *J. Vac. Sci. Technol.* **A10,** 1483–1487.

Krauss, A. R., Gruen, D. M. (1977), *Appl. Phys.* **14,** 89–97.

Kruger, W., Scharmann, A. (1970), *Z. Phys.* **237,** 263–271.

Kusano, E., Goulart, D. M. (1990), *Thin Solid Films* **193/194,** 84–91.

Laegreid, N., Wehner, G. K. (1961), *J. Appl. Phys.* **32,** 365–369.

Lee, W. W., Oblas, D. (1970), *J. Vac. Sci. Technol.* **7,** 129–133.

Lee, W. W., Oblas, D. (1975), *J. Appl. Phys.* **46,** 1728–1732.

Maissel, L. (1970), in: L. Maissel, R. Glang (Eds.), *Handbook of Thin Film Technology,* New York: McGraw-Hill, p. 4-1.

McCracken, G. M., Patridge, J. W. (1976), *J. Nucl. Mater.* **63,** 373–379.

Musket, R. G., Smith, H. P., Jr. (1968), *J. Appl. Phys.* **39,** 3579–3586.

Nakano, T., Baba, S., Kobayashi, A., Kinbara, A., Kajiwara, T., Watanabe, K. (1991), *J. Vac. Sci. Technol.* **A9,** 547–549.

Naundorf, V., Macht, M. P. (1980), *Nucl. Instrum. Method* **168,** 405–409.

Niwa, T., Saito, T., Kinbara, A. (1993), *J. Vac. Sci. Technol.* **A11,** 2790–2795.

Oechsner, H. (1975), *Appl. Phys.* **8,** 185–198.

Onderdelinden, D. (1966), *Appl. Phys. Lett.* **8,** 189–190.

Pickering, H. W. (1976), *J. Vac. Sci. Technol.* **13,** 618–621.

Ritchie, T., Harding, G. L. (1979), *Thin Solid Films* **57,** 315–320.

Rol, P. K., Fluit, J. M., Kistemaker, J. (1960), *Physica* **26,** 1000–1008.

Robinson, M. T., Torrens, I. M. (1974), *Phys. Rev.* **B9,** 5008–5024.

Roth, J., Bohdansky, J., Poschenrieder, W., Sinha, M. K. (1976), *J. Nucl. Mater.* **63,** 222–229.

Shimizu, R., Yamamura, Y., Matunami, N., Morita, K., Saidoh, M., Itoh, N. (1981), *Oyo Butsuri* **50,** 470–488.

Sigmund, P. (1968), *Can. J. Phys.* **7,** 731–737.

Sigmund, P. (1969), *Phys. Rev.* **184,** 383–417.

Southern, A. L. S., Willis, W. R., Robinson, M. T. (1963), *J. Appl. Phys.* **34,** 153–163.

Stirling, A. J., Westwood, W. D. (1971), *Thin Solid Films* **7,** 1–10.

Thornton, J. A. (1974), *J. Vac. Sci. Technol.* **11,** 666–670.

Thornton, J. A., Tabock, J., Hoffman, D. W. (1979), *Thin Solid Films* **64,** 111–119.

Thornton, J. A., Cornog, D. G., Anderson, W. W. (1981), *J. Vac. Sci. Technol.* **18,** 199–202.

Vossen, J. L., Krommenhoek, S., Koss, V. A. (1991), *J. Vac. Sci. Technol.* **A9,** 600–603.

Webb, A. P., Brewer, R., Stuessi, H., Veprek, S. (1980), *J. Nucl. Mater.* **93/94,** 634–639.

Wehner, G. K. (1954), *Phys. Rev.* **93,** 633–634.

Wehner, G. K. (1955), *J. Appl. Phys.* **26,** 1056–1057.

Winters, H. F., Kay, E. (1967), *J. Appl. Phys.* **38,** 3928–3934.

Wu, C. T. (1979), *Thin Solid Films* **64,** 103–110.

Yonts, O. C., Normand, C. E., Harrison Jr., D. E. (1960), *J. Appl. Phys.* **31,** 447–450.

SQUIDS

William G. Jenks, Ian M. Thomas, and John P. Wikswo, Jr., *Department of Physics, Vanderbilt University, Nashville, Tennessee, U.S.A.*

Introduction 457
1. **The Principles of Operation** 457
1.1 The Josephson Junction 457
1.2 The dc SQUID 458
1.3 The rf SQUID 460
1.4 The Pickup Coil 460
1.5 Noise in the SQUID 460
1.6 The Digital SQUID 461
2. **Design and Manufacture of SQUIDs** 461
2.1 Conventional SQUIDs 461
2.2 High-T_c SQUIDs 463
3. **SQUID Systems** 463
3.1 Gradiometers 464
3.2 Shielded Rooms 464
3.3 Cryogenic Dewar Design 465
4. **Applications** 465
4.1 Magnetoencephalography 465
4.2 Magnetocardiography 466
4.3 Other Biomagnetic Studies 466
4.4 Geomagnetism 466
4.5 Nondestructive Testing 467
4.6 Other Applications 467
Glossary 467
Works Cited 468
Further Reading 468

INTRODUCTION

Superconductors are best known for their ability to conduct current without developing a corresponding voltage. Perhaps a more fundamental property is the pairing of conduction electrons (see Superconductivity, Low-Temperature). This property of the microscopic charge carriers leads to the macroscopic phenomena associated with superconductivity. It also led to the invention of the most sensitive detector of magnetic flux known, the superconducting quantum interference device, or SQUID. The SQUID is, to the engineer, a magnetic flux-to-voltage transducer of unparalleled sensitivity. It is arguably the most sensitive detector of any kind, with an equivalent energy sensitivity that approaches the quantum limit. This article should serve as a starting point to any engineer or physicist who wishes to understand SQUIDs. The principles of operation, the methods of manufacture, and the applications of SQUIDs are each discussed in turn.

1. THE PRINCIPLES OF OPERATION

1.1 The Josephson Junction

The Josephson junction (Josephson, 1962) is the heart of SQUID technology. A SQUID is in essence a superconducting loop interrupted by one (rf SQUID) or two (dc SQUID) Josephson junctions. Figure 1 shows a typical junction and the Josephson equations, which govern its electrical properties. While several physical configurations are possible, the essential feature is a thin insulating layer, or a narrow constriction, between two superconductors. The Josephson equations may be derived from Ginzburg–Landau theory (1950). The essential feature of the theory is that in any particular superconductor, an order parameter ψ exists and has associated with it an arbitrary phase θ. In two *weakly connected* superconductors, the relative phase, $\delta = \theta_1 - \theta_2$, affects the electrical properties of the junction. The dc Josephson equation relates the applied current passing through the junction to the relative phase and the critical current of the junction, I_0. Notice that if $I = 0$, then $\delta = 2\pi n$; thus a bias current must flow through the junction for a physically meaningful phase difference to exist. The ac Josephson equation relates the voltage across the junction (once $I > I_0$) to the temporal derivative of δ.

To build a SQUID, the superconductors on either side of the junction must be joined, either by a loop with a second junction (dc SQUID, shown in Fig. 2a) or by a continuous superconducting loop (rf SQUID, shown in

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

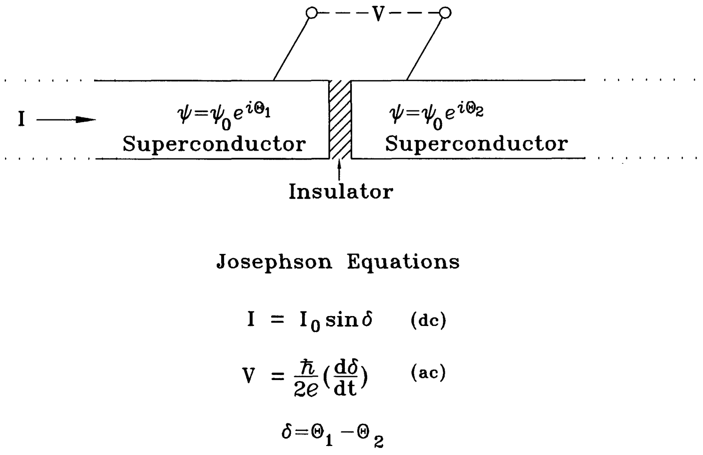
V
I
$\psi=\psi_0 e^{i\Theta_1}$
Superconductor
$\psi=\psi_0 e^{i\Theta_2}$
Superconductor
Insulator
Josephson Equations
$I = I_0 \sin\delta$ (dc)
$V = \frac{\hbar}{2e}\left(\frac{d\delta}{dt}\right)$ (ac)
$\delta=\Theta_1-\Theta_2$

FIG. 1. The Josephson junction and Josephson equations. A superconductor is interrupted by a thin insulating layer (cross-hatching). The phase of the superconducting order parameter ψ becomes a function of the current flow through the junction, which leads to the unique electrical properties of the junction.

Fig. 2b). A weakly connected superconducting loop is thus formed. The loop will contain flux only in multiples of the flux quantum $n\Phi_0$, where n is any integer and $\Phi_0 = 2.07 \times 10^{-15}$ Wb (see DIAMAGNETISM). Typically, the SQUID incorporates a resistor in parallel with the Josephson junction(s) to prevent hysteresis in the I-V characteristic. The equivalent circuit of the ideal, noise-free, resistively shunted junction (RSJ) consists of the Josephson junction, its physical capacitance, and the resistor, all joined in parallel, as shown in Fig. 2. Standard circuit analysis may be applied to the SQUID equivalent circuit, using Josephson's equations for the junction, and the resulting differential "equations of motion" can be solved numerically. This approach leads to the so-called washboard-potential model.

1.2 The dc SQUID

In the dc SQUID, a change in applied flux, Φ_A, leads to a phase difference across the junctions, giving rise to a voltage across the loop that we may detect. The characteristic I-V curves of the dc SQUID are shown in Fig. 3(a) for $\Phi_A = n\Phi_0$ and $\Phi_A = (n + \frac{1}{2})\Phi_0$, where n is any integer. If the bias current is held constant, the SQUID voltage will vary between two values, $V_{\min}$ and $V_{\max}$, as the flux applied to the SQUID varies between $n\Phi_0$ and $(n + \frac{1}{2})\Phi_0$. The V-Φ_A curve is thus roughly sinusoidal, as shown in Fig. 3(b), with period Φ_0. Figure 3(c) shows that a small flux modulation, $\Phi_m \sin(\omega t)$, where $\Phi_m \leq \Phi_0/2$, applied to the SQUID operating at three different points, A, B, and C, on the V-Φ_A curve will elicit three different voltage responses, V_s. At points A and C, V_s is at the same frequency as Φ_m, while at point B, where $\Phi_A = (n/2)\Phi_0$, V_s is at a frequency of 2ω. Thus the SQUID acts as a nonlinear flux-to-voltage transducer.

To linearize the V-Φ_A curve, we introduce the flux-locked loop (FLL) in Fig. 4. The FLL maintains lock by keeping the system at $\Phi = (n/2)\Phi_0$, one of the extrema in the V-Φ curve. A modulating flux $\Phi_m(\omega)$, where $\Phi_m \leq \Phi_0/2$, is applied to the SQUID, at frequency ω, by the oscillator. The response of the SQUID is fed into a lock-in amplifier, referenced to the oscillator. If the dc flux in the SQUID is a multiple of $(n/2)\Phi_0$, then the output of the SQUID is a periodic function of 2ω and the lock-in amplifier will put out 0 V dc. If the dc flux strays from an extremum by $\Delta\Phi$, then the SQUID output will contain a component at ω and the lock-in will put out a dc voltage, V_L, proportional to the amplitude of the signal at ω, as shown in the third line of Fig. 3(c). This dc signal is essentially an error signal; it is integrated and fed back into the modulation coil as $-\Delta\Phi$, returning the SQUID flux to $(n/2)\Phi_0$. The output of the loop, V_0, is a voltage proportional to the feedback current, controlled by V_L (which is proportional to $\Delta\Phi$.) The user records V_0 and relates that to the magnetic field through a prior calibration with a known field. The reader should note that many FLL oscillators put out a square-wave flux modulation with amplitude $\Phi_0/4$ and frequency ω. The system then shifts between $\Phi = (n/2 + \frac{1}{4})\Phi_0$ and $\Phi = (n/2 - \frac{1}{4})\Phi_0$, still centered on an extre-

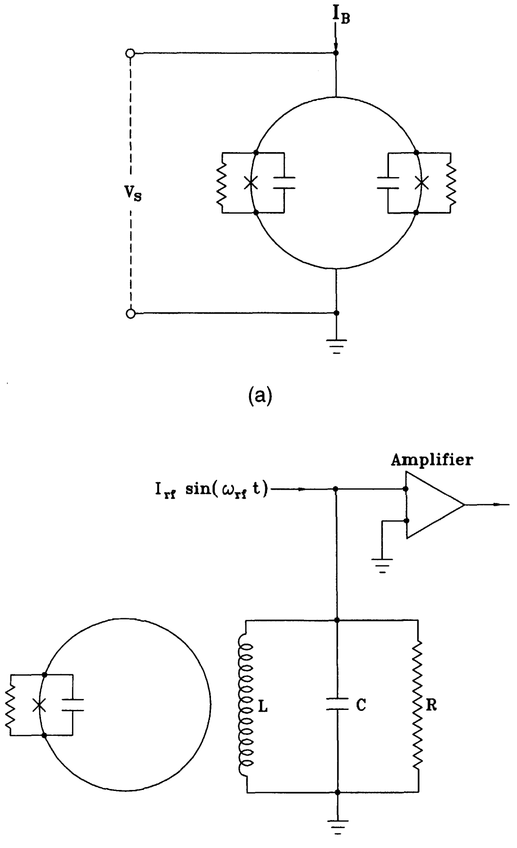

(a)

(b)

FIG. 2. (a) Schematic of a dc SQUID loop. The superconducting loop is interrupted by two Josephson junctions, marked by x. The dc SQUID is biased with a dc current. (b) Schematic of an rf SQUID loop. The rf SQUID has one Josephson junction. The rf SQUID is coupled to an rf tank circuit. The Josephson junctions are each shunted by a resistor. The physical capacitance of the junction is shown as well, because it must be considered in an accurate model of the electrical properties of the SQUID.

mum but spending very little time at the extremum.

The flux-locked SQUID has a bandwidth that is some fraction of the modulation frequency, typically 100-kHz modulation and 10-kHz bandwidth. If the applied flux changes too rapidly for the feedback electronics to track, then the SQUID jumps from one extremum to another, and the SQUID is said to have lost lock. The maximum change in flux per unit time that the system can tolerate while maintaining lock is the slew rate, typically $\approx 10^6 \Phi_0$/s.

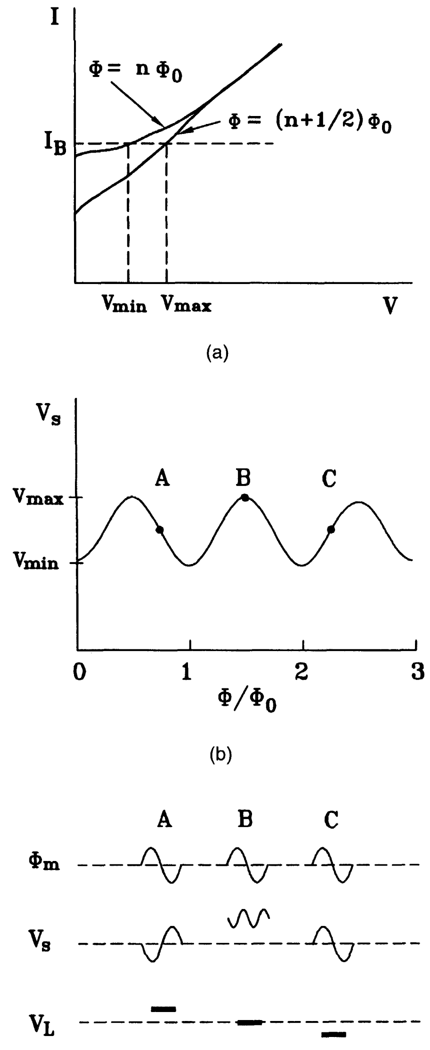

(a)

(b)

(c)

FIG. 3. (a) The *I-V* characteristics of a dc SQUID. The amount of applied flux, Φ_A, determines the voltage output V_s for a particular value of bias current I_B. As the applied flux varies between $\Phi_A = n\Phi_0$ and $\Phi_A = (n + 1/2)\Phi_0$, the output voltage changes between V_{min} and V_{max}. (Adapted from Clarke, 1989.) (b) The *V-Φ* curve of a dc SQUID, with constant bias current. (c) The voltage response of the SQUID to a modulating flux, Φ_m. The response varies greatly depending on the value of Φ_A. Three possible points are highlighted to illustrate the response of a SQUID and the feedback needed to "lock" the SQUID to operation at an extreme value. From top to bottom we see one cycle of the flux modulation, Φ_m, applied to the SQUID operating at three different locations on the *V-Φ* curve; the voltage response of the SQUID, V_s; and the necessary feedback of the flux-locked loop, V_L, which applies a counter flux to the SQUID loop (see text), returning the system to an extreme position on the V-Φ_A curve.

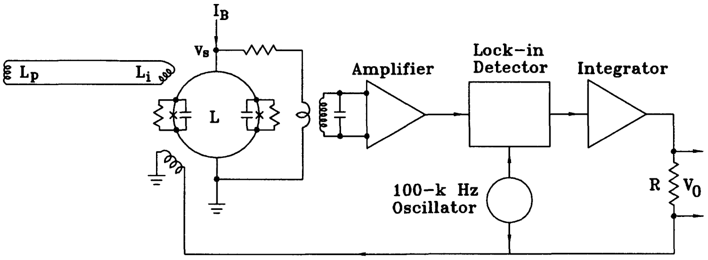

FIG. 4. Schematic of the pickup coil, input coil, SQUID and flux-locked loop (adapted from Clarke, 1989). A change in magnetic field at the pickup coil, with inductance L_p, induces a change in field at the input coil, with inductance L_i; the SQUID is inductively coupled to the input coil, and it detects a change in magnetic flux. The voltage response of the SQUID, V_s, is then fed into the flux-locked loop. The loop uses a modulating flux scheme, explained in the text, to maintain the SQUID at an extremum on the V-Φ curve. The user measures a change in V_0, which is proportional to the feedback from the loop and to the change in magnetic field at the pickup coil.

1.3 The rf SQUID

Conceptually, the operation of the rf SQUID is very similar to that of its dc counterpart. The rf SQUID is inductively coupled to a tank circuit, as shown in Fig. 2b, being driven at a frequency $f_{\mathrm{rf}} \approx$ 20–30 MHz. If the magnitude of the radio-frequency current in the tank circuit, I_{rf}, is kept constant, then the voltage across the circuit, V_{rf}, will be periodic with any flux applied to the SQUID. This holds true because of the hysteretic nature of flux in the SQUID loop. The V_{rf}-Φ curve is then linearized with a FLL in a manner similar to the dc SQUID. In this case, the modulating flux is applied through the inductor in the tank circuit.

1.4 The Pickup Coil

The SQUID is an excellent sensor of magnetic flux. To enhance its capabilities, most SQUID systems do not expose the bare SQUID to the magnetic field of interest; rather they typically employ a multiple-turn pickup coil inductively linked to the SQUID, as shown in Fig. 4. The pickup coil, with inductance L_p, is exposed to the environment, while the input coil, with inductance L_i, and the SQUID, with inductance L, are shielded from the ambient field by a superconducting niobium canister. Typical values for the inductances are $L_p = L_i = 1$ μH and $L = 0.1$ nH. Historically, all pickup coils were hand wound of superconducting wire, but in some thin-film configurations, the pickup coils are fabricated on the same chip as the SQUID or an adjacent "flip chip" using integrated circuit technology.

1.5 Noise in the SQUID

SQUID noise is quoted in different units by different members of the SQUID community. The given noise figure is generally a power spectral density in terms of energy (J/Hz), magnetic flux (Φ_0/Hz$^{1/2}$), field (fT/Hz$^{1/2}$) or field gradient [fT/(cm Hz$^{1/2}$)]. The distinction is important as one system (of special design) can have poor field sensitivity, despite superior energy sensitivity in the SQUIDs it contains. The energy and flux noise are figures of merit (FOM) for the bare SQUID, while the field and field-gradient noise are FOMs for the complete SQUID system. As a rule of thumb, the SQUID researcher quotes energy or flux noise, while the SQUID system builder begins with the flux figure and designs for the best possible field or field-gradient noise, which is of paramount interest to the end user. Typical orders of magnitude for commercial dc SQUIDs are noise energy $\epsilon = 10^{-31}$ J/Hz; magnetic flux noise $S_{\Phi}^{1/2} \sim 10^{-6}$ Φ_0/Hz$^{1/2}$; and magnetic field noise $B_N = 10$ fT/Hz$^{1/2}$.

Approximate formulas for the dc SQUID noise parameters are

$$S_{\Phi} = \frac{(4k_{\mathrm{B}}TR_D^2/R)[1 + \frac{1}{2}(I_0/I)^2]}{(R/L)^2}, \qquad (1)$$

$$\epsilon = \frac{S_\Phi}{2L}, \quad B_N = \frac{L_p + L_i}{\alpha\sqrt{L_i L}} \frac{S_\Phi^{1/2}}{A_e}, \tag{2}$$

where R_D is the dynamical resistance ($\partial V/\partial I$) of the junctions, A_e is the effective area of the pickup coil, and α is the coupling coefficient between the input coil and the SQUID. In discussing SQUID systems, the coupled energy sensitivity, ϵ/α^2, is often quoted rather than ϵ.

In the case of the rf SQUID, things are a bit more complex. The flux noise spectral density is given by

$$S_\Phi \approx \frac{(LI_0)^2}{\omega_{\mathrm{rf}}} \left(\frac{2\pi k_B T}{I_0 \Phi_0}\right)^{4/3}, \tag{3}$$

and the energy and field resolution have additional terms due to the room-temperature electronics.

Flicker or $1/f$ noise does exist in thin-film SQUIDs, but only below ~0.1 Hz does it begin to dominate in conventional niobium SQUIDs. At the current time, low-T_c dc SQUIDs are considerably less noisy than rf SQUIDS, except for rf SQUIDs operating at microwave frequencies. Thus most end users who need superior field resolution utilize dc SQUIDs. The rf SQUIDs dominated the market prior to the introduction of thin-film dc SQUIDs in the late 1970s; rf SQUIDs still command a substantial market share today.

1.6 The Digital SQUID

Digital SQUIDs attempt to replace the conventional room-temperature analog electronics of the FLL with either a room-temperature digital signal processor (DSP) or an on-chip (4.2 K) integrated Josephson-junction feedback circuit that involves flux counting. The DSP circuits are being used successfully by a growing number of research groups; the integrated digital SQUID is promising, but the technology is still immature and underperforms conventional systems. However, the interest in further development is spurred by the increasing number of channels being used in neuromagnetism. On-chip feedback would eliminate a great deal of the cost of electronics in systems of 100+ SQUIDs and allow for simpler cryogenic design as well, particularly by drastically reducing the number of wires between room temperature and the SQUIDs.

2. DESIGN AND MANUFACTURE OF SQUIDS

A single SQUID is not a particularly difficult item to manufacture. Consistently producing low-noise SQUIDs that integrate ideally into instrumentation is the technical challenge. Here we discuss the early SQUIDs, the modern thin-film SQUIDs, high-T_c SQUIDs, and the design considerations for manufacture.

2.1 Conventional SQUIDs

Some of the earliest SQUIDs were machined from blocks of niobium. Figure 5 shows two of these designs. Figure 5(a) depicts the simple "point-contact" dc SQUID (adapted from Zimmerman and Silver, 1966). The sharpened niobium screws could be adjusted, once the SQUID was immersed in liquid helium, until the proper I-V characteristic was obtained. Simple designs like this achieved flux resolution of $\sim 10^{-3}\Phi_0/\mathrm{Hz}^{1/2}$ in the late 1960s. Figure 5(b) shows a cross section of the very useful "toroidal" rf SQUID (Goodman *et al.*, 1973; Rifkin and Deaver, 1976). The smaller toroidal cavity contained the tank circuit, and the larger cavity contained the input coil, which is connected to the pickup coil. This self-screening design was a commercial success, and extensive work was done with this SQUID in biomagnetism during the late 1970s and early 1980s. The flux resolution was at least an order of magnitude better than contemporary dc SQUIDs, and the SQUID noise was white between 1 Hz and several kHz.

The modern thin-film SQUIDs are typically manufactured by sputtering of niobium thin films and patterning the films with photolithography or electron-beam lithography (see SPUTTERING; THIN-FILM DEPOSITION). In principle, any superconducting material can be used to manufacture SQUIDs, but in practice, niobium technology completely dominates the low-T_c commercial SQUID market. The junction itself may be a Nb/NbO$_x$/Pb or Nb/Al$_2$O$_3$/Nb trilayer, while the resistive shunts can be a thin film of copper, gold, or nearly any metal that does not be-

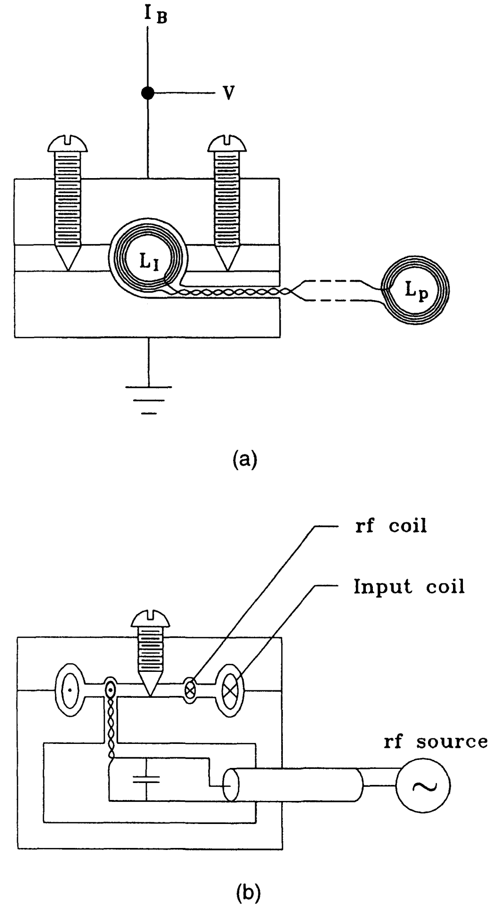

FIG. 5. Early SQUID designs machined from niobium. The Josephson junctions were made by adjusting the screws in liquid helium until the expected behavior was achieved. (a) The dc SQUID. The input coil, with inductance L_i, is linked to the pickup coil located near the magnetic sample. (b) The rf "toroidal" SQUID. The inner toroidal cavity contains the coil for the rf circuit, while a separate chamber houses the capacitor and the outer cavity holds L_i.

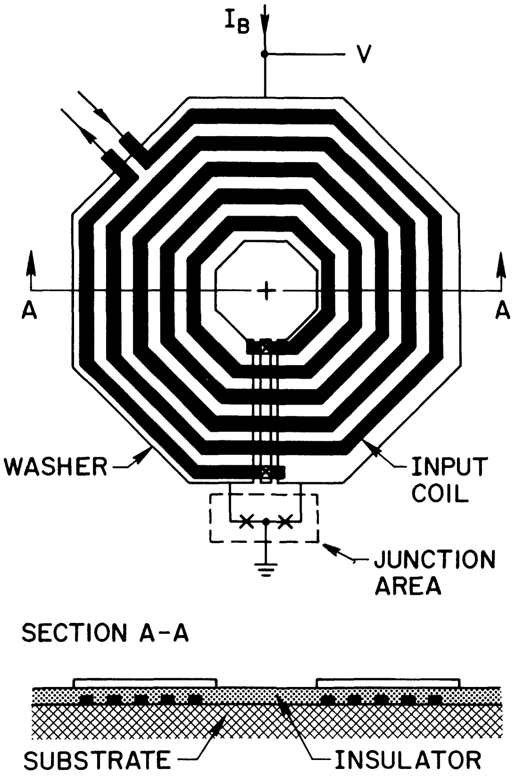

FIG. 6. A schematic drawing of a washer-type dc SQUID. The design has a very high mutual inductance between the input coil and the SQUID loop. The input coil, marked in black, is manufactured with the SQUID loop, which consists of the octagonal washer and the Josephson junctions, each marked by an x. Note that the washer is not a complete octagon; it is interrupted in two places, and the resulting bridge is used to complete the input coil. A cross section through the SQUID is shown at the bottom of the figure. The pickup coil, not shown, is connected to the input coil by the two leads in the upper left. In niobium, low-T_c technology, the entire structure is less than 1 mm across, while the junctions can be as small as 5 μm wide; the hole in the washer is typically 50–100 μm in diameter. (Courtesy of Mark Ketchen of IBM T.J. Watson Research Center.)

come superconducting at 4 K. Figure 6 shows an advanced design of the washer-type dc SQUID developed at IBM (Jaycox and Ketchen, 1981). This geometry provides SQUID loops with low inductance and hence high sensitivity, while allowing efficient coupling to an external flux transformer. Other dc SQUID designs have also been implemented for commercial SQUIDs.

The two characteristic physical parameters of a SQUID most often quoted in the literature are the reduced inductance β and the hysteresis parameter β_c, sometimes referred to as the Stewart–McCumber parameter:

$$\beta = \frac{2I_0L}{\Phi_0}, \quad \beta_c = \frac{2\pi I_0R^2C}{\Phi_0}, \qquad (4)$$

where R and C are the resistance and capacitance of the resistively shunted Josephson junction and L is the inductance of the SQUID loop. These parameters emerge quite naturally in an analysis of the differential "equations of motion" that govern operation of the dc SQUID. We see that these parameters are dimensionless, and that β is a function of the physical dimensions of the loop and the critical current of the junction, while β_c is a function of the resistively shunted junction alone. Numerical simulations of SQUIDs with various parameters have shown that for minimum noise energy β should approach unity. To avoid hysteresis β_c should be at or below 1. These conditions still give the manufacturer quite a bit of freedom, which may be used to design the SQUIDs for optimum *system* performance. To optimize the system performance, the manufacturer must achieve the highest possible signal-to-noise ratio for the signal of interest to the user in the environment of the user facility. Therefore, the system must be designed with foreknowledge of the signal of interest and the operating environment. The SQUID should be designed to couple as well as possible with the input coil, which should be impedance matched with the desired pickup coil (see below). The pickup coil must be designed to reject noise fields and to couple efficiently to the signal of interest. The designed SQUID noise should at all times be less than that of the amplifier used in the flux-locked loop. There is no single prescription for optimum performance, and SQUID design is still something of an art.

2.2 High-T_c SQUIDs

High-temperature superconductivity holds one great promise for SQUID builders and users: It will make the construction and operation of SQUID systems cheaper. The niobium SQUIDs described so far generally operate at liquid-helium temperature (4.2 K); the high-T_c systems being designed today operate at liquid-nitrogen temperature (77 K). The lower cost and greater cooling capacity of liquid nitrogen saves roughly a factor of 10^3 in the operating cost for systems in metropolitan areas and far more for remote systems. Construction costs will also be reduced as liquid-nitrogen systems are simpler and easier to build (see CRYOGENICS). The ultimate high-T_c SQUID performance will probably be inferior to the low-T_c counterpart. This is evident from the equations above for flux noise, since in dc and rf SQUIDs the flux noise scales with the temperature, i.e., by a factor of ~20.

Thin-film SQUIDs have been constructed from all the main families of high-T_c compounds: Y-Ba-Cu-O, Bi-Sr-Ca-Cu-O, and Tl-Ba-Ca-Cu-O. The material that has come to dominance is $YBa_2Cu_3O_{7-x}$. Josephson junctions have been manufactured in a number of ways using this material. The two main types of junction are grain-boundary and superconducting-normal-superconducting (SNS) junctions, i.e., $YBa_2Cu_3O_{7-x}$-N-$YBa_2Cu_3O_{7-x}$ junctions, where N represents a normal metal such as gold or silver. At the current state of the technology, the best grain-boundary SQUIDs have the lowest noise, but the best SNS junction SQUIDs are more rugged, reproducible, and manufacturable. The main difficulties with virtually all high-T_c SQUIDs fabricated to date are high $1/f$ noise below 1 Hz and poor performance when operated in weak environmental magnetic fields such as that of the Earth. The high $1/f$ noise precludes many applications in medicine and biophysics where the signals of interest have component frequencies as low as 0.05–1 Hz. Extreme sensitivity to environmental noise is a serious problem in the development of high-T_c SQUID systems, primarily because high-T_c magnetic shields and wire do not perform as well as the equivalent niobium components. This is likely to be a problem for some time, as all known high-T_c compounds are extreme type-II superconductors, which lose perfect diamagnetism at very low critical fields, H_{c1}. Electronic synthetic gradiometer systems, which incorporate multiple SQUIDs, are now being developed to allow high-T_c systems to function in noisy environments (Koch *et al.*, 1993). However, high-T_c SQUIDs already demonstrate lower noise than fluxgate magnetometers and may soon see commercial application in geophysics and nondestructive testing. The best high-T_c SQUIDs are getting better, and the optimum high-T_c SQUID is still to be made.

3. SQUID SYSTEMS

The vast majority of SQUID applications are concerned with the measurement of a magnetic field, in which the phenomenon of interest (usually a current, magnetization, or magnetic susceptibility distribution) generates a magnetic field or distorts an applied magnetic field. SQUID magnetometers, gradiometers, and susceptometers can be readily optimized for specific measurements. A typical SQUID instrument is depicted in Fig. 7. We see that the SQUIDs are located inside a small magnetic shield (e.g., superconducting niobium). Superconducting pickup coils are located at the bottom of the Dewar; the SQUID electronics are at room temperature, but close to the Dewar; and the magnetic object is placed beneath the instrument. If the experiment is conducted in an applied magnetic field, then the instrument is a susceptometer.

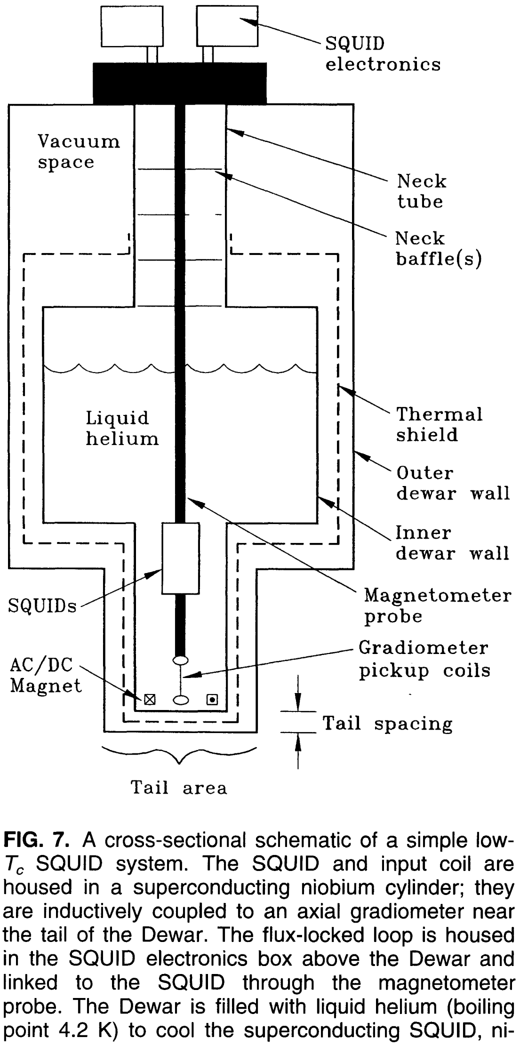

FIG. 7. A cross-sectional schematic of a simple low-T_c SQUID system. The SQUID and input coil are housed in a superconducting niobium cylinder; they are inductively coupled to an axial gradiometer near the tail of the Dewar. The flux-locked loop is housed in the SQUID electronics box above the Dewar and linked to the SQUID through the magnetometer probe. The Dewar is filled with liquid helium (boiling point 4.2 K) to cool the superconducting SQUID, niobium canister, and pickup coils. The Dewar is insulated by a vacuum space between the inner and outer walls. The instrument may be used to sense the intrinsic magnetic field of a sample or, with the magnet energized, it can measure the response of the sample to an ac or a dc magnetic field. (Adapted from Fagaly, 1990.)

3.1 Gradiometers

Although a simple single- or multiple-loop pickup coil (a magnetometer coil) can be used to measure one component of the magnetic field, such as B_z, there are a number of disadvantages to this approach, the most important of which are the contamination of the signal by environmental noise and extreme sensitivity to tilt in the Earth's magnetic field. Since the field of interest is generally orders of magnitude smaller than power-line noise, fields from passing vehicles, and radio-frequency interference, some means of background field rejection is required. This can be done either by shielding the entire experiment (as described in the next subsection), or with gradiometry. Gradiometers take advantage of the mathematical form of the falloff of magnetic fields with distance from the source. The local source of interest generates a much larger field gradient at the detector than does the more distant noise source, even though the absolute noise field is larger. Thus, by configuring coils to sense the magnetic field at two or more locations, it is possible to discriminate against the distant sources and in favor of the local one. The input coil is linked in series with two or more pickup coils, which are outside the SQUID's niobium shield. For example, we may sense $B_z(z_1)$ with one loop of the gradiometer coil and $B_z(z_2)$ with a second loop wound in the opposite sense, and the SQUID reports a voltage proportional to $\Delta B_z/\Delta z$. Commercial magnetometers are generally supplied with a hermetically sealed, shielded SQUID and two input terminals so that custom-designed gradiometers appropriate to the particular experiment and magnetic environment can be connected (see Fig. 8) (Wikswo, 1978).

3.2 Shielded Rooms

In hostile environments (such as hospitals and research facilities), the noise rejection achieved by a gradiometer may be insufficient. In these cases, the most reliable

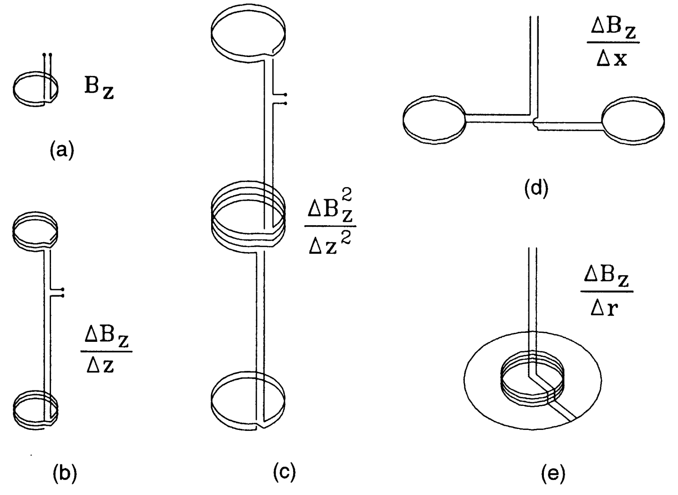
B_z
(a)
$\frac{\Delta B_z}{\Delta z}$
(b)
$\frac{\Delta B_z^2}{\Delta z^2}$
(c)
$\frac{\Delta B_z}{\Delta x}$
(d)
$\frac{\Delta B_z}{\Delta r}$
(e)

FIG. 8. A variety of pickup coil designs and the physical quantity they measure. (a) Magnetometer. (b) First-order axial gradiometer. (c) Second-order axial gradiometer. (d) First-order planar gradiometer. (e) First-order radial gradiometer.

(though expensive) solution is to carry out measurements inside a magnetically shielded room. A closed shell (access ports can be designed that do not significantly compromise the shielding) consisting of welded aluminum plates can provide eddy-current shielding of about 100 dB above 1 Hz. At low frequencies, however, structures made from high-permeability, soft ferromagnetic materials such as Mumetal or Permalloy are required. Many practical shielded rooms employ both methods.

3.3 Cryogenic Dewar Design

Liquid helium is required to maintain conventional (low-T_c) SQUIDs and input circuits in the superconducting state. Although some continuous-flow cryostats have been used, the superconducting components are generally immersed in a helium reservoir, which is thermally insulated with a vacuum jacket containing carefully designed radiation shields. The expense of liquid helium makes the boiloff rate an important consideration. Dewars are constructed with fiberglass-reinforced epoxy, which has been found to have suitable structural and thermal properties without introducing Johnson noise, distorting uniform noise fields, or screening the magnetometer from low-frequency signals. Most magnetometer measurements can be improved by minimizing the separation between the pickup coils and the room-temperature environment, and this has led to Dewars in which the superconducting components are in the vacuum space, thermally coupled to the helium. For neuromagnetic measurements, helmet-shaped Dewars containing large numbers of SQUID channels are used, while tubular Dewars, which allow a sample to be passed through the plane of the pickup coil, are popular for low-temperature and geomagnetic research.

Dewar design (see CRYOGENICS) is critical to recently developed SQUID imaging systems (Wikswo *et al.*, 1990). These high-resolution SQUID systems are designed to measure the magnetic field very close to the surface of quasi-two-dimensional objects. The resolution of the images is determined by the size of the pickup coil and its distance from the magnetic source: For high-resolution SQUIDs, magnetic sources in living tissue that are separated by as little as a millimeter can be discriminated. Some SQUID imaging systems are designed for samples, notably superconducting thin films or integrated circuits, that can be located within the cryogenic space, so that resolution of 10 μm or better is possible.

4. APPLICATIONS

4.1 Magnetoencephalography

The human brain, consisting of some 10^9 electrically active neurons, has been at the focus of biomagnetism research ever since it became apparent that SQUID magnetometers had sufficient sensitivity to measure the tiny neuromagnetic fields (10–100 fT). The measurement, named magnetoencephalography (MEG), supplies information about the same physical process as its counterpart, the electroencephalogram (EEG). The widely cited advantage of MEG is that the magnetic field generated by cortical neurons is not distorted by the intervening bone and tissue, as are the electrical voltages recorded by the scalp EEG. The primary disadvantage is the capital and operational cost of the MEG instrumentation. Most researchers agree that a complementary approach, using both the MEG and EEG, is optimum. One of the earliest clinical MEG applications was to epilepsy. It was shown by Modena *et al.* (1982) and Barth *et al.* (1982) that interictal spikes, occurring between seizures, could sometimes be detected magnetically but not electrically. Furthermore, the magnetic signals obtained often allowed localization of the source, which could be confirmed intraoperatively. These measurements were carried out using single-channel magnetometers and were unable to characterize asynchronous events occurring simultaneously at multiple brain locations. It became clear that progress would be extremely slow until instruments were developed that could simultaneously measure the magnetic field at a large number of sites.

Commercial MEG multi-SQUID systems have evolved from 7- and 37-channel units introduced in the late 1980s and early 1990s to 128-, 122-, and 148-channel helmet systems being produced by companies in Canada, Finland, and the United States. Many of the recent MEG studies have involved an evoked response, in which a characteristic cortical response is stimulated by an external stimulus supplied to the subject. By measuring the magnetic field associated with the response, the functionality of the visual, auditory, and somatosensory areas of the cortex are mapped. Additionally, higher cortical function, such as recognition and attention, may be investigated. Another area of study is spontaneous brain activity, such as the alpha rhythm, during waking and sleeping. MEG is also used in the assessment of sensory pathways between the periphery and the brain, and in the study of stroke, multiple sclerosis, and migraine.

4.2 Magnetocardiography

The first measurement of the magnetic field of the human heart was made in 1963 using room-temperature induction coils of several million turns (Baule and McFee, 1963). Just a few years later, the SQUID sensor became available, enabling a huge improvement in signal-to-noise ratio and stimulating worldwide interest in magnetocardiography (MCG) research. As with MEG, it is felt that combined magnetic and electric measurements are more useful than a single technique alone (Stroink *et al.*, 1992). Clinical applications such as mapping the sources of arrhythmias, ventricular tachycardia, and Wolff-Parkinson-White syndrome may benefit from the development of multichannel instruments. High-resolution SQUIDs have also been used in experimental studies of cardiac tissue slices (Staton *et al.*, 1993), which have contributed to the development of mathematical models of cardiac electrophysiology, and in studies of cortical activity (Okada *et al.*, 1992).

4.3 Other Biomagnetic Studies

SQUID magnetometers have also been used in a number of other biomagnetic applications. By applying a magnetic field to the liver and scanning it with a magnetometer, a susceptibility signal is measured that may be used to estimate the level of stored iron. Unlike the alternative, needle biopsy, this technique is noninvasive and may prove useful in the assessment of hemochromatosis. Magnetic tracers can be used in conjunction with SQUID magnetometers to measure gastrointestinal transit times and in the study of the internal dynamics of cells.

4.4 Geomagnetism

The geomagnetic work done with SQUIDs is primarily divided into paleomagnetism and prospecting. Paleomagnetism is the study of the magnetization of ancient rock for the purpose of understanding the geological history of the region from which the rock was excavated. Prospecting is the search for valuable ore or oil beneath the earth. This may be done by applying a low-frequency electromagnetic field to the earth and measuring the magnetic and electric response. By probing at different frequencies, it is possible to determine for the region a profile of the conductivity as a function of depth. This profile is then correlated with geological data to determine the composition of the various strata. Geological mapping of the magnetization of the earth may prove very useful in other ways. In particular changes in the maps over time may be useful in the prediction of earthquakes. Increased stress at the junction of tectonic plates could produce a measurable change in the magnetization of the region.

4.5 Nondestructive Testing

One new field of SQUID research that has attracted considerable interest is nondestructive evaluation (NDE). The NDE SQUID magnetometry efforts of the past decade include flaw characterization, analysis of magnetic properties of materials, and corrosion study. In most of these areas, a measurement of the magnetic field is made at multiple locations in the vicinity of a test object, and a map or image of the field is recorded. Magnetic field sources can be intrinsic in the test object as in the case of galvanic currents flowing in a corroding pipe, or may be excited by an external source. The external field excitation can be done either magnetically or electrically. With magnetic excitation, the test object is placed within a relatively strong magnetic field, whereas a dc or an ac current is established in the sample for electrical excitation using the current injection or current induction method. In the injection method, the metal is directly connected to a current source; for current induction, eddy currents are generated within the inspected region by a coil, wire, or plate carrying an ac current.

The advantages of SQUID NDE are high sensitivity, wide bandwidth (from dc to 10 kHz), and a large dynamic range. The high sensitivity of the SQUID allows one to make engineering and design compromises to reduce cost, enhance signal strength for particular types of flaws, and still remain orders of magnitude more sensitive than other means of magnetic anomaly detection. The ability of SQUID detectors to function down to zero frequency allows them to sense much deeper flaws than traditional eddy-current sensors and to detect and monitor the flow of steady-state corrosion currents. Demonstration experiments have been performed on a wide range of samples, from tracing current in a printed circuit to detecting parasites in fish, but no clear market niche has yet been established.

4.6 Other Applications

It is often said that there are as many SQUID applications as there are SQUID researchers; this is most likely a conservative estimate. This section attempts to show something of the breadth of SQUID research.

There are a number of applications of SQUIDs that do not involve an entire SQUID magnetometer system. A SQUID can be configured to perform as a sensitive picovoltmeter (Clarke, 1966) for a variety of applications, including the measurement of thermoelectric voltages and the study of superconductor mechanisms. The sensitivity of these devices can be on the order of tens of $\mathrm{fV/Hz^{1/2}}$. SQUIDs are also used in gravity gradiometers (Paik, 1981) and gravitational-wave detectors (Michelson *et al.,* 1987), where they are configured to sense minute displacements. There has been some development of SQUIDs as radio-frequency amplifiers (Hilbert and Clarke, 1985) and as magnetic resonance detectors (Hilbert *et al.,* 1985). Finally, SQUIDs have been used for a number of fundamental physical measurements, including an investigation into the possible existence of fractional flux quantization (Cabrera *et al.,* 1989), measurement of the Cooper pair mass (Cabrera and Peskin, 1989), and the search for magnetic monopoles (Huber *et al.,* 1990).

GLOSSARY

Bare SQUID: A SQUID operated in the ambient magnetic field, with no magnetic shielding. Typical for the current high-T_c SQUIDs.

Critical Current: The maximum current a superconductor can carry without dissipating energy.

Flicker Noise: Also known as $1/f$ noise. Noise, present in nearly all electrical circuits, with a characteristic rise in amplitude with decreasing frequency.

Fluxoid: The quantum unit of flux, $\Phi_0 = 2.07 \times 10^{-15}$ Wb.

High-T_c Superconductor: Any of a family of ceramic superconductors, discovered in the 1980s, with transition temperatures as high as 133 K.

Magnetocardiography: The study of the magnetic field of the heart.

Magnetoencephalography: The study of the magnetic field of the brain.

Nondestructive Evaluation: The examination of materials or components, without impairing their future usefulness, in order to detect defects and determine serviceability.

SQUID Triangles: The roughly sinusoidal curve of voltage vs applied flux, V_s-Φ_A, of the current-biased bare SQUID.

Tank Circuit: A resonant *L-C* circuit.

Tuning the SQUID: Adjusting the bias current I_B to achieve the largest voltage response V_s for a small change in applied flux, $\Delta\Phi \leq \frac{1}{2}\Phi_0$.

Works Cited

Barth, D. S., Sutherling, W., Engel, J., Jr., Beatty, J. (1982), *Science* **218,** 891–894.

Baule, G., McFee, R. (1963), *Am. Heart J.* **66,** 95–96.

Cabrera, B., Cunningham, C. E., Saroff, D. (1989), *Phys. Rev. Lett.* **62,** 2040–2043.

Cabrera, B., Peskin, M. E. (1989), *Phys. Rev. B* **39,** 6425–6430.

Clarke, J. (1966), *Philos. Mag.* **13,** 115–127.

Goodman, W. L., Hesterman, V. W., Rorden, L. H., Goree, W. S. (1973), *Proc. IEEE* **61,** 20–27.

Hilbert, C., Clarke, J. (1985), *J. Low Temp. Phys.* **61,** 263–280.

Hilbert, C., Clarke, J., Sleator, T., Hahn, E. L. (1985), *Appl. Phys. Lett.* **47,** 637–639.

Huber, M. E., Cabrera, B., Taber, M. A., Gardner, R. D. (1990), *Phys. Rev. Lett.* **64,** 835–838.

Jaycox, J. M., Ketchen, M. B. (1981), *IEEE Trans. Magn.* **MAG-17,** 400–403.

Josephson, B. D. (1962), *Phys. Lett.* **1,** 251–253.

Koch, R. H., Rozen, J. R., Sun, J. Z., Gallagher, W. J. (1993), *Appl. Phys. Lett.* **63,** 403–405.

Okada, Y. C., Kyuhou, S., Lähteenmäki, A., Xu, C. (1992), in: M. Hoke, S. N. Erné, Y. C. Okada, G.-L. Romani (Eds.), *Biomagnetism: Clinical Aspects,* New York: Elsevier, pp. 375–383.

Modena, I., Ricci, G. B., Barbanera, S., Leoni, R., Romani, G. L., Carelli, P. (1982), *Electroencephalography and Clinical Neurophysiology* **54,** 622–628.

Paik, H. J. (1981), in: H. Weinstock, W. C. Overton (Eds.), *SQUID Applications to Geophysics,* Tulsa: Society of Exploration Geophysicists, pp. 3–12.

Rifkin, R., Deaver, B. S. (1976), *Phys. Rev. B.* **13,** 3894–3901.

Staton, D. J., Friedman, R. N., Wikswo, J. P., Jr. (1993), *IEEE Trans. Appl. Supercond.* **3,** 1934–1936.

Stroink, G., Lant, J., Elliot, P., Lamothe, R., Gardner, M. (1992), in: M. Hoke, S. N. Erńe, Y. C. Okada, G.-L. Romani (Eds.), *Biomagnetism: Clinical Aspects,* New York: Elsevier, pp. 471–474.

Wikswo, J. P., Jr. (1978), in: B. S. Deaver, Jr., C. M. Salko, J. H. Harris, S. A. Wolf (Eds.), *Future Trends in Superconductive Electronics,* AIP Conference Proceedings No. 44, pp. 145–149.

Wikswo, J. P., Jr., van Egeraat, J., Ma, Y. P., Sepulvada, N., Staton, D. J., Tan, S., Wijesinghe, R. S. (1990), in: A. F. Gmitro, P. S. Idell, I. J. LaHaie (Eds.), *Digital Image Synthesis and Inverse Optics,* SPIE Proceedings No. 1351, Bellingham, WA: SPIE, pp. 438–470.

Zimmerman, J. E., Silver, A. H. (1966), *Phys. Rev.* **141,** 367–375.

Further Reading

General

Clarke, J. (1989), in: H. Weinstock, M. Nisenoff (Eds.), *Superconducting Electronics,* Berlin: Springer-Verlag, pp. 87–148.

Simon, R., Smith, A. (1988), *Superconductors: Conquering Technology's New Frontier,* New York: Plenum Press.

Tinkham, M. (1975), *Introduction to Superconductivity,* New York: McGraw-Hill.

Applications

Donaldson, G. B. (1989), in: H. Weinstock, M. Nisenoff (Eds.), *Superconducting Electronics,* Berlin: Springer-Verlag, pp. 175–207.

Fagaly, R. L. (1990), in: S. Sato (Ed.), *Advances in Neurology* Vol. 54, New York: Raven Press, pp. 11–32.

Swithenby, S. J. (1980), *J. Phys. E: Sci. Instrum.* 801–813.

Wikswo, J. P., Jr. (1995), *IEEE Trans. Appl. Supercond.* **5,** 74–120.

STARK EFFECT

See ZEEMAN AND STARK EFFECTS

STATISTICAL DATA AND ERROR ANALYSIS

Michael Axelrod, *Lawrence Livermore National Laboratory, Livermore, California, U.S.A.*

E. Richard Cohen, *Rockwell International (Ret.), Thousand Oaks, California, U.S.A.*

Introduction: The Nature of Data Analysis 470
1. **Measurements in the Physical Sciences** 470
1.1 Measurement as a Control Process 470
1.2 Measurement Error 471
2. **Basic Statistical Concepts** 471
2.1 Population and Sample 472
2.2 Random Variables 472
2.3 Expectations 473
2.4 Moments of Distributions 473
2.5 The Ubiquitous Normal Distribution 475
2.6 Estimation and Efficiency 475
3. **Exploratory Data Analysis** 476
3.1 Resistant and Robust Estimators 476
3.1.1 Resistant Estimates of Location 477
3.1.2 Resistant Estimates of Scale 477
3.1.3 The Resistant Line 478
3.2 Looking at Residuals in an Exploratory Analysis 478
3.3 Transformation of Data 478
3.4 Transformations to Induce Normality 479
4. **Estimation and Error Analysis** 479
4.1 Parametric Estimation 479
4.1.1 Single Estimate with Known Variance 480
4.1.2 Single Estimate with Unknown Variance 480
4.1.3 Multiple Estimates—Least-Squares Fitting 480
4.1.3.1 The Linear Case 481
4.1.3.2 Example: Fitting Data by a Straight Line 482
4.1.3.3 Nonlinear Data, Linear Approximation 483
4.1.3.4 Matrix Formulation 483
4.1.4 Student's Distribution 483
4.2 Error Propagation 483
4.3 Bayesian Statistics, Inverse Probability 484
4.4 Measurement Uncertainty 484
5. **Statistical Graphics** 485
5.1 Description and Communication 486
5.1.1 Pie Charts, Dot Plots, and Bar Graphs 486
5.1.2 Histograms 486
5.1.3 Probability-Density Plots 487
5.2 Probability Plotting Methods 487
5.2.1 The Notion of a Quantile 487
5.2.2 Empirical Quantile–Quantile Plots 488
5.2.3 Theoretical Quantile–Quantile Plots 488
5.2.4 Interpreting QQ Plots 489
5.2.5 Poissonness Plot 489
5.3 Exploration 489
5.3.1 Scatter Plots 489
5.3.2 Time-Series Plots 490
5.3.3 Box Plots 490
5.4 The Use of Color in Statistical Graphics 490
5.5 Dynamic Graphics 490
6. **The Impact of Computers on Data Analysis** 491
6.1 Computer-Intensive Methods for Parameter Estimation 491
6.2 Modern Computing Resources for Data Analysis 492
6.2.1 Software Packages for Data Analysis 493
Glossary 494
Works Cited 494
Further Reading 496

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION: THE NATURE OF DATA ANALYSIS

The term *data analysis* admits no universally accepted or understood definition. It differs from applied statistics in that the usual restrictive assumptions, such as normality and independence, are not expected to hold. Here we think of data analysis as a collection of techniques, having their origins in a variety of disciplines, that are used in a coordinated fashion to elucidate the useful structure in a (possibly) complicated collection of recorded information. Many of the techniques used in data analysis were first developed by nonstatisticians to attack specific problems in their respective fields of interest. For example, the method of least squares (introduced by Legendre in 1805) was first used in astronomy and geodesy. Similarly, researchers in high-energy physics have created a number of novel techniques such as the *projection pursuit* algorithm (Friedman and Tukey, 1974) for looking at high-dimensional data. Unfortunately, we can discuss only a limited number of topics in this short survey article, but we hope that our selection will prove interesting and useful to professionals in the physical sciences who work with experimental or observational data. We strongly recommend that the reader seek out more detailed discussions of the topics presented here. The Works Cited and Further Reading lists should provide a good start in that direction.

We first give a brief treatment of measurement errors to motivate the need for statistical techniques in the physical sciences. Then we discuss three broad, but interconnected, topics: exploratory data analysis, classical and modern mathematical statistics, and statistical graphics. Finally, we briefly review the use of computers in data analysis and the resources available for doing data analysis in a modern computing environment.

1. MEASUREMENTS IN THE PHYSICAL SCIENCES

The scientific method requires that a proposed theory of nature be falsifiable. We usually test a theory by comparing its predictions with experimental or observational measurements. Ideally, we would like an experimental design that would produce unequivocal indications when the theory is false, but we rarely get that. Errors will inevitably accompany even our best experiments, and we must accept some uncertainty when we decide whether the measurements do or do not confirm the predictions of the theory. Therefore, it is important that we quantify the measurement uncertainty. An example of the importance of calculating and reporting measurement uncertainty was an early test of the general theory of relativity. Einstein predicted that starlight just grazing the rim of the sun would deflect by 1.75 s of arc (Einstein, 1915, p. 831; Misner *et al.*, 1971). A crucial test of his prediction came on 29 May 1919, during a total eclipse of the sun, observed by English scientific expeditions to Sobral in Brazil and to Principe Island off the west coast of Equatorial Guinea. On 6 November 1919, Sir Arthur Eddington reported that the measured deflection was $1''.98 \pm 0''.30$ at Sobral and $1''.61 \pm 0''.30$ at Principe (Bertotti *et al.*, 1962; Moyer, 1979). The scientific community regarded the eclipse data as confirmation of the general theory of relativity, and Einstein became an instant worldwide celebrity (Pais, 1982, pp. 299–324). Note that the two uncertainty intervals overlap each other, and both cover Einstein's prediction of $1''.75$, but not the Newtonian value of $0''.875$, based on a classical particle moving with a speed c in a hyperbolic orbit past the Sun. Without the report of the uncertainty of the measurements, this early confirmation of the general theory of relativity would have had less credibility. In fact, had Eddington averaged the solar deflection values to get $1''.80$ and reported a combined uncertainty of $0''.21$, the prediction would have been even more spectacular. Eddington's reason for not combining the values may have been that he thought it possible that one or both of them were biased.

1.1 Measurement as a Control Process

Making measurements is similar to a production process: A sequence of prescribed operations produces an end product. Sometimes the end product is merely a number from a sensor; more often it is the result of a data reduction procedure, which could be as simple as an arithmetic mean or so complex it takes many hours to execute on a fast computing machine. Like the production process, the measurement process needs to be in a state of *statistical control*, with statistical properties that are invariant in time. The use of *control charts* (Shewhart, 1931; Grant and Leavenworth, 1980) to monitor statistical control is recommended. When a state of control is achieved, we can assume that the universe of observations is stable, and it makes sense to describe the observations in terms of a probability distribution with fixed parameters that can be claimed to represent physical reality.

Statistical control is just as important in the scientific laboratory as in the factory, and the dicta laid down by Deming (1950) for statistical process monitoring apply. In particular, we should not attempt to use statistical methods to compensate for a measurement process that lacks control. Instead, the root cause of the problem (e.g., instrument failure, procedures, environment, personal, etc.) should be identified and corrected.

1.2 Measurement Error

We use the 1993 definitions of measurement error provided by International Standards Organization (ISO, 1993a, 1993b), presented in Table 1.

Table 1. ISO definitions relative to modeling measurement systems.

Term	Definition
Error of a measurement	Result of a measurement minus the true value of the measurand.[a]
Random error of a measurement	Result of a measurement minus the mean that would result from an infinite number of measurements of the same measurand carried out under conditions of repeatability.[b]
Systematic error of a measurement	Mean that would result from an infinite number of measurements minus a true value of the measurand.[c]

[a] Since a true value cannot be determined in practice, a conventional value is used.

[b] (1) Random error is equal to error minus systematic error. (2) Random error can only be estimated for a finite set of measurements.

[c] (1) Systematic error is error minus random error. (2) Systematic error cannot be completely quantified. Like the true value, it cannot be completely known.

The simplest model for a measurement process in a state of statistical control without systematic error is given by the equation

$$x_i = \mu + \delta + \epsilon_i; \quad i = 1, 2, \ldots, n. \tag{1}$$

The parameter μ represents the true value of the quantity being measured; δ is a constant (systematic) error; and the ϵ_i represent the random errors (in the sense defined in Table 1) in the measurement system. The ϵ_i are zero-mean, independent, identically distributed random variables with a common finite variance σ^2. [Although Eisenhart (1963) has criticized the concept of error as the difference between the measured value and the true value as operationally meaningless, it is often useful to consider a true value as being a result that would be provided by an idealization of the measurement process.] In many cases μ itself is a random variable because the value of the quantity under measurement is not constant. The optimum estimate of μ (in the sense of maximum likelihood) depends on the probability distribution of the errors ϵ_i. If the distribution is known, then the analyst can craft the form of the estimator to have optimal statistical properties. However, in practice the analyst rarely knows the distribution of errors; frequently he simply assumes it to be normal—thus the ubiquitous use of the mean in reducing experimental data. However, even in a state of statistical control, the distribution of errors need not be normal and could be asymmetrical with a long tail; in this case the arithmetic mean will perform very poorly. In such cases *robust* statistical procedures are helpful. Robustness is a somewhat vague property of statistical procedures. Informally, a robust procedure performs well under both the assumed distribution and departures from the assumptions. This subject is treated in more detail in Sec. 3.

2. BASIC STATISTICAL CONCEPTS

What is the difference between the subject of probability and the subject of statistics? While statistical methods use the theory of probability, the two subjects are different. It is possible to take courses in probability and yet learn little about the subject of statistics. In a nutshell, probability is the forward problem, while statistics is the inverse problem. In probability problems, we postulate the parameters of various distributions, and we make predictions of outcomes (data). The conceptual flow is from model to data. In doing statistical problems, we are given outcomes (data), and we make inferences about a model that describes the process that created the data. The conceptual flow is from data to model. The subject of probability tends to use deductive reasoning, while statistics uses inductive reasoning. In this sense statistics is really similar to the basic scientific endeavor: make observations and inductively proceed to learn about the world.

2.1 Population and Sample

One of the most basic distinctions in statistics is that between population and sample. A *population* is a finite or infinite set of objects, where each object has one or more attributes of interest. Usually, we study a population by making observations on a finite *sample* from the population, and then make inferences to the population as a whole. In some situations, the population is an abstraction; we pretend that the data we have is but one realization of what might have been. The philosophical problems with this conceptualization of statistical inference have caused some (DeFinetti, 1972; Ramsey, 1980; Lindley, 1985) to adopt a completely subjective interpretation of statistics—the Bayesian approach, as opposed to the frequentist approach. The frequentists consider the parameters in statistical models as fixed (but unknown) quantities that are to be estimated from the observed data. It is the parameter *estimator* that is random and not the parameter. In contrast, Bayesians consider the estimators as the given data and the parameters as random. These must be assigned a prior distribution of possible values based on the information about the parameters that is available before observations. The observations (data) then modify the prior distribution to produce a new distribution called the posterior distribution. While the vast majority of professional statisticians would regard themselves as frequentists, it must be said that the Bayesian approach is both well thought out and logical. A number of statisticians have opted for an eclectic approach, using one or the other approach depending on the nature of the problem at issue. Our discussions here are frequentist; Efron (1978) provides an excellent review on the controversy between the subjective and objective viewpoints.

2.2 Random Variables

Mathematically we define a random variable as a mapping that assigns a number to a random outcome. The term "random variable" is somewhat confusing; a more appropriate term would be "random function," where the domain would be the set of experimental outcomes and the range would be the set of real numbers. Random variables provide the basic machinery for the mathematical theory of statistics. The idea of a variable that can appear with varying degrees of probability does not occur in ordinary analysis or in mathematical logic. Nevertheless, random variables are an extremely useful concept in statistics because we can operate on them as if they were ordinary numbers, subject to the laws of arithmetic. Random variables can appear in sums, products, ratios, or powers, or as the elements in infinite series, vectors, matrices, or tensors. Random variables can also appear as the arguments of functions, or as the coefficients of differential equations. The common practice in mathematical statistics is to use upper-case roman letters to denote random variables and lower-case roman letters to denote the numerical values that the random variables can assume.

A random variable is completely defined by its *distribution function* $F(x)$ (also called the cumulative distribution function), which is the probability that the random variable X has a value less than some given number, say x. Thus we can make statements such as

$$P[X < x] = F(x). \qquad (2)$$

In this equation, the term on the left is a statement of probability, the symbol P being shorthand for using the word "probability." The symbol X denotes some random variable, while x denotes a real number. In contrast, the term on the right is conventional, in the sense that we can substitute some function for the variable F and some number for the variable x and get a meaningful statement. Thus if we substitute "3" for x and "multiply by 0.02 and square" for F, we get a meaningful statement that has a truth value:

$$P[X < 3] = (0.2 \times 3)^2. \tag{3}$$

On the other hand, we cannot substitute any of the common mathematical objects for X and get a meaningful statement.

If the set of possible numerical values for a random variable is finite or countably infinite, then the random variable is said to be *discrete,* and we can write

$$P[X = x_k] = p_k, \quad k = 1, 2, \ldots. \tag{4}$$

If there is a function $f(\cdot)$ such that we can write the distribution as

$$F(x) = \int_{-\infty}^{x} f(t)dt, \tag{5}$$

then the random variable is said to be continuous, and $f(x)$ is called the *probability density* function of the random variable whose distribution function is F.

2.3 Expectations

An important topic in the mathematical theory of statistics is the *expectation* $E[X]$, the expected or mean value of a random variable X; it is the first moment of its probability density function:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx \text{ (continuous variable)}, \tag{6}$$

$$E[X] = \sum_{k=1}^{\infty} p_k x_k \text{ (discrete variable)}. \tag{7}$$

For a discrete distribution the expected value need not be a realization (example: the mean number of children per family is 2.31). Indeed, for some distributions, the expected value does not exist because the integral does not converge.

Is the expectation value in some sense expected? Yes. According to the (weak) *law of large numbers,* the average becomes less dispersed about the expectation value of the parent distribution as n increases:

$$\lim_{n\to\infty} \{P(|\bar{x}_n - E[X]|)\} > \epsilon] = 0, \quad \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{8}$$

where ϵ is any positive number.

The *variance* is the expectation of the square of the random variable about its mean:

$$\mathrm{Var}[X] = E[(x - E[X])^2]. \tag{9}$$

The variance provides a measure of how dispersed the realizations of the random variable are about the mean. The mean and the variance give a summary description of the random variable, and when the distribution function is Gaussian, these two quantities completely characterize it.

2.4 Moments of Distributions

There are other important numerical characteristics of a random variable, namely, its higher-order moments. The rth moment μ_r' of X is defined as

$$\mu_r' = E[x^r], \tag{10}$$

and the moments centered about the mean are given by

$$\mu_r = E[(X - \mu_1')^r]. \tag{11}$$

The second centered moment is the variance.

Do the moments—when they exist—completely characterize a random variable? For most distributions of practical interest the answer is yes. Two distributions with equal moments for all orders would be identical. A notable exception is the lognormal (frequently used to model particle sizes). This distribution is not uniquely determined by its moments in the sense that the infinite sequence of moments for any given lognormal distribution could also belong to another distinct distribution.

Why are moments important in statistics? Why not stay with distributions? The answer is that we need moments because in general it is much more difficult to get an estimate of the distribution function from data than it is to get estimates of the moments. Data sampling a single random variable (univariate data) with a Gaussian distribution, for example, requires knowledge of only the first two moments (two numbers). On the other hand, multivariate data (i.e., data sampling distributions of multiple random variables) are far more complex. For Gaussian data of dimension k, we need k first-order moments, and $k(k + 1)/2$ independent second-order moments and joint moments (the elements of a symmetric matrix of dimension k), for a total of $k(k + 3)/2$ independent numbers [$k(k + 1)$ numbers in all]. Finally, for non-Gaussian multivariate data, we need knowledge of higher-order moments and joint moments—tensor quantities (McCullagh, 1987).

The moments of a distribution (when they exist) may be found from the *characteristic function,* which is essentially the Fourier transform of its density function:

$$\phi(t) = E(e^{itX}) = \int_{-\infty}^{\infty} f(x)e^{itX}dx = 1 + \sum_{k=1}^{\infty} \mu_k' \frac{(it)^k}{k!}, \tag{12}$$

and

$$\mu_r' = (-i)^r\left[\frac{d^r\phi(t)}{dt^r}\right]_{t=0}. \tag{13}$$

In some cases the moments may exist and not the characteristic function. For example, the lognormal distribution has moments given by

$$\mu_r' = \exp(r\zeta + r^2\sigma^2/2), \tag{14}$$

where ζ and σ are parameters of the distribution, but the characteristic function does not exist (the moment expansion series does not converge). The problem lies in the fact that other distributions have the same moment sequence. Kendall *et al.* (1983) present a detailed discussion of when the moments uniquely determine a distribution.

The logarithm of the characteristic function defines the *cumulants* or *semi-invariants* κ_r of the distribution:

$$\ln \phi(t) = \sum_{r=1}^{\infty} \kappa^r \frac{(it)^r}{r!}, \tag{15}$$

$$\kappa_r = (-i)^r\left[\frac{d^r \ln \phi(t)}{dt^r}\right]_{t=0}. \tag{16}$$

The first two cumulants, κ_1 (the mean of the distribution) and κ_2 (the variance), provide measures of the location and spread of the distribution, while those of higher order give more subtle information. For example, the third cumulant, κ_3, the third central moment, measures asymmetry in the sense that for all symmetrical distributions, κ_3 vanishes. However, to guarantee symmetry, all odd-order cumulants must vanish, and the distribution must be determined by its moments.

There is a one-to-one correspondence between the cumulants κ_r and the moments μ_r or μ_r'. The conversion formulas for the first four orders in each direction are

$$\kappa_1 = \mu_1', \tag{17a}$$

$$\kappa_2 = \mu_2 = \mu_2' - \mu_1'^2, \tag{17b}$$

$$\kappa_3 = \mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3, \tag{17c}$$

$$\kappa_4 = \mu_4 - 3\mu_2^2 = \mu_4' - 4\mu_3'\mu_1' - 3\mu_2'^2 + 12\mu_2'\mu_1'^2 - 6\mu_1'^4, \tag{17d}$$

and

$$\mu_1' = \kappa_1, \quad \mu_1 = 0; \tag{18a}$$

$$\mu_2' = \kappa_2 + \kappa_1^2, \quad \mu_2 = \kappa_2; \tag{18b}$$

$$\mu_3' = \kappa_3 + 3\kappa_1\kappa_2 + \kappa_1^3, \quad \mu_3 = \kappa_3; \tag{18c}$$

$$\mu_4' = \kappa_4 + 4\kappa_1\kappa_3 + 3\kappa_2^2 + 6\kappa_2\kappa_1^2 + \kappa_1^4, \quad \mu_4 = \kappa_4 + 3\kappa_2^2. \tag{18d}$$

Kendall *et al.* (1983) provide conversions between moments and cumulants up to order ten.

Two important properties of cumulants are scaling,

$$\kappa_r(aX + b) = a^r\kappa_r(X), \quad r > 1, \tag{19a}$$

$$\kappa_1(aX + b) = a\kappa_1(X) + b, \tag{19b}$$

and additivity,

$$\kappa_r\left(\sum X_i\right) = \sum \kappa_r(X_i). \tag{19c}$$

Central moments of all orders are scalable in this sense, but only μ_2 and μ_1 are additive.

2.5 The Ubiquitous Normal Distribution

The normal distribution (also called the Gaussian distribution) plays a key role in the theory of statistics and data analysis. It seems to dominate the subject both in theory and in practice. The normal or Gaussian probability density of mean μ and variance σ^2 is

$$f(x) = \frac{1}{(2\pi)^{1/2}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{20}$$

A popular shorthand way to denote that a random variable X is normally distributed with mean μ and variance σ^2 is $X \sim N(\mu,\sigma^2)$.

What are the salient mathematical properties of the normal distribution that make it so useful? First, we note that the normal distribution is uniquely determined by its mean and variance, since all its higher-order cumulants vanish. Hence any other continuous distribution with appropriately small higher-order cumulants can be approximated by a normal distribution with the same mean and variance. We can gain more insight from the following observation: The cumulants of the average of n independent identically distributed random variables are given simply by $n\kappa_r/n^r$. Therefore, as n increases, the higher-order cumulants vanish much faster than the first two, and the distribution of the average approaches that of the normal distribution. This explains the famous central-limit theorem (Gnedenko and Kolmogorov, 1954).

Another important mathematical property of the normal distribution is its entropy. The normal distribution has the maximum entropy for all continuous distributions that vanish nowhere and have finite second moments (Cover and Thomas, 1991). Since entropy is a measure of uncertainty, the normal distribution is a maximum-uncertainty distribution. This makes the normal distribution the appropriate model for many random phenomena. See the seminal article on this subject by Jaynes (1957).

2.6 Estimation and Efficiency

A common problem in statistical analysis is the estimation of the parameters of the distribution of a parent population from a sample of observations. In general the estimate is expressed as some function of the observations. The numerical value that the function takes is called a *statistic.* Since the observations are random variables, the statistic is random. If we use the statistic to estimate the unknown parameter(s) say, θ, then in a sequence of repeated realizations of the observations, the statistic will fluctuate about the true value of the parameter. A statistical estimator is an algorithm that operates on the data under consideration to yield one (or more) numerical value that corresponds to the distribution parameter(s). Good estimators will give values that are close (in some well-defined sense) to the true value of the parameter. The estimator, being itself a random variable, has a distribution, and the quality of the estimator will depend on this distribution. The distributions of many estimators approach normality as the sample size increases (asymptotic normality) because of the central-limit theorem. Hence the asymptotic properties of the estimator are described by the expectation and variance of the distribution of the estimator. If the mean equals the unknown parameter, the estimator is unbiased; if the mean tends to the parameter as the sample size increases, the estimator is asymptotically unbiased. The variance of the estimator provides a measure of the quality known as its efficiency. The asymptotically normal estimator with the smallest variance for a given sample size n as $n \to \infty$ is known as the *efficient* estimator. The asymptotic relative efficiency (ARE) of a given parameter estimator is the ratio of the variance of the efficient estimator for that parameter to the variance of the given estimator. The ARE is then a number less than 1 (100%) and measures the effectiveness of an estimator relative to the best estimator for the same parameter. The ARE is an important concept for exploratory data analysis.

A powerful parametric estimation technique whose extensive use in statistical analysis was pioneered by Fischer uses the maximum-likelihood principle. Assume n statistically independent observations sampling the same distribution, $\{x_1,x_2,\ldots,x_n\}$. Furthermore, assume for simplicity that the distribution is a function of a single parameter θ. The joint probability of the n observations as a function of θ is called the likelihood function and can be written (in a Bayesian sense) as

$$L(x|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta), \tag{21}$$

where $f(x|\theta)$ stands for a generic, univariate or multivariate, continuous or discrete distribution. The maximum-likelihood principle chooses θ, the maximum-likelihood estimator of θ, as the value that maximizes $L(x|\theta)$. For a large class of distributions satisfying certain restrictions, the maximum-likelihood estimator is asymptotically normal and efficient.

3. EXPLORATORY DATA ANALYSIS

The analysis of data usually consists of two phases, exploratory and confirmatory. In the exploratory phase the analyst seeks to discover important patterns and features in the data without the knowledge of a specific probability model. On the other hand, the confirmatory phase is both more formal and more technical; it uses the mathematical methods of statistical inference to draw conclusions from observations and to assign a measure of uncertainty to the conclusions. Sometimes the results of the exploratory phase are so convincing that the analyst may skip the confirmatory phase. We emphasize that there is no set recipe for doing exploratory data analysis. Scientists and statisticians have always done some form of exploratory analysis in the course of their work. Four basic themes permeate the practice of exploratory data analysis: resistance, residuals, transformation, and display (Hoaglin, 1983). Analysts usually use all four themes in combination in a given analysis. We present various tools that we believe are especially helpful.

3.1 Resistant and Robust Estimators

The working style of classical parametric statistical estimation is to derive optimal estimates of parameters in distributional models. For example, the sample variance s^2 of the population variance σ^2, defined by

$$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2, \tag{22}$$

is an unbiased and (asymptotically) optimally efficient estimator. Unfortunately, in classical statistical estimation, we are dealing with functionals on the space of probability distributions that are nowhere continuous in any statistically reasonable topology (Hampel, 1973). Because of this lack of continuity, the classical estimators often become extremely suboptimal for even small changes in the assumed distribution. Hampel points out that mathematical theorems on optimality can be dangerous and misleading when applied to real-world data. For example, if we sample from a contaminated normal distribution, the estimate of the population standard deviation becomes extremely suboptimal (Tukey, 1960). Just how small a contamination will induce a significant drop in efficiency is surprising. An example of a contaminated normal is given by

$$F(x) = (1 - \epsilon)\Phi(x) + \epsilon\Phi(x/3), \tag{23}$$

where Φ is the standard (mean 0, standard deviation 1) normal distribution. The contaminated normal given by Eq. (23) has slightly thicker tails than a normal distribution, depending on the value of ϵ. For $\epsilon = 0.0018$ (which means that we sample from the outlier distribution only once in 550 trials), the mean deviation, given by

$$\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|, \tag{24}$$

becomes more efficient than the standard deviation. For many situations this amount of contamination would be undetectable from the data alone.

Some theorems from mathematical statistics appear to give results that require no assumption of a distribution. For example, the Gauss–Markov theorem states that for any distribution, the arithmetic mean provides a minimum-variance estimate of the mean of the distribution. But this requires the existence of a mean and a finite variance, and cases arise where this requirement is not satisfied. An extreme example, but nevertheless one that often approximates reality, is the Cauchy distribution with density function

$$f(x) = \frac{1}{\pi}\frac{b}{b^2 + (x - a)^2}\,. \tag{25}$$

The sample mean does not estimate the center of symmetry a because the first moment about the point a is undefined. In this case we would need a robust estimator, such as the median, to estimate the parameter a.

The subject of robust estimation attempts to deal in a systematic fashion with these problems. Tukey (Andrews *et al.*, 1972) refers to robust statistics (in analogy with computers) as "third-generation statistics," after parametric statistics and nonparametric statistics. In general, robustness refers to the property of resistivity to changes from assumed "standard conditions." A *resistant* estimator is one that is insensitive to aberrations or outliers in the data. Resistance can be considered a numerical property, while robustness is more of a statistical property. The most important robustness property is the robustness of *efficiency*. An estimator has robustness of efficiency if its variance is minimal or nearly so across a range of distributions. Estimators that are both resistant and robust are essential for exploratory data analysis.

An *outlier* is an observation that deviates strongly from the bulk of the data—it appears not to belong to the rest of the data. Frequent causes of outliers in physical science data are equipment failures, software bugs, transcription errors, and faulty experimental procedure. However, sometimes such observations arise from the very nature of the population under study. An outlier is definable only with respect to a model distribution, and if the observation represents an improbable sample from the assumed distribution, the analyst must decide if the observation is extraneous, and does not belong to the population, or if the model must be changed. Thus, the distinction between outliers and data can be subtle; a sample from some non-normal distributions will naturally have observations that look like outliers. Only if we know the underlying distribution can we craft the appropriate estimator, with optimal or nearly optimal properties such as efficiency, using the mathematical theory of statistics. But this is not within the spirit of exploratory data analysis, where we usually do not know the distribution. Mathematically this means that we need estimators that are nearly optimal (or at least not severely suboptimal) for the normal distribution as well as for a broad class of non-normal distributions. This problem was the subject of much empirical and theoretical research in statistics during the 1970s. Before we discuss those results, we review a few basic concepts.

3.1.1 Resistant Estimates of Location

First we review estimates of *location*, or the central tendency of the data. The mean is an estimate of location. Given a set of observations that we label $x_1, x_2, \ldots, x_N$, the sample mean $\bar{x}$, calculated by

$$\bar{x} = \frac{1}{N}\sum_{i-1}^{N} x_i, \tag{26}$$

is an estimate of the first moment of the probability density $f(x)$:

$$\int_{-\infty}^{\infty} xf(x)dx. \tag{27}$$

When the density function is the normal or Gaussian density [Eq. (20)], the sample mean is an optimal estimator for the first moment μ. However, even one aberrant observation can sometimes cause the mean to produce misleading results. Thus for exploratory data analysis, we need to replace the mean by a resistant estimator such as the *median*. The median is calculated by finding the value such that half of the observations are larger and half are smaller. While the median is resistant, it is not the most statistically efficient estimator across a broad range of distributions. For example, when the data are normal, the median has only $2/\pi$ the efficiency of the mean. However, if the median is substituted for the mean only when the difference between the two is large, then an efficiency of 90%–95% can be achieved, depending on the sample size (Andrews *et al.*, 1972). A number of very sophisticated computer-intensive estimators of location are also available (Mosteller and Tukey, 1977; Hoaglin, 1983).

3.1.2 Resistant Estimates of Scale

The sample standard deviation

$$\text{s.d.} = \left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{1/2} \tag{28}$$

provides an estimate of the spread or *scale* of the data, but like the mean it lacks both resistance and robustness, and we need a substitute for exploratory data analysis. Scale estimators should be location invariant and scale equivalent: If w is measure of scale, then

$$w(a + bx_1, a + bx_2, \ldots, a + bx_N) = |b| w(x_1, x_2, \ldots, x_N). \quad (29)$$

The standard deviation has this property, as well the median absolute deviation (MAD) defined by

$$\mathrm{MAD} = \mathrm{median}\{|x_1 - M|, |x_2 - M|, \times \ldots, |x_N - M|\}, \quad (30a)$$

$$M = \mathrm{median}\{x_1, x_2, \ldots, x_N\}. \quad (30b)$$

An estimate of the standard deviation of an assumed normal population is then given by

$$s = (2/\pi)\mathrm{MAD}. \quad (31)$$

Simulation studies (Lax, 1975; Tukey *et al.*, 1977) indicate that MAD has good robustness and resistance for a broad range of heavy-tailed non-Gaussian distributions, but it is not highly efficient for Gaussian data.

3.1.3 The Resistant Line

Given bivariate data (x_i, y_i), we often want to fit a straight line of the form

$$y_i = \alpha + \beta x_i, \quad i = 1, 2, \ldots, N. \quad (32)$$

The usual technique is to use the method of least squares, which has no resistance to aberrant observations—a single point can dominate the fit. Fortunately, there is a technique for bivariate data based on medians that does provide a resistant fit.

The technique works by dividing the data into three groups, a left, middle, and right group, each with about the same number of points. For each group the medians of the x and y values are calculated. This produces three summary points, (x_L, y_L), (x_M, y_M), and (x_R, y_R). The slope is then calculated by

$$\hat{\beta} = (y_R - y_L)/(x_R - x_L), \quad (33)$$

and the intercept is given by

$$\hat{\alpha} = \tfrac{1}{3}[(y_L - \hat{\beta}x_L) + (y_M - \hat{\beta}x_M) + (y_R - \hat{\beta}x_R)]. \quad (34)$$

However, it is more convenient to work with the line recentered about the median of all the x values:

$$\hat{y} = \alpha^* + \hat{\beta}(x - x_M). \quad (35)$$

The quantity α^* is called the *level*. The residuals from the conventional least-squares fit,

$$\hat{\epsilon}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i), \quad (36)$$

have the following property: When another line is fitted to the residuals by least squares, the slope and level are zero. However, this will not be the case with the resistant line fit. To preserve this desirable property, we must adjust the level and slope by an iterative procedure until the residuals become nearly zero. Each step in the iteration uses the residuals from the prior step to get a new resistant fit with slope and level updated as follows:

$$\hat{\beta}_1 = \hat{\beta}_0 + \delta_1, \quad \hat{\beta}_2 = \hat{\beta}_1 + \delta_2, \ldots. \quad (37)$$

Here δ_1 is the slope from the resistant fit to the residuals from the initial fit, and so on. The levels are updated in a similar fashion. Convergence problems can occur for some data sets, but safe iteration schemes are available (Johnstone and Velleman, 1981), as well as published software for doing resistant fits (Velleman and Hoaglin, 1981).

3.2 Looking at Residuals in an Exploratory Analysis

The residuals are what remain after a summary or fitted model is subtracted from the data. A careful examination of the residuals is an essential step in an exploratory analysis. When resistant techniques (such as the resistant line) are used to do fitting, aberrant points will produce dramatically extreme residuals. Graphical display of the residuals is also essential to reveal any systemic patterns in the residuals, which would indicate that the chosen fit model is not appropriate.

3.3 Transformation of Data

Transformation of the data (also called *re-expression*) is the process of finding the appropriate scale (e.g., logarithmic or square root) that simplifies the analysis. Transformations are used to make data with a skewed distribution more symmetrical, to stabilize the spread across different groups, or to make the variance more nearly constant. For example, data that consist of counts (e.g., neutron counting) often fit the Poisson distribution,

$$p(k) = e^{-\lambda}\lambda^k/k!, \tag{38}$$

where k is the number of counts detected in a fixed interval, and λ is the corresponding mean. The variance of the Poisson distribution equals its mean, and therefore Poisson-distributed counts exhibit increasing variance with increasing intensity. This often complicates the analysis or processing of such data. However, the square root of the Poisson counts has a variance asymptotically equal to 1/4.

3.4 Transformations to Induce Normality

Since the normal distribution has very convenient properties and has been extensively studied, it is frequently desirable to transform random variables to a form that is normally or approximately normally distributed. The transformation of the correlation coefficient provides a good example. If $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_n$ are samples from a bivariate normal distribution with population correlation coefficient ρ, then the sample correlation coefficient r is bounded, $-1 \le r \le 1$, and thus poorly represented by a normal distribution. Therefore the following transformation (Fisher, 1915) is appropriate:

$$Z = \tfrac{1}{2}\ln\frac{1+r}{1-r} = \operatorname{artanh}(r). \tag{39}$$

The transformed variable Z has the range $-\infty \le Z \le \infty$; it is approximately normally distributed and has mean and variance given by (Gayen 1951)

$$E(Z) = \operatorname{artanh}(\rho) + \frac{\rho}{2(n-1)} \times \left(1 + \frac{5+\rho^2}{4(n-1)}\right) + o(n^{-2}), \tag{40a}$$

$$\operatorname{var}(Z) = \frac{1}{n-1} \times \left\{1 + \frac{4-\rho^2}{2(n-1)} + \frac{22-6\rho^2-3\rho^4}{16(n-1)^2}\right\} + o(n^{-3}). \tag{40b}$$

Using the Fisher Z transform, the analyst can test the sample correlation against any assumed value for the true correlation ρ. For example, how large a sample correlation r do we need to have a reasonable (2σ) confidence that the true correlation ρ is nonzero? The Fisher Z transform readily answers this question: For $n = 10$, $|r| > 0.6$; for $n = 30$, $|r| > 0.4$; and for $n = 100$, $|r| > 0.2$ (assuming, again, a normal population). The Fisher Z transform should be used with care, however; r is neither robust nor resistant, and outliers could render this test useless.

4. ESTIMATION AND ERROR ANALYSIS

4.1 Parametric Estimation

Generally when we make observations (measurements) in the physical sciences, we think of the observations as a sample from some conceptual (if not actual) population of possible observations. Our interest focuses on the *variation* within the population. Sometimes we regard the variation as a nuisance effect (noise), while at other times the variation is itself the object of study. These ideas lead to the subject of *parameter estimation*. In the frequentist approach to statistics, we describe the population by some probability distribution with fixed (nonrandom) unknown parameters. Here the sample space is on the data. In contrast, in the Bayesian approach, the sample space is defined on the parameters; the observations are taken as fixed quantities, and a probability distribution is assigned to the parameters to be estimated. In this section, we adopt the frequentist approach, where the parameter *estimate* is a random quantity calculated from the random data, as opposed to the parameter *value* that we wish to estimate.

Given a set of data (the sample), we need to deal with two basic questions:

1. What function of the data do we use to estimate the unknown parameter?
2. How accurate is the estimate?

First, we discuss the classical approach using linear unbiased minimum-variance estimates (LUMVE) [or best linear unbiased estimates (BLUE)]. The classical approach requires some knowledge of the parametric form of the distribution that describes the data. Later, we briefly discuss computer-intensive approaches to parameter estimation that require much less knowledge of the distribution that describes the data—usually only statistical independence of the samples.

4.1.1 Single Estimate with Known Variance If the set $\{y_1,y_2,\ldots y_N\}$ represents values assumed to be drawn from a population with variance σ^2 and a fixed, but unknown, mean μ, then

$$\bar{y} = m'_1 = \frac{1}{N}\sum_{i=1}^{N} y_n \tag{41}$$

is the best estimate of μ in the sense that of all the linear combinations of the observations $\{y_1,y_2,\ldots,y_N\}$ with expectation μ, it has the smallest variance. The variance of $\bar{y}$ is σ^2/N.

If the observations $y_1, y_2, \ldots, y_N$ are drawn from a set of populations with different variances but a common mean μ, then the weighted average

$$\bar{y} = \frac{1}{W}\sum_{i=1}^{N} w_n y_n \text{ with } W = \sum_{i=1}^{N} w_n \tag{42}$$

is an unbiased estimate of μ. Furthermore, if we set $w_n = 1/\sigma^2$, then the weighted average becomes a minimum-variance unbiased estimate with variance σ^2. This statement of the problem is somewhat unrealistic because usually we do not know σ^2, although we might have a good estimate from prior experiments or from the underlying physical theory. The reciprocal of the variance is often called the precision (Frodesen *et al.*, 1979) of the measurement. A commonly made error is to combine measurements or estimates with unequal precisions without using the appropriate weighting factors.

4.1.2 Single Estimate with Unknown Variance If neither the mean nor the variance of the population is known, both must be estimated from the data. An unbiased estimate of σ^2 is provided by

$$s^2 = \frac{1}{N-1}\sum_{n=1}^{N}(y_n - \bar{y})^2 = \frac{N}{N-1}(m'_2 - m'^2_1), \tag{43}$$

where

$$m_k = \frac{1}{N}\sum_{n=1}^{N} y_n^k. \tag{44}$$

Sample moments are denoted by m_k and m'_k to distinguish them from the population moments μ_k and μ'_k. Strictly speaking, s^2 is the sample estimate of the variance of the distribution of y, but it is common to call it simply the sample variance. The variance of the sample, $m_2 = m'_2 - m'^2_1$, gives the maximum-likelihood estimate of the variance of the distribution of y, but it is a biased estimate. The variance of $\bar{y}$ as an estimate of μ (the variance of $\bar{y} - \mu$) is σ^2/n; this itself is estimated by s^2/n. The variance of s^2 is $2\sigma^4/(N - 1)$ for y_n drawn from a normal distribution.

4.1.3 Multiple Estimates—Least-Squares Fitting An important problem in data analysis is the fitting of models to data. By "model," we mean some linear or nonlinear relation between one or more independent variables (also called explanatory variables, regressors, or carriers) x_i and a dependent variable (or response) y_i, usually expressed in the form of an explicit equation. Least squares has been the predominant method of fitting models to data for almost 200 years despite the many criticisms of its shortcomings. Virtually every general textbook on statistics treats the subject, sometimes to the exclusion of all others, and any treatment of data analysis would be deficient without a thorough discussion of this important topic. The method owes its popularity to its simplicity and mathematical tractability. It is basically a numerical technique and can be treated without recourse to its statistical connections, as is typically done in textbooks on numerical analysis, where least squares is generally discussed in the context of solving overdetermined systems of linear equations.

The method of least squares had its origins in scientific data analysis, specifically in various empirical investigations in astronomy and geodesy where it was necessary to combine measurements in order to estimate certain quantities. While most scientists and mathematicians would regard Gauss (1809) as the progenitor of the method of least squares, priority in publication belongs to Legendre (1805). Evidently conditions were ripe for its discovery because Adrain (1808) independently discovered and published the method in America at about the same time. Many statisticians still feel that Legendre's exposition is unsurpassed in its lucidity (Stigler, 1986). He proposed the idea of minimizing the sum of squared errors and derived what we now call the normal equations. Legendre offered no proofs for the superiority of this technique, whereas Gauss deduced his first proof of the normal law of error from the postulate that the arithmetic mean is the most probable value associated with a set of equally good measurements.

In many situations the set of data $\{y_1,y_2,\ldots y_N\}$ does not consist of observations of a single quantity y, but values of different quantities that may often be reduced to a set of observations y_n ($n = 1, 2, \ldots N$) of expressions $g_n(x_1,x_2,\ldots,x_M;p_j)$ containing M unknown quantities x_i as well as known parameters p_j. The functions g_n might be the same function evaluated for different values of the known parameters, or the set $y_n = g_n(x_i)$ may be physically different quantities. For the application of least squares it is only necessary that at least one subset of M expressions $y_n = g_n(x_i)$ can be solved for a unique set of values of the M unknowns x_i. Then, if $N > M$, the least-squares algorithm will provide the best estimate of the set of M unknown quantities, in the same sense that an appropriately weighted mean provides the best estimate of the value of a single variable.

4.1.3.1 The Linear Case. The simplest situation involves expressions in which the quantities to be determined appear linearly. A typical example is the determination of the coefficients a_i in the expansion of an expression in terms of a set of base functions:

$$y(z) \approx Y(z) = \sum a_i f_i(z). \tag{45}$$

The given data are the pairs (z_n,y_n), $n = 1, 2, \ldots, N$, and the weights w_n. If the observations are independent (two different observations y_n and y_m are uncorrelated), then we have $w_n = 1/\sigma_N^2$, the reciprocal of the variance of the distribution from which the observation y_n is assumed to be drawn, under the assumption that the value of z_n is exact. The weights may also be relative values whose normalization s_0^2 (the variance to be assigned to unit weight) will be determined from the "goodness of fit" of the data. The estimate of the variance σ_n^2 is then given by $s_n^2 = s_0^2/w_n$.

The fitted coefficients a_i are found from

$$a_i = \sum_{j=1}^{M} V_{ij}Z_j, \tag{46}$$

where V_{ij} are elements of the covariance matrix V, the inverse of the weight matrix W:

$$(V^{-1})_{ij} = W_{ij} = \sum_{n=1}^{N} w_n f_i(z_n) f_j(z_n), \tag{47}$$

and the z_j are given by

$$z_j = \sum_{n=1}^{N} w_n f_j(z_n). \tag{48}$$

The variance of a fitted value of $y(z)$ (interpolated or extrapolated) is given by

$$\sigma^2(y) = \sum_{ij} f_i(z) V_{ij} f_j(z). \tag{49}$$

It is convenient to use an orthonormal basis for the set $f_i(z)$ with the inner product defined in terms of a finite summation, giving the orthonormality condition expressed by

$$\sum_n f_i(z_n) f_j(z_n) = \delta_{ij}; \tag{50}$$

however, the standard orthogonal functions (such as Legendre polynomials) are usually defined over a continuous interval, with the inner product defined by integration over that interval, and in this case orthonormality is expressed by

$$\int f_i(z) f_j(z) dz = \delta_{ij}. \tag{51}$$

Since the finite sum only approximates the integral even if the data points are uniformly spaced, the matrix W fails to be diagonal. Appropriate selection of the data points z_n can preserve orthogonality, but this is usually not an option. Alternatively, polynomials that are orthogonal over the specific data set can be generated. If the basis set is orthogonal over the data, the weight matrix **W** is diagonal and the expansion coefficients are independently evaluated: $a_i = Z_i/W_{ii}$. In general, with an approximately orthogonal basis, the weight matrix **W** will be approximately diagonal, and the off-diagonal terms will be small ($W_{ij}^2 \ll W_{ii}W_{jj}$). This has advantages for computational precision ($V_{ii} = W_{ii}^{-1}$), as well as in the physical interpretation of the fitted coefficients, which are then approximately statistically independent.

The "goodness of fit" of the data is measured by

$$S^2 = \sum_{n=1}^{N} w_n(y_n - \hat{Y}_n)^2 = \sum_{n=1}^{N} y_n w_n y_n - \sum_{ij=1}^{M} Z_i V_{ij} Z_j, \tag{52}$$

where $\hat{Y}_n$ is the fitted value of y_n. If the weights w_n are the reciprocals of the variances of the input data y_n, the distribution of S^2, considered as a sampling from a population defined by the distributions of the y_n, is the χ^2 distribution with $\nu = N - M$ degrees of freedom.

If $(S^2 - \nu)/\sqrt{2\nu - 1} \gg 1$, the fitted residuals are inconsistent with the precision of the observations and the fit can be improved by including additional functions f_i in the basis set. Basis functions should be retained or added to the fit if their inclusion produces a significant decrease in the value of S^2/ν. (If the basis set is not orthogonal, the presence of an additional basis function will alter the contributions of the existing functions to the value of S^2).

If w_n are only relative weights, S^2 is only proportional to a χ^2-distributed variable; the weights (and the variances) can be normalized by multiplying each w_n by the factor ν/S^2, and the covariance matrix **V** by the factor S^2/ν. In this case, χ^2 can no longer be used as a basis for evaluating "goodness of fit."

4.1.3.2 Example: Fitting Data by a Straight Line. If the data set (x_n, y_n, w_n) is to be fitted by a straight line $y = a + bx$, the fitted values of a and b are

$$a^* = (1/D)(S_y S_{xx} - S_{xy} S_x), \tag{53a}$$

$$b^* = (1/D)(S_1 S_{xy} - S_x S_y), \tag{53b}$$

with $S_1 = \Sigma_n w_n$, $S_x = \Sigma_n w_n x_n$, $S_{xx} = \Sigma_n w_n x_n^2$, $S_y = \Sigma_n w_n y_n$, $S_{xy} = \Sigma_n w_n x_n y_n$, $S_{yy} = \Sigma_n w_n y_n^2$, and $D = S_1 S_{xx} - S_x^2$. The covariance matrix of the fitted coefficients is

$$\begin{pmatrix} V_{aa} & V_{ab} \\ V_{ba} & V_{bb} \end{pmatrix} = \frac{1}{D}\begin{pmatrix} S_{xx} & -S_x \\ -S_x & S_1 \end{pmatrix}. \tag{54}$$

The goodness of fit is

$$S^2 = \sum_n w_n(y_n - a^* - b^* x_n)^2 = S_{yy} - a^* S_y - b^* S_{xy} \tag{55}$$

with $\nu = N - 2$ degrees of freedom.

The variance $s^2(y)$ of a point on the fitted line $y = a^* + b^* x$ is given by

$$s^2(y) = \left(\frac{S^2}{\nu}\right)\frac{1}{S_1}\left(1 + \frac{(xS_1 - S_x)^2}{D}\right) = \left(\frac{S^2}{\nu}\right)\left(\frac{1}{S_1} + \frac{(x - \bar{x})^2}{S_{xx} - S_x^2/S_1}\right), \tag{56}$$

where

$$\bar{x} = \frac{S_x}{S_1} = \frac{\sum w_n x_n}{\sum w_n} \tag{57}$$

is the mean value of x in the data. If the weights w_n are the reciprocals of the variances of the observed values y_n, the first factor in parentheses may be omitted and the variance evaluated on the basis of "internal consistency"; including the first factor produces the variances evaluated on the basis of "external consistency." If a straight line is a valid representation of the data, and if the data are in fact independent with variances $1/w_n$, the two expressions are equally valid estimates of the variance since in this case the expectation of $S^2 = \chi^2$ is $\nu = N - 2$. If χ^2/ν lies outside acceptable limits, the data should be examined for possible inconsistencies, or refitted by some other curve, or the weights may be accepted as relative values. When the assigned weights are not based on *a priori* information on the variances of the data but have only relative significance, the factor S^2/ν (which is an evaluation of the uncertainty to be associated with unit weight) normalizes the variances of the data based on the mean square deviation of the data from the fitted straight line.

4.1.3.3 Nonlinear Data, Linear Approximation. When the functional relationships between the observations and the unknown quantities are not linear, the least-squares algorithm may be applied to a linearized set of equations. Let x_i^0 be a set of approximations to the unknown quantities x_i, and introduce a new set of quantities u_i and z_n. The simplest choice is $x_i = x_i^0 + u_i$ and $y_n = g_n(x_i^0) + z_n$. A first-order Taylor expansion of the observational equations then leads to a linearized set

$$z_n = \sum_{i=3}^{M} \left.\frac{\partial g_n}{\partial x_i}\right|_{x_j} u_i + e_n, \tag{58}$$

where e_n is the residual error. Equivalent expressions are obtained by writing either $x_i = x_i^0(1 + u_i')$ and $y_n = g_n(x_i^0)(1 + z_n)$ or $u_i'' = \ln(x_i/x_i^0)$ and $z_n = \ln(y_n/g_n(x_i^0))$.

The solution of this set of equations has the same form as the solution given above for linear least squares with the identification

$$W_{ij} = \sum_{n=1}^{N} \frac{\partial g_n}{\partial x_i}, \quad Z_j = \sum_{n=1}^{N} \frac{\partial g_n}{\partial x_i} w_n z_n,$$
$$u_i = \sum_{i=1}^{M} V_{ij} Z_j, \quad \text{where } \mathbf{V} = \mathbf{W}^{-1}. \tag{59}$$

4.1.3.4 Matrix Formulation. Let the linearized approximation of the observational equations be expressed as

$$\mathbf{z} = \mathbf{y} - \mathbf{y}^0 = \mathbf{Au} + \mathbf{e}, \tag{60}$$

where $\mathbf{y}$ is a vector of observed data, $\mathbf{x}^0 + \mathbf{u}$ is the vector of parameters to be fitted in the least-squares adjustment, and $\mathbf{A}$ is a known matrix. The elements of $\mathbf{A}$ are $\partial y_n/\partial x_i$ ($N \times M$ matrix), evaluated at the set of fixed values, from which the data z_n and u_i represent deviations. The *input covariance matrix* $\mathbf{v}$ comprises the variances and covariances of $\mathbf{e}$ for the input data ($N \times M$). The input weight matrix is the reciprocal matrix $\mathbf{w} = \mathbf{v}^{-1}$. The output estimates of the least-squares fit are

$$\mathbf{u}^* + \mathbf{VA'wz} \text{ with } \mathbf{V} = (\mathbf{A'wA})^{-1}, \tag{61}$$

where the prime indicates the transposed matrix. The weight matrix for the fitted quantities is $\mathbf{W} = \mathbf{A'wA} = \mathbf{V}^{-1}$ and the covariance matrix is $\mathbf{V} = \mathbf{W}^{-1}$. The adjusted data are $\mathbf{y}^* = \mathbf{y}^0 + \mathbf{Au}^*$; the residuals of the fitting are $\mathbf{r} = \mathbf{y} - \mathbf{y}^* = \mathbf{Dz} - \mathbf{z}^* = \mathbf{z} - \mathbf{Au}^*$.

The data $\eta = \mathbf{y} - \mathbf{y}^0$ span a vector space with metric $\mathbf{w}$; in this space length can be defined: $|\mathbf{a}| = \mathbf{a'wa}$, and the scalar product of $\mathbf{a}$ and $\mathbf{b}$ is $\mathbf{a'wb}$. The matrix $\mathbf{P} = \mathbf{A}(\mathbf{A'wA})^{-1}\mathbf{A'w}$ is a projection operator:

$$\mathbf{PP} = \mathbf{P} \text{ and clearly } \mathbf{wP} = \mathbf{P'w}. \tag{62}$$

The space of the observed data, $S(z)$, is separated into two orthogonal subspaces, $S(Pz)$ and $S(r)$;

$$z = Pz + r, \quad r = z - Pz,$$
$$z'wPr = z'wP(z - Pz) = 0. \tag{63}$$

The goodness of fit is given by

$$S^2 = r'wr = z'wr = z'wz - (u^*)'Wu^*. \tag{64}$$

4.1.4 Student's Distribution When x is drawn from a Gaussian distribution, the probability of a value of $|\bar{x}|/\sigma_{\bar{x}}$ larger than t_p may be easily calculated from the tabulated values of the normal distribution; but for small sample sizes, where we have only an estimate of the variance, we must take into account the uncertainty in that estimate. For a normal distribution, the variance of s^2 is $2\sigma^4/n$. When the data are normal and the estimate s^2 is based on a sample of size n, the ratio $(\bar{y} - \mu)/\bar{y}$ has a Student's distribution (Student, 1908) with $\nu = n - 1$ degrees of freedom. For $\nu \gg 1$, the ratio

$$\sqrt{\frac{\nu - 2}{\nu}}\,\frac{\bar{y} - \mu}{s} \tag{65}$$

has a distribution that is approximately normal with zero mean and unit variance, whereas for $\nu = 1$, the Student distribution is a Cauchy distribution.

4.2 Error Propagation

Errors in the variables of a function will produce errors in its evaluation. If the errors in a set of measurements are assumed to be random variables, the result will be a random variable. Let the error in x_i be ϵ_i; then η, the error in y, is

$$\eta = \sum_i \frac{\partial f}{\partial w_i} \epsilon_i + \tfrac{1}{2} \sum_{i,j} \frac{\partial^2 f}{\partial w_i \partial w_j} \epsilon_i \epsilon_j + \ldots . \tag{66}$$

If the errors ϵ_i have means δ_i (systematic error) and a covariance matrix (random error)

$$v_{ij} = \overline{\epsilon_i \epsilon_j} - \delta_i \delta_j, \tag{67}$$

the mean value and variance of y are

$$\bar{y} = f(x_i^0) + \sum_i \frac{\partial f}{\partial x_i} \delta_i + \tfrac{1}{2} \sum_{i,j} \frac{\partial^2 f}{\partial x_i \partial x_j} (\delta_i \delta_j + v_{ij}) + \ldots = f(x_i^0 + \delta_i) + \tfrac{1}{2} \sum_{i,j} \frac{\partial^2 f}{\partial x_i \partial x_j} v_{ij} + \ldots \tag{68}$$

and

$$\sigma^2(y) = \sum_{i,j} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} v_{ij} + \ldots . \tag{69}$$

If the errors are uncorrelated, then $v_{ij} = \sigma_i^2 \delta_{ij}$, and Eq. (69) becomes

$$\sigma^2(y) = \sum_{i,j} \left(\frac{\partial f}{\partial x_i} \right)^2 \sigma_i^2 + \ldots . \tag{70}$$

For a function of a single variable x with variance σ^2, the mean and variance of y are

$$\bar{y} = f(x) + \tfrac{1}{2} f''(x) \sigma^2 + \ldots , \tag{71a}$$

$$\sigma^2(y) = [f'(x)]^2 \sigma^2. \tag{71b}$$

Since, in the evaluation of experimental data, the true means x_i are not known but are only approximated, the second-derivative terms in $\bar{y}$ may often be omitted, since they can be much smaller than the unknown contributions from the first-order errors.

4.3 Bayesian Statistics, Inverse Probability

The joint probability density of x and y, $p(x,y)$, can be written in two equivalent forms:

$$p(x,y)dxdy = p(x|y)dxp_2(y)dy = p(y|x)dyp_1dx, \tag{72}$$

where $p(x|y)dx$ is the conditional probability of the occurrence of x when it is known that y has occurred, and $p(y|x)dy$ is the conditional probability of the occurrence of y when it is known that x has occurred. When the distribution of x depends on a parameter (or set of parameters) denoted by y, the probability density of x for a fixed set of parameters is $p(x|y)$. The unconditioned (or *a priori*) probability densities of x and y are

$$p_1(x) = \int p(x|y)p_2(y)dy, \tag{73a}$$

$$p_2(y) = \int p(y|x)p_1(x)dx, \tag{73b}$$

and hence

$$p(y|x) = \frac{p(x|y)p(y)}{\int p(x|y')p(y')dy'} \tag{74}$$

(Bayes's theorem). The conditional probability $p(y|x)$ is also known as the *a posteriori* probability.

Probability theory is concerned with the properties of the distribution of x for a given set of parameters (the mean, variance, etc.). Bayes's theorem provides the basis for determining estimates of the values of the unknown parameters from a set of observations $p(y|x)$ when the observed data are considered to be given quantities rather than sampled random variables (inverse probability).

4.4 Measurement Uncertainty

The observed variability of the results of a physical measurement is often related to the measurement process itself rather than to a variability in the value of the quantity being measured; the data do not represent a sampling from a population of different objects, but from a population of approximate observations of a unique quantity. In the evaluation of a measurement, one is presented with a set of observations y_j from which, on the basis of the available information, an inference is to be made about the "best" value to be assigned to that quantity. The stated result is incomplete unless it is accompanied by an indication of the *uncertainty* to be attached to the stated value. This uncertainty describes neither the *error* in the measurement nor the distribution of possible errors—a given measurement result has only a single, unknown, error.

The *analysis of error* of a measurement process is based on the introduction of a statistical distribution, considered as an ensemble of results that could be obtained in repetitions of the measurement process. The distribution of errors is then described in terms of the means and variances of the given process. By itself, the error distribution is unable to describe fixed *systematic errors* that are unchanged in every repetition. These must be treated separately.

Uncertainty is the incompleteness of the knowledge of the value of the measurand associated with the given result of a measurement. It comprises, in general, many components, some of which may be evaluated from the statistical distribution of the results of series of measurements; other components are evaluated from assumed probability distributions on the basis of experience and external information. The quantitative measure of this (also called "uncertainty") is a parameter that characterizes the dispersion of the values that could reasonably be attributed to the measurand, recognizing that the stated measurement result itself should be the best available estimate of that value.

The evaluation of uncertainty is based on *inverse probability*. Whereas error analysis cannot treat fixed systematic errors to which a random variable cannot be assigned, uncertainty is based on the degree of belief in the values of quantities, both "random" and "systematic," that contribute to the lack of knowledge of the quantity being measured (the measurand).

A measurement of the quantity Y yields a result y. The error in this measurement is $\epsilon = Y - y$, although the concept of measurement error as the difference between true value and measured value can be criticized as being not meaningful operationally. If the measurement is repeated many times, it is appropriate to consider the distribution of the errors ϵ_i. The error ϵ can be expressed as the sum of a random error ϵ_r with zero mean and a specified variance, and a fixed systematic error ϵ_s with zero variance and nonzero mean. In order to specify this decomposition, it is necessary to consider "faithful realizations" of a defined measurement process, such that the statistical frequency distribution remains fixed. On this basis one can describe the error of the measurement process.

For a given set of observations, the observed mean $\bar{y}$ may be taken as the best estimate of $Y + \epsilon_s$; an estimate δ of the systematic error is also required in order to obtain $\bar{y} - \delta$, the estimate of Y that is the result of the measurement. Specifying the measurement uncertainty requires an evaluation of the uncertainty of each of the components. If all that is known about the systematic error is that ϵ_s lies between a and b, the best estimate is $\delta = (a + b)/2$, and the uncertainty of this estimate is $u^2(\epsilon_s) = (b - a)^2/12$. This result is equivalent to assuming a uniform distribution for ϵ_s between the limits a and b, but it can also be defined from a Bayesian approach based on maximum-entropy considerations, without the specific introduction of a probability distribution. The Bayesian viewpoint considers the data to be fixed quantities, not samples from a distribution.

Uncertainty components, which are equivalent to variances, and more generally to the elements of a covariance matrix, combine like variances. When $y = F(\xi_1, \xi_2, \ldots)$, where $\xi_1, \xi_2, \ldots$ represents the full set of all quantities (random or systematic) contributing to the uncertainty of F, the uncertainty is given by

$$u^2(F) = \sum_j \left(\frac{\partial F}{\partial \xi_j}\right)^2 u_j^2 + 2 \sum_{k<j} \frac{\partial F}{\partial \xi_j} \frac{\partial F}{\partial \xi_k} u_{jk}. \qquad (75)$$

All uncertainties are treated equally, whether evaluated from information based on a statistical evaluation of repeated observations or from other information.

The parameter used to indicate the uncertainty may be a standard deviation (or a given multiple of it), or the width of an estimated confidence interval, but any propagation of uncertainty or composition of uncertainty components must be based on the uncertainty variance.

5. STATISTICAL GRAPHICS

One of the most powerful statistical tools is an appropriate graph. The human psychovisual system is unmatched in its ability to process pictorial information, and thereby to detect patterns and relationships. The important word is "appropriate." While the psychovisual system is powerful, it can be misled, and making informative statistical graphs is an acquired skill. While we lack a detailed theory for good statistical graphics, we do have the fruits of much research on the relation between human perception and statistical graphics that can guide the search for effective displays (Cleveland and McGill, 1984).

The subject of statistical graphics is too large for a comprehensive treatment here; thus, we limit the discussion to a few topics appropriate for physical science data. These topics are expanded in works by Chambers *et al.* (1983), Cleveland and McGill (1988), and Cleveland (1994). For a brief history of the subject, see Beniger and Robyn (1978).

We can identify at least three broad categories of statistical graphics: description and communication, analysis, and exploration. Some graphs will fall into more than one category, depending on the context of the data.

5.1 Description and Communication

Pie charts, bar graphs, iconographic displays, and color maps are popular graphics. They are typically found in technical and business reports, in the electronic and print media, and in advertisements, where they are generally used to communicate a concise summary of statistical information. Many computer packages can make these plots and embellish them with color, shading, fill, perspective views, and intensity variations. Some embellishments will distort the data presented and should be avoided. For example, using a perspective view to display a bar graph usually obscures the quantitative information presented, because the data are inherently two dimensional. Another common display technique is to encode numerical magnitudes into iconographic symbols. This can dramatically distort data having a wide range of values when the area of the icon is made proportional to the magnitude being presented.

5.1.1 Pie Charts, Dot Plots, and Bar Graphs The data analyst should avoid pie charts. Pie charts are seldom used by scientists, but are all too common in business and the print media. The problem is that the viewer must make judgments of angles, and the psychovisual system is not good at this; one cannot easily order the variables displayed by their magnitudes, unless the discrepancies are large (Cleveland, 1994). The bar chart is a good alternative when the length of the label is not a problem.

A dot plot is a graphical technique for displaying labeled measurement data. The labels appear as horizontal text to the left of the ordinate axis. Large dots indicate measurement values and are connected by horizontal dotted lines to the corresponding label. Thus, the length of the horizontal line is proportional to the value of the datum. Perception of the data is generally enhanced by ordering the data—say, from smallest at the top to largest at the bottom. One exception is if the labels themselves have a natural ordering. There are a number of variations on the basic dot plots. The bar chart is also a good alternative when the length of the label is not a problem.

5.1.2 Histograms The histogram is the most popular method of displaying the distributional characteristics of a set of data. A histogram is essentially an estimate of the probability density function and is formed by partitioning the data range into a set of intervals or bins (not necessarily equally spaced) and plotting the count of the number of data values that fall in each interval versus the midvalue of the respective bin. The histogram is convenient because it is simple to construct and can be presented with little or no explanation. However, there are serious drawbacks. The shape of the histogram can be sensitive to the number and location of the bins, and it can give an extremely poor estimate of the probability density. Often the bin widths are chosen arbitrarily.

The following formula determines the approximate optimal width of the bins (Diaconis and Freedman, 1981):

$$h = 1.8s(n^{-1}\ln n)^{1/3}, \tag{76}$$

where h is the bin width, n is the number of data points, and s is an estimate of the standard deviation. This formula will approximately minimize the maximum absolute deviation between the resulting histogram and the underlying density distribution. If a robust measure of spread is used instead of the standard deviation (see Sec. 4.3), s should be rescaled to match the standard deviation. For example, the *interquartile range* IQ is the distance between the upper and lower quartiles of the data. In this case, put s = IQ/ 1.349.

5.1.3 Probability-Density Plots Density plots are excellent presentations of the distribution of a set of measurements. Like the histogram, the density plot will usually require little or no explanation and will reveal skewness and multimodality in the distribution of the data.

We discuss nonparametric density estimation, the situation where the analyst has no *a priori* knowledge of the distribution. An excellent method of density estimation is the kernel estimate, discussed in detail by Silverman (1986a). Here the estimate of the density function f at any point x is given by

$$\hat{f}(x) = \frac{1}{nh} \sum K\left(\frac{x - X_i}{h}\right), \tag{77}$$

where n is the number of data points, h is called the window width, and X_i ranges over all the data points. The kernel function K is normalized so that

$$\int_{-\infty}^{\infty} K(x)dx = 1. \tag{78}$$

The choice of the kernel function K is of secondary importance; the Gaussian kernel

$$K(x) = (2\pi)^{-1/2} \exp(-x^2/2) \tag{79}$$

is usually satisfactory. The bias and variance of the density estimate depend critically on the window width h; in fact, the integrated square bias is proportional to h^4 but does not depend on the sample size n, and therefore large n will not necessarily make the bias negligible. On the other hand, the variance is proportional to the inverse of the product nh. Thus the window width provides a trade-off between systematic and random error: As h increases the estimate becomes smoother.

The optimum value of h would minimize the total mean-squared error, but this value depends on the unknown density (Silverman, 1986a). The following rule of thumb for choosing h (Silverman, 1986b) seems to work well:

$$s = \left[\frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \hat{x})^2\right]^{1/2}, \tag{80a}$$

$$A = \min(s, IQ/1.349), \tag{80b}$$

$$h = 1.06An^{-1/5}. \tag{80c}$$

If an accurate estimate of the density is of critical importance, then the analyst should apply more refined techniques (Silverman, 1986a, 1986b).

5.2 Probability Plotting Methods

Probability plotting methods are especially useful tools for data analysis. They are used to compare samples and distributions. We limit the discussion here to one-dimensional data; generalizations for multivariate data are available (Gnanadesikan, 1977; Eastern and McCulloch, 1990). The major probability plots are quantile–quantile plot, Poissonness plot, empirical cumulative distribution plot, percentile plot, and symmetry plot. We confine the discussion here to the quantile–quantile plot and the Poissonness plot; for additional information, the reader should consult the excellent article by Wilk and Gnanadesikan (1968).

5.2.1 The Notion of a Quantile Quantiles are useful for graphing distributions. The p *theoretical* quantile is a number q on the abscissa of the distribution function such that $F(q) = p$; the 0.5 quantile would be median of the distribution. On the other hand, the p *empirical* quantile is defined from the observed data, and it is that observation such that approximately a fraction p of the data is less than the observation and a fraction $1 - p$ is greater. A set of empirical quantiles $\{q_i\}$ for a one-dimensional data sample may be easily calculated. Given a set of samples $\{x_1, x_2, \ldots, x_n\}$, first the data are sorted in ascending order: $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$; then q_i is set equal to $x_{(i)}$, and the p_i associated with q_i is calculated according to

$$p_i = (i - c)/(n - 2c + 1). \tag{81}$$

The choice of a value for c is a minor technical consideration; usually $c = 0.5$, yielding

$$p_i = (i - 0.5)/n, \tag{82}$$

but other values are also suitable depending on the assumption for the underlying distribution (Kimbell, 1960). Given an arbitrary probability p, the associated quantile may be calculated by linear interpolation between the two nearest sample quantiles. After the analyst calculates the quantiles and probabilities, he can construct a number of useful probability plots.

5.2.2 Empirical Quantile–Quantile Plots An especially powerful graphical method for comparing the distributions of two data sets is the empirical quantile–quantile (EQQ) plot. Apparently, this technique was first introduced in 1914 (Eastern and McCulloch, 1990), but was not extensively used until it was reintroduced in the 1960s (Wilk and Gnanadesikan, 1968). We construct an EQQ plot by plotting the quantiles of one data set against the quantiles of the other. If the two data sets have unequal numbers of points, then one uses the probabilities associated with the smaller set, and then adjusts the quantiles in the larger set by interpolation to match the probabilities.

If the two random samples of data come from the same distribution, then the EQQ plot will produce points that fluctuate about the line $y = x$. Departures from this line give detailed information on how the two distributions differ. The EQQ plot is sensitive to differences between the distribution tails. If the points scatter about the line $y = b + ax$, then the two distributions differ as to mean and standard deviation. For example, when data sets come from two different normal distributions, the plotted points will scatter about the line

$$y = (\sigma_y/\sigma_x)(x - \mu_x) + \mu_y. \tag{83}$$

The EQQ plot is the best technique for comparing two distributions when *a priori* knowledge is lacking, better than comparing various numerically calculated statistics or comparing histograms or density functions. When the two distributions are equivalent in the sense that the associated random variables are related by a linear transformation, most methods of comparison will disguise this fact; the EQQ plot will not.

The data analyst should first try to transform each of the data sets, to make them as normally distributed as possible, as this will facilitate interpretation. In some cases this transformation is essential. For example, suppose we have two random variables Y_1 and Y_2, each lognormally distributed as follows:

$$Y_1 = e^X, X \sim N(0,1); \tag{84a}$$

$$Y_2 = e^{aX+b}. \tag{84b}$$

We see that the first lognormal is derived from a normal distribution with zero mean and unit variance. The second is derived from a normal with mean b and standard deviation a. If we compare Y_1 to Y_2 using an EQQ plot, we will not get a straight line. However, if we first transform both Y_1 and Y_2 using the logarithm, then the plotted points will scatter about the line $y = ax + b$ because we are comparing two normal distributions.

5.2.3 Theoretical Quantile–Quantile Plots The theoretical quantile–quantile plot or probability plot can be used to compare an empirical data set against a specified reference distribution. If $F(\cdot)$ is the cumulative distribution of the reference distribution, then the QQ plot is constructed by plotting the empirical quantiles of the data against the theoretical quantiles of the reference distribution given by

$$F^{-1}\left(\frac{i - 0.5}{n}\right). \tag{85}$$

When the reference distribution is the normal distribution, then constructing the theoretical QQ plot is equivalent to using probability paper. However, in the case of probability paper, the probability values are displayed instead of the normal quantiles. For the normal distribution,

$$F^{-1}(x) = \Phi^{-1}(x), \tag{86a}$$

where

$$\Phi(x) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{x} \exp\left(\frac{-t^2}{2}\right) dt. \tag{86b}$$

Rational approximations for the inverse normal distribution Φ^{-1} are available (Abramowitz and Stegun, 1970). These approximations are derived by a rearrangement of truncated Chebyshev expansions—the method of "telescoping" of power series (Lanczos, 1956).

5.2.4 Interpreting QQ Plots Interpreting both empirical and theoretical QQ plots takes some practice. The beginner tends to expect more order and linearity than he usually gets. Training plots to help the beginner gain experience are available (Daniel, 1976; Daniel and Wood, 1980). In some cases, analysts will find it useful to have a summary statistic for a probability plot. In fact, the theoretical QQ plot with a normal distribution is related to one of the most powerful omnibus statistical tests for normality—the Shapiro–Wilk test (Shapiro and Wilk, 1965), based on the correlation between the observed order statistics of the sample and the expected order statistics from the standard normal distribution (zero mean, unit variance). High values of correlation (near unity) accept normality; low values reject it. Published tables provide critical values (Shapiro and Francia, 1972) of the correlation coefficient using the following approximation for the expected order statistics for the normal distribution (normal quantiles):

$$q_i = \Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right). \tag{87}$$

The critical values of the correlation coefficient are near unity.

5.2.5 Poissonness Plot The Poissonness plot (Hoaglin, 1980) is a probability plot designed to help diagnose whether the Poisson probability distribution is the appropriate model for a set of counted data (non-negative integer quantities). This plot is an excellent way of detecting aberrant points in an otherwise well-behaved distribution.

To construct the plot, the measurements must first be sorted into bins and the number of counts in each bin recorded. The bins should sequentially span the integers from zero to the maximum number of observed counts. Thus, K_0 is the number of measurements that are 0, K_1 is the number of measurements that are 1, etc.; some of the K_i may be zero. The total number of measurements N must satisfy

$$N = \sum K_k. \tag{88}$$

If the counts are Poisson distributed, then $\text{Prob}[K = k] = p(k)$ given by Eq. (37). The expected value of K_k is

$$E[K_k] = Ne^{-\lambda}\lambda^k/k!, k = 0, 1, \ldots. \tag{89}$$

If we take the natural logarithm of both sides, then

$$\ln(K_k) = \ln(N) - \lambda + k\ln(\lambda) - \ln(k!). \tag{90}$$

We plot $\ln(k!K_k)$ against k, leaving out points where $K_k = 0$. If the resulting plot looks straight, then the Poisson distribution is appropriate. The advantage of using the Poissonness plot is the aberrant points may be easily identified as those that obviously depart from a straight line. If the Poissonness plot indicates that the Poisson model is appropriate, and there are not aberrant points, then the maximum-likelihood estimate for λ,

$$\hat{\lambda} = \frac{1}{N}\sum kK_k, \tag{91}$$

can be used. Otherwise, we fit a straight line that ignores (or down weights) the aberrant points; the estimate of $\ln(\lambda)$ is the slope of the fitted line. If possible, the cause of the aberrant points should be determined.

5.3 Exploration

The data analyst uses exploratory graphs to uncover features and patterns in the data without assumptions about the behavior or the mechanism that generated the data. They help the analyst discover appropriate transformations and distributional models. There is an extensive literature of graphics for exploratory data analysis (Tukey, 1977; Chambers *et al.*, 1983; Snee and Pfeifer, 1983). The discussion here is limited to scatter plots.

5.3.1 Scatter Plots The scatter plot is used to study the relationship between two variables; to be used effectively, it must be carefully planned and embellished to bring out features in the data. Sometimes one of the variables is a response and the other a factor, and the analyst seeks to discover a "lawlike" relationship between the variables; in other cases, the goal is to determine the statistical dependence between two variables, where neither is regarded as a response or factor. It is important to choose the appropriate plotting symbol, aspect ratio, tic-mark spacing, and data span for the axes (Cleveland and McGill, 1984). When one is looking for statistical dependence, the plotting symbols should not be connected by lines, and no grid lines should appear in the body of the plots (Cleveland, 1994).

5.3.2 Time-Series Plots A special type of scatter plot obtains when one of the variables consists of uniformly spaced values—the time-series plot. Here we have special considerations that do not apply to scatter plots where neither variable is uniformly spaced. When we study time-series data, we look for trends, seasonal effects, short-term correlation, oscillations, jump phenomena, spikes, and outliers. We especially want to know whether the series is evolving with statistical uniformity. What we discover from the time-series plot influences how we model the data. There are four basic types of time-series plots, each enhancing a different aspect of the data:

1. The symbol plot. Here the data are plotted as symbols without connecting lines. This plot is best for studying long-term trends in the data because it minimizes the visual impact of outliers. However, it also inhibits the perception of local time order and any periodic components.
2. The connected-symbol plot. Here the observations are plotted as symbols with connecting lines. This plot enhances local time order and works well for noisy data since the individual points are perceived. However, plots with many points look cluttered.
3. The connected-point plot: The observations are plotted as points with connecting lines. This type of plot, while popular, is usually the least appropriate choice. The individual points lose their identities, and detection of positive autocorrelation (short-term smoothness) is difficult. Noise and outlier observations tend to create a cluttered and distracting picture. Use this plot only for data with a high degree of overall smoothness.
4. The vertical-line plot. Here, the observations are plotted as vertical lines to the reference baseline. This type of plot is appropriate when local time order is important and the number of points is large. The disadvantage of this plot is that outliers have enhanced visual impact.

5.3.3 Box Plots Box plots are good for an initial look at data (McGill and Larsen, 1978; Chambers *et al.*, 1983). The order statistics are coded into a rectangular icon, with a horizontal line segment marking the median and vertical lines extending to outliers. The analyst can quickly identify aberrant points and skewness in the underlying distribution. If the data can be grouped into various categories, then a series of icons is appropriate.

5.4 The Use of Color in Statistical Graphics

Using color effectively requires considerable skill. The beginner should use one color for the basic plotting frame and axis labeling (black or white) and strictly limit the use of color when embellishing the plot—better no color than misused color. Using too many colors, which results in a confusing jumble, is the most common mistake. While the color encoding of categorical variables works well, assigning different hues to continuously varying numerical magnitudes should be avoided because it usually makes the plot more difficult to interpret. Using gray levels works better. There are other potential problems in using color; size distortion effects can occur in the coloring of statistical maps, since saturated primary colors apparently cause misperception of the relative sizes of colored areas (Cleveland and McGill, 1983). These authors also point out that we do not fully understand the effect of color on people's perceptions.

5.5 Dynamic Graphics

Dynamic graphics is a technique for the visualization of three- (or higher) dimensional data by using interactive computer animation. In most implementations, the data are presented as a rotating point cloud where the user can select the axis of rotation and various features of the display; subsets can be highlighted, marked or labeled, or masked out. The movement of the point cloud in real time gives an effective illusion of depth.

One of the original implementations was the PRIM-9 system (Fishkeller *et al.*, 1974), designed for the analysis of high-energy physics research data. This system used special display hardware connected to a mainframe computer (IBM 360/91) and is the prototype for modern interactive graphics workstations. All of the features of PRIM-9 (and more) have become available since the mid-1980s on stand-alone applications that run on personal computers and workstations.

6. THE IMPACT OF COMPUTERS ON DATA ANALYSIS

In the 1950s and 1960s the routine (as opposed to research) use of computers in statistics was generally limited to automating classical procedures. That state of affairs began to change in the 1970s with the introduction of computer-intensive methods for doing statistics, such as the *bootstrap* (Efron, 1979), and techniques for handling high-dimensional data (Friedman and Tukey, 1974). The quality of the numerical software also began to improve significantly in the late 1970s. For example, Golub had introduced orthogonal techniques for doing least-squares problems as early as 1968, but it took nearly a decade for that superior and essential numerical technology to appear in the regression routines in statistical packages. By the mid-1990s computers had a significant impact on all phases of data analysis activity—data collection, storage, organization, retrieval, editing, display, analysis, and report generation. Moreover, the cost of doing calculations has become almost trivial, and so millions of calculations are feasible even for routine problems. We discuss a few computer-intensive methods that are especially interesting and useful to the data analyst.

Not all statistical computer techniques are new; some had to wait for the computer to be implemented. A good example is Fisher's *method of randomization.* First proposed in the 1930s by R. A. Fisher, tests based on the method of randomization using a maximally appropriate test statistic have very desirable mathematical properties—most powerful, consistent, unbiased, etc. However, despite these excellent mathematical properties, the method of randomization remained little more than a statistical curiosity, having been used by Fisher in a purely conceptual manner in doing certain proofs. The problem is that the distribution of a test statistic is conditional on the observed data. Each new application of the method has a new and unique distribution, making it impossible to compute a set of tables that may be used repeatedly. This forced the use of nonparametric techniques, where the observed data are replaced by their ranks and the test is carried out on the ranked data. This causes a loss of information and diminishes the power of the test as compared to the corresponding distribution-based parametric test, or the test using the method of randomization. With fast and inexpensive computing, Fisher's method has been resurrected (Good, 1994).

6.1 Computer-Intensive Methods for Parameter Estimation

In many cases we lack knowledge of the distribution that describes the data under consideration, and we have too few samples to fit a distribution with any confidence. The sample mean and its standard error $S/\sqrt{N}$, where S is given by Eq. (58), provide an *automatic algorithm* for estimating the mean and the accuracy of the estimate, regardless of the actual distribution that describes the data—the distribution need only have finite first and second moments. Unfortunately, the minimum-variance unbiased-estimator approach does not provide automatic algorithms for parameters other than the mean. However, if we are willing to perform many numerical calculations, we can find the accuracy of parameter estimates from the sample data alone.

Suppose we have a sample $\{y_1, y_2, \ldots y_N\}$ from a distribution F, and θ is the parameter of interest. If we use maximum-likelihood estimation to determine θ, the accuracy of the estimate is $1/\sqrt{(\text{Fisher information})}$. The *Fisher information* is defined as the reciprocal of the variance of an efficient estimate of a parameter. It is a fundamental quantity that measures the amount of information about the parameter being estimated that is contained in the sample data. Mathematically, the Fisher information $J(\theta)$ is given by

$$J(\theta) = E_0\left(\frac{\partial}{\partial\theta}\log[f(x;\theta)]\right)^2. \tag{92}$$

The Fisher information has the right property as an information measure because $J(\theta)$ for N independent identically distributed random variables is N times the $J(\theta)$ of one of the variables. The important connection to the accuracy of estimators is given by the Cramér–Rao inequality, which states that for unbiased $\hat{\theta}$,

$$\mathrm{var}[\hat{\theta}] \geq 1/J(\theta). \tag{93}$$

Therefore, $J(\theta)$ gives us a lower bound on the mean-squared error of the estimator and thus provides a measure of the accuracy of estimators. The problem is that often we cannot calculate $J(\theta)$ because the functional form of the estimator is too complicated, and we do not know F. If we knew F, then we could use computer simulation to resample from F, and the simulation would provide information on the variance of the estimator.

The difficulty of not knowing F can be overcome as follows. We can calculate a surrogate $\hat{F}_n$ for F, known as the *empirical cumulative distribution function* (Wilk and Gnanadesikan, 1968), from the primary sample $\{y_1, y_2, . . . y_N\}$. And we can get a surrogate sample $\{y_1^*, y_2^*, . . . , y_N^*\}$ by a process of resampling from $\{y_1, y_2, . . . y_N\}$. Efron recognized that the relationship between $\{y_1, y_2, . . . y_N\}$ and F is similar to the relation between $\{y_1^*, y_2^*, . . . , y_N^*\}$ and $\hat{F}_n$ (Efron, 1979). If we generate many realizations of $\{y_1^*, y_2^*, . . . , y_N^*\}$ by a repeated resampling from $\{y_1, y_2, . . . y_N\}$, we can calculate a new parameter estimate for each realization. Efron coined the term *bootstrap sampling* for this process, and the distribution of $\hat{\theta}$ is called the *bootstrap distribution* of the parameter estimate. While the bootstrap distribution of the parameter estimate is not identical to the distribution that would be obtained by actually sampling from F, it is usually close, and it can be used to calculate many useful things, such as the standard error of the estimate.

Application of the bootstrap technique to compute the standard error of estimates is straightforward. Given a sample $\{y_1, y_2, . . . y_N\}$, B bootstrap replications are generated by use of simple random sampling with replacement from the primary sample. This means that each of the y_n has probability equal to $1/N$ of being drawn in any one replication. Estimate values of $\hat{\theta}_i{}^*$ are generated by using each of the B replications as if it were the primary data, and the bootstrap estimate and its standard error are then given by

$$\hat{\theta}^* = \frac{1}{B}\sum_{i=1}^{B} \hat{\theta}_i^*, \tag{94a}$$

$$\mathrm{SE}(\hat{\theta}) = \left(\frac{1}{B-1}\sum_{i=1}^{B} (\hat{\theta}_i^* - \hat{\theta}^*)^2\right)^{1/2}. \tag{94b}$$

The above gives an automatic algorithm for finding the standard error for a wide and useful class of estimates, including very complicated functions of the sample data such as an eigenvalue of a random matrix. However, the bootstrap procedure can fail under some circumstances. If the estimate is not a smooth enough function of the data, the empirical distribution will not be a good approximation to the actual distribution. If resources permit, it is a good idea to inspect a histogram of the bootstrap replications of the sample. For refinements to the procedure to be used when something goes wrong, see Efron and Tibshirani (1993).

How many replications are necessary? Surprisingly few! Efron says that B rarely needs to exceed 200, and $B = 50$ often produces satisfactory estimates of the standard error (Efron and Tibshirani, 1993). However, large numbers of replications are needed to compute confidence intervals.

6.2 Modern Computing Resources for Data Analysis

Commercial data analysis software originally came in the form of packages designed to be run on mainframe machines and minicomputers in batch environments. Many of these original packages then migrated to the time-sharing environment and finally to workstations and personal computers. The early statistical software packages often failed to provide for data organization, retrieval, and editing, or they would implement these functions with a difficult and limited command syntax—a legacy of their evolution from mainframe machines. Statistical graphics capabilities were either nonexistent or primitive. By the 1990s the situation had changed. Packages that exploit graphical user interfaces (GUIs) became available for workstations and personal computers, thereby making the tasks of data analysis much easier to carry out.

The selection of an appropriate computer platform for doing data analysis is a daunting task, given the variety of hardware, operating systems, and application software available. The situation is further complicated by the nature of the computer industry itself. Both hardware and software are constantly changing, and both can become obsolete very quickly. Nevertheless, a computer platform is essential for doing data analysis, and some choices must be made. Here we try to give some guidelines based on our personal experience, with the caveat that we cannot guarantee the accuracy, timeliness, or completeness of this information.

There are three basic (but not independent) choices to be made: hardware, operating system, and application software. Usually the choice of hardware determines the operating system, and so we will describe the major operating systems. As of 1995, the three dominant operating systems for workstations and personal computers are Unix, DOS/Windows, and MacOS, with DOS/Windows having about 90% of the personal computer market. Unix is the operating system generally found on workstations, which are more costly than personal computers. The applications that run under Unix on workstations also tend to be much more expensive than those that run on personal computers. Unix is the most stable operating system, but it is also the most difficult for the inexperienced user. More applications are available for DOS/Windows than for MacOS, but many of the useful programs are available for both. MacOS and the associated hardware tend to be easier to use and configure. Almost any decision made will be rendered obsolete by technical changes in a few months. We believe that the prudent strategy is to select the most stable system, which generally means choosing hardware and software that have been operating reliably for at least a year, rather than the latest technology.

6.2.1 Software Packages for Doing Data Analysis Some of the data-analysis software packages developed for modern workstations and personal computers provide for almost all of the necessary functions to do good data analysis. They are strong on graphics and are highly interactive. A good example is the *S* system for data analysis and graphics, originally developed by Bell Laboratories but now only available as SPLUS from Statistical Sciences Inc., which holds exclusive license. *S* is also a full language where the user can easily define new functions and extend the capabilities of the package. Compiled FORTRAN and C code can also be linked. SPLUS is available for most popular Unix-based platforms. We have no experience with the available DOS/Windows version, and no version is currently available for MacOS. There is extensive documentation available for SPLUS, including four text books. An active news group (s-new@utstat.toronto.edu) provides regular help on using *S*. The s-news traffic has also been archived at statlib with a good search engine.

A MacOS application designed especially for exploratory data analysis is *Data Desk,* available from the Data Description Company. This package exploits the Macintosh graphical interface operating system. Other heavily marketed packages are *JMP, Systat,* and *Statview,* available for both MacOS and DOS/Windows. *Matlab* is mainly a numerical linear algebra package; it can be used for data analysis, but it lacks any extensive set of specific built-in functions. *Mathematica* and *Maple* are both packages for doing symbolic and numerical algebra and calculus. Mathematica has versions that run on all the major platforms. Both are more useful for mathematical statistics than for data analysis. Each has a powerful programming language with a steep learning curve. From what is currently available, we believe the system of choice for doing data analysis is *S* running on a workstation.

An alternative to commercial data-analysis packages is provided by subroutine libraries, such as those distributed by The International Mathematical and Statistical Libraries Inc. (IMSL) and the National Algorithms Group Library (NAG). Both are costly, but NAG is known to distribute reliable, high-quality software. The disadvantage of these libraries is that the user must write his own software around the subroutines, which can be costly and time consuming. This approach is recommended for either specialized problems or production applications.

GLOSSARY

Bayesian: One who interprets probability subjectively. A Bayesian can assign a probability, in the sense of a "degree of belief," to a nonrepetitive event (contrast with **Frequentist**).

Bootstrap: A technique for calculating estimation error without knowing the distribution of the data.

Confidence Interval: A measure of the uncertainty of a parameter estimate, in the form of an interval within which the value of the parameter is expected to lie with a stated probability.

Cumulants: A set of numbers that describe a probability distribution, particularly how much it deviates from the normal distribution. An often simpler alternative to moments. The cumulants are, in a sense, the orthogonal components of the moments.

Entropy: An information-theoretic concept measuring the uncertainty of a distribution. Related to, but distinct from, thermodynamic entropy.

Estimator: A statistic (a function of the data) that provides an approximation to the value of a parameter of the probability distribution that describes the data.

Expectation: The average or mean value of a probability distribution.

Fisher Information Matrix: A measure of the fundamental amount of information in a random sample. It is additive over independent experiments; the information in n data points is n times the information in one data point.

Frequentist: One who interprets probability as the limit of the ratio of the number of times an event occurs to the number of experiments in which it could occur. A frequentist will not assign a probability to an event that is fundamentally nonrepetitive (contrast with **Bayesian**).

Independent Events: Two events are independent if the occurrence of one provides no information about the probability of occurrence of the other. The joint distribution of independent variables is the product of the independent distributions.

Inference: The process of drawing conclusions about the distribution of a population on the basis of a sample from that population. Inference is frequently with respect to the determination of the parameters of a family of distributions.

Likelihood: A parameter in statistical inference, associated with an assumed model, proportional to the probability of the data given the truth of the model. The likelihood is not necessarily a probability.

Measurand: A particular physical quantity subject to measurement.

Moment: One of a set of parameters of a probability distribution that specifies its properties and that can often completely define it. The most common moments are defined in a manner analogous to moments in mechanics.

Parameter: A scalar or vector quantity, either discrete or continuous, that indexes the probability distribution with respect to some family of distributions.

Random Variable: A mapping from outcomes of an experiment to numbers. A variable that may take any of a set of values, and with which is associated a probability distribution.

Residual: The difference between the observed value and the fitted (estimated) value.

Resistant Estimate: A data-analysis estimate of a model parameter not sensitive to outliers in the data. A numerical concept.

Robust Estimate: An estimate (of a parameter) that is not sensitive to the form of the distribution of the data. The robustness is with respect to the inference to be drawn from the available data. A statistical concept.

Statistical Control: The condition achieved when the parameters defining the distribution can be assumed to be constant, so that the statistical characteristics of the process are stable in time.

Statistical Model: A set of probability laws that describe or determine experimental or observational outcomes. Often synonymous with assumed distribution.

Works Cited

Abramowitz, M. and I. A. Stegun (1970), *Handbook of Mathematical Functions,* Washington, D.C.: U.S. GPO.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., Tukey, J. W. (1972), *Robust Estimates of Location: Survey and Advances,* Princeton, NJ: Princeton University Press.

Beniger, J. R., Robyn, D. L. (1978), *Am. Stat.* **31,** 1–11.

Bertotti, B., Brill, D., Krotkov, R. (1962), *Gravitation,* New York: Wiley.

Chambers, J. M., Cleveland, W. S., Kleiner, B., Tukey, P. A. (1983), *Graphical Methods for Data Analysis,* Bellmont, CA: Wadsworth.

Cleveland, W. S., McGill, M. E. (1983), *Am. Stat.* **37,** 101–105.

Cleveland, W. S., McGill, M. E. (1984), *J. Am. Stat. Assoc.* **79,** 807–822.

Cleveland, W. S., McGill, M. E. (Ed.) (1988), *Dynamic Graphics for Statistics,* Bellmont, CA: Wadsworth & Brooks/Cole.

Cleveland, W. S. (1994), *The Elements of Graphing Data,* Summit, NJ: Hobart Press.

Cover, T. M., Thomas, J. A. (1991), *Elements of Information Theory,* New York: Wiley.

Daniel, C. (1976), *Applications of Statistics to Industrial Experimentation,* New York: Wiley.

Daniel, C., Wood, F. S. (1980), *Fitting Equations to Data,* New York: Wiley.

DeFinetti, B. (1972), *Probability, Induction, and Statistics: The Art of Guessing,* New York: Wiley.

Deming, W. E. (1950), *Some Theory of Sampling,* New York: Wiley.

Diaconis, P., Freedman, D. (1981), *Z. Wahrscheinlichkeitstheorie und verwandte Gebiete* **58,** 139–167.

Eastern, G. S., McCulloch, R. E. (1990), *J. Am. Stat. Assoc.* **85,** 376–386.

Efron, B. (1978), *Am. Math. Monthly* **85,** 231–246.

Efron, B. (1979), *Ann. Math. Stat.* **7,** 1–26.

Efron, B., Tibshirani, R. J. (1993), *An Introduction to the Bootstrap,* New York: Chapman and Hall.

Einstein, A. (1915), *Sitzungsber. Preuss. Akad. Wiss. (Berlin)* **47,** 831–839.

Eisenhart, C. (1963), *J. Res. Natl. Bur. Stds.* **67C,** 161–187.

Fisher, R. A. (1915), *Biometrika* **10,** 507–521.

Fishkeller, M. A., Friedman, J. H., Tukey, J. W. (1974), "PRIM-9: An Inter-Active Multidimensional Data Display and Analysis System," Stanford Linear Acclerator Center Report.

Friedman, J. H., Tukey, J. W. (1974), *IEEE Trans. Comput.* **C-23,** 881–890.

Frodesen, A. G., Skjeggestad, O. *et al.* (1979), *Probability and Statistics in Particle Physics,* Bergen: Universitetsforlaget.

Gayen, A. K. (1951), *Biometrika* **38,** 219–247.

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations,* New York: Wiley.

Gnedenko, B. V., Kolmogorov, A. N. (1954), *Limit Theorems for Sums of Independent Random Variables,* Reading, MA: Addison Wesley.

Good, P. (1994), *Permutation Tests,* New York: Springer-Verlag.

Grant, E. L., Leavenworth, R. S. (1980), *Statistical Quality Control,* New York: McGraw-Hill.

Hampel, F. R. (1973), *Z. Wahrscheinlichkeitstheorie verwandte Gebiete* **27,** 87–104.

Hoaglin, D. C. (1980), *Am. Stat.* **34,** (3), 146–149.

Hoaglin, D. (1983), *Understanding Robust and Exploratory Data Analysis,* New York: Wiley.

ISO (1993a), *International Vocabulary of Basic and General Terms in Metrology,* Geneva: International Organization for Standardization.

ISO (1993b), *Statistics—Vocabulary and Symbols. Part 1, Probability and General Statistical Terms,* Geneva: International Organization for Standardization.

Jaynes, E. T. (1957), *Phys. Rev.* **106,** 620–630.

Johnstone, I., Velleman, P. (1981), ASA Annual Conference, Washington, D.C., Section of Statistical Computing, American Statistical Association.

Kendall, M., Stuart, A. *et al.* (1983), *The Advanced Theory of Statistics,* New York: MacMillan Publishing Co.

Kimbell, B. F. (1960), *J. Am. Stat. Assoc.* **55,** 546–560.

Lanczos, C. (1956), *Applied Analysis,* Englewood Cliffs, NJ: Prentice Hall.

Lax, D. A. (1975), "An Interim Report of a Monte Carlo Study of Robust Estimators of Width," Department of Statistics, Princeton University, Report.

Lindley, D. (1985), *Making Decisions,* New York: Wiley.

Misner, C. W., Thorne, K. S., Wheeler, J. A., (1973), *Gravitation,* San Francisco: W. H. Freeman.

McCullagh, P. (1987), *Tensor Methods in Statistics,* London: Chapman and Hall.

McGill, R. T., Tukey, J. W., Larsen, W. A. (1978), *Am. Stat.* **32,** 2–16.

Mosteller, F., Tukey, J. W. (1977), *Data Analysis and Regression,* Reading, MA: Addison Wesley.

Moyer, D. F. (1979), *On the Path of Albert Einstein,* New York: Plenum, p. 55.

Pais, A. (1982), *Subtle Is the Lord,* Oxford: Clarendon Press.

Ramsey, F. P. (1980), "Truth and Probability," in: H. E. Kyberg, Jr., H. E. Smokler (Eds.), *Studies in Subjective Probability,* Huntington, NY: Robert E. Krieger Publ. Co.

Shapiro, S. S., Wilk, M. B. (1965), *Biometrika* **52,** 591–611.

Shapiro, S. S., Francia, R. S. (1972), *J. Am. Stat. Assoc.* **67,** 215–216.

Shewhart, W. A. (1931), *Economic Control of*

Quality of Manufactured Product, New York: Van Nostrand.

Silverman, B. S. (1986a), *Density Estimation in Statistics and Data Analysis,* London: Chapman and Hall.

Silverman, B. S. (1986b), Short Course Presented at the ASA Annual Meeting, Chicago, IL, August, 1986.

Snee, R. D., Pfeifer, C. G. (1983), "Graphical Representation of Data," in: S. Kotz, N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences,* Vol. 3, New York: Wiley, pp. 448–511.

Stigler, S. M. (1986), *The History of Statistics,* Cambridge, MA: Belknap Press.

Student (Gosset, Wm. S.) (1908), *Biometrika* **6,** 1–25.

Tukey, J. W. (1960), "A Survey of Sampling from Contaminated Distributions," in: S. Olkin, W. Ghurye, W. Hoeffding, W. Madow, H. Mann (Eds.), *Contributions to Probability and Statistics,* Ames, IA: Iowa State University Press, pp. 293–316.

Tukey, J. W. (1977), *Exploratory Data Analysis,* Reading, MA: Addison Wesley.

Tukey, J. W., Braun, H. I. *et al.* (1977), Further Progress on Robust/Resistant Widthing, Department of Statistics, Princeton University.

Velleman, P., Hoaglin, D. (1981), *Applications, Basics and Computing of Exploratory Data Analysis,* Boston: Duxbury Press.

Wilk, M. B., Gnanadesikan, R. (1968), *Biometrika* **55,** 1–17.

Further Reading

Akaike, H. (1985), "Prediction and Entropy," in: A. C. Atkinson, S. E. Fienberg (Eds.), *A Celebration of Statistics,* New York: Springer-Verlag.

Bradley, J. V. (1968), *Distribution-Free Statistical Tests,* Englewood Cliffs, NJ: Prentice Hall.

Cleveland, W. S. (1994), *The Elements of Graphing Data,* Summit, NJ: Hobart Press.

Efron, B., Tibshirani, R. J. (1993), *An Introduction to the Bootstrap,* New York: Chapman and Hall.

Kapur, J. N., Kesavan, H. K. (1992), *Entropy Optimization Principles with Applications,* New York: Academic Press.

Lindgren, B. W. (1976), *Statistical Theory,* New York: MacMillan.

Madansky, A. (1988), *Prescriptions for Working Statisticians,* New York: Springer-Verlag.

McCullagh, P. (1987), *Tensor Methods in Statistics,* London: Chapman and Hall.

Miller, R. G. J. (1986), *Beyond Anova, Basis of Applied Statistics,* New York: Wiley.

Rabinovich, S. (1995), *Measurement Errors: Theory and Practice,* New York: American Institute of Physics.

Sakamoto, Y., Ishiguro, M. *et al.* (1986), *Akaike Information Criterion Statistics,* Boston: D. Reidel Publishing Co.

Stuart, A., Ord, J. K. (1994), *Kendall's Advanced Theory of Statistics,* New York: Halsted Press.

Velleman, P., Hoaglin, D. (1981), *Applications, Basics and Computing of Exploratory Data Analysis,* Boston: Duxbury, Press.

Whittle, P. (1976), *Probability via Expectation,* New York: Springer-Verlag.

Wilk, M. B., Gnanadesikan, R. (1968), *Biometrika* **55,** 1–17.

STATISTICAL MECHANICS, CLASSICAL

J. R. DORFMAN, *Institute for Physical Science and Technology and Department of Physics, University of Maryland, College Park, Maryland, U.S.A.*

Introduction 497

1. **Phase Space and Ensemble Distribution Functions** 499

2. **The Microcanonical Ensemble** 499

3. **The Ergodic Hypothesis** 501

4. **The Canonical and Grand Canonical Ensembles** 502

5. **Applications of Equilibrium Statistical Mechanics** 505

5.1 The Properties of Ideal Systems 505

5.2 Fluctuations 506

5.3 Distribution Functions 507

5.4 Systems of Interacting Particles 508

5.5 Phase Transitions 509

6. **Nonequilibrium Statistical Mechanics** 510

7. **Conclusion** 513

Glossary 513

Works Cited 514

Further Reading 515

INTRODUCTION

A remarkable feature of bulk matter—gases, liquids, solids, plasmas, and so on—is that the properties of systems composed of large numbers of particles ($\sim 10^{23}$) can be often described by very simple physical laws. Examples of such laws are the equations of equilibrium thermodynamics, of fluid dynamics, of solid mechanics, and of magnetohydrodynamics. The aim of statistical mechanics is to predict the properties of systems composed of large numbers of particles in terms of the mechanical properties of the individual particles and of the forces between them. One of the goals of this field is to explain why the properties of macroscopic systems obey simple laws and to calculate the specific properties of each system of interest.

Statistical mechanics can be divided into two main areas, each of which attempt to explain a body of experimental results for large systems. *Equilibrium statistical mechanics* attempts to derive the laws of thermodynamics and to calculate thermodynamic functions and other equilibrium properties of matter in terms of the microscopic properties of the system of particles. *Nonequilibrium statistical mechanics* attempts to derive the equations that describe transport of mass, energy, and momentum in the system; to describe the approach to thermodynamic equilibrium; and to compute the values of macroscopic, nonequilibrium properties such as the coefficients appearing in the transport equations in terms of the microscopic behavior of the system of particles.

The central idea of statistical mechanics is the following: Suppose one is interested in computing some property of a system composed of a large number of particles that depends in some way on the dynamical processes in the system. For example, one might want to compute the force that a fluid exerts on a piston placed in the fluid, or the flow of energy across a small area in the system over a small time interval. It is not possible to compute these properties exactly for any individual system, since the equations of motion are too complex, and the initial conditions needed to solve them are not usually known. Instead the property of interest is computed by determining its *average* value, taken over a suitable ensemble of similarly prepared systems.

Statistical mechanics is successful in predicting properties of large systems because

1. the values of most of the quantities of experimental interest are not sensitive to

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

the particular microscopic state of the system, and
2. the ensemble average is usually much easier to compute than the exact value.

Fluctuations about the average value are usually small provided the system is sufficiently large and the ensemble has been properly constructed. The fact that the system is large, while leading to enormous complexity on the microscopic scale, leads to regular, simply describable behavior on a macroscopic scale.

The science of statistical mechanics has its origin in the mid-nineteenth century. Scientists such as J. Herapath and J. J. Waterston tried, with some degree of success, to explain experimental observations of Boyle's law, and of the ratio of specific heats of gases, on the basis of a simple, atomic description of a gas. Their ideas were extended by Joule and by Clausius, and the subject became a major research area with notable successes due to work of Maxwell and Boltzmann on the kinetic theory of gases, toward the end of the century (Brush, 1976).

Maxwell and Boltzmann both realized that the arguments they had advanced to understand the thermodynamic and transport properties of dilute gases could be extended to more complicated systems. In 1877 Boltzmann published the formula

$$S = k_B \log W, \tag{1}$$

where W is the probability of a particular macroscopic state of the system, k_B is Boltzmann's constant,

$$k_B = 1.380\,658 \times 10^{-23}\,\mathrm{J\,K^{-1}}, \tag{2}$$

and S is the entropy of the system in the particular macroscopic state. In order to provide a foundation for this formula and to justify its use in computing the thermodynamic properties of large systems for particles, Maxwell and Boltzmann introduced the "ergodic hypothesis," namely, that every microscopic state of the system, consistent with an overall constraint of fixed energy, is as likely to occur as any other. Using this formulation for the relation between the thermodynamic entropy and the underlying microscopic description of a system, Maxwell and Boltzmann were able to provide a basis for the computation of the thermodynamic properties of large systems of interesting particles that included, but was not restricted to, dilute gases. This, together with a study of the ergodic hypothesis, has been a major topic of research since their time. The modern formulation of equilibrium statistical mechanics is due to J. W. Gibbs, who based the calculation of thermodynamic properties of a system on the use of an appropriate ensemble and described the properties of the microcanonical, the canonical, and the grand canonical ensembles (Gibbs, 1960). Because of its elegance and simplicity, Gibbs's formulation of statistical mechanics is the one used today.

Of course, even with the methods of Maxwell, Boltzmann, and Gibbs, certain features of the thermodynamic properties of systems could not be explained until the correct mechanical picture of microscopic systems become clear with the advent of quantum mechanics in the twentieth century.

While the work of Maxwell, Boltzmann, and Gibbs led rather directly to a complete formulation of equilibrium statistical mechanics, the development of nonequilibrium statistical mechanics has taken a much longer time. This branch of statistical mechanics was, in fact, a central focus of the study of the properties of dilute gases by Maxwell and Boltzmann. The Boltzmann transport equation for dilute gas is still a topic of modern research, in both physics and mathematics, and much recent work has been devoted to extending it to higher densities, and to finding appropriate descriptions of transport in other, more complicated systems.

In this article the foundations of both equilibrium and nonequilibrium statistical mechanics for classical systems are presented. The extension to quantum mechanical systems and the application of statistical mechanics to specific problems are discussed in other articles (see STATISTICAL MECHANICS, QUANTUM). For many purposes the full quantum treatment of the microscopic state of the system is not required. For example, we can use classical mechanics to treat the translational motions of particles in gases and liquids at room temperatures and quantum mechanics to treat their internal states. When there is a slow exchange of translational and internal energies, a treatment of many of the macroscopic properties of a fluid can be based entirely on classical statistical mechanics.

1. PHASE SPACE AND ENSEMBLE DISTRIBUTION FUNCTIONS

As stated above, the essential idea of statistical mechanics is to replace a computation of the exact value of some mechanical quantity of interest for a large system by the computation of the ensemble average of that quantity, taken over an ensemble of mechanically identical, and appropriately prepared, systems. To set up the ensemble we consider a collection of N particles, each with f canonical coordinates and f canonical momenta, and with these we construct a $2Nf$-dimensional space, called phase space or Γ space. One point in Γ space specifies the complete set of coordinates and momenta for every particle at one instant of time. The time development of the system is represented by a trajectory in Γ space, which is determined by the solution of the equations of motion of the system.

An ensemble is constructed by considering all possible replicas of the system that satisfy the same macroscopic conditions at some initial instant of time, and assigning an initial probability distribution $\rho(\Gamma)$ such that $\rho(\Gamma)d\Gamma$ is the probability that one of the replicas, chosen at random, will be in the $2Nf$-dimensional volume element $d\Gamma$ about the point Γ in Γ space at the initial instant. The fact that each system in the ensemble develops in the course of time according to the equations of motion implies that the probability distribution function $\rho(\Gamma,t)$ is a function of time, having the value $\rho(\Gamma)$ only at the initial instant. If the classical equations of motion for the canonical coordinates and momenta are Hamilton's equations

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \qquad \dot{p}_i = -\frac{\partial H}{\partial q_i}, \tag{3}$$

for $i = 1, \cdots, Nf$, where (q_i,p_i) represent the canonical coordinates and momenta, respectively, for the Nf degrees of freedom for the system, and $H(q_1,p_1,\cdots q_{Nf},p_{Nf})$ is the Hamiltonian for the system of N particles, then $\rho(\Gamma,t)$ satisfies the equation

$$\frac{\partial \rho(\Gamma,t)}{\partial t} + \{\rho(\Gamma,t),H\} = 0. \tag{4}$$

Here $\{\rho(\Gamma,t),H\}$ is the Poisson bracket of ρ and H. Equation (4) is called the classical Liouville equation. The ensemble average of a dynamical quantity $A(\Gamma)$ is, in general, time dependent and is denoted by $\langle A(\Gamma,t)\rangle$ where

$$\langle A(\Gamma,t)\rangle = \int d\Gamma A(\Gamma)\rho(\Gamma,t), \tag{5}$$

where we use

$$\int d\Gamma \rho(\Gamma,t) = \int d\Gamma \rho(\Gamma) = 1. \tag{6}$$

Equation (6) is a consequence of Liouville's equation and shows that the interpretation of $\rho(\Gamma,t)$ as a probability density holds for all time, not only at the initial instant. One further important consequence of Liouville's equation is the *conservation of volume in phase space.* That is, consider a region in Γ space occupying some Γ-space volume $V_\Gamma(0)$, and follow the region in time, as the points in the region follow their trajectories according to the laws of motion. Then the volume of the region that the points occupy at a time t later, $V_\Gamma(t)$, is equal to the initial volume $V_\Gamma(0)$.

A special class of ensemble distribution functions is of great importance for equilibrium statistical mechanics. These are the *stationary* ensembles such that $\rho(\Gamma,t)$ does not depend on time. It follows from Liouville's equation, Eq. (4), that $\rho(\Gamma,t)$ will remain constant in time, if the Poisson bracket of ρ with the Hamiltonian vanishes. The most important set of such functions are those where the distribution function is a function only of the Hamiltonian H, and of other constants of the motion.

2. THE MICROCANONICAL ENSEMBLE

The formulation of equilibrium statistical mechanics established by Maxwell and Boltzmann is based on the following stationary ensemble. We consider a mechanically isolated system whose energy lies in the range

E to $E + \delta E$. Then we choose a stationary phase-space distribution function $\rho(\Gamma)$ to have the form

$$\rho(\Gamma) = \begin{cases} \text{const. for } E \leq H(\Gamma) \leq E + \delta E, \\ 0 \quad \text{otherwise,} \end{cases} \tag{7}$$

where $H(\Gamma)$ is the value of the Hamiltonian at the phase point Γ and is the total energy of the system. This ensemble is called the *microcanonical ensemble.* In many formulations of the microcanonical ensemble, Eq. (7) is replaced by the equivalent relation, for large systems,

$$\rho(\Gamma) = (1/\Omega)\delta(H(\Gamma) - E), \tag{8}$$

where $\delta(H(\Gamma) - E)$ is the Dirac delta function that defines a $(2Nf - 1)$-dimensional surface, the constant-energy surface, in Γ space. The principal justification for the use of the ensemble to describe the macroscopic properties of a (virtually) isolated system is that the results so obtained are entirely in accord with experimental results, apart, of course, from quantum effects, which require a more quantum mechanical formulation of the microcanonical ensemble. In the next section we discuss a more fundamental justification of the microcanonical ensemble based on *ergodic theory.*

The constant Ω^{-1} appearing in Eq. (8), or equivalently, the constant appearing in Eq. (7), is determined by the normalization conditions, Eq. (6), imposed on the phase-space distribution function. By considering Eq. (8), for example, we determine Ω as

$$\Omega = \int_{A(E,V\{N_\alpha\},z_1\cdots z_j)} \frac{dA(\Gamma)}{|\mathrm{grad}\, H|_E}, \tag{9}$$

where $dA(\Gamma)$ denotes a differential "area" element about the point Γ on the surface in Γ space determined by the condition that $H(\Gamma) = E$. This surface has one less dimension than the total number of dimensions of Γ space. In Eq. (8), $|\mathrm{grad}\, H|_E$ is the magnitude of the $2Nf$-dimensional gradient of the Hamiltonian, evaluated on the surface $H(\Gamma) = E$, that is

$$|\mathrm{grad}\, H|_E = \left\{ \sum_{i=1}^{Nf} \left[\left(\frac{\partial H}{\partial q_i}\right)^2 + \left(\frac{\partial H}{\partial p_i}\right)^2 \right] \right\}^{1/2}_{H(\Gamma)=E}. \tag{10}$$

The value of this gradient may depend on the particular point on the energy surface in the integration in Eq. (9). The region of integration in Eq. (9) is over the constant-energy surface defined by the condition $H(\Gamma) = E$ but subject to the physical constraints imposed on the system. That is, the particles of the system may each be confined to a spatial volume V, and the number of particles of each species α, N_α, must be specified. If the different species of particles have different numbers of mechanical degrees of freedom, then the total dimension of Γ space is given by the sum $2\Sigma N_\alpha f_\alpha$, where f_α is the number of degrees of freedom of each particle of species α. In addition it may be necessary to specify certain intensive quantities, denoted by $z_1, \cdots, z_j$, such as the values of electric, magnetic, or gravitational fields, to determine completely the thermodynamic state of the system. Thus $A(E,V,\{N_\alpha\},z_1,\cdots,z_j)$ denotes that part of the constant-energy surface that is consistent with these constraints.

The microcanonical distribution is used to provide a statistical description of the thermodynamics of isolated systems where the total energy E, the volume of the (physical) system V, the number of particles of each species α, N_α, and the intensive quantities $z_1, \cdots, z_j$ are all specified.

Boltzmann's postulate provides the essential connection between the thermodynamic properties of an isolated system in equilibrium and quantities evaluated in microcanonical ensemble. That is, the *statistical thermodynamics* of an isolated system is determined by the relation

$$S(E,V,\{N_\alpha\},z_1,\cdots z_j) = k_\mathrm{B} \ln\Omega(E,V,\{N_\alpha\},z_1\cdots z_j), \tag{11}$$

which was anticipated by Eq. (1). The quantity Ω is called the *microcanonical partition function,* and it represents the total "measure" in Γ space occupied by systems with energy E. Because of the presence of the gradient in the integral of Eq. (9), one cannot identify Ω with the $(2Nf - 1)$-dimensional area of the constant-energy surface, but rather with a measure on the surface. In Eq. (11), S is to be identified with the thermodynamic entropy as a function of the indicated variables. The statistical expressions for other thermodynamic quantities can be obtained by using ordinary thermodynamic relations. For example, the thermodynamic temperature, T, is given by

$$T^{-1} = \left(\frac{\partial S}{\partial E}\right)_{V,\{N_\alpha\},z_1,\cdots,z_j}. \tag{12}$$

The equilibrium average of any dynamical quantity $f(\Gamma)$ is given in the microcanonical ensemble by

$$\langle f(\Gamma)\rangle = \frac{1}{\Omega(E,V,\cdots)} \int_{A(E,V,\cdots)} \frac{dA(\Gamma)f(\Gamma)}{|\text{grad } H|_{H=E}}. \tag{13}$$

As an example we consider a nearly ideal gas of N identical particles of mass m contained in a volume V and compute the number of such particles with momentum p at a point r, in equilibrium. For this quantity, the microscopic density of particles at point r with momentum p is given by

$$f(\Gamma) = \sum_{i=1}^{N} \delta(r_i - r)\delta(p_i - p). \tag{14}$$

The equilibrium average can be computed using Eq. (13) and is

$$n(r,p) = \langle f(\Gamma)\rangle = \frac{N}{V}\left(\frac{3N}{4\pi Em}\right)^{3/2}\left(1 - \frac{p^2}{2mE}\right)^{3N/2} \tag{15}$$

for large N. Then using Eq. (12) to identify the temperature, we obtain, as $N \to \infty$, $V \to \infty$, $N/V = n$, the number density, and $k_BT = 2E/3N$,

$$n(r,p) = n[2\pi m k_B T]^{-3/2} \exp\left(-\frac{p^2}{2mk_BT}\right), \tag{16}$$

which is the Maxwell–Boltzmann distribution function for an ideal gas.

The fundamental relation between the macroscopic and the statistical formulations of thermodynamics, $S = k_B \ln\Omega$, provides a microscopic interpretation for the fact, embodied in the second law of thermodynamics, that the entropy of an isolated system increases whenever any of the constraints on the system are relaxed without changing the energy and the system spontaneously proceeds to a new equilibrium state. Very simply, the relaxing of the constraint has the effect of increasing the total measure Ω of the constant-energy surface because more of the surface becomes available as a possible physical state for the members of the ensemble.

3. THE ERGODIC HYPOTHESIS

The use of the microcanonical ensemble to compute the thermodynamic equilibrium properties of macroscopic systems, while successful, requires justification. Two problems need to be addressed:

1. Can one prove that a large isolated system will, in the course of time, approach a state of thermodynamic equilibrium?
2. Can one prove that this state of thermodynamic equilibrium is properly described by the microcanonical ensemble?

One approach to the resolution these questions was formulated by Boltzmann and forms the subject of ergodic theory. One identifies the measurement of the dynamical property of a single system, $f(\Gamma)$, with the time average of the quantity over some time interval, denoted $\bar{f}_T$, as

$$\bar{f}_T = \frac{1}{T}\int_{t_0}^{t_0+T} dt f(\Gamma(t)), \tag{17}$$

where $\Gamma(t)$ denotes the point on the trajectory of the system in Γ space at time t, t_0 is some initial time, and T is the duration of the experiment. If the system does approach an equilibrium state, then as $T \to \infty$, $\bar{f}_T$ should approach the value of $f(\Gamma)$ as measured in the thermodynamic equilibrium state where a steady value of the measurement is expected. The ergodic hypothesis states that as the trajectory of the phase point develops in time, it spends a certain amount of time in every region of the constant-energy surface, consistent with the macroscopic constraints discussed earlier. The amount of time that the system spends in a particular region is, according to the ergodic hypothesis, proportional to the measure of that region, if one follows the trajectory for a long enough time. This is the heart of the hypothesis, and it is often referred to as the hypothesis of equal *a priori* probabilities. By dividing up the constant-energy surface into a large number of regions, each with a small but nonzero measure, one can conclude that

$$\lim_{T\to\infty}\frac{1}{T}\int_{t_0}^{t_0+T} dt f(\Gamma(t)) = \frac{1}{\Omega(E,V,\cdots)}\int\frac{dA(\Gamma)f(\Gamma)}{|\mathrm{grad}\,H|_{H=E}} \tag{18}$$

if the ergodic hypothesis is correct. Therefore, if one can prove the ergodic hypothesis, one can conclude that whatever the value of $\bar{f}_T(\Gamma)$ may be for some finite time T, this value will always approach the value as computed by statistical mechanics using a microcanonical ensemble distribution function, Eq. (8). This approach has nothing to say about the details of the approach of $\bar{f}_T(\Gamma)$ to its equilibrium value as $T \to \infty$, and that forms the subject of nonequilibrium statistical mechanics to be discussed in Sec. 6.

To justify the ergodic hypothesis, one must carefully examine the microscopic dynamical behavior of the system, and this study has provided a number of physical and mathematical researches, generally called ergodic theory (Halmos, 1956). The first important result in this direction was provided by Birkhoff's individual ergodic theorem, which asserts that for a Hamiltonian dynamical system, the limit as $T \to \infty$ of the time average $\bar{f}_T(\Gamma)$ exists, for almost all initial points for the trajectories and for integrable functions $f(\Gamma)$. However, since the theorem does not rule out the possibility that the limiting function may vary from one trajectory to another, and thus not be equal to the microcanonical average of f, it is not enough to ensure the correctness of the ergodic hypothesis. Therefore one needs to study the circumstances under which the limiting time average is equal to a constant, independent of trajectories, and thus, with some simple arguments, equal to the microcanonical ensemble average. Such systems are called ergodic. Recently Ya. G. Sinai and co-workers have shown that some simplified hard-sphere systems are ergodic (Sinai, 1970). In fact, they were able to show that despite the purely deterministic nature of these systems, they exhibit behavior that is similar in many respects to a purely random system typified by a coin-toss experiment. This feature of the systems studied by Sinai and co-workers is called the Bernoulli property, and it, in turn, implies the ergodic behavior of the system. Because of the considerable mathematical difficulties involved, very little is known about the ergodic behavior of more realistic systems.

Other approaches to justifying the use of the microcanonical ensemble distribution rely on the fact that no physical system can be truly isolated. Rather, any system is always subjected to random perturbation from its surroundings. These random perturbations may affect the phase-space trajectories in such a way that the time average of a dynamical quantity approaches the ensemble average, whether or not the dynamical system, considered in strict isolation, has the ergodic property.

In any event, the microcanonical ensemble has proven, by its success in predicting thermodynamic properties of large systems, that it is a simple and reasonable starting point for a statistical theory of thermodynamic equilibrium.

4. THE CANONICAL AND GRAND CANONICAL ENSEMBLES

Very often in experimental situations one is interested in the thermodynamic properties of a system that is kept in contact with a large heat bath with which the system can exchange energy, or energy and particles. Examples are readily found in experiments that explore the equation of state and phase equilibrium phenomena of pure substances or mixtures (see PHASE EQUILIBRIA), or in light-scattering experiments that explore the properties of a small subregion, the scattering volume, of a large system. In this latter case, the subregion is not separated from the rest of the system by any sort of boundaries so that an exchange of particles and energy with the rest of the system is taking place at all times.

The microcanonical ensemble, discussed above, can be used to obtain a statistical mechanical description of a system in contact with a heat bath with which it can exchange energy or energy and particles. The procedure is to regard the system plus the heat bath as a large isolated system and to use the microcanonical distribution for the whole system. In carrying out this procedure it is important to distinguish between the case where the system and bath exchange only energy, and the cases where both energy and particles are exchanged. In the former case we construct an ensemble distribution for the system (not the system plus bath) called the *canonical ensemble,* and in the latter, where the system exchanges both energy and particles with the bath, we construct an ensemble distribution for the system called the grand canonical ensemble, after Gibbs.

If the system exchanges only energy with the heat bath, the microcanonical distribution function for the entire system is $\rho_{\text{tot}}(\Gamma_1,\Gamma_2)$, given in a $(2N_1 f_1 + 2N_2 f_2)$-dimensional phase space compared by combining the Γ space for the system, Γ_1, and that for heat bath, Γ_2, by

$$\begin{aligned}\rho_{\text{tot}}(\Gamma_1,\Gamma_2) = \Omega_{\text{tot}}^{-1}\delta(H_1(\Gamma_1) + H_2(\Gamma_2)\\ + H_{12}(\Gamma_1,\Gamma_2) - E).\end{aligned} \tag{19}$$

Here the subscripts 1 and 2 denote the system of interest and the heat bath, respectively, and we ignore, for the moment, that the system of interest may be composed of particles of different species, α. The $H_i(\Gamma_i)$, $i = 1,2$, are the Hamiltonian functions of system i, and Γ_i is the phase-space point for system i, and $H_{12}(\Gamma_1,\Gamma_2)$ is a Hamiltonian function that describes the interaction between the system and the heat bath. Next we assume that both the system and heat bath are sufficiently large that the interaction energy H_{12} can be neglected in Eq. (19). As we are not at all interested in the heat bath, we integrate $\rho_{\text{tot}}(\Gamma_1,\Gamma_2)$ over all the variables of the heat bath to obtain a normalized phase-space distribution function for the system of interest alone. This is easily seen to be

$$\begin{aligned}\rho(\Gamma_1) = \Omega_2(E - H_1(\Gamma_1))\\ \times\left[\int d\Gamma_1\Omega_2(E - H_1(\Gamma_1))\right]^{-1},\end{aligned} \tag{20}$$

where Ω_2 is the microcanonical partition function for the heat bath. We consider the heat bath to be very large, large enough that we can expand the entropy of the heat bath about its value at energy E, so as to obtain

$$\begin{aligned}\Omega_2(E - H_1(\Gamma_1)) &= \exp[S_2(E - H_1(\Gamma_1))/k_{\text{B}}]\\ &\cong \exp\{[S_2(E) - H_1/T + \cdots]/k_{\text{B}}\}\\ &\cong e^{S_2/k_{\text{B}}}e^{-\beta H_1},\end{aligned} \tag{21a}$$

where $\beta = (k_{\text{B}}T)^{-1}$, T is the temperature of the heat bath. For the distribution function, we obtain, using Eq. (20),

$$\rho(\Gamma_1) = e^{-\beta H_1(\Gamma_1)}\left[\int d\Gamma_1 e^{-\beta H(\Gamma_1)}\right]^{-1}. \tag{21b}$$

This resulting distribution function is the normalized canonical ensemble distribution function for a system in contact with a heat bath at temperature T. The statistical thermodynamics for this system is most easily obtained by considering the variation of the average energy of the system with respect to β, the volume V, and the intensive variables $z_1, \ldots, z_j$, where the average energy is $U = \langle E\rangle_c$, with

$$\langle E\rangle_c = \int d\Gamma_1 H(\Gamma_1)e^{-\beta H(\Gamma_1)}\left[\int d\Gamma_1 e^{-\beta H(\Gamma_1)}\right]^{-1}. \tag{22}$$

By examining the variation of U with respect to the variables β, V, z_1, $\cdots$, z_j and comparing it with the thermodynamic formula

$$dU = TdS - PdV + \sum_j X_j dz_j, \tag{23}$$

one determines the proper relation between statistical and thermodynamic properties. In Eq. (23), P is the pressure of the system, and $X_i dz_i$ is the work done on the system when the parameter z_i is varied. The result of this analysis is that the temperature of the heat bath and the system are the same, as expected in thermodynamic equilibrium, and that the connection between the thermodynamic properties of the system and statistical mechanical ensemble averages is most easily expressed by the relation

$$e^{-\beta F} = Z. \tag{24}$$

Here F is the Helmholtz free energy of the system, and Z, the canonical partition function of the system, is, almost, but not quite, the normalization factor in Eq. (22). Instead it is

$$Z = \frac{1}{\prod_\alpha N_\alpha! h^{NF}}\int d\Gamma_1 e^{-\beta H(\Gamma_1)} \tag{25}$$

The additional prefactors $[\Pi_\alpha N_\alpha! h^{NF}]^{-1}$, where h is Planck's constant, are not required by the thermodynamic identity, Eq. (23), since the identity determines $\ln Z$ up to an arbitrary constant. This arbitrary constant is determined by the condition that the classical partition function agree with the limit of the quantum mechanical version of the canonical partition function in the limit $h \to 0$. This accounts for the prefactor in Eq. (24). The $N_\alpha!$ appear because all the particles of species are identical and permutations of their particle indices do not lead to a new quantum state, and the h^{NF} counts the number of quantum states, such that the number of states in $dq_i dp_i$ is $dq_i dp_i/h$.

An earlier discussion of the canonical partition function by Gibbs also produced the $(\Pi_\alpha N_\alpha!)^{-1}$ in Eq. (25). Gibbs argued that if this prefactor were not present in Z in Eq. (24), the entropy so obtained would not be extensive and would lead to an entropy of mixing if two containers of an identical ideal gas were combined into one. Such an entropy of mixing is inconsistent with thermodynamics, and its paradoxical occurrence in the earlier formulation of the canonical ensemble was referred to as the *Gibbs paradox.* The paradox does not appear if the partition function is formulated properly. Of course, until the advent of quantum theory and the notion of discrete quantum states, it was not possible to obtain an absolute value of thermodynamic functions from statistical thermodynamics, and Planck's constant did not appear in the partition function.

In the case that chemical reactions are taking place in the system, the set of particle numbers $\{N_\alpha\}$ is defined to be the numbers of the constituent particles that are unaffected by the reaction (e.g., protons, electrons, or atoms of each element), and the numbers of the reacting chemical species are regarded as thermodynamic parameters to be fixed by the equilibrium conditions.

If the system has a fixed volume V but can exchange energy and particles with a heat bath, and the heat bath is sufficiently large, we can again use the microcanonical distribution function for the entire system to show that the probability of finding the system when it has N_α particles of each species α, and at phase point $\Gamma(\{N_\alpha\})$, is

$$\rho(\Gamma(\{N_\alpha\})) = \Xi^{-1} \times \exp\left(\sum_\alpha \beta N_\alpha \mu_\alpha\right) \exp[-\beta H(\Gamma(\{N_\alpha\}))]; \tag{26}$$

here $\Gamma(\{n_\alpha\})$ is a point in the $(\Sigma 2N_\alpha f_\alpha)$-dimensional Γ space for a system with N_α particles of each species α, where f_α is the number of mechanical degrees of freedom of species α. This distribution is the grand canonical distribution, and Ξ, the normalization factor in Eq. (26), is the grand canonical partition function given by

$$\Xi = \sum_{N_1=0}^{\infty} \cdots \sum_{N_k=0}^{\infty} \exp\left(\beta \sum_\alpha N_\alpha \mu_\alpha\right) \times \int d\Gamma(\{N_\alpha\}) e^{-\beta H(\Gamma(\{N_\alpha\}))}. \tag{27}$$

In Eqs. (26) and (27) we suppose that there are k distinct types of particles, $\alpha = 1, \cdots, k$; the parameters β and $\{\mu_\alpha\}$ are characteristics of the heat bath.

The grand canonical distribution function provides a description of open systems, and the relation between the statistical and thermodynamic description of such systems is provided by

$$P(T,V,\{\mu_\alpha\},z_1 \cdots z_j)V = k_B T \ln\Xi, \tag{28}$$

where P is the pressure, $T = (k_B\beta)^{-1}$ is the temperature, and the μ_α are the chemical potentials of the species α.

It is generally believed that for large systems, the thermodynamic functions should not depend sensitively on whether it is exchanging particles and/or energy with a heat bath. Consequently the computed thermodynamic properties of a large system should be the same regardless of the ensemble used in their computation. Therefore, one would expect to be able to prove the *equivalence of ensembles* for very large systems of physical interest. One does not expect, however, that the equivalence of ensembles will hold for a computation of fluctuations in the thermodynamic quantities. For example, the grand canonical ensemble allows for fluctuations in the species numbers N_α, but the microcanonical and canonical ensembles do not.

We are interested in computing values of thermodynamic functions for systems that are large enough that the effects of boundaries can be neglected. Then we consider the *thermodynamic limit,* where for the microcanonical ensemble $E \to \infty$ and $V \to \infty$ in such a way that the energy per unit volume E/V remains fixed at its value for the finite system;

for the canonical ensemble $V \rightarrow \infty$, and the particle numbers $N_\alpha \rightarrow \infty$ in such a way that the particle densities N_α/V remain fixed; and for the grand canonical ensemble, $V \rightarrow \infty$, and the mean particle numbers $\langle N_\alpha \rangle \rightarrow \infty$ such that the mean densities $\langle N_\alpha \rangle/V$ are fixed. In the thermodynamic limit one computes specific values of the extensive thermodynamic functions such as the entropy or free energies per particle, or per unit mass. There is considerable literature detailing the microscopic properties that a system must have for its specific thermodynamic properties to exist in the thermodynamic limit, and establishing the equivalence of the various ensembles for calculating the specific quantities. The conditions on microscopic properties, such as the interaction potentials between particles, are sufficiently general that the existence of the specific value of the extensive quantities in the thermodynamic limit is ensured for a wide variety of systems of physical interest (Ruelle, 1969).

5. APPLICATIONS OF EQUILIBRIUM STATISTICAL MECHANICS

The methods of equilibrium statistical mechanics have been applied to calculate thermodynamic and other properties of a wide variety of systems. The simplest applications are the studies of properties of system of noninteracting particles, such as very dilute gases of monatomic or polyatomic molecules, or harmonic crystals, where the noninteracting "particles" are the normal modes of the crystal. The computation of the properties of systems of interacting particles is more difficult, but notable progress has been made in the computation of the properties of dense gases, and in developing techniques to treat phase transitions, particularly the liquid–gas and paramagnetic–ferromagnetic phase transitions (see PHASE TRANSITIONS IN SOLIDS). The liquid–solid phase transition, i.e., "freezing," though, is still poorly understood. Various solid–solid transitions, and phase changes in the arrangement of atoms absorbed on solid surfaces, have been under active study with considerable success. Among the many other problems that have been successfully treated by the methods of classical equilibrium statistical mechanics are the conformational properties of polymers, the helix-coil transition in DNA-like molecules, the properties of various types of magnetic systems, and the properties of plasmas and of stellar matter. The theoretical understanding of the equilibrium properties of such systems has been aided considerably by the use of computers. In addition to the fact that many difficult calculations are a routine matter for computers, the use of computer-simulated molecular dynamics has led to a deeper understanding of the properties of fluid and solids (see SIMULATION BY MOLECULAR DYNAMICS). In this section we discuss briefly some of the abovementioned applications.

5.1 The Properties of Ideal Systems

Ideal systems are characterized by the property that the Hamiltonian function for them can be written as a sum of energies that are a positive definite quadratic form in the canonical coordinates and momenta. Typical examples are a dilute gas of atoms where the interactions between the atoms can be neglected, and the total energy is simply the sum of the kinetic energies of each particle. Other examples are a dilute gas of polyatomic molecules whose energies can each be written as the sum of a translational kinetic energy, a rotational kinetic energy, quadratic in the angular momenta about each of the principal axes of rotation of the molecule, and a vibrational energy comprising the kinetic and potential energies of each of the normal modes of the molecule. Any couplings between translational, vibrational, and rotational motions are neglected. Another example of an ideal system is a purely harmonic solid with an energy that can be written as a sum of the kinetic and potential energies of each of the normal modes.

The properties of such systems are easy to evaluate because the various partition functions are quite simple. For example, the canonical partition function of N noninteracting molecules or of an ideal solid with N normal modes can be written as

$$Z_{\text{ideal}} = \frac{1}{N! h^{NF}} \left[\int d^s p \int d^s q \; e^{-\beta H_1(\mathbf{q},\mathbf{p})} \right]^N, \qquad (29)$$

where $\mathbf{p},\mathbf{q}$ denote all the canonical momenta and coordinates for one particle (or normal mode), and $H_1(\mathbf{q},\mathbf{p})$ is the Hamiltonian function for one particle (or normal mode), quadratic in the appropriate coordinates and momenta. Since the integrals appearing in Eq. (29) are at most the integrals of Gaussian forms, Z_{ideal} is easily determined. The internal energy of the system, U, is obtained from the general relation

$$U = \langle E\rangle = -\frac{\partial \ln Z}{\partial\beta} = N\sum_{i=1}^{s}\frac{k_{\mathrm{B}}T}{2} = Nsk_{\mathrm{B}}T/2. \tag{30}$$

In Eq. (30), s denotes the number of quadratic terms in the Hamiltonian function for each particle. For example, $s = 3$ for an atom, $s = 7$ for a diatomic molecule (three translational degrees of freedom, two rotational degrees of freedom, and two to account for the kinetic and potential energies of vibration of the molecule). For a harmonic crystal N is equal to the total number of normal modes of the crystal, and $s = 2$ for each mode. Equation (30) is often referred to as the *law of equipartition of energy.* It predicts that the specific heat at constant volume C_v is a constant for such systems, given by

$$C_v = Nsk_{\mathrm{B}}/2. \tag{31}$$

The failure of Eq. (31) to hold for low temperatures, and the variation of C_v with temperatures even for very dilute noninteracting systems, was one of the indications that classical physics does not adequately describe molecules and crystals and that quantum effects are important.

5.2 Fluctuations

The canonical ensemble describes a system in contact with a thermal reservoir. As a result the energy of the system does not maintain a fixed value, but rather it undergoes fluctuations. The mean square fluctuations of the energy of a system in contact with a heat bath can easily be calculated by means of the canonical ensemble distribution function, as follows. We are interested in the quantity $\langle(E - \langle E\rangle)^2\rangle$, which expresses the mean square fluctuations of energy. This quantity is given, in the canonical ensemble (for a system of identical particles), by

$$\begin{aligned}\langle(E - \langle E\rangle)^2\rangle &= \frac{1}{ZN!h^{NF}}\int d\Gamma[H(\Gamma) - \langle E\rangle]^2 e^{-BH(\Gamma)}\\ &= \frac{1}{ZN!h^{NF}}\int d\Gamma H^2(\Gamma)e^{-BH(\Gamma)} - \langle E\rangle^2\\ &= \frac{1}{ZN!h^{NF}}\frac{\partial^2}{\partial\beta^2}\int d\Gamma e^{-BH(\Gamma)} - \langle E\rangle^2\\ &= k_{\mathrm{B}}T^2\left(\frac{\partial\langle E\rangle}{\partial T}\right)_{N,V} = k_{\mathrm{B}}T^2C_v.\end{aligned} \tag{32}$$

Thus the mean square energy fluctuations are proportional to the specific heat at constant volume. This result also obtains for quantum systems as well. One of the many interesting consequences of the result is that the mean square fluctuation in the energy per particle vanishes as N^{-1} in the thermodynamic limit.

Similarly, a system in contact with a reservoir with which it can exchange both energy and particles undergoes fluctuations of both energy and density. The energy fluctuations are still given by the last term on the right-hand side of Eq. (32), with the various thermodynamic functions evaluated in the grand canonical ensemble. The density fluctuations can be determined from the mean square fluctuations in particle number $\langle(N - \langle N\rangle)^2\rangle$, where the mean number of particles in the system $\langle N\rangle$ is given, in the grand canonical ensemble (for a system of identical particles), by

$$\langle N\rangle = k_{\mathrm{B}}T\left(\frac{\partial n\Xi}{\partial\mu}\right)_{V,T,z_1\cdots z_i}. \tag{33}$$

Then the fluctuation in particle number is given by

$$\langle(N - \langle N\rangle)^2\rangle = k_{\mathrm{B}}T\left(\frac{\partial\langle N\rangle}{\partial\mu}\right). \tag{34a}$$

The right-hand side of Eq. (34a) can be expressed in terms of more customary thermodynamic functions, as

$$\langle(N - \langle N\rangle)^2\rangle = \langle N\rangle^2k_{\mathrm{B}}T\kappa_T/V, \tag{34b}$$

where κ_T is the isothermal compressibility of the system,

$$\kappa_T = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T \tag{35}$$

with P the pressure of the system. To relate this result to density fluctuations, one need only write the density $n = N/V$ in Eq. (34) to obtain

$$\langle (n - \langle n\rangle)^2\rangle = \langle n\rangle^2 k_B T\kappa_T/V, \tag{36}$$

a result that shows that the density fluctuations vanish as $V \to \infty$.

Density fluctuations in an open system are responsible for the scattering of light by the system. This accounts for, among other things, the blue color of the sky, and the phenomenon of critical opalescence in the vicinity of the critical point of a phase transition, where the isothermal compressibility becomes large (see CRITICAL PHENOMENA).

5.3 Distribution Functions

Using the appropriate phase-space distribution functions for the microcanonical, canonical, and grand canonical ensembles, one can compute the distribution functions that describe the properties of an equilibrium system in more detail than do the thermodynamic functions. Furthermore, the thermodynamic functions themselves can be expressed in terms of integrals of products of the equilibrium distribution functions with functions such as the potential energy of interaction of two particles in the system, or other appropriate microscopic functions.

For most applications of statistical mechanics to fluid systems, the most important distribution functions are the one-particle and two-particle distribution functions $n_1(\mathbf{r},\mathbf{p})$ and $n_2(\mathbf{r},\mathbf{p},\mathbf{r}'',\mathbf{p}'')$, respectively, defined by

$$n_1(\mathbf{r},\mathbf{p}) = \left\langle \sum_{i=1}^{N} \delta(\mathbf{r}_i - \mathbf{r})\delta(\mathbf{p}_i - \mathbf{p})\right\rangle \tag{37a}$$

and

$$n_2(\mathbf{r},\mathbf{p},\mathbf{r}',\mathbf{p}') = \left\langle \sum_{i=1}^{N} \sum_{j\neq i} \delta(\mathbf{r}_i - \mathbf{r})\delta(\mathbf{p}_i - \mathbf{p}) \times \delta(\mathbf{r}_j - \mathbf{r}')\delta(\mathbf{p}_j - \mathbf{p}')\right\rangle, \tag{37b}$$

where, as above, the angular brackets denote an average over one of the equilibrium ensemble distribution functions. Here we consider a fluid system of monatomic particles where the canonical variables are the positions of the particles $\mathbf{r}_1, \cdots, \mathbf{r}_N$, and the corresponding translational momenta $\mathbf{p}_1, \cdots, \mathbf{p}_N$. The quantity $n_1(\mathbf{r},\mathbf{p})$ is the probability of finding a particle at point $\mathbf{r}$ in the fluid with momentum $\mathbf{p}$, and $n_2(\mathbf{r},\mathbf{p},\mathbf{r}'',\mathbf{p}'')$ is the probability of finding one particle at point $\mathbf{r}$ in the fluid with momentum $\mathbf{p}$, and another at point $\mathbf{r}''$ with momentum $\mathbf{p}''$ (Uhlenbeck and Ford, 1962).

If we consider a fluid system with N-particle Hamiltonian

$$H(\Gamma) = \sum_{i=1}^{N} \frac{p_i^2}{2m} + \Phi(\mathbf{r}_1,\cdots,\mathbf{r}_N),$$

where $\Phi(\mathbf{r}_1,\cdots,\mathbf{r}_N)$ is the potential energy of interaction between the particles, usually assumed to be the sum of pairwise central potentials $\phi(|\mathbf{r}_i - \mathbf{r}_j|)$,

$$\Phi(\mathbf{r}_1,\cdots,\mathbf{r}_N) = \sum_{i<j} \phi(|\mathbf{r}_i - \mathbf{r}_j|) \tag{38}$$

in the absence of external fields, then the one-particle distribution function is the Maxwell–Boltzmann distribution function

$$n_1(\mathbf{r},\mathbf{p}) = n\phi_0(\mathbf{p}), \tag{39a}$$

where $n = N/V$ in the microcanonical and canonical ensembles, and $\langle N\rangle/V$ in the grand canonical ensemble, and $\phi_0(\mathbf{p})$ is the equilibrium momentum distribution function

$$\phi_0(\mathbf{p}) = \left(\frac{\beta}{2\pi m}\right)^{3/2} \exp\left(-\frac{\beta p^2}{2m}\right). \tag{39b}$$

Equation (39a) must be corrected if the system is placed in an external field, such as a gravitational field. The latter case leads to the barometric formula for the variation of density with height.

The two-particle distribution function can be expressed as

$$n_2(\mathbf{r},\mathbf{p},\mathbf{r}',\mathbf{p}') = \phi_0(\mathbf{p})\phi(\mathbf{p}')\langle N(N - 1) \times \delta(\mathbf{r}_1 - \mathbf{r})\delta(\mathbf{r}_2 - \mathbf{r}')\rangle. \tag{40a}$$

The quantity in the angular brackets on the right-hand side of Eq. (39) cannot be calculated further without specifying the interaction between the particles. It is customary to write this quantity—the probability of finding one particle at $\mathbf{r}$ and another at $\mathbf{r}''$—in terms of the so-called radial distribution function $g^{(2)}(\mathbf{r},\mathbf{r}'')$ as

$$\langle N(N-1)\delta(\mathbf{r}_1-\mathbf{r})\delta(\mathbf{r}_2-\mathbf{r}')\rangle = n^2 g^{(2)}(\mathbf{r},\mathbf{r}'), \tag{40b}$$

so that n_2 becomes

$$n_2(\mathbf{r},\mathbf{p},\mathbf{r}',\mathbf{r}') = n^2\phi_0(\mathbf{p})\phi_0(\mathbf{p}')g^{(2)}(\mathbf{r},\mathbf{r}'). \tag{41}$$

The radial distribution function differs slightly among the ensembles, since its integrals over $\mathbf{r}$ and $\mathbf{r}''$ lead to average quantities that are different in the grand canonical ensemble from their values in the other ensembles. For a large enough system $g^{(2)}(\mathbf{r},\mathbf{r}'')$ is only a function of $|\mathbf{r}-\mathbf{r}''|$, and it can be determined as a function of spatial separation, temperature, and density by theoretical analysis, by computer studies, and by light-scattering experiments. In Fig. 1 we illustrate this function for hard-sphere molecules. One can easily see the oscillations in this function, which show that there is a short-range ordering of particles in a fluid that is sufficiently dense, which is one indication of the more detailed ordering of particles that is found in a crystalline solid (Hansen and McDonald, 1986).

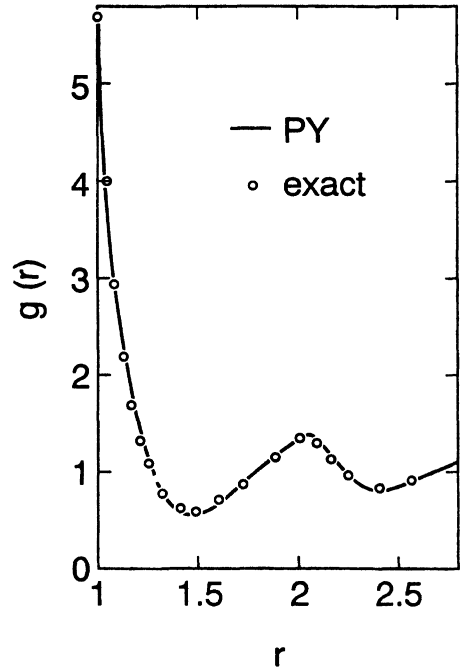

FIG. 1. A plot of the pair correlation function $g(r)$ for a hard-sphere fluid at $V/V_0 = 1.51$ where V_0 is the volume of close packing of the spheres. The points are the results of computer simulations, and the curve is the result of an approximate theory due to Percus and Yevick. Here r is measured in units of the hard-sphere diameter. [Figure from Hansen and McDonald (1986).]

In the case that the potential energy of interactions among the particles can be expressed as a sum of central, pair potentials, the thermodynamic functions of the fluid can be expressed in terms of the radial distribution function. For example, the pressure of the fluid P is

$$P = nk_BT - \frac{4\pi n^2}{6}\int_0^\infty dr r^2\phi'(r)g^{(2)}(r), \tag{42}$$

where $\phi'(r)$ is the derivative of $\phi(r)$ with respect to the separation r. Equation (42) can also be obtained from the virial theorem of classical mechanics and is often called the virial equation of state.

Although we have chosen to illustrate the use of distribution functions here with examples appropriate for fluid systems, very similar discussions can be given for magnetic or other systems.

5.4 Systems of Interacting Particles

The statistical theory of ideal systems described in Sec. 5.1 is, of course, only useful under very limited circumstances, because it is formulated under the assumption that the constituent particles, or modes, of the system do not interact. Methods have been developed in order to incorporate the particle interactions properly into the calculations of thermodynamic properties and equilibrium distribution functions. As long as one is not concerned with describing the properties of a system near a phase transition, which requires a more elaborate treatment, it is often possible to expand the properties of the system about some limiting ideal value in powers of an appropriate small parameter.

Typical of such expansions is the *virial expansion* of the thermodynamic properties of a real gas as a power series in the density of the gas. The virial expansion is motivated by the fact that if the gas is sufficiently dilute, the average distance between the molecules of the gas $\sim(V/N)^{1/3} = n^{-1/3}$ is much greater than the range of the intermolecular potential, σ, say. Thus the reduced density, $n\sigma^3 \ll 1$, is an appropriate small parameter for the representation of the properties of a dilute or moderately dense gas as a power series. J. Mayer and later workers have developed a very refined theory for developing such virial expansions for nonpolar, polar, and even ionic systems (Mayer and Mayer, 1981). In the last case, the virial expansion must be resummed term by term in order to obtain a well-behaved representation for the properties of an ionic system, such as an ionic solution or a plasma.

Typical of such expansions is the virial expansion of the pressure for a dilute or moderately dense gas, which, assuming that the particles interact with pairwise, central forces, has the form

$$\beta P = n + B_2(T)n^2 + B_3(T)n^3 + \cdots, \tag{43}$$

where the second virial coefficient, $B_2(T)$, is given by

$$B_2(T) = -\tfrac{1}{2}\int d\mathbf{r} f(|\mathbf{r}|), \tag{44a}$$

and the third virial coefficient $B_3(T)$ is given by

$$B_3(T) = -\tfrac{1}{3}\int d\mathbf{r}\int d\mathbf{r}'\, f(|\mathbf{r}'|)f(|\mathbf{r}-\mathbf{r}'|) \tag{44b}$$

and similar expressions for the higher virial coefficients B_4, B_5, $\cdots$. Here the function $f(|\mathbf{r}|)$ is the Mayer f function

$$f(|\mathbf{r}|) = e^{-\beta\phi(r)} - 1, \tag{45}$$

where $\phi(r)$ is the potential energy of interaction of two particles separated by a distance $|\mathbf{r}|$. If the potential has a range of σ, then, the Mayer f function vanishes for $|\mathbf{r}| \gg \sigma$, and consequently $B_2(T)$ is of order σ^3, $B_3(T)$ is of order $(\sigma^3)^2$, and the virial expansion becomes a power series in the reduced density $n\sigma^3$. The second virial coefficient, $B_2(T)$, represents the first correction to the ideal-gas equation of state, due to the influence of binary interactions taking place in the gas; B_3 is the next correction due to three-body interactions of three particles in the gas. The higher virial coefficients can be understood in a similar way.

While the virial expansion for the thermodynamic properties of gases is typical of the methods used to compute the equilibrium properties of nonideal systems, it by no means exhausts the wide variety of methods and expansions that have proved to be useful in the direction. High-temperature, low-temperature, weak-coupling, Debye-Hückel, . . . expansions abound in this area. It is safe to say that for many materials there now exists a suitable method for the computation of their thermodynamic properties. However, a suitable understanding and fundamental description of the properties of complicated liquids, solids, or other systems where our understanding of the microscopic details is incomplete is yet to be found.

5.5 Phase Transitions

One of the major areas of research in equilibrium statistical mechanics, for both classical and quantum systems, is to provide a description of the various phase transformations that thermodynamic systems exhibit. For the liquid–gas phase transition it is now possible to make quantitative predictions about the behavior of the thermodynamic and distribution functions near the critical point. The progress was made possible by the application of renormalization-group methods to the problem by Wilson and Kogut (1974), Wilson (1975), Fisher (1974), Kadanoff (1966), Kadanoff *et al.* (1967), Widom (1965), and others (see PHASE TRANSITIONS: RENORMALIZATION AND SCALING).

The theories for phase transitions indicate that near a critical point there are correlations between particles that extend over distances considerably larger than the average spacing between particles or the range of the interparticle potential. These correlations are propagated, so to speak, by the interparticle potential, but they are not very sensitive to the details of the potential. As a result many qualitatively different systems have similar phase-transition properties.

For example, it has been shown that the liquid–gas phase transition has many properties in common with other phase-transition phenomena such as the paramagnetic–ferromagnetic phase transition in metals and the one phase–two phase transition in binary fluid mixtures. It has been possible, in fact, to organize a large number of apparently different types of phase transitions into *universality classes,* such that the critical properties of all phase transitions in any one class are describable by the same set of fundamental parameters, the *critical exponents* (see UNIVERSALITY). These critical exponents describe the way that the various thermodynamic functions either diverge or approach zero as the critical point is approached from different directions in the space of the appropriate thermodynamic variables.

6. NONEQUILIBRIUM STATISTICAL MECHANICS

Nonequilibrium statistical mechanics deals with the time development of an ensemble of similar systems described at some initial time by a nonequilibrium distribution function, and attempts to describe the average behavior of this ensemble. As stated in the Introduction, the goal is to explain why the nonequilibrium behavior of large systems often follows simple laws, such as the equations of fluid dynamics or similar equations for other systems, generally described by the equations of irreversible, or nonequilibrium, thermodynamics.

The thermodynamics of nonequilibrium processes is essentially based on two principal ideas:

1. One can describe the behavior of large systems that are not in equilibrium in terms of local densities of "conserved" quantities, such as the local mass density $\rho(\mathbf{r},t)$, the local momentum density $g(\mathbf{r},t)$, and the local energy density $\epsilon(\mathbf{r},t)$ in a fluid. These local densities satisfy "conservation" laws, which express the fact that mass, momentum, and energy are not created within the system, but instead flow from one region of the system to another.
2. The fluxes of mass, momentum, and energy from one region of the system to an adjacent region are functions, often but not necessarily linear functions, of the gradients of the local densities.

While the first idea is strictly a consequence of the mechanical, macroscopic properties of the system, the second idea embodies some description of a microscopic system that is not often obvious from a knowledge of its microscopic properties, and indeed reflects a kind of randomness in the behavior of large systems. The flux–gradient relations include Fick's law of diffusion, Newton's law of viscosity, Fourier's law of heat conduction, and Ohm's law of electrical conductivity, as well as similar laws appropriate to multicomponent systems, chemically reacting systems, plasmas, and so on. The reason that one suspects that a kind of random behavior of the particles of the system is responsible for these laws is that flux–gradient relations of this form are typical of random-walk models for diffusion, and are not typical of simple mechanical systems where particles do not collide with each other, or with some obstacles placed in the system. Thus one expects that collisions of some sort—between particles, nonlinear interactions between normal modes of a crystal, etc.—are at the heart of the behavior described by nonequilibrium thermodynamics, and thus need to be taken into account in a statistical mechanical theory of nonequilibrium processes.

Therefore the problems to which nonequilibrium statistical mechanics addresses itself are to explain and to describe the approach of macroscopic systems to a state of total equilibrium, to describe the behavior of time-dependent fluctuations about equilibrium, and, perhaps the most fundamental problem, to explain why large systems with reversible microscopic dynamics obey irreversible macroscopic laws.

The foundation for the nonequilibrium statistical mechanics of classical systems is Liouville's equation for the time-dependent Γ-space distribution function $\rho(\Gamma,t)$, Eq. (4). In the main, two approaches have been developed for a discussion of irreversible processes. One approach applies to dilute or dense gases, and is part of the subject matter of the kinetic theory of gases. In this approach one tries to develop an equation for the single-particle distribution function $f(\mathbf{r},\mathbf{v},t)$ defined so that $f(\mathbf{r},\mathbf{v},t)d^3r d^3v$ is the number of particles in the system in the spatial region d^3r about $\mathbf{r}$ and in the velocity region d^3v about $\mathbf{v}$, at time t. The major contribution to the theory of this function was made by Boltzmann, who used plausible arguments about the number of collisions taking place in the gas in a small time interval to derive an equation for $f(\mathbf{r},\mathbf{v},t)$ in the case that the gas is dilute. This equation, the *Boltzmann transport equation,* is

$$\frac{\partial f(\mathbf{r},\mathbf{v},t)}{\partial t} + \mathbf{v}\cdot\nabla_r f(\mathbf{r},\mathbf{v},t) = \int d^3v_1 \int_0^{\sigma} b\,db \int_0^{2\pi} d\epsilon\, |\mathbf{v} - \mathbf{v}_1| \times [f(\mathbf{r},\mathbf{v}_1',t)f(\mathbf{r},\mathbf{v}',t) - f(\mathbf{r},\mathbf{v},t)f(\mathbf{r},\mathbf{v}_1,t)]. \tag{46}$$

The left-hand side of Eq. (46) describes the rate of change of f due to free-particle motions, that is, motion where every particle at $(\mathbf{r},\mathbf{v})$ at some time t moves to the point $(\mathbf{r} + \mathbf{v}dt,\ \mathbf{v})$ after a time interval dt. If there were no collisions in the gas, the right-hand side of Eq. (46) would be zero. Instead, some of the particles that started at $(\mathbf{r},\mathbf{v})$ collided with other particles in the gas before they could arrive at $(\mathbf{r} + \mathbf{v}dt,\ \mathbf{v})$. These are counted in the terms on the right-hand side with the negative sign. Still other particles arrived at $(\mathbf{r} + \mathbf{v}dt,\ \mathbf{v})$ without starting at $(\mathbf{r},\mathbf{v})$ but got there as a result of binary collisions with other particles. These are counted in the first term in the integrand on the right-hand side of Eq. (46). b and ϵ represent impact parameters and an azimuthal angle for a binary collision, and $\mathbf{v}'$ and $\mathbf{v}_1'$ are velocities of two particles that collide with collision parameters (b,ϵ) in such a way that one of the particles has velocity $\mathbf{v}$ after the collision. These collisions, which produce particles with velocity $\mathbf{v}$, are called restituting collisions, while those that reduce the number of particles with velocity $\mathbf{v}$ are called direct collisions. This situation is illustrated in Fig. 2.

The Boltzmann transport equation predicts that for an isolated system $f(\mathbf{r},\mathbf{v},t)$ decays in the course of time to its equilibrium value and allows the construction of a nonequilibrium analog for the thermodynamic entropy, the negative of the so-called H function, which increases monotonically in time as the system approaches equilibrium. Moreover, one can derive the hydrodynamic equations for a dilute gas together with expressions for the associated transport coefficients, such as the coefficients of thermal conductivity and viscosity, in terms of the intermolecular potential. These expressions describe the experimental results for dilute gases so well that they are used now to test various intermolecular potential models. In spite of the considerable success of the Boltzmann equation, a complete theory of nonequilibrium processes in gases requires

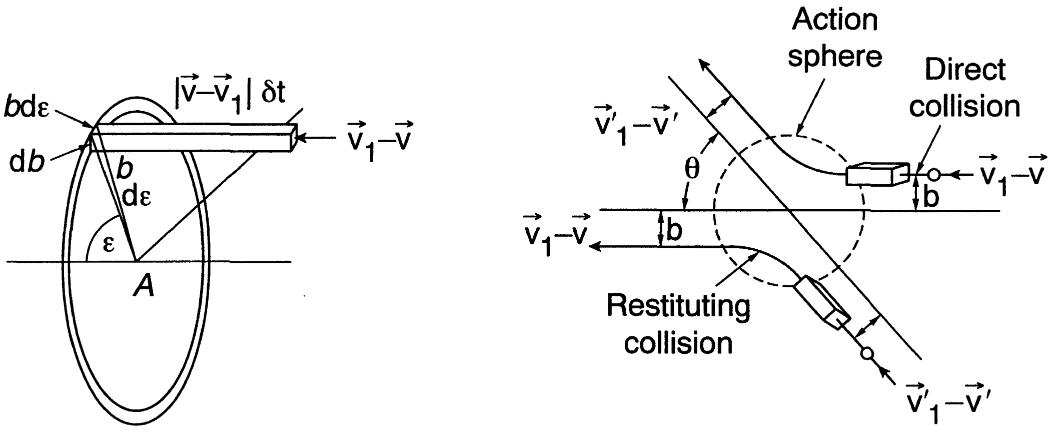

(a) (b)

FIG. 2. (a) The impact parameter b and azimuthal angle ϵ that describe a binary collision. The origin of the coordinate system is placed at the center of one of the colliding particles. (b) The direct and restituting collisions included in the Boltzmann equation. The origin of the coordinate system is placed on the center of the particle that has velocity v, either before or after the collision. The action sphere denotes the range of the interaction forces between the two particles.

that one be able to derive the Boltzmann equation from the Liouville equation directly, i.e., to replace the plausible argument about the number of collisions by a plausible construction of an initial ensemble and to extend the Boltzmann equation to higher orders in the density so as to treat dense gases. Such a derivation is necessary in order to understand how we can reconcile the irreversible behavior of macroscopic systems with the reversible microscopic equations of motion.

There have been satisfactory derivations of the Boltzmann equation from the Liouville equation (Bogoliubov, 1962; Green, 1958; Green and Piccerelli, 1963; Cohen, 1962). However, the extension of the Boltzmann equation to higher orders in the density is quite involved and not yet completely clear; e.g., the transport coefficients of a dense gas cannot be expanded in a power series in the density; instead, logarithmic terms appear (Dorfman and Cohen, 1967). Computer simulations of molecular dynamics have considerably aided the development of the kinetic theory of dense gases by allowing a detailed examination of the dynamical processes taking place in the gas and by providing quantitative information as to the effects of these processes.

The other approach to the theory of nonequilibrium processes that has been especially successful is called the time-correlation function theory and is due to Green (1952, 1954) and Kubo (1957). This approach seeks special solutions of the Liouville equation for a system that is close to a total equilibrium state from which it has been perturbed by an imposed external field or a gradient in the velocity or temperature of the system. One then studies the behavior of this perturbed equilibrium system with time; shows that, on the average, the hydrodynamic equations are obeyed; and obtains explicit expressions for the transport coefficients. The expression for a typical transport coefficient τ for a classical system has the form

$$\tau = \int_0^\infty dt \langle J(0)J(t)\rangle. \qquad (47)$$

Here J represents an N-particle current that depends on the positions and velocities of all N particles; $J(0)$ is the value of this current at some initial time $t = 0$; $J(t)$ is its value a time t later, obtained by solving the equations of motion for the N-particle system; and the angular brackets denote an average over an equilibrium ensemble, taken in the thermodynamic limit. A similar expression can be given for quantum systems. These expressions are very similar to the Einstein formula for the diffusion coefficient of a particle undergoing Brownian motion. Since the derivations of the time-correlation function formulas are quite general, they apply to a wide variety of systems. However, because one must use the dynamics of an N-particle system to evaluate them, explicit calculation of these formulas presents a formidable problem, similar to, but more difficult than, the evaluation of the partition function. Nevertheless, approximation methods have been developed to evaluate these functions for a number of systems, particularly gases and liquids. Moreover, these functions are particularly suited for computer studies, where the computer follows the dynamics of the system and calculates $\langle J(0)J(t)\rangle$ directly. In addition, the time-correlation functions can be studied experimentally by means of light-scattering or neutron-scattering methods. Each of these various approaches—theory, computer studies, and experiment—stimulates progress in the other two areas.

One of the most striking examples of the progress possible by making use of the interplay between theoretical, computer, and experimental studies is the development of the mode-coupling picture for understanding transport processes in fluid and other systems (Kadanoff and Swift, 1968). This picture was stimulated by two discoveries, one experimental and the other by means of computer-simulated molecular dynamics. Experimentally it was discovered that transport coefficients such as the thermal conductivity show marked anomalies in the vicinity of the critical point of a phase transition (see CRITICAL PHENOMENA). By means of computer simulations, it was found that time-correlation functions such as those appearing in the integrand of Eq. (47) show a characteristic behavior in fluid systems, where for long times, $t \gg t_0$,

$$\langle J(0)J(t)\rangle \approx \alpha_J(t/t_0)^{-d/2}, \qquad (48)$$

where a_J is a numerical coefficient depending on density and temperature, t_0 is some characteristic microscopic time scale such as the mean free time between collisions for a particle in the fluid, and d is the dimension of the system (Alder and Wainwright, 1970). These "long-time tails" have important consequences for the understanding of transport in fluids; for example, one can see that the time-correlation function method for determining the transport coefficients does not lead to sensible results in two dimensions since the integral in Eq. (47) diverges (Dorfman and Cohen, 1970; Ernst *et al.*, 1970). This and similar observations have lead to a considerable body of research into the general structure of hydrodynamic equations for fluids and plasmas.

Mode-coupling theory is able to account for both the critical-point anomalies in transport coefficients (Kawasaki, 1970) and the long-time tails in the time-correlation functions, and has been applied to a study of other nonequilibrium phenomena such as the formation and properties of glasses. The central idea of mode-coupling theory is that microscopic processes in fluids are, to a large extent, governed by hydrodynamic-like processes whereby transport of conserved quantities on a microscopic scale takes place according to hydrodynamic laws, but with large fluctuations about them. The proper treatment of the "molecular hydrodynamics" and the microscopic fluctuations leads to the mode-coupling theory (Boon and Yip, 1991).

To conclude this brief description of nonequilibrium statistical mechanics, we mention that a very basic understanding of the foundations of statistical mechanics is likely to develop from modern research on ergodic theory as well as on nonlinear dynamics and chaos. Some important areas in which new and interesting applications of nonequilibrium statistical mechanics have been made include such topics as crystal growth and pattern formation, the theory of electric conduction in solids, flow properties of superconductors and superfluids, the dynamics of polymers, and the theory of coagulation and polymerization.

7. CONCLUSION

A wide range of equilibrium and nonequilibrium behavior of macroscopic systems can be explained and described quantitatively by the methods of classical statistical mechanics. This approach breaks down only when a quantum mechanical description of the microscopic behavior of the system is necessary.

Practically every area of physics, chemistry, and engineering has been influenced by the ideas of statistical mechanics, and in recent years we have seen the utility of a statistical mechanical approach to problems in biology, in computer science (e.g., simulated annealing, neural networks), and in cosmology (e.g., the "inflationary" cosmology). There is every reason to expect that the ideas and methods of statistical mechanics will continue to develop and to infuse and inform other scientific endeavors, and vice versa.

GLOSSARY

Boltzmann Transport Equation: An equation that describes the nonequilibrium properties of a dilute gas.

Canonical Ensemble: A phase-space distribution that describes the equilibrium properties of a system in contact with a heat reservoir with which it can exchange energy.

Ensemble: A fictitious collection of mechanically identical systems, e.g., a collection of bottles of a fluid, containers of a plasma, etc., which are all initially prepared according to some general macroscopic constraint, but which may have different microscopic configurations. Statistical mechanics replaces the computation of the exact properties of one system by a calculation of the average behavior of an ensemble of similar systems.

Ensemble Distribution Function: A phase-space distribution function that describes the properties of an ensemble of systems at a particular time.

Equilibrium Ensemble Distribution Function: see **Canonical, Grand Canonical,** and **Microcanonical Ensembles.**

Equilibrium One-, Two-, . . . Particle Distribution Function: Functions that provide the average probabilities for finding one, two, . . . particles in small regions of position and momenta for a system in equilibrium.

Equipartition of Energies: A result of equilibrium statistical mechanics for classical systems that have a Hamiltonian function that is a positive definite quadratic form in the canonical coordinates and momenta. According to this result, each quadratic term in the Hamiltonian leads to a contribution of $k_B T/2$ to the total thermodynamic internal energy, where k_B is Boltzmann's constant and T is the temperature.

Ergodic Hypothesis: The hypothesis that the trajectory in phase space describing the dynamical behavior of an isolated system spends equal amounts of time in regions of equal measure on the constraint energy surface (in phase space).

Gamma (or Γ) Space: Phase space.

Gibbs Paradox: A problem pointed out by J. W. Gibbs with early formulations of the classical statistical mechanical prescription for computing the thermodynamic properties, whereby nonzero values were obtained for the entropy of mixing of two identical systems. This problem was corrected by a careful study of the classical limit of the quantum mechanical partition function.

Grand Canonical Ensemble: A phase-space distribution that describes the equilibrium properties of a system in contact with a heat reservoir with which it can exchange energy and particles.

Long-Time Tails: Algebraically decaying terms that appear in the time-correlation functions of microscopic currents for times long compared to some characteristic microscopic times.

Maxwell–Boltzmann Distribution Function: A function of momentum, density, and temperature that describes the momentum distribution of particles in a classical fluid system in thermodynamic equilibrium.

Microcanonical Ensemble: A phase-space distribution that describes the equilibrium properties of an isolated system.

Microscopic Current: A mechanical quantity that describes the microscopic flux of mass, momentum, or energy in a system.

Partition Function: A function determined for each of the equilibrium ensembles by means of which the thermodynamic properties of a system can be computed.

Phase Space: A multidimensional space, with one dimension for each canonical coordinate and one for each canonical momentum of a mechanical system. One point in phase space gives the values of all canonical coordinates and momenta of a mechanical system. An ensemble can be described by a collection of points in phase space.

Time-Correlation Functions: Equilibrium ensemble averages of products of two or more microscopic currents, each evaluated at a possibly different time.

Virial Equation of State: An expression relating the thermodynamic pressure of a fluid system to the temperature, density, pair potential, and the equilibrium radial distribution function of the system.

Works Cited

Alder, B. J., Wainwright, T. E. (1970), *Phys. Rev. A* **1,** 18–21.

Bogoliubov, N. N. (1962), in G. E. Uhlenbeck, J. de Boer (Eds.), *Studies in Statistical Mechanics,* Vol. I, Amsterdam: North-Holland, pp. 1–118.

Boon, J. P., Yip, S. (1991), *Molecular Hydrodynamics,* reprinted by Dover Publ. Inc., New York.

Brush, S. G. (1976), *The Kind of Motion We Call Heat,* 2 Vols., Amsterdam: North-Holland.

Cohen, E. G. D. (1962), *Physica* **28,** 1025–1044, 1045–1059, 1060–1073.

Dorfman, J. R., Cohen, E. G. D. (1967), *J. Math. Phys.* **8,** 282–287.

Dorfman, J. R., Cohen, E. G. D. (1970), *Phys. Rev. Lett.* **25,** 1257–1260.

Ernst, M. H., Hauge, E. H., van Leeuwen, J. M. J. (1970), *Phys. Rev. Lett.* **12,** 1254–1256.

Fisher, M. E. (1974), *Rev. Mod. Phys.* **46,** 597–616.

Gibbs, J. W. (1960), *The Collected Works of J. Willard Gibbs,* 2 Vols., reprinted by Dover Publishers, Inc., New York.

Green, M. S. (1952), *J. Chem. Phys.* **20,** 1281–1295.

Green, M. S. (1954), *J. Chem. Phys.* **22,** 398–413.

Green, M. S. (1958), *Physica* **24,** 393–403.

Green, M. S., Piccerelli, R. A. (1963), *Phys. Rev.* **132,** 1388–1410.

Halmos, P. (1956), *Lectures on Ergodic Theory,* New York: Chelsea Publ. Co.

Hansen, J. P., McDonald, I. R. (1986), *Theory of Simple Liquids,* London: Academic Press.

Kadanoff, L. (1966), *Physics* **2,** 263–272.

Kadanoff, L. P., Götze, W., Hamblen, D., Hecht, R., Lewis, E., Palciauskas, V., Rayl, M., Swift, J. (1967), *Rev. Mod. Phys.* **39,** 395–431.

Kadanoff, L., Swift, J. (1968), *Phys. Rev.* **166,** 89–101.

Kawasaki, K. (1970), *Ann. Phys.* **61,** 1–56.

Kubo, R. (1957), *J. Phys. Soc. Jpn.* **12,** 570–586.

Mayer, J. E., Mayer, M. G. (1981), *Statistical Mechanics,* 2nd Ed., New York: Wiley.

Ruelle, D. (1969), *Statistical Mechanics: Rigor-*

ous Results, W. P. Benjamin Publ. Co., reprinted by Addison-Wesley, Reading, MA.

Sinai, Ya. G. (1970), *Russian Math. Surv.* **25,** 137–189.

Uhlenbeck, G. E., Ford, G. W. (1962), in G. E. Uhlenbeck, J. de Boer (Eds.), *Studies in Statistical Mechanics,* Vol. I, Amsterdam: North-Holland, pp. 123–211.

Widom, B. (1965), *J. Chem. Phys.* **43,** 3892–3897, 3898–3905.

Wilson, K. G. (1975), *Rev. Mod. Phys.* **47,** 773–840.

Wilson, K. G., Kogut, K. (1974), *Phys. Rep.* **12C,** 75–199.

Further Reading

General

Balescu, R. (1975), *Equilibrium and Nonequilibrium Statistical Mechanics,* New York: Wiley.

Chandler, D. (1987), *Introduction to Modern Statistical Mechanics,* New York: Oxford University Press.

Huang, K. (1987), *Statistical Mechanics,* 2nd Ed., New York: Wiley.

Kestin, J., Dorfman, J. R. (1971), *A Course in Statistical Thermodynamics,* New York: Academic.

Lifshitz, E. M., Pitaevskii, L. P. (1980), *Statistical Physics,* Part I, Part II, New York: Pergamon.

McQuarrie, D. (1976), *Statistical Mechanics,* New York: Harper and Row.

Reichl, L. E. (1980), *A Modern Course in Statistical Physics,* Austin, TX: University of Texas Press.

Reif, F. (1985), *Fundamentals of Statistical and Thermal Physics,* New York: McGraw Hill.

Toda, M., Kubo, R., Saito, N. (1991, 1992), *Statistical Mechanics,* 2 Vols., 2nd Ed., Berlin: Springer-Verlag.

Equilibrium Statistical Mechanics

Amit, D. J. (1984), *Field Theory, the Renormalization Group, and Critical Phenomena,* Singapore: World Scientific.

Binney, J. J., Dowrick, N. J., Fisher, A. J., Newman, M. E. J. (1992), *The Theory of Critical Phenomena,* Oxford: Oxford University Press.

Domb, C. S., Green, M. S., Domb, C. S., Lebowitz, J. (Eds.), *Phase Transitions and Critical Phenomena,* many volumes, New York: Academic.

Hansen, J. P., McDonald, I. R. (1986), *Theory of Simple Liquids,* 2nd Ed., London: Academic.

Thompson, C. J. (1988), *Classical Equilibrium Statistical Mechanics,* Oxford: Oxford University Press.

Yeomans, J. M. (1992), *Statistical Mechanics of Phase Transitions,* Oxford: Oxford University Press.

Ergodic Theory

Arnold, V. I., Avez, A. (1968), *Ergodic Problems of Classical Mechanics,* New York: W. A. Benjamin.

Cornfeld, I. P., Fomin, S. V., Sinai, Ya. G. (1986), *Ergodic Theory,* Berlin: Springer-Verlag.

Lebowitz, J. L. (1993), *Physica* **A194,** 1–27.

Lebowitz, J. L., Penrose, O. (1973), *Phys. Today* **26** (2), 23–29.

Peterson, K. E. (1983), *Ergodic Theory,* New York: Cambridge University Press.

Sinai, Ya. G. (1976), *Introduction to Ergodic Theory,* Princeton, NJ: Princeton University Press.

Nonequilibrium Statistical Mechanics

Boon, J. P., Yip, S. (1991), *Molecular Hydrodynamics,* reprinted by Dover Publications Inc., New York.

Chapman, S., Cowling, T. G. (1970), *The Mathematical Theory of Non-Uniform Gases,* 3rd Ed., Cambridge, U.K.: Cambridge University Press.

Cohen, E. G. D. (1993), *Physica* **A194,** 229–257.

Cohen, E. G. D. (1993), *Am. J. Phys.* **61,** 524–533.

Dorfman, J. R., van Beijeren, H. (1977), "The Kinetic Theory of Gases," in: B. J. Berne (Ed.), *Statistical Mechanics,* Part B, New York: Plenum Press.

McLennan, J. A. (1989), *Introduction to Nonequilibrium Statistical Mechanics,* Englewood Cliffs, NJ: Prentice Hall.

Pomeau, Y., Resibois, P. (1975), *Phys. Rep.* **19C,** 63–139.

Resibois, P., de Leener, M. (1977), *Classical Kinetic Theory of Fluids,* New York: Wiley.

Spohn, H. (1991), *Large Scale Dynamics of Interacting Particles,* Berlin: Springer-Verlag.

STATISTICAL MECHANICS, QUANTUM

Masuo Suzuki, *Department of Physics, University of Tokyo, Tokyo, Japan*

Introduction 518
1. **Density Matrix and Ensemble Theory** 518
1.1 Definition of Density Matrix 518
1.2 Definition of Exponential Operators 519
1.3 Density Matrix as a Projection Operator 519
2. **Average, Response, Fluctuation, and Generalized Partition Function** 519
3. **Statistical Mechanical Expressions of Thermodynamic Functions** 520
3.1 Helmholtz Free Energy 520
3.2 Equation of State for Gas and Grand Partition Function 521
4. **Fermi Statistics** 521
4.1 Symmetry of the Wave Function and Statistics of Quantum Particles 521
4.2 Fermi Statistics and Fermi Distribution 522
4.3 Dilute Fermi Gas and Maxwell–Boltzmann Distribution 523
4.4 Degenerate Fermi-Particle System 523
5. **Bose Statistics** 524
5.1 Bose Distribution Function 524
5.2 Bose–Einstein Condensation 524
5.3 Bose–Einstein Condensation of Quasiparticles in *d* Dimensions 525
5.4 Pressures for Ideal Fermi-Particle, Bose-Particle, and Classical-Particle Systems 525
6. **Harmonic Oscillators and Blackbody Radiation** 525
7. **Paramagnetism in Noninteracting Spin Systems** 526
8. **Exponential Product Formula and Quantum Monte Carlo Method** 527
8.1 Trotter Formula 527
8.2 Quantum Monte Carlo Method 527
8.3 Explicit Formulations of the Quantum Monte Carlo Method in Spin Systems 528
8.4 Generalized Decomposition Theory of Exponential Operators 529
8.5 Superoperator Formula of Ordered Exponentials 530
9. **Theory of Phase Transitions in Quantum Systems** 531
9.1 Landau's Phenomenological Theory of Phase Transitions 531
9.2 Order Parameters in Quantum Phase Transitions 532
9.3 Weiss Mean-Field Theory 532
9.4 Generalized Cluster Mean-Field Approximations 532
9.5 Coherent-Anomaly Method 533
9.6 Scaling Law and Renormalization-Group Theory 534
10. **Linear Response Theory** 535
10.1 Kubo Formula 535
10.2 Fluctuation–Dissipation Theorem and Kubo–Martin–Schwinger Condition 536
10.3 Onsager's Reciprocity Theorem 537
10.4 Exponential Perturbation Theory 537
Glossary 537
Works Cited 538
Further Reading 539

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

The density matrix plays a main role in quantum statistical mechanics, as Landau formulated it. The canonical and grand canonical ensembles introduced by Gibbs (see STATISTICAL MECHANICS, CLASSICAL) are expressed by the density matrix using the Hamiltonian and the number operator of the relevant quantum system. The basic concept of ensemble theory is the same in both classical and quantum statistical mechanics. The readers who are not familiar with this concept should look first at the article STATISTICAL MECHANICS, CLASSICAL. Fermi statistics is vital in condensed matter physics. Bose statistics plays an essential role in superconductivity and lasers. The Bose–Einstein condensation for quasiparticles with the dispersion $\epsilon(p) = ap^k$ with $k = \frac{3}{2}$ for three-dimensional helium is also explained in general d dimensions. A general decomposition theory of exponential operators such as the density matrix is also presented, which is shown to be of practical importance to quantum Monte Carlo simulations, as was formulated first in quantum spin systems (Suzuki, 1976). The quantum Monte Carlo method is used in Fermi and Bose systems as well as in spin systems. The simplest theory of phase transitions is the Weiss mean-field approximation (Weiss, 1907), which is equivalent to Landau's phenomenological theory of phase transitions. Bethe (1935) extended the Weiss theory to a cluster effective-field approximation. Even modern critical phenomena can be analyzed using the coherent-anomaly method (Suzuki, 1986) based on a systematic (or coherent) series of mean-field approximations. The linear response theory on transport phenomena was established by Kubo (1957) in the framework of quantum statistical mechanics.

1. DENSITY MATRIX AND ENSEMBLE THEORY

1.1 Definition of Density Matrix

Landau introduced the density matrix ρ by taking into account both the probabilistic character of quantum mechanics and statistics of ensemble theory. First the wave function of the relevant system, ψ, is expanded as a linear combination of the eigenfunctions $\{\psi_j\}$ of the Hamiltonian $\mathcal{H}$: $\psi = \Sigma c_j\psi_j$. The corresponding state in the Hilbert space is expressed as $|\psi\rangle = \Sigma c_j|\psi_j\rangle$. The expectation value of a physical quantity (operator) A over this state is written as

$$\langle\psi|A|\psi\rangle = \sum_{i,j} c_j c_i^*\langle\psi_i|A|\psi_j\rangle, \tag{1}$$

where c_i^* denotes the complex conjugate of c_i. The coefficients $\{c_jc_i^*\}$ in Eq. (1) are statistical quantities in the ensemble theory of statistical mechanics, as was mentioned above. Following Landau, we introduce here the statistical weight w_{ji} as the statistical average of the coefficient $c_jc_i^*$, and consequently the average of the physical quantity A over the corresponding statistical ensemble is expressed in the following matrix form:

$$\bar{A} = \sum_{i,j} w_{ji}\langle i|A|j\rangle = \mathrm{Tr}\,\rho A;\ \rho = (w_{ji}). \tag{2}$$

If we make use of the energy representation diagonalizing the Hamiltonian $\mathcal{H}$ (namely, $\mathcal{H}|\psi_j\rangle = E_j|\psi_j\rangle$), the off-diagonal matrix elements $\{w_{ji}\}$ for $j \neq i$ are considered to be vanishing because of statistical cancellation. In particular, the diagonal element w_{jj} is expressed in the form

$$w_{jj} = e^{-\beta E_j}/Z(\beta);\ Z(\beta) = \sum e^{-\beta E_j} \tag{3}$$

with $\beta = 1/k_BT$ in the canonical ensemble specified by the temperature T (see STATISTICAL MECHANICS, CLASSICAL). Here, k_B denotes the Boltzmann constant given by $k_B = 1.380\,658 \times 10^{-23}$ J K^{-1}. Thus, the density matrix ρ is given by

$$\rho = e^{-\beta\mathcal{H}}/Z(\beta);\ Z(\beta) = \mathrm{Tr}\,e^{-\beta\mathcal{H}}. \tag{4}$$

Similarly, the density matrix of the grand canonical ensemble is given by

$$\rho = \frac{e^{-\beta(\mathcal{H}-\mu\mathcal{N})}}{\Xi(\beta\mu,\beta)};\ \Xi(\alpha,\beta) = \mathrm{Tr}\,e^{-\beta\mathcal{H}+\alpha\mathcal{N}}, \tag{5}$$

where $\mathcal{N}$ denotes the number operator of particles. Here μ denotes the chemical potential determined by the condition

$$N = \mathrm{Tr}\,\mathcal{N}\rho \tag{6}$$

for the given average number of particles, N.

1.2 Definition of Exponential Operators

It is instructive to define here an exponential operator $\exp(-\beta\mathcal{H})$ as

$$e^{-\beta\mathcal{H}} = \sum_{n=0}^{\infty} \frac{(-\beta)^n}{n!}\mathcal{H}^n \tag{7}$$

or equivalently

$$e^{-\beta\mathcal{H}} = (e^{-1/n\beta\mathcal{H}})^n = \lim_{n\to\infty}\left(1 - \frac{\beta}{n}\mathcal{H}\right)^n. \tag{8}$$

The convergence of the above formulas (7) and (8) is assured for bounded operators $\mathcal{H}$, namely, in Banach space.

More detailed explanation of exponential operators will be given later.

1.3 Density Matrix as a Projection Operator

In order to discuss the physical meaning of the density matrix, we consider the projection operator P_j defined by $P_j = |j\rangle\langle j|$ in terms of the eigenstate $|j\rangle(\equiv|\psi_j\rangle)$ of the Hamiltonian $\mathcal{H}$. Thus, we have

$$P_j|\psi\rangle = (\langle j|\psi\rangle)|j\rangle, \tag{9}$$

as shown in Fig. 1.

The density matrix ρ is expressed in the form

$$\rho = \sum_j w(E_j)P_j = \sum_j w(E_j)|j\rangle\langle j|, \tag{10}$$

using the projection operator P_j. Here, $w(E_j)$ denotes the statistical weight of the eigenstate $|j\rangle$ in each ensemble. As shown in Eq. (3), $w(E_j) = e^{-\beta E_j}/Z(\beta)$ in the canonical ensemble and $w(E_j) = \exp[-\beta(E_j - \mu N)] \times \Xi(\beta\mu,\beta)$ in the grand canonical ensemble.

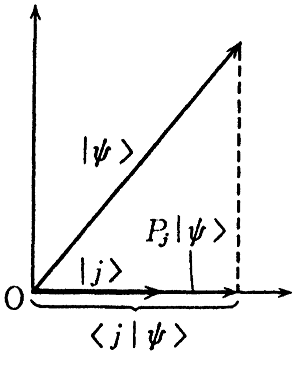

FIG. 1. Schematic explanation of the projection operator P_j.

2. AVERAGE, RESPONSE, FLUCTUATION, AND GENERALIZED PARTITION FUNCTION

In order to evaluate the average and response of an arbitrary physical quantity Q, we introduce here the following generalized partition function:

$$\Omega(\beta,h) = \operatorname{Tr}\hat{\rho} \equiv \operatorname{Tr} e^{-\beta(\hat{\mathcal{H}}-hQ)} \tag{11}$$

with a parameter h conjugate to Q. Here, $\hat{\mathcal{H}}$ denotes $\mathcal{H}$ or $\mathcal{H} - \mu\mathcal{N}$, corresponding to the canonical or grand canonical ensemble.

Then, the average of Q is given by

$$\bar{Q} = \frac{\operatorname{Tr} Qe^{-\beta\hat{\mathcal{H}}}}{\Omega(\beta,0)} = \left[\frac{\partial}{\partial(\beta h)}\log\Omega(\beta,h)\right]_{h=0}. \tag{12}$$

Here, we have used the formula

$$\frac{d}{dx}\operatorname{Tr} e^{A+xB} = \operatorname{Tr} Be^{A+xB}. \tag{13}$$

By the way, it is instructive to remark the following operator identity:

$$\frac{d}{dx}e^{A+xB} = \int_0^1 e^{(1-\lambda)(A+xB)}Be^{\lambda(A+xB)}d\lambda. \tag{14}$$

This yields the following perturbational formula up to the first order of x:

$$e^{A+xB} = e^A\left(1 + x\int_0^1 e^{-\lambda A}Be^{\lambda A}d\lambda + \cdots\right). \tag{15}$$

Now we consider a certain set of physical quantities $\{Q_j\}$. The response of Q_j with respect to the corresponding conjugate field h_k is expressed by

$$\bar{Q}_j(\{h_j\}) = \bar{Q}_j(\{0\}) + \sum_k \chi_{jk}h_k + O(h_k^2), \tag{16}$$

using the formula (14). Here, the response function χ_{jk} is given by

$$\chi_{jk} = \int_0^\beta (\langle Q_k(-i\hbar\lambda)Q_j\rangle - \langle Q_k\rangle\langle Q_j\rangle)d\lambda, \tag{17}$$

where $Q(z)$ denotes the following interaction representation:

$$Q(z) = e^{iz\hat{\mathcal{H}}/\hbar}Qe^{-iz\hat{\mathcal{H}}/\hbar}. \tag{18}$$

Here, $\hbar = h/2\pi$ and h denotes the Planck constant. In particular, if Q_k commutes with the Hamiltonian $\hat{\mathcal{H}}$, then we have

$$\begin{aligned}\chi_{jk} &= \beta(\langle Q_kQ_j\rangle - \langle Q_k\rangle\langle Q_j\rangle)\\ &= \beta\langle (Q_k - \langle Q_k\rangle)(Q_j - \langle Q_j\rangle)\rangle.\end{aligned} \tag{19}$$

Furthermore, χ_{jj} is given by

$$\chi_{jj} = \beta\langle (Q_j - \langle Q_j\rangle)^2\rangle. \tag{20}$$

Thus, the response function χ_{jj} is expressed in terms of the fluctuation of Q_j in the absence of the external field h_j.

For example, the susceptibility of a ferromagnet, $\chi_0(T)$, is expressed in terms of spin correlation functions as

$$\chi_0(T) = \lim_{H\to 0}\frac{1}{H}\mu_B\sum_j \langle S_j^z\rangle_H = \frac{g\mu_B^2}{k_BT}\sum_{i,j}\langle S_i^zS_j^z\rangle. \tag{21}$$

Here, μ_B and g denote the Bohr magneton and the Landé factor, respectively. The total magnetization is expressed by the operator $M = g\mu_B\Sigma_j S_j^z$, where S_j^z denotes the z component of the spin operator S_j^z at the site j. The parameter H in Eq. (21) denotes an external magnetic field, and the average $\langle\cdots\rangle_H$ is defined using the canonical ensemble in the presence of the magnetic field H. Similarly, the specific heat C_v is given by the response of the energy with respect to the inverse temperature β as

$$C_v = \frac{1}{k_BT^2}\langle(\mathcal{H} - \langle\mathcal{H}\rangle)^2\rangle. \tag{22}$$

3. STATISTICAL MECHANICAL EXPRESSIONS OF THERMODYNAMIC FUNCTIONS

Thermodynamic functions such as the Helmholtz and Gibbs free energies are expressed using the partition functions in the canonical and grand canonical ensembles.

3.1 Helmholtz Free Energy

The Helmholtz free energy defined thermodynamically by $F = E - TS$ for the energy E and the entropy S is given microscopically by

$$F = -k_BT\ln Z(\beta), \tag{23}$$

where $Z(\beta)$ denotes the partition function of the canonical ensemble, namely

$$Z(\beta) = \operatorname{Tr} e^{-\beta\mathcal{H}}. \tag{24}$$

Expression (23) is easily derived using the saddle-point method from the Boltzmann principle

$$S = k_B\ln W, \tag{25}$$

where W is the number of states for the specified energy range $E \sim E + \Delta E$. Thus, the entropy (25) is a function of the energy E as $S = S(E)$. The thermodynamic relation

$$\frac{\partial S(E)}{\partial E} = \frac{1}{T} \tag{26}$$

gives the temperature dependence of the energy and entropy. This is the scheme of the microcanonical ensemble. The number of states W is also expressed in terms of the density of states $D(E)$ as

$$W = D(E)\Delta E. \tag{27}$$

Thus, the partition function $Z(\beta)$ is written as

$$\begin{aligned}Z(\beta) &= \operatorname{Tr} e^{-\beta\mathcal{H}} = \sum e^{-\beta E_j} = \int_0^\infty e^{-\beta E}D(E)dE\\ &= \int_0^\infty e^{-\beta(E-TS(E))}dE \simeq \exp[-\beta(E - TS)],\end{aligned} \tag{28}$$

using the saddle-point method. This yields the relation (23).

3.2 Equation of State for Gas and Grand Partition Function

Thermodynamically, the pressure p of a gas in the volume V at the temperature T is given by

$$pV = G - F = \mu N - F, \tag{29}$$

where G is the Gibbs free energy. The grand partition function defined by

$$\Xi(\lambda) \equiv \Xi(\ln \lambda, \beta) = \mathrm{Tr}\, \lambda^{\mathcal{N}} e^{-\beta\mathcal{H}} \tag{30}$$

is related to the partition function $Z_N(\beta)$ for the particle number N as

$$\Xi(\lambda) = \sum_{N=0}^{\infty} \lambda^N Z_N(\beta). \tag{31}$$

Conversely, the partition function $Z_N(\beta)$ is expressed by the following integral:

$$Z_N(\beta) = \frac{1}{2\pi i} \oint \frac{\Xi(\lambda)}{\lambda^{N+1}}\, d\lambda \tag{32}$$

taken counterclockwise over a circle surrounding the origin. This integral is also rewritten as

$$Z_N(\beta) = \frac{1}{2\pi i} \oint \exp[\ln \Xi(\lambda) - N \ln \lambda] \frac{d\lambda}{\lambda}. \tag{33}$$

Using again the saddle-point method in Eq. (33), we arrive at the relation

$$F = -k_{\mathrm{B}}T \ln Z_{\bar{N}}(\beta) = -k_{\mathrm{B}}T \ln \Xi(\lambda) + \mu N, \tag{34}$$

where the number N is given by

$$N = \lambda \frac{\partial}{\partial \lambda} \ln \Xi(\lambda). \tag{35}$$

Thus we obtain the formula (29). Namely, the equation of state for a gas is expressed as

$$pV = k_{\mathrm{B}}T \ln \Xi. \tag{36}$$

As a simple application of Eq. (36), we derive the equation of state for an ideal classical gas as follows. The partition function for a classical gas (see STATISTICAL MECHANICS, CLASSICAL) is given by

$$\begin{aligned} Z_N(\beta) &= \frac{1}{N! h^{3N}} \int \cdots \int e^{-\beta\mathcal{H}} d\Gamma \\ &= \frac{1}{N! h^{3N}} \left(\int \exp\left(-\frac{\beta p^2}{2m} \right) d^3p\, d^3q \right)^N \\ &= \frac{1}{N!} \left[V \left(\frac{2\pi m k_{\mathrm{B}} T}{h^2} \right)^{3/2} \right]^N . \end{aligned} \tag{37}$$

Then, the grand partition function Ξ for this classical system is obtained as

$$\Xi = \sum_{N=0}^{\infty} \lambda^n Z_N(\beta) = \exp\left[\lambda V \left(\frac{2\pi m k_{\mathrm{B}} T}{h^2} \right)^{3/2} \right]. \tag{38}$$

On the other hand, the number of particles is expressed as

$$N = \frac{\partial}{\partial \alpha} \ln \Xi = \lambda \frac{\partial}{\partial \lambda} \ln \Xi = \lambda V \left(\frac{2\pi m k_{\mathrm{B}} T}{h^2} \right)^{3/2}. \tag{39}$$

Thus, we arrive at the equation of state for an ideal classical gas:

$$pV = k_{\mathrm{B}} T \lambda V \left(\frac{2\pi m k_{\mathrm{B}} T}{h^2} \right)^{3/2} = N k_{\mathrm{B}} T. \tag{40}$$

This is the well-known Boyle–Charles law. This result will be used in discussing later the relation among the three expressions p_{cl}, p_{F}, and p_{B} of pressure for an ideal classical gas, the ideal Fermi-particle system, and the ideal Bose-particle system.

4. FERMI STATISTICS

As a typical example of quantum statistical mechanics, we explain here Fermi statistics.

4.1 Symmetry of the Wave Function and Statistics of Quantum Particles

According to quantum mechanics, the state of the relevant quantum system with N particles is described by the wave function $\psi(r_1, r_2, \cdots, r_N)$. Since the quantum particles are indistinguishable among each other, the wave function $\psi(\cdots,r_j,\cdots,r_i,\cdots)$ obtained by the interchange of the two particles specified by the coordinates r_i and r_j is not independent; namely, it can be expressed as

$$\psi(\cdots,r_j,\cdots,r_i,\cdots) = C\psi(\cdots,r_i,\cdots,r_j,\cdots). \tag{41}$$

If we apply this relation once more, then we obtain $C^2 = 1$. Therefore, the constant C in Eq. (41) has to satisfy the requirement $C = \pm 1$. In the case of $C = 1$, the wave function ψ is symmetric and the corresponding particles satisfying this symmetry relation are called Bose particles. The statistics of Bose particles is called Bose statistics. In the case of $C = -1$, the wave function ψ is antisymmetric, and the corresponding particles are called Fermi particles. The statistics of Fermi particles is called Fermi statistics. According to this Fermi statistics, two particles cannot take the same state. This is also called the Pauli principle. These two kinds of statistics are classified by the values of spins of particles. It is derived generally that particles with integer spins obey Bose statistics and particles with half-odd integer spins obey Fermi statistics. For example, electrons (e^-), positrons (e^+), protons (p), and neutrons (n) have spin $\frac{1}{2}$, and they obey Fermi statistics. Photons obey Bose statistics. Composite particles of Bose particles obey Bose statistics. Composite particles of an even number of Fermi particles are easily shown to obey also Bose statistics. Composite particles of an odd number of Fermi particles are shown to be Fermi particles.

4.2 Fermi Statistics and Fermi Distribution

Here we consider an ideal Fermi-particle system whose states are specified using the following one-particle states:

$$\begin{aligned} &\text{states} \rightarrow 1, 2, 3, \ldots, j, \ldots, \\ &\text{energy} \rightarrow \epsilon_1, \epsilon_2, \epsilon_3, \ldots, \epsilon_j, \ldots, \\ &\text{number} \rightarrow n_1, n_2, n_3, \ldots, n_j, \ldots. \end{aligned} \tag{42}$$

In the case of Fermi statistics, the Pauli principle requires the condition that $n_j = 0$ or 1. The total number N and the total energy E of Fermi particles satisfy the following relations:

$$\sum_j n_j = N \quad \text{and} \quad \sum_j n_j\epsilon_j = E, \tag{43}$$

respectively.

The grand partition function $\Xi(\alpha,\beta)$ of an ideal Fermi-particle system is given by

$$\begin{aligned} \Xi(\alpha,\beta) &= \sum_{n_j=0,1} \exp\left(-\beta \sum_j n_j\epsilon_j + \alpha \sum_j n_j\right) \\ &= \prod_{j=1}^{\infty} (1 + e^{-\beta\epsilon_j+\alpha});\ \alpha = \beta\mu. \end{aligned} \tag{44}$$

The average number of particles in the state j is obtained as

$$\begin{aligned} \bar{n}_j &= -\frac{\partial}{\partial(\beta\epsilon_j)} \ln \Xi(\alpha,\beta) \\ &= \frac{e^{-\beta\epsilon_j+\alpha}}{1 + e^{-\beta\epsilon_j+\alpha}} = \frac{1}{e^{\beta(\epsilon_j-\mu)} + 1}. \end{aligned} \tag{45}$$

The chemical potential μ is determined by the condition

$$\bar{N} = \sum_j \bar{n}_j = \sum_j \frac{1}{e^{\beta(\epsilon_j-\mu)} + 1}. \tag{46}$$

The above distribution function (45) of a Fermi particle is called the Fermi distribution, and it is written as

$$f_{\mathrm{F}}(\epsilon) = \frac{1}{e^{\beta(\epsilon-\mu)} + 1}. \tag{47}$$

The chemical potential μ for the Fermi distribution function (47) is called the Fermi level or Fermi energy. This depends on the temperature T. The Fermi energy at $T = 0$ is written as μ_0. The Fermi distribution function $f_{\mathrm{F}}(\epsilon)$ at $T = 0$ becomes a step function, as shown in Fig. 2. At finite temperatures, $f_{\mathrm{F}}(\epsilon)$ becomes a smooth function, as shown in Fig. 2.

The Helmholtz free energy of this ideal Fermi-particle system is given in the form

$$\begin{aligned} \mathrm{F} &= E - TS = -k_{\mathrm{B}}T \ln Z(\beta) \\ &= -k_{\mathrm{B}}T \ln \Xi(\alpha,\beta) + \mu N \\ &= \mu N - k_{\mathrm{B}}T \sum_j \ln(1 + e^{-\beta(\epsilon_j-\mu)}). \end{aligned} \tag{48}$$

The energy of this system is expressed by

$$E = \sum_j \bar{n}_j\epsilon_j = \sum_j \frac{\epsilon_j}{e^{\beta(\epsilon_j-\mu)} + 1}. \tag{49}$$

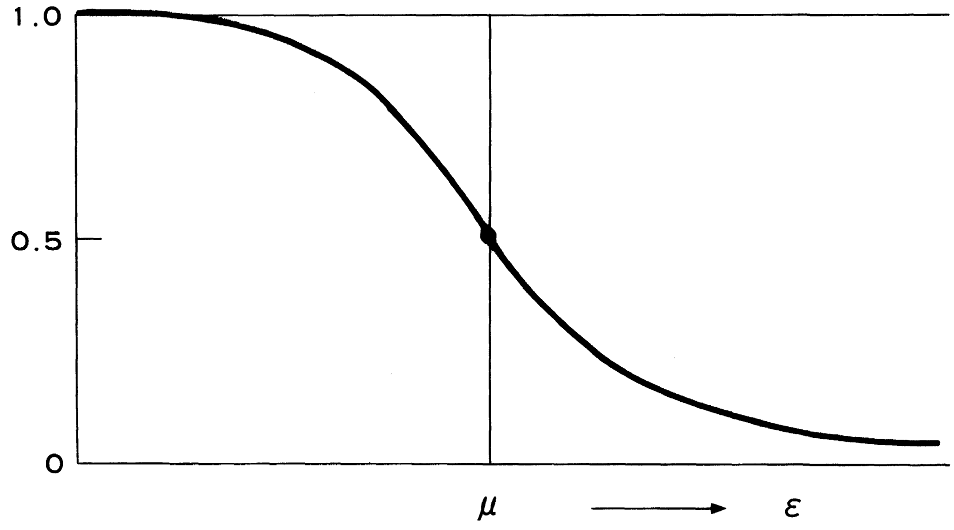

FIG. 2. Fermi distribution function as a function of the energy ϵ. Here μ is the chemical potential or the Fermi level.

The entropy of this Fermi-particle system is given in the form

$$S = -k_B \sum_j \{\bar{n}_j \ln \bar{n}_j + (1 - \bar{n}_j) \ln(1 - \bar{n}_j)\}. \tag{50}$$

If we introduce the density of states $D(\epsilon)$, the energy, the number, and the Helmholtz free energy are expressed by the following integrals:

$$\begin{aligned} E &= \int_0^\infty \epsilon f_F(\epsilon) D(\epsilon) d\epsilon, \\ N &= \int_0^\infty f_F(\epsilon) D(\epsilon) d\epsilon, \\ F &= \mu N - k_B T \int_0^\infty \ln(1 + e^{-\beta(\epsilon - \mu)}) D(\epsilon) d\epsilon. \end{aligned} \tag{51}$$

4.3 Dilute Fermi Gas and Maxwell–Boltzmann Distribution

When the number density of Fermi particles becomes very low, the Fermi distribution $f_F(\epsilon)$ reduces to the classical distribution

$$f_{cl}(\epsilon) = \lambda e^{-\beta\epsilon}, \tag{52}$$

neglecting $+1$ in the denominator of (47) under the condition that $\lambda \equiv e^{\beta\mu} \ll 1$. Equation (52) is called the Maxwell–Boltzmann distribution. The validity of this approximation (classical approximation), namely $\lambda \ll 1$, is assured by the condition

$$\lambda_D \ll r, \tag{53}$$

where r denotes the mean distance of particles, namely $r \simeq \rho^{-1/3}$ for $\rho = N/V$, and λ_D is the thermal de Broglie wavelength defined by

$$\lambda_D = h/\sqrt{3mk_BT} \tag{54}$$

for the mass of a Fermi particle equal to m.

4.4 Degenerate Fermi-Particle System

The state of the Fermi-particle system near $T = 0$ is called degenerate, because almost all levels up to the Fermi level are occupied by Fermi particles. The Fermi level μ_0 at $T = 0$ is determined by the condition

$$N = \int_0^{\mu_0} D(\epsilon) d\epsilon. \tag{55}$$

On the other hand, the energy of an ideal particle is given by $\epsilon = p^2/2m$. Thus, the density of states is evaluated as follows:

$$D(\epsilon) = g_s \frac{d}{d\epsilon}\left(\frac{V}{h^3}\int_0^{\sqrt{2m\epsilon}} 4\pi p^2 dp\right) = 2\pi g_s V\left(\frac{2m}{h^2}\right)^{3/2}\sqrt{\epsilon}. \quad (56)$$

Here, g_s is the spin degeneracy; namely, $g_s = 2$ for electrons. Therefore, the Fermi level μ_0 is given explicitly as

$$\mu_0 = \frac{h^2}{2m}\left(\frac{3N}{4\pi g_s V}\right)^{2/3} = \frac{3N}{2D(\mu_0)}. \quad (57)$$

The following formula is convenient for expanding μ, E, and C_v with respect to T:

$$\int_0^\infty g(\epsilon)f(\epsilon)d\epsilon = \int_0^\mu g(\epsilon)d\epsilon + \frac{(\pi k_B T)^2}{6}g^{(1)}(\mu) + \frac{7(\pi k_B T)^4}{360}g^{(3)}(\mu) + \cdots, \quad (58)$$

where $g^{(n)}(\epsilon)$ denotes the nth derivative of $g(\epsilon)$. Using this formula, we obtain

$$\begin{cases} \mu = \mu_0 - \dfrac{\pi^2}{6}\dfrac{D'(\mu_0)}{D(\mu_0)}(k_B T)^2 + \cdots, \\ E = E_0 + \dfrac{\pi^2}{6}D(\mu_0)(k_B T)^2 + \cdots. \end{cases} \quad (59)$$

Thus, the specific heat of this system near $T = 0$ is shown to be proportional to T as

$$C_v = \frac{dE}{dT} = \frac{\pi^2}{3}D(\mu_0)k_B^2 T + \cdots. \quad (60)$$

The specific heat of electrons has been observed experimentally to behave like Eq. (60) even up to room temperatures.

5. BOSE STATISTICS

Since the number of Bose particles in the state j is allowed to take all the values $n_j = 0, 1, 2, \ldots$, the grand partition function $\Xi(\alpha,\beta)$ of ideal Bose systems is given by

$$\Xi(\alpha,\beta) = \prod_{j=1}^{\infty}\left\{\sum_{n=0}^{\infty}(e^{-\beta\epsilon_j+\alpha})^n\right\} = \prod_{j=1}^{\infty}(1 - e^{-\beta\epsilon_j+\alpha})^{-1}. \quad (61)$$

5.1 Bose Distribution Function

The average number of n_j is thus evaluated as

$$\bar{n}_j \equiv f_B(\epsilon_j) = -\frac{\partial}{\partial(\beta\epsilon_j)}\ln \Xi(\alpha,\beta) = \frac{1}{e^{\beta(\epsilon_j-\mu)} - 1}. \quad (62)$$

Here, the chemical potential is determined by the condition

$$\sum_j \frac{1}{e^{\beta(\epsilon_j-\mu)} - 1} = N. \quad (63)$$

It is seen from (62) that $\mu \le 0$ because of the requirement that $\bar{n}_j \ge 0$ for all $\epsilon_j \ge 0$. The above Bose distribution function (62) also reduces to the Maxwell–Boltzmann distribution function (52) in the low-density limit.

5.2 Bose–Einstein Condensation

An ideal Bose-particle system is shown to show a phase transition at low temperatures as follows. The density of states for ideal particles with mass m is given by

$$D(\epsilon) = 2\pi V\left(\frac{2m}{h^2}\right)^{3/2}\sqrt{\epsilon} \quad (64)$$

or

$$D(p)dp = \frac{V}{h^3}4\pi p^2 dp. \quad (65)$$

Here, we have assumed that the spin degeneracy g_s equals unity (i.e., $g_s = 1$). The total number N is expressed as

$$N = \sum_j \bar{n}_j = \frac{V}{h^3}\int_0^\infty \frac{4\pi p^2 dp}{e^{\beta(p^2/2m-\mu)} - 1} \equiv Vh(T,\mu), \quad (66)$$

where the function $h(T,\mu)$ satisfies the inequality $h(T,\mu) \ge h(T,0)$. This means that Eq. (66) is valid only in the range of density given by $\rho \equiv N/V \ge h(T,0)$. For the density $\rho > h(T,0)$, the number of particles for $\epsilon = 0$ (or $p = 0$) has to be macroscopic. For this requirement, we have $\mu = 0$, and Eq. (66) should be replaced by

$$N = N_0 + Vh(T,0). \tag{67}$$

Here, N_0 denotes the number of particles in the state $\epsilon = 0$. This situation occurs below the critical point T_c defined by

$$h(T_c,0) = \rho. \tag{68}$$

This phenomenon is called the Bose–Einstein condensation. The superfluidity of ^{4}He observed experimentally is closely related to this Bose–Einstein condensation. The function $h(T,0)$ is given more explicitly as

$$h(T,0) = 2.612\left(\frac{2\pi m k_B T}{h^2}\right)^{3/2}. \tag{69}$$

Thus, the critical point of the Bose–Einstein condensation depends on the number density ρ as

$$T_c = \frac{h^2}{2\pi k_B m}\left(\frac{\rho}{2.612}\right)^{2/3} \propto \frac{h^2}{m}\rho^{2/3}. \tag{70}$$

In liquid helium, we obtain $T_c = 3.14$ K from Eq. (70). The condensation density $\rho_0 = N_0/V$ has the following temperature dependence:

$$\rho_0 = h(T,0) - h(T_c,0) = \rho[1 - (T/T_c)^{3/2}]. \tag{71}$$

It is also easily shown that the specific heat of this system behaves like $C_v \propto T^{3/2}$. The temperature derivative of C_v is discontinuous at T_c, and consequently the Bose–Einstein condensation is of third order.

5.3 Bose–Einstein Condensation of Quasiparticles in *d* Dimensions

When there exists an interaction among particles, such interacting particles may be treated as quasiparticles with the new spectrum $\epsilon_p = ap^k$ modified by the interaction, where $a > 0$ and $k > 0$. In most cases, we have $k = 2$, but strong many-body interaction may change the value of k, as in superfluid liquids that show a spectrum with $k = 3/2$. In such a generalized Bose quasiparticle system, the total number of such particles, N, is given in the form

$$N = N_0 + \frac{g_S L^d}{h^d}\int_0^\infty \frac{S_d p^{d-1}dp}{e^{\beta(ap^k-\mu)} - 1}. \tag{72}$$

in d dimensions, where $g_s = 2S + 1$ (spin degeneracy), L denotes the system size, and $S_d = 2\pi^{d/2}/\Gamma(d/2)$. Here $\Gamma(x)$ is the gamma function. The critical point below which this system shows the Bose–Einstein condensation is given by

$$\begin{aligned} T_c &= \frac{a}{k_B}\left(\frac{k\Gamma(d/2)h^d\rho}{2\pi^{d/2}g_s\Gamma(d/k)\zeta(d/k)}\right)^{k/d} \\ &\propto ah^k[\rho/(2S + 1)]^{k/d}. \end{aligned} \tag{73}$$

Here $\zeta(x)$ is the zeta function. This transition occurs for $d > k$.

5.4 Pressures for Ideal Fermi-Particle, Bose-Particle, and Classical-Particle Systems

It is quite interesting to compare the pressures p_F, p_B, and p_{cl} of the ideal Fermi-particle, Bose-particle, and classical-particle systems, respectively, at the same number density. It has been proved (Suzuki, 1994a) that

$$P_F > P_{cl} > P_B \tag{74}$$

for the same number density. Here we have used the inequalities $\ln(1 + x) > x/(1 + x)$ for $x > 0$ and $\ln[1/(1 + x)] < x/(1 - x)$ for $0 < x < 1$. Thus, an effective repulsive interaction appears in an ideal Fermi-particle system because of the Pauli principle and an effective attractive interaction appears in an ideal Bose-particle system.

6. HARMONIC OSCILLATORS AND BLACKBODY RADIATION

According to quantum mechanics, the eigenvalues of the one-dimensional harmonic oscillator described by the Hamiltonian

$$\mathcal{H} = p^2/2m + \tfrac{1}{2}kq^2 \tag{75}$$

are given by $\epsilon_n = (n + \frac{1}{2})h\nu$, where the eigenfrequency ν is expressed as $\nu = (k/m)^{1/2}/2\pi$. The partition function of the system (75), $Z_1(\beta)$, is given as

$$Z_1(\beta) = \sum_{n=0}^{\infty} e^{-\beta\epsilon_n} = \frac{1}{2\sinh(\beta h\nu/2)}. \tag{76}$$

Thus, the average energy of one harmonic oscillator is calculated as

$$\epsilon = -\frac{\partial}{\partial\beta}\ln Z_1(\beta) = \tfrac{1}{2}h\nu + \frac{h\nu}{e^{\beta h\nu}-1}. \tag{77}$$

The specific heat of N independent harmonic oscillators is given by

$$C = N\frac{d\epsilon}{dT} = \frac{N(h\nu)^2}{k_B T^2}\frac{e^{\beta h\nu}}{(e^{\beta h\nu}-1)^2}. \tag{78}$$

This approaches the classical value $C_{\mathrm{cl}} = Nk_B$ in the limit $\beta h\nu \to 0$.

Now we discuss blackbody radiation using the result (77). The photon energy of frequency ν is given by $\epsilon(\nu) = h\nu/[\exp(\beta h\nu) - 1]$. The density of states of photons is written as

$$D(\nu)d\nu = (8\pi V/c^3)\nu^2 d\nu, \tag{79}$$

where c denotes the light velocity. Thus, the energy density of the blackbody radiation is expressed as

$$\rho(\nu)d\nu = \frac{1}{V}\epsilon(\nu)D(\nu)d\nu = \frac{8\pi h}{c^3}\frac{\nu^3 d\nu}{e^{\beta h\nu}-1}. \tag{80}$$

In the low-frequency region (i.e., $h\nu \ll k_B T$), this reduces to the Rayleigh–Jeans law

$$\rho(\nu) = (8\pi\nu^2/c^3)k_B T, \tag{81a}$$

and in the high-frequency region (i.e., $h\nu \gg k_B T$), Eq. (80) reduces to Wien's law

$$\rho(\nu) = (8\pi h/c^3)\nu^3 e^{-h\nu/k_B T}. \tag{81b}$$

7. PARAMAGNETISM IN NONINTERACTING SPIN SYSTEMS

When we treat the magnetism of an atom or an ion, we have to consider the total angular momentum $J\hbar = (L + S)\hbar$, where $L\hbar$ and $S\hbar$ denote the orbital angular momentum and spin angular momentum, respectively. According to quantum mechanical calculations, the magnetic moment operator μ is expressed as

$$\mu = (-e\hbar/2mc)(L + 2S);\ \hbar \equiv h/2\pi. \tag{82}$$

As J is a constant of motion, the average moment $\bar{\mu}$ projected on the direction of J plays a physical role, and it is easily given by

$$\bar{\mu} = -g\frac{e\hbar}{2mc}J;\quad g = \frac{3}{2} + \frac{S(S+1) - L(L+1)}{2J(J+1)}, \tag{83}$$

where

$$\mathbb{S}^2 = S(S+1),\ \mathbb{L}^2 = L(L+1),\ \text{and}\ \mathbb{J}^2 = J(J+1). \tag{84}$$

The quantity g is called the Landé factor. Then the Zeeman energy due to this magnetic moment $\bar{\mu}$ in a magnetic field H is given by

$$-\bar{\boldsymbol{\mu}}\cdot\mathbf{H} = g\mu_B\mathbf{J}\cdot\mathbf{H}, \tag{85}$$

where

$$\mu_B = \frac{e\hbar}{2mc} = \frac{eh}{4\pi mc}. \tag{86}$$

Now we consider a system of N independent spins in a magnetic field H along the z axis. The Hamiltonian of this system is expressed as

$$\mathcal{H} = -g\mu_B\sum_{j=1}^{N} S_j^z H, \tag{87}$$

where S_j^z denotes the z component of the spin operator S_j; namely, $S_j^z = S, S-1, S-2, \ldots, -S+1, -S$. The partition function of this system is calculated as

$$Z_N(\beta) = Z_1^N(\beta);$$
$$Z_1(\beta) = \sum_{m=-S}^{S} e^{\beta g\mu_B m H} = \frac{\sinh[\frac{1}{2}(2S+1)\beta g\mu_B H]}{\sinh(\frac{1}{2}\beta g\mu_B H)}. \tag{88}$$

The average energy of this system is then given by

$$E(T) = -Ng\mu_B SHB_S(\beta g\mu_B SH), \tag{89}$$

where $B_s(x)$ is the Brillouin function defined by

$$B_S(x) = \frac{2S+1}{2S}\coth\left(\frac{2S+1}{2S}x\right) - \frac{1}{2S}\coth\left(\frac{1}{2S}x\right). \tag{90}$$

In the limit of large S, this reduces to the Langevin function

$$L(x) = \coth x - 1/x. \tag{91}$$

The magnetization $M(T,H)$ is thus given by

$$M(T,H) = k_BT\frac{\partial}{\partial H}\ln Z_N(\beta) = Ng\mu_BSB_S(\beta g\mu_BSH). \tag{92}$$

The susceptibility of this system is obtained as

$$\chi_0(T) \equiv \lim_{H\to 0}\frac{M(T,H)}{H} = \frac{N(g\mu_B)^2S(S+1)}{3k_BT}. \tag{93}$$

This is called the Hund equation, and the behavior of $\chi_0(T)$ shown in (93) [namely, $\chi_0(T) \propto 1/T$] is called the Curie law. In particular, it reduces to

$$\chi_0(T) = N\mu_B^2/k_BT \tag{94}$$

in the case of $S = \frac{1}{2}$ for $g = 2$.

8. EXPONENTIAL PRODUCT FORMULA AND QUANTUM MONTE CARLO METHOD

In quantum statistical mechanics, exponential operators such as the density matrix (4) play an important role. It is a crucial problem how to manipulate such exponential operators.

In most cases, the Hamiltonian $\mathcal{H}$ of the relevant quantum system is composed of two or more partial Hamiltonians as $\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2 + \cdots + \mathcal{H}_q$. In many physical problems, each of these partial Hamiltonians can be easily diagonalized independently, though it is extremely difficult to diagonalize the total Hamiltonian $\mathcal{H}$.

In these situations, it is very convenient to decompose the exponential operator $e^{-\beta\mathcal{H}}$ into the product of the partial density matrices $\{e^{-\beta'\mathcal{H}_j}\}$ with different inverse temperatures β', because it is quite easy to evaluate explicitly the matrix elements of the partial density matrices $\{e^{-\beta'\mathcal{H}_j}\}$ when each of the $\{\mathcal{H}_j\}$ can be diagonalized.

8.1 Trotter Formula

One of the simplest decompositions of an exponential operator $e^{x(A+B)}$ is the Trotter (1959) formula

$$e^{x(A+B)} = (e^{(x/n)(A+B)})^n = \lim_{n\to\infty}(e^{(x/n)A}e^{(x/n)B})^n. \tag{95}$$

The convergence of the limit (95) is easily given in a Banach (namely normed) space, as will be shown later. The Trotter formula (95) is easily extended to the following formula:

$$e^{x(A_1+A_2+\ldots+A_q)} = \lim_{n\to\infty}(e^{(x/n)A_1}e^{(x/n)A_2}\ldots e^{(x/n)A_q})^n. \tag{96}$$

These formulas will be used conveniently in quantum Monte Carlo simulations (Suzuki *et al.*, 1977, 1987; Suzuki, 1993a).

8.2 Quantum Monte Carlo Method

It is extremely difficult to calculate exactly the free energy of the relevant quantum system, because the Hamiltonian $\mathcal{H}$ cannot be diagonalized in most realistic cases. Thus, we need some approximation that can be manipulated explicitly. This is realized by the following approximate mapping.

The partition function $Z(\beta)$ of the relevant system described by the Hamiltonian $\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2 + \cdots + \mathcal{H}_q$ is expressed as

$$\begin{aligned} Z &= \mathrm{Tr}\exp[-\beta(H_1+H_2+\cdots+H_q)\} \\ &= \lim_{n\to\infty}\mathrm{Tr}\left\{\exp\left(-\frac{\beta}{n}H_1\right)\exp\left(-\frac{\beta}{n}H_2\right)\right. \\ &\quad \left.\times\cdots\exp\left(-\frac{\beta}{n}H_q\right)\right\}^n \\ &= \lim_{n\to\infty}\sum_{\alpha_1}\cdots\sum_{\alpha_{nq}}\langle\alpha_1|\rho_1|\alpha_2\rangle\langle\alpha_2|\rho_2|\alpha_3\rangle \\ &\quad \times\cdots\langle\alpha_q|\rho_q|\alpha_{q+1}\rangle\langle\alpha_{q+1}|\rho_1|\alpha_{q+2}\rangle \\ &\quad \times\cdots\langle\alpha_{nq}|\rho_q|\alpha_1\rangle \\ &\equiv \mathrm{Tr}^{(d+1)}\exp\mathcal{H}_{\mathrm{eff}}^{(d+1)};\quad \rho_j \equiv \exp(-\beta H_j/n), \end{aligned} \tag{97}$$

where $\{|\alpha_j\rangle\}$ denote orthogonal complete sets. Here, $\mathscr{H}_{\text{eff}}^{(d+1)}$ denotes a $(d + 1)$-dimensional effective Hamiltonian defined by

$$\begin{aligned}\mathscr{H}_{\text{eff}}^{(d+1)} &= H_1(\alpha_1,\alpha_2) + H_2(\alpha_2,\alpha_3) \\ &\quad + \cdots + H_q(\alpha_q,\alpha_{q+1}) + H_1(\alpha_{q+1},\alpha_{q+2}) \\ &\quad + H_2(\alpha_{q+2},\alpha_{q+3}) + \cdots + H_q(\alpha_{2q},\alpha_{2q+1}) \\ &\quad \cdots\cdots\cdots\cdots \\ &\quad + H_1(\alpha_{(n-1)q+1},\alpha_{(n-1)q+2}) \\ &\quad + \cdots + H_q(\alpha_{nq},\alpha_1), \end{aligned} \tag{98}$$

with the partial Hamiltonians $H_j(\alpha,\alpha')$ defined by

$$\langle\alpha_q|\exp(-\beta H_j/n)|\alpha_{q+1}\rangle = \exp H_j(\alpha_q,\alpha_{q+1}). \tag{99}$$

Thus, a d-dimensional quantum system can be transformed to the corresponding $(d + 1)$-dimensional quantum system (equivalence theorem; Suzuki, 1976). This transformation is often called the Suzuki–Trotter transformation. Now, we can apply classical Monte Carlo methods such as the Metropolis algorithm to the transformed classical system. Of course, the effective Hamiltonian contains a new type of many-body interactions shown in Eq. (99), and consequently the corresponding algorithms are much more complicated than the ordinary Metropolis algorithm is.

8.3 Explicit Formulations of the Quantum Monte Carlo Method in Spin Systems

It is instructive to formulate the above quantum Monte Carlo method in the one-dimensional transverse Ising model described by the Hamiltonian $\mathscr{H} = \mathscr{H}_1 + \mathscr{H}_2$, where

$$\mathscr{H}_1 = -J\sum_{j=1}^{M}\sigma_j^z\sigma_{j+1}^z \quad\text{and}\quad \mathscr{H}_2 = -\Gamma\sum_{j=1}^{M}\sigma_j^x. \tag{100}$$

Here, the operators σ_j^x and σ_j^z denote the Pauli operators

$$\sigma_j^x = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} \quad\text{and}\quad \sigma_j^z = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}. \tag{101}$$

The parameters J and Γ denote the spin interaction and the transverse external field, respectively. Clearly, $\mathscr{H}_1$ does not commute with $\mathscr{H}_2$.

Following the general formulation (97), the partition function of this system is expressed as

$$\begin{aligned} Z &= \operatorname{Tr} e^{-\beta H} = \operatorname{Tr}\exp\left(K\sum_{j=1}^{M}\sigma_j^z\sigma_{j+1}^z + \gamma\sum_{j=1}^{M}\sigma_j^x\right) \\ &= \sum_{\{\sigma_{jk}=\pm 1\}} \langle\{\sigma_{j1}\}|e^{A/n}|\{\sigma'_{j1}\}\rangle\langle\{\sigma'_{j1}\}|e^{B/n}|\{\sigma_{j2}\}\rangle \\ &\quad\times\langle\{\sigma_{j2}\}|e^{A/n}|\{\sigma'_{j2}\}\rangle\langle\{\sigma'_{j2}\}|e^{B/n}|\{\sigma_{j3}\}\rangle \\ &\quad\times\cdots\cdots\cdots\cdots \\ &\quad\times\langle\{\sigma_{jn}\}|e^{A/n}|\{\sigma'_{jn}\}\rangle\langle\{\sigma'_{jn}\}|e^{B/n}|\{\sigma_{j1}\}\rangle, \end{aligned} \tag{102}$$

where $K \equiv \beta J$, $\gamma \equiv \beta\Gamma$, and

$$A \equiv K\sum_{j=1}^{M}\sigma_j^z\sigma_{j+1}^z, \quad B \equiv \gamma\sum_{j=1}^{M}\sigma_j^x. \tag{103}$$

Here $|\{\sigma_{jk}\}\rangle$ denote n sets of the product space $|\{\sigma_{jk}\}\rangle = |\sigma_{1k}\rangle\otimes|\sigma_{2k}\rangle\otimes\cdots\otimes|\sigma_{Mk}\rangle$ with the eigenvectors defined by

$$\sigma_j^z|\{\sigma_{jk}\}\rangle = \sigma_{jk}|\{\sigma_{jk}\}\rangle \quad (k = 1, 2, \ldots, n). \tag{104}$$

More explicitly we have

$$\begin{aligned}\langle\{\sigma_{jk}\}|e^{A/n}|\{\sigma'_{jk}\}\rangle &= \exp\left(\frac{K}{n}\sum_{j=1}^{M}\sigma_{jk}\sigma_{j+1,k}\right) \\ &\quad\times\prod_{j=1}^{M}\delta(\sigma_{jk},\sigma'_{jk}), \end{aligned} \tag{105}$$

and

$$\langle\{\sigma_{jk}\}|e^{B/n}|\{\sigma_{j,k+1}\}\rangle = a_n^M\exp\left(\gamma_n\sum_{j=1}^{M}\sigma_{jk}\sigma_{j,k+1}\right), \tag{106}$$

where

$$a_n \equiv \left\{\tfrac{1}{2}\sinh\left(\frac{2\gamma}{n}\right)\right\}^{1/2} \quad\text{and}\quad \gamma_n \equiv \tfrac{1}{2}\ln\coth\left(\frac{\gamma}{n}\right). \tag{107}$$

Thus, we arrive finally at the partition function of the two-dimensional Ising model:

$$\begin{aligned} Z &= \lim_{n\to\infty} a_n^{Mn}\sum_{\{\sigma_{jk}=\pm 1\}} \\ &\quad\times\exp\left\{\sum_{j=1}^{M}\sum_{k=1}^{n}\left(\frac{K}{n}\sigma_{jk}\sigma_{j+1,k} + \gamma_n\sigma_{jk}\sigma_{j,k+1}\right)\right\}. \end{aligned} \tag{108}$$

This is shown in Fig. 3.

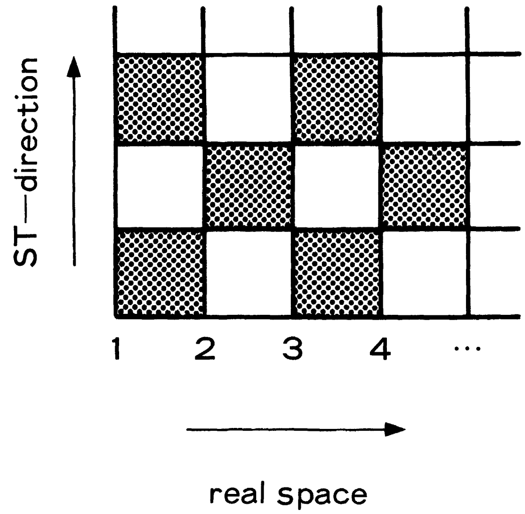

FIG. 3. An example of the two-dimensional Ising model equivalent to the one-dimensional quantum Heisenberg model (figure from Suzuki, 1976). The shaded plaquettes denote a new four-body interaction appearing through the Suzuki–Trotter transformation.

This formulation can be easily extended to other quantum spin systems, and there have been published many articles on explicit applications of the above quantum Monte Carlo method and as well even to Fermi-particle and Bose-particle systems. For example, the magnetic moment $\langle S_i^z \rangle$ at site i of a spin $S = 1$ Heisenberg antiferromagnetic chain has been obtained using the above quantum Monte Carlo method, as shown in Fig. 4.

8.4 General Decomposition Theory of Exponential Operators

The above quantum Monte Carlo method is based on the exponential product formula (96), and consequently the accuracy of simulations for a finite Trotter number n depends on the degree of approximation for decomposition.

For example, we have

$$e^{x(A+B)} = (e^{(x/n)A}e^{(x/n)B})^n + O\left(\frac{x^2}{n}\right). \tag{109}$$

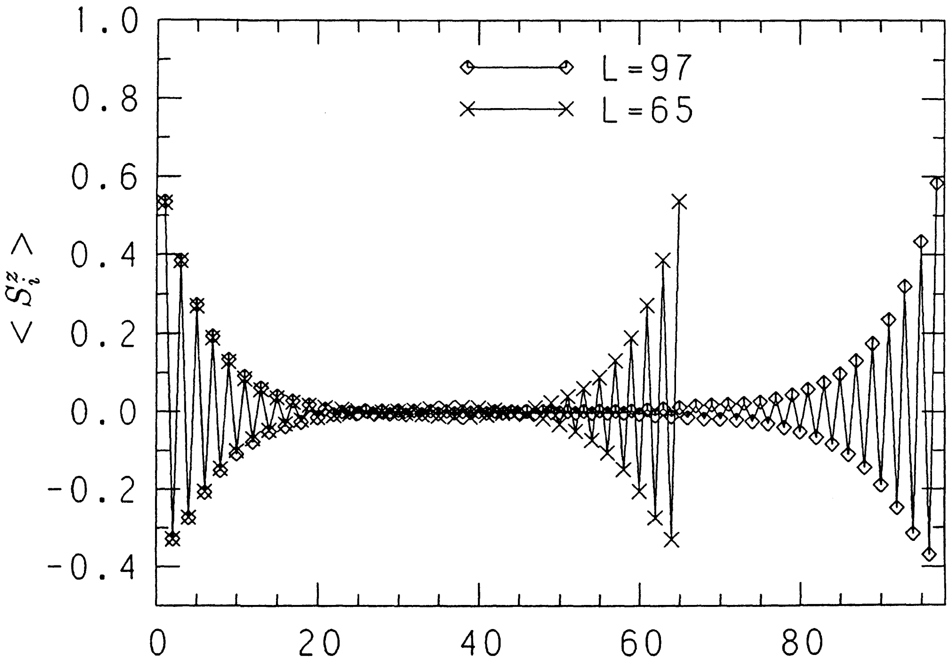

FIG. 4. The spin averages $\langle S_i^z \rangle$ at site i of the $S = 1$ Heisenberg antiferromagnetic chain for the length $L = 65$ and for $L = 97$. The abscissa denotes the position i of the relevant spin S_i^z. (Figure from Miyashita and Yamamoto, 1993.)

The symmetrized product gives a better approximation as

$$e^{x(A+B)} = (e^{(x/2n)A}e^{(x/n)B}e^{(x/2n)A})^n + O(x^3/n^2). \tag{110}$$

If we find a higher decomposition formula

$$e^{x(A+B)} = F_m(x) + O(x^{m+1}) \tag{111}$$

with the form

$$F_m(x) = e^{t_1A}e^{t_2B}e^{t_3A}e^{t_4B} \cdots e^{t_MA}, \tag{112}$$

using some appropriate parameters $\{t_j\}$ proportional to x, then we obtain the following generalized Trotter formula (called the Suzuki–Trotter formula):

$$e^{x(A+B)} = [F_m(x/n)]^n + O(x^{m+1}/n^m). \tag{113}$$

Then, we show how to construct a general higher-order decomposition formula $F_m(x)$ explicitly. One of the most convenient procedures is to make use of the following recursive scheme (Suzuki, 1990): If the mth-order decomposition $Q_m(x)$ is known, then the $(m+1)$th-order decomposition $Q_{m+1}(x)$ is constructed recursively as

$$Q_{m+1}(x) = Q_m(p_1x)Q_m(p_2x) \cdots Q_m(p_sx) \tag{114}$$

with some parameters $\{p_j\}$ $(j = 1, 2, \ldots, 3)$ satisfying the conditions that

$$p_1 + p_2 + \cdots + p_s = 1, \quad \text{and} \quad p_1^{m+1} + p_2^{m+1} + \cdots + p_s^{m+1} = 0. \tag{115}$$

For example, if we start from a second-order symmetric decomposition

$$S_2(x) = e^{(x/2)A}e^{xB}e^{(x/2)A} = S_2^{-1}(-x), \tag{116}$$

then the fourth-order decomposition $S_4(x)$ is obtained as

$$S_4(x) = e^{(p/2)xA}e^{pxB}e^{[(1-p)/2]xA}e^{(1-2p)xB} \times e^{[(1-p)/2]xA}e^{pxB}e^{(p/2)xA}. \tag{117}$$

Here, the parameter p is given by the solution of the equation

$$p^3 + (1-2p)^3 + p^3 = 0; \tag{118}$$

namely,

$$p = \frac{1}{2 - 2^{1/3}}. \tag{119}$$

It should be remarked here that if $S_{2m-1}(x)$ is symmetric, then it is correct up to the $2m$th order, namely,

$$S_{2m-1}(x) = S_{2m}(x). \tag{120}$$

In general, the $2m$th-order decomposition $S_{2m}(x)$ is constructed recursively as

$$S_{2m}(x) = S_{2m-2}(p_mx)S_{2m-2}((1-2p_m)x) \times S_{2m-2}(p_mx) \tag{121}$$

with the parameter $p_m = (2 - 2^{1/m})^{-1}$.

There are many other formulations of exponential products and explicit applications to quantum mechanics, quantum statistical mechanics, laser physics, nonlinear dynamics, and other nonlinear problems.

8.5 Superoperator Formula of Ordered Exponentials

In quantum mechanics and statistical mechanics, the following ordered exponential is frequently used:

$$\begin{aligned} \mathrm{P}\left(\exp \int_0^t \mathcal{H}(s)ds\right) &= 1 + \int_0^t \mathcal{H}(s)ds + \int_0^t dt_1 \int_0^{t_1} dt_2 \mathcal{H}(t_1)\mathcal{H}(t_2) \\ &\quad + \cdots + \int_0^t dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{n-1}} dt_n \mathcal{H}(t_1)\mathcal{H}(t_2) \\ &\quad \times \cdots \mathcal{H}(t_n) + \cdots. \end{aligned} \tag{122}$$

When $\mathcal{H}(t) = \mathcal{H}_1(t) + \mathcal{H}_2(t)$, we can decompose the above-ordered exponential, using the following superoperator $\mathcal{T}$:

$$F(t)e^{s\mathcal{T}}G(t) = F(t+s)G(t) \tag{123}$$

for any operators $F(t)$ and $G(t)$. Then, the following formula holds:

Any ordered exponential can be expressed by an ordinary exponential operator in terms of the super-operator $\mathcal{T}$ as

$$\mathrm{P}\left(\exp \int_t^{t+s} \mathcal{H}(u)du\right) = \exp[s(\mathcal{H}(t) + \mathcal{T})]. \tag{124}$$

Thus, the decomposition of ordered exponentials is reduced to that of ordinary exponential operators.

For example, the first-order and second-order decompositions of the ordered exponential are given by

$$\begin{aligned}\mathrm{P}\left(\exp \int_t^{t+\Delta t} (\mathcal{H}_1(u) + \mathcal{H}_2(u))du\right) &= e^{\Delta t\mathcal{H}_1(t+\Delta t/2)}e^{\Delta t\mathcal{H}_2(t+\Delta t/2)} + O(\Delta t)^2 \\ &= e^{(\Delta t/2)\mathcal{H}_1(t+\Delta t/2)}e^{\Delta t\mathcal{H}_2(t+\Delta t/2)}e^{(\Delta t/2)\mathcal{H}_1(t+\Delta t/2)} \\ &\quad + O(\Delta t^3),\end{aligned} \tag{125}$$

respectively. Any higher-order decomposition of ordered exponentials can be obtained easily from the decomposition formulas of ordinary exponential operators (Suzuki, 1993b).

9. THEORY OF PHASE TRANSITIONS IN QUANTUM SYSTEMS

To study phase transitions theoretically is one of the most important subjects in statistical mechanics. Concerning conceptual arguments of phase transitions, see the article STATISTICAL MECHANICS, CLASSICAL.

There are two categories of phase transitions in quantum systems. Phase transitions in the first category are called quantum phase transitions, which occur as a result of quantum effects. Typical examples of such quantum phase transitions are superconductivity and superfluidity (or the Bose–Einstein condensation). Phase transitions in the second category are such transitions as occur even in classical systems. The gas–liquid phase transition belongs to the second category.

9.1 Landau's Phenomenological Theory of Phase Transitions

The concept of symmetry is crucial in the theory of phase transitions. Landau (1937) introduced an order parameter, say M, for specifying the symmetry of the relevant system. The phase transition is characterized by symmetry breaking. The Helmholtz free energy F is expanded as

$$F = F_0 + AM^2 + BM^4 + \cdots - HM, \tag{126}$$

because the relevant system is assumed to be symmetric as shown in Fig. 5. The equation of state is obtained by setting the derivative of F equal to zero:

$$2AM + 4BM^3 + \cdots = H. \tag{127}$$

As is easily seen from Fig. 5, the coefficient A may assume the following temperature dependence, namely, $A \simeq a(T - T_c)$ near the critical point T_c, where $a > 0$. Then, the susceptibility $\chi_0(T)$ defined by $\chi_0(T) = \lim_{H\to 0}(M/H)$ shows the following Curie–Weiss law:

$$\chi_0(T) = \frac{1}{2A} \simeq \frac{1}{2a(T - T_c)}. \tag{128}$$

The spontaneous order parameter M_s is given by the solution of the equation $2AM_s + 4BM_s^3 = 0$; namely,

$$M_s = \pm\left(\frac{-A(T)}{2B(T)}\right)^{1/2} \simeq \pm\left(\frac{a(T_c - T)}{2B(T_c)}\right)^{1/2} \tag{129}$$

for $T \leq T_c$, using $B(T_c) > 0$.

The above Landau theory is valid even in

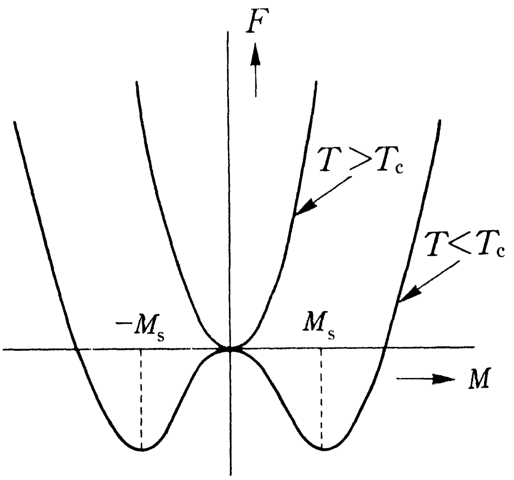

FIG. 5. Landau's free energy as a function of the order parameter M.

quantum phase transitions. His theory has historically played a leading role in finding the mechanism of a new type of phase transitions such as superconductivity, though it is essentially equivalent to the mean-field theory first proposed by P. Weiss (1907).

9.2 Order Parameters in Quantum Phase Transitions

The Bose–Einstein condensation explained in Sec. 5.2 is characterized by the order parameter $b = b_{k=0}$ for the Bose annihilation operator b_k with the momentum k. The eigenstate of the order parameter b is given by the coherent state

$$|\alpha\rangle_c = \mathcal{N}\exp(\alpha b^\dagger)|0\rangle \tag{130}$$

with the normalization factor $\mathcal{N}$. That is, we have

$$b|\alpha\rangle_c = \alpha|\alpha\rangle_c. \tag{131}$$

In Eq. (130), the state $|0\rangle$ denotes the vacuum of the Bose system; namely, $b|0\rangle = 0$, and $b^\dagger b|n\rangle = n|n\rangle$. Thus, the order parameter M of the Bose–Einstein condensation is defined by the average of the annihilation operator b in the coherent state $|\alpha\rangle_c$ as

$$M = {}_c\langle\alpha|b|\alpha\rangle_c = \alpha. \tag{132}$$

Another example of quantum phase transitions is superconductivity, whose order parameter is given by the number of Cooper pairs (Bardeen *et al.*, 1957); namely,

$$\Psi \equiv \sum_k \langle a_{k\uparrow}^\dagger a_{-k\downarrow}^\dagger\rangle \tag{133}$$

for the Fermi operators $\{a_{k\sigma}\}$ with spins $\sigma = \uparrow$ and $\downarrow$. The Cooper pair $b^\dagger \equiv \Sigma_k a_{k\uparrow}^\dagger a_{-k\downarrow}^\dagger$ is essentially a Bose operator, and consequently Ψ denotes the order parameter of the condensation of Cooper pairs.

9.3 Weiss Mean-Field Theory

The simplest theory of phase transition is the Weiss mean-field theory, in which an infinite number of degrees of freedom is replaced by the mean field applied at a specific single spin or particle.

For simplicity, we explain here the mean-field theory of magnetic phase transitions. We consider the Heisenberg model, which is described by the Hamiltonian

$$\mathcal{H} = -J\sum_{\langle ij\rangle} \mathbb{S}_i\cdot\mathbb{S}_j - g\mu_B H\sum_{j=1}^{N} S_j^z, \tag{134}$$

where $\mathbf{H}$ is parallel to z. The mean-field theory replaces the above total Hamiltonian by the following effective Hamiltonian:

$$\mathcal{H}_{\text{mf}} = -(Jz\langle S_i^z\rangle + g\mu_B H)S_j^z, \tag{135}$$

where z denotes the number of nearest neighbors. The average $\langle S_i^z\rangle$ is determined by the self-consistency condition that $\langle S_i^z\rangle = \langle S_j^z\rangle$ because of the uniformity of the system.

The average $\langle S_j^z\rangle$ is easily calculated as

$$\langle S_j^z\rangle = SB_S(\beta g\mu_B SH + \beta Jz\langle S_i^z\rangle S), \tag{136}$$

using the Brillouin function (90). The magnetization M is thus given by

$$\begin{aligned} M &= Ng\mu_{\text{B}}\langle S_j^z\rangle = \beta\frac{N(g\mu_{\text{B}})^2S(S+1)}{3-\beta zJS(S+1)}H \\ &= \frac{N(g\mu_{\text{B}})^2S(S+1)}{3k_{\text{B}}(T-T_c)}H;\ T_c = \frac{zJS(S+1)}{3k_{\text{B}}}, \end{aligned} \tag{137}$$

by expanding the Brillouin function $B_S(x)$ with respect to x, as far as the linear term in H is concerned. Thus, the susceptibility $\chi_0(T)$ is shown to satisfy the Curie law

$$\chi_0(T) = \frac{N(g\mu_{\text{B}})^2S(S+1)}{3k_{\text{B}}}\frac{1}{T-T_c} \tag{138}$$

in the Weiss mean-field approximation.

9.4 Generalized Cluster Mean-Field Approximations

For simplicity, we consider here a ferromagnet. In order to take fluctuations into account, we consider a generalized cluster at whose boundary a mean field H_{mf} is applied, as shown in Fig. 6. We assume that the mean field H_{mf} is proportional to the magnetization m_b at the boundary as $H_{\text{mf}} = k_{\text{mf}}m_b$ with the mean-field coefficient k_{mf}. It is expressed as $k_{\text{mf}} = \bar{z}J$ with the interaction strength J and with the number $\bar{z}$ of nearest neighbors outside of the cluster.

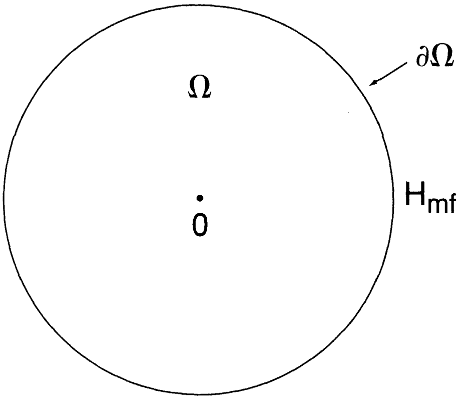

FIG. 6. Generalized cluster Ω with a mean field H_{mf} at the boundary $\partial\Omega$.

Now, the order parameter, namely, the magnetization at the center of the cluster Ω, is expressed in the form

$$m_0 = \chi_L(T)H + \mathscr{F}_L(T)m_b, \tag{139}$$

as far as the terms linear in the external field H and the mean field H_{mf} are concerned. Here L denotes the system size L of the cluster Ω. The cluster susceptibility $\chi_L(T)$ is expressed by the correlation function

$$\chi_L(T) = \beta\mu_{\mathrm{B}}^2\langle S_0^z;S_\Omega^z\rangle, \tag{140}$$

where $S_\Omega^z = \Sigma_{j\in\Omega}S_j^z$ and $\langle A;B\rangle$ denotes the following canonical correlation:

$$\langle A;B\rangle = \frac{1}{\beta}\int_0^\beta \langle AB(i\hbar\lambda)\rangle_\Omega d\lambda. \tag{141}$$

Here $B(z)$ denotes the interaction representation defined by

$$B(z) = e^{it\mathscr{H}_\Omega/\hbar}Be^{-it\mathscr{H}_\Omega/\hbar} \tag{142}$$

with the cluster Hamiltonian $\mathscr{H}_\Omega$ in the absence of a mean field H_{mf}. The average $\langle Q\rangle_\Omega$ denotes the canonical average over the cluster Hamiltonian as

$$\langle Q\rangle_\Omega = \mathrm{Tr}\,Qe^{-\beta\mathscr{H}_\Omega}/\mathrm{Tr}\,e^{-\beta\mathscr{H}_\Omega}. \tag{143}$$

The coefficient $\mathscr{F}_L(T)$ in (139) denotes the feedback function defined by

$$\mathscr{F}_L(T) = \beta k_{\mathrm{mf}}\langle S_0;S_{\partial\Omega}\rangle;\ S_{\partial\Omega} = \sum_{j\in\partial\Omega} S_j^z. \tag{144}$$

Now we impose the self-consistency condition $m_0 \equiv m_b \equiv m$. Then we obtain the magnetization in the form

$$m = \frac{\chi_L(T)}{1 - \mathscr{F}_L(T)}H \equiv \chi_0(T)H. \tag{145}$$

The critical point T_c in this approximation is determined by the solution of the equation $\mathscr{F}_L(T_c) = 1$. Thus, the susceptibility $\chi_0(T)$ is derived to show the following classical anomaly, namely, the Curie-Weiss law:

$$\chi_0(T) \simeq \frac{\bar{\chi}(T_c)}{\epsilon};\ \epsilon = \frac{T - T_c}{T_c}. \tag{146}$$

This is called a "skeletonized susceptibility." The mean-field critical coefficient $\bar{\chi}(T_c)$ is given by

$$\bar{\chi}(T_c) = \chi_L(T_c)/[-T_c\mathscr{F}_L'(T_c)]. \tag{147}$$

This expression is very useful in the coherent-anomaly method (CAM) explained in the succeeding subsection.

There are many other cluster effective-field approximations such as the Bethe-type one (Bethe, 1935) and the double cluster approximation (Oguchi and Kitatani, 1988)

9.5 Coherent-Anomaly Method

As is seen from the preceding subsection, the critical exponent γ of the susceptibility defined in

$$\chi_0(T) \sim 1/(T - T_c^*)^\gamma \tag{148}$$

near the critical point T_c^* is classical; namely, it is given by $\gamma = 1$. However, recent theoretical and experimental results show that $\gamma > 1$. This difference comes from the existence of intrinsic fluctuations in realistic systems.

The main purpose here is to present a general scheme for studying these intrinsic fluctuations and to estimate accurately nonclassical critical exponents by using a systematic series of generalized critical mean-field approximations. The crucial point of the present general scheme, called the coherent-anomaly method (CAM) (Suzuki, 1986), is to note that the mean-field critical coefficient $\bar{\chi}(T_c)$ becomes anomalously large as the approximate critical point T_c approaches the true critical point T_c^*. This is called the coherent anomaly. By "coherent," we mean "systematic." Thus, we have

$$\lim_{T_c \to T_c^*} \bar{\chi}(T_c) = \infty. \tag{149}$$

Then, we may assume the singularity of the mean-field critical coefficient $\bar{\chi}(T_c)$ in the form

$$\bar{\chi}(T_c) \sim 1/(T_c - T_c^*)^{\psi} \tag{150}$$

near the true critical point T_c^*. Then, the critical exponent γ defined in Eq. (148) is shown to be related to the coherent-anomaly exponent ψ estimated through the coherent-anomaly data $(T_c, \bar{\chi})$ in Eq. (150). This situation is explained by the envelope theory shown in Fig. 7.

This CAM theory has been already applied to many classical and quantum systems for evaluating the critical points and critical exponents of such systems (Suzuki *et al.*, 1987). This can be also applied to quantum phase transitions at the ground state (Nonomura and Suzuki, 1992).

9.6 Scaling Law and Renormalization-Group Theory

Critical phenomena are characterized by the scaling law (Fisher, 1964; Widom, 1965), which was justified by the Kadanoff (1966) cell analysis. The scaling law provides the following equation of state:

$$M \simeq t^{\beta} f_{\mathrm{sc}}(H/t^{\Delta}) \tag{151}$$

for the order parameter M and the temperature difference $t = T - T_c^*$. Here, the critical exponent β denotes the singularity of the spontaneous order parameter (namely, the spontaneous magnetization in a ferromagnet) as $M_s \sim (T_c^* - T)^{\beta}$. The exponent Δ is related to β as $\Delta = \beta + \gamma$.

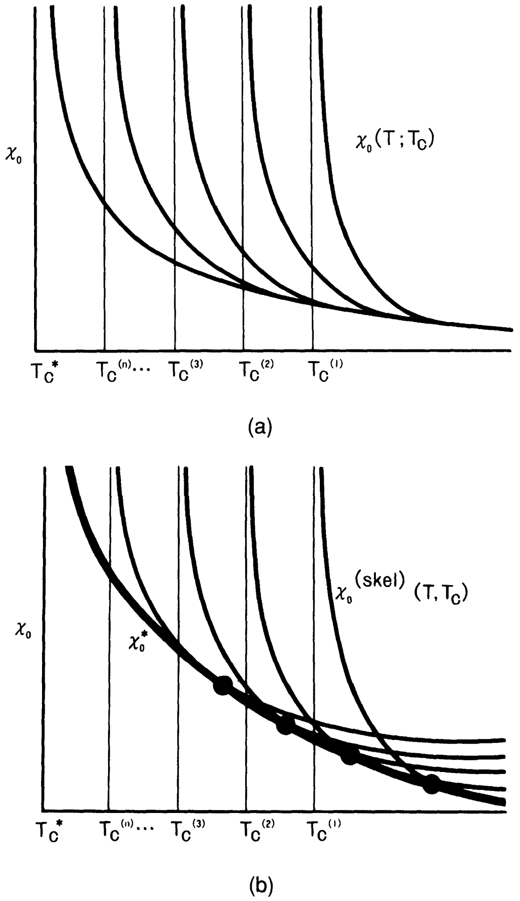

FIG. 7. (a) Susceptibilities obtained in a systematic series of cluster mean-field approximations. (b) Skeletonized susceptibilities of (a) and the envelope χ_0^*. (Figure from Suzuki, 1986.)

The renormalization-group theory (Wilson, 1971) gives microscopic evaluations of critical exponents. In particular, the correlation-length exponent defined in the singularity of the correlation of $\xi \sim t^{-\nu}$ is given by the Wilson formula

$$\nu = (\log b)/(\log \Lambda). \tag{152}$$

Here b denotes the scale factor for the renormalization and Λ the derivative of the renormalized interaction $K' = f_b(K)$ with respect to the original interaction K at the fixed point $K^* = f_b(K^*)$. This renormalization scheme has been used in many systems (see STATISTICAL MECHANICS, CLASSICAL).

For example, in the ferromagnetic Heisenberg model described by the Hamiltonian

$$\mathcal{H} = -J \sum_{\langle ij \rangle} \mathbb{S}_i \cdot \mathbb{S}_j, \tag{153}$$

the renormalized interaction K' is given in the form

$$K' = 2^{d-2} \ln\left(\frac{2e^k + e^{-2k}}{3}\right) \tag{154}$$

for $K = J/2k_BT$ in d dimensions, using the Migdal–Kadanoff approximation (Takano and Suzuki, 1981). The fixed point K^* in three dimensions is calculated easily to give $K^* = 0.68$. However, there exists no fixed point except the trivial one $K^* = 0$ in two dimensions, as is expected.

10. LINEAR RESPONSE THEORY

Up to now, equilibrium statistical mechanics has been presented. In this section, we discuss nonequilibrium statistical mechanics, focusing on Kubo's linear response theory (Kubo, 1957).

Einstein (1905) proposed a theory of Brownian motion, in which a transport coefficient is related to the fluctuation of the corresponding current in equilibrium. For example, the mobility of a particle, μ, is expressed as

$$\mu = \frac{1}{k_BT}\int_0^\infty \langle v(t)v(0)\rangle dt, \tag{155}$$

with $v(t)$ the velocity of the particle. The electric conductivity σ is also given in the form

$$\sigma = \frac{1}{k_BT}\int_0^\infty \langle j(t)j(0)\rangle dt. \tag{156}$$

These expressions are easily derived phenomenologically in the classical approximation.

10.1 Kubo Formula

Now we present here quantum-statistical expressions of transport coefficients. We start from the von Neumann equation

$$i\hbar \frac{\partial}{\partial t} \rho(t) = [\mathcal{H}(t), \rho(t)] \tag{157}$$

for the density matrix $\rho(t)$, where $\mathcal{H}(t)$ is the time-dependent Hamiltonian of the relevant system, and it is assumed to be given in the form

$$\mathcal{H}(t) = \mathcal{H} - AF(t). \tag{158}$$

Here, $F(t)$ denotes an external force and A its conjugate physical quantity operator. The expectation value of any physical quantity B at time t is expressed as

$$\langle B\rangle_t = \mathrm{Tr}\, B\rho(t). \tag{159}$$

We are now interested in a perturbational expansion of $\langle B\rangle_t$ with respect to the external force. To derive it, we study the unitary operator $U(t_1,t_0)$ defined by

$$\rho(t) = U(t,t_0)\rho(t_0)U^\dagger(t,t_0). \tag{160}$$

It satisfies the equation

$$i\hbar \frac{\partial}{\partial t} M(t,t_0) = \mathcal{H}(t)U(t,t_0) \tag{161}$$

with the initial condition $U(t_0,t_0) = 1$. This is also expressed as the product of the following two unitary operators:

$$U(t,t_0) = U_0(t,t_0)\, U_1(t), \tag{162}$$

where $U_0(t,t_0)$ is the unperturbed unitary operator defined by

$$U_0(t,t_0) = \exp[(1/i\hbar)(t - t_0)\mathcal{H}]. \tag{163}$$

Then, the perturbed unitary operator $U_1(t)$ satisfies the following equation:

$$i\hbar \frac{d}{dt} U_1(t) = \mathcal{H}'(t)U_1(t), \tag{164}$$

where $\mathcal{H}'(t)$ denotes the interaction representation

$$\mathcal{H}'(t) = U_0^\dagger(t,t_0)[-AF(t)]U_0(t,t_0). \tag{165}$$

Thus, the expectation value $\langle B\rangle_t$ is expressed as

$$\langle B\rangle_t = \mathrm{Tr}\, \rho(t_0)B - \frac{i}{\hbar}\int_{t_0}^t \mathrm{Tr}\, \rho(t_0)[A, B(t - s)]F(s)ds. \tag{166}$$

Here, we have used the condition that $[\rho(t_0),U_0(t,t_0)] = 0$ and the definition

$$B(t - s) = U_0^{\dagger}(t,s)BU_0(t,s). \tag{167}$$

If we put $\rho(t_0) = \rho_{\text{eq}} \propto \exp(-\beta\mathscr{H})$, the above condition that $[\rho(t_0),U_0(t,t_0)] = 0$ is satisfied. Thus, the first-order deviation or response $\Delta\langle B\rangle_t$ is expressed in the form

$$\Delta\langle B\rangle_t = \text{Re}[\chi_{BA}(\omega)Fe^{i\omega t}], \tag{168}$$

when the external force is given by $F(t) = F \times \cos\omega t$. Here, Re denotes the real part of the relevant quantity, and the complex admittance $\chi_{BA}(\omega)$ is given by

$$\chi_{BA}(\omega) = -\frac{i}{\hbar}\int_{-\infty}^{t} \langle[A, B(t - s)]\rangle_{\text{eq}}e^{i\omega s}ds \tag{169}$$

with the definition (167). When Eq. (169) is divergent, we add the adiabatic factor $e^{\epsilon t}$ as

$$\chi_{BA}(\omega) = -\lim_{\epsilon\to+0}\frac{i}{\hbar}\int_0^{\infty} \langle[A,B(t)]\rangle_{\text{eq}}e^{-i\omega t-\epsilon t}dt. \tag{170}$$

Using Kubo's identity

$$[A,e^{-\beta\mathscr{H}}] = \frac{\hbar}{i}e^{-\beta\mathscr{H}}\int_0^{\beta} \dot{A}(-i\hbar\lambda)d\lambda, \tag{171}$$

we arrive at the expression

$$\chi_{BA}(\omega) = \lim_{\epsilon\to+0}\int_0^{\infty} dte^{-i\omega t-\epsilon t}\int_0^{\beta} d\lambda\langle\dot{A}(-i\hbar\lambda)B(t)\rangle_{\text{eq}}. \tag{172}$$

Now we define a current operator J by $J = \dot{A}$ and we observe the current, namely $B = J$. Then, we may put

$$\langle J\rangle_t = \sigma(\omega)Fe^{i\omega t}. \tag{173}$$

Here, the transport coefficient $\sigma(\omega)$ is expressed as

$$\sigma(\omega) = \int_0^{\infty} dte^{-i\omega t}\int_0^{\beta} d\lambda\langle J(-i\hbar\lambda)J(t)\rangle_{\text{eq}}. \tag{174}$$

If we use the canonical correlation

$$\langle J;J(t)\rangle = \frac{1}{\beta}\int_0^{\beta} \langle J(-i\hbar\lambda)J(t)\rangle_{\text{eq}}d\lambda, \tag{175}$$

then $\sigma(\omega)$ is expressed as

$$\sigma(\omega) = \frac{1}{k_{\text{B}}T}\int_0^{\infty} \langle J;J(t)\rangle e^{-i\omega t}dt. \tag{176}$$

In particular, the static coefficient $\sigma(0)$ is given in the form

$$\sigma(0) = \frac{1}{k_{\text{B}}T}\int_0^{\infty} \langle J;J(t)\rangle dt. \tag{177}$$

In the classical limit $\beta \to 0$, we obtain the result $\langle J;J(t)\rangle \to \langle JJ(t)\rangle_{\text{eq}}$, and consequently we rederive the classical expression (156).

10.2 Fluctuation–Dissipation Theorem and Kubo–Martin–Schwinger Condition

The relaxation function $\Phi_{BA}(t)$ is defined by

$$\Phi_{BA}(t) = \lim_{\epsilon\to+0}\int_t^{\infty} \phi_{BA}(s)e^{-\epsilon s}ds \tag{178}$$

using the response function

$$\phi_{BA}(t) = -(i/\hbar)\langle[A,B(t)]\rangle_{\text{eq}}. \tag{179}$$

The Fourier transform $f_{BA}(\omega)$ of $\Phi_{BA}(t)$ is introduced as

$$\Phi_{BA}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f_{BA}(\omega)e^{i\omega t}dw. \tag{180}$$

Next, we introduce the symmetrized time correlation $\Psi_{BA}(t)$ as

$$\Psi_{BA}(t - t') = \tfrac{1}{2}\langle(\Delta A(t)\Delta B(t') + \Delta B(t')\Delta A(t))\rangle_{\text{eq}} \tag{181}$$

with $\Delta A(t) = A(t) - \langle A\rangle_{eq}$. The Fourier form $g_{BA}(\omega)$ of $\Psi_{BA}(t)$ is related to $f_{BA}(\omega)$ as

$$g_{BA}(\omega) = E_{\beta}(\omega)f_{BA}(\omega) \tag{182}$$

with the energy $E_{\beta}(\omega)$ of a harmonic oscillator in equilibrium at temperature T, namely with

$$E_\beta(\omega) = \tfrac{1}{2}\hbar\omega \coth(\tfrac{1}{2}\beta\hbar\omega). \qquad (183)$$

Relation (182) is called the fluctuation–dissipation theorem. This is very important conceptually in nonequilibrium statistical mechanics.

The following relation is also easily derived (Martin and Schwinger, 1959):

$$\langle \Delta A \Delta B(t)\rangle_\omega = e^{\beta\hbar\omega}\langle \Delta B(t)\Delta A\rangle_\omega \qquad (184)$$

concerning the Fourier transform $\langle Q(t)\rangle_\omega$ of a time-dependent operator $Q(t)$.

10.3 Onsager's Reciprocity Theorem

When we consider several external forces $\{F_j(t)\}$ and their conjugate physical quantities $\{A_j\}$, we can define the following response functions $\{\Phi_{ij}(t)\}$:

$$\Phi_{ij}(t) = \beta\langle A_j; A_i(t)\rangle. \qquad (185)$$

If the relevant system (namely, the corresponding Hamiltonian $\mathcal{H}$) is symmetric with respect to time inversion, then we have the Onsager (1931) reciprocity relation

$$\Phi_{ij}(t) = \Phi_{ji}(t). \qquad (186)$$

That is, we have

$$\chi_{ij}(\omega) = \chi_{ji}(\omega). \qquad (187)$$

These symmetry relations are practically very useful, for example, in chemical engineering.

10.4 Exponential Perturbation Theory

If we put $\rho(t) = e^{\eta(t)}$, then the "entropy" operator $\eta(t)$ satisfies the same von Neumann equation as ρ:

$$i\hbar\frac{\partial}{\partial t}\eta(t) = [\mathcal{H}(t), \eta(t)]. \qquad (188)$$

To solve this equation, we put

$$\eta(t) = -\Phi(t) - \beta\mathcal{H} + \Delta\eta(t). \qquad (189)$$

Using the time-dependent Hamiltonian $\mathcal{H}(t) = \mathcal{H} - AF(t)$, we obtain the equation

$$i\hbar\frac{\partial}{\partial t}\Delta\eta(t) = [\mathcal{H} - AF(t), \Delta\eta(t)] + \beta F(t)\dot{A}. \qquad (190)$$

The operator $\Delta\eta(t)$ may be expanded with respect to $F(t)$ as

$$\Delta\eta(t) = \Delta\eta_1(t) + \Delta\eta_2(t) + \cdots, \qquad (191)$$

where

$$\Delta\eta_1(t) = \beta\int_{-\infty}^{0} F(t+s)\dot{A}(s)ds, \qquad (192)$$

$$\Delta\eta_2(t) = \frac{\beta}{i\hbar}\int_{-\infty}^{0} dsF(t+s)\int_0^s F(t+s') \times [A(s'), \dot{A}(s)]ds', \qquad (193)$$

and in general we have

$$\Delta\eta_n(t) = \frac{\beta}{(i\hbar)^{n-1}}\int_{-\infty}^{0} dsF(t+s)\int_0^s dt_1\int_0^{t_1} dt_2 \cdots \int_0^{t_{n-2}} dt_{n-1}F(t+t_1)F(t+t_2) \cdots F(t+t_{n-1}) \times [A(t_1),[A(t_2),\cdots[A(t_{n-1}),\dot{A}(s)]\cdots]. \qquad (194)$$

This scheme is called exponential perturbation theory. The first-order expression

$$\rho_1(t) = \exp[-\Phi(t) - \beta\mathcal{H} + \Delta\eta_1(t)] \qquad (195)$$

gives the Zubarev (1971) statistical operator. This is often used for studying nonlinear relaxation phenomena in many physical systems.

GLOSSARY

Bose–Einstein Condensation: The fact that Bose particles occupy the zero-momentum state macroscopically at low temperatures.

Bose Particle: A particle obeying Bose statistics.

Bose Statistics: Quantum statistics coming from the even symmetry of the many-particle wave function for integer-spin particles, in which any number of particles can occupy the same state simultaneously.

Density Matrix: A statistical operator expressing a statistical-mechanical ensemble such as the canonical or grand canonical ensembles, and having both the probabilistic character of quantum mechanics and statistics of the ensemble theory.

Exponential Product Formula: A formula expressing an exponential operator such as $\exp[x(A + B)]$ for the noncommutating operators A and B as the product of the exponential operators $\exp(xA)$ and $\exp(xB)$.

Fermi Particle: A particle obeying Fermi statistics.

Fermi Statistics: Quantum statistics coming from the odd symmetry of the many-particle wave function for half-integer-spin particles, in which only single particle can occupy each state.

Fluctuation–Dissipation Theorem: A theorem expressing that dissipation is related to fluctuation in equilibrium.

Kubo Formula: A general formula expressing transport coefficients in terms of time-correlation functions in equilibrium.

Kubo–Matrix–Schwinger Condition: A condition regulating an equilibrium state in terms of time-correlation functions.

Linear Response Theory: A theory expressing responses with respect to an external force in its first order, namely, fluctuations in the absence of the external force.

Mean-Field Theory: A theory expressing approximately a system of infinite degrees of freedom in terms of finite degrees of freedom by introducing some effective fields.

Ordered Exponential: An exponential operator prescribed by the Dyson time-ordering operator.

Quantum Monte Carlo Method: A Monte Carlo method for quantum systems based on the Suzuki–Trotter transformation which transforms a d-dimensional quantum system with short-range interactions to a $(d + 1)$-dimensional classical system with short-range interactions.

Quasiparticle: A particle deformed in mass or momentum dependence by the interaction with other particles.

Thermal de Broglie Wavelength: A characteristic length λ_D of a Fermi particle in thermal equilibrium determined by the condition $\lambda_D = h/p$ with $p^2/2m = 3k_BT/2$.

Trotter Formula: A product formula on an exponential operator, namely, $\exp[(A + B)] = \lim_{n\to\infty}[\exp(A/n)\exp(B/n)]^n$

von Neumann Equation: A basic formula that the time-dependent density matrix satisfies.

Works Cited

Bardeen, J., Cooper, L. N., Schrieffer, J. R. (1957), *Phys. Rev.* **108,** 1175.

Bethe, H. A. (1935), *Proc. R. Soc. London A* **150,** 552.

Einstein, A. (1905), *Ann. Phys. (Leipzig)* **17,** 549.

Einstein, A. (1907), *Ann. Phys. (Leipzig)* **19,** 371.

Fisher, M. E. (1964), *J. Math. Phys.* **5,** 944.

Kadanoff, L. P. (1966), *Physics (LI City)* **2,** 263.

Kadanoff, L. P. (1975), *Phys. Rev. Lett.* **34,** 1005.

Kubo, R. (1957), *J. Phys. Soc. Jpn.* **12,** 570.

Landau, L. D. (1937), *Phys. Z. Sowjetunion* **11,** 26, 545.

Martin, P., Schwinger, J. (1959), *Phys. Rev.* **115,** 1342.

Migdal, A. A. (1975), *Zh. Eksp. Teor. Fiz.* **69,** 810.

Miyashita, S., Yamamoto, S. (1993), *Phys. Rev. B* **48,** 913.

Nonomura, Y., Suzuki, M. (1992), *J. Phys. A: Math. Gen.* **25,** 85.

Oguchi, T., Kitatani, H. (1988), *J. Phys. Soc. Jpn.* **57,** 3973.

Onsager, L. (1931), *Phys. Rev.* **37,** 405; **38,** 2265.

Suzuki, M. (1976), *Prog. Theor. Phys.* **56,** 1454.

Suzuki, M. (1986), *J. Phys. Soc. Jpn.* **55,** 4205.

Suzuki, M. (Ed.) (1987), *Quantum Monte Carlo Methods in Equilibrium and Nonequilibrium Systems,* Berlin: Springer.

Suzuki, M. (1988), *J. Phys. Soc. Jpn.* **57,** 2310.

Suzuki, M. (1990), *Phys. Lett.* **A146,** 319.

Suzuki, M. (1991), *J. Math. Phys.* **32,** 400.

Suzuki, M. (1992), *Phys. Lett.* **A165,** 387.

Suzuki, M. (Ed.) (1993a), *Quantum Monte Carlo Methods in Condensed Matter Physics,* Singapore: World Scientific.

Suzuki, M. (1993b), *Proc. Jpn. Acad. Ser. B* **69,** 161.

Suzuki, M. (1994a), *Statistical Mechanics,* Tokyo: Iwanami Shoten.

Suzuki, M. (1996), *Rev. Math. Phys.* **8,** 487.

Suzuki, M., Miyashita, S., Kuroda, A. (1977), *Prog. Theor. Phys.* **58,** 1377.

Suzuki, M., Katori, M., Hu, X. (1987), *J. Phys. Soc. Jpn.* **56,** 3092.

Takano, H., Suzuki, M. (1981), *J. Stat. Phys.* **26,** 635.

Trotter, H. F. (1959), *Proc. Am. Math. Soc.* **10,** 545.

Weiss, P. (1907), *J. Phys. Radium* **6,** 661.

Widom, B. (1965), *J. Chem. Phys.* **41,** 1633.

Wilson, K. G. (1971), *Phys. Rev. B* **4,** 3174, 3184.

Zubarev, D. N. (1971), *Nonequilibrium Statistical Mechanics,* Moscow: Nauka.

Further Reading

General

Feynman, R. P. (1972), *Statistical Mechanics—A Set of Lectures,* New York: Benjamin.

Itzykson, C., Drouffe, J.-M. (1989), *Statistical Field Theory—From Brownian Motion to Renormalization and Lattice Gauge Theory,* Vols. 1 and 2, Cambridge, U.K.: Cambridge Univ. Press.

Landau, L. D., Lifshitz, E. M. (1958), *Statistical Physics,* Oxford: Pergamon.

Mayer, J. E., Mayer, M. G. (1940), *Statistical Mechanics,* New York: Wiley.

Parisi, G. (1988), *Statistical Field Theory,* Reading, MA: Addison-Wesley.

Reichl, L. E. (1980), *A Modern Course in Statistical Physics,* Austin, TX: Univ. of Texas Press.

Reif, F. (1965), *Statistical Physics,* Berkeley Physics Course 5, New York: McGraw-Hill.

ter Haar, D. (1961), *Elements of Statistical Mechanics,* New York: Holt, Rinehart and Winston.

Toda, M., Kubo, R., Saito, N. (1991), *Statistical Physics I, Equilibrium Statistical Mechanics,* Springer Series in Solid State Science No. 30, Berlin: Springer.

Phase Transitions

Amit, D. J. (1974), *Field Theory, the Renormalization Group, and Critical Phenomena,* New York: McGraw-Hill; (1984), 2nd ed., *ibid.*

Baker, G. A., Jr. (1990), *Quantitative Theory of Critical Phenomena,* New York: Academic.

Baxter, R. J. (1982), *Exactly Solved Models in Statistical Mechanics,* New York: Academic.

Burkhardt, J. W., van Leewen, J. M. J. (1982), *Real-Space Renormalization,* Berlin: Springer.

Cardy, J. L. (1988), *Finite-Size Scaling,* Amsterdam: North-Holland.

Christe, P., Henkel, M. (1993), *Introduction to Conformal Invariance and Its Applications to Critical Phenomena,* Berlin: Springer.

Domb, C., Green, M. S. (1972–1976), *Phase Transitions and Critical Phenomena,* Vols. 1–6, New York: Academic.

Domb, C., Lebowitz, J. L. (1983–), *Phase Transitions and Critical Phenomena,* Vols. 7–, New York: Academic.

Fisher, M. E. (1968), *Rep. Prog. Phys.* **30,** 615.

Itzykson, C., Saleur, H., Zuber, J.-B. (Eds.) (1988), *Conformal Invariance and Its Applications to Statistical Mechanics,* Singapore: World Scientific.

Ma, S. K. (1976), *Modern Theory of Critical Phenomena,* Reading, MA: Benjamin.

McCoy, B. M., Wu, T. T. (1973), *The Two-Dimensional Ising Model,* Cambridge, MA: Harvard Univ. Press.

Pfeuty, P., Toulouse, G. (1975), *Introduction to the Renormalization Group and Critical Phenomena,* New York: Wiley.

Privman, V. (Ed.) (1990), *Finite Size Scaling and Numerical Simulation of Statistical Systems,* Singapore: World Scientific.

Stanley, H. E. (1973), *Introduction to Phase Transitions and Critical Phenomena,* Oxford: Oxford Univ. Press.

Suzuki, M., *et al.* (1995), *Coherent Anomaly Method—Mean Field, Fluctuations, and Systematics,* Singapore: World Scientific.

Suzuki, M., Kubo, R. (Eds.) (1991), *Evolutionary Trends in the Physical Sciences,* Springer Proceedings in Physics No. 57, Berlin: Springer.

Uzunov, D. I. (1993), *Introduction to the Theory of Critical Phenomena—Mean-Field, Fluctuations and Renormalization,* Singapore: World Scientific.

Wilson, K. G., Kogut, J. (1974), *Phys. Rep.* **C12,** 75.

Yeomans, J. (1992), *Statistical Mechanics of Phase Transitions,* Oxford: Oxford Univ. Press.

Zinn-Justin, J. (1989), *Quantum Field Theory and Critical Phenomena,* Oxford: Clarendon Press.

Nonequilibrium Statistical Mechanics

Glansdorff, P., Prigogine, I. (1971), *Thermodynamics of Structure, Stability and Fluctuations,* New York: Wiley.

Hida, T., Kuo, H.-H., Potthoff, J., Streit, L. (1993), *White Noise—An Infinite-Dimensional Calculus,* Dordrecht: Kluwer Academic.

Katchalsky, A., Curran, P. (1965), *Nonequilibrium Thermodynamics in Biophysics,* Cambridge, MA: Harvard Univ. Press.

Kubo, R., Toda, M., Hashitsume, N. (1991), *Statistical Physics II, Nonequilibrium Statistical Mechanics,* Springer Series in Solid State Physics No. 31, Berlin: Springer.

Nicolis, G., Prigogine, I. (1977), *Self-Organization in Nonequilibrium Systems,* New York: Wiley.

Nicolis, G., Dewel, G., Turner, J. W. (Eds.) (1981), *Order and Fluctuations in Equilibrium and Nonequilibrium Statistical Mechanics,* New York: Wiley.

Risken, H. (1984), *The Fokker–Planck Equation—Methods of Solution and Applications,* Berlin: Springer.

Suzuki, M. (1981), *Adv. Chem. Phys.* **46,** 195.

van Kampen, N. G. (1981), *Stochastic Processes in Physics and Chemistry,* Amsterdam: North-Holland.

Complex Systems and Other Topics

Chowdhury, D. (1986), *Spin Glasses and Other Frustrated Systems*, Singapore: World Scientific.

Kawasaki, K., Suzuki, M., Onuki, A. (Eds.) (1990, 1993), *Formation, Dynamics and Statistics of Patterns*, Vols. 1 and 2, Singapore: World Scientific.

Lewin, R. (1992), *Complexity*, New York: Macmillan.

Magnus, W., Karras, A., Solitar, D. (1976), *Combinatorial Group Theory*, New York: Dover.

Mezard, M., Parisi, G., Virasoro, M. A. (1987), *Spin Glass Theory and Beyond*, Singapore: World Scientific.

Palmer, R. G. (1983), *Heisenberg Colloquium on Spin Glasses*, Springer Lecture Notes in Physics No. 192, Berlin: Springer.

Prigogine, I. (1980), *From Being to Becoming*, San Francisco: Freeman.

Weiss, U. (1993), *Quantum Dissipative Systems*, Singapore: World Scientific.

STATISTICAL PHYSICS OF MEMBRANES AND LAMELLAR SYSTEMS

David H. Boal, *Department of Physics, Simon Fraser University, Burnaby, British Columbia, Canada*

Introduction 541
1. **Membranes and Networks** 542
1.1 Fluid Membranes 542
1.2 Polymerized Membranes and Networks 544
1.3 Multilamellar Systems 546
2. **Elasticity** 546
2.1 Deformations and the Strain Tensor 546
2.2 Elastic Moduli 547
2.3 Bending Resistance 548
2.4 Elasticity and Fluctuations 549
3. **Flexible Chains** 550
3.1 Ideal Chains 550
3.2 Self-Avoiding Chains 550
3.3 Self-Avoiding Rings 552
3.4 Chain Elasticity 552
4. **Fluctuating Membranes** 553
4.1 Fluid Membranes 553
4.2 Polymerized Membranes 555
4.3 Membranes under Stress 556
4.3.1 Closed Fluid Membranes 556
4.3.2 Open Membranes at a Wall 556
4.3.3 Planar Membranes under Tension 556
5. **Systems with Defects** 557
5.1 Percolation Phenomena 557
5.2 Defective Two-Dimensional Networks 557
5.3 Defective Membranes in Three Dimensions 558
6. **Interactions between Membranes** 558
6.1 Interactions between Rigid Surfaces 558
6.1.1 van der Waals Forces 559
6.1.2 Double-Layer Forces 559
6.1.3 Hydration Forces 559
6.1.4 Steric Interactions 559
6.2 Entropic Repulsion of Fluctuating Membranes 559
6.3 Unbinding Transition 561
Glossary 561
Works Cited 561
Further Reading 562

INTRODUCTION

> "Like gold to ayery thinnesse beate"
> —John Donne, *A Valediction: forbidding mourning*, 1611

In the nineteenth century, Lord Raleigh speculated that oil spread on water could form a layer only one molecule thick. By spreading a known volume of oil on a calm lake and estimating, by observing reflected light, the area that the oil covered, he was able to determine that the layer had a thickness in the range of 10^{-9} m. Today, we know of many examples of thin sheets, networks, and membranes that occur naturally in biological systems or that can be produced by a variety of physical and chemical means. These membranes and networks are very flexible and have small elastic moduli. For example, the volumetric compression resistance of a protein network in a cell may be more than ten million times smaller than that of a typical liquid.

The properties of highly flexible materials are a natural area of application for statistical mechanics. Many interesting characteristics of rubber, such as the observation first made by John Gough in 1805 that natural rubber contracts when heated, arise from the entropic properties of polymeric chains. The chains in such materials are flexible at the molecular level and possess a large number of equal-energy conformations, or equivalently, have substantial conformational entropy. Statistical mechanics explains Gough's observations by demonstrating that a flexible chain behaves like a spring with a spring

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

constant that increases linearly with temperature.

Viewed as two-dimensional generalizations of one-dimensional chains, membranes show a rich behavior on mesoscopic length scales. Computer simulations argue that *fluid* membranes, with vanishing in-plane shear modulus, exhibit highly convoluted configurations on long length scales, whereas *polymerized* membranes, which resist an in-plane shear, are asymptotically flat. Further, the elastic moduli of membranes are predicted to be wavelength dependent. In their interaction with each other, and with rigid objects such as flat walls, membranes unbind at high temperatures because of thermal undulations.

The study of the elastic and geometrical properties of membranes is a rapidly advancing field with an interplay of experiment, analytical theory, and computer simulation. Insights obtained from each of these physical approaches are contained in the following pages, which also include a short review of some concepts from polymer physics for both comparative and illustrative purposes. Space does not permit the presentation here of the application of statistical mechanics to the phase behavior of surfactant/solvent systems. We describe several of these phases in Sec. 1, where their existence is taken as given, but refer the interested reader to the monographs by Israelachvili (1992), Mouritsen (1988), and Safran (1994) for detailed experimental and theoretical treatments.

1. MEMBRANES AND NETWORKS

1.1 Fluid Membranes

Surface-active agents (or *surfactants*) owe their grease-cutting success to their affinity to both polar and nonpolar solvents. Surfactants are referred to as *amphiphilic*, reflecting their common attraction to otherwise immiscible liquids. Viewed at the molecular level, common soap (sodium stearate) has a long hydrocarbon chain that is attractive to nonpolar liquids such as oil and has a polar carboxylate group at the end of the chain that is attractive to polar liquids such as water; these two interactions are referred to as *hydrophobic* and *hydrophilic*, respectively. The molecular structures of two families of amphiphiles, fatty acids and dual-chain lipids, are illustrated in Fig. 1. Most lipids found in biomembranes have two chains with 16–18 carbon atoms per chain and head groups that are geometrically smaller than the hydrocarbon chains. Some phospholipids, for example, have just a three- or four-segment group (e.g., $-CH_2-CH_2-NH_2$) linked to the carboxylate end of the hydrocarbon chains through a phosphate group ($-PO_4-$) [see Gennis (1989) or Alberts *et al.* (1989) for further details].

While amphiphiles play an important role at interfaces in ternary systems such as oil/water/amphiphile, complex phases are found even in binary systems such as water plus amphiphile. Concentration, temperature, and *p*H are among the conditions that influence the phase structure of binary and ternary systems containing amphiphiles. Among other things, what makes these phases of interest biologically is that the amphiphiles self-assemble into structures with mesoscopic dimensions. Some aspects of the mesoscopic geometry have been interpreted in terms of the molecular geometry of the amphiphile, as we now attempt to demonstrate by example [see Israelachvili (1992) for further details].

Consider first the simple system of water and a single-chain lipid such as a fatty acid. At very low concentrations, the amphiphile dissolves in the conventional manner into the water solvent. In spite of the hydrophobic nature of the amphiphile's hydrocarbon

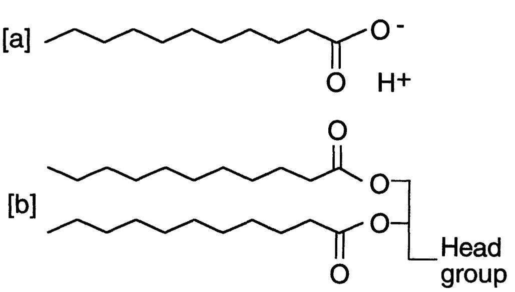

FIG. 1. Molecular structure of two families of amphiphiles. Molecule (a) is a fatty acid, having a single hydrocarbon chain terminating in a carboxylic acid group. Molecule (b) is a dual-chain lipid, with two hydrocarbon chains and a polar head group. In a typical phospholipid, the head group is much smaller than the hydrocarbon chains (e.g., $-PO_4-CH_2-CH_2-NH_2$ for phosphatidyl ethanolamine).

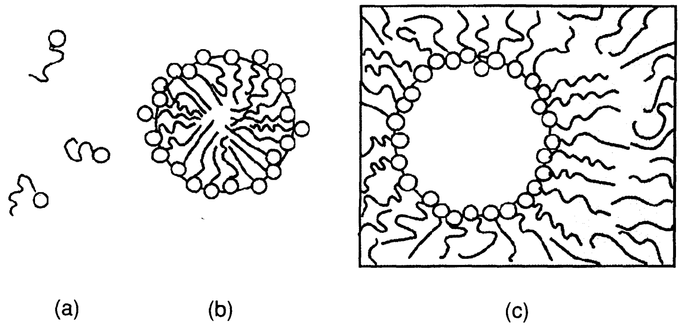

FIG. 2. Sample of phases that may be present in amphiphile/water systems: (a) conventional solution at very low amphiphile concentration, (b) micelle formation for concentrations above the critical micelle concentration (CMC), and (c) inverted micelles at high amphiphile concentrations. Other geometries are possible as well. In the diagram, the polar group of the molecule is indicated by a circle.

chain, the hydrophilic polar group at the head of the chain favors solvation. However, even at very modest concentrations of amphiphile, some fraction of the amphiphilic molecules undergo phase separation into *micelles,* as shown in Fig. 2. The abrupt onset of micelle formation occurs at the *critical micelle concentration* (or CMC). In an aqueous medium, the amphiphiles in a micelle are arranged so that their polar groups are in contact with water, while the nonpolar acyl chains aggregate in the interior of the micelle and are shielded from water. While micelle formation is favored energetically, it does not occur at very low amphiphile concentrations because of the low entropy of mixing compared to a conventional solution. Finally, at large amphiphile concentration, a phase of inverted micelles may form, also as indicated in Fig. 2.

The hydrocarbon chains of the micelle in Fig. 2(b) are packed together rather tightly, surrounded by the polar groups, which form a relatively close-packed array on the micelle surface to minimize the contact between the micelle interior and its aqueous environment. Thus, micelle formation is favored for amphiphiles whose polar head-group area is large compared to the transverse dimension of the hydrocarbon chains. In dual-chain lipids, the chains occupy proportionately more area than the head group does, and other phases, such as the lipid bilayers of Fig. 3, compete with micelle formation. Depending on the molecular geometry of the lipid, the most favored bilayer may be flat or curved. The ability of dual-chain lipids to form bilayers should not be interpreted as meaning that such lipids do not form micelles. In fact, because of the large hydrocarbon areas, the CMC of dual-chain lipids (typically 10^{-8} mol L^{-1}) is orders of magnitude smaller than that of single-chain lipids (typically 10^{-3} mol L^{-1}). Dual-chain lipids also are capable of forming phases of inverted micelles. As discussed further in Sec. 1.2, living cells are bounded by a plasma membrane whose structure includes a fluid lipid bilayer.

It is possible to obtain the elastic properties of biomembranes and related structures by several different techniques. An example of the micropipette aspiration technique is shown in Fig. 4, in which a cell or synthetic vesicle is drawn under pressure into a micropipette with a diameter of about a micron. Elastic moduli can be determined from the strain observed in the membrane when it is subject to a known stress. Another technique for measuring elastic moduli uses the fluctuations in cell geometry (see Sec. 2.4 for a de-

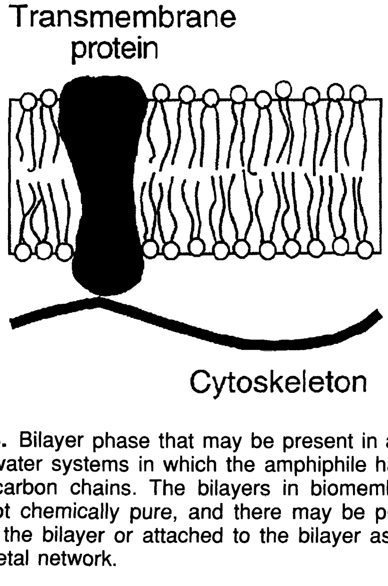

FIG. 3. Bilayer phase that may be present in amphiphile/water systems in which the amphiphile has two hydrocarbon chains. The bilayers in biomembranes are not chemically pure, and there may be proteins within the bilayer or attached to the bilayer as a cytoskeletal network.

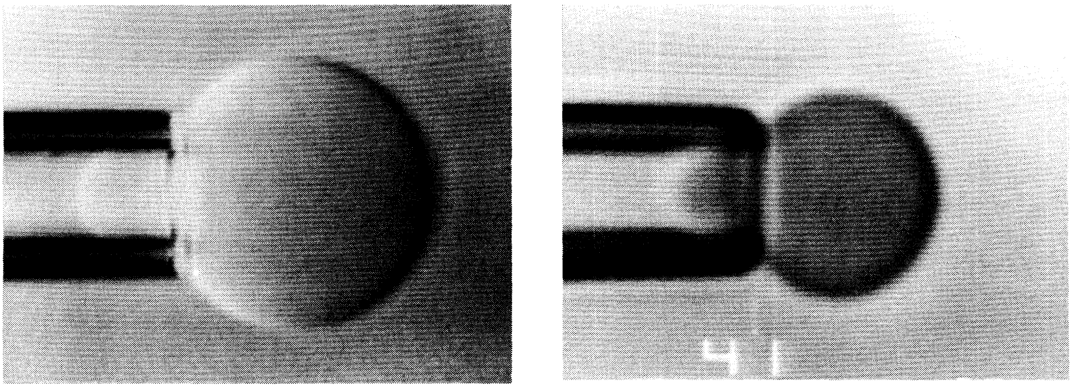

(a) (b)

FIG. 4. Micropipette aspiration of a giant phospholipid bilayer vesicle (left, diameter ~20 μm) and an osmotically preswollen red cell (right, diameter 6 μm). Several of the elastic constants of bilayers can be measured using variations of a technique in which the vesicle is subject to a known stress (from Mohandas and Evans, 1994).

scription of the bulk modulus in terms of volume fluctuations).

The bending moduli (k_b of Sec. 2.3) of phospholipid bilayers of biological importance typically are in the range (3–20) × 10^{-20} J, depending on chemical composition, and is in the same range as that reported for the red blood-cell plasma membrane using different techniques. While small compared to thin inorganic films, this modulus is about an order of magnitude larger than that of surfactant monolayers at an oil–water interface. The two-dimensional area compression modulus of a flat red-cell bilayer is 0.5 J/m^2, which may be more than an order of magnitude smaller than the product of the bilayer thickness and the volume compression modulus of a simple liquid such as water. Since bilayers are fluid, their shear modulus vanishes.

Depending on their geometry and other properties, lipids may form into bilayers that are flat or have a tendency to bend spontaneously (*spontaneous curvature;* see Sec. 2.3). A naturally curved bilayer may form into a closed geometry such as a sphere with no energy penalty. A spherical *vesicle* whose surface energy density depends only on the mean surface curvature has a relatively modest total energy of $8\pi k_b$ independent of the vesicle radius. This means that the deformation energy of the whole vesicle may be just $(250–2000)k_BT$ for the range of moduli quoted above (where k_B is Boltzmann's constant and T is the absolute temperature). The size of the mesoscopic structures adopted by lipid bilayers covers a broad range: Stable vesicular systems that are found in Nature or that can be prepared in the laboratory vary from tens of nanometers to tens of microns.

Ternary mixtures of amphiphiles, oil, and water have very complicated phase diagrams, as one would expect. Micellar phases may be present, of course, and there are structures that can form with amphiphile at the interface between oil and water. Sponge phases also exist, in which oil and water are separated from each other in a labyrinth of three-dimensional channels running throughout the system. In this article, we concentrate on the self-assembled bilayers formed in binary systems.

1.2 Polymerized Membranes and Networks

The lipid bilayers described in Sec. 1.1 are fluid: they do not resist an in-plane shear. However, there are many inorganic monolayers or polymerized organic monolayers and bilayers that can support an in-plane shear (at least on small length scales, see Sec. 4.2). These materials are predicted to have different properties than fluid membranes, so we describe several examples here.

Graphite derives its lubricating properties from its molecular geometry consisting of carbon rings linked together in a planar honeycomb structure. Other examples of planar inorganic monolayers can be obtained by a process known as exfoliation (Divigalpitiya *et al.,* 1989). This is a two-step process in which potentially reactive ions such as lithium are *intercalated* into layered structures such as MoS_2. In a second step, the intercalated material is exposed to water, which reacts strongly with the lithium and forces the monolayers apart. A section of an MoS_2 monolayer is illustrated in Fig. 5. Polymerized organic membranes can be obtained by exposing to ultraviolet radiation monolayers or bilayers of certain reactive organic compounds. The resulting materials are not always uniform, since polymerization may occur in local patches. A variety of materials have shown the ability to polymerize, many examples of which can be found in Ringsdorf (1988).

Two-dimensional networks also are examples of polymerized membranes. These networks need not be "solid" in the chemical sense—they only need to have a two-dimensional structure and fixed connectivity resulting in shear resistance. An example of a polymerized network is the spectrin cytoskeleton of the red blood cell (see Alberts *et al.*, 1989). The erythrocyte plasma membrane consists of three regions: the lipid bilayer forming the cell boundary, the glycocalix exterior to the cell, and the cytoskeleton on the interior or *cytoplasmic* side of the membrane. Tetramers of the spectrin protein (with a contour length of about 200 nm) form a two-dimensional network with largely triangular connectivity at the junction points of the spectrin tetrameres. The network is loose, with a mean separation between the junction complexes of 70 nm. A schematic representation of the cytoskeleton from a computer simulation is shown in Fig. 6, in which view (a) is perpendicular to the bilayer and view (b) is along the bilayer plane. The bilayer lies below the cytoskeleton in view (b). The in-plane shear modulus of this network, as measured by the micropipette experiments, is (6–9) × 10^{-6} J/m^2, although other experiments based on long-wavelength fluctuations have yielded lower values.

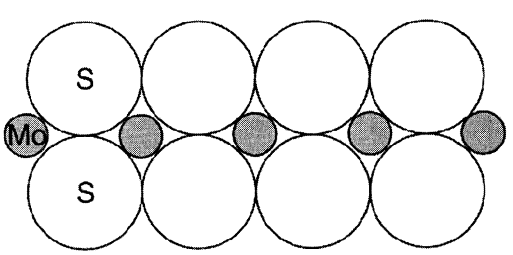

FIG. 5. Schematic drawing of a section through an MoS_2 monolayer. Such monolayers are formed by first intercalating the bulk material with reactive ions such as lithium, and then adding a solvent such as water that undergoes a reaction with the intercalant.

There is no reason to expect that all polymerized membranes or networks have uniform connectivity. For example, the erythrocyte cytoskeleton has approximately sixfold connectivity at its junction vertices, but the fraction of sixfold vertices is not 100%, and the connectivity of a given spectrin tetramer may not be fixed: There may be detachment and reattachment of spectrin, even if the rate is very slow. Further, there are hereditary

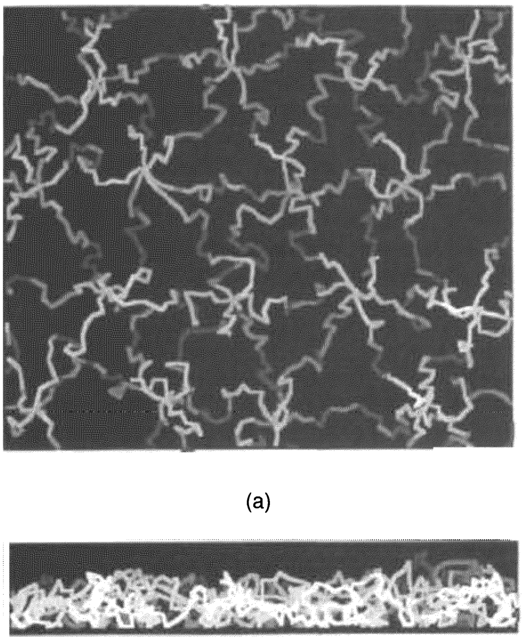

FIG. 6. Computer simulation of the membrane-associated cytoskeleton of the human erythrocyte. (a) The network seen from the cytoplasmic side of the membrane showing the sixfold connectivity of the cytoskeleton. (b) A section through the cytoskeleton. The lipid bilayer, not shown in (b), lies below the cytoskeleton.

diseases in which the spectrin concentration is depleted, resulting in incomplete networks and red cells with reduced elasticity (Mohandas and Evans, 1994).

1.3 Multilamellar Systems

One technique for vesicle production involves the swelling of multilayer (or *multilamellar*) stacks of dried lipid. Similar to, but much slower than, the production of inorganic monolayers by exfoliation described in Sec. 1.2, the invasion of water laterally through the stack forces the individual membranes apart. Not all bilayers in this procedure become separated from each other, and some remain in multilamellar structures. The bilayers in multilamellar systems are not always tightly bound to each other; there may be solvent trapped between lamellae, or several stacks of membranes may be linked by a single membrane. It is also possible to obtain more uniform multiple layers of material through the Langmuir–Blodgett technique (see Gaines, 1966). Small multilamellar stacks of membranes can be produced by the less gentle technique of applying high-frequency mechanical vibrations to lipid/solvent dispersions (sonication—see Fromherz *et al.*, 1986).

Part of the interest in multilamellar systems is the study of adhesion between bilayers. In some systems, thermal fluctuations of a membrane may prevent two membranes from adhering. These phenomena are treated in more detail in Sec. 6.

2. ELASTICITY

2.1 Deformations and the Strain Tensor

In one formulation of continuum mechanics, force is replaced by stress, and displacement is replaced by strain. Stress and strain can be represented by tensors, with the stress tensor having the dimensions of an energy density, while the strain tensor is dimensionless. In the statistical mechanics applications here, we use only the strain tensor; the reader interested in the more general formulation of continuum mechanics is referred to Fung (1994), Landau and Lifshitz (1986), or MECHANICAL PROPERTIES OF SOLIDS.

Consider the deformation of a two-dimensional square as shown in Fig. 7. Two-dimensional deformations are used here for diagrammatic simplicity; the notation is three dimensional. Two positions, marked by a and b on the undeformed square, define a vector $d\mathbf{x}$. A stress is applied to the square to change its shape and the points a and b move to new absolute positions, changing the vector $d\mathbf{x}$ to $d\mathbf{x}'$. Vectors $\mathbf{u}_a$ and $\mathbf{u}_b$ describe the displacement of positions a and b from the undeformed to the deformed state. The coordinate system does not change with the deformation.

Adding a constant displacement $\mathbf{u}$ to all positions on the object simply results in a translation of the object. In Fig. 7, the direction and magnitude of $\mathbf{u}$ vary locally, and the change of $\mathbf{u}$ with $\mathbf{x}$ provides a description of how vectors $\mathbf{x}$ at different locations on the object change with respect to each other. In our example, the vectors $d\mathbf{x}$ and $d\mathbf{x}'$ are related through

$$d\mathbf{x}' = d\mathbf{x} + \mathbf{u}_b - \mathbf{u}_a, \tag{1}$$

which can be written in the infinitesimal limit as

$$dx_i' = dx_i + \sum_j (\partial u_i/\partial x_j)dx_j, \tag{2}$$

where the components of each vector are indicated explicitly by the subscripts i and j.

The square of the length between neighboring points, dl^2, changes under the deformation to

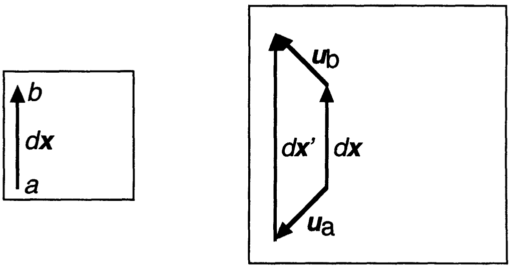

FIG. 7. Change in vector *dx* to *dx'* during a shape-preserving deformation. The initial object is shown on the left, while the deformed object is shown on the right. The initial position vector *dx* is superimposed on the deformed object to show the changes more clearly.

$$dl'^2 = dl^2 + 2\sum_{i,j} \epsilon_{ij} dx_i dx_j, \tag{3}$$

where

$$\epsilon_{ij} = \frac{1}{2}\left[\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} + \sum_k \frac{\partial u_k}{\partial x_i}\frac{\partial u_k}{\partial x_j}\right] \tag{4}$$

is the *strain tensor*. For deformations sufficiently small that terms quadratic in $\partial u_i/\partial x_k$ can be neglected, then the strain tensor can be approximated by

$$\epsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right). \tag{5}$$

2.2 Elastic Moduli

Objects may deform in response to applied forces that, in general, may have complex orientations with respect to the object. For example, while an object may have a simple cubic shape, the forces on the object may be at oblique angles with respect to the surface of the cube. A spatial component of an applied force can be written in terms of the stress tensor σ_{ij} via $\Sigma_i \sigma_{ij} s_j$, where $\mathbf{s}$ is a vector whose magnitude is equal to the surface area to which the force is applied and whose direction is normal to the surface.

A spring is a simple example of a Hooke's-law object in which the strain (the displacement from equilibrium) is proportional to the stress (the applied force). Hooke's law can be generalized to the stress and strain tensors defined above, although the single spring constant must be replaced by more complex proportionality constants. One way of writing Hooke's law is

$$\sigma_{ij} = \sum_{p,q} c_{ijpq}\epsilon_{pq}, \tag{6}$$

where the constants c_{ijpq} are the elastic constants and have units of energy density. In principle, there are 36 elastic constants for three-dimensional materials, but if the elastic energy is a quadratic function of the strain, as it is in Hooke's law, then the maximum number of independent constants is 21. Further, the number of independent constants depends on the symmetry of the system, with isotropic systems having only two elastic constants.

We now consider an isotropic system in more detail. The change in free-energy density of deformation, ΔF, can be expanded in powers of the strain tensor, and the lowest-order terms in the expansion are quadratic in ϵ. Two independent scalar combinations of ϵ are available: the squared sum of the diagonal elements, $(\mathrm{tr}\epsilon)^2$, and the sum of the elements squared, $\Sigma_{i,j}\epsilon_{ij}^2$. If we assume that the deformations are small, then ΔF can be written as

$$\Delta F = (\lambda/2)(\mathrm{tr}\epsilon)^2 + \mu \sum_{i,j} \epsilon_{ij}^2. \tag{7}$$

The constants λ and μ are called the Lamé coefficients, of which μ is the shear modulus. The compression modulus K can be expressed in terms of the Lamé coefficients by

$$K_A = \lambda + \mu \quad \text{(two dimensions)}, \tag{8a}$$

$$K_V = \lambda + 2\mu/3 \quad \text{(three dimensions)}. \tag{8b}$$

Both K and μ must be positive, or else the system could spontaneously deform by a shear or compression mode. The Lamé coefficient λ is not constrained to be positive, and systems are known in which λ is negative.

A further quantity of interest is the Poisson ratio σ_P, which is a measure of how much a material *contracts* in the transverse direction when it is stretched in the longitudinal direction. The geometrical definition of σ_P is then $-\epsilon_{yy}/\epsilon_{xx}$ for a deformation u_x in the x direction that results in a displacement u_y in the y direction. Poisson's ratio can be expressed in terms of elastic constants by

$$\sigma_P = (K_A - \mu)/(K_A + \mu) \quad \text{(two dimensions)}, \tag{9a}$$

$$\sigma_P = (3K_V - 2\mu)/(6K_V + 2\mu) \quad \text{(three dimensions)}. \tag{9b}$$

As an application, we consider a triangular network of springs in two dimensions, as shown in Fig. 8. The elastic moduli can be found by evaluating the energy change of a specific deformation and comparing the result with Eq. (7). At zero temperature, one finds

$$K_A = \sqrt{3}k_{sp}/2 \tag{10}$$

and

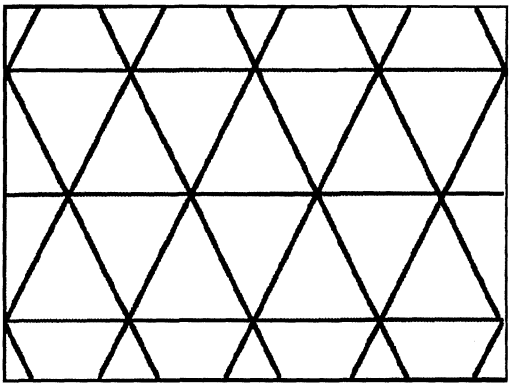

FIG. 8. Triangular network of springs in two dimensions.

$$\mu = \sqrt{3}k_{sp}/4 = K_A/2, \tag{11}$$

where k_{sp} is the spring constant. Note that both moduli are independent of the unstretched spring length. The Poisson ratio for this network is then

$$\sigma_P = \tfrac{1}{3}. \tag{12}$$

This sign of the Poisson ratio indicates that when the network is stretched in one direction, it contracts in the transverse direction. Many solids have a Poisson ratio of about $\frac{1}{3}$ in three dimensions. Some networks under tension are predicted to have a negative Poisson ratio.

2.3 Bending Resistance

The compression modulus is a measure of the free-energy change when the elements of a structure are forced closer together under pressure, or pulled further apart under tension. When the membranes illustrated in Figs. 3 and 5 are forced to bend, say about an axis pointing out of the plane of the diagram, then elements on one side of the midplane are forced closer together, while elements on the other side move farther apart. The free-energy change associated with the bending deformation is a measure of the bending resistance.

Just as with the shear and compression deformations of Secs. 2.1 and 2.2, we first need to develop a geometrical description of the bending of a surface. Following the approach of Safran (1994), we begin by reviewing the description of a curved line. In a coordinate system external to the line, each point on the curve is specified by a vector **R**, as in Fig. 9. The distance in space dl between two points separated by an infinitesimal arc length ds is then

$$dl = \left|\frac{\partial \mathbf{R}(s)}{\partial s}\right| ds, \tag{13}$$

where $|\mathbf{R}|$ represents the length of vector **R**. The unit vector normal to the curve, **n**, is proportional to the second derivative of the position vector:

$$C\mathbf{n}(s) = \frac{d^2\mathbf{R}(s)}{dl^2}, \tag{14}$$

where C is the curvature.

Now, the above description must be generalized for a two-dimensional surface embedded in three dimensions. Consider the surface shown in Fig. 10. Each point on the surface is specified by a height $h(x,y)$, which is a function of the fixed coordinate system x and y. In the figure, there is only one surface element for each (x,y) pair—there are no overhangs, and $h(x,y)$ is single-valued. We assume that the membrane of interest is sufficiently flat that this representation, referred to as the Monge representation, is valid.

The figure also shows vectors **n** normal to the surface. At each point (x,y), the normal has coordinates

$$\mathbf{n}(x,y) = (-\partial_x h, -\partial_y h, 1)/(1 + |\nabla h|^2)^{1/2}, \tag{15}$$

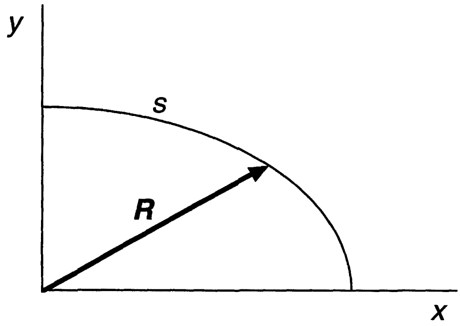

FIG. 9. Coordinates of a curve. The arc length s is the coordinate along the curve, while **R** is the position of a particular point on the curve, as measured with respect to an external coordinate system.

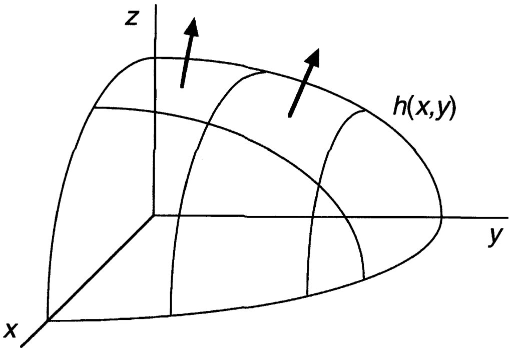

FIG. 10. Coordinates of a surface. Each element on the surface is described by its height $h(x,y)$ above the xy plane. Normals to the surface are indicated by bold arrows.

where ∂_x is $\partial/\partial x$. While there is a unique curvature associated with a given point on a line in Eq. (14), the curvature on a surface has a direction dependence that reflects the orientation of the vector $d\mathbf{R}$ along the surface:

$$C = -(d\mathbf{R}\cdot d\mathbf{n})/(d\mathbf{R}\cdot d\mathbf{R}). \tag{16}$$

Clearly, one has to establish a set of directions along which to evaluate the curvature, and the method of choice is to construct a curvature tensor $\mathbf{C}$ such that

$$d\mathbf{n} = d\mathbf{R}\cdot\mathbf{C}. \tag{17}$$

The eigenvalues of $\mathbf{C}$ are referred to as the principal curvatures, which we denote by C_1 and C_2. Note that there is no requirement that C_1 and C_2 have the same sign: a saddle point, for example, has principal curvatures of opposite sign.

The curvature tensor provides the geometrical description of a bending deformation, much as the strain tensor describes compression and shear in Sec. 2.1. Further, the free energy can be written as a function of the principal curvatures, as is done for the strain tensor in Sec. 2.2. The curvatures are usually written in the combinations $(C_1 + C_2)/2$ and C_1C_2, which are the mean curvature and Gaussian curvature, respectively. Expressed in terms of the principal curvatures, the leading-order terms in the free-energy density for bending, F, are

$$F = (k_b/2)(C_1 + C_2 - 2C_0)^2 + k_gC_1C_2, \tag{18}$$

where k_b is the bending rigidity and k_g is the Gaussian rigidity. The addition of the parameter C_0, referred to as the spontaneous curvature, allows for situations in which the membrane intrinsically bends to one side. The Gaussian rigidity is often omitted from expressions for the free energy in situations where the topology of the surface is closed, since the integral of the Gaussian curvature over a closed surface is a topology-dependent constant. The shapes of closed surfaces subject to bending resistance, as in Eq. (18), are discussed in VESICLES AND BIOMEMBRANES.

2.4 Elasticity and Fluctuations

The elastic moduli are introduced in Sec. 2.2 as terms in an expansion of the free-energy density. This approach is useful in statistical mechanics problems, since the weight accorded to a specific configuration in an ensemble is $\exp(-\beta E)$, where E is the energy of the configuration and β is the inverse temperature $(k_BT)^{-1}$. Statistical mechanics provides a description of the elastic moduli based on ensemble averages.

Consider the isothermal compression modulus, K_V, as an example. The physical definition of K_V is

$$K_V = -\frac{V}{\partial V/\partial P}, \tag{19}$$

where V is the volume of the system and P is the applied pressure. In an alternative approach that emphasizes the fluctuations in the system at nonzero temperature, the Gibbs free energy is expanded about the equilibrium value at a fixed pressure (Reif, 1965):

$$\Delta G = G(V) - G(V_{eq}) = (V - V_{eq})^2/(2V_{eq}K_V), \tag{20}$$

where V_{eq} is the equilibrium value of the volume. The corresponding partition function can be evaluated analytically, since the weight for a given volume V is proportional to $\exp(-\beta\Delta G)$. As a result, the compression modulus also can be expressed as

$$\beta K_V = \langle V\rangle/(\langle V^2\rangle - \langle V\rangle^2), \tag{21}$$

where $\langle V\rangle$ is the ensemble average of V.

A formally identical problem to this is the relation between the fluctuations in the length of a spring and the spring constant k_{sp}. The expression analogous to Eq. (21) is

$$\beta k_{sp} = (\langle l^2 \rangle - \langle l \rangle^2)^{-1}, \tag{22}$$

where l is the length of the spring. Note that Eq. (22) is independent of the unstretched spring length.

3. FLEXIBLE CHAINS

3.1 Ideal Chains

Linear and branched chains provide an intuitive introduction to the geometrical and elastic properties of membranes. We pause to assemble a number of results from polymer physics that will put the behavior of membranes in context.

The simplest geometry of a single polymer chain is one in which a chain element, or monomer, is freely jointed and can assume any orientation irrespective of the orientation of its neighbors. The conformations of the chain are equivalent to a random walk if the chain elements are permitted to intersect each other. Consider the situation in which the elements of the chain are represented by vectors $\mathbf{a}_i$, each of which has the same length a. A chain with N segments has a contour length along the chain of Na. Of interest is the end-to-end displacement vector $\mathbf{r}_{ee}$, illustrated in Fig. 11, which is just the sum of the individual vectors: $\mathbf{r}_{ee} = \Sigma_i \mathbf{a}_i$. The ensemble average over all possible conformations at fixed N yields the well-known result

$$\langle r_{ee}^2 \rangle = Na^2. \tag{23}$$

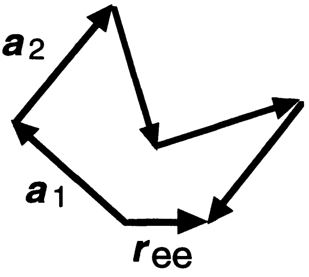

FIG. 11. End-to-end displacement $\mathbf{r}_{ee}$ for a chain whose elements have a common length and random orientation.

Thus, the end-to-end displacement grows with contour length like $N^{1/2}$. Other geometrical quantities, such as the radius of gyration, also scale like $N^{1/2}$.

Consider now the slightly more complicated problem of chains with a fixed polar angle θ between successive segments i and $i + 1$, such that

$$\mathbf{a}_i \cdot \mathbf{a}_{i+1} = -a^2 \cos\theta, \tag{24}$$

where the convention is that $\theta = 0$ if neighboring segment vectors $\mathbf{a}_i$ point in opposite directions. Although the polar angle between segments is fixed, the azimuthal angle is not, and the segments are free to rotate around one another. As shown in Flory (1953), the end-to-end displacement of this chain at large N is

$$\langle r_{ee}^2 \rangle = Na^2(1 - \cos\theta)/(1 + \cos\theta). \tag{25}$$

The important feature to note in Eq. (25) is that the end-to-end displacement scales like $N^{1/2}$, in spite of the restrictions on the chain's conformations. In fact, the $N^{1/2}$ scaling is found for all chains on sufficiently large length scales, so long as the chain path can cross itself. It is then useful to introduce the idea of an effective bond length b which reflects the length scale at which the chains appear to be ideal; that is, $\langle r_{ee}^2 \rangle = Nb^2$. The effective bond length has the same order of magnitude as the persistence length, which measures the length scale over which the spatial orientation of the monomers is correlated [see Eq. (43)]. For the restricted orientation situation of Eq. (25), the effective bond length is

$$b = a[(1 - \cos\theta)/(1 + \cos\theta)]^{1/2}. \tag{26}$$

In alkane chains (the hydrocarbon chains present in soap, for example), θ is close to the tetrahedral angle of 109.5°, and the effective bond length is $\sqrt{2}a$, according to Eq. (26), or about 0.2 nm. Their rigidity and monomer size dictate that protein chains have considerably longer effective bond lengths, several nanometers or more, depending on the protein. Of course, steric interactions prevent physical systems, such as proteins, from intersecting themselves, and so their behavior is not described fully by ideal-chain scaling.

3.2 Self-Avoiding Chains

While the "ideal chains" of Sec. 3.1 may intersect themselves, physical systems have an excluded volume that enforces self-avoidance of the chain. This steric interaction among the chain elements is important for chains in one-, two-, and three-dimensional systems. Consider the simple situation in which a chain lies along the x axis. Self-avoidance forbids the chain from reversing on itself from one step to the next, so that the end-to-end distance must be just the contour length Na. But Eq. (23) shows that $\langle r_{ee} \rangle$ scales like $N^{1/2}$ for ideal chains, *independent of embedding dimension.* Thus, we conclude that in one dimension, self-avoidance of the chain dramatically affects its scaling properties. The same conclusion can be drawn for chains in two and three dimensions.

A simple model for the length-scaling exponent of self-avoiding chains was proposed by Flory (1953). The calculation evaluates the power-law dependence of the free energy on the effective chain size r and number of segments N at both large and small r. Minimizing the free energy yields r as a function of N. Since our goal is to extract the scaling exponent, we do not pay close attention to numerical factors like $4\pi/3$, and we use r to represent the effective size of the chain, as characterized by an end-to-end length or a root mean square radius. Also, our model chains have no explicit energy scale other than the temperature. Following de Gennes (1979), the behavior of the free energy is evaluated in two regimes:

1. Short distances. Steric repulsion between the chain segments causes the chain to swell compared to an ideal chain. The repulsive energy experienced by one segment through its interaction with other segments is proportional to the concentration of segments, roughly N/r^d for a chain in a d-dimensional space. Thus, the total repulsive energy experienced by all N segments is proportional to N^2/r^d. Now, the repulsive energy also will be proportional to the excluded volume of the segment–segment interaction, which we characterize by a parameter v_{ex}. Taking the excluded volume as a hard-core interaction, then the energy scale of the interaction is set by the temperature $k_B T$. Thus, the steric contribution to the free energy should behave like

$$F = k_B T v_{ex} N^2 / r^d, \tag{27}$$

where all constants have been absorbed into v_{ex}.

2. Long distances. As a chain is stretched, the number of configurations that it can adopt at a fixed end-to-end distance decreases rapidly. As shown in Sec. 3.4, the probability of finding a given end-to-end distance r for an ideal chain decays exponentially as $\exp(-dr^2/2Na^2)$, where a is the elementary segment length. Recalling that the entropy S is proportional to the logarithm of the probability, then to within a constant

$$S/k_B = -dr^2/2Na^2. \tag{28}$$

The entropic contribution to the free energy at long distances can be found through $F = E - TS$. Combining Eqs. (27) and (28) and discarding overall normalization constants, the free energy of the self-avoiding chain behaves like

$$F = k_B T v_{ex} N^2 / r^d + k_B T d r^2 / 2Na^2. \tag{29}$$

This expression shows that there is a penalty for pushing the chain elements close together (small r), and there is a penalty for stretching out the chains (large r). The value for r that minimizes F can be found by taking the derivative of Eq. (29) with respect to r, holding other quantities fixed, and this value scales like

$$r \approx N^{3/(2+d)}. \tag{30}$$

The scaling behavior of Eq. (30) is expected for any length scale r that characterizes the linear dimension of the system as a whole, such as the end-to-end distance r_{ee} and the root mean square radius of the system $(\langle r_{ee}^2 \rangle^{1/2}/\sqrt{6})$. The exponent on the right-hand side of Eq. (30) is called the Flory exponent.

The systems discussed thus far in this section are linear chains. Chains with side branches, referred to as branched polymers, have different scaling with system mass N than do linear chains, since branching will add monomers to a given position along the chain. Because a branched polymer has more than two ends, the end-to-end displacement is replaced as a measure of the polymer size by some other quantity such as the radius of gyration R_g, defined by

$$R_g^2 = \sum_i (\mathbf{r}_i - \mathbf{r}_{\text{c.m.}})^2, \tag{31}$$

where $\mathbf{r}_{\text{c.m.}}$ is the center-of-mass position of the chain and the sum is over the positions of the segments. The radius of gyration for branched polymers is found to have a scaling form

$$\langle R_g^2 \rangle \approx N^{2\eta}. \tag{32}$$

The branched-polymer scaling exponents are smaller than the self-avoiding walk exponents, being 0.64 and 0.5 in two and three dimensions respectively, compared to 0.75 and 0.59 for self-avoiding walks. The scaling exponents for linear and branched polymers are summarized in Table 1.

In real polymers, the chains may be rigid because of bond-angle restrictions, and there may be attractive van der Waals interactions between monomers. What effects do these interactions have on scaling behavior? As shown in Sec. 3.1, making chains stiffer only increases the persistence length of the chain; it does not change the scaling behavior in the large-N limit. Of course, for small N the introduction of rigidity may change the appearance of the chain configuration substantially. The situation is different if there is attraction between chain elements. If the attraction is strong enough, then the chains collapse to a dense phase in which the chain size scales like $N^{1/d}$. For example, self-avoiding walks in three dimensions without attractive interactions scale like $N^{3/5}$, but for strong attraction they scale like $N^{1/3}$. The transition between the two scaling regimes is a phase transition and occurs at a temperature (or attraction strength) called the Θ point.

3.3 Self-Avoiding Rings

Let us close our linear chain to form a loop in two dimensions, which will serve as a two-dimensional vesicle. Closing the chain introduces a new characteristic, namely, the enclosed ring area, as well as the conjugate thermodynamic variable of pressure P. The parameter space of the rings then includes bending rigidity, van der Waals attraction, and pressure.

We consider the effects of each parameter in turn, with the others set equal to zero.

Table 1. Scaling exponents η for the scaling law $\langle R_g^2 \rangle \sim N^{2\eta}$ for ideal or random walks, self-avoiding walks, and branched polymers as a function of embedding dimension d.

Configuration	$d = 2$	$d = 3$
Random walk	1/2	1/2
Self-avoiding walk	3/4	0.59
Branched polymer	0.64	1/2

1. Rigidity. The presence of rigidity changes the persistence length, but not the asymptotic scaling behavior at large N, although the scaling behavior at finite N is nontrivial.
2. Pressure. Defining

   $$\Delta P = P_{\text{inside}} - P_{\text{outside}}, \tag{33}$$

   then at $\Delta P > 0$, the ring is always inflated at large N and the scaling behavior is $\langle R_g^2 \rangle \approx N^2$ and $\langle A \rangle \approx N^2$. At zero pressure, the ring should behave like a self-avoiding walk, with $\langle R_g^2 \rangle \approx N^{3/2}$ and $\langle A \rangle \approx N^{3/2}$. When the outside pressure exceeds the inside pressure ($\Delta P < 0$), the ring belongs to the branched polymer family but is not dense. For the ring to look like a branched polymer, it must have little enclosed area, so that $\langle R_g^2 \rangle \approx N^{1.28}$ and $\langle A \rangle \approx N$.
3. Attraction. If the attraction between monomers on the chain is strong enough, then the configurations may become dense, and $\langle R_g^2 \rangle \approx N$ and $\langle A \rangle \approx N$. This is similar to the Θ point of linear polymers.

A summary of the scaling exponents is given in Fig. 12 for two-dimensional rings simultaneously subject to pressure and attraction (Boal, 1991).

3.4 Chain Elasticity

There are comparatively few configurations in which an ideal chain is fully stretched: The most common configurations are ones in which the chain is highly convoluted. Since entropy is proportional to the logarithm of the number of configurations, then another way of stating this observation is that the entropy of the chain is reduced, and the free energy is increased, as a chain is stretched. This implies that the chain has elasticity arising from its entropy.

To find the relationship between the elas-

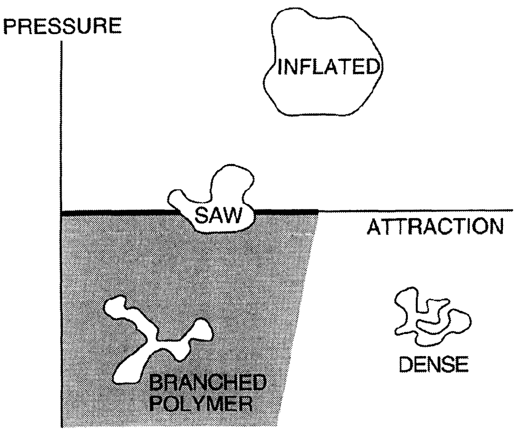

FIG. 12. Phase diagram for two-dimensional self-avoiding rings subject to pressure and attraction. There may be considerable finite-size scaling effects near the phase boundaries.

ticity of an ideal chain and its geometry, we consider the distribution of end-to-end vectors $\mathbf{r}_{ee}$. Let us break up the three-dimensional distribution into three one-dimensional distributions. The x component of the end-to-end displacement, $r_{ee,x}$, is just the sum of the individual monomer segment lengths projected onto the x axis: $r_{ee,x} = \Sigma_i a_{i,x}$. For ideal chains, $a_{i,x}$ is uncorrelated with $a_{i+1,x}$, and the situation is the same as the one-dimensional random walk with steps of variable size. In the continuum limit, the probability of finding a one-dimensional walk with end-to-end displacement between x and $x + dx$ is just

$$\rho(x)dx = (2\pi Na^2/d)^{-1/2}\exp(-dx^2/2Na^2)dx, \tag{34}$$

where $\rho(x)$ is the probability density (probability per unit length) and d is the embedding dimension. This distribution is normalized to unity and has expectations

$$\langle r_{ee,x}\rangle = \int x\rho(x)dx = 0 \tag{35}$$

and

$$\langle r^2_{ee,x}\rangle = \int x^2\rho(x)dx = Na^2/3 \tag{36}$$

for the x component of the end-to-end displacement vector in three dimensions.

Equations (35) and (36) can be used with Eq. (22), which relates the spring constant to the fluctuations in spring length, to obtain an expression for the effective spring constant of the chain:

$$k_{sp} = dk_BT/Na^2. \tag{37}$$

Equation (37) shows that the effective stiffness of the chain increases with d and the temperature T. The elastic behavior of the chain is entropic in origin, since there is no potential-based restoring force at the microscopic level.

4. FLUCTUATING MEMBRANES

4.1 Fluid Membranes

The lipid bilayer component of a cell's plasma membrane is a fluid in the sense that it cannot resist an in-plane shear. In the absence of bending rigidity or in-plane shear resistance, fluid membranes can assume convoluted shapes and are, in a sense, the two-dimensional analogs of freely jointed chains. The geometry of membranes can be described using similar scaling relationships to those used to describe linear and branched chains (see Secs. 3.1–3.3).

Closed configurations, such as spherical or toroidal geometries, have a larger number of scaling observables than do open configurations. An example of a closed fluid membrane with the topology of a sphere is shown in Fig. 13. If the membrane had a sufficiently large value for k_b, then the spherical shape would be obvious, but in fact, the membrane shown in Fig. 13 has $k_b = 0$. The shape is reminiscent of a branched polymer, in that the enclosed volume is proportional to the area, rather than the area to the $\frac{3}{2}$ power expected for a sphere. An analysis performed via computer simulation confirms that closed fluid membranes obey the same branched polymer scaling as self-avoiding surfaces (Kroll and Gompper, 1992):

$$\langle R_g^2\rangle \sim A^1, \tag{38a}$$

$$\langle V\rangle \sim A^1. \tag{38b}$$

The generalized Flory exponent [see Eq. (30)] for self-avoiding surfaces is

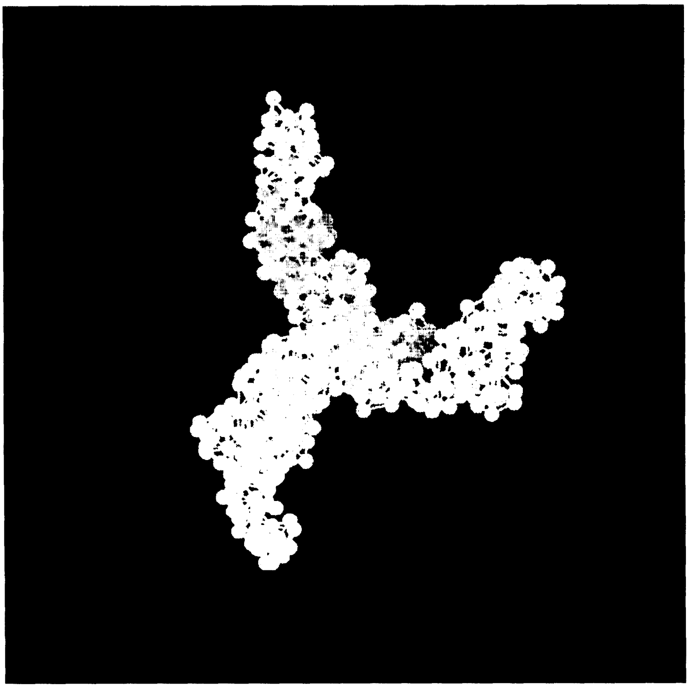

FIG. 13. Sample configuration from a computer simulation of a fluid membrane with spherical topology. The surface, which has no bending resistance, is represented by discrete points: hard beads linked together by tethers of a maximal extension. The beads are not attached to a fixed set of neighbors, but can migrate through the membrane.

$$\nu_F = (2 + D)/(2 + d), \tag{39}$$

where D is the dimension of the surface ($D = 1$ for chains, $D = 2$ for membranes) and d is the embedding dimension (see lectures by Kantor in Nelson *et al.,* 1989). Equation (39) predicts that $\langle R_g^2 \rangle \sim A^{0.8}$ for a fluid membrane in three dimensions, but this scaling behavior has been seen only for moderate system sizes, not large systems. Fluid membranes with bending resistance also belong to the class of branched polymers, although the distance at which the asymptotic scaling behavior becomes apparent increases with k_b. Again, this is like the scaling of chains with bending resistance described in Sec. 3.1.

Fluid-membrane undulations, at distances short compared to the persistence length, can be treated using the Monge representation (see Sec. 2.3). The mean square fluctuation in the height $h(\mathbf{r})$ at position $\mathbf{r}$ is predicted to vary as

$$\langle h(\mathbf{r})^2 \rangle \sim TL^2/k_b, \tag{40}$$

where T is the temperature and L is the linear dimension of the system (see Safran, 1994). This relationship demonstrates that the fluctuations increase with temperature and system size. In contrast, the mean square undulations at an interface dominated by surface tension (between water and air, for example) grow only logarithmically with the system dimension L.

The membrane height can be expressed in terms of its Fourier components $h_\mathbf{q}$ via

$$\langle h(\mathbf{r})^2 \rangle = A^{-1} \sum_\mathbf{q} \langle |h_\mathbf{q}|^2 \rangle \tag{41}$$

where A is the membrane area. The Fourier components corresponding to Eq. (40) scale as

$$\langle |h_\mathbf{q}|^2 \rangle \sim T/k_b q^4. \tag{42}$$

In comparison, the Fourier components of the height of an interface dominated by surface tension scale like $\langle |h_\mathbf{q}|^2 \rangle \sim T/q^2$. Experimental observations of membranes under tension are consistent with the undulation spectrum of Eq. (42) (see references in Mohandas and Evans, 1994).

Because of surface undulations, the normals to the surface $\mathbf{n}(\mathbf{r})$ are correlated only at short distances. That is, the product of two normals decays with displacement $\mathbf{r}$ as

$$\langle \mathbf{n}(0)\cdot\mathbf{n}(\mathbf{r})\rangle \sim \exp(-|\mathbf{r}|/\xi_p), \tag{43}$$

where ξ_p is the persistence length. Perturbation theory can be used to calculate the low-temperature fluctuations of a membrane whose energy is characterized by a bending modulus k_b, as in Eq. (18), yielding a persistence length of (see lectures by Leibler in Nelson *et al.*, 1989)

$$\xi_p \sim a \exp(4\pi\beta k_b/3), \tag{44}$$

where a is a microscopic length scale and β is the inverse temperature $(k_B T)^{-1}$. The numerical factors in the exponential depend on the method used for the calculation. What is important about Eq. (44) is that the persistence length is exponential in the bending rigidity.

The persistence length of a chain sets the length scale above which bending resistance and other geometrical constraints are sufficiently unimportant that the chain conformations are described by a random walk. Another way of saying this is that the effective bending resistance of the chain decreases with the increasing length scale on which the resistance is measured. The behavior of fluid sheets can be viewed in the same way, that the effective bending modulus $k_{b,\mathrm{eff}}$ decreases with the increasing length scale of the system. Again, perturbation theory consistent with Eq. (44) indicates that to leading order, the effective bending resistance has the form

$$k_{b,\mathrm{eff}}(L) = k_b[1 - (3/4\pi\beta k_b)\ln(L/a)], \tag{45}$$

where L is the length scale of interest (see lectures by Leibler in Nelson *et al.*, 1989).

4.2 Polymerized Membranes

Polymerized membranes and networks have fixed connectivity, by which we mean that a given element of the membrane has a fixed set of neighbors, as distinct from fluid membranes in which an element can diffuse through the membrane. The shear rigidity imparted by fixed connectivity results in constraints on the configuration space of the membrane, which, in turn, affect its scaling behavior. The shape of the fluid membrane in Fig. 13 easily can be seen to be different from the shape of the polymerized membrane in Fig. 14: The fluid membrane is convoluted whereas the polymerized membrane is flat.

Scaling analyses of open polymerized membranes show that

$$\langle R_g^2\rangle \sim A^1, \tag{46a}$$

$$\langle h^2\rangle \sim A^\zeta, \tag{46b}$$

where A is the contour area (or mass) of the membrane and ζ is the roughness exponent. The mean square height $\langle h^2\rangle$ in Eq. (46b) measures the thickness of the membrane and is the smallest eigenvalue of the membrane's inertia tensor. First measured to be 0.65 ± 0.05 (Plischke and Boal, 1988), the roughness exponent obtained from recent computer simulations is close to the field-theoretic prediction of 0.59 (Le Doussal and Radzihovsky, 1992).

The polymerized membrane in Fig. 14 is flat in spite of the fact that it has no local bending resistance: $k_b = 0$. The membrane becomes increasingly rigid at long length scales because of geometrical constraints im-

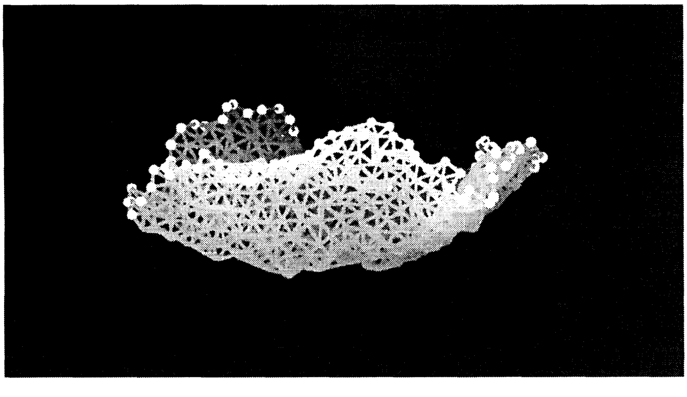

FIG. 14. Sample configuration from a computer simulation of an open polymerized membrane with no local bending resistance. As in Fig. 13, the surface is represented by discrete points; but here, the beads are attached to a fixed set of neighbors.

posed by polymerization. As experimentation with aluminum foil will convince the reader, an initially flat polymerized membrane cannot even approximate the configuration shown for the fluid membrane in Fig. 13 without extensive folding of the surface, and such folding is strongly suppressed by entropy. Thus, the effective bending resistance $k_{b,\mathrm{eff}}$ of a polymerized membrane should increase with length scale L, rather than decrease as found for fluid membranes. Several calculations show that $k_{b,\mathrm{eff}}$ diverges like $L^{2-2\zeta} \sim L^{0.8}$ (see Nelson *et al.*, 1989).

Although the bending resistance of a polymerized membrane increases with L, other in-plane moduli are expected to decrease with L because of undulations. Calculations of the in-plane area compression modulus and shear modulus both show moduli that vanish at large L, as, for example,

$$\mu \sim L^{-(4\zeta-2)} \sim L^{-0.4}. \tag{47}$$

Most of these predictions have been confirmed by computer simulation. Further, the Poisson ratio of polymerized networks is predicted to be universally equal to $-\frac{1}{3}$ (Le Doussal and Radzihovsky, 1992).

Experimental tests of the analytic and simulation results thus far have addressed whether polymerized membranes are flat, and not all of the measurements are in agreement. One set of measurements on the isolated two-dimensional cytoskeleton of the red blood cell indicates that the network is flat on long length scales. Other work on inorganic monolayers does not support this finding.

4.3 Membranes under Stress

The membranes discussed in Secs. 4.1 and 4.2 are isolated, in the sense that they are not subject to external stress nor do they interact with a medium or with other membranes. However, a complete description of physical membrane behavior must include the characteristics of interacting membranes. In this section, the effects of compression and tension on membranes will be described, while interactions between membranes are presented in Sec. 6.

4.3.1 Closed Fluid Membranes Closed fluid membranes with spherical topology (a computational *vesicle*) subject to pressure have been simulated both for surfaces whose elements are plaquettes on a lattice and for triangulated surfaces that are not constrained to lie on a lattice, and whose connectivity can evolve with time (as in Fig. 13). Both rings in two dimensions (Sec. 3.3) and fluid vesicles in three dimensions obey branched-polymer scaling under compression, but only vesicles show branched-polymer scaling at zero pressure. Depending on the vesicle mass, the scaling may remain branched-polymer even if the pressure inside the vesicle moderately exceeds the pressure outside; however, at large pressure differences, there is a transition to an inflated state.

4.3.2 Open Membranes at a Wall The interaction between a membrane and a wall has been investigated for two situations: a sheet forced against a flat surface by pressure and a sheet confined between two flat surfaces. The presence of the wall removes the divergence in the fluctuation spectrum of fluid membranes observed at long wavelengths [Eq. (42), see Safran, 1994]. The behavior of polymerized membranes near walls can be used to determine the roughness exponent.

4.3.3 Planar Membranes under Tension Observed at sufficiently short length scales, membranes under tension, or membranes with bending resistance, are relatively flat. The elasticity of a planar, polymerized network of springs under tension has been investigated by simulation. At zero temperature, the in-plane elastic moduli vary with stress σ as

$$K_A/k_{sp} = (3^{1/2}/2)(1 - \sigma/3^{1/2}k_{sp}), \tag{48a}$$

$$\mu/k_{sp} = (3^{1/2}/4)(1 + 3^{1/2}\sigma/k_{sp}), \tag{48b}$$

where positive σ corresponds to a network under tension and k_{sp} is the spring constant of an individual spring. These expressions, which have Eqs. (10) and (11) as their zero-stress limit, predict that the Poisson ratio of the network becomes negative when $\sigma/k_{sp} > \sqrt{3}/5$. At finite temperature, in-plane fluctuations increase the moduli above their zero-temperature values.

5. SYSTEMS WITH DEFECTS

5.1 Percolation Phenomena

The networks and membranes typically found in Nature are not the ideal, uniform systems that are considered in Sec. 4. Rather, they may be chemically inhomogeneous and/or may contain structural defects. Two examples of defective membranes and networks are mentioned in Sec. 1.2: partially polymerized membranes in which polymerization is random or localized, and spectrin-depleted erythrocyte cytoskeleta in which the coordination of some junction complexes may be much less than sixfold. Further, defects may develop in otherwise regular membranes and networks when they are subject to stress or elevated temperatures, leading ultimately to mechanical failure. These structural aspects of membranes are an active area of study in statistical mechanics, although the investigation of systems with defects dates back more than a century to Maxwell.

One very useful concept that emerged from computational studies of defective systems is that of *percolation.* Consider the two-dimensional triangulated network shown in Fig. 15, in which the light gray "bonds" are fluid (as in Fig. 13), while the white "bonds" are polymerized and connected to a fixed set of vertices (as in Fig. 14). If most or all of the bonds are fluid, then the shear modulus of the network vanishes. There must be a reasonable concentration of polymerized bonds present before the network exhibits shear resistance. Computer simulations of systems such as this have shown that

1. there is a well-defined concentration, referred to as the *rigidity percolation* threshold p_r, above which networks resist shear; and
2. there is a well-defined concentration, referred to as the *connectivity percolation* threshold p_c, above which a connected sequence of polymerized bonds completely traverses the network,

where p may be defined as the fraction of polymerized bonds present in the network compared to the fully polymerized network, for which $p = 1$. In Fig. 15, p is greater than p_c, and uninterrupted paths of polymerized bonds can be seen in both the x and y directions. There are many examples in which the connectivity and rigidity thresholds are different (Thorpe, 1986).

5.2 Defective Two-Dimensional Networks

Connectivity and rigidity percolation occur at the same polymerized bond concentration in the mixed-composition network of

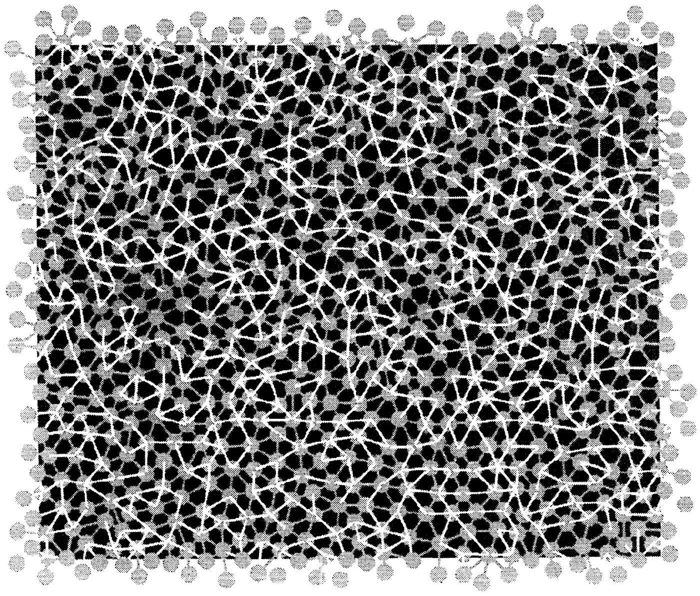

FIG. 15. Sample configuration of partially polymerized membrane with polymer bond fraction p = 0.5, viewed perpendicular to the membrane plane. The fluid bonds are gray, while the polymerized bonds are white.

Fig. 15, namely, $p_r = p_c = 2\sin(\pi/18) = 0.35$ expected for triangulated networks in two dimensions. However, if the fluid bonds of Fig. 15 are completely removed and not replaced by polymerized bonds, thus leaving open holes in the membrane, then the membrane is floppier and p_r increases to about $\frac{2}{3}$, while p_c, of course, remains the same. The mixed-composition (fluid/polymerized) networks with no open holes have higher rigidity at a given polymerized bond concentration than the single-component (polymerized only) networks with holes because of the presence of isolated pockets of fluid that cannot diffuse past the polymerized regions. Single-component bond- or site-diluted networks have been studied in both two and three dimensions, and p_r is typically found to be d times p_c in the systems studied, where d is the embedding dimension (Thorpe, 1986).

Once the rigidity percolation threshold has been crossed from below p_r, the elastic moduli do not rise immediately to their values for a fully connected network. Rather, to a first approximation, the shear modulus rises linearly with $p - p_r$:

$$\begin{aligned} \mu(p)/\mu(p=1) &= (p-p_r)/(1-p_r) \quad \text{for}\, p > p_r, && \text{(49a)} \\ &= 0 \quad \text{for}\, p < p_r, && \text{(49b)} \end{aligned}$$

where, again, $p = 1$ for a fully polymerized network.

Both the mixed-composition and bond-depleted networks may be candidates to describe various partially polymerized networks. Consider the model erythrocyte cytoskeleton of Fig. 6, in which the chains represent the spectrin tetramers of the cytoskeleton and give the model network its elasticity [see Eq. (22)]. In some hereditary blood diseases, the erythrocytes have a lower spectrin concentration and may be examples of bond-diluted, single-component networks (but not mixed polymer/fluid networks since there is no trapped fluid component to the cytoskeleton). The shear modulus of these depleted cytoskeleta is found to decrease linearly as the spectrin content is reduced, as expected qualitatively from Eq. (49a). Although data are not available for highly depleted cytoskeleta (such cells are not viable), the extrapolated value for p_r is much lower than $p_r \sim \frac{2}{3}$ of a bond-depleted triangular network. This does not necessarily indicate that percolation theory is inapplicable to spectrin-depleted erythrocytes; it may be that the coordination of such cytoskeleta is still relatively uniform, but less than sixfold.

5.3 Defective Membranes in Three Dimensions

Bond-depleted networks in two dimensions possess reduced in-plane rigidity, as described in Sec. 5.2. When allowed to fluctuate out of plane into the third dimension, defective networks have fewer geometrical constraints and may adopt configurations in which one region of the membrane passes through a hole in another region. It is known that *phantom* polymerized membranes, in which the surface is free to intersect itself, are not flat but highly crumpled if $k_b = 0$ (see lectures by Nelson in Nelson *et al.*, 1989). Are bond-depleted membranes also crumpled?

The flatness characteristics of two types of bond-depleted (but otherwise self-avoiding) network have been investigated. In one case, bonds are randomly removed from a polymerized triangulated network, as described in the previous section on percolation. This type of network is observed to be flat for any $p > p_c$. In contrast, the geometry of a bond-depleted network known as a Sierpinski gasket is found to be described by the Flory exponent in Eq. (39).

A different defect structure is one in which network connectivity is uniform, but the interactions between junctions are variable. For example, the interaction strength or geometry of the intervertex potential might vary randomly across a network, corresponding to a glassy ground state. Simulation studies of systems with randomly varying interactions are challenging, because the network ground states are difficult to determine uniquely, and because many configurations must be sampled to construct accurate ensemble averages. Thus far, the temperature dependence of the membrane phases between the ground state and high-temperature configurations has not shown novel behavior in the limited studies of inhomogeneous networks with glassy ground states.

6. INTERACTIONS BETWEEN MEMBRANES

6.1 Interactions between Rigid Surfaces

The interactions of flexible membranes, particularly biological systems, involve forces of several different origins. Although this article focuses on effects arising from thermal fluctuations, it is important to appreciate the type and strength of chemical, entropic, and other components of membrane interactions. Here, we briefly review several features of the forces between rigid surfaces, before discussing the effects of undulations in Secs. 6.2 and 6.3. A more extensive treatment of these forces can be found in Israelachvili (1992).

6.1.1 van der Waals Forces Attractive, but relatively weak in strength, van der Waals forces may extend several nanometers away from a surface. The interaction arises from fluctuations in charge distributions at the atomic level: A quantum mechanical fluctuation that results in a dipole moment in one atom induces a dipole moment in a second atom. Thus, an attractive dipole–dipole interaction driven by fluctuations can be present between atoms with no permanent dipole moments. The van der Waals energy per unit surface area between two rigid surfaces with separation s decreases like s^{-2} at modest separations, and like s^{-4} at large separations.

6.1.2 Double-Layer Forces As introduced in Sec. 1.1, fluid membranes typically have a polar headgroup exposed to a polar solvent. Many polar groups can dissociate, leaving a charged region surrounded by counterions in the neighboring solvent, as illustrated in Fig. 16. In other situations, head groups may be capable of adsorbing ions from the solvent. Consequently, the membrane may carry a surface charge, which, together with the counterions in solution, constitute an electric double layer. Double-layer forces, which have been investigated theoretically and experimentally, are longer ranged than van der Waals forces, and are predominantly repulsive. The energy per unit surface area for the interaction between two double layers decays exponentially with separation.

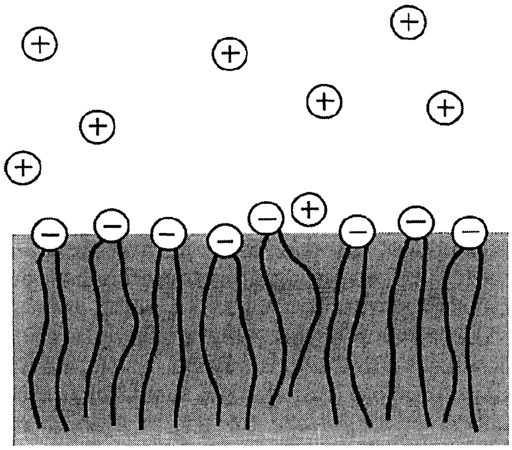

FIG. 16. Double layer at the boundary of a lipid monolayer with a solvent.

6.1.3 Hydration Forces The polar head groups of amphiphiles in a free biomembrane are surrounded by solvent molecules, which may be "bound" in the sense that the amphiphile and solvent have a strong mutual attraction. These solvent molecules must be removed if two membranes are to approach each other closely, suggesting that there is a repulsive hydration force associated with the hydrated amphiphile (the word "hydration" refers to water as a solvent). The pressure required to force bilayers together at small interlayer separation s is approximately exponential, $\exp(-s/s_0)$, and s_0 is found to be in the 1–3-nm range (Rand and Parsegian, 1989). The range of the hydration force is thus of the same magnitude as the attractive van der Waals force or repulsive thermal fluctuations of the bilayer itself.

6.1.4 Steric Interactions All of the above forces may apply to rigid or flexible membranes. However, flexible membranes also may undergo thermal fluctuations, resulting in a repulsive steric interaction between a rigid wall and a membrane or between membranes themselves. Membrane components, such as proteins, that extend into the solvent may give rise to additional steric interactions that are short-ranged compared to undulations but are more repulsive than undulations at small length scales.

The combined van der Waals and double-layer forces comprise the DLVO theory (Derjaguin and Landau, 1941; Verwey and Overbeek, 1948) of colloidal stability. As we have indicated, there may be several short-ranged interactions present in biomembranes in addition to the DLVO forces.

6.2 Entropic Repulsion of Fluctuating Membranes

The idea that membrane fluctuations result in repulsive interactions has support from several experimental observations:

1. vesicles composed of certain lipids show increased binding when their surfaces are under lateral tension;
2. in some situations, repulsion between lipid bilayers is reduced when one of the bilayers is attached to a rigid substrate; and
3. interbilayer repulsion may increase with temperature.

There are several fluctuation modes for a membrane, including out-of-plane undulations and fluctuations in membrane thickness. Here, we consider undulations, whose importance was pointed out by Helfrich (1978), and refer the interested reader to Israelachvili (1992) for a discussion of other fluctuation modes.

Consider a stack of n membranes as shown in Fig. 17, in which each membrane is taken to be parallel to the xy plane on average. A set of n functions is needed to describe the displacement of the membrane stack in the z direction. We label these functions $\psi_n(x,y)$ to show explicitly their x, y, and n dependence, and we assume that they are single valued as in the Monge representation. If the membranes are equally spaced on average, then

$$\int \psi_n(x,y)\, dxdy = ns, \tag{50}$$

where s is the average spacing between successive membranes and where the $n = 1$ membrane has an average height s above the xy plane. One can change the functions to regain a more familiar form by introducing a set of height functions $h_n(x,y)$

$$h_n(x,y) = \psi_n(x,y) - ns, \tag{51}$$

such that $\int h_n(x,y)\, dxdy = 0$. For gentle undulations, the energy density of the membrane stack is

$$H = \int \Big[(B/2) \sum_n (h_{n+1} - h_n)^2 + (k_b/2) \sum_n (h_{n,xx} + h_{n,yy})^2 \Big] dxdy, \tag{52}$$

where $h_{n,xx} + h_{n,yy}$ is the mean curvature of the nth membrane and k_b is the bending modulus [see Eqs. (14) and (18)]. For notational simplicity, the xy arguments have been omitted from Eq. (52).

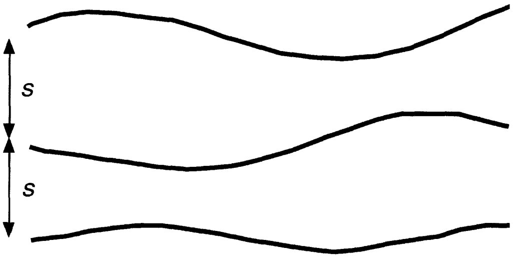

FIG. 17. Cross section through a membrane stack illustrating intermembrane repulsion arising from thermal fluctuations. The average separation between successive membranes is s.

Compared to Eq. (18), the new feature of Eq. (52) is the elastic modulus B, which represents the intermembrane repulsion. Helfrich (1978) uses a self-consistent argument for obtaining B by (1) determining as a function of B the difference in the free energy per unit volume Δf_v between the many-membrane and single-membrane systems, and (2) imposing the elasticity relationship

$$B = s \frac{\partial^2 \Delta f_v}{\partial s^2}, \tag{53}$$

to the free energy found by procedure (1). This method yields (Safran, 1994)

$$B = 36/\pi^2 \beta^2 k_{\mathrm{B}} s^4. \tag{54}$$

and a corresponding difference in the free energy per unit area $\Delta f_A = s \Delta f_v$ of

$$\Delta f_A = 3/\pi^2 \beta^2 k_{\mathrm{B}} s^2. \tag{55}$$

There are two aspects of the free-energy expression in Eq. (55) worth noting. First, the s^{-2} dependence of the entropic repulsion is the same as that of the van der Waals interaction between rigid surfaces at moderate separations. Second, the repulsion increases with temperature. Both of these features are consistent with observations **1** and **2** listed at the beginning of this section. Helfrich (1978) estimates that the repulsion from undulations is about $\frac{1}{4}$ of the van der Waals interaction in magnitude for egg lecithin membranes.

A mean-field approach that avoids the introduction of the modulus B has been used by Evans and Parsegian (1986) to obtain the undulation modes analogous to Eq. (42) and the free-energy density Δf_A. At large separations, their expressions have the same functional forms as those obtained by Helfrich and illustrate the substantial shift in the free-energy density across the 2–3-nm range in s arising from undulations.

6.3 Unbinding Transition

Fluctuations cause membranes to repel each other, and the magnitude of the repulsion increases with temperature. In principle, such fluctuations can overcome an attractive interaction between membranes that might be sufficient to bind them at low temperatures. Indeed, bilayer membranes have been observed to dissociate from a closely bound membrane stack over a very narrow temperature range (see Lipowsky in Lipowsky and Sackmann, 1995).

Since van der Waals interactions decrease more rapidly than s^{-2} at large separation s, then, for the purpose of investigating membrane binding, it is sufficient to use a square-well potential to represent the attractive interaction between two membranes:

$$\begin{aligned} V(s) &= -U \text{ for } 0 < s < s_{\max}, \\ V(s) &= 0 \text{ for } s > s_{\max}, \end{aligned} \tag{56}$$

where U and $s_{\max}$ are positive parameters that set the energy and length scales of the potential. At zero temperature, the membranes are confined to distances less than $s_{\max}$, as dictated by Eq. (56). At temperatures greater than zero, the membranes repel but do not unbind until the transition temperature T^* given by

$$T^* \sim (k_b U s_{\max}^2)^{1/2}, \tag{57}$$

which predicts that the more rigid the membranes or the stronger their attraction, the higher the unbinding temperature.

Systematic calculations have shown that the unbinding transition is continuous and that the mean separation diverges with temperature like $|T - T^*|^{-1}$ as the transition is approached from below T^*. For representative lipid bilayers, the estimated theoretical unbinding temperature varies by 50% about room temperature, depending on the membrane rigidity. Whether the suppression of thermal undulations is responsible for the tension-induced adhesion observed experimentally remains to be seen. Recent calculations, based on a description of the steric interaction that encompasses both the rigidity-dominated and tension-dominated regimes, show that the applied tension must be very large to permit tension-induced adhesion of fluid membranes (Seifert, 1995).

GLOSSARY

Cytoskeleton: A two- or three-dimensional protein network in a cell's interior.

Erythrocyte: A red blood cell; human erythrocytes are structurally simple since they do not have a nucleus and their cytoskeleton is connected only to the plasma membrane.

Fatty Acid: A carboxylic acid (-COOH) with a long hydrocarbon chain, found in natural fats and oils.

Fluid Membrane: Two-dimensional system with no resistance to in-plane shear.

Lipid: A small water-insoluble molecule with one or more hydrocarbon chains.

Lipid Bilayer: A double layer of oriented amphiphilic molecules; in biomembranes, the lipid bilayer is fluid.

Monolayer: A sheetlike structure with a thickness of one molecular unit.

Persistence Length: Length scale at which the relative orientation of chains or surfaces is uncorrelated.

Polymerized Membrane: Two-dimensional system with resistance to in-plane shear.

Protein: A macromolecule composed of one or more polypeptide chains (amino acids linked by peptide bonds); actin, spectrin, and tubulin are examples of proteins that form two- and three-dimensional networks in cells.

Works Cited

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J. D. (1989), *Molecular Biology of the Cell,* 2nd Ed., New York: Garland.

Boal, D. H. (1991), "Phases of attractive two-dimensional vesicles under pressure," *Phys. Rev. A* **43,** 6771–6777.

Derjaguin, B. V., Landau, L. (1941), *Acta Physicochim. URSS* **14,** 633–662.

Divigalpitiya, W. M. R., Frindt, R. F., Morrison, S. R. (1989), "Inclusion systems of organic molecules in restacked single-layer molybdenum disulphide," *Science* **246,** 369–371.

Evans, E. A., Parsegian, V. A. (1986), "Thermal-mechanical fluctuations enhance repulsion between bimolecular layers," *Proc. Natl. Acad. Sci. U.S.A.* **83,** 7132–7136.

Flory, P. J. (1953), *Principles of Polymer Chemistry,* Ithaca: Cornell University Press.

Fromherz, P., Röcker, C., Rüppel, D. (1986), "From discoid micelles to spherical vesicles: The concept of edge activity," *Faraday Discuss. Chem. Soc.* **81,** 39–48.

Fung, Y. C. (1994), *A First Course in Continuum Mechanics,* Englewood Cliffs, NJ: Prentice-Hall.

Gaines, G. L. Jr. (1966), *Insoluble Monolayers at Liquid-Gas Interfaces,* New York: Wiley.

Helfrich, W. (1978), "Steric interaction of fluid membranes in multilayer systems," *Z. Naturforsch.* **33a,** 305–315.

Israelachvili, J. N. (1992), *Intermolecular and Surface Forces,* 2nd Ed., New York: Academic.

Kroll, D. M., Gompper, G. (1992), "The conformation of fluid membranes: Monte Carlo simulations," *Science* **255,** 968–971.

Landau, L. D., Lifshitz, E. M. (1986), *Theory of Elasticity,* 3rd Ed., Oxford: Pergamon.

Le Doussal, P., Radzihovsky, L. (1992), "Self-consistent theory of polymerized membranes," *Phys. Rev. Lett.* **69,** 1209–1212.

Lipowsky, R., Sackmann, E. (Eds.) (1995), *Structure and Dynamics of Membranes,* Amsterdam: Elsevier.

Mohandas, N., Evans, E. (1994), "Mechanical properties of the red cell membrane in relation to molecular structure and genetic defects," *Annu. Rev. Biophys. Biomol. Struct.* **23,** 787–818.

Nelson, D., Piran, T., Weinberg, S. (Eds.) (1989), *Statistical Mechanics of Membranes and Surfaces,* Singapore: World Scientific.

Plischke, M., Boal, D. H. (1988), "Absence of a crumpling transition in strongly self-avoiding tethered membranes," *Phys. Rev. A* **38,** 4943–4945.

Rand, R. P., Parsegian, V. A. (1989), "Hydration forces between phospholipid bilayers," *Biochim. Biophys. Acta* **988,** 351–376.

Reif, F. (1965), *Fundamentals of Statistical and Thermal Physics,* New York: McGraw-Hill.

Ringsdorf, H., Schlarb, B., Venzmer, J. (1988), "Molecular architecture and function of polymeric oriented systems: models for the study of organization, surface recognition and dynamics of biomembranes," *Angew. Chem. Int. Ed. Engl.* **27,** 113–158.

Safran, S. A. (1994), *Statistical Thermodynamics of Surfaces, Interfaces and Membranes,* Reading, MA: Addison Wesley.

Seifert, U. (1995), "Self-consistent theory of bound vesicles," *Phys. Rev. Lett.* **74,** 5060–5063.

Thorpe, M. (1986), "Elastic properties of network glasses," *Ann. N. Y. Acad. Sci.* **484,** 206–213.

Verwey, E. J. W., Overbeek, J. Th. G. (1948), *Theory of Stability of Lyophobic Colloids,* Elsevier: Amsterdam.

Further Reading

de Gennes, P.-G. (1979), *Scaling Concepts in Polymer Physics,* Ithaca: Cornell University Press.

Gennis, R. B. (1989), *Biomembranes: Molecular Structure and Function,* New York: Springer-Verlag.

Gompper, G., Schick, M. (1994), "Self-assembling amphiphilic systems", in: C. Domb, J. L. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena,* Vol. 16, New York: Academic.

Israelachvili, J. N. (1992), *Intermolecular and Surface Forces,* 2nd ed., New York: Academic.

Mouritsen, O. G. (1988), "Computer simulation of cooperative phenomena in lipid membranes," in: *Molecular Description of Biological Membrane Components by Computer-Aided Conformational Analysis,* Boca Raton, FL: CRC Press.

Nelson, D., Piran, T., Weinberg, S. (Eds.) (1989), *Statistical Mechanics of Membranes and Surfaces,* Singapore: World Scientific.

Safran, S. A. (1994), *Statistical Thermodynamics of Surfaces, Interfaces and Membranes,* Reading, MA: Addison Wesley.

STEAM ENGINES

STEPHEN LUCHTER, *U.S. Department of Transportation, Washington, D.C., U.S.A.*

	Introduction	563
1.	**System Considerations**	564
2.	**Expanders**	565
2.1	Positive Displacement	565
2.1.1	Configurations	566
2.1.2	Valving	568
2.1.3	Pistons	568
2.1.4	Lubrication	569
2.1.5	Other Components	569
2.1.6	Design Considerations	569
2.2	Dynamic (Turbomachinery)	570
2.2.1	Impulse Turbines	570
2.2.2	Reaction Turbines	570
2.2.3	Losses	570
2.2.4	Control	570
2.2.5	Design Considerations	571
3.	**Condensers**	571
4.	**Pumps**	571
4.1	Positive Displacement	571
4.2	Dynamic	572
5.	**Vapor Generators**	572
5.1	Heat-Transfer Subsystem	572
5.2	Combustion Subsystem	573
6.	**Applications**	573
6.1	Electric-Power Generation	573
6.2	Stationary Power Plants	574
6.3	Motor-Vehicle Propulsion	574
6.4	Rail Propulsion	574
6.5	Ship Propulsion	575
7.	**Other Working Fluids**	576
8.	**Other Cycles**	577
	Glossary	577
	Works Cited	578
	Further Reading	578

INTRODUCTION

Steam engines can trace their lineage to the early 1700s, when the need for a means of pumping water out of English coal mines generated a burst of creative development. In the earliest practical steam engine, patented by Thomas Savery 1698, steam admitted to a cylinder pushed water in the cylinder through a pipe that had a check valve to stop return flow. The steam supply was stopped, and cold water flowing over the cylinder caused a vacuum that in turn allowed water being pumped to be sucked in. In 1712 Thomas Newcomen developed an engine that also condensed steam in a cylinder, but here the vacuum was used to move a piston, which, connected to a beam, in turn moved another piston to pump the water. This engine was used to drive machinery in addition to pumping water and is reported to have been a commercial success. James Watt improved on the Newcomen engine in the 1770s by removing the condensing function from the working cylinder as well as introducing a number of innovations, including pistons that allowed steam to be admitted to both sides of the piston (double-acting pistons) and the concept of allowing the steam to expand after being admitted to the cylinder. In the late 1700s, Richard Trevithick developed engines using steam at pressures above atmospheric and began adapting his designs to steam-driven road vehicles and locomotive propulsion. In 1804 he demonstrated a steam locomotive. A variation of the Trevithick design was applied by George Stephenson in his "Rocket" locomotive in 1829. This launched the practical application of steam for rail propulsion. A practical marine propulsion engine was demonstrated in 1801 by William Symington. In 1838 the *Great Western* crossed the Atlantic ocean powered by steam.

Steam engines have been applied to a variety of other purposes, including farm machinery, electrical-generating plants, and auxiliary drives, as well as propulsion for automobiles, ships, and rail locomotives. There were even steam-powered bicycles in the late 1890s and a steam-powered airplane in the 1930s.

digital Encyclopedia of Applied Physics

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

With a few important exceptions, such as electrical-power generation and marine propulsion, the steam engine is of historical significance at this time. The literature of the past two decades includes reports of refurbishing old engines, celebration of centennials or bicentennials of famous developers, and design exercises or experimental applications as part of waste-heat recovery equipment, in solar and geothermal power applications, and in developing countries where indigenous fuels are available.

This article discusses the mechanical manifestation of the Rankine cycle, including the thermodynamic, fluid mechanical, and mechanical design considerations of the overall system and of each of the major components. It also summarizes the characteristics of some of the more notable applications and of efforts to use fluids other than water in a variety of Rankine-cycle applications.

1. SYSTEM CONSIDERATIONS

The ideal Rankine cycle described in HEAT ENGINES AND REFRIGERATORS consists of isentropic compression of a saturated liquid to boiler pressure, followed by isobaric heat addition, first to a saturated liquid, then phase change to a saturated vapor. Work is performed during an isentropic expansion, and heat is rejected in an isobaric phase change (condensation) back to liquid. One reason the Rankine cycle has been so widely applied is that there are many practical ways to approximate this theoretical cycle.

The efficiency of the cycle is the net work divided by the heat added. For a desired amount of work out, the most efficient cycle is the one with the least heat added. On the basis of thermodynamic fundamentals (the Carnot cycle), the cycle efficiency also increases with higher average temperatures for the heat addition.

One commonly applied method of reducing the amount of heat added is to use some form of internal heat transfer between the expander exhaust and the feed water. In large central-station electrical-generating steam plants, feed-water heating usually takes the form of extraction of some steam from the turbine at a number of places along the expansion, using it to heat water on its way to the vapor generator.

One way to increase the average temperature of heat addition is to superheat the vapor, that is, heat the saturated vapor to a higher temperature than the boiling temperature. Another way is to use reheat, that is, to return the partially expanded steam to the vapor generator for further heat addition.

Although ideally it is desirable to have as high a peak temperature as possible in a Rankine cycle, the actual choice is limited by vapor-generator design considerations. Even though combustion-gas temperatures may be on the order of 1400 °C, very few steam power plants have peak steam temperatures as high as 600 °C, with most somewhat lower than 550 °C. Peak pressures are typically on the order of 7 MPa (except for some large central-station plants operating in the supercritical region, where pressures will be on the order of 22 MPa). One reason for this is that boiler tubes must be designed to withstand the pressure on the basis of the properties of materials suitable for use at these high temperatures. Few materials are available that meet the strength criteria. Another limitation on the choice of peak temperature, when a positive-displacement expander is used, is the properties of available lubricants.

At the heat-rejection part of the cycle, in most cases the steam will be condensed. The exceptions are where the expander exhaust steam can economically be discharged directly to the atmosphere, where the steam is usable for other purposes, or where there is no room for a condenser. Although thermodynamic considerations dictate as low a condensing temperature as possible, this temperature is limited by the available heat sink. For large central-station electrical-generating plants, this is usually a large body of water. For surface transportation applications, the heat sink is the ambient air temperature. Marine propulsion engines use the body of water they are traversing.

When considering real engines as compared to ideal cycles, a number of other practical limitations must be taken into account. Each of the subsystems contains a number of auxiliary components. The vapor generator not only includes the heat-transfer components but it also contains the components necessary for combustion: a fuel pump and its drive for liquid fuels or equivalent devices for solid fuels, a combuster, an air blower and its drive, and the controls for these units. It often also contains a heat exchanger that preheats the combustion air and a heat exchanger that uses the exhaust combustion products to heat the water. The expander includes a lubricant pump and its auxiliaries. The condenser usually contains a lubricant separator to remove any lubricant in the expander exhaust so that it does not coat the surfaces of the condenser or get carried over to the boiler, where in either case it would affect the heat transfer. If the condenser is air cooled, there will be some kind of fan drive plus possibly some additional air-flow control such as louvers. If the condenser is water cooled, there will be a cooling-water pump and its controls. If only minimal cooling water is available, a cooling tower may also be employed. The pumping subsystem often includes a boost pump and its drive as well as the main feed pump and its drive. For propulsion applications, a transmission or other speed-changing devices may also be necessary.

There are other factors that differentiate the real steam engine from an ideal Rankine cycle. Expansions and compressions are not isentropic—expanders produce less work on expansion and pumps require more work on compression than the ideal components. Flow through heat exchangers results in pressure drops, increasing the amount of pump work required. Heat is lost through piping and heat-exchanger walls. Some steam leaks past valve stems and piston rings, or turbine shaft seals, doing no work. Also, ideal cycles do not take into account the work that must be provided to drive the numerous auxiliaries. In small plants, these auxiliary loads tend to be a large part of the total work output.

There are further inherent limitations on the overall power-plant design in propulsion applications. One of these is the requirement to operate at part load for major portions of the time. This makes it difficult to optimize component design for high efficiency. Engines for naval combat vessels and for motor vehicle propulsion must also be able to change output rapidly. Probably the most severe limitation for most propulsion applications is the strict limits on the available volume, weight, or frontal area, all of which tend to limit attainable efficiencies.

2. EXPANDERS

There are two fundamental types of expanders: positive displacement and dynamic or turbine expanders. Each has its own advantages and disadvantages.

Positive-displacement expanders have relatively high efficiency at part loads. Countering this advantage is a pressure drop across the valving, a limited volumetric expansion ratio, the friction losses as a piston or other positive-displacement device moves in its chamber, and the difficulty in finding practical lubricants for use with high-temperature superheated steam.

Turbomachinery has the advantage of a potentially higher design-point efficiency than a positive-displacement expander, and the ability of having the highest temperatures in stationary components so that turbine blades and disk and the lubricants are not exposed to these temperatures. Turbines are typically lighter than positive-displacement expanders; however, the inherent high speed of the output shaft often requires gearing to bring shaft speed to a usable range, which may negate any weight and size advantage. Turbines also have poor part-load efficiency, and sealing the high-speed shafting may be difficult.

2.1 Positive Displacement

A pressure–volume (p-V) diagram is helpful in visualizing what happens in a positive-displacement expander. In an ideal positive-displacement expander (Fig. 1), steam is admitted at constant pressure to a point where the ensuing expansion would occur adiabatically to the exhaust pressure. The steam is exhausted to a point where the steam remaining in the cylinder would be compressed adiabatically to just fill the clearance space at the pressure of the incoming steam. No heat losses or friction are assumed to occur.

Real engines, of course, do not follow this ideal cycle exactly. "Good" design would have the steam enter the cylinder when the piston is nearing the top of its stroke, and steam admission would end shortly after the piston began its downward stroke. Similarly, exhaust would begin shortly before the piston reached the bottom of its stroke, and a

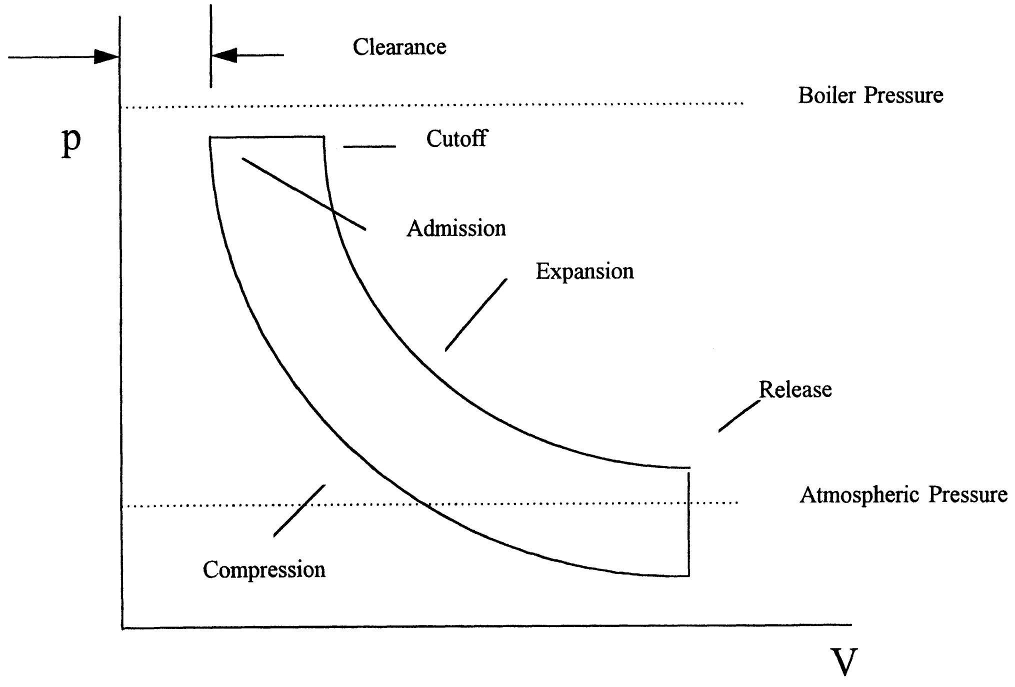

FIG. 1. Schematic pressure–volume diagram of positive-displacement steam engine expander.

minimum of steam would remain in the cylinder to be recompressed.

Although p-V diagrams of operating engines now are obtained by means of electronic pressure transducers, early practice going back to the time of Watt used a great variety of mechanical devices that could provide accurate indications of what was occurring inside the cylinder at the low rotational speeds employed. The "reading" of pressure–volume diagrams (indicator cards) of steam engines developed to a high art, with numerous publications showing how such diagrams could tell the operator where in the cycle the variations from the ideal pressure–volume relationship were occurring. For example, an indicator would show pressure drop in the valving, the cutoff point not at the ideal point, the expansion pressure below the exhaust pressure, or recompression pressure exceeding inlet pressure.

2.1.1 Configurations The simplest positive-displacement expander consists of a piston moving in a cylinder (Fig. 2). A connecting rod is attached to a crank of a drive shaft. A more widely applied design employs a crosshead interposed between the piston and the crankshaft (Fig. 3). The crosshead can be thought of as a sliding bearing. Attached to the crosshead is one end of the connecting rod, the opposite end of which is attached to a crank. This allows the piston rod to move in reciprocating motion with the connecting rod now moving in a complex motion to drive the crank of the output shaft.

Various arrangements of steam flow in the cylinder are possible. In a single-acting engine, steam expands against only one side of the piston, whereas in a double-acting engine, it expands against both sides. An advantage of the double-acting design is that there are two "working strokes" per revolution of the crankshaft. Steam flow into the cylinder can be either uniflow, where it enters at one end and leaves at the other, or counterflow, where it enters and leaves from the same end.

The uniflow concept was first successfully demonstrated by Stumpf in 1908 as a means of reducing the tendency for the incoming steam to condense on the cold cylinder walls. With superheated steam, initial con-

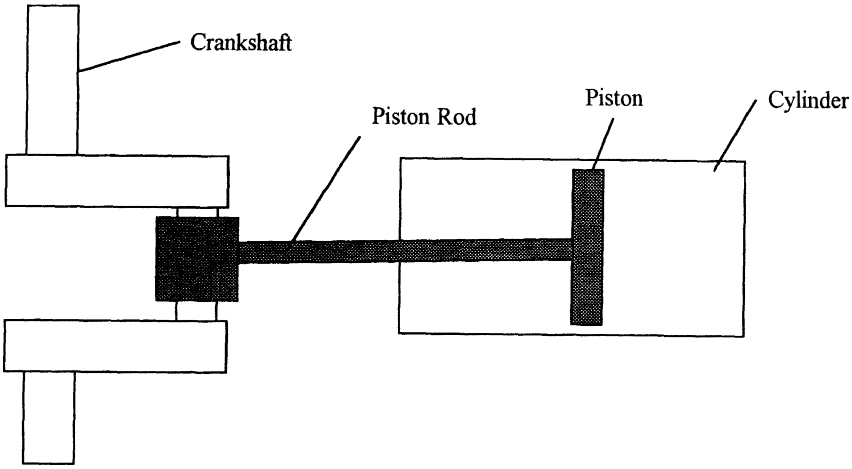

FIG. 2. Diagram of simple expander.

densation is not a significant problem. However, there are other advantages of this expander design. Since in a uniflow expander the exhaust steam exits the cylinder through ports cut into the side of the cylinder, there is no need for exhaust valves, and thermal stresses are reduced by allowing a fairly uniform temperature gradient from the cylinder head down to the crankcase. Another advantage of the uniflow design is that the engine efficiency remains high across the load range, which makes it particularly attractive for applications with widely varying loads.

At higher steam pressure more than one stage of expansion can result in higher overall efficiency as well as a more practical engine design. Engines with two stages of expansion (compound engines) were common in many applications, including some locomotive designs. Triple expansion was widely applied in ship propulsion engines, with some marine engines employing quadruple expansion.

Other positive-displacement expander configurations have been used at low power levels or for auxiliary drives. These include

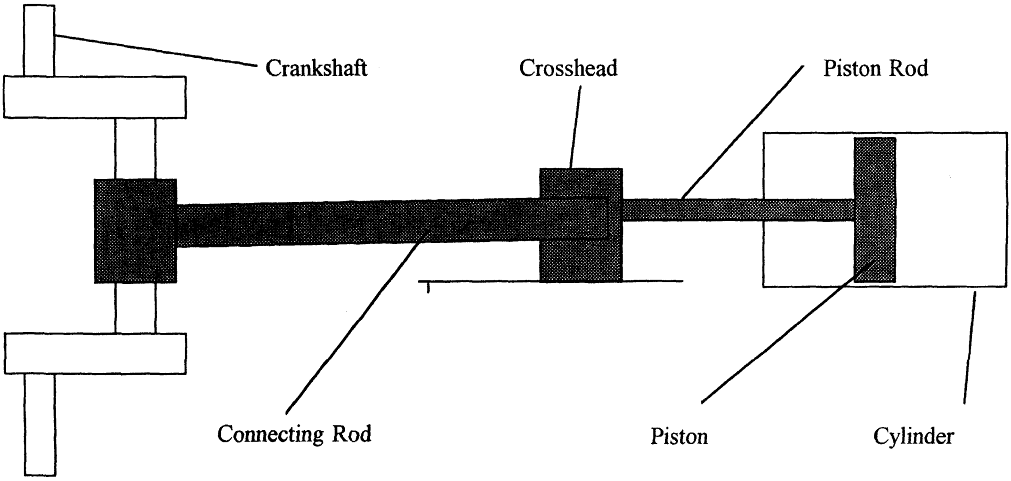

FIG. 3. Diagram of expander with crosshead.

the rotary-vane expander, which has movable vanes in slots of an off-center rotating shaft. As the working fluid expands, it drives the vanes, which, in turn, rotate the shaft. Trochoidal (Wankel) and multiple-cylinder piston designs resembling hydraulic motors have also been applied but not frequently. Considerations of friction and leakage, although present in the piston–crank arrangement, are of greater concern in these other configurations, thus limiting their application. Unless otherwise noted, the remainder of this article considers only the piston–crank arrangement of positive-displacement expanders.

2.1.2 Valving When valves open, steam enters the expansion chamber. Controlling the portion of the cycle during which a valve is open can be used to control the amount of steam admitted and thus the power output of the engine. A cutoff valve remains open to admit steam for a designated portion of the rotation. An early cutoff allows less steam into the cylinder, and thus more of the stroke is available for expansion. A late cutoff allows more steam into the cylinder at the high pressure. This does work, but without expansion. A variable cutoff can be used to change the point where expansion begins.

An alternate approach to controlling engine output is to use a throttling valve at the engine inlet. This valve adjusts the incoming steam pressure and thus the amount of expansion.

For decades, sliding valves were the common design, driven by an eccentric off the main output shaft. Figure 4 shows the principle of operation of this type of valve. As the valve moves down, steam is admitted to the top of the cylinder while it is exhausted from the bottom of the cylinder. As the valve continues to move down, it comes to a point where the openings for both inlet steam and exhaust steam are the greatest. The valve then reverses and closes the steam flow to the top of the piston while allowing exhaust to continue. As the valve continues to move upward, it closes both inlet and exhaust. With further upward motion, steam begins to exhaust from the top of the piston and begins to be admitted to the bottom of the piston. Using a similar principle, cylinder or spool valves were also used in some engines.

A disadvantage of the slide valve is that when the valve plate nears the end of its admission or exhaust positions, there is a loss as the steam has to flow through a small opening. The Corliss valve introduced a quick-acting rotary valve design that reduced the throttling loss significantly. In this design four rotary valves were used. The design was used in many stationary engines and increased the overall efficiency.

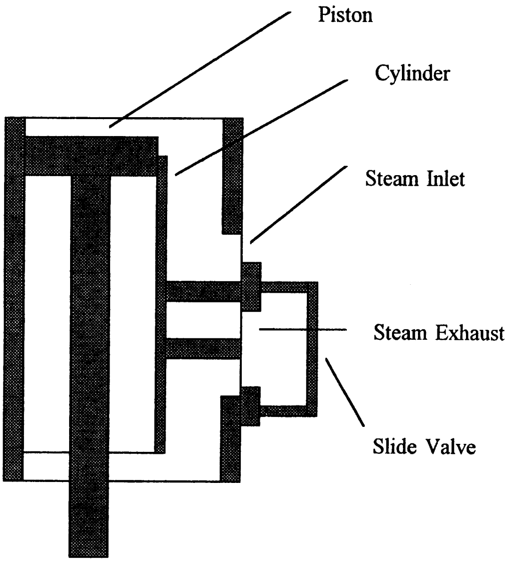

FIG. 4. Diagram of steam engine slide valve.

In a uniflow expander, the piston can be considered to be a valve, as the outlet ports are uncovered when the piston moves to a particular point in the expansion process.

Depending on the details of the expander design, some portion of the entering steam will remain in the cylinder after the exhaust valves or ports have been closed but before the admission valve has been opened. The steam remaining in the cylinder will then be compressed, possibly to a pressure higher than the inlet pressure. To eliminate any problems that may result from overpressure, a recompression valve can be fitted to open at a pressure near the incoming steam pressure.

2.1.3 Pistons The piston is the mechanical interface with the expanding steam. The unbalanced pressure force in the cylinder causes the piston to move, doing work.

Pistons are usually fairly thin disks relative to their diameter. The pistons operate in a hostile environment, with their tops exposed to the highest temperature and pressure.

A critical part of the piston assembly is the rings to seal the pressure. These rings must seal the high pressure in a high-temperature environment against a back pressure that may be close to or possibly below atmospheric. These rings create friction work and must be carefully designed to minimize the loss. Their placement along the axis of the piston must be optimized as well. Rings placed near the top of the piston reduce the overall clearance volume; however, this design puts the rings in the highest-temperature environment. Placing the rings further down on the piston skirt has the reverse effect.

Another limitation on the piston design is the need to meet the dynamic requirements of the engine. Piston mass must be kept low for this reason.

2.1.4 Lubrication When steam engines used saturated steam, the condensing water along the cylinder surface was often sufficient lubricant. For superheated steam, lubricants must retain their lubricating properties in a high-temperature steam environment where many conventional lubricants thermally decompose, losing their lubricating properties. Since some of the lubricant will inevitably become entrained in the working steam, the lubricant must be separable from the steam prior to being returned to the boiler so that the boiler tubes do not become contaminated. The composition of many steam-engine lubricants is proprietary, but it is known that they usually contain additives for extreme pressure operation, as well as rust inhibitors and demulsifiers to assist in separation from the working fluid.

2.1.5 Other Components The cylinder in which the piston slides also must be compatible with the operating conditions. The surface must be smooth and be able to withstand the temperatures and pressures encountered in these engines. In some designs using saturated steam, the cylinders were surrounded by steam jackets in order to minimize the problem of initial condensation.

Cylinder heads are exposed to the highest pressures and temperatures and must be of a material that can withstand those conditions. The gasket between the cylinder head and the expander body is also a critical component, as it too is exposed to the high temperatures of the inlet steam.

Other components, such as connecting rods, crank shafts, bearings, seals, and engine mounts, are designed in a similar manner as any other reciprocating machinery components, taking into account the temperature environment in which they must operate.

2.1.6 Design Considerations Determining the dimensions for most positive-displacement machinery is based largely on very simple fundamental relationships adjusted by empirical factors. The basic relationship is that the power is a function of the mean effective pressure, the displacement, and the number of cycles per unit time. This is useful because the desired power output is the starting point for any design, and the rotational speed is generally known, based on the application.

Mean effective pressure is the integrated average pressure over the cycle. The expansion of the steam can be approximated by a polytropic expansion where pV^n is a constant. The exponent has been found to range from about 0.7 for wet steam to 1.0 for dry steam to 1.5 for superheated steam. Under the assumption that n equals 1.0, the mean effective pressure can be approximated by

$$p_m = p_1(1 + \ln r_e)/r_e - p_3,$$

where p_m is the mean effective pressure, p_1 is the pressure at the cutoff point, p_3 is the pressure at the point of release, and r_e is the ratio of the volume at the release point to the volume at the cutoff point. The cutoff point refers to the point where the steam admission stops, that is, cuts off. The release point is the point at which the steam begins to exhaust from the cylinder. The value of the mean effective pressure found by this equation is modified by a factor that experience has shown to relate this theoretical value to an actual value. The factor varies from about 0.75 to 0.85 depending on the type of valving as well as such factors as the clearance volume, the size of the passageways, and the speed of operation.

The displacement is the volume of the cylinder. In a double-acting engine, the number of cycles is twice the number of rotations per unit time (revolutions per minute or second).

Mechanical design must also take into account several boundary conditions, including cylinder-wall temperature (which limits the choice of cylinder material and available lubricants and contributes to heat losses), and piston sliding velocity (which is correlated with frictional losses).

2.2 Dynamic (Turbomachinery)

Practical steam turbines were developed in the late 19th and early 20th century, even though the principles were demonstrated much earlier—Hero's reaction turbine in 150 B.C. and Branca's impulse turbine in 1629. Today steam turbines are the primary drivers of electrical generators in large central stations, either fossil or nuclear fueled. Small steam turbines have a number of industrial applications, largely for auxiliary drives in process plants where steam is otherwise available.

In a turbine, energy of a fluid is converted into velocity in a nozzle, with the velocity a function of the square root of the enthalpy change. The jet passes across a set of blades, which are attached to a disk, which, in turn, is attached to a shaft, with rotary motion the final result. The most common turbine design has the flow roughly parallel to the shaft center line, although radial-flow turbines are used in some applications. Only axial-flow turbines are considered here. The two main axial-flow turbine types are impulse and reaction.

2.2.1 Impulse Turbines In an impulse turbine the steam expands in a stationary nozzle and is directed through a set of blades. An impulsive force is imparted to the blades as the steam leaves in the opposite direction. Impulse turbines typically have a large pressure drop in the nozzle, with no additional expansion across the moving blades. In a pure impulse turbine the pressure on the downstream side of the blades is the same as on the upstream side. An advantage of the impulse turbine is that the steam expands in a stationary nozzle, lowering its temperature. This allows the turbine disk, which may be highly stressed, to run at a lower temperature. It can be shown theoretically that the best efficiency of an impulse turbine occurs when the blade velocity is half the velocity of the jet exiting from the nozzle. In practice, slightly less than half the velocity is commonly used for turbine design.

There are a number of variations of the impulse turbine. When the steam passes across one set of moving blades and is then redirected by a stationary set of blades to a second set of moving blades, it is called a Curtis or velocity stage. Other variations of the impulse turbine include turbines that have reentry of the steam after passing through the moving blades, and tangential flow types of turbines where the steam "corkscrews" around several blades in the circumference of the disk.

2.2.2 Reaction Turbines In a reaction turbine, part of the expansion occurs in the nozzle and part in the rotating blades. The rotating blades have a different shape than impulse blades. They are designed as nozzles rather than turning vanes. Reaction stages typically have a relatively low pressure drop across each stage. The best efficiency of a reaction turbine occurs when the blade velocity equals the jet velocity. As a result, reaction turbines tend to operate at higher rotational speeds than impulse turbines.

In large multistage turbines, it is common for the first few stages to be impulse stages to reduce the pressure and temperature the blades and disks are exposed to, with the later stages reaction.

2.2.3 Losses In any turbine, there are a number of losses that reduce the actual output below what is theoretically attainable. Some heat is lost through the turbine case. Nozzles, even well-designed ones, do not convert energy adiabatically. Windage losses are the result of the friction of the steam against the rotating disk. Leakage losses occur when some of the steam goes between the blades and the turbine housing. This can be minimized by having a shroud on the outer circumference of the wheel. There is also an exhaust loss, as the steam exiting the turbine has some velocity that was not converted to work.

2.2.4 Control Power-level control of a turbine can be accomplished in a number of ways. Flow into a turbine can be partial admission, wherein steam is directed at only a portion of the blades. As more power is desired, the mass flow is increased by opening additional nozzles. Partial admission reduces the attainable efficiency as only a portion of the blades is doing any work, and the rest of the disk is exposed to windage and other losses. Many small turbines are controlled by partial admission. Full-admission turbines, wherein steam flows across all of the blades, can be controlled by throttling the inlet flow.

2.2.5 Design Considerations The design limitations for turbomachinery are usually the inlet temperature and the turbine tip speed. This can be correlated with the stresses in the turbine disk and blades. Another consideration is the corrosion properties of the working fluid, which can reduce the available material choices. Design charts suitable for preliminary design of single-stage turbines are available in the literature (Baljé, 1962). These charts show a relationship between the turbine tip diameter, volumetric flow rate, rotational speed, enthalpy change, and turbine efficiency. A more detailed presentation of turbine design can be found in Stodola (1927).

3. CONDENSERS

When the steam exhausts from the expander, it flows to the condenser where the cycle heat is removed and the water returns to its liquid state. There are two basic types of condenser: direct contact and surface. As the name implies, in the direct-contact condenser, cooling water is sprayed into the exhaust steam. These condensers are typically suited for smaller plants and were common in early marine steam-propulsion plants. The surface condenser is the type most commonly applied in modern plants. In this design, cooling water flows inside tubes, and the steam condenses on the outside of the tubes. In addition to provision for condensation, some condensers may include provision for cooling the expander exhaust to the saturation temperature (desuperheating) prior to condensation and cooling the condensate to slightly below saturated liquid temperatures (subcooling) in order to take into account pump design considerations. The condensed water collects in the hot well, which is usually at the lowest level of the plant.

Since the pressure in the condenser is generally considerably below atmospheric, there is a chance that air may leak into it, adversely affecting the heat-transfer characteristics. This air is continuously removed by a steam pump or a mechanical pump.

In many stationary applications, the coolant is relatively cold water, which is a good heat-transfer medium. Steam condensing pressures on the order of 10 kPa absolute are common, with temperatures often less than 30 °C. Where cooling water is scarce, cooling towers may be used, with the ambient air becoming the final heat sink. The cooling tower operates on the principle of evaporative cooling, so that the cooling water approaches the wet-bulb temperature of the air.

For smaller plants that use ambient air as the heat sink, the condenser is usually a fin-tube exchanger that resembles an automotive radiator, except that there are a greater number of fins in order to increase the heat-exchanger effectiveness.

Another condenser design consideration is the fact that the heat exchanger may be exposed not only to the subatmospheric pressure during most of the operating envelope but also to the average pressure of the system when it is shut down. This may be greater than atmospheric. As a minimum, this complicates the pressure-vessel aspects of the design.

4. PUMPS

The pump in a steam engine raises the pressure of the condensed water from the condenser pressure to the boiler inlet pressure. Since the pressure at the pump inlet may be below the boiling pressure at that temperature, a separate boost pump may be required between the condenser and the main feed-water pump.

As with expanders, there are two basic kinds of pumps, positive displacement and dynamic. No matter which type of pump, the key characteristics are the pressure rise and flow rate.

4.1 Positive Displacement

As with positive-displacement expanders, there are numerous types, each with its own characteristics. The overall operation can best be visualized by considering a piston moving in a cylinder. As the piston moves away from the head, water flows through the inlet (or suction) valve to fill the cylinder volume. As the piston starts its return stroke, the pressure begins to rise and the inlet valve closes. When the pressure has risen to the value set by the exhaust (or discharge) valve, the valve opens, and the piston forces the water out at the higher pressure. With a piston pump there are pulsations in the flow, and some type of capacitance is often introduced to allow the mean pressure to stabilize.

Other positive-displacement pumps work on the same principle, except that the pressure rise and flow characteristics differ for each pump type. A variety of gear pumps, vane pumps, and screw pumps have been applied to various pumping tasks, but for feed pumps for relatively low–power-level steam engines, the piston pump is the frequent choice.

4.2 Dynamic

The majority of dynamic pumps used in steam engines are centrifugal pumps as they provide the greatest pressure rise per stage. Axial pumps are used in some applications to achieve higher efficiency than possible with the centrifugal type but are rarely applicable to low-power steam engines.

In the centrifugal pump, water enters at or near the center of a vaned disk and flows outward toward the periphery. The velocity thus generated in the flow is converted into pressure in the volute (the space surrounding the disk) or, for higher efficiency, in diffuser vanes. The vanes are carefully shaped on both the rotating disk and the diffuser. The disk is also shaped so that the flow area tends to remain constant as the diameter increases.

Generalized design charts for dynamic pumps, based on similitude considerations, are available in the literature.

5. VAPOR GENERATORS

Vapor generators have two main subsystems: combustion, where fuel is burned to generate a high-temperature gas, and the heat exchanger or "boiler," which transfers the heat from the combustion gas to the water. In small power plants the boiler is usually a combined feed heater, boiler, and superheater. In it the pressurized feed water is heated to saturated-liquid conditions, the mixed water/vapor mixture is heated to saturated-vapor conditions, and additional heat is added to superheat the steam to the desired expander inlet conditions.

5.1 Heat-Transfer Subsystem

The simplest type of boiler is a heated pot with a pipe for the steam to escape—for example, a tea kettle. An advance over the pot boiler is the "fire-tube" design, where the combustion products are on the inside of tubes going through a large container filled with water. In some designs the hot gas makes two passes through the water, going first through a bank of tubes on the bottom of the container and then turning and going through a second set of tubes on the top portion of the boiler. The combustion products then exit through a smokestack. The heat transfer is from the hot gases through the tube walls to the water in the large container. The steam bubbles from around the tubes and rises to the surface of the water and accumulates in the space above the water level. The steam exits from the boiler through a "dry tube," which is designed to minimize the amount of liquid water carried over with the steam.

The fire-tube boiler was the most widely applied design for many years for both stationary and marine propulsion applications. As the desire to utilize steam at higher temperatures and pressures increased, the practical limit of boiler shell fabrication was reached at about 2.0 MPa.

In the "water-tube" design the combustion products are on the outside of the tubes and water is on the inside. A large number of configurations are available, the details of which depend on the intended application. For example, naval boilers must rapidly respond to varying steam demands, whereas merchant-ship boilers are usually designed with economy of operation in mind.

Most water-tube boilers use natural circulation of the water in the tubes, with the steam collecting in drums. Separate banks of tubes are placed relative to the combustion gas and the radiant walls to perform the feed-water heating, boiling, and superheater functions. In the bank of tubes in which boiling occurs, boiling is usually at low quality to minimize the effects of any unevenness in the heating, which could lead to unstable flow oscillations. (Quality is the term describing the proportion of vapor in a vapor/liquid mixture such as occurs during the boiling process.) Some boilers use a monotube design, in which all of the water flows through a single tube, entering as feed water, then boiling and superheating. An advantage of this design is flow stability. Forced-circulation boilers have a circulating pump to increase the velocity of the water in the tubes.

Modern vapor generators also frequently include an air preheater placed in the path of the exhaust combustion products after they flow over the water tubes. This can add to the overall efficiency by removing the need to heat the combustion air with primary fuel. A feed-water heater, called an economizer, may also be in the path of the by then cooler combustion gases.

Most modern boilers use quite pure water. The water chemistry is checked frequently to be sure that no impurities are introduced.

5.2 Combustion Subsystem

There are a number of important design considerations for vapor-generator combustion subsystems. If coal or other solid fuels are used, the concerns are the geometry of the furnace, as both radiant and convective heat transfer are important. The design depends on the type of coal used, as some are more volatile than others. For fire-tube boilers the grate area relative to the combustion-gas flow area becomes an important consideration. For liquid fuels, especially when low emissions are desired, the approach that has been found to satisfy the myriad requirements best is to have a premixed, prevaporized air–fuel mixture, with a lean fuel/air ratio.

6. APPLICATIONS

6.1 Electric-Power Generation

Large electrical-generating plants are individually designed taking into account the available fuel, the climate, and the economic conditions in the service area as well as numerous technical details. Design decisions are usually based on thorough economic analysis to insure that any additional increase in efficiency is worth the investment. Only the rudiments of this application of steam power are presented here, as a comprehensive treatment is beyond the scope of the article.

Boilers of central stations are designed to extract as much of the thermal energy in the combustion products as possible. The feed water enters the boiler at close to the boiling point, as it already has been heated in several stages of feed-water heating. Boiling usually takes place at low quality in order to assure stable flow. The steam collects in drums. Separate tube sections for superheat and reheat are placed in an appropriate place to insure efficient heat transfer. An air preheater is commonly applied in order to preheat the combustion air.

The expander is a multistage turbine. Some of the steam is extracted at different portions of the expansion and used to heat the feed water on its way to the boiler. Care is also taken to extract any water droplets that form as the steam expands and cools, as they can cause erosion of the blades.

Condensers for large, steam power plants are of the surface type and are typically placed directly below the turbine.

Feed pumps are multistage centrifugal pumps. Since the water exiting from the condenser is below atmospheric pressure, a boost pump is placed ahead of the main feed pump. This pump brings the pressure to a point where the water can enter the main feed pump without vapor bubbles forming.

Many large plants are usually designed to operate at a fixed power level. In large systems comprising many generating plants, there are usually specially designed plants to take care of the varying load.

An application that is increasingly widespread is the cogeneration concept. In these plants the exhaust from a gas turbine supplies the heat source for a steam vapor generator. The steam is then used either for process heat or for generation of electricity. Although there is a significant capital investment required because there are essentially two power plants, the overall efficiency has often been found to be high enough to warrant the extra expense.

6.2 Stationary Power Plants

Stationary power plants are often designed using commercially available components. The design is based on the power-level needs, the need for steam in the plant process, the available fuel, and whether the plant will be operating seasonally such as in some agricultural processing applications or full time as in a manufacturing facility. Historically, many local plants produced shaft power that was transmitted throughout the factory via leather belts. This has been largely replaced by electrical power being distributed with local motors.

6.3 Motor-Vehicle Propulsion

The earliest propulsion application of steam occurred in 1770, when Cugnot in France designed a steam-powered vehicle intended to haul artillery. Steam-powered road vehicles were built in the early 1800s. By the early 1900s, steam-powered automobiles were being produced in significant numbers. Serpollet in France and Stanley and White in the United States developed sophisticated engines operating at high temperatures with complex expanders. Serpollet claimed success in lubricating his expanders at 500 °C. The Stanley vehicles used a two-cylinder, double-acting, piston expander and a fire-tube boiler. The White vehicles used a two-cylinder, compound-expansion, double-acting, piston expander and a monotube boiler. Steam conditions at the vapor generator outlet of both the Stanley and White engines were in the 370 °C, 4.1-MPa range. Production of the Stanley Steamer ended in 1925. Some efforts continued into the 1930s, with designs by Doble, who built passenger cars with four-cylinder, compound-expansion, double-acting, piston expanders and monotube boilers. The expander inlet conditions were typically 400 °C, 5.2 MPa. Although there were some further efforts at steam-powered vehicles, interest waned as internal-combustion engine design matured.

Another burst of activity in steam engine development began in the late 1960s, when concerns over air pollution became a significant issue. Programs sponsored by the U.S. Environmental Protection Agency, the State of California, General Motors, and many others culminated in the demonstration of several sophisticated engine designs in both automobiles and busses. Very compact vapor generators with very low emissions were demonstrated, with peak conditions of 590 °C and 6.9 MPa in several designs.

Fuels for automotive steam engines have included kerosene, diesel fuel, and gasoline. Fuel atomization to about 10 μm is necessary for low emissions, as large droplets have a "wake" that burns at stoichiometric conditions. The mixture must be uniformly prepared, and there must be good flame stabilization, sufficient time for completion of combustion, and precise control during transients.

Although most early steam vehicles did not employ condensation, in order to be competitive with modern spark ignition engines, the experimental steam engines of the 1970s were designed with condensers. Unfortunately, the air neither is a good heat-transfer medium nor is it necessarily very cold. In addition, the frontal area of the vehicle may be the limitation that determines the condenser size and thus what the condensing-state point will be. These practical limitations often mean that the condensing pressure is above atmospheric, on the order of 140 kPa absolute, with temperatures on the order of 100 °C. These conditions have a major effect on the attainable cycle efficiency.

Although Rankine-cycle engines can attain very high efficiencies, requiring such an engine to fit within the available volume required for automotive propulsion, and with weight similar to current internal-combustion engines, severely limits the attainable efficiency and thus fuel economy.

When fuel economy became an important issue in the mid-1970s, interest in steam engine power for automobiles ceased. There was a short-lived interest in using the heat remaining in the exhaust of large over-the-road diesel engines as the heat source for a Rankine-cycle waste-heat recovery plant (bottoming cycle). Demonstration devices were built but were not commercialized.

6.4 Rail Propulsion

The basic design of the locomotive steam engine remained very much the same from its inception in the 1830s until it was largely replaced by diesel engines in the post World

War II era. The details of the design varied from country to country, taking into account local conditions, such as fuel availability, the type of terrain the locomotive would traverse, the limitations on width and height due to existing structures, and the type of loads it would carry.

Except for a few experimental locomotives, locomotive boilers were of the fire-tube design, with pressures in the neighborhood of 1.4 to 2.0 MPa. Early boilers were designed to produce saturated steam. In the early 1900s a practical superheater design was developed, and its application became widespread by the 1920s. In this superheater design, saturated steam was led through small-diameter U-shaped pipes placed in enlarged fire tubes. The usual diameter for the ordinary fire tubes was about 5 cm, whereas the tubes in which the superheater elements were placed were about 12 cm in diameter. The use of superheated steam eliminated the problems with condensation of the saturated steam as it entered the cylinder (initial condensation) as well as providing an increase in cycle efficiency.

An interesting feature of the steam locomotive is the close coupling between the combustion and steam expansion processes. The steam exhausted from the working cylinders goes to a blast pipe and thence out through the smokestack. This flow tends to act as a pump for the combustion products, with increased steam flow creating additional draft that is a reduced pressure at the outlet end of the fire tubes. This in turn draws more air through the combustion region of the boiler, which increases the thermal output of the furnace and increases the amount of steam being produced.

Until very late in its history, the primary fuel for steam locomotives was coal. Wood was used on some locomotives, and toward the end of their active usage, some locomotives used oil.

The most common locomotive expander configuration employed two double-acting simple expansion cylinders. Some three- and four-cylinder locomotives saw service. Some engines with two stages of expansion were used, but the increased efficiency did not prove attractive.

Early steam locomotives used slide valves. With the advent of superheated steam, these were generally replaced by cylindrical valves, which were easier to lubricate in the dry-steam environment. Toward the end of the steam era, some locomotives used poppet valves.

In order to turn the reciprocating motion of the piston into rotary motion of the wheels, the piston was attached to a crosshead, which in turn had a connecting rod to the drive wheel. The drive wheel was connected to the remaining wheels with coupling rods.

Although steam locomotives reigned supreme for more than 100 years, and were economically viable, they never were very efficient in the thermodynamic sense. The reason for this is that the noncondensing basic cycle with quite moderate temperatures and pressures were not more than about 5% efficient in converting the thermal energy in the fuel to work at the wheel. There were some experimental locomotives that used water tube boilers, some that condensed the steam, some that used turbine expanders, and some with exotic variants of the basic Rankine cycle; however, these engines did not prove practical.

6.5 Ship Propulsion

In 1785, James Fitch built a 10-m steam-powered boat propelled by vertical oars. The first steam-powered vessel that can be considered the progenitor of modern design was demonstrated in 1801 by William Symington. His vessel had a single, double-acting horizontal engine directly driving the crankshaft of a stern paddle wheel. The cylinder was 0.56 m in diameter and had a 1.2-m stroke.

Early marine steam engines used saltwater as the working fluid with the pressure at the vapor generator outlet in the range of about 13–20 kPa above atmospheric. The boiler had to be cleaned periodically as deposits built up as a result of the use of saltwater. A remedy to this was found by replacing a small portion of the water every few hours. Superheat was introduced in 1856. Direct-acting condensers were the usual design in the earlier applications, and surface condensers came into service in the 1870s.

Boilers in the early days of marine propulsion by steam were flue boilers; that is, the combustion products followed a tortuous path around walls filled with water. These walls were about 13 cm thick. It was not long before fire-tube boilers came into use, eventually followed by water-tube boilers. Very compact designs of water-tube boilers were developed. Boilers for naval combat vessels were developed so that rapid changes in steam rates were possible.

Triple-expansion reciprocating expanders were the most common design until the steam turbine became the expander of choice. Some of the triple-expansion expanders had three cylinders, and some had four. In either case, the intermediate- and low-pressure cylinders were larger in diameter than the high-pressure cylinder to accommodate the increased volume of the steam as it expanded.

7. OTHER WORKING FLUIDS

Water is not the only possible fluid to use in a Rankine-cycle power plant. There are a number of reasons to consider alternatives. Water freezes at a temperature commonly encountered during part of the year in many parts of the world, whereas other fluids can remain liquid at subzero temperatures. The properties of water are not especially attractive for low–power-level Rankine-cycle engines, as the high enthalpy change per unit mass results in relatively low overall flow rates. Efficient expanders may be difficult to design at these flow rates. Probably the most intriguing reason for considering other fluids is that, at least theoretically, it is possible to design a fluid for a particular application.

On the other hand, many fluids that would be desirable from a thermodynamic, heat-transfer, or expander design viewpoint are toxic, flammable, or with poor thermal stability, and in some cases very costly. Another problem with alternative working fluids is that at the higher molecular weights, the saturated-vapor line has a positive slope on the temperature–entropy diagram. This means that expander exhaust goes far into the superheat region, and thus a considerable amount of superheat must be removed prior to condensing. In practice this can mean that a large heat exchanger would be required, possibly negating any other weight or size advantages the fluid might show.

As the molecular weight of a fluid increases, the enthalpy change per unit mass decreases, which for turbomachinery design means that the nozzle velocity (but not necessarily the Mach number) will decrease, since it is proportional to the square root of the enthalpy change. The result is that turbomachinery with higher molecular-weight fluids can be designed to operate at slower rotational velocities. For the same power level, however, there will tend to be greater mass flow, which will tend to increase the size of the system heat exchangers.

Designing for fluids other than water can be difficult because the thermodynamic and fluid-mechanical properties may not be known with the same degree of precision. However, assuming a perfect gas for the expansion can provide a preliminary idea of the dimensions and rotational velocity for a given power level if a turbine expander is being considered.

A closer approximation to the thermodynamic properties of the working fluid can be obtained by using the principle of corresponding states. Rather than the usual expression relating the pressure, specific volume, and temperature of a perfect gas, the equation

$$pv = ZRT$$

can be used, where Z is the compressibility factor, usually defined as the ratio of the actual fluid volume to the volume that the same fluid would occupy at the same temperature and pressure if it were an ideal gas; p is pressure; v is specific volume; R is the gas constant for the particular gas; and T is the absolute temperature. Z is a highly nonlinear function of the temperature and pressure. Generalized charts of Z are available based on the reduced pressure and temperature, that is, the ratio of the actual pressure and temperature to the critical pressure and temperature. Even if there are meager data on the particular fluid, the critical point is generally known.

A vast number of fluids other than water have been demonstrated in Rankine-cycle plants. Mercury was demonstrated in an experimental central station. Mercury and other liquid metals were also demonstrated in space-power applications. A variety of other fluids were demonstrated or at least considered, including toluene, mixtures of trifluoroethanol and water, hexaflourobenzene and mixtures of hexaflourobenzene and perflourotoluene, and dichlorobenzene, as well as proprietary fluids. The choice was typically made on the basis of the thermal decomposition and corrosion properties of the fluid in contact with the materials that would be used for the mechanical parts of the system.

8. OTHER CYCLES

Depending on where the compression, heat addition, expansion, and heat rejection occur with respect to a fluid's vapor dome, a number of cycle variations are possible (Angelino and Moroni, 1973). One variant would have the entire cycle take place in the supercritical region. In another, heat addition would take place in the supercritical region, but expansion and condensation as in the conventional Rankine cycle. The cycle can also take place entirely in the superheat region, which in the case of air would be the nearly perfect gas region. This cycle is actually the Brayton cycle (see HEAT ENGINES AND REFRIGERATORS) applicable to gas turbine engines. If the compression takes place close to the saturated vapor curve, large real-gas effects may be introduced. An even more exotic cycle can take place when the properties of the working fluid change with temperature and pressure. The thermally dissociating cycle is an example of this type of cycle.

GLOSSARY

Boiler: The device in which combustion products vaporize the steam. This may also include superheat and reheat.

Compound Engine: An engine with two stages of expansion.

Condenser: The device in which heat is removed from the steam after expansion, and the steam returns to a liquid state.

Connecting Rod: The shaft that transmits the force from the crosshead or the piston rod to the crankshaft.

Counterflow Expander: An expander in which the steam enters the cylinder at one end of the cylinder and exhausts at the same end. (See **Uniflow.**)

Crankshaft: A shaft upon which cranks are placed. The action of the connecting rod against the cranks causes the shaft to move in rotary motion.

Crosshead: A sliding bearing that is intermediate between the piston rod and the crank arm of the crankshaft. The crosshead allows the piston rod to move in reciprocating motion while it executes the complex motion required to transmit the force generated by the piston to the crankshaft.

Cutoff: The point where steam admission into the cylinder ceases. Some engines have variable cutoff as a means of controlling power output.

Cylinder: The component in which the steam expansion takes place in a positive-displacement expander.

Enthalpy: A property of a substance equal to the sum of the substance's internal energy and the product of its pressure and specific volume. Changes in enthalpy are governed by the initial and final states rather than by the path between the states.

Expander: The device in which the steam transforms thermal energy into work. In a positive-displacement expander, the work is the result of an unbalanced pressure across a piston, vane, or other similar component. In a dynamic expander, the work is the result of a momentum change as steam exits a nozzle at high velocity and impinges upon movable blades.

Feed-Water Heater: A heat exchanger in which heat is added to the pressurized water prior to boiling.

Fire-Tube Boiler: A vapor generator in which the products of combustion flow through tubes surrounded by water. (See **Water-Tube Boiler.**)

Full Admission: The design of dynamic expander wherein steam flows through all of the stationary nozzles. (See **Partial Admission.**)

Impulse Turbine: A dynamic expander in which the steam expands in a stationary nozzle and impinges on a set of blades, transferring momentum. (See **Reaction Turbine.**)

Indicator Card: A pressure–volume diagram of the process within a positive-displacement steam expander.

Partial Admission: The design of a dynamic expander in which steam is admitted to only a subset of the full circumference of the turbine blades. Partial admission can be used as a power-level control by having more than one set of blades available for admitting steam to the rotating turbine disk. (See **Full Admission.**)

Piston: A disk upon which the steam acts in a positive-displacement expander. The steam exerts a force on one side of a single-acting piston and on both sides of a double-acting piston.

Piston Rod: The shaft that is connected at one end to the piston and at the other end to the crank or crosshead.

Pump: A device that raises the pressure of the water. A circulating pump causes a fluid to flow through a heat exchanger—for example, the cooling water through the tubes of a condenser. A feed pump raises the pressure of the water prior to heat addition. A boost pump raises the pressure of the water from the condenser to the main feed-pump inlet.

Reaction Turbine: A dynamic expander in which the steam expands in stationary nozzles and continues to expand in rotating blades. (See **Impulse Turbine.**)

Reheat: Steam that has been partially expanded and returned to the vapor generator for further heat addition.

Superheat: The state of the steam when it is heated to a temperature greater than its saturation temperature.

Superheater: The portion of the vapor generator in which additional heat is added to saturated steam in order to raise its temperature.

Triple-Expansion Engine: An engine with three stages of expansion. Three-cylinder and four-cylinder versions of this engine type were used. This engine type was commonly used in marine propulsion applications.

Uniflow Expander: An expander in which steam enters the cylinder at one end of the cylinder and flows in one direction toward exhaust ports. This design minimizes initial condensation. This type of engine also has a fairly constant part-load efficiency. (See **Counterflow Expander.**)

Valve: A mechanical device that opens or closes a flow path. Slide valves slide back and forth over slots in the cylinder, allowing steam to enter the cylinder or to be exhausted from it. They are usually D shaped, with the flat side of the D sliding across the cylinder and the round part of the D in communication with the steam inlet and exhaust. Cylindrical valves have a similar action to slide valves, except they are cylindrical. Corliss valves are rotary quick-acting valves of a unique design. Poppet valves are mushroom-shaped valves that move up and down to allow steam into or out of the cylinder. This is the common design found in internal-combustion engines.

Vapor Generator: The device in which water is heated, boiled, and superheated by combustion products. The term boiler is often used as a synonym. The complete vapor generator includes the furnace in which the fuel is burned, blowers and fuel admission components, and other heat exchangers that may be present such as air preheaters and economizers, as well as controls for the various subcomponents.

Water-Tube Boiler: A vapor generator in which the products of combustion flow across water or vapor-filled tubes.

Works Cited

Angelino, G., Moroni, V. (1973) *J. Eng. Power* **95** (2), 75–83.

Baljé, O. E. (1962), *J. Eng. Power* **84** (1), 83–114.

Stodola, A. (Translated by L. C. Lowenstein) (1927), *Steam and Gas Turbines,* New York: McGraw-Hill.

Further Reading

Luchter, S., Renner, R. A. (1977), *An Assessment of the Technology of Rankine Engines for Automobiles,* U.S. Energy Research and Development Administration, Report Number ERDA-77-54.

Stodola, A. (Translated by L. C. Lowenstein) (1927), *Steam and Gas Turbines,* New York: McGraw-Hill.

Stepanhoff, A. J. (1948), *Centrifugal and Axial Flow Pumps,* New York: Wiley.

Taylor, C. F., Taylor, E. S. (1961), *The Internal Combustion Engine,* 2nd Ed., Scranton, PA: International Textbook Company.

Tuplin, W. A. (1974), *The Steam Locomotive, Its Form and Function,* New York: Scribner's.

United States Naval Institute (1941), *Naval Machinery,* Annapolis, MD.

STEEL

H. P. Hougardy, *Max-Planck-Institut für Eisenforschung GmbH, Düsseldorf, Germany*

Introduction 579
1. **Some Properties of Pure Iron** 580
2. **Phase Diagrams with Iron** 581
3. **Factors That Determine Properties of Steels** 584
3.1 Yield Stress 584
3.1.1 Definitions 584
3.1.2 Influence of Grain Size 586
3.1.3 Influence of Solid Solution 587
3.1.4 Influence of Particles 587
3.1.5 Influence of Dislocation Density 587
3.2 Toughness 588
3.3 Creep Resistance 589
3.4 Fatigue Strength 589
3.5 Corrosion and Oxidation Resistance 589
4. **Microstructures of Steels with α–γ–α Transformation** 591
4.1 Definitions 591
4.2 Austenite 592
4.3 Ferrite and Pearlite 592
4.4 Martensite 593
4.5 Bainite 594
4.6 Microstructures after Tempering 595
5. **Grain Size and Recrystallization** 595
6. **Correlations between Microstructures and Properties** 596
7. **Description of the α–γ–α Transformation** 597
7.1 Austenitization 597
7.2 Transformation during Cooling 598
8. **Technical Steel Production and Treating** 600
8.1 Steelmaking 600
8.2 Rolling 601
8.3 Heat Treating 601
8.3.1 Influences on Transformation 601
8.3.2 Austenitization 602
8.3.3 Effect of Dimensions 602
8.3.4 Transformation in the Pearlite Range 603
8.3.5 Transformation in the Bainite Range 603
8.3.6 Transformation in the Martensite Range 603
8.3.7 Quenching and Tempering 604
8.3.8 Annealing for Spheroidization 604
8.3.9 Surface Hardening 604
8.3.10 Simulation of Heat Treatment 604
9. **Steel Grades** 605
9.1 Low-Alloy Steels 605
9.2 Heat-Treated Engineering Steel 605
9.3 Tool Steels 606
9.4 Stainless Steels 606
10. **Standards** 607
Works Cited 607
Further Reading 610

INTRODUCTION

Iron is the chemical element Fe, but the word is also used for an alloy of Fe with low contents of Mn, Ni, and Cr. Iron has a strength slightly above that of copper. By addition of carbon up to 2 wt% and other elements, the strength can be increased by a factor of 10; this is the range of steels that can be hot and cold deformed. With carbon contents above about 2 wt%, hot and cold deformation is nearly impossible; this is the range of cast iron that in general is used in the cast state (Elliott, 1992).

Native meteoric iron has been used for at least 5000 years. Reduction from iron ores has been feasible for about 4000 years, but because of the high melting point of iron

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

(about 1500 °C), iron and steel could only be produced in the solid state. This was very time consuming, and the volume of steel that could be produced was limited. With improved techniques using water power in the 13th century, temperatures about 1200 °C could be reached, giving liquid pig iron with 2.5–5 wt% C from a blast furnace with a higher capacity than earlier processes. Furnace plates or cannon balls could be cast directly, but on the extraction of carbon by oxidation to the range of forgeable steel, i.e., to below 2 wt% C, the melting point still increases above the temperature that could be realized at that time, and steel production was possible only in the solid state. Iron production at that time in Europe was about 25 000 tonnes per annum. Today one blast furnace produces about 10 000 tonnes per day. People knew nothing about crystal structure, phase transformation, or mechanisms of hardenability. So much the more admirable in retrospect is what was made of steel. One example is the pattern of the Damascus swords (Fig. 1). The striation was realized by welding and forging together steel plates with high and low carbon contents, giving different reflectivity after etching (Sachse, 1994). Patterns such as that shown in Fig. 1 were reproducible; they were not a matter of chance. Such swords are—from today's point of view—composites built up by hard and soft layers of steels giving an excellent combination of high strength, high hardness, and good toughness. Very modern methods such as the creation of compressive residual stresses at the surface were already realized with the finishing of such swords.

With the invention of air refining in a converter and air preheating, i.e., regenerative heating, about 1860 temperature within the furnaces could be raised to produce steel in the liquid state. By combination of this with steam power, producing sufficient energy, the production of steel was enlarged, among other things, one of the requirements for the development of railways. Starting with that time, up to today as a result of investigations of the processes of steel production and heat treatment, not only the quality of steels was improved but also the reproducibility of a defined steel grade. Today more than 2000 different steel grades are available, giving a wide range of applications as structural steels, corrosion-resistant steels, tool steels, or materials with special physical properties. Today the worldwide production of steel is about 700 million tonnes per annum, about 30 000 times more than in the 13th century.

In the following, starting with the properties of pure iron and an introduction to some alloy systems, the strengthening mechanisms of steel and the basic microstructures are discussed. A discussion of the principles of the transformation of steels is followed by a short description of steel production and heat treatment. Engineering steels, tool steels, and stainless steels are taken as examples to give some information on special steel grades. In all cases only basic information about steel production and properties is given. Excellent books are available describing details, and are listed in Further Reading.

1. SOME PROPERTIES OF PURE IRON

Except on a laboratory scale, it is not possible to produce an alloy with 99.999% iron. A typical analysis of commercial pure iron is given in Table 1. The interpretation of such an analysis is not easy, for today we know that a content of 15 ppm B is already a boron-alloyed steel. But such a low level could not be analyzed 50 years ago. Therefore, the values given below and elsewhere refer to an iron as pure as available at the time; experiments are done or are extrapolated from alloyed steels to an alloy content of zero.

Iron forms two crystal structures, the body-centered cubic (bcc) and the face-centered cubic (fcc) (see CRYSTALLOGRAPHY). The bcc α phase is stable at room tempera-

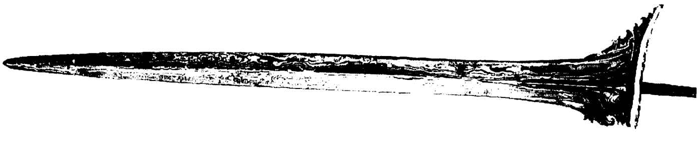

FIG. 1. Part of a dagger with Damascus pattern (Horstmann, 1983).

Table 1. Some elements of the chemical composition of commercial pure electrolytical iron.

Element	Wt%
Fe	>99.9
Si, As, Cu	>0.001
Mn, Cr, N	0.0015
Pb, Zn	0.0004
P, S	0.0001
B	0.0002

ture and ferromagnetic up to the Curie temperature of 768 °C. The transition of the ferromagnetic state to the paramagnetic state at the Curie temperature often is referred to as the A_2 temperature, but it is a transformation of second order; see MAGNETS; MAGNETIC MATERIALS; and PARAMAGNETISM. At 911 °C, the A_3 temperature, the bcc α iron changes to the fcc γ iron. At 1392 °C, the A_4 temperature, the fcc γ iron changes to the bcc δ iron, which is stable up to the melting point of 1536 °C. The bcc lattice has two atoms per unit cell; the fcc lattice, four atoms per unit cell. Therefore, the lattice constants of α iron and γ iron are very different; see Table 2, which also gives values for the density. The atomic volume of iron (Fig. 2) indicates the change of the structure with temperature. The heat capacity shown in Fig. 3 has a maximum at the Curie temperature T_c. The heat of transformation is 820 J/mol for the α–γ transformation and 850 J/mol for the γ–δ transformation. The heat of melting is about 14 000 J/mol.

Values for the thermal expansion are given in Table 2. The magnetization curve for single crystals at room temperature has the same saturation polarization of 2.16 T at 20 °C for all directions, but below saturation there are differences by a factor of 2 between the (100) and (111) orientation. This effect is used in the production of textured steels for electric machines. The electrical resistance is 10 $\mu\Omega\cdot$cm at 20 °C. The ratio of the resistivity at 273 K to that at 4.2 K is 5000 for iron with 99.999 wt% Fe. This value is used to describe the purity of iron. The thermal conductivity is about 70 J/ms·K at 20 °C. In the visible spectrum the optical reflectivity of iron is about 0.55. The value depends on the quality of the surface. For mechanical properties, see Sec. 3.

The physical properties of pure iron are changed severely by alloying elements. So, fcc γ iron may be stable at room temperature, or bcc α iron may be stable up to the melting point (see Sec. 2), giving ferromagnetic as well as nonmagnetizable steels. In an iron–nickel alloy with about 35% Ni, the thermal expansion coefficient is extremely low, about 1×10^{-6} K^{-1}. In this case magnetostriction compensates thermal expansion. On the other hand, an alloy with 0.6 wt% C, 1 wt% Si, 6 wt% Mn, and 15 wt% Ni has a thermal expansion coefficient as high as 20×10^{-6} K^{-1} at room temperature. Any value in between can be produced (Pepperhoff, 1992; Stäblein and Arntz, 1993; Weßling and Heimann, 1993).

2. PHASE DIAGRAMS WITH IRON

Steels are alloys of iron with elements such as C, Mn, Si, Cr, Ni, Mo, W, etc. For the understanding of the kinetic processes, the knowledge of the phase equilibria is a necessary basis. For the principles of phase diagrams and solid solutions, see THERMODYNAMICS, EQUILIBRIUM and PHASE EQUILIBRIA. Figure 4 shows the system Fe–Mn as an example for alloying elements such as Ni, C, and Mn, which widen the temperature range where γ iron is the stable phase. The

Table 2. Some physical data of pure iron. The coefficient of thermal expansion is the average for the temperature range given in the same row.

Temperature (°C)	Structure	Number of atoms per unit cell	Lattice constant (nm)	Density, calculated (g/cm^3)	Coefficient of thermal expansion (K^{-1})
20	bcc	2	0.286 62	7.876	15.3×10^{-6}
911	bcc	2	0.290 41	7.572	
911	fcc	4	0.364 62	7.652	22.0×10^{-6}
1392	fcc	4	0.368 76	7.397	
1392	bcc	2	0.293 19	7.359	18.5×10^{-6}
1536	bcc	2	0.294 13	7.288	

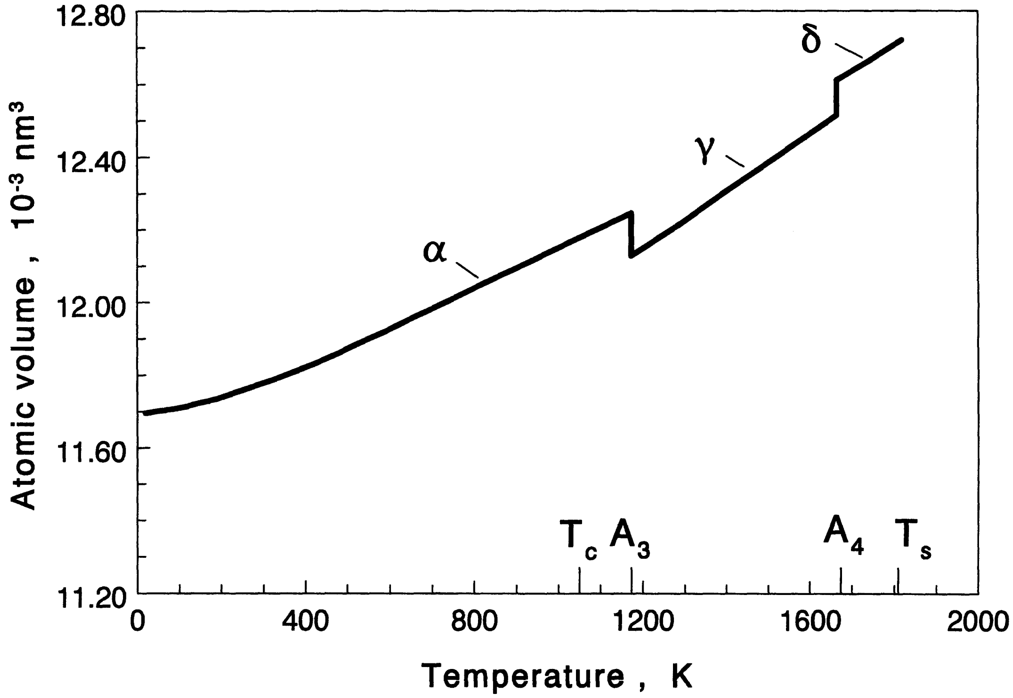

FIG. 2. Atomic volume of pure iron as a function of temperature. T_c, Curie temperature. Transformation temperatures: A_3, α–γ; A_4, γ–δ. T_s, melting temperature. (Pepperhoff, 1992).

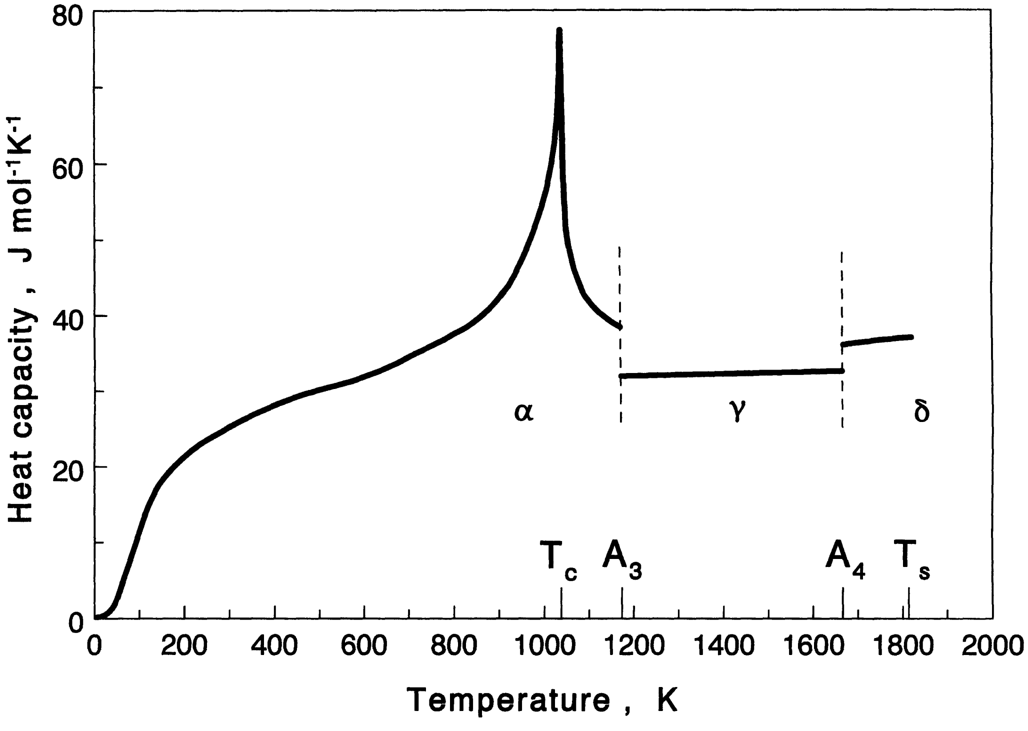

FIG. 3. Heat capacity of pure iron as a function of temperature. Characteristic temperatures, see Fig. 2 (Pepperhoff, 1992).

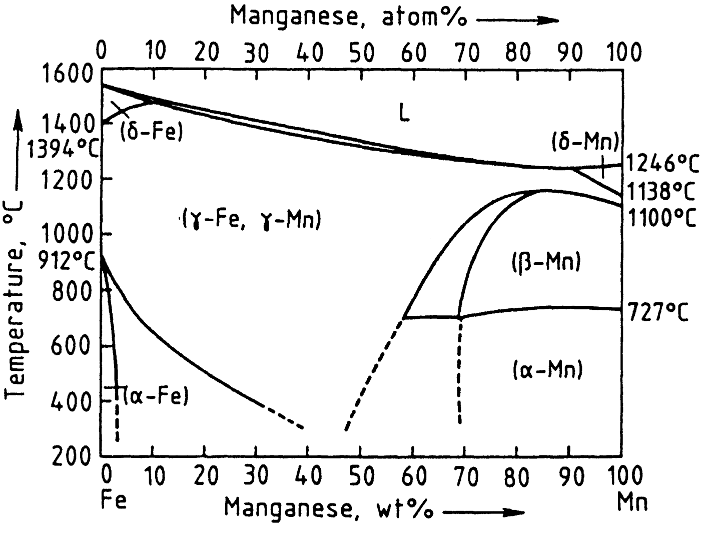

FIG. 4. Iron–manganese phase diagram (Massalski, 1990).

opposite effect is produced by elements like Cr and Si, which extend the temperature range where α iron is the stable phase (Fig. 5). In iron with more than 13 wt% Cr, ferrite is stable between room temperature and the solidus temperature. However, the effects of the alloying elements are not additive. As an example, in a steel containing 18 wt% chromium and 10 wt% nickel, γ iron is stable from room temperature to the melting point. Phase diagrams can be calculated from thermodynamic data (Predel, 1993). Compilations of the most important two- and three-component systems are published (Massalski, 1990; Petzow and Effenberg, 1991; Raghavan, 1987–1992).

In plain carbon steels, the metastable phase Fe_3C, called cementite, is of great im-

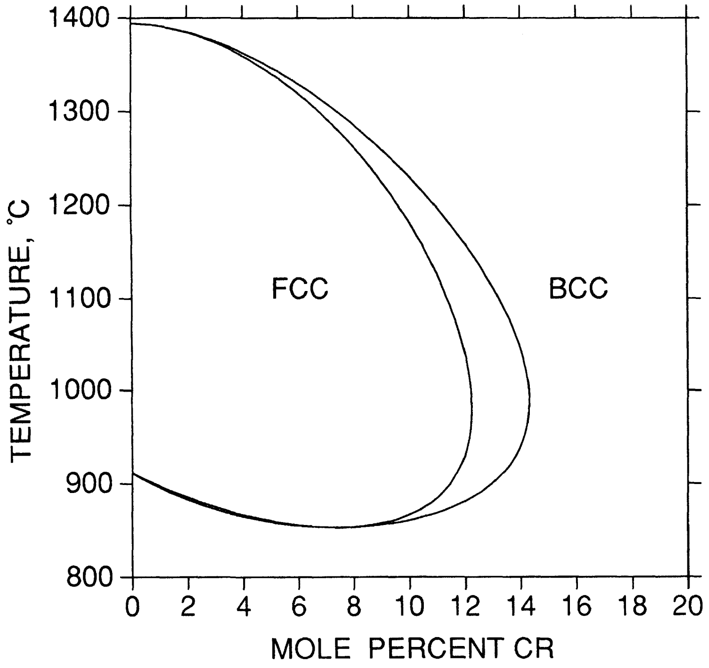

FIG. 5. Section of the iron–chromium phase diagram calculated with the program Thermocalc (Franke, 1995).

portance (Fig. 6). In the system Fe–C, the stable phase is graphite, formed in gray cast iron. These are alloys of Fe, e.g., with carbon contents of 4 wt% C, 1 wt% Mn, and 2 wt% Si. Under technical conditions of heating and cooling in steels, cementite is stable, especially if a solid solution $(Fe,Mn)_3C$ is formed in the presence of the usual manganese content in steels (see Solubility, Alloys). In Fe-Cr-C systems, $(Fe,Cr)_3C$ is stable if more than about 0.7 wt% chromium is present. Figure 7 shows a quasibinary section through an Fe-M-M_3C system, where M represents a metal. For technical applications, the temperature limits of the phase fields are measured during heating with 3 °C/min. For these temperature limits the technical term "Ac temperature" is used. The relationship of the Ac temperatures to the boundaries of the phase fields can be seen from Fig. 7. Alloy steels, which are multicomponent systems containing Fe and C, may contain not only M_3C but also the carbides M_7C_3, $M_{23}C_6$, M_6C, M_2C, MC, and intermetallic phases. In these cases, the simple scheme shown in Fig. 7 is not valid (Pitsch and Sauthoff, 1992).

3. FACTORS THAT DETERMINE PROPERTIES OF STEELS

3.1 Yield Stress

3.1.1 Definitions. In a uniaxial tension test, with increasing load first there is a reversible, linear, elastic deformation with $\sigma = E\epsilon$ N/mm^2 (see Fig. 8). E is the Young's modulus, which for α iron is in the range of 210 000 N/mm^2. The stress σ is defined by $\sigma = K/F_0$, if K is the load and F_0 the initial cross section of the cylindrical specimen. If l_0 is the initial length and Δl the elongation under load, then the strain is defined as $\epsilon = \Delta l/l_0$. At loads higher than the so-called yield stress, the correlation between stress and strain is no longer linear, and plastic deformation starts with irreversible glide of dislocations; see Point and Extended Defects in Crystals. During deloading from point a in Fig. 8, the strain goes back following the dotted line. Point b represents the plastic deformation. The slope of the line a–b is E. Loading again, the material deforms elasti-

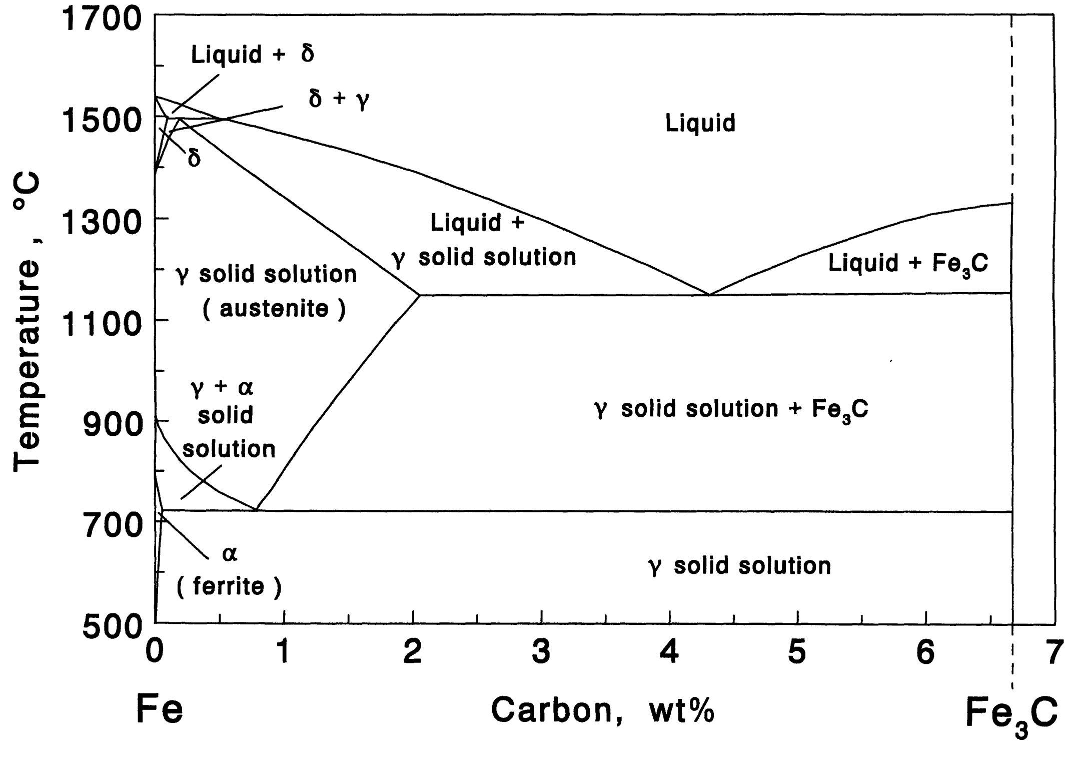

FIG. 6. Phase diagram of the metastable equilibrium iron cementite (Massalski, 1990).

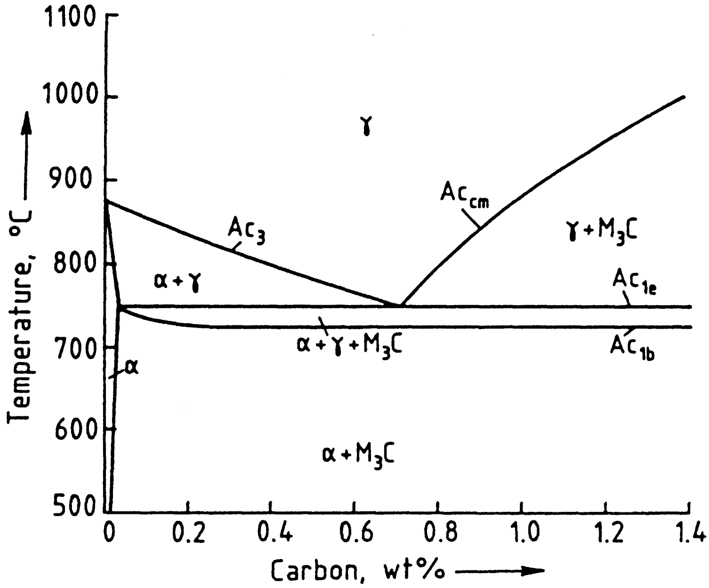

FIG. 7. Quasibinary section of a ternary phase diagram Fe-*M*-M_3C, where *M* represents a metal like Cr, Mn, etc., and M_3C stands for (Fe,*M*)$_3$C (Hougardy, 1990).

cally following the line *b–a*. At loads greater than *a* the stress–strain behavior follows the original curve. For many steels the yield stress cannot be measured with sufficient accuracy. Therefore, the stress giving 0.2% permanent deformation is taken as proof stress, $R_{p0.2}$ in Fig. 8. The increase of yield stress *a* above $R_{p0.2}$ is called work hardening. In iron, dislocations have a density of about 10^8 cm^{-2}; this is the total length of dislocations in cm/cm^3. After plastic deformation the density may go up to 10^{12} cm^{-2}. With increasing load, the cross section of the test specimen decreases; therefore, the true strain ϵ_W must

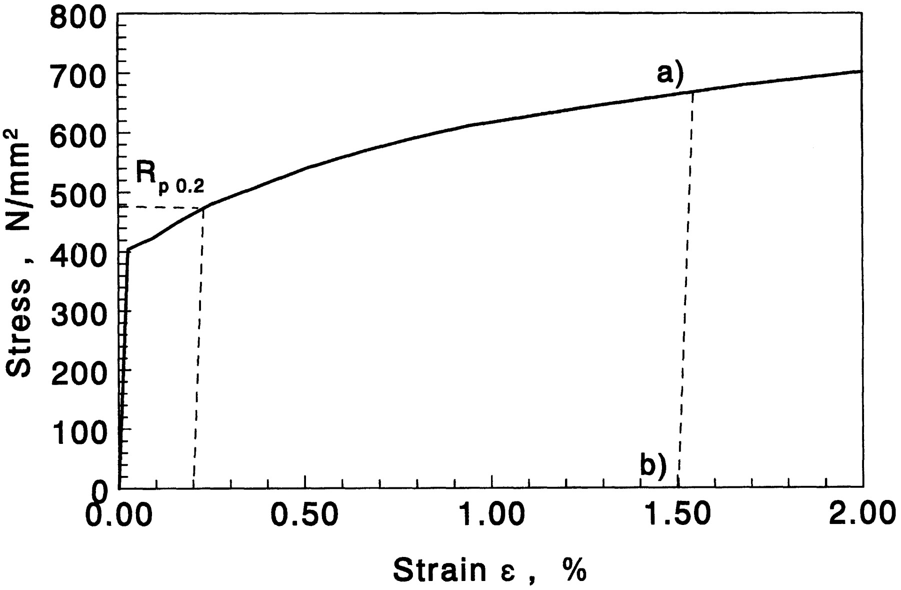

FIG. 8. Initial part of a stress-strain diagram of a steel as obtained by a tensile test.

be calculated by $\epsilon_W = \ln(F_0/F)$, if F is the actual cross section. The subscript denotes the true strain and the true stress, which is defined as $\sigma_W = K/F$. At a stress above $R_{p0.2}$, a first approximation yields $\sigma_W = \sigma_0 + K\epsilon_W^n$. At $\sigma_W = d\sigma_W/d\epsilon_W$ necking of the specimen starts, and the strain increases with decreasing load until rupture; see Fig. 9. After rupture of the specimen there is a reduction of area described by Z, with $Z = (F_0 - F_E)/F_0$, if F_E is the cross section at the fracture surface, and an elongation after fracture $A = (L - L_0)/L_0$, if L is the length after fracture. The maximum stress in the σ-ϵ curve before necking starts is called tensile strength (Fig. 9).

Since the plastic deformation is caused by the movement of dislocations, the yield stress and the exponent n depend on temperature and loading rate. Figure 10 shows the stress–strain curves for a plain carbon steel with 0.45 wt% C for various temperatures. Increasing the strain rate from $\dot{\epsilon} = 10^{-4}\ \text{s}^{-1}$ up to $\dot{\epsilon} = 10^2\ \text{s}^{-1}$ may raise the yield stress by a factor of 1.8.

Pure iron has a yield stress at room temperature of about 100 N/mm^2. The yield stress of high-purity iron with 99.999 wt% Fe is about 50 N/mm^2 at 273 K but about 600 N/mm^2 at 4.2 K, measured with a strain rate of $8 \times 10^{-5}\ \text{s}^{-1}$. For technical applications, steels often should have higher strength to reduce the necessary cross section in constructions or to increase wear resistance. There are four ways of increasing the strength of pure iron. Theoretically the increase of a shear stress is calculated, but in the following this is transposed to an increase in the yield stress measured in a uniaxial tensile test (Dahl, 1992; Mughrabi, 1993; Cahn and Haasen, 1996).

3.1.2 Influence of Grain Size. If d is the average diameter of ferrite grains in the microstructure, the yield stress R_{el} is increased by $\Delta R_{el} = K_y d^{-0.5}$. Here K_y is the grain-boundary resistance. In technical steels d can be varied between 0.5 and 100 μm. K_y is in the range of 20 N/mm$^{3/2}$. For grain sizes in the range of nanometers, see NANOPHASE MATERIALS.

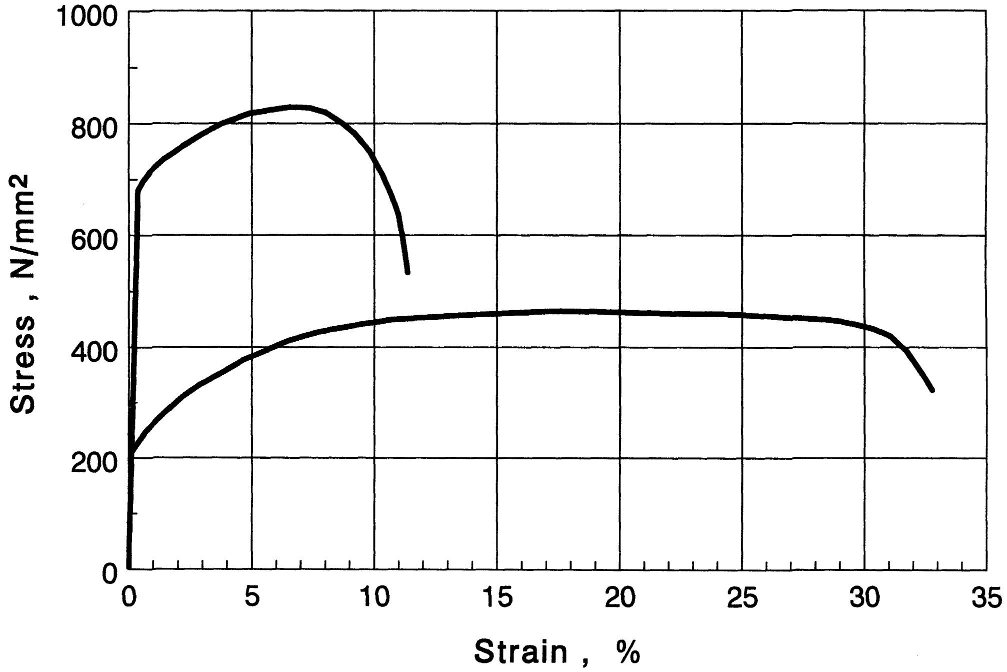

FIG. 9. Stress-strain diagram of two different steels. Strain rate after passing $R_{p0.2}$: $\dot{\epsilon} = 10^{-4}\ \text{s}^{-1}$. Upper curve, steel with a yield stress of 690 N/mm^2 and a tensile strength of 825 N/mm^2; lower curve, steel with a yield stress of 200 N/mm^2 and a tensile strength of 475 N/mm^2.

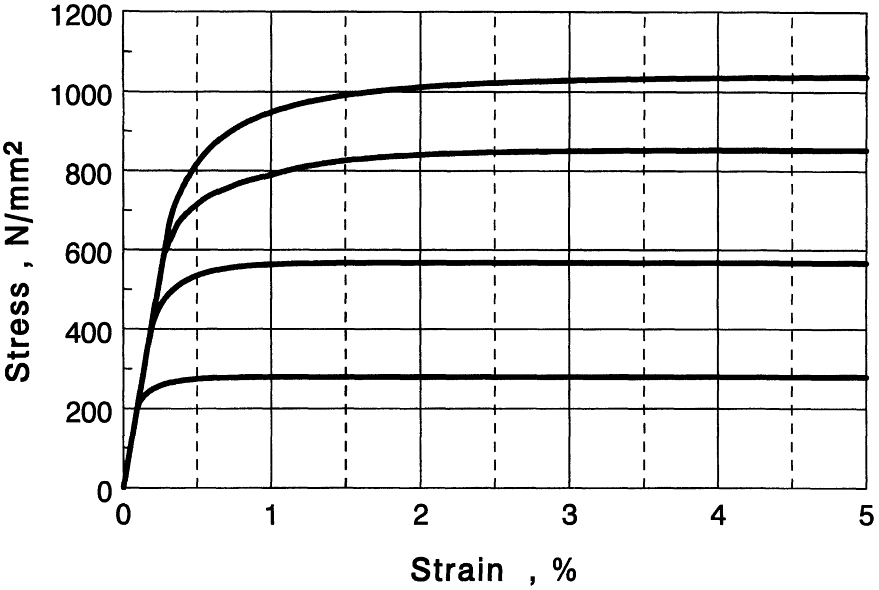

FIG. 10. Stress-strain curves for four temperatures. Steel Ck 45 isothermally transformed to bainite at 425 °C. Testing temperature from top to bottom: 20 °C, 400 °C, 500 °C, and 650 °C.

3.1.3 Influence of Solid Solution. Atoms of alloying elements, such as Mn, Cr, or Si, form a solid solution with iron, at least up to contents of 2 at.%; see Sec. 2. But the lattice of the iron atoms is disturbed by local strain fields around the atoms of the alloying elements. This gives a higher force necessary for the movement of dislocations. The yield stress is increased by $\Delta R_{el} = M\sqrt{c}$, where c is the weight percentage of the foreign atoms. In steels up to about 2 wt% of alloy content yields $\Delta R_{el} = Kc$. For an iron alloyed with Ni, K is 1 N/mm^2; an alloy of iron with Si yields $K = 125$ N/mm^2. Carbon and nitrogen have maximum values of $K = 300$ N/mm^2; they form interstitial solid solutions but have a solubility limited to a few parts per million.

3.1.4 Influence of Particles. Particles are obstacles for the movement of dislocations. For a uniform distribution of particles with a constant diameter d_p and a volume fraction V_v, the increase in yield stress is given by $\Delta R_{el} = KV_V{}^{1/3}/d_p$. The coefficient K is in the range of 0.027 N/mm. There are different equations for nonrandom arrangement and size distribution of particles as well as the kind of particles. If the particles giving an increase in yield stress are produced by precipitation, it is called precipitation hardening.

3.1.5 Influence of Dislocation Density. With increasing plastic deformation the density of dislocations increases. After loading to a stress of a in Fig. 8 and reducing the stress to zero, the steel deforms elastically. Loading again gives a yield stress of a because of the increased density of dislocations ρ by the first load. This increase in yield stress is given by $\Delta R_{el} = K\sqrt{\rho}$. For pure α iron one has $K = 155$ cm·N/mm^2; in steels K depends on the alloying content and the microstructure.

In Table 3 an approximation is given for the maximum yield stress that can be got by

Table 3. Estimation of the maximum increase in yield stress ΔR_{el} and direction of change in toughness of pure iron by different mechanisms. Grain size of the reference pure iron is about 50 μm.

Mechanism	ΔR_{el} (N/mm^2)	Toughness
Solid solution	500	↓
Precipitation	3000	↓
Grain size	500	↑
Dislocation density	1500	↓
Sum of all effects	5500	
Available	3700	

the four mechanisms of increasing the strength of steels. The maximum yield stress got in steel cord wires is about 3500 N/mm^2, less than the sum given in Table 3. The reason is that there is no microstructure combining all elements giving the maximum increase of strength given in Table 3 simultaneously. Above all, steels with the maximum strength possible would have no toughness.

3.2 Toughness

If a steel component is loaded beyond its yield stress or proof stress, it does not break immediately. The plastic deformation and the associated work hardening, see Fig. 8, provide safety against catastrophic fracture of steel constructions. At the end of plastic deformation, ductile fracture takes place, giving a reduction in area of the tensile specimen up to 60%, sometimes even more. If the conditions for superplasticity are met, the elongation after fracture reaches values of 200% or more. The toughness, which can be described qualitatively as resistance to fracture, is reduced with increasing yield stress. There is an upper limit, the microscopic cleavage fracture stress σ^*, where transcrystalline separation of crystal faces by brittle crack propagation or separation of grain boundaries occurs, giving a brittle fracture; see Fig. 11. The temperature where the curve for the yield stress crosses the curve for the microscopic cleavage fracture stress is the transition temperature between ductile and brittle fracture. If the strength of iron is increased by particles, alloying elements, or dislocation density (see Table 3), the transition temperature is raised (line 3). Only refinement of grain size increases the microscopic cleavage fracture stress. As a consequence, only grain refining gives an increase in strength of steel combined with a shift of the transition temperature to lower values (line 2 in Fig. 11). This makes evident the importance of producing fine-grained steels.

The rupture of the specimen is induced by cracks. In ductile fracture the propagation of cracks is accompanied by large plastic deformations with large energies necessary to produce the fracture. In brittle fracture, cracks grow with only very limited plastic deformation and hence limited energy of the mate-

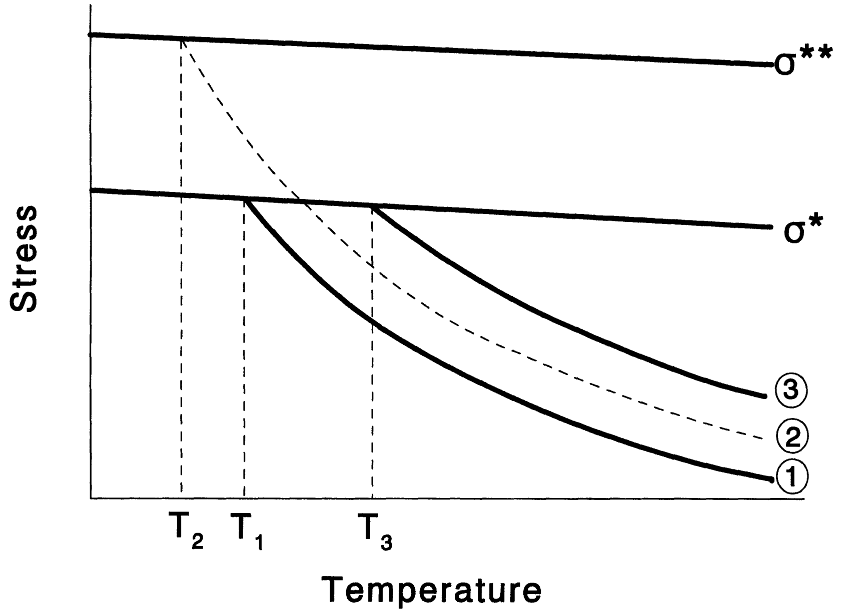

FIG. 11. Effect of hardening mechanisms on the brittle–ductile transition temperature *T*. Microscopic cleavage fracture stress of pure iron σ^*, of fine grained steel σ^{**}. Curve 1, yield stress of pure iron; curve 2, yield stress of iron with hardening by grain fining; and curve 3, yield stress of coarse-grained iron with precipitation hardening.

rial. This is described quantitatively by fracture mechanics (Atluri *et al.*, 1992; Ernst *et al.*, 1992; Mughrabi, 1993; Flügge, 1996).

3.3 Creep Resistance

At high temperatures, at a constant load below the yield stress, the strain increases time dependently. After an initial phase I (Fig. 12), the creep rate is constant, state II. In the last state, III, the creep rate increases, and internally voids are formed until rupture is initiated. In state II the creep rate $\dot{\epsilon}$ is given by $\dot{\epsilon}\ (\sigma,T) = K\sigma^n$. The creep rate increases for a given load with increasing temperature. Diagrams as given in Fig. 12 for a defined steel are used for the design of components in industrial applications working under elevated temperature. Ferritic steels are designed to be used up to temperatures about 600 °C, austenitic steels up to about 800 °C, depending on the operating conditions. Ni-based alloys may be used up to 1100 °C, e.g., for turbine blades (Gemmil, 1992; Evans and Wilshire, 1993; Fabritius *et al.*, 1993).

3.4 Fatigue Strength

During cyclic loading the applied stress σ is varied with time t with $\sigma = \sigma_m + \Delta\sigma \times \sin\omega t$. Thus $\sigma_m + \Delta\sigma$ is the maximum and $\sigma_m - \Delta\sigma$ the minimum load. Also if $\sigma_m + \Delta\sigma$ is below the yield stress, after n cycles cracks with length a are formed, which grow with $da/dn = c\Delta K^m$. Here da/dn is the crack propagation rate, ΔK the stress intensity amplitude, and c and m are constants. According to Fig. 13, there is a stress level below which no damage occurs. This level must be known for technical applications; as a first approximation it is correlated to the yield stress. The stress level below which no damage occurs is influenced by residual stresses (Boller and Seeger, 1987; Löhe, 1996; Mitchell and Landgraft, 1993).

3.5 Corrosion and Oxidation Resistance

In aqueous electrolytes iron is dissolved as $Fe \rightarrow Fe^{++} + 2e^-$ (anodic reaction) with $2H^+ + 2e^- \rightarrow H_2$ (cathodic reaction). If oxygen is dissolved there results, in a neutral electrolyte, $O_2 + 2H_2O + 4e^- \rightarrow 4OH^-$. The flow of the electrons e^- is the anodic and cathodic current, respectively. In a system of steel and a liquid corrosive environment, a potential is built up, giving the same current for the cathodic and the anodic reaction. If iron goes into solution as Fe^{++}, the corrosion loss L is given as $L = Ki_a$ mm/a (mm

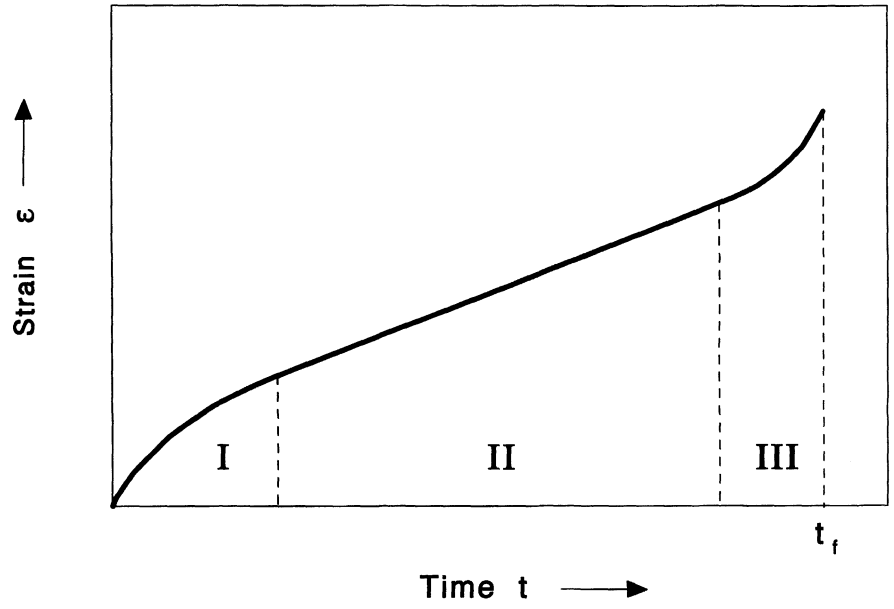

FIG. 12. Strain–time diagram for a constant load measured at elevated temperatures showing three stages of creep.

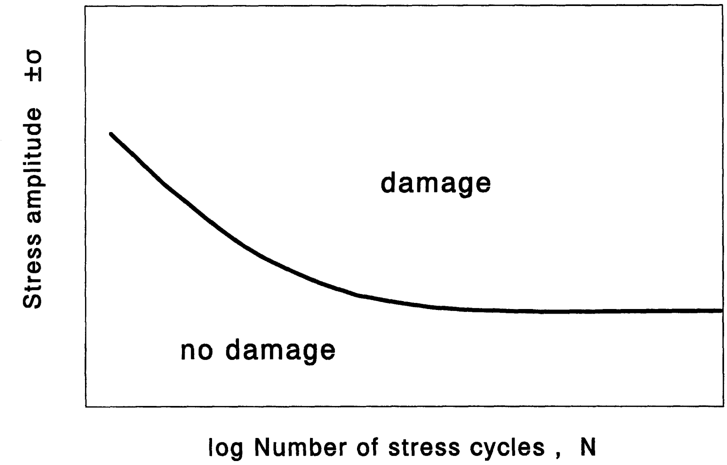

FIG. 13. Influence of cyclic loading with $\sigma < R_{p0.2}$ on lifetime.

per year); i_a is the anodic current density in μA/cm^2 (see Fig. 14), and $K = 0.01$ mm·cm^2/a·μA. In neutral water, dissolution of iron is very low, but in neutral water saturated with oxygen, the corrosion rate is about a factor of 1000 higher, giving a corrosion loss of 1.5 mm per year. In air, the corrosion loss is less because oxyhydrates and oxides are precipitated at the surface, forming rust. The corrosion rate increases with the number of dry/wet cycles at the surface and the presence of other molecules such as SO_2 in the air. As rust does not adhere, unprotected iron gets in repeated contact with the atmosphere, accelerating the corrosion rate. With addition of more than 12 wt% Cr, the steel forms a thin, adherent oxidic layer, which gives passivation against corrosion when oxygen is present. But as described above for water with and without oxygen, the corrosion rate and passivation depend strongly on the chemical composition

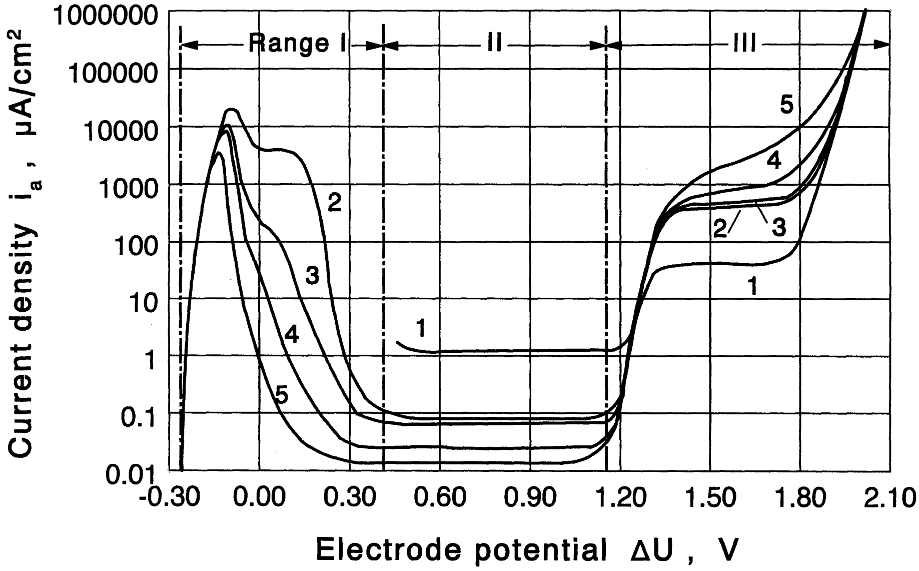

FIG. 14. Influence of Cr on the anodic current density i_a as a function of electrode potential with passivation in range II (Engell and Grabke, 1992). Weight percentages of Cr: Curve 1, 2.8; curve 2, 9.5; curve 3, 12; curve 4, 14; curve 5, 18.

of the steel and the corrosive environment. Figure 14 shows that for steels with high chromium content there is a range of a potential between 0.5 and 1.1 V over which the current density i_a is very low. In weak oxidizing agents such as water, this potential is built up, giving passivation. For this condition, a mechanically damaged passivation layer will heal immediately. For other combinations of steel and corrosive environment, passivation may be metastable, and damage of the passivation layer will start corrosion. In acid solutions without oxygen the potential may be at lower values, and, therefore, other steels must be used; see Sec. 9.4. In sulfuric and phosphoric acid, there is no passivation, but with choice of an appropriate steel the corrosion loss may be no more than 0.3 mm per year.

At high temperatures unalloyed steel is oxidized by an atmospheric environment, forming scale. At temperatures above 550 °C, the growth rate of the scale is high, because of the high diffusivity of vacancies and cations. In ferritic steels with more than 13 wt% Cr, a protective layer of $(Fe,Cr)_2O_3$ is formed, growing only slowly, for the diffusivity of oxygen, cations, and vacancies is much lower than in steels without Cr. Adding about 7 wt% Al gives, at temperatures above 900 °C, a similar protection by the formation of an Al_2O_3 coating. Ferritic steels with 25 wt% Cr, 1.5 wt% Al, and 1 wt% Si or austenitic steels with 25 wt% Cr, 20 wt% Ni, and 2.5 wt% Si can be used up to temperatures of 1150 °C in air. Again, if there are still other gases present containing S, CO, or H_2, the composition of steels must be adapted. Heat-resisting steels are used for such parts as radiation tubes, furnace equipment, and devices for dust cleaning at elevated temperatures. The steel used must be adapted not only to the temperature but also to the environmental gases (Engell and Grabke, 1992; Weßling and Oppenheim, 1993; Craig and Anderson, 1994).

4. MICROSTRUCTURES OF STEELS WITH α–γ–α TRANSFORMATION

4.1 Definitions

Steels are polycrystalline materials. A crystallite is a region of a phase with uniform crystal orientation separated from others by grain boundaries (*q.v.*). In addition, the microstructure includes lattice defects, such as vacancies and dislocations (see POINT AND EXTENDED DEFECTS IN CRYSTALS). For a quantitative calculation of properties according to Sec. 3, the microstructure must be described quantitatively by parameters given in Fig. 15 for each crystallite and each defect. Describing a crystallite, a typical example for a lattice structure is α iron; its chemical composition is the content of Mn, Cr, Si, and so on, which sometimes must be given as a function of location. Volume fraction, shape, and size distribution of the α-iron grains, combined with their arrangement and—except for spheres—the orientation distribution define the spatial distribution of α iron in the material. If the orientation of the lattices of the α-iron grains is not random, the crystallographic texture must be described additionally. Microstructures with texture are produced—among other applications—to optimize the properties of sheets for deep drawing. For a complete characterization, this description ac-

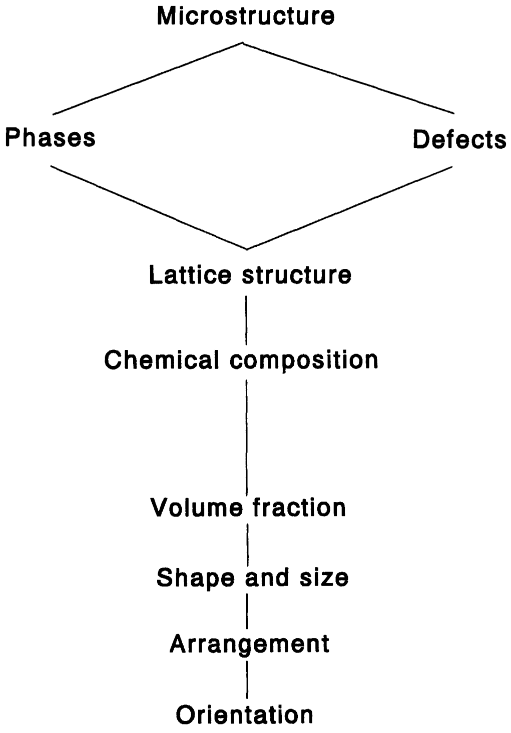

FIG. 15. Parameters for a quantitative description of microstructures.

cording to Fig. 15 must be realized for all phases and defects building up the microstructure of a material. This is quite impossible with the methods we have today. Therefore, for example, the properties of α iron or that of pearlite are measured as given by its alloy content, its density of dislocations, and the interlamellar distance of ferrite and carbide within the pearlite (see Sec. 4.3). The change of properties of a ferritic–pearlitic steel can then be described by the different volume fractions of ferrite and pearlite. The compromise is thus to measure only that detail of the microstructure, in this case the volume fraction of ferrite and pearlite, that has the decisive influence on the property under investigation (Exner and Hougardy, 1988; Diaz, 1994; Hatherley and Hutchinson, 1994).

In steel microstructures, γ iron is referred to as austenite and α iron as ferrite. The term cementite only is used for the carbide $(Fe,M)_3C$. M stands for Mn or Cr, or another metal. Induced by the metallurgical processes—see Sec. 8.1—steel contains some nonmetallic inclusions. These are small particles of sulfides as (Fe,Mn)S and oxides such as SiO_2 or Al_2O_3 (Cubberly *et al.*, 1985; Nicholson and Thornton, 1992; Hougardy, 1994, pp. 231–236).

4.2 Austenite

The microstructure of austenite is characterized by annealing twins that form parallel lines within the grains; see Fig. 16. The curved black lines are the grain boundaries. The picture is taken of an austenitic steel. In plain carbon steels austenite is not stable at room temperature; see Fig. 6.

4.3 Ferrite and Pearlite

In pure iron, austenite, cooled from 1000 °C to room temperature (see Fig. 6), transforms to ferrite by nucleation and growth, which are diffusion-controlled processes; see SOLIDS, DIFFUSION IN. Ferrite has a more irregular shape of grains than austenite; see Fig. 17. The grains may be pure iron or a solid solution of iron with alloying elements such as Cr, Mn, Si, etc. In any case, the carbon concentration within the ferrite is in the range of a few ppm. Figure 17 may also be the microstructure of a ferritic steel.

In a plain carbon steel with 0.5 wt% C at 600 °C, according to Fig. 6, ferrite and the carbide Fe_3C are in equilibrium. The driving force for the transformation is proportional to the free energy of the transformation ΔG with $\Delta G = G_A - (G_p + G_b)$. Here G_A is the free energy of the austenite, G_p that of the mixture of ferrite and carbide, and G_b is the energy stored in the phase boundaries. According to thermodynamics, in equilibrium there is a minimum of grain-boundary energy; i.e., there exist only one grain of ferrite and one grain of cementite. But under technical conditions with high cooling rates, this equilibrium is not established. Instead a lamellar arrangement of ferrite and cementite, called pearlite, is formed, which gives the highest transformation rate from austenite, Fig. 18. The maximum growth rate is found if $\lambda(T) = 4\gamma/[\Delta S_V(T_E)(T_E - T)]$, where λ is the interlamellar distance, γ is the phase-

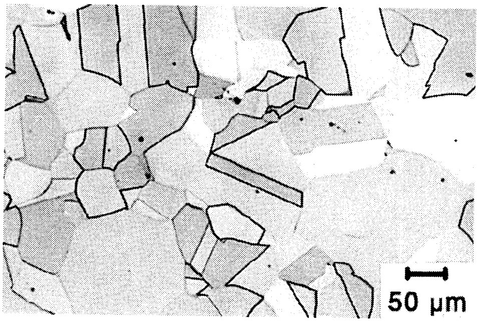

FIG. 16. Microstructure of austenite in an austenitic steel.

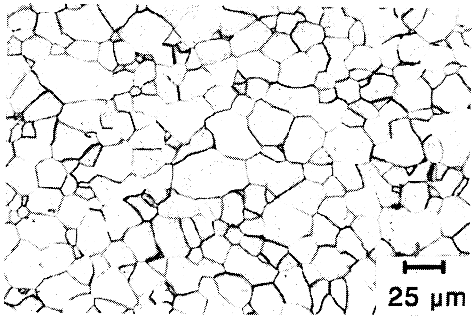

FIG. 17. Microstructure of ferrite, etched with nital (nital is 3% HNO_3 in alcoholic solution; see Petzow, 1994).

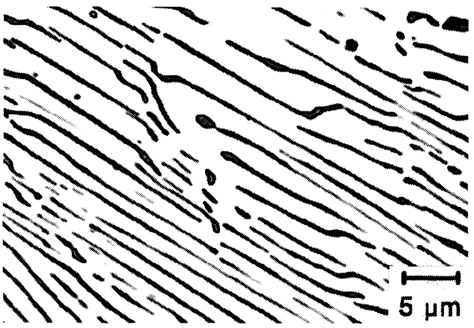

FIG. 18. Pearlite, a mixture of ferrite (white) and cementite (black). Plain carbon steel with 0.7 wt% C, austenitized 800 °C 15 min, cooling with $t_{8/5}$ = 10 000 s (see Sec. 7.2) etched with sodium picrate (for the etching, see Petzow, 1994).

boundary energy between ferrite and cementite, ΔS_V is the entropy change per mol iron at the equilibrium temperature T_E, and T is the actual temperature.

4.4 Martensite

At very high cooling rates, and at temperatures in general below 400 °C, martensite is formed from austenite without diffusion by an ordered movement of atoms over short distances. This movement is defined by the crystal structures of γ iron and α iron. The transformation propagates with the speed of sound. The carbon dissolved in the austenite remains in solution during martensite formation, which leads to very marked lattice distortions, i.e., solution hardening. Therefore, the hardness of martensite increases with the carbon content (see Fig. 19, line *a*). Of all possible microstructures of any steel, martensite has the highest hardness and yield stress. The transformation of austenite to martensite starts if the temperature goes below the martensite start temperature M_s at which ΔG is increased to a value sufficient to allow lattice distortion induced by the formation of martensite. At a given temperature, the energy available for lattice distortion is limited. Therefore, martensite formation is incomplete; its volume fraction is time independent. Depending on the alloy content, the temperature M_s varies between

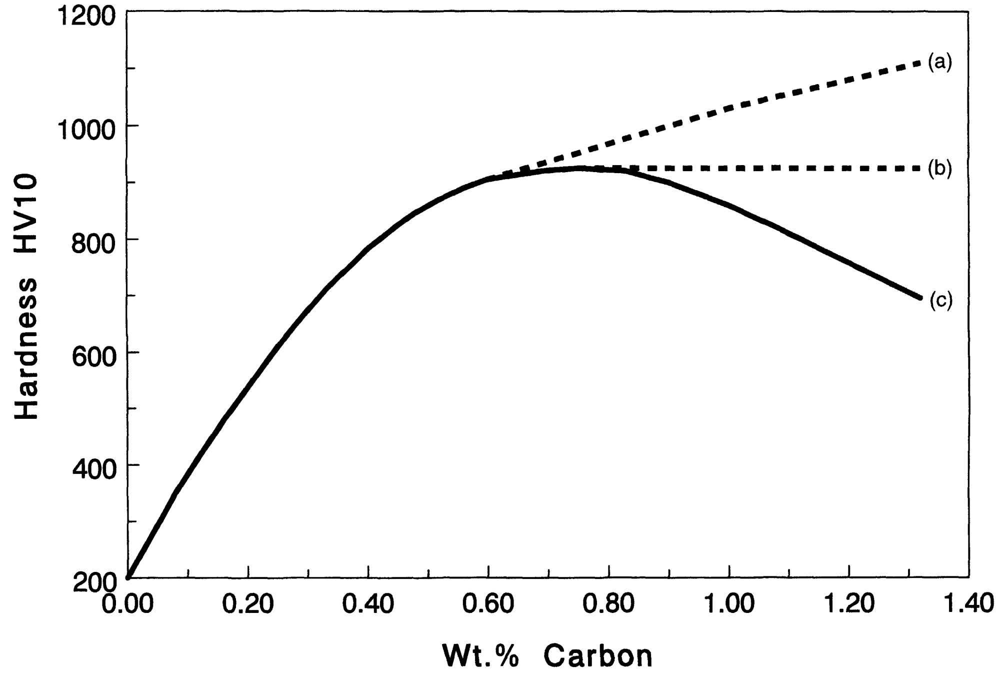

FIG. 19. Hardness of martensite as a function of carbon content of steels. Curve a, pure martensite (estimate); curve b, martensite and undissolved carbides; curve c, martensite and retained austenite. For details, see text.

450 and ~ −100 °C. Steel must be cooled down to approximately 250 °C below M_s to transform 99% of the austenite to martensite. The temperature M_s decreases with carbon content. A first approximation yields $M_s = 550 - 350c_c - 40c_{Mn} - 20c_{Cr}$, the alloy content being in wt%. If the martensite finish temperature M_f is below room temperature, after quenching there is some retained austenite. In plain carbon steels with more than ~0.7 wt% C, the hardness after quenching decreases with increasing carbon content according to the increasing volume fraction of soft retained austenite coexisting with the hard martensite (Fig. 19, line *c*). The transformation of austenite to martensite is associated with a volume increase of 0.5%–1%, depending on the alloy content. In steels with a carbon content >0.2 wt%, martensite forms mainly plates with a high density of twins, which appear needlelike in the polished section. Figure 20 shows a microstructure that is unusual for steels but makes evident the shape of the martensite. This plate martensite has a low toughness relative to its yield stress. In steels with a carbon content <0.2 wt%, a different morphology, lath martensite, is formed with a high density of dislocations. This kind of martensite has very good toughness properties (Schulz-Beenken and Hougardy, 1989; Olson and Owen, 1991; Wayman and Perkins, 1993).

4.5 Bainite

Bainite is formed by nucleation and growth at a temperature between that of the transformation of austenite to ferrite and pearlite and that of its transformation to martensite. The transformation is partly diffusion controlled and partly proceeds by a lattice-shearing mechanism similar to that for martensite. By the combination of these mechanisms, at a temperature just above M_s, the transformation of austenite to bainite is much more rapid than the transformation to pearlite at the same temperature. Below M_s, of course the formation of martensite is the fastest reaction, but after the formation of martensite, transformation may continue with the formation of bainite. Bainite is an arrangement of ferrite and carbides detectable by optical microscopy only, if formed at high transformation temperatures (Fig. 21). The ferrite within the bainite is supersaturated in respect to carbon and has a high dislocation density. After formation just above M_s, the bainite is very fine (Fig. 22)

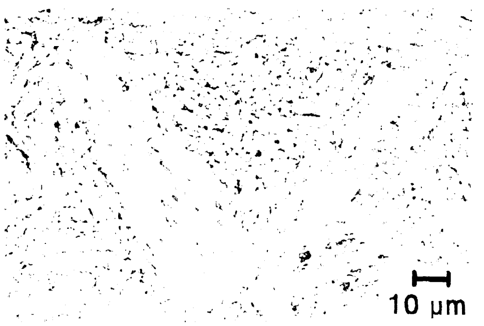

FIG. 21. Coarse bainite, formed at 500 °C. Steel with 0.2 wt% C and 2 wt% Mo, austenitized 950 °C 15 min, etched with nital.

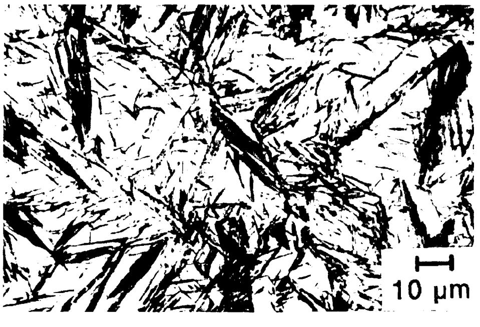

FIG. 20. Martensite (needles and plates) and retained austenite (gray) in a plain carbon steel with 1.0 wt% C. Austenitized 850 °C 15 min, water quench. Etched with nital.

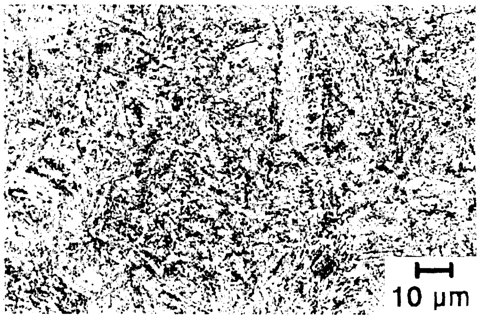

FIG. 22. Fine bainite formed at 375 °C. Steel with 0.34 wt% C and 1 wt% Cr, austenitized 850 °C 15 min, etched with nital.

and cannot be distinguished from tempered martensite; see below (Bhadeshia, 1992).

4.6 Microstructures after Tempering

The carbon held in solution within the lattice on martensite formation can form carbide precipitates by diffusion, if the temperature is increased above the temperature at which martensite is formed. Precipitation takes place more rapidly and to a greater extent as the annealing temperature is increased. The microstructure is very similar to that of bainite formed at low temperatures (Fig. 22). In alloyed steels, carbides like *M*C and M_7C_3 as described in Sec. 2 may precipitate from a supersaturated solid solution during tempering. These carbides or carbonitrides have a size ranging from a few nanometers to 100 nm, giving a precipitation hardening; see Fig. 23. The mechanisms of microstructure formation described above also apply to alloy steels, which may have several carbide phases and precipitations during cooling (Schrader and Rose, 1989; Speich, 1992).

5. GRAIN SIZE AND RECRYSTALLIZATION

During the production process, steels are deformed to get the final shape by rolling or forging in the austenite range, according to Fig. 6 between about 850 °C and 1200 °C. At least for plain carbon steels, above 900 °C, recrystallization of austenite is often so rapid that it can take place during the deformation process. This is referred to as dynamic recrystallization (see RECRYSTALLIZATION, DYNAMIC), in contrast to static recrystallization, which only occurs after deformation ceases (see RECRYSTALLIZATION, STATIC). Figure 24 shows, as an example, the fundamental influence of the grain size before deformation, deformation temperature, and degree of deformation on the final austenite grain size. The addition of alloying elements, e.g., Nb and Ti, which form carbonitrides, can cause recrystallization in the austenite domain to be slowed down so much that, after the deformation by rolling, the steel can be transformed during cooling without recrystallization. This is called thermomechanical treatment (TMT). It requires very precise control of the temperature–time curve, especially at the end of a rolling operation. The

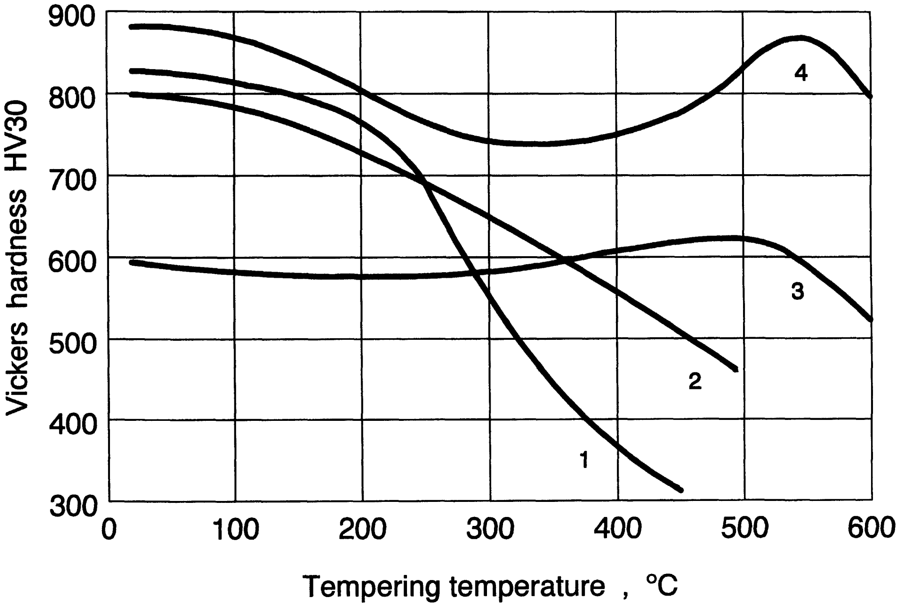

FIG. 23. Hardness as a function of temperature of tempered martensite. Curve 1, plain carbon steel with 0.80 wt% C; curve 2, steel with 0.8 wt% C, 0.6 wt% Cr, and 0.2 wt% V; curve 3, steel with 0.4 wt% C, 5 wt% Cr, 1.5 wt% Mo, and 1.0 wt% V; curve 4, steel with 0.9 wt% C, 4 wt% Cr, 5 wt% Mo, 2.0 wt% V, and 6.5 wt% W.

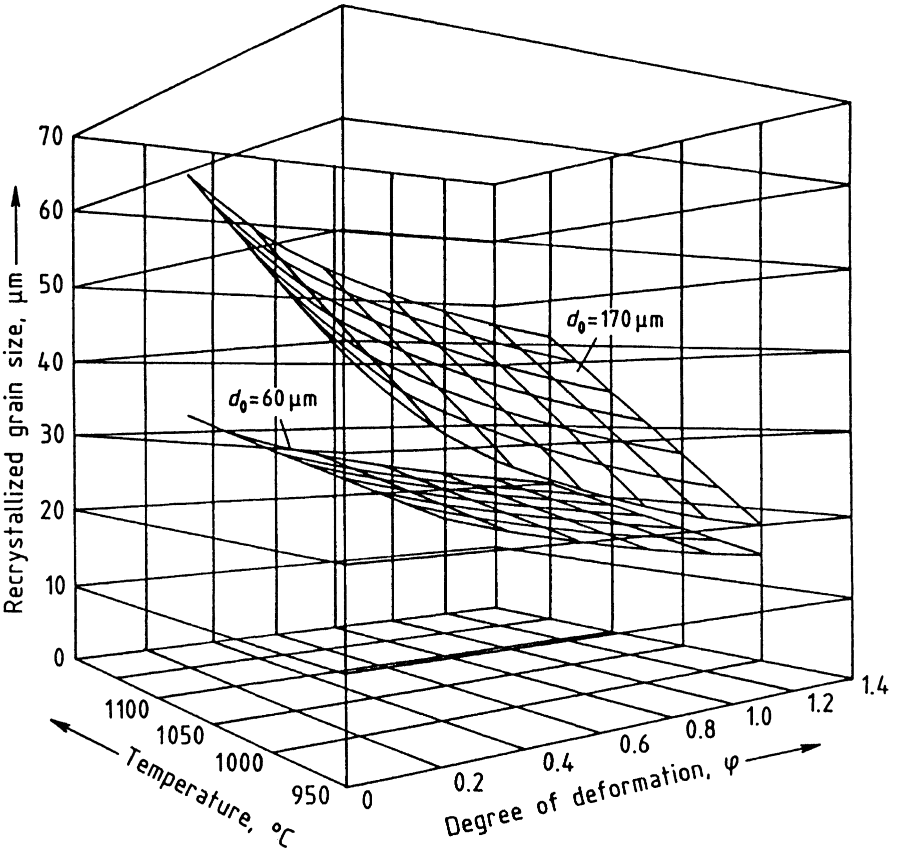

FIG. 24. Grain size of austenite as a function of deformation and initial grain size d_0. Steel with 0.038 wt% C, 1.8 wt% Mn, 0.28 wt% Mo, 0.07 wt% Nb, and 0.009 wt% N. $\varphi = \ln(h_0/h)$, h_0 initial thickness (Kaspar and Pawelski, 1985).

transformation of an unrecrystallized austenite proceeds more quickly than that of a recrystallized austenite, giving extreme fine-grained microstructures. The chemical compositions of the steels must be adapted to the parameters of the available equipment, which determines feasible cooling rate and degree of deformation (Kozasu, 1992).

Lathes and plates of martensite grow across the whole austenite grain. Therefore, the needle length of martensite is defined by the austenite grain size. The grain size of the other microstructures decreases the lower the transformation temperature. Additionally, by repeated austenitization and transformation, the average grain size of the microstructures can be reduced. This is not possible for ferritic and austenitic steels, which have no transformation. In these steels grain refining can be realized only by recrystallization.

6. CORRELATIONS BETWEEN MICROSTRUCTURES AND PROPERTIES

From Sec. 3 it follows that a fine-grained, homogeneous microstructure gives a high ratio of toughness to strength. If the yield stress is increased by precipitation of particles (see Sec. 3.1.4), and the arrangement of particles is not homogeneous, plastic deformation starts locally where particle distances are greatest, giving inhomogeneous gliding. In general, all elements influencing the mechanical properties (see Table 3) must be as homogeneous as possible. The best microstructure in this sense is tempered martensite. Plate martensite has a yield stress very close to the microscopic cleavage fracture stress (see Fig. 11) and, therefore, an insufficient toughness. With increasing tempering temperature, hardness and yield stress steadily decrease (see Fig. 23) and the toughness increases, and this enables any desired intermediate stage to be produced between the properties of martensite and that of a soft-annealed microstructure with a uniform carbide distribution; see Sec. 4.6. Bainite formed at temperatures just above M_s has a similar microstructure and, therefore, similar properties as tempered martensite. Compared with the very fine distribution of carbides in tempered martensite, the carbide distribution in bainite formed at high temperatures or in pearlite is much more inhomogeneous, and therefore the toughness-to-strength ratio of those microstructures is much less than for tempered martensite. For pearlite itself, toughness is improved with decreasing interlamellar distance. Ferrite has excellent toughness but limited strength. Unfortunately, in technical production, microstructures containing ferrite and pearlite are the cheapest ones. Therefore, for practical applications the microstructure is chosen that for a defined level of yield stress gives a sufficient toughness for the purpose in mind. These often are ferritic–pearlitic microstructures, especially if additional properties as weldability must be guaranteed. Other properties are also improved by a homogeneous microstructure. If by a heat treatment or a welding process of a steel with 13 wt% Cr locally a carbide $(Fe,Cr)_3C$ or $(Fe,Cr)_7\ C_3$ is precipitated, the Cr concentration in the surrounding matrix may be lowered below 10 wt% Cr, causing local corrosion, while farther away the matrix has 13 wt% Cr and is corrosion resistant; see Fig. 14. By adding elements such as Nb, carbon is precipitated as NbC, and the steel is stabilized against corrosion induced by precipitation of carbides containing Cr (Tenckhoff and Vöhringer, 1990; Dahl, 1992; Hougardy, 1992, pp. 461–503; Pickering, 1992, pp. 41–94).

7. DESCRIPTION OF THE α–γ–α TRANSFORMATION

If it is intended to produce microstructures such as ferrite, pearlite, bainite, or martensite by a heat treatment, first austenite is formed, which is cooled, typically by quenching, to get the desired microstructure.

7.1 Austenitization

During heating, austenite forms by nucleation and diffusion-controlled growth. Figure 25 illustrates how austenite is formed isothermally from an arrangement of ferrite and cementite in a plain carbon steel with 0.45 wt% C. The first line Ac_{1b} (see also Fig. 7), which is reached at 800 °C after 0.4 s, represents the formation of 1 vol% γ iron (austenite). The last line Ac_3, reached at 800 °C after 2000 s, represents the formation of 99 vol% γ iron. The area $\alpha + \gamma + M_3C$ corresponds to the three-phase equilibrium in Fig. 7. After a short annealing time, the carbon is still enriched in positions where carbides have gone into solution; away from these positions the carbon concentration is

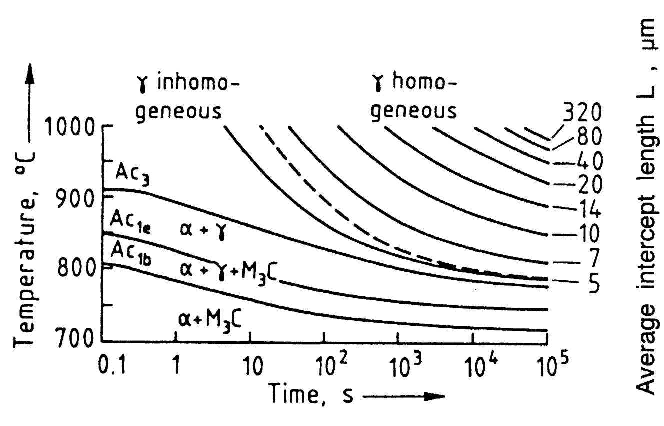

FIG. 25. Diagram for isothermal austenitization for a plain carbon steel with 0.45 wt% C including lines for equal grain size of austenite (Hougardy, 1990).

less than the average value of the steel, giving insufficient or inhomogeneous hardness after quenching (see Fig. 19). With increasing temperature and annealing time, the austenite becomes homogeneous (dotted line in Fig. 25), but the grain size, given as average intercept length, increases (see Fig. 25). In practice, austenitization is carried out at a temperature 30–50 °C above the temperature Ac₃, as defined in the relevant standards; see Sec. 10 (Orlich and Rose, 1973; Orlich and Pietrzeniuk, 1976; Hougardy, 1992, pp. 167–200; Ohtani, 1992).

7.2 Transformation during Cooling

On cooling austenite, the microstructures ferrite, pearlite, bainite, and martensite are formed as described in Sec. 4. A diagram such as that presented in Fig. 26 is called a time–temperature transformation diagram for isothermal transformation (TTT diagram). It shows the time for the beginning and the end of the transformation of austenite in the pearlite and the bainite range as a function of temperature. Thus, at 550 °C, the transformation begins after 1 s, with formation of ferrite. It ends with the transformation to pearlite and is complete after 6 s. Below 480 °C, the transformation of austenite to bainite is superseded by a transformation to pearlite. The maximum in the transformation rate to pearlite at 550 °C occurs because the nucleation rate passes through a maximum as the temperature decreases, while the growth rate, as only diffusion-controlled process, continuously decreases with decreasing temperature. At 300 °C, 50 vol% martensite is formed followed by a transformation to bainite.

Industrial heat-treating processes to produce a microstructure usually involve continuous cooling, for cost reasons. Figure 27(a) shows the time–temperature transformation diagram for continuous cooling (CCT diagram) of a plain carbon steel. As the cooling curves are all similar, they may be characterized by quoting the duration of the cooling between 800 and 500 °C ($t_{8/5}$) or the cooling time to 500 °C [Fig. 27(b)]. For each cooling curve the temperatures and times can be read off for the start of the formation of fer-

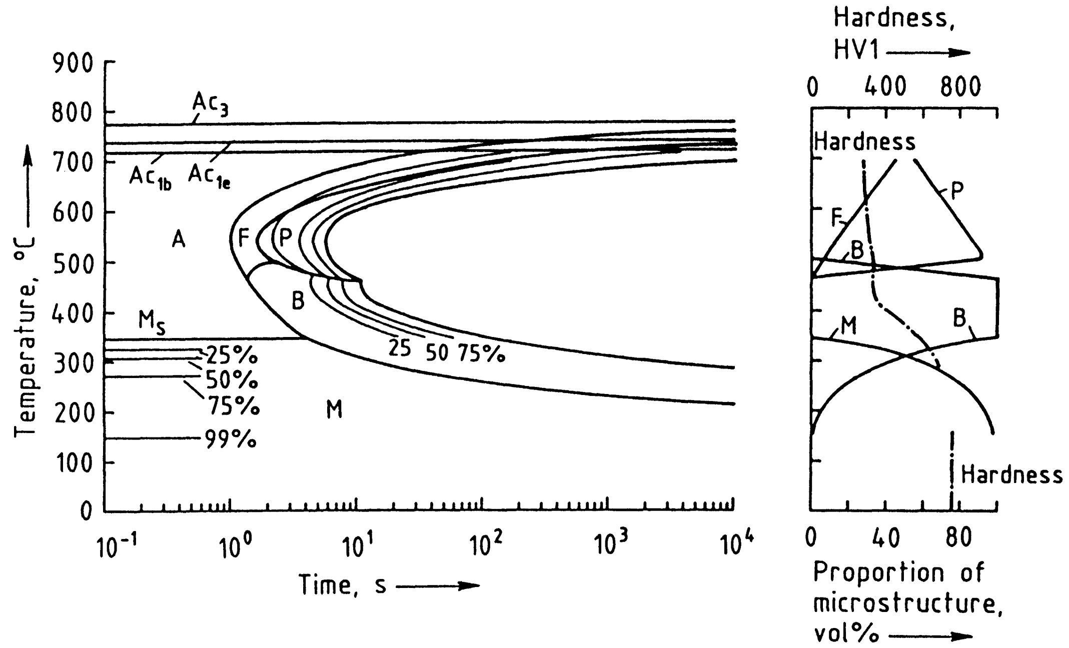

FIG. 26. Left side: diagram for isothermal transformation for a plain carbon steel with 0.45 wt% C. Austenitization 850 °C 15 min. Right side: Volume fraction of the microstructures as formed at the different temperatures. Hardness measured at room temperature. A, austenite; F, ferrite; P, pearlite; B, bainite; M, martensite (Hougardy, 1990).

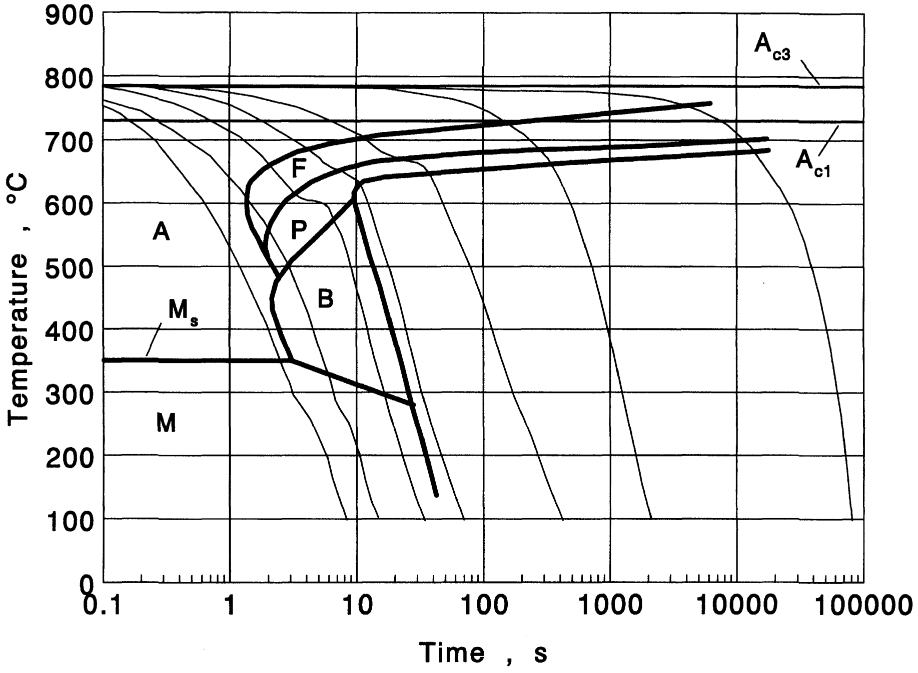

(a)

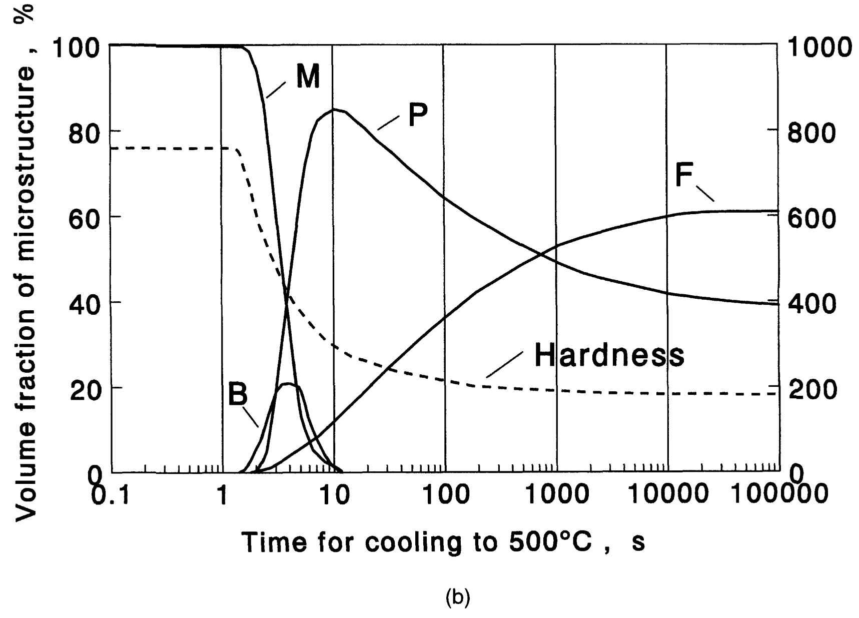

(b)

FIG. 27. (a) Transformation diagram for continuous cooling for a plain carbon steel with 0.45 wt% C. (b) Volume fraction of microstructures and hardness at room temperature as a function of cooling time $t_{8/5}$. A, austenite; F, ferrite; P, pearlite; B, bainite; M, martensite (Hougardy, 1990).

rite, pearlite, bainite, or martensite and also for the end of the transformation process.

Figure 27 shows that with $t_{8/5} = 2 \times 10^4$ s first ferrite forms, with carbon content fairly close to that predicted from the phase diagram. This ferrite forms either as a film on the austenite grain boundaries or—in steels with low carbon content—as separate grains (Fig. 28). Its properties mainly correspond to those of pure iron. After the formation of ferrite, pearlite is formed (Fig. 27). The volume ratio of ferrite to cementite in the pearlite can fluctuate over a wide range, as can be seen from the volume fractions [Fig. 27(b)]. As the volume fraction of carbide is constant, with decreasing volume fraction of ferrite precipitated first, the volume fraction of ferrite within the pearlite must increase. Nevertheless, pearlite in Fig. 28 is so fine that the ferrite-cementite lamellae are below the resolution. Figure 27 shows that $t_{8/5}$ must be <1.2 s to get martensite alone. CCT diagrams are valid only for the time–temperature sequence given in the diagram. But nevertheless, the time–temperature sequence for a technical heat treatment of a steel, to get a desired microstructure, can at least be estimated. For a more detailed calculation see Sec. 8.3.10. Collections of transformation diagrams for different steels are available (Rose and Hougardy, 1972; Van der Voort, 1991).

8. TECHNICAL STEEL PRODUCTION AND TREATING

8.1 Steelmaking

Steelmaking starts with the reduction of iron ores, which contain oxides (Fe_2O_3 or Fe_3O_4), e.g., by the reaction $2Fe_2O_3 + 3C = 4Fe + 3CO_2$. Instead of carbon also CO or H_2 is used, derived from coke, oil, or gases. Additional C is used to produce the high temperature necessary to get liquid iron by $C + 2O \rightarrow CO_2$. But depending on the costs of energy, a lot of process variants have been developed working between the liquid and the solid state of iron. Reduction with coke in the blast furnaces at present has the highest capacity. It produces hot iron or, after solidification, pig iron with about 4 wt% C. In the steelmaking process this carbon content is reduced to the desired level in the steel, between 1 and 0.01 wt%, by injection of pure oxygen in a vessel containing 200–400 *t*. The oxygen is forced into the melt with a Mach number about 2 causing a high turbulence. Hence a large contact surface is achieved giving a high reaction rate. In this vessel also 40%–60% scrap is added—the correct figure depends on details of the process—giving the excellent recycling of steel. In this process the combustion of accompanying elements such as C and Si in the iron, coming from the blast furnace, produces the necessary heat. In electric steel process the heat required for the process is obtained from an electric arc or a plasma. In electric arc furnaces only scrap or a mixture of scrap, hot iron, and pig iron can be melted. Hot and pig irons contain more or less phosphorus and sulfur that must be removed to a level less than 0.1 wt%. This is realized by slags containing CaO, SiO_2, MnO, and FeO. Volume, temperature, and composition of the slag are decisive for the level of sulfur and phosphorus achieved. If the carbon is blown down to the desired level, the oxygen content of the melt is reduced by reactions such as $xM + yO \rightarrow M_xO_y$, where M stands for Mn, Si, Al. This deoxidation reaction may be combined with vacuum degassing. After deoxidation, alloying elements such as chromium are added, which are highly sensitive for oxidation. This treatment is generally realized in a ladle outside the steelmaking furnace, called secondary metallurgy. Using slags with special compositions combined with the vacuum metallurgy, carbon contents down to 10 ppm, sulfur contents of 10 ppm, and contents of phosphorus down to 20 ppm including levels of oxygen < 10 ppm, nitrogen < 20 ppm, and hydrogen < 0.8 ppm can be achieved. Of course these low levels of clean steels are very expensive and realized

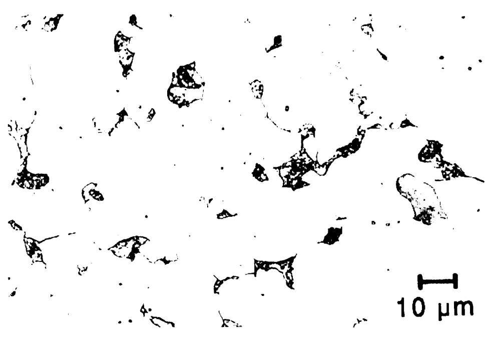

FIG. 28. Ferrite (white) and pearlite (dark). Plain carbon steel with 0.15 wt% C. Austenitization 950 °C 15 min. Cooling with $t_{8/5} = 100$ s.

only as far as really necessary for the steel quality (Nilles, 1992; Nicholson and Thornton, 1992; Deo and Boom, 1993; Oeters, 1994).

The result of steelmaking is—ideally—a liquid, homogenous solution of iron with the alloying elements. In practice, before pouring, steel contains more or less small (<1 μm) nonmetallic inclusions like (Fe,Mn) S, Al_2O_3, or SiO_2 from the desulfurization and the deoxidation reactions. On casting, the homogeneous nature of the liquid steel should be maintained in the solid solution. The aim of casting is the production of a dense, homogeneous ingot without discontinuities and with a smooth surface. But during solidification the density of steel increases; i.e., the specific volume decreases. This and the formation of dendrites, growing into the melt like lances, favor the formation of holes. Because of the change of equilibria with temperature during solidification, segregation occurs; see SOLUBILITY, ALLOYS. In equilibrium, these differences are equalized by diffusion in the solid state, but under technical conditions the time necessary for this diffusion is by far too small. In the solid steel the chemical composition may differ locally from a minimum value, e.g., of 0.7 wt% Cr to a maximum value of 1.5 wt% Cr. In the final microstructure this may be seen as banded structure.

For solidification, steel is poured into ingot molds via a refractory-lined ladle. During solidification the solubility of oxygen in steel decreases, and by reaction with the carbon in solution, CO gas bubbles are formed, partly entrapped between the growing dendrites, partly ascending, forming bubbles at the rim that gave the name "rimming steel." No CO bubbles arise after the content of oxygen is lowered by deoxidation below the solubility in the solid state, giving "killed steels." Today continuous casting is used for a wide range of steel grades and dimensions. In this process, steel is cast in an oscillating mold, and the strand is withdrawn continuously from the mold. Near net shape casting is in development. The limitation is that in casting the final dimension, the surface has not such a high quality as after rolling, internal discontinuities are not closed, and the microstructure is "as cast" concerning local segregation. All these disadvantages can be eliminated by rolling, but this should be reduced to a minimum to save costs (Schrewe, 1989; Hiebler, 1992; Hammerschmid, 1993; Irving, 1993).

8.2 Rolling

By rolling, the cast slab dimension is reduced to the final dimension, e.g., of a sheet. Internal discontinuities are closed, and a smooth surface is produced. Before starting, the slabs are heated in a furnace to reduce temperature gradients remaining from casting and to establish a unique temperature across the whole section. This is done under air as an environmental gas, which produces oxidation of the steel at the surface. Before rolling, the scale, with thickness up to 1 mm, is removed and with it all casting defects on the surface. The starting temperature of hot rolling is between 1150 and 1250 °C; the finishing temperature depends on the final dimension and on the desired state of recrystallization. To get sufficient reproducibility in dimensions below about 2 mm final thickness, cold rolling is used, e.g., for deep drawing sheets used for car bodies with a final thickness about 0.7 mm (Täffner, 1993; Wiegels, 1994).

8.3 Heat Treating

8.3.1 Influences on Transformation. In technical heat treatments temperature–time cycles are often used that do not follow isothermal or continuous cooling as implied in the TTT or CCT diagrams. Therefore, a general representation of transformation behavior is possible only by computer simulation of the transformations (see Sec. 8.3.10). The following treatment is mainly restricted to the principles of a continuous cooling process, as in Fig. 27.

The addition of alloying elements enables the temperature M_s as well as the time required for a transformation in the pearlite and bainite ranges to be shifted within wide limits. In a plain carbon steel, 25 vol % ferrite and 75 vol % pearlite are formed during a cooling from 800 to 500 °C within 30 s; see Fig. 27. By addition of alloying elements such as Mn, Cr, Ni, but also C, the diffusion-controlled transformation to ferrite, pearlite, and bainite becomes much slower. This effect is used to design tool steels, which on air cooling of sections of up to 30 mm diameter form only martensite. However, in any given application, steels with the lowest possible alloy contents are used for reasons of cost (Davis, 1991; Hougardy, 1992, pp. 461–503; Othani, 1992; Vöge, 1993).

8.3.2 Austenitization. The industrial austenitization of heat-treated engineering steels 30–50 °C above Ac_3 usually leads to the formation of a homogeneous austenite with small grain size; see Fig. 25. Special treatments are necessary for tool steels (see Sec. 9.3).

8.3.3 Effect of Dimensions. Industrial cooling operations are usually carried out continuously in air or liquid media. Figure 29 shows the comparison of temperature–time curves for the outer and inner zones of two cylinders with different diameters. The temperature differences between the zones

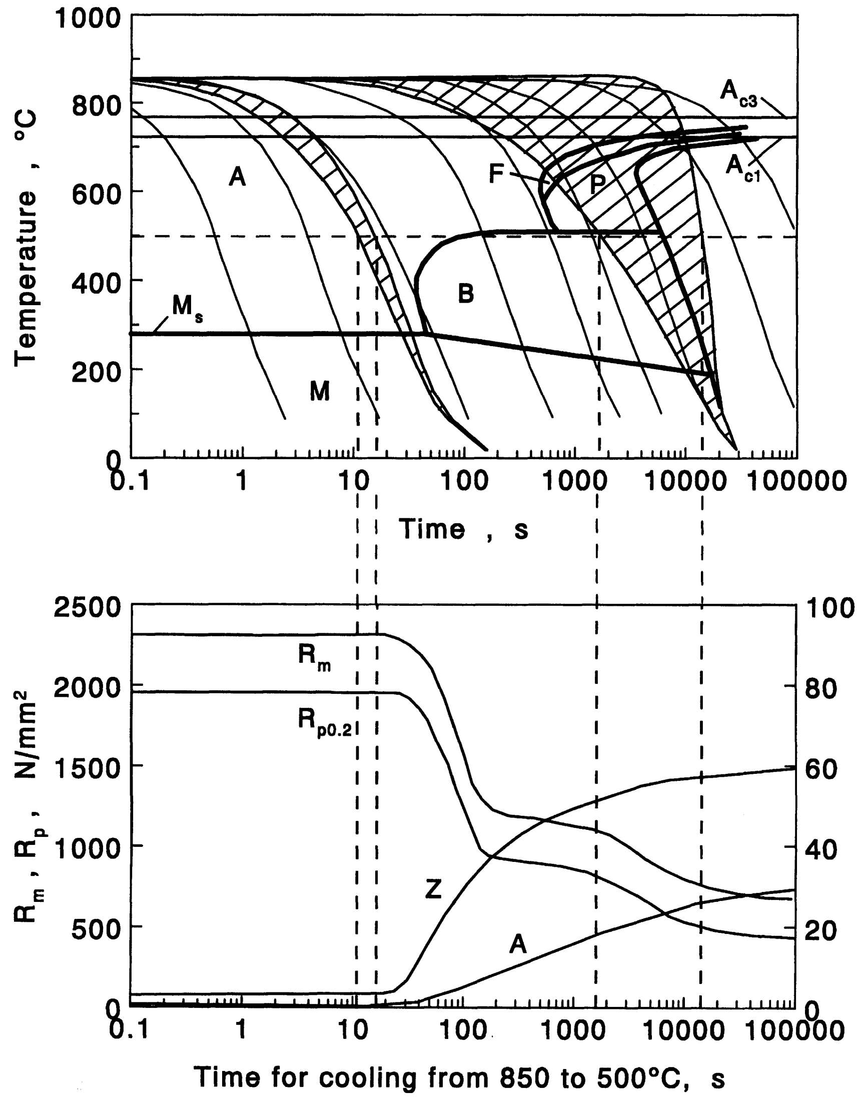

FIG. 29. Upper: Transformation diagram for continuous cooling of a steel with 0.5 wt% C, 1 wt% Cr, and 0.25 wt% Mo. Dashed area: cooling curves for surface (left line) and core (right line) of cylinders with 20 (left area) and 1000 (right area) mm diameter, respectively, during oil quenching. Lower: results of tensile tests (see Sec. 3.1) for different microstructures plotted as function of the cooling time $t_{8/5}$ (Hougardy, 1990).

can be very great, especially during quenching. This leads to the formation of different microstructures in the outer and inner zones. An increase in the cooling time $t_{8/5}$ decreases the proof stress $R_{p0.2}$ and increases the elongation after fracture A and reduction in area after fracture Z (Fig. 29). Figure 30 shows how the hardness varies over the cross section of cylinders of different steels. For an oil-quenched plain carbon steel, maximum hardness is not reached, even near the surface. Maximum hardness with a purely martensitic microstructure can be obtained only near the outer surface for steel *c*. This effect limits the generation of only one microstructure across a whole section (Hougardy, 1992, pp. 461–503; Liscic *et al.*, 1992).

8.3.4 Transformation in the Pearlite Range. A transformation in the lower region of the pearlite range is used in wire patenting. This forms pearlite with very small interlamellar distances, which has very good wire-drawing properties. As the diameter is small, the transformation can be carried out isothermally. Other applications include the production of rails, in which pearlite is formed at low temperatures from the heat of rolling. In plain carbon steels, a ferritic–pearlitic microstructure can be produced by air cooling (Fig. 27). The corresponding heat treatment is known as normalization. Delayed cooling is necessary with alloy steels, especially in thin sections, if a ferritic–pearlitic microstructure is desired (Othani, 1992; Williams, 1992).

8.3.5 Transformation in the Bainite Range. In bainitization, a microstructure consisting only of bainite is produced by an isothermal transformation (Fig. 26). Continuous cooling cannot in principle give a 100% bainitic microstructure; see Fig. 27. The fine microstructures formed in the lower region of the bainite range (see Fig. 22), have a favorable combination of toughness and yield stress, often matching the properties of tempered martensite (Pickering, 1992, pp. 41–94).

8.3.6 Transformation in the Martensite Range. Martensite can be obtained by cooling so rapidly that the formation of microstructures of the pearlite and bainite range is suppressed, and this is simply called hardening. Plain carbon steels must usually be water-quenched to achieve martensite at

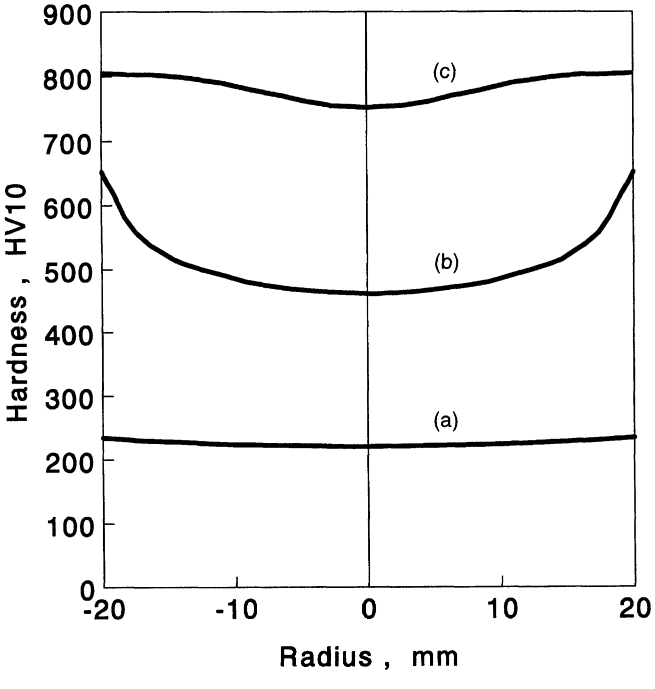

FIG. 30. Variation of hardness across the diameter for three steels, after oil quenching of cylinders with 40 mm diameter. Curve a, plain carbon steel with 0.45 wt% C; curve b, steel with 0.41 wt% C and 1 wt% Cr; curve c, steel with 0.50 wt% C, 1 wt% Cr, and 0.25 wt% Mo.

least at the surface. Alloyed steels may be cooled more slowly by quenching in oil or cooling in air. Because of the high rates of cooling obtained by water quench, large temperature differences between the outer and inner regions of large workpieces occur (see Fig. 29), so that stresses and residual stresses are produced that can induce cracking and distortion. These hardening cracks impose a limit to heat-treatment processes involving rapid cooling of large or complex workpieces (Fletcher, 1989; Totten, 1992; Hauk *et al.*, 1993).

8.3.7 Quenching and Tempering. In general, quenching and tempering consists of hardening followed by annealing, which can give any yield stress between that of martensite and that of a soft-annealed microstructure; see Fig. 23. The relationship between tempering temperature and yield stress is known for many grades of steel. After tempering of martensite, the arrangement of carbides in the ferrite is very uniform and homogeneous; therefore, this process gives very good toughness properties for a given yield stress (Pickering, 1992, pp. 41–94; Speich, 1992).

8.3.8 Annealing for Spheroidization. In components of small dimensions fabricated of high-alloy steels, microstructures of appreciable hardness are produced even with air cooling, which can give machining problems. To get a microstructure suitably weak for machining and to get a standardization of shape and size of carbides before the final heat treatment of tool steels, the carbides are often spheroidized by means of a prolonged annealing process at temperatures just below Ac_1; i.e., coarse carbides are formed in the ferrite (Fig. 31). This process, annealing for spheroidizing, can be accelerated by an imposed temperature oscillation around A_1 or by a prior deformation process (Gasterich *et al.*, 1993).

8.3.9 Surface Hardening. For technical applications, it is often desired that the surface of a workpiece should be hard, in order to provide wear resistance. However, as the load is comparatively small, the average yield stress across the total section can be low. Therefore, often only the surface zone is hardened by two methods (Stevens and Davies, 1992; Tacke *et al.*, 1993; Grosch *et al.*, 1994):

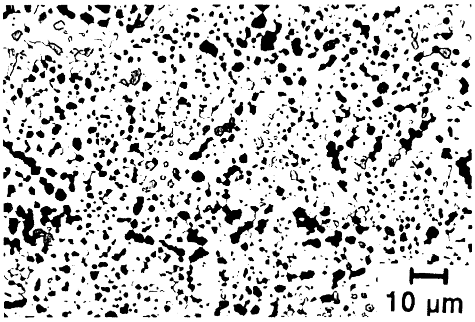

FIG. 31. Microstructure after spheroidize annealing. Plain carbon steel with 1.3 wt% C. White, ferrite; black, carbide. After austenitization 850 °C 15 min and water quench annealed at 720 °C for 8 h. Etched with sodium picrate.

1. Only the surface zone of a workpiece is austenized and hardened by quenching; the interior temperature does not exceed Ac_1. After tempering at about 150 °C, a hardened and tempered high-strength surface zone is produced, surrounding a comparatively ductile core. A process of this kind currently used is induction hardening.
2. In case hardening, steels with about 0.15 wt% C are placed in a carbon-containing medium at 900–950 °C, and carbon diffuses into the surface. The most common technique is gas carburization, which is very easy to control to get a carbon content of about 0.8 wt% at the surface, giving maximum hardness after quenching (see Fig. 19). This treatment is used for gear wheels for motor vehicles and sometimes for roller bearings. Surface hardening in very thin layers can be produced by nitriding, chromizing, and boronizing.

By such treatments, residual compressive stresses are usually produced at the surface, and this also improves fatigue resistance. Because of the internal stresses, the properties of the workpiece can no longer be described by the properties of the material only (Hauk *et al.*, 1993).

8.3.10 Simulation of Heat Treatment. For transformable steels, there are limitless possibilities for transforming austenite in the pearlite, bainite, or martensite range, using controlled temperature–time regimes. Finding an optimum temperature–time regime to produce a desired microstructure, and hence the specified properties, requires long-time experience or a very large number of experiments. Models based on empirical equations or equations based on physical principles have been developed that can be used to calculate the progress of transformation (SteCal 2.0; Hougardy and Yamazaki, 1986; Denis *et al.*, 1992; Bodin and Segerberg, 1992; Inden *et al.*, 1992; Kirkaldy, 1993). Similarly, knowing the temperature and degree of deformation, grain sizes of the austenite after heat treatment of particular grades of steel can be calculated (Kern and Schriever, 1993). If the mechanical properties of the individual microstructure of a steel are known, the final properties produced by a given process can also be calculated (Degenkolbe *et al.*, 1992). In this way, a proposed process can be simulated in advance. Some models are now available but must still be improved. For defined conditions, the properties can be calculated from the chemical composition (Pickering, 1992, pp. 41–94). Additionally, the residual stresses arising during a heat treatment can be estimated (Fletcher, 1989; Hougardy, 1991).

9. STEEL GRADES

Besides mechanical properties, cold workability, weldability, resistance to corrosion, or machining properties may be important for an application. The classification of steels is not simple, for one steel may be used for different applications. Therefore, one classification focuses on the chemical composition, another on the application (e.g., tool steels), and a third on the microstructure (as example, ferritic or austenitic steels). Compilations can be found in standards, or in the general literature. In the following only some examples are given, to have an impression of the range of chemical compositions used and the variety of properties achieved (Pickering, 1992, pp. 41–94; Jäniche *et al.*, 1993; Kalla, 1994).

9.1 Low-Alloy Steels

Low-alloy steels are used as structural steels with a yield stress <355 N/mm^2. If the yield stress is >355 N/mm^2, the steels are often called high-strength low-alloy steel. These steels are supplied with the mechanical properties as standardized; they are not heat treated by the customer. In most cases these steels are designed for welding (Hougardy, 1992, pp. 504–534; Houldcroft, 1992; Olson *et al.*, 1993). Levels of yield stress up to about 500 N/mm^2 can be achieved with ferritic–pearlitic or bainitic microstructures by normalizing or thermomechanical treatment. The chemical analysis is in the range <0.2 wt% C, 0.1–0.6 wt% Si, and 0.5–1.8 wt% Mn. For a yield stress up to 900 N/mm^2, steels have additions of Cr, Ni, and Mo and are delivered in a quenched and tempered condition. Especially sheets from these steels are protected during the production process by different coatings against corrosion (Stevens and Davies, 1992; Tenhaven *et al.*, 1992; Müsgen *et al.*, 1993).

9.2 Heat-Treated Engineering Steels

The main characteristic of heat-treated steels is the chemical composition, which is a decisive factor for the formation of austenite and its transformation during cooling. The first point is the grain size of austenite during austenitization that can be controlled by precipitation of AlN and/or Ti(C,N). These fine-grained steels allow complete austenitization of parts with great thickness without formation of a coarse grain. As described in Sec. 6, for a defined yield stress the optimal ratio of toughness to strength is achieved by quenching to martensite and tempering. But to produce martensite across the whole section of a part has two limitations. One is the finite thermal conductivity, the other is the presence of stresses arising during quenching in water. Therefore, if the maximum toughness must be achieved in thick cross sections, the ferrite, pearlite, and bainite transformation must be shifted to long times by appropriate alloying. In this case, martensite is produced even at modest cooling rates in the center of the part. As far as it is allowed by the desired combination of toughness and strength necessary for an application, steels with a ferritic–pearlitic or bainitic microstructure are taken or surface hardening is used. Therefore, the chemical analysis ranges from 0.45 wt% C and 0.7 wt% Mn for small dimensions to a steel with 0.35 wt% C, 0.5 wt% Mn, 2 wt% Cr, 0.4 wt% Mo, and 4 wt%

Ni for heavy parts and case-hardening steels with <0.2 wt% C, <1 wt% Mn, 1.8 wt% Cr, 0–0.5 wt% Mo, and 0–3.5 wt% Ni (Naylor and Cook, 1992; Schauwinhold, 1994).

9.3 Tool Steels

Tool steels have a wide range of applications such as cold and hot forming and cutting as well as die casting of metals and polymers. Therefore, their strength and hardness must be above that of the materials being formed or cut. Besides a high wear resistance, a minimum of toughness must be realized. For the selection of chemical analyses, the same principles are valid as for steels designed for heat treatment. The transformation characteristic must be adapted to the diameter to get the formation of martensite through the desired part of the diameter. A further aspect used in design of one subset of tool steels is to increase the content of carbides from maximal 11 vol% in eutectoid steels up to 30 vol%, to improve the abrasive resistance including the formation of hard carbides such as WC. Carbides such as $(Fe,Cr)_7C_3$, $(Fe,Cr)_{23}C_6$, MoC, and VC precipitate at higher temperatures than Fe_3C (see Fig. 23), giving better high-temperature strength. Nevertheless, according to the application the hardness of tool steels varies from about 200 HV 10 for glass-mold steels to about 900 HV 10 for machining tools. For some applications it is most efficient to apply surface hardening including carburizing and nitriding or surface coating by, e.g., TiN.

In general, tool steels are quenched and tempered. Quenching a plain carbon tool steel with 1 wt% C from 1000 °C would give retained austenite and a decrease in the hardness according to Fig. 19. In service, this austenite will transform under the service load, and the dimension of the tool would change, which is not tolerable. Therefore, plain carbon tool steels are austenitized at temperatures 30–50 °C above Ac_1, just to get about 0.8 wt% carbon in solid solution in austenite, sufficient to get the maximum hardness of martensite according to Fig. 19. In alloyed steels, the method is, in principle, the same, but the temperature for austenitization varies with the steel grade and may be as high as 1200 °C to get sufficient carbon in solid solution as for high-speed steels. Besides the effect on carbide formation, the alloying elements reduce the transformation rate, allowing oil quenching or even air cooling to still produce martensite at the surface. This slow cooling is important to prevent cracking or dimensional changes during quenching and tempering of thick sections. For the effect of tempering on hardness, see Fig. 24. During this tempering—in alloyed steels it is repeated several times—the volume fraction of retained austenite is also reduced.

Carbon tool steels have carbon contents between 0.6 and 1.4 wt% C. Applications include cutting blades, hammers, and axes. Steels containing up to 5 wt% Cr, 1.5 wt% Mo, 1 wt% W, and 0.5 wt% V are used for cold-work dies. The chemical analysis is adapted to get a transformation characteristic that allows oil quenching or air cooling to get martensite. Applications include reamers, punches, drills, and in some cases plastics molding dies. Heat-resistant die steels are needed—for instance for Al die casting or hot deformation and plastics molding—and have up to 0.5 wt% C, 9 wt% Cr, 3 wt% Mo, 2 wt% V, and 8 wt% W. High-speed steels, used for cutting metals, e.g., by turning, drilling, or punching, have up to 1.4 wt% C, 4 wt% Cr, 10 wt% W, 9 wt% Mo, 4 wt% V, and 10 wt% Co. Quenched and tempered, they have a hardness of about 65 HRc.

Maraging steels contain about 18 wt% Ni, 8 wt% Co, 4 wt% Mo, and 0.5 wt% Ti. After quenching, the microstructure is a lath martensite with a proof stress of about 500 MPa. During tempering about 500 °C there is a precipitation hardening by Ni_3Mo and Ni_3Ti within the martensite, giving an increase of yield stress to about 2000 MPa. Applications include parts with extreme high strength as well as parts for plastics and light metal molding for high accuracy (Rayson, 1992; Wilmes *et al.*, 1993).

9.4 Stainless Steels

In corrosion resistance mainly the concentration of alloying elements in solid solution is important; see Sec. 3.5. The other parameters of the microstructure have much less influence, except some effects in stress corrosion and hydrogen-induced corrosion. The combination of alloying elements needed to get the desired corrosion resistance is in concurrence with that necessary to guarantee other properties such as strength, toughness, ease of forming, and weldability. In general, there are martensitic steels with 0.3–1 wt% C and about 15 wt% Cr, giving a 0.2% proof stress up to 700 N/mm^2 as used for some knife blades or spindles. Ferritic steels with 15–30 wt% Cr and <0.1 wt% C have a proof stress between 300 and 500 N/mm^2 and a tensile strength between 500 and 700 N/mm^2 and are used, for instance, for heat exchangers under very aggressive conditions. Carbon contents of <0.003 wt% in these steels give an improved toughness. Austenitic steels with about 18 wt% Cr and more than 9 wt% Ni have a 0.2% proof stress of about 300 N/mm^2 and a tensile strength of about 700 N/mm^2. This indicates the great work hardenability of these steels. Applications range from kitchen equipment to vessels to resist organic and sulfuric acids in the chemical industry. Lastly, there are ferritic–austenitic steels, which have a 0.2% proof stress of about 400 N/mm^2. All steels containing martensite and ferrite may be sensitive to temper embrittlement at temperatures above 400 °C. Therefore, for high-temperature application in general, austenitic steels are used.

For the corrosion resistance of a steel with more than 12 wt% Cr, see Sec. 3.5 and Fig. 14. In solutions with Cl^- a nonuniform corrosion loss, pitting or a crevice corrosion, occurs. Also crevices between stainless steel and nonconductors, such as plastics or ceramics, may cause this corrosion. The higher the so-called "pitting corrosion equivalent," PRE, the higher the critical pitting and crevice corrosion temperature with PRE = wt% Cr + 3.3 × (wt% Mo) + 30 × (wt% N). Ferritic steels are used with 30 wt% Cr and 5 wt% Mo including some Nb and low C and N contents to prevent precipitation of Cr carbide (Truman, 1992; Chawla and Gupta, 1993; Heinemann *et al.*, 1993; Davis, 1994; Bogaerts and Agema, 1995).

10. STANDARDS

To make trade and application of steels more easy, there are standards that describe the chemical analyses as well as the properties of a steel grade. Further standards define testing methods. International standards are ISO (International Standardization Organisation) and ECISS (European Committee for Iron and Steel Standards). The standards are published as ISO *xxx* and EN *xxx*, respectively, where *xxx* stands for a number. There are national standards published by ASTM (American Society for Testing and Materials) and AISI (American Iron and Steel Institute) in the U.S.A. or JIS (Japanese Institute for Standardization) in Japan. In general, a steel with, for instance, 0.25 wt% C, 1 wt% Cr, and 0.25 wt% Mo standardized according to an EN standard is similar to but not identical with a steel standardized according to an ASTM standard having similar chemical composition. The reason for this divergence is that a new steel in general is developed following the demands in one country and standardized for the application by a national standard. If application is widened, an international standard may be sensible, but now the demands of different countries and a wider field of application must be taken into concern, leading to changes compared to the first national standard. Or, the steel remains of national interest only, and there is no international standard at all. A comparison of steel grades is summarized in various books (International Technical Information Institute, 1990; Verein Deutscher Eisenhüttenleute, 1994; DIN Handbooks, 1993; Wegst, 1995; Unterweiser, 1993; ASTM, 1994).

Works Cited

ASTM (1994), "Iron and Steel Products," *Annual Book of ASTM Standards*, New York: American Society for Testing and Materials.

Atluri, S. N., Newman, J. C., Jr., Rajmu, I., Epstein, J. S. (Eds.) (1992), *Fracture Mechanics*, Vol. 2, New York: American Society for Testing and Materials.

Bhadeshia, H. K. D. H. (1992), *Bainite in Steels*, London: Institute of Materials.

Bodin, J., Segerberg, S. (1992), "Benchmark Testing of Computer Programs for Determination of Hardening Performance," in: *Proceedings of the First International Conference on Quenching and Control of Distortion*, Chicago, Illinois, 22–25 September 1992, Metals Park, OH: American Society for Metals, pp. 133–139.

Bogaerts, W. F., Agema, K. S. (Eds.) (1995), *The Corrosion Reference on CD-ROM*, National Association of Corrosion Engineers.

Boller, C., Seeger, T. (1987), *Materials Data for Cyclic Loading*, Amsterdam: Elsevier.

Cahn, R. W., Haasen, P. (1996), *Physical Metallurgy*, Amsterdam: North-Holland.

Chawla, S. L., Gupta, R. K. (1993), *Material Selection for Corrosion Control,* Metals Park, OH: American Society for Metals.

Craig, B., Anderson, D. (Eds.) (1994), *Handbook of Corrosion Data,* Metals Park, OH: American Society for Metals.

Cubberly, W. H., Baker, H., Benjamin, D. (Eds.) (1985), *Metals Handbook,* Vol. 9, *Metallography and Microstructures,* Metals Park, OH: American Society for Metals.

Dahl, W. (1992), in: W. Jäniche, W. Dahl, H. F. Klärner, W. Pitsch, D. Schauwinhold, W. Schlüter, H. Schmitz (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 1, *Fundamentals,* Düsseldorf: Verlag Stahleisen, pp. 201–371.

Davis, J. R. (1991), *ASM Handbook,* Vol. 4, *Heat Treating,* Metals Park, OH: American Society for Metals.

Davis, J. R. (Ed.) (1994), *Stainless Steels,* Metals Park, OH: American Society for Metals.

Degenkolbe, A. K. J., Müsgen, B., Schriever, U. (1992), *Iron Steel Inst. Jpn.* **32,** 387–394.

Denis, S., Farias, D., Simon, A. (1992), *Iron Steel Inst. Jpn.* **32,** 316–325.

Deo, B., Boom, R. (1993), *Fundamentals of Steelmaking Metallurgy,* Englewood Cliffs, NJ: Prentice-Hall.

Diaz, D. J. (1994), *Quantitative Microscopy and Image Analysis,* Metals Park, OH: American Society for Metals.

DIN Handbooks (1993), *Iron and Steel,* Düsseldorf: Verlag Stahleisen.

Elliott, R. (1992), in: F. B. Pickering (Ed.), *Constitution and Properties of Steels,* Materials Science and Technology Vol. 7, Weinheim: VCH, pp. 693–738.

Engel, H.-J., Grabke, J. J. (1992), in: W. Jäniche, W. Dahl, H. F. Klärner, W. Pitsch, D. Schauwinhold, W. Schlüter, H. Schmitz (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 1, *Fundamentals,* Düsseldorf: Verlag Stahleisen, pp. 412–460.

Ernst, A., Saxena, D. L., McDowell, M. R. C. (Eds.) (1992), *Fracture Mechanics,* New York: American Society for Testing and Materials.

Evans, R. W., Wilshire, B. (1993), *Introduction to Creep,* London: Institute of Materials.

Exner, H. E., Hougardy, H. P. (1988), *Quantitative Microscopy,* Oberursel, Germany: DGM Informationsgesellschaft.

Fabritius, H., Christianus, D., Forch, K., Krause, M., Küller, H., von den Steinen, A. (1993), in: W. Jäniche, W. Dahl, H. F. Klärner, W. Pitsch, D. Schauwinhold, W. Schlüter, H. Schmitz (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 2, *Applications,* Düsseldorf: Verlag Stahleisen, pp. 226–272.

Fletcher, A. J. (1989), *Thermal Stress and Strain Generation in Heat Treatment,* New York: Elsevier Applied Science.

Flügge, M. (Ed.) (1996), *The Appearance of Cracks and Fractures in Metallic Materials,* Düsseldorf: Verlag Stahleisen.

Franke, P. (1995), Max-Planck-Institut für Eisenforschung, Düsseldorf, personal communication.

Gasterich, H.-J., Hagedorn, K. E., Kaspar, R., Pawelski, O. (1993), *Steel Res.* **64,** 414–419.

Gemmill, M. G. (1992), in: F. B. Pickering (Ed.), *Constitution and Properties of Steels,* Materials Science and Technology Vol. 7, Weinheim: VCH, pp. 489–526.

Grosch, J. *et al.* (1994), *Einsatzhärten,* Renningen-Malmsheim, Germany: Expert-Verlag.

Hammerschmid, P. (1993), in: W. Jäniche, W. Dahl, H. F. Klärner, W. Pitsch, D. Schauwinhold, W. Schlüter, H. Schmitz (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 2, *Applications,* Düsseldorf: Verlag Stahleisen, pp. 650–677.

Hatherley, M., Hutchinson, W. B. (1994), *An Introduction to Textures in Metals,* London: Institute of Materials.

Hauk, V., Hougardy, H., Macherauch, E., Tietz, H.-D. (Eds.) (1993), *Residual Stresses,* Oberursel, Germany: DGM Informationsgesellschaft.

Heinemann, W., Oppenheim, R., Weßling, W. (1993), in: W. Jäniche, W. Dahl, H. F. Klärner, W. Pitsch, D. Schauwinhold, W. Schlüter, H. Schmitz (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 2, *Applications,* Düsseldorf: Verlag Stahleisen, pp. 382–422.

Hiebler, H. (Ed.) (1992), *Gmelin-Durrere Metallurgy of Iron,* Vol. 11, Berlin: Springer.

Hodgeson, P. (1993), *Quenching and Carburizing,* London: Institute of Metals.

Hoffmann, J. E., Löhe, D., Macherauch, E. (1987), in: E. Macherauch, V. Hauk (Eds.), *Proceedings of the International Conference on Residual Stresses in Science and Technology,* Vol. 2, Oberursel: DGM Informationsgesellschaft, pp. 801–808.

Horstmann, D. (1983), *Damaszenerstahl,* Fachaußbericht 9.008 des VDEh, Düsseldorf: Verein deutscher Eisenhüttenleute, pp. 87–105.

Hougardy, H. P. (1990), *Umwandlung und Gefüge unlegierter Stähle. Eine Einführung,* Düsseldorf: Verlag Stahleisen.

Hougardy, H. (1991), in: V. Hauk, H. Hougardy, E. Macherauch, H.-D. Tietz (Eds.), *Residual Stresses, Measurement, Calculation, Evaluation,* Oberursel, Germany: DGM informationsgesellschaft, pp. 87–98.

Hougardy, H. (1992), in: W. Jäniche *et al.* (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 1, *Applications,* Düsseldorf: Verlag Stahleisen.

Hougardy, H. (1994), in: W. Gerhartz (Ed.), *Ullmann's Encyclopedia of Industrial Chemistry,* Vol. A25, *Steel,* Weinheim: VCH.

Hougardy, H. P., Yamazaki, K. (1986), *Steel Res.* **57,** 466–471.

Houldcroft, P. T. (1992), in: F. B. Pickering

(Ed.), *Constitution and Properties of Steels,* Materials Science and Technology Vol. 7, Weinheim: VCH, pp. 739–776.

Inden, G., Franke, P., Knoop, U. (1992), "Computer simulation of phase transformations and microstructure evolution," in: *Proceedings of CAMSE '92, the Second International Conference on Computer Applications to Materials and Molecular Science and Engineering,* Yokohama: Elsevier Science Publishers.

International Technical Information Institute (1990), *Handbook of Comparative World Steel Standards,* New York: American Society for Testing and Materials.

Irving, W. R. (1993), *Continuous Casting of Steel,* Bd. 584, London: Institute of Metals.

Kalla, U. (1994), "Steel," in: W. Gerhartz (Ed.), *Ullmann's Encyclopedia of Industrial Chemistry,* Vol. A25, *Steel,* Weinheim: VCH.

Kaspar, R., Pawelski, O. (1985), *Steel Res.* **56,** 477–481.

Kern, A., Schriever, U. (1993), "Computer-simulation of the austenite grain growth and recrystallization during thermomechanical rolling of microalloyed HSLA-steels," in: *Recrystallization '92, Proceedings of the International Conference on Recrystallization and Related Phenomena,* Mathematical Science Forum Vols. 113–115, San Sebastian, Spain: Trans-Technical Publishers, pp. 349–354.

Kirkaldy, J. S. (1993), "Development of Thermodynamic and Kinetic Models as a Basis for Microstructure and Mechanical Property Prediction in Steels," in: *Proceedings of COMP '93, International Conference on Computer-Assisted Materials Design and Process Simulation,* Tokyo: ISIJ.

Kozasu, I. (1992), in: F. B. Pickering (Ed.), *Constitution and Properties of Steels,* Materials Science and Technology Vol. 7, Weinheim: VCH, pp. 183–217.

Liscic, B., Tensi, H. M., Luty, W. (1992), *Theory and Technology of Quenching,* Berlin: Springer.

Massalski, B. (1990), *Binary Alloy Phase Diagrams,* 2nd ed., Vols. 1–3, Metals Park, OH: American Society for Metals.

Mitchell, M. R., Landgraft, R. W. (Eds.) (1993), *Advances in Fatigue Lifetime Predictive Techniques,* Vol. 2, New York: American Society for Testing and Materials.

Mughrabi, H. (Ed.) (1993), *Plastic Deformation and Fracture of Materials,* Materials Science and Technology Vol. 6, Weinheim: VCH.

Müsgen, B., de Boer, H., Fröber, H., Peteresen, J. (1993), in: W. Jäniche, W. Dahl, H. F. Klärner, W. Pitsch, D. Schauwinhold, W. Schlüter, H. Schmitz (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 2, *Applications,* Düsseldorf: Verlag Stahleisen, pp. 6–60.

Naylor, D. J., Cook, W. T. (1992), in: F. B. Pickering (Ed.), *Constitution and Properties of Steels,* Materials Science and Technology Vol. 7, Weinheim: VCH, pp. 433–488.

Nicholson, A., Thornton, D. S. (1992), in: F. B. Pickering (Ed.), *Constitution and Properties of Steels,* Materials Science and Technology Vol. 7, Weinheim: VCH, pp. 95–146.

Nilles, P. (Ed.) (1992), *Gmelin-Durrer Metallurgy of Iron,* Vol. 12, Berlin: Springer.

Oeters, F. (1994), *The Metallurgy of Steelmaking,* Düsseldorf: Verlag Stahleisen.

Olson, G. B., Owen, W. S. (1991), *Martensite,* Metals Park, OH: American Society for Metals.

Olson, D. L., Siewert, Th. A., Liu, St., Edwards, G. R. (1993), in: W. H. Cubberly, H. Baker, D. Benjamin (Eds.), *ASM Handbook,* Vol. 6, *Welding, Brazing, and Soldering,* Metals Park, OH: American Society for Metals.

Orlich, J., Pietrzeniuk, H.-J. (1976), *Atlas zur Wärmebehandlung der Stähle,* Vol. 4, Düsseldorf: Verlag Stahleisen.

Orlich, J., Rose, A. (1973), *Atlas zur Wärmebehandlung der Stähle,* Vol. 3, Düsseldorf: Verlag Stahleisen.

Othani, H. (1992), in: F. B. Pickering (Ed.), *Constitution and Properties of Steels,* Materials Science and Technology Vol. 7, Weinheim: VCH, pp. 147–182.

Pepperhoff, W. (1992), in: W. Jäniche *et al.* (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 1, *Fundamentals,* Düsseldorf: Verlag Stahleisen, pp. 379–411.

Petzow, G. (1994), *Metallographisches, Keramographisches Plastographisches Ätzen,* 6. überarb. Aufl., Stuttgart: Verlag Gebr. Borntraeger.

Petzow, G., Effenberg, G. (1991), *Ternary Alloys,* Weinheim: VCH.

Pickering, F. B. (1992), in: F. B. Pickering (Ed.), *Constitution and Properties of Steels,* Materials Science and Technology Vol. 7, Weinheim: VCH.

Pitsch, W., Sauthoff, G. (1992), in: W. Jäniche, W. Dahl, H. F. Klärner, W. Pitsch, D. Schauwinhold, W. Schlüter, H. Schmitz (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 1, *Fundamentals,* Düsseldorf: Verlag Stahleisen, pp. 1–166.

Predel, B. (Ed.) (1993), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology,* Group IV, *Macroscopic and Technical Properties of Matter,* Vol. 5, *Phase Equilibria, Crystallographic and Thermodynamic Data of Binary Alloys,* Subvol. X, Berlin: Springer.

Raghavan, V. (1987–1992), *Phase Diagrams of Ternary Iron Alloys,* Pts. 1–6, London: Institute of Metals.

Rayson, H. W. (1992), in: F. B. Pickering (Ed.), *Constitution and Properties of Steels,* Materials Science and Technology Vol. 7, Weinheim: VCH, pp. 583–640.

Rose, A., Hougardy, H. P. (1972), *Atlas zur Wärmebehandlung der Stähle,* Vol. 2, Düsseldorf: Verlag Stahleisen.

Sachse, M. (1994), *Damascus Steel, Myth—History—Technology—Application,* Düsseldorf: Verlag Stahleisen.

Schauwinhold, D. (1994), in: W. Gerhartz (Ed.), *Ullmann's Encyclopedia of Industrial Chemistry,* Vol. A25, *Steel,* Weinheim: VCH, pp. 237–277.

Schrader, A., Rose, A. (1989), *De Ferri Metallographia,* Vol. 2, *Structure of Steels,* Düsseldorf: Verlag Stahleisen.

Schrewe, H. F. (1989), *Continuous Casting of Steel,* Düsseldorf: Verlag Stahleisen.

Schulz-Beenken, A. S., Hougardy, H. P. (1989), in: E. Hornbogen, H. Jost (Eds.), *The Martensitic Transformation in Science and Technology,* Oberursel, Germany: DGM Informationsgesellschaft, p. 473.

Speich, G. R. (1992), *Fundamentals of Aging and Tempering in Bainitic and Martensitic Steel Products,* Warrendale, PA: AIME Iron & Steel Society.

Stäblein, H., Arntz, H.-E. (1993), in: W. Jäniche, W. Dahl, H. F. Klärner, W. Pitsch, D. Schauwinhold, W. Schlüter, H. Schmitz (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 2, *Applications,* Düsseldorf: Verlag Stahleisen, pp. 534–546.

SteCal 2.0 (software), Metals Park, OH: American Society for Metals.

Stevens, K. T., Davies, A. (1992), in: F. B. Pickering (Ed.), *Constitution and Properties of Steels,* Materials Science and Technology Vol. 7, Weinheim: VCH, pp. 777–804.

Tacke, B., Forch, F., Sartorius, K., von den Steinen, A., Vetter, K. (1993), in: W. Jäniche, W. Dahl, H. F. Klärner, W. Pitsch, D. Schauwinhold, W. Schlüter, H. Schmitz (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 2, *Applications,* Düsseldorf: Verlag Stahleisen, pp. 118–176.

Täffner, K. (1993), in: W. Jäniche, W. Dahl, H. F. Klärner, W. Pitsch, D. Schauwinhold, W. Schlüter, H. Schmitz (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 2, *Applications,* Düsseldorf: Verlag Stahleisen, pp. 677–690.

Tenckhoff, E., Vöhringer, O. (1990), *Microstructure and Mechanical Properties of Materials,* Oberursel, Germany: DGM Informationsgesellschaft.

Tenhaven, U., Guinomet, Y., Horstmann, D., Meyer, L., Pappert, W. (1992), in: W. Jäniche, W. Dahl, H. F. Klärner, W. Pitsch, D. Schauwinhold, W. Schlüter, H. Schmitz (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 1, *Fundamentals,* Düsseldorf: Verlag Stahleisen, pp. 628–647.

Totten, G. (1992), *Quenching and Distortion Control,* Metals Park, OH: American Society for Metals.

Trumann, J. E. (1992), in: F. B. Pickering (Ed.), *Constitution and Properties of Steels,* Materials Science and Technology Vol. 7, Weinheim: VCH, pp. 527–582.

Unterweiser, P. M. (1993), *Worldwide Guide to Equivalent Irons and Steels,* Metals Park, OH: American Society for Metals.

Verein Deutscher Eisenhüttenleute (Ed.) (1994), *Register of European Steels,* Düsseldorf: Verlag Stahleisen.

Vöge, H. (1993), in: W. Jäniche, W. Dahl, H. F. Klärner, W. Pitsch, D. Schauwinhold, W. Schlüter, H. Schmitz (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 2, *Applications,* Düsseldorf: Verlag Stahleisen, pp. 707–727.

Van der Voort, G. (1991), *Atlas of Time-Temperature Diagrams for Irons and Steels,* Metals Park, OH: American Society for Metals.

Wayman, C. M., Perkins, J. (Eds.) (1993), *Proceedings of the International Conference on Martensitic Transformations (ICOMAT-92),* Monterey, CA: Institute of Advanced Studies.

Wegst, C. W. (1995), *Key to Steel,* Düsseldorf: Verlag Stahleisen.

Weßling, W., Heimann, W. (1993), in: W. Jäniche *et al.* (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 2, *Applications,* Düsseldorf: Verlag Stahleisen, pp. 546–555.

Weßling, W., Oppenheim, R. (1993), in: W. Jäniche *et al.* (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 2, *Applications,* Düsseldorf: Verlag Stahleisen, pp. 433–444.

Wiegels, H. (1994), in: W. Gerhartz (Ed.), *Ullmann's Encyclopedia of Industrial Chemistry,* Vol. A25, *Steel,* Weinheim: VCH, pp. 133–137.

Wilmes, S., Becker, H.-J., Krumpholz, R., Verderber, W. (1993), in: W. Jäniche *et al.* (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 2, *Applications,* Düsseldorf: Verlag Stahleisen, pp. 302–373.

Williams, S. L. (1992), in: W. H. Cubberly, H. Baker, D. Benjamin (Eds.), *Metals Handbook* Vol. 4, *Heat Treating,* Metals Park, OH: American Society for Metals.

Further Reading

Bunge, J. (1988), *Directional Properties of Materials,* Oberursel, Germany: DGM Informationsgesellschaft.

Cahn, R. W., Haasen, P. (1983), *Physical Metallurgy,* Amsterdam: North-Holland.

Cottrell, A. (1988), *Introduction to the Modern Theory of Metals,* London: The Institute of Materials.

Davis, J. R. (1991), in: W. H. Cubberly, H. Baker, D. Benjamin (Eds.), *Metals Handbook,* Vol. 4, *Heat Treating,* Metals Park, OH: American Society for Metals.

Deo, B., Boom, R. (1993), *Fundamentals of Steelmaking Metallurgy,* Englewood Cliffs, NJ: Prentice-Hall.

Evans, R. W., Wilshire, B. (1993), *Introduction to Creep,* London: The Institute of Materials.

Gerhartz, W. (Ed.) (1994), *Ullmann's Encyclopedia of Industrial Chemistry,* Vol. A25, *Steel,* Weinheim: VCH.

International Technical Information Institute

(1990), *Handbook of Comparative World Steel Standards,* New York: American Society for Testing and Materials.

Irving, W. R. (1993), *Continuous Casting of Steel,* London: The Institute of Materials.

Jäniche, W., Dahl, W., Klärner, H. F., Pitsch, W., Schauwinhold, D., Schlüter, W., Schmitz, H. (Eds.), *Steel—A Handbook for Materials Research and Engineering,* Vol. 2, *Applications,* Düsseldorf: Verlag Stahleisen.

Krauss, G. (1990), *Steels: Heat Treatment and Processing Principles,* Metals Park, OH: American Society for Metals.

Liscic, B., Tensi, H. M., Luty, W. (1992), *Theory and Technology of Quenching,* Berlin: Springer.

Nilles, P. (Ed.) (1992), *Gmelin-Durrer Metallurgy of Iron,* Vol. 12, Berlin: Springer.

Oeters, F. (1994), *The Metallurgy of Steelmaking,* Düsseldorf: Verlag Stahleisen.

Pickering, F. B. (1978), *Physical Metallurgy and the Design of Steels,* London: Applied Science Publishers.

Pickering, F. B. (Ed.), *Constitution and Properties of Steels,* Materials Science and Technology Vol. 7, Weinheim: VCH.

Tenckhoff, E., Vöhringer, O. (1990), *Microstructure and Mechanical Properties of Materials,* Oberursel, Germany: DGM Informationsgesellschaft.

Thelning, K.-E. (1984), *Steel and Its Heat Treatment,* 2nd Ed., London: Butterworths.

Totten, G. (1992), *Quenching and Distortion Control,* Metals Park, OH: American Society for Metals.

Tyrkiel, E. (1986), *Multilingual Glossary of Heat Treatment Terminology,* London: The Institute of Materials.

STIRLING ENGINES

ROELF J. MEIJER, *Product Innovation Center, Inc., Saline, Michigan, U.S.A.*
BENJAMIN ZIPH, *Stirling Thermal Motors, Inc. Ann Arbor, Michigan, U.S.A.*

Introduction 1
1. **Physical and Technical Principles** 2
1.1 The Ideal (Isothermal, Reversible) Stirling Cycle 3
1.1.1 Harmonic Volume Variations, Finite-Volume Regenerator 3
1.2 The "Real" Cycle 6
1.2.1 Adiabatic Variable Volumes with Dedicated Heat Exchangers 6
1.2.2 The Major Loss Mechanisms 6
1.2.3 The Effect of "Dead Volume" 7
2. **Design, Function, and Operation** 8
2.1 Mechanical Arrangements 8
2.1.1 The Displacer Engine 8
2.1.1.1 The Free-Piston Engine 8
2.1.2 The Double-Acting Engine 9
2.1.3 Other Configurations 9
2.2 Operational Controls 9
2.2.1 Variable Charge 9
2.2.2 Variable Displacement 10
2.3 Critical Components 11
2.3.1 Regenerator 11
2.3.2 Seals 11
2.3.3 High-Temperature Materials 12
3. **Manufacture** 12
3.1 V Engine (SPS V-160) 12
3.2 Displacer Engine with Rhombic Drive (GM's GPU-3) 12
3.3 Double-Acting, Variable-Charge, Crank Drive (MTI MOD II) 12
3.4 Double-Acting, Variable-Charge, Swashplate Drive (Philips' 4-215) 13
3.5 Double-Acting, Variable-Displacement, Swashplate Drive (STM4-120) 14
3.6 Free-Piston Engine (Sunpower, Inc.) 14
4. **Applications** 15
4.1 Special Energy Sources 16
4.1.1 Solar-Energy Conversion 16
4.1.2 Submarine Vessels 16
4.1.3 Power from Biomass 16
4.2 Remote, Maintenance-Free Operation 17
4.2.1 Space Power 17
4.3 Quiet, Clean, Efficient Power 17
4.3.1 Generator Sets 17
4.3.2 Vehicle Propulsion 17
5. **Economic Aspects** 17
Further Reading 17

INTRODUCTION

In the beginning of the 19th century, there was only one heat-driven machine (prime mover), the steam engine. However, the components of such a steam engine were rather primitive with items such as the boiler constructed of inadequate materials like cast iron or bronze. Many accidents occurred with those pressure vessels made of such brittle material, even when the pressure level was only atmospheres. Therefore, several inventors were looking to replace the working medium of water with air in order to avoid storing so much energy.

The invention in 1816 by Robert Stirling, a Scottish clergyman, was the most remarkable. It was one of the first practical regenerative engine cycles. Figure 1 shows the 1816 patent drawing of the first hot-air engine, later called a Stirling engine. The first model built according to the pictured scheme was in use by 1818 pumping water from a quarry.

Hot-air engines were used in considerable numbers during the 19th century. It was

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

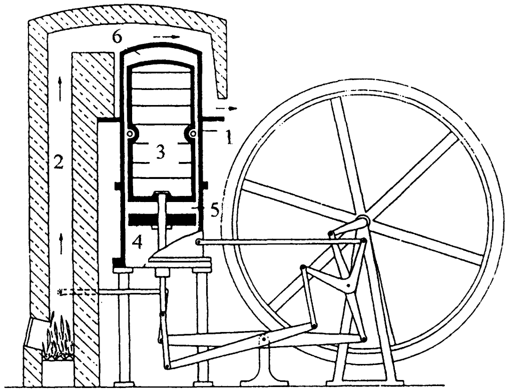

FIG. 1. A long, vertical cylinder 1 is heated on top by the flue gases from the furnace 2. The bottom is cooled either with water or by convection with the atmosphere. The cylinder contains the displacer 3, centered by small rollers, and the power piston 4. As the displacer moves up and down, air in the cylinder 1 passes alternately from the cold space 5 and the hot space 6 via a regenerator and back. The regenerator (not shown here) is in the annular space between the displacer and the cylinder and probably consisted of thin wire wound around the displacer. The air is thus alternately in the cold and the hot spaces and undergoes temperature and pressure variations that cause it to perform work on the piston 4.

even thought, for a time, that this type of prime mover might rival the steam engine as a source of power. However, the comparative weight and volume of the device per unit of delivered power were huge. It is no small wonder that the internal combustion engine that arrived in the late 19th century, with its much greater specific power and higher efficiency, displaced this hot-air engine completely—despite a certain charm that the latter possessed.

A renewed interest came in 1938 when N. V. Philips in the Netherlands started an investigation of very small generator sets and chose the hot-air engine for that purpose. Later Philips was involved in much bigger engines, up to 300 kW per cylinder, during 40 years of research and development. In 1954 the name "hot-air engine" was changed to "Stirling engine" with the introduction of the rhombic-drive engine. This device used hydrogen as the working medium, which rendered the reference to "hot air" irrelevant.

During that time, N. V. Philips acquired several licensees to the technology worldwide including General Motors and Ford in the United States, M.A.N. and M.W.M. in Germany, and United Stirling in Sweden. When the last licensee, Ford, terminated its activity in 1979, N. V. Philips also stopped their research. Stirling Thermal Motors, Inc., a small developing company in the United States, continued the work of Philips on a small scale. In addition, many universities and companies all over the world became interested in the Stirling engine and made worthy contributions to the technology, many times supported in the United States by the Department of Energy (DOE) or National Aeronautics and Space Administration (NASA). Japan also became interested in this cycle and has produced numerous articles describing its research work and prototypes.

1. PHYSICAL AND TECHNICAL PRINCIPLES

The Stirling engine is a prime mover much like the well-known gasoline and diesel engines, where heat of sufficiently high temperature is transformed into mechanical energy. The Stirling engine, however, is based on a closed cycle containing three heat exchangers—heater, regenerator, and cooler—with a variable volume on the cooler side called "cold space" and a variable volume on the heater side called "hot space." These variable volumes are caused by pistons moving with a phase difference to each other. An amount of pressurized working gas, such as helium, is captured between the pistons and manipulated by this arrangement in such a way that heat is supplied to the gas through the wall of the heater; the hot gas is allowed to expand into the hot space, then is pushed through the regenerator where it yields heat in order to be compressed in the cold space at lower temperature. The heat developed by compression is transported through the wall of the cooler into the surrounding space, or ambient. The mechanical energy obtained by expansion at high temperature is always larger than the mechanical energy needed for compression at lower temperature. Therefore, an excess of mechanical energy is obtained.

Because the Stirling engine is based on a closed cycle, its characteristics are different from those of the internal combustion engine. The Stirling engine has a multiple-fuel capability since the heat supplied to the hot heat exchanger can come from any source of sufficiently high temperature. Another advantage occurs when fossil fuel is burned in a well-designed burner. The emissions can be very low since the burner is at atmospheric pressure and exhaust-gas recirculation can be easily applied.

Operational characteristics of a Stirling engine include silent operation because the pressure variation within the engine is sinusoidal (having no sharp peaks) and because combustion is continuous. Another effect of the sinusoidal pressure variation is the constant torque occurring within an engine with four or more cylinders.

Performancewise, an ideal Stirling engine offers the highest efficiency possible, matching that of a Carnot-cycle device operated between the same temperature limits, along with higher specific power. The power is proportional to the pressure of the working gas. With a compact drive design and modern materials, the specific power can be the same as that of the modern gasoline engine. Finally, fast power control is now available with recently developed stroke control of the pistons, making it possible to change the power levels in a fraction of a second.

There have been some issues in the development of Stirling engines, however, that have stopped the technology from reaching the manufacture point. With the closed, pressurized system, the cycles are more difficult to build and operate than an open system. The pressurization of the Stirling engine also makes accessibility to the machine difficult, requiring that the working gas must be vented before there is access to the engine.

An ideal Stirling design requires an oil-free cycle since a lubricating oil and its byproducts will block the gas flow in the heat exchangers. Formerly, the pistons were lubricated with small amounts of oil, but this has shown to be explosive when combined with air as the working medium. Hydrogen has proven to be the best working medium with respect to the performance of the Stirling cycle, and it does not adversely react with any lubrication oil that may be required. However, hydrogen is flammable outside the engine and, if used in the engine, it slowly diffuses through the hot metals of the components. More work has to be done to reduce this leakage.

Piston leakage also has a great influence on the performance of the engine and creates a need for great attention to be given to the seals in this area.

1.1 The Ideal (Isothermal, Reversible) Stirling Cycle

1.1.1 Harmonic Volume Variations, Finite-Volume Regenerator. The ideal Stirling cycle is embodied in the configuration shown schematically in Fig. 2. Two variable-

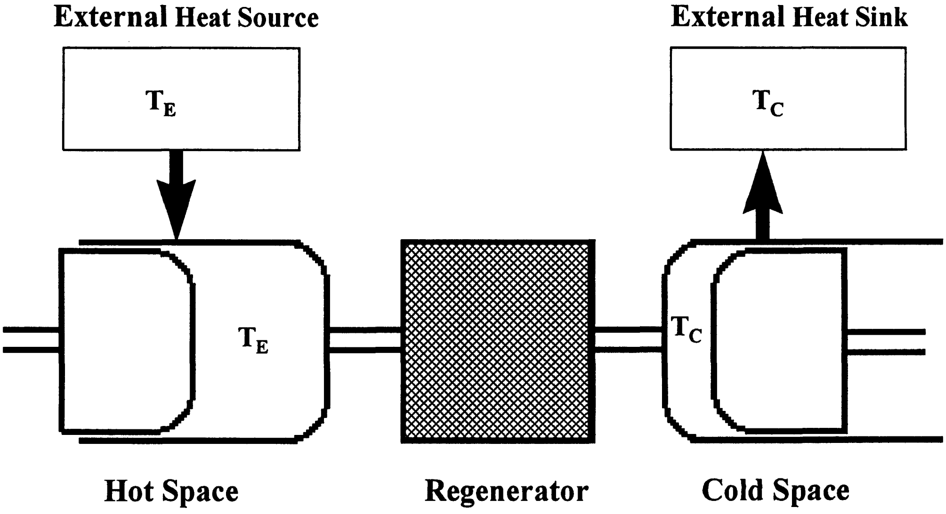

FIG. 2. Stirling engine configuration (schematic).

volume chambers, a cold space and a hot space, are interconnected via a regenerative heat exchanger called a "regenerator." The two pistons undergo harmonic motion in such a way that the volume variation of the cold space lags in phase behind the volume variation of the hot space. The cold and hot spaces are maintained at constant temperatures, T_C and T_E, respectively, by exchanging heat with external hot and cold reservoirs at the same temperatures, T_C and T_E, respectively. The ideal cycle occurs when

1. the gas in the cold and hot spaces undergoes isothermal thermodynamic processes;
2. the regenerator is ideal, i.e., its matrix is in perfect thermal contact with the gas; and
3. the gas is frictionless and flows without dissipating energy so that the instantaneous pressure is uniform throughout the enclosure.

The cold-space volume lags behind the hot-space volume by a phase angle ϕ ($0 < \phi < \pi$) as shown in Fig. 3. This ensures that during expansion (when the total volume increases) most of the gas is in the hot space and during compression (when the total volume decreases) most of the gas is in the cold space, as shown in the figure. Thus, the bulk temperature, and hence the pressure during expansion, is higher than it is during compression, and a surplus of mechanical work is produced by the machine. At the same time, the amount of heat absorbed by the gas from the outside through the enclosure wall during the expansion process is greater than that rejected by the gas to the outside during the compression process. In this way the Stirling engine converts heat supplied from the outside into work done on the pistons by the enclosed gas.

In contrast to other possible representations of the ideal Stirling cycle, the compression and expansion processes here are not separate from the transfer processes: During expansion some of the gas is transferred from the hot space to the cold space, and vice versa during compression. Gas that is

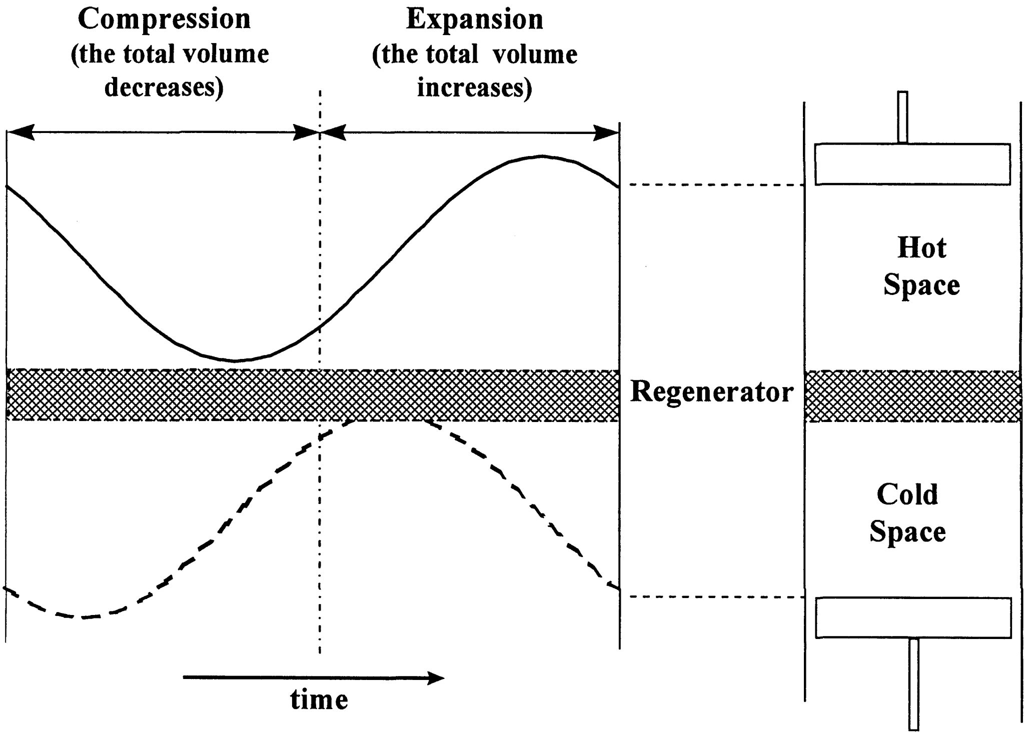

FIG. 3. Piston motion and volume variations in an ideal cycle.

transferred from the hot space to the cold space is cooled en route by storing its heat in the regenerator. This heat is picked up by the gas later in the cycle during the transfer from the cold space to the hot space.

With an ideal regenerator, the processes of storing heat and picking it up are thermodynamically reversible because they occur with no temperature difference between the heat-exchanging elements—the gas and the regenerator matrix. Similarly, the isothermal compression and expansion processes in the variable-volume chambers are reversible because the gas is (ideally) in thermal equilibrium with the enclosure walls with which it exchanges heat. Finally, mechanical equilibrium also prevails throughout the cycle because, with nondissipating gas flow, the instantaneous pressure is uniform throughout the enclosure. In consequence of the thermal and mechanical equilibria, the ideal Stirling cycle is a reversible thermodynamic cycle that converts heat into work at the ideal Carnot efficiency: Efficiency = (net work output)/(heat absorbed from the external hot source) = $1 - T_C/T_E$.

The ideal regenerator prevents the net transfer of enthalpy from the hot space to the cold space: With an ideal regenerator, the enthalpy flow through any imaginary partition in the regenerator, integrated over the part of the cycle during which the flow direction is, say, from the hot space to the cold space, is exactly equal to that integrated over the balance of the cycle when the flow direction is reversed. Therefore, all the heat Q_E absorbed by the gas in the hot space at the source temperature T_E over a cycle is converted into positive work W_E. Similarly, all the work W_C done on the gas in the cold space over a cycle is converted into heat Q_C, which is rejected by the gas in the cold space at the sink temperature T_C, as shown in Fig. 4. Therefore, with $Q_E = W_E$ and $Q_C = W_C$, the net work can be calculated as

$$W_{\text{net}} = W_E - W_C = Q_E - Q_C. \tag{1}$$

Because the compression and expansion processes are isothermal and reversible, the ratio Q_C/Q_E is equal to the absolute temperature ratio T_C/T_E, and the energy conversion efficiency, η, for the ideal Stirling cycle is given by

$$\begin{aligned} \eta &= W_{\text{net}}/Q_E = (Q_E - Q_C)/Q_E \\ &= 1 - Q_C/Q_E = 1 - T_C/T_E. \end{aligned} \tag{2}$$

Different elements of gas in the Stirling cycle undergo, in general, different cyclical processes. Clearly, some of the gas that is in the hot space when the hot space is at its maximum volume will never reach the cold space. Rather, some gas elements will forever shuttle back and forth between the hot space and the regenerator, others will shuttle between the cold space and the regenerator, and some may go completely through the re-

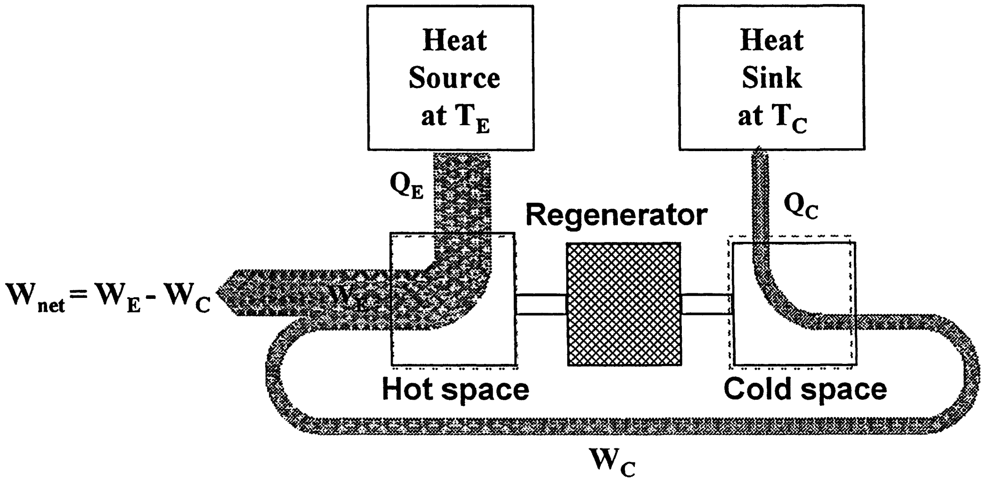

FIG. 4. Energy flow in an ideal Stirling engine.

generator. This is simply a consequence of the heat exchangers containing finite volumes of gas.

1.2 The "Real" Cycle

1.2.1 Adiabatic Variable Volumes with Dedicated Heat Exchangers. In all but the smallest Stirling engines, the wetted areas, the surface areas that allow heat to transfer to the ambient, in the variable-volume chambers are too small to maintain isothermal processes in these chambers. To accommodate the required heat transfer, the hot and cold spaces are each split into variable-volume portions in which little heat exchange takes place and fixed-volume portions where heat exchanges with the outside takes place, as shown in Fig. 5.

The thermodynamic processes in the heat exchangers are ideally isothermal, while in the variable-volume chambers they are practically adiabatic. This makes the cycle no longer reversible. Because the heat exchangers are fixed-volume spaces, energy transport between a heat exchanger and the adjacent variable-volume space must take the form of a net enthalpy flow. This is shown in Fig. 6, which depicts the temperature of a gas element along its path through the source heat exchanger ("heater"). The gas element in Fig. 6 transports the following amount of enthalpy h_i, per unit of mass, from the heater to the expansion space:

$$h_i = c_p(T_H - T_i), \tag{3}$$

where T_H and T_i are the temperatures shown in the figure and c_p is the heat capacity of the gas.

Thus, thermal equilibrium cannot prevail at the interface between the heat exchanger and the adjacent variable-volume space. Even with an ideal heater the mean bulk temperature in the expansion space must be lower than the source temperature. Similarly, the mean bulk temperature in the compression space must be higher than that of the sink. In consequence, the pressure during expansion is lower and during compression higher than the corresponding values of the ideal Stirling cycle. Less heat will be absorbed from the external source, and more heat will be rejected to the external sink, and so a smaller portion of the heat absorbed is converted into net work. In other words, the efficiency of the engine with adiabatic variable-volume chambers and dedicated heat exchangers is lower than the Carnot efficiency of the reversible ideal cycle.

1.2.2 The Major Loss Mechanisms. The efficiency of a real, practical Stirling engine is further reduced by several effects.

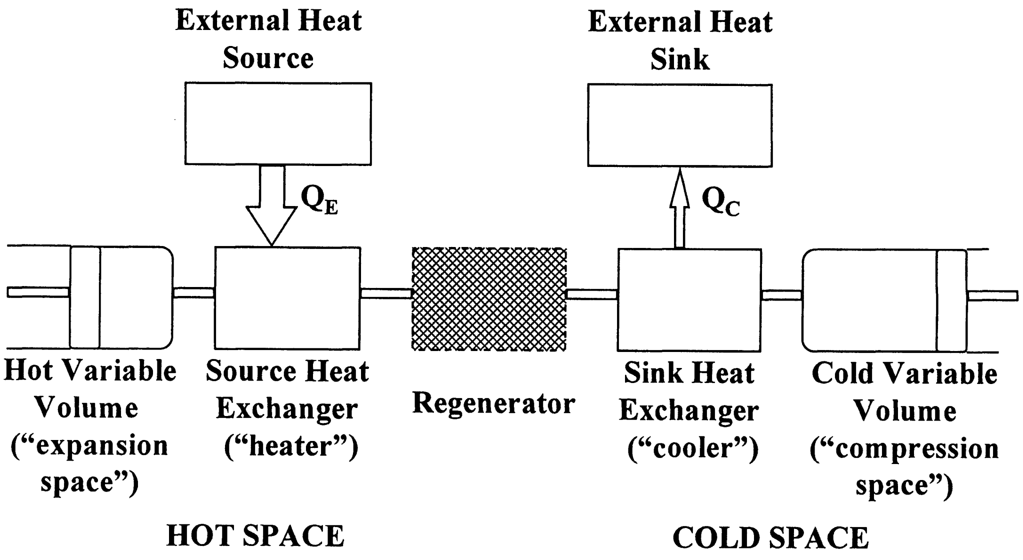
External Heat Source
External Heat Sink
Q_E
Q_C
Hot Variable Volume ("expansion space")
Source Heat Exchanger ("heater")
Regenerator
Sink Heat Exchanger ("cooler")
Cold Variable Volume ("compression space")
HOT SPACE
COLD SPACE

FIG. 5. Stirling engine with adiabatic variable chambers and dedicated heat exchangers.

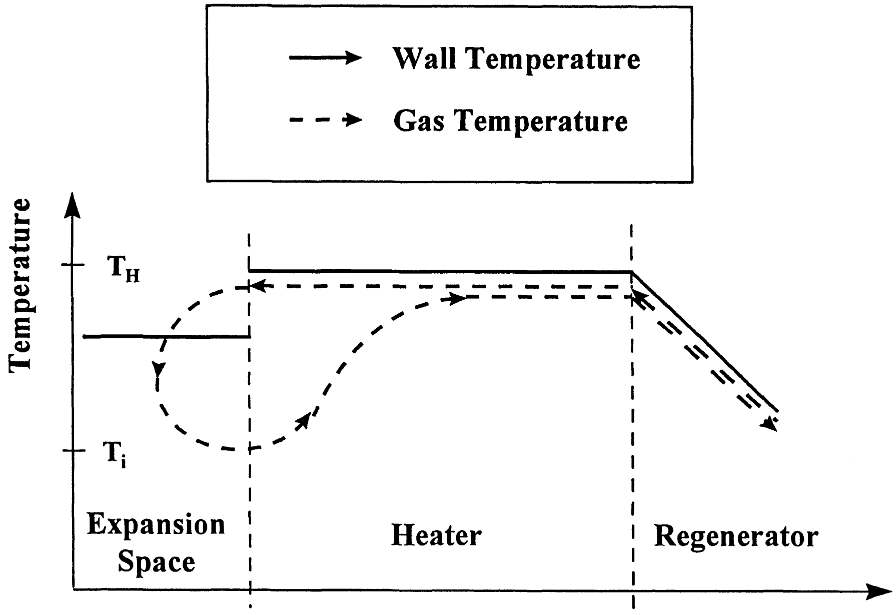

FIG. 6. Temperature of a gas element along its path in a real Stirling engine.

One such item is fluid friction in the heat exchangers, whereby cycle mechanical work is converted back into heat.

A component that can greatly affect the efficiency is a nonideal regenerator. In a typical Stirling engine the amount of heat stored in the regenerator per cycle is about five times larger than the heat absorbed from the external source. Thus every 1% drop in the regenerator effectiveness gives rise to a 5% drop in the engine efficiency. The regenerator is also the major contributor to fluid friction loss. A well-designed regenerator represents the best compromise of effectiveness and flow loss.

Additional losses are associated with a nonideal heater and cooler resulting in a lower bulk expansion temperature, a higher bulk compression temperature, and hence lower efficiency.

Natural and augmented heat conduction from the hot side to the cold side is a direct efficiency loss. Natural conduction takes place primarily in the regenerator, the matrix and housing of which are subjected to a sharp axial thermal gradient. Augmented (and natural) conduction takes place in the annular, blind gap called "appendix gap" between the expansion piston and its cylinder. The expansion cylinder has to be cooled at the end opposite the heater to permit the use of a dry-running piston ring and to protect the lubricated piston drive mechanism. Therefore, it experiences an axial thermal gradient. Natural conduction from the heater to the cooled end of the cylinder is augmented by the hot gas flowing into and out of the appendix gap, as well as by the piston motion.

The exact nature of the piston leakage loss depends on the engine configuration and is difficult to describe in general. Usually, the piston ring separates spaces in which the pressures fluctuate about a common mean pressure but at different phases and/or amplitude. Leakage across the piston ring shifts the phase of the pressure fluctuation closer to that of the total volume variation and reduces the power and the efficiency of the engine.

1.2.3 The Effect of "Dead Volume." Dead volume, or void volume, is that volume enclosed between the two pistons of Fig. 5 that is not swept by either piston. It is made up of the internal volumes of the heater and cooler, the internal void volume of the regenerator, and the volumes of the ducts and passages connecting these components. The larger the ratio of the dead volume to the swept volume (or "displacement") of the engine, the lower the pressure variation amplitude and, hence, the power developed by the engine. Therefore, it may be advantageous to employ compact heat exchangers with small internal volumes.

On the other hand, the dead volume has little or no direct effect on the efficiency. The heat absorbed from the external source is, just as is the power, proportional to the pressure-variation amplitude and is reduced by the same proportion with increasing dead volume. There are various indirect effects of the dead volume on efficiency. Many of these, however, give rise to higher efficiency with increasing dead volume because of smaller pressure variations, making the cycle become more isothermal and sacrificing power.

Dead volume, in itself, is not necessarily a disadvantage. By judiciously sizing and distributing the various dead volumes, the engine designer can trade power for efficiency and obtain an optimal match of the engine to its application and duty cycle.

2. DESIGN, FUNCTION, AND OPERATION

2.1 Mechanical Arrangements

2.1.1 The Displacer Engine. The volume variations necessary for the Stirling cycle can be obtained in several ways. For a single-cylinder engine the most common is the displacer type, the original arrangement designed by Robert Stirling. Figure 7 shows a modern version of the displacer arrangement.

The piston moves up and down by means of a drive mechanism. The displacer is mounted on a rod running through the piston and also moves up and down, by means of this same drive mechanism but with a prescribed phase difference from the piston. The variable volume on the heater side is formed by the stationary wall of the top of the cylinder (labeled "6" in Fig. 7) and the moving top of the displacer (labeled "1" in the figure)—a piston with an insulated dome on top of it—separating the hot space and the cold space. The displacer moves the gas between the hot (or expansion) space and the cold (or compression) space via the heat exchangers, which heat or cool the gas. In the meantime the piston compresses the gas when it is in the cold space or causes the expansion when the gas is in the hot space.

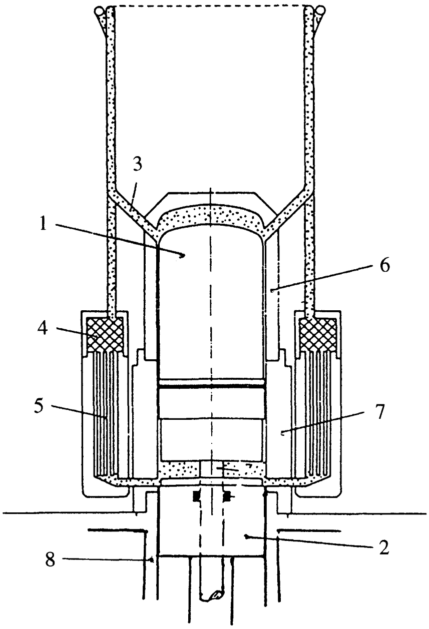

FIG. 7. Displacer and piston arrangement of a single-cylinder displacer engine. 1, Displacer; 2, piston; 3, heater; 4, regenerator; 5, cooler; 6, hot cylinder; 7, cold displacer cylinder; 8, cold piston cylinder.

With a displacer engine the compression work is obtained from the inertia of a moving mass such as a flywheel, and the expansion work is partially stored in this mass with the balance creating useful energy.

2.1.1.1 The Free-Piston Engine. The *free-piston* Stirling engine began its evolution in the early 1960s largely as a result of the work of Professor W. T. Beale of Sunpower, Inc., attempting to avoid the sealing and lubrication difficulties of the crank-type Stirling engines. Since then, efforts by many designers have resulted in a machine that is mechanically simple and designed for long life, as shown in Fig. 8.

The free piston is of the displacer type, but the displacer and the working piston are designated as "free" to indicate that they are not connected to any drive; rather, they are suspended within the "bounce space" by a

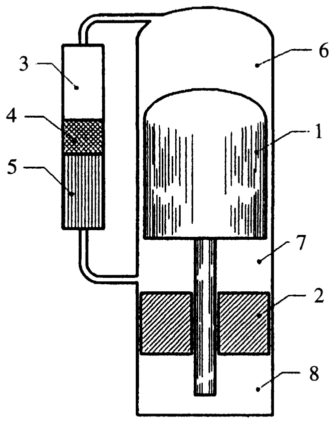

FIG. 8. Typical free-piston Stirling engine arrangement. 1, Free displacer; 2, free piston; 3, heater; 4, regenerator; 5, cooler; 6, hot (expansion) space; 7, cold (compression) space; 8, buffer (bounce) space.

gas buffer. As the displacer oscillates at a rate of 50 to 60 Hz, the working gas shuttles between the cold space and hot space causing pressure variations that move the piston up and down. To accomplish the desired gas cycle, the displacer and piston are tuned as mechanical spring-mass-damper resonators, which ensure the correct phase relation. The power is obtained from the piston by connecting it with a permanent-magnet linear alternator within the sealed cylinder, thus converting piston motion into electricity.

2.1.2 The Double-Acting Engine. A modern multiple-cylinder Stirling engine is a "double-acting" configuration and can be a three-, four-, five-, or six-cylinder engine. A four-cylinder type is shown schematically in Fig. 9. There is only one piston per cylinder. The compression of the gas takes place under the piston (the cold space), and its expansion takes place on top of the piston (the hot space) of the adjacent cylinder. These events are accomplished by the movement of the pistons, each of which differs in phase from its successor by the same angle. After the expansion in the hot space, the gas is shuttled through the heater, regenerator, and cooler into the cold space. There the gas is compressed and forced back to the hot space. Completion of one cycle of the engine results in one revolution and an excess of work.

2.1.3 Other Configurations. In the first days of renewed interest in the hot-air engine (1938), several small models were born. One example of these configurations that survived is shown in Fig. 10. The variable volumes, with a phase difference of 90°, are obtained by a two-cylinder V drive.

2.2 Operational Controls

2.2.1 Variable Charge. The power of a Stirling engine can be controlled by changing the torque of the engine. The torque is

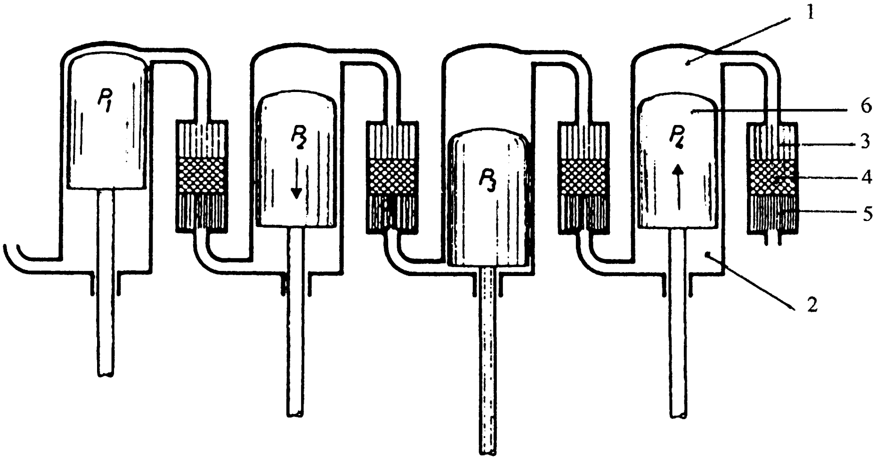

FIG. 9. Schematic arrangement of a four-cylinder double-acting Stirling engine. 1, Hot (expansion) space; 2, cold (compression) space; 3, heater; 4, regenerator; 5, cooler; 6, piston.

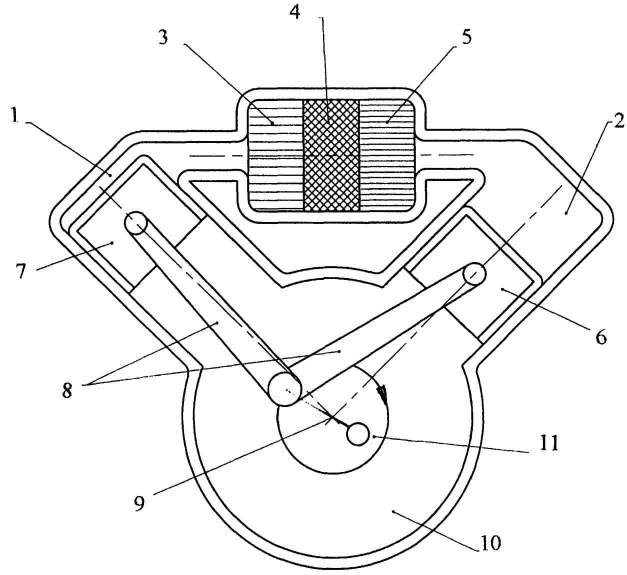

FIG. 10. V engine with cylinders at 90°. 1, Hot (expansion) space; 2, cold (compression) space; 3, heater; 4, regenerator; 5, cooler; 6, compression piston; 7, expansion piston; 8, connecting rods; 9, crankshaft; 10, crankcase; 11, counter weight.

approximately proportional to the pressure of the working gas. Thus, by an increase in the pressure, the torque and, therefore, the power are increased and vice versa. Figure 11 explains how this is done with a single-cylinder engine. This system works quite satisfactorily for single-cylinder engines but becomes rather complex for multiple-cylinder engines.

2.2.2 Variable Displacement. When it became clear that the power control of multiple-cylinder Stirling engines was one of the bottlenecks for commercialization of the en-

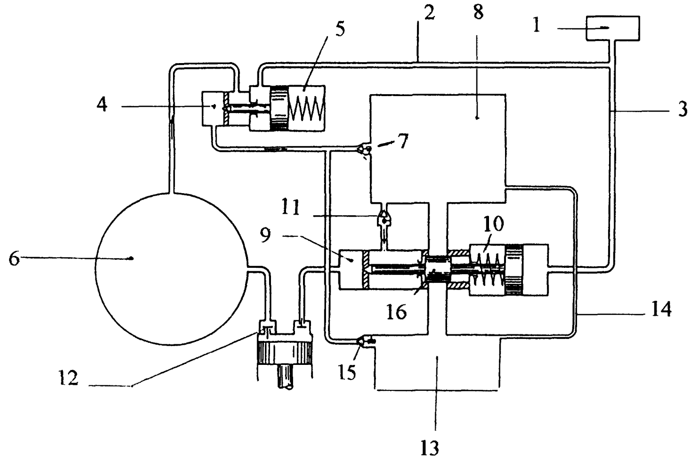

FIG. 11. Variable charge of single-cylinder engine. When the power has to increase, the regulator 1 opens the valve 4, by increasing the hydraulic pressure in the pipes 2 and 3 and in the valve mechanism 5, by which high-pressure gas flows from the high-pressure vessel 6 into the engine 8 and buffer 13 via check valves 7 and 15. This can be done relatively quickly. When the power has to decrease, the regulator opens the valve 9, by decreasing the hydraulic pressure in the valve mechanism 10, and the gas flows via the check valves 11 and 16 and compressor 12 from the engine and buffer back into the vessel 6. However, this would occur slowly with a small compressor, and, therefore, a slide throttle 16 is introduced that connects the cycle with the buffer space 13 to ensure a rapid response when the load is reduced suddenly. The capillary 14 is a connection between the cold space and the buffer to ensure mean pressure in the buffer space in a stationary situation.

gine, a mechanical device was found that could change the stroke of the pistons, thus changing the displacement and power, while keeping all of the working gas inside the engine.

The *swashplate* drive incorporates a disk fixed at a certain angle to the shaft, over which piston rods slide to cause reciprocating movement. By changing the angle of the swashplate with respect to the shaft, the stroke of the pistons is changed causing a change in the power. This mechanism has shown to be very successful in the Stirling engine and is shown schematically in Fig. 12.

2.3 Critical Components

2.3.1 Regenerator. The regenerator is the heart of the high-performance Stirling engine. It ensures that the two processes, compression and expansion, occur over the greatest temperature difference possible. Therefore, this heat exchanger should have a structure allowing very high heat transfer as well as high heat capacity per unit of volume, low heat conduction in an axial direction, and, finally, very low flow resistance for the gas shuttling between the cold space and the hot space. The best performance to date is a regenerator made of layers of wire cloth where the wire diameter, the mesh, and the number of layers are optimized for a given engine.

2.3.2 Seals. The two types of seals in the Stirling engine can be distinguished as either internal (sealing to the pistons) or external (with relation to the ambient). The internal piston rings must be made of dry-running material in order to keep the heat exchangers free of oil. Not only should the leakage along the pistons be small, but in a double-acting engine, the gas mass should not be allowed to migrate to adjacent cylinders. While different designs have been found to obtain such a balance, most come with the cost of increased friction.

External leakage to the ambient has always posed one of the most severe problems with the Stirling engine: preventing gas leakage out of the high-pressure cylinder along a lubricated moving rod while avoiding oil penetration into the cylinder. One of the latest solutions requires a pressure vessel around the drive and uses a rotating shaft seal. In the case of a generator set, the generator can be included within the pressure vessel as is done in the free-piston machine.

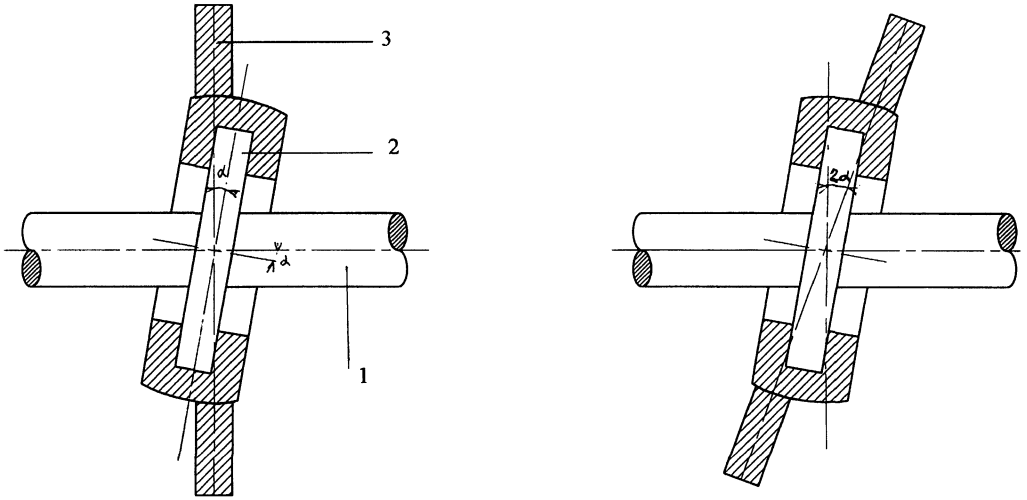

FIG. 12. The variable swashplate mechanism. 1, Crankshaft; 2, fixed disk; 3, variable swashplate. The fixed disk makes an angle α with a plane perpendicular to the shaft. The variable swashplate has a bearing fitting around the fixed disk, which also makes an angle α with the swashplate itself. By turning the swashplate relative to the shaft, the angle that the swashplate makes with the perpendicular plane can be changed from 0° to 2α, thus changing the piston stroke from zero to maximum. See Fig. 15 for mode of application.

2.3.3 High-Temperature Materials. In Robert Stirling's time, cast iron or bronze were the only high-temperature materials available, yet these proved to be brittle and unsuitable for a high-pressure device. Today, there is a wide choice of high-strength, high-temperature materials. The use of heat-resistant materials for a modern Stirling engine can be divided into two categories: castings for the hot cylinder and regenerator cups, and drawn tubes for the hot heat exchanger. However, since the cost of the hot parts of the engine is most critical, there are also attempts to cast the tubes, with or without fins.

In the future, heat- and creep-resistant ceramics are also expected to play a major role as ideal Stirling engine materials. Innovative fabrication techniques are also being investigated in relation to common material properties that could be exploited.

3. MANUFACTURE

3.1 V Engine (SPS V-160)

Stirling Power Systems, Inc., a subsidiary of United Stirling, has developed a generator set that produces 5 to 10 kW, based on a V-type configuration. This machine was also used in recreational vehicles as a cogenerator, heating the water and providing electric power without the use and dangerous storage of propane fuel.

3.2 Displacer Engine with Rhombic Drive (GM's GPU-3)

The rhombic drive has played a very important role in the development of the Stirling engine. It was invented in 1953 at Philips. It led the way to engines with high power because it allowed a completely balanced, single-cylinder engine suitable for high pressure. In addition, this engine was the first optimized for hydrogen as the working gas.

Figure 13 shows a schematic of the rhombic engine. The desired phase difference is obtained from the design of the crank-connecting rod mechanism and, if the reciprocating weight of the piston part is equal to that of the displacer part, only a first harmonic of the inertia forces is left that can be balanced by two counter-rotating weights.

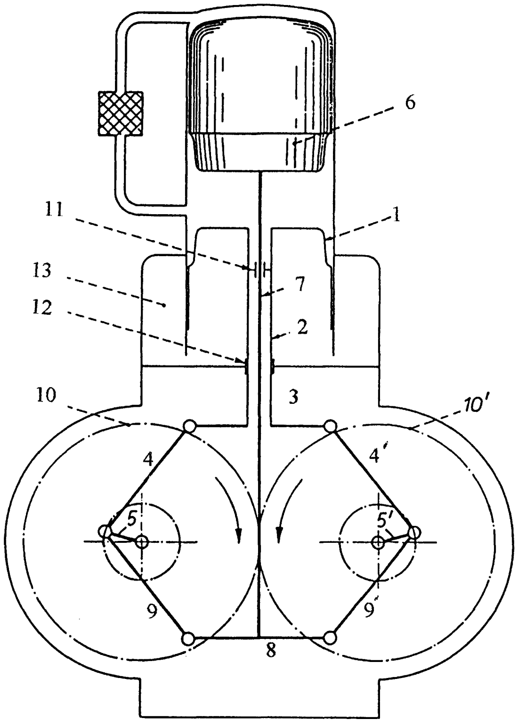

FIG. 13. The rhombic drive mechanism. 1, Piston; 6, displacer; 5 and 5′, cranks on two shafts that rotate in opposite directions and are coupled by the gear wheels 10 and 10′; 4 and 4′, connecting rods pivoting on the ends of a yoke 3, fixed to the hollow piston rod 2; 9 and 9′, connecting rods pivoting on the ends of a yoke 8, fixed to the displacer rod 7, which passes through the hollow piston rod; 11 and 12, stuffing boxes; 13, buffer space filled with gas at the mean pressure of the cycle.

A generator set was manufactured by General Motors according to this principle. This ground power unit (GPU) was produced for the U.S. Army. It was a 3- to 5-kW generator set with a 10-HP (7.5 kW) single-cylinder engine with rhombic drive.

3.3 Double-Acting, Variable-Charge, Crank Drive (MTI MOD II)

Funded by contracts from the U.S. Department of Energy via the NASA Lewis Research Center, Mechanical Technology, Inc. (MTI), has developed a double-acting engine on a V drive for direct propulsion of an automobile, shown in Fig. 14. This is a contin-

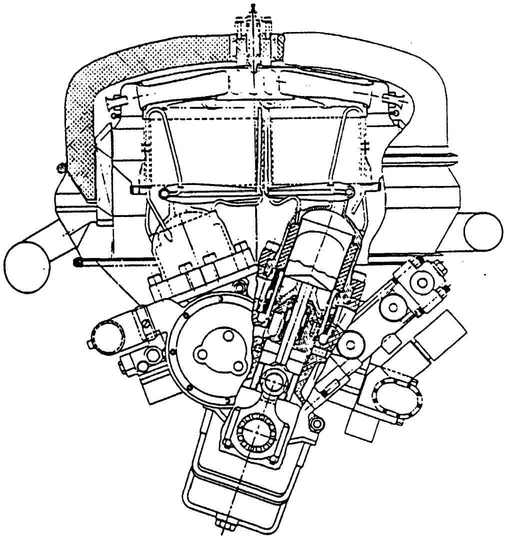

FIG. 14. Double-acting, variable-charge, crank drive engine by United Stirling and MTI (MOD II).

uation of the research originated by United Stirling of Sweden, a licensee of N. V. Philips. United Stirling had successfully developed this same configuration for submarine propulsion, which is in manufacture today.

3.4 Double-Acting, Variable-Charge, Swashplate Drive (Philips' 4-215)

Acting through its licensee, N. V. Philips created a few direct-propulsion engines producing 130 kW for the Ford Torino automobile, which was demonstrated in 1976 in the United States. In order to make the engine as compact as possible, a double-acting swashplate configuration was chosen together with a rotating, regenerative, ceramic preheater, drawing upon Ford's experience with gas-turbine technology. Figure 15 shows a schematic of this configuration with the actual engine appearing in Fig. 16. The power control was obtained by the variable-charge method of Fig. 11, but the Torino system required much more complexity to achieve the response times needed for the automotive application.

Ford's interest in the Stirling engine was its low-emission capability. Table 1 shows the emission levels obtained without the use of a catalytic converter. While these figures were remarkable against the standards of 1970, the emissions levels of today are still impressively low when compared to other technologies. This same characteristic is driving the quest for Stirling engines even more today.

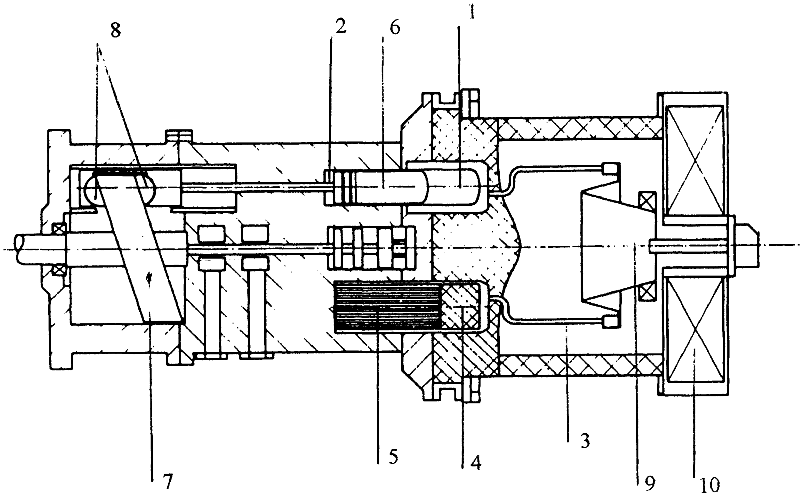

FIG. 15. Double-acting, variable-charge, swashplate drive engine (Philips' 4-215), shown here schematically. As part of the Ford–Philips license agreement, N. V. Philips manufactured four of these 130-kW engines for Ford's Torino automobiles. 1, Hot space; 2, cold space; 3, heater; 4, regenerator; 5, cooler; 6, piston; 7, swashplate; 8, sliders; 9, burner; 10, preheater.

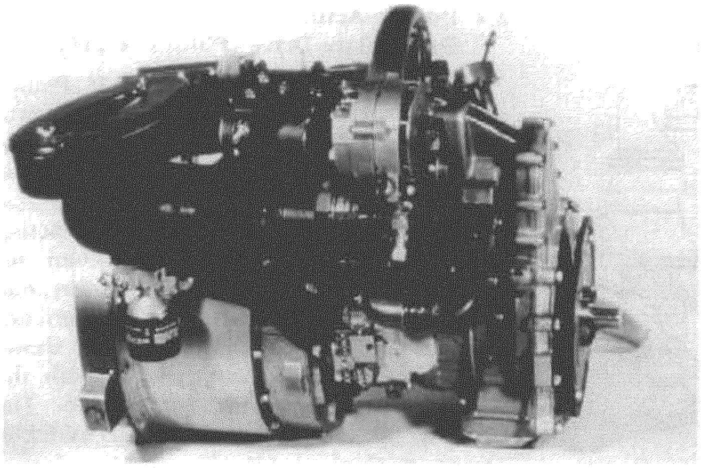

FIG. 16. The Ford–Philips 4-215 engine.

3.5 Double-Acting, Variable-Displacement, Swashplate Drive (STM4-120)

When N. V. Philips terminated its Stirling engine activities in 1979, Stirling Thermal Motors, Inc. (STM), was established in the United States and acquired all rights to the past technology and computer simulation programs used in the design process. Rather than addressing the difficulty of introducing a new, competitive engine to the world, STM sought economical applications where internal-combustion engines could not be used. The focus was to concentrate on one engine size and configuration, solve all the existing problems of the technology, and make the design absolutely reliable.

The chosen engine size was 25 kW at 1800 rpm with an 800 °C heater-wall temperature. This configuration was designated the STM4-120: four cylinders (double-acting) and 120-cm^3 swept volume per piston. The concept was to change only the heater for different applications: direct heating with natural gas or heat from biomass of waste energies, and remote heating via heat pipe for solar or chemical resources.

The two remaining problems were power control and sealing to the ambient. These were solved with the development of the variable swashplate and use of a pressure hull around the relatively small drive, with a rotating shaft seal or including the generator within the pressure hull. The angle position of the swashplate is obtained with a hydraulic vane motor and bevel gears or electrically by means of small planetary gears and bevel gears.

Figure 17 shows a cross section of the STM4-120, and Fig. 18 shows the actual engine with heat pipe heaters.

3.6 Free-Piston Engine (Sunpower, Inc.)

Figure 19 shows the typical features of the current free-piston Stirling engine. This engine delivers about 2.5 kW of electric

Table 1. Measured emission levels of Stirling engines without catalytic converters.

	Federal Clean Air Act of 1970 as measured on the Federal Driving Cycle (CVS-CH)	1998 Ultra Low Emissions Vehicle (ULEV) requirements	Measured emissions in 1975 of the 130-kW Ford-Philips Stirling engine	1998 projected Stirling engine based on recent measurements by Stirling Thermal Motors
	(g/mi.)	(g/mi.)	(g/mi.)	(g/mi.)
Carbon monoxide (CO)	3.40	1.70	1.20	0.20
Hydrocarbons (HC)	0.41		0.2	0.02
Oxides of nitrogen (NO_x)	0.40	0.20	0.14	0.10

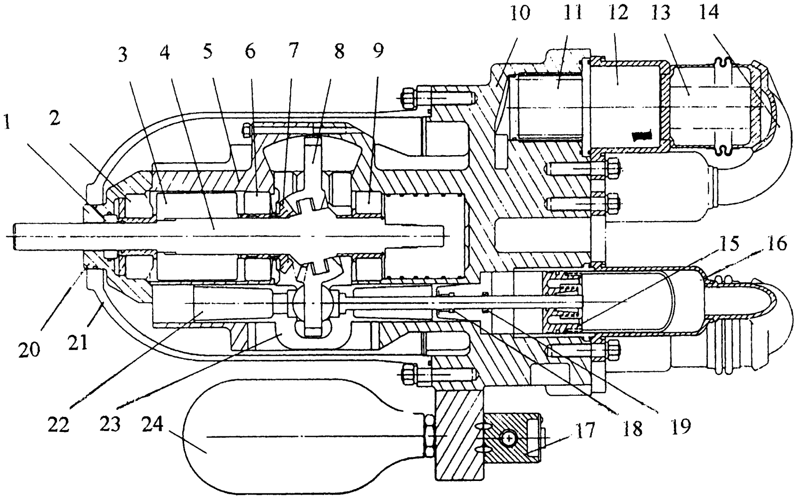

FIG. 17. Cross section of the double-acting, variable-displacement, swashplate drive engine (STM4-120). 1, Rotating shaft seal; 2, thrust bearing; 3, stroke converter; 4, main shaft; 5, rear crankcase; 6, rear main bearing; 7, bevel gear; 8, swashplate; 9, front main bearing; 10, front crankcase; 11, cooler; 12, regenerator; 13, heater; 14, hot connecting duct; 15, piston; 16, cylinder; 17, power control valve; 18, oil scraper; 19, cap seal; 20, seal housing; 21, pressure hull; 22, crosshead; 23, crosshead bridge; 24, hydraulic accumulator.

power at approximately 25% fuel-electric efficiency. It operates with helium at 4 MPa and 650 °C heater-wall temperature. The piston power is delivered directly to the magnets of a permanent-magnet alternator to produce alternating current. The engine operates at an approximately constant frequency regardless of load or amplitude.

4. APPLICATIONS

It is clear that the Stirling engine as a new prime mover can technically replace internal-combustion engines in many applications when it is proven reliable. At introduction into the market, price will be the biggest handicap. This creates the need for applica-

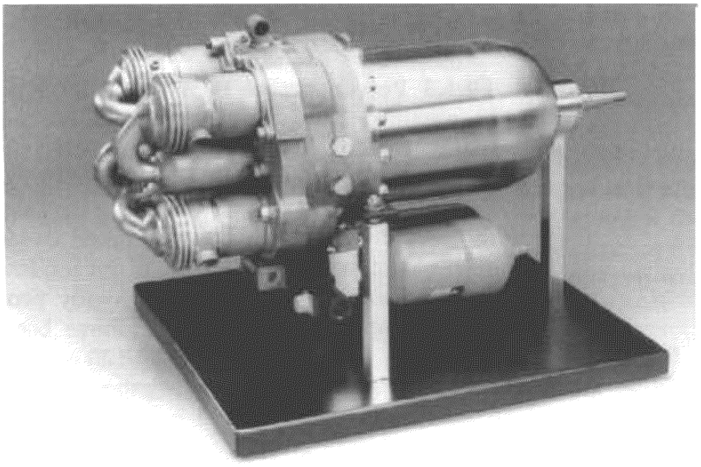

FIG. 18. The STM4-120 engine with heat-pipe heater heads.

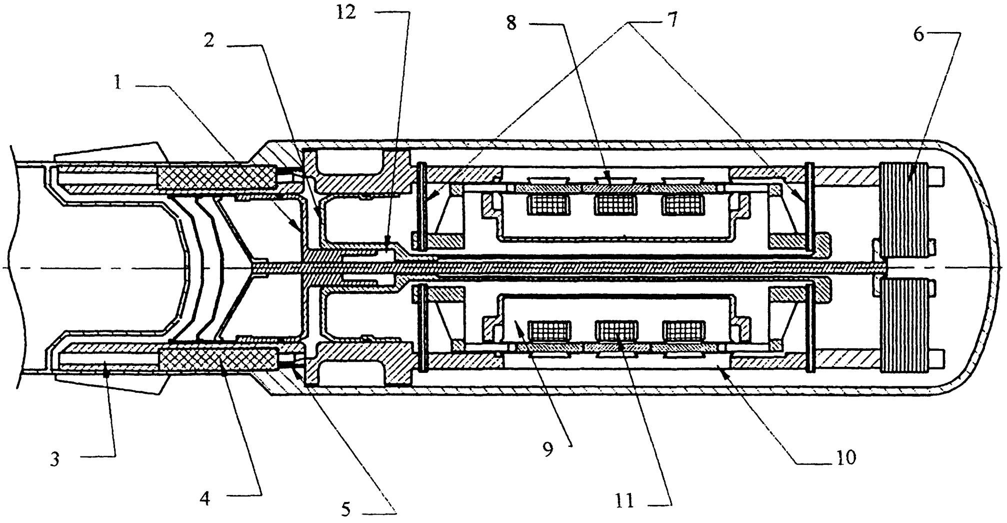

FIG. 19. Cross section of Sunpower's free-piston engine with a linear permanent-magnet alternator connected to the piston. 1, displacer; 2, piston; 3, heater; 4, regenerator; 5, cooler; 6, displacer springs; 7, magnetic-ring guide springs; 8, magnetic ring; 9, alternator inner iron; 10, alternator outer iron; 11, winding; 12, power control.

tion niches to be found where price is a smaller part of the buying decision. Volume production could overcome this handicap, and further markets could be penetrated.

4.1 Special Energy Sources

4.1.1 Solar-Energy Conversion. The use of solar energy is one of the most promising applications. The Stirling engine has already been tested for many hours in large parabolic collectors (80 m^2 or greater) with an energy conversion efficiency of more than 30%. The STM unit, complete with conventional radiators for cooling, is being tested at the proving grounds of Sandia National Laboratories, U.S.A. In this model, the receiver, that part of the engine containing the heater, is heated only by concentrated solar insolation. In order to be accepted by the utilities, the receiver had to be changed to a hybrid receiver that can also be heated by natural gas, making the unit independent of the sunshine.

4.1.2 Submarine Vessels. Since the Stirling engine has an external heating system, it is possible to use any suitable chemical reaction at sufficiently high temperature. A lot of research into this concept was performed by Philips and M.A.N. in the 1970s for the German government with the reaction of the metal Li and the oxidizer SF_6 (sulfur hexafluoride). The heat was transported by heat pipes with sodium as the working medium. This reaction produces no gaseous product, just salts, making this heating system very suitable for underwater applications.

Another successful nongaseous heating system uses stored heat from molten salt, such as LiF, also transported by sodium heat pipes. Finally, United Stirling was very successful in this market using liquid oxygen and diesel fuel.

4.1.3 Power from Biomass. Biomass gas obtained from an incompletely burned organic solid is difficult to use in internal-combustion engines but is quite suitable for Stirling engines with high efficiency because the gases do not needed to be cooled down and the composition of the gas is not too important. Biomass furnaces are already producing clean exhaust gases and have been successfully tested with an STM4-120 engine. Fast-growing wood in the neighborhood of small villages could be the solution for the need for remote electricity. This would not contribute to global warming.

4.2 Remote, Maintenance-Free Operation

4.2.1 Space Power. Photovoltaic cells are the common energy source in space using solar energy. However, sometimes there is need for higher energy density when an exotic heat source, such as that of a nuclear reaction, is available. Therefore, the U.S. Government, under supervision of NASA, has awarded many contracts to companies for the development of a Stirling generator set for space. One result is the 25-kW free-piston engine developed by Mechanical Technology, Incorporated.

4.3 Quiet, Clean, Efficient Power

4.3.1 Generator Sets. Calculations have shown that the cost of a modern Stirling engine, produced in the same numbers as the automotive gasoline engine, is not much different than that of the internal-combustion engine. This fact, together with the same or better reliability of a Stirling engine, makes it is easy to see the potential for a Stirling-engine generator set. In the case of cogeneration, Stirling devices have the advantage because the heat rejection in the cooling water is about twice as high as with the internal combustion engine.

4.3.2 Vehicle Propulsion. There have been a few trials to use the Stirling engine as direct propulsion for vehicles. This was never a success because the existing engines were too sophisticated and not reliable enough to meet the specifications of manufacturability and performance, although the Stirling engine did have a demonstrable superiority in some areas, such as lower emissions.

However, with the successful introduction of the variable swashplate, the situation has changed. Now the modern automotive, multiple-valve gasoline engine looks sophisticated compared to the four-cylinder variable swashplate engine. With the previous handicap eliminated, Stirling machine characteristics can directly compete with those of internal combustion engines as well as being demonstrated in areas unfulfilled by much of today's technology.

5. ECONOMIC ASPECTS

As mentioned, the cost of manufacturing a Stirling engine has prohibited the introduction of the technology into many areas where the advantages are great. Unfortunately, overcoming that obstacle has caused many companies and agencies to give up on the concept. With fewer than a dozen companies in the world maintaining the idea that Stirling devices are a product of tomorrow, many of the materials and subsystems needed to overcome this problem have been developed. In addition, new technologies have introduced new cost savings throughout the engine designs. With the decreased costs, it has allowed reliability, performance, and field tests to be completed in various arenas, letting the data speak for itself.

Finally, with the increased need for environmentally friendly sources of power, heat, and cold, the changing times have created a niche for Stirling engines. With the ability to change quickly from one energy source to another, Stirling has proven to be an old technology ready to face a new world.

Further Reading

Hargreaves, C. M. (1991), *The Philips Stirling Engine,* New York: Elsevier Science Publishing Company, Inc.

West, Colin D. (1986), *Principles and Applications of Stirling Engines,* New York: Van Nostrand Reinhold Company, Inc.

STOCHASTIC PROCESSES

Melvin Lax, *Physics Department, City College of the City University of New York, New York, New York, U.S.A., and Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey, U.S.A.*

Introduction 20
What Is a Stochastic Process? 20
Kinds of Dynamical Systems 20
An Approach to the Spectrum of Stationary Stochastic Processes 21
1. **The Treatment of Stationary Stochastic Processes** 22
1.1 Correlation Functions and the Regression Theorem 22
2. **Spectral Measurements and Correlations** 23
2.1 Introduction: An Approach to the Spectrum of Stochastic Processes 23
2.2 Standard Engineering Definition of Noise 23
2.3 The Wiener–Khinchin Theorem 25
2.4 Noise Measurements Using Filters 25
2.4.1 A Realizable Filter 26
3. **Thermal Noise** 26
3.1 Johnson Noise 26
3.2 Equipartition 28
3.3 Thermodynamic Derivation of Johnson Noise 28
3.4 Nyquist's Theorem 29
3.5 Relation between Nyquist and Einstein 30
4. **Shot Noise** 31
4.1 The Poisson Process 31
4.2 Pure Shot Noise 32
4.3 Generalized Characteristic Functions 33
4.4 Rice's Generalized Campbell Theorem 34
4.5 Campbell's Theorems 34
4.6 Equivalence of Shot Noise to the Poisson Process 35
4.7 Transit-Time Effects 35
4.8 Generation–Recombination Process 36
5. **Resistance Modulation Fluctuations** 37
5.1 Conductivity Fluctuations 37
5.2 A Thermodynamic Treatment of Total Carrier Fluctuations 37
5.3 Einstein Derivation of Carrier Fluctuations with Traps 38
5.4 The Spectrum of Resistance Modulation Fluctuations 40
6. **Concentration Fluctuations in Semiconductors** 41
6.1 General Theory of Concentration Fluctuations 41
6.1.1 Application to Semiconductors with Electrons, Holes, and Traps 42
6.2 Effect of Drift and Diffusion on Resistance Modulation Noise 42
7. **Langevin Processes** 44
7.1 Simplicity of Langevin Processes 44
7.2 Relation to the Fokker–Planck Equation 45
7.3 An Exactly Solvable Gaussian Example 46
7.4 Stochastic Integrals: The Ito–Stratonovich Controversy 47
8. **Further Contributions to Stochastic Processes** 50
8.1 Overview 50
8.2 Random-Walk Problems 51
8.3 Linear with Time-Dependent Decay 51
8.4 The Nonlinear (Fokker–Planck) Case: Reaction-Rate Theory 51
8.5 Stochastic Resonance 54
8.6 Self-Sustained Oscillators 54
Glossary 56
Works Cited 58
Further Reading 60

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

What Is a Stochastic Process?

If there are n people in a room and we ask what is the probability that any two have the same birthday, we have a pure probability problem since there is no change with time. Person j has birthday b_j, which can take any of 365 values, presumably with equal probability. There are n random variables, each of which has a discrete state space of 365 points. And with a modest effort, one can show that a coincidence of birthdays will acquire a probability more than 1/2 if $n > 25$.

If a man starts with a capital of \$100 and aims to end up with \$1000 by betting a dollar once a minute on coin tosses, we are dealing with a stochastic process because we have a random variable (the bettor's current capital) that changes (makes transitions) with time. The state space is discrete with the integral values from 0 to 1000, and the time variable is also discrete. This problem can be mapped onto a random-walk problem in which the man's capital is indicated by a position on a line, with absorbing barriers at each end.

The total number of electrons in a cell, in a semiconductor, increases by one in a generation process and decreases by one in a recombination processes. In this case, we have a stochastic process for a single, discrete random variable N as a function of a continuous time.

The current flowing through a resistor is a random process involving a continuous random variable $I(t)$ as a function of a continuous time.

The concept of a stochastic process is, of course, immediately extended to the case of several, or even an infinite number of, random variables. Each cell in a crystal could contain atomic-like states whose occupancy is a random variable, indexed by a discrete triple index. A carrier density, $n(\mathbf{r},t)$, provides a continuous index to an infinite number of continuous-time variables. In this case, carriers can be generated, recombined, or modified by drift and diffusion.

Kinds of Dynamical Systems

The simplest dynamical systems are linear. When the noise is weak, a stable nonlinear system will stray only slightly from a motion, possibly a time-independent state of the nonlinear state. Many important electronic and semiconductor devices are of this nature. We attempt to display some techniques that are universally applicable to fluctuations from a stable nonequilibrium state. We summarize briefly, here and in Sec. 8, problems that violate the requirements just mentioned that permit quasilinear solutions.

We are usually concerned with physical systems in which the noise is intrinsic to the system and related to the dissipative response of the system. There is also a vast literature on the response of nonlinear systems to external noise whose properties are unrelated to the system. See an early summary by Deutsch (1962), and the book by Van Kampen (1992) that provides an overview of stochastic processes with an emphasis on nonlinear processes.

Aside from the distinction between weak and strong noise, there is also a distinction between noise that is white (that is, possesses a flat spectrum) and colored noise whose spectrum is frequency dependent.

If the noise source is white, the problem, even if nonlinear, reduces to a Markoffian one. A Markoff process is a stochastic process whose future depends only on the latest information and is independent of prior information. This absence of memory is guaranteed by the fact that the Fourier transform of a noise constant in frequency is a delta function in time. The zero width of the delta function is the length of the memory time. If $a(t)$ is a random process, and t_j constitute a set of increasing times $t_0 < t_1 < t \cdots < t_n$, then the probability density of $a(t_n)$ conditional on all the values at all the earlier times obeys

$$P(a_n|a_{n-1},a_{n-2},\cdots a_0) = P(a_n|a_{n-1}) \qquad (1)$$

for a Markoffian process, like a person who remembers only the last thing he was told. Markoff processes then have the simplicity that a general multiple-time probability can be written

$$P(t_n,t_{n-1},\cdots t_0) = P(t_n|t_{n-1})P(t_{n-1}|t_{n-2}|) \cdots P(t_1|t_0)P(t_0) \qquad (2)$$

in terms of an initial probability multiplied by a series of two-time transition probabilities. The latter must obey a self-consistency condition known as the Chapman–Kolmogorov relation to insure the Markoffian nature of the process (see Sec. 7.2). See the articles PROBABILITY and RANDOM PROCESSES for more background on these subjects.

If a system is linear (or quasilinear, with a weak noise source), analytic solutions are possible for the probability densities even in the presence of colored noise sources. If the system is nonlinear, but with a white noise source, the probability will obey a generalized Fokker–Planck equation. The problem is thus reduced to (nontrivial) analysis. If the noise sources are also Gaussian, the problem reduces to a standard Fokker–Planck equation (see Sec. 7.2). Then analytic and numerical methods are feasible if the number of variables is small.

A system can be described as stationary if it is stable and none of the parameters describing the system is time dependent. In this case, there will be a steady (time independent) state, and all probability distributions that depend on variables at more than one time will be a function only of the time differences. We added the caveat about stability to exclude such cases as a free particle executing a Brownian motion that permits excursions that will grow indefinitely. Another example is that of an autonomous self-sustained oscillator. The latter has been shown by Lax (1968a, 1968b) always to contain one mode of instability that is usually related simply to an overall phase of the system that can grow at will. (The reason is given in Sec. 8.5.)

Quantum systems also require special handling. By the use of ordered operators, Lax (1968a,b) and Haken and Weidlich (1967) showed that quantum stochastic processes could be exactly expressed in terms of associated c-number ("classical") processes.

An Approach to the Spectrum of Stationary Stochastic Processes

Many experimental articles on semiconductor devices prior to 1960 included a systematic discussion of transport in these devices but an *ad hoc* theory of the associated noise. Lax (1960) recognized that a large class of devices that involved weak fluctuations from a (possibly nonequilibrium) steady state could be treated by a general approach that related noise to transport by an Einstein relation and a regression theorem, outlined as follows:

1. The response of most nonlinear systems to noise sources is sufficiently weak to permit a quasilinear approximation. Generalization of this procedure to self-sustained oscillators including quantum lasers is reviewed elsewhere (Lax, 1991).
2. The time decay of a correlation such as $\langle A(t)B(0)\rangle$ is the same as the time dependence of $\langle A(t)\rangle$, the dissipative relaxation of an initial deviation from the steady state. This is referred to as Onsager's regression hypothesis for deviation from an equilibrium state. Lax (1963, 1968b) proposed it as a theorem for Markoffian systems for fluctuations from a steady, possibly nonequilibrium, steady state. The angular brackets denote an average over an ensemble. In probability, ⟨something⟩ is often written E[something] and is called an expectation.
3. To the equations of time dependence determined from transport theory, we must add an initial condition, namely the value of $\langle A(0)B(0)\rangle$. For fluctuations from the equilibrium state, the second moments are known thermodynamically (Callen, 1985):

$$\langle A(0)B(0)\rangle \equiv \langle AB\rangle = -k\partial\langle A\rangle/\partial F_B. \qquad (3)$$

When regarded as a thermodynamic quantity, it is customary to rewrite $\langle A\rangle$ simply as A. Here, F_B is the force "conjugate" to the variable B in the sense that the negative of the pressure, $-P$, is conjugate to the volume V. (The negative sign is necessary since pressure decreases volume rather than increasing it.)

Sometimes, we must deal with quasiequilibrium states. For example, in a semiconductor, the electrons come to equilibrium with each other rapidly. So do the holes. But the radiative and nonradiative processes that produce recombination and generation of electron–hole pairs are slow. It is then customary to introduce quasi-Fermi levels, one for the electrons and one for the holes. In that case, Eq. (3) is useful, even though complete equilibrium is absent. In more complicated cases, one must make use of more detailed knowledge of the noise processes involved.

In Sec. 1 we discuss the above-mentioned approach to obtaining correlation functions for stochastic processes. In Sec. 2 we develop the relation between the Fourier transform of the correlation functions, the ideal noise measurement, and real noise measurements. In Sec. 3 we discuss the set of noise sources that arise from thermal origins. In Sec. 4 we discuss the "shot noise sources" that arise from the discrete nature of the charges that give rise to this noise. In Sec. 5 we discuss examples of these noise sources. In Sec. 6 we discuss noise in homogeneous semiconductors. In Sec. 7 we review the Langevin noise-source approach and describe its relation to the Fokker–Planck approach. Pitfalls in the use of the Ito definition of a stochastic integral are elucidated. In Sec. 8 we discuss a variety of nonlinear examples including random-walk theory, linear noise with time-dependent decay, reaction-rate theory, stochastic resonance, and self-sustained oscillators.

1. THE TREATMENT OF STATIONARY STOCHASTIC PROCESSES

1.1 Correlation Functions and the Regression Theorem

When equilibrium statistics cannot be utilized, the second moments can be obtained by using a generalization of the Einstein relation.

If $\mathbf{a}(t)$ is a set of random variables, $a_1(t)$, $a_2(t), \cdots$, then Lax (1960) has derived a generalized Einstein relation of the form

$$\frac{d}{dt}\langle\mathbf{aa}\rangle = 2\langle\mathbf{D}(\mathbf{a},t)\rangle + \langle\mathbf{aA}(\mathbf{a},t)\rangle + \langle\mathbf{A}(\mathbf{a},t)\mathbf{a}\rangle. \tag{4}$$

Here, **aa** is a dyadic abbreviation for a set of products a_ia_j, and **aA** is a corresponding abbreviation. The "drift vector," $\mathbf{A}(\mathbf{a},t)$, is defined by

$$\mathbf{A}(\mathbf{a},t) \equiv \lim_{\Delta t\to 0}\frac{\langle\mathbf{a}(t+\Delta t)-\mathbf{a}(t)\rangle}{\Delta t} \text{ so that}$$
$$\frac{d\langle\mathbf{a}(t)\rangle}{dt} = \langle\mathbf{A}(\mathbf{a},t)\rangle, \tag{5}$$

and the "diffusion matrix," $\mathbf{D}(\mathbf{a},t)$, is defined by

$$2!\mathbf{D}(\mathbf{a}) \equiv \lim_{\Delta t\to 0}\frac{\langle\Delta\mathbf{a}(t)\Delta\mathbf{a}(t)\rangle}{\Delta t}. \tag{6}$$

We understand the abbreviation

$$\Delta\mathbf{a}(t) = \mathbf{a}(t+\Delta t) - \mathbf{a}(t). \tag{7}$$

The 11 component

$$D_{11} = \langle[\Delta a_1(t)]^2\rangle/2\Delta t \tag{8}$$

reduces to the conventional diffusion constant when a_1 has the meaning of a position, x. If the diffusion coefficient D_{11}, labeled D_{xx} to remind us of its nature, is a constant, the mean square displacement $\langle[\Delta x]^2\rangle$ grows linearly with the time interval, the conventional Brownian-motion result (see RANDOM PROCESSES).

The evaluation of the averages in Eqs. (5) and (6) will make use of the nature of the noise process. For example, the system may be in interaction with a thermal reservoir. Or the noise may have shot-noise character, that is, it may arise because discrete particles are involved, and all the occupancies of states must be integers. We illustrate these mechanisms in Secs. 3 and 4.

The important point for the purposes of this section is that two major simplifications are possible. In the quasilinear case, there will be a set of associated stationary variables $\mathbf{a}_0$ and a set of small deviations

$$\boldsymbol{\alpha}(t) \equiv \mathbf{a}(\mathrm{t}) - \mathbf{a}_0. \tag{9}$$

The stationary points $\mathbf{a}_0$ are those at which the drift vectors vanish:

$$\mathbf{A}(\mathbf{a}_0) = 0. \tag{10}$$

Expanding around the steady operating point yields

$$\mathbf{A}(\mathbf{a}) \approx -\boldsymbol{\Lambda}\cdot\boldsymbol{\alpha},\ \langle\mathbf{A}(\mathbf{a})\rangle \approx -\boldsymbol{\Lambda}\cdot\langle\boldsymbol{\alpha}\rangle,$$
$$\frac{d\langle\boldsymbol{\alpha}(t)\rangle}{dt} = -\boldsymbol{\Lambda}\cdot\langle\boldsymbol{\alpha}(t)\rangle, \tag{11}$$

$$\mathbf{D}(\mathbf{a}) \approx \mathbf{D}(\mathbf{a}_0) \equiv \mathbf{D}. \tag{12}$$

In that case, the generalized Einstein relation, Eq. (4), simplifies to

$$\frac{d\langle\boldsymbol{\alpha\alpha}\rangle}{dt} = 2\mathbf{D} - \boldsymbol{\Lambda}\cdot\langle\boldsymbol{\alpha\alpha}\rangle - \langle\boldsymbol{\alpha\alpha}\rangle\cdot\boldsymbol{\Lambda}^{\dagger} \tag{13}$$

where $\mathbf{\Lambda}^\dagger$ is the transpose of $\mathbf{\Lambda}$. (In the complex case, the Hermitian adjoint would be needed, which is why we use the dagger symbol even though we are dealing with real variables here.)

The second major simplification is that we are dealing with fluctuations from a stationary state. In that case, a correlation such as $\langle \alpha_i(t)\alpha_j(t)\rangle$ is a function only of the time difference $t - t$. Hence, the time derivative vanishes, and we arrive at the stationary Einstein relation

$$2\mathbf{D} = \mathbf{\Lambda}\cdot\langle\boldsymbol{\alpha\alpha}\rangle + \langle\boldsymbol{\alpha\alpha}\rangle\cdot\mathbf{\Lambda}^\dagger. \tag{14}$$

For the case of a single variable, Eq. (14) reduces to the original Einstein relation between the diffusion constant and mobility; see Eq. (89).

In the presence of time reversal, the two terms on the right-hand side in Eq. (14) have been shown to be equal; see Onsager (1931) and Eq. (6.18) of Lax (1960). Thus the initial value $\langle\boldsymbol{\alpha}(0)\boldsymbol{\alpha}(0)\rangle$ is given by

$$\langle\boldsymbol{\alpha}(0)\boldsymbol{\alpha}(0)\rangle = \mathbf{\Lambda}^{-1}\cdot\mathbf{D}. \tag{15}$$

Equation (11) yields the conditional mean motion (transport)

$$\langle\boldsymbol{\alpha}(t)\rangle_{\boldsymbol{\alpha}(0)} = \exp(-\mathbf{\Lambda}t)\cdot\boldsymbol{\alpha}(0). \tag{16}$$

By calculating the autocorrelation in two steps—first an average conditional on the initial condition, and second an average over the initial condition,

$$\begin{aligned}\langle\boldsymbol{\alpha}(t)\boldsymbol{\alpha}(0)\rangle &= \langle\langle\boldsymbol{\alpha}(t)\rangle_{\boldsymbol{\alpha}(0)}\boldsymbol{\alpha}(0)\rangle\\ &= \exp(-\mathbf{\Lambda}t)\cdot\langle\boldsymbol{\alpha}(0)\boldsymbol{\alpha}(0)\rangle\end{aligned} \tag{17}$$

—we have established the regression theorem, that the fluctuation contains the same time dependence as the transport. The Markoffian assumption was tacitly made, since only in the Markoffian case is the conditional average independent of any earlier history that could affect the second average.

Equation (16) is restricted to positive times, and so necessarily is Eq. (17). However, stationarity permits us to write

$$\begin{aligned}\langle\boldsymbol{\alpha}(-t)\boldsymbol{\alpha}(0)\rangle &= \langle\boldsymbol{\alpha}(0)\boldsymbol{\alpha}(t)\rangle\\ &= \langle\boldsymbol{\alpha}(0)\boldsymbol{\alpha}(0)\rangle\cdot\exp(-\Lambda^\dagger t).\end{aligned} \tag{18}$$

2. SPECTRAL MEASUREMENTS AND CORRELATIONS

2.1 Introduction: An Approach to the Spectrum of Stochastic Processes

In this section we compare three definitions of noise: The standard engineering (SE) definition takes a Fourier transform over a finite time interval, squares it, divides by the time, and then takes the limit as the time approaches infinity. The second definition is the Fourier transform of the autocorrelation function. The equality between these two definitions is known as the Wiener (1930)–Khinchin (1934) theorem. The third procedure, introduced by Lax (1968a), passes the signal through a realizable filter of finite bandwidth, squares it, and averages over some large finite time. This simulates an actual measurement of the noise spectrum. As the bandwidth is allowed to approach zero, the result will (aside from a normalization factor) approach the ideal value of the two preceding definitions.

2.2 Standard Engineering Definition of Noise

The spectrum of noise $G_s(\alpha,\omega)$ in a single random variable $a(t)$ is a measure of the fluctuation energy in the frequency interval $[\omega, \omega + d\omega]$ associated with the fluctuating part $\alpha(t)$, where

$$\alpha(t) = a(t) - \langle a(t)\rangle;\ \langle\alpha(t)\rangle = 0. \tag{19}$$

The SE definition of noise, denoted by G_s, is chosen to obey the normalization

$$\int_0^\infty G_s(\alpha,\omega)df = \langle|\alpha(t)|^2\rangle, \tag{20}$$

because it is customary in engineering to emphasize positive frequencies $f = (\omega/2\pi) > 0$ only. For this reason, we adopt the definition

$$\begin{aligned}\tfrac{1}{2}G_s(\alpha,\omega) &\equiv \tfrac{1}{2}G_s(\omega) \equiv (\alpha^2)_\omega\\ &= \lim_{T\to\infty}\frac{1}{2T}\left\langle\left|\int_{-T}^{T}\alpha(t)e^{-j\omega t}dt\right|^2\right\rangle\end{aligned} \tag{21}$$

and verify the normalization later.

The letter j is used to denote imaginary unit customarily in electrical engineering. The SE convention is that $\exp(j\omega t)$ describes positive frequencies and $R + j\omega L + 1/j\omega C$ is the impedance of a series circuit of a resistance R, an inductance L, and a capacity C. Because propagating waves are described by $\exp(ikx - i\omega t)$ in physical problems, we regard $\exp(-i\omega t)$ as describing positive frequencies so that the physics convention is equivalent to setting $j = -i$ consistently.

In this definition, Eq. (21), the interval on t is truncated to the region $-T \le t \le T$, its Fourier transform is taken, and the result is squared. Since such a measurement would attempt to filter out one component ω and square it, this definition is reasonable. Since the integral is not expected to converge, it must be normalized by T. What is not clear yet is why one divides by T rather than T^2. The angular brackets denote an ensemble average. It is curious that both the limit $T \to \infty$ and an ensemble average are taken. For an ergodic process, a time average—by definition—is equal to an ensemble average. One might guess that the ensemble average could have been eliminated—that its use is primarily for convenience of computation. The above assumption is, however, wrong! Middleton (1960) shows that if the ensemble average is not performed, substantial fluctuations occur in the value of $G_s(\omega)$.

Presumably, this sensitivity occurs because we are asking for the noise at a precise frequency. Because of the Fourier relation between frequency and time, a measurement accurate to $\Delta\omega$ requires a time $t > 1/\Delta\omega$. Realistic noise measurements, to be discussed below, using filters of finite width, are presumably ergodic.

The above definition assumes that the noise is stationary. A more general definition of the noise at a frequency ω and time t is given by the Fourier transform

$$\tfrac{1}{2}G(\alpha,\omega,t) = \int_{-\infty}^{\infty} e^{j\omega\tau}R(\tau,t)d\tau \tag{22}$$

of the Wigner (1932)–Moyal (1949) type of autocorrelation function,

$$R(\tau,t) = \langle \alpha(t + \tau/2)^*\alpha(t - \tau/2)\rangle, \tag{23}$$

guaranteed to yield a real $G(\alpha,\omega,t)$. For future use, we note that the inverse of Eq. (22) is

$$\langle \alpha(v)^*\alpha(u)\rangle = \frac{1}{4\pi}\int_{-\infty}^{\infty} e^{-j\omega(v-u)} \times G(\alpha, \omega, (u + v)/2)d\omega. \tag{24}$$

The nonstationary case is discussed further in Lax (1968a). In the stationary case, to which we now restrict ourself in this section, the autocorrelation, Eq. (23), is invariant under a shift of time origin, hence to a change in t. Thus both $R(t,\tau)$ and $G(\alpha,\omega,t)$ are independent of t. Equation (22) can then be simplified to

$$G(\alpha,\omega) = 2\int_{-\infty}^{\infty} e^{j\omega t}\langle \alpha^*(t)\alpha(0)\rangle dt. \tag{25}$$

According to the SE definition, Eq. (21), the noise is manifestly real. Equation (25) can also be shown to be real by taking its complex conjugate and introducing $-t$ as a new variable of integration. This conclusion remains true even if α is a complex variable, or a non-Hermitian operator. In the classical case when time-reversal invariance holds, the ensemble average is an even function of time and

$$\begin{aligned} G(\alpha,\omega) &= 4\int_0^{\infty} \cos\omega t\langle \alpha^*(t)\alpha(0)\rangle dt \\ &= 4\mathrm{Re}\int_0^{\infty} \exp(j\omega t)\langle \alpha^*(t)\alpha(0)\rangle dt. \end{aligned} \tag{26}$$

It is easy to verify from Eq. (25) that the normalization integral

$$\int_{-\infty}^{\infty} G(\alpha,\omega)d\omega/4\pi = \langle |\alpha(0)|^2\rangle \tag{27}$$

is consistent with our original aim in Eq. (20).

If we have a classical variable V expressed as a linear combination of variables,

$$V(t) = \sum_m C_m\alpha_m(t), \tag{28}$$

then the noise in V is given by

$$G(V,f) = \sum_{mn} C_m^*C_nG_{mn}(\boldsymbol{\alpha},f), \tag{29}$$

where the noise matrix

$$G_{mn}(\boldsymbol{\alpha}) = 4\mathrm{Re}\int_0^\infty \exp(j\omega t)\langle\alpha_m^*(t)\alpha_n(0)\rangle dt. \tag{30}$$

If the variables $\boldsymbol{\alpha}$ have the decay matrix $\boldsymbol{\Lambda}$ of Eq. (11), the noise correlation G_{mn} is given by

$$G_{mn}(\boldsymbol{\alpha}) = 4\mathrm{Re}\sum_s [(-j\omega\mathbf{1} + \boldsymbol{\Lambda}^*)^{-1}]_{ms}\langle\alpha_s\alpha_n\rangle. \tag{31}$$

2.3 The Wiener–Khinchin Theorem

We shall prove the Wiener–Khinchin theorem by evaluating $G_s(\alpha,\omega)$ in terms of $G(\alpha,\omega)$:

$$G_s(\alpha,\omega) = \lim_{T\to\infty}\frac{1}{T} \times \left\langle\int_{-T}^{T}\alpha(v)^*\exp(j\omega v)dv\int_{-T}^{T}\alpha(u)\exp(-j\omega u)du\right\rangle. \tag{32}$$

In the stationary case (for which the theorem is valid), Eq. (24) can be written

$$\langle\alpha(v)^*\alpha(u)\rangle = \frac{1}{4\pi}\int_{-\infty}^{\infty} e^{j\omega(v-u)}G(\alpha,\omega)d\omega. \tag{33}$$

If Eq. (33) is inserted into Eq. (32), with ω replaced by ω', we have

$$\begin{aligned} G_s(\alpha,\omega) &= \lim_{T\to\infty}\frac{1}{T}\int_{-\infty}^{\infty} G(\alpha,\omega')d\omega' \times\frac{1}{4\pi}\left|\int_{-T}^{T}\exp[j(\omega-\omega')u]du\right|^2 \\ &= \lim_{T\to\infty}\int_{-\infty}^{\infty} G(\alpha,\omega')d\omega'\,\frac{\sin^2(\omega'-\omega)T}{\pi(\omega'-\omega)^2T} \\ &= G(\alpha,\omega). \end{aligned} \tag{34}$$

This completes our heuristic proof of the Wiener–Khinchin theorem. The last step moved the limiting procedure under the integral sign and used

$$\lim_{T\to\infty}\frac{\sin^2(\omega'-\omega)T}{\pi(\omega'-\omega)^2T} = \delta(\omega'-\omega). \tag{35}$$

The appropriateness of the limit, Eq. (35), is based on the facts that

1. the integral of the left-hand side, for any T, is 1;
2. the width of the function is of order $1/T$, and the maximum height, at $\omega' = \omega$, is of order T. This function becomes very tall and narrow. An integration against this function by any function $G(\omega')$ of bounded variation will be sensitive only to its value at the peak $\omega' = \omega$.

Note that Eq. (33) with $u = v = t$ leads to the normalization condition

$$\langle\alpha(t)^*\alpha(t)\rangle = \frac{1}{4\pi}\int_{-\infty}^{\infty} G(\alpha,\omega)d\omega. \tag{36}$$

This normalization is equivalent to the customary choice, Eq. (20), when $G(\alpha,\omega)$ is even in ω, but is generally valid even when it is not. It follows easily from time-reversal invariance that for classical variables evenness follows, but this is not true for quantum mechanical variables. (Our definitions apply to the quantum case if α^* is replaced in the quantum case by the Hermitian conjugate, $\alpha^\dagger$.)

2.4 Noise Measurements Using Filters

An actual measurement of noise at a frequency ω_0 passes the signal $\alpha(t)$ through a filter described in the time domain by

$$\alpha_{\text{out}}(t) = \int_{-\infty}^{t} K(t-t')\alpha(t')dt', \tag{37}$$

where $K(t)$ is known as the indicial response of the filter, or its response to a $\delta(t)$ input pulse. In order that the filter be realizable, hence causal, output can only appear after input, so that

$$K(t) = 0 \text{ for } t < 0. \tag{38}$$

The upper limit in Eq. (37) can thus be extended to infinity. In terms of Fourier components,

$$\alpha(\omega) \equiv \int_{-\infty}^{\infty}\exp(-j\omega t)\alpha(t)dt, \tag{39}$$

Eq. (37) yields the convolution-theorem result

$$\alpha_{\text{out}}(\omega) = k(\omega,\omega_0)\alpha(\omega), \tag{40}$$

where

$$k(\omega,\omega_0) \equiv \int_0^\infty \exp(-j\omega t)K(t)dt. \tag{41}$$

The ω_0 is arbitrarily introduced into our notation to be a reminder that we shall be dealing with a filter that emphasizes the frequency region near ω_0. Thus we expect the output spectrum to be $|k(\omega,\omega_0)|^2$ times the input spectrum:

$$(\alpha_{\text{out}}^2)_\omega = |k(\omega,\omega_0)|^2(\alpha^2)_\omega. \tag{42}$$

However, this argument is heuristic, since the integral for α_ω does not converge in the usual sense.

What is actually measured is

$$G_m(\omega_0) = \frac{1}{2T}\int_{t_0-T}^{t_0+T} |\alpha_{\text{out}}(t)|^2 dt, \tag{43}$$

the time average of the squared signal. For long enough T we expect ergodicity and can replace the time average by the ensemble average. Because of stationarity, this result simplifies to

$$\begin{aligned} G_m(\omega_0) &= \langle|\alpha_{\text{out}}(t)|^2\rangle = \int_{-\infty}^{t} K(t-t')^*dt' \\ &\quad \times \int_{-\infty}^{t} K(t-t'')dt''\langle\alpha(t')^*\alpha(t'')\rangle \\ &= \int_0^\infty\int_0^\infty K(u)^*K(v) \\ &\quad \times \langle\alpha(t-u)^*\alpha(t-v)\rangle dudv \\ &= \int_0^\infty K(u)^*du \int_0^\infty K(v)dv\langle\alpha(v)^*\alpha(u)\rangle. \end{aligned} \tag{44}$$

Order has been preserved in the above steps so that they remain valid for noncommuting operators. Using the Wiener–Khinchin theorem in reverse, Eq. (33), to eliminate the autocorrelation, we obtain

$$G_m(\omega_0) = \frac{1}{4\pi}\int_{-\infty}^{\infty} G(\alpha,\omega)d\omega|k(\omega,\omega_0)|^2. \tag{45}$$

The factor 4π arises because of the convention followed in Eq. (36).

2.4.1 A Realizable Filter. The simplest example of a realizable filter is to regard $\alpha(t)$ as a voltage placed across an R-L-C circuit, with the output $\alpha_{\text{out}}(t)$ obtained across the resistance. The differential equations describing this filter are

$$RI(t) = \frac{1}{C}\frac{dQ}{dt} + L\frac{dI}{dt} = \alpha(t), \tag{46}$$

$$\alpha_{\text{out}}(t) = I(t)R. \tag{47}$$

These equations result in the Fourier relation

$$\alpha_{\text{out}}(\omega) = k(\omega,\omega_0)\alpha(\omega) \tag{48}$$

with

$$k(\omega,\omega_0) = \frac{R}{R + j(\omega L - 1/\omega C)}, \tag{49}$$

where $\omega_0 = 1/\sqrt{LC}$ is the resonance frequency.

The measured spectrum $G_m(\omega_0)$ continues to be given by Eq. (45). In the limit when the $Q \equiv \omega_0 L/R$ of the oscillator becomes large, there are two sharp resonances at $\pm\omega_0$, and we can approximate

$$|k(\omega,\omega_0)|^2 = [\delta(\omega-\omega_0) + \delta(\omega+\omega_0)]\pi R/2L, \tag{50}$$

where the coefficient $\pi R/(2L)$ was chosen to yield the correct integral

$$\int_{-\infty}^{\infty} \frac{R^2 d\omega}{R^2 + (\omega L - 1/\omega C)^2} = \frac{\pi R}{L}. \tag{51}$$

This integral was evaluated using formula 031.10 in Gröbner and Hofreiter (1950). Equations (45) and (50) combine to yield

$$G_m(\omega_0) = RG(\alpha,\omega_0)/4L, \tag{52}$$

a result sensitive only to the measurement frequency with a known renormalization factor $R/(4L)$.

3. THERMAL NOISE

3.1 Johnson Noise

Johnson (1928) measured the voltage noise in a variety of materials, as a function of resistance. He found the results shown in Fig. 1, namely, that $\langle V^2\rangle$ is proportional to the resistance R, independent of the mate-

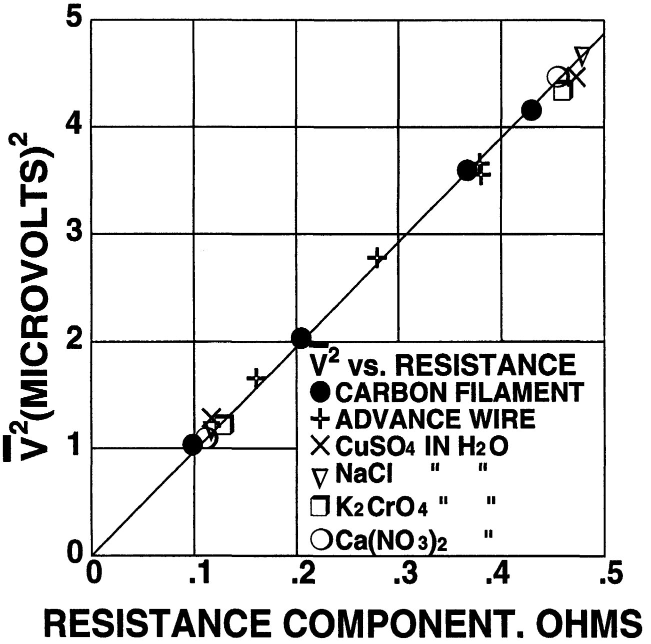

FIG. 1. The noise measured by Johnson (1928) versus resistance in six diverse materials.

rial. See also Kittel (1958). He also found that the measured noise power in the frequency interval df is proportional to the temperature of the resistor from which the noise emanates:

$$\langle v^2 \rangle = 4kdfRT = G(v,f)df. \tag{53}$$

To within his experimental accuracy, k is found to agree with the Boltzmann constant. The theoretical support for this choice is based on the equipartition theorem and Nyquist's theorem discussed in Secs. 3.2 and 3.4.

Moullin (1938) gives a review of the experimental measurements of noise. He considers two resistors in parallel at two different temperatures and concludes that the noise in a frequency interval df is given by

$$\begin{aligned}\langle v^2 \rangle &= 4kdf\frac{R_1R_2}{(R_1+R_2)^2}[T_1R_2 + T_2R_1] \\ &\equiv 4kdfR_eT_a, \end{aligned}\tag{54}$$

where k is Boltzmann's constant and R_e is an effective resistance that is compared with experimental data in Fig. 2 by Williams (1937). Williams chooses T_a to be T_1, except in the limiting case of $R_1 = \infty$, in which there is only one resistance, and he then takes T_a to be room temperature.

Moullin (1938, p. 31) generalizes this result to the case of an arbitrary number of impedances in parallel:

$$Z = \left[\sum_n (1/Z_n)\right]^{-1}, \tag{55}$$

$$\langle v^2 \rangle = 4kdf|Z|^2 \sum_n \frac{R_nT_n}{|Z_n|^2}\,. \tag{56}$$

If we remember that admittance Y_n is related to the impedance $Z_n \equiv R_n + jX_n$ by

$$Y_n = \frac{1}{Z_n} = \frac{R_n - jX_n}{R_n^2 + X_n^2} \equiv g_n - js_n \tag{57}$$

and write $V = ZI$, this result takes a simpler form in terms of current fluctuations,

$$\langle I^2 \rangle = 4kdf\sum_n g_nT_n. \tag{58}$$

These alternative expressions are not surprising in terms of Thevenin's and Norton's theorems. The terms g_n and s_n are referred to as conductance and susceptance. The result, Eq. (54) or (58), used equilibrium theory to

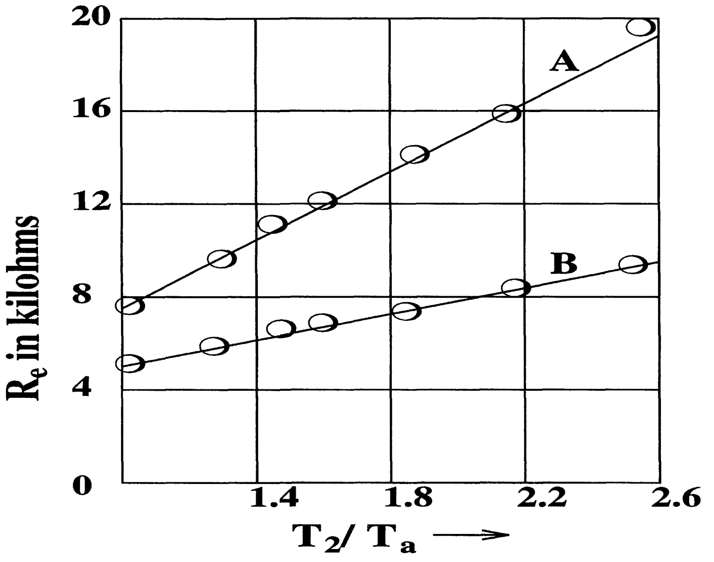

FIG. 2. Thermal noise for two resistors in parallel versus temperature obtained by Williams (1937) is plotted as an effective resistance defined by $R_e = \langle v^2\rangle/[4kdfT_a]$ against T_2/T_a. Williams takes T_a to be T_1, except in the one-resistor case, for which T_1 is infinite, and he then chooses T_a to be room temperature. Both theory, in Eq. (54), and experiment are linear functions of the abscissa. Line B is the two-resistance case, and line A is the one-resistance case.

obtain results when true equilibrium (all temperatures equal) is absent.

3.2 Equipartition

In this section we establish a fundamental compatibility between Johnson noise and thermodynamic equilibrium by demonstrating that this noise source generates an energy $kT/2$ in inductances and capacitances to which it is connected.

Consider a series circuit of resistor R, inductance L, and capacitor C, with a Johnson noise voltage v in the resistor. Then the resulting current and charge fluctuations i and q are

$$i = \frac{v}{R + j(\omega L - 1/\omega C)}; \quad q = i/j\omega. \tag{59}$$

The fluctuation energy in the inductance is, from Eq. (53),

$$\begin{aligned}\tfrac{1}{2}L\langle i^2\rangle &= \tfrac{1}{2}L\int \frac{G(v,f)df}{R^2 + (\omega L - 1/\omega C)^2} \\ &= 2kTRL\int_0^\infty \frac{df}{R^2 + (\omega L - 1/\omega C)^2} = \tfrac{1}{2}kT.\end{aligned} \tag{60}$$

Similarly, the energy stored on the capacitance is

$$\begin{aligned}\tfrac{1}{2}\frac{1}{C}\langle q^2\rangle &= \frac{2RkT}{C}\int_0^\infty \frac{df}{\omega^2[R^2 + (\omega L - 1/\omega C)^2]} \\ &= \tfrac{1}{2}kT.\end{aligned} \tag{61}$$

Thus, for both the inductance and the capacitor, the energy stored because of the noise in the resistor is precisely that expected by the equipartition theorem ($kT/2$ for each degree of freedom quadratic in the coordinate or the velocity.) Note that Eq. (59) is a relation between Fourier components, so that, for example, i and q should be written i_ω and q_ω, whereas in Eqs. (60) and (61), we are really dealing with the time-dependent quantities $\langle i(t)^2\rangle$ and $\langle q(t)^2\rangle$, respectively.

A fundamental truth now emerges. Fluctuations must be associated with dissipation in order that $\langle q^2\rangle$ does not decay to zero, but maintains the appropriate thermal equilibrium energy.

3.3 Thermodynamic Derivation of Johnson Noise

In view of the compatibility with thermal equilibrium shown in the preceding section, it is not surprising that a simple thermodynamic argument can be used to demonstrate that the noise emanating from a resistor must be proportional to its resistance.

Consider two resistors in a series circuit shown in Fig. 3. The current i_1 through resis-

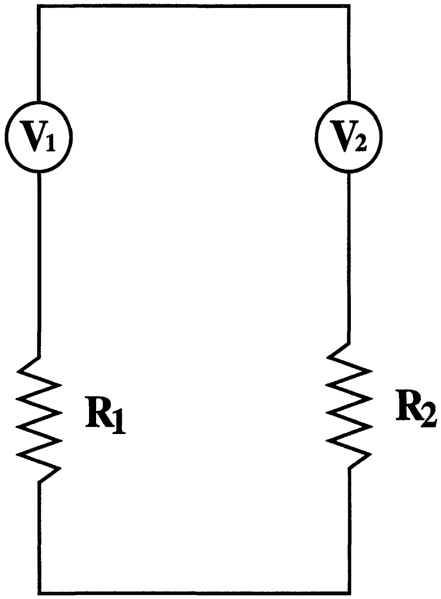

FIG. 3. The power transfer between two resistances in parallel, where v_j is the Johnson noise voltage in resistor j.

tor 2 produced by the noise voltage v_1 in the first resistor is

$$i_1 = v_1/(R_1 + R_2). \tag{62}$$

The power from v_1 into R_2 is given by

$$P_{2\leftarrow 1} = \left(\frac{v_1}{R_1 + R_2}\right)^2 R_2 = i_1^2 R_2. \tag{63}$$

Conversely, the power from 2 into 1 is given by

$$P_{1\leftarrow 2} = i_2^2 R_1 = \frac{v_2^2 R_1}{(R_1 + R_2)^2}\,. \tag{64}$$

If both resistors are at the same temperature, the second law of thermodynamics requires that there can be no steady net flow (in either direction). Equating Eqs. (63) and (64), we obtain

$$\frac{v_1^2}{R_1} = \frac{v_2^2}{R_2} \text{ or } \frac{G(v_1,f)}{R_1} = \frac{G(v_2,f)}{R_2} = W(f), \tag{65}$$

where $W(f)$ is a possibly frequency-dependent factor that is independent of the various resistances. In summary, the noise spectrum associated with an arbitrary resistance R is given by

$$G(v,f) = W(f)R. \tag{66}$$

If we connect a resistor R_1 to an impedance $R(f) + jX(f)$, the same equality requires

$$R(f)\left|\frac{v_1}{R_1 + R(f) + jX(f)}\right|^2 = \left|\frac{v_2}{R_1 + R(f) + jX(f)}\right|^2 R_1. \tag{67}$$

Thus we can conclude that the noise $G(v,f)$ associated with impedance $Z(f)$ is given by

$$G(v,f) = W(f)R(f) = W(f)\mathrm{Re}Z(f). \tag{68}$$

Thus the noise is proportional to $R(f) = \mathrm{Re}Z(f)$ even when the impedance $Z(f)$ is frequency dependent.

3.4 Nyquist's Theorem

A more complete derivation of the fluctuation–dissipation relation including a determination of the universal coefficient $W(f)$ is provided by Nyquist (1927, 1928). Nyquist's procedure is to calculate the power dissipated in a load connected to the end of a transmission line in two different ways and compare the results. Equation (63) for the power from resistor R_1 into R_2 in the frequency interval df reduces with Eq. (66) to

$$P_{2\leftarrow 1} = \frac{W(f)R_1R_2}{(R_1 + R_2)^2}\,df. \tag{69}$$

The maximum power is transferred when the impedance is matched, $R_1 = R_2$:

$$P_{2\leftarrow 1} = \frac{1}{4}W(f)\Delta f. \tag{70}$$

A transmission line can be terminated with its "characteristic impedance"

$$R_0 = \sqrt{L_1/C}, \tag{71}$$

where L_1 is the inductance per unit length of the line and C is its shunt capacitance per unit length. In this case, waves down the line are not reflected. The line acts as if it were infinite. Nyquist therefore chooses as his proof vehicle a transmission line terminated by R_0 at both ends. The line is assumed to have length L. The transmission line can be described in terms of harmonic-oscillator modes. If U is the energy density per mode, then the energy per mode is

$$UL = kT, \tag{72}$$

where we have made use of the equipartition theorem.

If the modes are described as plane waves $\exp(\pm ikx)$, then k takes the discrete values

$$k = 2\pi n/L, \quad \text{whence } \Delta k = 2\pi/L \tag{73}$$

is the mode spacing. With $\omega \equiv 2\pi f$ related to k by the group velocity, $v = \partial\omega/\partial k$, the number of modes propagating to the right in a given frequency interval is

$$\begin{aligned}\text{No. of modes} &= \frac{\text{modes}}{\delta k}\cdot\frac{\delta k}{\delta f}\Delta f\\ &= \frac{L}{2\pi}\frac{2\pi}{v}\Delta f = \frac{L}{v}\Delta f.\end{aligned} \tag{74}$$

Since each mode carries an energy U with the group velocity v, the power transmission down the line is

$$\begin{aligned}\text{flux} &= (\text{flux/mode})(\text{No. of modes})\\ &= (Uv)(L/v)\Delta f = UL\Delta f.\end{aligned} \tag{75}$$

Comparison with Eq. (70) yields

$$W(f) = 4UL. \tag{76}$$

In the limit of classical physics, Eq. (72) applies, and

$$W(f) = 4kT. \tag{77}$$

This result yields Nyquist's theorem

$$G(v,f)\Delta f = 4kTR\Delta f, \tag{78}$$

a result in agreement with Johnson's experimental results.

The beautifully simple Nyquist proof yields a result independent of frequency because all the harmonic-oscillator traveling modes have the same energy kT. The normalization must, of course, agree with that found in order to obtain the agreement with equipartition found in Sec. 3.2.

An apparent problem with Nyquist–Johnson noise is that the total voltage fluctuation

$$\langle v^2\rangle = \int_0^\infty G(v,f)df \tag{79}$$

diverges. Nyquist suggested that this problem could be removed if the classical energy, kT, associated with a harmonic oscillator were replaced by the quantum energy,

$$kT \rightarrow \hbar\omega\bar{n}(\omega) = \hbar\omega[\exp(\hbar\omega/kT) - 1]^{-1}, \tag{80}$$

which approaches kT at low frequencies and vanishes experimentally at high frequencies.

Of course, the actual energy associated with a harmonic oscillator,

$$E = \hbar\omega[\bar{n}(\omega) + \tfrac{1}{2}], \tag{81}$$

includes the zero-point energy. If the latter is retained, the divergence in the integrated energy reappears.

It is appropriate to argue, however, that the zero-point energy, while real, can never be absorbed and hence should be omitted in the Nyquist proof. This point, however, requires further discussion, as it relates to fundamental considerations of the effects of zero-point motion of the electromagnetic field in the area of quantum optics.

Callen and Welton (1951) considered a general class of systems (quantum mechanically) and established that Eq. (78), and its dual form

$$G(I,f)\Delta f = 4kTG\Delta f = 4kT\,\mathrm{Re}\,Y(\omega)\Delta f, \tag{82}$$

apply to all systems near equilibrium with the replacement of kT by Eq. (80) or (81) when necessary. The importance of the Callen–Welton work is the great generality of potential applications. The fact that all dissipative systems have corresponding noises associated with them is necessary in order that the second law of thermodynamics not be violated when such systems are connected.

3.5 Relation between Nyquist and Einstein

Consider a mechanical system with velocity v. Then the noise, Eq. (21), associated with v can be written

$$G(v,f) = \lim_{T\to\infty} \frac{2}{T}\left\langle \left| \int_{-T/2}^{T/2} v(t)e^{-j\omega t}dt \right|^2 \right\rangle. \tag{83}$$

The zero-frequency noise is

$$G(v,0) = \frac{2}{T}\langle[\Delta x]^2\rangle, \text{ where}$$
$$\Delta x = x(T/2) - x(-T/2). \tag{84}$$

But the usual diffusion constant, D, is defined by $\langle[\Delta x]^2\rangle = 2DT$, where T is the total time traveled. Thus the zero-frequency velocity noise is directly determined by the diffusion constant,

$$G(v,0) = 4D. \tag{85}$$

Conversely, the fluctuation–dissipation theorem, Eq. (82), for the velocity (which is analogous to a current rather than a voltage, since the current is proportional to ev) is given by

$$G(v,f) = 4kT \operatorname{Re} Y(\omega), \tag{86}$$

where

$$Y(\omega) = v/F \tag{87}$$

is the admittance, or velocity per unit applied force. At zero frequency, we refer to v/F as the mechanical mobility B and v/E as the (electrical) mobility μ, so that

$$Y(0) = B = \mu/e. \tag{88}$$

The fluctuation–dissipation theorem at zero frequency now reads

$$D = kTB = (kT/e)\mu, \tag{89}$$

which is simply the Einstein (1905) relation between diffusion and mobility. For a verification of this relation for electrons and holes in semiconductors, see Transistor Teachers' Summer School (1953).

By relating the mechanical mobility to the viscosity determined by Stokes's law, Einstein was able to make a macroscopic determination of Avogadro's number in good agreement with current values. Although Einstein indicates that he had not seen the article by Brown (1828) describing microscopic observations on tiny pollen grains in water, in his second article Einstein (1906) refers to the work of Gouy (1888) with the quote that the "so-called Brownian motion is caused by the irregular thermal movements of the molecules of the liquid."

4. SHOT NOISE

4.1 The Poisson Process

There are two physical problems describable by the same random process. The first process is the radioactive decay of a collection of nuclei. The second is the production of photoelectrons by a *steady beam* of light on a photodetector. In both cases, we can let a discrete, positive, integer-valued variable $n(t)$ represent the number of counts emitted in the time interval between 0 and t. In both cases there is a constant probability per unit time ν such that νdt is the expected number of counts in $[t, t + dt]$ for small dt. We use the initial condition

$$n(0) = 0. \tag{90}$$

Then $n(t)$ will be the number of counts in the interval $[0,t]$. When we talk of $P(n,t)$ we can understand this to mean, $P(n, t|n = 0, 0)$, the conditional density distribution. Since the state $n(t) = n$ is supplied by transitions from the state $n - 1$ with production of photoelectrons at a rate νdt and is diminished by transitions from state n to $n + 1$, we have the rate equation

$$\frac{\partial P(n,t)}{\partial t} = \nu[P(n-1,t) - P(n,t)] \quad \text{for } n > 0. \tag{91}$$

In the first term, n increases from $n - 1$ to n, and in the second, from n to $n + 1$. Since $n \geq 0$, we have no supply from the state $P(-1,t)$ so that

$$\frac{\partial P(0,t)}{\partial t} = -\nu P(0,t), \tag{92}$$

whose solution is

$$P(0,t) = P(0,0)\exp(-\nu t) = \exp(-\nu t) \tag{93}$$

since $P(n,0) = \delta_{n,0}$ at time $t = 0$, corresponding to the certainty that there are 0 particles at time $t = 0$.

The form, Eq. (93), of this solution suggests the transformation

$$P(n,t) = \exp(-\nu t)Q(n,t) \tag{94}$$

with the resultant equation

$$\frac{\partial Q(n,t)}{\partial t} = \nu Q(n-1,t) \tag{95}$$

subject to the initial condition

$$P(n,0) = Q(n,0) = \delta_{n,0}. \tag{96}$$

Thus any $Q(n,t)$ may be readily obtained if $Q(n-1)$ is known. But n, as described by Eq. (91), can only increase. The result is the closed-form solution

$$P(n,t) = P(n,t|0,0) = \frac{(\nu t)^n}{n!}e^{-\nu t} \tag{97}$$

for $n \geq 0$ with a vanishing result for $n < 0$.

Distribution functions, such as Eq. (97), are also characterized by their moments μ_j', and their central moments μ_j, defined by

$$\mu_j' = \langle n^j\rangle;\ \mu_j = \langle (n - \langle n\rangle)^j\rangle. \tag{98}$$

A third set of moments, introduced by statisticians because they indicate clearly the deviation of a random process from that of a Gaussian, are called Thiele (1903) semi-invariants or cumulants by Kendall and Stuart (1969). The cumulants or linked moments $\kappa_j \equiv \langle n^j\rangle_L$ are defined in terms of the characteristic function (see PROBABILITY),

$$\begin{aligned}\langle \exp(ikn)\rangle &\equiv \sum_n \exp(ikn)P(n,t)\\ &= \exp\left[\sum_{j=1}^{\infty}\frac{(ik)^j}{j!}\kappa_j\right]\\ &= \exp\langle \exp(ikn) - 1\rangle_L.\end{aligned} \tag{99}$$

By taking the logarithm of this equation, expanding in powers of k, and comparing coefficients, one can read off the cumulants. The first four are particularly simple:

$$\kappa_1 = \langle n\rangle;\ \kappa_2 = \langle n^2\rangle_L = \langle (\Delta n)^2\rangle; \tag{100}$$

$$\kappa_3 = \langle n^3\rangle_L = \langle (\Delta n)^3\rangle; \tag{101}$$

$$\langle n^4\rangle_L = \langle (\Delta n)^4\rangle - 3\langle (\Delta n)^2\rangle^2; \tag{102}$$

where $\Delta n \equiv n - \langle n\rangle$ is the deviation from the mean. Cumulants beyond the second describe deviations from a Gaussian distribution.

The Poisson process is stationary. But no limit exists as $t \to \infty$, so that there is no time-independent $P(n)$. We can, however, evaluate the characteristic function of the conditional probability density:

$$\begin{aligned}\langle \exp(ikn)\rangle &\equiv \phi(k, t|n = 0, t = 0)\\ &= \sum_n e^{ikn}\frac{(\nu t)^n}{n!}e^{-\nu t}\\ &= \exp[\nu t(e^{ik} - 1)].\end{aligned} \tag{103}$$

Comparison of Eq. (103) for the Poisson process with Eq. (99) shows that for this process all cumulants have the same value,

$$\kappa_j = \nu t. \tag{104}$$

4.2 Pure Shot Noise

The Poisson process just described is, in fact, the simplest example of shot noise. We can picture the actual number $n(t)$ as a staircase function that is flat except at a set of times, t_j, at which a jump of unity occurs. Mathematically, this is describable as the solution of the stochastic differential equation

$$\frac{dn}{dt} = \sum_j \delta(t - t_j) \equiv \hat{\nu}(t). \tag{105}$$

We shall first show that the noise source, $\hat{\nu}(t)$, in Eq. (105) is indeed white noise and demonstrate its relation to conventional shot noise.

The average pulse rate over the large time interval $[0,T]$ is

$$\nu(t) = \langle \hat{\nu}(t)\rangle = \frac{1}{T}\sum_j \int_0^T \delta(t - t_j)dt_j = \frac{N}{T} = \nu, \tag{106}$$

where N is the number of pulses in the interval $0 < t_j \leq T$. We call the ratio ν to conform with the Poisson process in Sec. 3.1 and assume that this ratio is independent of t, and of the location of the time interval of length T.

The correlation in the fluctuation $\alpha(t) = \hat{\nu}(t) - \nu$ can be calculated in a similar way:

$$\langle \alpha(t)\alpha(t')\rangle = \langle \hat{\nu}(t)\hat{\nu}(t')\rangle - \langle \hat{\nu}(t)\rangle\langle \hat{\nu}(t')\rangle. \tag{107}$$

The first term involves a double sum

$$\langle \hat{\nu}(t)\hat{\nu}(t')\rangle = \sum_{i,j} \langle \delta(t - t_i)\delta(t' - t_j)\rangle, \tag{108}$$

but the diagonal ($j = i$) term involves the simpler average

$$\frac{1}{T}\sum_i \int_0^T \delta(t - t_i)\delta(t' - t_i)dt_i = \frac{N}{T}\delta(t - t'). \tag{109}$$

The terms with $j \neq i$ involve an uncorrelated average, which cancels against the last term in Eq. (107) in the limit as T goes to infinity, with the result

$$\langle \Delta\hat{\nu}(t)\Delta\hat{\nu}(t')\rangle = \nu\delta(t - t'). \tag{110}$$

The noise, by the Wiener–Khinchin formula, is

$$G(\nu,\omega) = 2\int_{-\infty}^{\infty} \langle \Delta\hat{\nu}(t)\Delta\hat{\nu}(t')\rangle dt = 2\nu, \tag{111}$$

a constant, or white noise.

If the pulses carry a charge e, the current is

$$\hat{I}(t) = e\hat{\nu}(t), \tag{112}$$

and we arrive at the traditional shot-noise formula

$$G(I,\omega) = 2e^2\nu = 2eI, \tag{113}$$

where $I = \langle \hat{I}\rangle$ is the average current.

4.3 Generalized Characteristic Functions

We must establish the equivalence of this pulse description in Eq. (105) with the Poisson distribution characterized by Eq. (91), by calculating the characteristic function by a direct method, and demonstrate its equality with that of the Poisson process given in Eq. (103). The characteristic function is defined by

$$\phi(k,t) = \langle \exp[ikn(t)]\rangle = \left\langle \exp\left[ik\int_0^t \hat{\nu}(s)ds\right]\right\rangle. \tag{114}$$

To save duplication later, we consider an appreciably more general problem. In this problem, the sequence of pulses that arrive can have a localized shape $f(t - t_j)$ that need not be a delta function. In the original problem considered by Campbell (1909), the discontinuity involved the charge on the electron, and so a factor q was added. The model was further generalized by Rice (1944, 1945, 1948) to permit the jumps to contain a random factor η. Thus Rice considered the process

$$\Theta(t) = q\sum \eta_j f(t - t_j), \tag{115}$$

where the η_j's are random jumps with a distribution independent of t_j, of t, and of j.

Our procedure for dealing with the same problem consists in writing

$$\Theta(t) = \int f(t - s)S(s)ds, \tag{116}$$

where the shot-noise function $S(s)$ is now given by

$$S(s) = q\sum_j \eta_j\delta(s - t_j). \tag{117}$$

Equation (116) describes $\Theta(t)$ as filtered shot noise. We can always remove the filtering to obtain results appropriate to pure shot noise.

We can now relate the ordinary characteristic function of Θ to the generalized characterized function of the shot-noise function, $S(s)$,

$$\langle \exp ik\Theta(t)\rangle = \left\langle \exp i\int_{-\infty}^{\infty} y(s)S(s)ds\right\rangle \tag{118}$$

by setting

$$y(s) \equiv kqf(t - s). \tag{119}$$

The average in Eq. (118), for general $y(s)$, was evaluated in two ways in Lax (1966b). The first made explicit use of Langevin techniques. The second, which follows Rice, is presented here. It makes use of the fact that the average can be factored:

$$\left\langle \exp i\int_{-\infty}^{\infty} y(s)S(s)ds\right\rangle = \prod_{j=1}^{N} \langle \exp[i\eta_j y(t_j)]\rangle. \tag{120}$$

Here, we have supposed that N pulses are distributed uniformly over a time interval T at the rate $\nu = N/T$. All N factors are independent of each other and have equal averages, so that after adding and subtracting 1 inside the brackets, the result is

$$\begin{aligned}\prod_{j=1}^{N}\int g(\eta_j)d\eta_j\left[1+\frac{1}{T}\int_{-T/2}^{T/2}\{\exp[i\eta_j y(t_j)]-1\}dt_j\right] \\ =\left[1+\frac{1}{T}\int g(\eta)d\eta\int_{-T/2}^{T/2}\{\exp[i\eta y(s)]-1\}ds\right]^N \\ =\exp\left[\nu\int g(\eta)d\eta\int_{-\infty}^{\infty}\{\exp[i\eta y(s)]-1\}ds\right],\end{aligned} \tag{121}$$

where $g(\eta)$ is the normalized probability density for the random variable η. In the last step, we assumed that the integral over s converges and replaced it by its limit before taking a final limit in which N and T approach infinity simultaneously with the fixed ratio $N/T = \nu$.

4.4 Rice's Generalized Campbell Theorem

Campbell's two theorems are the first and second moments, $\langle\Theta\rangle$ and $\langle(\Delta\Theta)^2\rangle$, evaluated explicitly in Eqs. (126) and (127) below. Rice's generalized Campbell's theorem is obtained by setting $y(s) = kqf(t - s)$ in Eq. (118). The result is not just the first two moments, but the complete characteristic function

$$\begin{aligned}\langle\exp ik\Theta(t)\rangle &= \left\langle\exp\left[ikq\sum_j \eta_j f(t-t_j)\right]\right\rangle \\ &= \exp\left(\nu\int g(\eta)d\eta\int_{-\infty}^{\infty}\{\exp[ikq\eta f(t-s)]-1\}ds\right).\end{aligned} \tag{122}$$

Let us now specialize to the case of the original Campbell process by setting $g(\eta) = \delta(\eta - 1)$:

$$\begin{aligned}&\langle\exp ik\Theta(t)\rangle \\ &= \exp\left[\nu\int_{-\infty}^{\infty}\{\exp[ikqf(t-s)]-1\}ds\right].\end{aligned} \tag{123}$$

The probability density of the variable Θ of Eq. (115) may then be obtained by taking the inverse Fourier transform of the characteristic function in Eq. (122):

$$\begin{aligned}P(\Theta) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-ik\Theta)dk \\ &\times \exp\left[\nu\int g(\eta)d\eta\int_{-\infty}^{\infty}\{\exp[ikq\eta f(s)]-1\}ds\right].\end{aligned} \tag{124}$$

This form of generalized Campbell's theorem, like its antecedents, assumes that the t_j are randomly (and on the average uniformly) distributed in time. Moreover, there is assumed to be no correlation between successive pulse times. Lax and Phillips (1958) have found it convenient to exploit Eq. (124) in studying one-dimensional impurity bands.

The cumulants of the Rice process can be obtained from Eq. (122) as the coefficients of $k^r/r!$ in the exponent:

$$\begin{aligned}\kappa_r &= \left\langle\left[q\sum_j \eta_j f(t-t_j)\right]^r\right\rangle \\ &= \nu q^r\int g(\eta)\eta^r d\eta\int_{-\infty}^{\infty} f(s)^r ds \\ &= \nu q^r\langle\eta^r\rangle\int_{-\infty}^{\infty} f(s)^r ds.\end{aligned} \tag{125}$$

4.5 Campbell's Theorems

The choice $g(\eta) = \delta(\eta - 1)$ reduces the Rice process to the original Campbell process. Campbell's theorem itself includes only the cases $r = 1,2$. These take the form

$$\langle\Theta\rangle = \nu q\int_{-\infty}^{\infty} f(s)ds, \tag{126}$$

$$\langle[\Delta\Theta]^2\rangle = \nu q^2\int_{-\infty}^{\infty} f(s)^2 ds. \tag{127}$$

We close this section with an application of Campbell's theorem to the RC circuit shown in Fig. 4. A charge e from a vacuum tube generates a voltage pulse

$$ef(t) = (e/C)\exp(-t/RC) \tag{128}$$

across the resistor R. Campbell's theorems then yield

$$\langle V\rangle = \nu(e/C)\int_0^{\infty} e^{-t/RC}dt = \nu eR = IR \tag{129}$$

and

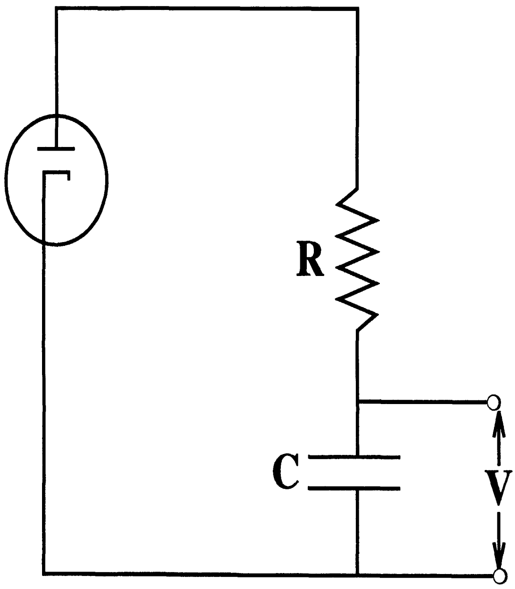

FIG. 4. The voltage fluctuation induced by shot noise into an *RC* circuit.

$$\langle(\Delta V)^2\rangle = \nu\frac{e^2}{C^2}\int_0^\infty e^{-2t/(RC)}dt = \frac{IeR}{2C}. \tag{130}$$

The electronic charge can be determined by comparison of $\langle(\Delta V)^2\rangle$ with $\langle V\rangle$:

$$e = 2C\langle(\Delta V)^2\rangle/\langle V\rangle. \tag{131}$$

4.6 Equivalence of Shot Noise to the Poisson Process

If we further specialize Eq. (123) by setting

$$qf(t - s) = 1 \quad \text{for } 0 \le s \le t \tag{132}$$

and zero elsewhere, then Eq. (116) reduces to

$$\Theta(t) = \int_0^t \hat{\nu}(s)ds = n(t). \tag{133}$$

The characteristic function, Eq. (132), of Θ is then identical to that of n and is given by

$$\langle \exp ikn(t)\rangle = \exp\left[\nu\int_0^t [\exp(ik) - 1]ds\right], \tag{134}$$

which reduces immediately to Eq. (103) as desired. This completes our proof that the shot-noise process of Eq. (105) is equivalent to the Poisson process of Eq. (91).

4.7 Transit-Time Effects

A set of carriers, j, leave a source (e.g., a cathode) at a set of times t_j and move to a destination (e.g., a plate) with a velocity $v(t - t_j)$. All carriers travel the fixed distance L and take the same travel time T, such that

$$\int_0^T v(t)dt = L. \tag{135}$$

The current in the external circuit will be

$$I(t) = \frac{e}{L}\sum_j v(t - t_j). \tag{136}$$

This intuitive result can be justified by a detailed calculation using Maxwell's equations.

The current can then be written as a convolution

$$I(t) = \frac{1}{L}\int_{-\infty}^{\infty} v(t - s)S(s)ds, \tag{137}$$

where the shot-noise function is

$$S(s) = e\hat{\nu}(s) = e\sum_j \delta(s - t_j) \tag{138}$$

as in Eq. (112) or (117) with $\eta = 1$. This convolution translates into a product of Fourier transforms that are squared to get the desired noise result

$$G(I,\omega) = G(S,\omega)W(\omega) = 2eIW(\omega), \tag{139}$$

where we have used the pure shot-noise result, Eq. (113), and the window factor

$$W(\omega) = \left|\frac{1}{L}\int_0^T v(t)e^{(-j\omega t)}dt\right|^2 \tag{140}$$

accounts for the transit-time effects and reduces the noise from the maximum value of $W(0) = 1$ eventually to $W(\infty) = 0$, although not necessarily monotonically. The time dependence of the velocity will be determined by the nature of the forces acting on the electron.

4.8 Generation–Recombination Process

The generation–recombination process in semiconductors is a second example of a shot-noise process in which the diffusion constants can be calculated from first principles on the basis of an understanding of the physics of the process. It is also an example of what statisticians refer to as a birth and death process.

Let us define the random integer variable n as the occupancy of some state. We assume that the particles are generated at the rate $G(n)$ and disappear at the rate $R(n)$ (recombination). Then our mean equation of motion is

$$\frac{\partial\langle n\rangle}{\partial t} = \langle G(n)\rangle - \langle R(n)\rangle \equiv \langle A(n)\rangle, \tag{141}$$

where $\langle A(n)\rangle$ is our drift vector. Since the occupancy of state n is increased by generation from the state $n - 1$ and reduced by generation out of the state n, whereas it is increased by recombination out of state $n + 1$ and reduced by recombination out of state n, the probability distribution function $P(n,t)$ obeys the following equation:

$$\begin{aligned}\frac{\partial P(n,t)}{\partial t} &= G(n-1)P(n-1,t) - G(n)P(n,t)\\ &\quad + R(n+1)P(n+1,t) - R(n)P(n,t).\end{aligned} \tag{142}$$

The Poisson process is a special case of the present one with $R = 0$, $G = \nu$. If the derivative is written as a difference, and the symbol n is replaced by n', Eq. (142) can be written as a master equation

$$\begin{aligned}&\frac{P(n',t+\Delta t) - P(n,t)}{\Delta t} = \\ &\sum_n P(n',t+\Delta t|n,t)/\Delta t P(n,t)\end{aligned} \tag{143}$$

with the transition rate

$$\begin{aligned}&\frac{P(n',t+\Delta t|n,t)}{\Delta t}\\ &\quad = \sum_n \{G(n)\delta_{n,n'-1} + R(n)\delta_{n,n'+1}\\ &\quad - [G(n) + R(n)]\delta_{n,n'}\}.\end{aligned} \tag{144}$$

The rth diffusion coefficient D_r, defined by (see Secs. 7.1 and 7.2)

$$\begin{aligned}r!D_r &= \frac{\langle (n'-n)^r\rangle}{\Delta t}\\ &= \sum_{n'} \frac{(n'-n)^r P(n',t+\Delta t|n,t)}{\Delta t}\end{aligned} \tag{145}$$

takes the value

$$\begin{aligned}r!D_r &= \sum_{n'} (n'-n)^r[G(n)\delta_{n,n'-1} + R(n)\delta_{n,n'+1}]\\ &= G(n) + (-1)^r R(n)\end{aligned} \tag{146}$$

after omission of noncontributing terms from Eq. (144).

A simpler procedure is to use Taylor's theorem in the operator form

$$\exp\left(h\frac{\partial}{\partial n}\right)F(n) = F(n+h), \tag{147}$$

with h set to $+1$ or -1, to rewrite Eq. (142) in the operational form

$$\begin{aligned}\frac{\partial P}{\partial t} &= \left[\exp\left(-\frac{\partial}{\partial n}\right) - 1\right]G(n)P(n)\\ &\quad + \left[\exp\left(\frac{\partial}{\partial n}\right) - 1\right]R(n)P(n).\end{aligned}$$

Expanding the Taylor series leads *immediately* to

$$\frac{\partial P}{\partial t} = \sum_{r=1}^{\infty}\left(-\frac{\partial}{\partial n}\right)^r \frac{[G(n) + (-1)^r R(n)P(n)]}{r!}P(n). \tag{148}$$

The rth-order diffusion constant derivative term read off from the coefficient of the rth derivative, in accord with Eq. (273), agrees with Eq. (146). [See Sec. 7.2 and Lax (1966a) for a detailed discussion of the generalized Fokker–Planck equation.] Thus all the even-numbered diffusion constants are proportional to the sum of the rate in plus the rate out, while all the odd-numbered *diffusion* constants are proportional to the difference, i.e., the rate in minus the rate out. In particular, the first and second moments, D_1 and D_2, obey

$$D_1 = A(a) = G - R, \tag{149}$$

$$2!D_2 = 2D = G + R = \text{rate in} + \text{rate out}. \tag{150}$$

These results are characteristic of shot noise.

Thus, in the quasilinear approximation, the operating point n_0 and decay parameter Λ are determined by Eqs. (10) and (11):

$$A(n_0) = G(n_0) - R(n_0) = 0, \tag{151}$$

$$\Lambda \equiv -\frac{\partial A}{\partial n}\bigg|_{n=n_0} = R'(n_0) - G'(n_0), \tag{152}$$

and the diffusion coefficient is

$$D = \tfrac{1}{2}[G(n_0) + R(n_0)] = G(n_0). \tag{153}$$

Then the Einstein relation, Eq. (15), in the case of a single variable yields the mean square fluctuation from the average value in terms of the diffusion constant D and the decay constant Λ:

$$\langle(\Delta n)^2\rangle = \frac{D}{\Lambda} = \frac{G(n_0)}{R'(n_0) - G'(n_0)}. \tag{154}$$

We now want to calculate the autocorrelation function, $\langle\Delta n(t)\Delta n(0)\rangle$. Since the transport equation and its solution are

$$\begin{aligned}\frac{\partial\langle\Delta n(t)\rangle}{\partial t} &= \langle A\rangle \approx -\Lambda\langle\Delta n\rangle,\\ \langle\Delta n(t)\rangle &= e^{-\Lambda t}\Delta n(0),\end{aligned} \tag{155}$$

the regression theorem and stationarity determine the correlations, via Eq. (18), to be

$$\langle\Delta n(t)\Delta n(0)\rangle = e^{-\Lambda|t|}\langle[\Delta n(0)]^2\rangle. \tag{156}$$

The noise in n is then given by Eq. (26),

$$G(n,\omega) = \frac{4\Lambda}{\Lambda^2 + \omega^2}\langle(\Delta n)^2\rangle. \tag{157}$$

5. RESISTANCE MODULATION FLUCTUATIONS

5.1 Conductivity Fluctuations

Information about carrier-number fluctuations can be obtained by injecting a constant current and measuring the voltage fluctuations induced by the conductivity modulation caused by carrier-concentration fluctuations. The admittance can be written

$$Y = A(pe\mu_p + ne\mu_n) = (Pe\mu_p + Ne\mu_n)/L, \tag{158}$$

where μ_p and μ_n are the hole and electron mobilities, p and n are the hole and electron concentrations, and P and N are the total hole and electron numbers over the volume AL between the electrodes, of area A and separation L. Thus the fractional voltage fluctuations are given by

$$\frac{\Delta V}{V} = -\frac{\Delta Y}{Y} = -\frac{\Delta P\mu_p + \Delta N\mu_n}{P\mu_p + N\mu_n}. \tag{159}$$

If only electrons and holes are present (and not traps), charge neutrality will be enforced up to the (very high) dielectric relaxation frequency, so that to an excellent approximation

$$\Delta N(t) = \Delta P(t). \tag{160}$$

Thus the voltage autocorrelation is given by

$$\begin{aligned}\langle\Delta V(t)\Delta V(0)\rangle &= V^2[(1 + b)/(P + Nb)]^2\\ &\quad\times\langle\Delta P(t)\Delta P(0)\rangle,\end{aligned} \tag{161}$$

where $b = \mu_n/\mu_p$. It follows that the voltage noise is given by

$$G(V,\omega) = V^2\left(\frac{1 + b}{P + Nb}\right)^2\langle(\Delta P)^2\rangle\int_0^\infty 4\cos\omega t dt\,\Phi(t), \tag{162}$$

and the total voltage fluctuation may be obtained by replacing the integral by unity. The "after-effect function" $\Phi(t)$ is defined by

$$\langle\Delta P(t)\Delta P(0)\rangle = \langle(\Delta P)^2\rangle\Phi(t). \tag{163}$$

The total noise, which only involves $\Phi(0) = 1$, is consistent with the normalization condition, Eq. (27), in the noise spectrum. The after-effect function will be calculated in later sections.

5.2 A Thermodynamic Treatment of Total Carrier Fluctuations

We first consider a set of free electrons in the nondegenerate case when Boltzmann (rather than Fermi) statistics are applicable. Then Shockley (1950) has shown that the total number of electrons obeys the statistical mechanical relation

$$N = N_c \exp[(\mu - E_c)/kT], \tag{164}$$

where E_c is the energy at the bottom of the conduction band and N_c is a temperature-dependent effective density of states. The conventional symbol μ is used to represent the electron Fermi level or chemical potential. Since it will not appear later, there will be no confusion with the mobility. The thermodynamic formula, Eq. (3), then yields

$$\langle(\Delta N)^2\rangle = kT\left(\frac{\partial N}{\partial \mu}\right)_T = N. \tag{165}$$

This result is not entirely surprising, since the total number of carriers is an integral

$$N = \int_V \sum_j \delta(\mathbf{r} - \mathbf{r}_j)d^3r; \tag{166}$$

this is the three-dimensional analog of Eqs. (105) and (133). Indeed, Eq. (165) can be derived directly from Eq. (166) using only the assumption that the $\mathbf{r}_j$ are uniformly distributed in space.

A less obvious case is that of a set of N_t traps interacting with a reservoir of chemical potential μ_t. We assume that the trap occupancy is sufficiently high that Fermi statistics are necessary. In that case, the number of filled traps by use of Fermi–Dirac statistics is

$$\hat{N} = \frac{N_t}{1 + \exp[(E_t - \mu_t)/kT]}. \tag{167}$$

In that case, the thermodynamic formula, Eq. (165), becomes

$$\langle(\Delta \hat{N})^2\rangle = kT\left(\frac{\partial \hat{N}}{\partial \mu_t}\right)_T = \hat{N}\left[1 - \frac{\hat{N}}{N_t}\right]. \tag{168}$$

It can be seen that the fluctuations are reduced by a factor equal to the fraction of empty states. The reason for this result is made clear in the next section, in which a kinetic approach is used for the same problem. If both N and $\hat{N}$ are allowed to vary simultaneously, the simplest distribution consistent with these second moments is

$$W \propto \exp\left[-\frac{(\Delta N)^2}{2\langle(\Delta N)^2\rangle} - \frac{(\Delta \hat{N})^2}{2\langle(\Delta \hat{N})^2\rangle}\right]. \tag{169}$$

The term in $\Delta N \Delta \hat{N}$ vanishes because N does not depend on μ_t and $\hat{N}$ does not depend on μ. Within the quasilinear approximation, it is appropriate to ignore higher cumulants than the second and stop at the Gaussian approximation.

Suppose, now, that the electrons in traps do not have an independent reservoir, but are obtained from the free carriers. Then we must impose the conservation condition

$$\Delta \hat{N} = -\Delta N. \tag{170}$$

If this constraint is inserted into Eq. (169), we obtain a Gaussian in a single variable with the second moment

$$\begin{aligned}\langle(\Delta N)^2\rangle = \langle(\Delta \hat{N})^2\rangle &= \frac{\langle(\Delta N)^2\rangle\langle(\Delta \hat{N})^2\rangle}{\langle(\Delta N)^2\rangle + \langle(\Delta \hat{N})^2\rangle} \\ &= \frac{N\hat{N}(N_t - \hat{N})}{NN_t + \hat{N}(N_t - \hat{N})}.\end{aligned} \tag{171}$$

In the next section, we show that this result can be derived by the Einstein relation.

The situation for holes is similar to that for electrons. If the holes have their own reservoir, then the typical Poisson process prevails:

$$\langle(\Delta P)^2\rangle = P. \tag{172}$$

If holes, traps, and electrons are all present and coupled to each other, then charge neutrality imposes the constraint

$$\Delta N + \Delta \hat{N} = \Delta P. \tag{173}$$

In the presence of compensating centers, N_{co}, there is also a neutrality condition for the steady state:

$$N + \hat{N} = P + N_t - N_{\text{co}}. \tag{174}$$

The influence of these constraints on the fluctuations is developed in the next section.

5.3 Einstein Derivation of Carrier Fluctuations with Traps

As in the previous section, we consider traps interacting with free electrons. This problem is a generation–recombination problem with a generation rate proportional to the number of carriers in the traps and a recombination rate proportional to the product of the number of free carriers and the number of empty traps,

$$g(N) = \gamma\hat{N}, \qquad r(N) = \rho N(N_t - \hat{N}). \tag{175}$$

We regard g and r as functions only of N since $\hat{N}$ is a function of N given in Eq. (174). The diffusion coefficient in the steady state is then

$$D(N) = r(N) \tag{176}$$

in view of Eqs. (151) and (153); and the decay coefficient, in accord with Eq. (152), is

$$\Lambda = r'(N) - g'(N) = \rho[(N_t - \hat{N}) + N] - (-\gamma), \tag{177}$$

where the derivatives were taken using $\Delta\hat{N} = -\Delta N$. The electron carrier fluctuation $\langle(\Delta N)^2\rangle = D/\Lambda$ is in agreement with Eq. (171).

We next consider carrier fluctuations in a p-type material in the presence of electrons and traps. We will regard N and $\hat{N}$ as the determining variables, with P obtained from the neutrality constraint, Eq. (174). Our discussion will follow Lax and Mengert (1960) with the replacement of C/Ω in that article by ρ, and of g in that article by γ, to conform with our notation in Eq. (175). Our nonlinear transport equation is taken to be

$$\begin{aligned}\frac{d\langle N\rangle}{dt} &= \langle G - R + \gamma\hat{N} - \rho(N_t - \hat{N})N\rangle,\\ \frac{d\langle\hat{N}\rangle}{dt} &= \langle\rho(N_t - \hat{N})N - \gamma\hat{N}\rangle.\end{aligned} \tag{178}$$

The terms in G and R refer to generation and recombination of electron–hole pairs, and the other terms refer to generation and recombination from the traps, as in Eq. (175).

The underlying transition rate per unit time for this model is

$$\begin{aligned}&w(N',\hat{N}';N,\hat{N})\\ &\quad= [G\delta(N',N+1) + R\delta(N',N-1)]\delta(\hat{N}',\hat{N})\\ &\qquad+ \rho\delta(N',N+1)\delta(\hat{N}',\hat{N}-1)\\ &\qquad+ \gamma(N_t - N)N\delta(N',N-1)\delta(\hat{N}',\hat{N}+1).\end{aligned} \tag{179}$$

The first moments of the transition rate are obtained by summing over the primed variables:

$$\begin{aligned}&\sum(N' - N)w(N',\hat{N}';N,\hat{N}) = G - R + \rho\hat{N} - \nu N,\\ &\sum(\hat{N}',-\hat{N})w(N',\hat{N}';N,\hat{N}) = \nu N - \rho\hat{N},\end{aligned} \tag{180}$$

where we use the abbreviation

$$\nu = \gamma(N_t - \hat{N}). \tag{181}$$

The results agree with our phenomenological transport equations, Eq. (178). If we linearize these equations, we obtain the two-by-two system

$$\frac{d}{dt}\begin{bmatrix}\langle\Delta N\rangle\\ \langle\Delta N\rangle\end{bmatrix} = -\mathbf{\Lambda}\begin{bmatrix}\langle\Delta N\rangle\\ \langle\Delta N\rangle\end{bmatrix} \tag{182}$$

with

$$\mathbf{\Lambda} = \begin{bmatrix}\nu + r & -g'\\ -\nu & g'\end{bmatrix}, \tag{183}$$

where the parameters are now evaluated under the steady-state conditions

$$\gamma\hat{N} = \rho(N_t - \hat{N})N = \nu N, \quad R = G. \tag{184}$$

The symbol r now stands for the experimental electron–hole recombination rate

$$r = \frac{\partial(R - G)}{\partial N}, \tag{185}$$

and

$$g' = \gamma + \rho N. \tag{186}$$

The diffusion constants may be obtained from the second moments of the transition rate:

$$\begin{aligned}D_{11} &= \sum(N' - N)^2 w(N',\hat{N}';N,\hat{N})/2\\ D_{22} &= \sum(\hat{N}' - \hat{N})^2 w(N',\hat{N}';N,\hat{N})/2\\ D_{12} &= \sum(N' - N)(\hat{N}' - \hat{N})w(N',\hat{N}';N,\hat{N})/2,\end{aligned} \tag{187}$$

so that

$$D_{11} = [(G + R) + (\gamma\hat{N} + \nu N)]/2, \tag{188}$$

where the evaluation of D_{11} was done at the steady state using the conditions

$$G = R, \quad \nu N = \gamma \hat{N}. \tag{189}$$

The complete matrix at the steady state is thus found to be

$$\mathbf{D} = \left| \begin{matrix} R + \gamma\hat{N} & -\gamma\hat{N} \\ -\gamma\hat{N} & \gamma\hat{N} \end{matrix} \right]. \tag{190}$$

The population fluctuations that result from Eq. (15) are then

$$\mathbf{\Lambda}^{-1}\mathbf{D} = \begin{bmatrix} R/r & 0 \\ (\nu R/r - \gamma\hat{N})/g' & \hat{N}\gamma/g' \end{bmatrix}. \tag{191}$$

There is a coupling between the free and bound electron populations. However, if the hole (majority) population P is much larger than N or $\hat{N}$, it would provide uncoupled reservoirs for these two populations, and we would expect the 21 element of the matrix to disappear. If we write the recombination as proportional to the product of the number of holes and electrons,

$$R = \alpha(N + \Delta N)(P + \Delta P), \quad G = NP, \tag{192}$$

then

$$R - G = \alpha[(P + N)\Delta N + N\Delta\hat{N}] \tag{193}$$

after use of charge neutrality, Eq. (173). Thus Eq. (185) yields

$$r = \alpha(P + N) \approx \alpha P = R/N. \tag{194}$$

In view of Eq. (189), the 21 elements vanish. With the help of Eq. (181), we find

$$\gamma/g' = 1 - (\hat{N}/N_t), \tag{195}$$

so that

$$\mathbf{\Lambda}^{-1}\mathbf{D} = \begin{bmatrix} N & 0 \\ 0 & \hat{N}' \end{bmatrix}, \tag{196}$$

where in this connection we have introduced the shorthand notation

$$\hat{N}' = \hat{N}[1 - (\hat{N}/N_t)]. \tag{197}$$

5.4 The Spectrum of Resistance Modulation Fluctuations

We return to the problem of the last section with a large number of majority carriers, P, a smaller number of trapped carriers, $\hat{N}$, and a still smaller number of free electrons, N. Because of the neutrality condition, Eq. (173), the voltage fluctuation of Eq. (159) can be written

$$\Delta V = C_1\Delta N + C_2\Delta\hat{N}, \tag{198}$$

where

$$C_2 = V/(P + bN), \quad C_1 = (1 + b)C_2 \tag{199}$$

and $b = \mu_n/\mu_p$.

In view of the total population fluctuations given in Eqs. (196) and (197), the total voltage noise is given by

$$\begin{aligned} \int_0^\infty G(V,f)df &= \langle(\Delta V)^2\rangle, \\ &= |C_2|^2[\hat{N}' + (1 + b)^2 N]. \end{aligned} \tag{200}$$

The spectrum of voltage fluctuations is given by Eqs. (29) and (31):

$$\begin{aligned} G(V,f) = 4C_2^2\,\mathrm{Re}\{&(i\omega\mathbf{1} + \mathbf{\Lambda})_{22}^{-1}\hat{N}' + (1 + b) \\ &\times [(i\omega\mathbf{1} + \mathbf{\Lambda})_{12}^{-1}\hat{N}' + (i\omega\mathbf{1} + \mathbf{\Lambda})_{21}^{-1}N] \\ &+ (1 + b)^2(i\omega\mathbf{1} + \mathbf{\Lambda})_{11}^{-1}N\}. \end{aligned} \tag{201}$$

Here we have omitted the star, since $\mathbf{\Lambda}$ is real, and replaced $-j$ by i to avoid minus signs. The elements of $\mathbf{\Lambda}$ were given in Eq. (183). An exact evaluation by Lax and Mengert (1960) yields a ratio of a quadratic to a quartic in ω. Lax and Mengert re-expressed their results by a partial-fraction analysis into the more understandable form

$$G(V,f) = \frac{4V^2}{(P + Nb)^2}\left[\frac{A_S\lambda_S}{\omega^2 + \lambda_S^2} + \frac{A_F\lambda_F}{\omega^2 + \lambda_F^2}\right] \tag{202}$$

of a sum of two Lorentzians whose integrated contributions are A_S and A_F associated with the slow and fast modes with eigenvalues λ_S and λ_F. The trace and determinant of $\mathbf{\Lambda}$, Eq. (183), yield the sum and product relations

$$\lambda_F = \nu + r + g' - \lambda_S, \quad \lambda_S = rg'/\lambda_F. \tag{203}$$

Typical parameters from the experiments of Hornbeck and Haynes (1955) and Haynes and Hornbeck (1955), as revised by Lax and Mengert (1960), indicate that the trapping rate $\nu = 3 \times 10^7$/s for free electrons greatly exceeds the effective release rate $g' = 3/\mathrm{s}$ from traps. These would be the basic decay rates for N and $\hat{N}$ respectively. It is then appropriate to assume that the slow rate $\lambda_S \ll \lambda_F$. If we iterate starting with $\lambda_S = 0$, we obtain a first approximation

$$\lambda_F \approx \nu + r + g', \quad \lambda_S \approx g'r/(\nu + r + g') \tag{204}$$

Since $\lambda_F \approx \nu = 3 \times 10^7$, and $\lambda_S \approx g'r/\nu \approx 3 \times 10^{-3}$, Eq. (204) has a relative accuracy of $\lambda_S/\lambda_F \approx 10^{-10}$. Lax and Mengert (1960) use this large decay ratio to simplify the formulas for A_S and A_F to

$$A_S \approx \hat{N}' + (1 + b)N, \quad A_F \approx (b^2 + b)N. \tag{205}$$

Perhaps the surprise in these numbers is that the slow eigenvalue λ_S of the combined system is 1000 times smaller than the emission rate from traps, which represents the decay of $\hat{N}$. The reason this is so is that the electrons rapidly ("adiabatically") respond to the bound trap density and generate captures that almost cancel the release rate. To verify this, we note that the adiabatic procedure in lowest order, as described in Lax (1967), is to set the time derivative of the rapidly changing quantity, $d\langle \Delta N \rangle/dt = 0$. This yields

$$\Delta N = [g'/(\nu + r)]\Delta\hat{N}. \tag{206}$$

When this is inserted into the second equation (for $\Delta\hat{N}$), the effective slow equation takes the form

$$d\langle\Delta\hat{N}\rangle = \lambda_S\langle\Delta\hat{N}\rangle \tag{207}$$

with the reduced value

$$\lambda_S \approx g'r/(\nu + r) \tag{208}$$

obtained after a near cancellation between the two terms.

6. CONCENTRATION FLUCTUATIONS IN SEMICONDUCTORS

6.1 General Theory of Concentration Fluctuations

The thermodynamic discussion of occupancy fluctuations in Sec. 5.2 can be generalized by noting that the average occupancy of a state of energy E is given by

$$\langle n \rangle = 1/\{\exp[\beta(E - \mu) + \epsilon]\}, \tag{209}$$

where for Fermi, Boltzmann, and Bose particles

$$\epsilon = \begin{cases} 1 \text{ for Fermi particles,} \\ 0 \text{ for Boltzmann particles,} \\ -1 \text{ for Bose particles.} \end{cases} \tag{210}$$

For application to a particular state a, replace n by $n(a)$ and E by $E(a)$. The fluctuation in occupancy of that state is given by the thermodynamic formula, Eq. (3), or the first part of Eq. (165), to be

$$\begin{aligned}\langle\Delta n(a)\Delta n(c)\rangle &= n(a)'\delta(a,c) \\ &\equiv n(a)[1 + \epsilon n(a)]\delta(a,c),\end{aligned} \tag{211}$$

which includes all three statistical cases, Fermi, Boltzmann and Bose, with the three choices of ϵ above. This result is true in equilibrium.

We have also established the truth of Eq. (211) for the nonequilibrium steady state in Lax (1960) by explicitly constructing a model in which there are transition probabilities for the transfer of particles between states, by determining the diffusion coefficients for this model, and then by solving the Einstein relation, Eq. (14), for the second moments. Since the details of the analysis are not particularly illuminating, they will not be repeated here except to note that if there is a constraint, such as the total number of particles in all the states being fixed, the simple formula, Eq. (211) above, is replaced by

$$\langle\Delta n(a)\Delta n(c)\rangle = n(a)\delta(a,c) - n(a)'n(c)' \Big/ \sum_b n(b)'. \tag{212}$$

This result clearly obeys the constraint

$$\left\langle\left[\sum_a \Delta n(a)\right]\Delta n(c)\right\rangle = 0. \tag{213}$$

Another point of interest is that the diffusion matrix in Lax (1960), Sec. 12, was found to have the diagonal elements

$$\begin{aligned}2D_{bb} = &\text{ total jump rate out of } b \\ &+ \text{total jump rate into } b,\end{aligned} \tag{214}$$

and the nondiagonal elements are

$$2D_{ab} = -[\text{transition rate from } a \text{ to } b + \text{rate from } b \text{ to } a]. \tag{215}$$

We therefore regard such diffusion constants as characteristic of shot noise because of particle-number quantization.

6.1.1 Application to Semiconductors with Electrons, Holes, and Traps. Equation (212) can be readily applied to the case in which $n(1) = N =$ the number of conduction electrons, $n(2) = \hat{N} =$ the number of trapped electrons, and $n(3) =$ number of electrons in the valence band $= N_v - P$, where N_v is the number of valence-band states and $P =$ the number of holes. Thus

$$n(1)' = N[1 - N/N_c] \approx N, \tag{216}$$

$$n(2)' = \hat{N}[1 - \hat{N}/N_t] = \hat{N}', \tag{217}$$

$$n(3)' = (N_v - P)(P/N_v) \approx P. \tag{218}$$

We have assumed nondegeneracy for the holes and the free electrons, but not for the trapped electrons. Since $\Delta n(3) = -P$, we can write the second moments in the form

$$\langle(\Delta N)^2\rangle = N - N^2/(P + N + \hat{N}'), \tag{219}$$

$$\langle(\Delta\hat{N})^2\rangle = \hat{N}' - (\hat{N}')^2/(P + N + \hat{N}), \tag{220}$$

$$\langle(\Delta P)^2\rangle = P - P^2/(P + N + \hat{N}), \tag{221}$$

$$\langle\Delta N\Delta\hat{N}\rangle = -N\hat{N}/(P + N + \hat{N}), \tag{222}$$

$$\langle\Delta P\Delta N\rangle = PN/(P + N + \hat{N}), \tag{223}$$

$$\langle\Delta P\Delta\hat{N}\rangle = P\hat{N}/(P + N + \hat{N}). \tag{224}$$

These results are consistent with the constraint $\Delta P = \Delta N + \Delta\hat{N}$ of charge neutrality. If the number of traps is zero, they reduce to $\Delta N = \Delta P$ and

$$\langle(\Delta N)^2\rangle = \langle(\Delta P)^2\rangle = NP/(N + P). \tag{225}$$

When $P \gg N$ and $P \gg \hat{N}$, these second moments reduce to those in Eq. (196), for the case of electrons and traps in the presence of majority holes.

6.2 The Influence of Drift and Diffusion on Resistance Modulation Noise

To concentrate on the influence of drift and diffusion on density fluctuations and modulation noise, let us return to the trap-free case discussed in Sec. 5.1. The spectrum of voltage noise, already given in Eq. (162), can be rewritten as a product

$$G(V,\omega) = \langle(\Delta V)^2\rangle g(\omega) \tag{226}$$

of the total noise

$$\langle(\Delta V)^2\rangle = V^2\left(\frac{1 + b}{P + Nb}\right)^2 \langle(\Delta P)^2\rangle \tag{227}$$

and the normalized spectrum

$$g(\omega) = \int_0^\infty 4 \cos\omega t dt \Phi(t), \tag{228}$$

where

$$1 = \int_0^\infty g(\omega) df = \int_{-\infty}^\infty g(\omega) \frac{d\omega}{4\pi}. \tag{229}$$

Note that this normalization is four times that used for $g(\omega)$ in Lax and Mengert (1960). For simplicity, we confine ourselves to a one-dimensional geometry, as was done by Hill and van Vliet (1958), and calculate the total hole fluctuation as

$$\Delta P(t) = \int_0^L \Delta p(x,t) dx. \tag{230}$$

We can now apply our techniques to continuous-parameter systems by replacing the sum

$$\alpha_i(t) = \sum_i [\exp(-\Lambda t)]_{ij}\alpha_j(0) \tag{231}$$

by the integral

$$\alpha_x(t) = \int [\exp(-\Lambda t)]_{xx'} dx' \alpha_{x'}(0). \tag{232}$$

Introducing a more convenient notation for the Green's function

$$[\exp(-\Lambda t)]_{xx'} = K(x,x',t), \tag{233}$$

we can write

$$\Delta p(x,t) = \int K(x,x'',t)dx''\Delta p(x'',0), \tag{234}$$

so that the correlation at two times is, as usual, related to the pair correlation at the initial time:

$$\langle \Delta p(x,t)\Delta p(x',0)\rangle = \int K(x,x'',t)dx'' \times \langle \Delta p(x'',0)\Delta p(x',0)\rangle. \tag{235}$$

It is customary to treat fluctuations at the same time at two places as uncorrelated. This is clearly the case, for independent carriers. It is less obvious when Coulomb attractions (say between electrons and holes) are included. It was shown, however, in Appendix C of Lax and Mengert (1960) that a delta-function correlation is valid, as long as we are dealing with distances greater than the screening radius. Thus we can take

$$\langle \Delta p(x'',0)\Delta p(x',0)\rangle = \langle (\Delta P)^2\rangle L^{-1}\delta(x'' - x'), \tag{236}$$

where the coefficient of the delta function is chosen so that the fluctuation in the total number of carriers $\langle (\Delta P)^2\rangle$ is given correctly by Eq. (225). Here L is the distance between the electrodes.

The definition, Eq. (163), of $\Phi(t)$ yields the expression

$$\Phi(t) = \frac{1}{L}\int_0^L dx \int_0^L dx' K(x,x',t). \tag{237}$$

If the Green's function is defined, appropriately as in Lax (1960), to vanish for $t < 0$, it will obey an equation of the form

$$\left(\frac{\partial}{\partial t} + \Lambda\right)K(x,x',t) = \delta(x - x')\delta(t), \tag{238}$$

where, in the continuous-variable case,

$$\Lambda = r + v\frac{\partial}{\partial x} - D\frac{\partial^2}{\partial x^2} \tag{239}$$

is an operator. Here, v and D are the ambipolar drift velocity and diffusion constants found by Van Roosebroeck (1953) to describe the coupled motion of electrons and holes while maintaining charge neutrality:

$$v = e\mu_a E, \quad \mu_a = (N - P)\mu_n\mu_p/(N\mu_n + P\mu_p); \tag{240}$$

$$D_a = (N + P)D_pD_n/(ND_n + PD_p), \tag{241}$$

where the individual diffusion constants and mobilities are related by the Einstein relation.

Equation (238) for the Green's function can be solved by a Fourier-transform method:

$$K(x,x',t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp[ik(x - x') - \lambda(k)t]dk, \tag{242}$$

where

$$\lambda(k) = r + ivk + Dk^2. \tag{243}$$

Here $\lambda(k)$ are the eigenvalues of the Λ operator,

$$\Lambda \exp(ikx) = \lambda(k)\exp(ikx). \tag{244}$$

With Eq. (242) for K, the after-effect function can be calculated from Eq. (237):

$$\Phi(t) = \frac{L}{2\pi}\int_{-\infty}^{\infty} dk\,\frac{\sin^2(z)}{z^2}\exp[-\lambda(k)t], \tag{245}$$

where $z = kL/2$. Thus the spectrum, Eq. (228), is

$$g(\omega) = \frac{4}{\pi}\int_{-\infty}^{\infty} dz\,\frac{\sin^2 z}{z^2}\frac{1}{i\omega + \lambda(z)}, \tag{246}$$

where λ has been re-expressed as a function of z,

$$\lambda(z) = r + 2i(v/L)z + 4(D/L^2)z^2. \tag{247}$$

Lax and Mengert (1960) provide an exact evaluation of this integral. However, the resulting expressions are complicated. It is therefore worth while to treat some limiting cases. For example, if there is no diffusion, then

$$K(x,x',t) = \exp(-rt)\delta(x - x' - vt), \tag{248}$$

and the after-effect function is given by

$$\Phi(t) = \exp(-rt)[1 - (t/T_a)], \tag{249}$$

where $T_a = L/v$ is the transit time and the spectrum is governed by a windowing factor W:

$$g(\omega) = 4T_aW \tag{250}$$

with the window factor given by

$$W = \int_0^{T_a} \cos\omega t e^{-rt}\left[1 - \frac{t}{T_a}\right]d\left(\frac{t}{T_a}\right). \tag{251}$$

Indeed, the current noise, in this special case, can be written in the form given by Hill and van Vliet (1958),

$$G(I,\omega) \equiv (I^2/V^2)G(I,\omega) = 2eI_{\text{eq}}, \tag{252}$$

which emphasizes the similarity to shot noise. The equivalent current is defined by

$$I_{\text{eq}} = 2I\frac{b + 2 + b^{-1}}{|P/N - N/P|}W. \tag{253}$$

The window factor still takes a complicated form,

$$W = \frac{(\tau/T_a)^2}{1 + (\omega\tau)^2}\left\{\frac{T_a}{\tau} - \frac{1 - (\omega\tau)^2}{1 + (\omega\tau)^2} + \frac{\exp(-T_a/\tau)}{1 + (\omega\tau)^2}\right.$$
$$\left.\times \{[1 - (\omega\tau)^2]\cos(\omega T_a) - 2\omega\tau\sin(\omega T_a)\}\right\}. \tag{254}$$

Here $\tau = 1/r$. Even this result is complicated to understand. If we take the limiting case when recombination is unimportant over the transit time, the result simplifies to

$$W = \frac{1}{2}\frac{\sin^2(\omega T_a/2)}{(\omega T_a/2)^2}, \tag{255}$$

a windowing factor similar to that found associated with the effect of transit time on shot noise. See Sec. 4.2.

In the opposite limit, in which diffusion is retained but drift is neglected, the exact result for the spectrum is given by

$$g(\omega) = 4\text{Re}J, \tag{256}$$

where

$$J = (r + i\omega)^{-1} - [D/(r + i\omega)^3]^{1/2}[1 - \exp(-\Gamma L)]/L \tag{257}$$

and

$$\Gamma = [(r + i\omega)/D]^{1/2} \tag{258}$$

is the reciprocal of the diffusion length. The exponential term represents an interference term between the two boundaries that is usually negligible since they are separated by substantially more than a diffusion length. A simple approximate form over intermediate frequencies is

$$g(\omega)/4 \approx r/(\omega^2 + r^2) + (1/L)(D/2\omega^3)^{1/2}. \tag{259}$$

In summary, in addition to the first term, which represents the volume noise easily computed just by using the total carrier, $P(t)$, effects, we get an inverse frequency to the three-halves power that arises from diffusion across the boundary at the electrodes.

7. LANGEVIN PROCESSES

7.1 Simplicity of Langevin Processes

Langevin treated noise by adding a random noise source to the linear equations describing the transport. Langevin methods, at least for a linear or quasilinear system, have the simplicity of the circuit equation of electrical engineering. The noise source may arise from thermal reservoirs as in Johnson noise, or shot noise from the discreteness of particles. But once the noise is represented as a voltage source with known moments, the physical nature of the source is no longer important. We have the analog for noise sources of current sources and Norton's theorem. The sources can be thought of as a black box, with an impedance in series with a voltage source, or an admittance in parallel with a current source. And the original nature of the sources will not enter into the solution of problem.

For the quasilinear case, we can write our set of Langevin equations in the form

$$\frac{d\boldsymbol{\alpha}}{dt} + \boldsymbol{\Lambda}\cdot\boldsymbol{\alpha} = \mathbf{F}(t), \tag{260}$$

where $\boldsymbol{\alpha} = \mathbf{a} - \langle \mathbf{a} \rangle$ is a multicomponent object, as is the force $\mathbf{F}(t)$. Equation (260) can be regarded as the definition of $\mathbf{F}(t)$. What condition must be imposed on $\mathbf{F}(t)$ in order that the resulting process $\boldsymbol{\alpha}(t)$ be a Markoffian process? (A Markoffian process is analogous to a student who only remembers the last thing he was told.) It was shown in Sec. 1 of Lax (1960) that a necessary and sufficient condition for $\boldsymbol{\alpha}(t)$ to be a Markoffian set of variables is for the Langevin forces $\mathbf{F}(t)$ to be delta correlated,

$$\langle \mathbf{F}(t)\mathbf{F}(u)\rangle = 2\mathbf{D}\delta(t-u). \tag{261}$$

The second moments in Eq. (261) are sufficient to calculate all second-order correlation functions of the α's for linear processes. Since these second-order correlations are insensitive to the higher moments of the noise sources, it is then permissible to regard the forces as Gaussian random variables. Then all linked moments beyond the second have been set equal to zero.

Shot-noise sources, however, do have higher linked moments than the second, and, at the linear level, we can consider a complete set of linked moments for any linear Markoffian process:

$$\langle F(t_1)F(t_2)\cdots F(t_n)\rangle^L = n!\mathbf{D}_n\delta(t_1-t_2)\cdots\delta(t_{n-1}-t_n). \tag{262}$$

The delta-function nature of these correlations guarantees the absence of memory required by a Markoffian process.

Indeed, if we choose

$$n!D_n = \langle(\Delta\alpha)^n\rangle/\Delta t = \nu\int \eta^n g(\eta)d\eta, \tag{263}$$

then the Langevin process, Eq. (260), reduces to the generalized one-dimensional shot-noise process of Sec. 4.2, associated with

$$\frac{da}{dt} = \sum_k \eta_k\delta(t-t_k), \tag{264}$$

where $g(\eta)$ is the probability of jumps in a of size η.

7.2 Relation to the Fokker–Planck Equation

To provide a bridge between the Langevin and Fokker–Planck points of view, we must show how the nth-order diffusion constants just introduced, by Eq. (262), into a Langevin description determine the motion of a general function $M(a)$ of the random process $a(t)$.

A general random process (not necessarily a Markoff one) necessarily obeys the relation on conditional probabilities

$$P(a', t+\Delta t|a_0, t_0) = \int P(a', t+\Delta t|a, t; a_0, t_0) \times da P(a,t|a_0,t_0). \tag{265}$$

The transition probability $P(a', t+\Delta t|a, t; a_0, t_0)$ describes the probability of arriving at a' at $t+\Delta t$ if one starts at a at time t, remembering that one started the entire process at a_0 at time t_0. This last bit of information may be omitted if the process is Markoffian. In that case Eq. (265) reduces to the Chapman–Kolmogoroff condition.

The nth-order conditional diffusion constant is defined by the relation

$$\begin{aligned} D_n(a,t|a_0t_0) &= \frac{1}{n!}\lim_{\Delta t\to 0}\int (a'-a)^n \\ &\quad\times P(a', t+\Delta t|a, t; a_0, t_0)\frac{da'}{\Delta t} \\ &= \frac{1}{n!}\lim_{\Delta t\to 0}\frac{1}{\Delta t} \\ &\quad\times \langle[a(t+\Delta t)-a(t)]^n\rangle|_{\substack{a(t)=a\\ a(t_0)=a_0}}. \end{aligned} \tag{266}$$

For a Markoffian process the dependence on a_0 can be omitted everywhere. Equation (265) can then be replaced by

$$P(a', t+\Delta t) = \int P(a', t+\Delta t|a, t)da P(a,t). \tag{267}$$

The average motion of an arbitrary function $M(a)$ of the random process $a(t)$ may be obtained by integrating $M(a')$ against $P(a', t+\Delta t)$. On the right-hand side of the equation, we shall replace $M(a')$ by its Taylor expansion

$$M(a') = M(a) + \sum_{n=1}^{\infty} (a' - a)^n \frac{M^{(n)}(a)}{n!} \tag{268}$$

The integrals over $(a' - a)^n$ give rise to the diffusion coefficients, so that we obtain

$$\langle M(a)\rangle_{t+\Delta t} = \langle M(a)\rangle_t + \Delta t \sum_{n=1}^{\infty} \int D_n(a,t) \times M^{(n)}(a)P(a,t)da. \tag{269}$$

Thus

$$\frac{d}{dt}\langle M(a)\rangle = \sum_{n=1}^{\infty} \left\langle D_n(a,t) \frac{\partial^n (M(a))}{\partial a^n}\right\rangle, \tag{270}$$

where for any function $g(a)$ the average means

$$\langle g(a)\rangle \equiv \int g(a)P(a,t)da. \tag{271}$$

To obtain the equation of motion for $P(a,t)$, we write Eq. (270) in the explicit from

$$\int \left[\sum_{n=1}^{\infty} D_n(a,t)P(a,t) \frac{\partial^n M(a)}{\partial a^n} - M(a) \frac{\partial P}{\partial t}(a,t)\right] \times da = 0. \tag{272}$$

After an integration by parts, we obtain

$$\int \left\{\sum_{n=1}^{\infty} (-1)^n \left(\frac{\partial}{\partial a}\right)^n [D_n(a,t)P(a,t)] - \frac{\partial P(a,t)}{\partial t}\right\} \times M(a)da = 0. \tag{273}$$

Since this equation is to be valid for any choice of $M(a)$, the coefficient of $M(a)$ in the above equations must vanish, yielding the generalized Fokker–Planck equation:

$$\frac{\partial P(a,t)}{\partial t} = \sum_{n=1}^{\infty} (-1)^n \left(\frac{\partial}{\partial a}\right)^n [D_n(a,t)P(a,t)]. \tag{274}$$

The ordinary Fokker–Planck equation is the special case in which the series terminates at $n = 2$.

Since Eq. (267) was true for all processes, including non-Markoffian, the existence of a generalized Fokker–Planck equation does not guarantee that the associated process is Markoffian. If we had done all our averages retaining the initial condition at t_0, we would have obtained the equation of motion

$$\frac{d}{dt}\langle f(a)\rangle_{a(t_0)=a_0} = \sum_{n=1}^{\infty} \left\langle D_n(at|a_0t_0) \frac{\partial^n f(a)}{\partial a^n}\right\rangle_{a(t_0)=a_0}, \tag{275}$$

and the Fokker–Planck equation is

$$\frac{\partial P(at|a_0t_0)}{\partial t} = \sum_{n=1}^{\infty} (-1)^n \left(\frac{\partial}{\partial a}\right)^n \times [D_n(a,t|a_0t_0)P(a,t|a_0t_0)]. \tag{276}$$

In the Markoffian case, Eq. (276) agrees with Eq. (274). Thus if (and only if) the process is Markoffian, $P(at|a_0t_0)$ obeys the same equation of motion as $P(a,t)$. Then, we can calculate $P(at|a_0t_0)$ by solving Eq. (274) subject to the initial condition

$$P(a,t) = \delta(a - a_0) \text{ at } t = t_0. \tag{277}$$

7.3 An Exactly Solvable Gaussian Example

We start with the simple example

$$\frac{dx}{dt} = \mu + \sigma f(t), \tag{278}$$

where μ and σ are constants and $f(t)$ is a Gaussian random process with mean zero and known autocorrelation

$$\langle f(t)f(u)\rangle = 2R(t - u). \tag{279}$$

Then the simpler variable $y = x - \mu t$, which obeys

$$\frac{dy}{dt} = \sigma f(t), \tag{280}$$

represents an integral over Gaussian variables of mean zero; hence it too is Gaussian with mean zero and variance

$$H(t) = \langle [y(t)]^2\rangle = 2\sigma^2 \int_0^t \int_0^t drdsR(r - s). \tag{281}$$

We have used $x(0) = y(0) = 0$ as an initial condition. Thus the Gaussian probability distribution for y is given by

$$P(y,t) = \frac{1}{[2\pi H(t)]^{1/2}} \exp\left[-\frac{1}{2}\frac{y^2}{H(t)}\right]. \tag{282}$$

The special case in which

$$H(t) = Ct^{2H} \tag{283}$$

in which C is an arbitrary constant and H is a constant between zero and one, describes "fractional Brownian motion." Ordinary Brownian motion is the special case $H = 1/2$.

A more interesting generalization occurs when one sets

$$S/S(0) = \exp(x) = \exp(y + \mu t). \tag{284}$$

Using $dx = dS/S$, Eq. (278) becomes

$$\frac{dS}{dt} = \mu S(t) + \sigma S(t) f(t). \tag{285}$$

The relations

$$y = \ln[S/S(0)\exp(-\mu t)], \quad dy = dS/S, \tag{286}$$

can be substituted into Eq. (282) to obtain the distribution for S:

$$\begin{aligned}\hat{P}(S,t|S(0),0)dS &= P(y,t|0,0)dy \\ &= \frac{1}{[2\pi H(t)]^{1/2}} \exp\left[-\frac{\ln^2[Se^{-\mu t}/S(0)]}{2H(t)}\right]\frac{dS}{S}.\end{aligned} \tag{287}$$

In the Brownian-motion limit, the noise is white, as represented by a delta correlation, $R(t - u) = \delta(t - u)$. Equation (281) specializes to $H(t) = 2Dt$ with $D = \sigma^2$. Equation (287) then reduces to

$$\begin{aligned}\hat{P}(S,t|S_0,0) &= \frac{S(0)}{S}\frac{1}{(4\pi Dt)^{1/2}} \\ &\quad \times \exp\left[-\frac{\ln^2[Se^{-\mu t}/S(0)]}{4Dt}\right].\end{aligned} \tag{288}$$

The Fokker–Planck equation obeyed by $P(x,t)$ follows from Eqs. (278) and (276) in the Brownian-motion limit in which $D_n = 0$ for $n > 2$,

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}[\mu P] + \frac{\partial^2}{\partial x^2}[\sigma^2 P]; \tag{289}$$

it contains the constant diffusion term $D = \sigma^2$, and the constant drift term $A = \mu$. One can obtain the equation for

$$\hat{P}(S,t) = P(x,t)S(0)/S \tag{290}$$

by starting from the previous equation for $P(x,t)$, introducing the relation, Eq. (284), of x to S. The result after some labor is

$$\frac{\partial \hat{P}}{\partial t} = -\frac{\partial}{\partial S}[(\mu + \sigma^2)S\hat{P}] + \frac{\partial^2}{\partial S^2}[\sigma^2 S^2 \hat{P}]. \tag{291}$$

The diffusion term [in the notation of Eq. (8)] $D_{SS} = \sigma^2 S^2$ is easy to understand. We can write

$$D_{SS} \equiv \frac{(\Delta S)^2}{2\Delta t} = \left(\frac{\partial S}{\partial x}\right)^2 \frac{(\Delta x)^2}{2\Delta t}. \tag{292}$$

Thus the diffusion constants are related by

$$D_{SS} = S^2 D_{xx} = S^2\sigma^2. \tag{293}$$

One might think that

$$\frac{\Delta S}{\Delta t} = \frac{\partial S}{\partial x}\frac{\Delta x}{\Delta t}. \tag{294}$$

However, this choice omits the second term in the drift vector. There is a correction because the diffusion coefficient D depends on S:

$$\frac{\Delta S}{\Delta t} = \frac{\partial S}{\partial x}\frac{\Delta x}{\Delta t} + \frac{1}{2}\frac{\partial^2 S}{\partial x^2}\frac{(\Delta x)^2}{\Delta t} \tag{295}$$

or

$$A_S = SA_x + \tfrac{1}{2}S(2D) = S(\mu + \sigma^2) \tag{296}$$

in agreement with Eq. (291).

7.4 Stochastic Integrals: The Ito–Stratonovich Controversy

Because mathematicians concentrate on Brownian motion, which is rather singular in behavior, it is not clear how to define the integral of a product of a random variable and a random force. In particular, how should we convert the differential equation

$$\frac{da}{dt} = B(a(t)) + \sigma(a)f(t), \tag{297}$$

where $f(t)$ is a standard white-noise source with

$$\langle f(t)f(u)\rangle = 2\delta(t-u), \tag{298}$$

to an integral? The Riemann sum of integral calculus would be

$$a(t) = a(0) + \sum_j [B(a(\bar{t}_j)) + \sigma(a(\bar{t}_j))f(\bar{t}_j)] \times (t_{j+1} - t_j), \tag{299}$$

where

$$t_j \leq \bar{t}_j \leq t_{j+1}. \tag{300}$$

The Riemann integral exists if the sum approaches a limit independent of the placement of $\bar{t}_j$ in the interval in Eq. (300). But the conditions of boundedness required for the existence of the Riemann integral (see Jeffries and Jeffreys, 1950) are violated. Thus the value of the sum approaches different limits, according to how the point $\bar{t}_j$ is chosen. [This is not true for the B term, but only the one involving the white-noise source $f(t)$.] Ito (1951), Ito and McKean (1965), and Doob (1953) avoid the difficulty by evaluating σ at the beginning of the interval. They also convert the integral to Stieltjes form by introducing

$$w(t) \equiv \int_0^t f(s)ds; \qquad \langle[w(t)]^2\rangle = 2t. \tag{301}$$

Some authors omit the factor of 2 in Eqs. (298) and (301). The result is

$$\int^t \sigma(a(s))f(s)ds = \sum \sigma(a(t_j))[w(t_{j+1}) - w(t_j)]. \tag{302}$$

However, even the latter does not converge to a unique integral, and the evaluation at the beginning is an arbitrary choice. The effect of this choice, since $f(s)$ is independent of $a(t)$ for $t < s$, is that the average of the second term in Eq. (297) vanishes, so that the Ito drift vector that follows from Eq. (297) and the integration rule, Eq. (302), is

$$\text{Ito: } A(a) = B(a) + \langle\sigma(t)f(t)\rangle = B(a). \tag{303}$$

Stratonovich (1963) makes the fortuitous choice of using

$$\overline{a(t_j)} = \tfrac{1}{2}[a(t_j) + a(t_{j+1})], \tag{304}$$

the average of the values at the two end points of each interval. In our notation, Stratonovich's choice is equivalent to equating

$$A(a) = B(a) + \langle\sigma(a)f(t)\rangle = B(a) + \sigma(a)\frac{\partial\sigma(a)}{\partial a}, \tag{305}$$

a result in agreement with the outcome, Eq. (311), below, of the iterative procedure we advocate. It is intuitively clear that an average of the end points is better than using either one. But is the average value always the best choice?

Our viewpoint, expressed in Lax (1966b), Sec. 3, is that mathematicians have concentrated too exclusively on the Brownian-motion white-noise process, which has delta correlation functions. Real processes can have a sharp correlation time of finite width. Thus their spectrum is flat, but not up to infinite frequency. Thus, for real processes, the Riemann sums do converge, and no ambiguity exists. After the integration is performed, the correlation time can be allowed to go to zero, that is, one can then approach the white-noise limit. Thus the ambiguity is removed by approaching the integration limit and the white-noise limit in the correct order.

To obtain the drift vector $A(a)$, the differential equation, Eq. (297), is rewritten as an integral equation:

$$a(t+\Delta t) - a(t) = \int_t^{t+\Delta t} B(a(s))ds + \int_t^{t+\Delta t} \sigma(a(s))f(s)ds. \tag{306}$$

This equation can be solved by iteration. The lowest approximation is obtained by replacing $a(s)$ by $a(t)$ under the integral. This first approximation,

$$(\Delta a)_1 = B(a(t))\Delta t + \sigma(a(t))\int_t^{t+\Delta t} f(s)ds, \tag{307}$$

is equivalent to the Ito choice. The first term is already of order Δt and need not be improved. In the second term, let us insert the first approximation,

$$a(s) = a(t) + B(a(t))(s - t) + \sigma(a(t)) \int_t^s f(u)du, \tag{308}$$

into Eq. (306) to get

$$\begin{aligned}(\Delta a)_2 &= B(a(t))\Delta t + \sigma(a(t)) \int_t^{t+\Delta t} f(s)ds \\ &\quad + \sigma(a(t)) \frac{\partial \sigma}{\partial a(t)} \int_t^{t+\Delta t} ds \int_t^s du f(s)f(u).\end{aligned} \tag{309}$$

We have retained only terms of order Δt, or f^2, but not $\Delta t f$ or higher. If we restrict ourselves to Gaussian processes, their linked moments of order higher than two vanish. Thus the process terminates. Indeed,

$$\langle (\Delta a)^n \rangle_L = 0 \text{ for } n > 2. \tag{310}$$

For $n = 2$ it is sufficient to use the first approximation,

$$\begin{aligned}\langle (\Delta a)^2 \rangle_L &= \sigma^2 \int_t^{t+\Delta t} ds \int_t^{t+\Delta t} du \langle f(s)f(u) \rangle \\ &= 2\sigma^2 \Delta t = 2!D_2 \Delta t.\end{aligned} \tag{311}$$

In the last step, we have specialized to the Brownian, white-noise limit. For $n = 1$ it is necessary and sufficient to use the second approximation of Eq. (309) to get

$$\begin{aligned}\langle \Delta a \rangle &= B(a(t))\Delta t \\ &\quad + \sigma(a(t)) \frac{\partial \sigma}{\partial a(t)} \int_t^{t+\Delta t} ds \int_t^s du 2\delta(s - u),\end{aligned}$$

so that

$$A(a)\Delta t = B(a)\Delta t + \sigma \frac{\partial \sigma}{\partial a} \Delta t. \tag{312}$$

The factor of 2 disappears because only half the area of the delta function contributes at the boundary point s. This result is clear if the correlation function $R(s - u)$ is any sharp symmetric function. In summary, our Langevin process is equivalent to an ordinary (no derivatives higher than second) Fokker–Planck process with coefficients

$$D(a) = [\sigma(a)]^2, \tag{313}$$

$$A(a) = B(a) + \sigma \frac{\partial \sigma}{\partial a} = B(a) + \frac{1}{2} \frac{\partial D}{\partial a}. \tag{314}$$

These results agree completely with those found in the previous section for our exactly solvable example.

We note that the Fokker–Planck can be written in two completely equivalent forms:

$$\begin{aligned}\partial P(a)/\partial t &= -\frac{\partial}{\partial a}[A(a)P(a)] + \frac{1}{2}\frac{\partial^2}{\partial a^2}[D(a)P(a)] \\ &= -\frac{\partial}{\partial a}[B(a)P(a)] \\ &\quad + \frac{\partial}{\partial a}\left[\sigma(a)\frac{\partial}{\partial a}(\sigma(a)P(a))\right].\end{aligned} \tag{315}$$

The Ito choice can be compensated for by choosing the correct Fokker–Planck equation, but his procedure is sufficiently counterintuitive that the wrong choice is often made. In applying Ito's procedure to the financial world Hull (1992) almost always makes the correct choice. Hull uses the equation

$$dx = dS/S = \mu dt + \sigma dw, \tag{316}$$

where w is the Wiener process defined in Eq. (299), which is simply Eq. (278) written in the notation preferred by mathematicians. Since σ is a constant in Hull's application, the second term on the right-hand side of Eq. (314) vanishes, and there is no distinction between the Ito and Stratonovich viewpoints. Hull (1992) then states correctly that $\Delta S/S$ is normally distributed with mean "$\mu \Delta t$ and standard deviation $\sigma\sqrt{\Delta t}$. In other words,

$$\frac{\Delta S}{S} \approx \phi(\mu \Delta t, \sigma\sqrt{\Delta t}), \tag{317}$$

where $\phi(m,s)$ denotes a normal distribution with mean m and standard deviation s." Equation (316) cannot be simply rewritten (Hull, 1992) as an Ito equation in the form

$$dS = \mu S dt + \sigma S dw \tag{318}$$

because Ito's variables do not obey the usual rules of calculus. Equation (318) is correct, as a Stratonovich equation, but the second term does not average to zero but modifies the drift term as in Eqs. (305) and (309). To be an Ito equation, the first term should contain A, the modified *drift term, and not B*, the unmodified drift term. Equation (291) shows that the correct Ito equation is

$$dS = (\mu + \sigma^2)Sdt + \sigma Sdw. \tag{319}$$

This result could also have been obtained using Ito's lemma, which describes how the drift vector changes under a transformation of variables. If, however, one makes the natural error of regarding Eq. (318) as a valid Ito equation and then makes the inverse transformation $x = \ln S/S(0)$ using Ito's lemma (now carefully), one obtains (Hull, 1992)

$$dx = [\mu - \sigma^2]dt + \sigma dw, \tag{320}$$

which disagrees with the natural starting point, Eq. (316), as does the associated distribution after a finite time interval $\Delta t = T - t$, which takes the form

$$\ln S_T - \ln S \approx \phi(\mu - \sigma^2\Delta t, \sigma\sqrt{\Delta t}) \tag{321}$$

that disagrees with Hull's earlier intuitively correct result, Eq. (317). The Ito procedure can be done correctly by using Eq. (319) instead of (318), but its counterintuitive nature can be misleading even to experts.

For the record, we note that the behavior of diffusion constants and drift vectors under a transformation from one set of variables a_j to a new set of variables a'_i was derived in Lax (1966b) from Langevin considerations that are a generalization of the above arguments to the multivariable case. The diffusion constants transform simply as

$$D'_{ij} = \frac{\partial a'_i}{\partial a_k}\frac{\partial a'_j}{\partial a_l}D_{kl}. \tag{322}$$

The B drift vectors also transform simply:

$$B'_i = \frac{\partial a'_i}{\partial t} + \frac{\partial a'_i}{\partial a_k}B_k, \tag{323}$$

where the second term is the natural transformation of a vector, and the first term enters only if the transformation is time dependent. Under a nonlinear transformation, however, there is a change in the contribution from the nonconstancy of the diffusion coefficient as shown by the last term in Eqs. (309)–(314). The A drift vectors, which enter the Fokker–Planck equation, possess therefore the more complicated transformation

$$A'_i = \frac{\partial a'_i}{\partial t} + \frac{\partial a'_i}{\partial a_k}A_k + \frac{\partial^2 a'_i}{\partial a_m \partial a_n}D_{mn}, \tag{324}$$

which can be understood simply from the relation

$$\Delta a'_i = \frac{\partial a'_i}{\partial t}\Delta t + \frac{\partial a'_i}{\partial a_k}\Delta a_k + \frac{\partial^2 a'_i}{\partial a_m \partial a_n}\frac{\Delta a_m \Delta a_n}{2}. \tag{325}$$

In summary, there are two kinds of Langevin equations, those whose random term need not average to zero [as used in Eq. (297)] and those used by Ito in which the average of the random term vanishes by Ito's definition of the stochastic integral. Both can be used correctly, but Ito's choice requires more care because the usually permissible way in which we handle equations is no longer valid. An able analysis of this situation, with references to earlier discussion, is presented by Van Kampen (1992).

Hull's (1992) book is an extremely well-written text on *Options, Futures and Other Derivative Securatives*. The Ito–Stratonovich controversy applies to physics, chemistry, and other fields. We have analyzed Hull's treatment in detail because his use of the Black–Scholes (1973) work involving an application of the Langevin Eq. (316) to the pricing of options coincides with the one example, in Sec. (7.3), for which we have an exact solution.

Although the Ito choice can be dangerous, as shown above, we trust that by now the Ito choice is used consistently in practice. However, there may be more serious problems, since real options may have statistics based on more wildly fluctuating processes than Brownian motion, such as Lévy and fractal processes. This has recently been emphasized by Peters (1994) and Bouchard and Sornette (1994). For an excellent review of Brownian and fractal walks, see RANDOM PROCESSES. Gashghaie *et al.* (1996) have also found a parallelism between prices in foreign exchange markets and turbulent fluid flow. Thus the conventional Brownian-motion approach will be invalid in such markets.

8. FURTHER CONTRIBUTIONS TO STOCHASTIC PROCESSES

8.1 Overview

In the following sections we review current contributions. We start with random walks, since the approach remains linear, and the applications are usually stationary. Then we discuss contributions in which one or more of our assumptions of (1) stationarity, (2) linearity, and (3) white noise are eliminated.

8.2 Random-Walk Problems

Many problems can be mapped into a random-walk problem on a lattice, or actually is a problem on a lattice. These problems are discussed first since the master equation for the probability density is linear, and simple techniques based on discrete periodicity can be used. Our description of the random-walk case is especially brief because the article RANDOM PROCESSES has covered the random-walk problem in detail.

Recent contributions have been made to diffusion and reaction kinetics by considering one or more walkers on a lattice. Discrete-time random walks have been generalized to continuous-time walks. For such walks the distribution $\Psi(t)$ of time intervals to a hop from a given site are not exponential. If the first or second moment of $\Psi(t)$ is not finite, one generates a "Lévy" process. Other choices lead to stretched-exponential processes. Analytic solutions are often possible by taking advantage of the regularity of the lattice, as has been shown by Montroll, Weiss, Scher, Shlesinger, and others as summarized in RANDOM PROCESSES. An extension was also made to walks on fractals and to disordered lattices. Scher and Lax (1973) showed that a disordered lattice could be replaced by a continuous-time random walk that yields a conductivity varying approximately as $\omega^{0.9}$, in agreement with the experimental conductivity found in semiconductor impurity bands.

8.3 Linear with Time-Dependent Decay

The case of homogeneous noise with linear (time-dependent) damping [namely, $\mathbf{\Lambda}(t)$, but $\mathbf{D}(t) = \mathbf{D}$] is only mentioned here since an analytic evaluation was found for the generalized characterized functional

$$M[q] = \left\langle \exp\left[i \int_0^t q(u)a(u)du \right] \right\rangle, \tag{326}$$

where $q(u)$ is an arbitrary function of u, in Eq. (2.16) of Lax (1966b). Many desired multiple-time averages can be deduced from this generalized characterized function. For example, the multiple-time moments are given by a functional derivative

$$m(t_1,t_2,\ldots t_n) = i^{-n} \left. \frac{\delta^n M[q]}{\delta q(t_1) \cdots \delta q(t_n)} \right|_{q=0}, \tag{327}$$

as discussed by Hänggi (1989).

Lax (1966a) introduced a more general functional involving an arbitrary nonlinear function of $q(u)$, and Lax and Zwanziger (1973) evaluated it for quadratic functions to obtain the photocount distribution in lasers near threshold.

8.4 The Nonlinear (Fokker–Planck) Case: Reaction-Rate Theory

If we eliminate the linearity assumption, including the case of weak noise and a quasilinearity approximation, many possibilities arise. If the variables are continuous and the noise is white and Gaussian, the problem is properly described by a Fokker–Planck equation.

Many applications have been made of Fokker–Planck theory, often with approximations that are special to a particular application. One example is to the theory of self-sustained oscillators discussed in Sec. 8.5. Applications have been made to systems subject to bifurcation, and to quantum mechanical systems. An application of long standing is to reaction-rate theory, which is described briefly here, and to stochastic resonance, which is described briefly in Sec. 8.5. Many of these applications are reviewed in the three-volume series by Moss and McClintock (1989). An excellent review of reaction-rate theory, which we cannot summarize adequately here, is the 90-page review by Hänggi *et al.* (1990).

The modern surge of effort in reaction-rate theory stems from the widely read article by Kramers (1940), since the author made a simple Fokker–Plank model for the escape of a particle from a metastable state. However, Landauer (1989), in a review, points out many examples of the rediscovery of previously published ideas. In particular, he remarks on an earlier article by Pontryagin *et al.* (1933) (PAV) that contained the essential ideas in Kramers's article.

Kramers's model, in the notation of Jung (1993), is

$$\frac{dx}{dt} = \frac{p}{m} \qquad (328)$$

$$\frac{dp}{dt} = -\gamma p - V'(x) + \sqrt{m\gamma k_B T}\,\xi(t) + E\sin\theta;$$
$$\theta \equiv \Omega t + \text{const.}, \qquad (329)$$

where $E \sin\theta$ is the periodic term needed in the next section on stochastic resonance, and the Brownian motion is induced by the noise with second moment

$$\langle \xi(t)\xi(u)\rangle = 2\delta(t-u). \qquad (330)$$

(The periodic term is not included in the articles by Kramers or PAV.)

Kramers considered both weak and strong damping cases. We review only the strong-damping case here because Kramers's extensive mathematical treatment was reviewed by Hänggi *et al.* (1990) and the main ideas, contained in the strong-damping case, are easier to explain because they lead to a closed-form solution.

In the strong-damping case, we assume that γ is faster than any other decay mechanisms, or reciprocal times. Then d/dt can be neglected compared to γ. Thus the term in dp/dt can be omitted. It is thus possible to solve Eq. (329) for p and substitute the result into Eq. (328). Discussions of the adiabatic approximation are given by Lax (1967) and Van Kampen (1985). The resulting equation for x is

$$\frac{dx}{dt} = A(x) + \sqrt{D}\xi(t), \qquad (331)$$

where

$$A(x) = -V'(x)/m\gamma + (E/m\omega)\sin\theta, \qquad (332)$$

$$D = k_B T/m\gamma. \qquad (333)$$

The Fokker–Planck equation associated with Eqs. (332) and (333) is

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}[A(x,t)P(x,t)] + \frac{\partial^2}{\partial x^2}[D(x)P(x,t)]. \qquad (334)$$

In the PAV article, there is no time-dependent force, and $A(x)$ and $D(x)$ are regarded as known arbitrary functions of x [with $D(x)$ written as $b(x)/2$].

The Fokker–Planck equation can be regarded as an equation of continuity:

$$\frac{\partial P(x,t)}{\partial t} + \frac{\partial J(x,t)}{\partial x} = 0, \qquad (335)$$

where

$$J(x,t) = A(x,t)P(x,t) - \frac{\partial D(x)P(x,t)}{\partial x}. \qquad (336)$$

If the periodic force is absent, there will be a stationary solution if $A(x)$ is a restoring force, namely, negative for large positive x and positive for large negative x. Such a stationary solution obeys

$$\frac{dJ(x)}{dx} = 0; \qquad J(x) = AP(x) - \frac{d[D(x)P(x)]}{dx}, \qquad (337)$$

and the current must be a constant,

$$J(x) = C_1. \qquad (338)$$

Although Lax (1966b), Sec. 4, has encountered a case in which a current flow is present, we are normally concerned with stationary states that carry no current. If we set $C_1 = 0$, $P(x)$ obeys a simple first-order equation whose solution will be denoted $P_0(x)$:

$$P_0(x) = \frac{C}{D(x)}\exp[-U(x)]U(x) \equiv -\int_{x_{st}}^{x}\frac{A(y)}{D(y)}dy. \qquad (339)$$

Here it is customary to choose x_{st} to be the stationary point, which obeys

$$A'(x_{st}) = 0. \qquad (340)$$

This is the point, Eq. (10), about which a quasilinear expansion would be made. PAV then introduce a probability $\phi(x,t)$ that a particle starting at $t = 0$ in the interval $a < x < b$ will reach a boundary in the time interval $[0,t]$. Instead of a direct attack on this problem, they introduce the modern approach of invariant embedding made popular by Bellman (1968) and Casti and Kalaba

(1973) by writing a recursion relation relating $\phi(x, t + \tau)$ to $\phi(x,t)$ for small τ:

$$\phi(x, t + \tau) = \int_a^b \phi(y,t)dyP(y, t + \tau|x, t), \tag{341}$$

which assumes that a limit in which $\tau \to 0$ will be taken and that "the probability of the random point leaving the interval *ab* during the short time τ is very small." PAV then expand $\phi(y,t)$ about $\phi(x,t)$. The results involve the incomplete moments

$$\mu_n = \int_a^b (y - x)^n dy. \tag{342}$$

Because the conditional probability is so narrow in y (of order $\sqrt{D\tau}$), the limits of integration can be extended to full interval $[-\infty,\infty]$ for almost all x. Then

$$\mu_0 \to 1; \qquad \mu_1 \to A(x)\tau; \qquad \mu_2 \to D(x)\tau; \tag{343}$$

with higher moments vanishing faster than linearly in τ. The result is that ϕ obeys the equation adjoint to Eq. (334) for $P(x,t)$:

$$\frac{\partial\phi(x,t)}{\partial t} = \frac{\partial\phi}{\partial x}A(x) + \frac{\partial^2\phi}{\partial x^2}D(x). \tag{344}$$

Equation (344) is no easier to solve analytically than the original Fokker–Planck equation (334). However, $dt\ \partial\phi/\partial t$ is the probability of a crossing in the interval $[t, t + dt]$, so that

$$M(x) = \int_0^\infty t\frac{\partial\phi}{\partial t}dt \tag{345}$$

is the mean time to a crossing, starting at x. By taking the time derivative of Eq. (344), multiplying by t, and integrating over t, PAV obtain

$$\int_0^\infty t\frac{\partial^2\phi}{\partial t^2}dt = A(x)\frac{dM(x)}{dx} + D(x)\frac{d^2M(x)}{dx^2}. \tag{346}$$

Since $\phi(x,0) = 0$ and $\phi(x,\infty) = 1$, integration by parts shows that the left-hand side of Eq. (346) is -1. Thus $dM(x)/dx$ obeys a first-order ordinary differential equation. The integral of this equation subject to the boundary conditions

$$M'(a) = 0, \qquad M(b) = 0, \tag{347}$$

appropriate to the case of a being a reflecting boundary and b being an absorbing boundary, is

$$M'(x) = -\exp[U(x)]\int_a^x \exp[-U(z)]\frac{dz}{D(z)}. \tag{348}$$

As pointed out by Stratonovich (1989), the final integral can be written in terms of the steady-state distribution itself:

$$M(x) = \int_x^b \frac{dy}{D(y)P_0(y)}\int_a^y dzP_0(z). \tag{349}$$

More explicit formulas are available only after making some approximations. Stratonovich (1989) simplifies the mean lifetime to

$$\begin{aligned}\bar{T} &= 2M(x_{st})\\ &= 2\pi[D(b)/D(x_{st})]^{1/2}[|A'(x_{st})|A'(b)]^{-1/2}\\ &\quad\times \exp[U(b) - U(x_{st})] = 2T_{\mathrm{K}},\end{aligned} \tag{350}$$

where T_K is referred to as the Kramers time. The factor of 2 was inserted on the assumption that after reaching the point b there is a 50% chance of escaping and a 50% chance of returning to the point x_{st} of stable equilibrium. If we revert to the original definitions of A and D in Eqs. (332) and (333), the Kramers time is given by

$$T_{\mathrm{K}} = \pi\frac{\gamma}{\omega_{st}\omega_b}\exp\left\{\frac{V(b) - V(x_{st})}{k_{\mathrm{B}}T}\right\}, \tag{351}$$

where $(\omega_j)^2 = |V''(x_j)|/m$ are the frequencies associated with quadratic approximations to the local potential at the points in question.

The results display the expected Arrhenius law associated with the activation energy. The mean rate of transition vanishes with γ demonstrating explicitly that the noise associated with the damping induces the transitions.

Although we have demonstrated the result for the one-dimensional case, the original authors all considered x to be a one-dimensional coordinate along the direction of the reaction embedded in a multidimensional space. The qualitative ideas remain the same, but the results are, of course, sensitive to details such as symmetry that are not present in the one-dimensional case. Contributions by Brinkman (1956), Landauer and Swanson (1961), and Langer (1968) are summarized in detail by Hänggi *et al.* (1990). A recent review of the current status of Kramers's reaction-rate theory is given by Talkner and Hänggi (1995).

8.5 Stochastic Resonance

The subject of stochastic resonance is concerned with systems such as those discussed by Kramers, Pontryagin *et al.*, Stratonovich, and others, to which has been added an applied periodic force. Such a term was already included in Eq. (329). Since this field is more complex than the original reaction-rate theory, there are more limits and approximations to consider. The aim of this section is to indicate the different regions for which approximate solutions have been found and to explain the nature of stochastic resonance, namely, the way in which noise-induced transitions are enhanced by a periodic force, and conversely the way in which noise can enhance the response to a periodic force. For a detailed examination of stochastic resonance, the reader is referred to the review of this subject by Jung (1993) and to the NATO Conference Proceedings (Bulsara *et al.*, 1993).

The simplest case is the application of a periodic force to an Uhlenbeck–Ornstein (1930) system. Since this system is linear, an analytical solution is possible, both of the Fokker–Planck equations and of the Langevin equations themselves. Since the linear response to Gaussian forces is Gaussian, the full time-dependent distribution function is a Gaussian in all its variables. The form is the same as if there were no force applied, except that each variable has a mean shifted by the force. The shift is, in fact, simply calculable by ignoring the noise. The Gaussian distribution function is readily expressible in terms of the second moments (taken relative to the mean). And these are unchanged from those in the original Uhlenbeck–Ornstein article. For details see Jung (1993).

The next simplest case is that in which the frequency of the driving force is low compared to all the other rates in the problem. Then the adiabatic procedure can be applied just as when no periodic forces were present. We discuss one such example later.

The case of a high-frequency force can only be handled (by perturbation theory) if the force is weak. Since the force is periodic, the solution for the Fokker–Planck probability distribution takes the Floquet form. Perturbation theory has been applied to obtain correlation functions and/or the Green's function for the Fokker–Planck equation in terms of the corresponding solutions in the absence of a force by Presilla *et al.* (1989), Fox (1989), and Gang *et al.* (1990). For the escape problem, an extensive set of numerical results are plotted in Jung (1993). For large driving frequencies, the enhancement of escape rates, according to Jung, Eq. (8.43), decrease inversely with the square of the driving frequency. Resonance activation appears to occur. But the results are largely numerical, except for one result quoted by Jung from an unpublished thesis by Linkowitz (1989).

A simple qualitative explanation for the fairly broad stochastic resonance is that it occurs when the potential is modulated at a period T that is synchronized with the Kramers reaction time T_K. In a double-well system, the potential modulation must complete one period in a time T just large enough for a Kramers transition from one well to the other and back. See Fig. 5.

Thus we require

$$T = 2T_K,$$

where T_K is the Kramers time given in Eq. (350). This resonance not only appears in the escape rate but also appears as an enhancement in the output spectrum at the driving frequency. The signal-to-noise ratio in a bistable ring laser displays this smooth resonance in Fig. 6

8.6 Self-Sustained Oscillators

The quasilinear approximation to a nonlinear system can break down because of an instability in some coordinate. Lax (1968a) and Hempstead and Lax (1967) have argued that all autonomous self-sustained oscillators possess one degree of freedom that is neutrally unstable. These are oscillators described by a set of ordinary differential equa-

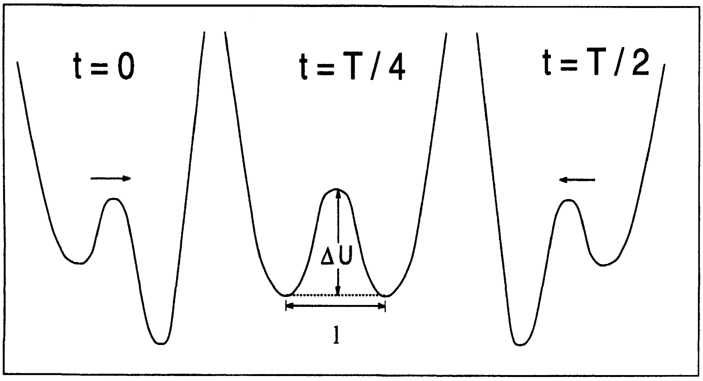

FIG. 5. Bistable potential with external modulation, shown at three times [after Jung (1993), but recomputed].

tions containing no coefficients that depend explicitly on the time. For this reason, if $a_j(t)$ is a solution (in the absence of noise), then so is $a_j(t + \tau)$ for arbitrary τ. This instability against a shift of the time origin is equivalent to an instability in a phaselike variable.

The onset of oscillation is like a phase transition characterized by a long time constant. Thus most internal degrees are relatively fast and can be adiabatically eliminated, leaving one complex (field) amplitude to describe the behavior near threshold.

A self-sustained oscillator operates by connecting an absorptive system, such as a laser cavity containing absorbing materials and windows for energy to escape, to a gain system such as a set of pumped atoms. In equivalent-circuit terms, the absorptive system has a positive resistance, and the pumped system has a negative resistance. The oscillator will stabilize at a level of operation at which these two (nonlinear) resistances cancel. The frequency of operation will be that at which the reactances cancel. Lax and Louisell (1967) found an equation for a laser of the form

$$\frac{d\beta}{dt} = -(1 - i\alpha)R(|\beta|^2)\beta + f(t) \qquad (352)$$

for the complex electromagnetic field ampli-

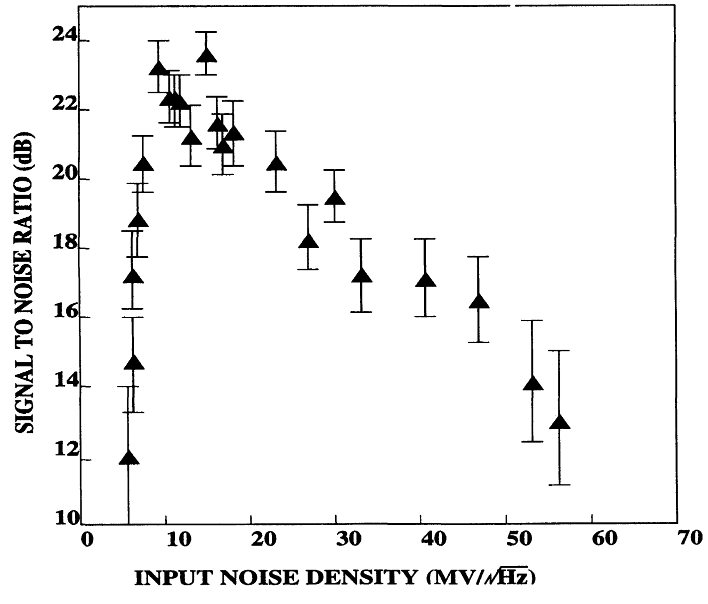

FIG. 6. Signal-to-noise ratio of one mode in a bistable laser. Taken from Jung (1993), based on experimental results of Roy and Mandel (1980) and Lett and Mandel (1985).

tude β after the rotating-wave approximation has been made.

The parameter α, which first appeared as the detuning parameter in Lax (1965) also appears in Lax (1967) as a coupling parameter between amplitude and phase fluctuations. [The parameter α of Lax (1965) is identical to $\tan\beta$ of Lax (1967). Thus in both cases, the linewidth is enhanced by a factor $1 + \alpha^2$.] It is this coupling between intensity and phase fluctuations that produces the strong enhancement of linewidth in semiconductor lasers found by Fleming and Mooradian (1981) and explained by Henry (1986). The function R describes the dependence of the net resistance on the intensity $|\beta|^2$. At the operating point, $|\beta|^2 = p$, this resistance must vanish. For any well-designed self-sustained oscillator, it was shown in Appendix B of Hempstead and Lax (1967) that it is adequate to expand the resistance to the linear term $|\beta|^2 - p$. The relative importance of the first neglected term is shown to be of the order of the ratio of the noise in the oscillator to the signal. For a well-designed oscillator, this ratio is small. In Appendix A of Hempstead and Lax, the ratio is shown to be one over the number of photons in the cavity. Once this expansion is made, it is possible to rescale both amplitude and time to obtain a universal equation that we have called the rotating-wave van der Pol with noise:

$$\frac{d\beta}{dt} = (1 - i\alpha)(p - |\beta|^2)\beta + h(t). \tag{353}$$

The rescaling in amplitude was chosen to make the coefficient of $|\beta|^2$ unity. The scaling in time was chosen to yield the simple correlation function for the c-number noise source:

$$\langle h(t)h(u)\rangle = 4\delta(t - u), \tag{354}$$

ignoring, here, the coupling between reactive and resistive effects, by setting the parameter α equal to zero.

Away from threshold, Lax (1967) carried out a quasilinear treatment of self-sustained oscillators. Near threshold, fluctuations of the order of 100% are possible, and a numerical solution of the associated Fokker–Planck equation becomes necessary. Risken and Vollmer (1967) and Hempstead and Lax (1967) carried out such calculations to obtain the linewidth associated with field-amplitude fluctuations and the linewidth associated with intensity fluctuations. Comparisons of these theoretical predictions with subsequent experimental results are shown in Fig. 7 and Fig. 8, respectively.

GLOSSARY

Brownian Motion: The limit of a random-walk process in which the individual steps become negligible. The position is continuous but not differentiable. The mean square displacement grows linearly with time. The particle density obeys the diffusion equation.

Characteristic Function: The Fourier transform of the probability density.

Chapman–Kolmogoroff Condition: A statement that conditional probabilities must obey if the process is to be Markoffian.

Fokker–Planck Process: A stochastic process that generalizes the Brownian process by having drift and diffusion coefficients that can be position dependent.

Fokker–Planck Equation: The partial differential equation, involving a single time derivative and up to two space derivatives, obeyed by a Fokker–Planck process. The independent (position) variables can be replaced by other variables—for example, spin angles in the description of spin diffusion.

Generalized Fokker–Planck Equation: A Fokker–Planck equation that may contain derivatives to all orders in the independent (position) variables.

Langevin Equation: An equation for a random variable in which the randomness is generated by a random ("force") term added usually on the right-hand side. The noise source is typically delta correlated in time (white noise) but nondelta correlation (colored noise) is permitted.

Markoff Process: A random process in which the probability of future events depends on present information and no prior information.

Noise Spectrum: The power as a function of frequency obtained by passing the signal through a narrow filter, squaring the output, and averaging it.

Shot Noise: The input noise assumed to be associated with a series of Dirac delta-function pulses arriving at random at a uni-

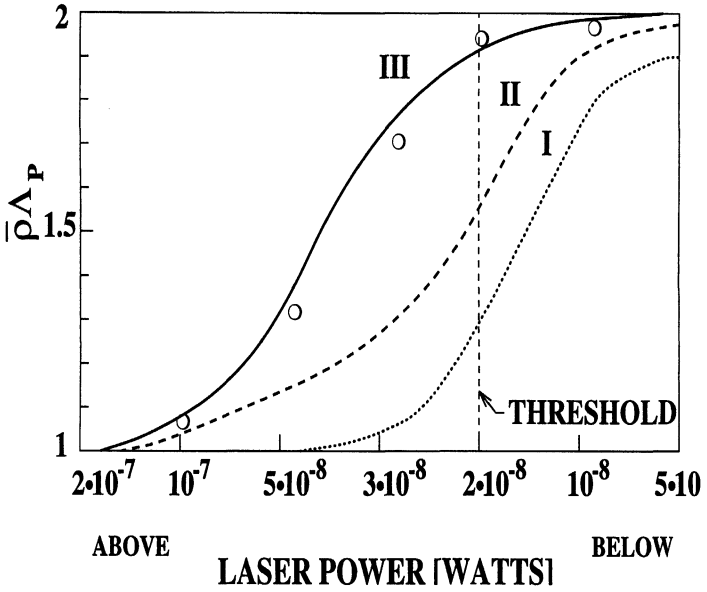

FIG. 7. Spectrum for phase and amplitude fluctuations of a laser is Lorentzian to a good approximation. The dimensionless half-width Λ_p times the dimensionless mean power $\bar{\rho} \equiv \langle|\beta|^2\rangle$ is plotted against laser power. The experimental results are those of Gerhardt *et al.* (1972). The theoretical curves I and II are due to Grossman and Richter (1971), and curve III is due to Risken and Vollmer (1967) and Hempstead and Lax (1967).

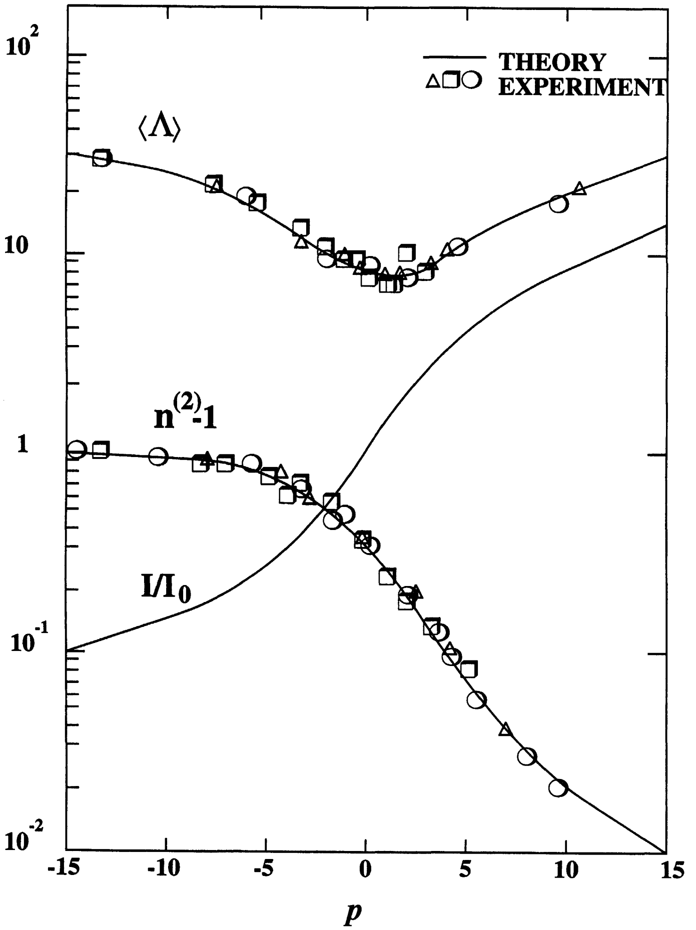

FIG. 8. Experimental verification of the Van der Pol model for lasers operating near threshold. The theoretical curves are from Hempstead and Lax (1967). The experimental points △, □, and ○ are from Gamo *et al.* (1968), Davidson and Mandel (1967), and Arecchi *et al.* (1967a, 1967b), respectively. Lower-case p is the pump parameter of the model, $\langle\Lambda\rangle$ is the effective linewidth of intensity fluctuations, and $n^{(2)}$ is the normalized second factorial moment $\langle n(n-1)\rangle/\langle n\rangle^2$ of the photocount distribution $p(n,T)$ as $T \to 0$. In this limit, $n^{(2)} - 1 = \langle(\Delta\rho)^2\rangle/\langle\rho\rangle^2$, and I/I_0 is the light intensity normalized to threshold value ($p = 0$); i.e., $I/I_0 = \langle\rho\rangle/\langle\rho\rangle_{th} = \langle\rho\rangle/1.128$. Here ρ is an abbreviation for $|\beta|^2$.

form average rate. The term is also used to describe the response a physical system to such input white noise.

Wiener–Khinchin Theorem: Equality between the noise spectrum and the Fourier transform of the autocorrelation function.

Works Cited

Arecchi, F. T., Giglio, M., Sona, A. (1967a), "Dynamics of the laser radiation at threshold," *Phys. Lett.* **A25,** 341–342.

Arecchi, F. T., Rodari, G. S., Sona, A. (1967b), "Statistics of the laser radiation at threshold," *Phys. Lett.* **A25,** 59–60.

Bellman, R. (1968), *Some Vistas of Modern Mathematics; Dynamic Programming, Invariant Imbedding, and the Mathematical Biosciences,* Univ. of Kentucky Press.

Black, F., Scholes, M. (1973), "The pricing of options and corporate liabilities," *J. Polit. Econ.* **81,** 637–654.

Bouchard, Jean-Philippe, Sornette, Didier (1994), "The Black-Scholes option pricing problem in mathematical finance: generalization and extensions for a large class of stochastic processes," *J. Phys. I (Paris)* **4,** 863–881.

Brinkman, H. C. (1956), "Brownian motion in a field of force and the diffusion theory of chemical reactions," *Physica* **22,** 149–155.

Brown, R. (1828), *Philos. Mag.* **4,** 161; *Ann. Phys. Chem.* **14,** 294.

Bulsara, A., Hänggi, P., Marchesoni, F., Moss, F., Shlesinger, M. (Eds.) (1993), Proceedings of the NATO Advanced Research Workshop, *J. Stat. Phys.* **70,** 1–512.

Callen, H. B., Welton, T. A. (1951), "Irreversibility and generalized noise," *Phys. Rev.* **83,** 34–40.

Callen, Herbert B. (1985), *Thermodynamics,* New York: Wiley.

Campbell, N. (1909), "The Study of discontinuous phenomena," *Proc. Camb. Philos. Soc.* **15,** 117–136.

Casti, J. L., Kalaba, R. E. (1973), *Imbedding Methods in Applied Mathematics,* Reading, MA: Addison-Wesley.

Davidson, F., Mandel, L. (1967), "Correlation measurements of laser beam fluctuations near threshold," *Phys. Lett.* **A25,** 700–701.

Deutsch, Ralph (1962), *Nonlinear Transformations of Random Processes,* Englewood Cliffs, NJ: Prentice Hall.

Doob, J. L. (1953), *Stochastic Processes,* New York: Wiley, Chap. VI, Eqs. (3.1) and (3.4), and Chap. IX, Eq. (2.6).

Einstein, A. (1906), "On the theory of the brownian movement," *Ann. Phys. (Leipzig)* **19,** 371–381.

Einstein, Albert (1905), "On the movement of small particles suspended in a stationary liquid demanded by the molecular-kinetic theory of heat," *Ann. Phys. (Leipzig)* **17,** 549–560. See the English translation of this and four later, related articles in Einstein, A. (1956), *Investigations on the Theory of the Brownian Movement,* New York: Dover.

Fleming, M. W., Mooradian, A. (1981), "Fundamental line broadening of single-mode (GaAl)As diode lasers," *Appl. Phys. Lett.* **38,** 511–513.

Fox, R. F. (1989), "Stochastic resonance in a double well," *Phys. Rev. A* **39,** 4148–4153.

Gamo, H., Grace, R. E., Walter, T. J. (1968), "Statistical analysis of intensity fluctuations in single mode laser radiation near the oscillation threshold," *IEEE J. Quantum Electron.* **QE-4,** 344.

Gang, H., Nicolis, G., Nicolis, C. (1990), "Periodically forced Fokker–Planck equation and stochastic resonance," *Phys. Rev. A* **42,** 2030–2041.

Gerhardt, H., Welling, H., Güttner, A. (1972), "Measurements of the laser line width due to quantum phase and quantum amplitude noise above and below threshold," *Z. Phys.* **253,** 113–126.

Ghashghaie, S., Breymann, W., Peinke, J., Talkner, P., Dodge, Y. (1996), "Turbulent cascades in foreign exchange markets," *Nature* **381,** 767–770.

Gouy, M. (1888), *J. Phys. (Paris)* **7,** 561.

Gröbner, W., Hofreiter, N. (1950), *Integraltafel, Zweiter Teil, Bestimmte Integrale,* Wien: Springer.

Grossman, S., Richter, P. H. (1971), "Laser threshold and nonlinear Landau fluctuation theory of phase transitions," *Z. Phys.* **242,** 458–475.

Hänggi, P., Talkner, P., Borkovec, M. (1990), "Reaction-rate theory: fifty years after Kramers," *Rev. Mod. Phys.* **62,** 251–341.

Hänggi, Peter (1989), "Colored noise in continuous dynamical systems: a functional calculus approach," in: F. Moss, P. V. E. McClintock (Eds.), *Noise in Dynamical Nonlinear Dynamical Systems,* Vol. 1, *Theory of continuous Fokker-Planck Systems,* Cambridge, U.K.: Cambridge Univ. Press.

Haken, H., Weidlich, W. (1967), "A theorem on the calculation of multi-time-correlation functions by the single-time density matrix," *Z. Phys.* **205,** 96–102.

Haynes, J. R., Hornbeck, J. A. (1955), "Trapping of minority carriers in Si.II.n type," *Phys. Rev.* **100,** 606–615.

Hempstead, R., Lax, M. (1967), "Classical noise VI. Noise in self-sustained oscillators near threshold," *Phys. Rev.* **161,** 350–366.

Henry, C. H. (1986), "Phase noise in semiconductor lasers," *IEEE J. Lightwave Technol.* **LT-4,** 298–311.

Hill, J. E., van Vliet, K. M. (1958), "Ambipolar transport of carrier density fluctuations in germanium," *Physica* **24,** 709–720.

Hornbeck, J. A., Haynes, J. R. (1955), "Trapping of minority carriers in Si.I.p type," *Phys. Rev.* **97,** 311–321.

Hull, John (1992), *Options, Futures and Other*

Derivative Securities, Englewood Cliffs: Prentice Hill.

Ito, K. (1951), "On stochastic differential equations," *Am. Math. Soc. Memoirs* No. 4, 1–51.

Ito, K., McKean, H. P. (1965), *Diffusion Processes and Their Sample Paths,* Grundlehren der mathematischen Wissenschaften, Vol. 125, Berlin: Springer.

Jeffries, Harold, Jeffreys, Bertha Swirles (1950), *Methods of Mathematical Physics,* Cambridge, U.K., Cambridge University Press, Sec 1.10.

Johnson, J. B. (1928), "Thermal agitation of electricity in conductors," *Phys. Rev.* **32,** 97–109.

Jung, Peter (1993), "Periodically driven stochastic systems," *Phys. Rept.* **234,** 175–295.

Kendall, M. G., Stuart, A. (1969), *The Advanced Theory of Statistics,* Vol. I, 3rd ed., New York: Hafner.

Khinchin, A. (1934), "Korrelationstheorie der stationären stochastischen Prozesse," *Math. Ann.* **109,** 604–615.

Kittel, C. (1958), *Elementary Statistical Physics,* New York: Wiley.

Kramers, H. A. (1940), "Brownian motian in a field of force and the diffusion model of chemical reactions," *Physica* **7,** 284–304.

Landauer, R. (1989), "Noise-activated escape from metastable states: an historical review," in: F. Moss, P. V. E. McClintock (Eds.) *Noise in Nonlinear Dynamical Systems,* Vol. 1, *Theory of Continuous Fokker–Planck Systems,* Cambridge, U.K.: Cambridge Univ. Press.

Landauer, R., Swanson, J. A. (1961), *Phys. Rev.* **121,** 1668–1674.

Langer, J. (1968), "Theory of nucleation rates," *Phys. Rev. Lett.* **21,** 973–976.

Lax, M., Zwanziger, M. (1973), "Exact photocount statistics: lasers near threshold," *Phys. Rev. A* **7,** 750–771.

Lax, M. (1960), "Fluctuations from the nonequilibrium steady state," *Rev. Mod. Phys.* **32,** 25–64.

Lax, M. (1963), "Formal theory of quantum fluctuations from a driven state," *Phys. Rev.* **129,** 2342–2348.

Lax, M. (1965), "Quantum noise V: phase noise in a homogeneously broadened maser," in: P. L. Kelley, B. Lax, P. E. Tannenwald (Eds.), *Physics of Quantum Electronics,* New York: McGraw-Hill.

Lax, M. (1966a), "Classical noise III: nonlinear markoff processes," *Rev. Mod. Phys.* **38,** 359–379.

Lax, M. (1966b), "Classical noise IV: Langevin methods," *Rev. Mod. Phys.* **38,** 541–566.

Lax, M. (1967), "Classical noise V: noise in self-sustained oscillators," *Phys. Rev.* **160,** 290–307.

Lax, M. (1968a), "Fluctuations and coherence phenomena in classical and quantum physics," in: M. Chretien, E. P. Gross, S. Deser (Eds.), *Statistical Physics,* 1966 Brandeis Summer Lecture Series, Vol. 2, New York: Gordon and Breach Science Publishers, pp. 270–478.

Lax, M. (1968b), "Quantum noise XI. multitime correspondence between quantum and classical stochastic processes," *Phys. Rev.* **172,** 350–361.

Lax, M. (1991), "The theory of laser noise," 1990 Conference on Laser Science and Optics Applications—Boston, MA, *Proc. SPIE* **1376,** 2–20.

Lax, M., Louisell, W. H. (1967), "Quantum noise IX: quantum Fokker–Planck solution for laser noise," *J. Quantum Electronics* **QE-3,** 47–58.

Lax, M., Mengert, P. (1960), "Influence of trapping, diffusion and recombination on carrier concentration fluctuations," *J. Phys. Chem. Solids* **14,** 248–267.

Lax, M., Phillips, J. C. (1958), "One dimensional impurity bands," *Phys. Rev.* **110,** 41–49.

Lett, P., Mandel, L. (1985), "Investigation of time-dependent correlation properties of the bidirectional dye ring laser," *J. Opt. Soc. Am. B* **2,** 1615–1625.

Linkowitz, S. (1989), Ph.D. Thesis.

Middleton, D. (1960), *Statistical Communications Theory,* New York: McGraw-Hill.

Moss, F., McClintock, P. V. E. (1989), *Noise in Nonlinear Dynamical Systems,* Vol. I, *Theory of continuous Fokker–Planck Systems;* Vol. II, *Theory of Noise Induced Processes in Special Applications;* Vol. III, *Experiments and Simulations,* Cambridge, U.K.: Cambridge Univ. Press.

Moullin, E. B. (1938), *Spontaneous Fluctuations of Voltage,* Oxford: Clarendon Press.

Moyal, J. E. (1949), "Quantum mechanics as a statistical theory," *Proc. Camb. Philos. Soc.* **45,** 99–124.

Nyquist, H. (1927), *Phys. Rev.* **29,** 614.

Nyquist, H. (1928), "Thermal agitation of electric charge in conductors," *Phys. Rev.* **32,** 110–113.

Onsager, L. (1931), "Reciprocal relations in irreversible processes," Part 1, *Phys. Rev.* **37,** 405–426; Part 2, **38,** 2265–2679.

Peters, Edgar, E. (1994), *Fractal Market Analysis,* New York: Wiley.

Pontryagin, L., Andronov, A., Vitt, A. (1933), "On the statistical treatments of dynamical systems," *Zh. Eksp. Teor. Fiz.* **3,** 165–180. An English translation appears on pp. 329–348 of F. Moss, P. V. E. McClintock (Eds.) (1989), *Noise in Nonlinear Dynamical Systems,* Vol. 3, *Experiments and Simulations,* Cambridge, U.K.: Cambridge Univ. Press.

Presilla, C., Marchesoni, F., Gammaitoni, L. (1989), "Random Time-Modulated Bistable Systems: Non-stationary Statistical Properties," *Phys. Rev. A* **40,** 2105–2113.

Rice, S. O. (1944), "Mathematical analysis of random noise, part I," *Bell Syst. Technol. J.* **23,** 282–332.

Rice, S. O. (1945), "Mathematical analysis of random noise, part II," *Bell Syst. Technol. J.* **24,** 46–156.

Rice, S. O. (1948), "Mathematical analysis of random noise, part III," *Noise Through Nonlinear Devices* **27,** 109–157.

Risken, H., Vollmer, H. D. (1967), "Correlation function of the amplitude and of the intensity fluctuation near threshold," *Z. Phys.* **191,** 301–312.

Roy, R., Mandel, L. (1980), "Optical bistability and first order phase transition in a ring dye laser," *Opt. Commun.* **34,** 133–136.

Scher, H., Lax, M. (1973), "Stochastic transport in a disordered solid: I and II," *Phys. Rev. B* **7,** 4491–4502, 4502–4591.

Shockley, W. (1950), *Electrons and Holes in Semiconductors,* Princeton, NJ: Van Nostrand.

Stratonovich, R. L. (1963), *Topics in the Theory of Random Noise,* Vol. I, New York: Gordon and Breach. See also Vol. II (1967).

Stratonovich, R. L. (1989), "Some Markoff methods in the theory of stochastic processes in nonlinear dynamical systems," in: F. Moss, P. V. E. McClintock (Eds.), *Noise in Nonlinear Dynamical Systems,* Vol. I, Cambridge, U.K., Cambridge Univ. Press.

Talkner, Peter, Hänggi, Peter (1995), *New Trends in Kramers' Reaction Rate Theory,* Dordrecht/Boston/London: Kluwer Academic Publishers.

Thiele, T. N. (1903), "Theory of observations," *Ann. Math. Stat.* **2,** 165.

Transistor Teachers' Summer School (1953), "Experimental verification of the relation between diffusion constant and mobility of electrons and holes," *Phys. Rev.* **88,** 1368–1369.

Uhlenbeck, G. E., Ornstein, L. S. (1930), "On the theory of the Brownian motion," *Phys. Rev.* **36,** 823–841.

Van Kampen, N. (1985), "Elimination of fast variables," *Phys. Rep.* **124,** 69–160.

Van Kampen, N. (1992), *Stochastic Processes in Physics and Chemistry,* Amsterdam: North Holland.

Van Roosbroeck, W. (1953), "Transport of added current carriers in a homogeneous semiconductors," *Phys. Rev.* **91,** 282–289.

Wiener, N. (1930), "Generalized harmonic analysis," *Acta Math.* **55,** 117–258.

Wigner, E. (1932), "Quantum corrections for thermodynamic equilibrium," *Phys. Rev.* **40,** 749–760.

Williams, E. C. (1937), "Thermal fluctuations in complex networks," *J. Electrical Eng.* **81,** 751.

Further Reading

Bell, D. A. (1960), *Electrical Noise,* Princeton: Van Nostrand.

Burgess, R. E. (1965), *Fluctuation Phenomena in Solids,* New York: Academic.

Chandrasekhar, S. (1943), "Stochastic problems in physics and astronomy," *Rev. Mod. Phys.* **15,** 1–89.

Feller, William (1971), *An Introduction to Probability Theory and Its Applications,* Vol. I, New York: Wiley.

Gillespie, Daniel T. (1996), "The mathematics of Brownian motion and Johnson noise," *Am. J. Phys.* **64,** 225–240.

Hamming, Richard W. (1991), *The Art of Probability for Scientists and Engineers,* Reading, MA: Addison Wesley.

Lawson, J. L., Uhlenbeck, G. E. (1950), *Threshold Signals,* M.I.T. Radiation Lab Series, Vol. 24, New York: McGraw Hill, Dover.

Lax, M. (1964), "Quantum relaxation, the shape of lattice absorption and inelastic neutron scattering lines," *J. Phys. Chem. Solids* **25,** 467–503.

Lax, M. (1966), "Quantum Noise IV: Quantum Theory of Noise Sources," *Phys. Rev.* **145,** 110–129.

Lax, M. (1974), *Symmetry Principles in Solid State and Molecular Physics,* New York: Wiley.

Louisell, W. H. (1973), *Quantum Statistical Properties of Radiation,* New York: Wiley, Appendix A.

Lukacs, E. (1960), *Characteristic Functions,* London: Griffith.

MacDonald, D. K. C. (1962), *Noise and Fluctuations,* New York: Wiley.

Middleton, David (1960), *Introduction to Statistical Communication Theory,* New York: McGraw-Hill.

Moullin, E. B. (1938), *Spontaneous Fluctuations of Voltage,* Oxford: Clarendon.

Robinson, F. N. H. (1962), *Noise in Electrical Circuits,* Oxford: Clarendon.

Robinson, F. N. H. (1974), *Noise and Fluctuations in Electronic Devices and Circuits,* Oxford: Clarendon Press.

Valley, George E., Jr., Wallman, Henry (1948), *Vacuum Tube Amplifiers,* M.I.T. Radiation Lab Series, Vol. 13, New York: McGraw-Hill.

Wang, Ming Chen, Uhlenbeck, G. E. (1945), "On the theory of Brownian motion II," *Rev. Modern Phys.* **17,** 323–342.

STORAGE OF ENERGY

Osman K. Mawardi, *Collaborative Planners, Inc., Cleveland Heights, Ohio, U.S.A.*

Introduction 61
1. **Classification of Energy-Storage Systems** 61
1.1 Energy Storage with Slow Energy Transfer 63
1.2 Pulsed Energy-Storage Systems 63
2. **Mechanical Energy-Storage Systems** 63
2.1 Flywheel Design 64
2.1.1 Disk Flywheels 64
2.1.2 Rim Flywheels 65
2.2 Inertial Storage and Electric Machines 66
2.2.1 Pulsed Homopolar Generators 66
2.2.2 Design of Pulsed Homopolar Generators 68
2.2.3 Operational Experience 68
3. **Thermal Energy Storage** 69
3.1 Thermal-Storage Processes 70
3.2 Heat Exchangers and Thermal-Storage Design 70
4. **Compressed-Air Storage** 71
4.1 Principle of Operation 71
4.2 Operational Experience 72
5. **Pumped-Hydro Storage** 73
6. **Chemical Energy Storage** 73
6.1 Hydrogen-Cycle System 73
6.1.1 Hydrogen Production 73
6.1.2 Storage of Hydrogen 75
6.2 Electrical Batteries 75
7. **Magnetic Storage** 75
7.1 Components of SMES 76
7.1.1 The Superconductors 76
7.1.2 The Coil and Its Accessories 77
7.1.3 The Graetz Circuit 77
7.1.4 Conclusion 77
7.1.5 SMES and Transmission Stabilization 80
7.2 Pulsed Inductive Storage 80
Glossary 81
Works Cited 82

INTRODUCTION

The industrial nations of the world have become conscious of their energy requirements. The spiraling escalation of the cost of fuel and of capital investment in new power plants has instigated an intensive effort to identify more efficient methods of processing energy. A study sponsored by the U.S. National Academy of Sciences (NAS, 1976) has indicated that if economically efficient storage devices can be developed, as much as 10% of the installed generating capacity required by the electric utilities can be displaced by less costly storage systems. The size of these storage units would range from 15 MW to several thousand megawatts with a storage duration from 2 h to 2 d.

The subject matter of this article is concerned with a review of the salient features of storage systems of the scale and size mentioned in the National Academy report. As will be seen later in the discussion, large-scale storage devices display unique characteristics that give rise to complex technical issues. Another family of storage systems that are covered in the present article comprises systems used with sources capable of delivering energy pulses of great intensity and short durations. The increasing number of practical applications that require such sources has warranted the inclusion of such devices in the list of storage systems reviewed here.

1. CLASSIFICATION OF ENERGY-STORAGE SYSTEMS

Energy can be stored in many ways: in electrostatic form as in a charged capacitor, as magnetic energy in an inductor, as kinetic energy in a flywheel, or as chemical energy in a battery. It is possible to store energy in

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

a compressed gas, in hydrogen, or in an explodable liner of a vessel. The extraction of energy in a controlled fashion from the latter systems is no longer a simple matter. Special machinery or devices are needed to execute the appropriate energy transfer from the store to the load.

As a result of advances in materials science, new compounds with unusual physical and chemical properties have been produced. These innovations have led to the invention of several novel energy stores with impressive figures of cost, energy stored per unit weight, and efficiency of energy transfer.

A question of paramount importance to the design of an energy system is the selection of the energy storage best suited to the specific application. Now there is no single general law capable of categorizing all energy-storage devices. This is because of the vast conceptual difference in the governing principles for the mode of operation of these devices.

The task of selecting suitable storage may be facilitated by examining its intended function. It is noticed, for instance, that most storage devices that have been used fall in one of three major groups of classes that are defined by the temporal behavior of the transfer of energy from the storage to the load.

An example of the first group is the energy storage systems used in electric power grids. In such systems, the storage devices are distributed over several substations and are installed for the purpose of accepting energy from base-load generators when the load demand falls below the base-load generating capacity. This stored energy is then released to the power grid when peak load conditions demand it. This transfer of energy takes place *slowly.* In this manner the electric network can deliver power to a peak load that exceeds the installed capacity. The addition of storage thus effectively improves the capability and reliability of the service of the electric grid. Analogs to the storage systems used in the power grid are the flywheel storage systems installed in electric cars and trains.

A subclass of this group of applications comprises the so-called uninterrupted power systems (UPS). In these systems, it is required to maintain a continuous supply of power whenever the primary source of power is interrupted. Typical illustrations for this class are solar-energy power plants (land based or in space) and power supplies for computers. It is obvious that unlike the case of the electric power grid, the capacity of the primary source of power must be greater than that of the power consumption in the load.

In the second group the storage devices are used in *oscillatory* systems. The frequency of oscillations could be in the microwave range or very low. An example of the latter case is the use of a superconducting inductive energy-storage system to damp out the power oscillations in an ac intertie transmission line.

The third and final group belongs to a growing number of applications that require *bursts* of nonoscillatory power of short duration but of great intensity. For pulses between 1 and 100 ms long, as needed in nuclear fusion research, there is an advantage in using electromechanical storage devices such as pulsed homopolar machines or explodable liners. For shorter durations as required for the operation of lasers, radar, plasma guns, etc., the pulse generator consists of a stationary capacitive or inductive energy-storage device, sometimes combined with a pulse-shaping network.

Other factors that influence the selection procedure of a specific device are the cost, the level of energy to be stored, and the weight and the size of the storage. For the case of pulsed sources, additional consideration is given to the duty cycle and rise time of the energy delivered to the load.

Energy storage is too vast a subject to do it justice in one article. Accordingly, the discussion has been narrowed down to a few selected systems of great practical importance. This article has been divided into seven sections that deal with the features, characteristics, and performance of mechanical, thermal, electrical, magnetic, and chemical devices used in energy storage.

The transfer of energy from a storage system to the load is not the central topic covered here. A discussion of a general nature on the features of storage systems with slow and impulsive energy transfer will be given first. These comments will be helpful to the understanding of the operation of some of the storage schemes considered here.

1.1 Energy Storage with Slow Energy Transfer

Fernandes (1976) suggested a convenient method of classification of energy-storage systems used in the electric utility industry. These typify the storage systems with slow transfer time. He identified two categories of storage systems.

The first is the single-input shaving storage system that is charged with off-peak energy. Examples of the single-input systems are batteries, flywheels, pumped hydro storage, and superconducting magnetic storage. Analysis of the off-peak energy requirement in the electric system indicates that a storage system with a capacity between 6% and 14% of that of the grid can handle most of the off-peak demand. Interestingly enough, this also applies to a storage system for electric cars or trains.

The second category is a dual system that generates electric power from a charge cycle obtained with off-peak electricity as well as from a primary or secondary fuel supply. Examples of such systems are hydrogen storage systems and compressed-air/natural gas storage. Dual systems are specially useful when large (18% to 20%) reserve capacity margin is needed.

1.2 Pulsed Energy-Storage Systems

Carruthers (1962) recognized the economic advantage of using inductors over capacitors when the energy stored exceeds 100 kJ. Unfortunately, in pulsed applications the load is inductive. Now a generalized equation governing the impulsive transfer of energy from charged inductors to another inductor can be obtained from a theorem due to Carnot (Sturm, 1861). It can be shown (Mawardi, 1982) that

$$E_1 - E_2 = \sum \tfrac{1}{2} L_1 (I_{i,1} - I_{i,2})^2, \tag{1}$$

where E_1 is the energy stored in the system before the transfer, E_2 the energy in the system after the transfer, L_i the inductance of the ith element, $I_{i,1}$ the current in L_i before the transfer, and $I_{i,2}$ the current after the transfer. An important consequence of this theorem, which was further investigated by Lord Kelvin (1863), is that no matter how clever the scheme, the transfer of energy from a purely inductive system to another inductance always will be accompanied by a loss of energy.

To remove the limitations imposed by Carnot's theorem, a number of methods have been developed, which introduce a capacitor in the circuit and trigger the transfer of energy by opening or closing a switch. Another option has been to initiate the transfer of energy from charged inductors to an inductive load by altering the shape of the storage inductors (Fowler *et al.*, 1962) or their relative position (Smith and Lewin, 1970) without interrupting the current. Either one of these two options will permit a lossless transfer of energy.

It is interesting to note that in a pulsed electric rotating machine the inertia of the rotor does act as the dynamical analog of a capacitor. Hence in such a system the energy transfer is in accordance with the first procedure mentioned in the previous paragraph. On the other hand, for the case of an imploding metallic liner embracing a magnetic field, the resulting flux compression that leads to the transfer of energy operates along the lines of the second option described above.

2. MECHANICAL ENERGY-STORAGE SYSTEMS

Energy in mechanical form can be stored as kinetic or as potential energy. The most common mechanical potential-energy storage devices are the springs. The state of the art of springs has been fairly stable for decades, and for this reason this will not be discussed here. The interested reader is referred to several classic texts on this subject (Timoshenko, 1951). On the other hand, there have been several impressive innovations in mechanical kinetic-energy–storage devices, specially in the rotational devices, viz., flywheels.

Flywheels have been used in a variety of applications that fall in the classification schemes discussed in Sec. 1. For example, a flywheel when installed on the shaft of an internal-combustion engine helps to maintain an even average angular torque delivered by the engine. Another example is the drive of steel rolling mills: the Ward–Leonard system (Thaler and Wilcox, 1966), which consists of a combination of a flywheel and an electric motor. The flywheel acts as an energy accumulator during the no-load cycle of the drive. In its discharge mode, the flywheel conspires with the electric motor to overcome the initial surge of starting the process of forcing the white-hot steel ingot through the rollers.

2.1 Flywheel Design

Modern flywheel "accumulators" are incorporated nowadays in systems considerably more complex than the two examples given above. Flywheels of advanced design use rotors of composite materials, and they spin much faster. Furthermore, since the flywheels often are enclosed in burst-containment casings (Coppa and Zweben, 1980), they can operate in a low-pressure environment, which reduces the windage losses (Barlow, 1980). An additional feature of very high-speed rotors is the introduction of magnetic bearings (Pourbeau, 1980), which further reduce the friction losses.

A figure of merit for the performance of a flywheel is its specific stored energy, which is defined as the energy stored per unit mass. This figure of merit can be related to the maximum energy that can be stored in the flywheel or, more specifically, to the bursting elastic stress in the rotor. It can be shown (Genta, 1985) that

$$E/M = K(\sigma_c)_{\max}/\rho, \tag{2}$$

where E is the energy stored in the rotor of mass M, K is a constant related to the rotor shape, $(\sigma_c)_{\max}$ is the maximum allowable stress, and ρ is the density of the rotor material.

It is helpful to the analysis of flywheel dynamics to render ω, the angular speed of the rotor, dimensionless by referring it to ω_0, the angular speed of a thin rim that has the same outer diameter and density and is equally stressed. The dimensionless speed is thus

$$\xi = \omega/\omega_0. \tag{3}$$

The evaluation of K and ξ require laborious calculations because solutions in closed forms are not available for rotors of practical shapes. The stress analysis must consequently be performed by means of numerical computations.

2.1.1 Disk Flywheels.

Disk flywheels have a most advantageous disposition of mass and consequently display a high specific energy. Values of 50 W·h/kg are common. For this reason they are of special interest to energy-storage systems. But disk flywheels must operate at high speeds to be used to their full advantage.

When the thickness of the disk is small compared to its diameter and its thickness profile is a monotonically decreasing function of the radius, then it is reasonable to approximate the stress field by a plane stress. An intriguing concept due to de Laval (Stodola, 1934) is to consider constant stress profiles. De Laval's hypothesis implies that the radial stress is equal to the circumferential stress everywhere in the rotor. Under these conditions it is found that (Genta, 1985)

$$\xi = \sqrt{2F}, \tag{4}$$

where $F = \rho(\omega r_0)^2/(2\sigma)$ in which σ is the constant stress in the disk. The de Laval profile is too thin at the rim. To be practical an outer rim has to be added that will exert a radial stress equal to σ. An example of such a configuration is shown in Fig. 1, where the rim has been selected to have a constant width.

The K and ξ factors for a constant-stress flywheel are illustrated in graphical form in Fig. 2. The dependence of K on the width of the rim is indicated on the same figure by α, defined by

$$\alpha = (r_0 - r_D)/h_0; \tag{5}$$

the meaning of the radii r_0 and r_D and the thickness h_0 are given in Fig. 1.

Constant-stress disks should not have a central hole, as it would cause a departure from the plane stress distribution, and the local strength of the stress can increase beyond the design value for the maximum stress in the disk. A row of equally spaced holes makes matters worse. It was estimated by Johnson (1956) that the stress intensification can exceed a factor of 2. The disk should not be attached to a flange with

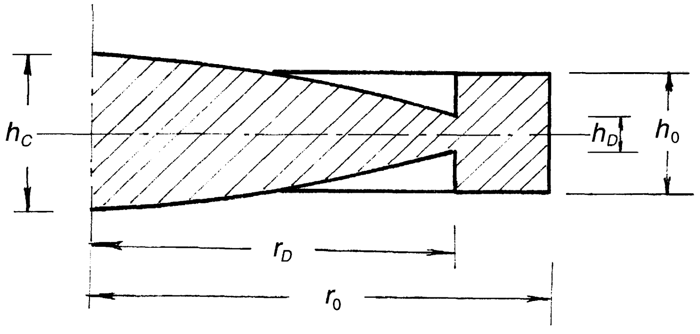

FIG. 1. Profile of disk flywheel with constant radial stress. The rim has a uniform thickness h_0 and width $r_0 - r_D$. The maximum thickness of the flywheel at its axis of rotation is h_C. (After Genta, 1985.)

equally spaced bolts. Instead, it is better to use a short stub shaft or flange made as an integral part of the disk.

2.1.2 Rim Flywheels. There is a range of applications that require rotating power delivered from an energy storage at relatively low angular speeds. Rim flywheels are ideal devices for this type of load. In these flywheels the major fraction of the mass is in the rim. Spokes attached to the hub of the wheel are used to support the rim. To optimize the energy stored in the flywheel, its radius is made as large as possible. The limits on the size and the peripheral velocity of the rotor are set by the stresses in the wheel. Because of the difference in the radial displacements of the rim and of the spokes under the influence of the centrifugal forces, the rim will sense bending moments. These bending moments are highest in the regions where the spokes are attached to the rim. The corresponding stresses from the bending moments can be quite severe and will be the factors determining the limiting speed of rotation of the flywheel.

Various configurations of commonly used rim flywheels are shown in Fig. 3. These have been compiled by Oberg and Jones (1976). The maximum safe speed that can be reached for the rim with such construction and with stacked steel plates has been 160 m/s. Higher speeds have been reported by investigators using rims made of Kevlar (Rabenhorst, 1979).

A considerable amount of work has been done in adapting composite materials (*q.v.*) to rim flywheels. The search for improved specific energy densities has been instigated by the adaption of these flywheels to electric cars and more recently to the uninterrupted-power-supply market for computers. Some of the materials investigated have been fiberglass–epoxy (Boutman and Krock, 1974), graphite–epoxy (Loud, 1977), and quasi-iso-

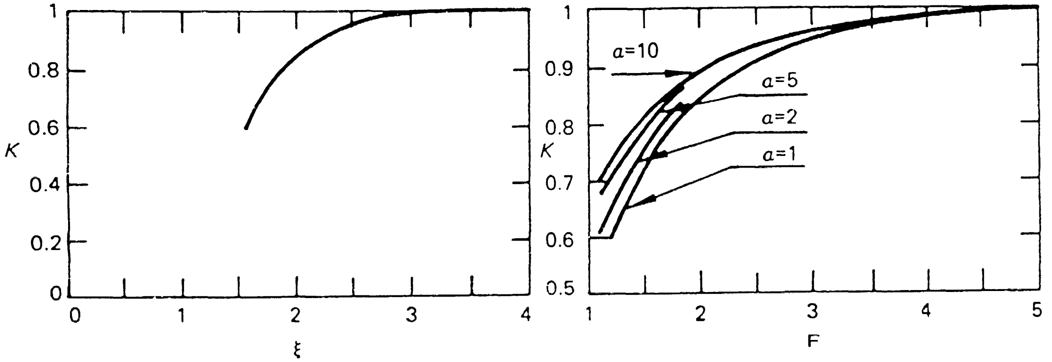

FIG. 2. The graph on the left shows K, the rotor shape factor, vs. ξ, a normalized angular rotational speed; that on the right, K vs. F, the normalized peripheral force at the rim, for different values of the aspect ratio α (after Genta, 1985).

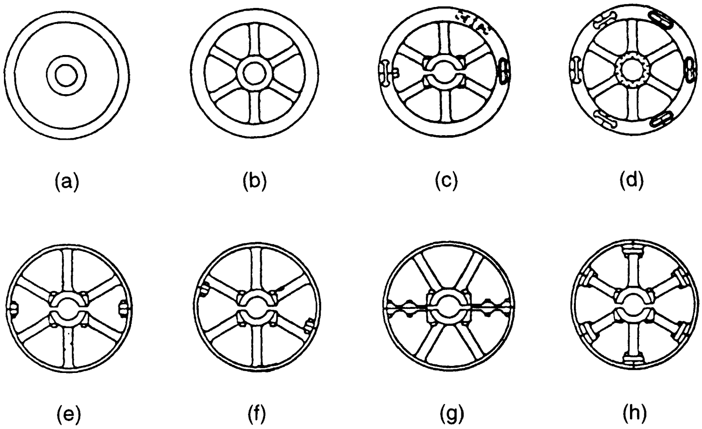

FIG. 3. Schematic representation of some rim flywheels. (a), (b) Solid cast wheels; (c), (d) rim in halves or in sections; (e), (f) rims with bolted flanges; (g) a wheel in two halves; (h) a wheel with joints. (After Oberg and Jones, 1976.)

tropic laminate disks (Sapowith and Hardy, 1981).

Several flywheel designs with composite materials have exploited the high tensile strength of these synthetic fibers and assumed that the stresses are mainly circumferential. In rim flywheels with relatively thin rims, the stresses indeed are chiefly hoop stresses. But the radial stresses in the rim, small as they are by comparison to the hoop stresses, still can induce failure of the rim by forcing the layer of fibers to split apart.

A quantitative estimate of the maximum allowable stress of an isotropic composite material when used in a rim flywheel has been evaluated (Lupton *et al.*, 1976). As expected, the radius of the flywheel and its angular speed of rotation affect the onset of failure. The net effect of the isotropy of the medium is to lower the shape factor K (which defines the figure of merit) to well below unity. For a fiberglass–epoxy combination, K rarely exceeds 0.6.

2.2 Inertial Storage and Electric Machines

The extraction of stored energy in electrical form from flywheels is not an easy task. This is because the flywheel in its discharge mode is in a decelerating mode and continuously losing speed. Consequently, if an electric dc generator is attached to the shaft of the flywheel and energy is extracted *slowly,* means must be provided to continuously adjust the field excitation of the generator to maintain a constant output voltage from the generator. From practical considerations no more than 10% of the energy stored in the flywheel can be extracted, as otherwise it would be difficult for the control circuit to track the drooping voltage output of the generator. Should a *steady* ac electric output be required, then an intermediate dc/ac converter has to be added to the dc generator. This approach has been proposed for uninterrupted power-supply systems with the flywheel storage system taking the place of a bank of electrical batteries.

A different situation is encountered when the flywheel storage system is used in a pulsed mode and the transfer of energy is to take place through an electric generator as before. Since the time of discharge is rapid, the electrical conductors in the generator are subjected to such large mechanical forces that they could be pulled out of the slots of the machine unless special reinforcement is provided. Furthermore, these same conductors must be braced also for the centrifugal forces due to rotation. The combination of these two factors will lead to an increase of the mass and cost of the system. On the other hand, if the flywheel is replaced with the rotor of a homopolar generator, one obtains a rugged machine of simple construction and specially suited to the delivery of pulsed power of great magnitude.

2.2.1 Pulsed Homopolar Generators. The rotor of a homopolar machine consists of a solid axially symmetric conductor rotating about its axis in a unidirectional magnetic field. The conductor can be a disk rotating in an axial magnetic field with radially

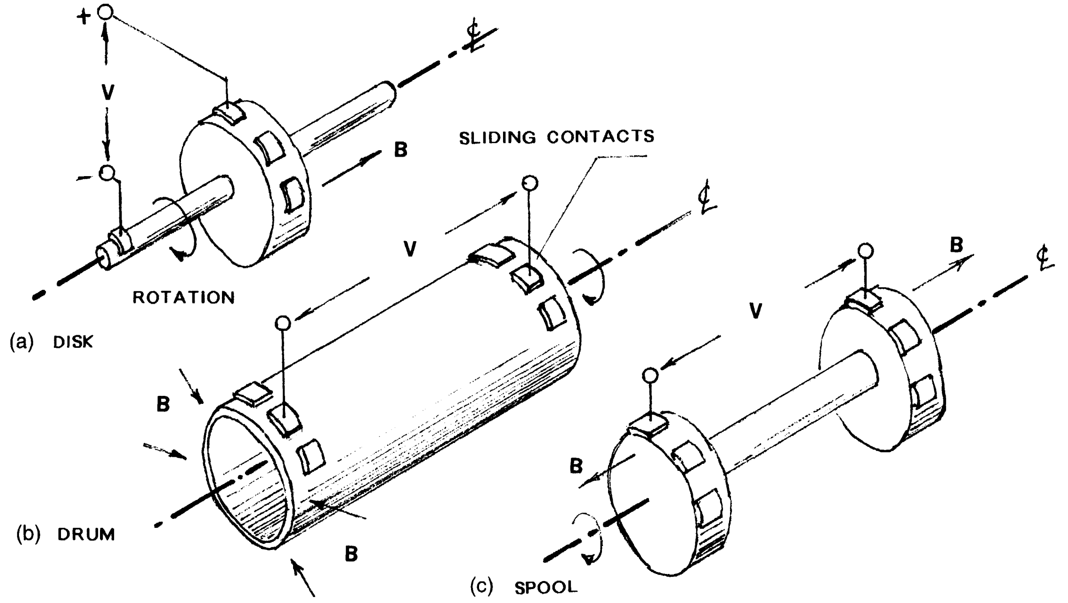

FIG. 4. Possible configurations for the rotor of three classes of homopolar machines. In all of the cases illustrated, the sliding contacts are stationary brushes attached to the frame of the machine. The rotor in (b) is an electrically conducting cylindrical shell mounted on and electrically insulated from a hollow shaft (not shown).

separated contacts [Fig. 4(a)] or a drum rotating in a radial field with axially separated contacts [Fig. 4(b)]. Other configurations are also possible [Fig. 4(c)].

The governing equations for the pulsed generator can be easily derived from basic considerations of electromagnetics. Indeed, it can be shown (Slemon and Straughen, 1980) that this equation is

$$V_t = L\frac{di}{dt} + Ri + \left(\frac{\phi}{4\pi^2 J}\right)\int i\,dt, \tag{6}$$

where i is the current circulating in the homopolar generator whose effective inductance and Ohmic resistance are L and R, respectively, and J is the moment of inertia about the axis of rotation. The quantity Φ is the magnetic flux that links with the rotor. The magnetic field contributing this flux is produced externally and is assumed to be constant with time. The above equation has made use of Kirchhoff's law for an electric circuit whose terminal voltage V_t is the potential difference appearing across the brushes of the homopolar generator. The generator can thus be modeled by the electric circuit of Fig. 5.

It is important to note that the quantity $4\pi^2(J/\Phi)$ associated with the inertia of the rotor is the magnitude of the capacitance of an effective condenser C in Kirchhoff's circuit equation. Now, it is well known from elementary circuit theory that the energy stored in the capacitor C of the circuit shown in Fig. 5 can be totally transferred to a load inductor (except for a small fraction that re-

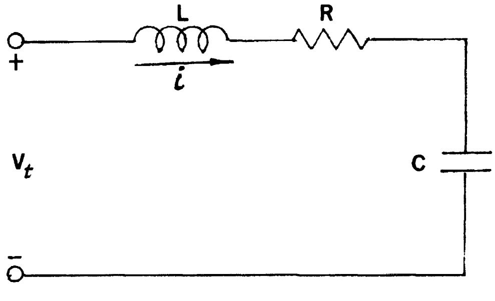

FIG. 5. Equivalent circuit for the pulsed homopolar generator. The inductance L represents the self-inductance of the rotor, R is the equivalent Ohmic resistance of the generator, while the mechanical inertia of the rotor, J, is reflected as an effective capacitance $C = 4\pi^2 J/\Phi$. The quantity Φ stands for the magnetic flux linking the rotor.

mains stored in the self-inductor L and another dissipated in the internal resistor R). These derivations corroborate the statement made earlier that the rotor inertia of a pulsed electric generator will prevent the loss of energy transfer discussed by Carnot.

2.2.2 Design of Pulsed Homopolar Generators. The interest in pulsed homopolar generators is due in great part to their applications to fusion research, rail and plasma guns, and, more recently, welding and metal processing. This warrants a brief review of design innovations that are scattered throughout the technical literature.

The criteria for an efficient pulsed machine are minimal resistance and self-inductance of the generator, magnetic fields of high value concentrated at the rim of the rotor (i.e., at that region that contributes the most to the moment of inertia), low losses in the field excitation, and, finally, selection of the output voltage that will enable the generator to deliver the required peak current.

The introduction of an iron yoke reduces the energy loss by an order of magnitude over that required by an air-core excitation coil. With this arrangement the field coils can be driven with a reasonably small dc supply. The reluctance of the path followed by the magnetic field excitation is further reduced by making the rotor an electrically conducting solid rotor of ferromagnetic material. A band of the rotor where the peripheral velocity is the highest is also the active portion of the generator. This shortened path for the current effectively reduces the inductance of the rotor. The current leads out of the machine are coaxial to minimize their self-inductance. The greatest contributor to the electric resistance of the generator are the brushes. Several materials have been investigated (MacNab *et al.*, 1980). Copper-plated graphite fiber brushes are found to be able to follow surface irregularities of the rotor rim and to maintain good contact.

Homopolar machines are essentially low-voltage generators. The magnitude of the peak current resulting from a given voltage V_t can be estimated by means of the circuit of Fig. 5 that models the pulse generator. Suppose that the energy stored in the rotor is to be discharged in a time τ. Then, from circuit theory (Guillemin, 1953) we find that $i_{\max}$, the maximum value of the current, is approximately

$$i_{\max} = V_t C(2\pi/\tau). \tag{7}$$

But we have also

$$V_t = \tfrac{1}{2} F v_{\text{tip}} R_0, \tag{8}$$

where v_{tip} stands for the peripheral velocity of the flywheel, B for the excitation flux density, and R_0 for the radius of the rotor. By inspection of the above relations, one infers that the peak current will increase with the open-circuit voltage, which, in turn, depends on the magnetic field, the tip velocity, and the size of the rotor. The issue is to find the design that yields the largest value of the open-circuit voltage within constraints of cost and size of the machine.

There is a practical limit to the increase that can be gained with conventional coils. Excessive Ohmic losses and restrictions of space impose a ceiling on the number of ampere-turns that can be fitted in a finite volume. The alternative would be to opt for superconducting or cryogenic coils. Either one of these solutions is expensive. Increasing the size of the radius entails major alterations in the dimensions of the yoke and the excitation coils. If the energy rating of the machine is to remain constant, the axial length of the machine must be reduced. When the length of the machine is too short, the profile of the rotor must be altered to make it stiffer. This is to prevent parasitic flutter vibrations. Clearly, these alterations too are costly.

Finally, the rotor can be spun faster. Here again the velocity of the rim must be kept below the critical speed to avoid failure of the rotor. An alternative would be to construct a new rotor from nonferromagnetic materials that are stronger. Unfortunately, the magnetic excitation system becomes complex and hence costly. The preferred solution must balance the various alternatives and arrive at an optimum trade-off.

2.2.3 Operational Experience. Several research groups have developed and constructed pulsed homopolar machines with a wide spectrum of rotational speeds, of energy ratings, and of time constants. Detailed descriptions of collections of such experimental investigations on these machines have been reported in the technical literature (Intersociety Energy Conversion Conference, 1988). A machine (Gully *et al.,* 1984) conceived by the Center for Electro-mechanics of the University of Texas at Austin is shown in the cut-away Fig. 6. Unlike almost all of the existing generators, which are one-of-a-kind, experimental machines, the Texas device has been refined to the point that it has been commercialized by OIME, Inc.

This machine was designed for a discharge time of 15 ms. To achieve this time constant, two stubby counter-rotating rotors were used. This method of construction not only reduces the response time but improves also the stability against vibrations.

The copper finger brushes used were actuated hydraulically, and when in contact with the rotor they were subjected to a 225-m/s slip speed. The designers felt that the brushes could sustain a current density of 570 A/cm^2. This high density had been verified by Marshall (1980) in an experimental test rig. The actual performance parameters are as indicated in Table 1.

Table 1. Performance parameters for the OIME machine developed by the Center of Electromechanics of the University of Texas (Austin).

Stored energy	6.2 MJ
Rotational speed	6245 rpm
Weight	1545 kg
Terminal voltage	50 V
Effective capacitance	4960 F
Internal inductance	30 mH
Internal resistance	7.5 Ω
Peak current	750 kA

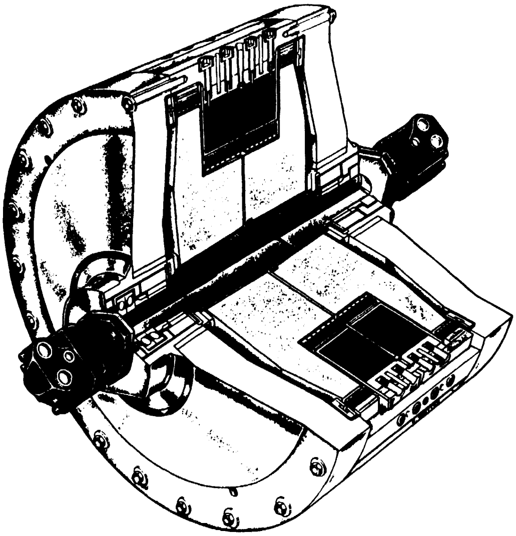

FIG. 6. Sectional view through the pulsed homopolar generator built by OIME, Inc. The cut-away shows the location of the brushes and of the two counter-rotating rotors. (After Gully *et al.*, 1984.)

3. THERMAL ENERGY STORAGE

Approximately 85% of the demand for commercial and residential energy requirements is for heating and/or cooling purposes. In countries that use electrical energy to supply hot water or space heating, thermal storage units installed in individual residences or business establishments can contribute appreciably to easing the demand on the electric utility power plant during the peak hours. These units consists of electrically heated ceramic storage materials (Polland, 1970). Consumers are encouraged to turn on these storage units during the off-peak hours by being offered a much cheaper charging rate. Thermal storage units have been incorporated also in heat-pump installations. In this case the ability of the pump to meet the heat load is improved considerably when the incoming cold air is preheated by a few degrees by passing it through a thermal storage unit. Another application for thermal storage is in air-conditioning plants for large office or apartment buildings where absorption rather than compression-cycle refrigerators are used. These systems are very practical since there are absorption chillers on the market that can be energized with sources as low as 85 °C.

The economic importance of energy savings in the industrial arena through thermal storage has been recognized for a long time. Some of the early work in this connection was initiated in Germany many decades ago. Several innovations related to thermal storage were developed at the Institute of Fuel Economy in Düsseldorf, in particular the two basic concepts of counterflow regenerators and waste heat recovery (Schmidt and Wilmott, 1981). Almost all of the metal-processing industries today make intense use of thermal storage to conserve waste heat. Similarly, the cement, glass, and ceramic industries also store exhaust heat for later use.

3.1 Thermal-Storage Processes

Thermal energy storage can be achieved by the storing of energy through heating, melting, or vaporizing a substance, the energy becoming available as heat when the process is reversed. Storage by causing a material to rise in temperature is referred to as sensible heat storage. Another method of thermal storage, which exploits the heat absorbed by phase transition at constant temperature, is known as latent-heat storage. Both sensible and latent-heat storage may occur in the same material, as when a solid is heated, melted, and then raised in temperature.

There are three basic types of storage systems using the sensible heat concept (Brumleve, 1974). In the first one, the working fluid already at high temperature and at low or high pressure is stored in a thermally insulated container until the energy stored is extracted. Hot water storage as used in homes for domestic water systems falls in this category. The second method relies on the transfer of heat from a stored material at high temperature to the working fluid. One of the most promising materials now available for storage is ferric oxide, Fe_2O_3. This material has a very high volume enthalpy and a close atomic packing and is inexpensive. At ideal single-crystal density, it would have the same thermal energy storage per unit volume as cast iron up to 700 °C. The ferric oxide usually is installed in a packed solid bed and heated. Finally, the third process consists in transferring heat from one fluid to another fluid via a stored material.

The latent-heat storage devices are mostly based on solid–liquid phase change concept because a substantial quantity of heat can be stored over a narrow change of temperature and without a large change of volume of the working substance (Schulten *et al.*, 1974).

Thermal energy-storage systems of the types discussed above can be designed to provide output temperatures between 100 and 1250 °C. The most favored technology, however, is the one based on sensible heat. It has been used with great success (Christensen *et al.*, 1985) in pressurized water storage tanks, in organic-liquid reservoirs for feed-water heating, in packed solid beds, and in fluidized solids. On the other hand, the latent-heat storage method has been applied successfully to the process associated with inorganic phase-change materials like the fluoride eutectics. The latter have been used often in solar-energy systems.

3.2 Heat Exchangers and Thermal-Storage Design

Storage systems that rely on sensible heat are used often enough that they merit some special discussion in this review. There are two classes of such systems that represent two complementary functional operation points of view. The first kind can be viewed as a thermal regenerator used to transfer heat between two fluids. This transfer is accomplished in two discrete steps, which are repeated periodically. Heat is first stored in a permeable packing by flowing a hot fluid through it. During this portion of the cycle, sensible heat is deposited in the packing. Subsequently, the stored heat is regained or "regenerated" by passing a cold fluid through the packing. This mode of operation is often referred to as the "single-blow" procedure because the surface of the packing is submitted to a succession of *intermittent* hot and cold blows.

The second type of thermal storage operates in such a fashion as to allow a continuous exchange of heat between a hot and cold fluid. Several schemes have been developed to achieve this mode of operation. The simplest of these schemes makes use of two packing beds. By means of a combination of valves the cold fluid is diverted alternately to one then the other of the two packings. While one of the beds is being discharged and heating the fluid, the other is heated and made ready for the next phase of the cycle. The result is an almost continuous heating of the cold fluid. In a more complex arrangement, the heated packing is placed on a rotating tray. One-half of the tray is exposed to the hot fluid and the other to the incoming cold fluid. As the tray rotates, heat temporarily stored into the packing is physically moved into the cold stream. Considerable attention must be given to provide seals to minimize the leakage between the counterflowing cold and hot fluids.

The prediction of the performance dynamics of any one of these storage systems is a difficult task. To render the analysis tractable, the packing bed is idealized by a model consisting of a number of conduits or passages imbedded in a matrix of high heat capacity. Now it can be shown (Szego and Schmidt, 1978) that the optimum performance of a thermal-storage system conforming to the idealized model is obtained when the temperature gradient along the direction of flow of the fluid in the conduits is considerably greater than the gradient normal to the flow. A considerable amount of work has been done (Gröber and Erk, 1961; Petukhov, 1970) to evaluate the film thermal resistance at the interface between the flowing fluid and the walls of the conduits in order to satisfy the above-mentioned criterion. Thermal regenerators have been in use for several years for a number of applications. A substantial amount of documentation on their design is readily available in the technical literature (Kays and London, 1964).

4. COMPRESSED-AIR STORAGE

Air storage system energy transfer (ASSET) uses a process of great interest to the electric utilities. ASSET systems are simple to install and operate. They are less costly than gas turbines or old steam power plants reserved exclusively for peak load demands. Other advantages of ASSET systems is their ability to start fast (about 6 min) and to operate efficiently even at a fraction of their capacity. Unlike hydroelectric pumped-storage systems, they do not create environmental disturbances.

4.1 Principle of Operation

In an ASSET system (Stys, 1983) the pressure of ambient air is raised by compressors to a high value (up to 70 bars). The air is then cooled and subsequently stored in an underground cavity. During the peaking cycle, air is led from the underground store, heated, and then allowed to expand in a turbine, whereby the energy stored is retrieved. Such a system is sketched in Fig. 7.

Numerous variations in the design of such systems have been suggested. Most of the proposals rely on an air storage that can be an aquifer, a mined cavern, or a salt dome. The air storage can be maintained at constant or variable pressure. The heat of com-

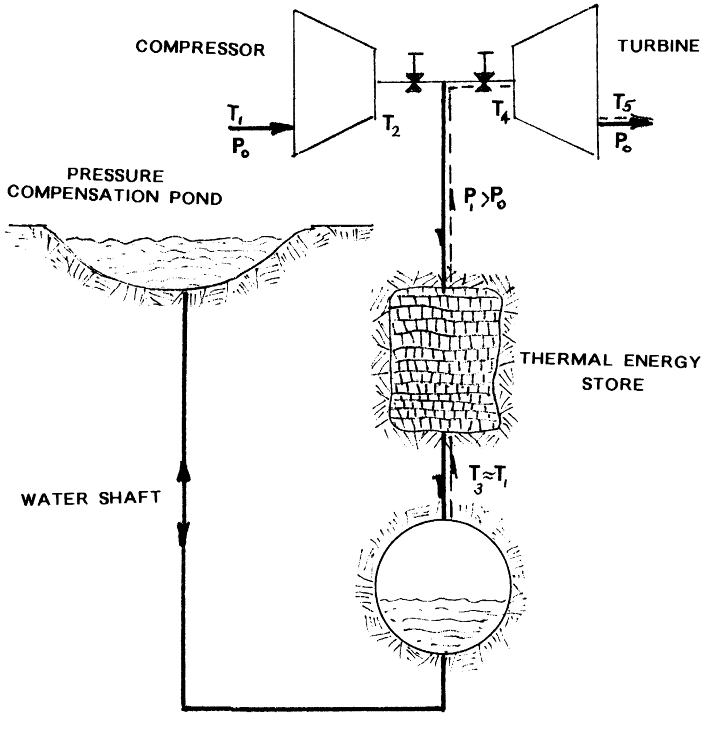

FIG. 7. Operational diagram for an ASSET system (after Stys, 1983). During the charging stage, the motor/generator machine connected to the grid acts as a motor to actuate the compressor. The compressed air raised to 70 atm (≈1000 psi) is stored in an underground cavity. In the discharge phase the high-pressure air is preheated prior to being directed to the two turbines HGT (high temperature) and LGT (low temperature) connected in tandem. The electrical machine attached to the turbines shaft now is run as a generator returning electric power to the grid.

pression may or may not be stored. If it is not stored, combustion of additional fuel is required to make up for the heat lost. When the heat of compression is stored, it is returned to the air upon discharge. The overall process may be considered adiabatic. The advantage of such a system is that it has a high stage efficiency. It is also a stand-alone system that does not require additional fuel (to reheat the air).

The performance advantage of storing the heat of compression was recognized early in the development of this class of storage systems. The proposed procedure included an underground rock-based bed, which allows recuperation in a single (Glendening, 1974) or two-stage thermal cycle (Stephens, 1976), as encountered in the ASSET system. The basic scheme for a single stage is shown in Fig. 8.

4.2 Operational Experience

The Nordwestdeutsche Kraftwerke (NWK) of Hamburg, Germany, installed the first ASSET plant in 1978. A large fraction of NWK generating capacity is from nuclear reactors, which require system management. Since the subsoil of Northwest Germany contains a large number of salt domes, it was the obvious solution for the location of a storage system. Two such caverns with a volume of 150 000 m^3 apiece were combined to provide the air reservoir for the system. The two domes could provide two hours' coverage at full-load peak. The NWK storage system has performed very well, as indicated by the operation parameters given in Table 2.

In the United States, Soyland Power Cooperative, Inc. of Decatur, IL, is a complex of 15 individual Illinois cooperatives. A por-

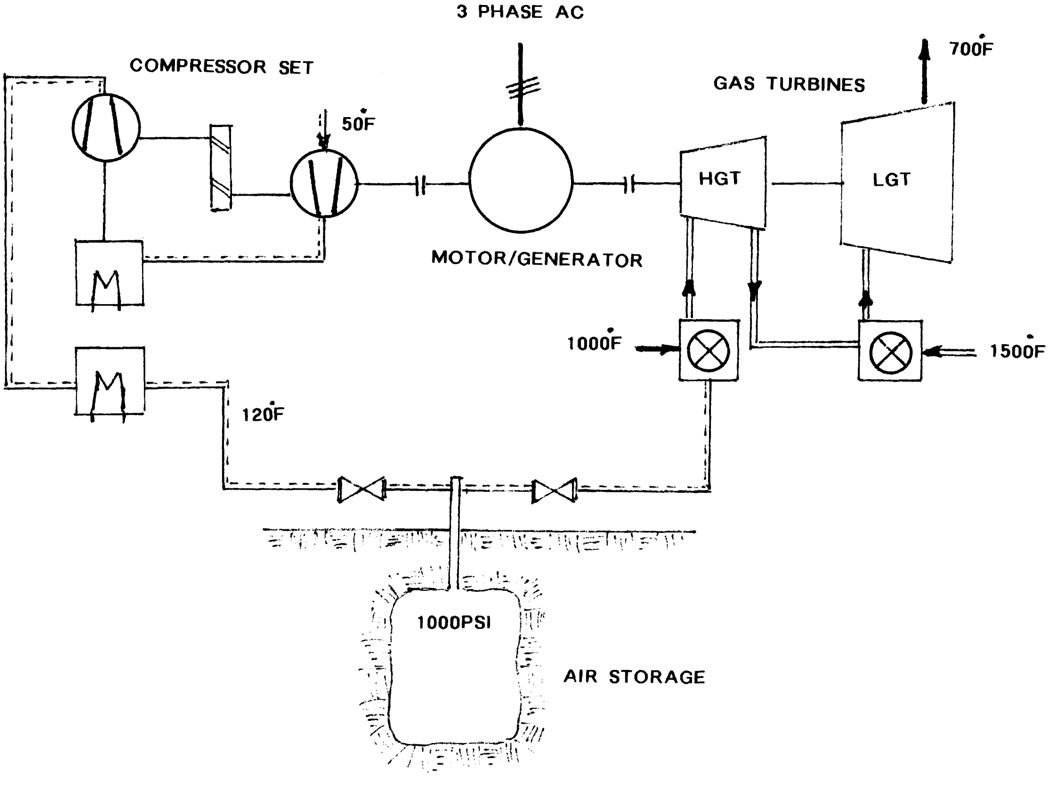

FIG. 8. Single-stage heat-of-compression storage adiabatic compression system (after Glendening, 1974). As the air is compressed, its temperature is raised above its intake temperature. The sensible heat associated with the air is deposited in a thermal store prior to being stored in an underground cavern. The pressure compensation pond maintains the air-pressure constant. Upon discharge, the air is diverted to a turbo-generator to regain the energy stored in the air. The operating temperatures and pressures are indicated on the sketch.

Table 2. Operating parameters of the ASSET plant in Hamburg, Germany (Stys, 1983).

Interval considered	One year
Charging time	1 874 h
Discharging time	840 h
Capital invested (including salt domes)	\$230/kW
Energy delivered	86 475 MWh

tion of Soyland's capacity is nuclear. Following a system analysis of their power generation management, Soyland has decided to include in the planning the addition of an ASSET system. Such a system is being constructed.

5. PUMPED-HYDRO STORAGE

Pumped-hydro storage is based on the principle that the potential energy of water can be increased, if surplus electric power is available, by pumping the water from a lower elevation to a higher one, thus indirectly storing energy. A variation on the conventional pumped storage is the underground pumped-hydro system. The impetus in adapting pumped storage to electric utility systems is attributed to the development of the reversible Francis-type impeller wheel, operating in one direction as a pump and in the other as a turbine. This has reduced the cost of medium-head pumped storage most prevalent in the United States.

Pumped-hydro systems have been in operation for many years in the United States and overseas. Their technology is well established and extensive documentation (Hunt, 1968) on the subject is readily available. This important method of storage will not be reviewed here in detail; it is mentioned only for completeness. Although new pumped-hydro plants are still being installed overseas, pumped-hydro storage plants are losing their popularity in the United States because of their adverse impact on the environment by flooding large tracts of land and altering the level of the underground water table.

6. CHEMICAL ENERGY STORAGE

A basic criterion for a chemical energy-storage system is its ability to be reversible. It should be able to absorb energy from a primary power source, store it, and deliver upon demand this energy in a usable form: usually electricity or heat. As a result of the reversibility requirements, chemical storage systems tend to be elaborate. Chemical reactions offer large energy-storage densities as a result of the highly energetic processes involved in the reversible destruction and reformation of the chemical bonds. Catalytic reactions are particularly attractive because the products of the endothermic reactions contain thermal energy as chemical bond energy that is retrievable over an appropriate catalyst.

The emphasis here has been placed on chemical storage in a broad sense and considers systems more complex than the specific chemical reactions mentioned above. This statement is clarified by two such examples that have been commercialized, namely, the hydrogen-cycle system and electrical batteries.

6.1 Hydrogen-Cycle System

The hydrogen-cycle peak shaving system has received a great deal of attention from the electric utility energy systems. It consists of a complex mix of chemical, thermal, and electromechanical energy conversion components so combined as to optimize performance and cost. A typical hydrogen-storage system is shown in Fig. 9 (Salzano *et al.*, 1974). During the off-peak hours of the utility system the output of the converter–inverter is directed to electrolysis cells. The hydrogen produced by electrolysis is dried first and then stored in an iron–titanium storage tank.

Under peak load conditions the electrolysis cells are disconnected from the network, and the fuel cells are restored on line. The dc output from the fuel cells is processed through the inverter–converter prior to being connected to the ac network. In an alternative scheme one starts with the combustion of hydrogen and recovers the energy stored by means of a gas turbine or a magnetohydrodynamics (MHD) generator operated in an open-cycle mode. The main features of the hydrogen-cycle system are discussed below.

6.1.1 Hydrogen Production. The cost of producing hydrogen by electrolysis represents an important fraction of the capital in-

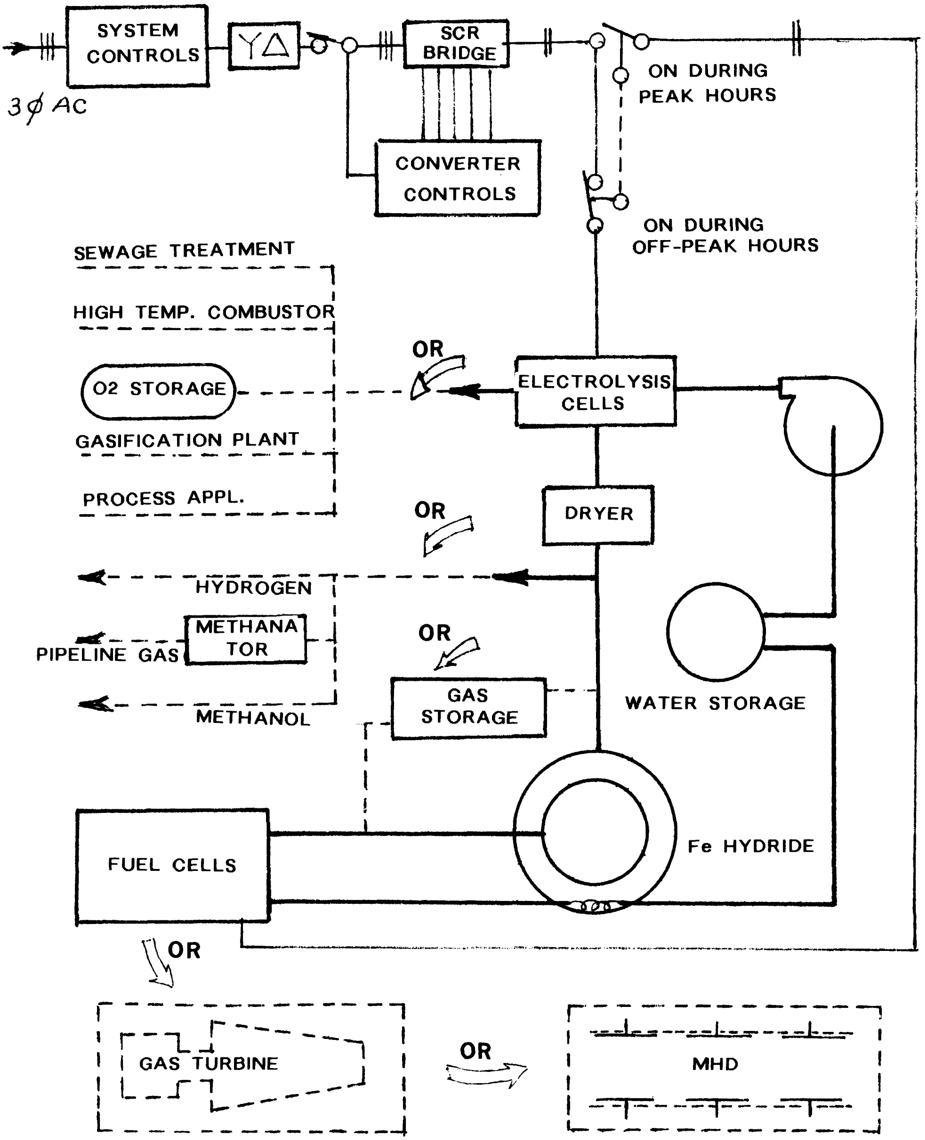
SYSTEM CONTROLS
3ϕ AC
YΔ
SCR BRIDGE
CONVERTER CONTROLS
ON DURING PEAK HOURS
ON DURING OFF-PEAK HOURS
SEWAGE TREATMENT
HIGH TEMP. COMBUSTOR
O2 STORAGE
GASIFICATION PLANT
PROCESS APPL.
OR
ELECTROLYSIS CELLS
OR
DRYER
HYDROGEN
METHANATOR
PIPELINE GAS
METHANOL
OR
GAS STORAGE
WATER STORAGE
Fe HYDRIDE
FUEL CELLS
OR
GAS TURBINE
OR
MHD

FIG. 9. Schematic of a hydrogen peak shaving system. The output hydrogen emerging from the storage can be directed to a fuel cell for direct conversion to electricity, or after burning used in a MHD generator or channeled to a gas turbine. (After Fernandes, 1976.)

vestment on the system. One of two competing schemes is used for the electrolysis. The first is a solid polymer sulfonic acid cell developed mainly by the General Electric Co., and the other is an alkaline cell produced by Teledyne Isotopes.

The solid polymer electrolyte is a sulfonic acid plastic membrane (Nultall and Fernandes, 1974) through which ionic conductivity is contributed by hydrated hydrogen ions. The maintenance of the cells is simple because of the absence of free acids or caustic liquids. The membrane acts as a barrier to the hydrogen generated and can withstand pressures as high as 35 bars.

Teledyne electrolysis cells (Murray and Yaffe, 1980) with capacities as high as 4 tons per day have been designed. Unlike the polymer cells, which require an operating temperature of about 300 °F, the alkaline cells function at room temperature and need light catalyst loading.

6.1.2 Storage of Hydrogen. A key to the success of the hydrogen-cycle system is the ability to store hydrogen safely for a long time, at reasonable cost, and to charge and discharge the hydrogen gas quickly. Several solutions have been found acceptable. One is to store the hydrogen in cylinders at high pressure (70 to 105 bars). The recovery performance is close to 98%. A second option is the iron hydride storage scheme (Strickland *et al.*, 1974). In this method the hydrogen is absorbed by iron–titanium in an exothermic reaction described by

$$\mathrm{FeTi} + \frac{x}{2}\mathrm{H}_2 = \mathrm{FeTiH}_x + \Delta, \qquad (9)$$

where Δ, the heat of reaction, is approximately 16 000 kJ/kg. The advantage of this method is that the hydrogen can be stored at room temperature, quite safely and at an efficiency of 95%. Attempts have also been made to store the hydrogen as liquid hydrogen. The process unfortunately is costly and inefficient (66%).

6.2 Electrical Batteries

As a result of a study (Crouch and Perry, 1973) sponsored by the U.S. Department of Energy, the Electric Power Research Institute, and the International Lead Zinc Research Institute Organization, it was found that lead-acid batteries could be used as storage batteries in an electric network. On the other hand, the cost of these batteries designed for the required cycle life is prohibitive as compared to other energy-storage systems such as pumped hydro storage.

The need for electric batteries for use in electric cars or in uninterrupted power supplies has instigated the development of advanced light-weight batteries capable of long life and reasonably priced. The general guidelines in the search for these batteries have been a preference for soluble or liquid reactants to achieve reproducibility in morphology of active materials, and high temperatures for improved reversibility. Some of the most promising systems being considered are sodium–sulfur (and antimony trichloride), lithium-alloy/metal sulfide, and redox. Several research reports on the latest developments of these advanced batteries are reviewed in the proceedings of an annual conference: the IEEE International Power Source Symposium.

A typical example of an advanced battery system is the molten sodium chloride aluminate cell. This battery uses inexpensive materials in its construction; it operates at 200 °C so that one can use Teflon seals and aluminum containers. Furthermore, it has a long cycle life because there are no solid–solid transformations.

7. MAGNETIC STORAGE

In a superconducting magnetic energy-storage system, the energy is stored in the magnetic field associated with a coil. If the coil is wound with superconducting material, then once the coil is "charged" the magnetic energy stored can be preserved indefinitely provided the coil can be maintained in the superconducting state.

Historically, when the art of manufacturing superconductors on a large scale first became a practical reality, applications were quick to start in experimental plasma physics (Robson and Lupton, 1975), solid state, and particle physics (Hammel *et al.*, 1970). Ferrier (1970) was the first to propose the use of a superconducting coil as an energy storage for the Electricité de France. Ferrier's scheme required the installation of a very large central superconducting magnetic energy storage (SMES) for the whole network. Ferrier's concept was judged impractical, and the proposal was abandoned.

In 1971, a program was started in the United States at the University of Wisconsin to re-evaluate the SMES concept and to examine in detail the operation of the components in the system. The project leaders, Boom and Peterson, came to the conclusion (Peterson *et al.*, 1975) that there was great merit in having the SMES interface the electric network through a Graetz bridge. This circuit, discussed in some detail in Sec. 7.1.3, has the attractive feature of effectively stabilizing the system against oscillations while acting as an inverter–converter.

The definitive study for the feasibility of SMES was conducted at the Los Alamos Scientific Laboratory in 1972 (Hassenzahl *et al.*, 1973). This was as a result of a directive by the U.S. Department of Energy to conduct a realistic cost analysis to provide guidelines on the course of action to take on SMES.

One component of the SMES unit, the structure required to deal with the magnetic forces on the conductors, was found to be so expensive as to make the system impractical. However, two concepts developed in 1972 completely changed this picture. The use of warm supports for the superconducting coil (Powell and Bezler, 1973) not only reduced the refrigeration load but also made it feasible to consider a warm structure such as *in situ* rock as a less costly support to take care of the magnetic forces.

Since the reinforcing structure represents by far the major cost of a SMES system, attempts were made to identify coil configurations in which the fields and currents conspire to compensate for the internal forces. Mawardi (1975) proposed a combination of poloidal and toroidal coils that would achieve quasi force-free conditions, thus eliminating the need of external support. The coil, however, deemed not cost-effective per unit of energy stored, was not pursued any further. The scheme, however, is well suited to space applications where different criteria apply and where rock support is not available.

The early estimates of SMES indicated that only very large units (greater than 1000 MW·h) would be economical. It was difficult, therefore, to justify the construction of a smaller unit on which a SMES system could be made economically viable. As a result of a series of conferences and workshops in the United States and Japan (SMES Review Panel, 1976), a suggestion was made to construct a small SMES to stabilize long interties between power plants.

7.1 Components of SMES

Reference designs of a SMES unit of 1 GW·h rating have been compiled by the Los Alamos group (Rogers and Turner, 1980) and by the University of Wisconsin team (Wisconsin Superconducting Energy Storage, 1974). The two designs contained novel ideas for the construction of the superconductors, the coil configuration, the restraining structure, the Dewar vessel, and the inverter–converter interfacing circuit. These findings are summarized below.

7.1.1 The Superconductors. The cost of the superconductors is an important fraction of the capital invested on a SMES system. To evaluate accurately the minimum mass of conductors required, one needs to determine the maximum permissible current density in the conductors, the cooling method, and the heat capacity of the conductors.

Two possible methods of cooling are possible: the pool-cooled conductors (Mullan, 1977) and the force-cooled conductors (Hoening, 1976). In the first method a cable of conductors with many strands (500–2000) is immersed in a bath of liquid helium. In the second the helium is forced through the conductor itself.

Conductors should be designed to be cryostable; i.e., the heat generation, when all the current flows in the stabilizing material, should be less than the heat-removal rate capacity of the system. The ratio of stabilizing material (Cu, Al, . . .) to superconducting material is usually chosen so that the operating current density is one-half of critical. This conservative rule of thumb has been abandoned in favor of the "grading" procedure. This latter procedure, which has been incorporated in the SMES design, rests on the fact that the critical current density increases inversely with the strength of the magnetic field. Accordingly, one adjusts the quantity of superconductors in the cable so that the current density throughout the windings always is at 90% of the critical density. This procedure leads to substantial saving of conductors.

The superconductors used in the SMES are multifilament (Rogers, 1979) extruded copper with a copper-to-NbTi ratio of 1.33. The composite is then drawn to a strand. A flat transposed cable is made of 23 such strands and subsequently surrounded by a stabilizer of aluminum clad with a thin layer of copper. Drawing of the assembly produces a conductor as shown in Fig. 10. The current density in the aluminum is reckoned to be no more than 15 kA/cm^2. At that value of the current, 1% of the stored energy is dissipated in the aluminum, which must not exceed a temperature of 100 K.

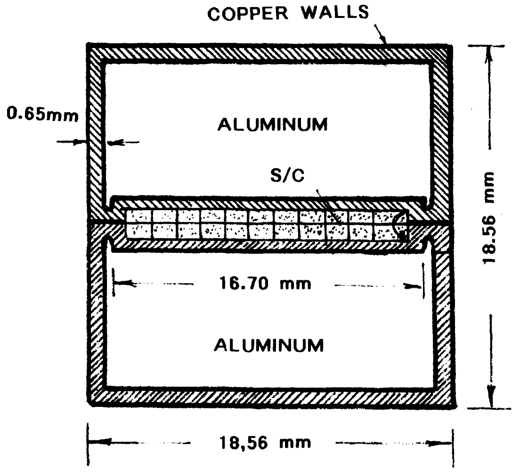

FIG. 10. Cross-sectional view of the conductors used in the windings of the SMES (after Rogers, 1979). The superconductors are located in the center strip. The aluminum bars contribute mostly to the stabilization and to the mechanical strength of the conductors.

7.1.2 The Coil and Its Accessories. The coil is cylindrical and is pool-cooled by liquid helium at 1.9 K. The discharge requirement of the coil is that the maximum turn-to-turn voltage does not exceed 100 V. The coil has accordingly been subdivided into 13 segments, each with its individual Dewar vessel. The coil characteristics are summarized in Table 3.

The layout for the coil is drawn in Fig. 11(a). Figure 11(b), the cross-sectional view across the plane "AA" in Fig. 11(a), illustrates the manner in which the helium vessel is installed. Figure 11(c), on the other hand, gives a detailed view of the construction of the warm supports. The radial and axial forces on the conductors are transmitted through the conductor stack to the helium vessel and to the support. The axial forces accumulate. The windings into each of the segments have been divided vertically into two parts. One abuts against the Dewar vessel and the other on ledges fitted to the inner and outer walls of the Dewars. The coils and their Dewar are installed into an underground excavation whose profile is shown in Fig. 12. The axial pressures (tending to compress the coil) are constant and are evaluated to be 108 bars. The radial pressures vary with height. They are a minimum at the top of the coil (12.5 bars) and reach a maximum about mid-height (100 bars).

Table 3. Characteristics of superconducting coil designed by the Los Alamos Group (after Rogers and Turner, 1980).

Average radius	66 m
Height	44 m
Radial thickness	0.20 m
Inductance	3170 H
Current at full charge	50 kA
Coolant	Helium @ 1.8 K and 1 atm
Number of turns	4280 (in 5 layers)

Cooling of the coil is accomplished by means of a two-fluid system. The heat load is due mainly to losses through the supports and the leads in and out of the SMES coil. To reduce these losses the supports and the leads are vapor-cooled to 63 K by the return helium.

7.1.3 The Graetz Circuit. Some very brief comments will be made on the novel use of this classic circuit in the SMES. This circuit, shown in Fig. 13, behaves as an ac-to-dc rectifier and a dc-to-ac converter. In the charging mode, the direction of the current I_d in the coil is fixed by the SCRs. In the discharging phase, the polarity is reversed, while the voltage across the coil is made null when the storage coil is idle.

The three-phase bridge consists of SCRs, which are controlled by a firing circuit. The firing pulses use for reference the 60-Hz power source. By altering the phase of the firing signal with respect to the 60-Hz reference, the voltage can be made to change in magnitude and polarity. The converter operation is explained along the same lines.

The pulsed bridge generates harmonics, which can be transmitted to the network. The amplitude of the harmonics can be reduced by increasing the frequency of the firing signals or by filtering.

7.1.4 Conclusion. The fast response time, high efficiency, and environmentally safe attributes of the SMES make it a system of great interest to the utilities. Continuous technical improvements in the cost and performance of the components are bound to bring the operation of the SMES to a level that will be competitive with other methods of storage. There is a great deal of expectation that it will happen in the very near future.

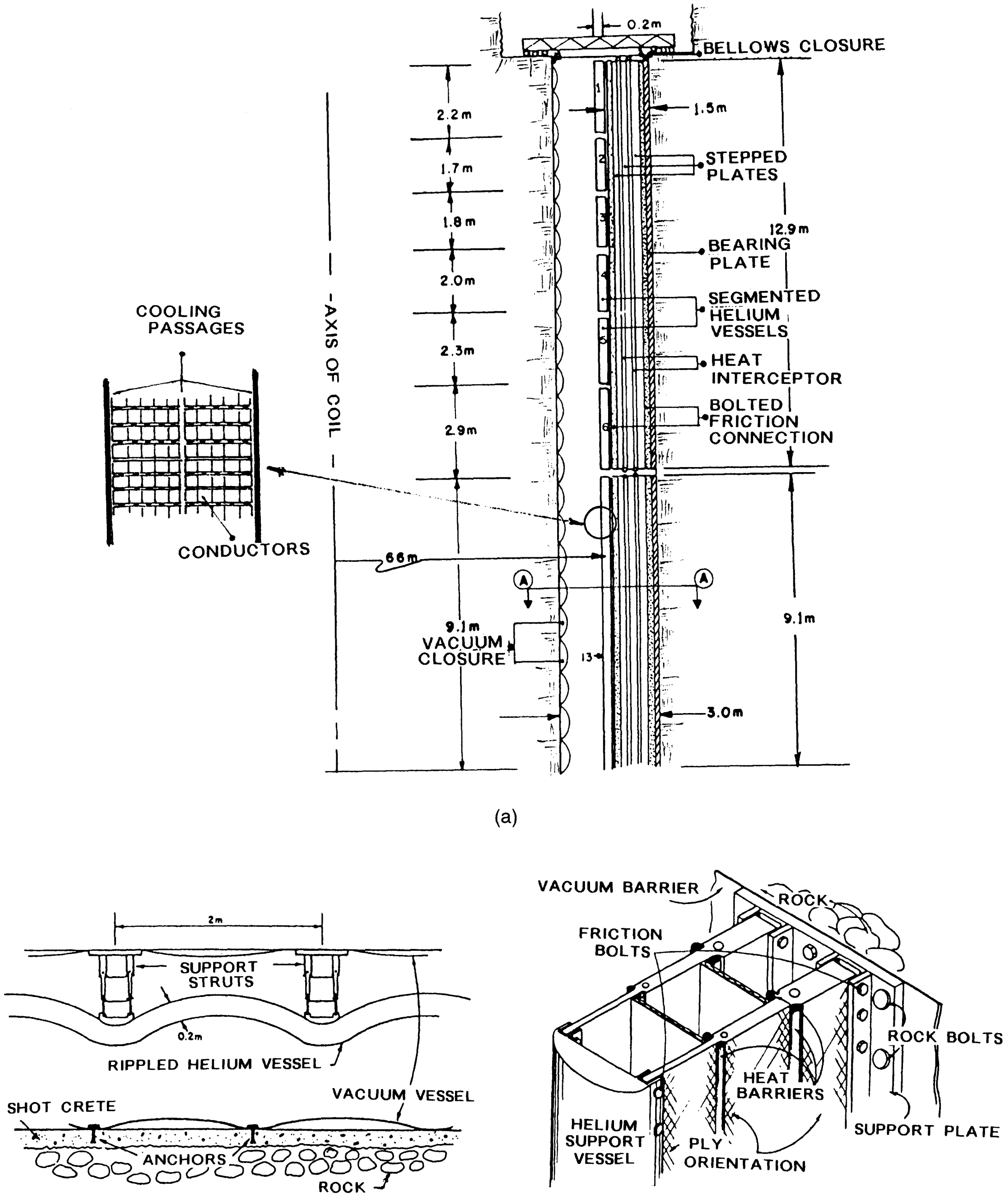
0.2m
BELLOWS CLOSURE
2.2m
1.7m
1.8m
2.0m
2.3m
2.9m
1.5m
STEPPED PLATES
12.9m
BEARING PLATE
SEGMENTED HELIUM VESSELS
HEAT INTERCEPTOR
BOLTED FRICTION CONNECTION
AXIS OF COIL
COOLING PASSAGES
CONDUCTORS
66m
A
A
9.1m
VACUUM CLOSURE
13
9.1m
3.0m
2m
SUPPORT STRUTS
0.2m
RIPPLED HELIUM VESSEL
VACUUM VESSEL
SHOT CRETE
ANCHORS
ROCK
VACUUM BARRIER
ROCK
FRICTION BOLTS
ROCK BOLTS
HEAT BARRIERS
HELIUM SUPPORT VESSEL
PLY ORIENTATION
SUPPORT PLATE

(a)

(b)

(c)

FIG. 11. (a) Cross section of a SMES coil in an underground tunnel showing dimensions of various segments. Only half a coil is shown. The midplane coil depth is about 200 m from the surface. Inset: The stacking of the conductors in the liquid-helium Dewar vessel. As stated in the text, the conductors are anchored to the walls of the Dewar. (After Hassenzahl, 1983.) (b) Projected top view in the plane AA in (a). The view shows a section through the vacuum vessel. This illustrates the method used in attaching the vacuum vessel to the rock. The rippled construction of the liquid-helium container permits it to expand or contract freely. (After Hassenzahl, 1983.) (c) Schematic view for the warm supports to which the liquid-helium Dewar vessel containing the superconducting conductors is attached. The supports minimize the heat leaks to the bed rock. (After Rogers, 1979.)

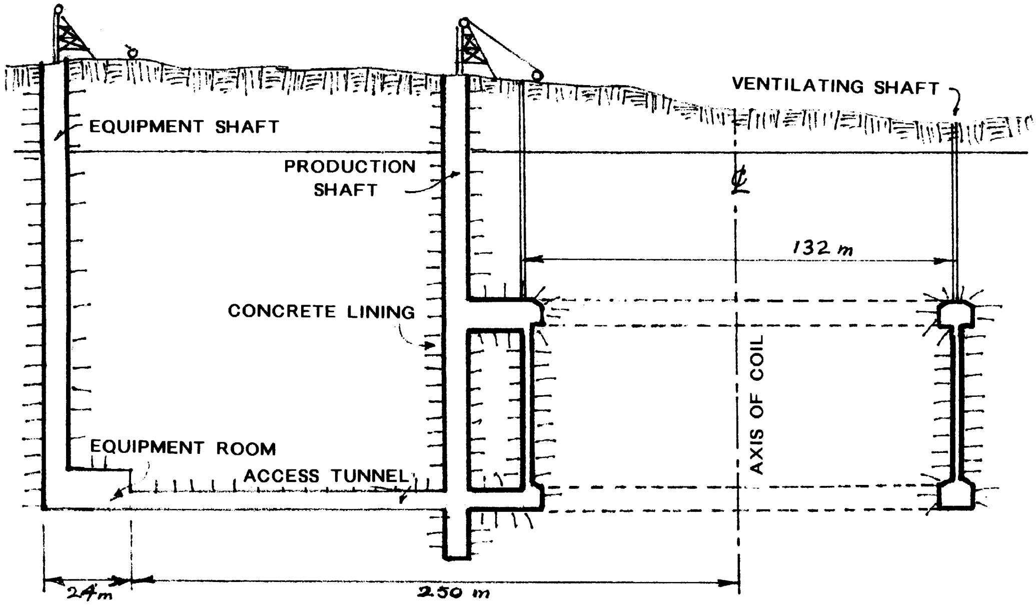
VENTILATING SHAFT
EQUIPMENT SHAFT
PRODUCTION SHAFT
132 m
CONCRETE LINING
AXIS OF COIL
EQUIPMENT ROOM
ACCESS TUNNEL
24 m
250 m

FIG. 12. Profile of tunnel excavated in the bedrock to accomodate the SMES coil (after Rogers, 1979).

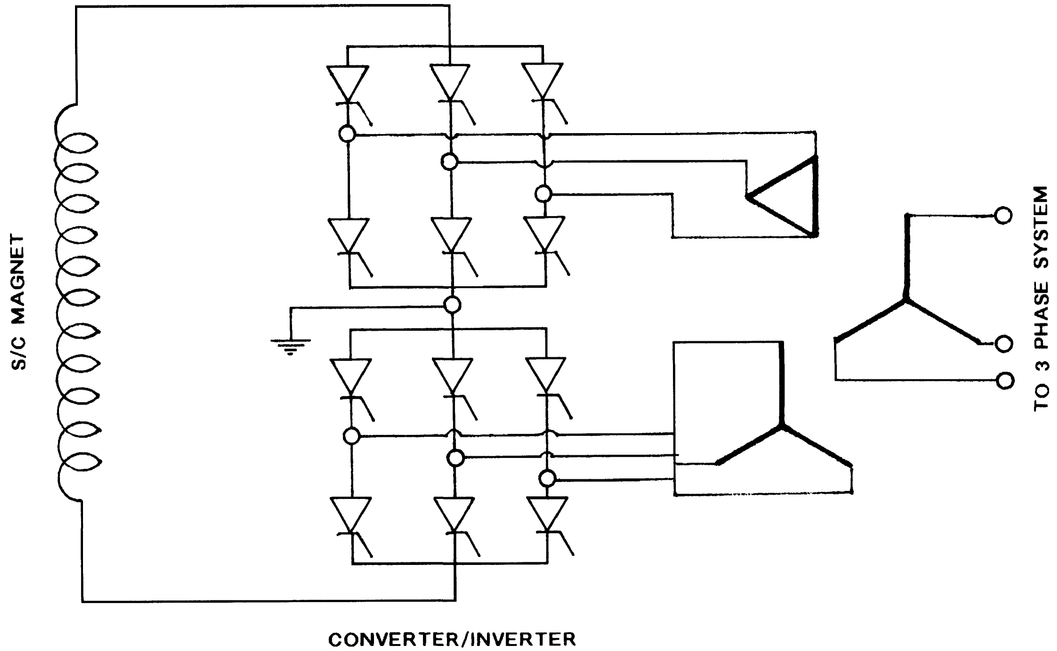
S/C MAGNET
TO 3 PHASE SYSTEM
CONVERTER/INVERTER

FIG. 13. Schematic diagram for the Graetz circuit. This inverter/converter is specially suited for high-voltage, high-power operation. (After Wisconsin Superconducting Energy Storage, 1976.)

7.1.5 SMES and Transmission Stabilization. The Pacific Northwest and southern California are part of the Western U.S. Power System. The electric-generating centers of these regions are connected to the load by two intertie complexes: the 900-mile-long Pacific AC Intertie and the Pacific HVDC Intertie.

Studies completed prior to the installation of the AC Intertie predicted (Cresap and Mittlestadt, 1976) that the line had the potential of exhibiting self-excited oscillations at the frequency of 0.33 Hz. On 7 August 1974 such oscillations did occur with a power amplitude swing of 300 MW and a frequency of 0.35 Hz.

To forestall the occurrence of such oscillations, the Bonneville Power Administration (BPA) installed equipment on the high-voltage direct current (HVDC) line to modulate the power flow and damp the oscillations. Unfortunately, the HVDC line is available for only 89% of the time and furthermore is subject to being down on account of earthquakes.

It had occurred to a number of investigators (Byerly, 1980) that the addition of a SMES to the network would add enough inertia in the system to stabilize it. The SMES is a reliable alternative to and less expensive than a redundent HVDC transmission line. Los Alamos thus was commissioned to build a 30-MJ SMES.

The design of this unit was similar to the 1-GW·h reference unit except for two features. Since the SMES would operate under time-varying conditions, the conductors had to be designed to withstand such conditions. Accordingly, the construction shown in Fig. 14 (Rogers *et al.*, 1983) was adopted. The second variation in the design was the use of a nonmetallic Dewar vessel to avoid the excitation of eddy currents in the walls of the vessel.

7.2 Pulsed Inductive Storage

The transfer of energy from a SMES takes place over times that can be as short as a second and that can extend to periods of the order of hours. On the other hand, for the pulsed homopolar generator, we found that the time constant ranged from milliseconds to a few seconds. The transfer of energy from a charged air-core inductor to a load, however, can take place in times of the order of microseconds or even, with proper care, of nanoseconds. In view of the many applications that are in need of these rapid pulses of energy, it was considered worth while to devote a special section on this class of energy storage.

One of the simplest arrangements for the transfer of energy from a charged inductor to a load is shown in Fig. 15. Initially the switch is closed to charge the storage inductor *L*. Upon opening the switch, energy is transferred from the inductor *L* to the load. As was discussed in Sec. 1.2, if the load is inductive, then adding a capacitance to the circuit improves the efficiency of transfer.

There are several features of the above circuit that must be considered. The voltage *e*(*t*) can climb to very high values that may reach the megavolt range. The design criterion for the inductor is that it can withstand

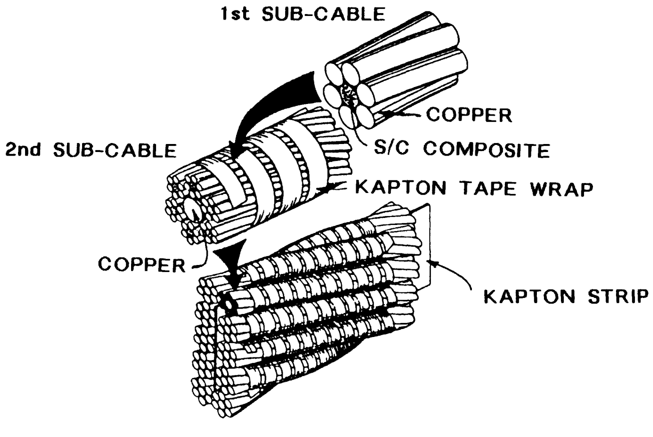

FIG. 14. Superconducting braided conductor used in a SMES unit operating as a stabilizer (after Rogers and Turner, 1980).

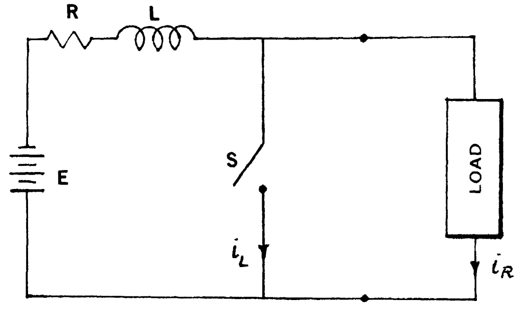

FIG. 15. Classic circuit for a pulsed inductive source. The inductor is charged when the switch S is closed. Discharge takes place upon opening S.

high voltages and avoid electric flashover. The switch is not ideal, i.e., it does not instantly acquire an infinite resistance once it opens. Quite often the switch arcs and the current interruption takes place only after extinction of the arc. Trost *et al.* (1976) calculated the performance of the circuit of Fig. 15 when the ideal switch was replaced with a time-dependent resistor representing an arc. It was shown that the peak voltage was reduced by almost a third (of the ideal case), when the switch "opens" in a time equal to twice the decay period (R/L) of the circuit.

At first glance it would appear that there would be a definite advantage in making inductors out of superconducting materials. This would yield $R/L \rightarrow 0$. Unfortunately, when such conductors are subjected to rapidly varying fields, severe eddy currents are induced in the stabilizing material. As a result the coil may quench to the normal state. Many attempts have been made to construct superconductors for ac fields (Yamamoto, 1986). So far, most of the results have been disappointing.

Several variations on the simple circuit of Fig. 15 have been used with the purpose of shaping the pulse. One such circuit is illustrated in Fig. 16 (Holland, 1984). In it the inductors are successively brought on line. The technology of pulse shaping is highly developed, and the interested reader can find additional information in the literature. Up-to-date information on this subject can be found in the proceedings of the IEEE Pulsed Power annual conferences.

GLOSSARY

ASSET: An acronym for air-storage system energy transfer. This refers to a system comprising both the compressed-air installation and the air turbine to extract the energy from the compressed air.

Force-Free Coil: A coil, first discussed by E. Fermi and S. Chandrasekhar, in which

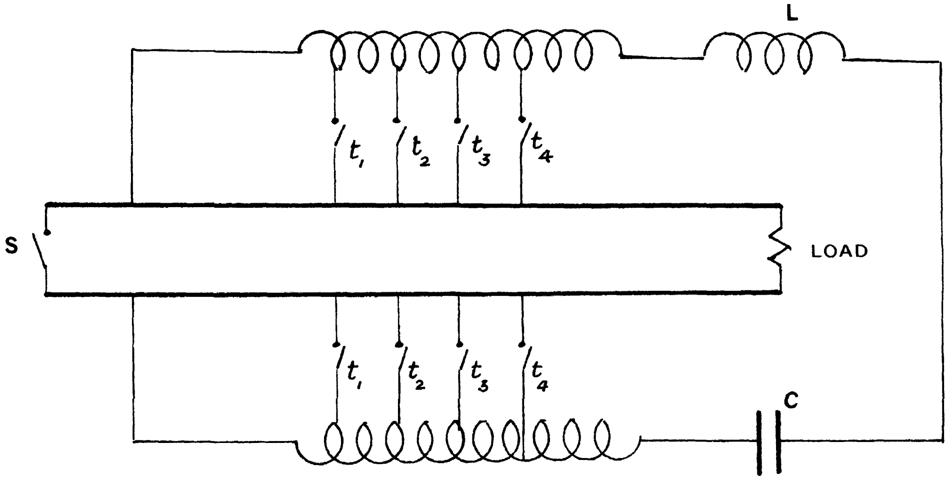

FIG. 16. Pulse-shaping circuit that doubles as an inductive storage system (after Holland, 1984). Charging of the inductor L takes place by closing S and opening the switches t_1, t_2, ... To discharge the inductor and start a current pulse down the transmission line, S is closed first, and subsequently, the switches t_1, t_2, . . . are closed in succession.

the current in the coil windings is everywhere parallel to the induced magnetic field. As a result, there are no net forces on the conductors.

Francis Turbine: A reversible machine that can operate as a pump and as a turbine.

Graetz Circuit: An inverter/converter with very low harmonic content.

Homopolar Generator: Originally invented by Faraday, consists of a solid conductor in the form of a cylinder or disk that is spun in a magnetic field. The device behaves as a generator with a one-turn winding. As a result, large currents can be produced.

Peak Load Shaving: A term applied to the procedure of reducing the output power from a generating plant by relying on a storage system to return stored energy and thus supplement the power plant during the "peak" load.

Pumped Hydro: Short for pumped hydroelectric.

SMES: Superconducting magnetic energy storage.

Stable Superconductor: A composite of superconductors embedded in normal conductors with sufficient heat capacity that they can act as a heat sink and prevent the temperature of the superconductor from rising as a result of a surge inducing eddy currents in islands of normal material in the superconductor.

Ward Leonard: Name given to the combination of a flywheel and motor used mostly in rolling mills and mining shafts.

Warm Supports: Thermally insulated mounts fixed to the rock and to which the vacuum vessel of the SMES is attached.

Works Cited

Barlow, T. M. (1980), in: *Proceedings of the Mechanical, Magnetic, & Underground Energy Storage Annual Contractors Review Meeting,* Washington, D.C. (U.S. DOE) Conf. 801128, pp. 299–306.

Boutman, L. J., Krock, R. H. (1974), *Composite Materials,* New York: Academic Press, Vols. 1–8.

Brumleve, J. D. (1974), *Sensible Heat Storage In Liquids,* Sandia Labs., SLL-733-0263.

Byerly, R. T. (1980), in: *Proceedings of the Mechanical, Magnetic, & Underground Energy Storage Annual Contractors Review Meeting,* Washington, D.C. (U.S. DOE) Conf. 801128, pp. 249–254.

Carruthers, R. (1962), *High Magnetic Fields,* New York: MIT Press–Wiley, pp. 307–318.

Christensen, C. B., Burrows, R. K., Shinton, Y. D. (1985), *3rd International Conference on Energy Storage for Building Heating & Cooling,* Toronto, Canada, September 22–26, 1985. Publ. by Energy, Mines & Resources, Communications Branch, Ottawa, ON, Canada. Also available from U.S. DOE (ITL-8509), pp. 416–420.

Coppa, A. P., Zweben, C. H. (1980), *Flywheel Containment Technology Assessment,* Lawrence Livermore Laboratory Report, July 1980.

Cresap, R. L., Mittlestadt, W. A. (1976), *IEEE Trans. Power Appl. Soc.* **PAS-95,** 536–541.

Crouch, D. A., Perry, E. R. (1973), *IEEE Power Eng. Soc. Meeting,* Paper C 73, pp. 213–216.

Fernandes, R. A. (1976), *Analysis of Hydrogen Cycle Peak Shaving for New York State,* Report from R & D Department, Niagara Mohawk Power Corp., Syracuse, New York.

Ferrier, M. (1970), *Low Temperature and Electric Power,* London: Pergamon Press, p. 425.

Fowler, C. M., Caird, R. S., Garn, W. B., Thomson, D. B. (1962), *High Magnetic Fields,* New York: Wiley, pp. 269–276.

Genta, G. (1985), *Kinetic Energy Storage,* London, England: Butterworths, p. 56.

Glendening, L. (1974), *Symposium on Long Term Studies: Compressed Air Storage,* Central Electricity Generating Board (UK).

Gröber, H., Erk, S. (1961), *Fundamentals of Heat Transfer,* New York: McGraw-Hill.

Guillemin, E. (1953), *Circuit Theory,* New York: Wiley, p. 261.

Gully, J. H., Estes, E. G., Walls, W. A., Weldon, W. F. (1984), *IEEE Trans. Magnetics* **MAG-20,** 203–206.

Hammel, E. F., Rogers, J. D., Hassenzahl, W. E. (1970), *Cryogenics* **10,** 186–198.

Hassenzahl, W. H. (1983), *Proc. IEEE* **71,** 1089–1098.

Hassenzahl, W. H., Baker, B. L., Keller, W. E. (1973), *Superconducting Magnetic Energy Storage Systems for Load Levelling,* Los Alamos Nat. Lab., Rept. LA-5377-MS.

Hoening, M. O. (1976), *Evaluation of Forced Flow Cooled Superconductors,* Tech. Rept. Francis Bitter National Magnet Lab., Cambridge, MA.

Holland, L. D. (1984), *IEEE Trans. Magnetics* **MAG-20,** 272–275.

Hunt, H. W. (1968), *Civil Engineering ASCE* **48** (3).

Intersociety Energy Conversion Conference Proceedings (1988), *IEEE Publications,* Vol. 2.

Johnson, A. E. (1956), *Aircraft Eng.* **46,** 187.

Kays, W. M., London, A. L. (1964), *Compact Heat Exchangers,* 2nd Ed., New York: McGraw-Hill.

Lord Kelvin (1863), *Proc. Roy. Soc. (London)* **19,** 113.

Loud, S. N. (1977), *Glass Fiber for Energy Stor-*

age Flywheel, San Francisco: Flywheel Symposium, 403.

Lupton, W. H., Robson, A. E., Warnick, W. L. (1976), *Proc. Int. Pulsed Power Conf.* (IEEE Publ. 76CH1147-8 REG5) paper no. IID-3.

MacNab, I. R., Johnson, J. L., Schreurs, J. (1980), *Wear* **80,** 259.

Marshall, R. A. (1980), *Holm Conf. on Electric Contacts,* Chicago, IL, p. 151.

Mawardi, O. K. (1982), *IEEE Trans. Magnetics* **MAG-18,** 151–156.

Mawardi, O. K. (1975), *Design of Force Free Inductive Storage Coil,* Los Alamos Nat. Lab., Report LA 5935-MS.

Mullan, E. (1977), *Design & Fabrication of a 300 kJ Superconducting Energy Storage Cell,* Westinghouse Electric Corp., Tech. Report EM 5077.

Murray, J. N., Yaffe, M. R. (1980), *Advanced Alkaline Electrolysis Cell Development,* Brookhaven National Laboratory, Tech. Report BNL 51374.

National Academy of Sciences (1976), Committee on Advanced Storage Systems, Criteria for Energy Storage R&D, Washington, D.C.

Nultall, L. J., Fernandes, R. A. (1974), *The Hydrogen Economy Miami Energy Conf.* NSF, Paper S14.

Oberg, E., Jones, F. D. (1976), *Machining Handbook,* 20th Ed., New York: Industrial Press.

Peterson, H. H., Mohan, N., Boom, R. W. (1975), *IEEE Trans Power Appl. Sys.* **PAS-94,** 1337–1348.

Petukhov, B. S. (1970), *Adv. Heat Transfer* **6,** 503–564.

Polland, R. (1970), *Engineering* **209,** 473.

Pourbeau, P. C. (1980), *Flywheel Operating on Magnetic Bearings,* Flywheel Tech. Symposium, Scottsdale, AR, October 1980.

Powell, J. R., Bezler, P. (1973), *Proceedings of Technology of Controlled Thermonuclear Fusion Experiments and the Engineering Aspects of Fusion Research,* Austin, TX, p. 358.

Rabenhorst, D. W. (1979), *Proceedings of the Mechanical and Magnetic Energy Storage Contractors—Review Meeting,* U.S. DOE Conf. 790854, pp. 407–414.

Robson, A. E., Lupton, W. H. (1975), *Proc. IEEE Sixth Symposium on Engineering Problems of Fusion Research,* (IEEE Pub. 75CH1097-5-NPS), p. 298.

Rogers, J. D. (1979), *Proceedings of Mechanical & Magnetic Energy Storage Contractors Review Meeting* U.S. DOE Conf. 790854, pp. 27–46.

Rogers, J. D., Turner, R. D. (1980), *1 GWh Diurnal Load Leveling Superconducting Magnetic Energy Storage,* Los Alamos Natl. Lab. Report LA 7885 MS.

Rogers, J. D., Schermer, R. I., Miller, B. L., Hauer, J. F. (1983), *Proc. IEEE* **71,** 1099–1107.

Salzano, F. J., Burger, J. M., Lewis, P. A., Isler, R. J. (1974), *Intersoc. Energy Convers. Energy Conf., Proc.,* San Francisco, CA, August 1974, pp. 428–434.

Sapowith, A. A., Handy, W. E. (1981), *Evaluation and Design of Woven Composite Materials* Lawrence Livermore Lab. Tech. Report UCRL 151415.

Schmidt, F. W., Wilmott, A. I. (1981), *Thermal Energy Storage & Regeneration,* New York: Hemisphere Publ. Corp.

Schulten, R., Kraemer, H., Arndt, E., Barnest, H., Schlenker, H. V., van d. Decken, C. B., Kuegler, K. (1974), Am. Nucl. Soc. Topical Meeting, Tennessee, May 7–10, 1974, Publ. AEC(CONF.-740501), pp. 378–396.

Slemon, G. R., Straughen, A. (1980), *Electrical Machines,* Reading, MA: Addison-Wesley.

SMES Review Panel (1976), Proc. DOE, December 13, 1976, Washington, D.C.

Smith, P. F., Lewin, J. D. (1970), *Part. Accel.* **1,** 155–172.

Stephens, T. (1976), *Proceedings of Engineering Foundations Conference on Energy Storage,* Pacific Grove, CA.

Stodola, A. (1934), *Dampf- und Gas-Turbinen,* Berlin: Springer Verlag.

Strickland, G., Reilly, J. J., Wiswell, R. H. Jr. (1974), *The Hydrogen Economy Miami Energy (THEME) Conf. Proc.,* Miami Beach, FL, March 1974, pp. 611–620.

Sturm, J. C. F. (1861), *C. R. Acad. Sci.* **13,** 1046.

Stys, S. Z. (1983), *Proc. IEEE* **71,** 1079–1085.

Szego, J., Schmidt, F. W. (1978), *J. Heat Trans. Trans. ASME* **100,** 740–742.

Thaler, G., Wilcox, M. (1966), *Electric Machines; Dynamics and Steady State,* New York: Wiley, p. 56.

Timoshenko, S. (1951), *Theory of Elasticity,* New York: McGraw-Hill Book Co.

Trost, T. F., Garisson, P. E., Burkes, T. R. (1976), *Proc. Int. Pulsed Power Conf.* (IEEE Publ. CH1147-8 REG5), paper no. IID-1.

Wisconsin Superconducting Energy Storage (1974), Vol. I, (1976), Vol. II, University of Wisconsin Publications.

Yamamoto, M., Ikeda, M., Tanaka, Y., Masumoto, K., Ishigokha, T., Tsukamoto, O. (1986), *IEEE Trans. Magnetics* **MAG-23,** 557–560.

3

Strong Interaction

Frithjof Karsch and Werner Vogelsang

3.1 Overview 97
3.2 Eightfold Way and Quark Model 100
3.3 Quantum Chromodynamics 101
3.3.1 Asymptotic Freedom 102
3.3.2 Confinement 103
3.3.3 Chiral Symmetry 105
3.3.4 Glueballs 105
3.4 Experimental Evidence for QCD 106
3.4.1 High-Energy Collisions 106
3.4.2 QCD in Electron–Positron Annihilation 107
3.4.3 Deep-Inelastic Scattering 108
3.4.4 Hadron–Hadron Scattering 111
3.5 Phases of Strongly Interacting Matter 112
3.5.1 Quarks and Gluons at High Temperature 113
3.5.2 Quarks and Gluons at High Baryon Number Density 114
3.5.3 QCD at High Energy 115
References 116

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

3.1
Overview

In this chapter, we give an overview of our theory of the strong interactions, *quantum chromodynamics* (QCD) and its properties. We also briefly review the history of the study of the strong interactions, and the discoveries that ultimately led to the formulation of QCD.

The strong force is one of the four known fundamental forces in nature, the others being the electromagnetic, the weak, and the gravitational forces. The strong force, usually referred to by scientists as the *strong interaction*, is relevant at the subatomic level, where it is responsible for the binding of protons and neutrons to atomic nuclei (see Figure 3.1). To do this, it must overcome the electric repulsion between the protons in an atomic nucleus and be the most powerful force over distances of a few femtometers ($1\,\text{fm} = 1\ \text{femtometer} = 1\ \text{fermi} = 10^{-15}\,\text{m}$), the typical size of a nucleus. This property gave the strong force its name.

The first quantitative theory of the strong interactions was proposed by Yukawa (1935). Yukawa postulated that the strong force arises from the exchange of new particles, now called the *pions*, between protons and neutrons. From the known range of the strong interaction, he could estimate the mass of these particles. The pions were, indeed, discovered by Lattes *et al.* (1947). In the following years, many new strongly interacting particles were discovered at new particle accelerators as well as in cosmic ray showers. They are collectively referred to as *hadrons*. It was found that hadrons could be grouped into two classes by whether or not they carried a conserved quantum number called the *baryon number*. Particles that carry baryon number, examples of which are the proton and neutron, are called *baryons*. Among the particles with vanishing baryon number, known as *mesons*, are the pions. Some of the discovered hadrons showed an unexpectedly long (strange) lifetime, like the Λ baryon which had already been observed in 1947 in cosmic ray showers (Rochester and Butler, 1947).

The discovery of the large array of strongly interacting particles implied that Yukawa's theory could not be the fundamental theory of the strong interactions. The pursuit of finding an underlying order and understanding the regularities observed in experiment eventually led to the proposal that there are only a few truly fundamental particles of the strong interactions, of which all hadrons are composed. This proposal was made independently in 1964 by Gell-Mann (1964) and Zweig, and Gell-Mann coined the name

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

quarks for these new particles. To take into account the observed systematics of baryons and mesons, one had to introduce different types, or *flavors*, of quarks. The basic constituents of the nucleus – the proton and the neutron – are built up from three quarks with two different flavors called *up* (*u*) and *down* (*d*) (see Figure 3.1). Mesons consist of a quark and an antiquark. The strangeness of the Λ particle (Gell-Mann, 1956) could be explained through the introduction of a third quark flavor – the strange quark (*s*). Another observation that became crucial for the further development of strong interaction theory was made by Greenberg (1964) and Han and Nambu (1965) soon after the introduction of quarks: to satisfy the Pauli exclusion principle for baryons such as the Δ^{++} or the Ω^{-}, which are made up of three quarks of the same flavor and spin orientation, the spin-$\frac{1}{2}$ quarks had to carry a new quantum number, later termed *color*. It was proposed that quarks came in three different colors.

Quarks were originally introduced simply on the basis of symmetry considerations. A modern rendition of Rutherford's experiment then showed that quarks are real (Bloom *et al.*, 1990). This experiment is the *deep-inelastic scattering* (DIS) of electrons (or, later, muons) off the nucleon, a program that was started in the late 1960s at the Standford Linear Accelerator Center (SLAC). The early DIS results compelled an interpretation as elastic scattering of the electron off pointlike, spin-$\frac{1}{2}$, constituents of the nucleon, carrying fractional electric charge (Bjorken, 1969). These constituents, called *partons* by Feynman, were subsequently identified with the quarks.

In 1974, a new meson, soon called the J/ψ, was observed simultaneously in experiments at Brookhaven National Laboratory (BNL) (Aubert *et al.*, 1974) and at SLAC (Augustin *et al.*, 1974). Its surprisingly long lifetime made it clear that there was yet another quark flavor – now called the *charm* quark (*c*). Currently, we know six different quark flavors. In addition to the *u, d, s*, and *c* quarks, the very heavy bottom (*b*) (Herb *et al.*, 1997) and top quarks (*t*) (Abe *et al.*, 1995) have been discovered experimentally in 1977 and 1995, respectively, at the Fermi National Accelerator Laboratory (Fermilab). There is currently no evidence of the existence of further quark flavors.

Remarkably, quarks have never been observed in isolation, or as free particle states, like those familiar in an atom or the proton. They only seem to exist bound inside hadrons or in larger entities called *quark matter*, which is presumed to have existed in the early universe and may still exist in the interior of compact stars. This striking phenomenon is known as *confinement*. It is clear that a true theoretical understanding of the strong interactions requires a quantitative explanation for the confinement of quarks, which has remained elusive till today.

QCD, the modern theory of strong interactions, is a quantum field theory. It was formulated by Fritzsch, Gell-Mann, and Leutwyler (1973). QCD rests on a number of the concepts and experimental findings just mentioned: QCD specifies the interactions among the quarks introduced in the quark model of Gell-Mann and Zweig. It turns the relatively ad hoc color quantum number into the conserved charge of a local gauge theory, identical in spirit to the electric charge of quantum electrodynamics (QED). And, finally, the features seen in the DIS experiments paved the way to QCD as a viable contender for the theory of the strong interactions. The experimental data pointed to a theory that would have the remarkable

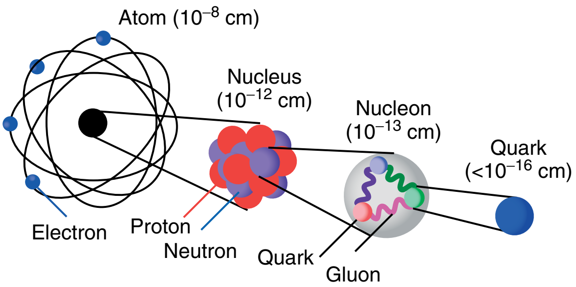

Fig. 3.1 The substructure of matter made of strongly interacting quarks and gluons. They form protons and neutrons that bind through the strong force to form the atomic nucleus.

property of becoming weakly coupled at large momentum transfer or, equivalently, short distances. QCD precisely has this property of *asymptotic freedom* (Gross and Wilczek, 1973).

QCD has been constructed in close analogy to QED, which very successfully describes the electromagnetic forces among electrically charged particles in terms of the exchange of a massless particle, the photon. The force carrier in QCD is called the *gluon*. The structure of the theory demands that there are actually eight gluons. Gluons are responsible, for example, for binding quarks together to form hadrons. Direct evidence for the existence of gluons was found in 1979 at the Deutsches Elektronen Synchrotron (DESY) in Hamburg, Germany, through the discovery of so-called three-jet events (Barber *et al.*, 1979).

The asymptotic freedom of QCD at short distances makes the theory in this regime accessible to standard perturbative calculational techniques and allows for direct comparisons with experiments performed at high-energy accelerators. This has allowed to subject QCD to stringent tests, and to validate it as the correct theory of the strong interactions. Perturbative techniques fail, however, when the theory is probed at lower energies. It is in this so-called nonperturbative regime that the binding of quarks and gluons to hadrons, and that of protons and neutrons to heavier nuclei, occurs. It has so far not been possible to understand and derive the masses of these bound states in detail from the basic Lagrangian of QCD. Among the important phenomena QCD shows here, is, in particular, the confinement of quarks that we have mentioned earlier. A consequence of this is the rich spectrum of hadrons arising as bound states of quarks and gluons. To calculate the properties of these states, for example, their masses, requires calculational techniques that go beyond perturbation theory. A method that allows studies of some aspects of nonperturbative QCD was proposed by Wilson (1974), who introduced a formulation of QCD on a regular discrete four-dimensional space–time lattice. *Lattice QCD* is one of many ways to regularize a quantum field theory. It has the advantage that it makes QCD accessible to numerical calculations. Creutz (1980) could show through numerical calculations that, in an approximation in which only the gluon sector of QCD is taken into account, confinement does indeed occur. Today, lattice QCD calculations are performed mostly on large-scale computers, and many important insights into QCD have been gained in this way.

Another important phenomenon in QCD is *chiral symmetry breaking*, which is believed to be responsible for certain aspects of the observed hadron spectrum, for example, the small value of the pion mass compared to other hadron masses. In the limit of vanishing up and down quark masses, the Lagrangian that describes QCD possesses a certain symmetry in the *u, d* quark sector, known as *chiral symmetry*. This name comes from the fact that one can define left- and right-handed parts of the quark fields, which become independent of one another in the QCD Lagrangian, once the quark masses are neglected. However, it turns out that the chiral symmetry is broken spontaneously by paired quarks, so-called quark condensates, in the QCD vacuum. This again is a nonperturbative feature of QCD that has been verified to exist by lattice calculations. Spontaneous breaking of a symmetry leads to the existence of massless (Nambu–Goldstone, 1961) bosons, in this case, the pions. Since the quark masses are not zero, the pions are not in fact massless, but remain much lighter than all the other hadrons. As the quark masses are small, one can formulate a systematic perturbative expansion that is valid in the vicinity of the massless limit. This *chiral perturbation theory* (Weinberg, 1967) leads to many predictions in the low-energy domain of QCD.

The spontaneous breaking of chiral symmetry as well as confinement are striking nonperturbative aspects of the theory of strong interaction, which are absent in the asymptotically free regime of QCD that can be addressed perturbatively. It has therefore been speculated that QCD has different phases that are separated by phase transitions very much like the ones known to occur in many-body systems. In fact, from QCD one should, at least in principle, be able to derive the properties of bulk matter formed from strongly interacting particles, for example, the properties of nuclei, properties of compact stars as well as of the hot and dense matter that existed in the early universe. In the latter case, there exists strong evidence from heavy-ion collision experiments that at temperatures about 100 000 times larger than that in the center of our sun, that is, under conditions that existed in the early universe at a time of about 10^{-5} seconds after the Big Bang, a hot and dense form of strongly interacting matter can exist in which chiral symmetry is restored and quarks and gluons are deconfined. The exploration of the properties of this so-called quark–gluon plasma is being pursued intensively in relativistic heavy-ion experiments, lattice QCD calculations, as well as analytic calculations that make use of modern perturbative techniques.

3.2
Eightfold Way and Quark Model

Prior to the quark model, it was realized independently by Gell-Mann and Ne'eman (1961) that the observed regularities in the masses, interactions, spin, and charge of hadrons could be explained by grouping them according to rules dictated by the mathematical special unitary group in three dimensions, SU(3). The observed hadron states appear in higher dimensional representations of this group. The proton and neutron, for instance, are members of an octet representation along with other hadrons of similar properties, and a meson octet contains the pions, among others. This organization scheme, which has been called the *eightfold way* (Ne'eman, 1961) owing to the presence of octet representations, is illustrated in Figure 3.2. It was of crucial importance that it predicted the existence of a hitherto unknown hadron, the Ω^-, which was subsequently observed experimentally at BNL (Barnes *et al.*, 1964).

Since the hadrons were related to higher dimensional representations of the SU(3)

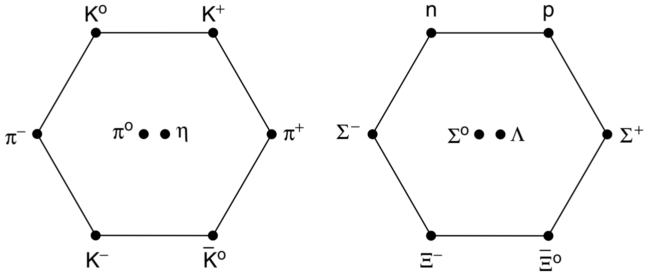

Fig. 3.2 Eightfold way: meson and baryon octets.

group, it was natural to ask what the particles corresponding to the fundamental representation might be. As this is mathematically a triplet representation, there would need to be three elementary particles. These were proposed by Gell-Mann (1964) and Zweig and are known as *up* (u), *down* (d), and *strange* (s) quarks. The quark model explains the eightfold way: the baryons consist of three quarks, while the mesons are made up of a quark and an antiquark. Mathematically, the baryons are products of three-quark triplet representations, while the mesons are the product of a triplet and an antitriplet representation. We have listed the quark content of some hadrons in Table 3.1. Each quark is characterized by a baryon number of $\frac{1}{3}$, spin $\frac{1}{2}$, an electric charge of $-\frac{1}{3}$ or $+\frac{2}{3}$, and a new quantum number called *color* (Greenberg, 1964; Han and Nambu, 1965), which comes in three different variants, say red, blue, and green. This color is the strong charge of a quark, similar to the positive or negative charge of a particle in electrodynamics. The hadrons observed in experiment are color singlets, that is, are white combinations of the colored quarks.

Over the last 50 years, studies at high-energy particle accelerators have led to the conclusion that there exist six different quark flavors (up, down, strange, charm (c), top (t), and bottom (b)). The three additional c, b, t quarks are much heavier than the originally proposed u, d, s quarks. The properties of quarks are summarized in Table 3.2. The six quarks have been grouped into three generations of two quarks with electric charge $\frac{2}{3}$ and $-\frac{1}{3}$, respectively. This ordering principle goes beyond the framework of QCD and indicates the role of QCD in the standard model of particle physics (Glashow, 1961) that unites the strong, electromagnetic, and weak forces. Within the standard model, precision studies of the decay width of the Z-boson, a carrier of the weak force, have established that there are just three generations of particles. Dedicated searches have not produced any evidence for the existence of further quark generations or for a substructure of quarks. We finally note that the quarks are the only known elementary particles that participate in all of the four fundamental interactions.

3.3
Quantum Chromodynamics

QCD incorporates the basic properties of quarks in the framework of a quantum field theory. The structure of a local quantum field theory, its local and global symmetries, have important consequences for the interaction among particles as well as for their properties, which can be tested experimentally.

Tab. 3.1 Some hadrons and their properties. Hadron masses are given in units of giga electron volt divided by the velocity of light squared (GeV/c^2)[1 GeV = 1 billion electron volts]. For further information, see (Yao *et al.*, 2006).

Hadron	Quark content	Mass (GeV/c^2)	Electric charge	Spin
Mesons				
Pion (π^+)	$u\bar{d}$	0.139	1	0
Kaon (K^-)	$s\bar{u}$	0.494	−1	0
Rho (ρ^+)	$u\bar{d}$	0.775	1	1
J/ψ	$c\bar{c}$	3.097	0	1
Upsilon (Υ)	$b\bar{b}$	9.460	0	1
Baryons				
Proton (p)	*uud*	0.938	1	$\frac{1}{2}$
Neutron (n)	*udd*	0.940	0	$\frac{1}{2}$
Lambda (Λ)	*uds*	1.116	0	$\frac{1}{2}$
Xi$_c$ (Ξ_c^+)	*usc*	2.468	1	$\frac{1}{2}$
Lambda$_b$ (Λ_b)	*udb*	5.624	0	$\frac{1}{2}$

QCD has been constructed in close analogy to QED. The basic degrees of freedom are the quarks, and the interaction among them is transmitted by massless bosons, the gluons. In contrast to QED, the three color degrees of freedom of the quarks, however, suggested to formulate QCD in terms of a nonabelian gauge group, the special unitary group SU(3)$_c$. We have used the subscript *c* here to distinguish it from the SU(3) that we used earlier for the quark flavors. Like in QED, the interaction among quarks and gluons is determined by the requirement of local gauge invariance: the theory is invariant under local transformations that rotate the charges of the quarks in the color space defined by the SU(3)$_c$ gauge group. This controls the basic properties of QCD. A direct consequence of this is that gluons not only interact with quarks but also have a self-coupling and interact directly among themselves. This is very different from QED. The existence of these specific self-coupling terms has been verified experimentally through the observation of high-energy jets (see Section 3.4). The self-interaction of gluons is responsible for much of the complexity of QCD and the richness of its phenomena.

3.3.1
Asymptotic Freedom

One of the peculiar consequences of the self-interaction of gluons is that QCD becomes *asymptotically free* at short distances, which means that the coupling among quarks and gluons becomes weaker with decreasing distance. This property of QCD was discovered by Gross and Wilczek (1973). The precise behavior of the coupling between quarks and gluons as a function of distance (or momentum transfer between them) follows from the renormalization properties of QCD. One finds that the coupling vanishes as an inverse logarithm of the momentum

Tab. 3.2 The six quark flavors. All quarks are fermions that have spin $\frac{1}{2}$, baryon number $\frac{1}{3}$, and an additional color charge red, blue, or green. For each quark, there exists an antiparticle (antiquark) with opposite quantum numbers. Their properties are given in Yao *et al.* (2006).

Flavor	Mass (GeV/c^2)	Electric charge
Up (u)	$0.0015 - 0.003$	$\frac{2}{3}$
Down (d)	$0.0025 - 0.0055$	$-\frac{1}{3}$
Strange (s)	0.095 ± 0.025	$-\frac{1}{3}$
Charm (c)	1.25 ± 0.09	$\frac{2}{3}$
Bottom (b)	4.2 ± 0.07	$-\frac{1}{3}$
Top (t)	174.2 ± 3.3	$\frac{2}{3}$

transfer. This behavior has been exquisitely verified by a variety of measurements and has been confirmed in lattice QCD calculations.

Large momentum transfer may be realized in particle collisions at high energy or in media at very high temperature where the average momentum transfer in a reaction is proportional to the temperature of the medium (see Section 3.5). When the coupling becomes weak, processes in QCD can often be analyzed well in terms of *perturbation theory*. This led to numerous predictions for the properties of high-energy processes that have been tested experimentally (see (Bethke, 2007; Ellis, Stirling and Webber, 1996) and Section 3.4). The good agreement found between theoretical predictions and experimental results found here cemented our belief that QCD is the correct theory of the strong interaction. A recent compilation of determinations of $\alpha_s(Q)$ is given in Figure 3.3.

3.3.2
Confinement

As one can see from Figure 3.3, the QCD coupling $\alpha_s(Q)$ appears to become large for small momentum transfer. However, the result shown in the figure is based on a perturbative calculation and is therefore no longer self-consistent when $\alpha_s(Q)$ becomes large. Nonetheless, it is consistent with the fact that QCD phenomena become strongly coupled and confining. The analysis of heavy-quark (c, b) bound states and the sequence of their excitations suggests that the potential between quarks rises linearly at large distances. This gave rise to nonrelativistic potential models that describe the spectrum of heavy-quark bound states like the J/ψ and its excitations and the Υ family. The origin of the linearly rising potential again can be attributed to the self-interaction among gluons, which allows them to build up an almost one-dimensional color flux tube (string) that transmits the interaction between quarks over large distances. Thus, it would cost an infinite amount of energy to completely separate a quark antiquark pair. The qualitative picture behind confinement is that after separation by a certain distance the energy contained in the field between the quark and the antiquark becomes so large that a further quark–antiquark pair can be created from the vacuum, so that two color singlet hadrons – for

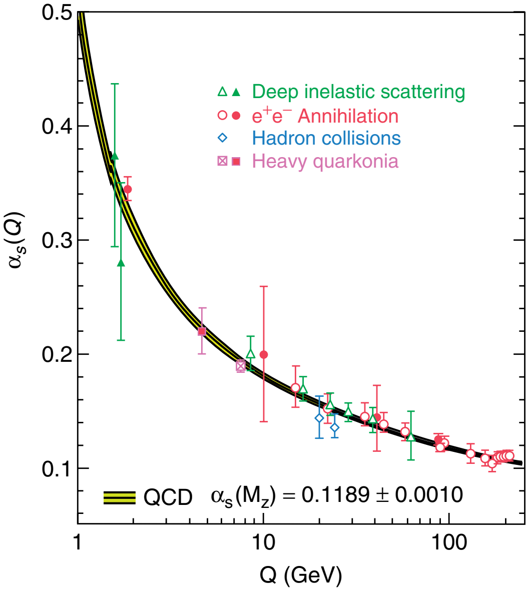

Fig. 3.3 Summary of experimental results on the QCD running coupling constant $\alpha_s(Q)$ as function of the momentum transfer, Q, in an elementary interaction. To determine the running coupling, experimental data are compared with calculations performed in leading (open symbols) and next-to-leading (filled symbols) order of QCD perturbation theory. The figure also includes a purely theoretical data point obtained from lattice calculations (boxed square). The figure is taken from Bethke (2007).

example, pions – can be formed. This explains why free quarks are not observed experimentally and why observed hadrons only appear as color singlet states.

Unlike the short-distance properties of QCD, which can be studied using perturbative techniques, the long-distance properties are entirely nonperturbative. An appropriate tool for performing calculations in this nonperturbative regime is provided by lattice QCD, a formulation of QCD on a discrete four-dimensional space–time grid (lattice) that makes the analysis of large-distance properties of QCD, or equivalently of processes with low momentum transfer, accessible to numerical calculations. Quarks occupy the sites of the lattice, while the gluons, as the mediators of the strong force, correspond to the links from one lattice site to another. Lattices used for modern calculations typically have a few million sites. Currently, this allows to analyze physics in a spatial volume of about (10–20) fm^3 and to resolve details on a distance scale of about 0.05 fm. Ideally, one would like the volume to be as large as possible, and the distance between the lattice sites to be as small as possible, to match the true continuum. This requires large computational resources. Progress in lattice QCD calculations thus is closely coupled to the rapid development in computer technology. Lattice QCD allows to calculate the potential between static quark sources. These calculations indeed demonstrate that the potential between quarks increases linearly at large distances, as shown in Figure 3.4(a). Lattice QCD also allows a direct calculation of hadron masses. A comparison of lattice calculations for the light hadron spectrum with experimental results is shown in Figure 3.4(b). Although this calculation has been performed in an approximation that only takes into account the contribution of valence quarks and gluons (quenched QCD), the agreement with experimental results is striking. It is expected that this agreement will further improve once calculations that also include sea quark contributions become feasible and accurate enough.

3.3.3
Chiral Symmetry

If some, say n_f, of the quark flavors had vanishing masses, QCD would have an additional exact symmetry property. It would be possible to interchange different spin components (chiralities) of the quark fields without changing physical properties of the theory. In fact, the left and right chiral components of quark fields could be transformed independently in flavor space by a nonabelian rotation with $\mathrm{SU}(n_f)$ matrices. This chiral symmetry is broken owing to the nonvanishing quark masses. However, this explicit breaking is rather mild for the two lightest (up and down) quarks, which have a very small mass. The QCD Lagrangian thus has an almost exact $\mathrm{SU}(2)$ chiral symmetry for each of the two quark chiralities.

A symmetry of a physical system is said to be *broken spontaneously* when its ground state (or, vacuum) does not obey the symmetry. The chiral symmetry of QCD is spontaneously broken by a nonvanishing *chiral quark condensate* in its vacuum, which connects left and right quark chiralities. If an exact global symmetry is spontaneously broken, it gives rise to massless particles in the theory – the Goldstone bosons. If the symmetry is only almost exact, these Goldstone bosons will acquire a mass, but typically remain light. The fact that the light pseudoscalar pions have a mass that is small on the scale of all the other hadron masses suggests that they are the Goldstone bosons arising from the spontaneous breaking of chiral symmetry in QCD. Indeed, a nonvanishing chiral condensate is found in lattice QCD calculations. As chiral symmetry is broken explicitly in QCD owing to the presence of small quark masses, the pions have a mass. There are many consequences of this spontaneous symmetry breaking and the small perturbations that arise owing to the small, nonzero up and down quark masses. They can be analyzed systematically in terms of an expansion in the small quark masses known as *chiral perturbation theory* (Weinberg, 1967). Among these results is the famous (Gell-Mann–Oakes–Renner, 1968) relation that connects the chiral condensate and light quark masses with the pion mass and the pion decay constant.

3.3.4
Glueballs

As discussed above, an essential property of QCD, which distinguishes the strong force from the electromagnetic one, is that its force carriers can interact among themselves. As a consequence of this, even in the absence of quarks, the pure-gauge sector of QCD already is a complex nontrivial theory that shows many of the features of full QCD. It is asymptotically free and leads to a linearly rising potential (confinement) between static external color sources. Self-interacting gluons should also be able to form bound states, so-called glueballs. QCD predicts that in addition to the experimentally well-explored hadron spectrum there should be further bound states, predominantly built up from gluons, with no valence quarks. These glueball states have been analyzed extensively in numerical studies of QCD on the lattice, and quite accurate predictions exist for the masses of these states. Experimentally they are, however, difficult to identify as the glueball states can mix with

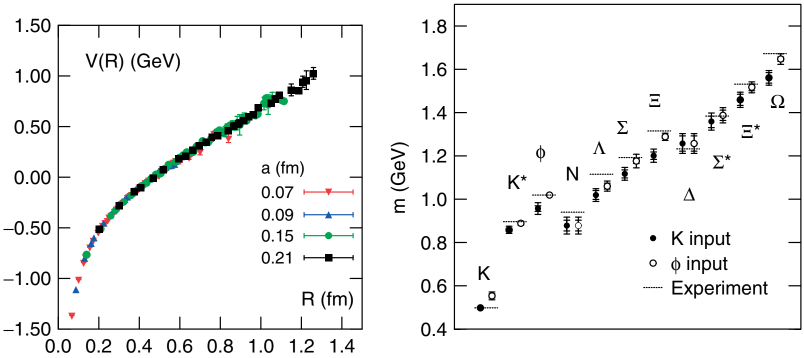

Fig. 3.4 The potential energy between a pair of static quark and antiquark sources as a function of their separation (a). Shown is the result of a lattice calculation performed for several values of the discretization length scale, that is, the lattice spacing "a" (from Cheng *et al.* (2008)). (b). The result of a spectrum calculation for the masses of light quark bound states (from Aoki *et al.* (2003)).

ordinary mesons with the same quantum numbers. Large experimental programs are being set up at the Thomas Jefferson National Accelerator Facility (Jefferson Lab) and the Facility for Antiproton and Ion Research at the Gesellschaft für Schwerionenforschung (GSI FAIR) to search for glueballs.

3.4
Experimental Evidence for QCD

3.4.1
High-Energy Collisions

Experiments at particle accelerators carried out over the past three decades have established that QCD is the correct theory of the strong interactions (Yao *et al.*, 2006; Bethke, 2007; Ellis, Stirling and Webber, 1996). Particle accelerators allow the collision of high-energy particles such as electrons, protons, or heavier atomic nuclei. One can, for example, accelerate a beam of particles and scatter it off a stationary target of other particles. This is known as a fixed-target experiment. Higher collision energies are typically reached when two separate accelerated beams of particles are collided. The result of individual particle collisions can be observed and reconstructed using arrays of detector systems around the collision point. These detectors record, for example, the species and momenta, or perhaps decays, of particles receding from the collision.

The basic unit that characterizes accelerator energies is the electron volt (eV), the kinetic energy acquired by an electron accelerated through a potential difference of 1 V. Major technological breakthroughs have been made to achieve the acceleration of particles to ever higher energies. In the 1950s, it became possible to accelerate protons to energies of a few billion electron volts (one 1 billion eV = 1 giga electron volt = 1 GeV), at the cosmotron machine at BNL. Soon after, new techniques allowed to reach proton energies of 25 GeV at the proton synchrotron at the European high-energy physics laboratory CERN, and of 33 GeV at Brookhaven's alternating gradient synchrotron. At such energies, a proton with its rest mass of 938 MeV/c^2 is already less than 0.1% away from the speed of light. In many of today's accelerators, energies are even much higher. At the DESY in Hamburg, Germany, one routinely studied head-on collisions of protons and electrons with a collision energy of about 300 GeV. The Tevatron at the Fermi National Accelerator Laboratory in Batavia, Illinois, has colliding beams of protons and antiprotons (their antimatter partners) at a total energy of nearly 2000 GeV. Finally, the *Large Hadron Collider* (LHC) at CERN, whose construction has just been completed in 2008, will have proton–proton collisions at 14 000 GeV.

In the study of the strong interactions, essentially all types of particle collisions have played an important role. Electrons (e^-) and their antimatter partners, positrons (e^+), do not participate in the strong interactions. The creation of strongly interacting particles in the annihilation of electrons with positrons therefore offers a particularly pristine environment for studying QCD. The experiments that first paved the way to QCD were performed in scattering of electrons off protons in the process known as *deep-inelastic scattering DIS*. Finally, collisions of two hadrons, such as two protons or a proton and an antiproton, has brought further important information about the strong interactions. In the following, we briefly review some of the most important insights into QCD that have been brought by these three basic types of collisions.

3.4.2
QCD in Electron–Positron Annihilation

A lot has been learned about QCD from measurements at high-energy electron–positron colliders. A graphical representation of the production of strongly interacting particles in e^+e^--annihilation is given in Figure 3.5. The fundamental underlying reaction is $e^+e^- \to \gamma \to q\bar{q}$, that is, the electron and positron annihilate into a photon (γ), which then decays into a quark–antiquark pair. If the electron–positron collision happens at energies in excess of several tens of giga electron volts, the intermediate particle will also often be the Z boson of the weak interactions, instead of the photon γ. The process by which the conversion of the color-charged quark and antiquark into colorless hadrons happens, involves QCD in the confinement regime described in Section 3.3 and is known as *hadronization*.

The simplest strong interaction observable in e^+e^--annihilation is the total rate for the production of hadrons such as pions,

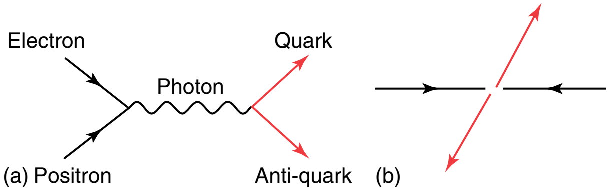

Fig. 3.5 Lowest-order diagram for the production of strongly interacting particles in e^+e^--annihilation. (a) shows a Feynman diagram, used in the theoretical computation of the process and (b) depicts the situation in an actual experiment where the electron and the positron collide head-on. The wiggly line represents the intermediate particle, photon, or Z-boson, created in the annihilation of the electron–positron pair. Because of confinement, the quark and the antiquark will eventually hadronize, that is, convert to colorless hadrons that are seen in the detector.

kaons, and so on. For this, one counts any event with hadrons in the final state. It is customary to normalize the rate to the one for the simple QED annihilation into muons, $e^+e^- \rightarrow \mu^+\mu^-$. The corresponding ratio is denoted by R_{QCD}. At sufficiently high energy, the process $e^+e^- \rightarrow q\bar{q}$ shown in Figure 3.5 becomes the main source for the generation of hadrons. One then finds $R_{\text{QCD}} = 3 \cdot (\frac{4}{9} + \frac{1}{9} + \frac{1}{9})$, where the factor 3 is the number of colors that quarks have in QCD (see Section 3.1), and where the terms in brackets are the squares of the fractional quark electric charges for the light up quarks (fractional charge $\frac{2}{3}$), down quarks $(-\frac{1}{3})$, and strange quarks $(-\frac{1}{3})$ (see Table 3.2). If the collision energy is above the threshold for charm production, another term $\frac{4}{9}$ will need to be added, and if bottom quarks can be produced, an additional $\frac{1}{9}$. The early e^+e^--annihilation experiments performed in the mid-to-late 1970s had already been able to establish the need for the color factor 3, as well as for the factor associated with the fractional quark charges that increased by a step when the charm threshold was crossed. However, the simple form of R given above is only an approximation and is, for example, modified by the radiation of energetic gluons from the $q\bar{q}$ pair. Over the years, it became possible to develop a controlled framework, known as *QCD perturbation theory* that allows very precise calculations of R_{QCD} and in fact of many other observables in high-energy scattering.

At high e^+e^- collision energies, the produced quark and antiquark will also be very energetic, and when they recede from one another in opposite directions, they are expected to each produce a spray of hadrons, termed *jets* (Sterman and Weinberg, 1977). Again, such jet structures had already been seen in the early experiments. The underlying process $e^+e^- \rightarrow q\bar{q}$ predicts a particular pattern of angular distribution of these jets with respect to the initial beams, which depends on the spin of the produced quark and antiquark. The distribution found in experiment has confirmed that quarks and antiquarks have spin-$\frac{1}{2}$, as they should have if QCD is right. The use of jet final states has gone yet further. If the quark or antiquark radiates an energetic gluon, one typically expects a third jet in the detector. Three-jet events were first seen in 1979 at the PETRA machine at DESY (Barber *et al.*, 1979), marking the discovery of the force carriers of QCD, the gluons. Again, detailed studies of angular distributions in three-jet events confirmed that gluons have spin 1, as in QCD. If yet another hard gluon is radiated, one may have four jets in the event. Some of these four jets will arise from a gluon itself radiating a gluon. This is possible in QCD thanks to the fact that the gluons themselves are color charged and hence have a self-coupling, as described in Section 3.1. This markedly distinguishes QCD from abelian theories like QED. Experimental studies of four-jet events have confirmed the presence of the gluon self-interaction. They have also allowed to verify the color charges that quarks and gluons have in QCD.

Finally, these, and related, studies in e^+e^--annihilation have produced very precise measurements of the strong coupling constant and have verified its running with the energy scale predicted by QCD, as shown in Figure 3.3.

3.4.3
Deep-Inelastic Scattering

Electron–positron annihilation has been of great importance in subjecting QCD to precision tests. The DIS process (Bloom *et al.*, 1990), a modern-day Rutherford

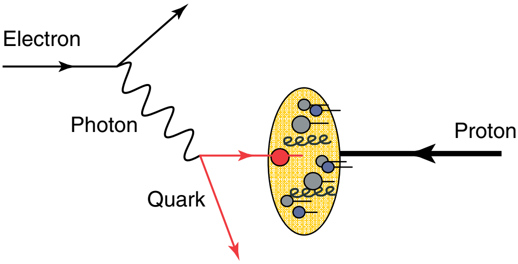

Fig. 3.6 Diagram for deep-inelastic scattering of an electron off a proton, seeing a quark inside the proton.

scattering, has been pivotal in identifying the asymptotically free QCD as our theory of the strong interactions. It also gave us fundamental insights into the internal structure of the proton through the discovery of quarks as its fundamental building blocks. Figure 3.6 shows a graphical representation of the DIS process. An electron scatters off the proton by exchange of a photon with very high momentum transfer. The electron is deflected at a large angle, and the proton breaks apart in the course of this interaction. In typical DIS experiments, the momentum transfer is so large that, according to Heisenberg's uncertainty principle, one probes distances from 0.2 down to 0.002 fm or so, far smaller than the size of the proton.

In the parton model (Bjorken, 1969), the DIS process is viewed as a scattering of the electron off a quark inside the proton, as indicated in Figure 3.6. The struck quark carries a fraction x of the proton's momentum. As such, DIS allows to map out the momentum distributions of quarks in the proton. As discussed in Section 3.2, in the simple quark model, the proton is made of just two up quarks and one down quark. However, a relativistic probe – which, in effect, takes a snapshot of the proton – is expected to also reveal quark–antiquark pairs and gluons inside the proton. These must be present because of QCD interactions that, for example, hold the proton together. In the context of high-energy scattering, all constituents of the proton are collectively referred to as *partons*. The parton model predicts that the distributions of quarks, antiquarks, and gluons depend only on their momentum fraction x, and not on the momentum transfer in the DIS process. The early DIS data were consistent with this so-called scaling behavior. However, it was soon realized that QCD predicted a gentle logarithmic dependence on momentum transfer, known as *scaling violation* (Christ, Hasslacher and Mueller, 1972). The effect of scaling violation can be reliably calculated in QCD and has subsequently been verified with great precision in fixed-target experiments at SLAC and CERN, and at DESY's collider HERA. In this way, DIS has led to one of the best direct tests of QCD (Bethke, 2007). Another early finding of the DIS program was that the quarks and antiquarks carry roughly half of the proton's momentum. The conclusion was that the other half must be carried by gluons (which are not directly seen by the photon since they do not carry electric charge) to be regarded as first indirect evidence for the existence of the gluons of QCD. Figure 3.7 represents the information on the parton distribution functions of the proton extracted from DIS measurements by the H1 and ZEUS experiments at HERA Cooper-Sarkar and Gwenlan. They are shown as functions of the fraction x of the proton's momentum carried by the parton. One can see that two

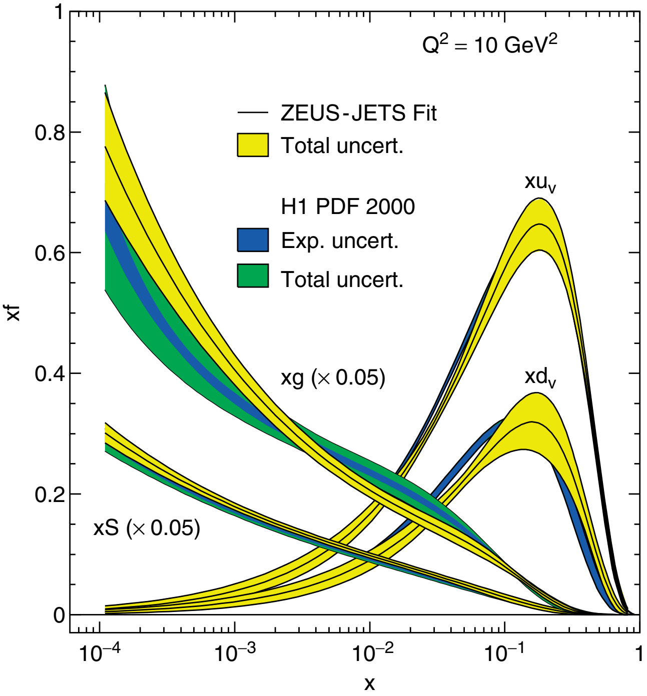

Fig. 3.7 Compilation of parton distribution functions extracted from DIS measurements at HERA. Taken from Cooper-Sarkar and Gwenlan.

distributions, referred to as u_V and d_V have their maximum at rather high x. These are the so-called valence distributions, which represent the two up quarks and the one down quark in the proton as specified by the quark model. For example, one can see that indeed the u-valence distribution is roughly twice as large as the d-valence one. The other distributions, the gluon distribution $g(x)$ and the antiquark distribution $S(x)$, are mostly concentrated at small momentum fractions. This reflects the fact that, in contrast to the valence quarks, it is quite unlikely that gluons or antiquarks carrying a large momentum fraction are found in the proton. All distributions shown have an uncertainty band, which is indicative of the uncertainties in the experimental measurements.

Starting from the late 1970s, DIS measurements were performed with spin-polarized electrons and protons. The electrons and the proton are both spin-$\frac{1}{2}$ particles, and one can set up DIS measurements in such a way that their spins have fixed relative orientations in the scattering, for example, parallel or antiparallel. An important question that one may address in this way is how the quarks, antiquarks, and gluons inside the proton conspire to provide the spin of the proton. After all, these particles themselves have spin, and may also have orbital angular momenta thanks to their motion in the proton. A spectacular discovery made by the European Muon Collaboration (EMC) experiment at CERN was that the quarks and antiquarks combined carry only a small fraction of the proton spin (Ashman *et al.*, 1988), now established to be about a quarter. This came as a surprise to many researchers in the field, because models of the proton predicted a much larger contribution. It clearly means that other contributions to the proton spin, for example, by gluons and/or orbital motion, must be very significant. The resolution of this proton spin puzzle is still an open issue and currently the focus of a number of experiments in this field.

3.4.4
Hadron–Hadron Scattering

The scattering of two hadrons, such as two protons, is another important process in QCD. A typical diagram is shown in Figure 3.8. Certain final states may be selected for which the interaction is very inelastic and both protons break apart in the interaction. Like in DIS, one may then use the parton model concepts to describe the cross section in terms of the distributions of quarks, antiquarks, and gluons in each proton, and their short-distance interaction. Thanks to the asymptotic freedom of QCD (see Section 3.3), the latter may be evaluated in QCD perturbation theory. So-called factorization theorems (Collins, Soper and Sterman, 1988) make this relation quantitative in QCD and show that the distribution functions relevant here are the same as those measured in DIS. In other words, the structure of the nucleon is universal in all these very inelastic reactions.

The prototype for inelastic proton–proton scattering is the Drell–Yan process (Drell and Yan, 1970), in which an electron–positron (or, muon–antimuon) pair is produced. This pair may be created by a quark from one proton annihilating with an antiquark from another. As such, the underlying reaction $q\bar{q} \to \gamma \to e^+e^-$ is the inverse reaction to the e^+e^- process discussed earlier. The Drell–Yan process gives evidence for the presence of antiquarks in the proton. One may also consider final states produced at very high momentum perpendicular to the beam axis, so-called high-$p_\perp$ processes.

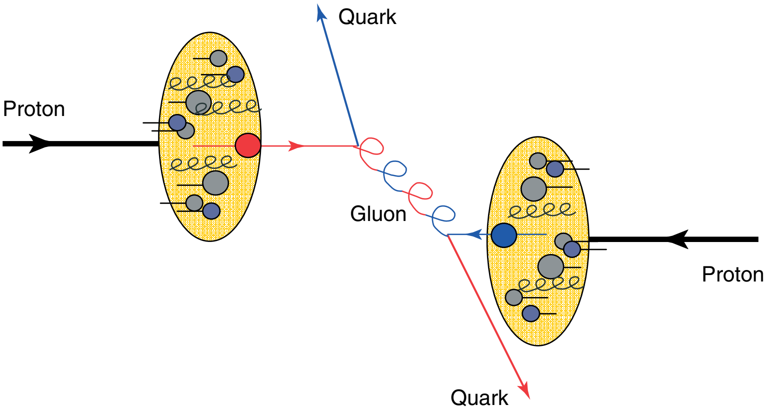

Fig. 3.8 Diagram for an inelastic reaction in proton–proton scattering. Two quarks, one from each proton, scatter by exchange of a colored gluon and produce two jets with large momentum, perpendicular to the directions of the initial protons.

An example is shown in Figure 3.8. The outgoing quarks will lead to high-$p_\perp$ jets, similar to the ones we described in the case of e^+e^- annihilation. Such reactions can serve as probes of the gluon distribution in the proton. For example, at the Relativistic Heavy-Ion Collider (RHIC) at Brookhaven, one studies high-$p_\perp$ reactions in collisions of spin-polarized protons. In this way, it becomes possible to explore the contribution by gluons to the proton spin, helping to solve the proton spin puzzle described above.

Today, our overall knowledge of the distributions of quarks, antiquarks, and gluons inside the proton, as Figure 3.7 shows, is quite good. In fact, we know them so well that we have confidence that we can use the information to make quantitative predictions for signals and background processes at the present and future high-energy colliders, Tevatron and LHC, which try to find and explore phenomena beyond our standard model of particle physics.

3.5
Phases of Strongly Interacting Matter

As has been discussed above, one of the most remarkable properties of QCD is that it incorporates asymptotic freedom as well as confinement: while quarks and gluons behave as almost free particles once they are close to each other, they are never observed individually as free particles. To reach conditions that allow quarks and gluons to behave as almost free particles, one has to ensure that their average separation is small or the average momentum exchanged during their interactions is large. These conditions are met in the high-energy experiments discussed in the previous section. They also existed in the early universe and may occur even today in the interior of dense stellar objects – the neutron stars.

In the early universe, just after the Big Bang, the temperatures were high enough and the density large, so that quarks and gluons were asymptotically free. They carried momenta of the order of a few times the temperature (times the factor k/c with k the Boltzmann constant and c the speed of light), and their average separation was of the order of the inverse of the temperature (times the factor $\hbar c/k$ with $\hbar$ the Planck constant). Unlike in collision experiments performed with protons or electrons, where only a few partons participate, quarks and gluons in the early universe formed large systems of hot and dense matter that has been called the *quark–gluon plasma*. Modern-day experiments with colliding heavy ions, for example, lead or gold nuclei, aim at recreating conditions necessary for the formation of the quark–gluon plasma.

In the interior of neutron stars, the density may also become sufficiently large for quark matter to form. Unlike that in the early universe, however, this matter has quite different properties. While in the early universe the temperature was high and the excess of matter over antimatter (the net baryon number density) was small, the situation is just the opposite inside dense stars: the baryon number density is high but the temperature is rather low.

Much like other many-body systems, macroscopic systems of quarks and gluons have thermodynamic properties. In particular, matter made of quarks and gluons can also exist in different states of aggregation. These different states of matter may be separated through phase transitions, like the liquid–gas phase transition in ordinary fluids. Again, very much like in other many-body systems, the properties of various phases of matter are controlled by

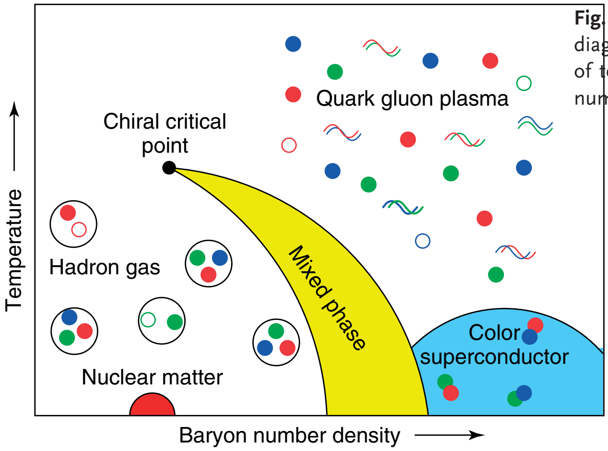

Fig. 3.9 Schematic phase diagram of QCD as function of temperature and baryon number density.

global symmetries of the system, and the transition from one phase to another is controlled by the spontaneous breaking or restoration of these symmetries. In QCD, it is the chiral symmetry that plays a central role for the understanding of the phase structure of strongly interacting matter. A schematic phase diagram of QCD as we think of it today is shown in Figure 3.9.

3.5.1
Quarks and Gluons at High Temperature

To be answered quantitatively, many of the questions related to the phase structure of QCD require nonperturbative techniques. Lattice-regularized QCD and numerical calculations that use this formulation thus play a central role in the theoretical exploration of the phase structure of QCD. These calculations are successfully used at vanishing baryon number density. They thus allow an exploration of QCD thermodynamics under conditions appropriate to describe the hot and dense matter formed in the early universe (Karsch, 2002). These calculations yield evidence for the existence of a rapid transition from a gas phase, where quarks and gluons are still confined inside hadrons and chiral symmetry is broken, to a plasma in which quarks and gluons are the most relevant degrees of freedom and become indeed asymptotically free at very high temperature. The transition between both regimes would indeed be a true phase transition, related to the restoration of chiral symmetry, if chiral symmetry were an exact symmetry of QCD, that is, if the light up and down quarks were massless. The slight explicit breaking of the chiral symmetry due to the nonvanishing up and down quark masses makes the QCD transition a rapid, but continuous, crossover from one phase to the other. Lattice calculations indicate that the transition to the quark–gluon plasma occurs at a temperature of roughly $2 \cdot 10^{12}$ K or almost 200 MeV/k (where k is the Boltzmann constant), which is a temperature about 100 000 times larger than that in the center of our sun. Such conditions have been met in the early universe at about 10^{-5} seconds after the Big Bang.

Although it is not yet confirmed by direct calculations, several theoretical considerations suggest that the continuous transition at vanishing baryon number density becomes a true phase transition with long-range correlations and a divergent correlation length at some critical value of the baryon number density (Stephanov, 2004). At densities larger than that indicated by this chiral critical point, the transition becomes a first-order phase transition, that is, it is expected that an intermediate mixed phase exists in which hadrons can coexist with already deconfined quarks and gluons.

Hot and dense elementary particle matter can be studied experimentally in heavy-ion experiments. In the collision of relativistic heavy ions, that is, the nuclei of heavy atoms such as gold or lead whose electrons have been stripped off, dense matter is generated that rapidly expands, but also seems to interact strongly enough to form an equilibrated hot medium. Experiments at the SPS collider at CERN found first indications for the formation of a quark–gluon plasma in such collisions Heinz and Jacob. The RHIC at BNL provides ground-breaking insights into the properties of strongly interacting matter under these extreme conditions. While it is difficult to establish direct evidence for the restoration of chiral symmetry, there is evidence for the formation of hot and dense quark–gluon matter with striking properties (Arsene *et al.*, 2005). The matter is very opaque: jets that are created in the collision are strongly obstructed and often absorbed while traversing the produced medium, a phenomenon known as *jet quenching*. In addition, the matter shows remarkable elliptic flow properties: an initial spatial anisotropy that is present when the two colliding nuclei hit each other not completely centrally is translated by the medium into a strong momentum anisotropy that persists all the way out to the detectors. Among other important signatures of a quark–gluon plasma that are under investigation at RHIC are the suppression of heavy-quark bound states like the J/ψ, which are expected to no longer be able to form in dense matter because of the strong modification (screening) of the heavy-quark potential, and the thermal radiation of photons, which leads to an enhancement of the photon spectrum at low transverse momenta.

One of the most important scientific discoveries of the past few years, which follows from the observations of jet quenching and elliptic flow, was that the hot matter that is formed when gold nuclei collide in RHIC appears to have the properties of an almost perfect liquid, characterized by a ratio of viscosity to entropy that is lower than that for any other known liquid. This indicates that the quark–gluon plasma formed in the collisions at RHIC is strongly coupled and not a dilute gaseous plasma as originally expected. Further exploration of this new state of matter is currently under way at RHIC and will also be an important task for heavy-ion experiments at the LHC.

3.5.2
Quarks and Gluons at High Baryon Number Density

A change from the low-temperature, low-density hadronic matter to forms of matter in which quarks are the relevant degrees of freedom is also expected to occur at large baryon number density. Again, it is asymptotic freedom that dictates the interactions of quarks at very high baryon number density and low temperatures. As quarks are fermions, they will, at low temperatures, show the characteristic properties of a Fermi gas. At high density, they cannot all occupy the ground state but must fill up higher energy levels, that is, states where individual quarks carry a large momentum. The highest momentum states near the Fermi surface will behave as almost free particles since they have to transfer large momenta when interacting with other quarks. Because of asymptotic freedom this interaction is weak. It can be analyzed systematically within perturbation theory and, in fact, is found to be attractive. An attractive interaction among fermions is the prerequisite for the occurrence of superconductivity through the formation of Cooper pairs. Just like in electrical superconductivity, the attractive color force between quarks leads to color-superconductivity in cold but dense quark matter. Such form of matter may exist in the interior of neutron stars.

As we discussed in Section 3.2, in addition to color and spin, quarks also carry a flavor degree of freedom. This introduces new possibilities for the pairing of quarks into Cooper pairs, which do not exist in ordinary superconductors. The high-density, low-temperature phase diagram of QCD thus may lead to several different phases that are characterized by different pairing mechanisms. In the idealized case of three degenerate light quark flavors, that is, in a limit where it is ignored that the strange quark is much heavier than the light quarks, it is well understood that the so-called color-flavor locked phase, in which quarks of all different colors and flavors can form Cooper pairs, is the preferred state of quark matter at high density. As the density is lowered, other symmetry breaking patterns may become more relevant. In reality, as the u, d, s quarks have different masses, flavor symmetry is not exact. This may lead to a rather complicated phase diagram of QCD at high density (Rajagopal and Wilczek, 2001).

3.5.3

QCD at High Energy

The exchange of gluons between quarks mediates the color force that binds all hadrons. A major discovery of the last decade, made in DIS at the HERA collider, is the dominant role of gluons in protons when viewed by a high-energy probe. The number density of gluons grows extremely rapidly as one probes them at progressively smaller fractions x of a proton's momentum. We have seen this strong rise of the gluon distribution in Figure 3.7. The growth reflects a QCD cascade in which higher momentum (harder) parent gluons successively split into two or more lower momentum (softer) daughter gluons. According to Einstein's theory of relativity, a high-energy proton appears strongly length contracted, or compressed, along its direction of motion, looking much like a disk to a stationary observer. At very high energies, the density of the gluons on this disk is seen to increase greatly. But, if this growth were to continue without moderation, it would eventually lead to probabilities for the scattering process in excess of unity, which is physically impossible. Instead, one expects the growth to be tamed because, once sufficiently high gluon densities have been reached, softer gluons will start to recombine again into harder ones. This competition between splitting and recombination processes should stop the growth of the density of gluons, a phenomenon known as *saturation*.

At very high gluon density, all gluons begin to act collectively and their color forces add coherently. At the same time, typical distances between the gluons become small, where QCD becomes weakly coupled thanks to its asymptotic freedom. As a result, the gluons behave like a very strong classical field. An important recent realization has been that a systematic solution becomes feasible in this domain. The theory describing this high-density gluon matter arising at high energies is called the *color glass condensate (CGC)* theory Iancu, Leonidov and McLerran. Here, the word "color" refers to the fundamental charge of QCD which the gluons carry. "Glass" is borrowed from the term for silica and other materials that are disordered and act like solids on short timescales, but as liquids on long timescales. In the "gluon disk" the gluons themselves are disordered and do not change their positions rapidly because of time dilation. Finally, "condensate" alludes to the very high density of the gluons. A striking prediction of the CGC theory is that eventually all hadrons and all nuclei should appear the same when viewed at high enough energy, that is, should be predominantly made of a universal form of gluonic matter. Possible hints of gluon saturation phenomena have been found in measurements of electron–proton collisions at HERA and of deuteron–nucleus and nucleus–nucleus collisions at RHIC. Saturated gluon densities could have a profound influence on heavy-ion collisions planned at the LHC. The CGC theory predicts that for heavy nuclei the onset of saturation phenomena would already have occurred at lower energies than for the proton. It is believed that a definitive search for universal gluonic matter would become feasible at a future high-energy electron–ion collider.

References

(a) Abe, F. *et al.* (CDF Collaboration) (**1995**) Observation of top quark production in $\overline{p}p$ collisions. *Phys. Rev. Lett.*, **74**, 2626; (b) Abachi, S. *et al.*(D0 Collaboration) (**1995**) Observation of the top quark. *Phys. Rev. Lett.*, **74**, 2632.

Aoki, S., *et al.* (CP-PACS Collaboration) (**2003**) Light hadron spectrum and quark masses from quenched lattice QCD. *Phys. Rev. D*, **67**, 034503.

(a) Arsene, I., *et al.* (BRAHMS Collaboration) (**2005**) Quark-gluon plasma and color glass condensate at RHIC? The perspective from the BRAHMS experiment. *Nucl. Phys. A*, **757**, 1; (b) Back, B.B., *et al.* (PHOBOS Collaboration) (**2005**) The PHOBOS perspective on discoveries at RHIC. *Nucl. Phys. A*, **757**, 28; (c) Adams, J. *et al.* (STAR Collaboration) (**2005**) Experimental and theoretical challenges in the search for the quark–gluon plasma: the STAR Collaboration's critical assessment of the evidence from RHIC collisions. *Nucl. Phys. A*, **757**, 102; (d) Adcox, K., *et al.* (PHENIX Collaboration) (**2005**) Formation of dense partonic matter in relativistic nucleus-nucleus collisions at RHIC: experimental evaluation by the PHENIX collaboration. *Nucl. Phys. A*, **757**, 184.

(a) Ashman, J., *et al.* (European Muon Collaboration) (**1988**) A measurement of the spin asymmetry and determination of the structure function g(1) in deep inelastic muon proton scattering. *Phys. Lett. B*, **206**, 364; (b) Ashman, J., *et al.* (**1989**) An investigation of the spin structure of the proton in deep inelastic scattering of polarized muons on polarized protons. *Nucl. Phys. B*, **328**, 1.

(a) Aubert, J.J., *et al.* (E598 Collaboration) (**1974**) Experimental observation of a heavy particle. *J. Phys. Rev. Lett.*, **33**, 1404; (b) Ting, S.C.C. (**1976**) *The Discovery of the J Particle*, Nobel lecture.

(a) Augustin, J.E. *et al.* (SLAC-SP-017 Collaboration) (**1974**) Discovery of a narrow resonance in e^+e^- annihilation. *Phys. Rev. Lett.*, **33**, 1406; (b) Richter, B. (**1976**) *From the ψ to Charm – the Experiments of 1975 and 1976*, Nobel lecture 1976.

(a) Barber, D.P., *et al.* (Mark J Collaboration) (**1979**) Discovery of three jet events and a test of Quantum Chromodynamics at Petra energies. *Phys. Rev. Lett.*, **43**, 830; (b) Brandelik, R., *et al.* (TASSO Collaboration) (**1979**) Evidence for planar events in e^+e^- annihilation at high-energies. *Phys. Lett. B*, **86**, 243; (c) Bartel, W., *et al.*

(JADE Collaboration) (**1980**) Observation of planar three jet events in e^+e^- annihilation and evidence for gluon bremsstrahlung. *Phys. Lett. B*, **91**, 142; (d) Berger, C. *et al.* (PLUTO Collaboration) (**1979**) Evidence for gluon bremsstrahlung in e^+e^- annihilations at high-energies. *Phys. Lett. B*, **86**, 418.

Barnes, V.E., *et al.* (**1964**) Observation of a hyperon with strangeness −3. *Phys. Rev. Lett.*, **12**, 204.

Bethke, S. (**2007**) Experimental tests of asymptotic freedom. *Prog. Part. Nucl. Phys.*, **58**, 351. [arXiv:hep-ex/0606035].

(a) Bloom, E.D. *et al.* (**1969**) High-energy inelastic e p scattering at 6-degrees and 10-degrees. *Phys. Rev. Lett.*, **23**, 930; (b) Observed behavior of highly inelastic electron-proton scattering. *Phys. Rev. Lett.*, *ibid.* 935; (c) Friedman, J.I. (**1990**) *Deep Inelastic Scattering: Comparisons with the Quark Model*; (d) Kendall, H.W. (**1990**) *Deep Inelastic Scattering: Experiments on the Proton and the Observation of Scaling*; (e) Taylor, R.E. (**1990**) *Deep Inelastic Scattering: The Early Years*, Nobel lectures.

(a) Bjorken, J.D. (**1969**) Asymptotic sum rules at infinite momentum. *Phys. Rev.*, **179**, 1547; (b) Feynman, R.P. (**1969**) Very high-energy collisions of hadrons. *Phys. Rev. Lett.*, **23**, 1415; (c) Bjorken, J.D. and Paschos, E.A. (**1969**) Inelastic electron proton and gamma proton scattering, and the structure of the nucleon. *Phys. Rev.*, **185**, 1975.

Cheng, M., *et al.* (RBC-Bielefeld Collaboration) (**2008**) The QCD equation of state with almost physical quark masses. *Phys. Rev. D*, **77**, 014511.

(a) Christ, N.H., Hasslacher, B., and Mueller, A.H. (**1972**) Light cone behavior of perturbation theory. *Phys. Rev. D*, **6**, 3543; (b) Georgi, H. and Politzer, H.D. (**1974**) Electroproduction scaling in an asymptotically free theory of strong interactions. *Phys. Rev. D*, **9**, 416.

Collins, J.C., Soper, D.E., and Sterman, G. (**1988**) Factorization of hard processes in QCD. *Adv. Ser. Direct. High Energy Phys.*, **5**, 1. *and references therein*.

Cooper-Sarkar, A. and Gwenlan, C. *Proceedings of "HERA and the LHC: A Workshop on the Implications of HERA and LHC Physics"*. Comparison and Combination of ZEUS and H1 PDF Analyses, Hamburg, Germany, 21–24 Mar 2005. arXiv:hep-ph/0508304.

Creutz, M. (**1980**) Monte Carlo study of quantized SU(2) gauge theory. *Phys. Rev. D*, **21**, 2308.

Drell, S.D. and Yan, T.M. (**1970**) Massive lepton pair production in hadron-hadron collisions at high-energies. *Phys. Rev. Lett.*, **25**, 316.

Ellis, R.K., Stirling, W.J., and Webber, B.R. (**1996**) QCD and collider physics. *Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol.*, **8**, 1.

Fritzsch, H., Gell-Mann, M., and Leutwyler, H. (**1973**) Advantages of the color octet gluon picture. *Phys. Lett. B*, **47**, 365.

(a) Gell-Mann, M. (**1956**) The interpretation of the new particles as displaced charged multiplets. *Nuovo Cimento Suppl.*, **4**, 848; (b) Nishijima, K. (**1955**) Charge independence theory of V particles. *Prog. Theor. Physics*, **13**, 285.

(a) Gell-Mann, M. (**1964**) A Schematic model of baryons and mesons. *Phys. Lett.*, **8**, 214; (b) Gell-Mann, M. (**1969**) *Symmetry and Currents in Particle Physics*, Nobel Lecture.

Gell-Mann, M., Oakes, R.J., and Renner, B. (**1968**) Behavior of current divergences under SU(3) × SU(3). *Phys. Rev.*, **175**, 2195.

(a) Glashow, S.L. (**1961**) Partial symmetries of weak interactions. *Nucl. Phys.*, **22**, 579; (b) Weinberg, S. (**1967**) A model of Leptons. *Phys. Rev. Lett.*, **19**, 1264; (c) Salam, A. (**1968**) Weak and electromagnetic interactions, in *Elementary Particle Physics (Nobel Symposium No. 8)*, (ed. N. Svartholm), Almqvist and Wiksell, Stockholm, p. 367; (d) Glashow, S.L. (**1979**) *Towards a Unified Theory – Threads in a Tapestry*; (e) Salam, A. (**1979**) *Gauge Unification of Fundamental Forces*; (f) Weinberg, S. *Conceptual Foundations of the Unified Theory of Weak and Electromagnetic Interactions*, Nobel lectures 1979.

(a) Goldstone, J. (**1961**) Field theories with superconductor solutions. *Nuovo Cim.*, **19**, 154; (b) Nambu, Y. and Jona-Lasinio, G. (**1961**) Dynamical model of elementary particles based on an analogy with superconductivity I. *Phys. Rev.*, **122**, 345.

Greenberg, O.W. (**1964**) Spin and unitary spin independence in a paraquark model of baryons and mesons. *Phys. Rev. Lett.*, **13**, 598.

(a) Gross, D.J. and Wilczek, F. (**1973**) Ultraviolet behavior of non-abelian gauge theories. *Phys. Rev. Lett.*, **30**, 1343; (b) Politzer, H.D. Reliable perturbative results for strong interactions? *Phys. Rev. Lett.*, **30**, 1346; (c) Gross, D.J. (**2004**) *The Discovery of Asymptotic Freedom and the Emergence of*

QCD; (d) Politzer, H.D. (**2004**) *The Dilemma of Attribution*; (e) Wilczek, F. (**2004**) *Asymptotic Freedom: from Paradox to Paradigm*, Nobel lectures.

Han, M.Y. and Nambu, Y. (**1965**) Three-triplet model with double SU(3) symmetry. *Phys. Rev.*, **139**, B1006.

(a) Lattes, C.M.G., Muirhead, H., Occhialini, G.P.S., and Powell, C.F. (**1947**) Processes involving charged mesons. *Nature*, **159**, 694; (b) Powell, C.F. (**1950**) *The Cosmic Radiation*, Nobel Lecture.

Heinz, U.W. and Jacob, M. *Evidence for a new state of matter: An assessment of the results from the CERN lead beam programme*, arXiv:nucl-th/0002042.

Herb, S.W., *et al.* (**1997**) Observation of a dimuon resonance at 9.5-GeV in 400-GeV proton – nucleus collisions. *Phys. Rev. Lett.*, **39**, 252.

Iancu, E., Leonidov, A., and McLerran, L.(**2002**) The colour glass condensate: an introduction to Cargese summer school on "QCD perspectives on hot and dense matter", 6–18 Aug 2001, Cargese, France, *Proceedings: QCD Perspectives on Hot and Dense Matter: Proceedings*. (eds. J.-P. Blaizot and E. Iancu, Dordrecht), NATO Science Series, II, Mathematics, Physics and Chemistry, Vol. **87**, Kluwer, The Netherlands, pp. 522.

Karsch, F. (**2002**) Lattice QCD at high temperature and density. *Lect. Notes Phys.*, **583**, 209.

(a) Ne'eman, Y. (**1961**) Derivation of strong interactions from a gauge invariance. *Nucl. Phys.*, **26**, 222; (b) Gell-Mann, M. (**1962**) The eightfold way: a theory of strong interaction symmetry, Caltech preprint CTSL-20 Symmetries of baryons and mesons. *Phys. Rev.*, **125**, 1067; (c) Gell-Mann, M. (**1969**) *Symmetry and Currents in Particle Physics*, Nobel Lecture 1969.

Rajagopal, K. and Wilczek, F. (**2001**) The condensed matter physics of QCD, in *At the Frontier of Particle Physics*, (ed. M. Shifman), Vol. **3**, World Science, Singapore, p. 2061.

Rochester, G.D. and Butler, C.C. (**1947**) Evidence for the existence of new unstable elementary particles. *Nature*, **160**, 855.

Stephanov, M. (**2004**) The phase diagram of QCD and the critical point. *Acta Phys. Polon. B*, **35**, 2939.

Sterman, G. and Weinberg, S. (**1977**) Jets from quantum chromodynamics. *Phys. Rev. Lett.*, **39**, 1436.

(a) Weinberg, S. (**1967**) Precise relations between the spectra of vector and axial vector mesons. *Phys. Rev. Lett.*, **18**, 507; (b) Weinberg, S. (**1968**) Nonlinear realizations of chiral symmetry. *Phys. Rev.*, **166**, 1568; (c) Gasser, J. and Leutwyler, H. (**1984**) Chiral Perturbation Theory to one loop. *Ann. Phys.*, **158**, 142.

Wilson, K.G. (**1974**) Confinement of quarks. *Phys. Rev. D*, **10**, 2445.

Yao, W.M., *et al.* (Particle Data Group) (**2006**) Review of particle physics. *J. Phys. G*, **33**, 1.

(a) Yukawa, H. (**1935**) On the interaction of elementary particles. *Proc. Phys. Math. Soc. Jpn.*, **17**, 4. 8; (b) Yukawa, H. (**1949**) *Meson Theory in its Developments*, Nobel Lecture.

Zweig, G. (**1964**) *An SU(3) Model for Strong Interaction Symmetry and its Breaking*, preprint CERN-TH-401.

STRUCTURAL AND PHYSICAL GEOLOGY

Robert D. Hatcher, Jr., *Department of Geological Sciences, University of Tennessee–Knoxville, Knoxville, Tennessee, U.S.A.*

	Introduction	97
1.	**Tectonics**	97
2.	**Minerals**	101
3.	**Igneous Rocks**	101
4.	**Surficial Processes**	104
4.1	Subaerial Processes	104
4.2	Gravity-Related Slope Processes	105
4.3	Stream Processes	106
5.	**Sedimentation**	107
6.	**Metamorphism**	107
7.	**Deformation Processes**	109
7.1	Physical Properties of Rocks	110
7.2	Folds	110
7.3	Fractures and Faults	112
7.4	Foliations	114
	Glossary	115
	Works Cited	116
	Further Reading	116

INTRODUCTION

Physical geology is the study of the physical processes, both internal and surficial, that function upon and within the Earth and other planetary bodies. They include mantle and lithospheric convection processes that drive plate motion (see Plate Tectonics) that in turn produces volcanism, plutonism (a process whereby molten rock is generated and emplaced into the crust of the Earth and crystallized beneath the surface), and deformation within plates, but particularly concentrated at plate boundaries. Structural geology is a branch of geology that deals with the origin, geometry, and kinematics of emplacement of tectonic structures—folds, faults and other fractures, foliations, and microstructures—produced by deformation processes within the Earth. In addition, impact structures and structures formed near the surface by gravity-related processes are sometimes included in the collection of geologic structures. Historical geology places these processes in the continuum of geologic time. Internal processes are driven largely by thermal energy derived from radioactive decay, with smaller components from residual primordial heat, friction, and gravity. Surficial processes are driven primarily by solar and gravitational energy.

Many of the individual processes studied by geologists are components or steps in natural cycles or subcycles that absorb and utilize thermal energy released from the Earth's interior to bring about physical or chemical transformations. The tectonic cycle embodies all of the processes that produce the phenomena associated with both internal and surficial processes and incorporates the rock or geochemical cycle that also contains the hydrologic and biological cycles (Figs. 1 and 2). The article outline is constructed primarily around the rock cycle, with the addition of increments of the tectonic cycle and rock deformation.

1. TECTONICS

Tectonics is the study of large-scale structural features on the Earth ranging from the scale of a few square kilometers to continent or whole-Earth-scale features (also see Plate Tectonics). The dividing line between structural geology and tectonics is not clear because many attributes of structural geology contribute to the interpretation of tectonic features and conversely.

The lithosphere—the 100-km-thick outer shell of the Earth, including all the crust and the uppermost rigid part of the mantle—can be divided into a number of large and several smaller plates (Fig. 3). Plate tectonics is the major paradigm by which we explain many of the processes of the evolution of

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

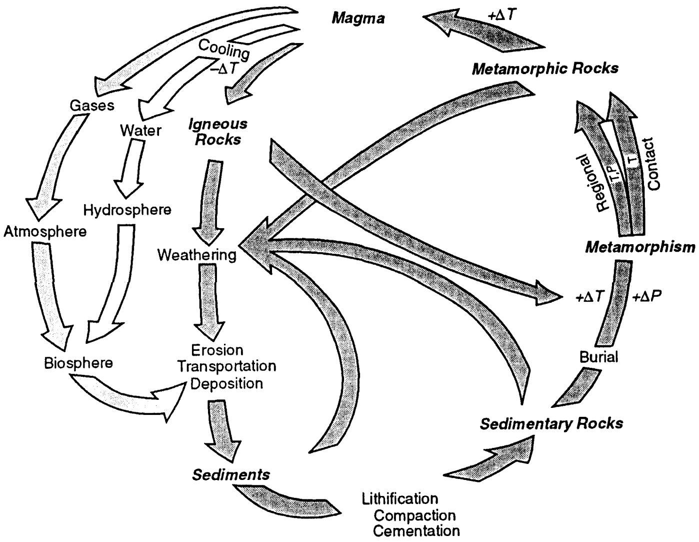
Magma
+ΔT
Cooling
−ΔT
Metamorphic Rocks
Gases
Water
Igneous Rocks
Hydrosphere
Atmosphere
Weathering
Regional
T,P
T
Contact
Metamorphism
+ΔT
+ΔP
Burial
Biosphere
Erosion
Transportation
Deposition
Sedimentary Rocks
Sediments
Lithification
Compaction
Cementation

FIG. 1. The rock, or geochemical, cycle—a thermally and mechanically driven equilibrium cycle involving many intermediate stages and shorter cycles. (Modified from Hatcher, 1995, © 1995 Prentice-Hall Inc.)

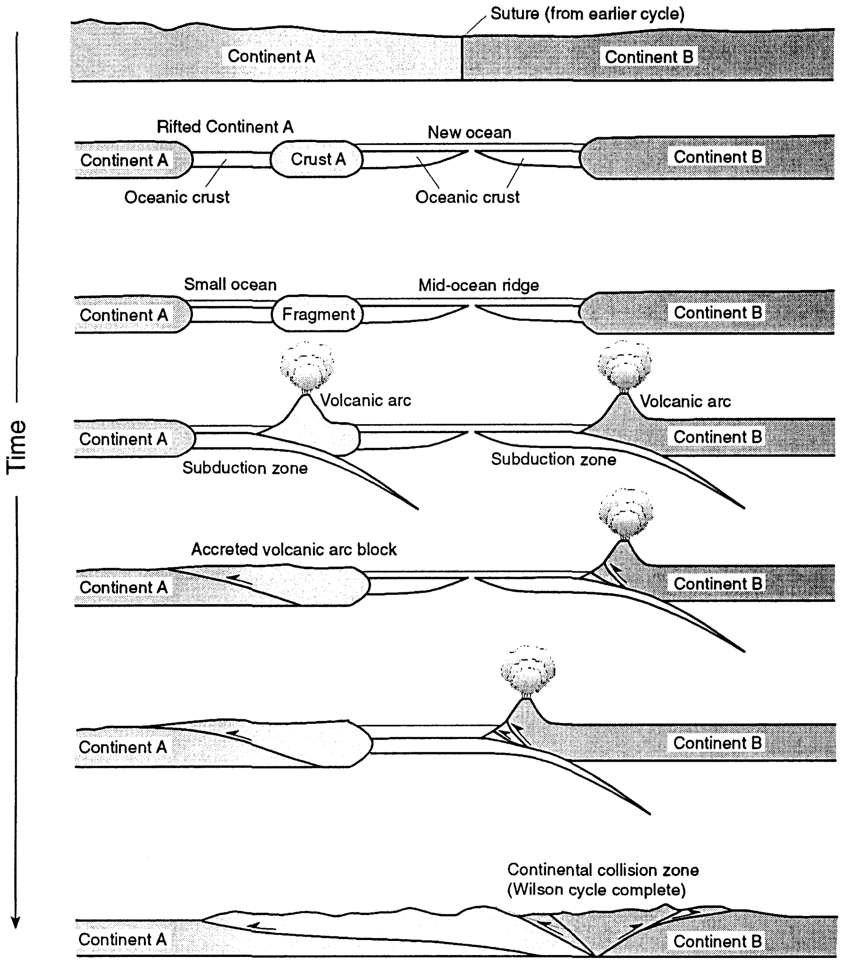
Time
Suture (from earlier cycle)
Continent A
Continent B
Rifted Continent A
New ocean
Crust A
Oceanic crust
Oceanic crust
Small ocean
Mid-ocean ridge
Fragment
Volcanic arc
Volcanic arc
Subduction zone
Subduction zone
Accreted volcanic arc block
Continental collision zone
(Wilson cycle complete)

FIG. 2. The Wilson cycle of the opening and closing of an ocean basin. The cycle may be complicated by the formation and movement of suspect terranes, partial closing of small oceans, and the lack of continent–continent collision to terminate the cycle. (From Hatcher, 1995, © 1995 Prentice Hall Inc.)

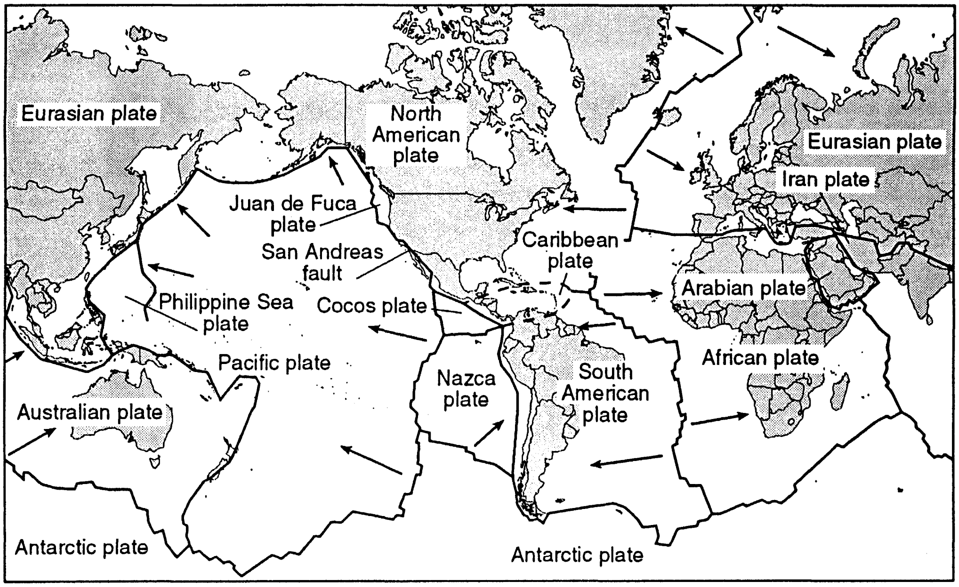

FIG. 3. Distribution of tectonic plates on the Earth's surface. Arrows indicate principal directions of plate motion. Solid lines indicate plate boundaries. (From Hatcher, 1995, © 1995 Prentice Hall Inc.)

continents, ocean basins, mountain chains, and other features that dominate the surface of the Earth today. The continents are composed of light (quartz- and Na-feldspar-rich, commonly $>65\%$ SiO_2) materials with an average granitic composition and, as a consequence, tend to float and not be consumed in subduction zones, regions where crustal or lithospheric material is conveyed downward into the mantle. Oceanic crust that has an average basaltic composition (Mg- and Fe-rich minerals plus Ca-rich feldspar; commonly $<45\%$ SiO_2), in contrast, because of its higher density, is consumed at convergent plate boundaries and is generated at divergent plate boundaries (Fig. 2). The present-day cycle for consumption of ocean crust appears to be about 200 million years, because no ocean crust is identifiable in the present-day oceans that is older than mid-Mesozoic ($\sim$180 Ma), but the time required for the reappearance of individual atoms or ions on the surface may be on the order of billions of years (Anderson, 1995).

Continental crust, because of its lower density, floats and is not consumed. Therefore, it is a scorekeeper that records events through time; it has recorded the plate tectonic history of the Earth since the first continental crust was formed about four billion years ago, and it is still evolving today. The older parts of the continents consist of isostatically adjusted crust that formed the continental nuclei to which younger and younger crust was added through time. This younger continental crust is still being formed today at the western margins of both North and South America and in other parts of the world where plate convergence and collision are taking place and crust-forming processes are active.

Ocean crust forms at the oceanic ridges—spreading centers—where two plates are being pulled apart and new crust is generated as upwelled molten rock derived from melting at lower pressures crystallizes to form basalt. The oceanic ridges commonly have a high topography because the crust beneath them is warm and thus of slightly lower density. The deeper abyssal plains of the oceans form as the crust cools, becomes denser, thickens, becomes more rigid, and sinks. Ocean crust is destroyed by being consumed in trench-subduction zones.

Mountain chains are the product of continental convergence and the completion of Wilson cycles (opening and closing of ocean basins), provided the convergence process involves continent–continent collision (Figs. 2 and 4). Modern mountain chains that involve

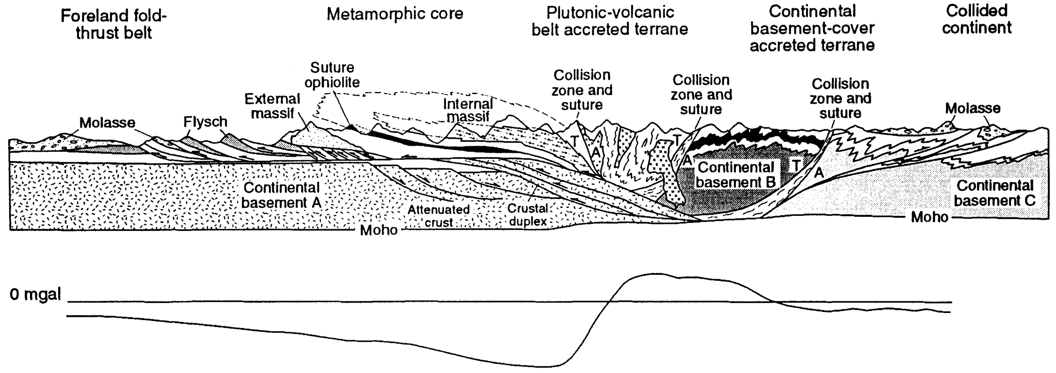
Foreland fold-thrust belt
Metamorphic core
Plutonic-volcanic belt accreted terrane
Continental basement-cover accreted terrane
Collided continent
Suture ophiolite
External massif
Internal massif
Collision zone and suture
Collision zone and suture
Collision zone and suture
Molasse
Flysch
Molasse
T
A
T
A
T
A
Continental basement A
Attenuated crust
Crustal duplex
Continental basement B
Continental basement C
Moho
Moho
0 mgal

FIG. 4. Ideal mountain chain. Masses of continental or oceanic crust or island arcs may be accreted by either head-on or oblique collision, producing a variety of deformational styles, metamorphism, and plutonism. The external massif is a block of basement caught up and transported in the thrust stack. Continental basements A, B, and C represent parts of different continents now amalgamated by collision. T—movement toward observer; A—movement away from observer. Curve is a hypothetical gravity profile of the kind observable in most mountain chains. (From Hatcher and Williams, 1986.)

the completed Wilson cycles include the Alps and the Himalayas. Other mountain chains that are still active today, like the Andes and North American Cordillera, either involve Wilson cycles that are incomplete or the concept of the Wilson cycle is inapplicable to consideration of these mountain chains. This may be the case with all of the island arc and mountain chains around the Pacific margin because there is no indication that continent–continent collision will take place any time in the near- or far-distant geologic future.

Mountain chains have a number of features in common that form during the mountain building cycle (Fig. 4). These include a foreland fold-thrust belt that occurs in most mountain chains produced by deformation of the sediments deposited on the continental margin during the rift-to-drift transition following the opening of a new ocean. Rifted-margin clastic (mechanically deposited sands and muds) sediments are succeeded by carbonate rocks that formed when the margin consistently bounded open ocean. This is followed by the initiation of closing, recognized in the sedimentary record by a reversal of the source of sediments from the old continent to the newly rising mountain chain that is forming offshore or near shore as plate convergence becomes dominant. Island arcs frequently form during this process, and they too contribute sediments to the nearby forearc and backarc regions (the regions toward and away from the open ocean, respectively) and frequently are accreted to continental margins as subduction consumes the intervening ocean crust.

Foreland fold-thrust belts are commonly deformed during the latter stages of mountain building, a statement not true everywhere even in the same mountain chain (that in the northern Appalachians formed during the Ordovician, whereas that in the southern and central Appalachians formed at the end of the cycle). Farther into the chain is commonly a metamorphic core composed of the metamorphosed equivalents of continental rifted-margin sediments, along with oceanic components, either accreted arc material, ocean-floor material, ocean crust, ophiolites (mantle, ocean crust, and overlying ocean-floor sediments), or even rifted pieces of continent that were accreted to the original continental mass, as well as exotic elements derived from volcanic arcs and pieces of distant continents. Pluton generation (a pluton is the mass formed by crystallization of magma) accompanies subduction, and, as the molten material moves or is forced upward into the crust, it intrudes the mountain chain as part of the formation of island or continental arcs. Some late plutons are sometimes considered anorogenic. Plutons that were derived from continental crust commonly have a greater than 0.706 $^{87}Sr/^{86}Sr$ initial ratio, whereas plutons that formed in island arcs, in oceanic material, in the lower crust, or in mantle have a less than 0.706 $^{87}Sr/^{86}Sr$ ratio.

The plate convergence process that involves the accretion of both suspect and exotic terranes—blocks of crustal dimensions whose source either is questionable or is demonstrably alien to the block to which it is presently attached—to the continental mass may involve lateral, oblique, or head-on accretion or a combination of all of these. There may also be rotational components involved in the accretion process.

2. MINERALS

Minerals are naturally occurring, crystalline, inorganic compounds with a composition that is constant or varies within definable limits. Exceptions include water and mercury, which are liquids or gases at normal surface conditions, and several minerals, like opal, that are amorphous or have very open structures like liquids. The most common elements in the crust are mostly light elements, including O, Si, Al, Fe, Ca, Na, K, Mg, and Ti, in order of decreasing abundance. The cations are mostly strongly electropositive, while oxygen is strongly electronegative, and so the bonds in many minerals have a large ionic component. Bonding in minerals varies from pure end-member types of metallic (native metals), ionic (halides), and covalent (quartz, diamond) to complex mixed-bond types, such as the micas that contain covalent, ionic, and weak electrostatic (mostly van der Waals) bonds.

Minerals are commonly grouped according to chemical composition and crystal structure. They are also broadly grouped into silicate and nonsilicate minerals. Silicates are the most common rock-forming minerals in the crust, with the carbonate minerals, principally calcite ($CaCO_3$) and dolomite [$CaMg(CO_3)_2$], forming a small but significant proportion of the minerals in sedimentary rocks. Clay minerals consist of poorly crystallized, weakly bonded, sheet silicate minerals.

Among the silicate minerals (Fig. 5), the feldspars (potash feldspars, $KAlSi_3O_8$, and the plagioclase feldspars where compositions range from $NaAlSi_3O_8$ to $CaAl_2Si_2O_8$) are the most common in the crust, with quartz (SiO_2) second in abundance. Both feldspars and quartz have strongly bonded polymeric three-dimensional network structures in which SiO_4 tetrahedra are linked to each other. If Al is substituted for one of four Si atoms, a negative charge imbalance results that is neutralized by a positive cation of Na, K, or Ca to form the feldspars, feldspathoids, and zeolites. In addition, other important rock-forming minerals include the micas—mostly biotite and muscovite, which have sheet structures; the chain silicates (also called inosilicates); amphiboles [most commonly hornblende, less commonly tremolite, glaucophane, and anthophyllite, based on the composition $X_2Y_5(Si_8O_{22})(OH)_4$, where, if X is a divalent cation (Ca,Mg,Fe^{II}), Y is also, but if X is univalent (mostly Na), Y will be a trivalent cation (Fe^{III}, Al)]; pyroxenes (most commonly augite, hypersthene, and diopside); double tetrahedra including epidote, $Ca_2(Al,Fe)Al_2O(SiO_4)(Si_2O_7)(OH)$; isolated tetrahedra, such as garnet [$X_3Y_2(SiO_4)_3$, where X is a divalent cation like Fe^{II}, Mg^{II}, Mn^{II}, or Ca^{II}, and Y is a trivalent cation like Al^{III}, Cr^{III}, and Fe^{III}]; and olivine [$(Fe,Mg)_2SiO_4$].

Common minerals in oceanic crust are silicates rich in Mg, Fe, and Ca, in contrast with those in continental crust, which are rich in Na, K, and Al. Moreover, the most common minerals in oceanic crust have structures that are less complex with respect to SiO_4 linkages and polymerization than minerals in continental crust, which has an abundance of strongly polymerized three-dimensional structure silicate minerals (mostly quartz and feldspars). The minerals in oceanic crust (pyroxene, plagioclase, and olivine) have simpler silicate structures, and higher densities (average density 3.0 g/cm^3) and seismic velocities (6.0–7.2 km/s), than those in continental crust (average density 2.6 g/cm^3, seismic velocities <6.0 km/s), trends that become even more pronounced in the mantle where even simpler silicate structures (olivine, pyroxene, very little feldspar) and higher densities (average density 4.0 g/cm^3) and seismic velocities (≥ 8.0 km/s) prevail.

3. IGNEOUS ROCKS

Magma—molten rock—is generated by partial melting of crust in subduction zones and by pressure decreases accompanying

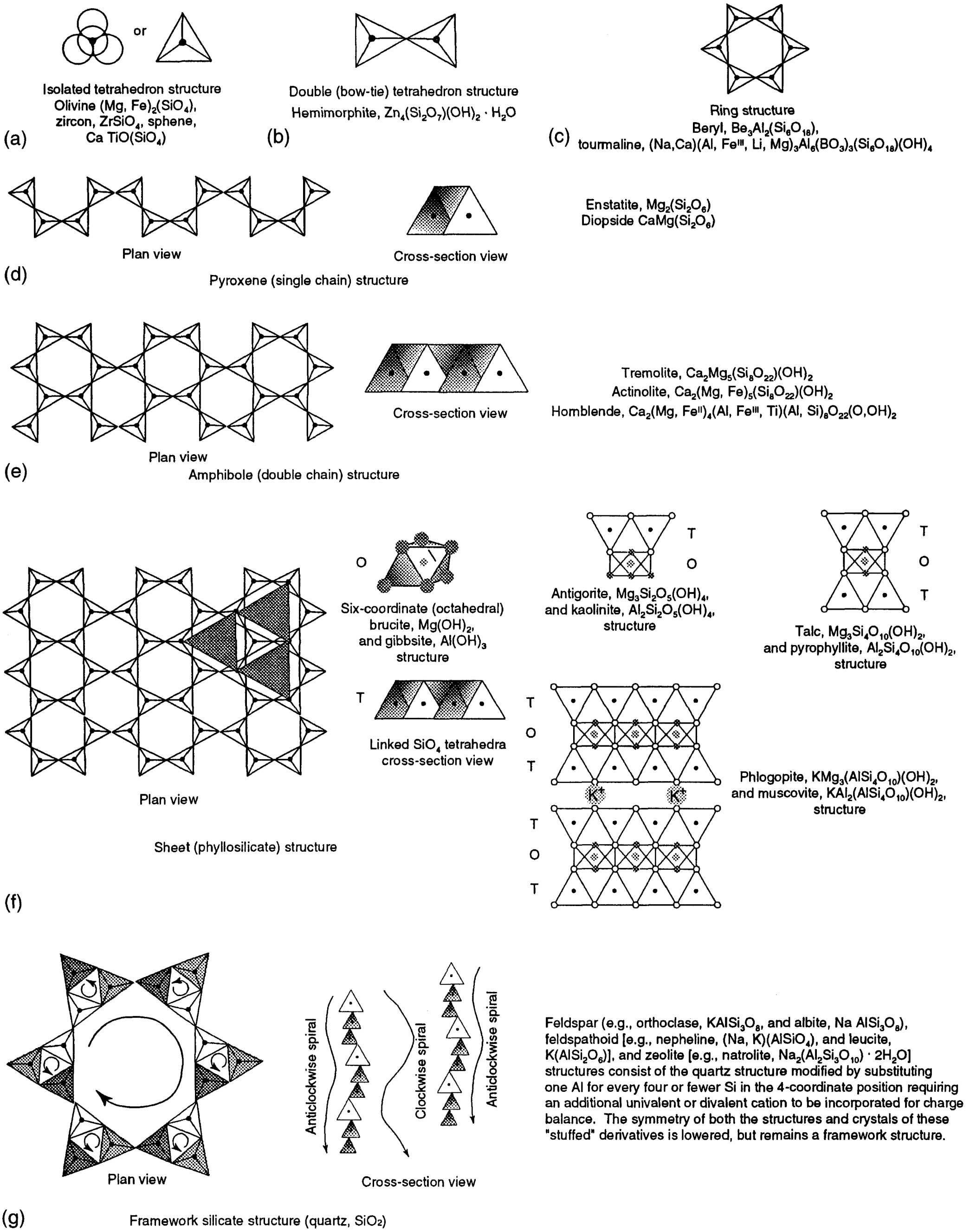
or
Isolated tetrahedron structure
Olivine (Mg, Fe)$_2$(SiO$_4$), zircon, ZrSiO$_4$, sphene, Ca TiO(SiO$_4$)
(a)
Double (bow-tie) tetrahedron structure
Hemimorphite, Zn$_4$(Si$_2$O$_7$)(OH)$_2$ · H$_2$O
(b)
Ring structure
Beryl, Be$_3$Al$_2$(Si$_6$O$_{18}$),
tourmaline, (Na,Ca)(Al, Fe^{III}, Li, Mg)$_3$Al$_6$(BO$_3$)$_3$(Si$_6$O$_{18}$)(OH)$_4$
(c)
Plan view
Cross-section view
Enstatite, Mg$_2$(Si$_2$O$_6$)
Diopside CaMg(Si$_2$O$_6$)
(d)
Pyroxene (single chain) structure
Cross-section view
Tremolite, Ca$_2$Mg$_5$(Si$_8$O$_{22}$)(OH)$_2$
Actinolite, Ca$_2$(Mg, Fe)$_5$(Si$_8$O$_{22}$)(OH)$_2$
Hornblende, Ca$_2$(Mg, Fe^{II})$_4$(Al, Fe^{III}, Ti)(Al, Si)$_8$O$_{22}$(O,OH)$_2$
Plan view
(e)
Amphibole (double chain) structure
O
Six-coordinate (octahedral) brucite, Mg(OH)$_2$, and gibbsite, Al(OH)$_3$ structure
T
Linked SiO$_4$ tetrahedra cross-section view
Plan view
Sheet (phyllosilicate) structure
(f)
T
O
Antigorite, Mg$_3$Si$_2$O$_5$(OH)$_4$, and kaolinite, Al$_2$Si$_2$O$_5$(OH)$_4$, structure
T
O
T
Talc, Mg$_3$Si$_4$O$_{10}$(OH)$_2$, and pyrophyllite, Al$_2$Si$_4$O$_{10}$(OH)$_2$, structure
T
O
T
K$^+$
K$^+$
T
O
T
Phlogopite, KMg$_3$(AlSi$_4$O$_{10}$)(OH)$_2$, and muscovite, KAl$_2$(AlSi$_4$O$_{10}$)(OH)$_2$, structure
Anticlockwise spiral
Clockwise spiral
Anticlockwise spiral
Feldspar (e.g., orthoclase, KAlSi$_3$O$_8$, and albite, Na AlSi$_3$O$_8$), feldspathoid [e.g., nepheline, (Na, K)(AlSiO$_4$), and leucite, K(AlSi$_2$O$_6$)], and zeolite [e.g., natrolite, Na$_2$(Al$_2$Si$_3$O$_{10}$) · 2H$_2$O] structures consist of the quartz structure modified by substituting one Al for every four or fewer Si in the 4-coordinate position requiring an additional univalent or divalent cation to be incorporated for charge balance. The symmetry of both the structures and crystals of these "stuffed" derivatives is lowered, but remains a framework structure.
Plan view
Cross-section view
(g)
Framework silicate structure (quartz, SiO$_2$)

FIG. 5. Common silicate structures and representative examples. Silicate structures, with the exception of the isolated tetrahedral structures (a), consist of tetrahedra linked in various ways from double tetrahedra (b), to four- and six-membered rings (c), to single and double chains (d and e), to sheets (f), and three-dimensional framework structures (g). Solid dots represent Si atoms, open circles are oxygen atoms, dark-shaded dots represent OH ions, light-shaded small circles are Al atoms, T identifies a tetrahedral (four-coordinate) layer in the phyllosilicates, and O identifies an octahedral (six-coordinate) layer in the phyllosilicates. The shaded triangles in the plan view sheet structure represent the sites to which octahedral layers are attached to the tetrahedral layers throughout the phyllosilicate structures. Large, light-shaded dots with K$^+$ in them represent large cation sites in the phlogopite and muscovite structures that provide increased bond strength to permit large sheets to form; divalent cations (e.g., Ca^{2+}) in these sites further increase bond strength between layers so that more equant crystals form (these are the "brittle micas," such as margarite and ottrelite).

crustal extension (at mid-ocean ridges or by extension of continental or oceanic crust). Magma moves upward by pushing aside the rocks above, driven by the lower density of the magma; by melting, removal, and assimilation of solid blocks of wallrock (rock adjacent to an igneous body or a mineral deposit)—a thermally inefficient process; and by magma fracturing (analogous to hydrofracturing) whereby magma is forced into zones of lower pressure by forming tension fractures that provide a conduit system into which magma flows. Physical properties of the magma, such as viscosity, volume, temperature, water content, and composition, control the emplacement mechanism.

While temperature has a strong influence on viscosity of magma, silica content of magma also controls the viscosity. The higher the percentage of SiO_2 in the magma, the higher the viscosity and the lower the melting temperature. Thus granitic magmas have much higher viscosities than basaltic magmas. The higher water and other gas content of granitic magmas also lowers the viscosity. Granitic magma that reaches the surface as rhyolitic volcanic rocks, however, commonly flows very slowly because the water and other gases that were dissolved in the magma under pressure at depth then are exsolved as pressure decreases near the surface. As a result, glasses of granitic composition (obsidian, pitchstone, and pumice) are much more common than basalt glass. Basalt, low-silica volcanic rock, has a much higher melting temperature (~1000 °C) than granitic magma at minimum-melting conditions (~650 °C, water saturated). Near-surface water provides a source for much of the water in magmas and is also responsible for much of the explosive volcanic activity observed. High-silica magmas, however, are more likely to produce explosive activity than low-silica magmas if there is no surface-water contamination.

Contamination of magmas by melting and assimilation of wallrocks provides a way for alteration of the composition of magmas. Assimilation and melting of appreciable quantities of wallrocks can significantly modify magma composition, but this also is thermally a very inefficient process, because of the heat energy necessary for assimilation. Turcotte and Emerman (1985) concluded that, because of the heat-exchange requirements of melting, assimilation, and related processes, magma-induced fracturing is a much more efficient process for intrusion into a cold crust.

Initial ratios of $^{87}Sr/^{86}Sr$ may be used to determine the source of the original magma. The upper continental crust commonly has an average $^{87}Sr/^{86}Sr$ of >0.706; the lower continental crust and the mantle have $^{87}Sr/^{86}Sr$ <0.706, so it is possible to distinguish broadly magmas derived from the upper continental crust from those derived from the mantle and lower crust. Ocean crust also has low initial $^{87}Sr/^{86}Sr$ ratios, because of its derivation from the mantle.

Crystallization of magma begins as soon as enough heat is lost for the temperature to fall below the melting temperatures of the highest–melting-point constituents and proceeds until the last remaining liquid is solidified. The original composition of the magma largely determines the range of constituent minerals and rock types that will crystallize from the magma. Additional factors include the amount of water present, oxidation state of the magma, and, to a lesser extent, other dissolved gases. Dissolved gases accumulate in the magma as the high–melting-point constituents crystallize, lower the viscosity of the residual magma, permit greater element mobility, and allow larger crystals to form. During this process, if the magma is basaltic, crystals of olivine and calcium-rich plagioclase will be the first to crystallize. These crystals will sink toward the bottom of the magma chamber, and, as the temperature decreases, some will react with the liquid to form clinopyroxene and plagioclase that contain less Ca and more Na, respectively (Fig. 6). This reaction process, called Bowen's reaction series, will continue until all of the liquid is consumed and the outer rims of plagioclase crystals are dominated by Na-rich plagioclase, and the hydrous phases hornblende and biotite dominate the mafic constituents—those composed of Mg- and Fe-rich minerals—that crystallize. If any liquid remains, a small amount of K-spar, quartz, and muscovite may crystallize.

An original magma of granitic composition likewise would crystallize the highest-melting constituents first: amphibole and/or biotite and plagioclase of intermediate composition (but still with Na > Ca). Crystallization of K-spar and quartz would consume

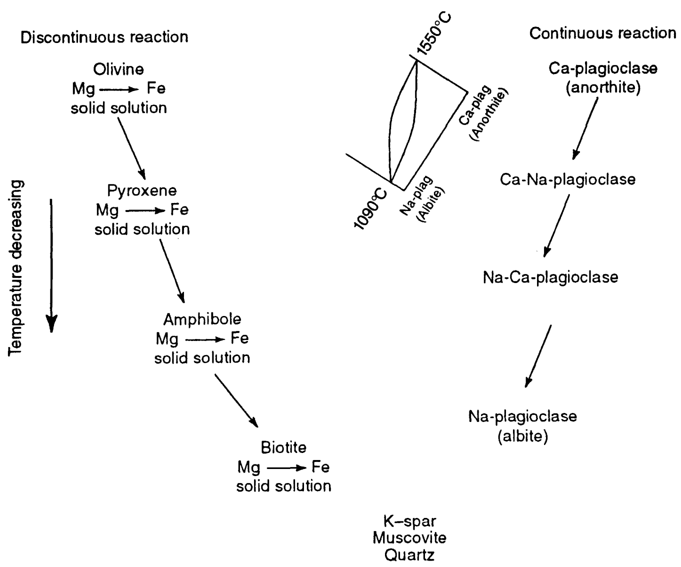
Discontinuous reaction
Olivine
Mg → Fe
solid solution
Pyroxene
Mg → Fe
solid solution
Amphibole
Mg → Fe
solid solution
Biotite
Mg → Fe
solid solution
Temperature decreasing
1550°C
1090°C
Ca-plag (Anorthite)
Na-plag (Albite)
Continuous reaction
Ca-plagioclase
(anorthite)
Ca-Na-plagioclase
Na-Ca-plagioclase
Na-plagioclase
(albite)
K–spar
Muscovite
Quartz

FIG. 6. Bowen's reaction series illustrating phase changes accompanying the crystallization of an initially basaltic magma.

the remaining magma, forming rock of granitic composition.

Variation in crystallization rate determines the texture of the resulting igneous rock. Uniform texture usually indicates a uniform crystallization rate. Presence of megacrysts (large crystals) in a finer groundmass, yielding a porphyritic texture, usually indicates that the crystallization rate changed during crystallization of the magma. Because of the lower viscosity of basaltic magmas (despite their lower water content), porphyritic textures are less common, except in mafic (basaltic) volcanic rocks. Glasses commonly indicate quenching of magma but also indicate faster cooling (not necessarily quenching) of very high-silica (high-viscosity) magma producing obsidian (glass of granitic composition). Basalt glass is relatively uncommon, despite the overall greater abundance of basalt than the more felsic (granitic) volcanic rocks, but obsidian and other glassy rocks are much more common in areas of felsic volcanism.

Large volumes of gases are released during volcanic eruptions. Some of these are primordial gases from the interior of the Earth; the majority of the water in volcanic eruptions since the late Proterozoic, however, is water vapor derived from surface water or near-surface ground water that comes into contact with the rising magma. During the original cooling of the Earth into a solid mass, the atmosphere and hydrosphere may have been derived from outgassing of the water and other gases trapped in the interior. The primordial atmosphere and hydrosphere are thus probably originally by-products of igneous processes.

4. SURFICIAL PROCESSES

Surficial processes modify topographic features that occur above sea level by in-place weathering of rocks and sediments formed in different environments and by moving surficial materials into transport and subsequent deposition. Subsea processes operating on and near the ocean bottoms modify the shape of ocean-floor topography and the distribution of sediments. Geomorphology is the study of surficial processes and features.

4.1 Subaerial Processes

Weathering and erosion are processes that go hand in hand on the surface of the Earth. Weathering may be either chemical or mechanical or a combination of both, determined mostly by climatic conditions and also determined in part by the locations on slopes of the affected materials. Erosion involves the removal and transportation of weathered materials. Rocks weather at different rates, depending on their composition and the weathering environment; they also erode at different rates. For example, the same materials, such as limestone, produce valleys in the humid environment of the eastern United States, but form ridges in arid to semiarid climates of the West, where mechanical weathering and erosion processes dominate.

Chemical weathering, or decomposition, involves chemical breakdown of minerals and the formation of new minerals. Intuitively, one might predict that minerals that form at high temperatures should break down most readily in the surface environment, because high-temperature minerals are far removed from the environment in which they formed. An additional variable is involved, however, so that many high-temperature minerals survive in even intense tropical weathering conditions. This variable is the amount of energy involved in the formation of these minerals, forming strong chemical bonds, and the amount of activation energy needed to transform them to other species under surface conditions. Some minerals, such as quartz (SiO_2), form and are stable in both surface and high-temperature conditions. Feldspars also form under surface conditions in the marine environment but readily break down under tropical weathering conditions to form kaolinite [$Al_2Si_2O_5(OH)_4$] and quartz. Ferromagnesian minerals (e.g., biotite, hornblende, and pyroxene) break down into other clays, quartz, and iron oxides under tropical weathering conditions. The rate of chemical weathering is increased by presence of water with a high or low *p*H, warm temperatures, and a large available surface area (fine grain size).

Mechanical weathering involves the breaking apart—disaggregation—of rock masses into smaller particles without a change of chemical composition of the constituent minerals. For example, a limestone or a granite may ideally be broken apart by frost wedging and plant roots without any change in the amount of calcite ($CaCO_3$) or feldspar, quartz, and biotite, respectively, present in any of the resulting smaller particles. Although mechanical weathering occurs in all environments and aids chemical weathering by increasing surface area, mechanical weathering dominates in arid regions where little water is available for decomposition of minerals, and in arctic and alpine regions where water freezes and thaws frequently, or remains frozen much of the year.

4.2 Gravity-Related Slope Processes

Slope processes driven by gravity may be divided into two groups: those that move slowly and those that move rapidly. Slow-moving processes involve downslope movement under the influence of gravity. The rates of these processes vary from thousands to millions of years and involve the transport of materials from places where they are derived to the site where they are deposited on lower slopes. In a humid climate, soil, including suspended partially decomposed rock, may move slowly downslope. It may move at a very slow rate, producing a tongue of colluvium (unsorted, unconsolidated, nonglacially derived slope material), or rapidly—even catastrophically—if the toe of the colluvial tongue is removed and the slope oversteepened either naturally (e.g., by stream erosion) or artificially (e.g., by highway construction). In dry regions, transport occurs during short but intense rainfall events that move large amounts of material, but the water becomes dispersed on the lower slopes forming an alluvial fan. In arid to semiarid regions, water accomplishes most of the work of erosion and transportation of materials, and, although a desert region may appear static to the casual observer, large amounts of energy are expended over short periods of time to effect erosion, transportation, and deposition. The desert landscape may thus be changed gradually by the effects of many catastrophic events over a period of several thousand years.

Rapid processes in humid regions involve the catastrophic movement of material from upper to more gentle lower slopes. There, oversteepened slopes may occur on the upper slopes and earthquakes, freeze–thaw, or other mechanisms may trigger landslides, wedge failures, rock falls, rock and snow avalanches, and other rapid types of motion.

Glaciers are gravity-driven slow downslope-moving masses of ice that occur either in Alpine regions or as icecaps—continental glaciers. Glaciers are less important agents of erosion, transportation, and deposition today than they have been in the recent geologic past (the Pleistocene epoch). They were also important during the late Paleozoic (Carboniferous deposits in the Gondwanan continents) and early Paleozoic (Ordovician deposits in North Africa and southern South America) as well as during the Upper and Middle Proterozoic. Relict continental glaciers occur in Greenland and Antarctica, and numerous Alpine glaciers exist today mostly in high mountains in different parts of the world. These are remnants of the most recent Pleistocene glacial maximum (actually four maxima separated by four interglacial events) that began to retreat from the northern hemisphere some 8 000–10 000 years ago. A large amount of evidence has been cited by climate model research that "global warming" has occurred during the past several decades and that this is attributable principally to increases of CO_2 and other gases released into the atmosphere from combustion of coal and hydrocarbons (e.g., Houghton *et al.*, 1990, 1992; Jager and Ferguson, 1991). Rigorous climate models, however, span less than two decades—an instant in geologic time. Therefore, it is still not possible to say whether or not the small increases in global climate measured in the last several decades will form a long-term warming trend toward interglacial (or postglacial) times wherein all glacial ice will melt and present climatic zones will be significantly altered, or if the Earth has gone through a short interglacial period, and the climate will return to a glacial maximum during the next 10 to 20 ka. Regardless of the outcome of the debate about the present and near-term scenario, there is little doubt that continued combustion of hydrocarbons producing a buildup of CO_2 and other harmful gases in the atmosphere will be destructive in the long term.

4.3 Stream Processes

Streams erode, transport, and deposit material, continuously change their channels and floodplains, and finally reach the primary base level, or sea level, as they empty into an ocean or sea. Temporary base level may be established naturally where streams flow into larger streams, into natural or artificial lakes, or onto alluvial fans (Fig. 7). These attributes change the dynamics of streams and permit them, at the outlets from lakes or reservoirs, to erode their channels and increase the amount of sediment being transported, because water emerging from natural or artificial lakes normally contains very little sediment but does possess a great deal of potential energy. Therefore, the large amount of potential energy can be converted to kinetic energy to erode the channel and transport sediment below a dam where no energy is presently being consumed by transportation of sediment by water above the dam.

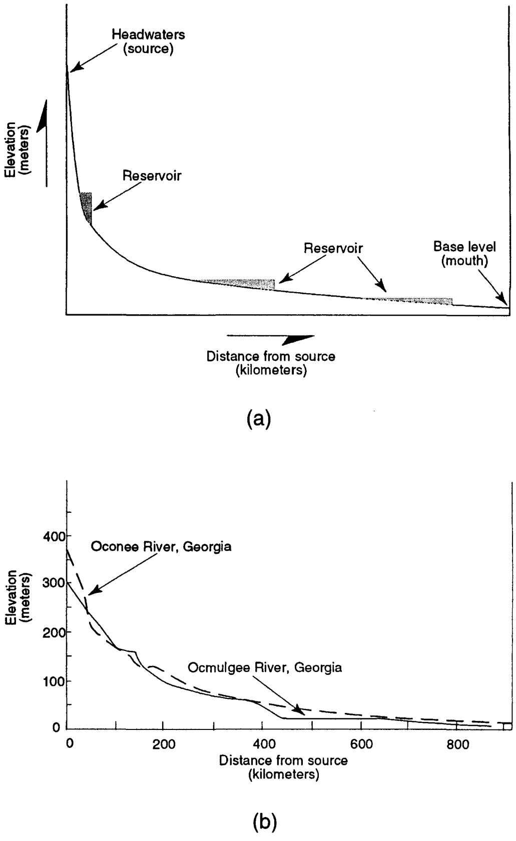

FIG. 7. (a) Equilibrium stream profile showing the superposition of temporary base level by construction of reservoirs. (b) Examples of actual stream profiles of two streams in the Blue Ridge and Piedmont of Georgia. (From Hack, 1982.)

Stream discharge (Q) at a point on a stream is described by $Q = VA$, where V is velocity and A is cross-sectional area. The cross-sectional area, velocity, and discharge of a stream increase downstream, but the stream gradient decreases, a floodplain develops, and the stream channel becomes more sinuous, or meandering. The channel becomes more asymmetric in cross section, deposition occurs on the inner arcs of the channel, and erosion occurs on the outer arcs, thus causing the channel to migrate both laterally (increasing the amplitude of meanders) and downstream. Erosion, transportation, and deposition become linked more to the local dynamics of the stream. A delta may form at the mouth of the stream if the body of water into which the stream empties does not have enough energy to remove the sediment as it is deposited. Deposition is accelerated if the stream flows into the ocean or other body of saltwater, because saltwater flocculates suspended clays and other fine sediment. Large accumulations of sediment occur in deltas of major rivers where strong currents are not present; proportionally smaller deltas form where strong currents do exist or at the mouths of streams with smaller discharge.

5. SEDIMENTATION

Sediment derived from the continents and smaller land masses, such as islands, is ultimately dispersed into the marine environment. This sediment is then dispersed in the ocean by longshore currents, or, as currents build unstable sediment accumulations along the continental slope, gravity-driven turbidity flows may carry sediment for long distances down the continental slopes and into the abyssal plains of the oceans. Most of the sediment deposited in this way is mechanically derived—clastic sediment. It is dispersed and accumulates to varying depths on the continental shelves, slopes, and in the deep-ocean basin.

Shallow seas under the right conditions of temperature, lack of clastic sediment, and low latitude will deposit carbonate sediment. This carbonate sedimentation, if voluminous and assisted by organisms like coral and algae, forms reef-dominated carbonate banks that commonly face the open ocean. In the back-reef area where poor circulation exists, the limestone will be converted to dolomite [$CaMg(CO_3)_2$].

Deep-ocean sediments are predominantly clastics at depths of greater than 4000 m, because the water at greater than these depths is commonly undersaturated with respect to carbonate, and carbonate minerals will redissolve. Most of the sediments of the deep-ocean basins are also relatively fine-grained because of the processes that transport, disperse, and deposit them. In some cases, however, large blocks derived from nearby areas of high topographic relief on adjacent continents or islands, or along even the submarine parts of the oceanic ridges, occur in the deep-ocean basins. Active faults may produce topographic relief and cliffs that will shed blocks into the deep-ocean basins either in island arc systems or along the oceanic ridges. Icebergs that are carried by currents southward or northward from the polar icecaps also transport large blocks that are dropped and become part of the deep-ocean sediments as the icebergs melt. These are relatively easy to identify either from the source material or from the dynamics of accumulation where they disturb the sediment around them and contrast markedly in grain size and otherwise with the enclosing sediments.

6. METAMORPHISM

Metamorphism is a process that involves solid-state recrystallization of minerals largely under the influence of increased temperature. Increased pressure, both lithostatic (confining) stress and simple shear (deviatoric or nonlithostatic) stress, and pressures from minor amounts of fluids (e.g., water and CO_2) can also influence the products of metamorphism, along with the original composition and grain size.

Recognition of the grade, or degree, of metamorphism—assessment of the T and P conditions of metamorphism—is made by means of characteristic minerals and mineral assemblages recognized either in the field, under the microscope, or both, as well as determination of the compositions of particular minerals in metamorphic rocks in the laboratory (Fig. 8). Determination of metamorphic grade can be most easily accomplished in metamorphosed shales because they form characteristic assemblages over a broad range of temperatures and pressures. For example, the suite of characteristic minerals formed in shale from lowest to highest grade under conditions of increasing temperature and pressure is chlorite, biotite, garnet, stau-

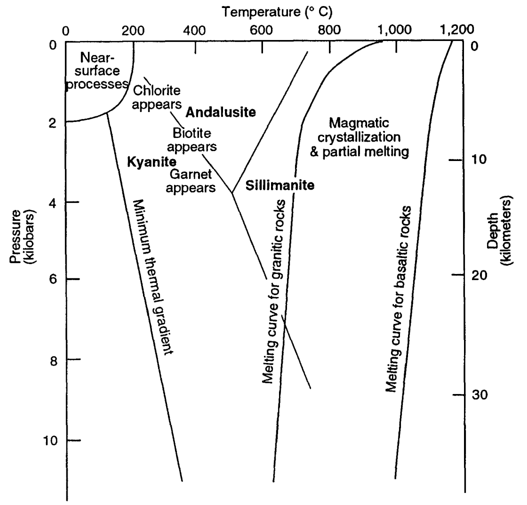

FIG. 8. Relationships between pressure (depth), temperature, and the formation of several characteristic mineral assemblages in metamorphic rocks.

rolite, kyanite (Al_2SiO_5), sillimanite (a high-temperature moderate-pressure polymorph of kyanite), and sillimanite-K-spar. If only temperature increases, the suite is chlorite, biotite, garnet, andalusite (another high-temperature, but low-pressure, polymorph of kyanite), cordierite [$Mg_2Al_3(AlSi_5O_{18})$], and K-spar. Impure carbonate rocks are nearly as useful as shales. Sequences of characteristic minerals in other rock types are not as diverse, so where shales are absent or rare it may not be possible to determine the metamorphic grade as exactly. Regardless of the rock types present, assessment of metamorphic grade is best made using the minerals and mineral compositions present in the entire assemblage under investigation.

In addition to the suites and conditions outlined above, low-temperature (<300 °C), high-pressure regimes exist in the near-surface parts of subduction zones that produce other characteristic minerals like the Na-amphibole glaucophane. Very high-pressure, high-temperature metamorphism of basalt, like that in the upper mantle, produces an assemblage of Mg–Al garnet (pyrope) and a pyroxene (omphacite) that does not form in the near-surface regimes of the middle to upper crust. Moreover, there are ultrahigh-pressure polymorphs of quartz that form in impact structures at near-surface temperatures.

As metamorphic grade (*T*,*P*) increases, dehydration increases the number of anhydrous phases and decreases the number of hydrous minerals, until at the highest grades (high *T*, ±*P*) only a few hydrous minerals, like biotite, are stable. At lower grades (low *T*, ±*P*), micas and amphiboles are abundant; as the grade increases, however, micas are converted to feldspars, and the aluminum silicates (kyanite, andalusite, and sillimanite, Al_2SiO_5) and amphiboles (e.g., actinolite, tremolite, and hornblende) are converted to pyroxenes (e.g., diopside and hypersthene). Feldspars and quartz persist at all grades, with quartz being either the product of several metamorphic reactions or consumed as a starting material in many reactions. An example of a common metamorphic reaction involves impure dolostone as a starting material, composed of dolomite, quartz, and clay minerals, which reacts to form wollastonite ($CaSiO_3$), diopside [$(Ca,Mg)_2Si_2O_6$], and either grossularite [$Ca_3Al_2(SiO_4)_3$] or andradite [$Ca_3Fe_2(SiO_4)_3$]; both of the latter are garnets. Water released from clay minerals may initially form hydrous minerals like phlogopite (a mica) or tremolite (an amphibole), but these too dehydrate at higher temperatures.

Regional metamorphism, the most widespread metamorphic process, involves recrystallization under conditions of increased temperature, both confining and directed (simple or pure shear) stress, and the presence of fluids. Water produced by dehydration of clays is commonly a product of metamorphic reactions and can affect greater mobility of elements, particularly during the early stages (lower grades) of metamorphism. Because regional metamorphism and deformation are linked in the formation of mountain chains during subduction, terrane accretion, and collisional processes, numerous structures, ranging from foliations (textures imparted by the parallel orientation of platy or elongate minerals) and lineations to folds and ductile faults, form during metamorphism.

Contact metamorphism is thermally driven and is commonly related to zones where magma and the host rocks are in closest contact. Zones of contact metamorphism range from a few centimeters to several kilometers thick, depending on the amount of heat and fluid available, and the bulk compositions of both the magma and the host rocks. Basaltic magmas are hotter but are commonly dry and less voluminous; granitic magmas are cooler but usually contain more water and are more voluminous. Impure carbonate and fine-grained clastic rocks, because of their diverse composition and fine grain size, produce the most extensive contact metamorphic zones, whereas pure sandstones and intrusive igneous rocks produce the least.

7. DEFORMATION PROCESSES

Most rock deformation within the Earth or the other planets is the product of tectonic forces. Within the Earth, deformation is driven by plate tectonic processes (Figs. 2 and 3; also see PLATE TECTONICS), which are in turn driven by processes (ultimately thermally driven flow) that produce convection in the mantle. Deformation can be defined in terms of strain: the change in length of reference lines (elongation strain, $\epsilon = dl/l$), change of volume ($\Delta = dV/V$), or distortion of the angular relationships within a reference object (shear strain $\gamma = \tan\psi$, where ψ is the rotation angle). It can also be defined in terms of displacement. In the simplest sense, forces affecting deformation can be thought of as occurring in two forms: compression, in which particles are forced closer together, and tension, in which particles are pulled farther apart. Compression may be produced by gravity-driven vertical forces related to depth of burial either beneath the atmosphere (atmospheric pressure, 1 bar at sea level), under water (hydrostatic pressure that varies with water depth), or within the Earth (lithostatic, related to the thickness of the column of rock above). Compression may also be produced by laterally directed tectonic forces related to plate motion. Tension (tensile deviatoric stresses) may be produced tectonically in parts of the crust, such as at the midocean ridges or the Basin and Range in western North America, where the crust is being pulled apart, and by gravitational forces that produce near-surface readjustments in the upper crust (such as the north-dipping detachment related to gravitational collapse of the high Himalaya), flow in glaciers, and in meteorite impact structures related to elastic rebound following impact.

Deformation in the shallow crust is commonly brittle and produces only granulation (cataclastic deformation) or pressure solution (diffusive mass transfer at grain boundaries) without any recrystallization or other internal deformation of constituent grains composing the rock. Granulation is commonly the product of linear-elastic behavior described by Hooke's law,

$$\epsilon \propto \sigma \quad \text{or} \quad \epsilon = Y\sigma,$$

where ϵ is strain, σ is stress, and Y is Young's modulus, one of several elastic constants that may be used to convert the proportionality to an equality.

Pressure solution is another process important in deformation below 400 °C in the presence of fluid. Quartz and calcite are generally the most soluble common rock-forming minerals. Under shallow-burial conditions of less than 1 km, irregular surfaces called stylolites form in limestones, with the orientation of the surfaces generally parallel to bedding, indicating the force that produced the dissolution was oriented vertically. Similar structures form in sandstones at greater depth and in both rock types under tectonic compression at greater depth. Tectonic stylolites are differentiated from those forming under static burial load by forming at a high angle to bedding. A flow law for pressure solution (assuming Newtonian behavior) is

$$\dot{\epsilon} = \frac{KD_{gb}\Omega C_0}{x^3 kT\rho}\tau_d,$$

where $\dot{\epsilon}$ is strain rate (rate of deformation; geologic rates are commonly on the order of 10^{-6} to 10^{-9} sec^{-1}), K is a material constant, D_{gb} is the coefficient of grain-boundary diffusion, Ω is molar volume, C_0 is concentration of the solution outside the grain boundary, x is grain size, k is Boltzmann's constant, T is absolute temperature, ρ is density, and τ_d is differential (shear) stress (Rutter, 1976). This, and the other flow law presented below, relates strain rate to deviatoric (shear) stress rather than to normal stress because of the complex nature of geologic systems.

Deformation in the lower crust, the mantle, and slow deformation of ice and rock salt at the surface occur plastically where the constituent minerals in a rock undergo recrystallization and the rock flows by a variety of mechanisms (most commonly dislocation creep within individual grains). A flow law that describes the most common recrystallization mechanism in rocks, or dislocation creep, is

$$\dot{\epsilon} = \dot{\epsilon}_0 \tau^n e^{-(Q/RT)}$$

where $\dot{\epsilon}_0$ is the initial strain rate for the crystal being deformed, n is the stress exponent, defined as $d\ln\dot{\epsilon}/d\ln\tau$, Q is the apparent activation energy for the process, R is the universal gas constant (1.987 $\text{cal/deg}^{-1}/\text{mol}^{-1}$), and T is absolute temperature (Poirier, 1985).

The change from brittle to ductile deformation in the crust and mantle is partially a function of composition but is more a function of temperature (Fig. 9). The boundary is not sharp but is a transition zone where the change gradually takes place, and it is called the ductile–brittle transition, or DBT.

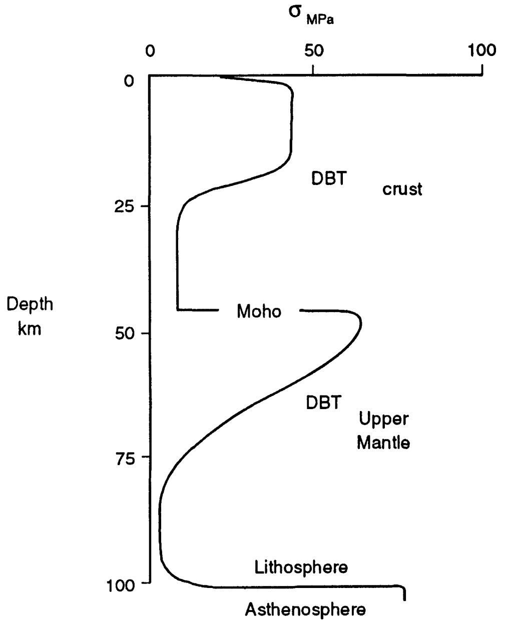

FIG. 9. Stress vs. depth curve showing possible relationships of behavior of quartz (in the crust) and olivine (in the mantle) as a model for brittle-plastic behavior of crustal and mantle rocks. Variation in differential stress is directly related to the strength of the material. DBT: ductile–brittle transition. (From Hatcher, 1995, © 1995 Prentice Hall Inc.)

7.1 Physical Properties of Rocks

Mineral composition of a rock (mineralogy) is a primary control of the physical properties of rocks, but grain size and shape, layering, composition of layers (all layers the same or different), layer thickness, temperature, confining pressure, size and interconnectedness of pore spaces between grains, and presence or absence of fluids all determine the physical properties of rocks. In order to minimize and control as many of these variables as possible, until recently most experimental work has been conducted using monomineralic rocks, rocks with uniform grain size and texture, and rocks without many other anisotropies.

Massive rocks—rocks that have properties that are largely isotropic—deform differently than rocks that contain anisotropies (e.g., bedding or foliation). Layer thickness also has a strong influence on properties, as does the presence of layers of contrasting composition.

Deformation itself may play a role in the subsequent properties of rocks, particularly shallow depths where brittle deformation dominates. Limestone, sandstone, and shale have markedly different properties under most conditions. If, however, each is deformed dry at surface (or near-surface) conditions under ideal (experimental) conditions, and then the products deformed a second time, the strength of the shale would remain nearly constant, and the strengths of both the limestone and sandstone would be much lower for the second cycle of deformation. The reason for this marked change in behavior under surface conditions is that neither the grain size nor porosity of the shale change markedly, whereas the average particle size of the limestone and sandstone both decrease markedly, and the porosity increases. Additional changes may be noted with the addition of fluid during these experiments.

7.2 Folds

Folds are wavelike structures that result from the distortion of originally planar structures—bedding, foliations, faults, joints, or other surfaces—into a curved shape (Hudleston, 1986). They occur on all scales and form in all deformational environments, from brittle to mixed brittle–ductile to ductile environments. Factors that affect the formation of folds, fold wavelength, amplitude, and fold mechanisms include the inherent properties of the rocks (composition, texture, and anisotropy), layer thickness, whether or not layers will slip past one another, and the ambient pressure–temperature conditions of the rock mass. Stress orientation that produces the folds is also a factor, wherein forces may be applied parallel to (producing buckle folds) or across the layering (producing bending folds). Whether or not layers are confined, layer contrast—the relative viscosity of layers—and the thickness of layers can also affect the shapes and mechanisms of the folds that form (Johnson, 1977). Changes in temperature, pressure, and presence of fluids also change the rheological state of the rocks. The most commonly recognized mechanisms include buckling (Currie *et al.*, 1962), where the stresses are applied from the ends of a layer or beam; bending, in which the stresses are applied perpendicular to the beam; kinking, in which the stresses are applied from the ends but the layering cannot slip past itself (Johnson and Ellen, 1974); and various flow mechanisms in which layering either participates to effect the shape of the structures or does not participate.

Folds produced by bending of a layer that is pinned at both ends can be described by $dM/dx = V_a x$ or $M = V_a x +$ constant, where M is deflection, V_a is vertical force, and x is amount of shortening.

A direct relationship exists between the wavelength of folds formed by buckling and the thickness of the dominant strong layer in the sequence (Currie *et al.*, 1962) and can be expressed by

$$\lambda_d = 2\pi t(\mu_1/6\mu_2)^{1/3},$$

where λ_d is the dominant wavelength (of folds in the strong layer), t the thickness of the strong layer, μ_1 the viscosity of the strong layer, and μ_2 the viscosity of the weak supporting medium or matrix.

Depending on the properties of the rocks, nature and properties of layers, temperature, and pressure, folds may form that have a very uniform, gently curved geometry, or they may form as very tight and closely spaced structures. Within the same fold there may be layers that change thickness, while others do not, depending on the relative ductility of individual layers. Other folds may exist where all layers either change thickness by the same amount (similar folds) or do not change thickness at all and produce folds in which the layers are all parallel to one another.

The nucleation and propagation of folds in initially undeformed layers have been the subject of numerous investigations and experiments (Dubey and Cobbold, 1977). Again, depending on conditions, the folds that form may transfer motion to another, they may involve the formation of additional folds, or, if the deformation proceeds far enough, they may be superposed by folds having the same trend or crossing trends if the orientation of the strain field changes. Most folds change properties and geometry along their length, providing evidence of a component of simple shear involved in formation of most folds. Where a large component of simple shear is directed perpendicular to the initially formed fold hinges (the hinge is the part of the fold that has the greatest curvature), a large amount of transport commonly develops in the direction from which the original strain was applied, producing folds that are very elongate parallel to the direction of transport called strongly noncylindrical to sheath folds (Fig. 10).

Depending on the orientations of principal strain axes ($X > Y > Z$, related to the strain ellipsoid) through time, deformation may produce folds that interfere with one another like waves on a water surface and produce the combined effects of two or more events. Many different options arise for creating different geometric variations in the shapes of folds based on the angular relationships between the first and second deformations. Ideally, there are three fundamental fold-interference types (Ramsay, 1967):

1. Type I: The axial surfaces of folds were vertical in both cases with respect to each other, and the deformation occurred at right angles, producing fold hinges at right angles to each other; the interference pattern resembles an egg carton or a series of domes and basins.
2. Type II: The axial surfaces of the first folds are nearly horizontal, and those of

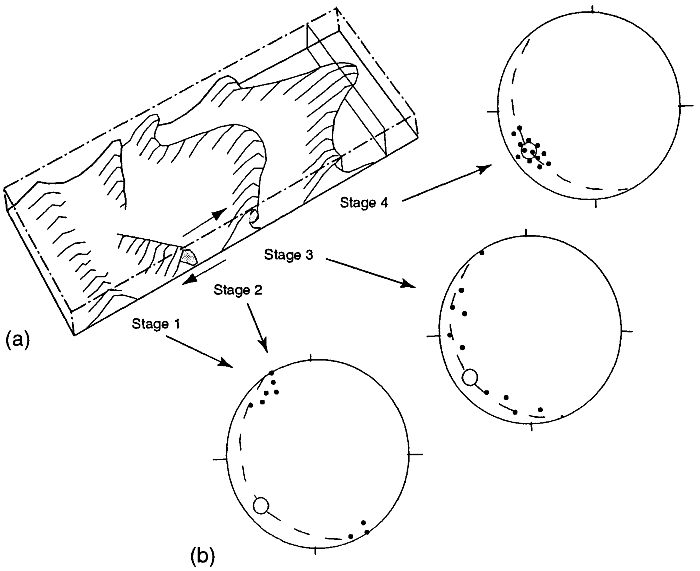

FIG. 10. (a) Progressive amplification of folds (stages 1–4) from cylindrical to sheath geometry in a shear zone, with (b) the corresponding fabric diagrams for each fold. Note the clustering of hinges of the noncylindrical folds about great-circle girdles (dashed lines) that approximate the shear plane. The open circles represent the sheath-fold axes. (Modified from Hudleston, 1986; from Hatcher, 1995, © 1995 Prentice Hall Inc.)

the second set of folds are nearly vertical with respect to the first; the hinges are oriented at right angles to each other, producing a boomerang pattern.

3. Type III: The fold hinges both have the same trend, and the later folds continue to fold and refold the first folds, producing a hook pattern in outcrop, or jellyroll structure, as the layering simply is rolled about itself.

Thiessen and Means (1980) determined that there are at least 16 possible alternative geometries that might be produced with variations in the three basic types described above.

7.3 Fractures and Faults

Geologists use the term fracture in ways different from the usage employed by physicists, probably more akin to that employed by materials scientists: We call any surface across which a rock has lost cohesion a "fracture" (a crack). Following that usage, geological fractures include joints, shear fractures, and faults. Fractures are thought to form by three different mechanisms (Fig. 11): Mode 1 involves opening and therefore a tensional field (joints); mode 2 involves sliding parallel to the fracture surface; and mode 3 involves tearing, again parallel to the fracture surface, but the motion is perpendicular to that in mode 2 (Pollard and Aydin, 1988). Both mode 2 and mode 3 fractures are analogs to the three classes of faults, although there is no requirement here of significant displacement parallel to the fracture surface. There is also another analog with the propagation of a dislocation surface that involves comparing motion of an edge dislocation to mode 2 fractures and the motion of mode 3 fractures to the propagation of screw dislocations (Fig. 12).

Faults are fractures along which there has been appreciable movement parallel to the plane of the fracture. Fractures (mode 1 joints) are structures that form in the brittle realm that have no displacement parallel to the plane of the fracture and have all of the movement perpendicular to it. We recognize three basic types of faults—thrust, normal, and strike-slip—but combinations with

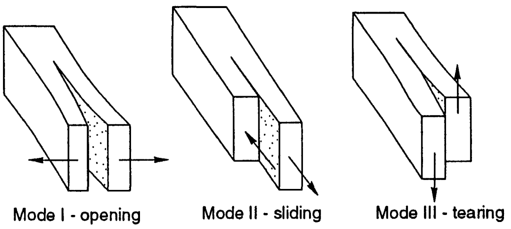

FIG. 11. Three modes of fracture formation. (From Hatcher, 1995, © 1995 Prentice Hall Inc.)

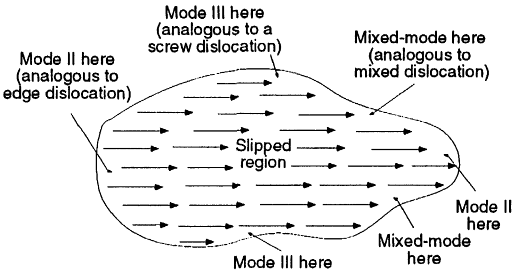
Mode III here (analogous to a screw dislocation)
Mode II here (analogous to edge dislocation)
Mixed-mode here (analogous to mixed dislocation)
Slipped region
Mode II here
Mixed-mode here
Mode III here

FIG. 12. Mode II and III behaviors may occur during fracturing on different parts of the same shear fracture, analogous to the formation of screw and edge dislocations. (From Hatcher, 1995, © 1995 Prentice Hall Inc.)

oblique slip are probably more common than the pure end-member types. In an attempt to understand rigorously the mechanics of faults and how they form, Anderson (1942, 1951) defined the three groups of faults and the mechanical conditions that would be required to produce each one. Inherent in each of his definitions is a standard state that he defined as a "condition of stress which is the same in all directions at any point and equal to that which would be caused by the weight of the superincumbent material [the overlying rock mass, lithostatic stress] across a horizontal plane, at the particular level in the rock." It is assumed in this definition that the surface is flat and that the strata are of uniform specific gravity. He also made several other assumptions that are applicable to understanding the mechanics of how faults form. These assumptions oversimplify the Earth or other planets but are essential for a uniform definition of mechanics:

1. The crust was initially intact and unfractured.
2. One of the principal stresses is vertical.
3. The two other principal stresses must be horizontal and are perpendicular to the first and to each other. They also probably maintain a constant orientation for a long period of time.
4. Forces must balance each other in a body in equilibrium (Newton's third law).

In Anderson's ideal model, he defined three basic stress conditions that would produce the deviatoric stresses necessary to produce faulting:

1. Thrust faulting: Principal stresses σ_1 and σ_2 are horizontal, the least principal stress σ_3 is vertical, and the dip on the fault plane will be $45° - \phi/2$ (about 30°; ϕ is the angle of internal friction, tan ϕ is the coefficient of internal friction).
2. Strike-slip faulting: The maximum and minimal principal stresses σ_1 and σ_2 are horizontal, implying that the intermediate stress σ_2 is vertical and that the fault plane will have a near-vertical dip.
3. Normal faulting: Principal stress σ_1 is vertical; therefore, the fault plane will dip $45° + \phi/2$ (about 60°).

In a classical sense, thrust faults and strike-slip faults are thought to form in compressional environments and normal faults to form in tensional environments. While this conclusion is generally correct, we know that most faults are not ideal thrust, normal, or strike-slip faults but have components of each with one dominating in the fault system. We also know that the other types may form within a fault system that is dominated by one fault type simply by changing the orientation of the fault plane, changing the position of the fault relative to external blocks, or causing the fault to step over from one segment to another, thereby producing a zone of tension or compression in an otherwise compressional or tensional environment.

We commonly think of the motion on faults as being the product of frictional sliding involving either stick-slip or stable sliding. Both are produced by the requirement that the fault will move if the stresses applied are sufficient to overcome the coefficient of friction (μ) of the fault surface or to break the rock if the rock mass was not previously broken. Renewed movement on the fault is produced by overcoming the coefficient of friction again and again. Faults may move by sudden motion that produces earthquakes, called stick slip, where the energy that is accumulated along the fault zone is relieved catastrophically and elastically so that the rebound of the blocks on either side of the fault produces a sudden release of elastic energy in the form of an earthquake. The alternative is stable sliding, where a fault moves continuously to relieve the accumulated elastic strain so that the deformation occurs uniformly through time and the accumulated strain is released without catastrophic rebound. Both mechanisms may occur in segments of the same fault, or one may dominate along the entire length of a fault, although movement does not occur along the entire length of a fault at the same time. Large faults such as the San Andreas commonly have segments that are undergoing stick-slip–type deformation and other segments that undergo stable sliding (Sylvester, 1988).

The motion on faults may be enhanced by the presence of fluid along the fault zone. Fluid pressure generally reduces the coefficient of friction on the fault and promotes motion at lower values of accumulated elastic strain than would otherwise occur. Terzaghi (1923) first recognized the importance of fluids in his studies of soil mechanics and introduced the concept of effective normal stress ($\sigma_n - P_w$, where P_w is the buoyant pressure of water in the material or along a fault) to modify Amontons' first law to $\tau = \mu \times (\sigma_n - P_w)$. In detail, motion on a fault occurs whenever the coefficient of friction is overcome, which physically means that the fault is moving. For this to occur, sometimes microscopic, sometimes visible, irregularities called asperities serve as barriers to motion and must be either broken or bypassed for the fault to move. If fluid is present in the fault zone, the stress may be borne by the fluid and fewer asperities broken for movement to occur.

Faults in the outer several kilometers of the crust move by breaking asperities, and deformation proceeds by a brittle mechanism. Movement along faults in the deeper crust occurs along zones of ductile deformation below the DBT. The position of the DBT is a function of temperature, combinations of temperature and fluid pressure, or these combined with other factors. Ironically, we are learning today that many large earthquakes, such as the deep-focus (640 km) magnitude 8.2 Bolivian earthquake in June 1994, are nucleated at depths where rocks are quite hot and should be relieving strain by aseismic creep, indicating that Amontons' law (brittle failure) is not the only mechanism by which earthquakes may be generated. For example, earthquakes between 75 and 300 km in depth may be produced by rapid (catastrophic) dehydration of serpentine [$(Mg,Fe)_3Si_2O_5(OH)_4$] to olivine [$(Mg,Fe)_2SiO_4$] (Meade and Jeanloz, 1991). Earthquakes at depths greater than 300 km are today thought to occur by nucleation of anticracks (lensoidal microfractures that form perpendicular to the greatest principal stress, σ_1) during the phase change from olivine to an equivalent Mg (–Fe) silicate with a much denser spinel ($MgAl_2O_4$) structure (Green *et al.*, 1990).

Direction of motion on a fault may be determined by several means. At the map scale, faults may be identified by displacements of rock units, or generation of earthquakes if the fault is active today, and by displacement of topographic features and other structures. On the microscopic to hand-specimen scale, features are developed within fault zones that also indicate motion. Grooves or other features on a brittle fault surface may indicate direction of motion; rotation sense on small folds that are developed in association with the fault zone are very good indicators of motion sense on the fault. In addition, the presence of rotated porphyroclasts (relicts of mechanically resistant minerals, such as feldspars and garnets) and other features has been used to some advantage, but it must be used with care in the determination of motion. In addition, shear foliations may develop in fault zones at an angle to existing foliation that will indicate the sense of motion on the fault. The angular relationships of the shear foliation relative to the original foliation indicate shear sense on the fault.

7.4 Foliations

Foliations consist of one or more structures that permit rocks to split along planes other than bedding. These include cleavage, schistosity, gneissic banding, and the general term foliation. Foliations may be divided into two different groups. One group consists of structures that do not permeate the entire rock mass. These are called nonpenetrative foliations and consist of structures that are developed within fault zones only locally or that have a wide spacing in a rock mass. The other kind—penetrative—is pervasive throughout the entire rock mass on all scales and includes slaty cleavage, schistosity, and the foliations observed in gneisses and consisting of parallel alignment of nonequant minerals (e.g., micas, amphiboles, and quartz). This is the classic foliation of most rocks and has been known for several hundred years. On the microscopic scale, many foliations become spaced and are separated by small areas of undeformed rock. Therefore, they are nonpenetrative on the microscopic scale. There are, however, structures that permeate the rock on the scale of a hand specimen, a road cut, or a map. Slaty cleavage is the foliation that occurs in fine-grained rocks at low metamorphic grade; gneissic banding and schistosity dominate at high grades. All kinds of foliations, except those associated with flow during the latter stages of crystallization of an igneous body, are tectonic structures. Once again, the properties of the rock determine to some extent the spacing and the angular relationships of the cleavage or foliation in adjacent layers. Fine-grained rocks produce very closely spaced penetrative cleavage, whereas coarser-grained rocks involve cleavage at much greater spacing. In addition, the angle of cleavage to some other reference plane such as bedding may change drastically at the interface between one layer composed of a coarse-grained rock like sandstone and another layer composed of a different rock type like shale.

Most foliations that are pervasive throughout large rock masses were traditionally thought to form by pure shear—flattening. Williams (1976), however, challenged this view by demonstrating, because of the variable distribution of strain in folds, that there must be components of simple shear present.

GLOSSARY

Amphibole: Group of orthorhombic and monoclinic, double-chain (SiO_4 tetrahedra) silicate minerals that contain divalent cations (Mg^{II},Fe^{II},Ca^{II}) in particular sites. Some also contain large univalent cations (Na,K) in other sites in structures where Al^{III} has substituted for Si^{IV}, leaving a charge imbalance. Amphiboles form at high temperatures and pressures and thus are common in igneous and metamorphic rocks. Hornblende is one of the more complex amphiboles; tremolite has a simpler structure and composition. Glaucophane is one of the Na-rich amphiboles.

Basalt: Fine-grained, dark-colored igneous rock (commonly volcanic) composed of calcic plagioclase and pyroxene, sometimes with lesser amounts of olivine, amphibole, and biotite.

Clinopyroxene: Pyroxene group mineral that crystallizes with monoclinic symmetry.

Dolostone: Sedimentary rock composed predominantly of the mineral dolomite [$CaMg(CO_3)$].

Fabric Diagram: An equal-area projection of a hemisphere onto a plane circle to represent orientations of plot lines or planes that represent structures in rocks, such as foliations, lineations, or small folds, that contribute to the fabric (texture plus structure) in rocks.

Felsic: Light-colored igneous rock composed of quartz, plagioclase, and potassium feldspar, and one or more micas (occasionally hornblende).

Fold: Geologic structure defined by layering in rocks that is continuously curved. Folds may be very tight, gentle, or disharmonic, with different curvature from layer to layer.

Foliation: Texture in deformed and metamorphosed rocks imparted by the parallel orientation of platy or elongate minerals. Schistosity, gneissic banding, and slaty cleavage are all foliations.

Garnet: Group of silicate minerals that contain isolated SiO_4 tetrahedra and trivalent (Fe^{III},Al^{III},Cr^{III}) and divalent cations (Mn^{II},Mg^{II},Fe^{II},Ca^{II}) in particular sites. Almandine (Fe) is the most common garnet; other common garnets include grossalurite (Ca), spessartine (Mn), and pyrope (Mg).

Gondwana: Supercontinent that existed during the Paleozoic and the latest Precambrian, consisting of Africa, Antarctica, Australia, India, and South America.

Granite: Coarse-grained igneous rock composed of potassium and plagioclase feldspar (K-spar dominant), quartz, and minor amounts of biotite (and/or hornblende, or muscovite).

Hinge: Part of a fold that has the greatest curvature.

Laurentia: North America.

Plagioclase: Silicate mineral group with a composition that ranges continuously from $NaAlSi_3O_8$ to $CaAl_2Si_2O_8$. Structures consist of three-dimensional polymeric networks of SiO_4 tetrahedra, with one Al^{III} substituting for every third Si^{IV} in a four-coordinate site, requiring a univalent or divalent cation to neutralize the charge imbalance. Mostly forms in igneous and metamorphic rocks.

Pyroxene: Group of single SiO_4 tetrahedra chain silicate minerals that commonly contain Mg^{II}, Fe^{II}, and Ca^{II} as principal cations.

Ultramafic: Igneous rock rich in Mg and Fe, poor in K and Na, and containing <40% by weight of SiO_2. The mantle is composed mostly of ultramafic rocks.

Works Cited

Anderson, D. L. (1995), "Lithosphere, asthenosphere, and perisphere," *Rev. Geophys.* **33,** 125–149.

Anderson, E. M. (1942, 1951), *The Dynamics of Faulting and Dyke Formation, with Applications to Britain,* Edinburgh: Oliver and Boyd.

Currie, J. B., Patnode, H. W., Trump, R. P. (1962), "Development of folds in sedimentary strata," *Geol. Soc. Am. Bull.* **73,** 655–674.

Dubey, A. K., Cobbold, P. R. (1977), "Noncylindrical flexural slip folds in nature and experiment," *Tectonophysics* **38,** 223–239.

Green, H. W., II, Young, T. E., Walker, D., Scholz, C. H. (1990), "Anticrack-associated faulting at very high pressure in natural olivine," *Nature* **348,** 720–722.

Hack, J. T. (1982), *U.S. Geological Survey Professional Paper 1265,* Washington, DC: U.S. GPO.

Hatcher, R. D., Jr. (1995), *Structural Geology: Principles, Processes, and Problems,* Englewood Cliffs, NJ: Prentice-Hall.

Hatcher, R. D., Jr., Williams, R. T. (1986), *Geol. Soc. Am. Bull.* **97,** 975–985.

Houghton, J. T., Jenkins, J. G., Ephramus, J. J. (1990), *Climate Change: The IPCC Scientific Assessment,* Cambridge, U.K.: Cambridge Univ. Press.

Houghton, J. T., Callender, B. A., Varney, S. K. (1992), *Climate Change 1992: The Supplementary Report to the IPCC Scientific Assessment,* Cambridge, U.K.: Cambridge Univ. Press.

Hudleston, P. J. (1986), "Extracting information from folds in rocks," *J. Geol. Ed.* **34,** 237–245.

Jager, J. Ferguson, H. L. (1991), *Climate Change: Science, Impacts, and Policy,* Cambridge, U.K.: Cambridge Univ. Press.

Johnson, A. M. (1977), *Styles of Folding: Mechanics and Mechanisms of Folding of Natural Elastic Materials,* Amsterdam: Elsevier.

Johnson, A. M., Ellen, S. D. (1974), "A theory of concentric, kink, and sinusoidal folding and of monoclinal flexuring of compressible, elastic multilayers. I. Introduction," *Tectonophysics* **21,** 301–339.

Meade, C., Jeanloz, R. (1991), "Deep-focus earthquakes and recycling of water into the Earth's mantle," *Science* **252,** 68–72.

Poirier, J.-P. (1985), *Creep of Crystals: High-Temperature Deformation Processes in Metals, Ceramics, and Minerals,* Cambridge, U.K.: Cambridge Univ. Press.

Pollard, D. D., Aydin, A. (1988), "Progress in understanding joints over the past century," *Geol. Soc. Am. Bull.* **100,** 1181–1204.

Ramsay, J. G. (1967), *Folding and Fracturing of Rocks,* New York: McGraw-Hill.

Rutter, E. H. (1976), "The kinetics of rock deformation by pressure solution," *Philos. Trans. R. Soc. A* **283,** 203–219.

Sylvester, A. G. (1988), "Strike-slip faults," *Geol. Soc. Am. Bull.* **100,** 1666–1703.

Terzaghi, K. V. (1923), "Die Berechnung der Durchlässigkeitsziffer der Zonen aus dem Verlauf der hydrodynamischen Spannungserscheinungen," *Sitzungsber. Akad. Wiss. Wien* **132,** 105.

Thiessen, R. L., Means, W. D. (1980), "Classification of fold interference patterns: a reexamination," *J. Struct. Geol.* **2,** 311–316.

Turcotte, D. L., Emerman, S. H. (1985), "Magma fracture as a mechanism for magma migration," *Eos* **66,** 361.

Williams, P. F. (1976), "Relationships between axial-plane foliations and strain," *Tectonophysics* **30,** 181–196.

Further Reading

Anderson, D. L. (1989), *Theory of the Earth,* Boston: Blackwell Scientific Publications.

Aydin, A., Nur, A. (1982), "Evolution of pull-apart basins and their scale independence," *Tectonics* **1,** 91–105.

Best, M. G. (1982), *Igneous and Metamorphic Petrology,* San Francisco: W. H. Freeman Co.

Chapple, W. M. (1978), "Mechanics of thin-skinned fold-and-thrust belts," *Geol. Soc. Am. Bull.* **89,** 1189–1198.

Chorley, R. J., Schumm, S. A., Sugden, D. E. (1984), *Geomorphology,* New York: Methuen Inc.

Davis, D., Suppe, J., Dahlen, F. A. (1983), "Mechanics of fold-and-thrust belts and accretionary wedges," *J. Geophys. Res.* **88,** 1153–1172.

Green, H. W., II (1994), "Solving the paradox of deep earthquakes," *Sci. Am.* **265** (9), 64–70.

Hatcher, R. D., Jr. (1995), *Structural Geology: Principles, Processes, and Problems,* 2nd ed., Englewood Cliffs, NJ: Prentice Hall.

Kerrich, R., Allison, I. (1978), "Flow mechanisms in rocks: microscopic and mesoscopic structures, and their relation to physical conditions of deformation in the crust," *Geosci. Canada* **5,** 109–118.

Knipe, R. J. (1989), "Deformation mechanisms—recognition from natural tectonites," *J. Struct. Geol.* **11,** 127–146.

Marder, M., Fineberg, J. (1996), "How things break," *Phys. Today* **49** (9), 24–29.

McClay, K. R. (Ed.) (1992), *Thrust Tectonics,* London: Chapman and Hall.

Nicolas, A. (1987), *Principles of Rock Deformation,* Dordrecht: D. Reidel Publishing.

Raymond, L. A. (1995), *The Study of Igneous, Sedimentary, and Metamorphic Rocks,* Dubuque, IA: W. C. Brown Co.

Scholz, C. H. (1990), *The Mechanics of Earthquakes and Faulting,* Cambridge, U.K.: Cambridge Univ. Press.

Selby, M. J. (1985), *Earth's Changing Surface,* Oxford: Oxford Univ. Press.

Turcotte, D. L., Schubert, G. (1982), *Geodynamics: Applications of Continuum Mechanics to Geological Problems,* New York: Wiley.

STRUCTURE DETERMINATION BY NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

A. MARGARET CHIPPENDALE, *ZENECA Specialties, Manchester, United Kingdom*

Introduction 120
1. **Principles of Nuclear Magnetic Resonance** 120
1.1 Nuclei in a Stationary Magnetic Field 120
1.2 Basic Principles of the NMR Experiment 121
2. **High-Resolution Solution NMR Spectroscopy** 122
2.1 Spectral Parameters 123
2.1.1 Chemical Shift 123
2.1.2 Spin–Spin Coupling 124
2.1.3 Signal Intensity 127
2.2 NMR and Structure 128
2.2.1 Hydrogen (^{1}H and ^{2}H) 128
2.2.1.1 Chemical Shifts 129
2.2.1.2 Spin–Spin Coupling 130
2.2.2 Carbon (^{13}C) 131
2.2.2.1 Chemical Shifts 131
2.2.2.2 Spin–Spin Coupling 131
2.2.3 Fluorine (^{19}F) 133
2.2.3.1 Chemical Shifts 133
2.2.3.2 Spin–Spin Coupling 133
2.2.4 Phosphorus (^{31}P) 133
2.2.4.1 Chemical Shifts 133
2.2.4.2 Spin–Spin Coupling 134
2.2.5 Nitrogen (^{14}N and ^{15}N) 134
2.2.5.1 Chemical Shifts 134
2.2.5.2 Spin–Spin Coupling 134
2.2.6 Oxygen (^{17}O) 135
2.2.6.1 Chemical Shifts 135
2.2.6.2 Spin–Spin Coupling 136
2.2.7 Silicon (^{29}Si) 136
2.2.7.1 Chemical Shifts 136
2.2.7.2 Spin–Spin Coupling 137
2.3 Double-Resonance Techniques 137
2.3.1 Homonuclear Spin Decoupling 137
2.3.2 Heteronuclear Spin Decoupling 137
2.3.2.1 Selective Spin Decoupling 137
2.3.2.2 Broadband Spin Decoupling 137
2.3.2.3 Gated Decoupling 138
2.3.2.4 Off-Resonance Decoupling 138
2.3.3 NOE Difference Spectroscopy 139
2.4 Spectral Editing Experiments 140
2.5 Multidimensional NMR 142
2.5.1 Basic Principles of 2D NMR 142
2.5.2 Homonuclear Chemical-Shift Correlation 142
2.5.3 Homonuclear *J*-Resolved Spectra 144
2.5.4 Heteronuclear Chemical-Shift Correlation (HETCOR) 144
2.5.5 Heteronuclear *J*-Resolved Spectra 145
2.5.6 Homonuclear NOE Correlation 146
2.5.7 3D and 4D NMR 146
2.6 Application to Chemical Structure Determination 147
3. **NMR of Solids** 148
3.1 Magnetic Interactions in the Solid State 149
3.1.1 Dipole–Dipole Interaction 149
3.1.2 Chemical-Shift Anisotropy 149
3.1.3 Quadrupole Interactions 149
3.2 NMR Spectra of Solids 150
3.2.1 Abundant Spin-$\frac{1}{2}$ Nuclei 150
3.2.2 Dilute Spin-$\frac{1}{2}$ Nuclei 150
3.2.3 Quadrupolar Nuclei 151
3.3 High-Resolution NMR of Solids 151
3.3.1 Resolution-Enhancement Techniques 151
3.3.2 Dilute Spin-$\frac{1}{2}$ Nuclei 152
3.3.3 Abundant Spin-$\frac{1}{2}$ Nuclei 153
3.3.4 Two-Dimensional Experiments with Spin-$\frac{1}{2}$ Nuclei 153
3.3.5 Quadrupolar Nuclei 154
Glossary 156
Works Cited 156
Further Reading 158

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

It is now nearly 50 years since the phenomenon of nuclear magnetic resonance (NMR) was first discovered. For much of that time NMR has, arguably, been regarded as the single most important tool for obtaining detailed information about chemical structures. The continuing development of NMR has resulted from a series of step changes. NMR was first developed using continuous-wave (cw) methodology. By the late 1960s the development of pulsed Fourier-transform (FT) techniques extended the scope of NMR, enabling many low-abundance nuclei such as ^{13}C to be observed. By the late 1970s two-dimensional solution NMR techniques were allowing ever more complex molecules to be studied. In the last two decades NMR techniques have been developed that obtain solution-type spectra from solids. Over the same time scale the technique of magnetic resonance imaging (MRI), which has revolutionized diagnostic medicine, has also been developed. This is described in a separate article, MAGNETIC RESONANCE IMAGING.

The first part of this article contains a very brief summary of the principles of NMR and the NMR experiment. Reflecting the major role that solution NMR plays in the determination of chemical structures, the main part is concerned with high-resolution solution NMR. The parameters measured, the nuclei studied, the experiments used, and some applications are discussed. The final section covers NMR of solids. This briefly introduces the additional interactions observed in the solid state and summarizes the type of information that can be obtained from both low- and high-resolution experiments.

1. PRINCIPLES OF NUCLEAR MAGNETIC RESONANCE

To exhibit magnetic resonance the system concerned must have a magnetic moment. For NMR this is provided by a nucleus with a nonzero nuclear spin. The nuclear spin I of an atomic nucleus has one of the values 0, $\frac{1}{2}$, 1, $\frac{3}{2}$, etc. in units of $h/2\pi$, where h is the Planck constant. The mass and atomic numbers of nuclei constrain the possible nuclear-spin values as shown in Table 1. Thus some important, common nuclei such as ^{12}C and ^{16}O that have both even mass and atomic numbers have zero spin. Nuclear magnetic moments μ_N are proportional to the magnitude of the spin, i.e.,

Table 1. Nuclear properties.

Mass number	Atomic number	Nuclear spin I
Odd	Even or odd	$\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots$
Even	Even	0
Even	Odd	1, 2, 3, . . .

$$\mu_N = \gamma_N hI/2\pi, \tag{1}$$

where γ_N is the magnetogyric ratio of the nucleus. In SI units μ_N is in ampere meter^2 and γ_N in radian tesla^{-1} second^{-1}. Alternatively, the magnetic moment can be expressed as

$$\mu_N = g_N \beta_N I, \tag{2}$$

where g_N is a dimensionless constant, the nuclear g factor, and β_N the nuclear magneton,

$$\beta_N = eh/4\pi Mc, \tag{3}$$

where e and M are respectively the charge and mass of the proton and c is the velocity of light. g_N and I are the quantities that distinguish nuclei from one another.

1.1 Nuclei in a Stationary Magnetic Field

When a nucleus with nonzero spin is placed in a strong magnetic field B, the orientation of the spin becomes quantized, each possible orientation having a different energy. A nucleus of spin I has $2I + 1$ possible orientations, with associated spin quantum numbers m_I of $-I, -I + 1, \ldots, I - 1, I$. The energy of interaction is proportional to the nuclear moment and the applied field. For a nucleus this can be written as

$$E_I = -g_N \beta_N m_I B = -\gamma_N h m_I B/2\pi. \tag{4}$$

Energy-level diagrams for nuclei with spins of $\frac{1}{2}$ and $\frac{3}{2}$ are given in Fig. 1. The selection rule governing transitions for a single nucleus is $\Delta m_I = \pm 1$. Thus

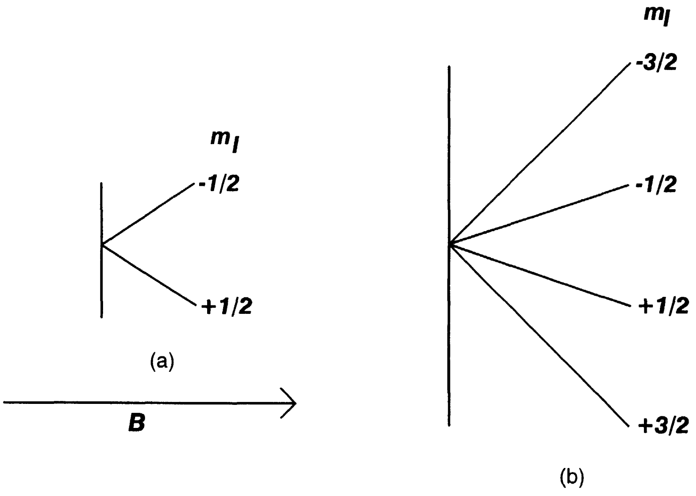

FIG. 1. Energy levels for (a) a nucleus with $I = \frac{1}{2}$, (b) a nucleus with $I = \frac{3}{2}$.

$$\Delta E_I = g_N \beta_N B = \gamma_N hB/2\pi. \tag{5}$$

1.2 Basic Principles of the NMR Experiment

To observe transitions, radiation given by $\Delta E = h\nu$ must be applied. The fundamental resonance condition for the NMR experiment is

$$h\nu = g_N \beta_N B = \gamma_N hB/2\pi, \tag{6}$$

where ν is the resonance frequency. For a proton in a magnetic field of 2.349 T, ν is 100 MHz.

Resonance absorption can only be detected if the spin levels involved differ in population. Consider a spin-$\frac{1}{2}$ nucleus, where the lower, more stable energy level is denoted by α and the higher one by β. In thermal equilibrium the Boltzmann distribution of energies is maintained and

$$N_\beta/N_\alpha = \exp(-h\nu/kT). \tag{7}$$

For NMR the excess of spins in the lower state is only ca. 1 in 10^5. This is one reason for the low sensitivity of NMR compared with IR and UV spectroscopy and even electron spin resonance.

The magnetic-resonance experiment itself is best explained and visualized in classical as opposed to quantum-mechanical terms. How is the energy transferred to nuclei that are in a magnetic field, and how is the energy absorbed measured? The following description is for nuclei of spin $\frac{1}{2}$. In a magnetic field B_0, applied in the z direction, the nuclei will precess in random phase around the z axis with a frequency ω_0 known as the Larmor frequency, where $\omega_0 = 2\pi\nu$. For a spin-$\frac{1}{2}$ nucleus there are two precessional cones, as shown in Fig. 2, one with and the other against B_0. There is a slight excess of nuclei processing with B_0, which gives rise to a small net magnetization vector M_0 along

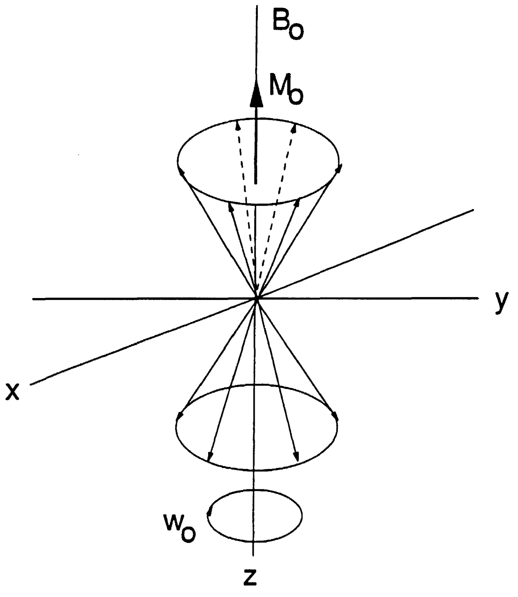

FIG. 2. Motion of spin-$\frac{1}{2}$ nuclei in a magnetic field.

the *z* axis but not in the *x*-*y* plane. The quantum description of the lower- and higher-energy levels can be equated with the classical description of nuclei processing with and against the applied field.

In the magnetic-resonance experiment, the net magnetization M_0 is tipped toward the *x*-*y* plane. Radio-frequency (rf) electromagnetic energy is applied such that its magnetic component, B_1, is at right angles to B_0 and is rotating with the processing nuclei. The rf irradiation is applied along the *x* axis, and at the resonance frequency it tilts the net magnetization toward the *x*-*y* plane. Conventionally, the receiver is located on the *y* axis of the Cartesian coordinates.

To obtain an NMR spectrum, which may contain more than one peak at slightly different resonance positions, either the rf frequency or the magnetic field B_0 can be scanned over a narrow range to bring the various peaks into resonance. Historically, this mode of excitation, known as continuous wave (cw), was used by NMR instruments. It has now been almost completely superseded by the more sensitive pulsed FT method. For more detailed descriptions of the NMR experiment and instrumentation, the reader is referred to the articles NUCLEAR MAGNETIC RESONANCE AND QUADRUPOLE RESONANCE and MAGNETIC RESONANCE SPECTROMETERS, respectively, and books by Harris (1986), Shaw (1984), and Farrar (1987). In particular, the reader should consult these sources for information on the FT experiment since this knowledge is assumed in the description of multiple-pulse experiments later in this article. In such experiments, "pulse angles" are specified. A pulse whose duration is exactly adjusted to turn the double cone of precession through 90° is called a 90° pulse. When the cone is turned about the *x* axis, it is a 90°_x pulse.

In the absence of any mechanism enabling the nuclear spins to return to the ground state, the populations of the lower- and higher-energy levels will become equal, and no more energy will be absorbed. However, two such mechanisms do exist. One is the spin–lattice or longitudinal relaxation process, characterized by a rate constant T_1, which occurs by the loss of energy from excited nuclear spins to the surrounding molecules. The other is the spin–spin or transverse relaxation process with a rate constant T_2, which involves a transfer of energy from one nucleus to another. An important feature of spin–spin relaxation is that it determines the natural linewidth of the signals in a spectrum.

2. HIGH-RESOLUTION SOLUTION NMR SPECTROSCOPY

In general, solids give broad NMR signals devoid of the fine structure usually of interest to the chemist. The interactions (see Sec. 3.1) that cause the broadness are generally averaged to zero in the gas phase or in solution as a result of the random motion of the molecules.

The magnetogyric ratio (γ) of a nucleus enables the resonance frequency to be calculated from Eq. (6) for any magnetic-field strength. Table 2 lists useful, commonly studied nuclei and their magnetic properties. The frequencies are for a 2.35-T applied magnetic field, i.e., relative to ^1H at 100 MHz. Table 2 gives the relative sensitivity for equal numbers of each nucleus and their relative receptivities, i.e., the product of the relative sensitivity and natural abundance.

The abundant nuclei, ^1H, ^{19}F, and ^{31}P, with $I = \frac{1}{2}$ were the first to be studied by cw NMR. On account of sensitivity problems, less abundant isotopes of important spin-$\frac{1}{2}$ nuclei such as ^{13}C and ^{15}N require FT techniques for their observation in natural abundance. Nuclei with spin $I \geq 1$ have an electric quadrupole moment in addition to their magnetic dipole moment. The interaction of the nuclear quadrupole moment with the electric field gradient at the nucleus provides a very efficient process for nuclear relaxation and can result in very broad signals.

The sample, normally a solution in a deuterated solvent in a 5- or 10-mm glass tube, is placed in the instrument probe between the poles of an electromagnet or permanent magnet or inside the solenoid of a superconducting magnet. The probe contains the rf transmitter and receiver coils and a spinner to spin the tube about its vertical axis. Sample spinning is used to average out magnetic-field inhomogeneity across the sample. High-resolution NMR requires both the magnetic field and the rf source to be homogeneous and stable to better than 1 part in 10^8.

The ideal NMR solvent should dissolve

Table 2. Magnetic properties of some nuclei.

Isotope	Spin	Natural abundance %	NMR frequency at B = 2.35 T (MHz)	Relative Sensitivity	Relative Receptivity
^{1}H	$\frac{1}{2}$	99.98	100.000	1.00	1.00
^{2}H	1	0.015	15.351	9.65×10^{-3}	1.45×10^{-6}
^{11}B	$\frac{3}{2}$	80.42	32.084	0.17	0.13
^{13}C	$\frac{1}{2}$	1.108	25.144	1.59×10^{-2}	1.76×10^{-4}
^{14}N	1	99.63	7.224	1.01×10^{-3}	1.01×10^{-3}
^{15}N	$\frac{1}{2}$	0.37	10.133	1.04×10^{-3}	3.85×10^{-6}
^{17}O	$\frac{5}{2}$	0.037	13.557	2.91×10^{-2}	1.08×10^{-5}
^{19}F	$\frac{1}{2}$	100	94.077	0.83	0.83
^{23}Na	$\frac{3}{2}$	100	26.451	9.25×10^{-2}	9.25×10^{-2}
^{29}Si	$\frac{1}{2}$	4.70	19.865	7.84×10^{-3}	3.69×10^{-4}
^{31}P	$\frac{1}{2}$	100	40.481	6.63×10^{-2}	6.63×10^{-2}

the sample, be inert, nonpolar, liquid over the temperature range of interest, and inexpensive, and not give rise to large interfering peaks in the spectrum being recorded. Deuterated solvents such as deuterochloroform ($CDCl_3$) and deuterodimethylsulfoxide (D_6DMSO) are most commonly used. They offer the advantage of being "transparent" in proton spectra and provide a deuterium signal that can be used to stabilize and lock the spectrometer system. A clear mobile solution is required for a high-quality spectrum. Solid particles or viscous solutions will degrade the resolution. Traces of paramagnetic impurities in a solution can dramatically reduce relaxation times and cause peak broadening.

The amount of sample required depends on the method of detection (cw or FT) and the receptivity of the nucleus being studied. For most routine cw proton spectrometers, 10–20 mg of compound of molecular mass 200–300 Da [1 Da (dalton) = 1/12 the mass of a ^{12}C atom], dissolved in 0.4 mL of solvent, should give a reasonable proton spectrum. On a 500-MHz FT instrument it should be possible to obtain a proton spectrum from a few micrograms of such a compound. The low receptivity of nuclei such as ^{13}C means that it is important to use as concentrated a solution as possible in order to obtain a spectrum in as short a time as possible. Larger-diameter sample tubes (10 and 15 mm) and consequently larger sample volumes can be used to increase the amount of sample used.

2.1 Spectral Parameters

All applications of solution NMR spectroscopy to structure determination make use of one or more of the spectral parameters chemical shift, spin–spin coupling, and signal intensity. In the following sections, examples from ^{1}H and ^{13}C NMR are used, but the concepts described apply to magnetic nuclei in general. Books by Harris (1986), Abraham *et al.* (1988), and Emsley *et al.* (1965) will provide further information.

2.1.1 Chemical Shift Equation (6) might suggest that only a single peak would be obtainable from each magnetic isotope. However, when the molecule under observation is placed in the magnetic field, the electrons in the molecule shield the nuclei from the applied magnetic field to a small extent. For a particular nucleus within a molecule, the degree of shielding depends on the density of the electrons circulating about that nucleus. Protons in different chemical environments therefore have different shieldings. Figure 3 shows diagrammatically what happens to spherically symmetrical *s* electrons at a nucleus under the influence of an external magnetic field B_0. They circulate around the nucleus, producing their own magnetic field that opposes the applied field. The effective field B at the nucleus is then given by

$$B = B_0(1 - \sigma). \qquad (8)$$

Thus to obtain the resonant condition, Eq. (6), it is necessary to increase the applied

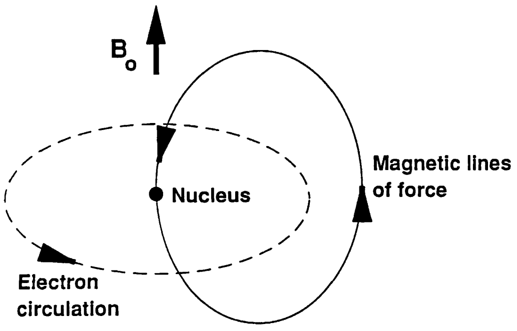

FIG. 3. Electronic shielding of a nucleus by circulating electrons.

field above that for an isolated nucleus. The shielding constant σ is a small fraction, usually of the order of parts per million. For electrons in *p* orbitals where there is no spherical symmetry, the situation is more complex since shielding and deshielding effects depend on the orientation of the molecule with respect to the applied field. Such effects are considered in more detail in Sec. 2.2.1.

The magnitude of the nuclear shielding is proportional to the applied magnetic field. The difference in the peak position of a particular nucleus from the peak position of a reference nucleus is called the chemical shift. In ^{1}H NMR, if two peaks are separated by 100 Hz in a magnetic field corresponding to 100 MHz, they will be separated by 400 Hz in a field corresponding to 400 MHz. Chemical shifts can be expressed in dimensionless units δ as parts per million of the applied frequency as follows:

$$\delta = \frac{(\nu_{\text{sample}} - \nu_{\text{reference}}) \times 10^6}{\text{applied frequency}} \,. \qquad (9)$$

Thus for the example given above, at 100 MHz,

$$\delta = \frac{100 \times 10^6}{100 \times 10^6} = 1.0;$$

at 400 MHz,

$$\delta = \frac{400 \times 10^6}{400 \times 10^6} = 1.0.$$

The chemical shift δ is therefore a molecular parameter that does not depend on the spectrometer frequency. ^{1}H and ^{13}C nuclei in different environments in organic molecules give peaks that cover ranges of approximately 20δ_H and 220δ_C, respectively.

Tetramethylsilane (TMS) is the most commonly used chemical-shift reference compound for ^{1}H and ^{13}C NMR. In both types of spectra, TMS gives a single sharp resonance at lower frequency than most other proton or carbon resonances. The TMS is usually added to the sample solution as an internal reference. The accepted sign convention is for chemical shifts to high and low frequency of the reference peak to be positive and negative, respectively. Figure 4 shows the 250-MHz ^{1}H spectrum of methylacetate, CH_3COOCH_3, referenced to TMS, the peaks at 2.0δ and 3.6δ arising from the protons of the acetyl and methoxy groups respectively, respectively.

2.1.2 Spin–Spin Coupling

Chemical shift is one source of information on fine structure in NMR spectra. The second is spin–spin coupling. Figure 5 shows the resonances from the aromatic protons in the 4 and 5 positions of 3-methoxy-6-nitro-o-xylene,

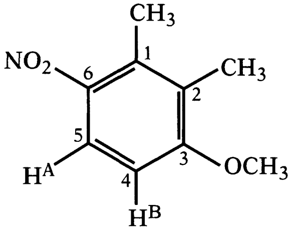

One might expect two peaks from the two different aromatic protons in the molecule. In fact four peaks are observed. This results from the indirect coupling of proton spins through the intervening bonding electrons. Thus the magnetic field experienced by nucleus A has a small component arising from nucleus B. Nucleus B has two equally probable orientations in the magnetic field, $m_B = +\frac{1}{2}$ or α and $m_B = -\frac{1}{2}$ or β. As a result nucleus A gives two equally intense peaks corresponding to the two orientations of B. By the same reasoning B itself also gives two equally intense peaks with the same separation as those for A. This splitting is called the spin–spin coupling *J*, where J_{AB} is the

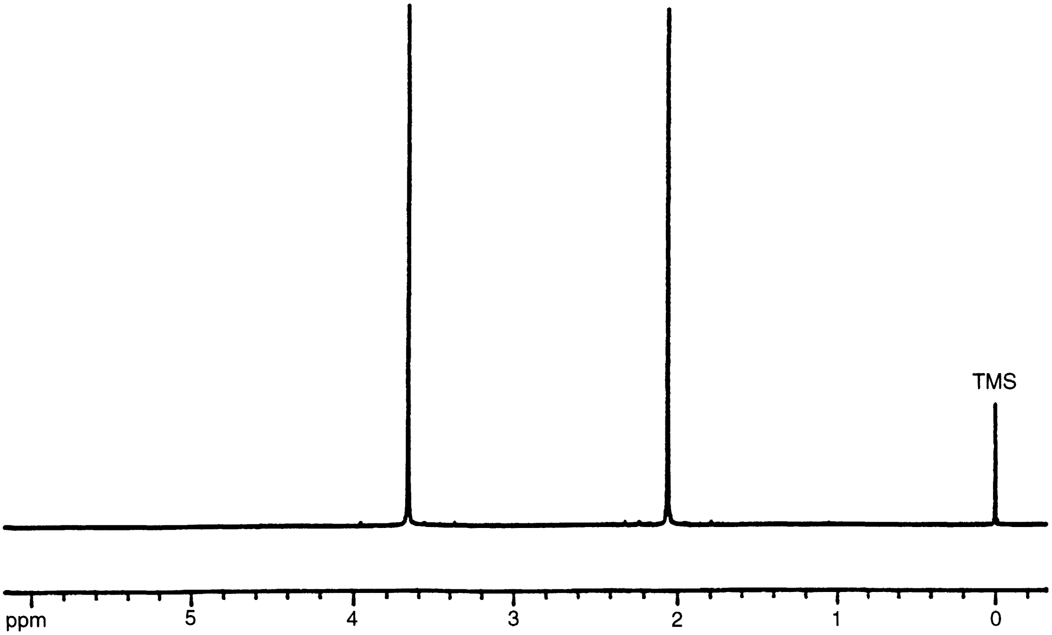

FIG. 4. 250-MHz ^{1}H NMR spectrum of methylacetate.

coupling between nuclei A and B. The magnitude of the coupling, which is measured in hertz, is independent of the applied magnetic field. Couplings between protons may be observed from nuclei separated by up to four bonds. Their magnitude rarely exceeds 20 Hz.

If more than two nuclei are interacting with each other, all the possible orientations of the coupling nuclei must be considered. Figure 6 shows the splitting patterns from two different combinations of three interacting nuclei. When all three nuclei have different chemical shifts, as for the olefinic protons of vinyl acetate, $CH_2=CHOCOCH_3$, then each nuclear signal is split by coupling to both of the other nuclei. Thus each resonance can be envisaged as being split into a

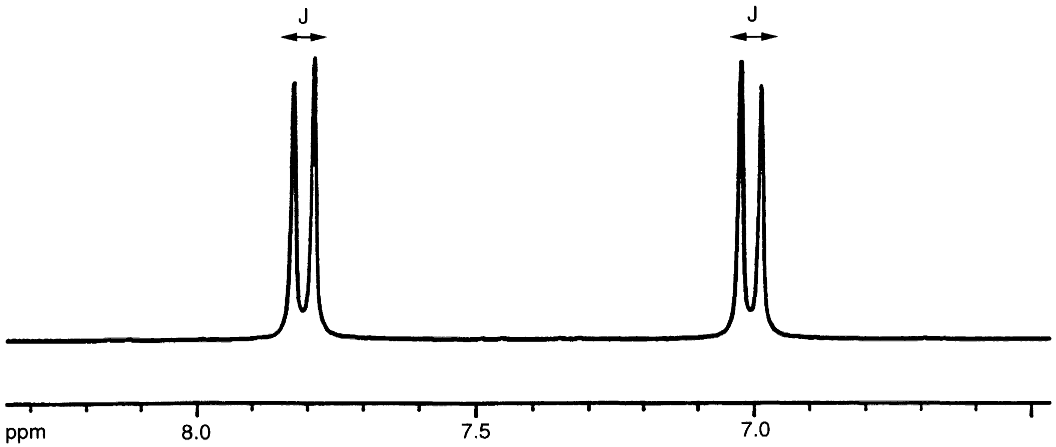

FIG. 5. Aromatic region of the 250-MHz ^{1}H NMR spectrum of 3-methoxy-6-nitro-orthoxylene.

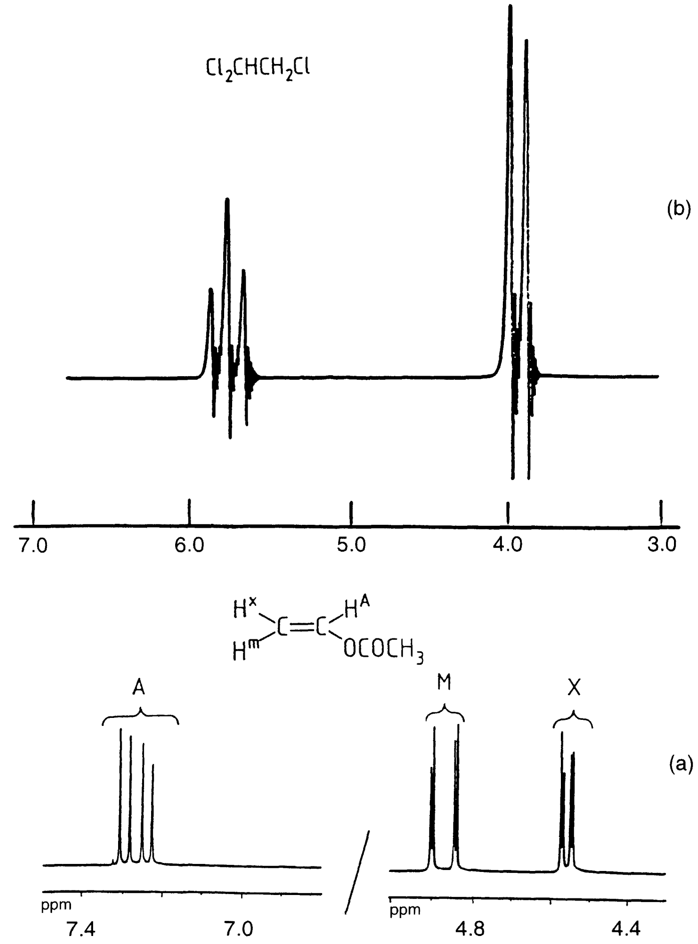

FIG. 6. (a) Olefinic region of the 250-MHz ^1H NMR spectrum of vinyl acetate; (b) 60-MHz ^1H NMR spectrum of 1,1,2-trichloroethane.

doublet by the first coupling, each line of which is further split into a doublet, to give a quartet pattern for each nucleus. When two of the three nuclei are equivalent as in 1,1,2-trichloroethane, Cl_2CHCH_2Cl, the total number of lines in the spectrum is reduced because some of the couplings are identical. The methylene resonance is split into a doublet with a coupling J to the methine resonance. However, the methine resonance is a 1:2:1 triplet pattern with the same splitting J. This arises because the two equivalent methylene protons have three possible orientations ($\alpha\alpha$, $\alpha\beta$ and $\beta\alpha$, $\beta\beta$). Note that the two chemically equivalent methylene protons do not exhibit any splitting from coupling to each other.

Simple splitting patterns such as these, for nuclei of spin $\frac{1}{2}$, obey the following general rules:

1. If a nucleus interacts with n other nuclei with a different coupling to each, then the signal will consist of 2^n lines of equal intensity.
2. If a nucleus A couples to n equivalent nuclei B, then the pattern for A comprises $n + 1$ lines with relative intensities given by the coefficients of the binomial expansion of $(1 + x)^n$.

These simple rules are only strictly applicable to so-called "first-order spectra." For a spectrum to be first order, a single criterion must be obeyed: The chemical shift difference $\Delta\nu$ for the coupled nuclei must be much larger than the coupling between them ($\Delta\nu/J > 10$). If this condition is not met, the spectrum may show more lines than the rules predict, and their intensities are distorted, with a buildup of intensity toward the middle of the coupling pattern. It is beyond the scope of this article to describe second-order spectra. In the limit where two coupled nuclei have exactly the same chemical shift, either because they are chemically equivalent or because they just happen to have the same chemical shift, the coupling is not observed, and in the absence of coupling to other nuclei, a single peak is observed.

Because the magnitude of the spin–spin coupling interaction is independent of the applied magnetic field whereas the chemical shift separation is proportional to it, a complex NMR spectrum can be simplified by going to higher applied fields.

All the above examples of spin–spin coupling involve homonuclear coupling in proton spectra. In natural-abundance ^{13}C spectra homonuclear coupling is not normally observed since the probability of two adjacent ^{13}C atoms is extremely low, viz., 10^{-4}. For organic molecules the most commonly encountered couplings are those to protons. One-bond couplings range from about 110 to 250 Hz. Most longer-range couplings are <10 Hz. Proton-coupled ^{13}C spectra often show complex overlapping multiplets. In the interests of simplifying the spectra and obtaining a spectrum in a shorter time, routine ^{13}C spectra are usually proton-noise decoupled (see Sec. 2.3.2). Thus in the absence of other coupling nuclei such as ^{31}P or ^{19}F, a single sharp peak is observed for each chemically inequivalent ^{13}C atom. Figure 7 shows both the proton coupled and decoupled ^{13}C spectra of ethyl benzene,

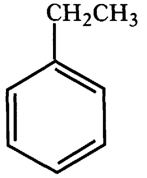

.

2.1.3 Signal Intensity NMR is an inherently quantitative technique. For any nucleus the NMR signal obtained, at least for a single-scan cw experiment or a single-pulse FT experiment, is directly proportional to the number of nuclei producing it. For FT experiments in general to be quantitative, the spin system must be given sufficient time to relax completely between successive pulses. Most proton spectra are recorded under essentially

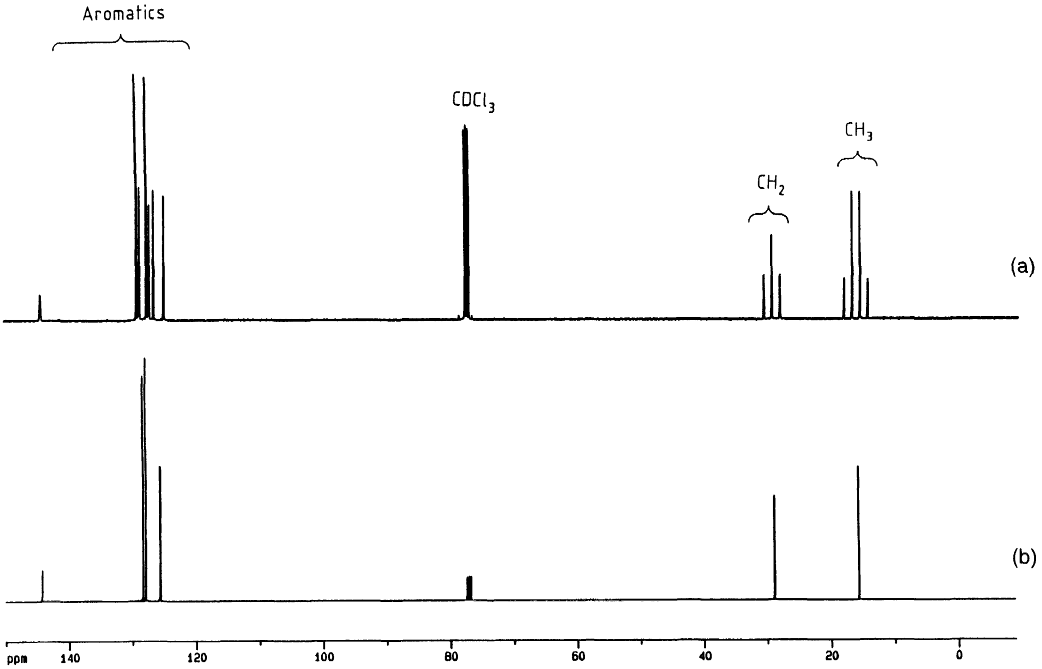

FIG. 7. (a) Proton-coupled and (b) proton-broadband decoupled 100-MHz ^{13}C NMR spectrum of ethylbenzene in deuterochloroform ($CDCl_3$).

quantitative conditions and are integrated to give the relative number of protons contributing to each resonance. Figure 8 shows the integrated proton spectrum of ethylbenzene with the integral traces and relative peak areas of 5:2:3 displayed for the aromatic, methylene, and methyl protons, respectively.

^{13}C spectra and those of other low-natural-abundance nuclei are not normally recorded under quantitative conditions. For structural identification work with such nuclei, conditions are usually selected for optimum sensitivity, not quantification. There is an additional problem in spectra such as ^{13}C where broadband proton decoupling is usually used. Apart from collapsing the multiplet structure, the decoupling irradiation causes an increase in signal intensity due to a phenomenon known as nuclear Overhauser enhancement (NOE—see Sec. 2.3.2). The enhancement factor is often different for the various ^{13}C resonances in the spectrum. Quantitative ^{13}C work is possible, but the experimental conditions have to be carefully selected.

2.2 NMR and Structure

In this section the relationship between chemical structure and the NMR parameters of chemical shift and spin–spin coupling are discussed for some of the more commonly studied nuclei. Information about other nuclei can be found in books by Harris and Mann (1978), Akitt (1983), and Mason (1987).

2.2.1 Hydrogen (^1H and ^2H) The two naturally occurring isotopes of hydrogen are the proton and the deuteron, with natural abundances of 99.985% and 0.015%, respectively. While the former with $I = \frac{1}{2}$ is the most commonly studied of all magnetic nuclei, the latter with $I = 1$ is much less frequently studied in natural abundance and then only for a few specialist applications. One such application is the determination of deuterium levels in samples. If small isotope effects are disregarded, corresponding ^1H and ^2H chemical shifts are the same. Resonances in ^2H spectra are less well dispersed than the corresponding ^1H spectra by a fac-

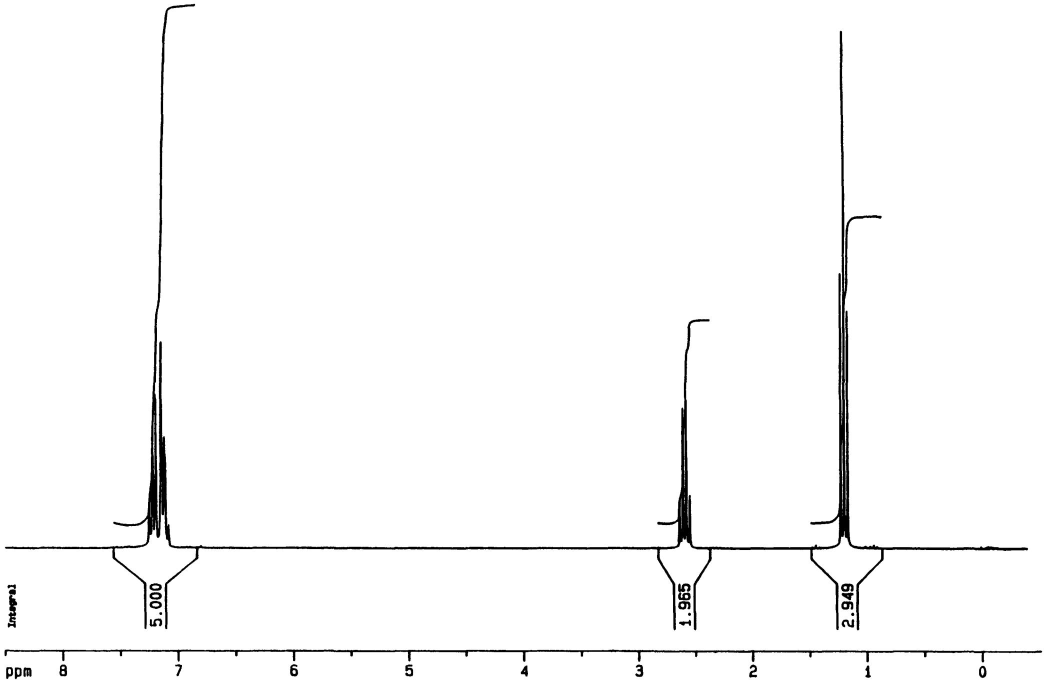

FIG. 8. Integrated 250-MHz ^1H NMR spectrum of ethylbenzene.

tor of 6.51, the ratio of the magnetogyric ratios. The smaller magnetogyric ratio also results in correspondingly smaller coupling constants. While the following discussion is exemplified by reference to the proton, most of the content is equally applicable to the deuteron. However, since deuterium homonuclear coupling constants are very small, ^{2}H spectra with broadband ^{1}H decoupling (see Sec. 2.3.2) give a single peak from each chemically inequivalent deuteron even for deuterium-enriched samples.

2.2.1.1 Chemical Shifts. The shielding effect of spherically symmetrical *s* electrons was discussed in Sec. 2.1.1. This diamagnetic upfield shift affects all nuclei since every molecule has *s* electrons. For electrons in *p* orbitals there is no spherical symmetry, and the phenomenon of diamagnetic anisotropy is used to explain some otherwise anomalous chemical shifts. Diamagnetic anisotropy results in shielding and deshielding, which depend on the orientation of the molecule with respect to the applied magnetic field.

The hydrogen atom is a special case since it is the only commonly studied nucleus with no *p* electrons. A consequence of this is that in proton NMR the direct influence of the diamagnetic term can be seen. For example, in substituted methanes CH_3X, as X becomes more electronegative, the electron density around the protons decreases, and they resonate at lower fields. For the same reason, acidic protons in carboxylic acid groups resonate at very low fields. Figure 9 shows the chemical-shift ranges for a selection of organic groups.

While the electronegativity concept can be used to explain chemical shifts for saturated aliphatic compounds, the effects of diamagnetic anisotropy must also be considered for some other classes of compounds. For example, aromatic compounds such as benzene are strongly deshielded, the protons of acetylene are much more strongly shielded than those of ethylene (acetylene 1.48δ compared with ethylene 5.31δ), and aldehydic protons resonate at low fields. In the case of benzene the deshielding results from the so-called "ring current effect." When a molecule of

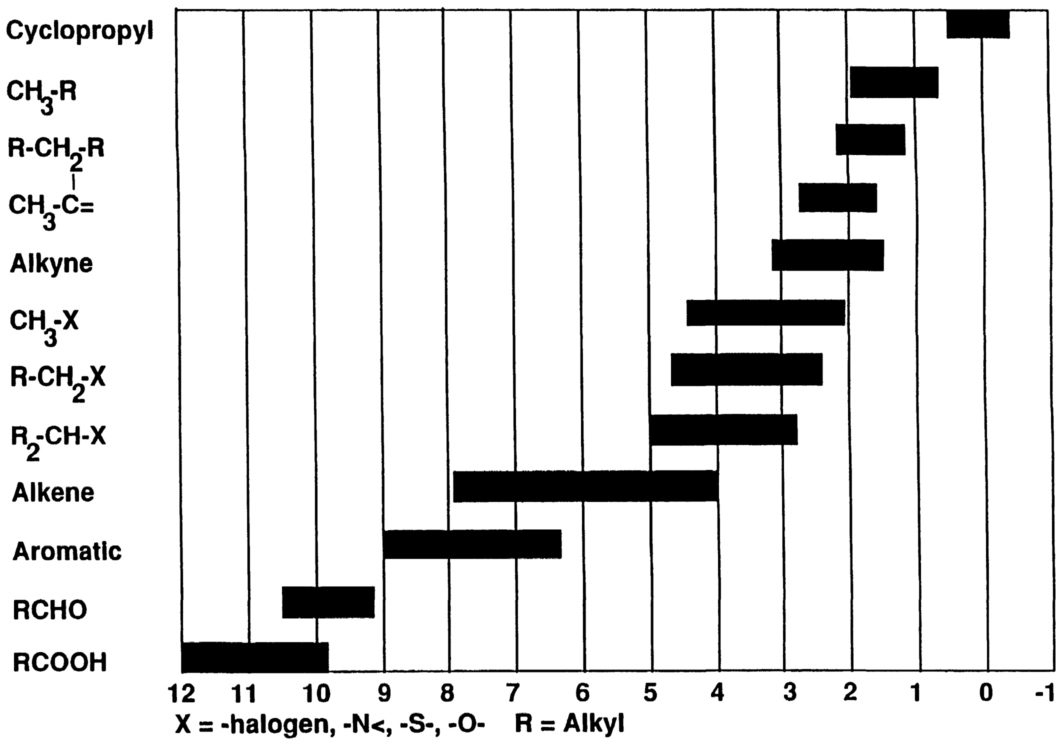

FIG. 9. ^{1}H NMR chemical-shift ranges.

benzene is oriented perpendicular to the applied magnetic field B_0, there is a molecular circulation of electrons about the field as shown in Fig. 10. The resulting ring current produces an additional field that opposes the applied magnetic field along the sixfold axis of the benzene ring but adds to it at the protons on the benzene ring, giving a low field shift. The ring current is only induced when the applied magnetic field is perpendicular to the benzene ring. In practice only a proportion of the rapidly tumbling molecules are perpendicular to the magnetic field, but the overall averaged shift is affected by them. Similar arguments can be invoked to explain the apparently anomalous shifts of acetylenic and aldehydic protons.

The effect of substituents on the proton chemical shifts of hydrocarbons has been extensively investigated. Schoolery's rules enable the prediction of the chemical shift of any CH_2XY and $CHXYZ$ proton. The chemical shift is the shift for methane plus the sum of substituent contributions that are largely determined by the electronegativity of the substituent. In aromatic compounds, strongly electron-withdrawing groups such as nitro and carboxyl deshield all the protons, but the effect is largest at the ortho and para positions. The opposite is true for strongly electron-donating groups such as NH_2 and OH, while the halogens show smaller effects.

2.2.1.2 Spin–Spin Coupling. The origin of spin–spin coupling and the resulting multiplet structure in proton NMR spectra was introduced in Sec. 2.1.2. The magnitude of couplings plays an important role in chemical structure determination, particularly in differentiating between isomeric structures. In proton NMR couplings are observed between protons and to other magnetic nuclei in the molecule such as ^{19}F and ^{31}P. A vicinal proton–proton coupling between protons separated by three bonds is denoted by $^3J_{HH}$, whereas $^1J_{PH}$ represents a one-bond coupling between a proton and a phosphorus atom.

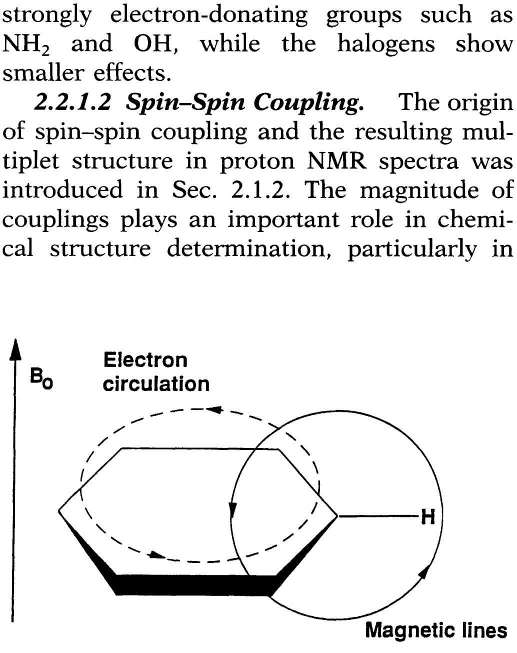

FIG. 10. Ring-current effects in benzene.

The most commonly encountered coupling in NMR is the vicinal proton–proton coupling. For protons on adjacent carbon atoms in saturated aliphatic groups, this depends primarily on the dihedral angle ϕ between the CH bonds. The angular dependence may be written approximately as

$$^3J(\mathrm{CH\cdot CH}) = 10\cos^2\phi\ \mathrm{Hz} \qquad (10)$$

and is best exemplified by couplings in six-membered rings of fixed geometry, such as the chair conformation of cyclohexane,

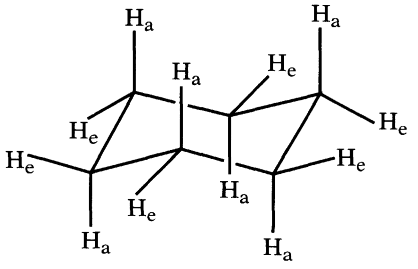

.

Thus an axial–axial coupling $^3J_{aa}$ in a cyclohexane ring where $\phi = 180°$ would be ca. 10 Hz. Experimentally, $^3J_{aa}$ varies from 8 to 13 Hz, the axial–equatorial coupling $^3J_{ae}$ from 2 to 4 Hz, and the equatorial–equatorial $^3J_{ee}$ from 1 to 3 Hz depending on the nature of the substituents present. In acyclic systems the same angular dependence exists with *gauche* and *trans* couplings typically in the ranges of 2–4 and 8–12 Hz, respectively. However, in systems where there is free rotation about the intervening C–C bond, an average coupling of ca. 7 Hz is observed.

The magnitude of geminal couplings $^2J_{HH}$ over two bonds depends to a first approximation on the H–C–H angle between the coupled protons. For tetrahedrally substituted carbon atoms, a coupling in the range of 12–15 Hz is typical. In aliphatic systems where there is free rotation about bonds, geminal couplings are only observed where methylene protons are diastereotopic, that is, when the molecule contains a chiral center:

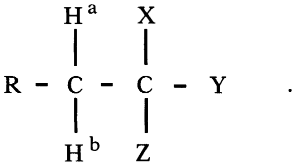

$$\begin{array}{ccccccc} & & H^a & & X & & \\ & & | & & | & & \\ R & - & C & - & C & - & Y \\ & & | & & | & & \\ & & H^b & & Z & & \end{array}$$

In such compounds, protons H^a and H^b cannot be interchanged by symmetry operations or by rapid rotation and therefore may have different chemical shifts and exhibit coupling to each other.

In olefinic compounds the $^3J_{HH}$ coupling constant is used to differentiate *cis* and *trans* isomers, e.g. of a disubstituted ethylene with substituents A and B,

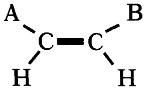

CIS

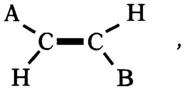

TRANS

since $J_{trans} > J_{cis}$. Ranges for *trans* and *cis* couplings are 12–18 and 6–12 Hz, respectively. Geminal couplings from terminal $=CH_2$ groups in olefins lie in the range 0–3 Hz. In aromatic compounds the observed couplings are also characteristic of the disposition of the hydrogen atoms. For benzene derivatives the following ranges are typical: $^3J_{\text{ortho}} = 6\text{–}10$ Hz, $^4J_{\text{meta}} = 1\text{–}3$ Hz, $^5J_{\text{para}} = 0\text{–}1$ Hz, where the ortho (o), meta (m), and para (p) protons relative to the proton of interest, H′, are shown for the benzene molecule,

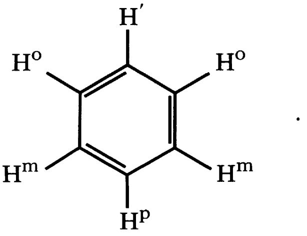

Couplings of protons to other nuclei are discussed in the sections concerned with those nuclei.

Further information about proton chemical shifts and coupling constants can be found in books such as Abraham *et al.* (1988) and Silverstein *et al.* (1991) as well as data collections that include "The Sadtler Collection of High Resolution Spectra," Pouchert (1983) and Brugel (1979).

2.2.2 Carbon (^{13}C) Although present in natural abundance at only 1.1%, the ^{13}C nucleus has, since the advent of FT methodology, been extensively studied. Only ^{1}H NMR is more widely used. In organic chemistry the information obtained from the two nuclei is often complementary for solving structural problems. As mentioned in Sec. 2.1.2, proton–carbon spin–spin couplings in ^{13}C spectra are usually removed by broadband proton decoupling.

2.2.2.1 Chemical Shifts. For organic chemicals most shifts are spread over a range of 0–200 ppm, about 20 times that for the proton. Figure 11 shows chemical-shift ranges for some common organic groups. By comparison with Fig. 9 it can be seen that there is an overall similarity in ordering on going downfield from TMS to that observed for proton chemical shifts. However, in ^{13}C NMR the effects of substitution are often significant several bonds away from the point of substitution. In simple alkanes, α and β carbons result in deshielding effects of approximately 9 ppm each while a γ carbon shields by about 2.5 ppm. As a result the C-1 shifts for methane (CH_4), ethane (CH_3CH_3), and propane ($CH_3CH_2CH_3$) are -2.3, 5.7, and 15.4 ppm, respectively. For alkanes the effects are additive, and a set of rules has been derived by Lindeman and Adams (1971). Similar analyses have been developed for alkenes.

Polar substituents have large effects as can be seen by comparing the shifts for *n*-pentane and 1-pentanol:

$$\begin{array}{ccccccccc} 13.7 & & 22.6 & & 34.5 & & 22.6 & & 13.7 \\ CH_3 & - & CH_2 & - & CH_2 & - & CH_2 & - & CH_3 \end{array}$$

$$\begin{array}{ccccccccccc} 13.9 & & 22.9 & & 28.5 & & 32.7 & & 62.0 & & \\ CH_3 & - & CH_2 & - & CH_2 & - & CH_2 & - & CH_2 & - & OH \end{array}$$

Substituent chemical shifts in benzene derivatives have been collated by Ewing (1979).

2.2.2.2 Spin–Spin Coupling. Carbon–proton couplings are usually sacrificed by broadband proton decoupling in the interests of spectral simplification and obtaining an acceptable spectrum more quickly. However, such couplings can provide valuable structural information.

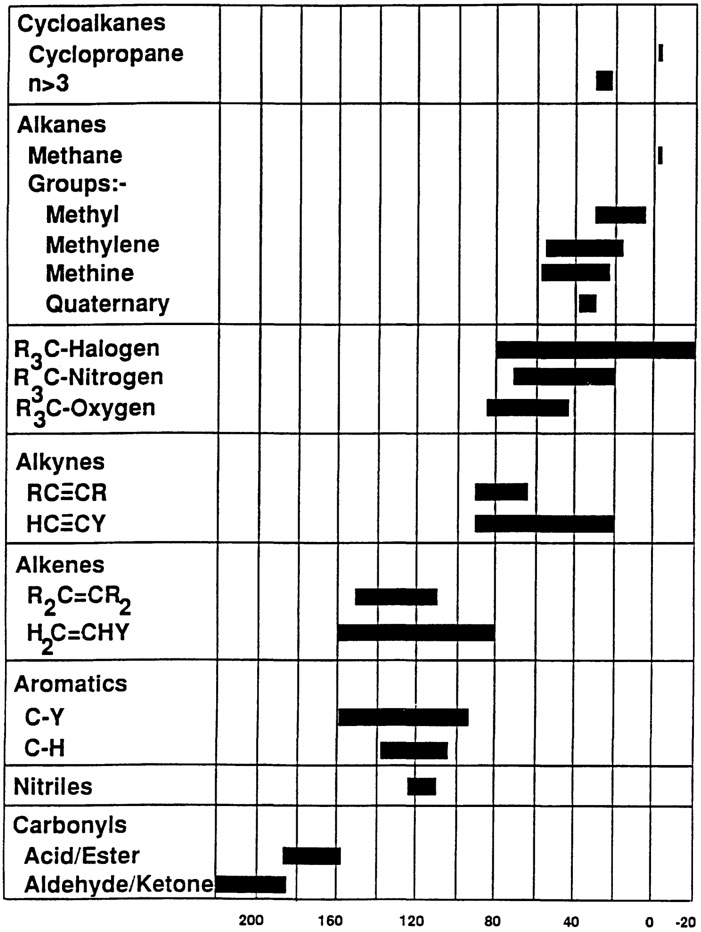

Where R = H or alkyl, Y = all substituents

FIG. 11. ^{13}C NMR chemical-shift ranges.

One-bond carbon–proton couplings $^1J_{CH}$ are of particular interest. At the simplest level the resulting multiplet provides information about the type of carbon atom; for example, a methyl group gives a quartet resulting from coupling to three equivalent hydrogen atoms. However, such information can be more conveniently obtained from editing experiments (see Sec. 2.4). $^1J_{CH}$ values vary from ca. 110 to 320 Hz, increasing with increasing *s* character of the C–H bond. Thus the couplings for methane, CH_4 (sp^3), ethylene, $CH_2{=}CH_2$ (sp^2), and acetylene, $CH{=}CH$ (sp) are 125, 156, and 248 Hz, respectively. Polar substituents also have significant effects on the magnitudes of couplings.

$^2J_{CH}$ values in aliphatic hydrocarbons range from 5 to 60 Hz. Couplings are often very characteristic of substitution and/or bond hybridization. Thus while $^2J_{CH}$ for ethane is 4.5 Hz, the coupling to the methyl carbon of acetaldehyde, $CH_3CH = O$, is 26.7 Hz. Acetylene has a very large $^2J_{CH}$ of 49.3 Hz.

$^3J_{CH}$ couplings show orientational effects analogous to those observed in the corresponding HH couplings. Thus $^3J_{cis}$ is always less than $^3J_{trans}$ in olefins, and the same type of dihedral-angle dependence exists for saturated aliphatic groups. In aromatic rings $^3J_{CH}$ values are characteristically larger than $^2J_{CH}$ values. For benzene itself, $^3J_{CH} = 7.4$ Hz and $^2J_{CH} = 1.0$ Hz.

Note that heteronuclear couplings to magnetic nuclei other than ^{1}H (e.g., ^{19}F, ^{31}P, and ^{2}H) are observed in proton-decoupled spectra. Thus couplings to deuterium will be seen on signals from deuterated solvents.

Further information about ^{13}C chemical shifts and coupling constants can be found in several books, including Abraham *et al.* (1988), Breitmaier and Voelter (1987), Stothers (1972), and Wehrli *et al.* (1988). Large, computer-searchable ^{13}C libraries available from Chemical Concepts (SPECINFO) and BIO-RAD/Sadtler (CSEARCH) probably contain the largest available collections of ^{13}C reference data. The SPECINFO database contains ^{13}C data from over 80 000 compounds.

2.2.3 Fluorine (^{19}F) With 100% natural abundance and $I = \frac{1}{2}$, ^{19}F is ideally suited to NMR investigations. Since the earliest days of NMR spectroscopy, ^{19}F applications in organic, inorganic, and organometallic chemistry have been widespread. As in the case of the ^{13}C nucleus, broadband proton decoupling is often used to simplify spectra.

2.2.3.1 Chemical Shifts. ^{19}F NMR spectroscopy has the advantage of a very large chemical-shift range. Organofluorine compounds alone cover over 400 ppm, while the range is extended to some 700 ppm if inorganic compounds are included. $CFCl_3$ is now the generally adopted internal reference. Most organofluorine resonances are to the high-field side of the $CFCl_3$ peak. In NMR spectroscopy the accepted convention is for shifts to high field of the reference to be assigned negative values. While this sign convention is now generally observed, there are many publications where authors have used the opposite convention without explanation. Moreover, the confusion was frequently compounded by the use of alternative reference compounds. Table 3 gives some representative ^{19}F chemical shifts. The shifts are sensitive not only to directly attached substituents but also to long-range effects and thus are very good probes of changes in molecular structure. The chemical shifts in ppm given for the following compound provide an idea of the wide chemical-shift range observed for some fluorocarbons:

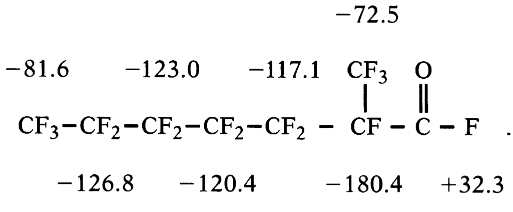

$$CF_3-CF_2-CF_2-CF_2-CF_2-\underset{}{\overset{CF_3}{\overset{|}{CF}}}-\overset{O}{\overset{\|}{C}}-F\ .$$

Table 3. Representative ^{19}F chemical shifts[a].

Compound	δ/ppm	Compound	δ/ppm
CH_3F	−271.9	CF_2Cl	−6.9
C_2H_5F	−211.5	CF_3CO_2H	−78.5
CF_3CF_3	−88.2	$C_6H_5 \cdot CF_3$	−63.9
$FCH=CH_2$	−114.0	AsF_3	−43.5
$F_2C=CH_2$	−81.3	ASF_5	−67.2
$F_2C=CF_2$	−135.2	SF_6	48.5
C_6F_6	−162.9	PF_3	−36.2
C_6H_5F	−113.2	PF_5	−77.8
$C_6H_5SO_2F$	65.5		

[a] Referenced to $CFCl_3$.

2.2.3.2 Spin–Spin Coupling. Large compilations of data are available for both fluorine–fluorine and fluorine–hydrogen coupling constants. The presence of several relatively large couplings can lead to complicated spectra even from relatively simple molecules if they are highly fluorinated. However, the spectra are often amenable to first-order analysis, and in general the coupling constants are very sensitive to structure and substituents.

Three-bond fluorine–hydrogen couplings show orientational effects similar to those observed in the corresponding HH couplings, e.g., the dihedral-angle dependence in saturated aliphatics and that $^3J_{cis}$ is less than $^3J_{trans}$ in olefins.

Wray (1980, 1983) has reviewed published ^{19}F data.

2.2.4 Phosphorus (^{31}P) The ^{31}P nucleus has 100% natural abundance and $I = \frac{1}{2}$. However, a relatively small magnetogyric ratio results in the sensitivity of the ^{31}P nucleus being only 0.066 that of the proton. Nevertheless, with Fourier-transform methodology, the nucleus is easy to detect even at low levels. There is much data in the literature covering applications involving inorganic compounds and complexes, organophosphorus chemistry, and biological studies.

2.2.4.1 Chemical Shifts. The accepted standard for ^{31}P NMR work is 85% H_3PO_4, which is normally used as an external reference. Phosphorus compounds are often reactive, and no suitable internal reference standard has been found for general use. Relative to H_3PO_4, ^{31}P shifts cover the range from ca. 250 ppm to −470 ppm. As in the case of ^{19}F NMR, there is a lack of uniformity in the sign convention adopted in earlier literature. Phosphorus(III) compounds cover the complete range, showing that the chemical shifts are very substituent dependent, while phosphorus(V) compounds cover a smaller range. To a first approximation the substituents directly attached to the phosphorus atom determine the chemical shift. Three variables are important: the bond angle, the electronegativity of the substituent, and the π-bonding character of the substituent. Some representative ^{31}P shifts are given in Table 4.

2.2.4.2 Spin–Spin Coupling. Coupling constants have been studied between ^{31}P and a wide variety of other magnetic nuclei. Couplings to protons, fluorine, and other phosphorus nuclei have been extensively used for structure elucidation. One-bond phosphorus–hydrogen couplings depend on the oxidation state of the phosphorus atom. Thus for P(III) compounds $^{1}J(^{31}P,^{1}H)$ values are of the order of 200 Hz, while for P(V) compounds they are usually in the range 400–1100 Hz. By contrast, one-bond phosphorus–fluorine coupling constants for both P(III) and P(V) compounds have a common range of ca. 800–1500 Hz. Phosphines are widely used as ligands in inorganic complexes, and therefore phosphorus couplings to some transition metals are important aids to structure determination.

2.2.5 Nitrogen (^{14}N and ^{15}N) There are two magnetic isotopes of nitrogen, ^{14}N and ^{15}N, with natural abundances of 99.63% and 0.37%, respectively. In a magnetic field of 9.4 T, the ^{14}N resonance frequency is 28.9 MHz, while that for ^{15}N is 40.5 MHz. There are problems associated with observing both nuclei. The ^{14}N nucleus possesses a quadrupole moment since $I = 1$. As a result, although ^{14}N signals can be detected easily, the lines can be very broad with widths often between 100 and 1000 Hz. With a 9.4-T applied magnetic field, this is equivalent to 3.5–34.6 ppm. This can cause resolution problems when a molecule contains more than one type of nitrogen atom.

Although $I = \frac{1}{2}$ for the ^{15}N nucleus, its very low natural abundance results in low receptivity. It also has a negative magnetogyric ratio, which results in a negative NOE when broadband proton decoupling is used. This reduces the intensity of peaks and can result in negative and even null resonances. It has therefore been necessary to use experimental procedures that minimize or eliminate the NOE. The two principal ones are inverse-gated proton decoupling (see Sec. 2.3.2.3) and the use of a paramagnetic relaxation reagent such as Cr(acetylacetonate)$_3$, which reduces T_1 and consequently the NOE. With modern FT equipment it is now possible to obtain ^{15}N spectra from most small and medium-sized nitrogen-containing molecules at natural abundance level. However, for work with large biological macromolecules, ^{15}N enrichment is commonly used.

2.2.5.1 Chemical Shifts. In principle, the use of two isotopes raises the possibility of a primary isotope shift, but in practice ^{14}N and ^{15}N shieldings differ by no more than a few tenths of a ppm and are interchangeable for practical purposes. For many years there was a reluctance to adopt a single reference standard for calibration purposes, and both nitromethane and ammonia were frequently used, in the former case with both sign conventions. Nitromethane is now generally used, and here the convention of negative values to high field of the reference is used. However, the chemical shift of nitromethane is somewhat solvent and concentration dependent. The total nitrogen chemical-shift range covers about 1000 ppm, and ranges for a selection of chemical classes are shown in Fig. 12.

2.2.5.2 Spin–Spin Coupling. It is unusual to observe coupling in ^{14}N spectra,

Table 4. Representative ^{31}P shifts[a].

P(III) compounds		P(V) compounds	
Compound	δ/ppm	Compound	δ/ppm
PF_3	97	PF_5	−35
PCl_3	219	PCl_5	−80
PBr_3	227	PBr_5	−101
$P(OMe)_3$	141	$P(OEt)_3$	−71
$P(OPh)_3$	127	$P(OPh)_5$	−86
PH_3	−240	$(MeO)_3P=O$	2
$MePH_2$	−164	$(PhO)_3P=O$	−18
Me_2PH	−99	$Me_3P=S$	59
Me_3P	−62	$Me_3P=O$	36
Ph_3P	−6	$Ph_3P=O$	25

---

[a] Referenced to 85% H_3PO_4.

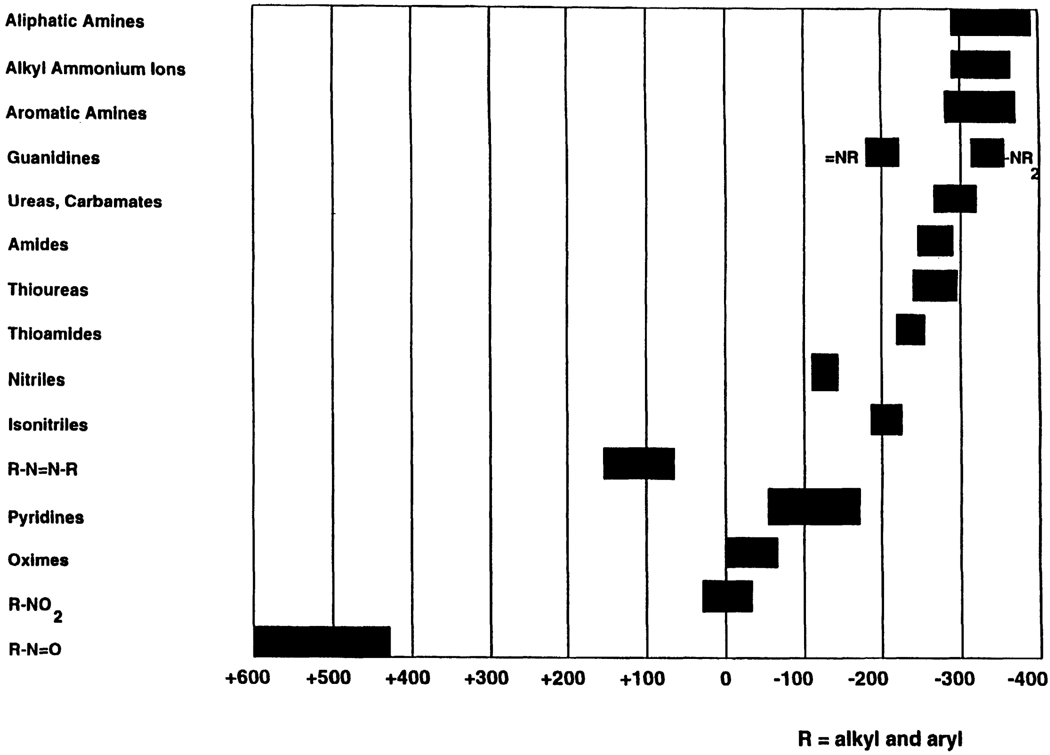

FIG. 12. $^{14}N/^{15}N$ NMR chemical-shift ranges.

where resonances are very broad as a result of quadrupolar relaxation. It is, however, often resolved for ammonium salts. The relative magnitudes of the corresponding ^{14}N and ^{15}N coupling constants are given by the ratio of their magnetogyric ratios. The discussion here will be restricted to ^{15}N coupling. The observation of coupling to low-abundance nuclei such as ^{13}C or another ^{15}N nucleus normally requires isotopic enrichment. Couplings to abundant nuclei such as ^{1}H, ^{19}F, and ^{31}P are more readily observed as in the case of ^{13}C spectroscopy.

One-bond couplings to hydrogen depend on hybridization and electronegativity. Thus the $^{1}J(^{15}N,^{1}H)$ values for $[NH_4]^+$, $O=C(NH_2)_2$, and $HC\equiv NH^+$ are 73.3, 89.0, and 134.0 Hz, respectively. One-bond couplings to ^{13}C also show a reasonable correlation with hybridization. The magnitudes of one- and two-bond couplings to ^{13}C are often comparable.

Witanowski *et al.* (1986) have reviewed ^{15}N data.

2.2.6 Oxygen (^{17}O) The only magnetic isotope of oxygen is ^{17}O. With a natural abundance of only 0.037%, a receptivity that is 0.61 that of ^{13}C, and a quadrupole moment ($I = \frac{5}{2}$), ^{17}O is relatively difficult to study. Nevertheless, because of the importance of oxygen in both inorganic and organic chemistry as well as biology, the ^{17}O nucleus has been quite widely studied both in natural abundance and in isotopically enriched samples. For small molecules, linewidths are typically in the range of several tens to several hundreds of hertz. With large molecules or viscous solutions much larger linewidths may be encountered.

2.2.6.1 Chemical Shifts. Water is the usual chemical-shift reference, being readily available in enriched form. A chemical-shift range of over 1000 ppm affords wide spectral dispersion. Figure 13 shows representative chemical-shift ranges for some organic groups. There is a clear distinction between single and double carbon–oxygen bonds. However, a single resonance is observed

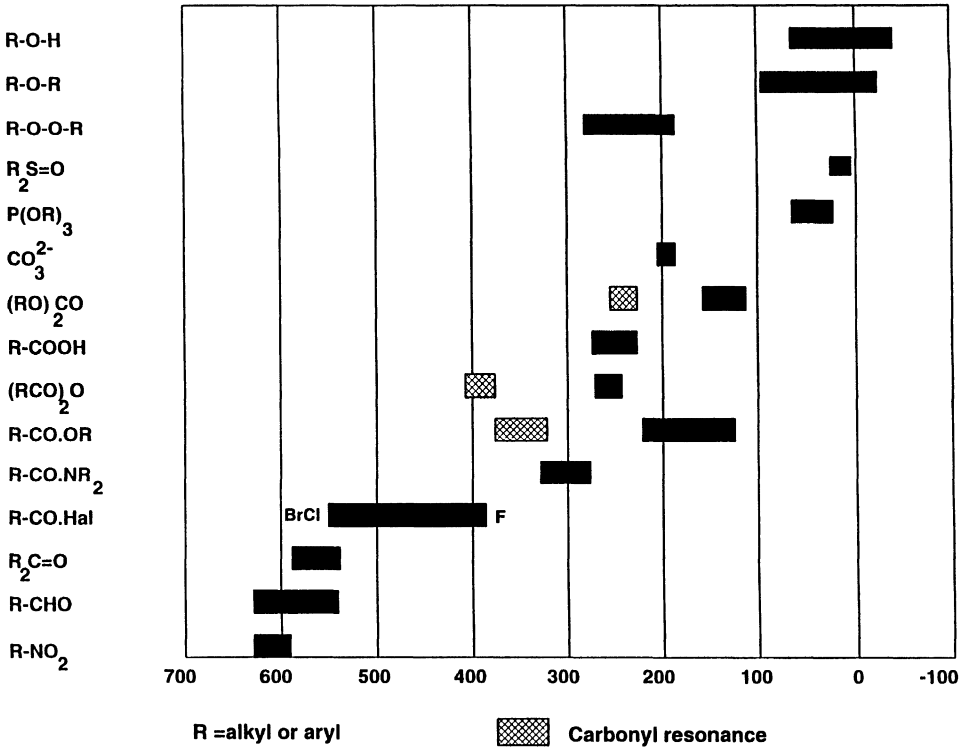

FIG. 13. ^{17}O NMR chemical-shift ranges.

from a carboxylic group at the average position expected for the carbonyl and OH groups as a result of rapid proton exchange facilitated by dimeric species. Primary alcohols and ethers resonate close to the water position, while lower field shifts are observed for secondary and tertiary species.

2.2.6.2 Spin–Spin Coupling. Since ^{17}O resonances are usually broad, the majority of reported values are for couplings between directly bonded nuclei. Values of $^1J(^{17}O,H)$ in water and alcohols are ca. 80 Hz. The $^1J(P,^{17}O)$ couplings involving P=O groups cover the range 145–210 Hz. In contrast, couplings involving P–O groups are usually smaller (90–100 Hz). Some two-bond couplings have been reported. Examples of two-bond hydrogen–oxygen couplings are 38, 7.5, and 10.5 Hz for the O–CH, O–CH_3, and HC=O interactions in methyl formate.

2.2.7 Silicon (^{29}Si) ^{29}Si, with an abundance of 4.7%, is the only naturally occurring isotope of silicon with a nonzero spin ($I = \frac{1}{2}$). While its receptivity is 2.1 times that of the ^{13}C nucleus, there are problems associated with its detection. In common with the ^{15}N nucleus, a negative magnetogyric ratio results in negative NOEs when broadband proton decoupling is employed as is usually the case. Inverse-gated decoupling and relaxation reagents are again used to circumvent the problem. Broad signals from the glass in the NMR probe head or tube can also cause problems. The glass can be replaced by alternative materials such as Teflon, or the spectrum obtained from a blank run (i.e., without sample) can be subtracted from the spectrum of the sample.

2.2.7.1 Chemical Shifts. The ^{29}Si resonance of TMS is the usual chemical-shift reference. ^{29}Si chemical shifts cover a range of ca. 400 ppm. The large dispersion arising from structural effects makes ^{29}Si NMR a valuable tool for molecular structure determination. The nature of the atoms directly bonded to silicon is of particular importance. Table 5 lists chemical shifts for a series of compounds SiX_nMe_{4-n} (Me denotes methyl). ^{29}Si NMR has proved to be an ideal tool for

Table 5. ^{29}Si chemical shifts for SiX_nMe$_{4-n}$.

X[a]	$n = 1$	$n = 2$	$n = 3$	$n = 4$
H	−15.5	−37.3	−65.2	−93.1
Et	1.6	4.6	6.5	8.4
Ph	−5.1	−9.4	−11.9	−14.0
OMe	17.2	−2.5	−41.4	−79.2
NMe$_2$	5.9	−1.7	−17.5	−28.1
Cl	30.2	32.2	12.5	−18.5

[a] Et denotes ethyl; Ph denotes phenyl.

structure determination of polysiloxane macromolecules.

2.2.7.2 Spin–Spin Coupling. Couplings to abundant nuclei such as the proton and fluorine can be measured either in a coupled ^{29}Si spectrum or by observing the ^{29}Si satellites in the spectrum of the abundant nucleus. The latter are relatively weak compared with peaks from the main ^{28}Si-containing species. Nevertheless, many determinations were made in this way during the 1960s using cw spectrometers. In a well-resolved ^{1}H NMR spectrum, the satellites due to a $^{2}J_{SiH}$ coupling of 6.8 Hz are readily observed at the base of the TMS reference peak. One-bond proton–silicon coupling constants in silanes are usually in the range 150–380 Hz, the magnitude of J increasing with substituent electronegativity. The literature contains data about coupling to a range of other nuclei, including ^{31}P, ^{13}C, ^{15}N, and ^{29}Si itself.

Published data have been reviewed by Williams and Cargioli (1979).

2.3 Double-Resonance Techniques

In Sec. 2.1.2, reference is made to the use of proton-noise (or broadband) decoupling to simplify ^{13}C spectra and improve the signal-to-noise ratio. This is just one example of a large number of double-resonance techniques that are available either to simplify spectra or as spectral interpretation tools to aid structure elucidation.

2.3.1 Homonuclear Spin Decoupling

The principles of homonuclear double resonance can be illustrated by considering two coupled protons, A and B, as shown in Fig. 5. In the double-resonance experiment, nucleus A is observed in the usual way in a cw or FT experiment during simultaneous selective irradiation at the nucleus-B resonance frequency. This irradiation induces transitions between the two spin states of nucleus B. If sufficient irradiating power is applied, B flips between the α and β spin states so rapidly that nucleus A can no longer distinguish between the two orientations of B and perceives an average orientation. The coupling J_{AB} disappears and the A signal doublet collapses to give a single line. In the same way, irradiation of the methyl triplet of ethylbenzene (Fig. 8) results in the methylene quartet collapsing to a singlet.

2.3.2 Heteronuclear Spin Decoupling

In the case of heteronuclear spin decoupling, the decoupled and observed nuclei are of different chemical types. The experiments are illustrated with examples in which protons are decoupled and the effect is observed on the ^{13}C spectrum. However, the methodologies are equally applicable to other heteronuclear spin systems.

2.3.2.1 Selective Spin Decoupling. Selective spin decoupling is analogous to the homonuclear decoupling experiment described above. In general, higher rf power is required than for homonuclear decoupling but lower than is generally used for broadband heteronuclear decoupling. An individual proton signal is selected and irradiated to remove all couplings to that specific proton in the ^{13}C spectrum.

2.3.2.2 Broadband Spin Decoupling. If the proton decoupler frequency is set to the center of the proton spectrum and modulated by using a noise generator with a bandwidth wide enough to cover the complete proton region, then this is equivalent to simultaneously irradiating every proton frequency. This effectively decouples all the protons in the molecule, producing the effect shown in Fig. 7. Two problems are, however, associated with this type of broadband decoupling. First, the high rf power required heats the sample, and second, the decoupling is ineffective if the bandwidth to be irradiated is large. The latter is more of a problem on higher-field spectrometers and where broadband decoupling of other nuclei such as ^{19}F or ^{31}P with their large spectral widths is required. In recent years alternative methods—e.g. WALTZ-16 (see Shaka *et al.*, 1983)—have largely overcome these problems. As has been mentioned in previous sections, broadband decoupling also produces a change in intensity, known as the nuclear Overhauser enhancement (NOE), of the observed signals. This occurs because the decoupling causes additional relaxation of the observed nucleus, resulting in increased population of the ground-state energy levels. For ^{13}C the increase in intensity can be as much as 200%, but resonances from quaternary carbon atoms that do not have attached protons are little affected.

2.3.2.3 Gated Decoupling. When decoupling is turned on or off, coupling effects appear and disappear instantaneously. The NOE effects, however, grow and decay at rates related to some relaxation process. As a result it is possible by computer control of the decoupler to separate the two effects completely.

Sometimes it is necessary to record broadband proton-decoupled spectra without any NOE. The NOE may be undesirable either because quantitative information is needed about the relative number of atoms contributing to each peak in, for example, a ^{13}C spectrum or because the nucleus being studied has a negative NOE (e.g., ^{15}N or ^{29}Si). Such spectra are also used as controls when measuring NOE values. In other instances coupled spectra with NOE are used as a means of improving sensitivity.

Figure 14 shows an FT ^{13}C experiment with three possible broadband decoupling schemes. The repetitive pulse experiment consists of three parts: a relaxation delay *t*, a ^{13}C rf pulse, and the acquisition time *AT* during which time the FID is sampled. *AT* and *t* are of the order of seconds, while the pulse duration is a few microseconds. For conventional broadband decoupling (a) the decoupler is on throughout the experiment. If the decoupler is gated off just before the rf pulse as in (b), NOE effects can be observed without any decoupling. Conversely, if the decoupler is switched on when the pulse is applied and off as soon as the acquisition is complete as shown in (c), the resultant decoupled spectrum does not exhibit an NOE. The latter experiment is referred to as inverse gated decoupling.

2.3.2.4 Off-Resonance Decoupling. While the broadband proton decoupling described above produces considerable simpli-

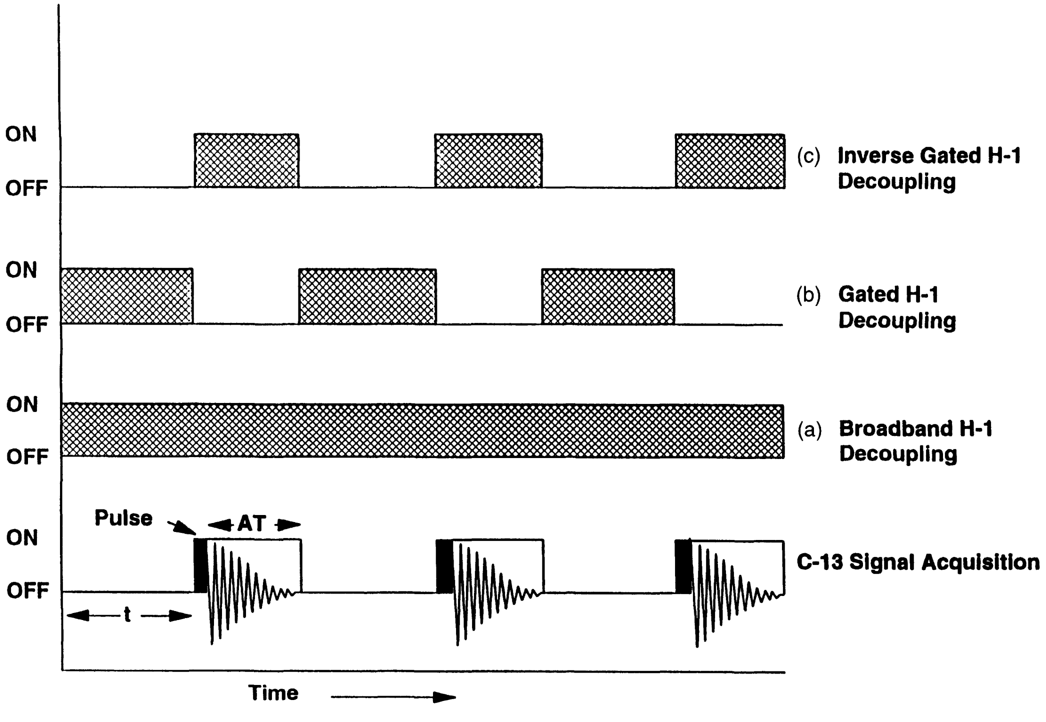

FIG. 14. ^{1}H broadband and gated decoupling schemes.

fication of ^{13}C and other spectra, it removes all the coupling information. As a result the spectra may be difficult to assign. In the early days of ^{13}C NMR, a technique known as off-resonance decoupling was used to circumvent this problem. Off-resonance decoupling is achieved by setting the central frequency of the broadband proton decoupler 1000–2000 Hz outside the proton spectrum. This results in residual couplings from protons directly bonded to ^{13}C atoms, whereas longer-range couplings are lost. The magnitude of the residual coupling, which is smaller than the true coupling J, is determined by the amount of the offset and the power of the decoupler. In such spectra it is usually possible to determine the multiplicities of all the individual carbon atoms; i.e., methyl carbons appear as quartets, methylenes as triplets, etc. Nowadays the use of gated decoupling for multiplicity determination has largely been replaced by spectral editing experiments (Sec. 2.4).

2.3.3 NOE Difference Spectroscopy

The phenomenon of NOE was introduced in Sec. 2.1.3 to explain the increase in intensity of ^{13}C peaks when broadband proton decoupling is used. In general terms the NOE is defined as the change in signal intensity of a nucleus when a second nucleus is irradiated. It is beyond the scope of this article to give a theoretical treatment of the origin of the effect, which is caused by through-space nuclear relaxation by dipolar interaction. Among other factors the magnitude of the enhancement has a $1/r^6$ dependence on the distance between the two nuclei concerned. Therefore, homonuclear NOEs between protons have been used extensively to obtain information about internuclear distances and thereby to distinguish between possible structures.

The most convenient way of observing such enhancements, which rarely exceed 20% in the most favorable cases and are sometimes <1%, is by a technique known as NOE difference spectroscopy. In such an experiment two spectra are recorded. The first is a conventional ^1H spectrum, while in the second a chosen resonance is selectively irradiated during the interpulse delay of an FT experiment. The second spectrum is then subtracted from the first to give a "difference" spectrum.

Figure 15(a) shows the ^1H NMR spectrum for the following structure, denoted as structure **1**:

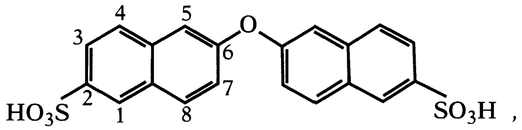

From this spectrum alone it is not obvious whether the ether linkage is at C-6 or C-7. The peak at 8.2δ is clearly due to H-1, while that at ca. 8.0δ is from H-5 or H-8, depending on which isomer is present. Figure 15(b) shows the difference spectrum obtained when the resonance at 8.0δ is selectively irradiated. The irradiated resonance, which is partially or fully saturated, gives a negative signal in the difference spectrum. Positive signals are obtained from those resonances that exhibit an enhancement. Since an en-

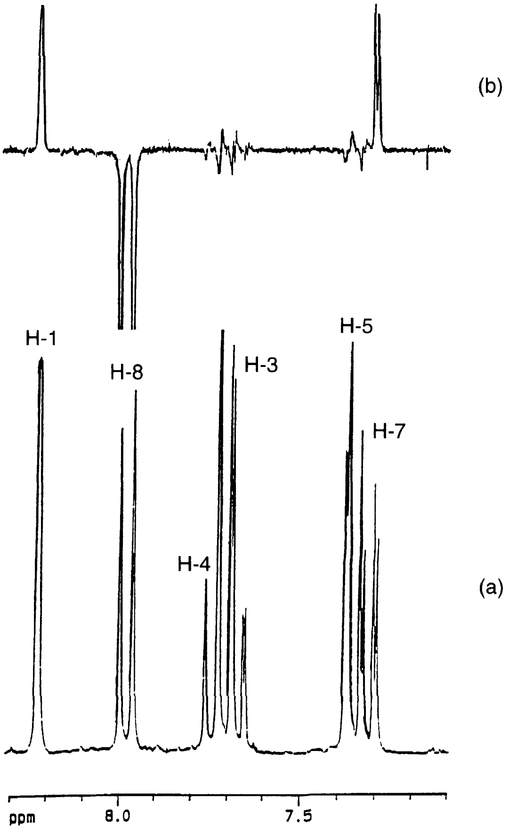

FIG. 15. (a) 250-MHz ^1H NMR spectrum of structure 1; (b) NOE difference spectrum on irradiation of H-8.

hancement is observed for H-1, this is a good indication that H-1 and H-8 are close together. Measurable effects can be observed between atoms up to about 400 pm apart. Since the above description is rather simplistic, it should be stressed that because dipolar relaxation of a particular nucleus may involve interaction with more than one other nucleus, a basic understanding of the processes involved is a prerequisite to applying the technique successfully. The reader is referred to Noggle and Shirmer (1971) and Neuhaus and Williamson (1989).

2.4 Spectral Editing Experiments

The basic FT experiment is a single-pulse experiment. This means that only one RF pulse is applied prior to recording the free-induction decay (FID) response. A large number of experiments have been developed, for a variety of purposes, in which more than one pulse is applied prior to recording the FID. The most frequently used multiple-pulse experiments are probably those for multiplicity determination in ^{13}C NMR. Both the *J*-modulated spin-echo (LeCocq and Lallemand, 1981) and the distortionless enhancement by polarization transfer (DEPT) (Doddrell *et al.*, 1982) experiments are extensively used for this purpose.

Figure 16 shows the pulse sequence used for the former experiment, which includes the three periods characteristic of all modern pulse experiments, namely preparation, evolution, and detection. The preparation period is a relaxation delay ending with a 90°_x pulse along the *x* axis of the rotating frame. The decoupler is switched off, and during the first half of the evolution period, the magnetization in the *xy* plane evolves under the influence of the proton–carbon spin–spin (*J*) couplings. At the beginning of the second period the decoupler is switched on again, and a 180°_x pulse refocusses all effects other than *J* modulation. In the detection period a proton-decoupled FID signal with NOE is observed. If $t = 1/J$, signals from quaternary and CH_2 carbons are in antiphase to those from CH and CH_3 carbons. Figure 17 shows such a *J*-modulated spin-echo ^{13}C spectrum for limonene (structure **2**),

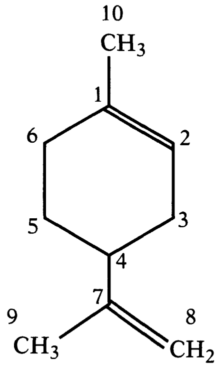

,

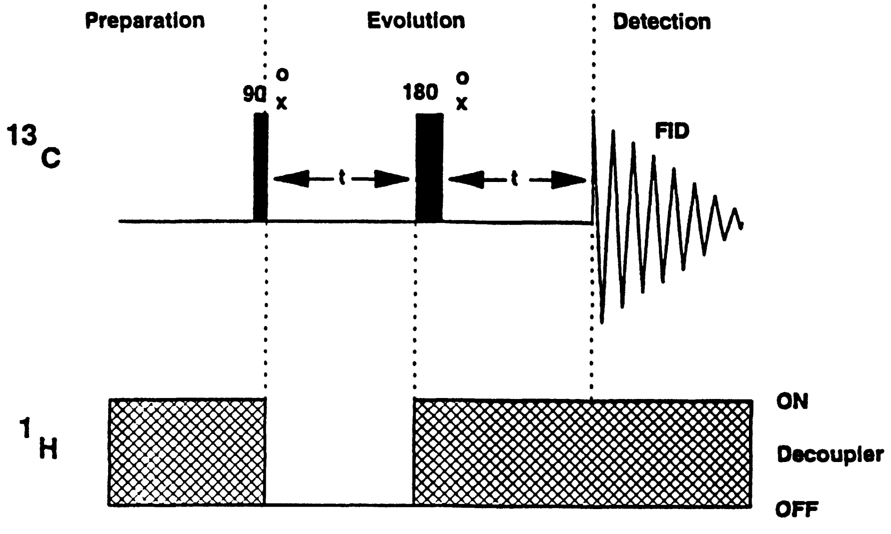

FIG. 16. *J*-modulated spin-echo pulse sequence.

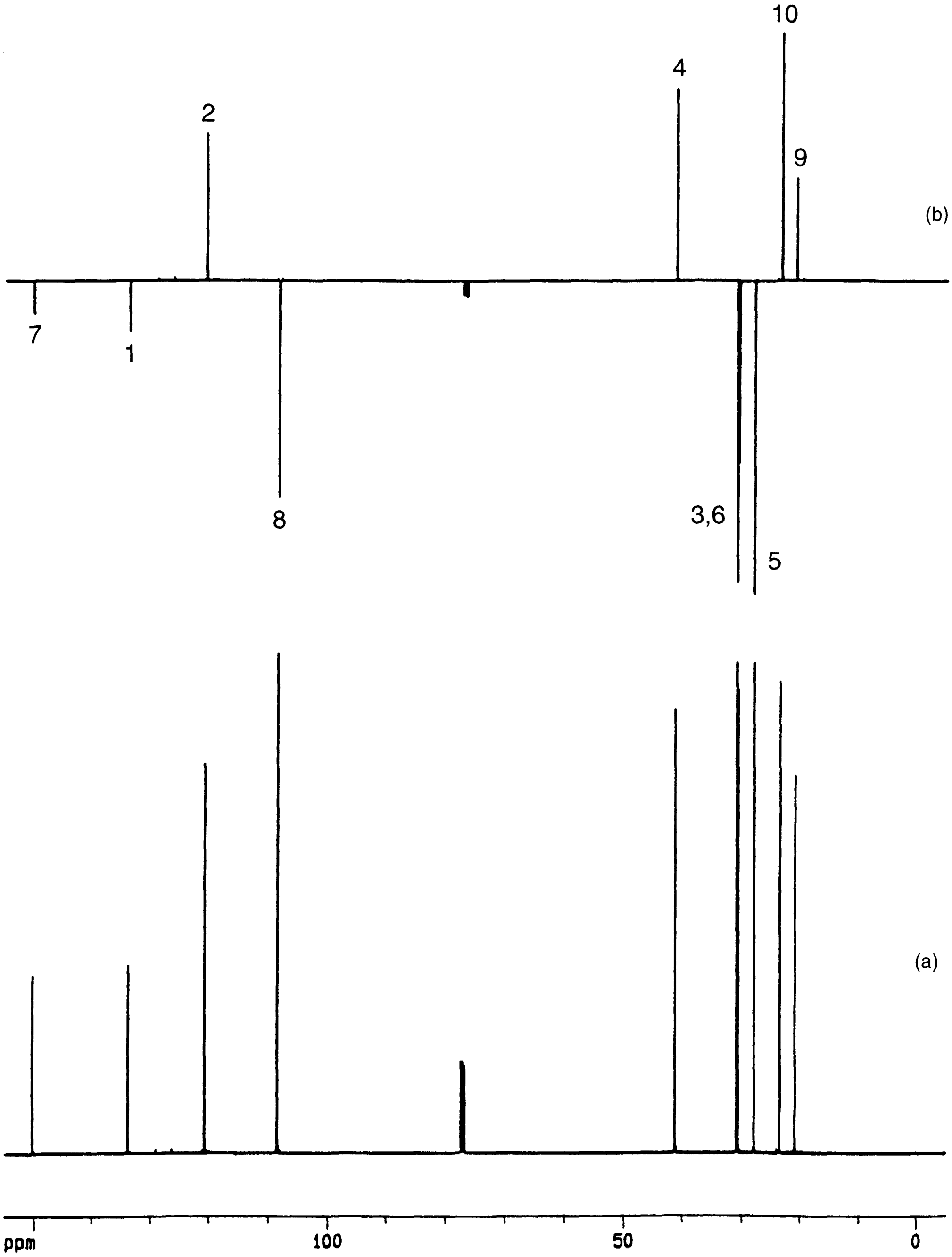

FIG. 17. 100 MHz ^{13}C NMR spectra of limonene (structure 2): (a) broadband decoupled, (b) *J*-modulated spin-echo.

together with its conventional broadband proton decoupled spectrum. When $t = 1/2J$, only signals from quaternary carbons are observed. While widely used, this experiment has two shortcomings for multiplicity assignment. First, signals from CH and CH_3 carbons cannot be distinguished, and second, it is sensitive to variations in J values. One-bond carbon–proton couplings are typically ca. 125 and ca. 160 Hz for saturated aliphatic groups and aromatic/olefinic carbon atoms, respectively. For such compounds use of a compromise value for t works well. However, for carbon atoms with much larger couplings such as alkynyl carbons and some unsaturated heterocyclic carbons, it is less successful. In such instances the DEPT technique is more suitable. Compared with the J-modulated spin-echo experiment, it is insensitive to variations in $^1J_{CH}$. The method is based on the use of polarization transfer whereby magnetization is transferred from ^1H to ^{13}C nuclei, and as a result quaternary carbon atoms are not detected in such experiments. This experiment involves pulsing both the ^1H and ^{13}C channels. The final proton pulse is along the y axis, and by performing two experiments with angles for this pulse of 90° and 135°, full multiplicity data can be obtained. The 135° experiment gives negative CH_2 but positive CH and CH_3 carbon signals, while in the 90° experiment only CH signals are observed.

2.5 Multidimensional NMR

A wide range of two-dimensional experiments are now available as relatively routine tools, and more recently three- and four-dimensional methods have been developed. In general the experiments are used to extract information from complex spectra (e.g., about which nuclei are J coupled to each other and to measure J couplings). In this article it is only possible to give a very brief introduction to two-dimensional methodology together with a few examples of its application in ^1H and ^{13}C NMR. Further information about this most important field is available from Derome (1987), Sanders and Hunter (1987), Morris (1986), and Benn and Gunther (1983).

2.5.1 Basic Principles of 2D NMR The common feature of 1D multiple-pulse experiments is the time sequence preparation–evolution–detection, where the detected signal is only a function of the detection time t_2. The important difference in 2D NMR is that the evolution time t_1 is now a variable. In a 2D experiment, n separate experiments are performed with incremented values of t_1. For each experiment a free-induction decay $S(t_2)$ is measured. In this way a matrix $S(t_1,t_2)$ is built up. Fourier transformation of this matrix with respect to t_2 gives a series of spectra $S(t_1,f_2)$. Further transformation with respect to t_1 gives a matrix $S(f_1,f_2)$, which is a spectrum in two independent frequency dimensions. The dispersion of the signals as a function of f_1 depends on the effect of the pulse sequence used on the spin system as t_1 is varied. In general there are two distinct types of 2D experiments. These are referred to as J-resolved and correlated spectroscopy. In a J-resolved spectrum one frequency axis (f_1) contains spin–spin coupling information and the other (f_2) chemical-shift information. In correlated spectroscopy both axes contain chemical-shift information, the connection between the two being via spin–spin coupling, NOE, or exchange effects.

2.5.2 Homonuclear Chemical-Shift Correlation The most frequently applied 2D technique is proton homonuclear correlation using one of the many variants of the COSY (correlated spectroscopy) experiment. The basic COSY experiment consists of two 90°_x pulses separated by a time t_1 as shown in Fig. 18. Figure 19 shows the aliphatic part of the COSY spectrum obtained from the compound having structure **3**,

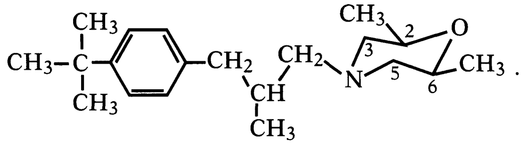

The spectrum is presented as a contour plot in which intensities above a chosen threshold level are plotted like height contours on a map. Since both frequency domains are proton chemical shifts, the matrix is square. Contours along the diagonal of the square correspond to the peaks in a one-dimensional spectrum. Homonuclear couplings give rise to the off-diagonal contours or cross peaks. The one-dimensional spectrum

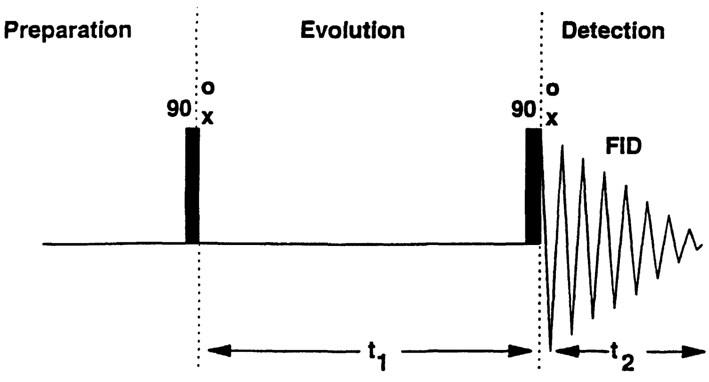
Preparation
Evolution
Detection
90°x
90°x
FID
t1
t2

FIG. 18. Basic COSY pulse sequence.

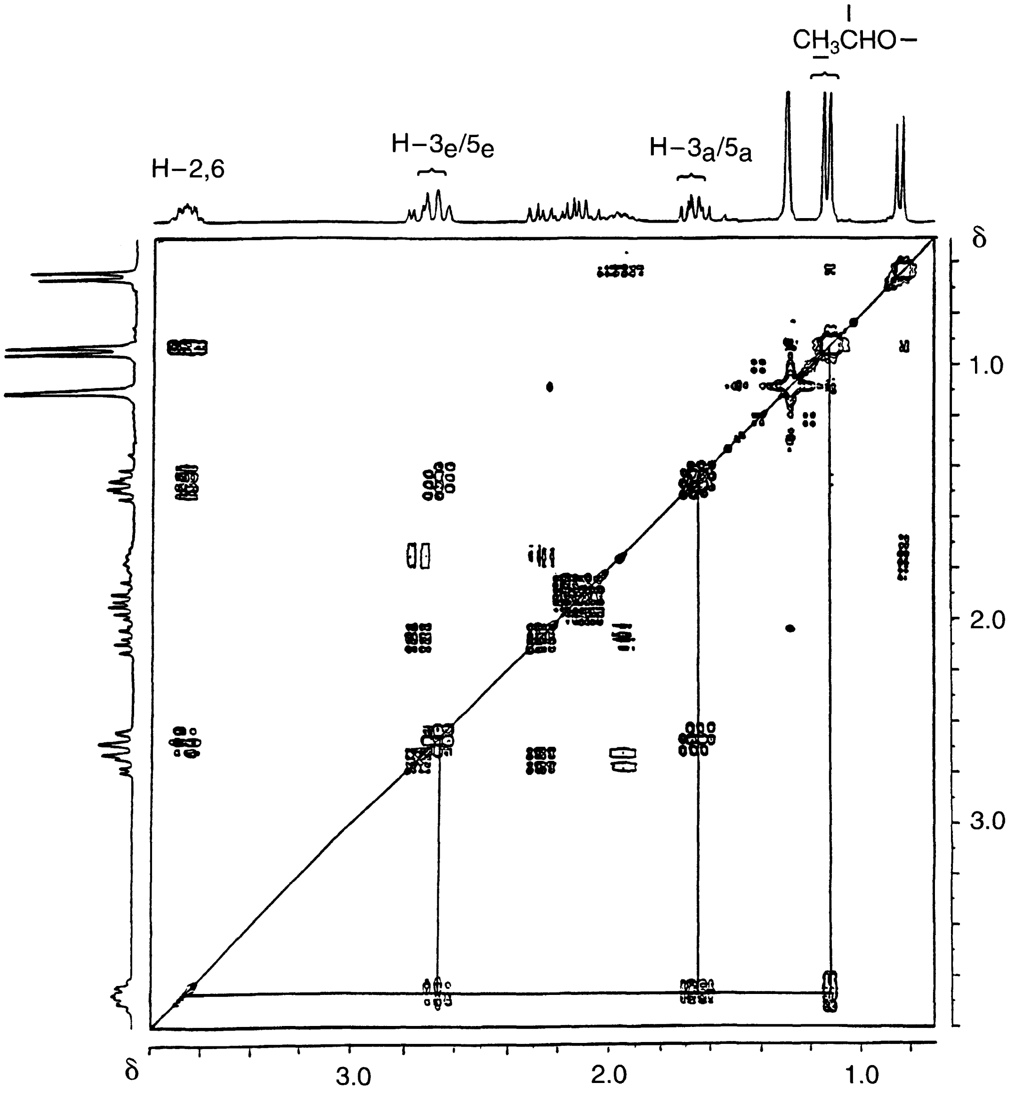

FIG. 19. 250-MHz ^{1}H COSY spectrum of structure 3.

is also plotted as a projection along both axes. The cross peaks provide the same information about proton–proton connectivities as can be obtained from a series of homonuclear decoupling experiments, but all in a single spectrum!

The connectivities for a particular proton can be extracted by drawing a horizontal line from the relevant diagonal peak that intercepts cross peaks corresponding to correlations. In Fig. 19 the multiplet at 3.68δ is due to H-2/6, and from the cross peaks it shows correlations to three other multiplets. By drawing a vertical line from each cross peak back to the diagonal, it can be determined with which peak the first diagonal peak is correlated (i.e., coupled). In this case correlations are observed at 1.11δ, 1.65δ, and 2.67δ from the attached methyl groups, H-3_a/5_a, and H-3_e/5_e, respectively. Note that the cross peaks are found symmetrically on both sides of the diagonal.

Two-dimensional methodology is inherently more time consuming than 1D principally because of the requirement to perform a large number of individual experiments. However, because of the relatively high sensitivity of the proton, COSY spectra can be readily obtained if sufficient sample is available, and a few milligrams will yield a good spectrum on a modern high-field instrument.

2.5.3 Homonuclear *J*-Resolved Spectra Overlapping absorptions in a 1D proton spectrum can sometimes be resolved by a homonuclear *J*-resolved experiment, in which chemical shifts are presented on one axis and coupling constants on the other. The pulse sequence used is

$$90^\circ_x - t_1/2 - 180^\circ_x - t_1/2 - \text{acquire}.$$

Figure 20 shows the aliphatic part of the spectrum obtained from the structure shown in Sec. 2.5.2. The spectrum was recorded to resolve the multiplet from H-3_a/5_a at 1.65δ. In the *J*-resolved spectrum the presence of two columns in this region, each from a doublet of doublets, clearly shows that H-3_a and H-5_a are magnetically inequivalent. Similarly two columns are observed for H-3_e and H-5_e from the multiplet at 2.67δ. In each of these cases the two nuclei involved are not coupled to each other and as such are ideal for this type of study. Where the overlapping resonances are from nuclei that are strongly coupled to each other, they cannot be resolved in this manner because second-order effects are observed in the 2D spectrum in the form of additional columns of peaks. Such a situation exists in Fig. 20 in the region 2.05δ to 2.15δ from the protons of the methylene group attached to the benzene ring.

2.5.4 Heteronuclear Chemical Shift Correlation (HETCOR) In the basic HETCOR experiment the peaks of a ^{13}C spectrum are correlated with those of a ^1H spectrum showing the specific protons that are attached to each ^{13}C nucleus. Figure 21 shows part of the spectrum of 2-carboxy-6-hydroxynaphthalene (structure **4**),

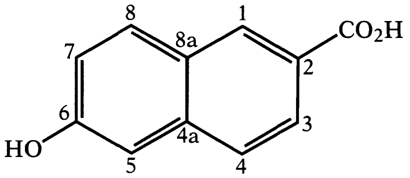

The relevant parts of the corresponding 1D spectra are plotted along the axes. A correlation is observed as a cross peak at the intersection of a horizontal line drawn from a proton resonance and a vertical line drawn from a carbon peak. Correlations are not observed for quaternary carbon atoms. The technique is a very important tool for chemical-shift assignment and thus structure elucidation. A number of variants of this basic experiment have been developed, some of which enable long-range proton–carbon correlations to be established. The original ^1H/^{13}C HETCOR experiments involved detection of a ^{13}C signal at the end of the pulse sequence. However, a significant enhancement in sensitivity can be achieved if nuclei with small magnetogyric ratios such as ^{13}C are detected through their effects on the more sensitive ^1H nuclei. In ^1H/^{13}C correlation experiments where optimum sensitivity is required, the proton-detected heteronuclear multiple-quantum coherence (HMQC) (see Bax and Subramanian, 1986) and heteronuclear multiple-bond connectivity (HMBC) (see Bax and Summers, 1986) are frequently used for one-bond and long-range correlations, respectively.

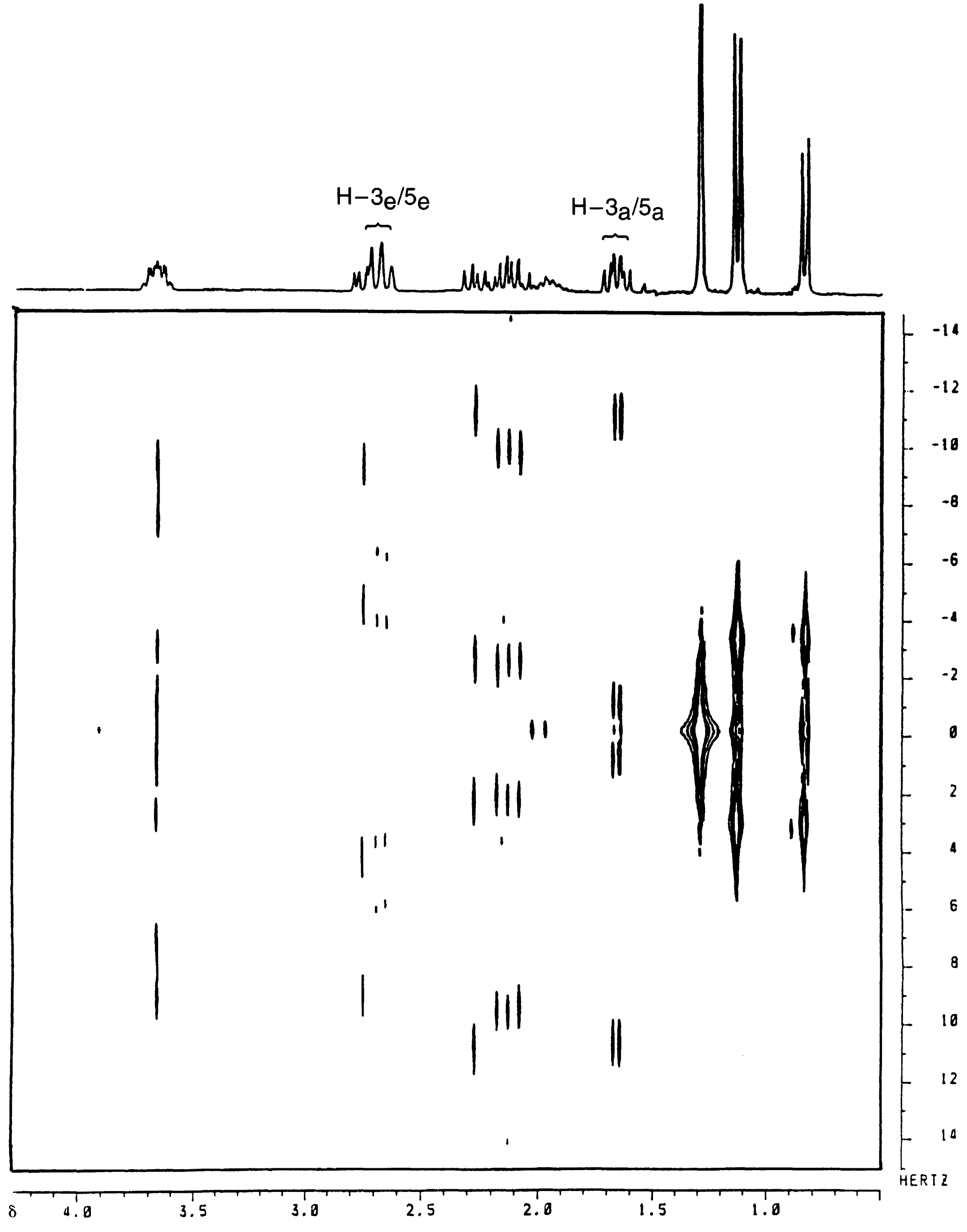

FIG. 20. 250-MHz ^{1}H 2D *J*-spectrum of structure 3.

2.5.5 Heteronuclear *J*-Resolved Spectra

In heteronuclear *J*-resolved spectra, ^{13}C chemical shifts are presented on one axis and proton–carbon *J* couplings on the other. The information content is equivalent to that in a proton-coupled ^{13}C spectrum but without the severe overlap of multiplets that is usually encountered in the latter. In common

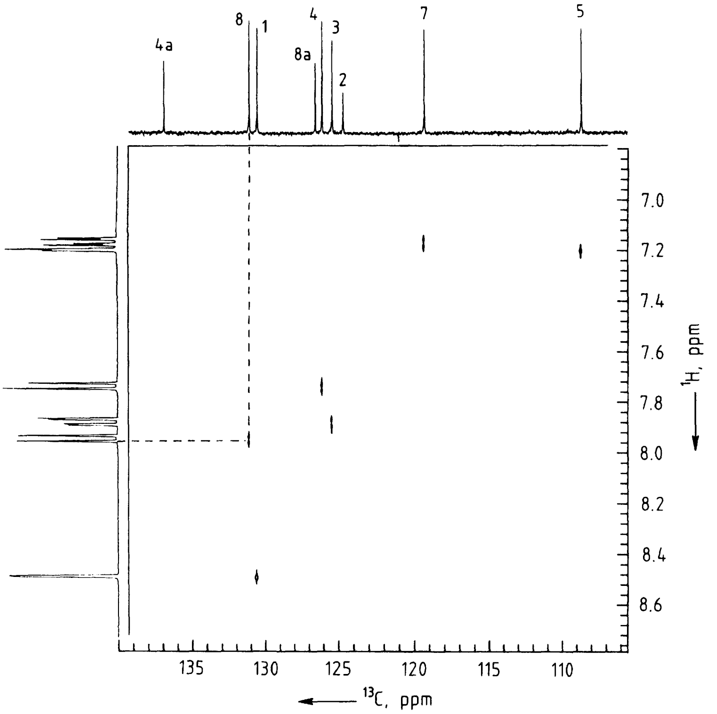

FIG. 21. Carbon–proton (400 MHz) shift-correlated spectrum of structure **4**.

with off-resonance proton decoupling, *J*-modulated spin-echo, and DEPT experiments, it facilitates multiplicity determination. In addition, it enables proton–carbon coupling constants to be measured.

2.5.6 Homonuclear NOE Correlation The 2D NOESY (nuclear Overhauser enhancement spectroscopy) technique in principle enables all the proton–proton NOE effects to be assembled in a single spectrum. Such spectra are very similar in appearance to COSY spectra, with contours on the diagonal representing the 1D spectrum and off-diagonal contours arising from through-space proton–proton interactions between protons that are nearby in space. The technique is extensively used in the determination of structures and conformations of very large molecules such as proteins. For smaller molecules with shorter correlation times where NOEs are small and slow to build up, 1D NOE difference spectroscopy is more widely used.

2.5.7 3D and 4D NMR The development of 2D spectroscopic techniques led to NMR being used as a tool for structure determination of small proteins and DNA fragments (see Wüthrich, 1986). Early investigations of proteins were limited to relatively small molecules up to a size of about 80 amino acid units since signal overlap in 2D spectra of larger proteins was too great. The impetus to study larger systems has led to a massive investment in the design of three- and four-dimensional experiments and the associated instrumental developments. It is now possible to examine systems with molecular weights up to about 35 kDa when the proteins have been uniformly labeled with ^{13}C and ^{15}N. Oschkinat *et al.* (1994) have recently reviewed the field of protein structure determination with 3D and 4D NMR. A very brief description of the type of methodology used will be given here.

Structure determination in biological macromolecules such as proteins and oligonucleotides is based mainly on the determination of *J*-coupling connectivities from homonuclear and heteronuclear chemical-shift correlation experiments, NOEs from NOESY experiments, and *J* couplings from a variety of correlation experiments. NOEs and *J* couplings correlate with distances between nuclei and dihedral angles, respectively. The following information is required from an NMR study:

1. the complete sequence assignment,
2. the complete assignment of side-chain signals,
3. the collection of a sufficient number of precisely measured NOEs, and
4. long-range homonuclear and heteronuclear *J* couplings where possible.

These data can then be used to calculate three-dimensional structures using distance geometry and molecular-dynamics computer programs.

The principles of 3D and 4D NMR spectroscopy are an extension of those used in 2D spectroscopy. While a 2D experiment has one evolution time t_1, 3D and 4D experiments have two and three evolution times, respectively. In a 3D experiment the signals resulting from the different evolution times t_1 and t_2 are ordered in a 3D matrix. The 3D spectrum is then obtained by Fourier transformation with respect to t_1, t_2, and the detection time t_3. Many 3D experiments are combinations of 2D experiments and are designed to reduce the overlapping of signals in a related 2D experiment. An example of such a 3D experiment is ^{15}N-NOESY-HMQC, which is used in the sequential assignment of α-helical regions in proteins and uses uniformly ^{15}N-labeled samples. The experiment is used to find cross peaks between NH signals from neighboring amino-acid groups. If two NH proton signals have the same chemical shift, then four cross peaks to other amino acids are expected at this frequency in a 2D NOESY experiment, and an unambiguous assignment is not possible. In the ^{15}N-NOESY-HMQC spectrum, where the f_1, f_2, and f_3 axes represent ^{1}H, ^{15}N, and ^{1}H chemical shifts, respectively, the signals are found on different ^{15}N planes if the ^{15}N nuclei have different chemical shifts.

The very large number of individual, constituent experiments in a 3D or 4D experiment make such studies very time consuming, and a lot of research work is being carried out to design more efficient experiments. One approach is to minimize the total number of multidimensional experiments required to obtain the required correlations, NOEs, etc. needed for a protein study. Another approach that is being very successfully employed in many multidimensional experiments uses pulsed magnetic-field gradients instead of phase cycling to obtain artifact-free spectra. Norwood (1994) has recently reviewed the use of magnetic gradients in multidimensional solution NMR and in other areas of NMR such as imaging and diffusion studies.

2.6 Application to Chemical Structure Determination

Solution NMR spectra usually contain a wealth of information. Previous sections have provided an overview of the parameters that can be determined from NMR spectra, and techniques that can be used for the assignment of signals in spectra. Here some of the applications of solution NMR spectroscopy to chemical structure determination are briefly summarized. The list is by no means exhaustive, and further information about applications can be found in many of the reference books as well as the specialist literature referred to in this section.

NMR spectroscopy is the most powerful tool available for the unambiguous determination of molecular structures of compounds in solution. In a modern chemical research laboratory, the principal tools for structure determination are usually mass spectrometry, to obtain molecular mass data, and NMR spectroscopy, which provides detailed information about the groups present and how they are assembled. The abundant nuclei ^{1}H, ^{19}F, and ^{31}P have been used for well over 30 years to study the structures of organic and inorganic chemicals. In the early 1970s the advent of FT methodology enabled a much wider range of nuclei to be studied and in particular resulted in an explosive growth in the use of ^{13}C NMR as a complementary technique to ^{1}H NMR in the identification of organic compounds. The most useful NMR parameters are, in general, chemical shifts, coupling constants, and integrals from which structural information can be deduced.

In addition to the study of monomeric species, NMR is a long established tool for studying the structures of synthetic polymers. In the case of copolymers the various groups present can be identified, their relative concentrations determined, and often information about sequence distributions obtained. The latter provides a measure of the random/block character of the polymer. While ^1H and ^{13}C are the most commonly studied nuclei for such work, several others such as ^{19}F and ^{29}Si have been used where appropriate. For vinyl polymers, ^{13}C and ^1H NMR spectra provide information on tacticity. Chain branching and end groups can also be identified from NMR spectra of polymers.

During the last decade NMR has developed as a powerful tool for biochemical structure determination. The most impressive applications are in the determination of protein structures with molecular masses up to the 25 000–35 000 range (see Bax and Grzesiek, 1993). Such work has been made possible by the development of multidimensional NMR techniques. 2D and 3D techniques are widely used for such work, and 4D experiments are being developed (see Sec. 2.5.7). Since NMR allows the study of biological macromolecules in solution, it has proved complementary to solid-state studies by x-ray crystallography. The value of the solution structure lies in the fact that it provides a starting point for the study of protein/substrate interactions, a knowledge of which may then aid drug design.

NMR has also become established as a valuable technique in the field of clinical biochemistry where ^1H NMR, in particular, is used to study complex biochemical fluids such as plasma, urine, and bile. The major problem associated with such work is the removal or reduction of the very large water signal. One method involves freeze drying the sample and redissolving it in deuterated water. NMR techniques are also available for selective suppression of the water signal. Under the conditions used only low-molecular-mass constituents are observable, the signals from proteins and other large molecules not being resolved. Applications include screening and monitoring for metabolic diseases, the study of biochemical mechanisms associated with disease processes, and the identification of drug metabolites. Nicholson and Wilson (1989) have reviewed applications.

During the past decade NMR has become an increasingly important tool for the determination of enantiomeric purity (see Parker, 1991). The four different substituents around a tetrahedral chiral center (see Sec. 2.2.1.2), often a carbon atom, can be arranged in two ways. The two arrangements, which are not superimposable, are mirror images of each other. Isomers that differ only in their arrangement at a single chiral center are known as enantiomers. Lactic acid is an example of a chiral compound that exists as two enantiomers:

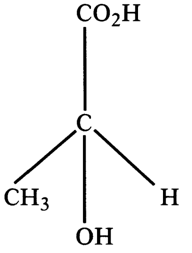

,

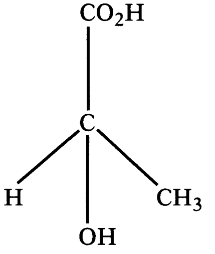

.

Many pharmaceutical and agrochemical products are chiral with only one enantiomer having the required effect. At best the other is inactive; at worst it may have undesirable properties. This situation has led to a surge in enantioselective synthesis, the products of which require analysis. In an achiral medium, enantiomers cannot be distinguished from their NMR spectra, which are identical. Molecules that contain two chiral centers have four isomeric forms, which comprise of two pairs of enantiomers, known as diastereoisomers. Diastereoisomers have different physical properties and can be distinguished by NMR. Therefore, the determination of enantiomeric purity by NMR requires the use of a chiral auxiliary that converts a mixture of enantiomers into a mixture of diastereoisomers. Chiral lanthanide shift reagents and chiral solvating agents form diastereomeric complexes *in situ* and can be used directly for NMR analysis. In favorable cases absolute configuration can be determined from the NMR spectrum of the diastereoisomer formed.

3. NMR OF SOLIDS

Solids often give very broad NMR spectra devoid of the fine structure that provides detailed information about chemical structures. The restriction of NMR to the liquid/solution or gas phase does, however, impose considerable limitations on the applicability of the NMR technique. Solid-state spectra can be useful to

1. study insoluble materials,

2. compare solution and solid-state molecular structures and conformations, and
3. study solid-state (e.g., crystallographic) properties.

Various aspects of NMR studies of solid samples up the early 1980s are covered by Fyfe (1983).

3.1 Magnetic Interactions in the Solid State

The principal interactions are

1. interaction with the applied magnetic field,
2. magnetic shielding by the surrounding electrons,
3. spin–spin coupling to other nuclei,
4. direct dipole–dipole interactions with other nuclei, and
5. quadrupolar interactions for nuclei with spin $>\frac{1}{2}$.

The first three have been discussed in the context of solution-state NMR. In solids all the interactions are anisotropic; i.e., they depend on molecular orientation, whereas in solution fast isotropic motion results in averaging. For chemical shifts and spin–spin couplings, discrete, average, isotropic values are obtained in solution. Dipolar and quadrupolar couplings are not observed in solution since the average values are zero. In solids spin–spin coupling interaction is usually much smaller than all the other interactions, and its effects are rarely observed. A more in-depth treatment on magnetic interactions in solids and low-resolution spectra can be found in the article NUCLEAR MAGNETIC RESONANCE AND QUADRUPOLE RESONANCE.

3.1.1 Dipole–Dipole Interactions Dipolar coupling results from direct dipole–dipole interactions between nuclei. Its magnitude is independent of the applied magnetic field H_0, proportional to the magnetogyric ratios of the nuclei involved, and inversely proportional to the cube of the distance between the two nuclei. Its effects are most significant for spin-$\frac{1}{2}$ nuclei such as ^1H with large magnetic moments. The spectrum produced by a pair of identical spin-$\frac{1}{2}$ nuclei consists of two lines whose positions are given by

$$\nu = \nu_0 \pm \tfrac{3}{4}R(3\cos^2\theta - 1), \qquad (11)$$

where R is the dipolar coupling constant and θ is the angle between the internuclear vector and the applied magnetic field. Such a situation exists for isolated H_2O molecules present as water of crystallization in single crystals of gypsum, $CaSO_4 \cdot 2H_2O$. A study of the orientation dependence of the splitting as a function of the angle between H_0 and the x crystallographic axis gives information about the orientation of the water molecules in the unit cell and the internuclear distance between the water protons. For polycrystalline material, averaging over all possible orientations to the magnetic field gives a "Pake doublet" (Pake, 1948).

3.1.2 Chemical-Shift Anisotropy If dipolar interactions are absent or can be removed, the NMR spectrum of a single crystal will contain one sharp line for each magnetically unique orientation of a nucleus relative to the direction of the applied magnetic field. However, the positions of the lines will change as the orientation of the crystal is changed. Shielding is a tensor property characterized by three principal components σ_{11}, σ_{22}, and σ_{33} specified in a molecule-fixed coordinate system. The observed shielding constant is given by a linear combination of the principal components, which have a $\cos^2$ dependence on the angle between the component and the applied magnetic field. For a microcrystalline powder sample, a broad line, known as a powder pattern, will result whose shape will depend on the principal components of the shielding tensor. For examples of chemical shift powder patterns, see Fig. 4 in the article NUCLEAR MAGNETIC RESONANCE AND QUADRUPOLE RESONANCE. Shielding anisotropies can be very large. For example, those for ^{13}C nuclei can be as large as 438 ppm (see Spiess *et al.*, 1971). While the principal components can be determined from a powder pattern, a single-crystal study is required to relate them to the molecular-axis system.

3.1.3 Quadrupole Interactions For nuclei with spin $> \frac{1}{2}$, solid-state NMR spectra are usually dominated by nuclear quadrupole interactions; even the effects of anisotropic chemical shifts and dipole–dipole interactions can be neglected to a first approximation. The latter may produce some line broadening. Quadrupolar effects result from the interaction of the nuclear electric quadrupole moment with the non-spherically-symmetrical field gradient around the nucleus. The effect of the quadrupolar interaction is to perturb the energy levels resulting from the interaction of the nuclei with the applied magnetic field. For instance, a spin-1 nucleus has three energy levels corresponding to $m_I = 1$, 0, and -1. For a single spin-1 nucleus, such as ^{2}H, in the absence of quadrupolar interaction, there would be two allowed, degenerate transitions giving a single peak. The quadrupolar interaction perturbs the energy levels such that the transitions are no longer degenerate, and a doublet results with the peak separation given by

$$\Delta v = \tfrac{3}{4}\chi(3\cos^2\theta - 1), \tag{12}$$

where χ is the quadrupole coupling constant and θ is the angle between the principal component of the symmetric electric-field-gradient tensor and the magnetic field. For deuterons attached to carbon the quadrupole coupling constant is in the range 170–200 kHz, while for ^{14}N and ^{35}Cl, values of up to 5 and 80 MHz, respectively, have been determined. For a powder all values of θ are possible and, as in the case of anisotropic dipolar and chemical shift effects, a powder pattern results. See Sec. 1.1.3 in the article NUCLEAR MAGNETIC RESONANCE AND QUADRUPOLE RESONANCE.

3.2 NMR Spectra of Solids

The type of experiments that can be performed and the structural information that can be obtained from a solid depend largely on the relative magnitudes of the interactions described in the preceding section. These will depend on the characteristics of the chosen nucleus, its abundance in the system under study, and the local and long-range molecular ordering present. Nuclei can be conveniently grouped into three types: abundant spin-$\frac{1}{2}$, dilute spin-$\frac{1}{2}$, and quadrupolar.

3.2.1 Abundant Spin-$\frac{1}{2}$ Nuclei In the absence of a quadrupole moment, homonuclear dipole–dipole interactions between abundant nuclei are the dominant contribution to the spectra of spin-$\frac{1}{2}$ nuclei such as ^{1}H and ^{19}F. For all but the simplest spin systems, spectra are broad and featureless (10–100 kHz). The relatively high sensitivity of experiments with abundant nuclei enables them to be studied with cw methodology. While little structural information can be obtained from such spectra, they are sensitive to molecular motions. NMR variable-temperature and relaxation studies can therefore provide information about such processes.

3.2.2 Dilute Spin-$\frac{1}{2}$ Nuclei Homonuclear dipolar interactions are essentially eliminated if the nucleus of interest is "dilute" in the system being studied. Dilution can result either from the spin-$\frac{1}{2}$ isotope being present in low natural abundance—e.g., ^{13}C—or from a high-abundance spin-$\frac{1}{2}$ nucleus being present in low concentration, as for ^{31}P in organophosphates. Any heteronuclear—e.g., ^{1}H–^{13}C—dipolar interactions present can also be removed by the application of powerful decoupling fields. With the elimination of all dipolar interactions, chemical-shift anisotropy is the dominant line-broadening interaction.

However, while magnetic dilution is an effective way of removing homonuclear dipolar interactions, it also drastically reduces NMR sensitivity such that FT techniques are essential to observe nuclei such as ^{13}C. Even with these techniques, averaging of several thousand signals is often required. To compound the problem, many of the nuclei of interest also have very long relaxation times. A technique known as cross polarization (CP) (see Pines *et al.,* 1973) is used to largely overcome these sensitivity problems. It is based on the transfer of magnetization from abundant spins, usually ^{1}H, to dilute spins such as ^{13}C. Briefly, CP is achieved in the following manner, illustrated by a ^{1}H/^{13}C system using the rotating frame of reference. A 90° pulse is first applied to the proton channel, and the ^{1}H magnetization is spin-locked in the y direction. The ^{13}C channel is then switched on, also orienting the ^{13}C spins in the y direction. The magnitudes of the proton and carbon spin-locking fields are chosen so that the following condition is fulfilled:

$$\gamma_H B_{1H} = \gamma_C B_{1C}. \tag{13}$$

Magnetization transfer takes place during a contact period after which the ^{13}C field is switched off and the ^{13}C FID recorded. A maximum enhancement equal to the ratio of the magnetogyric ratios of the abundant and rare spins may be obtained.

The shift anisotropy is a potential source of information on structure and bonding interactions. It is best studied in single-crystal systems (see Mehring, 1976) where an assignment of the principal tensor components with respect to the molecular geometry may be made. Principal shift-tensor components obtained from single crystals of model compounds can be used to assist in the interpretation of spectra from more complex polycrystalline samples. While ^{13}C has been the most commonly studied nucleus, others include ^{31}P, ^{29}Si, ^{207}Pb, and ^{77}Se.

A very simple, elegant example of the use of chemical-shift anisotropy patterns to probe solid-state structures is the use of xenon gas intercalated into solid systems such as β-quinol clathrate, reported by Ripmeester (1982). The ^{129}Xe nucleus is very sensitive to induced-shift anisotropy. The isolated atom is spherically symmetrical and shows no chemical-shift anisotropy. Any anisotropy observed therefore results from its environment, making it a sensitive probe to the interior shape of cages in clathrates.

3.2.3 Quadrupolar Nuclei Solid-state spectra of quadrupolar nuclei are normally dominated by quadrupolar interaction. In a single crystal the spectrum obtained from nuclei with $I = 1$, such as ^{2}H and ^{14}N, consists of two lines from each nucleus in each crystalline environment. Variation of crystal orientation enables quadrupolar coupling constants and asymmetry parameters of the electric field gradient to be determined. For polycrystalline materials all possible orientations are represented, and, in the case of deuterium, a powder pattern with a doublet splitting is obtained from which the quadrupolar coupling constant can be determined. The quadrupolar interaction is too large for powder patterns to be observed for other quadrupolar nuclei with integral spins, and studies have therefore been limited mainly to deuterium. Extensive ^{2}H studies (see Mantsch *et al.*, 1977) have been performed on a wide range of deuterium-labeled systems that include membranes, polymers (see Spiess, 1983), and liquid crystals. Information about structure and dynamics can be obtained in this way. While early solid-state ^{2}H spectra were recorded on cw spectrometers, more recent studies have been facilitated by the use of FT methodology and in particular the quadrupolar spin-echo technique (Sternin *et al.*, 1983). For most quadrupolar nuclei with nonintegral spins (e.g., ^{11}B, $I = \frac{3}{2}$; and ^{17}O, ^{27}Al, $I = \frac{5}{2}$) the central $M_I = \frac{1}{2}$ to $m_I = -\frac{1}{2}$ transition is usually the only one to be observed since it is much less affected than the others by quadrupolar interaction. The shape of the spectrum obtained from a polycrystalline sample depends on the values of the quadrupolar coupling constant and asymmetry parameter. ^{11}B and ^{17}O spectra of this type obtained by cw techniques have been used to investigate the structures of borosilicate glasses (see Jellison and Bray, 1977).

3.3 High-Resolution NMR of Solids

In the last 20 years techniques have been developed to obtain solution-type spectra from solids. For spin-$\frac{1}{2}$ nuclei this requires the removal of the effects of dipolar coupling and chemical shift anisotropy. As described in Sec. 3.2.2, high-power decoupling can be used to remove heteronuclear dipolar interactions.

3.3.1 Resolution-Enhancement Techniques In solution molecular tumbling averages dipolar and quadrupolar interactions to zero and chemical shifts to their isotropic values. For solids a technique known as magic-angle spinning (MAS) simulates such motion and averages out angular dependences. It involves spinning the sample at an angle of 54° 44′ to the applied magnetic field. To be effective the rate of rotation must be comparable to the frequency range of the interactions involved. The effects of chemical-shift anisotropy in ^{13}C spectra can usually be removed adequately with rates in the range of 3–10 kHz. However, for dipolar and many quadrupolar interactions the speeds required are much higher and cannot be achieved. The effects of MAS and related resolution-enhancement techniques applied to quadrupolar nuclei are further discussed in Sec. 3.3.5.

For abundant spin systems, homonuclear dipolar interactions dominate the spectra. Multiple-pulse line-narrowing techniques have been developed to average out the effects of homonuclear dipolar interactions. The simplest sequence, known as WAHUHA, was developed by Waugh *et al.* (1968) and consists of a series of four pulses.

Recent developments in high-resolution techniques have been reviewed by Clayden (1992).

3.3.2 Dilute Spin-$\frac{1}{2}$ Nuclei The combined application of CP, MAS, and high-power proton decoupling can give high-resolution spectra from powders and amorphous solids that approach the quality to those obtained in solution. Nuclei studied include ^{13}C, ^{29}Si, ^{15}N, and ^{31}P. There are two principal reasons for obtaining a high-resolution NMR spectrum from a solid. Either it is insoluble or information is being sought about the solid-state structure. In the latter case cystallographic information is required either because a single-crystal x-ray structure is not available or to enable a comparison of solid- and solution-state structures/conformations to be made when spectra are available in both phases. One example of such a comparison is the ^{13}C study of a series of penicillins by Clayden *et al.* (1986).

Crystallographic effects frequently give rise to splitting of lines in spectra of solids when compared with those obtained from the sample in solution. For a particular type of atom to yield a single line in a spectrum obtained from a solid, all the atoms of that type must be related by symmetry in the crystal. Nonequivalence is not an uncommon occurrence and may be intramolecular or intermolecular in origin. Figure 22 shows the

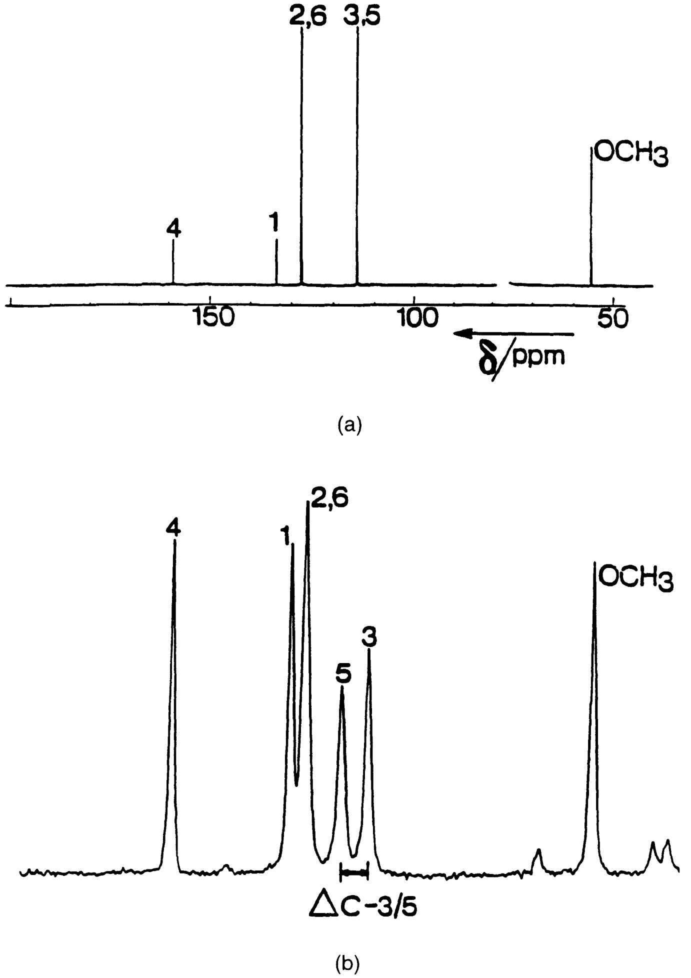

(a)

(b)

FIG. 22. ^{13}C NMR spectra of 4,4′-dimethoxybiphenyl: (a) solution, (b) solid (CP/MAS).

solution and CP/MAS solid-state ^{13}C spectra of 4,4′-dimethoxybiphenyl. In solution a single peak is observed for C-3 and C-5, whereas for the solid, two peaks separated by 6.4 ppm are present. This is attributed by Chippendale *et al.* (1986) to the fact that in the solid the compound exists in the locked conformation described by

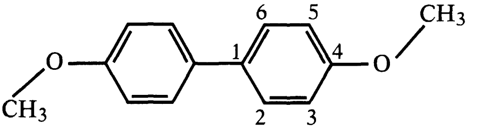

,

whereas in solution rotation about the C-4–O bond causes an average shift to be observed.

Solid-state NMR spectroscopy, particularly of ^{13}C, has been widely used to study polymorphism since polymorphs invariably give spectra that are discernably different as for the six crystalline forms of cortisone acetate reported by Harris *et al.* (1990).

^{29}Si NMR, together with ^{27}Al NMR, has been used extensively to study structures of zeolitic molecular sieves. Klinowski (1993) has recently reviewed such applications.

3.3.3 Abundant Spin-$\frac{1}{2}$ Nuclei The combination of multiple-pulse homonuclear dipolar line narrowing with MAS to average out chemical-shift anisotropy is known as CRAMPS (combined rotation and multiple-pulse spectroscopy) and is the most effective approach for obtaining high-resolution ^{1}H spectra from organic solids. Figure 23 shows a comparison of spectra obtained by Dec *et al.* (1989) from a sample of powdered adipic acid, $HO_2CCH_2(CH_2)_2CH_2CO_2H$: (a) static, (b)–(d) MAS only at the indicated rates, (e) CRAMPS, and (f) CRAMPS on an expanded scale. Even with use of CRAMPS residual linewidths are still too great to facilitate widespread application of high-resolution ^{1}H NMR of solids because of the small proton chemical-shift range. Some applications to the study of hydrogen bonding and crystallographic nonequivalence have been reported.

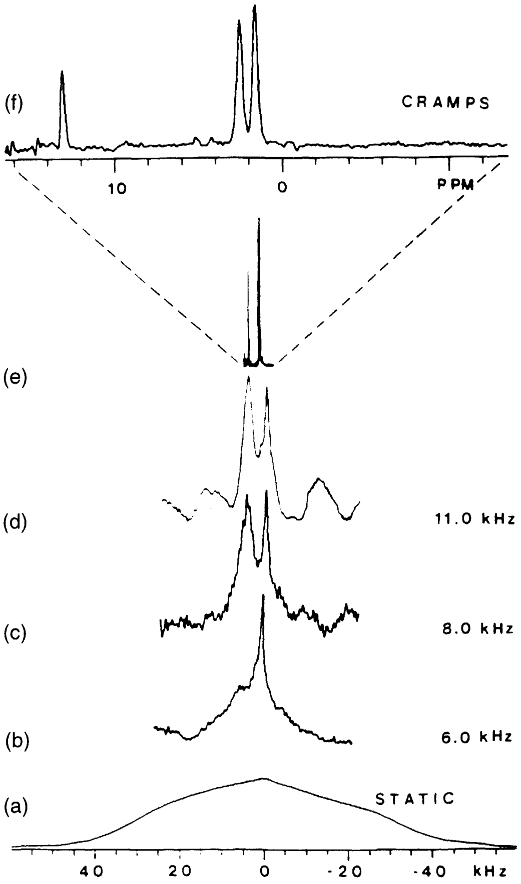

FIG. 23. ^{1}H NMR spectra of powdered, solid adipic acid, $HO_2CCH_2(CH_2)_2CH_2CO_2H$: (a) static sample, (b)–(d) MAS-only with indicated MAS speed, (e) CRAMPS, (f) CRAMPS, on an expanded scale. Reproduced with permission from Dec *et al.*, (1989).

3.3.4 Two-Dimensional Experiments with Spin-$\frac{1}{2}$ Nuclei Following the example of solution NMR, two-dimensional methodology (see Sec. 2.5) for solid-state NMR has begun to be developed in recent years. Such techniques are enabling the study of processes that include diffusion and exchange in solids as well as chemical-shift correlation and even resolution of *J* couplings in favorable cases. A few such experiments are mentioned briefly in this section. While still not in widespread use, $^{1}H/^{13}C$ 2D heteronuclear correlation experiments have been reported. In such experiments in the solid state, correlations are made through heteronuclear dipole interactions. Bronnimann *et al.* (1992) have applied the method to representative organic solids ranging from pure powdered crystalline compounds to lignite. Figure 24 shows the result obtained for monoethyl fumarate. For comparison, the 1D ^{1}H CRAMPS spectrum is displayed along the ^{1}H chemical-

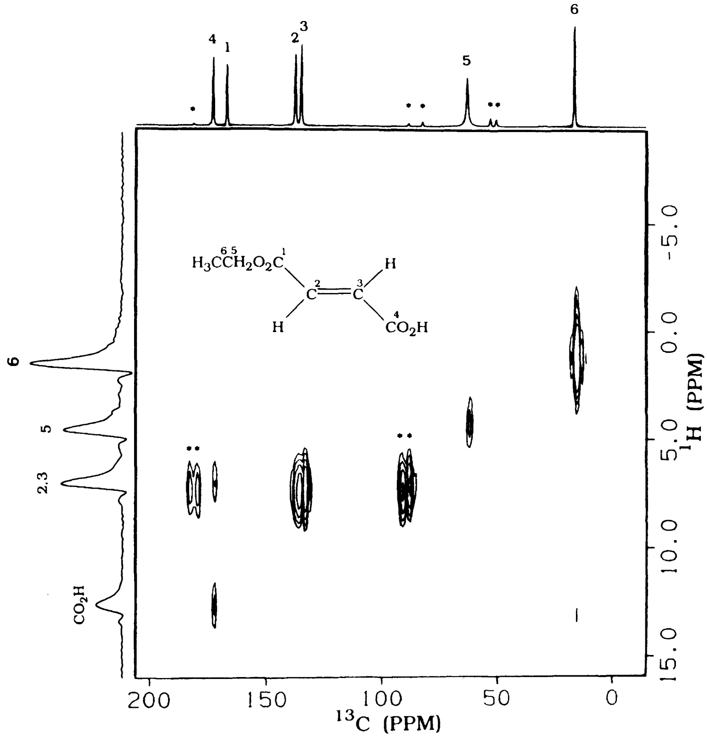

FIG. 24. The 2D HetCor contour spectrum, 1D ^{13}C CP/MAS spectrum (top), and 1D ^{1}H CRAMPS spectrum (left) of monoethylfumarate. Reproduced by permission from Bronniman *et al.*, (1992).

shift axis and the 1D ^{13}C CP/MAS spectrum along the ^{13}C chemical-shift axis.

A variety of methods has been used to achieve homonuclear correlations between dilute spin-$\frac{1}{2}$ nuclei, some through dipolar interactions and others through *J* couplings. Fyfe *et al.* (1992) have used the 2D ^{29}Si INADEQUATE experiment to obtain a complete Si–O–Si connectivity pattern for the zeolite ZSM-23.

While spin–spin coupling is rarely observed, even in high-resolution spectra, because it is usually so much smaller than other interactions, 2D *J*-resolved experiments can, in favorable cases, offer a means of significantly enhancing spectral resolution relative to 1D experiments. In a ^{31}P homonuclear experiment, Wu and Wasylischen (1992) have resolved $^2J_{PP}$ coupling constants in $RhCl(PPh)_3$.

High-resolution NMR is a valuable tool for studying exchange processes in solid samples in much the same way as it has been used in solution-state NMR for many years. 2D exchange spectroscopy (EXSY) is one of the NMR experiments used in such studies. Cross peaks link the signals of exchanging atoms, and exchange rates can also be obtained from spectra. Aliev and Harris (1993) have used 2D ^{13}C CP/EXSY to study conformational exchange properties of monosubstituted cyclohexanes, $C_6H_{11}X$, in thiourea inclusion complexes.

3.3.5 Quadrupolar Nuclei Line-narrowing techniques involving fast sample spinning can usefully be carried out on quadrupolar nuclei provided the magnitude of the quadrupolar interaction is not too large. In discussing techniques it is helpful to distinguish between quadrupolar nuclei with integral and half-integral spins. Half-integer spins have a central $m_I = \frac{1}{2}$ to $m_I = -\frac{1}{2}$ transition that is unaffected to first order by the quadrupolar interactions, whereas for integer spins all transitions are affected. In the latter case even fast MAS (~20 kHz) is of little value in significantly removing or reducing the quadrupolar contribution, except for ^{2}H where the interaction is relatively small or for a nucleus in a high-symmetry site. While several applications of MAS deuterium NMR to isotopically enriched samples have been described, reports on natural-abundance studies are now also beginning to appear, and Poupko *et al.* (1994) have reported the determination of deuterium interaction constants from such studies.

Most high-resolution studies of quadrupolar nuclei with nonintegral spins have been concerned with the observation of the central transition. While this is not affected by first-order quadrupolar interactions, there is a shift and broadening due to second-order effects. A description of the origins of these is beyond the scope of this article. The second-order interaction is inversely proportional to the static magnetic field, and resolution can be improved by working at higher magnetic fields. MAS will partially but not completely average the second-order effects, and Alemany (1993) has reviewed the factors involved in obtaining meaningful, fast MAS NMR spectra from such nuclei with particular emphasis on ^{27}Al ($I = \frac{5}{2}$) NMR of catalysts and minerals. In addition, Klinowski (1993) has described applications of the study of the same nucleus to obtain information about aluminum coordination sites in zeolitic and aluminophosphate molecular sieves. In some cases (see Ganapathy *et al.*, 1982), spinning at angles other than the magic angle gives better resolution, but in most cases lines remain broad in comparison with the results obtained for spin-$\frac{1}{2}$ nuclei.

In recent years Pines's group (see Chmelka *et al.*, 1989) have devised and applied two elegant experiments, double rotation (DOR) and dynamic-angle spinning (DAS), which average the anisotropic second-order, quadrupolar broadening and give spectra with much narrower signals than MAS. DOR involves the simultaneous rotation of the sample around two axes inclined at angles corresponding to zeros of the second and fourth Legendre polynomials, 54.74° (magic angle) and 30.6°. Experimentally, this is achieved with a large outer rotor spinning at the magic angle accommodating a smaller inner rotor containing the sample that spins at an angle of 30.6° with respect to the spinning axis of the outer rotor. Figure 25 shows a comparison of the ^{17}O NMR spectra obtained (a) by MAS, and by DOR with the outer rotor spinning at (b) 540 Hz and (c) 680 Hz, for an ^{17}O-enriched sample of diopside, $CaMgSi_2O_6$. Three isotropic center-band peaks, corresponding to different oxygen sites, are observed and labeled on the spectra.

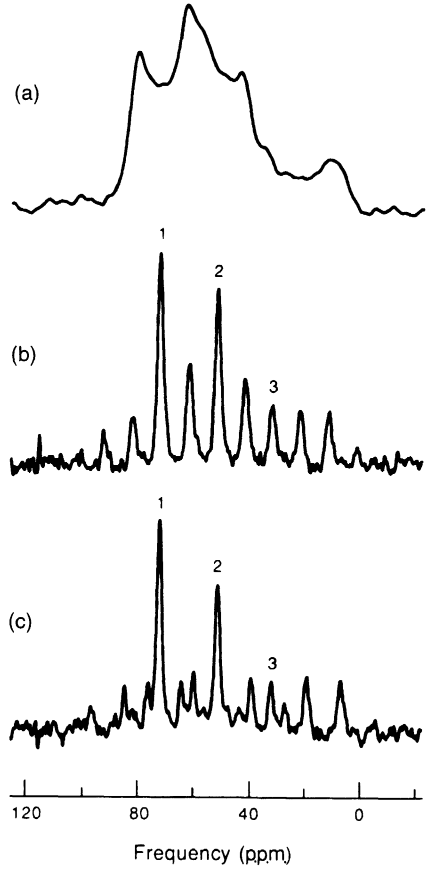

FIG. 25. ^{17}O NMR spectra of diopside, $CaMgSi_2O_6$: (a) MAS, (b), (c) DOR with outer rotor spinning at (b) 540 Hz and (c) 680 Hz. Reproduced with permission from Chmelka *et al.*, (1989).

In the DAS experiment second-order quadrupolar interactions are averaged by reorienting a sample about a time-dependent axis. Spin evolution occurs at two different angles of the sample rotor in order for the broadening from both the quadrupolar and chemical shift anisotropies to disappear. Baltisberger *et al.* (1992) have reported ^{87}Rb DAS NMR results for five inorganic rubidium salts. Narrow lines were obtained from each distinct rubidium site. The experiment was performed at four magnetic-field strengths for $RbNO_3$, which enabled the isotropic chemical shift and a product of the quadrupolar coupling constant and the quadrupolar asymmetry parameter to be calculated. A two-dimensional DAS experiment is also described in the same article that involved three orientations of the rotor relative to the magnetic field. In this experiment, the isotropic signals from the different rubidium sites in $RbNO_3$ were resolved in one dimension and the powder pattern cross sections in the other. Simulation of the latter for each site enabled the quadrupolar coupling constant and asymmetry parameter to be obtained for the various sites.

GLOSSARY

Chemical Shift: The displacement of an NMR signal relative to that of a reference material. Chemical shifts are expressed on a field-independent scale in ppm, by the displacement in hertz divided by the applied magnetic field in megahertz.

Cross Polarization: A process that enhances the magnetization of rare spins in a magnetically dilute spin system by transferring magnetization from abundant nuclei such as ^1H and hence increases the NMR signal.

Dipole–Dipole Coupling: Direct interaction between the magnetic moments of two nuclei.

Fourier Transformation: The mathematical operation that converts an FID time-dependent signal into the frequency domain.

Free-Induction Decay (FID): The decaying signal that is recorded as a function of time at the end of a pulse experiment.

Magic-Angle Spinning: Rotating a sample at an angle of 54° 44′ to the applied magnetic field.

Magnetogyric Ratio: The proportionality constant relating the magnetic moment of a nucleus to the spin I of the nucleus.

Nuclear Overhauser Enhancement (NOE): The change in the intensity of the NMR signal from a nucleus A caused by irradiation at the resonance frequency of a second nucleus B. NOEs can be either homonuclear or heteronuclear in origin.

Powder Pattern: An envelope of NMR responses from all possible orientations to the applied magnetic field of a spin system in a solid sample.

Quadrupole Interaction: For nuclei with spin $>\frac{1}{2}$, interaction between the nuclear electric quadrupole moment and the non-spherically-symmetrical field gradient around the nucleus.

Relaxation Time: The time taken by spins to relax to thermal equilibrium after radio-frequency excitation. The exponential decay of magnetization back along the x axis is characterized by the spin–lattice relaxation time T_1. The spin–spin relaxation time T_2 characterizes the process by which spins fan out in the xy plane.

Spin–Spin Coupling: Indirect coupling between nuclear spins transmitted by electrons and characterized by a coupling J.

Works Cited

Abraham, R. J., Fisher, J., Loftus, P. (1988), *Introduction to NMR Spectroscopy,* Chichester: Wiley.

Akitt, J. W. (1983), *NMR and Chemistry: An Introduction to the Multinuclear Fourier Transform Era,* London: Chapman and Hall.

Alemany, L. B. (1993), *Appl. Magn. Reson.* **4,** 179–201.

Aliev, A. E., Harris, K. D. M. (1993), *J. Am. Chem. Soc.* **115,** 6369–6377.

Baltisberger, J. H., Gann S. L., Wooten, E. W., Chang, T. H., Mueller, K. T., Pines, A. (1992), *J. Am. Chem. Soc.* **114,** 7489–7493.

Bax, A., Grzesiek, S. (1993), *Acc. Chem. Res.* **26,** 131–138.

Bax, A., Subramanian, S. (1986), *J. Magn. Reson.* **67,** 565–569.

Bax, A., Summers, M. F. (1986), *J. Am. Chem. Soc.* **108,** 2093–2094.

Benn, R., Gunther, H. (1983), *Angew. Chem. Int. Ed. Engl.* **22,** 350–380.

Breitmaier, E., Voelter, W. (1987), *Carbon-13 NMR Spectroscopy,* 3rd ed., New York: VCH.

Bronnimann, C. E., Ridenour, C. F., Kinney, D. R., Maciel, G. E. (1992), *J. Magn. Reson.* **97,** 522–534.

Brugel, W. (1979), *Handbook of NMR Spectral Parameters,* Vols. 1-3, Philadelphia: Heyden.

Chmelka, B. F., Mueller, K. T., Pines, A., Stebbins, J., Wu, Y., Zwanziger, J. W. (1989), *Nature* **339,** 42–43.

Chippendale, A. M., Aujla, R. S., Harris, R. K., Packer K. J., Purser, S. (1986), *Magn. Reson. Chem.* **24,** 81–88.

Clayden, N. J., Dobson, C. M., Lian, L.-Y., Twyman, J. M. (1986), *J. Chem. Soc. Perkin Trans. II,* 1933–1940.

Clayden, N. J. (1992), in: G. A. Webb (Ed.), *Annual Reports on NMR Spectroscopy,* Vol. 24, New York: Academic, p. 1.

CSEARCH, Philadelphia: BIO-RAD/Sadtler.

Dec, S. F., Bronniman, C. E., Wind, R. A., Maciel, G. E. (1989), *J. Magn. Reson.* **82,** 454–466.

Derome, A. E. (1987), *Modern NMR Techniques for Chemistry Research,* Oxford: Pergamon Press.

Doddrell, D. M., Pegg, D. T., Bendall, M. R. (1982), *J. Magn. Reson.* **48,** 323–327.

Emsley, J. W., Feeney, J., Sutcliffe, L. H. (1965), *High Resolution Nuclear Magnetic Resonance Spectroscopy* (2 vols), Oxford: Pergamon Press.

Ewing, D. E. (1979), *Org. Magn. Reson.* **12,** 499–524.

Farrar, T. C. (1987), *An Introduction to Pulse NMR Spectroscopy,* Chicago: Farragut Press.

Fyfe, C. A. (1983), *Solid State NMR for Chemists,* Guelph: CFC Press.

Fyfe, C. A., Grondey, H., Feng, Y., Kokotailo, G. T., Ernst, S., Weitkamp, J. (1992), *Zeolites* **12,** 50–53.

Ganapathy, S., Schramm, S., Oldfield, E. (1982), *J. Chem. Phys.* **77,** 4360–4365.

Harris, R. K., Mann, B. E. (1978), *NMR and the Periodic Table,* London–New York–San Francisco: Academic.

Harris, R. K. (1986), *Nuclear Magnetic Resonance Spectroscopy,* Harlow: Longman Scientific & Technical.

Harris, R. K., Kenwright, A. M., Say, B. J., Yeung, R. R., Fletton, R. A., Lancaster, R. W., Hardgrove, G. L. (1990), *Spectrochim. Acta* **46A,** 927–935.

Jellison, G. E., Bray, P. J. (1977) in: L. D. Pye *et al.* (Eds.), *Borate Glasses: Structure, Properties, Applications,* New York: Plenum Press.

Klinowski, J. (1993), *Anal. Chim. Acta* **283,** 929–965.

LeCocq, C., Lallemand, J. Y. (1981), *J. Chem. Soc. Chem. Commun.* 150–152.

Lindeman, L. P., Adams, J. Q. (1971), *Anal. Chem.* **43,** 1245–1252.

Mantsch, H., Saito, H., Smith, I. C. P. (1977), in: J. W. Emsley *et al.* (Eds.), *Progress in Nuclear Magnetic Resonance Spectroscopy,* Vol. 11, Oxford: Pergamon Press, p. 211.

Mason, J. (1987), *Multinuclear NMR,* New York: Plenum.

Mehring, M. (1976), in: P. Diehl *et al.* (Eds.), *NMR—Basic Principles and Progress,* Vol. 11, New York: Springer-Verlag.

Morris, G. A. (1986), *Magn. Reson. Chem.* **24,** 371–403.

Neuhaus, D., Williamson, M. P. (1989), *The Nuclear Overhauser Effect in Structural and Conformational Analysis,* New York: VCH.

Nicholson, J. K., Wilson I. D. (1989), in: J. W. Emsley *et al.* (Eds.), *Progress in Nuclear Magnetic Resonance Spectroscopy,* Vol. 21, Oxford: Pergamon Press, pp. 449-501.

Noggle, J. H., Schirmer, R. E. (1971), *The Nuclear Overhauser Effect—Chemical Applications,* New York–London: Academic.

Norwood, T. J. (1994), *Chem. Soc. Rev.* **23,** 59–66.

Oschkinat, H., Müller, T., Dieckmann, T. (1994), *Angew. Chem. Int. Ed. Engl.* **1994,** 277–293.

Pake, G. E. (1948), *J. Chem. Phys.* **16,** 327–336.

Parker, D. (1991), *Chem. Rev.* **91,** 1441–1457.

Pines, A., Gibby, M. G., Waugh, J. S. (1973), *J. Chem. Phys.* **59,** 569–590.

Pouchert, C. J. (1983), *Aldrich Library of NMR Spectra,* Vols. 1 and 2, Milwaukee, WI: Aldrich.

Poupko, R., Olender, Z., Reichert, D., Luz, Z. (1994), *J. Magn. Reson. Ser. A* **106,** 113–115.

Ripmeester, J. A. (1982), *J. Am. Chem. Soc.* **104,** 289–290.

Sanders, J. K. M., Hunter, B. K. (1987), *Modern NMR Spectroscopy,* Oxford: Oxford Univ. Press.

Shaka, A. J., Keeler, J., Freeman, R. (1983), *J. Magn. Reson.* **53,** 313–340.

Shaw, D. (1984), *Fourier Transform NMR Spectroscopy,* 2nd ed., Amsterdam: Elsevier.

Silverstein, R. M., Bassler, G. C., Morrill, T. C. (1991), *Spectrometric Identification of Organic Compounds,* 5th ed., New York: Wiley.

SPECINFO, Weinheim: Chemical Concepts.

Spiess, H. W., Schweitzer, D., Haeberlen, U., Hausser, K., (1971), *J. Magn. Reson.* **5,** 101–108.

Spiess, H. W. (1983), *Colloid Polym. Sci.* **261,** 193–209.

Sternin, E., Bloom, M., Mackay, A. L. (1983), *J. Magn. Reson.* **55,** 274–282.

Stothers, J. B. (1972), *Carbon-13 NMR Spectroscopy,* New York: Academic.

Waugh, J. S., Huber, L. M., Haeberlen, U. (1968), *Phys. Rev. Lett.* **20,** 180–182.

Wehrli, F. W., Marchand, A. P., Wehrli, S. (1988), *Interpretation of Carbon-13 NMR Spectra,* 2nd ed., New York: Wiley.

Williams, E. A., Cargioli, J. D. (1979), in: G. A. Webb (Ed.), *Annual Reports on NMR Spectroscopy,* Vol. 9, New York: Academic, p. 221.

Witanowski, M., Stefaniak, L., Webb, G. A. (1986), in: G. A. Webb (Ed.), *Annual Reports on NMR Spectroscopy,* Vol. 18, New York: Academic, p. 1.

Wray, V. (1980), in: G. A. Webb (Ed.), *Annual Reports on NMR Spectroscopy,* Vol. 10B, New York: Academic, p. 1.

Wray, V. (1983), in: G. A. Webb (Ed.), *Annual Reports on NMR Spectroscopy,* Vol. 14, New York: Academic, p. 1.

Wüthrich, K. (1986), *NMR of Proteins and Nucleic Acids,* 1st ed., New York: Wiley.

Wu, G., Wasylishen, R. E. (1992), *Inorg. Chem.* **31,** 145–148.

Further Reading

Abraham, R. J., Fisher, J., Loftus, P. (1988), *Introduction to NMR Spectroscopy,* Chichester: Wiley.

Breitmaier, E., Voelter, W. (1987), *Carbon-13 NMR Spectroscopy,* 3rd ed., New York: VCH.

Derome, A. E. (1987), *Modern NMR Techniques for Chemistry Research,* Oxford: Pergamon Press.

Fyfe, C. A. (1983), *Solid State NMR for Chemists,* Guelph: CFC Press.

Gorenstein, D. G. (1984), *Phosphorus-31 NMR: Principles and Applications,* New York: Academic.

Harris, R. K. (1986), *Nuclear Magnetic Resonance Spectroscopy,* Harlow: Longman Scientific & Technical.

Koenig, J. L. (1992), *Spectroscopy of Polymers,* Washington, D.C.: American Chemical Society.

Levy, G. C., Lichter, R. L. (1979), *Nitrogen-15 NMR Spectroscopy,* New York: Wiley.

Mason, J. (1987), *Multinuclear NMR,* New York: Plenum Publishing Co.

Mehring, M. (1983), *Principles of High Resolution NMR in Solids,* Berlin: Springer-Verlag.

Mooney, E. F. (1970), *An Introduction to ^{19}F NMR Spectroscopy,* London: Heyden-Sadtler.

Neuhaus, D., Williamson, M. P. (1989), *The Nuclear Overhauser Effect in Structural and Conformational Analysis,* New York: VCH.

Sanders, J. K. M., Hunter, B. K. (1987), *Modern NMR Spectroscopy,* Oxford: Oxford Univ. Press.

Silverstein, R. M., Bassler, G. C., Morrill, T. C. (1991), *Spectrometric Identification of Organic Compounds,* 5th ed., New York: Wiley.

Wehrli, F. W., Marchand, A. P., Wehrli, S. (1988), *Interpretation of Carbon-13 NMR Spectra,* 2nd ed., New York: Wiley.

Wüthrich, K. (1986), *NMR of Proteins and Nucleic Acids,* 1st ed., New York: Wiley.

STRUCTURE OF CERAMICS

Chan Park and Robert L. Snyder, *New York State College of Ceramics at Alfred University, Alfred, New York, U.S.A.*

Introduction 159
1. **Bonding** 160
2. **Crystal Structures** 160
2.1 AX 162
2.1.1 Rock-Salt Structure ($Fm3m$) .. 162
2.1.2 Nickel Arsenide Structure ($P6_3/mmc$) 162
2.1.3 Wurtzite Structure ($P6_3mc$) 162
2.1.4 Zinc Blende Structure ($F43m$) 162
2.1.5 Cesium Chloride Structure ($Pm3m$) 163
2.2 AX_2 163
2.2.1 Fluorite Structure ($Fm3m$) 163
2.2.2 Rutile Structure ($P4_2/mnm$) 163
2.2.3 Cadmium Iodide Structure ($P\bar{3}m1$) 163
2.2.4 Cadmium Chloride Structure ($R\bar{3}m$) 164
2.2.5 Structures of Silica 164
2.3 A_2X 165
2.3.1 Antifluorite Structure ($Fm3m$) 165
2.4 A_2X_3 165
2.4.1 Corundum Structure ($R\bar{3}c$) 165
2.5 AB_2X_4 165
2.5.1 Spinel Structure ($Fd3m$) 165
2.5.2 Inverse Spinel Structure 165
2.6 ABX_3 166
2.6.1 Perovskite Structure ($Pm3m$) 166
2.6.2 Ilmenite Structure ($R\bar{3}$) 166
2.6.3 Lithium Niobate Structure ($R3c$) 166
2.7 A_2BX_4 167
2.7.1 Olivine Structure ($Pbnm$) 167
2.8 Others 167
2.8.1 Graphite ($P6_3/mmc$) 167
2.8.2 Diamond ($Fd3m$) 167
2.8.3 High-Temperature Superconductors 167
3. **Defect Structures** 167
3.1 Defect Structures in Stoichiometric Oxides 168
3.2 Defect Structures in Nonstoichiometric Oxides 168
3.2.1 Oxygen-Deficient Oxides 168
3.2.2 Metal-Deficient Oxides 168
4. **Microstructure** 169
5. **Structures and Mechanical Behavior** 170
5.1 Brittleness 170
5.2 Toughening Mechanisms 171
5.2.1 Transformation Toughening ... 171
5.2.2 Microcracking 172
5.2.3 Whisker or Fiber Reinforcement 172
5.2.4 Ductile Dispersion Toughening 172
6. **Structures of Electroceramics** 173
6.1 $BaTiO_3$ 173
6.2 ZnO Varistor 174
7. **Structures of Ferrites** 175
7.1 Spinel Ferrites 175
7.2 Magnetoplumbite 176
7.3 Microstructures of Ferrites 176
Glossary 177
Works Cited 177
Further Reading 178

INTRODUCTION

Ceramics are any of a class of inorganic, nonmetallic products that have been subjected to a high temperature during manufacture or use. Typical ceramics are metallic oxides, although borides, carbides, nitrides, silicides, covalent compounds of nonmetals (e.g., B, C, Si), and mixtures or composites of such materials are also ceramics.

The fundamental principle of materials science is that the properties of a material originate from its structure. The structure of a material may be considered on many dif-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ferent levels: microstructure, defect structure, atomic structure, electronic structure, and nuclear structure, each of which may influence the physical and chemical properties of the material.

At the most fundamental level, the atoms of the various elements differ in the amount of charge on their nuclei and the number of electrons they attract to neutralize this charge. The nature of the electron configuration that shields the nuclear charge dictates the type of chemical bonding (see CRYSTALLINE STATE) in which an atom may engage and directs such fundamental properties as thermal, optical, electrical, and magnetic behavior. The nature of the chemical bonding in turn determines the type of material that may form (metal, ceramic, or polymer).

Atoms and molecules arrange themselves in space in patterns that will minimize the electrostatic energy. These arrangements (crystal structure) may be periodic (crystals) or nonperiodic (amorphous or quasicrystalline materials). Disturbances of the strict periodicity of the crystal arise from the presence of various kinds of defects. Multiple-element materials may also exhibit marked deviations from stoichiometry. The defect structure can frequently play an important role in the properties of ceramics.

The microstructure of a material describes the way crystallites pack together and involves the grains, grain boundaries, size and shape of grains, grain orientation, and porosity. Microstructure also determines many properties, e.g., mechanical, electrical, optical, and magnetic properties. Most of the time, even when the property itself originates from the electronic, crystal, or defect structure, the property of the final ceramic material is more closely related to the control of microstructure. When more than two phases are present, the multiple-phase structure, which includes the type, size, distribution, connectivity, and amount of the phases, is also a part of the microstructure.

In this article, structures of ceramics are described in terms of crystal structure, defect structure, and microstructure. The relations of structures and properties are also discussed. For mechanical behavior, the brittle nature of the ceramics and the toughening mechanisms are mentioned. The role of structures in electrical and magnetic properties are discussed in some practically important ceramic systems.

1. BONDING

Bonding in ceramics spans the range from nearly completely ionic (LiF) to purely covalent (C). Metallic and van der Waals bonds may be present but rarely predominate. The tendency toward ionic bonding between atoms increases as the difference in the electronegativity of the constituent atoms increases. Materials comprising a single group-IV element such as carbon or silicon, with zero electronegativity difference, can thus be associated with pure covalent bonding. The fractions of ionic and covalent characters for simple ceramic compounds are given in Table 1.

2. CRYSTAL STRUCTURES

The ways in which individual atoms in ceramic compounds are arranged geometrically depends on several factors, including the type of bonding, the relative sizes of the atoms, and the need to balance electrostatic charges.

In covalent crystals, each atom is surrounded by a specific number of other elements arranged according to the bond directionality. In purely ionic crystals, the bond is not directional, and the relative atomic sizes and charge-balance factors become of increasing importance. In most oxides, the oxygen ions form close-packed arrays, and the metallic ions, which are usually smaller, are arranged in the interstices (Fig. 1). Thus the larger the relative difference in the radius of

Table 1. Fractions of ionic and covalent bondings for simple compounds (Davidge, 1979).

Compound	Electronegativity difference	Fraction ionic bonding	Fraction covalent bonding
LiF	3.0	0.89	0.11
MgO	2.3	0.73	0.27
Al_2O_3	2.0	0.63	0.37
SiO_2	1.7	0.51	0.49
Si_3N_4	1.2	0.30	0.70
SiC	0.7	0.11	0.89
Si	0	0.00	1.00

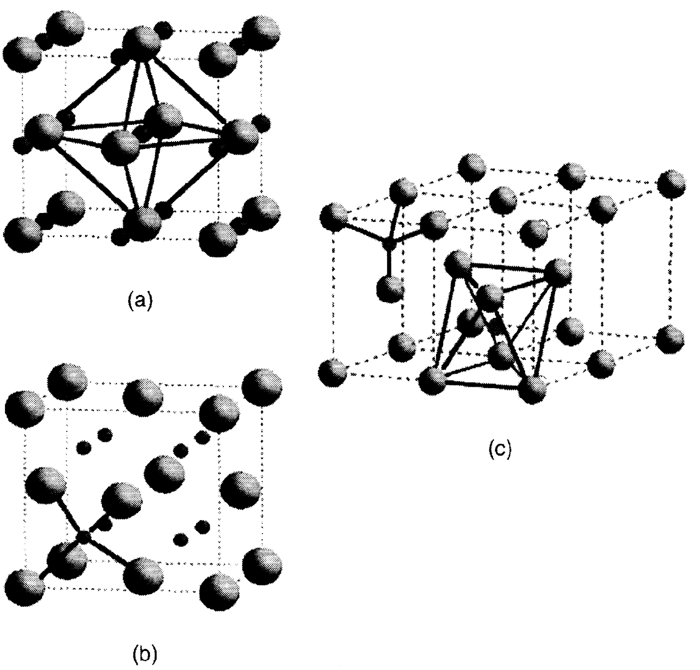

FIG. 1. (a), (c) Octahedral and (b), (c) tetrahedral interstices in (a), (b) cubic close-packed and (c) hexagonal close-packed lattice. Large spheres are anions.

the ions, the smaller the expected coordination number. A wide range of ceramic structures can be classified in terms of anion packing, the pattern of cation sites, or coordination polyhedra and their linkage. They will be categorized according to cation–anion (A:X or B:X) ratios and will be grouped as AX, AX_2, A_2X, A_2X_3, ABX_3, etc. (Table 2). The names of structure types are usually (but not always) taken from the simplest representative of the family. Some are taken from the first member of the family that was carefully

Table 2. Summary of ceramic crystal structures (after ASM Handbook, 1991; Megaw, 1973).

Name	Anion packing	Sites by cations[a]	CNs[b]
Rock salt	Cubic close-packed	All oct.	6:6 AX
Cesium chloride	Simple cubic	All cubic	8:8 AX
Zinc blende	Cubic close-packed	$\frac{1}{2}$ tet.	4:4 AX
Wurtzite	Hexagonal close-packed	$\frac{1}{2}$ tet.	4:4 AX
Nickel arsenide	Hexagonal close-packed	All oct.	6:6 AX
Rutile	Tetragonal close-packed	$\frac{1}{2}$ oct.	6:3 AX_2
Fluorite	Simple cubic	$\frac{1}{2}$ cubic	8:4 AX_2
Antifluorite	Cubic close-packed	All tet.	4:8 A_2X
Silica	Connected tetrahedra		4:2 AX_2
Spinel	Cubic close-packed	$\frac{1}{8}$ tet. (A), $\frac{1}{2}$ oct. (B)	4:6:4 AB_2O_4
Inverse spinel	Cubic close-packed	$\frac{1}{8}$ tet. (B), $\frac{1}{2}$ oct. (A,B)	4:6:4 $B(AB)O_4$
Corundum	Hexagonal close-packed	$\frac{2}{3}$ oct.	6:4 A_2X_3
Ilmenite	Hexagonal close-packed	$\frac{2}{3}$ oct.	6:6:4 ABX_3
Lithium niobate	Hexagonal close-packed	$\frac{2}{3}$ oct.	6:6:4 ABX_3
Perovskite	Cubic close-packed (A,X)	$\frac{1}{4}$ oct. (B)	12:6:6 ABX_3
Olivine	Hexagonal close-packed	$\frac{1}{2}$ oct. (A), $\frac{1}{8}$ tet. (B)	6:4:4 A_2BO_4

---

[a] Tet.: tetrahedral sites; oct.: octahedral sites.
[b] CNs: coordination numbers ("6:3 AX_2" means the CNs of A and X are 6 and 3, respectively).

studied. For more detailed information about the space groups provided for each structure type, see CRYSTALLOGRAPHY.

2.1 AX

2.1.1 Rock-Salt Structure (*Fm*3*m*) This structure consists of a cubic close-packed anion lattice in which all octahedral interstitial sites are filled with cations [Fig. 2(a)]. Both anions and cations are octahedrally coordinated. Ionically bonded binary compounds with radius ratios between 0.73 and 0.41 favor octahedral coordination and crystallize in this structure. Compounds exhibiting this structure include most alkali halides (except CsCl, CsBr, and CsI); oxides, sulfides, selenides, and tellurides of Mg, Ca, Sr, Ba, etc.

2.1.2 Nickel Arsenide Structure ($P6_3/mmc$) Binary compounds with radius ratios between 0.73 and 0.41 possessing some degree of covalent bonding adopt the nickel arsenide structure. This structure consists of a hexagonal close-packed lattice of anions with cations filling all octahedral interstices [Fig. 2(b)]. Cations are octahedrally coordinated, and each anion is surrounded by six cations at the corners of a trigonal prism. Compounds exhibiting this structure include NiAs, FeS, NiS, VS, CoS, FeSe, CoSe, etc.

2.1.3 Wurtzite Structure ($P6_3mc$) Binary compounds with radius ratios between 0.22 and 0.41 and a strong covalent contribution to bonding favor tetrahedral coordination for both constituents. These requirements can be met with hexagonal packing of the large oxygen ions with half the tetrahedral interstices filled with cations [Fig. 2(c)]. Compounds exhibiting this structure include ZnS, CdS, ZnO, BeO, SiC, AlN, GaN, InN, etc.

2.1.4 Zinc Blende Structure (*F*43*m*) Binary compounds with radius ratios between 0.22 and 0.41 and a strong covalent contribution to bonding can also be stable in

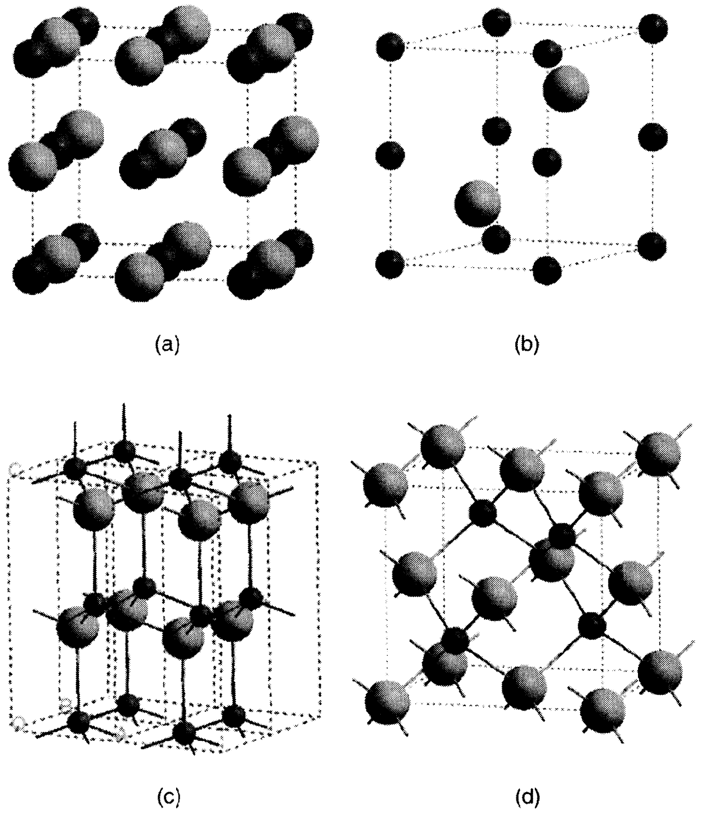
(a) (b) (c) (d)

FIG. 2. (a) Rock-salt, (b) nickel arsenide, (c) wurtzite, and (d) zinc blende structures.

the zinc blende structure. This structure [Fig. 2(d)] is similar to the diamond structure, except that each atom is tetrahedrally coordinated by the other element. Cations fill half the tetrahedral interstices of the cubic close-packed anions. Compounds exhibiting this structure include ZnS, BeS, CdS, GaAs, BeO, SiC, InSb, etc.

2.1.5 Cesium Chloride Structure (*Pm*3*m*) For binary compounds with radius ratios greater than 0.73, stability requires that the atom should be eight-fold coordinated, which is accomplished in a simple cubic lattice of one element with the large interstice at the body-centered position filled with the other element [Fig. 3(a)]. Compounds exhibiting this structure include chlorides, bromides, and iodides of Cs and Tl.

2.2 *AX*$_2$

2.2.1 Fluorite Structure (*Fm*3*m*) For compounds with radius ratios greater than 0.73, eightfold coordination is favored. This is realized in the fluorite structure, which consists of a simple cubic anion lattice with half of the body-centered positions filled with cations [Fig. 3(b)]. This structure is the same as CsCl with the exception that only half of the cation sites are occupied. Compounds exhibiting this structure include ZrO_2, HfO_2, CeO_2, ThO_2, UO_2, PrO_2, CaF_2, SrF_2, BaF_2, ZrF_2, HfF_2, HgF_2, SrF_2, CdF_2, PbF_2, BaF_2, $BaCl_2$, etc.

2.2.2 Rutile Structure (*P*4$_2$/*mnm*) *AX*$_2$ compounds with radius ratios between 0.41 and 0.73 can have the rutile structure, in which cations are octahedrally coordinated by anions, and each anion is bonded to three cations (Fig. 4). Cations fill half the available octahedral sites, and the closer packing of oxygen ions around the filled cation sites leads to the distortion of the nearly close-packed anion lattice. Compounds exhibiting this structure include TiO_2 (rutile), SnO_2, MnO_2, GeO_2, PbO_2, VO_2, NbO_2, WO_2, CrO_2, MoO_2, CoO_2, RuO_2, IrO_2, OsO_2, MgF_2, ZnF_2, MnF_2, FeF_2, NiF_2, $CaBr_2$, etc.

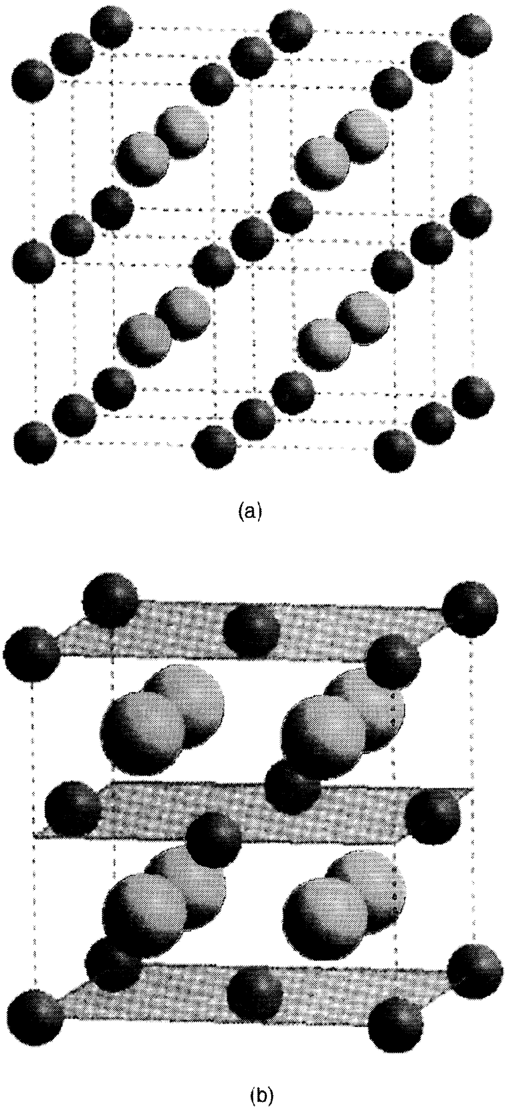

FIG. 3. (a) Cesium chloride structure (eight unit cells) and (b) fluorite structure (one unit cell). All the body-centered positions are occupied in cesium chloride, while in fluorite only half are occupied.

2.2.3 Cadmium Iodide Structure (*P*$\bar{3}$*m*1) This structure is a layer structure consisting of planes of cations sandwiched between planes of anions. The sandwich is the repeat unit in the *z* direction. Each cation is surrounded by six anions forming an octahedron. Each anion also has six neighbors, three cations on one side of it and three anions on the other. Anions form a hexagonally packed array. The octahedral sites between pairs of anion planes are occupied by cations. Compounds exhibiting this structure include CdI_2, MgI_2, CaI_2, PbI_2, $MgBr_2$, $FeBr_2$, $Mg(OH)_2$, $Ca(OH)_2$, etc.

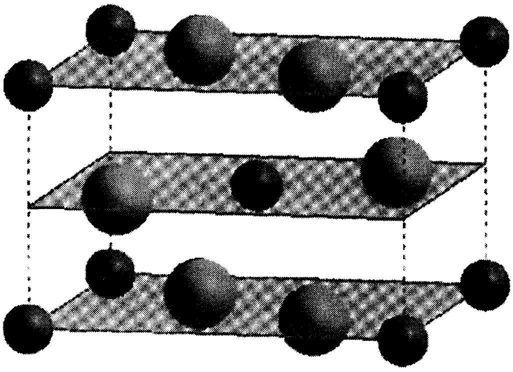

FIG. 4. A unit cell of rutile structure.

2.2.4 Cadmium Chloride Structure ($R\bar{3}m$)

This is also a layer structure like the CdI_2 structure. The layers are the same in both structures, but they pack differently. The cation environment is the same as in the CdI_2 structure. However, the anion still has three cation neighbors on one side and three anions on the other, the six taken together constituting a trigonal prism. Compounds exhibiting this structure include $CdCl_2$, $MgCl_2$, $FeCl_2$, ZnI_2, etc.

2.2.5 Structures of Silica

Crystalline silica (SiO_2) exists in several different polymorphic forms corresponding to different ways of combining corner-sharing SiO_4 tetrahedra. There are three basic structures: quartz, tridymite, and cristobalite, all of which exist in high- and low-temperature modifications.

The ideal cristobalite structure ($Fd3m$) can be described in terms of the zinc blende lattice. The A atoms occupy both the A and X sites of the zinc blende structure with one X atom midway between each pair of A atoms [Fig. 5(a)]. Each A atom is surrounded by four X atoms that form a tetrahedron. Each X atom is coordinated by two A atoms.

The ideal tridymite structure ($P6_3/mmc$) is related to wurtzite in the same way that cristobalite is to zinc blende. If all A and X positions in wurtzite are replaced by A atoms, and X atoms are introduced midway between the A atoms, it is the structure of tridymite. The A atoms are again tetrahedrally coordinated, while the X atoms are surrounded by two A atoms.

In quartz ($P6_222$ or $P3_221$), the atoms are coordinated in the same manner, and the silica tetrahedra are arranged in a hexagonal

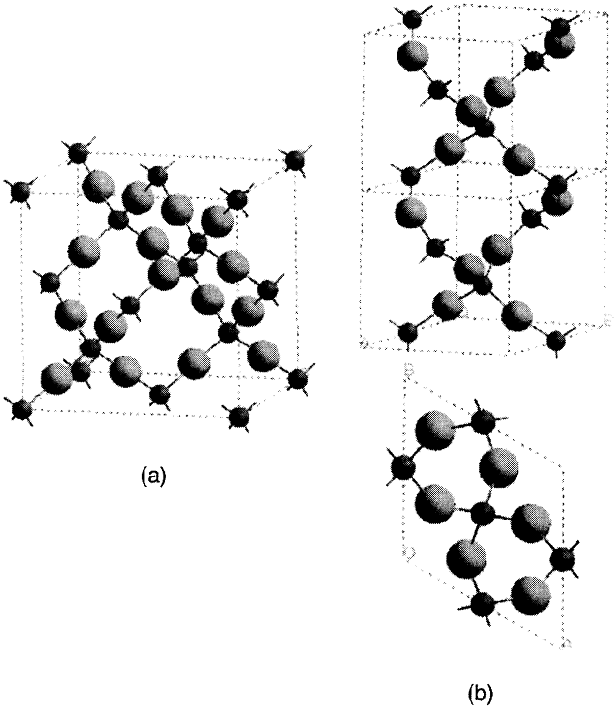

FIG. 5. The structures of (a) cristobalite and (b) quartz. Both are high-temperature forms. A projection on the xy plane of quartz is also included.

array [Fig. 5(b)]. For more detailed information on silica, see SILICON DIOXIDE; QUARTZ.

2.3 A_2X

2.3.1 Antifluorite Structure (*Fm*3*m*) This structure consists of a cubic close-packed lattice of anions with cations arranged in the tetrahedral sites. This structure has cation and anion positions exactly reversed from the normal fluorite lattice. Compounds exhibiting this structure include Li_2O, Na_2O, K_2O, Rb_2O, and various sulfides.

2.4 A_2X_3

2.4.1 Corundum Structure ($R\bar{3}c$) This structure consists of a nearly hexagonal close-packed anion lattice with cations filling two-thirds of the octahedral sites [Fig. 6(a)]. The cations are in approximate octahedral coordination. When cations fill adjacent octahedra, they repel each other so that neither is exactly centered on its site. The anion positions are also slightly distorted. Compounds exhibiting the corundum structure include α-Al_2O_3 (corundum), Cr_2O_3, Fe_2O_3, V_2O_3, Ti_2O_3, Rh_2O_3, Ga_2O_3, etc.

2.5 AB_2X_4

2.5.1 Spinel Structure (*Fd*3*m*) The spinel structure consists of a cubic close-packed anion lattice with divalent metal cations occupying one-eighth of the tetrahedral sites and trivalent metal cations occupying one-half of the octahedral sites (Fig. 7). The cubic unit cell is large, comprising eight formula units and containing 64 tetrahedral (*A*) and 32 octahedral (*B*) sites; eight of the *A* sites and 16 of the *B* sites are occupied. Compounds exhibiting the spinel structure include $MgAl_2O_4$ (spinel), $FeAl_2O_4$, $ZnAl_2O_4$, $MnFe_2O_4$, $ZnFe_2O_4$, $CdFe_2O_4$, $TiMg_2O_4$, $MoNa_2O_4$, $NiLi_2F_4$, $CuTi_2S_4$, Mn_3O_4, Co_3O_4, etc.

2.5.2 Inverse Spinel Structure The divalent ions switch sites with half of the trivalent ions in the normal spinel structure. As a consequence, the A^{2+} ions and half the B^{3+} ions are on the octahedral (*B*) sites, while the other half of the B^{3+} are on tetrahedral (*A*) sites [B(AB)O_4]. Compounds exhibiting inverse spinel structure include $FeMgFeO_4$, $FeNiFeO_4$, $FeCoFeO_4$, $FeCuFeO_4$, $FeTiFeO_4$, $InFeInO_4$, $MgTiMgO_4$, $ZnSnZnO_4$, $ZnTiZnO_4$, $CoSnCoO_4$, $GaMgGaO_4$, $InMgInO_4$, $InCoInO_4$, $InNiInO_4$, Fe_3O_4, etc.

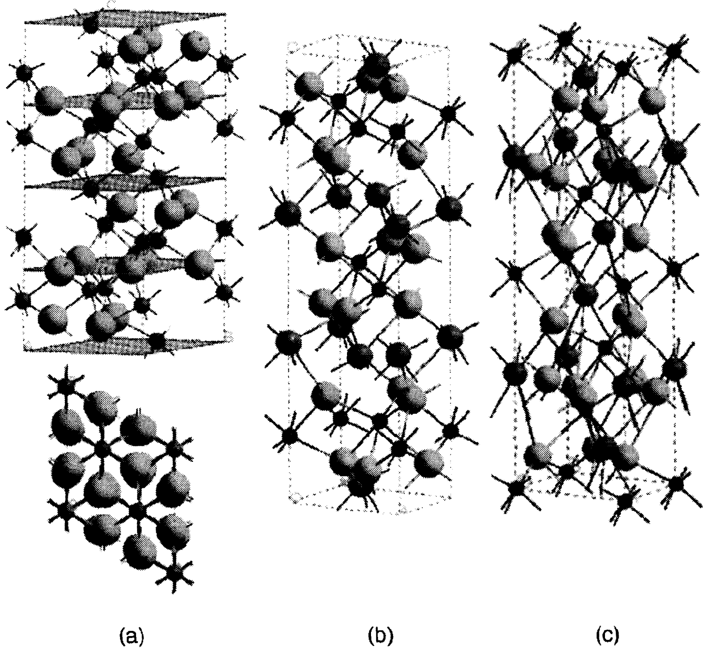

FIG. 6. (a) Corundum, (b) ilmenite, and (c) lithium-niobate structure.

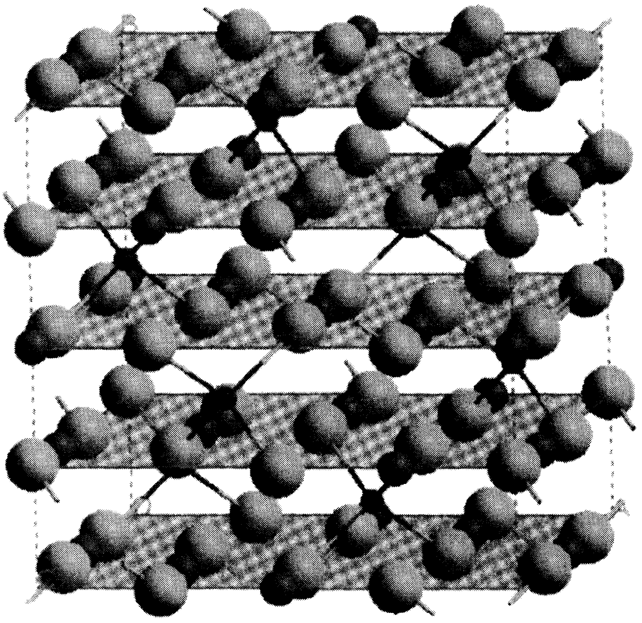

FIG. 7. Spinel structure. The four-fold coordination of the cations in *A* sites are indicated with lines.

2.6 ABX_3

2.6.1 Perovskite Structure (*Pm*3*m*) This structure consists of a cubic close-packed anion lattice in which one-fourth of the anions are replaced by a large cation (*A*), and a smaller highly charged cation (*B*) occupies one-fourth of the octahedral interstices (Plate 1). Alternatively, it can be described as a simple cubic unit cell with a large cation on the corners, a smaller cation in the body center, and oxygens in the centers of the faces. The structure is a network of corner-linked oxygen octahedra, with the smaller cations filling the octahedral holes and the large cation filling the dodecahedral holes.

The stability of this structure depends critically on the relative sizes of the cations. Most compounds with this structure form distorted perovskite cells in which the atoms are slightly displaced from the ideal positions. The flexibility of the perovskite structure allows a wide range of solid solutions. It is possible to substitute an element of a different valence, in which case the resulting charge imbalance is usually accommodated by the formation of oxygen vacancies. Some of the compounds can accommodate a large number of point defects and still remain stable. Perovskites and spinels typify many ceramic materials in which electrical or magnetic properties are important.

Compounds exhibiting perovskite structure include $CaTiO_3$ (perovskite), $BaTiO_3$, $SrTiO_3$, $PbTiO_3$, $CdTiO_3$, $SrSnO_3$, $CaSnO_3$, $BaSnO_3$, $CaZrO_3$, $SrZrO_3$, $PbZrO_3$, $BaZrO_3$, $CdCeO_3$, $CaCeO_3$, $SrCeO_3$, $PbCeO_3$, $BaCeO_3$, $SrHfO_3$, $BaHfO_3$, $BaPrO_3$, $BaThO_3$, $KNbO_3$, $NaNbO_3$, $LaAlO_3$, $YAlO_3$, $LaMnO_3$, $LaFeO_3$, $LaCrO_3$, $PbMgF_3$, $KZnF_3$, $KNiF_3$, $NaWO_3$, $KMgF_3$, $PbMg_{1/3}Nb_{2/3}O_3$, etc., and the structures of most of the cuprate superconductors are based on the perovskite structure.

2.6.2 Ilmenite Structure (*R*$\bar{3}$) ABX_3 compounds in which the relative sizes of the cations do not allow the perovskite structure to be stable adopt the ilmenite structure. This structure is similar to the corundum structure except that the two cations reside in the octahedral sites in alternating layers [Fig. 6(b)]. Compounds exhibiting ilmenite structure include $NiTiO_3$, $MnTiO_3$, $NiMnO_3$, $CoMnO_3$, $MgTiO_3$, $CoTiO_3$, $FeTiO_3$, $CdTiO_3$, etc.

2.6.3 Lithium Niobate Structure (*R*3*c*) This structure is the same as the ilmenite structure except that each layer of cations contain an ordered arrangement of two cations [Fig. 6(c)]. Compounds exhibiting this structure include $LiNbO_3$, $LiTaO_3$, etc.

2.7 A_2BX_4

2.7.1 Olivine Structure (*Pbnm*) The olivine structure is based on a hexagonal close packing of anions in which half of the octahedral sites and one-eighth of the tetrahedral sites are occupied by *A* and *B* cations, respectively. Compounds exhibiting this structure include $(Mg,Fe)_2SiO_4$ (olivine), Mg_2SiO_4, Fe_2SiO_4 (fayalite), γ-Ca_2SiO_4, Mg_2GeO_4, Al_2BeO_4, Na_2BeF_4, $CaMgSiO_4$, $LiMnPO_4$.

2.8 Others

2.8.1 Graphite (*P6₃/mmc*) The sp^2 hybridized bonding in graphite results in a hexagonal layered structure with strong σ bonds between carbon atoms in the plane and weaker interaction between π bonds between the planes [Fig. 8(a)]. Compounds/elements exhibiting this structure include carbon and BN.

2.8.2 Diamond (*Fd3m*) When carbon bonds consist of hybrid sp^3 orbitals, the atoms must be tetrahedrally coordinated [Fig. 8(b)]. This structure is essentially the same as the zinc blende structure except that only one element is present.

2.8.3 High-Temperature Superconductors In $YBa_2Cu_3O_7$ (*Pmmm*), the Cu atoms occupy two sites on both sides of the Ba–O planes. On one side, CuO_2 sheets sandwich the Y atoms with the Cu atoms having a square pyramidal (fivefold) coordination. The Cu–O atoms on the other side of the Ba–O planes form linear chains with Cu atoms having fourfold planar coordination, which leads to an orthorhombic lattice. The structures can be derived from three perovskite unit cells ($ACuO_3$) by stacking them with the cation sequence of -Y-Ba-Ba-Y-Ba-Ba- on the *A* site and removing two (out of nine) oxygens per unit cell. In fact, all of the high-T_c cuprate superconductors have structures related to perovskite (Park and Snyder, 1995).

3. DEFECT STRUCTURES

Disturbances of the strict periodicity of the crystal arise because of defects. These are characterized by dimensions as point defects (vacancies, interstitial atoms, substitutional atoms), one-dimensional (dislocations), two-dimensional (low- and high-angle grain boundaries, stacking faults, external surfaces), or three-dimensional defects (inclusions, pores). In addition to structural defects, crystals also contain electronic imperfections (i.e., electrons and holes), which either can be formed intrinsically through ionization of electrons or may be formed in association with point defects.

Numerous oxides exhibit marked deviations from stoichiometry. Some oxides are even known to be unstable when stoichiometric. Wustite, e.g., which is nominally written as FeO, has under all equilibrium conditions a deficiency of iron ($Fe_{1-y}O$). An exact stoichiometric composition in ceramic compounds is in principle the exception rather than the rule (Kofstad, 1972). For

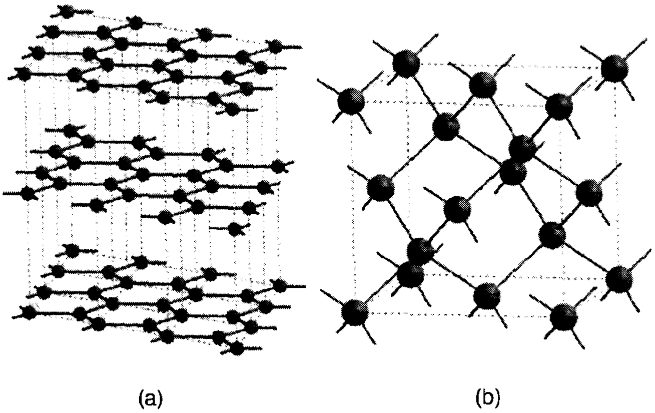

FIG. 8. Structures of (a) graphite and (b) diamond.

more detailed information, see POINT AND OTHER DEFECTS IN CRYSTALS.

3.1 Defect Structures in Stoichiometric Oxides

If a charged point defect is formed in a stoichiometric crystal with composition MO, a complementary point defect with opposite effective charge must be formed to maintain the electrical neutrality of the stoichiometric crystal. In general, the types of defect found in a binary ionic crystal are Schottky ($V_M + V_O$), Frenkel ($V_M + M_i$), and anion-Frenkel ($V_O + O_i$) disorder.

3.2 Defect Structures in Nonstoichiometric Oxides

Nonstoichiometry in a compound is equivalent to the presence of point defects, and the extent of nonstoichiometry is a direct measure of the net concentration of the corresponding defect(s) in a compound. While complementary point defects are formed in stoichiometric crystals, the electrical neutrality of nonstoichiometric compounds is conserved through formation of point defects and complementary electronic defects.

3.2.1 Oxygen-Deficient Oxides When a nonstoichiometric oxide is oxygen deficient, the predominating defects may be oxygen vacancies (V_O), interstitial metal ions (M_i), or both. If oxygen vacancies predominate, the formula of an oxide, say MO_2, may be written as MO_{2-x} [Fig. 9(a)]. If interstitial cations predominate, the composition can be written as $M_{1+y}O_2$ [Fig. 9(b)]. If both types are present, the formula may be written as $M_{1+y}O_{2-x}$. The formation of oxygen vacancies or interstitial cations leads to the formation of complementary free electrons, and electrical conductivity by transport of electrons can result (n-type conductor). Examples of oxides exhibiting overall oxygen deficiency are TiO_2, ZrO_2, CeO_2, Nb_2O_5, etc.

3.2.2 Metal-Deficient Oxides When a nonstoichiometric oxide is metal deficient, the predominating defects may be metal vacancies (V_M), interstitial oxygen atoms (O_i), or both. The formation of charged metal-ion vacancies results in complementary positive electronic defects (holes). In such oxides the electrical conductivity involves the transport of holes (p-type conductor).

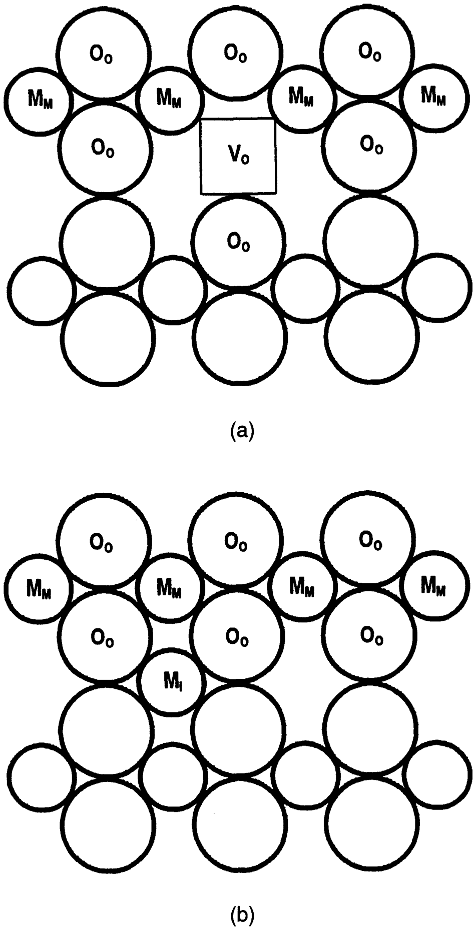

FIG. 9. Oxygen-deficient oxides. (a) MO_{2-x} in which oxygen vacancies (V_O) predominate, and (b) $M_{1+y}O_2$ in which interstitial cations (M_i) predominate.

In metal-deficient oxides where the disorder is limited to the cation lattice, the point defects can be associated and form complex combinations. In the case of $Fe_{1-y}O$ (rock-salt structure), for each Fe ion vacancy created as a result of nonstoichiometry, an equal number of Frenkel defect pairs are simultaneously formed in the metal lattice resulting in the formation of complex defects ($V_{Fe}Fe_iV_{Fe}$) [Fig. 10(a)]. Well-known examples of oxides with cation vacancy-type defects are MnO, FeO, CoO, NiO, Cu_2O, etc.

If the disorder in a metal-deficient oxide is limited to the oxygen lattice, the point de-

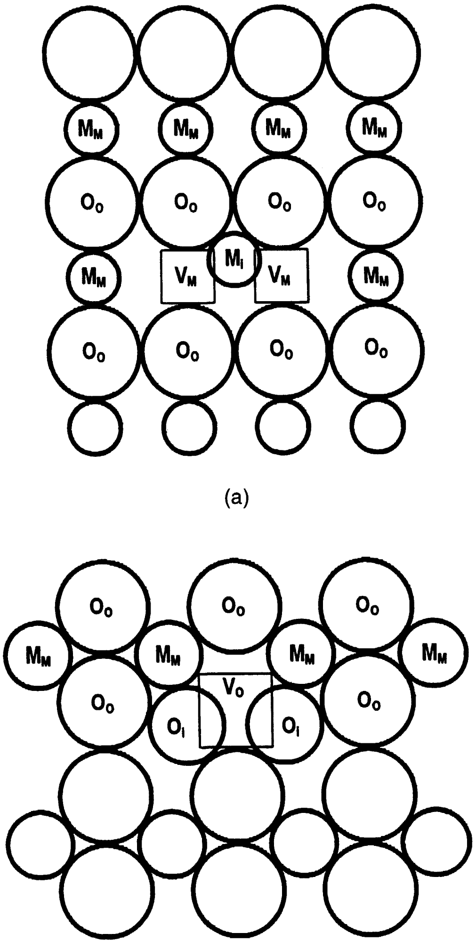

(a)

(b)

FIG. 10. Metal-deficient oxides. (a) $M_{1-y}O$ with complex metal vacancy defect, and (b) a complex oxygen interstitial defect in an oxide M_{2+x} with excess oxygen.

fects may constitute single interstitial oxygen atoms or ions in a manner illustrated in Fig. 9(b) for interstitial cations. But, as in the case of the metal-vacancy defects, interstitial oxygen can form complex interstitial-oxygen defects. In the case of UO_{2+x}, vacant oxygen sites are also created as a result of the introduction of the interstitial oxygen ions. The creation of these defects can be visualized by assuming that each excess interstitial oxygen ion displaces one of the neighboring oxygen ions into a neighboring interstitial position [Fig. 10(b)].

4. MICROSTRUCTURE

Generally, a polycrystalline ceramic material is a complex system composed of crystallites, grain boundaries, and pores. In addition, secondary phases and inhomogeneity in compositions may be present, particularly at grain boundaries. Such microstructural parameters greatly affect the performance of the material even when the property itself originates from the structural parameters on an atomic level. The type of atomic bonding and crystalline structure that determine whether a material is brittle, semiconducting, magnetic, etc. may be masked or enhanced by the microstructure.

The fundamental microstructural parameter in polycrystalline ceramic materials is the grain boundary. The study of the effect of grain size and shape, grain orientation, grain-boundary chemistry, and porosity is of considerable importance to understanding the material properties.

Porosity is almost always present as a phase in ceramics. This results primarily from the relatively high solubility of gases at high temperatures and the slow rate of sintering, especially when the initial material is coarse grained. Pores can be open to the surface and connected to one another (open porosity) or isolated from the surface (closed porosity).

Grain boundaries have a particularly profound effect on the microstructure-sensitive properties of ceramics such as mechanical behavior. In pure single-phase materials, grain boundaries are simply the surfaces separating the individual crystallites, and they can be regarded as thin regions of atomic disarray where the atomic density is slightly lower than normal. Grain boundaries can thus act as sources of and sinks for structural defects such as point defects and dislocations and as sites for pores. In impure materials, the impurities and second phases are almost always localized at grain boundaries, which can lead to a variety of phenomena.

Grain-boundary–pore interactions—whether pores move when attached to the grain boundaries or grain boundaries break away from the pores—largely determine the degree of densification because grain boundaries provide an easy route for diffusion of pores during sintering. Therefore, grain growth in ceramics is seldom determined by the intrinsic grain-boundary mobility but is usually controlled by the presence of impurities and pores. The specific problems encountered in studies of grain-growth phenomena in ceramics include the dominant role of impurities, the effects resulting from a charged grain boundary, and the effects of solute segregation and phase separation near grain boundaries.

The crystallographic orientation of the individual grains may not be randomly distributed but can exhibit a preferred orientation. This effect can be critical when there is anisotropy in properties.

Chemical homogeneity is another factor to be considered when correlating microstructure and property. Chemical inhomogeneity can result in a nonuniform microstructure and thus make interpretation of the effects of other microstructural parameters ambiguous.

The parameters affecting the microstructure are usually correlated so that optimization of a particular property is a complex process. The various microstructural parameters can be controlled by processing conditions and also by changing the starting composition of the material. Many additives are used to adjust the microstructure for desired properties. The concentration and distribution of the additives (and/or impurities) also affect the evolution of the microstructure.

A typical and very well-known example of a microstructural control in ceramics is MgO-doped Al_2O_3, where MgO acts as a grain-growth inhibitor and can lead to a fine-grained, pore-free, microstructure. The detailed mechanisms of how MgO acts to change the microstructure evolution in Al_2O_3 are still not fully understood (Bennison and Harmer, 1990). The amount of MgO required is a few hundred ppm and is sometimes even lower in concentration than the amount of background impurities. Al_2O_3 responds to MgO doping in different degrees, depending on its synthesis and the background impurity level (MgO has been observed to increase densification directly in some powders and to decrease densification directly in others). The single most important effect of MgO doping is its ability to reduce significantly the grain-boundary mobility, regardless of the presence of varying background impurities, through a solid-solution pinning mechanism. The pinning action of MgO reduces the tendency for pore–grain-boundary separation and abnormal grain growth arising from inhomogeneous densification and nonuniform liquid phase distribution.

Various properties will respond to the microstructure differently. The magnetic permeability, e.g., has a definite linear relationship with the mean grain size (if the pores remain at the grain boundary). On the other hand, the critical flaw size is roughly equal to the largest microstructural feature, i.e., not the mean grain size, but the largest grain size in the structure (when the grain size is the dominant parameter). Thus small mean grain size is not sufficient to give high strength. A narrow grain-size distribution together with small mean grain size is required for high mechanical strength. This is a typical example of the influence of the microstructure on a "bulk" property (permeability) and a "critical" property (mechanical strength) (Stuijts, 1977).

5. STRUCTURES AND MECHANICAL BEHAVIOR

5.1 Brittleness

The most distinguishing mechanical property of ceramics is their brittleness—that is, catastrophic failure following an almost entirely elastic deformation. Strong covalent/ionic bonding, which is dominant in the crystal lattice of ceramic materials, leads to a high melting point, high corrosion resistance, high wear resistance against abrasion, high hardness, high Young's modulus, etc. On the other hand the strong covalent/ionic bonding hinders the movement of adjacent atoms/ions so that the motion of a dislocation is either absent or present to a very limited extent except at very high temperatures.

A much smaller number of independent slip systems is available in ceramics than in metals. The slip systems for a number of common ceramic crystals are summarized in Table 3. In ionic crystals, a dislocation must also maintain the cation–anion site ratio, which makes the dislocation motion even more difficult. The Burgers vector in ceramics is greater than that found in metals, which means that more energy is required to initiate the dislocation motion in ceramics.

Table 3. Primary slip systems of ceramics at low temperature. At high temperature, slip can occur on other planes. Those for fcc metal systems are included for comparison. (After Davidge, 1979; Kingery *et al.*, 1975.)

Material	Crystal structure	Primary slip system	NSS[a]
Al_2O_3	Hexagonal	$0001\langle 11\bar{2}0\rangle$	2
BeO	Hexagonal	$0001\langle 11\bar{2}0\rangle$	2
MgO	Cubic (NaCl)	$110\langle 1\bar{1}0\rangle$	2
$MgAl_2O_4$	Cubic (spinel)	$111\langle 1\bar{1}0\rangle$	5
β-SiC	Cubic (ZnS)	$111\langle 1\bar{1}0\rangle$	5
β-Si_3N_4	Hexagonal	$10\bar{1}0\langle 0001\rangle$	2
TiC	Cubic (NaCl)	$111\langle 1\bar{1}0\rangle$	5
UO_2	Cubic (CaF_2)	$001\langle 1\bar{1}0\rangle$	3
Cu, Ag, etc.	fcc metals	$111\langle 110\rangle$	12

[a] Number of associated independent slip systems.

The fewer slip systems, problems associated with charge neutrality, and longer Burgers vector add up to a much greater resistance to plastic deformation in crystalline ceramics than in their metallic counterparts.

The high toughness of metals, in comparison to ceramics, is a result of plastic deformation, which occurs by movement of dislocations. The plasticity found in metals is not possible in ceramics, resulting in the brittleness of ceramics.

On loading, the ceramic will deform elastically until it fractures catastrophically. This behavior can be understood in terms of the stress that is concentrated at the crack-tip region. The level of these stress concentrations can be described by introducing a stress intensity factor (K): $\sigma \sim K\alpha^{-1/2}$, where σ and α are applied stress and crack length, respectively. A crack can proceed when K becomes higher than a critical value (critical stress intensity factor or fracture toughness, K_{IC}): $\sigma_f \sim K_{IC}\alpha^{-1/2}$ where σ_f is the fracture strength. Improvement of fracture strength basically can be achieved by either a smaller starting crack length (by modifying the process) or a higher value of K_{IC} (by toughening) (Schubert and Petzow, 1988). For more detailed information on mechanical properties, see the related articles in this encyclopedia (MECHANICAL PROPERTIES OF SOLIDS; CERAMICS, MATERIAL PROPERTIES OF)

The strengthening of metals can be accomplished by controlling the movement of dislocations. In ceramics that have high shear strength combined with crack sensitivity and low fracture strength, toughening can be obtained by controlling the propagation of cracks. Just as in metals, high toughness and high strength do not generally co-exist. Toughening mechanisms in ceramics can lower the strength. The strength, however, is no more sensitive to damage or handling because the toughened ceramics become more crack resistant and can retain strength better.

5.2 Toughening Mechanisms

5.2.1 Transformation Toughening

The transformation-toughening concept exploits the martensitic transformation of ZrO_2 [Fig. 11(a)]. The stable form of ZrO_2 at room temperature has monoclinic symmetry with seven-fold coordination of the cation. The following transformations take place at elevated temperatures: monoclinic–1200°C–tetragonal–2300°C–cubic–2600°C–melt. Partially stabilized zirconia (PSZ) is a two-phase material that consists of precipitates of tetragonal and/or monoclinic phase (depending on the heat treatment) dispersed in a cubic matrix. CaO,

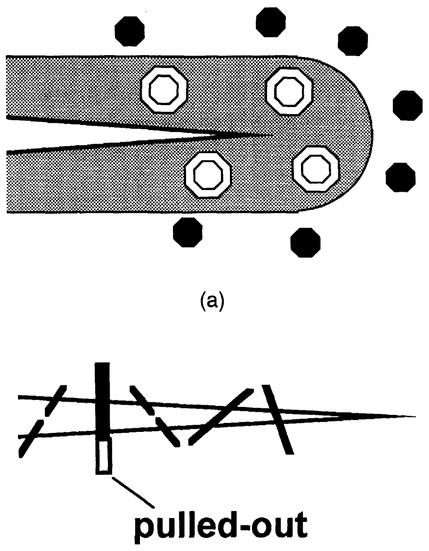

FIG. 11. (a) Stress-induced transformation toughening. Black particles are original metastable tetragonal ZrO_2 particles. The white ones are martensitically transformed monoclinic particles. (b) Fiber reinforcement toughening.

MgO, or Y_2O_3 is commonly used as stabilizing agent.

The tetragonal-to-monoclinic phase transformation (t-m transformation) at ~1200 °C is accompanied by a volume increase of 3–5%. Both shear and tension promote the transformation, while compression hinders it. On cooling from high temperature, the t-m transformation takes place accompanied by a volume change. If the grains are embedded in a stiff matrix (cubic ZrO_2 or another ceramic such as Al_2O_3), the volume increase would create compressive stress and therefore hinder the transformation. In practice, the t-m transformation temperature can be depressed by several hundred degrees Celsius. In the systems that are used in practical applications such as Mg–PSZ, Al_2O_3–ZrO_2, and Y–TZP (tetragonal zirconia polycrystal), the tetragonal phase can be retained down to temperatures even below room temperature.

When shear or tensile stresses are applied, however, the transformation is triggered. In the stress field ahead of a crack tip, shear and tensile stresses are always present and are usually strong enough to start the t-m transformation. There is a critical stress required to transform the tetragonal grains. The zone in which the transformation occurs has a shape similar to the plastic zone in metals. The tetragonal ZrO_2 grains in this zone undergo transformation with a volume increase. The resulting compressive stress acts on the crack tip, reducing the stress concentration, and, thus, quenches the fracture.

5.2.2 Microcracking Another toughening mechanism is induced by the dispersion of unstabilized ZrO_2 particles in a matrix. Upon cooling, the unstabilized particles transform to the monoclinic phase, and the resultant stress field (due to the volume expansion) causes localized microcracking around the ZrO_2 particles. These microcracks can absorb or dissipate the energy of a propagating crack by extending in its stress field or by deflecting it, thereby increasing the toughness of the ceramic. The toughness will increase to a maximum, above which microcracks generated by the ZrO_2 particles will interact with each other resulting in a decrease in strength.

In addition to the phase transformation, residual stress can be caused by the thermal expansion mismatches (in multiple-phase systems), thermal expansion anisotropy, or elastic anisotropy. These residual stresses can also contribute to the toughening by forming microcracks.

5.2.3 Whisker or Fiber Reinforcement Whiskers or fibers densely embedded within a matrix (see COMPOSITE MATERIALS) may have thermal expansion coefficients or Young's modulus values different from those of the matrix material [Fig. 11(b)]. Consequently, on cooling from the sintering temperature, the different shrinkages of the matrix and the inclusions lead to residual stress. For inclusions with lower expansion coefficients, the inclusion is under compression and induces radial compressive stress and tangential tensile stress in the matrix. This residual stress field is superposed on the applied stress and results in crack deflection. The crack is deflected onto the inclusions, creating a zig-zag-shaped crack path. If the amplitude of this deflection path is large enough, additional frictional forces will appear over the crack surfaces. This changes the stress concentrations around the crack tip and reduces the driving force for crack propagation.

In addition to the above effect, a pullout mechanism is possible if the whiskers/fibers are oriented orthogonally to the crack plane and to the direction of crack propagation and if the interfaces between the inclusions and the matrix are weak. On loading, the matrix material will fracture, but the stronger inclusion will stay intact. At a critical stress, the inclusions are pulled out of their original position. This movement requires the friction at the interface to be overcome, and the frictional forces "bridge" the crack surfaces, which changes the stress concentrations ahead of the crack tip. The pullout effect also leads to crack-energy dissipation and, hence, material toughening.

5.2.4 Ductile Dispersion Toughening The incorporation of a ductile metal phase can also improve the toughness (e.g., Al_2O_3/Al, WC/Co, etc.). The brittle matrix preserves stiffness and lightness, and the metal phase enhances toughness and energy absorption capacity. However, the properties of the two phases must be such that the propagating crack is attracted to the ductile phase.

6. STRUCTURES OF ELECTROCERAMICS

6.1 $BaTiO_3$

$BaTiO_3$ exhibits a perovskite structure. Both Ba^{2+} and O^{2-} ions have similar ionic radii and together form a face-centered cubic array. Octahedrally coordinated Ti^{4+} ions located at the center of the cubic perovskite cell are the polarizable active ions promoting ferroelectricity and the high permittivity of $BaTiO_3$. On cooling from high temperature, the crystal structure of $BaTiO_3$ undergoes phase transformation from paraelectric cubic to ferroelectric tetragonal at ~130 K (Curie point, T_c).

The interesting electrical properties of this material arise from the Ti–O plane. The octahedral site for Ti is large compared to the size of the Ti ion. The rattling titanium model suggests the presence of minimum-energy positions for the Ti ion that are off center (Fig. 12). The off-centered Ti ions give rise to the electric dipole. At temperatures higher than T_c, the thermal energy is such that the Ti ion has no fixed unsymmetrical position, leaving the structure with an average cubic symmetry. However, even at high temperatures, the open octahedral site allows the Ti ion to develop a large dipole moment in an applied field.

The most interesting properties of $BaTiO_3$ are found at or around T_c. In the tetragonal phase, the linear expansion coefficient along the *c* axis is negative, which means that the crystal structure becomes less and less tetragonal as temperature increases toward the transition temperature. The material is much less polarizable (thus low permittivity) along the *c* axis than in the perpendicular directions. In the vicinity of T_c, the stability of the lattice decreases, and the amplitude of the displacement of the Ti^{4+} ions inside the oxygen octahedral becomes very large. This induces a strong variation of the polarizability and, consequently, a large increase in the permittivity.

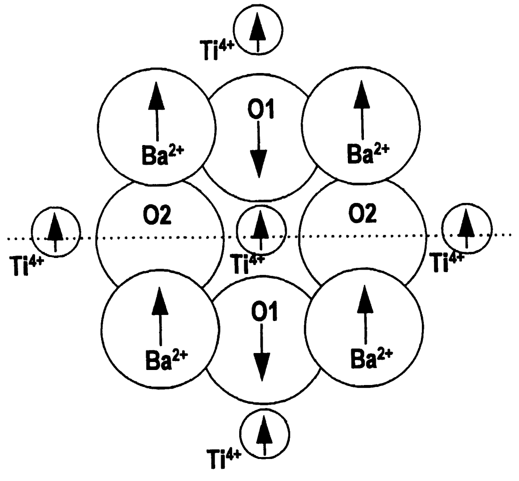

FIG. 12. Ion displacements in the cubic–tetragonal distortion in $BaTiO_3$.

Additions or substitutions of cations can shift the T_c, introduce second phases, restrict domain-wall motion, control crystallite size, or change the defect structure. Shifting the T_c enables the peak permittivity to be put in a temperature range in which it can be exploited. The substitution of Sr^{2+} for Ba^{2+} or of Sn^{4+} or Zr^{4+} for Ti^{4+} lowers T_c, while the substitution of Pb^{2+} for Ba^{2+} increases it. Cations that have a higher valency than those they replace (e.g., La^{3+} in place of Ba^{2+} or Nb^{5+} in place of Ti^{4+}) generally enhance grain growth when dopant level is small but inhibit grain growth when concentration of the dopant gets high. Figure 13 shows the $BaTiO_3$ grains of different sizes obtained in a thin trilaminate. The middle layer with small grains is undoped, and the left and right layers with larger grains are doped with La^{3+} and Mn^{3+}.

$BaTiO_3$, when doped with ions of higher valence such as La^{3+}, Ce^{3+} (for Ba^{2+}), or

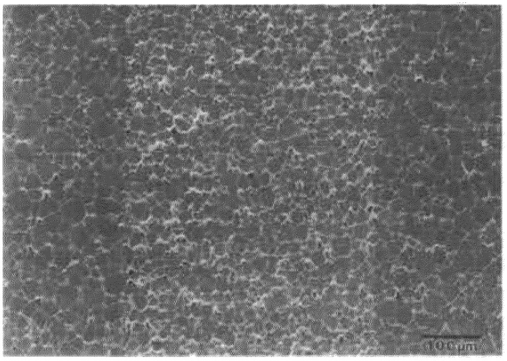

FIG. 13. SEM micrograph of a thin $BaTiO_3$ trilaminate. The large grains in left/right layers were obtained by adding La^{3+} and Mn^{3+} (Capurso, *et al.*, 1995).

Nb^{5+} (for Ti^{4+}), also exhibits interesting semiconducting properties. A large increase in electrical resistance (PTCR effect: positive temperature coefficient of resistivity) occurs over a limited temperature range near the T_c. This semiconducting property of $BaTiO_3$ is connected with grain-boundary resistivity effects in combination with the ferroelectric–paraelectric phase transition.

When sintered at high temperature, La-doped $BaTiO_3$ becomes an *n*-type semiconductor with conduction taking place via transfer of electrons between Ti ions (Ti^{4+} and Ti^{3+}). Thus the $BaTiO_3$ grains are conducting and remain conducting on cooling to room temperature. But the grain-boundary region changes during cooling. Oxygen is adsorbed on the surface of the ceramic and diffuses along grain boundaries. The added oxygen ions assist cation vacancies to form at the grain boundary and to diffuse slowly into the grains. They form a thin layer at the grain boundaries in which they compensate the donors, thereby creating an insulating barrier between grains. The net result of this process is that the ceramic consists of semiconducting grains separated by thin insulating grain boundaries. The electrical resistance of the ceramic is inversely proportional to grain size because smaller grains mean more insulating grain boundaries and, therefore, higher resistance.

Below T_c, the resistivity is low because the electron charge trapped in the grain-boundary region is partially compensated by spontaneous polarization. Wherever the domain structure is favorably positioned, a positive polarization charge will cancel the negatively charged barriers between conductive grains, thereby establishing low-resistivity paths across the ceramic.

Above T_c, the spontaneous polarization disappears, and the resistivity increases, giving rise to the PTCR effect. At first the increase is slow because of the high permittivity at T_c. The barrier height is inversely proportional to the permittivity of the surrounding medium. A highly polarizable medium shields the charges trapped at the grain boundary, reducing the barrier height and the electrical resistance. As the temperature increases further above T_c, the permittivity decreases rapidly according to the Curie–Weiss law. Decreases in permittivity cause rapid increase in the barrier between grains and an increase in the electrical resistance. Eventually the resistance increase levels off as the decrease in permittivity slows down and the normal NTC (negative temperature coefficient) effect takes over. High density, very fine-grained material is desirable for high dielectric constant, while porosity and fine grain size is required for high PTCR effect (Moulson and Herbert, 1990; Newnham, 1983, 1984).

6.2 ZnO Varistor

ZnO varistors (voltage-dependent resistors) are ZnO-based ceramic semiconductor devices with highly nonlinear current (*I*)-voltage (*V*) characteristics. The electrical behavior depends both on the type and amount of oxide additives and on the ceramic microstructure. The microstructure can be visualized as shown in Fig. 14. The resistance is controlled by the region in which the ceramic grains contact one another in much the same way as in the semiconducting $BaTiO_3$ showing PTCR effect.

The basic building block of the varistor is the ZnO grains. During sintering, various cations are distributed in such a way that the near–grain-boundary region becomes highly resistive (~10^{12} Ω cm), and the grain interior becomes highly conductive (~10 Ω cm). A sharp drop in resistivity from the grain boundary to grain occurs within a distance of ~100 nm. This area is known as the depletion layer. Thus, at each grain boundary, there exists a depletion layer on both

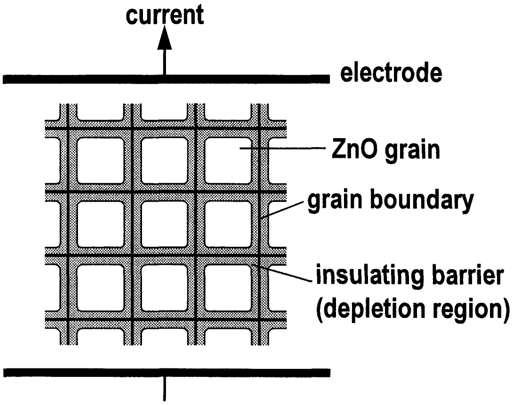

FIG. 14. Schematic of the microstructure of ZnO varistor.

sides of the grain boundary extending into the adjacent grains. The varistor action arises as a result of the presence of this depletion layer within the grains. This depletion layer provides a barrier to the transport of electrons across the grain boundary.

The theoretical and experimental studies of this barrier conform to a classical Schottky-barrier model, with positive charges residing in the depletion layer and the compensating negative charges at the grain-boundary interface as traps. Since the region near the grain boundary is depleted of electrons, a voltage drop appears across the grain boundary upon application of an external voltage. This macroscopic breakdown voltage per intergranular barrier (barrier voltage) is typically of the order of ~2–4 V/barrier.

The electrical characteristics of the ZnO varistors are therefore related to the bulk structure; i.e., the device is inherently multijunction with varistor action shared between the various ZnO grain boundaries. This implies that tailoring the device breakdown voltage is simply a matter of fabricating a varistor with the appropriate number of grains in series between the electrodes. Thus to achieve a given breakdown voltage, one can change the varistor thickness (for fixed grain size) or alternatively one can vary the grain size (for fixed thickness) to change the number of barriers.

Many oxides (Bi, Co, Cr, Mn, Sb, Al, Nb, Ba, Ni, etc.) can be added to ZnO either to impart or to improve or adjust the nonlinear *I-V* characteristics. These oxide additions can affect the varistor properties both directly and indirectly: directly by segregating at the grain boundaries and changing the *I-V* characteristics, indirectly by affecting the microstructural development. Because the varistor breakdown voltage is inversely proportional to the average grain size of the ZnO, it is important to understand the effects of the individual oxide additions on the microstructure of ZnO varistors. For example, Sb_2O_3 and Al_2O_3 react with ZnO and form the $Zn_7Sb_2O_{12}$ spinel-type and $ZnAl_2O_4$ spinel structures, respectively. These spinel particles are insulating and play only a secondary role in determining device properties, but they reduce the rate of ZnO grain growth (Levinson and Philipp, 1986; Newnham, 1984).

7. STRUCTURES OF FERRITES

7.1 Spinel Ferrites

An important family of ferrites are spinels with the chemical formula MFe_2O_4, in which M represents a divalent ion such as Mn^{2+}, Fe^{2+}, Co^{2+}, Ni^{2+}, Cu^{2+}, Zn^{2+}, or a combination of ions with an average valence of 2.

The cation distribution can be disordered, particularly at high temperatures, and all the ferrimagnetic spinels are more or less inverse. The site preference of the metallic ions determines the saturation magnetization of the material and can be explained by the ionic radius and the size of the lattice site. The divalent ions are generally larger than the trivalent, and the *B* sites are larger than *A*. Therefore, it would be reasonable that the trivalent ions and the divalent ions would go into the *A* and *B* sites, respectively. Two exceptions are found in Zn^{2+} and Cd^{2+}, which prefer *A* sites because of their electronic configurations.

The exchange force between the moments of two metal ions on different sites depends on the distances between them, the oxygen ion that links them, and also the angle between the three ions. The interaction between moments on *A* and *B* sites is greatest. The *BB* and *AA* interactions are much weaker.

In case of Zn ferrite ($ZnFe_2O_4$), the *A* sites are occupied by Zn^{2+} ions, which, being nonmagnetic (having no unpaired electronic spins), do no ordering of the ions on the *B* sites occupied by Fe^{3+} ions. The interactions between Fe^{3+} ions (*BB* interaction) are negligible. In Ni ferrite ($NiFe_2O_4$), the Fe^{3+} ions fill the eight *A* sites while the other eight Fe^{2+} ions and eight Ni^{2+} ions go on the *B* sites. The antiferromagnetic interaction orients the eight Fe^{2+} moments and eight Ni^{2+} moments antiparallel to the eight Fe^{3+} moments on the *A* sites. The Fe^{3+} ion moments will cancel, but the moments on the Ni^{2+} ions give rise to an uncompensated moment or magnetization.

In mixed Zn–Ni ferrite where Zn replaces some (e.g., half) of the Ni, four Zn^{2+} and four Fe^{3+} ions fill the eight *A* sites and twelve Fe^{3+} and four Ni^{2+} ions go on the *B* sites. The Fe^{3+} moments in the *A* sites neutralize only one-third of those in *B* sites, leaving the other two-thirds oriented but uncompensated, giving a net magnetic moment. The total uncompensated oriented ions consist of one-half of the Fe^{3+} ions plus all the Ni^{2+} ions, giving a large magnetic moment (Buchanan, 1986; Kingery *et al.*, 1975; Moulson and Herbert, 1990). For more detailed information, see MAGNETIC MATERIALS; MAGNETIC ORDERING IN SOLIDS.

7.2 Magnetoplumbite

The magnetoplumbite structure is hexagonal, and therefore it shows strong uniaxial anisotropy along the *c* axis, which is the preferred axis (often called the easy direction) for magnetization to be oriented. The general formula is $M\text{Fe}_{12}\text{O}_{19}$ (or $M\text{O}\cdot 6\text{Fe}_2\text{O}_3$ giving rise to the common name "hexaferrites"), where *M* can be Ba, Sr, or Pb. A unit cell contains two formula units, i.e., a total of 64 ions. The M^{2+} and O^{2-} ions together form a close-packed structure with some of the layers cubic close packed and others hexagonal close packed. Of the twelve Fe^{3+} ions in a formula unit, nine are on octahedral sites, two on tetrahedral sites, and one on a five-coordinated site. Seven of the ions on octahedral sites and the one on a five-coordinated site have their spins in one direction, and the remainder are oppositely directed. Thus there are four more ions with spins in the one sense than there are with opposing spins. Since there are five electrons with parallel spins in each Fe^{3+} ion, there are 20 unpaired spins per formula unit. The presence of a unique axis in hexagonal ferrites, which cannot be found in spinel structures, contributes to the high coercivities that make these materials useful as hard ferrites (Buchanan, 1986; Moulson and Herbert, 1990).

7.3 Microstructures of Ferrites

Some factors such as saturation magnetic moment are a function only of a chemical composition and crystal structure. Others, such as permeability and losses, are strongly dependent on microstructural conditions.

Both the grain size, porosity, and the distribution of porosity determine the initial permeability. Grain boundaries and other defects such as pores, second phases, cracks, and impurities can impede domain-wall motion and decrease the permeability. Therefore, the permeability increases with increasing grain size and decreasing porosity. In practice the permeability increases with increasing average grain size up to a certain point and drastically decreases with increasing percentage of grains possessing pores (intrangranular pores) despite increasing grain size.

Exaggerated grain growth can also trap pores within grains during growth, and hence pin domain walls and finally lower permeability and increase losses. This would happen by inhomogeneous distribution or segregation of impurities or dopants such as SiO_2 and CaO, which produce rapid grain growth locally.

When grain size is very small, the grains cannot support domain walls. Therefore, only the rotational motion of domains contribute to permeability. Larger grains contain domain walls, and their movement is only limited by the grain boundaries and the defects included inside the grain (such as intragranular pores). Therefore, motion of domain walls is sensitive to grain size and the distribution of pores. If pores are suppressed or located at the grain boundaries, the permeability will increase with grain size.

Hysteresis loss is controlled by the same factors that control the low-frequency permeability and coercivity (grain size, porosity, impurities, etc.). The lowest hysteresis losses are achieved with minimum anisotropy, large and uniform grain sizes associated with reduced grain-boundary areas, and low intergranular porosity.

The eddy-current losses that are important for high-frequency operation can be reduced by increasing the resistivity. Resistivity can be increased by adding CaO and SiO_2. Ca and Si ions substitute in the ferrite crystal lattice close to the grain boundary, rendering the grain-boundary region insulating. Resistivity can also be increased by sintering in oxidizing atmospheres, which induces preferential grain-boundary oxidation by lowering Fe^{2+} concentrations (this suppresses electron hopping between Fe^{2+} and Fe^{3+}). Eddy-current losses can be further reduced by reducing grain size and increasing porosity, which increases the total volume of the resistive region.

Therefore, there is a compromise between the large-grain, dense ferrites for high permeability at low frequencies and the small-grain, porous ferrites for low losses at high frequencies (Buchanan, 1986; Moulson and Herbert, 1990).

GLOSSARY

Burgers Vector: Displacement vector necessary to close a stepwise loop around a dislocation.

Coercive Force: The magnetizing force that must be applied in the direction opposite to that of the previous magnetizing force in order to reduce magnetic flux density to zero.

Coordination Number: The number of anions surrounding each metal ion.

Covalent Bond: A primary bond in which two atoms share a pair of electrons.

Curie Temperature: The temperature marking the transition between ferromagnetism and paramagnetism, or between the ferroelectric phase and paraelectric phase.

Defect Structure: A complete description of the point and electronic defects in a compound and their concentrations as a function of temperature and partial pressure of the constituent atoms or molecules.

Domain: A part of a crystal in which all of the magnetic moments are cooperatively aligned.

Electronegativity: A qualitative property that is a measure of the power of an atom in a molecule to attract electrons to itself.

Fracture Toughness: The ability to carry loads or deform plastically in the presence of a notch.

Ionic Bond: A primary bond arising from the electrostatic attraction between two oppositely charged ions; a type of chemical bonding in which one or more electrons are transferred completely from one atom to another, thus converting neutral atoms into electrically charged ions.

Partially Stabilized Zirconias: Two-phase materials that consist of precipitates of tetragonal and/or monoclinic phase (depending on the heat treatment) dispersed in a cubic matrix. CaO, MgO, or Y_2O_3 are commonly used as stabilizing agent.

Permeability: A generic term used to express various dynamic relations between magnetic induction (B) and magnetizing force (H), under various conditions of magnetic excitation.

Remanent Magnetization: The magnetic induction remaining in a magnetic material after removal of the applied magnetizing force.

Slip System: A combination of families of crystallographic planes and directions corresponding to dislocation motion.

Stoichiometric: Having the precise weight relation of the elements in a chemical compound; or (quantities of reacting elements or compounds) being in the same weight relation as the theoretical combining weight of the elements involved.

Works Cited

ASM Handbook (1991), *Engineered Materials Handbook,* Vol. 4, *Ceramics and Glasses,* Metals Park, OH: ASM International.

Bennison, S. J., Harmer, M. P. (1990), in: C. A. Handwerker, J. E. Blendell, W. A. Kaysser (Eds.), *Sintering of Advanced Ceramics,* Westerville, OH: American Ceramic Society, pp. 13–49.

Buchanan, R. C. (Ed.) (1986), *Ceramic Materials for Electronics,* New York: Marcel Dekker.

Capurso, J. S., Alles, A. B., Schulze, W. A. (1995), *J. Am. Ceram. Soc.* **78,** 2476–2480.

Davidge, R. W. (1979), *Mechanical Behavior of Ceramics,* Cambridge, U.K.: Cambridge Univ. Press.

Kingery, W. D., Bowen, H. K., Uhlmann D. R. (1975), *Introduction to Ceramics,* New York: Wiley.

Kofstad, P. (1972), *Nonstoichiometry, Diffusion, and Electrical Conductivity in Binary Metal Oxides,* New York: Wiley.

Levinson, L. M., Philipp, H. R. (1986), *Am. Ceram. Soc. Bull.* **65,** 639–646.

Megaw, H. D. (1973), *Crystal Structures: A Working Approach,* London: W. B. Saunders Company, pp. 173–232.

Moulson, A. J., Herbert, J. M. (1990), *Electroceramics,* New York: Chapman and Hall.

Newnham, R. E. (1983), *J. Mater. Educ.* **5,** 941–982.

Newnham, R. E. (1984), *J. Mater. Educ.* **6,** 807–839.

Park, C., Snyder, R. L. (1995), *J. Am. Ceram. Soc.* **78,** 3171-3194 (1995).

Schubert, H., Petzow, G. (1988), *J. Mater. Educ.* **10,** 601–620.

Stuijts, A. L. (1977), in: R. M. Fulrath, J. A. Pask (Eds.), *Ceramic Microstructures '76: With Emphasis on Energy Related Applications,* Boulder, CO: Westview Press, pp. 1–26.

Further Reading

Askeland, D. R. (1984), *The Science and Engineering of Materials,* Monterey, CA: Brooks/Cole.

Azaroff, L. V. (1960), *Introduction to Solids,* New York: McGraw-Hill.

Burns, G., Glazer, A. M. (1990), *Space Groups for Solid State Scientists,* Boston: Academic.

Evans, A. G. (Ed.) (1984), *Fracture in Ceramic Materials,* Park Ridge, NJ: Noyes Publications.

Evans, R. C. (1964), *An Introduction to Crystal Chemistry,* Cambridge, U.K.: Cambridge Univ. Press.

Fulrath, R. M., Pask, J. A. (Eds.) (1977), *Ceramic Microstructures '76: With Emphasis on Energy Related Applications,* Boulder, CO: Westview Press.

Green, D. J., Hannink, R. H., Swain, M. V. (1989), *Transformation Toughening of Ceramics,* Boca Raton, FL: CRC Press.

Greuter, F., Blatter, G. (1990), *Semicond. Sci. Technol.* **5,** 111–137.

Handwerker, C. A., Blendell, J. E., Kaysser, W. A. (Eds.) (1990), *Sintering of Advanced Ceramics,* Westerville, OH: American Ceramic Society.

Hench, L. L., West, J. K. (1990), *Principles of Electronic Ceramics,* New York: Wiley-Interscience.

Jaffe, B., Cook Jr., W. R., Jaffe, H. L. (1971), *Piezoelectric Ceramics,* London: R.A.N. Publishers.

Lawn, B. (1993), *Fracture of Brittle Solids,* Cambridge, U.K.: Cambridge Univ. Press.

Levinson, L. M. (Ed.) (1988), *Electronic Ceramics,* New York: Marcel Dekker.

Niihara, K., Ishizaki, K., Isotani, M., (Eds.) (1994), *Materials Processing and Design: Grain-Boundary-Controlled Properties of Fine Ceramics II,* Westerville, OH: American Ceramic Society.

Pask, J. A., Evans, A. G. (Eds.) (1987), *Ceramic Microstructures '86: Role of Interfaces,* New York: Plenum.

Reed, J. S. (1995), *Principles of Ceramics Processing,* New York: Wiley-Interscience.

Muller, O., Roy, R. (1974), *The Major Ternary Structural Families,* New York: Springer-Verlag.

Xu, Y. (1991), *Ferroelectric Materials and Their Applications,* New York: North-Holland.

Yan, M. F., Heuer, A. H. (Eds.) (1983), *Advances in Ceramics,* Vol. 7, *Additives and Interfaces in Electronic Ceramics,* Columbus, OH: American Ceramic Society.

SUN, STRUCTURE OF

See Supplement Volumes

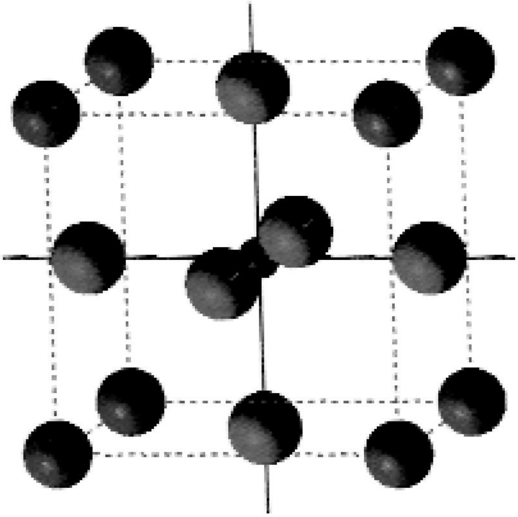

Vol. 20 – PLATE 1. A unit cell of perovskite structure.

SUPERCONDUCTING DEVICES

Kevin A. Delin, *Center for Space Microelectronics Technology, Jet Propulsion Laboratory, Pasadena, California, U.S.A.*

Introduction 179
1. **Basic Physics of Superconductors** 179
2. **General Electromagnetic Properties of Superconductors** 180
3. **Passive Superconducting Devices** 181
3.1 Microwave Filters 182
3.2 Nuclear Magnetic Resonance Spectroscopy Coils 183
3.3 Microwave Delay Lines 183
4. **Superconducting Weak Links** 183
4.1 Elementary Model of Weak Links 184
4.2 Fabrication of Practical Weak Links 184
4.2.1 Low-T_c Devices 185
4.2.2 High-T_c Devices 185
5. **Active Superconducting Devices** 187
5.1 Superconducting Quantum Interference Device 187
5.2 Logic Circuits 188
5.2.1 Latching Logic 189
5.2.2 Single–Flux-Quantum Logic 189
5.3 Three-Terminal Devices 190
5.4 Mixers 191
6. **Summary** 192
Further Reading 193

INTRODUCTION

The number of potential applications for superconductors in electrical circuits, both lumped and continuous, is large. Some applications merely require replacing an Ohmic material with a superconductor to reduce system loss and improve sensitivity. Microwave circuits made with superconducting striplines are one such example. However, like the laser, superconductivity is a quantum phenomenon that is accessible on a macroscopic scale. This unusual property allows engineers to build, for instance, magnetometers of great sensitivity by exploiting quantum interference. This article surveys both the passive devices used in microwave applications and the active devices used in applications as diverse as magnetometry, high-speed computation, and heterodyne mixing. In addition to the electrical engineering concerns, some of the materials issues associated with device fabrication are addressed.

1. BASIC PHYSICS OF SUPERCONDUCTORS

When a superconductor is cooled to a sufficiently low temperature, the motions of all its electrons will correlate into a *single* quantum state. All superconducting devices, in one way or another, take advantage of this property. According to the well-tested theory of John Bardeen, Leon Cooper, and Robert Schrieffer (BCS theory) (see Superconductivity, Low-Temperature), this correlation takes place when two electrons couple to form a Cooper pair. We may therefore think of the electrical charge carriers in a superconductor to be Cooper pairs (or more colloquially, superelectrons) with a mass m^* and a charge q^* that are twice those of normal electrons. The average distance between the two electrons forming a Cooper pair is known as the coherence length ξ_s. Both the coherence length and the binding energy of two electrons in a pair, 2Δ, depend on the particular superconducting material.

In a clean superconductor, where ξ_s is

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

much smaller than the carrier mean free path l, the BCS theory gives the coherence length as

$$\xi_{sc} = \hbar v_{\mathrm{F}} / \pi \Delta. \tag{1}$$

Here, $\hbar$ is the reduced Planck's constant, $h/2\pi$, and v_{F} is the Fermi velocity in the material. Equation (1) can be intuitively understood from the uncertainty principle in which the product of the position and momentum uncertainties of a Cooper pair must exceed $\hbar$. The position uncertainty is ξ_s, while the momentum uncertainty is approximately $\delta E / v_{\mathrm{F}}$, where δE is the uncertainty of the energy about the Fermi surface. Since $\delta E \approx \Delta$, Eq. (1) follows immediately. In a dirty superconductor, where ξ_{sc} is much larger than l, the usual random-walk argument (in three dimensions) leads to the dirty-limit coherence-length expression

$$\xi_{sd} = \sqrt{\xi_{sc} l / 3}. \tag{2}$$

It is inaccurate to think of Cooper pairs as tightly bound electron molecules. Instead, there are usually many other electrons between those of a specific Cooper pair, allowing for the paired electrons to change partners on a time scale of $h/2\Delta$. High-temperature superconductors form a significant exception to this observation, since their coherence lengths are typically on the order of the crystalline lattice spacing.

The pair correlation, and hence superconductivity, can be destroyed by providing energy to the Cooper pairs in excess of their binding energy. This energy can be thermal, and so superconductivity exists only below a critical temperature T_c. Other types of depairing energy are kinetic, resulting in a critical current density J_c, and magnetic, resulting in a critical field H_c.

2. GENERAL ELECTROMAGNETIC PROPERTIES OF SUPERCONDUCTORS

The two hallmark electromagnetic properties of all superconductors are the ability to support a static current without any electrical resistance and the ability to *exclude* a static magnetic flux from the interior of the material. Although the first property is the better known, it is the second property, known as the Meissner effect, that distinguishes a superconductor from a perfect conductor (which only *conserves* a static magnetic flux in its interior).

An extremely useful, yet simple, model for superconductivity can be developed through a set of constitutive relations that describe these two hallmark properties. The zero dc resistance criterion implies that the superelectrons move unimpeded in a superconductor. Therefore, by assuming that the charge carriers do not scatter incoherently, the standard Ohmic conduction model extended to superconductivity becomes

$$\frac{\partial}{\partial t} (\mu_0 \lambda^2 \mathbf{J}_s) = \mathbf{E}, \tag{3}$$

where $\mathbf{J}_s$ is the supercurrent density, $\mathbf{E}$ is the electric field, μ_0 is the permeability of the material, and λ is the penetration depth, defined as $\lambda \equiv \sqrt{m^* / n^* (q^*)^2 \mu_0}$. Here, n^* is the density of the supercurrent carriers. Equation (3) is known as the first London equation after its originators, Heinz and Fritz London. This expression reveals that the superelectrons' inertia creates a lag between their motion and that of an applied electric field. As a result, a superconductor will support a time-varying voltage drop. The impedance associated with the supercurrent is therefore inductive, and it will be useful to think of $\mu_0 \lambda^2$ as a kinetic inductance created by the correlated motion of the Cooper pairs.

Substituting the first London equation into Faraday's law and integrating with respect to time yields the second London equation,

$$\nabla \times (\mu_0 \lambda^2 \mathbf{J}_s) = -\mathbf{B}, \tag{4}$$

where $\mathbf{B}$ is the magnetic flux density and the integration constant has been set equal to zero. This latter choice is necessary so that the second London equation is consistent with the Meissner effect. This can be demonstrated by combining Eq. (4) with Maxwell's equations to obtain the vector Helmholtz equation:

$$\nabla^2 \mathbf{B} - \frac{1}{\lambda^2} \mathbf{B} = 0. \tag{5}$$

Physically, this result means that a superconductor will screen out an applied flux on a length scale of λ. Equation (5) reveals that the penetration depth λ plays a role in a superconductor analogous to that of the skin depth δ in an Ohmic conductor. The crucial difference is that the penetration depth does not depend on frequency.

The penetration depth is a function of temperature because the Cooper pairs can be destroyed thermally. Fitting to experimental data on conventional (low-T_c) superconductors results in the phenomenological expression

$$\lambda(T) = \frac{\lambda_0}{\sqrt{1 - (T/T_c)^4}} \quad \text{for } T \leq T_c, \tag{6}$$

where λ_0 is the penetration depth at zero temperature. Since broken Cooper pairs become normal electrons, Eq. (6) reveals how the conductance of a superconductor should behave:

$$\tilde{\sigma}(T) = \sigma_c (T/T_c)^4 \quad \text{for } T \leq T_c, \tag{7}$$

where σ_c is the dc conductance of the material at the critical temperature. Thus we find that transport in the superconductor occurs from two interpenetrating fluids created by the normal and superelectrons. Since the normal and supercurrent channels of transport are parallel, the effective conductance of a superconductor in the sinusoidal steady state is

$$\sigma = \tilde{\sigma} + \frac{1}{j\omega\mu_0\lambda^2}, \tag{8}$$

where the explicit temperature-dependent notation has been suppressed. The real part of σ is a result of the Ohmic loss in the normal transport channel, and the imaginary portion of σ is a result of the kinetic inductance in the supercurrent transport channel.

It should be noted that the temperature dependences given in Eqs. (6) and (7) are not precisely correct for the high-T_c materials. It has been suggested that this is because the angular momentum of the Cooper pairs in high-T_c materials is different than that in low-T_c ones. Nevertheless, the two-fluid picture of transport and its associated constitutive law, Eq. (8), are still valid for high-T_c superconductors.

Equation (8) is the central constitutive relation for the general electromagnetic properties of a superconductor. Although the model is relatively simple, it contains the essential features for making many engineering decisions. For example, typical electromagnetic calculations can be performed in the usual manner by simply including an additional energy storage term accounting for the lossless kinetic motion of the superelectrons. Table 1 shows the inductance L_0 and resistance R_0 per unit length of a single superconducting plate. Here the skin depth is calculated from the conductance $\tilde{\sigma}$: $\delta \equiv \sqrt{2/\omega\mu_0\tilde{\sigma}}$. For contrast, Table 2 shows the same quantities for an Ohmic plate. Both tables show quantities intrinsically associated with the plates only; inductance arising from geometrical considerations is not included.

3. PASSIVE SUPERCONDUCTING DEVICES

In many applications, the advantage of using a superconductor is a direct result of its low-loss properties. This is most obvious

Table 1. Summary of relevant lumped-element parameters per unit length associated with a single superconducting waveguide plate of thickness b and width d. It is assumed that the current is traveling along the length of the plate.

Case	Inductance/length L_0	Resistance/length R_0
$\lambda \ll \delta$, b arbitrary	$\frac{\mu_0\lambda}{d}\coth\frac{b}{\lambda}$	$\frac{1}{d\delta\tilde{\sigma}_0}\left(\frac{\lambda}{\delta}\right)^3 \frac{\sinh(2b/\lambda) + (\delta/\lambda)^2 \sin(2b\lambda/\delta^2)}{\sinh^2(b/\lambda)}$
$\lambda \ll \delta$, $\lambda \gg b$	$\frac{\mu_0\lambda^2}{db}$	$\frac{4}{db\tilde{\sigma}_0}\left(\frac{\lambda}{\delta}\right)^4$
$\lambda \ll \delta$, $\lambda \ll b$	$\frac{\mu_0\lambda}{d}$	$\frac{2}{d\delta\tilde{\sigma}_0}\left(\frac{\lambda}{\delta}\right)^3$

Table 2. Summary of relevant lumped-element parameters per unit length associated with a single Ohmic waveguide plate of thickness b and width d. It is assumed that the current is traveling along the length of the plate.

Case	Inductance/length L_0	Resistance/length R_0
$\delta \gg b$	$\mu b/3d$	$1/db\sigma_0$
$\delta \ll b$	$\mu\delta/2d$	$1/d\delta\sigma_0$

in microstrip circuits, where the microwave devices are patterned thin films. Such devices are termed "passive" since the films merely guide the electromagnetic energy in the circuit. Of all the superconducting materials that have been successfully grown in film form, Nb is the most common low-T_c material found in applications, while $YBa_2Cu_3O_7$ is the most common high-T_c material. Relevant material properties of these superconductors are found in Table 3. Some care is required when making estimates on anisotropic materials like $YBa_2Cu_3O_7$. Since $YBa_2Cu_3O_7$ is typically grown with its CuO planes parallel to the substrate (c axis perpendicular to the substrate), electrical transport occurs in the a-b planes. Consequently, although the magnetic fields inside a superconducting microstrip device are spatially decaying along the c direction, the appropriate value of λ is that for the a-b plane, the direction of electrical transport.

To gain the full advantage of a superconducting film in a microcircuit device, the supercurrent channel must dominate the normal channel. In other words, it is primarily the superelectrons, rather than the normal electrons, that screen electromagnetic fields from the material: $\lambda \ll \delta$. Inspecting Tables 1 and 2 shows that when this criterion is satisfied, the resistance per unit length for a superconducting plate drops from that of an Ohmic plate by an impressive factor of $(\lambda/\delta)^4$ in the thin-plate limit and $(\lambda/\delta)^3$ in the thick-plate limit. Experiments confirm this. At 1 GHz, the surface resistance of Nb at 7.7 K and $YBa_2Cu_3O_7$ at 77 K is about 2×10^{-6} Ω, over two orders of magnitude lower than that of Cu at 77 K. Of course, at higher frequencies the surface resistance of the superconductors is larger since it scales as ω^2, as seen in Table 1. At 10 GHz, for example, the measured surface resistance of $YBa_2Cu_3O_7$ at 77 K is roughly 200×10^{-6} Ω, but this value is still quite small. In fact, even at frequencies of 100 GHz, the surface resistance of $YBa_2Cu_3O_7$ at 77 K is smaller than that of cooled Cu.

3.1 Microwave Filters

The loss of surface resistance immediately translates into device performance improvement. For example, the insertion loss of a superconducting microwave device is much lower than that of a similar device made from Ohmic materials. Filters with sharper bandpass skirts are possible using superconducting materials. This is because the number of poles in the transfer function of a microcircuit filter is directly related to the number of resonating pieces of microstrip in the device: The lower the loss per piece, the more pieces are possible. It is common to fabricate 19-pole filters out of $YBa_2Cu_3O_7$ films. These filters operate at a center frequency of 900 MHz with a bandpass of 25 MHz and create less than a 1 dB of insertion loss. Such filters can be extremely important in the cellular communication industry, where efficient use of an allocated frequency band without cross-talk to other channels is of central interest. The issues involved with having to cool the superconducting filters be-

Table 3. Typical material properties for both low- and high-T_c superconductors important for devices. These numbers should be taken only as a general guide, since growth and processing conditions can affect them. Because $YBa_2Cu_3O_7$ is anisotropic, it is necessary to specify λ and ξ for each crystallographic direction. The a and b directions lie in the CuO planes; the c direction is perpendicular to them.

Material	T_c (K)	λ at $T = 0$ (nm)	ξ at $T = 0$ (nm)
Nb (low-T_c)	9.25	85	40
$YBa_2Cu_3O_7$ (high-T_c)	95	150 (a-b plane) 1350 (c direction)	3 (a-b plane) 0.2 (c direction)

low T_c in cellular communication base stations may not be overwhelming, since it is desirable to use cooled low-noise amplifiers in such systems anyway. Moreover, the inconvenience of cooling is offset by the economic benefits of using the smaller, thin-film superconducting devices instead of bulky, nonsuperconducting microwave cavities.

3.2 Nuclear Magnetic Resonance Spectroscopy Coils

Obviously the reduced surface resistance raises the Q value for the superconducting microcircuit since the two quantities are inversely proportional. This fact has profound implications for high-resolution NMR spectroscopy. In typical systems, the way to increase signal sensitivity is to use larger samples, average out the sample response over more measurement iterations, or perform the measurements in ever higher magnetic fields (the best typical systems already use 9.4-T superconducting magnets). An economically viable alternative, however, is obtained by using $YBa_2Cu_3O_7$ superconducting coils both to excite the rf radiation into and to detect it from the sample. With superconducting coils that have Q values exceeding 10^4 at temperatures of 25 K, it becomes possible to demonstrate an increase of the signal-to-noise ratio of 4 times or more. Such superconducting NMR probes are commercially available and can be retrofitted into existing NMR systems. Again, the attendant cooling issues are not of central concern from either an economic or engineering viewpoint; such cooling can already exist for low-noise amplifiers in the system. The increased probe performance also justifies the cooling requirements.

3.3 Microwave Delay Lines

Microcircuit delay lines are another obvious application for passive superconducting devices. First, the power attenuation along a superconducting stripline, $2\alpha = R_0\sqrt{C_0/L_0}$ (C_0 is the capacitance per unit length), is greatly reduced from that of a Cu stripline. This allows for greater line lengths. More importantly, the signal phase velocity $u_p = 1/\sqrt{L_0C_0}$ is frequency independent in a superconducting plate, unlike that in a thick Ohmic plate. This is simply because the penetration depth, unlike the skin depth, does not depend on frequency. Thus electrical pulses moving on a superconducting line hold their shape better, again conducive for longer delay lines. In addition, the nondispersive nature of the line allows for tighter timing between successive incoming signals. Indeed, from an engineering viewpoint, the reduction of signal dispersion is as important as the reduction of surface impedance in many superconducting microstrip devices.

4. SUPERCONDUCTING WEAK LINKS

Because all the itinerant electrons in a superconductor exhibit correlated motion, the usual wave–particle duality normally associated with a single quantum particle can be applied to the entire ensemble of Cooper pairs. Thus, there is a spatiotemporal phase associated with the ensemble that characterizes the supercurrent flowing in the material.

If the overall electron correlation is broken, this phase is lost, and the material is no longer a superconductor. However, there is a broad class of structures, known as *weak links,* where the correlation is merely perturbed locally in space rather than outright destroyed. Colloquially, we say that the phase "slips" across the weak link to acknowledge the perturbation.

The unusual properties of this phase slippage were first investigated by Brian Josephson and constitute the central principles behind superconducting electronic devices. Josephson found that the phase slippage could be defined as the difference between the macroscopic phases on the two sides of the weak link. This phase difference, denoted as ϕ, is related to the supercurrent i_s through and the voltage v across the weak link according to the Josephson equations

$$i_s = I_c \sin\phi, \tag{9}$$

$$v = \frac{\Phi_0}{2\pi}\frac{d\phi}{dt}, \tag{10}$$

where I_c is the critical (maximum) current of the junction and Φ_0 is the quantum unit of flux. (The flux quantum has a precise definition in terms of Planck's constant, h, and the electron charge, e: $\Phi_0 \equiv h/2e \approx 2.068 \times 10^{-15}$ Wb.) Intuitively, the Josephson effect can be envisioned as a small overlap of the quantum-mechanical wavefunctions between the two weakly coupled superconductors, with I_c reflecting the strength of the coupling.

As in the case of the two-fluid model in a bulk superconductor, the correlated motion of the superelectrons, here represented through the phase ϕ, manifests itself as an inductance. This is demonstrated by taking the time derivative of Eq. (9) and combining it with Eq. (10). Although the resulting inductance is nonlinear (it depends on $\cos\phi$), its relative scale is determined by $L_j = \Phi_0/2\pi I_c$, a useful quantity for making engineering estimates. For example, the energy scale associated with Josephson coupling is $L_j I_c^2 = I_c \Phi_0/2\pi$.

4.1 Elementary Model of Weak Links

It is convenient to model a superconducting weak link in terms of elementary lumped circuit elements. The resistively and capacitively shunted junction (RCSJ) model, shown in Fig. 1, is relatively simple and yet accurate enough to understand most practical aspects of weak-link behavior. The flow of superelectrons through the device is described by a nonlinear inductance as determined from Eqs. (9) and (10). A resistor R_n describes the flow of normal electrons through the device. Finally, a capacitor C_j represents a physical capacitance in the system, such as in a tunnel junction where two superconducting electrodes are separated by a very thin (typically less than 2 nm) insulating layer. Parasitic capacitance created by the fields around a device interacting with a dielectric substrate is also included in this lumped element.

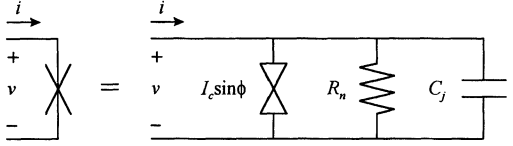

FIG. 1. The resistively and capacitively shunted junction (RCSJ) model of a Josephson device. If the shunted junction is envisioned as an inductance, many energy and time scales may be calculated as in elementary circuit theory. In a more sophisticated model, the resistive element would be nonlinear allowing for the subgap resistance to differ from that of the normal channel.

Because the weak link has been modeled from simple lumped elements, circuit intuition based on a knowledge of time constants is immediately applicable to understanding device behavior. For example, the quality factor of the circuit is

$$Q^2 = \frac{R_n C_j}{L_j/R_n} = \frac{2\pi I_c R_n^2 C_j}{\Phi_0} \equiv \beta_c, \tag{11}$$

where β_c is known as the Stewart–McCumber parameter. Clearly, if $\beta_c \gg 1$, the capacitive time constant $R_n C_j$ dominates the dynamics of the circuit. As the bias current is raised from zero, no time-average voltage is created until the critical current I_c is exceeded. At this point, the junction switches to a voltage consistent with the breaking of the Cooper pairs, $2\Delta/e$, with a time constant $\sqrt{L_j C_j}$. Once the junction has latched in the voltage state, however, the capacitor has charged up, and the only way for it to discharge is to lower the bias current to zero again. As a result, a device with $\beta_c \gg 1$ will have a hysteretic current–voltage curve as shown in Fig. 2(a). Conversely, $\beta_c \ll 1$ implies that the capacitance of the device is unimportant, and so the current–voltage curve is not hysteretic [see Fig. 2(b)]. In fact, the time-averaged voltage $\langle v \rangle$ for such a RSJ device is

$$\langle v \rangle = iR_n\sqrt{1 - (I_c/i)^2} \quad \text{for } |i| > I_c. \tag{12}$$

In other words, once the supercurrent channel carries its maximum amount of current, the rest of the current is carried through the normal channel.

4.2 Fabrication of Practical Weak Links

There are many ways to fabricate a weak link. For example, microbridges can be fabricated from a single superconducting film where an intentional constriction, characterized by a size that is not much larger than ξ_s, is lithographically patterned. Because the Cooper pairs cannot stay correlated as they pass through the small constriction, the phase slips. However, using such a device in a commercial application is not realistic be-

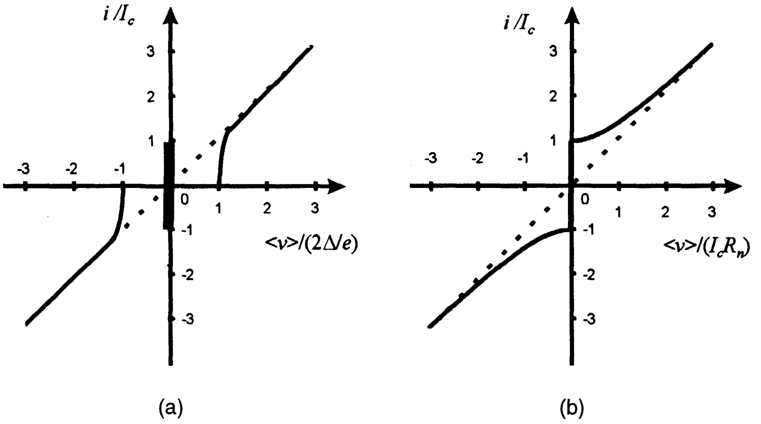

(a) (b)

FIG. 2. The current–voltage characteristics of ideal Josephson junctions. Notice that the time-averaged voltage is the plotted quantity. (a) An underdamped, hysteretic junction. (b) An overdamped nonhysteretic junction. In both cases, the dashed line indicates the resistance R_n associated with the normal electrons.

cause the coherence lengths of practical superconductors are typically less than 40 nm. Using present lithographic techniques, this length scale puts unrealizable demands on the fabrication control and uniformity of a group of microbridges. Fortunately, several other geometries produce more manufacturable devices.

4.2.1 Low-T_c Devices

The low-T_c device used in most applications is the Josephson tunnel junction. The device consists of two superconducting electrodes separated by a very thin (typically 2 nm or less) insulating layer, resulting in a superconducting–insulating–superconducting (SIS) structure. A common, robust fabrication scheme for such devices is to deposit an Al/Nb bilayer, oxidize the Al *in situ* to form the insulating tunnel barrier, and complete the junction by immediately depositing another layer of Nb for the counterelectrode. Typically the Nb films are about 250 nm thick. It is possible to deposit these trilayer sandwiches reliably with critical current densities of up to 10 kA/cm^2, although experimental trilayers with critical current densities of up to 400 kA/cm^2 have been demonstrated. The junctions are then formed by patterning the trilayer sandwiches either with reactive ion etching or selective anodization of the Nb counterelectrode into an insulating oxide.

Nb SIS devices can be described by the RCSJ model with negligible conductance in the subgap region defined by $\langle v \rangle < 2\Delta/e$. At 4.2 K, a standard target J_c is 1 kA/cm^2 so that devices that are several square microns in size will have critical currents on the order of tens of microamperes. The gap voltage for such Nb devices is about 2.9 mV. The subgap resistance, defined as the ratio of the voltage to current at 2 mV, is about $40R_n$ in a well-made device. The specific capacitance is typically 4.5 μF/cm^2. The value of I_cR_n is expected to be about 1.8 mV. Of course, with increasing J_c all these values will change. It is routine to fabricate quality Nb Josephson junctions on silicon wafers up to 4 in. (~10 cm) in size.

4.2.2 High-T_c Devices

In contrast to the Nb trilayer, the most reliable form of a high-T_c weak link is the relatively simple bicrystal grain-boundary junction. This structure is fabricated by simply growing an epitaxial high-T_c film on a bicrystal substrate such as strontium titanate. The bicrystal substrate is formed by fusing two single-crystal substrates together at a specified angle. The grain boundary that forms in the epitaxial film over the fused region is a natural weak link for the oxide superconductors. (Such a technique would not work with the typical low-T_c superconductor because the grain boundary formed would not be a weak link.) The critical current of a bicrystal junction can be tuned by varying the relative angle of the two fused crystalline substrates. For $YBa_2Cu_3O_7$ applications, the optimum angle is about 24°. At 77 K, a 2-μm-wide

$YBa_2Cu_3O_7$ bicrystal device suitable for applications typically has a critical current of 10 μA and a resistance of 10–20 Ω. As a result, I_cR_n is 100–200 μV. Experimentally, the scaling law $I_c \propto R_n^{-2}$ is observed in high-T_c grain-boundary devices.

Although quite reliable from a production standpoint, the bicrystal junction can only be used in limited applications that involve superconducting devices that are linearly distributed along the grain boundary in the substrate. Moreover, any superconducting wiring that crosses the grain boundary would itself contain a weak-link device embedded in it. As a result, there has been a substantial effort to develop a different high-T_c device technology that is more suitable for circuit applications. The inherently short coherence length of the high-T_c materials makes them unlikely candidates for fabricated SIS junctions because the device properties are sensitive to the superconducting electrode properties within a distance of ξ_s of the barrier. As a result, it becomes difficult to control the quality of the important superconductor–insulator interface in the device. In addition, the angular momentum of a Cooper pair in the high-T_c materials may prevent the type of SIS tunneling expected from low-T_c devices.

These considerations have pushed workers to develop instead other types of high-T_c weak links, especially superconducting–normal–superconducting (SNS) devices, for circuit applications. The fundamental physics of both SIS and SNS junctions are similar: They both represent weak-link devices that operate because of overlapping superconducting wave functions from the two electrodes. There is an important difference, however: In a SIS junction, Cooper pairs must tunnel through a classically forbidden barrier (the insulator), while in a SNS device, Cooper pairs can be found for a time inside the metallic interlayer. Depairing in the normal metal occurs because the Fermi surface is smeared out by an energy kT. As a result, superconductivity is induced in the normal metal and the respective clean- and dirty-limit coherence lengths are given by

$$\xi_{nc} = \hbar v_F / 2\pi kT, \qquad (13)$$

$$\xi_{nd} = \sqrt{\xi_{nc} l / 3}. \qquad (14)$$

It is no coincidence that these expressions are similar to Eqs. (1) and (2), for the physics is essentially the same. The proximity of the normal material to the superconductor allows for a supercurrent channel to exist in the N interlayer. In general, the normal material will be in neither the clean nor the dirty limit. In such a case, the coherence length in the N interlayer is given by $\xi_n^{-2} \cong \xi_{nc}^{-2} + \xi_{nd}^{-2}$. Notice that the true value of ξ_n is always less then either the extreme limits given by ξ_{nc} and ξ_{nd}. Typical numerical values of the normal coherence length range from a few nanometers to about a micron depending on material and temperature.

High-T_c SNS devices are often made using an edge junction geometry. First, an epitaxial $YBa_2Cu_3O_7$ base electrode, typically 200 nm thick, is deposited and then covered by some insulator, often strontium titinate or yttria-stabilized zirconia. This bilayer is then patterned by ion milling. The patterning not only defines the base electrode but also forms the active edge of the device (the exposed edge of the $YBa_2Cu_3O_7$ film) because the incidence of the ion beam is approximately 60° from the normal of the film. A normal metal/$YBa_2Cu_3O_7$ counterelectrode bilayer is then deposited and patterned, completing the structure as shown in Fig. 3. Because the Cooper pairs are not classically forbidden from the normal material, the interlayer thickness of a SNS device can be much greater than that of a SIS device. A useful SNS device can have an interlayer thickness on the order of tens of nanometers, which relaxes fabrication requirements dramatically. This is the central appeal of developing SNS junctions for use in high-T_c circuits.

The conventional theory of the proximity effect in a SNS junction is well tested for low-T_c devices. It predicts a critical current that has a dramatic, exponential-like increase with decreasing temperature and is exponentially

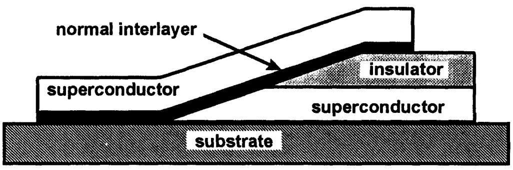

FIG. 3. SNS edge junction structure.

dependent on the thickness of the interlayer. Both of these observations hold in the clean and dirty limits and are a result of overlapping of the exponentially decaying superconducting wavefunctions in the N region. However, I_c of most high-T_c SNS junctions does not show the basic functional dependences predicted by conventional theory. For example, $I_c(T)$ in these devices is similar to the quasilinear behavior of grain-boundary junctions. This is most likely a result of pinholes in the N interlayer rather than an inappropriate application of proximity theory to high-T_c devices. The observation is justified by noting that if substituted forms of $YBa_2Cu_3O_7$, such as $Y_xCa_{1-x}Ba_2Cu_3O_7$ and $YBa_2Cu_{3-y}Co_yO_7$, are used as the N interlayer, then I_c as a function of temperature and interlayer thickness is reasonably consistent with conventional proximity effect theory. As a result, it appears that the N interlayer must be well matched structurally to the superconducting electrodes to grow properly and create a reproducible high-T_c SNS device.

Promising high-T_c SNS junctions for circuit applications can be made with an interlayer of $YBa_2Cu_{2.79}Co_{0.21}O_7$, typically 15 nm thick. Because this N material is a superconductor with $T_{cn} \approx 57$ K, the operating temperature of these SNS devices is restricted from below: $T_{cn} < T < T_c$. Typical critical current densities at 65 K for quality devices are 1.3×10^5 A/cm^2 so that junctions that are several microns wide have critical currents on the order of a milliampere. Corresponding specific resistances are approximately $4.8 \times 10^{-9}\ \Omega\ \text{cm}^2$. A large specific resistance is important because the devices should have I_cR_n values on the order of 500 μV at 65 K to be useful for circuit applications.

5. ACTIVE SUPERCONDUCTING DEVICES

Generally speaking, while we may think of passive superconducting devices as associated with microwave structures, active superconducting devices are usually associated with electronic applications. Moreover, active superconducting devices exploit the special quantum nature of the Cooper pairs more directly than passive ones. Thus the general applications of active superconducting devices are not merely an extension of existing technology but involve new concepts altogether.

5.1 Superconducting Quantum Interference Device

Just as the correlated motion of the superconducting electrons creates the frequency-independent Meissner effect in a bulk superconductor, so too the macroscopic quantum nature of superconductivity allows the possibility of a device whose output voltage is a function of a static magnetic field. If two weak links are connected in parallel, the lumped version of Faraday's law gives the voltage across the second weak link as $v_2 = v_1 + (d\Phi/dt)$, where Φ is the total flux threading the loop between the links. Substituting Eq. (10), integrating with respect to time, and again setting the integration constant equal to zero yields

$$\phi_2 - \phi_1 = 2\pi\Phi/\Phi_0, \tag{15}$$

showing that the spatial change in the phase of the macroscopic wavefunction is proportional to the local magnetic flux. From this discussion, it is apparent that a duality exists in how fields interact with the macroscopic phase: Electrical fields are coupled to its rate of change in time and magnetic fields are coupled to its rate of change in space.

Quantum interference, as represented by Eq. (15), can be exploited to make useful structures. Consider a dc superconducting quantum interference device (dc SQUID), shown in Fig. 4, which consists of two identical overdamped ($\beta_c < 1$) Josephson junctions connected in parallel and with wires into and out of the loop so that a biasing current i_{bias} can be injected symmetrically into the junctions. The current–voltage characteristics of the SQUID are identical in shape to those of the individual overdamped junctions with the critical current of the SQUID twice I_c of the junctions and the normal state resistance of the SQUID half R_n of the junctions. Applying a magnetic field to the SQUID induces a circulating current in the loop that will add to the dc bias current for one of the junctions and subtract from the dc bias current for the other one. As a result, the applied magnetic field will cause

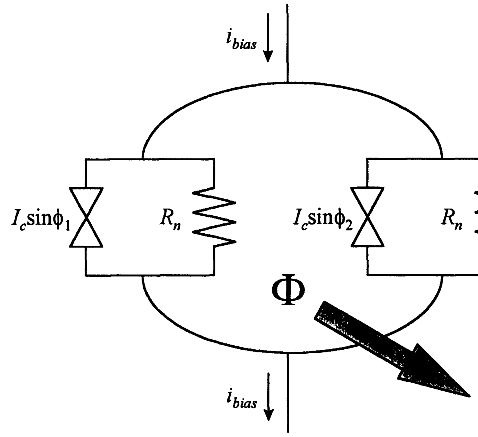

FIG. 4. A dc SQUID used to sense magnetic flux. The actual junctions are overdamped and have negligible capacitance as indicated in the model. The flux Φ threads the superconducting loop.

the junction where the induced and applied currents are adding to switch at a lower bias current. The magnetic field lowers the critical current of the SQUID from its maximum value of $2I_c$ while not influencing its normal-state resistance of $R_n/2$.

From Eq. (15) and the Josephson relations, we see that this magnetic-field modulation of the SQUID current–voltage characteristic is periodic in units of the flux quantum. Thus, the phase difference between the two junctions, and hence the SQUID critical current, is the same for all applied flux $\Phi + n\Phi_0$ where n is an integer. If we current bias the SQUID just higher than $2I_c$, then the change in magnetic field will produce a corresponding change in voltage across the SQUID. The SQUID is a magnetic flux-to-voltage transducer with a maximum peak-to-peak voltage modulation given by

$$\Delta\langle v\rangle = \alpha I_c R_n/(1 + \beta_L), \qquad (16)$$

where α is a constant of order unity, $\beta_L \equiv I_c L/(\Phi_0/2)$, and L is the SQUID loop inductance. Because the Φ_0–$\langle v\rangle$ transfer function is periodic in units of the flux quantum, a SQUID is an extremely sensitive magnetometer with an extraordinary dynamic range.

There are some engineering rules useful when designing a dc SQUID. First, the energy associated with the junctions should be much greater than that associated with thermal fluctuations to prevent unwanted switching: $I_c\Phi_0/2\pi \gg kT$. In other words, the junction critical current should be greater than 0.2 μA at 4.2 K and 3 μA at 77 K. Second, the energy of half a flux quantum in the loop should be greater than the thermal energy to prevent unwanted magnetic fluctuations: $(\Phi_0/2)^2/L \gg kT$. As a result, the inductance should be less than 18 nH at 4.2 K and 1 nH at 77 K. Note, however, that if L becomes too small, the SQUID will not couple to the applied field efficiently.

These criteria can be combined to reveal that in a well-designed SQUID, we expect $\beta_L \approx 1$. Usually L is chosen to satisfy this condition after I_c has been identified. To obtain the maximum voltage modulation, we see from Eq. (16) that we wish to maximize I_cR_n. There are two upper limits on this quantity, however. We do not want R_n so big as to become a major noise source in the SQUID. In addition, we do not want the product of I_cR_n and R_n so big as to violate the overdamped criterion $\beta_c < 1$.

In practice, SQUIDs are normally used in a feedback mode where their output is sent to a flux-locked loop, which in turn generates a current to create a small field to null the applied field. Such a scheme allows for accurate measurement of fluxes to less than $10^{-6}\Phi_0$ for a 1-Hz bandwidth. If the applied field is generated by a small current, it is then possible to use the SQUID as a precision voltmeter, capable of measuring picovolts. Such applications have been successfully demonstrated, and commercial SQUID systems are available in both low- and high-T_c technology. At present, magnetometers made from Nb SQUIDs operating at 4.2 K have an order of magnitude better field-noise performance than those based on $YBa_2Cu_3O_7$ bicrystal SQUIDs operating at 77 K. Nevertheless, $YBa_2Cu_3O_7$ magnetometers have improved in noise performance by three orders of magnitude since the discovery of the high-T_c materials and may soon provide a performance similar to that of Nb magnetometers.

5.2 Logic Circuits

Josephson devices have two properties that make them highly desirable for digital circuit applications: their switching speed is quite high, and their power dissipation is quite low. Originally, these properties motivated researchers to develop full-scale digital computers based on superconductor logic. The significant gains made in semiconductor technology since that time, however, have made general-purpose superconducting digital computers economically impractical. Nevertheless, niche high-speed applications, such as wide-band signal processing, and low-power applications, such as cross-bar switches, exist where superconducting circuits can still play an important role.

5.2.1 Latching Logic The first serious attempts to build digital circuits out of Josephson devices used underdamped junctions ($\beta_c \gg 1$). This choice is not surprising since the ideal hysteretic junction can go from a zero-voltage state, where all the current is carried by the superelectrons, to a finite-voltage, zero-current state, where the capacitor is charged and little current flows through the device. Thus we can envision a hysteretic junction as a switch: a logical 0 for the finite-current, zero-voltage state and a logical 1 for the finite-voltage, zero-current state.

The operation of the switch is straightforward. A bias current is applied to the device, $i_{\text{bias}} < I_c$. When an additional current signal flows through the junction, the critical current is exceeded, and the junction switches into the voltage state. By design, the load resistor on the junction is small enough that the switched state has minimal current. The switching time is estimated from the lumped-element model; first the current is steered from the inductor to the capacitor, and then the capacitor charges up. As a result, $\tau_{\text{switch}} \approx \sqrt{L_j C_j} + R_n C_j$. For a Nb SIS junction with a J_c of 1 kA/cm^2, the switching time is 9 ps. Higher values of J_c will increase the switching speed further. Resetting the device to its initial state requires a reverse current bias. Notice that some "amplification" occurs in the device because, if the bias current is close to the critical current, it will only take a small signal current to switch the junction. The power dissipated by the device is on the scale of $I_c^2 R_n$ and so is typically about a microwatt in a Nb device.

The central problem with latching logic is that the gates must be reset. The practical way to reset the devices in synchrony is to make the bias currents a common rf signal. Now, however, the designer must deal with clock skew; that is, the finite propagation speed of the clock signal to various parts of the chip. More importantly, the rf bias can interact in an unfavorable way with the internal nonlinear dynamics of the device itself (for example, those associated with the time $\sqrt{L_j C_j}$). This can leave the device at the end of the reset cycle with its capacitance charged in the reverse voltage state, an effect known as "punch through." To prevent this type of coupling, latching circuits typically are limited to clock speeds of a few gigahertz. This frequency is not high enough to make latching logic for general-purpose computing economically competitive with various present semiconductor technologies.

5.2.2 Single–Flux-Quantum Logic Overdamped, nonhysteretic junctions ($\beta_c < 1$) provide a way out of the reset problem because there is no significant capacitance to influence the device dynamics. Consider an overdamped junction, again current biased just less than I_c. A small signal current will switch the device into a finite-voltage state as before. Now, however, removal of the signal current will leave the junction in the initial superconducting, zero-voltage state. In a sense, the overdamped device is self-resetting.

Consider the dc SQUID. The time it takes for a single-flux quantum to pop into or out of the superconducting loop is equal to the switching time of the overdamped junctions. It takes little effort to envision the dc SQUID as a flux switch: a logical 0 for a loop containing no flux quantum and a logical 1 for a loop containing a quantum unit of flux. Thus we can view a logic based on overdamped junctions in terms of voltage pulses running around a circuit or, equivalently, as a circuit shuttling around flux quanta. The latter perspective lends its name to the architecture. In single–flux-quantum (SFQ) logic, the dc SQUID is the key lumped element.

SFQ circuits are typically built from Nb devices. Because standard Nb SIS junction processing is so robust, overdamped devices are made by externally shunting a SIS device. The Nb circuit is expected to operate at 4.2 K, and so a critical current of about 100 μA is chosen to minimize the effect of thermal fluctuations. The shunt resistance R_s is typically 1 Ω, much less than the value of about 18 Ω for R_n in such a device. From the lumped-element model, the overdamped junction switching time is about L_j/R_s or 3 ps in this case. As with the hysteretic device, τ_{switch} is quite fast. However, because the junction need not be reset, it is possible to take full advantage of this short switching time. Consequently, ultrahigh-speed performance on the order of 100 GHz could be possible. Moreover, the voltage-pulse nature of operation yields a power dissipation per device of about 100 nW. SFQ logic is therefore attractive because it is a low-power application that can be operated with dc, rather than rf, bias currents.

The dc SQUID loops found in SFQ logic are designed as previously described. For the Nb-based circuits discussed, a reasonable choice of L is about 15–20 pH. In general, to ensure that each SQUID loop contains either zero or one flux quantum, the I_cL product should be greater than Φ_0 but less than $2\Phi_0$.

SFQ logic is a natural choice for circuits using high-T_c materials for two reasons. First, latching logic does not scale well with temperature. As mentioned above, the required critical currents in an electronic application will scale linearly with temperature so that the devices will not switch because of thermal fluctuations. In latching logic, this implies that the power dissipated per device will scale as T^2, assuming a temperature-independent value of R_n. As a result, latching logic operating at 65 K will dissipate 200 times more power per gate than that at 4.2 K. In contrast, the power dissipation in a SFQ circuit only rises as T. Notice that the average voltage per cycle is determined by the nearly temperature-insensitive speed at which the flux quantum pops in and out of the superconducting loop. Thus, the average power dissipated by the overdamped junction will only be significantly affected by the size of critical current, which scales as T. The second reason why designers use SFQ architectures to implement high-T_c circuits is that only overdamped high-T_c devices can be fabricated as discussed above.

The higher operating temperature of high-T_c SFQ circuits, which makes them attractive, also causes problems. At 65 K, for example, I_c should be about 2 mA to minimize thermal fluctuations. This requires the SQUID inductance L to be about 0.5 pH if the condition $\beta_L \approx 1$ is to hold. The inductance per square of a $YBa_2Cu_3O_7$ stripline at 65 K, however, is only about 0.25 pH/square as can be estimated from Table 1. Thus, making a SQUID element for SFQ logic out of $YBa_2Cu_3O_7$ is not trivial since only a few squares of stripline can used for the loop. Practical designs will require engineering compromises, such as relaxing the choice of the magnitude of I_c, and engineering solutions, such as clever layouts. For instance, in video applications where ultralow bit error rates are not required, some thermal fluctuations can be tolerated and the magnitude of the critical current can be lowered. An example of a clever layout is to fabricate the SQUID loop out of two layers of $YBa_2Cu_3O_7$, one on top of the other, with a very thin insulating film between them.

Useful computational functions, such as high-speed sampling, multiplexing, and decimation, have been designed and fabricated in Nb-based SFQ technology. A special problem with such experiments involves performing the circuit validation tests at high speeds. Often, the designer will measure the circuit at some more modest speed, say 10 kHz, and check this measurement against a simulation. Nevertheless, experiments involving simple Nb SFQ circuits such as shift registers suggest that operation at frequencies of 60 GHz is possible. SFQ logic involving high-T_c devices is in its infancy, as the device fabrication technology has only just become available.

5.3 Three-Terminal Devices

Superconducting weak links are two-terminal devices and therefore cannot truly provide gain. Consequently, the use of two-terminal devices in an integrated circuit puts restrictions on fabrication tolerances of key parameters (such as I_c) that are more severe than those usually encountered in the semiconductor industry. There has thus been a strong desire to develop a superconducting three-terminal device that can deliver gain, high switching speed, and low power dissipation.

One class of three-terminal devices is analogous to bipolar transistors and treats the superelectrons and normal electrons as majority and minority carriers, respectively. In these devices, the superconductor would form the base of the transistor with normal metals or semiconductors forming the emitter and collector. Both ballistic and diffusive transport of the normal (minority) carriers in the superconducting base have been examined. Although such devices present major fabrication challenges, they can theoretically allow switching speeds approaching 1 THz.

Superconducting field-effect transistors form a second class of devices. Because the carrier density of high-T_c materials is relatively low, applied electric fields are not as easily screened as in ordinary metals. Consequently, one might expect a field-effect device to be attractive for high-T_c applications. Although gate-induced changes in transport properties have been demonstrated, thus far they are too small to warrant building a circuit from such elements.

Finally, there is the vortex transistor, which is unique to superconductivity. Here, the carriers are flux quanta that can be injected into the device from the magnetic field surrounding a separate current-carrying control line. As a result, flux can be independently introduced, for example, into a parallel array of overdamped junctions. When a bias current is applied to the array, the flux quanta will be shuttled from one end to the other, similar to the behavior in a SFQ shift register. Both the number of quanta moving through the array (given by the control line) and the speed with which they move (given by the bias current) determine the voltage across the device according to Faraday's law. These types of devices necessarily have a large area, thereby precluding their use as a common element in large-scale integrated circuits. Fortunately, most proposed applications of superconducting three-terminal devices do not require high integration density.

5.4 Mixers

In both space and atmospheric science, chemical constituents are typically identified from the electromagnetic spectrum created by molecular and atomic resonances. These spectra often have radiation lines with wavelengths of millimeters or less. For example, water will radiate at 325 GHz, carbon monoxide at 346 GHz. At the extreme end of the submillimeter spectrum there is an important radiation line associated with chemical processes involving atmospheric ozone at 2.5 THz. To detect such signals, it is desirable to mix the chemical spectral signal down to a more convenient frequency where traditional electrical components operate. A mixer is simply a nonlinear element that takes the incoming signal at a frequency ω_s and multiplies it with a local oscillating signal at a lower frequency ω_{lo} to produce a beat signal at an intermediate frequency $\omega_{if} = |\omega_s - \omega_{lo}|$.

The two charge-carrier channels in a superconducting device give rise to two classes of junction mixers. The first involves the interaction of the superelectrons (Cooper pairs) with the incident radiation to be downconverted. Suppose a dc bias voltage v_{bias} is applied to a Josephson weak link. Energy from rf radiation incident on the device can be used to assist the transfer of Cooper pairs from one side to the other. If a photon is absorbed by the junction, a Cooper pair can be moved from the low-bias side of the device to the high-bias side. A photon can also be emitted from the device, stimulated by the incident radiation. In this case, a Cooper pair is assisted toward the low-bias side of the junction. This photon-assisted transport occurs only at integral multiples of the Josephson frequency,

$$n\hbar\omega_J = 2ev_{bias}, \tag{17}$$

where, numerically, $\omega_J/2\pi v_{bias} = 483.6$ GHz/mV. Notice that the integer variable n can be either positive or negative corresponding to photon absorption or emission. The same relationship can be directly obtained from the Josephson relations, Eqs. (9) and (10), where a dc bias voltage will cause the magnitude of the supercurrent to oscillate at ω_J. Because Eq. (17) involves only fundamental physical constants, it provides a convenient and exceptionally accurate way to define a voltage standard on the basis of frequency alone.

From this simple picture, it is clear that rf radiation at a frequency ω_s will boost the supercurrent transport at bias voltages of $n\hbar\omega_s/2e$. Suppose that additional radiation, created by a local oscillator at a frequency ω_{lo}, is also incident on the device. The photons from this radiation will also assist Cooper pair transport, and so now the enhanced supercurrent transport occurs at bias voltages of $(n\hbar\omega_s + m\hbar\omega_{lo})/2e$, where m is an integer. Now consider a process in which a photon at frequency ω_s is absorbed at the same instant as a photon at frequency ω_{lo} is emitted. The supercurrent boost will come at a dc voltage associated with the intermediate frequency $\omega_{if} = \omega_s - \omega_{lo}$; the two rf sources have been mixed. To increase the probability of this particular photon process, the number of photons at ω_{lo} is increased by increasing the power of the local oscillator.

The enhanced-supercurrent mixing requires an overdamped device. Eliminating the effects of the junction capacitance prevents the incident rf radiation from coupling to the internal device resonances associated with the L_jC_j dynamics that would spoil the mixing. High-T_c weak links are a natural choice for Josephson mixers: Not only are they inherently overdamped but also they have displayed the highest I_cR_n values of any Josephson device ever made, exceeding 10 mV at 4.2 K. Indeed, the possibility of performing mixing at frequencies approaching 5 THz has motivated the application of high-T_c device technology in this area.

The second class of junction mixers involves the interaction of the normal electrons with the incident radiation. Again the quantum picture of photon-assisted transport is useful. Consider a SIS junction that has its supercurrent channel totally suppressed by a magnetic field. From the BCS theory, the occupied (unoccupied) energy states for the normal electrons are Δ below (above) those of the Cooper pairs in each of the superconducting electrodes. Incident radiation now assists the normal-electron transport across the energy gaps. As with the supercurrent case, the simultaneous absorption of photons at one frequency and emission of photons at another gives the nonlinear mixing effect. For negative biasing of the device to nearly $-2\Delta/e$, the maximum mixing frequency to assist normal carrier transport is $4\Delta/h$. Notice that again the incident rf radiation cannot couple to the internal device resonances associated with the L_jC_j dynamics because the inductance associated with the supercurrent has been suppressed.

SIS mixers have extremely good sensitivity and noise performance, even when compared to competing technologies. For example, Schottky diodes are nonlinear elements that can be used to mix signals in the same wavelength regime. However, SIS devices display sharper nonlinearities, as shown in the normal electron branch of Fig. 2(a). As a result, a superconducting mixer can provide greater sensitivity than a semiconductor one. In addition, superconductor mixers require much less local oscillator power. At $\omega_{lo}/2\pi \approx$ 100 GHz, a SIS mixer requires only about 10 nW of local oscillator power, about six orders of magnitude less than that for an equivalent semiconductor mixer.

For these reasons, Nb SIS mixers can be found on radio telescopes, such as the one in Owens Valley, California, and on high-altitude experiments where frequency detection in the range of 100–800 GHz is desired. (Cooling constraints are not an issue for these scientific applications.) At present, there is research activity to push toward the theoretical limit in Nb devices of 1.4 THz. Nb SIS mixers cooled to 4.2 K have been operated at 1 THz with a noise temperature 3 times better than that of GaAs Schottky diodes at the same frequency. Other ways of increasing the maximum mixing frequency is to use a low-T_c NbN SIS junction as the nonlinear element, which would theoretically allow mixing up to 2.2 THz. Indeed, the characteristics of low-T_c SIS mixers are so good that only their intrinsic frequency limit of $4\Delta/h$ makes high-T_c Josephson mixers attractive at all.

Finally, it should be noted that superconducting bolometers are also candidates for low- and high-T_c mixer applications. These devices are kept at temperatures just below T_c so that an incoming photon heats the superconductor up through its resistive transition. The corresponding large change in resistance is then detected. For small enough devices, the inherently nonlinear response of a bolometer can be made fast enough to be used in heterodyne applications. By definition, high-T_c devices have the advantage of operating at relatively high temperatures. As a result, a high-T_c bolometric mixer may be particularly useful when power constraints limit local oscillator power and prohibit cooling to ultralow temperatures, as on a satellite.

6. SUMMARY

Superconducting materials can be used to improve present systems dramatically by reducing loss and increasing system sensitivity. This is especially true with passive devices, such as microcircuit components. In contrast, active devices, usually formed from weak links, exploit the macroscopic quantum nature of superconductivity. Applications using active devices typically involve new concepts like superconducting quantum interference. In specific computational applications where ultrahigh speed is required, superconducting circuits may be economically competitive with semiconductor implementations. Although sophisticated circuit logic and ultrahigh-frequency mixing has been demonstrated using low-T_c materials only, high-T_c materials are the preferred choice for microwave applications. Indeed, commercial NMR spectroscopy probes built around $YBa_2Cu_3O_7$ coils are commercially available, and it appears that high-T_c superconducting microwave components may play a role in wireless communication systems. Finally, commercially available SQUID systems exist in both low- and high-T_c technologies.

In a brief article, it is only possible to hint at the many possible applications of superconductivity and the many issues involved in reducing those possibilities to practice. There are a number of important topics that are not covered in this article, including the voltage standard (see CONSTANTS, FUNDAMENTAL; UNITS), oscillators, bolometers, and applications of SQUIDs. In addition, device materials and fabrication is a constantly evolving subject. The reader is therefore strongly encouraged to consult the books and review articles listed below to appreciate applied superconductivity more fully. In addition, the peer-reviewed literature should also be examined, as superconductivity remains an active research field for both scientists and engineers.

Further Reading

Barone, A., Paterno, G. (1982), *Physics and Applications of the Josephson Effect,* New York: Wiley. Textbook covering all the basics of Josephson devices. Discusses some advanced topics with many references to the scientific literature prior to 1982.

Costabile, G., Pagano, S., Pedersen, N. F., Russo, M. (Eds.) (1991), *Nonlinear Superconductive Electronics and Josephson Devices,* New York: Plenum Press. Conference proceedings covering topics in detail not often found in other books. Extended articles on noise in junctions, particle detectors, oscillators, and low-noise SQUIDs. Both low- and high-T_c materials are discussed, although not necessarily in each article.

Delin, K. A., Kleinsasser, A. W. (1996), *Supercond. Sci. Technol.* **9,** 227–269. Comprehensive review of high-T_c SNS device development from an integrated physics and materials viewpoint. Includes a distillation of important theoretical results necessary for interpreting experimental data.

Fujitsu Limited (1991), *Fujitsu Sci. Technol. J.* **27,** 1–83. Special issue containing review articles on Josephson integrated circuits, especially as practiced at Fujitsu. Useful for examining the progress made in low-T_c device fabrication and superconducting latching logic.

IBM Corporation (1980), *IBM J. Res. Develop.* **24,** 105–264. Special issue devoted to the Josephson computer technology IBM was developing at the time using low-T_c materials. Both fabrication and circuit architecture issues are discussed.

Knuutila, J., Kajola, M., Seppä, H., Mutikainen, R., Salmi, J. (1988), *J. Low Temp. Phys.* **77,** 369–392. Practical guide for designing a low-T_c dc SQUID. Includes engineering and optimization details often not found in other review articles.

Likharev, K. K. (1979), *Rev. Mod. Phys.* **51,** 101–159. Classic review article on the physics behind weak-link devices.

Likharev, K. K. (1986), *Dynamics of Josephson Junctions and Circuits,* Philadelphia, PA: Gordon and Breach Science Publishers. Classic advanced monograph. Focuses more on devices than circuits.

Likharev, K. K., Semenov, V. K. (1991), *IEEE Trans. Appl. Supercond.* **1,** 3–28. One of the first review articles focused solely on single–flux-quantum logic.

Mannhart, J. (1996), *Supercond. Sci. Technol.* **9,** 49–67. Recent comprehensive review of progress in high-T_c three-terminal devices.

Newman, N., Lyons, W. G. (1993), *J. Supercond.* **6,** 119–160. Review article concerning the physics, materials, and engineering issues associated with high-T_c passive microwave devices.

Orlando, T. P., Delin, K. A. (1991), *Foundations of Applied Superconductivity,* Reading, MA: Addison-Wesley. Introductory textbook on superconductivity written especially for engineers. Focuses on simple modeling to apply superconductivity to useful situations. Covers both passive and active devices. Specifically includes issues associated with the high-T_c materials, such as anisotropy.

Richards, P. L. (1994), *J. Appl. Phys.* **76,** 1–24. Review article describing bolometric mixers, both semiconducting and superconducting.

Ruggiero, S. T., Rudman, D. A. (1990), *Superconducting Devices,* Boston, MA: Academic Press. Series of thorough review articles covering nearly all topics usually associated with superconducting devices: SQUIDs, oscillators, mixers, passive devices, and three-terminal devices. Many articles address both low- and high-T_c material issues.

Simon, R. W., Burns, M. J., Colclough, M. S., Zaharchuk, G., Cantor, R. H. (1996), *Mr. SQUID*

User's Guide Version 6, Sunnyvale, CA: Conductus, Incorporated. Practical guide to high-T_c SQUIDs and SQUID circuits. Includes circuit diagrams for building flux-locked loops.

Tucker, J. R., Feldman, M. J. (1985), *Rev. Mod. Phys.* **57,** 1055–1113. Advanced review article of low-T_c SIS mixers used in the millimeter wavelength regime. Covers both the physics and engineering aspects of the subject.

Van Duzer, T., Turner, C. W. (1981), *Principles of Superconductive Devices and Circuits,* New York: Elsevier North Holland. Introductory text on superconductivity. Covers both passive and active devices. Provides more coverage of traditional superconducting circuit design than found in comparable texts on this level.

Van Duzer, T., Taylor, C. E. (Eds.) (1989), *Proc. IEEE* **77,** 1107–1287. Special issue devoted to superconductivity. Excellent resource for low-T_c circuit design issues.

Weinstock, H., Nisenoff, M. (Eds.) (1989), *Superconducting Electronics,* New York: Springer-Verlag. Series of review articles covering SQUIDs and their uses, mixers, oscillators, circuits constructed from latching logic, and fabrication of low-T_c devices. Most articles do not address issues in the context of high-T_c materials.

Weinstock, H., Ralston, R. W. (Eds.) (1993), *The New Superconducting Electronics,* Boston, MA: Kluwer Academic Publishers. Sequel to the above reference that not only discusses progress made but also often focuses on the impact of high-T_c materials to this field. Additional topics include novel superconducting active devices, analog-to-digital converters, single–flux-quantum logic, and high-T_c device fabrication.

Every two years, an Applied Superconductivity Conference is held devoted to practical technological issues. The proceedings of these conferences have been published every other year from 1977 to 1991 in the *IEEE Transactions on Magnetics.* From 1993 onward, the proceedings are published in the *IEEE Transactions on Applied Superconductivity.*

SUPERCONDUCTING MAGNETS

Alvin V. Tollestrup, *Fermi National Accelerator Laboratory, Batavia, Illinois, U.S.A.*

Introduction 195
1. **Properties of Superconductors** 196
1.1 Type I and Type II Superconductors 197
1.2 Load Line for a Magnet 198
2. **Conductor Design** 199
2.1 Flux Jumps 199
2.2 Eddy-Current and Hysteresis Losses 199
2.3 Conductor Fabrication 200
2.4 NbTi Conductor 200
2.5 Nb_3Sn Conductor 201
2.6 Cryogenic Stability and Quench Waves 201
2.7 Conductor Splices 202
3. **Large Solenoids** 202
4. **Accelerator Magnets** 204
4.1 Coil Configuration 204
4.2 Forces 205
4.3 Magnet Fabrication 205
4.4 Training 207
5. **Quench Protection** 207
6. **Cooling the Magnet** 208
7. **Current Leads** 208
8. **Future Directions** 209
8.1 High-Temperature Superconductors 209
Glossary 210
Works Cited 210
Further Reading 211

INTRODUCTION

Superconductivity was discovered in 1911 by H. Kamerlingh Onnes, but it has taken nearly three-quarters of a century to make practical use of the phenomenon. The early materials investigated, mercury, lead, and tin, are type I superconductors and have a limited ability to carry high currents in the presence of a magnetic field. However, in 1930 type II superconductors were discovered by de Haas and Voogd. These materials showed promise of carrying very large current densities even in the presence of large magnetic fields. And over the following 30 years practical superconductors have been developed that can carry current densities of more than 1000 A/mm^2 in fields of more than 5 T. Equally important was the discovery of the correct theory of the phenomenon by Bardeen, Cooper, and Schrieffer in 1957, which placed the experimental facts on a firm theoretical basis. Commercially available conductors now make practical a wide variety of magnets useful for magnetic resonance imaging (*q.v.*) (MRI) systems, particle accelerators, thermonuclear machines, magnetic separators, and the levitation system for high-speed transportation. The operating temperature of these modern materials is generally about 4 K, and the cooling is provided by liquid helium (LHe). Some applications even make use of the high heat conductivity of superfluid helium below its lambda point. The delay in the deployment of superconductivity was undoubtedly tied to the difficulties in the metallurgy of the conductor on the one hand and the development of large cryogenic helium plants on the other hand.

The discovery of high-temperature superconductors (HTSs) in 1986 (see SUPERCONDUCTIVITY, HIGH-TEMPERATURE) has led to many exciting new prospects but, as before, practical conductor fabrication is facing many difficult problems. The major impact of HTSs will come about as the theory is developed and the practical problems are solved of fabricating these brittle, anisotropic materials into conductors possessing high current density and with mechanical properties suitable for the fabrication of magnets.

There are two important aspects of superconductivity that govern its practical applications. The first is the fact that the power loss in a conventional coil may be prohibi-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

tively expensive. The second is the need for fields higher than 2 T. The currents required to reach these higher fields may be so large that a conventional coil would destroy itself through Joule heating. Examples of these considerations are seen in fusion magnets, large bubble-chamber magnets, and MRI magnets. For instance, the International Thermonuclear Test Reactor has a magnet in which the current per conductor is 39 kA in a field of 13 T. Some idea of the size of the magnet is obtained from the average length of a turn around the toroidal coil, which is almost 43 m, and from the total weight of the coil, which is 500 t. The stored energy is over 100 GJ!

Medical MRI magnets make use of another feature of superconductivity. If a magnet is charged to full current and then shorted with a superconducting switch, the magnetic field will remain constant, as the currents will continue to flow in a zero-resistance circuit. The design of such switches along with that of exceedingly low-loss cryostats and the ready availability of liquid helium has contributed greatly to the wide spread use of MRI for medical diagnostics.

The magnets we have been describing are large magnets with almost no restriction on the amount of space that can be taken up by the coil. However, particle accelerator magnets and magnets used for very high-field research bring in different problems. A typical high-energy accelerator such as the Tevatron or the Large Hadron Collider (LHC) being built in Geneva requires a uniform magnetic field over a transverse region of only a few centimeters. However, the radius of the orbit may be many kilometers. For instance, the Superconducting Super Collider (SSC) had a design circumference of almost 50 km, but the useful region of the magnetic field was contained inside of a 5-cm-diameter circle.

At a fixed energy, the radius of the machine is inversely proportional to the magnetic field, and typical fields are of the order of 4–9 T. Magnet design places a high premium on locating the conductor very close to the useful field region in order to minimize the amount of conductor required. To obtain high fields, current densities in the coil of 2000–3000 A/mm^2 are required. Some appreciation of this current density can be obtained when one remembers that typical house wiring has current densities of the order of a few amperes per square millimeter. Thus, modern accelerator magnets are completely dependent upon superconducting technology. Similar considerations apply to very high-field magnets used for research purposes, where a very compact coil is necessary if fields of 20 T or more are desired.

The design of superconducting magnets is heavily dependent upon the design of the conductor, and there are two problems that must be solved. The first is that the conductor may be unstable in a high magnetic field and spontaneously make a transition to the resistive state. The second problem involves protecting the cable from destruction by heating if such a transition should take place. The specific resistance of superconducting alloys in their normal state is much higher than that of copper, and at the current densities that have been mentioned above, the conductor would destroy itself from Joule heating within seconds. Both of these problems are solved by embedding the superconducting filaments inside a matrix of copper or other high-conductivity material. Successful fabrication of rather complex conductors by industry has been the key to the construction of successful superconducting magnets.

1. PROPERTIES OF SUPERCONDUCTORS

The discovery of superconductivity at Leiden was made possible by the fact that Kamerlingh Onnes had liquefied helium, which has a boiling point of 4.2 K. Using helium as a cryogenic fluid, he was able to achieve temperatures down to about 1 K, and mercury was the first superconductor to be discovered. But soon many other metals showed the same property, and Kamerlingh Onnes was quick to realize the potential use of this discovery for constructing magnets that could generate a very high field without the concomitant high energy losses experienced by magnets with a copper winding. Further experiments soon showed that the property of superconductivity was a function not only of temperature but also of the field in which the conductor was embedded. The superconducting property would simply disappear at some critical value of the magnetic field. This effect precluded the attainment of high-field magnets, since the self-field would turn the conductor into its normal state. It would be another 50 more years before high-field superconducting magnets were a reality!

1.1 Type I and Type II Superconductors

The early superconductors such as lead, tin, and mercury are called type I and are characterized not only by zero resistance but also by being completely diamagnetic. That is, they completely expel any magnetic field from their interior when they are in the superconducting state. The state is defined by the following parabolic formula:

$$B_C = k[1 - (T/T_C)^2], \tag{1}$$

where T_C is the critical temperature, B_C is the critical field above which the superconducting state vanishes, and k is a constant. All of the superconductors that were studied in this early period had T_C less than about 4 K and B_C less than 0.1 T.

In 1930, it was discovered by de Haas and Voogd that an alloy of tin and bismuth behaved in a different manner. This was the first so-called type II superconductor. On a microscopic level these materials allow the magnetic field to penetrate, and one observes bundles of flux tunneling through the conductor, each surrounded by a superconducting vortex of current. These vortices are not free to move but are pinned in the material by imperfections in the lattice. Such imperfections are called "pinning centers" and are controlled by the metallurgy of the fabrication process. The major developments in the type II superconductors have come through understanding the manufacturing processes that establish a high density of pinning centers in the lattice of the material. This will be discussed more completely in Sec. 2.3.

Type II superconductors are characterized by three variables: the current density in the conductor, *J*; the magnetic field at the conductor, *B*; and the temperature, *T*. A plot of *J* vs. *B* and *T* is shown in Fig. 1 for a typical NbTi alloy. The superconducting state disappears for values of *J*, *B*, and *T* above the surface. A subscript "*C*" is used to define the critical values on the surface. A plot of J_C vs. *B* at various values of fixed *T* is shown in Fig. 2. The values of *B* at which the various curves intersect the axis is an intrinsic characteristic of the alloy, and the shape of the curve below this point is an extrinsic property dependent upon the manufacturing process. At 4.2 K, the maximum field for NbTi alloy is about 9.9 T, and for Nb_3Sn, about 19 T. Nb_3Sn is clearly a better material for a high-field magnet; however, it is a material that has proved very difficult to fabricate into a usable conductor. As a result, when higher B_C is required, NbTi alloy is frequently employed at lower temperatures than the 4.2 K of boiling He. One can raise the maximum field to 13 T by reducing the temperature to 1.8 K. The magnets that are being constructed for the Large Hadron Collider (LHC) in Geneva rely on operating at temperatures near 1.8 K, and at this temperature LHe is below the Λ point and is superfluid. Not only is B_C increased, but the cooling is vastly improved by the high heat conductivity of the helium.

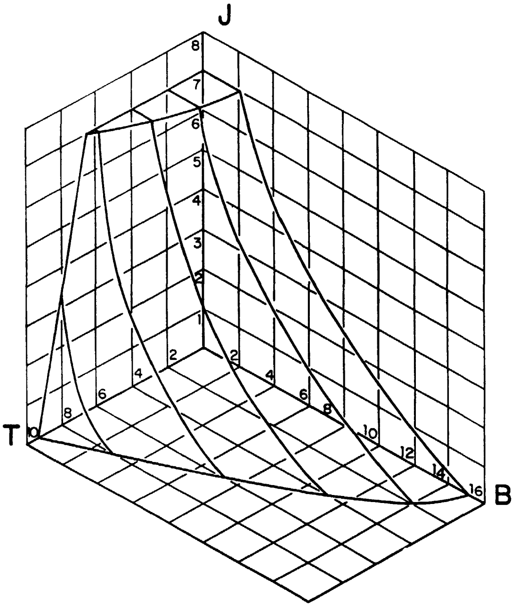

FIG. 1. The shape of the critical surface for a representative NbTi alloy. The material is superconducting only in the space under the surface. *J* is in kA/mm^2, *T* in K, and *B* in T.

In this article, two types of magnets will be used to illustrate the principles of superconducting magnet design. These examples will be solenoids and magnets for particle accelerators. There are many other uses for superconducting magnets, such as for fusion

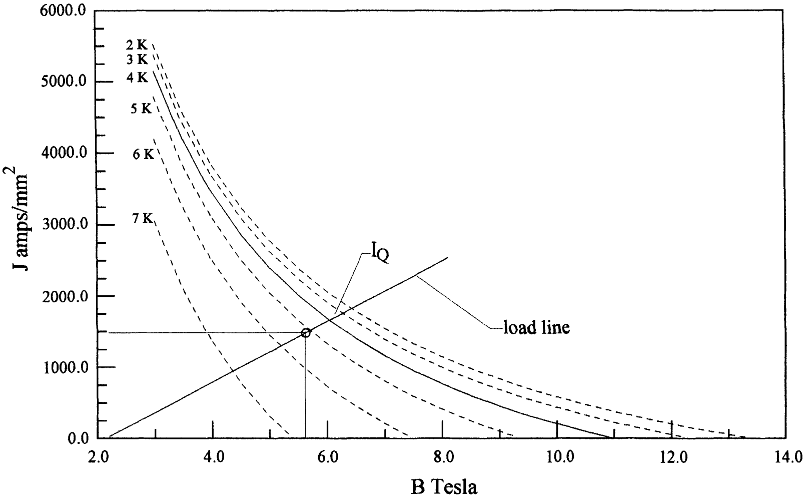

FIG. 2. This set of curves displays the properties of a typical NbTi alloy at a set of fixed temperatures between 2 and 7 K. The vertical axis is the current density in the superconductor, and B is the field in which it is immersed. The solid curve is for an ambient temperature of 4 K. Also shown is the load line for a typical magnet. If the ambient temperature is 4 K, then this magnet would quench at I_Q. The designed operating point is shown by the symbol "O."

devices, propulsion of ships, levitation and propulsion of trains, magnetic separators, and many other uses in instrumentation. However, the principles discussed here apply in general and serve as a tutorial approach to a complex subject. High-temperature superconductors bring in some new problems that are not discussed in this article, and the reader should consult other references (Shi, 1995; SUPERCONDUCTIVITY, HIGH-TEMPERATURE).

1.2 Load Line for a Magnet

The design of a superconducting magnet is dependent on the configuration of the conductor and its allowable current density. The required field shape will determine the geometrical placement of the turns, and the magnitude of the field will determine the current density. From the previous section, we have a relation between the field at the conductor and the limiting current density at the operating temperature. In general, the maximum field in the winding, B_m, will be higher than the required field B_0 in the useful region and must be calculated (frequently numerically) for the given winding geometry. We can plot the curve of B_m versus the current density in the winding as shown graphically in Fig. 2. This curve is called the "load line" for the magnet, and the conductor will turn normal at the point where the load line intersects the critical surface for the chosen operating temperature, which is shown as the solid line in Fig. 2. A successful design will have B_m less than the critical field at the chosen operating temperature and current density J_0, or a total conductor current of I_0. The maximum possible current is referred to as the "short-sample limit", I_Q, of the magnet. The term "short sample" refers to the fact that the characteristics of the conductor are determined using a short piece of the material, and it is assumed that this measurement will be representative of the whole winding. The safety factor I_0/I_Q can range from 0.9, which is typical for accelerator magnets, to 0.5 or less for magnets where a large stored energy could pose a real danger if the magnets should be accidentally driven into the normal state. As mentioned in the previous section, I_Q can be increased for a given conductor by lowering the temperature. For the large magnets that have been built so far, the two temperatures of choice are either that of boiling helium, which is 4.2 K, or superfluid helium at a temperature of about 1.8 K (Kingsbury and Van Sciver, 1995).

2. CONDUCTOR DESIGN

The design of a stable low-temperature conductor is a marvel of modern technology. The two materials that comprise the bulk of the commercially available conductor are NbTi and Nb_3Sn. The primary problem is to stabilize the conductor against the many sources of heat input that are present in an excited magnet. These consist of frictional motion of the winding, eddy currents in the conductor, small energy releases due to spontaneous reconfiguration of the field shape, and cracking of the epoxy insulation. Since the operating temperature is only a few kelvins, the heat capacity of the material is small and proportional to T^3. For instance, NbTi alloy has a heat capacity of only 3 mJ/cm^3 K at 4 K, almost 1000 times smaller than at room temperature. A secondary problem is that if the conductor goes normal, its resistance is very high and the Joule heating could immediately destroy it.

The solution to these two problems is to subdivide the superconductor into very fine filaments that are embedded in a copper or silver matrix and to assure that this matrix is in good thermal contact with the refrigerant. At low temperatures the electrical conductivity of copper may be over 100 times its room-temperature value, and this provides a means of carrying the current until the conductor can be recooled or the current reduced to zero by external means.

2.1 Flux Jumps

One source of heat input originates from small changes in the configuration of the magnetic field within the conductor. These changes are called flux jumps. To understand the process, consider a piece of superconductor in a magnetic field. The flux has penetrated, and there is some equilibrium distribution. However, Lenz's law states that during the charging process, currents will flow that resist the field penetration. When this opposing current density reaches J_C, the currents will no longer be able to resist further changes in the field, and more flux will enter the conductor. This reconfiguration of the flux in the conductor results in local heating. Another way of explaining the effect is that as the magnetic pressure builds up, it overcomes the strength of the pinning centers, and flux is forced into the conductor.

Now suppose we apply a small amount of heat dQ to a small volume dV. If no field were present, the temperature rise would be determined by the heat capacity. However, when a field is present, the temperature rise reduces the critical current J_C and consequently allows the field to change its distribution. This change in field distribution (redistribution of pinning forces) results in heating the conductor even more, and a runaway condition may exist that forces the conductor into its normal state. This is called a flux jump. The solution to the problem consists in subdividing the conductor into many fine filaments, which reduces the energy available for driving the instability. A detailed calculation shows that filament sizes of less than 100 μm are required. Figure 2 indicates that NbTi has a J_C at 4 K of about 2000 A/mm^2, and therefore a single 100 μm filament can carry no more than about 15 A. It becomes necessary to combine large numbers of filaments in parallel if the conductor is to be operated in a many-kiloampere range. The copper matrix structure provides the necessary mechanical support.

2.2 Eddy-Current and Hysteresis Losses

Additional sources of heat input are from eddy currents and from hysteresis losses. Eddy currents flow in conductors situated in a changing magnetic field. The source of particular concern here arises in multistrand cables where the strands are not insulated from each other and small loops can form using paths in the copper matrix. If the field changes, eddy currents will flow in these loops and produce local heating. These losses may be small, but, as was mentioned above, the heat capacity of the conductor is very low, and only a small amount of heat may increase the temperature enough to cause the conductor to quench.

Hysteresis losses are an additional source of heat that occurs in type II semiconductors. Consider a magnet that is driven up to its operating field and then returned to zero. It is found that there is a net energy loss in going around such a loop, much the same as for the hysteresis loss in a normal iron magnet. As the field increases, field lines are forced into the conductor, and as the field decreases, they move out. However, the pinning forces on the flux tubes make this an inelastic process, and heat is pumped into the lattice. This heat causes a small rise in the local temperature and must be carried away by the cooling system.

2.3 Conductor Fabrication

Satisfactory conductor construction depends upon the following considerations. First the superconductor must be divided into small filaments in order to insure flux-jump stability. Second, the filaments must be embedded in a copper matrix that can carry the current if the superconductor goes normal. The finished conductor must have good mechanical properties, as the forces in the magnet may be very high. Movement of the conductor can generate frictional heat, and strain in the conductor is known to reduce the J_C. Finally, there must be excellent cooling in order to insure stability as well as to remove any heat that is generated when the coil is excited.

2.4 NbTi Conductor

Niobium titanium superconductor construction is perhaps the easiest. Rods of the superconductor that are a few millimeters in diameter and up to a meter long are slid into a hole in a hexagonal copper rod. As many as 2000 rods are then closely packed together and inserted into a large copper cylinder that may be up to 12 in. (30.5 cm) in diameter. This assembly is heated and then extruded under high pressure through a die that is of the order of 3 in. (7.6 cm) in diameter. The composite is drawn down through successively smaller dies until finally a strand of conductor is produced that may have a diameter of a millimeter or less. A cross section of this strand will reveal NbTi filaments that are a few microns across embedded in a copper matrix. A photomicrograph of the type of strand used in the Tevatron is shown in Fig. 3. The details of the drawing of the strand and its annealing schedule control the pinning centers and thus the shape of the $J(B,T)$ critical current surface. Major improvements in the current-carrying capacity and the yield of unbroken strand have resulted from extensive metallurgical studies of the properties of this composite over the last 20 years.

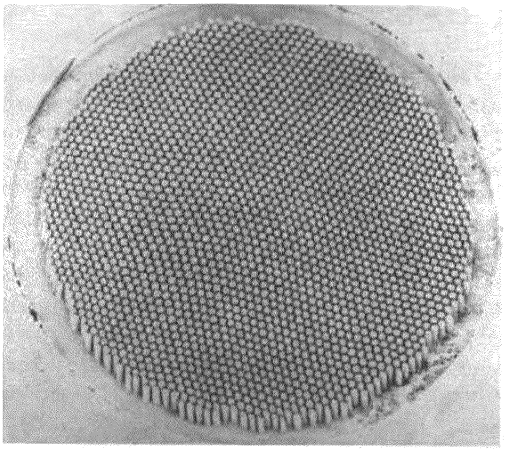

FIG. 3. Photomicrograph of a strand of superconducting wire used in the Tevatron. The copper matrix, which is about 0.6 mm in diameter, has been etched back, and the individual NbTi filaments, which are less than 10 μm in diameter, can be seen. At 4 K this strand will carry about 200 A in a magnetic field of 5 T.

Many smaller magnets are wound directly from the insulated strand. However, if currents larger than a few hundred amperes are required, the strand is woven into cable, which may have 30 or more strands. As the strands are cabled, they are twisted about their axes with a pitch of the order of a few centimeters in order to decouple the strand from any magnetic field in which it is embedded. In addition, the strands in the cable are wound in a spiral fashion around the axis of the cable in order to decouple the cable itself from the magnetic field. For particle accelerator magnets, this cable is flattened into a rectangular cross section that is a few centimeters wide and a few millimeters thick. It is insulated with Kapton film that is coated with epoxy on the side away from the conductor. After the coil is wound, a heat treatment bonds the epoxy-coated insulation together, forming a structure that can be handled.

However, other treatments of the strand are possible. Large solenoids have been constructed from a conductor manufactured by extruding very pure aluminum around the strands. A cross section of this conductor then shows a rectangular aluminum shape with a strand embedded inside. The shape provides a good means for mechanically supporting the forces on the conductor, and the additional cross section of very pure aluminum provides a large amount of very low-resistivity conductor to carry the current in case the coil should quench.

An additional technique that has been developed where space is not important is to incorporate the strand into a cable and then insert the whole cable into a stainless steel tube. Open channels are left within the cable so that LHe can then be circulated through the conductor. This technique is referred to as cable-in-conduit or CIC conductor and has been used for the large thermonuclear device magnets (Williams *et al.*, 1994; Mitchell *et al.*, 1994).

2.5 Nb_3Sn Conductor

Nb_3Sn conductor presents a new set of special problems. All of the preceding considerations apply, but in addition, the conductor is very brittle, and the filaments can easily fracture when the conductor is wound into a coil. As a result, the conductor is formed of unreacted niobium and tin. The billet is formed as for NbTi, but the niobium is enclosed in a bronze alloy instead of pure copper. The bronze provides a source of tin for forming the final alloy. The billet is then extruded and drawn into strand, which is cabled and wound into a coil. After the coil has been formed, it is then heat treated and the Nb_3Sn formed *in situ*. In order to do this, the conductor must be heated to about 650–700 °C for 175 to 200 h. The complexity of this process has limited its use to the fabrication of conductor for large rugged coils that can withstand the rigors of this heat treatment.

Nb_3Sn conductor is also being fabricated in the form of a thin tape that may be a few millimeters wide but only a fraction of a millimeter thick. This material is flexible and overcomes the obstacle of reacting the material in the completed coil. This conductor is particularly suited to solenoids where the topology of the winding does not require any tight bends and the forces are easily supported. There is a slight additional complication in that the material is anisotropic, and so attention must be paid to the field direction as well as magnitude during the design and subsequent construction of the coil.

2.6 Cryogenic Stability and Quench Waves

There are two general classes of conductor. Large bubble chambers and thermonuclear reactor coils are designed to be cryogenically stable. This means that there is sufficient LHe cooling to reestablish the superconducting state if the current should momentarily transfer to the copper matrix. This is arranged by having a sufficient amount of copper so that the Joule heating of the copper when forced to carry all of the current is less than the cooling provided by the LHe. This assures that if the temperature accidentally fluctuates above T_C, the cooling will be sufficient to reestablish the superconducting state. This type of conductor requires a large cross section of high-conductivity copper or aluminum. The problem is ameliorated by the fact that the conductivity of very pure copper or aluminum at 4 K can be several hundred times its value at room temperature.

As mentioned before, some magnets, such as particle accelerator magnets, place a high premium on coils that occupy smallest possible space. This requires a minimum amount of copper surrounding the superconductor, and if the conductor is driven into the normal state, the current flowing through the magnet will continue to heat it faster than it can be cooled. The spot that first turned normal will then spread longitudinally along the cable. This wave radiating away from the initial point is called a "quench wave." The initiating spot will be the hottest point, and the temperature rise can be calculated as follows. Consider a cable carrying a current I, and having a resistance per unit length $R(T)$ that is a function of the temperature. Let $C(T)$ be the temperature-dependent heat capacity per unit length of the cable. Then we can write the following equation:

$$I^2(t)R(T)dt = C(T)dT. \qquad (2)$$

This equation says that the Joule heat generated in a time dt only heats the cable and that no sources of external cooling are available. This is clearly a pessimistic assumption in that the helium in contact with the cable provides a major reservoir for cooling. By collecting together all of the terms that depend on temperature on one side, one can integrate the equation as follows:

$$\int_0^t I^2(t)dt = \int_{T_i}^{T_f} \frac{C(T)}{R(T)}\,dT. \tag{3}$$

The integral of I^2 on the left-hand side is under the control of the external power supplies, and the term on the right depends only upon the thermodynamic properties of the conductor and can be calculated from the known values of $C(T)$ and $R(T)$ and is only a function of the final temperature, T_f. This equation is extensively used in magnet design to calculate how fast the current must be reduced in order to limit the temperature rise of the conductor after a quench wave has started. Note that the point of highest temperature will be the point of the initiation of the quench, and the rest of the cable will be at a lower temperature. The low resistivity of cold copper helps to limit T_f. The designer also can control the ratio of the amount of copper to superconductor in the strand. This ratio is typically between 1.5 and 10.

2.7 Conductor Splices

Splicing conductor becomes necessary in order to obtain strands of conductor long enough for cable fabrication, for connecting coils internally within a magnet, or for connecting magnets in series. The strand itself can be ultrasonically welded, making a joint that reduces I_Q locally by about 30–40%. In a multistrand cable, this is only a small effect on the total cable current-carrying capacity but could be a serious defect in a single-strand conductor.

Conductor may be joined by soldering to the copper matrix that encloses the filaments. Frequently the joint is made by overlapping the two flat sides of the cable and soldering with lead tin solder. Joints in a 23-strand cable with an overlap of 30 cm can have a resistance of the order 10^{-9} Ω. Similarly, conductor coextruded with a pure aluminum jacket can be overlapped and welded along the edges. Both of these methods result in a compact splice that may be embedded smoothly within the winding. Other methods have been developed (McIntyre *et al.*, 1995). Careful attention must be paid to cooling because at currents of 10^4 A, a joint with 10^{-9} Ω will still dissipate 0.1 W. Coils for MRI magnets that operate in the persistent-current mode require careful attention to the joints, as a fractional magnetic field stability about 10^{-12}/s is desired. With an inductance of the order of 100 H, this requires a total coil resistance of less than 0.1 nΩ.

CIC cable used in large-volume magnets is a more difficult problem, and the joint is generally made in an external region, as the cryogenic circuit must also be kept intact. In all cases, the magnetic forces on the joint must be carefully calculated and provision made for absorbing them within the magnet structure.

3. LARGE SOLENOIDS

We now examine two solenoids as examples of the design techniques that have been developed. Table 1 gives the parameters for a MRI solenoid and for a very large magnet being designed for one of the particle detectors to be used with the LHC at CERN. A brief description of these two magnets follows.

The first is the MRI magnet developed by General Electric for image-guided invasive procedures (Laskaris *et al.*, 1995). It consists of two sets of coils that are split axially by 56 cm in order to provide access to the central region. There is no return yoke, and the large axial attractive force between the two coil assemblies is absorbed by four stainless steel posts. The conductor is copper-stabilized Nb_3Sn tape, and there is a superconducting switch to short the magnet and place it in a persistent-current mode (Dorri and Laskaris, 1995). It is charged through leads fabricated from high-temperature superconductor (HTS) from an external power supply. There are two cryocoolers that maintain the temperature at 10 K, thus effectively utilizing the high T_C of Nb_3Sn.

The field is constant over a sphere of 20 cm diameter to 1.6 ppm at the center of the magnet. In order to achieve such high uniformity, the two main coils are subdivided into three sections, and the positions and numbers of turns in the individual sections

Table 1. Properties of two modern solonoids. The first is a cryocooled solenoid designed for interventional MRI applications. The second is a large detector magnet intended for use with the LHC at CERN. The table illustrates the extreme range of parameters possible with SC solonoids.

	GE-MRI	CERN CMS detector
B (T)	0.5	4.0
Diam. × length (m^2)	0.96 × 1.48	5.9 × 14
Inductance (H)	464	14
Current (A)	65	20 000
Stored energy (MJ)	1	2850
Conductor type and size	Nb_3Sn 3 × 0.22 mm^2	NbTi 61 × 22.3 mm^2
Stabilizer to SC ratio	6	~9
Cooling	Cryocooler, ~10 K	LHe, ~4 K

are then used to optimize the size of the uniform volume and minimize the conductor required. The final trimming of the field for uniformity is done with iron shims placed in the warm bore of the cryostat.

The coils are wound on two stainless steel forms and then impregnated with epoxy resin. The large hoop stress is absorbed by heat-shrinking stainless steel rings over the winding, which provides a large compressive force. This illustrates a technique commonly used to limit the stress excursions at the winding. In many cases the initial pressure is great enough so that there is no stress reversal when the magnet is cooled and excited. This insures that the support will not have microfractures, which can release heat and cause the magnet to quench. In all cases a careful 3D computer analysis must be made in order to ensure safe magnet operation during cool-down, normal operation, and abnormal operation such as during a quench.

The initial cool-down is accomplished using cryogens and takes about 5 days. The two cryocoolers maintain it at the operating temperature of 10 K and also allow for servicing. The total heat load of the 70-K shield and intercept is 49 W and at 10 K is 2.1 W. HTS leads using grain-oriented YBCO are intercepted at 70 K and are copper from that point to room temperature. The 10-K load of the leads is 0.2 W and the 70-K load is 11 W.

For contrast we next consider a solenoid being constructed for the CMS detector at CERN (Desportes, 1995). Larger magnets have been designed or built, but this example will illustrate the techniques used. The design problems are dominated by the large forces that must be safely controlled. In this example, the return yoke provides a massive structural element for supporting the coil. However, the very large hoop and axial forces within the coil package have required new techniques to be developed.

First, a four-layer coil is used in order to obtain the ampere-turns necessary for a 4-T field. In order to distribute the stress uniformly through the four layers, it is required that each layer must support its own hoop stress. However, the customary superconductor coextruded with ultrapure aluminum is not strong enough, and an additional layer of high strength alloy is used to reinforce the conductor. The uniform stress distribution is necessary to keep the epoxy-insulation matrix from fracturing under excitation.

The coil is fabricated in four sections that are 3.35 m long and have 146 turns each, wound on the inside of a 50-mm-thick Al alloy cylinder before vacuum impregnation with epoxy. The four coils are then assembled in a common cryostat. Indirect cooling of the coil is accomplished by flowing LHe through a closed circuit thermally bonded to the outer aluminum shell. The support for the coil in the cryostat requires care in design. If the coil is exactly centered relative to the iron flux return, the net axial force is zero by symmetry. However, a small displacement can result in large magnetic forces, and the elastic constant of the support must be great enough to ensure stability. However, a very stiff support requires structural members of large cross section with a resultant high heat loss. Instrumentation to measure the forces on the support system during excitation can help to minimize the off centering and maintain an acceptable margin of safety while utilizing a less massive support system.

Quench protection is accomplished by dumping the energy of the coil into an external resistor. Equation (3) is used to control the maximum temperature reached by the conductor. However, as pointed out, this equation gives an upper limit since, in general, a quench wave will spread through the coil structure and shorten the discharge time by increasing the total circuit resistance.

4. ACCELERATOR MAGNETS

Superconductivity has played a dominant role in the design of high-energy particle accelerators. We will discuss the magnets for these machines in detail, as they provide many illustrations of the construction techniques that have been developed for magnets over the past 20 years. The size of an accelerator at fixed energy is determined by the radius of curvature of the orbit. Thus, a premium is placed on bending magnets with a high magnetic field. However, the forces within the magnet are proportional to B^2, and the complexity of the support system for magnetic fields over 10 T rapidly increases the cost. Studies show that the cost of a machine has a flat minimum for fields around 8–9 T. Part of this cost is due to the extreme accuracy required for the field shape. For fields greater than 2 T, the shape is no longer determined by the shape of the iron but rather by the accuracy of the coil dimensions, and winding tolerances of the order of 20 μm are not uncommon in modern accelerator magnets.

A large part of the cost of the magnet is due to that of the superconductor, and hence the amount of superconductor is minimized by operating at 90–95% of I_Q. As a result the magnet becomes very sensitive to any source of heat input such as conductor motion, beam loss, etc., and a great deal of attention must be paid to magnet protection. With this short introduction, we now examine some of these points in detail.

4.1 Coil Configuration

For the main bending magnets, one desires a uniform vertical field. In practical terms, this requires that the field be constant to an accuracy of one part in 10^4 over a central region of the magnet cross section, which may be several centimeters in size. The field in a long magnet is essentially two dimensional, and at any point (r,ϑ) the field in the vertical direction, B_y, and in the horizontal direction, B_x, can be written as

$$B_y = B_0\left[1 + \sum_{n=1}^{\infty}\left(\frac{r}{r_0}\right)^n (b_n \cos n\theta - a_n \sin n\theta)\right], \tag{4}$$

$$B_x = B_0\left[a_0 + \sum_{n=1}^{\infty}\left(\frac{r}{r_0}\right)^n (a_n \cos n\theta + b_n \sin n\theta)\right], \tag{5}$$

where r_0 is a reference radius that sets the scale and is customarily taken to be two-thirds of the inner coil radius. Terms in the series are referred to as dipole, quadrupole, sextupole, etc. The a_n and the b_n are referred to as the skew and normal terms, respectively. The geometry of the coil determines which terms of the series are excited.

For bending magnets, only the first term in B_y is desired, and all the other coefficients should be minimized. Figure 4 shows a drawing of the Tevatron magnet, and the coil cross section can be seen surrounding the beam pipe. The coils subtend half angles with respect to the center of the beam pipe of β_1 and β_2 measured from the horizontal. The z axis, which is the direction of the beam, is out of the figure. Symmetry about the x and y axes will eliminate all of the a_n and the odd b_n. The two terms representing a sextupole and decapole, b_2 and b_4, can be eliminated by adjusting the two angles β_1 and β_2. Thus, the first harmonic after the dipole that appears is the 14-pole, which varies as r^6, and the field is approximately constant over the inner two-thirds radius of the aperture to better than one part in 1000. By placing additional spacers between the turns of the winding, more of the harmonics can be canceled, as was done for the magnets developed for the SSC, where the 14- and 18-pole terms were also eliminated.

In addition to bending magnets, accelerators and particle beam lines frequently require quadrupoles for focusing, and sextupoles or even higher multipoles for correcting errors or making the beam achromatic. Such magnets are designed in the same manner as indicated above for the dipole, but the coil will have additional poles.

The region outside of the coil is surrounded by iron in order to make a low-reluctance flux return. At the high fields, the

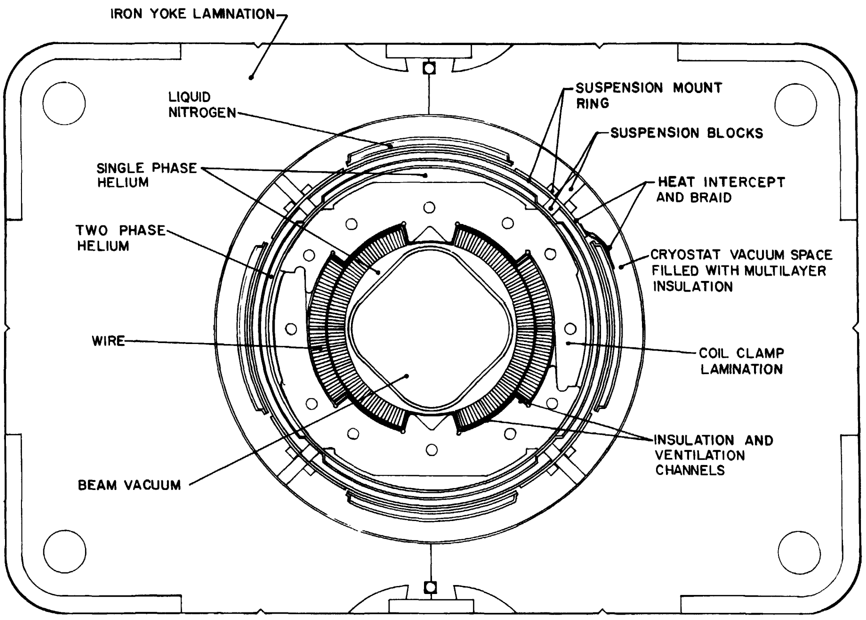
IRON YOKE LAMINATION
LIQUID NITROGEN
SINGLE PHASE HELIUM
TWO PHASE HELIUM
WIRE
BEAM VACUUM
SUSPENSION MOUNT RING
SUSPENSION BLOCKS
HEAT INTERCEPT AND BRAID
CRYOSTAT VACUUM SPACE FILLED WITH MULTILAYER INSULATION
COIL CLAMP LAMINATION
INSULATION AND VENTILATION CHANNELS

FIG. 4. Cross-section drawing of the Tevatron magnet. The iron yoke provides the flux return and is at room temperature. The coils are at 4 K and surrounded by LHe. The annular spaces around the collars carry the cryogens in a closed-loop system.

iron is frequently completely saturated, and the magnetic field will be a nonlinear function of the current. The properties of the iron are determined not only by its chemical composition but also by the fabrication processes in the steel mill such as rolling and heat treating. Thus careful control of the iron manufacturing process is necessary for accurate control of the magnetic field.

4.2 Forces

The forces in the coil are caused by the normal **J** × **B** forces. For a high-field magnet the field within the winding can produce large forces on the coil. For instance, a cable in a 10-T field carrying 10 000 A has a force of 10^5 N/m perpendicular to its axis. Another way of appreciating these forces is to note that the same 10-T field corresponds to a magnetic pressure of about 4×10^7 N/m^2.

For the coil that we are considering there are two components of force. The first is a large radial outward force on the coil, and the second is an azimuthal compression. This latter is caused by the magnetic field curling around the ends of the winding. These forces must be supported by an external structure that maintains the coil dimensions constant to about 1 part in 10^4. This support is referred to as a "collar" and is a nonmagnetic stainless steel yoke that provides a precision form for the coil and easily supports the radial forces.

The azimuthal compressive forces are particularly vexing. They cause the coil to compress and thus change the angles β_1 and β_2, which in turn affect the field shape. This problem is solved by collaring the coil under high pressure so that the mechanical forces are always greater than the magnetic ones. This guarantees that the contour of the collar will always be in contact with the coil and thus control its shape.

4.3 Magnet Fabrication

We can now describe the complete magnet. The coil is first fabricated from cable that has been flattened into a shape that has a thickness of about a millimeter and a width of 1 to 2 cm. It is wrapped in Kapton film that is 20 to 40 μm thick and serves as the electrical insulation. The outside is treated with a heat-sensitive adhesive. The coil is wound on a form, put into a high-pressure mold, and heated to set the adhesive. All of the dimensions of the components must be carefully controlled if the shape of the magnetic field is to be within tolerance. The coil after curing can be handled but is structurally rather weak.

A major invention for constructing the collars involved using laminations stamped from thin stainless steel sheet. Stampings can be made with very high accuracy (better than 10 μm) and with a very reproducible shape determined by a die. The industrial process is well developed and inexpensive. The collars are formed in two pieces that are assembled around the coil, pressed in a large press to the final dimensions, and then welded together. The heat conductivity of the stainless steel is so low that the welding does not damage the insulation of the cable.

The Tevatron uses a warm-iron flux return, as can be seen in Fig. 4. The iron laminations form a 10-in. (25.4 cm) diameter hole along the axis. The cryostat fits in this hole and holds the collared coil, which is $5\frac{1}{2}$ in. (14 cm) in diameter. The angular space between contains the vacuum jacket, inside of which there are layers of superinsulation, a liquid nitrogen shield of 77 K, and finally a LHe bath around the coil. The cryostat forms a pipe for the cryogens, and a chain of 16 magnets is cooled by the circulating liquid. The iron yoke increases the field by 1 T. The arrangement is awkward because of the mechanical complications caused by placing a very complicated cryostat in a small radial space. The advantage, of course, is that the large mass of iron does not have to be cooled to 4 K.

The electron–proton collider, HERA, at Hamburg uses a similarly collared coil, but it is surrounded directly by the iron return yoke, which is therefore at 4 K. Modern refrigerators are able to cool the resulting large mass to LHe temperature in a reasonable time (days), and there is a big advantage in being able to use the iron to help support the radial forces exerted by the coils on the collars. The vacuum cryostat then surrounds the whole mass.

The new accelerator being built in Geneva, the LHC, is a 14 × 14-TeV proton–proton (*pp*) collider that has evolved still further (Perin, 1995). A *pp* collider requires two rings to carry the two proton beams in opposite directions (unlike the Tevatron, which is a proton–antiproton collider and uses only a single ring). The two magnets are mounted near each other, as shown in Fig. 5. Since the field must be in opposite directions in the two apertures, the LHC uses a single iron return yoke enclosing both coils so that flux lines of one magnet return through the other, resulting in a very compact structure. This can be seen in the flux plot shown in Fig. 6. The operating field is 9 T, and the iron is carefully shaped so that as it goes

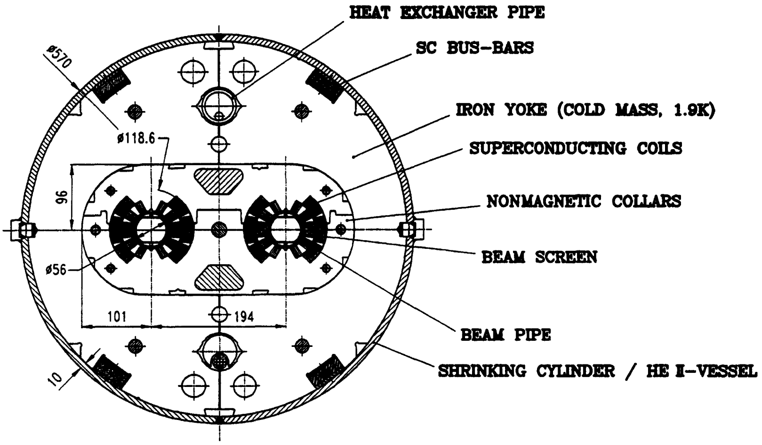

FIG. 5. Cross-section drawing of the LHC magnet. The cryostat surrounding the iron yoke is not shown. The magnetic flux goes up through the right-hand coil and down through the one on the left. The stainless steel collars that clamp the coil are nonmagnetic. The coils and iron yoke operate at 1.8 K. The dimensions are in millimeters.

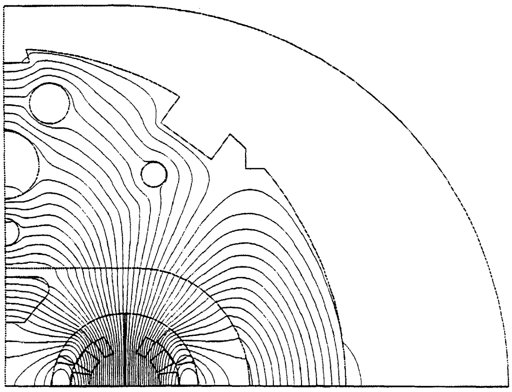

FIG. 6. Flux plot for the right-hand coil of the LHC dipole shown in Fig. 5. The flux going up through this coil is returned going down through the left coil.

into saturation the field in the two apertures remains uniform.

4.4 Training

The road to successful superconducting magnets has been plagued with the process called "training." A magnet designed to quench at current I_Q may only reach 50–90% of this value before spontaneously quenching on its first cycle of excitation. The second cycle and subsequent cycles will quench at increasingly higher values of current and asymptotically approach the value I_Q. The number of quenches to reach the limiting current can vary from 0 up to hundreds of cycles and is caused by heat generated from extremely small frictional motion of the conductors in the winding due to the increasing $\mathbf{J} \times \mathbf{B}$ forces during excitation. Each excitation cycle releases, in a ratchetlike fashion, the nonequilibrium forces that were generated during the initial coil fabrication until the coil eventually comes to an equilibrium state in which the ponderomotive forces are insufficient to break any more of the bonds.

This phenomenon has been particularly troublesome for accelerator magnets, as they are designed to operate very near I_Q, and since there may be several thousand magnets in a ring, it is imperative that all magnets operate at their designed value without training. The problem does not exist for cryostable magnets (as the conductor has enough cooling to return it to the superconducting state even if it is driven into the normal region) and is seldom a problem for large solenoids that are designed to operate with a very conservative safety margin. The source of the instability is due to the extremely small amount of heat that is necessary to drive the conductor from the superconducting state to the normal state as I approaches I_Q. Calculations and measurements both show less than a millijoule per gram of winding can cause a magnet to quench when it is operating at $0.9I_Q$. This small amount of energy is easily available from a number of sources. It may come from flux jumps within the conductor or from frictional motion of the conductors as the magnet is excited. A particularly bad source of heat can arise from fracture of the epoxy adhesive if a coil is completely impregnated. Accelerator magnets have thus avoided having epoxy in direct contact with the metallic conductor by completely wrapping it with a plastic film. Flux jumps are eliminated as previously described by proper conductor design. Finally, beam losses from the accelerated protons can easily deposit enough energy to quench a magnet, and so careful control of the orbit during all parts of the accelerating cycle is demanded.

Experience has shown that it is possible to prevent inelastic motion of the conductor, provided the mechanical pressure from the collaring operation is much larger than any of the magnetic pressures that the coil undergoes during excitation. Increases in collaring pressure are limited because if it becomes excessive, the turn-to-turn insulation will be ruptured. In spite of all of the above-mentioned difficulties, it is possible by careful design to construct large numbers of magnets that are stable at a current of $0.9I_Q$ and do not require training. However, in all cases, one must then be prepared to protect the coil if it starts to quench.

5. QUENCH PROTECTION

Once a quench wave starts in the conductor of a magnet, it will continue to grow as long as the current continues to flow. The temperature will rise according to Eq. (3) and must be limited to values that will not destroy the conductor. The method of disconnecting the power supply and connecting a resistor across the coil terminals works for a solenoid but does not work for an accelerator, for there may be hundreds of magnets in series, each storing one or more megajoules of energy. A different protection scheme is used. The offending magnet is first shorted after an electronic monitoring circuit detects a resistive component in the voltage. All of the energy in the magnet will now be absorbed in its winding. If the quench does not spread very fast, the energy may be deposited in only a few meters of the conductor with destructive results. However, a second mechanism can come into play. As the conductor heats, it can become hot enough to cause adjacent turns to also turn normal. If the thermal conductivity between turns is high, the whole winding may turn normal, and the energy will be deposited in a large mass. The resulting temperature rise for the coil may be less than 100 K. If the thermal conductivity is not high enough to spread the quench spontaneously, a second technique can be used. Each magnet is equipped with a heater that is thermally coupled to the coil. In the event a quench is detected, a pulse of current is supplied to the heater, which turns the whole winding normal. The amount of heat required is rather small, and the energy necessary for this pulse of current is stored in a large capacitor connected to a transistor switch.

6. COOLING THE MAGNET

Various techniques are used for cooling magnets, and the technique chosen is highly dependent on the anticipated heat load (see CRYOGENICS). MRI magnets for medical purposes can have very low losses as they are fixed-field magnets that are placed in a very low-loss cryostat and may even have the charging leads disconnected after the magnet is at full field. Pulsed magnets for large accelerators or the confinement magnets for thermonuclear machines must have large amounts of heat removed from the winding. The heat is generated by high-energy particles interacting in the coils and, in the case of pulsed magnets, from eddy-current and hysteresis losses as described earlier.

The simplest system is that used for the medical MRI magnets. The magnet is contained in an extremely low-loss cryostat, and LHe is supplied from a Dewar container. The total losses are so low that the He supply may only need to be serviced weekly or monthly. A second option for cooling low-loss magnets is a cryocooler (Laskaris *et al.*, 1995). This is simply a small closed-cycle refrigerator in which the last stage operates at a level of a few watts at 10 K. The 10-K operating temperature is nicely matched to magnets with Nb_3Sn coils and is now used extensively with medical MRI installations.

Large accelerator installations and thermonuclear machines require massive amounts of refrigeration to remove the heat generated within the magnets. Modern plants have an efficiency of the order of 0.5 MW of wall power per kilowatt of heat removal at 4 K. The cryogen is LHe at 4 K, but intermediate heat shields at 20–30 K or at liquid-nitrogen (LN) temperature may be used to intercept the room-temperature load and increase the overall Carnot efficiency. If superfluid He is required, it is generated from the 4-K liquid by passing it through a Joule–Thompson valve into a low-pressure region. This is the scheme that will be used at CERN for the LHC magnets.

7. CURRENT LEADS

A major heat load for the cryogenic system may be the leads that carry the current form the room-temperature power supply to the cold coil. If the superconductor were brought out of the cryostat to the power supply, its cold end would remain superconducting until reaching a point along the length where the temperature was greater than T_C, at which point it would become normal. It will conduct heat along its length and also generate Joule heat over the normal section. As mentioned before, once the superconductor goes normal, its resistance is very high, and the current is carried completely by the copper stabilizer. In general it is necessary to increase the amount of copper to enable the lead to safely carry the magnet current in a steady-state condition. Joule heating would dictate leads with a large cross section, but the Wiedemann–Franz law (see METALS AND ALLOYS, CONDUCTIVITY IN) guarantees that such leads will have a high heat conduction, and so careful attention to detail is necessary. In general, the design equates the Joule heating to the conduction heating. The combined source of heat is then carried away by passing cold vapor from the LHe through channels in the leads. Practical leads have a conduction loss to the 4-K system of the order of 1 mW per ampere of lead current.

HTS is being used to develop leads for a variety of magnets (Maehata *et al.*, 1995). Low-current magnets using cryocoolers and Nb_3Sn coils furnish an ideal application. As the intermediate shield is generally below 70 K, HTS provides an ideal conductor between this temperature and that of the coil. A section of the lead is made of bulk HTS. The field around the leads is frequently not high, and this provides an ideal use for the high-J_C, low-B_C material that is readily available. It is possible to make such leads with a heat leak that is only 0.1 mW per ampere of lead current, which is a factor of 10 better than the equivalent copper lead.

8. FUTURE DIRECTIONS

The use of low-temperature superconductor (LTS) materials has been completely dependent on the solution of some very complicated material design problems. At present, applications below 10 T are dominated by the use of NbTi, and even here cooling to 1.8 K is often employed. Nb_3Sn is used for higher fields, up to about 15 T at 4 K, or it can be used at lower fields with cryocoolers, thus utilizing its higher T_C to operate at 10 K. Improvements in both types of conductors continue to be made. Although B_C is fixed by the alloy and the operating temperature, J_C at a specified temperature and field can be optimized by controlling the details of the strand manufacturing process. The current-carrying capacity of NbTi at 5 T has been more than doubled over the last 15 years, and further improvements may still be expected. Recently a new method of strand fabrication has been developed that allows much more control over the density, spacing, and size of the pinning centers. At present, these variables are controlled by the schedule of the heat-treatment and drawing process. This is time consuming and must be carefully monitored. In the new technique, so-called "artificial pinning centers" are directly introduced before the material is extruded, which allows more direct optimization of the conductor (Miura *et al.*, 1995; Scanlan *et al.*, 1994).

8.1 High-Temperature Superconductors

The greatest promise for the future lies in the domain of HTSs. Materials are now known with a T_C of 125 K, which is well over the boiling point of LN. This opens up an enormous role for these materials in the electrical power and transportation industries. In addition, some of the material shows B_C values that are in excess of 50 T, which is well beyond anything available at present. However, the development of practical conductor for use in large-scale magnets is fraught with a myriad of difficulties as HTS materials are anisotropic crystalline ceramics. Progress is being made, and we list below some of the differences that HTS magnets will have relative to LTS types.

1. The operating temperature is higher, which makes the refrigeration much easier and cheaper to construct. The economic advantage is enormous and will have a great influence on the practical applications, especially if it can be done using LN as a refrigerant. Since the temperature is higher, the heat capacity of the conductor is also much greater than for material at 4 K, which makes the magnets intrinsically less sensitive to small influxes of heat from the surroundings.
2. The promise of higher B_C points to the possibility of practical very high-field magnets. For instance, some materials exhibit a critical field of over 50 T when cooled to 4 K. At present, the highest fields are obtained from hybrid magnets. An example is the 45-T hybrid being constructed at the National High Field Magnet Facility at Tallahassee, in which a LTS coil is designed to produce a field of about 14 T and an inserted, water-cooled copper coil generates an additional 31-T field using 24 MW of power (Miller *et al.*, 1994). There would be an enormous advantage to eliminating the inner copper coil and its huge power supply and replacing it with a low-loss HTS coil with a very high B_C.

This is a rapidly changing field, and the best overview is obtained from studying the transactions of various conference proceedings such as the INTERMAG Conference series published under the auspices of the IEEE. Some examples can be seen in the list of references.

GLOSSARY

Collars: The stainless pieces that are placed around the conductor and that absorb the magnetic forces.

Critical Current, J_C: The critical current density beyond which the superconducting state disappears. It is a function of field and temperature.

Critical Field, B_C: The critical field beyond which a material is no longer in the superconducting state.

Flux Jumps: Spontaneous rearrangements of the flux inside the body of a type II superconductor.

I_Q: The maximum superconducting current that a conductor can carry at a given field and temperature as measured with a short sample of the material.

HERA: A proton–electron collider using superconducting magnets for the proton ring. It is located at the DESY Laboratory in Hamburg, Germany.

HTS: High-temperature superconductor.

LHC: Large Hadron Collider. This proton–proton collider is being constructed at CERN in Geneva. It consists of two rings of superconducting magnets with a maximum dipole field of more than 8 T that guide two counter-rotating beams of protons into collision with a center-of-mass energy of 14 TeV.

LHe: Liquid helium. The normal temperature is 4.22 K at 1.0 atm.

Load Line: The relation between the field at some chosen point in the magnet and the current through the winding. The point used is frequently the central field, but for winding design, the high-field point in the conductor is the relevant point.

LN: Liquid nitrogen. Its temperature at 1.0 atm is 77.39 K.

LTS: Low-temperature superconductor. Generally NbTi or Nb_3Sn.

MRI: Magnetic resonance imaging.

Pinning Centers: Defects in the lattice of a type II superconductor that anchor the current vortices surrounding the quantized bundles of flux.

Quench: Refers to some point in the magnet coil making a transition to the normal state.

Quench Wave: The propagation through the magnet winding of the normal state caused by the Joule heating.

Tevatron: The first proton synchrotron accelerator constructed using superconducting magnets. It operates as a normal proton accelerator to 800 GeV and as a proton–antiproton collider at 1.8 TeV c.m. energy. It is located at Fermilab in Batavia, Illinois.

Training: If a newly cooled magnet is repeatedly excited to its quench point, it is frequently observed that the series of peak currents will show an increase with cycle number. The magnet is said to be "training." A well-designed magnet will reach its short-sample limit with no training quenches.

Short-Sample Limit: A magnet is said to reach its short sample limit if the current can be increased to the value predicted from measurements on the current-carrying capacity of the conductor combined with the calculation of the high-field point in the magnet coil.

SSC: Superconducting super collider. The design was for a proton–proton collider operating at 20 TeV in the center of mass system. Construction was started in Texas, but it was canceled before completion. However, the magnet was developed and prototypes constructed and tested.

Works Cited

Desportes, H. (1995) in: "Proceedings of the 1994 INTERMAG Conference," *IEEE Trans. Appl. Supercond.* **5,** 1525–1532.

Dorri, Bijan, Laskaris, T. (1995), *IEEE Trans. Appl. Supercond.* **5,** 177–180.

Kingsbury, D. L., Van Sciver, R. W. (1995), *Cryogenics* **35,** 573–579.

Laskaris, T., Ackermann, R., Dorri, B., Gross, D., Herd, K., Mimas, C. (1995), in: "Proceedings of the 1994 INTERMAG Conference," *IEEE Trans. Appl. Supercond.* **5,** 163–168.

Maehata, K., Nishioka, T., Ishibashi, K., Takeo, M. (1995), *IEEE Trans. Appl. Supercond.* **5,** 765–768.

McIntyre, P., Wu, Y., Liang, G., Meitzler, C. (1995), in: "Proceedings of the 1994 INTERMAG Conference," *IEEE Trans. Appl. Supercond.* **5,** 238–241.

Miller, J. B. *et al.,* (1994), *IEEE Trans. Magn.* **30,** 1633–1638.

Mitchell, N. (1994), *IEEE Trans. Magn.* **30,** 1633–1638.

Miura, O., Inoue, I., Suzuki, T., Matsumoto, K., Tanaka, Y., Funaki, K., Iwakuma, M., Yamafuji, K., Matsushita, T. (1995), *Cryogenics* **35,** 169–180 and 181–188.

Perin, P. (1995), in: "Proceedings of the 1994 INTERMAG Conference," *IEEE Trans. Appl. Supercond.* **5,** 189–195.

Scanlan, R. M., Lietzke, A., Royet, J., Taylor, C. E., Wandesforde, A., Wong, J., Rudziak, M. K. (1994), *IEEE Trans. Magn.* **30,** 1627–1632.

Shi, Donglu (Ed.) (1995), *High-Temperature Superconducting Materials Science and Engineering,* Tarrytown, NY: Pergamon Press.

Williams, J. E. C., Baker, M., Bobrov, E. S., Iwasa, Y., Leupold, M., Stejskal, V., Weggel, R., Zhukovsky, A. (1994), *IEEE Trans. Magn.* **30,** 1633–1638.

Further Reading

Brechna, H. (1973), *Superconducting Magnet Systems,* New York: Springer-Verlag. Probably one of the most complete and detailed references available on superconducting magnet design. It is becoming somewhat dated, but still remains an excellent resource.

Iwasa, Yukikazu (1994), *Case Studies in Superconducting Magnets,* New York: Plenum Press. The author uses case studies to illustrate the various aspects of superconducting magnet technology. Up to date and good for specific design principles some coverage of HTS use.

Palmer, R., Tollestrup, A. (1984), "Superconducting Magnet Technology for Accelerators," *Annu. Rev. Nucl. Part. Sci.* **34,** 247–284. A general overview of superconducting magnet design for accelerators.

Shi, Donglu (Ed.) (1995), *High-Temperature Superconducting Materials Science and Engineering,* Tarrytown, NY: Pergamon. An up-to-date book on high-temperature superconductors and their technology.

Tinkham, Michael (1975), *Introduction to Superconductivity,* New York: McGraw Hill. An excellent text book on the classical theory of superconductors.

Wilson, Martin N. (1983), *Superconducting Magnets,* Oxford: Clarendon Press. A comprehensive text for LTS magnet design.

In addition to the above, the following articles in this encyclopedia contain information relevant to superconducting magnet technology:

CRYOGENICS; ELECTRICAL POWER ENGINEERING; FUSION, MAGNETIC CONFINEMENT; INDUCTORS; MAGNETS; MAGNETOSTATICS; MAGNETIC RESONANCE IMAGING; PARTICLE COLLIDERS; SUPERCONDUCTIVITY, HIGH-TEMPERATURE; SUPERCONDUCTIVITY, LOW-TEMPERATURE.

SUPERCONDUCTIVITY, HIGH-TEMPERATURE

C. W. Chu, *Department of Physics and Texas Center for Superconductivity, University of Houston, Houston, Texas, U.S.A.*

Introduction 214
1. **General** 214
1.1 High-Temperature Superconducting Materials 216
1.1.1 Crystal Structures 217
1.1.1.1 The Fullerites 217
1.1.1.2 The Bismuthates 217
1.1.1.3 The Cuprates: General 217
1.1.1.4 The Cuprates: 214 Compounds 217
1.1.1.5 The Cuprates: R123 Compounds 218
1.1.1.6 The Cuprates: Bi-, Tl-, and Hg-Based Compounds 218
1.1.1.7 The Cuprates: Infinite-Layered Compounds 218
1.1.2 Main Features of High-Temperature Superconductors 219
1.1.2.1 Insulating Parents 220
1.1.2.2 Mixed Valence and CuO_2 Layers 220
1.1.2.3 Strong Anisotropy 220
1.1.2.4 High H_{c2}, Short ε, Low Carrier Density, and Long λ 220
1.1.2.5 Large Thermal-Fluctuation Effects 220
1.2 Ginzburg–Landau Phenomenological Theory 220
1.3 BCS Theory 221
2. **Synthesis** 222
2.1 Polycrystalline Materials 222
2.1.2 Solid-State Reaction 222
2.1.3 Solution-Chemistry Route 223
2.2 Textured Bulk Materials 223
2.3 Wires and Ribbons 223
2.4 Single Crystals 225
2.5 Thin Films 226
2.5.1 Evaporation 226
2.5.2 Sputtering 226
2.5.3 Laser Ablation 227
2.5.4 Molecular-Beam Epitaxy 227
3. **Experimental Properties** 228
3.1 Transport Properties 228
3.1.1 Normal-State Resistivity 228
3.1.2 Critical Current Density 228
3.1.3 Surface Impedance 229
3.1.4 Hall Effect above T_c 229
3.1.5 Thermal Conductivity 230
3.1.6 Thermoelectric Power 230
3.2 Magnetic Properties 230
3.3 Symmetry of the Superconducting Order Parameter 232
4. **Theoretical Status** 233
4.1 Magnetic Models 233
4.2 Electron–Phonon Interaction Models 234
4.3 Marginal Fermi-Liquid Model 234
5. **Applications** 234
5.1 Conductors 235
5.2 Josephson Junctions and SQUIDs 236
5.3 Passive Microwave Devices 238
5.4 Trapped-Field Magnets 239
5.5 Magnetic-Mechanical Devices 239
5.5.1 Magnetic Bearings 240
5.5.2 Vibration Dampers and Magnetic Clutches 241
5.6 Electrodes, Fault-Current Limiters, and Power-Transmission Cables 241
6. **Future Trends** 242
Acknowledgments 244
Works Cited 244
Further Reading 247

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

The discovery of high-temperature superconductivity (HTSy) in 1986 (Bednorz and Müller, 1986) in $La_{2-x}Ba_xCuO_4$ at 35 K and in 1987 (Wu *et al.*, 1987) in $YBa_2Cu_3O_7$ at 93 K, above the boiling point of liquid nitrogen (77 K), ranks among the most exciting developments in modern physics, with profound implications for future technologies. Before 1986, the highest superconducting transition temperature (T_c) was 23 K in Nb_3Ge. Many people considered superconductivity a laboratory curiosity, available only at very low temperatures with the aid of either the rare and expensive liquid helium or the not-so-safe liquid hydrogen. By 1986, the field seemed to have reached full maturity. Low-temperature superconductors (LTSrs) found some practical uses, such as superconducting quantum interference devices (SQUIDs) for detecting electromagnetic signals too weak for conventional sensors, and superconducting magnets for generating extremely strong magnetic fields more efficiently than conventional electromagnets. Nonetheless, their low T_c made it difficult to realize the full technological potential of conventional LTSrs.

With the advent of HTSy, the situation changed drastically. The superconductivity community was revitalized by the challenge to account for the occurrence of superconductivity at these high temperatures and to understand the many unusual normal-state properties of the high-temperature superconductors (HTSrs). These materials raised hope that a wide range of applications could become a practical reality, and they generated great enthusiasm among scientists, engineers, and the general public. Conflicting and poorly reproducible results soon led scientists to uncover significant materials problems and to raise serious questions about the technological future of HTSy. The dedicated efforts of many scientists, fortunately, has yielded steady progress, which has resulted in renewed optimism. Data from different experiments and different laboratories began to converge after scientists synthesized high-quality samples. Advances in the material processing and engineering of HTSrs have attracted the attention of design engineers to consider device integration and system development. Many application concepts have now been demonstrated successfully with HTSrs.

This article focuses on issues relevant to HTSy technology. It will also include a brief account of other interesting aspects of HTSy to give readers a bird's-eye view of this fascinating, rapidly evolving field, which is still in its infancy. As this article is being written, new high-temperature superconducting (HTSg) compounds are being discovered, new theoretical models of HTSy are being proposed, and new applications of HTSy are being demonstrated.

The first section provides a brief history of HTSy, discusses the materials that demonstrate HTSy, and summarizes the macroscopic theory of superconductivity by Ginzburg and Landau and the microscopic theory of superconductivity by Bardeen, Cooper, and Schrieffer. The second section deals with the synthesis of the complex HTSg materials. The third section summarizes physical properties crucial both for understanding HTSy and for applying these interesting materials. The fourth section describes our current theoretical understanding. The fifth section presents the promises and criteria for the application of HTSg devices. The last section briefly describes future scientific and technological trends in the study of HTSy. Finally, several review articles on various aspects of HTSy are listed, following the bibliography.

1. GENERAL

The term "high-temperature superconductors" (HTSrs) refers to materials with a high transition temperature (T_c), where a superconductor undergoes a transition to its superconducting state on cooling, evidenced by a drastic resistance drop and a diamagnetic shift. In this article, the T_c of a superconductor can refer either to the temperature of the onset (T_0) or the completion (T_c) of the superconducting transition, with the former higher than the latter, although they both always represent temperatures below which superconductivity exists. The difference between the two definitions diminishes with improved sample quality. However, the operating temperature of a device made of a superconducting material is usually lower than its T_c in order to enhance the stability, the sensitivity, and the supercurrent carrying capacity of the device. For instance, the supercurrent carrying capacity of a superconductor increases from 0 at T_c to about 90% of its maximum value at $0.7T_c$. Before 1986, superconductors with $T_c > 10$ K, the highest for an element, could be considered HTSrs (Allen, 1982). Since 1986, the term has generally applied only to materials with $T_c > 23$ K, the record of the intermetallic compounds, or above 77 K, the liquid nitrogen boiling point.

Throughout the history of superconductivity, one main goal has been to find compounds with higher T_c's to make applications more practical. However, T_c has risen dramatically only since the discovery of superconductivity in cuprates in 1986, as shown in Fig. 1, with a present record T_c of 134 K at ambient pressure set in early 1993 (Schilling *et al.*, 1993) and T_0 of 164 K demonstrated in a compressed compound in late 1993 (Gao *et al.*, 1994a). Before discovery of the cuprate HTSrs, high T_c was found mainly in intermetallic compounds and alloys, many of which contained transition metals. Hundreds of such compounds and alloys were synthesized (Roberts, 1976) with T_c up to 23 K, largely as a result of the work of Bernd Matthias, the leader in discovery of high-T_c intermetallics until his untimely death in 1980. In 1953, he proposed the most successful guide to high T_c during the ensuing 33 years: the "Matthias empirical rule" (Matthias, 1953), which correlates T_c with the ratio of number of valence electrons to number of atoms, with the maximum T_c to occur at valence electron-per-atom ratios of $\sim$4.75 or $\sim$6.4. For instance, Nb_3Ge (T_c = 23 K) has a valence electron-per-atom ratio of $\sim$4.75. During this period, many other researchers also tried to raise T_c, primarily with experiments based on the BCS theory (Bardeen *et al.*, 1957), but they had little success.

Disappointed by unsuccessful attempts to raise T_c of intermetallic compounds above 23 K, Bednorz and Müller (1986) decided to examine metallic oxides. According to BCS theory (Bardeen *et al.*, 1957), a strong electron–phonon coupling will lead to a high T_c. Since strong electron–phonon interactions can happen in oxides, they reasoned that higher T_c might occur in nickelates and cuprates. Indeed, they detected a T_c up to 35 K, as evidenced by a resistivity drop, in mixed-phase

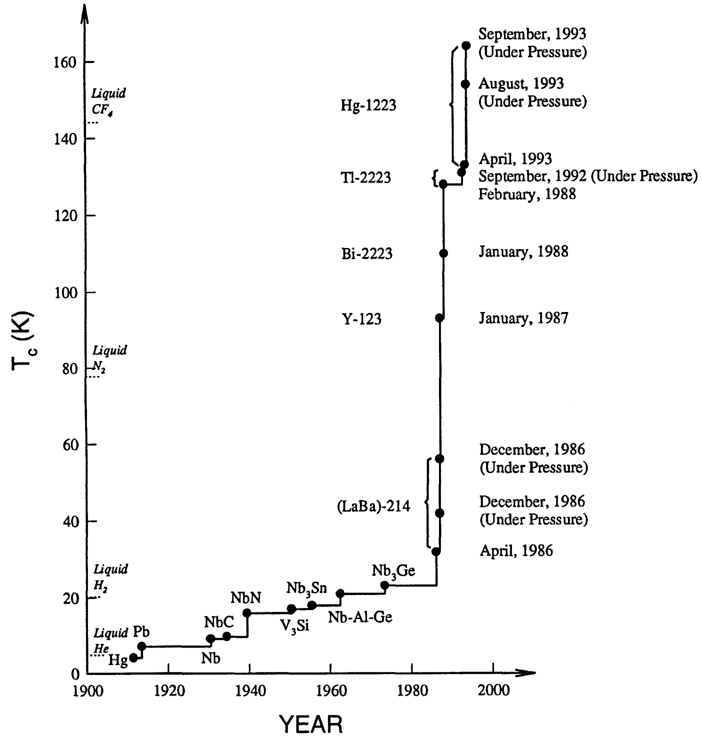

FIG. 1. Evolution of the maximum observed superconducting transition temperature (T_c) with time.

samples of La-Ba-Cu-O, and published their results in the September 1986 issue of *Zeitschrift für Physik* (Bednorz and Müller, 1986). The appearance of superconductivity at such a high temperature in an oxide surprised many experts. Partially because of its modest title, "Possible High T_c Superconductivity in the Ba-La-Cu-O System," their report did not attract much attention until the independent confirmations in December 1986 by research groups at the University of Tokyo and at the University of Houston. The Tokyo group identified the superconducting phase as $La_{2-x}Ba_xCuO_4$. Bednorz and Müller won the 1987 Nobel prize in physics a little over one year after the publication of their discovery.

In attempts to raise T_c, the Houston group made three interesting observations (Chu, 1987) in late 1986 and early 1987:

1. pressure enhancement of T_c to ~52 K in $La_{2-x}Ba_xCuO_4$ at an unusually high rate of ~1 K/kbar;
2. drastic resistivity drop at 80 K in mixed-phase La-Ba-Cu-O samples at ambient pressure, but not in $La_{2-x}Ba_xCuO_4$; and
3. a large diamagnetic signal persisting up to 90 K in a freshly made mixed-phase La-Ba-Cu-O sample at ambient pressure, but not in an aged one.

The observations suggested that a material different from $La_{2-x}Ba_xCuO_4$ with $T_c > 77$ K must be possible. The question was how to stabilize and identify the $T_c > 77$ K phase. By reviewing data at ambient and high pressures, the Houston group concluded that replacing La with chemically similar but smaller elements such as Y might help stabilize the high-T_c phase. In late January 1987, Wu *et al.* (1987) detected stable superconductivity at 93 K in a mixed-phase Y-Ba-Cu-O material. Several laboratories subsequently confirmed these results. Shortly afterward, the structure of the superconducting phase was identified as $YBa_2Cu_3O_7$ (Hazen, 1989). Within a few weeks, a series of isostructural compounds $RBa_2Cu_3O_7$ with R = Y, La, Nd, Sm, Eu, Gd, Dy, Ho, Er, Tm, Yb, or Lu was found (Hor *et al.*, 1987; Murphy *et al.*, 1987) to be superconducting with $T_c \sim 90$ K. A worldwide race to raise the T_c, to understand the science, and to make use of HTSrs was on.

Meanwhile, several research groups investigated the replacement of the trivalent rare earth element with trivalent Bi and Tl. Only a few weeks into 1988, Maeda *et al.* (1988) observed superconductivity above 100 K in mixed-phase Bi-Sr-Ca-Cu-O. Shortly afterward, the superconducting phases were identified as $Bi_2Sr_2Ca_{n-1}Cu_nO_{2n+4+\delta}$ with $n = 2$ and 3 (Hazen, 1989). Two weeks after the Tsukuba announcement, Sheng and Hermann (1988) reported a T_c above 120 K. The superconducting phases were immediately isolated and characterized (Parkin *et al.*, 1988) as $Tl_2Ba_2Ca_{n-1}Cu_nO_{2n+\delta}$ with $n = 1$, 2, and 3, with a T_c up to 125 K in the $n = 3$ member. In 1992, the T_c of this compound was raised to 131 K by the application of pressure (Berkeley *et al.*, 1993).

In March 1993, Putilin *et al.* (1993) synthesized $HgBa_2CuO_{4+\delta}$ with a $T_c = 94$ K. By adding Ca to the Hg-Ba-Cu-O material, Schilling *et al.* (1993) detected superconductivity at a new record T_c of 134 K in mixed-phase samples. The superconducting phases of $HgBa_2Ca_{n-1}Cu_nO_{2n+2+\delta}$, where $n = 1$, 2, and 3, were separated (Chu, 1994) with the record T_c in the $n = 3$ member. An unusually large increase in the T_c of this same compound to 164 K at a pressure of 30 GPa was observed (Gao *et al.*, 1994a), demonstrating that superconductivity can exist above 160 K in cuprates. The challenge is to retain such a high T_c without pressure by chemical and/or other physical means to make this material practical for applications.

There are still no rules for the discovery of higher T_c's. The search for new superconductors has always required people with strong intuition, courage, and even faith. The hard work of those few has resulted in the exciting progress shown in Fig. 1.

1.1 High-Temperature Superconducting Materials

Since 1987, more than 90 nonintermetallic superconducting compounds were found with T_c above the 23 K record for the intermetallics. HTSrs belong to three families: cuprates, bismuthates, and fullerites. The cuprates at ambient pressures (Schilling *et al.*, 1993) have an anisotropic layered structure and a maximum $T_c \sim 134$ K; the bismuthates (Cava *et al.*, 1988) and fullerites (Hebard *et al.*, 1991) have a cubic structure with a maximum $T_c \sim 30$–39 K. We describe the basic crystal structures of the HTSrs at room temperature and refer readers to special articles (Chen, 1990) for the temperature-induced structural transformation and detailed microstructure in these compounds. We also briefly discuss the main features of HTSrs that are responsible for some of their unusual properties to be described later in Sec. 3.

1.1.1 Crystal Structures

1.1.1.1 The Fullerites. Compounds A_3C_{60}, with A = K, Rb, or Cs, belong to the newest family of HTSrs. They are formed by filling the interstitial sites of the face-centered cubic lattices of C_{60} (Fig. 2). The crystal structure is cubic and therefore isotropic, with an increasing lattice parameter as the atomic radius of A increases (Hebard, 1993).

1.1.1.2 The Bismuthates. The compounds $Ba_{1-x}A_xBiO_3$ (BABO), where A = K or Rb, display the perovskite structure (Fig. 3). The undoped semiconducting $BaBiO_3$ has a monoclinic symmetry, but doping transforms the superconducting monoclinic phase to an orthorhombic phase and finally to a superconducting cubic one. The cubic lattice parameter decreases with doping (Pei *et al.*, 1990).

1.1.1.3 The Cuprates: General. Cuprates account for the great majority of the known HTSrs and include all materials that superconduct above 77 K. They have a layered structure closely related to the perovskite lattice. The structure consists of CuO_2 layers with intervening metal and/or metal–oxygen layers that serve as the carrier reservoir. The CuO_2 layers are considered to be the active components of the HTSrs, displaying three different types of Cu–O coordination. They form arrays of CuO_4 squares, or they form CuO_5 pyramids and CuO_6 octahedra, respectively, by including an apical oxygen only above and both above and below the CuO_2 layers (Fig. 4). The intervening layers, which serve as the charge reservoir, may have much greater structural complexity than the CuO_2 layers. The crystal structures of the most-studied cuprate HTSrs are shown below (Hazen, 1989). Others are basically derivatives of them.

1.1.1.4 The Cuprates: 214 Compounds. The compounds $(R, R')_{2-x}A_xCuO_4$ exhibit a T_c up to ~36 K where R and R' are rare-earth elements and can be the same element, and the dopant A = Ba, Sr, Ca, Ce, or Th. Depending on the ionic radius of R and R', three structural phases, T, T', and T^*, are stabilized at 300 K (Fig. 5). They contain one CuO_2 layer per formula unit. The CuO_2 layer comprises arrays of CuO_6 octahedra in the T phase, CuO_5 pyramids in the T^* phase, and

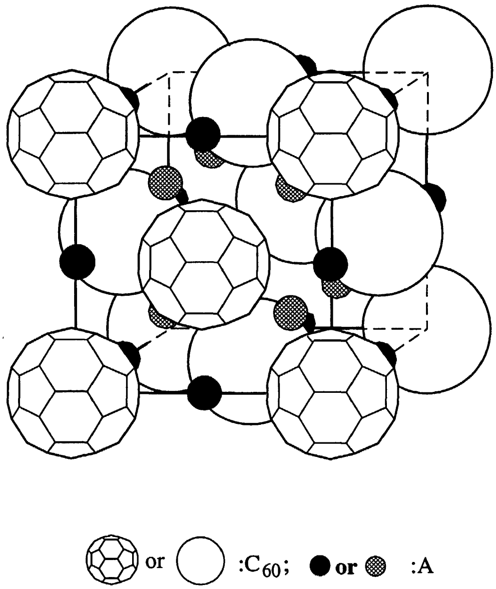

FIG. 2. The cubic crystal structure of fullerite HTSrs: A_3C_{60} (A = K, Rb, or Cs).

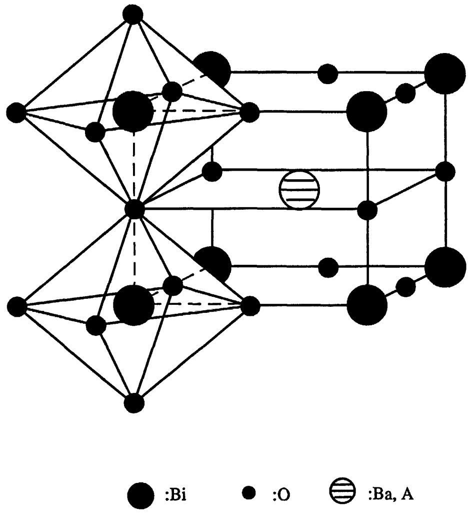

FIG. 3. The perovskite crystal structure of bismuthate HTSrs: $Ba_{1-x}A_xBiO_3$ (A = K or Rb).

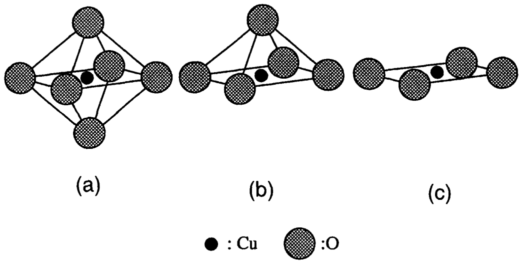

FIG. 4. Three types of CuO_2 layers displaying different Cu–O coordinations: (a) CuO_6 octahedron, (b) CuO_5 pyramid, and (c) CuO_4 square.

CuO_4 squares in the T' phase. These HTSrs display a tetragonal symmetry.

1.1.1.5 The Cuprates: R123 Compounds. $RBa_2Cu_3O_7$ (also known as RBCO) show a T_c up to ~100 K where R = Y or any of the rare-earth elements except Ce and Tb. The crystal structure of the 123 compound is orthorhombic (Fig. 6). It consists of two CuO_2 layers and one CuO-chain layer per unit formula. The oxygen in the CuO chain is only loosely bound. When the oxygen content is less than ~6.4, the oxygen atoms in the CuO chain lose their long-range order, and the compound becomes tetragonal and nonsuperconducting. $PrBa_2Cu_3O_7$ is superconducting only in thin-film form, and then only when Pr is partially replaced by Ca.

1.1.1.6 The Cuprates: Bi-, Tl-, and Hg--Based Compounds. The compound $Bi_2Sr_2Ca_{n-1}Cu_nO_{4+2n}$ [Bi-22(n − 1)n or BSCCO] displays a T_c up to ~120 K; $Tl_mBa_2Ca_{n-1}Cu_nO_{2+m+2n}$ [Tl-m2(n − 1)n or TBCCO] shows a T_c up to 125 K; and $HgBa_2Ca_{n-1}Cu_nO_{2n+2+\delta}$ [Hg-12(n − 1)n or HBCCO] exhibits a T_c up to ~134 K at ambient, where m = 1 or 2, n = 1, 2, 3, 4, ... The n represents the number of CuO_2 layers in the compound per unit formula. Figure 7 shows the crystal structures for the Tl-22(n − 1)n, Tl-12(n − 1)n, and Hg-12(n − 1)n for n = 3. Notice that the linear Hg–O coordination and the vacant site for oxygen in the HgO_δ layers of Hg-12(n − 1)n, which do not appear in other cuprate HTSrs, may be responsible for the record T_c of 164 K in Hg-1223 at 30 GPa. The crystal symmetry is either orthorhombic or monoclinic for Bi-2201, pseudotetragonal for Bi-2212 and -2223, and tetragonal for all Tl- and Hg-based compounds.

1.1.1.7 The Cuprates: Infinite-Layered Compounds. The compound $Sr_{1-x}M_xCuO_2$, where M = Ca, La, Nd, Pr, or vacancies, shows a T_c up to 100 K. These compounds can be considered doped members of

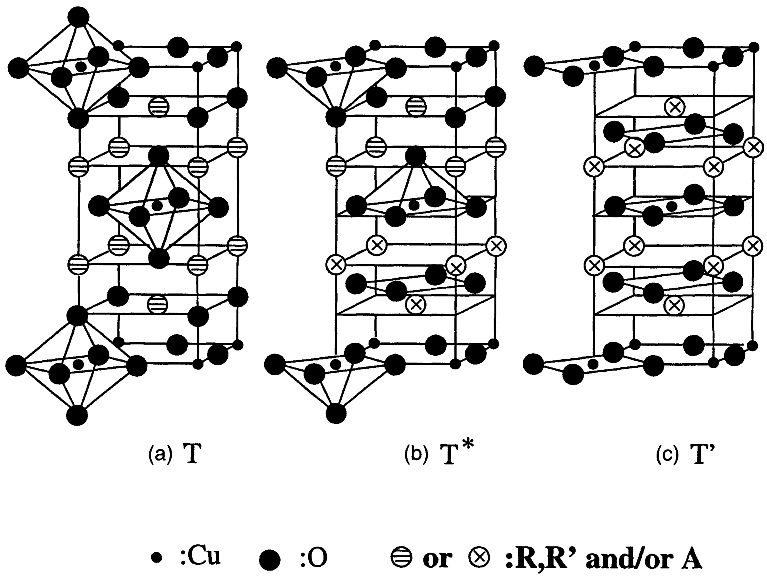

FIG. 5. The crystal structures of cuprate 214 compounds: T, T^*, and T' phases (R, R' are rare-earth elements and A = Ba, Sr, Ca, Ce, or Th).

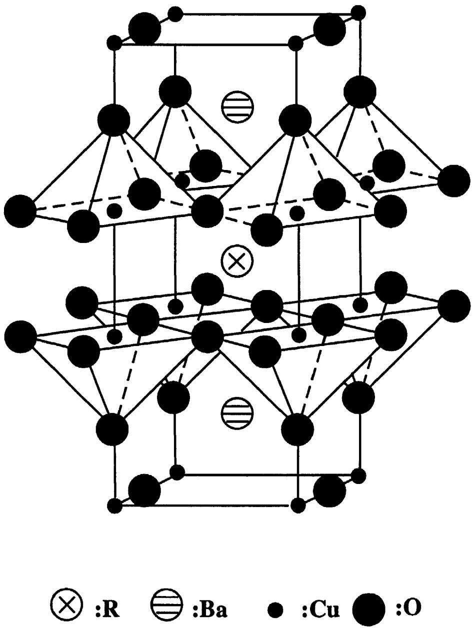

FIG. 6. The crystal structure of R123 compounds (*R* = rare earths other than Ce and Tb).

$SrCuO_2$, which belongs to the homologue series of $A_{n-1}Cu_{n+1}O_{2n}$ as $n \rightarrow \infty$, where A = Ca, Sr, or Ba. Therefore they are called the infinite-layered compounds. They consist of the simplest intervening Sr layers between CuO_2 layers with a tetragonal structure and may thus be considered the parent compounds of all HTSrs (Fig. 8). The superconducting phase can be stabilized in thin-film form by sputtering, or in bulk form under high pressures >2.5 GPa. The samples display a $T_c \sim 30$ K when doped with La, Nd, or Pr and up to 100 K when doped with Ca or vacancies. The detailed local structure of the $T_c > 30$ K phase is complex and has not yet been determined. Tantalizing resistive and magnetic anomalies suggestive of superconductivity up to 180 K have been reported (Li, *et al.*, 1992) in a thin-film sample of $Sr_{0.7}Ca_{0.3}CuO_2$ prepared by the laser MBE technique.

1.1.2 Main Features of High-Temperature Superconductors Similar to LTSrs, HTSrs below their T_c's display the same basic superconducting characteristics, i.e., resistivity $\rho = 0$, magnetic induction $B = 0$ when the applied field is smaller than the lower critical field H_{c1}, and electron pairing. However, HTSrs are rather different from LTSrs and pose serious challenges to the un-

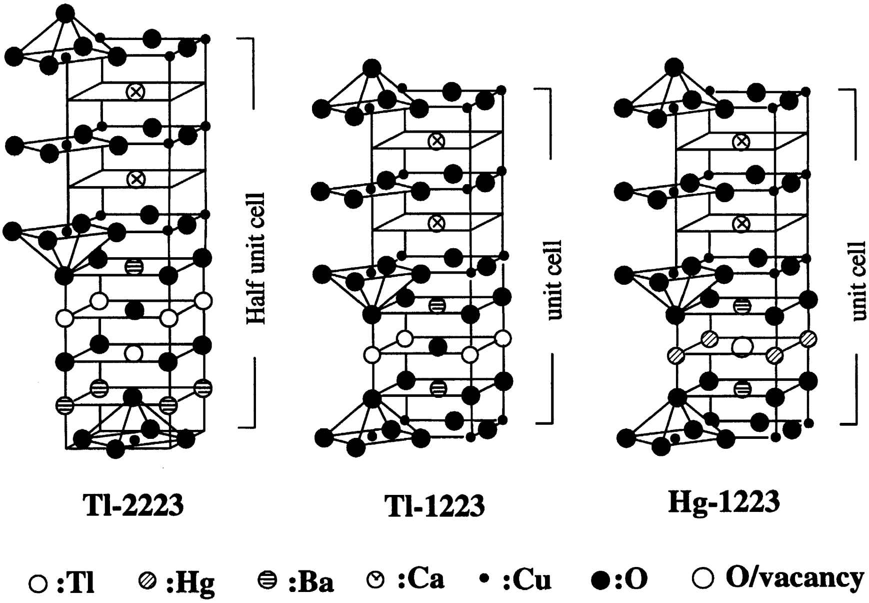

FIG. 7. The crystal structures of Bi-, Tl-, Hg-based cuprate HTSrs [Tl-22(*n* − 1)*n*, Tl-12(*n* − 1)*n*, Hg-12(*n* − 1)*n*] for *n* = 3 CuO_2 layers per unit formula.

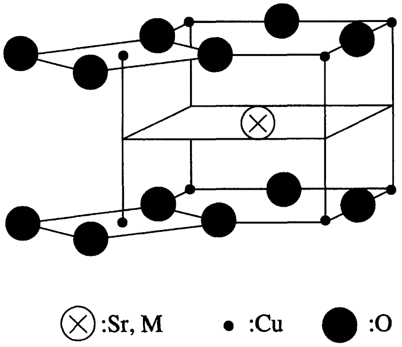

FIG. 8. The crystal structure of the infinite-layered cuprate $Sr_{1-x}M_xCuO_2$ (M = Ca, La, Nd, Pr, or vacancies).

derstanding and application of HTSy because of the following features.

1.1.2.1 Insulating Parents. Without exception, all HTSrs can be derived from their respective insulating parent compounds, through either hole or electron doping over a narrow doping range. The parent compound is an insulator when doping is below this narrow doping range and becomes a nonsuperconducting metal above. T_c depends sensitively on doping, which, in turn, is sensitive to the valence and ionic size of the dopant element, and on the synthesis conditions.

1.1.2.2 Mixed Valence and CuO_2 Layers. All HTSrs except fullerites contain elements in a mixed valence state, i.e., Cu in the cuprates and Bi in the bismuthates, and all cuprate HTSrs possess CuO_2 layers. Unfortunately, none of these common characteristics has been demonstrated to be sufficient for the existence of HTSy. More subtle factors must determine how HTSy occurs.

1.1.2.3 Strong Anisotropy. In contrast to most LTSrs, cuprate HTSrs are highly anisotropic electrically and magnetically, because of their layered crystal structure. To be more specific, electric current in both the normal and superconducting states flows more easily along the CuO_2 layers than perpendicular to these layers. The resistivity ratio between these two directions reaches $\sim 10^5$ for the case of Bi-2212 slightly above its T_c. A critical current density (J_c) ratio as high as $\sim 10^3$ has been obtained between these two directions.

1.1.2.4 High H_{c2}, Short ξ, Low Carrier Density, and Long λ (see Sec. 1.2 for definitions). HTSrs usually have a high upper critical field H_{c2}, e.g., >100 T for Y123 at 0 K and a low carrier density of ~(2 to 5) $\times 10^{21}$ cm^{-3}. As a result, the coherence length ξ is very short, ~17 and ~3 Å in and perpendicular to the CuO_2 plane, respectively; and the penetration depth λ is rather large, ~2 000 and ~10 000 Å, respectively. HTSrs are therefore extreme type-II superconductors with a Ginzburg–Landau parameter $\kappa \sim 100$.

1.1.2.5 Large Thermal-Fluctuation Effects. Because of the high T_c, short ξ, and layered structure of a HTSr, microdomains in a HTSr can fluctuate between their normal and superconducting states over a wide and easily detectable T region both above and below its T_c. This leads to an easily detectable complex magnetic-phase diagram for HTSrs, reduced magnetic flux pinning, and, hence, reduced J_c at high temperatures. For instance, J_c is almost zero at 77 K for Bi-2223 single crystals, in spite of their relatively high T_c of ~120 K.

1.2 Ginzburg–Landau Phenomenological Theory

The Ginzburg–Landau phenomenological theory (de Gennes, 1966) links the electrodynamic with the thermodynamic properties of a superconductor. In 1950, Ginzburg and Landau introduced a macroscopic quantum wave function $\Psi(r)$ with amplitude and phase to embody the quantum-mechanical nature of the superconducting state. The density of the superconducting electrons is proportional to $|\Psi(r)|^2$. Following the Landau theory of second-order phase transition, Ginzburg and Landau intuitively expressed the Gibbs free energy in a power series of $\Psi(r)$ and $\nabla\Psi(r)$ at temperatures close to T_c. They minimized the Gibbs free energy and obtained two intercoupled equations to describe an equilibrium superconducting system in the presence of a magnetic field.

The theory also provides many valuable estimates, such as the maximum critical current density (known as the depairing critical current density) when the kinetic energy of the carriers is equal to the superconducting condensation energy, $J_d = \Phi_0/12\sqrt{3}\pi^2\lambda^2\xi$, and the upper critical field $H_{c2} = \Phi_0/8\pi^2\xi^2$. In these relations, Φ_0 is a flux quantum that equals $h/2e \sim 2.07 \times 10^{-15}$ T m^2, where h is the Planck constant, e the electron charge, λ

the magnetic penetration depth, and ξ the coherence length. λ measures the extent that a magnetic field can penetrate into a superconductor, and ξ is roughly the distance between two electrons in a superconducting pair. The Ginzburg–Landau parameter also predicts a temperature-independent quantity $\kappa = \lambda/\xi$, called the Ginzburg–Landau parameter, which differentiates a type-I superconductor with $\kappa < 1/\sqrt{2}$ from a type-II superconductor with $\kappa > 1/\sqrt{2}$. A type-I superconductor is one that exhibits the Meissner effect; i.e., it completely expels magnetic flux, i.e., $B = 0$, until it becomes completely normal when the applied magnetic field (H) exceeds a critical value of H_c. Similar to a type-I superconductor, a type-II superconductor completely expels the magnetic flux but only when H is smaller than the lower critical field H_{c1}. However, the magnetic field starts to penetrate into a type-II superconductor when it is greater than H_{c1} until it reaches an upper critical field H_{c2}, beyond which superconductivity is completely destroyed (Fig. 9). The basic superconducting phenomena of HTSrs are similar to those of LTSrs and therefore have been accounted for quite well by the Ginzburg–Landau phenomenological theory.

1.3 BCS Theory

The Ginzburg–Landau phenomenological macroscopic theory can describe superconducting phenomena, but it cannot explain why and how they occur. Fröhlich (1950) predicted that lattice vibrations, known as phonons in their quantum form, can give rise to an effective attraction between the otherwise repulsive electrons and thus are important to superconductivity. Cooper (1956) showed that, in the presence of an attractive interaction, two electrons will form a bound state or a "Cooper pair" having a lower energy state, no matter how weak the interaction is. With these and other observations, together with the appropriate mathematical tools, the time was ripe in 1957 for Bardeen, Cooper, and Schrieffer to develop the first comprehensive microscopic theory of superconductivity (Bardeen *et al.*, 1957). Known as the BCS theory, it provided the first microscopic description of the origin of superconductivity. Later it was shown that the previous macroscopic theories can be derived from this microscopic theory. The three architects of the BCS theory were awarded the Nobel prize in physics in 1972. (See SUPERCONDUCTIVITY, LOW-TEMPERATURE, for details.)

According to the BCS theory, the ground state consists of Cooper pairs of electrons near the Fermi surface. It gives a superconducting transition temperature

$$T_c = 1.14\Theta_D \exp[-1/N(E_F)V],$$

where Θ_D is the Debye temperature, $N(E_F)$ the electron density of states at the Fermi energy (E_F), and V the strength of the attractive electron–phonon interaction. The BCS theory provides expressions for a wide range of properties of a superconductor, including the isotope effect, the penetration depth, and the energy gap (which is the minimum amount of energy required to split a Cooper

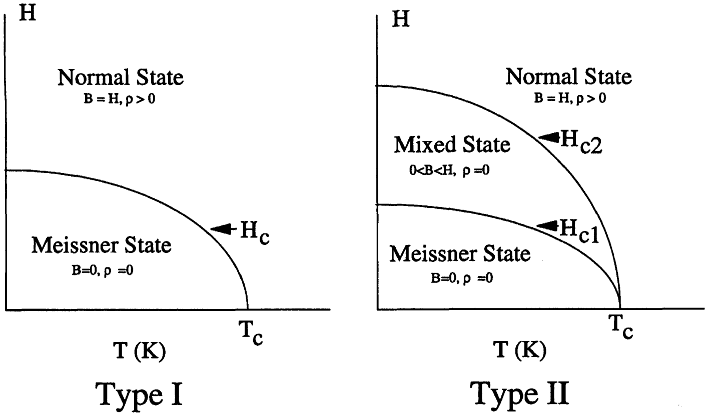

FIG. 9. Dependence of critical magnetic fields on temperature for type-I and type-II superconductors: in the Meissner state, $B = 0$ and $\rho = 0$; in the mixed state, $0 < B < H$ and $\rho = 0$; and in the normal state, $B = H$ and $\rho > 0$.

pair). The theory also accounts for many other superconducting characteristics such as tunneling, thermal conductivity, and flux quantization. Deviations from BCS predictions have been detected in LTSrs and explained over the years by modified models, which are still primarily based on the BCS theory. However, it remains an open question if BCS in its original or modified form can explain all properties of a HTSr and if an attractive electron–phonon interaction can give rise to the high T_c observed (see Sec. 4 for HTSy theoretical models).

2. SYNTHESIS

Compounds demonstrating the HTSg properties can be easily prepared in any laboratory with a mortar and pestle and a simple furnace. Perhaps this is why HTSy is the only Nobel-prize winning experiment that was quickly duplicated in high school laboratories. However, the synthesis of HTSg compounds with superconducting properties good enough for scientific and applied studies requires great effort. The complex chemistry, the anisotropic structure, and the short coherence length of the cuprate HTSrs complicate the preparation of high-quality materials. The general synthesis conditions for some representative HTSg compounds are listed in Table 1.

2.1 Polycrystalline Materials

2.1.2 Solid-State Reaction Solid-state reaction (Yan *et al.*, 1988), the simplest technique, can produce high-quality polycrystalline HTSrs. It consists of three basic steps:

1. mixing reactants in the appropriate cation stoichiometry to obtain a homogeneous precursor,
2. calcination (or reaction) of the precursor to obtain the HTSg compound, and
3. heat treatment to optimize the HTSg properties.

The reactants usually are constituent oxide, carbonate, or nitrate powders. To avoid possible residual carbon, which usually degrades the superconducting properties, oxides or nitrates are the better choice. Homogeneous polycrystalline HTSg samples can be pre-

Table 1. General synthesis parameters for polycrystalline HTSrs.

	Calcination			Sintering			Heat Treatment		
Compound	Temperature (°C)	Atmosphere	Time (h)	Temperature (°C)	Atmosphere	Time (h)	Temperature (°C)	Atmosphere	Time (h)
214	900–950	Air	10–20	1100–1170	Air or O_2	24–100	500	O_2	24–100
123	750–850	Air	10–24	920–950	O_2	12–24	400–500	O_2	10–24
BSCCO	750–850	Air	10–20	840–900	Air or reducing	20–100	750	Reducing or O_2	20–40
TBCCO	Not required			850–925	O_2	3–10	450–750	O_2 and/or Ar	1–15
BKBO	Not required			725–800	N_2 or inert gas	1–5	475–550	O_2	4
$Sr_{1-x}CuO_2$	Not required			1000–1200	High-pressure synthesis	0.5–2			
HBCCO-1201	Not required			780–820	Vacuum		240–300	O_2	3–24
HBCCO-1212	Not required			780–820	Vacuum (or high pressure)	0.5–5	240–300	O_2	3–24
HBCCO-1223	Not required			850–860	Vacuum (or high pressure)	0.5–5	240–300	O_2	3–24
HBCCO-(1201–1256)	850–1000 (with oxidizing agent $KClO_4$ during processing)	High pressure	0.5–5	Not required			Not required		

pared by thorough mixing and calcination, followed by repeated pulverization, compaction, and sintering. However, the solid-state reaction route is tedious and time consuming, and it risks contamination due to repeated grinding and pulverization.

2.1.3 Solution-Chemistry Route The solution-chemistry route (Yan *et al.*, 1988) offers a good way to mix the reactants at the molecular level in solution and provides good control of chemical homogeneity, stoichiometry, and particle size and morphology. The techniques usually start with a mixed solution of salts of the metal elements and produce the particles by precipitation or by solvent removal. The precipitated particles need further calcination to form the HTSg compound. The techniques can be grouped into three general categories: co-precipitation, sol-gel, and spray drying. In co-precipitation, the cations are precipitated simultaneously from solution. The sol-gel method involves the formation of ultrafine particles dispersed in a solvent that is slowly evaporated resulting in excellent homogeneity control. Spray drying involves the formation of small spherical droplets of the solution with an atomizer and drying them by spraying them either into liquid nitrogen or onto a hot surface.

2.2 Textured Bulk Materials

For applications, such as electrodes, electromechanical devices, and trapped-field magnets (see Sec. 5.6), it is essential to raise the critical current density (J_c) of the polycrystalline samples produced by the above techniques. Because of their large structural and electrical anisotropy and their complex chemistry, HTSrs are prone to formation of weak links, resulting from the presence of high-angle grain boundaries, microcracks, off-stoichiometric regions, chemical inhomogeneity, and impurity phases. These Josephson-coupled (see Sec. 5.2) weak links restrict J_c: The J_c of a polycrystalline HTSr is low and decreases rapidly with magnetic field.

Cuprates usually decompose to a mixture of solid and liquid phases above a temperature known as the peritectic decomposition temperature (T_p), which is typically 1320 K. The melt-texturing process (Murakami *et al.*, 1989; Salama *et al.*, 1989) was developed to reduce the number of weak links in cuprates. This process involves heating a HTSr rapidly to above T_p followed by slow cooling to below T_p. After melt texturing, Y123 grains are formed with their CuO_2 planes preferentially aligned. At 77 K, J_c's up to $\sim 8 \times 10^4$ A/cm^2 and $\sim 2 \times 10^3$ A/cm^2 have been achieved in melt-textured Y123 at zero and 30 T, respectively. Unfortunately, the grains so obtained are small. Melt-textured BSCCO, TBCCO, and HBCCO have not yet been achieved.

To enhance the grain size in a large sample of RBCO, one may adopt the directional solidification technique (Meng *et al.*, 1990), which requires that during melt texturing one maintains a temperature gradient, a compositional gradient of R, and/or a stack of thin layers of RBCO with different R's across the bulk precursor in the c-axis direction intended during melt texturing. Samples exhibiting a large J_c with a length of a few centimeters but only a small cross-sectional area of ~ 0.2 cm^2 have been obtained.

A "top-seeding technique" (Sawano *et al.*, 1991) can control the crystal nucleation and growth during melt texturing: It is essential for obtaining large quasi single-grain HTSrs. The technique involves introducing a small single-grain HTSr as a seed to the top of the HTSr precursor, with or without a temperature gradient, before melt texturing. The seed must have a melting or decomposition point above T_p of the specific HTSr disk processed. HTSr disks of 3 cm diam. × 1 cm thick have been obtained (Plate 1), to trap a field of 1.5 T at 77 K and 6 T at 40 K and to shield a field greater than 21 T at 4.2 K (Gao *et al.*, 1994b).

2.3 Wires and Ribbons

For many large-current applications, such as generators, motors, power-transmission lines, and high-field magnets, the most practical forms of HTSrs are wire and ribbon. For several applications, the requirements on J_c, the total critical current (i.e., $J_c \times$ cross section), and length of the HTSg wires and ribbons are shown in Fig. 10. Since HTSg materials are mechanically brittle and electrically anisotropic, the fabrication of HTSg wires or ribbons with large J_c and great length is a formidable task. The powder-in-tube (PIT) method (Sato *et al.*, 1991) has

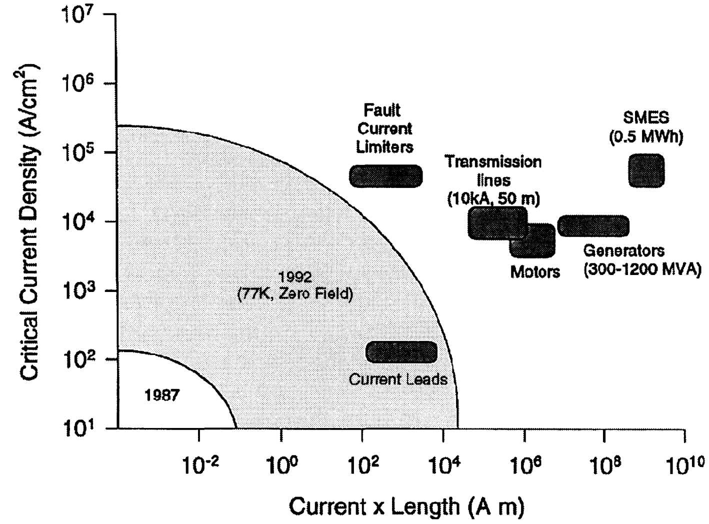

FIG. 10. The conductor characteristics required for various commercial applications. The shaded area represents the developments between 1987 and 1992, and SMES means superconducting magnetic-energy storage (Electric Power Research Institute Report, 1992).

yielded short BSCCO ribbons that display large $J_c \sim 10^5$ A/cm^2 at 4.2 K, even at fields higher than 20 T. BSCCO ribbons of length of more than 1 km have been prepared by American Superconductor Corporation in the United States and Sumitomo in Japan with an overall J_c of better than 8×10^3 A/cm^2 at 77 K. Very recently, American Superconductor Corporation made BSCCO multifilament wire over 1 km long with a $J_c \sim 14 \times 10^3$ A/cm^2 at 77 K. The PIT method involves filling a metal tube with powder of the HTSg compound, drawing the filled tube to wire of smaller diameter, and rolling the wire to ribbon, followed by repeated compression and heat treatment (Fig. 11). A multifilament wire can be made by bundling together the HTSg wire made by the PIT method through a multistage process. The starting HTSg material is ground into fine particles of uniform size. Pb dopant in BSCCO facilitates the formation of the pure Bi-2223 phase, which shows the best J_c of the BSCCO-homologous series. Repeated drawings not only reduce the diameter of the HTSg wire but also enhance the packing density of HTSg powder inside the tube for better texturing in subsequent processing. The rolling step further reduces the overall thickness of the HTSg component inside the ribbon and, more importantly, induces partial grain alignment with the c axis perpendicular to the ribbon

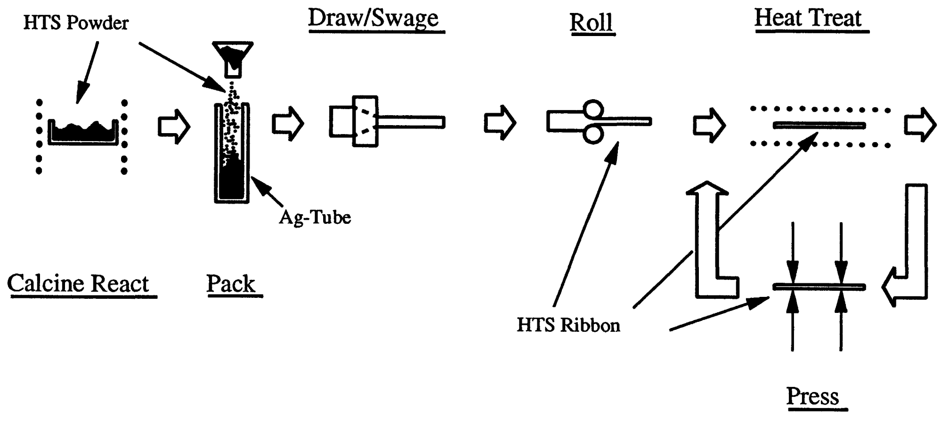

FIG. 11. The powder-in-tube technique for the production of HTS ribbons.

surface. The final compression is also important for grain alignment. A representative set of parameters and conditions for processing silver-sheated BSCCO is given in Table 2.

The cladding metal used should not melt or react with the HTSg material at the processing temperature. At the same time, it must be permeable to oxygen. Silver, palladium-doped silver, and gold are the only suitable choices for BSCCO ribbons.

The major problem with BSCCO wire is the drastic drop of its J_c at 77 K resulting from motion of the flux lines under the influence of current due to the weak flux-pinning force. Flux pinning at 77 K is much stronger in Y123, Tl-1223, and HBCCO. For Y123, the PIT method fails: because of the strong coupling between the CuO_2 layers in Y123, it cracks during rolling, rather than flowing as BSCCO does. The chemical instability of HBCCO at temperatures close to the ribbon processing temperature poses a serious challenge to ribbon making. Promising preliminary results have recently been reported for Tl-1223 thick films on a short Ag ribbon with a J_c up to ~6.7 × 10^4 A/cm^2 at 77 K by the National Renewable Energy Laboratory, General Electric, and Oak Ridge National Laboratory; and for Y123 thick films on a short flexible Ni strip with a J_c of ~1 × 10^6 A/cm^2 at 75 K by the Los Alamos National Laboratory. The Tl-1223 ribbon was made by properly heating the atomic-size precursor oxides deposited electrochemically on the Ag substrate. The Y123 thick film on the Ni strip was prepared by the laser ablation technique, to be described in Sec. 2.5.3, using a buffer layer with a good crystalline alignment.

2.4 Single Crystals

Large single crystals are best for determining the intrinsic physical properties of a material. The growth of large HTSg single crystals is difficult because the material melts incongruently; i.e., the concentration of the intended HTSr is different from that of the melt. Additional problems include the corrosive nature of the melt, which makes it difficult to find an inert crucible material, and the lack of detailed knowledge of the phase diagram. Relatively large 214-HTSg single crystals of high quality have been obtained. Crystals of Y123 (up to ~10 × 9 × 9

Table 2. Power-in-tube BSCCO-wire processing parameters.

Compound	Composition					Precalcination (°C)	Atmosphere	Wire Treatment (°C)	Time (h)	J_c (10^4 A/cm^2)	
	Bi	Pb	Sr	Ca	Cu					4.2 K	77 K
Bi-2223(*a*)	1.6–1.9	0.32–0.4	1.9–0.4	2.0–2.2	3.0–3.1	800–830	Air or ~8% O_2	820–840	100–200	20	5.4
Bi-2212(*b*)	2	0	2	1	2	800–820	Air or ~8% O_2	880–890	0.5–50	20	3.0

mm^3) and thin BSCCO and TBCCO crystals (~0.5 mm thick) are also available with an acceptable quality for detailed studies. However, few single crystals of practical size of other compound systems exist. Large, contamination-free, high-quality single crystals of materials other than 214 suitable for detailed neutron-spectroscopy studies have yet to be synthesized.

Various growth techniques for producing high-T_c crystals have been tested and yield some results (Schneemeyer, 1993). The common-flux technique has been used to grow single crystals from two possible types of HTSr saturated solution: one solution consisting of a flux of one component of the HTSg compound such as CuO or $BaCuO_2$ in the R123 or 214 compounds and the other an inorganic salt such as KCl or KOH. Single crystals of the R123 or 214 compound have been grown by slowly cooling the saturated solution inside a crucible (Al_2O_3, Zr_2O_3, Au, Pt, or Ag/Pd). Although the method is simple, it suffers from crucible contamination. The best crucible materials appear to be Au, Ag/Pd, and Zr_2O_3 alloy in terms of chemical stability, and Pt in terms of resistance to thermal shock.

Recently, a continuous pulling method has been developed to grow single crystals of Y123 from a solute-rich liquid. Y211 powder is first introduced as a solute at the bottom of a crucible that is made of Y_2O_3 or MgO. A constant temperature gradient of several degrees higher at the bottom of the crucible is maintained, while the surface of the solvent, melted oxides of Ba and Cu, is kept at ~1000 °C. The seed crystals used are single-crystal MgO with appropriate surface treatment. High-quality Y123 crystals with large facets ~10 mm along the *c* axis and 9×9 mm^2 along the *a-b* plane have been grown.

2.5 Thin Films

HTSg thin films are essential for many applications that involve small currents, such as microelectronics and ultrasensitive sensors. They are also crucial for fundamental studies such as tunneling spectroscopy, proximity effect, anisotropy effect, grain-boundary effect, and flux pinning. High-quality thin films must possess the correct stoichiometry, the correct crystal structure with a proper orientation, chemical homogeneity, uniform thickness, smooth surface morphology, proper microstructure, and no impurity phase. Epitaxially grown films that satisfy the above criteria display excellent superconducting properties, e.g., high T_c, narrow transition width, high J_c, low surface impedance (R_s) at high and low frequencies, and high chemical stability. Excellent Y123 and TBCCO thin films of a couple of thousand angstroms thick have been prepared with a $J_c \sim (3 \text{ to } 5) \times 10^6$ A/cm² at 77 K. Multilayered structures with up to 15 epitaxial layers consisting of Y123 with buffer layers and insulators have been fabricated by Conductus in California. High-quality BSCCO films have been made by Varian in California.

Careful consideration must be given to the selection of substrate materials. Lattice match, thermal compatibility, absence of HTSr/substrate interdiffusion, and, for high-frequency applications, low dielectric constant and low loss tangent are all issues to consider. Common substrate materials include $LaAlO_3$, MgO, $SrTiO_3$, Sr_2AlNbO_6, Sr_2AlTaO_6, and sapphire.

2.5.1 Evaporation The starting source materials, which can be elements and/or compounds, are thermally evaporated (Mankiewich *et al.*, 1987) from a crucible by resistive heating or by an electron beam. The atomic or molecular species of the material move freely toward the substrate located near the source, resulting in epitaxial deposition of the HTSg material on the substrate. The properties of the thin film are influenced by the substrate temperature, deposition rate, oxygen partial pressure, and the source-to-substrate distance. Often, films require post-annealing in an oxygen atmosphere *in situ* after deposition to acquire the proper oxygen content. Plasma- and ion-assisted coevaporation improve the oxygen content and yield superconducting films without post-annealing.

2.5.2 Sputtering Sputtering (*q.v.*) is a common thin-film deposition technique (Eom *et al.*, 1989). The dc-diode sputtering method (Fig. 12) illustrates the basic principle. When a dc voltage of 1–2 kV is applied between the cathode (the target) and the anode (the substrate) in a low-pressure gas environment, a discharge is generated. The positive ions in the discharge accelerate toward the target and impinge on it at high

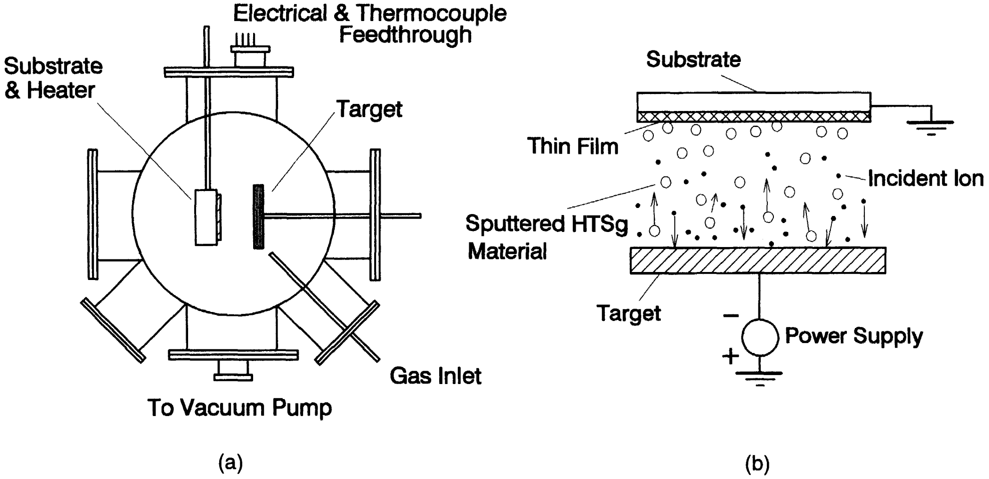
Electrical & Thermocouple Feedthrough
Substrate & Heater
Target
Gas Inlet
To Vacuum Pump
(a)
Substrate
Thin Film
Incident Ion
Sputtered HTSg Material
Target
−
+
Power Supply
(b)

FIG. 12. The dc-diode sputtering method for HTS thin films: (a) the sputtering chamber configuration and (b) the sputtering diode with the sputtered sample molecules drifting toward the substrate.

speed. They eject molecules from the surface of the target. These ejected molecules then deposit on a heated substrate to form a thin film. Single or multiple targets containing the appropriate components have been used to synthesize HTSg thin films. The target can be a compound or mixed powder, and the sputtering atmosphere can be argon, oxygen, or argon/oxygen mixed gas at different pressures. For an insulating target, e.g., one made of prereacted metal oxide powders, the dc sputtering technique fails, and so an rf source is used.

2.5.3 Laser Ablation Laser ablation (Inam *et al.*, 1988) is the simplest and most versatile method for preparing high-quality HTSg thin films, which are epitaxial and sometimes even single crystalline. A high-power laser system, e.g., a 50-W ArF excimer pulsed laser (~50 Hz) at 193 nm wavelength, is used. The laser beam strikes the HTSr target, generating a plume that consists of an active plasma and radical atoms. The particles ablated from the target are deposited at the heated substrate (Fig. 13). The kinetic energy (1–40 eV) and the internal excitation energy (a few electron volts) of these particles facilitate their rearrangement after they reach the substrate, resulting in a good growth of the HTSg thin film. The process requires no bias voltage and is operable in any atmosphere at any pressure. The film stoichiometry is the same as the target composition. The superconducting films are easily made *in situ* without post-annealing.

2.5.4 Molecular-Beam Epitaxy The layered structure of HTSrs enables one to fabricate HTSg films by condensing layer by layer of the metal-element vapors in a low-pressure oxidizing atmosphere. The complete deposition of each layer is evidenced by the appearance of an intensity peak in the reflection high-energy electron-diffraction pattern (Kawai *et al.*, 1988). The starting metal-element vapors can be generated by laser ablation or by heating the sources in well-controlled cells. The method has been

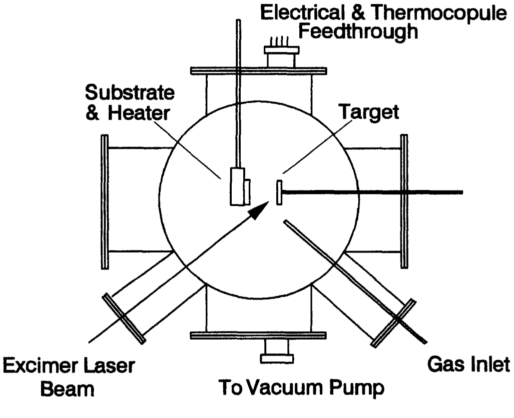
Electrical & Thermocopule Feedthrough
Substrate & Heater
Target
Excimer Laser Beam
To Vacuum Pump
Gas Inlet

FIG. 13. Chamber for HTS thin-film synthesis by laser ablation.

demonstrated to be very effective in producing metastable compounds, such as the $n \geq 4$ members of the homologous series of BSCCO, which are difficult to form using the other methods described above.

3. EXPERIMENTAL PROPERTIES

Since 1987, nearly all experimental tools available for the characterization of solids have been applied directly or with modifications to the studies of HTSrs. With the rapid progress in controlling the sample quality, properties of well-characterized HTSrs are now available. In this section, the transport and magnetic properties, which have a direct bearing on both the understanding and applications of HTSy, are discussed.

3.1 Transport Properties

3.1.1 Normal-State Resistivity The normal-state resistivity (ρ) of the cuprate HTSrs is highly anisotropic, reflecting the layered crystal structure. The resistivity in the CuO_2 planes ($\rho_{\parallel}$) at room temperature is $\sim 3 \times 10^{-4}\ \Omega$ cm, but the resistivity perpendicular to the planes ($\rho_{\perp}$) is much higher. The anisotropy ratio of ρ ($\rho_{\perp}/\rho_{\parallel}$) of Y123 is nearly 100. In Bi-2212, an anisotropy ratio on the order of 10^5 was observed (Martin *et al.*, 1988).

For compounds optimally doped for highest T_c, $\rho_{\parallel}$ varies linearly with the temperature (T) (Iye, 1992) [Fig. 14(a)]. The linear T behavior can extend over a wide range, from T_c up to the decomposition temperature of the material. In the case of Bi-2201, the range represents two orders of magnitude in temperature, from 6 to 600 K. In $La_{1.85}Sr_{0.15}CuO_4$ the linear T behavior of ρ in the CuO_2 plane continues to temperatures above 1000 K. Unlike normal metals, the $\rho_{\parallel}$ of these optimally doped compounds does not saturate at high temperatures nor flatten out for temperatures below the Debye temperature. The T dependence of $\rho_{\perp}$ remains controversial: both hopping and metallic behaviors have been observed [Fig. 14(b)]. The hopping behavior may result from oxygen deficiency. For compounds not optimally doped and for the electron-doped $Nd_{2-x}Ce_xCuO_{4-\delta}$ compounds, a T^2 term enters $\rho_{\parallel}$ prominently at low temperatures, before hopping or localization sets in.

3.1.2 Critical Current Density The critical current density (J_c) is ultimately limited by the depairing current density (J_d) (see Sec. 1.2). At 0 K, $J_d \sim 3 \times 10^8$ A/cm^2, similar to that for LTSrs. Above 0 K, thermal excitations come into play and suppress J_c to $\sim 10^7$ A/cm^2 at 77 K. This is the actual limit only if fluxoids are completely pinned.

Under the influence of a Lorentz force, the fluxoids inside a superconductor will move, resulting in a loss when the Lorentz force is greater than the pinning force $\propto \xi^2$. Defects of sizes $\leq \xi$ introduced by irradiation or by processing steps are required to act as pinning centers to enhance J_c. Some of the best representative results are summarized in

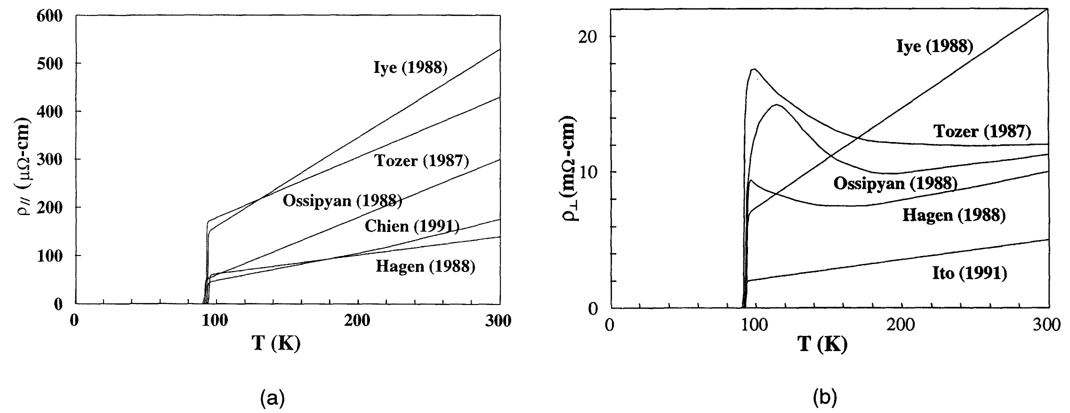

(a) (b)

FIG. 14. The temperature dependence of the normal-state resistivity of Y123: (a) $\rho_{\parallel}$ and (b) $\rho_{\perp}$.

Fig. 15, where Nb_3Sn, the best LTSr, is included for comparison. HTSrs clearly outperform LTSrs at high field. The J_c in thin films has approached the theoretical value set by the depairing limit.

3.1.3 Surface Impedance The surface impedance (R_s) below T_c measured at microwave frequencies (ν's) determines the energy loss of a HTSr and is a measure of its suitability for microwave devices. The penetration depth λ, the weak links present, and the magnetic flux trapped within a distance λ from the surface are the important parameters in determining the magnitude of R_s. For a BCS superconductor at low temperatures, the dependence of R_s on T and ν is $(\nu^2/T) \times \exp(-\Delta_0/kT)$, where Δ_0 is the energy gap at 0 K. For HTSrs, a deviation is expected because of the linear-T behavior of the normal-state resistivity ρ, the T-dependent motion of trapped flux, and the T dependence of the ever-present weak-link effect. Representative data points for R_s at 77 K and different ν's are displayed for Y123 in Fig. 16, which also shows the $R_s(\nu)$ dependence of superconducting Nb at 7.7 K and Cu at 77 K for comparison.

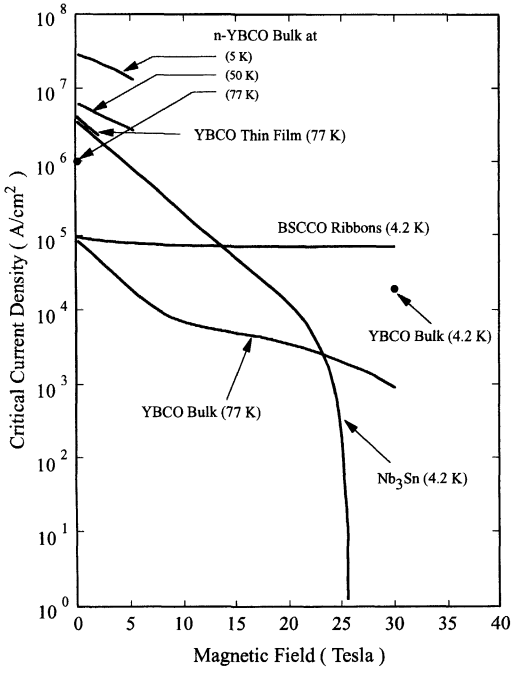

FIG. 15. The dependence of the critical current density on magnetic field for the best LTSr, Nb_3Sn, and for several HTSrs (where *n* stands for neutron irradiated); note the superior performance of HTSrs at high field compared with the LTSrs.

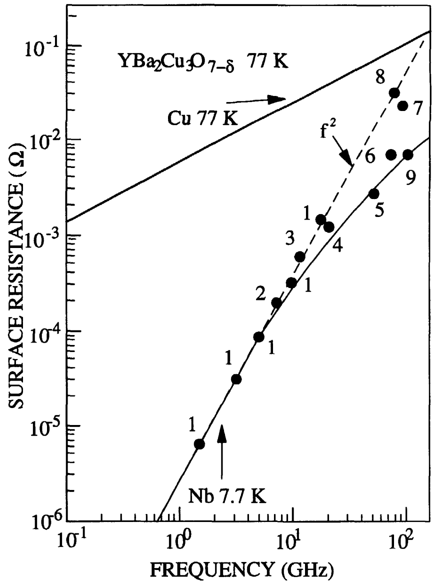

FIG. 16. The frequency dependence of the surface impedance R_s of Y123 at 77 K (the solid lines represent the behavior of Cu at 77 K) and the LTSr Nb at 7.7 K, respectively. (1) Oates *et al.* (1991); (2) Valenzuela *et al.* (1991); (3) Newman *et al.* (1991); (4) Klein (1991); (5) Asano *et al.* (1991); (6) Davis *et al.* (1991); (7) Klein *et al.*, (1990); (8) Rensch *et al.* (1991); (9) Drabek *et al.* (1990).

3.1.4 Hall Effect above T_c The normal-state Hall coefficient (R_H) of the HTSg materials (Ong, 1992) has positive sign for the hole-doped compounds and negative sign for the electron-doped Nd(Ce)-214. The R_H of the HTSg materials is, in general, strongly dependent on T, obscuring any simple interpretation as a measure of the charge-carrier density. The T dependence of R_H in many HTSg materials above T_c can be fitted by $(\alpha + T)/(\beta + T^2)$, where α and β are material-dependent parameters proportional to the density of crystal defects. In the case of pure Y123, α and β are close to zero, giving an R_H proportional to T^{-1} or a Hall number p_H proportional to T. p_H is defined as V/eR_H, where e is the electron charge and V is the unit cell volume. As the doping is reduced, this T dependence of p_H persists until the compound is no longer superconducting (Fig. 17).

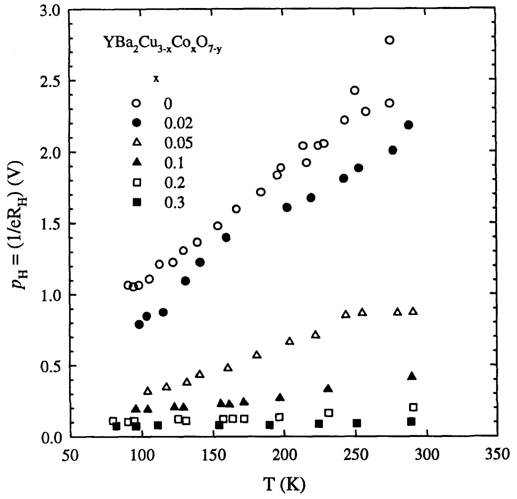

FIG. 17. The temperature dependence of the Hall number p_H for Co-doped Y123 (Clayhold *et al.*, 1989).

3.1.5 Thermal Conductivity

3.1.5 Thermal Conductivity The thermal conductivity of Y123 is almost T independent between room temperature and T_c and drops rapidly after an upturn below T_c (Uher, 1992). In single-crystal material, along the a-b plane, the thermal conductivity above T_c is κ_{ab} = 12.5 W/K m, and in the c direction it is smaller, κ_c = 0.4 W/K m; both are a few hundred to a thousand times smaller than that of Cu. Comparison with the Wiedemann–Franz law indicates that approximately 50% of the in-plane thermal conductivity is due to the charge carriers. In contrast, nearly all the c-axis heat conduction is due to phonons. The thermal conductivity of ceramic material depends upon its processing history (Fig. 18) but is often lower by a factor of 2–5 than the single-crystal values of κ_{ab}.

3.1.6 Thermoelectric Power Above T_c, the thermoelectric power (TEP) decreases almost linearly with increasing T for most hole-doped HTSrs and does not show any sign of phonon drag [Fig. 19(a)]. The underlying mechanism for such a T dependence is not yet known. However, a systematic trend of TEP variation with doping and a universal correspondence of the room-temperature TEP with the carrier per CuO_2 layer (p) was uncovered [Fig. 19(b)] (Obertelli *et al.*, 1992). The room-temperature TEP varies linearly with p for the overdoped HTSrs and exponentially for the underdoped ones, and thus provides a simple qualitative method for the characterization of HTSg materials. A quadratic T_c-p relation was also obtained to describe many of the HTSg systems [Fig. 19(c)] (Presland *et al.*, 1991).

3.2 Magnetic Properties

Extensive efforts have been directed toward the investigation of the flux dynamics of HTSrs. Although complete descriptions of

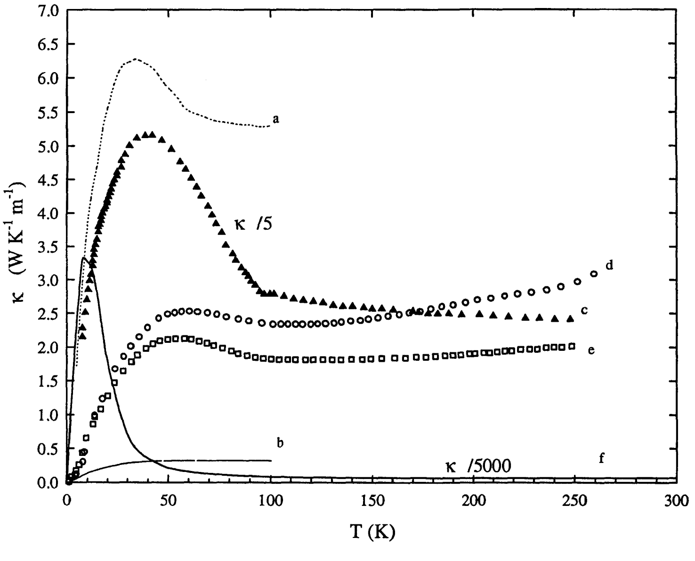

FIG. 18. The temperature dependence of the thermal conductivity of Y123: curve *a*, single crystal in the *a–b* plane; curve *b*, single crystal in the *c* direction; curve *c*, high-quality single crystal in the *a–b* plane; curve *d*, hot-pressed polycrystalline sample in the direction perpendicular to the direction of hot pressing; curve *e*, polycrystalline sample in the direction parallel to the direction of hot pressing. Also shown in curve *f* for comparison is the thermal conductivity of Cu, which is a few thousand times greater than that of Y123 (Uher, 1992).

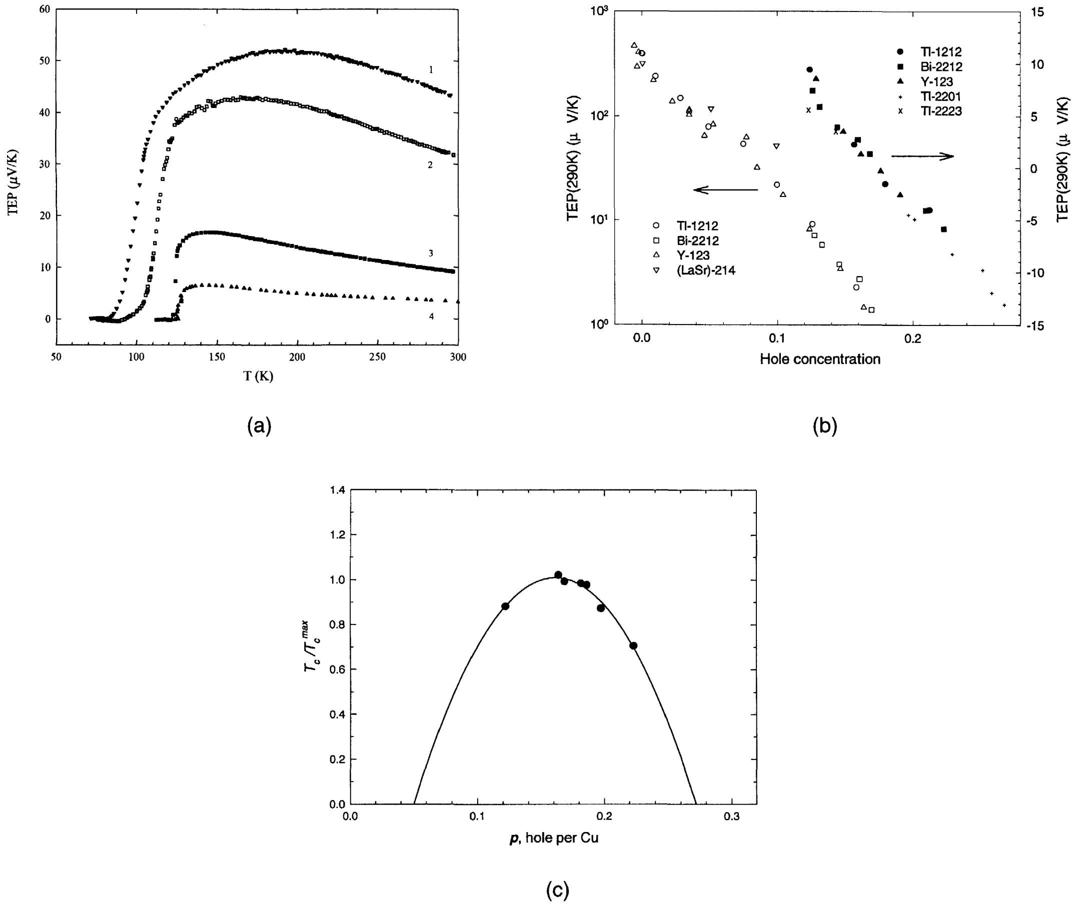

(a)

(b)

(c)

FIG. 19. (a) The temperature dependence of the thermoelectric power TEP of Hg-1212 in the numerical order of increasing doping (Chen *et al.*, 1994); (b) the universal hole dependence of the room-temperature thermoelectric power for various HTSrs (Obertelli *et al.*, 1992); and (c) T_c/T_c^{max} versus p where T_c^{max} is the T_c of the optimally doped sample of the HTSg system (Presland *et al.*, 1991).

the magnetic properties of HTSrs in the superconducting state have yet to be developed, insight has been gained into the movement of fluxoids (Huse *et al.*, 1992).

For $H_{c1} < H < H_{c2}$, a type-II superconductor is said to be in the mixed state (Fig. 9), where ρ remains asymptotically small as H enters the interior of the superconductor discretely in the form of fluxoids or flux quanta. Each flux quantum has a supercurrent circulating over a range λ around its normal core (of size $\sim\xi$). In a defect-free type-II superconductor, the fluxoids form a rigid flux lattice, which will move in the presence of an externally applied current in response to the Lorentz force (Fig. 20). The movement of these fluxoids through the sample produces an electric field opposite to the current flow and thus leads to energy dissipation.

The introduction of defects of a dimension $\sim\xi$ to a superconductor will destroy the periodicity of the flux lattice. The fluxoids may cluster and be pinned around the defects. These flux bundles may or may not move, depending on the balance between the flux-pinning force (associated with defects and interfluxoid interaction) and the depinning force (due to the Lorentz force and thermal activation).

The mixed state was generally considered a featureless phase region until the discovery of HTSrs. HTSrs are type-II superconductors. Their high T_c, short ξ, low carrier con-

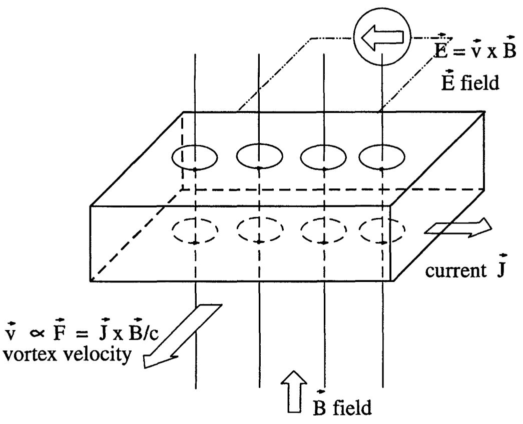

$\vec{E} = \vec{v} \times \vec{B}$
$\vec{E}$ field
current $\vec{J}$
$\vec{v} \propto \vec{F} = \vec{J} \times \vec{B}/c$
vortex velocity
$\vec{B}$ field

FIG. 20. Fluxoid movement (velocity v) due to the Lorentz force between the current (density J) and an applied magnetic field B. The resulting electric field E is opposite to J and causes loss.

centration (n), and large λ greatly enlarge the area under the critical-field line in the H-T phase diagram. The larger range of the parameter space enhances the likelihood of detecting any possible differences within the mixed state arising from the different detailed balances among the various forces acting on and between fluxoids. Indeed, an additional boundary (Fig. 21) has been observed (Brandt, 1991), i.e., the irreversibility line $H_i(T)$ [or $T_i(H)$] below $H_{c2}(T)$. Above $H_i(T)$, the response of a HTSr to an H change is reversible, i.e., the vortices behave as a fluid; below it, the response is irreversible. $H_i(T)$ can be treated operationally as a boundary below which the dissipation drops so rapidly that an upper bound of J_c can be defined. The nature of $H_i(T)$ is rather complex and depends on a combination of characteristics such as interlayer coupling, flux-pinning strength, and the defect morphology. As a result, a convergent interpretation of $H_i(T)$ is still lacking. Interpretations range from a first-order melting through a second-order vortex-glass transition to a Bose-glass transition. Additional features in the H-T phase diagram have also been reported, and their origins are unknown.

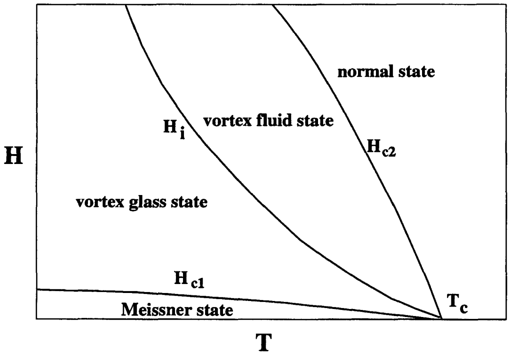

FIG. 21. The schematic H-T phase diagram for a HTSr: $B = 0$ and $\rho = 0$ in the Meissner state; $0 < B < H$ and $\rho = 0$ in the vortex glass state; $0 < B < H$ and $\rho \neq 0$ in the vortex fluid state; and $B = H$ and $\rho \neq 0$ in the normal state.

3.3 Symmetry of the Superconducting Order Parameter

The symmetry of the superconducting order parameter refers to the angular dependence of the superconducting energy gap ($\Delta_{\mathbf{k}}$) in the reciprocal lattice (**k**) space, which is uniquely related to the crystal structure. Since superconductivity in both LTSrs and HTSrs involves the pairing of electrons (or holes) of opposite spin, the angular dependence of $\Delta_{\mathbf{k}}$ is determined by the angular momentum (l) state of the pairing electrons (or holes). For $l = 0$ (the s wave), $\Delta_{\mathbf{k}}$ never vanishes in **k** space and is known as nodeless. It has the same symmetry as the crystal and therefore can be isotropic (same in all directions) or anisotropic (different values in different directions) in the **k** space. However, for $l = 2$ (d-wave pairing), $\Delta_{\mathbf{k}}$ vanishes at points or lines in certain directions of **k** and so is said to have nodes or node lines and needs not display the same symmetry as the crystal. The existence of nodes or node lines in $\Delta_{\mathbf{k}}$ enables a very small amount of energy to excite the electrons of a superconductor in its superconducting state. As a result, several distinct signatures in both the microscopic properties and macroscopic behavior are predicted for a d-wave superconductor (Annett *et al.*, 1989). Knowledge of the symmetry of $\Delta_{\mathbf{k}}$ has been thought by some to hold keys to the HTSr mechanism mystery.

No nodes or node lines have been detected in the conventional LTSrs, and so their $\Delta_{\mathbf{k}}$ has the s symmetry. However, d pairing is evidently present in a small but interesting class of LTSrs known as heavy-fermion superconductors such as UPt_3 and UBe_{13} (see HEAVY-FERMION PHENOMENA).

The situation in HTSrs remains controversial (Levi, 1993; Schrieffer, 1994) in spite of the many sophisticated experiments done on Y123 and Bi-1212. For instance, a linear T dependence of λ suggesting a d pairing in Y123 has been reported by some experiments in contrast to a quadratic T dependence, consistent with an s pairing, by others. Also considered are the results of tunneling experiments. Josephson pair tunneling between Y123 and Pb (a conventional LTSr) observed along the c axis of Y123 is consistent with an s pairing (Sun *et al.*, 1994). In contrast, in a so-called corner dc SQUID, which consists of a pair of Y123/Au/Pb tunnel junctions along the a and b axes of a Y123 single crystal, a critical-current dip in zero external field was observed, which is compatible with a d pairing (Wohlman *et al.*, 1993). Very recently, an elegant experiment was carried out (Tsuei *et al.*, 1994) by measuring the spontaneous magnetization of tricrystal superconducting rings of Y123 with various orientations of the substrate crystals. With proper orientations between the crystals, which determine the relative orientation of the order parameters with respect to the junction interface, a spontaneous magnetization of a half magnetic flux quantum has been unambiguously discovered, in agreement with a d-wave pairing symmetry. Angular resolved photoemission results (Shen *et al.*, 1993) showed the existence of lines of nodes in $\Delta_{\mathbf{k}}$ along the direction predicted by the theory. However, another set of results obtained by a different group (Gofron *et al.*, 1994) suggested an anisotropic s wave.

The above inconsistencies have been ascribed to possible different sample qualities, different junction conditions, a combination of s and d wave, or a highly anisotropic s wave. Adding to the uncertainties is the recent suggestion (Abrikosov, 1995) that some of the above tunneling results indicative of d-wave pairing may be associated with the orthorhombic structure symmetry of Y123. A tunneling experiment similar to that of Tsuei *et al.* (1994) using the tetragonal HTSr will resolve this uncertainty.

4. THEORETICAL STATUS

The Ginzburg–Landau theory satisfactorily describes the general superconducting phenomena of HTSrs on a macroscopic level without explaining the occurrence of HTSy or the mechanism responsible for it. On the microscopic level, many models have been proposed to explain HTSy. Some fall into the framework of the BCS theory while others do not. They may be combined into three groups, based on the origin of the interaction. Most of them are built on the assumption that HTSy occurs in a two-dimensional electron system, namely, the CuO_2 layers, although the exact dimensionality of the HTSy observed is still not known.

4.1 Magnetic Models

Immediately after the discovery of HTSy, Anderson (1987) pointed out that the copper-oxide layers are the key building blocks in these compounds. He proposed the large-positive-U Hubbard model to describe the HTSrs. In this model, the electrons move in a square lattice with a repulsive interaction (U) between them at each of the lattice sites. Subsequent discoveries of the long-range antiferromagnetic (AF) order and other magnetic properties in the insulating La_2CuO_4 turn out to be consistent with the low-temperature behavior of the Hubbard model in the strong-coupling limit. The most fascinating but difficult problem is associated with the doping, which results in a transition at low temperature from the insulating AF state to the superconducting state. As discussed earlier, the normal state of these HTSrs also exhibits some remarkable nonconventional metallic behaviors. Even within the general context of the Hubbard model, different theories exist.

In the strong-coupling regime (large U), the resonant valence-bond (RVB) state (Baskaran *et al.*, 1987) and its various generalizations have later been considered as possible candidates for the solution of the Hubbard model. In the RVB state, the electrons are considered to form some exotic short-range AF ordering. Although the prediction of charge (holon) and spin (spinon) separation in the RVB theory appears consistent with some experiments (Chien *et al.*, 1991b), the RVB theory has not been fully developed because of strong interactions existing between spinons and holons. HTSy arises from interlayer coupling in this model and is therefore more three-dimensional.

In an alternative approach, Montoux and

Pines (1992) obtained expressions for the electron band structure and the spin fluctuation spectrum using photoemission and nuclear magnetic resonance data. Using these expressions, they solved the basic equations for a superconducting system, known as the Eliashberg equations, and obtained an estimate of T_c in agreement with experiment and a "d-wave" pairing symmetry. This approach can be considered less "exotic" than RVB, since the charge–spin separation is not invoked.

4.2 Electron–Phonon Interaction Models

The conventional electron–phonon coupling theory for superconductivity (Bardeen *et al.*, 1957), which has been investigated extensively for the past 30 years, is mathematically rigorous and physically sound, if Landau's Fermi-liquid theory for the charge carriers is valid and the so-called Migdal approximation holds. Landau's theory asserts that even in the presence of strong interactions, such as the Coulomb interaction, the charge carriers behave similarly to a gas of noninteracting fermions, which follow a Fermi–Dirac distribution. The Migdal approximation is a simplification made on the calculations, possible only when the phonon energy is much smaller than the electron energy. Electron–phonon coupling has not been regarded as the main mechanism for causing HTSy in cuprates, since band-structure calculations do not reproduce the AF insulating phase together with other magnetic properties in La_2CuO_4, the unusually high T_c of HTSrs, and the extremely small isotope effect in Y123.

Recently, however, evidence for a nonnegligible role of the electron–phonon interaction has emerged. This evidence includes the proposition of anharmonicity that accounts for the small isotope effect observed in certain HTSg compounds (Crespi *et al.*, 1991), and the explanation of the linearly T-dependent resistivity in terms of electron–phonon interaction, although its contribution to T_c does not seem to be dominant (Xing *et al.*, 1988). A band-theory-like Fermi surface with van Hove singularities has also been suggested (Campuzano *et al.*, 1990) by photoemission and positron-annihilation experiments. Models based on the electron–phonon interactions have been proposed (Newns *et al.*, 1992) to account for the high T_c and some anomalous properties of HTSrs by taking advantage of the large density of states associated with the van Hove singularities near the Fermi surface.

4.3 Marginal Fermi-Liquid Model

Although it is possible to argue that the linearly T-dependent resistivity is due to electron–phonon interactions, it is difficult to explain the linearly frequency-dependent Drude scattering rate in terms of electron–phonon scattering. The linearity of the scattering rate over a large frequency range ($h\nu \sim$30–300 meV) suggests that the quasiparticles are being scattered by excitations whose spectral density covers a similarly wide range in energy. This is a characteristic of the marginal Fermi-liquid (MFL) spectrum proposed phenomenologically (Varma *et al.*, 1989) to explain the linear scattering rate, as well as other observations. Within the MFL framework, one obtains a scattering rate $1/\tau \propto \max(h\nu, k_BT)$, where h is the Planck constant and k_B the Boltzmann constant, assuming that the quasiparticles interact with fluctuations that have a flat bosonic spectrum. By definition, the MFL spectrum is a phenomenological form for the dressed electron–hole polarizability. This form for the polarizability appears to be directly observable in the Raman scattering as an unusual continuum in the Raman intensity extending out to energies comparable to an interband transition (Sugai, 1989). This interpretation is controversial, as the observed continuum is surprisingly insensitive to doping and has even been detected in the insulating phase (Reznik *et al.*, 1993). The MFL theory is a phenomenological theory and, as yet, has no microscopic basis of support.

5. APPLICATIONS

A superconductor can carry a large current without loss at low frequencies and with a low loss at high frequencies, trap and shield a strong dc magnetic field, shield electromagnetic signals, display a drastic resistivity change near T_c, and exhibit novel behavior when made into tunneling junctions. Many application concepts have exploited one or more of these characteristics. Some examples are given below.

5.1 Conductors

Many large-current applications require long lengths of HTSg wire capable of carrying large electric current with negligible or minimal loss, mechanically strong and flexible enough to withstand the manufacturing process, and cryogenically stable enough to avoid thermal runaway. Groups in Japan and the United States have fabricated long powder-in-tube (PIT) wires from the BSCCO 2223 phase with high critical current densities at 77 K in their own magnetic field. Most recently, American Superconductor Corporation has reported a 1.16-km length with a J_c of 13 800 A/cm^2 (Riley, 1994).

J_c of BSCCO at 77 K decreases rapidly in the presence of an external magnetic field, but, at temperatures up to 30 K, J_c remains relatively high in strong magnetic fields. As the operating temperature is reduced, the J_c increases and becomes less field dependent. At 4.2 K, a 3-cm-long Bi-2223 wire was reported to have a $J_c \sim 2.5 \times 10^5$ A/cm^2 in 0 T, which decreases only to $\sim 1 \times 10^5$ A/cm^2 at 25 T (Haldar and Motowidlo, 1992). As the wire length increases, the achieved value of J_c decreases.

Many groups in the United States and Japan have produced BSCCO wires and tapes that are flexible enough to be wound into coils. A five-horsepower synchronous motor was recently demonstrated by Reliance Electric (Fig. 22) using the American Superconductor Corporation HTSg coils. A magnet made by Intermagnetics General Corporation, which consists of 10 pancake coils and contains a total of 480 m of HTSg tape, has generated the highest fields: 2.6 T at 4.2 K, 1.8 T at 27 K, and 0.36 T at 77 K. This magnet and a horseshoe HTSg magnet made by American Superconductor Corporation are shown in Fig. 23. Many researchers are confident that HTSg magnets will eventually generate fields in excess of 20 T at temperatures of 20 K or higher. The first application will likely be in hybrid magnets that generate strong fields by combining LTSg coils with HTSg coils.

Since the ability of BSCCO coils to generate strong fields seems intrinsically restricted to temperatures below 30 K, some groups are pursuing various approaches to making conductors from the Tl-1223 or Y123 phases and even from Hg1223, all of which have higher irreversibility lines than BSCCO (Fig. 24). In principle, Tl-1223 and Y123 and Hg1223 coils could generate high fields at 77 K, if suitable processing techniques could be developed. Short Tl-1223 wires made by the PIT technique have J_c near 10 000 A/cm^2 at 77 K (Seido *et al.*, 1992), and Tl-1223 ribbons made by electrodeposition possess J_c up to 67 000 A/cm^2 at 77 K (Seido *et al.*, 1992). An impressive $J_c \sim 2 \times 10^6$ A/cm^2 at 75 K has recently been achieved by using the laser ablation technique to deposit Y123 thick films on a Ni substrate with a textured buffer layer (Wu *et al.*, 1994). Effort on Hg-1223 wires has just been started. The chemical instability of Hg-1223 is a challenge.

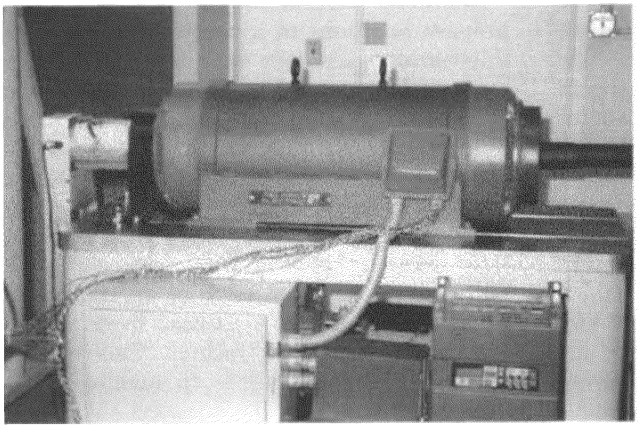

FIG. 22. A five-horsepower HTSg synchronous motor (Reliance Electric/American Superconducting Corporation/Electric Power Research Institute).

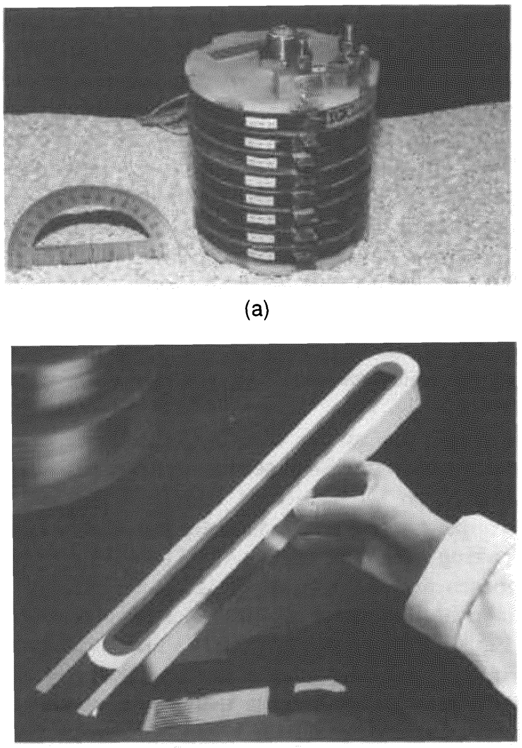
(a)

(b)

FIG. 23. (a) Intermagnetic General Corporation pancake magnet producing a record field of 2.6 T at 4.2 K and 0.36 T at 77 K (scale in inches) and (b) American Superconducting Corporation racetrack magnet.

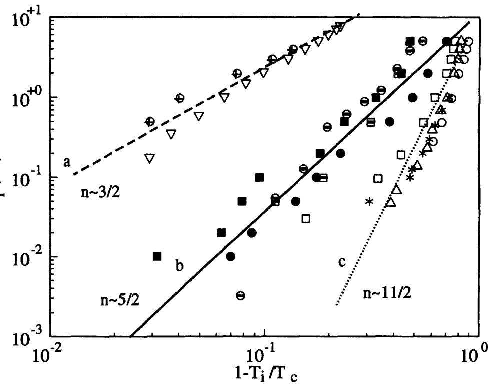

FIG. 24. A plot of the thermomagnetic irreversibility line $H_i(T)$, which is proportional to $[1-(T_i/T_c)]^n$ for HTSrs: (a) Y123, (b) HBCCO, and (c) BSCCO and TBCCO (Huang *et al.*, 1994). A smaller n represents a stronger flux-pinning force.

5.2 Josephson Junctions and SQUIDs

The Josephson junction is the key element of superconducting electronics, based on observations that led to the awarding of the 1973 Nobel prize in physics to its discoverer, Brian Josephson. It consists of two weakly coupled superconducting regions, usually in a superconductor–insulator–superconductor (SIS) or superconductor–normal metal–superconductor (SNS) trilayered structure. Only recently have SNS junctions of cuprate HTSrs been made to display successfully Josephson tunneling with hysteretic *I-V* characteristics or two well-defined voltage states for a switching device (Fink *et al.*, 1992; Virshup *et al.*, 1992). The difficulty may be attributed to the extremely short ξ of cuprate HTSrs, comparable to the unit cell dimension. The short ξ, especially perpendicular to the CuO_2 layers, poses stringent conditions on the fabrication of the superconducting layers. Planar junctions with weak links may alleviate the problem partially by tunneling electrons into the CuO_2 layer. Other reasons associated with possible anomalous electronic properties of the HTSg cuprates cannot be ruled out.

A variety of alternative Josephson devices made of Y123 films and using a diverse set of weak links have been fabricated and have yielded reasonable results. The first devices were based on weak links that occur naturally between grain boundaries in poor-quality films. The random appearance of grain boundaries throughout the films increases noise and makes device fabrication difficult to control.

Methods were subsequently developed to prepare junctions in a controllable way. SIS microbridge junctions were fabricated (Kingston *et al.*, 1990) using single-layer geometry where noble metals were employed to fill the narrow gap between two HTSg strips. Because of the proximity effect, the ξ near the gap increases to facilitate tunneling. Edge junctions (Gross *et al.*, 1991) were successfully created between one overlying HTSg film and the exposed edge of an underlying HTSg film, which is separated from the top one by an insulating barrier. This method overcomes the difficulties in making multilayer junctions.

More refined junction fabrication techniques were developed to define the location and sometimes even the angle of the grain boundary. One approach is to form a bicrystal by fusing two substrates with different orientations (Dimos *et al.*, 1990). An epitaxial HTSg film deposited onto the bicrystal substrate grows with a well-defined clean grain boundary above the interface between the two substrates. Junctions of excellent quality have been obtained. Unfortunately, the "bicrystal" method is not suitable for preparing a large array of small junctions for complex electronic circuits.

To produce an array of comparable small junctions over a large area with high yield, the "step-edge" approach (Daly *et al.*, 1991) has been used. An epitaxial HTSg film is deposited over a sharply etched step in a substrate. Two clean grain boundaries in the film are created.

Yet another well-controlled "biepitaxial" method with high yield has been recently developed (Char *et al.*, 1991). In this method, one takes advantage of the ability of a thin buffer layer to rotate the in-plane orientation of a HTSg film. For instance, a thin MgO underlying layer changes the orientation of the Y123 film by 45° relative to that of the film without such a buffer layer. Therefore, junctions at the 45°-tilt grain boundary can be formed at any chosen locations at the edges of the MgO seed patches. Unfortunately, 45° is not the optimal tilt angle for the lowest junction noise.

The superconducting quantum interference device (SQUID) (Clarke, 1989) is the most sensitive device available for measuring a variety of physical quantities such as magnetic field, magnetic-field gradient, magnetization, magnetic susceptibility, and many other quantities coupled to the above (see SQUIDs). A dc SQUID consists of two Josephson junctions that are connected in parallel in a superconducting loop. A change in the flux within the loop substantially smaller than a flux quantum Φ_0 with a resolution of $\sim 10^{-6}\ \Phi_0\ \mathrm{Hz}^{-1/2}$ has been detected by a Nb SQUID operated at 4.2 K and at frequencies above 1 Hz. A flux transformer is frequently used in conjunction with the SQUID in order to enhance its field sensitivity, to determine the field gradient, or to couple to other physical quantities to be determined. An rf SQUID consists of one Josephson junction in a superconducting loop that is coupled to a resonant circuit. It is significantly less sensitive than its dc counterpart, although easier to operate.

The LTSg SQUID has already had a significant impact on experimental physics, geophysical surveying, nondestructive testing, precision metrology, and biomagnetism study. SQUIDs that can operate at 77 K will undoubtedly make the applications more common and practical. Because a simpler cryostat can be used with the HTSg SQUID, an increase in sensitivity can also be achieved by reducing the distance between the SQUID and the object generating the magnetic field, e.g., the brain. A HTSg SQUID operated at 77 K has a substantially higher low-frequency noise level than a LTSg SQUID at 4.2 K. However, steady progress has been made in reducing the noise level, as shown in Fig. 25. Practical HTSg SQUIDs can be produced with high enough sensitivities for a wide range of applications. High-performance dc SQUID magnetometers (Koelle *et al.*, 1993) with single-layer Y123 flux transformers have been made with a very low noise level below 10^{-30} J/Hz at 77 K (Koelle, 1994), which is close to that

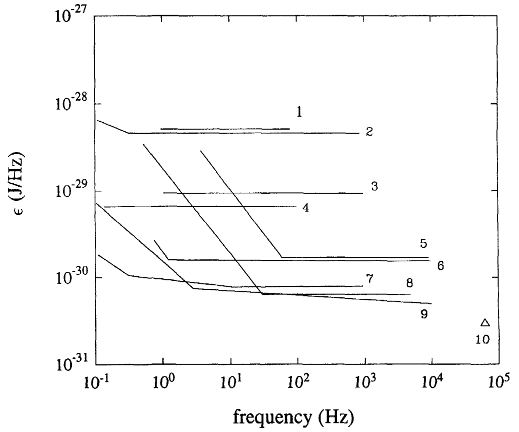

FIG. 25. Energy resolution ϵ for HTSg SQUIDs with different rf and dc SQUID inductances: (1) rf (150 MHz, 190 pH) (Zhang *et al.*, 1993), (2) rf (4.2 K) (Dilorio *et al.*, 1993), (3) (12 pH) (Dilorio *et al.*, 1993), (4) rf (3 GHz) (Mueck, 1993), (5) dc (80 pH) (Sun *et al.*, 1993), (6) dc (20 pH) (Koelle *et al.*, 1993), (7) dc (4.2 K) (Dilorio *et al.*, 1993), (8) dc (10 pH) (Friedl *et al.*, 1993), (9) dc (13 pH) (Koelle *et al.*, 1993), and (10) dc (very small inductance) (Kawasaki *et al.*, 1991).

achieved in commercial LTSg dc SQUIDs operated at 4.2 K.

HTSg multiturn coils have been fabricated by patterning two epitaxial HTSg thin films separated electrically by an insulating layer and by connecting the HTSg coils through openings in the insulating layer. HTSg SQUIDs with biepitaxial junctions were also made on the same substrate in addition to the coils. A SQUID magnetometer was thus fabricated by using a flux transformer to couple magnetic flux into the SQUID (Plate 2). A major step forward was made recently in processing epitaxial HTSr/insulator multilayer films and devices (Miklick *et al.,* 1991). The growth of six epitaxial layers and various steps of patterning, representing most of the technology needed in HTSg electronics, have been achieved by Conductus.

A simple version, "Mr. SQUID" by Conductus, is already commercially available (Fig. 26). Such a device cooled in a hand-held Styrofoam coffee cup has detected MKG signals. A HTSg magnetometer system using a special design has recently been developed to obtain clear 16-channel human heart signals by the Superconducting Sensor Laboratory in Itami, Japan.

5.3 Passive Microwave Devices

Some of the more successful small-current applications of HTSrs have been in the area of passive high-frequency devices. These devices depend on the low surface resistance R_s of HTSg materials to propagate high-frequency signals with lower dissipation and distortion than normal metals. These microwave devices (Newman and Lyons, 1993) include filters, resonators, delay lines, antennas, transmission lines for integrated-circuit chip carriers, etc. The advantages of HTSg materials are primarily in improved performance, which, particularly in the case of antennas and filters, permits the use of compact designs that would not be practical with normal metals.

The R_s of state-of-the-art HTSg thin films at 77 K has been shown in Fig. 16 as a function of frequency. The relative advantages of HTSrs operating at 77 K are not very dramatic in the millimeter-wave region near 100 GHz, but HTSg devices operating at lower frequencies can yield considerable gains. Thus, simple and compact stripline and microstripline thin-film devices can replace waveguide components for narrow-band applications such as delay lines and filters (Fig. 27). As a typical example, a four-pole microstrip band-pass filter at 4 GHz has an insertion loss of only 7% fabricated with HTSrs compared to about 50% when fabricated with Au (the Au filter is unusable in practice). A TBCCO compact dielectric resonator has shown (Shen *et al.,* 1992) quality factors of 2×10^6 with a circulating power of 1 kW at 5.5 GHz and 90 K, 100 times better than a similar resonator constructed using a Au film.

Superdirective antenna arrays are another application where low-loss HTSg materials may be important, especially when the elements are electrically small. Miniaturization is a major concern in antenna applications. Electrically small antennas, less than $\frac{1}{2}$ wavelength long, are attractive below 3 GHz. The

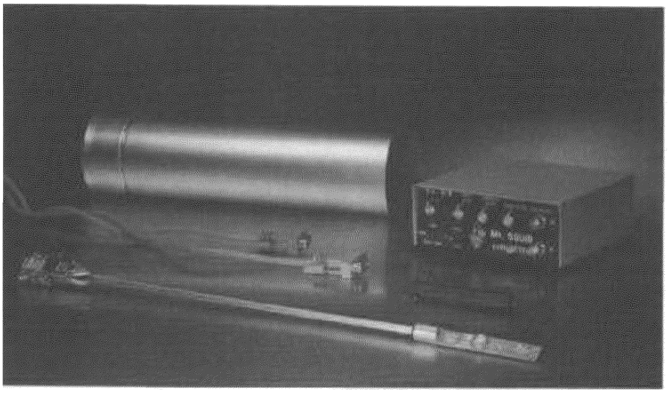

FIG. 26. Mr. SQUID (Conductus). The SQUID sensor is attached to the right-hand side of the holder at the bottom of the picture.

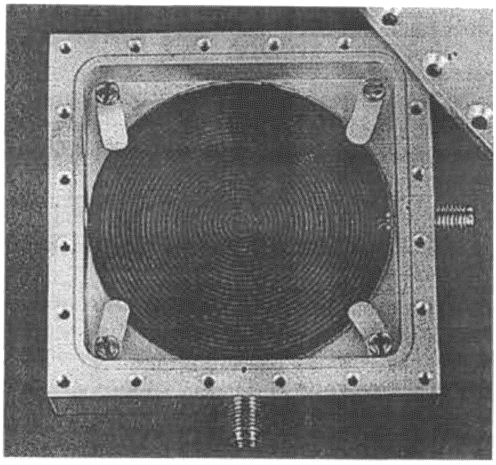

(a)

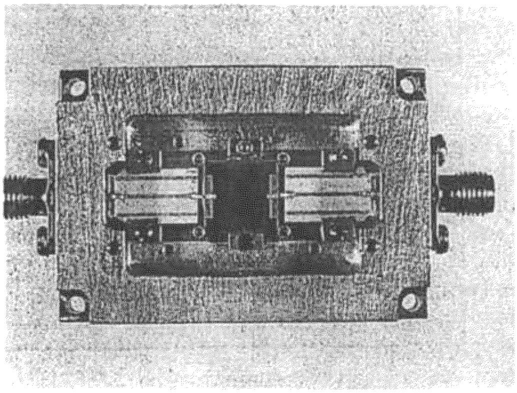

(b)

FIG. 27. (a) A HTSg delay line for a few nanoseconds on a 5-cm-diam. wafer, and (b) a microstrip resonator for the microwave electron system (Lincoln Laboratories).

price paid in reduced efficiency is largely due to resistive losses in the impedance-matching network between the transmitter and antenna. A Cu matching network for a $\frac{1}{10}$-wavelength dipole antenna would result in a radiation efficiency of the system of 2%, whereas a value of 37% has been achieved with HTSrs. Further increases in efficiency are expected.

5.4 Trapped-Field Magnets

HTSrs are type-II superconductors capable of trapping a magnetic field. This is done by withdrawing the energizing field after cooling a superconductor to below T_c in such a field or by cooling the superconductor in the absence of a field, followed by the application of the energizing field to above H_{c2}. Therefore, they can be used as permanent magnets due to the trapped magnetic field, sometimes even with the high fidelity of the field profile (Weinstein *et al.*, 1990). The advantages of a trapped-field magnet (TFM) include lighter weight, simpler geometry, lower energy consumption, and lower capital cost, in comparison with a conventional electromagnet. A TFM may become the first practical application of bulk HTSrs, bypassing complications associated with the fabrication of HTSg wires.

The trapped field (B_T) is approximately proportional to the J_c and the HTSg-grain size in the sample. Single-grain Y123 disks have been demonstrated by Nippon Steel and the Texas Center for Superconductivity at the University of Houston (TCSUH) to trap a $B_T \sim 1.6$ T at 77 K, a value many times greater than that generated by any permanent magnet. A further increase of B_T toward the limiting value of H_i at 77 K is not inconceivable by continuously increasing the dimensions of the HTSg sample to a critical value of 3 cm diameter (Meng *et al.*, 1994) and/or by enhancing the flux pinning through processing or thorough irradiation by high-energy heavy ions. A record 2.3-T B_T has recently been achieved in a heavy-ion–irradiated HTSg TFM at 77 K by TCSUH. The TFMs are being tested for applications such as magnetic clamps, magnetic rivet guns, magnetic dent pullers, homopolar generators and motors, ultrahigh-field research, and possibly table-top MRI's, which can perform three-dimensional imaging that an electron microscope cannot.

5.5 Magnetic-Mechanical Devices

The demonstration of levitation or suspension, of either a magnet by a HTSr chilled with liquid nitrogen or the inverse, has captured the imagination of experts and nonexperts alike and remains one of the most popular science exhibitions today. It embodies the most fundamental principle of a wide variety of magnetic-mechanical devices such as magnetic bearings, magnetic clutches, vibration dampers, vibration isolators, electromagnetic projectors, magnetic shields, etc.

The phenomenon of magnetic levitation or suspension by HTSrs is a manifestation of the ability of a type-II superconductor to expel or to trap, respectively, an external field originating from a conventional magnet. The prevention of such flux from entering the interior of a HTSr leads to the creation of a supercurrent in the HTSr (but not in the sense of Lenz's law), which generates a field opposite to the field of the magnet. The levitation effect is the result of the interaction between these two fields.

The magnetic flux already in the interior of a HTSr can be trapped by pinning centers and will not leave the HTSr upon the withdrawal of the external field. The interaction between this trapped field and the magnet gives rise to magnetic suspension. A few examples of the HTSg magnetic-mechanical devices are briefly described.

5.5.1 Magnetic Bearings To reduce the friction due to the contact between the surfaces of two objects in relative motion, or to eliminate the need for lubrication, which fails at low temperatures, magnetic bearings of different versions have long been proposed and used. One or more of the magnets act as the rotor and the rest as the stator. The basic concept is to make use of the repulsive force between two similar magnetic poles or the attractive force between two opposite poles. An ideal magnetic bearing should have high thrust (resistance to axial motion) and high stability (resistance to radial motion). Unfortunately, the Maxwell theory of electrodynamics shows that a pair of conventional magnets is intrinsically unstable on account of the existence of a saddle point in a constant-energy surface. In other words, a strong thrust between two magnets in a bearing system in one direction is invariably associated with a strong instability in the other direction (or directions). A complicated servo system is usually used to overcome the instability and makes the bearing bulky and expensive. Therefore, HTSrs, which are not susceptible to this type of instability because of their flux-trapping ability, have been used in conjunction with permanent magnets to construct magnetic levitating bearings with high thrusts. HTSg magnetic bearings are especially attractive for uses below 77 K, such as for liquid-hydrogen pumps, because both the levitation and trapping effects become stronger at these low temperatures. The magnitude of the thrust and the degree of stability depend on the specific arrangement of the magnets and HTSrs.

The simplest form of HTSg magnetic bearing [Fig. 28(a)] consists of a spinning magnet supported by a HTSg disk. Its thrust is determined by the limited levitation or suspension force, and its sideways stability by the field-trapping effect. When similar poles of two magnets are brought close to each other, the thrust increases, but they fly sideways as a result of their instability. On insertion of a slab of HTSr between them [Fig. 28(b)] (Chu *et al.*, 1992), the field-trapping effect overcomes such instability when proper slab thickness and pole strengths are chosen. Similarly, the instabilities can be removed by surrounding the same two opposing-pole magnets with a HTSg cylinder or HTSg blocks [Fig. 28(c)] (Chu *et al.*, 1994). Both the field-repulsive and trapping effects provide the stability to prevent the two magnets from flying away sideways, while high thrust due to the reduced separation between the magnetic poles is realized in this last version. A 19-kg prototype flywheel employing a multistage HTSg bearing (Fig. 29) with an energy loss of 1%/h when running at 6000 rpm has been constructed and tested by TCSUH. A prototype lunar telescope support was also constructed jointly by TCSUH and NASA to overcome difficulties, due to the cold and dusty lunar environment, associated with conventional ball bearings.

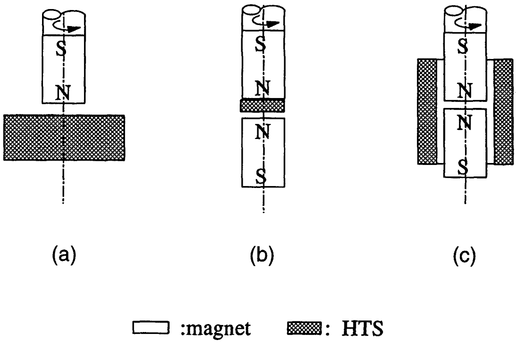

FIG. 28. Three basic types of HTS magnetic bearings: (a) Moon *et al.* (1990), (b) Chu *et al.* (1992), and (c) Chu *et al.* (1994).

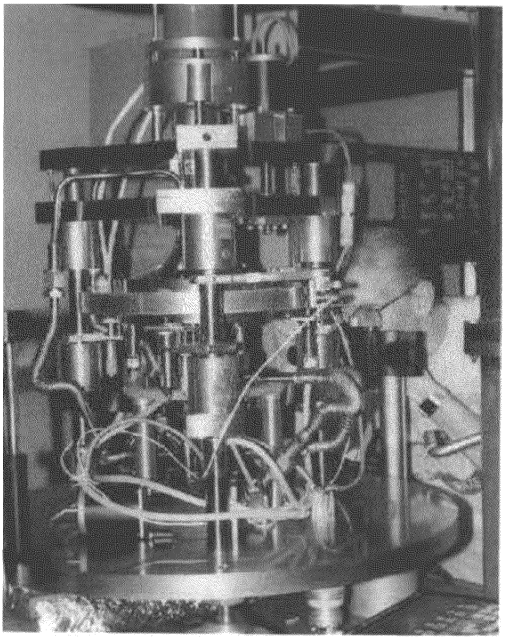

FIG. 29. A 19-kg flywheel supported by HTS magnetic bearings (Chu *et al.*, 1995).

5.5.2 Vibration Dampers and Magnetic Clutches Because of the field-trapping effect, any change of magnetic field experienced by a HTSr leads to energy dissipation. In other words, it requires a force to move a magnet across a HTSr. It was proposed (Ma *et al.*, 1992) to amplify such a force by adopting a multipole configuration. Any linear or rotational motion of the magnets with respect to the HTSrs generates rapid cyclic variations of fields across the HTSrs and therefore a large dissipative force. Clutches [Fig. 30(a)] and vibration dampers [Fig. 30(b)] have been constructed and tested, with the object to be damped or coupled attached to the multiple-magnet assembly, which is placed near but not in contact with a HTSg slab fastened to its surrounding. The various modes of coupling between the magnets and the HTSr can be tuned by varying the temperature of the HTSr or by auxiliary fields generated by simple small electromagnets. Using the vibration damper arrangement, a prototype was constructed by TCSUH to support a scanning tunneling microscope with a minimum vibration to enhance the resolution.

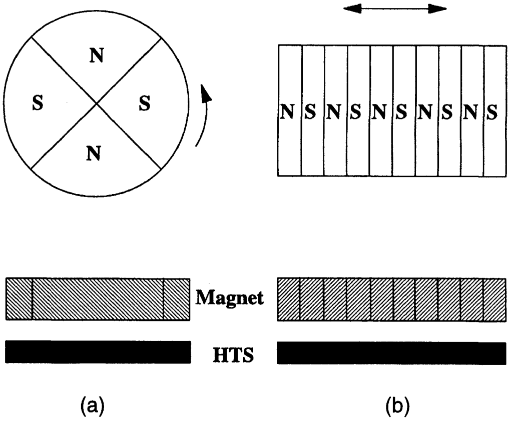

FIG. 30. Different HTS/magnet assemblies for (a) clutches and (b) vibration dampers (Chu *et al.*, 1992).

5.6 Electrodes, Fault-Current Limiters, and Power-Transmission Cables

In many applications of LTSrs, large currents have to be transferred from room temperature to the 4.2-K operating environment. Heavy Cu cables with low electrical resistance are usually used to reduce Joule heating and, thus, energy waste and liquid-helium evaporation. Unfortunately, Cu is also a good thermal conductor, which leads to a large helium boiloff. Melt-textured Y123 has a large J_c and very low thermal conductivity, several hundreds to a thousand times smaller than Cu (Fig. 18). Long melt-textured rods with a large cross section can be used as electrodes to transmit large currents from 77 to 4.2 K. Y123 bars ~10 cm long and with a $0.4 \times 0.4\text{-cm}^2$ cross section have been prepared at TCSUH to carry a total critical current up to 1000 A at 77 K. Rods are also easily made of sintered Y123 with high dimensional precision by extrusion to carry a total current of 125 A. An electrode consisting of many sintered Y123 rods was reported by Argonne National Laboratory as capable of carrying a critical current of 2×10^3 A. General Electric Company has developed a magnetic resonance imaging system requiring no liquid cryogen, using HTSg electrodes as current leads, as shown in Fig. 31.

Falling power lines or lightning can cause "fault currents," momentary spikes many times greater than normal operating currents, making it necessary to safeguard gen-

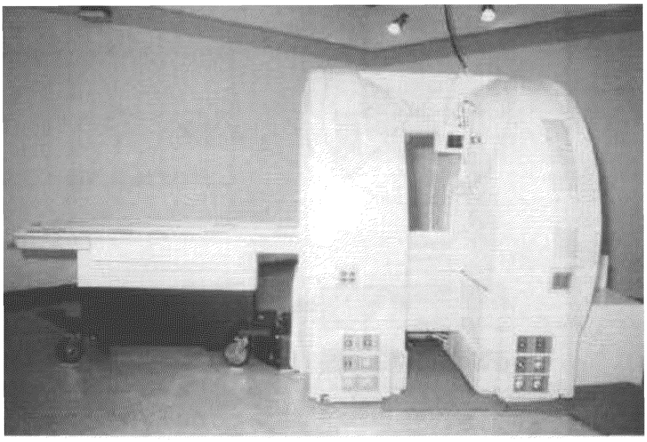

FIG. 31. An MRI that requires no liquid cryogen to operate, using HTSg electrodes for current leads (General Electric).

erators, transmission lines, and transformers and to minimize the down time to reduce customer inconvenience. An HTSg current limiter limits the current flow by switching from a nonresistive superconducting state to a resistive normal state when subjected to the fault current. An effective current limiter is required to consume no energy during its normal operation and to have a very short switching time, e.g., less than 0.1 s. The demand on HTSg material by such a device is rather stringent in terms of large J_c, large normal-state resistance, and small thermal volume. Several device concepts have been proposed and are being pursued (Giese *et al.*, 1994).

A HTSr-based power-transmission system must be a long-term effort. Because of the required refrigeration, use of such a system will be limited to underground lines. The Electrical Power Research Institute (EPRI) has identified two areas where HTSg power-transmission systems appear to have compelling advantages over conventional systems (Dales *et al.*, 1990). The first is to retrofit existing underground circuits in congested urban areas in order to increase the power ratings. The other is for underground lines longer than 25 miles, where conventional systems require large and expensive inductors every 20 miles or so. American Superconductor Corporation has fabricated a 1-km length of HTSg wire with a $J_c \sim 13\,800$ A/cm^2 and a total critical current of 16 A. They have also made a 1-m flexible cable (of diameter $\sim$5 cm, including the supporting structure) that can carry a total current of 4200 A at 77 K without loss. The target for major cable applications is HTSg cables with a $J_c \sim 20\,000$–$200\,000$ A/cm^2 in kilometer lengths.

6. FUTURE TRENDS

It has long been known that superconductivity offers both unusual scientific challenge and great technological potential. The discovery of superconductivity above 77 K in 1987 heralded a new era of high-temperature superconductivity. As a result, many applications of superconductors appear to be more practical than ever before. Enormous effort has thus been devoted to the study of HTSy, and impressive progress has subsequently been made at an unprecedentedly fast pace since 1987. As described in the previous sections, more than 90 compounds with a T_c higher than 23 K and more than 30 with a T_c higher than 77 K have been discovered. Many physical properties characteristic of the HTSg materials that are significant to the understanding and use of HTSrs have been determined, various theoretical models to account for these properties proposed, and numerous processing techniques to improve the performance of HTSrs for scientific pursuit and practical applications developed. Many old concepts for the application of superconductivity have been demonstrated using HTSrs, and new concepts are being tested, prototypes of some HTSg devices are being built and tested, and even a few HTSg products are already commercially available.

In spite of the great effort devoted to HTSy and the impressive progress made in HTSy, several important scientific questions have yet to be answered, and viable commercial products have yet to be identified. A few questions by way of example might illustrate this: What is the underlying mechanism responsible for HTSy? What is the symmetry of the HTSg order parameter? What is the dimensionality of HTSy? Can HTSy (especially for a $T_c > 77$ K) exist in a truly three-dimensional or isotropic electron system? Is there a T_c ceiling; if so, what is it? As T_c becomes higher and higher, will ξ become shorter and shorter, making $\rho = 0$ and $B \neq 0$ for a HTSr in an operational sense? What is the genuine H-T phase diagram of a HTSr, especially between H_i and H_{c2}? While answers to these questions might seem to be purely academic, they will have major impacts on HTSg technology. On the application front, many of the HTSg device prototypes that have been built and tested will no doubt have important roles in HTSg technology, but the economic impact of any device is still not clear.

To a great extent, the impasses that are faced to develop fully the HTSy science and technology are material based. There is a need for HTSg materials of high quality and large dimensions, and (if possible) noncuprate based. This will enable the research community to carry out the critical experiments to unravel the mystery of HTSy and thus make possible the development of a comprehensive HTSy theory that is simple, general, descriptive, and predictive as well. The chemistry of HTSg materials is rather complicated. A systematic study of the phase equilibria of these material systems will enable one to achieve the material quality needed for scientific study, to improve the material performance to that required for applications, and to provide guidance to the discovery of new HTSg compounds. The search for noncuprate HTSrs will help reveal if Cu plays an indispensable role in HTSy, a concept on which a great majority of the known HTSy theories is based. Attempts to synthesize these novel materials under extreme conditions, e.g., high pressure or fast quenching, may prove to be fruitful.

Many reports of anomalies in oxides suggestive of a superconducting transition at temperatures above 200 K have appeared (Levi, 1994; Service, 1994) after the discovery of HTSy. They usually involve a small drop in resistivity (but not to zero) or a small diamagnetic shift in susceptibility (but only superimposed on a large paramagnetic background). They display poor reproducibility, not just from laboratory to laboratory and from sample to sample, but also from run to run of the same sample. Some of these fleeting anomalies have been shown to arise from experimental artifacts. Others are known at best as unidentified superconducting objects (USOs). The numerous persistent reports of resistive (or magnetic) anomalies in multiphase cuprates from many credible laboratories in the temperature range of 150–300 K are too tantalizing to ignore. It has been conjectured that the absence of a clear Meissner effect may be a result of the ever-increasing difficulty to detect the effect due to the ever-decreasing coherence length as T_c increases. There presently exists no theory that sets a T_c limit in oxides. However, the possibility that the cause of the reported resistive anomalies is the sensitive dependence of oxide resistance on humidity and gases has not yet been ruled out completely.

Conflicting characteristics of HTSrs from different types, and sometimes the same type, of experiments have been reported, e.g., some results support d pairing but others favor s pairing. While sample quality has often been the suggested reason for the inconsistent reports, the complex chemistry of HTSg materials and the possible interaction of the experimental probe with the HTSg sample may have complicated the experiments. Extra caution is thus needed to perform the experiments and to interpret the observations. The implication of the pairing symmetry in the HTSg mechanism needs more study.

HTSg material processing will play a critical role in making applications of HTSrs practical. The material has to be made with the optimal superconducting properties (e.g., high T_c and J_c), the desired dimensions (e.g., long length and large cross section for larger-current applications, or large-area thin film for some small-current uses), and required forms (e.g., wires or films) for devices, all at low cost. Techniques for device fabrication with high yield at low cost are equally important. Ultimately, it will be the cost that differentiates success from failure in the commercialization of the product.

While a quantum leap is always welcome in science and technology, incremental improvement should never be overlooked. Insertion of HTSg devices into an existing system, to improve the system performance, such as the successful use of HTSg filters to enhance the capacity of a mobile phone base station, is one such example. Although many concepts for the application of HTSrs have been successfully demonstrated, other significant commercial uses of HTSrs that are currently unforeseen may be discovered. They may include the use of the normal-state properties of HTSrs.

ACKNOWLEDGMENTS

The author would like to thank the many scientists and engineers who have contributed to the progress made in HTSy to date, without whom this article could not have been written. He would like to express his great appreciation to his colleagues at TCSUH, particularly Y. Y. Xue and F. Chen for their help in compiling some of the information and preparing the figures. He is also grateful to George Trigg, the editor, and to Lorin Vant-Hull, Donn Forbes, and Mike Albert, who have made valuable suggestions that helped make the article more readable. The work is partially supported by NSF, EPRI, and TCSUH.

Works Cited

Abrikosov, A. A. (1995), *J. Phys. Chem. Solids* (in press).

Allen, P. B. (1982), *Solid State Physics,* Vol. 37, New York: Academic, pp. 2–92.

Anderson, P. W. (1987), *Science* **235,** 1196–1198.

Annett, J. F., Goldenfeld, N., Renn, S. R. (1989), in: D. Ginsberg (Ed.), *Physical Properties of High Temperature Superconductivity I,* Singapore: World Scientific, pp. 571–686.

Asano, H., Satoh, M., Konaha, T. (1991), *Appl. Phys. Lett.* **58,** 2981–2983.

Bardeen, J., Cooper, L. N., Schrieffer, J. R. (1957), *Phys. Rev.* **106,** 162–164.

Baskaran, G., Zou, Z. Z., Anderson, P. W. (1987), *Solid State Commun.* **63,** 973–976.

Bednorz, J. G., Müller, K. A. (1986), *Z. Phys. B* **64,** 189–193.

Berkeley, D. D., Skelton, E. F., Moulton, N. E., Osofsky, M. S., Lechten, W. T., Browning, V. M., Liebenberg, D. H. (1993), *Phys. Rev. B* **47,** 5524–5527.

Brandt, E. (1991), *Int. J. Mod. Phys. B* **5,** 751–795.

Campuzano, J. C., Jennings, G., Faiz, M., Beaulaique, L., Veal, B. W., Liu, J. Z., Paulikae, A. P., Vandervoort, K., Claus, C., List, R. S., Arko, A. J., Bartlett, R. J. (1990), *Phys. Rev. Lett.* **64,** 2308–2311.

Cava, R. J., Batlogg, B., Krajesky, T. J., Farrow, R., Rupp, Jr., L. W., White, A. E., Short, K., Peck, W. F., Kometani, T. (1988), *Nature* **332,** 814–816.

Char, K., Coclough, M. S., Garrison, S. M., Newman, N., Zabarchuki, G. (1991), *Appl. Phys. Lett.* **59,** 733–735.

Chen, C. H. (1990), in: D. M. Ginsberg (Ed.), *Physical Properties of High Temperature Superconductors II,* Singapore: World Scientific, pp. 199–267.

Chen, F., Xue, Y. Y., Chu, C. W. (1994), unpublished.

Chien, T. R., Brawner, D. A., Wang, Z. Z., Ong, N. P. (1991a) *Phys. Rev. B* **43,** 6242–6245.

Chien, T. R., Wang, Z. Z., Ong, N. P. (1991b), *Phys. Rev. Lett.* **67,** 2088–2091.

Chu, C. W. (1987), *Proc. Natl. Acad. Sci. USA* **84,** 4681.

Chu, C. W. (1994), *J. Superconductivity* **7,** 1–7.

Chu, C. W., Xue, Y. Y., Gao, L., Meng, R. L., Ramirez, D. (1994), patent pending.

Chu, W. K., Ma, K. B., McMichael, C. K. (1992), US Patent No. 5,159,219.

Chu, W. K., Xia, Z., Chen, Q., Ma, K., Cooley, R., Fowler, C., McMichael, C., Lamb, M. (1995), 1995 International Workshop on Superconductivity Co-Sponsored by ISTEC and MRS (Maui, HI), unpublished.

Clarke, J. (1989), in: H. Weinstock and M. Nisenoff (Eds.), *Superconducting Electronics,* Berlin: Springer-Verlag, pp. 87–148.

Clayhold, J., Ong, N. P., Wang, Z. Z., Tarascon, J. M., Barboux, P. (1989), *Phys. Rev. B* **39,** 7324–7327.

Cooper, L. (1956), *Phys. Rev.* **104,** 1189–1190.

Crespi, V., Cohen, M. L., Penn, D. R. (1991), *Phys. Rev. B* **43,** 12921–12924.

Dales, S. J. N., Wolf, S. M., Schneider, T. R. (1990), *Energy Applications of High Temperature Superconductivity,* EPRI Report ER-6682, Vol. 1 (February).

Daly, K. P., Dozier, W. D., Burch, J. F., Coons, B., Hu, R., Platt, C. E., Simon, R. (1991), *Appl. Phys. Lett.* **58,** 543–545.

Davis, M. F., Wosik, J., Forster, K., Deshmukh, S. C., Rampersad, H. R., Shah, S. (1991), *J. Appl. Phys.* **69,** 7182–7188.

de Gennes, P. G. (1966), *Superconductivity of Metals and Alloys,* translated by P. A. Pincus, New York: W. A. Benjamin, pp. 171–201.

DiIorio, M. S., Yang, K. Y., Maung, M., Zhang, J., Power, B. (1993), *IEEE Trans. Appl. Supercond.* **3,** 2011–2017.

Dimos, D., Chaudhari, P., Mannhart, J. (1990), *Phys. Rev. B* **41,** 4038–4049.

Drabek, L., Holczer, K., Grüner, G., Chang, J. J., Scalapino, D. J., Venkatesen, T. (1990), *J. Superconductivity* **3,** 317–322.

Eom, C. B., Sun, J. Z., Yamamoto, K., Marshall, A. F., Luther, K. E. (1989), *Appl. Phys. Lett.* **55,** 595–597.

Fink, R. L., Thompson, M., Hilbert, C., Kroger, H. (1992), *Appl. Phys. Lett.* **61,** 595–597.

Friedl, G., Daalmans, G., Vidldic, M., Uhl, D., Fömmel, F., Roas, B., Hillenbrand, B., Stritzker, B., Hoenig, H. E. (1993), *IEEE Trans. Appl. Supercond.* **3,** 2018–2021.

Frölich, H. (1950), *Phys. Rev.* **79,** 845–856.

Gao, L., Xue, Y. Y., Chen, F., Xiong, Q., Meng, R. L., Ramirez, D., Chu, C. W., Eggert, H., Mao, H. K. (1994a), *Physica C* **235–240,** 1493–1494.

Gao, L., Xue, Y. Y., Meng, R. L., Chu, C. W., Hannahs, S., Thang, V., Rubin, L. (1994b), *Appl. Phys. Lett.* **65,** 707–909.

Giese, R. F., Runde, M., Sheahen, T. P. (1994), in: T. P. Sheahen (Ed.), *Introduction to High-Temperature Superconductivity,* New York: Plenum, Ch. 23.

Gofron, K., Campuzano, J. C., Abrikosov, A. A., Lindoors, M., Bausil, A., Ding, H., Koelling, D., Dabrowski, B. (1994), *Phys. Rev. Lett.* **73,** 3302–3305.

Gross, R., Chaudhari, P., Kawasaki, M., Gupta, A. (1991), *IEEE Trans. Magn.* **27,** 3227–3230.

Hagen, S. J., Jing, T. W., Wang, Z. Z., Horvath, J., Ong, N. P. (1988), *Phys. Rev. B* **37,** 7928–7931.

Haldar, P., Motowidlo, L. J. (1992), *J. Met.* **44,** 54–58.

Hazen, R. M. (1989), in: D. M. Ginsberg (Ed.), *Physical Properties of High Temperature Superconductors II,* Singapore: World Scientific, pp. 121–198.

Hebard, A. F. (1993), in: S. Cooperman (Ed.), *Annual Review of Materials Science,* Vol. 23, Palo Alto, CA: Annual Reviews, pp. 159–191.

Hebard, A. F., Rosseinsky, M. J., Haddon, R. C., Murphy, D. W., Glarum, S. H., Palstra, T. T. M., Ramirez, A. P., Kartan, A. M. (1991), *Nature* **350,** 600–601.

Hor, P. H., Meng, R. L., Wang, Y. Q., Gao, L., Huang, Z. J., Bechtold, J., Forster, K., Chu, C. W. (1987), *Phys. Rev. Lett.* **58,** 1891–1894.

Huang, Z. J., Xue, Y. Y., Meng, R. L., Chu, C. W. (1994), *Phys. Rev. B* **49,** 4219–4231.

Huse, D. A., Fisher, M. P. A., Fisher, D. S. (1992), *Nature* **358,** 553–559.

Inam, A., Hegde, M. S., Wu, X. D., Venkatesan, T., England, P., Mieceli, P. F., Chase, E. W., Chang, C. C., Tarascon, J. M., Wachtman, J. B. (1988), *Appl. Phys. Lett.* **53,** 908–910.

Ito, T., Takagi, H., Ishibashi, S., Ido, T., Uchida, S. (1991), *Nature* **350,** 596–597.

Iye, Y. (1992), in: D. Ginsberg (Ed.), *Physical Properties of High Temperature Superconductors III,* Singapore: World Scientific, pp. 285–361.

Iye, Y., Tamegai, T., Sakakibara, T., Goto, T., Miura, N., Takeya, H., Takei, H. (1988), *Physica C* 153–155, 26–31.

Kawai, M., Kawai, T., Kawai, S. (1988), *Appl. Phys. Lett.* **54,** 1802–1804.

Kawasaki, M., Chaudhari, P., Newman, T. H., Gupta, A. (1991), *Appl. Phys. Lett.* **58,** 2555–2557.

Kingston, J. J., Wellstood, F. C., Lerch, P., Miklich, A. H., Clarke, J. (1990), *Appl. Phys. Lett.* **56,** 189–191.

Klein, N. (1991), KFA Jülich, private communication.

Klein, N., Chaloupka, H., Müller, G., Orbach, S., Piel, H., Roas, B., Schultz, L., Klein, U., Pienger, M. (1990), *J. Appl. Phys.* **69,** 6940–6945.

Koelle, D. (1994), OE/LASE'94 SPIE Conference (Los Angeles, CA), unpublished.

Koelle, D., Miklich, A. H., Dantsker, E., Ludwig, F., Nemeth, D. T., Clarke, J., Ruby, W., Char, K. (1993), *Appl. Phys. Lett.* **63,** 3630–3632.

Levi, B. G. (1993), *Phys. Today* **46** (5), 17–20.

Levi, B. G. (1994), *Phys. Today* **47** (2), 17–18.

Li, X., Kawai, T., Kawai, S. (1992), *Jpn. J. Appl. Phys.* **31,** L934–937.

Ma, K. B., McMichael, C. K., Lamb, M. A., Chu, W. K. (1992), in: C. W. Chu, W. K. Chu, P. H. Hor, K. Salama (Eds.), *Proceedings of 1992 TCSUH Workshop, HTS Materials, Bulk Processing and Bulk Applications,* Singapore: World Scientific, pp. 419–429.

Maeda, H., Tanaka, Y., Fukutomi, M., Asano, T. (1988), *Jpn. J. Appl. Phys.* **27,** L209–210.

Mankiewich, P. M., Scofield, J. H., Skocpol, W. J., Howard, R. E., Dayem, A. M., Good, E. (1987), *Appl. Phys. Lett.* **51,** 1753–1755.

Martin, S., Fiory, A. T., Fleming, R. M., Schneemeyer, L. F., Waszczak, J. V. (1988), *Phys. Rev. Lett.* **60,** 2194–2197.

Matthias, B. T. (1953), *Phys. Rev.* **92,** 874–876.

Meng, R. L., Gao, L., Gautier-Picard, P., Ramirez, D., Sun, Y. Y., Chu, C. W. (1994), *Phys. C* **232,** 337–346.

Meng, R. L., Kinalidis, C., Sun, Y. Y., Gao, L., Tao, Y. K., Hor, P. H., Chu, C. W. (1990), *Nature* **345,** 326–328.

Miklick, A. H., Kingston, J. J., Wellstood, F. C., Clarke, J., Colclough, M. S., Chat, K., Zaharchuk, G. (1991), *Appl. Phys. Lett.* **59,** 988–990.

Montoux, P., Pines, D. (1992), *Phys. Rev. Lett.* **69,** 961–964.

Moon, F. C., Chang, P. Z. (1990), *Appl. Phys. Lett.* **56,** 397–399.

Mueck, M. (1993), *IEEE Trans. Appl. Supercond.* **3,** 2003–2010.

Murakami, M., Morita, M., Doi, K., Miyamoto, K. (1989), *Jpn. J. Appl. Phys.* **28,** 1189–1194.

Murphy, D. W., Sunshine, S., van Dover, R. D., Cava, R. J., Batlogg, B., Zahurak, S. M., Schneemeyer, L. F. (1987), *Phys. Rev. Lett.* **58,** 1888–1890.

Newman, N., Cole, B. F., Garrison, S. M., Char, K., Taber, R. C. (1991), *IEEE Trans. Magn.* **27,** 1276–1279.

Newman, N., Lyons, W. G. (1993), *J. Supercond.* **6,** 119–160.

Newns, D. M., Krishnamurphy, H. R., Pattnaik, P. C., Tsuei, C. C., Kane, C. L. (1992), *Phys. Rev. Lett.* **69,** 1264–1267.

Oates, D. E., Anderson, A. C., Sheen, D. M., Ali, S. M. (1991), *IEEE Trans. Microwave Theory Tech.* **39,** 1522–1529.

Obertelli, S. D., Cooper, J. R., Tallon, J. L. (1992), *Phys. Rev. B* **46,** 14928–14931.

Ong, N. P. (1992), in: D. Ginsberg (Ed.), *Physical Properties of High Temperature Superconductors II,* Singapore: World Scientific, pp. 459–507.

Ossipyan, Yu. A., Timofeev, V. B., Schegolev, I. F. (1988), *Physica C* **153–155,** 1133–1137.

Parkin, S. S. P., Lee, V. Y., Nazzal, A. I., Savoy, R., Bayer, R., LaPlaca, S. (1988), *Phys. Rev. Lett.* **61,** 750–753.

Pei, S., Jorgensen, J. D., Dabrowski, B., Hinks, D. G., Richard, D. R., Mitchell, A. W., Newsam, J. W., Sinka, S. K., Vakuin, D., Jacobson, A. J. (1990), *Phys. Rev. B* **41,** 4126–4141.

Presland, M. R., Tallon, J. L., Buckley, R. G., Liu, R. S., Flower, N. E. (1991), *Physica C* **176,** 95–105.

Putilin, S. N., Antipov, E. V., Chmaissem, O., Marezio, M. (1993), *Nature* **362,** 226–228.

Rensch, D. B., Josvowicz, J. Y., MacDonald, P., Nich, C. W., Hoefer, W., Krajenbrink, F. (1991), *IEEE Trans. Magn.* **27,** 2553–2556.

Reznik, D., Cooper, S. L., Klein, M. V., Lee, W. C., Ginsberg, D. M., Maksimov, A. A., Puchkov, A. V., Tartakovski, I. I., Cheng, S. W. (1993), *Phys. Rev. B* **48,** 7624–7635.

Riley, G. N. (1994), *DoE Annual Review,* Washington, DC: U.S. GPO.

Roberts, B. W. (1976), *J. Phys. Chem. Ref. Data* **5,** 581.

Salama, K., Selvamanickam, V., Gao, L., Sun, K. (1989), *Appl. Phys. Lett.* **54,** 2352–2354.

Sato, K., Hikata, T., Mukai, H., Veyama, M., Shibuta, N., Kato, T., Matsuda, T., Nagata, M., Iwata, K., Mitsui, T. (1991), *IEEE Trans. Magn.* **27,** 1231–1238.

Sawano, K., Morita, M., Tanaka, M., Sasaki, T., Kimura, K., Takebayashi, S., Kimura, M., Miyamoto, K. (1991), *Jpn. J. Appl. Phys.* **30,** L1157–L1159.

Schilling, A., Cantoni, M., Guo, J. D., Ott, H. R. (1993), *Nature* **363,** 56–58.

Schneemeyer, L. F. (1993), in: S. Jin (Ed.), *Processing and Properties of High T_c Superconductors,* Vol. 1, Singapore: World Scientific, p. 45.

Schrieffer, J. R. (1994), *Solid State Commun.* **92,** 129–139.

Seido, M., Sato, J., Sasaoka, T., Nomoto, A., Kamo, T., Matsuda, S. (1992), *Proceedings of the International Workshop on Superconductivity,* Singapore: World Scientific, pp. 236–239.

Service, R. F. (1994), *Science* **265,** 2015.

Shen, Z. X., Dessau, D. S., Wells, B. O., King, D. M., Spicer, W. E., Arko, A. J., Marshall, D., Lombardo, L. W., Kapitulnik, A., Dickinson, P., Doniack, S., DiCarlo, J., Loeser, T., Park, C. H. (1993), *Phys. Rev. Lett.* **70,** 1553–1556.

Shen, Z. Y., Wilker, C., Pang, P., Holstein, W. L., Face, D. W., Kountz, D. J. (1992), *IEEE Trans. Microwave Theory Tech.* **40,** 2424–2432.

Sheng, Z. Z., Hermann, A. M. (1988), *Nature* **332,** 138–139.

Sugai, S. (1989), in: H. Kamimura, A. Oshiyama (Eds.), *Mechanisms of High Temperature Superconductivity,* Heidelberg: Springer-Verlag, p. 207.

Sun, A. G., Gajewski, D. A., Maple, M. B., Dynes, R. C. (1994), *Phys. Rev. Lett.* **72,** 2267–2270.

Sun, J. Z., Gallagher, W. J., Callegari, A. C., Foglietti, V., Koch, R. H. (1993), *Appl. Phys. Lett.* **63,** 1561–1563.

Tozer, S. W., Kleinsasser, A. W., Penny, T., Kaiser, D. Holtzberg, F. (1987), *Phys. Rev. Lett.* **59,** 1768–1771.

Tsuei, C. C., Kirtley, J. R., Chi, C. C., Yu-Jahnes, L. S., Gupta, A., Shaw, T., Sun, J. Z., Ketchen, M. (1994), *Phys. Rev. Lett.* **73,** 593–596.

Uher, C. (1992), "Thermal Conductivity of High-Temperature Superconductors," in: D. M. Ginsberg (Ed.), *Physical Properties of High Temperature Superconductors III,* Singapore: World Scientific, pp. 159–284.

Valenzuela, A. A., Daalmans, B., Roas, B. (1991), *Electron. Lett.* **25,** 1435–1436.

Varma, C. M., Littlewood, P. B., Schmitt-Rink, S., Abrahams, E., Ruckenstein, A. E. (1989), *Phys. Rev. Lett.* **63,** 1996–1999.

Virshup, G. F., Klausmeier-Brown, M. E., Bozavic, I., Eckstein, J. N. (1992), *Appl. Phys. Lett.* **60,** 2288–2290.

Weinstein, R., Chen, I. G., Liu, J., Parks, D., Selvamanickam, V., Salama, K. (1990), *Appl. Phys. Lett.* **56,** 1475–1477.

Wohlman, D. A., van Harlingen, D. J., Lee, W. C., Ginsberg, D. M., Leggett, A. J. (1993), *Phys. Rev. Lett.* **71,** 2134–2137.

Wu, M. K., Ashburn, J. R., Torng, C. J., Hor, P. H., Meng, R. L., Gao, L., Huang, Z. J., Wang, Y. Q., Chu, C. W. (1987), *Phys. Rev. Lett.* **58,** 908–910.

Wu, X. D., Foltyn, P., Arendt, P., Townsend, J., Adams, C., Campbell, I. H., Tiwari, P., Coulter, Y., Peterson, D. E. (1994), *Appl. Phys. Lett.* **65,** 1961–1963.

Xing, D. Y., Liu, M., Ting, C. S. (1988), *J. Phys. C* **21,** L591-597.

Yan, M. F., Ling, H. C., O'Bryan, H. M., Gallagher, P. K., Rhodes, W. W. (1988), *Mater. Res. Soc. Symp. Proc.* **121,** 385–400.

Zhang, Y., Mück, M., Hermann, K., Schubert, J., Zander, W., Braginski, A. I., Heiden, C. (1993), *IEEE Trans. Appl. Supercond.* **3,** 2465–2468.

Further Reading

Brief History

Chu, C. W. (1987), *Proc. Natl. Acad. Sci. U.S.A.* **84,** 4681.

Chu, C. W. (1988), in: W. Fickinger, K. L. Kowalski (Eds.), *Modern Physics in America: A Michelson–Morley Centennial Symposium, Cleveland, Ohio, 1987,* American Institute of Physics Conference Series No. 169, New York: AIP.

Hazen, R. M. (1988), *The Breakthrough,* New York: Ballantine.

Müller, K. A., Bednorz, J. G. (1987), *Science* **237,** 1133.

Schechter, B. M. (1989), *The Path of No Resistance,* New York: Simon and Schuster.

Simon, R., Smith, A. (1988), *Superconductors,* New York: Plenum.

Vidali, G. (1993), *Superconductivity: The Next Revolution?,* Cambridge, U.K.: Cambridge Univ. Press.

Brief General Overview

Cyrot, M., Paruma, D. (1992), *Introduction to Superconductivity and High T_c Materials,* Singapore: World Scientific.

Fossheim, K. (1991), *Superconducting Technology,* Singapore: World Scientific.

Rowell, J. (1991), *Phys. Today* **44** (6), 22–82.

Sheahan, T. P. (1994), *Introduction to High Temperature Superconductivity,* New York: Plenum.

Brief Summary on Current Trends

Lubkin, G. L. (1995), *Phys. Today* **48** (3), 20–23.

Yam, P. (1993), *Sci. Am.* **269** (6), 118–126.

Detailed Properties

Ginsberg, D. M. (Ed.) (1989, 1990, 1992, 1994), *Physical Properties of High Temperature Superconductors,* Vols. 1–4, Singapore: World Scientific.

Bulk Synthesis

Jin, S. (Ed.) (1993), *Processing and Properties of High T_c Superconductors,* Singapore: World Scientific.

Bulk Applications

Chu, C. W., Chu, W. K., Hor, P. H., Salama, K. (Eds.) (1992), *HTS Materials, Bulk Processing and Bulk Applications,* Singapore: World Scientific.

Energy Applications

EPRI (Eds.) (1990, 1992), *Energy Applications of High Temperature Superconductivity,* Vols. 1 and 2, Electric Power Research Institute Reports No. ER-6682 and TR-101635, Stanford, CA: EPRI.

Thin-Film Processing

Mohan, S. (1991), "High T_c Superconducting Thin Films, Processing Method and Properties," in: C. N. R. Rao (Ed.), *Chemistry of High Temperature Superconductors,* Singapore: World Scientific, p. 411.

Thin-Film Applications

Weinstock, H., Ralston, R. W. (1993), *New Superconducting Electronics,* Dordrecht: Kluwer Academic.

SQUIDs

Clarke, J. (1991), in: R. G. Lerner, G. L. Trigg (Eds.), *Encyclopedia of Physics,* 2nd ed., New York: VCH Publishers, pp. 1194–1198.

Mechanisms

Bedell, K. (Ed.) (1990), *High Temperature Superconductivity: Proceedings,* Redwood City, CA: Addison-Wesley.

HTSg Pairing Symmetry

Schrieffer, J. R. (1994), *Solid State Commun.* **92,** 129–139.

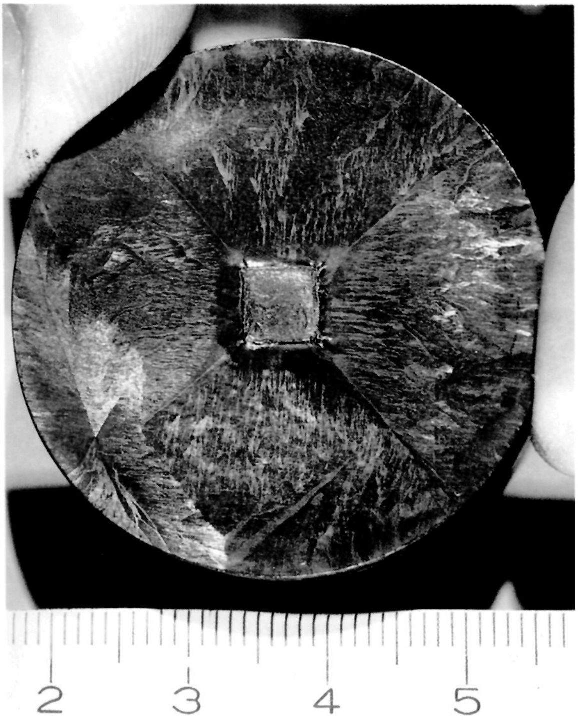

Vol. 20 – PLATE 1. A large quasi single-grain Y123 disk grown by the top-seed technique (Meng et al., 1994). Scale is in centimeters.

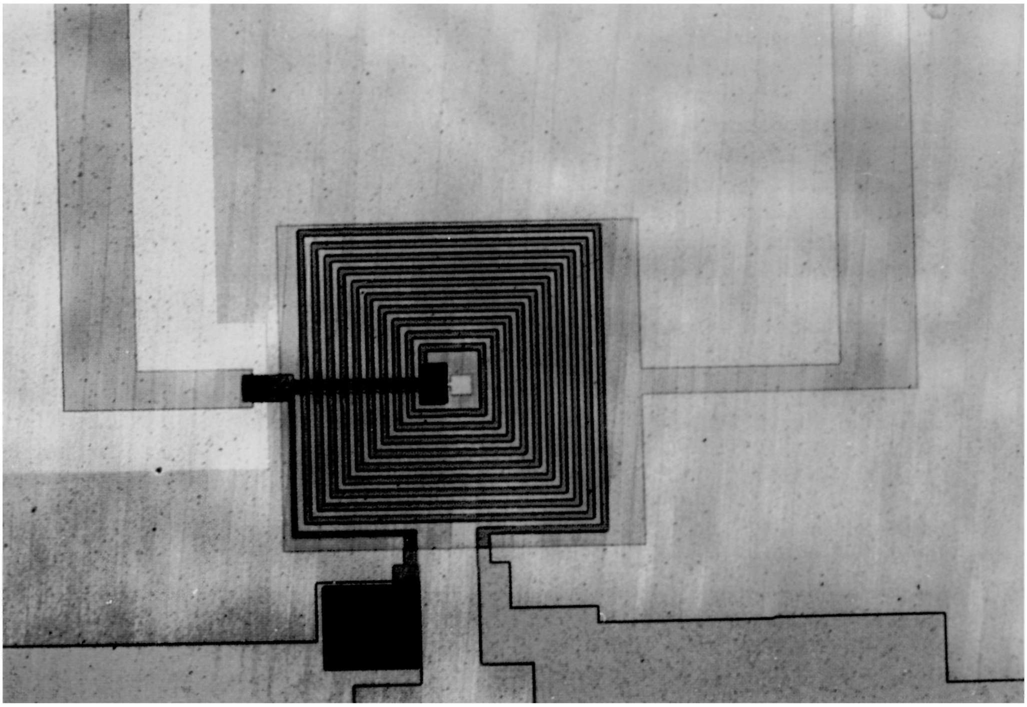

Vol. 20 – PLATE 2. A HTSg SQUID magnetometer with a multitum flux transformer (courtesy of Conductus).

SUPERCONDUCTIVITY, LOW-TEMPERATURE

JAN HERRMANN, *Institute for Pure and Applied Physical Sciences, University of California, San Diego, La Jolla, California, U.S.A.*

M. BRIAN MAPLE, *Institute for Pure and Applied Physical Sciences and Department of Physics, University of California, San Diego, La Jolla, California, U.S.A.*

Introduction 249
Brief History of Superconductivity 250
1. **Basic Properties of Superconductors** 252
1.1 Zero Resistivity (Infinite Conductivity) 252
1.2 Ideal Diamagnetism 252
1.2.1 Meissner Effect 252
1.2.2 Magnetic Penetration Depth 253
1.3 Critical Magnetic Fields 254
1.3.1 Type I Superconductivity 254
1.3.2 Intermediate State 254
1.3.3 Type II Superconductivity 255
1.3.3.1 Superconducting Coherence Length 255
1.3.3.2 Abrikosov Vortex Lattice 255
1.3.3.3 Magnetic Fluxoid 256
1.4 Critical Current Density 257
1.5 Energy Gap 257
1.5.1 Specific Heat 257
1.5.2 Electromagnetic and Acoustic Absorption 258
1.5.3 Electron Tunneling 258
1.6 Josephson Effects 258
1.7 Isotope Effect 259
2. **Theory of Superconductivity** 260
2.1 BCS Theory 260
2.1.1 Cooper Pairs 260
2.1.2 Electron-Pairing Mechanism 260
2.2 Ginzburg–Landau Theory 261
3. **Superconducting Materials** 262
3.1 Elements 262
3.2 Binary Alloys, *A*15 Compounds 262
3.3 Ternary Compounds 264
3.4 Magnetically Ordered Superconductors 264
3.5 Heavy-Fermion Superconductors 264
3.6 Low–Electron-Density Superconductors 264
3.7 Low-Dimensional Superconductors 265
3.8 Organic Superconductors 265
3.9 Carbon-Based Superconductors 266
4. **Applications of Superconductivity** 266
4.1 Large-Scale Applications 267
4.1.1 Superconducting Wires 267
4.1.2 Superconducting Magnets 268
4.1.3 Motors and Generators 268
4.1.4 Electrical Power Transmission and Distribution 268
4.2 Small-Scale Applications 269
4.2.1 Microwave Applications 269
4.2.2 Superconducting Quantum Interference Devices (SQUIDs) 269
4.2.3 Superconducting Digital Electronics 269
4.2.4 Josephson-Junction Noise Thermometry 270
4.2.5 Josephson-Junction Voltage Standard 270
Glossary 270
Further Reading 271

INTRODUCTION

Superconductivity, the disappearance of the dc electrical resistivity of a material cooled below a characteristic temperature, is a fascinating phenomenon that has attracted the interest of physicists since its discovery in 1911. From the very beginning, researchers have been attracted not only by the wealth of interesting physical phenomena associated with superconductivity, but also by the technological applications that are imag-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

inable with the prospect of circulating electrical current in superconducting wires with virtually no losses.

Today, with many well-established applications of superconductivity at both large and small scales and a renewed interest in the mechanisms and phenomena associated with superconductivity that was sparked by the discovery of new classes of superconductors in the years since 1986, superconductivity remains an exciting field of condensed-matter physics with many interesting questions remaining to be answered.

Brief History of Superconductivity

In the year 1908, the noble gas helium was liquefied for the first time by H. Kammerlingh Onnes at Leiden, thus allowing the achievement of low temperatures that were inaccessible before. In 1911, investigating the low-temperature electrical resistance of metals, Kammerlingh Onnes observed that the resistance of a mercury sample sharply dropped to immeasurably small values at about 4.15 K and called the new state of matter characterized by vanishing electrical resistance "superconductivity." (Mercury was chosen because it could be obtained in very pure form by simple distillation.) Shortly thereafter, he also observed the suppression of electrical superconductivity by currents, establishing the concept of critical currents (see Sec. 1.4). In 1912, Kammerlingh Onnes found tin and lead to be superconducting, a discovery that made experiments on wires and coils much easier compared to the brittle and mechanically unstable threads of frozen mercury. He was awarded the Nobel prize in physics for the year 1913 for "his investigations on the properties of matter at low temperatures." Early 1914 saw the discovery of the existence of a threshold magnetic field above which superconductivity is suppressed, the critical magnetic field (see Sec. 1.3). Furthermore, the first experiments devised to determine the time constant of the decay of a current circulating in a superconducting loop and thus of an upper limit for the residual resistance of the superconductor were performed, demonstrating the presence of persistent currents (see Sec. 1.1).

In 1916, F. B. Silsbee hypothesized that the critical current is the current that produces the critical field at the surface of a superconductor, thus explaining the critical current as a consequence of the existence of a critical magnetic field. New superconductors were discovered at Leiden in 1919 (thallium and the lead isotope ^{206}Pb) and in 1922 (indium). It was not possible, at that time, to detect any sizable variation of the superconducting transition temperature T_c with the atomic mass (the so-called isotope effect; see Sec. 1.7).

With further laboratories becoming able to perform experiments at liquid-helium temperatures, in the course of the following years, more elements and, to an increasing extent, binary alloys were found to be superconducting, some of which exhibited critical fields orders of magnitude higher than those of the superconducting elements. This put practical applications—e.g., superconducting magnets (see Sec. 4.1)—within reach.

In 1932, the specific-heat jump associated with the superconducting transition in zero field was observed in tin and thallium, again at Leiden.

The second characteristic property of the superconducting state, the complete expulsion of magnetic flux from the interior of a superconductor, was discovered in 1933 by W. Meissner and R. Ochsenfeld. The observation of the Meissner effect established the superconducting state as a distinct thermodynamic equilibrium phase (see Sec. 1.2.1). Shortly thereafter, the concept of the intermediate state (see Sec. 1.3.2) was established.

In 1935–1936, the magnetic properties of type II superconductors with their two characteristic fields above which field penetration and restoration of resistance set in were uncovered by L. V. Shubnikov. Magnetization measurements provided conclusive evidence for the existence of the mixed state where magnetic flux partially penetrates the superconductor.

The theoretical description of the electrodynamics of superconductors was attempted in a phenomenological theory by the brothers F. and H. London in the mid 1930s. Their theory incorporated the Meissner effect and defined a characteristic penetration depth for the magnetic field (see Sec. 1.2.2).

Another phenomenological description of the superconducting state was proposed by C. J. Gorter and H. B. G. Casimir in 1934. Their two-fluid model assumed that "superconducting" electrons that are responsible for the infinite conductivity and "normal" electrons that undergo scattering are present in the superconductor at the same time, with the fraction of superconducting electrons adjusting itself for any temperature and field so that the free energy is minimized, assuming the limiting values of 1 at zero temperature (only superconducting electrons) and 0 at the transition temperature (no superconducting electrons).

The first measurements of the magnetic penetration depth were performed by D. Shoenberg in 1940, who investigated the temperature dependence of the magnetic susceptibility of colloidal suspensions of mercury spheres.

On the basis of measurements of the surface resistance of tin and mercury resonators at microwave frequencies, A. B. Pippard developed the concept of a minimum coherence length for the spatial variation of the parameters characterizing the superconducting properties (see Sec. 1.3.3.1).

The year 1950 brought the publication of a comprehensive phenomenological theory by V. L. Ginzburg and L. D. Landau (GL), which modeled the free-energy density of a superconductor as a functional of a superconducting order parameter and was able to describe successfully both type I and type II superconductivity, with a characteristic quantity, the Ginzburg–Landau parameter, distinguishing between these two cases (see Sec. 2.2).

Also in 1950, the dependence of the superconducting transition temperature on the atomic mass of the superconducting element, the isotope effect that Kammerlingh Onnes had not been able to detect in the 1920s, was verified experimentally. Almost in parallel, H. Fröhlich had developed a theory that predicted the very same effect using the assumption of an indirect interaction between electrons mediated by lattice vibrations (phonons). Although this theory failed to explain the Meissner effect, its concept of an electron–phonon interaction was developed further by J. Bardeen, L. N. Cooper and J. R. Schrieffer (BCS), who published their microscopic theory of superconductivity in 1957. It predicted the formation of electron pairs of spin zero (Cooper pairs) as the consequence of an attractive electron interaction by virtue of the exchange of virtual phonons (see Sec. 2.1). The BCS theory successfully reproduced the crucial experimental facts such as the existence of an energy gap in the electronic excitation spectrum, the Meissner effect, and the isotope effect.

Also in 1957, A. A. Abrikosov predicted that the equilibrium flux configuration in the mixed state of type II superconductors is characterized by the penetration of the magnetic field into the superconducting material in the form of flux lines or Abrikosov vortices (see Sec. 1.3.3.2).

On the materials side, more superconductors with higher transition temperatures and critical fields were discovered in the 1950s, many of them by B. T. Matthias and J. K. Hulm. An empirical rule established by Matthias correlated high transition temperatures (and high electronic specific-heat coefficients) with peaks in the average ratio of valence electrons to the number of atoms.

In the early 1960s, the technological development was characterized by the construction of superconducting, iron-free solenoids from intermetallic materials such as Nb_3Sn and Nb-Ti that not only possessed extremely high critical fields but were also able to sustain large electrical currents in the presence of these fields.

In 1962, B. D. Josephson predicted the existence of a constant dc current across a junction where two superconductors are separated by a thin insulating or normal barrier. This current is due to the tunneling of electron pairs across the barrier and exhibits a characteristic modulation as a function of the magnetic flux threading the junction. Furthermore, in the presence of a potential difference across the barrier, there should be an alternating current flowing whose frequency is determined by the potential difference. These dc and ac Josephson effects were verified experimentally shortly thereafter and soon utilized in extremely sensitive magnetic-field sensors and other applications (see Sec. 4.2).

In 1967, magnetic-flux lines were imaged for the first time using a magnetic decoration technique (see Sec. 1.3.3.2), confirming the existence of the Abrikosov vortex lattice.

By 1973, the record value of the superconducting transition temperature T_c was 23.2 K for the intermetallic compound Nb_3Ge. It remained there for 13 more years, with many researchers believing that an upper limit had been reached that could not be surpassed. In early 1986, however, J. G. Bednorz and K. A. Müller discovered superconductivity at temperatures as high as 30 K in an oxide system, Ba-La-Cu-O, triggering an immense worldwide research effort that resulted in the discovery of ever new superconducting compounds with increasingly higher transition temperatures. The significance of this development is illustrated in Fig. 1, where the maximum superconducting transition temperature T_c is plotted as a function of time.

After this breakthrough, all the materials known before with transition temperatures below ~30 K are classified as low-temperature superconductors, as opposed to the new high-temperature superconductors that are the subject of a separate article (see SUPERCONDUCTIVITY, HIGH-TEMPERATURE).

1. BASIC PROPERTIES OF SUPERCONDUCTORS

1.1 Zero Resistivity (Infinite Conductivity)

When a superconductor is cooled below a characteristic temperature, called the superconducting transition temperature or critical temperature T_c, its dc electrical resistance R becomes immeasurably small and is assumed to be zero. Superconductors are therefore considered to be ideal conductors; their conductance G is infinite. Thus, the time constant $\tau = L/R$ for the decay of an electrical current I circulating in a ring made of superconducting material with self-inductance L is infinite, and such a current is termed a persistent current or supercurrent. From time constants that were found to be of the order of years, changes of up to 18 orders of magnitude in the resistance of samples undergoing the superconducting transition were inferred. This transition can be extremely sharp (a few microkelvins) in very clean samples and broadens with increasing density of impurities and/or structural defects (see Fig. 2). Values for T_c range from a few microkelvins to more than 150 K (see Sec. 3), spanning many decades of the thermal energy k_BT_c ($k_B = 1.38054 \times 10^{-23}$ J K^{-1} is the Boltzmann constant).

Superconductors may exhibit a finite ac resistance at frequencies that are higher than about $2\Delta(T)/\hbar$, where $2\Delta(T)$ denotes the superconducting energy gap (see Sec. 1.5) and $\hbar = 1.0545 \times 10^{-34}$ J s is Planck's constant.

1.2 Ideal Diamagnetism

1.2.1 Meissner Effect Superconductors are not only perfect conductors; they are also perfect diamagnets. In an external magnetic

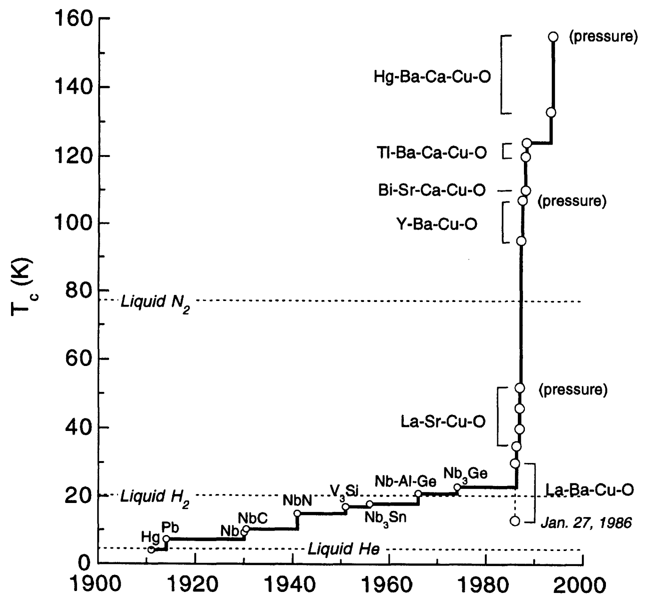

FIG. 1. Development of the maximum superconducting transition temperature over time.

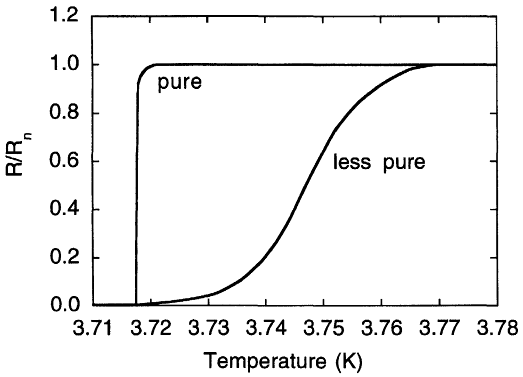

FIG. 2. Temperature dependence of the zero-field electrical resistance $R(T)$, normalized with respect to its normal-state value R_n, for a clean and a dirty superconductor (here: single-crystalline and polycrystalline tin). The resistive transition broadens with increasing number of impurities and/or defects.

field H_a that is smaller than a characteristic field H_c (see Sec. 1.3), the interior of a superconductor remains field free, i.e., any magnetic flux Φ is expelled, and hence the magnetic-flux density B is zero throughout the sample.

In principle, two different paths in the H-T phase space that bring the sample into a state with $H_a < H_c$ and $T < T_c$ can be followed (see Fig. 3). The sample is either cooled down to $T < T_c$ in zero applied field, and then $H_a < H_c$ is switched on (zero-field cooling, ZFC), or the cooldown process takes place with the external field H_a applied (field cooling, FC). In the first case, the superconductor is indistinguishable from a perfect conductor: According to Lenz's law, upon switching on the external field, supercurrents are induced in the sample that compensate the change of the magnetic flux through the sample, producing a diamagnetic magnetization M that cancels the applied field, so that $B = \mu_0(H_a + M) \equiv 0$ everywhere in the sample. Here, $\mu_0 = 4\pi \times 10^{-7}$ H/m (V s A^{-1} m^{-1}) is the vacuum permeability, and the vector character of **B**, **H**, and **M** is neglected, as are demagnetization effects due to the sample geometry (see Sec. 1.3.2). In the FC case, however, the perfect conductor, once penetrated by the external field at $T > T_c$, would remain in that state upon cooling, since there is no change in the magnetic-flux density that would induce screening currents, whereas the superconductor expels the flux from its interior upon cooling through the superconducting transition.

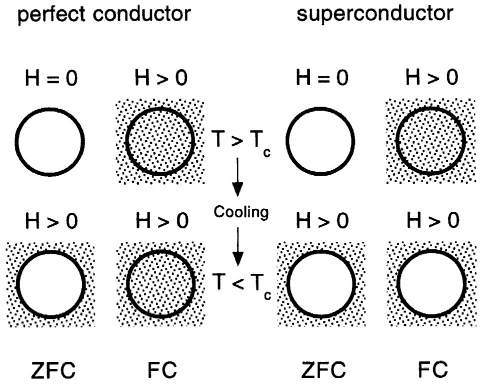

FIG. 3. Response of a perfect conductor (left) and of a superconductor (right) to an external magnetic field. Upon zero-field cooling (ZFC), both samples screen the external field. Under field cooling (FC), however, the field remains inside the perfect conductor, whereas it is expelled from the superconductor (Meissner effect).

The fact that $B = \mu_0(H_a + M) = \mu_0(1 + \chi)H_a \equiv 0$ in a superconductor implies that the volume susceptibility χ equals -1 (ideal diamagnetism). This fundamental property of superconductors that establishes the superconducting phase as a true thermodynamic equilibrium state (see Sec. 1.3.1) is named the Meissner effect. It provides the conclusive test to experimentally distinguish a superconductor from a perfect (or very, very good) conductor.

1.2.2 Magnetic Penetration Depth

The screening or shielding currents that create a magnetization canceling the external field and keeping the superconductor's interior field free circulate only in a thin surface layer. This fact is described by the phenomenological equation

$$\nabla^2 \mathbf{B} = (1/\lambda^2)\mathbf{B},$$

which incorporates the condition that $\mathbf{B} \equiv 0$ throughout the superconductor and, together with the Maxwell equation $\nabla \times \mathbf{B} = \mu_0 \mathbf{j}$, permits the calculation of the distribution of the magnetic-flux density **B** and the supercurrent density **j** in the sample. It yields the magnetic penetration depth λ as the characteristic length scale over which fields and currents decay in the interior of the superconductor. Typical values for λ for low-T_c superconductors—extrapolated to $T = 0$—are of the order of 10 nm, so that most macroscopic samples with dimensions much larger than λ can be considered field free. However, this may not be true for certain sample geometries (see Sec. 1.3.2).

The penetration depth λ varies for different superconducting materials, often changes with the orientation of the external magnetic field with respect to the sample, and depends on temperature, saturating at low temperatures with $\lambda(T = 0) \equiv \lambda_0$ and rapidly increasing while approaching T_c. For many superconductors, this dependence is described by the empirical relation

$$\lambda(T)^2 = \lambda_0^2 \frac{T_c^4}{T_c^4 - T^4}.$$

The magnetic penetration depth λ is an important physical quantity characterizing a specific superconductor. Among the methods employed to determine λ are the measurement of changes in the self or mutual inductance of the superconducting components of rf circuits, in the magnetic volume susceptibility, or in the surface impedance of superconducting films at microwave frequencies, as well as the measurement of the depolarization rate of spin-polarized muons, which provides information about the local magnetic field inside the superconductor.

1.3 Critical Magnetic Fields

1.3.1 Type I Superconductivity The superconducting state is strongly influenced by the presence of a magnetic field: For fields higher than a characteristic temperature-dependent field, the critical field H_c, superconductivity is suppressed, and the sample becomes normal conducting. The shielding supercurrents break down, and the external magnetic field completely penetrates the sample. A material characterized by this behavior is called a type I superconductor, and many of the metallic elements belong to this class (see Sec. 3). The critical field decreases with increasing temperature, typically following the empirical relation

$$H_c(T) = H_c(0)[1 - T^2/T_c^2].$$

The extrapolated zero-temperature critical field $H_c \equiv H_c(0)$ is of the order of 10^4–10^5 A m^{-1} ($\sim 10^2$–10^3 Oe) for typical type I superconductors (see Sec. 3).

The $H_c(T)$ line separates the superconducting and normal states in the H-T phase plane. Since the difference in the Gibbs free-energy density $\Delta g \equiv g_n(T) - g_s(T)$ between these two phases is equivalent to the magnetic energy density,

$$\Delta g(T) = -\mu_0 \int_0^{H_c(T)} M(H,T)dH = \frac{\mu_0}{2} H_c^2(T),$$

knowledge of the $H_c(T)$ curve allows the determination of important thermodynamic quantities such as the entropy change, the specific-heat jump, and the latent heat associated with the phase transition. Therefore, H_c is also referred to as the thermodynamic critical field. In zero magnetic field, the phase transition is of second order, whereas in the presence of a magnetic field, it becomes a first-order transition.

1.3.2 Intermediate State The simple picture of magnetic flux uniformly penetrating a type I superconductor in fields higher than the critical field H_c is strictly valid only for samples whose dimension parallel to the external field is infinite (long, thin slabs with the field along the axis or thin plates with the field in the plane are good approximations). If the magnetization of the sample has components perpendicular to the direction of the external field, surface magnetic charges or free magnetic poles create stray fields known as "demagnetizing" fields. If the sample is homogeneously magnetized (which is the case for ellipsoidal sample geometries), the demagnetizing field H_d is proportional to the magnetization M and given by $H_d = NM$, with the demagnetization factor N assuming its maximum value of unity for the case of an infinite plate perpendicular to the field and its minimum value of zero when the field is along the axis of a long, thin slab or in the plane of a thin plate (again, the vector character of **H** and **M** as well as the tensor character of **N** and χ are neglected here). Then, the effective field experienced by the sample can be expressed as

$$H_{\text{eff}} = H_a - NM = H_a - N\chi H_{\text{eff}},$$

or

$$H_{\text{eff}} = \frac{1}{1 + N\chi} H_a.$$

In superconductors, $\chi < 0$, so that $H_{\text{eff}} > H_a$. Since the tangential field component has to be continuous, this implies local field enhancement, e.g., at the equator of spherical samples. Thus, there exists a field interval in which the external field is larger than H_c on some points at the sample surface and, at the same time, lower than H_c on other points. In this case, it is energetically most favorable for the superconductor to split up into superconducting and normal domains, and the sample is said to be in the intermediate state. The size of the normal domains can be of the order of millimeters and increases with field. Demagnetization effects due to the sample geometry can completely dominate the magnetic response of a superconductor, especially in the case of thin platelets oriented perpendicular to the field.

1.3.3 Type II Superconductivity Below a characteristic, temperature-dependent field called the lower critical field $H_{c_1}(T)$, a type II superconductor is completely equivalent to a type I superconductor for $H < H_c$: $B = 0$, $M = -H$; the sample is considered to be in the so-called Meissner state. If the applied field exceeds another characteristic field, the upper critical field $H_{c_2}(T)$, the sample enters the normal state, and magnetic flux completely penetrates its interior. In the intermediate field interval $H_{c_1} < H < H_{c_2}$, however, a new phase develops that is characterized by the partial entry of magnetic flux into the sample and is called mixed state or Shubnikov phase. Schematic H-T phase diagrams for a type I and a type II superconductor are shown in Fig. 4.

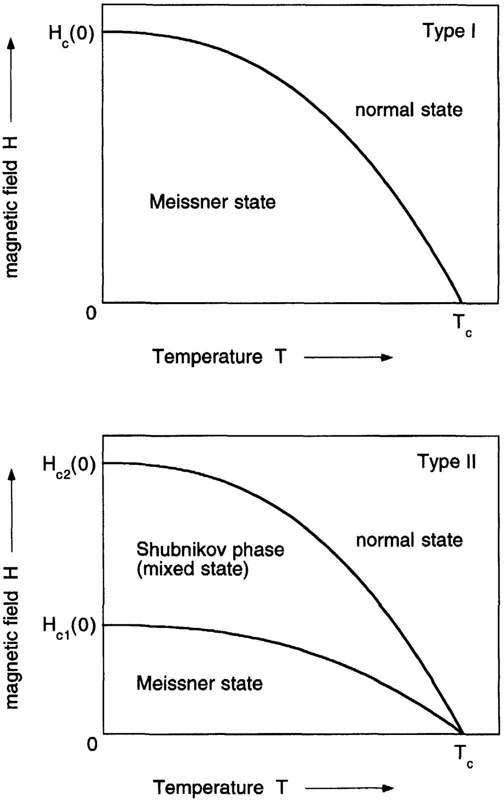

FIG. 4. Schematic illustration of the *H-T* phase diagram for type I and type II superconductors.

1.3.3.1 Superconducting Coherence Length. Aside from the penetration depth λ, which describes the distance over which screening supercurrents and magnetic fields decay inside a superconductor, there is another important length scale characterizing the spatial extent of the attractive correlation between the electrons that gives rise to superconductivity (see Sec. 2). This coherence length ξ, which is of the order of tens to thousands of nanometers for low-T_c superconductors (see Sec. 3) also represents a typical distance over which changes in the density of the superconducting charge carriers occur.

The ratio of the two characteristic lengths $\kappa = \lambda/\xi$ is called the Ginzburg–Landau parameter of the superconductor. Materials with $\kappa < 1/\sqrt{2}$ are type I superconductors, whereas $\kappa > 1/\sqrt{2}$ describes type II materials. Figure 5 illustrates these two situations and demonstrates that in the latter case, a negative surface-energy density may arise. It is due to this negative surface energy that a type II superconductor splits into superconducting and normal regions, the latter being penetrated by the external magnetic field.

1.3.3.2 Abrikosov Vortex Lattice. The energetically most favorable configuration of superconducting and normal domains in the mixed state corresponds to the penetration of magnetic flux in the form of so-called Abrikosov vortices or flux lines. These vortices are comprised of a normal "core" region

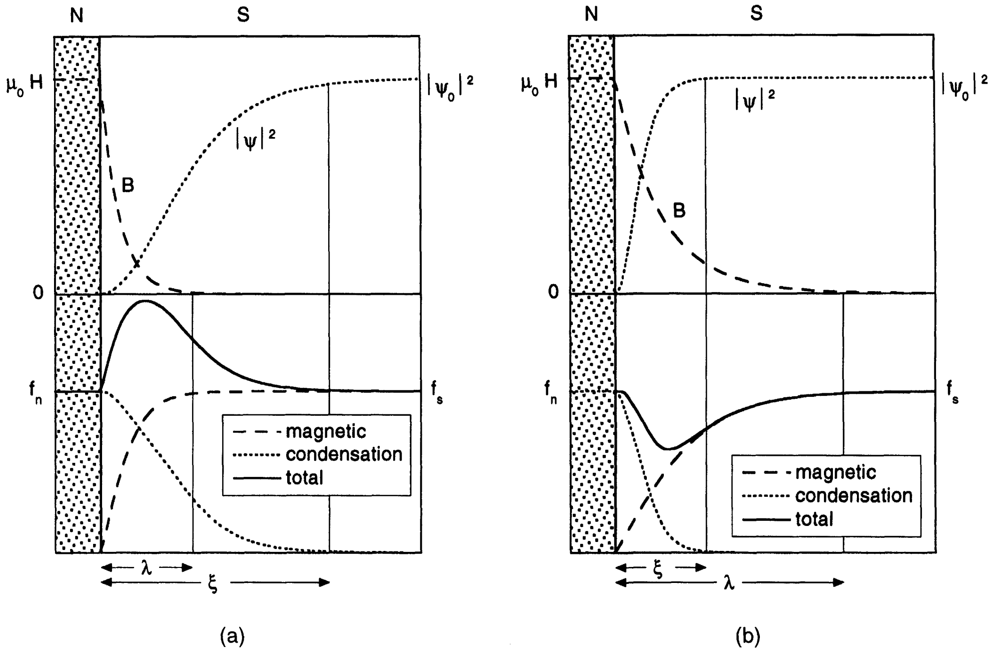

FIG. 5. Schematic illustration of the spatial variation of the magnetic flux density B and Cooper-pair number density $|\psi|^2$ at the interface between a normal metal N and (a) a type I and (b) a type II superconductor S. Also shown are schematic representations of the contributions to the free-energy density due to magnetic and condensation energy, and the resulting positive and negative surface energies.

of radius $\sim\xi$ parallel to the external field, where the density of the superconducting charge carriers is strongly reduced, and supercurrents circulating around the core, which screen the field over a distance λ. In clean (i.e., defect-free) type II superconductors, the vortices arrange themselves in a regular hexagonal pattern known as the Abrikosov vortex lattice, as illustrated in Fig. 6. Experimentally, vortices are imaged using magnetic decoration techniques where fine magnetic particles that accumulate at the regions of high-flux gradients on the sample surface corresponding to the vortex positions are observed by electron microscopy. Other methods include small-angle neutron scattering, scanning probe microscopy with miniature Hall probes as the magnetic sensor, and magneto-optical techniques (the latter can only resolve bundles of flux lines but are capable of imaging their dynamics).

1.3.3.3 Magnetic Fluxoid. It is not the magnetic flux threading a superconducting loop that is conserved by supercurrents counterbalancing flux changes, but rather the so-called fluxoid

$$\Phi \equiv \iint_S \mathbf{B}\cdot d\mathbf{s} + \mu_0\lambda^2 \oint_L \mathbf{j}\cdot d\mathbf{l},$$

where the first term represents the magnetic flux through the loop cross section S and the

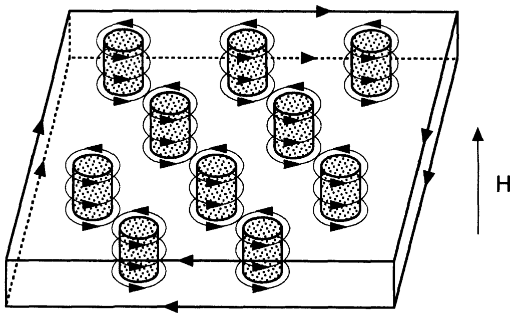

FIG. 6. Schematic illustration of the penetration of an external magnetic field H into an ideal type II superconductor in the form of hexagonally arranged flux lines or vortices, forming the Abrikosov flux lattice. Screening currents flowing on the sample periphery create a diamagnetic magnetization, whereas the vortices can be imagined to be surrounded by opposite supercurrents.

second term is the line integral of the supercurrent density around the loop L. The fluxoid is quantized in units of the magnetic-flux quantum $\Phi_0 = h/2e = 2.07 \times 10^{-15}$ Wb (V s), where $e = 1.602 \times 10^{-19}$ C (A s) is the elementary charge. Flux quantization is a consequence of the requirement that the macroscopic superconducting wave function (see Sec. 2) is single valued, so that its phase can only change by integer multiples of 2π upon circumfering the loop on a path well inside the superconductor where the field is zero.

In the mixed state, each flux line carries one flux quantum.

1.4 Critical Current Density

The superconducting properties of a conductor carrying a transport electrical current I with current density j are suppressed when j exceeds a characteristic value, the critical current density j_c. Aside from the transition temperature $T_c(H)$, the temperature- and field-dependent critical current density $j_c(T,H)$ is that parameter characterizing a superconductor that is most relevant for large-scale technological applications (see Sec. 4), since it marks the onset of dissipation in the form of resistive losses.

For type I superconductors, $j_c > 0$ for $T < T_c$ and $H < H_c$. In this case, a finite resistance is observed whenever the total magnetic field corresponding to the superposition of the applied field H and the field generated by the transport current exceeds the critical field H_c anywhere at the surface of the sample. For $H = 0$, the critical current will be that current that creates the critical field H_c at the conductor's surface (Silsbee's rule).

For type II materials, a modification of Silsbee's rule applies to the Meissner state ($H < H_{c1}$), when H_{c1} is substituted for H_c. In the mixed state, however, j_c is not necessarily nonzero for $T < T_c$ and $H < H_{c2}$. For an ideal sample, where the interacting flux lines form an undisturbed Abrikosov lattice, an arbitrarily small force acting on the vortices (such as arising from a transport current or from a changing field) induces vortex motion and, therefore, dissipation. For this reason, in clean type II superconductors, j_c is very small in the mixed state. On the other hand, if there exist regions inside the superconductor where the superconducting order parameter (see Sec. 2.2) and the density of superconducting charge carriers are reduced locally, the presence of a flux line's normal core at such a location is energetically favored compared to its presence in a fully superconducting region because, in the latter case, the energy necessary to suppress superconductivity in that region is higher than in the former case. Therefore, such regions represent pinning centers for the flux lines that prevent vortex motion, and the associated pinning forces determine the temperature and field dependences of the critical current density. These forces (or the respective pinning potentials) depend on the nature of the defects, especially on their size. High pinning forces have been achieved in various alloys, resulting in superconducting materials with very good current-carrying capabilities in high magnetic fields (see Secs. 3 and 4). These materials are referred to as "hard" type II superconductors.

1.5 Energy Gap

In the framework of the microscopic theory for low-T_c superconductivity (see Sec. 2.1), the superconducting charge carriers are Cooper pairs of electrons that form as a result of an attractive interaction between electrons that is mediated by lattice vibrations (phonons). Because the binding or condensation energy of these pairs has to be overcome in order to break them apart, there is a temperature-dependent superconducting energy gap $2\Delta(T)$ in the electronic excitation spectrum of these superconductors. The temperature dependence of the energy gap $\Delta(T)$ close to T_c is predicted by the microscopic theory (see Sec. 2.1) as

$$\Delta(T) \propto T_c(1 - T/T_c)^{1/2}.$$

The existence of the energy gap manifests itself in a variety of experiments.

1.5.1 Specific Heat The low-temperature specific heat $C(T)$ of a metal can be described by the sum of a lattice contribution $C_1(T)$ and an electronic contribution $C_e(T)$. The lattice term $C_1 \propto T^3$ does not change when the metal undergoes a superconducting transition. The electronic contribution, however, which is proportional to temperature for a normal metal, $C_e^{(n)}(T) \propto T$, is found to vanish exponentially in the superconducting state: $C_e^{(s)}(T) \propto \exp(-\Delta/k_BT)$ (see Fig. 7). Such a dependence is characteristic for systems with an energy gap. The measurement of the normal-state specific heat can be performed after application of an external magnetic field that is larger than the critical field.

1.5.2 Electromagnetic and Acoustic Absorption As is the case in other materials with an energy gap, such as insulators and semiconductors, in superconductors the gap manifests itself in the optical absorption spectra as an absorption edge at the frequency $\omega = 2\Delta/\hbar$ since the photon energy has to exceed the energy 2Δ necessary to break up a Cooper pair. These frequencies are of the order of 10^{10}–10^{12} Hz (far infrared), which corresponds to the frequency interval where the ac electrical resistance becomes finite (see Sec. 1.1).

The energy of the low-frequency ultrasonic phonons employed in ultrasonic attenuation experiments is too small to break up the Cooper pairs; instead, these phonons couple to low-energy excitations already present, the number of which is proportional to $\exp(-\Delta/k_BT)$. The ratio of the ultrasonic attenuation coefficients α in the superconducting and normal states, respectively, is given by

$$\frac{\alpha_s}{\alpha_n} = \frac{2}{\exp(\Delta/k_BT) + 1}.$$

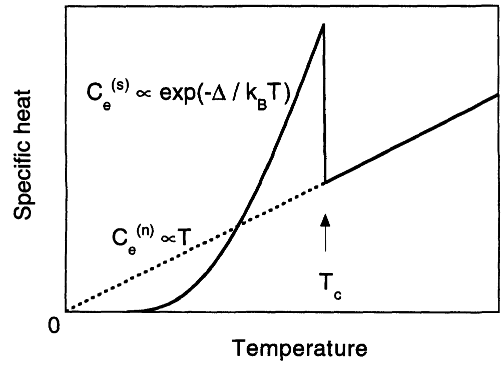

FIG. 7. Temperature dependence of the low-temperature electronic specific heat $C_e(T)$ for a metal in the normal (dotted line) and superconducting (solid curve) states. In the normal state, the electronic contribution to the specific heat is proportional to temperature, $C_e^{(n)} \propto T$, whereas the electronic specific heat vanishes exponentially in the superconducting state: $C_e^{(s)} \propto \exp(-\Delta/k_BT)$. The lattice contribution to the specific heat is negligible at low temperatures.

Measurements of the ultrasonic attenuation have confirmed the predictions of the BCS theory (see Sec. 2.1) for the temperature dependence of the energy gap and its correlation with the critical temperature.

Low-frequency electromagnetic absorption is probed by nuclear magnetic resonance (NMR) experiments that determine the spin–lattice relaxation rate T_1^{-1}. In analogy to the ultrasonic attenuation experiments, only scattering of quasiparticles can contribute to the scattering for $h\nu \ll \Delta(T)$, where ν is the precession frequency of the nuclear spins. Thus, for $T \ll T_c$, T_1^{-1} decreases exponentially with decreasing temperature. (For temperatures slightly below T_c, the relaxation rate increases, which is predicted correctly by the BCS theory.)

1.5.3 Electron Tunneling When a normal metal and a superconductor are separated by a thin insulating barrier (for instance, an oxide layer with a thickness of a few nanometers), electrons can cross this barrier by the quantum mechanical tunnel effect. If the electrons are bound in Cooper pairs with a binding energy of 2Δ per pair, an energy Δ has to be supplied to cause an electron to break free and traverse the barrier. This becomes apparent in the current–voltage (I-V) characteristics that are illustrated in Fig. 8. When the superconductor is in the normal state, $I \propto V$ (Ohmic behavior). In the superconducting state, on the other hand, essentially no current flows until the voltage reaches the value $V = \Delta/e$ (for $T = 0$, the current is zero; at finite temperatures, there are always individual electrons that are not condensed into Cooper pairs and can tunnel under the influence of an arbitrarily small potential difference, thus giving rise to a background current). The gap is preferably estimated from the maximum in the differential conductivity dI/dV.

1.6 Josephson Effects

Two superconductors separated by a thin region in which the Cooper-pair density is locally reduced form a weak link or Josephson junction. It can be realized either as a tunnel junction, where an insulating or nor-

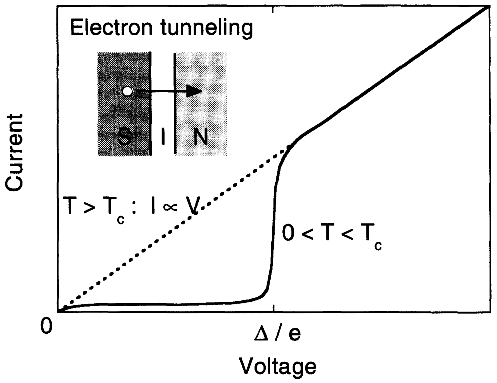

FIG. 8. Current–voltage (*I*-*V*) characteristics of a superconductor–insulator–normal metal (SIN) tunnel junction. With the superconductor in the normal state, the junction exhibits Ohmic behavior, $I \propto V$, whereas in the superconducting state, a voltage of the order of Δ/e is necessary to excite an electron across the energy gap.

mal conducting layer with a thickness of typically less than 20 Å is sandwiched between superconductors, as a point-contact junction between a flat superconducting surface and a superconducting tip, or as a bridge junction in the form of a thin constriction within a superconductor. In addition to the single-electron tunneling current (see Sec. 1.5.3) flowing across a Josephson junction, there may be a supercurrent of Cooper pairs traversing the barrier.

This phenomenon is known as the dc Josephson effect. In the absence of a magnetic field, the current density j_J associated with the Josephson current is determined by the phase difference $\Delta\phi$ of the macroscopic superconducting wave functions of the two superconductors:

$$j_\mathrm{J} = j_c \sin \Delta\phi,$$

where j_c is the critical current density of the junction, which depends on the value of the energy gap of the superconductors and on the normal-state resistivity of the weak link. Its value is much smaller than the typical bulk critical current density.

The Josephson current density j_J exhibits a dramatic dependence on the magnetic field through the junction, varying sinusoidally along the coordinate perpendicular to the field. The magnetic-field (or rather magnetic-flux) dependence of the maximum average current density j_max across the entire junction is given by

$$j_\mathrm{max} = j_c \left| \frac{\sin(\pi\Phi/\Phi_0)}{\pi\Phi/\Phi_0} \right|,$$

which implies that the Josephson current vanishes whenever the external field corresponds to a magnetic flux Φ through the junction that is equal to an integer number of flux quanta, $n\Phi_0$. This field dependence is the basis of the importance of Josephson junctions for magnetic-field–sensing applications (see Sec. 4.2).

If a constant electric potential difference is present across the weak link (e.g., through the application of a dc voltage V, or because the current passing through the junction is higher than the critical current), an oscillating supercurrent will appear,

$$j(t) = j_c \sin\left(\Delta\Phi_0 + \frac{2eV}{\hbar} t\right),$$

where the Josephson frequency $\omega_\mathrm{J} = 2eV/\hbar$ of this alternating current is of the order of 10^{11} Hz for voltages in the millivolt range. This phenomenon is referred to as the ac Josephson effect. If a junction biased with a constant current to a voltage V is subjected to microwave irradiation of frequency ω_0, the oscillating supercurrent contains a stationary component whenever $\omega_\mathrm{J} = n\omega_0$. This leads to the appearance of constant-voltage steps in the *I*-*V* characteristics of the junction at finite voltage values $V_n = n\omega_0\hbar/2e$. These features are known as Shapiro steps. Applications of the ac Josephson effect are described in Sec. 4.2.

1.7 Isotope Effect

The attractive electron interaction that leads to the formation of Cooper pairs is mediated by exchange of virtual lattice vibrations or phonons (see Sec. 2.1.2). The superconducting properties are, therefore, influenced by changes in the masses of the constituents of the crystal lattice. This becomes apparent in the variations of the transition temperature T_c when one isotope of a superconducting element is replaced by another isotope:

$$T_c \propto M^{-\alpha},$$

where M is the atomic weight and $\alpha = \frac{1}{2}$. This expression describes the so-called isotope effect. It has been verified experimentally for a variety of materials. Deviations of the value of α from $\frac{1}{2}$ can be explained by modifications of the BCS theory taking into account electron–electron Coulomb interaction.

2. THEORY OF SUPERCONDUCTIVITY

2.1 BCS Theory

The microscopic theory of superconductivity was developed by J. Bardeen, L. N. Cooper and J. R. Schrieffer and is known as the BCS theory.

2.1.1 Cooper Pairs In the BCS theory, it was shown that the Fermi-sea ground state is, in the presence of a weak attractive interaction between the electrons, unstable with respect to the formation of bound pairs of electrons occupying states with equal and opposite momentum and spin; i.e. $(\mathbf{k}\uparrow; -\mathbf{k}\downarrow)$. The spatial extent of the superconducting electron pairs, commonly referred to as Cooper pairs, determines the superconducting coherence length $\xi \approx hv_F/k_BT_c$, where v_F is the Fermi velocity. According to the theory, the binding energy, which corresponds to the minimum energy required to break a pair, creating two quasiparticle excitations, is given by $E_g(T) = 2\Delta(T)$, where $\Delta(T)$ is the temperature-dependent energy gap in the quasiparticle excitation spectrum. The plot of the energy gap as a function of temperature in Fig. 9 illustrates that Δ vanishes at T_c and achieves its maximum value $\Delta(0) = 1.764k_BT_c$ at $T = 0$. The energy gap is manifested in various physical properties such as electron-tunneling spectra, ultrasonic attenuation, nuclear relaxation, electromagnetic absorption, and specific heat (see Sec. 1.5).

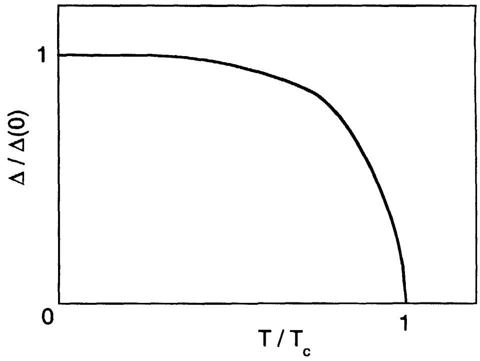

FIG. 9. Temperature dependence of the energy gap $\Delta(T)$ as given by the BCS theory.

2.1.2 Electron-Pairing Mechanism In conventional superconductors, the weak attractive interaction that binds the two electrons of a Cooper pair together, overcoming the strong Coulomb repulsion between them, is mediated by the exchange of virtual quantized lattice vibrations or phonons, which ordinarily scatter the electrons in the normal state. The way the attractive electron–phonon interaction works can be visualized in the following way: A negatively charged conduction electron passing in the vicinity of a positively charged ion exerts an attractive force on the ion and displaces it from its equilibrium lattice position, producing a local net positive charge that attracts a second negatively charged electron traveling through that region. This results in a weak attractive interaction between the two electrons that can overcome the Coulomb repulsion because the electrons move much faster than the ions (i.e., the lattice distortion persists for a relatively long time compared to the transit time of the electron that produced it).

The BCS theory successfully accounts for the properties of the conventional superconductors and provides an expression for the transition temperature T_c in terms of the properties of the normal state, given by

$$T_c \approx \Theta_D \exp\left[\frac{1}{\mathcal{N}(E_F)\mathcal{V}}\right],$$

where Θ_D is the Debye temperature, $\mathcal{N}(E_F)$ is the density of electronic states at the Fermi level E_F, and $\mathcal{V}$ is the strength of the attractive electron–phonon interaction. Since $\Theta_D \propto M^{-1/2}$ for a monatomic lattice of atoms of mass M, it follows that $T_c \propto M^{-\alpha}$ with $\alpha = \frac{1}{2}$, which provides an explanation for the isotope effect (see Sec. 1.7). This expression also accounts for the minimum in T_c observed for the transition-metal superconductors near the half-filled d-electron shell where $\mathcal{N}(E_F)$ passes through a minimum. Refinements of the BCS theory that include the effect of the Coulomb repulsion between the electrons are capable of accounting for the reduced values of α found for transition-metal element superconductors (e.g., $\alpha = 0.37$ for Mo, 0.38 for Re, and 0.21 for Os, and ~0 for Ru and Zr). An approach that has achieved some success in accounting for the values of T_c and the isotope-effect exponent α for transition-metal elements and alloys, which are in the strong-coupling limit $T_c/\Theta_D \gtrsim 0.05$, is due to McMillan. The McMillan theory gives the following approximate expressions for T_c and α:

$$T_c = \frac{\Theta_D}{1.45}\exp\left[-\frac{1.04(1+\lambda)}{\lambda - \mu^*(1+0.62\lambda)}\right]$$

and

$$\alpha = \tfrac{1}{2}\left[1 - \left(\mu^* \ln\frac{\langle\Omega\rangle}{1.2T_c}\right)^2 \frac{1+0.62\lambda}{1+\lambda}\right],$$

where λ is an electron–phonon interaction parameter that ranges from 0.1 to 1.5 in various metals, μ^* is a Coulomb repulsion parameter that is often found to be ~0.1, and $\langle\Omega\rangle$ is an average phonon frequency defined by

$$\langle\Omega\rangle = \frac{\int \alpha^2(\Omega)d\Omega}{\int \alpha^2(\Omega)F(\Omega)d\Omega/\Omega},$$

where $\alpha^2(\Omega)$ is the effective electron–phonon coupling constant and $F(\Omega)$ is the phonon density of states, which can be extracted from quasiparticle tunneling spectra and inelastic neutron-scattering experiments. The electron–phonon interaction enhances $\mathcal{N}(E_F)$ or the electron mass m^* with respect to their bare values $\mathcal{N}_0(E_F)$ and m by $\mathcal{N}(E_F)/\mathcal{N}_0(E_F) = m^*/m = 1 + \lambda$.

2.2 Ginzburg–Landau Theory

Prior to the publication of the BCS theory, V. L. Ginzburg and L. D. Landau (GL) developed a phenomenological theory based on a pseudo wave function Ψ that serves as a complex order parameter such that $|\Psi(\mathbf{r})|^2$ represents the local density of superconducting electrons $n_s(\mathbf{r})$. The wave function can be viewed as the wave function of the center of mass of the Cooper pairs. The theory is capable of treating situations in which n_s varies in space and where a nonlinear field response is strong enough to change n_s. Predictions of critical fields and of the spatial structure of $\Psi(\mathbf{r})$ in nonuniform situations have been made based on the GL theory. The wave function Ψ satisfies a differential equation that is analogous to the Schrödinger equation for a free particle, but with a nonlinear term, and is given by

$$\frac{1}{2m^*}\left(\frac{\hbar}{i}\nabla - \frac{e^*}{c}\mathbf{A}\right)^2 \Psi + \beta|\Psi|^2\Psi = -\alpha(T)\Psi.$$

The equation for the supercurrent,

$$\mathbf{J}_s = \frac{e^*\hbar}{i2m^*}(\Psi^*\nabla\Psi - \Psi\nabla\Psi^*) - \frac{e^{*2}}{m^*c}|\Psi|^2\mathbf{A},$$

is the same as the quantum-mechanical expression for a particle of charge e^* and mass m^*. For a superconductor, the relevant particle is a Cooper pair for which $e^* = 2e$ and $m^* = 2m$. One of the parameters that emerges from the GL theory is the temperature-dependent superconducting coherence length

$$\xi(T) = \hbar/|2m^*\alpha(T)|^{1/2},$$

which characterizes the distance over which $\Psi(\mathbf{r})$ can vary without appreciable energy change. L. P. Gor'kov showed that the phenomenological theory is a limiting form of the microscopic BCS theory (suitably generalized to deal with situations where n_s is varying in space), which is valid near T_c, and in which Ψ is directly proportional to the gap parameter Δ.

Near T_c, $\xi(T)$ diverges as $(T - T_c)^{-1/2}$, and α vanishes as $T - T_c$. The temperature-dependent GL coherence length ξ and penetration depth λ in the clean ($\xi_0 \gg l$) and dirty ($\xi_0 \ll l$) limits, respectively, can be expressed in terms of the BCS coherence length $\xi_0 = \hbar v_F/\pi\Delta(0)$ and the transport mean free path l:

$$\xi_{\text{clean}} = 0.74\xi_0(1 - T/T_c)^{-1/2},$$

$$\lambda_{\text{clean}} = \lambda_L(0)[2(1 - T/T_c)]^{-1/2},$$

$$\xi_{\text{dirty}} = 0.85(\xi_0 l)^{1/2}(1 - T/T_c)^{-1/2},$$

$$\lambda_{\text{dirty}} = 0.615\lambda_L(0)(\xi_0/l)^{1/2}(1 - T/T_c)^{-1/2},$$

where $\lambda_L(0) = (mc^2/4\pi|\Psi|^2e^2)^{1/2}$.

The Ginzburg–Landau parameter $\kappa = \lambda/\xi$ determines the sign of the surface energy associated with an interface between superconducting and normal material, thus distinguishing between type I and type II superconductors (see Sec. 1.3.3.1). The GL theory is capable of describing both the intermediate state for type I materials ($\kappa \ll 1$), where a domain pattern is stabilized by the positive surface energy associated with the domain walls (see Sec. 1.3.2), and the mixed state in type II systems ($\kappa \gg 1$), where the negative surface energy gives rise to flux penetration in the form of fluxoids. It is applicable to many situations including the calculation of the bulk critical fields H_{c1} and H_{c2} and the surface critical field H_{c3}, the structure of an isolated vortex and of the vortex lattice (Abrikosov vortex lattice), and superconducting fluctuations. It has been very useful in treating the properties of ferromagnets, as well as the interaction between superconductivity and magnetic order.

3. SUPERCONDUCTING MATERIALS

3.1 Elements

Most of the metallic elements that do not exhibit magnetic order become superconducting when cooled to a sufficiently low temperature. The element Nb has the highest transition temperature T_c at atmospheric pressure ($T_c = 9.3$ K), while La has the highest T_c when subjected to high pressure ($T_c =$ 11.2 K at 16 GPa compared to ~6 K at atmospheric pressure). Many of the metallic elements have been tested for superconductivity down to temperatures as low as ~1 mK and at high pressures of several tens of gigapascals and higher. Table 1 lists the transition temperatures of some of the superconducting elements. Shown in Fig. 10 is a periodic table of the elements in which the known superconducting elements are marked.

The critical temperatures of the transition-metal elements exhibit a characteristic variation with the valence-electron concentration (vec) that is characterized by pronounced maxima near vec values of 5 and 8, respectively. Here, the vec is the number of outermost electrons, including those in the d shell and the s and p shells. This characteristic dependence of T_c on the vec of the transition-metal elements and of alloys formed between neighboring elements in the periodic table is part of "Matthias's rules," named after the late pioneer in superconducting materials, B. T. Matthias. It is illustrated in Fig. 11.

Table 1. Superconducting transition temperature T_c of some elements. The greek letters indicate different modifications for polymorphic elements; the corresponding crystal structure is indicated.

Element	T_c (K)	Structure
Nb	9.2	Cubic
Pb	7.2	Cubic
La (α)	6.0	Cubic
V	5.3	Cubic
La (β)	4.9	Hexagonal
Ta	4.6	Cubic
Hg (α)	4.2	Rhombohedral
Hg (β)	4.0	Tetragonal
Sn (β)	3.7	Tetragonal
In	3.4	Tetragonal
Tl	2.4	Cubic
Re	1.7	Hexagonal
Th (α)	1.4	Cubic
Al	1.2	Cubic
Ga	1.1	Orthorhombic
Mo	0.92	Cubic
Zn	0.87	Hexagonal
Os	0.66	Hexagonal
Cd	0.54	Hexagonal
Ru	0.49	Hexagonal
Ti (α)	0.39	Hexagonal
Ir	0.10	Cubic
W	0.012	Cubic

Some elements, such as Be, Bi, and Ga, become superconducting only in the form of thin films after being vacuum deposited at temperatures below ~10 K ("collapsed" phases). After a temperature increase, they undergo an irreversible transition into a nonsuperconducting phase.

3.2 Binary Alloys, *A*15 Compounds

Many of the compounds formed by alloying transition-metal elements and other metallic elements exhibit variations in T_c with the vec that conforms to Matthias's rule. One particularly important group included in this category consists of the *A*15 compounds, the materials that exhibited the highest values of T_c prior to the discovery of layered high-T_c cuprate compounds. The *A*15 materials are

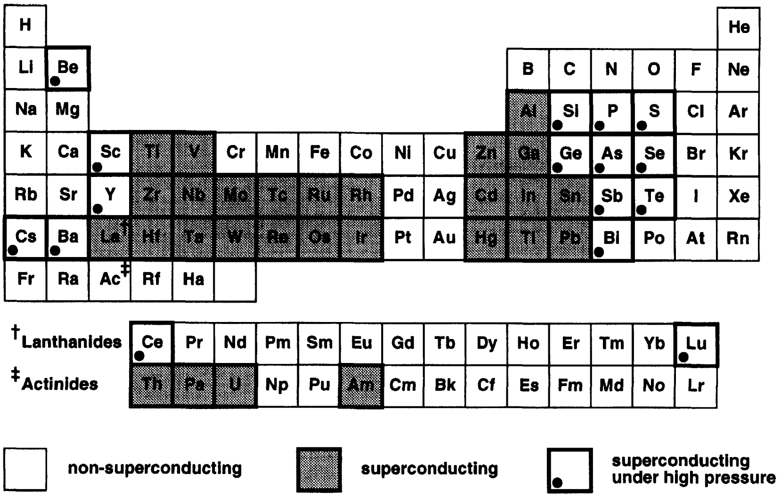
H He
Li Be B C N O F Ne
Na Mg Al Si P S Cl Ar
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I Xe
Cs Ba La† Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi Po At Rn
Fr Ra Ac‡ Rf Ha
†Lanthanides Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb Lu
‡Actinides Th Pa U Np Pu Am Cm Bk Cf Es Fm Md No Lr
non-superconducting
superconducting
superconducting under high pressure

FIG. 10. Location of the superconducting elements in the periodic table.

described by the general chemical formula A_3B, where A is Nb or V and B is a non-transition-metal element such as Al, Ge, Ga, Sn, or Bi; they crystallize in a cubic crystal structure. The unit cell consists of a non-transition-metal element in a body-centered cubic lattice with the Nb or V atoms forming nonintersecting chains closely spaced on each cubic face as shown in Fig. 12.

Table 2 gives T_c values for some $A15$ compounds. Many of the materials belonging to this group exhibit very high upper critical fields H_{c2}, which can reach values in the range of 2×10^7 A m^{-1} (240 kOe) for Nb_3Sn

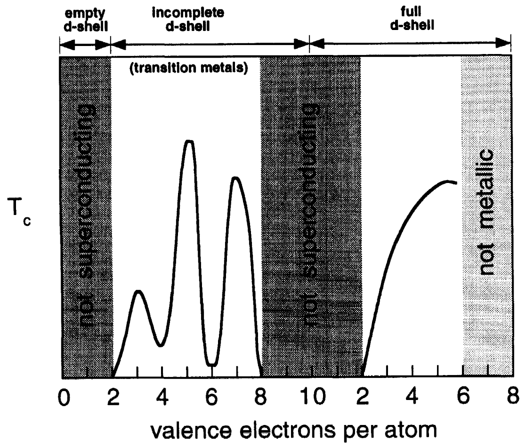

FIG. 11. Matthias's empirical rule for the superconducting transition temperature T_c throughout the periodic system.

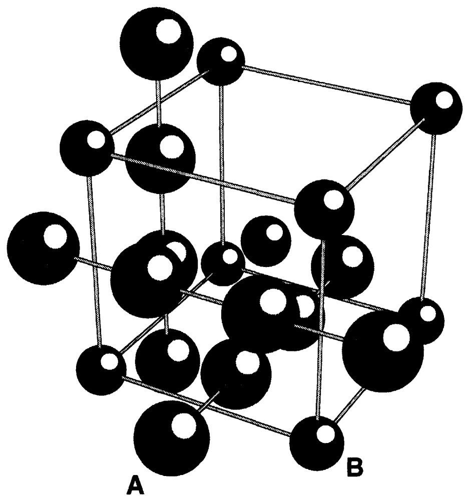

FIG. 12. Unit cell of an $A15$ compound with the general chemical formula A_3B, where A is Nb or V and B is a non-transition-metal element such as Al, Ge, Ga, Sn, or Bi.

Table 2. Superconducting transition temperature T_c of some $A15$ compounds.

Compound	T_c (K)
Nb_3Ge	23
$Nb_3(Al,Ge)$	22
Nb_3Al	21
Nb_3Sn	18
V_3Si	17

to 3×10^7 A m^{-1} (380 kOe) for Nb_3Ge. This is the basis for the application of some *A*15 compounds in the fabrication of wires for the production of superconducting electromagnets and in other applications of superconductivity (see Sec. 4.1). Other superconducting materials with relatively high T_c's that are currently used in technological applications include Nb-Ti and Nb-Zr alloys, which have T_c values of ~12 K. Another interesting compound with $T_c \sim 15$ K is NbN, which has the *B*-1 (NaCl) structure. A large number of superconductors crystallize in this structure. The superconductivity in these materials is rather resistive to disorder, in contrast to the *A*15 compounds whose superconducting properties are very sensitive to disorder. A particularly interesting *B*-1 superconductor is PdH, which can be synthesized by low-temperature ion implantation. The T_c of this material reaches 17 K when it is alloyed with Ag.

3.3 Ternary Compounds

An important class of ternary compounds has crystal structures that consist of molecular units or clusters of atoms within which the superconducting charge carriers are confined. This produces large values of the electronic density of states at the Fermi level, which apparently contributes to the relatively high T_c values of these materials. Two of the best known examples are the Chevrel-phase compounds MMo_6X_8, where *M* is one of many metals and *X* is a chalcogenide, and the rare-earth (*R*) rhodium borides RRh_4B_4. The highest T_c value for the former materials is ~15 K for $PbMo_6S_8$ and, for the latter compounds, ~11 K for YRh_4B_4. In these materials, the atomic clusters are Mo_6X_8 or Rh_4B_4 molecular units. The compound $PbMo_6S_8$ is the low-T_c superconductor with the highest H_{c2}, which reaches about 7×10^7 A m^{-1} (620 kOe) for $T \to 0$.

3.4 Magnetically Ordered Superconductors

Magnetically ordered superconductors are materials in which superconductivity and magnetic order occur within the chemical unit cell of the material. Two classes of *f*-electron materials fall into this category: molecular compounds and heavy-fermion compounds. In the molecular compounds, the electrons involved in the superconductivity reside primarily within the molecular units and interact only weakly with the magnetic moments of the rare-earth ions with partially filled 4*f* electron shells that occupy an ordered sublattice.

Antiferromagnetic order has been found to coexist with superconductivity in many materials, while ferromagnetic order has been observed to destroy superconductivity at a temperature T_{c2} near the Curie temperature and to produce a sinusoidally modulated magnetic state that coexists with superconductivity in a narrow temperature interval above T_{c2}. The materials in which these phenomena were first discovered include the rare-earth (*R*) molybdenum chalcogenides RMo_6X_8 (X = S, Se) and the rare-earth rhodium borides RRh_4B_4.

3.5 Heavy-Fermion Superconductors

Certain rare-earth and actinide intermetallic materials, known as heavy-fermion compounds, have effective electron masses of several hundred times the free electron mass, due to hybridization between the conduction electron states and the localized *f*-electron states (see HEAVY-FERMION PHENOMENA). A small number of Ce and U heavy-fermion compounds such as $CeCu_2Si_2$, UBe_{13}, UPt_3, URu_2Si_2, and UPd_2Al_3 exhibit superconducting behavior with values of T_c below ~2 K. The superconductivity in these materials appears to involve an energy gap that vanishes at points or along lines on the Fermi surface. The mechanism responsible for the pairing of the superconducting charge carriers may be mediated by magnetic excitations, rather than phonons. Several of the heavy-fermion compounds exhibit the coexistence of superconductivity and antiferromagnetism where both phenomena involve the same set of *f* electrons.

3.6 Low-Electron-Density Superconductors

This class of superconductors consists of materials with a low electronic density of states compared to the metallic systems described thus far. Among these compounds are several tellurides such as GeTe, SnTe, InTe, and Tl-doped PbTe. Another prominent group of materials includes superconducting oxides, in particular $SrTiO_3$ and $Ba(Bi_{1-x}Pb_x)O_3$, with the latter compound exhibiting a relatively high transition temperature of $T_c = 13$ K. In general, the transition temperature increases with increasing number density of the charge carriers. It is interesting to note that for $SrTiO_3$, T_c decreases after an initial increase when the carrier density is enhanced by doping. The decrease has been attributed either to screening of optical modes involved in the pairing by conduction electrons or to paramagnetic defects introduced with the doping.

3.7 Low-Dimensional Superconductors

Systems with pronounced structural and electronic anisotropies resulting in an effective reduced dimensionality have been of great interest, since superconducting electron pairing mechanisms mediated by excitations different from phonons have been suggested to be relevant in these cases, with a high dielectric polarizability giving rise to the attractive electron interaction and possibly to very high T_c. Furthermore, the occurrence of spin- and charge-density waves and of superconductivity as a result of the instability of the Fermi surface in low-dimensional systems and the competition between these interaction effects have attracted a lot of interest. Charge-density waves were first observed in superconducting layered transition-metal dichalcogenides, such as $NbSe_2$ or TaS_2. By intercalation of various amides and amines into TaS_2, the charge-density wave can be suppressed, resulting in an increase in T_c. Some of these and other intercalated compounds, such as MoS_2 intercalated with alkali-metal and alkaline-earth elements, exhibit very high initial upper critical field slopes close to T_c.

Other examples of low-dimensional materials exhibiting superconducting behavior include quasipolymeric strands of sulfur and nitrogen $(SN)_x$, organic charge-transfer salts (see Sec. 3.8), graphite intercalation compounds (see Sec. 3.9), and the high-T_c cuprates.

3.8 Organic Superconductors

A variety of organic compounds conducts electricity, the most prominent group being the so-called charge-transfer salts that contain donor (cation) and acceptor (anion) molecules (see SEMICONDUCTORS TO SUPERCONDUCTORS: ORGANIC LOWER-DIMENSIONAL SYSTEMS). Figure 13 shows several examples of donor and acceptor groups and gives the common abbreviations.

The conducting properties arise from the charge delocalization due to the strong overlap of π orbitals achieved by stacking the mostly planar molecules on top of each other. This layered structure gives rise to a pronounced anisotropy in the conductivity.

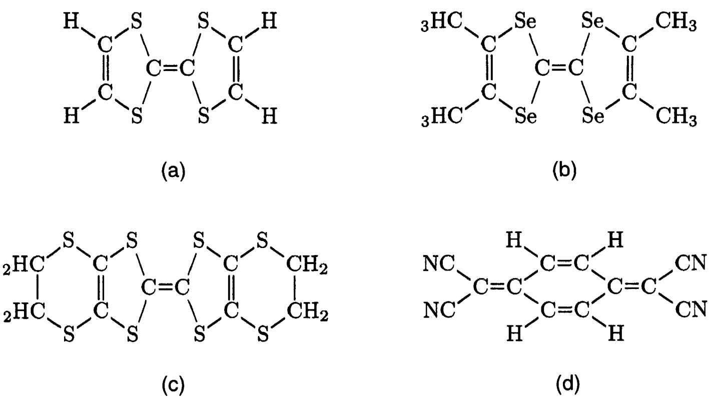
(a) (b) (c) (d)

FIG. 13. Typical donor and acceptor molecules forming organic charge-transfer conductors. (a) Tetrathiafulvalene (TTF). (b) Tetramethyltetraselenafulvalene (TMTSF). (c) Bis(ethylenedithia)-tetrathiafulvalene (BEDT-TTF or ET). (d) Tetracyanoquinodimethane (TCNQ). (a)–(c) are donors; (d) is an acceptor.

In 1980, the first organic superconductor, $(TMTSF)_2PF_6$, was discovered, with a T_c of 0.9 K at a pressure of 1.2 GPa (12 kbar). This material is a member of the family of compounds with the general formula $(TMTSF)_2X$ known as Bechgaard salts, where X is an inorganic anion such as PF_6^-, AsF_6^-, ReO_4^-, or ClO_4^-. The highest T_c achieved in this class is about 2 K.

Another important group of organic superconductors is based on the BEDT-TTF (ET) molecule (see Fig. 13) and described by the general formula $(ET)_2X$. The first superconducting compound of this class, $(ET)_2ReO_4$, was discovered in 1983. It has a T_c of about 2 K. Subsequently, higher transition temperatures of up to 12 K (at ambient pressure) were observed in compounds with X = $Cu(SCN)_2^-$, $Cu[N(CN)_2]Br^-$, and $Cu[N(CN)_2]Cl^-$.

Because of the strongly anisotropic character of the conductivity, organic superconductors are considered examples of "reduced dimensionality" systems. They exhibit very rich phase diagrams characterized by ground states where charge- and spin-density modulations are competing with superconductivity.

3.9 Carbon-Based Superconductors

Various graphite intercalation compounds (see INTERCALATION COMPOUNDS), which are formed by the insertion of atomic or molecular layers (intercalants) between the layers of graphite, exhibit superconductivity. Examples for these materials, which are often characterized by strongly anisotropic, quasi two-dimensional superconducting properties, include C_4KHg, C_4RbHg, C_8KHg, C_8RbHg, and C_8K, with T_c ranging from 150 mK to 19 K. Surprisingly, the critical temperatures of the C_8XHg compounds (X = K, Rb), where the intercalant sheets are separated by two graphite layers (two-staged), are higher than those for the one-staged C_4XHg materials.

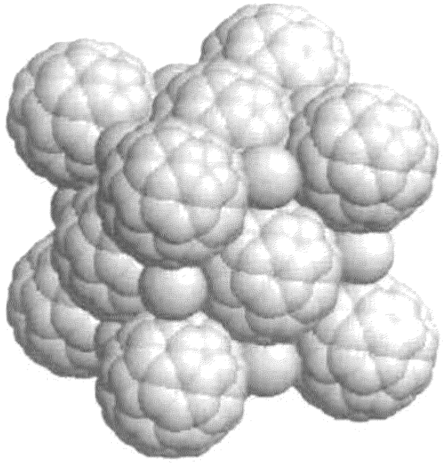

FIG. 14. Structure of the superconducting alkali-metal fulleride A_3C_{60}. The C_{60} molecules form a fcc lattice, with the smaller A cations occupying tetrahedral and octahedral sites.

Another fascinating group of superconductors that was discovered in 1991 is based on the intercalation of alkali-metal atoms into the interstitial sites of a spherically close-packed, face-centered cubic (fcc) solid of C_{60} molecules (see Fig. 14). C_{60} is a member of the fullerene family of carbon compounds forming soccer-ball–shaped molecules of high symmetry (see FULLERENES). If each of the two tetrahedral and the one octahedral sites per C_{60} unit are occupied by a single alkali-metal atom A, the resulting compound can be described by the general chemical formula A_3C_{60}. For higher alkali-metal concentrations, the crystal expands and undergoes a structural phase transition into body-centered tetragonal and body-centered cubic structures. The superconducting transition temperature T_c increases nearly linearly with increasing lattice constant a, and the application of hydrostatic pressure causes a contraction of the lattice and an associated suppression of T_c. While K_3C_{60}, the first superconducting fullerene to be discovered, exhibits a T_c of 18 K, a maximum T_c of 33 K was reported for the compound $RbCs_2C_{60}$. The observation of a distinct isotope effect (see Sec. 1.7) suggests phonon-mediated pairing of electrons as described by the BCS theory (see Sec. 2.1) as the mechanism responsible for superconductivity in these compounds, although pairing based on electron–electron correlations has also been considered.

4. APPLICATIONS OF SUPERCONDUCTIVITY

The large variety of physical phenomena associated with superconductivity are the basis of many applications involving superconducting materials and/or devices.

4.1 Large-Scale Applications

The capability of "hard" type II superconductors to carry large electrical currents in the presence of high magnetic fields naturally suggests their application for cables and conductors and in devices based on such conductors.

4.1.1 Superconducting Wires Wires based on superconducting materials that are to be used in conductors for energy or magnet technology have to meet a variety of requirements (see SUPERCONDUCTING MAGNETS): The transition temperature T_c, upper critical field H_{c2}, and critical current density j_c of the superconducting material should be as high as possible. The conductor should be stable against thermal fluctuations and flux jumps and tolerate sudden breakdown of superconductivity ("quenching"). The ac losses associated with changes in current and/or fields—e.g., during charging and discharging or pulsed operation of magnets, or in ac power applications—should be minimized. The mechanical properties should allow realization of complex conductor geometries (especially magnet coils) and, at the same time, provide sufficient strength to withstand the immense magnetic forces that may be present during operation.

Only two superconducting materials exhibit high enough critical current densities at elevated fields to be technologically relevant *and* have also been successfully integrated into the manufacturing process: the ductile alloy Nb-Ti and the brittle *A*15 compound Nb_3Sn. Because of its much more difficult processing, Nb_3Sn is only used at fields above $\sim 8 \times 10^6$ A m^{-1} (100 kOe), where the critical current density of Nb-Ti drops very rapidly with increasing field.

The conductors are produced in the form of composite wires consisting of up to 100 000 superconducting filaments that are embedded within a ductile, metallic matrix (usually copper). This subdivision into filaments with diameters ranging from ~ 1 to ~ 100 μm helps to suppress magnetic instabilities. Furthermore, the increased surface area between the superconducting filament material with its very low thermal conductivity and the Cu matrix, which is an excellent thermal conductor, provides more efficient cooling and increased thermal stability. In addition, the low-resistivity Cu matrix can carry the current in the event of a breakdown of superconductivity without developing dangerously large amounts of Joule heating. Finally, the matrix provides additional mechanical stability.

A typical manufacturing process for Nb-Ti wires includes the bundling of the superconducting and metallic components into a billet, followed by warm extrusion and drawing. In the case of the brittle Nb_3Sn, a "wind-and-react" technique is employed, where ductile composites consisting of Cu, Nb, and Sn are fabricated, whereas the superconducting phase is formed by solid-state reaction only after the conductor is given its final geometry. The filament diameter has to be limited to a few microns in order to facilitate the formation of Nb_3Sn within a reasonable time span. Thus, a high number of filaments are required, which can be achieved by repeating the bundling and drawing process.

Usually, the strands of multifilamentary wire are twisted to reduce ac losses, to limit the coupling between currents in neighboring filaments, and to increase magnetic stability. Figure 15 shows a cross-sectional image of a Nb-Ti multifilamentary wire. For applications requiring very large fields, additional stabilizing material (Cu, Al), mechanical reinforcements, and cooling channels for

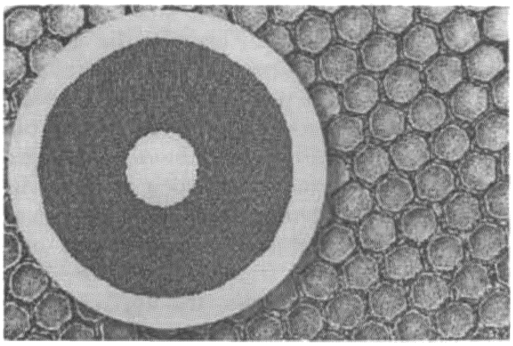

FIG. 15. Multifilamentary superconducting composite strand produced by IGC Advanced Superconductors for the SSC (superconducting supercollider) project. The strand (inset) is composed of a high-purity Cu stabilizing matrix and a single stack of over 4000 Nb–47 wt.% Ti filaments. Each Nb–Ti filament is wrapped in a Nb diffusion barrier (detail). Under thermomechanical processing, a fine dispersion of α-Ti precipitates is produced in the Nb–Ti, which are the primary flux pinners. (Micrograph by P. J. Lee. Reproduced with permission.)

the cryogens may be required. In ac applications, the Cu matrix may be subdivided by CuNi barriers to reduce eddy-current losses. Current-carrying capacities of more than 1 kA in fields up to $\sim 2 \times 10^7$ A m^{-1} (>200 kOe) can be achieved.

4.1.2 Superconducting Magnets By far, the most widespread application of conductors based on superconducting materials is in electromagnets. Superconducting magnets offer a number of advantages in comparison to conventional electromagnets made from normal metals: As a result of the absent (or very low) losses due to Joule heating, very high currents can be passed through the superconducting magnet windings, creating magnetic fields that are hardly achievable with conventional magnets. The lower losses drastically reduce the amount of electrical energy required to produce and maintain the magnetic fields. Superconducting magnets are much more compact and lightweight than conventional magnets that generate comparable fields. Since many applications require the magnetic field to be constant as a function of time, superconducting magnets are very often operated in the so-called persistent mode: The magnet terminals are shunted by a superconducting bridge (usually Nb-Ti) that can be driven normal by a heater wound coaxially around it. This assembly is referred to as a persistent switch. When the heater is activated and the switch is normal ("open"), the current flowing through the magnet may be changed using a connected power supply to generate the desired field. After turning the heater off, the persistent switch cools down, creating a superconducting loop consisting of the magnet windings and the switch. The persistent current that circulates in this circuit produces the extremely stable magnetic field.

Here, we only mention the application of superconducting magnets as laboratory research magnets, for nuclear magnetic resonance (NMR) spectrometers and magnetic resonance imaging (MRI); for particle acceleration, guidance, and detection; for plasma confinement in thermonuclear fusion reactors; and for magnetic levitation (see TRANSPORT BY MAGNETIC LEVITATION), magnetic separation of materials, or superconducting magnet energy storage (SMES) (see STORAGE OF ENERGY). For more details about superconducting magnets we refer the reader to the corresponding article.

4.1.3 Motors and Generators Superconducting generators and motors are considerably more efficient and lighter than their normal conducting counterparts. Both 5- and 10-MVA superconducting generators have been successfully tested; however, the breakeven point for economical usage is estimated to be reached only at 500–600 MVA. Combinations of superconducting generators, driven by gas turbines, with superconducting motors powered by the generators have been considered for ship propulsion, and a 3000-HP (2250-kW) system using a homopolar motor has been successfully tested by the U.S. Navy.

4.1.4 Electrical Power Transmission and Distribution Because of the vanishing dc losses and (for low ac field amplitudes) low 50/60-Hz losses, the transmission efficiency of superconducting power cables is much higher than for conventional cables. The feasibility and performance of metallic 4-K superconducting cables with lengths of up to 115 m has been demonstrated. However, the economic breakeven point would only be reached at transmission rates of 4000–5000 MVA, which is far beyond the values achieved in the tests.

Another potential application of superconductors in power networks is the superconducting fault-current limiter, which, in the case of a short-circuit, would enter the resistive normal state, thereby effectively limiting the current. Such a limitation would prevent damages to network components and reduce the load on generators, transformers, and switches. The lack of mechanically actuated parts should improve the reliability of the current limiter. Prototype single-phase current limiters consisting of Nb-Ti filaments embedded in highly resistive Cu-Ni have been successfully tested. As in the case of the cables, however, the high investment and operating costs have prevented large-scale implementation.

Transformers with superconducting windings also offer lower ac losses, higher efficiency, and a drastic reduction of size and weight, compared to normal conducting components. At the same time, the short-circuit current is limited to about double the normal operating current, which reduces the forces in the windings and thus allows simplification of the mechanical construction. However, care has to be taken to avoid eddy-current losses in the metallic matrix. While prototypes of 220-kV transformers constructed using submicron Nb-Ti wires embedded in Cu-Ni have been successfully tested, the efficiency did not surpass that of comparable conventional systems because of high thermal losses at the current feedthroughs. Nevertheless, the fault-current limiting capability of the superconducting windings was demonstrated, and a tenfold reduction in mass was achieved.

The application of low-T_c superconductors in electrical power transmission and distribution networks requires a further improvement in their superconducting properties, since only very large-scale units can be operated economically.

4.2 Small-Scale Applications

In contrast to the applications discussed above, the devices and techniques of this group are characterized by much smaller currents. They are mostly based on the Josephson effects (see Sec. 1.6) and are typically realized using thin-film technology.

4.2.1 Microwave Applications Passive high-frequency devices rely on the low surface resistance of superconductors, which enables them to transmit high-frequency signals with lower dissipation and less distortion than normal conductors, and allows reduction of the size of the elements. Among the devices realized are superconducting cavities and resonators, delay lines, filters, and antennas. Point-contact Josephson junctions, when biased to a steep part of a Shapiro step in their *I-V* characteristics (see Sec. 1.6), are very sensitive detectors of electromagnetic radiation in the microwave and submillimeter region. They can also function as efficient mixers.

4.2.2 Superconducting Quantum Interference Devices (SQUIDs) SQUIDs (*q.v.*) are the most sensitive detectors for magnetic flux. They are based on flux quantization and Josephson tunneling (see Sec. 1.6) and consist of a superconducting loop with one or more Josephson junctions.

With appropriate circuitry connected, SQUIDs can be used for the measurement of physical quantities that can be transformed into a magnetic flux, such as current, voltage, magnetic field gradient, magnetic susceptibility, or displacement. SQUIDs have been used to construct extremely sensitive voltmeters and galvanometers, as well as magnetometers in which the magnetic moments of very small samples can be determined with high accuracy. They are the basis of a primary thermometer measuring the Johnson noise of a resistor to determine temperature (see Sec. 4.2.4). Other applications include nondestructive evaluation of materials (see TESTING, NONDESTRUCTIVE), where SQUIDs measure the distortion of magnetic signals due to defects; geophysics (magnetotellurics, paleomagnetism, and piezomagnetics); and detectors for gravitational waves, where they are used to register the minute oscillations induced in large, freely suspended antennas by passing gravitational waves, as well as biomagnetism, i.e., the detection of the magnetic fields generated by the activity of biological systems, such as the human heart and brain (cardio- and neuromagnetism). For details of the principle, operation, and applications of SQUIDs, we refer the reader to the corresponding article.

4.2.3 Superconducting Digital Electronics The fact that Josephson junctions are capable of switching from the superconducting into the resistive state very rapidly (typical switching times are lower than 10 ps) makes Josephson-junction circuits very attractive for ultrahigh-speed digital signal processing. Other advantages are the integration of superconducting strip lines that are capable of transmitting terahertz signals without distorting them, and the very low power dissipation, which is of the order of a few microwatts and thus orders of magnitude lower than for semiconductor devices. Less power dissipation allows higher integration of the components, which reduces wiring delays. In addition, the quantization of magnetic flux in superconductors suggests the implementation of analog–digital converters and memory cells.

A variety of logic and memory circuits have been realized, initially based on lead-alloy technology; however, stability and reliability problems have induced increased efforts to produce integrated Josephson-junction circuits on the basis of other material combinations, such as NbN or Nb/Al/Al-oxide/Nb multilayers. Different types of logical gates have been proposed and realized: magnetic-coupled gates, where the state of Josephson junctions in a SQUID loop is switched by an external magnetic field associated with a current in a control line, and direct-coupled gates, in which the switching is triggered by injection of current directly into the circuit. Four- and eight-bit adders with subnanosecond addition times and (4 × 4)-bit multipliers containing more than 2800 individual Josephson junctions are among the logical circuits successfully tested, as well as four-bit counters, shift registers, and multiplexers. The realization of kilobit-sized random-access memory structures with subnanosecond access times is one of the main goals to be achieved in order to make progress on the way toward integrated superconducting digital systems.

4.2.4 Josephson-Junction Noise Thermometry Noise thermometers are based on the temperature proportionality of the Johnson-noise voltage fluctuations across a resistor. One implementation uses a Josephson junction that is biased by a dc current at a voltage V and is shunted by a resistor R ("R-SQUID"). The noise voltage across R modulates the high-frequency oscillation in the junction, and a measurement of the linewidth of this modulation, usually by detection of the frequency variance of the radiation emanating from the junction, yields the temperature information. In contrast to this absolute temperature measurement, which was used to establish a temperature scale between 6 and ~700 mK, an alternative method that uses a SQUID as an ultralow-noise preamplifier in a conventional noise thermometer circuit realizes a secondary thermometer that has to be calibrated against a primary standard.

4.2.5 Josephson-Junction Voltage Standard The presence of Shapiro steps in the I-V characteristics of Josephson junctions irradiated with microwaves of frequency ν_0 (see Sec. 1.6) provides a relationship between the voltage V_n, where the nth step is observed, and the frequency: $V_n = nh\nu_0/2e$. Since the frequency of the radiation can be stabilized to a very high degree, this opens up the possibility of using these steps to define reproducible reference voltages. A typical voltage value obtained using a single junction at a frequency of 70 GHz is 145 μV. For low microwave power, the constant-voltage steps overlap for a certain current bias, so that the voltage is multiple-valued with steps of $\Delta V = h\nu_0/2e$ between neighboring levels. All possible multiples of ΔV can be obtained in series arrays of Josephson junctions with appropriate combinations of the outputs of the individual junctions. Voltage standards up to the 10-V level have been achieved using series arrays of more than 10 000 individual junctions. Tuning the frequency of the microwave radiation allows the continuous change of the voltage, thus realizing a Josephson potentiometer. For details on the Josephson voltage standard, the reader is referred to the article CONSTANTS, FUNDAMENTAL. We only mention the use of superconductors as cryogenic current comparators, which are especially attractive for highly precise resistance-ratio measurements.

GLOSSARY

Abrikosov Vortex: Same as **Flux Line.**

Cooper Pair: Pair of electrons bound together in a singlet state (characterized by opposite spins) with a finite binding energy due to an attractive interaction.

Critical Magnetic Field: Characteristic magnetic field at which superconductivity in a type I superconductor is destroyed.

Critical Temperature: Transition temperature for the phase transition from the normal state with average zero superconducting order parameter to the superconducting state characterized by a nonzero order parameter.

Demagnetization Factor: Tensor quantity describing the difference between the external field outside a magnetized sample and the internal field. Only defined for ellipsoidal geometries, where the magnetization is homogeneous.

Energy Gap: Characteristic gap in the quasiparticle excitation spectrum of a superconductor corresponding to the binding energy of the Cooper pairs.

Flux Line: Magnetic flux penetrating a superconductor in the mixed state. Corresponds to the volume in which the superconducting order parameter is locally suppressed. In clean superconductors, flux lines arrange themselves in a regular triangular lattice (Abrikosov lattice).

High-Temperature Superconductors: New class of superconductors discovered in 1986. See SUPERCONDUCTIVITY, HIGH TEMPERATURE.

Intermediate State: Equilibrium state assumed by a type I superconductor with nonzero demagnetization factor in magnetic fields close to the critical magnetic field, characterized by the coexistence of superconducting and normal domains inside the superconductor.

Isotope Effect: Dependence of the superconducting transition temperature T_c on the atomic mass M of a constituent of a superconductor. Indication for a phonon-mediated pairing mechanism.

Josephson Effects: Appearance of a dc supercurrent across an unbiased Josephson junction (dc Josephson effect) and of an oscillating supercurrent across a junction biased to a voltage (ac Josephson effect).

Josephson Junction: Area of locally reduced order parameter separating two superconducting regions.

Low-Temperature Superconductors: Superconductors with transition temperatures T_c below ~23 K, the highest T_c achieved before the discovery of the cuprate high-temperature superconductors.

Magnetic Penetration Depth: Characteristic length over which magnetic fields decay inside superconductors.

Meissner Effect: The expulsion of magnetic field H from the interior of a superconductor (except for a surface sheet with a thickness corresponding to the magnetic penetration depth), so that the magnetic induction B remains zero inside the superconducting sample.

Meissner Phase: Equilibrium state of a superconductor characterized by complete flux expulsion.

Mixed State: Equilibrium state assumed by a type II superconductor above a characteristic field H_{c1}. Characterized by the penetration of magnetic flux into the superconductor in the form of flux lines, which, in a clean superconductor, arrange themselves in a triangular lattice (Abrikosov flux lattice).

Order Parameter: Phenomenological complex quantity for the description of the superconducting phase transition corresponding to the degree of superconducting ordering.

Persistent Current: Electrical current loop in a superconductor, which decays only with an extremely long time constant.

Pinning Center: Area of locally reduced order parameter due to defects or impurities within a type II superconductor, where it is energetically favorable for flux lines to be located, resulting in an effective force acting on the flux line.

Shapiro Steps: Constant-voltage steps in the *I-V* characteristics of a biased Josephson junction irradiated with microwaves.

Shubnikov Phase: Same as **Mixed State.**

SQUID (Superconducting Quantum Interference Device): Superconducting loop with one or more Josephson junctions exhibiting characteristic dependences of the critical current on the magnetic flux threading the loop. See SQUIDs.

Superconducting Coherence Length: Characteristic length describing the spatial range of the superconducting electron pair wave function.

Thermodynamic Critical Field: Characteristic magnetic field defined by the magnetic energy density that corresponds to the difference in free-energy densities between the superconducting and normal states in the absence of a magnetic field.

Type I Superconductor: Superconductor characterized by complete flux expulsion (see **Meissner Phase**) below a critical magnetic field.

Type II Superconductor: Superconductor characterized by complete flux expulsion (see **Meissner Phase**) below a characteristic lower critical field, the penetration of magnetic flux in the form of flux lines (see **Mixed State**) above this field, and the transition to the normal state at a second characteristic field, the upper critical field.

Weak Link: see **Josephson Junction.**

Further Reading

Abrikosov, A. A. (1988), *Fundamentals of the Theory of Metals*, Amsterdam, Oxford, New York, Tokyo: North Holland.

Barone, A., Paterno, G. (1982), *Physics and Applications of the Josephson Effect*, New York: Wiley.

Buckel, W. (1991), *Superconductivity: Fundamentals and Applications*, New York: VCH.

Dahl, P. F. (1992), *Superconductivity: Its Historical Roots and Development from Mercury to the*

Ceramic Oxides, New York: American Institute of Physics.

de Gennes, P. G. (1966), *Superconductivity of Metals and Alloys,* New York: W. A. Benjamin.

Huebener, R. P. (1979), *Magnetic Flux Structures in Superconductors,* Berlin, New York: Springer.

London, F. (1950), *Superfluids,* New York: Wiley.

Lynton, E. A. (1969), *Superconductivity,* 3rd ed., New York: Wiley.

Parks, R. D. (Ed.) (1969), *Superconductivity,* New York: Marcel Dekker.

Rowell, J. (Ed.) (1991), Special Issue: High Temperature Superconductivity, *Phys. Today* **44** (6), 22–82.

Savitskii, E. M., Baron, V. V., Efimov, Yu. V., Bychkova, M. I., Myzenkova, L. F. (1973), *Superconducting Materials,* New York, London: Plenum Press.

Schrieffer, J. R. (1964), *Theory of Superconductivity,* New York: W. A. Benjamin.

Shoenberg, D. (1952), *Superconductivity,* New York: Cambridge Univ. Press.

Tinkham, M. (1995), *Introduction to Superconductivity,* 2nd ed., New York: McGraw-Hill.

Tinkham, M. (Ed.) (1986), Special Issue: Superconductivity, *Phys. Today* **39** (3), 22–80.

Wilson, M. N. (1983), *Superconducting Magnets,* New York: Oxford Univ. Press.

SUPERFLUIDITY: LIQUID HELIUM SYSTEMS

See Supplement Volumes

SUPERLATTICES

See HETEROSTRUCTURES AND SUPERLATTICES, METALLIC; HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR

SURFACE CONDUCTIVITY

See INSULATORS AND SEMICONDUCTORS, SURFACE CONDUCTIVITY IN

SURFACE PROCESSES

HARALD BUSSE AND BRUCE E. KOEL, *Department of Chemistry, University of Southern California, Los Angeles, California, U.S.A.*

Introduction 273
1. **Surface Structure** 275
1.1 Thermodynamics of Surfaces 275
1.2 Surface Defects and Dislocations 275
1.3 Relaxation and Reconstruction 276
1.4 Surface Segregation 278
1.5 Surface Alloy Formation 278
1.6 Electronic Structure at Surfaces 278
2. **Energetic-Particle–Surface Interactions** 279
2.1 Photon–Surface Interactions 279
2.1.1 Photoelectron Emission (Photoemission) 280
2.1.2 Electromagnetic Fields at Surfaces 281
2.2 Electron–Surface Interactions 281
2.2.1 Diffraction 282
2.2.2 Electron Emission 282
2.3 Ion–Surface Interactions 283
2.3.1 Scattering 283
2.3.2 Sputtering and Implantation 284
3. **Gas–Surface Interactions** 285
3.1 Surface Collisions 285
3.1.1 Energy Exchange 285
3.1.2 Atomic- and Molecular-Beam Scattering 286
3.2 Adsorption 287
3.2.1 Physisorption and Chemisorption 288
3.2.2 Potential-Energy Diagrams 289
3.3 Diffusion 291
3.4 Desorption 292
4. **Surface Reactions** 293
4.1 Heterogeneous Catalysis 293
4.2 Oxidation and Corrosion 295
5. **Deposition and Surface Processing** 296
5.1 Nucleation 296
5.2 Growth Mechanisms 297
5.3 Deposition Methods 299
5.4 Etching 300
5.5 Lithography and Nanotechnology 301
6. **Mechanical Processes** 302
6.1 Adhesion 302
6.2 Friction and Lubrication 303
7. **Complex Surfaces and Environments** 303
7.1 Amorphous and Porous Materials 304
7.2 Polymers and Biological Materials 305
7.3 Electrochemistry and Fuel Cells 305
7.4 Space Exploration 306
Glossary 307
Works Cited 308
Further Reading 309

INTRODUCTION

It has been said that surfaces are the "seats of communication" between two phases. Atoms and molecules to the greatest extent sense or know about other energetic or reactive atoms or molecules that constitute a different phase only if they are at the surface. Thus processes at surfaces control the changes that occur for a large fraction of all possible physical and chemical processes. Upon formation of a surface, or more generally an interface, the surface atoms have a unique electronic structure and often geometric arrangement that is different from those atoms in the bulk phase because of the abrupt termination of the homogeneous, bulk environment. This altered structure induces new, altered chemical and physical properties for surfaces compared to the corresponding bulk phases, such as enhanced reactivity or lower melting point. Also, it is necessary to introduce separate descriptions

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

of the thermodynamics of the interfacial regions, such as indexing separately the surface and bulk contributions to the entropy, and to catalog new thermodynamic state variables such as the surface tension for liquids and stress tensor for solids.

It is exceedingly important to understand surfaces, and the adsorbed layers present at surfaces, since the physical changes and chemical reactions that occur there underpin practical applications in many scientific fields, important technologies, and industrial processes. Table 1 illustrates just a few of the myriad of applications where surfaces play a key role. Historically, heterogeneous catalysis, which is at the heart of the huge chemical and oil-refining industries, provided impetus to understand the structure of solid surfaces and the chemical reactions that occur there. Perhaps the greatest efforts currently involve applications in materials science and engineering, such as the processing of Si and SiO_2 for microelectronic device manufacture, development of high-temperature superconducting films, and synthesis of ultrathin metal films and quantum dots with unique electronic and magnetic properties. These driving forces along with great instrumental developments in physics and physical chemistry have led to the emergence since the mid-1960s of the field of surface science—the study of surface processes. A great deal is now known about surfaces, mostly solid surfaces, and about surface processes, but a complete description and understanding of any but the most elementary process is still not possible. The surface science field is a vibrant one as advances are made toward an overarching goal of atomic-level design and manipulation of materials.

Since gases are miscible at all proportions, possible interfaces are S/S, S/L, S/G, L/L, and L/G, where S, L, and G denote solid, liquid, and gas phases, respectively. Table 2 shows a partial listing of processes that occur at the different interfaces. Primarily as a result of the analytical capabilities currently available, the most fundamental information is known about solid surfaces and interfaces involving solids. This article emphasizes an elementary and atomistic point of view in describing some of the major types of surface processes and so focuses on solid surfaces and in particular on the gas–solid or vacuum–solid interface. For a detailed treatment of the nature and thermodynamics of liquid interfaces, surface films on liquid substrates, and the solid–liquid interface, the reader is referred to the classic text by Adamson (1990). The interested

Table 1. Surfaces play a key role in a wide variety of technological processes.

Field	Processes or materials
Biochemistry	Membranes, colloids
Catalysis	Oil refining, natural gas conversion, chemical production
Combustion	Catalytic ignition, pollution control, soot formation
Electronics	Semiconductor and solid-state device fabrication and performance
Fusion	Plasma/wall reactions
Lubrication	Friction, high-temperature service
Mining	Geochemistry, recovery
Paints and Coatings	Adhesion, corrosion inhibitors
Plastics	Production, molding
Solar Energy	Mirror technology, photochemistry
Space Exploration	Vacuum/material interfaces, space shuttle "glow"
Surfactants	Soaps, foams, emulsions

Table 2. Processes and effects at different interfaces between the solid (S), liquid (L), and gas (G) phases.

Interface		Surface processes and effects
$\frac{\text{Gas}}{\text{Solid}}$	G/S	Adsorption, catalysis, condensation and nucleation, corrosion, CVD, desorption, diffusion, energy transfer, etching, oxidation, permeation, reaction, surface states, thin-film growth
$\frac{\text{Liquid}}{\text{Solid}}$	L/S	Adsorption, capillarity, catalysis, cleaning, corrosion, electrochemical effects, etching, friction, ion electromigration, lubrication, nucleation and growth, optical properties, reaction, spreading, wetting
$\frac{\text{Solid A}}{\text{Solid B}}$	S/S	Abrasion, adhesion, corrosion, creep, delamination, diffusion, epitaxial growth, friction, grain-boundary, passivation, nucleation, reaction, segregation, solid-state device fabrication, thin-film growth, wear

reader is also referred to the recent monograph by Sutton and Balluffi (1995) for the basic ideas and principles relating to solid–solid interfaces.

1. SURFACE STRUCTURE

The cleavage of a solid forms two additional surfaces. The new surface atoms that were previously embedded in a well-defined and ordered bulk environment now suddenly lack some of their neighboring atoms and consequently some of their interactions. Thus, there exist large changes in the coordination number (number of bonding neighbors) and electrical potential at an interface or surface compared to that in the bulk phase. This dramatically affects the electronic and geometric structure of surfaces, along with the surface composition in multicomponent systems. Since the surface structure imparts the specific chemical and physical properties of the surface, we begin with a discussion of changes that are inherent to a solid surface.

1.1 Thermodynamics of Surfaces

A proper thermodynamic description of a system with a free surface must include an extensive variable A, the surface area, and an intensive variable γ, the surface tension. Additional surface contributions to the Gibbs free energy G per atom may be defined by

$$N^{\mathrm{tot}}G^{\mathrm{tot}} = N_{\mathrm{B}}G + AG^{\mathrm{S}}, \qquad (1)$$

where N_{B} is the number of atoms in the bulk and G^{S} is the Gibbs free energy per unit area. Similar expressions exist for other thermodynamic variables, e.g., enthalpy H and entropy S, where H^{S} and S^{S} indicate their respective surface contributions. Since solids do not spontaneously cleave, formation of a surface must increase the free energy. The change in total free energy due to increasing the surface area of an isotropic solid is equal to the reversible work dW^{S}, given by

$$dW^{\mathrm{S}} = \left(\frac{\partial(AG^{\mathrm{S}})}{\partial A}\right)_{T,P} dA = G^{\mathrm{S}}dA + A\left(\frac{\partial G^{\mathrm{S}}}{\partial A}\right)_{T,P} dA. \qquad (2)$$

The first term of the sum in Eq. (2) is the work spent to form a unit area of surface, and the last term is the work required to stretch the surface, which involves the surface stress. The last term of Eq. (2) vanishes for liquids because G^{S} does not depend on the area; work is only required to form additional unit surface area, and G^{S} is identified as the surface tension γ. In the case of solids, the surface tension and the surface stress are not the same. Also, for anisotropic solids we require formulation of a surface stress tensor since the extent of stretching (strain) induced by a given force depends on the direction of the force. Tables for liquid surface tensions are commonly available, and these can provide a rough guide for solids. Refractory metals have some of the highest surface tensions (≈150 kJ/mol), with lower values for oxides and other compounds, and the lowest values are for organic solids and polymers ($\gamma_{\mathrm{Teflon}} < 0.5$ kJ/mol). The presence of surface tension has several important consequences for solids: Small particles assume a spherical shape (within the constraints of their crystalline structure), and nanometer-size particles have increased vapor pressure and reactivity.

1.2 Surface Defects and Dislocations

Practical surfaces of crystalline solids exhibit simple structural imperfections (defects) on an atomic scale. Steps and kinks (at the intersection of steps) occur as the macroscopic angle of the surface with respect to a crystalline symmetry direction is produced with low–free-energy atomic planes. Isolated atoms (adatoms) and missing atoms (vacancies) on the flat terraces also exist as point defects as shown in Fig. 1(a). The reduced coordination number (number of bonding neighbors) of defect atoms induces changes in the electronic structure at these locations and imparts special reactivity. The presence and distribution of steps affects the growth rate of thin films because steps serve as preferential adsorption sites.

Line defects of the bulk structure, such as edge or screw dislocations, may extend to the surface. Figure 1(b) illustrates how a screw dislocation produces a ledge on the surface, providing altered adsorption and nucleation sites. Grain boundaries at the intersection of separate crystalline regions are

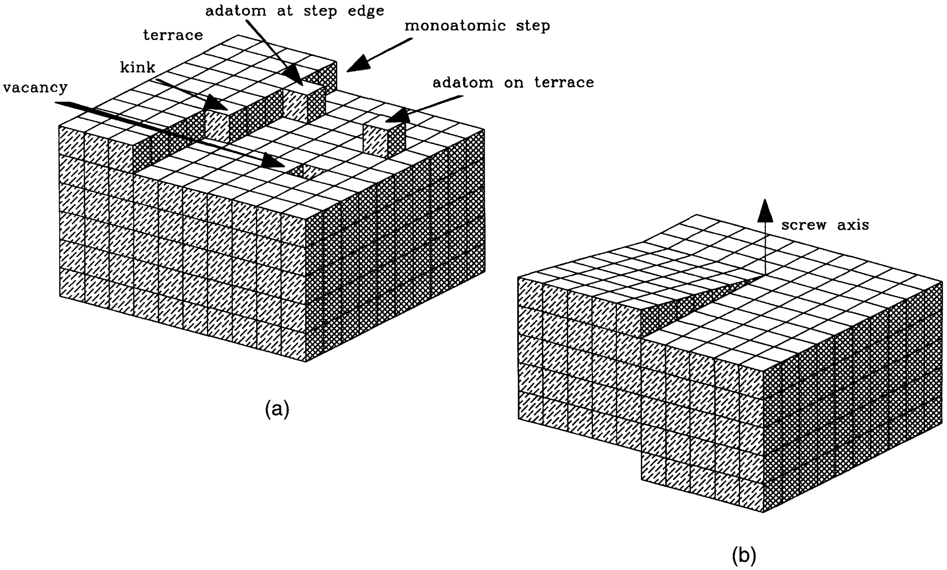

FIG. 1. Schematic representation of surface defect structures. (a) Terraces, steps, kinks, adatoms, and vacancies. (b) A screw dislocation emerges at the surface.

preferential pathways for the interlayer diffusion of atoms.

Heterogeneous defect structures are formed or induced by impurity atoms (such as C, S, and O for metals and alloys) or aggregates that affect the surface stress and the equilibrium arrangement of surface atoms. Minimization of the total energy of the actual surface structure (including defects) is the driving force for structural changes. A common defect evolution via the diffusion of adatoms, however, is often controlled kinetically, so that the surface does not necessarily correspond to its equilibrium structure.

1.3 Relaxation and Reconstruction

Surface atoms rearrange into positions away from those expected from the ideal termination of the bulk lattice in order to achieve a new, more energetically favorable configuration. The properties and reactivity of the surface are affected by this reordering. Two different phenomena are distinguished: relaxation and reconstruction.

Relaxation is the movement of surface atoms perpendicular to the surface plane. This is a common phenomenon for the closest-packed (i.e., lowest Miller index) planes of metal surfaces, such as the fcc (111) surfaces of Pt or Ni. The interlayer spacings between the first and the second layer, d_{12}, the second and the third layer, d_{23}, etc., deviate slightly from the bulk value d_B. Generally, the relaxation $(d_{12} - d_B)/d_B$ is (1–4)% of the nearest-neighbor distance for transition metals, but the more open the surface (increasing surface roughness) the larger the relative deviation can be, and these values can exceed 15%. In some cases it has been found that the values for d_{12} and d_{34} are smaller than d_B, while the value for d_{23} is increased compared to d_B. Relaxation and this oscillatory behavior of the different interlayer spacings have been measured extensively by low-energy electron diffraction (LEED). Prediction of these oscillations and comparisons to experiment provide an active test bed for theoretical treatments of the energy minimization problem at metal surfaces.

Reconstruction is a term describing large displacements of the surface atoms in three dimensions from those positions expected from the bulk lattice. This is a dominant process for solids characterized by highly directional, covalent bonding of the atoms, such as in semiconductors. The creation of a Si(100) or Si(111) surface initially leaves the Si surface atoms with an unsaturated valence. Eliminating these broken bonds, called "dangling bonds," drives the reconstruction of Si surfaces: atoms change positions to saturate these bonds, i.e., minimize the number of dangling bonds. This is achieved via a so-called 2 × 1 reconstruction to give rows of Si dimers (pairs of atoms) with a 2 × 1 surface net for Si(100) and by a very complex 7 × 7 reconstruction of the clean Si(111) surface that extends over three layers. The latter surface structure was unresolved for many years until it was determined using electron diffraction and verified by several techniques including scanning tunneling microscopy (STM). Figure 2 shows a structural model along with a STM image of this surface.

Even some metal surfaces with low–Miller-index planes, e.g., the fcc (110) surfaces of Au, Pt, and Ir, reconstruct forming "added-row" structures, and those with high–Miller-index planes reconstruct to form ordered step arrays in order to maximize the concentration of the low-free-energy, low–Miller-index planes.

Surface reconstructions occur for all compound semiconductors and for surfaces of oxides and other compounds, but much less information exists on the structures of these materials on account of the difficult problem of determining and controlling their surface composition and stoichiometry under conditions suitable for surface analysis. It is of great interest to know more about the surface structure of molecular solids, e.g., solid water (ice), energetic materials such as propellants, polymers, and model membranes, but these materials are very sensitive to damage by probing electron or ion beams

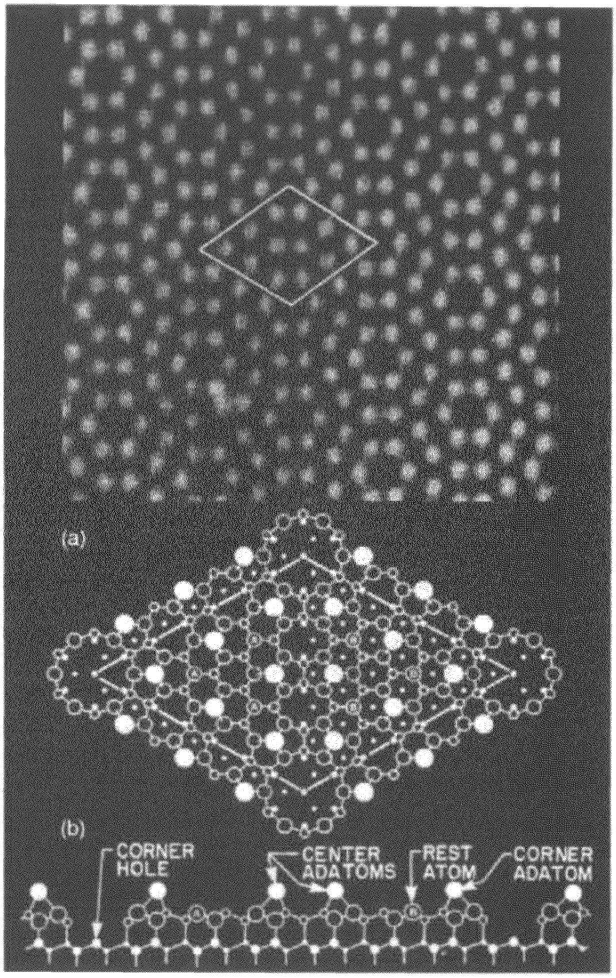

FIG. 2. The "7 × 7" reconstruction of the Si(111) single-crystal surface. Top: A STM image on which a surface unit cell (rhombus) has been overlaid (from Wiesendanger, 1994). Bottom: Schematic drawing of this surface showing large-scale atomic displacements away from those positions for a simple termination of the bulk lattice. (From Takayanagi *et al.*, 1985.)

used for analysis and are just now starting to yield to detailed investigations.

1.4 Surface Segregation

Multicomponent systems have a different surface composition compared to the bulk. This phenomenon is called surface segregation and is largely due to different surface free energies for the individual components. In the simplest approach, using ideal-solution theory for a binary system AB, the component with the lowest surface free energy will be found at the surface in a concentration given in terms of mole fractions x_i according to

$$\frac{x_A^{\text{surf}}}{x_B^{\text{surf}}} = \frac{x_A^{\text{bulk}}}{x_B^{\text{bulk}}} \exp\left(\frac{\gamma_b - \gamma_A}{RT}\right), \tag{3}$$

where R is the gas constant and T is the temperature. Since for liquid metals the surface tension is nearly proportional to the heat of sublimation ΔH_{subl} ($\gamma \approx 0.16\Delta H_{\text{subl}}$), then the metal with the lowest value of ΔH_{subl} segregates to the surface; and the bigger the difference $\gamma_B - \gamma_A$, the more different are the concentrations of the surface and bulk. For most metal alloys these effects are quite large, leading to nearly complete coverage of the surface by a component that may be present in the bulk at negligible concentrations. More accurate expressions can be developed using correction terms to account for enthalpy changes upon mixing in the regular solution model and to account for strain induced in the lattice by size differences in the components. Enhanced surface concentrations are found for the largest atoms. Segregation will occur preferentially to surface planes and positions with the highest free energies, and this has a strong effect on surface reactivity and structure, especially for small, crystalline particles typical of catalysts.

1.5 Surface Alloy Formation

Interdiffusion or intermixing (alloying) can occur to form a surface alloy film or a bulk alloy (solid solution) in the deposition of one miscible metal onto another. However, the altered environment inherent to surfaces can lead to alloying in the surface layer even when the bulk metals are completely immiscible, like in the case of Au on a Ni(110) single crystal. This phenomenon results in surface structures with novel properties.

Strong chemical bonding between the constituent atoms can lead to intermetallic compounds with a well-defined structure in the bulk and at the surface. These compounds usually have the formula A_mB_n, where m and n are integers, and the surface phase often corresponds to a structure of a crystalline plane of the bulk material. By reason of surface free energies, ordered surface phases can be formed that do not correspond to bulk crystal planes, such as in the case of an ordered, metallic γ-$FeSi_2$ surface alloy film formed by high-temperature annealing of an Fe film on a Si single-crystal substrate. Surface science methods must determine the existence and properties of such surface intermetallic phases, since they cannot be predicted by the bulk structures.

1.6 Electronic Structure at Surfaces

The 3D periodicity of the electronic potential is lost at the surface. This means that new traveling-wave solutions of the Schrödinger equation exist, and these are known as surface states—states with wave functions localized at the surface that can have energies that are not allowed within the bulk band structure. Surface states can also be created by changes in the potential due to relaxation, reconstruction, defects (e.g., emerging screw dislocations), or adsorbed impurities. These special states affect the electrical properties and the chemical reactivity of the surface by modifying the availability of filled or empty electronic states.

The localization of charge in these surface states leads to another important effect—band bending. The band structure of the bulk solid is distorted by the varying electrostatic potential on passing from the surface into deeper layers. The existence of occupied surface states can result in surface potentials on the order of millivolts or volts, and band bending can extend on the order of microns into the solid for insulators and semiconductors. Metals have negligible band bending on account of efficient screening of the surface charge by the conduction-band electrons. Surface capacitance and surface conductivity are used to probe these electronic properties (Prutton, 1983).

The work function of a solid surface controls the removal of electrons from a solid surface, affecting the contact potential between two solids, thermionic and field emission of electrons, and bonding of adsorbates. The work function ϕ is the energy difference between an electron at rest in the vacuum just outside the solid and the most loosely bound electrons inside the solid (at the Fermi energy). In addition to the attractive potential for conduction electrons in the bulk, the value of ϕ has several surface contributions: the image potential, due to electrostatic screening of charge outside of a conductor; and the surface dipole layer, arising from the overextension of electron wave functions beyond the ion cores at the surface, which leads to a charge imbalance perpendicular to the surface. The work function depends on the geometry and composition of the surface, the presence of surface roughness and defects, and the concentration and nature of adsorbed layers. Typically the value of ϕ is measured via the photoelectric effect, and work-function changes $\Delta\phi$ are sensitively monitored by capacitance methods such as the Kelvin probe.

2. ENERGETIC-PARTICLE–SURFACE INTERACTIONS

Surface processes involving energetic particles (photons, electrons, and ions) form the basis for surface analytical spectroscopies and for state-of-the-art processing methods for producing solid-state electronic devices. Depending on the energy and momentum transfer of the interaction of these particles with condensed matter, these processes may be quite varied, ranging from nondestructive probing of surface vibrations with IR radiation to atom ejection caused by impact of high-energy ions. We provide here a brief and phenomenological overview of several of these important interactions. We discuss the interactions of energetic particles with clean and *adsorbate-covered* surfaces. This is appropriate since surfaces are usually covered with monomolecular layers (monolayers) or thin films of adsorbed atoms and molecules, and these layers are important to many processes.

2.1 Photon–Surface Interactions

The interaction of photons with surfaces may be divided into photoabsorption, reflection, and scattering. Photoabsorption causes specific elementary excitations (ionization or electronic and vibrational transitions) in the substrate or adsorbed layers. Additional transitions are possible at surfaces as a result of the presence of surface states, surface plasmons (collective electron excitations), and surface phonons (collective lattice vibrations). The excitation energies and excited-state lifetimes of molecules in adsorbed layers are modified through the bonding to the surface. Photodesorption, photofragmentation, and photon-stimulated desorption (PSD) alter surface layers by breaking bonds. PSD causes desorption of atoms and molecules from surfaces due to nonthermal mechanisms under irradiation of ultraviolet (UV) or higher photon energies. The bulk photoabsorption technique of extended x-ray absorption fine structure (EXAFS) can be applied to surface atoms in SEXAFS where the photoabsorption signal is monitored with surface sensitivity by means of the secondary-electron yield or Auger electrons. In SEXAFS, as in EXAFS, oscillations appear as fine structure above the absorption edge due to interference between the emitted photoelectron wave and the waves backscattered from adjacent atoms, and subsequent Fourier analysis yields surface structural information.

The dielectric response of the surface can be derived by determining the surface contribution to the reflectivity. The simplest approach is to consider the substrate, surface, and ambient as three phases separated by sharp boundaries. The surface contribution may then be from the outermost layer(s) of the substrate or from molecular layers adsorbed on the substrate. Assuming that the thickness of the surface layer is small compared to the wavelength of the light, one can calculate the differential reflectivity $\Delta R/R$ with and without the surface layer. The value of $\Delta R/R$ is related to the imaginary part of the complex dielectric function, Im ϵ, that is associated with the surface layer. This model has been successfully applied to some electronic transitions on semiconductors and vibrational excitations in adsorbed layers. Ellipsometry (*q.v.*), one of the oldest surface-sensitive techniques, utilizes measurements of the elliptically polarized light that results from reflection of plane-polarized light from a solid surface to determine concentrations of adsorbed layers and surface film thicknesses. In recent years, using IR reflection from surfaces has become one of the most powerful probes of vibrations and identifying the species present in adsorbed layers.

The scattering of high-energy photons as applied in grazing-incidence x-ray scattering or x-ray standing-wave (XSW) analysis is an increasingly important tool in surface science. This method uses intense synchrotron radiation and very low angles of incidence (within 12° of the surface plane) to achieve total external reflection accompanied by an evanescent wave propagating along the surface to enhance the signal from the surface in x-ray diffraction studies.

Photochemical and photoprocessing technologies involve surfaces. For example, light absorption in the near-surface region of semiconductors is an elementary step in the operation of photovoltaic (solar) cells, and photocatalysts utilize light absorption by the surface region of an oxide such as TiO_2 to generate very reactive surface species that can oxidize and decompose biomaterials and chlorinated organic pollutants. Absorption of intense laser light by a solid heats surfaces and thin films to the melting point (or beyond) in 10^{-12} s. This interaction is the basis for sources used to inject nonvolatile materials (e.g., large biomolecules) into mass spectrometers, flash-annealing and quenching materials processing, and laser-induced thermal desorption (LITD) that has been used to study surface diffusion and reactions.

2.1.1 Photoelectron Emission (Photoemission) Photoelectron emission spectroscopy is based on the photoelectric effect. Light incident on solid surfaces produces free electrons from the surfaces when the energy of the incident light exceeds the work function ϕ of the sample. Energy conservation requires that the kinetic energy E_K of the emitted photoelectron is given by the difference between the photon energy $\hbar\omega$ and the binding energy E_B of the electron in the solid, according to the equation $E_K = \hbar\omega - \phi - E_B$. Ultraviolet photoelectron spectroscopy (UPS) uses 5–40-eV photons to probe sensitively the valence electronic structure of both the substrate surface and adsorbed layers at the surface. Angle-resolved photoemission has been a powerful tool for determining the energy and dispersion of the electronic band structure of crystalline solids and for studying surface states. A dipole selection rule governs photoemission, and a simple form of the matrix elements can be expressed as $\langle f|\mathbf{p}|i\rangle \cdot \mathbf{A}$, where i and f denote the initial and final state, respectively, $\mathbf{p}$ is the momentum operator, and $\mathbf{A}$ is the vector potential of the electromagnetic field. By constraining the outgoing photoelectron final state (with the detector direction) and controlling the polarization vector of the incident light, one can derive straightforwardly selection rules for the initial states that are dipole-allowed and thus the orientation of adsorbed molecules can be determined. The majority- and minority-spin states in ferromagnetic systems can be distinguished if one uses spin-polarized sources and detectors. In inverse photoemission, electrons incident on the surface transiently occupy empty electronic states and provide a mapping of the unoccupied density of states.

Binding energies of electrons in atomic core levels are characteristic of the probed atom. X-ray photoelectron spectroscopy (XPS), also known as electron spectroscopy for chemical analysis (ESCA), is used for quantitative elemental analysis of surfaces. Small changes in these core-level binding energies due to changes in the number of valence electrons (oxidation state) make XPS a rather unique probe of the chemical state of atoms at surfaces, e.g., distinguishing Fe^0 in metallic Fe from the oxidized Fe^{+3} in Fe_2O_3. A different binding energy occurs for electrons in core levels of atoms at the surface compared to the bulk—the surface core level shift. In x-ray photoelectron diffraction (XPD), the photoemission intensity at a particular energy is measured as a function of the incident photon energy (using synchrotron radiation) and/or the emission angle. Modulation of this intensity occurs by interference of the directly emitted photoelectron wave with parts that are backscattered from neighboring atoms. This gives local information on distances to adjacent atoms in the substrate or adsorbate, orientation of molecules on surfaces, and overlayer growth modes.

2.1.2 Electromagnetic Fields at Surfaces

There are a range of phenomena particularly associated with the electromagnetic fields at surfaces. For example, giant enhancements (by factors of 10^6) can occur in the Raman-scattering cross section on surfaces, especially for roughened silver surfaces. Surface-enhanced Raman scattering (SERS) thus became a viable vibrational spectroscopy for surfaces even though the Raman effect is a weak one. The large surface enhancement is primarily caused by the strong local electric-field enhancements from local plasmons associated with small structural features on the surface but also includes some additional local property associated with changes in the electronic states of the system due to chemical bonding.

Second-order nonlinear optical processes are allowed necessarily at surfaces or interfaces, but they are forbidden in media containing inversion symmetry. This specificity makes two second-order nonlinear optical spectroscopies, second harmonic generation (SHG) and sum-frequency generation (SFG), unusually surface sensitive. In SHG one measures a signal emerging from the sample at twice the frequency of the incident light due to a nonlinear optical response of the surface. In SFG one obtains a signal at the sum of the two incident light frequencies (typically visible and infrared) in order to measure vibrational spectra. Both methods allow for the determination of the second-order susceptibility $\chi^{(2)}_{\mathrm{surf}}$, from which one can obtain spectroscopic information and molecular orientation for adsorbed molecules. The high spatial, temporal, and spectral resolutions associated with lasers used in these techniques have led to a number of unique applications for probing surfaces.

2.2 Electron–Surface Interactions

The interactions of electrons with matter are usually characterized by elastic or inelastic scattering. Scattering events may induce electronic excitations that decay via the emission of light or electrons and related secondary processes. The inelastic mean free path of electrons with kinetic energies in the range between 1 eV and several keV traveling through condensed matter is typically smaller than 10 nm (100 Å). Electron spectroscopies based on the detection of electrons at those energies are surface-sensitive techniques. Inelastic-scattering events in solids include collective excitations of the electron plasma (5–20 eV), electronic (single-electron-like intraband and interband) excitations (5–20 eV), and core-level ionization (20 eV–10 keV). Incident electrons below 10 eV efficiently excite lattice vibrations (phonons) and vibrational modes of molecules at surfaces, and this causes inelastic losses that are below 1 eV.

Auger-electron spectroscopy (*q.v.*) (AES) is the most widely used quantitative probe of surface composition. In AES, an incident electron beam of 1–5 keV causes core-level ionization, and the ionized atoms relax efficiently via a two-electron nonradiative transition (the Auger process) in which another electron (the Auger electron) is emitted with an energy characteristic of the ionized atom. This technique is carried out using a finely focused, rastered incident electron beam in scanning Auger-electron microscopy (SAM) in order to obtain elemental composition maps of the surface with <500-Å resolution.

Much of the early information on surface structure and morphology and surface diffusion came from field-emission microscopy (FEM) (see FIELD EMISSION AND FIELD IONIZATION). Field-emitted electrons from a metallic tip are accelerated toward a fluorescent screen to produce a highly magnified image of the tip. The related technique of field-ion microscopy (FIM) typically uses He as an imaging gas, which is ionized by the high electrical field gradient at the sample tip, obtaining higher resolution than FEM. Electron microscopies use a finely focused beam of high-energy electrons rastered across the sample to obtain a highly magnified image of the probed area. Scanning electron microscopy (SEM) and scanning transmission electron microscopy (STEM) reveal details about surface morphology for a wide range of materials, with a resolution down to about 50 Å. Transmission electron microscopy (TEM) has atomic resolution and requires thin films as substrates, i.e., it is not inherently surface sensitive. It can be used to determine surface structure when a cross section of the film is analyzed. Electron tunneling between the sample surface and a conducting tip scanned across the surface by piezoelectric ceramic tubes is the basis for the scanning tunneling microscope (STM), a revolutionary probe of surface structure and morphology.

Electrons are useful in surface modification processes, since electron bombardment at energies exceeding a few electronvolts breaks chemical bonds at the surface. Electron-induced dissociation (EID) refers to the production of new surface species in adsorbed layers by dissociation of surface and adsorbate bonds. Following ionization of an atom or molecule, relaxation (sometimes via Auger decay) causes desorption of surface atoms or adsorbed species in a process called electron-stimulated desorption (ESD) analogous to PSD. In ESD and PSD, molecules are excited to a strongly repulsive part of a dissociative electronic state that leads to ion ejection predominantly along the ground-state bond axis. This forms the basis for the ESD ion angular distribution (ESDIAD) technique, which provides patterns like those in Fig. 3.

2.2.1 Diffraction Diffraction occurs as a special case of coherent elastic scattering of electrons by atoms in a crystalline solid and requires that the de Broglie wavelength of the electron ($\lambda = h/p$) is less than or equal to the lattice constant d. For metals, a typical d value is 3 Å, and so incident electrons with kinetic energies in the 10–200-eV range (λ = 1–4 Å) undergo diffraction. This forms the basis for an analytical technique called LEED. The diffraction (spot) pattern reveals information about the size and orientation of the unit cell of surface structures with long-range order. More detailed analysis of the two-dimensional (2D) structure of the surface is possible when the width of the diffraction spots is measured (spot profile analysis, SPA-LEED). The measurement of the spot intensity I as a function of the incident electron energy (proportional to the gun acceleration voltage V) is called IV-LEED, and dynamical scattering simulations of these curves can provide bond distances at surfaces to within 0.1 Å for small and simple unit cells. In reflection high-energy electron diffraction (RHEED), the electron beam hits the surface at grazing incidence, and the diffraction spots form a semicircle-like pattern around the specular beam spot. RHEED is performed at incident energies of 20 keV and can be used for *in situ* monitoring and investigation of film deposition and growth under vacuum conditions.

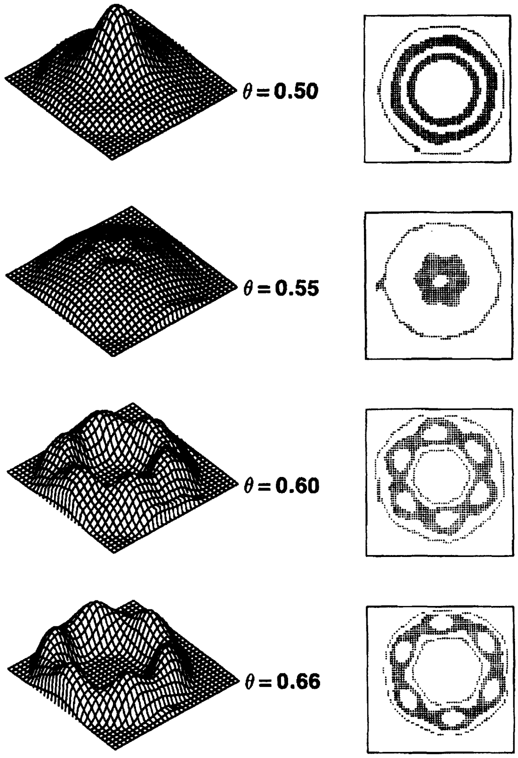

FIG. 3. ESDIAD patterns of CO^+ ions indicating particular tilt directions of the molecular CO axis with respect to the surface due to surface crowding of chemisorbed CO species (from Kiskinova *et al.*, 1988).

2.2.2 Electron Emission Historically, much of surface physics was concerned with electron emission, and electron sources are essential to a wide array of devices, such as vacuum-tube amplifiers, and all electron-beam applications, e.g., electron microscopy. Electrons emitted from a solid have to overcome the work function, which is typically about 5 eV, and so high temperatures (and thus W filaments) are required for thermionic emission of electrons. The coating of these "hot cathodes" with thorium oxide, or the use of LaB_6, reduces the work function and allows operation at lower temperatures. This reduces problems with outgassing, a desorption of adsorbed gases from thermal heating of the surrounding parts. Field emission occurs from very sharp metallic tips via electron tunneling through the narrow potential barrier at the surface caused by the application of a high negative voltage to the tip. The combination of a small tip radius and high potential creates a large electrical field gradient allowing electron emission at very low temperatures ($T \leq 300$ K). These sources are characterized by narrow electron energy distributions and small source sizes. The surface of a negative–electron-affinity (NEA) source emits electrons under the influence of only modest electric fields at low temperatures in a "cold cathode" source. This is achieved by the combination of a low–work-function material, like a Cs-covered Si surface, with a semiconductor device, where the energy level of electrons at the heterojunction (Si-semiconductor interface) lies above the Cs-reduced vacuum energy level of Si. These NEA sources are extremely interesting for a number of technologies like microwave tube applications, high-definition monitors, and flat-panel displays.

2.3 Ion–Surface Interactions

Ions incident on a surface at appreciable kinetic energies (>5 eV) impart significant momentum to the collision partner at the surface, and this energy available from ion scattering causes surface changes. In addition, neutralization via charge transfer upon collision with a surface releases energy into the incident species equal to the ionization potential of the corresponding atom or molecule. Since ions can easily be accelerated, deflected, or focused with electrostatic fields and lenses, a number of surface analytical spectroscopies have been developed using ions, most notably ion-scattering spectroscopy (ISS), Rutherford backscattering spectroscopy (RBS), and secondary-ion mass spectrometry (SIMS). The chemically reactive nature of ions has resulted in a variety of surface-processing technologies involving ions generated in plasmas near the surface or in focused beams.

2.3.1 Scattering The interaction of ions with surfaces depends primarily on the ion kinetic energy. The higher the incident ion kinetic energy, the smaller the distance of closest approach $d_{\min}$ between an incoming ion and a scattering atom. For thermal energies $d_{\min}$ is of the order of interatomic distances, and scattering happens at a potential slightly modulated by the geometry of the surface atoms. For relatively low-energy ions (0.1–1 keV) $d_{\min}$ is smaller than interatomic distances, and ions slightly penetrate into the surface but are predominantly scattered from the topmost surface layer. Severe neutralization of ions in conductors insures that scattered ions have trajectories that only probe atoms exposed at the surface. Ion interactions with surface atoms may be described in terms of a collision process between two isolated atoms of masses m_1 and m_2, under the assumption that the energy transfer is much bigger than the binding energy of the atom within the solid (a few electron volts). Simultaneous conservation of momentum and energy leads to the basic scattering equation that gives the final energy of the scattered ion (relative to the initial energy of the incoming ion) as a function of the mass ratio $A = m_2/m_1$ of target mass m_2 and projectile mass m_1 and the total scattering angle θ:

$$\frac{E_f}{E_i} = \frac{1}{(1 + A)^2}(\cos\theta \pm \sqrt{A^2 - \sin^2\theta})^2 \qquad (4)$$

This relation forms the basis for a surface analytical technique called ISS, also known as low-energy ion scattering (LEIS), that is used to obtain information on the elemental composition and the geometrical arrangement of atoms at a surface. In the simplest application the energy distribution of scattered ions is recorded using a monoenergetic incoming ion beam (0.2–5 keV) and a fixed scattering geometry. Different peaks in the final ion energy distribution belong to scatterers with different masses.

It is also possible to determine the relative geometrical position of surface atoms by utilizing the angular dependence of the scattered-ion intensity. The main phenomena that control the intensity dependence are shadowing and blocking. A shadow cone results in a region behind a scattering target atom where no ions may be found for the allowed ion trajectories. Shadowing (of other scattering atoms) by atoms along the incident ion trajectories and blocking (of the detector) by atoms along the exiting ion trajectories cause large angular variations in the scattered-ion intensities for a given geometrical arrangement of the near-surface atoms when alkali ions (which have low neutralization cross sections) are used.

The analytical technique high-energy ion scattering (HEIS), called Rutherford backscattering spectroscopy (RBS) when using α particles (He^{++} ions) typically at 1 MeV, provides unique structural information about buried interfaces and concentration/depth profiles, yet can still reveal near-surface–sensitive information. The stopping power of matter for α particles and other high-energy ions is well known, and this leads to an absolute determination of concentration in the solid. Since the average kinetic energy of the He^{++} ion that is dissipated (by long-range interactions) is proportional to the distance traveled within the solid, observation of the He^{++} ion energy loss gives a very powerful tool for high-sensitivity, nondestructive depth analysis. If the incoming beam is aligned with a high-symmetry direction of a single-crystal sample, which is often the case for many practical Si-based devices, then the scattering at the surface atoms will give rise to a so-called surface peak with a well-defined energy loss. All other ions will enter deep into the crystal—a process called channeling—and have only a low probability of being reflected into the detector. The reflected-ion intensity, however, will not be zero since scattering still occurs at imperfections or dislocations, or on account of thermal motion of the atom cores about their equilibrium positions. The geometric structure of the solid can be determined since for an arbitrary ion-beam direction all layers will contribute to the reflected ion intensity, which will be much higher than for the channeling directions.

2.3.2 Sputtering and Implantation

Ion–surface collision at ion kinetic energies above a few electronvolts can cause sufficient momentum transfer to eject surface atoms. This removal of atoms is called sputtering; it abrades the surface like an atomic-scale sandblaster. Theoretical simulations that apply energy and momentum conservation provide information on the fundamental processes that occur upon collision of an ion and a target atom, ranging from scattering and defect formation to sputtering. A sample (calculated) ion trajectory is shown in Fig. 4.

Sputtering (*q.v.*) is a common method for the cleaning of surfaces in ultrahigh vacuum (UHV), where the pressure is below 10^{-7} Torr. A beam of typically 2-keV Ar^+ ions with a current density of 10 $\mu A/cm^2$ on a sample would typically remove material from the surface at rates of 20–1000 Å/min depending on the sputtering yield, which is determined by the "hardness" of the material. When more aggressive conditions are used, i.e., very high (milliampere) ion currents, this process is called ion milling and can cause macroscopic changes in the sample morphology. Sputtering is also utilized in depth-profiling analysis of heterogeneous systems, where the surface composition is monitored at the base of an etch pit formed after gradual removal of surface layers. Secondary-ion mass spectrometry (SIMS) is based upon sputtering of the sample. In SIMS, the distribution of positive and negative secondary ions (primary ions are those of the incident beam) ejected from the surface by ion-beam bombardment is recorded.

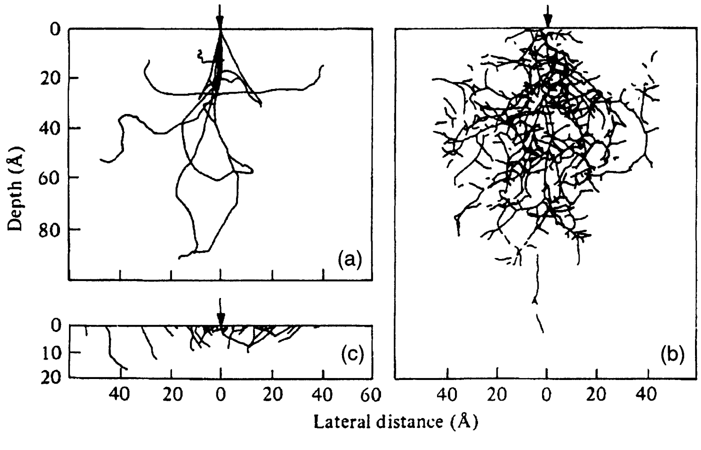

FIG. 4. Monte Carlo simulation of 4-keV Ar^+ ions incident on a target of randomly distributed Cu atoms. (a) 10 incident Ar^+ ion trajectories, (b) resulting Cu recoil distribution, and (c) the Cu sputtering events resulting from 50 incident ions. (From Ishitani and Schimizu, 1974.)

The elemental sensitivity and dynamic range of SIMS are several orders of magnitude higher than for AES and XPS, but huge variations in the neutralization efficiency of the secondary ions means that SIMS is not a quantitative probe in general.

Ion implantation can be carried out at incident ion energies of several MeV to provide a known distribution of ions (which become dopant atoms) at a specific depth within a solid substrate. At very high incident ion energies (100 keV–4 MeV) $d_{\min}$ gets so small that ions can enter deep (several microns) into a solid; the shadow cone becomes very narrow, and neutralization processes are dramatically reduced. This is a crucial commercial manufacturing process since doping of semiconductors is required for the production of microelectronic devices. For example, boron (B) is introduced into Si to produce p-doped material. SIMS using a calibration provided by the known depth profile of high-energy ion implantation, e.g., B in Si, can be quantitative and finds widespread use in quality control in semiconductor fabrication lines.

3. GAS–SURFACE INTERACTIONS

While incoming charged particles (electrons and ions) interact with a surface via a long-range attractive Coulomb potential, gases (neutral atoms and molecules) exhibit much shorter-range interactions with the surface. Four elementary processes in gas–surface interactions are adsorption, diffusion, reaction, and desorption. All surface reactions begin with collision of the gas with the surface or adsorbed layers. Gas localized at the surface is trapped or adsorbed, but still may be mobile in two dimensions and diffuse across the surface. Reactions can occur within a colliding or adsorbed molecule (e.g., isomerization or dissociation) or between coadsorbed species. Desorption is the opposite process to adsorption and results in the liberation of species from the surface into the gas phase. A schematic of these gas–surface interactions is shown in Fig. 5. Analogous processes exist for the L/G interface, and also for the S/L and L/L interfaces, where adsorptive molecules are furnished by the liquid phase(s), but these interfaces have not been studied in as much detail.

3.1 Surface Collisions

Gases impinging on a surface may be either adsorbed or reflected from the surface (not necessarily into a specular direction). The incoming flux Z (number of particles per unit time per unit area) may be rewritten as a function of the partial pressure p of the molecule in the gas phase as $Z = p(2\pi mkT)^{-1/2}$, where m is the mass and k is the Boltzmann constant, and provides an absolute upper limit to the adsorption rate. The following mainly considers collisions at thermal velocities. However, incoming gases with hyperthermal energies (0.5–5 eV) are also of interest. These interactions lead to collision-induced dissociation (CID), where momentum transfer high up on the repulsive potential, the "hard wall," dissociates the colliding molecule, but this dissociation occurs often as the molecule is leaving (has left) the surface region and all of the fragments appear in the gas phase.

3.1.1 Energy Exchange The adsorption of a gas requires minimally the loss of incident translational energy perpendicular to the surface, and this requires energy exchange with the surface. Even for gases not adsorbed, the scattered gas will generally leave the surface with a somewhat different energy and momentum as a result of the interaction. If T_i and T_f denote the temperatures in the gas phase before and after the collision, respectively, and T_s is the temperature of the surface, then the accommodation coefficient, α_T, is defined by

$$\alpha_T = \frac{T_i - T_f}{T_i - T_s}. \tag{5}$$

If there is little or no interaction between the gas and the surface, the temperature in the gas phase will not change much and α_T will be near zero. If, on the other hand, the gas resides on the surface for a time that is large compared to molecular vibration times (i.e., the gas–surface interaction energy is large), the molecule will leave the surface with $T_f = T_s$, and α_T becomes 1. Analogous expressions may be defined separately for the energy accommodation coefficient in each of the individual degrees of freedom of the gas, and quantum-state–specific information about surface-mediated energy transfer and relaxa-

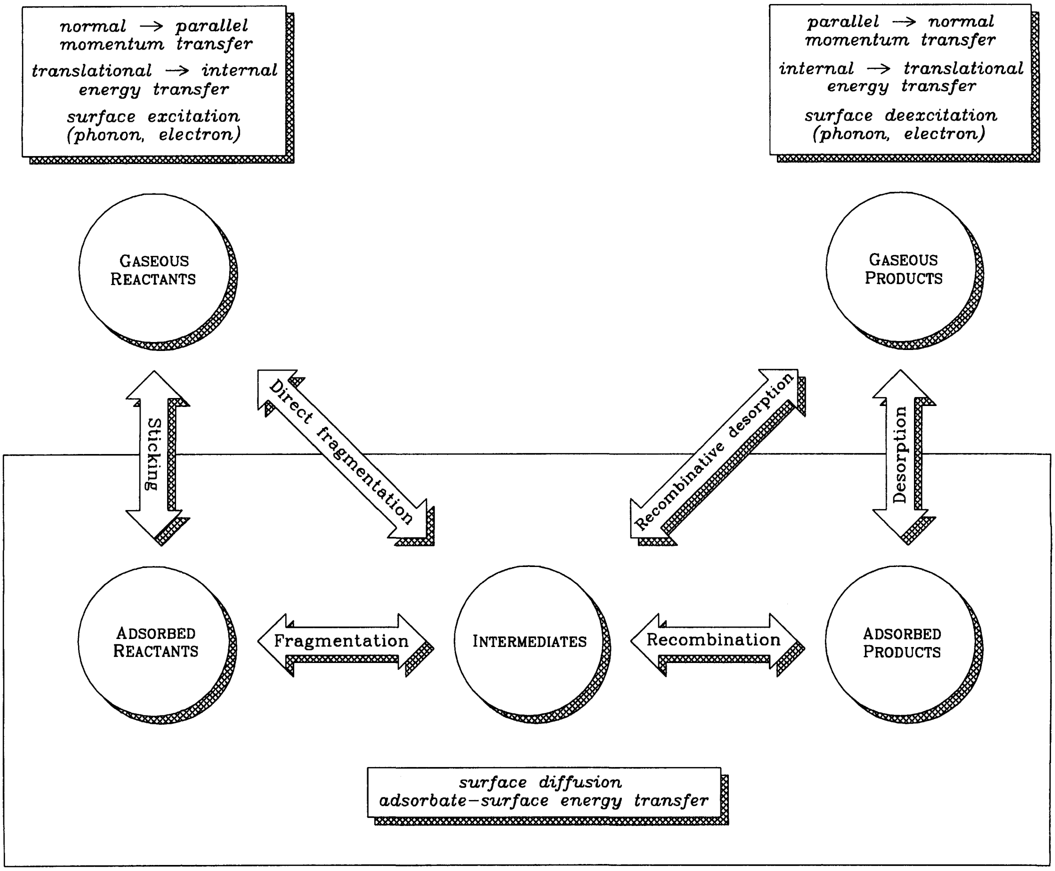
normal —> parallel momentum transfer
translational —> internal energy transfer
surface excitation (phonon, electron)
parallel —> normal momentum transfer
internal —> translational energy transfer
surface deexcitation (phonon, electron)
GASEOUS REACTANTS
GASEOUS PRODUCTS
Sticking
Direct fragmentation
Recombinative desorption
Desorption
ADSORBED REACTANTS
Fragmentation
INTERMEDIATES
Recombination
ADSORBED PRODUCTS
surface diffusion
adsorbate–surface energy transfer

FIG. 5. Possible steps for a surface-mediated reaction (after Tully and Cardillo, 1984).

tion of vibrational, rotational, and translational energy of incident molecules is emerging.

3.1.2 Atomic- and Molecular-Beam Scattering Atoms with thermal energies (<0.1 eV) do not penetrate into the surface but are deflected at some distance above the surface. This atom scattering can be used as a surface structure probe. An atom beam, mostly using thermal-energy He gas, can be made monoenergetic with the aid of a nozzle beam, a skimmer, and a rotating chopper disk, and its energy can be measured with time-of-flight (TOF) techniques. This experimental setup allows measurements of the diffraction pattern from ordered hydrogen adatoms, which are usually hard to detect by electron diffraction, and a determination of the adsorbed layer structure. One can also measure the amount of the specularly reflected beam intensity compared to the diffuse scattering intensity as a function of changing surface conditions. For example, a good-quality, closed-packed fcc (111) surface shows almost no diffuse scattering, so that the intensity of the scattered atoms appears all in the specular beam. However, a surface defect like a vacancy or an adatom leads to a decrease of the specularly reflected beam intensity I_{SR}. Since I_{SR} depends nonlinearly on the number of defect structures, atom scattering is a very powerful tool for the investigation of the distribution of defect structures, growth and diffusion mechanisms, and 2D phase transitions of adsorbates.

The dynamics of gas–solid interactions can be probed by using a similar experimental setup to that for atom scattering in molecular-beam scattering (MBS) studies. In this case, a beam of molecules with well-defined parameters like kinetic energy, vibrational or rotational excitation, and in some cases even molecular orientation and alignment is directed toward a surface. Moreover, the angular and energy distribution of the reflected beam can be monitored via mass spectroscopy and other laser-based spectroscopic probes, such as laser-induced fluorescence (LIF), to determine the population of specific quantum states of the scattered product to elucidate specific energy-transfer processes occurring at the surface. This yields detailed information on elementary interactions such as energy accommodation, adsorption mechanisms, differential reaction cross sections, and the potential energy surface describing the gas–surface interaction. Activation energies for adsorption and reaction can be probed by varying the translational energy (and that of other internal degrees of freedom such as vibrational energy) of the incident molecules and the surface temperature.

3.2 Adsorption

The thermodynamics of adsorption is often evaluated phenomenologically in terms of an adsorption function $n = n(p,T)$, where n is the number of adsorbed species (per unit area) at a given temperature T and partial pressure p of the species in the gas phase. Whenever the system reacts reversibly upon changes in p or T, i.e., whenever n is a function uniquely defined by p and T, the adsorption system can be described by equilibrium thermodynamics. In analogy to gas-phase systems $n = n_T(p)$, $n = n_p(T)$, and $p = p_n(T)$ are then called adsorption isotherms, isobars, and isosteres, respectively. The shape of the isotherm was used historically to supply evidence about the interactions in the system. A plot of the isotherm is also a convenient form for the adsorption data to compare with theoretical predictions based on the assumptions in various microscopic models.

An expression for the simplest isotherm can be derived by assuming that the surface is homogeneous and the adsorbed species show no interaction with each other so that the adsorption energy is a constant. At equilibrium, the adsorption rate ($r_a = k_a p$), which is proportional to the partial pressure p in the gas phase, is equal to the desorption rate ($r_d = k_d\theta$), which is proportional to the concentration (coverage) θ of species on the surface, and this may be written in the form $\theta = kp$, with the constant $k = k_a/k_d$ being characteristic for the particular adsorption system. The resulting isotherm is simply a straight line in a plot of coverage versus partial pressure. This so-called Henry's-law isotherm has been found experimentally for only a few adsorption (physisorption) systems at very low coverages, and such behavior is not expected to hold for many real adsorption systems on the basis of the assumptions used in the derivation.

Consider now the surface to have a total number of adsorption sites n_0, of which n are occupied and hence $n_0 - n$ are empty. The adsorption rate r_a should then be proportional to the partial pressure p of the adsorptive in the gas phase and the number of empty adsorption sites, and the desorption rate r_d will be proportional to the number of adsorbed species, θ. At equilibrium the adsorption and the desorption rates are equal, and this results in the Langmuir adsorption isotherm

$$\theta = \frac{kp}{1 + kp} \quad \text{or} \quad \frac{1}{n} = \frac{1}{kn_0 p} + \frac{1}{n_0}. \qquad (6)$$

This expression implies that for low p the adsorbed amount θ is proportional to p (this gives the Henry isotherm) and that at high pressures n reaches the saturation value n_0 (i.e., $\theta = 1$). The right-hand form shows that n_0 and k may be determined from the slope and intercept of a plot of $1/n$ versus $1/p$. An independent derivation from statistical thermodynamics results in a functionally identical relation. Both derivations may formally be made equivalent if the internal partition functions of the molecule in the gas phase and on the surface are assumed to be the same. In general this condition is not met because the adsorption process will at least affect the rotational degrees of freedom on account of the interaction with the surface, and the Langmuir adsorption isotherm has to be modified for a more accurate description.

One can easily derive a Langmuir isotherm also for the case of dissociative adsorption. Considering the adsorption of

A_2(gas) at an adsorption site on the surface, denoted as *, to form 2A(ads) as in

$$A_2(\mathrm{gas}) + 2^* \leftrightarrow 2\mathrm{A(ads)}, \tag{7}$$

the Langmuir isotherm for dissociative adsorption then becomes

$$\theta = \frac{k\sqrt{p}}{1 + k\sqrt{p}}. \tag{8}$$

Brunauer *et al.* (1938) (BET) have extended the approach by Langmuir to the important case of multilayer adsorption. They assumed a Langmuir-type behavior for each layer, with the heat of adsorption being Q for the first layer and Q_v (heat of condensation of the liquid layer) for the following layers, with a further assumption that evaporation and condensation can only occur from or on exposed surfaces. With just one parameter it is possible to determine surface areas from adsorption data and knowledge of the adsorbate size (cross-sectional area). Although this BET isotherm does not describe the experimental data over the whole pressure range, it has found widespread application for commercial rapid determination of the surface area of solids. More refined model isotherms have been derived, but they also introduce one or more additional parameters.

Surface lifetimes are key factors in a wide variety of surface processes. Under the limiting conditions of Henry's isotherm, n may be rewritten as $n = Z\tau$ with τ denoting the lifetime of the molecule on the surface. In the presence of attractive forces the lifetime is given by $\tau = \tau_0 \exp(Q/kT)$, where $\tau_0 = h/kT$ is a characteristic vibrational time (e.g., $\tau_0 \approx 10^{-13}$ s for $T \approx 500$ K), Q is the enthalpy of adsorption, and T is the substrate temperature. Thus, τ is very short for small values of Q and high T, and increases strongly with increasing Q and decreasing T.

The number of adsorbed species n can be related more generally to the incoming gas flux Z with the introduction of a sticking coefficient S defined as $S = r_a/Z$, where the adsorption rate $r_a = dn/dt$. Recalling that Z may be written as a function of the partial pressure p of the molecule in the gas phase, S is determined from the experimental measurement of the adsorption rate as a function of the partial pressure. Expressions for S are generally of the form $S(\theta,T) = S_0(T)S(\theta)$, with S_0 denoting the initial sticking coefficient on the clean surface and $S(\theta)$ accounting for the influence of coadsorbed (see below) molecules on the adsorption (uptake) rate. S_0 is a temperature- and structure-sensitive characteristic of the molecule–surface bonding, with values ranging from 10^{-8} to 1. Simple expressions for $S(\theta)$ are $1 - \theta$ and $(1 - \theta)^2$ for nondissociative and dissociative Langmuir adsorption, respectively. The presence of weakly bound "precursor" adsorption states allows for diffusion prior to desorption and causes values of S on adsorbate-covered or chemically modified surfaces to be higher than that predicted by Langmuirian kinetics. At low coverages, S can even be a constant, i.e., independent of the adsorbate coverage.

Coadsorption, the presence of two or more species on a surface, is a common requirement for surface reactions. A second species in the adlayer also alters the local environment of adjacent adsorption sites and modifies the reactivity of the surface by simply "blocking" (occupying) potentially reactive sites and/or by directly or indirectly altering the electronic structure of the surface. For example, submonolayer coverages of alkali metal atoms on transition metals redistribute electron density at the surface and cause a large decrease of the work function. This affects the probability and extent of electron-transfer reactions at the surface and the nature and energetics of adsorption of coadsorbed molecules. The influence of such "modifiers" on reactions at the surface is the origin of promotion (enhancement) and poisoning (reduction) in heterogeneous catalysis and affects the designed modification of surface properties.

3.2.1 Physisorption and Chemisorption

The nature of the gas–surface adsorption interaction distinguishes two processes: physisorption and chemisorption. Physisorption is adsorption on a surface caused by long-range electrostatic, van der Waals forces, i.e., weak, nondirectional interactions between closed-shell atoms and molecules. This includes dipole ↔ dipole, dipole ↔ induced dipole, and dispersion forces (induced dipole ↔ induced dipole) with an interaction potential U proportional to $1/r^6$. Interaction energies are typically lower than 30–40 kJ/mol, being only a little larger than the heat of vaporization of the bulk (condensed) phase of the adsorptive. Physisorption forces operate for any system, including on top of a chemisorbed layer (giving rise to multilayer adsorption). There is minimal evidence for perturbations of the electronic states of adsorbate and adsorbent, the adsorption is not an activated process, equilibration is rapid and often limited by transport processes, and the coverage responds to changes in the partial pressure p and the temperature T.

On the other hand, chemisorption involves chemical bonding interactions (covalent, ionic, or dative) usually greater than 40 kJ/mol, i.e., it is chemically specific and therefore depends strongly on the surface and adsorbate identity and structure. The adsorbate occupies specific locations (adsorption sites) on the surface. This does not necessarily imply that the adsorbate is frozen at these sites, because mobile adsorbates can diffuse across the surface, but as a time average the adspecies will be preferentially located in specific adsorption sites and not be randomly distributed. The chemical nature of the adsorptive gas may be altered by surface dissociation or reaction (and this may not be reversible), and changes in the electronic states of the substrate surface are often detectable. The elementary adsorption step is often activated (for dissociative adsorption), and a true equilibrium is only slowly achieved if the activation energy is large. The saturation coverage (maximum surface concentration in the monolayer) is on the order of $\theta = 10^{14}$ molecules/cm^2. This is obtained even at very low pressures and is insensitive to small changes in p and T.

3.2.2 Potential-Energy Diagrams The interaction of a molecule and a surface is usually discussed in terms of a one-dimensional (1D) potential function $U(r)$, which is similar to a pairwise potential for two molecules. The Lennard–Jones potential

$$U(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{9}$$

can be expressed with only two parameters, ϵ and σ, and is a reasonably realistic potential for nonpolar, spherical molecules. It can be used to describe adsorption interactions, with r set now equal to the height of the adsorbate above the surface. For a complete description of adsorption, physisorption and chemisorption interactions must be considered, and a discussion of dissociative chemisorption must consider the interactions of both the molecular and dissociated species. Two potential curves describe the two processes, each with a different interaction energy (well depth) and height of the adsorbate above the surface (equilibrium bond distance). Consider the dissociative adsorption $A_2(\text{gas}) \rightarrow 2\ A(\text{ads})$ as depicted in Fig. 6. Incoming molecules usually follow an adiabatic pathway along the lowest-energy curve, passing through the molecularly chemisorbed well and on into the dissociatively chemisorbed well if energetically possible. Commonly, the plot of the potential energy versus the distance z perpendicular to the surface will exhibit a local potential maximum between the minima of these two chemisorption wells. The position of this maximum with respect to the reference potential, where the free molecule is at some large separation from the surface, determines the activation energy E^* for dissociative chemisorption. If E^* is negligible, i.e., the temperature coefficient of the adsorption rate is very small, the chemisorption is said to be unactivated and adsorption rates can be high even at low temperatures. Activated adsorption is required if there is a significant E^*, and adsorption rates are low at modest temperatures and strongly temperature dependent. This energetic barrier results from a correlation of the dissociatively chemisorbed species to the free molecular fragments (or to an excited state of the free molecule in cases of activated nondissociative chemisorption where a barrier emerges between the physisorbed and chemisorbed wells).

These diagrams are also useful to discuss the system thermochemistry, showing clearly endothermic or exothermic adsorption enthalpies, Q, the activation energies for the respective desorption processes, and derived surface-adsorbate bond strengths. Consider some experimental results on hydrogen adsorption on transition metal surfaces (Gasser, 1985). The dissociative chemisorption of H_2 on Pt surfaces is not activated, i.e., $E^* \approx 0$, and so the measured activation energy for desorption corresponds to the heat of ad-

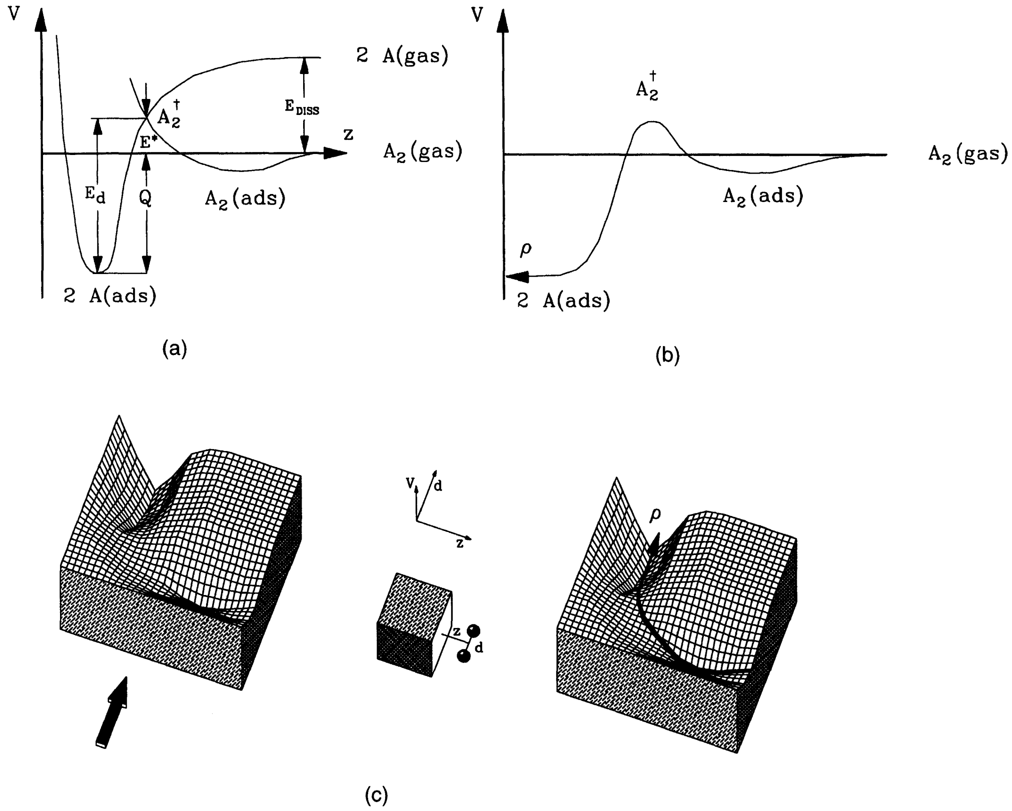

(a)

(b)

(c)

FIG. 6. Schematic potential energy diagrams for a diatomic molecule A_2 interacting with a surface to form an adsorbed A_2 molecule, A_2(ads), and dissociating to form two adatoms, 2A(ads), with its bond axis (bond length *d*) parallel to the surface at distance *z*. (a) 1D Lennard–Jones diabatic curves *V*(*z*), where *z* is the distance of the adsorbate perpendicular to the surface, showing the correlation between gaseous and adsorbed species, (b) variation of the potential energy *V*(*ρ*) along an adiabatic path for the reaction coordinate *ρ*, and (c) 2D potential-energy surface *V*(*z*,*d*), where *d* is the A_2 intramolecular bond length.

sorption. For the dissociative chemisorption of H_2 on copper, $E^* > 0$ and the energy required for the desorption process is given by $E_d = Q + E^*$. Endothermic ($Q > 0$) dissociative chemisorption of H_2 on gold results in a desorption energy given by $E_d = E^* - Q$. In this latter case it is not possible to form H(ads) from H_2(gas), but one still can still form H(ads) from gas-phase H radicals or from sufficiently activated (excited) hydrogen molecules H_2^*(gas).

These 1D models may be used to evaluate some of the basic interactions, but a more accurate description of surface processes and dynamics of surface interactions certainly has to include more than one coordinate. A plot of the potential energy versus both the distance *z* of the molecule from the surface and the interatomic separation *d* within the molecule gives a 2D potential energy surface. The local minimum at large *z* corresponds to the molecular precursor state, the local maximum to an activated intermediate complex and the valley at low *z* to the dissociated state. Analogous plots are used to describe the progress of chemical reactions (see Sec. 4) as introduced by Eyring and Polanyi (Eyring *et al.*, 1974–1975). The dynamics of surface reactions are best discussed in terms of these 2D potential-energy surfaces. Trajectories on the surface are associated with a continuous transformation of energy between the different degrees of freedom of the adsorbed complex, and one can evaluate the influence of translational, rotational, and vibrational excitations in the incident gaseous molecules on reaction probabilities and paths.

3.3 Diffusion

The diffusion coefficient $D(B)$ for moving particles B on a surface A is defined by Fick's first law:

$$J(B) = -D(B)\frac{\partial[B]}{\partial x}, \qquad (10)$$

where $J(B)$ is the flux (in one dimension) from diffusion due to a coverage gradient (density gradient in two dimensions) $\partial[B]/\partial x$. Diffusion of surface atoms or molecules at moderate temperatures is an activated process and experimental results are often expressed in the form

$$D(B) = D_0 \exp(-E_m/kT), \qquad (11)$$

where D_0 is the preexponential factor and E_m is the activation energy for diffusion (migration). The energetic barriers to diffusion are usually a small fraction (0.1–0.5) of the activation energy for desorption or sublimation, and so diffusion occurs at much lower temperatures than desorption or sublimation; and many surface species are mobile at room temperature or above. Adatom mobility on semiconductor surfaces is smaller because of the highly directional covalent bonding present that results in nearly equal activation-energy barriers for diffusion and desorption.

Two different microscopic mechanisms for activated diffusion may be distinguished: a hop-site mechanism and a place-exchange mechanism, both illustrated in Fig. 7. In the former, most common situation (necessarily for admolecules) diffusion occurs by the particle hopping from site to site (minima in the surface or adsorption potential energy wells) with a nearly instantaneous transition over an energetic barrier that separates the well minima in two dimensions and normally coming to rest in a new equivalent position one lattice site away. In the second mechanism, atoms insert into interstitial sites, and migration occurs via displacement of an adjacent atom. This involves a local strain deformation of the surrounding atoms along the diffusion path (Kellogg, 1994).

Within the context of the hop-site mechanism, the mean intersite jump frequency Γ may be expressed by absolute rate theory as

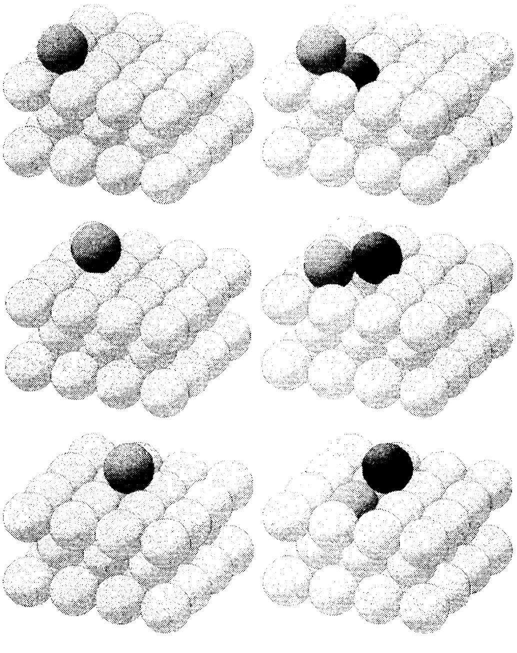
(a) (b)

FIG. 7. Two mechanisms for surface diffusion (from top to bottom). (a) Hop-site mechanism. (b) Place exchange mechanism.

$$\Gamma = \frac{kT}{h}\exp\left(\frac{\Delta S_m^\dagger}{k}\right)\exp\left(\frac{-E_m}{kT}\right) \qquad (12)$$

assuming unity transmission coefficient at the barrier, where $\Delta S_m^\dagger$ is the activation entropy and E_m is the intersite activation energy barrier. The diffusion coefficient is related to the jump frequency by $D = a^2\Gamma/2\alpha$, where a is the intersite jump distance and $\alpha = 1$ or 2 for diffusion in one or two dimensions, respectively. Assuming a random walk, the mean square displacement $\langle x^2\rangle^{1/2}$ in time t is given by

$$\langle x^2\rangle^{1/2} = (\Gamma t)^{1/2}a. \qquad (13)$$

The preexponential D_0 in Eq. (11) can be determined from Eqs. (11) and (12) to be $D_0 = \nu a^2/2\alpha$, where ν is the frequency factor in Eq. (12), i.e., $\nu = (kT/h)\exp(\Delta S_m^\dagger/k)$. A rough estimate of the value of D_0 can be made by assuming $\Delta S_m^\dagger = 0$, with $D_0 \approx 2 \times 10^{-3}$ cm^2 s^{-1} for $T = 500$ K, $a = 3$ Å, and $\alpha = 2$. The efficiency of this process also depends on the number of vacancies, in addition to the frequency of jumping to a neighboring vacancy, and so one expects a coverage-dependent value of D_0 in addition to a coverage-dependent value of E_m due to lateral interactions.

Experimental measurements of diffusion mechanisms (including the role of tunneling), diffusion coefficients, and diffusion activation energies on well-defined single-crystal surfaces have now been performed utilizing many different surface spectroscopies. Most of our information on metal-on-metal diffusion has come from observation of individual adatoms using FIM. These studies revealed that $E_m \approx 0.3E_d$. Direct adatom observation can now also be carried out by STM. Surface diffusion of adsorbates has mainly been studied with LITD using coverage-gradient methods in which a sharp boundary between clean and adsorbate-covered regions is formed at low temperatures, and a surface probe is utilized to follow diffusion across the boundary as the temperature is raised. Now, diffusion can also be probed by grating methods utilizing lasers and other structure-sensitive probes like atom scattering.

Diffusion is a dominant mechanism for material transport and can be a rate-determining step in a variety of processes. Reconstructive phase transitions, annealing of surface roughness, segregation, and crystal growth are processes that all involve material (atom, molecule, etc.) transport. In many practical examples, as in polycrystalline materials, diffusion of atoms along grain boundaries occurs simultaneously with diffusion across the lattice, and this complicates the observed kinetics from that discussed above.

3.4 Desorption

Desorption is the removal of adsorbed species from a surface into a free adsorptive phase. Adsorption/desorption equilibrium is established when the desorption rate (always an activated process) is equal to the adsorption rate. Increasing the surface temperature increases the desorption rate by providing more thermal energy (primarily due to phonon coupling with the degrees of freedom of the adsorbate-surface complex) that can be used to overcome the desorption activation energy barrier.

The desorption rate expression can be written in the form

$$r_d = k_d \theta^n, \tag{14}$$

where k_d is the desorption rate constant, usually given by an Arrhenius expression $k_d = \nu_n \exp(-E_d/RT)$, θ is the surface coverage of desorbing species, and n is the kinetic order of the desorption reaction. For sublimation of multilayer films, $n = 0$, while $n = 1$ for atomic or molecular desorption from an adsorbed monolayer, and $n = 2$ for desorption processes where bimolecular recombination at the surface occurs prior to desorption. The desorption activation energy E_d often shows strongly coverage-dependent changes, reflecting lateral interaction effects of adsorbates on adsorption energies. While this expression describes the kinetics of the desorption process, the dynamics of the event is not included. Detailed dynamical information now exists for a variety of desorption processes, and atoms and molecules are often observed to emerge from the surface without being fully equilibrated to the surface temperature. This is manifested by noncosine angular distributions and translational, rotational, and vibrational energy-level populations that do not reflect the surface temperature and are not necessarily described well by Boltzmann distributions. This is especially true for the desorption of the products of exothermic surface reactions.

Temperature-programmed desorption (TPD) or thermal-desorption spectroscopy (TDS), in which the surface temperature is increased linearly with time, is a widely used technique in surface science to determine activation barriers for desorption or surface reaction and binding energies of adsorbates in specific adsorption sites. During the temperature ramp, a mass spectrometer is used to monitor the desorption flux of previously adsorbed species, determining the identity of all desorbing gases and measuring their desorption rates.

Energy can be provided to the adsorbate-surface complex by other (than thermal) means, and this can be used to overcome the activation barrier to desorption. These other excitation mechanisms are more specific in their activation of adsorbed molecules compared to thermal desorption and give rise to

ESD, PSD, ion-induced desorption, and collision-induced desorption.

4. SURFACE REACTIONS

The simplest surface reactions at the gas–solid interface are dissociative chemisorption, in which bonds are broken in the incoming molecule to produce surface-bound atoms, and adatom recombination, in which bonds are made between surface-bound atoms. Surface reactions can be divided into two primary categories. In the first case, the solid surface catalyzes a chemical reaction but does not undergo any net chemical change itself, e.g.,

$$CO(g) + H_2O(g) \xrightarrow{Fe_2O_3 \cdot Cr_2O_3} CO_2(g) + H_2(g). \tag{15}$$

Heterogeneous catalysis relies on this class of reactions to carry out the majority of the processes required for the chemical and petroleum industries. In the second case, the solid surface is directly involved in the reaction by compound formation. Reactions like

$$Ni(s) + CO(g) \rightarrow NiC(s) + NiO(s) \tag{16a}$$

are involved in describing oxidation and corrosion, while reactions like

$$Ni(s) + 4CO(g) \rightarrow Ni(CO)_4(g) \tag{16b}$$

are typical for etching processes. The elementary processes described in Sec. 3 are important for both categories of reactions.

4.1 Heterogeneous Catalysis

Catalysis occurs by the acceleration of the rates of chemical reactions without net consumption of the catalyst and without any influence on the overall reaction thermodynamics (equilibrium balance). Heterogeneous catalysis means that catalyst, reactants, and products involve more than one phase. This is the case for the majority of industrial catalysts, where the catalyst is a solid (surface) and the reactants and products are in gaseous or liquid phases. A catalyst is characterized by its activity—the rate of a catalyzed surface reaction—and its selectivity—the tendency of one desired reaction product to be produced over all others. Dispersion is the term used to define the fraction of total catalyst atoms that are present at the surface. Most catalysts are characterized by a very high dispersion of the active phase to increase the contact area between the catalyst and the reactant/product fluid. In the case of metal-based catalysts, this usually occurs by the formation of small separated crystallites (<20 Å) on a porous, high–surface-area oxide support (which also greatly reduces sintering—the agglomeration of small particles to form a larger one), typically leaving more than 50% of the active catalyst exposed at the surface.

Three major types of catalysts may be distinguished:

1. electronic—oxidation and reduction,
2. acidic—ionic, and
3. structural—molecular sieves.

Oxidation and reduction catalysts usually involve electron transfer to or from the solid surface. Solids with free or easily excited electrons, like transition-metal surfaces (Fe, Pt, Ni) or semiconductors (NiO, ZnO) are likely catalysts. Metal surfaces commonly exhibit spontaneous H_2 and O_2 chemisorption and are the preferred catalysts for (de)hydrogenation, reduction, or oxidation. The activity of catalysts using different transition metals varies over several orders of magnitude, e.g., eight for the hydrogenolysis of ethylene ($CH_2=CH_2 \rightarrow CH_4$). Often this variation is described by a "volcano" plot in which one transition metal is the best (at the top of the volcano) and metals to the left or right in the periodic table have lower activity. Ionic catalysts, like the insulators Al_2O_3 or SiO_2, with acidic functional groups are used for cracking and isomerization (skeletal rearrangements) that occur via carbonium ion mechanisms common in organic chemistry. Shape-selective catalysts, like zeolites, have high internal areas (100–500 m^2/g) and well-defined pore structures with sizes on the order of molecular dimensions. The control exerted by the zeolite on diffusion of reactants and products through the pores significantly alters the overall activity and selectivity from that of the acidic functional groups of the zeolite itself (or of the metal catalyst particles placed into the zeolite cages). Finally, catalysts contain (usually but not always by design) chemical additives that dramatically modify the observed activity and selectivity. These are called promoters or poisons depending on the desirability of their action.

In part, surfaces enhance reaction rates simply on account of dimensionality (adsorbed reactants are confined to movements within two dimensions only). Of course the main effect of a catalyst is to lower rate-limiting energetic barriers due to stabilization of transition states and reaction intermediates, as shown in Fig. 8.

Two main reaction processes are distinguished in heterogeneous catalysis (Adamson, 1990). The Langmuir–Hinshelwood (LH) mechanism

$$A(\text{ads}) + B(\text{ads}) \rightarrow \text{Products} \qquad (17a)$$

involves the reaction of two adsorbed species. A simple reaction rate law can be formulated that is proportional to the surface coverages $\theta(A)$ and $\theta(B)$, and these can be expressed by Langmuir adsorption isotherms. The Eley–Rideal (ER) reaction mechanism

$$A(\text{ads}) + B(\text{gas}) \rightarrow \text{Products} \qquad (17b)$$

takes place as the direct result of a collision between an impinging gas molecule and a molecule adsorbed on the surface. In this case, the rate expression can be given as proportional to the surface coverage $\theta(A)$ and the pressure of B. The occurrence of these two mechanisms in a variety of surface reactions has been debated over the years, with the result that nearly all thermal surface reactions are thought to proceed via a LH mechanism and coadsorbed reactants. However, recently several reactions have been demonstrated in carefully controlled studies to occur via an ER process (Rettner and Auerbach, 1996).

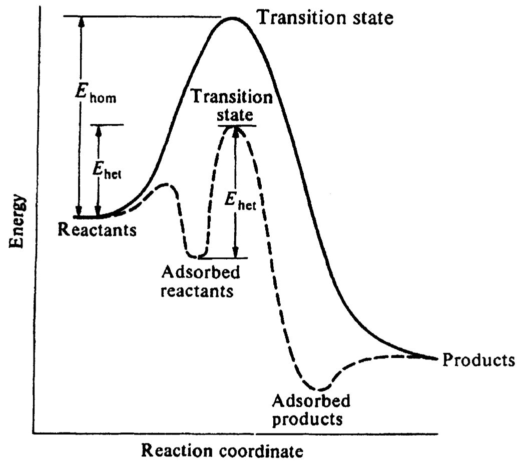

FIG. 8. Schematic representation according to transition-state theory of a homogeneous reaction (solid line) and a heterogeneous reaction (dashed line), illustrating how catalysts increase reaction rates by lowering the energetic barrier to reaction (from Gasser, 1985).

A surface-science approach to catalysis offers the possibility of investigating catalytic surface processes at a molecular level. The most powerful probe of reaction mechanisms is modulated molecular-beam reactive scattering (MBRS), where an incident molecular beam is chopped and the TOF of incident molecules is determined by phase-shift measurements. These time-dependent studies are capable of differentiating between different reaction mechanisms by measuring (through the phase lag) the residence times of reacting molecules on the surface. Model systems have been studied to understand the role of individual reaction steps in the overall activity or selectivity of a given catalyst system. A challenge for current research is to bridge the "pressure gap" that exists between UHV surface-science studies and the high pressures (1–100 atm) of commercial chemical processing. Providing the basic physicochemical foundation for these processes is important for the many catalytic challenges that lie ahead, e.g., the activation of methane (CH_4) to produce useful fuels and chemicals from the vast quantities of CH_4 that are available.

An example that illustrates the interplay between elementary surface processes and their effect on the catalytic reaction mechanism and the observed rate law is the catalytic reaction of NO and CO over Rh. This is a key reaction for control of pollution from engine exhaust gases. One proposed mechanism on Rh(100) surfaces is shown in Table 3 (Hendershot and Hansen, 1986). The rate law shown for the overall reaction is "simple" for a first-order rate-determining step (rds) even though the reaction mechanism is rather complicated. If the dissociation of NO (step 3) is the slow step, then it is straightforward to derive the LH rate law given.

Table 3. Mechanism of the NO + CO reaction on Rh(100) surfaces. Overall rate is

$$r = \frac{d[\mathrm{CO_2}]}{dt} = \frac{k_3K_2p(\mathrm{CO})}{1 + K_1p(\mathrm{CO}) + K_2p(\mathrm{NO})}.$$

	Reaction	Rate constant
(1)	CO(gas) ↔ CO(ads)	K_1
(2)	NO(gas) ↔ NO(ads)	K_2
(3)	NO(ads) → N(ads) + O(ads)	k_3 (rds)
(4)	2N(ads) → N_2(gas)	k_4
(5)	CO(ads) + O(ads) → CO_2(gas)	k_5

Surface-science studies have elucidated a detailed reaction mechanism for the NH_3 synthesis reaction (Ertl, 1983), as shown in Fig. 9. The reactants N_2 and H_2 chemisorb on the Fe surface, surface intermediates are formed that sequentially hydrogenate to form NH_3, and the products desorb and free the Fe surface for further reaction. In this case, the catalyst greatly speeds up the reaction via lowering the barrier to the dissociation of dinitrogen (bond dissociation enthalpy = 945 kJ/mol or 226 kcal/mol) in the gas phase by forming stable Fe–N bonds for adsorbed nitrogen adatoms.

4.2 Oxidation and Corrosion

Topologically, the growth of surface oxide films involves elementary reactions at the oxide/adlayer interface, inside the oxide layer, and at the metal/oxide interface. In the oxidation of metals, the oxygen-containing molecule (typically O_2) first has to chemisorb dissociatively on the surface. Adsorbed oxygen may then migrate on the surface in the form of a polarized or charged adparticle, may occupy a subsurface site, or may diffuse into the crystal. Rearrangement of surface atoms occurs with dissolution of oxygen into the near-surface region, and oxide islands nucleate, grow laterally, and eventually coalesce to produce a continuous oxide film. Growth of surface oxide films is dependent on the diffusion of cations, anions, and electrons through the oxide layer. Already in 1949, Cabrera and Mott (1949) developed a mechanism that described the growth kinetics of very thin (10–50 Å) films in which electrons tunnel through the oxide film to produce oxide ions from the oxygen atoms at the surface and induce a large electric field gradient at the surface, which enhances cation diffusion into interstitial sites of the oxide from the metal-oxide interface. Typical rate laws for the film thickness d with oxidation time t show a logarithmic dependence, $d_{ox} = k_1\log(k_2t + k_3)$ (low-T oxidation of Fe), parabolic dependence, $d_{ox}^2 = k_1t + k_2$ (high-T oxidation of Fe), or unimpeded dependence, $d_{ox} = k_1t + k_2$ (light metals, like Na, Ca, Mg).

Many other oxidation reactions involving halogens, chalcogenides, and carbon-containing molecules proceed to form surface compounds on solids. Examples include the formation of metal chlorides, sulfides, and carbides following chemisorption of Cl_2, H_2S, and CO, respectively, the conversion of CuO catalysts to $CuCl_2$ upon exposure to HCl, and the formation of volatile fluorides of Si following exposure to XeF_2 (etching).

Corrosion reactions on metals are extensive oxidation processes consuming large amounts of the metal, and corrosion often refers to the loss of material from the solid due to dissolution or flaking of the oxidation products at the surface. Corrosion is an elec-

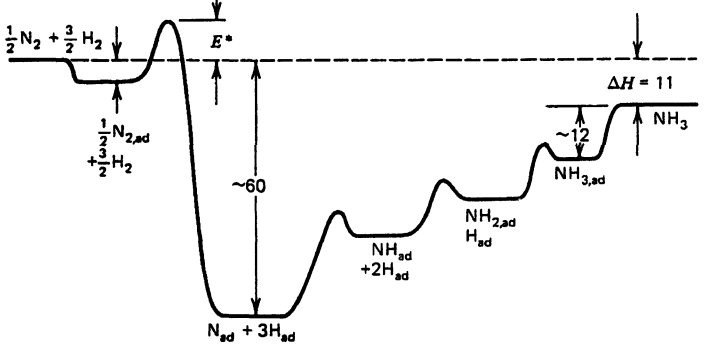

FIG. 9. Potential-energy diagram showing the thermochemistry of catalytic NH_3 synthesis over Fe surfaces. The activation energy E^* for the rate-limiting reaction step actually depends on the Fe surface structure and the amount of adsorbates (coverage) on the surface. Energies are given in kcal/mol. (From Ertl, 1983.)

trochemical process, as shown in Fig. 10, where O_2 and liquid water are depicted as essential ingredients for the rusting of iron. In some cases, the oxidation products may themselves offer corrosion protection. The oxidation of aluminum alloys, extensively used in aerospace, commercial, and semiconductor industries, produces a 20 Å film of Al_2O_3, which forms a protective coating against further oxidation below 400 °C. Stainless steels, known for their high strength and superior workability, obtain their corrosion resistance by the formation of a thin passivating oxide layer of Cr_2O_3. In most cases, however, corrosion inhibition is achieved by added adsorbates, such as in the use of imidazoles to inhibit copper corrosion, or by dense polymeric protective films (paints and coatings).

5. DEPOSITION AND SURFACE PROCESSING

Deposition and surface modification processes are used extensively in the semiconductor industry for the production of solid-state microelectronic devices. In this section we briefly describe atomic processes occurring during crystal growth, focusing on nucleation and growth mechanisms in deposition processes for the production of epitaxial thin films, and comment on important applications in etching, lithography, and the emerging field of nanotechnology. The interested reader is referred to the book by Sze (1988), in particular to the chapters on epitaxy by Pearce, etching by Schutz, and lithography by Watts.

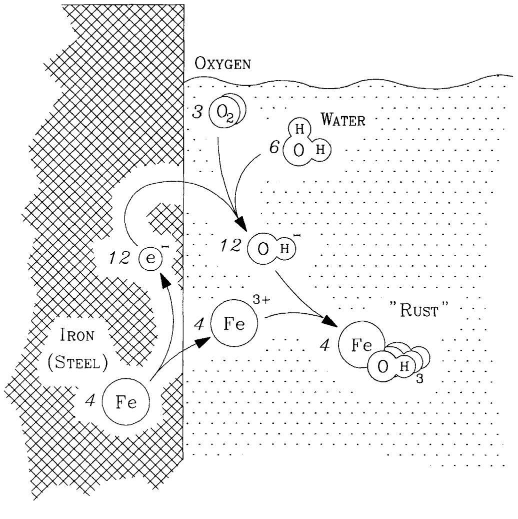

FIG. 10. The oxidation of Fe produces Fe ions and electrons that are then combined with oxygen and water at the cathode to form $(OH)^-$ ions. Rust formation occurs due to recombination of Fe^{3+} and OH^- ions that can be formed if water and oxygen are present simultaneously.

5.1 Nucleation

Nucleation is the formation of very small particles of a new phase due to random fluctuations in the density. Particles larger than a critical particle size tend to grow, and those below it tend to dissolve. The critical size balances the free energy of formation of the particle and the surface free energy of the particle, and particles greater than the critical size tend to grow because the surface-to-volume ratio decreases with increasing particle size. In nucleation on surfaces under nonequilibrium conditions (which is usual), the nature of the surface plays a significant role. The critical particle size then depends on the adsorption, diffusion, and interaction parameters of the particular system. In the case of adparticles on a surface, at high $\Delta\mu$ (chemical potential difference, or thermodynamic driving force), extreme values of the critical particle size i are $i = 1$, i.e., adatoms only have to meet each other, and $i = 0$ in the case of defective surfaces, i.e., adatoms "just" have to occupy a special adsorption site to initiate growth. Flat surfaces have different nucleation rates than stepped surfaces. In the absence of island nucleation the condensation coefficient is a function of the step density. At low $\Delta\mu$ growth does not occur at a measurable rate on flat surfaces but is mediated by screw dislocations.

With the aid of STM, and the related technique atomic force microscopy (AFM), it is now possible to monitor nucleation at an atomic scale in the vapor deposition of thin films. Systematic investigations with atomic resolution of the adsorption of atoms and nucleation on "perfectly" flat terraces and studies of the influence of step structures, defect structures, impurities or coadsorbed particles, temperature, and incident flux have allowed new insight into nucleation phenomena.

A sketch of atomistic processes responsible for nucleation and growth of thin films on a solid substrate is shown in Fig. 11 (Venables, 1994). Particles from a dilute gaseous phase arrive at the surface with a rate r to form adatoms. Their mean displacement x from their arrival site during their residence time on the surface, τ, i.e., prior to evaporation, will depend on the kinetic parameters of diffusion and desorption and may be written as

$$x = (D\tau)^{1/2} = a\left(\frac{\nu_m}{\nu_d}\right)^{1/2} \exp\left(\frac{E_d - E_m}{2kT}\right), \qquad (18)$$

where D is the diffusion coefficient, a is the jump distance, and ν_m and ν_d are the frequency factors and E_m and E_d the activation energies for diffusion and desorption, respectively. Since typically $E_d \approx 3E_m$, x/a can be large at suitably low temperatures, and this allows the adatom to encounter specific sites and other adatom structures to which it might bind before re-evaporation. Critical factors affecting the concentration of nucleated clusters are the density of adatoms on the surface and the binding energy between two atoms.

5.2 Growth Mechanisms

Three different growth modes for the deposition of material A on a substrate B are distinguished as shown in Fig. 12:

1. a perfect layer-by-layer growth, Frank–van der Merwe (FM);
2. 3D growth directly ontop of the substrate, Volmer-Weber (VW); and
3. a combined growth where the first layer(s) grow(s) in two dimensions and every subsequent layer in three dimensions—Stranski-Krastanov (SK).

These growth behaviors can be rationalized in terms of a simple thermodynamic model that takes into account the relative surface free energies γ_A and γ_B, the free energy of the overlayer–substrate interface γ_{AB}, and the crystallographic misfit between the deposit and the underlying substrate surface. If γ_A +

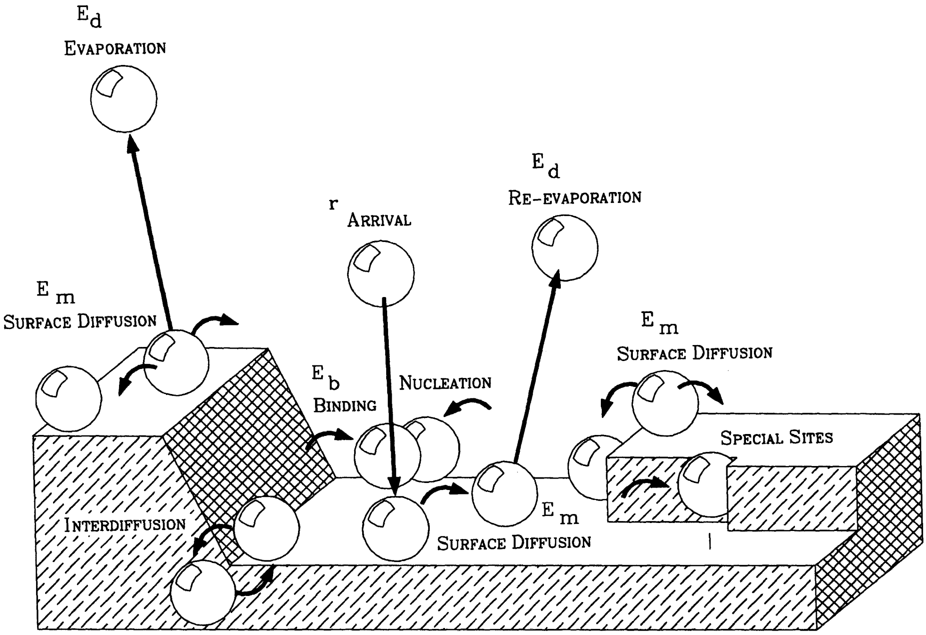

FIG. 11. Schematic diagram of key processes controlling nucleation and growth on surfaces, highlighting the importance of the rate of arrival r of species to the surface and the activation energy barriers for evaporation (desorption) E_d, diffusion E_m, and binding E_b (after Venables, 1994).

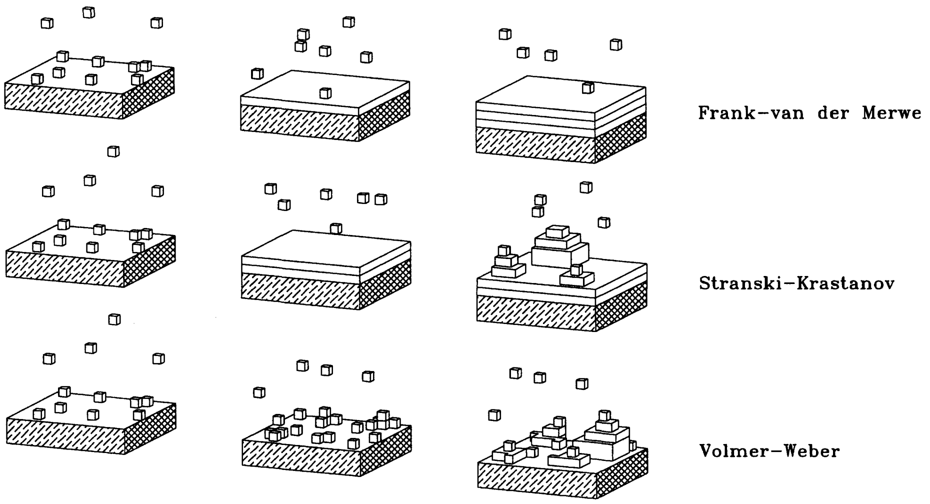

FIG. 12. Schematic view of the three topologically distinct epitaxial growth modes. Growth proceeds from left to right according to each mechanism.

$\gamma_{AB} < \gamma_B$, island growth (VW) is favorable. If $\gamma_B < \gamma_A + \gamma_{AB}$, then B atoms tend to cover the surface completely, and a layer-by-layer growth (FM) is obtained. In the latter case, however, γ_{AB} often increases with the thickness of the layer because of strain in this layer. Beyond a critical layer thickness, island formation is thermodynamically favored, resulting in layer-plus-island growth (SK), a rather common growth mode. Many practical layered structures of technological importance, most notably semiconductor heterostructures and magnetic superlattices of alternating layers (ABAB . . .) of a magnetic metal (Fe, Co, Cr, etc.) and a noble metal (Cu, Ag, etc.) are grown under kinetically limited conditions, i.e., they are far from being equilibrium structures, and so growth modes in these systems cannot be so simply classified.

The 2D and 3D islands formed during growth exhibit a wide variety of structures depending on a complex relationship among all of the parameters affecting nucleation and growth, such as the incident flux, surface temperature, diffusion constants, and so on. Island shapes range from ramified (fractal) structures due to low diffusion mobility along the island edges to polyhedral and triangular structures. Figure 13 illustrates how temperature affects the density and shapes of 2D Si islands grown on the 7×7-Si(111) surface. Stepped (vicinal) surfaces have been shown to exhibit strong anisotropic effects for diffusion along or across the steps, and this strongly affects the shapes of these islands. Reconstructed surfaces with row structures exhibit similar effects.

A detailed description of film growth has to include a kinetic model of the elementary microscopic processes. Theoretical modeling, calculations, and simulations are especially useful since measurement and control of many of the elementary processes are not directly accessible from experiments. In kinetic Monte Carlo simulations (see MONTE CARLO METHODS), atoms are deposited randomly on a well-defined surface and any of the kinetic parameters can be varied to explore their influence on interface structures and growth rates. Using either calculated or model interaction potentials, molecular dynamics calculations (see SIMULATION BY MOLECULAR DYNAMICS) can provide 3D trajectories of individual atoms in large systems of atoms interacting with each other and the surface.

Comparable nucleation and growth mechanisms operate in electrocrystallization of metals on electrodes. In the electrochemical environment a complete monolayer may easily be grown by a process called underpotential deposition (UPD). When the chosen potential is slightly "under" the potential necessary for bulk deposition, e.g., Ag on a

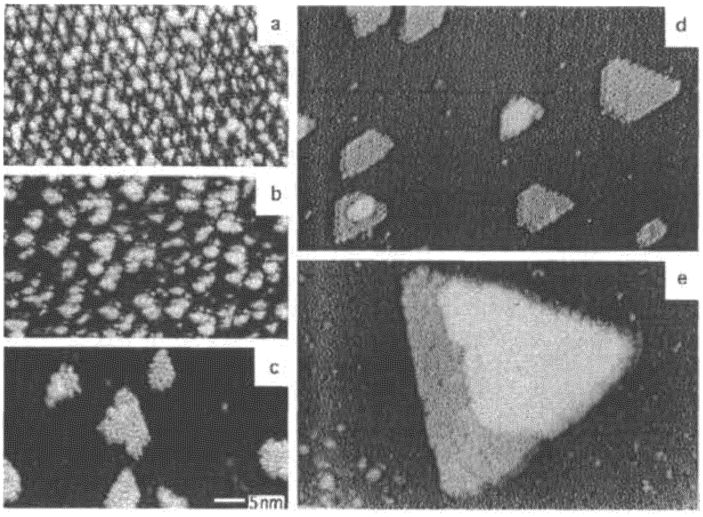

FIG. 13. Sequence of STM images showing the results of nucleation and growth of Si islands on a Si(111) 7 × 7 surface at different substrate temperatures: (a) 20 °C, (b) 250 °C, (c) 350 °C, (d) 450 °C, and (e) 550 °C. On the perfectly triangular island in (e), the second epitaxial layer is visible. (From Köhler *et al.*, 1989.)

Ag electrode, it is possible to deposit a monolayer of Ag on Au but impossible to deposit more Ag at the same potential.

Epitaxy is a special type of growth process where the orientation of the grown film is determined by the single-crystal substrate. A major application for epitaxy is in semiconductor device fabrication. Homoepitaxy means that the grown layer is identical to the single-crystal substrate (e.g., Si on Si, or diamond on diamond) while heteroepitaxy refers to growth of one material on a substrate of another material, and this can be done even when their lattice constants are different (e.g., Si on Ge, or $Al_xGa_{1-x}As$ on GaAs). Heteroepitaxy, in particular, gives the opportunity for modifying electronic and magnetic properties of thin layers with respect to bulk properties. For example, the required high resistivity for high breakdown voltages of transistors simultaneously limited the frequency behavior and power dissipation in the case of the bulk wafers used historically, but the epitaxial growth of a high-resistivity layer on a low-resistivity substrate solved this problem. The strain induced in the film grown in the case of a significant lattice mismatch between the bulk lattices of the film and substrate, however, restricts heteroepitaxy to occur only below a critical thickness.

5.3 Deposition Methods

A wide variety of processes have been used to deposit materials, including those using liquids, ion-beam (sputter) deposition, laser vaporization deposition, molecular-beam epitaxy (*q.v.*) (MBE), and chemical vapor deposition (CVD). Here we can only consider two methods. CVD processes have found widespread application for the preparation of thin films of conducting, semiconducting, and insulating materials needed for the fabrication of integrated circuits. CVD uses gas-phase molecules [precursor gas(es)] that contain the elements required for the deposited materials and is carried out at pressures from 10^{-3} to 760 Torr (1 atm). The precursor gas can be "activated," i.e., made more reactive, photochemically (laser CVD) or by a radio-frequency (rf) or microwave plasma reactor [plasma-enhanced CVD (PECVD), or plasma deposition]. The morphology (crystallinity, defects, impurities, etc.) and the growth rate are found to depend on the substrate, precursor gas, additional reaction gases, etching agents, gas concentration, total pressure, and temperature. For a discussion of the fundamental surface chemistry and kinetics in an exemplary CVD process (W metal deposition using WF_6 and H_2), the reader is referred to Creighton (1994); see also CVD (CHEMICAL VAPOR DEPOSITION).

Some other materials that have been deposited by CVD techniques can be mentioned. The CVD growth of Si epitaxial thin films can be carried out using $SiHCl_3$ and H_2. Silane (SiH_4) and germane (GeH_4) are common precursor gases for thin-film deposition of $Si_{1-x}Ge_x$ heterostructures, while metal carbonyls and organometallic compounds are suitable for the deposition of metal films (Al, Cd, Ni, Fe, etc.). Polycrystalline Si thin films can be deposited at room temperature by PECVD using SiH_4, SiF_4, and H_2. The crystalline fraction increases significantly with addition of SiF_4, indicating the role of etching agents that preferentially remove amorphous material. Low-pressure synthesis of diamond can also be carried out with a variety of molecules as carbon sources in CVD. One possible layout of a PECVD system for diamond growth consists of a parallel-plate rf plasma reactor operated with methanol, a hydrogen radical source using a microwave discharge for the etching of nondiamond carbon phases, and a laser for substrate heating.

MBE is essentially a highly controlled evaporation process that occurs in UHV. MBE has a number of inherent advantages over CVD techniques. Its main advantage for growing thin films is low-temperature processing, which minimizes unintentional diffusion processes. This is especially important for the growth of quantum wells and quantum wires, which are made from ultrathin-layer structures that generate electron-confining potential wells on such a small scale that particle-in-a-box electronic quantum states are produced. For example, the growth of AlGaAs on vicinal (110) GaAs surfaces forming quantum wires along the step edges can be accomplished with MBE techniques. Other advantages include precise control of doping, allowing the generation of complicated doping profiles in semiconductors, and the ability to monitor film growth and the quality of a growing layer *in situ* by RHEED. The main technological disadvantages of MBE compared to CVD are its low growth rates and its expense.

5.4 Etching

Etching is the process of removing surface material and thus, like dissolution, is the opposite of crystal growth. Etching is widely used for surface treatments and as a fundamental step in semiconductor processing. For example, wet chemical etching using HF, HCl, or sulfuric acid (H_2SO_4) in combination with passivation and rinse treatments can produce very flat surfaces with subnanometer surface undulation for Si and GaAs surfaces. Various techniques depending on the substrate and the desired etching results have been developed. In all of these techniques, corrosive surface reactions occur by breaking chemical bonds within the substrate surface layer by attack of highly reactive chemical species and/or by momentum transfer. A complete microscopic description of the elementary mechanisms remains elusive because etching occurs in very complex environments that are difficult to probe. For example, etching with a reactive plasma (see PLASMA ETCHING) involves a variety of neutral and ionic species and collision, dissociation, recombination, and interaction processes both in the gas-phase plasma and on the exposed surfaces. Processing with nanoscale dimensions requires an improved understanding and new methodology.

The majority of etching methods can be grouped under the heading reactive plasma etching, e.g., reactive ion etching (RIE), reactive ion-beam etching (RIBE), and microwave plasma etching. In such processes, a plasma discharge generates ions and many other chemically active neutral species. Fluorine and chlorine plasmas contain atomic fluorine and chlorine, respectively, which are far more chemically active than the naturally occurring molecular species. In the special case of fluorine etching of Si, atomic F and molecular F_2 etch Si spontaneously at room temperature without a plasma. This can occur through efficient fluorine adsorption and subsequent desorption of volatile SiF_4. When ions strike the surface an increased etch rate is observed, which is not simply due to the additional sputtering effect, in a reaction usually referred to as ion-enhanced etching. Increased surface roughness and synergistic effects during desorption have been discussed as possible causes. In the etching of Si with a plasma using Cl_2, chlorine adsorption produces a surface $SiCl_2$ species that is a precursor to the formation of volatile $SiCl_4$. In contrast to F_2, Cl_2 does not etch Si spontaneously, and so this type of reaction is termed ion-induced etching because it occurs only in the presence of ion bombardment.

Isotropic etching during pattern transfer in lithography (see below) usually results in "undercutting" of the mask, which narrows the feature. This is a major problem for linewidths and spaces of features that are significantly smaller than the feature height, and production of these structures requires an anisotropic etching. Wet etching is isotropic, but ion-enhanced processes are expected to produce anisotropic etching with some degree of mask undercutting, and ion-induced processes should produce perfectly anisotropic etching in the vertical direction.

5.5 Lithography and Nanotechnology

The manufacturing of integrated circuits requires sequential formation of patterned layers of material on a substrate surface, most commonly a Si wafer. These layers alternate between insulators, semiconductors, and conductors, and the intricate pattern that will eventually be in the respective layers is first defined in a polymeric resist film via lithography and then transferred to the layer usually with etching techniques (see MICROLITHOGRAPHY). Historically lithography started out as photolithography, where an image is recorded into a resist (an etching-proof material) by means of light exposure to chromic groups. The resulting changes are then responsible for a different solubility or volatility of those parts compared to the unexposed areas in a subsequent treatment called resist development. In modern microlithography, optical methods are complemented by electron, x-ray, and ion techniques, each with particular advantages. In x-ray lithography, x rays are used to minimize diffraction limitations during the writing process. It is not surprising that the masks themselves are produced by x-ray microlithography. The basic structure of these masks is an adsorber film (Ta,W,Re) on top of a membrane (Si, SiN, SiC, diamond) that is transparent for soft x rays. Figure 14 shows the basic fabrication process of a Ta/SiN mask used for x-ray lithography illustrating how several different deposition and etching methods are needed to construct such a structure. Novel approaches for the

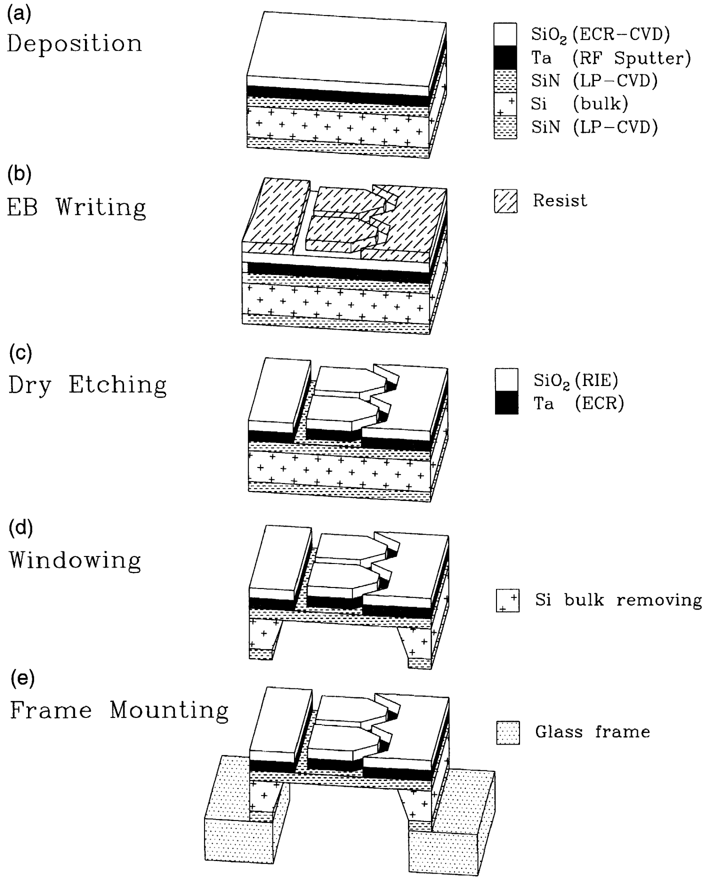

FIG. 14. Fabrication steps in producing a Ta/SiN mask starting with a 3-in. Si wafer as the base material. (a) Deposition of 2-mm-thick SiN films via low-pressure CVD (LPCVD), a 0.7-mm-thick Ta film by rf sputtering a Ta target onto the substrate, and a SiO_2 etching mask via electron cyclotron resonance CVD (ECR-CVD). (b) Resist pattern production by electron-beam writing following deposition of the resist film. (c) SiO_2 etching in regions not protected by the resist film using RIE and Ta film etching in regions unprotected by the SiO_2 mask using ECR followed by removal of the resist film. (d) Underside etching of SiN by using RIE and bulk Si removal using an alkaline solution in a "wet chemical" procedure. (e) Mounting on a glass frame. (After Oda and Yoshihara, 1993.)

formation of nanostructures (with dimensions of nanometers) utilize STM or AFM.

Realization of the fantastic promises of nanotechnology—whereby nanoscopic devices are constructed at the atomic or molecular level, biomolecular modification is carried out with bond specificity by direct manipulation of the atoms of the molecule by cell-repairing nanorobots, and the evolution of ultrafast and ultrasmall computers, smart materials, and self-replicating machines occurs—obviously requires extensions of scanning probe microscopy beyond imaging and lithographic patterning. The most impressive results in direct manipulation of atomic structures have been achieved by using STMs. STM techniques have been extended from positioning individual xenon atoms on a cold (4 K) nickel substrate at 0.5-nm resolution, as demonstrated in the pioneering work of Eigler and Schweizer (1990) who spelled IBM by direct placement of 35 Xe atoms, to the construction with an STM of a "quantum corral" for electrons by placing 65 Fe atoms in a ring about a center atom on a Cu surface (Crommie *et al.*, 1993). The atoms in these experiments were attached to and detached from the STM tip electrostatically by controlling the tip bias voltage. Molecular manipulation has recently been demonstrated in the controlled movement of individual tetrabutylphenyl porphyrins on Si at room temperature. In these studies the STM tip was used to "push" the molecules across the surface. Movement of larger structures such as metal discs and colloidal particles can also now be accomplished by contact AFM and imaged by noncontact AFM. AFM is a scanning probe microscopy technique that relies on surface forces to cause a displacement of a microscopic lever to produce an image. In contact AFM, the lever "touches" the surface and senses the repulsive potential, while in noncontact AFM the lever oscillates just above the surface sensing the attractive potential. Scanning probe technology is therefore capable of enough spatial resolution for both placing and sensing atomic-scale elements on surfaces, but further developments will require increased understanding and control of surface forces.

6. MECHANICAL PROCESSES

Interfaces strongly affect the mechanical properties of polycrystalline materials. These properties define the response of materials to external and internal forces. Deformation phenomena like flow and creep (continuous deformations at constant load) are influenced by interfaces, and surface forces play a role in understanding these distortions. The reader is referred to Sutton and Balluffi (1995) for an in-depth discussion of this topic. Friction, lubrication, and adhesion are important mechanical processes that involve surfaces and adsorbed layers and that are controlled by surface forces. Below we highlight the changes that occur when new interfaces are created as an adherent film is peeled away from a substrate and when two solid objects slide over each other with and without lubricating layers.

6.1 Adhesion

The work necessary to separate two phases A and B with a common interface is known as the work of adhesion W_{AB} between the two phases and is given (most simply) by

$$W_{AB} = \gamma_A + \gamma_B - \gamma_{AB}. \tag{19}$$

An analogous expression for the energy of adhesion can be given, although these equations ignore important electrostatic contributions that can occur due to charging of the surfaces upon separation. In the cases of strong chemical bonding at the interface, these energies are large and can exceed the cohesive energies of the pure phases (i.e., adhesive joints can fail by formation of an interface within the bulk of one of the two phases). This is often desirable since this corresponds to the maximum strength of an adhesive joint. "Peel tests" commonly reveal that failing loads of joints do not approach this value. In these tests, the normal force to separate the joint (tensile strength) is measured, but alternatively one can also determine the force parallel to the surface that is needed to break the joint (shear strength). The simplest model for adhesive joint failure is a classical mechanical picture based on the work of crack propagation and stored elastic energy. However, more detailed models attempt to account accurately for the structural distortions of the surface layers, adsorbate layers at the interface, and the actual bonding interactions present at the interface.

Monolayer or even submonolayer

amounts of adsorbates present at the interface can dramatically influence adhesion, either positively or negatively. Adhesion promoters such as amines for Cu/CuO localize at the interface for a beneficial effect. Segregation of extremely small (ppm) concentrations of impurities (e.g., C, O, S, P, As) in steels and other alloys to interfaces negatively affect adhesion (and can cause bulk material failure when grain boundaries of the pure component are decorated). Polymer–metal interfaces have a great importance, and current research is involved in investigating deposition of metals at polymer surfaces at the monolayer level with surface-science techniques to probe the nature of the chemical bonds at the interface. Reactive processing (e.g., plasma etching) of polymers can break chemical bonds at the interface, generating reactive sites that can form new strong bonds to a second phase.

6.2 Friction and Lubrication

Given the tremendous practical and economic importance of friction (and the related wear), it is somewhat remarkable that so little is known about the microscopic-level processes responsible. However, within the last few years, experimental capabilities for studying sliding friction and wear (in the field of tribology, *q.v.*) have been rapidly advancing. For example, it is now possible to directly measure forces between macroscopic objects using crossed cylinders coated with mica (which can be atomically smooth). Using AFM, one can probe atom–atom forces and friction and sliding friction on atomically smooth, well-characterized surfaces with lubrication films of a known and controllable thickness (within a few angstroms).

Even smooth-appearing surfaces have irregular shapes at an atomic level, and these defects control the nature of contact between two solid surfaces. Understanding the tremendous forces and high temperatures generated at these localized contacts (making and breaking chemical bonds) is necessary for a description of macroscopic friction. This involves the influence of shearing and plowing phenomena within the context of deformations and plastic flow at the atomic level. A phenomenon that has been demonstrated amply at surfaces and measured directly by using an AFM is stick-slip friction motion.

Lubrication is used to decrease the friction between two surfaces. Two types of lubrication are recognized: boundary lubrication and hydrodynamic lubrication. In hydrodynamic lubrication, where the lubricating film is thick enough so that the surface regions are essentially independent of each other, the coefficient of friction depends on the hydrodynamic properties (mainly the viscosity) of the lubricant and not much on the surfaces involved. The specific nature of the solid surface is important in boundary lubrication, which describes the situation for high loads and low speeds where very thin (indeed monolayer) films are present at the sliding surfaces. Understanding this situation requires consideration of the structure, orientation, and bonding of the adsorbed layer at the surface, surface diffusion mobility, and the influence of large mechanical compressive forces on this layer. Modern surface-science tribology research is aimed at developing lubricants that strongly bond to the solid surface but have minimal chemical interactions with other lubricant molecules present on the other solid surface.

7. COMPLEX SURFACES AND ENVIRONMENTS

Many technologically important or fundamentally interesting interfaces are very difficult to study, and thus relatively little molecular-level information is known about surface processes in these systems. In this section, we briefly highlight four examples of these "complex" surfaces and environments to illustrate the wide range of applications for interfacial phenomena. Complex surfaces include those for amorphous materials, which have an irregular structure; porous materials, which have "internal" surfaces; polymer and biomembrane surfaces, which have a complicated morphology; and very heterogeneous surfaces with a multiplicity of surface species. Complex environments mainly mean those that are not compatible with the UHV conditions required by many surface analytical techniques: liquid solutions such as those in electrochemical devices and fuel cells, "wet" solids such as biological samples, high–vapor-pressure materials, and high-pressure catalytic reactors. We also include outer space, which exposes surfaces to vacuum and bombardment by a myriad of particles.

7.1 Amorphous and Porous Materials

Amorphous materials are often used for their chemical or heat resistance, electrical isolation, and spectral transmission properties. The most common amorphous materials are glasses (supercooled liquids). Amorphous means literally "without shape," and so these materials do not exhibit the regular, well-defined atomic arrangement of crystalline solids. The irregular surface structure is not accessible via diffraction measurements. Adsorption and reaction at the surface can be studied by AES, XPS, and SIMS, which are all methods that do not rely on long-range order of the substrate. If the sample is an electrical insulator, charging of the surface limits analysis. SEXAFS probes only local atomic order and is suitable for studies of these surfaces. Also, Raman spectroscopy with its ability to fingerprint oxide phonon modes provides near-surface information on very thin films. Theoretical calculations for amorphous surfaces are more difficult than for crystalline materials since they have to include a higher number of independent particles and interactions.

Porous materials consist of extensive channels or pores that allow the accessible solid surface area to exceed greatly the geometrical or apparent area. The high activity required for most catalytic conversion processes necessitates the use of porous materials with a high surface area. Oxides such as V_2O_5, Al_2O_3, SiO_2, and TiO_2 are often used directly and catalyze many reactions. When metal-based catalysts are needed, the transition metal (e.g., Ni, Pd, Pt) is placed as a highly dispersed phase of small nanoparticles on high–surface-area oxide supports. Zeolites, inorganic molecular cage structures as illustrated in Fig. 15, are very important as acid catalysts, catalyst supports, and filtration/separation media.

Microporous materials are used as ion-exchange membranes. In applications such as dialysis, the membrane surface has large open pores with diameters of around 300 nm and a system of connected pores in the membrane volume with diameters in the 10–100-nm range. The filtration/separation properties are governed by the pore-size distribution of the membrane surface and volume. Membrane surface charges and the surface pore sizes influence the adhesion properties of molecules. The use of macroporous polymeric media provides extensive control of the surface chemistry that determines the separation selectivity and capacity via the incorporation of specific functional groups that modify the hydrophilicity and chemical affinity in the large and small pores.

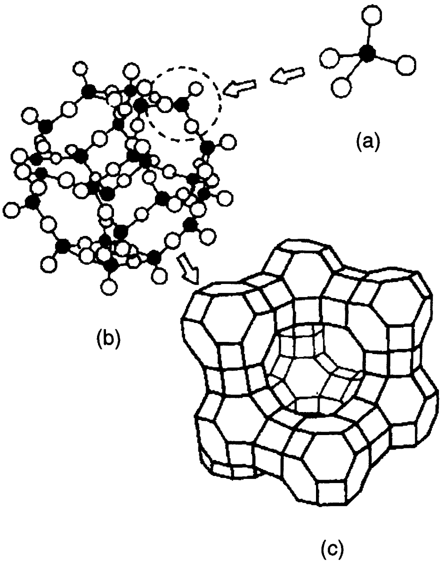

FIG. 15. Structural units showing the framework of a zeolite (faujasite) that gives rise to channels with molecular dimensions permeating the solid. (a) The tetrahedral arrangement of Si (or Al) atoms bonded to O atoms to form (b) a sodalite unit consisting of 24 SiO_4^- and AlO_4^- tetrahedra that are connected in this case to (c) the zeolite superstructure. (From Sleight, 1980.)

Most surface analytical probes only detect external surfaces and not the internal surfaces of the pores, which are relevant. High surface-to-volume ratios for these high–surface-area materials mean, however, that traditional bulk analytical methods such as FTIR and Raman spectroscopy can be used nicely to probe adsorption and chemical changes in the surface layers. The surface area may be determined by measurement of the BET isotherm, and the pore-size distribution may be determined with SEM or mercury porosimetry.

7.2 Polymers and Biological Materials

Plastics, synthetic elastomers (rubber), and thermosetting resins are composed of polymers. The use of these materials is extremely widespread and varied, with new applications being developed continuously. Even though interfacial chemical bonding and structure controls adhesion, thermal stability, and many of the other materials properties, our understanding of surface structure and processes on these materials is relatively poor. They are often amorphous, porous, and nonconductive and have large cross sections for electron- and ion-induced damage, and thus they are quite difficult to analyze by standard surface analytical techniques. XPS has found the most extensive use in identifying functional groups at the surface and studying their reactivity. SIMS also is valuable. FTIR and Raman spectroscopy sample the first few microns of the substrate to provide unique information on morphology (e.g., chain conformation, crystallinity, and orientation). AFM is just beginning to be applied to polymeric surfaces.

Investigations of surfaces of biological samples are a major challenge in terms of finding appropriate preparation and analysis methods. The presence of water inherent to biological systems represents a problem for standard UHV analysis. Below 120 K, however, the partial pressure of water is negligible, and biological samples have been investigated with electron microscopy, SIMS, and XPS under UHV conditions using a cryogenic sample-transfer system.

Surface structure and molecule–solid interactions play important roles in processes on biological membranes. A lipid bilayer structure with hydrophilic functional groups on the outside and hydrophobic functional groups inside represents the basic element of a biological membrane (Fig. 16). The specific binding of large molecules, such as proteins, to the membrane is generally affected by the proximity of polar areas, while the nonspecific adsorption is governed largely by the number of defects and by the availability of hydrophobic binding sites in the interfacial region. Knowledge of these interactions can be used to design artificial lipid membranes with desirable surface properties. A parallel approach uses monomers to insert specific functional groups at the surface of thin polymer films as membrane models. These efforts may ultimately lead to an improved understanding of the molecular recognition of proteins, which controls biological processes like protein separation, biosensing, or drug delivery. TOF-SIMS (for high-molecular-weight capability), XPS measurements, and vibrational spectroscopies are commonly used in these studies. Dramatically increased modeling capabilities are now emerging from improved theory and simulation methods and the enormous increase in availability and speed of computers.

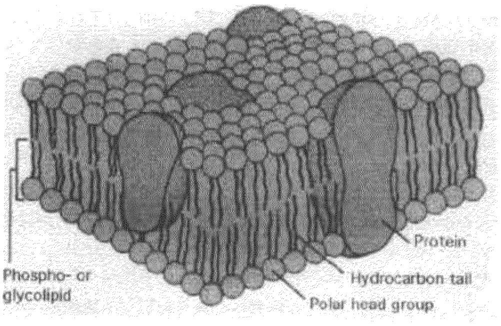

FIG. 16. Illustration of a simple bilayer membrane (from Fendler, 1984).

7.3 Electrochemistry and Fuel Cells

Electrochemical devices convert chemical energy to electrical energy and vice versa. The classical example for an electrochemical process is the electrolysis of water, where water molecules are decomposed by means of a direct electric current. Electrochemical methods are routinely applied in surface treatments such as HF etching of Si samples or for plating purposes, e.g., Pd electrodeposition for electrical contacts. Both neutral molecules and charged molecular species (ions) are important in these processes. Classical investigation methods, such as cyclic voltammetry or polarography (*q.v.*) are used to reveal information on net charge transfers at electrode surfaces. Features in the curves of current versus potential (*I*/*V*) are produced by oxidation, reduction, and capacitive currents and have often been interpreted in terms of surface roughness and morphology without access to the detailed microscopic structure of the electrode interface.

With the development of surface-sensitive optical spectroscopic techniques such as reflection FTIR, SHG, SFG, and XSW, a molecular-level characterization is now possible *in situ*. Also, STM yields images of surfaces with atomic resolution in the liquid, electrochemical environment. Other studies of the electrode surface with AES, XPS, and SIMS via *ex situ* analysis in UHV use an integrated experimental setup that combines the electrochemical cell with a UHV analytical chamber and sample transfer from one environment to the other without exposure to air. This equipment may be used for either UHV surface characterization of a "real" electrode surface or an *in situ* performance test of a model electrode surface that has been prepared and characterized under UHV conditions prior to use.

A primary interest is to describe the electrode–solution interface, i.e., the details of the charge separation responsible for the potential difference at the interface. Figure 17 shows the simplest model for this interface, treating it as an electric double layer, and current research attempts to improve this model. Advances are now being made concerning the relationship between macroscopic kinetic parameters of electrochemical processes and the structural dynamics on a nanometer scale. Chemisorption issues control the nature of molecular adlayer structures and the deposition of metal atoms on metal surfaces in the electrochemical environment. Thus, in many cases a comparison of the electrochemical adlayer structure to its UHV analog is quite useful. For example, the CO adlayer on a Pt(111) electrode in solution shows a potential-induced phase transition that is very similar to that observed for CO adsorption on a Pt(111) surface in UHV experiments.

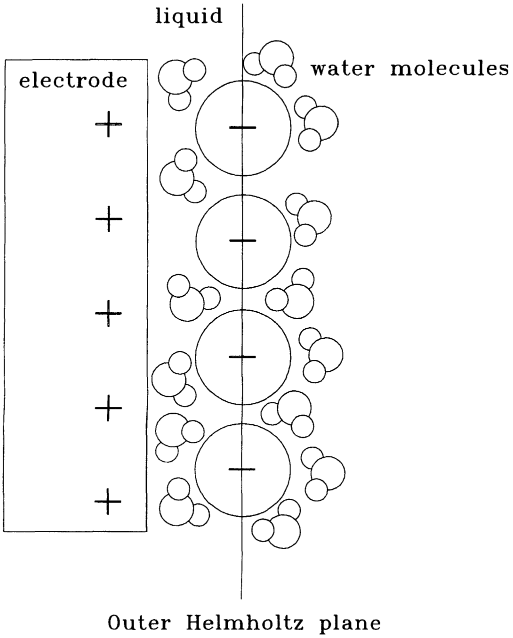

FIG. 17. The simplest model of the electric double layer. The outer Helmholtz plane is due to the ions in the solution, and the inner Helmholtz plane is due to the charge on the electrode. The simplicity of the model arises from the "simple" arrangement of ions and water molecules near the electrode surface.

A fuel cell operates in reverse of electrolysis. The electrodes of a H_2 fuel cell catalyze the controlled oxidation of hydrogen yielding a direct electric current. Fuel cells can utilize other fuels, such as methanol, and have other constructions of solid electrode/liquid electrolyte or solid electrode/solid electrolyte. However, all suffer from limited operation time, which is likely a consequence of surface processes such as an electrolyte-induced surface modification/degradation and poisoning of the electrode surface. Pd and Pt, for example, which are considered as highly efficient catalysts for hydrogen conversion reactions in a low-*p*H environment, also exhibit a strong affinity for CO, which is known to poison the electrode surfaces. Thus, while methanol has some advantages over hydrogen, like its higher energy density, the small amount of CO formed during the reforming process of the methanol poisons the electrodes. The development of CO-tolerant catalysts is of crucial importance for long-term operation of this and several other fuel cells.

7.4 Space Exploration

In high Earth orbit or in outer space, the surfaces of exposed materials undergo collisions with a myriad of energetic particles and are bombarded by intense cosmic radiation, creating a very heterogeneous and complex surface. *In situ* experiments have been performed at NASA's Long Duration Exposure Facility (LDEF) that remained in space for nearly 6 years (Benton *et al.*, 1992), but much of the fundamental interaction physics can be studied in the laboratory under simulated conditions. Atomic oxygen is the dominant atmospheric species in low Earth orbit, and hyperthermal O-atom–surface interactions are central to many materials problems and also account for the nuisance "glow" of the space shuttle. Hyperthermal oxygen sources are used in the laboratory to simulate the kinetic energies of collisions that occur at orbital velocities. Laboratory studies combining FTIR and XPS on the UV degradation of polymer (polyimide) films in air and in vacuum showed that UV irradiation results extensively in surface degradation, leaving the bulk of the polymer intact. Synergistic effects on polymer surfaces were found for UV radiation and oxygen atoms (Hill *et al.*, 1995). While the exposure of fluorinated polyimide films to UV radiation alone caused no observable damage to the polymer surface, atomic oxygen resulted in substantial oxidation of the surface, and a simultaneous exposure to UV radiation caused rapid extensive oxidation. The oxidation process was found to occur via aromatic ring-opening reactions. The effects of hyperthermal (3–7 eV) O atoms, 30-eV O^+, and 30-eV Ne^+ ions on two polymers commonly used for space applications (Kapton and Teflon) have also been investigated with respect to the surface morphology (SEM and AFM) and surface composition (XPS). The damage from its initial atomic-scale roughness (approximately 1 nm) to its final macroscopic-scale roughness (>1 mm) was imaged with AFM (Vered *et al.*, 1994).

Another important consideration in the space environment is the high vacuum that exists. This leads to evaporation and desorption of lubricating layers of water and organic films normally present on all surfaces. Maintenance of good lubrication of moving joints and metal surfaces in contact is essential because the energy gain by eliminating two surfaces with high surface tensions via "vacuum welding" can be very large ($\gamma \approx 150$ kJ/mol, which is comparable to the metal–metal bond strength). In fact, this joining procedure is now used commercially in a variety of bonding applications.

GLOSSARY

Absorption: The penetration of the bulk of one phase by the other in the contact of immiscible phases.

Adsorbate: A species (atom, molecule, radical, or ion) in an adsorbed state.

Adsorptive: A species in a contacting phase that is capable of being adsorbed.

Adsorbent: The surface on which adsorption is occurring. Also called the substrate.

Adspecies: An atom, molecule, or fragment adsorbed on a surface.

Adsorption: Process in which atoms or molecules accumulate (localize) at an interface. A specific location is called an adsorption site. If an energetic barrier has to be surmounted prior to adsorption, the process is denoted as activated adsorption.

Adsorption Complex: The entity constituted by the adsorbate and the part of the adsorbent to which it is bound.

Catalyst: A substance that increases the rate of a chemical reaction without being consumed.

Chemical Vapor Deposition (CVD): Deposition process that proceeds via decomposition of a gaseous species.

Chemisorption (Chemical Adsorption): Adsorption in which the forces responsible for linking the species to the surface are valence forces (covalent, ionic, and dative bonds) similar to those between the atoms in molecules and chemical compounds.

Coadsorption: The simultaneous presence of more than one adspecies.

Corrosion: Material deterioration and removal by chemical reactions.

Deposition: Transferring material from the gas or liquid phase onto a surface.

Desorption: The converse of adsorption.

Diffusion: Large-scale translational displacements of atoms or molecules in a material.

Epitaxy: Growth process where the orientation of a single-crystal layer is determined by that of the single-crystal substrate.

Etching: Removal of surface atoms by reactive particles or chemical reactions.

Implantation: Deposition of foreign atoms deep (microns) into the bulk of a material, usually carried out by high-energy ion bombardment.

Interface: Boundary between two phases; the region of space comprising and adjoining the phase boundary within which the properties of matter are significantly different from the values in the adjoining bulk phases.

Lithography: The transfer of structural information from a template or mask to a substrate by inducing reactions in exposed regions of the substrate using UV light, electrons, ions, or x rays.

Molecular-Beam Epitaxy (MBE): A technique for the vacuum deposition of epitaxial thin films from molecular precursor gases.

Monolayer: Film in which all adsorbed species are in contact with the adsorbent (substrate) surface; i.e., the film has a thickness of one atomic or molecular layer.

Multilayer: Film with a thickness of more than one atomic or molecular layer and thus having not all adsorbed species in contact with the adsorbent surface.

Nucleation: Fluctuation in particle concentration above ambient responsible for initiation of the growth of a new phase.

Phase: A state of matter uniform in chemical composition and physical state throughout.

Photoelectron: Free electron produced by light irradiation in a process denoted as photoelectron emission or photoemission.

Physisorption (Physical Adsorption): Adsorption in which the forces responsible for linking the species to the surface are weak van der Waals forces such as those responsible for imperfections of real gases and condensation of vapors.

Reconstruction: 2D rearrangement of the positions of the surface atoms from those expected for the continuation of the bulk lattice.

Relaxation: Vertical displacement of atomic layers at a surface without significant accompanying horizontal displacements.

Segregation: Movement of one component to an interface resulting in a surface concentration not equal to that in the bulk phase.

Sputtering: Removal of surface material during the impact of energetic particles (usually ions).

Surface: Outermost layer(s) of a condensed phase at an interface.

Works Cited

Adamson, A. W. (1990), *Physical Chemistry of Surfaces*, 5th ed., New York: Wiley.

Benton, E. V., Heinrich, W., Parnell, T. A., Armstrong, T. W., Derrickson, J. H., Fishman, G. J., Frank, A. L., Watts, J. W., Wiegel, B. (1992), *Int. J. Rad. Appl. Instr. D* **20,** 75–100.

Brunauer, S., Emmett, P. H., Teller, E. (1938), *J. Am. Chem. Soc.* **60,** 309–319.

Cabrera, N., Mott, N. F. (1949), *Rep. Prog. Phys.* **12,** 163–184.

Creighton, J. R. (1994), *Thin Solid Films* **241,** 310–317.

Crommie, M. F., Lutz, C. P., Eigler, D. M. (1993), *Nature* **363,** 524–527.

Eigler, D. M., Schweiter, E. K. (1990), *Nature* **344,** 524–526.

Ertl, G. (1983), *J. Vac. Sci. Technol. A* **1,** 1247–1253.

Eyring, H., Jost, W., Henderson, D. (Eds.) (1974–1975), *Physical Chemistry—An Advanced Treatise,* Vol. 6, New York: Academic.

Fendler, J. H. (1984), *Chem. Eng. News* **62,** 25–38.

Gasser, R. P. H. (1985), *An Introduction to Chemisorption and Catalysis by Metals,* Oxford: Clarendon.

Hendershot, R. E., Hansen, R. S. (1986), *J. Catal.* **98,** 150–165.

Hill, D. J. T., Rasoul, F. A., Forsythe, J. S., O'Donnell, J. H., Pomery, P. J., George, G. A., Young, P. R., Connell, J. W. (1995), *J. Appl. Polym. Sci.* **58,** 1847–1856.

Ishitani, T., Schimizu, R. (1974), *Phys. Lett.* **46A,** 487–488.

Kellogg, G. L. (1994), *Surf. Sci. Rep.* **21,** 1–88.

Kiskinova, M., Szabo, A., Yates, J. T., Jr. (1988), *Surf. Sci.* **205,** 215–229.

Köhler, U. K., Demuth, J. E., Hamers, R. J. (1989), *J. Vac. Sci. Technol. A* **7,** 2860–2867.

Oda, M., Yoshihara, H. (1993), in: G. K. Celler, J. R. Maldonado (Eds.), *Materials Aspects of X-Ray Lithography,* Materials Research Society Symposium Proceedings No. 306, Pittsburgh: Materials Research Society, pp. 69–80.

Prutton, M. (1983), *Surface Physics,* Oxford: Clarendon.

Rettner, C. T., Auerbach, D. J. (1996), *Surf. Sci.* **357–358,** 602–608.

Sleight, A. W. (1980), *Science* **208,** 895–900.

Sutton, A. P., Balluffi, R. W. (1995), *Interfaces in Crystalline Materials,* New York: Oxford Univ. Press.

Sze, S. M. (1988), *VLSI Technology,* 2nd ed., New York: McGraw-Hill.

Takayanagi, K., Tanishiro, Y., Takahashi, M., Takahashi, S. (1985), *J. Vac. Sci. Technol. A* **3,** 1502–1506.

Tully, J. C., Cardillo, M. J. (1984), *Science* **223,** 445–450.

Venables, J. A. (1994), *Surf. Sci.* **300,** 798–817.

Vered, R., Matlis, S., Nahor, G., Lempert, G. D., Grossman, E., Marom, G., Lifshitz, Y. (1994), *Surf. Interf. Anal.* **22,** 532–537.

Wiesendanger, R. (1994), *Scanning Probe Microscopy and Spectroscopy, Methods and Applications,* Cambridge, U.K.: Cambridge Univ. Press.

Further Reading

Adamson, A. W. (1990), *Physical Chemistry of Surfaces*, 5th ed., New York: Wiley.

Duke, C. B. (Ed.) (1994), *Surface Science: The First Thirty Years*, New York: North-Holland.

Garcia-Moliner, F., Flores, F. (1979), *Introduction to the Theory of Solid Surfaces*, New York: Cambridge Univ. Press.

Lüth, H. (1993), *Surfaces and Interfaces of Solids*, New York: Springer-Verlag.

Morrison, S. R. (1990), *The Chemical Physics of Surfaces*, New York: Plenum.

Somorjai, G. A. (1994), *Introduction to Surface Chemistry and Catalysis*, New York: Wiley.

Woodruff, D. P., Delchar, T. A. (1994), *Modern Techniques of Surface Science*, 2nd ed., New York: Cambridge Univ. Press.

Zangwill, A. (1988), *Physics at Surfaces*, New York: Cambridge Univ. Press.

SURFACES AND INTERFACES OF FLUIDS, STRUCTURE OF

Mark L. Schlossman, *Department of Physics, University of Illinois, Chicago, Illinois, U.S.A.*

Introduction 311
1. **Interfacial Free Energy and Macroscopic Descriptions of Liquid Interfaces** 312
2. **Interfacial Profiles: The Microscopic View of Interfacial Structure** 315
2.1 Density Profiles 315
2.2 Interfacial Fluctuations 316
3. **Adsorption and Wetting Phenomena** 318
4. **Interfacial-Profile Measurements** 321
4.1 X-Ray Measurements of Interfaces far from Criticality 321
4.2 Optical Measurements of Interfaces near Criticality 324
4.3 Nonclassical or Nondielectric Fluid Interfaces 326
4.3.1 Quantum Liquids 326
4.3.2 Liquid-Metal Surfaces 326
4.4 Complex-Fluid Interfaces 328
4.4.1 Microemulsions 328
4.4.2 Alkanes and Alcohols 330
4.4.3 Liquid Crystals 330
5. **Some Other Surface Phenomena** 332
Glossary 333
Works Cited 333
Further Reading 336

INTRODUCTION

As is evident from everyday observation, the water surface is macroscopically sharp, clearly separating the bulk water phase from a bulk vapor phase. For many liquids near their triple point, the inhomogeneous, transition region of the surface occurs over a length scale approximately equal to two molecular diameters. However, by adjusting the pressure and temperature to approach the liquid's critical point (e.g., $T = 374.15$ °C and $p = 2.212 \times 10^7$ Pa for water), the surface or interfacial width can be increased substantially to thousands of molecular diameters or larger.

Also evident from macroscopic observations of poking a pin into the water surface is the existence of a surface tension. This tension, when balanced against the gravitational forces, determines the macroscopic behavior of the interface. As an example, water rises when in contact with a vertical wall of clean glass. The shape of the water meniscus is determined by gravity and the interfacial tensions at the glass–liquid, glass–vapor, and liquid–vapor interfaces. The rise indicates that the liquid water is attracted to the glass.

The surface or interfacial tension can be determined by macroscopic measurements, such as the shape of a drop, the rise of a column of liquid in a capillary, or the force on a thin plate dipped in a liquid. This tension is determined by the free energy required to take molecules from the bulk of the liquid and create a unit area of surface or interface (the term interface refers to the transition region between any two bulk phases in contact; the term surface is reserved for the liquid–vapor interface). The energy of a molecule at the surface is altered because it has fewer nearest neighbors than in the bulk. Unless confined or preferentially ordered at the surface, a surface molecule has additional entropy over molecules in the bulk. By use of statistical mechanics the tension can be computed from the variation in density through the interface (the longitudinal density profile). The connection between the interfacial tension and the microscopic density profile can be investigated experimentally and is of fundamental importance.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Multicomponent solutions have interfaces between two bulk liquid phases. For example, a binary mixture consisting of two different chemical components, A and B, can exhibit coexistence behavior of two phases, A^* and B^*, each consisting of primarily one component mixed with a lesser amount of the other component. The interface between these two phases can be described by a longitudinal composition profile whose width is of molecular dimensions far from the mixing critical point. As the critical point is approached, the two bulk phases approach each other in composition, and the interfacial width increases.

The surface or interface does not have to consist of the same components as the bounding bulk phases. Surfactants minimize their free energy by going to the interface and are often found primarily at the interface. Surfactant molecules consist of a hydrophilic moiety (typically a polar group) that prefers to be in contact with a polar solvent, such as water, and a hydrophobic moiety (typically nonpolar) that prefers a nonpolar environment, such as oil or air. An extreme example of this is the formation of monolayers of insoluble surfactants at the air–water interface (denoted Langmuir monolayers). These monolayers have interesting structure in the plane of the surface and exhibit a variety of solid and liquid phases.

Soluble surfactant molecules form a great variety of interfaces internal to the bulk liquids. In a three-component system of oil, water, and a surfactant, the surfactant can partition the water and oil on the microscopic level. For example, random sheets of surfactant separate microscopic regions of oil and water in the bicontinuous microemulsion phase. Additionally, these internal structures and interfaces may be ordered at the interfaces between the bulk phases, thereby influencing the interfacial tension and wetting properties.

Although the structure and properties of liquid interfaces have been studied both theoretically and experimentally for many years, there still remain some fundamental questions that have not been conclusively addressed. For example, the electron and molecular longitudinal density profiles (normal to the interface) and the transverse correlations (within the plane) of the interface are not adequately understood. This is partially due to the theoretical difficulties of understanding inhomogeneous fluid systems. Liquid–liquid and liquid–vapor interfaces can be viewed as model systems whose study will influence our general understanding of inhomogeneous materials. Experimental progress has been impeded by the lack of structural information on the molecular length scale for these interfaces. Until recently, much of the available information was from scattering techniques such as ellipsometry, reflectivity, and dynamic light scattering using wavelengths on the order of those of visible light or longer. The recent application of scattering of shorter-wavelength radiation—in particular, x rays, and also neutrons—has provided detailed microscopic information previously unavailable. It is expected that this rapid recent progress in understanding the structure of liquid interfaces will continue as a greater variety of systems are studied.

This article surveys determinations of structure on the molecular length scale at liquid–liquid and liquid–vapor interfaces. Important phenomena also occur at the liquid–solid interface and are reviewed elsewhere in this Encyclopedia (see SURFACE PROCESSES; SURFACES AND INTERFACES OF SOLIDS, STRUCTURE OF).

1. INTERFACIAL FREE ENERGY AND MACROSCOPIC DESCRIPTIONS OF LIQUID INTERFACES

Josiah W. Gibbs (1957) pioneered the formulation of surface thermodynamics most often used. For a system consisting of two fluid phases separated by an interface, the differential of the Helmholtz free energy F can be written as

$$dF = -SdT - pdV + \sum \mu_i dn_i + \gamma dA, \qquad (1)$$

where S is the entropy, T the temperature, p the pressure, V the volume, μ_i the chemical potential of molecular species i, n_i the number of moles of species i, γ the surface or interfacial tension, and A the interfacial area. Equation (1) supplements the expression for dF of a bulk system by the addition of the last term, γdA. This term represents the free energy required to create a differential area of interface, dA, thereby defining the proportionality constant, the surface or interfacial tension γ, given in units of energy per area. As seen from Eq. (1), the tension γ is given by $\gamma = (\partial F/\partial A)_{T,V,n_i}$. Alternatively, Eq. (1) can be rewritten in terms of the grand canonical potential Ω, where $\Omega = F - \Sigma\mu_i n_i$, resulting in

$$d\Omega = -SdT - pdV - \sum n_i d\mu_i + \gamma dA. \tag{2}$$

This leads to an alternative definition of the tension as $\gamma = (\partial\Omega/\partial A)_{T,V,\mu_i}$.

In Gibbs's formulation, thermodynamic surface functions are described in terms of a mathematical surface, a dividing surface, chosen to be near the macroscopically observable interface and everywhere parallel to it (Fig. 1). Once chosen, the dividing surface determines the volume V^α for each phase α (e.g., $\alpha = 1$ or 2), with the total volume given by $V^{(1)} + V^{(2)} = V$. Far from the dividing surface it is expected that densities of extensive thermodynamic quantities are constant. These quantities can then be written as, for example, $n_i^\alpha = \rho_{i,\mathrm{mol}}^\alpha V^\alpha$, where $\rho_{i,\mathrm{mol}}^\alpha$ is the molar density of component i far from the dividing surface in phase α. When these quantities for the two phases are summed, for example $n_i^{(1)} + n_i^{(2)}$, it is found generally that they are not equal to the total n_i. The remainder is referred to as the surface quantity n_i^s, where $n_i^{(1)} + n_i^{(2)} + n_i^s = n_i$. This is related to the excess or surface density, also known as the adsorption, $\Gamma_i = n_i^s/A$.

For insight into the microscopic nature of excess surface quantities, consider a two-component, two-phase system with a planar interface located at $z = 0$, as shown in Fig. 1. The dividing surface is also chosen to be at $z = 0$. If we assume that within each phase the molar densities are independent of position, that is, constant for $z > 0$ (phase 1) and for $z < 0$ (phase 2), then, in this special case, the sum $n_i^{(1)} + n_i^{(2)}$ is equal to the total n_i. If the molar density profile at the interface is not a step function, i.e., the molar densities of the components near the interface do not abruptly change from one bulk value to the other, but rather change smoothly (see Sec. 2.1), then the surface quantity n_i^s will not usually be zero and may represent a net excess or deficit of component i at the interface (note that n_i^s also depends upon the choice of dividing surface).

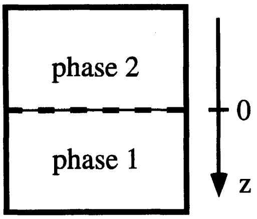

FIG. 1. Cartoon of two-phase liquid system in a jar with an idealized planar interface. The dashed line, directly over the interface, indicates one of many possible locations of the Gibbs dividing surface.

Similar expressions can be written for other extensive quantities such as F and Ω, yielding the surface functions F^s and Ω^s. Under suitable conditions of integration, $F = -pV + \gamma A + \Sigma\mu_i n_i$. The surface Helmholtz free energy F^s is given by $F^s = \gamma A + \Sigma\mu_i n_i^s$, where $F^s + F^{(1)} + F^{(2)} = F$ and F^α for each homogeneous phase can be defined in a manner similar to the definition for n_i^α. If the Gibbs dividing surface is chosen such that $\Sigma\mu_i n_i^s = 0$, then the interfacial tension is given as the interfacial free energy per unit area, $\gamma = F^s/A$. A similar integration of Eq. (2) yields $\Omega = -pV + \gamma A$. The interfacial tension in the grand canonical ensemble is independent of the position of the dividing surface, $\gamma = \Omega^s/A$. A complete discussion of surface thermodynamics is presented by Rowlinson and Widom (1982) and by Defay *et al.* (1966). (Note that the classical thermodynamic arguments do not explicitly include gravity, but assume that it is weak enough not to distort the density gradients across the interface, yet strong enough to produce a stable planar, or otherwise shaped, interface).

Prior to Gibbs's development of interfacial thermodynamics, liquid interfaces were treated as if they were mechanical membranes under tension. Laplace showed that there is a differential pressure Δp across a curved interface. For example, a spherical gas bubble in a liquid would collapse because of the surface tension if the pressure inside was not greater that the pressure outside. The Laplace pressure is given by $\Delta p = \gamma(1/R_1 + 1/R_2)$, where R_1 and R_2 are orthogonal radii of curvature of the interface ($R_1 = R_2 \equiv R$ for the case of the spherical bubble of radius R). Describing interfaces as membranes under tension leads naturally to important relations between interfacial tension and gravity that determine the macroscopic shape of interfaces, thereby allowing for the measurement of interfacial tensions. For example, the capillary rise of a liquid interface in a small-bore cylindrical tube above that of the bulk interface (see Fig. 2) is given by a balance of the hydrostatic pressure determined by gravity and the Laplace pressure. For a bore of sufficiently small radius r, the liquid–vapor (or liquid–liquid) interface in the tube follows the shape of a spherical surface. The surface or interfacial tension between the lower and upper phase is then given in terms of the capillary rise h as $\gamma = \Delta\rho_m ghr/2\cos\theta$, where θ is the contact angle between the lower phase and the capillary wall, g is the gravitational acceleration, $R_1 = R_2 \equiv R = r/\cos\theta$, and $\Delta\rho_m$ is the mass density difference between the upper and lower phases (Adamson, 1990).

The capillary rise can also be determined by noting that the rise of the lower phase results in the replacement of an interface between the capillary wall and the upper phase with that of a wall–lower-phase interface. This can be quantified by using Young's equation that relates the surface tensions at the contact line of a liquid in contact with a solid wall, $\gamma_{sv} = \gamma_{ls} + \gamma_{lv}\cos\theta$, where s, l, and v stand for solid, liquid, and vapor (the upper phase), and θ is the contact angle in Fig. 2. The difference between the interfacial free energies, $\gamma_{sv} - \gamma_{ls}$, can be introduced into the equation for the capillary rise, $h = 2(\gamma_{sv} - \gamma_{ls})/\Delta\rho_m gr$. If it is more favorable for the wall to be in contact with the lower phase ($\gamma_{ls} < \gamma_{sv}$), then the lower phase rises (as is true for capillary rise of a water–vapor interface in a glass capillary). If the opposite is true, as for mercury, then the level of the lower phase in a glass capillary drops below the height of the bulk interface.

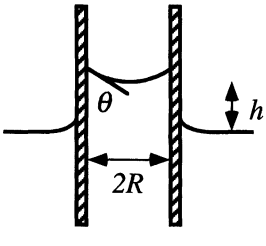

FIG. 2. Capillary rise.

Typical values for liquid–vapor surface tensions are 20–40 mN/m. The water–vapor surface tension of ~72 mN/m at 25 °C is unusually high, though surface tensions for liquid metals are much higher, e.g., ~710 mN/m for the liquid-gallium–vapor interface near 30 °C. The liquid–liquid interfacial tension of *n*-heptane and water is 50.2 mN/m at 20 °C. Addition of a surfactant can reduce the water surface tension by tens of mN/m. At the extreme, the interfacial tensions of liquid–liquid interfaces in oil, water, and amphiphile microemulsion systems have been measured as low as 0.001 mN/m. Tables of values can be found in Bonnet and Pike (1972), Jasper (1972), Weast (1989), and Adamson (1990).

Waves are easily observed on liquid interfaces. A hydrodynamic description of these waves leads to the dispersion relation given by

$$\omega^2 = gq + (\gamma/\Delta\rho_m)q^3, \tag{3}$$

where ω is the frequency, $q = 2\pi/\lambda$ is the wave number (λ, the wavelength) of the surface waves, and g is the gravitational acceleration (Barber and Loudon, 1989). Gravity and surface tension have an equal influence on the surface waves for $q_x = (\Delta\rho_m g/\gamma)^{1/2} \equiv 2^{1/2}/a$ (where a is the capillary constant) or, equivalently, for $\lambda_x = 2\pi/q_x$. For the water surface at 0 °C, $a = 3.93$ mm and $\lambda_x \cong 17$ mm. For surface waves of wavelength $\lambda \ll \lambda_x$ (denoted capillary waves), the frequency is determined primarily by surface tension, while for wavelengths $\lambda \gg \lambda_x$ (denoted gravity waves), gravity dominates. The full treatment of the hydrodynamics of liquid interfaces relies on solutions of the Navier–Stokes equation and is beyond the scope of this article (Landau and Lifshitz, 1986).

These three descriptions of liquid interfaces (classical thermodynamic, mechanical membrane, and hydrodynamic) are adequate to describe most macroscopic behavior. To venture beyond this to describe the microscopic structure requires the use of statistical thermodynamics and new experimental methods capable of probing on the molecular length scale. The study of the density variation when passing from one bulk phase through the interface to the other bulk phase has received the most emphasis and is discussed in the next section.

2. INTERFACIAL PROFILES: THE MICROSCOPIC VIEW OF INTERFACIAL STRUCTURE

The origin of the interfacial tension is the microscopic arrangement of molecules in the transition region between bulk phases. Van der Waals related the interfacial tension to interfacial structure, specifically to the longitudinal number density profile $\rho(z)$ along the coordinate z normal to the interface. The theory was extended by Cahn and Hilliard (1958), and modern versions of van der Waals–type theories have been reviewed by Rowlinson and Widom (1982). Although it is recognized that the molecules are in constant thermal motion, these theories are concerned only with time-averaged profiles.

Buff *et al.* (1965) suggested an alternative viewpoint in which the interface is treated as a membrane whose shape varies as a result of thermal molecular motion. The shape is analyzed using capillary waves at the interface as the relevant collective coordinates. The longer-wavelength fluctuations of the interface cost very little energy and act to increase the effective width of the interface. This width is determined from an effective longitudinal density profile $\rho(z)$, calculated by averaging the structure over the transverse coordinates, x and y, at a particular height z. A hybrid approach, proposed by Weeks (1977), combines the intrinsic longitudinal density profile predicted from van der Waals theories and capillary-wave roughening of this intrinsic interface.

The characteristic length scales of the longitudinal interfacial profile can be estimated by dimensional analysis of the interfacial tension. The characteristic interfacial free energy may be estimated from the entropic loss of confining a molecule to the interface and is given by k_BT (k_B is Boltzmann's constant). For $\gamma \cong 40$ mN/m, the characteristic interfacial length scale ($\sim\sqrt{k_BT/\gamma}$) corresponding to the interfacial width is 0.3 nm; for small tensions characteristic of microemulsions such as $\gamma \cong 0.04$ mN/m, the width is 10 nm (Safran, 1994). For fluids near a critical point where the interfacial width is determined by a diverging correlation length (and $\gamma \rightarrow 0$), the width can be made as large as hundreds of nanometers. X-ray scattering has been most effective at studying interfacial profiles of simple liquids away from the critical point, and light scattering has been used near criticality (Langevin, 1992).

2.1 Density Profiles

In the classical van der Waals theory for the liquid–vapor interface, the inhomogeneity of the density at the interface is accounted for by adding to the free energy a term proportional to the square of the gradient of the density. Without this term the classical theory predicts a step-function profile for which the interface is an infinitely thin region representing an abrupt change from one bulk density to the other. However, the presence of the squared-gradient term implies an infinite free-energy cost for a step-function profile and, instead, predicts a diffuse interface with a gradual, continuous change in density from one bulk phase to the other as illustrated in Fig. 3.

Following the development in Rowlinson and Widom (1982), the number-density profile $\rho(z)$ along the coordinate z normal to the surface is given by

$$m\frac{d^2\rho(z)}{dz^2} = \mu[\rho(z),T] - \mu_0 = -\frac{dW(\rho)}{d\rho}, \tag{4}$$

where $\mu[\rho(z),T]$ is the chemical potential as a function of density and temperature and μ_0 is the uniform chemical potential in the bulk two-phase system. $W(\rho)$ is, therefore, an excess free-energy density due to the presence of the interface, specifically, the excess density of the grand potential Ω. The profile $\rho(z)$ is determined by solving Eq. (4) subject to the boundary conditions that $\rho = \rho_{l,v}$ at $z =$

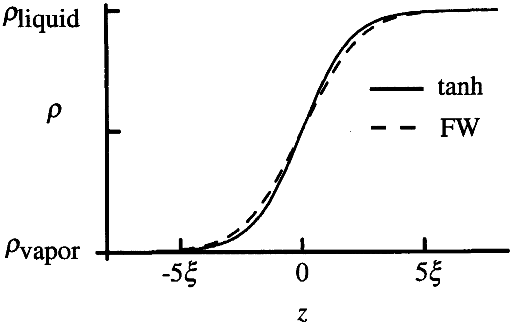

FIG. 3. The hyperbolic tangent (tanh) and Fisk–Widom (FW) intrinsic interfacial profiles.

$\pm\infty$. [The quantity m is given by the second moment of the attractive part of the intermolecular potential $u(r)$,

$$m = -\tfrac{1}{6}\int_{r>d} r^2 u(r)dr, \tag{5}$$

where $r > d$ indicates that the integral is taken only over the attractive part of the potential.] The surface tension is calculated from the density profile as

$$\gamma = \int_{-\infty}^{\infty} m\rho'(z)^2 dz = -2\int_{-\infty}^{\infty} W[\rho(z)]dz.$$

At a critical point (e.g., the liquid–vapor critical point) the two bulk phases become one (Hahne, 1983). As the critical point is approached, the densities of the two phases approach each other continuously. In a van der Waals mean-field theory, $W(\rho)$ near the critical point can be approximated as a quartic, which has two minima representing the two densities of the bulk phases. In this case, an explicit form for the profile (Fig. 3) can be given as

$$\rho(z) = \rho_c + \tfrac{1}{2}(\rho_l - \rho_v)\tanh[(z - z_0)/2\xi], \tag{6}$$

where ρ_c is the critical density, z_0 is arbitrary, and ξ is the Ornstein–Zernike correlation length of either bulk phase (ξ represents the exponential decay length of correlations between pairs of molecules). The interfacial width is of the same order as the bulk correlation length, which diverges as the critical point is approached in the manner $\xi \sim t^{-\nu}$, where $t = |T_c - T|/T_c$. The van der Waals theory is a mean-field theory with $\nu = \frac{1}{2}$ and is not valid very near the critical point. However, a slightly more accurate profile is given by substituting the nonclassical value for this exponent, $\nu \cong 0.63$. Fisk and Widom (1969) included realistic critical-point singularities in $W(\rho)$ and obtained

$$\rho(z) = \rho_c + \tfrac{1}{2}(\rho_l - \rho_v) \times \left[\frac{\sqrt{2}\tanh[(z - z_0)/2\xi]}{[3 - \tanh^2[(z - z_0)/2\xi]]^{1/2}}\right]. \tag{7}$$

This profile is similar to the classical profile given in Eq. (6), although slightly less steep (see Fig. 3). Experimental results to be discussed indicate a strong preference for the Fisk–Widom profile near criticality.

Van der Waals theories rely on a local thermodynamics, assuming that a local free-energy density can describe the inhomogeneous region of the interface that contains densities different from the stable bulk densities (at the chosen T and p). This local free-energy density is then a function of z, and it is assumed that it can be related to a bulk free energy evaluated at that intermediate density. The validity of these assumptions is not firmly established, and modern statistical mechanical theories of inhomogeneous liquid structure are written in terms of correlation functions that describe the correlations between pairs (and larger groups) of molecules. For example, the attractive part of the intermolecular potential, $u(r)$, in Eq. (5) is replaced by $-k_BTc(r)$, where $c(r)$ is the two-particle direct correlation function. In the modern theory, it is seen that Eq. (4) is an approximation to an integral equation of a functional of the entire density profile and is, therefore, nonlocal (Rowlinson and Widom, 1982). Integral equations have been used to determine numerical solutions for interfacial profiles. However, these solutions are not used often in the analysis of experimental data (though surface-tension values computed from integral equations have been compared, in some cases favorably, with measurements; see Rowlinson and Widom, 1982, for a discussion). In this article, we follow the experimental research literature in comparing measurements with the simpler analytic forms for the profiles.

2.2 Interfacial Fluctuations

In the original formulation of the capillary-wave theory of interfaces away from critical points, Buff *et al.* (1965) calculated the work W_{cw} necessary to create transverse oscillations in an interface of zero thickness. This consists of the work needed to deform the interface in a gravitational field and the work needed to create additional surface area of an interface with surface tension γ_{cw}. The interface is modeled as a taut membrane whose vibration modes (surface capillary waves) occur with a probability given by the Boltzmann distribution, $\exp(-W_{cw}/k_BT)$. In this theory the interface has a step-function *intrinsic* profile as given by van der Waals–

type theories discussed in the previous section. However, an effective profile can be calculated from the mean density $\rho(z)$ at any height z averaged over the values of the two bulk phases.

The capillary-wave theory postulates an effective Hamiltonian,

$$H_{\mathrm{cw}} = \gamma_{\mathrm{cw}} \int d^2s\left(\sqrt{1 + |\nabla\zeta|^2} - 1\right) + \tfrac{1}{2}\,\Delta\rho_m g \int d^2s\,\zeta^2, \tag{8}$$

where the first term accounts for the free-energy cost of the area change and the second term describes the gravitational-energy cost of a fluctuation. The interface lies, on average, in a two-dimensional plane whose coordinates are given by $\mathbf{s} \equiv (x,y)$, and the configuration of the interface is given by $z = \zeta(\mathbf{s})$, which represents the displacement of the interface from the x-y plane (Fig. 4). The capillary-wave theory can be extended to other dimensions, with interesting results reviewed by Gelfand and Fisher (1990).

Buff *et al.* identified the surface tension γ_{cw} used in their model with a bare surface tension of the interface without capillary waves. In an exact calculation for capillary waves in a one-dimensional interface of the two-dimensional Ising model, Abraham and Reed (1974, 1976) demonstrated that γ_{cw} is the actual surface tension, γ. Besides this result, there is not a firm theoretical foundation to guide the interpretation of γ_{cw} in the analysis of experimental data. In practice, most data analysis assumes $\gamma_{\mathrm{cw}} = \gamma$.

If the fluctuations are of long wavelength or small amplitude, the square root in Eq. (8) can be expanded, resulting in a Gaussian form for the effective Hamiltonian,

$$H_{\mathrm{cw}} = \tfrac{1}{2}\,\gamma_{\mathrm{cw}} \int d^2s\left[|\nabla\zeta|^2 + \left(\frac{\zeta}{\xi_{\|,\mathrm{cw}}}\right)^2\right], \tag{9}$$

where the capillary-wave interfacial correlation length $\xi_{\|,\mathrm{cw}}$ is given by $\xi^2_{\|,\mathrm{cw}} = \gamma_{\mathrm{cw}}/\Delta\rho_m g$ and determines the exponential decay of the interfacial correlations given by the height–height correlation function $\langle\zeta(\mathbf{s})\zeta(\mathbf{s}')\rangle$ (Gelfand and Fisher, 1990). The actual correlation length, $\xi_{\|}$, is given by $\xi_{\|}^2 = \gamma/\Delta\rho_m g = a^2/2$, where a is the capillary length introduced earlier and is of macroscopic dimensions.

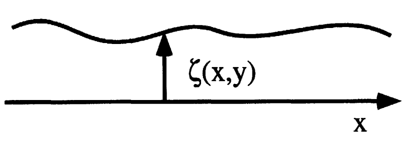

FIG. 4. Interfacial configuration along *x*-axis.

Equation (9) can be analyzed by expressing the interfacial configuration $\zeta(\mathbf{s})$ in terms of decoupled harmonic interfacial waves (capillary waves). The average over the canonical ensemble of all accessible interfacial waves, $\langle\zeta^2\rangle$, is given (in three dimensions) as

$$\langle\zeta^2\rangle = \frac{k_{\mathrm{B}}T}{2\pi\gamma_{\mathrm{cw}}} \int_{q_{\min}}^{q_{\max}} \frac{q\,dq}{q^2 + \xi^{-2}_{\|,\mathrm{cw}}}, \tag{10}$$

where q is the wave number of the surface capillary wave. The cutoff $q_{\max}$ represents the shortest-wavelength capillary mode on the interface. It is reasonable to expect that the shortest wavelength should be no smaller than the molecular size, thereby setting an upper limit for $q_{\max}$. This short-wavelength behavior of the capillary-wave model is under debate, and currently, there is not a consensus on how to choose $q_{\max}$ from either theoretical or experimental considerations. In the Buff *et al.* model, $q_{\max}$ is chosen as inversely proportional to the interfacial width, $q_{\max} = \pi/\langle\zeta^2\rangle^{1/2}$. The cutoff $q_{\min}$, corresponding to the largest-wavelength capillary mode, is determined by the system size as $q_{\min} = \pi/D$ (e.g., for a square interface of area D^2). Individual experimental techniques may not be sensitive to wavelengths as large as the system size, thereby increasing the effective $q_{\min}$. By integrating Eq. (10) we have

$$\langle\zeta^2\rangle = \frac{k_{\mathrm{B}}T}{4\pi\gamma_{\mathrm{cw}}} \ln\left(\frac{q^2_{\max} + \xi^{-2}_{\|,\mathrm{cw}}}{q^2_{\min} + \xi^{-2}_{\|,\mathrm{cw}}}\right). \tag{11}$$

In the thermodynamic limit ($D = \infty$) the interfacial width $\langle\zeta^2\rangle$ diverges as $g \to 0$. Similarly for fixed $g = 0$, $\langle\zeta^2\rangle$ diverges as $D \to \infty$. This delocalization of the interface is not consistent with the van der Waals theories.

The capillary theory of Buff *et al.* with a step-function intrinsic profile leads to an error-function profile

$$\rho(z) = \tfrac{1}{2}(\rho_l + \rho_v) + \tfrac{1}{2}(\rho_l - \rho_v)\,\mathrm{erf}(\sqrt{\pi}z/L) \text{ with}$$
$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}dt, \tag{12}$$

where $\sqrt{24}L = \xi$ [to be compared with Eq. (7)]. The central region of this profile is almost identical to the Fisk–Widom profile in Eq. (7) and only differs slightly in the wings (for $|z| \cong 4\xi$).

An important extension of the capillary-wave theory was made by Weeks (1977) in which a smooth intrinsic profile is further roughened by capillary waves. Here, $q_{\max} \cong \pi/\xi$, where ξ is the bulk correlation length that sets the width of the intrinsic profile. This hybrid model includes an intrinsic profile for in-plane length scales less than ξ and capillary waves on longer length scales. The profile $\rho(z)$ is a convolution of the error-function capillary-wave profile and the intrinsic profile. The surface tension γ_{cw} is taken to be the actual surface tension, γ.

Meunier (1987) introduced coupling between surface waves of different q that results in a q-dependent surface tension $\gamma(q)$. The shorter-wavelength capillary waves renormalize the surface tension associated with longer-wavelength waves, as first suggested by Stillinger (1967). Meunier's analysis leads to the expression for the interfacial width

$$\langle \zeta^2 \rangle \propto k_{\mathrm{B}} T \int_{q_{\min}}^{q_{\max}} \frac{q\,dq}{\gamma_{\mathrm{cw}}(q) q^2 + \Delta\rho_m g}, \tag{13}$$

where $\gamma_{\mathrm{cw}}(q) \approx \gamma + \tilde{\nu} q^2 + \ldots$. In this theory the intrinsic interface is a step function, and the coupling between the capillary waves introduces a natural cutoff, $q_{\max}$. Sengers and van Leeuwen (1989) built upon this coupled-mode analysis to introduce a hybrid model accounting for both an intrinsic interface and capillary waves.

The q-dependent surface tension has been used recently to analyze optical measurements of interfaces. However, a direct experimental verification of this q dependence is not available. On a related topic, the analysis of measurements of surfactant-covered surfaces utilizes a q^4 term in the denominator of Eq. (13) to account for the elasticity of the interface. In this case there will be an additional energy cost associated with bending under area-conserving conditions, and the prefactor of the q^4 term is then proportional to an elastic bending modulus (Safran, 1994).

Napiórkowski and Dietrich (1993) used microscopic density-functional theory for inhomogeneous media to derive a nonlocal effective Hamiltonian for a fluctuating liquid–vapor interface with a step-function intrinsic profile. The Hamiltonian is expressed in terms of $\tilde{w}(|\mathbf{r}' - \mathbf{r}|)$, a particle–particle interaction potential of, e.g., long-range van der Waals forces, and is written as

$$\begin{aligned} H([\zeta]) &= \int_{\mathbf{s}} d^2 s\, h_0(\mathbf{s},[\zeta]) \\ &\quad + (m/2)(\Delta\rho) g \int d^2 s\, \zeta^2, \text{ with} \\ h_0(\mathbf{s},[\zeta]) &= -\tfrac{1}{2}(\Delta\rho)^2 \int_{\mathbf{s}} d^2 s \int_0^{\infty} dz \int_0^{\zeta(\mathbf{s}')-\zeta(\mathbf{s})} \\ &\quad \times dz'\, \tilde{w}(|\mathbf{r}' - \mathbf{r}|), \end{aligned} \tag{14}$$

where m is the mass of the particles. The second term in $H([\zeta])$ corresponds to the gravitational term in Eq. (8). The first term in Eq. (8) can be derived from the first term of $H([\zeta])$ by expanding $\zeta(x',y') - \zeta(x,y)$ in powers of $x' - x$ and $y' - y$ and keeping only the linear terms; that is, the Hamiltonians are equivalent in the limit of slowly varying interfaces. This nonlocal Hamiltonian behaves very differently from the standard local Hamiltonian H_{cw} [Eq. (8)] in the region of large q. Similar to the theory of Meunier, this Hamiltonian predicts a q-dependent surface tension. Napiórkowski and Dietrich have shown that this Hamiltonian implies that the interpretation of experimental measurements is strongly dependent on the wave-vector cutoff, $q_{\max}$.

3. ADSORPTION AND WETTING PHENOMENA

In a single-phase liquid mixture of two or more components the component whose surface tension as a pure system is lowest will be adsorbed preferentially at the liquid–vapor interface. Similarly, addition of a low-surface-tension impurity to a liquid will lower the surface tension of the mixture because of adsorption of the impurity at the surface. These, and related, adsorption phenomena are described by the Gibbs adsorption equation, given by

$$d\gamma + \eta_s dT + \Sigma \Gamma_i d\mu_i = 0, \tag{15}$$

which relates the change in the surface tension to changes in the temperature T and the chemical potentials of the i components in the bulk. Here η_s is the excess or surface entropy density, $\eta_s = S^s/A$. The adsorption can be defined in terms of the interfacial profile $\rho(z)$:

$$\Gamma_i = \int_{-\infty}^{\infty} [\rho_i(z) - \rho_{\alpha,\beta}]dz, \tag{16}$$

where $\rho_{\alpha,\beta}$ is the value of ρ on the α or β side of an arbitrarily chosen dividing surface. As an example, consider a two-component system at constant temperature with a dividing surface chosen so that $\Gamma_1 = 0$. Equation (15) is then rewritten as $\Gamma_2 = -(\partial\gamma/\partial\mu_2)_T$. In the case of water–ethanol mixtures (where water is component 1 and ethanol is component 2), measurements of the surface tension as a function of ethanol content yield $(\partial\gamma/\partial\mu_2)_T < 0$. Therefore $\Gamma_2 > 0$, indicating a surface excess of ethanol (Adamson, 1990). Using Eqs. (15) and (16), measured changes in surface tension can, in principle, be related to measurements of the microscopic density profiles.

Fisher and de Gennes (1978) proposed that adsorption at an interface can be significantly enhanced near a critical point. For example, consider a single-phase binary liquid mixture consisting of components A and B in a sample cell whose walls preferentially attract component A (this attraction is typically characterized in terms of a surface field h_1). When the mixture is in a single phase far from the critical point (where the single phase separates into two phases), a thin layer of component A (typically on the order of a monolayer) will be found adsorbed to the walls. As the critical point is approached, the pair correlation length between A molecules in the bulk diverges, and Fisher and de Gennes predict that the adsorption of A will be enhanced at the walls. Heuristically, the walls provide a boundary condition that pins bulk critical fluctuations of component A to the wall. This critical adsorption profile should scale as

$$\rho(z) - \rho_{\text{bulk}} \cong \rho_0 t^{\beta} f(z/\xi, h_1/t^{\Delta_1}), \tag{17}$$

where $\Delta_1 > 0$ implies that the scaled surface field h_1/t^{Δ_1} diverges as $t \to 0$ ($t = |T_c - T|/T_c$), strongly enhancing the attraction of the preferentially adsorbed component. In this strong-field limit, the scaling function $f_{\text{sat}}(z/\xi) \equiv f(z/\xi,\infty)$. At criticality, $t = 0$, the scaling function $f(z/\xi) \approx f_0(z/\xi)^{-\beta/\nu}$ as $z/\xi \to 0$, giving an algebraically decaying interfacial profile. Away from criticality, the decay of the interfacial profile is exponential and is determined by the bulk correlation length with $f(z/\xi) \approx f_\infty e^{-z/\xi}$ as $z/\xi \to \infty$. This leads to the expectation that the adsorption diverges as $t^{\beta-\nu}$ when $t \to 0$ (where $\beta = 0.325$ and $\nu = 0.63$). Several groups made the earliest measurements on organic binary liquids (Beaglehole, 1980; Beysens and Leibler, 1982; Franck and Schnatterly, 1982), and others are reviewed by Liu and Fisher (1989). Measurements by Franck and co-workers have explored the role of the surface field h_1, including measurements of the strong- and weak-field limits (Schlossman *et al.*, 1985; Desai *et al.*, 1995). Although there are numerous measurements of critical adsorption, the full scaling of the profile has not been completely confirmed (Smith and Law, 1995).

A wetting transition is illustrated in Fig. 5 where a droplet of phase 3 situated at the interface between phases 1 and 2 [Fig. 5(a)] spreads to form a continuous film separating phases 1 and 2 [Fig. 5(b)]. In the latter situation, phase 3 is said to have wetted the interface, and Antonow's rule gives the relationship between the interfacial tensions:

$$\gamma_{12} = \gamma_{13} + \gamma_{32}. \tag{18}$$

If phase 3 does not wet the 1–2 interface, then $\gamma_{12} < \gamma_{13} + \gamma_{32}$ [Fig. 5(a)]. Cahn (1977) introduced the idea of a wetting transition expected to occur ubiquitously in systems near their critical point. Since then numerous theoretical and experimental studies of wetting transitions have been published. Here we mention a few highlights of these

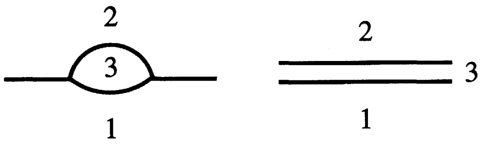

(a) (b)

FIG. 5. (a) Incomplete wetting. (b) Complete wetting of phase 3 at the 1–2 interface.

studies and refer to them later in the discussion of experimental data (for a review, refer to Sullivan and da Gama, 1986, and Dietrich, 1988).

Some of the phenomena associated with wetting transitions are summarized in Fig. 6, which illustrates the $(\Delta\mu,T)$ plane in the neighborhood of the liquid–vapor critical point for a one-component system, where $\Delta\mu = \mu - \mu_{\text{coexistence}}$ (the chemical-potential difference from liquid–vapor coexistence) (Pandit *et al.*, 1982). In this circumstance, phase 1 in Fig. 5(a) is an inert phase such as a container wall, and wetting occurs when a macroscopic layer of liquid spreads at the wall in the presence of bulk vapor. Along the horizontal axis, $\Delta\mu = 0$, for $T < T_c$ bulk liquid and vapor coexist. As $T \rightarrow T_w$, the wetting transition temperature, from below (along the $\Delta\mu = 0$ line), a first-order transition at the vapor-wall interface occurs at which the adsorption of the liquid phase at the wall diverges to give a macroscopic liquid layer (complete wetting). With further increase of the temperature (for $\Delta\mu = 0$), the system reaches the bulk critical point at $T = T_c$. Similarly, following path (a) the adsorption increases by a finite amount at the prewetting line. The prewetting transition that occurs at the prewetting line consists of a finite change from a thin layer adsorbed at the wall to a thick adsorbed layer. Along the prewetting line, there is predicted to be coexistence of thin and thick layers on the surface. The magnitude of this finite change in thickness increases for temperature scans at progressively smaller $\Delta\mu$ and diverges as $\Delta\mu \rightarrow 0$, at the point where the prewetting line joins the coexistence line at T_w. The magnitude of the change in thickness vanishes at C_{pre}, the prewetting critical point.

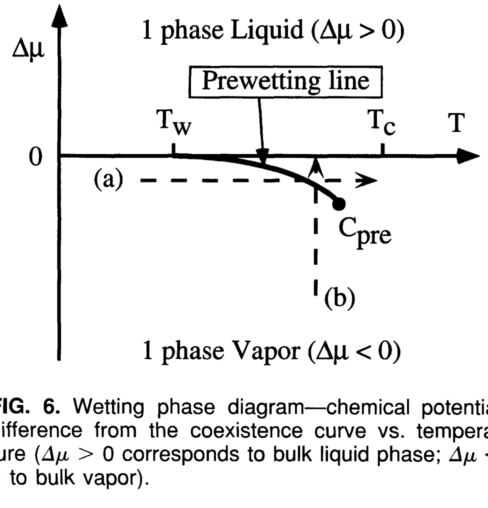

FIG. 6. Wetting phase diagram—chemical potential difference from the coexistence curve vs. temperature ($\Delta\mu > 0$ corresponds to bulk liquid phase; $\Delta\mu < 0$ to bulk vapor).

Alternatively, on fixing the temperature at $T < T_w$ and scanning $\Delta\mu$ from $\Delta\mu < 0$ to $\Delta\mu = 0$, the adsorption is expected to increase only slightly to a finite microscopic value at the coexistence curve (incomplete wetting). For $T > T_w$ along path (b) in Fig. 6, the same variation of $\Delta\mu$ results first in a finite change in adsorbed layer thickness as the prewetting line is crossed and then in a diverging layer thickness as the coexistence curve is approached (complete wetting). For many experimental systems first-order wetting transitions are observed, yet the prewetting line is not. It is believed that this occurs because the prewetting line is too close to the coexistence line to be observable. However, for several systems that are very weakly bound to the substrate, the prewetting transition has been unambiguously observed. This is the case for ^{4}He on the alkali-metal surfaces Rb and Cs (and also observed for H_2 and D_2 on Rb and, possibly, for Ne on Cs) (Taborek and Rutledge, 1994; Cole, 1995).

This phenomenology of wetting and prewetting transitions was originally developed using theories incorporating short-range interactions only. These surface phenomena are very sensitive to the form and magnitude of the interactions, which include both the interparticle potentials and wall or surface potentials. For the appropriate choice of potentials, the first-order wetting transition is expected to become second order. In this case the prewetting line does not appear. The adsorption and the interfacial profiles are also very sensitive to the potentials. For example, long-range forces such as van der Waals forces, electrostatic forces, and gravity can have a significant effect on the surface adsorption.

Similar phenomenology also occurs in the study of multilayer adsorption of, for example, low-temperature gases on solid graphite or, as discussed more extensively below, liquid-crystal layering at the free surface. In these cases, complete wetting corresponds to a buildup of a macroscopic number of adsorbed layers. Incomplete wetting corresponds to the surface saturating at a finite number of adsorbed layers. Closer approach to the bulk coexistence line then results in the growth of bulk islands that do not spread over the surface to form complete layers.

Understanding wetting phenomena is also of relevance for the study of first-order bulk transitions. Nucleation of a new phase may be more favorable at the walls of the container, especially if the new phase completely wets the walls. Wall effects on both spinodal decomposition and nucleation play a particularly important role for fluids confined in one or more directions such as in thin cells or in porous materials.

Because of limited space, this article does not review the vast number of experiments in this field. However, a few experimental results on wetting in liquid crystal systems are discussed for which measurements of the microscopic structure are available.

4. INTERFACIAL-PROFILE MEASUREMENTS

X-ray and light scattering have been particularly useful probes for determining interfacial profiles at liquid interfaces. X-ray reflectivity measurements probe the gradient of the longitudinal electron density profile, $d\rho_e/dz$. X-ray surface diffuse scattering directly probes the interfacial fluctuations. Similarly light-reflectivity measurements probe the width of the profile, but on longer length scales than for x-ray reflectivity, making it a useful technique closer to the critical point where the profile width is expected to diverge.

4.1 X-Ray Measurements of Interfaces far from Criticality

Specularly reflected x rays have a wave-vector transfer normal to the interface; therefore, they probe the longitudinal profile (specifically, the electron-density profile) averaged over the transverse directions. The kinematics for this reflection is shown in Fig. 7 where $\theta_{\text{in}} = \theta_{\text{scat}}$ for specular reflection and the wave-vector transfer $\mathbf{Q} = \mathbf{k}_{\text{scat}} - \mathbf{k}_{\text{in}}$ (the difference between the scattered and incident wave vectors) is given by $Q_z = (4\pi/\lambda)\sin\theta$ and $Q_x = Q_y = 0$ (λ is the x-ray wavelength, and the reflection angle $\theta \equiv \theta_{\text{in}}$).

With neglect of multiple scattering, the first Born approximation can be used to calculate the specularly reflected x-ray intensity $R(Q_z)$ (Pershan, 1990):

$$\frac{R(Q_z)}{R_F(Q_z)} \cong \left| \frac{1}{\rho_{e,\text{bulk}}} \int dz \, \frac{d\langle\rho_e(z)\rangle}{dz} \exp(iQ_z z) \right|, \qquad (19)$$

where $d\langle\rho_e(z)\rangle/dz$ is the electron-density gradient averaged over the x-y plane, $\rho_{e,\text{bulk}}$ is the electron density far below the surface, and $R_F(Q_z)$ is the Fresnel reflectivity for a smooth step-function interface. This approximation is accurate for reflection angles $\theta \gtrsim 4\theta_c$, where the critical angle for total x-ray reflection, $\theta_c \cong \sqrt{\rho_e r_e \lambda^2/\pi}$ ($r_e = e^2/mc^2 \cong 2.818$ fm), is small (for example, $\theta_{c,\text{water}} \cong 0.15°$ for $\lambda = 0.154$ nm). Note that θ is measured from the surface plane, in contrast to optical measurements for which it is conventional to measure θ from the surface normal. X-ray scattering from liquid surfaces is reviewed by Schlossman and Pershan (1992).

If the error-function profile in Eq. (12) is used in Eq. (19), the x-ray reflectivity is written as

$$R(Q_z)/R_F(Q_z) = \exp(-Q_z^2\sigma^2), \qquad (20)$$

where the constant σ is a measure of the interfacial width and is related to the L in Eq. (12) by $\sigma = L/\sqrt{2\pi}$. In the spirit of a hybrid model of the interface that describes an intrinsic profile further roughened by capillary waves, the interfacial width σ can be represented as a combination of an intrinsic width σ_0 and the capillary-wave contribution [Eq. (10)] (Braslau *et al.*, 1985, 1988; Schwartz *et al.*, 1990):

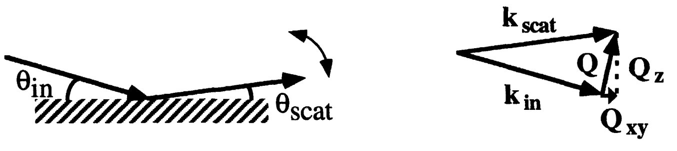

FIG. 7. Kinematics for surface scattering in the plane of incidence.

$$\sigma^2 = \sigma_0^2 + \frac{k_B T}{2\pi\gamma}\int_{q_{\min}}^{q_{\max}} \frac{q\,dq}{q^2 + \xi_\parallel^{-2}}, \qquad (21)$$

where $q_{\min}$ is determined by the instrumental resolution that sets the largest in-plane wavelength that the measurement probes. Consequently, once the instrumental resolution and the incident angle have been set, σ^2 is specified in terms of only two adjustable parameters: the intrinsic width of the electron-density profile σ_0 and the limit $q_{\max}$. Although the integration in Eq. (21) actually results in a logarithmic dependence of σ^2 on Q_z, the variation is slow, and data are often well represented by a value of σ^2 that is independent of Q_z but logarithmically dependent on the instrumental resolution.

X-ray specular reflectivity data from the surface of H_2O far from a critical point ($T \cong$ 20 °C), as shown in Fig. 8(a), are typical of the reflectivity from pure liquid surfaces (Schwartz *et al.*, 1990). These data were taken from the surface of H_2O in a vapor- and temperature-controlled Langmuir trough in which the surface tension was monitored *in situ* to be 72.5 ± 0.4 mN/m. There is a small region for $\theta < \theta_c$ where the reflectivity is essentially 100% (corresponding to total reflection), followed by a rapid fall. As shown in Fig. (8b) where the same data are normalized to $R_F(Q_z)$ and plotted as $R(Q_z)/R_F(Q_z)$ versus Q_z^2 on a semilogarithmic scale, the data are well represented by the form in Eq. (20) where $\sigma^2 = (0.270 \pm 0.003\ \text{nm})^2$. The two parameters, σ_0 and $q_{\max}$, cannot be distinguished by this set of measurements; therefore, the lines displayed in Fig. 8 were determined by fits to $q_{\max}$ only (with the intrinsic width, σ_0, set to zero). The lines in Figs. 8(a) and 8(b) display this fit, with $q_{\max} = \pi/(0.14 \pm 0.01\ \text{nm})$. By changing the slit size that determines the angular acceptance of the detector, the instrumental resolution can be varied, thereby varying $q_{\min}$. This allows the dependence of the integral in Eq. (21) on the limit $q_{\min}$ to be tested. Figure (8c) displays the comparison of the measured values of σ^2, as obtained from the best fits like that represented by the solid line in Fig. (8b), and those calculated for $Q_z = 4\ \text{nm}^{-1}$. This shows good agreement with the theoretical form of Eq. (21). However, these x-ray measurements cannot distinguish between the use of one particular value of $q_{\max}$ for an interface that is locally sharp and alternative hypotheses about the nature of the intrinsic interfacial structure.

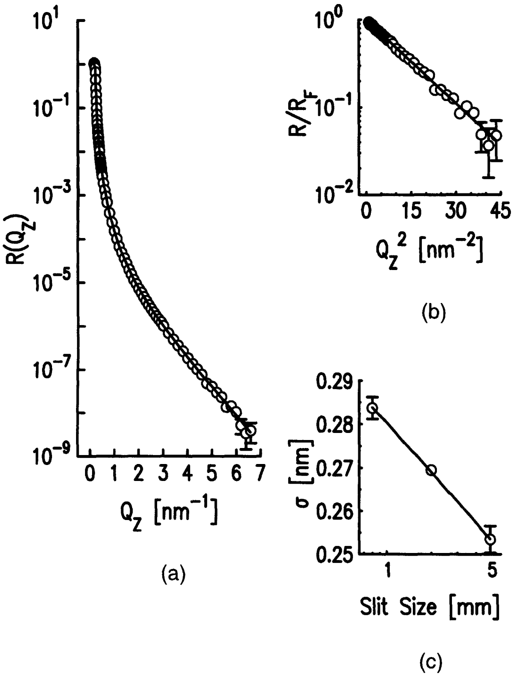

FIG. 8. Measured x-ray reflectivity from a water–vapor interface (see text for explanation).

In a related x-ray reflectivity measurement by Ocko *et al.* (1994) of the liquid–vapor interface of two long-chain alkanes, $C_{20}H_{42}$ and $C_{36}H_{74}$, the temperature was varied from 40 °C to 140 °C to distinguish the effects of the two terms in Eq. (21). The measurements are well described by the hybrid model represented by Eq. (21). The intrinsic width, σ_0, was determined to be $\sigma_0 = 0.11$ nm with a $q_{\max} = \pi/(0.71\ \text{nm})$ for $C_{20}H_{42}$ and $q_{\max} = \pi/(1.16\ \text{nm})$ for $C_{36}H_{74}$. These measurements could not be explained by the Buff *et al.* (1965) theory for which $\sigma_0 = 0$, or by the mode-coupling theory of Meunier (1987). Earlier measurements by Braslau *et al.* (1988) on the surfaces of CCl_4 and CH_3OH fixed $\pi/q_{\max}$ to be the molecular radius and fitted the reflectivity to σ_o. For CCl_4, σ_0 varied from 0.25 to 0.39 nm as the instrumental resolution was varied. These values are close to the molecular radius of 0.388 nm, but the variation with instrumental resolution is unexplained. For the surface of methanol the intrinsic profile was very small, $\sigma_0 = 0.07 \pm$

0.04 nm, much smaller than the molecular radius of 0.252 nm.

To summarize these reflectivity measurements, some indicate that both the intrinsic profile and the capillary-wave profile as given in Eq. (21) are necessary to fit the measurements. However, the physical significance of the numerical results for the fitting parameters, σ_0 and $q_{\max}$, is unclear and is not directly related to the predictions of hybrid capillary-wave models such as that given by Weeks for which both σ_0 and $\pi/q_{\max}$ should be of the order of the bulk correlation length.

Both neutron and x-ray specular reflectivity probe the scattering density averaged over the interfacial plane. For this reason, specular reflectivity measurements cannot distinguish between a rough interface and one with a graded scattering density profile of a certain width, but without roughness or fluctuations. In the model of thermally excited capillary waves, the reflectivity differs from the Fresnel reflectivity because the surface waves scatter the radiation out of the specular condition (the Fresnel reflectivity is expected for an ideal flat and sharp interface). This surface diffuse scattering directly measures the presence of fluctuations of the surface (recent reviews are given by Schlossman and Pershan, 1992, and Dietrich and Haase, 1995).

Figure 9 shows the measured intensity at a fixed angle of incidence (shown for two different incident angles) as the scattering angle θ_{scat} is varied (see Fig. 7). When the incident angle θ_{in} is equal to the scattering angle θ_{scat} there is a peak corresponding to the specular scattering. For other scattering angles, $\theta_{\text{scat}} \neq \theta_{\text{in}}$, the wave-vector transfer $\mathbf{Q} = \mathbf{k}_{\text{scat}} - \mathbf{k}_{\text{in}}$ (see Fig. 7) has a component in the plane of the interface, Q_{xy}, that probes the in-plane structure. In Fig. 9 a small peak appears when $\theta_{\text{scat}} = \theta_c$ (the angle for total reflection) at which the surface diffuse scattering is enhanced above the bulk diffuse scattering. On account of the multiple scattering that occurs near total reflection, the distorted-wave Born approximation is needed to calculate the scattering intensity near the critical angle θ_c as

$$I(\mathbf{Q}) \propto |T(\theta_{\text{in}})|^2 |S(\mathbf{Q})|^2 |T(\theta_{\text{scat}})|^2, \qquad (22)$$

where $S(\mathbf{Q})$ is the scattering structure factor, and $T(\theta)$ is the Fresnel transmission coefficient (Becker *et al.*, 1983; Deitrich and Wagner, 1984, 1985; Dosch, 1987). The function $|T(\theta)|^2$ varies from 0 to 4 as θ increases from 0 to θ_c and then falls to unity for $\theta \gg \theta_c$. The angular dependence of the two factors $T(\theta_{\text{in}})$ and $T(\theta_{\text{scat}})$ enables the surface scattering to be distinguished from other diffuse-scattering processes by giving rise to the peak at $\theta_{\text{scat}} = \theta_c$ for a rough interface. Note that an ideally smooth interface would not exhibit this surface enhancement peak.

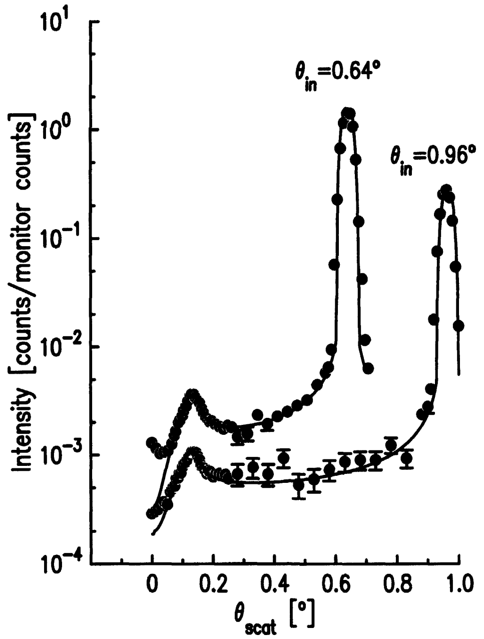

FIG. 9. Scattered x-ray intensity, in the plane of incidence, as a function of scattered angle for two fixed incident angles.

The solid lines in Fig. 9 represent the prediction for the scattered intensity I_D, written as

$$I_D \cong I_0 \frac{(r_e\rho_e)^2}{\theta_{\text{in}}\Delta\theta_{\text{in}}} \frac{k_B T}{\gamma} \int_{\theta_{\text{in}}-\Delta\theta_{\text{in}}/2}^{\theta_{\text{in}}+\Delta\theta_{\text{in}}/2} d\theta_{\text{in}} \int_{-\Delta\phi/2}^{\Delta\phi/2} d\phi \times \int_{\theta_{\text{scat}}-\Delta\theta_{\text{scat}}/2}^{\theta_{\text{scat}}+\Delta\theta_{\text{scat}}/2} d\theta_{\text{scat}} \frac{|T(\theta_{\text{in}})|^2 |T(\theta_{\text{scat}})|^2}{q^2 + \xi_{\|,\text{cw}}^{-2}}, \qquad (23)$$

where the integrations are over the instrumental resolution $\Delta\theta_{\text{in}}$, $\Delta\theta_{\text{scat}}$, and $\Delta\phi$ (ϕ is the azimuthal in-plane angle of the detector), and I_0 is the incident intensity. This expression for the scattered diffuse intensity assumes that the intrinsic profile is a step function, and is also an approximation for weak height–height correlations. More general expressions are discussed by Sinha *et al.* (1988) and Dietrich and Haase (1995). The only adjustable parameter for the fit in Fig. 9 is a small constant background, of the order of 10% of the peak at $\theta_{\text{scat}} = \theta_c$. The high quality of the fit indicates that the capillary-wave model is a good representation of the interfacial fluctuations with wavelengths from 40 to 8000 nm (Schwartz *et al.*, 1990). Combining these measurements with optical light scattering allows the capillary-wave theory to be tested over capillary wavelengths from 40 nm to ~1 mm (Loudon, 1984).

Sinha and co-workers have pointed out that the height–height correlation function $\langle \zeta(\mathbf{s})\zeta(\mathbf{s}')\rangle$ for capillary waves should exhibit a power-law decay as seen generally for displacement–displacement correlations in two-dimensional systems (Sinha *et al.*, 1988; Sanyal *et al.*, 1991). In general, power-law singularities of the form $Q^{\eta-2}$ appear in the structure factor $S(Q)$ of two-dimensional systems. For the experiments of Sanyal *et al.* on the surface of ethanol, this singularity appears in the scattered intensity as $I_D(Q_{xy}) \sim Q_y^{\eta-1}$ because of integration over the Q_x direction due to a large angular acceptance normal to the scattering plane (Sanyal *et al.*, 1991). In this case the exponent η is given by $\eta = (k_BT/2\pi\gamma)Q_z^2$. Their data are consistent with capillary-wave theory with a step-function intrinsic profile and show clearly the predicted power-law region over slightly more than a decade in Q_y. Their analysis yields an interfacial width of 0.69 ± 0.02 nm. Zhao *et al.* (1992) used similar techniques to study the liquid–vapor interface of toluene and of semidilute solutions of polybromostyrene and toluene. For sufficiently concentrated polymer solutions it is expected that the liquid-surface modes will be altered by the viscoelasticity of the polymer solution. However, in this case, even though the solutions were well into the entangled regime, they exhibited behavior in agreement with the standard capillary-wave theory (with a step-function intrinsic interface), as did the pure toluene surface.

4.2 Optical Measurements of Interfaces near Criticality

Optical reflectivity, ellipsometry, and dynamic light scattering have been used to study liquid interfaces. Optical reflectivity is sensitive to structures on length scales of ~10 nm or longer and has been used primarily to study near-critical interfaces where the interfacial thicknesses are on the order of 100 nm or more. Ellipsometry is sensitive to thinner layers and, as shown by Meunier, is more sensitive to the effect of smaller-wavelength capillary waves than is reflectivity. Dynamic light scattering is a useful probe of structure and dynamics on length scales for which $Q\lambda < 1$ (~10 to ~100 μm). The information on interfacial dynamics measured by dynamic light scattering from liquid interfaces is outside the scope of this article and has been reviewed recently by Langevin (1992).

Early optical reflectivity measurements near the critical temperature were made by Webb and co-workers on the liquid–vapor interface of SF_6 and the liquid–liquid interface of cyclohexane and methanol mixtures (Huang and Webb, 1969; Wu and Webb, 1973). These measurements were analyzed to determine both the functional form of the interfacial profile and the temperature scaling of the profile width. Since the profile width is expected to scale as the bulk correlation length, $\xi \sim t^{-\nu}$, where $t = |T_c - T|/T_c$, these measurements provided one of the early experimental determinations of the critical exponent ν. Similar to x-ray reflection, the optical reflectivity is interpreted using a Born approximation resulting in an equation very similar to Eq. (19) where the gradient of the index of refraction replaces the electron-density gradient, $d\rho(z)/dz \Rightarrow dn(z)/dz$. These measurements were not consistent with the van der Waals hyperbolic tangent profile given in Eq. (6), but were accurately described by the Fisk–Widom profile in Eq. (7) and the error-function profile in Eq. (12), with a slight preference for the error-function profile predicted for capillary waves on a step-function intrinsic profile. Values of ν determined by the fits by the error-function profile (e.g., $\nu = 0.62 \pm 0.01$) are also closer to the currently accepted value of $\nu = 0.63$ than those determined by the fits by the Fisk–Widom profile (e.g., $\nu = 0.663 \pm 0.02$) (Wu and Webb, 1973).

These early reflectivity measurements have been reanalyzed by several authors using hybrid models for the interface that combine the effects of capillary waves and the intrinsic profile as well as taking advantage of more recent measurements of bulk parameters, such as ν, that affect the analysis of the interfacial profile. Beysens and Robert (1987) determined that the reflectivity measurements on SF_6 and methanol-cyclohexane solutions could not be explained using only an intrinsic profile, but it was not clear whether a hybrid theory was required or if capillary-wave theory alone was adequate (with a step-function intrinsic profile) (Beysens and Robert, 1987). Sengers *et al.* (1991) used a hybrid theory that incorporated a q-dependent surface tension and determined that the predicted temperature dependence (near the critical point) of the profile was consistent with the measurements on the binary fluid, but not with those on SF_6. In addition, using their prediction for $q_{\max}$, they determined that the theory overestimates the measured reflectivity by 15–20% for the binary fluid.

Ellipsometry has been used to study the liquid–vapor interface and the liquid–liquid interface of many fluids, including a few measurements near the critical point (Beaglehole, 1986, 1987; Croxton, 1986; Schmidt, 1988, 1991; Miles and Schmidt, 1990; Schmidt and Moldover, 1993). Here, we discuss some recent measurements by Schmidt and co-workers on single-component and binary liquid mixtures near criticality.

The ellipticity of an interface is given by the ratio of the complex reflectivity coefficients, r_p/r_s, given by

$$\rho(\theta) = r_p(\theta)/r_s(\theta) = |r_p(\theta)/r_s(\theta)|e^{i\Delta(\theta)}, \qquad (24)$$

where the reflectivity coefficients for p- or s-polarized light relate the reflected electric field E_r to the incident electric field E_i as $E_i^{p,s} = r_{p,s}E_i^{p,s}$; θ is the angle of incidence (θ is measured from the interfacial normal), and Δ is a phase shift. The ellipticity is often measured at the Brewster angle θ_B, where $\Delta = \pi/2$. This results in a measurement of $\bar{\rho} = \mathrm{Im}[\rho(\theta)]_{\theta=\theta_B}$. In this notation, the intensity of the p- or s-polarized specular reflectivity is given by $R_{p,s}(\theta) = |r_{p,s}(\theta)|^2$.

Schmidt and co-workers interpret their measurements of $\bar{\rho}$ in terms of a combination of ellipticities due to the Fisk–Widom intrinsic interface and to a sharp interface roughened by capillary waves, $\bar{\rho} = \bar{\rho}_{FW} + \bar{\rho}_{cw}$. The ellipticity is expected to scale with temperature as

$$\bar{\rho} \approx \frac{\sqrt{2}\pi}{\lambda} \Delta n_0 \xi_0^- t^{\beta-\nu}(\chi_{FW} + \chi_{cw}), \qquad (25)$$

where Δn_0 is an amplitude in the expression for the index of refraction near the critical point given by $n_\pm = n_c \pm \Delta n_0 t^\beta/2$, ξ_0^- is given by the expression for the bulk correlation length near the critical point in the two-phase region by $\xi^- = \xi_0^- t^{-\nu}$, and χ_{FW} and χ_{cw} are constants determined, respectively, by the choice of intrinsic profile and by the choice of capillary-wave theory. Schmidt and co-workers chose to determine χ_{cw} from either the theory of Meunier ($\chi_{cw} \cong 1.43$) or that of Sengers and van Leeuwen ($\chi_{cw} \cong 1.39$) and used $\chi_{FW} \cong 2.28$ from the Fisk–Widom profile (Schmidt and Moldover, 1993). Equation (25) indicates that the quantity $\bar{\rho}\lambda/\Delta n_0 \xi^- \sqrt{2}\pi$ is universal and should yield a universal constant, $\chi = \chi_{FW} + \chi_{cw}$. This is expected to be true for both the liquid–vapor and the liquid–liquid interfaces near criticality because these two systems fall in the same universality class. For three pure liquid–vapor interfaces and five low-molecular-weight binary liquid–liquid interfaces measured over a range of approximately two decades in reduced temperature, Schmidt and co-workers measure $\chi \cong 3.0 \pm 0.2$ (Schmidt and Moldover, 1993). The temperature scaling agrees with Eq. (25), but the theoretical prediction is 20% too high (similar to the predictions for the optical reflectivity data previously discussed). For three high-molecular-weight polymer mixtures, Schmidt and co-workers measure a similar temperature scaling, but with a slightly lower χ (Miles and Schmidt, 1990).

Two subsequent theoretical analyses have produced values of χ in agreement with experiment. Blokhuis and Bedeaux (1990, 1992) extended Eq. (25) by adding the effects of a bending modulus, κq^4 with $\kappa = \mu k_B T$, where μ is predicted to be universal at criticality. The amplitude is adjusted to the value $\mu = 1.1$ resulting in a value for χ in agreement with the ellipsometry data. An alternative approach by Bonn and Wegdam (1992) utilizes a q^4 term in a q-dependent surface tension to obtain similar agreement with the experimental results. Therefore, the ellipsometry measurements provide indirect support for the necessity of a q^4 term, due to either a q-dependent surface tension or an interfacial stiffness.

4.3 Nonclassical or Nondielectric Fluid Interfaces

4.3.1 Quantum Liquids As a result of the quantum statistics of liquid helium, the nature of the liquid–vapor interface is different from that for classical liquids. Calculation of the capillary-wave spectrum requires replacing the classical thermal occupation numbers by those given by Bose–Einstein statistics. On the basis of a theory for the capillary-wave modes given by Atkins (1953), Cole (1970) calculated the broadening of the $T = 0$ intrinsic interface due to the capillary modes. Density-functional theories by Stringari and Treiner (1987) and others (see Lurio *et al.*, 1993, for references) predict an asymmetric form for the ^{4}He liquid–vapor interface in which the decay of the density into the vapor phase is more rapid than the decay into the liquid phase.

Measurements of the profile have been made using atom scattering, ellipsometry, and x-ray reflectivity. The atom scattering is most sensitive to the outer, vapor part of the profile, and determination of the full profile into the liquid is difficult with this technique. Ellipsometry has been used to measure the overall width of the profile, though details of its functional form were not obtained.

Fits to x-ray reflectivity measurements are consistent with an asymmetric hyperbolic secant form given by

$$\frac{d\langle\rho\rangle_0}{dz} = \frac{C}{\exp(az/\sigma_0) + \exp(-bz/\sigma_0)}, \tag{26a}$$

with

$$a = 1 + \exp(-\eta), b = 1 + \exp(\eta), \tag{26b}$$

where σ_0 is a measure of the interfacial width, η represents the asymmetry of the interface ($\eta = 0$ corresponds to a symmetric profile), and $d\langle\rho\rangle_0/dz$ is the *intrinsic* profile at $T = 0$. This $T = 0$ profile is convoluted with the capillary-mode spectrum to analyze the data. The fitted profile consistent with the x-ray reflectivity closely matches the theoretical prediction of Stringari and Treiner (1987). The best fit obtained is an asymmetric profile, though these measurements cannot conclusively exclude the possibility of a symmetric profile. The $T = 0$ profile has a 90–10% width of 0.76 nm and is broadened by capillary waves to 0.92 ± 0.1 nm at 1.13 K, for example. As shown by Lurio *et al.* (1993), these x-ray reflectivity data are consistent with the interfacial profile width determined by Osborne (1989) using ellipsometry and with the outer vapor portion of the profile determined by Nayak *et al.* (1983) using atom scattering.

4.3.2 Liquid-Metal Surfaces As we have seen, the profile at the liquid–vapor interface for a dielectric fluid changes monotonically from the bulk liquid density to the bulk density of the vapor. This is a consequence of the approximate density independence of the intermolecular potential. In a liquid metal the particles are electrons and ions, while the vapor above the surface contains atoms or molecules, resulting in an interparticle potential for the liquid metal that is strongly density dependent and varies significantly across the interface. It is reasonable to expect that liquid–vapor interfaces of dielectric and metallic fluids are very different (a review of the older literature is given by Rice *et al.*, 1986).

Early theoretical work by Evans and others combined an effective Hamiltonian for the inhomogeneous liquid metal with the assumption of a monotonic density profile and determined that the liquid–vapor transition region is very narrow, smaller than an atomic diameter (Evans, 1974; Kumaravadivel and Evans, 1975; Evans and Kumaravadivel, 1976; Evans and Hasegawa, 1981; Hasegawa and Watabe, 1982). Relaxing the assumption of monotonicity, Rice and coworkers used computer simulations to predict that the liquid–vapor interface of the metals sodium, cesium, magnesium, and mercury is stratified for approximately three atomic diameters into the bulk (D'Evelyn and Rice, 1983a,b; Rice *et al.*, 1986; Harris *et al.*, 1987a; Gomez and Rice, 1994). Heuristically, this stratification occurs because the strong electron-density dependence of the pseudopotential plays the role of a confining wall that creates atomic layering in a manner analogous to atomic layering at a liquid–solid interface.

Until recently, only indirect experimental evidence for this stratification was available. As examples, ellipsometry, visible-light reflectivity, dynamic visible-light scattering, surface viscosity, and surface-plasmon dispersion measurements at the mercury surface yielded anomalous results that could, however, be explained with a structured liquid–vapor interface (see Sluis and Rice, 1983, for references). In addition, measurements on several metals of the surface-tension variation with temperature yielded a positive $d\gamma/dT$ over a range in temperature, indicating a negative excess surface entropy (White, 1966, 1971) [see Eq. (15)]. This is consistent with local atomic ordering in the interface not present in the bulk.

If stratification occurs at the interface, a peak in the x-ray reflectivity should be observed at $Q_z \cong 2\pi/d \cong 20$–$25\ \mathrm{nm}^{-1}$, where d is the layer spacing (approximately the atomic diameter of the liquid metal). Early x-ray reflectivity measurements on laboratory sources were unable to attain these high values of Q_z and provided only indirect evidence of stratification (Lu and Rice, 1978; Bosio *et al.*, 1984; Barton *et al.*, 1986; Thomas *et al.*, 1987). Recent synchrotron x-ray reflectivity measurements have reached these high values of Q_z and confirmed the existence of the interface stratification at the surfaces of pure liquid gallium and mercury (Magnussen *et al.*, 1995; Regan *et al.*, 1995). In gallium, the measured layering decays with a $1/e$ length of 3–4 atomic diameters and is broadened slightly by capillary waves, in essential agreement with the general predictions from simulations of other liquid metal interfaces (simulations for the gallium interface have not been published). The interfacial width of the top layer with the vapor is approximately 0.087 nm, in agreement with values calculated solely from the surface tension and the resolution of the x-ray measurements. An additional contribution to the interfacial width from an intrinsic profile is not necessary. Similar synchrotron x-ray reflectivity measurements from the surface of mercury yield results in qualitative agreement with the predictions from simulation (Magnussen *et al.*, 1995).

Rice and co-workers (1986) have also predicted the behavior of the transverse structure function, i.e., the radial distribution function within the plane of the interface. In the point local-density approximation, the transverse structure function at a height z in the inhomogeneous interfacial region is given by the corresponding structure function in a homogeneous liquid at the same temperature and density. Because of the strong variation of density with height z, this approximation predicts a strong dependence of the transverse structure function in the inhomogeneous region of the interface on the position normal to the interface. An alternative approximation introduced by Fischer and Methfessel (1980) for hard-sphere liquids predicts the structure function in an inhomogeneous liquid to be that of the homogeneous liquid at the same temperature and at a density that corresponds to averaging over regions of one sphere diameter about a point in the inhomogeneous liquid. This leads to a greatly reduced dependence of the transverse structure function on position in the interface (Rice, 1987). This reduction is a consequence of the fact that the longitudinal structure at the interface is on the same length scale, the atomic diameter, as the averaging. Note that these considerations lead to the prediction that the liquid–vapor interface of dielectric liquids (which has inhomogeneities in the longitudinal structure on length scales of several atomic diameters) should exhibit a sizable variation in the transverse structure function. This prediction has not been subject to experimental verification (Flom *et al.*, 1992).

Recent measurements using x-ray grazing-incidence diffraction of the transverse structure function at the liquid–vapor gallium interface by Flom *et al.* (1992) are consistent with the prediction of the Fischer–Methfessel approximation and indicate that there is no measurable difference between the transverse structure function in the interface and the bulk structure function. These measurements are limited by the weak sensitivity to the topmost layer because the x rays probe a depth of ~5 nm, thereby averaging the structure function over this region. Unambiguous measurements of the transverse structure function at the interfaces of metallic and nonmetallic liquids await further experimental investigation.

In the case of liquid-metal alloys for which one of the components preferentially segregates at the interface, the pseudopotential is dependent upon both density and composition. As expected from the Gibbs adsorption equation [Eq. (15)], simulations on a Cs:Na alloy indicate that the lower-tension component, Cs, is preferentially adsorbed at the interface. In contrast to adsorption at a dielectric-mixture liquid–vapor interface for which the excess surface concentration is expected to decay monotonically over several atomic distances, the Cs formed essentially a single layer at the interface with subsequent modulations into the bulk alloy fluid similar to those expected for a pure metal (Harris *et al.*, 1987b). This view of the liquid-metal interface is supported by recent x-ray reflection and grazing-incidence diffraction experiments on a Bi-doped (~0.2%) Ga sample for which the Bi forms a segregated almost complete monolayer at the interface. The liquid-type transverse structure function of the Bi is similar to that expected for Bi supercooled by 271 °C (the transverse structure function was measured at 36 °C and the freezing temperature of Bi is 307 °C!) (Flom *et al.*, 1993; Lei *et al.*, 1996).

4.4 Complex-Fluid Interfaces

Complex fluids often exhibit more complex interfacial phenomena including layering, adsorption, and wetting, to name a few. In this section interfacial structure in surfactant solutions, long-chain liquids such as alkanes, and liquid crystals is used to illustrate the variety of phenomena studied. In this discussion, we concentrate on experimental results from x-ray or neutron scattering that directly probe these interfacial structures.

4.4.1 Microemulsions Although water and oil do not mix to an appreciable extent, the addition of a small amount of amphiphilic surfactant allows the three components to form a single phase. The mechanism for mixing is usually the formation of internal interfaces that partition the water and oil microscopically. Examples of these internal structures are micelles that consist of droplets of oil coated by surfactants and surrounded by water (this phase is denoted a droplet microemulsion) or large random sheets of surfactant that separate microscopic regions of oil and water (denoted a bicontinuous microemulsion). On variation of the relative composition and the temperature, these ternary mixtures often form three phases consisting of a middle-phase microemulsion sandwiched (by gravity) between the heavier water-rich phase and the lighter oil-rich phase. In the single-phase microemulsion the microemulsion-vapor (or -solid) interface can be studied; in the three-phase system the microemulsion-oil (or -water) interface can be studied.

In the three-phase system, wetting of the oil–water interface by the microemulsion can be discussed in terms of an effective interface potential of the oil-microemulsion interface with the microemulsion-water interface. For short-range forces, the effective interaction between these interfaces a distance d apart $V(d) \rightarrow \cos(kd + \phi)\exp(-d/\xi)$, in the asymptotic limit when d is large (ϕ is an unimportant phase) (Gompper and Schick, 1990a). The effective interface potential exhibits oscillations similar to those in the bulk phase, which represent locally structured microscopic regions of surfactant sheets separating oil and water. The exponential decay of this interaction mirrors the short-range character of the bulk structure.

Importantly, the interaction potential is influenced by the structure at the interfaces. Lattice models utilizing short-range forces predict the existence of nonmonotonic concentration profiles normal to the oil-microemulsion and microemulsion-water interfaces. This structure mimics the structure in the bulk microemulsion and is predicted, for short-range forces, to vary as $f(z) \sim \sin(kz) \times \exp(-z/\xi)$, where k and ξ are expected to take the same values as the bulk parameters (Dawson, 1987; Gompper and Schick, 1989, 1990a, b).

An important prediction of the lattice models is the nonmonotonic interfacial profile, which plays a role in determining the wetting properties of the microemulsion. Historically, Meunier (1985) used ellipsometry to study the structure at the oil-microemulsion and microemulsion-water interfaces of a five-component mixture. He concluded that "On the microemulsion side of the film we find a perturbed zone, the thickness of which is comparable with the size of the structure in the bulk."

Zhou *et al.* (1992) used neutron reflectivity to study the microemulsion-*vapor* interface of mixtures of the surfactant tetraethylene glycol monodecyl ether ($C_{10}E_4$), D_2O, and octane in the one-phase microemulsion region of the phase diagram. These measurements are consistent with the form of the interfacial profile given above, $f(z) \sim \sin(kz)$

$\times \exp(-z/\xi)$, with the surface and bulk values of k and ξ in agreement.

D. D. Lee *et al.* (1995) have made similar neutron-reflectivity measurements at the microemulsion–*solid silicon* interface for the same one-phase microemulsion system in the same region of the phase diagram. They measured the profile to be of the form given above; however, the surface and bulk values of k and ξ were not in agreement. Measurements as a function of temperature indicated that the surface values, k_s and ξ_s, varied in a systematic manner, while the bulk values were essentially constant. The authors hypothesize that these effects are due to the influence of the spontaneous curvature of the amphiphilic layers on the surface profile.

The microemulsion-oil (or -water) interfaces have an ultralow surface tension, measured to be as small as 1/10 000 of the pure water–oil interfacial tension (Fleming and Vinatieri, 1981; Chatenay *et al.*, 1985; Lee *et al.*, 1990). This low tension can be understood in terms of an unbinding transition of the near-surface lamellar structure (Gompper and Schick, 1994).

Because of the small interfacial tension, it is expected that the oil–microemulsion and microemulsion–water interfaces will be rough with large-amplitude thermal fluctuations. A similar expectation holds for a (out-of-equilibrium) surfactant-monolayer–covered oil–water interface as recently investigated. L. T. Lee *et al.* (1991) used a thin (~300 nm) layer of oil on top of a surfactant-monolayer–covered subphase of water. Neutron reflectivity was used to probe an octane–($C_{10}E_4$ monolayer)–water interface. Analyzing the overall drop in reflected intensity as in Eq. (20), they showed that their neutron reflectivity data are consistent with an interfacial roughness of 9 nm, as determined from their previous interfacial tension measurements using surface light scattering (L. T. Lee *et al.*, 1990). This roughness amplitude is much larger than the molecular lengths of the three components and is consistent with the presence of thermally excited surface waves.

McClain *et al.* (1994) used x-ray reflectivity and x-ray surface diffuse scattering to study a triethylene glycol monooctyl ether (C_8E_3) monolayer at a decane–water interface. In this case, the monolayer was formed by removing most of the nonwetting middle-phase microemulsion from a three-phase decane, water, and C_8E_3 solution, leaving a small amount of microemulsion in equilibrium with a C_8E_3 monolayer covering most of the interface. The use of diffuse scattering and reflectivity allowed for the surface tension to be determined independently from the interfacial width. Capillary-wave theory with a step-function intrinsic profile was used to fit their data and revealed the presence of large surface fluctuations (8.5-nm interfacial width).

Within the context of Ginzburg–Landau theories it was expected that the local interfacial density profile of the (oil-rich)–(water-rich) interface could be approximated by the mean-field profile for an (oil-rich)–(water-rich) interface covered by a surfactant monolayer (Gompper and Kraus, 1993). However, the microemulsion–water and oil–microemulsion interfaces are expected to be more complicated, with a nonmonotonic interfacial profile that will be strongly distorted by the interfacial curvature (Gompper and Kraus, 1993). These predictions have yet to be confirmed experimentally.

The first x-ray scattering (reflectivity and surface diffuse scattering) measurements from an equilibrium microemulsion–oil interface were recently reported on a 2-butoxyethanol (C_4E_1), water, and decane system that has very weak bulk structure and, therefore, is not expected to have nonmonotonic interfacial profiles (Williams *et al.*, 1996). The measured interfacial roughness and interfacial tension are respectively $\sigma = 5.7$ nm [Eq. (20)] and $\gamma = 0.14$ mN/m. This interfacial roughness is much larger than the largest molecule (~0.9 nm). No evidence for a nonmonotonic profile was reported; however, the range in Q_z was not large enough to detect it, if present.

Related structural measurements at the vapor interface of droplet microemulsions have been made (see Langevin, 1992, for a review of the optical measurements). Here we mention an x-ray reflectivity measurement by Schwartz *et al.* (1988) on the AOT (sodium di-2-ethylsulfosuccinate), D_2O, and decane system that has spherical micelles in the bulk (Kotlarchyk *et al.*, 1984; Huang and Kotlarchyk, 1986). Considering the planar symmetry of the interface, two possibilities are likely for the surface structure of this system:

1. one or more surfactant layers at the interface followed by a transition zone leading into the bulk fcc ordering of micelles or
2. extension of the fcc-ordered micelles directly up to the interface.

Schwartz *et al.* showed that the latter is true. The structure consists of two well-defined layers mimicking the surface of a fcc crystal, then decaying to the more random orientation of the bulk.

4.4.2 Alkanes and Alcohols The surface tension of most materials decreases with increasing temperature, indicating a positive excess surface entropy [see Eq. (15)]. As discussed, some liquid metals have $d\gamma/dT > 0$, which indicates an additional ordering at the surface. Earnshaw and Hughes (1992) measured $d\gamma/dT$ for temperatures slightly above the melting points of the *n*-alkanes, C15–C18 [where C_n refers to $CH_3(CH_2)_{n-2}$-CH_3]. For all of these alkanes, $d\gamma/dT > 0$ immediately above the bulk melting transition with an abrupt change to $d\gamma/dT < 0$ approximately 2°–3° above the melting transition. Subsequently, Wu and co-workers (1993a,b,c) used x-ray reflectivity and in-plane grazing-incidence diffraction to demonstrate that the first-order transition occurring a few degrees above the bulk melting transition is a surface freezing transition in which a single layer at the surface orders. Prior to these measurements, surface ordering or freezing had been discovered in liquid crystals (discussed in Sec. 4.4.3). However, for most materials the tendency is for surface melting because of the excess surface entropy at a solid–vapor interface.

Wu *et al.* (1993a,b,c) demonstrated that the surface freezing occurs only for chain lengths of 14–50 carbon atoms. The x-ray scattering from the surface is well explained by the top, single layer of molecules being orientationally ordered normal to the interface and positionally ordered in a hexagonal lattice within the plane of the interface with correlation lengths exceeding several hundred nanometers. For longer chain lengths, $n > 30$, the molecules in the top layer are tilted toward their nearest neighbors. This top layer has the density and x-ray signature of a rotator phase found in bulk alkanes for which the molecules are free to rotate about the long molecular axis. Only the top layer is ordered until the temperature is lowered to the bulk melting temperature at which the entire sample freezes. Deutsch *et al.* (1995) have seen similar freezing of the top layer of *n*-alcohol liquids for even carbon numbers between 18 and 26. The surfaces of binary mixtures of *n*-alkanes exhibit an additional surface phase, a monolayer of crystalline alkane, if the difference in chain lengths, Δn, is large enough (Wu *et al.*, 1995).

Currently there is not a consensus on the theoretical explanation for this surface freezing. Leermakers and Cohen-Stuart (1996) used a lattice model to demonstrate orientational surface ordering due to a hypothesized greater surface affinity of the CH_3 moieties. Tkachenko and Rabin (1996) propose that the free energy of a top solid layer is reduced by strong fluctuations along the molecular axis while the layer still remains a solid. This free-energy reduction is not available to molecules in the bulk since these fluctuations will be suppressed by neighboring molecules. This mechanism allows the top layer to solidify while the bulk remains a liquid. They successfully fit two sets of data from the experiments of Wu *et al.*:

1. the shift of the surface freezing temperature from the bulk freezing temperature as a function of the chain length *n* and
2. the correction to the liquid-surface tension due to the existence of the ordered top layer as a function of chain length.

As pointed out by Tkachenko and Rabin, this mechanism for surface freezing is only appropriate for materials whose ordered bulk phases have a high degree of three-dimensional ordering with out-of-plane fluctuations suppressed. For that reason it is not expected to apply to smectic liquid crystals for which strong out-of-plane fluctuations in the bulk phase are suppressed at the surface. As discussed in the next section, surface ordering in liquid crystals takes a different form.

4.4.3 Liquid Crystals Liquid crystals exhibit complex ordering phenomena in the bulk that has been extensively reviewed (Litster and Birgeneau, 1982; Pershan, 1988; Brock *et al.*, 1989; de Gennes and Prost, 1993). Early measurements by Miyano (1979a,b) and Beaglehole (1982) showed that the molecular order near the surface differs from that in the bulk. Theoretical work by Gubbins (1986) demonstrated that small, nonspherical molecules preferentially orient at surfaces, although these effects for small molecules are much harder to detect experimentally. Extensive theoretical work has demonstrated a variety of surface transitions for the longer molecules typically considered to be liquid crystals (for an older review, see Sluckin and Poniewierski, 1986; more recent theoretical work is discussed in Somoza *et al.*, 1995, and del Río *et al.*, 1995). The surfaces of liquid crystals exhibit wetting transitions, critical adsorption, and orientational transitions. Here, we discuss representative measurements of a few of these phenomena.

Early measurements of complex surface phenomena that directly probed structure on the molecular length scale were made by Pershan, Als-Nielsen, and co-workers on the liquid–vapor interface of liquid crystals. X-ray and neutron reflection has been used to study the surfaces of thermotropic and lyotropic crystals (Pershan, 1989; Penfold and Thomas, 1990). Synchrotron x-ray sources have much higher intensity than neutron sources. This allows for larger wave-vector transfer for x-ray scattering and, consequently, higher spatial resolution. However, by isotopic substitution the neutron scattering-length density of different regions of the interface can be enhanced. This is particularly useful in probing the structure of water molecules near ordered alkane chains of surfactant tail groups for which the x-ray scattering contrast is weak. The combined use of these techniques promises to be very powerful.

Although several different liquid crystals have been used as model systems to study surface effects, this discussion will be restricted primarily to studies of the isomorphic series of cyanobiphenyl molecules referred to as *n*CB. The *n*CB molecule has two adjacent phenyl groups providing a rigid core attached on one side to a cyano group and on the other to an alkyl chain with n carbon atoms. Integral values of n refer to the pure compound while fractional values refer to mixtures of *n*CB and $(n + 1)$CB. In the bulk material the phase transitions of interest here are as follows (these transitions occur in the range of 30 °C $< T <$ 60 °C for $8 < n < 12$): For $n \gtrsim 9.55$ there is a first-order transition from the isotropic to the smectic-A phase as the temperature is lowered; at $n \cong 9.55$ the isotropic–nematic–smectic-A triple point occurs; for $9 < n \lesssim 9.55$, first-order transitions from the isotropic to the nematic and from the nematic to the smectic A occur as the temperature is lowered, while for $n < 9$ the same sequence occurs, except that the nematic–smectic-A transition is second order.

The general tendency of *n*CB and other liquid-crystal systems is to form layers of smectic phase at the liquid–vapor interface that gradually decay into the bulk (Pershan, 1989). This is in contrast to the single layer at alkane surfaces previously discussed. Similar phenomena are also seen at a liquid-crystal interface with a solid wall (Ocko, 1990; Kellogg *et al.*, 1995). For example, as the reduced temperature $t = (T - T_{IA})/T_{IA}$, measuring the deviation from the first-order bulk isotropic to smectic-A transition, is reduced from $t = 0.1$ to $t = 10^{-5}$, the number of smectic layers at the liquid–vapor interface of the isotropic phase of 12CB increases from 1 to ~6 or 7. This increase in the number of adsorbed smectic layers occurs at well-defined temperatures in steps of one layer. Similar phenomena are seen for 9.55CB, 10CB, and 11CB as well as for a related isomorphic series 12OCB, 14OCB, and 16OCB (for *n*OCB the alkyl chain in *n*CB is replaced by an alkoxy chain). Pershan and co-workers argue that these transitions are most likely incomplete wetting transitions with the saturation at two layers for 16OCB unambiguously indicating incomplete wetting. For the shorter-chain samples (9.55CB, 10CB, and 11CB) that are closer to the triple point, the transition from the surface layering to the bulk becomes progressively more continuous in temperature. Pershan and Braslau argue that at any particular temperature the number of surface layers is well defined. These data suggest the possibility of a roughening transition as the triple point is approached.

When approaching a second-order nematic to smectic-A transition, these systems exhibit critical adsorption of smectic-A layers at the free surface of the nematic phase. In this case the surface layering extends exponentially into the bulk with a decay length equal to $\xi_{\parallel}$, the bulk decay length for correlations normal to the smectic layers. At the bulk first-order nematic to smectic-A transition, a pretransition growth of smectic correlations in the nematic phase occurs. This growth is cut off at the length scale $\xi_{\parallel}(T_{\mathrm{NA}})$ because of the first-order transition. Measurements on 9.15CB and 9.3CB show that the surface correlation length $\xi(T)$ is slightly larger than $\xi_{\parallel}(T)$ with a steadily increasing, but not diverging, ratio $\xi(T)/\xi_{\parallel}(T)$ as $T \rightarrow T_{\mathrm{NA}}$. This does not appear to be a true wetting transition.

5. SOME OTHER SURFACE PHENOMENA

Here, we briefly mention a few areas of importance that space has not permitted us to discuss more fully.

The in-plane or transverse structure factor at the surface of liquid crystals is most easily studied on freely suspended liquid crystal films (similar to soap films). The film thickness can be varied from two molecular layers to microns. These films have a large surface-to-volume ratio, allowing for the study of phase transitions in two and three dimensions (including the dimensional crossover). Transmission optical and x-ray techniques have been used to study wetting and adsorption phenomena, including surface freezing and layering transitions (Swanson *et al.*, 1989), molecular alignment surface transitions, and thin-film and surface hexatic phases (including fundamental studies of melting; see Brock *et al.*, 1986).

Langmuir monolayers of insoluble surfactants at the liquid (generally water)–vapor interface have been studied for many years (Gaines, 1966; Ulman, 1991). Many theoretical and experimental techniques have been used to study these films. Early equation-of-state measurements established some of the basic concepts concerning molecular orientation at interfaces and the (typically) short-range nature of intermolecular interactions, and deduced the cross-sectional area of long-chain fatty-acid molecules (Harkins, 1917; Langmuir, 1917). More recent measurements using nonlinear optical techniques and x-ray and neutron scattering have been able to probe directly the longitudinal and transverse density distributions of these monolayers (Shen, 1989; Knobler, 1990; Penfold and Thomas, 1990; Als-Nielsen *et al.*, 1994). An extensive variety of monolayer solid phases have been recently discovered, including the existence of two-dimensional solid islands at very low densities where it was formerly believed that two-dimensional gas and liquid phases coexisted. These direct probes also allow for molecular-scale studies of the kinetics of phase transformations that are important for determining the equilibrium equation of state for these materials (Bommarito *et al.*, 1996).

Very thin liquid films bounded by one or more solid walls or filling porous solid materials have also been studied (Gelfand and Fisher, 1990; Israelachvili, 1992). Although the solid–liquid interface is discussed elsewhere in this Encyclopedia, there is one result of particular interest for liquid interfaces. The structure of a liquid surface of a very thin (≤10 nm) liquid film on a rough solid surface is very different from the bulk surface structure. For thin enough films, the roughness of the liquid–vapor interface is correlated with the underlying rough liquid–solid interface (Garoff *et al.*, 1989; Tidswell *et al.*, 1991). Andelman and co-workers explained this correlation in terms of the nonretarded van der Waals forces important at these small distances (Andelman *et al.*, 1988; Robbins *et al.*, 1991).

The thermal energy responsible for capillary waves is also responsible for a continual exchange of fluid and vapor molecules across the interface. This type of surface process is discussed in the article SURFACE PROCESSES. Recent experiments that directly explore the atomic exchange at liquid surfaces include atom scattering from a liquid surface (King *et al.*, 1994).

Extensive studies of liquid surfaces and a few recent studies at the liquid–liquid interface have been made with nonlinear optical techniques that are surface sensitive. Second-harmonic generation is sensitive to the nonlinear polarizability of the molecule and determines molecular orientations at fluid interfaces (Shen, 1989; Eisenthal, 1992; Naujok *et al.*, 1995). Sum-frequency generation can provide information on alkane chain disorder in interfacial molecules by distinguishing between *trans* and *gauche* conformations in the chain (Guyot-Sionnest *et al.*, 1987; Messmer *et al.*, 1995).

In addition to the computer simulations of liquid-metal surfaces previously mentioned, numerous simulations of aqueous interfaces have been made (recently reviewed by Pohorille and Wilson, 1993). These simulations have been used to study, for example, the orientation of water molecules at the pure water surface, the water structure in hydrated ions and adjacent to surfactants at interfaces, the interfacial density profiles at liquid–liquid interfaces (such as water–oil interfaces), the structure of amphiphilic monolayers and bilayers at interfaces, and the surface tension and surface potential. Many of these predictions can be tested experimentally. For example, the prediction for a water molecule at the water surface indicates that its dipole moment is in the plane of the surface with one hydrogen pointing toward the vapor and one toward the liquid. The orientational distribution is broad and slightly asymmetric, resulting in an excess net dipole moment pointing into the liquid. This prediction is consistent with second-harmonic generation studies of the interface by Eisenthal (1992).

Although this article has focused on the structure at interfaces between bulk phases, complex fluids often contain internal interfaces that separate microscopic regions. These materials can be understood in terms of the interactions between these fluctuating, internal interfaces, a viewpoint very different from classical solution chemistry. This field is reviewed by Gompper and Schick (1994) and by Safran (1994).

GLOSSARY

Adsorption: The surface excess density.

Capillary Waves: Interfacial waves whose return force is determined by surface tension. Often combined with gravity waves whose return force is determined by gravity.

Criticality: At a critical point characterized by the divergence of pair correlation functions.

Interface: The transition region between any two bulk phases in contact.

Interfacial Profile (Longitudinal Density Profile): Electron, molecular, or number density profile through and normal to the interface between two bulk phases.

Interfacial or Surface Tension: The interfacial (or surface) tension is given by the free energy required to take molecules from the bulk of the liquid and create a unit area of interface (or surface).

Surface: The liquid–vapor interface.

Wetting: The intervening of a third equilibrium phase between two other equilibrium phases.

Works Cited

Abraham, D. B., Reed, P. (1974), *Phys. Rev. Lett.* **33,** 377–379.

Abraham, D. B., Reed, P. (1976), *Comm. Math. Phys.* **49,** 35–46.

Adamson, A. W. (1990), *Physical Chemistry of Surfaces,* New York: Wiley.

Als-Nielsen, J., Jacquemain, D., Kjaer, K., Leveiller, F., Lahav, M., Leiserowitz, L. (1994), *Phys. Rep.* **246,** 251–313.

Andelman, D., Joanny, J. F., Robbins, M. O. (1988), *Europhys. Lett.* **7,** 731–736.

Atkins, K. R. (1953), *Can. J. Phys.* **31,** 1165–1169.

Barber, D. J., Loudon, R. (1989), *An Introduction to the Properties of Condensed Matter,* Cambridge, U.K.: Cambridge Univ. Press.

Barton, S. W., Thomas, B. N., Novak, F., Harris, J., Dolmer, P., Bloch, J. M., Rice, S. A. (1986), *Nature* **321,** 685–687.

Beaglehole, D. (1980), *J. Chem. Phys.* **73,** 3366–3371.

Beaglehole, D. (1982), *Mol. Cryst. Liq. Cryst.* **89,** 319–325.

Beaglehole, D. (1986), in: C. A. Croxton (Ed.), *Fluid Interfacial Phenomena,* Chichester: Wiley, p. 523.

Beaglehole, D. (1987), *Phys. Rev. Lett.* **58,** 1434–1436.

Becker, R. S., Golovchenko, J. A., Patel, J. R. (1983), *Phys. Rev. Lett.* **50,** 153–156.

Beysens, D., Leibler, S. (1982), *J. Phys. Lett. (Paris)* **43,** L133–L136.

Beysens, D., Robert, M. (1987), *J. Chem. Phys.* **87,** 3056–3061.

Blokhuis, E. M., Bedeaux, D. (1990), *Phys. A* **164,** 515–548.

Blokhuis, E. M., Bedeaux, D. (1992), *Phys. A* **184,** 42–70.

Bommarito, G. M., Foster, W. J., Pershan, P. S., Schlossman, M. L. (1996), *J. Chem. Phys.* **105,** 5265–5284.

Bonn, D., Wegdam, G. H. (1992), *J. Phys. I France* **2,** 1755–1764.

Bonnet, J. C., Pike, F. P. (1972), *J. Chem. Eng. Data* **17,** 145–150.

Bosio, L., Cortes, R., Defrain, A., Oumezine, M. (1984), *J. Non-Cryst. Solids* **61/62,** 697–702.

Braslau, A., Deutsch, M., Pershan, P. S., Weiss, A. H., Als-Nielsen, J., Bohr, J. (1985), *Phys. Rev. Lett.* **54,** 114–117.

Braslau, A., Pershan, P. S., Swislow, G., Ocko,

B. M., Als-Nielsen, J. (1988), *Phys. Rev. A* **38,** 2457–2470.

Brock, J. B., Birgeneau, R. J., Litster, J. D., Aharony, A. (1989), *Phys. Today* **42** (7), 52–59.

Brock, J. D., Aharony, A., Birgeneau, R. J., Evans-Lutterodt, K. W., Litster, J. D., Horn, P. M., Stephenson, G. B., Tajbakhsh, A. R. (1986), *Phys. Rev. Lett.* **57,** 98–101.

Buff, F. P., Lovett, R. A., Stillinger, F. H. (1965), *Phys. Rev. Lett.* **15,** 621–623.

Cahn, J. W. (1977), *J. Chem. Phys.* **66,** 3667–3672.

Cahn, J. W., Hilliard, J. E. (1958), *J. Chem. Phys.* **28,** 258–267.

Chatenay, D., Abillon, O., Meunier, J., Langevin, D., Cazabat, A. M. (1985), in: D. O. Shah (Ed.), *Macro- and Microemulsions,* Washington, D.C.: ACS, p. 119.

Cole, M. W. (1970), *Phys. Rev. A* **1,** 1838–1840.

Cole, M. W. (1995), *J. Low Temp. Phys.* **101,** 25–40.

Croxton, C. A. (Ed.) (1986), *Fluid Interfacial Phenomena,* Chichester: Wiley.

D'Evelyn, M. P., Rice, S. A. (1983a), *J. Chem. Phys.* **78,** 5081–5095.

D'Evelyn, M. P., Rice, S. A. (1983b), *J. Chem. Phys.* **78,** 5225–5249.

Dawson, K. A. (1987), *Phys. Rev. A* **35,** 1766–1773.

de Gennes, P. G., Prost, J. (1993), *The Physics of Liquid Crystals,* Oxford: Clarendon Press.

Defay, R., Prigogine, I., Bellemans, A., Everett, D. H. (1966), *Surface Tension and Adsorption,* London: Longmans, Green & Co. Ltd.

del Río, E. M., da Gama, M. M. T., Miquel, E. d., Rull, L. F. (1995), *Phys. Rev. E* **52,** 5028–5039.

Desai, N. S., Peach, S., Franck, C. (1995), *Phys. Rev. E* **52,** 4129–4133.

Deutsch, M., Wu, X. Z., Sirota, E. B., Sinha, S. K., Ocko, B. M., Magnussen, O. M. (1995), *Europhys. Lett.* **30,** 283–288.

Dietrich, S. (1988), in: C. Domb, J. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena,* London: Academic Press, Vol. 12, p. 2.

Dietrich, S., Haase, A. (1995), *Phys. Rep.* **260,** 1–138.

Dietrich, S., Wagner, H. (1984), *Z. Phys. B* **56,** 207–215.

Dietrich, S., Wagner, H. (1985), *Z. Phys. B* **59,** 35–42.

Dosch, H. (1987), *Phys. Rev. B* **35,** 2137–2143.

Earnshaw, J. C., Hughes, C. J. (1992), *Phys. Rev. A* **46,** R4494–R4496.

Eisenthal, K. B. (1992), *Annu. Rev. Phys. Chem.* **43,** 627–661.

Evans, R. (1974), *J. Phys. C* **7,** 2808–2830.

Evans, R., Hasegawa, M. (1981), *J. Phys. C* **14,** 5225–5246.

Evans, R., Kumaravadivel, R. (1976), *J. Phys. C* **9,** 1891–1906.

Fischer, J., Methfessel, M. (1980), *Phys. Rev. A* **22,** 2836–2843.

Fisher, M. E., de Gennes, P. G. (1978), *C. R. Acad. Sci. Paris Ser. B* **287,** 207–209.

Fisk, S., Widom, B. (1969), *J. Chem. Phys.* **50,** 3219–3227.

Fleming, P. D., Vinatieri, J. E. (1981), *J. Coll. Int. Sci.* **81,** 319–331.

Flom, E. B., Cai, Z.-h., Acero, A., Lin, B., Maskil, N., Liu, L., Rice, S. A. (1992), *J. Chem. Phys.* **96,** 4743–4749.

Flom, E. B., Li, M., Acero, A., Maskil, N., Rice, S. A. (1993), *Science* **260,** 332–335.

Franck, C., Schnatterly, S. E. (1982), *Phys. Rev. Lett.* **48,** 763–766.

Gaines, G. L. (1966), *Insoluble Monolayers at a Liquid Gas Interface,* New York: Wiley.

Garoff, S., Sirota, E. B., Sinha, S. K., Stanley, H. B. (1989), *J. Chem. Phys.* **90,** 7505–7515.

Gelfand, M. P., Fisher, M. E. (1990), *Physica A* **166,** 1–74.

Gibbs, J. W. (1957), *The Collected Works of J. Willard Gibbs,* New Haven: Yale Univ. Press.

Gomez, A., Rice, S. A. (1994), *J. Chem. Phys.* **101,** 8094–8097.

Gompper, G., Kraus, M. (1993), *Phys. Rev. E* **47,** 4289–4300.

Gompper, G., Schick, M. (1989), *Phys. Rev. Lett.* **62,** 1647–1650.

Gompper, G., Schick, M. (1990a), *Phys. Rev. Lett.* **65,** 1116–1119.

Gompper, G., Schick, M. (1990b), *Phys. Rev. B* **41,** 9148–9162.

Gompper, G., Schick, M. (1994), in: C. Domb, J. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena,* Vol. 16, London: Academic Press.

Gubbins, K. E. (1986), in: C. A. Croxton (Ed.), *Fluid Interfacial Phenomena,* Chichester: Wiley, p. 469.

Guyot-Sionnest, P., Hunt, J. H., Shen, Y. R. (1987), *Phys. Rev. Lett.* **59,** 1597–1600.

Hahne, F. J. W. (Ed.) (1983), *Critical Phenomena,* Berlin: Springer-Verlag.

Harkins, W. D. (1917), *J. Am. Chem. Soc.* **39,** 541–596.

Harris, J. G., Gryko, J., Rice, S. A. (1987a), *J. Chem. Phys.* **87,** 3069–3081.

Harris, J. G., Gryko, J., Rice, S. A. (1987b), *J. Stat. Phys.* **48,** 1109–1128.

Hasegawa, M., Watabe, M. (1982), *J. Phys. C* **15,** 353–375.

Huang, J. S., Kotlarchyk, M. (1986), *Phys. Rev. Lett.* **57,** 2587–2589.

Huang, J. S., Webb, W. W. (1969), *J. Chem. Phys.* **50,** 3677–3693.

Israelachvili, J. N. (1992), *Intermolecular and Surface Forces,* London: Academic Press.

Jasper, J. J. (1972), *J. Phys. Chem. Ref. Data* **1,** 841–1009.

Kellogg, G. J., Pershan, P. S., Kawamoto, E. H., Foster, W. F., Deutsch, M., Ocko, B. M. (1995), *Phys. Rev. E* **51,** 4709–4726.

King, M. E., Saecker, M. E., Nathanson, G. M. (1994), *J. Chem. Phys.* **101,** 2539–2547.

Knobler, C. M. (1990), *Adv. Chem. Phys.* **77,** 397–449.

Kotlarchyk, M., Chen, S. H., Huang, J. S., Kim, M. W. (1984), *Phys. Rev. Lett.* **53,** 941–944.

Kumaravadivel, R., Evans, R. (1975), *J. Phys. C* **8,** 793–808.

Landau, L. D., Lifshitz, E. M. (1986), *Fluid Dynamics,* Oxford, England: Pergamon Press.

Langevin, D. (1992), *Light Scattering by Liquid Surfaces and Complementary Techniques,* New York: Marcel Dekker Inc.

Langmuir, I. (1917), *J. Am. Chem. Soc.* **39,** 1848–1906.

Lee, D. D., Chen, S. H., Majkrzak, C. F., Satija, S. K. (1995), *Phys. Rev. E* **52,** R29–R32.

Lee, L. T., Langevin, D., Meunier, J., Wong, K., Cabane, B. (1990), *Prog. Coll. Poly. Sci.* **81,** 209–214.

Lee, L. T., Langevin, D., Farnoux, B. (1991), *Phys. Rev. Lett.* **67,** 2678–2681.

Leermakers, F. A. M., Cohen-Stuart, M. A. (1996), *Phys. Rev. Lett.* **76,** 82–85.

Lei, N., Huang, Z., Rice, S. A. (1996), *J. Chem. Phys.* **104,** 4802–4805.

Litster, J. D., Birgeneau, R. J. (1982), *Phys. Today* **35** (5), 26–33.

Liu, A. J., Fisher, M. E. (1989), *Phys. Rev. A* **40,** 7202–7221.

Loudon, R. (1984), in: V. M. Agranovich, R. Loudon (Eds.), *Surface Excitations,* New York: Elsevier, p. 589.

Lu, B. C., Rice, S. A. (1978), *J. Chem. Phys.* **68,** 5558–5567.

Lurio, L. B., Rabedeau, T. A., Pershan, P. S., Silvera, I. F., Deutsch, M., Kosowsky, S. D., Ocko, B. M. (1993), *Phys. Rev. B* **48,** 9644–9659.

Magnussen, O. M., Ocko, B. M., Regan, M. J., Penanen, K., Pershan, P. S., Deutsch, M. (1995), *Phys. Rev. Lett.* **74,** 4444–4447.

McClain, B. R., Lee, D. D., Carvalho, B. L., Mochrie, S. G. J., Chen, S. H., Litster, J. D. (1994), *Phys. Rev. Lett.* **72,** 246–249.

Messmer, M. C., Conboy, J. C., Richmond, G. L. (1995), *J. Am. Chem. Soc.* **117,** 8039–8040.

Meunier, J. (1985), *J. Phys. Lett. (France)* **46,** L1005–L1014.

Meunier, J. (1987), *J. Phys. (Paris)* **48,** 1819–1831.

Miles, D. G., Schmidt, J. W. (1990), *J. Chem. Phys.* **92,** 3881–3885.

Miyano, K. (1979a), *J. Chem. Phys.* **71,** 4108–4111.

Miyano, K. (1979b), *Phys. Rev. Lett.* **43,** 51–54.

Napiórkowski, M., Dietrich, S. (1993), *Phys. Rev. E* **47,** 1836–1849.

Naujok, R. R., Higgins, D. A., Hanken, D. G., Corn, R. M. (1995), *J. Chem. Soc. Faraday Trans.* **91,** 1411–1420.

Nayak, V. U., Edwards, D. O., Masuhara, N. (1983), *Phys. Rev. Lett.* **50,** 990–992.

Ocko, B. M. (1990), *Phys. Rev. Lett.* **64,** 2160–2163.

Ocko, B. M., Wu, X. Z., Sirota, E. B., Sinha, S. K., Deutsch, M. (1994), *Phys. Rev. Lett.* **72,** 242–245.

Osborne, D. V. (1989), *J. Phys. Cond. Mat.* **1,** 289–297.

Pandit, R., Schick, M., Wortis, M. (1982), *Phys. Rev. B* **26,** 5112–5140.

Penfold, J., Thomas, R. K. (1990), *J. Phys. Cond. Mat.* **2,** 1369–1412.

Pershan, P. S. (1988), *Structure of Liquid Crystal Phases,* Singapore: World Scientific.

Pershan, P. S. (1989), *J. Phys. (Paris) Colloq. C-7* **50,** C7-1–C7-19.

Pershan, P. S. (1990), *Far. Disc. Chem. Soc.* **89,** 231–245.

Pohorille, A., Wilson, M. A. (1993), *J. Mol. Struct.* **284,** 271–298.

Regan, M. J., Kawamoto, E. H., Lee, S., Pershan, P. S., Maskil, N., Deutsch, M., Magnussen, O. M., Ocko, B. M. (1995), *Phys. Rev. Lett.* **75,** 2498–2501.

Rice, S. A. (1987), *Proc. Natl. Acad. Sci. USA* **84,** 4709–4716.

Rice, S. A., Gryko, J., Mohanty, U. (1986), in: C. A. Croxton (Ed.), *Fluid Interfacial Phenomena,* Chichester: Wiley, p. 255.

Robbins, M. O., Andelman, D., Joanny, J.-F. (1991), *Phys. Rev. A* **43,** 4344–4354.

Rowlinson, J. S., Widom, B. (1982), *Molecular Theory of Capillarity,* Oxford: Clarendon Press.

Safran, S. A. (1994), *Statistical Thermodynamics of Surfaces, Interfaces, and Membranes,* Reading, MA: Addison-Wesley.

Sanyal, M. K., Sinha, S. K., Huang, K. G., Ocko, B. M. (1991), *Phys. Rev. Lett.* **66,** 628–631.

Schlossman, M., Wu, X.-L., Franck, C. (1985), *Phys. Rev. B* **31,** 1478–1485.

Schlossman, M. L., Pershan, P. S. (1992), in: D. Langevin (Ed.), *Light Scattering by Liquid Surfaces and Complementary Techniques,* New York: Marcel Dekker Inc., p. 365.

Schmidt, J. W. (1988), *Phys. Rev. A* **38,** 567–570.

Schmidt, J. W. (1991), *Physica A* **172,** 40–52.

Schmidt, J. W., Moldover, M. R. (1993), *J. Chem. Phys.* **99,** 582–589.

Schwartz, D., Braslau, A., Ocko, B., Pershan, P. S., Als-Nielsen, J., Huang, J. S. (1988), *Phys. Rev. A* **38,** 5817–5824.

Schwartz, D. K., Schlossman, M. L., Kawamoto, E. H., Kellogg, G. J., Pershan, P. S., Ocko, B. M. (1990), *Phys. Rev. A* **41,** 5687–5690.

Sengers, J. V., van Leeuwen, J. M. J. (1989), *Phys. Rev. A* **39,** 6346–6355.

Sengers, J. V., van Leeuwen, J. M. J., Schmidt, J. W. (1991), *Physica A* **172,** 20–39.

Shen, Y. R. (1989), *Annu. Rev. Phys. Chem.* **40,** 327–350.

Sinha, S. K., Sirota, E. B., Garoff, S., Stanley, H. B. (1988), *Phys. Rev. B* **38,** 2297–2311.

Sluckin, T. J., Poniewierski, A. (1986), in: C. A. Croxton (Ed.), *Fluid Interfacial Phenomena,* Chichester: Wiley, p. 215.

Sluis, D., Rice, S. A. (1983), *J. Chem. Phys.* **79,** 5658–5672.

Smith, D. S. P., Law, B. M. (1995), *Phys. Rev. E* **52,** 580–595.

Somoza, A. M., Mederos, L., Sullivan, D. E. (1995), *Phys. Rev. E* **52,** 5017–5027.

Stillinger, F. H. (1967), *J. Chem. Phys.* **47,** 2513–2533.

Stringari, S., Treiner, J. (1987), *Phys. Rev. B* **36,** 8369–8375.

Sullivan, D. E., da Gama, M. M. T. (1986), in: C. A. Croxton (Ed.), *Fluid Interfacial Phenomena,* Chichester: Wiley, p. 45.

Swanson, B. D., Stragier, H., Tweet, D. J., Sorensen, L. B. (1989), *Phys. Rev. Lett.* **62,** 909–912.

Taborek, P., Rutledge, J. E. (1994), *Phys. B* **197,** 283–289.

Thomas, B. N., Barton, S. W., Novak, F., Rice, S. A. (1987), *J. Chem. Phys.* **86,** 1036–1047.

Tidswell, I. M., Rabedeau, T. A., Pershan, P. S., Kosowsky, S. D. (1991), *Phys. Rev. Lett.* **66,** 2108–2111.

Tkachenko, A. V., Rabin, Y. (1996), *Phys. Rev. Lett.* **76,** 2527–2530.

Ulman, A. (1991), *An Introduction to Ultrathin Organic Films,* San Diego, CA: Academic Press.

Weast, R. C. (Ed.) (1989), *CRC Handbook of Chemistry and Physics,* Boca Raton, FL: CRC Press.

Weeks, J. D. (1977), *J. Chem. Phys.* **67,** 3106–3121.

White, D. W. G. (1966), *Trans. Metall. Soc. AIME* **236,** 796–803.

White, D. W. G. (1971), *J. Inst. Met.* **99,** 287–289.

Williams, S. M., Zhang, Z., Mitrinovic, D., Schlossman, M. L. (1996), *Bull. Am. Phys. Soc.* **41,** 174.

Wu, E. S., Webb, W. W. (1973), *Phys. Rev. A* **8,** 2065–2076.

Wu, X. Z., Ocko, B. M., Sirota, E. B., Sinha, S. K., Deutsch, M. (1993a), *Physica A* **200,** 751–758.

Wu, X. Z., Ocko, B. M., Sirota, E. B., Sinha, S. K., Deutsch, M., Cao, B. H., Kim, M. W. (1993b), *Science* **261,** 1018–1021.

Wu, X. Z., Ocko, B. M., Tang, H., Sirota, E. B., Sinha, S. K., Deutsch, M. (1995), *Phys. Rev. Lett.* **75,** 1332–1335.

Wu, X. Z., Sirota, E. B., Sinha, S. K., Ocko, B. M., Deutsch, M. (1993c), *Phys. Rev. Lett.* **70,** 958–961.

Zhao, W., Zhao, X., Sokolov, J., Rafailovich, M. H., Sanyal, M. K., Sinha, S. K., Cao, B. H., Kim, M. W., Sauer, B. B. (1992), *J. Chem. Phys.* **97,** 8536–8542.

Zhou, X.-L., Lee, L.-T., Chen, S.-H., Strey, R. (1992), *Phys. Rev. A* **46,** 6479–6489.

Further Reading

Adamson, A. W. (1990), *Physical Chemistry of Surfaces,* New York: Wiley.

Croxton, C. A. (1980), *Statistical Mechanics of the Liquid Surface,* Chichester: Wiley.

Croxton, C. A. (Ed.) (1986), *Fluid Interfacial Phenomena,* Chichester: Wiley.

Defay, R., Prigogine, I., Bellemans, A., Everett, D. H. (1966), *Surface Tension and Adsorption,* London: Longmans, Green & Co. Ltd.

Dietrich, S. (1988), in: C. Domb, J. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena,* Vol. 12, London: Academic Press, p. 2.

Gompper, G., Schick, M. (1994), in: C. Domb, J. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena,* Vol. 14, London: Academic Press.

Langevin, D. (1992), *Light Scattering by Liquid Surfaces and Complementary Techniques,* New York: Marcel Dekker Inc.

Rowlinson, J. S., Widom, B. (1982), *Molecular Theory of Capillarity,* Oxford: Clarendon Press.

Safran, S. A. (1994), *Statistical Thermodynamics of Surfaces, Interfaces, and Membranes,* Reading, MA: Addison-Wesley Publishing Co.

SURFACES AND INTERFACES OF SOLIDS

See Supplement Volumes

SURFACES, ELECTRONIC STRUCTURE

See ELECTRONIC STRUCTURE OF SURFACES

SYMMETRIES AND CONSERVATION LAWS

See Supplement Volumes

SYMMETRY, MOLECULAR

MAGDOLNA HARGITTAI, *Hungarian Academy of Sciences, Budapest, Hungary*
ISTVAN HARGITTAI, *Budapest Technical University, Budapest, Hungary*

Introduction 337
1. **Description of Molecular Symmetry** 339
1.1 Symmetry Notations 340
1.2 Establishing the Molecular Point Group 341
1.3 Examples 341
2. **Determination of Molecular Symmetry** 343
2.1 Diffraction Techniques 343
2.2 Spectroscopic Techniques 344
2.3 Molecular Motion and Symmetry 344
2.4 Computational Techniques 345
2.5 A Qualitative Model: VSEPR .. 345
3. **Applications** 346
3.1 Molecular Vibrations 347
3.1.1 Normal Modes 347
3.1.2 Symmetry Coordinates 348
3.1.3 Selection Rules 348
3.2 Electronic Structure of Molecules 349
3.2.1 Symmetry Properties of the Wave Function 349
3.2.2 Construction of Molecular Orbitals 349
3.2.3 Electronic States 350
3.3 Chemical Reactions 351
3.3.1 Reaction Coordinate 351
3.3.2 Electronic Structure 352
3.3.2.1 Frontier Orbitals: HOMO and LUMO 352
3.3.2.2 Conservation of Orbital Symmetry 352
3.3.2.3 State Correlation 352
3.3.2.4 Generalized Woodward–Hoffmann Rules 352
Acronyms 353
Glossary 353
Works Cited 354
Further Reading 354

INTRODUCTION

A molecule consists of the nuclei of its constituent atoms and an electron density distribution. Symmetry considerations are important practically everywhere that molecules are considered. Among different branches of science it is mostly chemistry that is concerned with molecules, but other disciplines such as physics, crystallography, materials science, molecular biology, spectroscopy, just to mention the most obvious ones, are intimately related.

When we talk about the symmetry of a molecule we mean the symmetry of the geometrical arrangement of the atomic nuclei. A molecule may be "very symmetrical" or may have no symmetry at all. The possibility of applying different physical techniques to study molecules depends on their symmetry just as many other properties.

Molecular symmetry is a basic characteristic of molecular structure, which, in turn, determines all physical, chemical, and biological properties of a molecular substance. A recent conspicuous example of the importance of molecular symmetry is the discovery of buckminsterfullerene, the C_{60} molecule of the truncated icosahedral shape. The discovery originated from a synergistic interaction of astrophysical and cluster research and opened a whole new field of fullerene science and technology. The discovery was preceded by a prediction of the superstable C_{60} structure purely on molecular symmetry considerations.

But what is symmetry? It is one of the fundamental concepts of science, extending also to the humanities and the arts (Hargittai, 1986, 1989; Hargittai and Hargittai, 1994; Weyl, 1952). In a broad sense it refers to harmony and proportion. In a strict sense, though, it means a one-to-one correspondence among parts of an object. The word it-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

self comes from the Greek *symmetria*, meaning "the same measure." Thus, e.g., the common mirror symmetry occurs when two halves of a whole are each other's mirror images. The action that characterizes a *symmetry type* (e.g., mirror symmetry) is called a *symmetry operation* (e.g., reflection by a mirror). The tool (e.g., the *mirror*, or *symmetry plane*) by whose means the operation is performed is called a *symmetry element*. Another symmetry element is a *rotation axis*. Its *order* tells us how many times the original figure recurs during a complete rotation. The *elemental angle* is the smallest angle of rotation by which the original figure can be reproduced. Yet another type of symmetry is called *inversion*, and the corresponding symmetry element is the *inversion point* or *center of symmetry*. In this case equivalent positions are related along a line going through the center. Inversion can also be looked at as a consecutive application of two simple symmetry operations, namely, a twofold rotation followed by a reflection through a plane perpendicular to the twofold rotation axis. Sometimes a fourth symmetry element, the *mirror–rotation axis* (called also *improper axis*) is distinguished. The corresponding symmetry operation, mirror–rotation or improper rotation, consists of two steps, a proper rotation followed by a reflection through a plane perpendicular to the rotation axis. When the rotation is twofold this symmetry is the inversion. The first three basic symmetry elements are indicated in Figure 1.

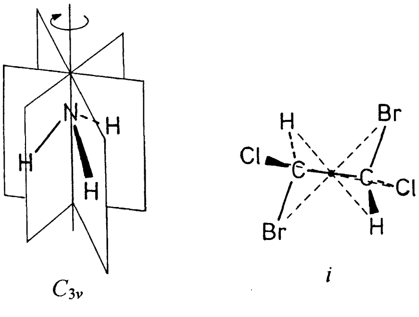

FIG. 1. Examples of molecules with (a) mirror symmetry (three mirror planes) and rotational symmetry (threefold rotation), and (b) inversion symmetry.

There are two large groups of symmetry, *point groups* and *space groups*. Point groups have at least one special point in the object or pattern that differs from all the others. It is also called a unique or singular point. Such a point or points remain unchanged during the symmetry operation. In space groups there is no unique point. Space groups are characterized by infinite repetition in one, two, or three directions in which case they are called one-, two-, or three-dimensional space groups, respectively. Extended arrays and networks of atoms and molecules, and, ultimately, crystal structures are described by space-group symmetries.

Molecules are finite with at least one singular point in their symmetry description. Thus point groups are applicable to them. There is no inherent limitation on the available symmetries for molecules, while severe restrictions apply to the symmetries of crystals. In fact, molecules occupy a more fundamental level in the hierarchy of structures than do crystals, and many crystals themselves are built from molecules.

Molecules are never motionless. They are performing vibrations all the time. Molecular vibrations constitute relative displacements of the atomic nuclei with respect to their equilibrium positions and occur at all times and in all states of matter, including the crystalline state, and even at the lowest possible temperatures. The magnitude of molecular vibrations is not negligible at all; they amount to several percent of the internuclear distances. Typically, there are about 10^{12}–10^{14} vibrations per second.

Symmetry considerations are fundamental in any description of molecular vibrations, as will be seen later in detail. First, however, the molecular symmetries are discussed ignoring entirely the motion of the molecules. Various molecular symmetries are illustrated by their notations and examples. The most important techniques of the determination of the symmetry of molecules are enumerated with some comments, and a simple model is also discussed to gain some insight into the origins of the various shapes and symmetries of molecules. Our considerations are restricted, however, to relatively simple systems. Finally, the potentials of applications of the symmetry concept and description are demonstrated for the discussion of molecular vibrations, the electronic structure of molecules, and chemical reactions. For a more detailed account of the symmetry aspects of chemistry, we refer the reader to Hargittai and Hargittai (1995).

1. DESCRIPTION OF MOLECULAR SYMMETRY

Molecules with the same formula but in which the distances between corresponding atoms are not all the same are called *structural isomers* (Fig. 2). They are of two types. If their atomic connectivities are the same, they are *diastereomers,* and if their atomic connectivities are different, they are *constitutional isomers.* Connectivity is the order by which the atoms are linked through chemical bonds in the molecule. Some diastereomers can become superimposable by rotation about a bond, and they are called *rotational isomers.* Depending on the magnitude of the barrier to rotation, *geometrical isomers* (high barrier) and *conformers* (low barrier) are distinguished (see CONFORMATIONAL ANALYSIS).

Identical molecules have the same formula, the same atomic connectivity, and the same distances between corresponding atoms. Besides, they are superimposable (homomers). Enantiomers have the same formula, the same atomic connectivity, and the same distances between corresponding atoms, but they have no symmetry plane, and they are not superimposable. On the other hand, they are mirror images of each other. They are also called *chiral* (possessing handedness) pairs. Thus chirality is related to the absence of a certain symmetry element (the symmetry plane). This is a case of dissymmetry, a term first used by Louis Pasteur. Pierre Curie (1894) suggested a broad application of this term: "Dissymmetry creates the phenomenon" ["C'est la disymétrie qui crée le phénomène"], that is, a phenomenon is observable because of the absence of some symmetries from the system. The simplest chiral molecules are those in which a carbon atom is surrounded by four different atoms at the vertices of a tetrahedron. Molecules of many substances can be right-handed and left-handed. Many biologically important compounds exist in either the left-handed or the right-handed form in the living organism.

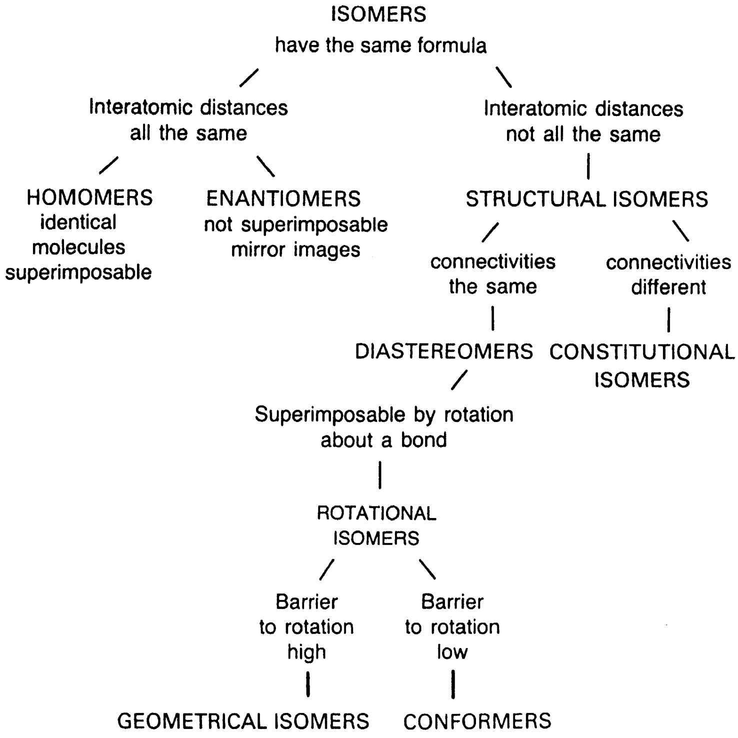
ISOMERS
have the same formula
Interatomic distances all the same
Interatomic distances not all the same
HOMOMERS identical molecules superimposable
ENANTIOMERS not superimposable mirror images
STRUCTURAL ISOMERS
connectivities the same
connectivities different
DIASTEREOMERS
CONSTITUTIONAL ISOMERS
Superimposable by rotation about a bond
ROTATIONAL ISOMERS
Barrier to rotation high
Barrier to rotation low
GEOMETRICAL ISOMERS
CONFORMERS

FIG. 2. The hierarchy of isomers.

For example, all naturally occurring amino acids are left-handed (except glycine). The two forms of a compound differing in handedness only may have very different properties. Figure 3 shows the formulas of the chiral pair of bitter and sweet asparagine.

1.1 Symmetry Notations

The so-called International or Hermann–Mauguin symmetry notations have been widely used in crystallography (see also CRYSTALLOGRAPHY). Another, older system, by Schoenflies, is generally used in spectroscopy to describe the molecular point-group symmetries. The Schoenflies notation has the advantage of succinct expression for even complicated symmetry classes combining various symmetry elements. The two systems are compiled in Table 1 for a selected set of symmetry classes. The set includes all point-group symmetries in the world of crystals, which are restricted to 32 classes. There are no restrictions on the point-group symmetries for individual molecules, and a few further, so-called *limiting,* classes are also listed in Table 1.

The Schoenflies notation for rotation axes is C_n, and for mirror–rotation axes the notation is S_{2n}, where n is the order of the rotation. The symbol i refers to the center of symmetry. Symmetry planes are labeled σ; σ_v is a vertical plane that always coincides with the rotation axis with an order of two or higher, and σ_h is a horizontal plane, which is always perpendicular to the rotation axis when it has an order of two or higher. The Hermann–Mauguin notations are not used for indicating molecular symmetries, and so they will only be mentioned briefly. Here n is used to indicate an n-fold rotation axis, m to indicate a mirror plane; a bar above a number indicates mirror–rotation axis, a slash after a number indicates rotation axis, and the letter m indicates that the reflection plane is perpendicular to the rotation axis.

Table 1. Examples of symmetry notations.

Hermann–Mauguin	Schoenflies
Crystallographic groups	
1	C_1
$\bar{1}$	C_i
m	C_s
2	C_2
$2/m$	C_{2h}
mm	C_{2v}
222	D_2
mmm	D_{2h}
4	C_4
$\bar{4}$	S_4
$4/m$	C_{4h}
$4mm$	C_{4v}
$\bar{4}2m$	D_{2d}
422	D_4
$4/mmm$	D_{4h}
3	C_3
$\bar{3}$	S_6
$3m$	C_{3v}
32	D_3
$\bar{3}m$	D_{3d}
$\bar{6}$	C_{3h}
6	C_6
$6/m$	C_{6h}
$\bar{6}m2$	D_{3h}
$6mm$	C_{6v}
622	D_6
$6/mmm$	D_{6h}
23	T
$m\bar{3}$	T_h
$43m$	T_d
432	O
$m\bar{3}m$	Q_h
Limiting groups	
∞	C_∞
$\infty 2$	D_∞
∞/m	$C_{\infty h}$
∞mm	$C_{\infty v}$
∞/mm	$D_{\infty h}$

Point-group symmetries not listed in Table 1 may easily be assigned the appropriate Schoenflies notation by analogy. Thus, e.g.,

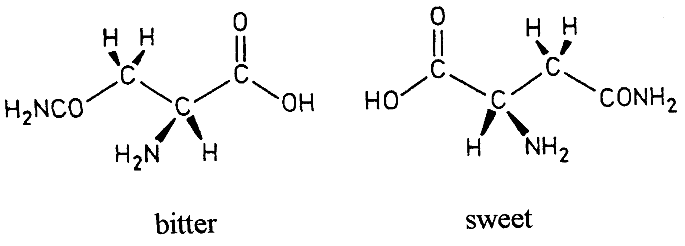

FIG. 3. Chiral pair of bitter and sweet asparagin.

C_{5v}, C_{5h}, C_7, C_8, etc. can be established. Such symmetries may well occur among real molecules.

1.2 Establishing the Molecular Point Group

Figure 4 shows a possible scheme for establishing the molecular point group. The symmetry of most molecules may be reliably established by this scheme.

First an examination is carried out whether the molecule belongs to some "special" group. If the molecule is linear, it may have a perpendicular symmetry plane ($D_{\infty h}$), or it may not have one ($C_{\infty v}$). Very high symmetries, such as tetrahedral, octahedral, or icosahedral, are easy to recognize. Each of the tetrahedral (T, T_h, and T_d) and octahedral (O and O_h) groups has four threefold rotation axes. Both icosahedral groups (I and I_h) require ten threefold rotation axes and six fivefold rotation axes. The molecules belonging to these groups have a central tetrahedron, octahedron, cube, or icosahedron.

If the molecule does not belong to one of these "special" groups, a systematic approach is followed. First, the possible presence of rotation axes in the molecule is checked. If there is no rotation axis, then it is determined whether there is a symmetry plane (C_s). In the absence of rotational axes and mirror planes, there may only be a center of symmetry (C_i), or there may be no symmetry element at all (C_1). If the molecule has rotation axes, it may have a mirror–rotation axis with even-number order (S_{2n}) coinciding with the rotation axis. For S_4 there will be a coinciding C_2, for S_6 a coinciding C_3, and for S_8, both C_2 and C_4.

In any case the search is for the highest-order C_n axis. Then it is examined whether there are n C_2 axes present perpendicular to the C_n axis. If such C_2 axes are present, then there is D symmetry. If in addition to D symmetry there is a σ_n plane, the point group is D_{nh}, while if there are n symmetry planes (σ_d) bisecting the twofold axes, the point group is D_{nd}. If there are no symmetry planes in a molecule with D symmetry, the point group is D_n.

Finally, if no C_2 axes perpendicular to C_n are present, then the lowest symmetry will be C_n; when a perpendicular symmetry plane is present, it will be C_{nh}; and when there are n coinciding symmetry planes, the point group will be C_{nv}.

1.3 Examples

Actual molecular structures are shown for various point groups in Fig. 5 in addition to those in Fig. 1, which correspond to the point groups (a) C_{3v} and (b) C_i. The Schoen-

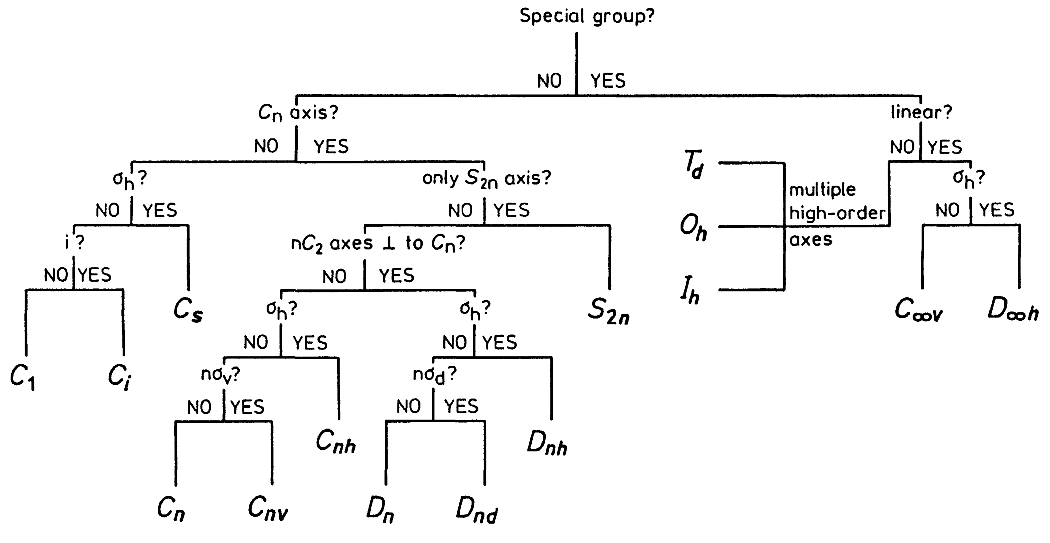

FIG. 4. Scheme for establishing the molecular point group.

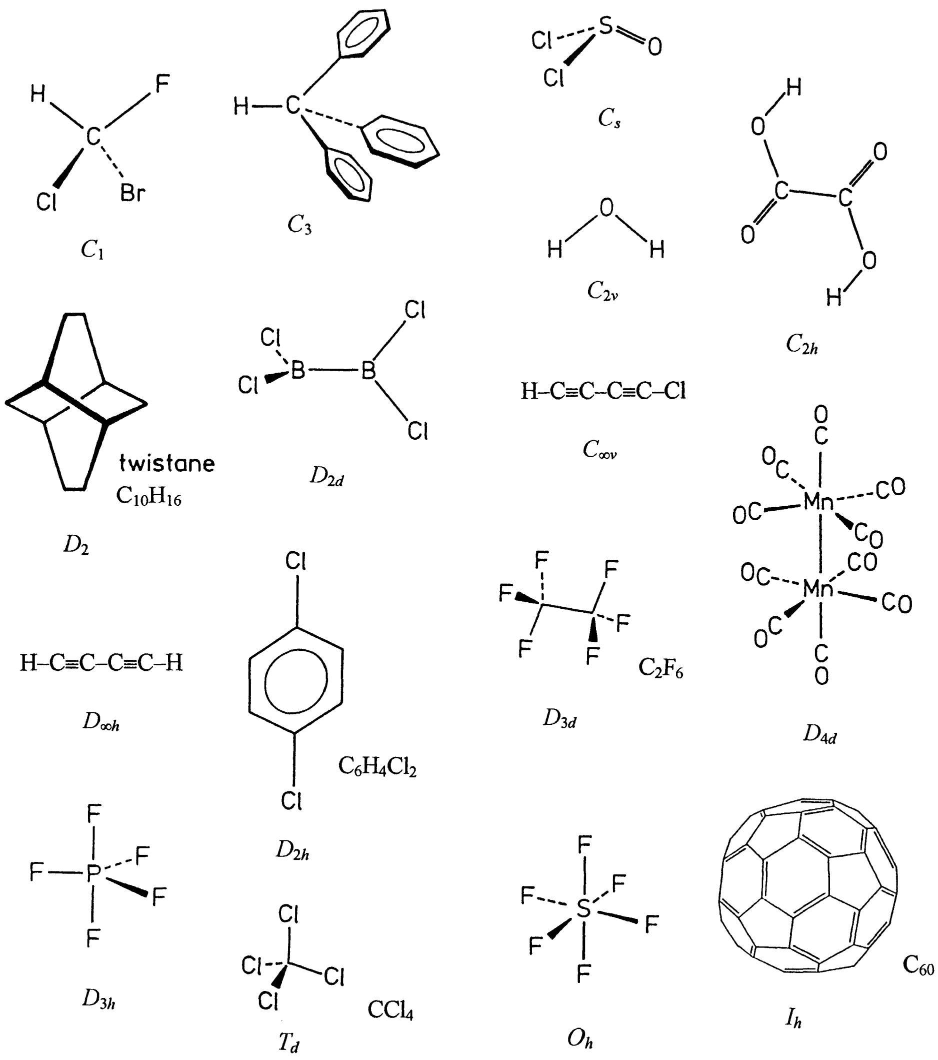

H
F
C
Cl
Br
C_1
H—C
C_3
Cl
S
O
Cl
C_s
O
H
H
C_{2v}
H
O
C
C
O
O
O
H
C_{2h}
twistane
$C_{10}H_{16}$
D_2
Cl
Cl
Cl
B
B
Cl
D_{2d}
H–C≡C–C≡C–Cl
$C_{\infty v}$
O
C
OC
Mn
CO
OC
CO
CO
OC
Mn
CO
OC
C
O
D_{4d}
H–C≡C–C≡C–H
$D_{\infty h}$
Cl
Cl
$C_6H_4Cl_2$
D_{2h}
F
F
F
F
F
F
C_2F_6
D_{3d}
F
F
P
F
F
F
D_{3h}
Cl
Cl
Cl
Cl
CCl_4
T_d
F
F
F
S
F
F
F
O_h
C_{60}
I_h

FIG. 5. Examples of molecular symmetries.

flies notation is used. In some cases the characteristic symmetry elements are enumerated.

C_1: There are no symmetry elements except the onefold rotation axis, or identity, of course. C_1 symmetry is *asymmetry*, that is, the complete lack of any symmetry. An example is shown in Fig. 5.

C_n: One n-fold rotation axis. Example: Fig. 5 with C_3 symmetry.

C_i: Center of symmetry. Example: Fig. 1(b).

C_s: One symmetry plane. Example: Fig. 5.

C_{nh}: One n-fold rotation axis with a symmetry plane perpendicular to it. Example: Fig. 5 with a molecule of C_{2h} symmetry.

C_{nv}: One n-fold rotation axis with n symmetry planes that include the rotation axis. When n is even, there are two sets of symmetry planes. One set is rotated relative to the other set by $(180/n)°$. The angle between the planes within each set is $(360/n)°$. When n is odd, the angle between the symmetry planes is $(180/n)°$. Examples: Fig. 1 (C_{3v}) and Fig. 5 (C_{2v} and $C_{\infty v}$).

D_n: One n-fold rotation axis and n twofold rotation axes perpendicular to the n-fold axis. When n is even, there are two sets of twofold rotation axes. One set is rotated relative to the other set by $(180/n)°$. The angle between the axes within each set is $(360/n)°$. When n is odd, the angle between the twofold axes is $(180/n)°$. Example: Fig. 5 (D_2).

D_{nd}: One n-fold rotation axis with n twofold rotation axes perpendicular to it, and n symmetry planes. The angle between the twofold axes is $180/n°$. The symmetry planes include the n-fold axis and bisect the angles between the twofold axes. Example: Fig. 5 (D_{2d}, D_{3d}, and D_{4d}).

D_{nh}: There is one n-fold rotation axis, one symmetry plane perpendicular to it, and n symmetry planes that include the n-fold axis. When n is even, there are two sets of symmetry planes. One set is rotated relative to the other set by $(180/n)°$. The angle between the planes within each set is $(360/n)°$. When n is odd, the angle between the symmetry planes is $(180/n)°$. Example: Fig. 5 (D_{2h}, D_{3h}, and $D_{\infty h}$).

T: Three mutually perpendicular twofold rotation axes and four threefold rotation axes. The threefold axes all go through a vertex of a tetrahedron and the midpoint of the opposite face center. The twofold axes connect the midpoints of opposite edges of this tetrahedron.

T_d: In addition to the symmetry elements of symmetry T, there are six symmetry planes, each two of them being mutually perpendicular. Each of these symmetry planes contains two threefold axes. Example: Fig. 5.

O_h: Three mutually perpendicular fourfold rotation axes and four threefold rotation axes that are tilted with respect to the fourfold axes in a uniform manner, and a center of symmetry. Example: Fig. 5.

I_h: The most characteristic feature of this point group is the presence of six fivefold rotation axes. Example: Fig. 5.

2. DETERMINATION OF MOLECULAR SYMMETRY

In this section the most important methods are briefly introduced by which the symmetry of a molecule may be determined. Emphasis is placed on the accuracy of information and the main potentials and limitations of each method (Domenicano and Hargittai, 1992). In conclusion a qualitative model will be mentioned that has been used successfully for predicting the shapes and symmetries of large classes of simple molecules (Gillespie and Hargittai, 1991).

Some methods give only qualitative information on the general shape and symmetry of a molecule, while others give this information as part of a full, quantitative determination of molecular geometry.

2.1 Diffraction Techniques

These techniques are based on scattering phenomena in which a portion of the radiation changes direction without energy transfer. Diffraction effects arise from interference within the scattered radiation. This radiation may be electromagnetic, such as x rays, or very fast-moving small particles, such as electrons or neutrons. All three methods give quantitative information on molecular geometry.

X rays are scattered by the electron cloud of an atom. From the intensities and directions of the scattered beams of x rays, the three-dimensional pattern of electron-density distribution of the molecule can be deduced. From this the positions of the atomic nuclei can be determined, particularly for heavy atoms, while the accuracy is considerably lower for lighter atoms. Asphericity of the electron-density distribution of atoms introduces uncertainty into the determination of the nuclear position. Another limitation originates from the vibrational motion of the atoms relative to each other and the librational motion of the molecule as a whole. *X-ray diffraction* is primarily applied to crystalline solids, where the structure of a molecule may be affected by interactions with its neighbors.

In *neutron diffraction,* neutrons are scattered directly by the atomic nuclei, and they are scattered about as strongly by light atoms as by heavy atoms. Neutron diffraction is also primarily applied to crystalline solids. Its biggest disadvantage is that a suitable beam of neutrons may not be readily available.

The electrostatic interaction between matter and electrons is very strong in *electron diffraction.* The relative positions of the nuclei can be determined directly. Electron diffraction is applied to molecules in the gaseous state. An obvious limitation is that it is restricted to substances that have a vapor pressure of a few Torr at a temperature at which they are stable. The structure of only relatively simple molecules can be determined with this technique. The reliability of structural information may be enhanced considerably by using complementary data from other sources.

2.2 Spectroscopic Techniques

These are based on interactions that are accompanied by a transfer of energy between electromagnetic radiation and matter.

A rotating permanent dipole, such as a polar molecule in the gas phase, interacts with electromagnetic radiation if the frequency of the radiation is commensurate with the rotational frequencies of the dipole. The rotational frequencies of a molecule depend on the three moments of inertia. Depending on the symmetry of a molecule, spherical tops, symmetric tops, and asymmetric tops are distinguished, differing in the number of different values for the moments of inertia about three mutually perpendicular axes. For all, except the lightest, molecules the spectrum is in the microwave region, and so this spectroscopy is usually called *microwave spectroscopy*. It yields a complete geometry for simple molecules in favorable cases.

The most common determination of molecular symmetry is by *vibrational spectroscopy*. Displacement of the atoms during molecular vibrations leads to oscillations of the electric charge distribution in a molecule. The oscillating charge distribution interacts with the oscillating electric vector of the electromagnetic radiation. If the frequency of the vibration is commensurate with the frequency of the electromagnetic radiation, a quantum of energy may be absorbed by the molecule, raising it to a higher vibrational level. The vibrational frequencies of a molecule are generally in the range of 10^{14}–10^{12} Hz, and the corresponding frequencies are in the infrared region of the spectrum.

The vibrational frequencies of a molecule may also be obtained from the so-called Raman spectrum. When a molecule interacts with radiation of much higher frequency than its vibrational frequencies, it may absorb a quantum of radiation. This brings the molecule to an excited electronic state, but the molecule almost immediately re-emits a photon. The energy of the re-emitted photon may differ from the energy of the incident photon by a quantum of vibrational energy. The resulting spectrum is the Raman spectrum. A vibration will give rise to a Raman line if it leads to a change in the polarizability of the molecule.

2.3 Molecular Motion and Symmetry

Molecules are permanently in motion. The translational motion has no effect on the shape of the molecule. There is usually a slight centrifugal distortion due to rotation. The consequences of intramolecular vibrations may be truly important in the determination of molecular symmetry. As a consequence of molecular vibrations a molecule may appear to have a lower symmetry than the one ascribed to its equilibrium configuration in which the molecule is supposed to be motionless. Since the averaging over molecular vibrations is not the same in each technique, the apparent structures originating from different experiments may be different for the same molecule. Metal dihalide molecules, with their floppy nature, are typical examples. Thus, e.g., $MgCl_2$ is a linear molecule in its equilibrium state, and this is what we measure by spectroscopic techniques. Electron diffraction, however, because of vibrational averaging, "sees" the molecule as bent with a Cl–Mg–Cl angle of about 155°. The consequences of bending motion need to be eliminated via vibrational corrections in order to derive the equilibrium structure and symmetry of such a molecule.

An extreme but not uncommon case of intramolecular motion is the exchange, or permutation, of the atomic nuclei in the molecule. An example is the pseudorotation of the cyclopentane molecule. Pseudorotation refers to a process in which identical atoms permute among nonequivalent sites, and the net result looks like a rotation of the molecule.

The large-amplitude pseudorotation of the PF_5 molecule is called Berry pseudorotation. The fluorine nuclear magnetic resonance spectrum of PF_5 displays a single fluorine frequency, pointing to the equivalence of the fluorine sites. On the other hand, electron diffraction has established the trigonal bipyramidal structure of this molecule with D_{3h} symmetry in which the axial and equatorial sites are distinctly different. According to the Berry pseudorotation the axial pair of fluorines bends and forms a triangle together with one of the equatorial fluorines. The other two equatorial fluorines become the new axial atoms. There is a transition square pyramidal structure of C_{4v} symmetry in between the two trigonal bipyramidal structures (Fig. 6).

The Berry pseudorotation occurs fast compared with the nuclear magnetic resonance time scale, and slowly compared with the electron-diffraction time scale. Thus the understanding of the time-scale relationships between the magnitudes of interaction time in a physical measurement and the lifetime of a particular structure is fundamental to our understanding of the information on molecular symmetry from different physical techniques. Although, in principle, permutation of the nuclei may occur in all molecules, it happens billions of times a second in a PF_5 (D_{3h}) molecule, while it may happen only once or twice in 10^{15} years in a methane (CH_4) molecule (T_d).

2.4 Computational Techniques

The computational determination of molecular geometry, including molecular symmetry, has developed considerably in recent years as the capabilities of computers and the theoretical understanding of molecular structure have increased. It is possible, at least in principle, to compute all molecular properties, including symmetry, from the knowledge of the composition and atomic connectivity of the molecule by solving the corresponding Schrödinger equation. When this is done without invoking any additional information, the technique is called *ab initio* ("from the beginning"). These calculations yield the so-called equilibrium geometry corresponding to the minimum position of the molecular potential energy function. Thus it may be different from the average geometries yielded by the various experimental techniques. The computational determination of molecular symmetry, at least for the simplest systems, is as good today as that by the best experimental techniques.

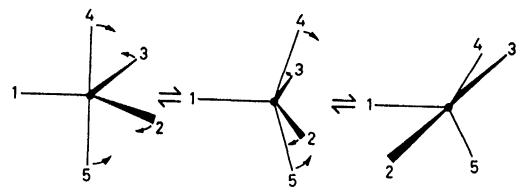

FIG. 6. Pseudorotation of a trigonal bipyramidal AX_5 molecule.

2.5 A Qualitative Model: VSEPR

In addition to the experimental and computational determination of molecular symmetry and structure, qualitative models are useful to predict molecular symmetry and its variations for series of substances. Such qualitative models are usually based on stressing a few effects, if not just one, in the complexity of intramolecular interactions and ignoring the rest. Thus it is always important to clearly delineate the scope of applicability of a qualitative model. A very successful and simple qualitative model of considerable predictive potential has been the so-called valence-shell electron-pair repulsion (VSEPR) model based on considerations of space requirements of the electron pairs.

It has been recognized that the number of electron pairs in the valence shell of atom A will determine the bond configuration of an AX_n molecule. If the valence shell contains unshared electron pairs as well, they also have to be considered. A formula such as AX_nE_m describes a molecule with a central atom A; n is the number of X ligands, meaning also that there are n A–X bonds, and m is the number of lone pairs in the valence shell of A. The arrangements of the electron pairs will always be such as to minimize their distances apart.

For two, three, four, five, and six electron pairs, i.e., when $n + m = 2, 3, 4, 5$, and 6, the arrangements of the electron pairs will be linear, trigonal planar, tetrahedral, trigonal bipyramidal, and octahedral. It is supposed here that the core beneath the valence shell has spherical symmetry, and the distribution of the electron pairs in the valence shell is the same as when mutually repelling points are distributed along the surface of the sphere. The power exponent in the expression of the potential energy of such repulsions is not known in our case of electron pairs in the valence shell. The utility of this model is in the invariance of the solutions to changes in the power exponent.

If all the electron pairs are bonds, the bond configuration will be the same as the arrangement of the electron pairs (Fig. 7). If one or more bonding pairs (bp) are replaced by lone pairs (lp), the bond configuration will be what corresponds to the original electron-pair arrangement being a part thereof (Fig. 7).

Establishing variations of molecular symmetry in series of substances, the relative spaces available at different sites in a given configuration, and the relative space requirements of bonds to various ligands, lone pairs vs. bonds, the multiplicity of bonds and other factors need additional consideration (Gillespie and Hargittai, 1991).

3. APPLICATIONS

Molecular properties can be of either static or dynamic nature. A static property remains unchanged by every symmetry

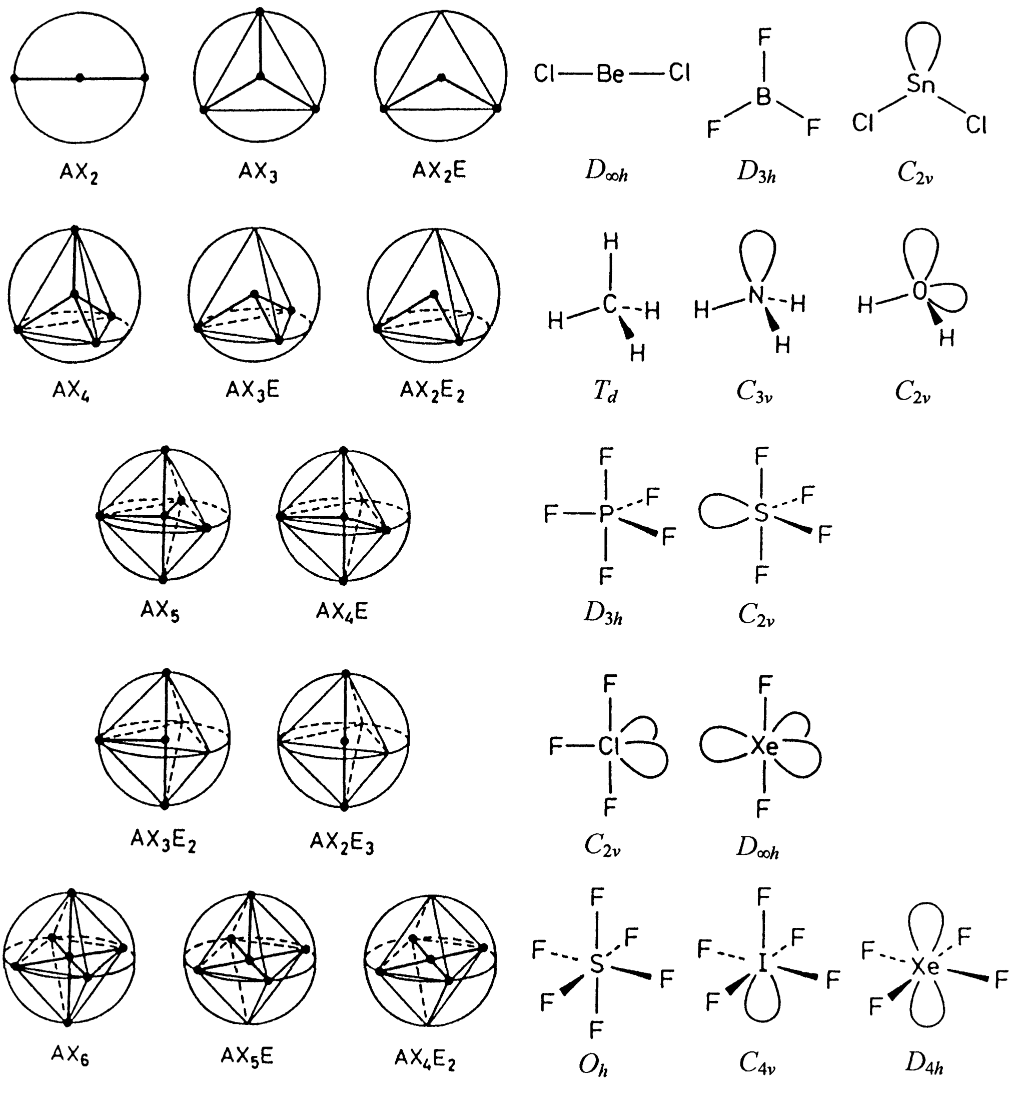
AX_2
AX_3
AX_2E
Cl—Be—Cl
$D_{\infty h}$
F B F F
D_{3h}
Sn Cl Cl
C_{2v}
AX_4
AX_3E
AX_2E_2
H C H H H
T_d
H N H H
C_{3v}
H O H
C_{2v}
AX_5
AX_4E
F P F F F F
D_{3h}
F S F F F
C_{2v}
AX_3E_2
AX_2E_3
F Cl F F
C_{2v}
F Xe F
$D_{\infty h}$
AX_6
AX_5E
AX_4E_2
F S F F F F F
O_h
F I F F F F
C_{4v}
F Xe F F F
D_{4h}

FIG. 7. Molecular shapes and symmetries based on the arrangements of two–six valence-shell electron pairs (after Gillespie and Hargittai, 1991).

operation carried out on the molecule. The geometry of the nuclear arrangement in the molecule is such a property: A symmetry operation transforms the nuclear arrangement into another, which will be indistinguishable from the initial.

Dynamic properties, on the other hand, may change under symmetry operations. Molecular motion itself is a dynamic property, and since real molecules are never motionless, the symmetry properties of their motion may greatly influence their chemical behavior. The two main kinds of motion in a molecule are nuclear and electronic.

Dynamic properties can be either *symmetric* or *antisymmetric* with respect to the various symmetry operations of the molecular point group. Group theory (*q.v.*) is the mathematical apparatus for describing symmetry operations. The first step in the symmetry determination of dynamic properties is the selection of the appropriate basis for the representation. Appropriate here means the correct representation of the changes in the properties examined. In the investigation of molecular vibrations (Sec. 3.1), either Cartesian displacement vectors or internal coordinate vectors are used. In the description of the molecular electronic structure (Sec. 3.2), the angular components of the atomic orbitals are frequently used bases. Since the angular wave function changes its "sign" under certain symmetry operations, its behavior will be characteristic of the spatial symmetry of a particular orbital. Molecular orbitals can also be used as a basis of representation. Considerations of symmetry properties play a role in determining whether a quantum mechanical integral of the type $\int \psi_i$ op. $\psi_j d\tau$ is different from zero or not. This is important in such areas as selection rules for electronic transitions, chemical reactions, infrared and Raman spectroscopy, and other spectroscopies.

3.1 Molecular Vibrations

Vibration is a special kind of motion: the atoms of every molecule are permanently changing their relative positions at every temperature (even at absolute zero) without changing the position of the molecular center of mass. In terms of the molecular geometry these vibrations amount to continuously changing bond lengths and bond angles. (Molecular vibrations are also dealt with under the heading MOLECULAR SPECTROSCOPY.)

Symmetry considerations are instrumental in discussing and understanding molecular vibrations. Here we give only a short account on what kind of information can be deduced about the internal motion of the molecule from the mere knowledge of its point-group symmetry. For more detailed treatment we refer to monographs by Herzberg (1959), Wilson *et al.* (1955), and Hargittai and Hargittai (1995).

3.1.1 Normal Modes The seemingly random motion of molecular vibrations can always be decomposed into the sum of relatively simple components, called *normal modes of vibration.* Each of the normal modes is associated with a certain frequency. Thus for a normal mode every atom of the molecule moves with the same frequency and in phase. There are $3N - 6$ (for the linear case $3N - 5$) normal modes of vibration in a molecule.

The close relationship between symmetry and vibration is expressed by the following rule: Each normal mode of vibration forms a basis for an irreducible representation of the point group of the molecule. (For each point group the irreducible representations can be found in the so-called character tables; see GROUP THEORY). From the symmetry group of the molecule the symmetry species of the normal modes can be determined without any additional information.

For this purpose the so-called *Cartesian displacement vectors* are used as bases. The obtained reducible representation can be reduced to a sum of irreducible representations using the reduction formula (see GROUP THEORY). Subtracting those irreducible representations that correspond to translation and rotation (on the basis of character tables), the symmetry of the normal modes can be determined. For example, the water molecule has three normal modes of vibration. The molecular point group is C_{2v}, and two of the normal modes belong to the A_1 and one to the B_2 irreducible representation (Fig. 8) of this point group.

The normal modes can usually—though not always—be associated with a certain kind of motion. Those connected mainly with changes in bond lengths are the *stretch-*

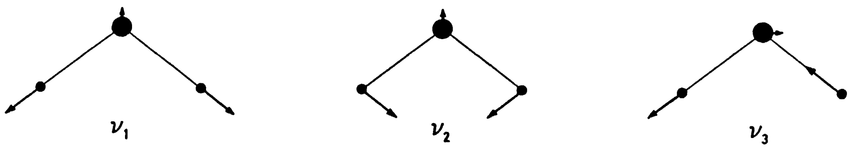

FIG. 8. Normal modes of vibration for the water molecule: ν_1, symmetric stretch (A_1); ν_2, bending (A_1); and ν_3, asymmetric stretch (B_2). The lengths of the arrows indicate the relative displacements of the atoms.

ing modes. The ones connected mainly with changes of bond angles are the *deformation modes*. These are mostly either in-plane or out-of-plane deformation modes. The simplest deformation mode is the *bending mode*. In case of water, one of the A_1- and the B_2-symmetry normal modes are stretching modes, and the other A_1-symmetry normal mode is a bending mode. The A_1-symmetry stretching mode is called the symmetric stretch, while the B_2 mode is the antisymmetric stretch (see Fig. 8).

3.1.2 Symmetry Coordinates The symmetry coordinates are symmetry-adapted linear combinations of the *internal coordinates*. The internal coordinates are the geometrical parameters of a molecule, and their displacement is used as a basis for representations. Symmetry coordinates can be generated from the internal coordinates by the use of the projection operator (see GROUP THEORY).

Both the symmetry coordinates and the normal modes of vibration belong to an irreducible representation of the point group of the molecule. A symmetry coordinate is always associated with one or another type of internal coordinate—that is, pure stretch, pure bend, etc.—whereas a normal mode can be a mixture of different internal coordinate changes of the same symmetry. For the water molecule, e.g., the symmetry coordinates are good approximations of the normal vibrations, and so their pictorial representation is the same.

The relationship between the symmetry coordinates and the normal vibrations can be established only by calculations called normal-coordinate analysis (see, e.g., Wilson *et al.*, 1955). These calculations necessitate further data in addition to the knowledge of molecular symmetry.

3.1.3 Selection Rules Depending on the symmetry of the molecule, not all normal modes will be observed in the infrared and Raman spectra because they do not all lead to a change in the dipole moment or in the polarizability of the molecule. In general, the number of bands observed in the infrared and Raman spectra of a molecule increases with decreasing molecular symmetry.

The considerations on the symmetries of the ground and excited states are instrumental in determining the selection rule for infrared spectroscopy: A fundamental vibration will be infrared active—i.e., appear as an absorption band in the infrared spectrum—if the corresponding normal mode belongs to the same irreducible representation as one or more of the Cartesian coordinates.

The selection rule for Raman spectroscopy can also be derived by similar reasoning. It says: A fundamental vibration will be Raman active if the normal mode undergoing the vibration belongs to the same irreducible representation as one or more of the components of the polarizability tensor of the molecule. These components are the quadratic functions of the Cartesian coordinates given in the fourth area of the character tables. The Cartesian coordinates themselves are given in the third area. Thus the symmetry of the normal modes of a molecule is sufficient information to tell what transitions will be infrared and what transitions will be Raman active. The normal modes of the water molecule belong to the A_1 and B_2 irreducible representation of the C_{2v} point group. Using merely the C_{2v} character table, it can be deduced that all three vibrational modes will be active in both the infrared and Raman spectra.

Since a particular normal mode may belong to different symmetry species in different point groups, its behavior depends strongly on the molecular symmetry. Just to mention one example, the ν_1 symmetric stretching mode of an AX_3 molecule is not infrared active if the molecule is planar (D_{3h} symmetry), i.e., it does not appear in the spectrum. It is infrared active and thus appears in the spectrum, however, if the molecule is pyramidal (C_{3v} symmetry). Vibrational spectroscopy is obviously one of the best experimental tools to determine the symmetry of molecules.

3.2 Electronic Structure of Molecules

Everything that counts in chemistry is related to the electronic structure of atoms and molecules (see ELECTRONIC STRUCTURE OF ATOMS AND MOLECULES). The formation of molecules from atoms, their behavior, and their reactivity all depend on their electronic structure. What is the role of symmetry in all this? In various aspects of the electronic structure, symmetry can tell us a good deal: why certain bonds can form and others cannot, why certain electronic transitions are allowed and others are not, and why certain chemical reactions occur and others do not.

3.2.1 Symmetry Properties of the Wave Function The symmetry properties of the electronic wave function and the energy of the system are two determining factors in chemical behavior. The relationship between the wave function characterizing the behavior of the electrons and the energy of the system—atoms and molecules—is expressed by the Schrödinger equation. In its general and time-independent form, it is usually written as

$$H\psi = E\psi, \tag{1}$$

where H is the Hamiltonian operator and E is the energy of the system.

With respect to the interchange of like particles (either nuclei or electrons), the Hamiltonian must be unchanged under a symmetry operation. This means that it belongs to the totally symmetric representation of the molecular point group.

A fundamental property of the wave function is that it can be used as basis for irreducible representations of the point group of a molecule. This property establishes the connection between the symmetry of a molecule and its wave function.

The electronic wave function of a one-electron system can be represented as a product of a radial and an angular component:

$$\psi_e = R(r)A(\Theta,\Phi). \tag{2}$$

The separation of the wave function into two parts is convenient since these two parts relate to different properties. The radial part determines the energy of the system and is invariant to symmetry operations. The angular part, on the other hand, says nothing about the energy of the system, but it can be altered by symmetry operations. The function $A(\Theta,\Phi)$ may have different signs (+ and −) in different spatial regions. A change in sign indicates a drastic change in the wave function. These signs might be thought of as signs of the amplitudes of the wave function. The places where the wave function changes sign are called *nodes*. The number of nodes is $n - 1$, where n is the principal quantum number.

The well-known shapes of some one-electron orbitals are presented in Fig. 9; these are, in fact, representations of the angular wave functions. Such representations are used commonly for illustrations because they describe accurately the symmetry properties of the wave function. In order to give the total wave function, however, they must be multiplied by an appropriate radial function.

What are the symmetry properties of an orbital? An orbital can be either symmetric or antisymmetric with respect to a certain symmetry operation. In Fig. 10 are shown the $d_{x^2-y^2}$ and p_z orbitals and their behavior under the inversion operation. The d orbital is transformed into itself by inversion; both the magnitude and the "sign" of the wave function will remain the same; it is symmetric with respect to inversion. The effect of the inversion operation on the p orbital is different: its magnitude does not change, but its "sign" does upon inversion; the p orbital is antisymmetric with respect to inversion. In the character tables, this is indicated by +1 for symmetric and −1 for antisymmetric behavior under each symmetry operation. The atomic orbitals (AOs) always belong to the same irreducible representations of the given point group as do their subscripts (x, y, z, xy, $x^2 - y^2$, etc.).

3.2.2 Construction of Molecular Orbitals A molecule is not simply a collection of its constituting atoms; rather, it is a system of atomic nuclei and a common electron distribution. Nevertheless, in describing the

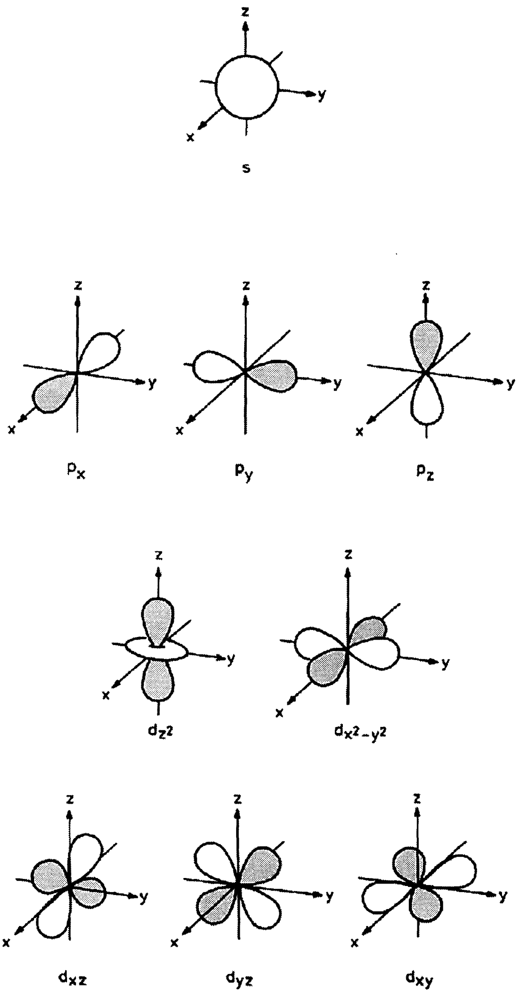

FIG. 9. Shapes of one-electron orbitals. They are representations of the angular wave function, $A(\Theta,\Phi)$.

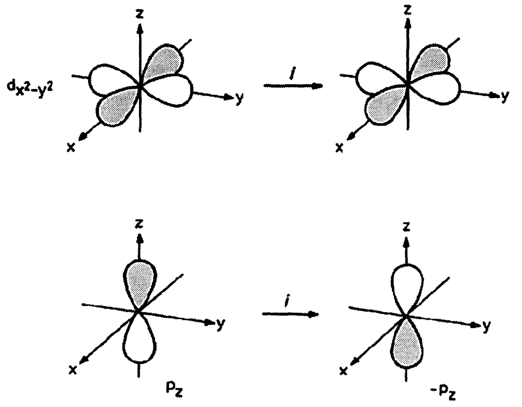

FIG. 10. The effect of inversion on the $d_{x^2-y^2}$ and p_z orbitals.

electronic structure of a molecule the most convenient way is to approximate the molecular electron distribution by the sum of atomic electron distributions. This approach is called the *linear combination of atomic orbitals* (LCAO) method. The orbitals produced by the LCAO procedure are called *molecular orbitals* (MOs).

Although both AOs and MOs are one-electron wave functions, the shape and symmetry of the MOs are different from those of the AOs of the isolated atom. Molecular orbitals extend over the entire molecule, and their spatial symmetry must conform to that of the molecular framework.

When constructing MOs from AOs, there may be a large number of possible linear combinations of AOs. Many of these linear combinations, however, are unnecessary. Symmetry is instrumental as a criterion in choosing among them. All those linear combinations must be eliminated that do not belong to any irreducible representation of the molecular point group. The reverse of this statement constitutes the fundamental principle of forming MOs: Each possible MO must belong to an irreducible representation of the molecular point group. Another equally important rule for the construction of MOs is that only those AOs can form a MO that belong to the same irreducible representation of the molecular point group.

When AOs are used for MO construction, their symmetry must be considered in the symmetry group of the particular molecule. A MO can only be formed from AOs of the same symmetry, and the symmetry of this new MO will be the same as that of the component AOs.

In addition to complying with symmetry rules, successful MO construction requires certain energy conditions. Two AOs will only combine for a MO if their energies are not too different.

3.2.3 Electronic States The orbitals and electronic configurations are useful descriptions. In many-electron systems, however, the energy of the orbitals loses its physical meaning, and only the energies of the (ground and excited) states are real. It is these states that are described by the total electronic wave functions. Electronic transitions, in fact, represent changes in the state of an atom or a molecule and not necessarily in the electronic configurations.

The symmetry of the electronic states can be determined from the symmetry of the occupied orbitals. There are two different cases:

1. States with fully occupied orbitals. An electronic configuration in which all orbitals are completely filled possesses only one electronic state, and it will be totally symmetric.
2. States with partially occupied orbitals. The completely filled orbitals are ignored (for reasons described above), and the symmetry of the state will be given by the direct product of the partially filled orbitals.

3.3 Chemical Reactions

The first application of symmetry considerations to chemical reactions were the Wigner–Witmer rules that are concerned with the conservation of spin and orbital angular momentum in the reaction of diatomic molecules. The real breakthrough in recognizing the role that symmetry plays in determining the course of chemical reactions has occurred only recently, mainly through the activities of Woodward and Hoffmann (1970), Fukui (1975), Bader (1962), Pearson (1976), and others. For other than symmetry aspects of chemical reactions, see CHEMICAL REACTIONS.

Symmetry phenomena may play as important a role in chemical reactions as they do in the construction of molecular orbitals or in molecular spectroscopy. It is even possible to make certain symmetry-based "selection rules" for the "allowedness" and "forbiddenness" of a chemical reaction, just as is done for spectroscopic transitions.

There are, of course, limitations to the applicability of symmetry rules for chemical reactions. Symmetry rules can usually be applied to comparatively simple reactions, the so-called *concerted reactions.* In a concerted reaction all relevant changes occur simultaneously; the transformation of reactants into products happens in one step with no intermediates.

The statement that a chemical reaction is "symmetry allowed" or "symmetry forbidden" should not be taken literally. When a reaction is symmetry allowed, it means that it has a low activation energy. This makes it possible for the given reaction to occur, though it does not mean that it always will. There are other factors that can impose a substantial activation barrier. Such factors may be steric repulsions, difficulties in approach and unfavorable relative energies of orbitals. Similarly, "symmetry forbidden" means that the reaction, as a concerted one, would have a high activation barrier. However, various factors may make the reaction still possible; for example, it may happen as a stepwise reaction through intermediates. In this case, of course, it is no longer a concerted reaction.

Most of the symmetry rules explaining and predicting chemical reactions deal with changes in the electronic structure. However, a chemical reaction is more than just that. Breakage of bonds and formation of new ones are also accompanied by nuclear rearrangements and changes in the vibrational behavior of the molecule.

As has been shown previously, both the vibrational motion and the electronic structure of the molecules strongly depend on symmetry. This dependence can be fully utilized when discussing chemical reactions. It is not only the reactant and product molecules whose electronic structure and vibrations can be described with the help of symmetry. Even more importantly, certain symmetry rules can be applied to the *transition state* between the reactants and products.

3.3.1 Reaction Coordinate The changing nuclear positions during the course of a reaction are collectively described by the term *reaction coordinate.* In simple cases we may assume that the chemical reaction is dominated by one of the normal modes of vibration, and thus this vibrational mode is the reaction coordinate. All the other normal vibrations are at their optimum value, so their energy is at minimum. Therefore, this reaction path can be taken as a *minimum-energy path.* Concerning symmetry properties, the reaction coordinate belongs to the totally symmetric irreducible representation of the molecular point group except at a maximum or at a minimum.

3.3.2 Electronic Structure A chemical reaction is a consequence of interactions between molecules. Before a reaction takes place (i.e., while the reacting molecules are still far apart), their electron distribution is unperturbed. When they approach each other, their orbitals begin to overlap, and distortion of the original electron distribution takes place. There are two requirements for a constructive interaction between molecules: symmetry matching and energy matching. These two factors can be treated in different ways. The approaches of Fukui (1975) and of Woodward and Hoffmann (1970) differ somewhat.

3.3.2.1 Frontier Orbitals: HOMO and LUMO. A successful chemical reaction requires both energy and symmetry matching between the MOs of the reactants. The requirements are essentially the same as in case of constructing MOs from AOs. Only orbitals of the same symmetry and comparable energy can overlap successfully. The most significant interactions can be expected between the highest occupied molecular orbital (HOMO) of one molecule and the lowest unoccupied molecular orbital (LUMO) of the other. The concept of these so-called *frontier orbitals* simplifies the MO description of chemical reactions enormously, since only these MOs of the reactant molecules need to be considered.

3.3.2.2 Conservation of Orbital Symmetry. Within this approach a correlation diagram is constructed from the orbitals of the reactants and the product(s). On one side of the diagram the approximate energy levels of the reactants are drawn, while on the other side those of the product(s) are indicated. A particular geometry of approach must be assumed. Furthermore, the symmetry properties of the molecular orbitals must be considered in the framework of the point group of the transition structure. Those molecular orbitals are taken into account that are associated with bonds that are broken or formed during the chemical reaction. Energy levels on the two sides of the diagram are to be connected in such a way that the so-called *noncrossing rule* is not violated. According to this rule, two orbitals of the same symmetry cannot intersect.

3.3.2.3 State Correlation. The correlation diagram discussed above refers to molecular orbitals. Since electronic transitions occur physically between electronic states, the correlation of these states is of interest. The rules for the state correlation diagrams are the same as for the orbital correlation diagrams: Only states that possess the same symmetry can be connected. Figure 11 shows the state correlation diagram for ethylene dimerization.

An obvious connection between states that possess the same electronic configuration would be the one indicated by the dashed lines in the figure. This does not occur, however, because states of the same symmetry cannot cross. Instead of crossing, when two states are coming too close to each other they will turn away, and so the two ground states, both of A_g symmetry, and also two A_g-symmetry excited states will each mutually correlate. The full line connecting the two ground states in Fig. 11 indicates that there is a substantial energy barrier to the ground-state-to-ground-state process; this reaction is said to be "thermally forbidden."

Consider now one electron in the reactant system excited photochemically to the B_{2g} state. Since this state correlates directly with the B_{2g} state of the product, this reaction does not have any energy barrier and may occur directly. It is said that the reaction is "photochemically allowed." Indeed, it is an experimental fact that olefin dimerization occurs smoothly under irradiation.

This observation can be generalized. If a concerted reaction is thermally forbidden, it is photochemically allowed and vice versa; if it is thermally allowed, then it is photochemically forbidden.

3.3.2.4 Generalized Woodward–Hoffmann Rules. *Cycloaddition.* The reaction between two molecules is thermally allowed

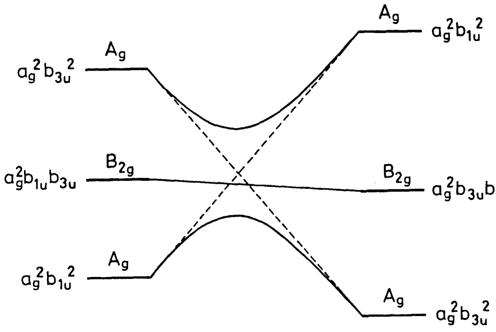

FIG. 11. State correlation diagram for the ethylene dimerization.

if the total number of electrons in the system is $4n + 2$ (n is an integer) and both components are either suprafacial or antarafacial. If one component is suprafacial and the other is antarafacial, the reaction will be thermally allowed if the total number of electrons is $4n$.

Electrocyclic reactions. The rules are similar to those given above. A disrotatory process is thermally allowed if the total number of electrons is $4n + 2$, and a conrotatory process is allowed thermally if the number of delocalized electrons is $4n$. For a photochemical reaction, both sets of rules are reversed.

ACRONYMS

AO: Atomic orbital.

HOMO: Highest occupied molecular orbital.

LCAO: Linear combination of atomic orbitals.

LUMO: Lowest unoccupied molecular orbital.

MO: Molecular orbital.

VSEPR: Valence-shell electron-pair repulsion.

GLOSSARY

Ab Initio Calculations: Nonempirical determination of molecular properties by computational techniques.

Antisymmetry: The symmetry of opposites.

Asymmetry: The absence of any symmetry.

Buckminsterfullerene: C_{60} molecule of truncated icosahedral shape.

Chiral: Not superimposable with its mirror image.

Chiral Pairs: Enantiomeric molecules.

Concerted Reaction: A reaction in which relevant changes occur simultaneously; the transformation of reactants into products happens in one step.

Connectivity: The order by which the atoms are linked by chemical bonds in the molecule.

Constitutional Isomers: Molecules having the same sum formula but different connectivities.

Dissymmetry: The absence of certain symmetry.

Elemental Angle: The smallest angle of rotation by which the original figure can be reproduced during rotation.

Enantiomers: Molecules that have the same formula, same connectivity, and same distances but have no symmetry plane and are not superimposable but are mirror images of each other.

Frontier Orbitals: The highest occupied molecular orbitals (HOMO) and lowest unoccupied molecular orbitals (LUMO) in a molecule.

Internal Coordinates: The geometrical parameters of a molecule (bond lengths, bond angles, and angles of torsion).

Inversion Point: The point remaining invariant under the operation of inversion. Each pair of points equivalent to one another under the operation can be joined by a straight line passing through the point.

Molecular Vibrations: Relative displacements of the atomic nuclei with respect to their equilibrium positions in the molecule.

Normal Modes of Vibration: Simple vibrations whose sum make up the seemingly random motion of molecular vibrations.

Order of Rotational Symmetry: The integer telling how many times the original figure recurs during a complete rotation.

Point Groups: Those symmetry groups for which there is at least one unique point remaining unchanged during a symmetry operation.

Pseudorotation: A process in which identical atoms permute among inequivalent sites and the net result looks like a rotation of the molecule.

Reaction Coordinate: Collective description of the changing nuclear positions in molecules during a chemical reaction.

Rotational Isomers: Diastereomers that can be superimposed by rotation about a bond.

Selection Rules: Rules that describe which normal modes will be active in the infrared and Raman spectra.

Space Groups: Symmetry groups that have no unique point.

Structural Isomers: Molecules with the same sum formula but in which the distances between corresponding atoms are not all the same.

Symmetry Coordinates: Symmetry-adapted linear combinations of internal coordinates.

Symmetry Element: The *tool* by whose means a certain symmetry operation is performed.

Symmetry Notations: A shorthand description of all symmetries.

Symmetry Operation: The *action* that characterizes a symmetry type.

VSEPR Model: Qualitative determination of molecular shape based on the space requirements of electron domains in the valence shell of the central atom of a molecule.

Works Cited

Bader, R. F. W. (1962), *Can. J. Chem.* **40,** 1164–1175.

Curie, P. (1894), *J. Phys. (Paris)* **3,** 393.

Domenicano, A., Hargittai, I. (Eds.) (1992), *Accurate Molecular Structures,* Oxford: Oxford Univ. Press.

Fukui, K. (1975), *Theory of Orientation and Stereoselection,* Berlin: Springer-Verlag.

Gillespie, R. J., Hargittai, I. (1991), *The VSEPR Model of Molecular Geometry,* Boston: Allyn and Bacon.

Hargittai, I. (Ed.) (1986), *Symmetry: Unifying Human Understanding,* New York: Pergamon Press.

Hargittai, I. (Ed.) (1989), *Symmetry 2: Unifying Human Understanding,* Oxford, UK: Pergamon Press.

Hargittai, I., Hargittai, M. (1994), *Symmetry—A Unifying Concept,* Bolinas, CA: Shelter Publ.

Hargittai, I., Hargittai, M. (1995), *Symmetry through the Eyes of a Chemist,* 2nd ed., New York: Plenum Press.

Herzberg, G. (1959), *Infrared and Raman Spectra,* Princeton, NJ: Van Nostrand Company.

Pearson, R. G. (1976), *Symmetry Rules for the Chemical Reactions,* New York: Wiley-Interscience.

Weyl, H. (1952), *Symmetry,* Princeton, NJ: Princeton Univ. Press.

Wilson, E. B., Decius, J. C. Jr., Cross, P. C. (1955), *Molecular Vibrations,* New York: McGraw-Hill Book Co.

Woodward, R. B., Hoffmann, R. (1970), *The Conservation of Orbital Symmetry,* Weinheim: Verlag Chemie.

Further Reading

Cotton, F. A. (1990), *Chemical Applications of Groups Theory,* 3rd ed., New York: Wiley.

Gillespie, R. J., Hargittai, I. (1991), *The VSEPR Model of Molecular Geometry,* Boston: Allyn and Bacon.

Hargittai, I. (Ed.) (1990), *Quasicrystals, Networks, and Molecules of Fivefold Symmetry,* New York: VCH.

Hargittai, I., Hargittai, M. (1994), *Symmetry—A Unifying Concept,* Bolinas, CA: Shelter Publ.

Hargittai, I., Hargittai, M. (1995), *Symmetry through the Eyes of a Chemist,* 2nd ed., New York: Plenum Press.

Harris, D. C., Bertolucci, M. D. (1978), *Symmetry and Spectroscopy,* New York: Oxford Univ. Press.

Heilbronner, E., Dunitz, J. D. (1993), *Reflections of Symmetry in Chemistry . . . and Elsewhere,* Basel: Verlag Helvetica Chimica Acta and Weinheim: VCH.

Person, R. G. (1976), *Symmetry Rules for the Chemical Reactions,* New York: Wiley-Interscience.

Woodward, R. B., Hoffmann, R. (1970), *The Conservation of Orbital Symmetry,* Weinheim: Verlag Chemie.

SYNCHROTRON RADIATION

Masaharu Oshima, *The University of Tokyo, Tokyo, Japan*

Introduction 355
1. **Properties of Synchrotron Radiation** 356
2. **Storage Rings** 358
3. **Microfabrication Techniques** 359
3.1 ULSI Fabrication Utilizing Photochemical Reactions 359
3.2 Micromachine Fabrication for Micromechanics 361
4. **SR-Excited Photochemistry for Process Technology** 363
5. **Materials Science and Technology** 365
5.1 Atomic and Molecular Spectroscopy 365
5.2 Photoelectron Spectroscopy of Solids 366
5.3 Extended X-Ray Absorption Fine Structure (EXAFS) and X-Ray Absorption Near-Edge Structure (XANES) 369
5.4 X-Ray Standing Waves 371
5.5 X-Ray Diffraction: Microdiffraction, Surfaces, and High-Pressure Conditions 373
5.6 Topography and Tomography 374
5.7 X-Ray Fluorescence Analysis 376
5.8 Infrared Spectroscopy 377
6. **Medical and Biological Science and Technology** 378
6.1 SR Angiography 379
6.2 Protein Crystallography 380
6.3 Small-Angle X-Ray Scattering 381
6.4 *In Vivo* Soft X-Ray Microscopy 382
6.5 Irradiation Effects on Biology 383
7. **Future Prospects** 384
Acronyms 384
Glossary 385
Works Cited 385
Further Reading 388

INTRODUCTION

Synchrotron radiation (SR) is an outstanding tool for materials science and life sciences. After the usefulness of SR was shown by the early experiments, many efforts have been devoted to develop new technologies for SR storage rings including compact storage rings for industrial use, insertion devices for brighter beams, advanced monochromators for higher resolution, beamlines for higher spatial resolution, and experimental apparatus for versatile applications. SR is extensively used for investigating chemical, biological, geological, and physical systems, and for characterizing new materials.

In materials science and technology, new materials analysis techniques have been developed utilizing unique properties of SR. Extended x-ray absorption fine structure (EXAFS), surface x-ray diffraction, high–energy-resolution photoelectron spectroscopy, and spatially resolved chemical analysis with a microbeam are good examples. These techniques have been employed to characterize trace impurities and surface/interface structures of semiconductors, which dominate the properties of new electronic and optoelectronic devices. Furthermore, SR has been utilized to fabricate structures with 0.1-μm lateral resolution (by x-ray lithography) and atomic-scale vertical resolution (by photochemical reaction). Atomically controlled film growth by means of real-time monitoring with SR is now possible.

In life sciences, SR is regarded as a key tool for analyzing the structures of newly developed small crystals of protein and other giant molecules using an x-ray microbeam. In this field, structure determination used to be very difficult because of the difficulty of

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

making crystals large enough for conventional x-ray diffraction. The structure determination of huge viruses such as the cancer virus SV40 of 500-Å diameter has become possible as a result of the high intensity and collimation at short wavelengths. The structure information is well correlated with medical or chemical properties. Nowadays the medical industry is showing great interest in these techniques. SR-induced reactions have reportedly been utilized to produce new proteins by selective synthesis. Further improvements can be expected when much brighter SR beams from third-generation SR sources are used.

Within the last decade many electronics companies have established their own proprietary beamlines for mainly x-ray lithography and crystallographic analysis of semiconductor materials. On the basis of the success of x-ray lithography, many compact storage rings have been constructed for device fabrication processes, where an SR ring is regarded as an indispensable process apparatus like an ion implanter, chemical vapor deposition (CVD) apparatus, or ion-etching apparatus. Thus, in order to foresee future trends in SR research, the technological prospects should be also better understood.

1. PROPERTIES OF SYNCHROTRON RADIATION

When charged particles are forced to move at speeds close to the speed of light in a circular orbit in vacuum, photons are emitted in the forward direction tangential to the orbit. This radiation is called synchrotron radiation. The radiation from the Crab Nebula has the same origin as SR.

The properties of SR sources are unique, producing radiation with high brightness, a high degree of linear polarization, a high degree of collimation in the forward direction, and pulse widths of about 200 ps and durations ranging from 100 ns to 10 μs. All of these properties enable us to do various experiments in materials science and technology and life sciences. The flux of radiation from storage rings as a function of photon energy can be calculated by applying classical electrodynamics to relativistic electrons on circular orbits. The falloff at high energies is determined by the characteristic energy ϵ_c (or the characteristic wavelength λ_c), which depends on the radius of curvature R, the magnetic field B, and the electron energy E. The relationship among them is as follows:

$$B(\mathrm{T})R(\mathrm{m}) = 3.34E^3(\mathrm{GeV}), \tag{1}$$

$$\begin{aligned}\epsilon_c(\mathrm{eV}) &= 2218E^2(\mathrm{GeV})/R(\mathrm{m})\\ &= 665.1B(\mathrm{T})E^2(\mathrm{GeV}),\end{aligned} \tag{2}$$

$$\lambda_c(\text{Å}) = 5.59R(\mathrm{m})/E^2(\mathrm{GeV}), \tag{3}$$

$$\lambda_c(\text{Å}) = 12\,400/\epsilon_c(\mathrm{eV}). \tag{4}$$

Furthermore, the flux of photons emitted from bending magnets at $\epsilon < \epsilon_c$ is given by

$$\begin{aligned}&I[\mathrm{photons/s\cdot mrad\cdot(0.1\%\ bandwidth)}]\\ &\quad\sim 4.5\times 10^9 R^{1/3}(\mathrm{m})\epsilon^{1/3}(\mathrm{eV})j(\mathrm{mA}),\end{aligned} \tag{5}$$

where ϵ is photon energy and j is the stored current of electrons.

The brilliance, which is another good indicator of storage rings, is defined as the number of photons per second, per 0.1% bandwidth, per unit area A of the cross section of the electron beam, and per solid angle. The horizontal and vertical beam sizes, respectively, are given by $2.35\sigma_x$ and $2.35\sigma_y$. Here σ_x and σ_y are square root of (emittance) × (β function). On the other hand, the beam divergences, σ_x' and σ_y', are almost equal to square root of (emittance)/(β function). Electrostatic and electromagnetic interactions of electrons within the beam usually increase the emittance with increasing stored current. In order to obtain a high-brilliance SR beam, the emittance ϵ_x, that is, the product of σ_x and σ_x', should be reduced.

In addition to the radiation from bending magnets, different types of radiation can be obtained from insertion devices such as wigglers and undulators, which are inserted at the straight sections of the ring. A wiggler, which is a kind of wavelength shifter, is a device in which high-energy photons are produced in low-energy storage rings because of a high magnetic field B. An undulator, which is a kind of multipole wiggler with N magnets, produces N times the photon flux of the bending magnet, because the electrons emit photons with a superposition of contributions from all points on the trajectory. The positions of interference peaks in the undulator radiation spectrum is given by

$$\lambda_m = \lambda_0/2m\gamma^2(1 + K^2/2 + \gamma^2\theta^2), \qquad (6)$$

where m is the harmonic number; λ_0 is the period length of undulator; γ is the electron-energy parameter, which is $1957E$(GeV); K is the undulator parameter, which is $93.4B$(T) λ_0; and θ is the angle at which the radiation is observed. Since the width of the undulator spectral lines is given by

$$\Delta\lambda/\lambda = 2/Nm, \qquad (7)$$

the relative spectral resolution in first order is the reciprocal of the number of periods. Thus, the first harmonic line at 100-Å wavelength with $N = 100$ is semimonochromatic with 2-Å bandwidth. While the normal undulator generates linearly polarized radiation, the helical undulator produces circular polarization, which is of great interest in the investigation of magnetic phenomena. This can be combined with electron-spin analysis.

The more advanced insertion device is a free-electron laser, which can produce a coherent laser beam with tunable wavelength by a combination of the undulator and a resonator consisting of two mirrors. The first free-electron laser experiment was performed using a linear accelerator (linac) in 1976 in the infrared region, and visible laser emission was realized in 1983. Many attempts have been carried out to produce vacuum ultraviolet (VUV) and soft x-ray laser emission using low-emittance storage rings and efficient reflection mirrors. In the near future, specific experiments using coherent x rays and extremely high-brilliance beams will become possible at the fourth-generation storage rings constructed mainly for these free-electron lasers.

By using synchrotron radiation, various kinds of advanced experiments for materials science and life science can be performed. Figure 1 shows various kinds of interactions between SR beams and materials. When the SR beam strikes the sample, various secondary emissions are generated including photoelectrons, Auger electrons, secondary electrons, fluorescence x rays, luminescence, ions desorbed from the surface, reflected x rays, elastic scattering, and inelastic scattering. In addition to these emissions, the SR beam is also absorbed in the sample, and some transmitted x rays emerge from the rear side of the sample.

By utilizing these interactions, we can obtain mainly three kinds of information. First, the atomic or crystal structure can be determined by diffraction and scattering. The structure of all kinds of materials can be analyzed with a high degree of spatial, temporal, and energy resolution. Second, the electronic state or chemical bondings can be determined by photon absorption or emission of photoelectrons and fluorescence; this

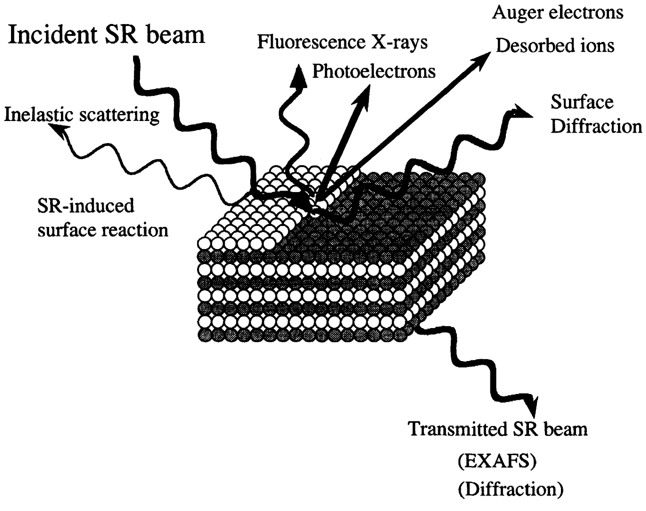

FIG. 1. Secondary emission from materials by interaction with synchrotron radiation beam.

is spectroscopy. Finally, the radiation effect can be intentionally used to change chemical bonds and atomic structure of materials, resulting in the formation of materials with different properties. One of the most important applications of this effect is x-ray lithography for ultralarge-scale integration (ULSI) devices.

2. STORAGE RINGS

The first SR beam was generated in 1947 (Edler *et al.,* 1947), and SR sources themselves have evolved, as shown in Fig. 2. The first-generation SR rings are parasitic SR sources generated in high-energy synchrotron accelerators. The second-generation rings are purpose-built dedicated SR sources with rather high emittance. The third-generation SR rings are brilliance-optimized SR sources with extremely low emittance of electron or positron orbits for insertion devices such as undulators. The fourth-generation SR rings are coherence-driven SR sources that will be suitable for free-electron lasers.

The European Synchrotron Radiation Facility (ESRF) in France, 6 GeV, Advanced Photon Source (APS) in the United States, 7 GeV, and Super Photon ring-8 (SPring-8) in Japan, 8 GeV, are outstanding SR facilities, representing the third-generation rings. These facilities can provide highly brilliant hard x-ray beams for protein crystallography, small-crystal analysis, angiography, and so on.

During the last 30 years, hundreds of SR beamlines and instruments have become available at many SR facilities in the world,

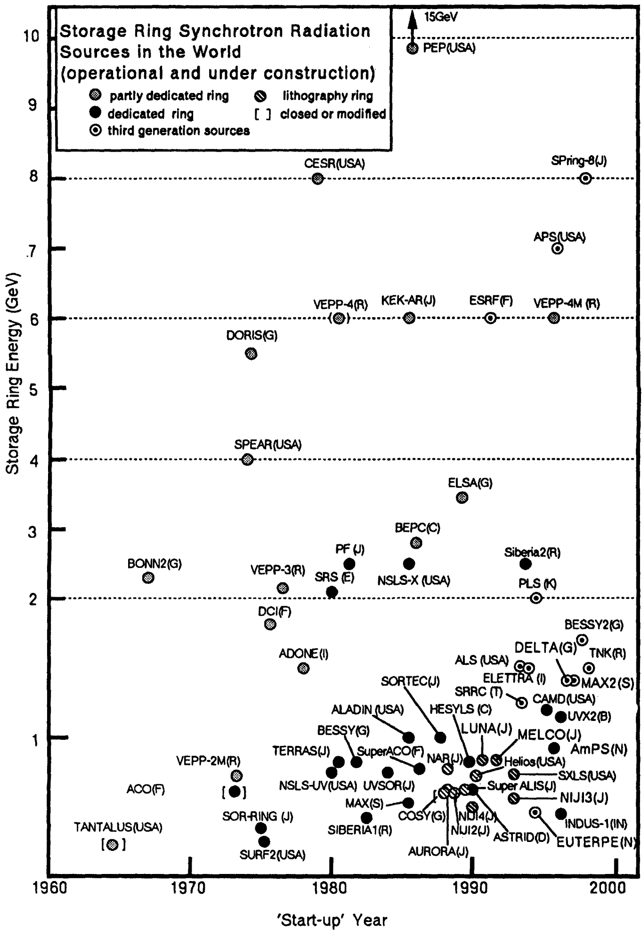

FIG. 2. The evolution of storage-ring synchrotron radiation sources over the decades, as illustrated by their increasing number and range of machine energies (Hasnain *et al.*, 1994).

and many scientists in universities and engineers in industries have recognized the usefulness of SR. At present, there seem to be two approaches in SR science and technology. One is to seek a highly brilliant SR beam as a tool for sophisticated science. This is a way to the third- and fourth-generation rings. The other is to construct more application-oriented SR rings. This is a way to the compact storage ring.

Based on the success of x-ray lithography at medium-sized SR storage rings with electron energies ranging from 2 to 3 GeV, many electronics companies have recognized the compact storage ring as an important process tool mainly for x-ray lithography for ULSI fabrication. The SR wavelength suitable for lithography, which is a kind of photochemical reaction with resist films, is about 7 Å, because photons with about 7-Å wavelength have a suitable penetration depth in resist films such as PMMA (polymethylmethacrylate). Therefore, compact storage rings of less than 800 MeV of electron energy emitting soft x rays and VUV are constructed for industrial applications. These rings are much easier than the medium SR rings from the viewpoint of radiation protection.

The first attempt at constructing a compact storage ring, named Compact Synchrotron (COSY), was made at Berlin Synchrotron Radiation (BESSY) in Germany early in the 1980s. Since there are many electronic industries engaged in dynamic random access memory (DRAM) fabrication, which requires fine patterns of less than 0.2 μm for more than 1-Gbit integration, x-ray lithography is regarded as a key technology. There are now several compact storage rings in Japan, such as Super-ALIS and NAR (NTT), LUNA (Ishikawajima-Harima Heavy Industry), AURORA (Sumitomo Heavy Industry), NIJI-2,3,4 (Sumitomo Electric Co.), Mitsubishi SOR ring (Mitsubishi Electric Co.), SORTEC, and TERAS (Electrotechnical Laboratory, Ministry of International Trade and Industry) (Ohta *et al.*, 1994). Thus, Japan has been dubbed "Synchrotron radiation island." In England, Oxford Instruments made two rings with superconducting magnets (Helios-1 and -2).

Figure 3 shows a photograph of the NTT SOR superconducting storage ring, Super-ALIS, which has ten beamline ports (Hosokawa *et al.*, 1989). The electron energy is 600 MeV with a magnetic field of 3 T, and the stored current reaches about 1.2 A. This ring is mainly utilized for x-ray lithography for ULSI fabrication.

3. MICROFABRICATION TECHNIQUES

3.1 ULSI Fabrication Utilizing Photochemical Reactions

Electronic devices such as Si ULSIs have been developed with an amazing speed following the scale-down rule. Currently 64-

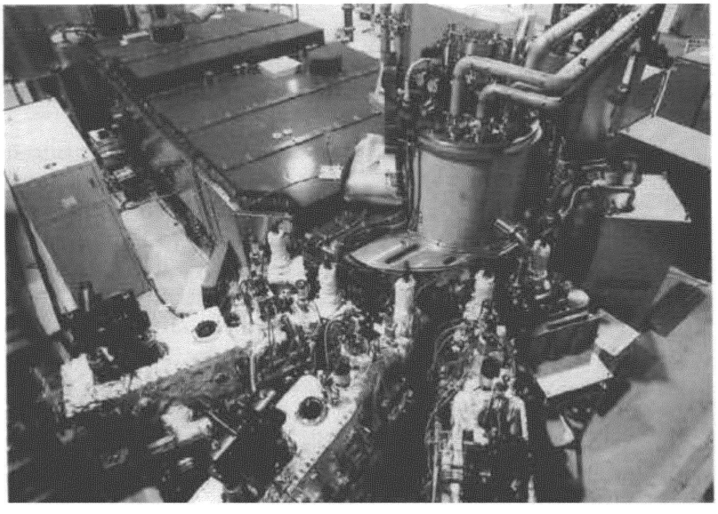

FIG. 3. Photograph of the NTT superconducting storage ring, Super-ALIS.

Mbit DRAMs are being fabricated, and the integration scale is quadrupling every three years, so that 1-Gbit DRAMs should reach commercial production in 1999 (Hayashi, 1994). In order to maintain this speed of development, process technologies providing smaller lateral scale, atomic-scale vertical abruptness, and well-controlled impurity doping are strongly required. For optoelectronic devices, more efficient light-emitting and -detecting devices are required. Figure 4 shows the trend of advanced electronic devices heading toward electron-wave devices. These devices will utilize the quantum size effect, which is being studied in the laboratory in quantum wire or dot lasers and resonant tunneling hot-electron transistors, which require atomically controlled thin-film technology together with an ultrafine lateral pattern delineation process.

In order to improve ULSI performance, there is a strong need to reduce the feature size of circuit patterns. Although various attempts to overcome the limit of optical lithography have been performed, including reducing the wavelength (by using KrF and ArF excimer lasers), increasing the numerical apertures (NA) of projection lenses (NA > 0.6), modifying illumination techniques, and using phase-shift masks, optical lithography is reaching its resolution limit, which is mostly determined by the wavelength and the focusing depth of the light.

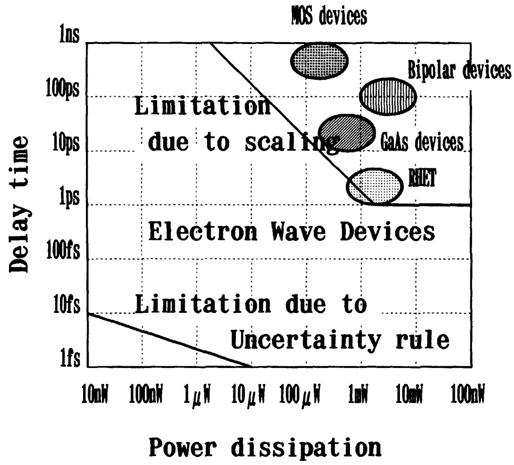

FIG. 4. Trend of advanced electronic devices plotted on a power-dissipation axis and a delay-time axis.

The first proposal for the use of x rays for lithography using a simple 1:1 proximity printing technology was made in 1972 (Spears and Smith, 1972). The x-ray mask was made of a thin membrane transparent to x rays, on which mask patterns were formed with a heavy metal that absorbs x rays. As the transmitted x rays expose the resist, the mask patterns are projected onto the wafer. When the resist films are irradiated by x rays, the chemical bonds are broken, resulting in smaller molecules that are soluble in a developing solution. This is the case for positive-type resist films. In negative-type resist films, on the other hand, x-ray irradiation makes bridges among small molecules, resulting in the formation of fine patterns that are insoluble in a developing solution. Since x-ray lithography uses a much shorter wavelength than that used in optical lithography, the resolution degrading factors such as light diffraction and interference are greatly reduced relative to optical lithography, resulting in a deeper depth of focus and a higher resolution. The resist patterns with high aspect ratio are used as etching masks for the subsequent reactive-ion etching processes to fabricate ULSI components.

Figure 5 shows a schematic view of the SR lithography system at the Super-ALIS beamline in NTT (Ishihara, 1994). In this system, the SR beam is collimated and converted to a parallel beam by two curved mirrors. To maintain the UHV in the SR ring, there is a thin Be window between the beamline and the SR stepper, which is placed in a He atmosphere. For SR lithography, a vertical stepper is used. X-ray masks consist of a finely patterned x-ray absorber film, a thin membrane film (SiN or SiC), and a Si substrate for supporting the patterned film.

Figure 6 shows a scanning electron micrograph of a 0.2-μm line-and-space resist pattern delineated by SR lithography (Deguchi *et al.*, 1992). Among nearly 30 lithography steps in the ULSI fabrication processes, more than ten steps are performed by SR lithography, while the others are done by conventional optical lithography. However, there are several difficulties in proximity x-ray lithography with one-to-one projection, such as the fabrication of x-ray masks and the axis alignment between Si wafers and x-ray masks with patterns smaller than 0.18 μm

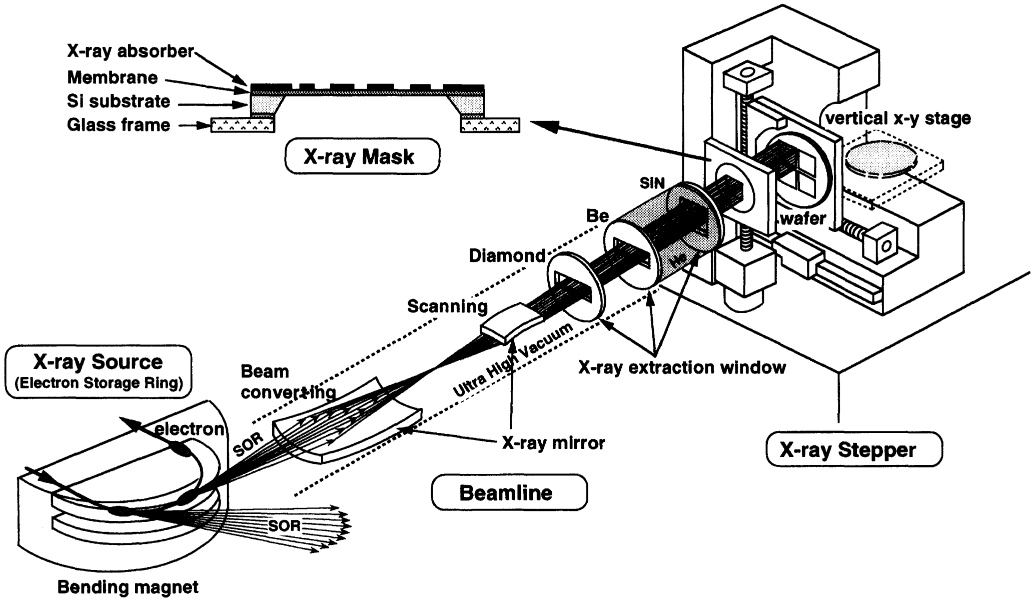
X-ray absorber
Membrane
Si substrate
Glass frame
X-ray Mask
vertical x-y stage
SIN
Be
wafer
Diamond
Scanning
X-ray Source
(Electron Storage Ring)
Beam converting
Ultra High Vacuum
X-ray extraction window
electron
SOR
X-ray mirror
X-ray Stepper
Beamline
SOR
Bending magnet

FIG. 5. Schematic view of the SR lithography system. Note that this figure is not to scale.

for more than 1-Gbit dynamic random access memory (DRAM); refer to the article MICROLITHOGRAPHY.

In order to solve these problems from a different viewpoint, a unique lithography system has been developed. Figure 7 shows a schematic illustration of the two-mirror system for reduction projection lithography taking advantage of the comparably high reflectivities of multilayer-coated spherical or nonspherical mirrors in the VUV region at about 100 eV of photon energy (Kinoshita *et al.*, 1989). Using this kind of projection reduction system, fine patterns of less than 0.1 μm have been successfully delineated (Jewell *et al.*, 1990). Although this system has superior characteristics to the proximity lithography, there are still many problems to be solved, such as the small exposure area at a time, and the low reflectivity of multilayer mirrors in the soft x-ray region at several angstroms, resulting in low throughput.

3.2 Micromachine Fabrication for Micromechanics

One of the new keywords for advanced technologies is "micromachines." Many national and corporate projects have devoted a

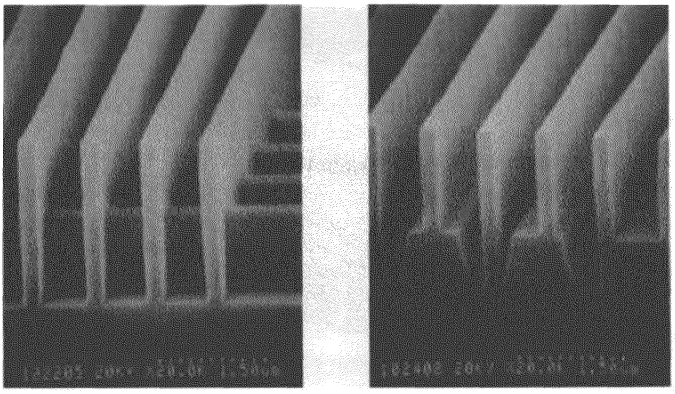

FIG. 6. Scanning electron micrograph of resist patterns of 0.2-μm-wide line and space replicated on Si by SR lithography.

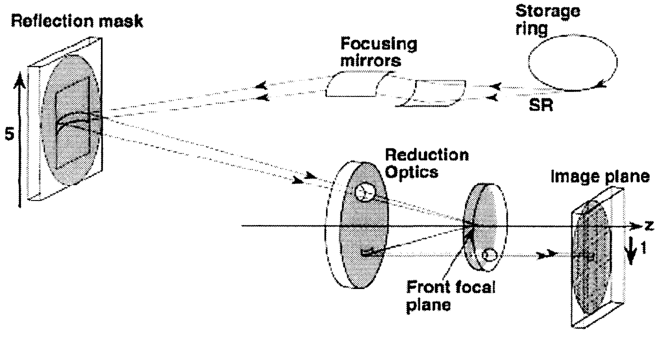
Reflection mask
5
Focusing mirrors
Storage ring
SR
Reduction Optics
Image plane
z
1
Front focal plane

FIG. 7. Schematic illustration of the two-mirror system for reduction projection lithography, which is expected to enable the formation of fine patterns less than 0.1 μm.

lot of effort to fabricate micromachines equipped with microactuators, microsensors, and microprocessors. Sizes at present are a few hundred micrometers.

There are two ways to fabricate these micromachines. One is microprocessing utilizing conventional optical lithography, ion-beam etching or plasma etching for anisotropic or isotropic etching, and CVD film growth techniques, which are well-established for LSI. Many microparts such as microvalves and microactuators have been successfully made this way.

The other approach is the "LIGA" process, which is a German acronym for the three basic processes of deep lithography, electroforming, and plastic molding. This technique was developed at the Karlsruhe Nuclear Research Center (KfK) in Germany to produce very small slotted nozzles for uranium isotope separation. Figure 8 shows a schematic diagram of the LIGA process (Menz *et al.*, 1991). The base material is a layer of plastics, less than 1 mm thick, which is sensitive to x rays. PMMA is used as an x-ray resist, and a mask pattern is transferred to the

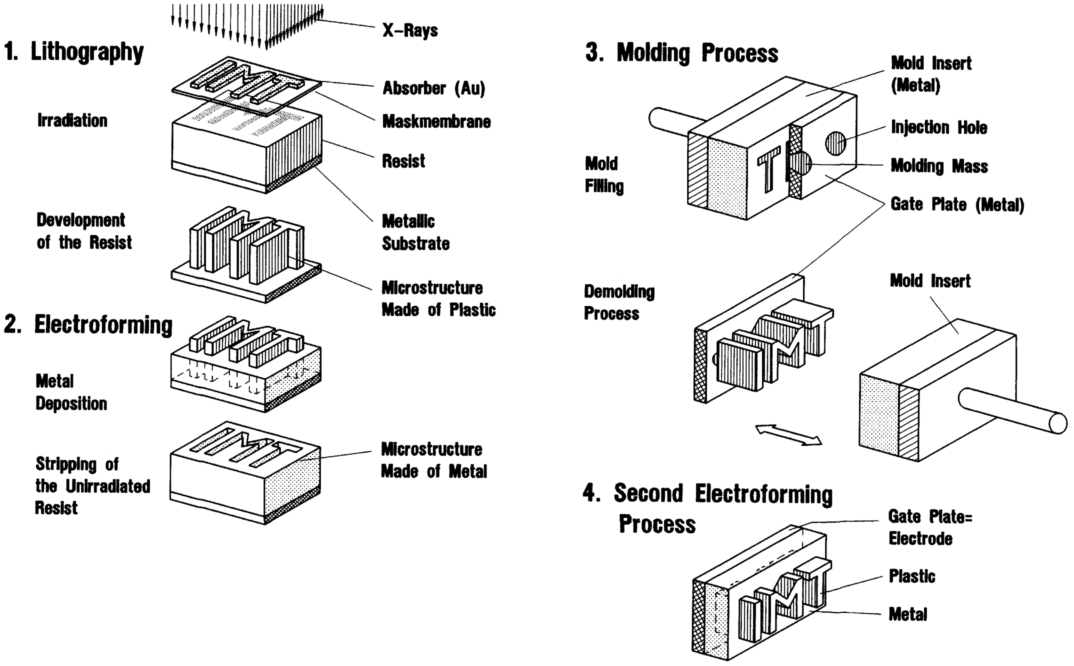
1. Lithography
X-Rays
Absorber (Au)
Irradiation
Maskmembrane
Resist
Development of the Resist
Metallic Substrate
Microstructure Made of Plastic
2. Electroforming
Metal Deposition
Stripping of the Unirradiated Resist
Microstructure Made of Metal
3. Molding Process
Mold Insert (Metal)
Injection Hole
Mold Filling
Molding Mass
Gate Plate (Metal)
Demolding Process
Mold Insert
4. Second Electroforming Process
Gate Plate= Electrode
Plastic
Metal

FIG. 8. Principal steps for the LIGA process for micromachining.

thick resist layer by an x-ray SR beam with a short wavelength of less than 0.2 nm. In SR-irradiated areas, the chemical stability of long molecular chains is changed, and they are removed in a developing solution. This is the same procedure as in the SR lithography for ULSI, except that the wavelength is short enough for deep lithography for thick resist films.

Electrodeposition on a microscopic scale is used to make a complementary structure in metals such as Cu and Ni. The metal pattern is then used to manufacture a number of plastic copies by injection molding. These plastic structures can then be filled up again with metal by a second electrodeposition process. Ceramic microstructures can also be produced. So far this LIGA technique has made various microparts including optical waveguides, microspectrometers, microturbines, acceleration sensors, and micromotors. Figure 9 shows an electrostatic motor fabricated by LIGA. The rotor is 500 μm in diameter and 100 μm high. In order to make the micromotor rotate, the underlying layer should be etched away by using a sacrificial layer that can be dissolved in an etching solution. When 65 V was applied to the stator, that is, an electrode, the rotor started to rotate along the central axis, overcoming the static friction and reaching a rotating speed of 80 rps at about 200 V. This kind of micromachining technique using SR will be extensively applied for micromachine fabrication, which will also open new possibilities in engineering, biology, medical applications, chemistry, and physics; see SYSTEMS TECHNOLOGY, MICROMECHANICAL.

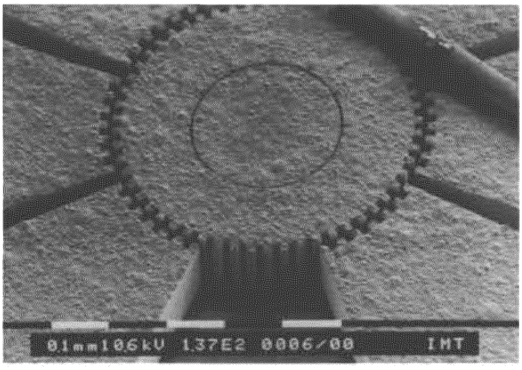

FIG. 9. Electrostatic micromotor fabricated by the LIGA process (courtesy of W. Menz).

4. SR-EXCITED PHOTOCHEMISTRY FOR PROCESS TECHNOLOGY

SR can also be utilized for surface excitation and gas decomposition for film growth of aluminum electrodes and epitaxial silicon for ULSI, and dry etching of silicon-related compound films. The merit of SR over other techniques such as thermal CVD and plasma CVD is that it enables fully selective film growth and much lower-temperature epitaxial growth, which are of importance in forming abrupt interfaces and controlling the impurity doping for fabricating advanced electronic and optoelectronic devices such as quantum-effect devices.

Figure 10 shows a cross-sectional transmission electron microscope (TEM) image of an SR-excited epitaxial Si film grown with a mixture of Si_2H_6 and H_2 at 100 °C after removal of native oxides on Si substrates (Nara *et al.*, 1992). Formation of epitaxial Si films with good quality at such a low temperature suggests the possibility of strongly suppressing interdiffusion at the interface, which is required to make abrupt interfaces on an atomic-layer scale.

The second important application is fully selective film growth, which may drastically reduce the number of fabrication process steps because of its directly patterned film growth. Figure 11 shows a schematic diagram for SR-excited–patterning aluminum CVD processes and patterned Al (Uesugi and

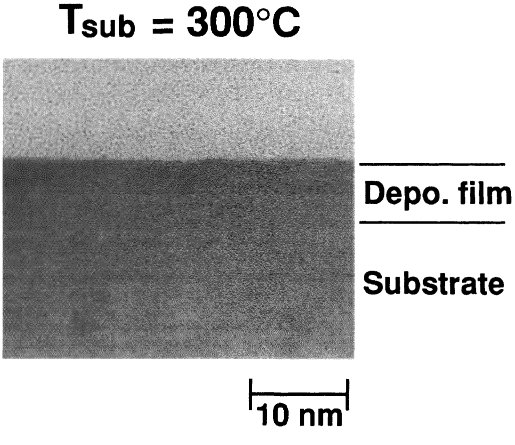

FIG. 10. TEM image of a SR-excited epitaxial Si film grown with a mixture of Si_2H_6 + H_2 at 300 °C after removal of Si native oxide at 300 °C.

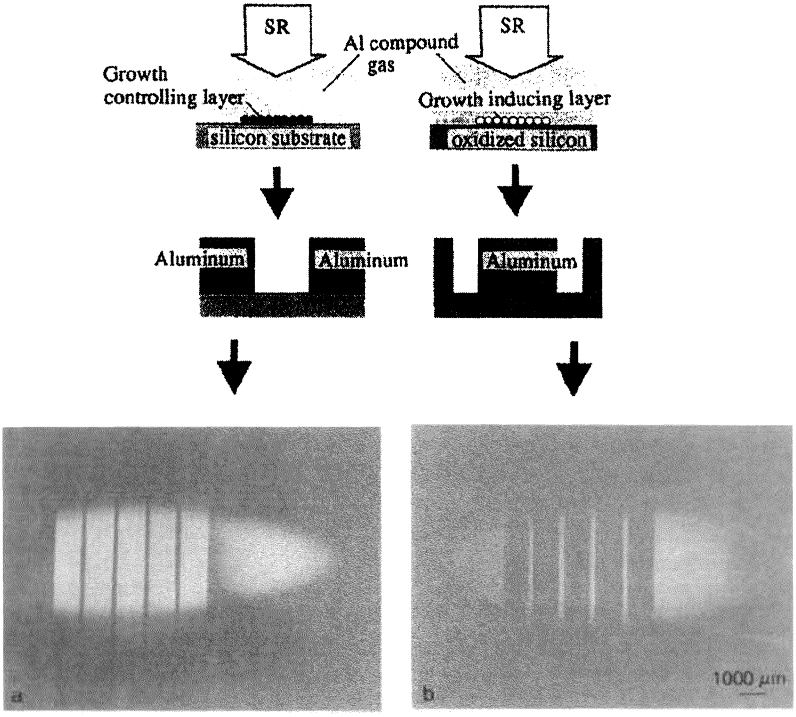

FIG. 11. Normal and inverse-projection patterning aluminum CVD by SR-induced surface photomodification.

Nishiyama, 1992). This is a kind of SR-induced surface photomodification. When SR is directed onto a Si surface that is exposed to a dimethyl-aluminum hydride (DMAH) gas, the Al compound gas containing carbon is decomposed, and an AlC-like growth-controlling layer is formed on Si, resulting in no further Al film growth. Therefore, this is called SR-excited patterning. In contrast, since such a carbide film is not formed on a Si oxide, the Al pattern is directly formed only on the SR-irradiated region of Si oxide preferentially. Thus, selective Al electrode formation is achieved with the SR-excited CVD. This leads to positive and negative direct-patterning processes by controlling the substrate surface properties.

Similar process technology was developed to grow Si epitaxial films on a Si substrate completely selectively with SR-excited CVD (Utsumi *et al.*, 1993). Here SR is directed onto a wide area where a SiO_2 film covers the Si substrate with finely patterned windows. This technique takes advantage of the property that the SR-decomposed Si_2H_6 gas fraction does not adsorb on the SiO_2 surface at all, and the epitaxial Si layer grows only on the patterned window region where the clean Si surface is exposed to the gas fraction.

The third application is atomically controlled film growth, or atomic-layer epitaxy (ALE). In the case of GaAs, the ALE is achieved only by a thermal process because adsorbed $Ga\text{-}CH_3$ fragments reject the subsequent TMG molecule adsorption. However, it was impossible to grow Si epitaxially in the ALE mode by conventional thermal CVD. Recently, ALE-mode Si growth on Ge(100) resulting from surface photochemical reaction was realized using alternating disilane (Si_2H_6) exposure and SR irradiation at a low substrate temperature of 350 °C (Akazawa *et al.*, 1993). The ALE process consists of

1. adsorption of Si_2H_6 on Ge for about 20 s from a pulse nozzle,
2. pumping the residual gas, and
3. SR irradiation for desorbing hydrogen atoms from the surface for about 200 s.

Although complete ALE growth was not achieved, this kind of surface reaction utilizing SR photochemical reaction will be useful in practice.

The mechanism of SR-excited CVD is believed to be as follows: First, SR excites core electrons in Si 2*p* or Al 2*p* core levels in Si_2H_6 or $Al(CH_3)_2H$ gases. Then, through the Auger de-excitation process such as Si_{LVV} or Al_{LVV}, two holes are created in the Si or Al valence band. Then Si–Si bonds or Al–C bonds are broken by repulsion force between Si^+–Si^+ or Al^+–C^+. Thus SR can effectively decompose metal compound gas. However, more precise analysis such as time-of-flight (TOF) analysis of desorbed ions or neutrals in terms of mass, angle distribution, and kinetic energy by changing photon energy will be needed to understand fully the SR photochemical processes.

Besides these CVD processes, the usefulness of SR etching has been also shown for Si, SiO_2, SiN, SiC, Ta, Ti, W, WO_3, TiO_2, GaAs, and diamond. Thus unique surface reactions that have been impossible by thermal process are possible through SR photochemical processes.

5. MATERIALS SCIENCE AND TECHNOLOGY

5.1 Atomic and Molecular Spectroscopy

Since the mass of the photon is zero, excitation due to momentum transfer does not happen in photon-induced excitation processes. Thus, the SR, having wavelength tunability, is an ideal probe to resonantly excite target atoms and molecules. Furthermore, the polarization property is very useful for analyzing dynamics of photoexcitation and photoionization. Various kinds of knowledge on photoninduced elemental processes of atoms and molecules are important for applications such as photo CVD, photochemical reaction, plasma physics, astrophysics, and the development of lasers.

The measurement system of atomic and molecular spectroscopy is rather simple. A monochromatic SR beam is directed onto a gaseous sample, and photoabsorption, photoelectrons, and photoejected ions are measured as a function of photon energy. For these experiments, noble gases and evaporated metal atoms are mainly used to investigate electron–electron correlation, which is one of the most important subjects in atomic physics.

The pioneering works on atomic and molecular spectroscopy using SR were done by Madden and Codling (1963) on observation of two-electron excitation in He atoms, and by Ederer (1964) on measurement of the 4*d* → ϵf giant resonance in Xe atoms.

The second important discovery was the 4*d* giant resonance of Ba, Ba^{+1}, and Ba^{+2} measured by the double-laser technique (Lucatorto *et al.*, 1981), showing that the shape of the 4*d* giant resonance strongly depends on the number of 6*s* electrons in the outermost shell. This kind of experiment is expected to extend the research field vastly to understand the phenomenon, which includes not only simple atoms and molecules in the ground states, but also ionized and excited atoms and molecules (Koizumi *et al.*, 1995). A new insight on electron–electron correlation will be obtained. However, more sensitive detectors together with much brighter SR beams are needed for ions, which have a very low density such as 10^7 ions/cm^3 because of space-charge limitation.

The third example is the measurement of the photoabsorption cross section of CO molecules. Stark *et al.* (1992) systematically measured photoabsorption cross sections of CO under the condition that the CO rotation temperature was frozen, by ejecting CO molecules from a supersonic molecular beam source to a vacuum chamber. In order to enable the precise measurement of density of CO molecules in space from the Hubble space telescope and so on, the SR photoabsorption measurement in a circumstance similar to space is necessary.

The fourth example is a new spectroscopy for inner-shell–excited molecules. Yagishita *et al.* (1989) developed a new method for the soft x-ray molecular spectroscopy of gaseous molecules by separating symmetry of inner-shell–excited states. When the photon energy is set at the $\sigma \rightarrow \sigma^*$ resonance, only the molecules that are oriented parallel to the SR electric vector are strongly excited, and the dissociated ions are emitted toward the molecular axis. Thus, the measurement of dissociated ions at $\theta = 0°$ extracts only the $\sigma \rightarrow \sigma^*$ transition, and that at $\theta = 90°$ extracts only the $\sigma \rightarrow \pi^*$ transition, as shown in Fig. 12. This "symmetry separation spectroscopy" can easily separate Rydberg states located in

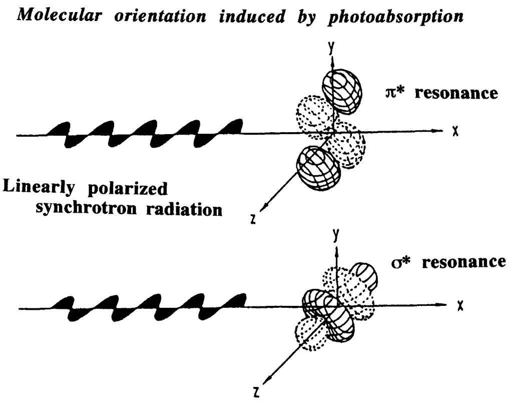
Molecular orientation induced by photoabsorption
y
π* resonance
x
Linearly polarized synchrotron radiation
z
y
σ* resonance
x
z

FIG. 12. Principle of excitation processes for $\sigma \rightarrow \pi^*$ resonance and $\sigma \rightarrow \sigma^*$ resonance. The electric vector of the incident SR beam is in the *x-z* plane. In the case of $\sigma \rightarrow \pi^*$ resonant absorption, the molecules that are aligned perpendicular to the electric vector are strongly excited, while those parallel to the electric vector are strongly excited in the $\sigma \rightarrow \sigma^*$ resonant absorption. (Yagishita *et al.*, 1989.)

many excited states and completely discriminate the σ symmetry from the π symmetry.

Furthermore, they succeeded in measuring angular distributions of photoelectrons from molecules by using a photoelectron-photodissociated ion coincidence technique (Shigemasa *et al.*, 1995). For molecule-adsorbed surfaces, the polarization vector of SR is very useful to determine the bonding direction of adsorbates. However, since gaseous molecules are rotating in all directions, only an averaged picture of the emitted photoelectron angular distribution can be obtained. If it is possible to detect photoelectrons of molecules by an angle-resolved analyzer only when dissociated ions are detected along the direction parallel to the SR electric vector, the precise angular distribution of the emitted photoelectrons can be obtained. They measured the angular distribution of N 1*s* photoelectrons from N_2 molecules at the photon energy of 419 eV, which corresponds to the shape resonance of N_2 molecules. They found that the photoelectron distribution in the parallel condition corresponds to the $f\sigma$ component, while that in the vertical condition corresponds to the $d\pi$ component.

Thus, fundamental understandings of atomic and molecular physics in noble gases are steadily proceeding, taking advantage of the linear polarization, wavelength tunability, and high photon flux of the SR beam together with a new coincidence measurement system. With further improvement of SR beam brilliance and new detectors, atomic and molecular physics in more complicated gases will be understood.

5.2 Photoelectron Spectroscopy of Solids

Photoelectron spectroscopy (PES) is one of the most widely used surface-analysis techniques (see PHOTOEMISSION AND PHOTOELECTRON SPECTRA). PES can provide various types of information about surface composition, chemical bonds, electronic properties, and surface atomic structures. One of the major advantages of PES with SR is the very surface-sensitive analysis because the photon energy is tunable to obtain the shortest escape depth of photoelectrons (Lindau and Spicer, 1974). To make use of this advantage over the conventional x-ray photoelectron spectroscopy, *in situ* analysis is required. *In situ* surface analysis consists of growing epitaxial layers on substrates, transferring the specimen to the analysis chamber in UHV, and analyzing the surface.

With the use of high–energy-resolution grazing-incidence monochromators together with a high-efficiency hemispherical electron analyzer, it became possible to discriminate surface Si atoms showing asymmetric dimers from the second-layer Si atoms and the bulk Si atoms on Si(100) 2 × 1 and *c*(4 × 2) reconstructed surfaces (Landemark *et al.*, 1992).

Various kinds of interfaces in metal/semiconductor, insulator/semiconductor, and semiconductor/semiconductor systems have been analyzed by *in situ* PES (Brillson, 1993). These PES data are compared to electrical properties such as Schottky-barrier heights (SBHs) of metal-semiconductor contacts. The dominant works by SR-PES have been done in metal deposition on clean Si and GaAs that were cleaved in vacuum for understanding Schottky-barrier formation and controlling SBHs.

For insulator-semiconductor interfaces, oxidation processes have been widely analyzed. Initial stages of oxidation of silicon substrates have been investigated vigorously with SR-PES because the initial oxidation features are technologically important for fabricating very thin gate oxides of less than 50 Å for ULSI. Figure 13 shows Si 2*p* photoelectron spectra taken from Si(100) and (111) surfaces with 5-Å SiO_2 (Himpsel *et al.*, 1988). The experimental results that the Si(111) surface has larger Si^+ and Si^{3+} components than the Si(100) surface, which has a larger Si^{2+} component in return, are explained by the Si(111) surface structure having one dangling bond and three dangling bonds, while Si(100) has two dangling bonds.

Various SiO_2 thin films grown by different dry and wet processes using HNO_3, H_2SO_4, H_2O_2, HCl, and other solutions have been analyzed by high-resolution PES to correlate the interfacial chemical bonds with the electrical properties. Although it is well known that the SiO_2 film passivates the Si surface dangling bonds, there are high-density surface states at the oxide/GaAs interface, which strongly hinder the fabrication of GaAs MISFETs (metal-insulator-semiconductor field-effect transistors) with good quality. SR-PES analysis revealed that there exist elemental As atoms at the interface, playing a role of mobile charge or surface states (Pianetta *et al.*, 1978).

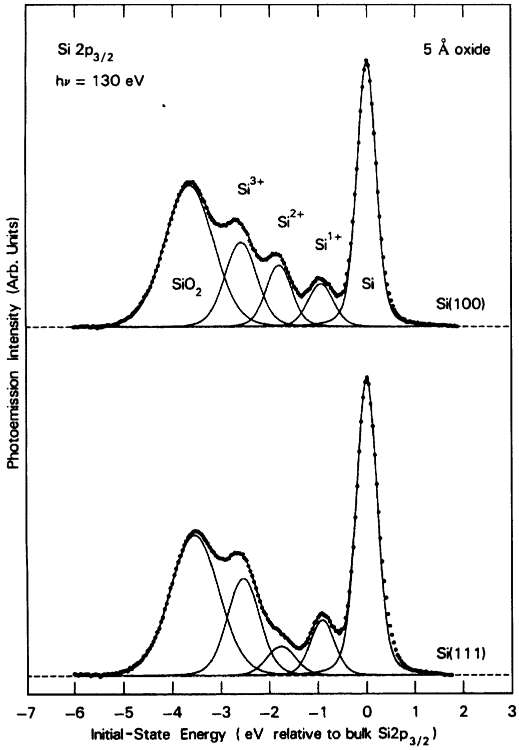

FIG. 13. Photoelectron spectra of very thin silicon oxide films on Si(100) and (111) surfaces measured with high-energy resolution.

CaF_2 is a promising candidate for an epitaxial insulator film grown on a GaAs substrate with a low interfacial state density. The analysis of interfacial chemical states by SR-PES and surface morphology by RHEED revealed that when CaF_2 is grown on GaAs(111)*B*, a layer-by-layer growth with an abrupt interface is obtained, while CaF_2 on GaAs(100) has decomposed Ca^0 at the interface (Yamada *et al.*, 1988). On the basis of these SR-PES results, for MIS diode applications, a perfectly lattice-matched $(Ca_xSr_{1-x})F_2$ epitaxial insulator was grown on GaAs(111)*B*, which showed a very low interface state density of $1 \times 10^{11}\ \mathrm{cm}^{-2}\ \mathrm{eV}^{-1}$ (Waho and Seki, 1991). This is a good example for the usefulness of *in situ* SR-PES for device applications.

With the drastically improved energy resolution in PES, various types of information about surface and bulk properties can be obtained in addition to chemical components and bondings. Photoelectrons contain chemical and electronic information about solids. By analyzing photoelectrons with a high–angular-resolution analyzer, band structure can be obtained. One example is the direct determination of the superconducting gap energy of high-T_c oxide superconductors. Using extremely high–energy-resolution angle-resolved PES, the gap energy was determined to be 24 meV for a $Bi_2Sr_2CaCu_2O_8$ sample using a 22-meV energy-resolution setup (Olson *et al.*, 1989). This result supports the Fermi-liquid model for the superconducting mechanism. The other example of high–energy-resolution PES is the direct observation of the quantum size effect of nanocrystals. Quantum confinement energy was determined to be several tens of meV for 24-Å CdS nanocrystals (Colvin *et al.*, 1991). In the angle-resolved PES, photoelectrons are analyzed in terms of their emission (azimuthal and polar) angle and their kinetic energies. The kinetic energy gives the binding energy as the energy difference from the Fermi energy. The internal wave vector of the photoelectrons must also be determined to test the calculated electron-energy dispersion. Figure 14 shows angle-resolved photoelectron spectra measured from a cleaved GaAs(110) surface, from which band mapping showing good agreement with the pseudopotential theory is achieved (Chiang *et al.*, 1979). This technique has been applied to epitaxial GaAs(100) layers, HF-treated Si surfaces, alkali-metal–deposited Si surfaces, and so on.

Besides these chemical and electronic state analyses, the photoelectron diffraction technique, which is a kind of angle-resolved PES, can give unique information about surface structures of adsorbates or the surface itself. By using a two-dimensional electron analyzer (Daimon, 1988), a 180° distribution of photoelectrons can be displayed on the screen. This photoelectron diffraction pattern can be regarded as a hologram in *k* space. Therefore, by analyzing the hologram by computational Fourier transformation, a photoelectron holograph in real space can be obtained (Barton, 1988). Figure 15 shows three photoelectron diffraction patterns of Si 2*p* obtained from a Si(111) 7 × 7 reconstructed surface by using incident SR beams of different energies, and a Fourier-transformed photoelectron holographic image (Daimon, 1993). The cross points correspond to the neighboring Si atom positions.

Furthermore, the microanalysis of surfaces and interfaces became possible, taking advantage of the low divergence of SR. A very small SR beam spot of about 0.1 μm has been achieved for microphotoelectron spectroscopy (Williams *et al.*, 1995). Figure 16 shows a schematic diagram of a micro-beam PES system with an undulator, a monochromator, a Schwarzschild objective lens, and a sample scanning stage for two-dimensional imaging (Ng *et al.*, 1990). Two-dimensional compositional and chemical-state

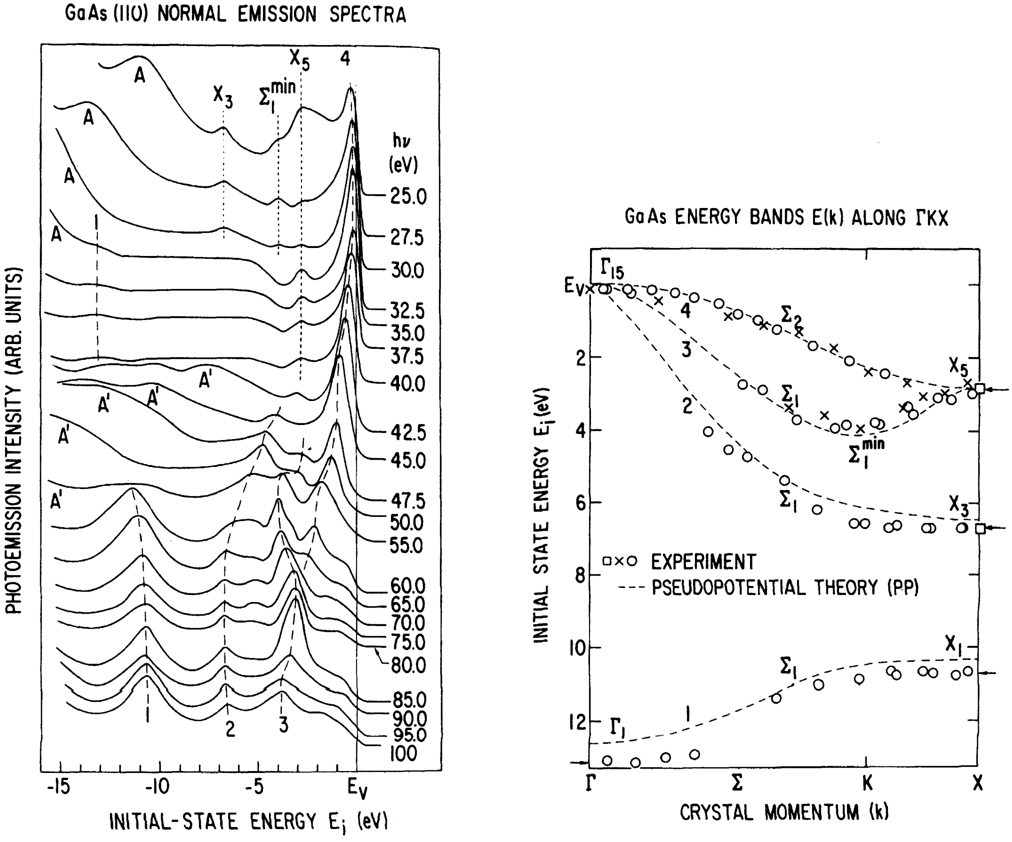

FIG. 14. Angle-resolved photoelectron spectra taken from a cleaved GaAs(110) surface and the derived band structure.

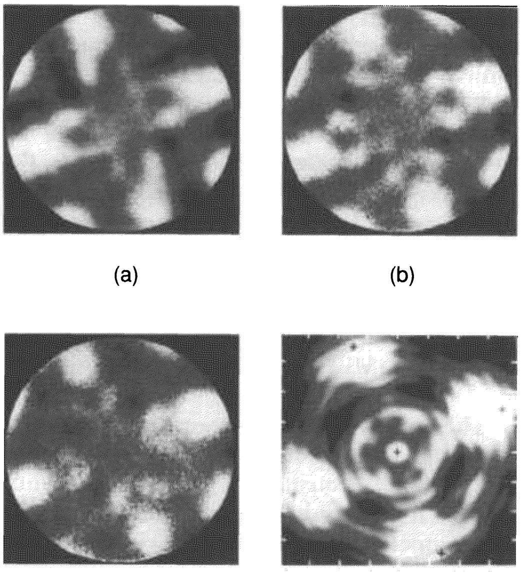
(a) (b) (c) (d)

FIG. 15. (a)–(c) Si 2*p* photoelectron diffraction patterns obtained at the photoelectron kinetic energies of 350, 400, and 650 eV, respectively, by using a two-dimensional photoelectron analyzer. (d) Photoelectron holography of Si(100) surface calculated from (a)–(d).

mapping was achieved, which should be useful for in-process diagnosis of microdevices during fabrication processes.

In the near future, SR-PES will be further improved by higher resolutions in energy (several meV), in lateral dimension (several tens of nanometers), in angle (less than 1 degree), and in time (several milliseconds). Third-generation SR sources with high brilliance will be required in order to realize these high-resolution PESs.

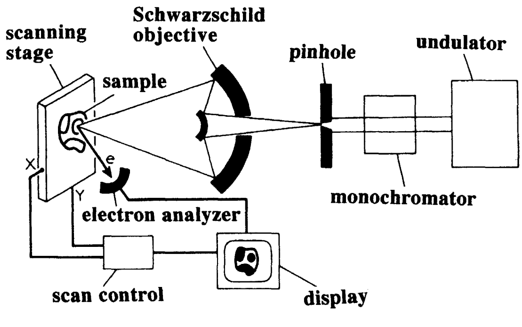

FIG. 16. Schematic diagram of a photoelectron microscope using a SR microbeam from an undulator source.

5.3 Extended X-Ray Absorption Fine Structure (EXAFS) and X-Ray Absorption Near-Edge Structure (XANES)

The EXAFS technique (Teo and Joy, 1986) provides information about the atomic structure in the immediate vicinity of an x-ray–absorbing atom. The photoelectron emitted after absorption can be regarded as a wave. This wave interacts with the atoms surrounding the absorbing atom, resulting in a modulation of the absorption spectrum at energies higher than the absorption edge. Deconvolution of these modulations provides information about the number, types, and distances of neighbors. Therefore, EXAFS is complementary to diffraction, which naturally reflects the translation symmetry of the material. Thus, EXAFS can be applied to materials of any phase, crystalline and noncrystalline solids, solutions, or gases.

One of the classes of materials most widely analyzed by EXAFS is catalysts. It is well known that catalytic effects strongly depend on the structure of the catalyst. For dispersed metal catalysts consisting of small metal clusters dispersed on the surface of silica or alumina (carrier), the interaction of oxygen with 1% Ru on a silica catalyst was studied *in situ,* and it was found that bulk oxidation occurred and destroyed the Ru lattice (Lytle *et al.,* 1977). Other catalysts such as metal oxide catalysts and homogeneous catalysts were also successfully analyzed by EXAFS, where reactions such as polymerization and hydrogenation processes were well correlated with structure changes. Such knowledge will increase our understanding of catalytic reactions, leading to significant improvements in catalysts.

Metal cluster particles in colloidal dispersions as well as small metal particles in supported catalysts were analyzed by EXAFS to determine the local structure. As an example for the characterization of the exact structure of the bimetallic clusters in supported metal catalysts, Sinfelt (1987) characterized Ir/Rh, Au/Cu, Ru/Cu, and Os/Cu bimetallic clusters supported on silica by means of EXAFS and found that the surface can consist entirely of Cu in the form of a monolayer and that the core consists of ruthenium.

With the increased x-ray flux available from a SR source, the range of target materials is expanding from the conventional bulk materials to targets containing many fewer atoms. One of the important materials in semiconductor devices is impurity atoms implanted into Si substrates. The local geometry or environment of impurities incorporated in Si substrates by ion implantation was analyzed by EXAFS. Figure 17 shows EXAFS spectra for the As-implanted Si(100) substrate with a dose of 3 × 10^{16} cm^{-2} at 100 keV followed by laser annealing. The Si substrate was then annealed at 350 °C and higher temperatures. It is found that the subsequent annealing changed the As–Si distance from 2.41 to 2.37 Å and the coordination number from 4 to about 3. From these results, a model was developed based on As-rich vacancy complexes such as As_4V at 350 °C and SiAs precipitates at higher temperature. The structure information was well associated with the marked decrease in the carrier concentration in Si during the annealing process (Allain *et al.*, 1993).

Surfaces and interfaces are also important targets of EXAFS where the number of target atoms is quite small compared to bulk atoms. The structure of Ge epitaxial overlayers on Si(001) has been studied by the *in situ* surface-sensitive EXAFS technique after growth by MBE. The oscillatory intensities of RHEED during the growth of Ge overlayers were utilized to control the number of Ge layers. The Ge EXAFS spectra for Ge overlayers on Si(001) have been obtained by detecting the Ge fluorescence signal in a total-reflection geometry where only the subsurface (50 Å) region is probed. Fourier-transform analysis of EXAFS oscillations revealed the expanded dimer structure of 1 monolayer (ML) Ge on Si(001) (Oyanagi *et al.*, 1994).

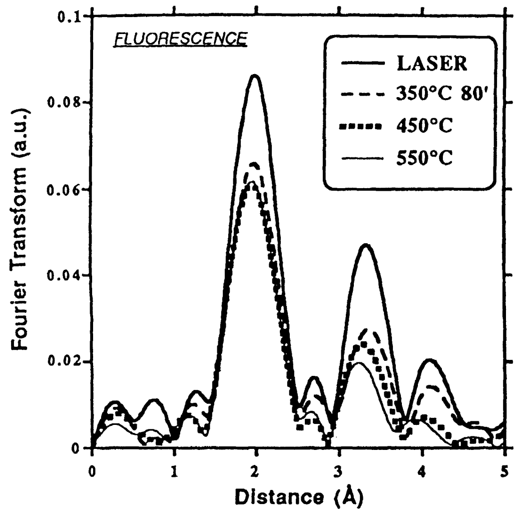

FIG. 17. EXAFS analysis of As-implanted and laser-annealed Si substrates.

The strained Ge–Si system is expected to be useful for optoelectronic devices because of its direct transition-type band structure and for high-speed electronic devices such as heterojunction bipolar transistors. By analyzing a 340-Å pseudomorphic Ge–Si film by EXAFS and x-ray diffraction for the short-range order and the long-range order, respectively, the Ge–Ge and Ge–Si first-neighbor bond lengths of 2.44 and 2.38 Å were found to coexist. Furthermore, the pseudomorphic epitaxy produces an expansion of the Ge_xSi_{1-x} lattice spacing perpendicular to the interface and a contraction parallel to the interface (Woicik *et al.*, 1991). This kind of interface analysis was also conducted for InGaAs/GaAs multiquantum wells, ZnSe/CdSe superlattices, SiO_2/GaSe/GaAs MIS structures, lattice-matched InGaAs/InP quantum-well heterostructures, and Si/GaAs interfaces.

Initial oxidation features of GaAs surfaces were analyzed by surface EXAFS utilizing secondary electrons (Stohr *et al.*, 1979). From the analysis of GaAs oxide of 1 monolayer (ML), which was formed by exposure to 5 × 10^9 Langmuir (1 Langmuir = 10^{-6} Torr-s) of oxygen, the As–O distance was determined. Furthermore, the O–As and O–Ga distances on the plasma-oxidized GaAs surface were determined for MIS device application (Oshima *et al.*, 1988). In order to analyze surface and interface structures, a gas-scintillation proportional counter (GSPC) was developed for effectively detecting soft x rays from light elements such as oxygen and carbon (Maeyama *et al.*, 1991). This detector was also applied to structure analysis of the surface sulfur layer, which is reported to play an important role of atomic-layer passivation for GaAs MESFETs. This passivation was performed by etching the GaAs substrate with an $(NH_4)_2S_x$ solution and annealing in vacuum, resulting in 1–2 monolayers of Ga–S bonds. An x-ray standing-wave (XSW) study of this system also determined the Ga–S atomic distance to be 2.30 Å, assuming no surface relaxation. Since the Ga–S distance determined by surface EXAFS was also 2.31 Å, it can be concluded that there is almost no surface relaxation caused by sulfur passivation (Maeyama *et al.*, 1995).

More spatially resolved information can be obtained by microbeam EXAFS. Using a Kirkpatrick–Baez mirror configuration as an x-ray focusing element, the x-ray beam size on the sample was reduced to around 5 μm with a photon flux of 10^7–10^8 photons/(s 300 mA) (Iida *et al.*, 1993). With the use of a highly brilliant SR beam from the third-generation rings, the microbeam EXAFS will be extensively applied to various electronic, chemical, and biological materials.

Finally, a unique application of x-ray absorption spectroscopy is reviewed—that is, the element-specific magnetic microscopy with circularly polarized x rays. The magnetic contrast arises from the fact that the x-ray absorption cross section at inner-shell absorption edges of aligned magnetic atoms depends on the relative orientation of the photon spin and the local magnetization direction. It is reported that magnetic imaging of a CoPtCr magnetic recording disk with a lateral resolution of 1 μm was obtained using the magnetic circular X-ray dichroism (MCXD) technique (Stohr *et al.*, 1993). The monochromatic soft x rays were incident at an angle of 30° from the surface, and the secondary electrons were accelerated into the microscope along the surface normal. Secondary electrons of about 6-eV kinetic energy were focused into an image using a two-stage electrostatic lens system. The pattern obtained at the photon energy of 778 eV, that is, the Co L_3 edge, shows the same image as the pattern at 793 eV, that is, the Co L_2 edge, but with reversed contrast. Thus, the magnetic contrast results from the alignment between photon spin and magnetization vectors. The feasibility of magnetic microscopy by using undulator insertion devices (factor of 100 in flux) and x-ray focusing optics (factor of 100 in spot size) was estimated, resulting in 100×100-Å^2 MCXD microscopy. This kind of magnetic microscopy with different photon energies may enable element specificity, chemical-state specificity, and variable probing depth. Furthermore, by combining the MCXD microscopy with NEXAFS and EXAFS, one can obtain magnetic, chemical, and even structural information on the same microscopic area.

5.4 X-Ray Standing Waves

While scanning tunneling microscopy (STM) is a useful tool for obtaining detailed information about the lateral distribution of surface atoms with atomic-scale resolution, it does not readily provide information on distances perpendicular to the surface. This kind of information can be obtained from surface EXAFS and x-ray standing waves (XSW). Therefore, these techniques are the ideal complement to STM studies. Since XSW is simpler than surface EXAFS in several ways such as simple mathematical analysis, much narrower photon energy range for energy scanning, and three-dimensional position determination (triangulation), XSW is becoming more popular for structure analysis.

When monochromatic x rays impinge onto crystals such as Si and GaAs having perfect crystallinity, the incident and the diffracted x-ray beams interfere to produce a standing wave with nodal and antinodal planes that are parallel to the diffraction plane. If the target atoms on the surface are located at the net plane, which in principle corresponds to the lattice plane as shown in Fig. 18, the electrons surrounding the atoms are strongly excited when the antinodes of the standing waves exist at the net plane, that is, condition 3 in Fig. 18, corresponding to the high-angle side of the Bragg condition. On the other hand, when the nodes exist at the net plane, that is, condition 1, the atoms are not excited at all in principle. Changing the x-ray incident angle causes the standing-wave nodes to move continuously from the diffraction plane to between the planes. By monitoring the secondary-emission yields from the target atoms, such as fluorescent x rays, Auger electrons, or photoelectrons, the position of these atoms relative to the diffraction planes can be determined. The analysis of the profile gives the exact position of the target atoms regardless of whether they

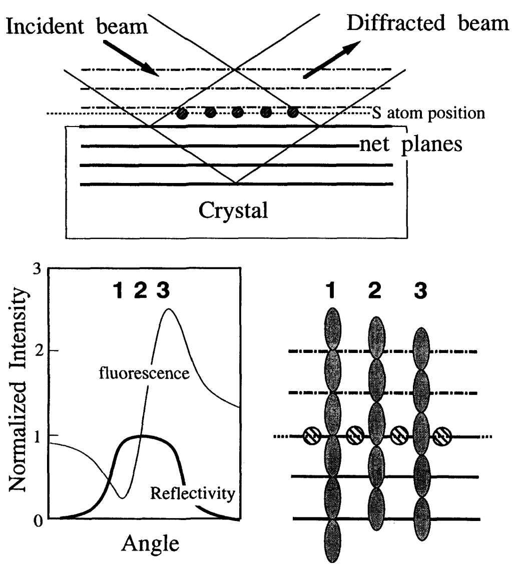

FIG. 18. Principle of x-ray standing wave (XSW) analysis.

exist at the surface, in the overlayer, or at the substituting atomic position in the crystals.

After the theoretical formulation of XSW analysis with the x-ray dynamical diffraction theory was established, XSW has been extensively applied for semiconductor surfaces and impurities such as As and Bi atoms doped into Si. In the 1980s many studies were performed using synchrotron radiation. With regard to chemisorption on Si, structure analysis was done on Br adsorbed on the cleaved Si(111) surface in UHV (Dev *et al.,* 1985). Submonolayer Ge (Patel *et al.,* 1985) and Au (Berman *et al.,* 1988) on Si(111) were also analyzed.

Heteroepitaxial overlayers such as $NiSi_2$/Si(111) (Vlieg *et al.,* 1986), CaF_2/Si(111), and GaAs/Si(111) (Kawamura *et al.,* 1991) are very important systems for device applications such as Schottky-barrier diodes, MIS devices, and solar cells. Since their electrical properties are sensitive to the interfacial structure and bonds, precise analysis of interfaces is required. XSW analysis of $NiSi_2$/Si(111) reveals that both A-type (untwinned) and B-type (twinned) $NiSi_2$ films epitaxially grown on Si(111) substrates have the same sevenfold coordinated interface, but the A type has larger inward relaxation (0.16 Å) than the B type (0.07 Å). This 0.09-Å difference in interface relaxation is thought to be responsible for the difference in Schottky-barrier heights observed for A- and B-type epitaxial films. Thus the fundamental electrical properties are reflected in interfacial structures. Furthermore, initial stages of epitaxial overgrowth such as GaAs on Si and Sb compounds on InP(110) (Kendelewicz *et al.,* 1991) were precisely investigated by XSW.

In order to fabricate GaAs MISFETs with a higher operating speed, the defect density at the interface should be drastically decreased. An epitaxial CaF_2 film with good interface chemistry was grown on sulfur-passivated GaAs (Scimeca *et al.,* 1991). In order to elucidate the sulfur passivation mechanism, the S atomic position was determined three-dimensionally by means of XSW triangulation (Sugiyama *et al.,* 1993). Plate 1 shows XSW profiles of CaF_2/S/GaAs(111)*B* and a structure model for the interfacial sulfur atoms constructed on the basis of combined photoelectron spectroscopy and XSW results (Oshima *et al.,* 1993). By analyzing the XSW profile, the S position was determined to be at the hollow site having three bonds with Ga, substituting for an As. The thermodynamically stable S–Ga bonds play a key role of drastically reducing GaAs surface defects. Thus the interfacial structure (obtained by XSW) together with the chemical bondings (obtained by SRPES) are successfully correlated with electrical properties (*I-V* or *C-V* methods) (Sugiyama *et al.,* 1992).

Furthermore, XSW is also used to determine the implanted impurity position of Bi in Si (Hertel *et al.,* 1985). Two reflection orders, (220) and (440), were used with SR to study the impurity distribution function at five doses ranging from (0.6 to 10) $\times 10^{14}$ Bi atoms/cm^2. The analysis reveals that the substitutional fraction of Bi position decreases from 100% to about 40% with increasing implantation dose, which can be related to a lattice expansion and the formation of precipitates.

A new application of the XSW was attempted (Lessmann *et al.,* 1995), in which not only Bragg reflection from the GaAs(001) substrate but also satellite diffraction from an AlAs/GaAs/AlAs/ . . . /GaAs superlattice were utilized. Photoelectrons for probing the shallow region and fluorescent x rays for probing deeper region were detected. When photons with 5.3 keV energy were incident on the short-period AlAs/GaAs superlattice (460 bilayers) on GaAs(001) substrate, a clearly modulated XSW profile of photoelectron yield was detected at the zeroth-order superlattice reflection, that is, about 55.82° of θ. The reflection can be interpreted as a Bragg reflection from the averaged unit cell of one superlattice period. Thus, the XSW method can give new insights into the structure of heteroepitaxial systems and buried interfaces.

5.5 X-Ray Diffraction: Microdiffraction, Surfaces, and High-Pressure Conditions

X-ray diffraction (XD) has been the most useful technique among hard x-ray applications such as diffraction, Compton scattering, absorption, fluorescence, nuclear resonance, and angiography. These techniques are now routinely used to characterize various crystals. By using the highly brilliant SR beam, it became possible to analyze crystals with small numbers of target atoms, such as microcrystals and surfaces (or interfaces).

A microdiffraction technique has been developed to analyze microcrystals by means of the Laue method using intense SR white x-ray beams. The performance of the extremely small Laue diffraction method was verified using a 0.8-μm-diameter Mo crystal. CVD diamond, $BaTiO_3$, $(Fe,Mg)_2SiO_4$, and interplanetary dust particles consisting of $(Fe,Ni)_{1-x}S$ with about 8-μm diameter have been analyzed. At present, it becomes possible to analyze about 2×10^{-2} μm^3 of microcrystal, which contains only 5×10^8 atoms. The crystallographic study of small crystals is expected to greatly advance protein and medical studies. Determination of crystal structures should lead directly to the development of new medicines. Recently, a glass capillary with an entrance inner diameter of 22 μm and an exit inner diameter of 3 μm was found to be very useful for obtaining a fine (less than 100 nm) beam of about 10-keV hard x rays, which will be applicable to microcrystallography and microphotoelectron microscopy (Bilderback *et al.*, 1994). Laue diffraction patterns were taken from a lysozyme crystal with a 5.6-μm-diameter x-ray beam and from a lithographically prepared gold pattern with a 360-nm beam. This is the smallest x-ray beam for protein crystallography experiments, taking advantage of the high flux of a low-divergence SR beam. This technique will be applicable to crystallographic diagnosis for finely patterned ULSI parts of submicron size.

The second application is surfaces and interfaces, where there are very few atoms compared to the bulk atoms. Because of the brilliant SR beam, sharp diffraction spots have been obtained from about 10^{14} atoms/cm^2. It is found that the superstructure that used to be believed to exist only on the clean surface in UHV also exists at the metal/semiconductor and insulator/semiconductor interfaces. Although many interface studies have been done on metal/semiconductor interfaces to control SBHs (Schottky-barrier heights) so far, atomic-level interface structures have not been elucidated yet. Akimoto *et al.* (1991) for the first time succeeded in the interface structure determination of Al/Yb/GaAs(100) systems having different reconstructions at the interface using grazing-incidence x-ray diffraction (GID). The Al/Yb/GaAs(100) with a 3-Å Yb interlayer having 0.76 eV of SBH showed a quite different atomic arrangement at the interface from that with a 32-Å Yb interlayer having 0.83 eV of SBH, although they both showed the same 4×1 superstructure. Thus, Schottky-barrier heights of these contacts are well correlated with the interfacial atomic arrangement.

Direct observation of Ge and Si ordering at the interface was performed by anomalous x-ray diffraction (Tweet *et al.*, 1992), which is a modification of GID at the Ge K absorption edge. The ordered $\sqrt{3} \times \sqrt{3}$ structure with interfacial boron atoms was analyzed for the heteroepitaxial $Ge_{1-x}Si_x$/Si(111) B($\sqrt{3} \times \sqrt{3}$)$R30°$ interface, which can be used for electronic devices with a strained-layer superlattice. By analyzing the diffracted intensities in a Patterson's map, the interfacial displacement of boron, silicon, and germanium atoms was precisely determined.

This surface diffraction technique was applied to the GaAs crystal growth process in the metallorganic chemical vapor deposition (MOCVD) process. During the crystal growth process using trimethylgallium (TMG) and tertiarybutylarsine (TBA) gases, x rays were directed onto the surface at a grazing-incidence angle to understand the growth mechanism and monitor the atomic-level layer-by-layer growth process (Kisker *et al.*, 1992). If the sample is atomically flat over the coherent length of the x rays, the surface scattering should be observed at the 110 reflection position, which is forbidden for the bulk. Thus the observation of 110 diffraction verifies the atomic layer-by-layer growth. As shown in Fig. 19, the oscillation of x-ray scattering intensity is also useful to monitor the layer-by-layer growth process in much the same fashion as RHEED oscillation in MBE growth (Fuoss *et al.*, 1994).

Under extremely high-pressure conditions, a lot of novel properties have been reported, such as insulator–metal transition and insulator–superconductor transition. Since these electrical properties are well correlated with crystal structures, it is of great importance to investigate the crystal structure change *in situ* under high pressure. So far, various kinds of materials such as graphite, BN, Mg_2SiO_4 and $MgSiO_3$, BaS, Ni_2SiO_4, Ga, Bi, Se, Te, I, Cs, S, O, Si, Ge, Sn, and III–IV compounds have been analyzed.

In the largest-volume press system, the energy-dispersive x-ray (EDX) diffraction technique with SR has been used in many studies, such as precise determination of pressure–temperature phase diagrams and equations of state. Although the EDX diffraction mode has an advantage of quick acquisition of diffraction data, it is difficult to derive reliable diffraction intensity for each reflection peak on account of uncertainty of the spectrum of incident radiation and difficulty in making precise correction of absorption effects. Thus, an angle-dispersive x-ray (ADX) diffraction measurement is necessary. A combined diffraction system with energy-dispersed XD (EDX) and angle-dispersive XD (ADX) was developed for high-pressure structure studies using a cubic-type multianvil press (Kikegawa *et al.*, 1995). Quick recording of the diffraction pattern is made in the EDX diffraction mode for pressure determination, and sample identification and high-quality diffraction patterns with reliable diffraction intensities are obtained in the ADX diffraction mode. The structure relation between Bi III and Bi III′ was investigated with this system. The volumetric compressibility was obtained by the EDX mode, and the atomic arrangement was determined from the data obtained in the ADX mode. The crystal structure of Bi III′ was found to be identical to that of Bi III.

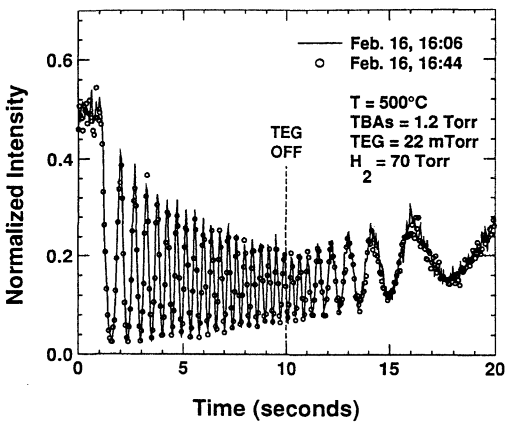

FIG. 19. Real-time monitoring of MOVPE homoepitaxial GaAs growth process by means of grazing-incidence x-ray diffraction. In the MOVPE growth, TBA (tertiarybutylarsine) gas and TEG (triethylgallium) gas were used (Fuoss *et al.*, 1994).

5.6 Topography and Tomography

As far as x-ray topography is concerned, a very bright SR beam opened two unique aspects in time-resolved analysis (real-time imaging) and high–angular-resolution analysis. X-ray topography with the very bright SR beam was applied to real-time monitoring of the GaAs and GaSb LPE (liquid-phase epitaxy) growth processes (Sato *et al.*, 1985). Real-time topography images of the solid–liquid interface taken with a Saticon image intensifier at (220) reflection revealed that a lot of slip zones are generated when the substrate temperature increases. The decrease of diffracted x-ray intensity would be entirely expected, since the positional ordering of the atoms disappears on melting.

The mechanism of primary recrystallization of steel sheets was investigated by *in situ* x-ray topography (see TOPOGRAPHY, X-RAY), which can provide rapid projection of the crystal grain orientation distribution in polycrystalline materials (Kawasaki *et al.*, 1995). The measurement system consists of an x-ray two-circle diffractometer and a translating imaging plate holder. Titanium-added ultralow carbon mild steel sheets with the thickness of 0.1 mm were used as a sample. A monochromatic SR beam of 0.06-nm wavelength was focused on the sample, which was rotated, while the imaging plate for detecting diffracted x rays two-dimensionally was translated. The results of measurements on some alloy sheets showed that the pattern recorded contains information about microstructures, size, and strain of crystal grains. Successive growing of crystal grains is seen as an increase in the size of diffraction spots. Thus, the rapid-projection method of crystal grain-orientation distribution is found very useful for investigating the primary recrystallization process of steel sheets for improving the mechanical properties of steel products.

A plane-wave topography technique with extremely high angular resolution was developed to characterize dislocation-free Si CZ (Czochralski method) and FZ (floating-zone method) wafers (Ishikawa *et al.,* 1989). The capability of producing a highly collimated x-ray beam at a desired wavelength enabled ultraplane-wave topography (UPWT) by a four-crystal arrangement. The interference pattern, which is usually smeared out in the conventional double-crystal system, can be clearly observed even for FZ Si wafers. Minute strain fields around defects are imaged on a topography as equal-inclination fringes. This technique with angle resolution of less than 0.01 arcs is also applicable to evaluation of process-induced microdefects or strains.

In addition to these two aspects in x-ray topography, a new aspect in x-ray computed tomography (CT) has appeared for nondestructive evaluation of industrial materials and medical diagnostics, after Hounsfield first developed this technique (Hounsfield, 1973). This CT becomes a more powerful tool using synchrotron radiation, because SR is an ideal x-ray source for a CT scanner since it can provide a highly collimated, high-intensity, tunable monochromatic x-ray beam. Thus the high–energy-resolution monochromatic x-ray CT (SR-CT) has enabled the imaging of element distribution and different chemical bonding states, *in situ* observation of transient phenomena, and various materials evaluation. One of the most important applications of this technique is medical diagnostics using K-edge subtraction for coronary angiography, which does not require high spatial resolution.

A new SR-CT method with the spatial resolution of about 10 μm has been developed (Hirano *et al.,* 1995). Figure 20 shows a three-dimensional CT image of a SiC/A 2024 composite rendered from 20 CT images taken at the maximum stress. The arrow *a* indicates a carbon fiber, which SiC fibers are surrounding. The arrow *b*, the transparent part, indicates the void that was formed by extraction of a fiber. The neighboring arrow *c* is also the void. Thus, the SR-CT method is

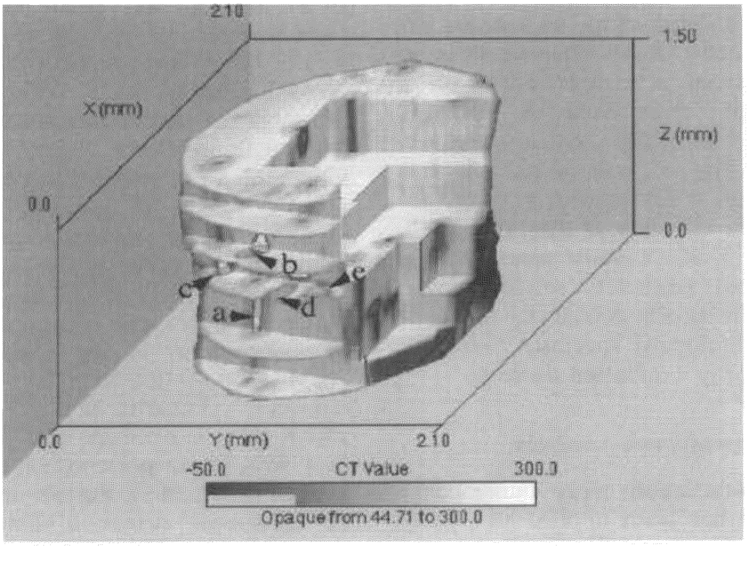

FIG. 20. A three-dimensional CT image of a SiC/A 2101 composite (Hirano *et al.*, 1995).

useful for nondestructive evaluation of materials. In order to improve the lateral resolution of SR-CT, more parallel incident x rays and submicron x-ray detectors should be developed. Furthermore, *in situ* observation of materials under heating and stress-field conditions will be important for elucidating destruction, deformation, and growth mechanisms of materials.

However, it is difficult to obtain high-contrast x-ray transmission images of organic materials using the ordinary CT method, since x-ray absorption coefficients of light elements are extremely small. In this respect, phase-contrast x-ray imaging using an x-ray interferometer is promising for observing the inner structure of organic materials. This is because even light elements cause x-ray phase shifts large enough to be detected. A phase-contrast x-ray CT (PCX-CT) using an x-ray interferometer to detect phase shifts has been developed (Momose *et al.*, 1996). The researchers used the fringe-scanning method to read the phase shift precisely from several interference patterns obtained by varying the external phase shift. Figure 21(a) shows an x-ray interferometer having three crystal plates for a phase-contrast CT image. When the incident beam to the x-ray interferometer satisfies the Bragg condition, an object beam and a reference beam are created. A liquid cell is inserted in the object-beam path. A wedge-shaped phase shifter in the reference-beam path is moved vertically to read the phase shift from interference patterns detected by an imaging detecter. Figure 21(b) shows phase-contrast x-ray CT images of a sample cut from a cancerous rabbit liver fixed in formalin. Shown in **a** is one of the reconstructed sectional images. The left half of the image is the cancer lesion. A 3D expression of the entire body of the CT data is displayed in **b**. Thus, PCX-CT is much more sensitive than the conventional x-ray CT, suggesting the possibility of drastically decreasing the x-ray dose into the biological specimen, which is vulnerable to x-ray irradiation damage.

5.7 X-Ray Fluorescence Analysis

Synchrotron-radiation x-ray fluorescence analysis (XFA) has been applied to nondestructive trace-element analysis of samples of biological, environmental, mineralogical, and industrial interest. The minimum detection limits are of the order of parts per billion. Using an SR x-ray microbeam, obtained with a Kirkpatrick–Baez mirror system (Noma and Iida, 1994), two-dimensional distributions of trace elements (elemental images) can be obtained with a lateral resolution of a few micrometers.

The x-ray microbeam technique has been applied also to surface and chemical-state analysis on a spatial scale of a few micrometers. Elemental distributions of K and Zn from the cross section of a human hair were used to demonstrate the lateral resolution (Iida and Noma, 1994). Furthermore, this technique is applicable for insulators, because XFA is free from the charging-up effect.

When x rays with good parallelism impinge onto a flat surface with a grazing angle lower than the critical angle such as 3.9 mrad for Si at the photon energy of 8 keV, total reflection occurs, resulting in a very shallow penetration depth such as 32 Å for Si. This total-reflection XFA was recognized as a useful analytical technique for trace analysis in a solution dropped onto a flat substrate surface. Since the background originating from the substrate can be drastically reduced because of the extremely shallow penetration depth, a very low detection limit such as parts per billion was obtained. This technique has been applied to surface impurity analysis on Si (Brennan *et al.*, 1994), because heavy-metal surface impurities such as Fe and Ni may create atomic-scale defects that lead to leakage currents, gate insulator breakdown, or poor threshold-voltage control, causing device reliability problems. Although the best conventional total x-ray reflection fluorescence (TXRF) using a rotating-anode x-ray source has achieved a sensitivity of 5×10^9 atoms/cm^2, the rapidly evolving ULSIs require a better detection limit, on the order of 10^8 atoms/cm^2. The combination of the bright SR beam from the 8-pole wiggler and the double-multilayer monochromator in the TXRF arrangement enabled a detection limit for Ni of 2.5×10^8 atoms/cm^2 (Pianetta *et al.*, 1995). This SR-TXRF takes advantage of a SR beam with over 95% linear polarization and a high degree of collimation. Further improvement to 10^7 atoms/cm^2 detection should be possible with a brighter SR source and a faster detector.

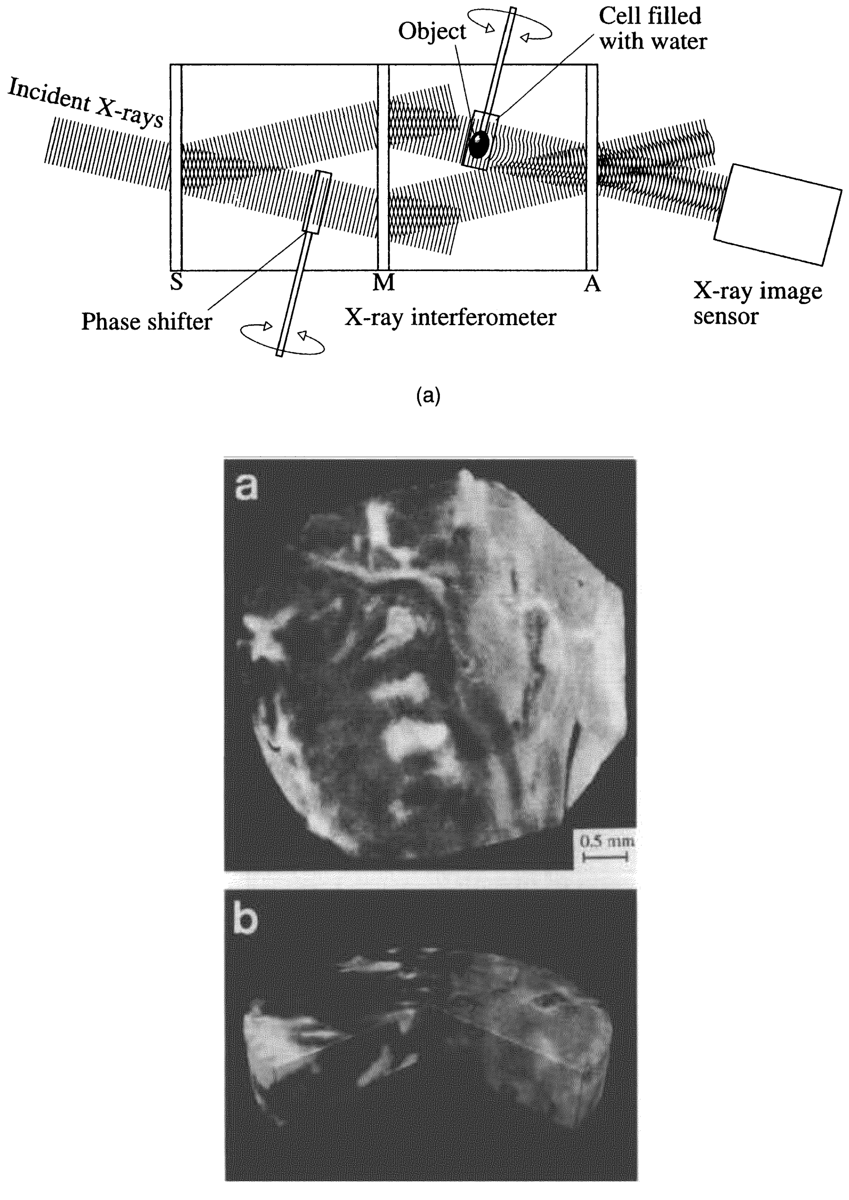
Object
Cell filled with water
Incident X-rays
S
M
A
Phase shifter
X-ray interferometer
X-ray image sensor
a
0.5 mm
b

(a)

(b)

FIG. 21. (a) An x-ray interferometer having three crystal plates for a phase-contrast CT image. (b) Phase contrast x-ray CT images of a sample cut from a cancerous rabbit liver fixed in formalin. Part **a**: One of the reconstructed sectional images. The thickness corresponding to the image is 12 μm. The left half of the image is the cancer lesion. Part **b**: A 3D expression of the entire body of the CT data. One quadrant and outer part have been cropped. (Momose *et al.*, 1996.)

5.8 Infrared Spectroscopy

Infrared (IR) spectroscopy provides the means for analyzing the molecular chemistry of materials in detail. However, because of its longer wavelengths, IR radiation limits the spatial resolution, but its ability of resolving chemistry is extremely useful. Thus, IR microspectroscopy (IMS) has been developed in combination with light microscopy and IR spectroscopy to correlate microstructure with molecular chemistry. Synchrotron radiation source has 100–1000 times higher brilliance than a traditional thermal-emission source, and because of its nonthermal source, it produces less Boltzmann source noise. Thus, high brightness, low noise, and improved optical performance make the SR an ideal source for IMS.

A scanning IR microprobe consisting of an integrated Fourier transform infrared (FTIR) spectrometer and microscope optical module was developed (Reffner *et al.*, 1995). They adopted Schwarzschild reflecting designs as the microscopic optics using 0.65-NA (numerical aperture) and 0.58-NA objective lenses. A rapid-scan Michelson interferometer with a germanium-coated KBr beam splitter was used. Spectral resolution is from 2 to 16 cm^{-1}. The beam-profile measurement revealed that the IR beam was focused to a diffraction-limited spot size of less than 12.5 μm. Furthermore, it is found that scanning across a sharp boundary with a 6-μm-diameter aperture produced a maximum spatial resolution of 6 μm. The signal-to-noise ratio was found to be 250:1 in the IR spectra, where the total acquisition time was 5 s per spectrum.

So far, IMS, which is a basic technology for the molecular characterization of materials, has been applied in various research fields such as biological science, polymer science, earth science, forensic science, and cultural materials. In biological science, the single human hair erythrocyte was studied using FTIR microspectroscopy (Dong *et al.*, 1988). The single-cell spectrum contained all of the pertinent information about both the characteristic protein's secondary structure and bound carbon monoxide. Further studies of single red blood cells showed that different mutant forms were identifiable through IMS. In polymer science, polymer-dispersed liquid-crystal films were studied by FTIR microspectroscopy to compare the polymer and liquid-crystal droplet regions (McFarland *et al.*, 1993). Spatially resolved functional-group maps of a droplet showed characteristic textures corresponding to the visually observed morphology of the droplet. It was possible to correlate the microscopic texture with the chemistry and with the changes that occur when electric fields are applied. In earth science, synchrotron IMS provided the high signal-to-noise ratio necessary to quantitate the trace amount of hydrogen in nominally anhydrous perovskite. $MgSiO_3$ single crystals were quantitatively analyzed, which were grown directly from H_2O-rich melts of $MgSiO_3$ at a pressure of 27 GPa and a temperature of 1830 °C (Meade *et al.*, 1994). They found that the hydrogen concentration is comparable to 12% of the mass of hydrogen in the earth's atmosphere. In forensic science, a new approach to forensic analysis with IMS was performed (Bartick *et al.*, 1994). The microscopic attenuated internal-reflection spectroscopy (ATR) technique is an extension of IMS that permits the analysis of materials' surface with little or no sample preparation. This method will open a new dimension to the evaluation of forensic evidence. Thus, with SR, IMS can be performed at the diffraction limit of the IR radiation, and spectra with high signal-to-noise quality can be obtained.

6. MEDICAL AND BIOLOGICAL SCIENCE AND TECHNOLOGY

It is said that the twenty-first century is a century of bioscience or life science. SR is expected to play an important role in life science, especially in studies on biological substances in the crystalline state, the noncrystalline state, the solution state, and the amorphous state. Among various kinds of biological substances, it is believed that protein crystallography is the most important target for bioscience.

Protein is in charge of translation of heredity information, energy conversion, information transmission, action, and substance synthesis. In order to elucidate the characteristic function of a protein or nucleic acid, it is necessary to analyze the molecular structure at atomic resolution and understand how the amino-acid arrangement forms its three-dimensional structure. On the basis of this information, it becomes possible to design new functional proteins.

Furthermore, the medical diagnosis system using the *K* edge of injected iodine, contraction of muscle, *in vivo* soft x-ray microscopy, and evaluation of irradiation damage are also important subjects in life science. Without brilliant SR beams, these techniques cannot be developed.

6.1 SR Angiography

X-ray imaging and x-ray crystallography are promising for medical and biological applications. After the usefulness and prospects for practical applications in these fields were recognized, many researchers in medical industries became involved in R&D of SR analysis techniques mainly in the following four systems. One is the *K*-absorption–edge subtraction method using monochromatic x rays for a heart vascular system. The second target is the high–spatial-resolution monochromatic x-ray CT system for high-precision diagnosis of physiological circulation in the brain. The third target is the imaging examination system for the fine vascular system for early finding of malign tumor or cancer. The last target is the new imaging system with fluorescence x rays emitted from a target element for trace elements in the human body.

Among these the most promising application is the new real-time imaging technique for the heart vascular system using *K*-edge subtraction. This technique has been developed by using an iodine filter and an imaging focusing and processing system. As shown in Fig. 22, a monochromatic SR beam with photon energy below and above the iodine *K* edge irradiates an object injected with iodinated material, and real-time images are processed by subtraction to obtain clearly distinguished images of heart.

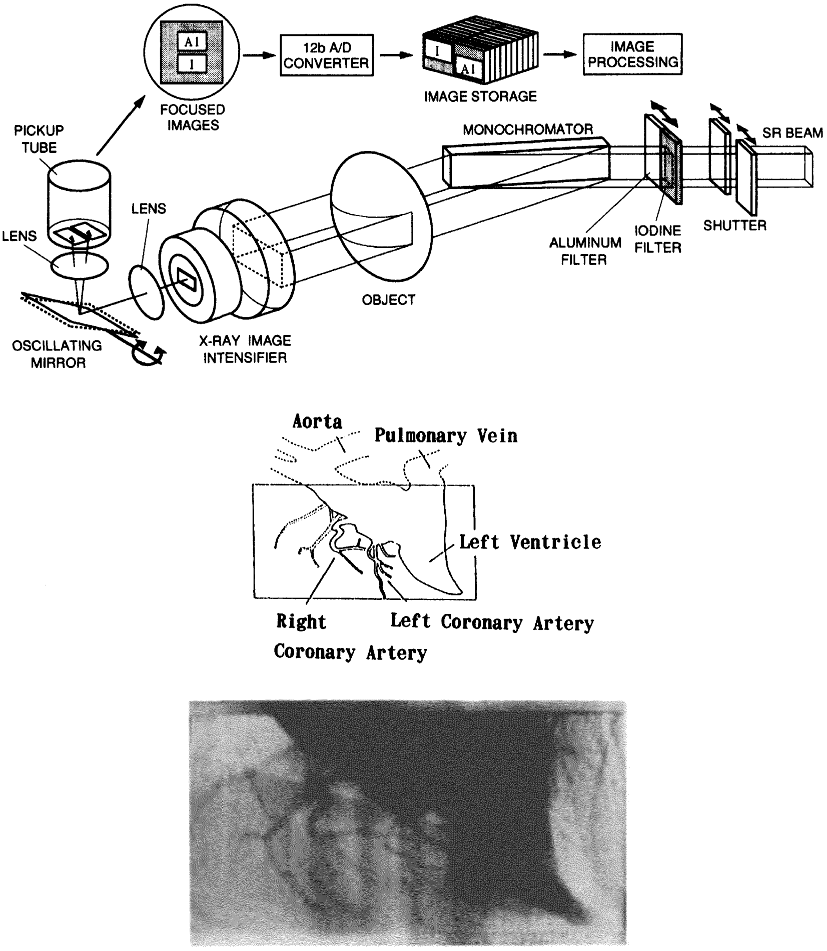
AI
I
12b A/D CONVERTER
I
AI
IMAGE STORAGE
IMAGE PROCESSING
FOCUSED IMAGES
PICKUP TUBE
MONOCHROMATOR
SR BEAM
LENS
LENS
ALUMINUM FILTER
IODINE FILTER
SHUTTER
OBJECT
OSCILLATING MIRROR
X-RAY IMAGE INTENSIFIER
Aorta
Pulmonary Vein
Left Ventricle
Right Coronary Artery
Left Coronary Artery

FIG. 22. Schematic diagram of the dual-energy real-time imaging system and a *K*-edge substraction image of a 12-kg dog's heart (Ueda *et al.*, 1989).

This *K*-edge subtraction system is under development for intravenous coronary angiography capable of providing a screening test for ischemic heart diseases. X-ray images of a dog taken using monochromatic x rays at above the *K*-edge energy of iodine clearly distinguished the coronary arteries 7–8 s after the iodinated contrast materials were injected into a leg vein (Ueda *et al.*, 1989). The optimal condition for transvenous coronary angiography to obtain clear images of human coronary arteries was reported (Chapman *et al.*, 1995). They used a 1200-element dual-array detector with a pixel-to-pixel spacing of 0.25 mm. In the near future this intravenous coronary angiography will be performed routinely for heart disease patients. This will need a large storage ring such as SPring-8 at 8 GeV to provide iodine *K*-edge hard x rays of about 33 keV.

The evaluation system for blood flowing in the brain is also important for diagnosis of brain tumor before and after the operation. Although the RI (radioisotope) imaging system and the magnetic-resonance imaging (MRI) system have been used for that purpose, the spatial resolution is still as poor as several millimeters. In order to improve the spatial resolution for the brain blood-flow diagnosis system, a monochromatic x-ray CT system was developed using an about 50-keV monochromatic SR beam whose width was expanded by asymmetric reflection, and a one-dimensional charge-coupled device (CCD) detector (see CHARGE-COUPLED DEVICES). The spatial resolution has reached 7 μm. Thus, the medical diagnosis system using SR is regarded as a significantly important technique for human life. In the near future, compact SR rings that are specified for these medical diagnosis systems will be developed.

6.2 Protein Crystallography

The extremely bright SR beam has opened new fields, such as structural analysis of macromolecules. Knowing their three-dimensional structures is vital for revealing the mechanism of activities of biological macromolecules such as proteins and viruses. Since biotechnology or protein engineering is expected to become a huge industry in the next century, many efforts are being devoted to establish a data bank of protein structures. For that purpose, x-ray crystallography is one of the most powerful techniques to determine three-dimensional structures of biological macromolecules at the atomic level. Nowadays new three-dimensional structures of proteins are reported in many journals such as *Nature* and *Science.* Actually, SR has drastically reduced the measurement time and enabled the measurement of small protein crystals. Furthermore, new methods to determine protein crystal structures, such as the multiple-wavelength anomalous-dispersion (MAD) method for phase calculation and a time-resolved Laue method using white x rays, have been developed.

A data collection system for macromolecular crystals using a Weissenberg camera and an imaging plate was developed at the Photon Factory at the National High Energy Physics Laboratory, Tsukuba (KEK), for high-resolution, high-precision, and high-speed data collection. It takes only a few hours to collect all the diffraction data from one crystal. Figure 23 shows a data collection system consisting of a Weissenberg camera and an imaging plate (Sakabe, 1983; Sakabe *et al.*, 1995), and Plate 2 shows an atomic model of Actin-DNase 1 complex constructed on the basis of x-ray diffraction patterns (Sasaki *et al.*, 1993). This complex consists of 634 amino-acid bases, and the total

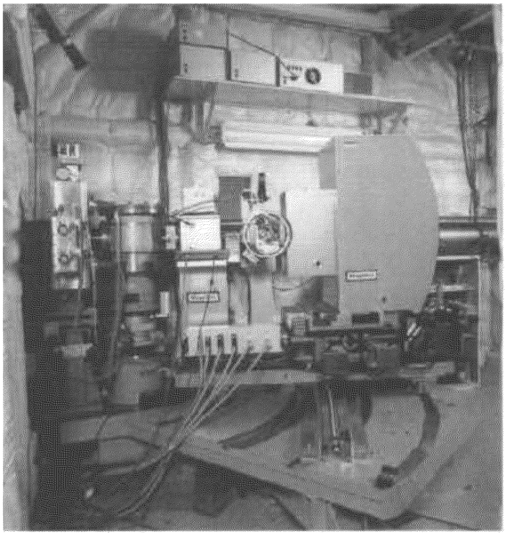

FIG. 23. A protein crystallography system with a Weissenberg camera and an imaging plate (courtesy of N. Sakabe).

molecular number is 73 000 Da. The crystal structure is rhombohedral with the space group of $P2_12_12_1$, and the lattice parameters a, b, and c are 42.0, 225.3, and 77.4 Å, respectively. This analysis was done with an extremely high resolution, 1.8 Å.

The HIV reverse transcriptase, which is an AIDS-related protein, was analyzed (Jones *et al.*, 1993). The work revealed that the space group of this crystal is $P4_222$ and that the lattice parameters a and b are 120 Å and c is 320 Å. The structure of interferon, which is an important protein for medical applications such as medicines for brain cancer and B- and C-type liver diseases, was determined (Mitsui *et al.*, 1993). A double-shelled spherical virus that has a huge molecular weight, as high as 65 200 000 Da, was analyzed (Mizuno *et al.*, 1991). Crystals of the virus belong to the space group $I23$, and the lattice parameter a is 789 Å. This result gave important information about the interaction between nucleic acids and proteins, the flexibility of proteins, and the interactions between proteins.

The very bright SR beam has also enabled dynamic structure analysis of protein crystals using white x rays. The reaction time of enzymes with their substrates ranges from sub-milliseconds to minutes in solution, while it becomes about 100 times slower in crystals. A new time-resolved Laue camera system is being developed to measure diffraction with a 10-ms exposure, where millisecond x-ray pulses together with an imaging plate that is synchronously moved are used.

Thus, SR protein crystallography is expected to open the key to understanding the mechanism of various life phenomena and to strongly support the design of new drugs.

6.3 Small-Angle X-Ray Scattering

Elucidation of the relationship between the structure change of living substances and their functions is one of the most important subjects in biophysics. Among various chemical reactions and structure changes, contraction of muscle is regarded as one of the most suitable systems for this subject, because the structure of muscle is directly correlated with the function. As shown in Fig. 24, the dominant proteins constituting the muscle are actin and myosin, and the fundamental process of muscle contraction is the interaction between the two proteins by extracting energy from the hydrolysis of adenosine triphosphate (ATP). Although the structures of actin and myosin, which are regularly arranged like liquid crystals in the form of filament, have been analyzed by x-ray diffraction, the clear answers about how actin and myosin interact, how the structure change happens, and how they are correlated with the generation of tension have not been obtained yet.

In order to answer these questions, small-angle x-ray scattering using SR has been utilized for understanding the dynamic processes of chemical bonding between actin and myosin, and the transfer of chemical energy to molecules on an atomic scale. Furthermore, the mechanisms of the structure change of regulatory proteins (such as troponon and tropomyosin) by Ca ions and the information transfer to contractile proteins should be also elucidated on a molecular scale.

Time-resolved analysis of the intensity change of a specific diffraction spot together with the generation of tension can provide important information. The myosin heads were found to bond to actin during contraction by measuring the 5.9-nm reflection corresponding to the pitch of the left-handed screw of actin filament and the 5.1-nm reflection corresponding to the pitch of the right-handed screw of actin filament (Wakabayashi *et al.*, 1985). In this experiment, a one-dimensional position-sensitive proportional counter (PSPC) with the time resolution of millisecond order was used.

In order to realize more precise analysis of the structure change during muscle contraction and relaxation, a new two-dimensional detector system called an imaging plate, which consists of a photostimulated phosphor, was introduced (Amemiya and Miyahara, 1988). This system has a large dynamic range of more than five orders of magnitude. By using this system, x-ray diffraction images of frog skeletal muscle during contraction and relaxation were obtained with exposure time as short as 10 s with high spatial resolution. They found that the reflection corresponding to the helical structure of the myosin heads is drastically decreasing during contraction, but the 14.3-nm meridional reflection of the myosin heads showed a strong intensity. These results in-

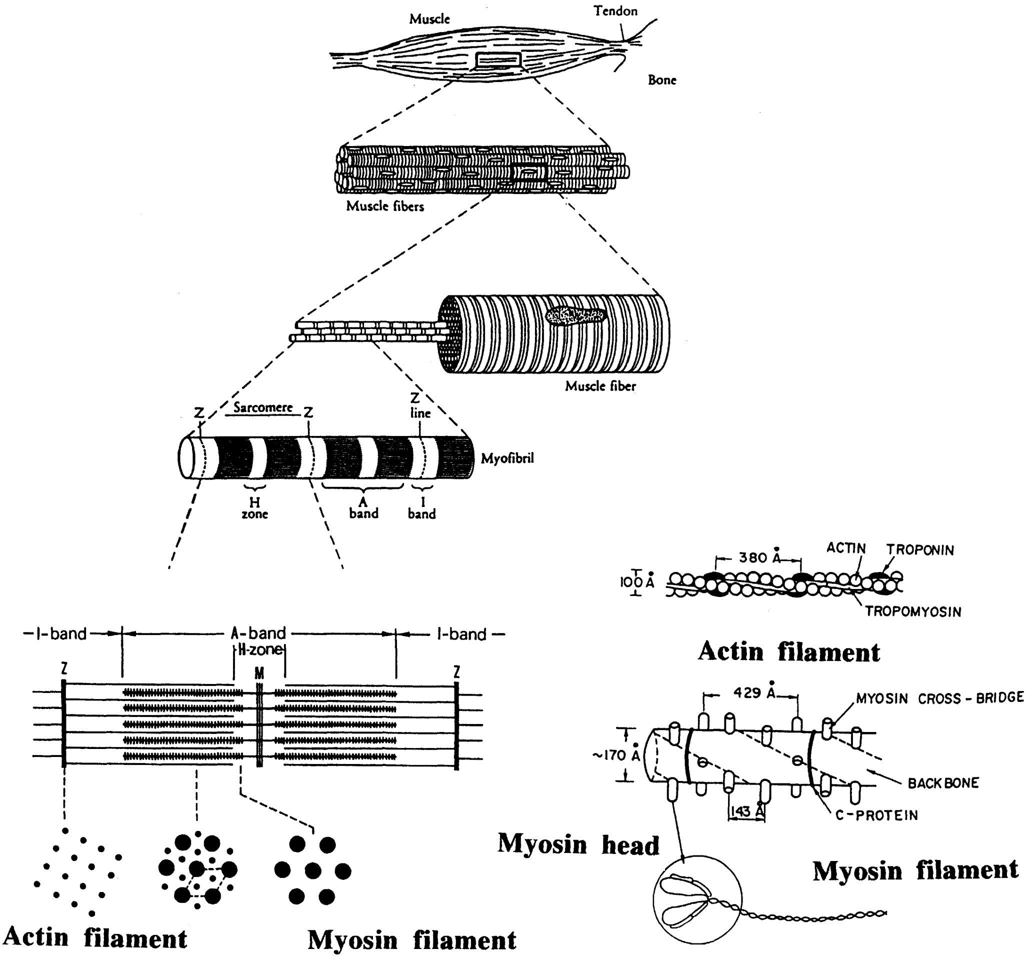
Muscle
Tendon
Bone
Muscle fibers
Muscle fiber
Z
Sarcomere
Z
Z line
Myofibril
H zone
A band
I band
380 Å
ACTIN
TROPONIN
100 Å
TROPOMYOSIN
Actin filament
–I-band
A-band
I-band –
H-zone
Z
M
Z
429 Å
MYOSIN CROSS - BRIDGE
~170 Å
BACKBONE
143 Å
C-PROTEIN
Myosin head
Myosin filament
Actin filament
Myosin filament

FIG. 24. A schematic diagram of structure of muscle (courtesy of Y. Amemiya).

dicate that myosin heads move toward the azimuthal direction but still interact with actin with a 14.3-nm periodicity along the actin filament axis.

Furthermore, more precise analysis of actin reflection, which is extremely weak (about 10^{-4} of the equatorial reflection), revealed that the actin filament is extended by 0.3–0.4% during contraction and that the energy from ATP, which is decomposed at the myosin heads, causes a structure change of myosin heads themselves and is converted into the elastic energy to extend actin.

Thus, many important phenomena have been discovered by the SR small-angle x-ray scattering method, and a real image of the muscle contraction mechanism is being elucidated. However, these measurements provide information about averaged phenomena in several tens of thousands of fibers. In the near future, time-resolved measurement of a single muscle fiber of about 100 μm in diameter would be realized by development of a more brilliant SR beam and higher-sensitivity x-ray detectors in order to provide information about the correlation between ATP decomposition and mechanical response of molecular structure.

6.4 *In Vivo* Soft X-Ray Microscopy

X-ray microscopy can directly analyze living biological substances in water. It is well known that water has a window between the O *K* edge (about 530 eV) and the C *K* edge (about 280 eV), which is the useful energy region for analyzing living biological samples using soft x rays. Soft x-ray microscopy has several advantages over electron microscopy. First, the sharp x-ray absorption dependence on elements and/or wavelength enables one to realize microscopy that is sensitive to elemental composition. Second, specimens can be observed directly in their natural state without any fixing and staining.

A new imaging technique taking advantage of this water window has been developed by using a Fresnel zone plate (FZP). The FZP is a circular transmission diffraction grating consisting of alternating opaque and transparent rings. It gives a spatial resolution almost equal to its outermost zone width. The FZPs are fabricated by using the current ULSI microprocessing technology such as electron-beam lithography and reactive-ion etching. Since the outermost zone is about 50 nm wide, a corresponding resolution of about 50 nm can be expected (Rarback *et al.*, 1990).

As an example of scanning x-ray microscopy (SXM), the SXM images of a chromosome taken at the photon energies below and above the C *K* edge were shown to be different from each other, indicating that chemical-state mapping is possible by changing the wavelength to discriminate successfully C=O bonds from C—N bonds in DNA (deoxyribonucleic acid) (Ade *et al.*, 1992). A two-dimensional Ca distribution in mineralized tissues was obtained by the absorption-difference imaging method at the Ca *K* absorption edge (Buckley, 1995). If the photon energy can be increased up to the Al *K* edge (about 1.56 keV), it is possible to obtain the Al distribution in brain cells, which may be correlated with the Alzheimer disease.

In contrast, the highest spatial resolution in soft x-ray microscopy was achieved by means of an imaging x-ray microscope (IXM) using the FZP system. Very clear IXM images of Kupfer cells of rat liver were obtained using the phase-contrast method (Schmahl *et al.*, 1995). Thus, soft x-ray microscopes are expected to open new medical and biological fields.

6.5 Irradiation Effects on Biology

Radiobiology is an important field from the viewpoints of radiotherapy of cancer using heavy particles and protection from environmental radiation of low dose rates. Radiobiological effects are induced with energy deposited by radiation. The deposition events occur randomly and discretely. The average amount of energy deposited per event is about 30 eV, which produces two ion pairs on average. The consequences of each energy deposition event should be understood. Although deactivation processes take place, molecular changes produce the biological damage. We also have to understand how these consequences change depending upon the density of energy deposition events in space and time through the interaction of the reactive species produced by them.

In order to elucidate these radiation processes, it is necessary to deposit energy into biological samples in a well-controlled manner. One way is to irradiate samples with low-energy monochromatic photons. Since vacuum ultraviolet photons of about 40-nm wavelength give an energy quantum of about 30 eV around the absorption site, it is possible to study the processes and their efficiencies in producing molecular changes induced with this energy. In the higher photon-energy region, the processes of absorbing energy by *K*-shell electrons of oxygen in water molecules contained in living cells occur, and a shower of almost monoenergetic electrons is produced in the cells.

Using plasmid DNA, it was revealed that the base damage recognized by an UV endonuclease (with specificity for pyrimidine dimers) was not produced by photons below 180-nm wavelength and that single-strand breaks (SSB) were produced more effectively with shorter-wavelength photons (Hieda *et al.*, 1994). It was noted that the action spectrum of SSB production was similar to that of the absorption of the sugar-phosphate moiety of DNA, indicating that SSB might be produced along with the photoabsorption of the backbone (sugar) of DNA. Further studies using dinucleotides have revealed that SSB is produced associated with the destruction of the sugar moiety, releasing fragments having a 5′ phosphate end (Ito, 1991). Recently the energy dependence for the production of SSB and of double-strand breaks (DSB) in pBR322 DNA was shown in the energy region from 60 to 550 eV (Hieda, 1994). The yield for DSB increased with the increase in photon energy, when compared with the yield for SSB. Thus a photon energy of about 300 eV is needed for DSB induction.

At a higher-energy region, most works have been focused on photoabsorption at the *K* shell of phosphorus in DNA because of the strategic position in the biologically important molecule, taking advantage of the complete tunability of SR. The enhancement in the biological effects at the resonance absorption peak can be considered to be initiated by the photoabsorption by phosphorus. This enhancement was observed in several organisms including lethality and induction of genetic changes in yeast, lethality in bacteria, chromosome aberration in mammalian cells, and mutation and killing in mammalian cells (Kobayashi *et al.,* 1991).

As a close simulation of aqueous systems, SSB and DSB were measured with the plasmid DNA irradiated in aqueous solutions. Strand-break yields were shown to be enhanced by the photoabsorption of phosphorus in the DNA molecule. Similar enhancements were observed for the release of adenine from ATP in a concentrated aqueous solution. These data suggest that water molecules may play an important role in the production mechanisms of the damage in aqueous systems.

Further studies on molecular mechanism of biological action using inner-shell ionization and energy dependence of molecular change due to SR irradiation ranging from VUV to x rays should be done to understand the radiobiology mechanism.

7. FUTURE PROSPECTS

Based on more than 30 years of experience in SR research, third-generation SR sources have been constructed or are now being constructed, including high-energy SR sources such as ESRF (6 GeV, Grenoble, France), APS (7 GeV, Chicago, Illinois, U.S.A.), and SPring-8 (8 GeV, Nishi-Harima, Japan). Furthermore, medium-energy SR sources of the third generation such as ALS (Advanced Light Source at U.C. Berkeley, California, U.S.A.), Super ACO (Paris, France), PLS (Pohang Light Source, Pohang, Korea), SRRC (Synchrotron Radiation Research Center, Hsinchu, Taiwan), and Elettra (Trieste, Italy) with energies of about 1–2 GeV are now available for experiments with a very low emittance.

These advanced storage rings together with undulators (linearly and circularly polarized), high-efficiency monochromators, and high-throughput optics will make it possible to focus the SR beam to much less than 0.1 μm with extremely high energy resolution, greater than 10 000. Thus, experiments with high spatial resolution, time resolution, energy resolution, and circular polarization will become possible, leading to complete understanding of phenomena in materials and life sciences and new development of atomic-scale characterization, real-time process monitoring, and advanced fabrication processes using SR-excited film growth and etching. Furthermore, new physical or electronic properties such as quantum size effects in quantum nanostructures will be discovered. These advanced SR techniques will be also combined with other advanced tools such as STM, focused ion beams, electron beams, and laser beams in a complementary manner in order to discover novel phenomena and control properties on an atomic/molecular scale.

ACRONYMS

ALE: Atomic layer epitaxy.

ATP: Adenosine triphosphate.

ATR: Attenuated internal reflection.

CCD: Charge-coupled device.

CT: Computed tomography.

CVD: Chemical vapor deposition.

CZ: Czochralski method.

DMAH: Dimethyl aluminum hydride.

DNA: Deoxyribonucleic acid.

DRAM: Dynamic random access memory.

EXAFS: Extended x-ray absorption fine structure.

FZ: Floating zone method.

FZP: Fresnel zone plate.

GID: Grazing-incidence diffraction.

GSPC: Gas-scintillation proportional counter.

IMS: Infrared microspectroscopy.

LIGA: German acronym for three basic processes of deep-etch lithography with synchrotron radiation, electroforming, and plastic molding.

LPE: Liquid-phase epitaxy.

MBE: Molecular-beam epitaxy.

MESFET: Metal-semiconductor field-effect transistor.

MISFET: Metal-insulator-semiconductor field-effect transistor.

PCX-CT: Phase-contrast x-ray computed tomography.

RHEED: Reflection high-energy electron diffraction.

SBH: Schottky-barrier height.

SRPES: Synchrotron-radiation photoelectron spectroscopy.

SSB: Single-strand break.

TBA: Tertiarybutylarsine.

TEM: Transmission electron microscope.

TMG: Trimethylgallium.

TOF: Time-of-flight.

TXRF: Total x-ray reflection fluorescence.

ULSI: Ultralarge-scale integration.

UHV: Ultrahigh vacuum.

UPWT: Ultraplane-wave topography.

VUV: Vacuum ultraviolet.

XAFS: X-ray absorption fine structure.

XANES: X-ray absorption near-edge structure.

XFA: X-ray fluorescence analysis.

XSW: X-ray standing waves.

GLOSSARY

Micromachine: Mechanical system with the size of less than millimeters, usually consisting of microactuators, microsensors, and microprocessors. Micromechanical parts such as microgears, micromotors, etc. are also called micromachines.

Photochemistry: Chemical reactions induced by photoabsorption, not thermal effect. Lithography and photo CVD are examples. There are two types of photochemical reaction: one is the direct excitation of valence electrons by several-eV photons such as lasers or halogen lamps, and the other is core-electron excitation followed by Auger de-excitation, which takes place by SR irradiation.

Photoelectron Spectroscopy: A form of electron spectroscopy for measuring the energies of core-level and valence electrons that are emitted from atoms or solids when monochromatic x rays or VUV photons are absorbed.

Storage Ring: A circular vacuum ring to store charged particles such as electrons or positrons moving with extremely high speed, close to the speed of light, to produce synchrotron radiation.

Synchrotron Radiation: A kind of bremsstrahlung radiation emitted in the forward direction tangential to the orbit when charged particles such as electrons are forced to move at speeds close to the speed of light in a circular orbit.

Tomography: The use of x rays to photograph a selected plane of a human body with other planes eliminated. The computerized tomography scanner is a ring-shaped x-ray machine that rotates through 180°. A three-dimensional image of the tissues under examination is formed.

X-Ray Absorption Fine Structure: Technique to provide information about the short-range order such as the number, types, and distances of neighbors by analyzing the x-ray absorption spectra at energies higher than the absorption edge of the target atom. XAFS is useful in analyzing amorphous materials, surfaces, and so on.

X-Ray Diffraction: Diffraction phenomena of x rays as an electromagnetic wave in crystals, which can be used to determine structures of crystals including protein crystals by means of x-ray diffraction camera and diffractometers.

X-Ray Microscopy: Microscopy using soft x rays with photon energy between the carbon K edge and the oxygen K edge mainly for biological specimens. There are two types of x-ray microscopes, namely, the projection type and the scanning type.

Works Cited

Abrahams, J. P., Leslie, A. G. W., Lutter, R., Walker, J. E. (1994), *Nature* **370,** 621–628.

Ade, H., Zhang, X., Cameron, S., Costello, C., Kirz, J., Williams, S. (1992), *Science* **258,** 972–975.

Akazawa, H., Utsumi, Y., Urisu, T., Nagase, M. (1993), *Phys. Rev. B* **47,** 15946–15949.

Akimoto, K., Hirose, K., Mizuki, J. (1991), *Phys. Rev. B* **44,** 1622–1627.

Allain, L. L., Regnard, J. R., Bourret, A., Tourillon, G., Pizzini, S., Parisini, A., Armigliato, A. (1993), *Jpn. J. Appl. Phys.* **32,** 625–627.

Amemiya, Y., Miyahara, J. (1988), *Nature* **336,** 89–93.

Bartick, E. G., Tungol, M. W., Reffner, J. A. (1994), *Anal. Chem. Acta* **288,** 35–42.

Barton, J. J. (1988), *Phys. Rev. Lett.* **61,** 1356–1359.

Berman, L. E., Batterman, B. W., Blakely, J. M. (1988), *Phys. Rev. B* **38,** 5397–5405.

Bilderback, D. H., Hoffman, S. A., Thiel, D. J. (1994), *Science* **263,** 201–203.

Brennan, S., Tompkins, W., Takaura, N., Pianetta, P. J., Kortright, B., Laderman, S., Fischer-Colbrie, A., Ishii, H., Madden, M. C., Wherry, D. C. (1994), *Nucl. Instrum. Methods* **A347,** 417–421.

Brillson, L. J. (1993), *Contacts to Semiconductors,* Park Ridge, NJ: Noyes Publications.

Buckley, C. J. (1995), *Rev. Sci. Instrum.* **66,** 1318–1321.

Chapman, D., Thomlinson, W. C., Gmur, N. F., Dervan, J. P., Stavala, T., Giacomini, J., Gordon, H., Rubenstein, E., Lavender, W., Schulze, C., Thompson, A. C. (1995), *Rev. Sci. Instrum.* **66,** 1329–1331.

Chiang, T.-C., Knapp, J. A., Eastman, D. E., Aono, M. (1979), *Solid State Commun.* **31,** 917–920.

Colvin, V. L., Alivisatos, A. P., Tobin, J. G. (1991), *Phys. Rev. Lett.* **66,** 2786–2789.

Daimon, H. (1988), *Rev. Sci. Instrum.* **59,** 545–549.

Daimon, H. (1993), *J. Jpn. Synchrotron Radiat. Soc.* **6,** 323–330 (in Japanese).

Deguchi, K., Miyoshi, K., Ishii, T., Matsuda, T. (1992), *Jpn. J. Appl. Phys.* **31,** 2954–2958.

Denlinger, L. D., Rotenberg, E., Warwick, T., Visser, G., Nordgren, J., Guo, J.-H., Skytt, P., Kevan, S. D., McCutcheon, K. S., Shuh, D., Bucger, J., Edelstein, N., Tonner, B. P. (1995), *Rev. Sci. Instrum.* **66,** 1342–1345.

Dev, B. N., Aristov, V., Hertel, N., Thundat, T., Gibson, W. M. (1985), *Surf. Sci.* **163,** 457–477.

Dong, A. *et al.* (1988), *Biochem. Biophys. Res. Commun.* **156,** 752.

Ederer, D. L. (1964), *Phys. Rev. Lett.* **13,** 760.

Elder, F., Gurewitsch, A. M., Langmuir, R. V., Pollack, H. C. (1947), *Phys. Rev.* **71,** 829.

Fuoss, P. H., Kisker, D. W., Payne, A. P., Stephenson, G. B., Brennan, S. (1994), in: I. Ohdomari, M. Oshima, A. Hiraki (Eds.), *Control of Semiconductor Interfaces,* Amsterdam: Elsevier.

Hasnain, S. S., Helliwell, J. D., Kamitsubo, H. (1994), *J. Synchrotron Radiat.* **1,** 1–4.

Hayashi, Y. (1994), *Extended Abstracts of the 1994 Solid State Devices and Materials Conference, Yokohama,* Tokyo: Japan Society of Applied Physics, pp. 403–405.

Hertel, N., Materlik, G., Zegenhagen, J. (1985), *Z. Phys. B—Condensed Matter* **58,** 199–204.

Hieda, K. (1994), *Int. J. Radiat. Biol.* **66,** 561–567.

Hieda, K., Suzuki, M., Hirano, T., Suzuki, M., Furusawa, Y. (1994), *J. Radiat. Res.* **35,** 104.

Himpsel, F. J., McFeely, F. R., Taleb-Ibrahimi, A., Yamoff, J. A., Hollinger, G. (1988), *Phys. Rev.* **B38,** 6084–6096.

Hirano, T., Usami, K., Tanaka, Y., Matsuda, C. (1995), *J. Mater. Res.* **10,** 381–386.

Hosokawa, T., Kitayama, T., Hayasaka, T., Ido, S., Uno, Y., Shibayama, A., Nakata, J., Nishimura, K., Nakajima, M. (1989), *Rev. Sci. Instrum.* **60,** 1783–1785.

Hounsfield, G. N. (1973), *Brit. J. Radiol.* **46,** 1016.

Iida, A., Noda, T., Hayakawa, S., Takahashi, M., Gohshi, Y. (1993), *Jpn. J. Appl. Phys.* **32,** 160–164.

Iida, A., Noma, T. (1994), *Nucl. Instrum. Methods* **B82,** 129–138.

Ishihara, S. (1994), *NTT R&D* **43,** 585–596.

Ishikawa, T. (1988), *Acta Crystall.* **A44,** 496–499.

Ito, T. (1991), in: R. M. Sweet, A. D. Woodhead (Eds.), *Synchrotron Radiation in Structural Biology,* New York: Plenum, 355 pp.

Jewell, T. E., Rodgers, J. M., Thompson, K. P. (1990), *J. Vac. Sci. Technol.* **B8,** 1519–1523.

Jones, Y. *et al.* (1993), *J. Cryst. Growth* **126,** 261–266.

Kawamura, T., Oshima, M., Fukuda, Y., Ohmachi, Y., Izumi, K., Hirano, K., Ishikawa, T., Kikuta, S. (1991), *Surf. Sci.* **251/252,** 185–190.

Kawasaki, K., Iwasaki, H., Yoshinaga, N. (1995), *Rev. Sci. Instrum.* **66,** 1404–1406.

Kendelewicz, T., Woicik, J. C., Miyano, K. E., Cowan, P. L., Karlin, B. A., Bouldin, C. E., Pianetta, P., Spicer, W. E. (1991), *J. Vac. Sci. Technol.* **B9,** 2290–2293.

Kikegawa, T., Chen, J.-H., Yaoita, K., Shimomura, O. (1995), *Rev. Sci. Instrum.* **66,** 1335–1337.

Kinoshita, H., Kurihara, K., Ishii, Y., Torii, Y. (1989), *J. Vac. Sci. Technol.* **B7,** 1648–1651.

Kisker, D. W., Stephenson, G. B., Fuoss, P. H., Lamelas, F. J., Brennan, S., Imperatori, P. (1992), *J. Cryst. Growth* **124,** 1–9.

Kobayashi, K., Hieda, K., Maezawa, H., Furusawa, Y., Suzuki, M., Ito, T. (1991), *Int. J. Radiat. Biol.* **59,** 643–650.

Koizumi, T., Itoh, Y., Sano, M., Kimura, M., Kojima, T. M., Kravis, S., Matsumoto, A., Oura, M., Sekioka, T., Awaya, A. (1995), *J. Phys.* **B28,** 609–616.

Landemark, E., Karlsson, C. J., Chao, Y.-C., Uhrberg, R. I. G. (1992), *Phys. Rev. Lett.* **69,** 1588–1591.

Lessmann, A., Brennan, S., Materlik, B., Schuster, M., Riechert, H. (1995), *Rev. Sci. Instrum.* **66,** 1428–1430.

Lindau, I., Spicer, W. E. (1974), *J. Electron Spectrosc. Rel. Phenom.* **3,** 409.

Lombardi, V., Piazzesi, G., Ferenczi, M. A., Thirlwell, H., Dobble, I., Irving, M. (1995), *Nature* **374,** 553–555.

Lucatorto, T., McIlrath, T. J., Suger, J., Younger, S. M. (1981), *Phys. Rev. Lett.* **47,** 1124–1128.

Lytle, F. W., Via, G. H., Smith, J. H. (1977), *J. Chem. Phys.* **67,** 3831.

MacFarland, C. A., Koenig, J. L., West, J. L. (1993), *Appl. Spectrosc.* **47,** 321–329.

Madden, R. P., Codling, K. (1963), *Phys. Rev. Lett.* **10,** 516.

Maeyama, S., Oshima, M., Shoji, T., Yamamoto, H. (1991), *Rev. Sci. Instrum.* **62,** 58–61.

Maeyama, S., Sugiyama, M., Heun, S., Oshima, M. (1995), *J. Cryst. Growth* **150,** 1122–1125.

Meade, C., Reffner, J. A., Ito, E. (1994), *Science* **264,** 1558–1560.

Menz, W., Bacher, W., Harmening, M., Michel, A. (1991), *Micro Electro Mechanical Systems* Nara, Japan, 30.1–2.2.1991, *Proc. IEEE,* Cat. Nr. 91CH2957-9, S.69.

Mitsui, Y., Senda, T., Shimazu, Matsuda, S., Utsumi, J. (1993), *Pharmacol. Therapeut.* **53,** 93.

Mizuno, H., Kano, H., Omura, T., Koizumi, M., Kondoh, M., Tsukihara, T. (1991), *J. Mol. Biol.* **219,** 665.

Momose, A., Takeda, T., Itai, Y., Hirano, K. (1996), *Nature Med.* **2,** 473–465.

Nara, Y., Sugita, Y., Horiuchi, K., Ito, T., Hirano, K. (1992), *Jpn. J. Appl. Phys.* **31,** 2333–2337.

Ng, W., Ray-Chaudhuri, A. K., Crossley, S., Crossley, D., Gong, C., Guo, J., Hansen, R., Margaritondo, G., Cerrina, F., Underwood, J. H., Perera, R. C. C., Kortwright, J. B. (1990), *J. Vac. Sci. Technol.* **A8,** 2563–2565.

Noma, T., Iida, A. (1994), *Rev. Sci. Instrum.* **65,** 837–844.

Ohsumi, K., Hagiya, K., Uchida, M., Suda, N., Miyamoto, M., Kitamura, M., Ohmasa, M. (1995), *Rev. Sci. Instrum.* **66,** 1448–1450.

Ohta, T., Suga, S., Kikuta, S. (1994), *Synchrotron Radiation Facilities in Asia,* Tokyo: Ionics Publishing Co.

Olson, C. G., Liu, R., Yang, A.-B., Lynch, D. W., Arko, A. J., List, R. S., Veal, B. W., Chang, Y. C., Jlang, P. Z., Paulikas, A. P. (1989), *Science* **245,** 731–733.

Oshima, M., Kawamura, T., Maeyama, S., Miyahara, T. (1988), *J. Vac. Sci. Technol.* **A6,** 1451–1455.

Oshima, M., Scimeca, T., Sugiyama, M., Maeyama, S., Oigawa, H., Nannichi, Y., Hashizume, H. (1993), *Appl. Surf. Sci.* **70/71,** 496–501.

Oyanagi, H., Sakamoto, K., Shioda, R., Sakamoto, T. (1994), *Jpn. J. Appl. Phys.* **33,** 3545–3552.

Patel, J. R., Golovchenko, J. A., Bean, J. C., Morris, R. J. (1985), *Phys. Rev. B* **31,** 6884–6886.

Pianetta, P., Lindau, I., Garner, C. M., Spicer, W. E. (1978), *Phys. Rev. B* **18,** 2792–2806.

Pianetta, P., Takaura, N., Brennan, S., Tompkins, W., Laderman, S. S., Fischer-Colbrie, A., Madden, M., Wherry, D. C., Kortright, J. B. (1995), *Rev. Sci. Instrum.* **66,** 1293–1297.

Rarback, H., Buckley, C., Goncz, K., Ade, H., Anderson, E., Attwood, D., Batson, O., Hellman, S., Jacobsen, C., Kern, D., Kirz, J., Lindaas, S., McNulty, I., Oversluizen, M., Rothman, S., Shu, D., Tang, E. (1990), *Nucl. Instrum. Methods* **A291,** 54–59.

Reffner, J., Martoglio, P. A., Williams, G. P. (1995), *Rev. Sci. Instrum.* **66,** 1298–1302.

Rubenstein, E., Giacomini, J. C., Gordon, H. J. (1990), *Nucl. Instrum. Methods Phys. Res.* **A291,** 80–85.

Sakabe, N. (1983), *J. Appl. Crystallogr.* **16,** 542–547.

Sakabe, N., Ikemizu, S., Sakabe, K., Higashi, T., Nakagawa, A., Watanabe, N., Adachi, S., Sasaki, K. (1995), *Rev. Sci. Instrum.* **66,** 1276–1281.

Sasaki, K., Sakabe, K., Sakabe, N., Kondo, H., Shimomura, M. (1993), *Acta Crystallogr.* **A49,** Suppl., 111–112.

Sato, F., Matsui, J., Kawata, H. (1985), *Int. Symp. GaAs and Related Compounds,* Karuizawa, 193 pp.

Schmahl, G., Rudolph, D., Guttman, P., Schneider, G., Thieme, J., Niemann, B. (1995), *Rev. Sci. Instrum.* **66,** 1282–1286.

Scimeca, T., Muramatsu, Y., Oshima, M., Oigawa, H., Nannichi, Y. (1991), *Phys. Rev. B* **44,** 12927–12932.

Shigemasa, E., Adachi, J., Oura, M., Yagishita, A. (1995), *Phys. Rev. Lett.* **74,** 359–362.

Sinfelt, J. H. (1987), *Acc. Chem. Res.* **20,** 134–139.

Spears, D. L., Smith, H. I. (1972), *Electron. Lett.* **8,** 102–104.

Stark, G., Smith, P. L., Ito, K., Yoshino, K. (1992), *Astrophys. J.* **395,** 705.

Stohr, J., Bauer, R. S., McMenamin, J. C., Johansson, L. I., Brennan, S. (1979), *J. Vac. Sci. Technol.* **16,** 1195–1199.

Stohr, J., Wu, Y., Hermsmeier, B. D., Samant, M. G., Harp, G. R., Koranda, S., Dunham, D., Tonner, B. P. (1993), *Science* **259,** 658–661.

Sugiyama, M., Maeyama, S., Oshima, M., Hashizume, H., Oigawa, H., Nannichi, Y. (1992), *Appl. Phys. Lett.* **60,** 3247–3249.

Sugiyama, M., Maeyama, S., Oshima, M. (1993), *Phys. Rev. B* **48,** 11037–11042.

Teo, B. K., Joy, D. C. (1986), *EXAFS Spectroscopy,* New York: Plenum Press.

Tweet, D. J., Akimoto, K., Tatsumi, T., Hirosawa, I., Mizuki, J., Matsui, J. (1992), *Phys. Rev. Lett.* **69,** 2236–2239.

Ueda, K., Umetani, K., Takeda, T., Akisada, M., Nakajima, T., Anno, I., Yamaguchi, C. (1989), *Rev. Sci. Instrum.* **60,** 2272–2275.

Uesugi, F., Nishiyama, I. (1992), *Appl. Surf. Sci.* **60/61,** 587–591.

Utsumi, Y., Akazawa, H., Nagase, M., Urisu, T., Kawashima, I. (1993), *Appl. Phys. Lett.* **62,** 1647–1649.

Vlieg, E., Fischer, A. E. M. J., van der Veen, J. F., Dev, B. N., Materlik, G. (1986), *Surf. Sci.* **178,** 36–46.

Waho, T., Seki, H. (1991), *Jpn. J. Appl. Phys.* **30,** 221–227.

Wakabayashi, K., Tanaka, H., Amemiya, Y., Fijishima, A., Kobayashi, T., Hamanaka, T., Shigi, H., Mitsui, T. (1985), *Biophys. J.* **47,** 847.

Williams, S., Jacobsen, C., Kirz, J., Maser, J., Wirick, S., Zhang, X., Ade, H., Rivers, M. (1995), *Rev. Sci. Instrum.* **66,** 1271–1275.

Woicik, J. C., Bouldin, C. E., Bell, M. I., Cross, J. O., Tweet, D. J., Swanson, B. D., Zhang, T. M., Sorenson, I. B., King, C. A., Hoyt, J. L., Pianetta, P., Gibbons, J. F. (1991), *Phys. Rev.* **B43,** 2419–2422.

Yagishita, A., Maezawa, H., Ukai, M., Shigemasa, E. (1989), *Phys. Rev. Lett.* **62,** 36–39.

Yamada, Y., Oshima, M., Waho, T., Kawamura, T., Maeyama, S., Miyahara, T. (1988), *Jpn. J. Appl. Phys.* **27,** L1196–L1198.

Further Reading

Chance, B. *et al.* (1994), *Synchrotron Radiation in the Biosciences,* Oxford: Oxford Science Publications.

Ebashi, S., Koch, M., Rubenstein, E. (1991), *Handbook on Synchrotron Radiation,* Vol. 4, Amsterdam: North-Holland.

Hasnain, S. S. (1990), *Synchrotron Radiation and Biophysics,* Chichester: Ellis Horwood Limited.

Helliwell, J. (1992), *Macromolecular Crystallography with Synchrotron Radiation,* Cambridge: Cambridge University Press.

Koch, E. E. (1983), *Handbook on Synchrotron Radiation,* Vols. 1A, 1B, Amsterdam: North-Holland Publishing.

Koch, E. E., Sasaki, T., Winick, H. (1987), *Handbook on Synchrotron Radiation,* Vol. 2, Amsterdam: North-Holland Publishing.

Koch, E. E., Sasaki, T., Winick, H. (1991), *Handbook on Synchrotron Radiation,* Vol. 3, Amsterdam: North-Holland Publishing.

Kunz, C. (1979), *Synchrotron Radiation,* Berlin: Springer.

Mandelkow, E. (1988), *Synchrotron Radiation in Chemistry and Biology II,* Topics in Current Chemistry, Vol. 147, Berlin: Springer-Verlag.

Prins, R., Koningsburger, D. C. (1988), *X-ray Absorption; Principles, Applications and Techniques of EXAFS, SEXAFS and XANES,* New York: Wiley.

Samson, J. A. R. (1967), *Techniques of Vacuum Ultraviolet Spectroscopy,* New York: Wiley.

Schmahl, G., Rudolph, D. (1984), *X-ray Microscopy,* Berlin: Springer.

Shinohara, K. *et al.* (1990), *X-ray Microscopy in Biology and Medicine,* Japan Scientific Societies Press, Berlin: Springer-Verlag.

Teo, B. K. (1986), *EXAFS Basic Principles and Data Analysis, Inorganic Chemistry Concepts,* Vol. 9, Berlin: Springer-Verlag.

Weissman, S., Balibar, F., Petroff, J. F. (1984), *Applications of X-ray Topographics Methods to Materials Science,* New York: Plenum Press.

Winick, H., Doniach, S. (1980), *Synchrotron Radiation Research,* New York: Plenum Press.

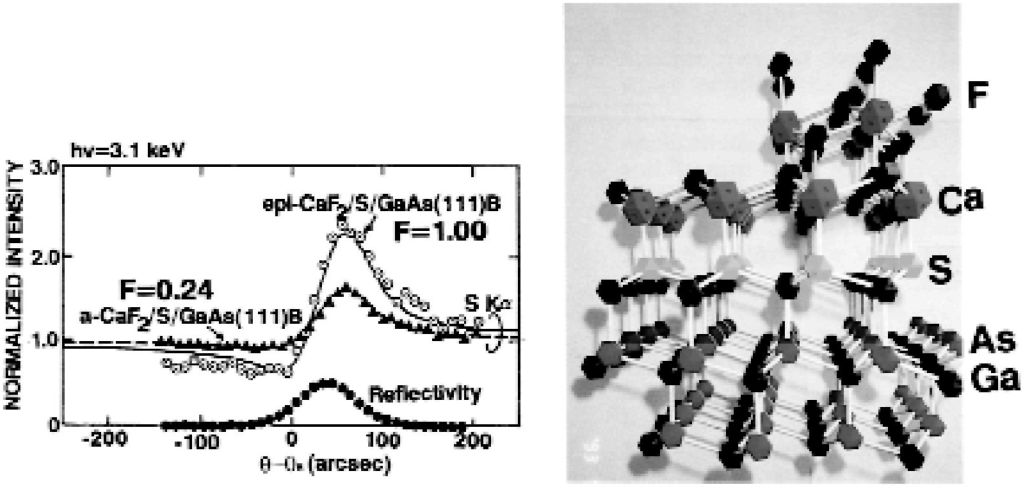

Vol. 20 – PLATE 1. XSW analysis of an atomic sulfur layer at the CaF$_2$/S/GaAs(111)B interface.

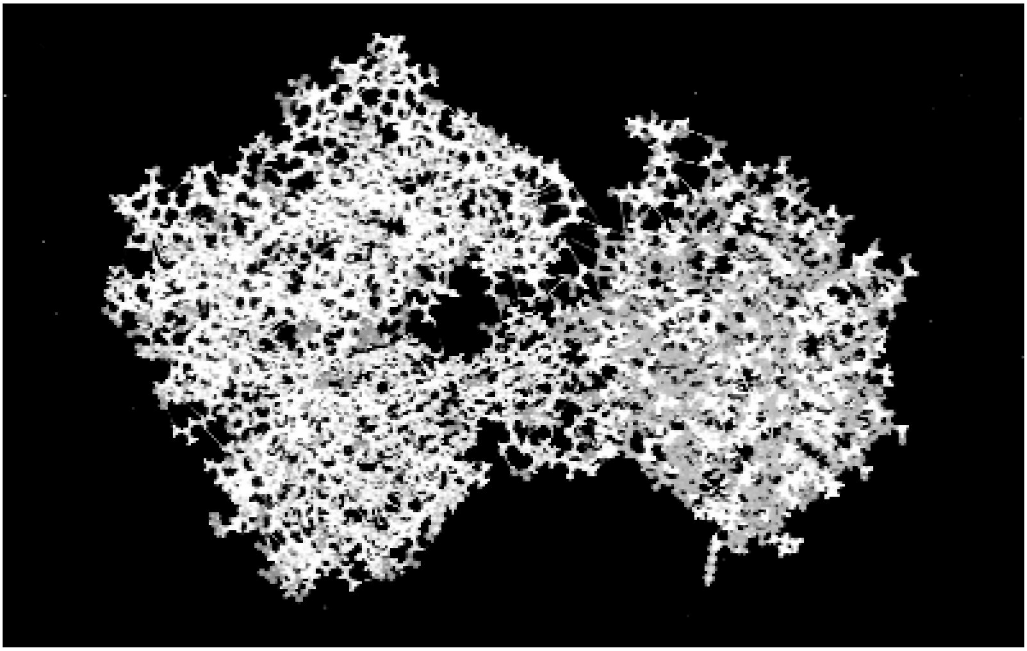

Vol. 20 – PLATE 2. A molecular structure model of actin-Dase 1 complex at 1.8-A resolution, which was derived from diffraction patterns detected by using the system shown in Figure 23.

18
Synchrotrons and Storage Rings

Hisashi Kobayakawa and Oliver Boine-Frankenheim

18.1 Introduction 563
18.2 Brief History 564
18.3 General Principles 566
18.3.1 Operating Principle 566
18.3.2 Particle Motion in a Circular Accelerator 568
18.3.3 Influence of Field Errors and Working Point on Resonance Diagram 571
18.3.4 Phase Stability 571
18.3.5 Momentum Compaction Factor, Transition Energy 573
18.3.6 Beam Injection and Extraction 574
18.3.7 Effects of Synchrotron Radiation 575
18.3.7.1 Radiation Damping and Quantum Fluctuation 575
18.3.7.2 Quantum Lifetime 577
18.4 Applications of Synchrotrons and Storage Rings 577
18.4.1 Synchrotrons and Storage Rings for Nuclear and for High-Energy Physics 577
18.4.2 Applications of Synchrotrons and Storage Rings for Materials Science and Others 578
18.4.2.1 Neutron Sources for Materials Science 578
18.4.2.2 Synchrotron-Radiation Sources 579
18.4.2.3 Radiotherapy 580
18.5 Technology and Engineering Aspects 580
18.5.1 Magnet System 580
18.5.2 Radio-Frequency Engineering 581
18.5.3 Computers and Automatic Controls 581
18.5.4 High-Vacuum Engineering 581
18.5.5 Civil Engineering 582
18.5.6 Superconducting System 582
Glossary 583
References 583
Further Reading 584

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

18.1
Introduction

The synchrotron is a particle accelerator, in which charged particles take a circular orbit and are accelerated to a high energy during circulation. Among other circular accelerators, the synchrotron method is, at present, unique in that it enables charged particles to be accelerated to energies of about 1 TeV. Both electron and proton synchrotrons have been mostly employed in basic research, especially in high-energy physics. They have also played essential roles in the advance of charged-particle acceleration techniques. The synchrotron is closely related to storage rings and the development of the colliding-beam method. Since the acceleration principle of storage rings is basically the same as that of the synchrotron, they have a similar configuration to that of synchrotrons. Beams are injected into a storage ring from a linear accelerator (linac) or a synchrotron after preacceleration. Synchrotrons accelerate the beam with a high repetition rate (up to 50 Hz). Storage rings and colliders can also accelerate, but at much slower rates.

Initially, storage rings were prominently used for particle colliders, in which two particle beams traveling in opposite directions are made to collide. The particles used in such experiments are electrons (positrons), protons (antiprotons), and heavy ions. Storage rings also serve as particle accumulators in which successive particle bunches from an injector are stacked so as to store relatively high-current beams. They are also used for beam cooling.

Although the synchrotrons, as well as other particle accelerators, are developed for the purpose of particle physics, the applications of synchrotrons have expanded to a wide variety of research fields, such as materials science, medical and biological sciences, and studies in radiotherapy. Besides basic research, synchrotrons are now also being used for industrial applications.

New technological developments have boosted the progress of accelerators, and there are now a great many technological spinoffs. For example, new large synchrotrons require high-field superconducting magnets that push further development of superconducting materials. The requirement of an extremely low pressure in storage-ring vacuum chambers having very large dimensions has greatly stimulated high-vacuum technologies. The accurate control of a large number of accelerator components was followed by progress in computer technology,

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

which in turn stimulated progress in these fields.

18.2
Brief History

As the acceleration energy increased, accelerator technologies have inevitably encountered practical limitations in both size and cost. The linac has reached some practical construction limit because of its length. Lawrence (1930) proposed a circular-acceleration method that can conveniently overcome this problem of length by accelerating particles in a circular path maintained by magnetic fields. Charged particles move in a spiral orbit in a homogeneous magnetic field such that they are accelerated many times while traversing an accelerating gap. Since the revolution period is independent of the particle energy in the nonrelativistic approximation, the frequency of the acceleration field is kept constant. This machine, called the *cyclotron*, was first built by Livingston (1931), who demonstrated the principle by accelerating hydrogen ions to 80 keV.

However, as the particle energies increased further, the normal cyclotron method reached its limit, because relativistic effects caused the synchronism of the accelerating field to be lost. A *synchrocyclotron* is an accelerator in which a synchronous acceleration method is applied to a cyclotron. Many such machines were built for energies in the 100–1000 MeV range. The cyclotron, or synchrocyclotron, is arranged so as to provide a constant magnetic field over relatively large areas, which allow the particles to travel in a spiral orbit of gradually increasing diameter as they increase in energy.

Kerst and Serber (1941); Kerst (1941, 1942) proposed a circular accelerator called a *betatron*, which utilizes a time-varying magnetic field that, in turn, produces an electric field for acceleration through induction, which has been used only for the acceleration of electrons. The betatron comprises a doughnut-shaped vacuum chamber for the beam path located between the poles of an electromagnet. The magnetic field acts to bend the electron beams in a circular orbit and to induce a voltage that accelerates the electron tangentially to the orbit. They are limited to acceleration not much above 300 MeV.

McMillan (1945) and Veksler (1945) independently discovered the principle of *phase stability* and invented the *synchrotron*. In the synchrotron, the guide field increases with particle energy so that the beam takes a circular orbit, and acceleration takes place owing to a radio-frequency (rf) electric field, rather than an induction field, as in the case of betatrons.

In early synchrotrons, the focusing fields were weak quadrupole fields superimposed on dipole magnetic fields; this is called the *weak-focusing* method. The first such synchrotron, the 3-GeV Cosmotron at Brookhaven National Laboratory (BNL), was built in 1952; the largest of these is the 10-GeV Syncrophasotron at Dubna.

Modern synchrotrons use a *strong-focusing* method, usually by means of separate quadrupole magnets that constrain the particles to a narrow path. A pair of such magnets exerts a net focusing force in both the vertical and horizontal directions. These strong-focusing machines are known as *alternating-gradient* accelerators. The method of the alternating-gradient synchrotron (AGS) was discovered by Christofilos (1950), and independently by Courant, Livingston, and Snyder (1952); see also Courant and Snyder (1958). The invention of the strong-focusing method resulted in a real breakthrough

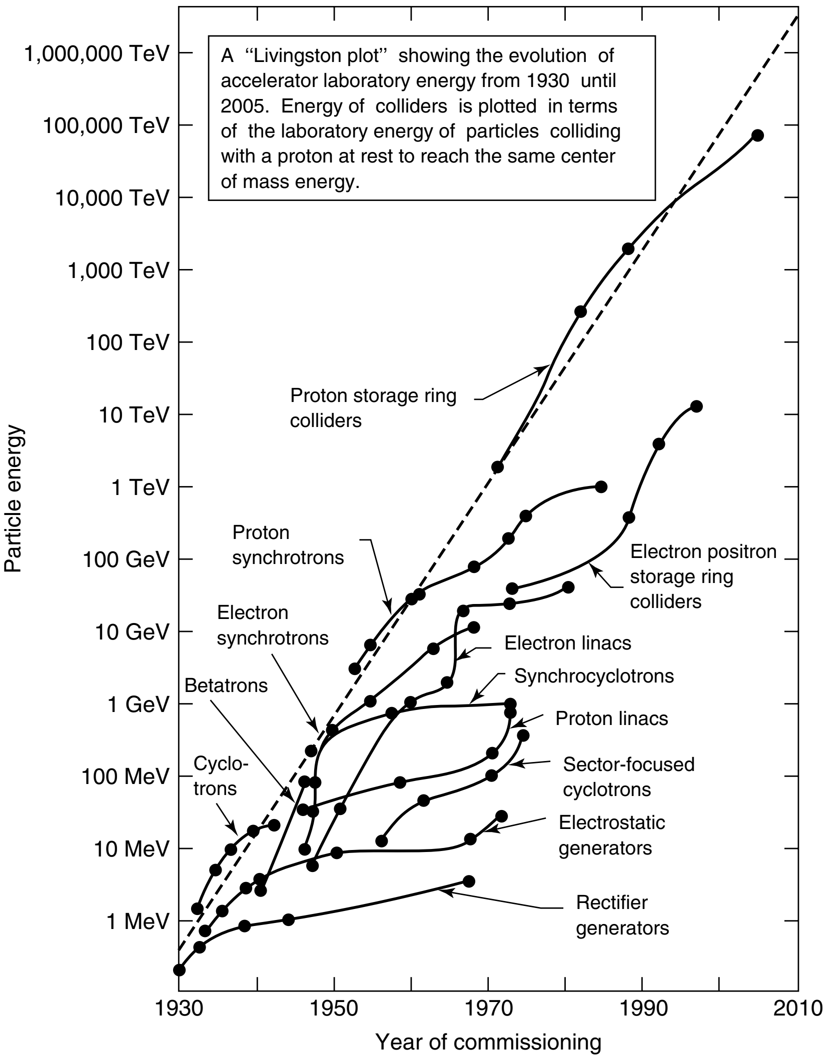

Fig. 18.1 "Livingston chart" showing the maximum energy as a function of the year for various accelerator technologies.

in synchrotrons. In this method, since particles are able to travel in a narrow beam chamber, the sizes as well as the construction costs of the synchrotron magnets became considerably reduced. This overcame the practical limitations of the particle energy and pushed machines to higher energies; synchrotrons having energies above 10 GeV could thus be constructed. The AGS at BNL was the first synchrotron using this method and was built in 1961. In this machine, protons are accelerated up to 33 GeV. The first two synchrotrons of this type were *combined-function* accelerators; the bending magnets had additional quadrupole-field components. Modern synchrotrons are built with *separate-function* type magnets. Simple dipole magnets are used for bending the particle path, and quadrupole magnets are used for focusing the beam. The two largest synchrotrons for protons are the Large Hadron Collider (LHC) (7 TeV) at European Laboratory for Particle Physics (CERN) and the Tevatron (900 GeV) at Fermi National Accelerator Laboratory (FNAL).

For high-energy physics experiments, accelerated particles in a synchrotron are extracted and bombarded onto atoms in a fixed target. The efficiency of the energy used for this process rapidly decreases as the particle energy increases. This is because a large part of the energy of a particle colliding with a fixed atom is wasted in the forward motion of both particles, that is, in the direction of motion of the incident particle; a small part is converted into the relative motion of the two particles.

On the other hand, in the *colliding-beam* method the efficiency of the accelerated-particle energy utilization has been drastically increased, particularly in very high energy collisions. Two independent bunches of particles accelerated in synchrotron(s) are injected into storage ring(s); those bunches travel in opposite directions and make head-on collisions.

Figure 18.1 is a ‘‘Livingston chart,’’ which shows the maximum energy as a function of the year for various accelerator technologies. Cockcroft and Walton accelerated protons to about 400 keV using a static high-voltage (700 keV) generator (1932). Since then, dramatic progress has been achieved in various accelerator technologies. The inventions of the linac and circular-acceleration methods, such as cyclotrons, synchrocyclotrons, betatrons, and synchrotrons, have brought about great progress in particle energy. The maximum particle energies increased by more than one and a half orders of magnitude during one decade (in storage-ring colliders, the energies are given as the energies that would be needed in a fixed-target machine to provide the same available interaction energy).

Although synchrotrons and storage rings with very high energies belong almost exclusively to the field of high-energy physics, they have played a prominent role in utility applications. The development of intensive neutron sources and synchrotron-radiation sources, which are becoming ever more important in various fields of materials science and industrial technologies, has been made possible by advances in synchrotron and storage-ring design.

18.3
General Principles

18.3.1
Operating Principle

Figure 18.2 is a schematic illustration of a synchrotron. Charged particles are injected into a synchrotron chamber provided

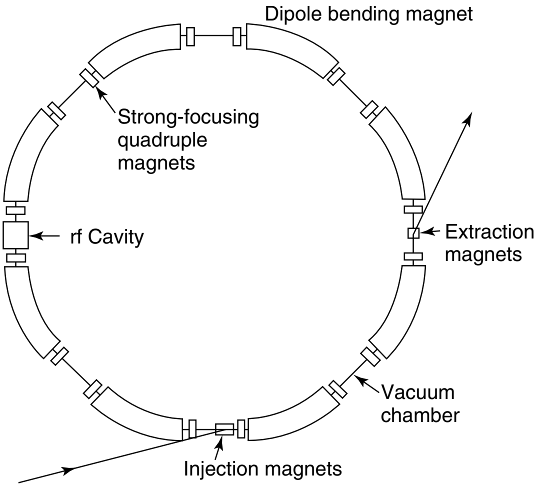

Fig. 18.2 Schematics of a synchrotron accelerator.

with a number of electromagnets to bend the particle path by the Lorentz force. The injection beams come from a preaccelerator, such as a linac. In rf cavities, acceleration takes place with electric fields supplied from rf-power sources, which commonly operate at between several and several hundred megahertz. In a synchrotron, particles take an orbit having a constant radius, like in a betatron, nominally at the center of the beam pipe around the circumference of the synchrotron.

Figure 18.3 shows the operating cycle of a synchrotron. After injection is completed, the dipole magnetic field (B) increases linearly, thus causing the particles to stay in a constant orbit as their energy (E) increases. Here, E is the total energy ($E^2 = (pc)^2 + (mc^2)^2$, where p is momentum of a particle, m its rest mass, and c the speed of light). Since protons or ions, which are much heavier than electrons, gain velocities relatively slowly, frequency modulation of the rf accelerating voltage has to be applied to make certain that the accelerating voltage and frequency of particle revolution are synchronous. Therefore, during a single acceleration cycle (when particles take a circular orbit), both the magnetic field (B) and the rf frequency (f) increase as

$$f = \frac{ec^2 B}{2\pi E} \tag{1}$$

Since in an electron synchrotron the particles are injected at a kinetic energy higher than some tens of megavolts, they always have a velocity close to the speed of light. As a consequence, the revolution frequency of the particles is almost constant, and it is possible to use a fixed-frequency accelerating cavity. This is one of the primary differences between electron synchrotrons and proton synchrotrons, where the frequency of the accelerating rf fields must be adjusted during acceleration by a factor of 10 or more, because the velocity of the much heavier protons at typical injection energies is much slower than that of light.

At the end of the accelerating cycle (Figure 18.3), accelerated particles are either directed at an internal target in the accelerator chamber or are extracted outside and guided to experimental

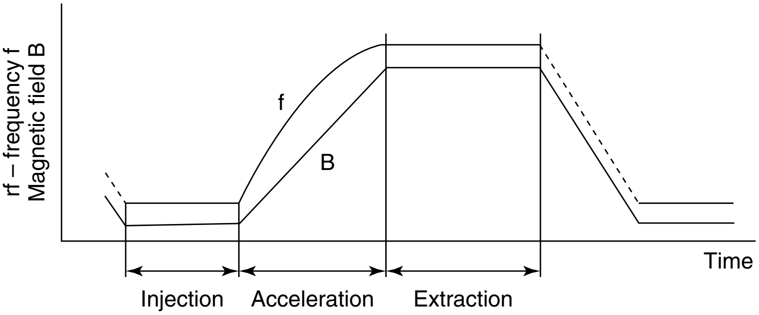

Fig. 18.3 Typical operating cycle of the magnetic field and acceleration frequency of a proton synchrotron.

stations, or to additional accelerators. In a large proton synchrotron, the cycle lasts for between a few seconds and a few tens of seconds.

18.3.2
Particle Motion in a Circular Accelerator

Practically, since circulating particles have a small transverse momentum, they oscillate around an ideal circular orbit with a frequency determined by the focusing force of the magnetic fields. This motion is known as the *betatron oscillation*, and the frequency is called the *betatron frequency*.

We assume that the circular accelerator has a median plane, taken to be the horizontal plane ($y = 0$). The bending fields are perpendicular and the focusing fields are perpendicular or parallel to this plane. There is a closed orbit in this plane, called the *equilibrium orbit*, on which a particle with reference momentum p_0 can circulate. The trajectory of an arbitrary particle is characterized by its coordinates: x, y, and s, defined by Figure 18.4

The equation of motion of $x(s)$ and $y(s)$ can be obtained from the Lorentz force. If all terms of higher order than the first are neglected (linear approximation), the equations take the form

$$\frac{d^2x}{ds^2} - \left[k(s) - \frac{1}{\rho^2(s)}\right]x = \frac{1}{\rho(s)}\frac{\Delta p}{p_0},$$
$$\frac{d^2y}{ds^2} + k(s)y = 0 \qquad (2)$$

Here, $\rho(s)$ is the radius of curvature due to the dipole fields, and $k(s)$ defines the focusing strength due to the quadrupole fields:

$$k(s) = -\frac{1}{B\rho}\frac{dB_y}{dx} \qquad (3)$$

An important property of the quadrupole magnet is that when the force is focusing in the horizontal plane, it is defocusing in the vertical plane, or vice versa. Note that both $k(s)$ and $\rho(s)$ are periodic functions of s; $\Delta p/p_0$ is the momentum error of the particle. While discussing betatron oscillations, we put $\Delta p/p_0 = 0$. In this case, both equations have the same form:

$$\frac{d^2u}{ds^2} + K(s)u = 0 \qquad (4)$$

where $K(s)$ depends on the plane under consideration and $u(s)$ means $x(s)$ or $y(s)$. Since the equilibrium orbit is closed, $K(s)$

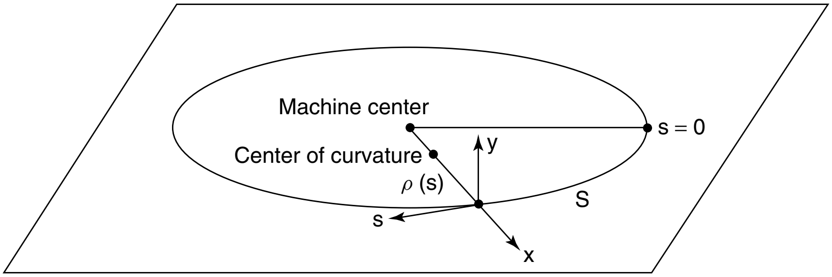

Fig. 18.4 Coordinate system for description of beam.

is a periodic function with a period of the machine circumference (L),

$$K(s+L) = K(s) \tag{5}$$

(If the machine has N identical sections of length $l(L = Nl)$, $K(s)$ satisfies the periodic condition $K(s + l) = K(s)$.)

A particle trajectory is described by a solution of the differential equation (4), which is called *Hill's equation*. According to standard convention, we use the general solution in the form

$$u(s) = A\sqrt{\beta(s)}\cos[\psi(s) + \delta] \tag{6}$$

Substitution of the general solution into the differential equation gives a relation between $\beta(s)$ and $\psi(s)$:

$$\psi(s) = \int_0^s \frac{ds}{\beta(s)} \tag{7}$$

Here, $\beta(s)$ is called the *β-function*, which is interpreted as the local wavelength of the oscillation divided by 2π. In particular, the number of oscillations per turn,

$$\nu = \frac{1}{2\pi}\int_0^L \frac{ds}{\beta(s)} \tag{8}$$

is called the *betatron tune* of the accelerator.

Most high-energy accelerators use a periodic sequence of quadrupole magnets of alternating polarity. Such configuration is called a FODO lattice, in which charged-particle beams are focused or defocused by quadrupole magnet fields. An example of the β-function and particle trajectories is shown in Figure 18.5 in a regular FODO lattice, in which quadrupole are arranged repeatedly, such as equally spaced focusing and defocusing lenses.

Each time that particle passes position s in the ring, its betatron oscillation appears as a point on an ellipse, as shown in Figure 18.6, where $u(s)$ and $du(s)/ds$ are plotted in the abscissa and ordinate, respectively. Since the area of the ellipse is invariant when the particle moves around the accelerator, we obtain the important result that the area of the ellipse enclosing the beam in phase space is invariant (*Liouville's theorem*). This is true under the condition that stochastic effects like beam–gas scattering or radiation can be neglected. The area ($\pi\varepsilon$) is an important quantity, and ε is called the *emittance*.

Practically, particles in a synchrotron have some energy spread, that is, $\Delta p/p_0$ is not zero. In this case, the motion of a particle is described by the inhomogeneous Hill's equation for the horizontal motion (Eq. (2)). Since the particle orbit in bending magnets is determined by the momentum of each particle, there is a radial spread of particles in the beam. The shift in the radial position from the central orbit is related to

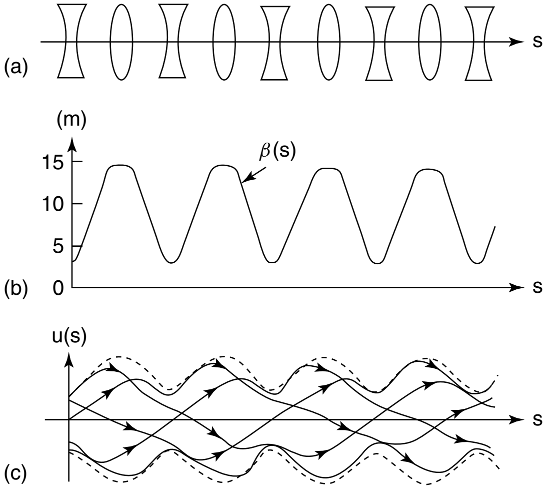

Fig. 18.5 (a) A regular FODO lattice of focusing and defocusing lenses. (b) β-Function. (c) One trajectory on several revolutions (according to Sands (1970)).

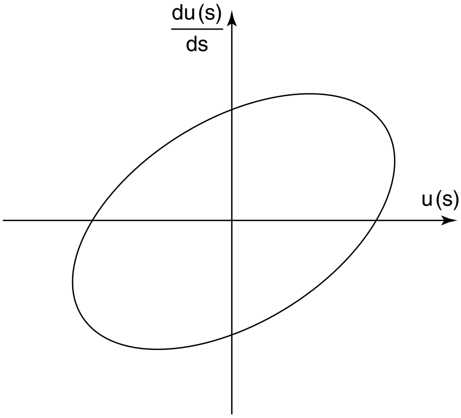

Fig. 18.6 Phase-space ellipse at point *s*.

the momentum deviation as

$$\Delta x(s) = \eta(s)\left(\frac{\Delta p}{p}\right) \tag{9}$$

Here, $\eta(s)$ is called the *dispersion function*.

Although a more rigorous treatment has to be made, the transverse beam size and divergence can be derived using the β-functions, dispersion functions, and emittances in both planes. In electron (positron) storage rings, although the emission of synchrotron radiation determines the beam emittance (Section 18.3.7), the beam size $(\sigma_x(s), \sigma_y(s))$ can be written as

$$\sigma_x(s) = \left[\beta_x(s)\varepsilon_x + \eta(s)^2\left(\frac{\Delta p}{p}\right)^2\right]^{1/2}$$

and

$$\sigma_y(s) = \left[\beta_y(s)\varepsilon_y\right]^{1/2} \tag{10}$$

When the particles are accelerated, the emittance decreases inversely proportional to the momentum. This can be interpreted as meaning that only the longitudinal component of the momentum vector is increased during acceleration, whereas the transverse components remain invariant, so that the angular divergence decreases (*adiabatic damping*).

Here, we define the *normalized emittance* as $\varepsilon_n = (p_0/mc)\varepsilon$. For proton and heavy-ion beams, and for electron beams in linacs, the normalized emittance stays constant during acceleration. It even remains invariant when the beam is passed from an ion source through a whole chain of preaccelerators into a storage ring, provided the transitions are properly matched.

In electron (positron) storage rings, the emission of synchrotron radiation and the subsequent reacceleration are the dominant factors in determining the beam dimensions. In an ideal flat ring with complete decoupling of horizontal and vertical motion, the vertical emittance tends to zero while the horizontal emittance grows with the square of the energy.

18.3.3
Influence of Field Errors and Working Point on Resonance Diagram

In practical machines, bending magnets and quadrupole magnets are not ideal but have some field imperfections, and they contain more higher order field components due to sextupole and octupole magnets (the role of sextupole and octupole magnets is described in, for example, Rossbach and Schmüser, 1994). These higher order field components disturb ideal betatron motions. Dipole errors lead to amplitude growth of betatron oscillations at the point where the betatron tunes ν_x, ν_y (defined in the horizontal and vertical planes, respectively) are integer values, quadrupole errors excite at half-integer values, and sextupoles at one-third-integer values. One has a coupling between the horizontal and vertical betatron oscillations due to a misalignment of the quadrupoles or sextupoles. In more general terms, if tunes are based on a relation such as

$$l\nu_x + m\nu_y = n \quad (l, m, n = 0, \pm 1, \pm 2, \ldots) \tag{11}$$

the amplitudes of the oscillations can become large. This numerology defines the so-called *betatron resonance*. Figure 18.7 is a resonance diagram; one has to choose the working point (ν_x, ν_y) with a reasonable separation from the most dangerous resonance lines.

18.3.4
Phase Stability

The acceleration field in the rf cavity must be in the accelerating phase every time a particle bunch arrives. Thus, the time taken for a bunch to make one cycle around the ring circumference (revolution period of T) must be an integral multiple of the rf period. The voltage in the rf cavity is

$$V(t) = V\sin(2\pi ft) \tag{12}$$

Each particle gains an energy of ΔE, where

$$\Delta E = eV\sin\varphi \tag{13}$$

Here, φ is the phase of the rf voltage at the moment of particle passage through the cavity and is called the *acceleration phase*.

To maintain a particle on the same orbit, the magnetic field (B) is varied as

$$\frac{dB}{dt} = \frac{\Delta E}{2\pi e\rho R} \tag{14}$$

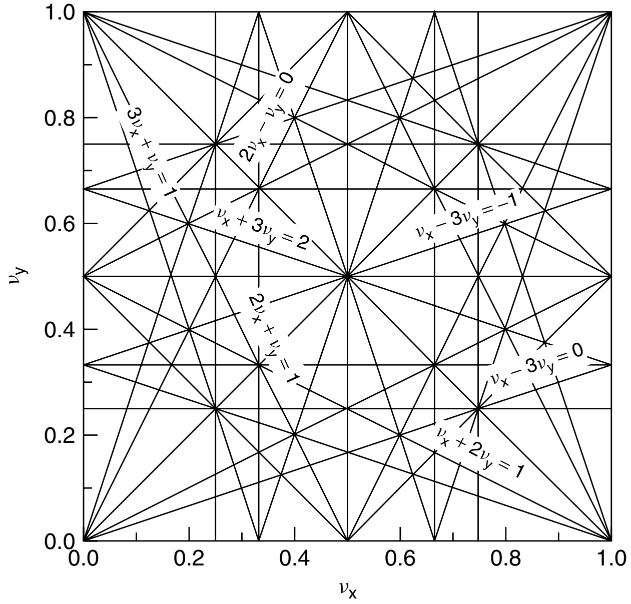

Fig. 18.7 Resonance lines in a unit square of betatron tune plane for the fourth integer and below.

where ρ is the bending radius ($\rho = p/eB$) and R is the average radius of the ring ($R = L/2\pi$; here, L is the length of the nominal orbit in circumference). The acceleration phase that satisfies these two equations is called the *equilibrium phase* or *synchronous phase*. In a synchrotron, particles oscillate in phase around the synchronous phase (φ_s) because of the focusing property of the phase (*phase stability*); they also oscillate in energy around the central energy (E_s). This oscillation in both phase and energy is called *synchrotron oscillation*.

The voltage in the accelerating gaps generally has a sinusoidal variation with time, as Eq. (12). When the equilibrium phase (φ_s) is between $\pi/2$ and π, it means that rf voltage decreases with time, and a particle at point A having an energy of E_s of which the phase (φ) is ahead of φ_s acquires more energy than at the equilibrium-phase point S ($\Delta E > \Delta E_s$), as shown in Figure 18.8(a). Since a particle having a higher energy takes a larger orbit radius and takes a longer time for one revolution, the particle at A tends to come close to point S and then go across it. The phase (φ) is behind φ_s in this case; the particle, in turn, obtains less energy than at point $S(\Delta E < \Delta E_s)$, and the particle energy tends to decrease until below E_s. Then, the phase (φ) begins to return to point A. This means that particles are not very far apart from the equilibrium phase and energy (φ_s and E_s) during acceleration. A particle takes, roughly speaking, an elliptical trajectory (Figure 18.8(b)) on the phase–energy plane (φ–E plane). Figure 18.8(c) shows the particle trajectory in longitudinal phase space in more detail. A curve that crosses

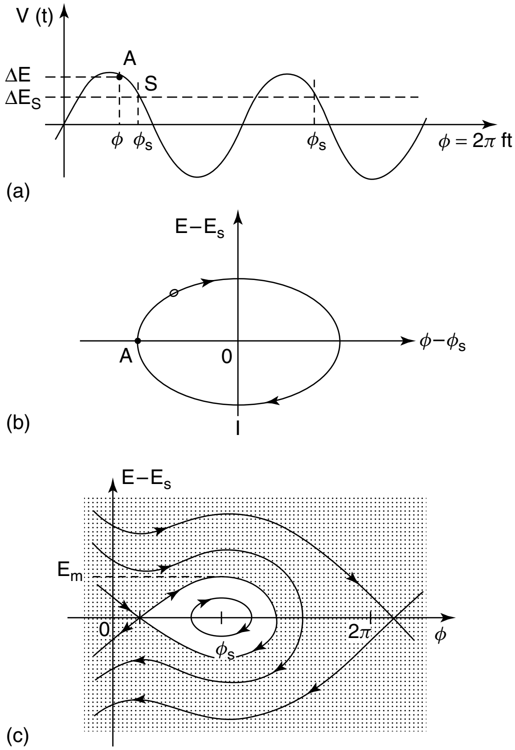

Fig. 18.8 (a) Accelerating voltage and particle phase. (b) Particle trajectory on the $\phi - E$ plane (longitudinal phase space). (c) Particle trajectory of the synchrotron motion above the transition energy in longitudinal phase space in more detail. In the shaded region, particles are unstable.

an unstable point (shaded region in the figure) is called a *separatrix*. The stable region surrounded by the separatrix is called an *rf bucket*, and the energy (E_m) of the rf bucket is called the *bucket height*. These explanations are restricted to the case that a particle energy is above the transition energy (see the next section). Below the transition energy, similar discussions are made when the equilibrium phase (φ_s) is between 0 and $\pi/2$.

18.3.5
Momentum Compaction Factor, Transition Energy

There is a closed orbit for every value of the particle momentum p; this orbit is called the *equilibrium orbit* in each momentum. The length L of the equilibrium orbit during one revolution varies with the momentum; the relation between the deviation in momentum from p and that in the equilibrium orbit length is written as

$$\frac{\Delta L}{L} = \frac{\alpha \Delta p}{p} \tag{15}$$

where the constant α is called the *momentum compaction factor*, which gives the dependence of the equilibrium orbit length on the momentum of the particles. In standard circular accelerators, α is positive. Although the value depends on the focusing methods, it is approximately given as the inverse square of the horizontal betatron tune. In weak-focusing synchrotrons $\alpha > 1$, and in strong-focusing synchrotrons $\alpha < 1$.

The revolution period T equals L/v, where v is the particle velocity. Therefore, the relation among small deviations in these quantities is given by

$$\frac{\Delta T}{T} = \frac{\Delta L}{L} - \frac{\Delta v}{v} \tag{16}$$

Using the relation

$$\frac{\Delta v}{v} = \left(\frac{1}{\gamma^2}\right)\frac{\Delta p}{p} \tag{17}$$

between the velocity and momentum (here, $\gamma = E/mc^2$), we obtain

$$\frac{\Delta T}{T} = \left(\frac{\alpha - 1}{\gamma^2}\right)\frac{\Delta p}{p} \tag{18}$$

This is an important relation for discussing the stability in rf acceleration. At low energies, $\alpha - 1/\gamma^2$ is negative, and the revolution period T decreases as the particle momentum p increases. On the other hand, when the particles exceed a certain energy, called the *transition energy*, $\alpha - 1/\gamma^2$ becomes positive. Then, the relation between T and p becomes reversed to that mentioned above; as the momentum becomes higher, a particle takes a longer time to undergo one revolution. During the acceleration process, at the moment that particles pass the transition energy, the value of the equilibrium phase suddenly makes a big change. At the transition energy, the phase stability condition is unsatisfied, the bunch length becomes short, and the momentum spread increases. The transition energy is approximately given as the horizontal betatron tune times the rest mass of the accelerating particle.

This phenomenon occurs only in strong-focusing proton (or ion) synchrotrons, because weak-focusing machines have $\alpha > 1$, and, therefore, $\alpha - 1/\gamma^2$ is always positive. For electron synchrotrons, particles are usually injected at energies higher than the transition energy.

18.3.6
Beam Injection and Extraction

A linac or microtron is usually used as a preaccelerator to inject particles into the synchrotron. In the first stage of the preaccelerator, ion sources or electron guns are used for the particle sources. Injection usually takes place at an energy of tens or hundreds of mega electronvolts. The field strength of the injection magnets is chosen so as to make the equilibrium orbit of the injection particle energy coincide with the central orbit of the synchrotron. The design aims of the injection system are to achieve transfers with little beam loss and with a minimum dilution of the beam emittances. The area of the region in phase space that can be used for injection of the accelerator is called the *acceptance*, which is determined by both the physical aperture (such as the width of the beam pipes) and the orbit parameters. In practice, errors in the magnetic fields and nonlinear oscillations cause a reduction of the acceptance (*dynamic aperture*).

Single-turn injection schemes are relatively straightforward (Figure 18.9). The beam is injected onto a central orbit through a septum unit and a fast kicker element with appropriate matching arranged in both the transverse and longitudinal planes. There should be no beam loss and very little emittance dilution. For the longitudinal plane, the momentum spread and phase parameters of the incoming bunch must be matched to those of the bucket defined by the rf fields of the ring.

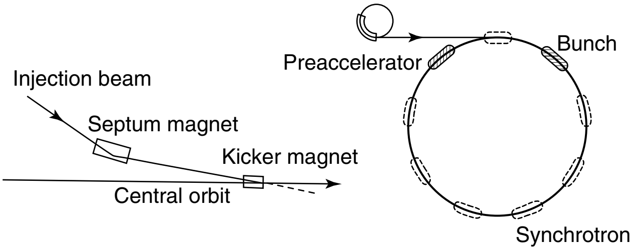

Fig. 18.9 Example of injection scheme.

In a *multi-turn injection,* the beam is injected for a period of time longer than the revolution time, thus stacking the beam in many turns. The process for electron accelerators is significantly different from that for proton or heavy-ion machines. The electron rings take advantage of *radiation-damping* effects to reduce the emittance of the beam. Proton and heavy-ion machines generally operate at a beam emittance that is set by space-charge effects. The hardware required for multiple-turn injection is not greatly different from that used for single-turn injection, but the role of the fast kicker is replaced by that of orbit bump magnets.

Proton machines have utilized a type of multiple-turn scheme known as *charge-exchange injection*, which requires micron or submicron stripping foils.

Particles in the synchrotron are most often extracted and led to the experimental areas outside the ring, or led to other machines, such as storage rings. *Fast extraction* is commonly used for beam transfer between synchrotrons and storage rings. Special kicker magnets remove individual bunches from the ring, or distort the beam orbit so that the entire beam can be extracted during a single turn.

There is also a *slow-extraction* process, requiring one to several seconds, and called *resonant extraction*. In this process, the dipole magnetic field is usually maintained constant. Then, the quadrupole magnets are tuned until the betatron oscillations are close to resonance. At the same time, a local perturbation field is applied, often using a sextupole magnet, and then varied slowly to sweep the beam through the resonance. During this sweep, particles are brought continuously into resonance, resulting in a controlled growth of the beam size. The extraction then takes place using a septum to scrape the beam.

18.3.7
Effects of Synchrotron Radiation

18.3.7.1 Radiation Damping and Quantum Fluctuation

The properties of synchrotron radiation can be derived by applying the methods of classical electrodynamics to the motion of relativistic charged particles in circular orbits. The theory of synchrotron radiation is based on the work of Iwanenko and Pomeranchuk (1944) as well as Schwinger (1946). Because of the relativistic speed of the particles, the angular distribution of the synchrotron radiation is strongly concentrated into a narrow cone in the instantaneous direction of motion. The spectrum extends from the microwave region to the hard X-ray range. The Lorentz transformation of the radiation pattern from the rest frame of the electron into the relativistically moving frame leads to a distribution of the intensity into higher energy.

The energy radiated in one revolution by a monoenergetic relativistic charged particle in a circular orbit (bending radius ρ) is expressed as

$$U_0 = \frac{e^2}{3\varepsilon_0}\frac{\beta^3\gamma^4}{\rho} \tag{19}$$

where $\beta = v/c$ and $\gamma = E/mc^2$. The total radiated power also depends on γ^4, which explains why electrons are used to produce synchrotron radiation. Electron machines require a large radius to obtain higher energies to keep the radiated power at a reasonable value. For example, the Large Electron Positron (LEP) collider at CERN was designed with a maximum energy of more than 100 GeV and a circumference of 27 km. It can also be seen that the emission of synchrotron radiation from protons is very small for energies up to several hundred giga electronvolts.

Synchrotron radiation is emitted primarily along the direction of motion into a narrow cone within an approximate angle of $1/\gamma$. The reaction force and the accompanying momentum change are opposite to the direction of motion. The only effect of the radiation is then to decrease the energy of the electron without changing its direction of motion. Energy losses are replaced by rf acceleration. As a consequence of the acceleration, the amplitudes of the synchrotron oscillation and the betatron oscillations are damped. The processes are known as *radiation damping* in electron machines.

Damping of the Energy Oscillations Small energy oscillations, synchrotron oscillations, are damped at a rate proportional to the change in the radiation loss with the energy. Since the energy loss per turn (U_0) due to synchrotron radiation depends on E^4, a particle having higher energy radiates more, and one having lower energy radiates less than the ideal particle. Therefore, in both cases the energy deviation tends to be reduced. The phase oscillation is also damped because of their conjugate relation. We assume that the amplitude decays as $\exp(-\lambda t)$, where λ is a damping coefficient approximately written as U_0/T_0E_0. (For rigorous discussions one should consult the literature – for example, Sands (1970).)

Damping of the Betatron Oscillations The emission of a photon implies a momentum loss in both the transverse and longitudinal directions. The transverse oscillations (vertical and horizontal betatron oscillations) are damped because the rf acceleration compensates for energy loss with an rf field in the longitudinal direction. The damping coefficients of the betatron oscillations in both the vertical and horizontal directions are approximately written as $U_0/2T_0E_0$.

Quantum Fluctuation and Beam Sizes Until now, we have considered only the total energy loss due to synchrotron radiation, assuming implicitly that the energy loss is a continuous process. However, all electromagnetic radiation occurs in quanta of discrete energy. This quantization of the energy loss has significant effects on the behavior of electron beams.

Each time a quantum is emitted, the energy of the electron makes a small discontinuous jump. This discontinuous energy change disturbs the trajectory of the electron (*quantum fluctuation* or *quantum* *excitation*). Such disturbances are a kind of random noise, which causes their amplitudes to grow until the quantum excitation is, on the average, balanced by radiation damping of the oscillations. The beam sizes in the transverse and longitudinal directions, beam emittances, and spread in energy are thus determined in an equilibrium state.

18.3.7.2 Quantum Lifetime

The transverse distributions of particles in a stored beam can be assumed to have a Gaussian form and, thus, a tail that extends to infinity. The beam is actually limited at least to the vacuum chamber, which causes particle loss. When the particle loss is small, the loss rate is proportional to the number of particles, N. Therefore, N decreases exponentially with a time constant τ_q according to

$$\frac{1}{\tau_q} = -\frac{1}{N}\frac{dN}{dt} \tag{20}$$

The constant τ_q is referred to as the *quantum lifetime* and, in an ideal case, is given as

$$\tau_g = \frac{\tau}{2}\frac{\exp\xi_q}{\xi_q} \tag{21}$$

where τ and ξ_q are given values for the transverse and longitudinal directions. The quantum lifetime, which depends on the transverse direction, is related to the betatron oscillations; τ is the damping time of the betatron oscillation (τ_x or τ_y), and

$$\xi_q = \frac{1}{2}\left(\frac{a}{\sigma}\right)^2 \tag{22}$$

where a is the aperture size and σ (σ_x or σ_y) is beam size in the horizontal or vertical direction, respectively.

A bunch loses particles through longitudinal motion in the rf bucket. The lifetime due to synchrotron oscillation is obtained by putting τ for the damping time of the synchrotron oscillation, and

$$\xi_q = \frac{1}{2}\left(\frac{E_m}{\sigma_E}\right)^2 \tag{23}$$

where E_m is the bucket height and σ_E the standard deviation in the particle energy.

18.4
Applications of Synchrotrons and Storage Rings

18.4.1
Synchrotrons and Storage Rings for Nuclear and for High-Energy Physics

For medium-energy or high-energy experiments, accelerated protons or heavy ions are most often extracted from a synchrotron and transported to the experimental stations. The beam particles bombard various fixed targets where interactions among particles are studied or where beams of secondary particles are generated. Secondary beams, such as antiprotons or radioactive ions are also guided to the experimental areas and are fed to fixed targets. Beams of neutrinos and muons, which are decay products of π-mesons, are also important tools for basic research.

At present, storage-ring colliders dominate the high-energy field. For particle–antiparticle collisions, one can use a single ring, where they circulate in opposite directions in the same magnetic guide field. For proton–proton collisions or electron–proton collisions, a double ring must be used. The largest colliders are the LHC at CERN, Tevatron at FNAL and Relativistic Heavy Ion Collider RHIC at BNL. The technologies of *beam cooling* were developed for such facilities and are now being applied in smaller storage rings. The large synchrotrons and storage-ring colliders presently in operation are listed in Table 18.1.

The development of heavy-ion accelerators is closely related to that of proton accelerators. Most heavy-ion accelerators, including cyclotrons, linacs, and electrostatic machines, are used to accelerate both protons and heavier nuclei. The acceleration principle of heavy-ion synchrotrons is exactly analogous to that of protons. However, since the velocity of ions is slow during injection, the frequency range of the rf system must be wide to maintain synchronous acceleration as the ion velocity increases. The beam-monitoring system must be very sensitive because of the low beam intensity. Since the charge state of stripped ions changes during collisions with the residual gas molecules in the beam pipe, which leads to a beam loss, an extremely good vacuum is required, typically 10^{-8}–10^{-10} Pa.

18.4.2
Applications of Synchrotrons and Storage Rings for Materials Science and Others

18.4.2.1 Neutron Sources for Materials Science

Relatively low-energy (0.5–1.5 GeV) proton synchrotrons with high currents and high repetition rates are useful as sources of intense neutron beams. Slow-neutron beams are excellent tools for studying condensed matter on account of their penetration ability due to the neutral charge.

Tab. 18.1 Proton and heavy-ion synchrotrons, storage rings and colliders for nuclear and for high-energy physics.

Accelerator	Institution[a]	Particles (energy)	Luminosity ($\text{cm}^{-2}\ \text{s}^{-1}$)	Particles/cycle	Remarks
PS	CERN	p (26 GeV), heavy ions	–	10^{13} (p)	Injector chain for LHC
SPS	CERN	p (450 GeV), heavy ions	–	2×10^{14} (p)	Injector chain for LHC
LHC	CERN	p (7 + 7 TeV), Pb (2.7+2.7 GeV/u)	10^{34} (p), 2×10^{27} (Pb)		SC magnets
AGS	BNL	p (29 GeV), heavy ions	–	7×10^{13} (p)	Injector for RHIC
RHIC	BNL	p (250 + 250 GeV) Au (100 + 100 GeV/u)	4×10^{30} (p) 2×10^{26}(Au)		SC magnets
Tevatron	FNAL	p, p^- (900 + 900 GeV)	2×10^{32}		SC magnets
SIS 100/300 (proposed)	GSI	U^{28+} (2.7 GeV/u), U^{92+} (34 GeV/u)	–	5×10^{11} (U^{28+})	Double-ring facility

[a] BNL: Brookhaven National Laboratory, Upton, NY, USA. CERN: European Laboratory for Particle Physics, Geneva, Switzerland. FNAL: Fermi National Laboratory, Batavia, IL, USA. GSI: Gesellschaft fuer Schwerionenforschung, Darmstadt, Germany. KEK: National Laboratory for High Energy Physics, Ibaraki-ken, Japan.

Tab. 18.2 Operating proton synchrotrons and accumulator rings for pulsed spallation neutron sources.

Accelerator	Institution[a]	Particles (energy)	Cycle	Beam power	Remarks
ISIS	RAL	p (800 MeV)	50 Hz	160 kW	Rapid cycling synchrotron
SNS	ORNL	p (1 GeV)	60 Hz	1.4 MW (design)	Accumulator ring

[a] RAL: CCLRC Rutherford Appleton Laboratory, Chilton, UK. ORNL: Oak Ridge National Laboratory, TN, USA.

An extracted proton beam produces intense neutron bursts by spallation in a heavy-metal target. In such collisions, the created neutrons of rather high energy are emitted in the forward direction. However, most of the neutrons, having an energy peak at around 1 MeV, are almost uniformly distributed in angle. The velocities of such neutrons are slowed down by a moderator that surrounds the target. The slowed neutrons are called *cold neutrons* (0.1–10 MeV), thermal neutrons (5–100 MeV), and epithermal neutrons (100–500 MeV), respectively. Spallation neutrons form a pulsed beam of microsecond or submicrosecond duration having a high peak current. Such a time structure has a greatly different property compared to neutrons delivered from nuclear reactors, and is well used to investigate the time-varying nature of the materials. Examples of pulsed neutron sources are given in Table 18.2.

18.4.2.2 Synchrotron-Radiation Sources

Synchrotron radiation (see above) is one of the most valuable tools among the application fields of circular accelerators. Synchrotron radiation has been used for basic studies of materials science. Recently, applications are being widely and rapidly developed for chemical and biological research, as well as for industrial purposes. In the early stage, electron synchrotrons were used as sources for synchrotron-radiation research, which was a by-product of high-energy physics. At present, they are mostly being replaced by storage rings constructed as dedicated synchrotron-radiation sources.

Since the use of storage rings as synchrotron-radiation sources is spreading into various fields, a large number of storage rings have been constructed, or are under construction, worldwide. Electron storage rings of 1–2 GeV are built for using radiation from the vacuum-ultraviolet (VUV) to the soft X-ray range. For X-ray users, storage rings having energies of 2–8 GeV are constructed or under construction. Modern synchrotron-radiation sources include electron storage rings designed to circulate high-current beams with very low emittance to produce high-brightness radiation. The storage ring contains several or some tens of long straight sections for installing insertion devices, such as undulators or wigglers, whose periodic magnetic fields are designed to produce radiation with particular properties. Modern rings generally have large circumferences of several hundred meters up to more than one thousand meters.

Besides academic research, the application of synchrotron radiation has also spread to materials research and development in industry. Semiconductor

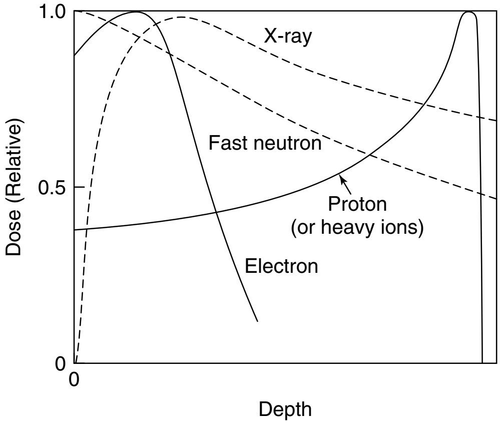

Fig. 18.10 Relative dose as a function of the depth in water for various types of beams.

lithography is being studied, which is going to realize semiconductors on the submicron scale for producing the next generation of microchips or microcircuit components. In addition to microelectronics technology, very fine fabrication to make tiny mechanical parts is under development using the technique of lithography. Small-size, low-cost machines, so-called compact storage rings, with energies usually below 1 GeV, are being built for industrial use. In the design of such machines, superconducting magnets are often used for the purpose of size reduction.

18.4.2.3 **Radiotherapy**

Heavy-ion synchrotrons as well as proton synchrotrons in the range of several hundred mega electronvolts are used for radiotherapy, such as cancer treatment. The aim of radiotherapy is to destroy the malignant cells with radiation beams. However, when radiation travels through the body, the normal tissues along the path are also damaged. Protons and heavy ions have well-defined particle ranges. The maximum energy is deposited in a region only a few millimeters in length near the end of range (known as the Bragg peak), as shown in Figure 18.10 This is an excellent property for radiation therapy, compared with X-ray, electron, or neutron beams. This allows a high dose to be delivered while concentrating on a diseased part with less damage to the overlying healthy tissues. Although the present status of this field is still experimental, dedicated facilities are under design or construction.

18.5
Technology and Engineering Aspects

18.5.1
Magnet System

The magnet is the most important and expensive item in circular accelerators. Various types of magnets are required, such as dipoles for bending the beam path and quadrupoles for focusing. Such magnets in synchrotrons are constructed under the requirements for a high degree of field quality and homogeneity over very large dimensions. For beam injection and extraction, they require kickers for the fast deflection of beams and septa for separation between the synchrotron chamber and the pipe for beam transport.

The progress of superconductivity technology has resulted in a new generation of magnet systems (see 18.5.6).

18.5.2
Radio-Frequency Engineering

Rf generators are needed to provide accelerating power for circulating beams and to compensate for the energy emitted as synchrotron radiation of the stored beams in the electron storage rings. In an electron storage ring having an energy of more than 1 GeV, several megawatts of rf power at 300–500 MHz are typically delivered to several rf cavities located in the ring. In this case, a UHF klystron of 100–1000 kW (continuous wave) is usually used as an rf generator.

In addition to the high-power systems, low-power rf technologies are very important because an accurate control of the frequency, voltage, and phase in the rf cavity is required for stable beam acceleration.

18.5.3
Computers and Automatic Controls

A large accelerator comprises a very complicated system, which contains a large number of components that must be controlled by computers. Such components are the rf accelerating system, the magnets and power supplies, the injection and extraction magnets, the vacuum system, the beam monitors, the timing circuits, the safety interlock system, and so on. Start-up, operation, and stopping of the entire machine system are in general automatically and finely controlled by computers.

Information concerning those machine components and beam parameters in the accelerators must be presented and displayed to the person who needs to interact with the machines. Practically, beams in accelerators are always changing their parameters, such as beam currents, beam positions, energies, and beam sizes. These parameters are always monitored during injection, acceleration, and extraction, and have to provide a feedback to the beams; otherwise, the machine would lose its stable operation. Such parameters of the beams and the current status of the machine components are recorded and displayed on screens. The control of the safety system is an important role of computers. It contains protection against radiation hazards, electricity accidents, and vacuum troubles.

18.5.4
High-Vacuum Engineering

The beam chambers of synchrotrons and storage rings require a vacuum of very low pressure over very large areas. Chambers are usually made of nonmagnetic metals (e.g., stainless steel, aluminum alloys, and copper), and insulation materials, such as ceramics, in special cases. Synchrotrons require a pressure in the range 10^{-5}–10^{-7} Pa to preserve the beam quality. A very low pressure of 10^{-7}–10^{-10} Pa is required for storage rings, and collider rings in which a beam must be stored for many hours, or even days, without any significant beam loss caused by scattering with residual gas molecules. The time in which the number of particles becomes half, or $1/e$, is called the *beam lifetime*. The beam lifetime is, roughly speaking, inversely proportional to the vacuum pressure in the beam chamber (the lifetime is often determined by reasons other than the finite vacuum pressure).

In electron (positron) storage rings, special techniques are used to achieve such low pressures in the presence of high-power synchrotron radiation hitting the chamber walls. Outgassing due to the photodesorption process results in difficulty in maintaining a sufficiently low pressure under operation of high-current electron beams. Special considerations are necessary for choosing the materials of the chambers and for methods of surface treatment to obtain a good vacuum.

The vacuum chambers of synchrotrons or storage rings require some special performance. The electromagnetic interaction between the beam-chamber walls and circulating beams becomes crucial when the beam current is large. Electromagnetic fields (wake fields) are generated inside the chamber by interactions of the beam charge with the conducting wall surface. The wake fields sometimes shake the beams, thus causing beam instabilities.

18.5.5
Civil Engineering

The magnets and other components of synchrotrons must be aligned at a tolerance level of 0.1 mm over a large extent. This tolerance level must be maintained for 10–20 years. The construction of the building in which the machine is to be installed thus carefully takes into account such factors as distortion due to thermal stress and the placement of heavy components, and vibration from utilities.

18.5.6
Superconducting System

The technologies of superconductivity are already widely utilized in accelerators, such as magnets and rf cavities. The development of superconducting magnets (see above) in accelerator applications started around 1970 with the first available multifilamentary composite material. The main advantage of using superconducting magnets is that they allow the field to be increased, so that the sizes of accelerators, and, therefore, the construction costs, can be considerably reduced. Since heating loss is very low, operating costs also decrease. The Tevatron at FNAL is the first superconducting accelerator, using NbTi as a superconductor in the 5-T range at 4.2 K.

The LHC (CERN) employs superconducting magnets and rf cavities. The LHC dipole magnets (1.9 K) have to produce a field of around 8.4 T. Each dipole magnet is 14.3 m long (see Figure 18.11). In total, there are 1232 dipole magnets in the ring.

Superconductive rf cavities, for example, with Nb as a superconductor, are widely used for saving electric power. The power

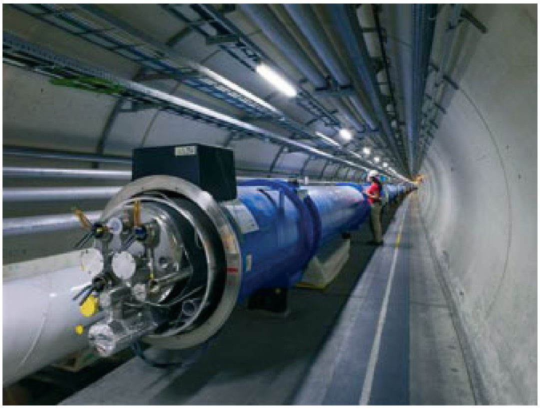

Fig. 18.11 LHC dipole magnet in its cryostat.

loss on the Nb cavity wall is $\frac{1}{10}$ of that on copper, and the accelerating voltage can be an order of magnitude larger than that of copper.

The storage rings for synchrotron-radiation sources built for industrial purposes widely use superconductive magnets for reducing their size and cost.

The superconducting systems in accelerators require sophisticated cryogenic systems. Technological progress in high-temperature superconductive materials is important and will influence future accelerator technologies.

Glossary

Beam Cooling: In storage rings, the beam emittances and momentum spreads are important parameters and are generally expected to be maintained at small values. The manipulation used to reduce the beam emittances and momentum spreads in accelerators is called *beam cooling*. The following methods have been developed: (i) radiation cooling for electron storage rings using radiation damping, (ii) electron cooling for ion beams, proposed by Budker (1966), and (iii) stochastic cooling using a feedback system (Van der Meer, 1972).

Beam Monitor: The equipment used to observe the beam intensities, beam positions, beam profiles, bunch lengths, and their variations in time are called *beam monitors*. The observation of the beam intensity commonly uses current transformers. The positions of the beams in the synchrotron are usually measured with pickup electrodes (e.g., bottom-type electrodes or strip-line electrodes).

Bunch: A bunch is defined as the group of particles formed due to the period of the rf accelerating field.

Luminosity: The performance of a storage ring for colliding-beam experiments is measured by the luminosity L, defined by $L = R/\sigma$, where R is the reaction rate, or number of interactions per second; σ is the interaction cross section. Typical values are in the range $L = 10^{29}$–$10^{32}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$.

Space-Charge Effect: When the beam intensity in a synchrotron is high, and particularly when the charged particles are tightly bunched and their velocities are slow, the Coulomb repulsion between them has two noticeable effects: (i) in the transverse direction, it counteracts the focusing forces of the synchrotron magnets. This results in shifts in the betatron tune. If the tune is drawn into a nearby resonance, a large loss of particles results. The shift in tune varies as $\beta^{-2}\gamma^{-3}$ (here, $\beta = v/c$, and $\gamma = E/mc^2$). (ii) In the longitudinal motion, it reduces the rf bucket area. A sufficiently high injection energy can cause the beams to be bound.)

References

Budker, G.I. (**1967**) *Symposium International sur les anneaux de Collisions dèlectrons et positrons* (eds R.A. Beck and E. Crémieu-Alcan), Presses Universitaires de France, Paris, pp. II/1–1–3.

Christofilos, N.C. (**1950**) unpublished report and U.S. Patent No. 2,736,799.

Courant, E.D., Livingston, M.S., and Snyder, H.S. (**1952**) *Phys. Rev.*, **88**, 1190–1196.

Courant, E.D. and Snyder, H.S. (**1958**) *Ann. Phys. (New York)*, **3**, 1–48.

Iwanenko, D. and Pomeranchuk, I. (**1944**) *Phys. Rev.*, **65**, 343.

Kerst, D.W. (**1941**) *Phys. Rev.*, **60**, 47–53.

Kerst, D.W. (**1942**) *Phys. Rev.*, **61**, 93–94.

Kerst, D.W. and Serber, R. (**1941**) *Phys. Rev.*, **60**, 53–58.

Lawrence, E.O. and Livingston, M.S. (**1931**) *Phys. Rev.*, **38**, 834.

McMillan, E.M. (**1945**) *Phys. Rev.*, **68**, 143–144.

Montanet L., *et al.* (**1994**) (Particle Data Group) *Phys. Rev. D* **50**, 1173–1875.

Rossbach, J. and Schmüser, P. (**1994**) Basic course in accelerator optics, in *Fifth General Accelerator Physics Course (CERN Accelerator School)*, (ed. S. Turner), CERN Report No. 94-01, Vol. 1, CERN, Geneva, Switzerland.

Sands, M. (**1970**) The Physics of Electron Storage Rings, Stanford Linear Accelerator Center Report No. 121, Stanford, CA.

Schwinger, J. (**1946**) *Phys. Rev.*, **70**, 798–799.

Van der Meer, S. (**1972**) Stochastic Damping of Betatron Oscillations in the ISR, CERN/ISR-PO/72-31.

Veksler, V. (**1945**) *J. Phys. USSR*, **9**, 153–158.

Further Reading

Bryant, P.J. (**1994**) *Fifth General Accelerator Physics Course (CERN Accelerator School)*, (ed. S Turner), CERN Report No. 94-01, Vol. **1**, CERN, Geneva, Switzerland.

Current developments can be found in the Proceedings of the U.S. Particle Accelerator Conferences and of the European Particle Accelerator Conferences and other Conferences (see *http://www.jacow.org*).

Edwards, D.A. and Syphers, M.J. (**1993**) *An Introduction to the Physics of High Energy Accelerators*, Wiley, New York.

Livingston, M.S. and Blewett, J.P. (**1962**) *Particle Accelerators*, McGraw-Hill, New York.

Publications of the CERN Accelerator School series, for example, Balbalat, O., *et al.* (**1994**) *Fifth General Accelerator Physics Course/CERN Accelerator School*, (ed. S. Turner), Vols. 1 and 2, CERN, CERN 94-01, are very useful readings.

SYNERGETICS AND CYBERNETICS

H. HAKEN, *Institute for Theoretical Physics and Synergetics, Stuttgart, Germany (Sec. 1)*
R. VALLÉE, *Université Paris-Nord, Paris, France (Sec. 2)*

Introduction 407
1. **Synergetics** 408
1.1 Goals of Synergetics 408
1.2 Basic Concepts of Mathematical Treatment 409
1.2.1 Microscopic or Mesoscopic Approach 409
1.2.2 Mesoscopic Approach 410
1.2.3 Order-Parameter Equations and Slaving Principle 410
1.3 Nonequilibrium Phase Transition 412
1.4 Instabilities and Instability Hierarchies 413
1.5 Phenomenological Synergetics 413
1.6 Macroscopic Synergetics 413
1.7 The Synergetic Computer 414
1.8 Some Further Applications 416
1.9 Relationships to Other Fields and Limitations 417
2. **Cybernetics** 417
2.1 Historical Background 418
2.1.1 The Initiators 418
2.1.2 The Precursors 418
2.2 Fundamental Concepts 418
2.2.1 Negative Feedback and Generalizations 418
2.2.2 Signal and Information 419
2.2.3 Isomorphism 421
2.3 Further Developments 421
2.3.1 Requisite Variety 421
2.3.2 Second-Order Cybernetics 421
2.3.3 Cognition and Action 422
2.3.4 Links with Systems Theory and Synergetics 422
2.4 A Tentative Chronology of Cybernetics 423
2.5 Fields of Application 423
Glossary 424
Works Cited 425
Further Reading 427

INTRODUCTION

Synergetics and cybernetics are two interdisciplinary fields that deal with the behavior of systems in both the animate and the inanimate worlds. Largely based on mathematics, both fields have found widespread applications to physics, chemistry, electrical and mechanical engineering, biology, medicine, economics, psychology, and other disciplines. Roughly speaking, cybernetics focuses on the problem of how to control a system so that it shows a required behavior, while synergetics deals with self-organization of systems, where the various possibilities of new states of a system are studied that may occur when external but rather unspecific control parameters are changed. Both synergetics and cybernetics study deterministic and stochastic processes and their interplay.

This article elucidates both the overlap between these two fields and their distinctions. While in cybernetics negative feedback is in the foreground of interest, in synergetics both positive and negative feedbacks are studied. While in cybernetics the controlling agent is given from the outside, in synergetics the controlling agent in the form of order parameters (or informators) is generated by the cooperation of the individual parts of the system. Here the feedback loop is closed by the slaving principle. In cybernetics the problem of the stability of a system plays a major role. Synergetics focuses its attention on instabilities and qualitative changes of a system's behavior close to instability points. Nevertheless, as it will transpire, the transition between the two fields is rather smooth.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. SYNERGETICS

1.1 Goals of Synergetics

Synergetics (S.) was initiated by Haken (1969) (see also Haken and Graham, 1971). It deals with systems that are composed of many individual parts (components) and that can spontaneously produce spatial, temporal, or functional structures by self-organization. "Self-organization" means that the specific structures are not prescribed from the outside by specific means. (Some examples are provided by Figs. 1–5.) S. searches for *general* (*unifying*) *principles of self-organization* irrespective of the nature of the parts of a system and establishes analogies between the macroscopic behavior of otherwise different systems. This is achieved by a mathematical description and the general concepts of stability, instability, control parameters, order parameters, and the slaving principle. S. has numerous applications in physics (including astrophysics, meteorology, and geology), chemistry, biology, computer science, psychology, economy, sociology, and other fields. In physics, chemistry, and biology, S. deals with open systems whose state is maintained by a flux of energy and/or matter into and out of the system under consideration. An example of a synergetic system is the light-source laser, whose principle may be described by the model of a gas laser. A glass tube is filled with a gas. The end faces of the tube carry mirrors that can reflect light running in the axial direction sufficiently often so that the damping of the light field is small enough and the laser condition can be met. The gas atoms are excited by an electrical current. In a usual lamp, the excited atoms emit individual wave trains of light entirely independently of each other so that microscopically chaotic light is produced. When the electric current is increased beyond a critical value, *coherent* light is emitted, i.e., a qualitatively new (highly ordered) state of the light field emerges. The coherence of laser light is brought about by the stimulated emission of radiation by the atoms. (Note the acronym "laser": light amplification by stimulated emission of radiation.)

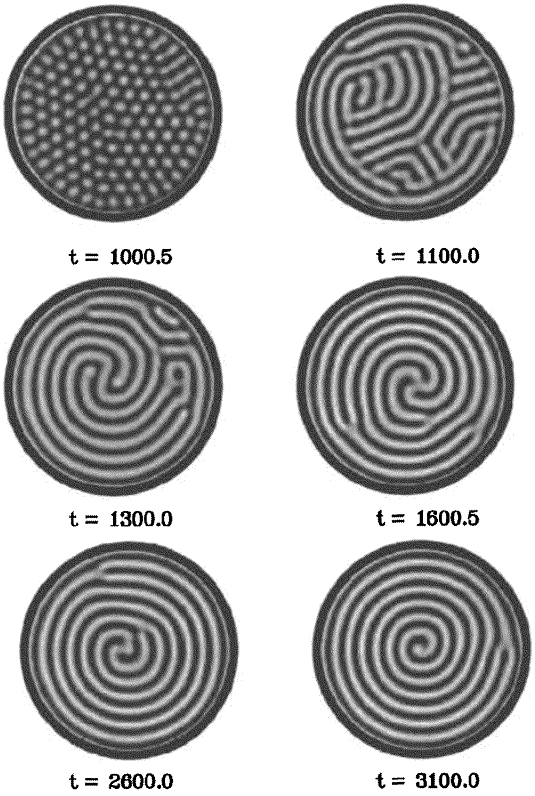

FIG. 2. Computer simulation of the transition from a hexagonal pattern to a spiral pattern in a liquid that is heated from below and cooled from above. This transition occurs when the walls are also heated (after Bestehorn *et al.*, 1993).

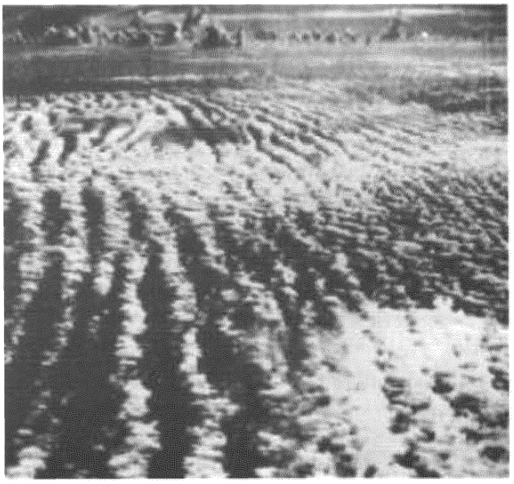

FIG. 1. Formation of cloud streets. These patterns are dynamic in that the air rises and sinks down spatially periodically (after Scorer, 1972).

The theoretical treatment of self-organiz-

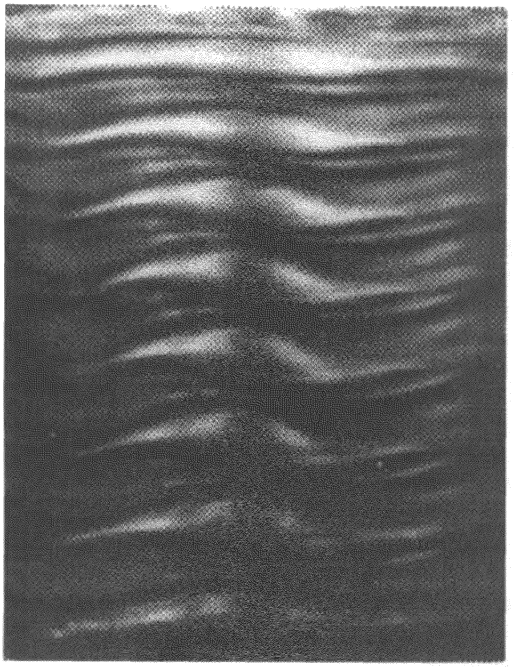

FIG. 3. Formation of periodic vortices in the so-called Taylor instability. A fluid is kept between two concentric cylinders of which the outer rotates. When the rotation speed exceeds a critical value, the patterns shown are formed. With increasing rotation speed, the patterns can become more complicated. For instance the pattern of Fig. 4 results (Swinney *et al.*, 1977).

ing systems distinguishes between microscopic (or mesoscopic), macroscopic, and phenomenological approaches.

1.2 Basic Concepts of Mathematical Treatment

1.2.1 Microscopic or Mesoscopic Approach Since the systems are open, their coupling to reservoirs (heat baths) must be taken into account. In the usual treatment, in a first step the reservoir variables are eliminated from the equations of motion, which may be classical or quantum mechanical. The resulting equations (see below) contain both damping of the system variables and fluctuations. The state, i.e., the spatiotemporal structure of a system, is represented by the—in general, time-dependent—state vector

$$\mathbf{q}(t) = (q_1, \ldots, q_n) \tag{1}$$

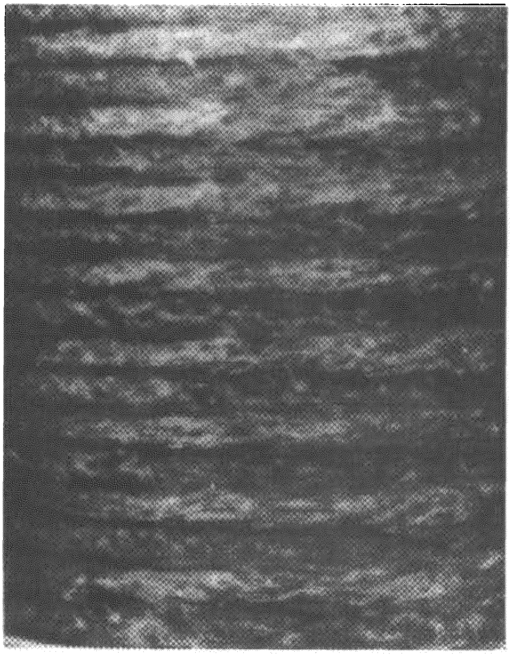

FIG. 4. Taylor instability leading to weak turbulence (Swinney *et al.*, 1977).

with its components q_j, $j = 1, \ldots, n$. Examples for these components in microscopic laser theory and nonlinear optics are atomic dipole moments, atomic inversion, and amplitudes of the modes of the electric-field strength. These quantities are classical variables in semiclassical theory and operators in quantum theory.

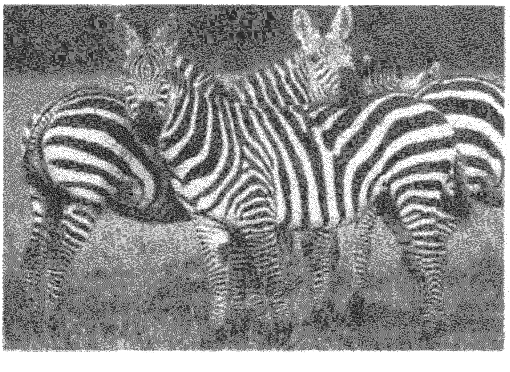

FIG. 5. A simple example of pattern formation in biology: formation of stripes on zebras. Such patterns can be accounted for by reaction-diffusion models based on the Turing instability (Turing, 1952; see also Meinhardt, 1982).

1.2.2 Mesoscopic Approach Here, a fluid is not described by the motion of its individual molecules but by the densities of its molecules, the velocity field, and the temperature field. In this case, the components of (1) become both space and time dependent. In plasmas the components of (1) are densities and velocity fields of electrons and ions, electric and magnetic fields, and temperature fields (see PLASMA PHYSICS; PLASMA WAVES AND INSTABILITIES). In semiconductors (1) refers to densities of electrons and holes, to occupation numbers of donors and acceptors, and to electric and magnetic fields.

In classical physics, chemical kinetics, a number of biological model systems, and other fields, **q** obeys evolution equations of the form (Haken, 1987)

$$\dot{\mathbf{q}} = \mathbf{N}(\mathbf{q},\nabla,\alpha,\mathbf{x}) + \mathbf{F}(t), \text{ where } \frac{d\mathbf{q}}{dt} \equiv \dot{\mathbf{q}}. \tag{2}$$

N is a *nonlinear* function of **q**. It may depend on spatial derivatives of **q** and may contain nonlocal couplings between the individual components of **q**. **N** depends on fixed control parameters α, which describe the influence of the surrounding "reservoirs" on the system, and it may explicitly depend on the space coordinate **x**. **F** represents fluctuating forces that may be caused from internal or external fluctuations (for instance, by coupling to the reservoirs). If **F** depends on the state variable **q**, the Îto or Stratonovich calculus (see Haken, 1987) must be used. In (2) it is assumed that $\dot{\mathbf{q}}$ and **q** refer to the same time. However, delay equations of the form

$$\dot{\mathbf{q}}(t) = \mathbf{N}(\mathbf{q}(t - T)) \tag{3}$$

or, more generally, of the form

$$\dot{\mathbf{q}}(t) = \mathbf{N}(\mathbf{q}(\tau)), \quad -\infty < \tau \leq t, \tag{4}$$

are also considered. In quantum theory, (2) is replaced by Heisenberg equations for operators **q**,

$$\dot{\mathbf{q}} = \frac{i}{\hbar}[H,\mathbf{q}] + \mathbf{R}(\mathbf{q}), \tag{5}$$

where H is the Hamiltonian of the system under consideration and **R** represents the effect of the (eliminated) reservoir variables. Alternative but equivalent descriptions of the system's evolution are the Fokker–Planck equation instead of (2) (for an example see below) and the density-matrix equation instead of (5).

1.2.3 Order-Parameter Equations and Slaving Principle Equations (2)–(5) are too general to be treated from a unifying point of view with results that are explicit enough for practical use. However, in many practical applications, the following strategy that is suggested by the laser example has proved highly successful. Assume that for a control parameter value α_0 a solution of (2) is known,

$$\alpha_0\colon \mathbf{q}_0(\mathbf{x},t). \tag{6}$$

$\mathbf{q}_0$ may be time independent, time periodic, time quasiperiodic, or chaotic. When the control parameter α_0 is changed, $\alpha_0 \to \alpha$, a linear stability analysis is performed, where the hypothesis

$$\mathbf{q}(\mathbf{x},t) = \mathbf{q}_0(\mathbf{x},t) + \mathbf{w}(\mathbf{x},t) \tag{7}$$

is made. Inserting (7) into (2), neglecting the fluctuating forces, and linearizing the resulting equation with respect to **w** leads to

$$\dot{\mathbf{w}} = L\mathbf{w}, \tag{8}$$

where L is a linear operator. The solutions of (8) have the general form

$$\mathbf{w} = \exp(\lambda_j t)\mathbf{v}_j(\mathbf{x},t), \tag{9}$$

where j is an index that distinguishes the different solutions of (8). If $\mathbf{q}_0$ is time independent, $\mathbf{v}_j$ is time independent or may contain powers of time t in case of degeneracies. If $\mathbf{q}_0$ is periodic, $\mathbf{v}_j$ is periodic and may contain powers of t in case of degeneracies. If $\mathbf{q}_0$ is quasiperiodic or chaotic, in large classes of systems $\mathbf{v}_j$ is also quasiperiodic or chaotic. In the case that (6) is time independent, the solution of (2) may be represented in the form

$$\mathbf{q}(\mathbf{x},t) = \mathbf{q}_0(\mathbf{x},t) + \sum_j \xi_j(t)\mathbf{v}_j(\mathbf{x},t). \tag{10}$$

(For the general case see Haken, 1987.) Inserting (10) into (2), and projecting on the

$\mathbf{v}_j$'s, Eqs. (2) may be cast into the general form

$$\dot{\xi}_j = \lambda_j\xi_j + \sum_{l,m} A_{jlm}\xi_l\xi_m + \sum_{l,m,n} B_{jlmn}\xi_l\xi_m\xi_n + \cdots + \hat{F}_j(t), \tag{11}$$

provided **N** can be expanded into a power series in **q**. Depending on the size of the real part of λ_j, unstable and stable mode amplitudes ξ_u and ξ_s, respectively, may be distinguished:

$$\xi_u\text{: } \mathrm{Re}\lambda_u \geq 0, \tag{12}$$

$$\xi_s\text{: } \mathrm{Re}\lambda_s < 0. \tag{13}$$

The *slaving principle* of synergetics (Haken, 1987) states that the stable mode amplitudes ξ_s can be uniquely expressed by the unstable mode amplitudes ξ_u at the same time to any desired degree of accuracy provided $|\boldsymbol{\xi}_u|$ is small enough:

$$\xi_s(t) = f_s(\boldsymbol{\xi}_u,(t),t). \tag{14}$$

In general, f_s can be approximated by a low-order polynomial for the components of $\boldsymbol{\xi}_u$. The explicit time dependence of $f_s(\boldsymbol{\xi}_u,t)$ stems from the time dependence of the fluctuating forces. By means of (10) and (14), $\mathbf{q}(\mathbf{x},t)$ can be expressed by the *order parameters* ξ_u alone. Since in general the number of $\boldsymbol{\xi}_u$ is much smaller than the number of the stable mode amplitudes ξ_s, the number of degrees of freedom of the system is greatly reduced and an enormous information compression achieved. Furthermore, in many practical applications it is sufficient to confine the sum over j in (10) to that over the order parameters ("mode skeleton"). The slaving principle contains a number of mathematical theorems or procedures as special cases. The adiabatic elimination procedure is based on the assumption that the λ_j's in (11) are real (or that by suitable transformations the imaginary part has been transformed away) and that $|\lambda_u| \ll |\lambda_s|$. (The fluctuating forces are neglected.) Then Eqs. (11) are split into those for $j = u$ ("unstable") and $j = s$ ("stable"). The *adiabatic approximation* means that in the equations for ξ_s, the assumption

$$\dot{\xi}_s \approx 0 \tag{15}$$

is made, by which these differential equations are converted into algebraic equations that can be solved by conventional means so that (14) can be established. In the *center manifold theorem* (Pliss, 1964; Kelley, 1967), it is assumed that $\lambda_u = 0$ and that there are no fluctuations. A further method of elimination of mode amplitudes ξ_s is by means of scaling procedures in which a specific scaling behavior of space, time, ξ_u, and ξ_s must be anticipated (see for instance Cross and Hohenberg, 1993).

By means of (14) the amplitudes ξ_s can be eliminated from Eqs. (11) so that order-parameter equations of the general form

$$\dot{\boldsymbol{\xi}}_u = \hat{\mathbf{N}}(\boldsymbol{\xi}_u,\alpha) + \hat{\mathbf{F}}(\boldsymbol{\xi}_u,t) \tag{16}$$

result. In a number of cases (16) can be simplified by use of normal-form theory (Arnold, 1983, 1993) and group theory (Sattinger, 1980) that exploits symmetry properties of the system.

If the spectrum λ_u of the linear operator L in (8) is continuous, the order parameters ξ_u become slowly space dependent,

$$\xi_u(t) \rightarrow \xi_u(\mathbf{x},t). \tag{17}$$

In this case the eigenvalues λ_u are transformed into operators, $\lambda_u(\partial/\partial x,\partial/\partial y,\partial/\partial z)$, that act on the space coordinate **x**, and Eqs. (16) acquire the form (Haken, 1987)

$$\dot{\boldsymbol{\xi}}_u = \lambda_u(\nabla)\boldsymbol{\xi}_u + \sum_{u',u''}\int dV'dV''A_{uu'u''}(\mathbf{x},\mathbf{x}',\mathbf{x}'')\,\xi_{u'}(\mathbf{x}')\xi_{u''}(\mathbf{x}'') + \cdots + \mathbf{F}(\mathbf{x},t). \tag{18}$$

Equations (18) are called *generalized Ginzburg–Landau* equations. Among special cases are the Swift–Hohenberg–Haken equation of the form

$$\dot{\boldsymbol{\Psi}}(\mathbf{x},t) = (\alpha + \beta(1 + \Delta)^2) + a\boldsymbol{\Psi}^2 - b\boldsymbol{\Psi}^3, \tag{19}$$

the original Swift–Hohenberg equation (Swift and Hohenberg, 1977) in which the quadratic term is lacking, and the nonlinear Schrödinger equation. The analogies between the behavior of different systems become manifest through Eqs. (16) or (18) with their special cases and the mode skeleton of (10). For instance the single-mode laser, a liquid developing a roll pattern, and a chemical reaction showing a stripe pattern have the same order-parameter equation.

1.3 Nonequilibrium Phase Transitions

Equations (16) and (18) may be considered as *generalized Langevin equations* to which the corresponding Fokker–Planck equation may be established (see Risken, 1989; Gardiner, 1983). Explicit examples of (16) and its corresponding Fokker–Planck equation are provided by

$$\dot{\xi}_u = \alpha\xi_u + \beta\xi_u^2 - \gamma\xi_u^3 + F(t) \qquad (20)$$

and

$$\dot{f}(\xi_u,t) = -\frac{\partial}{\partial\xi_u}[(\alpha\xi_u + \beta\xi_u^2 - \gamma\xi_u^3)f] + \frac{Q}{2}\frac{\partial^2 f}{\partial\xi_u^2}, \qquad (21)$$

respectively. Here f is the distribution function and Q is the strength of the fluctuating forces defined by

$$\langle F(t)F(t')\rangle = Q\delta(t - t'). \qquad (22)$$

The fluctuating forces are assumed to be white noise and Gaussian distributed with

$$\langle F(t)\rangle = 0. \qquad (23)$$

Examples for (20) and (21) (with $\beta = 0$) are provided by the order-parameter equations of a single-mode laser, of the convection instability with a single-roll pattern, and of stripe formation in some chemical reactions.

The behavior of ξ_u may be visualized by means of a ball sliding down in a potential as depicted in Fig. 6 (for the case $\beta = 0$) and subject to fluctuating forces. Close to $\alpha = 0$ *critical fluctuations* and *critical slowing down* occur. The occurrence of a *symmetry-breaking instability* is visible from Fig. 6. The stationary solution of (21) reads

$$f(\xi_u) = N\exp\left[\frac{2}{Q}\left(\frac{\alpha}{2}\xi_u^2 + \frac{\beta}{3}\xi_u^3 - \frac{\gamma}{4}\xi_u^4\right)\right] \qquad (24)$$

and puts in evidence the analogy between the behavior of a nonequilibrium system close to its instability point $\alpha = 0$ with the Landau theory of phase transitions. Equations (20), (21), and (24) describe a second-order *nonequilibrium phase transition* for $\beta = 0$, and a first-order nonequilibrium phase transition for $\beta \neq 0$ with hysteresis effect. In both cases, i.e., equilibrium and nonequilibrium phase transition, ξ_u serves as the order parameter, but while in the Landau theory the phase transition (Landau and Lifshitz, 1959) is caused by the temperature dependence of α, in the case of nonequilibrium systems α depends on rate constants. Note that in the case of a discrete spectrum of L, (20) and (21) are appropriate representations, and the Wilson theory of phase transitions of systems in thermal equilibrium (see PHASE TRANSITIONS: RENORMALIZATION AND SCALING) is not needed here.

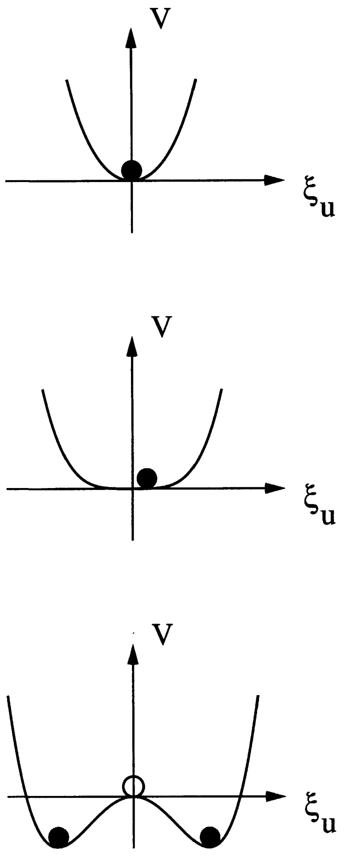

FIG. 6. Visualization of the solution of the order-parameter equation (20) with $\beta = 0$ by means of a ball sliding down the potential landscape $V = -(\alpha/2)\xi_u^2 + (\gamma/4)\xi_u^4$ for $\alpha < 0$ (upper part), $\alpha = 0$ (middle part), $\alpha > 0$ (lower part).

1.4 Instabilities and Instability Hierarchies

When fluctuating forces in (16) are neglected, concepts and results of dynamic systems theory (see Arnold, 1993; Guckenheimer and Holmes, 1993) may be invoked, and solutions of the order-parameter equations (16) may be classified as fixed points (time-independent ξ_u), limit cycles (time-periodic ξ_u), tori (time-quasiperiodic ξ_u), and chaotic ξ_u. If these objects are stable, they are called *attractors.* The region in phase space surrounding the attractor from which the trajectories are pulled into the attractor is called *basin of attraction.* When one or several control parameters are changed, the type of a solution of (2) or (16) may change. For instance, one stable fixed point may be replaced by an unstable and two stable fixed points (compare Fig. 6), or by a limit cycle that can be replaced by a torus or chaotic attractor, and so on. These transitions are studied by bifurcation theory and synergetics. A few phenomena that were derived from the solutions of equations (16) or (18) are the following: From Eqs. (16), lasers: cw single-mode operation, multiple-mode operation including steady states, frequency locking, ultrashort pulses, and chaotic laser light (Haken, 1985; Weiss and Klische, 1984). From Eqs. (18), transverse spatial and spatiotemporal patterns of the near and the far field of lasers and nonlinear optical devices (Arecchi and Harrison, 1987); in semiconductors current oscillations by impact ionization, current filaments (Schöll, 1987), Gunn oscillations, and chaotic oscillations.

In fluids, plasmas, and chemical reactions, spatio-temporal patterns can be treated via (16), (18), or (19). The patterns may depend on geometry (for instance, plane, square, or circular layers) or spherical geometry (such as in astrophysics, meteorology, or geophysics), and on initial conditions, fluctuations, intrinsic mechanisms, or driving conditions (e.g., heating from below). Among the patterns treated are rolls, hexagons, zigzag waves, single- and multiple-armed spirals, period-doubling routes to chaos, solitary waves, and defect structures.

1.5 Phenomenological Synergetics

In a number of cases, such as in complex physical or chemical systems, and in particular in biology, the behavior of the microscopic parts is not directly accessible to a mathematical description. Nevertheless, in a number of cases it is possible to describe the behavior of the system phenomenologically by an appropriate identification of control and order parameters. In this case, Eqs. (16) or (18) are formulated heuristically. Examples are provided by fluid dynamics (for instance, Ginzburg–Landau equations and some generalizations) or in biology by models of movement coordination (Haken *et al.*, 1985), Eigen equations of evolution (Eigen, 1971), Gierer–Meinhardt model of morphogenesis (see Meinhardt, 1982), etc.

1.6 Macroscopic Synergetics

The aim is to derive underlying equations of motion for the order parameters or even for enslaved mode amplitudes using the available experimental data. In the Friedrich–Uhl–Haken method (Friedrich *et al.*, 1995a,b), the leading terms of both the order-parameter equations (16) and the slaving principle (14) are taken into account, and the unknown coefficients that appear in (14) and (16) are determined by a least-squares fit where the experimental data enter. A further approach is based on the maximum-information entropy principle (Jaynes, 1957a,b, 1962) that can be formulated as follows: Let p_j be the probability of finding a system in state j, $j = 1, \dots, M$, and set $K = 1/\ln 2$. Then the information—or information entropy—

$$i = -K \sum_j p_j \ln p_j \tag{25}$$

is maximized by an appropriate choice of the p_j's under the constraints

$$\sum_j p_j = 1 \tag{26}$$

and

$$\langle f^{(k)} \rangle \equiv \sum_j f_j^{(k)} p_j = f_k. \tag{27}$$

While the $f_j^{(k)}$'s are given functions, the f_k's are experimentally observed average values. In most applications to synergetics, the index j is identified with the (continuous) state vector $\mathbf{q}$, Eq. (1), and the constraints (27) to be used are moments of the components of $\mathbf{q}$ mostly up to the fourth order of the form

$$\langle q_k \rangle = \bar{q}_k, \langle q_k q_{k'} \rangle = \overline{q_k q_{k'}}, \cdots. \tag{28}$$

While the left-hand sides of Eqs. (28) are defined according to (27), the right-hand sides are averages over experimentally observed time series of moments, i.e., of products of q_k, as indicated by the bars. Maximizing (25) under the constraints (26) and (27) yields

$$P(\mathbf{q}) = N \exp\left\{-\sum_j \lambda_j q_j - \sum_{jj'} \lambda_{jj'} q_j q_{j'} \cdots\right\} \tag{29}$$

for the time-independent distribution function. Note the analogy with (24). Also, time-dependent processes can be analyzed using the maximum-information entropy principle provided the process is Markovian (Haken, 1988). Based on conditional moments, the short-time propagator is thus determined.

1.7 The Synergetic Computer

The synergetic computer (Haken, 1990) represents an alternative to neural computers [see NEURAL NETWORKS including spin-glass models (Hopfield, 1982)]. It is used, in particular, for pattern recognition in the form of an associative memory (Kohonen, 1987) and based on the idea that pattern recognition can be considered as pattern formation. The prototype patterns to be stored are encoded as vectors

$$\mathbf{v} = (v_1, \cdots, v_j, \cdots, v_N), \tag{30}$$

whose components may be, for instance, the gray values of the corresponding pixels of a picture. The prototype vectors are distinguished by the index k:

$$\mathbf{v}_k, \quad k = 1, \cdots, M. \tag{31}$$

They are assumed to be normalized,

$$\mathbf{v}_k^2 = 1. \tag{32}$$

The adjoint vectors $\mathbf{v}_{k'}^{\dagger}$ obey

$$\mathbf{v}_{k'}^{\dagger} \cdot \mathbf{v}_k = \delta_{kk'}, \tag{33}$$

where δ is the Kronecker symbol. The synergetic computer uses a dynamics in which an offered test pattern vector $\mathbf{q}$ becomes time dependent and is, eventually, identified with one of the stored prototype patterns to which it had been most similar in the beginning:

$$\mathbf{q} \rightarrow \mathbf{q}(t) \rightarrow \mathbf{v}_k. \tag{34}$$

This dynamics is described by the equations

$$\dot{\mathbf{q}} = \sum_k \lambda_k (\mathbf{v}_k^{\dagger} \mathbf{q}) \mathbf{v}_k - \sum_{kk'} B_{kk'} (\mathbf{v}_{k'}^{\dagger} \mathbf{q})^2 (\mathbf{v}_k^{\dagger} q_k) \mathbf{v}_k - C \mathbf{q}^2 \mathbf{q}. \tag{35}$$

Equations (35) can be derived from a potential function V

$$\dot{\mathbf{q}} = -\frac{\partial V}{\partial \mathbf{q}^{\dagger}}. \tag{36}$$

Examples of prototype patterns $\mathbf{v}_k$ stored in the computer and the recognition process are shown in Figs. 7 and 8. A learning algorithm (Haken *et al.*, 1989) allows the synergetic computer to learn the prototype patterns even if the offered patterns are noisy or partly hidden (Fig. 9), or if classifications must be made. To this end, the potential function

$$V_{\text{tot}} = \sum_j V(\mathbf{q}_j, \mathbf{v}_k, \mathbf{v}_k^{\dagger}), \tag{37}$$

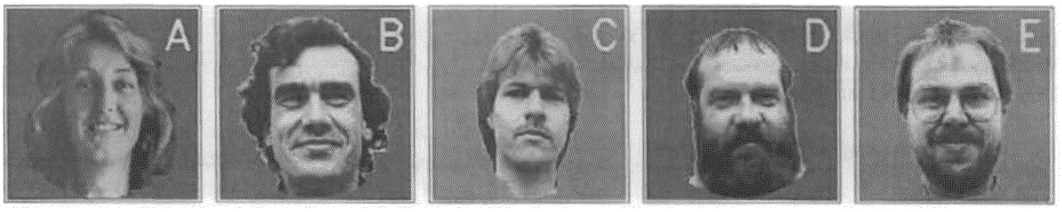

FIG. 7. Examples of stored prototype patterns including the encoded family names.

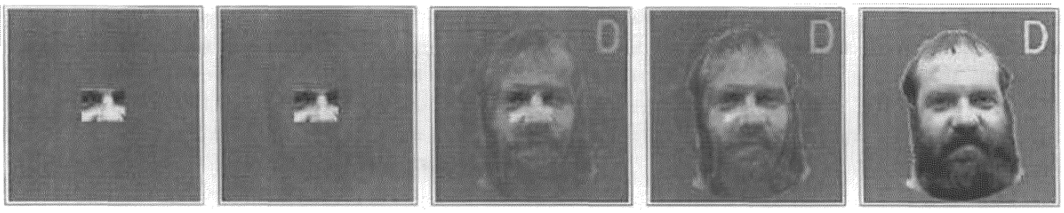

FIG. 8. Example of a recognition process, where part of a face is prescribed as initial state.

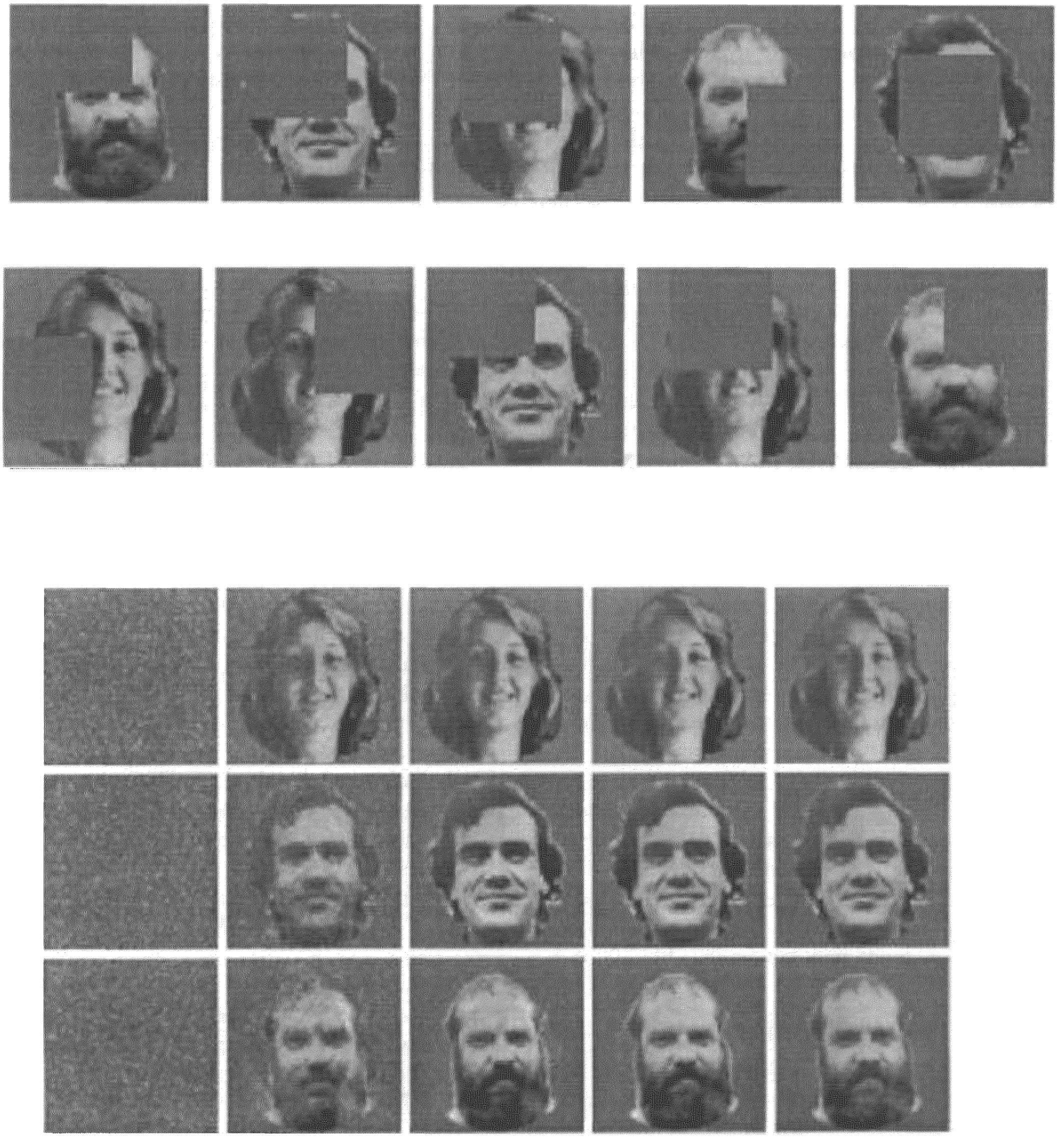

FIG. 9. Learning of partly hidden faces. Upper part: A random sequence of such patterns, each time with a different position of the squares, was offered to the computer. Lower part: Simultaneous learning of the three offered faces.

where the sum runs over all offered patterns, is minimized with respect to $\mathbf{v}_k$ and $\mathbf{v}_k^\dagger$ by a gradient dynamics, where the constraints (32) and (33) are taken into account. The sum in (37) runs over all patterns offered to the computer.

1.8 Some Further Applications

In chemistry synergetics is applied to pattern formation in the Belousov–Zhabotinski reaction and other reactions (for reviews see Ross *et al.*, 1988; Nicolis, 1989). In that chemical reaction $Ce_2(SO_4)_3$, $KBrO_3$, $CH_2(COOH)_2$, H_2SO_4 and a few drops of Ferroin (redox indicator) are mixed. The thus resulting homogeneous mixture can show oscillations or develops spirals or concentric rings depending on the geometry of the vessel (and on the continuation of stirring in the case of oscillations).

In biology and medicine, synergetics applies to the Turing model of morphogenesis (Turing, 1952). This model proposes a mechanism of cell differentiation during the development of an organism. In its original form the model considers two undifferentiated cells (for instance, of an embryo in its early stage). In both cells the same chemical reactions (metabolisms) are going on. When an exchange of molecules between the cells is permitted, a small fluctuation of the concentration of one kind of molecules can cause an instability whereby the concentrations of molecules change and become dramatically different in the two cells. It is assumed that high concentrations of so-called activator molecules can switch on genes, which leads to cell differentiation, for instance, to the production of pigments on furs, or to specific cells of organs. The Turing model was extended to sheets (or other ensembles) of cells by Gierer, Meinhardt, and others (see Meinhardt, 1982). Stripe and hexagonal ("spot") patterns on furs or on skins of fish were derived from this model by Olbrich and Haken (see Haken, 1983) by use of the concepts of synergetics.

Models of evolution of biomolecules (Eigen, 1971; Eigen and Schuster, 1977, 1978) utilize synergetics. These models deal with the early stage of the evolution of life at the level of biomolecules. It is assumed that these molecules (probably RNA) can multiply, mutate, and compete with each other. Because of a specific fitness factor, one kind of molecule ("species") wins the competition so that selection is achieved. There is a one-to-one correspondence between the original Eigen equations and the photon-number equations of the laser so that the evolution dynamics can be mimicked by the latter.

Synergetics applies to movement coordination in humans and animals (Kelso, 1995; Haken, 1996). The movement of limbs is highly correlated and shows up in specific movement patterns ("gaits"). The concepts of synergetics allow one to model mathematically the transitions between movement patterns. A classical experiment is that by Kelso (1995), where a person is asked to move the index fingers of both hands in parallel. When the frequency of the movement is enhanced, suddenly and involuntarily a switch to an antiparallel, symmetric movement pattern occurs. The movement frequency plays the role of a control parameter and the relative phase that of an order parameter. Phenomena such as hysteresis, critical fluctuations, and critical slowing down could be predicted. The phenomena described by this model seem to be of a universal nature in movement coordination.

Synergetics may enter into the analysis of electroencephalograms (EEGs) and magnetoencephalograms (MEGs) (Haken, 1996). In EEGs (MEGs) the spatio-temporal distribution of electric (magnetic) fields over the scalp is measured. It is assumed that hereby the electric (magnetic) activity of the brain is measured. If it is assumed that the brain operates close to instability points, the concepts of synergetics as outlined in this article become applicable. The observed spatiotemporal (EEG or MEG) patterns can be decomposed into a superposition of a few spatial modes with time-dependent amplitudes that act as order parameters and obey typical order-parameter equations. Examples are the analysis of α-waves (with five order parameters), petit-mal epilepsy (with three order parameters that show so-called Shilnikov chaos), or MEGs in sensorimotor control tasks, such as finger tapping on acoustic signals (with two order parameters).

Models of perception [including that of ambiguous figures (Haken, 1996)] are based on the synergetic computer (see Sec. 1.7) but take into account the time dependence of the so-called attention parameters λ_k that occur in (35). This time dependence is caused by the saturation of attention and leads to an oscillation between different percepts associated with ambiguous figures. Qualitatively, perhaps in a more speculative fashion, synergetic concepts were applied to psychology, psychiatry, and group therapy (see, for instance, Langthaler and Schiepek, 1995).

In sociology, synergetics is applied to formation of public opinion (Weidlich and Haag, 1983). The numbers of persons with specific political opinions serve as order parameters that obey typical order-parameter equations. As a consequence of changes of specific control parameters (e.g., political "climate"), dramatic changes of public opinion can occur. Transitions between democratic behavior and dictatorship become accessible to mathematical modeling. For more intricate situations and their models, see Weidlich and Haag (1983).

In urban development (Sanders, 1992), in studies related to synergetics, the development of cities is conceived as a process of self-organization. Among basic questions is this: How can planning agencies fix appropriate control parameters that allow for self-organization of cities into desired structures?

In models of economy (Zhang, 1991), questions treated are these: How can an economy be described (modeled) by a few characteristic quantities (order parameters)? What is their dynamics depending on specific control parameters (e.g., interest rates, oil price, etc.), and what typical changes can occur (e.g., transitions between full employment and underemployment)? Somewhat related problems occur in ecology—for instance, the impact of pollution on vegetation.

1.9 Relations to Other Fields and Limitations

Cybernetics (cf. Sec. 2): While in cybernetics the concepts of control, steering, regulations, etc. are in the foreground of interest, in synergetics the impact of unspecific changes of control parameters on the system's dynamics is studied. Synergetics shares the goal of general systems theory (founded by von Bertalanffi: von Bertalanffi, 1950, 1953) in that it looks for analogies in the behavior of different systems. While general systems theory looks for such analogies at the level of the individual parts, synergetics establishes such analogies at the level of the order-parameter equations (16) or (18). In this way, for instance, analogies between phenomena (chaos, spatiotemporal patterns) in fluid dynamics and lasers (Haken, 1985) were established. Relationships between synergetics and the theory of phase transitions of systems in thermal equilibrium were mentioned above.

S. shares a goal with the concept of dissipative structures in the sense of Prigogine (Glansdorff and Prigogine, 1971), in that it looks for general principles underlying self-organization. Principles proposed by the latter theory, namely, the principles of minimum entropy production and of excess entropy production, were not successful, however. Later work by the Prigogine school is mainly based on bifurcation theory (Nicolis and Prigogine, 1977; Nicolis, 1989). Dynamic systems theory including bifurcation theory (Guckenheimer and Holmes, 1993), catastrophe theory (*q.v.*; see also Thom, 1975), and chaos theory (Schuster, 1988; CHAOTIC PHENOMENA) provide synergetics with important mathematical tools, though, on the other hand, these theories ignore entirely the role of fluctuations and the statistical aspects that play a fundamental role in phase transitions far away from thermal equilibrium. Finally, information theory plays a role in synergetics via the concept of maximum information entropy. An interesting future task will be to establish profound links between fractal geometry (Mandelbrot, 1983; FRACTAL GEOMETRY), which is a static theory, and synergetics, which is a dynamic theory.

In its present form the mathematical approach outlined above applies to situations close to instability points (in control-parameter space), where the order parameters change smoothly. There are both experimental and theoretical indications that in a number of cases the order-parameter concept and the slaving principle are also valid away from the instability points. A phenomenon that has not (yet?) been accessible to these methods is turbulence. The same may be true for cases in which several or many instabilities coincide or follow closely (for instance, in plasmas).

2. CYBERNETICS

Modern cybernetics was introduced by Norbert Wiener as the science of control and communication in the animal, the machine, and even societies. Negative feedback, quantity of information, isomorphism, and requisite variety are among the important concepts of cybernetics. Perception and action are linked to cybernetics, which is close to systems theory and synergetics. Applications of cybernetics include robotics and information theory and may also be found in biological modeling, economics, and ecology.

2.1 Historical Background

2.1.1 The Initiators The modern start of cybernetics dates back from the first edition of the book of Norbert Wiener (1894–1964), *Cybernetics or Control and Communication in the Animal and the Machine* (Wiener, 1948). Norbert Wiener, Warren S. McCulloch, John von Neumann, and other colleagues, belonging to a group sponsored by the Macy Foundation and called "Teleological Society," recognized the importance of a new field of research (1943) that they decided, in 1947, to call *cybernetics* from the Greek *κυβερνήτικὴ*, which means the art of steering. The Teleological Society was also nicknamed "Cybernetical Club," an expression similar to "Cercle d'Etudes Cybernétiques" created in 1950 in Paris. The role of McCulloch (1898–1969) is not to be underestimated. His first contribution to the field was an article, with N. Pitts, devoted to the algebra of logics applied to nervous activity (McCulloch and Pitts, 1943). In opposition to Wiener, who considered Leibniz as the patron of cybernetics because of his attempt to build up a general calculus of logics, he put this new science under the aegis of Descartes, who proposed a kind of feedback in the framework of his theory of nervous transmission.

2.1.2 The Precursors In fact, the French word *cybernétique* had been introduced by Ampère (1843), in his classification of sciences. For him it represented the art of government, the same meaning Plato gave metaphorically to *κυβερνήτικὴ*, particularly in his dialogue on politics. The same year as Ampère, Trentowski introduced in Polish the word *kibernetiki*.

To this list we may add S. Odobleja who had an intuition of the universal role of feedback presented in a book (Odobleja, 1938, 1939) on what he called "psychologie consonantiste." Also from Romania, P. Postelnicu had a clear view of the general importance of feedback in a paper proposed in 1944 but accepted for publication only in 1968 (Postelnicu, 1968). Another precursor is Lafitte (1932), who had seen the philosophical importance of "machines réflexes."

2.2 Fundamental Concepts

2.2.1 Negative Feedback and Generalizations Control and communication are very important in Wiener's cybernetics. Depending upon both of them, regulation by *negative feedback*, already known for a long time, plays a prominent role. Watt's flyball governor gives a famous example of it, as does the clepsydra water supply. In all these kinds of regulation the problem is to maintain, as well as possible, an essential parameter of a dynamical system at a given value: the rotation speed in a steam engine for Watt's governor, the level of water in a reservoir for the clepsydra. In the case of Watt's governor, when the speed is too high, the steam supply is reduced through a well-known device based upon the action of centrifugal forces upon the flyballs. More generally the value of the parameter to be regulated is compared to its desired level. Based upon this comparison, a countermeasure is generated tending to decrease the parameter when it is too big and to increase it when too small. The fact that the actual value of the parameter (considered as the output) is compared to its desired value (considered as an input) justifies the word feedback. This feedback is said to be negative because of the action, generated by the value of the difference between the output and the input, tends to decrease this difference when it is positive and to increase it when it is negative.

Closer to us, the thermostat, which is used to maintain the temperature of a place close to a prescribed value, is another kind of regulator. More generally a device designed to make a temperature follow a given program as well as possible, despite perturbations, is a more refined type of regulator involving also a negative feedback. If $u(t)$ is the desired temperature at instant t, and $\theta(t)$ the temperature actually realized under the influence of a perturbation $v(t)$, $\theta(t) - u(t)$ is the gap whose absolute value is to be maintained as small as possible. Here $\theta(t)$ is the *output* of this system, $u(t)$ is an *input*, the one that is at the disposal of the user of the device, and $v(t)$ may be considered as another input but out of reach of the user. When the gap $\theta(t) - u(t)$, considered as a signal, is positive, it must generate some absorption of heat in order to make the gap decrease despite the perturbing influence of $v(t)$. This behavior may be represented, in the most simple case, by a differential equation such as

$$\begin{aligned}\dot{\theta}(t) &= \alpha[\theta(t) - u(t)] + \beta v(t)\\ &= \alpha\theta(t) - \alpha u(t) + \beta v(t), \qquad (38)\end{aligned}$$

where $\alpha < 0$ and β is sufficiently small.

In the above example the state $\theta(t)$ of the considered device is a mere scalar, but it could be a vector (represented by a column matrix). In the same way $u(t)$ and $v(t)$ also could be column matrices, α and β becoming respectively a square and a rectangular matrix. In this case, condition $\alpha < 0$ would be replaced by the strict negativity of the eigenvalues of matrix α, and the smallness of β by that of the elements of matrix β. These considerations open the way to more general cases, with or without feedback, represented, in the linear differential case, by the following multivariable equations:

$$X(t) = A(t)X(t) + B(t)U(t), \qquad (39)$$

$$Y(t) = C(t)X(t) + D(t)U(t), \qquad (40)$$

where input $U(t)$, state $X(t)$, and output $Y(t)$ are column matrices, $A(t)$ is a square matrix, and $B(t)$, $C(t)$, and $D(t)$ are generally rectangular matrices, all depending upon time. In such a differential system, input $U(t)$ is considered as a control upon state $X(t)$ and output $Y(t)$ as what is observed of $X(t)$. Then this differential system is said to be completely *controllable* if, for any initial and final instant t_0 and t_1 and for any state X_0 at t_0 and X_1 at t_1, there exists at least one control that brings the state X_0 to X_1. In fact there is an infinity of such controls, a situation that opens the way to the problem of *optimal control* with the works of Bellman (1957) and Pontryagin (1962). The system is said to be completely *observable* if, for any initial and final instant t_0 and t_1, the knowledge of input and output functions on the interval $[t_0,t_1]$ enables us to reconstruct the state function on this interval. Kálmán (1960) gave necessary and sufficient conditions for both controllability and observability. Of course, these results may be adapted to the linear recurrent case and to nonlinear equations, even with delay, whether deterministic or stochastic.

2.2.2 Signal and Information The second important concept upon which Wiener based cybernetics is that of communication or more precisely *signal* theory. His main results in this field are exposed in Wiener (1949), but founded upon more ancient results he obtained on Brownian motion and his generalized harmonic analysis. In this book, Wiener solves, in the framework of stationary time series, the problem of optimal elimination of an additive noise perturbing a signal and of optimal prediction of this signal. The solution, anticipating those initiated later, was achieved by Wiener by means of a frequency filter, but had been obtained earlier quite independently by Kolmogoroff (1941) in comparable but different terms.

Wiener considered the works of Claude E. Shannon on *information theory* as an essential part of cybernetics. Presented in Shannon's seminal article of 1948, they were published also in his book *The Mathematical Theory of Communication*, with an additional chapter by W. Weaver (Shannon and Weaver, 1949). Shannon considered, as did Wiener, a signal perturbed by an additive noise. But instead of trying to eliminate the noise as well as possible, his aim was to study the ability of a channel to transmit information, in both the discrete and continuous cases.

According to Shannon, a general communication system can be represented by a source of information producing an input message on which operates a transmitter giving an input signal suitable for transmission over a channel. The addition to this signal of a noise, coming from a source of noise, generates the output signal, which enters into a receiver, giving an output message that then reaches its destination (be it a conscious being or not) (see Fig. 10).

Let us consider the case of a discrete source of information giving, as an input

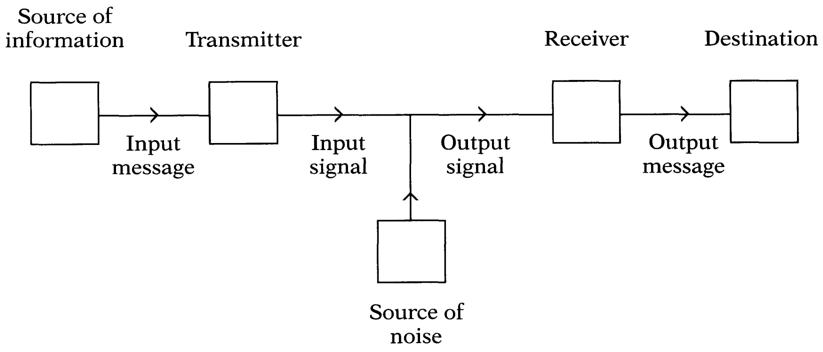

FIG. 10. Schematic view of a communication system.

message, a sequence of symbols chosen according to certain probabilities depending generally upon the symbol itself and the preceding choice. This input message is encoded into an input signal that is a sequence of other symbols chosen according to certain other probabilities. This input signal is perturbed by a discrete noise also of probabilistic nature. In such a case, the only one we shall present here, Shannon has shown that there exists a parameter C, the *channel capacity,* such that, if the *quantity of information* per unit of time produced by the source is less than or equal to C, there is a coding of the input message into the input signal such that the transmission over the channel, up to the output message, is realized with an arbitrarily small frequency of errors.

In the above statement, Shannon introduces what he calls quantity of information. The idea originated in statistics with the works of R. Fisher and continued with those of Nyquist (1924), also known for his diagram concerning feedback and stability, and Hartley (1928). The quantity of information of Shannon has no semantic meaning. It is only an index that increases with unexpectedness. So the quantity of information attributed to a message announcing that an event, with *a priori* probability p, has occurred is

$$I = -\log_2 p, \quad 0 \le p \le 1, \tag{41}$$

where $\log_2$ means logarithm with base 2. Of course $I > 0$, and when $p = \frac{1}{2}$, then $I = 1$. Consequently, the unit of quantity of information corresponds to the case where realization and nonrealization of the event are equiprobable. It initially had been given the name of Hartley and is currently called bit (binary digit).

Coming back to communication theory, we consider a discrete source of information generating a sequence of symbols chosen from a finite set $\{s_1, s_2, \ldots, s_n\}$ according to probabilities supposed here, for the sake of simplicity, independent of the preceding choices. So the probability of symbol s_i is p_i with $0 \le p_i \le 1$, $\Sigma_i p_i = 1$. Consequently, the quantity of information associated to the occurrence of symbol s_i is $-\log_2 p_i$ and the average quantity of information per symbol is, according to Eq. (41),

$$J = -\sum_i p_i \log_2 p_i, \tag{42}$$

with, by convention, $0 \log_2 0 = 0$. We easily see that $0 \le J \le \log_2 n$, the maximum being obtained when all the p_i are equal to $1/n$. If there are only two symbols, we have $p_1 = p$, $p_2 = 1 - p$, and the average quantity of information per symbol is

$$J = -[p \log_2 p + (1 - p) \log_2 (1 - p)]. \tag{43}$$

The maximum of 1 bit is obtained for $p = 1 - p = \frac{1}{2}$ (see Fig. 11). Coming back to the case of n symbols, the average quantity of information, per unit of time, produced by the source is J/τ if τ is the mean duration of the emission of a symbol. It is this quantity that is considered in the proposition of Shannon given above.

In statistical mechanics the *entropy* of a probability distribution $p_1, p_2, \ldots, p_n$ of the

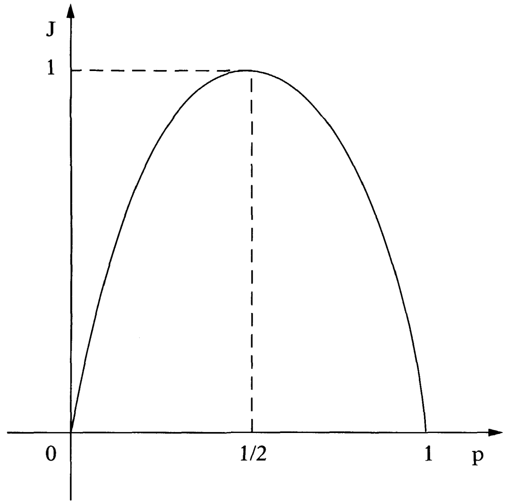

FIG. 11. Average quantity of information per symbol for an "alphabet" of two symbols of probabilities p and $1 - p$.

n possible states of a thermodynamical system is given by

$$S = -k \sum_i p_i \ln p_i,$$

where ln means the natural logarithm and k is Boltzmann's constant. The formal analogy between Eq. (42) and (43) was remarked by Shannon, who used the word "entropy" in place of average quantity of information per symbol. It also generated researches by Szílard (1929) and then by Mandelbrot (1953) (known also for his theory of fractals) and Brillouin (1956), which aroused the interest of L. de Broglie. They proposed explanations of the so-called Maxwell demon paradox, involved in a thought experiment supposed to show that the second law of thermodynamics (increase of entropy in a closed system) may be wrong at the microscopic level. These approaches are still controversial, and new theories have been proposed.

2.2.3 Isomorphism *Isomorphism* is another feature of cybernetics. In the very title of Wiener's book, the two words "animal" and "machine," to which we may add "society" if we consider an important chapter of this work, show that cybernetics aims, with at least the help of the concepts of feedback and quantity of information or more generally control and communication, at the recognition of structural analogies between, for example, the fields of the artificial, of biology, and of sociology. These structural analogies concern forms if we consider this word with its broadest meaning; they may imply, for example, a common kind of modelization. This is why isomorphism, the strongest case of analogy, that is to say a one-to-one correspondance between elements and relations between elements, from a field to another, is so important for cybernetics, even if in many cases isomorphisms are not perfectly realized.

2.3. Further Developments

2.3.1 Requisite Variety After the first modern approach to game theory by E. Borel, von Neumann greatly improved the initial results, mainly in his *Theory of Games and Economic Behavior,* in collaboration with Morgenstern (von Neumann and Morgenstern, 1944). Mandelbrot (1953) introduced the idea of a link between games theory and communication. He observed that in communication theory we have a transmitter who wants to communicate information with the help of a signal; the line; the receptor who expects to receive the information; and "nature," which passively spoils the signal with a noise. So we have a duel between transmitter and nature, a situation considered with great details in games theory.

The contribution to cybernetics of Ashby (1903–1972) involves his so-called principle of *requisite variety* proposed in Ashby (1956). This principle, when made more precise in a special case by Conant (1969), tells that the greater the capacity of a regulator to transmit information per unit time, through its feedback chain, the better its ability to resist successfully a perturbing noise. We see here concepts coming both from information and game theories. In its most general form, Ashby's principle says that in order to resist a great variety of possible aggressions it is necessary to dispose of at least the same variety of countermeasures—in short, that only variety can destroy variety.

2.3.2 Second-Order Cybernetics In *second-order cybernetics,* an expression due to Heinz von Förster, emphasis is put on observation not merely of a cybernetical device, or system, by an external observer but performed by the system upon itself and the outside world. It involves, in mathematical terms, the introduction of an *observation operator* (Vallée, 1951) and obviously auto-observation as well as *self-reference,* which was considered first by von Förster and developed later by Bråten (1982), Löfgren (1987), and Pask (1975), to whom we are also indebted for the concept of *conversation theory.* In the same context was introduced the concept of *perceptron,* a device able to reproduce, at least schematically, the perceptive properties of an observing system (Cellérier *et al.,* 1968). The iteration of the process of perception may give rise to a fixed point, that is to say an element of a set that is invariant under a given transformation of the set into itself. Here the fixed point represents, for the observing system, a stable way of acting, which von Förster called *eigen-behavior,* an expression already used by Jerne (1974), with a somehow different meaning, in the field of immunology. This approach enabled von Förster (1976) to propose an interesting interpretation, in terms of observation and *self-organization* (or *autopoiesis*), of Piaget's (1937) genetic epistemology, according to which the way an individual sees the world is built up progressively. Later eigenbehaviors were also considered by Varela (1980) in his researches on *autonomy* and self-organization, initiated by Maturana (1969), who introduced the idea of *operational closure.* We may note also the notion of *double bind,* or contradictory constraints, introduced in 1956 by Bateson (1972) in his theory of schizophrenia.

2.3.3 Cognition and Action We have seen that there are some links between cybernetics and game theory (*q.v.*). This is one of the reasons why cybernetics may be placed, at least partly, under the aegis of *action.* L. Couffignal even considered cybernetics as "the art of making action efficient." In an article entitled "Behavior, purpose and teleology," Wiener and some colleagues felt that the field that was to become cybernetics had something to do with action, that teleology, by definition, is involved each time a negative-feedback device tends to realize some predetermined purpose, or when a conscious being tries to achieve a given end (Rosenblueth *et al.,* 1943). The connotation with action is also to be found in the concept of controllability as well as optimal control.

The theory of *ago-antagonism,* presented by E. Bernard-Weil in his medical thesis of 1955 and much developed later, shows the role of simultaneous actions, of opposite natures, in the control of nonlinear systems, particularly in the field of medicine (Bernard-Weil, 1975). But action is not to be considered alone since it must be based upon *cognition.* This leads to *epistemo-praxiology,* where one considers a cybernetical system able to observe, decide, and act and so modify itself and the rest of the universe (Vallée, 1974, 1995). This approach involves new aspects of subjectivity due to *inverse transfer* of both epistemological and praxiological traits of the system considered. Obviously cognition and action lead to the science of the artificial as foreseen by Simon (1969).

2.3.4 Links with Systems Theory and Synergetics Ludwig von Bertalanffy (1901–1972) introduced the concept of *general systems* (or *system*) *theory,* a translation of the German words "allgemeine Systemlehre" (von Bertalanffy, 1945), often improperly seen as a theory of "general systems" instead of a general theory of systems (von Bertalanffy, 1968). For von Bertalanffy, a system is a set of elements in interaction. So we may consider cybernetics, involving elements interacting through feedbacks and flows of information, as a part of general system theory. In brief we can say *systems theory* or *systemics,* remembering, from a historical point of view, that A. Bogdanoff had partly forseen it, in 1912, with his tectology or universal science of organizations.

Among the main concepts, or ideas, of systems theory that have links with cybernetics or with cybernetics of the second order, we can quote homeostasis, local and global points of view, open systems, order and noise, ... The concept of *homeostasis* was used by C. Bernard, in his introduction to experimental medicine (1865), to express the stability of physiological parameters such as blood pressure or temperature, but the word was chosen by Cannon (1930) and used by von Bertalanffy. It is typically cybernetical and implies negative feedbacks. It led Vendryès (1942) to a theory of general autonomy. Up to a certain point, the opposition between *local* and *global* points of view exists already in cybernetics if we consider the existence of long informational transmission chains. Systems theory is more radical and often insisted upon the importance of a global view, opposed to the mere local one leading to the so-called reductionist approach that proposes decomposition of physical systems into parts as the best way to study them. This attitude has been often attributed to Descartes who, in fact, had in mind only the decomposition of logical difficulties into parts. Local and global points of view have not to be taken too literally; both are necessary, as Pascal pointed out. For von Bertalanffy in an *open system* there is a flow of matter and energy (and even information) through its boundary. A cell and a living being are examples of open systems as well as *dissipative structures* considered by Prigogine (1969). This concept helps cybernetics to extend its domain of application. The idea that it is possible to get *order from noise* was proposed by von Förster (1960). This is not to be understood in a naive way. It must mean that noise can play a triggering role in the emergence of a potential ordered structure, so giving rise to self-organization, a concept of second-order cybernetics. This happens with a physical system, subject to a possible *bifurcation* in the evolution of its state, when, because of some fluctuation generated by noise, a certain parameter reaches a critical value.

With this type of emergence we can see some links of cybernetics with *synergetics*, an interdisciplinary field created, in 1969, by Hermann Haken (see Sec. 1), dealing with a cooperation of parts of a system able to generate self-organization. In particular, synergetics deals with open systems and order parameters that act both as control and information agents.

2.4 A Tentative Chronology of Cybernetics

Obviously the concepts upon which cybernetics is based have interfered in a very complex way, generating new concepts and interacting with other fields more or less close to it, as we have seen in the previous sections. Neglecting details and too remote historical influences, we present now a tentative diagram (Fig. 12) showing the main interactions between concepts and fields taking chronology into account.

2.5 Fields of Application and Directions of Development

One of the roles of cybernetics in applications is the inspiration it gave in many

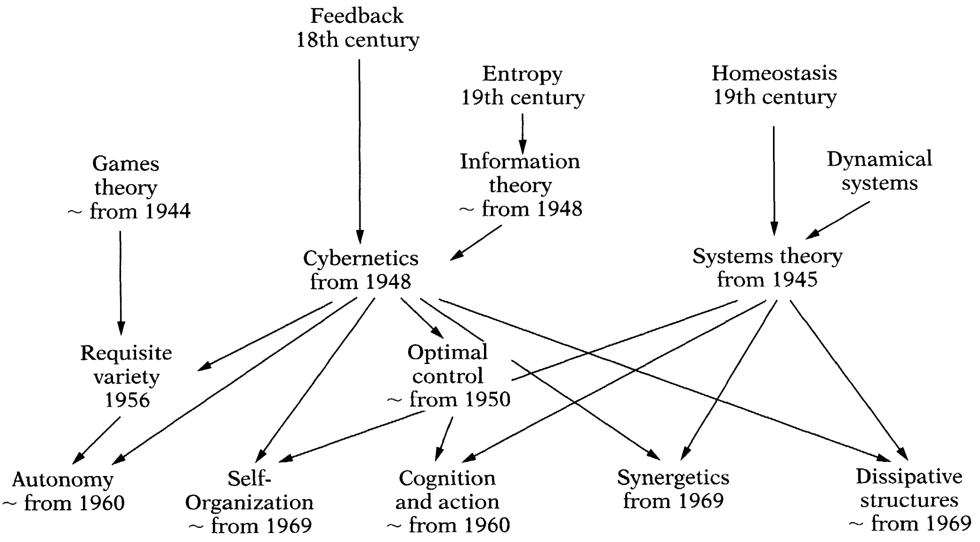
Feedback
18th century
Entropy
19th century
Homeostasis
19th century
Games theory
~ from 1944
Information theory
~ from 1948
Dynamical systems
Cybernetics
from 1948
Systems theory
from 1945
Requisite variety
1956
Optimal control
~ from 1950
Autonomy
~ from 1960
Self-Organization
~ from 1969
Cognition and action
~ from 1960
Synergetics
from 1969
Dissipative structures
~ from 1969

FIG. 12. Chronology of cybernetics.

fields, from philosophy to pedagogy. Here we shall only mention briefly some of them.

The whole domains of automatics, control engineering, and even *fuzzy control,* based upon the theory of fuzzy sets where elements have a degree of belonging to a set (Zadeh, 1965; see THEORY AND APPLICATIONS OF FUZZY LOGIC), are greatly indebted to cybernetics just as cybernetics has been indebted to the theory of regulators at its beginning. More precisely, *robotics,* with the role played by artificial organs of perception and effectuation, may be seen as inspired by cybernetics. We can include here *computer science* (or *informatics*) and also *artificial intelligence,* pattern recognition, and more recently *artificial life,* which aims at the realization of artifacts reproducing traits of natural life. The whole field of communications, with the part played by information theory and signal processing, and the importance taken by the possibilities opened, under the name of cyberspace, by worldwide networks, is essentially cybernetical.

Biological modeling, and biomedical systems, with homeostasis and self-organization, are domains of application of cybernetics as Wiener had foreseen it, as well as neurology and brain studies.

Economics and *management,* with the works of Lange (1962) and Beer (1972), have also found interesting bases in cybernetics, as well as the fields of *ecology* and *sociology* where modeling plays a part.

In what directions can we expect cybernetics to be developed? Obviously artificial intelligence, artificial life, robotics, and all the world of the artificial offers a very favorable field, as well as those of biology, sociology, economics, and management. This is not surprising if we look at the table of contents of Wiener's book on cybernetics, which is partly a program of research not yet completed. But there are other possibilities in the domains of control and communication. Feedback has been mainly studied in the linear case, be it unidimensional or multidimensional. Indeed, the nonlinear case has not been forgotten, but it is far from being completely known if indeed it ever can be: New examples of unstability of cybernetical systems have arisen involving "chaotic" behaviors. Classical information theory suffers from the important limitation, pointed out by Shannon, that it does not take into account the semantical aspects. Despite interesting approaches, semantics of information still needs to be developed. The links between quantity of information, entropy, and even complexity are still arousing a great interest and offer challenging problems.

GLOSSARY

Associative Memory: Memory that complements an incomplete set of data to a complete set on account of stored data or rules.

Attractor: A set of points in state space toward which the state vector is attracted.

Basin of Attraction: Set of points in state space from where the state vector can be attracted toward the attractor.

Center Manifold: Set of variables with neutral stability.

Center Manifold Theorem: A mathematical theorem that allows one to express the variables of the stable manifold by means of those of the center manifold.

Chaos: In this article chaos is meant as deterministic chaos.

Critical Fluctuations: Close to a phase transition or nonequilibrium phase transition the state vector undergoes particularly large fluctuations.

Critical Slowing Down: Close to a phase transition or nonequilibrium phase transition the state vector relaxes very slowly.

Fixed Point, Stable: Point attractor.

Information: This word is used in the sense of Shannon and measures the number of bits.

Limit Cycle, Stable: Attractor connected with a periodic motion.

Lyapunov Exponent: Exponent characterizing the convergence or divergence of initially nearby trajectories.

Quasiperiodic Motion: The time dependence of the state vector can be described as a superposition of exponential functions with exponents containing linear combinations of several frequencies with integer coefficients.

Stability: A system is stable when after a perturbation it relaxes toward its former state.

Symmetry Breaking: Close to a phase transition or nonequilibrium phase transition the order parameter(s) make a choice between different equally possible states.

Synergetics: Word constructed from Greek meaning science of cooperation.

Trajectory: Curve in state space along which the state vector moves according to the evolution equation.

Works Cited

Ampère, A. M. (1843), *Essai sur la Philosophie des Sciences,* Vol. 2, Paris: Bachelier.

Arecchi, F. T., Harrison, R. G. (1987), *Instabilities and Chaos in Quantum Optics,* Berlin: Springer.

Arnol'd, V. I. (1983), *Geometrical Methods in the Theory of Ordinary Differential Equations,* Berlin: Springer.

Arnol'd, V. I. (Ed.), Kazarinoff, N. D. (Transl.) (1993), "Dynamical Systems V. Bifurcation Theory and Catastrophe Theory," in: *Encyclopedia of Mathematical Sciences,* Vol. 5, Berlin: Springer.

Ashby, W. R. (1956), *An Introduction to Cybernetics,* London: Chapman and Hall.

Bateson, G. (1972), *Steps to an Ecology of Mind,* New York: Ballantine.

Beer, S. (1972), *The Brain of the Firm,* London: Penguin.

Bellman, R. G. (1957), *Dynamic Programming,* Princeton, NJ: Princeton Univ. Press.

Bernard-Weil, É. (1975), *L'Arc et la Corde, un Modèle d'antagonisme Dialectique en Biologie et Sciences Humaines,* Paris: Maloine.

Bestehorn, M., Fantz, M., Friedrich, R., Haken, H. (1993), "Hexagonal and Spiral Patterns of Thermal Convection," *Phys. Lett. A* **174,** 48–52.

Bråten, S. (1982), "Simulation and Self-Organization of the Mind," in: G. Floisted (Ed.), *Contemporary Philosophy, A New Survey,* Vol. 2, The Hague: Martinus Nijhoff, pp. 189–218.

Brillouin, L. (1956), *Science and Information Theory,* New York: George Braziller.

Cannon, W. B. (1930), *The Wisdom of the Body,* London: Norton.

Cellérier, G., Papert, S., Voyat, G. (1968), *Cybernétique et Epistémologie,* Paris: Presses Universitaires de France.

Conant, R. C. (1969), "The Information Transfer Required in Regulatory Processes," *IEEE Trans. Syst. Sci. Cybern.* **5,** 334–336.

Cross, M. C., Hohenberg, P. C. (1993), "Pattern Formation outside of Equilibrium," *Rev. Mod. Phys.* **65,** 851–1112.

Eigen, M. (1971), *Naturwissenschaften* **58,** 465–523.

Eigen, M., Schuster, P. (1977), *Naturwissenschaften* **64,** 541–565.

Eigen, M., Schuster, P. (1978), *Naturwissenschaften* **65,** 7–41, 341–369.

Friedrich, R., Uhl, C., Haken, H. (1995a), *Z. Phys. B* **92,** 211–219.

Friedrich, R., Uhl, C., Haken, H. (1995b), *Phys. Rev. E* **51,** 3890–3900.

Gardiner, C. W. (1983), *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences,* Berlin: Springer.

Glansdorff, P., Prigogine, I. (1971), *Thermodynamics of Structure, Stability and Fluctuations,* London: Wiley.

Guckenheimer, J., Holmes, P. J. (1993), *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields,* Berlin: Springer.

Haken, H. (1969), Lectures on Nonequilibrium Phase Transitions at Stuttgart University (unpublished).

Haken, H. (1983), *Synergetics, an Introduction,* 3rd ed., Berlin: Springer.

Haken, H. (1985), *Laser Light Dynamics,* Amsterdam: North-Holland.

Haken, H. (1987), *Advanced Synergetics,* 2nd ed., Berlin: Springer.

Haken, H. (1988), *Information and Self-Organization,* Berlin: Springer.

Haken, H. (1990), *Synergetic Computers and Cognition,* Berlin: Springer.

Haken, H. (1996), *Principles of Brain Functioning,* Berlin: Springer.

Haken, H., Graham, R. (1971), *Umschau* **6,** 191–195.

Haken, H., Kelso, J. A. S., Bunz, H. (1985), *Biol. Cybern.* **51,** 347–356.

Haken, H., Haas, R., Banzhaf, W. (1989), *Biol. Cybern.* **62,** 107–111.

Hartley, R. (1928), "Transmission of Information," *Bell Syst. Tech. J.* **7,** 535–563.

Hopfield, J. J. (1982), *Proc. Natl. Acad. Sci. U.S.A.* **79,** 2554–2558.

Jaynes, E. T. (1957a), *Phys. Rev.* **106,** 620–630.

Jaynes, E. T. (1957b), *Phys. Rev.* **108,** 171–190.

Jaynes, E. T. (1962), "Information Theory," in: K. W. Ford (Ed.), *Lectures in Theoretical Physics, Brandeis Summer Institute, 1962,* Vol. 3, *Statistical Physics,* New York: Benjamin.

Jerne, N. (1974), "Towards a Network Theory of the Immune System," *Ann. Immun. Inst. Pasteur* **125c,** 373–389.

Kálmán, R. E. (1960), "Contributions to the Theory of Optimal Control," *Bol. Mat. Mex.* **5,** 102–109.

Kelley, A. (1967), in: R. Abraham, J. Robbin (Eds.), *Transversal Mappings and Flows,* New York: Benjamin.

Kelso, J. A. S. (1995), *Dynamic Patterns,* Cambridge, MA: MIT Press.

Kohonen, T. (1987), *Associative Memory and Self-Organization,* 2nd ed., Berlin: Springer.

Kolmogoroff, V. A. (1941), "Interpolation und Extrapolation von Stationären Zufälligen Folgen," *Bull. Acad. Sci. URSS, Sér. Math.* **5,** 3–14.

Lafitte, J. (1932), "Réflexions sur la Science des Machines," *Cah. Nouv. J.* **1932,** 21.

Landau, L. D., Lifshitz, E. M. (1959), *Statistical Physics,* Course of Theoretical Physics Vol. 5, London: Pergamon.

Langthaler, W., Schiepek, G. (Eds.) (1995), *Selbstorganisation und Dynamik in Gruppen,* Münster: LIT Verlag.

Löfgren, L. (1997), "Autology," in: M. D. Singh (Ed.), *Systems and Control Encyclopedia: Theory, Technology, Applications,* Oxford: Pergamon.

Mandelbrot, B. (1953), "Contribution à la Théorie Mathématique des Jeux de Communication," *Publ. Inst. Statist. Univ. Paris* **2,** 80–102.

Mandelbrot, B. (1983), *The Fractal Geometry of Nature,* New York: Freeman.

Maturana, H. (1969), "The Neurophysiology of Cognition," in: P. Garvin (Ed.), *Cognition: A Multiple View,* New York: Spartan Books.

McCulloch, W. S., Pitts, W. (1943), "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bull. Math. Biophys.* **5,** 1–90.

Meinhardt, H. (1982), *Models of Biological Pattern Formation,* London: Academic.

Nicolis, G. (1984), "Physics of Far-from-Equilibrium Systems and Self-Organization," in: P. Davies (Ed.), *The New Physics,* Cambridge, U.K.: Cambridge Univ. Press.

Nicolis, G., Prigogine, I. (1977), *Self-Organization in Nonequilibrium Systems,* New York: Wiley.

Nyquist, H. (1924), "Certain Factors Affecting Telegraph Speed," *Bell Syst. Tech. J.* **3,** 324–346.

Odobleja, S. (1938, 1939), *Psychologie Consonantiste,* Vols. 1 and 2, Paris: Maloine.

Pask, G. (1975), "Cybernetics Theory of Cognition and Learning," *J. Cybern.* **5,** 1–90.

Piaget, J. (1937), *La Construction du Réel chez l'Enfant,* Neuchâtel: Delachaux et Nestlé.

Pliss, V. A. (1964), *Izv. Akad. Nauk SSSR, Mat. Ser.* **28,** 1297–1324.

Pontryagin, L. S. (1962), *The Mathematical Theory of Optimal Processes,* New York: Interscience.

Postelnicu, P. (1968), "Ipotesa Compexului Vicios sau a Sistemelor cu Conexiune Inversa," *Telecomunicatii* **12,** 455–458.

Prigogine, I. (1969), *Etude Thermodynamique des Phénomènes Irréversibles,* Paris: Dunod.

Risken, H. (1989), *The Fokker–Planck Equation,* Berlin: Springer.

Rosenblueth, A., Wiener, N., Bigelow, J. (1943), "Behavior, Purpose and Teleology," *Philos. Sci.* **10,** 18–24.

Ross, J., Müller, S. C., Vidal, C. (1988), *Science* **240,** 460–465.

Sanders, L. (1992), *Système de Villes et Synergétique,* Paris: Anthropos.

Sattinger, D. H. (1980), *Group Theoretic Methods in Bifurcation Theory,* Lecture Notes in Mathematics Vol. 762, Berlin: Springer.

Schöll, E. (1987), *Nonequilibrium Phase Transitions in Semiconductors,* Berlin: Springer.

Schuster, H. G. (1988), *Deterministic Chaos: An Introduction,* 2nd rev. ed., Weinheim: VCH.

Scorer, R. (1972), *Clouds of the World,* Melbourne: Lothian.

Shannon, C. E., Weaver, W. (1949), *The Mathematical Theory of Communication,* Urbana, IL: Univ. of Illinois Press.

Simon, H. (1969), *The Science of the Artificial,* Cambridge, MA: MIT Press.

Swift, J., Hohenberg, P. C. (1977), *Phys. Rev. A* **15,** 319–328.

Swinney, H. L., Fenstermacher, P. R., Gollub, J. P. (1977), in: H. Haken (Ed.), *Synergetics, A Workshop,* Springer Series in Synergetics Vol. 2, Berlin: Springer.

Szílard, L. (1929), "Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen," *Z. Phys.* **59,** 840–856.

Thom, R. (1975), *Structural Stability and Morphogenesis,* Reading, MA: Benjamin.

Turing, A. M. (1952), *Philos. Trans. R. Soc. London B* **237,** 37–52.

Vallée, R. (1951), "Sur Deux Classes d' 'Opérateurs d'Observation'," *C. R. Acad. Sci.* **233,** 1350–1351.

Vallée, R. (1974), "Observation, Decision and Structure Transfers in Systems Theory," in: R. Trappl, F. Pichler (Eds.), *Progress in Cybernetics and Systems Research,* Vol. 1, Washington: Hemisphere Publishing, pp. 15–20.

Vallée, R. (1995), *Cognition et Système, Essai d'Epistémo-praxéologie,* Lyon-Limonest: L'Interdisciplinaire.

Varela, F. (1980), *Principles of Biological Autonomy,* New York: Elsevier North-Holland.

Vendryès, P. (1942), *Vie et Probabilité,* Paris: Albin Michel.

von Bertalanffy, L. (1945), "Zu einer allgemeinen Systemlehre," *Blaett. Dtsch. Philos.* **18** (3, 4).

von Bertalanffy, L. (1950), *Science* **111,** 23–29.

von Bertalanffy, L. (1953), *Biophysik des Fließgleichgewichts,* Braunschweig: Vieweg.

von Bertalanffy, L. (1968), *General System Theory,* New York: George Braziller.

von Förster, H. (1960), "On Self-Organizing Systems and Their Environments," in: M. C. Yovits, E. Cameron (Eds.), *Self-Organizing Systems,* London: Pergamon, pp. 31–50.

von Förster, H. (1976), "Objects: Tokens for Eigen-behaviors," *Cybern. Forum* **8,** 91–96.

von Neumann, J., Morgenstern, O. (1944), *Theory of Games and Economic Behavior,* Princeton, NJ: Princeton Univ. Press.

Weidlich, W., Haag, G. (1983), *Concepts and Models of a Quantitative Sociology,* Berlin: Springer.

Weiss, C. O., Klische, W. (1984), *Opt. Commun.* **51,** 47–48.

Wiener, N. (1948), *Cybernetics or Control and Communication in the Animal and the Machine,* Paris: Hermann et Cie.; Cambridge, MA: The Technology Press; New York: Wiley; 2nd ed., with two additional chapters, Cambridge, MA: MIT Press.

Wiener, N. (1949), *Extrapolation, Interpolation and Smoothing of Stationary Time Series,* New York: Wiley.

Zadeh, L. (1965), "Fuzzy Sets and Systems," in: J. Fox (Ed.), *Systems Theory,* Brooklyn: Polytechnic Press.

Further Reading

Andrew, A. M. (1989), *Self-Organizing Systems,* New York: Gordon and Breach.

Auger, P. (1989), *Dynamics and Thermodynamics in Hierarchically Ordered Systems,* London: Pergamon.

Carrier, B. (1992), "Multidimensional Capital and Dynamical Systems," *Kybernetes* **21** (6), 31–42. Cybernetics applied to economics.

Ebeling, W., Feistel, R. (1982), *Physik der Selbstorganisation und Evolution,* Berlin: Akademie-Verlag.

Haken, H. (Ed.), Springer Series in Synergetics, Vols. 1–67, Berlin: Springer.

Jarón, J. (1978), *Systemic Prolegomena to Theoretical Cybernetics,* Scientific Papers of the Institute of Technical Cybernetics of Wrocław Technical University Vol. 45, Monograph 8, Wrocław: Wrocław Technical University. A general text; an original approach taking action into account.

Klir, G. (1969), *An Approach to General Systems Theory,* New York: Van Nostrand. A general text in cybernetics.

Lange, O. (1962), *Introduction to Economic Cybernetics,* Oxford: Pergamon. A classical text translated from Polish.

Lorigny, J. (1992), *Les Systèmes Autonomes,* Paris: Dunod. New aspects of autonomy.

Mesarović, M. D. (1964), "Foundation for a General Systems Theory," in: M. D. Mesarović (Ed.), *Views on General Systems Theory,* New York: Wiley. A general text on cybernetics.

Mesarović, M. D., Macko, D., Takahara, Y. (1970), *Theory of Hierarchical, Multilevel Systems,* New York: Academic.

Mikhailov, A. S., Loskutov, A. Yu. (1996), *Foundations of Synergetics I and II,* 2nd ed., Berlin: Springer.

Van Gigch, J. P. (1987), *Decision Making about Decision Making. Metamodels and Metasystems,* Tunbridge Wells, U.K.: Abacus.

Velarde, M., Normant, C. (1980), "Convection," *Sci. Am.* **243** (1), 92–108.

Walliser, B. (1977), *Systèmes et Modèles,* Paris: Seuil. A general text on cybernetics; a very comprehensive presentation.

Zhang, Wei-Bin (1991), *Synergetic Economics,* Berlin: Springer.

SYNTHESIS, CHARACTERIZATION, AND PROPERTIES OF MODEL POLYMERS

KAREN I. WINEY, *Materials Science and Engineering Department, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.*

LEWIS J. FETTERS, *Corporate Research Science Laboratory, Exxon Research and Engineering Co., Clinton Township, Annandale, New Jersey, U.S.A.*

EDWIN L. THOMAS, *Materials Science and Engineering Department, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*

Introduction 429
1. **Synthesis of Model Polymers** 430
1.1 Living Anionic Polymerizations 430
1.1.1 Linear Polymers 431
1.1.2 Functionalized Polymers 433
1.1.3 Polymers with Complex Architectures 434
1.2 Stereospecific and Sequence-Specific Polymerizations 435
1.3 Recombinant DNA Methods ... 436
2. **Characterization of Model Polymers** 437
2.1 Molecular Weight and Polydispersity 437
2.2 Functionality and Architecture 439
3. **Model Polymers in Polymer Physics** 440
3.1 Morphology 440
3.1.1 Crystallinity in Stereospecific Polymers 441
3.1.2 Self-Assembly in Block Copolymers 442
3.2 Polymer Dynamics in the Melt 444
3.3 Chain Conformations in Solutions 445
Glossary 447
Works Cited 448
Further Reading 449

INTRODUCTION

From the earliest days of polymer science, the novel and diverse behavior of newly synthesized materials was immediately recognized—for example, after the discovery of the polymerization of nylon, researchers realized that a wide variety of properties was available from a single polymer by careful manipulation of processing conditions. Thus, nylon fibers could be spun into a family of fibers with properties conveniently controlled by extent of draw, draw temperature, post heat-setting treatment, etc.

The wide range in polymeric behavior stems from the enormous spectrum of structural states that polymers can assume: solutions, melts, glasses, crystals, liquid crystals, gels, and rubbers. This in turn stems from the chemical diversity of the macromolecules. Moreover, because of the tremendous conformational diversity of a polymer, a given polymer can yield an immense variety of structures! It is hard to imagine a material replacing rubber's unique physical properties, which essentially derive from conformational diversity.

In addition to technically relevant properties such as reversible deformation of several hundred percent (rubber elasticity) and Young's moduli 1.5 times that of steel (rigid rod fibers), polymer molecules offer a scientist a fertile area for basic study. From flexible chains to rigid rods, scientists have been seeking to understand polymeric behavior. This task is challenging, because polymeric materials are highly complex. They are complex because polymer synthesis often does

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

not result in simple, single-component species, they are complex because frequently their organization in the solid state takes place on several length scales, and they are complex because the relatively long time scales required to reach equilibrium are frequently much longer than experimentally feasible. While a good deal of the complexity of polymers is unavoidable, the difficulties arising from their synthetic complexities have been reasonably addressed via the production of model materials by the modern synthetic polymer chemist.

By *model material,* we mean a species for which the parameters of molecular characterization are very well represented by the ensemble average. That is, every molecule is essentially identical to every other (monodisperse ideal population), or the collection of molecules corresponds to a valid statistical assembly of species arising from a given stochastic process (ideal random population). Experiments conducted on model materials benefit from the specificity of response to the molecular variables and allow much more precise interpretation and testing of structure–property relationships.

Over the past fifty years, polymer chemists have made important advances in the control of macromolecular chain structure. Initially, most polymers were produced by polymerizations in which the chemical linking events were more or less randomly selected from a large number of possibilities, hence making the product macromolecules highly complex species. The chemist's prime goals were high yields and high molecular weight. This fact of dispersity in composition, in stereochemistry, in sequence distribution, and in molecular length meant that the resultant material could only be characterized for *average* molecular properties. And, as became evident, different physical probes were sensitive to different averages or different aspects of the molecular distribution. For example, polydispersity causes the progression of molecular-weight averages $\bar{M}_z > \bar{M}_w > \bar{M}_n$, where $\bar{M}_z$ is the z-average molecular weight (see Sec. 2.1) (measured via sedimentation equilibrium), $\bar{M}_w$ is the weight-average molecular weight (via light scattering), and $\bar{M}_n$ is the number-average molecular weight (via osmometry). Molecular weight and polydispersity will be discussed more in Sec. 2.1.

Given that polymer materials contained fairly broad molecular-weight distributions, early workers by necessity satisfied themselves with more qualitative measurements than were possible with pure materials composed of a collection of identical macromolecules. Quantitative materials science (akin to that currently practiced in, say, the field of semiconductors) is just beginning to take place for polymers with the careful, precise characterization of highly specific behavior replacing the former measurements, which were almost always significantly broadened by the spread in the molecular response stemming from the collection of diverse species in a given "pure" sample. The availability of these model materials has facilitated studies of, for example, polymer dynamics, adsorption, block-copolymer morphology, crystalline-polymer morphology, blend miscibility, diffusion (both tracer and self-diffusion), rheology, and unperturbed chain dimensions.

As we approach the twenty-first century, three activities are especially important in the changeover from such statistically complex materials of the past to the new, precisely defined, materials of the future. First, activity in novel synthetic methods, including recombinant DNA technology, promises synthesis of uniquely specific macromolecules. Second, advances in instrumentation to measure molecular properties of these substances accurately are continually being made. Third, the significant growth of computational power to model and of theoretical power to predict, ever more realistically, such molecules provides good prospects for future understanding. All three of these endeavors, plus the merger of ideas and techniques of the polymer scientist with those of the molecular biologist and the closer coupling of the polymer chemist with the materials scientist and physicist, will contribute in major ways to the success of the "physics of macromolecules."

1. SYNTHESIS OF MODEL POLYMERS

1.1 Living Anionic Polymerizations

Chain polymerizations occur by the successive addition of monomers to a free-radical, cationic, or anionic reactive center.

These methods are commonly applied to the polymerization of vinyl monomers CH_2═CHR, where R represents a simple or complex substituent group. Living polymerizations refer to chain polymerizations that do not exhibit spontaneous termination and consequently maintain their reactive centers indefinitely. While there are pseudoliving polymers having free-radical or cationic reactive centers, strictly living polymers are limited to anionic systems under special reaction conditions. The characteristics of living anionic polymerization have been exploited to create nearly monodisperse homopolymers, linear polymers along which the monomer type changes, end-functionalized polymers, and a host of nonlinear polymer architectures (see Fig. 1).

1.1.1 Linear Polymers Living anionic polymerization systems are obtainable with organolithium initiators and hydrocarbon solvents under conditions that exclude oxygen and proton-donating impurities, such as water. These conditions provide rapid initiation of the propagating chains and a stable anionic reactive center with no effective termination mechanism. Because the propagating chains begin polymerizing nearly simultaneously and grow until all the monomer is consumed, the polymer chains have nearly monodisperse molecular-weight distributions. (Molecular-weight averages and determination are discussed more fully in Sec. 2.1.) The molecular weights of these materials are controlled by the simple expedient of manipulating the ratio of monomer $[M]$ to initiator $[I]$ because each molecule of the latter will generate a growing chain. Thus, from the standpoint of the number-average degree of polymerization, DP_n, we have

$$DP_n = [M]_0/[I]_0, \qquad (1)$$

where the subscript 0 denotes the initial reactant concentrations. These two characteristics, nearly monodisperse homopolymers and controllable molecular weight, are paramount to polymer physicists studying molecular-weight–dependent properties such as viscosity, diffusion, and chain conformation.

Another characteristic of the organolithium-based polymerization systems in hydrocarbon solvents is the possible control of the microstructure of polymers prepared from diene monomers (CH_2═CR—CH═CR'H). The relative amounts of 1,2 and 1,4 addition in a polydiene are altered by modifiers such as ethers (ROR′), amines (NR_3), and diamines (R_2N—CH_2CH_2—NR_2):

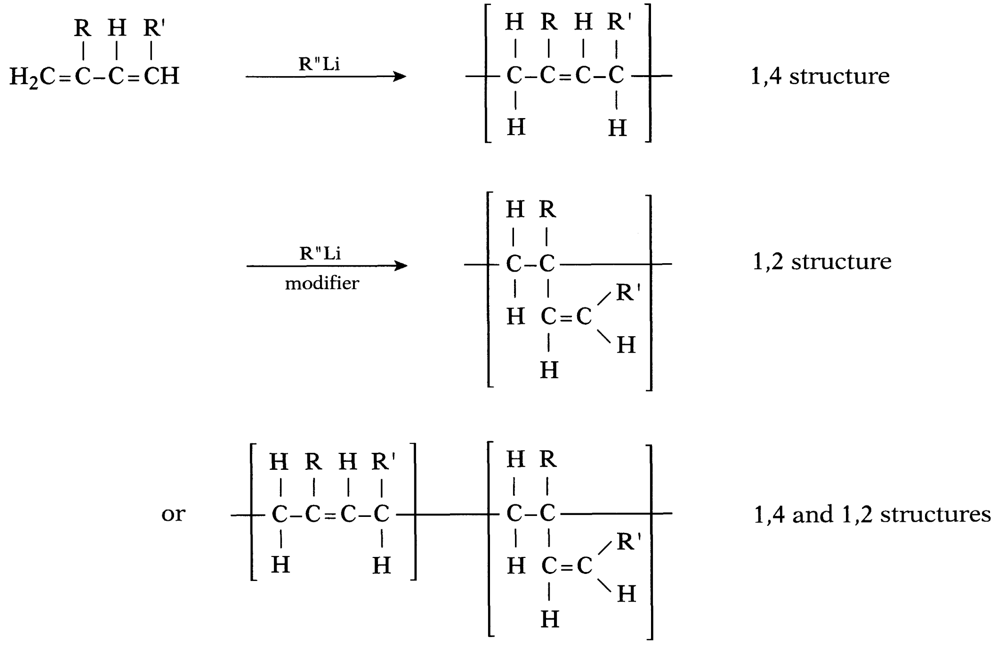

(2)

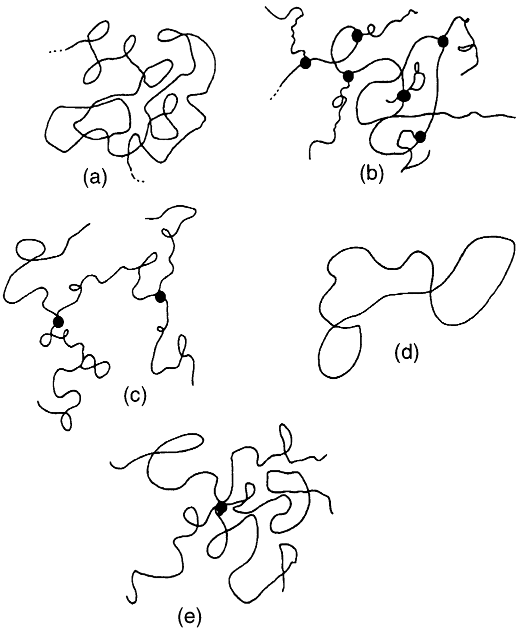

FIG. 1. Schematic of macromolecules with various architectures: (a) linear, flexible chain; (b) graft; (c) H-shaped; (d) ring; (e) star (after Fetters and Thomas, 1993).

The polymers containing both 1,2 and 1,4 addition possess a specific type and level of branching that is not readily accessible by other synthetic routes. The carbon–carbon double bonds of the polydienes can be saturated to produce polyolefins that retain the narrow molecular-weight distribution and the branching of the parent material. Though typically accomplished by catalytic hydrogenation, the saturation reaction can also be performed with deuterium or a mixture of deuterium and hydrogen in order to label the polymers for neutron scattering, forward recoil spectrometry, secondary-ion mass spectrometry, or other experiments using isotopic contrast.

The retention of chain-end reactivity has allowed the preparation of block copolymers through the sequential polymerization of different monomers. The sequential addition of monomers to a reactive center is used routinely in industry to produce diblock and triblock copolymers (two monomer types) and triblock terpolymers (three monomer types). In theory there is no limit to the number of sequential additions of monomer and thereby only practical limits to the number of blocks in a copolymer. Two or more different monomers are used to build up a block copolymer, though there are constraints as to the type of monomers that can be added. For example, methyl methacrylate monomers can be added to a polystyrene propagating reactive center, but styrene monomers cannot be added to poly(methyl methacrylate) propagating chain ends.

The conditions under which these polymerizations are conducted require attention. Generally, an acceptable temperature range is 0 to 50 °C. Higher temperatures will lead to active-center deactivation via, for example, lithium hydride elimination or metallation events (Kern *et al.*, 1972; Broske *et al.*, 1987; Pennisi and Fetters, 1988). If an active center on a propagating polymer chain is displaced to another substance, the reaction is termed chain transfer and results in a decrease in the molecular size of the polymer product. In particular, solvents that can cause chain transfer are to be avoided. Toluene is often cited as the polymerization solvent. This is an unsuitable choice, since the following occurs (Gatzke, 1969):

$$\mathrm{P^-Li^+ + CH_3}\text{—⬡} \rightarrow \mathrm{PH + LiCH_2}\text{—⬡}, \qquad (3)$$

where P^- and P represent the polymer chain with and without a propagating anionic center, respectively. Such reactions clearly will distort the molecular-weight distribution.

Care must also be exercised in the termination of these active centers, since trace amounts of oxygen in, for example, methanol will lead to chain coupling (Fetters and Firer, 1977). Ideally, deactivation of the active centers occurs as follows:

$$\mathrm{P^-Li^+ + CH_3OH \rightarrow PH + CH_3OLi}. \qquad (4)$$

The presence of oxygen can cause the side reaction

$$\mathrm{(P^-Li^+)_2 + \tfrac{1}{2}\,O_2 \rightarrow PP + Li_2O}. \qquad (5)$$

The transition state of the chain ends involves free radicals, which readily couple to yield "doublet" material, with a subsequent distortion of the molecular-weight distribution. This dimerization event is also promoted by the fact that the polymers themselves are associated as dimers in hydrocarbon solvents as a consequence of electron-deficient bonding present in organolithiums.

1.1.2 Functionalized Polymers In recent years there has been a growing interest in the preparation of well-defined polymers with terminal functional groups. The primary motivation for this interest in functional groups is their potential for the following applications:

1. cross-linking reactions in which polymers are bonded to two or more other polymers to form a network of polymers;
2. chain extension or branching;
3. reactions with functional groups on other polymer chains;
4. polymer modification to produce surfactant-like properties—that is, two distinct chemical affinities on a polymer;
5. initiation of polymerization of other monomers; and
6. reversible ionic association.

Some of the feasible structures are schematically given in Fig. 2 where X and Y indicate the locations of functional groups. The most well-defined and exploited route to these materials to date has been lithium-based living anionic polymerizations. Some examples of the functional groups are given in Table 1. The foregoing also serves to summarize some of the nonfunctionalized structures that have been prepared via anionic polymerizations.

One of the most useful and widely used functionalization reactions has been the carboxylation of polymeric anions using carbon dioxide. Although intrinsically a simple reaction, it yields a complex mixture of products involving carboxylated polymer, the dimeric ketone, and trimeric alcohol, as noted previously.

Initially, it was found that the coupling and linking side reactions could be virtually eliminated in systems where the polystyrillithium was complexed with a base such as tetrahydrofuran or tetramethylethylenediamine (Quirk *et al.*, 1989). An alternative route involved the solid-state (achieved by freeze-drying the polystyryllithium) exposure of the active centers to CO_2 at atmospheric pressure. In both cases the yield of carboxylated polystyrene was >99% (Quirk and Yin, 1992; Quirk *et al.*, 1992).

The use of $^{13}CO_2$ revealed that approximately 15% of the —COOH groups were present on the aromatic ring. In other words, although carboxylation efficiency was 99% or greater, nonetheless two types of carboxylic acid groups were present. Ring carboxylation was nearly eliminated for the reaction scheme that involved addition of CO_2 to the modifier-complexed polystyryllithium, in benzene solution.

The foregoing serves to highlight the necessity of a thorough and exhaustive characterization of functionalized polymers in order to establish confidence in the efficacy of the chemistry used. The review of Quirk (1992) is recommended as a thorough brief-

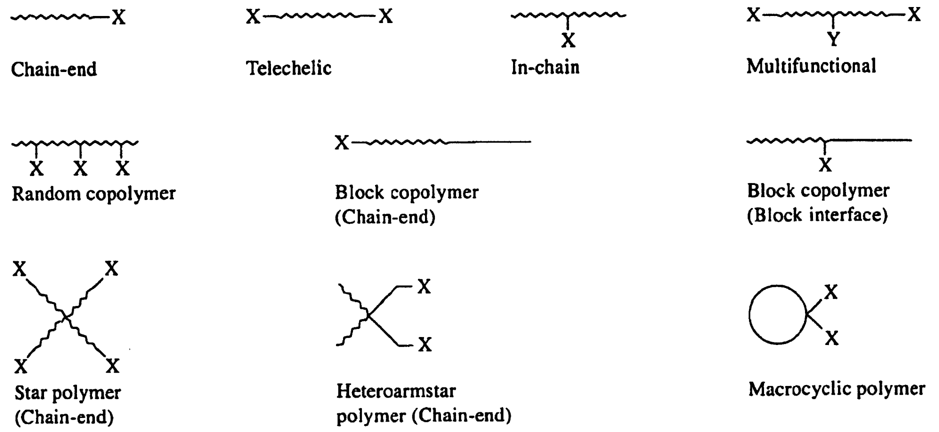

FIG. 2. Schematic of functionalized model polymers with various architectures (after Quirk, 1992). X and Y represent functional groups. Straight and zigzag lines represent different linear polymer chains.

Table 1. Polvmer functional groups via anionic reactions (Quirk, 1992).

Reactant	Group	Side reactions
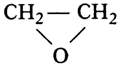	$-OH$	No
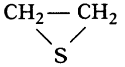	$-SH$	No
CO_2	$-\overset{\overset{O}{\|}}{C}OH$	Yes
SO_2	$-\overset{\overset{O}{\|}}{S}OH$	Yes
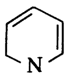	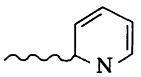	Yes
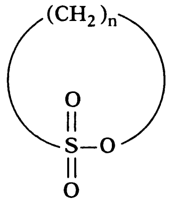	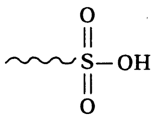	Yes
O_2	$-\overset{\vert}{\underset{\vert}{C}}OOH$	Yes
Br_2	$-Br$	Yes
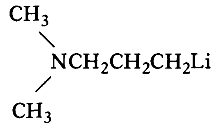	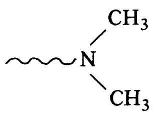	No

ing on this topic. For more about characterization of functionalized polymers see Sec. 2.2.

1.1.3 Polymers with Complex Architectures Living anionic polymerization methods can be used to produce well-defined polymers of complex architecture. The retention of chain-end reactivity allows living polymers (homopolymers or block polymers) to be joined to an existing chain to form a grafted structure [Fig. 1(b)] with well-defined branch length and branch composition. Alternatively, homopolymers or block polymers with reactive chain ends can be bound to one another forming H-shaped or star structures [Figs. 1(c) and 1(e)]. There is a variety of chemical reactions that can be used to build these complex architectures.

The preparation of model star polymers is best approached via the application of chlorosilanes as the linking agents and the lithium counterion in place on the terminal active center. This approach was first used by Morton *et al.* (1962) for the preparation of three- and four-armed polystyrene stars. For example, the reaction for a four-armed star is

$$4P^-Li^+ + SiCl_4 \rightarrow P_4Si + 4LiCl, \qquad (6)$$

where P denotes a linear polymer. Using chlorosilane linking agents, controlled functionalities as high as 128 have been obtained, which are multiply branched "dendrimer"-type molecules (Roovers *et al.*, 1993).

The chemistry in question has allowed the sequential buildup of star structures including

1. asymmetric homopolymeric materials in which the arms of the star are of the same monomer type but of different lengths (Pennisi and Fetters, 1988);
2. single-graft, multicomponent A_2B polymers having a linear backbone of component A and a single-grafted branch of component B (Mays, 1990); and
3. mixed-armed stars, or mikta-armed stars, ABC, in which each of the arms in a three-armed star contains a different monomer type (Hadjichristidis *et al.*, 1993), where A is polyisoprene (PI), B is polystyrene (PS), and C is polybutadiene (PB).

The mikta-armed stars rely on the reaction sequences

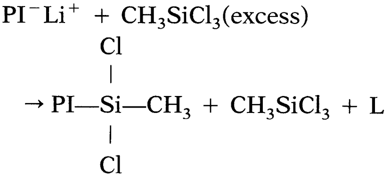

$$PI^-Li^+ + CH_3SiCl_3(\text{excess}) \rightarrow PI\text{—}\overset{\overset{Cl}{\vert}}{\underset{\underset{Cl}{\vert}}{Si}}\text{—}CH_3 + CH_3SiCl_3 + LiCl, \qquad (7a)$$

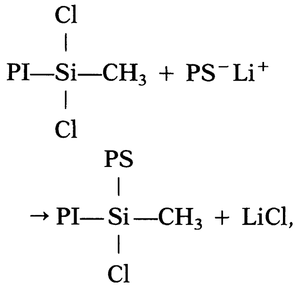

$$\mathrm{PI}-\underset{\mathrm{Cl}}{\overset{\mathrm{Cl}}{\mathrm{Si}}}-\mathrm{CH_3} + \mathrm{PS^-Li^+} \rightarrow \mathrm{PI}-\underset{\mathrm{Cl}}{\overset{\mathrm{PS}}{\mathrm{Si}}}-\mathrm{CH_3} + \mathrm{LiCl}, \tag{7b}$$

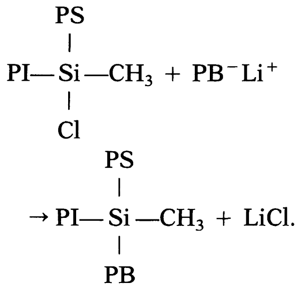

$$\mathrm{PI}-\underset{\mathrm{Cl}}{\overset{\mathrm{PS}}{\mathrm{Si}}}-\mathrm{CH_3} + \mathrm{PB^-Li^+} \rightarrow \mathrm{PI}-\underset{\mathrm{PB}}{\overset{\mathrm{PS}}{\mathrm{Si}}}-\mathrm{CH_3} + \mathrm{LiCl}. \tag{7c}$$

This approach permits the individual arms to be independently characterized along with the resultant mikta-armed star. The combination of size-exclusion chromatography (SEC), nuclear magnetic resonance (NMR) spectroscopy, refractive-index measurements, and light scattering provided the pedigree that unequivocally demonstrated that the desired ABC structure was obtained. The work of Iatraou and Hadjichristidis (1992) is a demonstration of how a model material can be prepared *and* properly characterized.

1.2 Stereospecific and Sequence-Specific Polymerizations

Monomers having a chiral center can be connected head to tail to produce atactic, isotactic, or syndiotactic polymers; see Fig. 3. While living anionic polymerizations are uniquely capable of producing nearly monodisperse molecular-weight polymers, well-defined block polymers, and specified architectures, the anionic synthetic route is not capable of producing stereospecific polymers. Nonetheless, the atactic polymers synthesized by anionic polymerizations are indeed model polymers, because the mean tacticity of the ensemble, i.e. random, sufficiently defines the tacticity of individual molecules.

Stereospecificity occurs when the polymerization reaction includes a coordination mechanism, as is possible with Ziegler–Natta initiators. Since the original discoveries of Ziegler and Natta in the 1950s, an ever-increasing number of initiators of the Ziegler–Natta type has emerged (Boor, 1979). Ziegler–Natta initiators encompass literally thousands of different combinations of a group I–III organometallic compound and a compound of a group IV–VIII transition metal. Subtle changes in these initiators can significantly alter the stereospecificity and the reaction rate, allowing polymerizations to be tailored for specific monomers. Regrettably, much of this tailoring is accomplished empirically on account of the limited knowledge of the mechanisms.

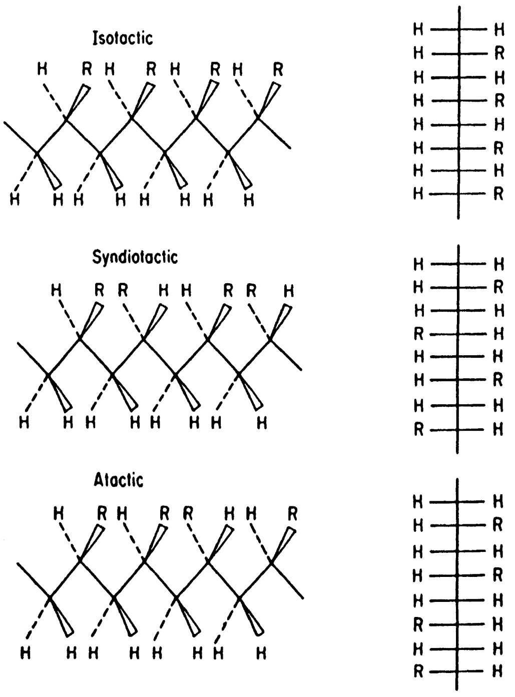

FIG. 3. Different polymer configurations of monosubstituted polyethylene $(CH_2—CHR)_n$. The dotted and triangular lines represent bonds to substituents below and above the plane of the carbon–carbon chain, respectively. (After Odian, 1981.)

As a specific example of stereoregular polymerization, Ishihara *et al.* (1988) have recently succeeded in preparing and characterizing highly syndiotactic polystyrene and substituted polystyrenes. A variety of mixtures containing methylaluminoxane and a titanium compound [$TiCl_4$, $Ti(OEt)_4$, or (η-C_5H_5)$TiCl_3$] yielded syndiotactic product at room temperature. As in the case of living anionic polymerization, measures must be taken to eliminate oxygen and moisture from the materials and the reaction vessel. The ^{1}H NMR spectra are used to delineate clearly between atactic, isotactic, and the new syndiotactic polystyrene. This new material is an important semicrystalline thermoplastic polymer with a high melting temperature (~270 °C), high crystallinity, and excellent chemical resistance.

In addition to stereoregular polymers, metal-catalyzed polymerizations can produce perfectly alternating copolymers of monomers A and B, i.e., . . .ABABABAB. . . (Hirai *et al.*, 1979; Galvin, 1991). For example, Galvin added methyl methacrylate to a toluene solution of an aluminum catalyst ($Et_3Al_2Cl_3$). The subsequent addition of styrene yields a strictly alternating copolymer as determined by ^{1}H and ^{13}C NMR. These materials have been used to investigate the importance of sequence distribution on blend miscibility by comparing to a randomly sequenced copolymer produced by free-radical polymerization (Galvin, 1991; Winey and Galvin, 1993; Winey *et al.*, 1996). Figure 4 shows transmission electron micrographs of two ternary blends containing equal amounts of poly(styrene-*co*-methyl methacrylate) copolymer, homopolystyrene, and homopoly(methyl methacrylate). The blends containing the random and alternating copolymers exhibit three and two phases, respectively, indicating that this small-scale difference in monomer sequence is sufficient to alter bulk miscibility. The study of sequence-specific copolymers awaits synthetic routes that can produce a full range of sequences (random, alternating, multiblock, block) for a selection of monomers having different interaction characteristics.

Note that, to date, these stereoregular and sequence-regular polymers can only be produced with broad molecular-weight ranges. Therefore, fractionation to reduce the polydispersity is required prior to studying molecular-weight–dependent properties.

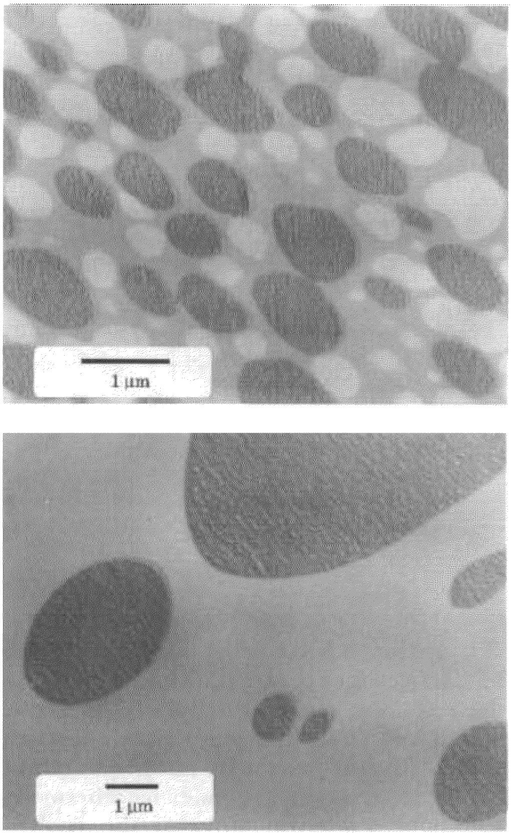

FIG. 4. Transmission electron micrographs of ternary blends containing 33% poly(styrene-*co*-methyl methacrylate), 34% poly(methyl methacrylate), and 33% polystyrene. (top) Copolymer has a random monomer sequence; (bottom) copolymer has a strictly alternating monomer sequence. Striations in the darkest phase are due to knife chatter while microtoming.

1.3 Recombinant DNA Methods

The last class of model polymer synthesis to consider is one in which every aspect of the resultant polymer is model. Recombinant DNA methods produce essentially absolute uniformity of molecular weight, stereochemistry, and sequencing. The biological sciences originally developed these techniques to study the biological function of genes and proteins and to duplicate the properties of naturally occurring structural proteins such as silks and collagens. Recombinant DNA methods are now being used to synthesize new materials designed either to mimic biological activity or to investigate polymer physics.

The recombinant DNA procedure begins with the selection of a specific sequence of amino acids, and then a complementary DNA strand is synthesized producing a synthetic gene. Through a sequence of steps, this synthetic gene replaces a naturally occurring gene sequence in bacterial cells such as *E. coli.* As the cells grow, an amino acid polymer is expressed and subsequently purified. The success of this method requires limited loss of the synthetic gene material from the bacteria, efficient transcription from the synthetic DNA to RNA, efficient translation from the RNA to the new protein, and good stability with limited toxicity of the new protein in the bacteria. Current strategies exist to combat these challenges, and multigram quantities can be successfully prepared by cloning and expressing artificial genes (Tirrell *et al.,* 1991).

Recombinant DNA methods are beginning to be used to construct chemically simple peptides that are capable of mimicking specific biological functions. Such studies attempt to distill the essential aspects of the structure from the highly complex natural peptides and to engineer these aspects into synthetic peptides. Synthetic proteins are being studied as novel electron-transfer and catalytic systems. For example, DeGrado and Dutton have designed and prepared water-soluble, four-helix bundle peptides containing 62 amino-acid residues in an effort to reproduce the extremely high electron-transfer efficiency of photosynthesis (Robertson *et al.,* 1994; Choma *et al.,* 1994). These peptides have specific metal-binding sites, which

1. bind a large rigid macrocycle (heme) to form redox sites and
2. provide tertiary structure, which is common in natural proteins.

The synthetic heme-binding protein can self-assemble and mimic the spectral and electrochemical properties of natural heme-binding proteins. This and other work have established the feasibility of designing functional synthetic proteins. However, a more complete understanding of the complex interdependence of the peptide structure and properties is required to guide the design and selection of peptides prior to synthesis by recombinant DNA methods.

Recombinant DNA methods can also produce new proteinlike polymers for the study of and uses as polymeric materials. In general, these polypeptides are longer and more periodic in their sequencing than natural polypeptides, which has required modification of the biological synthesis methods to accommodate their propensity to crystallize. Of current interest is the extension of the recombinant DNA methods to include non-natural amino acids. Tirrell and coworkers have used selenomethionine in place of methionine in the sequence, $[(\mathrm{Gly}\cdot\mathrm{Ala})_3\cdot\mathrm{Gly}\cdot\mathrm{SeMet}]_9$, which was designed to form lamellar crystals with selenium atoms on the fold surfaces (Krejchi *et al.,* 1994). The extensive incorporation of fluorine has also been accomplished by using *p*-fluorophenylalanine (pfF) in the sequence $[(\mathrm{Ala}\cdot\mathrm{Gly})_3\cdot\mathrm{pfF}\cdot\mathrm{Gly}]_{13}$ (Yoshikawa *et al.,* 1994). These fluoropeptides may exhibit properties related to conventional fluoropolymers such as low surface energy, low coefficient of friction, and excellent solvent resistance.

2. CHARACTERIZATION OF MODEL POLYMERS

Thorough characterization of a polymer is necessary for it to be classified as a model material. The details of polymer characterization are dictated in part by the synthesis method used; thus this section addresses primarily the living anionic synthetic route.

2.1 Molecular Weight and Polydispersity

Molecular-weight characterization of polymers is complicated by the fact that every polymerization route produces species that differ in molecular weight, with the notable exception of the recombinant DNA techniques. Model polymerizations strive to produce materials in which the various molecular-weight moments are, within experimental error, virtually identical.

The advent of size-exclusion chromatography (SEC) (also called gel permeation chromatography, GPC) allows detailed and rapid insight into molecular-weight distributions. SEC is based on the principle that the path a molecule takes through a porous medium depends on its molecular size. In the SEC experiment a polymer solution is injected into a series of columns that are packed with porous beads. A carrier solvent is precisely pumped through the columns, and a detection system at the end of the columns, typically intrinsic viscosity and/or index of refraction, determines the amount of polymer in the elution volume relative to the pure solvent as a function of time. The elution time, that is, the time between the injection of the solution and the elution of polymer, is shortest for the highest molecular-weight fraction of a polymer sample, because larger molecules flow past rather than reside in many of the cavities. Molecular weights are determined from the elution times by calibrating the apparatus with a set of molecular-weight standards.

A nearly monodisperse homopolymer polymerized using organolithium initiators produces a monomodal, narrow peak shape using SEC. Six 30-cm SEC columns are recommended to obtain suitable resolution; the columns may have mixed or single porosity without substantially altering the results. Fewer than six columns will mask complexities in the molecular-weight distributions that arise from less than ideal polymerization or coordination reactions. Figure 5 compares chromatographs obtained using two, four, and six columns, in which the bimodal nature of the particular molecular-weight distribution is most evident with six columns. The SEC data are used to calculate the molecular-weight moments or molecular-weight averages defined as follows: number-average molecular weight,

$$\bar{M}_n = \sum N_i M_i \Big/ \sum N_i; \tag{8}$$

weight-average molecular weight,

$$\bar{M}_w = \sum w_i M_i = \frac{\sum N_i M_i^2}{\sum N_i M_i}; \tag{9}$$

z-average molecular weight,

$$\bar{M}_z = \frac{\sum w_i M_i^2}{\sum w_i M_i} = \frac{\sum N_i M_i^3}{\sum N_i M_i^2}; \tag{10}$$

(z + 1)-average molecular weight,

$$\bar{M}_{z+1} = \frac{\sum w_i M_i^3}{\sum w_i M_i^2} = \frac{\sum N_i M_i^4}{\sum N_i M_i^3}, \tag{11}$$

where $w_i = c_i/c$ denotes the weight fraction of species i in the polymer and N_i is the number of moles of species i with the molecular weight M_i. Note that the other common methods for determining molecular weight determine only one of these moments.

Table 2 shows the ratios of these various moments for anionically polymerized 1,4-polybutadienes with molecular weights as low as 10^3. As indicated by ratios near unity, these SEC results indicate that the polybutadienes have nearly monodisperse molecular-weight distributions. As expected, the heterogeneity indices decrease as molecular weight increases. These three ratios or heterogeneity indices serve as the *complete and necessary* pedigree insofar as the length heterogeneity characteristics of a polymer are concerned.

A cautionary note is in order regarding the oft-cited, and often misleading, ratio of $\bar{M}_w/\bar{M}_n$, the so-called polydispersity index, as

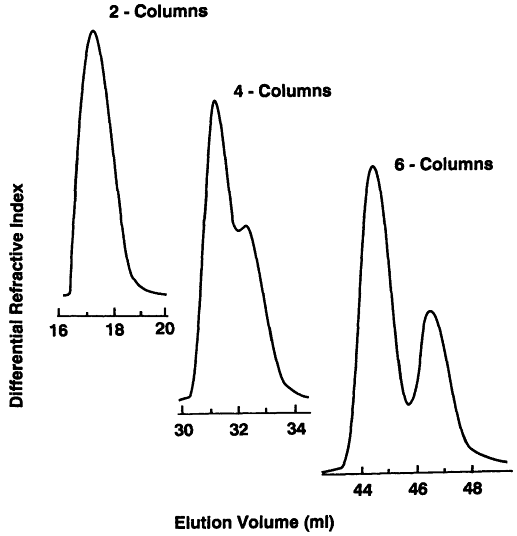

FIG. 5. SEC chromatographs showing improved resolution with increasing number of columns.

Table 2. SEC heterogeneity evaluations for low-molecular-weight polybutadienes.[a]

$\bar{M}_w \times 10^{-3}$ (g/mol)	$\bar{M}_w/\bar{M}_n$	$\bar{M}_z/\bar{M}_w$	$\bar{M}_{z+1}/\bar{M}_z$
1.00	1.08_0	1.07_2	1.06_2
1.05	1.09_8	1.07_8	1.07_6
1.05	1.07_4	1.06_6	1.05_9
1.15	1.07_4	1.06_9	1.06_9
2.00	1.07_4	1.06_6	1.06_6
3.60	1.03_5	1.03_3	1.03_3
4.30	1.02_9	1.02_8	1.02_8
5.90	1.02_4	1.02_4	1.02_4
6.50	1.02_3	1.02_3	1.02_3
13.0	1.02_3	1.02_2	1.02_2
18.0	1.02_4	1.02_3	1.02_3
36.0	1.01_8	1.01_8	1.01_8

---

[a] Polymerization of polybutadiene in cyclohexane at 20 °C with *tert*-butyl lithium initiator.

Table 3. Sample calculations: Equal weights of monodisperse polymers. A = 50 000 g/mole; B = 100 000 g/mole; C = 150 000 g/mole.

Sample	Weight fraction of component			$\bar{M}_n$	$\bar{M}_w$	$\bar{M}_w/\bar{M}_n$
	A	B	C			
I	0.50	0.50	0.00	66 667	75 000	1.125
II	0.33	0.33	0.33	81 818	100 000	1.222

a hallmark of near monodispersity. Table 3 summarizes calculations performed on distinctly bimodal and trimodal systems. The calculated values of $\bar{M}_w/\bar{M}_n$ are 1.125 and 1.22 and will suggest to some a suitably narrow chain-length distribution. However, the SEC chromatographs calculated for these mixtures illustrate the unacceptable character in the molecular-weight distribution; see Fig. 6. In light of the potentially deceptive nature of $\bar{M}_w/\bar{M}_n$ ratios, examples of SEC data must be presented to verify near monodispersity. The term *narrow molecular-weight distribution* must be limited to polymers with $\bar{M}_{z+1}/\bar{M}_z$, $\bar{M}_z/\bar{M}_w$, and $\bar{M}_w/\bar{M}_n$ ratios of <1.05.

A final example of the fallacy of relying only upon an $\bar{M}_w/\bar{M}_n$ ratio as an indicator of monodispersity is given for the polyisobutylene sample of Fig. 7. This polymer, prepared by "living" polymerization, has bimodal character coupled with a high-molecular-weight component. The measured $\bar{M}_w/\bar{M}_n$ value of 1.11 fails completely in being a quantitative description of the distribution profile of that particular material.

Clearly, the mere comment, which is often encountered, that anionic polymerization was used to prepare the polymer in question is an inadequate pedigree. Reaction conditions must be described and their limitations understood. The companion information needed, as has been noted, is proper characterization.

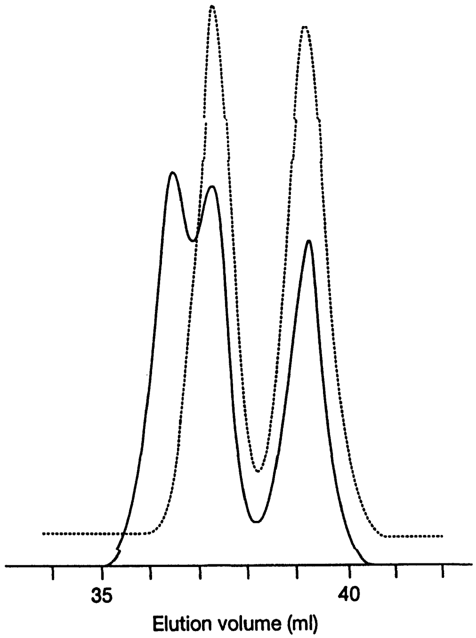

FIG. 6. SEC chromatographs of the polymer mixtures of Table 3; sample I (·······) is bimodal and sample II (———) is trimodal (after Fetters and Thomas, 1993).

2.2 Functionality and Architecture

Unfortunately, many of the functionalization reactions that have been described are not well characterized, as is discussed in a

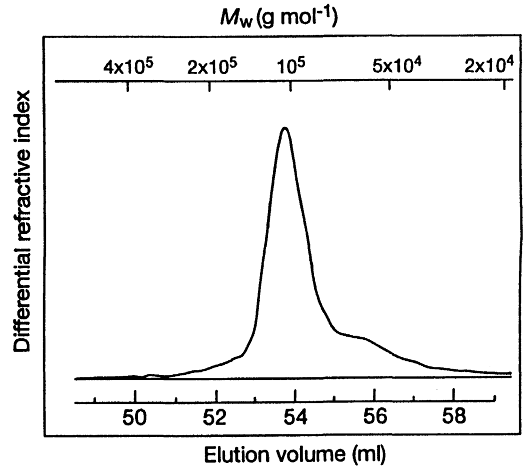

FIG. 7. SEC chromatograph of polyisobutylene standard prepared by a "living" polymerization (after Fetters and Thomas, 1993).

recent detailed review by Quirk (1992). Precedent from the reactions of model small-molecule analogs is often misleading. Thus, adequate, and necessary, characterization probes embrace chain molecular weight, functional group titration, NMR, UV-visible, and FTIR spectroscopy, thin-layer and column chromatography, and chemical reactions. The primary need in chain-end characterization is an accurate evaluation of functionalization efficiencies. To this end, knowledge of functional-group titration, spectroscopic probes, and chemical reactions can all contribute to determining the efficiency of a given reaction. Chromatography has proven useful (see, for example, Quirk and Yin, 1992) for analyzing unfunctionalized polymer mixed with the functionalized product as a consequence of the different migration rates that the two species exhibit. Chain molecular weight, when combined with these techniques, allows a quantitative assessment of the functionalization reaction in question.

As noted earlier, a high-resolution SEC column arrangement must be in place if meaningful distribution profiles are to be obtained. This is particularly true for the case where sequential monomer addition, chain coupling, linking, or branching can occur either intentionally to form complex architectures or unintentionally, as during polymerization or via a post-polymerization event. AB, ABA, or ABC block polymers are simple examples of the former. Accurate insight into the efficiency and success of each step in the preparation of these multicomponent systems can, and must, be evaluated by high-resolution SEC. This is shown in Fig. 8, where a poly(styrene-isoprene) diblock (Quirk and Lee, 1992) and the polystyrene segment (polymerized first) are shown. The lack of a signal for homopolystyrene in the SEC of the diblock clearly demonstrates that the resulting diblock is free of the homopolystyrene contaminant.

An example of unintentional chain coupling occurs in the presence of oxygen during the termination event (Fetters and Firer, 1977). As has been shown, inadequate SEC resolution will miss the presence of such coupled products (Ambler *et al.*, 1977). A similar situation can be encountered in functionalization reactions. For example, the carboxylation of poly(styryllithium), PS^-Li^+, in benzene using high-purity CO_2 at atmospheric pressure,

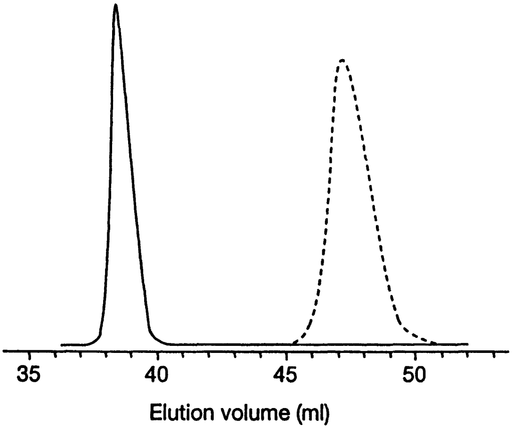

FIG. 8. SEC chromatographs of a poly(styrene-isoprene) diblock (—) and the polystyrene segment (---------) (after Quirk and Lee, 1992).

$$PS^-Li^+ + CO_2 + H_3O^+ \rightarrow PS\text{—}CO_2H + Li^+ + H_2O, \qquad (12)$$

produces the carboxylated polymer, PCO_2H, in only 46–66% yield. However, the contaminating products include the dimeric ketone, P_2CO (25–30%), and the trimeric alcohol, P_3COH (7–24%). Procedures that virtually eliminate the ketone and alcohol components have been developed and presented (Quirk and Yin, 1992; Quirk *et al.*, 1992). The presence of the trimeric alcohol and the dimeric ketone is shown in Fig. 9, with the former component unresolved by the μ-Styragel SEC column arrangement. The higher-plate-count ultra-Styragel column set reveals the presence of that branched component.

3. MODEL POLYMERS IN POLYMER PHYSICS

3.1 Morphology

There are many examples of the use of model polymers to develop novel materials structure. Rational control of microstructure is a particular challenge in polymers, in that one needs to control packing at several length scales simultaneously. Furthermore, each molecule has an enormous number of

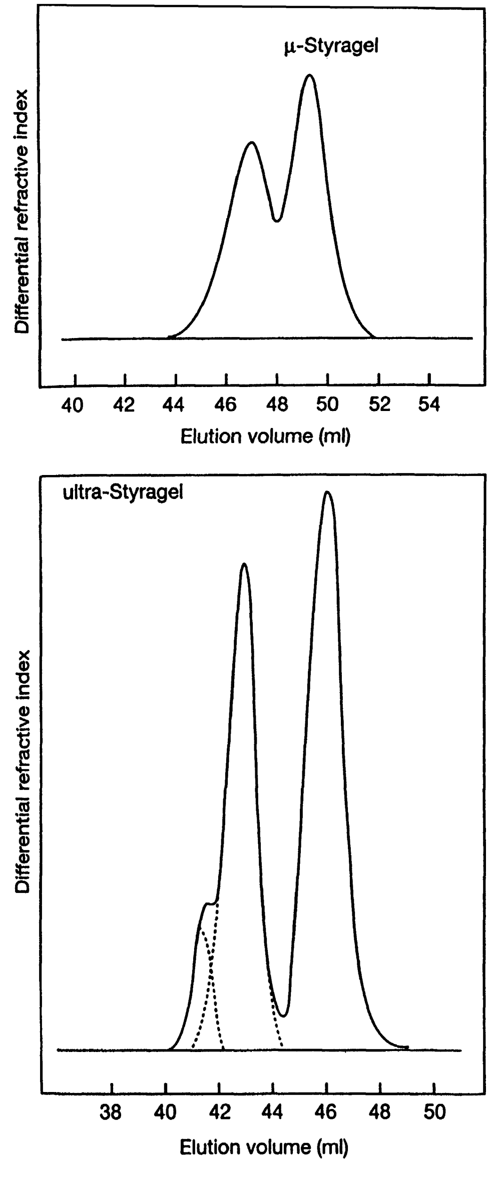

FIG. 9. SEC chromatographs of trimodal polystyrene evaluated by a μ-Styragel set and an ultra-Styragel set. The latter column arrangement reveals the presence of the three-armed star. (After Quirk *et al.*, 1989.)

structural states of similar energy from which to select, as a consequence of their inherent conformational diversity. As indicated, stereochemical and monomer sequence control during species synthesis is the essential aspect for molecular-level control of structure and function.

3.1.1 Crystallinity in Stereospecific Polymers The realization of desirable superstructures arising from the self-assembly of individual species far more complex than simple spheres requires intermolecular assembly to proceed in a uniform manner in order to allow the carefully built-in specificity of chain structure and the applied environmental conditions to produce the prescribed, desired structural state. With polymers, one discovers even with the most simple macromolecules a wide array of structural states and length scales; e.g., for semicrystalline linear polyethylene, the orthorhombic crystalline unit cell has edge lengths about 1 nm and the all-*trans* chains pack into lamellae typically 20 nm thick, which are separated by interlamellar noncrystalline regions wherein a combination of *trans* and *gauche* states results in a random chain structure.

Moreover, crystallization proceeds from nucleation centers to develop spherulites of average diameter of several microns. Inside a spherulite, the lamellae are radially arranged. Such a complex microstructural geometry involving both crystalline and noncrystalline phases makes establishment of structure–property relations difficult; furthermore, the important and practical study of the evolution of structure and physical properties with applied force or imposed strain is challenging both to characterize and to model. For this reason, simplified geometries and uniform structures have been produced to enable exploration of fundamental structure–property relations.

With the advent of synthetic schemes to produce stereoregular polymers, many polyolefins previously restricted to the amorphous state are now available in the semicrystalline state. These new polymers are rapidly becoming important thermoplastic elastomers with enhanced temperature stability and chemical resistance. However, studies of syndiotactic and isotactic polymers have frequently been hampered by the less than perfect configurational regularity of the materials. A prime example of what is possible with highly stereoregular polymers exists for syndiotactic polypropylene, sPP. The improved stereoregularity is the result of switching from vanadium-based catalysts (Boor and Youngman, 1966) to new metallocene catalysts based on Zr and Hf (Ewen *et* *al.*, 1988). These new catalysts have increased the percentage of racemic triads as measured by ^{13}C NMR from 77% to >99%.

More than 25 years of investigation of the less than fully syndiotactic material concluded that the chains formed $(t_2g_2)_2$ helices that pack isochirally in a face-centered unit cell. Recently, Lovinger *et al.* (1993) used x-ray diffraction, electron diffraction, and molecular modeling to identify the stable, high-temperature structure as a fully antichiral, face-centered unit cell (*Ibca* space group) of sPP. This new (and correct) unit cell has dimensions double that of the earlier unit cells. Furthermore, Lovinger *et al.* (1993) showed that the previously proposed models were the result of studying mixtures of syndiotactic and isotactic PP, as illustrated in Fig. 10. The incorporation of defects increases as the crystallizing temperature decreases, as evidenced by streaking of the *h*20 layer of the reciprocal lattice in electron diffraction. This streaking corresponds to a *b*/4 vacancy along the crystal-growth direction to produce isochiral packing. Work by other groups has confirmed these morphological findings, and studies are now underway to understand the structure–property relationships in sPP.

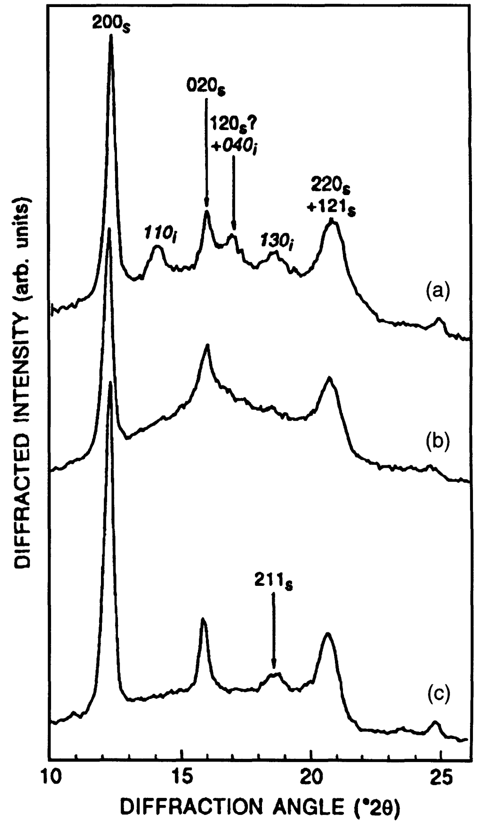

FIG. 10. X-ray diffractograms of (a) poorly syndiotactic polypropylene containing 77% racemic triads, typical of all previously studied sPP samples, (b) fully syndiotactic polypropylene as received, and (c) fully syndiotactic polypropylene crystallized isothermally at 135 °C. Subscripts *i* and *s* refer to isotactic and syndiotactic polypropylene, respectively. (After Lovinger *et al.*, 1993.)

3.1.2 Self-Assembly in Block Copolymers Self-assembly in block polymers produces microdomains composed of the respective blocks. Typical domain sizes are several hundred angstroms, and these domains, in turn, assemble into grains with typical diameters of several microns. The grains of the cylindrical or lamellar microdomain-forming block copolymer have a preferred orientation, resulting in local anisotropy of the material properties. However, since the grains are relatively small in comparison to the whole sample and since they are normally in random orientations, the resultant macroscopic sample exhibits isotropic properties. When such materials are examined by scattering or subjected to mechanical testing, the net response is the superposition of the different responses to each of the various grain orientations. Fundamental understanding of structure–property relationships is thus difficult to achieve.

To overcome such limitations, one can employ special processing to convert the isotropic multigranular sample into a single crystal. An applied bias field (mechanical, electrical, etc.) can result in materials with single-crystal-like texture. Such an approach was first taken by Keller and co-workers (Keller *et al.*, 1970; Folkes *et al.*, 1973), utilizing extrusion to produce highly oriented single-crystal-like material of a triblock copolymer. This process is akin to the texturing of polycrystalline metals by die drawing. While providing a highly oriented sample, this process suffers from the presence of grain-boundary structures and other morphological defects because the alignment process is initiated from an already microphase-separated material. The influence of such defects and boundaries is unknown at present, but in analogy to work with other materials, one can expect a strong influence on properties. Recently, Albalak and Thomas (1994) devised a roll-casting process that eliminates grain boundaries by causing the self-assembly of the block copolymer molecules into the microdomains to take place in an orienting flow field.

The roll-casting process is based on the rolling of an initially homogeneous polymer solution between two counter-rotating cylinders while removing solvent at a controlled rate. The experimental system is schematically pictured in Fig. 11 and consists of two adjacent rollers and is similar to the conventional setup of calenders or roll mills. The polymer solution flows between the rolls through a minimum gap, called the "nip," and coats either one or both of the rolls. The detailed velocity profile in such a geometry has been analyzed (Gaskell, 1950) and contains both shear and elongational flow components. The key to this process is that the onset of microphase separation occurs in the applied flow field so that the microdomains form in a globally oriented structure. Figure 12(a) is a TEM micrograph of a [0001] axial view showing the well-ordered microdomain structure of a PS/PB/PS triblock copolymer containing PS cylinders in the PB matrix. The perfection of the packing of the PS cylinders is quite evident. Low-magnification TEM images show that the single-crystal structure is unperturbed for at least many tens of microns. The degree of order over larger distances can also be assessed by small-angle x-ray scattering (SAXS). Figure 12(b) is a two-dimensional SAXS pattern taken with the incident beam along the [0001] direction. The pattern originates from a sample area on the order of 1 mm^2, demonstrating the single-crystal nature of the macroscopic film. Such specimens offer significant advantages to test phase-morphology and phase-transformation theories and to further the basic understanding of transport and chain dimensions in block-copolymer systems.

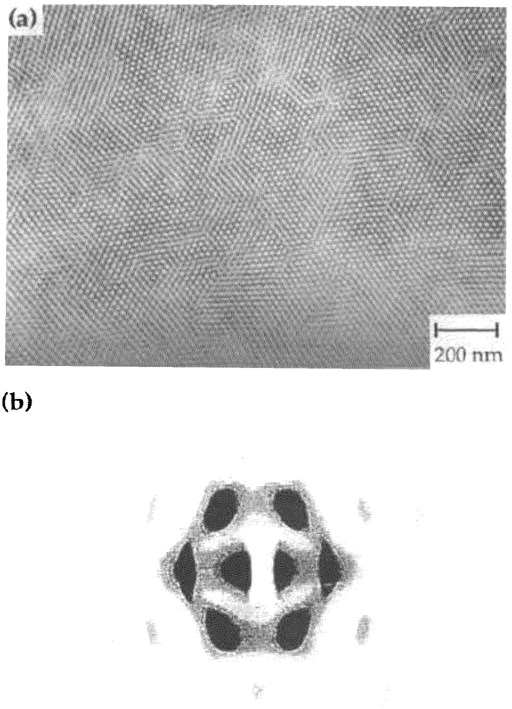

FIG. 12. (a) Bright-field TEM micrograph of an OsO_4 stained, microtomed section of a roll-cast PS/PB/PS film. View is along flow direction and shows the two-dimensional hexagonal lattice packing of the PS cylinders. (b) Two-dimensional small-angle x-ray scattering pattern of roll-cast PS/PB/PS film with incident beam along the flow direction. Bragg reflections of a well-ordered two-dimensional hexagonal structure are evident. (After Albalak and Thomas, 1994.)

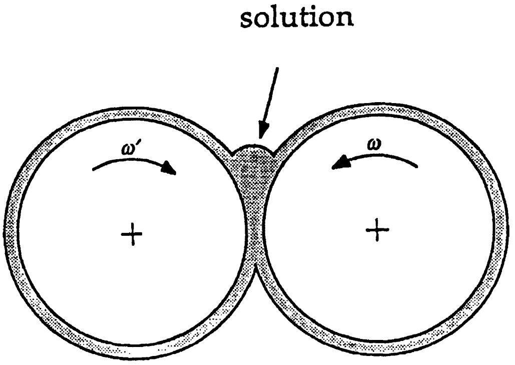

FIG. 11. A schematic representation of the roll-casting apparatus. A concentrated solution of block copolymer is processed between two adjacent counter-rotating cylinders. (After Albalak and Thomas, 1994.)

Model diblock copolymers produced by anionic polymerization and processed into single-crystal specimens are quite useful to probe how the microstructure influences chain dimensions. SAXS and small-angle neutron-scattering (SANS) measurements, as well as theoretical predictions of the domain spacing in A/B amorphous/amorphous block copolymers, indicate that the spacing varies approximately as $M^{2/3}$ in the strong-segregation regime where the individual phases are essentially pure components and the A/B junction is strongly localized in the narrow interfacial region between domains. The individual component block dimensions will therefore be distorted from the Gaussian behavior of a chain in a single-phase melt, where R_g scales as $M^{1/2}$. Measurement of the chain dimensions in a phase-separated block copolymer is, however, complicated by the strong scattering from the periodic domain microstructure, which occurs in the same angular range as the single-chain scattering.

Recently, Matsushita *et al.* (1990) used anionic synthesis to carefully deuterium label *parts* of a block. The selective use of perdeutero monomer provides neutron contrast so as to examine the dimension of a particular portion of the chain with SANS. Two types of triblock copolymers composed of deuterated PS, PS, and poly(2-vinylpyridine) (PVP) were prepared: "end labeled" (dPS-PS-PVP) and "centered labeled" (PS-dPS-PVP). Such studies permit the evaluation of polymer chains in a polymer melt environment. Lamellar orientation resulted from the solvent casting process rather than by shearing. The center portion of the chain adjacent to the A/B diblock junction was found to be contracted along the direction parallel to the interface to the same extent as previous researchers had found for the entire block (i.e., approximately 70%). The portion of the chain at the free end of the block was localized in the middle of the PS domain, and its dimension was unperturbed by the microdomain morphology. Thus, the gradient in chain stretching can be explored in microstructured systems.

3.2 Polymer Dynamics in the Melt

There are many examples of the use of model polymers to evaluate theoretical concepts in polymer physics. Here we detail the employment of model polymers in testing the elegant reptation hypothesis of de Gennes, Doi, and Edwards and the tube concept in reptation (de Gennes, 1990; Doi and Edwards, 1986). According to reptation theory, individual chains are assumed to be constrained by their neighbors and are thus compelled to migrate primarily along their own contours in a snakelike manner. Thus, a polymer melt is made up of an assembly of chains all simultaneously involved in disengagement from their current cage of strands. In this disentanglement process, each chain creates a new cage and configuration as it undergoes its movement through the mesh. The cage can be visualized as a tube of diameter d and contour length l, which encloses each chain in the mesh. The chains are assumed Gaussian, with v chains per unit volume. (Gaussian behavior of polymer molecules is discussed more fully in the next section.) The tube diameter is independent of molecular weight M, while the tube contour length is directly proportional to the molecular weight.

The distance between entanglements, d_t, is expressible in terms of the measurable properties as

$$d_t^2 = 0.8\left(\frac{\langle R^2\rangle_0}{M}\right)\left(\frac{R_gT}{G_N^0}\right), \tag{13}$$

where $\langle R^2\rangle_0$ is the square of the end-to-end distance of the polymer, G_N^0 is the plateau modulus (the short-time modulus of an entangled polymer liquid), R_g is the gas constant, and T is the absolute temperature. The above equation can be reduced to

$$d_t^2 = 0.8[\langle R^2\rangle_0/M]M_e, \tag{14}$$

where M_e denotes the entanglement molecular weight that is determined from the plateau modulus (Doi and Edwards, 1986). Thus, a combination of the static- ($\langle R^2\rangle_0/M$) and the dynamic-based (M_e) parameters allows an evaluation of d_t. Until recently, the *direct* measurement of d_t had not been accomplished.

For that purpose, nearly monodisperse, amorphous poly(ethylene-propylene) (PEP) was used (Butera *et al.*, 1991; Richter *et al.*, 1992). The route to that polymer was via the hydrogenation saturation of anionically prepared polyisoprene. The analytical tool of choice was the IN-11 neutron spin-echo (NSE) spectrometer at the Institut Laue-Langevin in Grenoble. The unique feature of the NSE technique is its ability to evaluate in a direct way the energy changes induced in neutrons during the scattering process. In essence, the dynamic data are gathered by examining the decay in an NSE as a function of an imposed precessional frequency. Therein, a magnetic field of varying strength is applied to the incident polarized neutron beam. The small energy change of each scattered neutron is analyzed as a shift in the precessional phase. In view of the limited time resolution (ca. 4 × 10^{-8} s) and distance resolution (5–100 Å) of the IN-11 instrument, only the most dynamically flexible chains with rapid local motion are candidates. PEP and polyethylene (from the saturation of 1,4-polybutadiene) fulfill these criteria.

The results of such an NSE investigation on PEP are summarized in Fig. 13, which also contains rheological and SANS findings. The tube diameter (NSE) was found to increase with temperature, while the radius of gyration (SANS) and plateau modulus (rheology) were observed to decrease. The plateau modulus also signaled an increase in the entanglement molecular weight, which, when combined with the change in chain dimensions, led to an increase in the tube diameter with increasing temperature. It is important to note that the values of d_t determined by various techniques were in essential agreement, a picture also found for polyethylene (Richter *et al.*, 1992). These results fortify the elegant *Ansatz* of de Gennes's reptation model.

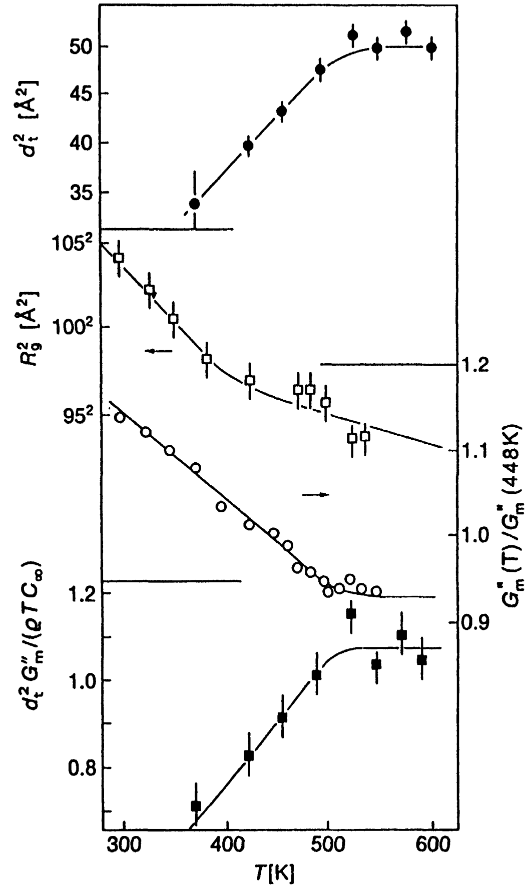

FIG. 13. Entanglement distance d_t, radius of gyration R_G^2, ratio of G''_m to G''_m (448 K), and $d_t^2 G''_m/rTC_\infty$ for alternating poly(ethylene-propylene) as a function of temperature, where G''_m is the maximum in the shear loss modulus of the entanglement network (after Fetters and Thomas, 1993).

3.3 Chain Conformations in Solutions

As mentioned in the previous section, Gaussian statistics are used to describe the dimensions of a polymer chain. Such random-walk statistics imply that each step of the process (i.e., along the polymer chain) is of the same length and fully independent of all other steps. The simplest application of Gaussian statistics to polymers uses the length of a monomer unit, l, and the number of monomer units, $N_w = M/M_0$, where M_0 is the molecular weight of a monomer unit and end effects are neglected. In this case the square of the end-to-end distance of the polymer chain is $N_w l^2$. A more realistic approach includes the bond-angle and bond-rotation restrictions associated with the chemical structure of the polymer chain, which add prefactors to the expression. In lieu of these structural details, the short-ranged steric hindrances can be accounted for by defining a statistical segment length a to be the step length of a random walk, such that the unperturbed square of the end-to-end distance is given as

$$\langle R^2 \rangle_0 = N_w a^2. \tag{15}$$

Experimentally, the statistical segment length is determined by measuring the radius of gyration R_g, using a scattering technique and applying the expression

$$a^2 = 6\langle R_g^2 \rangle / N_w, \tag{16}$$

where $\langle R_g^2 \rangle = \langle R^2 \rangle_0 / 6$, which is an exact solution for a random walk. The characteristic ratio, C_∞, describes the stiffness of a polymer molecule, which is associated with fixed bond angles and short-range steric hindrances and is defined as $\langle R^2\rangle_0/N_w l^2$. The short-range modifications to the random walk are sufficient when the polymer–polymer and polymer–solvent interactions are equivalent, that is, in the θ condition. The θ condition or theta condition refers to a specific combination of polymer, solvent, and temperature in which the polymer chain exhibits Gaussian behavior. Otherwise, longer-ranged interactions, which give rise to excluded-volume effects, need to be incorporated, and Gaussian statistics fail.

Flory (1953) was the first to predict that chain dimensions in a θ condition are identical to those in a polymer melt and that the chain behaves like an ideal random walk. This prediction has been confirmed by SANS and other techniques. Another prediction, the rotational isomeric state (RIS) prediction, involves the interdependence of the chain conformation and the bond-rotation angle. Rather than assign a finite probability to all possible bond-rotation angles as defined by four backbone atoms, the RIS prediction restricts the bond-rotation angle to three discrete values. A probability is assigned to each of the three values, which ignores the possibility that severe steric interactions do occur for certain sequences of bond angles. The use of model polymers has been of great value in the detailed evaluation of RIS predictions pertaining to the polymer chain dimensions and their respective temperature coefficients. The advent of SANS has expanded the range of these evaluations, because both parameters can be measured in the melt state. Some recent findings are presented in Table 4.

The combined results from light scattering in the θ condition, RIS theory, and SANS in the melt show excellent agreement with regard to the parameters in question for PE and PEP. However, such agreement is absent for *a*-PP and *a*-PEE. It can be seen that neither RIS theory nor θ-condition measurements yield chain-dimension temperature coefficients that may be related to what is observed in the melt state. From the RIS standpoint there seemingly is a problem in correctly visualizing the role of nonbonded interactions as a function of temperature. The failure of the RIS approach is duplicated in the failure of the θ-condition results to mimic melt behavior. Thus, the assumption that modeling a polymer in a vacuum universally yields its behavior under θ conditions, which in turn serves as a mimic of melt behavior, is erroneous. Solvent influences on "unperturbed" chain dimensions have been previously predicted (Lifson and Oppenheim, 1960). Our current understanding lacks predictive capacities regarding which solvents can exert so-called specific solvent effects, or, conversely, which polymers are susceptible to such influences. Nonetheless, a partial rationalization as to why solution θ condition is not equivalent to melt behavior is possible.

Rotational isomeric states are identified with the minima of the conformational energy maps, minima that can be influenced by local-polymer solvent interactions—the so-called "specific solvent effect." This effect can be viewed as one that does not create

Table 4. Polyolefin chain parameters at 298 K.

	$R_g/M^{0.50}$ (Å mol$^{0.50}$ g$^{-0.50}$)		C_∞		a (Å)		$d \ln\langle R^2\rangle_0/dT$ (10^3/deg)		
Polymers	LS	SANS	LS	SANS	LS	SANS	LS	SANS	RIS
PE(PEB-2)	0.470	0.485	7.9	8.4	6.1	6.3	−1.1	−1.15	−1.2
PEP	0.381	0.397	6.5	7.1	5.6	5.8	−1.1	−1.16	−1.0
a-PP	0.340	0.337	6.6	6.1	5.6	5.3	−2.9	−0.1	0 to −1.8
a-PEE	0.279	0.283	5.5	5.7	5.1	5.2	−2.3	0.4	−0.1

PE = polyethylene; PEB-2 = poly(ethylene-butene) with ca. 2 ethyl branches per 100 backbone carbons; PEP = alternating poly(ethylene-propylene); a-PP = atactic polypropylene with Bernoullian chain statistics; a-PEE = atactic poly(ethylene ethylene) with Bernoullian chain statistics. LS = solution light scattering. SANS = small-angle neutron scattering of melt. RIS = rotational isomeric state (theory).

new rotational isomeric states but modifies the potential energy of those already existing in the melt. Thus, the locations of the various minima remain unchanged while their magnitudes can be altered by the prevailing solvent–polymer interaction for any particular state.

This scenario requires that the probabilities of occurrence of the most stable conformations—those isomeric states required by the short-range intramolecular interactions—can be modified by solvent interaction. Under theta conditions such short-range interactions can be neglected *if solvent discrimination between conformations is absent.* This criterion is apparently met for polyethylene and alternating-poly(ethylene-propylene) copolymer (Table 4). Conversely, the difference between the solution and melt behavior for *a*-PP and *a*-PEE seemingly demonstrates that their local conformational characteristics are considerably influenced by theta-solvent (*n*-alkyl alcohols) interactions, interactions that lead to enhanced *gauche* populations with increasing temperature (and hence a negative temperature coefficient for chain dimensions). It appears that the configuration of the skeletal bonds will lead in some cases to solvent discrimination regarding the attached side groups with *gauche* sequences being more susceptible to such interactions than their *trans* counterparts. Given the presence of such favorable interactions, the net effect is a decrease in the *trans/gauche* ratio with increasing temperature.

Similar behavior is evidenced for *a*-PP in either 2,4-dimethyl heptane or 2,2,4-trimethyl-pentane, which gave a $d \ln\langle R^2\rangle_0/dT$ of -2.4×10^{-3} deg^{-1} over the temperature range of 10–70 °C. The use of those alkanes was based on their assumed capacity to act as athermal solvents, an assumption that the SANS work has shown to be incorrect. Interestingly, like the *n*-alkyl alcohol theta solvents, these branched alkanes alter the *a*-PP *trans/gauche* ratio with increasing temperature. The solution-induced behavior of *a*-PP and *a*-PEE is strikingly different from that observed in the melt, where the former is found to exhibit a virtually invariant population of conformer types, while the latter shows that increasing temperature favors the *trans* sequences.

Attempts have been made to account for solvent effects within the framework of the RIS model (Brant and Flory, 1965; Bahar *et al.*, 1986; Yoon *et al.*, 1975). This involved the adoption of a truncation procedure involving an adjustable parameter, *s*, which Brant and Flory (1965) took to be the sum of the van der Waals radii of the interacting atoms. Variations of this approach have subsequently led to a procedure (Bahar *et al.*, 1986) that does not utilize adjustable parameters. However, although *a*-PP and *a*-PEE apparently acquiesce to the influence of their solvent shells, the nature and identity of these local interactions are not generally understood (Janik *et al.*, 1987).

The foregoing SANS studies were facilitated by small heterogeneity indices of the subject polymers. Thus, the measured R_g (z-average–dependent dimension) is equal to the weight-average value. In other words, polydispersity corrections are unnecessary.

GLOSSARY

Chain Transfer: The active site at a chain end that propagates during chain polymerizations transfers to a monomer, solvent, initiator, or impurity molecule in the reaction medium.

Configuration: Distinct polymer molecules, typically referring to a specified arrangement of substituent groups along the backbone. Different configurations are delineated by the tacticity and sequence distribution of the polymers.

Conformations: Distinct three-dimensional shapes of a polymer molecule as accessed by rotating around backbone atoms. A Gaussian conformation is a conformation that obeys random-walk statistics.

Copolymer: A polymer having two or more monomer units within the polymer molecule. The relative amounts and the sequence (random, alternating, block, etc.) of these different monomers need to be specified to define the copolymer fully.

Mikta-Arm Polymer: A three-arm star molecule in which each arm contains a different monomer repeat unit.

Model Materials: Materials for which the parameters of molecular characterization are very well represented by the ensemble average.

Molecular-Weight Distribution: The relative amounts of the different molecular-weight polymers in a single sample.

SANS: Small-angle neutron scattering. Contrast for these experiments is typically provided by using perdeutero monomer or by saturating carbon–carbon double bonds with deuterium or a mixture of deuterium and hydrogen.

Size-Exclusion Chromatography: An experimental tool to evaluate the molecular weight and molecular-weight distribution of polymers as compared to standards (typically polystyrene).

Tacticity: Distinct polymer configurations for polymers with the monomer type CH_2—CXY, where X and Y are substituent groups. See Fig. 3 for an illustration of isotactic, syndiotactic, and atactic polymers.

Terpolymer: A polymer having three monomer units within the polymer molecule. The relative amounts, sequence (random, alternating, block, etc.), and architecture (linear, star, etc.) of these different monomers need to be specified to define fully the copolymer.

Theta Condition: The specific condition of polymer, solvent, and temperature at which the polymer chain behaves according to Gaussian statistics and the polymer–polymer interactions are enthalpically equivalent to polymer–solvent interactions.

Works Cited

Albalak, R., Thomas, E. L. (1994), *J. Polym. Sci., Polym. Phys. Ed.* **32,** 341–350.

Ambler, M. R., Fetters, L. J., Kesten, Y. (1977), *J. Appl. Polym. Sci.* **21,** 2439–2451.

Bahar, I., Baysal, B. M., Erman, B. (1986), *Macromolecules* **19,** 1703–1709.

Boor, J., Jr., (1979), *Ziegler-Natta Catalysts and Polymerizations,* New York: Academic.

Boor, J., Jr., Youngman, E. A. (1966), *J. Polym. Sci., Polym. Chem. Ed.* **4,** 1861–1880.

Brant, D. A., Flory, P. J. (1965), *J. Am. Chem. Soc.* **87,** 2791–2800.

Broske, A. D., Huang, T. L., Allen, R. D., Hoover, J. M., McGrath, J. E. (1987), in: T. E. Hogen-Esch, J. Smid (Eds.), *Recent Advances in Anionic Polymerization,* New York: Elsevier, pp. 363–380.

Butera, R., Fetters, L. J., Huang, J. S., Richter, D., Pyckhout-Hiutzen, W., Zirekl, A., Farago, B., Ewen, B. (1991), *Phys. Rev. Lett.* **66,** 2088–2091.

Choma, C. T., Lear, J. D., Nelson, M. J., Dutton, P. L., Robertson, D. E., DeGrado, W. F. (1994), *J. Am. Chem. Soc.* **116,** 856–865.

de Gennes, P. G. (1990), *Introduction to Polymer Dynamics,* New York: Cambridge Univ. Press.

Doi, M., Edwards, S. F. (1986), *The Theory of Polymer Dynamics,* Oxford: Clarendon.

Ewen, J. A., Jones, R. L., Razavi, A., Ferrara, J. D. (1988), *J. Am. Chem. Soc.* **110,** 6255–6256.

Fetters, L. J., Firer, E. M. (1977), *Polymer* **18,** 306–307.

Fetters, L. J., Thomas, E. L. (1993), in: E. L. Thomas, E. J. Kramer (Eds.), *Structure and Properties of Polymers,* Vol. 12, Weinheim: VCH Verlag, pp. 1–31.

Flory, P. J. (1953), *Principles of Polymer Chemistry,* Ithaca, NY: Cornell Univ. Press.

Folkes, M. J., Keller, A., Scalisi, F. P. (1973), *Colloid Polym. Sci.* **251,** 1–4.

Galvin, M. E. (1991), *Macromolecules* **24,** 6354–6356.

Gaskell, R. E. (1950), *J. Appl. Mech.* **17,** 334–336.

Gatzke, A. J. (1969), *J. Polym. Sci., Polym. Chem. Ed.* **7,** 2281–3392.

Hadjichristidis, N., Iatraou, H., Behal, S. K., Chludzinski, J. J., Disko, M. M., Garner, R. T., Liang, K. S., Lohse, D. J., Milner, S. T. (1993), *Macromolecules* **26,** 5812–5815.

Hirai, H., Tanabe, T., Koinuma, H. (1979), *J. Polym. Sci., Polym. Chem. Ed.* **17,** 843–856.

Iatraou, H., Hadjichristidis, N. (1992), *Macromolecules* **25,** 4649–4651.

Ishihara, N., Kuramoto, M., Uoi, M. (1988), *Macromolecules* **21,** 3356–3360.

Janik, B., Samulski, E. T., Toriumi, H. (1987), *J. Phys. Chem.* **91,** 1842–1850.

Keller, A., Pedemonte, E., Willmouth, F. M. (1970), *Nature* **225,** 538–539.

Kern, W. J., Anderson, J. A., Adams, H. E., Bouton, T. C., Bethea, T. W. (1972), *J. Appl. Polym. Sci.* **16,** 3123–3131.

Krejchi, M. T., Atkins, E. D. T., Waddon, A. J., Fournier, M. J., Mason, T. L., Tirrell, D. A. (1994), *Science* **265,** 1427–1432.

Lifson, S., Oppenheim, I. (1960), *J. Chem. Phys.* **33,** 109–115.

Lovinger, A. J., Lotz, B., Davis, D. D., Padden, F. J. (1993), *Macromolecules* **26,** 3494–3503.

Matsushita, Y., Mori, K., Saguchi, R., Noda, I., Nagasawa, M., Chang, T., Glinka, C. J., Han, C. C. (1990), *Macromolecules* **23,** 4387–4391.

Mays, J. W. (1990), *Polym. Bull.* **23,** 247–250.

Morton, M., Helminiak, T. E., Gadkary, S. D., Bueche, F. (1962), *J. Polym. Sci.* **57,** 471–482.

Odian, G. (1981), *Principles of Polymerization,* 2nd ed., New York: Wiley.

Pennisi, R. E., Fetters, L. J. (1988), *Macromolecules* **21,** 1094–1099.

Quirk, R. P. (1992), in: S. L. Aggarwal, S. Russo (Eds.), *Comprehensive Polymer Science, First Supplement,* New York: Pergamon Press, pp. 83–106.

Quirk, R. P., Lee, B. (1992), *Polym. Int.* **27,** 359–367.

Quirk, R. P., Yin, J. (1992), *J. Polym. Sci., Polym. Chem. Ed.* **30,** 2349–2355.

Quirk, R. P., Yin, J., Fetters, L. J. (1989), *Macromolecules* **22,** 85–90.

Quirk, R. P., Yin, J., Fetters, L. J., Kastrup, R. V. (1992), *Macromolecules* **25,** 2262–2267.

Richter, D., Butera, R., Fetters, L. J., Huang, J. S., Farago, B., Ewen, B. (1992), *Macromolecules* **25,** 6148–6155.

Robertson, D. E., Farid, R. S., Moser, C. C., Urbauer, J. L., Mulholland, S. T., Pidikiti, R., Lea, J. D., Wand, A. J., DeGrado, W. F., Dutton, P. L. (1994), *Nature* **368,** 425–431.

Roovers, J., Zhou, L. L., Toporowski, P. M., van der Zwan, M., Iatrou, H., Hadjichristidis, N. (1993), *Macromolecules* **26,** 4324–4331.

Tirrell, D. A., Fournier, M. J., Mason, T. L. (1991), *MRS Bull.* 23–28.

Winey, K. I., Berba, M. L., Galvin, M. E. (1996), *Macromolecules* **29,** 2868–2877.

Winey, K. I., Galvin, M. E. (1993), in: G. W. Bailey, C. L. Reider (Eds.), *Proceedings of the 51st Annual Meeting of the Microscopy Society of America,* San Francisco: San Francisco Press, Inc., pp. 896–897.

Yoon, D. Y., Sundarajan, P. R., Flory, P. J. (1975), *Macromolecules* **8,** 766–783.

Yoshikawa, E., Fournier, M. J., Mason, T. L., Tirrell, D. A. (1994), *Macromolecules* **27,** 5471–5475.

Further Reading

Boor, J., Jr., (1979), *Ziegler-Natta Catalysts and Polymerizations,* New York: Academic.

Broske, A. D., Huang, T. L., Allen, R. D., Hoover, J. M., McGrath, J. E. (1987), in: T. E. Hogen-Esch, J. Smid (Eds.), *Recent Advances in Anionic Polymerization,* New York: Elsevier, p. 363.

de Gennes, P. G. (1979), *Scaling Concepts in Polymer Physics,* Ithaca, NY: Cornell Univ. Press.

Flory, P. J. (1953), *Principles of Polymer Chemistry,* Ithaca, NY: Cornell Univ. Press.

Mark, J. E., Eisenberg, A., Graessley, W. W., Mandelkern, L., Samulski, E. T., Koenig, J. L., Wignall, G. D. (1993), *Physical Properties of Polymers,* 2nd ed., Washington, DC: American Chemical Society.

Odian, G. (1981), *Principles of Polymerization,* 2nd ed., New York: Wiley.

Young, R. J., Lovell, P. A. (1991), *Introduction to Polymers,* 2nd ed., London: Chapman and Hall.

SYSTEMS TECHNOLOGY, MICROMECHANICAL

Kiyoshi Itao, *University of Tokyo, Tokyo, Japan*

Hiroo Ukita, *Ritsumeikan University, Shiga, Japan*

Noriyoshi Yamauchi, *NTT Laboratories, Tokyo, Japan*

Eduardo S. Vera, *Science and Technology Information Center (ICT), University of Chile, Santiago, Chile*

Introduction 451
1. General Principles and Overview of Microsystems Technologies 451
1.1 Historical Perspective 452
1.2 Categories of Systems Technologies 452
2. Interdisciplinary Approaches to Microsystems Technologies 453
2.1 Microenergy Systems 453
2.2 Microdesign Methods: CAD 454
2.3 Microfabrication Process Technologies 455
3. Microelectronic Mechanical Devices: Micrometer Machines 457
3.1 Microactuators 457
3.1.1 Micromotors 458
3.1.2 Digital Micromirror Devices 459
3.2 Microsensors 461
3.3 Intelligent-Microsystem Technologies 461
4. Nanoelectronic Mechanical Devices: Nanometer Machines 462
4.1 Optical Manipulation 463
4.2 Molecular Self-Assembly 465
4.3 Near-Field Scanning Probes 467
5. Summary and Conclusions 467
Glossary 468
Works Cited 469
Further Reading 471

INTRODUCTION

The main focus of this article is the study of *micromechanical devices,* defined as precision micromachines built by microelectromechanical technologies in the size range between a micrometer and a millimeter. The roles of these types of machines and the underlying science on which they are based are discussed in Sec. 3. Other interdisciplinary aspects specific to micromachine design and fabrication technologies are described earlier, in Sec. 2. Some of the technological challenges that the next frontier of submicron devices face are addressed in Sec. 4. The final section of the article discusses the present-day status and the future prospects of micromechanical devices.

1. GENERAL PRINCIPLES AND OVERVIEW OF MICROSYSTEMS TECHNOLOGIES

The relatively recent development of microelectromechanical devices—or simply microsystems, as they are also referred to—crosses many conventional borders of different disciplines of science and technology. Indeed, microsystems research has become a model of the modern interdisciplinary efforts that have gained such prominence in the past two decades. Silicon fabrication techniques have been the real foundations of microsystems, and many impressive devices have been fabricated by direct adaptation of microelectronic techniques into this field. Many other techniques, such as surface mi-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

cromachining and LIGA processes (see Sec. 2.3) have developed new capabilities to enhance further the versatility of semiconductor technology. It is the richness of different ideas and techniques that makes the microelectromechanical field so dynamic.

1.1 Historical Perspective

The historical development of microelectromechanical technologies can be viewed in three stages. First, each specific technology based on mechanics, electronics, optics, information, and physical sciences developed independently. Machine systems were quite independent of electronics technology at this earlier stage (see MECHANICAL INSTRUMENTATION). However, with the rapid progress of sensors and processors brought by the microelectronics revolution (see MICROELECTRONICS), the relationship changed, resulting in the development of truly electromechanical systems, thus creating the interdisciplinary field of mechatronics. This second stage witnessed the parallel combination of optics and electronics into what is now widely known as optoelectronics (see QUANTUM OPTOELECTRONIC DEVICES). Furthermore, the recent developments in optical technology, especially semiconductor lasers and advanced lithography, are now allowing the beginning of a third stage, where complete technology fusion is bringing these fields to merge fully in the new field of optomechatronics. New physical frameworks such as micromechanics, microdynamics, or even photomicrodynamics appear at this point (Ukita *et al.*, 1993).

These developments can be described by the chronological emergence of new technologies. As shown in Table 1, process automation started in the early 1960s. In the late 1960s mechanical automation started making use of electronic control technology. In the early 1970s the age of mechatronics was in an embryonic stage and grew into a mature discipline in the late 1970s, when the use of integrated circuits (ICs) with large-scale integration capabilities became widespread. In the meantime, the semiconductor laser, which had been developed in 1962, became able to emit light continuously at room temperature in 1979. This resulted in the birth of optomechatronics, which combines mechanics, electronics, and photonics, during the early 1980s. In the late 1980s, with the use of advanced semiconductor-device fabrication technologies, microsystems were actively studied. New kinds of microsensors and microactuators became feasible. In addition, novel probing techniques such as the scanning tunneling microscope (STM) and atomic force microscope (AFM) have greatly advanced our knowledge of microscopic mechanisms to the molecular and the atomic levels. Thus, in the 1990s the fusion of information and energy technologies has started, as a prelude to the first stages of molecular nanomachines.

Table 1. Chronology of microsystems technology developments.

Years	Development of technology and new trends
1960	Process automation in chemical industry and heavy machine industry
1965	Mechanical automation by electrical control
1970	Early stage of mechatronics that makes use of IC and LSI
1975	Mechatronics using microprocessors
1980	Optomechatronics using semiconductor lasers
1985	Micromechatronics using process technology
1990	Age of mechatronics that achieves fusion of information and energy

1.2 Categories of Systems Technologies

A conventional precision machine usually consists of a main body and many components. Therefore, specific design, fabrication, and measurement technologies are necessary to assemble all these pieces together in a systematic way. According to their size dimension, machines are usually classified in three categories: millimeter, micrometer, and nanometer machines. Although there is no obvious abrupt transition between these arbitrarily defined types, each broad category uses different materials and has a distinctive set of systems technologies. The millisystems use metals, polymers, or ceramic materials. Various ultraprecision methods such as discharge and beam processes, as well as bulk etching, are employed to approach the miniaturization limit of these conventional precision machines. The microsystems are based on modified semiconductor fabrication technologies such as wafer bonding and advanced photolithography, which can be applied to silicon, gallium arsenide, and thin-film materials. In the new frontiers of miniaturization, nanosystems usually consist of macromolecules such as proteins or other organic polymers in which matter is manipulated at the molecular and the atomic levels.

2. INTERDISCIPLINARY APPROACHES TO MICROSYSTEMS TECHNOLOGIES

As we have underscored, the design and fabrication of micromechanical systems requires wide exchange of information and joint efforts from researchers in many divergent fields of science and engineering. In this section we emphasize those interdisciplinary aspects of microsystems technologies that are of special importance for the successful development of microelectronic mechanical devices.

2.1 Microenergy Systems

The relentless pace of miniaturization is forcing drastic changes in the way electronic systems are built. As more compact and larger numbers of components with high-level functions are built in ever-shrinking dimensions, interconnections, packaging, and assembly become increasingly complex and important aspects of the fabrication process. Among other effects, this is motivating designers to rethink drastically how to power such systems (Govindarajan and Forrest, 1991). Local autonomy from a central power supply becomes a highly desirable feature to avoid excess power dissipation, high-frequency crosstalk between adjacent elements, and the general reliability problems associated with ultrahigh-density systems. Having distributed microenergy capabilities also becomes more feasible as the energy level required to perform complex electronic or electromechanical functions on such small scales decreases accordingly. To achieve this goal, three main aspects will require much attention in the years to come, namely, the power generation, the energy storage, and the power supply of microenergy systems.

There are several approaches to the power generation in microsystems, depending on the nature of the energy source employed. Electrical power can be obtained from optical, mechanical, or thermal sources using light, pressure, or infrared sensors, respectively. For instance, photovoltaic, microwave, and infrared devices can generate electrical currents from optical, microwave, and infrared radiation absorption existing within or externally supplied to the environment. The electrical energy thus generated can directly power different microelectromechanical functions or be stored by a microstorage battery system for later use. Mechanical pressure sensors, accelerometers (Allen *et al.*, 1989), and flow-measurement microdevices (Cho *et al.*, 1989) are rapidly emerging with enhanced sensitivity and high-precision capabilities. Vice versa, electrostatic micromotors that convert electrical signals into mechanical motion have been successfully fabricated from thin-film technologies compatible with silicon processing (Fan *et al.*, 1988), as described in Sec. 3.

The development of power-integrated technologies has brought the ability to provide the interface between digital control logic and the power load (Baliga, 1990). Thus, power ICs monolithically integrate power control with sensor and logic-circuit functions. This interface allows the intelligent use of power in microelectronic systems and micromachines. Smart-power technology combines complementary metal-oxide semiconductor (CMOS) logic, high-performance bipolar analog circuits, and high-voltage power devices all on the same chip. Indeed, smart-power chips can perform load monitoring, diagnostic functions, self-protection, and information feedback to a microprocessor, in addition to handling large amounts of power to the load. This allows highly efficient energy management in a variety of microelectronic systems by connecting a distributed network of smart power supplies with a multiplexed bus. One of the most critical elements in such a system is the power switch, which is usually a metal-oxide-semiconductor field-effect transistor (MOSFET) or MOS/bipolar device depending on the current-handling levels required. The main advantages of power MOSFETs over bipolar power transistors are the paralleling capability, high switching speed, high input impedance, and superior safe operating area

(Baliga, 1991). For the realization of ultracompact power supplies, high-frequency switching is essential in order to reduce the size of other electronic and magnetic components to provide full integration, as shown by Mino *et al.* (1992) (see MICROMAGNETICS). Thus, low capacitance and ultralow "on" resistance are critical requirements of advanced power MOSFET design (Vera *et al.*, 1992).

One area that has yet to produce significant progress is that of microenergy storage. Indeed, in terms of the energy that batteries store for a given size and weight, not much improvement has been achieved. For example, today's nickel-cadmium batteries are only about 40% more effective than the first ones about 100 years ago. Lithium-based batteries provide about two or three times the energy of standard nickel-cadmium or lead-acid cells (Yotsumoto, 1992), as shown in Fig. 1. However, the lithium metal is highly reactive, and some associated safety concerns that have prevented their extensive use remain among the most important problems to be solved. Thin-film batteries are of great interest for their potential use in monolithically integrated electromechanical systems. A rechargeable microbattery that could power a single chip, or even single electronic components within, would have a significant impact on the development of advanced electromechanical devices. Preliminary steps in that direction have been taken by Li *et al.* (1992), who have reported a nanometer-scale electrochemical cell built using an STM to deposit silver and copper pillars electrochemically on a graphite substrate.

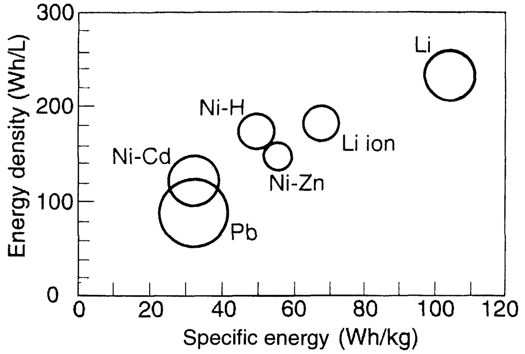

FIG. 1. Energy density in battery systems (Yotsumoto, 1992).

2.2 Microdesign Methods: CAD

Over the last decade, we have observed dramatic progress in the development of computer-aided design (CAD) tools for ICs. Indeed, the ever-increasing computational power has allowed the development of extensive numerical codes that perform very sophisticated calculations at relatively low cost. These advanced CAD capabilities have had a tremendous impact on the development of very large-scale integrated (VLSI) and ultra-large-scale integrated (ULSI) technologies because they have allowed the full simulation of a wide range of electronic systems at the process, device, and circuit levels. The virtual fabrication and testing of complex electronic systems in a computer environment makes it possible to bypass the real implementation of all but the optimal electronic structures, with significant savings of resources and time. However, most unfortunately, many of these tools are of limited applicability to the design of microelectromechanical systems.

Although microelectromechanical systems are usually fabricated using refinements of the planar photolithographic techniques used in the IC industry, they require very different CAD tools than those used by the IC designers because of their intrinsic three-dimensional (3D) nature. The most important requirement that such a microdesign CAD system should have is to be able to analyze realistic 3D structures whose geometry is directly provided or derived from a fabrication process simulation that uses a process flow chart and mask-level description. Since 1987, several research groups have reported some progress in this area. Koppelman (1989) has created a program that gives a 3D solid-geometric model from the IC process description and mask data. Crary and Zhang (1990) and Zhang *et al.* (1990) have developed an overall user interface that can access modules simulating the mechanical behavior of specific structures. Senturia (1987) has addressed both the system and the specific tool-development levels.

However, all these approaches must satisfy some rather stringent requirements. For instance, they must work from the actual design. Usually, IC CAD does not provide tools for automatic generation of 3D models that are needed for subsequent mechanical and electrical analysis. Hence, the structure simulator link between the IC layout and fabrication process world and the three-dimensional model is an essential feature of microelectromechanical CAD systems. Simulation typically involves finding solutions for quantities not only in the interior but also in the exterior of the structures. It is possible to use finite-element modeling methods to analyze both interior and exterior problems. However, this can be computationally very demanding. Consequently, sometimes accelerated boundary-element procedures that require only surface and interior meshes are employed (Conti *et al.*, 1991). Despite recent progress, generation of a satisfactory 3D volume mesh shape from a boundary representation is still an open question.

The long-term aim should be to provide a structure simulator with great geometric flexibility as well as effective links to general 3D analysis tools within the same computational environment. Maximum portability and compatibility with existing IC CAD tools is of utmost importance in any microelectromechanical CAD system. Thus, a structure simulator needs to interface with independent mechanical and electrical modeling software packages. Many specific implementation decisions must be made in the early stages of a system development in order to guarantee an architecture that permits the substitution of different software models. To achieve this degree of flexibility, an object-oriented architecture is most convenient because it can use the same data structure regardless of the specific data type as long as the type can be recognized and evaluated by the user interface (Shulman *et al.*, 1991). Following this approach a structure simulator can merge the mask layout and process information with the construction of a 3D solid model whose mechanical as well as material and electromagnetic properties can then be simulated by independent modules well integrated into a complete microelectromechanical CAD system that can perform comprehensive design, analysis, and test functions.

2.3 Microfabrication Process Technologies

With the recent surge of available microactuators in addition to the steady improvement of microsensors and interface circuitry, many of the elements for the development of complete integrated microsystems appear to be in place. Accordingly, there have been many proposals for the creation of such systems with a wide range of applications (Minami *et al.*, 1993). These range from microrobots for medical and several other applications to sophisticated control or positioning systems of great accuracy. If the proponents are successful, the coming years should witness the practical realization of multiple structures merging mechanical, thermal, electromagnetic, and optical components on a common substrate. Advanced assembly, interconnections, and packaging techniques will be particularly crucial to the cost-effective implementation of such schemes.

Many research groups around the world are intensively developing a variety of microfabrication technologies for a diverse number of applications. Most of them are inspired by or derived from the vast pool of know-how accumulated during the past decades of explosive growth in the microelectronic industry. However, translating IC-based technologies for micromachine fabrication sometimes is neither possible nor advisable. Thus, creatively adapting these technologies or attempting completely new approaches is of great importance for the development of this emerging field.

Micromachining was born in the 1960s when precise silicon etching techniques were first developed in the early stages of ICs (Lepselter, 1966). The initial work was based on isotropic etchants, which were straightforward to use but difficult to control. Later, anisotropic etchants that make highly directional and reproducible etching possible allowed significant progress in selective control. By and large, they are the basis of micromachining today. With the use of photolithography, an etch mask is defined on the back of the wafer, aligning it with patterns on the front surface. The wafer bulk is then selectively etched from the back. The existence of impurity-based etch stops in silicon has allowed micromachining to become a high-yield production process in the past decade (Petersen, 1982). A well-known example of this technique is the wide use of heavily boron-doped (p^+) silicon with ethylene diamine pyrocatechol (EDP) as anisotropic chemical etchant for the fabrication of diaphragm or beam sensors. Indeed, bulk micromachining has been so far the widest used technique for most integrated sensors produced by industry (Yoon and Wise, 1990). However, this method requires very long times to etch the silicon substrate from the back, usually several hours. An approach to overcome this serious limitation that has recently gained much acceptance is surface micromachining (Howe, 1988).

Surface micromachining utilizes a number of thin-film technologies. This technique has been used to fabricate various microstructures, including resonant microsensors and electrostatically driven micromotors (Fan *et al.*, 1988). In this method, a sacrificial layer (usually silicon dioxide or silicon nitride) is initially deposited and patterned on a silicon wafer. Next, the film for the desired microstructure is deposited and patterned as well. The sacrificial layer is then etched, leaving the undercut microstructure freely suspended but connected to the substrate through those areas beyond the patterned sacrificial layer. Several research groups have demonstrated the significant potential of this technology. Howe and Muller (Howe and Muller, 1986) fabricated a resonant vapor sensor with a free-standing polysilicon bridge. Rotary micromotors have been developed by Mehregany *et al.* (1990), Bart *et al.* (1990), and Flynn *et al.* (1992), among others. Resonant microsensors can be easily fabricated with this technique, and a wide range of accelerometers, pressure sensors, and microactuators are currently under development for automotive and consumer electronics. The successful development of surface-micromachined micromechanical structures has been made possible by the progress in understanding the mechanical properties of thin films that are crucial for the fabrication of devices with free-moving parts (Suzuki *et al.*, 1992). Most structures use polysilicon as the mechanical material, but other films, including polymers and metals, have also been employed. The availability of reactive ion-etching processes to define patterned features precisely in the deposited thin films has also been an important factor in the progress achieved.

In spite of its many desirable features, surface micromachining is limited to the fabrication of structures with vertical features of the order of few microns because of the intrinsic limitations of thin-film deposition methods. When micromechanical devices require structures of tens of microns or more, alternative technologies such as wafer bonding or dissolved-wafer processes must be employed. For instance, microstructures as thick as 20 μm have been fabricated by Suzuki (1990) by using a bulk dissolved-wafer process in conjunction with reactive ion etching. Microstructures with much higher thickness-to-width ratio can be fabricated with x-ray lithography and electroforming (LIGA) processes (Munchmeyer and Langen, 1992).

Originally developed in Germany, where the acronym "LIGA" was coined (from *litographische galvanoformung*), this method uses x-ray synchrotron radiation lithography to produce extremely precise polymer molds of three-dimensional microstructures on a supportive substrate to generate them subsequently by electroplating metal films. The main attraction of this technology is the large thickness-to-width tolerance (up to 300:1) that results from the short photon wavelength and the thickness of photosensitive polymers. No other technique allows the fabrication of truly 3D microstructures that have the ability to generate large forces and torques.

All these microstructures should be efficiently integrated at the system level. A number of important efforts are presently under way to establish interface standards for next-generation smart microelectromechanical devices that can operate with a VLSI interface chip to implement local intelligence, self-testing, and autocalibration capabilities, as well as bus compatibility at the system level.

Although significant progress has been made in some classes of microstructures (most notably, microsensors), an immense number of challenging problems remain to be answered before microsystems technology can really take off. Greater understanding, characterization, and control of the mechanical properties of thin films, such as stress and strain, is badly needed. The removal of stress by new annealing techniques will have to be achieved in order to design, simulate, and fabricate new and more complex microsystems that can perform as predicted. Improved connectivity, packaging, and assembly techniques will need to be developed. All these issues must be the focus of great attention ahead. Solving these problems will demand years of interdisciplinary effort from leaders in engineering, physics, chemistry, biology, and business in universities as well as industrial laboratories.

3. MICROELECTRONIC MECHANICAL DEVICES: MICROMETER MACHINES

In this section, we discuss machines in the size range between 1 and 10^{-3} mm. To describe these machines accurately, we call them microelectronic mechanical devices. Here, we focus on two types of microelectronic mechanical devices that are playing an increasingly important role in the development of this field: *microactuators,* which convert electrical signals into mechanical motion, and *micromechanical sensors,* which convert mechanical motion into electrical signals. We discuss the principles, the most common structures, and the fabrication processes for these typical microelectronic mechanical devices, illustrating state-of-the-art technologies with the first version of commercial products.

Three important factors have motivated a high level of interest in microelectronical mechanical devices. They are summarized as follows:

1. High efficiency: If we consider a machine of feature size L, we find that the mass of the movable part is proportional to L^3. Here we assume that the velocity of the movable part is proportional to L. Although there is no direct link between the feature size and the velocity, it gives a reasonable relation in the size dependence of velocities, which is effective in the dimension analysis. Under this assumption, the momentum is proportional to L^4, and the kinetic energy $\frac{1}{2}mv^2$ is proportional to L^5. This means that the momentum and the kinetic energy can be reduced by the order of 10^{-4} and 10^{-5}, respectively, by reducing the size. In other words, conversion from electricity to motion in actuators can be achieved with extremely small energy compared with larger machines. By the same argument, the efficiency of motion to electricity conversion in sensors can also be extremely high.
2. Low cost and high reliability: In terms of practical feasibility, massive production is possible with batch fabrication processes similar to those employed in the industrial production of VLSI chips. This naturally leads to significant cost reduction. Indeed, fabrication processes with precise quality-control methods translate into high-quality and cost-effective manufacturing alternatives with high reliability.
3. Integration with electronic circuits: The basic fabrication process for the microelectronic mechanical devices is fully compatible with silicon integrated-circuit technology. This permits the monolithic integration of electromechanical devices and electronic microcircuits in the same piece of silicon, enabling a high degree of device functionality, such as self-test and self-calibration, which could not be achieved otherwise. It should be noted that in the two examples described in this section, namely a digital micromirror device and a microaccelerometer device, the micromechanical elements are monolithically integrated with silicon CMOS circuits that electronically control the function of the micromechanical devices. This "smart actuator/sensor" approach is essential to achieve a systems design with high functionality and massive parallelism of the microelectronical mechanical devices.

The above factors have significantly contributed to accelerating the pace of the research and development effort in this field. Commercial production of microelectronic mechanical devices has entered its initial stages, and a variety of novel devices have started to appear in the market (Cho *et al.*, 1989).

3.1 Microactuators

Many research groups have been trying to exploit various types of forces in the design of microactuators: electrostatic (Fan *et al.*, 1988; Mehregany *et al.*, 1990; Kikuya *et al.*, 1993), magnetic (Yanagisawa *et al.*, 1991), piezoelectric, and the force derived from shape-memory alloy, to name a few. All these forces usually originate from electricity. Thus, microelectronic mechanical devices effectively cover most microactuators.

Among the various types of microactuators using the above forces, those employing electrostatic forces have been the devices most extensively studied. They provide good examples of successful electricity-to-motion conversion. In the following subsections, we describe two different types of microactuators that use the electrostatic force: micromotors and digital micromirror devices.

3.1.1 Micromotors The demonstration of the side-drive micromotor developed by Fan *et al.* (1988) at the University of California, Berkeley, caused public sensation, triggering substantial research activity in microelectronic mechanical devices. Figure 2 shows the top view and the cross section of the stepping motor they developed. The biasing and the phasing schemes are shown here. The motor consists of a rotor and a stator, both made of 1.0- to 1.5-μm-thick polycrystalline silicon thin film. The rotor rotates around a hub, which is also made of polycrystalline silicon thin film. The diameter of the rotor is 120 μm and the gap between the rotor and stator is 2 μm. When pulse voltages of around 100 V are applied to the stator poles from the peripheral circuits with different phases ϕ_1, ϕ_2, ϕ_3, a torque of the order of piconewton meters arises due to the electrostatic force between the rotor and the stator, which are offset, as shown in Fig. 2. A synchronous motor was also developed by the same researchers, achieving an experimental speed of 500 rpm.

Shortly after the Berkeley micromotor demonstration, Mehregany *et al.* (1990) at the Massachusetts Institute of Technology fabricated and demonstrated a side-drive micromotor with a 100-μm diameter that rotates at the much higher speed of 15 000 rpm with a continuous operation lifetime of about a week. By carefully analyzing the motion of the micromotor, these researchers developed a structure that significantly reduced the friction between the rotor and the stator, which were fabricated with heavily phosphorus-doped polysilicon thin films. They also demonstrated that 7000 micromotors can be built on a 4-in. (10-cm) silicon wafer in a batch process compatible with CMOS VLSI fabrication. Figure 3 shows the different fabrication steps for this micromotor (Mehregany *et al.*, 1992). The fabrication process employs the sacrificial oxide technique and uses polycrystalline-silicon thin-film technology. First, the initial sacrificial layer is deposited and patterned. Then, the bushing mold is etched, as shown in Fig. 3(a). Next, a 2.5-μm-thick phosphorus-doped polysilicon film is deposited and patterned, as illustrated by Fig. 3(b). The second sacrificial layer is

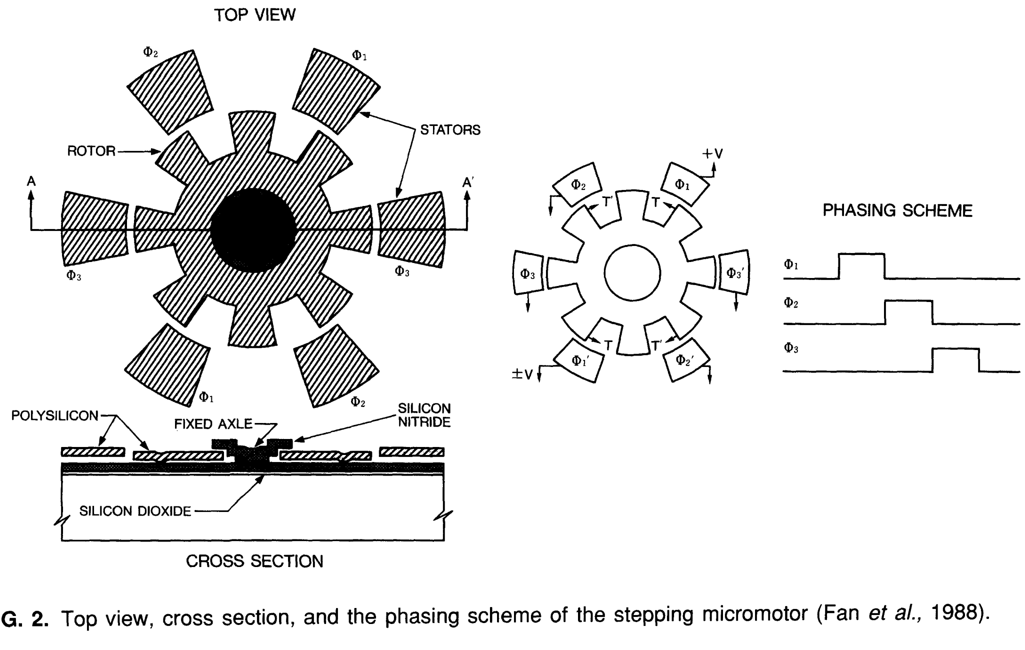

FIG. 2. Top view, cross section, and the phasing scheme of the stepping micromotor (Fan *et al.*, 1988).

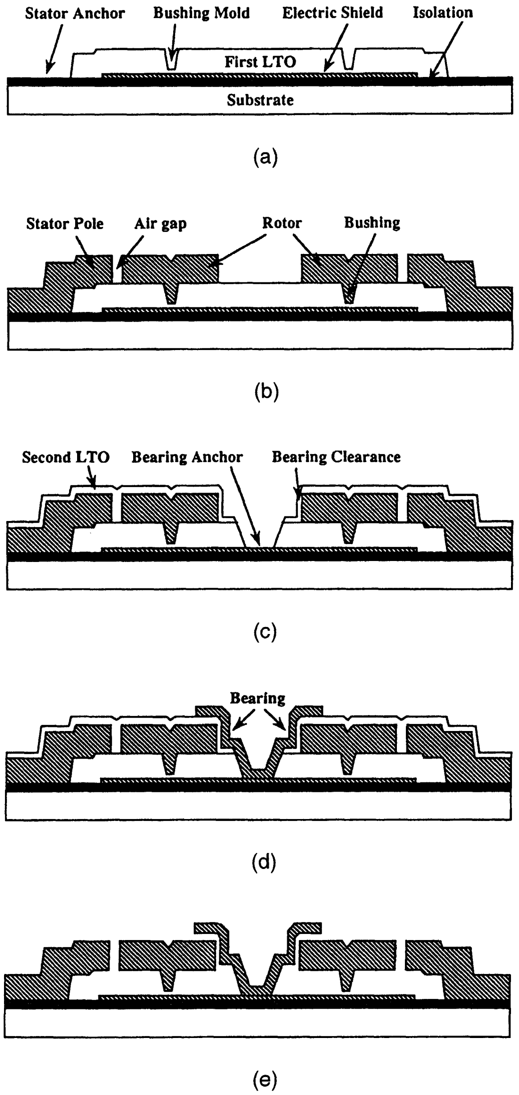
Stator Anchor
Bushing Mold
Electric Shield
Isolation
First LTO
Substrate
(a)
Stator Pole
Air gap
Rotor
Bushing
(b)
Second LTO
Bearing Anchor
Bearing Clearance
(c)
Bearing
(d)
(e)

FIG. 3. Side-drive micromotor fabrication process demonstrated by schematic cross-sectional figures. (a) After stator definition. (b) After rotor-stator definition. (c) After bearing anchor definition. (d) After bearing definition. (e) After release. (Mehregany *et al.*, 1992.)

then deposited and a 1-μm-thick polysilicon thin film is deposited, and patterned, as shown in Fig. 3(c). The bearing is patterned, according to Fig. 3(d). Finally, the first and second sacrificial layers are removed by HF etching to complete the structure shown in Fig. 3(e). Figure 4 shows a scanning electron micrograph of the micromotor.

Micromotors have not found practical applications yet. The friction between the rotor and the axle still needs to be reduced significantly in order to guarantee a practical lifetime. However, the broad impact of micromotor laboratory demonstrations has activated wide interest in the world of micromachines, accelerating significantly the investigations of electronic mechanical devices.

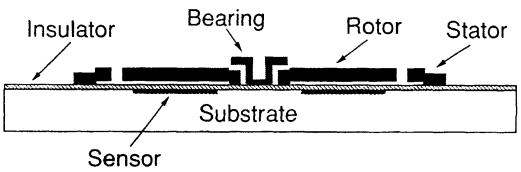
Insulator
Bearing
Rotor
Stator
Substrate
Sensor

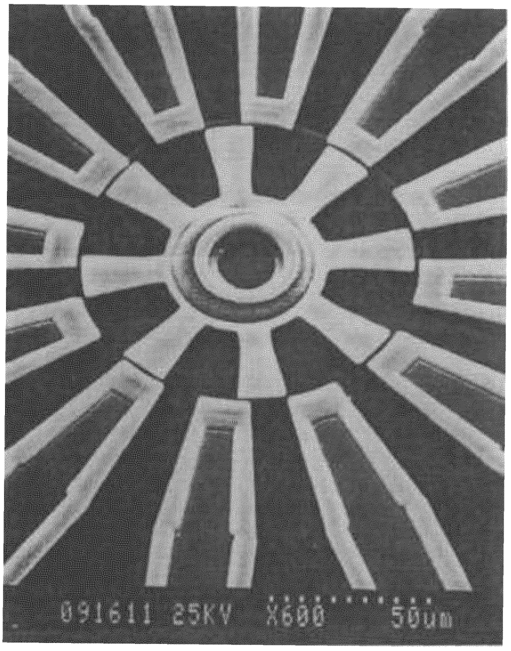
091611 25KV X600 50um

FIG. 4. Scanning electron micrograph of the micromotor (Mehregany *et al.*, 1992).

3.1.2 Digital Micromirror Devices The digital micromirror device (DMD) is one of the few microactuators that has found a clear practical application: the use in large-screen projection displays (Hornbeck, 1993; Younce, 1993).

The DMD is a micromechanical, reflective spatial light modulator consisting of an array of rotatable aluminum mirrors, monolithically fabricated over a conventional CMOS SRAM address circuit. It was invented in 1987 by Hornbeck (1989) at Texas Instruments. Figure 5 shows the schematic cross

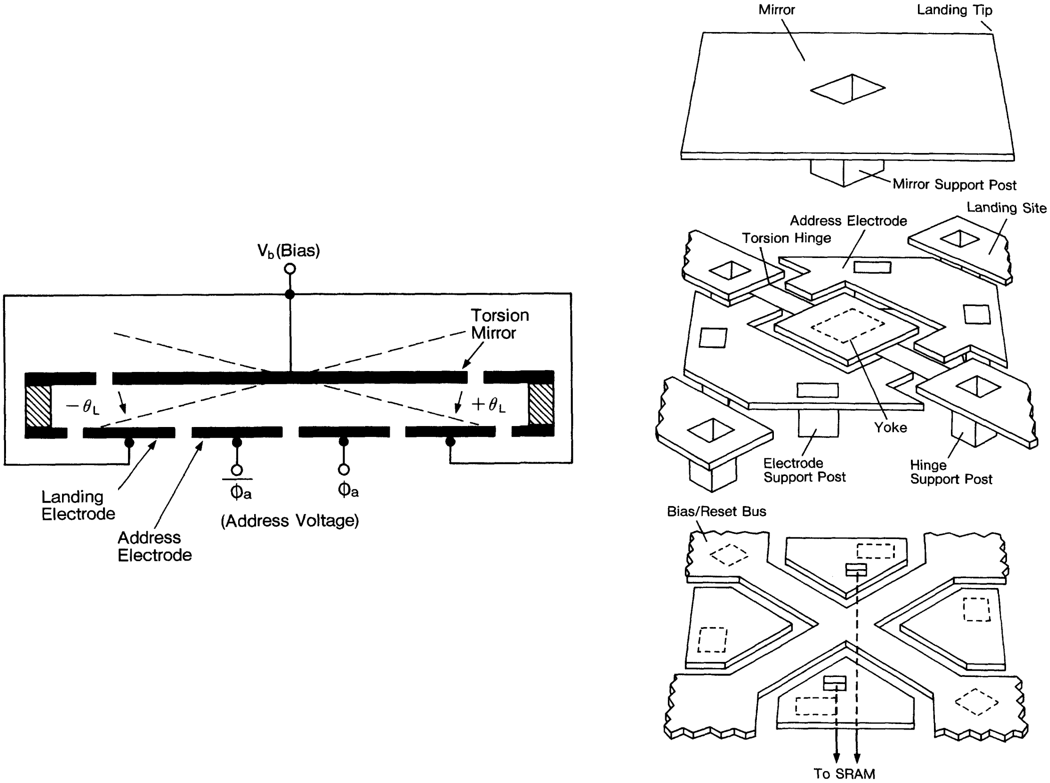
V_b(Bias)
Torsion Mirror
−θ_L
+θ_L
Landing Electrode
Address Electrode
Φa
Φa
(Address Voltage)
Mirror
Landing Tip
Mirror Support Post
Landing Site
Address Electrode
Torsion Hinge
Yoke
Electrode Support Post
Hinge Support Post
Bias/Reset Bus
To SRAM

FIG. 5. Schematic cross section and structure of the hidden-hinge digital micromirror device (DMD) pixel (Hornbeck, 1993).

section and the latest structure of the DMD (Hornbeck, 1993). The aluminum mirror is suspended over an air gap by two thin, mechanically compliant torsion hinges supported by posts that are electrically connected to an underlying bias/reset bus. The address electrodes and the landing electrodes are on the same plane as the hinge. There are two air gaps: one between the mirror and the underlying hinges and a second one between the address electrodes and the underlying third metal level of the CMOS SRAM structure. The structure is fabricated by the sacrificial layer technique in which organic material is etched away. Because the hinges are hidden from the surface, they cannot diffract light. Therefore, they do not degrade the contrast ratio. Each aluminum mirror rotates independently +10° or −10° by an electrostatic force between the address electrode and the aluminum mirror. Although the DMD can be operated in an analog mode, the mirrors are biased in such a way that only the digital landing states of +10° or −10° are possible. The response speed is 10 μs, which is fast enough to use the device digitally in the full-color projection display system. By addressing the SRAM, any pattern of +10° or −10° rotated pixels can be arranged in the two-dimensional micromirror array.

A full-color projection display using one DMD is possible using an optical system that includes a light source, a color filter wheel, condenser lenses, and a zoom projection lens, as shown in Fig. 6 (Hornbeck, 1993). The pixel structure has been improved to achieve high contrast ratio and high optical efficiency. A 768 × 576 hidden-hinge DMD has been tested, and system contrast rations up to 100:1 have been measured. Very high optical efficiency (>60%) is obtained on account of high reflectivity of the aluminum mirror (>90%) and the high ratio of the mirror area (>75%) in the pixel. In durability tests at Texas Instruments, the DMD has

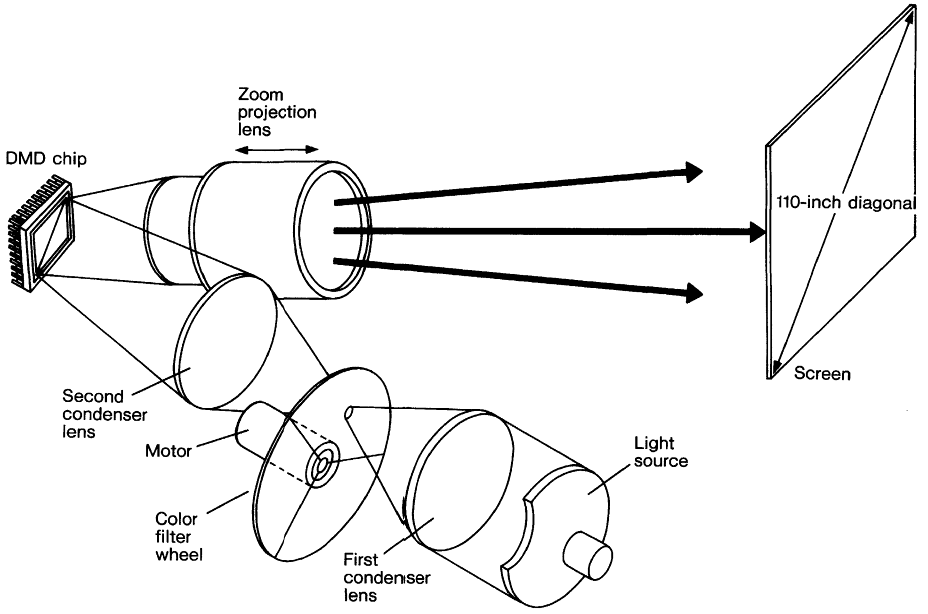

FIG. 6. Full-color projection display using the DMD (Hornbeck, 1993).

shown no fatigue failure after several hundred billion cycles (the equivalent of about four years of continuous operation), proving the ruggedness of this electronic mechanical device.

3.2 Microsensors

Micromechanical sensors convert mechanical motion to electrical signals (see MICROSENSORS). Among a variety of smart sensors presently being studied, the integrated microaccelerometer has found recent applications in the automobile industry, where it is now employed as a key component of airbag safety systems.

Figure 7 summarizes the schematic structure and the function of an integrated microaccelerometer developed at Analog Devices (Goodenough, 1991; Core *et al.,* 1993). The device consists of a micromechanical sensor monolithically integrated with a CMOS electronic circuit on the same silicon substrate. The sensor has fixed capacitor plates and a moving capacitor plate, both made of doped polycrystalline silicon thin film. The moving capacitor plate is formed by the sacrificial etching technique. Special efforts were made to reduce the residual stress in the polycrystalline silicon thin films (Core *et al.,* 1993). When the sensing part is accelerated, the moving part moves and the capacitance between the fixed plates and the moving plate changes. This change in capacitance is detected and electrically processed by the CMOS circuits. A linear output is obtained in the (0–50)g range.

3.3 Intelligent-Microsystem Technologies

As we have seen in the previous examples, the first generation of commercial products using microelectronic mechanical devices is beginning to emerge and state-of-the-art technologies are being developed to mass produce several other intelligent microdevices in the near future. Novel applications may be found in optical-communication hardware systems, in which precise alignment with accuracies of the order of 1 μm is required (see OPTICAL INTERCONNECTIONS). Kikuya *et al.* (1993) have successfully demonstrated an optical-fiber alignment mechanism using electrostatic forces to align optical fibers to optical devices such as lasers or photodiodes. This technology could be applied to the alignment of one- or two-dimensional optical-fiber arrays in various mas-

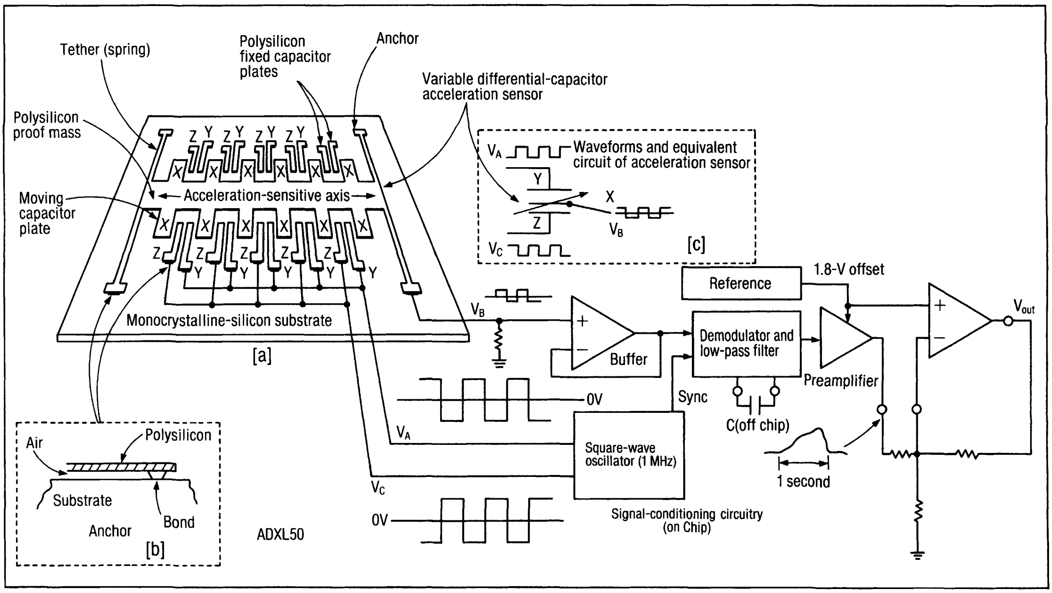
Tether (spring)
Polysilicon fixed capacitor plates
Anchor
Variable differential-capacitor acceleration sensor
Polysilicon proof mass
Moving capacitor plate
Acceleration-sensitive axis
Z Y
X
Monocrystalline-silicon substrate
[a]
Air
Polysilicon
Substrate
Anchor
Bond
[b]
ADXL50
VA
Waveforms and equivalent circuit of acceleration sensor
Y
X
Z
VB
VC
[c]
Reference
1.8-V offset
VB
Buffer
Demodulator and low-pass filter
Preamplifier
Vout
0V
Sync
C(off chip)
VA
Square-wave oscillator (1 MHz)
1 second
VC
0V
Signal-conditioning circuitry (on Chip)

FIG. 7. Integrated micromechanical accelerometer (Core *et al.*, 1993).

sively parallel optical interconnection schemes (Shirai *et al.*, 1993). Driven by the strong market demand for advanced optical-communication systems, further applications of smart microelectronic mechanical devices may increase rapidly.

Traditionally, most microelectronic technologies such as thin-film deposition, patterning, and photolithography have been used in the fabrication of micromechanical structures, either directly or with minor modifications. As VLSI technologies evolve to smaller and smaller ULSI size dimensions, micromechanical technologies will continue to catch up with these ever-higher levels of integration. It is thought that micromechanical technologies will fully merge with VLSI technologies, allowing new sophisticated functionalities that result only from the technology fusion. From what we are presently seeing, it is reasonable to expect that various technologies from different fields will merge into truly intelligent-microsystems technologies, creating totally new capabilities for future sensors and actuators. Furthermore, these advanced microelectronic mechanical devices will give birth to intelligent micromachines, which should eventually be an integral part of the next-generation communications networks that will act as interface between human beings and their environment.

4. NANOELECTRONIC MECHANICAL DEVICES: NANOMETER MACHINES

We define "nanomachines" as ultrasmall structures whose feature size is smaller than one micron (1 μm = 10^{-6} m) but greater than one nanometer (1 nm = 10^{-9} m), a range that corresponds to the so-called "mesoscopic" realm. Nanoscale fabrication is expected to play a central role in the next technological frontier after the enormously successful development of microelectronics. Thus, nanotechnologies are likely to affect many research areas, such as precision engineering, automated manufacturing, environmental sciences, and medicine, in a very profound way. Although we have already seen the experimental fabrication of nanoscale devices in many laboratories, nanotechnologies are still only in an embryonic stage. Therefore, it is quite unclear what form nanomachines will take as nanotechnologies mature in the years to come, possibly well into the twenty-first century (Ball and Garwin, 1992). Figure 8 compares several possible fabrication processes that are presently being ex-

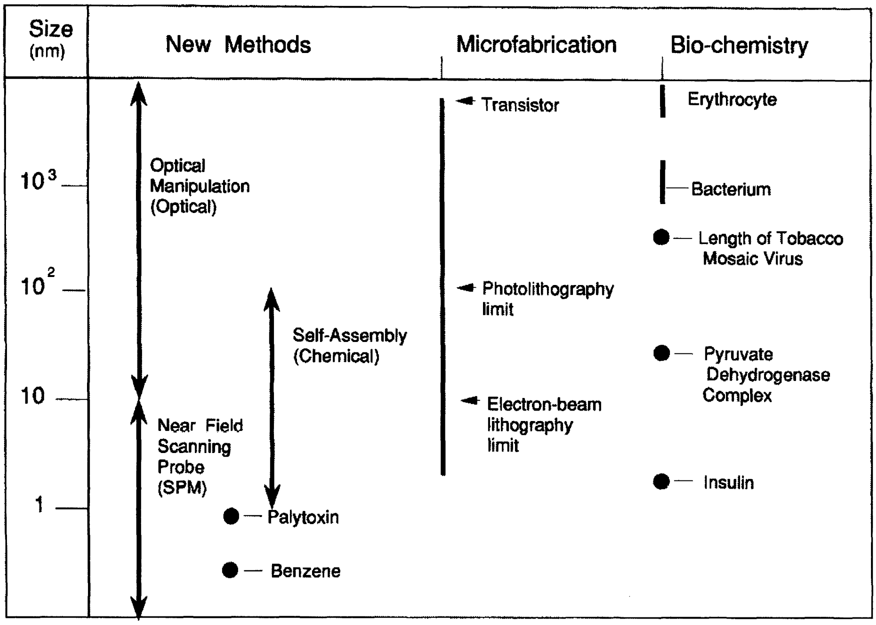
Size (nm)
New Methods
Microfabrication
Bio-chemistry
10³
10²
10
1
Optical Manipulation (Optical)
Self-Assembly (Chemical)
Near Field Scanning Probe (SPM)
Palytoxin
Benzene
Transistor
Photolithography limit
Electron-beam lithography limit
Erythrocyte
Bacterium
Length of Tobacco Mosaic Virus
Pyruvate Dehydrogenase Complex
Insulin

FIG. 8. Fabrication methods for nanostructures in relation to the sizes of various well-known structures.

plored as alternative approaches to assemble nanostructures: conventional microelectronics, biochemistry, and other novel methods.

Recent developments in microelectronics have already allowed the fabrication of mesoscopic devices. Ultrasmall structures, such as quantum wells, quantum wires, and quantum dots, in which electrons are confined to potential wells in two, one, or zero dimensions, are leading to new electrical and optical properties (see ELECTRON STATES IN ZERO-, ONE-, AND TWO-DIMENSIONAL STRUCTURES; HETEROSTRUCTURES AND SUPERLATTICES, SEMICONDUCTOR). By using advanced epitaxial deposition techniques, one can readily make artificially quantized structures. By means of atomic steps or terraces on highly pure single-crystal surfaces, it is possible to make very small (diameter < 100 Å) wires (see MOLECULAR-BEAM EPITAXY). Thus, electrons can be confined and controlled or released in this dimension. At present, however, this atomic-scale manipulation is usually restricted to one dimension. On the device surface plane, the minimum feature size is limited by the lithographic resolution. Thus, microelectronic fabrication methods have a resolution limit of about 100 nm due to the photon wavelength used in x-ray lithographic patterning. Using electron-beam lithography, the patterning resolution can be extended down to 20 nm, a limit enforced by the physical properties of the polymers employed as resist masks in this process. Therefore, further feature size reduction and greater design flexibility in the fabrication of nanoscale structures and device systems will require a radical departure from the conventional microelectronic technologies so successfully being used today. In this section, we describe three promising new approaches that might lead to the development of such nanotechnologies: optical manipulation of small particles, molecular self-assembly or chemical synthesis, and near-field scanning probing.

4.1 Optical Manipulation

Optical manipulation uses photon pressure to trap and assemble small-size particles, and ultraviolet light to link them together by adhesion. Optical pressure arises from changes in photon momentum. When we assume light power P of 1 mW (a typical laser light power), the magnitude P/c, where c is the speed of light, is about 3×10^{-12} N. This force is nearly equal to the force of gravity on a sphere whose radius is 4.3 μm and its density is 1 g/cm^3. Therefore, we can use a laser beam to levitate small particles. For particles whose diameter is larger than the laser wavelength, the principle of optical trapping is explained as follows (Ashkin, 1980). Let a sphere with refractive index higher than that of the surrounding medium be placed off-axis at position Q in a Gaussian beam, as shown in Fig. 9(a). Rays a and b refract through the sphere, giving rise to forces F_a and F_b in the direction of the momentum change. Since there is more light at a than at b, $F_a > F_b$. This results in a (transverse) gradient force F_{tr} and an axial force F_{ax}. The gradient force pulls the sphere into the high-intensity region and traps it at the laser focus. Optical trapping by the axial force is also possible in air with a single, vertically directed laser beam, as shown in Fig. 9(b). For particles whose diameter is smaller than the laser wavelength, the trapping force is calculated by electromagnetic theory (Wright, 1993; in this calculation, the microsphere was assumed to be polystyrene in water). The maximum axial force decreases proportional to r^3 for a radius less than 0.1 μm.

The use of optical tweezers, a useful extension of the phenomenon of optical trapping, was first demonstrated by Ashkin *et al.* (1986). The optical manipulation of individual small particles is achieved by noncontact and nondestructive methods using the equipment shown in Fig. 10. A Gaussian-mode Nd:YAG laser, with a wavelength of 1.06 μm, is used for micro-object trapping and rotation. The beam is strongly focused by a microscope objective lens (NCF Plan 100x, N.A. = 1.25) onto the micro-object. The laser power on the micro-object surface is varied from 1 to 100 mW for the experiment. The trapping behavior is recorded with a CCD camera.

Optical tweezers have found diverse applications in fields such as biotechnology (Berns *et al.*, 1992) and microchemistry (Kuo and Sheetz, 1993), where they have been used to demonstrate cell fusion and to measure mechanical properties of intracellular components (Block *et al.*, 1989) as well as to perform microchemical spectroscopic analyses, among other uses. State-of-the-art optical systems presently allow the manipulation of particles whose diameter is greater than 20 nm and smaller than 50 μm.

In order to build three-dimensional nanostructures, a multibeam laser scanning manipulation system was developed and used to demonstrate the simultaneous trapping and

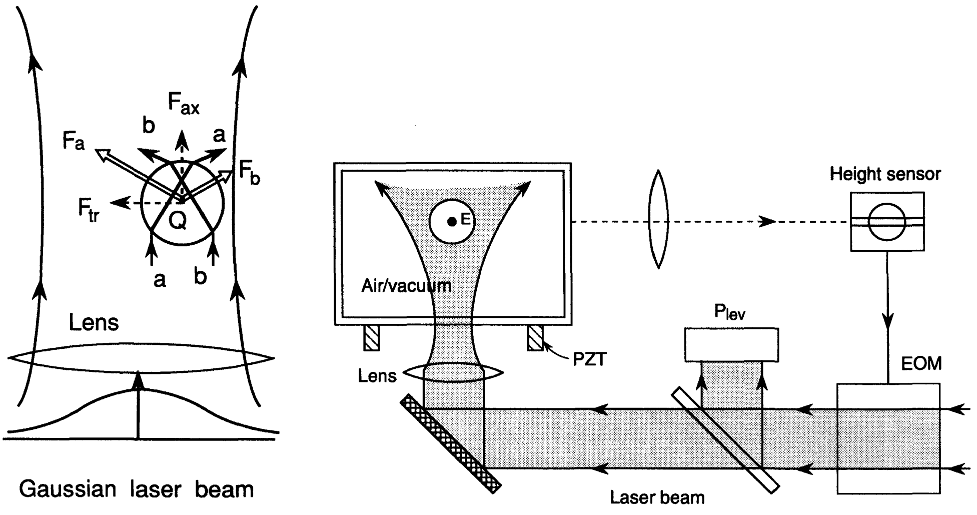

(a) (b)

FIG. 9. (a) Optical pressure on a dielectric sphere by a Gaussian laser beam and (b) the feedback apparatus for levitating the sphere (Ashkin, 1980).

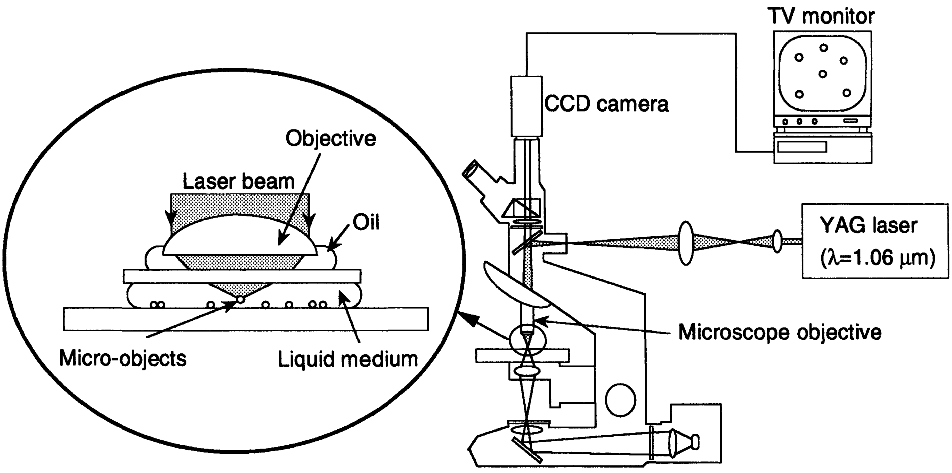

FIG. 10. Block diagram of the experimental apparatus used for optical tweezers. The micro-objects are optically trapped near the focus of the laser beam.

adhesion of multiple particles (Misawa *et al.*, 1992). Any well-defined polystyrene structure can be assembled with this system by sequentially repeating these optical manipulation and photopolymerization procedures.

A fabrication process uses physical motion to shape, move, and assemble different components into a predesigned system. If we could use optical pressure to rotate shaped elements of the mesoscopic world, we could drive such components and eventually assemble them together, manufacturing nanomachines.

As an initial step in such direction, dissymetric micro-objects fabricated by reactive ion-beam etching (RIBE) have been rotated in water by optical pressure (Higurashi *et al.*, 1994). When a dissymetric micro-object is trapped, it stands perpendicular to the substrate and parallel to the optical beam axis. In this position, it begins to rotate around the optical axis in the desired direction. Figure 11(a) shows a schematic view of the optically induced rotation of a microfabricated SiO_2 object. The rotational mechanism is described as follows. The anisotropic shape and scale of the designed micro-object are shown in Fig. 11(b). The optical pressure is based on the photon momentum change shown in Fig. 11(c). The ray refraction on the object side surfaces (i) and (ii) produces a momentum change. Thus, radiation pressure is conveyed to the object, which gives rise to a torque, as illustrated by Fig. 11(d). A microscope photograph of the trapped micro-object is shown in Fig. 11(e).

As was discussed in Sec. 3.1.1, electrostatic micromotors usually have a short lifetime on account of the surface friction between the rotor and the axle body. Since optically induced rotation requires no mechanical or electrical contacts, we believe it will play a significant role in novel photomicrodynamical systems (Ukita *et al.*, 1993). Furthermore, future research should pave the way for the optical drive technique to be applied to nanoscale mesoscopic objects with the long-term goal of making the fabrication of nanomachines possible.

4.2 Molecular Self-Assembly

Molecular self-assembly uses weak and nondirectional bonds, such as ionic bonds, hydrogen bonds, and van der Waals interactions, to obtain a final structure that is a thermodynamically minimum state. Examples of nanostructures based on self-assembly are shown in Fig. 12. The oligomer in Fig. 12(a) has an interesting helical structure based on the coordination of bipyridyl units to copper ions (Whitesides *et al.*, 1991). The toroidal *bis*-bipyridinium cyclophane in Fig. 12(b) is able to move back and forth along the backbone, a fact that has suggested its existence in the fanciful "molecular abacus" (Whitesides *et al.*, 1991). The molecular necklace in Fig. 12(c) is fabricated from

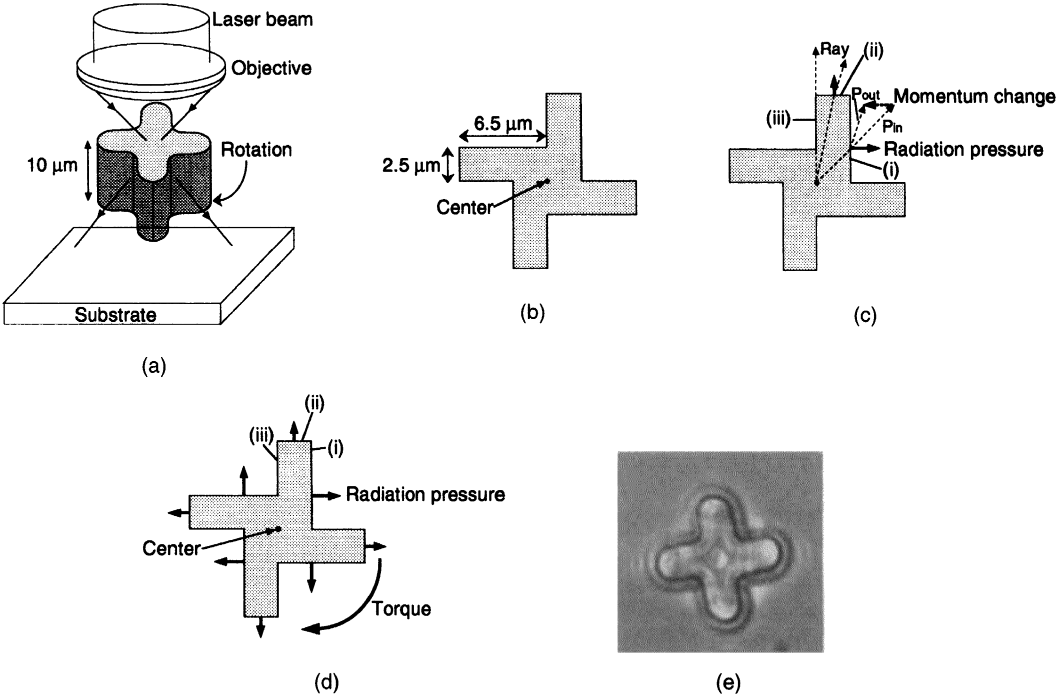
Laser beam
Objective
Rotation
10 μm
Substrate
(a)
6.5 μm
2.5 μm
Center
(b)
Ray
(ii)
(iii)
Pout
Momentum change
Pin
Radiation pressure
(i)
(c)
(ii)
(iii)
(i)
Radiation pressure
Center
Torque
(d)
(e)

FIG. 11. Optically induced rotation of an SiO_2 micro-object fabricated by surface micromachining.

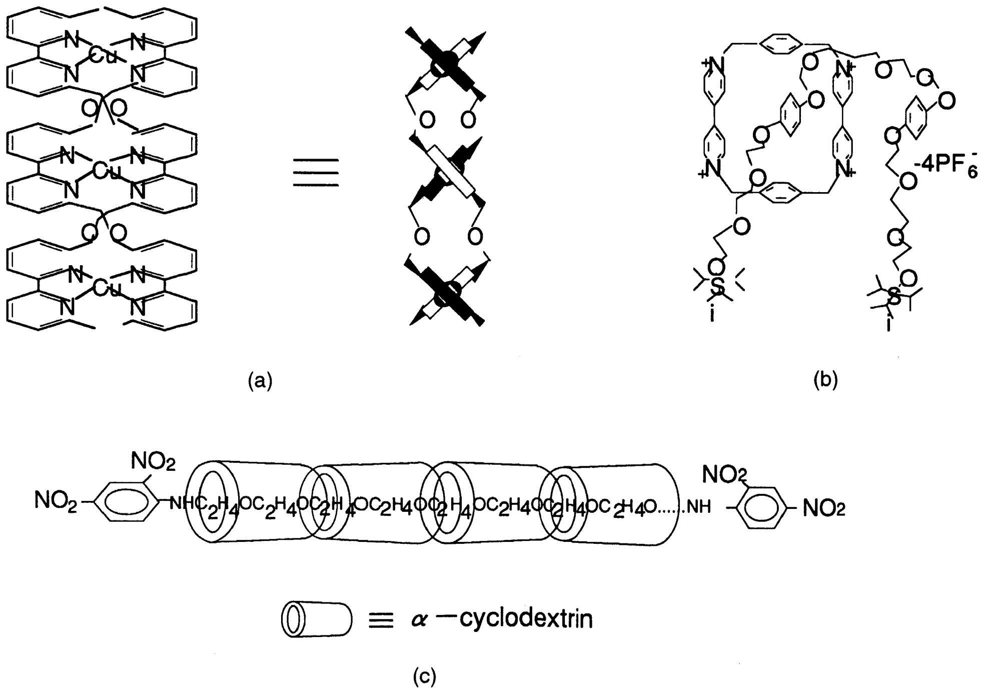
N
Cu
O
(a)
N+
O
-4PF6-
Si
(b)
NO2
NHC2H4OC2H4OC2H4OC2H4OC2H4OC2H4OC2H4OC2H4OC2H4O.....NH
NO2
≡ α—cyclodextrin
(c)

FIG. 12. Examples of nanostructure based on self-assembly. (a) Oligomer (Whitesides *et al.*, 1991). (b) Toroidal *bis*-bipyridinium cyclophane (Whitesides *et al.*, 1991). (c) α-Cyclodextrin with polyethylene glycol bisamine (Harada and Kamachi, 1992).

complexes of poly (ethylene glycol) bisamine with α-cyclodextrins by capping the complexes with 2,4-dinitroafluor-benzene. Polyethylene glycol penetrates the "breaker-like tunnel" of the α-cylodextrins (Harada *et al.*, 1992).

The use of reversible, noncovalent interactions to assemble molecules into aggregates of nanometer size will most likely be eventually successful. Tachibana *et al.* (1992) demonstrated an example where molecular self-assembly can be controlled. In the future, we should be able to assemble artificial nanoelectronic mechanical devices that are stronger and lighter than any microelectronic mechanical devices available with present-day manufacturing processes (Ball and Garwin, 1992).

4.3 Near-Field Scanning Probes

Condensed-matter physics has supported the extraordinary progress in electronics, and physicists are now extending their wide knowledge of bulk solids to the physics of surfaces, which become increasingly more important as device dimensions shrink. The invention of near-field scanning probes (SPM), such as the scanning tunneling microscope (STM) by Binnig and Rohrer in 1982 (Binnig *et al.*, 1982) and the atomic force microscope (AFM) four years later (Binnig *et al.*, 1986), has dramatically advanced our understanding of the physics of surfaces (see TUNNELING MICROSCOPY).

The near-field scanning probes use the near-field interaction between a sharp probe and the sample of interest for atomic-level surface modification (Leung and Goh, 1992). These techniques are able to handle individual atoms and molecules with extraordinary control and precision (Betzig and Trautman, 1992). For instance, surface scanning mechanisms for the STM have been developed, as shown in Fig. 13. A small probe is placed on a single-head beam. During experimental observation, the precision mechanism is scanned in the x,y direction, while the motion in the z direction is servo-controlled to hold a chosen electrical quantity constant (e.g., the tunneling current from the tip of the probe to the surface of a conducting sample). The head detects the probe's movement to observe the surface morphologies.

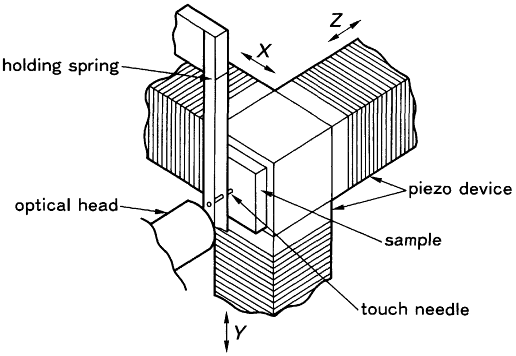

FIG. 13. Surface scanning precision mechanism for STM.

Using xenon atoms, IBM scientists drew the letters of "IBM" (Eigler and Schweizer, 1990), and researchers at NTT drew a cartoon of Albert Einstein (Utsugi, 1990) using similar STM techniques (see Plate 1). These remarkable achievements were in fact the world's first atomic-level designs. The force used to remove atoms from surfaces is the origin of friction and wear, and SPM has become a powerful tool for exploring these forces (Mate *et al.*, 1987). On the other hand, robust computer simulations of molecular dynamics that can be used to trace the movement of each atom between contacting or sliding surfaces have been developed (Landmann *et al.*, 1990). The synergy between advanced experimental techniques and powerful computer analysis has allowed the recent study of atomic-level microtribology on well-defined solid surfaces (see TRIBOLOGY). It is still quite difficult to apply these studies to "real world" surfaces, which are usually not well defined, contaminated, and not homogeneous. However, untangling these complicated mechanisms will no doubt play an important role not only in achieving no-wear micromechanical conditions but also allowing atomic-level structure modification and nanometer machining.

5. SUMMARY AND CONCLUSIONS

The miniaturization of machines has long been an important goal of technology development in many fields. In fact, it was over three decades ago that Richard Feynman delivered the now famous speech "There's Plenty of Room at the Bottom" (Feynman, reprinted 1992). His view of developments to come was remarkable in many ways. Feynman anticipated the sacrificial-layer method of making silicon micromotors, the use of electrostatic actuators, and the importance of friction and lubrication in such devices. The speech, presented in 1960 to an audience of scientists and engineers at the Jet Propulsion Laboratory (JPL), addressed the problem of manipulating and controlling things on a small scale. The question Feynman posed in the opening remarks of his lecture (Why cannot we write the entire 24 volumes of the *Encyclopedia Britannica* on the head of a pin?) startled his audience by revealing the staggering small world that still remained beneath the miniaturization levels of those early days. With remarkable vision, the leading late physicist accurately predicted the coming age of information storage on an atomic level. Although we are still not quite there, present-day developments give a clear indication that artificial precision machines on such level will indeed one day exist. In fact, Feynman himself had the opportunity to discuss his early predictions in the light of many late developments when he gave a lecture on "Infinitesimal Machines" (Feynman, reprinted 1993) 23 years later to a similar audience at the same place, as a sequel to his earlier speech.

Many important scientific and technological achievements have helped advancing the road toward the realization of Feynman's ultimate dreams in a dramatic way. As we have discussed in the present article, this recent decade has brought many impressive developments in micromachine technology. However, at the current technology level, there are still a number of serious problems to overcome. Many important sensors are yet to be fully miniaturized. There is no straightforward way of using the force delivered by actuators. Energy sources are not compact and efficient enough. There are few suitable microsystems for processors. Communication tools between different subsystems are primitive, if existant. It is difficult to build truly three-dimensional structures with complex functional parts. With all these tough hurdles to face, practical applications of intelligent-microsystems technologies are not likely to become widespread in the near future. Indeed, despite significant progress, flexible artificial systems where fully self-sufficient micromachines can dynamically adapt to changing environments are still a long way from being realized. Nevertheless, despite the difficult challenges ahead, there seems to be little doubt that such technological scenarios will happen one day. The uncertainty is when. No one questions the direction in which technology is moving, but there are major discrepancies about the time frame it will take to reach such ambitious goals.

GLOSSARY

Accelerometer: A device that detects changes of acceleration.

Actuator: A device that creates mechanical motion from other forms of energy.

DMD (Digital Micromirror Device): A micromechanical two-dimensional array of rotatable aluminum mirrors, monolithically fabricated over address-integrated circuits.

Electrostatic Micromotor: A motor, monolithically fabricated using surface micromachining, that rotates as a result of electrostatic forces.

LIGA Process: A technique developed in Germany that uses x-ray synchrotron radiation lithography to produce extremely precise polymer molds of three-dimensional micromechanical structures on a supportive substrate to generate them subsequently by electroplating metal films. The acronym used derives from the German words for lithographic (*lithographische*) electroforming (*galvanoformung*).

Lithography: A method to delineate patterns by means of selectively exposing a resist material that is sensitive to irradiation by light, x-ray, or electron beams.

Micromachine: Small functional structure whose feature size is smaller than 1 mm but greater than 1 μm (1 $\mu\text{m} = 10^{-6}$ m).

Nanomachine: Ultrasmall functional structure whose feature size is smaller than 1 μm but greater than 1 nm (1 $\text{nm} = 10^{-9}$ m).

Near-Field Scanning Probes (SPM): Atomic-scale microscopy technique that uses the near-field interaction between a sharp probe and a sample for atomic-level surface modification, such as scanning tunneling microscopy (STM) and atomic-force microscopy (AFM).

Optical Interconnection: A connection between two nodes using optical paths.

Optoelectronics: The combination of optics and electronics.

Sacrificial Etching: A technique used to fabricate micromechanical structures by depositing a thin film on top of a sacrificial layer that is then removed by etching (typically chemical).

Sensor: A device that transforms a physical or chemical perturbation into an electrical signal.

Surface Micromachining: Micromechanical structure fabrication on the surface of a substrate.

Ultrasonic Micromotor: A motor monolithically fabricated using surface micromachining that uses, as driving force, the high-energy ultrasonic vibration caused by the mechanical distortion of piezoelectric material.

Wafer Bonding: A technique to stack and bond two substrates (typically silicon) by facing the clean surfaces and then annealing them. This technique, widely used to fabricate SOI (silicon on insulator) electronic devices, can also be used to build micromechanical structures.

Works Cited

Allen, H. V., Terry, S. C., de Bruin, D. W. (1989), *Sensors and Actuators* **20** (11), 1335–1346.

Ashkin, A. (1980), "Applications of Laser Radiation Pressure," *Science* **210,** 1081–1088.

Ashkin, A., Dziedzic, J. M., Bjorkholm, J. E., Chu, S. (1986), "Observation of a Single-Beam Gradient Force Optical Trap for Dielectric Particles," *Opt. Lett.* **11** (5), 288–290.

Baliga, B. J. (1990), "Impact of VLSI Technology on Power Devices," in: N. Mikoshiba (Ed.), *Extended Abstracts of the International Solid State Devices and Materials Conference, Sendai, Japan, August 22–24,* Part I, Abs. A-1-2, Tokyo: Japanese Society of Applied Physics, 5–9.

Baliga, B. J. (1991), "An Overview of Smart Power Technologies," *IEEE Trans. Electron Devices* **38** (7), 1568–1575.

Ball, P., Garwin, L. (1992), "Science at the Atomic Scale," *Nature* **255,** 761–766.

Bart, S. F., Mehregany, M., Tavrow, L. S., Lang, J. H., Senturia, S. D. (1990), "Measurements of Electric Micromotor Dynamics," *Micro-structures, Sensors and Actuators, ASME,* **DSC-19,** 19–29.

Berns, M. W., Aist, J. R., Wright, W. H., Liang, H. (1992), *Expl. Cell Res.* **190,** 375.

Betzig, E., Trautman, J. K. (1992), "Near-Field Optics: Microscopy, Spectroscopy, and Surface Modification beyond the Diffraction Limit," *Science* **257,** 189–195.

Binnig, G., Rohrer, H., Gerber, C., Weibel, E. (1982), "Surface Studies by Scanning Tunneling Microscopy," *Phys. Rev. Lett.* **49** (1), 57–61.

Binnig, G., Quate, C. F., Gerber, C. (1986), "Atomic Force Microscope," *Phys. Rev. Lett.* **56-9,** 930–933.

Block, S. M., Blair, D. F., Berg, H. C. (1989), "Compliance of Bacterial Flagella Measured with Optical Tweezers," *Nature* **338,** 514–518.

Cho, S. T., Najafi, K., Lowman, C. L., Wise, K. D. (1989), "An Ultra-sensitive Silicon Pressure-Based Flowmeter," in: H. Geipel, A. Hartman (Eds.), *Technical Digest of the 1989 IEEE International Electron Devices Meeting, Washington, D.C., December 1–4,* New York: IEEE, pp. 499–502.

Conti, P., Hitschfeld, N., Fichtner, W. (1991), "W—An Octree-Based Mixed Element Grid Allocator for the Simulation of Complex 3-D Device Structures," *IEEE Trans. CAD* **10,** 1231–1241.

Core, T. A., Tsang, W. K., Sherman, S. J. (1993), "Fabrication Technology for an Integrated Surface-Micromechanical Sensor," *Solid-State Technol.* **36** (10), 39–47.

Crary, S. B., Zhang, Y. (1990), "CAEMENS: an Integrated Computer-Sided Engineering Workbench for Micro-Electro-Mechanical Systems," in: J. E. Wood, R. T. Howe (Eds.), *Proceedings of the IEEE Micro Electro Mechanical Systems Conference, Napa Valley, CA, February 11–14,* New York: IEEE, pp. 113–114.

Eigler, D. M., Schweizer, E. K. (1990), "Positioning Single Atoms with a Scanning Tunneling Microscope," *Nature* **344,** 524–526.

Fan, L. S., Tai, Y. C., Muller, R. S. (1988), "IC-Processed Electrostatic Micromotors," in: B. Griffing, C. Sodini (Eds.), *Technical Digest of the 1988 IEEE International Electron Devices Meeting, San Francisco, CA, December 11–14,* New York: IEEE, pp. 666–669.

Feynman, R., reprinted (1992), "There's Plenty of Room at the Bottom," *J. Microelectromech. Syst.* **1-1,** 60–66.

Feynman, R., reprinted (1993), "Infinitesimal Machines," *J. Microelectromech. Syst.* **2-1,** 4–14.

Flynn, A. M., Tavrow, L. S., Bart, R. A., Brooks, K. R., Udayakumar, K. R., Cross, L. E. (1992), "Piezolectric Micromotors for Micro-robots," *J. Microelectromech. Syst.* **1-1,** 44–51.

Goodenough, F. (1991), "Airbags Boom When IC Accelerometer Sees 50 G," *Electron. Design* **39-15,** 45–56.

Govindarajan, M., Forrest, S. R. (1991), "Optically Powered Arrays for Optoelectronic Interconnection Networks," *Appl. Opt.* **30-11,** 1335–1346.

Harada, A., Li, J., Kamachi, M. (1992), "The Molecular Necklace: A Rotaxane Containing Many Threaded α-Cyclodextrins," *Nature* **356,** 325–327.

Higurashi, E., Ukita, H., Tanaka, H., Ohguchi,

O. (1994), "Rotational Control of Anisotropic Micro-objects by Optical Pressure," in: T. Higuchi, K. Ikuta (Eds.), *Proceedings of the IEEE Micro Electro Mechanical Systems Conference, Oiso, Japan, January 25–28,* Piscataway, NJ: IEEE, pp. 291–296.

Hornbeck, L. J. (1989), "Deformable-Mirror Spatial Modulators," in: U. Efron (Ed.), *Spatial Light Modulators and Applications III,* SPIE Proceedings Vol. 1150, San Diego, CA, August 7–8, Bellingham, WA: Society of Photo-Optical Institute Engineering, pp. 86–102.

Hornbeck, L. J. (1993), "Current Status of the Digital Micrometer Device (DMD) for Projection Television Applications," in: M. Thompson (Ed.), *Technical Digest of the 1993 International Electron Devices Meeting, IEDM'93,* Piscataway, NJ: IEEE, pp. 381–384.

Howe, R. T. (1988), "Surface Micromachining for Microsensors and Microactuators," *J. Vac. Sci. Technol. B* **6** (6), 1809–1813.

Howe, R. T., Muller, R. S. (1986), "Resonant-microbridge vapor sensor," *IEEE Trans. Electron Devices* **33,** 499–506.

Kikuya, Y., Hirano, M., Koyabu, K., Ohira, F. (1993), "Micro Alignment Machine for Optical Coupling," in: A. P. Pisano, J. H. Lang (Eds.), *Proceedings of the IEEE Micro Electro Mechanical Systems Conference, Fort Lauterdale, FL, February 7–10,* Piscataway, NJ: IEEE, pp. 36–41.

Koppelman, G. M. (1989), "OYSTER, a Three-Dimensional Structural Simulator for Micromechanical Design," *Sensors and Actuators* **20,** 179–185.

Kuo, S. C., Sheetz, M. P. (1993), "Force of Single Kinesin Molecules Measured with Optical Tweezers," *Science* **26,** 232–234.

Landmann, U., Luedtke, W. D., Burham, N. A., Colton, R. J. (1990), "Atomic Mechanisms and Dynamics of Adhesion, Nanoidentation and Fracture," *Science* **248,** 454–461.

Lepselter, M. P. (1966), "Beam-Lead Technology," *Bell Syst. Tech. J.* **45,** 233–253.

Leung, O. M., Goh, M. C. (1992), "Orientational Ordering of Polymers by Atomic Force Microscope Tip-Surface Interaction," *Science* **255,** 64–66.

Li, W., Virtanen, J. A., Penner, R. M. (1992), "A Nanometer-Scale Galvanic Cell," *J. Phys. Chem.* **96,** 6529–6532.

Mate, C. M., McClelland, G. M., Erlandsson, R., Chiang, S. (1987), "Atomic-Scale Friction of a Tungsten Tip on a Graphite Surface," *Phys. Rev. Lett.* **59,** 1942–1945.

Mehregany, M., Bart, S. F., Tavrow, L. S., Lang, J. H., Senturia, S. D., Schlecht, M. F., reprinted (1990), "A Study of Three Microfabricated Variable-Capacitance Motors," *Sensors and Actuators* **A21,** 173–179.

Mehregany, M., Senturia, S. D., Lang, J. H., Nagarkar, P. (1992), "Micromotor Fabrication," *IEEE Trans. Electron Devices* **39,** 2060–2069.

Minami, K., Kawamura, S., Esasi, M. (1993), "Fabrication of Distributed Electrostatic Micro Actuator," *J. Microelectromech. Syst.* **2,** 121–127.

Mino, M., Yachi, T., Tago, A., Yanisawa, K., Sakakibara, K. (1992), "A New Planar Microtransformer for Use in Micro-switching Converters," *IEEE Trans. Magnetics* **28,** 1969–1973.

Misawa, H., Sasaki, K., Koshioka, M., Kitamura, N., Masihara, H. (1992), "Multibeam Laser Manipulation and Fixation of Microparticles," *Appl. Phys. Lett.* **60,** 310–312.

Munchmeyer, D., Langen, J. (1992), "Manufacture of 3D Microdevices Using Synchrotron Radiation," *Rev. Sci. Instrum.* **63,** 713–721.

Petersen, K. E. (1982), "Silicon as a Mechanical Material," *Proc. IEEE* **70,** 420–457.

Shulman, M. A., Ramaswamy, M., Heytens, M. L., Senturia, S. D. (1991), "An Object-Oriented Material-Property Database Architecture for Microelectro-Mechanical CAD," in: R. S. Muller (Ed.), *Proceedings of the 6th TRANSDUCERS'91,* San Francisco, CA, June 24–27, New York: IEEE, pp. 486–489.

Senturia, S. D. (1987), "Microfabricated Structures for the Measurement of Mechanical Properties and Adhesion of Thin Films," in: S. Kataoka (Ed.), *Proceedings of the 4th TRANSDUCERS'87,* Tokyo, June 2–5, Tokyo: Institute of Electrical Engineering of Japan, pp. 11–16.

Shirai, S., Kimura, K., Irita, Y., Naito, N., Srikawa, T., Kohda, S., Yamauchi, N. (1993), "Optical Beam Shifter Module with High Extinction Ratio for 1024-Input-Port Optical Switching Network," in: A. M. Goncharenko, F. V. Karpushke, G. V. Sintsyn, S. P. Apanaservich (Eds.), *1992 Topical Meeting on Photonic Switching,* SPIE Proceedings Vol. 1807, Bellingham, WA: Society of Photo-Optical Institute of Engineering, pp. 328–334.

Suzuki, K. (1990), "Single Crystal Silicon Micro-Actuators," in: L. Parrillo, A. Ipri (Eds.), *Technical Digest of the IEEE 1990 International Electron Devices Meeting, IEDM'90, San Francisco, CA, December 9–12,* New York: IEEE, pp. 625–628.

Suzuki, K., Shimoyama, I., Miura, H., Ezura, Y. (1992), "Creation of an Insect-Based Microrobot with an External Skeleton and Elastic Joints," in: W. Benecke, H. C. Petzold (Eds.), *Proceedings of the IEEE Micro Electro Mechanical Systems Conference, Travemunde, Germany, February 4–7,* New York: IEEE, pp. 190–195.

Tachibana, H., Manda, E., Azumi, R., Nakamura, T., Matsumoto, M., Kawabata, Y. (1992), "Multiple Photochemical Switching Device Based on Langmuir-Blodgett Films," *Appl. Phys. Lett.* **61,** 2420–2421.

Ukita, H., Uenishi, Y., Tanaka, H. (1993), "A Photomicrodynamical System with a Mechanical Resonator Monolithically Integrated with Laser Diodes on Gallium Arsenide," *Science* **260,** 786–789.

Utsugi, Y. (1990), "Nanometer-Scale Chemical Modification Using a Scanning Tunneling Microscope," *Nature* **347,** 747–749; (1991), "Einstein

Brought down to Size," news reported in *Nature* **349,** 556.

Vera, E. S., Yamashita, N., Yachi, T. (1992), "A Novel Power MOSFET Using Drain Trench Technology," in: M. Okamura (Ed.), *Proceedings of the 4th International Symposium on Power Semiconductor Devices and ICs, Tokyo, May 19–21,* Tokyo: Institute of Electrical Engineering of Japan, pp. 294–299.

Whitesides, G. M., Mathias, J. P., Seto, C. T. (1991), "Molecular Self-Assembly and Nanochemistry: A Chemical Strategy for the Synthesis of Nanostructures," *Science* **254,** 1312–1319.

Wright, W. H., Sonnek, G. J. (1993), "Radiation Trapping Forces on Microspheres with Optical Tweezers," *Appl. Phys. Lett.* **63,** 715–717.

Yanagisawa, K., Tago, A., Ohkubo, T., Kuwano, H. (1991), "Magnetic Micro-Actuator," in: M. Esashi, H. Fujita (Eds.), *Proceedings of the IEEE Micro Electro Mechanical Systems Conference, Nara, Japan, January 30–February 2,* New York: IEEE, pp. 120–124.

Yotsumoto, K. (1992), "The State of the Art in Telecommunications Energy Electronics," *NTT Rev.* **4** (3), 28–32.

Yoon, E., Wise, K. D. (1990), "A Multi-Element Monolithic Mass Flowmeter with On-Chip CMOS Readout Electronics," in: G. B. Hocker (Ed.), *Technical Digest of the TRANSDUCERS'90 Workshop, Hilton Head Island, SC, June 4–7,* New York: IEEE, pp. 161–164.

Younce, J. M. (1993), "Mirrors on a Chip," *IEEE Spectrum* **Nov.,** 27–31.

Zhang, S. B., Crary, S. B., Wise, K. D. (1990), "Pressure Sensor Design and Simulation Using the CAEMENS-D Module," in: G. B. Hocker (Ed.), *Technical Digest of the TRANSDUCERS'90 Workshop, Hilton Head Island, SC, June 4–7,* New York: IEEE, pp. 32–35.

Further Reading

Bryzek, J., Petersen, K., McCulley, W. (1994), "Micromachines on The March," *IEEE Spectrum* **31** (5), pp. 20–31.

Lundström, I. (Ed.), (1995), *Technical Digest of TRANSDUCERS'95 and Eurosensors IX, Stockholm, Sweden, June 25–29,* Stockholm: Royal Swedish Academy of Engineering Sciences, Vol. IVA.

Mehregany, M., (Ed.), *J. Micromech. Microeng.*, Bristol, UK: Institute of Physics Publishing.

Riezenman, M. J. (1995), "The Search for Better Batteries," *IEEE Spectrum* **32** (5), pp. 51–56.

Stix, G. (1992), "Micron Machinations," *Sci. Am.* **267** (5), pp. 107–117.

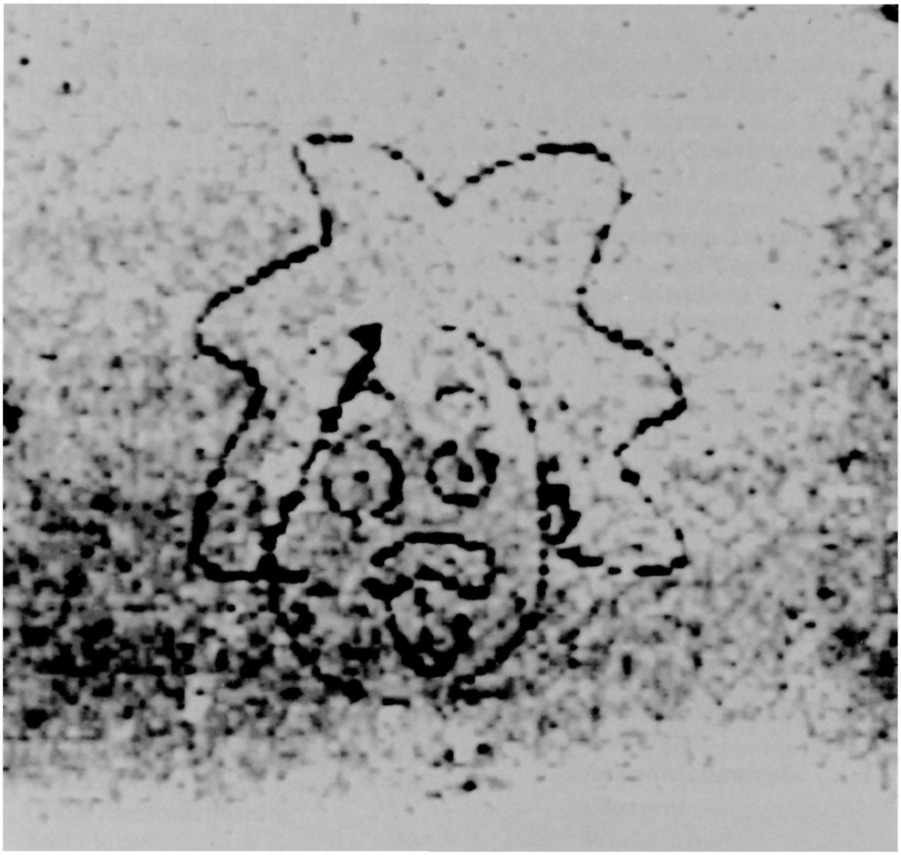

Vol. 20 – PLATE 1. Atomic-scale cartoon of Albert Einstein written by STM probe (Utsugi, 1990).

20
Targets for Particle Accelerators

Bettina Lommel and Birgit Kindler

20.1 Introduction 621
20.2 Thin-Film Targets 621
20.2.1 Thermal Evaporation 622
20.2.2 Electron-Beam Evaporation 624
20.2.3 Carbon Targets and Carbon Stripper Foils 624
20.2.3.1 Resistance Heating 625
20.2.3.2 Carbon Arc Evaporation 625
20.2.3.3 Deposition by Electrical Discharge or Pyrolysis 625
20.2.3.4 Laser Ablation 626
20.2.4 Sputtering Techniques 626
20.2.4.1 Focused Ion-Beam Sputtering 627
20.2.4.2 Saddlefield Ion-Source Sputtering 628
20.2.4.3 Magnetron Sputtering 628
20.2.5 Backings 629
20.2.6 Protective Layers 630
20.2.7 Parting Agents and Substrates for Self-Supporting Targets 630
20.3 Rolled Targets 632
20.4 Preparation of Targets by Electrochemical Deposition 632
20.4.1 Electroplating 633
20.4.2 Electrodeposition 633
20.4.3 Molecular Plating 633
20.5 Bulk Targets 634
20.6 Strategies for Increasing the Target Lifetime 636
20.7 Internal Targets 637
20.7.1 Gas Targets 637
20.7.2 Solid Targets 638
20.7.3 Cryogenically Cooled Liquid Targets 638
20.8 Liquid Targets 638
20.9 Enriched Isotopic Materials 640
20.9.1 Reduction with H_2 640
20.9.2 Reduction and Cleaning with the Hot-Wire-Iodide Process 641

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

20.9.3 Reduction with Ca, Th, or C and Extra Cleaning 641
20.9.4 The Reduction–Distillation method 641
20.9.5 Chemical Compounds 642
20.10 Characterization of Targets 642
20.10.1 Thickness 642
20.10.2 Purity and Homogeneity 643
20.11 Outlook 643
Acknowledgments 644
References 644

20.1
Introduction

To investigate the interaction between matter and matter there are, in principle, two complementary experimental setups known. In a collider, two nuclei accelerated in opposite directions, are brought to collision at a given point and the products of the reaction are detected. Alternatively, a beam of particles is accelerated onto a fixed target, and the reaction products are similarly detected; this is the subject of this chapter. The fixed target setup, as is depicted schematically in Figure 20.1, gives the possibility of combining any two elements of the periodic table.

This chapter is meant to give an overview of the different targets applied in particle accelerators and also to serve as a short introduction to the material demands and the techniques established for their production. It is not meant to be treated as a textbook, but rather as a guide for people new to this field, so we have provided an extensive bibliography furnishing details of the original publications.

The chapter is an update of the famous review by Heagney, Heagney and Maier (1997). The scientists who dedicate their work or at least a part of it in producing and developing targets are members of the International Target Development Society (INTS) (*www.intds.com*).

Sections 20.2–20.5 of this chapter are dedicated to the description of solid targets ordered by their production processes. Section 20.6 deals with strategies to increase their lifetime. Internal targets, applied in storage rings, which can be solid, cryogenically cooled, or gaseous, are discussed in Section 20.7. Section 20.8 describes liquid targets, that are mainly applied for higher beam intensities. High-precision measurements often require highly enriched isotopes as target material, which often implies special production processes, this topic is discussed in Section 20.9. The chapter concludes with a section on characterization of targets and a brief outlook of future developments in this field.

20.2
Thin-Film Targets

In this section, deposition techniques that are relevant for thin-film target production are described briefly and the necessary or optional accessories like backings, parting agents, and protective layers are explained.

Thin films are produced by different deposition techniques under vacuum. The methods differ generally in the way the

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

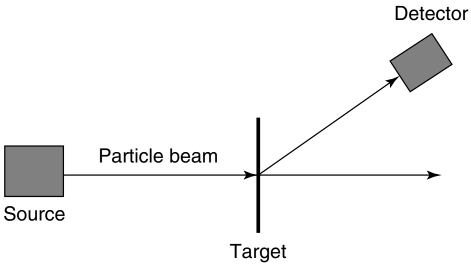

Fig. 20.1 Schematic view of a fixed target setup.

starting material is heated. If it is available as a solid piece or a solid rod, or if an inert crucible material for the process is available, thermal evaporation can be applied.

For high-melting substances or for those for which no crucible is available, either heating by an electron-beam gun or deposition by sputtering is suitable.

For a comprehensive review of this vast topic, please refer also to Muggleton (1987). Grossmann, Maier, and Friebel (1997) give an overview of the production of radioactive thin-film targets in the Munich hot-lab facility and describe specially designed resistive-heated microsources for evaporation as well as crucibles heated indirectly by electron bombardment. With these techniques, they reach a very high efficiency, which is crucial for the often scarce actinide target materials they work with and the high activities involved.

The yield differs widely for the different techniques discussed and additionally depends strongly on the applied geometry. The technique has to be chosen according to the necessary active target area and the required homogeneity of the target layer, taking into account the price and availability of the target material. Then the process has to be optimized in respect of material consumption, yield, and homogeneity of the layer.

An option to increase the thickness homogeneity and the yield of target material is to rotate the substrate or several substrates, respectively, above the evaporation source during the process; see Figure 20.2.

There is a group of targets that are specific not from the aspect of material or from the technique of deposition, but from the way they are mounted. For some experiments, for example, lifetime measurements of isomers with the recoil-distance method, it is essential to have a target surface as flat as possible to get a defined distance between the target foil and a detector. For these so-called plunger targets, the target layer has to be stretched over cone-shaped target frames, as is described, for example, by Grossmann, Maier, and Friebel (2006); alternatively, even two foils can be stretched parallel to each other in a fixed distance for a differential measurement (Zell and Dewald, 1989).

20.2.1
Thermal Evaporation

Some substances are available as massive pieces or rods to be clamped into electrodes and to be heated directly by resistance heating, as is described for carbon in Section 20.2.3. For all other materials, a crucible has to be found that does not react with the starting substance until the evaporation temperature is reached. Typical crucibles are either made of metals with a high melting point, for example, Mo,

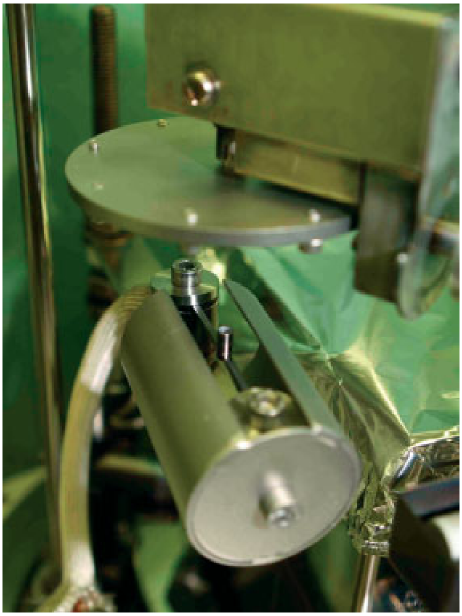

Fig. 20.2 Thermal evaporator: Tantalum crucible underneath a rotating substrate holder.

Pt, Ta, and W, or of carbon, or of ceramic compounds like BN or Al_2O_3.

At 10^{-6} or 10^{-5} mbar the material in the crucible is heated until it evaporates significantly, which happens at a vapor pressure of about 10^{-4} mbar. The vapor condenses as a thin film on the cold substrate. The difference between melting temperature and the temperature at which the material evaporates significantly defines the ease with which the process is controlled. Under high vacuum conditions, the mean free path of vapor atoms is of the same order as the vacuum chamber dimensions, so these particles travel in straight lines from the evaporation source toward the substrate. In thermal evaporation techniques, the average energy of vapor atoms reaching the substrate surface is generally low, which is advantageous for the removal of thin films from a substrate to obtain a self-supporting film.

Good candidates for thermal evaporation either out of a tungsten or of a tantalum crucible are, for example, Ag, Al, Au, Bi, Cu, Ga, Ge, Fe, In, Ti, Pb, Sn, and Zn. Though the amount of material needed for thin self-supporting targets (and the dimensions) are typically small, for highly toxic materials such as Be, Cd, or Th, the safety precautions are very important: see, for example, (Stoner, 1997; Grossmann, Maier, and Friebel, 2004) for the production of beryllium foils up to a thickness of 1 μm.

The evaporation of very volatile materials poses an additional difficulty, since they tend to stick preferably at the colder parts of evaporation chamber rather than on the substrate. They can be forced to condense on the required substrate either by high vapor stream density or by a low substrate temperature. Maier *et al.* describes this explicitly for Zn and Cd Maier *et al.* (1989) and Szerypo for HgS (Szerypo *et al.*, 2008).

While for natural materials the deposition process has to be optimized to reach a good homogeneity to obtain the required thickness, to reduce impurities from the crucible or substrate, and to get intact, pinhole-free foils, the deposition for isotopically enriched material poses an additional problem that has to be solved. Because of the limited availability of, and the often enormous prices for, highly enriched material, the process has to be optimized with regard to homogeneity and material consumption. With thermal evaporation, a yield of about 50% for a target homogeneity of ±5%, a target thickness significantly below 1 μm, and target dimensions of about 20 cm^2 can be reached. Good examples are targets of several erbium and lead isotopes, as described by Folger *et al.* (1999).

For several targets, a compound is applied instead of the pure material. This is done for two different reasons: For the rare earth materials and the low radioactive actinides, uranium and thorium, the fluoride has a much lower melting point and can be easily evaporated thermally (Thomas and Greene, 1995). The opposite is the reason for applying, for example, PbS instead of lead or Bi_2O_3 instead of the metal. The higher melting point of the compound leads to better performance and a longer lifetime of the target in the beam (Kindler *et al.*, 2006). Hg is liquid at room temperature, so HgS is one of the feasible compounds that may be used as a solid target (Szerypo *et al.*, 2008). This is discussed in detail in Section 20.6.

For materials with a very low melting point or materials that are typically available as compounds, the preferable process can be reductive evaporation. The reduction process and the choice of reductant is described in detail in Section 20.9. Reductive evaporation results in the deposition of the reduced material on a backing or a substrate in this process; see (Maier *et al.*, 1989; Maier and Grossmann, 1987; Lommel *et al.*, 2006) for Pb, Zn, Cd, and Sm.

20.2.2
Electron-Beam Evaporation

In electron-beam evaporation, the coating material rests in a water-cooled copper hearth so that it stays cold at the contact surface and a reaction or alloying of the target material with any hot crucible surface is prevented. An electron beam is generated in a heated filament, electrostatically accelerated, and – depending of the type of gun – deflected on the surface of the target material. The electron beam can either be focused to a point or can be wobbled across a larger area of the surface of the material.

Electron-beam evaporation is the ideal means for the production of targets out of refractory metals like W or Ta. The material consumption in this method, at least with commercially available standard electron guns, is rather high. This is a drawback when isotopically enriched targets are needed. To overcome this problem, there have been modifications to commercial guns, as described by Thomas *et al.* (1991) or even the development of a very compact electron gun that is optimized for the production of targets for accelerators by Heagney and Heagney (1985, 2004).

The method is applied preferably where refractories like tungsten are used as target material (Folger *et al.*, 1997; Stoner and Jolivet, 2006) but also for example, in magnetic layers of Gd (Maier-Komor *et al.*, 2004) and large area targets of ^{28}Si prepared by electron-beam evaporation (Greene and Berg, 2005).

20.2.3
Carbon Targets and Carbon Stripper Foils

Carbon is a very special material related to applications in accelerators not only as target and stripper foil but also as a backing for other materials, as is discussed in detail in Section 20.2.5. Though stripper foils are not really targets, they play a central role in cyclotrons, linacs, and tandem accelerators, and they are almost exclusively made of carbon. They are mostly developed and prepared by the same community, making use of almost the same techniques as targets, which justifies this issue also being addressed within the scope of this section.

Because of this broad range of applications of carbon foils, a number of investigations and developments are dedicated to the production of carbon with a special focus on high durability and long lifetime upon irradiation of the foils. Some preparation techniques are almost exclusively discussed in connection with carbon foils. These techniques are summarized in the following section. For a review of the different techniques for preparation of carbon foils, see also Stoner and Miller (2006) and the comprehensive citations therein.

20.2.3.1 **Resistance Heating**

In this method, spectroscopically pure carbon is heated directly by a high current in a high-vacuum recipient. There are, in principle, two different ways of arranging the carbon. Lommel *et al.* (2002) clamp a spectroscopically pure carbon rod vertically between water-cooled electrodes. Since in this arrangement the sublimation of the carbon upon resistive heating takes place over a zone of several centimeters of the rod, the method allows a deposition of large areas and a wide thickness range from some micrograms per square centimeter up to about $100\,\mu\mathrm{g\,cm^{-2}}$. Muggleton (1987), for example, applies two pointed carbon rods of high purity clamped in water-cooled electrodes that are held in contact while a current of approximately 200 A passes through them. The carbon sublimes preferably at the junction of the two rods. This method is mainly used for the production of thin carbon layers of up to $5\,\mu\mathrm{g\,cm^{-2}}$.

20.2.3.2 **Carbon Arc Evaporation**

Carbon arc evaporation is one of the earliest processes for producing carbon foils and strippers and has been investigated and improved over the decades. In this technique, two ultrapure carbon rods are mounted in a vacuum recipient. With one electrode fixed and the other one movable, a carbon arc is ignited and the carbon atoms are evaporated with significantly higher energy compared to conduction heating, resulting in quite differently structured carbon layers. As described, for example, by Takeuchi *et al.* (1979) different electrode designs and either an alternate current (ac) or a direct current (dc) are applied.

Sugai *et al.* (2006) developed stripper foils that consist of alternating layers of carbon prepared with ac and dc arc-discharge technique, respectively. Carbon foils from a dc arc discharge show a good durability against high-intensity beams, but are rather fragile, whereas foils ablated from both electrodes in an ac arc discharge have a comparably short lifetime in a high-intensity beam, but are mechanically very strong. Doping the carbon electrodes with different elements leads to an increase in the thickness achievable.

Jaggi *et al.* (2006) report on the production of carbon foils that have a thickness of $100–200\,\mu\mathrm{g\,cm^{-2}}$ with carbon arc deposition, which show an enhanced lifetime in high-power cyclotron applications.

20.2.3.3 **Deposition by Electrical Discharge or Pyrolysis**

Auble *et al.* (1980) produced carbon foils of thickness $2–10\,\mu\mathrm{g\,cm^{-2}}$ by vapor deposition and by cracking of a gas mixture of 10% argon and 90% ethylene in a dc glow discharge.

Liechtenstein *et al.* (1997, 1999) used glow discharge sputtering deposition with a low-density krypton plasma to produce thin diamond-like carbon foils (DLC foils) as stripper for tandem accelerators and compared the durability with those produced with other methods.

20.2.3.4 Laser Ablation

In pulsed laser deposition (PLD) or pulsed laser ablation, the source material is irradiated with a high-intensity laser beam of more than 100 MW cm^{-2} and instantaneously evaporated. The evaporated material is ionized and accelerated by the laser beam, and it condenses on a suitable substrate. An advantage of PLD compared to other deposition techniques is the possibility of controlling the amount deposited on the substrate by the number of laser pulses, but to guarantee the reproducibility of deposition it is mandatory to move the target perpendicular to the laser beam after each pulse to avoid burning of holes. Other advantages are that (i) the stoichiometry of the target is transferred exactly to the substrate, (ii) multilayers are produced easily, and (iii) the process is suited for computer control. Furthermore, it is possible to simultaneously prepare layers of different material classes like metals, semiconductors, ceramics, or polymers.

The disadvantages are that (i) the deposition rates are rather low and (ii) the formation of unwanted droplets is possible. The method is only suitable for small areas and the equipment is relatively expensive; the laser beam entrance window, especially, is problematic when the thickness of the required layer becomes too high.

Dollinger and Maier-Komor (1991) adapted the laser ablation technique for producing durable carbon stripper foils for application in tandem accelerators in an ultrahigh vacuum plant with a Nd:YAG laser. They later upgraded their set up with a vacuum interlock, thus increasing the obtainable foil thickness that had earlier been limited by the decreasing transmission of the laser beam entrance window. For details, please refer to (Maier-Komor, Dollinger, and Krücken, 2006) and citations therein.

20.2.4
Sputtering Techniques

It is very difficult, or almost impossible, to prepare some materials as thin targets with evaporation techniques such as, for example, high-melting elements like refractory metals when isotopically enriched material is needed, or for compounds that easily decompose upon heating.

It is common to all sputtering methods that the deposition of the target atoms does not depend on the vapor pressure since sputtering is a purely ballistic process that is relatively cold in the sense that substrate heating during deposition is negligible. This implies that the substrate can be placed relatively close to the source resulting in a high efficiency of the processes and also that heat-sensitive materials like plastics can be applied as substrates or backings.

Sputtering techniques are also advantageous for materials with a low ''sticking coefficient'', such as cadmium, which are extremely difficult to condensate on substrates with thermal methods. In general, sputtering methods lead to excellent condensation and good adhesion to the substrate. On the other hand, this can be a problem when self-supported layers are needed.

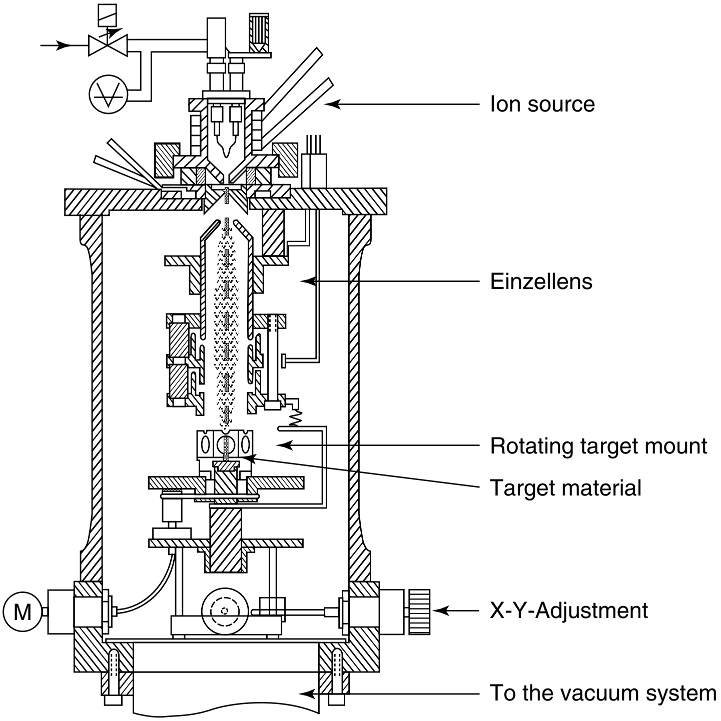

Fig. 20.3 Sketch of a Sletten-type focused ion-beam apparatus.

20.2.4.1 Focused Ion-Beam Sputtering

Heavy or focused ion beam sputtering was one of the first sputtering methods applied for the production of targets for accelerators in the case of scarce materials.

As shown in Figure 20.3, an argon ion beam is extracted from a duoplasmatron ion source into a high-vacuum chamber and focused to a beam spot of diameter smaller than 2 mm. The ions are accelerated to about 10 keV onto the cathode. The material to be deposited is placed on the cathode in the form of a bead or pressed powder inside the beam stop around which the substrates are arranged concentrically. Details of the process are described by Sletten and Knudsen (1972), who developed the technique especially for isotopic targets of refractories and oxides. Some further applications can be found, for example, in Folger, Klemm, and Müller (1983) and Kindler *et al.* (2004).

The major advantage of the focused ion-beam sputtering is the small beam focus possible and the extreme efficiency of the method that makes it possible to work with very small amounts of starting material. For this reason, focused ion-beam sputtering is particularly suitable in all cases where the target material is scarce or expensive.

On the other hand, the close geometry of the setup leads to a reasonable uniformity of the deposited layers only for small target areas and the small beam focus results in low deposition rates and in extremely long deposition times for thick or large layers. Another drawback of the technique is the rather complex and bulky electrical supply, which makes it expensive and complicated.

20.2.4.2 Saddlefield Ion-Source Sputtering

Instead of a duoplasmatron source, Drinkwater (1982) describes the application of a saddlefield ion source for the production of the above-mentioned problematic materials. A saddlefield ion source, as is described, for example, by Muggleton (1991) or in a modified version by Turkentine and Weisser (1997), is mounted in a vacuum chamber and the ion beam is focused on the surface of the sputtering target with 30° inclination.

Maier (1991) discuss the application of high-vacuum sputter deposition for the production of various nuclear accelerator targets. Here, a beam of xenon atoms or ions produced by a saddlefield-type source is applied. Targets of a wide range of different materials are made and characterized and the foils are reported to be of high purity and excellent mechanical strength (Greene *et al*. 1997). The method is applied not only to metals but also to oxides that are sputtered from the compound with an additional O_2 partial pressure and to nitrides that are sputtered reactively from metal targets with a nitrogen partial pressure present in the deposition chamber.

Sputtering with a saddlefield ion source can also be used for etching or thinning of foils down to 0.5 μm or even thinner, as is described by Thomas and Greene (1993, 2003).

Since this technique combines the advantages of a small beam focus and a high efficiency of the focused ion-beam sputtering with a rather compact and easy-to-operate setup that works reliably with respect to long-term runs, it has become a powerful and commonly used technique for the production of difficult small-size thin-film accelerator targets in many laboratories.

20.2.4.3 Magnetron Sputtering

In magnetron sputtering, ions, most commonly of argon or other noble gases, are produced in a low-pressure glow discharge. The ions are then concentrated by the magnetic field of a permanent magnet situated below the sputtering target and are accelerated on to the surface of the sputtering target by an electric field.

Conducting materials can be sputtered with a dc source whereas for insulators an RF magnetron is needed. The equipment is rather expensive and the material consumption high, especially since the starting material has to be available in the form of a sheet of at least 1 in. diameter. The method is advantageous for refractory materials and in all cases where a good adhesion to a substrate is required. Also large-area targets can be produced with a good homogeneity (Kindler *et al.*, 2004). The method also provides the opportunity to produce mixed targets by cosputtering different kind of materials from different sources in parallel or to do a reactive sputtering by using a suitable sputter-gas mixture.

Magnetron sputtering of magnetic materials like iron, cobalt, or nickel is very problematic since the magnetic susceptibility of the sputtering target interferes with the field of the permanent magnets of the magnetron source and the ignition of the plasma is impossible. Weidong *et al.* (2002) use an assisting magnetic field produced by a stack of permanent magnet rings above the sputtering target for the preparation of Fe targets with a thickness of up to 1000 nm. Kindler *et al.* (2004), on the other hand, use a stack of thin foils of Co or Ni each not thicker than 250–300 μm, to produce targets with a thickness of 100 $\mu g\,cm^{-2}$ for Co and 200 $\mu g\,cm^{-2}$ for Ni.

Sometimes, the availability of the wanted target material in the form of a suitable sputtering target can be problematic, especially in the case of ceramic materials, which are mostly pressed and/or sintered. The homogeneity and density of commercially available sputtering targets is difficult. For isotopically enriched materials, the only possibility for magnetron sputtering is to apply a rolled foil. But this is only an option for isotopes that are not too expensive.

20.2.5
Backings

For several materials, there is no substrate and no parting agent available to produce a self-supporting target or the desired target layer is too thin to make it self supporting. In these cases, a backing is needed, which supports the layer, does not affect the target material, and does not interfere too much with the intended measurement. The backing thickness should be homogeneous, and the surface should be clean and free from imperfections, as a deposition typically duplicates the quality of the substrate. The starting material should have a high purity and should not diffuse into the target material.

In many cases, carbon is an ideal backing, as it can be produced with high purity, made self-supporting over a wide thickness range, and prepared very thin, down to $5\,\mu g\,cm^{-2}$. Additionally, it has a very low Z, which allows, when dealing with significantly heavier reaction products, an easy separation of the relevant signals from the target material from those signals coming from nuclear reactions of the beam with the backing. But if the production process or the application is thermally or mechanically very stressful for the target-backing combination, a metal is preferable. With respect to a metal backing with a low Z, beryllium is often the material of first choice. However, for beryllium one must be able to ensure that no beryllium oxide is produced during the production process or in the application, as beryllium oxide is highly toxic. Alternatively, aluminum or titanium can be applied as a backing.

There is a large variety of plastic foils available, which have the same advantage of measurement as explained for carbon backings above. The disadvantages for plastics are their modest thickness homogeneity, no defined composition, a low heat resistivity, and a poor electrical conductivity, which can cause problems in their production as well as during the experiment. Polyethylene, polycarbonate, polyvinyl, polyimide are the plastics usually applied. The thinnest commercially available plastic is Mylar, a polyethylene terephthalate, which can be bought with a thickness down to $2\,\mu m$. Kapton, a polyimide, which is mechanically and thermally fairly stable, is commercially available down to $7\,\mu m$, but has been produced down to $44\,\mu g\,cm^{-2}$ ($\sim 0.5\,\mu m$) by Pauwels *et al.* (1979) or down to $10\,\mu g\,cm^{-2}$ ($\sim 0.07\,\mu m$) by Stolarz and van Gestel (2006).

For applications that do not observe the transmitted products, all good-quality thin films that are available can be feasible backings.

For in-beam γ-ray experiments, the backing has to be thick enough to stop the beam or the recoil, but it should not contribute to the expected γ-spectrum. In this case, the backing material and the backing thickness are determined by the expected energy range of the signals.

20.2.6
Protective Layers

To hinder sputtering in the beam or to reduce the oxidation of a sensitive target material, protective layers are applied.

Carbon and gold are commonly used as protective layers, depending on whether a very low Z or a fairly high Z is required, as they can be easily separated from the reaction products to be measured. Both materials are easily available and can be deposited by several techniques. Both materials are stable under air, even under moist conditions, and have a fairly high sublimation or melting point, respectively.

Depending on the experiment, all other materials feasible for evaporation can be applied. The deposition process for the protective layer should not heat the target too much or damage it. The material of the protective layer should preferably not interfere with the experiment or diffuse into the target material, which might be a problem with gold depending on the target material. In addition, the deposition should be feasible over a large thickness range.

In the beginning, for target production for heavy-ion accelerators, targets dipped in organic lacquer or paint were commonly used. As these organic substances often are neither suitable for ultrahigh vacuum conditions nor for an intense beam, this is no longer an option.

20.2.7
Parting Agents and Substrates for Self-Supporting Targets

As targets are typically needed to study the interaction of a projectile with the target, a defined material of a defined purity and a defined thickness is needed.

The ideal target is elemental and self-supporting. This is the main difference from the production process of commercially produced thin films. Even if most of the techniques are related, the demand for a self-supporting thin film makes the difference. A major task in the target production is preparing the substrate in such a way that the target layer can be removed and a complete intact film without any additional impurities is obtained.

In principle, two different techniques are applied: either there is a substrate available on which the desired target can be deposited as a film and which can be dissolved completely after the process, or a flat glass or metal substrate is coated with a soluble intermediate layer on which the target is subsequently evaporated.

The substrate or substrate plus intermediate layer has to be strainless, smooth, and homogeneous with as little structural unevenness as possible, since any structure of the substrate is reproduced by the deposited layer. It should be chemically stable and no trace of the material of the substrate should be finally found in the target.

In the early stages of heavy-ion targetry, depositions on collodion, which gives a celluloid-like film, were used for quasi-self-supporting target production. As the collodion film is not stable in a heavy-ion beam, the beam burns off the substrate during the experiment. Recent measurements in atomic physics show that the leftover collodion corresponds to a substrate thickness of 2–5 $\mu g\,cm^{-2}$ with an uneven thickness distribution, which cannot be tolerated by all experiments. So nowadays, collodion should be considered more like one of the feasible plastic backings (see Section 20.2.5). The papers by Braski and by Gao *et al.* (2007) give a good comparison between the target surface roughness obtained by applying a

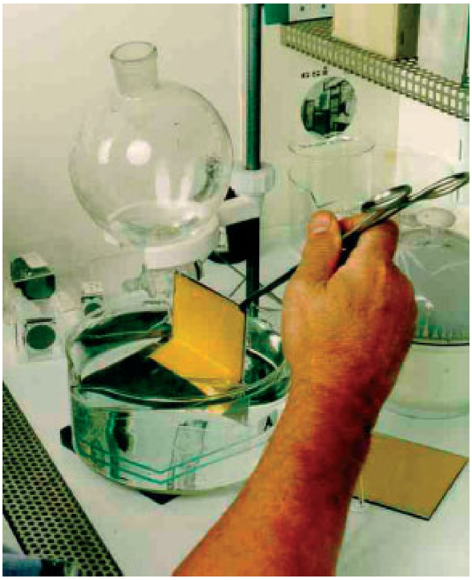

Fig. 20.4 Gold foil fetched with a frame.

crystallized intermediate layer and by using plastic backings.

For the alternative substrate plus soluble intermediate layer, a whole variety of water soluble intermediate layers mainly deposited on glass substrates exists. Evaporated alkali halides, for example, NaCl, CsI, or $BaCl_2$, tensides (detergents), or betaine–sucrose are frequently applied. The process of floating is shown for a gold foil in Figure 20.4. Scientists from all countries typically apply well-known branded detergents. The alkali halides and betaine–sucrose have the major advantage of being much more stable at higher temperatures, but the application process has to guarantee that the intermediate layer crystallizes in an acceptable smoothness as the structure is repeated by the deposited layer, and this should not lead to unbearable roughness. However, a little roughness allows the foil to release tension during drying and upon irradiation. For the alkali halides, it is an optimization between substrate, preferably glass, evaporation velocity, and distance. They are evaporated thermally on slides or very smooth glass; a comparison of the roughness achievable is given by Braski (1972). For betaine–sucrose, the production process, the mix ratio, and the humidity during the coating play a major role (Lommel *et al.*, 2002; Maier-Komor, 1972; Thalheimer *et al.*, 1999).

For the other possibility of depositing directly on a soluble substrate, copper is very often the material of choice. The copper foil should be fairly clean, 99.9% or better, flat, without any structure from recrystallization or cold rolling, strainless, preferably thermally stress relieved. The requirements are given by the need to release a complete, strainless, pinhole-free, and clean target. For the copper substrate, nitric acid is a typical solvent, depending on the deposited target material. The concentration of the acid is chosen to allow a slow solution process (Folger *et al.*, 1997; Kirchner *et al.*, 1981; Meens, Rossini, and Sens, 1993).

In principle, for a self-supporting target, any material for which there is a known solvent that will dissolve the substrate material without dissolving the target material is feasible as a substrate

20.3
Rolled Targets

Metal targets in the thickness range of micrometers are preferably prepared by rolling a thick foil in a sandwich of clean stainless steel plates in a mechanical rolling mill, with successive rolling giving thinner foils. Depending on the material, often, even on the charge available, foils down to several hundreds of micrograms per square centimeter can be achieved (Karasek, 1963; 1972; Folger, 1999).

For very ductile metals like indium or lead, the application of a pack of teflon sheets either instead of, or inside, the stainless-steel pack is advantageous to prevent the soft material from sticking to the inside surface of the pack, as described by Clifford *et al.* (2002).

Rolling metals sensitive to oxygen or moisture, for example, calcium, sodium, or uranium, down to a thickness of a micrometer or even thinner, is a difficult task (Maier, 1982). Good results are obtained by rolling in a glove box in an inert-gas atmosphere with an oxygen- and water-content below 1 ppm.

Some metals with a very low rollability at room temperature have to be prepared at higher temperatures. Therefore, the material is packed between stainless-steel plates, heated up to temperatures at which the ductility is higher, and then immediately rolled in the hot pack. This process has to be repeated until the desired thickness is achieved. After about 50% of thickness reduction, a heat treatment under vacuum for stress relief is necessary. It is important to keep the temperature below the recrystallization temperature. For beryllium, tungsten, and vanadium, this process is discussed in detail by Karasek (1979). The area achievable for a foil depends on the width of the rolls. Parameters that can be varied are the velocity of the rolls and the amount of blocking up of the rolls after each run, which determines the pressure exerted on the foil. Another way to influence the quality of the foil is to either roll it always along the same axis or to rotate the pack after each cycle through 90° or 180°, respectively.

Commercially available hot rollers can be heated up to 150 °C with hot liquid running through the rollers, thus keeping them at a constant temperature, which is suitable for several plastic foils such as polyethylene or polypropylene.

The process parameters for cold rolling metals out of isotopically enriched material can differ significantly from those obtained for natural ones, since the ductility of metals depends crucially on the impurities of the starting material. As the production process for isotopic material differs from the production process of commercially available high purity metal, the ductility and the behavior when rolling might differ. Cleaning processes in advance may be necessary (Manente and Blunt, 1995; Maier, 1989).

20.4
Preparation of Targets by Electrochemical Deposition

Electrochemical deposition (ED) techniques are widely used in cases where the desired target material is available only in limited amounts, is highly radioactive, or both. The advantages of these methods are their simplicity and deposition yields approaching 100% in select cases as well as the capability of producing a wide range in target thicknesses from a few micrograms per square centimeter up to several milligrams per square centimeter. Since for ED, a compound of the target material is dissolved in either aqueous or organic solvents, easy and complete recovery of precious target material is possible in case the deposition is incomplete. In principle, one can distinguish three different techniques:

1. Electroplating, where a thin metallic film is deposited onto a substrate that serves as the cathode in the deposition cell;
2. Electrodeposition of insoluble hydroxide compounds of highly electropositive elements, which cannot be deposited in metallic form from aqueous solutions; and
3. molecular plating, where deposition is performed from organic solutions such as isopropyl alcohol or acetone.

20.4.1
Electroplating

Here, the material of interest is deposited from ionic solutions by means of electrolysis. At the cathode, the dissolved ions are reduced to the elemental state and a metallic film of the desired element results (Paunovic and Schlesinger, 2006). Since electroplating is also widely used for industrial coatings on different substrates, a large number of procedures for the deposition of elements such as Cr, Ni, Zn, Ag, and Au are available and can be adapted for target fabrication. Target thicknesses of several milligrams per square centimeter are possible. The method is suitable for isotopic enriched material owing to high efficiency of up to 90%. Metallic foils or carbon might be used as substrates. Self-supporting targets can be produced by removing the substrate after completion of the deposition process by means of selected etching solutions. In this way, self-supporting Os targets with an average target thickness of 500 $\mu g\,cm^{-2}$ have been produced using Cu as substrate for electroplating and subsequent removal of the Cu backing by dissolving the Cu in a mixture containing trichloracetic acid and ammonia (Cecchi, 1991).

20.4.2
Electrodeposition

Metallic films of highly electropositive elements such as the lanthanides or the actinides cannot be produced by electroplating. In these cases, deposition of the hydroxide can be performed from aqueous acidic solutions containing NH_4Cl or NH_4NO_3 as electrolyte. At current densities up to several amperes per square centimeter, hydrogen is evolved at the cathode and the desired elements are deposited in the form of their hydroxides by the hydroxyl ions that are generated at the cathode (Hansen, 1959; Evans *et al.*, 1972; van der Eijk, Oldenhof and Zehner, 1973; Trautmann and Folger, 1989; Da Cruz and Poldena, 1990; Becerril-Vilchis *et al.*, 1996; Ingelbrecht *et al.*, 1997; Bajo and Eikenberg, 1999; Eberhardt *et al.*, 2001). For electrodeposition, the backing (the cathode) must be inert in highly acidic solutions and thus, metal foils of Ti, Ta, or Pt are most often used. Small target spots with a diameter of about 3 mm can be prepared.

20.4.3
Molecular Plating

For molecular plating, the lanthanide or actinide compound, typically the chloride or the nitrate, is dissolved in a small

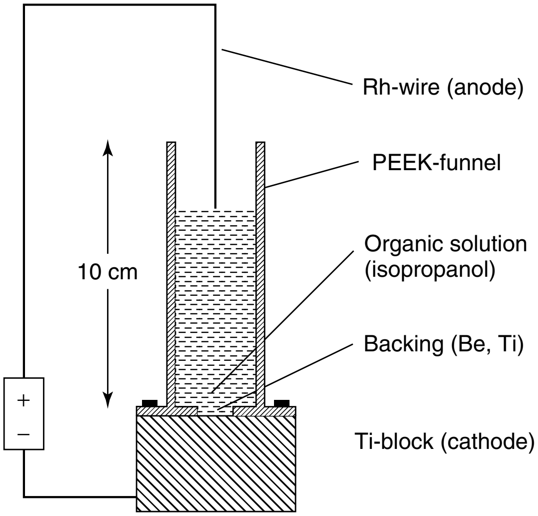

Fig. 20.5 Schematic view of the cell used for molecular plating (Eberhardt *et al.*, 2004). A funnel made of polyether-etherketone (PEEK) confines the area to be plated and acts as a solution container with a volume of 15 ml. The backing foil is sealed to the funnel by means of a Viton O-ring and a Ti block that serves as the cathode. An Rh wire is used as anode material.

volume (5–10 μl) of nitric acid and the aqueous phase is mixed with a surplus of an organic solvent (a few milliliters). The suitable solvents are isopropanol, isobutanol, or acetone. Under these conditions, no electrolytic dissociation occurs by applying an electric current at voltages up to 3000 V (Eberhardt *et al.*, 2001; Parker and Falk, 1962; Parker, Bildstein, and Getoff, 1964; Parker *et al.*, 1964; Getoff and Bildstein, 1965; Filippone and Wahlgren, 1986; Eberhardt *et al.*, 2004). It is assumed that the dissolved molecules form clusters that migrate in the electric field toward the cathode. However, the deposition process is not yet completely understood. In general, the composition of the deposition is complex and also contains possibly fragments of the solvent molecules. In cases where Be or Al is required as backing material, molecular plating must be applied, since these materials quickly dissolve in acidic solutions and thus electrodeposition is not possible. Figure 20.5 shows a schematic view of a cell used for molecular plating. The same setup can also be applied for electrodeposition.

20.5
Bulk Targets

There are applications where targets with a thickness of millimeters or even centimeters are needed. Some examples are production targets for pions or for exotic nuclei in fragment separators or parts related directly to the ion optics of the accelerator such as the energy degrader. In these cases, often elaborate mechanical preparation of shapes and surfaces is necessary. The samples have to be cut, turned, polished, and lathed, sometimes with high-precision requirements. Surface quality and accuracy of the target thickness over the whole relevant beam-spot area is especially a crucial point since the interaction that a particle beam experiences in a target and the energy loss of the particle beam in a degrader both depend directly on the amount of matter the beam passes. Therefore a lateral inhomogeneity in the thickness of the target or degrader inadvertently leads to a broadening of the reaction peak or a spreading of the momentum distribution of the beam particles. A good example for a complex variety of targets out

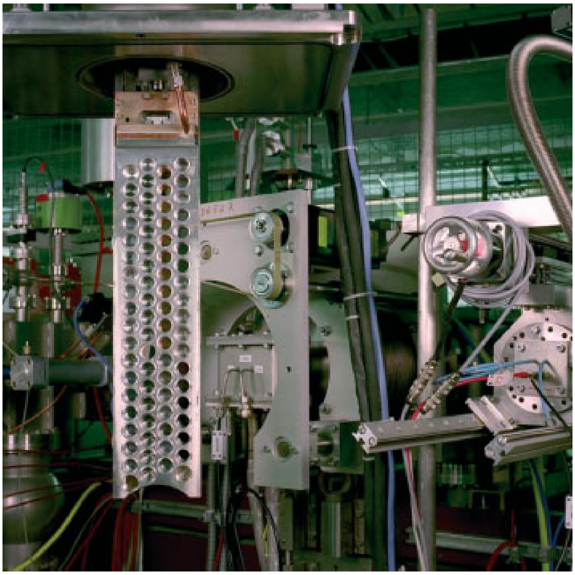

Fig. 20.6 Target ladder at the fragment separator FRS at GSI.

of different materials with different surface conditions and different thicknesses is the target ladder shown in Figure 20.6 in which 74 production targets for the fragment separator at GSI are mounted (Folger *et al.*, 1991).

Sometimes, the starting material is only available as a powder, for example, for some compounds like oxides or fluorides. In these cases, the pressing of tablets can be a rather straightforward method to get usable targets. For this purpose, a hydrostatic press with appropriate dies is typically used. Sometimes, simultaneously heating or evacuating the die is advantageous. Pressing a tablet of the starting material can also be a necessary preparation for applying the material in a thermal or electron beam-gun evaporation process.

Stolarz (1997) describes powder sedimentation from a glue solution as a very effective technique for preparing rather thick targets (25–65 $\mathrm{mg\,cm^{-2}}$) from enriched isotopic materials such as ruthenium-96 with high melting points and a hardness that makes pressing of pills inapplicable.

Sugai (1997, 1999) developed a target preparation method in which a powder is deposited by high-energy vibrational motion of particles in the electrostatic field between parallel electrodes, the so-called HiViP-method. The target material in form of a powder is placed in a cell consisting of two parallel metal discs serving as electrodes and separated by a pyrex tube. On the inner side of the two electrodes, a metallic backing foil of aluminum or copper is clamped. By increasing to high voltage slowly, the powder particles become charged, and at a particular voltage a vibrational motion of the particles sets in. The deposition takes place at about 10 kV over a period of 5 to 15 hours at room temperature and at ambient pressure. Data for various target materials and different thicknesses are given. The foils do not show a shiny metallic surface but are slick black. However, they are of high mechanical strength combined with a good homogeneity and high purity.

Using this method, free-standing targets of 50–125 mg cm^{-2} of natural silicon as well as enriched ones of Si-28 have been also prepared successfully; this is rather difficult with any other method (Sugai *et al.*, 2006).

20.6
Strategies for Increasing the Target Lifetime

While dealing with solid-state targets, there are cases where the intensity of the particle beam impinging on the target is too high for the target material. Evaporation or partial melting of the target layer leaving droplets on the backing or holes in the layer are the consequences, as described, for example, by Lommel *et al.* (2002). Different possibilities for the enhancement of target lifetime are discussed in this section.

One measure to prevent overheating of the target is to rotate it through the beam thus distributing the beam intensity over a larger target area and, at the same time, giving each target time for cooling before the next beam pulse hits. This method probably was mentioned for the first time by Organessian *et al.* (1975) at the Joint Institute for Nuclear Research in Dubna, Russia. To date, it has been applied and improved upon over there (Lobanov *et al.*, 1997) and at GSI in Darmstadt, Germany (Münzenberg *et al.*, 1989; Schädel, 2006) especially for experiments in nuclear physics and chemistry. The concept is adopted in a number of facilities worldwide, for example, the Argonne National Laboratory in Argonne, USA (Greene, Thomas, and Leonard, 1995; Roa *et al.*, 1996; Greene *et al.*, 2004) and at Grand Accelerateur National d Ions Lourds in Caen, France (Stodel *et al.*, 2006). Yoshida *et al.* (2004) at RIKEN, Japan also use the rotating-wheel concept and combine it with an additional cooling for two very different target types. In one case, they apply a massive rotating disc out of carbon, for example, which is in contact with a water-cooled plate, thus further increasing the tolerable beam intensity. In the other case, they operate a wheel with thin targets that rotates in a He-atmosphere providing a convective cooling. Antalic *et al.* (2004) describe measurements for various fixed targets and rotating target wheels in vacuum as well as in a stationary He-gas and a He-gas flow as cooling media.

For the highest intensities – where the energy deposition is too high for solid materials to survive – the use of liquid metals as targets is also under development, as is described in Section 20.8.

A totally different approach to enhance the durability and the lifetime of targets for some special cases is the substitution of the elemental material for a compound. For experiments in superheavy element synthesis, where high beam intensity is essential because of very low reaction cross sections, targets of lead and bismuth are frequently used. To overcome the problem arising from the very low melting temperatures of these metals, different higher melting compounds were investigated for the SHIP-experiment at GSI (Lommel *et al.*, 2002) and thereupon lead was substituted for PbS and bismuth was substituted for Bi_2O_3 (Kindler *et al.*, 2006). Similar problems emerge with materials that have a very high vapor pressure or are even liquid in the elemental form. Maier-Komor (1985) report on suitable compounds for phosphorus, sulfur, mercury, and iodine targets. Greene and Lister (2002) test different compounds for the application of sulfur targets and Szerypo *et al.* (2008) describe the preparation of HgS as an alternative. An application of compounds can also be necessary when the desired element is chemically unstable, such as some of the rare earth, alkaline, or alkaline earth metals that oxidize rather fast when exposed to air. In these cases, often the halides or the oxides can be used; see, for example, (Maier-Komor, 1989).

20.7
Internal Targets

For high-precision experiments, storage-ring operation with internal targets – in combination with phase-space cooling – is often an attractive alternative to single-pass experiments. Since the beam is passing an internal target typically with a frequency of 1 MHz, the target has to be several orders of magnitude thinner than necessary for a single-pass experiment.

Therefore internal targets installed in storage rings in combination with electron-cooling of stored ion beams are a well-established technique to perform high-precision experiments (Melissinos and Olsen, 1975; Stringfellow *et al.*, 1986; Ekström, 1992; Macri, 1984). The large number and variety of internal target types that have been developed and are in operation worldwide is the result of various technical and scientific boundary conditions (Baglin *et al.*, 1980; Kilian *et al.*, 1982; Burgei *et al.*, 1982; Pollock, Miller, and Singh, 1980; Popov, 1990). The design and the technical realization of each internal target has to optimize the conditions for the physics experiments to be made. Maximum luminosity (target thickness (atoms per square centimeter) times particle-beam intensity (particles passing through the target per second)) has to be reached under optimized duty-cycle conditions of the storage ring (particle-beam lifetime, accumulation time, and cooling time). Schematically, the large number of existing internal targets can be divided into two main types: gas targets and solid targets; additionally, cryogenically cooled liquid beams, which are a sort of intermediate type of internal targets, are also addressed.

20.7.1
Gas Targets

Gas targets are realized by the expansion of a gas through a nozzle into the vacuum system of the target installation. As storage rings are mainly operated under ultra-high vacuum (UHV) conditions ($p < 10^{-10}$ mbar), the effect of the gas load created by the gas inlet on the storage-ring system has to be minimized. Technical solutions are multistaged differential pumping and skimming sectors inside the installation (Dick and Kubischta, 1986; Gruber *et al.*, 1989; Reich *et al.*, 1997). Gas target density (target thickness) strongly depends on gas species, nozzle geometry, inlet pressure, and inlet temperature (Hagena and Obert, 1972; Krämer *et al.*, 2001). Adjusting these parameters, reachable target densities vary between 10^{10} atoms cm^{-2} (pure gaseous) and approximately 10^{15} atoms cm^{-2} (gas cluster). This flexibility in target species and target thickness makes gas targets a universal installation mainly chosen at existing storage rings. The existing gas target setups were usually designed to be used in a specific experimental setup (atomic physics and nuclear physics), optimized to deliver a well-defined target thickness of a specific gas species, but can be used with other gases species with decreased performance. A special type of construction are gas targets delivering polarized gas atoms created by an atomic beam source in combination with a storage cell (Steffens, 1991; Steffens, 1990; Mairon *et al.*, 1991).

20.7.2
Solid Targets

Solid targets in storage-ring environments are realized with the use of different materials in different forms (Ekström, 1992; Macri, 1984). Thin fibers or microribbons can be introduced into the circulating particle beam (Koch *et al.*, 1988; Lozowski and Hudson, 1991). Microparticles (e.g., dust) may be enclosed in a cell or injected as a beam, or pellets (frozen drops) are created in a generator and move across the particle beam (Bickel *et al.*, 1990; Tanabe, Sugai, and Yasue, 1987; Meyer *et al.*, 1990; Trostell and Kullander, 1990). As solid targets locally have a target thickness of approximately 10^{19} atoms cm^{-2}, problems could occur in particle-beam cooling performance and also in the storage-ring duty cycle.

The use of internal targets has been discussed at several conferences, workshops, and accelerator schools where detailed information is given in the form of reports (Ekström, 1992; Macri, 1984; Kilian *et al.*, 1982; Krämer *et al.*, 2001; Steffens, 1991; 1990; Holt, 1984; Oelert, 1987).

20.7.3
Cryogenically Cooled Liquid Targets

For many experiments, like those planned at the future Facility for Antiproton and Ion Research (FAIR) at Darmstadt, low-Z internal targets, that is, hydrogen and helium, are particularly important. Laval-type nozzles with a diameter of $d_0 = 100\,\mu\text{m}$ or greater are typically employed, providing a target density in the interaction region in the range $n{\sim}10^{11}-10^{14}\text{cm}^{-3}$ and a target spatial extension that spreads out over several millimeters. These features (low density and large interaction length) make gas-jet internal targets unsuitable for several storage-ring experiments that require a luminosity in the order of 10^{30} or greater. Larger target densities and smaller interaction lengths can be achieved by further cooling the gas to temperatures down into the liquid regime, below 20 K for hydrogen and below 4 K for helium.

Producing liquid beams of helium and hydrogen, however, is not straightforward. To make these beams entirely compatible with the high-vacuum conditions in a storage ring, the use of sub-10-micrometer diameter nozzles is mandatory. On the other hand, nozzle-clogging represents an experimental challenge when micrometer-size apertures are used. Thus, microscale liquid jet operation requires scrupulous cleaning at all stages of beam preparation, submicrometer filtering of the liquid, and the use of gases with ultrahigh-purity. These experimental issues have been extensively investigated recently and the feasibility of producing liquid beams from $d_0 = 2$-μm-orifice diameters has been demonstrated by Grisenti (2003) for both helium and hydrogen Grisenti *et al.* (2006).

20.8
Liquid Targets

Liquid targets are used sometimes when the target material needed actually is a liquid or a liquefied gas; in such cases the material is handled mostly in cells as for gases (see Section 20.7.1).

Additionally, in recent years, liquid targets have been discussed intensely in connection with high power applications at accelerators where the energy deposited in the target is too high for standard solid materials to withstand. Other concepts to withstand higher energy depositions are described in Section 20.6.

Liquid metals have been in use as coolants for more than 30 years in nuclear reactors worldwide; lithium, sodium, and lead–bismuth alloys have been in use and the techniques for handling and pumping them are well established. Other applications of liquid metals include cooling of detectors; in 1977, Grand and Goland even proposed a closed loop of liquid lithium as a windowless target for a high-power neutron generator in a material test facility for future fusion reactors (Grand and Goland, 1977). For an overview of the applications in the past as a basis for the new developments, see Nolen *et al.* (2002) and the citations therein.

In connection with high energy deposition in targets, there are basically two different fields of application where intense development for liquid targets is in progress. One is the production of radioactive isotopes for nuclear reaction experiments that had, in the past, only been possible with stable isotopes. The other field is the operation of high-intensity proton accelerators for the generation of intense neutron beams in spallation sources or for the production of neutrinos and muons for particle-physics research. In all applications, the liquid metals act as target and coolant simultaneously.

For the production of radioactive isotopes again there are, in principle, two different approaches: the in-flight method and the isotope separation online technique (ISOL), which were planned to be combined in the rare isotope accelerator (RIA). With the in-flight technique, exotic nuclei are produced by fragmentation and fission of heavy projectiles on low Z target materials. For high power fragmentation targets, Nolen *et al.* (2003) propose a windowless jet target of liquid lithium operated in a close loop suitable for intensities in the 400-kW range for a uranium beam. Such a target was planned for RIA and also for its follow-up version ''RIA-lite'' (Nolen, 2007); it is proposed as a possible solution for the fast extraction mode of the planned Super-FRS (Fragment Separator) of Gesellschaft für Schwerionenforschung GSI (Geissel *et al.*, 1977). For the new high acceptance fragment separator of the National Superconducting Cyclotron Laboratory (NSCL) at Michigan State University (MSU) – the A1900 – a hybrid target is being constructed as high power production target for lighter beams where thicker target layers are required. To reduce the target thickness, liquid lithium flows through a channel in a tapered beryllium block connected to a rigid stainless-steel loop thus acting as coolant and at the same time as target material (Nolen *et al.*, 2003b; Morrissey *et al.*, 2003).

In the ISOL technique, usually a proton or light-ion beam is directed onto a thick target to produce the rare isotope at rest and then a postaccelerator is employed to provide a secondary beam. For high power facilities of this type, Nolen *et al.*, (2002) proposed a two-step target where a porous tungsten block acting as primary neutron converter is cooled with liquid lithium and is surrounded by a secondary target of fissionable material. This concept involves some hazards with liquid lithium being in direct neighborhood of water-cooled and fissionable material. Bollen *et al.* (2006) describe, alternatively, a concept based on liquid mercury as primary target, which has the advantage of being not flammable.

Neutron spallation sources traditionally work with refractory metal targets. For new or upgraded facilities that will work with intense beams in the megawatt-regime, liquid metal targets are being studied or are under construction and material behavior and corrosion in high-intensity, high-radiation environments are being investigated (Mansur, 2006). The Swiss Spallation Neutron Source (SINQ) at the Paul Scherrer Institute (PSI), the most powerful accelerator-driven facility worldwide, is running a pilot experiment for a megawatt liquid metal spallation target (MEGAPIE) in a joint international initiative. It is planned to work with liquid lead–bismuth eutectic in a closed loop and will be operated at around 200 °C (Bauer, Salvatore, and Heusener, 2001). Amongst others, a closed loop of liquid mercury in a container as a spallation target is under construction for the Spallation Neutron Source (SNS) at Oak Ridge National Lab (ORNL) (Mansur and Haines, 2006) and in the Material Life Science Facility (MLF) of the new Japan Proton Accelerator Research Complex (J-PARC) (Oyama, 2006) as well as projected for the European Spallation Source (ESS) (Bauer and Ullmaier, 2003).

Nishimura (2002) propose a liquid metal target station without a window where liquid bismuth is pumped in the upward direction through a tube, spilling over the rim and then returning to the reservoir tank. The charged particle beam comes from above, thus hitting the free surface of the liquid metal.

Ravn *et al.* (2003b) also propose a free liquid mercury surface as production target for the neutrino factory at CERN. In this design, the proton beam is injected through an annular nozzle collinearly into a free surface mercury-jet flowing with high axial velocity.

20.9
Enriched Isotopic Materials

To understand the interaction of two different materials, it is important to apply starting material that is as clean as possible, and not to add any impurities during the production process. Quite often, this is not enough. For a number of experiments, isotopically enriched targets or source material is needed. The most economic way is to buy these as the chemical compound required, with the desired enrichment. However, the suitable compound is not always available or the required compound is disproportionately expensive. Therefore one has to adapt chemical synthesis on a small scale with a high yield. In the early days of targetry, most of the metals were obtained by electrolytic methods. As this results in low compact material with intrinsic impurities, cleaning processes had to be added or alternative reduction methods had to be found, as discussed in the following sections.

20.9.1
Reduction with H_2

As described in any chemistry text book, heated metal oxides can be reduced by hydrogen. This technique gives good results and relatively clean metals even when applied on a small scale, which is typical for isotopical enriched materials. For reduction with hydrogen, the metal oxide is placed in a quartz tube in a furnace. During heating, a little hydrogen flow is streamed through the tube. Good results are obtained for oxides of Cd, Cr, Cu, Pb, and Zn. With this technique, metal of a good and pure quality, also suitable for cold-rolling processes, can be achieved (Manente and Blunt, 1995).

20.9.2
Reduction and Cleaning with the Hot-Wire-Iodide Process

Isotopically enriched Ti is mostly available as oxide. As it needs a highly reactive reductant, it can be reduced with Ca or CaH_2. The resulting material may contain up to 2% oxygen, so cold rolling is impossible. With the hot-wire–iodide process (Van Arkel–de Boer process) the titanium is converted into TiI_4, the gaseous compound is cracked at a hot tungsten wire, the metal deposits at the tungsten wire, and the gas is applied again for forming the metal iodide out of the metal. With this process hafnium, thorium, titanium, and zirconium can be cleaned so that the metal is rollable (van Arkel and van Boer, 1925). For isotopic titanium, the process is described by Maier (1979) in detail.

20.9.3
Reduction with Ca, Th, or C and Extra Cleaning

Often simple reduction processes leave intrinsic impurities in the required isotope. Quite often, an unacceptable amount of reductant or oxygen can be found in the foil, which makes the subsequent processes for the target production difficult or impossible. Spirlet, Müller, and van Audenhove (1985) describe the reduction of Pu with Ca, with Th, and with C. After the reduction they recommend either electrorefining or an extra evaporation and deposition of the material obtained. The following section combines these processes in one.

20.9.4
The Reduction–Distillation method

As direct reduction processes often show impurities of the reductant, a reduction and distillation process is a clean alternative. The isotopically enriched isotope and the reductant must have a significantly higher melting point and accordingly a very low vapor pressure in the temperature range in which the resulting isotopically enriched metal already shows a significant vapor pressure. So the mixture of the oxide and the reductant perform a solid–solid reaction, the oxide is reduced, and the resulting metal condenses on a water-cooled collector above the setup, as shown in Figure 20.7. Depending on the metal to be reduced, different reductants are used. This method has been primarily developed and discussed by Kobisk *et al.* at the ORNL (Kobisk and Adair, 1979;

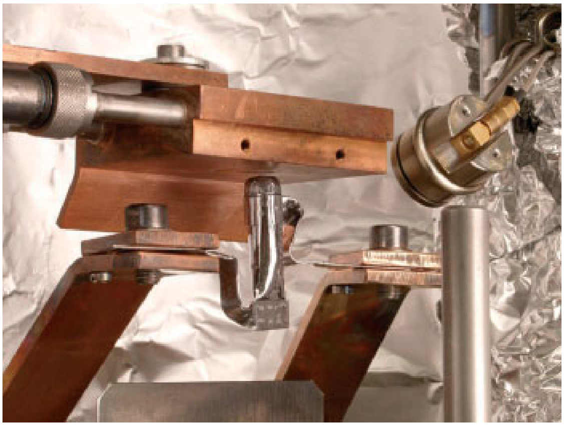

Fig. 20.7 ^{48}CaO mixed with Zr in a tantalum crucible. The reduced Ca is deposited on a water-cooled copper plate (Lommel *et al.*, 2006).

Kobisk and Grisham, 1969). They apply the technique for K, Ca, Sr, Ba, most of the rare earth oxides, Np, Pu, Am, and Cm. As reductants, they work with Ta, Zr, La, At, and Th. With this technique, pure isotopically enriched metal, with a yield of up to 92% and of a quality required for cold rolling, can be obtained. For each isotopically enriched oxide and the available amount of material, the process has to be optimized. For different materials, different reductants can be used. Today, Hf, La, and Zr are mostly the reductant of choice. The quality of the reductant, the size of the surface, and the enthalpy of both metal partners and the vapor pressure of both metals play an important role (Pengo *et al.*, 1991).

Marcus and Zevenbergen (1999) apply the reduction–distillation method with carbon as reductant, as known from steel production, for different zinc isotopes and obtain metal of good quality.

Sugai *et al.* describe the process of reducing the rare earth oxide via the fluoride. Then in an analogous process, calcium is applied to remove the fluorine and the metal is deposited on a cooled lid above the setup. Fairly clean isotopically enriched rare earth metal with a yield of 50–95% can be obtained (Sugai and Kato, 1985).

20.9.5
Chemical Compounds

In several situations where the isotope cannot be applied directly, compounds have to be used. The reasons are manifold: Sometimes, the compound is available more easily or is easier to produce in a target of the desired thickness; sometimes, the compound can be processed with a more favorable material consumption; and sometimes the applied compound shows a higher melting point or a lower sputter yield resulting in a better behavior in the beam as discussed in Section 20.6. In those cases, isotopic material has to be converted in the desired compound. Lommel *et al.* discuss the substitution of low melting metals. They replace lead by lead sulfide, and several rare earth metals by their fluorides. For this conversion of pure isotopic material, classical chemistry has to be adapted on a small scale for micro- or milligram amounts with a high yield under very clean conditions (Lommel *et al.*, 2006).

20.10
Characterization of Targets

The characterization of the target before application as well as after an irradiation is essential for the success and the accuracy of the experiment. The relevant parameters such as absolute thickness and homogeneity of the layer, variation in thickness across the active target area, and purity of the target have to be determined and controlled as accurately as possible. A survey of only the most commonly used methods is given in this chapter.

20.10.1
Thickness

The main parameter for a target is the absolute thickness of the layer. For most applications, weighing the substrate before and after the coating process is the most straightforward method to determine the target thickness in milli- or micrograms per square centimeter since this gives a direct measure for the number of nuclei interacting with a given beam diameter. In some cases, it may be not possible to determine the correct thickness of the target directly, for example, if the substrate is too heavy in relation to the target layer or if the target has to be coated additionally with a protective layer. When this is so, a dummy substrate can be measured that is either coated in parallel to the same process and is situated in an equivalent place with respect to the coating geometry or is coated in an extra process with equivalent conditions. Generally, balances of high accuracy are the workhorse for every target maker.

For an online measurement, oscillating quartz crystals are very commonly used as a monitor for evaporation or sputtering processes. They give a good qualitative idea of the achieved thickness, but for the determination of absolute thickness, a careful calibration is mandatory, as reported by Maier-Komor, Dollinger and Hammann (1991).

For thin targets that are transparent in visible or ultraviolet light, absorption with a spectrophotometer can be a fast method to measure the absolute thickness, provided that a prior calibration for the respective element was done. Lommel *et al.* (2002) discuss the accuracy of optical absorption and compare it with Rutherford backscattering measurements; the latter is a very accurate but rather complex method that is not applicable for daily quality control. Profilometers are the devices that have more commonly been in use during the last years for thickness determination of thin targets. They are suitable for very accurate measurements, but only on a very small scale.

For somewhat thicker layers, energy loss measurements with α-particles is a common method, as described, for example, by Alexander and Wirth (1985).

For radioactive targets, the radioactive decay itself can be detected either simply with a detector or with a radiographic imager, which gives a superior spatial resolution, as described by Liebe *et al.* (2008). The signals can be converted to a corresponding target thickness with a suitable calibration.

For bulk targets, the thickness can be measured with incremental length gauges with an accuracy of about 0.5 μm. In combination with an automatic setup, the thickness distribution of very large areas also can be measured with a high lateral accuracy, as reported by Kindler *et al.* (2002).

20.10.2
Purity and Homogeneity

In the first place, one has to make sure that the starting material is as pure as is necessary for the experiment and work in an environment that is as clean as possible. To prevent any contamination during the deposition process, the vacuum should be in the region of 10^{-7} hPa region, and attention has to be paid to the crucible material applied.

To check the appearance of the target layer and search for inclusions, holes, or other defects, optical microscopy is a method that is available in almost every laboratory. A more fancy but very powerful instrument is a scanning electron microscope that in combination with an energy dispersive X-ray analysis even allows for element identification, at least for impurities existing in the range of percent.

20.11
Outlook

In numerous accelerator facilities worldwide, the intensity of the beams available is increasing constantly and this has to be accompanied in equal measure by a target development. New concepts for the enhancement of target durability have to be found for future experiments to profit from the more and more intense beams, as was described in Sections 20.6 and 20.8. In this context, the label *high power target* has already found its way into literature.

The science and art of fabricating targets that are adapted to the experiment as optimally as possible remain a challenging task in the twenty-first century, and facilities dedicating manpower and funds to this task will secure a competitive advantage.

Acknowledgments

Special thanks to Klaus Eberhardt, Universität Mainz, Institut für Kernchemie, Fritz-Strassmann-Weg 2, D-55128 Mainz, Germany, who wrote Section 20.4, *Preparation of Targets by Electrochemical Deposition*. We thank Hartmut Reich-Sprenger from GSI, Vacuum Group for the Section 20.7, *Internal Targets*, and Robert Grisenti, Universität Frankfurt a. M., Institut für Kernphysik, Max-von-Laue-Str. 1, 60438 Frankfurt am Main for the Section 20.7.3 *Cryogenically cooled liquid targets*. We want to express our gratitude to John Stuart Winfield, GSI, for looking closely at the final version regarding English style and grammar, correcting both and offering suggestions for improvement.

References

Alexander, J.R. and Wirth, H.L. (**1985**) *Nucl. Inst. Methods*, **A 236**, 545–48.

Antalic, S., Cagarda, P., Ackermann, D., Burkhard, H.-G., Heß, F.-P., Hofmann, S., Kindler, B., Kojouharova, J., Lommel, B., Mann, R., Saro, S., and Schött, H.-J. (**2004**) *Nucl. Instrum. Methods*, **A 530**, 185–93.

van Arkel, A.E. and van Boer, J.H. (**1925**) *Z. anorg. algem. Chemie*, **148**, 345–50.

Auble, R.L., Baier, J.K., Galbraight, D.M., Jones, C.M., Stelson, P.H., and Weiser, D.C. (**1980**) *Nucl. Instrum. Methods*, **177**, 289–94.

Baglin, C. *et al.* (**1980**) Charmonium Spectroscopy at the ISR Using an Antiproton Beam and a Hydrogen Jet Target, CERN/ISRC/80-14.

Bajo, S. and Eikenberg, J. (**1999**) *J. Radioanal. Nucl. Chem.*, **242**, 745–51.

Bauer, G.S., Salvatore, M., and Heusener, G. (**2001**) *J. Nucl. Mater.*, **296**, 17–33.

Bauer, H. and Ullmaier, G.S. (**2003**) *J. Nucl. Mater.*, **318**, 26–37.

Becerril-Vilchis, A., Cortes, A., Dayras, F., and de Sanoit, J. (**1996**) *Nucl. Instrum. Meth. Phys. Res.*, **A 369**, 613–16.

Bollen, G., Baek, I., Blideanu, V., Lawton, D., Mantica, P.F., Morrissey, D.J., Ronningen, R.M., Sherrill, B.S., Zeller, A.F., Beene, J.R., Burgess, T., Carter, K., Carrol, A., Conner, D., Gabriel, T., Mansur, L., Remec, I., Rennich, M., Stracener, D., Wendel, M., Ahle, L., Boles, J., Reyes, S., Stein, W., and Heilbronn, L. (**2006**) *Nucl. Inst. Methods*, **A 562**, 915–20.

Braski, D.N. (**1972**) *Nucl. Instrum. Methods*, **A 102**, 553–66.

Burgei, R., Garcon, M., Grand, M., B. Gonel, Maillard, R., Malthiery, A., and Martin, J. (**1982**) *Nucl. Instrum. Methods*, **204**, 53–57.

Buschmann, W., Bickel, M., Dombrowski, H., Gaul, G., Grzonka, D., Hölker, G., Santo, R., and Wähning, M. (**1990**) *Nucl. Instrum. Methods*, **A 295**, 44–52.

Cecchi, A. (**1991**) *Nucl. Instrum. Meth. Phys. Res.*, **A 303**, 102–5.

Clifford, S., Guo-ji, Xu., Ingelbrecht, C., and Pomeroy, M.J. (**2002**) *Nucl. Instrum. Methods*, **A 480**, 29–35.

Da Cruz, P.A.L. and Poldena, R. (**1990**) *Nucl. Instrum. Methods*, **A 286**, 453–56.

Dick, L. and Kubischta, W. (**1986**) *Physics with Jet Targets at the SPS Antiproton-Proton Collider, Hadronic Physics at Intermediate Energy* (ed. T. Bressani and R.A. Ricci), Elsevier Science Publishers B.V., p. 209.

Dollinger, G. and Maier-Komor, P. (**1991**) *Nucl. Instrum. Methods*, **A 303**, 50–58.

Drinkwater, R.J. (**1982**) *Proceedings of 11th World Conference of the INTDS*, Seattle, 217–26.

Eberhardt, K., Schädel, M., Schimpf, E., Thörle, P., and Trautmann, N. (**2004**) *Nucl. Instrum. Methods*, **A 521**, 41–44.

Eberhardt, K., Thörle, P., Nähler, A., and Trautmann, N. (**2001**) *AIP Conference*

Proceedings (eds J.L. Duggan and I.L. Morgan) Vol. **5**, 76, pp. 1144–47.

van der Eijk, W., Oldenhof, W., and Zehner, W. (**1973**) *Nucl. Instrum. Methods*, **112**, 343–51.

Ekström, C. (**1992**) Internal Targsets, CERN Accelerator School 1991, CERN Report 92-01, 120.

Evans, E.J., Lougheed, R.W., Coops, M.S., Hoff, R.W., and Hulet, E.K. (**1972**) *Nucl. Instrum. Methods*, **102**, 389–401.

Filippone, B.W. and Wahlgren, M. (**1986**) *Nucl. Instrum. and Methods*, **A 243**, 41–44.

Folger, H. (**1999**) *Nucl. Instrum. Methods*, **A 438**, 131–51.

Folger, H., Geissel, H., Hartmann, W., Klemm, J., Münzenberg, G., Schardt, D., Schmidt, K.-H., and Thalheimer, W. (**1991**) *Nucl. Instrum. Methods*, **A 303**, 24–33.

Folger, H., Hartmann, W., Heß, F.P., Hofmann, S., Klemm, J., Münzenberg, G., Ninov, V., Thalheimer, W., and Armbruster, P. (**1995**) *Nucl. Instrum. Methods*, **A 362**, 64–69.

Folger, H., Hartmann, W., Klemm, J., Lommel, B., and Thalheimer, W. (**1997**) *Nucl. Instrum. Methods*, **A 397**, 55–58.

Folger, H., Klemm, J., and Müller, M. (**1983**) *IEEE Trans. Nucl. Sci.*, **NS-30** (2), 1568.

Gao, F., Zheng, R., Cheng, G., and Yang, T. (**2007**) *Nucl. Instrum. Methods*, **A 577**, 397–401.

Geissel, H., Weick, H., Winkler, M., Münzenberg, G., Chichkine, V., Yavor, M., Aumann, T., Behr, K.H., Böhmer, M., Brünle, A., Burkard, K., Benlliure, J., Cortina-Gil, D., Chulkov, L., Dael, A., Ducret, J.-E., Emling, H., Franczak, B., Friese, J., Gastineau, B., Gerl, J., R. Gernhäuser, Hellström, M., Jonson, B., Kohouharova, J., Kulessa, R., Kindler, B., Kurz, N., Lommel, B., Mittig, W., Moritz, G., Mühle, C., Nolen, J.A., Nyman, G., Roussell-Chomaz, P., Scheidenberger, C., Schmidt, K.-H., Schrieder, G., Sherrill, B.M., Simon, H., Sümmerer, K., Tahir, N.A., Vysotsky, V., Wollnik, H., and Zeller, A.F. (**1977**) *Nucl. Instrum. Methods*, **145**, 49–76.

Getoff, N. and Bildstein, H. (**1965**) *Nucl. Instrum. Methods*, **36**, 173–75.

Grand, P. and Goland, A.N. (**1977**) *Nucl. Instrum. Methods*, **145**, 49–76.

Greene, J.P. and Berg, G.P.A. (**2005**) *Nucl. Instrum. Methods*, **B 241**, 1006–8.

Greene, J.P., Heinz, A., Falout, J., and Janssens, R.V.F. (**2004**) *Nucl. Instrum. Methods*, **A 521**, 214–21.

Greene, J.P. and Lister, C.J. (**2002**) *Nucl. Instrum. Methods*, **A 480**, 79–83.

Greene, J.P., Thomas, G.E., and Leonard, R.H. (**1995**) *Nucl. Instrum. Methods*, **A 362**, 81–89.

Greene, J.P., Thomas, G.E., and Schiel, S.L. (**1997**) *Nucl. Instrum. Methods*, **A 397**, 99–104.

Grisenti, R.E., CostaFraga, R.A., Petridis, N., Dörner, R., and Deppe, J. (**2006**) *Europhys. Lett.*, **73**, 540–46.

Grisenti, R.E. and Toennies, J.P. (**2003**) *Phys. Rev. Lett.*, **90**, 234501.

Grossmann, R., Maier, H.J., and Friebel, H.U. (**1997**) *Nucl. Instrum. Methods*, **A 397**, 39–45.

Grossmann, R., Maier, H.J., and Friebel, H.U. (**2004**) *Nucl. Instrum. Methods*, **A 521**, 1–4.

Grossmann, R., Maier, H.J., and Friebel, H.U. (**2006**) *Nucl. Inst. Methods*, **A 561**, 62–65.

Gruber, A., Bourgeois, W., Franzke, B., Kritzer, A., and Treffert, C. (**1989**) *Nucl Instrum. Methods*, **A 282**, 87–93.

Hagena, E.W. and Obert, W. (**1972**) *J. Chem. Phys.*, **56**, 1793.

Hansen, P.G. (**1959**) *J. Inorg. Nucl. Chem.*, **12**, 30–37.

Heagney, J.M., Heagney, J.S., and Maier, J.J. (**1997**) *Encyclopedia Appl. Phys.*, **20**, 473–96.

Heagney, J.M. and Heagney, J.S. (**1985**) *Nucl. Instrum. Methods*, **A 236**, 530–32.

Heagney, J.M. and Heagney, J.S. (**2004**) *Nucl. Instrum. Methods*, **A 521**, 235–38.

Holt, R.J. (ed.) (**1984**) *Workshop on Polarized Targets in Storage Rings, ANL*, Argonne, ANL-84-50.

www.intds.com

Ingelbrecht, C., Moens, A., Eykens, R., and Dean, A. (**1997**) *Nucl. Instrum. Meth. Phys. Res.*, **A 397**, 34–38.

Jaggi, V., Pavan, R.A., and Zeisler, S.K. (**2006**) *Nucl. Instrum. Methods*, **A 561** (1), 1–3.

Karasek, F.J. (**1963**) *Nucl. Sci. Eng.*, **17** (3), 16–19.

Karasek, F.J. (**1972**) *Nucl. Instrum. Methods*, **102**, 457–58.

Karasek, F.J. (**1979**) *Nucl. Instrum. Methods*, **167**, 165–66.

Kilian, K., Möhl, D., Gspann, J., and Poth, H. (**1982**) *Internal targets for lEAR, workshop on physics at lear with low-energy cooled antiprotons*, Erice.

Kindler, B., Ackermann, D., Hartmann, W., Heß, F.P., Hofmann, S., Lommel, B., Mann, R., and Steiner, J. (**2006**) *Nucl. Instrum. Methods*, **A 561**, 107–11.

Kindler, B., Brand, H., Hartmann, W., Steiner, J., Klemm, J., Lommel, B. (**2002**) *Nucl. Inst. Methods*, **A 480**, 160–65.

Kindler, B., Hüller, B., Hartmann, W., Klemm, J., Lommel, B., and Steiner, J. (**2004**) *Nucl. Instrum. Methods*, **A 521**, 222–26.

Kirchner, R., Burkhardt, K.H., Bajo, W., and Klepper, O. (**1981**) *Nucl. Instrum. Methods*, **186**, 295–305.

Kobisk, E.H. and Adair, H.L. (**1979**) *Nucl. Instrum. Methods*, **167**, 153–60.

Kobisk, E.H. and Grisham, W.B. (**1969**) *Mat. Res. Bull.*, **4**, 651–62.

Koch, H.R., Riepe, G., Hamacher, A., Pfeiffer, J., Schult, O.W.B., Nakach, A., Olivier, M., Dewald, A., and von Bretano, P. (**1988**) *Nucl. Instrum. Methods*, **A 271**, 375–82.

Krämer, A., Kritzer, A., Reich, H., and Stöhlker, Th. (**2001**) *Nucl. Inst. Methods*, **B 174**, 205–11.

Liebe, D., Eberhardt, K., Hartmann, W., Häger, T., Hübner, A., Kratz, J.V., Kindler, B., Lommel, B., Thörle, P., Schädel, M., and Steiner, J. (**2008**) *Nucl. Inst. Meth. A*, In Press, Accepted Manuscript, Available online 4 March 2008.

Liechtenstein, V.Kh., Ivkova, T.M., Olshanski, E.D., Baranov, A.M., Repnow, R., Hellborg, R., Wellere, R.A., and Wirth, H.L. (**1999**) *Nucl. Instrum. Methods*, **A 438**, 79–85.

Liechtenstein, V.Kh., Ivkova, T.M., Olshanski, E.D., Feigenbaum, I., DiNardo, R., and Dubeli, M. (**1997**) *Nucl. Instrum. Methods*, **A 397**, 140–45.

Lobanov, Yu.V., Buklanov, G.V., Abdullin, F.Sh., Polyakov, A.N., Shirokovsky, I.V., Tsyganov, Yu.S., and Utyonkov, V.K. (**1997**) *Nucl. Instrum. Methods*, **A 397**, 26–29.

Lommel, B., Gembalies-Datz, D., Hartmann, W., Hofmann, S., Kindler, B., Klemm, J., Kojouharova, J., and Steiner, J. (**2002**) *Nucl. Instrum. Methods*, **A 480**, 16–21.

Lommel, B., Hartmann, W., Kindler, B., Klemm, J., and Steiner, J. (**2002**) *Nucl. Instrum. Methods*, **A 480**, 199–203.

Lommel, B., Hartmann, W., Kindler, B., and Steiner, J. (**2006**) *Nucl. Instrum. Methods*, **A 561**, 100–3.

Lozowski, W.R. and Hudson, J.D. (**1991**) *Nucl. Instrum. Methods*, **A 303**, 34–42.

Macri, M. (**1984**) Gas Jet Targets, CERN Accelerator School - Antiprotons for Colliding Beam Facilities, Geneva, Switzerland, 1983, Report CER 84-15 (eds P. Bryant and S. Newman), p. 469.

Maier, H.J. (**1979**) *Nucl. Instrum. Methods*, **167**, 167–70.

Maier, H.J. (**1982**) *Nucl. Instrum. Methods*, **200**, 113–20.

Maier, H.J. (**1989**) *Nucl. Instrum. Methods*, **A 282**, 113–15.

Maier, H.J. (**1991**) *Nucl. Inst. Methods*, **A 303**, 172–81.

Maier, H.J., Friebel, H.U., Frischke, D., and Grossmann, R. (**1989**) *Nucl. Instrum. Methods*, **A 282**, 128–32.

Maier, H.J. and Grossmann, R. (**1987**) *Nucl. Instrum. Methods*, **A 257**, 11–14.

Maier-Komor, P. (**1972**) *Nucl. Instrum. Methods*, **A 102**, 485–86.

Maier-Komor, P. (**1985**) *Nucl. Instrum. Methods*, **A 236**, 539–44.

Maier-Komor, P. (**1989**) *Nucl. Instrum. Methods*, **A 282**, 67–70.

Maier-Komor, P., Dollinger, G., and Hammann, E. (**1991**) *Nucl. Inst. Methods*, **A 303**, 88–93.

Maier-Komor, P., Dollinger, G., and Krücken, R. (**2006**) *Nucl. Inst. Methods*, **A 561**, 4–10.

Maier-Komor, P., Krücken, R., Speidel, K.-H., and Kenn, O. (**2004**) *Nucl. Instrum. Methods*, **A 521**, 17–21.

Mairon, H., Stock, F., Gaul, H.G., Moroz, Z., Povh, B., Shibata, T.A., Toporkov, D., Zapfe, K., Steffens, E., Rathmann, F., Korsch, W., Fick, D., Schiemenz, P., Graw, G., and Haeberli, W. (**1991**) *Status of the FILTEX Atomic Beam Source, Workshop on Polarized Gas Targets for Storage Rings, MPI* (ed. E. Steffens), Heidelberg.

Manente, G. and Blunt, D. (**1995**) *Nucl. Instrum. Methods*, **A 362**, 94–97.

Mansur, L.K. (**2006**) *Nucl. Inst. Methods*, **A 562**, 666–75.

Mansur, L.K. and Haines, J.R. (**2006**) *J. Nucl. Mater.*, **356**, 1–15.

Marcus, C. and Zevenbergen, L.A. (**1999**) *Nucl. Instrum. Methods*, **A 438**, 30–35.

Méens, A., Rossini, I., and Sens, J.C. (**1993**) *Nucl. Instrum. Methods*, **A 334**, 200–2.

Melissinos, A.C. and Olsen, S.L. (**1975**) *Phys. Rep.*, **17**, 77–132.

Meyer, H.O., Rohdjess, A. Berdoz, H., Doskow, J., and Sperisen, F. (**1990**) *Nucl. Instrum. Methods*, **A 295**, 53–64.

Morrissey, D.J., Sherrill, B.M., Steiner, M., Stolz, A., and Wiedenhoever, I. (**2003**) *Nucl. Inst. Methods*, **B 204**, 90–96.

Muggleton, H.A.F. (**1987**) *Vacuum*, **37**, 785–817.

Muggleton, A.H.F. (**1991**) *Nucl. Instrum. Methods*, **A 303**, 157–61.

Münzenberg, G., Heß, F.P., Faust, W., Folger, H., Hofmann, S., Schmidt, K.-H., Schött, H.-J., and Armbruster, P. (**1989**) *Nucl. Instrum. Methods*, **A 282**, 28–35.

Nishimura, E. (**2002**) *Appl. Phys.*, **A 74**, (Suppl) S46–S48.

Nolen, J.A. (**2007**) *Nucl. Phys.*, **A 787**, 84–93.

Nolen, J.A., Reed, C.B., Hassanein, A., and Gomes, I.C. (**2002**) *Nucl. Phys.*, **A 701**, 312c–322c.

Nolen, J.A., Reed, C.B., Hassanein, A., Novick, V.J., Plotkin, P., and Specht, J.R. (**2003**) *Nucl. Instrum. Methods*, **B 204**, 293–97.

Nolen, J.A., Reed, C.B., Hassanein, A., Novick, V.J., Plotkin, P., Specht, J.R., Morrissey, D.J., Ottarson, J.H., Sherrill, B.M. (**2003b**) *Nucl. Instrum. Methods*, **B 204**, 298–302.,

Oelert, W. (ed.) (**1987**) *Workshop on Internal Targets for COSY, KFA*, Jülich, Jül-Spez-409.

Oganessian, Yu.Ts., Demin, A.G., Iljinov, A.S., Tretyakova, S.P., Pleve, A.A.,Penionzhkevich, Yu.E., Ivanov, M.P., and Tretyakov, Yu.P. (**1975**) *Nucl. Phys.*, **A 239**, 157–71.

Oyama, Y. (**2006**) *Nucl. Instrum. and Methods*, **A 562**, 548–52.

Parker, W., Bildstein, H., and Getoff, N. (**1964**) *Nucl. Instrum. Methods*, **26**, 55–60.

Parker, W., Bildstein, H., Getoff, N., Fischer-Colbrie, H., and Regal, H. (**1964**) *Nucl. Instrum. Methods*, **26**, 61–65.

Parker, W. and Falk, R. (**1962**) *Nucl. Instrum. Methods*, **16**, 355–57.

Paunovic, M. and Schlesinger, M. (**2006**) *Fundamentals of Electrochemical Deposition*, 2nd edn, John Wiley & Sons.

Pauwels, J., van Craen, J., van Gestel, J., and van Audenhove, J. (**1979**) *Nucl. Instrum. Methods*, **167**, 109–12.

Pengo, R., Manente, P., Favaron, G., Cecchi, A., Pieraccini, L. (**1991**) *Nucl. Instrum. Methods*, **A 303**, 146–51.

Pollock, R., Miller, D., and Singh, P. (**1980**) The IUCF Cooler-Tripler, Proposal for an Advanced Light-Ion Physics Facility, Indiana University.

Popov, S.G. (**1990**) *Physics and Technology of Superthin Internal Targets in Storage Rings, Topical Conference on Electronuclear Physics with Internal Targets, SLAC, Stanford, California, 1989* (ed. R.G. Arnold), World Scientific, Singapore, p. 37.

Ravn, H.L. The ISOLDE Collaboration and the Neutrino-Factory Working Group (**2003**) *Nucl. Instrum. Methods*, **A 503**, 340–343.

Ravn, H.L. (**2003b**) *Nucl. Instrum. Methods*, **A 503**, 340–43.

Reich, H., Bourgeois, W., Franzle, B., Kritzer, A., and Varentsov, V. (**1997**) *Nuclear Physics*, **A 626**, 417–25.

Roa, D.E., Wright, L., Fox, J.D., Greene, J.P., Back, B.B., and Nardi, B.G. (**1996**) *Nucl. Instrum. Methods*, **A 368**, 307–13.

Schädel, M. (**2006**) *Angew. Chem. Ent. Ed.*, **45**, 368–401.

Sletten, G. and Knudsen, P. (**1972**) *Nucl. Instrum. Methods*, **102**, 459–63.

Spirlet, J.C., Müller, W., and Van Audenhove, J. (**1985**) *Nucl. Instrum. Methods*, **A 236**, 500–2.

Steffens, E. (**1990**) Experiments with dense polarized internal targets, 7th international conference on polarization phenomena in nuclear physics, *Colloque de Physique C6*, Paris.

Steffens, E. (ed.) (**1991**) *Steffens Workshop on Polarized Gas Targets for Storage Rings, MPI*, Heidelberg.

Stodel, Ch., Frémont, G., Auger, G., and Spitaëls, C. (**2006**) *Nucl. Instrum. Methods*, **A 561**, 112–14.

Stolarz, A. (**1997**) *Nucl. Instrum. Methods*, **A 397**, 114–16.

Stolarz, A., and van Gestel, J. (**2006**) *Nucl. Inst. Methods*, **A 561**, 115–19.

Stoner, J.O. Jr (**1997**) *Nucl. Instrum. Methods*, **A 397**, 68–74.

Stoner, J.O. Jr and Jolivet, C.S. (**2006**) *Nucl. Instrum. Methods*, **A 561**, 11–15.

Stoner, J.O. Jr and Miller, S.A. (**2006**) *Nucl. Instrum. Methods*, **A 561**, 24–37.

Stringfellow, B.C., Bujak, A.T., Duane Carmonu, D., Chung, Y.H., Finn, J.E., Gutay, L.J., Hirsch, A.S., Mahi, M., Paderewski, G.L., Porile, N.T., Sangster, T.C., Scharenberg, R.P., and Turkot, F. (**1986**) *Nucl. Instrum. Methods*, **A 251**, 242–253.

Sugai, I. (**1997**) *Nucl. Inst. Methods*, **A 397**, 81–90.

Sugai, I. (**1999**) *Nucl. Inst. Methods*, **A 438**, 58–64.

Sugai, I. and Kato, H. (**1985**) *Nucl. Instrum. Methods*, **A 236**, 549–54.

Sugai, I., Takeda, Y., Oyaizu, M., Kawakami, H., Irie, Y., Arakida, Y., Hara, K., Hattori, H., Kawasaki, K., Kamiya, J., Kinsho, M., and

Kuramochi, K. (**2006**) *Nucl. Instrum. Methods*, **A 561**, 16–23.

Szerypo, J., Friebel, H.U., Frischke, D., Grossmann, R., and Maier, H.J. (**2008**) *Nucl. Instrum. Meth. A*, **590**, 73–75.

Takeuchi, S., Kobayashi, C., Satoh, Y., Yoshida, T., Takekoshi, E., and Maruyama, M. (**1979**) *Nucl. Instrum. Methods*, **158**, 333–38.

Tanabe, T., Sugai, I., and Yasue, M. (**1987**) *Nucl. Instrum. Methods*, **A 256**, 439–43.

Thalheimer, W., Hartmann, W., Klemm, J., and Lommel, B. (**1999**) *Cryst. Res. Technol.*, **34**, 175–79.

Thomas, G.E. and Greene, J.P. (**1993**) *Nucl. Inst. Methods*, **A 334**, 132–36.

Thomas, G.E. and Greene, J.P. (**1995**) *Nucl. Instrum. Methods*, **A 362**, 201–4.

Thomas, G.E., Greene, J.P., Maier-Komor, P., and Leonard, R.H. (**1991**) *Nucl. Instrum. Methods*, **A 303**, 162–64.

Trautmann, N. and Folger, H. (**1989**) *Nucl. Instrum. Methods*, **A 282**, 102–6.

Trostell, B. and Kullander, F. (**1990**) *Hydrogen Pellet Target Developments for CELSIUS, 2nd European Particle Conference, Nice* (ed. P. Marin and P. Mandrillon), Editions Frontiers, Gif-sur-Yvette, p. 1306.

Turkentine, R.B. and Weisser, D.C. (**1997**) *Nucl. Instrum. Methods*, **A 397**, 200–3.

Weidong, W., Zhimei, C., Dangzhong, G., Yong, H., Yongming, Z., and Yongijian, T. (**2002**) *Nucl. Inst. Methods*, **A 480**, 98–103.

Yoshida, A., Morita, K., Morimoto, K., Kaji, D., Kubo, T., Takahashi, Y., and Tanihata, A.I. (**2004**) *Nucl. Instrum. Methods*, **A 521**, 65–71.

Zell, K.O., and Dewald, A. (**1989**) *Nucl. Inst. Methods*, **A 282**, 142–46.

TECHNOLOGY OF POLYMER PROCESSING

JAMES L. WHITE, *Institute of Polymer Engineering, University of Akron, Akron, Ohio, U.S.A.*

Introduction 499
1. **Historical Overview** 499
2. **Properties of Polymeric Materials Affecting Processing** 500
2.1 General 500
2.2 Melt Flow 501
2.3 Orientation and Solidification 502
3. **Processing Operations** 502
3.1 General 502
3.2 Polymer and Compound Preparation 502
3.3 Shaping: Extrusion 504
3.4 Shaping: Post-Extruder Processing 505
3.5 Shaping: Molding 506
3.6 Computer-Aided Engineering 506
4. **Part Design** 508
5. **Future Outlook** 508
Further Reading 508

INTRODUCTION

Polymers including plastics and elastomers are an important part of the world economy. They are widely used not only as final products but also as major components or modules in composite products such as automobiles. One sees fabricated polymer products in roles such as textile fibers, film for packaging and videotape, bottles, refrigerator liners, computer housings, automobile/truck/aircraft tires, automobile bumpers, and gasoline tanks.

It is our intention in this article to describe the development and status of polymer processing technology.

1. HISTORICAL OVERVIEW

Polymers have always been part of commercial and daily life to remotest antiquity in all civilizations. Cotton, wool, silk, leather, wood, and natural rubber are all polymeric materials that have been commercially fabricated into useful products from the dawn of civilization to the present. Some perspective is needed. Generally, the fabrication of textile structures from fibers such as cotton, wool, and silk is not classified as part of polymer processing but as textile processing. The shaping of wood and its conversion into paper is similarly classified as a technology distinct from polymer processing. The beginnings of modern polymer processing are generally considered to be associated with technologies arising from the introduction of natural rubber in large quantities into Europe and the United States from South America after 1815 and the introduction of gutta percha from Malaya after 1840. Processing operations including mastication, mixing, solution coating, and compression molding were practiced in the 1820s. The 1830s saw the invention of the two-roll mill and the four-roll calender by Edwin M. Chaffee. Progress accelerated in the 1840s with the introduction of vulcanization and the invention of injection molding, blow molding, and wire coating, which were applied to rubber and naturally occurring gutta percha. These inventions are associated with Charles Goodyear and the brothers Thomas Hancock and Charles Hancock.

The 1860s saw the introduction of cellulose nitrate, the first man-made polymer, into industry. This led to major advances in molding technology, which are largely associated with John W. Hyatt and his brother Isaiah S. Hyatt. Cellulose nitrate was developed into the first commercial man-made fiber by the Count de Chardonnet by extruding small-diameter filaments of its solution through spinneret dies. In one process developed by Chardonnet, the filaments were coagulated by extruding them into a suitable

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

liquid bath (wet spinning), and in a second they were extruded into an air stream that evaporated the solvents (dry spinning).

The screw extruder was first introduced to the polymer industry in the 1870s. It was first applied to wire coating with gutta percha and subsequently used for extruding profiles from natural rubber compounds. The early screw extruders, like the roll mills and calenders, were steam heated.

The twentieth century saw the development of totally synthetic polymers. Both synthetic rubber and synthetic plastics were developed by chemists by the 1930s. These included synthetic thermosets, such as Leo Baekeland's Bakelite®, and melt-processable thermoplastics including polystyrene and polyvinyl chloride. These two new polymers had much higher melt-processing temperatures than rubber or gutta percha, as did other polymers that soon were developed, including polymethyl methacrylate, polyethylene, polyamide-66, and polyamide-6. The 1930s and early 1940s saw the development of electrically heated screw extruders for these polymers. The late 1930s saw the invention of screw injection molding machines for rubber compounds and the early 1940s for thermoplastics. The first twin-screw extruders were commercially developed in Germany and Italy in the late 1930s and early 1940s.

The post World War II period witnessed the development and expansion of polyolefins. The branched low-density polyethylene of the 1930s was supplemented in the 1950s by linear high-density polyethylene and stereoregular isotactic polypropylene. Polyethylene terephthalates were introduced in the early 1950s. The 1940s and early 1950s also saw the development of rubber-modified or high-impact polystyrene (HIPS) and rubber-modified acrylonitrile–styrene copolymers (ABS resins). The end of the decade saw the introduction of tough transparent polycarbonate.

The period following World War II has also seen great advances in polymer processing technology. Single-layer plastic film, largely produced by a tubular film extrusion using air inflation, has become an important industrial process, largely based upon low-density polyethylene. Multilayer film extrusion using polymer melt streams from three or more extruders to produce film has become increasingly important. Biaxially stretched films are made primarily from polyethylene terephthalate by first extruding film from a slit die onto a chilled roll and from there onto a tentering frame where films are stretched biaxially at controlled rates and temperatures. Modular twin-screw extruders were developed for compounding.

Screw-injection molding has become the most important plastics processing operation in terms of numbers of machines and employment. However, other molding processes have also become important. Thermoplastic bottles are today produced by different molding processes in which injection-molded preforms are inflated with pressurized air and pressed into molds.

2. PROPERTIES OF POLYMERIC MATERIALS AFFECTING PROCESSING

2.1 General

Generally, thermoplastics and elastomers are processed in the molten state. It is first necessary to describe the structural and thermal characteristics of polymers. The thermal transition characteristics are largely independent of molecular weight, when such molecular weights exceed the so-called entanglement transition value (which is needed to achieve desirable mechanical properties). Many important commercial thermoplastics are largely crystalline and exhibit moderately strong melting transitions. These include the following (where the first parentheses contain a commonly used acronym for the polymer and the second parentheses indicate melting temperatures): high-density polyethylene (HDPE) (T_m = 135 °C), isotactic polypropylene (PP) (T_m = 165 °C), polyamide-6 (PA-6) (T_m = 215 °C), polybutylene terephthalate (PBT) (T_m = 220 °C), polyethylene terephthalate (PET) (T_m = 260 °C), polyamide-66 (PA 66) (T_m = 260 °C), poly *p*-phenylene sulfide (PPS) (T_m = 285 °C), and polyether ether ketone (PEEK) (T_m = 340 °C). Other important commercial polymers do not crystallize but rather vitrify into glasses. Here the glass transition temperature T_g is relevant. Important vitrifying thermoplastics include polystyrene (PS) (T_g = 100 °C) and polycarbonate (PC) (T_g = 140 °C).

Crystallizing thermoplastics can be melt processed at temperatures slightly (≥20 °C) above their melting points, but for vitrifying thermoplastics one must be roughly 80 °C or more above their glass transition temperatures.

2.2 Melt Flow

Unlike thermal characteristics, the rheological or melt flow properties of polymers are very dependent upon molecular weight and its distribution. Generally, the viscosity in shear flow is constant at low shear rates and decreases with increasing shear rate. The zero-shear viscosity η_0 of a linear polymer increases with the 3.4 power of molecular weight, above the entanglement transition. For polydisperse polymers the molecular-weight dependence is through the weight-average molecular weight M_w. Specifically,

$$\eta_0 = \mathrm{K} M_w^{3.4}, \tag{1a}$$

where

$$M_w = \frac{\sum w_i M_i}{\sum w_i} = \frac{\sum N_i M_i^2}{\sum N_i M_i}. \tag{1b}$$

Here N_i refers to numbers of molecules and w_i to weights or masses of polymer.

At higher shear rates where the viscosity decreases, it was found later by Vinogradov and Malkin that the reduced shear viscosity $\eta(\dot{\gamma})/\eta_0$ is independent of temperature and may be represented as a product of zero-shear viscosity times shear rate, i.e. $\eta_0\dot{\gamma}$. However, the functionality is sensitive to the breadth of the molecular-weight distribution (as measured by the ratio of different moments of the distribution, M_w/M_n or M_z/M_w):

$$\frac{\eta}{\eta_0} = F(\eta_0\dot{\gamma}, M_w/M_n \text{ or } M_z/M_w), \tag{2a}$$

where M_n and M_z are the number- and weight-average molecular weights,

$$M_n = \frac{\sum N_i M_i}{\sum N_i}, \quad M_z = \frac{\sum N_i M_i^3}{\sum N_i M_i^2}. \tag{2b}$$

At high shear rate, the shear viscosity is a decreasing function of shear rate on a logarithmic plot:

$$\eta = K\dot{\gamma}^{n-1}, \tag{3}$$

where n is a constant called the power-law exponent. Over the entire shear rate range, an expression such as

$$\eta = \frac{\eta_0}{1 + (\eta_0/K)\dot{\gamma}^{1-n}}, \tag{4}$$

would seem more appropriate.

A few words on the experimental determination of shear viscosity are necessary. At high shear rates, i.e. $\dot{\gamma} > 10\ \mathrm{s}^{-1}$, measurements are usually made in a capillary extrusion instrument. At lower shear rates, 0.01 to 1.0 s^{-1}, the cone-plate geometry is generally used. Here the polymer melt is usually tested in a nitrogen atmosphere to protect it from degradation. If lower shear rates are of concern, drag-flow sandwich or cylindrical (Pochettino) instruments are used.

Other rheological responses of polymer melts have been made and used to interpret processing behavior. These include

1. transient shear stresses at the startup of shear flow;
2. normal stresses, a nonlinear tensile stress effect that occurs in polymer melts in shear flow;
3. stress responses to oscillatory shear flow;
4. elongational viscosity arising from uniaxial extension measurements, and
5. stress relaxation.

It must be noted that many industrial polymer materials contain large amounts of fillers such as carbon black (in the case of rubber compounds) and (for thermoplastics) glass fibers and minerals such as talc, mica, and calcium carbonate. The addition of such materials can substantially increase the melt viscosity of the system. For compounds with more than a volume fraction of 0.15 carbon black or other submicron mineral particles, yield values develop. This means that the shear viscosity increases to infinity at a finite shear stress. At stresses below the yield value, there is no flow.

2.3 Orientation and Solidification

When cooled from melt as indicated above, many polymers (polyethylene, polypropylene, polybutylene terephthalate) crystallize. Other polymers (polycarbonate, polystyrene) vitrify into glasses. Some polymers (e.g., polyethylene terephthalate, polyphenylene sulfide) can vitrify or crystallize depending upon rates of cooling and states of stress during processing. The latter type of polymers often undergo two-stage processing, in which they are preshaped in a vitrified form and then undergo a second-stage processing in which they are shaped and crystallized.

Polymers are long-chain molecules. During processing, these chains can orient. Usually such orientation is in directions determined by applied stress fields. Thus if a filament is formed and uniaxially stretched, a state of uniaxial polymer chain orientation is developed along the direction of the filament axis. If a film is formed by biaxial stretching, the applied stresses tend to create a state of biaxial orientation in the plane of the film.

3. PROCESSING OPERATIONS

3.1 General

Following polymerization, polymers generally proceed through a series of processing operations before achieving their final product state. These include

1. post-reactor processing to remove volatiles and to introduce stabilizers;
2. compounding with fillers, reinforcing agents, and colorants to prepare the polymer for its final application;
3. molding or extrusion of the polymer melt to form final form; and
4. post-die processing in extrusion operations wherein extrudates are further shaped, such as inflation of melt emerging from annular dies to form a thin film (tubular film extrusion) or into bottles (extrusion blow molding).

In post-reactor processing, extrusion, and molding operations, there is a common usage of a screw, which occurs in all of the processing machines. Machines for producing profiles, extrusion blow molding, injection molding, post-reactor processing, and continuous compounding all contain screws. Usually, shaping machines involved in profile extrusion or molding operations use single screws. Profile extrusion of polyvinyl chloride for pipe and window profiles is, however, often carried out with intermeshing counter-rotating twin-screw extruders. Continuous compounding and post-reactor processing are usually carried out in modular corotating twin-screw extruders, but counter-rotating twin-screw machines are also used.

The screw principle is basic to polymer processing operations and deserves separate treatment. A screw is a steel shaft with a helical spine winding about its axis (Fig. 1). In a single-screw extruder, as the screw rotates in a barrel, the friction exerted on the material between the screw flights causes it to advance around the helix and move axially toward the die. Single-screw extruders are generally plasticating devices, in which thermoplastics are introduced into the entrance (or hopper) of the extruder and are conveyed toward the die. Barrel heating and pressurization leads to the initiation of melting, which is greatly enhanced by viscous dissipation of the highly viscous polymer as it is forced to flow. Generally, screws for single-screw extruders have deep channels in the feed/hopper region and tapered screw root diameters in the melting region. In the region nearest the die, known as the metering region, the screw channel is shallow. Often to enhance extruder productivity, the barrel in the solid conveying region is grooved to increase friction and pressure development.

Turning to twin-screw extruders (Fig. 2), self-wiping corotating twin-screw extruders and tangential counter-rotating twin-screw extruders are also drag-flow pumping machines. Fully intermeshing counter-rotating twin-screw machines are positive-displacement pumps.

3.2 Polymer and Compound Preparation

The first steps following polymerization, as described earlier, are post-reactor processing, carried out by a resin manufacturer, and compounding that may be carried out by either a resin manufacturer or a separate company usually called a compounder. The two processes are similar and are usually carried out in twin-screw extruders or in other complex mixing machines such as the Buss

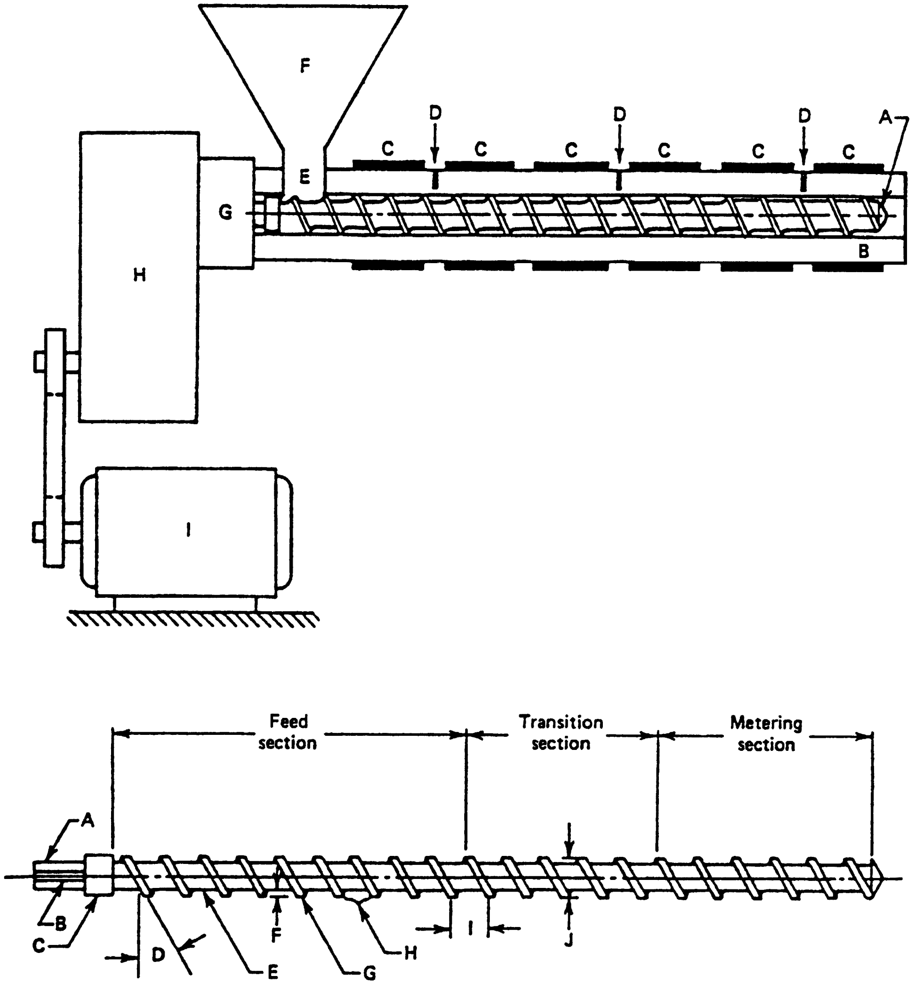

FIG. 1. (top) Parts of a single-screw extruder. A, screw; B, barrel; C, heater; D, thermocouple; E, feed throat; F, hopper; G, thrust bearing; H, gear reducer; I, motor. Courtesy of Society of Plastics Engineers. (bottom) Extruder screw. A, shank; B, key; C, hub; D, helix angle; E, root; F, channel depth; G, flight; H, channel; I, pitch; J, diameter; K, tip.

Kneader in which an interrupted screw oscillates axially and rotates within a barrel containing inwardly projecting teeth.

Most important are modular self-wiping corotating twin-screw extruders (Fig. 3). These machines have modular screws and a modular barrel construction. The machines have two parallel shafts with a fixed distance between them. Pairs of identical elements are placed on the two shafts until the shafts are totally filled. These elements are right-handed (forward pumping) screws, left-handed (backward pumping) screws, and various types of mixing elements, notably kneading discs. Twin-screw extruders generally have multiple feed parts so that polymers, stabilizers, and reinforcing agents may be introduced into different feed parts.

Modular intermeshing counter-rotating and tangential counter-rotating twin-screw extruders are also used for compounding. The intermeshing machines exhibit high levels of shear stress and are very good dispersive mixers for pigments. Tangential counter-rotating twin-screw extruders have large volumes for a given screw diameter and are often conveniently used for devolatilization and reactive extrusion.

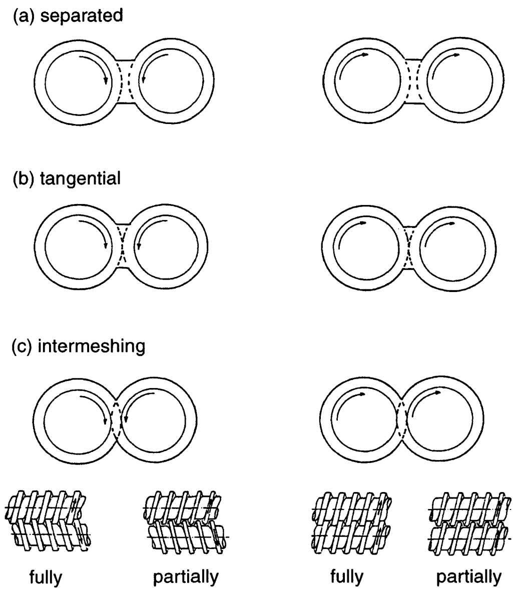
(a) separated
(b) tangential
(c) intermeshing
fully
partially
fully
partially

FIG. 2. Twin-screw extruders.

3.3 Shaping: Extrusion

One of the most important shaping operations is the manufacture of continuous profiles, pipe, and tube from thermoplastics. Either a single-screw or a fully intermeshed counter-rotating twin-screw extruder is used to plasticate the polymer and generate the pressure required for extrusion through the shaping die. The dies used to form profiles must be specifically designed to ensure uniform distribution of the melt to the die so that uniform extrudates (or extrudates of desired design) are obtained.

We have seen increasingly through the years the development of multilayer extrusion, where the feeds of polymer melts from several different extruders are combined in a die. The most important product produced in this way is multilayer film, which is widely used for packaging. Generally at least five melt layers are required in such a process. There are two outer layers on the surface of the film or sheet, a core layer (usually a barrier material) and two adhesive layers

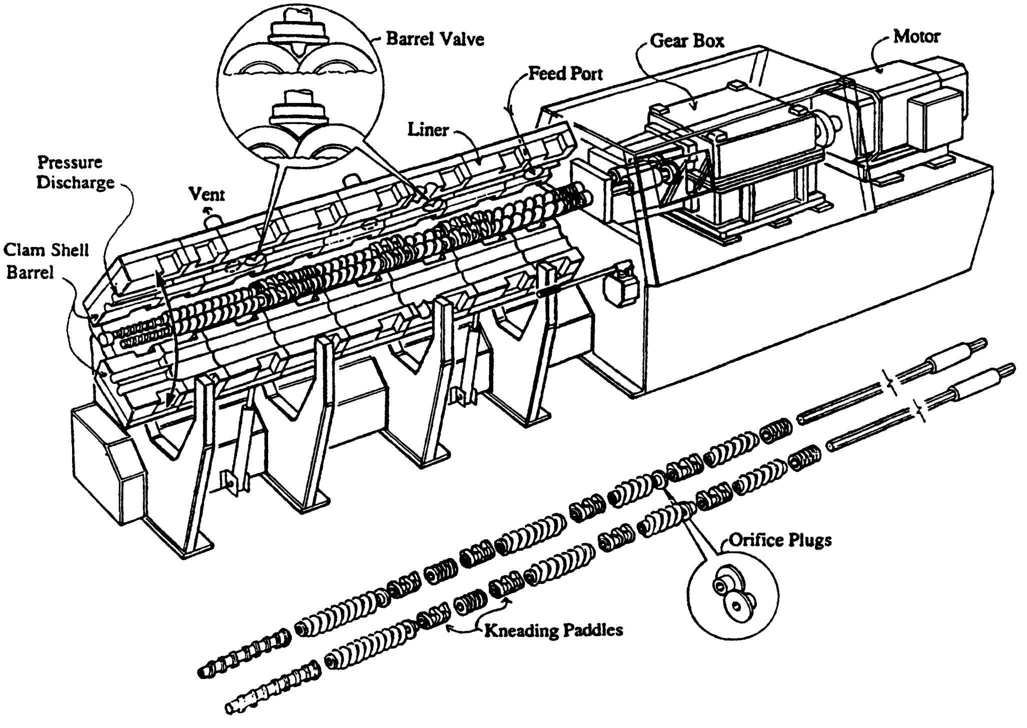
Barrel Valve
Gear Box
Motor
Feed Port
Liner
Pressure Discharge
Vent
Clam Shell Barrel
Orifice Plugs
Kneading Paddles

FIG. 3. Modular co-rotating twin-screw extruder.

that bind the external film surfaces to the core. The five melt layers may be seen to involve these materials: core, adhesive, and outer layer.

3.4 Shaping: Post-Extruder Processing

In some important processing operations, extrudates continuously emerging from dies are further shaped by applied tensions and pressures. Thus, extrudates emerging from dies are drawn down by applied tensions to form greatly reduced cross sections (Fig. 4). This is notably the case in melt spinning of molten polymers such as poly(ethylene terephthalate), polyamide-6 (and -66), and polypropylene to form continuous filaments or fibers. The take-up velocities today are quite large, being many thousands of meters per minute. Associated with the high velocities are high draw-down ratios and high spinline tensions. Generally the dies have a very large number of small orifices so that numerous filaments—a yarn—and not a monofilament are produced. The design of such dies, known as spinnerettes, is quite complex because of the problem of uniformly distributing polymer melt.

A second polymer processing operation that involves post-extruder processing involves the manufacture of film by the tubular blowing process (Fig. 5). Generally a thin annular tube is extruded vertically upward. This tube is both inflated with pressurized air and drawn down by applied tensions. The tubular film process is most widely used with low-density polyethylene.

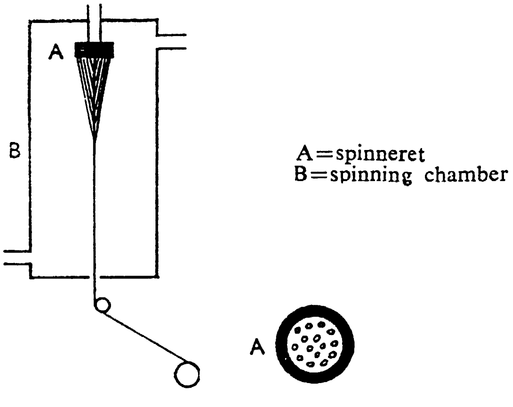

FIG. 4. Melt spinning of filaments.

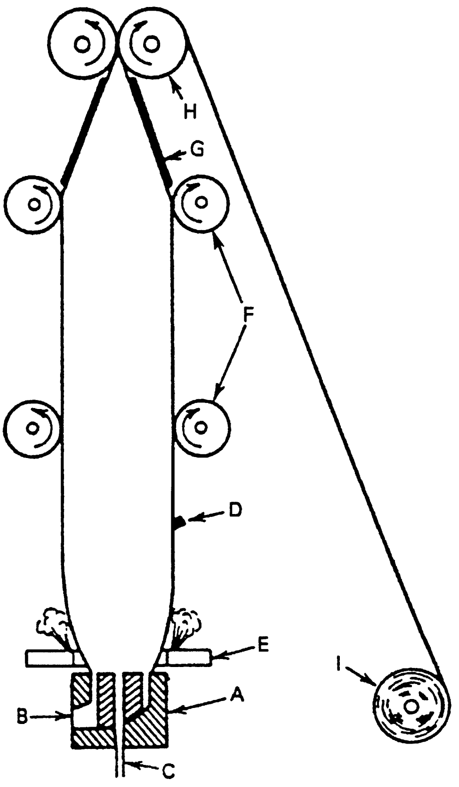

FIG. 5. Tubular film process for extrusion of blown film. A, blown-film die; B, die inlet; C, air hole and valve; D, plastic tube (bubble); E, air ring for cooling; F, guide rolls; G, collapsing frame; H, pull rolls; I, windup roll.

In addition to the tubular film process described in the above paragraph, double-bubble tubular film processes have been developed and used commercially. The first of the two bubbles is used to shape the film. The second of the two bubbles biaxially stretches the film in a controlled environment, generally both developing biaxial orientation and crystallizing the melt. This process was first devised for polyvinylidene chloride by Dow Chemical but has subsequently been used notably for polypropylene and polyamide-6.

Film may also be produced by slit dies that extrude vertically downward onto rotating, chilled rolls. It is possible to convert such "cast film" to biaxially oriented film by means of a tentering frame. This apparatus mechanically grasps the edges of film during passage through a controlled-environment oven. The film is usually stretched uniaxially in different directions by stages.

3.5 Shaping: Molding

The most important of polymer processing operations is the manufacture of discrete thermoplastic, rubber, and thermoset parts. This general technology is called molding. There are many variations of molding technology.

The simplest type of molding technology is compression molding, where the polymeric material is placed in an open mold. This technology is generally applied to elastomers and thermoset materials. The upper (male) portion of the mold is then moved into place. The mold is heated and the material placed into the cavity softens and flows out radially under the pressure of the moving member to fill the mold. This flow is usually considered shearing in character. Once the mold is filled, the polymer solidifies through chemical cross-linking for elastomers and thermosets. For thermoplastics, solidification requires cooling the mold. When the polymer is solidified, the mold is opened, the part removed, and the process repeated.

In injection molding, which is primarily applied to thermoplastics, a polymer is melted and pressured, usually with a rotating screw, and then injected into a mold by the reciprocating motion of the screw (Fig. 6). In thermoplastic injection molding, the mold is cold and solidifies the melt filling it. In injection molding of rubber compounds, the mold is maintained hot, and the rubber is vulcanized when it fills the mold. After solidification the mold is then opened and the molded part removed. The process is then repeated.

In rotational molding, a thermoplastic powder is introduced into a mold. The mold is then rotated biaxially in an oven. The powder melts and coats the surface of the mold. The mold is then moved to a quench chamber where it is cooled by water shower or mist. The mold is subsequently opened and the molded part removed. Generally, rotational molding produces hollow parts. It has a longer production cycle than blow molding, but the equipment is simpler and easier to construct.

In blow molding (Fig. 7), a screw extruder and die are used to produce an annular polymer melt sheet, known as a parison, which is extruded vertically downward. A mold of a bottle then closes around the parison, and air is injected into the melt parison until the melt fills the mold. The mold is cooled, the melt solidifies, the mold opens, and the fabricated bottle is ejected.

Stretch-blow molding is a two-stage molding process developed originally for polyethylene terephthalate. Tubelike preforms are first injection molded and extruded. These are noncrystalline glassy in character. The preforms are introduced into the stretch molding machine where the temperature is raised above T_g. The preform is then inflated (often with the aid of a moving mandrel) to fill the mold. During the mold filling, the inflating preform polymer develops biaxial orientation and crystallizes.

3.6 Computer-Aided Engineering

It is possible to model mathematically to varying extents the various polymer processing operations. In many cases such modeling

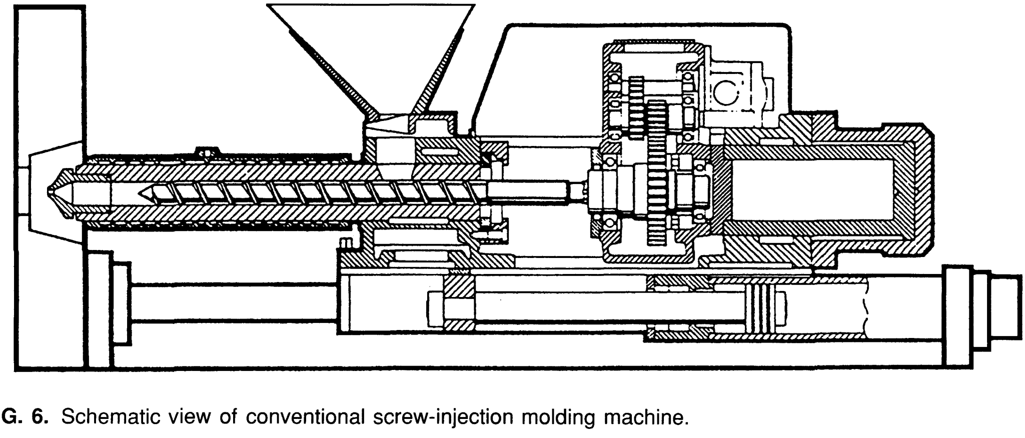

FIG. 6. Schematic view of conventional screw-injection molding machine.

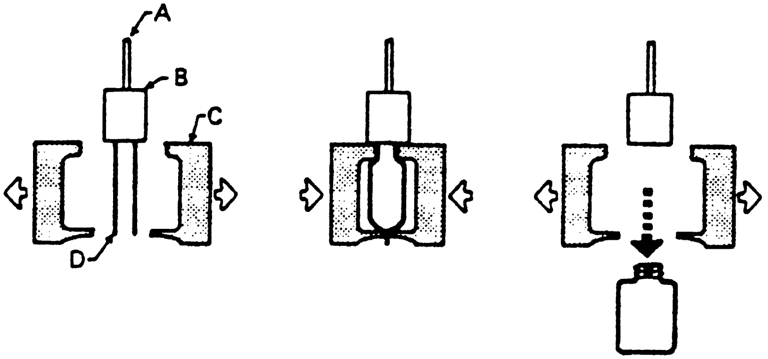

FIG. 7. Blow molding. A, air line; B, die; C, mold; D, parison.

is highly useful as it can provide information on the detailed behavior of melt flows in molds, dies, and extruder screws. Many of these models have been converted into user-friendly computer software, which is marketed by firms all over the world.

Probably the earliest activities of this type involved melt spinning of synthetic fibers and were the result of activities of Kase, Matsuo, and their co-workers with Toyobo in Ohtsu, Japan. At about the same time, Tadmor and Klein with Western Electric and later Scientific Process and Research developed simulations and software for flow in plasticating screw extruders.

More successful than these activities were modeling efforts of the late 1970s and 1980s directed toward injection-mold filling and later packing and residual-stress effects. In the 1970s, various researchers presented simulations of mold filling. Colin Austin founded the firm Moldflow-Australia. In later years, extensive activity has been given to simulation of flow in dies, blow molding, and other processing operations. The development of special finite-element packages such as "FIDAP" and "Polyflow" have allowed individual engineers to do calculations. These have been applied to problems ranging from die design to blow molding. Modular corotating twin-screw extruders have been simulated in the 1980s and 1990s, and software is available from The University of Akron and the University of Paderborn.

Such simulations are based upon solving the basic force-balance, energy-balance, and continuity equations of continuous mechanics together with suitable constitutive equations for the deforming polymer melt. Constant density is usually assumed. As the viscosity levels of polymer melts are very high, one usually neglects inertial force terms. The basic balance equations are force balance,

$$0 = \nabla \cdot \boldsymbol{\sigma}; \tag{5}$$

energy balance,

$$\rho c\left[\frac{\partial T}{\partial t} + (\mathbf{v}\cdot\nabla)T\right] = k\nabla^2 T + \sum_i \sum_j \sigma_{ij}\frac{\partial v_i}{\partial x_j}; \tag{6}$$

and continuity,

$$\nabla \cdot \mathbf{v} = 0, \tag{7}$$

where $\boldsymbol{\sigma}$ is stress, T temperature, $\mathbf{v}$ velocity, and k thermal conductivity. The constitutive equations used usually have the form

$$\boldsymbol{\sigma} = p\mathbf{I} + 2\eta(\operatorname{tr}\mathbf{d}^2, T)\mathbf{d}, \tag{8}$$

where $\mathbf{d}$ is the rate-of-deformation tensor, p is pressure, and tr $\mathbf{d}^2$ is the sum of the diagonal components of the matrix $\mathbf{d}^2$. Equation (8) represents polymer melts as viscous non-Newtonian fluids. There have also been efforts to include viscoelastic effects in constitutive equations to solve processing problems. However, no consensus exists on how to do this.

Most analyses of processing equipment, especially for dies, injection molds, and extruder screw channels, use hydrodynamic lubrication theory in which Eqs. (5) and (6) are simplified to force balance,

$$\begin{cases} 0 = -\dfrac{\partial p}{\partial x_1} + \dfrac{\partial}{\partial x_2} \eta \dfrac{\partial v_1}{\partial x_2}, & \text{(9a)} \\ 0 = -\dfrac{\partial p}{\partial x_3} + \dfrac{\partial}{\partial x_2} \eta \dfrac{\partial v_3}{\partial x_2}, & \text{(9b)} \end{cases}$$

and energy balance,

$$\rho c\left[\frac{\partial T}{\partial t} + v_1 \frac{\partial T}{\partial x_1}\right] = k \frac{\partial^2 T}{\partial x_2^2} + \eta\left[\left(\frac{\partial v_1}{\partial x_2}\right)^2 + \left(\frac{\partial v_3}{\partial x_2}\right)^2\right], \tag{10}$$

and the shear viscosity η is considered to be nonisothermal and non-Newtonian, depending upon T and tr $\mathbf{d}^2$ through:

$$\eta = F\left(\left(\frac{\partial v_1}{\partial x_2}\right)^2 + \left(\frac{\partial v_3}{\partial x_2}\right)^2\right) \exp[-B(T - T_0)], \tag{11}$$

where the functionality might be that of Eqs. (3) and (4). Such simulations may be used to design injection molds by considering the influence of the design of cooling channels in molds on melt filling.

In more recent years, mold design methods (including computer software) have allowed the calculation or better estimation of residual stresses and warpage of parts removed from molds. Such developments allow the problems of part design and mold design to be treated.

4. PART DESIGN

Thermoplastic and thermoset molded parts developed a "bad name" in the early years of the polymer industry. This was because parts were badly designed, with designers generally not recognizing that the polymers they were dealing with had lower Young's moduli and failure stresses than the metals they were replacing.

It is desirable to use design procedures that take into consideration not only the mechanical properties of the polymeric materials used but the thickness of sections and cross-sectional moments of inertia. While correct design procedures based on the theory of elasticity have always been available during the life of the polymer industry, they have frequently been ignored. It is only in recent years that computer-aided design procedures have been developed to accomplish this and attention has been given by engineers to these problems.

5. FUTURE OUTLOOK

Any view to the future of polymer processing technology must envisage the manufacture of more complex polymeric parts, which will be produced with greater precision. Greater complexity will involve on the material side sophisticated polymer structures, and better controlled molecular-weight distributions and types of branching. On the machine side, one will see processing machinery, especially injection molding machines, that will be able to produce "engineering parts" with more precise dimensions and with multiple components.

To a growing extent the future is already with us. The newer-generation Ziegler–Natta catalysts known as "metallocenes" are producing polyolefins with more controlled structures. Injection-molding technologies exist that produce engineering parts with more precise dimensions have been developed. This is accomplished by greater control of the *PVT* characteristics during the cooling cycle to limit shrinkage and warpage. In "gas-assist injection molding," hollow molded parts with carefully located void regions are developed. This allows the development of parts containing not only less material but lower residual stress. More industrially molded products contain inner layers that give greater barrier protection in packaging or contain recycled materials. Molding technologies exist where different thermoplastics and elastomers (usually thermoplastic elastomers) may be combined together in a single mold to give parts with both hard and soft sections.

Further Reading

Chen, Z., White, J. L. (1994), *Int. Polym. Proc.* **9,** 310–318.

Isayev, A. I. (Ed.) (1987), *Injection and Compression Molding Fundamentals,* New York: Marcel Dekker.

Kase, S., Matsuo, T. (1965), *J. Polym. Sci. A* **3,** 2541–2554.

Kase, S., Matsuo, T. (1967), *J. Appl. Polym. Sci.* **11,** 251–287.

Michaeli, W. (1991), *Extrusion Dies for Plastics and Rubber,* 2nd ed., Munich: Hanser.

Potente, H., Ansahl, J., Klarholz, B. (1994), *Int. Polym. Proc.* **9,** 11–25.

Rauwendaal, C. (1985), *Polymer Extrusion,* Munich: Hanser.

Tadmor, Z., Gogos, C. G. (1979), *Principles of Polymer Processing,* New York: Wiley.

Tadmor, Z., Klein, I. (1969), *Engineering Principles of Plasticating Extrusion,* New York: Van Nostrand Reinhold.

Wang, Y., White, J. L., Szydlowski, W. (1989), *Int. Polym. Proc.* **4,** 262–269.

White, J. L. (1990), *Twin Screw Extrusion: Technology and Principles,* Munich: Hanser.

White, J. L. (1995), *Rubber Processing: Technology, Materials and Principles,* Munich: Hanser.

Ziabicki, A. (1977), *Fundamentals of Fibre Formation,* New York: Wiley.

TELECOMMUNICATIONS

LEONARD BERGSTEIN AND RICHARD VAN SLYKE, *Center for Advanced Technology in Telecommunications, Polytechnic University, Brooklyn, New York, U.S.A.*

	Introduction	511
1.	**Applications**	513
2.	**Digital and Analog Transmission**	514
2.1	Analog Encoding of Digital Information	514
2.2	Digital Encoding of Analog Signals	515
2.3	Digital Encoding of Digital Data	516
2.4	Analog Encoding of Analog Information	516
3.	**Efficient Use of Channels**	516
3.1	Multiplexing	517
3.2	Concentration, Statistical Multiplexing	518
3.3	Compression, Error Control, and Encryption	519
3.3.1	Compression	519
3.3.1.1	Lossless Compression	519
3.3.1.2	Lossy Compression	520
3.3.1.3	Uses of Compression in Telecommunications	520
3.3.2	Error Detection and Correction	520
3.3.3	Encryption	521
4.	**Efficient Use of Networks**	522
4.1	Circuit, Message, and Packet Switching	522
4.2	Architectures, Standards, and Protocols	523
5.	**Transmission Media**	523
5.1	Wireless	523
5.1.1	Limitations on Wireless Communications	523
5.1.1.1	Free-Space Loss	523
5.1.1.2	Absorption	524
5.1.1.3	Multipath	524
5.1.1.4	Refraction	524
5.1.1.5	Thermal Noise	524
5.1.2	Satellite Communication	524
5.1.3	Point-to-Point Wireless Communication	527
5.1.4	Mobile Telephony	528
5.2	Constrained Media	529
5.2.1	Limitations on Constrained Media	529
5.2.1.1	Attenuation	529
5.2.1.2	Noise	530
5.2.1.3	Delay	530
5.2.1.4	Distortion	530
5.2.2	Twisted-Pair Cables	530
5.2.3	Coaxial Cable	530
5.2.4	Optical-Fiber Communications	531
6.	**The Telecommunications Industry and Its Future**	535
	Glossary	535
	Works Cited	537
	Further Reading	537

INTRODUCTION

"Telecommunications" means communication at a distance. While pigeons, human couriers, written messages, and sound systems of various sorts have been used for distance communications for millennia, the vehicle of rapid, long-distance communication is almost exclusively electromagnetic, ranging from very low-frequency radio for submarine communication to light in optical fibers.

We can do worse than start with Shannon. Figure 1, based on the famous figure by Shannon (1948), is a schematic communication model for a source seeking to communicate a message, *M*, to a destination over a communication channel. The rest of this article consists of elaborations or generalizations of this simple model.

To fix ideas, let us consider an example of a message being left in a voice-mail system. The originator of the message might speak the message into a microphone. The micro-

digital Encyclopedia of Applied Physics

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

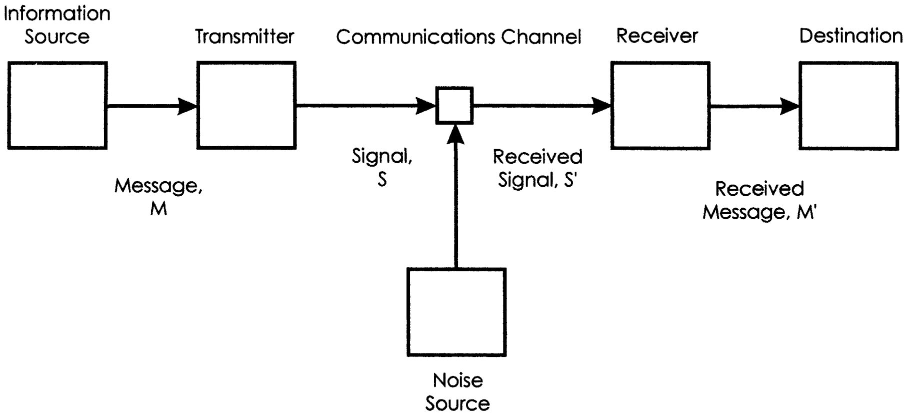

FIG. 1. Schematic diagram of a general communication system.

phone converts the acoustic waves of the speech into an electrical analog *message*. We will consider the acoustic message to be external to our system so that the *information source* is the microphone and the *message* is the electrical wave form from the microphone. Quite often, the message is converted from the analog electrical message into a digital form at the transmitter. The resulting digital signal, *S*, is sent over a communication channel. At the other end of the channel, the receiver, *R*, receives a signal, *S'*, which, it is hoped, represents faithfully the sent signal *S*. This will not always be the case because of possible noise in the system. The received signal is then decoded into a received message, *M'*, which, in this case, is still digital so that the message can be stored in the computer for later retrieval.

To preview the ensuing discussion let us consider the process illustrated by Fig. 1 in more detail. The first issue is how to characterize the "message". Some of the aspects that need to be addressed is how much communication capacity various types of messages require and how the communication requirements vary in time. These are addressed in Sec. 1. One can also say much about the conversion from the initial message *M* to the transmitted signal *S*. There are, potentially, three objectives in this transformation. The message may be *compressed* in order to use less channel capacity; on the other hand, additional redundant information may be added to enhance the ability to detect errors and, in some cases, even to correct them. Finally, the signal is converted to a form that is appropriate to the channel being used. The first two points are addressed in Sec. 3.3; the last point, in Sec. 2. The "channel" may be a simple point-to-point channel such as a wire pair or optical fiber, it may be a broadcast transmission facility such as a radio or satellite channel, or it may be derived from a complex network of facilities shared with other users. In either case, various techniques are used to gain efficiencies. We discuss the techniques that are applied to individual channels in Sec. 3 and the techniques that apply to networks of shared facilities in Sec. 4. The physical properties of the basic transmission facilities are then treated in Sec. 5. "Noise" in this simple model represents all the various phenomena that can interfere with communication. This includes noise in the narrow sense, such as thermal noise, as well as echos, distortions of various sorts, and even deliberate interference such as jamming in military applications. Despite the form of Fig. 1, noise and other transmission impairments can arise in the receivers and transmitters as well as the channel. For example, in space communication, receiver noise is very important. Issues dealing with communication impairments for wireless channels are addressed in Sec. 5.1.1, and for constrained media, in Sec. 5.2.1.

We end our article with Sec. 6, which provides a few remarks on the future of telecommunications technology.

1. APPLICATIONS

Information can be analog or digital. Communication capacity is measured in terms of bandwidth for analog communication and by data rate for digital communication. A simple relation between the two was given by Shannon (1948):

$$C = W \log_2(1 + S/N),$$

where C is the highest data rate that can be approached on a channel with bandwidth W with signal power S and noise power N (assumed to be white).

The message or information that is sent in communication systems can be characterized as one of four categories or *services:* voice, data, image, or video. It is useful to have a rough idea of the relative magnitudes of the communication capacity required by the four basic services. Since most modern telecommunications systems deal with information in digital form, we give our estimates in that form, even for information such as voice, which is intrinsically analog.

The most common example of data services is the communication of text, and numerical information. Typically, in a communication system, an individual alphanumeric character is represented by a byte or an octet (8 bits). So, for example, representing an A4 size, single-spaced page of text would require about 3500 octets.

Images can basically be represented digitally in two ways. The most straightforward way is by a *bit map.* The image to be communicated is divided into a rectangular grid. The intersections on the grid are represented by pixels (picture elements). Associated with each pixel is a color and/or intensity value. For black and white high contrast images, the "value" consists of only one bit representing black or white. A black and white bit map with a gray scale, using, say, b bits per pixel can be used to represent 2^b values ranging from white through gray to black. For color, a variety of schemes can be used to represent the color at the pixel. One straightforward way is to assign three equal-sized groups of bits, one for each of three primary colors to characterize the color as a mix of the three primary colors.

Consider a typical laser printer that might print 300 pixels per inch in the vertical and horizontal directions or 90 000 pixels per square inch. To represent the same A4 sheet of paper as before, except this time as a bit-mapped image, would require about 7 500 000 bits. In general, however, there is much redundancy in an image that can be taken advantage of in communication. For example, often in documents there are large areas of white space that can be summarized by coding rather than by sending large numbers of identical pixel representations. Moreover, there are more efficient ways to *send* the image to the printer than by a bit map. This brings us to the second representation of images. Rather than specifying the image at each of the pixels, we can give instructions on how to draw the image. For many types of images, for example, Fig. 1 of this article, these instructions can be encoded to communicate the image much more efficiently than by using bit maps. To print, the laser printer then carries out the "instructions" it receives to obtain a bit map for printing. In either case, the number of bits required to represent an image is typically many times that required by a data message representing the same information.

Voice is another important class of communication application or service (we also often include in this category communication of music and other sounds). Voice is an intrinsically analog source, but in modern systems it is typically communicated by digital signals. As with images there are two approaches to representing voice digitally: a conceptually simple scheme that is relatively expensive of channel capacity and a more complicated approach that can be used to reduce communication requirements. The analog of a bit map for images is simply to sample the voice (at a rate higher than twice the bandwidth to avoid aliasing), discretize the resulting samples, and then represent each sample's amplitude by a given number of bits. This is called *pulse code modulation* (PCM) and is commonly used for telephone voice. For telephone quality voice, one commonly assumes a bandwidth of less than 4000 Hz. The voice signal is sampled at 8000 times a second, and the signal amplitude of each sample is represented by an 8-bit octet. This results in a source rate of 64 000 bits per second (bps). As another example, sound on a music compact disc (CD) is sampled at 44 100 times per second, and 16 bits are used to represent the amplitude of each sample. The sound is stereo, and so we have two such sources of bits. Thus the total rate is 2 × 44 100 × 16 = 1 411 200 bps. Telephone-quality voice can be communicated using data rates that are $\frac{1}{2}$ or $\frac{1}{4}$ or less of the source rate of 64 000 bps by use of various compression schemes. Even lower rates can be used if a different strategy that takes into account the specific nature of the vocal tract is used. Such compression schemes—for example, linear predictive coding—are used to make efficient use of the broadcast spectrum for digital mobile telephony. By means of such techniques, recognizable, if relatively poor quality, speech can be communicated at rates in the neighborhood of 2400 bps.

For video, there is again a straightforward form and a more complicated yet efficient form. The straightforward form is to simply send a number of "bit maps" per second (in the range from 26 to 30 per second). The required communication capacity is formidable even using high-capacity fiber-optic transmission facilities. To be practicable, substantial compression is needed. The compression process is aided by the fact that, in addition to the geometric redundancy inherited from the individual bit-map images, there is also temporal redundancy in the sense that many of the successive images share much. Nevertheless, the compression required is so extreme that, typically, lossy compression is needed; see Sec. 5.3.1.2. A potentially more efficient form of video is to send drawings rather than bit maps. This form is usually called *animation.*

The variation in time of the arrival of bits generated by a message or sequence of messages is also important. For some types of communication the arrival of bits from the source is relatively constant, or *smooth*—for example, the sending of long files, or the sending of uncompressed digital voice. Other types of traffic are much more erratic, or *bursty.* For example, traffic generated by a human-operated terminal can be quite erratic with long (in electronic terms) periods of inactivity while the terminal operator ponders his/her next operation, followed by bursts of activity, say when the terminal receives a bit-mapped image. We can characterize the *burstiness* of traffic in various ways; one convenient way is in terms of the ratio of the peak data rate of the source to its mean rate. As we shall see, bursty traffic is best carried with completely different system architectures than for nonbursty traffic.

Messages in different applications often have differing fidelity measures or figures of merit. An important distinction here is the difference between real-time traffic and non-real-time traffic. Examples of real-time traffic include voice conversations and video conferencing. Such applications are sensitive to delay and, even more importantly, to variations in delay (jitter). The accuracy and even the arrival of individual message elements is relatively less important. On the other hand, for non-real-time applications such as data communication, accuracy is often paramount, so much so that if the message does not arrive correctly the first time, the relevant message is retransmitted. Real-time traffic often does not permit this luxury.

2. DIGITAL AND ANALOG TRANSMISSION

Both information and the representation of the information for transmission as signals can separately be digital or analog. This allows four possible combinations that we consider in turn.

We can encode digital information as an analog signal, we can digitally encode analog signals, digitally encode digital data, and encode in analog form analog information.

2.1 Analog Encoding of Digital Information

Analog encoding of digital information is called *modulation.* The device that performs the coding and decoding is called a *modem* (modulator-demodulator). The basic idea is to modulate or perturb a *carrier* sine wave. This can be done by varying each of the three defining parameters of the sine wave: the amplitude, the frequency, and the phase. The results are, respectively, *amplitude-shift modulation, frequency-shift keying,* and *phase-shift keying.*

Each change of the amplitude, frequency, or phase, respectively, is a signal element. For example, the phase may be shifted to one of four different values. Thus the signal element can have four realizations or 2 bits per signaling element. One should carefully distinguish between the *signaling* rate, which is measured in baud, and the data rate, which is measured in bits per second. So in the previous example, if the signaling is at 1000 baud, the data rate would be 2000 bps because each signaling element corresponds to 2 bits.

Notice that, for virtually all wireless communication of digital data, an analog encoding must be used since transmission can only take place at relatively high frequencies (using a carrier) compared to most digital information signaling rates. Similarly, communication of digital information over a fiber-optic link is usually amplitude-shift keying because the basic frequencies of light, around 10^{14} Hz, are much higher than any information data rates used today. Otherwise, amplitude-shift keying is rarely used except in combination with other modes because it is rather inefficient and is sensitive to sudden gain changes that commonly occur.

Combinations of amplitude-, frequency-, and phase-shift keying can be used as well. This approach is common for all but the lowest-performance modems. A combination of amplitude- and phase-shift keying is common and is called *quadrature amplitude modulation* (QAM). The *signaling constellation* or group of possible signal realizations can be plotted on a polar graph where the angle represents the phase shift and the radius, the amplitude of the realization. In practice, a signaling scheme with 16 possibilities (4 bits) might be used with a signaling rate of 2400 baud to produce a 9600-bps modem.

2.2 Digital Encoding of Analog Signals

We have discussed previously several advantages of digital transmission over analog. With the advent of the microprocessor in the late 1960s and early 1970s, digital transmission could be cost-effectively implemented. Some of the advantages of digital transmission are these:

1. The use of regeneration rather than amplification reduces the propagation of noise on the signal path.
2. Digital signals can be much more effectively routed, encrypted, and compressed.
3. More effective error control procedures are possible—e.g., cyclic redundancy checksums (CRCs).
4. Digital signals can be more effectively multiplexed [time-division multiplexing (TDM) is more efficient than frequency-division multiplexing (FDM)].

Thus since the early 1970s the transmission systems of choice have been digital. This makes it necessary to convert voice information and other analog information to digital signals. The most common method is pulse code modulation (PCM). The device that accomplishes this conversion is called a *codec* (coder-decoder). The conversion from analog information to digital signal by this method is a two-part process.

We assume that the information signal has a bandwidth of W. For telephone voice, W is between 3 and 4 kHz. For compact discs, $W = 22\,050$ Hz. The Nyquist sampling theorem tells us that if we sample an analog signal with bandwidth W at a rate of $2W$ or more samples per second, then the original signal can be reproduced exactly from the samples. Note that after this process, the samples are still analog; that is, they can take on a continuum of values rather than a discrete number. This part of the conversion is reversible with complete accuracy. Sometimes this sampling is used as a multiplexing method by itself without the second part. This method is called pulse amplitude multiplexing (PAM); for example, it was used in the Dimension line of PBX's by AT&T.

To make the analog samples digital, they are discretized to the closest of a number n of amplitudes. The number n is usually of the form 2^k with k (in bits) $= 8$ ($n = 256$) for telephone voice, and $k = 16$ ($n = 65\,536$) for CDs. Other values are used for specialized applications. This process is *not* reversible. The error resulting from the irreversibility of the process is called discretization noise.

Thus, the standard process for voice PCM is to assume (conservatively) that $W = 4000$ Hz, and to sample at 8000 samples per second. Then discretizing to 65 536 levels (8 bits) results in a digital signal of 64 000 bps. (Sometimes the least significant bit of the amplitude discretization is "stolen" for use in sending control information. In this case, the data rate is effectively 56 000 bps.) Other discretization methods can be used that take advantage of the fact that the differences between amplitude samples are usually less than the values of the sample themselves. Such methods, called differential encodings, can be used to carry a telephone voice message at 32 000 bps, easily, and with some loss of quality at 16 000 bps.

2.3 Digital Encoding of Digital Data

It would seem in this case that we would simply send the information down the channel. For example, one could simply use one nonzero voltage on a line to represent a binary 1 and another nonzero voltage to represent binary 0. This method is called *non-return-to-zero-level* (NRZ-L). This method is rarely used beyond a few feet; it has significant practical problems. The most significant is that it is difficult to synchronize the decoder with the encoder; that is, it is difficult for the receiver to determine when a bit begins in the signal. In most signaling systems the synchronization between sender and receiver is derived from the signal itself. If a long sequence of all ones or all zeros is sent, one can easily see that deriving the synchronization information is difficult. Moreover, when long sequences of ones or zeros are sent, the signal can have a significant dc component, which is often undesirable.

These problems can be alleviated by using a *biphase method* where a transition is forced each bit time so that synchronization information is always available. Two common biphase methods, used, for example, in local-area networks, are *Manchester* and *Differential Manchester.* Each bit time has average 0, so there is no dc component; there is also always a voltage transition each bit time, to aid synchronization. In the Manchester code a zero is recognized by, say, a transition from high to low in the middle of the bit, while a transition from low to high indicates a binary one. In the Differential Manchester scheme the middle transition has no meaning; it only provides synchronization. The code is whether the signal level changes at the beginning of the bit time or not. For example, a zero could be represented by a transition; a one, by no transition. Note that this scheme is insensitive to polarity of the signal, which is a convenience. Both these schemes have the disadvantage that their fundamental frequency is doubled, which makes the biphase more expensive in bandwidth.

2.4 Analog Encoding of Analog Information

Analog information can be directly carried on a wire without conversion. For example, a voice signal on the telephone local loop is commonly carried at "baseband." Nevertheless, it is common to use an analog signal to modulate a carrier, as in sending digital information on analog lines. There are two main reasons for this.

It is often difficult to send baseband. For example, voice communication uses the frequencies range from about 30 to 15 000 Hz; to send these frequencies directly would require antennas many miles in dimension. Similarly, optical fibers operate at frequencies approaching 10^{14} Hz, which will not support direct transmission of the frequencies of most communication traffic. In addition, the frequency bandwidth associated with the information source could only be used once for one communication. Modulation onto carriers allows many communications to be transmitted simultaneously using different carrier frequencies (frequency-division multiplexing).

The signal can be superimposed on the carrier sine wave by varying one of three defining parameters of the carrier: the amplitude, the frequency, or the phase. These three schemes are called *amplitude modulation* (AM), *frequency modulation* (FM), and *phase modulation.* For analog information, frequency modulation and phase modulation are essentially indistinguishable, and so usually only AM and FM are considered.

3. EFFICIENT USE OF CHANNELS

Even with the immense channel capacities possible with modern fiber-optic communications, it is necessary to make efficient use of channels. The most straightforward way is to multiplex several signals requiring relatively small amounts of capacity onto one larger-capacity channel that has enough capacity to carry all the signals. Often the multiple signals that are carried on a shared channel have characteristics that make greater efficiencies possible. If the signals are bursty, then concentration can be used. Here the shared channel may have a capacity that is less than the sum of the *peak* capacities of the subchannel signals. Of course, the sum of the *average* rates must be less than the channel capacity. The hope is that since the traffics are bursty they will not all be at their peak rates at the same time. Finally, if there is significant redundancy in the information being sent, compression can be used.

3.1 Multiplexing

In multiplexing, the channel capacity of the shared channel is divided among the subchannels. This can be done in three basic ways: by frequency, by time, or by space. In addition, "code division" can be considered to be a multiplexing scheme, although we will consider this as a form of concentration. Multiplexing is an actual partition of the shared channel's capacity. Even if there is no traffic for one or more subchannels, the capacity is still allocated.

The simplest method of multiplexing is to divide the bandwidth of the channel into subchannels by frequency (*frequency-division multiplexing,* FDM). If a transmitted signal is band limited and thus occupies a finite spectral region, other signals can be transmitted at the same time, provided that their frequencies do not overlap. Since frequencies can be shifted at will by means of modulation (or mixing), signals that occupy the same spectral region can be shifted in frequency relative to each other so that their spectra do not overlap. They thus can be added and transmitted over a single communication link. This is frequency-division multiplexing. Demultiplexing is achieved by filtering out the individual signals and shifting them back to their original spectral location. Common examples of FDM include over-the-air television, and AM and FM radio stations. In each case, the stations share a common spectrum (bandwidth), and each is allocated a carrier frequency and bandwidth. For U.S. analog television, the bandwidth allocation for a channel is 6 MHz. Frequency-division multiplexing is the simplest to accomplish since much of it is done by filters that can be passive. However, the filters are not ideal, and extra frequency allotments have to be assigned between channels (*guardbands*) to prevent the signal from one channel escaping into the bandwidth of an adjacent channel. This makes FDM relatively inefficient.

In optical communication, the information-carrying electrical signal modulates the power of the optical source, a laser diode or light-emitting diode (LED). We must thus differentiate between the frequency of the information-carrying electrical signal and the frequency of the optical source, which carries no information. Generally, "frequency" implies electrical signal frequency, whereas "wavelength" refers to the optical source.

Silica-based optical fibers have a transmission window that extends in wavelength between approximately 800 and 1600 nm. This corresponds to a spectral region of 800 nm (or 187 500 GHz). Commercially available laser diodes have a spectral width in the order of 0.1 to 2 nm (or 12.5 to 400 GHz).

A multiplicity of optical sources, adequately separated in wavelength, can thus be coupled to a single transmitting fiber, each source being modulated by different electrical signals and transmitting information independent of each other. This form of FDM (in which the individual carrier frequency sources are uncorrelated to each other) is known as wavelength-division multiplexing, or WDM. Of course, WDM does not preclude the additional use of either frequency division or time-division multiplexing to allow simultaneous transmission of a multiplicity of signals through the individual optical sources.

As the demand for channel capacity increases, WDM will be used extensively in long- and short-haul optical communications. It essentially converts a single fiber into the equivalent of a multiplicity of fibers and allows for network expansion without any major physical modifications of the existing links. It is also envisioned that WDM, in conjunction with photonic switching, will be used to provide to the house and office an assortment of information services, such as voice, video, computer data, and telemetry.

A scheme more efficient than FDM, when it can be utilized, is *time-division multiplexing,* TDM, where the subchannels are separated by time on a digital main channel. Thus, in the simplest case if there are, say, three signals, A, B, and C, one can send a bit from A, then a bit from B, then a bit from C, then a bit from A, . . . on the main channel to derive three equal-capacity subchannels. At the receiving end, a commutator is required to sort out the various signals. Clearly, for proper sorting or demultiplexing of the individual messages, the controlling clock at the receiving end must be synchronized with the clocks of the transmitters. This usually requires that all transmitters feeding the channel are nearly synchronized with each other. Most digital links that use TDM require such synchronization. The overhead for this process is minimal and, therefore, the efficiency high in the sense that the sum of the capacities of the subchannels can be essentially equal to the capacity of the main channel. However, this scheme, like all multiplexing schemes, wastes channel capacity if one or more of the subchannels has no information to send, since the capacity for the subchannel is reserved nevertheless. This scheme is also called synchronous multiplexing since the subchannels do not need to be addressed in any way to be identified but can be identified simply by counting, assuming the receiver and transmitter are synchronized.

Space-division multiplexing refers to the use of physical separation to distinguish subchannels. Examples include

1. wire cables that consist of many wire pairs;
2. frequency reuse in cellular telephony, where a frequency used in one cell can be reused for an entirely different call in another cell sufficiently distant so that the signals do not interfere;
3. the use of separate microwave dishes on the same towers to provide separate channels between the towers; and
4. the use of spot antennas for satellite communications to allow the carriers to be used in different places of the earth to carry different information.

3.2 Concentration, Statistical Multiplexing

Concentration is an extension of TDM. If the traffics to be carried on the subchannels are bursty, then one can rely on the statistical properties of the traffic to assume that, in general, the total amount of traffic summed over the traffics will be substantially less than the sum of the peak traffic rates. This leaves open the possibility of using a main channel with capacity substantially less than the sum of the *peak* rates of the subchannels (although, of course, not less than the total of the *average* traffics). Then one only sends the traffic on the subchannels when the traffic is offered. This approach has several consequences. First, since the traffic to the subchannels varies, one needs an *addressing scheme* so that the transmitter can tell the receiver where in the data stream to find the information for each subchannel that is active. For this reason, the subchannels cannot be identified simply by having a synchronized clock as was the case with TDM. Because of this, concentration is also called *asynchronous multiplexing.* Another very common term for this approach is *statistical multiplexing,* reflecting the dependence on the statistical properties of the traffic on the subchannels. A second consequence of concentration results from the fact that while in the long run there is enough capacity on the shared channel to support the subchannels, on occasion, hopefully rare, the total of the subchannel rates will exceed the shared channel rate for short periods of time. This is handled by providing *buffering* to store the excess traffic until the transmitting subchannels reduce their load on the system and the saved traffic can be metered out. This can give rise to two kinds of problems. First, the buffer may become filled up and data lost (in practice, signals are often sent back to the transmitters using the subchannels, inhibiting them from further transmission until the buffer clears). Second, the buffering can introduce substantial and varying delays (jitter). These latter problems are particularly onerous for real-time applications.

There is still another way of concentration that allows multiple digital users to share the same channel. This is code-division multiplexing (CDM). CDM is based on the rather counterintuitive notion of increasing the natural bandwidth, B_0, of each user to a larger one, B; that is, we spread the spectrum. This technique of spread spectrum arose out of military efforts to combat jamming and to provide secrecy. The ratio of B to B_0 is called the *spreading factor.* All users share the same time and the same spectral interval but use a different spectrum spreading code. This can be done in one of two ways (or both). Each bit can be translated in time into a coded sequence of B/B_0 digital signals called *chips.* Such spreading codes are called *direct sequence* codes. Alternatively, the signal can jump from carrier frequency to carrier frequency in time. These are called *frequency hopping* codes. A replica or replicas of the spreading codes are available at the receiving end. Correlation or other techniques are used that result in collapse of the spectrum and recovery of the signal. The spreading factors, whether direct or frequency hopping, are designed so that received signals that do not correspond to the desired signal when correlated with the replica of the desired signal results in a collapsed signal that is near zero and appears as low-level noise superimposed on the desired signal. This allows the desired signal to be recovered from the shared spread spectrum even in the presence of other communications in the same spectrum.

While CDM requires more bandwidth than either FDM or TDM, it often results in more efficient frequency utilization because a transmitter only uses bandwidth when it is active. This approach is especially useful in satellite and cellular radio communication links.

3.3 Compression, Error Control, and Encryption

Suppose we consider n bits of information from the same information source at various intervals of time. We can think of these n-bit strings as random variables with distribution determined by the source. If all possible 2^n n-bit strings are not equally likely, then the source has *redundancy* and the messages can be coded in such a way as to reduce the required data rate. Redundancy plays a key role in three areas of communications.

1. Compression: To the extent that "messages" are redundant they can be compressed to take up less communication or storage capacity.
2. Error detection and correction: One can add redundancy of a "good" sort to make it possible to detect or even correct transmission errors.
3. Cryptography: Redundancy is essential for breaking codes when all you have available is ciphertext (encoded messages). Conversely, the objective of encryption is to make the ciphertext appear as random as possible to minimize the redundancy.

Redundancy "measures, in a sense, how much text in the language can be reduced in length without losing any information" (Shannon, 1949). A little more precisely, a language is redundant if for some n, not all strings of length n have the same probability. Redundancy also measures the structure of a language, message, or source. Language structure limits the number of strings or their likelihood so that not all strings are equally likely. For example, in English, "q" is normally followed by "u," and "t" is frequently followed by "h."

3.3.1 Compression A simple, but frequently used, example of compression is *run-length coding.* For example, when sending a facsimile of a page with a lot of white space, long sequences of white pixels will appear. This can be much compressed by making use of a special code (escape character) followed by the character (pixel in this case) that is repeated, followed by a count of the number of repeats. At the received end the original bit map can be exactly recreated by replacing the code, the character, and the count by the appropriate number of characters. In general, there are two types of compression, *reversible* or *lossless* and *irreversible* or *lossy* compression. Run-length coding is an example of the first type because the exact signal before compression can be precisely reconstructed at decompression. For voice, image, and video applications, the compression ratios, ranging up to about 10:1, obtained by lossless compression are often not sufficient. Higher compression ratios, up to about 100:1, can be achieved at the cost of some loss in accuracy of the decompressed message compared to the original one. These inaccuracies are carefully controlled so as to not degrade the received message too much in the context of the application it is used for. However, for some applications, particularly for data services, only lossless compression will do.

3.3.1.1 Lossless Compression. After run-length coding, the two most commonly encountered lossless compression schemes are Huffman coding (Huffman, 1952), and Lempel–Ziv coding (Ziv and Lempel, 1978).

The latter approach, in particular, is used commonly for compressing computer files for storage and transmission. Huffman coding is a generalization of the Morse code. When Morse constructed his telegraph code of dots, dashes, and spaces, the codes for different letters were, in general, of different length. He was careful to assign the shortest codes, for example e = *dot* and t = *dash,* to common letters and long codes, e.g., j = *dot dash dash dash* and q = *dash dash dot dash* for less frequent letters. This gives compression because of the more frequent occurrence of short letters in common text. Note that this gain is lost if there is no redundancy in the original text; if each letter is equally likely, compression gain cannot be achieved in this way. Huffman coding is an optimal scheme in the following sense. We assume that we are transmitting messages expressed in an "alphabet," say the 7-bit ASCII character set, and that we know the frequency of occurrence of the various characters in the alphabet; that the frequency is stable in time; and that the characters are generated with these frequencies independently of one another. Unfortunately, the last assumption is blatantly false, and the others are difficult to achieve in practice. Nevertheless, Huffman provided an algorithm for assigning variable–bit-length codes for the characters so that the average expected message lengths would be minimal over such codes. Lempel and Ziv introduced a coding scheme that is adaptive. It learns the statistical properties of the source "on the fly," and it adapts to changes in these properties as well. In their system the encoded elements are of constant length, but the source text that they represent is variable; Huffman coding is exactly the other way around.

3.3.1.2 Lossy Compression. For many applications, particularly multimedia ones, the compression ratios achievable by lossless compression do not suffice. Two schemes commonly used for achieving higher ratios at the cost of some inaccuracies in the decompressed message were developed by the Joint Photographic Experts Group (JPEG) and the Moving Picture Experts Group (MPEG) of the International Standards Organization. Both methods treat images in a similar way. We sketch the JPEG scheme. It takes place in three steps. In the first step the image is divided into 8-pixel by 8-pixel blocks. The blocks are then individually converted into the "frequency" domain by a discrete cosine transformation (DCT). This is motivated by the fact that the higher-frequency elements of the block are usually less important for human perception than the lower ones. In the second step the DCT coefficients are discretized. Compression is achieved by discretizing the coefficients corresponding to higher frequencies more grossly. The amount that this is done is controlled by a quality parameter. The higher the parameter, the more compression and the less the fidelity of the reconstructed image. After the second step, many of the coefficients, especially those corresponding to high-frequency terms, will be zero. A form of run-length coding of zero coefficients plus either Huffman coding or arithmetic coding provides final compression. Note that only the second step introduces loss. The first and third steps are lossless (reversible).

MPEG uses these devices plus additional ones to take advantage of the correlation in time of successive images in movies and videos.

3.3.1.3 Uses of Compression in Telecommunications. Despite the seemingly unbounded bandwidth of fiber-optic channels, compression still plays a very important role in many aspects of modern telecommunication. Lossless compression is used to reduce the storage and communication capacity for data communications. Compression ratios near 10:1 are routinely obtained for facsimile transmission. Modern high-performance modems use a form of Lempel–Ziv coding to improve data rates. Lossy compression is used in many multimedia applications. To transmit full-motion video from CD-ROMs, over telephone or CATV channels (e.g., video on demand), or over satellite channels, lossy compression is virtually always used.

3.3.2 Error Detection and Correction

All communication channels and networks have errors due to noise and other factors. If special techniques are not used, these errors often are more frequent than the application can tolerate. Thus, one would like to correct errors as they are received or detect their existence and, if necessary, retransmit them. This can be done by adding redundancy to the messages. Thus, on the same channel one can be using compression techniques to reduce the required transmission and adding redundancy to detect and/or correct errors. To illustrate the ideas involved with the detection and correction of errors, we introduce a simple but inefficient method. Suppose we are attempting to transmit a message of n bits over a channel that is subject to transmission errors. One coding is to simply send the message more than once. For example, suppose we send the message twice. We then compare the two received messages bit by bit, and for those corresponding bits that do not match, we declare an error; otherwise, we say that the common value of the bits is the true value. There are at least three significant drawbacks to the scheme. First, on occasion, both of the corresponding bits could be wrong, in which case we would declare the wrong value correct. This problem can be particularly acute if the probability of a bit being in error is high or if the size of the message n is large. Second, while we can detect errors, we cannot correct them using this technique. Third, we are rather wasteful of communication resources since the effective length of each message is doubled. The first two problems can be eased at the cost of worsening the third by sending more than two copies. For example, we can send three copies of the message. Then we say a given bit of the message is received correctly if the corresponding bit of each of the three copies has the same value. We can correct a bit, if there are discrepancies by taking the bit value that appears at least twice in the three copies. If we wish more confidence, we can send even more copies. However, this is very inefficient. Fortunately, there are very much more efficient methods. Two common methods are *parity check* bits and *cyclic redundancy checksums* (CRCs). Parity checks are a rather weak form of error detection but appear commonly. The idea is that, given a sequence of bits from a message, typically 7 bits, redundancy in the form of an additional parity bit is added so that the total number of ones in the eight bits (byte or octet) is even in the case of *even parity* or odd in the case of *odd parity*. The overhead of this type of coding is 1/7 or about 14.3%. CRCs are much more effective. Typically, a message of 8 000 to 16 000 bits has added to it a checksum of 32 bits. The overhead here is then less than 1%, and the protection is strong. The general idea of the CRC is that we interpret the bits of a message M as a binary "number." Associated with a given redundancy system is a generator, G, which is also interpreted as a binary number. Typically, G has much fewer bits than M. Then one divides M by G, obtaining a remainder R, which will be smaller than G. Essentially, the message M is transmitted along with the remainder R. The receiver, who also knows the generator, then takes the received message M' and the received remainder R' and recomputes the remainder of M' when divided by G, to obtain R''. If $R' = R''$, the message is deemed correct. To make the process more efficient, the process is not exactly as described here. For example, in practical implementations M and G are not actually interpreted as ordinary numbers but rather as elements of a field over the integers modulo two. Moreover, things are arranged so that the receiver, instead of comparing two strings, compares a processed part of the received string with a string of zeros of the same length. For smart choices of G, most common errors can be detected and many corrected. In practice, if the remainder checks out okay, the receiving party sends an acknowledgment to the sender. If the sender does not receive an acknowledgment, it sends the message again. In some applications—for example, in communication with deep-space probes—this retransmission process is much too inefficient because of the long propagation delays. In these cases, forward error correction is more useful. Here, additional redundancy is added so that the errors can actually be corrected with very high confidence without retransmission. Similar techniques are used on CD-ROMs, so that minor imperfections in the discs can be corrected for.

3.3.3 Encryption For many applications, secrecy of communications is important. This is often provided by encryption. There are many other applications as well. For example, one often needs to authenticate the other party in a communication to provide confidence that one is communicating with the appropriate party or that the entity making use of computing or communication resources is entitled to do so. There are two basic types of encryption systems in use. The first type is variously called *symmetric, pri-*

vate key, or *single key.* The other approach is called *asymmetric, public key,* or *two key.* If Alice wishes to communicate a message *M* to Bob using the first system, they must share a common secret key *K*. Alice then encrypts *M* using *K* to obtain ciphertext, *C*, using an encryption algorithm. The algorithm must have the property that is very, very difficult to recover the *M* from *C* unless *K* is known. If *K* is known, then one can simply recover *M* from *C* using a decryption algorithm. The Data Encryption Standard (DES) is a commonly used example of a symmetric key system. In asymmetric systems, two keys are used. Alice encrypts *M* using a public key, which is well known to belong to Bob, obtaining ciphertext *C*. Bob then recovers *M* from *C* using a different private key that no one else knows. A common example of an asymmetric cryptography is the Rivest–Shamir–Adleman (RSA) algorithm (Rivest *et al.,* 1978). In general, symmetric systems such as DES can be implemented to run much faster than asymmetric systems such as RSA. On the other hand, asymmetric methods are more flexible and can be used for more applications. Quite often, hybrid systems using both types of encryption are used where the asymmetric method is used to establish the secret key for use by the more efficient symmetric algorithm.

4. EFFICIENT USE OF NETWORKS

4.1 Circuit, Message, and Packet Switching

In Sec. 3, we discussed the efficient use of individual telecommunications channels. Similar considerations obtain in networks of channels. We will think of a communication network as consisting of terminals, switches, and channels. The terminals originate or terminate traffic, the switches route information, and the channels carry the traffic between terminals and switches.

There are two basic ways that a network of channels can be used for communication. In the first way, *circuit switching,* logically a string of channels are hooked together at switches to form a path from the sender to the receiver. While communication is active, the entire path is dedicated to the communication. This use of networks has the advantage that the message is not delayed as it travels through the network. It simply sees what looks like a single channel derived from the network. It generalizes the notion of multiplexing for channels to networks. Moreover, like multiplexing, no addressing is needed to identify or route the traffic after the call is set up. This use of a network is most useful for traffics that are smooth, or for real-time traffic such as voice or video. Like multiplexing, in circuit switching the channel is dedicated so that when it is not being actively used the capacity is wasted and not available to others.

At the other extreme is *message switching.* Again the message works its way through the network. However, it does so only one switch at a time. In message switching, each channel is used only while the message is traversing it. The message is thus sent from the origin to the first switch. The message is then stored in that switch. Then the message is passed to the next switch. The channel from the origin to the first switch is no longer involved with the transmission. This procedure is continued until the message reaches the destination. This technique is related to statistical multiplexing where the channels are only used when there is traffic to be sent. However, latency or delay is encountered at each switch, since the message has to be read into the switch and then retransmitted. The longer the message, the longer the delay *at each switch.* To control this delay, *packet switching* was introduced in the late 1960s, where each message is broken into packets with a fixed maximum length. Thus since no packet can be larger than a given size, the retransmission delay at each switch is limited. In recent years this approach has been extended in asynchronous transfer mode (ATM) to where all communication takes place in cells that all have the same short (53 octet) length. In general, different packets between a given origin and destination need not follow the same path through the network. This makes addressing necessary to identify the packets and to route them. Packet switching is most useful for bursty traffic so that the statistical variations can be taken advantage of. An intermediate approach is the *virtual circuit.* In this case, the information is broken into packets, but the packets all take the same path through the network. However, in general, the path is not dedicated to the virtual circuit, and other traffics can use the channels in the path on occasion. This, for example, reduces the addressing requirements while still allowing the communications capacity to be dynamically shared.

4.2 Architectures, Standards, and Protocols

As telecommunications systems become more complex, agreements are being forged that help to organize the hardware and software components effectively. One useful way to organize one's view of the software and hardware of a communication network is according to the Open Systems Interconnection Model of the International Standards Organization. The functions of communication are partitioned, hierarchically, into seven layers. They are

1. the Physical layer,
2. the Data Link layer,
3. the Network layer,
4. the Transport layer,
5. the Session layer,
6. the Presentation layer, and
7. the Application layer.

These layers correspond to different views of the network. Layers 1 and 2 are only concerned with individual links within the network. Layer 3 actually deals with the way the information travels *through* the network, while the higher levels, levels 4–6, have the "cloud view" of the network. They only see information entering and leaving the network, but the workings within the network and its individual channels are hidden from these higher levels. In the early 1970s Vint Cerf and Bob Kahn (Cerf and Kahn, 1974) developed the theory of internetworking, which gave rise to the TCP/IP model of communication and, ultimately, the Internet.

5. TRANSMISSION MEDIA

Transmission media can be divided into two categories: wireless transmission into air or space and transmission into constrained channels such as wires or cables.

5.1 Wireless

For many reasons "tetherless" communication is desirable, that is, communication where the communicating entities do not have to be connected by wires or cables. For example, if one or both ends of a conversation are in motion, wire or cable connections are often impracticable. Likewise, great distance or difficult terrain may make wire connection infeasible. Finally, the broadcast nature of wireless communication supports the convenient transmission of the same information to many receivers. This is useful for radio and television. Because the environment is less controlled, wireless transmission is more difficult than using cables or wires. However, if the antennas for wireless communication are fixed and can be well situated, the performance can be quite predictably good. This is the case for satellite communication to fixed ground stations and point-to-point microwave communication between fixed antennas. The performance is limited mainly by the free-space loss inherent in any signal radiated into space or the atmosphere, by atmospheric absorption, and by refraction. On the other hand, mobile communication or communication with antennas that cannot be located for best transmission is much more complex. Fading, multipath interference, and a number of other additional performance degradations occur.

Cordless telephones, mobile telephones, and even phones in airplanes are examples of modern wireless communications. Paging and messaging services are other examples.

5.1.1 Limitations on Wireless Communications

5.1.1.1 Free-Space Loss. For any wireless communication, by the geometry of the radiation, the signal diverges with distance, decreasing the signal intensity. For satellite communication this is the primary mode of "attenuation." Under idealized conditions, the ratio of the power P_r received by the antenna to the radiated power P_t is given by $P_r/P_t = A_rA_t/(d\lambda)^2 = A_rA_tf^2/(cd)^2$, where A_r is the effective area of the receiving antenna, A_t is the effective area of the transmitting antenna, d is the distance between the antennas, f is the carrier frequency, $\lambda = f/c$ is the wavelength, and $c = 300\,000$ km/s is the speed of the electromagnetic wave. Thus, for the same antenna dimensions and separation, the higher the carrier frequency f, the lower is the free-space path loss, which is the fraction of the loss that depends only on the divergence with distance and not on the antennas. However, the ration P_r/P_t is most commonly expressed in a different form. With $G_r = 4\pi A_r/\lambda^2$ and $G_t = 4\pi A_t/\lambda^2$ being the "gains" of the receiving and transmitting antennas (relative to isotropically radiating antennas), respectively, $P_r/P_t = G_rG_t(\lambda/4\pi d)^2 = G_rG_t(c/4\pi df)^2$, giving a "free-space path loss" = [32.4 + 20 log(d/km) + 20 log(f/GHz)] dB. This seems to imply, erroneously, that the path loss increases with the square of carrier frequency if one loses sight of the dependence of antenna gain on frequency.

5.1.1.2 Absorption. An additional loss between the transmitting and receiving antennas is absorption. Two atmospheric gases contribute most to attenuation: water vapor and oxygen. A peak of approximately 0.2 dB/km for an absolute humidity of 15 g/m^3 occurs in the vicinity of 22 GHz due to water vapor. At frequencies below 15 GHz, the attenuation is reduced to less than 0.04 dB/km. The presence of oxygen results in an absorption peak of 10 dB/km in the vicinity of 60 GHz but contributes less than 0.02 dB/km at frequencies below 30 GHz. Accordingly, at frequencies below 15 GHz, typical path lengths of up to 50 km are essentially unaffected by atmospheric absorption.

Rain and fog (suspended water droplets) cause scattering of radio waves that results in attenuation. This can be a major cause of signal loss. The degree of attenuation depends on the rainfall rate and frequency. Light rain (1 mm/h rainfall rate) will cause attenuation of 0.05 dB/km at 10 GHz and 0.3 dB/km at 20 GHz; when the rain is heavy (15 mm/h), the respective attenuation values are 0.7 and 3.0 dB/km; cloudburst rain (100 mm/h) causes attenuation of 5 and 25 dB/km at 10 and 20 GHz, respectively. Neither space diversity nor frequency diversity are effective as protection against rainfall effects. Thus, in areas that experience significant amounts of precipitation, either path lengths have to be kept shorter or lower-frequency bands should be used.

5.1.1.3 Multipath. When an obstacle is below the line-of-sight path, the divergent radio wave can be partially reflected to form a second signal at the receiving antenna. At the point of reflection, the indirect signal undergoes attenuation and is shifted in phase by approximately 180°. The reflected signal can be particularly strong when the reflecting surface is water or smooth terrain. At the receiving antenna, the direct and reflected signals combine. Depending on the differences in the path lengths of the direct and reflected waves, the composite signal can be either larger or smaller than the direct signal. Reinforcement and cancellation of the signal resulting from the signal following multiple paths can be controlled for communication between fixed, well-sited antennas and between satellites and fixed ground stations, but for mobile telephony and communication to antennas that are not well sited, multipath considerations can be paramount, and no direct path may in fact exist.

5.1.1.4 Refraction. Additionally, radio waves are refracted (or bent) when they propagate through the atmosphere. The refraction is caused by changes in the dielectric constant or refractive index with altitude, or by other spatial changes in the atmospheric conditions. Normally, the refractive index decreases with altitude, causing radio waves to diffract downward. However, on occasion, weather conditions may lead to refractive index variations with height that differ significantly from the typical variations. This may result in a situation in which only a fraction or no part of the line-of-sight wave reaches the receiving antenna. When the refractive index variation is more rapid, anomalous wave propagation or "ducting" may occur. The result is attenuation of the line-of-sight signal and/or frequency-dependent changes in the interference between the multipath signals. In either case, the signal fades. This fading is usually of short duration. Both space diversity and frequency diversity are helpful in combating fading.

5.1.1.5 Thermal Noise. Thermal or "white" noise is inescapable. It arises from the thermal activity of the devices and media of the communication system. Because of the weakness of the signal received by satellite Earth stations, thermal noise is particularly important for satellite communication.

5.1.2 Satellite Communication Satellite transmission is unique in that it is capable of providing global communication, including transmission to and from moving terminals (such as trucks, ships, and airplanes), without the need of an extensive terrestrial infrastructure.

Geosynchronous satellites remain in a fixed position in the sky relative to the earth. For this to be the case, the satellite must be in the equatorial plane, and its circular orbit must have a period of one day, identical with that of the Earth's rotation. This requires an altitude of 35 838 km. From such an altitude, four satellites can cover most of the planet's surface. Only the polar regions, beyond 81° latitude, remain beyond the line of sight of geosynchronous satellites. This configuration has many advantages to recommend it. Because the satellite is stationary relative to the Earth, there is no problem with frequency changes due to relative motion of the satellite and earthbound antennas. Moreover, the tracking of the satellite by its Earth stations is simplified. On the other hand, there are some problems. The most obvious one is that the signal can get quite weak after traveling at least 35 838 km. Also, the polar regions and the far northern and southern hemispheres are poorly served by geosynchronous satellites. Moreover, even for signals traveling at the speed of light, 300 000 km/s, the delay in sending a signal to the satellite and back is considerable.

The distance from an Earth station to a geosynchronous satellite may range from a minimum of 35 800 km to a station on the equator beneath the satellite to 41 800 km for a station at the horizon relative to the satellite. Accordingly, the one-way propagation time from one Earth station to another may be as long as 280 ms, resulting in a round-trip delay time of 560 ms. In conversational voice communication this is acceptable. However, often a domestic satellite is used in tandem with an international satellite. In such a case the total trip delay is over 1 s, which might not be tolerable. For such reason, international satellite transmission is often implemented on a one-way basis, with the return path being terrestrial, such as an optical-fiber submarine link.

Geosynchronous satellites use their assigned frequencies over a very large area, although the area can be restricted to some extent by using special antennas on the satellite such as spot and steered beam antennas. For some broadcasting applications this is an advantage; however, it does limit the reuse of the satellite frequencies. To solve these problems, orbits other than geosynchronous have been designed for satellites. Two of importance are highly elliptic orbits and low Earth orbits (including space stations).

In northern Europe and in polar regions the angle of geosynchronous satellites above the horizon can be uncomfortably low. At low angles, satellite signals are easily obstructed by natural obstacles, hills and mountains, and by man-made ones, buildings. In northern Russia this is a particularly pronounced problem. To deal with it the Russians have used highly elliptic orbits for some of their satellites. The prototype of this type of orbit was used by the Russians in the Molniya series of satellites. The Molniya orbit was designed to offer a better angle of inclination of the satellite to the horizon in the far northern hemisphere and the polar region. The orbit is elliptical and is inclined 65° with respect to the equator. It is at its farthest from the Earth at the 40 000-km (slightly more than the 35 838-km height of a geosynchronous satellite) apogee over the northern hemisphere, and is closest, 500 km, at its perigee in the southern hemisphere. Thus it has very good coverage over the northern hemisphere. Moreover, the satellite spends most of its 12-h orbit over the northern hemisphere. Three satellites sufficed for 24-h coverage. A similar system was proposed by the European Space Agency for digital audio broadcasting.

This type of satellite has little to recommend it beyond its coverage of polar regions. Because the orbit is not geosynchronous, multiple satellites are needed for 24-h coverage. Moreover, to the extent that the apogee is greater than geosynchronous orbit altitude, the signal attenuation (loss) is even more.

The original AT&T satellite proposal was for low Earth-orbiting satellites (LEOs). However, most of the early satellites were geosynchronous. Nevertheless, low Earth orbits have advantages, and many recent satellite proposals (e.g., Motorola's Iridium) are based on them. The idea is to use constellations of low Earth-orbiting, inexpensive satellites sometimes called "lightsats." They orbit at altitudes of about 750–3000 km above the Earth's surface. Therefore, their signal is much stronger for given transmission power than geosynchronous satellites. Their coverage can be better localized so that spectrum can be conserved. For this reason, this technology is being currently proposed for communicating with mobile terminals and personal terminals, which need stronger signals to function. On the other hand, to provide broad coverage over 24 h, many satellites are needed. Sixty-six are being proposed by Motorola for their Iridium system.

A number of commercial proposals have been made to use clusters of LEOs to provide communications services. These proposals can be divided into those intended to work at communication frequencies less than 1 GHz and using no more than 5 MHz bandwidth and supporting data rates up to 10 kbps, the "little LEO's," and those working at frequencies above 1 GHz and supporting data rates up to a few megabits per second, the "big LEO's." LEO systems operating below 1 GHz typically are aimed at paging, tracking, and low-rate messaging applications. Big LEO systems tend to offer the same services as small LEO systems with the addition of voice and positioning services. The Iridium system is planned to be a big LEO.

The first satellite to use active repeaters was launched in 1962. These early satellites were placed in low orbits and required complex Earth stations to track the motion of the satellite. The first commercial geosynchronous satellite was placed in operation in 1965. By 1970, the international satellite consortium INTELSAT had established a global satellite-based telecommunications network.

The lifetime of a geosynchronous satellite is typically 10 to 20 years, while LEOs are designed for a 7- to 10-year life. The life of a geosynchronous satellite is limited by the amount of propellant stored onboard for firing thrusters that keep the satellite to within 0.5° of its assigned position.

The transmitted satellite power is orders of magnitude greater than power received by its antenna. To affect isolation between the transmitted and received signals, the received signal is shifted in frequency for retransmission. For a number of reasons, the lower frequency is assigned to the downlink.

The most desired frequencies for commercial use are in the 4- to 6-GHz band. The downlink is 3700 to 4200 MHz, and the uplink is 5925 to 6435 MHz. This frequency band is suitable for use under all possible weather conditions. Moreover, in this frequency range the equipment is relatively inexpensive, and the cosmic noise is small. However, since existing terrestrial microwave radio links operate in the 4- to 6-GHz bands, the FCC limits the power density on Earth from satellite transmitters (to 12 dbW effective isotropically radiated power in any 4 kHz band).

In a typical satellite, the 500-MHz bandwidth is subdivided into twelve 36-MHz bands, with 4.0-MHz guardband separation. Each of the 12 transponders aboard the satellite amplifies the signal received from the uplink and down-converts it by 2225 MHz for transmission on the downlink. Some satellites use polarization diversity to reuse the same frequency band. Such satellites have two sets of 12 transponders that operate with orthogonally polarized signals.

An antenna beamwidth of 17.3° would cover the entire surface of the Earth that is within view of the satellite. Within a factor of 1 to 2, the beamwidth of an antenna equals the ratio of the carrier wavelength to the antenna dimension. Thus, in the 4- to 6-GHz band, maximum coverage would be provided with an antenna 25 cm in diameter. However, satellite antenna dimensions are significantly larger than 25 cm, and, correspondingly, the beamwidths are much smaller. For one reason, the signal power density at the Earth's surface is proportional to the square of the antenna dimensions. Moreover, spot beaming is often used to increase the available bandwidth by allowing the reuse of frequencies in different areas of the Earth. The total available bandwidth is limited by the total number of satellites that can be placed in orbit. The U.S. standard allows synchronous satellites in the 6/4-GHz band to be assigned an orbital spacing of 2°.

Because there are few vacant spectral assignments in the 6/4-GHz band, newer satellites operate in the *Ku* band. These satellites use 11.7 to 12.2 GHz for the downlink and 14.0 to 14.5 GHz for the uplink. The allowed minimum orbital spacing is 3°. This commercial spectral band is suitable for use under most weather conditions. The bands above 17 GHz are more severely limited by precipitation.

Satellites are now used extensively to relay broadcast-quality television signals. Typically, a single 36-MHz transponder is used for each TV signal. This is accomplished by modulating the composite video and audio signals to generate an FM signal of 36-MHz bandwidth.

Satellite relays provide a channel for telephone and data signaling similar to conventional terrestrial links. Thus, either telephone signals can be frequency-division multiplexed and upconverted in frequency for transmission or the signals can be converted into 64-kbps PCM digital signal that are time-division multiplexed and modulated into a carrier for transmission. Correspondingly, either data can be time-division multiplexed or modems may be used to convert the data into signals that are compatible with analog transmission and FDM. Depending on the mode of transmission, three main methods are used to provide multiple access:

1. frequency-division multiple access (FDMA),
2. time-division multiple access (TDMA), or
3. code-division multiple access (CDMA) or other variants of spread-spectrum multiple access.

In all methods, the multiple access assignment can be either fixed (FAMA) or demand based (DAMA). In the DAMA mode, the assignment formats are changed in real time as needed to accommodate the traffic demands of the Earth stations involved. Clearly, this mode of operation uses the satellite capacity more efficiently but is more complex to implement than a fixed assignment.

5.1.3 Point-to-Point Wireless Communication Microwave radio systems provide cross-country communication links through the use of multiple repeater sites, as well as feeder service to long-haul coaxial and fiber cable systems. Individual circuit lengths range from a few kilometers to thousands of kilometers. Applications include telephone channels, wide-band data, and television transmission.

Microwave transmission uses line-of-sight free-space propagation of electromagnetic waves. Antennas are placed on towers of sufficient height to provide a direct and unobstructed path between the transmitting and receiving sites. The microwave spectral region extends from about 1.5 to 30 GHz. At these frequencies the predominant mode of signal transmission between the two radio antennas is direct line-of-sight propagation. Other modes of propagation, such as ground or surface waves that parallel the Earth's surface, ground reflection paths, and/or waves reflected off the troposphere or ionosphere, may cause destructive interference with the line-of-sight wave, and the signal will fade. Line-of-sight links are limited in length by the divergence of the electromagnetic wave ("free-space loss"), by obstacles along the path, and by curvature of the Earth. Terrain, atmosphere, and precipitation further affect the link length and signal quality.

Multipath propagation and atmospheric refraction changes cause time-dependent interference effects and result in signal fading.

Digital service constitutes an increasing portion of microwave transmission. Digital coding requires less power than analog signaling and is more immune to fading and other channel degradations. However, since digital modulation techniques result in broadband signals, filters must be used to limit the bandwidth of the modulated signals to the assigned radio channels while still allowing transmission with a minimum of intersymbol interference.

In the United States, radio channel assignments for commercial carriers are controlled by the Federal Communications Commission (FCC). FCC regulations specify the microwave frequency bands that are authorized for digital microwave and establish standards for efficient use of the radio spectrum and the control of interference between adjacent radio channels.

In the 2.11- to 2.13-GHz band, a radio channel bandwidth of 3.5 MHz is allowed with a minimum capacity of 36 PCM 64-kbps voice channels per radio channel. To achieve the required 6.312-Mbps rate, either 8-level phase-shift keying (PSK) or 16-level *quadrature amplitude modulation* (QAM) is typically used. The same radio channel bandwidth and minimum capacity requirements are specified for the 2.16- to 2.18-GHz band. For operation in the 3.7- to 4.2-GHz band, the radio channels are 20 MHz wide, with a minimum capacity of 1152 64-Kbps channels. A commercially available 64-level QAM system achieves the 90-Mbps rate at 15

Mbaud. The 30-MHz radio channel bandwidth available in the 5.925- to 6.425-GHz band allows a yet higher transmission rate if 64-level QAM is used. Three DS-3 streams, or 135 Mbps, corresponding to 2016 voice channels at 64-Kbps each, can be transmitted by such systems. The same or higher transmission rates are possible in the 10.7- to 11.7-GHz band with an allowed radio channel bandwidth of 40 MHz.

The conventional practice in frequency channel assignment for a particular band is to separate transmit and receive frequencies by placing all transmit channels in one half of the band and all return channels in the other half. Within each half-band, adjacent radio channels must be spaced far enough apart to avoid signal interference between channels. A common scheme used to increase adjacent channel discrimination is to alternate between vertical and horizontal polarization of the wave. This cross-polarization of adjacent channels provides isolation of more than 20 dB.

One commercial high-capacity radio-relay system in the 10.7- to 11.7-GHz band has six transmit and six receive radio channels, each having a 40-MHz bandwidth. Adjacent channels are separated by 27 MHz, and the separation between the transmit and receive half-bands is 140-MHz. The bit rate is 135 Mbps, providing 2016 digital voice channels per radio channel, for a total of 12 096 full duplex voice channels. Even higher capacities are possible with more recent systems.

Frequencies beyond the 10.7- to 11.7-GHz band have also been allocated by the FCC. They are, however, not in common use for transmission beyond metropolitan areas because of propagation problems during precipitation.

In the past, microwave radio systems provided over 70% of all the toll-message circuit distance within the United States. However, because of privacy concerns, rights of way and antenna location concerns, and capacity limits (to at most 50 000 full duplex circuits), the economic and engineering importance of analog and digital microwave radio is diminishing in comparison with optical-fiber communication links. Such links are next to impossible to tap, do not radiate, and are noise free, and the capacity of a cable with ten fibers is in excess of one million full duplex digital voice channels.

5.1.4 Mobile Telephony The first generation of mobile telephony was based on analog technology. Initially, conventional radio telephone service was used. Typically, in such a system perhaps only 666 voice channels would be available in a 40-MHz band. Service is provided through a high-power transmitter/receiver with an effective range of approximately 100 km, covering an area of 30 000 km^2. Consequently, when a channel is used by anyone in the area, it is not available for use by others. However, since the early 1980s, this approach has been largely supplanted by cellular systems such as the Advanced Mobile Phone Service (AMPS) developed by AT&T.

The cellular radio concept is to decrease the power of the transmitter and, correspondingly, reduce the range of the system to no more than a few kilometers. Thus, an area is divided into cells, each one served by its own antenna. Many more users can now be accommodated simultaneously, since the same radio-frequency channel can be used by another individual in a cell sufficiently distant. If more channels are needed, the existing cell sizes are decreased so that the existing channels can be reused more efficiently.

In the United States, two 25-MHz bands are allocated to AMPS, one for transmission from the base station to the mobile unit, 869–894 MHz, the other for transmission from the mobile unit to the base station, 824–849 MHz. Each of these bands is split in two to encourage competition; i.e., so that in each market two operators could be accommodated. An operator is allocated only 12.5 MHz in each direction for its system. The channels are spaced 30 KHz apart, which allows a total of 416 channels for an operator. Twenty-one are usually allocated for control, leaving 395 to carry calls. The control channels are data channels operating at 10 kbps. The conversation channels carry the conversations in analog using frequency multiplexing. At present, the telephone signals are frequency modulated with a frequency deviation of 12 kHz and bandwidth of 30 kHz. Frequency-shift keying is used for control signals. Control information is also sent on the conversation channels in bursts as data. Efforts are under way to use different encoding techniques, which will result in better spectral utilization.

No specific channel is assigned to a particular party. When a phone is on the air it automatically transmits its serial number to the Mobile Telephone Switching Office (MTSO). When a call is initiated, the MTSO assigns an available radio channel to the user's cellular phone and sends a ring signal to the called party. If the called party is land based, the connection is made via a central office to the telephone network. If the called party is mobile, the MTSO connects the user to the cell site that covers the area in which the called party is located. Available channels in that cell are used for base-station to mobile and mobile to base-station transmission. As the mobile users travel from one cell to another, the MTSO automatically switches them to available channels in the new cells, and the telephone conversation continues uninterrupted.

Second-generation mobile telephony, also generally based on the cellular concept, uses digital radio, which offers more efficient use of the spectrum as well as more functionality. One such system is the Global System for Mobile (GSM), which is very popular in Europe and is probably the fastest growing mobile telephony technology.

The third generation of mobile telephony will be the integration of mobile telephony into more general universal personal telecommunications and universal communications access. The first of these means the ability of a person to use conveniently any communication system in a large geographical area, perhaps globally, in terms of a single account. For example, one's phone number would travel with the person, not with a particular communication device. The second concept is that of a customer having one portable terminal that will work equally well in the office, on the street, and on airplanes or automobiles. Personal Communications Services (PCS) and Personal Communication Networks (PCN) are names attached to these concepts of global wireless communications.

PCS handsets are designed to be low power, and relatively small and light. Efforts are being made internationally to allow the same terminals to be used worldwide. Frequency allocations, worldwide, have been made for second-generation cordless telephones (CT-2) in the 800-MHz region and for more advanced personal communications (third generation) in the 1.7- to 2.2-GHz region of the spectrum.

The 1992 World Administrative Radio Conference (WARC 92) assigned worldwide allocations for "future public land mobile telecommunications systems (FPLMTS)." This concept includes both terrestrial and satellite-based services.

In addition, allocations were made for low Earth-orbiting (LEO) satellite services that can be used to support personal communications.

Some proposed technologies that come under the umbrella of PCS are American Digital Cellular System, Japanese Digital Cellular System, second-generation cordless telephones (CT-2), the European Community's GSM digital cellular service, and Digital European Cordless Telephone (DECT) for services such as advanced in-building wireless telephony, all supported by LEO satellites, and geosynchronous satellites, as well as terrestrial antennas.

5.2 Constrained Media

At some loss of flexibility because of the requirement to be attached to the physical media, guided or constraining media such as wires, cables, and optical fibers reduce or eliminate many kinds of channel distortion, attenuation, and interference.

5.2.1 Limitations on Constrained Media The communication of information in constrained media can be thwarted by four basic kinds of problems. First, the signals can become too weak to be useful—*attenuation.* The second kind is the corruption of the signal by noise, or interference; the third is *delay;* and the fourth is various kinds of *distortions,* including echos.

5.2.1.1 Attenuation. For wireless communications there are several ways in which signals can be attenuated. First, and most fundamental, is free-space loss that is an inverse-square law in distance and is the result of the geometric divergence of the signal from its source. This was discussed in Sec. 5.1.1. Also discussed there is attenuation due to absorption, principally by water vapor and oxygen. Finally, ducting and temperature inversions can cause attenuation where atmospheric variations divert the signal from the receiving antenna. In guided channels, attenuation is caused by absorption of the electromagnetic signals by the guiding media, such as metallic wires, cables, and/or optical fiber. The problem is not only the weakening of the signal but also the distortion caused by the variation of attenuation with frequency. Equalization techniques are often used to correct for this latter problem.

5.2.1.2 Noise. The two principal kinds of noise are *white* or *thermal noise* and *impulse noise*. The first is characterized by a power spectrum that is constant with respect to frequency at least over some band; the second, by random spikes of interference. The signal of interest can also be interfered with by other signals. *Cross talk* refers to the coupling of signals from other channels into the channel of interest. *Intermodulation* is interference resulting from the nonlinearity of frequency-division multiplexing giving rise to spurious frequencies. Finally, *echos* occur at impedance mismatches in channels. A common example occurs at the conversion of four-wire telephone transmission to two and conversely. Echos are particularly troublesome with satellite communication because of the long propagation delays. *Echo cancellers* and *suppressors* are devices designed to ameliorate this problem.

5.2.1.3 Delay. In a communication channel, signals are delayed, at a minimum, by the propagation time. In space or air, the signal propagates at very close to the speed of light, $c = 300\,000$ km/s. In optical fibers, the signal propagates at a speed determined by the index of refraction; the speed is c/ν, where ν is the index of refraction, which is about 1.47. For coaxial cable and wire pairs the situation is more complex. Especially for wire pairs, the speed of propagation varies strongly with frequency. This causes distortion as well as delay. Again, equalization techniques are used to combat the variability of delay with frequency.

5.2.1.4 Distortion. A common cause of signal distortion is the variation of features of signal transformation with frequency. For example, both propagation speed and attenuation vary with frequency, causing distortion. *Intersymbol interference* is the name given to the distortion causing the smearing of one pulse or other signal element into adjacent ones.

5.2.2 Twisted-Pair Cables

A *twisted-pair cable* consists of two insulated copper conductors twisted together. For multipair cable assemblies, the individual cable pairs are grouped in units or concentric layers to form the core of the cable, and adjacent pairs are twisted at different pitches to minimize interference between pairs.

The attenuation of twisted-pair cables depends on its series resistance, series inductance, and shunt capacitance. Other factors that influence the attenuation characteristics are the frequency-dependent skin effect and temperature change. To improve the frequency response for analog voice, loading coils (inductors) are usually inserted at 1.2-km intervals; however, these loading coils interfere with the use of twisted pair for digital communication. Typical attenuation values for a 0.4-mm (26-gauge) twisted pair is 8 dB/km at 10 kHz, 15 dB/km at 100 kHz, and 40 dB/km at 1000 kHz. Corresponding values for a 0.66-mm (22-gauge) twisted pair are 4 dB/km at 10 kHz, 6 dB/km at 100 kHz, and 20 dB/km at 1000 kHz.

Twisted pairs are the most common transmission medium linking individual telephones with a central office. Communication is analog with frequency-division multiplexing. In England, shielded twisted-pair cables are also used for cable TV distribution. Each pair carries a single TV channel, and channel assignment is demand based.

In the last 30 years, twisted-pair cables have also been used for digital transmission. In the United States, digital T-carrier systems operate over twisted-pair cables at rates up to 6.312 Mbps (T-2), with repeater distances of approximately 1.2 km. Because of higher loss and greater cross talk, the use of twisted-pair cables at transmission above 8 Mbps becomes undesirable, although twisted pair is used for short distances within buildings at up to 100 Mbps for local-area networks. Coaxial cables can be used for such applications, up to 500 Mbps. However, with the evolution of optical-fiber cables, new installation of twisted-pair or coaxial cable digital links is economically and operationally undesirable.

5.2.3 Coaxial Cable

A coaxial cable consists of an inner conductor surrounded by a concentric outer conductor. The two copper conductors are held in place and insulated by a dielectric material. A standard cable size for long-haul transmission has inner and outer conductor diameters of 2.6

and 9.5 mm, respectively, usually expressed as 2.6/9.5-mm cable. Other sizes commonly used are 1.2/4.4-mm and 0.7/2.9-mm cables.

The attenuation of a coaxial cable is proportional to its resistance. Alternating currents flow only in the outer layer of a conductor, with the thickness of the conducting layer being inversely proportional to the square root of frequency. This is known as the *skin effect*. As a consequence, the effective cross section of a coaxial cable decreases with frequency, and the attenuation correspondingly increases. One finds that the attenuation in dB/km of a coaxial cable is given by $\alpha = (A + Bf^{1/2} + Cf)$ dB/km, where A, B, and C are constants that depend on the cable geometry and f is the frequency in MHz. For a 2.6/9.5-mm cable, $A = 0.0150$, $B = 2.4235$, and $C = 0.0029$, giving 2.4 dB/km at 1 MHz and 19.0 dB/km at 60 MHz.

Prior to the evolution of optical-fiber technology, almost the entire long-haul telephone network in the United States was based on coaxial cables. The transmission mode was analog, with frequency-division multiplexing. Each voice channel was limited to 4.0 kHz bandwidth. Modulation was single sideband. The first long-haul L1 system was placed in service by AT&T in 1941. It used eight coaxial units per sheath, each carrying 720 voice channels, for a total capacity of 2160 full duplex channels. It was followed in 1953 by the L3 system with a capacity of 9300 voice channels. The most advanced system, the L5, was placed in operation in 1973. It uses 20 active coaxial units per sheath with two spares. Each cable has a capacity of 10 800 voice channels in one direction for a total capacity of 108 000 full duplex channels. AT&T's L multiplexing scheme uses frequency-division multiplexing in a hierarchal fashion to package the channels. A "group" of 12 message channels (voice or data) contains 12 inverted 4.0-kHz sidebands. Groups use 48 kHz of spectrum. A "supergroup" made up of five groups requires 240 kHz. A "mastergroup" of ten supergroups with guardbands is assigned 2.520 MHz. Finally, a "jumbogroup" of six mastergroups (and 36 000 voice channels) uses 16.894 MHz. Three multiplexed jumbogroups with adequate guardbands using the band from 3.124 to 60.556 MHz and additional single-frequency control signals form an L5 line.

Because of the strong frequency dependence of its attenuation, a coaxial cable has little bandwidth; however, its nominal bandwidth can be increased substantially by equalization techniques. For example, it is capable of carrying a multitude of frequency-division multiplexed signals since each is only 4.0 kHz in bandwidth. To ensure that all signals arrive at the receiving end with approximately equal power, equalizing amplifiers are used. To achieve toll-grade signal-to-noise quality, the L5 system requires equalization of 18 dB/km at the maximum frequency of 6.9 MHz. In addition, automatic adjustments in amplifier gain are needed to compensate for variations in cable attenuation with temperature. Even with those controls, there are still variations in attenuation and phase shift over the 4-kHz bandwidth of a single channel. As a consequence, dedicated conditioned lines must be used when amplitude-shift–keying or phase-shift–keying modems are used to transmit high-speed data over a voice channel. An alternative approach is to use hand-shake modems in which the modems negotiate at the initiation of a session equalization settings for the specific channel in use.

In view of the system complexity, large space requirements, high power consumption, limited capacity, and incompatibility with digital transmission, long-haul coaxial cable systems are being replaced worldwide by optical-fiber links. Cable television, which uses coaxial cable almost exclusively for its distribution system, is also switching over to optical fibers.

5.2.4 Optical-Fiber Communications

Light-wave communication over optical fibers is a major new technology that is having a profound impact on voice and video transmission, telephone systems, computer interconnections, instrumentation, data links, and telecommunication networks. The emergence of this technology was made possible by the dramatic progress achieved during the last 20 years in the development of low-loss and wide-bandwidth optical fibers, semiconductor laser diodes, and sensitive photodiode detectors.

An optical fiber is an all-dielectric waveguide that uses total internal reflection to confine the electromagnetic signal within its core (see FIBER OPTICS). This type of confinement at optical frequencies results in very low signal attenuation and extremely wide bandwidth. Moreover, since the wavelength of the propagating wave is of the order of 1 μm, the fiber cross-section dimensions are small, 125 μm in diameter. A schematic of an optical-fiber link is shown in Fig. 2. It consists of a laser transmitter, which converts the electronic signal into a corresponding optical signal that is coupled into the core of the fiber; a fiber link that guides the light wave to the receiving end; and an optical receiver, which converts the optical fluctuations back to an electronic signal. Thus, the communication, end to end, is electronic, but the link uses light waves for signal transmission. This results in a communication link that offers numerous advantages in comparison with other terrestrial or underwater transmission networks.

1. Optical transmission is immune to any ambient electrical noise, ringing, echoes, or electromagnetic interference and does not generate any of its own electrical noise. Moreover, fiber-optic communication does not present any problems of cross talk.
2. One fiber-optic filament, several mils in diameter, can replace a copper wire several millimeters in diameter. This savings in size and weight is especially important for underwater cables and in areas of overcrowded transmission wires such as in urban areas.
3. Optical cables are safe to use in explosive environments and eliminate the hazards of short circuits in metal wires and cables. Optical systems can be designed to have total electrical isolation. Thus, equipment used now for electronic cables that protects against grounding and voltage problems can be eliminated. Moreover, properly designed optical transmission lines and couplers are relatively immune to adverse temperature and moisture conditions.
4. Most importantly, optical fibers have very low attenuation and extremely wide bandwidth. It is this availability of wide bandwidths that made it possible to convert the common-carrier telephone system from the analog transmission mode to digital signaling and network integration, and, in addition, to increase the overall system capacity from 100 000 voice channels in the late 1960s to the multimillion-channel capacity of today. Also, low attenuation makes it possible to design repeaterless optical-fiber links up to 250 km in length.
5. The capacity of optical-fiber links, once installed, can be upgraded by increasing the bit rate of the transmitted signal (up to 10 Gbps with today's technology) and/or using wavelength division multiplexing; see Sec. 3.1.
6. The development of optical amplifiers made it possible to design "transparent

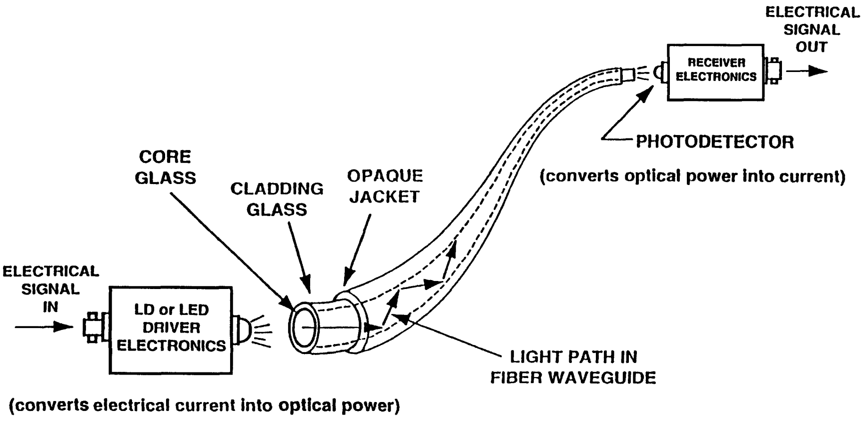
ELECTRICAL SIGNAL OUT
RECEIVER ELECTRONICS
PHOTODETECTOR
(converts optical power into current)
CORE GLASS
CLADDING GLASS
OPAQUE JACKET
ELECTRICAL SIGNAL IN
LD or LED DRIVER ELECTRONICS
LIGHT PATH IN FIBER WAVEGUIDE
(converts electrical current into optical power)

FIG. 2. Basic fiber-optic communication link. (Incoherent Optical Communication Link-Intensity Modulation)

pipe" undersea links of up to 6000 km in length. Such communication links do not use electronic regenerators and thus require no high-speed circuitry (except at the transmitting and receiving ends).

Optical fibers are cylindrical in shape. To confine the light wave, the dielectric inner core of the fiber is embedded within a dielectric material of a lower refractive index. The outside diameter of this cladding is standardized at 125 μm. When the refractive index of the core is uniform, the fiber is known as a *step-index fiber;* the fiber is *graded index* when the refractive index of the core region decreases continuously from its maximum value at the core axis.

The propagation characteristics of an optical signal guided by the fiber are determined by Maxwell's equations. When the core diameter is large, say 50 μm, a multitude of different field distributions, or modes, can propagate, and the fiber is "multimode."

To reproduce data accurately, light pulses must be kept separate and distinct with correct shape and spacing during transmission. In a multimode fiber the signal is spread among the many modes. Since each of the modes propagates with a slightly different speed, they will arrive at the receiving end at different times. This arrival-time variance results in a distorted signal and limits the frequency that can be transmitted. Since "modal dispersion," or pulse spreading, is linear with distance, the fiber is characterized by a length-bandwidth product, LB(3 dB), that is, by the product of its length L by the 3-dB bandwidth B(3 dB) of frequencies for which the transmitted signal power is greater than one-half of the signal power at low, near zero, frequencies. This results in a bandwidth that is inversely proportional to the fiber length. Multimode step-index fibers have a length-bandwidth product of only 10 MHz-km and are therefore of little use. However, multimode graded-index fibers have a length-bandwidth product of 200–1000 MHz-km and are at the present used extensively in the United States for local-area networks and other short-haul links.

When the core diameter of the fiber is reduced to approximately 8 μm, only a single electromagnetic-field distribution or mode can propagate. The fiber is then *single mode* and has, theoretically, at least, infinite bandwidth. All long-haul communication links use single-mode fibers.

Semiconductor lasers are used to convert the electrical signal into an optical signal carried by the fiber. Such sources are not single frequency but have a broad spectrum. For example, a typical laser diode operating at a wavelength of 1300 nm (corresponding to a frequency of 230 TH) has a spectral width of 3 nm, or 530 GHz. Even an expensive distributed-feedback laser diode used for bit rates above 1 Gbps and for cable television links has a spectral width of 0.1 nm, or 20 GHz. Moreover, the spectral components emitted by the source are not correlated in phase. Accordingly, optical communication is different from conventional terrestrial or wireless communication; it is incoherent. That is, what is being modulated by the signal is not the electric field or the magnetic field, but the optical power of the source. Correspondingly, direct detection is used to recover the signal at the receiving end. There is no local oscillator, and there is no mixing. The broad spectral width of the source leads to other consequences as well.

The speed of an electromagnetic signal within a dielectric medium depends on its refractive index, which is wavelength dependent. In addition, when a signal is guided, its velocity also depends on the physical dimensions of the waveguide relative to the wavelength. Accordingly, the different spectral components of the source propagate with slightly different speeds, and the signal will disperse. This is known as chromatic dispersion and limits the fiber bandwidth. The resulting length-bandwidth product of the fiber link depends on two factors: the properties of the fiber and the spectral width of the source. The narrower the source, the greater is the length-bandwidth product of the link.

Figure 3 shows the attenuation and dispersion of silica-based optical fibers including the newer dispersion shifted fibers (DSFs). For single-mode fibers there are two favorable transmission regions within the optical spectrum. The first, the (1300 ± 50)-nm window, is 18 000 GHz wide. Here, the fiber attenuation is 0.4 dB/km, and the chromatic dispersion is nearly zero for conventional step-index single-mode fibers labeled SSMF. Thus, the length-bandwidth product of a fiber link operating in the 1300-nm region is

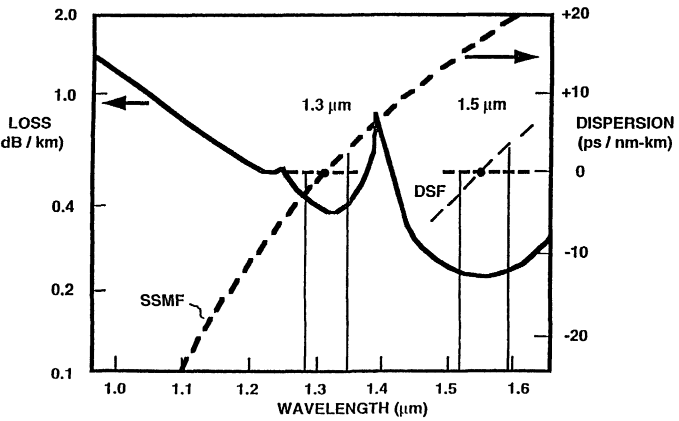

FIG. 3. Single-mode fiber attenuation and dispersion.

very high, regardless of the spectral width of the source. Initially, all fiber links operated in this window. However, lowest fiber attenuation occurs in the (1550 ± 50)-nm window, 12 500 GHz wide. Here, the chromatic dispersion is finite, but the attenuation is only 0.18 dB/km, allowing for an increase in the repeaterless link length to almost 250 km. Sources of very narrow spectral width are needed to operate at gigabit-per-second rates.

There has been extensive research to develop polarization-independent optical amplifiers for the 1300- and 1550-nm spectral regions. The effort became successful with the evolution of reliable erbium-doped optical-fiber amplifiers for the 1550-nm region. In contrast to electronic amplifiers, optical amplifiers raise insignificantly the noise level of the signal. This made it possible to implement passive fiber links up to 250 km in length and to design amplifier-based links thousands of kilometers in length. Such links provide a "transparent pipe" communication link and, with the exception of the transmitting and receiving ends, require no high-speed circuitry. In addition, erbium-doped amplifiers have a broad gain spectrum so that a single amplifier can be used to amplify a multiple-channel wavelength-division multiplexed (WDM) signal and eliminate the need for multiple wavelength-division splitters and combiners at each amplifier site.

It was pointed out earlier that the length-bandwidth product of optical fibers is inversely proportional to the product of the spectral width of the source and the chromatic dispersion constant of the fiber at the particular wavelength of operation. This seems to indicate that the fiber bandwidth becomes infinitely large when the spectral width of the source is reduced to zero. However, even a monochromatic source when modulated at, say, a 1-Gbps rate will result in a signal with a spectral width of 2 GHz. Accordingly, the length-bandwidth product with chromatic dispersion depends not only on the spectral width of the source but also on the information bandwidth. High-speed long-haul links must therefore use fibers that are nearly dispersion free at 1550 nm. All newly installed long-haul routes use such "dispersion shifted" or DSF single-mode fibers. In 1996, AT&T installed two transatlantic fiber cables, each 6000 km in length, and two transpacific cables, each 12 000 km in length. In addition, to complete the loop, a submarine cable was installed between England and Japan. Each of the five cables uses six dispersion-shifted fibers and in-line erbium-doped optical amplifiers (without regenerators), and is designed to operate at an initial rate of 5.0 Gbps to carry 64 512 duplex voice channels per fiber pair. Additional submarine cables to be installed in 1998 will use six WDM fibers at 5.0 Gbps each to carry 518 096 voice channels per fiber pair. It is of interest to note that for proper operation, the transpacific link requires a length-bandwidth product that is in excess of 30×10^{12} Hz-km. This is six orders of magnitude higher than the bandwidth that can be achieved with high-grade coaxial cables.

To facilitate global communications and to establish a flexible architecture capable of accommodating applications with a variety of transmission rates, an international Synchronous Optical Network (SONET) fiber transport standard was introduced by Bellcore in 1985. It establishes a standard multiplexing format using any number of 51.84-Mbps signals as building blocks. The lowest level is OC-1 with a data rate of 51.84 Mbps; OC-48 has a data rate of 48 × 51.84 Mbps = 2488.32 Mbps. SONET also defines a synchronous multiplexing format for carrying lower-level digital signals. The synchronous structure allows one-step bit-interleaved multiplexing and demultiplexing, and greatly simplifies the interface-to-digital switches and add-drop multiplexers.

6. THE TELECOMMUNICATIONS INDUSTRY AND ITS FUTURE

The two most important trends in modern telecommunications technology are toward the supporting of multiple services on broadband facilities and the use of wireless communication to untether the user from being wired to the network. Local-area networks have proliferated. Because they operate at rather high data rates, 3–100 Mbps, they are difficult to interconnect over long distances. Moreover, enhanced services beyond voice and data, such as image and video communication, have increased the demands. Too, the increasing use of the client–server approach to computing makes rapid communication essential. Fortunately, the increasing prevalence and high performance of fiber-optic transmission (up to 3 Gbps) make it plausible to support the required transmission. Switching is another matter. For most purposes that involve the processing of information on fiber links, the signal has to be converted from light to electronic forms for processing. This coupled with the high rates at which fibers operate make switching the bottleneck. Ultimately, the solution will be the development of optical processing so that the information can be processed in its form as light without conversion to electronic form. The erbium optical amplifiers discussed earlier are the first such devices. However, a full complement of optical processing devices is some time off. Developing such devices may require new, fundamental physical insights. In the near term, new protocols and standards are being developed to reduce the load on the switches. Not only is modern fiber-optic communications carried out at very high rates, but the channels are comparatively error free. This makes it possible to perform less error checking. The X.25-type packet switching of the 1970s and early 1980s usually involved error checking across each link of a network as well as end to end. The communication methods developed to meet the challenges of high-speed optical channels carrying integrated services do much less error control—detection and correction. The asynchronous transfer mode (ATM) protocol has been developed to reduce the load on the communication switches and to support real-time broadband services such as video conferencing. The primary data unit for this protocol is a 53-byte cell, 48 bytes of information and 5 bytes of overhead. Every cell is exactly the same length and format. This makes it possible to accomplish most of the switching functions in hardware. The small size reduces the switching latency so that high–data-rate real-time traffic can be supported. Much of the flow control and error control of the previous generation (X.25) of packet switching is missing in ATM because the relatively error-free performance of fiber-optic channels does not require it. It is believed that ATM will be able to support all services at very high-speed, broadband data rates.

The immense popularity of the Internet and its exponential growth—the number of computers (hosts) on the Internet has doubled annually for the last 15 years—will provide the demand for the high-performance and broadband networks made possible by fiber optics. In addition, users will demand the flexibility of wireless terminals, and so the technologies discussed here are not merely driven by what is technically possible but they will find a market as well.

GLOSSARY OF ACRONYMS AND TECHNICAL TERMS

Amplitude Modulation (AM): A method of superimposing a signal on a carrier wave by varying (modulating) the amplitude of the carrier wave.

AMPS: Advanced mobile phone system.

Analog Data: Data represented by a continuously variable physical quantity that is directly related to the data.

Asynchronous Transfer Mode (ATM): A special form of packet switching designed for high-speed networks. Transmission is based on small, fixed-size (53 octet) packets called cells. This reduces latency in switches so that multiple modes of information can be transmitted, including real-time video.

Attenuation: A decrease in the amplitude of a signal in transmission from one point to another.

Baseband Transmission: A method of using a transmission facility where the information is applied directly to the facility without the use of a carrier.

Carrier: A single frequency that is impressed with a second information-bearing signal by perturbing the characteristics of the carrier such as its frequency, amplitude, and/or phase.

CDM: Code-division multiplexing.

CDMA: Code-division multiple access.

Circuit Switching: A method of routing traffic through a communication network where a dedicated path is reserved from source to destination and bandwidth is guaranteed for the duration of the information transfer.

CT-2: Second-generation cordless telephone.

Cyclic Redundancy Checksum (CRC): A method for detecting errors in transmitted data. Additional redundant information in the form of the remainder of a modulo-two division is added to the information being transmitted to allow the detection of errors in transmission by the receiver performing a consistency computation.

DAMA: Demand assignment, multiple access.

Digital Data: Data consisting of a sequence of elements chosen from a discrete set of values.

Dispersion-Shifted Fiber (DSF): Optical fiber designed to make the region of minimum dispersion coincide with the region of minimum attenuation at 1550 nm.

Echo Canceler: A device to identify and filter out echoes.

Echo Suppressor: A device that turns off reverse transmission in a channel to eliminate echoes.

Equalization: Techniques to compensate for distortion due to frequency dependence of attenuation and/or delay.

FAMA: Fixed assignment, multiple access.

FCC: Federal Communications Commission (U.S.).

Free-Space Path Loss: The loss in signal strength in wireless communication resulting solely from the geometric divergence of the radiation as a function of distance.

Frequency Diversity: A feature of wireless systems where the received signal is selected from or derived from several signals that are transmitted on different carriers to provide more robust communication.

Frequency-Division Multiple Access (FDMA): The sharing of a transmission facility by dynamically allocating channels derived by frequency-division multiplexing to users.

Frequency-Division Multiplexing (FDM): The sharing of a transmission facility by dividing its bandwidth into narrower bands, each of which is used to carry a subchannel.

Frequency Modulation (FM): A method of superimposing a signal on a carrier wave by varying (modulating) the frequency of the carrier wave.

Frequency-Shift Keying (FSK): A method of superimposing a digital signal on a carrier wave by varying (modulating) the frequency of the carrier wave.

GSM: Global System for Mobile or Général Speciale Mobile.

Impulse Noise: Random noise characterized by high amplitude and short duration, often due to vehicle ignitions, lightning, and various other kinds of sparking phenomena.

Intersymbol Interference: The impairment in the transmission of digital information resulting from the overlap due to distortion of adjacent digital symbol representations.

LEO: Low Earth orbit for satellites.

Modem (Modulator/Demodulator): A device for converting a digital signal into analog form (modulation) for transmission and converting the analog form back to digital (demodulation) upon reception.

Modulation: The process, or its result, of perturbing characteristics such as amplitude, frequency, or phase of a carrier to transmit a message signal.

Multipath: In wireless transmission, the superposition of direct and reflected signals with varying delays.

Packet Switching: Routing of data in networks in blocks of data called packets that have a fixed upper bound in size. Each packet has its own address to specify its destination. Channels may be used by packets of different communications in an interleaved fashion. Packets are successively read into the memory of switches and then retransmitted to the next switch on the path from source to destination.

PCS: Personal communications services.

Phase-Shift Keying (PSK): A method of superimposing a digital signal on a carrier wave by varying the phase of a carrier wave among a finite set of phases.

Pixel (Picture Element): A bit-mapped image is represented by a rectangular grid of points. Associated with each point is a color value, black/white, gray-scale value, or color, depending on the type of image. Each of these points with the associated color value is a pixel.

Pulse Code Modulation (PCM): A two-step method for representing analog information digitally. In the first step, the analog information is sampled at a rate higher than twice the bandwidth of the information. Next, the samples are discretized into a representation by a fixed number of bits.

Quadrature Amplitude Modulation (QAM): A modulation scheme for sending binary information on an analog channel using a combination of phase-shift and amplitude-shift keying.

Single-Sideband Transmission: Transmission where one of the two sidebands of a carrier signal is suppressed.

SONET: Synchronous optical network.

Space Diversity: A feature of wireless systems where the received signal is selected from or derived from several signals that are physically separated in order to provide more robust communication.

Space-Division Multiplexing: A method for sharing communication resources where signals are separated by physical separation.

Statistical Multiplexing: A transmission facility shared by allocating time slots to channels upon demand using time division. The sum of the peak rates of the channels may be larger than the rate of the shared facility, but not the sum of the average information rates.

Thermal Noise: Ubiquitous noise resulting from the thermal movement within matter.

Time-Division Multiple Access (TDMA): The sharing of a transmission facility by dynamically allocating channels derived by time-division multiplexing to users (usually in wireless communication).

Time-Division Multiplexing (TDM): The sharing of a transmission facility by allocating it sequentially and in rotation among the users.

Wavelength-Division Multiplexing (WDM): A name for frequency-division multiplexing in the optical region of the spectrum.

Works Cited

Cerf, V., Kahn, R. (1974), *IEEE Trans. Commun.* **COM-22,** 637–648.

Huffman, D. A. (1952), *Proc. IRE* **40,** 1098–1101.

Rivest, R. L., Shamir, A., Adleman, L. M. (1978), *Commun. ACM* **21,** 120–126.

Shannon, Claude (1948), *Bell Syst. Tech. J.* **27,** 379–423 and 623–656.

Shannon, Claude (1949), *Bell Syst. Tech. J.* **28,** 656–715.

Ziv, J., Lempel, A. (1978), *IEEE Trans. Info. Theory* **IT-24,** 530–536.

Further Reading

Bertoni, H. L., Honcharenko, W., Maciel, L. R., Xia, H. (1994), *Proc. IEEE* **82,** 1333–1359.

Cover, T., Thomas, J. A. (1991), *Elements of Information Theory,* New York: Wiley.

Freeman, R. L. (1987), *Radio System Design for Telecommunications,* New York: Wiley.

Freeman, R. L. (1996), *Telecommunication System Engineering,* 3rd ed., New York: Wiley.

Jayant, N. (1992), *IEEE J. Sel. Areas Commun.* **10,** 796–818.

Holzman, G. J., Pehrson, B. (1995), *The Early History of Data Networks,* Los Alamitos, CA: IEEE Computer Society Press.

Lebow, I. (1995), *Information Highways and Byways,* New York: IEEE Press.

Lee, E. A., Messerschmitt, D. G. (1994), *Digital Communications,* 2nd ed., Boston: Kluwer Academic Publishers.

Lee, W. C. Y. (1993), *Mobile Communications Design Fundamentals,* 2nd ed., New York: Wiley.

Rappaport, T. S. (1991), *IEEE Commun. Magazine,* November, 52–71.

Salus, P. H. (1995), *Casting the Net,* Reading, MA: Addison-Wesley.

Simmons, G. J. (Ed.) (1992), *Contemporary Cryptology,* New York: IEEE Press.

Stallings, W. (1997), *Data and Computer Communications,* 5th ed., Upper Saddle River, NJ: Prentice-Hall.

U.S. Congress, Office of Technology Assessment, *Wireless Technologies and the National Information Infrastructure,* OTA-ITC-622, Appendix A, Washington, DC: U.S. GPO.

Viterbi, A. J. (1995), *CMDA, Principles of Spread Spectrum Communication,* Reading, MA: Addison-Wesley.

Special Issue on Mobile Radio Communication (1988), *IEEE Trans. Veh. Tech.* **37** (1).

TELESCOPES

See ASTRONOMICAL TELESCOPES; RADIO TELESCOPES

TELESCOPES, SPACE

See Supplement Volumes

TEMPERATURE CONTROL TECHNIQUES AND INSTRUMENTATION

EUGENE P. DOUGHERTY, *Research Laboratories, Rohm and Haas Company, Bristol, Pennsylvania, U.S.A.*

Introduction 539
The Concept of Temperature 539
Motivations: Why Good Temperature Measurement and Control Is Needed 539
Example of the Importance of Temperature: Molecular-Weight Control in Polymers ... 540
1. **Measurement Techniques** 541
1.1 Thermocouples 541
1.2 Resistance Temperature Detectors 544
1.3 Thermistors 545
1.4 Filled Thermal Systems 546
1.5 Bimetallic Thermometers 546
1.6 Infrared Temperature Sensors 546
1.7 Signal Conditioning and Transmitters 548
1.8 Sensor Selection Criteria, Applications, and Case Studies 550
2. **Control** 551
2.1 Standard Feedback-Control Techniques: On/off and Proportional–Integral–Derivative Controllers 552
2.2 Advanced Control Techniques: Feed-Forward and Model-Based Approaches 555
2.3 Applications: Two Very Different Temperature-Control Approaches 556
Acknowledgments 559
Glossary 560
Works Cited 560
Further Reading 562

INTRODUCTION

The Concept of Temperature

One of the most frequently asked questions in sci.physics, the physics user interest group on the Internet, is "What is temperature?" Fortunately, as the author responsible for sci.physics explains, we have a good intuitive concept of temperature: "Our intuitive notion is that two systems in thermal contact should exchange no heat, on average, if and only if they are at the same temperature." The author then describes the total energy of the system, the many possible internal states of a thermodynamic system, and division of energy in the two states that are in thermal contact.

This concept is a useful one, because it defines temperature by examining not one, but two systems in thermal contact. For temperature measurement and control, this is useful to bear in mind, because there are always at least two systems in thermal contact any time one wishes to measure and/or control temperature. Examples: a mercury thermometer (system 1) is used to determine the extent of a person's fever (system 2); thermocouples use the contact between the junction (system 1) and the object (system 2) to measure an object's temperature accurately; a water jacket (system 1) is used to cool controllably a hot reactor (system 2).

Motivations: Why Good Temperature Measurement and Control Is Needed

In many areas of science and industry, temperature measurement and control is crucial. Our bodies are a good example. Think of how poorly human beings would function if their body temperatures increased (or decreased) a few degrees away from the 37 °C set point. Weather conditions, which,

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

of course, are virtually impossible to control but very extensively measured, determine everything from how we dress to leave for the day to who might win an important political election.

Not only is temperature measurement and control important for its own sake; it is also important for what it can tell us about physics, chemistry, biology, or engineering. Calorimetric measurements, measures of heat and/or heat flux, are usually made by comparing two or more temperature measurements, often closely spaced in terms of time or physical location. Therefore, since the basic calorimetric measurement is a temperature difference, high accuracy and precision are required. Recently, calorimetric measurements have become very sophisticated (Mettler, 1991; Basalik, 1993), allowing scientists and engineers to determine accurately, for very complex processes, reaction-kinetic parameters (Moritz, 1988; Schmidt and Reichert, 1988), heat-transfer coefficients (Machado, 1994; Kumpinsky, 1992), mixing efficiency (Fasano *et al.*, 1991), and engineering scale-up parameters (Choudhury and Utiger, 1992).

The above examples are enough to illustrate the point of why good temperature measurement and control can be important. In what follows, a single specific scientific problem is examined in detail, to trace exactly why and how critical these measures and control approaches can be.

Example of the Importance of Temperature: Molecular-Weight Control in Polymers

Temperature is almost always a critical variable in controlling the molecular-weight distribution for polymers. Generally, the goal of a polymer scientist is to produce polymers with well-characterized (and usually high and narrow) molecular-weight distributions. This leads to well-characterized, well-defined polymer materials, with the required strength and resilience to meet applications needs. Sometimes, however, the properties required suggest that it is best to make a polymer consisting of a mix of different molecular-weight distributions; for example, narrow, short-chain polymers mixed with a second polymer exhibiting a much higher and broader molecular-weight distribution are usually good adhesives.

To get a low molecular-weight distribution, free-radical polymerizations are carried out at high temperatures, where many radicals produce many short chains. Conversely, to get high molecular-weight polymers, the same sort of polymerizations can be run, but at temperatures just low enough to produce a few free radicals, which in turn produce a few long chains of polymer. This is usually more difficult, because few radicals can also mean slow, sluggish kinetics.

Industrial processes involve many variables, many products, and many requirements, such as requiring rapid, safe, stable production of high molecular-weight polymers at very high yields. Such requirements make the task of producing a high molecular-weight polymer an even more daunting task, with many subtleties involved. Temperature measurement and control becomes even more crucial.

The article by Sack *et al.* (1973) describes some of these subtleties and provides reasons for why temperature control is so crucial. Their analysis of key experiments concludes, for instance, that those temperature variations that maximize the breadth of molecular-weight distribution are step changes between the minimum and maximum allowable temperatures in a system in order to achieve a bimodal distribution of sizes (Sack *et al.*, 1973). Moreover, these authors conclude that, on average, temperature variations have a much greater effect on breadth of molecular-weight distributions for chain-addition polymerizations than either the residence-time distributions or micromixing. Finally, they discuss different types of polymerizations and how temperature for these different types of processes should be tuned or manipulated in time so as to achieve just the right distribution of molecular weights.

Besides polymerization processes, there are many other areas where temperature control is even more critical: other reaction engineering disciplines, biotechnology, nuclear engineering, aerospace, materials processing, separation technologies, fermentation processes, and agriculture. The reference book *Temperature Control Principles for Process Engineers* (Dougherty, 1993) contains many such examples. Space constraints prohibit providing detailed examples from each of these areas of technology; but whenever possible, specific examples from these and other areas will be integrated into the following sections on measurements and control. In discussing applications, commercial products are occasionally mentioned, but this article by no means wishes to recommend one product over another. The point of the applications is to stress the main reasons to do good temperature measurement and control: to do or make something of value better, cheaper, and/or faster, or to do the kind of research that will allow others to improve how things are made or done at some time in the future.

1. MEASUREMENT TECHNIQUES

For the reasons outlined above, it is not surprising that temperature was one of the first variables ever to be measured in the process field. Over the years, it has been sensed in just about every way imaginable. Virtually any physical property that changes with temperature has, at one time or another, been used as a basis for this measurement (Smith, 1995). Many basic phenomena form the basis for most temperature measuring instruments. These include

1. the thermoelectric effect;
2. changes in electrical resistance;
3. radiant energy; and
4. changes in physical dimension or characteristics of liquids, metals, and gases.

Among the many methods still used in industry today are thermocouples (T/Cs), resistance temperature detectors (RTDs), thermistors, electronic temperature sensors, bimetallic devices, filled devices, infrared (IR) devices, and acoustic pyrometry (Smith, 1995). This article examines many of these methods in detail to explain how they work and to provide ideas on how and under what conditions each can be applied. In addition, examples of their use and applications to solve important problems are also provided.

1.1 Thermocouples

An electric current flows in a closed circuit consisting of two dissimilar conductors joined at both ends if the two junctions are at different temperatures (Anon., 1989a). Thomas Seebeck made this discovery in 1821 (Katz, 1993). If the circuit is broken, the net open-circuit voltage (the Seebeck voltage) is a function of the junction temperature and the composition of the two metals. That is, the magnitude of the voltage that causes current flow is a characteristic of the two metals and a function of the temperature difference, $T - T_R$. If T_R is known, the temperature T is determined for a given pair of metals by measuring the voltage developed (see Fig. 1). The voltage output is of the order of millivolts. A millivoltmeter, actuated by the current in a circuit of known resistance R, measures the voltage, which is related to the current I and resistance R by Ohm's law: $V = IR$. Manufacturers calibrate for a certain thermocouple resistance. They strive for the highest internal resistance consistent with the ruggedness required for measurements under all sorts of environments (Anon., 1992).

Output of a T/C is a function of the temperature difference, $T - T_R$. If the temperature is to be known accurately, the reference temperature must be known or its effect compensated for. Compensation shifts the

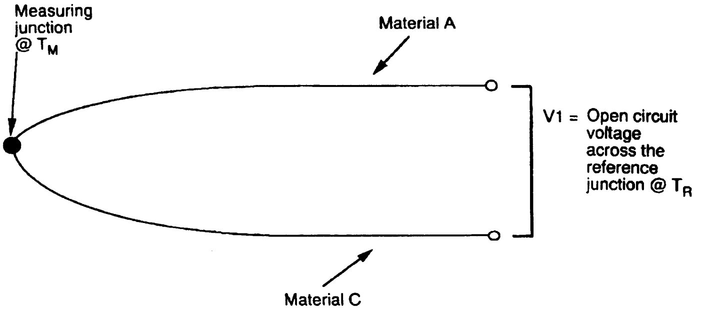
Measuring junction @ T_M
Material A
V1 = Open circuit voltage across the reference junction @ T_R
Material C

FIG. 1. Basic thermocouple (courtesy of Katz, 1993).

zero of the instrument such that it reads T_R whenever there is no input to the instrument.

Thermocouples are flexible; they are used to make sensors of special, often unusual shapes. They are generally recommended for use at very high temperatures or where temperature must be measured over very wide spans. When installing T/Cs, attention must be paid to the type of wire connecting the probe to the instrument. Special extension wire is used to ensure high measurement accuracy. Best accuracy is achieved when the measuring target is close to the junction point.

Thermocouples are characterized by the different types of materials used as the two dissimilar conductors. Listed in Table 1 are many of the most common types, together with the range of temperatures over which they are applicable for measurement (Hall, 1990; ASTM, 1987).

Thermocouple wire is usually color-coded in a standard way for the different thermocouple types; for example, type J thermocouple wire is usually red (positive) and white (negative). The different materials may be chosen because of environmental or corrosion considerations; for example, K- and E-type thermocouples are not recommended for sulfurous, reducing, or strongly oxidative environments (Wang, 1991).

Tables exist to show the variation of voltage with temperature for these and other different types of thermocouples. Generally, the voltage difference for a 1 °C temperature change is in the range of microvolts; this tiny voltage difference means that high accuracy is not easy to attain with T/Cs. Accuracy is typically only about ±2 °C, although this can be improved by using special high-accuracy probes and careful compensation. Some typical calibration tolerances are listed in Table 2 for selected thermocouple types (ASTM, 1987). At subzero temperatures, these tolerances will be slightly worse.

Table 1. Basic thermocouple types, materials of construction, and temperature ranges of applicability.

Type	Materials	Temperature range
B	Pt-30% Rh vs. Pt-6% Rh	0 to 1820 °C
E	Ni-10% Cr vs. Constantan	−270 to 1000 °C
J	Iron vs. Constantan	−210 to 1200 °C
K	Ni-10% Cr vs. Ni-5% Al, Si	−270 to 1372 °C
R	Pt-13% Rh vs. Pt	−50 to 1768 °C
S	Pt-10% Rh vs. Pt	−50 to 1768 °C
T	Cu vs. Constantan	−270 to 400 °C
N	Ni-14% Cr-1.5% Si vs. Ni-4.5% Si-0.1% Mg	−270 to 1300 °C
WRe	W-5% Rh vs. W	−270 to 2316 °C

Table 2. Specifications for tolerances for various types of thermocouples.

Type	Standard tolerance (°C or %)	Special tolerance (°C or %)
J	±2.2 or 0.75	±1.1 or 0.4
T	±1 or 0.75	±0.5 or 0.4
E	±1.7 or 0.5	±1 or 0.4
K	±2.2 or 2	±1 or 1.0

The average wire diameter is dictated by the thermocouple and the required measurement response time. Some suggested upper limits for the average wire sizes for various protected thermocouples are listed in Table 3 (Anon., 1989a).

T/C construction varies and is dependent on the required application. Figure 2 shows the wide variety of shapes and sheaths for various applications. The flexibility of thermocouples allows designers to fabricate them to make measurements at very high- or very low-temperature regimes, at the surface of a material, in highly corrosive environments, around corners, or at very special locations. There are ceramic-beaded, insulated, and metal-sheathed, mineral-insulated T/Cs. Ceramics allow use at very high temperatures but lack some flexibility because of reduced mechanical strength and slow response times. To protect thermocouples from highly hazardous or corrosive environments—say, those encountered in certain chemical

Table 3. Suggested upper temperature limits (°C) for various thermocouple types at two common wire diameters.

T/C	20 gage (0.8 mm diameter)	28 gage (0.3 mm diameter)
T	260	200
J	480	370
E	540	430

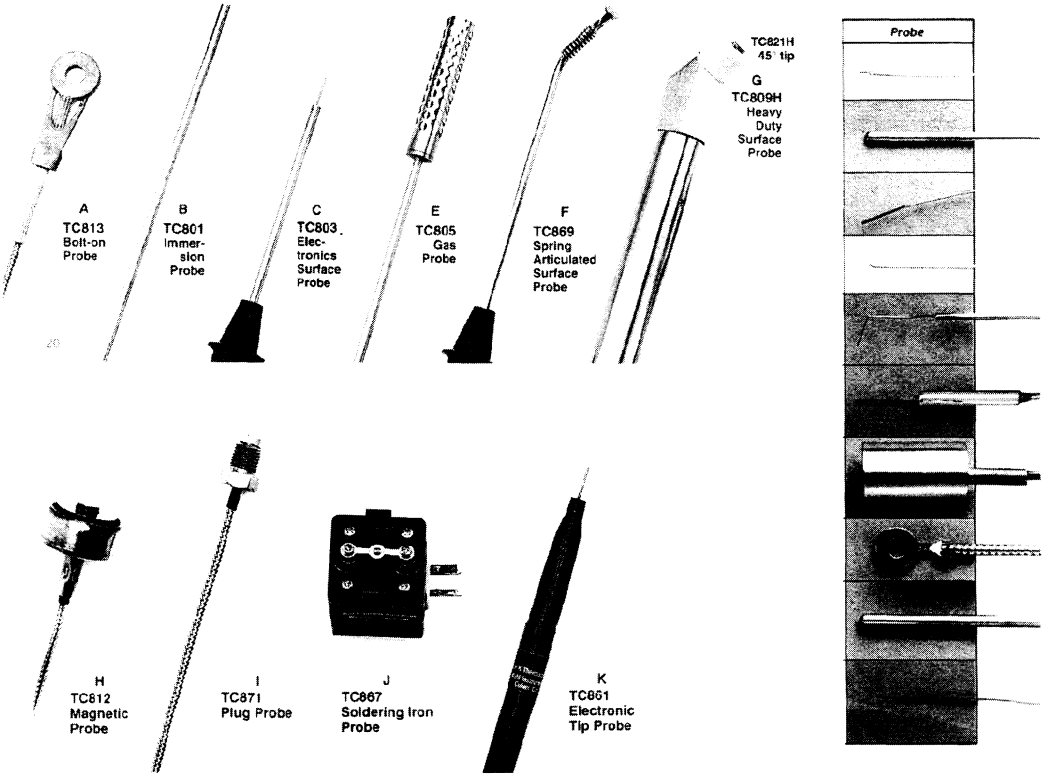
A
TC813
Bolt-on Probe
B
TC801
Immersion Probe
C
TC803
Electronics Surface Probe
E
TC805
Gas Probe
F
TC869
Spring Articulated Surface Probe
TC821H
45° tip
G
TC809H
Heavy Duty Surface Probe
Probe
H
TC812
Magnetic Probe
I
TC871
Plug Probe
J
TC867
Soldering Iron Probe
K
TC861
Electronic Tip Probe

FIG. 2. Thermocouples come in a wide variety of shapes and sizes to meet the required application needs (from Wahl, 1995).

applications—thermowells, which are simple, closed-end tubes, are used to encase the thermocouples.

Figure 2 shows some special-purpose thermocouples known as surface T/Cs. Some surface T/Cs use a special thermally conductive adhesive that allows them to be strapped onto pipes in order to infer temperatures inside those pipes very accurately (Anon., 1995). Alternatively, by using thermowells to sheath the thermocouples having the same thermal properties as the surface wall, surface T/Cs can be very useful for making accurate heat transfer or calorimetric measurements (Fasano *et al.*, 1991).

Because accuracy and precision are key issues with T/Cs, the very low electrical signals involved require the use of some special techniques. One article (Hashemian and Peterson, 1993) describes how to carry out cross-calibration of thermocouples, account for moisture and inhomogeneities in circuits, use redundant measurements effectively, provide noise analysis, and use empirical correlations—all steps to improve thermocouple accuracy.

To improve thermocouple accuracy and precision, a reference thermocouple should be provided at an appropriate, known temperature to compare measurements with. Originally, the "reference" was provided using an ice-water bath maintained at 0 °C; so the practice became known as "cold"-junction compensation (Liptak, 1993; Samuels, 1995). Nowadays temperature compensation is done electronically, using appropriate software. In Sec. 2.6, compensation and other temperature-signal conditioning approaches will be treated in greater detail.

Some very special thermocouples can be used for high-temperature applications, such as measuring temperature in incinerators, ovens, and furnaces. Special-purpose T/Cs, however, may lack the signal-conditioning devices needed for compensation, linearization, or calibration.

1.2 Resistance Temperature Detectors

Industrial resistance temperature detectors (RTDs) are generally constructed of platinum, nickel, or copper. The basis for measurement is knowledge of how the resistance of these materials changes with temperature; an accurate resistance measurement yields an accurate temperature (see Fig. 3). Calibration is done directly in terms of temperature. The range of temperature measured depends on the type of construction, the type of insulation used for the sensor, and only secondarily on the type of metal used. The manner of winding the wire to avoid the effect of mechanical strains is also important in determining the range of temperature measurement (Anon., 1992).

Despite its high cost, platinum is the most popular material for RTDs; its resistance change with temperature is 0.0390% per °C. Less expensive RTDs can be constructed from Balco (an iron–nickel alloy), nickel, and copper. However, nickel has stability problems, and both Balco and nickel have problems with reproducibility. The problem with a typical copper RTD is that it takes up a lot of space (Clayton, 1993).

Resistance thermometers are inherently accurate; repeatability can be as good as ±0.02 °C; accuracy, better than 0.05 °C. The ability of these thermometers to measure temperature accurately over very narrow temperature spans, say, less than 0.5 °C, is their greatest attraction for critical process control applications. Four-wire measurements can be especially accurate and precise (Lewis, 1995). One instrument company claims to be able to calibrate RTDs to an accuracy of ±0.001 °C (Labs, 1989).

Most circuits (Lewis, 1995) used with resistance thermometer bulbs are similar to the classical Wheatstone bridge circuit (see Fig. 4). Two slidewires driven in tandem are used. Slidewire No. 1 maintains the arm of the bridge containing resistor C at a value equal to the bulb resistance. Compensation of the connecting wire is theoretically perfect. Perfection is limited by the accuracy of matching the resistance for lead wire L_2 with that at L_3. The battery voltage does not affect calibration but does affect sensitivity. Instead of batteries, nowadays solid-state devices, together with ordinary power supplies, are used to minimize failures and maintenance (Anon., 1992). Finally, four-wire RTDs, which use a current source and an additional lead wire, provide more accurate temperature measurements because they reduce the (nonlinear) effects of lead-wire length and resistance on voltage (Anon., 1989a).

An RTD is usually constructed as a fixed length of fine resistance wire wound on a glass or ceramic mandrel. However, recently, thin-film (and even thick-film) platinum RTDs are challenging the wire-wound types in certain applications. Wire-wound types, the most accurate types, are common in military/aerospace, food, beverage, chemical,

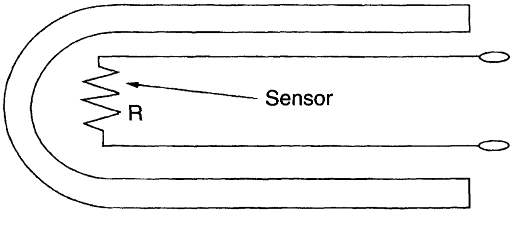

FIG. 3. Schematic of a resistance thermometer (courtesy of Katz, 1993).

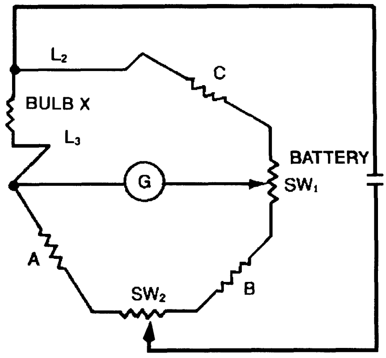

FIG. 4. A Wheatstone bridge used as a typical resistance bridge circuit for an RTD. The symbols L_i denote the three lead wires used; Sw_j denotes the two slidewires used to match the resistances (see text). G denotes a meter for measuring the voltage.

Table 4. Advantages of thermocouples and RTDs.

RTD advantages	T/C advantages
High accuracy; greatest over wide span	Modest accuracy
Relatively narrow spans (5 °C can be measured)	Very wide range (0–1000 °C)
Reproducibility unaffected by temperature change	Rugged
Short-term reproducibility better than T/Cs	Remote and versatile mounting
Compensation not necessary	Moderately stable
Handle suppressed ranges	Moderate cost
Relatively small sizes	Small sizes available also

Table 5. Disadvantages of thermocouples and RTDs.

RTD disadvantages	T/C disadvantages
More expensive	Low output signal
Limited above 1500 °F	Calibration changed by contamination
Less rugged	Electrical interference can be a problem
Self-heating can be problem	Calibration nonlinear over normal spans
Vibration can affect wire-wound types, whereas thin-film types are expensive	Higher installation costs (due to the need for special extension wires)

utility, plastic, and rubber applications. On the other hand, RTDs are not commonly used in metal processing, glass ceramics, or petrochemical applications. The reason: upper temperature limits of RTDs (about 850 °C) (Anon., 1989a) are well below the everyday working temperatures in these industries (Clayton, 1993).

Wire-wound RTD response times are fast (1–4 s), but the fastest thermocouples can respond much more rapidly. Still, with platinum (Pt) sensors and new, technologically advanced thin-film technology, it is now possible to create sensors that withstand 5 N of force and get Pt 100-Ω film elements down to $1.5 \times 5.0\ \text{mm}^2$. Response times can be as low as 0.07 s in water, or 5 s in air moving at a velocity of 1 m/s.

Ease of use is one factor that has led to RTDs growing in popularity relative to T/Cs. RTD signal levels are ten times higher than T/C levels (250 μV vs. 25 μV), so that they are less susceptible to noise. RTDs require neither compensation nor special lead wires. While T/Cs are generally priced lower, the extra accuracy afforded by RTDs can make costs nearly comparable in some applications.

On the other hand, making RTDs with stability, ruggedness, and accuracy is a complex process. One must take into account thermal expansion characteristics, material compatibility, heat transfer, ability to seal properly to prevent moisture penetration, and purity (it is especially important for platinum to be pure).

For temperature measurements in extremely cold environments (30–300 K), germanium resistance thermometers (GRTs), carbon-glass thermometers, platinum RTDs, and rhodium–iron resistance thermometers can all be used effectively. Germanium and carbon-glass sensors are nonmetallic RTDs. Below 30 K, special germanium-type sensors, while very expensive, can exhibit reproducibilities on the order of ±0.001 K (Labs, 1989).

Since RTDs and T/Cs are both popular means of measuring temperature, Tables 4 and 5 compare the advantages and disadvantages of RTDs and T/Cs (Kane, 1990). Table 6, a summary of typical RTD and T/C specifications, compares and contrasts these two popular methods of temperature measurement (Anon., 1992).

1.3 Thermistors

While most industrial resistance thermometers are RTDs, thermistors are resistance temperature sensors in which the resistance *decreases* with temperature, instead of

Table 6. A Summary of key thermocouple and RTD specifications.

Characteristic	RTD	T/C
Stability	0.005–0.05 °C	0.5–5 °C
Drift	Less than 0.1% in 5 years	0.5 °C/year
Range	−200–900 °C	−150–1100 °C
Sensitivity	0.2–20 Ω/°C	10–100 μV/°C
Output	1–6 V	0–60 mV
Power (100-Ω load)	0.04 W	2×10^{-7} W

increasing as most metals do. Additionally, the resistance change with temperature in thermistors can be enormous. When a device is required to have extreme sensitivity to changes in temperature, thermistors are to be recommended (Stockdale, 1992; Anon., 1992).

Thermistors are very sensitive probes that actually have not been very widely applied industrially. The best operating characteristics are for temperatures below 100 °C. Because they are so "sensitive" (negative temperature coefficient thermistors change as much as 4% per °C), they are often thought of as inaccurate, unstable devices. Yet, when used in a leg of an accurate Wheatstone bridge circuit, their sensitivity can be used to advantage to measure, precisely and accurately, very small temperature differences, mainly over narrow temperature spans (Trietley, 1989).

1.4 Filled Thermal Systems

Filled thermal systems can provide effective, efficient, and economical temperature measurement and control, provided such basic principles of selection, design, and installation are followed (Anon., 1992). Simple, inexpensive mercury thermometers can be both very accurate and very precise. Of course, response times can be very slow.

A filled-system thermometer makes use of the principle that most liquids and gases expand or try to expand as their temperature increases. A liquid or gas of known characteristics is totally enclosed, and the amount of expansion in the case of liquids or the pressure change resulting from the attempt to expand are measured. This gives a reading directly related to the temperature of the liquid or gas.

A bulb is exposed to the temperature being measured. It is connected by capillary tubing to a long, coiled, filled tube, known as a bourdon tube. Various types of materials are used as fill materials, depending on the application, but mercury, of course, is very commonly used.

A filled thermal application requires adherence to the following guidelines:

1. The right bulb sensing element for the temperature range to be measured must be selected.
2. The appropriate bulb shape and type must be chosen to allow proper installation, bulb immersion, and bulb protection; the smaller the temperature span, the larger the required bulb.
3. The bulb should be located in the right spot, i.e., exactly where the temperature is to be measured.
4. The bulb must be immersed completely in the medium of interest.
5. The bulb must be placed such that the medium surrounding the bulb and the material in the bulb (usually mercury) reach thermal equilibrium as quickly as possible.
6. The bulb should be protected from physical disturbance, damage, or sources of corrosion.

1.5 Bimetallic Thermometers

Bimetallic thermometers are also a useful, simple, and popular means to measure temperature accurately. Generally, bimetallic thermometers consist of a metal stem, which contains the bimetal sensing element wound up into a helical coil and located near the end of the stem. The metal stem is physically connected to a dial to provide an analog indicator of the temperature. The metal contacting the object or fluid whose temperature is to be measured is usually stainless steel and/or some other corrosion-resistant material. Accuracy is typically about ±1–2% of the scale reading of the dial. Typical bimetallic temperature ranges are −20–200 °C. Response times are reasonable: 40 s is typical. Some typical applications include boilers, cabinets, tanks, drums, piping, air conditioning units, refrigerators, and so on (Reotemp, 1987; Anon., 1989a). Bimetallic thermometers have limitations because they do not easily hook up to electronic computers and control systems and are somewhat limited in range and response times. Still, they represent a very accurate, low-cost, general-purpose sensor appropriate in a variety of different temperature measurement and control applications.

1.6 Infrared Temperature Sensors

Infrared thermometers, also called infrared pyrometers, measure the temperature of an object without touching it. This is useful when the object to be measured is in motion or too far away, too delicate, or too hot to touch. Radiation thermometry is often used in applications where contact thermometers fail. The major challenge is to minimize or compensate for such signal effects as target or background emittance, intervening flames, dust, and so on.

Radiation pyrometers make use of the fact that radiant energy is constantly being emitted by all materials (Love, 1993). The emitted radiation varies in a known way with temperature. Tables and graphs exist that show the shape of the radiated spectrum vs. temperature and wavelength. The peak power density moves to shorter wavelengths as temperature increases (see Fig. 5).

In practice, the energy from a hot object is collected by a lens and focused on a thermopile. This energy is a function of the absolute temperature of the hot object and its characteristic emissivity. Output of these devices is usually a voltage that varies with temperature raised to the fourth or fifth power.

Mathematically, the amount of this radiated power is determined by the following equation:

$$W = E\sigma T^4, \tag{1}$$

where W = spectral radiant exitance (radiation) (units of watts per square centimeter), E = emissivity, σ = Stefan–Boltzmann constant (5.67×10^{-12} W/cm^2), and T = temperature (K).

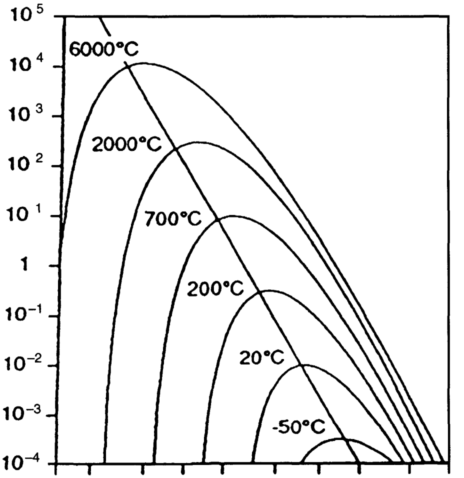

FIG. 5. Radiation power density decreases as a function of wavelength. Here maximum power densities for selected temperatures are shown.

Some appreciation should be gained for the dependence of radiation power on temperature. For example, a blackbody (emissivity of 100%) at 30 °C would have a radiation density of 5.4 mW/cm^2. That same blackbody at a temperature of 150 °C would have a radiation density of 139.2 mW/cm^2 (Walcutt, 1994).

Critical to the accuracy of an IR thermometer is the emissivity of the surface being measured. Emissivity is the measure of how effectively a surface radiates relative to an ideal, known as a blackbody radiator. A "blackbody radiator" (which is a material that exists only in theory) emits at 100%. Rough, porous materials, such as wood, paper, and powder, exhibit high emissivities ($E > 70\%$), while very shiny materials, such as stainless steel or aluminum, have low emissivities ($E < 40\%$). A mirror has the lowest emissivity—nearly 0. Conventional IR devices are designed and calibrated to theoretical blackbody conditions at a single ambient temperature; emissivity shifts must be accounted for by calibration at multiple temperatures. Accuracy can be as good as ±1%, although this is only possible with careful, repeated calibrations.

Originally, primitive optical pyrometers worked by visually comparing the color of a heated filament with that of the object to be measured. While this was a noncontact method, its disadvantages were that objects had to be hot enough to emit visible light, and human visual comparisons led to high measurement variability. Modern-day industrial radiation pyrometers measure the energy collected by a calibrated receiver electronically and work over both visible and invisible (IR) wavelengths (Carocci, 1995; Land Infrared Company, 1995).

Today, there are several types of radiation thermometers. Total radiation, or broadband, sensors measure the entire spectrum of energy emitted by the target. These require corrections for obscuring, interfering objects like smoke, as well as additional corrections for emissivity shifts. Single-wavelength, monochromatic sensors measure the intensity of radiation at one or a narrow range of wavelengths. They are accurate so long as the emissivity characteristics are well known at that wavelength. Finally, multiple-wavelength instruments can be used, and multiple-point and line-scan instruments can be used to determine an entire temperature profile over, say, a moving conveyor belt.

Not surprisingly, IR thermometers have been able to take advantage of technological advances in fiber-optic technology (Huber, 1993) and various computer-based technologies. For example, the use of flexible fiber optics allows users to locate targets that are obstructed or difficult to reach and, more important, provides immunity to electromagnetic interference (EMI). Moreover, while the first IR detectors had to be scanned mechanically, nowadays focal-plane or staring arrays (FPAs) can eliminate mechanical scanning (see Fig. 6). Using special, microprocessor-based technology, up to 78 000 detectors can be placed on one chip. Finally, computer-based statistical and image-analysis techniques have all been found to be useful in analyzing multiple-point, multiple-wavelength measurements (Finney, 1995).

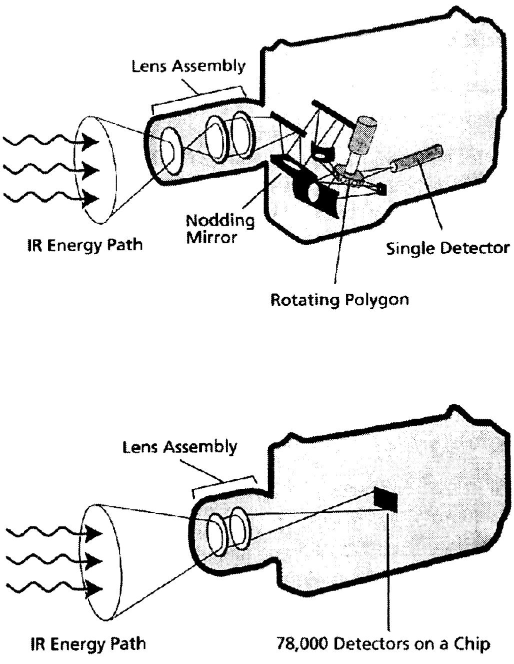

FIG. 6. Schematics for infrared thermometry cameras. Top: a mechanically scanning camera. Bottom: a staring array (focal plane) camera with 78 000 detectors on a single computer chip. (From Carocci, 1995.)

IR imaging systems produce a picture of the invisible thermal pattern of a component or process, using electronic cameras that see this thermal radiation just as a human eye sees visible light (Pompei, 1995). Such imaging systems exploit the power of multiple-point, multiple-wavelength, infrared thermography, despite its high cost. These measurement systems allow one to locate and monitor problems quickly and then to present critical decision-making information in a simple, visual form. IR imaging allows you to "see" the "hot spots" or temperature variations in complex equipment or machinery. Of course, with IR detection and imaging systems, a special type of resolution is an important specification. The resolution of the temperature of the image is a function of the sensor accuracy, the object's emissivity, the camera angle, geometry, and focus, and the distance of the camera from the object.

While the IR price per measurement is high, especially with respect to T/Cs, recent technological advances and improved manufacturing efficiencies have expanded the market for such sensors to make them cost competitive, at least on a life-cycle basis. Ultimately, the decision to use a single-point, multiple single-point, or multiple-point line scanner must be based on time, material, and energy savings, quality improvements, safety, and efficiency vs. the high sensor cost (Snell, 1995).

Typical applications of IR thermography include vacuum furnaces, induction heating, welding, and semiconductor processing. Table 7 includes a few general guidelines on choosing the right radiation thermometer (Tenney, 1988).

1.7 Signal Conditioning and Transmitters

Once a temperature signal is obtained, temperature control may be dependent on modifying or "conditioning" that signal electronically and transmitting that signal to a computer or controller. Filled thermal systems rarely use signal conditioning, but most other types do require some type of electronic manipulation. We have already introduced the concept of signal conditioning for thermocouples in discussing cold-junction compensation (Samuels, 1995). In practice,

Table 7. Guidelines for choosing the right radiation thermometer.

Material/Condition	Measuring solution
High-temperature metal, oxidizing, emittance changes	Use short wavelengths or a ratio thermometer
Glass annealing with flames	Use 4.8 to 5.2 μm wavelength
Plastic films	Use wavelength tuned to proper frequency

cold-junction compensation means that one must make electrical connections to an additional T/C junction, using the same T/C metal pair, opposite in polarity to the signal being measured. What is really measured by the electronics is the difference between the two thermocouple voltages. By convention (and for convenience), one T/C is maintained at 0 °C, so that the output voltage of the system is the output voltage of the original thermocouple. Actually, this cold-junction sensor could be a T/C, a current source, or a monolithic voltage-output temperature sensor. The errors introduced by the electronic signal conditioning as well as the accuracy of the cold-junction measurement itself may be issues with the accuracy for the desired measurement.

Inexpensive personal computers or smart microprocessors can be used with T/Cs and RTDs to improve their accuracy and precision immensely. Such devices can provide software for linearization, built-in lookup tables for popular T/C and RTD types, accurate cold-junction compensation, multiple channels for multiple temperature readings, and noise filtering to improve accuracy of repeated temperature readings. Software can be used for line-cycle averaging (filtering): several measurements per line cycle can be made to cancel positive and negative noise excursions.

In using personal computers, thermocouple leads must be appropriately shielded, and special electrical isolation techniques must be used to prevent ground-loop or common-mode errors. The computer itself must be ruggedized for its environment or properly enclosed to protect it from environmental contaminants.

Other important considerations for purchasing signal conditioners include input-to-output isolation, input protection, T/C cold-junction compensation for any and all T/C types that may be used, linearized RTD inputs, zero suppression, ambient operating range, output short-circuit protection, alarms, environmental shock, vibration, dust, military specifications, and so on. Safety can also be a concern for temperature sensors. If there is an electrical short or an electrical storm or a T/C junction "breaks," appropriate failure logic must be implemented. "Thermocouple break protection" is used to ensure that a control system will not result in a hazardous situation in response to the broken thermocouple. Fortunately, manufacturers of control and data-acquisition equipment often include hardware and software for temperature probes for safe operation, signal conditioning, and filtering (Schraff, 1995).

Transmission of the conditioned signal is also an integral part of many temperature measurements. "Smart transmitters" can be directly wired to modern transmitter units using the latest microprocessor-based digital technology to change low-level signals to high-level current signals. Use of low-cost, unshielded, twisted-pair copper wire in place of shielded cables and conduits and costly T/C extension wire often makes sense. Smart transmitters can yield overall lower total ownership cost; for, relative to conventional transmitters, they exhibit a longer lifetime, higher accuracy, and higher stability. Some smart transmitters accept many different types of sensor inputs and even conduct self-diagnostic tests to determine possible mechanical or maintenance problems (Sheng and Nodrop, 1990). Many modern temperature transmitters accept RTD and T/C signals, 4–20 mA, or digital-process variable signals on a field bus. Field bus (Berra, 1996) provides reduced wiring costs and improved accuracy. New process-control standards, such as SP88 (Sperber, 1995), should be consulted before designing the actual measurement and instrumentation systems for large-scale manufacturing plants (Craig, 1995).

Transmitters and signal conditioners may have special needs, depending on the environment they are subjected to. Some may require overvoltage protection, while others may only be useful over a limited range of environmental operating temperatures. Resistance to shock, vibration, and explosion, tightness of seal to moisture and/or dust, isolation capabilities, linearity, and the different types of sensors accepted may also be important. The turndown ratio for any device, which is the ratio of the highest value of a measurement that can be counted on to be accurate to the lowest accurate measurement value, is also an important consideration. Finally, the electronics must not be affected by stray radio-frequency signal interference—say, from passers-by carrying cellular phones or walkie-talkies. For these reasons, some people prefer the old-fashioned analog, rather than microprocessor-based, digital, transmitters.

Finally, once a signal has been conditioned and transmitted to a computer, modern-day personal computers allow users to take advantage of special data-processing aids. These include multitasking operating systems, such as Unix or Windows NT, and the dynamic data exchange (DDE) and object linking and embedding (OLE) capabilities of Windows 95 and 3.1x. Such software aids allow temperature signals from a variety of different signals to be easily compared with results from other processes and analytical measurements in real time, so that the appropriate digital filtering, signal averaging, and/or control can be applied (Wang, 1991).

1.8 Sensor Selection Criteria, Applications, and Case Studies

Temperature-measuring devices—thermocouples, RTDs, thermistors, IR thermometers, and filled thermometers—can be used for many applications, ranging from aerospace to zoology. The Works Cited and Further Reading abound in interesting and useful applications of these devices. An important goal is to use the sensor best suited to the measurement requirement of the process of interest.

In contact temperature sensing, the four most important selection criteria are response time, accuracy, size, and type of environment in which the probe will be used. Of course, these criteria often work against one another. A small probe will provide rapid response time, but may be fragile (Schraff, 1995). Also, if the probe must be protected from a harsh environment, the response time may be slowed.

Generally, stainless steel probes are a good choice for most applications, although even stainless steel may not always provide ample protection. For example, one manufacturer was using a 316 stainless steel thermowell to measure temperature of a 35% solution of caustic soda, but the thermowell had to be replaced about once every 15 days. The problem was not direct chemical corrosion, but instead, galvanic corrosion promoted by the chemically initiated electrical reaction between the thermowell and the vessel. The solution was to use a Teflon coating on one side of the thermowell.

If response time is a problem, one solution to improve response time is to make the sensor smaller. One way to do this is to use thin-film technology with platinum RTD sensors.

Noncontact sensors, while more expensive than T/Cs or RTDs, may be better suited to more difficult process-measurement requirements (Maczka, 1988). This section provides three applications of IR imaging technology: two from semiconductor manufacturing and one from the steel industry. IR imaging measurement examples are very well documented in the literature. This is likely because the IR imaging systems are so expensive that they are used mainly in applications for which less expensive T/Cs or RTDs are in some way deficient.

In many fields, heat can be a contributor to unreliability. In electronics, a key reliability factor is semiconductor junction temperature. The higher this temperature, the shorter the time to failure. When operating, a semiconductor is dissipating power and generating heat. So long as the device is generating heat, the temperature of the junction will rise in relation to the capacity of the path to conduct heat away from the device. At some point, heat flow from the device will be equal to the heat being generated, and thermal equilibrium will be reached. Thermal images can provide measurements of semiconductor case temperatures. Assuming one knows the value for thermal resistance, the junction temperatures can be determined (Walcutt, 1994).

A second and less well-recognized thermal problem occurs in manufacturing when circuit boards are reflow soldered. Today, many companies are experiencing problems with solder connections. Newer designs with their fine pitch and high component densities yield circuit boards that experience problems due to solder joint failures. This is particularly true for boards that go through reflow soldering. As in all solder processes, temperature is critical to the quality of solder connections. Not only is absolute temperature important; so is the uniformity of temperature across the board. If the thermal profile across the board is not uniform, the flux will not be cleaned, and the solder will not properly reflow. The results are inconsistent interconnect integrity and reduced reliability. Because of the high temperatures to which they are subjected and their complex structure, some boards become misshapen or warped. Others exhibit mechanical problems with components or solder connections due to thermal stresses. IR reflow ovens assume that a fairly homogeneous object is being heated, while in fact the temperatures across the board are frequently not uniform. In one application, an IR thermal image of a circuit board just removed from a reflow machine showed a large swing in temperature from minimum to maximum, even though the temperatures were expected to be uniform, since the board was just removed from the oven. Instead, the thermal image showed this board's temperatures exhibiting a ΔT of more than 70 °C. The value of seeing the whole temperature presentation on one display shows the power of thermal imaging (Williamson Corporation, 1995).

In the steel industry, noncontact, noninvasive IR measurements can save lives because they identify potentially catastrophic events before they happen. This industry uses IR cameras periodically to survey refractory in teeming ladles and gondolas that transport molten metal in steel making. Refractory, a heat-resistant ceramic metal, deteriorates over time, leading to burn-through and tons of molten metal being spilled near personnel. IR imaging can quickly and accurately determine the condition of equipment from a distance without process interruption (Carocci, 1995).

2. CONTROL

The purpose of any control system is to maintain one or more variables in a production process at a desired value. For example, keeping the temperature in an oven at 170 °C might be an important aspect of a commercial bakery. Sensors measure the temperature using one of the means described in the previous sections. The signal from the sensor is transmitted to the comparator, or error detector, within the controller. The comparator determines if there is any difference between the value measured by the sensor and the desired value. The desired value is usually referred to as the set point. Any difference between the actual value measured by the sensor and the set point is known as the error. When an error exists, the controller detects the magnitude of the error signal, makes a decision based on the error and transmits a signal to the final control device to correct the error in the controlled variable. For example, the controller may call for an increase in the flow of gas when the bakery oven temperature falls below 170 °C. In such a case, the flow of gas is the manipulated variable. The action of the controller produces a change in the manipulated variable causing a change in the temperature, the controlled variable.

Control systems are usually based on the principle of negative feedback: there is an error detected by a temperature-measuring system, and so the controller takes action to counteract that error. That is, a feedback-control loop uses the value of the error signal to make an appropriate change in the value of the manipulated variable. The new value of the controlled variable is continuously monitored by the sensor, whose signal is fed into the control system to complete the feedback loop. In this way, the control system acts continuously to maintain the controlled variable within acceptable limits.

Block diagrams (see Fig. 7) are often used to represent the actual pieces of hardware where these variables change from one form of energy or material flow to another (Shinskey, 1993). Figure 7 shows the concept of negative feedback control, indicated by inversion of sign at the summing junction on the block diagram. The input to the final control element is usually an electronic or pneumatic signal. The value of this signal determines whether the valve opens or closes. Block diagrams usually have one or more "comparators," which are used to compare two or more simultaneous electronic or pneumatic input signals. Often one input is the set point; the other may be the value of the measured variable. In the comparators,

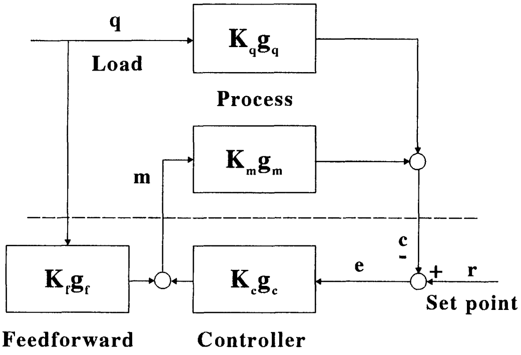

FIG. 7. A typical control block diagram, illustrating the load or disturbance, the process, the process controller, and the comparator for the set point. The *K*'s refer to gains, while the *g*'s refer to "process transfer functions" relating the variables for the load, the measurement, and the controller. (From Shinskey, 1993.)

one input is subtracted from another to determine the error. Although systems vary in complexity, they can all be represented using this simple method of connecting blocks with lines and arrows.

2.1 Standard Feedback-Control Techniques: On/off and Proportional–Integral–Derivative Controllers

On/off control is the simplest form of control, used by almost all domestic thermostats (Williams, 1995). On/off controllers are generally used either in processes that respond slowly to changes in the manipulated variable or in cases where the controlled variable can tolerate some deviation from the set point. On/off control may be used to turn on or off a furnace to keep the temperature of a house within reasonable limits. When the room gets cooler than the set-point temperature, the furnace is switched on. Once the room gets hotter than the set-point temperature, the furnace is switched off entirely. The turn-on and turn-off temperatures are deliberately made to differ by a small amount to prevent noise from switching the furnace rapidly and unnecessarily when the temperature is near the set point.

To control processes that respond quickly to changes in the manipulated variable or those that require fairly precise control of the controlled variable, it is necessary to have more than just on and off control. For example, in a continuous production process heated by steam, steam flow should be regulated by setting the control valve so that the steam flow will maintain the temperature at the set point. This means that the control valve must be set at some position between fully opened and fully closed. The control system can then make adjustments to the valve position—opening it more or less fully, to correct for variations from set point.

A proportional controller should perform better than an on/off controller, since it applies a change to the manipulated variable (e.g., a resistance heater) in proportion to the difference in temperature between the actual temperature and the set point (Williams, 1995).

The output of a proportional controller output is given by

$$O = Ke, \tag{2}$$

where the controller gain K is typically less than 1.0 and e is an appropriately normalized error signal (e usually is equal to the set point minus the measured variable). Sometimes the gain is expressed as the proportional band,

$$P = 100\%(1/K); \tag{3}$$

P (often expressed as PB) is the percentage change in controller output signal resulting from a process deviation. A 200% PB means that a 10% change in measurement signal produces a 5% change in controller output signal.

As the gain increases for a proportional controller, the system responds faster to changes in set point but becomes progressively underdamped and eventually unstable: The set point can be drastically overshot. Another problem with a proportional controller is that, even after a steady, controlled condition is established, there is some offset: The error never really reduces to zero. A proportional controller will always have an offset when working under load. This offset is $P(m - b)/100$, where P is the proportional band, m the manipulated variable, and b the output bias.

The stability and overshooting problems that arise when a proportional controller is used at high gain can be mitigated by adding a term proportional to the time derivative of the error signal:

$$O = Ke + K_d \frac{de}{dt}. \tag{4}$$

This technique is known as proportional–derivative (PD) control. The value of the damping constant, K_d, can be adjusted to achieve a critically damped response to changes in the set-point temperature.

The adjustment for derivative action is called the rate adjustment. If this adjustment is set too low, the control system will not respond quickly enough to load changes. If the rate action is too high, the response to even small changes in the controlled variable will be too great and the control system will become unstable. Instead of progressively reducing errors in the controlled variable, derivative action may actually compound them. The controller will need to be carefully tuned to prevent this from happening.

Although PD control deals neatly with the overshoot and ringing problems associated with proportional control, it does not cure the problem with the steady-state error. Fortunately it is possible to eliminate this by adding an integral term to the control function. The equation then becomes

$$O = K\left(e + K_d \frac{de}{dt} + K_i \int_0^t e\,dt\right), \tag{5}$$

where K_i, the integral gain parameter, is sometimes known as the controller reset level. The controller is known as a proportional–integral–differential, or PID, controller. The effect of the integral term is to change the manipulated variable until the time-averaged value of the temperature error is zero.

A PID controller is often called a three-mode controller. The three modes use the same error signal but each in a different way. The proportional mode uses information about the magnitude of the error. The derivative mode uses information about the rate of error change. The integral mode uses information about the average error over a specified period of time. In nonmathematical terms, the integral unit monitors the performance of the proportional unit. If it finds that there is an offset, it shifts, or resets, the proportional control action. Because of its ability to return a system to its set point, integral action is also known as reset action.

PID controllers work well, but the mathematical analysis of a PID controller is complicated since the system is third order. Unlike second-order systems, third-order systems are uncommon in physics, but the methods of control theory make analysis of them quite straightforward. By applying the so-called Routh–Hurwitz stability criterion, which is a systematic way of classifying the complex roots of the auxiliary equation for the model, one can show that, provided the integral gain is kept sufficiently small, parameter values can be found to give an acceptably damped response with the temperature error eventually tending to zero if the set point is changed by a step or linear ramp in time. Whereas derivative control should improve the system damping, integral control, especially at high gain, eliminates steady-state error at the expense of some stability (Williams, 1995).

In its raw form, integral control can be a mixed blessing: If the error $T_S - T_0$ is large for a long period—for example, after a large change in the set point or when the controller is switched on—the value of the integral term can become very large. The control action can result in overshooting or undershooting the set point and may take a long time to recover. More sophisticated controllers prevent excessive integral action, using aids like "antireset windup" usually implemented in the control software. The "windup" problem with some electronic controllers occurs because electronic controllers use capacitors for integral action, which can only store a maximum charge and thus become saturated. This saturation causes integral action to be deficient. Antireset windup ensures stability by only implementing the full integral action when the system is fairly close to equilibrium (far from saturation).

Any system that uses a resistive electrical heater to control temperature is inherently nonlinear, since an electrical heater can only generate, not absorb, heat. Thus such heaters artificially act as "integrators." When the oven temperature is higher than the set point, cooling occurs at a rate that depends on the oven and its temperature, not the controller. Dual PID controllers or special "split-range" controllers allow different heating and cooling parameter values to cope with such special conditions.

Commercial PID controllers may use parameters that do not exactly correspond to the parameters defined in this section. Values are sometimes specified in software in such a way that a long integral time constant is equivalent to a low value of K_i, whereas a long derivative time constant means a large value of K_d. The proportional gain is sometimes set by choosing a proportional band (PB), which is the fractional change in temperature that gives maximum change in the manipulated variable, so that a small number for this corresponds to a large K value. Commercial PID controllers have a variety of features; for example, some provide alarms or take special action if a temperature probe malfunctions; some provide digital, while others an analog temperature display; and most provide control for a variety of different sensor types. The commercial literature (Wahl, 1995; Anon., 1989a; Larson, 1995) is useful for obtaining a controller with the exact features and specifications required by the application of interest.

Once a controller is purchased, knowing "what buttons to push" in a three-mode, PID controller requires expertise to ensure that such a controller is tuned properly with the proper balance of control and robustness. First, the frequency domain behavior of the system should be assessed, preferably by mathematical simulation. The open- (uncontrolled) and closed-loop (controlled) frequency response for the system should be simulated, if possible. As the controller gain increases, the control system responds faster. This, of course, can cause the overshoot problem described earlier for proportional controllers. With three-mode controllers, the magnitude of the overshoots may be high. Moreover, the controller's response in time, which may be modeled as a "phase lag" between the time the disturbance takes place and the time that the controller responds, can produce a large overshoot resulting in a sharp, resonant peak. This derives from the characteristic frequency-domain response. Any additional lags in the system, for example, a nonnegligible time constant for the sensor to measure and transmit the signal, must also be accounted for in tuning to avoid overshoots or undershoots.

There are standard approaches for tuning PID controllers. In many ways, piano tuning and controller tuning are alike (St. Clair, 1991). A deep understanding of mathematics is not really required for either, but understanding *frequency* is. A piano tuner knows that the note A above middle C must be 440 Hz; other notes follow. Similarly, a controller must be tuned according to the dominant frequency of the process. This is especially true with three-mode controllers.

Ziegler and Nichols (1942) suggested the following approach to controller tuning. First, increase the controller gain until the control loop begins to cycle (Anon., 1989b). The period of the cycle is the natural period for the process. Formulas for the controller settings are based on this period and the gain that caused the cycle. In practice, this is done by placing the controller in manual and by making a small change, a step change, in the controller output. Response of the measured variable to this step change is made. After a delay, the measured variable will change, and, if the system is well behaved, it will eventually level out. Some process-variable gain will be produced, after some delay, or "dead time."

To understand dead time, imagine a freight train starting to move. All train cars are coupled. When the engine starts, the caboose does not start right away. It must wait some dead time while the other cars get going. Similarly, when the controller output moves, the controlled variable does not move immediately, but must wait for coupling elements in between to move. The output lags the true measurement, producing dead time, or the natural period of the process. Knowledge of the dead time or natural period and the gain can be used to assign the parameters of the PID controller using the methods calculations provided in the key articles (Ziegler and Nichols, 1942). The Ziegler–Nichols approach allows some overshoot, but the amplitude of the oscillation should decay rapidly, with the second peak being on the order of 25% of the first peak. So usually the Ziegler–Nichols PID parameters require some fine tuning, just as a piano does. More sophisticated tuning methods have been suggested (Gerry, 1991; Horton, 1991; Polonyi, 1989), but the simple Ziegler–Nichols tuning methods, properly applied in well-designed processes, can yield surprisingly good control. Autotuning methods are usually proprietary methods developed by commercial process-control companies, but it appears that controllers that tune themselves place themselves into an on/off control mode, then use Ziegler–Nichols tuning rules to refine the PID parameter estimates.

In the analysis of PID and related controllers, it is often useful to transform a controller from the time domain to the frequency domain. Laplace transforms and/or Fourier transforms of the time-domain control equations described here make it easier to assess the frequency response characteristics, i.e., resonant behavior or singularities. Laplace transfer-function representations, in particular, are convenient for tuning a controller to balance performance with robustness. This is very well illustrated in a recent technical article, which discusses tuning PID controllers for batch chemical reactors (Campbell, 1993).

2.2 Advanced Control Techniques: Feed-Forward and Model-Based Approaches

While feedback control is adequate for many temperature-control requirements, it has deficiencies. The biggest is that, with simple feedback control, the controller must wait for a change in the average error, current error, or error vs. time derivative before taking action. If tuning is done properly and if disturbances are relatively small and predictable, feedback control may suffice. But it certainly makes sense, whenever possible, to use knowledge of the process in order to control it. Process knowledge leads to a "proactive" control strategy, feed-forward control, in which the controller takes the appropriate action whenever there is a change in the system that one would predict to change the temperature. In other words, information is "fed forward," usually by some sort of computer-based model, to make the adjustment to control against an incoming load or disturbance.

There are many different types of models, but they can be roughly divided into three broad categories:

1. theoretical, physicochemical-based models,
2. empirical, statistically based models, and
3. heuristic or knowledge-based models.

Some models are mixtures of one or more of these general types, while others use both feed-forward and feedback elements.

Theoretical models provide profound physical insight to the process but are often the most time-consuming and difficult to develop. Process responses can be difficult to predict. For example, chemical processes are notoriously multivariable, nonlinear, nonstationary, and subject to disturbances, constraints, noise, instrument, and transport dead times. Even when a good model is developed, such models will never predict the outcome of every conceivable disturbance, and so these controllers often still require feedback elements as a "trim" for the feed-forward prediction.

Generally, theoretical-based models for control use a dynamic model (Lee and Sullivan, 1988; Rhinehart and Riggs, 1990; Wu and Dougherty, 1993):

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(\mathbf{y},\mathbf{u},\mathbf{d},\mathbf{k}), \tag{6}$$

where all variables listed are vectors. Here the transfer function **f** may be nonlinear or complicated, **y**'s are the variable(s) to be controlled, **u**'s are the manipulated variable(s), **k** is a vector of process parameters, and **d** is a vector of process disturbances.

The finite-difference approximation used by the controller is given in one dimension by

$$(y_{sp} - y_0)/\tau = f(y,u,d,k), \tag{7}$$

leading to the generic model law, expressed as

$$f(y,u,d,k) + K_1(y_0 - y_{sp}) + K_2\int_0^t (y_0 - y_{sp})dt = 0, \tag{8}$$

where *sp* stands for set point for *y*. The control parameters K_1 and K_2 are related to the parameters used in a PI controller, the proportional gain and the integral action, respectively. Model-based controllers should be implemented using on-line computers and can be updated periodically using on-line process data to keep the model tracking the process closely. While the controller looks very much like a PI controller, this process model has the advantage that it can at least approximately predict the process's response to the actions taken by the controller, over at least a short time. The better the model, the more accurate the prediction.

In internal model control (IMC), the difference between the model output and the process value serves as the feedback signal to the IMC controller. The transfer function of the IMC controller is just a single parameter, known as the filter time constant (Chia and Lefkowitz, 1991; Studebaker, 1995). This time constant "filters" the model output values in real time to match the real-time, measured process values reasonably closely. With just this one parameter, the IMC controller is relatively straightforward to tune to get a good balance between model prediction and performance and stability or robustness as conditions change.

Statistically based models use actual process data to predict a process dynamic model, which, in turn, can be used to generate a statistically based controller. Process data, of course, is correlated in time, so that the model generated is not a simple regression model, but instead, a time-series model. Often, a process that is in reasonably good control is "perturbed" from its steady state, either intentionally or unintentionally, by making deliberate changes in the manipulated variable u. Then, a statistically based transfer function model is constructed to predict the effect of a change in u on the control variable y. There may be one or more process "gains," and several significant delays or dead times, some short and some long, but all can be expressed as time constants or critical frequencies. Often, the statistically based models are relatively simple, like first order plus dead time. Such a structure may suggest the structure of the underlying physical model of the process.

Once a transfer function is known, the remaining "noise" (or deviations of the process from the model) also usually has a time-series structure; that is, the noise is often autocorrelated in time as well. By examining both the transfer function and the noise model, one can, using well-developed, statistically based analysis, construct rigorous, semiempirical controllers. MacGregor and others (MacGregor, 1972; Box and Jenkins, 1976; Dougherty *et al.*, 1989) have developed the mathematical equations for complex, multivariate, nonlinear models and have shown that such controllers developed from these equations can perform well but still be robust and stable. Dougherty *et al.* (1989) used a statistical, time-series–based approach to develop a controller, and then they approximated that statistically based controller as a PID so that simple control instrumentation hardware could be used to implement it.

Finally, "heuristic" or knowledge-based models can also be used in modern-day control schemes. Such controllers may involve a series of "rules" preset by "experts," or, alternatively, use directed "process learning." There are many approaches to such models. Almost every digital control system involves computer programs that involve rules developed by process "experts," and so expert system models are probably the most common knowledge-based models around. On the other hand, neural networks are being used in the control of very complex, highly nonlinear, multivariate systems. These, however, require large training sets for "process learning." Finally, "fuzzy logic" is now being used in "smart instruments" for control (George, 1996; Abdelrahman, 1990). Depending on the circumstances and a set of prespecified rules, fuzzy logic controllers may take one or more of several possible courses of action to maintain a good, stable robust process that remains under control. There are no "hard and fast" rules to follow but very often a "fuzzy" solution to handle all the "gray areas" characterizing complex systems.

2.3 Applications: Two Very Different Temperature-Control Approaches

There are many temperature-control applications: the Further Reading and Works Cited abound with them. However, this section outlines just two applications for the purposes of reinforcing the concepts introduced in Secs. 2.1 and 2.2.

The first application is for control of a jacketed, batch, industrial chemical reactor (Beatty, 1988; Verhaegen, 1991). Jacketed reactors are commonly used in chemical, polymer, and fermentation processes (see Fig. 8 for a typical control setup for a fermentation reactor). Most jacketed reactors are controlled using a cascade control scheme (Fig. 9). Control is called "cascade" because one controller cascades into the next; one feed-

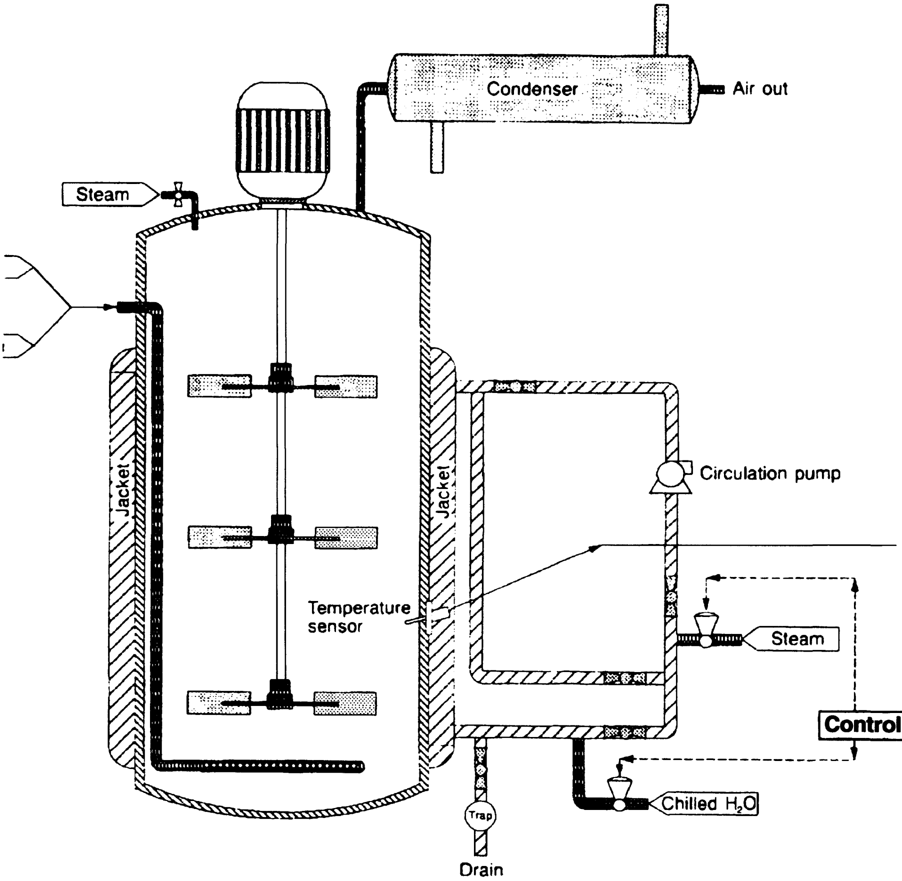
Condenser
Air out
Steam
Jacket
Jacket
Circulation pump
Temperature sensor
Steam
Controller
Chilled H₂O
Trap
Drain

FIG. 8. Temperature control system used in a fermentation process (From Lorbert and Pike, 1993).

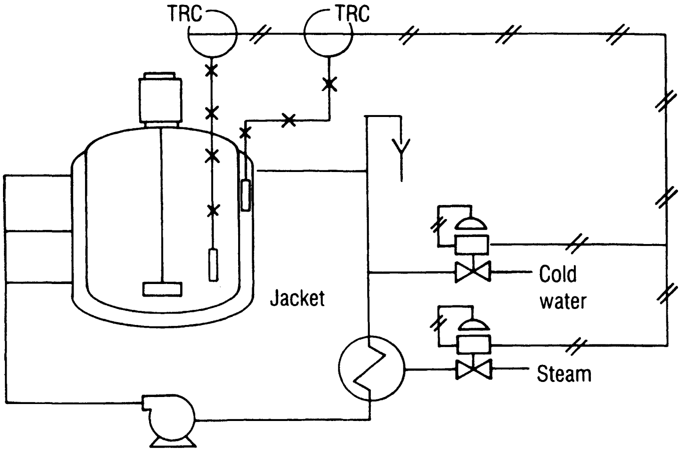
TRC
TRC
Jacket
Cold water
Steam

FIG. 9. Schematic for cascade control of a jacketed reactor. The temperature of both the batch reactor and its jacket are measured, and the signals transmitted to control valves regulating steam and cold water. Steam is added to a heat exchanger to increase the temperature, if necessary, while cold water is added to the jacket to reduce the temperature. Regulating the jacket temperature, by cascade control, regulates the batch temperature in a more stable fashion. (From Shinskey, 1993.)

back loop is "nested" inside another. In short, the process being controlled is split into two parts, and an intermediate variable for the inner loop is chosen, such as the jacket pressure or jacket outlet temperature. In cascade control, there is frequent manipulation of the set point of one controller by another; a secondary or so-called "slave" controller is responsive to the output of the primary or "master" controller.

The main reason that cascade control is used to control the temperature of large, industrial batch reactors is that these temperatures generally move very slowly in response to disturbances, which tend to be very frequent. The batch temperature is, of course, the primary cascade variable. But the choice of the secondary variable is key. Often, it is the position of a valve to let refrigerated water into the jacket. To summarize, in cascade temperature control, the temperature of the reactor (the variable that is most important to control) varies slowly with a very long time constant. When this reactor is cooled by a jacket, the jacket temperature (representing the secondary control "loop") moves more quickly in response to disturbances and generally exhibits a shorter time constant.

There are some "rules of thumb," i.e., a knowledge-based system, for determining when to use cascade control. Basically, when a number of various disturbances might be expected and overall process dynamics are slow, application of cascade should be considered. The following should be considered carefully in determining whether or not to use cascade control:

1. when a process reacts very slowly to a change in the valve position;
2. when the process value (i.e., the temperature) drifts around set point;
3. if process changes appear as disturbances on measured value (i.e., temperature);
4. if one or more variables (for example, jacket temperature or pressure) can be directly related to the measured set-point temperature; and/or
5. when, with a change of set point, quick parallel tracking between recorded and desired value is required with minimum overshoot.

Once a cascade system has been designed and is starting to work, there are some additional rules of thumb to ensure that the cascade controller works well:

1. the secondary variable should receive the maximum disturbances (there will be less chance to disturb the primary control variable, whose trajectory will move smoothly and stably toward the set point, with minimal overshoot);
2. the secondary loop must include "worst" disturbances to be expected;
3. there must be quick reaction to each change (i.e., the time constant for the secondary scale should be one fifth (i.e., faster than) that of the primary scale, so that it can rapidly respond to disturbances, with the primary variable responding more slowly and thus remaining stable);
4. the secondary controlled variable must have direct influence on the primary (e.g., a change in steam supply pressure should affect reboiler temperature);
5. there should be a directly proportional relation between the quantities in the primary and secondary loops; for example, a change in, say, the jacket temperature, the secondary variable, should be related proportionally to a change in the reactor temperature, the primary variable; and finally,
6. the variables should be linearized where possible by the use of appropriate variable transformations.

Advantages of cascade temperature control include these:

1. It reduces deviations around set point (i.e., drifting).
2. The system reacts faster.
3. The primary reaches a nearly constant value within a few time cycles.

Alarms can be made using the secondary loop, which allows for early action to be taken to prevent damages and disturbances from unnecessarily affecting the most important variable, the temperature of the large batch reactor, in the primary control loop.

While cascade control requires an analysis based on theoretical principles and statistical data, our second example illustrates a "knowledge-based" approach. Recently, commercial products have been introduced that achieve precision temperature and humidity control using fuzzy logic (Liebert Corporation, 1992; see also THEORY AND APPLICATIONS OF FUZZY LOGIC; Brubaken, 1992). This article does not endorse a particular product, but describing this particular product should help us understand some of the newer, knowledge-based approaches now being used in process-control applications.

Control of the environment for large computing systems is often a far greater challenge than for rooms inhabited by people. Not only do the systems themselves generate heat, but they are often specified by their manufacturers to be maintained in as little as a ±0.5 °C range. Humidity control is also a challenge, since corrosion rates and jamming of associated mechanical systems increases at high humidity levels, while static discharge is promoted at low levels. The humidity set point specified in this application is 50% relative humidity; but the maximum allowable change in relative humidity is only ±3% within 1 h.

In addition, the design of a precision environmental control system also faces nonlinearities, caused by such system behavior as air-flow delay and dead times, uneven airflow distribution patterns, and duct-work layouts. Uncertainties in system parameters are often present—for example, room size and shape, location of heat-producing equipment, thermal mass of equipment and walls, and amount and timing of external air introduction.

Recognizing these challenges, a control system was designed with the following objectives in mind:

1. precision temperature and humidity control;
2. minimization of cycling times (i.e., the opening and closing of the damper and turning on and off the compressor), thereby increasing reliability and component life, and also resulting in increased energy efficiency; and
3. straightforward and therefore inexpensive control electronics.

A fuzzy logic approach (Liebert Corporation, 1992; Brubaker, 1992) was investigated and ultimately implemented to achieve precise temperature control using relatively simple hardware. The actual controller consists of six fuzzy inputs, three fuzzy outputs, and 144 principles (rules). This particular control system runs on a small microprocessor; it is relatively easy to run, and new rules can be added.

Fuzzy input variables used in this system are the temperature relative to a set point, the rate of temperature change, the humidity relative to a set point, the rate of humidity change, and two proprietary variables associated with the action of the controllers.

Fuzzy outputs control the amount of cooling, the amount of dehumidification, and heat. Unlike conventional controllers, outputs can also be treated as fed-back input variables, while time delays are treated as fuzzy outputs as well. Each fuzzy variable is assigned seven membership functions as values—for example, Large Negative, Medium Negative, Small Negative, Near Zero, Small Positive, etc. Ranges for these membership functions are proprietary.

With this expert system consisting of 144 principles, more elaborate approaches to control were not necessary. Inputs are sampled, the principle base is queried, and outputs are updated once every second. The "long" intersample delay allows a simple eight-bit microprocessor to implement this rather large fuzzy system.

Relative to on/off control, fuzzy logic greatly reduces the number of system cycles, thereby significantly reducing wear on the damper. Moreover, with the rules set up the way they are, paradoxically, the outdoor ambient temperature actually assists (rather than interfering with) internal temperature control.

The control engineers wrote a PC-based simulator to test the fuzzy design under simulation prior to committing it to hardware. Their simulations and tests led them to the conclusion that fuzzy logic is able to control temperature and humidity precisely in rooms having uncertain and nonlinear characteristics. Moreover, installation includes no tuning procedure: The same set of principles evidently applies in many similar applications.

ACKNOWLEDGMENTS

The author would like to thank the Rohm and Haas Company for permission to publish this article. In addition, the author would like to thank Ed Sioma and Dick Wu for many helpful discussions on temperature control; and Greg Shinskey, Marcia Katz, Bruce Carocci, Steve Lorbert, and Jim Pike—those authors who have granted me permission to use their figures.

GLOSSARY

Accuracy: The closeness of agreement between an observed value and the true value.

Antireset Windup: Feature in a three-mode PID controller that prevents the integral circuit from functioning when the temperature is outside the proportional band.

Block Diagram: A diagram of a process-control system to show the relationships of all the loop elements.

Calibration: The procedure of comparing a particular instrument with either a primary standard or a secondary standard with a higher accuracy than the instrument to be calibrated.

Cascade Control: Control in which the output of one controller is the set point for another.

Cold-Junction Compensation: Use of a reference thermocouple, usually held at 0 °C (cold), to compare with to improve thermocouple accuracy.

Drift: An undesired change in the output–input relationship over time.

Emissivity: Ratio of energy emitted by an object to the energy emitted by a blackbody at the same temperature. Emissivity depends on the material and surface texture.

Feedback Control: Type of control in which a measured variable is compared to its desired value to produce a signal that reduces the magnitude of the error.

Feed-Forward Control: Type of control in which the control system anticipates the effect of measured disturbances and compensates for them.

Filled Thermal Systems: The method of measuring temperature that involves a filled system whose characteristics change with temperature.

Fuzzy Logic: Simple, intuitive concept invented by L. Zadeh in 1965 (George, 1996) that involves "membership" in a group as the basis for basic logic, action, and control. Membership in a group or several groups may be fractional or "fuzzy."

IR Sensor: A measurement system that examines the change in radiation power density to get a measure of the temperature.

PID Control: Feedback-control method that minimizes the error by making changes in a manipulated variable according to the current error (proportional term), the average error over a specified amount of time (integral term), and the rate of change of error in time (derivative term).

Radiation Thermometer: Method of measuring temperature that involves emission of electromagnetic radiation from a targeted source.

Repeatability: The closeness of agreement among a number of consecutive measurements of the output for the same value of the input under the same operating conditions.

Resistance Thermometer: Method of measuring temperature that uses the resistance developed across a metal. May be a thermistor or an RTD.

Signal Conditioning: Operations done on a measurement to improve the accuracy of the measurement either by filtering (taking several measurements over a short period of time), linearization, or calibrating against a reference measurement. Signal conditioning is usually done using a microprocessor or a personal computer.

Thermocouple: A device for measuring temperature, based on the junction of two dissimilar metals, which has a voltage output proportional to the difference in temperature between the hot junction and the lead wires (cold junction).

Thermowell: Closed-end tube designed to protect temperature sensors from harsh or corrosive environments, high pressures, and flows.

Transmitter: A device that responds to a measured variable, converting it to a standardized signal that is a function only of the measured variable.

Works Cited

Abdelrahman, M. (1990), "Fuzzy Sensors for Fuzzy Logic," *Control Eng.* **37,** 50–51.

Anon. (1989a), *Temperature Measurement Handbook,* Stamford, CT: Omega Engineering, Inc.

Anon. (1989b), "Understanding Controller Gain," *Control* **2,** 22–23.

Anon. (1992), "Plant Controls: Temperature Measurement," *Plant Eng.* **11,** 291–293.

Anon. (1995), "RdF: Specialists in Temperature Measurement," *Sensors* **6** (4), 39.

ASTM (1987), *T/C Standard: Temperature-Electromotive Force (EMF) Tables for Standardized T/Cs,* Standard E230-87, Gaithersburg, MD: National Bureau of Standards, pp. 101–103.

Basalik, T. (1993), "Calorimetry for Process Optimization," *PI Qual.* **7,** 22–23.

Beatty, B. D. (1988), "Techniques for Improving Control of Batch Reactors," *Control* **1** (10), 238–242.

Berra, J. (1996), "Fieldbus: Its True Potential," *Ind. Comput.* **8,** 20–22.

Box, G. E. P., Jenkins, G. M. (1976), *Time Series Analysis Forecasting and Control,* San Francisco, CA: Holden-Day, Chap. 11.

Brubaker, D. (1992), "The Huntington Brief," available on the Internet World Wide Web address http://www.austinlinks.com/Fuzzy/control.html, Menlo Park, CA: The Huntington Group.

Campbell, T. W. (1993), in: E. P. Dougherty (Ed.), *Temperature Control Principles for Process Engineers,* Munich: Hanser, Chap. 8.

Carocci, B. (1995), "Industrial IR Imaging Is Showing up Everywhere!" *Instrum. Control Syst.* **68** (6), 23–30.

Chia, T. L., Lefkowitz, I. (1991), "Model-Based Technique Helps Tune PID Loops," *Control* **4** (12), 36–40.

Choudhury, S., Utiger, L. (1992), "Heat Transport in Agitated Vessels: Scale-up Methods," Mettler Technical Application No. 3, Hightstown, NJ: Mettler Instrument Corp.

Clayton, W. (1993), "Thin Film Platinum RTDs Challenge Wire Wounds," *Instrum. Control Syst.* **66** (6), 33–36.

Craig, L. W. (1995), *Chem. Proc.* **58,** 32–34.

Dougherty, E. P. (Ed.) (1993), *Temperature Control Principles for Process Engineers,* Munich: Hanser.

Dougherty, E. P., Westkaemper, P. H., Wu, R. S. (1989), in: T. Provder (Ed.), *Computer Applications in Applied Polymer Science II,* ACS Symposium Series No. 404, Washington, DC: American Chemical Society, Chap. 37, pp. 478–489.

Fasano, J. B., Brodkey, R. S., Haam, S. J. (1992), "Local Wall Heat Transfer Coefficients Using Surface Calorimeters," in: R. King (Ed.) *Proceedings of the 7th European Conference on Mixing, Brugge, Belgium, September 18–20, 1991,* Dordrecht: Kluwer Academic, p. 253.

Finney, P. (1995), "Understanding the New IR Imaging Systems," *Plant Eng.* **49** (10), 108–110.

George, M. (1996), "DSPs Make Fuzzy Logic Practical for Industrial Control," *Instrum. Control Syst.* **69** (2), 49–57.

Gerry, J. (1991), "Tune Loops for Load Upsets, Not Set Point Changes," *Control* **4** (9), 46–52.

Hall, J. (1990), "Handy Temperature Tables," *Instrum. Control Syst.* **63** (6), 51–56.

Hashemian, H. M., Peterson, K. M. (1993), "In-Situ Tests Gauge Thermocouple Performance," *Intech* **20,** 31–33.

Horton, R. M. (1991), "Tuning a PID Control Loop Is Still Painful, but Special Software Removes the Sting," *Pers. Eng. Instrum. News* **4,** 67–70.

Huber, J. C. (1993), "Understanding Fiber Optics: The Basics," *Plant Eng.* **47** (11), 68–70.

Kane, T. (1990), "Thermocouples vs RTDs: The Battle Continues," *Intrum. Control Syst.* **63** (6), 43–45.

Katz, M. (1993), in: E. P. Dougherty (Ed.), *Temperature Control Principles for Process Engineers,* Munich: Hanser, Chap. 5.

Kumpinsky, E. (1992), "Experimental Determination of Overall Heat Transfer Coefficient in Jacketed Vessels," *Chem. Eng. Commun.* **115,** 13–17.

Labs, W. (1989), "Temperature Monitoring and Control: A Technology Update," *Instrum. Control Syst.* **62** (6), 25–31.

Land Infrared Company (1995), *A Basic Guide to Temperature Measurement Using Radiation Thermometers,* unpublished technical note.

Larson, K. (1995), "What to Look for in a Process Controller," *Control* **8** (11), 22–28.

Lee, P. L., Sullivan, G. R. (1988), *Comput. Chem. Eng.* **12,** 573–598.

Lewis, C. W. (1995), "Four-Wire Techniques Minimize Temperature Measurement Errors," *Intech* **22** (5), 40–43.

Liebert Corporation (1992), "LogiCool," Columbus, OH: Liebert Corporation.

Liptak, B. (Ed.) (1993), *Temperature Measurement,* Radnor, PA: Chilton.

Lorbert, S. J., Pike, J. P. (1993), in: E. P. Dougherty (Ed.), *Temperature Control Principles for Process Engineers,* Munich: Hanser, Chap. 2.

Love, J. (1993), "IR Temperature Sensors: Single-Point vs Multipoint," *Instrum. Control Syst.* **66** (6), 29–32.

MacGregor, J. F. (1972), "Topics in the Control of Linear Processes with Stochastic Disturbances," Ph.D. Thesis, University of Wisconsin, Madison.

Machado, R. (1994), "Application of Heat Flow Calorimetry to the Scale-Up of Polyurethane Prepolymer Processes," in: *Proceedings of the Engineering Foundation Conference on Polymer Reaction Engineering IV, St. Augustine, Florida, February 13–18, 1994,* (unpublished).

Maczka, W. J. (1988), "Infrared Use Heats up in Temperature Control," *Intech* **15** (8), 44–46.

Mettler Analytical Laboratory (1991), *The Calorimetric Principle of the Mettler RC,* Hightstown, NJ: Mettler Instrument Corp.

Moritz, H.-U. (1988), "Rechnergestutzter Laborreaktor," *Chem. Ind.* **5,** 100–108.

Polonyi, M. J. G. (1989), "PID Controller Tuning Using Standard Form Optimization," *Control Eng.* **36** (3), 102–106.

Pompei, F. (1995), *Instrum. Control Syst.* **68** (6), 35–46.

Reotemp Instrument Corporation (1987), *Reotemp Stainless Steel Bimetal Thermometers,* Technical Catalogs No. A-870 and No. LS-46SUR, San Diego, CA: Reotemp Instrument Corporation.

Rhinehart, R. R., Riggs, J. B. (1990), "Process Control through Nonlinear Modeling," *Control* **3** (7), 86–90.

Sack, M. E., Lee, S.-I., Biesenberger, J. A. (1973), "Effect of Temperature Variations on Molecular Weight Distributions: Batch, Chain Addition Polymerizations," *Chem. Eng. Sci.* **28,** 241–257.

St. Clair, D. W. (1991), *Intech* **18** (10), 35–37.

Samuels, H. (1995), "Intelligent Signal Conditioning," *Sensors* **6,** 28–32.

Schmidt, C.-U., Reichert, K.-H. (1988), "Reaction Calorimeter: A Contribution to Safe Operation of Exothermic Polymerizations," *Chem. Eng. Sci.* **42,** 2133–2139.

Schraff, F. (1995), "Put Temperature Measurement and Control Where It's Needed," *Instrum. Control Syst.* **68** (6), 47–51.

Sheng, R. S., Nodrop, R. W. (1990), "Smart Transmitters Make Their Move," *Instrum. Control Syst.* **63** (6), 47–51.

Shinskey, F. G. (1993), in: E. P. Dougherty (Ed.), *Temperature Control Principles for Process Engineers,* Munich: Hanser, Chap. 6.

Smith, J. (1995), "Matching Temperature Sensors with Process Tasks," *Process Control Course, Instrumentation and Control Systems,* Radnor, PA: Chilton, Chap. 4.

Snell, J. R., Jr. (1995), "The Power of Infrared Thermography," *Plant Serv.* **16,** 79–82.

Sperber, R. (1995), "Young Batch Package Matures with SP88 and Windows NT," *Control* **8** (11), 57.

Stockdale, R. B. (1992), "New Standards Squeeze Temperature Measurement and Control Specifications," *Control Eng.* **39,** 89–92.

Studebaker, S. P. (1995), "Advanced Control in the New World," *Control* **8** (11), 75–88.

Tenney, A. S. (1988), "Applying Radiation Thermometers to Process and Lab Measurements," *Intech* **15** (10), 49–55.

Trietley, H. L. (1989), "Applying Thermistors to Process Measurements and Control," *Control Eng.* **36,** 41–47.

Verhaegen, S. C. (1991), "When to Use Cascade Control," *Intech* **18** (10), 38–40.

Wahl Instruments (1995), *Wahl Temperature Instrument Catalog,* Culver City, CA: Wahl Instruments.

Walcutt, J. (1994), "Thermal Analysis for Today's Complex Design," The Internet World Wide Web address http://www.teleport.com:80/~compix/articles.npconint.html, Tualatin, OR: Compix, Inc.

Wang, T. P. (1991), "Thermocouple Selection and Maintenance—Part 1," *Intech* **18** (4), 44–46.

Williams, D. H. (1995), "Feedback and Temperature Controllers," The Internet World Wide Web address http://newton.ex.ac.uk:80/teaching/CDHW/Feedback, Exeter, U.K.: Exeter Academy.

Williamson Corporation (1995), "Everything for Noncontact Temperature Measurement" (Advertisement), *Instrum. Control Syst.* **68** (6), 60.

Wu, R. S., Dougherty, E. P. (1993), in: E. P. Dougherty (Ed.), *Temperature Control Principles for Process Engineers,* Munich: Hanser, Chap. 8.

Ziegler, J. G., Nichols, N. B. (1942), "Optimum Settings for Automatic Control," *ASME Trans.* **64,** 759–768.

Further Reading

ASTM (1987), *Standard Specification for Industrial Platinum Resistance Thermometers,* E 1137-87, Gaithersburg, MD: National Bureau of Standards.

ASTM (1993), *ASTM Manual on the Use of Thermocouples in Temperature Measurement MNL-12,* 4th ed., Philadelphia: American Society for Testing and Materials.

Baier, S. (1995), "Precise Temperature Measurements in a Hostile Environment," *Sensors* **6** (2), 13–14, 52.

Baker, H. D., Ryder, E. A., Baker, N. H. (1989), *Temperature Measurement in Engineering,* 2 vols., Stamford, CT: Omega Press.

Chafin, E. (1991), "Temperature Controllers are Getting Smaller and Better," *Instrum. Control Syst.* **64,** 27–30.

Curtis, D. (1990), "Fabric Producer Solves Cure-Oven Based Quality Control Problem with IR," *Instrum. Control. Syst.* **63,** 35–42.

Hewlett-Packard Co. (1990), *Practical Temperature Measurements,* Application Note No. 290, Palo Alto, CA: Hewlett-Packard Co.

Kerlin, T. W., Shepherd, R. L. (1989), *Industrial Temperature Measurement,* IRP Student Text, Stamford, CT: Omega Press.

Kleppe, J. A. (1991), "Acoustic Pyrometry Ups Boiler Efficiencies," *Instrum. Control Syst.* **64,** 35–39.

Kourtides, D. A. (1996), "Measuring High Temperatures in Ceramic-Fiber Blankets," *NASA Tech. Briefs* **15,** 46.

Riggs, J. B., Watts, J., Beauford, M. (1991), "Model-Based Control Streamlines Process," *Control* **4,** 53–57.

Ruppel, F. R., Carnal, C. L., Lind, R. F. (1990), "Tuning Control Loops: An Empirical Approach," *Control* **3,** 40–45.

Shinskey, F. G. (1979), *Process-Control Systems,* New York: McGraw-Hill.

Singer, P. (1996), "Thermal Processing: The RTP Race Heats Up," *Semicond. Int.* **11,** 64–70.

TEST EQUIPMENT, ELECTRICAL

Jim A. McEvers, *Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A.*

Introduction 564
1. **Fundamental Circuit** 565
2. **Component Measurements** 565
2.1 Passive Components 565
2.1.1 Resistance, Capacitance, Inductance, and Impedance Measurement 565
2.1.2 Multimeters 566
2.1.3 *LCR* Meters 566
2.1.4 Impedance Analyzers 566
2.2 Active Components 566
2.2.1 Semiconductor Measurement Systems 566
2.2.2 Integrated-Circuit Testers 566
3. **Power Sources** 567
3.1 Voltage Sources 567
3.1.1 AC Voltage Sources 567
3.1.1.1 AC Power Supplies 567
3.1.2 DC Voltage Sources 567
3.1.2.1 Constant-Voltage Supplies 567
3.1.2.2 Constant-Current Supplies 568
4. **Power-Supply Measurements** 568
4.1 Multimeters 569
4.1.1 Analog 569
4.1.2 Digital 569
4.2 Power-Line Monitors 569
4.3 Power Measurements 569
4.3.1 Electronic Watt Meters 570
4.4 Clamp-On ac Ammeters 570
4.5 Oscilloscopes 570
4.6 Spectrum Analyzers 570
5. **Signal Sources** 570
5.1 Sweep and Function Generators 570
5.2 Arbitrary-Function Generators 570
6. **Signal Measurement Equipment** 571
6.1 Oscilloscopes 571
6.1.1 Types of Oscilloscopes 571
6.1.1.1 Analog Oscilloscopes 572
6.1.1.2 Digital Oscilloscopes 572
6.1.1.3 Ranges 572
6.1.1.4 Triggering 572
6.2 Spectrum Analyzers 573
6.3 Network Analyzers 573
6.3.1 Multifunction Meters 574
6.3.1.1 Digital Multi-meter/ Oscilloscope 574
6.3.2 Frequency Measurement/ Counters 574
7. **Signal/Communications Link Testers** 574
8. **Probes** 575
8.1 Voltage Probes 575
8.1.1 Passive Voltage Probes 575
8.1.1.1 Low Impedance 575
8.1.1.2 High-Impedance Passive Probes 576
8.1.1.3 High-Voltage Probes 576
8.1.1.4 Differential Passive Voltage Probes 576
8.1.1.5 Active Voltage Probes 577
8.1.1.6 Differential Active Voltage Probes 577
8.2 Current Probes 578
8.2.1 Passive Current Probes 578
8.2.2 Active dc Current Probes 578
8.2.3 Active ac Current Probes 578
8.2.3.1 Clamp-On 578
8.2.3.2 Special-Purpose 579
8.2.3.3 Calibration/Compensation 579
9. **Factors Affecting Electrical Measurements** 579
9.1 Noise 579
9.1.1 Externally Generated Noise 580
9.1.2 Internally Generated Noise 580
9.1.3 White and Colored Noise 580
9.1.4 Signal-to-Noise Ratio 580
9.2 Grounding 581
9.3 Loading 581
9.4 Aliasing 582
10. **Electrical Testing Equipment Physical Configurations** 582
10.1 Handheld 582
10.2 Portable 583
10.3 Rack Mounted 583
10.4 Instrument-on-a-Card 583
11. **Display Configurations** 583
11.1 Analog 584

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

11.2 Digital 584
11.3 Graphic 584
12. Virtual Instrumentation 584
13. Interfacing 585
13.1 General-Purpose Instrument Bus (GPIB) 585
13.2 Dedicated Instrument Bus 585
13.3 Serial Instrument Communication 585
13.3.1 RS-232 585
13.3.2 RS-422 586
14. Conclusion 586
Glossary 586
Further Reading 587

INTRODUCTION

The application of electrical test equipment encompasses several areas. These include the measurement of properties and values of the components that make up the circuit, levels and characteristics of the potential and current sources used to power the circuit, the level and characteristics of the input and output signals associated with the circuit, the levels and characteristics of the noise present in the system, and the overall system characteristics. Table 1 provides an overview of the major areas of electrical testing, specific areas of interest by category, types of measurements required, the applicable types of electrical test equipment, and auxiliary equipment that may be required.

Electrical test equipment has undergone revolutionary changes during the past several years. Advances in microcircuitry and embedded-microprocessor technology have had a very significant impact on both the industry and those who must make electrical test measurements. Such advances have brought "field-grade" equipment almost to the level previously considered as belonging to laboratory-grade calibration equipment.

Most electrical test equipment utilizes the increased analysis and diagnostic capability provided by embedded computer technology. This provides the ability to perform detailed analysis as measurements are being made. The disadvantage is that the user must sometimes spend considerable time in understanding the capabilities that are available and how best to utilize them. This increased sophistication and capabilities of modern electrical test equipment, while providing

Table 1. Categories of Electrical Measurements and Corresponding Electrical Test Equipment

	Circuits/ components	Power sources	Power conversion/ conditioning	Signals	Nodes
Electrical parameter of interest	Passive elements Resistance Capacitance Inductance Impedance Active elements Diodes Transistors Integrated circuits	Voltage Current Power	Voltage Current Power	Voltage Current Power Frequency	Voltage Current Power Frequency
Specific areas of interest	Component values Component characteristics Circuit effects Circuit performance Circuit characteristics	dc ac Peak value Average value rms value Noise	dc ac Noise Ripple	dc ac Low freq. rf Instantaneous value Peak value rms value	Transients Pulses ac Low freq. & rf
Applicable equipment	Multimeter LCR meter Active component analyzers Network analyzer	Voltmeter Ammeter Multimeter Power meter Power-line monitor	dc supply ac supply Function generator Pulse generator Arbitrary function Generator	Arbitrary function Generator Pulse generator Oscilloscope Spectrum analyzer Time domain Reflectometer	Oscilloscope Spectrum analyzer
Auxiliary equipment		Probes	Probes	Probes	Probes

greatly increased functionality, does not relieve the user from having a basic understanding of electrical measurements, proper choice of equipment, proper interpretation of the results, and proper safety precautions to be followed while making such measurements. It is time well spent and results in the optimal use of the instrument and ultimately in time savings and more useful measurements.

Many of the manufacturers of electrical test equipment such as Tektronix, Hewlett Packard, and Fluke, to name just a few, provide excellent literature describing not only the latest in electrical test equipment but also literature and application notes that describe the latest measurement techniques and methods for performing many types of electrical measurements. Much of this literature is provided free of charge or at nominal cost.

1. FUNDAMENTAL CIRCUIT

All electric circuits possess common characteristics. The circuit is composed of some number of components, it must be powered from some source, the power generally must be converted and conditioned, it operates on and or produces some form of signal, it is subject to and produces electrical noise, and these circuits are combined to form complex systems. Electrical test equipment is available and used to make measurements in each of these areas. For reference, each area is illustrated in Fig. 1.

The *electrical circuit* is powered by the *power source* through the *power supply.* The electrical circuit may be a complex electronic circuit involving semiconductors, microwave devices, logic elements, electric heating equipment, or electric motors or other electromechanical devices. The input signal may be a complex communications signal, a signal from a process sensor, logic input from a preceding stage of a computational system, the set point for a process control function, or a simple on-off switch. The output signal may be a microwave signal, a current to a heating unit, a command to a process actuator such as a solenoid valve or motor, or a low-level logic pulse. Each of these elements has need to be accurately measured to ensure that the circuit is performing as designed and required. That is the function of electrical test equipment.

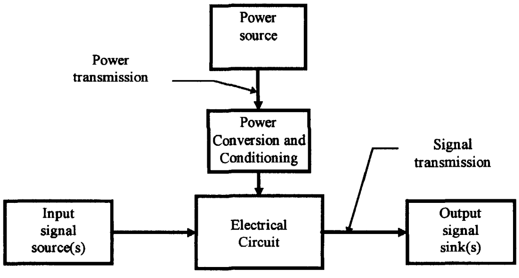

FIG. 1. Basic electric circuit.

The power source may provide direct current or alternating current, and virtually any voltage level and polarity as a function of time. The power supply may simply filter the power, or it may transform, rectify, filter, and/or regulate the power. The electrical circuit may perform any number of functions from simple to very complex. The signal source and sink may also be very simple or complex. All of these elements have common attributes: they may each require measurement of voltage, current, power, noise, frequency, ac and dc component content, phase, resistance, and impedance. The current levels may vary from a few milliamperes to thousands of amperes; the voltage levels may likewise extend over a broad range.

2. COMPONENT MEASUREMENTS

Accurate measurement and characterization of the components that make up a circuit is sometimes required to understand the performance of the circuit. Electrical test equipment has been designed for application to each specific category of component.

2.1 Passive Components

2.1.1 Resistance, Capacitance, Inductance, and Impedance Measurement Resistance, capacitance, and inductance are properties associated with all circuit elements to varying degrees. Impedance is the term applied to the combined effects of resistance, capacitance, and inductance and is frequency dependent. It can be the result of combined effects of discrete components such as resistors, capacitors, and inductors in a circuit or distributed effects such as in cables, wiring, and circuit boards.

2.1.2 Multimeters Digital multimeters (DMMs) offer a low-cost, generally rugged, method of making voltage, current, and resistance measurements over wide ranges and test conditions. Such instruments are well suited to everyday laboratory and field measurements and have generally replaced bridges for making even the most precise electrical measurements in the laboratory.

2.1.3 *LCR* Meters The *LCR* (inductance, resistance, and capacitance) meter is used for high-accuracy measurement of passive components. In addition to making simple resistance, capacitance, and inductance measurements, these meters may also be used to make measurements of other parameters such as admittance, phase angle, reactance, conductance, Q for inductors, dissipation factor, and Equivalent Series Resistance (ESR) for capacitors. These meters are typically capable of measuring over the range of 0.01 mΩ to 10^{16} Ω for resistance, 0.001 pF to several hundred microfarad for capacitance, and 0.01 nH to 100 kH for inductance (depending on the particular instrument). Specialty instruments may also be available for measurement ranges that exceed those cited. Many of these instruments also provide the ability to apply bias currents to inductors, bias voltages to capacitors, and rf to components as they are being tested.

Special adapters are also available that allow the measurement of various-component physical configurations such as surface-mount devices, the application of high bias voltages, high bias currents, and other special-purpose applications. In addition to the test and measurement of individual, discrete components, special configurations of these meters may be used for materials testing (e.g., circuit boards) and dielectric testing.

2.1.4 Impedance Analyzers Impedance analyzers, in addition to providing the measurement capabilities of *LCR* meters, also provide the ability to measure impedance, phase, and transmission parameters. These instruments generally cover a wider frequency range, and they have the ability to sweep over a selected frequency range.

2.2 Active Components

The measurement of active components such as transistors, diodes, and related components involves "driving" the component with a prescribed test condition and measuring the response of the device. For example, in the case of a diode, parameters of importance include forward voltage drop at a specific current or over a range of currents, reverse breakdown voltage, reverse leakage current at a specified voltage, reverse recovery time, and other properties based on the application. In testing forward voltage drop, the device is subjected to a specific level of current in the forward or conduction direction, and the voltage drop across the diode is measured. In some cases, the forward current may be of such magnitude that the diode would not survive if the current was applied continuously. The current must therefore be pulsed through the diode and a measurement taken within a relatively short time. Such measurements require generally more complex test equipment than that required for passive-component measurement.

2.2.1 Semiconductor Measurement Systems Semiconductor measurement systems provide the ability to expose discrete semiconductor devices to desired test conditions and measure the response of the device. These instruments may be used to measure dc gain, leakage current, and small-signal parameters in both pulsed and steady-state conditions. An example of such a test on a bipolar transistor would be the application of a specific base current level, swept over a predefined range, and the measurement of the corresponding collector current. In the case of a field-effect transistor, the gate voltage would be swept over a predefined range and the source current measured and plotted as a function of the gate voltage. The advantage of such an instrument lies in the ability to perform routine semiconductor measurements, under steady-state and pulsed conditions, without the need to fabricate and test custom test fixtures.

2.2.2 Integrated-Circuit Testers These systems are considerably more complex than the instruments used to measure individual components but provide the ability to measure and assess the performance of entire analog and digital integrated circuits. These test systems are programmable and can measure, store, display, and analyze entire circuits. They provide for the application of a complex set of excitation signals and conditions and the ability to record and analyze the output signals.

3. POWER SOURCES

All circuits require some form of power source. The source may be solar cells, batteries, line voltage, or any number of other such sources. Irrespective of the source, it must be capable of providing the required voltage and current level required by the circuit. It must produce very little electrical noise and must possess the required degree of regulation.

3.1 Voltage Sources

While some circuits may require a constant-current source, the most commonly used source provides a constant voltage to power the circuit. These are generally regulated to some degree but may also be completely unregulated. In the case of unregulated voltage sources, the output voltage level of the source will vary as a function of the load and the line supply voltage. In the case of regulated sources, the output voltage is held constant through the use of a sensing and feedback circuit. This feedback circuit utilizes active circuit components, e.g., regulators and error amplifiers, to maintain the desired output voltage at the desired level relatively independent of load and line voltage variations.

3.1.1 ac Voltage Sources The most common source of ac is that supplied through the 60- and 50-Hz commercial power grids. The power-line voltage is generally not used directly except for lighting and powering motors. In most electronic applications the power is transformed to a lower voltage, rectified, filtered, and regulated before being used to power a circuit. However, some applications may require an ac source that provides voltage at a specific frequency.

3.1.1.1 ac Power Supplies. High-power, single-phase and multi-phase, fixed- and variable-frequency ac voltage sources are available. These supplies typically operate from line power and rectify the line ac voltage to provide a source of dc. The dc power is then used to power oscillator and power-amplifier circuits, which in turn provide the ac power at the desired frequency and voltage level.

3.1.2 dc Voltage Sources

3.1.2.1 Constant-Voltage Supplies. Constant-voltage supplies, just as their name indicates, provide a source of constant voltage independent of the level of current required (within prescribed limits). Circuitry within the supply converts the ac line voltage into dc voltage through the use of voltage transformation and rectification. The resulting dc voltage is further filtered to reduce ripple effects that remain following the rectification process. The filtered dc voltage is then applied to voltage-regulating circuitry. The purpose of the regulating circuitry is to maintain the output voltage of the supply as close to the preset voltage as is reasonably possible, with the cost of the supply generally proportional to the degree of filtering and regulation required. The degree of regulation is specified for each particular type of supply and is a function of both the load current and the line supply voltage. The two primary types of regulation used are series/parallel or *linear* regulation and switching regulation.

In series regulation, a series pass element is placed between the filter stage of the supply and the output terminals. The arrangement of a typical circuit is shown in Fig. 2. A feedback circuit provides the series pass element with an appropriate regulating signal that is inversely proportional to the output voltage and is compared with the preset or desired reference voltage. The primary disadvantage of the linear regulating type power supply is its inefficiency. Since the supply operates from a fixed "up-stream" voltage source, i.e., the filter circuit, the only way in which it can maintain the desired output voltage is to drop voltage across the regulating element. This corresponds to power dissipation within the supply and becomes worse as more current is required by

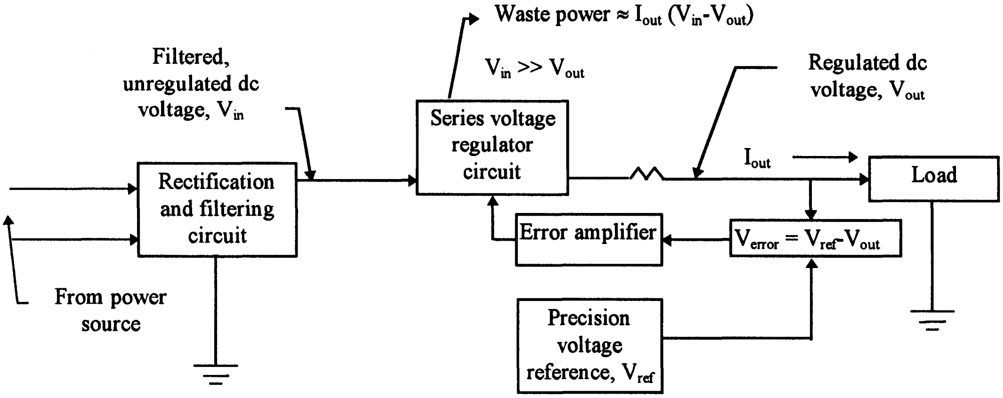
Waste power ≈ I_{out} (V_{in}-V_{out})
$V_{in} >> V_{out}$
Filtered, unregulated dc voltage, V_{in}
Regulated dc voltage, V_{out}
Series voltage regulator circuit
Rectification and filtering circuit
I_{out}
Load
Error amplifier
$V_{error} = V_{ref}-V_{out}$
From power source
Precision voltage reference, V_{ref}

FIG. 2. Typical series voltage-regulator circuit.

the load. The relative advantage of the linear supply is its simplicity and lower cost.

The more efficient method of regulating the output of a constant voltage supply is the *switching* regulator. Using the relationship $Q \cong CV$, or, rearranging, $V \cong Q/C$, one can see that it is possible to control the voltage across a capacitor by varying the quantity of charge present on the capacitor. This is the fundamental operating principle of the switching regulator. A high-frequency pulse stream is supplied to a capacitor through a *charge-pump* circuit. The amount of charge deposited on the capacitor is a function of the desired output voltage level. As the output voltage drops, the frequency of the charging circuit increases to replenish the charge on the capacitor, thereby increasing the voltage across the capacitor. The major advantage of this type of supply is its higher efficiency and lighter weight. A key concern in the use of this type supply, particularly with digital electronic circuits, is the level of ripple on the output voltage and the level of output regulation. This ripple is an artifact of the charging circuit and can have adverse effects on digital circuit applications, which may be particularly sensitive to switching transients. Significant effort is made to filter out such noise.

Almost all laboratory or bench-top power supplies incorporate some form of output-current limit protection including short-circuit protection. Some supplies are equipped with a variable current-limit capability that allows the user to preset the maximum allowable current available. This can prevent damage to test or experimental circuits during development. Other types of supplies are equipped with a simpler crow-bar circuit, which operates by "crow-barring" the input to the supply to zero when the output current exceeds the maximum limit. Both of these methods are aimed at preventing too much current from flowing through the load, thereby minimizing damage to the load and also to the supply. The variable current-limit supply is by far the more popular and functional since it allows the user to set the level at the most desirable level.

3.1.2.2 Constant-Current Supplies. In some applications it is necessary to maintain the current at a constant value and allow the voltage to adjust as necessary to maintain the desired current. Such supplies are referred to as *constant current* supplies. The current is maintained at a constant level by allowing the applied voltage to vary in proportion to the load. The voltage is referred to as the compliance voltage, and the preset current level may be maintained as long as the compliance voltage range is not exceeded. This is analogous to the maximum allowable current output from a constant voltage supply.

4. POWER-SUPPLY MEASUREMENTS

Most commercial power supplies are equipped with some form of meter or display to allow monitoring of output conditions. If such features are not provided, the output may obviously be measured or monitored using any of the equipment referenced in the signal-source measurement section. In addition to voltage and current measurements, it may be desirable to measure the output ripple, output power, or output noise or to observe the output wave form of a variable-frequency ac supply.

In assessing the quality of the power used for an application, it may be necessary to utilize any or all the instruments listed in the following sections.

4.1 Multimeters

Aside from dedicated equipment, the analog or digital multimeter is very useful for measuring the output voltage and current levels of power supplies. They are relatively low in cost and designed for portability.

The multipurpose meter or multimeter has probably been the most used, general-purpose piece of electrical test equipment on account of its handheld portability, simplicity, and application to the most common types of field measurements. It provides the ability to measure ac and dc voltages over the ranges from a few volts to 600 V, dc currents from a few milliamperes to typically 10 A, and dc resistance from milliohms (in some cases) to several megohms. The use of the field-effect transistor input stage provides even the analog multimeter with a relatively high input impedance.

4.1.1 Analog In its simplest form, the analog volt-ohmmeter (VOM) consists of an analog current-meter movement with appropriate current-limiting resistor combinations to provide several voltage ranges, several shunt-resistor configurations to provide various current ranges, and a battery circuit and resistors to provide a resistance measurement capability.

4.1.2 Digital The digital implementation of the traditional multimeter contains an analog-to-digital sampling circuit and digital display in place of the analog-meter portion of the conventional multimeter. The display is usually a liquid-crystal display unit to conserve power. A drawback of the digital multimeter is the inconsistency of readings due to the sampling of the analog-to-digital circuit, particularly on a noisy signal. While the analog meter mechanically "integrates" signal variations, the digital meter is very capable of sampling and displaying instantaneous signal values. In a noisy environment, this may prove to be very confusing as the display "hops" from one value to another. To help overcome this problem, some digital meters are equipped with an internal integrating circuit that filters out such instantaneous variations. A key advantage of digital instrumentation is its ability to autorange in response to the voltage being measured. In the simple analog meter-based instrument there is no provision to detect and adjust the meter's range, whereas in the digital version, it is possible for the more sophisticated input circuitry to detect the magnitude (and polarity in the case of dc voltages and currents) of the input voltage and adjust the display to the appropriate range.

4.2 Power-Line Monitors

The quality of the commercial power is generally assumed to be adequate for most applications. It is generally assumed that the line voltage is 120 VAC, 60 (or 50) Hz, that the power is clean (i.e., no noise, voltage "spikes," transients, etc.). This is generally not the case. Power-line monitors provide a means of measuring and recording fluctuations and transients on power lines. The units also detect and record sags and surges in line voltage, variations in frequency, and short-duration events such as spikes that occur on the line. They are usually equipped with an integral strip-chart printer and can be programmed to print results on demand or to print findings automatically at preset intervals. Transient events are generally reported as they occur, and variations in line frequency and voltage are generally reported at preset intervals. Multiphase units are also available, which record phase-to-phase information such as voltage differences. These instruments can be extremely useful when trying to determine the cause of intermittent problems in equipment when all other avenues have been explored.

4.3 Power Measurements

Power-measurement equipment is required when it is necessary to monitor ac input power to power-conversion equipment as well as to measure the output power of such equipment. Depending upon the application or requirement, a conventional watt meter may be sufficient for most ac power-source measurements; or, for more general applications, e.g., high-voltage, high-current, or floating measurements, a combination of current probes, voltage probes, and interface equipment to the oscilloscope can provide a very versatile power-measurement system.

4.3.1 Electronic Watt Meters Electronic watt meters are used to measure power consumption in many power-source applications. Conventional watt meters were based upon the interaction of the voltage and current within a mechanical meter arrangement. Electronic watt meters utilize electronic circuitry to determine instantaneous power, peak power, and average power, and to record power consumption as a function of time (kW-h).

4.4 Clamp-On ac Ammeters

These devices provide the ability to measure ac currents quickly, primarily in power sources. The meter is equipped with a pair of "jaws" that are opened, placed around the conductor in which the current of interest is flowing, then released. The jaws of the instrument couple to the magnetic flux encircling the conductor and transform this into a measure of the current flowing in the conductor. The meters are also equipped with a voltmeter capability, which allows the voltage to be measured with the same instrument (although not simultaneously). This provides a convenient means of determining not only the current but also the voltage and therefore the power in a power-source circuit. Since the instrument measures rms voltage and current, and is incapable of measuring the phase angle between the voltage and current, it is only useful in the determination of rms power and not reactive power.

4.5 Oscilloscopes

Although used primarily for observing signals into and out of circuits, oscilloscopes can be very useful for observing output levels, noise, and ripple in power-supply applications. When used to observe power-line input noise or voltage levels, care should be taken to avoid shorting the line to ground through the oscilloscope ground.

4.6 Spectrum Analyzers

Spectrum analyzers, like oscilloscopes, are primarily intended for observing signal information. However, spectrum analyzers can prove extremely useful in assessing the quality of input and output power from power supplies and sources.

5. SIGNAL SOURCES

In addition to the power supplies to power the circuit, some form of input signal is required for the circuit during normal operation or for testing. Several types of signal sources are available based on the type, frequency, magnitude, accuracy, and stability of the required signal.

5.1 Sweep and Function Generators

Sweep and function generators provide the ability to vary the output frequency or voltage function in a specified manner. In the case of the sweep generator, the user may select a start and stop frequency and initiate a sweep over the specified range. These instruments are generally used in signal analysis in radio-frequency (rf) and microwave applications.

Function generators produce a varied range of voltage values as a function of time. These are generally fixed wave forms and include sine, triangular, sawtooth, and square-wave voltage signals. Lower-priced versions of these instruments are designed for general use, while more expensive and therefore more precise instruments are available (see also Sec 5.2).

5.2 Arbitrary-Function Generators

The more common function generators are limited to generating single-pulse functions or relatively simple repetitive functions (triangle, sine, and sawtooth). In some applications it is necessary to produce more complex repetitive or nonrepetitive voltages as a function of time. Signals such as those obtained from mechanically vibrating systems, radar echoes, "random" nuclear decay events, and similar naturally occurring sources are not usually composed of simple functions but rather consist of complex combinations of many functions or randomly distributed pulses.

Where these signals must be reproduced or simulated, an arbitrary-function generator (AFG) can prove to be indispensable. An arbitrary-function generator is capable of generating virtually any wave form that can be programmed into the system, the only limitation being the required resolution and the amount of memory available in the AFG.

The AFG basically consists of a memory unit, a clocking unit, a counter unit, and a digital-to-analog conversion unit.

The wave form to be generated is stored in the AFG memory in digital form (usually 8 bit). Each memory location is then "clocked out" to the digital-to-analog conversion (DAC) circuitry. The DAC operation converts the signal to an analog representation of the corresponding signal. A block diagram of an AFG is shown in Fig. 3.

The wave form is stored in the random access memory (RAM) through the use of programming software in the form of 8-bit bytes (some systems allow 16-bit words). Following the programming of the memory, the oscillator is allowed to increment a counting circuit. The output of the counting circuit is used as the address pointer to the RAM. The content of each memory location of the RAM is then transferred to the DAC where it is converted to analog form. Essentially any wave form that can be defined as a series of digital bytes can be reproduced by the AFG.

Wave forms are programmed into the unit through several mechanisms. They may be generated by a library of mathematical functions, interpolation, graphic or tabular entry of instantaneous level, or tabular entry of the desired wave form in digital format. Wave forms may also be combined to produce even more complex functions. The system may also be programmed to produce noise functions.

Software is provided to allow the user to interact more easily with the system especially in the definition and generation of highly complex wave forms.

Key considerations in selecting an AFG include the number of output channels required, memory depth, output resolution, the need for simultaneous analog and digital outputs, wave-form entry modes, and peak-to-peak output voltage range.

These generators are capable of producing signals that range in frequency from dc to several megahertz. They can also provide a ± dc offset. The peak-to-peak output voltage capability varies from model to model but 10 V is common, with some units providing as little as 2 V and as much as 20 V. If higher or lower amplitudes are required, appropriate amplifiers or attenuators can be used to adjust the level. Single- and dual-channel configurations are available with both analog and digital data being available simultaneously. Essentially all units can be remotely programmed using the General Purpose Instrument Bus (GPIB) protocol (see Sec. 13; and BUS SYSTEMS AND COMPUTER INTERFACING).

6. SIGNAL MEASUREMENT EQUIPMENT

6.1 Oscilloscopes

Oscilloscopes are used to observe voltage and current wave forms directly in the time domain. They offer the advantage of observation of high-frequency and complex signals.

6.1.1 Types of Oscilloscopes There are two basic classes of oscilloscopes: analog and digital. Each is designed for specific applications with requirements vs cost being a principal consideration. Most general-purpose oscilloscopes are based on analog designs and are used for measurement and monitoring of commonly encountered conditions such as general circuit design and

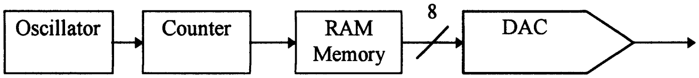

FIG. 3. Block diagram of arbitrary-function generator.

checkout; others are designed for observance and detection of specific high-speed phenomena or events. These include relatively simple analog scopes, analog storage scopes, digital scopes, digital storage scopes, and other more specialized scopes. See OSCILLOSCOPES ANALOG AND DIGITAL.

6.1.1.1 Analog Oscilloscopes. Analog oscilloscopes consist basically of input amplification circuitry, timing electronics, and support circuitry. The signal to be measured is scaled and displayed directly on the screen. The general-purpose analog scope is valued for its simplicity and low cost. It is used in circuit design activities, general troubleshooting, and other "everyday" activities. Analog storage scopes provide the ability to store the signal trace temporarily for longer viewing. It is useful for viewing single, high-speed events. The cathode-ray tube (CRT) uses a high-persistence phosphor that is capable of holding the image longer. This allows the image to be viewed longer and to be photographed easier.

6.1.1.2 Digital Oscilloscopes. The digital oscilloscope is replacing the analog scope in many areas because of its increased flexibility and decreasing cost. The digital oscilloscope utilizes a front end similar to that of an analog oscilloscope; there is an attenuator and a preamplifier section to condition the input signal as in the analog oscilloscope. However, that is where the similarity ends. In a digital oscilloscope, there is an analog-to-digital converter (ADC) immediately following the preamplifier. The ADC converts the incoming signal to its digital representation so that it may be processed, analyzed, and displayed. The high degree of accuracy and the relative complexity of the ADC generally results in a higher cost for that section. In lower-cost multichannel oscilloscopes there is a single ADC section that is shared between each of the input channels. Once the conversion has been accomplished, the digital data are stored and displayed. In the case of the digital storage scope the advantage is that the screen may be refreshed with the digital data by merely recalling it from memory.

Digital oscilloscopes also offer several types of mathematical functions. These include sum and difference of two channels, and the ability to compute and display histograms. A very powerful feature is the ability to transform the sampled wave form into the frequency domain and display the result. An example of this type of display is shown in Fig. 4.

There are two types of digital oscilloscopes: real-time (RT) and equivalent-time sampled (ETS). Those that use real-time, direct high-speed sampling circuits to capture the wave form provide a much more accurate representation of randomly occurring events such as glitches. ETS-based oscilloscopes acquire a number of samples of a wave form over several periods and "build up" the wave form that is then displayed. The ETS oscilloscope requires that the wave form be repetitive and is therefore not suited for finding extremely short-duration, nonrepetitive events (glitches).

6.1.1.3 Ranges. Oscilloscopes are equipped with several user-selectable time-base values and voltage ranges. The time bases range from several seconds down to a few nanoseconds. Most oscilloscopes are equipped with a 10× scale expansion of the time base, which is useful when trying to observe very fast events. The voltage ranges are front-panel selectable from 1 or 2 mV per division to 5 or 10 V per division. The voltage ranges may be extended upward through the use of a 10× or 100× attenuator probe.

6.1.1.4 Triggering. Triggering is the ability to synchronize the oscilloscope trace with some feature or characteristic of the signal to be observed. This allows the initiation of the trace to be "held off" until some

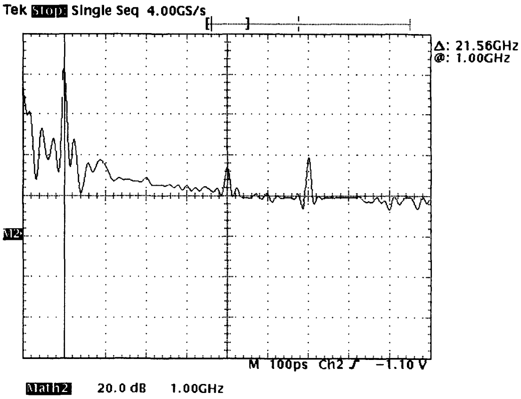

FIG. 4. Frequency-domain display of 1-GHz clock signal as measured on Tektronix 784 digital oscilloscope.

criterion is met. The criterion may be based on the magnitude of the voltage, the polarity of the voltage, the relative relationship between two voltages, or other events of interest. In the case of triggering the beam's movement on the basis of voltage level and polarity, the triggering control is set to either positive or negative depending upon the direction of interest, and the triggering level is adjusted accordingly. When the signal of interest moves in polarity in accordance with the criterion and reaches the preset trigger level, the beam moves across the screen and traces out the variations in the signal. In multichannel scopes, one input channel may be triggered on the basis of criteria established for one or more of the other channels. This triggering may also be delayed so that a trace does not begin until a specified time after the triggering criteria have been met.

In digital oscilloscopes, the triggering is performed quite differently than in analog oscilloscopes. In a digital oscilloscope the digitized signal is stored in a circular buffer. When the trigger condition is met, the stored samples are transferred to the display memory. The advantage of this method is that it is possible to look at the wave-form prior to the time the trigger occurred. This allows the analysis of a complex condition to include what has happened prior to the triggering condition being met.

6.2 Spectrum Analyzers

Spectrum analyzers are specifically designed to make possible the visualization and analysis of complex signals. This is accomplished by transforming the time-domain signal into the frequency domain. These instruments are able to accept a signal containing many frequency components and, through the application of Fourier analysis, to break the signal into its component frequencies. It displays the individual frequencies, the harmonics of those frequencies, and the relative magnitudes of each.

Such instruments are invaluable in the analysis of signals such as those obtained from mechanical equipment, biological and medical studies, radio astronomy, and any other area where it is necessary to identify the specific frequency components in a complex signal. Once a signal can be analyzed to determine its frequency content, appropriate filters and/or amplifiers can be acquired or designed to eliminate frequencies that lie outside of the area of interest or to amplify frequencies or bands of frequencies for further examination.

6.3 Network Analyzers

Network analyzers provide a convenient means for determining the transmission and reflection characteristics of electrical networks. These networks may consist of discrete components as in an electronic circuit or in distributed impedances such as cables. They may also be passive or active networks. A block diagram of a typical network analyzer arrangement is provided in Fig. 5.

Network analyzers operate by injecting a known signal into the network to be analyzed and comparing the network output signal with a known reference signal. The reference signal is obtained by passing the source signal through a power splitter, sending one half (by power level) to the network to be analyzed and the other half to the reference input of the analyzer. In this manner, the reference and test signals are known to be of the same frequency, phase, and magnitude. The signal that emerges from the network can then be compared to the reference signal to determine differences in magnitude and phase.

The source signal is swept through a range of frequencies in order to evaluate the network response or performance over a range of frequencies. The fact that the reference signal is derived from the same source minimizes the effect of any amplitude variations in the source. The network analyzer

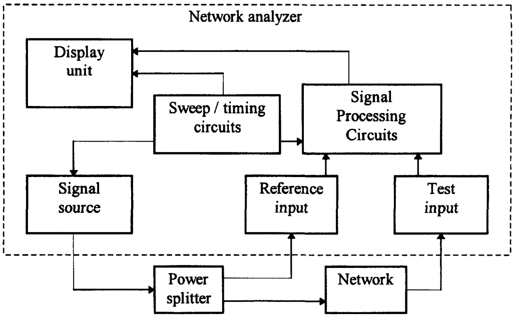

FIG. 5. Block diagram of network analyzer configuration.

may also be used to assess the degree of impedance matching in networks by measuring the reflectance.

On-screen markers are provided for ease of use and are controlled by a single front-panel knob, which allows the user to isolate points of interest for more detailed observation and analysis.

Network analyzers, like most of the current instruments, make extensive use of microprocessor and software technologies. Essentially all interaction with the system is through a CRT screen and software-programmed function keys. The extensive use of color CRT display technology allows for a better understanding and more rapid interpretation of the data being presented. The configuration settings may be saved on disk within the measurement system or on diskette as a record of the test setup; user programs and analysis procedures may be saved within the system as well as transferred to diskette. Extensive libraries of mathematical analysis routines are provided to aid in generation and analysis of test scenarios.

The sweep mode may be set for single-sweep mode, in which measurements are terminated following the initial sweep; continuous-sweep mode, in which case the sweep is automatically reinitiated following the last measurement in the current sweep; or a triggered-sweep mode, which requires a trigger condition to be met before the next sweep takes place.

6.3.1 Multifunction Meters The analog meter was the primary piece of easily transportable test equipment while equipment such as oscilloscopes, analyzers, and similar, more complicated instruments were generally considered to be laboratory-based equipment. Advances in microcircuit complexity and packaging have provided the ability to instill even more capability into the portable multimeter.

6.3.1.1 Digital Multimeter/Oscilloscope. An excellent example of increased functionality in portable equipment is the combined handheld multimeter, which incorporates the functionality of the multimeter (dc voltage, ac voltage, ac/dc current, etc.) and many of the capabilities of an oscilloscope. Such instruments are available from the Fluke Company, Tektronix, and other makers of test equipment. This equipment uses a relatively high-resolution liquid-crystal display (LCD) as the oscilloscope and conventional meter. The wave forms are "drawn" on the LCD display along with scale values and markers. A cursor may be moved along the wave form and the corresponding voltage value displayed on the screen in engineering units. Essentially all of the functions that are available on conventional CRT-based scopes and multimeters are being provided on handheld, digital-based configurations.

6.3.2 Frequency Measurement and Counters Frequency meters provide the ability to measure directly the frequency and period of an electrical signal. While the frequency of a signal may be determined by viewing on an oscilloscope and measuring the period of the wave form, it is generally not practical to do so in many applications. Since frequency measurement is based on the determination of the frequency with which a signal reaches a threshold level, similar circuits may also be used as counters. In the case of frequency measurement, the signal being measured must be periodic for the measurement to have any meaning. Frequency meters are available that can measure frequencies from 0.01 Hz to several gigahertz. Since counters only measure how many times a given signal meets a specified criterion (i.e., voltage level), the signal does not have to be periodic. It must, however, change sufficiently from the criterion to allow the circuit to reset and prepare for the next event. If the signal does not vary sufficiently from the triggering level, then, as far as the counter is concerned, there is no "new" event and therefore nothing to count. In both the frequency and counting modes, the detection threshold must be adjusted by the user threshold.

7. SIGNAL/COMMUNICATIONS LINK TESTERS

The measurement or verification of signal transmission parameters (refer to Fig. 1) may involve one or more types of measurement equipment. The signal being transmitted may be a simple dc voltage level (0–5 V), a current level (4–20 mA), an RF signal, a light signal (fiber optic), or a serial/parallel digital signal. The signal may be transmitted over a few feet or over several miles. The type of signal will determine the type of electrical test equipment used. In the case of a simple dc voltage level, a simple continuity test and voltage-level measurement using a portable multimeter will probably suffice; a current-level signal may be measured in a similar manner (it may be necessary to insert a small resistance in the circuit to convert the signal to a voltage since the series resistance of some multimeters may exceed the compliance voltage of the constant current signal source). The measurement of rf signals may require the use of a network analyzer or spectrum analyzer. Fiber-optic–based test equipment is available to support the troubleshooting and assessment of fiber-optic–based communication links. These include fiber-optic time-domain reflectometers to locate discontinuities in the transmission path. Digital communication "break out boxes" are available that allow a set of indicators and jumpers to be inserted in the communications path for observing transmission activity. Equipment is also available that allows the data transmission to be recorded for further analysis to aid in problem diagnosis. Traditional time-domain reflectometry test equipment is extremely useful in locating discontinuities in long (two to three miles) signal transmission cables. This is particularly useful if the signal path is buried or not directly visible or accessible.

8. PROBES

Probes are used to make actual contact with the electrical circuit or, in the case of noncontact measurements, to intercept signals emanating from conductors within the circuit for the purpose of making test measurements. There are several choices of probes, particularly for use with oscilloscopes. A probe may consist of nothing more than an insulated conductor such as a piece of wire, or may include active circuitry or clamp-on sensors used for making current measurements, or sophisticated, active probes used to make high-frequency or low-voltage measurements. An overview of the various types of probes commonly available is shown in Fig. 6.

8.1 Voltage Probes

Voltage probes are the most common type of probe. They can be relatively simple, as in the case of the passive probe, to more complex, special-purpose probes used for high-impedance, high-voltage, or differential-voltage measurements.

8.1.1 Passive Voltage Probes Passive probes do not provide any amplification of the measured signal and are the simplest and most common type of probe.

8.1.1.1 Low Impedance. Passive voltage probes can be as simple as an insulated conductor but more commonly employ some

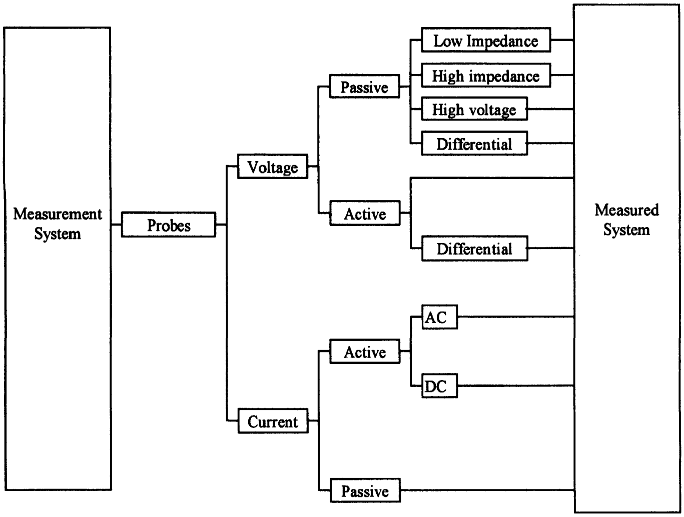

FIG. 6. Common probe types.

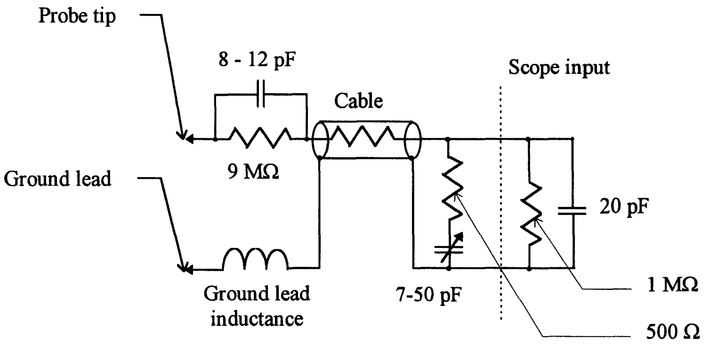

FIG. 7. Typical high-impedance 10× passive probe.

form of resistance configuration and are referred to as attenuating passive probes. These probes are designed to provide a convenient and inexpensive method of acquiring signals from most common circuit configuration and at voltages typically ranging up to 400 V. These probes present a 50-Ω impedance to the circuit under test. The common 10× passive voltage probe presents a higher impedance to the circuit and is useful up to frequencies of approximately 5 MHz. The main disadvantage of this type of probe is its relatively high input capacitance, which results in a decreasing input impedance with increasing frequency.

8.1.1.2 High-Impedance Passive Probes. High-impedance passive probes are similar in physical design to low-impedance probes. The difference is in the level of impedance presented to the circuit under test. The high-impedance probes are available in 1× and 10× attenuation and have a lower capacitance and therefore a higher impedance. This makes them suitable for higher-frequency applications, typically up to 500 MHz. A diagram of a typical 10× probe is provided in Fig. 7.

8.1.1.3 High-Voltage Probes. These probes are typically designed for use in measuring voltages in the range of 2.5–20 kV and frequency bandwidths to 250 MHz. High-voltage probes are available with 1×, 10×, 100×, and 1000× attenuation factors. This type of probe is generally used in power-conversion applications.

8.1.1.4 Differential Passive Voltage Probes. In some cases it is desired to know the value of a voltage that is referenced to some point other than circuit ground. Such measurements are referred to as differential-voltage measurements. An example is shown in Fig. 8. A measurement of this type is useful for determining voltage drops within a circuit, particularly in troubleshooting applications or design verification. Neither terminal of the instrument can be grounded (at or within the instrument) or it will effectively short R_2, resulting in a faulty measurement and possible damage to the circuit and/or instrument. This type of measurement can be made routinely using a handheld, battery-powered instrument. It is important, however, that the isolation voltage rating of the instrument not be exceeded. To measure accurately the voltage

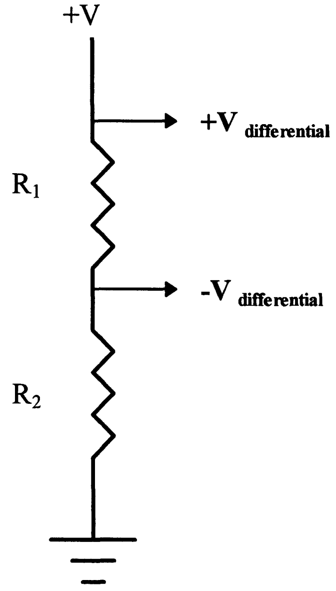

FIG. 8. Example of differential voltage measurement.

level difference between two points, the system used for the measurement must reject signals that are common to both test points and measure only the voltage difference. The degree to which the system rejects these common mode signals is called the Common-Mode Rejection Ratio (CMRR).

Such measurements can usually be made using battery-powered equipment since the electrical test equipment is effectively isolated from earth and circuit ground. Such measurements may also be made by using both channels of a dual-channel oscilloscope or with an active differential oscilloscope probe. On some oscilloscopes, it is necessary to invert one of the channels and subtract the measurement for the channels to obtain the differential. The need to invert and add the two channels may not be necessary on newer, high-end oscilloscopes that incorporate math capabilities. In this case one need only indicate that channel B is to be subtracted from channel A and the oscilloscope performs the required operations, thus yielding the differential signal.

Such a measurement is very useful in circuits that use differential drivers, such as in digital communications circuits. Differential driver circuits are used in such applications to avoid the problem of inconsistent ground levels between two widely spaced facilities. Since each side of the circuit is referenced only to the other side and not to a specific ground level, there is no uncertainty resulting from differences in ground potential between the two ends of the circuit. Such a configuration is shown in Fig. 9.

Another method of avoiding ground loops and also providing a very high degree of circuit isolation is optical isolators and fiber-optic links. In these applications, the electrical signal is converted to light, transmitted, then converted back into an electrical signal. Because glass fiber is used to carry the signal, there is no direct electrical connection between the circuits and therefore no possibility for grounding problems. Optical isolators are individual circuits that are commonly used within circuits to provide local isolation between sections of a circuit.

8.1.1.5 Active Voltage Probes. These probes contain a field-effect transistor (FET) input circuit, which presents a very high input impedance to the circuit being measured. These probes are generally available in 1×, 10×, and 100× attenuation configurations with input capacitances on the order of 0.4 to 3 pF and input resistance from 1 to 10 MΩ. Input voltage ranges for these probes are generally limited to a maximum of ±200 V (dc + peak ac). They have bandwidths to several gigahertz depending on the particular model and manufacturer.

Specific power supplies are available to allow this type probe to be used with equipment that does not directly support such probes. This type of probe is used in the design of high-speed analog and digital circuitry.

8.1.1.6 Differential Active Voltage Probes. These probes are used in the same manner as passive differential probes but are equipped with active FET amplifiers in each probe to reduce circuit loading. In addition, these probes are used with a dedicated differential amplifier to provide a single-ended

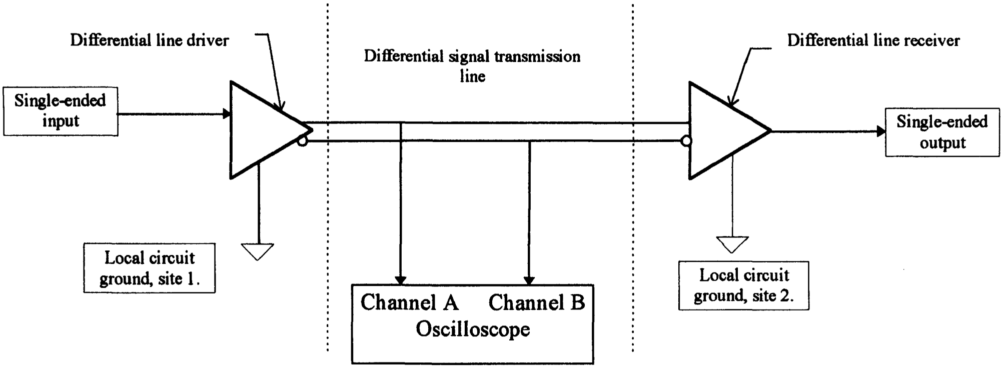
Differential line driver
Single-ended input
Local circuit ground, site 1.
Differential signal transmission line
Channel A Channel B
Oscilloscope
Differential line receiver
Single-ended output
Local circuit ground, site 2.

FIG. 9. Differential driver/receiver link used to avoid ground loops.

output. These probe/amplifier systems provide high sensitivity and high CMRR (typically 100 dB).

8.2 Current Probes

The measurement of current flowing within a conductor may involve insertion of the measurement device directly in series with the circuit or encircling the conductor with an appropriate probe that intercepts the magnetic field induced by the flowing current. The direct insertion of the measurement device into the circuit generally does not involve the use of a probe. The encircling of the conductor does require the use of some form of current probe. Current probes may be constructed in two styles; split core and solid core. In order for the current probe to couple with the maximum magnetic flux, it must encircle the conductor in which the current is to be measured.

The probes described in this section are generally available in current ranges from 200 to 300 ma to several hundred amperes dc, 1 to 2 amperes to 2 kA ac. They are useful at bandwidths of dc up to 1 GHz depending on the particular manufacturer and configuration.

8.2.1 Passive Current Probes This type of probe is used for the measurement of ac currents. The ac current causes a buildup and collapse of the magnetic-flux field surrounding a conductor. The passive probe utilizes a transformer coupled to this flux field to produce a voltage. This voltage is then used as a measure of the current flowing in the conductor. These probes are termed passive since they do not require any external power to operate. In a split-core configuration passive probe, the probe can be opened and clamped around the conductor. In a solid-core configuration, the conductor must be passed through the core, thereby requiring that the circuit be broken for the insertion of the probe element. When a solid-core configuration is used, it may be included as a permanent or semipermanent part of the circuit with the measurement instrument connected to the probe when a measurement is to be made.

A disadvantage of the transformer configuration is that the transformer core may become saturated as a result of a large current flowing in the conductor, thereby rendering the probe highly inaccurate. A key factor associated with current probes is the amp*second product. The saturation of the core material is a function of both the average amplitude of the current and the time for which it is applied. The product of these two quantities is called the amp*second product. Generally speaking, each current probe manufacturer will have a specified limit for the amp*second product. As along as the probe is operated in conditions that do not exceed this specification, the probe output will remain linear and provide an accurate measurement.

An approach taken to overcome this problem is the application of a reverse current, which is equal in magnitude to the current being sensed, through the probe. This has the net effect of minimizing the level of current "seen" by the probe and therefore avoiding core saturation.

8.2.2 Active dc Current Probes The varying flux field surrounding a conductor carrying an ac current provides a ready means of measuring the current. However, in a circuit in which a dc current is flowing, there is a steady-state magnetic flux surrounding the conductor. In this case, a Hall-effect device is used to sense the magnetic flux and convert it to a voltage signal. dc current probes are termed active probes in that they require external power to operate the Hall-effect device.

8.2.3 Active ac Current Probes Active ac current probes incorporate both the transformer principle used in passive ac probes and the Hall-effect principle used in dc current probes. This configuration is useful in applications that involve an ac current component in conjunction with a dc current component. This type of probe configuration allows such currents to be measured simultaneously with a single probe. This type of probe is also needed to measure signals such as a square wave where the signal may switch on and off and never go below zero. In this case, not only is there an ac component, but there is also a significant dc current component.

8.2.3.1 Clamp-On. While voltage measurements are made across terminals within a circuit, conventional current measurements require the breaking of the electrical circuit and the insertion of a shunt. Clamp-on probes eliminate the need to break the electrical circuit and are now the most generally used instrument for measuring both ac and dc current. These devices include both Hall-effect sensors and inductive-pickup types. Clamp-on ammeters are very common and are used for ac current measurements ranging from a few amperes to several hundred amperes. These are fully self-contained devices, which have a convenient hand grip that opens the jaws of the meter and allow its placement around the conductor. Hall-effect clamp-on probes are used in probes for oscilloscopes and allow the precise measurement of low-level, fast–rise-time currents.

8.2.3.2 Special-Purpose. Special-purpose probes include low-capacitance oscilloscope probes and active probes. Active probes include some form of amplification or impedance-matching circuitry, which is powered from the test equipment or a separate power supply. This type of probe allows for the measurement of high-frequency signals, typically to 1.5 GHz. These probes have an input capacitance of <1 pF and an input resistance of 1 MΩ, which allows for the measurement of extremely fast, low-level signals that would otherwise be altered by the relatively low impedance of ordinary probes and test equipment. These probes must be adjusted to specific circuit conditions in order to make correct and precise measurements.

8.2.3.3 Calibration/Compensation. A simple probe used to make voltage measurements usually does not require calibration since it is generally nothing more than a conductive shaft. However, in situations where the probe contains internal resistive, capacitive, and/or inductive components, it is important to check the probe periodically against manufacturer's specifications, particularly in rugged or abusive environments and where critical measurements are to be made.

The more sophisticated probes such as Hall-effect, inductive, and high-frequency, active probes require more frequent and comprehensive calibration since there are more components to drift out of calibration or be affected by adverse conditions. It is customary to calibrate some types of probes for use with a specific piece of equipment (probes and the associated equipment may not be readily interchangeable).

Probes used with oscilloscopes are generally equipped with an adjustment that allows for compensation for the probe's impedance effects. This may be accomplished by adjusting a small variable capacitor located within the probe body. On the more modern oscilloscopes, automated compensation and calibration features are provided. The user need only connect the probe to the output terminals of the calibrated signal source provided on the oscilloscope front panel and select *gain compensation.* The oscilloscope proceeds to perform the necessary gain compensation and verification of probe performance.

9. FACTORS AFFECTING ELECTRICAL MEASUREMENTS

In order for an electrical measurement to be of practical value, it must be repeatable, accurate, and stable. Several factors can have an adverse effect on making electrical measurements that meet these criteria. The most common include electrical and/or magnetic noise, poor or improper grounding of either the test equipment, the equipment to be measured, or both, and loading or otherwise influencing the circuit to be measured by the very application of the electrical test equipment. Each of these areas is addressed in more detail in the following sections.

9.1 Noise

For the purposes of this discussion, noise is defined as any signal other than the desired signal to be measured. There are several sources of noise that may be encountered in the course of making electrical measurements. Noise may be generated within the circuit being measured or may be generated external to the circuit and coupled into the circuit. The specific units of measurement of noise are dBm/Hz, dBμV/Hz$^{1/2}$, and μV/Hz$^{1/2}$. Noise is of serious concern in making accurate and repeatable electrical measurements. If a sensitive, low-level measurement is to be made, but the desired signal is "buried" in a background of noise, steps must be taken either to better identify the desired signal in the presence of the noise or minimize the noise (band-pass, band-limiting filters). In many cases, excessive noise may be attributable to poor or nonexistent grounding. An example might be in making measurements with an oscilloscope probe without properly grounding the probe or scope.

9.1.1 Externally Generated Noise In many situations the signals generated by the commercial power grid can be coupled into the test equipment and/or the circuit to be measured. The noise manifests itself at a frequency that corresponds to the power-line frequency, typically either 60 or 50 Hz (other frequencies may be encountered depending upon the ac power source; e.g., 400 Hz in military applications), and also at harmonics of the line frequency. Mechanisms for noise transmission and coupling include inductive pickup, capacitive coupling, electrostatic pickup, and ground loops. Inductive pickup occurs since any conductor carrying a varying current is capable of inducing a similar current in any conductors that lie in close proximity and are oriented so as to maximize coupling. A common mechanism for such coupling is the alternating fields generated around power wires installed in the walls, floors, and ceilings of buildings. Unless proper steps are taken to shunt or shield sensitive circuits and test equipment from such fields, low-level ac signals will be induced within the circuits. Electrostatic pickup occurs as a result of a circuit or test equipment being placed within a space that contains a high electrostatic gradient. Electrostatic pickup is most noticeably encountered when using high–input-impedance test equipment that is not properly connected to the circuit to be measured or is not properly grounded. Other sources of externally generated noise include fluorescent lighting, induction motors, and electrical contacts in motor and power equipment.

9.1.2 Internally Generated Noise Electrical circuits are also capable of generating noise internally. Noise internal to a circuit is generated by resistors, semiconductors, vacuum tubes, and virtually every component of the circuit. Although the magnitude of the noise may vary considerably depending upon the mode of generation, it is still necessary to have at least an awareness and understanding of noise sources internal to the circuit being measured. All conductors and semiconductors in a circuit generate *thermal noise* (also known as *Johnson* or *Nyquist* noise). This noise is produced as the result of thermal agitation of free electrons. The noise voltage generated across a resistor can be expressed as

$$\overline{e_n^2(t)} = 2kTRd\omega/\pi,$$

where $k = 1.38 \times 10^{-23}$ JK (Boltzmann's constant), T = temperature in Kelvins, R = resistor value in ohms, and $d\omega/2\pi$ = frequency band (centered at angular frequency ω). Thus, the resistor appears as a voltage source within the circuit with a mean squared open-circuit voltage of $\overline{e_n^2(t)}$.

9.1.3 White and Colored Noise The term white noise is derived from the concept of white light. White light contains a multitude of frequencies of light, red, violet, blue, green, etc. The term *white* noise as applied to electrical noise has the same connotation, the difference being that electrical noise does have a restricted bandwidth. The specific technical definition of white noise is noise such that the power in a frequency interval of fixed size is the same regardless of the interval chosen. Any other noise is *colored;* e.g., "pink" noise has higher power in the lower frequencies. Probably the most predominant noise is that found when trying to tune an AM short-wave radio, the ever present background white noise signal. Similar noise signals are generated when appliances that employ brush-type induction motors are operated. The noise generated by the arcing brushes generates a reasonably broad-band noise signal. Other sources of noise include fluorescent lamps, solid-state motor speed-control circuits, and automotive ignition systems.

9.1.4 Signal-to-Noise Ratio The magnitude of the desired signal as compared to the background noise level is referred to as the signal-to-noise ratio (SNR). It is desirable to maximize this ratio in order to obtain the clearest signal. This may be accomplished by employing properly designed filter circuits or by controlling the noise contribution of circuits such as amplifiers. An amplifier used to boost the desired signal also contributes noise since it too contains elements that are noise sources. The signal-to-noise ratio is measured in decibels, as this provides a wider dynamic range of measurement.

9.2 Grounding

There are basically two types of grounding: direct connection to the earth, and connection of power circuits to the circuit ground (which may in turn be connected to earth ground or not). Poor grounding can contribute significantly to the pickup and generation of noise. Such faults can arise from many sources, all associated with faulty grounding. In addition to erroneous measurements and confusing results, poor grounding may also result in personal injury. If a circuit is designed to be connected to earth ground, it must be properly connected to earth (safety) ground to eliminate the possibility of noise generation or electrocution. If a circuit is designed to "float" but is inadvertently connected to ground through dampness, contamination, or improperly connected equipment, erroneous measurements may arise. Many pieces of electrical test equipment are designed with internal shields around sensitive circuits within the instrument. These shields are grounded both to the internal circuit ground and to earth ground through the three-prong plug. This provides a direct path to ground for any spurious signals that might otherwise be picked up by such circuitry. A common problem in making electrical measurements is referred to as the "ground loop." Ground loops may originate from several situations but are generally attributed to leakage paths between circuits. Ground loops may be avoided by carefully reviewing grounding connections, and providing circuit isolation (e.g., optical isolators or transformer coupling).

Grounding problems can also arise as the result of routing the "return" circuits for power circuits back through a common, improperly sized ground conductor. The ground (negative) conductor has resistance and therefore results in a voltage drop across it. If the ground conductor is not of sufficient size, a substantial voltage drop across the ground conductor may result. This causes the circuit to rise to some potential level above ground potential. This may not pose a significant problem unless the circuit is connected to other circuits that are referenced to ground through alternate paths. An example of how such a situation may arise is illustrated in Fig. 10.

9.3 Loading

The direct attachment of electrical test equipment to the circuit to be measured may influence the measurement as a result of *loading*. Since the voltmeter (or similar types of equipment) is an electrical device, it possess the same characteristics as the circuit being measured. The degree to which the measured circuit is influenced depends upon how close the impedance of the device being used is to the impedance of the circuit being measured. An example is shown in Fig. 11. The meter being used to make the measurement has an input impedance of 1 MΩ. The portion of the circuit being measured has an equivalent impedance of 10 MΩ. The introduction of the test equipment in order to make the measurement alters the circuit being measured by effectively introducing a "new" circuit configuration. Such conditions not only can produce erroneous measurements, but in the case where a circuit may be oscillating, an undesired condition, the introduction of test equipment may alter the circuit to the point where it no longer oscillates. Under such circumstances, what is ob-

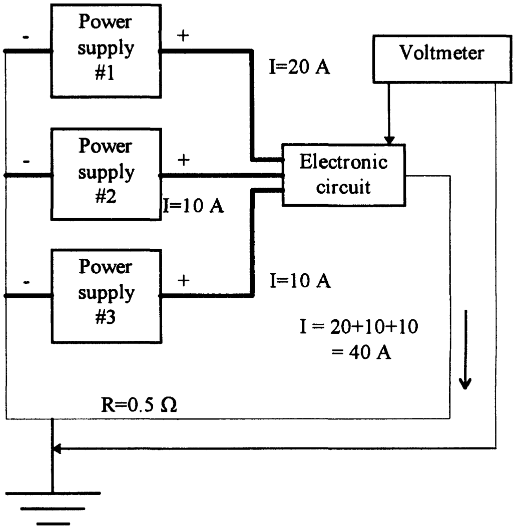

FIG. 10. Improper grounding through undersized conductor. Voltage drop across ground return is 0.5 Ω times 40 A or 20 V, therefore the circuit being measured is 20 V "above" ground.

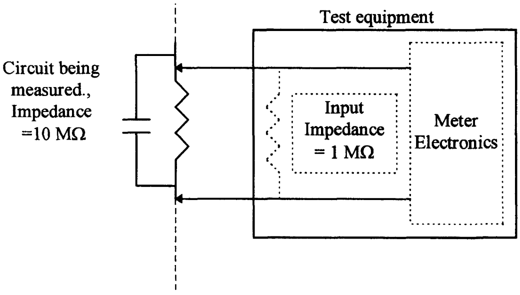

FIG. 11. Example of circuit loading by test equipment.

served when the test equipment is not connected may not agree at all with what is seen after the equipment is attached. Care should always be taken to observe the effects the test equipment may have on the system being measured.

9.4 Aliasing

Aliasing occurs when a measured signal has frequency components at or higher than the sampling rate of the digitizing equipment. Such high-frequency components can result from the inherent noise of the measurement system or from the system being measured. Sources of such noise include 50- or 60-Hz pickup from the power lines, broadcast stations, mechanical vibrations, and other sources. High-frequency components may also result from sharp transitions in the measured signal. If the signal being measured is not properly band-limited to prevent these higher-frequency components from being sampled, then they may "fold back" into the signal of interest and result in measurement error. The sampling frequency is less than twice the highest frequency of the signal being sampled; therefore, a lower-frequency component is introduced resulting in an erroneous measurement. This problem occurs more frequently in data-acquisition systems that do not incorporate proper anti-aliasing filtering at the input. Once the signal is sampled and the aliased signal is present in the data, it cannot be removed. It is therefore important to band limit the input to the measurement system or make certain that the sampling system is capable of sampling at least twice the frequency of the signal to be sampled (Nyquist criteria).

10. ELECTRICAL TESTING EQUIPMENT PHYSICAL CONFIGURATIONS

Electrical test equipment comes in several configurations owing to the rapid development of microcomputer, miniaturization, software, materials, battery, power supply, and integrated circuit technologies. Test equipment functionality once available only in bulky equipment that had to be powered from line voltage has evolved from bench-top and rack-mounted configurations to more portable and even handheld configurations. Sophisticated measurement capabilities including full-function oscilloscopes are now available in handheld, battery-powered configurations.

10.1 Handheld

Handheld test equipment refers to that category of equipment that can be transported and used while holding the instrument in one hand and probing the circuit to be measured with the other hand. Until the past few years, handheld electrical test equipment was generally limited to volt-ohmmeters (VOMs), or multimeters such as the familiar Simpson®/volt-ohmmeter. Advances in microcircuit, microcontroller, and microcomputer technologies have allowed more sophisticated functions to be incorporated into handheld test equipment. Examples of these include digital oscilloscopes, signal acquisition and analysis equipment (digital signal processing and spectral analysis), and handheld successors to the traditional multimeter that include meters with combined analog and digital displays. These meters provide the ability to measure very accurately ac and dc voltage, ac and dc current, true rms voltage, resistance, frequency, capacitance, and diode test, as well as other functions depending on the particular model. Handheld, LCD-based oscilloscopes are now widely available and can prove very useful in making field as well as bench-top measurements.

Although they may be line powered, this class of equipment is almost always battery powered in order to provide maximum utility and application in field environments. When using this type of equipment to make measurements where the entire instrument will be above ground potential (i.e., floating), it is *extremely* important to follow the manufacturer's warnings regarding maximum isolation voltage for the instrument. This is the maximum voltage to which the instrument's overall insulation is rated. If this voltage level is exceeded, it is possible for the electric current to pass from the test points, through the instrument, and into the body of the person holding the instrument. While such precautions also apply to portable and rack-mounted equipment, these types of equipment are usually connected to earth or safety ground and therefore provide an alternative path in the event of dielectric breakdown.

10.2 Portable

Portable electrical test equipment generally refers to larger–form-factor test equipment that can be transported by a single individual and is intended to be used in remote or field locations or merely for ease of transfer from bench to bench. Such equipment includes oscilloscopes, spectrum analyzers, power supplies, function generators, and similar equipment.

10.3 Rack-Mounted

Traditionally the term rack-mounted referred to test equipment that was semipermanently mounted into a metal rack. The opening of the rack was based on a standard 19-in. (48.3-cm) opening (to mounting centers). The instruments, while varying in height and depth, must conform to this standard horizonal rack-mounting requirement. Virtually every type of electrical test equipment is currently available in this form factor.

Newer test-equipment configurations are now also based on vertical rack arrangements referred to as card cages. The most common of these are the NIM, CAMAC, VME, VXI, and MMS configurations. The NIM and CAMAC formats have been used for many years, particularly in high-speed nuclear instrumentation. The VME, while considered to be more of a computer-oriented instrument format (digital input/output, analog-to-digital converter, and digital-to-analog output), is the predecessor to the more general class of modular instruments known as VXI. MMS or Modular Microwave System modules also include a wide assortment of instruments but do not seem to be as popular as the VXI because of their higher cost. Unlike the 19-in. rack-mounted configuration, which imposed only the physical width requirement, these rack-mount standards dictate module heights, widths, and depths, maximum power consumption, maximum cooling requirements, and the relative degree to which instruments may interfere or interact with their "neighbors." The standards also dictate how data and information are to be communicated between modules and between a module and the "crate controller."

10.4 Instrument-on-a-Card

The advent of instruments adapted to direct insertion into the personal-computer backplane and even into laptop and notebook computers has led to complete instruments being placed on a single card using the computer itself as the display and control interface to the instrument. The instrument-to-computer interface may be in the form of PCI or EISA for rack-mounted or desktop computer configurations, or PCMCIA for laptop and notebook computer configurations. While not yet providing the level of sophistication of dedicated, application-specific electrical test equipment, such implementations can prove useful in some applications. It is important to note that the performance of this type of instrument can be heavily dependent upon the type and performance of the computer to which it is attached and is not intended to replace high-performance specialized test equipment. Examples of such instrument-on-a-card configurations include spectrum analyzers, voltmeters, and oscilloscopes.

11. DISPLAY CONFIGURATIONS

The presentation of information obtained from electrical test equipment is important in interpreting the test results. The advent of new display technology has made it much easier to interpret data and even to present data in graphic form in handheld instruments.

11.1 Analog

The traditional method of displaying test measurement data has been through the use of a simple analog meter. The advantages of the analog meter include low cost, simplicity, and ease of reading. Although well suited for simple voltage and current display, the conventional analog meter display has been limited in application to single-parameter display and for "relative indication." This latter function, relative indication, refers to the ability to just glance at an analog display and quickly get an assessment of the magnitude of the parameter without actually having to "think" about the precise measurement. This had led to the implementation of bar-graph and other analog-type indicators using digital-display technology. Some instruments implement the traditional analog "meter" display using digital "icons," which mimic a traditional analog meter to provide the same "quick look" capability.

11.2 Digital

The advent of the light-emitting diode (LED) display and liquid-crystal display (LCD) have made it very easy to display data in digital form. In a digital-display–based instrument the voltage measurement is made through the use of an analog-to-digital conversion circuit. The measurement is quantified and sent to a digital-display circuit, which conditions the data for the LED or LCD display unit. The LED display offers the advantage of ease of viewing in low–light-level conditions at the expense of requiring significant power to energize the display elements. The LED display may also be difficult to read in high ambient light such as direct sunlight. The LCD display offers the advantage of extremely low power consumption. The primary disadvantage of the LCD is the difficulty in viewing the display in areas of low light level. Provisions have been made in some instruments to provide a back-lighting capability in which a small lamp may be energized to illuminate the display from behind, thereby making it viewable. Of course this is counter to the reason for using an LCD display, i.e., low power consumption. In field applications where instruments may be exposed to extremes of temperature, consideration should be given to the tendency of the LCD display to blank out or become inoperative at very low (freezing) or very high temperatures (e.g., inside of vehicles).

11.3 Graphic

Often the most useful form of data presentation is in graphic form. This is particularly useful where trends are to be observed or high-speed events need to be observed. Until recently the most common form of graphic data display has been the oscilloscope. These instruments provide a graphic plot of voltage variations as a function of time. Many instruments now utilize CRT (and LCD) technology to provide graphic presentations of complex data. Common examples of these include spectrum analyzers and network analyzers, which display the results of FFT transforms directly on the screen. Increased use of microprocessor and LCD technology have resulted in even more powerful methods of graphic data presentation. A very useful application of this technology has been in the display of complex information in handheld instrumentation such as handheld, battery-powered oscilloscopes. The introduction of a wide variety of plug-in boards and modules for the personal computer combined with the processing power and color display capabilities of the personal computer have made the graphic display of information in real or near-real time a relatively simple task.

12. VIRTUAL INSTRUMENTATION

Traditionally an electrical measurement was made by manually interacting with the front panel of the instrument to select type of measurement, set the desired measurement range, initiate the measurement, and record the results. The advent of the *virtual instrument* concept makes it possible to interact with electrical test equipment without ever having to touch or view the actual equipment directly. Essentially the virtual instrument is a graphic representation of the controls and display found on an actual instrument. The measurement to be made using the instrument is accomplished through the use of software commands to various modules or, in some cases, actual instruments. An example would be the display of the controls and displays of a conventional multimeter on the CRT display of a personal computer. The user interacts with the virtual instrument through the use of the keyboard or *mouse,* selecting the desired measurement, etc. The computer, which is connected to an electronics module (which has no front-panel display or controls—e.g., VXI module) through a communications link, transmits the control information to the module, acquires the measurement, and displays it on the computer monitor. The advantages of such instrumentation is that it minimizes desk or rack space that would be required for conventional rack-mounted equipment, allows sharing of instruments among several users, supports more sophisticated data analysis, recording, and sharing, decreases weight, permits virtual interconnection of equipment, and provides ease of measurement automation.

13. INTERFACING

In many cases it is desirable to acquire measurements from electrical test equipment that is situated in a remote or not easily accessible location. Or it is not economically feasible to measure and record the data manually. Rather than manually record data obtained by an instrument, it is possible to have the instrument automatically polled to obtain the data from a remote location and transmit the information to a central location for processing. This section describes some of the mechanisms that provide for easy interfacing between electrical test equipment.

13.1 General-Purpose Instrument Bus (GPIB)

Many instruments provide the ability to be interconnected by means of standard interface mechanisms. One of the most popular for individual instruments is the IEEE-488 protocol, also referred to as the *general-purpose instrument bus or GPIB.* This standard employs a well-defined standard cable and connector configuration to support interconnection of instruments and computers. It also employs a well-defined communication language protocol. This language consists of a library of specific function calls, which may be used both to program and to interrogate an instrument. The commands are listed in the user's manual, with some being unique to a particular instrument.

Through the application of such an interconnecting link, it is possible to build networks of instruments and to acquire data automatically from many instruments. A key feature of the GPIB interface is that it is a multidrop link. This means that many instruments may be linked to a single cable link and that new instruments may be added by just connecting them to the existing link by means of an additional cable. This ability to interrogate, configure, and control specific instruments over a computer link is in itself a very powerful feature. It also plays a critical role in the application of virtual-instrument concepts (see Sec. 12).

13.2 Dedicated Instrument Bus

Dedicated instrument buses (e.g., VXI) are those that are based on a back-plane configuration designed for specific instrument physical configurations. These buses generally provide a means for instruments to communicate with a computer to support configuration, control, and interrogation. These buses may be used to provide a compact configuration, high-speed measurement, multipurpose functionality, or a combination of these. Provision is usually made through the use of bus expanders to daisy-chain back-plane enclosures, thus providing for expansion.

13.3 Serial Instrument Communication

Serial data communication links (e.g., RS-232, RS-422) provide a low-cost and relatively simple method of obtaining data from electrical test equipment and in some cases providing a means to remotely configure the equipment.

13.3.1 RS-232 A very common instrument link is the RS-232 or Serial Link. This link is used primarily between a single instrument and a computer or data-logging equipment. The term "serial" refers to the fact that the information to be transmitted (in digital form) is sent one bit (binary digit) at a time in a serial fashion. This standard originated primarily for terminal-to-computer and computer-to-computer communication. Its use has since been extended to device-to-device and device-to-computer communication. The major disadvantage of RS-232 links is the inability to link more than one device to a communications path; i.e., it does not provide multidrop capability. Each device must have its own dedicated link to the central device (computer or data logger).

13.3.2 RS-422 The RS-422 serial communication standard was developed to counter the lack of multidrop capability in the RS-232 standard. The RS-422 standard allows up to 16 devices to be "dropped" from a common communications "backbone." The standard provides the mechanism for individually addressing each instrument on the bus and receiving responses from the instruments.

14. CONCLUSION

A wide variety of electrical test equipment is available, each well suited to specific areas of electrical circuit measurement. The advent of and continued progress in microprocessor and microcontroller technology, integrated circuits, and software have provided ever smaller and more powerful electrical test equipment. Much of the analysis that had to be performed manually prior to the advances in instrumentation is now performed routinely within the instrument. The ability to link a variety of electrical test equipment into a "measurement system" configuration and to operate in a virtual-instrument environment is a very powerful consideration when large numbers of measurements must be made or changes in measurement configuration are frequent.

Almost all electrical test equipment may be used to obtain measurements across the areas of power sources, power conversion and conditioning, component measurement, and signal measurements. However, some types of equipment are better suited to specific measurement areas than others. The user should carefully consider the specific application and determine the best type of equipment to use. A review of the free (or low-cost) applications notes and test-equipment data available from electrical test-equipment manufacturers can prove to be very helpful when deciding how best to make a measurement or the type of equipment to use.

GLOSSARY

Active Components: Electrical circuit elements such as transistors and tubes. Active components are used in the generation, control, and amplification of electrical signals. As examples, active components serve as switching elements, signal amplifiers, and oscillators. (See also **Passive Components.**)

Active Probe: An instrumentation probe that contains active elements such as field-effect transistors and is designed to provide signal amplification or to minimize loading effects on the circuit being measured.

Common-Mode Rejection Ratio (CMRR): A figure of merit for a differential amplifier. It is defined as $CMRR = |A_d/A_c|$, where A_d is the voltage gain for the difference signal and A_c is the voltage gain for the common-mode signal. A perfect differential amplifier has a CMRR equal to infinity. The higher the CMRR for the difference measurement system, the closer it is to the ideal. As an example, with a system CMRR of 10 000:1, a common mode input of 5 V will result in an output of 0.5 mV. CMRR may also be expressed in dB: CMRR (dB) = 20 log CMRR; e.g., a CMRR of 10 000:1 would be a CMRR of 80 dB.

Effective or rms Value: The effective or rms value of a varying power is the measure of the amount of energy being transferred in an electric circuit. The true rms value is the rms value of a voltage independent of the wave form of the voltage.

Hall Effect: The effect whereby a steady current, flowing in a steady magnetic field, gives rise to an electromotive force that is at right angles to both the magnetic field and the current. The electromotive force is proportional to the product of the intensity of the current, the magnetic field, and the sine of the angle between the direction of these quantities.

Instantaneous Power: Instantaneous power for an alternating voltage and current is the product of the instantaneous voltage and instantaneous current at the same instant.

Passive Components: Electrical circuit elements such as resistors, capacitors, and inductors. Passive components, by themselves, do not alter the frequency or wave form of an applied signal. They may only introduce phase shift and/or alter the magnitude of an applied signal.

Passive Probe: An instrumentation probe that consists simply of a conductor or, in the case of attenuated probes, resistive and capacitive elements.

Further Reading

Fink, D. G., Beaty, H. W. (1987), *Standard Handbook for Electrical Engineers,* 12th ed., New York: McGraw-Hill.

Hewlett Packard (1996). *Test and Measurement Catalog,* Boise, ID: Hewlett-Packard.

National Instruments (1995), *Instrumentation and Reference Catalogue,* Austin, TX: National Instruments.

Peebles, Jr., P. Z. (1980), *Probability, Random Variables, and Random Signal Principles,* New York: McGraw-Hill, pp. 196–202.

Sheingold, D. H. (1972), *Transducer Interfacing Handbook,* Norwood, MA: Analog Devices, Inc.

Strassberg, D. (1996), "Digital Oscilloscopes: For Best Results, Understand How They Work," *Electron. Design News,* July 4, 43–56.

Tektronix (1995), *ABC's of Probes,* Beaverton, OR: Tektronix, Inc.

Tektronix (1993), *Probing High Frequency Digital Circuitry,* Beaverton, OR: Tektronix, Inc.

Tektronix (1996), *The Effect of Probe Capacitance on Measurement Accuracy,* Beaverton, OR: Tektronix, Inc.

Tektronix (1996), *Measurement Products Catalog,* Beaverton, OR: Tektronix, Inc.

Tektronix Catalog (1996), *Probe Tutorial,* Beaverton, OR: Tektronix, pp. 436–496.

Witte, R. (1992), *Electronic Test Instruments, Theory and Applications,* New York: Simon and Schuster.

Witte, R. (1992), *Spectrum and Network Measurements,* New York: Simon and Schuster.

TESTING EQUIPMENT—MECHANICAL

ROBERT T. REESE, *Sandia National Laboratories, Albuquerque, New Mexico, U.S.A.*

	Introduction	1
1.	**Control Methods and Calibration**	3
2.	**Testing Systems Used for Determining Mechanical Responses and Properties of Materials**	3
2.1	Dead-Weight Testing System	3
2.2	Screw-Driven Testing Systems	4
2.3	Tension and Compression Hydraulic Testing Systems	5
2.4	Servo-Hydraulic Testing Systems	5
2.5	Structural Frames and Test Beds	6
2.6	Time-Dependent Tests and Equipment (Creep and Fatigue)	8
2.6.1	Creep	8
2.6.2	Fatigue	10
2.7	Fracture-Mechanics Tests and Specimens	11
2.8	Pressure Systems and Vessels	13
2.9	Thermal Systems	15
2.10	Modal Analyses	16
2.11	Inertial Systems (Centrifuges and Spinners)	16
2.12	Vibration Systems	17
2.13	Shock Systems	19
2.14	Hardness Tests and Equipment	21
3.	**Hazards Involved in Mechanical Testing**	22
	Glossary	23
	Works Cited	23
	Further Reading	23

INTRODUCTION

This article focuses on determining the mechanical properties of materials as well as the mechanical response or structural integrity of components and systems (assembled parts). Success in testing depends on how tests are planned, what insights are desired, and how tests are conducted. Specific reasons for performing tests include the following:

- to complement and confirm analyses of the response of mechanical systems to applied forces;
- to determine mechanical properties of various materials as functions of temperature, strain rate, and other test conditions;
- to satisfy laws or regulations such as acceptance of a part or system for service in commercial or military applications;
- to determine acceptance of mechanical parts by simple pass/fail evaluation criteria (e.g., a pressure proof test on a gas cylinder);
- to determine mechanical behavior independent of time (e.g., static strength tests on test items and assemblies);
- to determine time-dependent dynamic responses of materials and assemblies (e.g., fatigue resistance of critical composite aircraft assemblies); and
- to serve as the only source of understanding as to how assembled parts actually respond.

The minimum requirements to perform mechanical testing are that

1. a physical item (e.g., specimen or structure) must be available for testing,
2. test equipment must be available for use and capable of imposing a measurable stimulus to the test item, and
3. the test equipment and/or test item must be equipped with transducers or other devices needed to measure the stimulus to and response of the test item.

Reprinted from *The Handbook on Structural Testing* with permission from Fairmont Press, Lilburn, GA.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Guidance for performing tests is available from many sources. Standards and procedures for tests on metals are described in the ASM Metals Handbook (1985). Standards and procedures for tests on wood and timber products are described in the USDA Wood Handbook (1987). The American Society for Testing and Materials publishes standards, procedures, and references for tests on nearly all materials in seven broad categories:

1. iron and steel (series A),
2. materials for electrical applications (series B),
3. natural and man-made non-carbon-based materials (series C),
4. typical carbon-based materials with long-chain molecules (series D with some exceptions such as soil and rock mechanics),
5. characteristics of materials (series E),
6. applications including but not limited to tires, gaskets, paper (series F), and
7. corrosion and wear (series G).

The fundamentals of structural testing are described by Reese and Kawahara (1993). This basic handbook provides the details required to perform successfully a wide variety of tests. Relationships among engineering mechanics, materials science, and structures (mechanical assemblies) are described by Blake (1985). Relationships in materials sciences were reported by Van Vlack (1974) and Davis *et al.* (1982).

Attachment of test items, assemblies, and measurement equipment to test equipment or systems is essential. Much of the art of testing originates in cleverness in design and fabrication of fixtures. Fixtures provide the boundary conditions for the test items, the means to apply forces, pressures, vibrations, or shock loadings. Fixtures are used for tension, compression, shear, and bending tests on metals, wood, plastics, and concrete and include cylindrical, rectangular, and tubular geometries. Fixtures are also important for attaching assemblies to vibration and shock test systems. The ASTM Standards (current editions), ASM Metals Handbook (1985), USDA Wood Handbook (1987), and Reese and Kawahara (1993) describe many fixtures. The importance of fixtures in testing specimens and assemblies cannot be overstated as they are the essential link between the test items and the testing systems and equipment.

Measurements are the principal products of tests. Measurements are made with sensors and transducers directly in contact with the test item or by noncontact techniques. (The use of contact or noncontact measurement techniques may be important in planning a test.) Among the test measurements reported in Doebelin (1975), Hannah and Reed (1992), Harris (1988), Tse and Morse (1989), and Wright (1994) are acceleration (linear, angular), displacement (linear, angular), force and torque, frequency, pressure, sound, strain, temperature, and velocity. In addition, nondestructive measurement techniques can also be applied to mechanical testing and contribute the following types of diagnostic methods as given in chapter 12 of Kobayashi (1987) and Reese and Kawahara (1993): acoustic emissions (sonic and ultrasonic), electromagnetic induction, magnetic, optical, penetrants, photometrics, radiography, and thermal. Tests performed on prototype aircraft often include both broad categories of experimental measurements as described in chapter 15 of Reese and Kawahara (1993). Measurements are made at local points of interest or over a portion or all of the test item or assembly.

Measurements of a given physical quantity can be made in a variety of ways. For example, test-induced displacements of an assembly or changes in length of a material test specimen can be made using

1. a dial indicator,
2. a transducer (linear potentiometer or a differential transformer),
3. ultrasonic transmission and reflection through the length or thickness of materials or parts, or
4. photometrics with camera or video coverage.

Strains can be measured using strain gages, optical moiré interferometry, and photoelasticity (specimens and coatings). The description of experimental mechanics methods and techniques were reported by Kobayashi (1987). Instrumentation and measurement systems are described by Beckwith and Buck (1973), Bendat and Piersol (1986), Herzfeld (1962), and Wright (1994).

Mechanical testing equipment supplies forces or excitations to test items. Included in the broad categories of forces and excitations are pressures, thermal gradients, shocks, vibrations, energy-producing modal or frequency responses, and other phenomena (e.g., rapid energy deposition) causing assemblies and/or materials to respond mechanically. Representative test environments that can be simulated are transportation, crash, blast, impact, hydraulic, rocket motor burn and separation, parachute loads, earthquakes, acoustic noise loads, centrifugal turbine loads, etc.

1. CONTROL METHODS AND CALIBRATION

Mechanical testing equipment is controlled in one of two basic ways: open loop or closed loop. Open-loop testing means that the equipment operator has control and provides feedback to the test equipment by modifying the rate of loading, thermal input, etc. Open-loop control is applicable to tests where time is not a critical parameter. Closed-loop testing means that a portion or all of the control of the test is given to the testing system itself. Closed-loop systems typically require interaction with and controls supplied by computers, function generators, or other electronic means. In actual practice, most test activities involve some open- and some closed-loop portions. For example, a test stimulus could be applied open loop and test results recorded closed loop.

Calibration of equipment is essential in providing quality test data traceable to national standards. This includes the mechanical test equipment, measuring transducers, and recording electronic instrumentation. The estimate of the uncertainties in measurements is an essential step in reporting test results.

2. TESTING SYSTEMS USED FOR DETERMINING MECHANICAL RESPONSES AND PROPERTIES OF MATERIALS

Test systems apply loads by mechanical and hydraulic techniques. Mechanical testing machines can apply forces through dead weights, screw drives, or other motor-driven (e.g., eccentrics and electrodynamic) coupling. Hydraulic machines can operate as tension or compression (e.g., manually operated pilot valve controlled) or tension and compression systems (e.g., servo valve controlled).

Testing systems can also be combined. For example, ovens can be installed around materials test specimens or test assemblies to obtain responses at elevated or low temperatures. Pressure systems and test items containing pressure can be subjected to multiaxial stress conditions.

The testing machines described in the four following sections (dead-weight, screw-driven, hydraulic, and servo-hydraulic) are self-contained systems as far as the forces are concerned. That is, forces are applied and equivalent reaction forces occur internal to the machines. The reaction forces and machine stability are supplied by the columns and the crossheads. When these machines have two reactive columns, they are vulnerable to the adverse effects caused by possible eccentric loads. If the forces transmitted to a test item result in a reaction force located a column diameter away from the plane or axis between the columns, the force capability of the testing system can be reduced by a factor of 4. If bending or side loads are generated during a test using any of the test systems in Secs. 2.1 to 2.4, then the use of a four-column or equivalent stiffened testing machine is recommended.

2.1 Dead-Weight Testing System

The dead-weight machine uses a series of weights attached to a lever to apply forces to test items. A typical use for dead-weight testing machines is in creep testing, in which the forces must be maintained accurately over long time periods. Force calibration is another example of the use of a dead-weight testing machine.

In Fig. 1, a lever-type dead-weight testing machine is shown. Stacked weights are shown on the elevator on the right side. The weights apply a tension force to a train and lever. The arm pivots about the fulcrum. The fulcrum is often movable so that different ratios of lengths (L_1 and L_2) can be obtained. The tensile force in the left load train resists the applied moment. The test item or specimen can be subjected to tension or compression forces depending on how it is attached

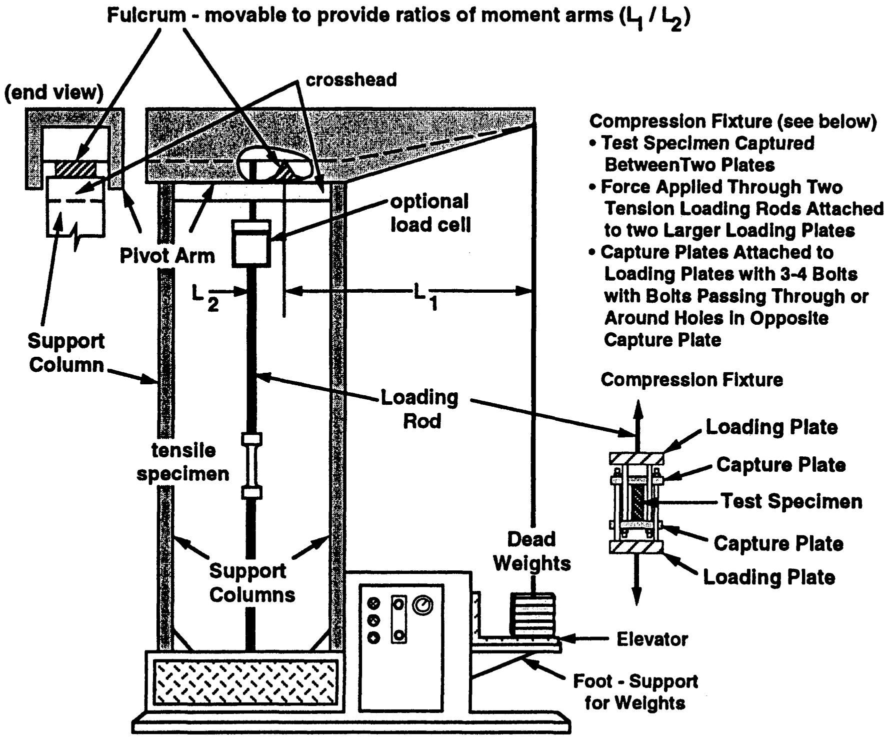
Fulcrum - movable to provide ratios of moment arms (L_1 / L_2)
(end view)
crosshead
Compression Fixture (see below)
• Test Specimen Captured BetweenTwo Plates
• Force Applied Through Two Tension Loading Rods Attached to two Larger Loading Plates
• Capture Plates Attached to Loading Plates with 3-4 Bolts with Bolts Passing Through or Around Holes in Opposite Capture Plate
optional load cell
Pivot Arm
L_2
L_1
Support Column
Loading Rod
Compression Fixture
Loading Plate
Capture Plate
Test Specimen
Capture Plate
Loading Plate
tensile specimen
Support Columns
Dead Weights
Elevator
Foot - Support for Weights

FIG. 1. Dead-weight testing system.

to this load train. Tension forces are applied by direct attachment to the load train. The applied tensile forces are resisted by a rod attached to the base plate and columns. Compression forces are applied through fixtures attached to the load train as shown in Fig. 1. A typical example of an associated fixture is a "compression cage" in which the test item is placed between two plates. The upper plate is attached to the lower tension rod, and the lower plate is attached to the upper rod. These plates have hole patterns that allow the connective hardware to pass freely through or to bypass the plates. In other variations to these machines the weights can also be applied directly to the test item.

2.2 Screw-Driven Testing Systems

An example of a screw-driven machine is shown in Fig. 2. The force is generated by moving the crosshead with lubricated, long, threaded columns or rods, which are turned by variable low-speed electric motors. Tension or compression forces are applied when

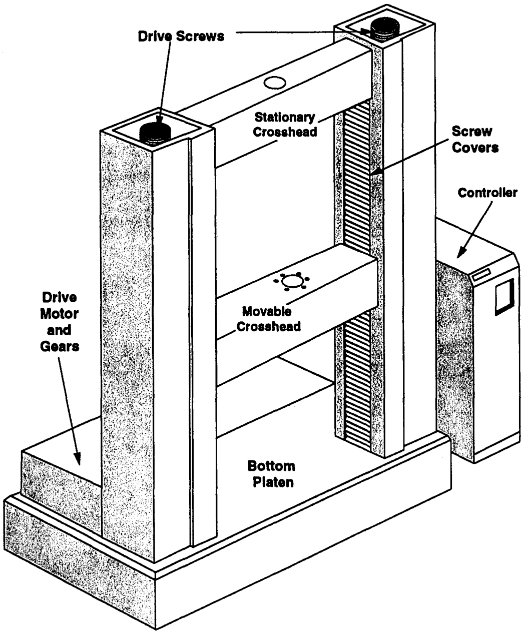
Drive Screws
Stationary Crosshead
Screw Covers
Controller
Drive Motor and Gears
Movable Crosshead
Bottom Platen

FIG. 2. Screw-driven testing system.

the crosshead moves respectively toward or away from the stationary top crosshead or the bottom platen. When tests are performed, the screws can be adjusted to provide crosshead motion at rates ranging from very slow to approximately 500 mm (20 in.) per minute. Tension or compression forces can be applied depending on the position of the test item, the direction of motion of the crosshead, and the attachment devices located in the main upper and lower support structures. For example, a test item would be in compression if it is placed in the bottom portion of the machine and the crosshead continues to move toward the base after making contact with the test item. Relatively recent advances have made it possible to apply both tension and compression forces (e.g., plus and minus torques) with very little inertial contributions using servo-controlled drive motors and linear direct-current amplifiers. The motors can be manually controlled with potentiometers, with function generators, and with interactive computer software. The motors and testing frame are protected with limit switches so that the movable crosshead does not become bound against either end and to prevent jamming the screw threads. The screws are often protected with movable covers to keep the lubricants from becoming contaminated, to keep the work area clean, to prevent entanglement of test items and hardware, and to reduce injuries to people.

2.3 Tension and Compression Hydraulic Testing Systems

These machines were originally called universal testing machines because they could supply either tension or compression forces but could not cycle between the two force modes. Since newer testing machines are capable of applying tension and compression forces, the term universal testing machine is now obsolete. These tension or compression hydraulic testing machines continue in service and have received many upgrades in terms of feedback, control, force, and displacement measurements. Figure 3 shows a typical two-column system. The hydraulic cylinder provides a force to the table. The table is attached to the upper crosshead by long columns and the applied forces can be transmitted to this upper crosshead. The movable crosshead is positioned for the test and, when stationary, provides the reaction point for the forces generated by the hydraulic cylinder. The tension space above indicates that the upper crosshead is moving away from the positioned movable crosshead during a test. The compression test area indicates that the table is moving toward the positioned movable crosshead. These machines typically have wedge grips or jaws built into or positioned in their crossheads. These grips or jaws are used to attach test items and have tension force capabilities equal to the maximum forces that can be generated by the machine.

These machines usually have 2.067-MPa (3000-psi) hydraulic cylinders located in the base of the machine. The hydraulic cylinder is machined to a tight tolerance within its case or capsule. A small amount of fluid is allowed to flow through the tolerance zone to lubricate the cylinder. The fluid thus used is returned to the reservoir, filtered, and recycled into the pump. The cylinder travels in one direction only and, since it is not sealed, can apply only compression forces on the top of the table, precluding cyclic tension and compression forces on a test item.

The control of these machines typically occurs through manual operation of pilot wheels controlling the flow of pressurized hydraulic oil into and out of the cylinder. One pilot wheel controls the loading, and one controls the unloading. Newer versions of these machines have been equipped with controls complementing manual operation so that load, stroke, and strain capabilities can be obtained.

2.4 Servo-Hydraulic Testing Systems

These testing systems are similar to the tension or compression testing machines in Sec. 2.3 in that the applied forces are transmitted hydraulically. Figure 4 shows an example of a servo-hydraulic testing system. The use of servo controls provides added test features when using this equipment. The sealed hydraulic cylinder provides actuator motion in two directions so that tension and compression forces can be applied. These systems are also equipped with built-in displacement gages (usually linear variable differential transformers, LVDTs) so that the displacement of the actuator can be defined

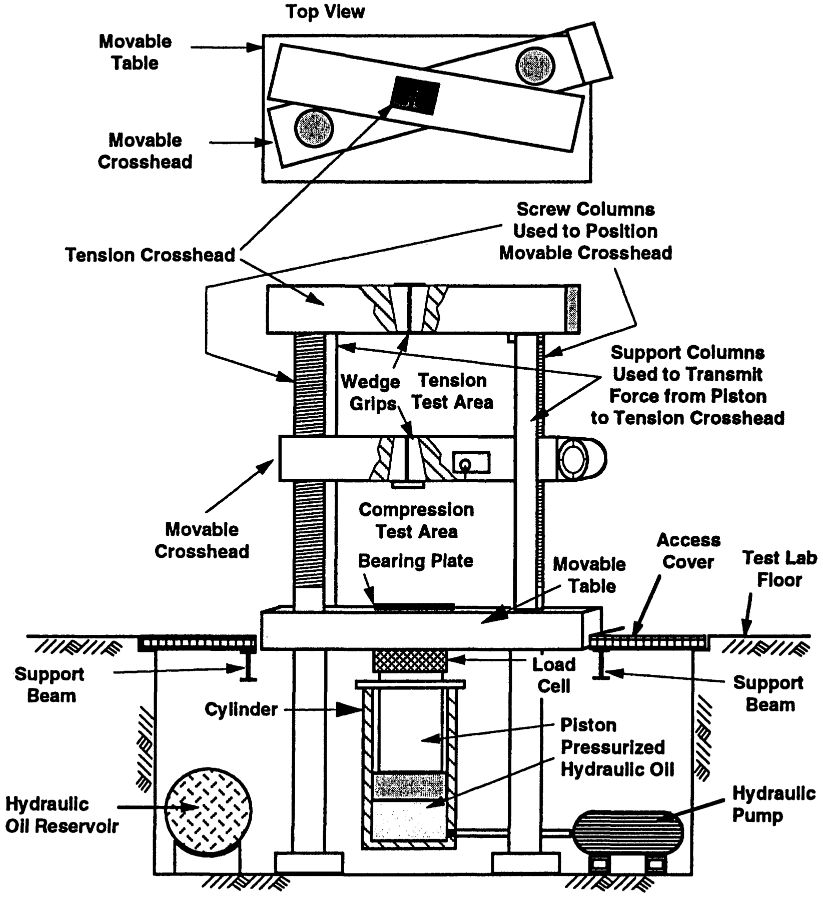
Top View
Movable Table
Movable Crosshead
Tension Crosshead
Screw Columns Used to Position Movable Crosshead
Support Columns Used to Transmit Force from Piston to Tension Crosshead
Wedge Grips
Tension Test Area
Movable Crosshead
Compression Test Area
Bearing Plate
Movable Table
Access Cover
Test Lab Floor
Support Beam
Load Cell
Support Beam
Cylinder
Piston
Pressurized Hydraulic Oil
Hydraulic Oil Reservoir
Hydraulic Pump

FIG. 3. Hydraulic testing system.

as well. Feedback controls are also available so that the load, stroke, or strain on a test item or specimen can be measured and used as part of the machine control. The crosshead can be moved manually, by using the hydraulic actuator and positioning blocks, or by hydraulic means. The hydraulic pump (not shown) may be built into the testing system or located at some other place for easier servicing, noise reduction, or space limitations.

These testing systems are equipped with load cells, displacement transducers, and feedback systems so that interactive control of load, stroke, and strain can be obtained. Strain control is obtained through extensometers or strain gages attached to test items. Some machines have the hydraulic actuator positioned in the crosshead, while others have the actuator in the base of the machine as shown in Fig. 4. These systems are capable of closed-loop operations because of the interactive controls, servo valves, and transducers. They can also be controlled by use of function generators or computer software interfaced with the hydraulics. A brief comparative summary of testing system capabilities (dead-weight, screw-driven, hydraulic, and servo-hydraulic) is given in Table 1. A servo-controlled actuator can also be attached directly to a structural frame or test bed as described in the following section.

2.5 Structural Frames and Test Beds

Structural test frames and beds accommodate large test items, test hardware having multiaxis loading requirements, large off-axis loading requirements, and other unique test requirements. A test frame or bed may

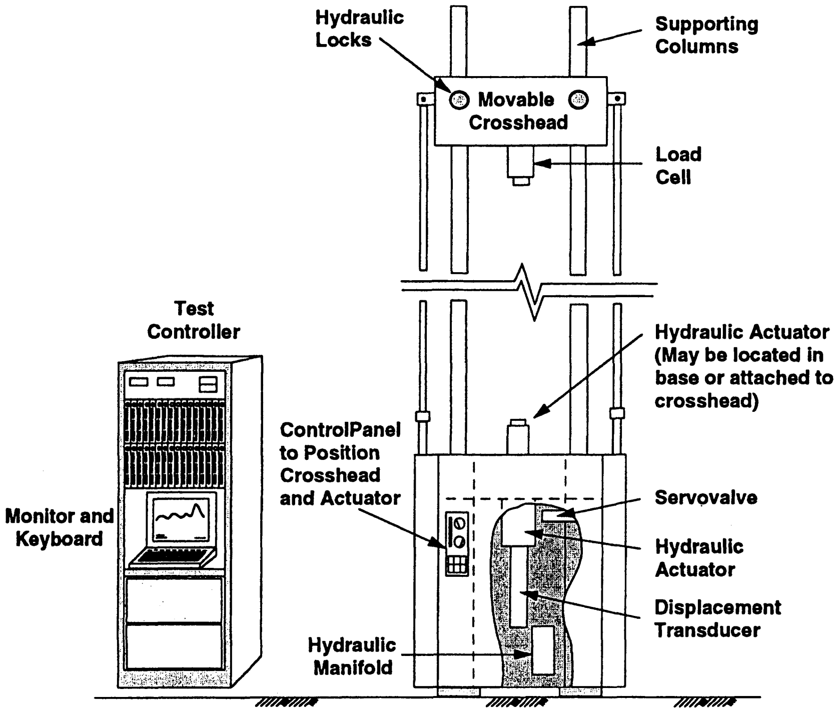
Hydraulic Locks
Supporting Columns
Movable Crosshead
Load Cell
Test Controller
Hydraulic Actuator (May be located in base or attached to crosshead)
ControlPanel to Position Crosshead and Actuator
Monitor and Keyboard
Servovalve
Hydraulic Actuator
Displacement Transducer
Hydraulic Manifold

FIG. 4. Servo-hydraulic testing system.

be small (e.g., milling-machine-table size) or large (e.g., aircraft-hangar size). Test items and hardware are attached to the bed or frame along with the loading systems. These beds or frames usually have a stiffened back or force reaction system. This force reaction system may be permanently attached or portable. Other beams, columns, and reaction systems are usually attached to the frame or bed. The large test frames or beds are serviced by overhead cranes, forklifts, and other related equipment for lifting and positioning test items, equipment, and loading systems.

Two examples of test frames and beds are shown in Fig. 5 as reported in Reese and Kawahara (1993). The first is a stiffened floor system (e.g., plates welded to beams) supported by four columns. The two rear columns also support a stiffened back. This back and floor system are the main force reaction systems. The only movable parts of this frame are the side rails and any portable beams, reaction systems, and loading equipment. The floor, beams, stiffened back, and columns have through or threaded holes on a uniform spacing to permit bolted attach-

Table 1. Comparison of testing system capabilities.

	Type of testing system			
Capability	Dead weight	Screw-driven	Tension or compression	Tension and compression
Maximum force, 10^3 lb	50	200	10 000	10 000
Maximum stroke, in.	5	Machine size minus crosshead	16–20	16–20
Strain rate	Variable	Low	Low	High
Load drift	None	±0.1%	±0.5%	±0.2%
Force measurement	Individual weights	Load cell	Bourdon tube or load cell	Load cell

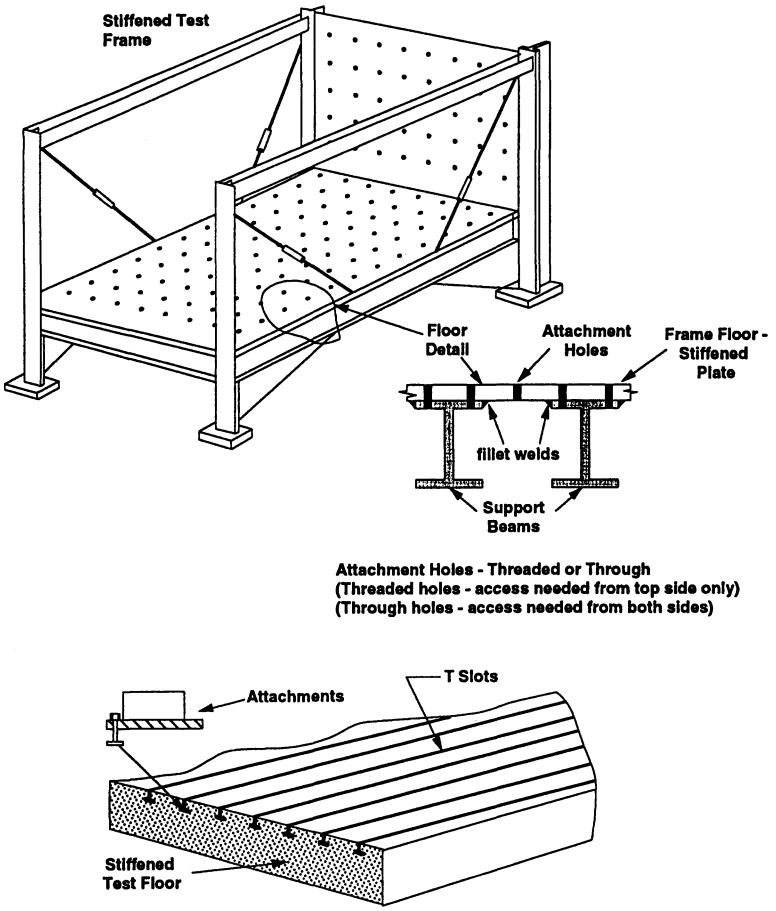
Stiffened Test Frame
Floor Detail
Attachment Holes
Frame Floor - Stiffened Plate
fillet welds
Support Beams
Attachment Holes - Threaded or Through
(Threaded holes - access needed from top side only)
(Through holes - access needed from both sides)
T Slots
Attachments
Stiffened Test Floor

FIG. 5. Test frames and test beds.

ment of test items and hardware. This type of frame has the appearance of a giant erector set and is used many different ways to accommodate test items and equipment. It is not uncommon to have this type of frame modified in many ways to satisfy specific test purposes. Many of these frames have had extra plates, stiffeners, and beams welded to them, holes have been drilled for attachment of loading equipment and test items, and other modifications have been made for test purposes.

The second example consists of a stiffened floor system. This floor can have through holes, threaded inserts, slots, rails, or tracks embedded in it for attachment of test items, equipment, and loading systems. These floor systems are typically large and are used to test aircraft, bridge beams, and sections of buildings. Testing machines and systems have also been integrated into these floor systems. An example of an aircraft wing subjected to vertical forces is shown in Fig. 6 with the whiffle trees supplying the distributed pressures simulating aerodynamic forces.

2.6 Time-Dependent Tests and Equipment (Creep and Fatigue)

2.6.1 Creep Nearly all materials and structures made of them exhibit time-dependent behavior to some extent, depending upon the test temperature and environment, load magnitude, and load duration. Prestressed concrete beams and floor systems experience sagging under their own weight. The total strain at any point consists of recoverable and permanent parts. Each part

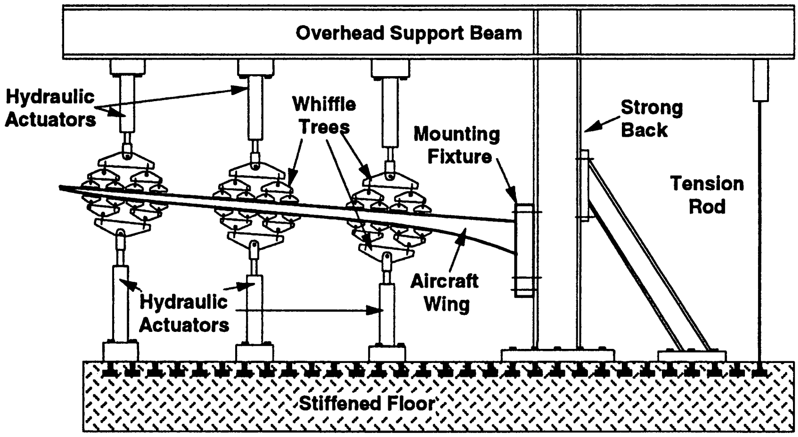

FIG. 6. Aircraft wing subjected to load tests on stiffened floor; vertical forces applied through whiffle trees.

may have time-independent and time-dependent components, again depending upon the test conditions.

For recoverable strain, the elastic component is relatively instantaneously recovered during unloading; its magnitude is usually proportional to the load. In stiffness testing, the load magnitude and duration are limited so that this elastic component dominates the total strain. For polymeric materials (and metals to a lesser extent) the elastic strain is often accompanied by a time-dependent anelastic strain component observed as an attempt by the structure to resume its original undeformed configuration for some time after unloading (i.e., residual strain at zero load that may decrease with time).

If the loading magnitude or duration (at a given test temperature) is sufficient, some of the total strain will not be recoverable even long after unloading; this is the permanent component of strain. The permanent strain can have a relatively time-independent component, the plastic strain, like that generated while yielding a steel tensile specimen at a constant velocity at room temperature. Or, it can be time dependent such as creep strain accumulating in the same specimen under contact stress at elevated temperature.

So, the total strain consists of time-independent and time-dependent recoverable and permanent parts. The proportions of these parts often change during a test. During a constant-stress tensile creep test on a metal, the elastic strain remains about constant while the creep strain increases with time. In a stress-relaxation test, a specimen is quickly strained a fixed amount; as the strain is held constant the load is observed to decrease. This is analogous to the preload decay in a nylon or soft lead washer after the bolt head is torqued. The decreasing preload means that the elastic strain is decreasing with time. Since the total strain is fixed, this elastic strain decrease is being balanced by an increase in anelastic and creep strain. For the nylon washer, it may be more anelastic than creep and vice versa for the lead washer.

The following identifies some failure modes that one should consider when creep behavior is anticipated. One should particularly be concerned when tests are requested at temperatures near or above one-half of the melting temperature (in kelvins) of any critical part under stress, especially for long loading durations, as reported by Bernasconi and Piatti (1979).

Excessive creep deformation leading to interference of parts is a failure mode. In turbine engines, rotor blades creep under radial centrifugal loads at elevated temperatures; the blade tips may engrave into the engine housing. As regards the grain-boundary sliding contribution to creep, one counter is to grow single-crystal blades; another is to form the blades superplastically (with very small grain size), then follow with a post heat treatment to dissolve most of the grain boundaries and recover the creep resistance inherent in the alloy.

Temperature cycling can cause excessive

deformation to accumulate as a result of incremental deformation. The source of this "ratcheting" can be time-independent plastic strains and/or time-dependent creep strains. For the latter, the phenomenon of "enhanced creep" may arise, wherein the creep strain accumulated during cycling exceeds that which would be calculated from continuous operation at the higher temperature boundary.

Stress relaxation in threaded connections can result in separation of parts and fatigue failure or reduction or loss of initial tension (preload). In aluminum electrical wiring, the loss of preload can result in a poor connection, high local electrical resistance, heating, and sometimes even fires.

Creep rupture is a catastrophic failure mode. For example, boiler vessels and high-pressure piping walls can reduce in thickness (a corrosive synergism may accelerate the process). Since the pressure is constant, the wall stress increases with thinning, thus accelerating further creep, thinning, and finally rupture. Empirical relations based on uniaxial tensile creep-rupture data can be used to help avoid creep rupture, although appropriate generalizations from uniaxial to multiaxial stress states is still a research topic as described in Reese and Kawahara (1993).

Creep crack growth under load can occur as the material near the crack tip progressively ruptures through void growth under high triaxial stress at elevated temperatures.

For slender members like columns and shell structures under compression, creep-induced buckling can be a catastrophic failure mode. That is, collapse can occur on account of creep deformation that causes the structure to be increasingly imperfect. Consider an elastic column under compression, and imagine a plot of compressive load versus midspan deflection perpendicular to the longitudinal axis of the column. For a perfectly straight column, the load-deflection curve will trace the load axis up to the Euler buckling load. For columns with an initial curvature (imperfection), the load-deflection curve will start at zero load with a slight positive deflection. As the load increases, the load-deflection curve will rise, bend over to the right, and follow a state of zero slope where the load deflection values at this point define an unstable state corresponding to that initial imperfection. Similar points in the load-deflection plane exist for other values of initial imperfection. One can anticipate that the locus of all such points is a curve starting at the Euler buckling load and decreasing as a function of deflection. In creep-induced buckling, the structure is initially loaded to a point safely under this instability locus; however, as creep proceeds under constant load, the deflection can increase with time until the locus is reached.

Thermomechanical fatigue (TMF) is another creep-related failure mode. Consider, for example, a service cycle where strain and temperature are cycled between two levels with a hold time at the higher level, all the while being proportional to each other. If the ratio of strain range to temperature range is other than the thermal-expansion coefficient of the material, a stress will be generated when the thermal-expansion strains are exceeded. During the dwell time, some of the stress will relax. After sufficient repetitions, a hysteresis loop in the stress–strain diagram may stabilize; this behavior is analogous to low-cycle fatigue involving plastic deformation where the number of cycles to failure can be significantly fewer than for high-cycle fatigue involving only elastic behavior.

The evaluation of certain types of electrical components (e.g., glass or ceramic insulators) can involve accelerated aging tests. These parts are typically expected to withstand long periods of use with environmental effects (e.g., temperature variations, humidity, and corrosives) having the major degrading influence on how they behave. These materials can deteriorate because of the condition of their surfaces. Small surface cracks can develop that weaken them and allow penetration of materials whose effects further weaken them.

2.6.2 Fatigue Structures are often subjected to fluctuating or oscillatory loads. The behavior of these loaded structures can differ fundamentally from those loaded statically or even dynamically if fatigue considerations are present. Fatigue is a structural behavior that is characterized by four features:

1. reduced strength,
2. reduced ductility,
3. reduced load ranges (minimums and maximums), and
4. reduced number of load repetitions expected in the service life.

The effects of fatigue can occur in widely different materials (e.g., metals, plastics, rubber, and concrete). Each material has some inhomogeneities; these imperfections (e.g., surface cracks, inclusions, voids, and interfaces in composite materials) are the source of fatigue. When a structure has some stress concentrations, the presence of these imperfections gives opportunities for cracks to grow. When the stress concentrations are combined with repeated loads, the resulting effect is to give greater emphasis to these imperfections. Characterizing these imperfections is an important but tedious process. Evaluating a structure (known to have some flaws) in a well-formulated fatigue test will integrate the effects and give insight into how a prototype or actual structure will respond.

Fatigue loads on structures may be cyclic, as in rotating machinery, or random, as in flutter on aircraft wings as given in Bendat and Piersol (1986). These loads may alternate between tension and compression, may increase or decrease around a mean value, and may be uniformly or randomly distributed over a structure. Experience shows that loads alternating about a mean value but *always* in tension or compression produce a certain damage potential; loads alternating from zero to a peak value (in tension or compression) produce a greater damage potential; and loads alternating from tension to compression produce even a greater damage potential. A common way of representing these stresses is through the stress ratio $R = \sigma_{\min}/\sigma_{\max}$; $R = -1$ for equally alternating tensile and compressive stresses, $-1 < R < 0$ for alternating tensile and compressive stresses, and $0 > R > 1$ for alternating tensile stresses. For alternating compressive stresses, R depends on the material (e.g., brittle or ductile) as the presence of any tension field governs the fatigue behavior as reported in the ASM Metals Handbook (1985).

The classic approach in fatigue studies is to generate a series of S–N curves for a specimen or portion of a structure exposed to cyclic loading. The ordinate S is for tensile stress and the abscissa N is for the number of cycles. Each point on the curve indicates the failure of a specimen at the number of cycles corresponding to the constant stress amplitude. These curves (for metals) typically exhibit a decreasing value for the failure condition until a horizontal asymptote is reached with the implication that the fatigue loads corresponding to this level of stress can be endured indefinitely.

Three areas of research have increased our understanding of fatigue loads and their impact on structural testing:

1. the development of relationships between plastic strain in materials and fatigue life,
2. the development of fracture mechanics to enable competent evaluations and assessments of structural designs, and
3. testing capabilities combined with statistically significant test methods, giving better information and data for design purposes.

It is now possible to assess the fatigue life of a structure or component on the basis of crack-growth data and strain-life data, along with the stress–strain relationships for materials. Areas or portions of structures that are "hot spots" known to be subjected to large fluctuating stresses are often isolated and tested separately. Aircraft test laboratories often use many testing machines devoted to fatigue testing of component sections. These include uniaxial (tensile specimens), biaxial (specially designed items, tension–torsion specimens), and triaxial tests.

With the continued development and use of high-strength materials, some composites, and minimum-weight structures, fatigue conditions are more prevalent than ever. Fatigue testing can be essential in the development of any structure that will experience fluctuating loads including structures designed for land transportation systems, aircraft, and railroad use.

Fatigue and fracture mechanics are often linked together. Fatigue may not result in fracture (although degradation is usually measured in this manner). Fracture can occur without fatigue. They are found together in fatigue testing because specimens and components usually fail by fracturing. Occasionally other failure modes are found by fatigue tests, but they are infrequent and do not result in a worst-case or even typical failure condition.

2.7 Fracture-Mechanics Tests and Specimens

Fracture is one of the ways failures occur in materials and mechanical assemblies.

Fracture rarely happens in one cycle of an overload. Fracture most often occurs because of repeated cycles of loading or because stressed materials are subjected to degrading environments (e.g., corrosion). Three basic developed approaches address fracture in materials: linear elastic fracture mechanics (LEFM), elastic-plastic fracture mechanics (EPFM), and fully plastic analysis. Each approach is an approximation of the actions within the material when a crack occurs. Linear theory is the most widely used as it is analogous to relating the stress state in a material or assembly to the tensile properties of the material. When a crack forms and propagates there is some plastic deformation that occurs in and around the crack tip.

The focus in fracture mechanics is on understanding how the generation and propagation of flaws in materials occurs. Three basic regimes of material behavior are relevant to understanding fracture mechanics:

1. at or below the threshold when cracks occur,
2. conditions in which cracks can develop, and
3. conditions in which cracks propagate.

Most materials have some types of microcracks that occur because of inclusions, surface roughness, or heat treatment. With the appropriate conditions of applied stress, microstructure, and environmental influences, microcracks begin to form into a crack. Three developed modes describe the propagation of a crack (opening, shearing, or tearing). The majority of cracks result from actions that cause opening, similar to driving a wedge into the material. The goal of fracture control is to ensure that strength of the material or assembly does not drop below a level safe for use. Cracks must be prevented from growing to a size that allows this to happen. The process of analysis and testing to determine the potential action of cracks and the degradation of an actual material or assembly is called damage tolerance design as described in Broek (1989). Among the items considered are the locations where cracks could develop, analysis and tests to determine the crack growth time, and the strength as a function of the crack size. Alternatively, periodic inspections are made of critical parts in which cracks could develop and reduce the strength of the material or assembly below the level safe for use. The focuses of the analysis and design efforts are on the effects of cracks on strength and the time required for crack growth.

Linear elastic fracture mechanics (LEFM) is based on the release of strain energy stored in the material. Crack propagation is predicted on the basis of the relationship of a material's plain strain fracture toughness (K_{Ic}) and the elastic stress in the vicinity of the crack in the material or assembly. This LEFM approach is often referred to as the defect-tolerant approach. In this approach, a flaw size is assumed to be present or is observed. The number of load cycles required to cause the component or specimen to fail is calculated. Tests for this fatigue-based fracture mechanics focus on two aspects:

1. determining the threshold below which crack growth is not observed (regime 1) and
2. determining the fatigue crack growth rate (ratio of growth to number of cycles) and the alternating (cyclic) stress intensity.

The material's fracture toughness is directly related to the threshold at which crack growth is not observed. The growth rate indicates the number of cycles the material can withstand before its strength drops below a safety level, as given in Rolfe and Barsom (1977) and Broek (1989).

The elastic-plastic fracture-mechanics approach extends the methods developed for linear elastic predictive capabilities into nonlinear elastic behavior. This approach uses refinements in predictions beyond the linear elastic range to measure a material's fracture toughness and tearing resistance, as cited in Rolfe and Barsom (1979).

Tests for a material's fracture toughness can be performed on testing systems capable of applying cyclic forces in a closed-loop configuration through either computers or function generators. Four standard specimens and test procedures have been developed and are defined in ASTM Standard E399. The elastic-plastic fracture-mechanics analysis method uses the crack opening displacement (COD) and the *J* integral. These two methods used in EPFM are defined in ASTM Standard E813 along with the test methods. Both E399 and E813 use bending and C-shaped precracked specimens.

2.8 Pressure Systems and Vessels

Pressures for test purposes are applied through two primary media: liquids and gases. These pressures are in turn transmitted directly or indirectly to a test item. Pressure systems consist primarily of pumps, regulators, intensifiers, accumulators, controls, and safety devices. Repeated cycles of a pump will achieve very high pressures (e.g., 1 GPa). An example pressure system is shown in Fig. 7. Each cycle of a pump produces a small step-function increase in pressure depending on the volume of the pump. Regulators, intensifiers, and accumulators can be installed on the outlet side of the pump. Regulators smooth out the steps in pressure increases. Accumulators store pressurized fluids or gases. An intensifier obtains higher pressures than the pump can deliver.

Control of pressure systems can be through manual valves or through air or electrically operated servo valves or through solenoid-controlled valves. Valve control is also possible with computers and function generators. It is possible for the output of a hydraulic testing machine (the pressure in the hydraulic system) to be the input (low-pressure side) of a pump so that higher pressures could be obtained. Pressure systems and vessels are described in detail in Reese and Kawahara (1993).

The pressure vessels are used in two basic ways:

1. to provide a pressure source for processes, operating and testing systems, and other applications requiring an uninterrupted pressure supply and
2. to contain the pressures applied to test items.

Examples of commercially available pressure sources that have pressure vessels include air compressors and systems using U.S. Department of Transportation (DOT) gas bottles. There are many different types of pressure vessels: two are shown in Fig. 8. On the left side is a vessel with a bolted closure with two ports (not shown) in the head of the vessel. The vessel is sealed by the use of two O-rings. The removable closure permits access to the vessel so test items can be placed inside. The ports are through holes in the closure. The ports are needed for the vessel inlet pressure supply, venting air while filling the vessel, a sealed pass-through for instrumentation, and removal of pressurized fluids or gases on conclusion of the test.

Figure 8 also shows a vessel with an elliptical removable closure (head). This closure is lowered onto a set of O-rings. The closure is tightened by pulling the tapered clamping rings around the vessel circumference using the ratcheting turnbuckles. This clamping action seals the vessel along the O-ring surfaces. This vessel also has ports (not shown) for inlet pressure, bleeding air if a liquid is used for the pressure medium, and pressure measurements. These ports may be in the vessel or closure or both. Both vessels can be equipped with additional sealed ports and electrical connectors that contain the pressure but allow passage of electrical wiring so that instrumented pressure tests can be performed. These sealed ports and connectors permit test items subjected to external pressure to have strain gages installed with the wires leading to the data-acquisition system. Test items are typically suspended from the closures while they are tested. These vessels are often placed below ground or are installed in pits to permit easier personnel access and reduced crane height needed to lower test items into the vessels.

The design and certification of pressure vessels for manned operations come under the guidance of the applicable sections (III and VIII) of the ASME pressure-vessel codes or equivalent requirements. Many commercial suppliers offer fittings, tubing, valves, manifolds, and other equipment needed to build pressure systems. The equipment supplied comes in various operating pressure

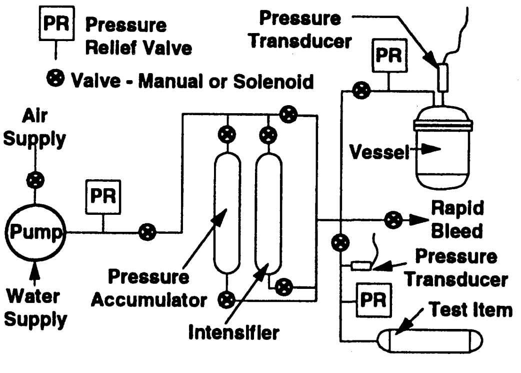

FIG. 7. Pressure-system layout.

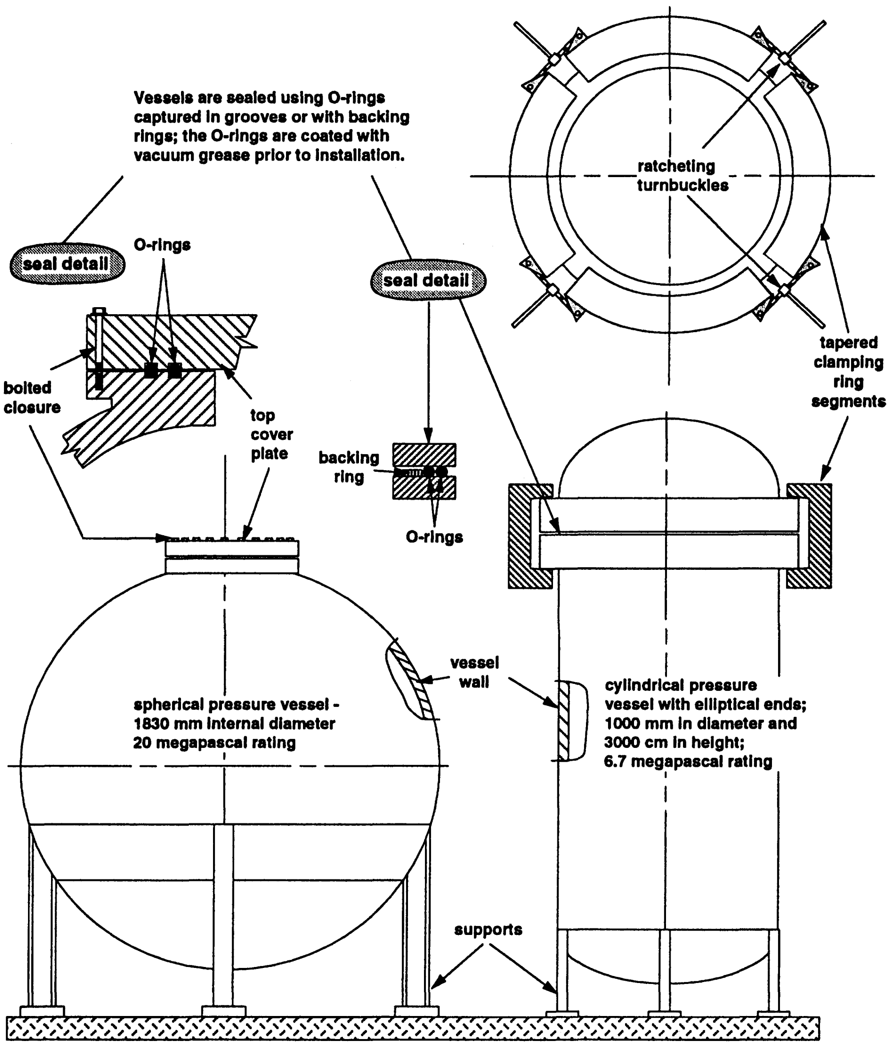
Vessels are sealed using O-rings captured in grooves or with backing rings; the O-rings are coated with vacuum grease prior to installation.
seal detail
O-rings
bolted closure
top cover plate
seal detail
backing ring
O-rings
ratcheting turnbuckles
tapered clamping ring segments
vessel wall
spherical pressure vessel - 1830 mm internal diameter 20 megapascal rating
cylindrical pressure vessel with elliptical ends; 1000 mm in diameter and 3000 cm in height; 6.7 megapascal rating
supports

FIG. 8. Pressure vessels: spherical and cylindrical with elliptical ends.

ranges, including the very high-pressure ranges in excess of 670 MPa.

In terms of containment, there are two concerns: pressure, and debris from tests in which failure occurred. Normally, a vessel used to contain pressure would not be used as a containment system for failure tests, particularly if debris from a failed test item is expected. Debris could impact the wall of the vessel and initiate cracks and other degradation of the inside surfaces of the vessel. For planned burst tests and proof tests that must be performed remotely, a containment system is needed to protect personnel from possible debris. Some examples of containment systems include:

1. bunkers and pits,
2. blast-proof test cells,
3. underwater operations, and, when all else fails,
4. sufficient distance.

Sufficient distance provides separation of personnel from possible debris. Bunkers, pits, and test cells contain the debris and released pressure by thick walls, soil embankments, and heavy access doors. Underwater operations involve lowering test items into the water and using the water to absorb the energies associated with ejected parts and released pressure.

Certification of pressure systems and vessels usually consists of a pressure test at 1.5 times the design working pressure. Most vessels so designed have much larger margins (typically 4) so that the proof test is very unlikely to cause initiation or propagation of material flaws reducing the pressure capability of the vessel. The major reason for the added safety requirements for pressure vessels is that the stored energies in the pressurizing medium, vessels, and pressure system are typically much greater than in other types of structures, and the potential for damage if ruptured is much greater. Failures of pressure vessels can be dramatic, involving large fragments that can be thrown great distances because of the energy contained. Further, there have been enough failures involving vessels that they require greater consideration in design and certification than other types of structures. Additional procedures are required for systems designed for use above 20 MPa. These certification procedures include detailed structural analyses, material tests results and certifications, certified fabrication procedures with special emphasis on machining, forming, and welding, and testing.

Pressure systems are protected using pressure-relieving devices. A variety of devices are used:

1. burst disks designed to fail at specific pressures,
2. spring-loaded valves that allow passage of fluid or gas when the force induced by the pressure overcomes the preload in the spring, and
3. other types of mechanically operated systems designed to open at or above specific pressures.

These devices should be set, installed, used, and periodically inspected to ensure safe operations involving pressure systems and vessels as described in ASME/ANSI (1988).

The nemesis of pressure systems and vessels is the presence of leaks. Leaks can occur in a myriad of places—welds, seals, connections, fittings, etc. Determining the leak location can be as simple as visually locating the hole or fitting or can require sophisticated techniques (e.g., helium leak detection). A leak can be observed by a drop in pressure during attempts to maintain it constant. Leak rates can be determined by monitoring the drop in pressure over time or by more sophisticated methods involving accurate flow measurements.

Before performing pressure tests, the following assessment should occur.

1. Determine the vessel and/or system to be tested and the test objectives.
2. Perform energy calculations as described in Reese and Kawahara (1993). These energies indicate the levels of protection and the type of facility needed in order to perform the tests safely (e.g., barriers, remote operations, and capability of containment systems).
3. Evaluate options—test using liquids whenever possible to reduce the energy in the pressurizing medium and facilities.
4. For units externally pressurized, provide some type of post-test pressure relief device (valve or removable plug) to vent any pressure accumulated during testing.
5. Check for leaks after the pressure level has been reduced if the test item cannot sustain pressure.
6. Provide other safety measures—personnel access denial, restraints on pressure lines in case of failure of test item, fittings, and proper disposal of test gases and fluids.

2.9 Thermal Systems

Thermal systems for structural tests involve three types of systems: temperature-controlled ovens for steady-state or programmable (variable) temperature conditions, radiant-heat systems for rapid temperature rises and temperature gradients, and burn facilities to create maximum temperatures and durations of fires. The ovens rely primarily on convection, conduction, and radiation as the means of heat transfer to the test item. The radiant-heat and burn facilities rely primarily on radiation as the means of heat transfer.

Temperature-controlled ovens typically operate between −75 and 535 °C (−100 and 1000 °F). They can control temperatures within a few degrees for nearly steady-state conditions. Ovens can also be programmed so that temperature cycles can be applied to test items. These ovens are typically heated with electrical resistance heaters, heat lamps, etc. Ovens can be used by themselves, installed in testing machines surrounding a test item, or placed on test frames and beds. In addition, most ovens can have other conditions introduced such as humidity, salt fog, or other environmental or climatic conditions, as described in Herzfeld (1962) and Reese and Kawahara (1993).

Radiant-heat facilities typically use some type of quartz lamps, which heat up rapidly when electrically activated. These facilities are capable of heat fluxes as large as 1380 kJ/m·s (400 Btu/ft·s) with rise times of 2 or 3 s. The amount of heat needed for a test depends on the test item, on its size and materials of construction including emissivities, and on the design or service requirement that is to be duplicated.

A burn facility involves the combustion of flammable materials, typically hydrocarbons. Various types of burn facilities exist including a chimneylike facility, an open pool, an enclosed pool, and an air-curtain burner. These facilities can have the supplies of air and fuel mixed so that there is an ideal burn (stoichiometric), fuel rich, or air rich. Rich means there is an excess supply of fuel or air in comparison with ideal burn conditions.

A chimney facility is constructed much like an ordinary fireplace with the test item positioned above the point where the fuel and air are combined and ignited. A pool facility consists of a lake or pond of fuel with the test item positioned near or above the surface of the fuel. An enclosed pool is similar except that a barrier surrounds the pool limiting or controlling the amount of air that can combine with the fuel vapor near the surface of the pool, depending on the position of the barrier. The temperatures in these burn facilities range from 760 to 1300 °C with a variation of about ±40 °C. The rate of heat production is limited to about 70–105 kJ/m·s (20–30 Btu/ft·s). The temperatures are determined with strategically placed thermocouples on the test item and in the fire.

2.10 Modal Analyses

The purpose of this type of equipment is to excite a structure to determine its mode shapes and frequencies. The loading conditions must be dynamic so that relative motions are created within the structure or component. Among the types of equipment used are instrumented hammers, vibration exciters, and others that result in forced or constrained displacements. An instrumented hammer consists of an ordinary hammer with a piezoelectric force transducer installed as shown in Fig. 9. The force transducer is calibrated so that the impact force–time history can be measured. The hammer tip is interchangeable so that different frequencies can be excited. With accelerometers and strain gages strategically placed on the test item, the excitation caused by the hammer blow will cause the structure or component to respond dynamically as measured by the transducers (typically accelerometers). Analyses of the data will yield the mode shapes and frequencies as given in Ewins (1985), Kobayashi (1987), and Reese and Kawahara (1993).

The structure can also be excited by an appropriately mounted vibration exciter. Similar instrumentation as with the hammer test would be used to measure structural response. The analysis of the data gives the mode shapes and frequencies. A forced or imposed displacement can also cause a structure to respond. For example, a cantilever beam could be preloaded into a deflected shape. If the preload is transmitted through an explosive bolt or some other type of restraint that can be released rapidly, the structure will respond dynamically at release until the vibrations damp out.

2.11 Inertial Systems (Centrifuges and Spinners)

Centrifuges create uniform acceleration conditions or gradients along a test item. Centrifuges are devices that have a rotating arm attached to a drive mechanism. The in-

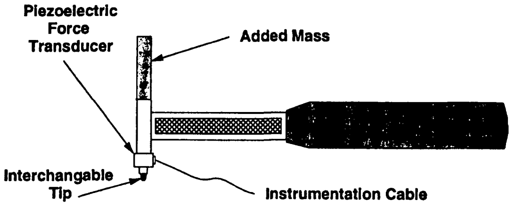

FIG. 9. Instrumented hammer used for modal analyses.

ertial forces are created by the combinations of the radial and tangential accelerations given by

$$a_r = \omega^2 r \tag{1}$$

and

$$a_t = r\alpha, \tag{2}$$

where a_r is the radial acceleration, a_t is the tangential acceleration, ω is the angular velocity, r is the radius to the point measured, and α is the rate of change of angular velocity. For nearly steady-state conditions, a_t is zero, and the inertial force is directly proportional to the mass of the test item and the radial acceleration.

Centrifuges come in a variety of sizes for testing structures and components ranging from a radius of nearly zero to a maximum of about 10 m. The accelerations vary from slightly over one g to more than $2000g$. An example centrifuge used for structural testing is shown in Fig. 10 as described in Reese and Kawahara (1993). This centrifuge has the drive motors located under subflooring in a pit. A containment structure is built around the perimeter of the centrifuge. A dynamic balance compensator is built into the two arms to allow for differences in the calculated balances required during operation. This compensator permits some vertical movement of the support arms while the centrifuge is rotating. The centrifuge also has slip rings so that data acquired during a test can be transmitted to data-acquisition and -recording systems. These slip rings also provide a means to operate test items (e.g., controls, actuators, and electrical systems). Larger centrifuges are usually serviced by an overhead crane.

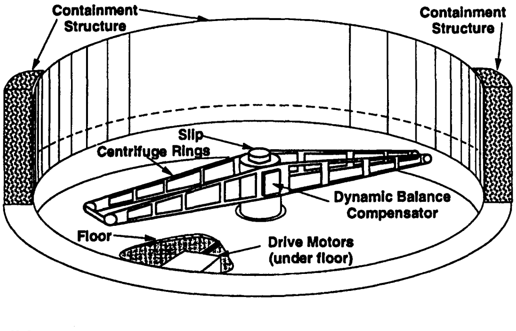

FIG. 10. Centrifuge.

Spinners are usually air-driven turbines placed in a housing. The housing is usually positioned in a sealed pit. The pit can be evacuated so that little or no air resistance is encountered while the spinner is in operation. Test items are attached to the turbine drive motor through fixtures and adapters. Spinners can operate at no-load speeds up to approximately 50 000 rpm. Spin testers are often used to evaluate the performance and operation of high-speed rotating assemblies and systems (e.g., artillery shells, and jet and turbine engine components—blades, compressors, and drive motors).

Among the assumptions made in using centrifuges and spin testers is that the test item will not fail. In order to rotate, centrifuges and spin testers must be dynamically balanced. If failure occurs in the test item, then the dynamic balancing of the arm allowing rotation of the spinner itself is upset when the structure deforms or changes position as it fails. When the dynamic balance of a centrifuge or spinner is upset, there is a danger of damaging bearings, containment systems, slip rings, counter systems, drive system, etc. A few centrifuges have been designed to tolerate ejection of or at least gross displacement of test items. However, this is not the usual case so that centrifuges usually need to be protected from displacement or gross deformation of test items. Among other features that can be incorporated in centrifuges and spinners are temperature chambers and vibration shakers.

2.12 Vibration Systems

Vibration equipment consists of three basic types of exciters: electromagnetic, hydraulic, and mechanical. Examples of each of these vibration test systems are given in Fig. 11. These exciters can produce steady-state, transient, and random vibrations depending on the controls used. Each type of equipment has advantages (e.g., larger frequency ranges) and some disadvantages (e.g., limited sizes for test items) as explained in this section. There are two ways vibration test equipment is used:

1. with items attached directly to the vibration table and shaker and

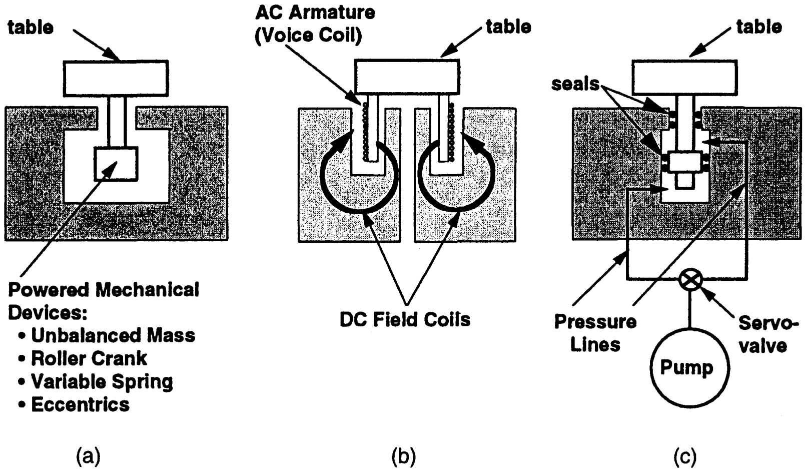

FIG. 11. Vibration testing systems: (a) mechanical, (b) electromagnetic, (c) hydraulic.

2. as an exciter attached to and capable of vibrating large structures such as buildings or bridges.

Test items are usually rigidly attached to a plate or mounting fixture that is connected directly to the shaker. The motion of the shaker is transmitted to the fixture and the attached test item. For test items much larger than the plates or mounting fixtures, portable exciters and driver systems (power sources) have been developed that can be installed at various locations on the structure or test item. Clever placement of these portable shakers can cause vibrations and resonances in the structure as reported in Reese and Kawahara (1993).

The mechanical exciters use cams or a variety of other devices (e.g., roller cranks, variable springs, eccentrics, and unbalanced masses) to induce small varying displacements to the table supporting the test item or directly to the test hardware. The hydraulic systems are similar to a testing machine with the motions of the support or base plate caused by moving the hydraulic fluid rapidly through servo valves as given in Harris (1988).

The electromagnetic exciters move a magnetic-field coil suspended in a driver coil. The driver coil is attached to the shaker support structure through flexures to minimize interactions of the field coil and the support structure. As the magnetic field changes, small displacements are induced in the field coil that cause vibrations in the attached test items. The motion of the field coil is very similar to the voice coil arrangement in a loudspeaker system. Vibrations can also be imposed on structures by acoustic systems, which can operate at 150 dB or more. Many flight structures have requirements to pass acoustic-induced vibrations simulating missile liftoff and engine noise.

Steady-state vibration consists of sinusoidally varying displacements of the plate or mounting fixture. Transient vibration is defined by a sinusoidal excitation multiplied by an exponentially decaying signal. Random vibration is defined by the superposition of many sinusoidal signals with random phase. Examples of these three basic types of acceleration–time histories are given in Fig. 12.

Vibration testing consists of four basic parts:

1. specification of the vibration environment

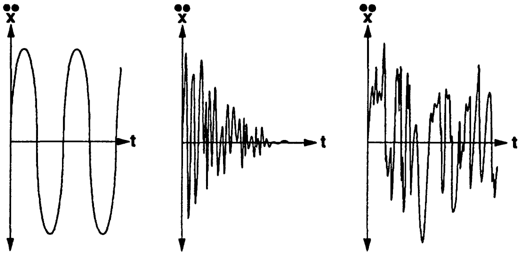

FIG. 12. Acceleration–time histories for vibration environments.

(sine, random, frequencies, magnitudes, and duration),
2. attachment of the test item or structure to the properly selected vibration machine through the use of fixtures and preloading techniques,
3. selection of the diagnostic transducers and their installation at critical and/or informative locations, and
4. performance of the test including the data acquisition, reduction, and analyses.

These four parts depend in varying degrees on the equipment. The input depends on a particular machine and/or laboratory capability. Attachments depend on the basic machine, modifications made to it, and proper fixtures. The diagnostic transducers chosen depend on the data-acquisition and control systems. Analysis of the data depends on the type of data and the computer systems and software available. Vibration-test equipment capabilities can be described in the following manner:

1. The frequency range capability of a shaker decreases with increasing design force capability.
2. Mechanical shakers can only be used for low-frequency tests (up to 100 Hz).
3. Hydraulic shakers can be used up to 2000 Hz.
4. The upper frequency ranges (well above 2000 Hz) can be reached only with electrodynamic shakers, which cover frequencies to 10 000 Hz.
5. The largest forces are delivered by hydraulic shaker systems.
6. Electromagnetic shakers are typically limited in displacement to 25 mm.
7. Hydraulic shakers have more displacement capability than electrodynamic shakers.
8. Mechanical shakers have the most displacement capabilities—about 150 mm.
9. Hydraulic and mechanical shakers perform sinusoidal vibration tests, but the precise control required for complex random or transient wave forms is difficult to achieve. Advances in electronic controls and servo valves will facilitate random-vibration tests on future hydraulic vibration systems.
10. Electromagnetic shakers can produce complex random motions in test items.

2.13 Shock Systems

Shocks can be induced in structures and test hardware by rapid accelerations or decelerations. Decelerations can be produced by dropping the structure or component to be tested onto a stationary target. The resulting velocity change causes a shock loading to the test item. Accelerations can also be produced by placing the structure or component in a type of gun. With a rapid creation of pressure from the burning of propellant or the sudden release of pressurized gases, the structure or component is subjected to a rapid increase in forces causing an acceleration. Shocks can also be produced by preloading a structure into a deflected position and then suddenly releasing the preload as reported in Reese and Kawahara (1993).

Shock testing machines are manufactured by many companies, and many laboratories have designed and built special and general use systems. They basically consist of an elevating table for mounting the test item. The table can move to a position above a contact surface. A shock loading is produced when the table is dropped along guide rails onto the contact surface. The shock pulse can be tailored according to the shock absorbing materials used on the contact surfaces, the velocity change produced, and the masses of the table and the test item. Greater velocities can be achieved if long elastic cords attached to the base of the machine are tensioned to provide a pulldown force to the table and test item.

An example of a shock test machine is shown in Fig. 13. The test item is located on the mounting plate. The lifting winch is used to raise the mounting plate to an elevated position. The mounting plate impacts onto a programmer of specially selected materials (e.g., felt pads and honeycomb) designed to supply a certain shock pulse (acceleration time history) to the component.

Gas guns, horizontal actuators, and other similar devices supply the sudden release of energy (typically pressure), which is captured behind a movable attachment system (e.g., piston) and test item. The release of the stored pressure causes rapid movement of the piston. The resulting acceleration can produce the desired shock depending on its rise time. The moving piston and test item can also be impacted into a stationary target,

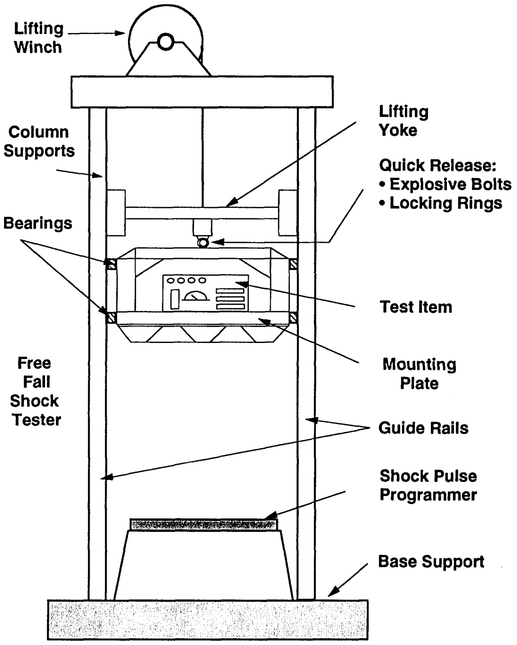
Lifting Winch
Column Supports
Bearings
Free Fall Shock Tester
Lifting Yoke
Quick Release:
• Explosive Bolts
• Locking Rings
Test Item
Mounting Plate
Guide Rails
Shock Pulse Programmer
Base Support

FIG. 13. Shock-test machine.

a movable mass, or other type of target depending on the type of shock pulse needed. An example of a gas gun is shown in Fig. 14. In this example the test item is attached directly to the piston. The shock pulse experienced by the test item is the same as that of the piston.

Shock tests are also performed using drop towers and cable suspension systems. A shock testing system used for large structures is shown in Fig. 15. This facility consists of a cable stretched between two hills or towers. An instrumented structure to be tested (shown in greatly exaggerated size) is raised toward the horizontal cables. The instrumentation cable travels with the test item. The test item can contain data-recording systems that can be recovered after the test, or it can telemeter data to a receiving station during the test. The left tower also shows a cantilevered section that has a target at its base.

The test item is typically allowed to fall freely either from the cables or from the drop tower. The disadvantage associated with free fall is that the test item may be acted on by lateral aerodynamic loads depending on the height of the tower and the prevailing wind conditions. Shock tests often employ diagnostic high-speed photometric coverage directed to a specific impact location, and these lateral loads can cause the test item to deviate from its course to the target. The test item can also be attached to two cables (not shown) that are located near the impact point. These director cables guide the test item to a predetermined location on the target in a similar manner to the guide columns in a shock machine. The impact velocity will be reduced slightly in this type of directed free fall because of frictional forces between the vertical cables and the attachment guides. Release systems (e.g., explosive bolts that fracture on detonation of a small directed explosive charge) allow the test item to fall freely a short distance prior to impact. The use of release systems prevents unwanted interactions among the guide cables, instrumentation lines, and the test item during impact, as reported in Reese and Kawahara (1993).

Impact velocities can be increased using

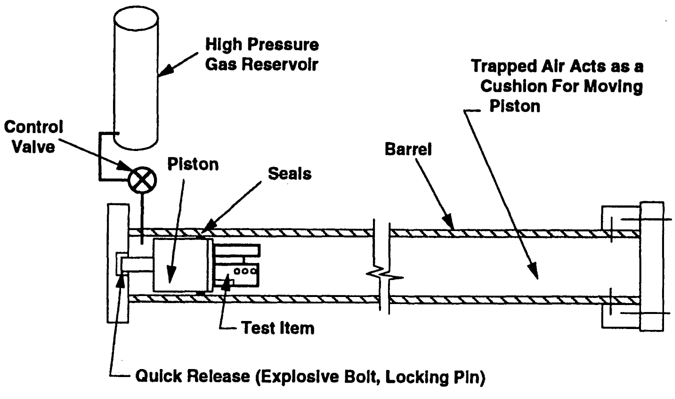
High Pressure Gas Reservoir
Trapped Air Acts as a Cushion For Moving Piston
Control Valve
Piston
Seals
Barrel
Test Item
Quick Release (Explosive Bolt, Locking Pin)

FIG. 14. Gas gun.

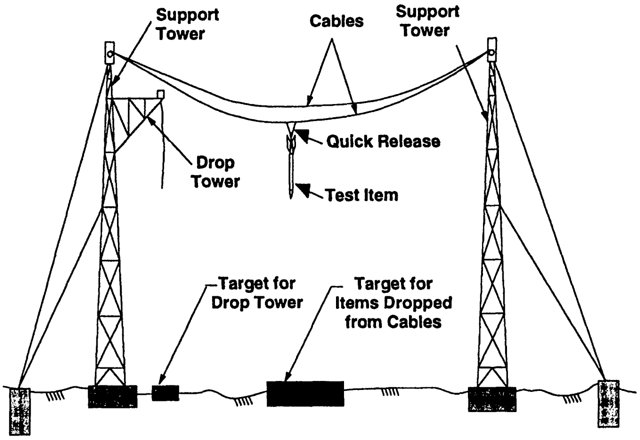

FIG. 15. Drop-tower/cable shock-test facility.

large elastic cords that will pull the test unit toward the target. The director cables also run through pulleys near the target. The ends of these director cables could be attached to a rocket sled. This sled would travel along a horizontal path away from the target area. The acceleration time history of the rocket sled could be designed to provide the tension force in the cable needed to pull the test item to the target at the additional desired velocity without cable whip or overload. The test item would also again be released (e.g., explosive cutters) from these guide cables to permit free fall into the target at or very near the desired location for camera coverage. The velocities achieved in these free-fall assisted tests are limited only by practical limits of height, elastic cord forces, or the accelerations delivered by a rocket sled.

Shocks can also be supplied by rocket-sled track systems—a sort of open-air gas gun. This type of loading system produces accelerations to test items attached to rocket sleds. The sleds are accelerated down a single- or dual-rail track by the ignition of rocket engines. Velocities of 1500 m/s can be achieved on tracks as short as 1500 m. The resulting shock pulses can be tailored by attaching the test item to the rocket sled and experiencing the resulting acceleration time history when the sled moves along the track. Shock can also be attained by impacting the sled into an object. Alternatively, the test item can be ejected from the sled after it has reached the desired velocity and then impacted into a target or slowed by deploying a parachute. Shock loads on a test item can alternatively be achieved by accelerating the target into the item. This approach of a moving target allows the test item to be hard wired, which usually aids instrumentation and data acquisition. This "reverse ballistic" technique was developed by Sandia National Laboratories, as described in Reese and Kawahara (1993). The purpose of these shock-generation systems is to impart enough energy or velocity change to the test item to expose it to the required or representative service, proof, or destructive conditions.

2.14 Hardness Tests and Equipment

The hardness of materials is largely a surface phenomenon as influenced by work hardening, yield strength, modulus of elasticity, and grain structures. The tests for hardness are comparative and have their basis in satisfying a need for some convenient measure to evaluate material resistance to scratching, wear, abrasion, and indentation. A hardness test can be used to grade material samples or lots and to provide evidence of quality control. Four hardness tests have been developed and continue to be used: Brinell, Rockwell, Vickers, and Shore scleroscope, as described in the ASTM Standards (current editions). The first three tests are made with indentation techniques while the Shore scleroscope uses the rebound height of a small hammer. The Knoop hardness test system has also been used, and relationships to these four enduring hardness tests are given in ASM Metals Handbook (1985) with primary applications to microhardness measurements. An ultrasonic hardness test is also available.

The indentation techniques use various types of indentors (e.g., balls and pyramids) that are placed in contact with the material surface. A force is applied, and a small permanent deformation occurs in the material surface. Because the indentations are small and usually do not degrade the material performance, these hardness tests including the Shore scleroscope are considered nondestructive. However, hardness measurements can be a detrimental influence on the material surface [e.g., hardness measurements made at the root (minor) diameter of a threaded fastener] where a discontinuity in the form of an indentation is made in a potentially high-stress area of a fastener. Sufficient use has been made of hardness tests so that the measurements made using the different approaches can be compared as given in the ASM Metals Handbook (1985) and the ASTM Standards (current editions). Table 2 summarizes some of the important features of hardness test equipment and practices, as given in Table 3 in the Hardness Testing chapter of the ASM Metals Handbook (1985).

A durometer is an instrument used to measure the hardness of rubber and plastics. Several types of durometers are available with only two (types A and D) conforming to ASTM Standard D2240 ("Indentation Hardness of Rubber and Plastic by Means of a Durometer"). A durometer operates similarly to a hardness tester in that the indentor is brought into surface contact, and then a force is applied and the reading taken using the built-in scale. A major source of discrepancies in durometer tests is the speed at which the durometer indentor is applied to the test item.

Hardness tests of all types need to be conducted on clean flat surfaces. The indentors need to be aligned at right angles to the surface. Several measurements will likely be needed to obtain a statistically significant hardness evaluation. Individual measurements need to be taken far enough from other measurements so there is no interference among them.

3. HAZARDS INVOLVED IN MECHANICAL TESTING

Mechanical and structural tests are often destructive. Protection of test personnel and testing equipment is important. Hazards that need to be considered are

1. the sudden releases of large forces and energies, shrapnel and other debris, and test fixtures that can occur in each destructive test;
2. the many pinch points and related hazards in working with test items and the fixtures, with particular caution required when performing disassembly and removal of damaged, destructively tested hardware;
3. the mechanical gears, electric motors, hydraulic lines, and lasers in proximity to test personnel;
4. the sudden release of high-pressure hydraulic fluids, liquids, or gases;
5. the lifting and rigging of heavy items into and out of the test spaces;
6. the congested spaces in and around the

Table 2. Comparisons of hardness test techniques and test equipment.

Comparison	Brinell	Rockwell	Vickers	Shore scleroscope
Indentor shape	Metallic ball (10 or 2.5 mm)	120° diamond cone 1.6 to 12 mm in diameter	Square-based pyramid with face angles of 136°	Diamond-tipped hammer
Measurements	Applied force/ diameter of indentation	Combination of specific penetrator and hardness scale (15 options)	Applied force and diagonal of indentation	Height of hammer on rebound
Applied force	100 kg$_f$ to 3000 kg$_f$	3 kg$_f$ to 150 kg$_f$	1–120 kg$_f$	2.4×10^{-3} kg from 250 mm
Applications	Large forgings and castings	Forgings to finished surfaces of parts	Fine finished surfaces, thin specimens	Fine finished surfaces, thin specimens

equipment with their associated trip hazards; and
7. the noise associated with equipment operation and test-item failure.

GLOSSARY

Bourdon Tube: A hollow tube that, when pressurized, deforms with repeatable accuracy and, when calibrated, accurately converts the deformed shape into measurements of applied pressures.

Fixtures: The mechanical parts used for attachment of items to mechanical test equipment.

Hardness: Resistance to indentation.

High-Cycle Fatigue: Fatigue involving only elastic behavior of the material.

Loads: The broad category of applied forces, torques, pressures, thermal effects, energy deposition, shock, vibrations, and other excitations that are simulated in mechanical or structural testing.

Low-Cycle Fatigue: A reduced number of cycles to failure because plastic deformation is present (opposed to **High-Cycle Fatigue**).

Photometrics: Descriptive photographic or video records of tests, designed to capture motion and displacement of test items.

Pressure Intensifier: A mechanical system used to increase pressure.

Pressure Regulator: A mechanical system used to control the flow of pressurized liquids or gases.

Servo Hydraulic: The operation and control of a testing machine through the use of electronically controlled servo valves acting on hydraulic cylinders capable of producing motion in two directions.

Slip Ring: Rotating shaft or disk containing a continuous electrical conductor attached to its surface.

Structural Testing: The work devoted to the understanding of how structural systems respond to the conditions imposed upon them.

Whiffle Trees: Loading fixtures used to simulate aerodynamic or distributed forces; the tree is a small beam loaded at each end and at the middle. These trees are typically used in series.

Works Cited

ASM (Eds.) (1985), *Metals Handbook,* Vol. 8, *Mechanical Testing,* Metals Park, OH: American Society for Metals.

ASME (1988), *Safety and Relief Valve, Test Performance Codes,* ASME/ANSI Standard No. PTC 25.3-1988, New York: ASME.

Beckwith, T. G., Buck, N. L. (1973), *Mechanical Measurements,* Reading, MA: Addison-Wesley.

Bendat, J. S., Piersol, A. G. (1986), *Random Data, Analysis and Measurement Procedures,* 2nd ed., New York: Wiley.

Bernasconi, G., Piatti, G. (Eds.) (1979), *Creep of Engineering Materials and Structures,* London: Elsevier Applied Science Publishers.

Blake, Alexander (1985), *Handbook of Mechanics, Materials, and Structures,* New York: Wiley.

Broek, D. (1989), *The Practical Use of Fracture Mechanics,* Dordrecht: Kluwer Academic.

Davis, H. E., Troxell, G. E., Hauck, G. F. W. (1982), *The Testing of Engineering Materials,* 4th ed., New York: McGraw-Hill.

Doebelin, E. O. (1975), *Measurement Systems: Application and Design,* rev. ed., New York: McGraw-Hill.

Ewins, D. J. (1985), *Modal Testing: Theory and Practice,* New York: Wiley.

Hannah, R. L., Reed, S. (Eds.) (1992), *Strain Gage User's Handbook,* London: Elsevier.

Harris, C. M. (1988), *Shock and Vibration Handbook,* 3rd ed., New York: McGraw-Hill.

Herzfeld, C. M. (Ed.) (1962), *Temperature, Its Measurement and Control in Science and Industry,* Vols. 1–3, New York: Reinhold.

Kobayashi, A. S. (1987), *Handbook on Experimental Mechanics,* Englewood Cliffs, NJ: Prentice-Hall.

Reese, R. T., Kawahara, W. A. (1993), *Handbook on Structural Testing,* Lilburn, GA: Fairmont Press.

Rolfe, S. T., Barsom, J. M. (1977), *Fracture and Fatigue Control in Structures,* Englewood Cliffs, NJ: Prentice-Hall.

Tse, F. S., Morse, I. E. (1989), *Measurement and Instrumentation Engineering,* New York: Marcel Dekker.

U.S. Forest Service (Eds.) (1987), *Wood Handbook,* Washington, DC: U.S. Department of Agriculture, Forest Service.

Van Vlack, N. L. (1974), *Elements of Materials Science,* Reading, MA: Addison-Wesley.

Wright, C. P. (1994), *Applied Measurement Engineering,* Englewood Cliffs, NJ: Prentice-Hall.

Further Reading

Abelkis, P. R., Hudson, C. M. (1982), *Design of Fatigue and Fracture Resistant Structures,* ASTM STP 761, Philadelphia: American Society for Testing and Materials.

Amzallig, C., Leis, B. N., Rabbe, P. (1982), *Low-Cycle Fatigue and Life Prediction,* ASTM STP 770, Philadelphia: American Society for Testing and Materials.

Boyer, H. E. (Ed.) (1987), *Atlas of Stress-Strain Curves,* Metals Park, OH: American Society for Metals.

Dummer, G. W. A., Griffin, N. B. (1962), *Environmental Testing Techniques for Electronics and Materials,* New York: Pergamon/Macmillan.

Garafalo, F. (1965), *Creep and Creep Rupture of Metals,* New York: Macmillan.

Metals and Ceramics Information Center (1983), *Damage Tolerant Design Handbook,* Columbus, OH: Batelle Columbus Laboratories.

Newland, D. E. (1975), *An Introduction to Random Vibrations and Spectral Analysis,* New York: Longman.

Pointer, A. R. S., Hayhurst, D. R. (Eds.) (1981), *Creep in Structures, Proceedings of the 3rd Symposium, Leicester, U.K., September 8–12, 1980,* Berlin: Springer.

Richards, C. W. (1961), *Engineering Materials Science,* San Francisco: Wadsworth.

Sabnis, G. M. *et al.* (1983), *Structural Modeling and Experimental Techniques,* Englewood Cliffs, NJ: Prentice-Hall.

Throop, J. F., Reemsnyder, H. S. (1981), *Residual Stress Effects in Fatigue,* ASTM STP 776, Philadelphia: American Society for Testing and Materials.

U.S. Department of Defense (1987), "Metallic Materials and Elements for Aerospace Vehicle Structures," in *Military Standardization Handbook,* MIL-HNDBK-5D, Washington, DC: U.S. GPO.

U.S. Department of Defense (1986), *Aerospace Structural Metals Handbook,* AFML-TR-68-115, Columbus, OH: Batelle Columbus Laboratories.

TESTING, NONDESTRUCTIVE

IRVING KAUFMAN, *Arizona State University, Tempe, Arizona, U.S.A.*

	Introduction	26
1.	**Visual Inspection**	27
2.	**Liquid-Penetrant Inspection**	28
3.	**Magnetic-Particle Inspection or Leakage-Field Measurement**	28
3.1	Magnetic-Particle Inspection	29
3.2	Magnetic-Field Measurement	31
4.	**Radiographic NDT**	31
4.1	Principle of Operation	31
4.2	Three-Dimensional Information	32
4.3	Unsharpness, Contrast, Sensitivity	33
4.3.1	Unsharpness	33
4.3.2	Contrast	33
4.3.3	Sensitivity	34
4.4	Radiation Wavelengths or Photon Energies Required	34
4.5	Sources of Radiation	34
4.5.1	X-Ray Sources	34
4.5.2	Gamma-Ray Sources	35
4.6	Sensing and Recording	35
4.6.1	Radiographic Film	36
4.6.1.1	Description	36
4.6.1.2	Film Parameters	36
4.6.1.3	Viewing the Image	36
4.6.1.4	Measurement of Radiographic Sensitivity	37
4.6.2	Other Sensing and Recording Media	37
4.7	Real-Time Radiography	37
5.	**Ultrasonic NDT**	39
5.1	Wave Types	39
5.1.1	Bulk Waves	39
5.1.2	Structure-Dependent Waves	39
5.1.2.1	Rayleigh Waves	39
5.1.2.2	Love Waves	39
5.1.2.3	Lamb Waves	39
5.1.2.4	Stonely Waves	40
5.2	Reflection, Refraction, and Mode Conversion	40
5.2.1	Reflection	40
5.2.2	Refraction	40
5.2.3	Mode Conversion	40
5.3	Acoustic-Wave Generation and Detection	40
5.3.1	Transduction—Longitudinal and Shear Waves	40
5.3.2	Generation of Structure-Dependent Waves	41
5.3.3	Geometric Ultrasonic Optics	41
5.4	Methods of NDT	41
5.4.1	Through-Transmission NDT	41
5.4.2	Resonance NDT	41
5.4.3	The Pulse-Echo Technique	42
5.4.3.1	Generation of Short Pulses	42
5.4.3.2	Pulse-Echo Action	42
5.4.3.3	Resolution	42
5.4.3.4	Presentation of Results	43
5.4.4	Noncontact Coupling	44
5.4.4.1	Magnetostriction	44
5.4.4.2	Electromagnetic Coupling	44
5.4.4.3	Electrostatic Coupling	45
5.4.4.4	Laser-Generated Acoustic Waves	45
6.	**Eddy-Current Testing**	45
6.1	Frequency-Domain (Sinusoidal) Eddy-Current Testing	46
6.1.1	Principle of Operation	46
6.1.2	Depth of Penetration	48
6.1.2.1	Geometric Factor	48
6.1.2.2	Skin Effect	49
6.1.3	The Importance of Phase	49
6.1.4	Instruments	50
6.1.5	Multiple-Frequency Testing	50
6.2	Time-Domain (Pulsed) Eddy-Current Testing	51
6.2.1	Measurement of Resistivity	51
6.2.2	Pulsed Eddy-Current NDT	51
7.	**Acoustic-Emission Testing (AET)**	52
7.1	AET—An Active Type of NDT	52
7.1.1	Acoustic Emission	52
7.1.2	Development of AET	52
7.1.3	AET and Other NDT Techniques	53

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

7.1.4 Signals Detected 53
7.1.5 The Kaiser Effect 54
7.2 Measurement Techniques 54
7.2.1 Basic Measurement Technique 54
7.2.2 Continuous Emission 54
7.2.3 Burst Emission 54
7.2.4 Interpretation of Data 54
7.3 Applications 56

8. Other NDT Procedures 56
8.1 Strain Gauge 56
8.2 Photoelasticity 56
8.3 Brittle Coating 56
8.4 Replication 57
8.5 Volatile-Liquid Method of Surface Crack Detection 57
8.6 Leak Detection 57
8.6.1 Tracer-Gas Method 57
8.6.2 Sealed-Package/Holographic Method 57
8.7 Radioactive-Gas Penetrant 57
8.8 Acoustic-Impact Response 57
8.9 Sonic Testing of Composite Structures 57
8.9.1 The Sonic-Resonator Method 57
8.9.2 The Eddy Sonic Method 57
8.10 Thermal Techniques 57
8.10.1 Temperature Mapping during Operation of Test Piece 57
8.10.2 Surface Flaw Detection by Scanning with a Heat Pulse ... 58
8.11 Electric-Current Injection 58
8.12 Static Magnetic Field 58
8.13 Corona Discharge 58
8.14 Capacitance Measurement 58
8.15 Microwaves 58
8.16 YIG Resonator 58
8.17 Nuclear Magnetic Resonance Imaging (NMRI) 58
8.18 Barkhausen-Noise Analysis (BNA) 58
8.19 Radiometric Surface Crack Detection in Moving Surfaces 59
8.20 Speckle Metrology 59
8.21 Optical-Holography Interference Method 59
8.22 Optical-Holography Vibrational Analysis 59
8.23 Ultrasonic Holography 59
8.24 Acoustic Microscopy 59
8.25 Exoelectron Emission from Metals 59
8.26 Compton-Scattered Radiography 59
8.27 Energy-Loss Ion Radiography 60
8.28 Neutron Radiography 60
8.29 Positron Annihilation 60
8.30 Tomographic Imaging 60
8.31 Surface-Crack Detection by Optical Scanning 60

Glossary 60
Works Cited 63
Further Reading 65

INTRODUCTION

The measurement of blood pressure, the determination of the characteristics of a transistor, or the flight test of an airplane are all procedures that can be called nondestructive testing (NDT), nondestructive evaluation (NDE), or nondestructive inspection (NDI). While these terms are certainly appropriate to such medical and other test procedures, it has become the practice to use them only to refer to techniques that deal with the detection of defects, or flaws, in mechanical structures. Examples are cracks, inclusions, and cavities, or the undesirable change in microstructural parameters of a material. NDT refers to the general method of finding such flaws without damaging the structure that is being examined; NDI refers to the actual testing; and NDE refers to the inspection procedure, a description of what was found, and an evaluation of possible results. For simplicity, in the following text only NDT is used.

Although some costs are involved in carrying out NDT, routine NDT of newly manufactured products has generally been found to be a desirable practice, in terms of both cost and reputation of reliability. NDT is also very important for equipment that operates over an extended period of time, for in the parts of a unit that are under constant or repeated stress there is always the possibility that so-called fatigue cracks can develop. Once formed, a crack can grow until a fracture occurs, with failure not only of the damaged part but, quite possibly, of the entire unit of which it is a component. To prevent such catastrophic failures, it has become the practice in some industries to carry out periodic NDT inspections, even though it means shutting down the operation of a unit for a period of time while the inspections are in progress. Indeed, the usefulness of periodic NDT was demonstrated in a recent survey of electric-power–generating stations in the United States, where it was found that periodic inspections carried out over a 15-year period had reduced expensive failures in these plants considerably from those that occurred before these procedures had become general practice.

The following text describes a number of different NDT procedures. Their aim is to determine the presence of flaws within or on the surface of a structure and, if any exist, their locations and descriptions. The techniques described deal chiefly with the detection of mechanical flaws. Not covered, therefore, are such topics as chemical analyses, measurement of tear strengths of materials, or such specialized topics as determination of concentration and energy levels of impurities in semiconductors.

Among the various NDT techniques, the following stand out as being most widely applied:

1. visual inspection,
2. liquid-penetrant inspection,
3. magnetic-particle inspection and leakage-field measurement,
4. radiographic NDT,
5. ultrasonic NDT,
6. eddy-current testing, and
7. acoustic-emission testing.

These techniques are covered in some detail in Secs. 1 through 7.

In addition, at least 30 other NDT procedures exist. Some of these are quite specialized. Others have been demonstrated in the laboratory but have not yet found wide application. Brief descriptions of these methods are given in Sec. 8.

The reader who would like more detailed information on nondestructive testing procedures than it is possible to present here may be interested in the *Nondestructive Testing Handbook*. This eight-volume set of books of the American Society of Metals (McMaster, 1982a,b) and the American Society for Nondestructive Testing (McIntire and McMaster, 1986; McIntire, 1985, 1987, 1989, 1991; McIntire and Moore, 1993) treats both theory and test procedures.

Finally, while the goals of NDT are to locate and characterize flaws and details of internal structure, in some cases the goal also is to use NDT results for estimating the time to failure, i.e., the interval from the time of the NDT procedure to the instant of failure of the component or structure. In many cases this is a rather difficult task, for it depends not only on the presence of a flaw but also on its size and location, the material, and the stresses to which the component is subjected. Nevertheless, a considerable amount of progress in such prediction methods has been made in recent years. The techniques that have been and are continuing to be developed for carrying out such predictions bear the name fracture mechanics (Hellan, 1984; Mitchell and Landgraf, 1992; Gdoutos, 1993).

1. VISUAL INSPECTION

Although the oldest method of inspection, visual inspection is still generally the first method of examining an object for deviations in shape or for surface faults. However, it is a very subjective method, for the ability to detect such surface faults as small cracks by simple visual inspection is very dependent on the person performing the inspection. It is for this reason that several other techniques of detecting surface cracks have evolved.

Well-known instruments that aid in visual inspection are the ordinary magnifying glass and the microscope. Somewhat less familiar is the borescope, which permits inspection of the inside surface of tubes, such as the cylinder of an automobile engine.

Two types of borescopes are in use. The rigid borescope is simply a small-diameter telescope, in which the image is transmitted toward and in front of the eyepiece by a series of lenses or a single solid fiber with a graded index of refraction. A focusing ring at the eyepiece allows fine focusing. A small lamp in the region of the eyepiece, whose light is transmitted to the distal end either through the lenses or, preferably, through a separate fiber-optic light guide, supplies the illumination for the area being observed. If desired, a mirror or prism at the distal end converts the instrument to a periscope, for viewing surfaces parallel to the borescope.

A more versatile instrument for visual observation of surfaces difficult to access is the fiber-optic borescope. This is a flexible unit that can be several meters in length. It has two sets of fiber-optic light guides between its eyepiece end and the distal end. One of these merely transmits the light from a lamp to the area being examined. The other one is a bundle consisting of many fibers that are precisely aligned to maintain the same relative positions over their entire length. Both ends of this bundle are cut flat and highly polished. An objective lens, whose position is remotely adjustable, is located between the fibers and the area under examination, so as to project an image of this area onto the distal end's surface of the fiber bundle. This divides the image into an array of pixels whose light is transmitted through the fibers to the eyepiece end, to form an image of the viewed area at the ends of the fibers. A lens placed between the fiber ends and the viewing port, with location adjustable to compensate for differences in the eyes of observers, permits viewing of an in-focus image of the area that is being examined.

2. LIQUID-PENETRANT INSPECTION

In this simple visual method of finding surface cracks or regions of porosity, the flaws are highlighted by the use of either a brightly colored liquid or a liquid that fluoresces when illuminated with ultraviolet light. The technique can be applied to virtually all kinds of materials, although its effectiveness is reduced by surface roughness or porosity. The technique is as follows:

1. The surface to be examined is cleaned and allowed to dry. The method of cleaning may include both mechanical and chemical means to remove oil, grease, oxides, and other deposits.
2. The surface is covered with the penetrant, i.e., with the liquid that contains the dye or fluorescent material.
3. After sufficient time has been allowed for penetration of the liquid into cracks or pores, the liquid that has covered the surface is removed, generally with a water spray.
4. A "developer" is applied. This is generally a very fine chalk powder that can be wetted by the penetrant liquid.
5. Around 10–30 min are allowed for the developer to draw the colored or fluorescent liquid out of the cracks or pores. The locations of these flaws are now visible or, in the case of a fluorescent penetrant, can be made visible with ultraviolet light.

Although the highlighting of faults provides a clear advantage over ordinary visual inspection, the method is still somewhat subjective. The fact that the method is not instantaneous may also be a drawback in some cases. Nevertheless, liquid-penetrant inspection is a very popular NDT technique of finding faults that penetrate the surface.

A variation of the method is the use of a fluid with suspended colored or fluorescent particles, used for locating defects in porous surfaces of materials such as unfired dried clay, certain fired ceramics, concrete, some powdered metals, carbon, and partially sintered tungsten and titanium carbide (McGonnagle, 1966). In another variation a surface is covered with a liquid containing a radioactive material, and photographic film is placed over the surface. When developed, the film shows the location of the pores in the surface.

3. MAGNETIC-PARTICLE INSPECTION OR LEAKAGE-FIELD MEASUREMENT

Magnetic nondestructive testing generally deals with the inspection of components constructed of ferromagnetic materials, such as iron and steel, although it could also be applied to ferrimagnetic materials, such as ferrites. These magnetic inspections aim either at finding and evaluating surface or subsurface cracks in a component or at finding possible differences in the magnetic material from the intended composition or microstructure. Methods of finding cracks are the very popular NDT technique of magnetic-particle inspection or an alternative method of leakage-field measurement. Both are treated below. A third technique that deals with magnetic effects, Barkhausen-noise analysis, yields microstructural information. It is briefly described in Sec. 8.

3.1 Magnetic-Particle Inspection

The magnetic-particle method of finding cracks on or just below a surface of ferromagnetic material is based on the principle that a discontinuity in a magnetized material will cause leakage magnetic fields, which will attract and hold fine iron particles.

In discussing magnetic fields that exist in or around materials, it is sometimes useful to recall that dc magnetic fields exist because of the presence of electric currents. Within a magnetic material these electric currents will actually be elemental circulating currents at the atomic level, i.e., elemental magnetic dipole moments, which can also be pictured as elemental bar magnets. For a steel bar these elemental dipole magnets are found to be organized into so-called "domains," i.e., small regions within which the dipoles are parallel to each other. In the absence of presently or previously applied magnetic-field intensity, the domains are randomly oriented, so that no net dipole moment exists. When a coil of wire is wrapped around the bar and an electric current is caused to flow through the wire and this current is gradually increased, more and more of the domains orient themselves parallel to the axis of the bar. To specify magnitude and direction of the density of these oriented magnetic dipoles, it is appropriate to define a vector quantity **M**, the magnetic dipole density, with units of amperes per meter. **M** is generally called the "magnetization."

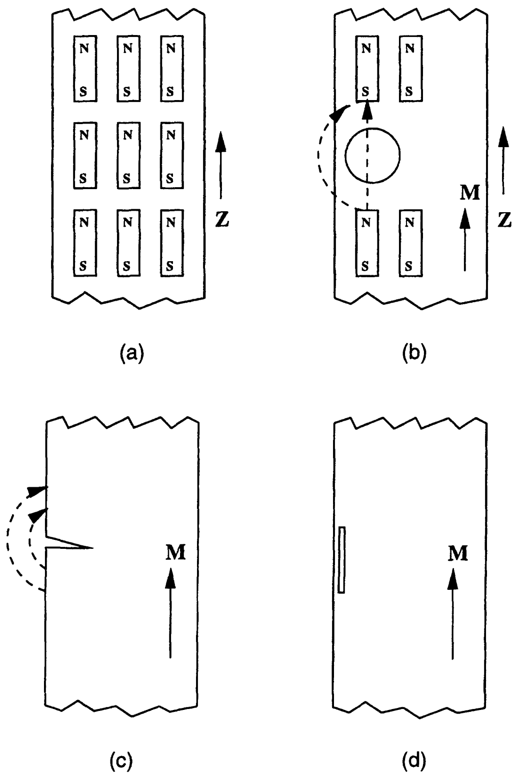

FIG. 1. Detection of surface and shallow subsurface cracks in magnetized ferromagnetic materials. (a) Aligned magnetic dipoles in bar without flaws. (b) Presence of cylindrical hole causes leakage of magnetic-field lines to outside of bar. (c) Surface crack normal to the direction of **M**. (d) Narrow crack parallel to **M**.

We now consider a bar of steel that is uniformly magnetized so that the dipole moments, and therefore **M**, point in the z direction, as in Fig. 1(a). Here, since each elemental south pole is adjacent to an elemental north pole, as shown in Fig. 1(a), there is no magnetic field external to the bar. Consequently, if small iron particles were blown from the left toward the bar, there would be no force to attract and hold them. Let us now suppose that a cylindrical hole were to exist near the edge of the bar, as in Fig. 1(b). The continuous north pole–south pole structure is now interrupted, leaving poles dangling at the hole. Since each of these poles acts as a source or sink of a magnetic field, the magnetic field in the hole is now different from that in the solid piece. Moreover, as seen in Fig. 1(b), there is now a curvature of magnetic-field lines within the bar, with a "leakage" magnetic field that reaches outside the bar. If any small iron particles were to be blown toward the bar now, or if the bar were immersed in a liquid that had iron particles suspended in it, this magnetic field would attract them and they would stick to the bar. If a surface crack normal to the direction of **M** were to exist in a magnetized bar, as in Fig. 1(c), the stray magnetic fields would again attract any iron particles and keep them stuck to the bar.

It is seen and borne out by analysis that a surface or near-surface crack normal to the direction of magnetization, as in Fig. 1(c), produces a leakage magnetic field that is readily detected. On the other hand, a narrow crack that runs parallel to **M**, as in Fig. 1(d), has very little external leakage field. Consequently, to detect such a crack or, in general, a crack oriented in an arbitrary direction, it is necessary to perform three tests, with the unit being tested successively magnetized in three orthogonal directions.

Actually, in most cases only two tests are necessary. An example is the section of steel pipe sketched in Fig. 2(a), with surface cracks A and B. Here current flowing through a wire on the axis of the pipe, as in Fig. 2(b), will generate a circumferential magnetization that will create appreciable external leakage fields for crack A and only negligible fields for crack B, thereby permitting detection of crack A. The converse is true for Fig. 2(c), where the longitudinal magnetization in the pipe will allow detection of crack B.

The techniques that are used for generating the magnetic fields depend on the size and configuration of the components to be tested. In the configurations of Fig. 2 the magnetizing currents flow through auxiliary wires. In some situations the current used to magnetize a unit under test can be caused to flow through the unit itself. Both dc and ac currents, as well as rectified ac, will generate magnetic fields that will attract and hold magnetic particles. In some cases it is possible to use ac polyphase methods that create rotating magnetic fields, thereby obviating the need for two or more test setups.

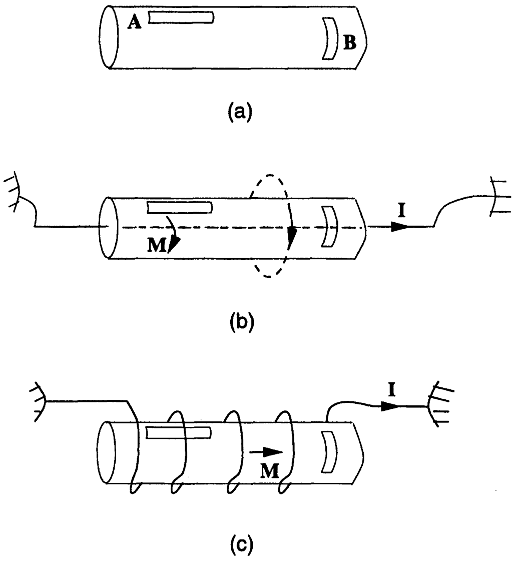

FIG. 2. Detection of cracks in a steel pipe. (a) Pipe with crack A parallel to axis and with circumferentially oriented crack B. (b) Axial current generates circumferential field and magnetization, thereby allowing detection of crack A. (c) Solenoidal current generates axial field and magnetization, thereby allowing detection of crack B.

For most steels, the finest crack widths observed by this magnetic-particle technique are of the order of 10 μm (Betz, 1967), although detection of a 1-μm crack in carburized steel has been reported (McIntire, 1989). The ability to detect subsurface flaws is a function of their shape, size, and orientation. Using dc magnetizing currents, detection depths of the order of 3 to 7 mm are reported possible, while the practical depth limit with ac magnetization is estimated to be less than 1 mm (Hull and John, 1988).

When a material has a high remanence, so that it stays magnetized after a dc current is switched off, the inspection can sometimes be carried out by using only this remanent magnetization. Known as the "residual method," this has the advantage that the test may be performed without the physical presence of the magnetizing equipment. However, the sensitivities in the residual method are generally lower than those achieved when magnetizing current is used.

The magnetic field resulting from any remanent magnetization after either type of inspection has been carried out could interfere with subsequent operations, such as arc welding or the use of magnetically sensitive instruments. It could also attract stray iron particles or chips that could damage such surface-sensitive components as bearing raceways. Moreover, the remanent magnetization could interfere with the electrical operation of a unit. It is therefore necessary to demagnetize the test piece after the NDT procedure.

A standard method of demagnetization is to subject the magnetized piece to a time-varying, gradually decreasing magnetic-field intensity by applying an ac voltage to a coil surrounding the magnetized component while gradually decreasing the amplitude of this voltage to zero. If the initial magnetization was due to an applied dc current, it may be necessary to carry out the demagnetization process by a combination of applying a dc current and the use of the ac demagnetizing technique (Hull and John, 1988).

In practice, the magnetic particles used to mark locations where leakage magnetic fields exist are usually of ferromagnetic material of low remanence, so that they do not adhere to the components under test because of their own magnetization. Both dry application of particles or particles in liquid suspension are in use. To make them highly visible, it is the practice to use particles that are either colored or coated with a material that will fluoresce when illuminated with ultraviolet light.

Since magnetic-particle inspection is an inexpensive and sensitive method of finding surface or near-surface cracks, it is widely used. Its chief drawbacks are these:

1. It is applicable only to ferromagnetic or ferrimagnetic materials.
2. Tests with at least two orientations of magnetization, or a polyphase system, are required.
3. To test large components, very large currents may be required to generate the necessary magnetizations.
4. Although relatively simple, complete success depends on the skill and experience of the person performing the test.

3.2 Magnetic-Field Measurement

While the magnetic-particle method is the preferred method of finding leakage fields that are due to flaws at or near the surface of ferromagnetic materials, these fields may also be discovered by direct measurement. Among the instruments that have been used for such sensitive magnetic-field measurements are

1. moving coil magnetometer—essentially a small rotating coil ac or dc generator;
2. Hall-effect probe—a semiconductor chip in which an electric field transverse to the direction of a drift current is developed because of the deflection of carriers by the magnetic field;
3. flux gate magnetometer—a sensitive instrument in which an ac-driven ferromagnetic bridge circuit develops a second-harmonic voltage whose amplitude is determined by the magnetic field in which it is located; and
4. magnetic tape—magnetic recording tape placed into region where field is to be measured, to be scanned after removal by magnetic-tape reader, to reveal field distribution.

4. RADIOGRAPHIC NDT

Radiographic NDT, also called radiology, uses x rays or gamma rays—electromagnetic waves whose wavelength is of the order of 1 Å or smaller—to penetrate materials opaque to visible radiation, to find flaws, and to map internal structure of bodies. As one of the oldest NDT methods, dating back to Roentgen's discovery of x rays in 1895, it has been popular since the late 1930s and remains as one of the most frequently used NDT techniques. This is despite the well-known fact that both x rays and gamma rays are ionizing radiation and their use requires proper protection for the person performing this type of NDT, either by appropriate shielding or by operation of sensing and recording instruments from a safe distance.

The wavelengths of x rays and gamma rays used for NDT are approximately in the same range of the spectrum. It is the practice to refer to the broadband radiation generated by high-velocity electrons impacting on metal surfaces as "x rays" and to the line spectrum radiation emitted by radioactive materials as "gamma rays."

4.1 Principle of Operation

Radiographic NDT is based on the facts that x rays and gamma rays travel through materials in straight-line paths and that the intensity of the radiation in passing through matter decays as $I = I_0 \exp(-\mu x)$. Here I_0 is the intensity at the beginning of the path, I is the intensity after passing through path length x, and μ is an attenuation constant that depends on the material.

It is the materials-dependent quantity μ that is the key to obtaining information about the internal structure of a test piece. For if a broad beam of x rays or gamma rays is sent through this test piece, any gaps or irregularities within the structure represent deviations in μ from that of the rest of the body and therefore produce local changes of the radiation intensity in the area of the beam as it exits the structure. Irregularities within the body are found by sensing and displaying the spatial distribution of radiation intensity within the beam area.

An example is seen in Fig. 3. In Fig. 3(a) an x-ray beam is directed to enter a metal bar that contains an inclusion, a small po-

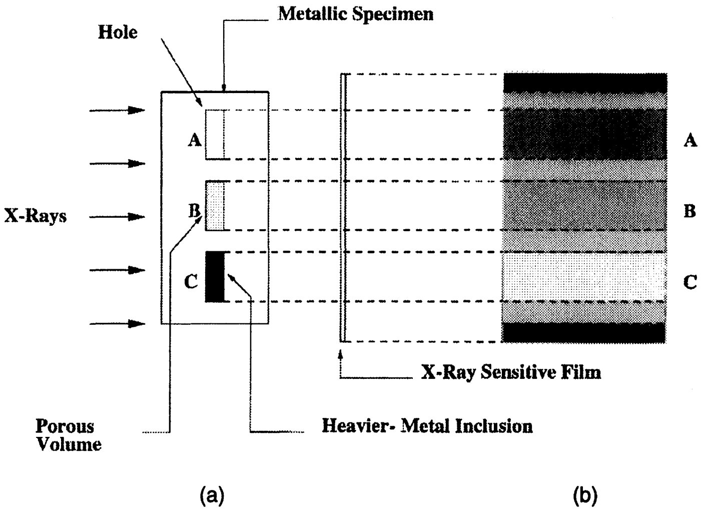

FIG. 3. Principle of radiography. (a) Side view of x rays entering a metallic block that has several regions extending perpendicularly to the plane of the paper throughout the specimen: Region A, a hole; region B, a porous volume; region C, an inclusion of metal of higher density than that of the block. (b) The various shades of gray resulting on the developed x-ray film.

rous volume, and a hole. A sketch of the relative intensity of the radiation on exit from this test piece is seen in Fig. 3(b). When a strip of photographic film sensitive to exposure by x rays is placed so that its plane is perpendicular to the beam, a latent image forms on this film. After development, the degree of "blackness" of a spot in the distribution of varying transparency of the film indicates the relative amount of radiation per unit area that has reached this spot. An experienced observer can conclude from this film strip that region A contains a hole, that there is a shallower hole or a porous volume in region B, and that an irregularity of some sort exists in region C.

Two points are evident in this NDT technique:

1. Accessibility to opposite sides of the test piece is required.
2. The developed film is essentially a shadowgraph that depicts the relative transparency to the radiation along the straight-line paths followed by the radiation. The "blackness" of a spot on the film is therefore an integrated response of the entire path traveled along the straight-line path of the radiation reaching this spot.

For the sake of demonstrating the basic idea of radiography, the rays representing the x-ray beam of Fig. 3 were drawn parallel. While a parallel x-ray (or gamma-ray) beam can be approximated by the use of a large separation between a physically small radiation source and the test piece, a more efficient method uses a small source, with subsequent data manipulation that takes beam divergence into account.

In the example of Fig. 3 the final step was interpretation of a pattern displayed on photographic film. Other types of visual display media are also in use and are mentioned later in this article. In all cases, though, radiographic NDT is based on interpretation of two-dimensional displays.

4.2 Three-Dimensional Information

Although a two-dimensional display such as that of Fig. 3(b) may be sufficient in many cases, it lacks information on the depth distribution of flaws or irregularities within the test piece. Additional information, though not a complete depth distribution, can be obtained by repeating the procedure with the test piece rotated by, say, 90° with respect to the first exposure. Consequently, for most cases one or two images can provide sufficient information about the integrity of the test piece.

For more detailed information, the operation of Fig. 3 can be repeated, with the test piece rotated with respect to the incoming radiation to *several* different angular positions. If the distributions of intensity at the film plane for each orientation are then used in a mathematical inversion process, detailed depth information of the internal structure can be obtained (Radon, 1917). One example of this technique is computer-aided tomography, well known in medical technology as "CAT scan," which pictures two-dimensional mass distributions within what are essentially slices of the scanned material (Kuhl and Edwards, 1963; Hounsfield, 1973).

4.3 Unsharpness, Contrast, Sensitivity

4.3.1 Unsharpness The detection of flaws of minute size or the accurate representation of internal structure requires high spatial resolution. Because it is generally not possible to focus x rays or gamma rays, it is important to have the source of the radiation as small as possible. This requirement is illustrated in Fig. 4, where the diameter of an aperture in a thin plate of material that transmits no x rays is to be determined by measurement of the diameter of the display on a photographic film.

Because of the short wavelengths, spatial uncertainty in the image due to diffraction effects are negligible, so that for the point source of Fig. 4(a), diameter d is almost precisely given by $d = (o/i)\ D$.

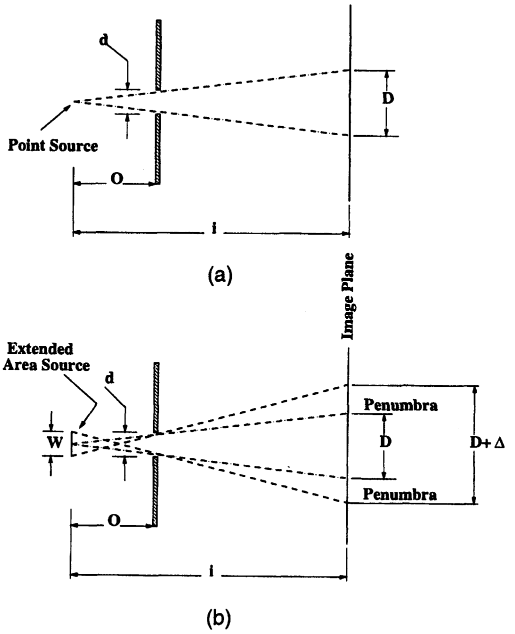

FIG. 4. Unsharpness caused by an x-ray source of extended area. (a) Point source causes illuminated region of precise width D. (b) Source of extended area produces a penumbra region of unsharpness.

In Fig. 4(b), on the other hand, the width W of the source creates a penumbra, or partial shadow region, which causes a blurriness in the image and therefore an uncertainty in the determination of dimensions and location and width of the aperture. In the language of radiographic NDT, this uncertainty is generally called the "unsharpness." Here this unsharpness is due to the size of the radiation source. It may also have other causes and is an important consideration in the design of a radiographic NDT system.

4.3.2 Contrast An example of the importance of the consideration of "contrast" is the following question: Since local changes in thickness of a specimen reflect as changes in the blackness of the film and should therefore be recognized visually by the contrast between various portions of the display on the film, what is the minimum change in thickness Δx that can be detected?

An estimate can be made by calculating a quantity called the "subject contrast," which is used to refer to the fractional change in the radiation transmitted because of a change in thickness or the presence of a foreign body within the test piece. Based on the expression $I = I_0 \exp(-\mu x)$, mentioned in Sec. 4.1, the subject contrast for a small change in thickness Δx is approximately $\mu \Delta x$. Similarly, if the radiation passes through a foreign body within the test piece whose thickness is x_0 and whose attenuation constant is $\mu + \Delta\mu$, where $\Delta\mu$ is considerably smaller than μ, the subject contrast is approximately $x_0 \Delta\mu$.

The calculation of subject contrast yields an optimistic estimate, for the actual contrast in a display depends on a number of factors, including fogging of the film by "noise" background caused by scattering of the radiation by the test piece and the characteristics of the film or other display system. More general expressions exist that also take these other factors into account. Some techniques to increase contrast have been devised and are mentioned in a later section.

On the basis of practical experience, it has been found that a gap that is less than 2% of the overall thickness of an item under test is difficult to observe (Cartz, 1995). A fine crack will therefore probably not be observed if the plane of the crack is perpendicular to the path of the radiation.

4.3.3 Sensitivity The term "radiographic sensitivity" refers to the smallest detail that can be seen on a radiograph, or the ease with which small details can be detected. It is a quantity that is somewhat subjective, for it refers to detail resolvability, not to spatial resolution. An example is the case of a small flaw that is an inclusion whose density is quite different from that of the region surrounding it. If this density difference is large, the flaw may yet be discovered, even if it is smaller than the spatial resolution of the system. A flaw whose density is quite close to that of the rest of the test piece would need to be considerably larger in order to be detected.

4.4 Radiation Wavelengths or Photon Energies Required

To produce a shadowgraph image within a reasonable time, the radiation incident on a test piece should be of readily achievable intensity. It also should have a wavelength that allows penetration, but with sufficient attenuation to permit the detection and imaging of incremental structural irregularities, such as a gap or an inclusion.

A quantity that has been used to determine the appropriate wavelength of the radiation required for a particular situation is the "half-value layer" (HVL), i.e., the thickness of material that reduces radiation intensity by one-half. Values of HVLs as a function of wavelength λ for different materials have been measured and are published. Instead of listing these as a function of λ, however, they are often listed as a function of the voltage V_a applied to the electron tube used to generate the radiation, since a photon of the shortest wavelength (and therefore of highest energy) in an x-ray beam so generated has wavelength $\lambda_0 = ch/eV_a$, where c is the velocity of light, h is Planck's constant, and e is the electronic charge.

As examples of the magnitudes involved, it is found that HVL values for concrete are 0.51 cm for $V_a = 50$ kV and 4.5 cm for 1 MV, whereas for steel and $V_a = 1$MV, HVL = 1.6 cm (Bryant, 1985). These numbers indicate that high voltages are required to generate the x rays for NDT of solid bodies. Thus, steel structures thicker than a few centimeters require radiation of shorter wavelengths, corresponding to photon energies greater than 1 MeV.

The maximum thickness of material that can be examined depends also on the method of detection and imaging. It is found that for typical commercial equipment, steel specimens examined with 1-MV x-ray equipment can have maximum thickness of 16 cm (Halmshaw, 1982).

Since the objective of radiographic NDT is to obtain an image that displays sufficient contrast for finding flaws or showing internal structure, the choice of energy of the radiation to be used depends on both the material and its thickness. In general, it has been found that a radiograph should be made with radiation of the longest wavelengths, i.e., the lowest photon energies, that will transmit adequate radiation intensities to the detecting medium, because long wavelengths tend to improve contrast. As an example, it would be fruitless to use 1-MeV radiation to examine a 1-mm-thick sheet of plastic, since virtually all of the radiation would pass through unattenuated, and no evidence of any internal holes would be found. On the other hand, radiation with wavelengths that are too low produce excessive scattered radiation, which reduces subject contrast. So there is an optimum range for each situation.

4.5 Sources of Radiation

4.5.1 X-Ray Sources X rays used in NDT are generated in vacuum tubes, in which electrons emitted from either hot or field-emission cathodes are formed into beams, accelerated by electric fields to energy eV_a, and caused to impact metal targets. Target materials most commonly used are tungsten or gold that is set in copper.

The interaction of the electrons with the target material generates radiation of a continuous electromagnetic spectrum, with the minimum wavelength λ_0 as given above. Some line-spectrum radiation is also possible but is unimportant in ordinary x-ray shadowgraph NDT. The continuous, or "white," spectrum has a broad peak around wavelength $1.5\lambda_0$. It is customary to refer to the x rays so produced in terms of the voltage

V_a. Thus, if the electrons on impact had maximum kinetic energy of 10^5 eV, the x rays produced would be called "100-kV x rays."

To minimize the unsharpness in the image produced by the x rays, the electron beam is generally focused so as to impact the target in a small spot. In conventional x-ray tubes, this spot size is of the order of 3 mm square. X rays are then emitted in a cone whose axis is perpendicular to the emitting surface. For specialized applications, such as producing a magnified image when the x-ray source is very close to the recording film, so-called microfocus tubes have been with fabricated, with spot sizes down to 10–20 μm.

Alternative techniques of effectively reducing the size of the x-ray emitting area or causing emission in preferred directions consist of tilting or appropriately shaping the target.

Among the equipment used to attain the required high voltages are power supplies containing voltage-doubling circuits in units operating up to 400 kV, air-core resonant transformer circuits for 1–2 MV, and Van de Graaff generators for 1–3 MV.

Vacuum tubes in which electrons are accelerated to MeV energies by dc or low-frequency fields, together with their high-voltage supplies, occupy so much space that they cannot be readily moved from one location to another. A successful alternative for generating million-volt x rays is the use of betatrons, electron cyclotrons (microtrons), or linear accelerators (linacs)—instruments that accelerate electrons to MeV energies without high electric fields external to the chamber. Such instruments can be made sufficiently compact so that they can be readily transported.

Betatrons have found considerable use in generating x rays from 2 to 31 MeV. Microtrons, though also good sources of x rays above 1 MeV, have not found great popularity. Linacs, with energy ranges from 1 to 25 MeV, are now the preferred instruments, for their x-ray output can be many times that of betatrons of the same energy (Halmshaw, 1987).

4.5.2 Gamma-Ray Sources Gamma-ray sources for NDT consist of small pellets of radioactive material that emit line-spectrum radiation of photon energies from 50 keV to several MeV, corresponding to wavelengths of 0.025 nm and shorter. The principal isotopes in use are cobalt-60, iridium-192, and ytterbium-169, with a much smaller usage also of cesium-137, thulium-170, and sodium-24. Cobalt-60, for example, emits photons of energy 1.17 and 1.33 MeV, corresponding to wavelengths of 0.00106 and 0.00093 nm. As used, a pellet is encased in shielding material, such as lead or depleted uranium, so that virtually no radiation exists external to the shielding. Mechanical systems have been devised to bring the pellet from its shielded location to the spot from which it is used to irradiate the test piece. The effective source size at this location can vary from 0.3 to 10 mm.

The advantage of using gamma rays over x-ray sources is that gamma-ray sources can be physically small, are portable, require no power or water supply, and are often less expensive. One possible disadvantage is that they could present more of a health hazard than x-ray equipment because they emit radiation continuously, whereas x-ray sources are switched to be active only during the NDT procedure. Another disadvantage in some cases is the high energy of the radiation, which, for example, causes the sensitivity of gamma-ray NDT for steel specimens less than 2.5 cm thick to be poorer than that obtainable with x rays of lower energy. Moreover, longer exposure times are required, because the radiation intensity is generally lower than achievable with x-ray apparatus.

Finally, another aspect of the use of gamma rays is the radioactive decay of the source, requiring periodic source replacement. A guide here is the half-life of the isotope used. For the three most popular materials these are for cobalt-60, 5.3 years; for iridium-192, 74.4 days; and for ytterbium-169, 31 days.

4.6 Sensing and Recording

The final steps in radiographic NDT are sensing and displaying the information in the beam that has passed through the test piece, and interpreting the results. This section deals with several methods of sensing and imaging.

4.6.1 Radiographic Film

4.6.1.1 Description. The technique of using photographic film designed for imaging spatial distributions of x rays and gamma rays remains as the most widely employed method of radiographic NDT. Since the intensities of the radiation striking the film are frequently quite weak, requiring from seconds to minutes of exposure times for x-ray sources and from a few minutes to an hour for gamma-ray sources to reach the necessary contrast in the image, means are employed to increase detection sensitivity. One of these is to use film that has a coating of photographic emulsion on *both* sides of the flexible film base.

Another technique is the use of intensifying screens. Here one distinguishes between "direct" type, also called "nonscreen" type, and "screen" type films. In the former, metallic screens, usually lead foils, are pressed on the two emulsions. Radiation absorbed in these metal screens generates free electrons, which pass into the emulsions, add to the photographic action, and thereby reduce exposure times by as much as a factor of 6. The foils used are very thin (0.02 to 0.2 mm), to prevent electrons from spreading sidewise before reaching the emulsion and increasing the image unsharpness.

An additional benefit accruing from the use of these metal screens is the reduction of "buildup," i.e., longer-wavelength radiation that arises through Compton scattering in the specimen and causes a type of fogging of the film.

Screen-type films have emulsions coated with fluorescent powders in a binder. When struck by x rays, these "salt screens" generate ultraviolet (UV) radiation. Since the emulsion is more sensitive to UV than to x rays, exposure times are reduced. However, while satisfactory for medical x-ray use, the image quality is not considered adequate for the precision needed in industrial NDT except for cases that do not require critical examinations, and for process control. The UV from these salt screens can also be used to expose photosensitive paper, to get positive prints of the x-ray pattern—again, without the great precision obtainable with film.

A successful combination of the two materials is the fluorometallic screen, which has lead layers outside of the fluorescent layers. The lead layers reduce the buildup, the fluorescent layer reduces the required exposure time from factors of 2 to 7, and the image quality exceeds that of the regular screen type of film.

4.6.1.2 Film Parameters. The following photographic parameters relate to the use of radiographic "direct" film (Halmshaw, 1987; Cartz, 1995):

1. Film density D: Films are generally exposed to a density between $D = 1.5$ and 3.5. [D relates to the viewing of the exposed and developed film with visible light and is defined as $\log_{10}$ {intensity of light incident on film/intensity of transmitted light}.]
2. The minimum density, the so-called "fog level," is usually around 0.2 to 0.3.
3. Film gradient G_D, also called "film contrast": G_D, a measure of the film's sensitivity to small changes in incident x-ray level, is the slope of the curve of D when plotted as a function of $\log_{10} E$, where E is the product of the x-ray intensity and film exposure time. It is found that G_D increases with D and that fine-grain film is more sensitive to changes in x-ray intensity than coarser-grain film.
4. Film speed: Radiographic films available from manufacturers cover a range of 20:1 in the measure of film sensitivity that is generally called "film speed."
5. Unsharpness: Among several factors that contribute to unsharpness of the image are size of radiation source and distance from test piece, thickness of intensifying screen, film unsharpness, relative motion of source and test piece, and secondary emission of electrons in the film. Surprisingly, results of experiments indicate film unsharpness to be almost independent of film graininess.
6. Resolving power: As can be expected, the greater the unsharpness, the poorer is the resolving power. A number quoted for 100-kV x rays and conventional radiographic film is 60 line pairs per mm.

4.6.1.3 Viewing the Image. A very important aspect of the conventional method of viewing the film and interpreting the image is an arrangement of good viewing conditions. This must include low ambient viewing-room illumination, no glare, and the necessary luminance. It has been found that to see small details and low contrast differences requires a minimum luminance of at least 30 cd/m^2. To see detail in areas on the film in which the density is 2.0, the luminance incident on the film is required to be at least 3000 cd/m^2.

An alternative method of film examination uses a television camera, which adds the features of permitting exploration of details in images of very low brightness, image magnification as desired, and performing computer-based image analysis.

4.6.1.4 Measurement of Radiographic Sensitivity. In industrial radiography the term "sensitivity" generally means the ability to show detail. Ever since the inception of radiographic NDT, special fixtures have been used to measure sensitivity. Originally called "penetrameters," they are now called "image quality indicators," or IQIs. Although the ability to detect a flaw in a casting or weld depends on the actual situation, the IQI sensitivity is a good measure of the size of flaw that can be detected.

Several types of IQIs are in use. A representative example is a fixture in which a series of wires of different diameters but of the same material as the specimen to be tested are strung in a frame, in essentially the same plane. To establish the sensitivity, this IQI is placed on top of the specimen, and the combination is irradiated by the x-ray or gamma-ray source of interest so that the beam passes perpendicularly through the plane of the wires. The diameter of the thinnest wire that can be discerned on the film image is a measure of the minimum size flaw that can be detected.

4.6.2 Other Sensing and Recording Media At the time of this writing, the method of recording on film has the greatest resolution over a large area. However, several other methods are also in use. Among these are the following:

1. Paper radiography—For some applications, photosensitive paper is used to record images formed by UV radiation that is emitted by salt screens struck by x rays.
2. Xeroradiography—In xerography a latent image is formed on a positively charged photoconducting plate by discharging areas of the plate to varying amounts with spatially varying x radiation that has passed through the test piece. Negatively charged pigmented particles that are brought in contact with the plate form a visible image, which can be transferred to a positively charged sheet of paper. The principal use of xeroradiography is in medical radiography.
3. Scintillating materials—Substances such as NaI(Tl), which can be fabricated in sheet form, emit light when struck by x rays. When backed by a fiber-optic plate, the structure becomes an area detector.
4. Semiconductor detector arrays—Semiconductor arrays make use of the x-ray photoconductivity of materials such as cadmium sulfide or of semiconductor junction devices. Since such components can be made microscopic in size, they can be built into line or matrix arrays of discrete x-ray detecting elements. Array-scanning circuitry develops an electrical signal by sensing changes in these elements as they are exposed to the radiation, such as changes in conductivity. This signal can be used for generating a video display, for signal processing, or for creating a "hard-copy" record.
5. Fluoroscopy—Fluorescent screens that emit visible light when struck by x rays or gamma rays present a visible image, although one that is generally of low brightness. Vacuum-tube or solid-state image intensifiers are available for improving brightness and contrast.

4.7 Real-Time Radiography

Real-time radiography, also known as radioscopy, is the technique of using results from x-ray or gamma-ray NDT procedures to create analog images for continuous viewing while these procedures are being carried out. Closely related is near-real-time radioscopy, in which the formation of images occurs after a brief time delay. One such radioscopy system is x-ray detection by the discrete components of an array, followed by array readout and the creation of an image on a television screen.

Since there is essentially real-time image formation, radioscopy techniques are applicable when there is motion involved. A well-known example is the system of luggage inspection at airports, in which a slit source of x rays shines through moving baggage toward a linear detector array. Another application is a system in which a test piece is rotated while being irradiated, to find optimum directions for finding and characterizing flaws.

The earliest system of radioscopy used a fluorescent screen, which converted incident x-ray spatial distribution to a visible-light display. A mirror placed at an angle reflected the light from the display, so that a person could see the image in the mirror without being in the path of the x-ray beam. Systems of this type are still used for examining items of low density, such as plastic or paper packets. For general inspection of metals, however, the image is too faint.

An improvement over this basic system substitutes a closed-circuit television system (CCTV) for direct viewing. Now the operator can be well removed from the test site, so that the x-ray hazard has been eliminated. Moreover, by using CCTV with image intensification, the problem of low light level has been alleviated.

An additional improvement is to use image intensification prior to image exposure to the CCTV camera. One such image intensifier is an electron tube in which the light intensity is amplified by causing electrons emitted by a layer of photocathode material adjacent to an x-ray-excitable fluoroscopic layer to form a smaller but much brighter image. A diagram illustrating such a system is seen in Fig. 5.

A real revolution in radioscopy has been image enhancement by the use of digital signal-processing techniques such as frame averaging, point brightness shifting, and spatial filtering. For example, contrast can be improved by shifting the gray scale of an image to where the human eye is most sensitive to small changes in contrast. Gray-scale steps normally too small for the eye to perceive can be stretched; others can be compressed, to preserve dynamic range. The image can be shifted to contrast and brightness ranges in which the eye is most sensitive. Fixed pattern noise can be removed. And the resolution can be improved by selectively amplifying high spatial frequencies.

A long-sought goal of radiographic NDT has been automatic inspection. So far it has been found difficult to replace an experienced human operator with a fully automatic inspection system. Nevertheless, there are expectations that new developments in expert systems and artificial intelligence will continue to make improvements toward the goal of automatic image evaluation.

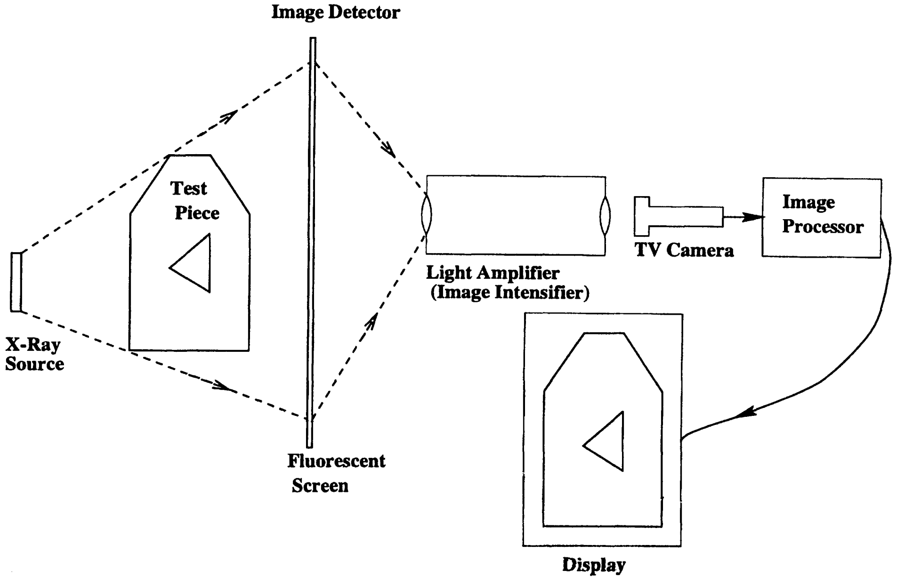

FIG. 5. Real-time radiography system.

5. ULTRASONIC NDT

Elastic waves, also called acoustic waves, can propagate in many liquids and solids over considerable distances without appreciable attentuation. Their velocity is five orders of magnitude lower than the velocity of light. These characteristics and the ease of generating, detecting, and precisely timing their progress by interaction with electrical systems have made them an important medium for carrying out noninvasive searches for imperfections in both solid and quasisolid bodies and in imaging internal structures. The latter includes living structures, such as the human body.

As shown in Sec. 5.4.3.3, to attain reasonable definition of the location of a flaw in many cases of interest, it is required that the waves used should have frequencies well above 20 kHz. Since this is above the audible, or sonic, range, the term "ultrasonic NDT" is appropriate.

5.1 Wave Types

A number of different types of waves, or modes of propagation of elastic energy, can exist in solids, each with its own velocity of propagation, particle motion, and pattern of stress and strain. Most often used in NDT are bulk waves, i.e., waves that exist without dependence on the thickness or shape of a specimen, and of wavelengths considerably shorter than specimen dimensions. Others either have actually also been used for NDT or have shown possibilities for use in special cases. Among the various modes are the following listed below. For comparison, when appropriate, the propagation velocity v of each mode in aluminum is given (McGonnagle, 1966).

5.1.1 Bulk Waves Bulk waves are of two types:

1. Longitudinal waves (LW). These are compressional/dilational waves. Particles that participate in the propagation oscillate parallel to the direction of propagation. In aluminum $v = 6.25 \times 10^3$ m/s.
2. Transverse or shear waves (TW). Particles oscillate transverse to the direction of propagation. In aluminum $v = 3.1 \times 10^3$ m/s. (The term "shear waves" is preferred in the NDT industry.)

LW can propagate in liquids as well as in solids. It is possible to generate LW in water, for transmission into a solid, *et vice versa.* TW will not propagate in liquids.

As stated briefly in Sec. 5.2.3, it is possible to convert energy from one type of wave to the other.

Both LW and TW may be used for NDT, with the choice depending on the application. For example, LW are used for measuring the thickness of a rectangular block or film; for searching for flaws in welds, TW are preferred.

5.1.2 Structure-Dependent Waves

Analysis of the possibility of acoustic-wave propagation for various structures has yielded several different wave types, or modes. The following are of interest.

5.1.2.1 Rayleigh Waves. Somewhat analogous to the propagation of water waves, Rayleigh waves are waves that propagate on smooth solids, with particle motion principally in a thin surface layer. Particle oscillations are elliptical, in the plane that is parallel to the direction of propagation and perpendicular to the surface. In aluminum $v = 2.8 \times 10^3$ m/s. Also known as surface acoustic waves, or SAW, Rayleigh waves have found uses in electronic circuits for carrying out functions of frequency filtering and high-frequency analog data processing.

A very useful feature of Rayleigh and other surface waves for NDT applications is that they can propagate on curved surfaces. Since propagation is very sensitive to surface flaws, surface waves can be used to detect surface or shallow subsurface defects, such as nicks or points of corrosion, not only on plane surfaces but also on such curved surfaces as cylinders or springs.

5.1.2.2 Love Waves. This type of surface wave can exist in a structure that is covered with a thin layer of material whose density differs from that of the substrate. Particle oscillations are in the plane parallel to the surface. Wave velocity is material and frequency dependent.

5.1.2.3 Lamb Waves. Whereas Rayleigh and Love waves propagate in a thin layer along *one* surface, Lamb-wave propagation is in thin layers that have *two* surfaces, i.e., it is propagation in thin sheets. A number of different patterns, or modes of propagation, are possible, each with a velocity that depends on the material, sheet thickness, and frequency. Anomalies in sheet thickness, such as the presence of corrosion or the presence of thin internal cavities, can be detected by measurement of wave velocity.

5.1.2.4 Stonely Waves. These are waves that propagate along the interface of two different materials. Since wave velocity is sensitive to surface conditions at the interface, the health of the bonding surface can be determined by comparative velocity measurements (Hsieh *et al.*, 1991).

5.2 Reflection, Refraction, and Mode Conversion

5.2.1 Reflection The principal ultrasonic NDT method used to find flaws in a specimen is the observation of reflections generated by these imperfections. Similarly, an ultrasonic map of a structure is based on observing and plotting the intensity and/or pulse-transit time of reflected signals. As in the treatment of electromagnetic waves, it can be said that reflections occur whenever a wave propagating in a medium of a particular "characteristic impedance Z_0" encounters a different characteristic impedance in its path. A characteristic impedance Z_0 can be defined for each mode of propagation. For LW and TW, the waves most often employed in NDT, Z_0 is defined as the product of wave velocity and material density.

For the simple case of plane waves in a medium of Z_{01} impinging perpendicularly on a boundary separating this medium from a medium of Z_{02}, the ratio of reflected energy to incident energy is $(Z_{02} - Z_{01})^2/(Z_{02} + Z_{01})^2$. More complicated expressions exist for oblique incidence.

5.2.2 Refraction Snell's well-known law of refraction for electromagnetic plane waves holds also for both longitudinal and shear acoustic waves. Thus, referring to Fig. 6, if v_1 and v_2 are the wave velocities in medium 1 and medium 2, respectively, it is found that $\sin \Theta_2 = (\sin \Theta_1)(v_2/v_1)$.

Of special interest is the case when $\sin \Theta_1$ has the value v_1/v_2, so that $\Theta_2 = 90°$. For this "critical angle" of incidence Θ_c and for angles larger than this, no waves of the original configuration are transmitted into medium 2. (However, through the phenomenon of mode conversion, waves of other configurations may be generated. Specifically, it is possible to generate a strong Rayleigh wave in medium 2 when Θ_1 is near the critical angle for shear waves.)

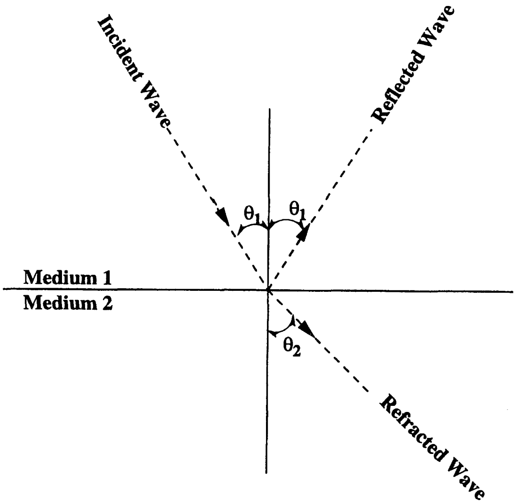

FIG. 6. Reflection and refraction of uniform plane waves at a boundary.

5.2.3 Mode Conversion In addition to the reflection and refraction that occurs at a boundary for both LW and TW, there is also a mode conversion. For when either type of wave propagates toward a boundary, reflected and refracted waves of both types exist (Knott, 1899). Since the velocity ratios of the two media generally differ for the two wave types, each of the waves propagates at its own angle. This makes it possible, for example, to generate shear waves (TW) from a source that would normally generate longitudinal waves (LW). By use of the critical-angle phenomenon, it is then possible to select only the type of wave desired and use it for NDT.

5.3 Acoustic-Wave Generation and Detection

5.3.1 Transduction—Longitudinal and Shear Waves In most cases both the generation and the detection of ultrasonic waves are carried out by the use of piezoelectric transducers that convert electrical to elastic energy, *et vice versa*. In early work these transducers were wafers of single-crystal quartz. In use for NDT now are poled ferroelectric ceramics, such as PZT (lead zirconate titanate) or PMN (lead metaniobate).

For single-crystal quartz with proper crystal orientation, when an alternating voltage is applied across the metallized plane surfaces of the wafer, it either expands and contracts axially or exhibits particle oscillations parallel to the end planes. By bonding one of the planes of the wafer to the test piece, either longitudinal or shear waves are sent into the structure. Conversely, when an elastic wave reaches the transducer from the structure, an alternating voltage appears across the two planes, so that the transducer is also a detector of elastic waves.

With a ferroelectric wafer bonded to the test piece, longitudinal waves are generated and will enter the test piece directly. If shear waves are desired, a dielectric wedge of proper angle is inserted between the transducer wafer and the test piece so that longitudinal waves cannot enter the test piece because their critical angle is exceeded, but shear waves that are generated by the mode conversion process *will* enter.

Instead of bonding the transducer firmly to the structure, it can be attached with a "couplant," such as a liquid or gel. And for the special case of longitudinal waves, the transducer can be inserted into a liquid to generate elastic waves, which can then be coupled into the solid structure. Again, coupling is reciprocal.

5.3.2 Generation of Structure-Dependent Waves Following is a brief summary of some methods used to generate various structure-dependent waves, as discussed by a number of authors. (See, for example, Krautkrämer and Krautkrämer, 1990; Lynnworth, 1989; McGonnagle, 1966; Szilard, 1982.)

1. Rayleigh waves can be generated by interdigital electrodes on a piezoelectric substrate, by transducers of proper crystal orientation (e.g., *X*-cut in single-crystal quartz), or by conical or knife-edge transducers.
2. Love waves can be launched by a shear-wave transducer.
3. Lamb waves can be launched by interdigital electrodes on a piezoelectric substrate.
4. Stonely waves have been generated by mode conversion from Rayleigh waves (Hsieh *et al.*, 1991).

5.3.3 Geometric Ultrasonic Optics The well-known techniques of geometric optics can be used to aid in ultrasonic NDT. For example, poled bowl-shaped ceramic transducers can generate converging beams. And since Snell's law holds, beams can be focused by the use of lenses with shapes similar to their optical counterparts or by electronic focusing by proper phasing of the various segments of Fresnel-zone plate structures. By using focusing structures, it is often possible to increase sensitivity when searching for small defects or when accurately sizing defects.

5.4 Methods of NDT

Three different methods of NDT have been in use: through transmission, resonance, and pulse echo.

5.4.1 Through-Transmission NDT Flaws in the path of a wave cause reflection and therefore reduce the energy transmitted onward. The transmission technique makes use of this fact by sending elastic waves of a convenient frequency through the structure under test and comparing the energy transmitted with that obtained by sending the same signal through an identical but fault-free solid. While this technique has found use in the past, it is subject to the fact that, as in an electromagnetic transmission line system, the energy transmitted depends on the thickness of the structure; so that, depending on the frequency, small deviations in thickness of thin materials can cause inaccurate results. In some early work, this problem was overcome by the use of a frequency-modulated or noise-modulated input signal (Krautkrämer and Rudiger, 1949; Schubring, 1955).

5.4.2 Resonance NDT The resonance technique makes use of the fact that a block with smooth, parallel sides is a resonator for elastic waves at frequencies at which its thickness is an integral number of acoustic half-wavelengths. When a swept-frequency sinusoidal signal is applied to a piezoelectric transducer that is fastened to one side of the block and the transducer voltage vs frequency is monitored on an oscilloscope, the reflected electrical impedance "seen" by the transducer causes peaks to appear on the oscilloscope display at the various resonant frequencies. A flaw, such as a crack that is not perpendicular to the sides of the block, can indicate its presence by weakening or destroying the resonance peaks.

5.4.3 The Pulse-Echo Technique

While the transmission and resonance methods can indicate the presence of a flaw, they do not generally indicate its location and configuration. The ultrasonic method that solves these problems is the pulse-echo technique.

In this technique a short pulse is sent from a source of elastic waves into the region to be explored. If an echo of this pulse arrives at the location of the source Δt seconds later, the imperfection that caused the echo is a distance d from the source, where $d = v(\Delta t)/2$. For example, if Δt for an aluminum specimen is found to be 100 μs, and if the wave is a longitudinal wave, for which v in aluminum is 6.2×10^3 m/s, the flaw is 0.31 m from the source.

5.4.3.1 Generation of Short Pulses. If a 1-MHz piezoelectric transducer that is completely isolated in air is excited by a delta-function voltage pulse, it will "ring" for many cycles, i.e., it will expand and contract and thereby generate a 1-MHz decaying sine-wave voltage. However, when one face is bonded to the specimen under test, while the other face is bonded to a damping mass, the transducer becomes broadband. As a result, the acoustic pulse sent into the specimen is not a "ringing" wave but a short pulse whose length in time is approximately the reciprocal of the bandwidth of the now "loaded" transducer. It is pulses generated in this manner that perform the action of the pulse-echo technique.

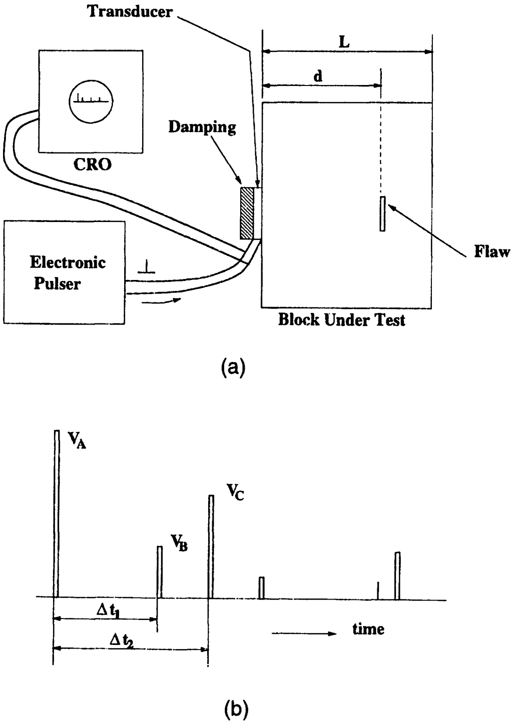

FIG. 7. Pulse-echo system. (a) Schematic. (b) Oscilloscope display.

5.4.3.2 Pulse-Echo Action. An example of the action in the pulse-echo technique is given in Fig. 7, where a transducer is fastened to the left wall of a block. As stated above, when a short voltage pulse V_A is applied to the transducer, a short acoustic pulse is sent into the block. Upon striking a flaw, a portion of the acoustic energy is reflected, thereby creating a pulse that moves back toward the transducer, which generates voltage V_B when struck by it. A second reflection, V_C, originating at the wall to the right of the flaw, is received a short time later. Additional signals, due to reflections from the left wall, the flaw, and the right wall, are also generated. These signals, which can be seen on an oscilloscope of sufficient bandwidth whose sweep is synchronized with the rate at which pulses are applied to the transducer, appear as sketched in Fig. 7(b). Here it is seen that the wave velocity v in the block shown is given by $v = 2L/\Delta t_2$, while the distance between transducer and flaw is $d = v\Delta t_1/2$.

To determine the lateral position or the extent of the flaw, the transducer can be moved along the wall, to generate displays such as Fig. 7(b) for various lateral locations.

5.4.3.3 Resolution. To find the lateral or angular location of the flaw requires that the probing wave is confined to lie in a reasonably narrow beam. Such a beam would generally be launched by a transducer of, say, diameter D, which can be considered an antenna for elastic waves and therefore follows the relation for angular beam aperture well known for electrical antennas. Specifically, the angular spread Θ in the far field for signals of wavelength λ launched by a source of diameter D is approximately $2\lambda/D$. As an example, for maximum angular spread of 0.2 radians from a source of $D = 0.01$ m, λ is found to have a maximum value of 0.001 m. For v of 6.2×10^3 m/s, this corresponds to a minimum frequency $f = v/\lambda = 6.2$ MHz. Consequently, it is the need for good resolution that is the reason for the use of frequencies of 50 kHz and higher.

The far field referred to above is the region in front of the transducer that lies at distances z larger than a distance $z_N = D^2/4\lambda$. In this region the energy density on the axis, i.e., on the center line that extends perpendicular to the transducer area, follows the inverse square law. On the other hand, in the near field, where $z < z_N$, the on-axis energy density passes through a number of maxima and minima as z increases from zero to z_N. Because of these fluctuations, it is not advisable to look for flaws in the near field of a transducer. For the values of D and λ in the example above, z_N has the value 0.025 m.

The relation of angular beam width to frequency can be seen to be readily applicable to finding imperfections in a block of material by transmission or resonance NDT. It is equally applicable to the pulse-echo technique. For by applying a short voltage pulse to a transducer designed to operate over a frequency band around, say, 1 MHz, the waves that are generated are principally those of spectral content around 1 MHz, so that the relations $\Theta = 2\lambda/D$ and $z_N = D^2/4\lambda$ still apply.

5.4.3.4 Presentation of Results. The aim of much of the development of ultrasonic NDT was to produce maps of the structure of a specimen from the data gathered by the ultrasonic test techniques. Attempts to achieve such maps included the use of chemical effects, temperature-sensitive substances, diffraction of light, orientation of microscopic particles, special phosphors, and the development of electronic tubes (McGonnagle, 1966). Ultimately, however, it has been the development of digital techniques and associated semiconductor-based devices that has resulted in ultrasonic maps of specimens. These techniques are based on digitizing and storing the data obtained by the pulse-echo method, such as that of Fig. 7, then presenting these data in map form.

It has become standard to refer to three different methods of visual presentation as A, B, and C scans, with the following meanings:

A scan: This is the display of echoes received vs time, as in Fig. 7(b). Obtained from a transducer at a specified location, it is a one-dimensional presentation that shows the existence of irregularities and their locations and relative magnitudes of echoes along a line that extends outward from the transducer into the specimen under test.

B scan: This presentation is a map of irregularities, or defects, in a cross section of the specimen. It therefore reveals the depths of such defects. An example is given in Fig. 8.

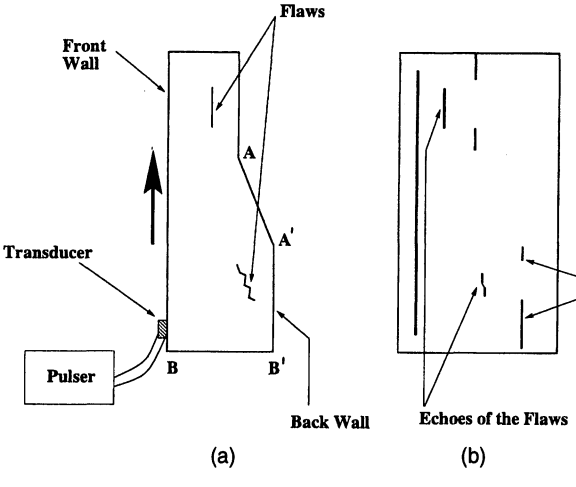

FIG. 8. B-Scan. (a) Physical arrangement. As indicated by the arrow, the transducer is moved upward along the front wall. (b) B-Scan display. For the geometry given, no echo exists for the portions of the back wall masked by the flaws. If the back wall is very smooth, the echo from portion A-A′ will be reflected to wall B-B′, where it will be reflected again, to bounce from wall to wall until it has spread or become attenuated to a level too low to be sensed by the transducer.

To produce this display, the transducer is moved along a straight line that then defines the cross section to be displayed, and time and amplitude of reflections are recorded. In a monochrome display system the intensity of the echo is indicated by the intensity of the spot or line in the display. In a multichrome display system, various colors can be used instead. An example of an actual B-scan image is seen in Plate 1.

C scan: This two-dimensional display is a map of the distribution of defects parallel to and beneath a surface of the item under test. It is generated by moving the transducer across this surface in a rasterlike manner. In any location, any echo that appears between the "transmit" pulse and the echo from the opposite side, i.e., the back wall, activates circuitry that controls the light output from a spot on a cathode-ray tube screen or a hard-copy display at the point that corresponds to the location of the transducer. The intensity of the echo determines the brightness of the spot in black-and-white display or the color in a color display. An example is seen in Plate 2.

If only defects from around a particular depth are to be shown, the display beam is turned on only around the particular time that echoes from the chosen depth are received.

5.4.4 Noncontact Coupling Although piezoelectric transducers are most commonly used in ultrasonic NDT, the requirement of mechanical contact with a test piece, either directly or through a couplant, is sometimes found to be a limitation. Examples are materials that are to be tested while at a high temperature, porous materials such as concrete that soak up liquids, and test pieces in a fast-moving production line. Noncontacting techniques for these cases can employ magnetostriction, electromagnetic coupling, electrostatic coupling, and some laser-based techniques (Birnbaum and White, 1984; Krautkrämer and Krautkrämer, 1990; Szilard, 1982).

5.4.4.1 Magnetostriction. The magnetostrictive effect in ferromagnetic materials, i.e., development of a strain upon application of a magnetic field, can be utilized by placing a small coil near a specimen and sending a high-current pulse through this coil. The result is a pulse of strain that propagates as an elastic-wave pulse within or on the surface layer of the specimen.

On the other hand, a strain in a magnetostrictive material is accompanied by a local change in permeability. Consequently, if a portion of the specimen has been placed in a region of applied magnetic-field intensity that does not change with time, the change in the permeability caused by a strain changes the magnetic-flux density, thereby generating a voltage pulse in a coil that is also in this field. The two coils therefore form a transmission/detection system of acoustic waves, similar to the system based on the use of piezoelectric transducers. Detection of flaws can be by either transmission or pulse-echo methods. Because of the high current that must be supplied to the transmitting coil, some arrangements use both a transmitting and a receiving coil. However, single-coil systems are also in use.

The magnetostrictive effect is found to increase with temperature, reaching its maximum near the Curie temperature of the ferromagnetic material. This means that for steel such a noncontacting NDT system can be used up to about 600 °C. Above this temperature the magnetostrictive effect drops rapidly to zero.

Results of tests from a particular magnetostrictive system report detection of a flaw whose area was 0.2% of the cross section of a rod, of a hole 0.8 mm in diameter and 0.5 mm deep in a 6-mm-thick plate, and of a surface scratch of depth 0.05 mm (Kaule, 1963, as reported by Szilard, 1982).

5.4.4.2 Electromagnetic Coupling. A coil carrying high-frequency currents near a metal object induces eddy currents in the surface layer of this object. If this surface layer is also immersed in a constant magnetic field, a Lorentz force $\mathbf{j} \times \mathbf{B}$ (N/m^3) exists. (Here $\mathbf{j}$ is the current density, in A/m^2; $\mathbf{B}$ is the magnetic-flux density, in T.) This force, in turn, causes elastic waves to propagate into the object. By proper orientation of the magnetic field, either longitudinal or shear waves, or both, can be generated. Moreover, the effect is reciprocal: If an acoustic wave reaches this surface layer, the particle oscillations are essentially the elements of a conductor moving in a constant magnetic field and will therefore generate voltages that cause circulating currents to flow. These will couple inductively to the external coil, which now is a detector of acoustic waves.

Using electromagnetic coupling to generate longitudinal acoustic waves of 0.6 to 1.2 MHz, a system that can detect spherical voids of 2–3 mm diameter in oblong bars at temperatures up to 1200 °C has been reported (Whittington, 1982, as reported by Szilard, 1982).

For similar pulse power and receiver gain, the sensitivity of electromagnetic systems has been reported to be about 60 dB below that of piezoelectric systems. However, by using a strong magnetic field and higher pulse power, it has been found possible to raise the received signal amplitudes and signal-to-noise ratios to levels similar to those of piezoelectric flaw-detecting systems (Szilard, 1982).

5.4.4.3 Electrostatic Coupling. If a voltage V is applied across the two plates of a parallel-plate capacitor separated by a gap of width d, an attractive force F given by $F = V^2\epsilon_0\epsilon_r A/2d$ exists. Here ϵ_r is the dielectric constant of the medium in the gap, ϵ_0 is the permittivity of free space, and A is the area of one side of one of the plates. If one of the plates is actually the side of a metallic block or the flat end of a rod, the application of a time-varying voltage will generate a time-varying force, which will generate elastic waves in the block or rod. If V varies sinusoidally in time with frequency f, the force will vary at frequency $2f$ and will have a dc component. However, by impressing a dc voltage of magnitude much larger than the time-varying voltage, the propagating wave will principally be of frequency f.

Conversely, since the voltage V across the plates is given by $V = Qd/A\epsilon_0\epsilon_r$, where Q is the charge on the positive plate, if Q is kept constant, changes in d will be reflected as changes in V. Q can be kept virtually constant by connecting the capacitor to a voltage source in series with a large resistance. Consequently, if one of the plates is the side of a metallic block or the end of a metallic rod, incident acoustic waves will generate a voltage, so that both generation and detection of acoustic waves are possible with this noncontacting scheme.

The literature describes a number of ways that this method has been used to perform NDT. While the sensitivity has been found to be about 30 dB below that of systems using piezoelectric transducers, it does have the feature of being noncontacting and broadband (10 kHz–10 MHz) (Szilard, 1982).

5.4.4.4 Laser-Generated Acoustic Waves. When a surface is heated from an external source, the energy entering the surface will spread into the body. In most cases the rise in temperature will cause the body to expand. If a short pulse of high heat is applied to a small spot on the surface and this heat cannot diffuse away rapidly enough, the expanding local volume adjacent to the heated spot will generate a stress pulse in the material and become the source of an elastic-wave pulse. As in the case of acoustic pulses generated by other means, when this pulse strikes a flaw, a portion of it is reflected. If sensed properly, the echo pulse will indicate the presence and location of the flaw.

To carry out this method of acoustic-pulse generation therefore requires a pulse that heats the surface very rapidly and with high energy density. This can be accomplished with a Q-switched laser, with beams of power density in range of 10^6 W/cm^2 or higher. Impacting a surface with such beams can generate longitudinal, transverse, and surface waves (Aindow *et al.*, 1981). Focusing and scanning of the ultrasonic energy by arrays produced by the use of multiple or scanned optical beams has also been demonstrated (Addison *et al.*, 1987). To prevent damage to the surface, it can be covered with a membrane, with a thin layer of liquid interposed.

To be a useful system, the transducer receiving the echo pulse must also be noncontacting. This can be accomplished by using reflections of the light of an auxiliary low-power laser from the surface of the test piece in an arrangement using optical interferometric techniques (Monchalin and Heon, 1986).

The cost of such an NDT system is considerably above that of methods described earlier. However, since it is based on the use of optical beams, it has the advantage of permitting nondestructive testing of a specimen from distances of several meters.

6. EDDY-CURRENT TESTING

The term "eddy currents" is used to refer to circulating flows of fluids and to circulating electric currents in conductors. The latter exist whenever time-varying magnetic fields penetrate surfaces of electrical conductors. In most situations efforts are made to minimize these electric eddy currents because they result in energy losses. However, they have found extensive use in the nondestructive testing of metallic components and structures. Applications include the detection of cracks, voids, and inclusions; sorting of components in accordance with their electric or magnetic properties; measurement of the thickness of metallic plates, tube walls, and coatings; precise dimensional measurements of symmetric parts such as ball bearings; and the detection and location of hidden metal objects.

An advantage of eddy-current testing over some other methods is that no physical contact is required. It can therefore be readily used to carry out automatic inspection of newly fabricated metallic tubing and other parts. As a result, the total number of NDT procedures using eddy-current inspection probably exceeds all other types.

A number of individuals have contributed to the development of the technique. The most prominent of these, Dr. Friedrich Foerster, who carried out this work in the period 1950 to 1965, has been cited as "the father of modern eddy current testing" (McIntire and McMaster, 1986).

6.1 Frequency-Domain (Sinusoidal) Eddy-Current Testing

6.1.1 Principle of Operation Eddy currents can be generated in electrically conducting materials by the flow of time-varying desired electric currents because

1. in accordance with Ampère's law, an electric current is always accompanied by a magnetic flux Φ adjacent to or surrounding it;
2. by Faraday's law, $V = -d\Phi/dt$, this time-varying flux causes a voltage V to exist around any closed path through which the flux passes.

Consider the arrangement of Fig. 9, in which a coil of wire is connected to a source that causes a sinusoidally time-varying current to flow through it. This current is accompanied by a sinusoidally time-varying magnetic flux that passes axially through the coil. If a loop of wire is located below the coil, as shown, some of the magnetic flux passes through this loop and generates a sinusoidally time-varying voltage. With the loop open-circuited, this voltage appears across the ends. If the two ends of the loop are connected to each other, on the other hand, a sinusoidally varying current flows in the loop, of magnitude and phase determined by its resistance and inductance and the frequency of the sinusoid. This current, in turn, generates magnetic flux that alters the net magnetic flux in the loop and in the coil. Sensing the amplitude and phase of the net magnetic flux passing through the coil, or of a quantity related to this flux, such as the coil voltage, provides an indication of the current flowing in the loop and thereby of its resistance and inductance.

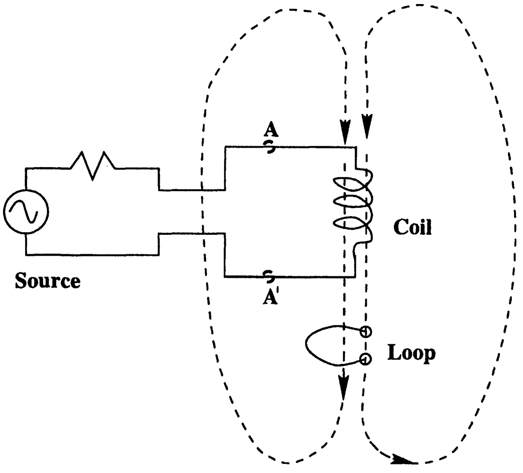

FIG. 9. Sinusoidally varying current in coil generates voltage across the ends of the loop.

An indirect method of sensing is the measurement of the input impedance Z_{in} seen when "looking" into the coil. This is because the coil/loop system can be represented by the simple equivalent circuit for a magnetically coupled linear system of Fig. 10, where R_1 and R_2, and L_1 and L_2, are the resistances and self-inductances of coil and loop, respectively. M, the "mutual inductance," is determined by the fraction of magnetic flux generated by the coil that also passes through the loop.

Using standard phasor analysis, in which all currents and voltages are assumed to vary as exp $(j\omega t)$, where $j = \sqrt{-1}$ and ω is the an-

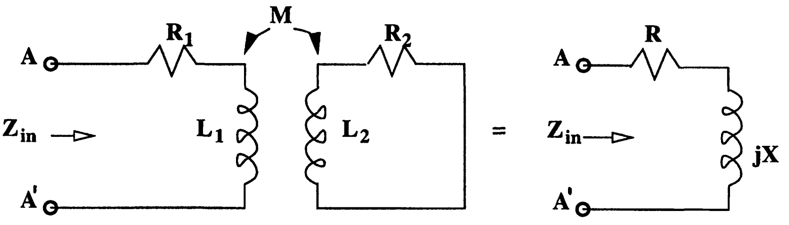

FIG. 10. Equivalent circuit of magnetically coupled coil–loop system.

gular frequency of these quantities, the complex input impedance Z_{in} at this frequency is found to be

$$\begin{aligned} Z_{\text{in}} &= R_1 + j\omega L_1 + \omega^2 M^2/(R_2 + j\omega L_2) \\ &= [R_1 + \omega^2 M^2 R_2/(R_2^2 + \omega^2 L_2^2)] \\ &\quad + j\omega[L_1 - \omega^2 M^2 L_2/(R_2^2 + \omega^2 L_2^2)] \\ &\equiv R + jX. \end{aligned} \tag{1}$$

As seen, both the real part R and imaginary part X of Z_{in} or, expressed alternatively, both magnitude and phase of Z_{in}, depend on R_2, L_2, and the magnetic coupling, expressed by mutual inductance M.

Since M depends on the separation between coil and loop, in eddy-current NDT it is often referred to as "lift-off."

Plots of the locus of Z_{in} vs R_2 in the complex impedance plane for three different values of M and continuously varying values of R_2 are given in Fig. 11. The top point of this figure for each curve corresponds to the case of infinite resistance of the loop, i.e., to the open loop. The bottom point for each curve is the loop of zero resistance, i.e., $R_2 = 0$. For $M = \sqrt{L_1 L_2}$, all of the magnetic flux of the coil passes through the loop.

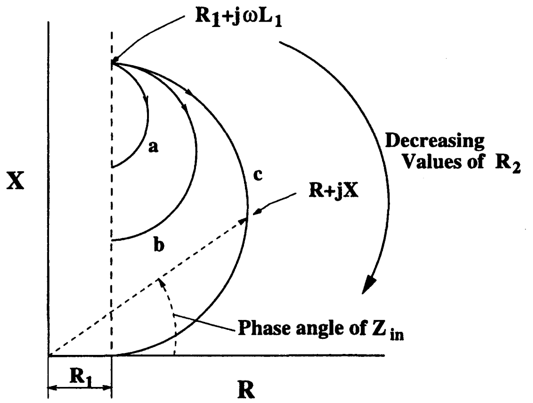

FIG. 11. Plot of Z_{in} vs R_2 in complex plane for three different values of the fraction of magnetic flux that passes through both coil and loop, represented by mutual inductance M: (a) $M \ll (L_1 L_2)^{1/2}$; (b) intermediate value of M; (c) $M = (L_1 L_2)^{1/2}$.

The measurement of R and X or of the magnitude and angle of this complex quantity, or of deviations of R and X from the corresponding values of a "standard," are the bases of sinusoidal eddy-current testing. In this NDT procedure the "loop" referred to above is actually an equivalent volume of a ring of metal through which magnetic flux will pass when a coil is placed adjacent to it and connected to a sinusoidal voltage source of angular frequency ω. Examples are seen in Fig. 12.

In Fig. 12(a) a small test coil, generally called "probe coil," is placed above a metallic block and connected to a sinusoidally varying voltage source. The current flowing in this coil generates a sinusoidally varying magnetic flux that passes into the block and induces eddy currents in the block in an "equivalent loop." The input impedance Z_{in} seen when looking into the test coil depends on the dimensions and conductivity of the volume and material of this "equivalent loop."

We now consider what will happen to Z_{in} when an operator moves the probe coil parallel to the surface of the block, with constant spacing above the surface. If the block is of uniform conductivity and there are no internal discontinuities, both the real and imaginary components of Z_{in} will remain the same, except when the probe coil is near an edge. Should the coil be passed over sections of the block of varying conductivity, however, the resistance of the "effective loop" will change, therefore changing both resistive and reactive components of Z_{in}. Similarly, if the test coil is passed over a section

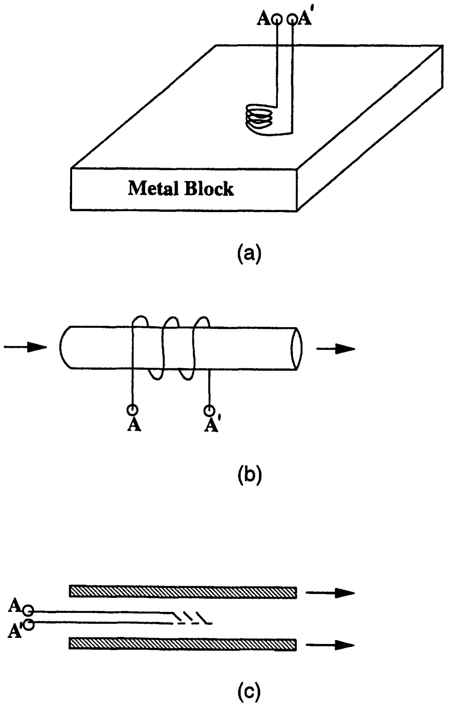

FIG. 12. Methods of carrying out eddy-current NDT: (a) Probe coil adjacent to a metal block. (b) Coil surrounding rod or tube. (c) Coil inside a metal tube.

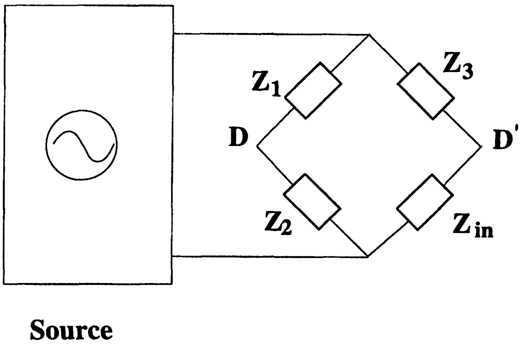

FIG. 13. Basic circuit for measurement of Z_{in} of probe coil.

of the block that contains a crack lying transverse to the direction in which eddy currents would normally flow, the effective path length for eddy currents is increased, again changing the value of Z_{in}. Continuous monitoring of Z_{in} will therefore alert the operator to a change in the material below the test coil.

The same considerations apply to the structure of Fig. 12(b), where a metal rod is moved through a test coil. Changes in the conductivity of the rod, the presence of cracks that impede the flow of circumferential currents, or changes in the shape or diameter will all be reflected into changes in Z_{in}.

A basic circuit that can be used to identify these changes is a bridge circuit, as that of Fig. 13, where the values of Z_1, Z_2, and Z_3 have been chosen so that the bridge is balanced and voltage $V_{DD'}$ is zero when Z_{in} has what can be considered the design center, or standard, value. The amplitude and phase of unbalance voltage $V_{DD'}$ can be used to identify the type of deviation, i.e., is it a change in conductivity or a change in diameter. Modern test apparatus essentially performs this basic function, but with added processing capabilities and such enhancements as memory, display of results, and, possibly, automated inspection of products flowing from a production line.

Another method of sensing changes is by the use of another, or secondary, coil, through which some of the magnetic flux is allowed to pass. For a thin sheet this secondary coil can be below the sheet. In both cases the desired information about the specimen can be extracted by measurement of amplitude and phase of the open-circuit voltage of this secondary coil.

A third method of sensing is by the use of one or more semiconductor probes, which measure the magnetic field by use of the Hall effect.

6.1.2 Depth of Penetration An important point in these measurements relates to the volume effectively probed by the magnetic field. The two aspects that determine this volume are a geometric factor and the skin effect.

6.1.2.1 Geometric Factor. One aspect of the geometric factor is evident in Fig. 12(a), from which it is rather obvious to see that for close spacing of coil and block the magnetic field from the test coil has appreciable magnitude principally in the area whose diameter is not much larger than the coil diameter. If the coil is some distance above the block, the lift-off effect spreads the area but reduces the intensity of the magnetic field within the block. This lift-off effect can be used to monitor the separation between test coil and block. For the case of a nonconducting layer on a metal block, it can therefore be used to measure the thickness of this layer.

A geometric effect is also found in Fig. 12(b), where induced currents flowing at the smaller radii of a rod under test are found to contribute less to influencing the input impedance or the voltage of a secondary coil than induced currents at the larger radii. This effect, which is relatively frequency independent, has been called the "coil shape effect" (Libby, 1971).

6.1.2.2 Skin Effect. The skin effect is a well-known effect that limits the penetration of electromagnetic fields into a good conductor. Specifically, if a uniform tangential electric field of form $E_0 e^{j\omega t}$ exists at surface $y = 0$ of a material that extends from $y = 0$ upward toward infinity, the electric field at any value of $y > 0$ is $E_0 e^{-y/\delta} e^{-jy/\delta} e^{j\omega t}$. Here, in SI units, $\delta = (2/\omega\sigma\mu_0\mu_r)^{0.5}$, where σ is the conductivity, $\mu_0 = 4\pi \times 10^{-7}$ H/m, and μ_r is the relative permeability. The quantity δ is generally called the "skin depth."

As an example, for copper at 1000 Hz, with $\sigma = 5.80 \times 10^7\ (\Omega\text{-m})^{-1}$, δ is found to be 2.1 mm. Consequently, if Z_{in} were measured at 1000 Hz while the test coil is moved along the surface of a copper block, flaws or changes in conductivity in the block at depths greater than a few millimeters would not change Z_{in} or the open-circuit voltage of a secondary coil significantly. Similarly, flaws that lie within the material of Fig. 12(b) a few skin depths below the cylindrical surface will not affect the value of Z_{in} and would therefore not be detected.

The dependence of depth of penetration on frequency allows tests to be performed that delineate changes in conductivity or the presence of flaws as a function of depth into a material. Thus, for example, a test frequency of 64 kHz will generally measure relative conductivity in nonmagnetic metal strips or slabs of thickness 3 mm or greater. To determine the presence of flaws or changes in conductivity for depths greater than 3 mm requires testing with lower frequencies. In practice, the test frequencies of commercial units vary from 5 Hz to 10 MHz.

The testing of conducting ferromagnetic materials of high relative permeability requires special consideration, since the skin depth varies as $\mu_r^{-0.5}$. For these materials, frequencies in the high kilohertz region restrict eddy currents to exist in a very shallow layer near the surface. This permits the detection of seams and laps or service fatigue cracks. If deeper penetration is desired, frequencies as low as 5 Hz can be used.

An alternative method for testing magnetically active conductors, such as steel, is to apply a strong dc magnetic field that saturates the material magnetically, thereby reducing the relative permeability μ_r close to unity for small ac signals.

6.1.3 The Importance of Phase This section discusses the changes in Z_{in} that are sensed as a test coil is scanned along a bar or cylinder that has deviations from a "standard" because of one or more of several properties. The differences in the changes are seen by looking at changes in Z_{in} due to lift-off and to changes in conductivity for Z_{in} of a test coil that surrounds a cylindrical nonmagnetic rod.

A number of individuals have performed calculations of the input impedance Z_{in} for a test coil for such a system. In applying these results to NDT, the results are often normalized by first subtracting the coil resistance, then dividing the values of reactance and resistance by the reactance of the empty coil, to arrive at a quantity $Z'_{\text{in}} = R' + jX'$. The general shape of the locus of Z'_{in} as the conductivity of the rod varies from zero to infinity is given by the curves of Fig. 14 (Libby, 1971). Curve (a) is for a tight-fitting coil. For curve (b) the rod diameter is smaller than that of the coil. As in the case of the loop, an increase in conductivity corresponds to clockwise progression along the curves. Points P, Q, and T relate to changes in the phase angle of Z'_{in}, i.e., $\tan^{-1}(X'/R')$, as a long rod is passed through the test coil.

Let us now assume that Z'_{in} for the section of rod of a certain conductivity σ_0 with tight-fitting test coil is given by the directed line O-P. If the next section of rod has the same diameter but has lower conductivity or contains a flaw, the change in Z'_{in} is given by directed line P-Q. If, on the other hand, the conductivity is the same but the diameter of the rod has decreased, the change in Z'_{in} is given by directed line P-T. Consequently, by monitoring the change in phase of Z'_{in}, it is possible to tell whether the change in Z'_{in} is

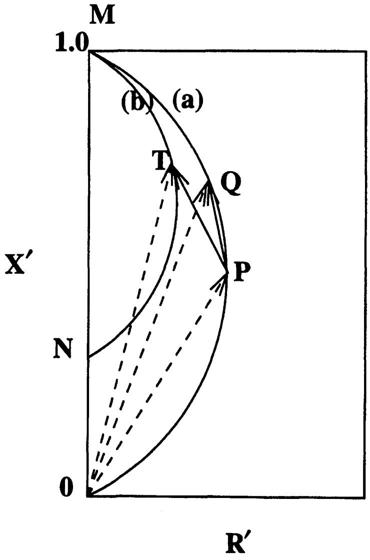

FIG. 14. General shape of locus of normalized complex impedance Z'_{in} for test coil surrounding a metal nonmagnetic rod, as conductivity of rod varies from zero (point M) to infinity (points O and N). Curve (a) is for a tight-fitting coil. For curve (b) the rod diameter is much smaller than the coil diameter. For the significance of points P, Q, and T, please see the text.

due to a change in separation of test coil from specimen, i.e., lift-off, or the presence of a flaw or change in conductivity.

While the above has dealt with the case of a rod within a coil, the same principles apply also to other geometries.

6.1.4 Instruments Over the years, the development of sinusoidal eddy-current NDT has reached a high level of sophistication. For example, an eddy-current probe in use since 1970, which is composed of two orthogonal windings, virtually ignores variations in material conductivity, lift-off, and permeability, to concentrate strictly on flaw detection. This probe is reported to detect surface flaws in magnetic materials that are only 0.13 mm deep (Kenton, 1970).

Instruments commercially available now combine the phase analysis discussed above with memory, computer, and display techniques. Examples of the results of the use of such an instrument are seen in Fig. 15. This figure is an oscilloscope display of the stored output as a probing coil is scanned along the surface of a metal plate for two cases. The nearly horizontal line shows the output as the probe that virtually ignores changes in lift-off is scanned along a metal strip at different spacings. Superimposed on this is another scan that was obtained as the same probe was scanned over a region that contained a flaw.

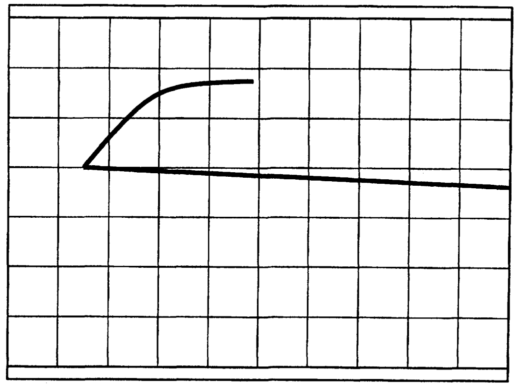

FIG. 15. Oscilloscope display from an instrument used to carry out sinusoidal eddy-current NDT. (Courtesy of Centurion NDT, Inc., Schaumburg, IL.)

6.1.5 Multiple-Frequency Testing Although most sinusoidal eddy-current measurements are conducted at a single frequency, in some cases several frequencies can be used. Specifically, conductivity could be measured by use of a high-frequency source, with only shallow penetration, while a low-frequency source would search for flaws at greater depths within the material. In another test, results with one frequency could determine the lift-off spacing, while another frequency would probe the specimen. The possibility of simultaneous monitoring with several frequencies to determine several parameters, such as the thicknesses of layers of different metals in a sandwich type of structure, or to separate interfering signals from portions of a structure that would mask desired signals, has also been considered.

The use of a number of frequencies for NDT of the multilayer structure is based on the fact that such a structure, for nonmagnetic materials, can be represented by a linear electric ladder circuit. For such a circuit the input impedance is a unique function of frequency. Consequently, by performing measurements of amplitude and phase at several frequencies and inserting the measured data into a set of simultaneous linearly independent algebraic equations, the desired parameters can be extracted (Libby, 1971).

6.2 Time-Domain (Pulsed) Eddy-Current Testing

While the sinusoidal method is the commonly accepted technique of eddy-current testing, there are cases in which time-domain methods have been found more appropriate. Examples are a noncontacting technique of determining the resistivity of a metal specimen, the measurement of the thicknesses of layers in a multilayer metal structure, and the search for irregularities in the fuel/cladding interface of nuclear fuel elements. These techniques are based on the fact that for good conductors such as metals, for which displacement currents are negligible when compared to Ohmic currents, the equation that governs the behavior of magnetic-flux density **B**, expressed in SI rationalized units, is

$$\frac{\partial \mathbf{B}}{\partial t} = \frac{\rho}{\mu_r \mu_0} \nabla^2 \mathbf{B}, \qquad (2)$$

where ρ is the resistivity, μ_r the relative permeability, and $\mu_0 = 4\pi \times 10^{-7}$ H/m. Solutions to this equation reveal that magnetic-flux density propagates in a good conductor by a diffusionlike process, with the velocity of propagation proportional to $t^{-0.5}$, where t is the time since the initiation of a nonequilibrium situation. Since magnetic-flux density and currents are intimately tied together, a similar result of diffusionlike propagation is obtained for the transient behavior of electric current in a good conductor.

For metal structures with thicknesses of a fraction of a millimeter to several millimeters, the effective propagation times are from a few microseconds to milliseconds. Transient events of such duration are readily observed in the laboratory.

The following sections describe some methods involving eddy currents that utilize the behavior that is governed by the above equation.

6.2.1 Measurement of Resistivity In this procedure, a cylinder of the material whose resistivity is to be measured is surrounded locally with a small "pickup" coil and then inserted into a larger "magnetizing" solenoid. A step function of current is sent through this solenoid, to produce a suddenly applied magnetic-field intensity H_0. The resulting time-varying magnetic-flux density generates a voltage in the pickup coil that eventually decays exponentially as $\exp(-t/T_R)$, where $T_R = K/\rho$, and K is a calculable constant determined by the cross-sectional dimensions of the cylinder. Resistivity ρ is found from the value of T_R extracted from the exponential decay curve (Bean *et al.*, 1959).

In a preferred method of carrying out this test, a steady magnetic field is established in the solenoid by the applied current, which is then suddenly interrupted. Although the polarity of the coil voltage is now reversed, the rate of the exponential decay remains the same.

Since the pickup coil is mainly sensitive to the material it encloses, average local resistivity may be measured by moving the pickup coil along the axis of the cylinder.

6.2.2 Pulsed Eddy-Current NDT The development of pulsed eddy-current NDT in the United States received its start in the 1950s (Waidelich, 1956). The objective of this initial work was to determine the thickness of a layer of a metal A that was cladding to a thick plate of a metal B. Magnetic-field pulses were sent into this cladding through the application of current pulses of high intensity to a small single-layer coil that was adjacent to the cladding and had its axis perpendicular to the plane of the cladding.

Had this coil been far removed from the metal, the application of a step function of current would have produced a high spike of coil voltage. Instead, the magnetic field produced by the coil current and coupled into the metal caused eddy currents to be generated in the layer of metal adjacent to the coil, thereby replacing the spike with a more gently rising voltage.

Now, with a ring of current density and its magnetic field in the top layer of the conductor, this current density and associated magnetic field would follow the diffusionlike action predicted by Eq. (2) and move toward the base metal, reaching it after time interval $t = d^2\mu\sigma$, where μ is the permeability, σ is the conductivity, and d is the thickness of the cladding.

Since the base metal had a different conductivity than the cladding, a reflection occurred, sending the current density and associated magnetic field back toward the front surface. The result was a voltage pulse across the coil that rose rather steeply, then decayed gradually and became negative, having passed through zero at time T given by $T = 2d^2\mu\sigma$.

While in the above discussion a step function of current and therefore of magnetic-field intensity H was assumed, Waidelich demonstrated that pulses of less abrupt rises could also be employed.

As demonstrated in this work, the basic idea of this method of eddy-current testing is that the duration of the "output" pulse voltage seen across the exciting coil is altered by changes of the conductivity or permeability encountered by the diffusing current and magnetic-flux densities. These changes could be due to actual changes in the metal, as in the change from cladding to base metal, or the presence of a crack or other irregularity.

When another coil is placed on the opposite side of a metal sheet, it is found that the signal transmitted through the sheet is attenuated by amounts that depend on the thickness of the sheet and the pulse length. In general, the thicker the sheet and the shorter the pulse, the higher the attenuation. Researchers have made use of these facts in efforts aimed at improving the spatial resolution while scanning an area over that attainable by an open exciting coil by enclosing this coil in a metal box that has a small aperture. Since virtually no magnetic flux passes through the metal walls of the box, the area probed is limited to the area of the aperture.

In another improvement a separate sensing coil is added and placed so that there is no direct coupling from the aperture to this sensing coil but that fields emanating from the metal are sensed. For other applications this sensing coil is placed on the other side of a sheet of metal that is to be scanned, so that anomalies in the transmission over that from a "standard" can be detected.

In adapting pulsed eddy-current testing to systems in which component parts move past an inspection station, systems have been developed that make use of the fact that the time that an anomaly in the "output" signal is observed depends on the depth of the defect in the metal. By electronically sampling the output signal at several distinct points of time and comparing these sampled signal intensities with those from a "standard," an alarm signal alerts the operator that an anomaly exists. By noting the point of time of sampling corresponding to the alarm, the approximate location of the anomaly is found.

7. ACOUSTIC EMISSION TESTING (AET)

7.1 AET—An Active Type of NDT

7.1.1 Acoustic Emission The use of acoustic emission, AE, also called stress-wave emission, stress-wave microseism, microseismic activity, and rock noise, as a method of nondestructive testing is essentially an adaptation to components and structures of the techniques of detecting, source locating, measuring the intensity, and assaying the damage associated with earthquakes. It is based on the observation that the rapid release of localized stored elastic energy during irreversible structure changes in materials subjected to mechanical stress generates transient acoustic waves that propagate through the material. Such release of potential energy occurs during the formation or extension of a crack or by the motions associated with dislocations, slip, twinning, grain-boundary sliding, and fracture and decohesion of inclusions. Composites are also found to emit such transient waves during debonding, cracking, and the fracture of fibers.

Since AET causes permanent modifications, one might argue that it should not be considered a method of nondestructive testing. Nevertheless, it has been accepted as one of the standard NDT techniques.

7.1.2 Development of AET Although there are some earlier reports of the emission of sonic or ultrasonic noise from materials under stress, AET had its origin in the pioneering work of Josef Kaiser, carried out shortly after World War II. In his first journal article on the subject (Kaiser, 1953) he discussed applying tensile stresses to various metals and detecting sonic or ultrasonic pulses as these stresses were applied in increasing amounts. He recognized that the detection of these pulses could be a big step forward in materials testing, which heretofore had been performed by destructive methods. His work sparked further research, thereby resulting in the foundation of AET (Schofield *et al.*, 1958).

7.1.3 AET and Other NDT Techniques By sensing changes that occur while a component or structure is loaded, AET has become a method of monitoring structural integrity, characterizing material behavior, and discovering the beginning of failures in components or structures. It is ultimately a method aimed at arriving at allowable safe limits of operation. In addition it can be used to find regions in a structure that exhibit weaknesses, so that these regions can be examined in more detail by other NDT techniques.

The equipment used to detect acoustic emission will also detect transient acoustic waves that arise from other sources, such as from friction associated with moving parts or from leaks or cavitation. The differing characteristics of signals from these various sources, as well as the extensive use of computers in the analysis of AET, has made it possible in many cases to distinguish between signals from stress-caused emissions from those of these *secondary* or *pseudo* sources.

Most other NDT methods are static methods. Their mission is to find and characterize irregularities or flaws that already exist at the initiation of a testing procedure, without creating any changes in the test piece. AET, a dynamic method, can locate some flaws already in existence, but its principal use is to detect changes in the test piece as it is stressed, with the aim of finding limits to the stresses that can be safely imposed. Often some of these changes occur because of flaws already in existence, such as in the enlargement of a crack.

AET also differs from other types of NDT in the method of finding flaws. In other NDT procedures a detailed search for flaws is carried out by observing the signals that are the response as a volume is scanned in some manner with a small amount of *injected* energy. In AET no scanning is necessary. Since the observed signals are elastic wave pulses created totally within the test piece, these sources can, in principle, be found by triangulation, based on the different times of signal arrival at several conveniently placed stationary transducers.

AET is frequently carried out by gradually increasing the applied stress up to some maximum, or "proof," value. If the results are satisfactory for repeated stresses up to this proof value, the unit under test is expected to function well in actual operation.

AET may also be carried out over long periods of time by the monitoring of elastic-wave signals that are generated while the unit is in service. Such continuous monitoring can reveal the formation and enlargement of fatigue or corrosion cracks and their effects on lifetime without interrupting the operation of the unit—a procedure not generally feasible with other methods of NDT.

An additional significant difference between AET and some other NDT methods is that a major aim of the latter generally is to find geometric data relating to irregularities, such as location, size, shape, and orientation. While spatial information is also an important part of the data obtained from AET, an even more important aspect is time. Specifically, what is detected by AET sensors is a large number of pulses of variable shape, amplitude, and time duration.

7.1.4 Signals Detected Two types of signals are detected. Continuous emission is characterized by emission "events" of comparable amplitude in which the average time interval between the beginning of consecutive emission events is less than or comparable to the duration of each of these events. This is the case of signals that are found to occur in the plastic deformation during and after yield of a material, when many low-level pulses interfere and result in an output signal that has the appearance of white noise. Such signals have been used in the study of the microstructure of materials. (Heiple and Carpenter, 1987). Of greater interest for practical AET is emission in individual bursts, associated with the formation or enlargement of cracks. To evaluate the effect of such burst emission requires rapid sensing, storing, and processing of a large amount of data. As a result, computer tie-in has become an essential part of AET.

7.1.5 The Kaiser Effect During Kaiser's exploratory work he discovered that when his specimens were stressed in increasing steps, he was able to detect AE. But when the stress was removed and then reapplied, no emission occurred until the maximum stress that had been applied earlier was exceeded. This behavior, which is now known as the Kaiser effect, was later found to exist only in certain materials and gave rise to additional research.

In what is now known as the Dunegan corollary (Dunegan *et al.*, 1970), it is now believed that for some materials emission below the previously maximum load indicates that the material had incurred some damage because the presence of a crack would increase the local stress for the same applied load.

7.2 Measurement Techniques

7.2.1 Basic Measurement Technique Measurement systems for AET consist of acoustic transducers followed by high-gain amplifiers, frequency filters, amplitude discriminators, counters, and computers to interpret the data and display results. The transducers generally are piezoelectric wafers of a thickness chosen to cause the transducer to be resonant at a specified frequency. They are attached to the test piece with an acoustic couplant, such as a thin layer of grease. If the structure to be tested is in an environment in which the transducer can become nonfunctioning, such as at high temperatures, an acoustic waveguide can be used to isolate the transducer from the structure. Such a waveguide can be a rod up to several meters in length that has one end firmly attached to the specimen under test and the transducer mounted on the opposite end.

Acoustic-emission signals can have appreciable spectral content from low audio frequencies to the megahertz range. The range from 100 to 300 kHz has been found to be optimum for detection. This avoids low-frequency noise from nearby machinery as well as high-frequency signals that could be misleading because of acoustic attenuation at high frequencies.

7.2.2 Continuous Emission Continuous emission signals, mentioned above, have the appearance of white noise. The signals emitted by the pseudo sources of leaks in a pressure vessel and cavitation also fall in this category. Because of the continuous nature of this type of emission, an accepted standard method of quantifying the amount of continuous emission is the rms value of the output voltage of a standardized transducer.

7.2.3 Burst Emission The two items of interest for burst emission are to find the locations of emission sources and to obtain data that will permit an assessment of the degree of integrity of the structure under test.

As stated above, in principle the location of a source can be found by the use of differences in arrival time of the same signal at several transducers. For example, for a source in a plane the use of three transducers permits the source of emission to be found at the intersection of hyperbolas whose equations contain coordinates of these transducers, the acoustic-wave velocity, and the differences in arrival time of a pulse at these transducers. While for arbitrary shapes the situation is not as straightforward, the locations of sources of burst emission can be generally found by judicious placement of several transducers.

7.2.4 Interpretation of Data Most important in the use of AET is the interpretation of experimental data. Here several quantities that provide a good measure of predictability have evolved. The first and oldest of these is the "number of counts," also called "ringdown count" or simply the "count," which is derived from the output voltage of a resonant transducer that monitors burst emission. Because transducers are chosen to be resonant systems, the voltage output for an event, i.e., when it is struck by a burst emission pulse, is a wave train, somewhat as in Fig. 16(a). To convert this into a signal that provides information about the condition of the test piece, the amplitude of this time-varying voltage wave is compared to a "threshold" voltage. Whenever the peak of any cycle of this wave exceeds this threshold voltage, the system generates a digital pulse, as seen in Figs. 16(b) and 16(c). The number of counts for this emission sig-

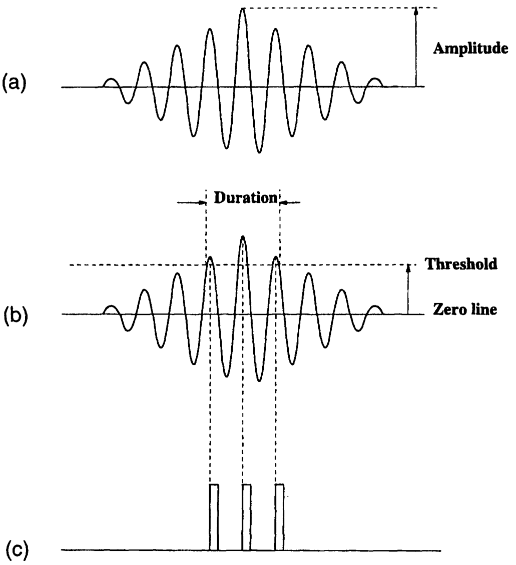

FIG. 16. Formation of counts: (a) Representative output voltage when resonant transducer is struck by an acoustic-emission pulse. (b) Threshold voltage used to formulate the "count." Also shown is the "duration." (c) Digital pulses formed by circuitry, as based on positive peak exceeding the threshold voltage. In this example the count is three.

nal is the number of pulses generated in this manner.

Since the number of counts depends on the arbitrarily set threshold voltage as well as on the acoustic properties of the item under test, the use of this number is clearly very subjective. Nevertheless, it has been found to be a satisfactory way of determining the integrity of structures under test. One way of using it is to plot the total number of counts accumulated, which is generally in the thousands, as a function of applied load while the load is gradually increased in steps. Experience has shown that as the load on a structure is increased in this manner, the count also increases. If the increase of the count is gradual with increase in the loading, up to and above (by a safety factor) the loading used during operation, the structure can be considered sound. On the other hand, if the count shows a steep increase during the test, it is quite probable that a microscopic crack that has been formed is increasing in size, thereby increasing the strain at its ends and eventually causing the crack to become large enough to cause the structure to fail.

A plot of accumulated counts as a function of load (or time, for a load that increases with time) is only one type of display that depicts the behavior of a structure. Among other displays in use are a histogram whose shape is essentially the derivative of the plot of accumulated counts, or similar plots in which the number of counts of an event is replaced by "energy counts." The latter is a number proportional to the area above the horizontal centerline of a wave train such as that of Fig. 16(a). This quantity, also known as MARSE, the acronym of "measured area under the rectified signal envelope," is sensitive to both amplitude and wave-train duration. Since there is no threshold level and MARSE is less frequency dependent, it is now preferred over the use of counts (Pollock, 1989).

Among other quantities used in the interpretation of AET data are duration and amplitude relating to an event. For example, duration, defined as the time interval between first and last threshold crossing, has been used to recognize special long-duration deleterious processes, such as delamination in composite materials.

The amplitude of an AE event, defined as the highest voltage reached by a transducer output signal for an event, is found to vary in the same test from tens of microvolts to volts. Because of this wide range, some AET instrumentation systems use logarithmic amplifiers.

A cumulative amplitude distribution displays the number of events whose amplitudes exceed voltages $A_1, A_2, \ldots, A_n$ as the load on a specimen is increased in steps. A simplified example illustrating this amplitude distribution is given in Fig. 17.

As they appear during a test, displays of the type of Fig. 17 present a virtually real-time view of the progress of the AE test. It has been found that for some metals the appearance of the beginning of large amplitudes is a sign of the enlargement of cracks. Similarly, the appearance of large-amplitude signals during progressive loading of fiberglass has revealed and will tell the observer the stress levels at which fiber breakage is initiated (Pollock, 1981).

In summary, thanks to the use of computers that are an integral part of the complete AE test setup, much AET information is now

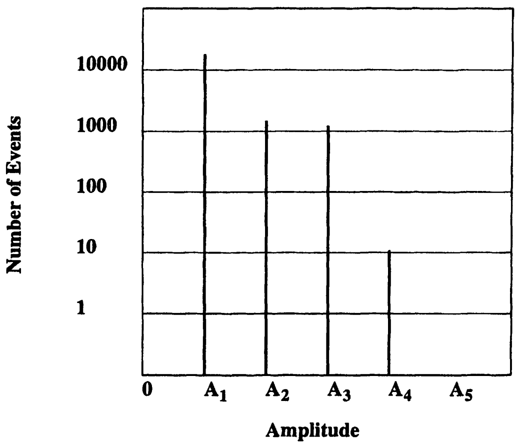

FIG. 17. Cumulative amplitude distribution of pulse heights. As shown here, 10 events have amplitudes exceeding A_4; 1010 have amplitudes exceeding A_3; 1110 have amplitudes exceeding A_2; and 11 110 have amplitudes exceeding A_1.

available for virtually instant monitoring and interpretation.

7.3 Applications

The effectiveness of AET has been well demonstrated in tests such as those dealing with the integrity of fiberglass storage tanks and pressure vessels (Pollock, 1989). In the 1970s the chemical industry was experiencing failures of fiberglass storage vessels. After the practice of AET of storage vessels was instituted by one manufacturer, the tank failure problem was eliminated. Another example of AET is its use during welding operations. Computer algorithms now exist that recognize various characteristics of acoustic-emission signals during the process of resistance welding and can control the welding current in a preprogrammed manner. AET has also been used for in-flight monitoring of aircraft (Scott, 1991).

In general, there are several features that have made AET attractive:

1. It can often be carried out while the unit tested is in normal operation.
2. Access to the whole structure is not required.
3. AET can highlight regions that are potential trouble spots, so that these can be examined in detail by other NDT methods.
4. AET responds to the behavior of materials under mechanical stress, no matter how this stress is produced—mechanical, thermal, or chemical.
5. There are no restrictions on such material properties as conductivity or magnetic permeability.

8. OTHER NDT PROCEDURES

This section lists and briefly describes NDT techniques that were not treated in earlier sections. Some of these are quite specialized; others have been demonstrated in the laboratory but have not yet found wide application.

If there are NDT procedures that are in existence but have not been mentioned in this article, the omission is unintentional and is regretted, with apologies to the developers of these procedures.

8.1 Strain Gauge

The use of strain gauges is not generally a flaw-finding technique, but a method of determining the state of strain in a surface, to determine experimentally the degree of safety in the operation of a mechanical structure. A strain gauge, fabricated of wire and thin layers of metal foil and bonded to a surface, measures strain by changes in its electrical resistance (Window and Holister, 1982).

8.2 Photoelasticity

This visual means for studying the results of applying various amounts of stress to an object makes use of the property of birefringence of some transparent materials when strained, to result in fringes that reveal strain patterns. The technique can be used with a model sheet of transparent material for design studies. For actual objects, the transparent material is bonded to a polished surface of the object as a transparent plastic coating, and the fringe, or strain pattern, is seen by reflection (Zandman *et al.*, 1977).

8.3 Brittle Coating

Fractures that occur in a thin coating that has been applied to a surface when a specimen is stressed indicate magnitudes and directions of strains in the surface (McMaster, 1959; Lineback, 1989).

8.4 Replication

The contours of a surface of a component are replicated by impressions with materials such as cellulose acetate or silicone. These impressions are observed and irregularities are quantified under examination with optical or electron microscopes, to attempt to predict the useful life of the component (McMaster, 1959; Marder, 1989).

8.5 Volatile-Liquid Method of Surface Crack Detection

In this simple technique of detecting regions of porosity or surface cracks, a volatile liquid is applied to the surface of a test piece. When the latter is heated, the liquid evaporates, leaving regions of porosity or spots with surface cracks colored or shaded (Oaks, 1972).

8.6 Leak Detection

The techniques of leak detection are aimed at finding microscopic channels or cracks in the walls of a volume that are supposed to be impervious to any gas transfer (McMaster, 1982a). The following techniques are in use.

8.6.1 Tracer-Gas Method This method is applicable to any structure that can be sealed, to become a closed vessel except for vacuum pumping through an orifice. This "closed vessel" is now vacuum pumped continuously through the orifice. To find a leak, helium is sprayed around the outside of the closed vessel while any gas remaining in the vessel is being examined for helium content by a mass spectrometer system. If sufficient helium is detected to sound an alarm, a leak exists in the wall of the structure in the vicinity of the location where the helium is being sprayed.

8.6.2 Sealed Package, Holographic Method A leak in a sealed package can be found by detecting small changes in the package dimensions by a holographic method (see Sec. 8.21).

8.7 Radioactive-Gas Penetrant

This method has been used to detect pores and cracks in the surface of a solid. When a radioactive gas such as krypton-85 is applied to the surface, some of the gas is absorbed in surface cracks or pores. Beta or gamma radiation emitted by this absorbed gas exposes x-ray–sensitive film placed on the surface, thereby generating a map of surface cracks and pores (Packer *et al.*, 1967).

8.8 Acoustic Impact Response

To search for cracks, debonding, and delaminating, an object under test is struck a blow with a hammer and the acoustic/vibrational transient response is compared with the response of a "perfect specimen" that has been struck in the same manner (Schroeer *et al.*, 1970).

8.9 Sonic Testing of Composite Structures

8.9.1 The Sonic-Resonator Method Flaws in a structure, such as disbonds, are detected by measuring the input impedance Z_{in} at audio frequencies to a piezoelectric transducer that is bonded to the structure and comparing this Z_{in} to that of a standard (Botsco, 1966).

8.9.2 The Eddy Sonic Method By setting up eddy currents of the frequency range 2–10 kHz in a partially electrically conducting structure, mechanical forces, and consequent vibrations at the eddy-current frequency, are created. Flaws are detected by monitoring the sonic output vs. frequency and comparing this with the output from a standard (Botsco, 1968).

8.10 Thermal Techniques

8.10.1 Temperature Mapping during Operation of Test Piece A temperature map of one or more of the surfaces of equipment during its actual operation can reveal regions whose temperature is excessive for safe operation. Such a map can be obtained by covering a surface with a film of liquid crystal or other types of material whose color is temperature sensitive, by the use of pyroelectric instruments, or by sensing and mapping the local thermal radiation emitted (McGonnagle, 1966; Brown, 1968).

8.10.2 Surface Flaw Detection by Scanning with a Heat Pulse Surface or near-surface flaws can be detected by scanning a surface with a pulse of heat and measuring the transient temperature rise in the vicinity of this scanning heat pulse. In the work that demonstrated this technique, the heat pulse was the heat from a watt-level laser beam; the transient temperature rise was obtained by measuring the far-infrared radiation emitted (Kaufman *et al.*, 1987).

8.11 Electric-Current Injection

Cracks, corrosion, or other flaws that reflect as irregularities of conductivity in bodies of regular shape of electrically conducting material, such as rail lines, can be found by resistance measurements by a four-point-probe method (Buchanan and Thurston, 1956).

8.12 Static Magnetic Field

Measurement of the local external magnetic-field intensity of magnetized ferromagnetic specimens of uniform and regular shapes can reveal the presence of surface cracks, subsurface discontinuities, and depths of case hardening. The technique can also be used to measure the thickness of nonmagnetic coating by comparison with a standard (Gardner and Barton, 1970).

8.13 Corona Discharge

The intent of this technique is to detect a void in dielectric material, such as in the insulation of cables or in glass-epoxy material. An imposed electric field of sufficiently high intensity ionizes the gas in the void, and electrons are accelerated to the wall of the void. The void can be detected either by the minute pulse of current in the secondary of the high-voltage transformer supplying the electric field or by the electromagnetic spectrum radiated (Dakin and Lim, 1957; Hendron *et al.*, 1964).

8.14 Capacitance Measurement

To check dielectric materials for leaky paths, electrodes are applied, and the capacitance and loss factor are measured and compared to a standard (Zubrick, 1968).

8.15 Microwaves

Microwaves, i.e., electromagnetic waves of range 1–100 GHz, can be used to measure the thickness of plates, as well as to locate voids and inclusions in dielectrics and cracks in metal surfaces, by measurement of changes in phase, amplitude, and polarization of transmitted or reflected waves (Bahr, 1982; Ida, 1992).

8.16 YIG Resonator

A sensitive method of detecting cracks in metal surfaces depends on observing the change in the ferrimagnetic resonance frequency of a small yttrium iron garnet sphere as it is moved along the surface (Auld, 1977, 1978).

8.17 Nuclear Magnetic Resonance Imaging (NMRI)

NMRI, the technique based on mapping local composition in the bulk of a substance by observing nuclear magnetic resonance frequencies and relaxation times in a spatially varying magnetic field, now a well-established procedure of medical diagnostics, can also be used to map composition and defects of spatial structure of materials (Listerud *et al.*, 1989; Sarkar and Komoroski, 1992).

8.18 Barkhausen-Noise Analysis (BNA)

When the magnetic-field intensity applied to a ferromagnetic material is swept smoothly up or down through a range of values, the magnetic-flux density changes in minute steps. A coil adjacent to the material translates these steps into voltage pulses. Signal analysis of these noise pulses reveals information about microstructural parameters of the material, stresses, fatigue, and creep damage (Sundstroem and Poerroenen, 1979).

8.19 Radiometric Surface Crack Detection in Moving Surfaces

Since a crack in a surface appears black, it has an emissivity for thermal radiation of virtually unity. This is generally higher than the emissivity of a surface. A crack in a moving surface can therefore be detected by monitoring the instantaneous thermal radiation received by a detector as the surface passes in front of it. (For a surface at room temperature, the detector must be sensitive to far-infrared radiation.) Changes in the pattern of the detected signal over a period of time indicate changes in the surface, such as the formation of new cracks (Kaufman and Choudhury, 1985).

8.20 Speckle Metrology

The speckle pattern of the reflected light that is seen when an optically rough surface is illuminated with highly coherent light can be used to measure displacements, detect cracks, and perform vibrational analysis (Sirohi, 1993).

8.21 Optical-Holography Interference Method

A holographic image of an object is stored on film. When this image is made to interfere optically with another holographic image that was taken while the body was stressed, or with the body itself, contour lines reveal surface displacements caused by such flaws as surface or internal cracks, or debonding (Erf, 1974; Vest, 1986).

8.22 Optical-Holography Vibrational Analysis

A hologram of an object taken while the object is vibrationally excited at some audio frequency yields a "time average" fringe pattern. Changes in internal structure are detected by comparison of this fringe pattern with that of a hologram recorded earlier with the same vibrational excitation (Powell and Stetson, 1965; Stetson, 1974).

8.23 Ultrasonic Holography

Ultrasonic holography substitutes ultrasonic acoustic waves for light waves, with the advantage of being able to "see" inside of opaque objects. Two basic ways of forming ultrasonic holograms exist:

1. an optical-analog scheme, in which surface displacements of a liquid surface, formed by interfering acoustic waves, are sensed optically and converted to an optical image;
2. point-by-point wave-front sampling by a moving transducer or by a number of transducers, with ultimate conversion to an optical image.

Ultrasonic holographic techniques offer the possibility of being a supplement to ordinary ultrasonic NDT, in which the latter can *find* defects, while acoustic holography can size them to an accuracy of 1–3 mm, even to a depth in steel of 30 cm (Szilard, 1982).

8.24 Acoustic Microscopy

With acoustic waves of frequencies in the range of 300 MHz and higher, pulse lengths of a few nanoseconds, and quasi-optical focusing techniques, internal and surface geometries of microscopic dimensions in test specimens can be sensed and converted to visual images (Briggs, 1995).

8.25 Exoelectron Emission from Metals

Emission of electrons in high vacuum, particularly if a surface is scanned by a spot of ultraviolet light, has been reported to provide quantitative measurements of distribution and extent of fatigue damage prior to formation of a fatigue crack (Hoenig *et al.*, 1971).

8.26 Compton-Scattered Radiography

This x-ray or gamma-ray technique is based on imaging the scattered, lower-energy photons resulting from Compton-scattering collisions. Since Compton collisions allow back-scattering, x-ray imaging with access to only one side of a specimen is possible. This suggests such applications as detecting and locating steel inside a concrete wall (Harding *et al.*, 1983/84).

8.27 Energy-Loss Ion Radiography

When finely focused beams of ions in the 1- to 5-MeV range are transmitted through thin slices of material, their energy loss suffered during the transmission can provide information of the material passed through. By rastering such a beam across a specimen, an area display similar to an x-ray radiograph can be obtained (Overley *et al.*, 1983).

8.28 Neutron Radiography

Neutron radiography complements x-ray or gamma-ray NDT by being insensitive to the presence of metals and, instead, revealing the presence of hydrogenous material. It can therefore be used for finding hydrogenous corrosion products; detecting rubber, plastics, and explosives; and imaging masses of these substances and finding imperfections in them (Bromley, 1983).

8.29 Positron Annihilation

When positrons are injected into a material from a radioactive source, they annihilate themselves by combining with electrons and generating gamma rays. Their mean lifetime before annihilation has been found to depend on the presence of microscopic defects. They can therefore be used to locate microscopic defects in situations where plastic deformation, fatigue cycling, or creep exist (Grosskreutz and Millet, 1969).

8.30 Tomographic Imaging

This covers the general technique of mapping the distribution of some material property inside a structure by finding the two-dimensional distributions in what are essentially "slices" of the structure. An example is the well-known CAT scan referred to in Sec. 4. Methods used include scanning with x rays (for material density), ultrasonic waves (for elastic properties), neutrons (for determining nuclei distribution), ions (for material density), and nuclear magnetic resonance (for nuclear-spin density and relaxation times) (Kuhl and Edwards, 1963; Hounsfield, 1973; Pontau *et al.*, 1989). The most successful of these methods, used extensively for medical diagnosis, are x-ray CAT scans, referred to in Sec. 4, and NMRI, of Sec. 8.17.

8.31 Surface-Crack Detection by Optical Scanning

Detection and mapping of surface cracks, measurement of crack width, and a pictorial display of cracks, even in a rough surface, by scanning with an optical beam make use of the fact that specular reflection in a fixed direction of a light-beam incident on a surface is interrupted by a crack (Chang *et al.*, 1987).

GLOSSARY

Acoustic Emission: The generation of transient acoustic waves by local rapid release of stored elastic energy during stress-induced irreversible structure changes.

Acoustic Waves: Also called "elastic waves." Mechanical waves that can propagate in liquids and solids by energy exchange between stress–strain potential energy and kinetic energy of particle motion.

A-Scan: From the pulse-echo method of ultrasonic NDT, display of echo signals vs. time that shows the location of sources of echoes and their relative magnitudes along a line that extends outward from the transducer into the specimen under test.

Barkhausen Noise: Small voltage pulses developed by a coil adjacent to a ferromagnetic material when an applied magnetic-field intensity is swept slowly through a range of values.

Borescope: Slender, straight-tube periscope or flexible fiber-optic system, used to visually inspect the interior of hollow components for defects.

B-Scan: For the pulse-echo method of ultrasonic NDT, a two-dimensional map of sources of echoes in the cross section of a specimen. Echo intensity can be indicated by relative brightness or by color.

Buildup: In radiographic NDT, undesirable fogging of x-ray film by radiation arising from Compton scattering in the test specimen.

Bulk Waves: Waves within a material specimen that do not depend on shape and dimensions of the body in which they propagate.

CAT Scan: See **Computer-Aided Tomography.**

Characteristic Impedance Z_0: A constant of a wave transmission system. It is a function of material properties and structure and determines relative amplitudes and phases of reflected and transmitted waves when waves arrive from another material or wave transmission system.

C-Scan: For the pulse-echo method of ultrasonic NDT, a two-dimensional map of the distribution of sources of echoes parallel to and beneath the surface of the item under test. Echo intensity can be indicated by relative brightness or by color.

Compton Scattering: Deflection from original path and reduction in energy of an x-ray photon by collision with a weakly bound electron.

Critical Angle of Incidence: The minimum angle of incidence of an acoustic wave that causes total reflection at a boundary between two media.

Decohesion: A brittle fracture, associated with rupture along grain boundaries.

Developer: In photography, a chemical that causes a latent image on a film to become a visible image. In liquid-penetrant inspection, a fine powder that aids in highlighting cracks by drawing colored or fluorescent fluid out of them.

Direct Type Film: A type of x-ray film, also called "nonscreen type," that is fabricated with a thin metal foil in contact with the emulsion.

Eddy Currents: Circulating currents in electrically conducting material caused by magnetic fields of time-varying currents flowing in adjacent conductors.

Elastic Waves: See **Acoustic Waves.**

Electromagnetic Coupling: Coupling between two systems by electromagnetic fields. In ultrasonic NDT this term means coupling between an electric circuit and acoustic waves in an electrically conducting test piece immersed in a magnetic field because of the Lorentz force $\mathbf{j} \times \mathbf{B}$.

Electrostatic Coupling: In ultrasonic NDT of a metal specimen, generation or detection of acoustic waves by utilization of the force on a conductor in an electric field.

Film Density, *D*: For a small spot of photographic or x-ray film that has been exposed and developed, $D = \log_{10}$ (intensity of light incident on film/intensity of light transmitted through film).

Film Gradient, G_D: Measure of film sensitivity to changes in incident light (or x-ray) level. G_D is slope of curve of film density D when plotted as a function of $\log_{10}E$, where E is the product of light (or x-ray) intensity and film exposure time.

Film Speed: Measure of relative sensitivity of a particular film. Example: To produce the same degree of blackening with the same incident light intensity, a film of speed 200 requires only half the exposure time of that required by a film of speed 100.

Fluoroscopy: Conversion of a radiographic image into a visible-light image by causing the radiation that has passed through a test specimen to impinge on a screen that emits visible light when struck by x rays or gamma rays.

Fracture Mechanics: Analysis of structures, to predict the degree of loading that will cause failure. Example: Study of the load-carrying capacity of a structure in the presence of an initial defect, such as a dominant crack.

Frame Averaging: Clarification of an image through a filtering technique based on averaging several image frames.

Half-Value Layer, HVL: The thickness of a material that reduces the intensity of radiation passing through this material to one-half of the incident value.

Image Quality Indicator, IQI: In radiographic NDT, fixture placed between x-ray source and specimen under test, to measure the minimum size of flaw that can be detected; also called "penetrameter."

Lamb Wave: An acoustic wave that can propagate in a thin solid sheet.

Leakage Field: The external magnetic field adjacent to the surface of a magnetized body in the vicinity of a surface crack or shallow subsurface defect.

Lift-Off Effect: In eddy-current testing, change in test coil impedance caused by separation of coil from specimen.

Liquid Penetrant: A liquid used to coat a surface in order to reveal cracks.

Longitudinal Wave: Acoustic wave in which particle motion is parallel to the direction of propagation.

Love Wave: An acoustic surface wave that can propagate on a solid covered with a thin layer of material of different density than that of the solid that it covers.

Magnetic Domain: A region in a solid within which elementary magnetic moments have arranged themselves parallel.

Magnetostriction: The mechanical strain developed in a ferromagnetic material on the application of a magnetic-field intensity.

Magnetic-Particle Inspection: Technique of searching for surface or shallow subsurface flaws of a magnetized ferromagnetic body by applying small iron particles and looking for spots on the surface where they accumulate.

Mode Conversion: The coupling of energy from one type of wave propagation to another one. Example: Conversion of a portion of the energy of a longitudinal to a transverse acoustic wave.

Nondestructive Evaluation, NDE: The procedure of examining a structure without damaging it, followed by evaluation of test results.

Nondestructive Inspection, NDI: The actual procedure of testing a structure.

Nondestructive testing, NDT: The procedure of examining a structure without damaging it. In the present article the initials NDT are used to include NDE and NDI.

Noninvasive: Procedure in which there is no penetration of the surface of a body or structure by a physical object.

Nonscreen Type Film: See **Direct Type Film.**

Penetrameter: Image quality indicator in radiographic NDT. See **Image Quality Indicator.**

Penumbra: Partially illuminated space between regions of perfect shadow and of full illumination. The penumbra regions in radiographic images set a limit on image resolving power.

Piezoelectric: Materials that are piezoelectric will generate electric polarization, and therefore electric fields, when mechanically strained. Conversely, they can become locally strained when placed in regions of electric field.

Plastic Deformation: Permanent change in shape or size of a solid body when it is subjected to a stress exceeding a particular value.

Pulse Echo: In ultrasonic NDT, the method of sending out short acoustic pulses and locating flaws and other irregularities by noting directions, times of arrival, and amplitudes of echoes.

Radiography: The science and technology of using x rays or gamma rays for imaging and analyzing internal structure of materials or organisms. The more common term used when dealing with living organisms is "radiology."

Radiometry: The technique of using spontaneous radiation emitted by a body or an area, sometimes called "blackbody radiation," to measure temperature or to map the area.

Radioscopy: The technique of creating analog images for continuous viewing of the results while a specimen is being illuminated with x rays or gamma rays.

Rayleigh Waves: Acoustic waves that propagate on smooth surfaces and can be used to detect surface or shallow subsurface defects. Rayleigh waves have also been developed to carry out high-frequency analog signal processing.

Remanence, Remanent Magnetization: The amount of magnetization remaining in a ferromagnetic structure after an impressed magnetic-field intensity that caused a net magnetization to exist has been removed. In practice, the remanence is the magnetization that remains after the exciting current in a coil surrounding the structure has been set equal to zero.

Resolving Power: A quantitative measure of the ability of an optical instrument to produce separable images.

Salt Screen Film: X-ray film whose emulsion is coated with a fluorescent powder that emits ultraviolet radiation when struck by x rays, thereby reducing needed exposure time, but at the expense of resolving power. Also called "screen-type film."

Scintillating Material: Materials such as NaI(Tl) that emit light when struck by x rays.

Screen-Type Film: See **Salt Screen Film.**

Shear Waves: Acoustic waves for which the particle motion is perpendicular to the direction of propagation. Also called transverse waves.

Skin Depth: For transverse electromagnetic waves propagating into a material, the depth at which the field intensity has decayed to 0.36787 of its value at the surface.

Spatial Filtering: In processing optical signals, the techniques of discriminating between various rates of spatial change (lines per mm) in light intensity.

Stonely Wave: An acoustic wave that can propagate along the interface of two different materials.

Tomography: The recording of internal structure of a body at a particular plane while blurring detail in other planes.

Transducer: A fixture or structure that converts energy from one type of distribution to another, or from one type of energy to another. Respective examples are radio antennas and the piezoelectric wafers used in ultrasonic NDT.

Transverse Waves: See **Shear Waves.**

Twinning: A process in which a portion of a crystal assumes an orientation that is symmetrically related to the basic crystal orientation.

Ultrasonic NDT: The method of NDT that is based on the use of acoustic waves whose frequencies lie in the ultrasonic range.

Unsharpness: An expression used in radiographic NDT to describe the amount of blur in the image of a specimen.

White Noise: A random signal whose power per unit frequency increment (up to some undefined high frequency) is the same for all frequencies.

Works Cited

Addison, R. C., Jr., Graham, L. J., Linebarger, R. S., Tittman, B. R. (1987), in: B. R. McAvoy (Ed.), *IEEE 1987 Ultrasonics Symposium, Proceedings*, Vol. 2, New York: Institute of Electrical and Electronics Engineers, Inc., pp. 1109–1113.

Aindow, A. M., Dewhirst, R. J., Hutchins, D. A., Palmer, S. B. (1981), *J. Acoust. Soc. Am.* **69,** 449–455.

Auld, B. A. (1977), in: *Proceedings of DARPA/AFML Review of Progress in Quantitative NDE*, Cornell University, Ithaca, NY: Defense Advanced Research Progress Agency.

Auld, B. A. (1978), in: G. L. Report 2839, E. L. Ginzton Laboratory, Stanford University, Stanford, CA, July, 1978.

Bahr, A. J. (1982), *Microwave Nondestructive Testing Methods*, New York: Gordon and Breach.

Bean, C. P., DeBlois, R. W., Nesbitt, L. B. (1959), *J. Appl. Phys.* **30,** 1976–1980.

Betz, C. E. (1967), *Principles of Magnetic Particle Testing*, Chicago, IL: Magnaflux Corp.

Birnbaum, G., White, G. S. (1984), in: R. S. Sharpe (Ed.), *Research Techniques in Nondestructive Testing*, Vol. VII, Orlando, FL: Academic.

Botsco, R. J. (1966), *Mater. Eval.* **24,** 617–623.

Botsco, R. (1968), *Mater. Eval.* **26,** 21–26.

Briggs, A. (Ed.) (1995), *Advances in Acoustic Microscopy*, New York: Plenum.

Bromley, D. A. (1983), *Physics Today* **36** (12), 31–39.

Brown, S. P. (1968), *Mater. Eval.* **26,** 163–166.

Bryant, L. E. (Technical Ed.) (1985), in: P. McIntire (Ed.), *Nondestructive Testing Handbook*, Vol. 3, *Radiography and Radiation Testing*, 2nd ed., Columbus, OH: American Society for Nondestructive Testing, p. 267.

Buchanan, J. G., Thurston, R. C. A. (1956), *Nondestr. Test.* **14,** 36–39.

Cartz, L. (1995), *Nondestructive Testing—Radiography, Ultrasonics, Liquid Penetrant, Eddy Current*, Materials Park, OH: ASM International.

Chang, P. T., Kaufman, I., Shyong, D.-Y. (1987), *Mater. Eval.* **45,** 943–950.

Dakin, T. W., Lim, J. (1957), *AIEE Trans. Power Apparatus Systems* **76,** 1059–1065.

Dunegan, H. L., Harris, D. O., Tetelman, A. S. (1970), *Mater. Eval.* **28,** 221–227.

Erf, R. K. (Ed.) (1974), *Holographic Nondestructive Testing*, New York: Academic.

Gardner, C. G., Barton, J. R. (1970), in: *Recent Advances in Magnetic Field Methods of Nondestructive Evaluation for Aerospace Applications, Advanced Technology for Production of Aerospace Engines*, Advisory Group for Aerospace Research and Development (NATO), AGARD-CP-64, Neuilly-sur-Seine, France.

Gdoutos, E. E. (1993), *Fracture Mechanics: An Introduction*, Dordrecht: Kluwer Academic.

Grosskreutz, J. C., Millet, W. E. (1969), *Phys. Lett.* **28A,** 621–622.

Halmshaw, R. (1987), *Nondestructive Testing*, London: Edward Arnold.

Harding, G., Strecker, H., Tischler, R. (1983/84), *Philips Tech. Rev.* **41,** 46–59.

Heiple, C. R., Carpenter, S. H. (1987), *J. Acoustic Emission* **6,** 177–204, 215–237.

Hellan, K. (1984), *Introduction to Fracture Mechanics*, New York: McGraw-Hill.

Hendron, J. A., Groble, K. K., Gruetzmacher, R. W., McClurg, G. O., Retsky, M. W. (1964), *Mater. Eval.* **22,** 311–314.

Hoenig, S. A., Ott, W. A., Ali, M. T., Russell, T. A. (1971), *New Techniques in Nondestructive Testing (Exo-Electron Emission Phase), Part 1*, University of Arizona, Tucson, AZ (AFML-TR-71-140-pt. 1, AD-748272).

Hounsfield, G. N. (1973), *Br. J. Radiol.* **46,** 1016.

Hsieh, T. M., Lindgren, E. A., Rosen, M. (1991), *Ultrasonics* **29,** 38–44.

Hull, B., John, V. (1988), *Non-Destructive Testing*, London: Macmillan Education Ltd.

Ida, N. (1992), *Microwave NDT*, Dordrecht: Kluwer Academic.

Kaiser, J. (1953), *Arch. Eisenhutten* **24,** 43.

Kaufman, I., Choudhury, A. K. (1985), *Appl. Phys. Lett.* **48,** 152–154.

Kaufman, I., Chang, P.-T., Hsu, H.-S., Huang, W.-J., Shyong, D.-Y. (1987), *J. Nondestructive Testing* **6,** 87–100.

Kaule, W. (1963), in: *Proceedings of the 4th International Conference on Nondestructive Testing, London, 1963,* London: Butterworth, pp. 316–318.

Kenton, R. H. (1970), presented at Sixth International Conference on Nondestructive Testing, sponsored by German Society for Nondestructive Testing Methods, Hanover, Germany, June 1–5, 1970. Not published.

Knott, C. G. (1899), *Philos. Mag.* **48,** 567–568.

Krautkrämer, J. H., Rudiger, O. (1949), *Arch. Eisenhuttenw.* **20,** 355–358.

Krautkrämer, J., Krautkrämer, H. (1990), *Ultrasonic Testing of Materials,* 4th Rev. Ed., New York: Springer-Verlag.

Kuhl, D. E., Edwards, R. Q. (1963), *Radiology* **80,** 653.

Libby, H. L. (1971), *Introduction to Electromagnetic Nondestructive Test Methods,* New York: Wiley-Interscience.

Lineback, L. D. (1989), in: S. R. Lampman, T. B. Zorc (Eds.), *Metals Handbook,* Vol. 17, *Nondestructive Testing and Quality Control,* 9th ed., Metals Park, OH: ASM International, p. 453.

Listerud, J. M., Sinton, S. W., Dobny, G. P. (1989), *Anal. Chem.* **61,** 23A–41A.

Lynnworth, L. L. (1989), *Ultrasonic Measurements for Process Control—Theory, Techniques, Applications, Boston:* Academic.

Marder, A. R. (1989), in: S. R. Lampman, T. B. Zorc (Eds.), *Metals Handbook,* Vol. 17, *Nondestructive Testing and Quality Control,* 9th ed., Metals Park, OH: ASM International, pp. 52–56.

McGonnagle, W. J. (1966), *Nondestructive Testing,* 2nd ed., New York: Gordon and Breach.

McIntire, P. (Ed.) (1985), *Nondestructive Testing Handbook,* Vol. 3, *Radiography and Radiation Testing,* Columbus, OH: American Society for Nondestructive Testing.

McIntire, P. (Ed.) (1987), *Nondestructive Testing Handbook,* Vol. 5, *Acoustic Emission Testing,* Columbus, OH: American Society for Nondestructive Testing.

McIntire, P. (Ed.) (1989), *Nondestructive Testing Handbook,* Vol. 6, *Magnetic Particle Testing,* Columbus, OH: American Society for Nondestructive Testing.

McIntire, P. (Ed.) (1991), *Nondestructive Testing Handbook,* Vol. 7, *Ultrasonic Testing,* Columbus, OH: American Society for Nondestructive Testing.

McIntire, P. (Ed.), McMaster, R. C. (Ed. Emeritus) (1986), *Electromagnetic Testing,* Vol. 4, Columbus, OH: American Society for Nondestructive Testing.

McIntire, P., Moore, P. O. (Eds.) (1993), *Nondestructive Testing Handbook,* Vol. 8, *Visual and Optical Testing,* Columbus, OH: American Society for Nondestructive Testing.

McMaster, R. C. (Ed.) (1959), *Nondestructive Testing Handbook,* Vols. I and II, Columbus, OH: American Society for Nondestructive Testing.

McMaster, R. C. (Ed.) (1982a), *Nondestructive Testing Handbook,* Vol. 1, *Leak Testing,* Metals Park, OH: American Society for Metals.

McMaster, R. C. (Ed.) (1982b), *Nondestructive Testing Handbook,* Vol. 2, *Liquid Penetrant Tests,* Metals Park, OH: American Society for Metals.

Mitchell, M. R., Landgraf (Eds.) (1992), *Advances in Fatigue Lifetime Predictive Techniques,* Philadelphia, PA: American Society of Testing Materials.

Monchalin, J. P., Heon, R. (1986), *Mater. Eval.* **44,** 1231–1237.

Oaks, A. E. (1972), *Nondestructive Testing Techniques of Scout Rocket Motors,* NASA CR-2013, Washington, DC: U.S. GPO.

Overley, J. C., Connolly, R. C., Singer, G. E., MacDonald, J. D., LeFevre, H. W. (1983), *Nucl. Instrum. Methods Phys. Res.* **218,** 43–46.

Packer, L. L., Bruton, W. A., Hecht, R. (1967), *Investigation of Radioactive Gas Radiography for Inspection of Brazed Insert Joints,* Report F-930511-1, United Aircraft Research Laboratory, East Hartford, CT.

Pollock, A. A. (1981), *International Advances in Nondestructive Testing,* Vol. 7, New York: Gordon and Breach, pp. 215–239.

Pollock, A. A. (1989), in: S. R. Lampman, T. B. Zorc (Eds.), *Metals Handbook,* Vol. 17, *Nondestructive Evaluation and Quality Control,* 9th ed., Metals Park, OH: ASM International, pp. 278–294.

Pontau, A. E., Antolak, A. J., Morse, D. H., Ver Berkmoes, A. A., Brase, J. M., Heikkinen, D. W., Martz, H. E., Proctor, I. D. (1989), *Nucl. Instrum. Methods Phys. Res.* **B40/41,** 646–650.

Powell, R. L., Stetson, K. A. (1965), *J. Opt. Soc. Am.* **55,** 1593–1598.

Radon, J. (1917), *Saechs. Akad. Wiss. Leipzig Ber.* **69,** 262.

Sarkar, S. N., Komoroski, R. A. (1992), *Macromolecules* **25,** 1420–1426.

Schofield, B. H., Bariss, R. A., Kyrala, A. A. (1958), *WADC TR-58-194,* Wright Air Development Center, Ohio. Later available from Armed Services Technical Information Agency (ASTIA) as ASTIA Doc. AD 155674. (Successor to ASTIA: Defense Documentation Center, Ft. Belvoir, VA.)

Schroeer, R., Rowand, R., Kamm, H. (1970), *Mater. Eval.* **28,** 237–248.

Schubring, N. W. (1955), *Iron Age* **176,** 87–90.

Scott, I. G. (1991), *Basic Acoustic Emission,* New York: Gordon and Breach.

Sirohi, R. S. (Ed.) (1993), *Speckle Metrology,* New York: Marcel Dekker.

Stetson, K. A. (1974), in: R. K. Erf (Ed.), *Holographic Nondestructive Testing,* New York: Academic.

Sundstroem, O., Poerroenen, K. (1979), *Mater. Eval.* **37,** 51–56.

Szilard, J. (Ed.) (1982), *Ultrasonic Testing, Nonconventional Testing Techniques,* New York: Wiley.

Vest, C. M. (Ed.) (1986), *Holographic Nondestructive Testing: Status and Comparison with Conventional Methods,* SPIE Proceedings, Vol. 604, Bellingham, WA: SPIE, The International Society for Optical Engineering.

Waidelich, D. L. (1956), *Nondestr. Test.* **14,** 14–16.

Whittington, K. R. (1982), private communication, reported in: J. Szilard (Ed.), *Ultrasonic Testing, Nonconventional Testing Techniques,* New York: Wiley.

Window, A. L., Holister, G. S. (1982), *Strain Gauge Technology,* Englewood, NJ: Applied Science.

Zandman, F., Redner, S., Dally, J. W. (1977), *Photoelastic Coatings,* Ames, IA: Iowa State University Press.

Zubrick, J. R. (1968), *Mater. Res. Standards* **8,** 25–36.

Further Reading

American Society for Metals and American Society for Nondestructive Testing: (1982–1993), *Nondestructive Testing Handbook:*

Vol. 1—McMaster, R. C. (Ed.) (1982), *Leak Testing,* Metals Park, OH: American Society for Metals.

Vol. 2—McMaster, R. C. (Ed.) (1982), *Liquid Penetrant Tests,* Metals Park, OH: American Society for Metals.

Vol. 3—McIntire, P. (Ed.) (1985), *Radiography and Radiation Testing,* Columbus, OH: American Society for Nondestructive Testing.

Vol. 4—McIntire, P. (Ed.), (1986), *Electromagnetic Testing,* Columbus, OH: American Society for Nondestructive Testing.

Vol. 5—McIntire, P. (Ed.) (1987), *Acoustic Emission Testing,* Columbus, OH: American Society for Nondestructive Testing.

Vol. 6—McIntire, P. (Ed.) (1989), *Magnetic Particle Testing,* Columbus, OH: American Society for Nondestructive Testing.

Vol. 7—McIntire, P. (Ed.) (1991), *Ultrasonic Testing,* Columbus, OH: American Society for Nondestructive Testing.

Vol. 8—McIntire, P., Moore, P. O. (Eds.) (1993), *Visual and Optical Testing,* Columbus, OH: American Society for Nondestructive Testing.

Broek, D. (1988), *The Practical Use of Fracture Mechanics,* Dordrecht: Kluwer Academic.

Bray, D. E., Stanley, R. K. (1989), *Nondestructive Evaluation,* New York: McGraw-Hill.

Cartz, L. (1995), *Nondestructive Testing—Radiography, Ultrasonics, Liquid Penetrant, Eddy Current,* Materials Park, OH: ASM International.

Erf, R. K. (Ed.) (1974), *Holographic Nondestructive Testing,* New York: Academic.

Gdoutos, E. E. (1993), *Fracture Mechanics: An Introduction,* Dordrecht: Kluwer Academic.

Halmshaw, R. (1987), *Nondestructive Testing,* London: Edward Arnold.

Lampman, S. R., Zorc, T. B. (Eds.) (1989), *Metals Handbook,* Vol. 17, *Nondestructive Testing and Quality Control,* Metals Park, OH: ASM International.

McGonnagle, W. J. (1966), *Nondestructive Testing,* 2nd ed., New York: Gordon and Breach.

Scott, I. G. (1991), *Basic Acoustic Emission,* New York: Gordon and Breach.

Sharpe, R. S. (1970), *Research Techniques in Nondestructive Testing,* New York: Academic.

Szilard, J. (Ed.) (1982), *Ultrasonic Testing,* New York: Wiley.

THEORY AND APPLICATIONS OF FUZZY LOGIC

See Supplement Volumes

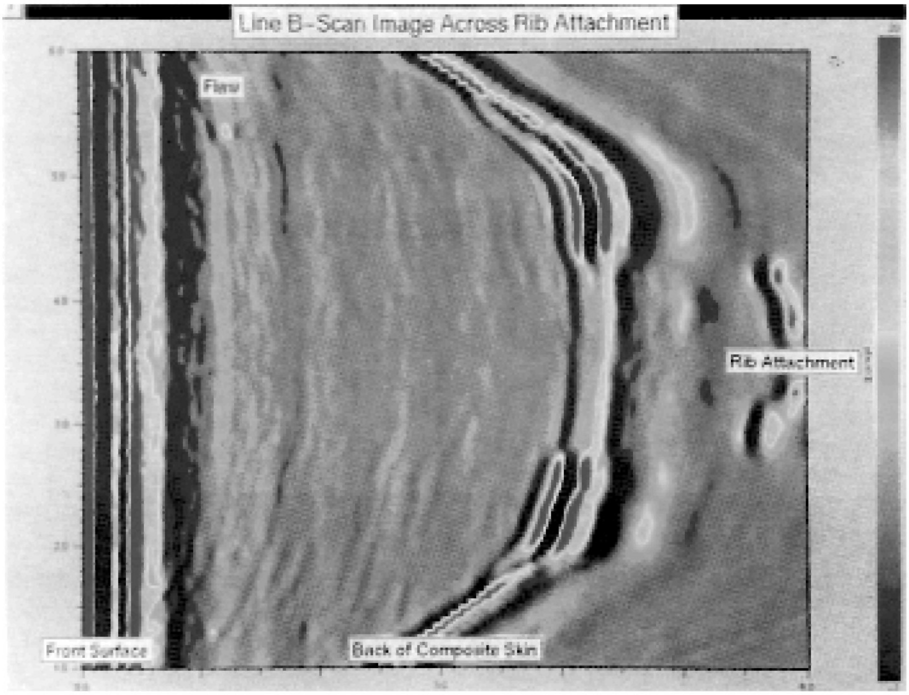
Line B-Scan Image Across Rib Attachment
Flaw
Rib Attachment
Front Surface
Back of Composite Skin

Vol. 21 – PLATE 1. B-Scan image from the rib attachment point of an air foil. (Courtesy of Panametrics, Waltham, MA 02154, U.S.A.)

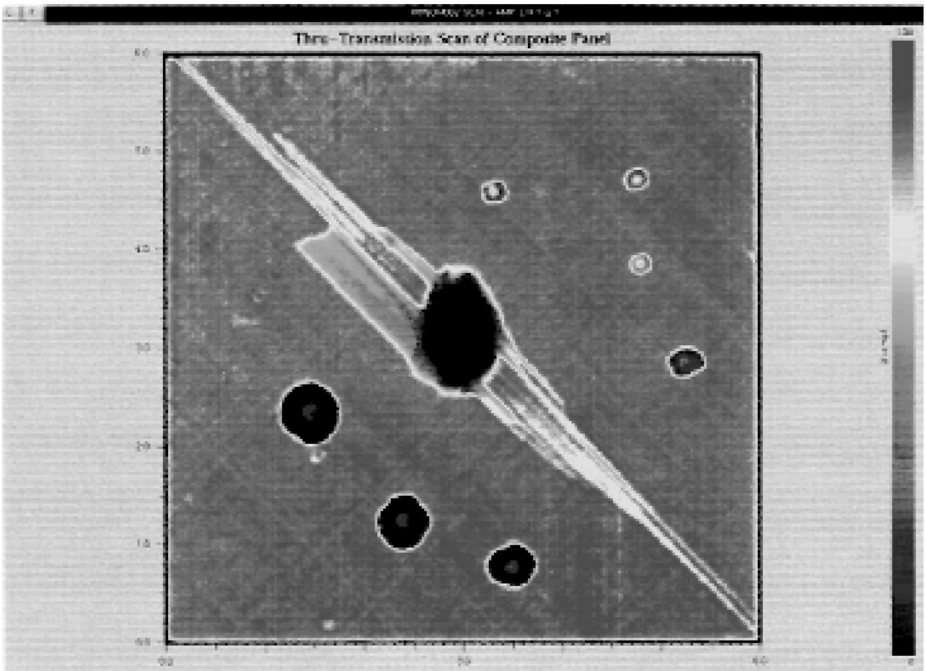
Thru-Transmission Scan of Composite Panel

Vol. 21 – PLATE 2. C-Scan image of fiber/epoxy composite sheet, showing internal impact damage. (Courtesy of Panametrics, Waltham, MA 02154, U.S.A.)

THERMAL ANNEALING

Rajiv K. Singh, D. Kumar, and B. Chen, *Department of Materials Science and Engineering, University of Florida, Gainesville, Florida, U.S.A.*

Introduction 67
1. **Defects in Materials** 68
1.1 Point Defects 69
1.2 Extended Defects 69
2. **Annealing of Metals** 71
2.1 Annealing Procedure 71
2.1.1 Normalizing 72
2.1.2 Full Annealing 72
2.1.3 Spheroidizing 72
2.2 Stress Relief, Recrystallization, and Grain Growth 72
2.3 Factors Affecting the Properties of Steel 73
3. **Annealing of Semiconductors** 73
3.1 Diffusion Effects during Thermal Annealing 74
3.2 Specialized Thermal Annealing Methods in Semiconductors 75
3.2.1 Furnace Annealing 75
3.2.2 Rapid Thermal Annealing 76
3.2.3 Pulsed Laser Annealing 78
3.3 Technology Trends in Specialized Annealing Methods 80
4. **Thermal Annealing in Oxide Materials** 81
4.1 High-Temperature Superconductors 81
4.2 Colossal-Magnetoresistive Materials 83
4.3 Ceramic Oxides 84
5. **Conclusions** 85
Glossary 86
Works Cited 86
Further Reading 87

INTRODUCTION

The electrical, thermal, optical, magnetic, and mechanical properties of materials are significantly dependent on its microstructure, i.e., structure of the material on a microscopic scale. All materials when examined at an atomic scale contain a large number of imperfections or defects, which have a profound effect on the behavior of the material. Typically, microstructural defects occur when atoms are displaced from their crystallographic lattice sites or when foreign atoms occupy lattice and nonlattice sites. The dimensions of the defects range from atomic size (point defects) to large extended defects such as grain boundaries, dislocations, stacking faults, voids, clusters, precipitates, etc. The properties of the materials are profoundly sensitive to these microstructural defects. The influence of defects on material properties is not always adverse, and often specific properties can be deliberately fashioned by controlled introduction of particular defects. For example, defects in semiconductor materials act as trapping sites for electron and hole transport, thereby significantly degrading the transport and optical properties and device performance, whereas controlled defects such as precipitates are introduced in metallic alloys to increase their yield strength. Thus, a controlled creation of microstructural defects may result in the formation of stronger metals and intermetallic alloys, long-lifetime lasers and light-emitting diodes, superconductors with enhanced current-carrying abilities, faster and more densely packed electronic circuits, more powerful magnets, and many other materials of practical importance.

One of the easiest and most widely used methods to alter and control the microstructure and crystalline defects in materials is to perform an elevated-temperature heat treatment or "thermal annealing" for an extended period of time. The annealing treatment is primarily meant to alter the microstructure,

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

produce a specific microstructure, and alter stresses and defect concentrations in the materials. Thermal annealing of materials is typically conducted in an isothermal, cylindrical or box furnace that is heated by resistive elements. The temperature and the ramp rate of the furnace can be precisely controlled. The near-equilibrium conductive, convective, and radiative heat transfer from the heating elements to the workpiece ensures isothermal heating of the materials with a high degree of precision in thermal uniformity. When the temperature of the material is increased, the motion of the atoms is significantly enhanced. This allows the atoms to migrate to the lattice sites. The stabilities of different defects are not the same; some defects can be annealed out at moderate temperatures (600–800 °C), while in other cases even very high temperatures (1000–1300 °C) may not be sufficient for complete defect removal. Thus the understanding of the thermal stability of defects is critical for controlling the microstructural development during the thermal processing of a given material.

Thermal annealing of materials may not only produce desirable changes in microstructure, but may also lead to unwanted effects such as precipitate growth in metals, dopant diffusion in semiconductors, and thin-film reactions in superconducting thin films. For example, an optimum goal for any silicon-based–semiconductor annealing process is to remove defects to ensure substitutional incorporation of the dopants so that a high degree of dopant activation and higher mobility of the carriers can be achieved. Besides obtaining these desirable properties, the annealing process should minimize unwanted effects such as dopant diffusion, which is especially critical for the next generation of very large-scale integrated (VLSI) circuits. In metal processing, the heat treatment reduces defects like dislocation density to improve ductility and reduce stress, but this treatment also causes grain growth, which results in detrimental mechanical properties. Thus, the optimization of the thermal annealing parameters (temperature, residence time, and gas environment) is an important engineering challenge. An overview of the critical issues related to application of thermal annealing to various material systems such as metals, semiconductors, ceramics, and high-temperature superconductors follows.

1. DEFECTS IN MATERIALS

A brief description of crystalline imperfections in solids has been provided in this section in order to understand the annealing processes and mechanisms of annealing in a better way. Solid materials may be classified according to the regularity in which the atoms or ions are arranged with respect to one another. A *crystalline material* is one in which the atoms are situated in a repeating or periodic array over large atomic distances; i.e., long-range ordering exists. To define the long-range ordering the term *lattice* is used in context of crystal structures. The term lattice means a three-dimensional array of points coinciding with the atom positions. For a crystalline solid, when the periodic and repeated arrangement in atoms is perfect and exists through the entirety of the specimen without interruption, the substance is called a *single crystal.* In a single crystal, the atoms have the same periodic structure or orientation throughout the whole specimen. Most crystalline solids are composed of a collection of small crystals or *grains;* such materials are termed *polycrystalline.* Thus throughout a grain, the material has the same crystallographic orientation. This orientation varies from grain to grain. Because of varying crystallographic orientation between grains, there exists an atomic mismatch within the region where two grains meet; this area is called the *grain boundary.* If a sufficient number of atoms are displaced from their lattice sites such that there is no long-range ordering, the material becomes *amorphous* or *noncrystalline.*

Any departure from the three-dimensional lattice grid represents a defect in the material. A crystalline defect is a lattice irregularity of the scale of atomic size in one or more dimensions. The classification of crystalline imperfections is made according to the geometry and the dimensionality of the crystalline imperfection. Thus one has point defects (associated with one of two atomic positions), linear (one-dimensional) defects such as *dislocations*, or two-dimensional defects such as *interfaces* and grain boundaries. See POINT AND EXTENDED DEFECTS IN CRYSTALS.

1.1 Point Defects

Point defects play a very significant role during thermal annealing of materials. Figure 1 shows some of the defects that are commonly observed in nonionic solids (metals, semiconductors, and covalently bonded ceramics). The simplest form of point defects is a vacancy, or a vacant lattice site that is normally occupied by the host atom. Thermodynamically an equilibrium concentration of vacancies exists at any given temperature; their concentration increases with increasing temperature. An interstitial atom is one crowded into the small space between lattice sites. If the interstitial atom is an atom from the crystal itself, it is called self-interstitial. Other types of point defect are formed when foreign atoms (impurities) are incorporated into the crystal. The impurity point defects can be either substitutional or interstitial. For substitutional, the impurity point defect replaces the host atom, while an interstitial atom fills the voids or interstices of the host atoms.

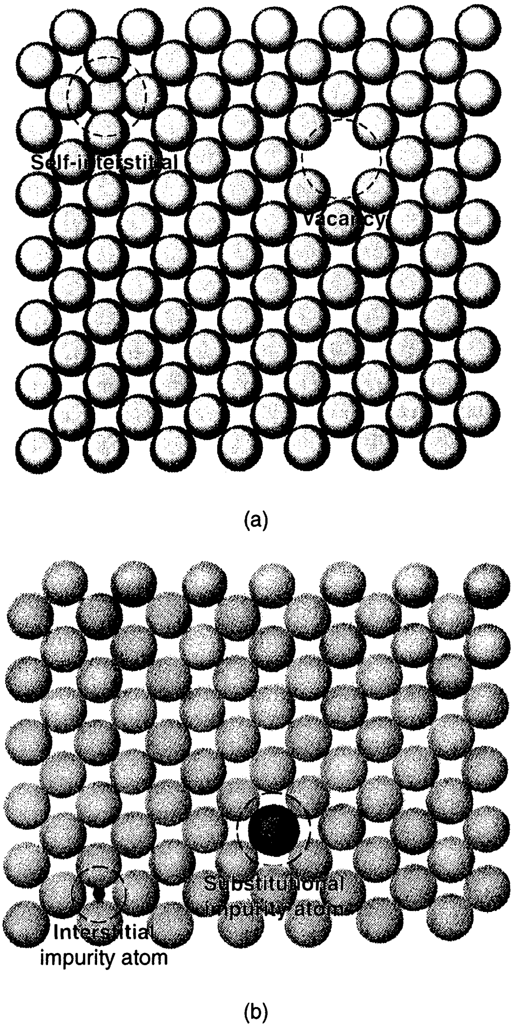

(a)

(b)

FIG. 1. Two-dimensional schematic representations of (a) a vacancy and self-interstitials and (b) substitutional and interstitial impurity atoms.

Atomic point defects in oxides or ceramic compounds become more complicated because two or more types of crystal sublattice exist in these materials. For example, in MgO, magnesium interstitials and vacancies and oxygen interstitials and vacancies may exist. As a consequence of electroneutrality, defects in ceramics may occur in pairs. One such type of defect involves a cation vacancy and a cation–interstitial pair and is termed the Frenkel defect (Fig. 2). A Schottky defect is a cation-vacancy, anion-vacancy pair, which is created by removing one cation and one anion from the interior of the crystal and placing them both at the external surface (Fig. 2).

1.2 Extended Defects

Defects that possess at least one dimension extending to large lengths (greater than several crystal lattice spacings) can be termed extended defects. A dislocation is a linear or one-dimensional defect around which some of the atoms are misaligned. One type of dislocation is represented in Fig. 3(a): an extra half-plane of atoms, the edge of which terminates within the crystal. This edge is termed the dislocation line, which in the case of edge dislocations is perpendicular to the plane of the page. Another type of dislocation is called the screw dislocation, shown in Fig. 3(b), in which the atomic distortion is linear along the dislocation line. Many dislocations in materials are mixed with some component being screw and the rest being edge type. In metals and to a lesser degree in semiconductors, the dislocations can move with the application of mechanical stress and thus are significantly involved in the plastic deformation of crystalline materials. The ductility (or the brittleness) of materials depends on the ability or inability of the dislocations to move.

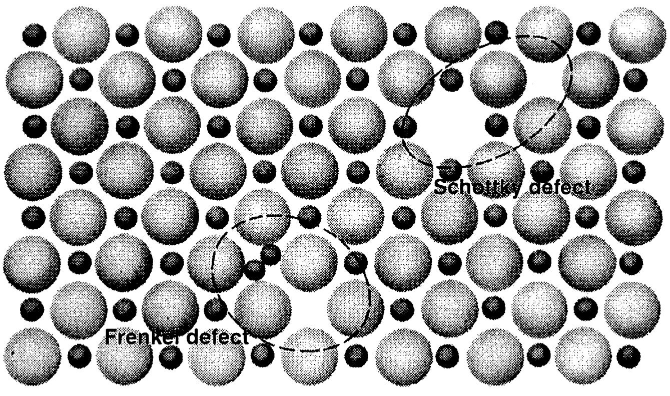

FIG. 2. Schematic diagram showing Frenkel and Schottky defects in ionic solids.

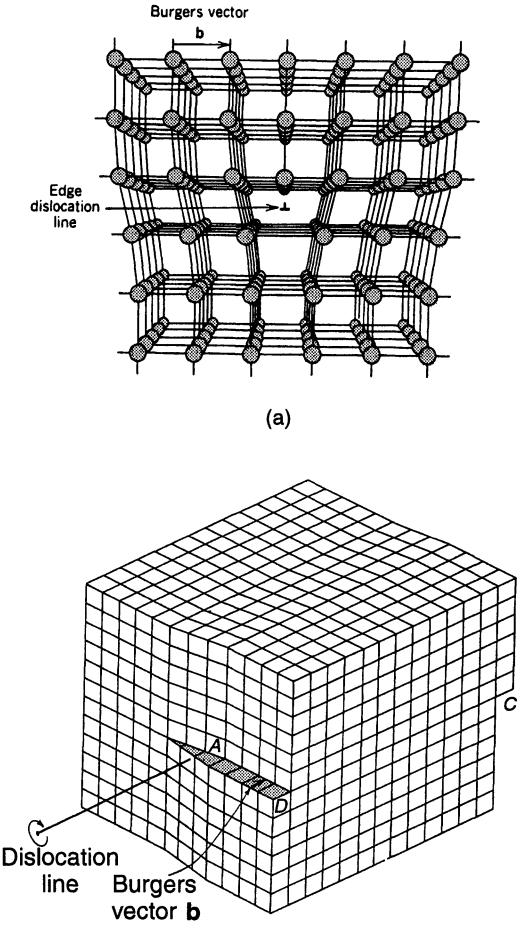

FIG. 3. Schematics of (a) an edge dislocation and (b) a screw dislocation within a crystal.

High-temperature annealing facilitates the movement of dislocations and can accelerate the annihilation process. In ceramic materials, dislocations are not mobile even at high temperatures and therefore do not play a significant role in the annealing process.

Another interfacial defect that plays a critical role in the annealing process is the grain boundary [Fig. 4(a)], which corresponds to the boundary separating two grains of different crystallographic orientation (see GRAIN BOUNDARIES). Within the boundary region, which is approximately a few atomic spacings wide, there is atomic mismatch due to transition from one crystallographic orientation to another. When the angular mismatch is small, on the order of a few degrees, then the term low-angle grain boundary is used. During thermal annealing, the grains increase in size or coarsen, resulting in a decrease in the total boundary energy. There is another type of grain boundary that exists commonly in many crystals, the twin boundary. A twin boundary is a special type of grain boundary across which there is a specific mirror lattice symmetry; i.e., atoms on one side of the boundary are located in mirror-image positions of the atoms on the other side [Fig. 4(b)]. Twins occur on a definite crystallographic orientation plane and in a specific direction, both of which depend on the crystal structure. Twins result from atomic displacement and are produced from applied mechanical shear forces (mechanical twins) or during annealing heat treatment

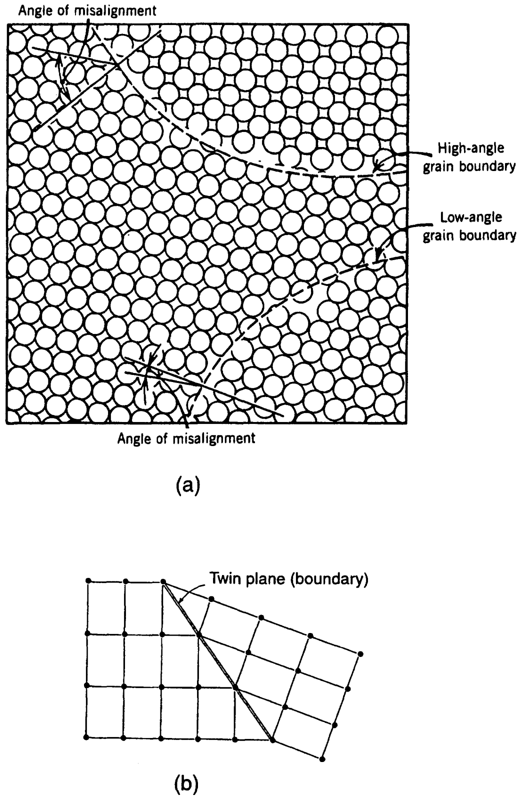

FIG. 4. Schematic diagram showing (a) low- and high-angle grain boundaries and (b) a twin boundary.

(annealing twins). Another type of defect that is important in metals, semiconductors, and superconductors is the stacking fault, which is formed when the stacking sequence of the close-packed planes is interrupted.

2. ANNEALING OF METALS

Annealing of metals is mostly used to eliminate the effect of cold working and to restore the cold-worked metal to the original soft, ductile condition. A variety of annealing heat treatments are possible, characterized by the changes that are induced. These changes are mostly related to microstructure resulting in the alteration of mechanical properties. The annealing process consists of three stages:

1. heating to the desired temperature,
2. holding at that temperature, and
3. cooling, usually to room temperature.

During annealing, diffusional processes are normally active. Therefore, temperature is an important consideration to accelerate the annealing process. Time is also an important parameter in these procedures as the actual annealing time must be long enough to allow for any necessary transformation reactions, particularly if the reaction involves diffusion. During heating and cooling, there exist temperature gradients between the outside and interior portion of the piece. The magnitudes of temperature gradients depend on the size and shape of the piece. If cooling is too fast or the rate of temperature change is too great, the temperature gradients cause internal stresses that may lead to warping or even cracking.

2.1 Annealing Procedure

Because of the technical importance and complexity of iron-based alloys, annealing procedure of only this system is discussed in the present article. However, many features associated with the annealing of iron-based alloys are common to annealing of other metals also (Clarebrough *et al.*, 1955). The main purpose of an annealing procedure is to enhance certain properties of steel alloys. Therefore, the annealing cycle and temperature of iron-based alloys should be designed in such a manner that the process results in the realization of steels with the optimum properties. The iron–iron-carbide phase diagram shown in Fig. 5 helps one to understand the annealing procedures more conveniently. The horizontal line at the eutectoid (three-phase reaction in which one solid phase transforms to two different solid phases) temperature, conventionally labeled A_1, is termed the *lower critical temperature,* below which, under equilibrium conditions, all austenite will have transformed into ferrite and cementite phases (see STEEL). Austenite is a face-centered-cubic iron or steel alloy. The phase boundaries denoted as A_3 and A_{cm} represent the upper critical temperature lines for hypoeutectoid and hypereutectoid steels, respectively. Hypo- and hypereutectoid refer to alloys for which the concentration of solute is less and greater than the eutectoid composition, respectively. For temperatures and compositions above these boundaries, only the austenite phase prevails.

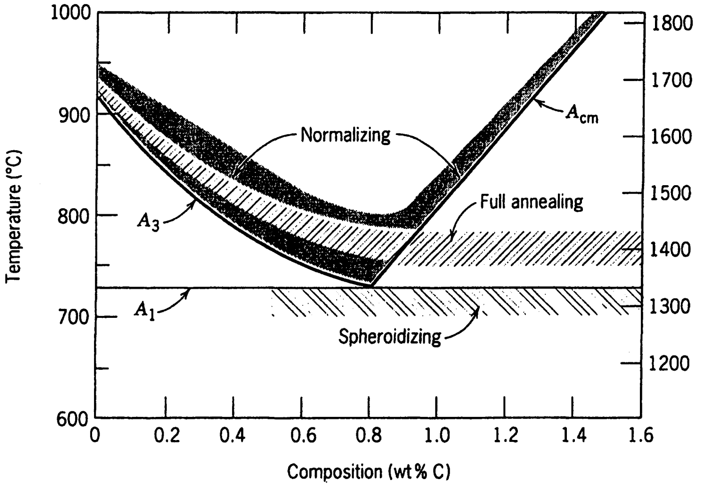

FIG. 5. The labeling of phase boundaries in the iron–iron-carbide phase diagram in the vicinity of the eutectoid indicating heat-treating temperature ranges for plain carbon steels. (Adapted from Lyman, 1948.)

2.1.1 Normalizing It has been commonly found that steels deformed plastically by a rolling operation consist of grains of pearlite, which are irregularly shaped and relatively large, but varying substantially in size. An annealing heat treatment called *normalizing* is used to refine the grains (i.e., to decrease the average grain size) and produce a more uniform and desirable size distribution. It is worth mentioning here that pearlitic steels are tougher than coarse-grained ones. Normalizing is accomplished by heating at approximately 55–85 °C above the upper critical temperature, which is dependent on composition as indicated in Fig. 5. After sufficient time has been allowed for the alloy to completely transform to austenite, the treatment is terminated by cooling in air.

2.1.2 Full Annealing A heat treatment known as *full annealing* is often utilized in low- and medium-carbon steels that will be machined or will experience extensive plastic deformation during a formation operation. The alloy is austenitized by heating to 15–45 °C above the A_3 or A_1 lines in Fig. 5 until equilibrium is achieved. The alloy is then furnace cooled for several hours. The microstructural product of this anneal is coarse pearlite that is relatively soft and ductile. Although the full-anneal cooling procedure is time-consuming, it yields a microstructure having small grains and uniform grain structure.

2.1.3 Spheroidizing Medium- and high-carbon steels, which may contain coarse pearlite, can still be too hard to machine or plastically deform conveniently. These steels may be annealed to develop the spheroidite structure. Spheroidized steels have a maximum softness and ductility and are easily machined or deformed. The spheroidizing heat treatment consists of heating the alloy at a temperature just below the eutectoid line A_1 in Fig. 5.

2.2 Stress Relief, Recrystallization, and Grain Growth

The original cold-worked microstructures of metals consist of deformed grains containing a large number of tangled dislocations. When the metal is first heat treated at low temperatures (low-temperature treatment is called recovery), the thermal energy forces the dislocations to move and form the boundaries of a polygonized subgrain structure. The density of dislocations, however, remains the same as shown schematically in Fig. 6. Because the number of dislocations is not changed during recovery, the mechanical properties remain practically unchanged. However, residual stresses are reduced or even eliminated when the dislocations are rearranged. During this annealing, the high electrical conductivity of the metal is also restored.

Recrystallization occurs by the nucleation and growth of new grains containing few

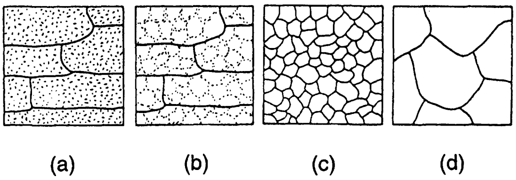

FIG. 6. The effect of annealing temperature on the microstructure of cold-worked metal; (a) cold worked, (b) after recovery, (c) after recrystallization, and (d) after grain growth.

dislocations. When the metal is heated above the recrystallization temperature, rapid recovery eliminates residual stresses and produces the polygonized dislocation structure. New grains then nucleate at the cell boundaries of the polygonized structure, eliminating most of the dislocations as shown in Fig. 6. Because of reduction in the number of dislocations, the recrystallized metal has a low strength but a high ductility.

At still higher annealing temperatures, both recovery and recrystallization occur rapidly producing the fine recrystallized grain structure. However, the energy associated with the large amount of grain-boundary area makes the fine structure unstable at high temperatures. To reduce this energy, the grains begin to grow, with favored grains consuming the smaller grains (Fig. 6). This phenomenon is called *grain growth.* A number of factors influence the size of the recrystallized grains. Reducing the annealing temperature, the time required to heat to the annealing temperature, or the annealing time will reduce the grain size by minimizing the opportunity for grain growth. Increasing the initial cold work also reduces the final grain size by providing a greater number of nucleation sites for new grains. Finally, the presence of a second phase in the microstructure helps prevent grain growth and keeps the recrystallized grain size small.

2.3 Factors Affecting the Properties of Steel

The optimum properties of a steel can be realized if during the quenching phase of the heat treatment, the specimen has been converted to high content of martensite. The formation of any pearlite and/or bainite will result in other than the best combination of mechanical characteristics. During the quenching process, it is impossible to cool the specimen at a uniform rate throughout the sample. The surface cools more rapidly than the interior regions, resulting in a possible variation of microstructure and properties with position within a specimen. The successful annealing of steels to produce the predominantly martensitic microstructure throughout the cross section depends primarily on alloy composition, type and character of the quenching medium, and size and shape of the specimen.

3. ANNEALING OF SEMICONDUCTORS

Defects in semiconductors act as trapping sites for electron and hole transport, thereby significantly degrading the transport properties and device performance. Point defects such as vacancies and interstitials can also significantly modify the diffusion characteristics of various impurities and dopants in the materials. Various semiconductor processing operations such as ion implantation, oxidation, lithography, and metallization create defects as a part of the process. In particular, ion implantation for dopant incorporation introduces a high density of defects. These defects range from point defects, Frenkel pairs, and dislocations to formation of amorphous layers and injection of interstitials, each of which can adversely affect device performance. The stability of defects controls the requirements of temperature and time of the annealing process. For example, amorphous thin films on surfaces requires low thermal budgets for crystallization and defect removal, whereas grain boundaries cannot be completely eliminated even by high-temperature annealing for extended time. Another example of incomplete removal of defects after thermal annealing is the solid-phase thermal annealing of ion-implanted silicon, leading to formation of a high density of extended defects such as dislocation loops, etc., which remain even after other defects have been removed.

Another important semiconductor thermal-annealing operation is solid-phase epitaxial (SPE) growth. Epitaxy refers to the ordered arrangement of atoms so as to mimic the structure of the substrate layer. Semiconductor materials are typically single-crystal in nature; thus annealing operations such as SPE growth are conducted so that the surface layer can be transformed into a single-crystal form. The SPE process occurs when a thin amorphous layer exists at the near-surface regions of the semiconductor. When the sample is annealed in a furnace at a fixed temperature, the amorphous layers reorder on the underlying single-crystal substrate. The thickness of the regrown layer increases linearly with time, thus indicating a constant growth velocity (see Fig. 7). The velocity can vary from 1 Å/s to 1 cm/s, depending on the processing temperature. This epitaxial ordering process occurs at temperatures much below the melting temperature. This is in sharp contrast to liquid-phase epitaxy, which occurs above melting temperature. The epitaxial regrowth velocity v_g can be expressed in form of an equation:

$$v_g = v_0 \exp(-E_a/kT), \tag{1}$$

where v_0 is called the preexponential factor and corresponds to 3.68×10^8 cm/s for ⟨100⟩ regrowth of silicon semiconductor. The activation energy for silicon has been determined to be $E_a = 2.76$ eV using this formula. In the case of silicon, the growth velocities in other directions have always been found to be lower than the [100] directions. The activation energy is approximately half of what is observed for dopant implantation. Thus SPE growth takes place at much lower processing temperatures. During the epitaxial growth process, implanted dopants move into substitutional lattice sites as the interfaces move past them. A second annealing treatment at high temperatures of 900–1000 °C is required to remove the residual defects in the material.

3.1 Diffusion Effects during Thermal Annealing

Concomitant with the removal of defects in thermal annealing, the redistribution of impurities is also observed after any thermal process. Typically, foreign atoms (called dopants) are incorporated into the semiconductor material via an ion-implantation process. The dopants make the semiconductor electrically active for formation of a useful electronic circuit. Dopants can be of two types: *p*-type like boron, which are positive carriers, and *n*-type like phosphorus, which are negative carriers. During the thermal annealing operation, significant diffusion of the dopants takes place. The implanted ion concentration can be described by a Gaussian function given by

$$C(x) = \frac{Q}{\sqrt{2\pi}\Delta R_p} \exp\left[\frac{-(x - R_p)^2}{2\Delta R_p^2}\right], \tag{2}$$

where Q is the dose, R_p is the projected-ion range, and ΔR_p is the range straggling. During heat treatment for dopant activation, the species are redistributed and the concentration $C(x)$ can be approximated by

$$C(x) = \frac{Q}{\sqrt{\pi(2\Delta R_p^2 + 4Dt)}} \left[\exp\left(\frac{-(x + R_p)^2}{2\Delta R_p^2 + 4Dt}\right) + \exp\left(\frac{-(x - R_p)^2}{2\Delta R_p^2 + 4Dt}\right)\right], \tag{3}$$

where D corresponds to the diffusion coefficient and t corresponds to the time. It is possible to make a quick estimate of the influence of the diffusion during annealing by comparing ΔR_p to the mass diffusion distance $(Dt)^{1/2}$. Figure 8 shows the composition profile of boron-implanted species after rapid thermal annealing (RTA). The as-implanted samples contain only dislocation loops with

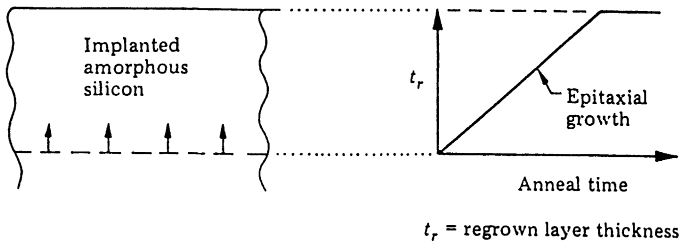

FIG. 7. Solid-phase epitaxial regrowth vs anneal time for an amorphous implanted layer on (100) Si.

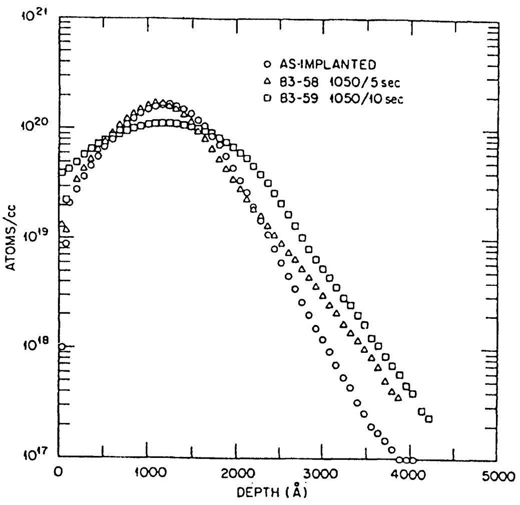

FIG. 8. Boron concentration profiles before and after annealing treatments for 5 and 10 at 1050 °C. (Adapted from Narayan *et al.*, 1984.)

a Gaussian profile peaked at 1180 Å. The figure shows that an average profile broadening of 100 Å is achieved after 1050 °C/(5 s) anneal, and increased to 330 Å after 1050 °C/(10 s) anneal.

3.2 Specialized Thermal Annealing Methods in Semiconductors

In semiconductor-based thermal annealing processes, elevated temperatures not only remove the damage but can also cause a number of unwanted effects such as dopant diffusion, thin-film reactions, etc. An optimum goal for any semiconductor annealing process is to remove the defects, while ensuring substitutional incorporation of the dopants so that a high degree of dopant activation and higher mobility of the carriers can be achieved. Besides obtaining these desirable properties, the annealing process should minimize unwanted effects such as dopant diffusion, which is especially critical for the next generation of VLSI circuits. In addition, in most semiconductors only the near-surface regions (<1 μm) represent the active device area; thus techniques that affect only the surface regions of a material can be applied. Section 3.2.1 deals with these different variations of standard furnace anneals that have been developed for semiconductor processing.

A schematic diagram of the processing temperature and the annealing times for each annealing method is shown in Fig. 9. Furnace annealing is typically conducted at temperatures of approximately 900 °C for several minutes, while in rapid thermal annealing (RTA) the wafer is typically heated to over 1000 °C for several seconds. In pulsed laser annealing (PLA), the maximum temperature is above the melting point of the silicon. However, the time scale for the whole process is less than 1 μs. Figure 10 shows different thermal methods for crystallization of amorphous layers in silicon. The figure shows that by use of either a furnace or a RTA system, solid-phase regrowth can be achieved, while in PLA the annealing phenomenon is accomplished by the formation of a thin melt layer on the surface. The depth of the melt controls the microstructure of the regrowth layer. If the melt depth is less than the thickness of the amorphous layer, polycrystalline surface regions are obtained. Table 1 lists the characteristics, advantages, and disadvantages of each of the annealing methods. As shown in this table, RTA represents the most promising method for systems that require low thermal budgets.

3.2.1 Furnace Annealing Furnace annealing has been the method of choice in the semiconductor industry since its inception. In this method the semiconductor wafers are placed in an isothermal cylindrical tube or box that is enclosed by an array of resistively heated elements. The temperature and the

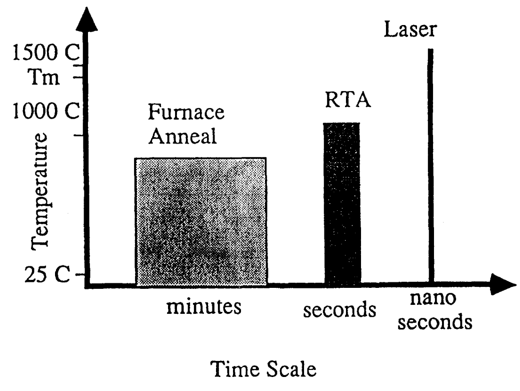

FIG. 9. Schematic diagram showing temperature and interaction time for various annealing processes.

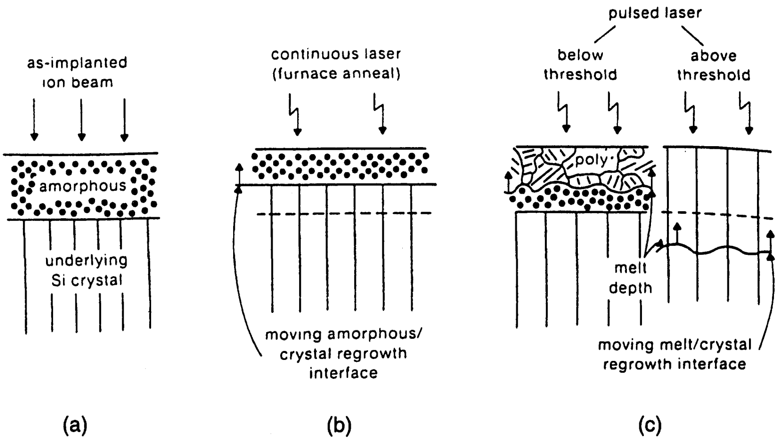
as-implanted
ion beam
amorphous
underlying
Si crystal
continuous laser
(furnace anneal)
moving amorphous/
crystal regrowth
interface
pulsed laser
below
threshold
above
threshold
poly
melt
depth
moving melt/crystal
regrowth interface
(a)
(b)
(c)

FIG. 10. Schematic representation of crystallization of amorphous layers in silicon using furnace and laser annealing.

ramp rate of the furnace can be precisely controlled. The near-equilibrium conductive, convective, and radiative heat transfer from the heating elements to the semiconductor ensures isothermal heating of the wafer with a high degree of precision in thermal uniformity. Because of the large thermal mass of the system, the ramp-up and ramp-down rates are severely limited. Another problem is that the relatively slow heat transfer results in a large time constant for the wafer to reach its equilibrium temperature. Thus furnace annealing is not very suitable for low–thermal-budget processing of semiconductors.

3.2.2 Rapid Thermal Annealing

Rapid thermal annealing (RTA) and pulsed laser annealing (PLA) both represent direct energy-transfer annealing methods. In both these methods optical energy is directly coupled with the semiconductor wafer, resulting in high ramp-up rates, thereby making them suitable for low–thermal-budget processing (Narayan *et al.,* 1985). Figure 11(a) shows a schematic diagram of a commercial RTA system manufactured by AG Associates. A bank of tungsten halogen lamps emitting incoherent radiation heats up a semiconductor wafer sitting in the chamber on quartz pins. The wavelength spread of this radiation is

Table 1. Characteristics of furnace, rapid thermal, and laser annealing.

	Furnace solid phase	RTA solid phase	PLA liquid phase
Temperature for Si processing	800–900 °C	1000–1100 °C	Above melting point
Time	Several minutes	~10 s	100–1000 ns
Coverage	Multiple wafer	Single wafer	Small area within the wafer
Dopant activation	Good	Very good	Very good
Defect reduction	Good	Very good	Good
Thermal gradient	Low	High	Highest
Implant redistribution	Significant dopant redistribution	Less dopant redistribution	Dopant redistribution only in melted region
Cost	Low	High	Highest
Critical technical barriers	Reduction in thermal budget	1. Control of temperature uniformity 2. Measurement of temperature	1. Large area uniformity 2. Reproductivity 3. Cost

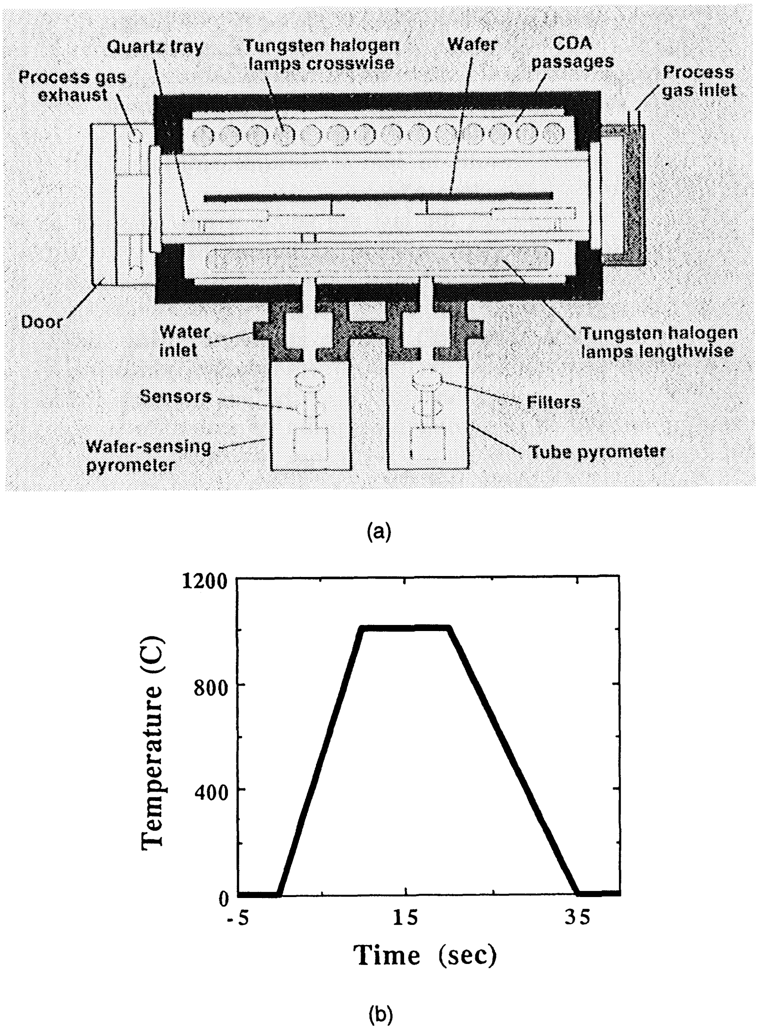
Quartz tray
Tungsten halogen lamps crosswise
Wafer
CDA passages
Process gas exhaust
Process gas inlet
Door
Water inlet
Tungsten halogen lamps lengthwise
Sensors
Filters
Wafer-sensing pyrometer
Tube pyrometer
1200
800
400
0
Temperature (C)
-5
15
35
Time (sec)

(a)

(b)

FIG. 11. Schematic of (a) rapid thermal processing system and (b) a typical RTA thermal cycle.

very broad, extending from the ultraviolet to the infrared regions. The peak of the wavelength depends on the operating temperature of the lamp (3000–3600 K). The electromagnetic radiation from the lamps is directly coupled with the silicon wafer, which heats up rapidly. It should be noted that the silicon at low temperatures (<600 °C) is transparent to the infrared component of the radiation. However, at higher temperatures the wafer becomes completely absorbing to the incident radiation. The chamber and the wafer surroundings are nonabsorbing to the lamp radiation. Thus the thermal mass of a RTA system is much smaller than that of a typical furnace annealing system.

A typical processing temperature profile of a RTA process is shown in Fig. 11(b). The wafer is rapidly heated (50–200 °C/s) to temperatures in the range of 900–1200 °C, kept under steady-state condition for several seconds, and then ramped-down back to room temperature. Although the RTA system can process materials in the order of 10 s, there are two critical system issues that have still prevented its widespread use in the semiconductor industry. These issues are related to

1. noncontact techniques to measure temperature and
2. temperature uniformity on the wafer.

Pyrometric techniques have been developed to measure the temperature; however, the presence of other radiation and other effects (dopants, surface roughness) can affect the accuracy of the pyrometer significantly. Presently, new methods such as ripple pyrometry and temperature-independent pyrometry are being developed to address some of these issues.

Another critical issue in RTA equipment is the temperature uniformity across the wafer. To ensure that no defects are generated by thermal stresses, and to achieve a high process yield, the temperature across the wafer should be very uniform (±5 °C) for all stages of the thermal cycle (ramp-up, steady state, and ramp-down). The temperature nonuniformities are primarily introduced by higher radiative losses at the wafer edges compared to its center. To achieve thermal uniformity the bank of the lamps is divided into several zones, each of which is controlled independently. The power to each zone is adjusted by a closed-loop system so that the wafer can be uniformly heated. In addition, slip-free rings are placed at the wafer edges so that thermal radiation from the edges can be considerably reduced. It should be noted that despite the sophisticated temperature control systems, the maintenance of thermal uniformity is still one of the outstanding critical issues in RTA systems.

3.2.3 Pulsed Laser Annealing Pulsed lasers have also been extensively applied in research for annealing of semiconductors. Typically in this method the laser-induced rapid heating causes melting of the surface with the propagation of the melt interface into the solid, followed by resolidification of the melt region. Because the pulse durations are extremely short, the thermal effects are over in less than a microsecond. Various types of lasers have been employed by researchers for laser annealing experiments. The wavelengths of the lasers range from far infrared (wavelength $\lambda = 10.6\ \mu m$) to deep ultraviolet ($\lambda = 193$–308 nm). In the mid-1980s most solid-state lasers such as ruby ($\lambda = 694$ nm) and frequency-doubled YAG ($\lambda = 532$ nm) were extensively used for the laser annealing process. However, the solid-state laser has certain limitations for high-throughout processing of large-area devices. Foremost among these limitations are spatial homogeneity, small processing area, and low pulse-repetition rates. In addition, the absorption depth of the infrared laser beam is quite large and strongly temperature dependent, which may lead to nonlinear effects. Finally, the high coherency of these lasers results in spatial nonuniformities due to diffraction effects.

In the last ten years, a new class of laser called excimer lasers has been used extensively for materials processing. A mixture of rare-gas halide excited dimers (ArF, XeF, and KrF) acts as an active lasing medium. These excited dimers are formed during electrical excitation of the gas mixture. Because of strong absorption of radiation in the ultraviolet (UV) range by most semiconductors, the UV wavelength of the excimer laser results in high coupling of the laser beam with a variety of semiconductors. In addition, recent advances in laser technology have made available large-area laser beam sources, which can be used for annealing. Excimer lasers also are characterized by low coherency, thus eliminating the diffraction-related spatial nonhomogeneity effects. It should be noted that despite tremendous advances in excimer laser technology, full-wafer single-shot processing is still not practical.

The mechanisms for removal of microstructural damage by laser annealing are based on surface melting and resolidification. Figure 12 shows TEM micrographs of the excimer-annealed regions of boron-implanted silicon. Each micrograph corresponds to a different laser intensity irradiation on the implanted surface. The implanted specimens contain a high concentration of defects up to a depth of 0.7 μm. As shown in the figure, the thickness of the annealed regions varies linearly with energy density. Laser irradiation of ion-implanted samples leads to the removal of implantation damage up to the maximum melt depth. A complete removal of the dislocation loops in the annealed regions together with the sharp transition between the annealed and the unannealed regions is consistent with the first-order melting transition. The V-shaped dislocations containing stacking faults are created because the melt front intersects the dislocation loops, and the two segments for the dislocation grow back to the surface because the dislocation cannot be terminated within a perfect lattice.

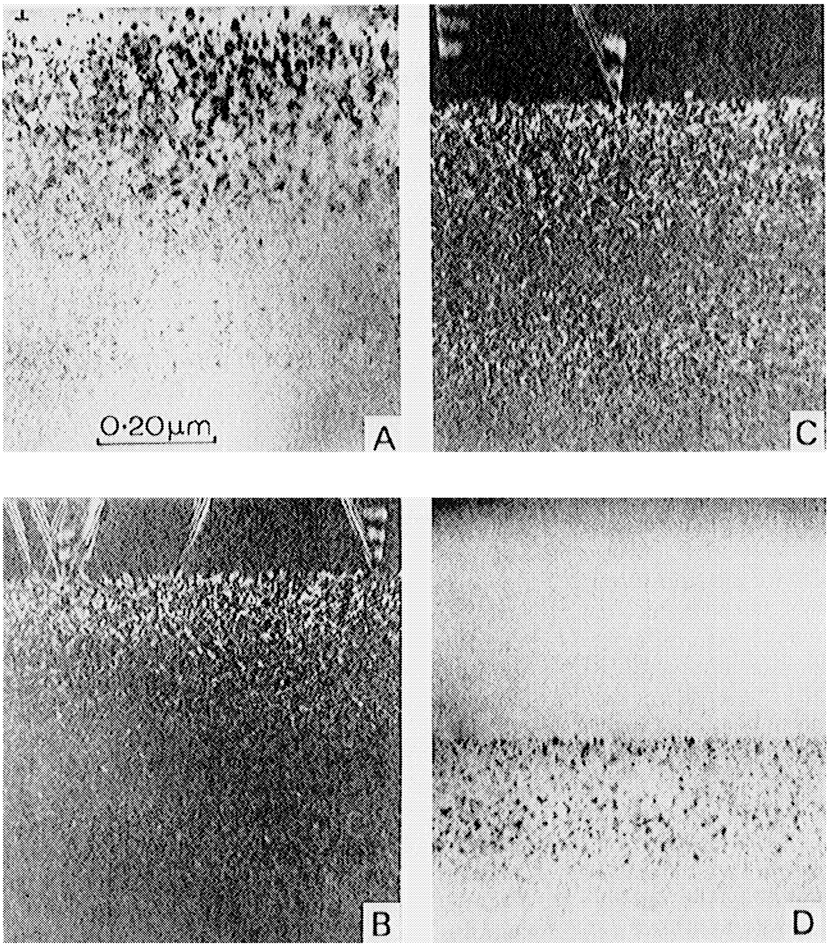

FIG. 12. TEM micrograph showing the depth of annealing or melting as a function of pulse energy density for 25- and 70-ns excimer laser pulses: (A) 1.0, (B) 1.5 J cm^{-2} (70 ns), (C) 1.0, and (D) 2.0 J cm^{-2} (25 ns). (Adapted from Singh and Narayan, 1989.)

The thickness of the melt region and the solidification velocities are controlled by the laser fluence and the pulse duration (Narayan and Holland, 1984). Figure 13 shows the position of the melt front with time for different pulse durations. From the curves it can be seen that the surface starts to melt slowly because of the sudden change in reflectivity of the materials. After the termination of the laser pulse, the melt front starts to recede back to the surface. At the initial stages of solidification, the solidification velocities are low but quickly reach a maximum when the steady-state curves are attained. The figure shows that as the pulse duration is increased, the maximum melt depth decreases accompanied with a concomitant decrease in the solidification velocity. From the slope of the solidification curves, the solidification velocities have been determined to be between 1 and 5 m/s.

Energy-balance considerations can be used to estimate the maximum melt depth during laser annealing of semiconductors (Singh and Narayan, 1989; Singh *et al.*, 1996). To determine the melt characteristics of the laser-irradiated materials, a threshold energy term E_{th} is defined, which corresponds to the energy that is lost by the solid part of the material undergoing laser irradi-

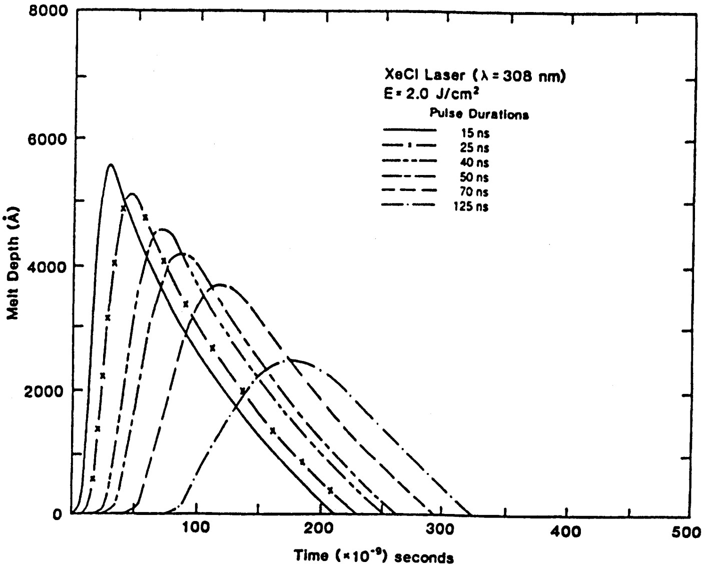

FIG. 13. Melt depth as a function of time for a 2.0-J-cm^{-2} excimer laser pulse irradiated on silicon for different pulse durations. (Adapted from Singh and Narayan, 1989.)

ation. Using energy balance considerations, the maximum melt depth, Δx, has been estimated to be

$$\Delta x = \frac{(1 - R_1)(E - E_{\text{th}})}{C_v T_m + L}. \tag{4}$$

The maximum melt depth, Δx, is plotted against incident fluence in Fig. 14. This figure shows a nearly linear relationship between the maximum melt depth and the incident fluence of the laser. Experimentally determined melt depths for various incident fluences are also shown in the same figure for comparison.

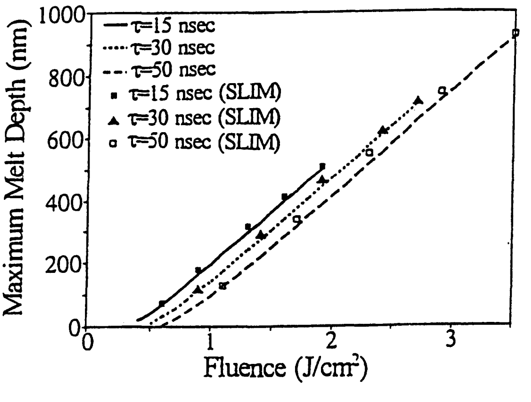

FIG. 14. Energy balance melt depth values for 15-, 30-, and 50-ns laser plotted with corresponding simulation-calculated values. (Adapted from Singh et al., 1996.)

3.3 Technology Trends in Specialized Annealing Methods

Furnace annealing to remove defects has been extensively used on an industrial scale since the development of VLSI technology. However, the reduction in device dimensions has been the principal driver for more advanced annealing techniques in semiconductor processing. In order to reduce device dimensions and maintain suitable electrical characteristics of devices, the vertical dimension of the device must be reduced along with the lateral dimensions. The reduction in lateral dimensions is normally achieved by lithography, while the reduction in vertical dimensions required the ability to form ultrashallow *p-n* junctions. It is expected that for the future of 0.15-μm devices the junction depths will be as shallow as 300 Å. Therefore, annealing techniques are required that not only can activate dopants like boron but also can ensure that the dopant diffusion is minimized. Nonequilibrium techniques such as rapid thermal annealing and laser annealing have been developed so that the microstructural damage due to dopant implantation can be removed while at the same time ensuring that the diffusion profiles are not significantly affected by the thermal anneals.

4. THERMAL ANNEALING OF OXIDE MATERIALS

4.1 High-Temperature Superconductors

High-temperature superconductors are copper-oxide–based compounds that were first discovered by Bednorz and Müller (1986). The structures of the copper oxide–based superconductors are in general classified in terms of well-known structural types. For example, the structure of $YBa_2Cu_3O_7$, which superconducts at 90 K, is related to that of perovskite with apical oxygen atoms removed from the structure, while the arrangement of Ba- or Sr-doped La_2CuO_4 is identical to that of K_2NiF_4. The superconductors belonging to the system $Bi_2Sr_2Ca_2Cu_3O_{10}$ (T_c = 110 K) and $Tl_2Ba_2Ca_2Cu_3O_{10}$ (T_c = 123 K) have been compared to Aurivillius phases. Synthesis of these materials can be straightforward since they have been prepared successfully from mixed oxides using standard solid-state reaction. However, in order to obtain optimum desirable properties, e.g., high density, high fraction of superconducting phase, sharper transition temperature, and high current density, an understanding of various processing parameters, including sintering and annealing conditions, is essential. After sintering of the mixture of constituent oxides to a dense body typically in the range 900–975 °C, cooling rate and annealing environment must be carefully controlled, since these factors have been shown to exercise important influences on structure and superconducting properties. A brief discussion on the effects of annealing on each class of high-T_c superconductors follows.

$YBa_2Cu_3O_{7-x}$ (YBCO) is probably the most important system among high-temperature superconductors because its single-phasic preparation is relatively easy and does not involve any poisonous or hazardous components. It was the first superconducting material that was found to show superconductivity above liquid-nitrogen temperature. The YBCO system in bulk form is prepared by sintering a homogenous and stoichiometric mixture of Y_2O_3, BaO or $BaCO_3$, and CuO at 900–950 °C. The sintered powder is pelletized and annealed in oxygen ambient to obtain a sample with good superconducting and structural properties. The annealing conditions have been proved to have a profound effect on the superconducting and structural properties of these materials (Shyu *et al.*, 1988). Shown in Fig. 15 are the plots of normalized resistance vs temperature for five YBCO bulk samples obtained by annealing them in O_2 at different temperatures and different times. As seen in the figure the sample annealed at 950 °C for 4 h has the highest T_c and the narrowest transition. Higher temperature and longer times raise T_c, but 930–950 °C is required for a sharp transition. This should be expected because above approximately 900 °C the composition becomes partially molten, allowing grain growth, densification, and sintering to occur more rapidly. To increase the oxygen content, the YBCO samples are further annealed at 400–500 °C for several hours (5 to 10 h) in oxygen ambient. The annealing procedure for YBCO films on different substrates differs from the procedure used for annealing bulk YBCO samples in terms of dwell time at high temperature. The high-temperature (920–930 °C) step in the processing of YBCO films is extremely short, 3 min at the maximum. Despite the high temperature needed to achieve desired orientation, the short processing time is more important to preclude significant interaction of the film with the substrate.

Studies by various researchers have shown that the phases of Bi-Sr-Ca-Cu-O films are strongly affected by the annealing

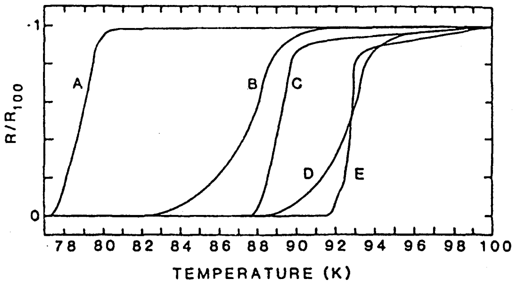

FIG. 15. Superconducting transitions for samples heat treated at (curve A) 890 °C, 30 min, (curve B) 910 °C, 4 h, (curve C) 930 °C, 4 h, (curve D) 910 °C, 23 h, and (curve E) 950 °C, 4 h.

temperature, time, ambient, and initial phases of the film. The annealing experiments show that the bismuth cuprate films melt at 865 °C in N_2, 875 °C in air, and 900 °C in oxygen. It has been observed that just melting the film is essential to obtain superconduction and for good adherence to the substrate. Table 2 shows the annealing conditions and conduction properties of Bi-Sr-Ca-Cu-O film on alumina substrates. The results presented in the table indicate that annealing at a low temperature is possible in nitrogen ambient. Since a complete annealing cycle in N_2 or O_2 does not produce superconduction in the film, a strategy of exposing the films in N_2 for a short time followed by air or air-oxygen flow in the furnace is employed. The advantages of adopting the N_2-air or N_2–air-O_2 sequence is that the film melts at a low temperature in nitrogen and subsequently solidifies at the same temperature when air or oxygen is introduced in the furnace. The film material absorbs the requisite amount of oxygen for superconduction when exposed to air or O_2 after melting in N_2. Thus the sequence of N_2–air-O_2 or N_2-air provides a kind of autocooling and perhaps is responsible for formation of the superconducting phase during regrowth of the melted film.

The preparation of thin films of Tl-based high-temperature superconductors is an important step to their potential applications in several devices. However, the fabrication of

Table 2. Annealing conditions and conduction properties of Bi-Sr-Ca-Cu-O films on alumina substrate.

Annealing conditions (temperature in °C)	Conduction property	T_c(K) $R = 0$
840–900/1–30 min N_2, slow or fast cool	Insulating or semiconducting	—
840–900/1–30 min O_2, slow or fast cool	Semiconducting up to 75 K then metallic	—
840/30 min air, fast cool	Semiconducting	—
875/30 s air, fast cool	Semiconducting	—
880/1 min + 825/15 min air, fast cool	Semiconducting	—
885/30 s + 865/4 min air, fast cool	Semiconducting	—
900/1 min air, slow up to 830 then fast cool	Metallic	—
875/1 min air, slow up to 830 then fast cool	Metallic	—
875/30 s air, slow up to 830 then fast cool	Metallic	—
860/30 min N_2,	Metallic	—
slow cool up to 700 air	Metallic	—
870/1 min + 840/15 min air, fast cool	Superconducting	69.5
870/1 min + 840/45 min air, fast cool	Superconducting	71
875 to 885/1 min air, slow cool up to 840, 840/30 min, fast cool	Superconducting	72–73
880/1 min N_2, slow cool up to 840 air 840/45 min air, fast cool	Superconducting	72
875/1 min N_2, slow cool up to 840 air 840/30 min air, fast cool	Superconducting	75
865/5 min N_2, 865/5 min air, 865/5 min O_2, slow up to 700 then fast cool	Superconducting	75

Tl-based superconducting films is very complicated because of the difficulty in maintaining proper stoichiometry, which arises from the volatile behavior of elemental Tl and its oxide compounds. One of the standard solutions is to deposit the stoichiometric films first at lower temperatures and then anneal the films at higher temperatures. The annealing at higher temperature is very crucial as the proper structure is realized during this step, not during the deposition at lower temperature. However, the low-temperature step provides the film with adequate adhesion and also influences the characteristics of subsequent annealing at high temperatures. Films deposited at higher temperatures tend to require a lower annealing temperature to achieve the highest T_c. Lee *et al.* (1989) obtained the Tl-based superconducting films with the highest T_c by annealing films deposited at ambient temperature and at 400, 895, and 805 °C for 1 h respectively. The highest T_c achieved tends to decrease with the increasing substrate temperature. For example, the highest T_c's observed were 121, 110, 103, 97, and 91 K for films deposited at ambient temperature and 200, 300, 400, and 500 °C, respectively. The four-point resistivity-versus-temperature data for the films deposited at ambient temperature and three at 300 °C, both post-annealed at 895 °C for 1 h, and for the film deposited at 400 °C and annealed at 805 °C for 1 h are shown in Fig. 16. Since these films have approximately the same starting composition before annealing, the different T_c's observed after annealing are likely related to microstructure-induced preferential growth of different superconducting phases during annealing. The lower-temperature annealing required for the films deposited at higher substrate temperatures is attributed to the film microstructure, which is more prone to material transport during annealing.

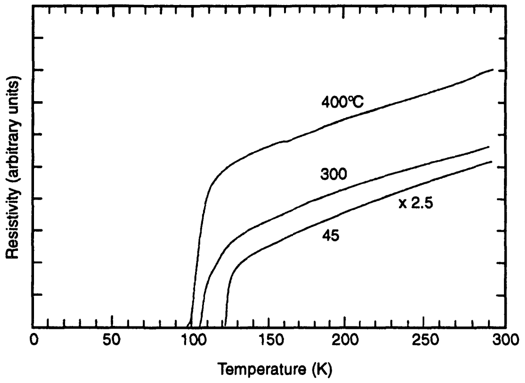

FIG. 16. Resistivity vs temperature for $Tl_2Ca_2Ba_2Cu_3O_{10}$ films sputter-deposited at ambient temperature and 300 °C both post-annealed at 895 °C for 1 h, and for the film deposited at 400 °C and annealed at 805 °C for 1 h.

4.2 Colossal-Magnetoresistive Materials

Magnetoresistance (MR), or the change in electrical resistance of a material in response to magnetic field, has a wide range of applications in digital information storage and retrieval. The MR phenomenon is the consequence of the ability of magnetic fields to change and thereby control the scattering of conduction electrons in metals so that the flow of electrical charge can be varied readily. The recent observation of colossal magnetoresistance (CMR) (up to 10^8 %) in perovskite lanthanum manganites $La_{1-x}M_xMnO_y$ (M = Ba, Ca, Sr) at high temperatures has generated further interest in the field (Jin *et al.*, 1994; Kumar *et al.*, 1996; Manoharan *et al.*, 1995). These CMR materials are both of technological interest due to their potential application as sensor materials and of scientific interest since the CMR mechanism is not yet fully understood, although explanations involving magnetic polaron scattering, double exchange, and metal–insulator transitions have been proposed. The MR ratio has been found to be strongly dependent on the microstructure of the oxide ceramic and the nature and amount of dopants in the lattice.

The CMR effect is more pronounced in epitaxial thin films and single-crystal materials. According to the literature, high-temperature post-annealing of CMR samples in oxygen or nitrogen ambient has been found to be beneficial in obtaining high CMR values. The post-annealing is critical in order to obtain high-quality films as this process raises metal–insulator transition temperature (T_c), drops resistivity by two or three orders of magnitude, and sharpens the magnetic signature of the ferromagnetic transition.

Shown in Fig. 17 are plots of MR ratio vs temperature for as-deposited and annealed $La_{0.6}Y_{0.07}Ca_{0.33}MnO_x$ (LYCMO) films. The as-deposited films were grown using pulsed la-

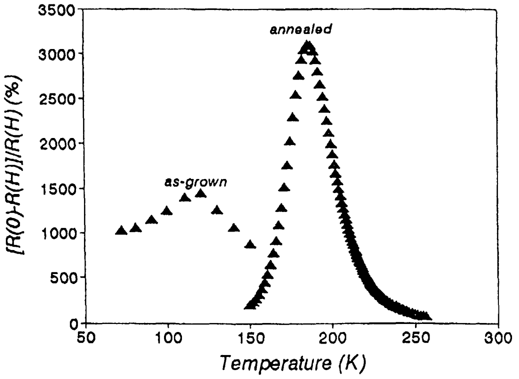

FIG. 17. The temperature dependence of MR ratio for as-deposited and 900 °C annealed films.

ser deposition (PLD) techniques in 300 mTorr of oxygen partial pressure and 650 °C substrate temperature. The same film, after the initial transport and magnetoresistance measurements, was annealed in oxygen for half an hour at 900 °C and characterized once again to see the effect of annealing. It is clear from this figure that a post-anneal process not only increases the metal–insulator transition temperature but also improves the MR ratio from 1500 to 3200%. Rutherford backscattering measurements indicate that the annealing process increases the oxygenation by 1–3%, whereas scanning tunneling microscopy (STM) and atomic force microscopy (AFM) reveal that the annealing process produces a tenfold increase in grain size. The improvement in the oxygenation process during post-annealing causes an increase in Mn^{4+}/Mn^{3+} ratio in the LYCMO films, which results in the realization of higher MR ratio at higher temperature. The enhancement in the grain size caused by the post-anneal process results in a decrease in the number of grain boundaries over a specified area. The grain growth is very important as it decreases the density of grain boundaries, which is responsible for the reduction in film resistivity. Therefore, annealing seems to be very important and advantageous for the realization of larger MR ratios at higher temperatures in CMR systems.

4.3 Ceramic Oxides

In ceramic oxides, such as the alkaline-earth oxide series (MgO, CaO, SrO, and BaO) and Al_2O_3, irradiation with energetic particles, such as neutrons, electrons, and ions, can produce lattice defects in the form of interstitials and vacancies in both the anion and cation sublattices. Annealing a ceramic oxide would generally return it to its initial virgin state except for vacancy mobility (Abraham *et al.*, 1976; Lee and Crawford, 1979; Halliburton *et al.*, 1973). The mobility of the interstitials and vacancies can be distinguished by performing experiments on two different types of crystals: those subjected to thermochemical reduction and those irradiated with energetic particles. Thermochemically reduced (TCR) crystals provide the means to determine vacancy mobilities, because they do not contain interstitials. They are prepared by heating the host crystals in a highly reducing atmosphere, such that oxygen ions in the crystal suffer stoichiometric deficiency, resulting in oxygen vacancies. The processing requirements can be stringent. For example, TCR of MgO crystals is performed at ~2000 K at ~7 atmospheres of magnesium vapor. The result of TCR is oxygen vacancies without oxygen interstitials. Thermal annealing experiments, monitored by optical absorption or EPR measurements, permit determination of the temperature at which these vacancies become mobile. In general, anion vacancies in ceramic oxides are mobile at temperatures in excess of 1400 K.

In the anion sublattice, stable anion interstitials and vacancies are produced. The anion vacancies annihilate at temperatures not much above room temperature. In MgO, CaO, and Al_2O_3, the temperatures at which anion vacancies begin to anneal are typically in the range of 400–600 K, well below the temperatures at which anion vacancies are known to be mobile (Lee and Crawford, 1977). Anion interstitials become mobile and recombine with the much more stable vacancies at these temperatures.

In simple oxides, such as MgO, CaO, and SrO, aliovalent impurities are always present. For example, MgO has the cubic rock-salt structure, and some of the defects can be represented in the following linear configurations (Lacy *et al.*, 1978; Unruh *et al.*, 1973):

$Mg^{2+}O^{2-}[Mg^{2+}$ vacancy$]O^{-}Mg^{2+}$,
V^{-} center, $E_A = 1.6$ eV;

$Al^{3+}O^{2-}[Mg^{2+}$ vacancy$]O^{-}Mg^{2+}$,
V_{Al} center, $E_A = 1.13$ eV;

$Mg^{2+}OH^{-}[Mg^{2+}$ vacancy$]O^{-}Mg^{2+}$,
V_{OH} center, $E_A = 0.84$ eV;

$Mg^{2+}O^{2-}Li^{+}O^{-}Mg^{2+}$, $[Li]^0$ center,
$E_A = 0.6$ eV,

where O^- represents a hole trapped at an O^{2-} ion site. The source of the holes can be from any ionizing radiation, such as electron irradiation, x rays, or gamma rays; these radiations create electron–hole pairs. Whereas the electrons are trapped by transition-metal impurities such as Fe ions, the holes are trapped at the V centers. The lithium impurity is not normally present in MgO crystals but can readily be doped in.

The holes are metastable at room temperature, varying from half-lives of order seconds in the $[Li]^0$ center, to a few hours in the V_{OH} and V_{Al} centers, to a few years in V^- centers. Given the different structures of the trapped-hole centers illustrated above, one would intuitively expect that the rate at which a hole thermally ionizes from a specific V center increases with the proximity of the O^- ion to the excess positive charge of the impurity. For example, the O^--to-H^+ distance is smaller than the O^--to-Al^{3+} distance. Therefore the stability of the hole should be less for the V_{OH} center. For the $[Li]^0$ center, the O^- ion is right next to the Li^+ ion, and the hole should be the least stable. Hence one would expect the order of increasing stability to be $[Li]^0$, V_{OH}, V_{Al}, and V^- centers. Indeed, this appears to be the case. The V^- center can also be regarded as a cation vacancy whose impurity charge compensator is at infinity.

The annealing of ceramic oxides may lead to recombination of anion vacancies and their subsequent migration to the grain boundaries or surfaces. The annealing may also cause recombination of interstitials with vacancies. While the former type of recombination is marked by a small increase in light scattering in the crystal, the latter type is marked by loss of displacement damage by the irradiating particles. Sometimes, the annealing may also bring about the release of holes from V-type centers, the holes recombining with electrons trapped earlier at the impurity sites such as Fe ions. The half-lives of the holes released are reflected in terms of experimentally obtained activation energies E_A showing a monotonic trend as mentioned above.

5. CONCLUSIONS

A brief review of the salient characteristics of furnace, rapid thermal, and laser annealing techniques is presented along with discussions focused on the roles they play during annealing of metals, semiconductors, and ceramic oxides. The annealing process consists of three stages, which involve heating the workpiece to the desired temperature, holding or soaking at that temperature, and cooling back to room temperature. The annealing treatment is primarily conducted to alter and produce specific microstructure. At elevated temperature, the defects or imperfections in solid materials, such as vacancies, interstitials, and dislocations, become mobile and thus can be tailored for specific requirements.

In metallic alloys, thermal annealing is carried out to relieve stresses, increase softness, ductility, and toughness, and/or produce a specific microstructure. A variety of annealing heat treatments in metals are possible, which are characterized by the changes that are induced. These changes are mostly related to microstructure, resulting in the alteration of the mechanical properties of metals or metal alloys. Annealing has been shown to play a critical role in semiconductors in the fabrication of very large-scale integrated (VLSI) circuits. The drive toward lower cost and larger-scale integration in electronic circuits has placed increasing demands on the development of annealing techniques that can remove the defects without significant diffusion of dopants. This has led to the development of newer thermal annealing techniques such as rapid thermal annealing (RTA) and pulsed laser annealing (PLA). Thermal annealing is also very useful for optimizing the desirable properties of high-temperature superconducting (HTS) and colossal magnetoresistive (CMR) oxide materials. It has been shown that annealing of bulk as well as thin-film high-temperature superconductors results in the realization of high density and high fraction of superconducting phase, sharper transition temperature, and high current density. The post-deposition annealing of colossal magnetoresistive films is also very critical in order to obtain high-quality films, as this process raises the metal–insulator transition temperature, drops resistivity drastically, and sharpens the magnetic signature of the ferromagnetic transition. Annealing of ceramic oxides has been found useful in the removal of lattice defects caused by irradiation with energetic particles and thereby returning the oxide lattice to its initial virgin state except for the vacancy mobility.

GLOSSARY

Austenites: Iron and steel alloys that have face-centered-cubic crystal structure.

Bainite: An austenitic transformation product found in some steels and cast irons.

Cold Working: The plastic deformation of a metal at temperature below that at which it recrystallizes.

Eutectoid Reaction: A reaction wherein, upon cooling, one solid phase transforms isothermally and reversibly into two new solid phases that are intimately mixed.

Martensite: A metastable iron phase supersaturated in carbon that is the product of a diffusionless transformation from austenite.

Pearlite: A two-phase microstructure found in some steels and cast irons that results from the transformation of austenite of eutectoid composition.

Shear: A force applied so as to cause or tend to cause two adjacent parts of the same body to slide relative to each other, in a direction parallel to their plane of contact.

Works Cited

Abraham, M. M., Chen, Y., Boater, L. A., Reynolds, R. W. (1976), "Stable [Li]^0 Defects in MgO Crystals," *Phys. Rev. Lett.* **37,** 849–852.

Bednorz, J. G., Müller, K. A. (1986), "Possible High T_c Superconductivity in Ba-La-Cu-O System," *Z. Phys. B* **64,** 189–193.

Clarebrough, L. M., Hargreaves, M. E., West, G. W. (1955), "The Release of Energy during Annealing of Deformed Metals," *Proc. R. Soc. London, Ser. A* **232,** 252–270.

Halliburton, L. E., Kappers, L. A., Cowen, D. L., Dravnicks, F., Wertz, J. E. (1973), "Isotropic ESR Line of the V^- Center in MgO at Room Temperature," *Phys. Rev. Lett.* **30,** 607–610.

Jin, S., Tiefel, T. H., McCromack, M., Fastnacht, R. A., Ramesh, R., Chen, L. H. (1994), "Thousandfold Change in Resistivity in Magnetoresistive La-Ca-Mn-O Films," *Science* **264,** 413.

Kumar, D., Kalyanraman, R., Narayan, J., Christen, D. K. (1996), "Giant Magnetoresistance Phenomenon in Laser Ablated La-Y-Ca-Mn-O Thin Films," in: R. Singh, D. Horton, L. Laude, J. Narayan, J. Cheung (Eds.), *Advanced Laser Processing of Materials—Fundamentals and Applications,* MRS Symposium Proceedings Vol. 397, Pittsburgh: Materials Research Society, pp. 241–245.

Lacy, J. B., Abraham, M. M., Boldu, J. L., Chen, Y., Narayan, J., Tohver, H. T. (1978), "Oxidation and Reduction of Lithium-Containing MgO at High Temperature," *Phys. Rev. B* **18,** 4136–4142.

Lee, K. H., Crawford, J. H., Jr. (1977), "Electron Centers in Single-Crystal Al_2O_3," *Phys. Rev. B* **15,** 4065–4070.

Lee, K. H., Crawford, J. H., Jr. (1979), "Luminescence of the F Center in Sapphire," *Phys. Rev. B* **19,** 3217–3221.

Lee, W. Y., Salem, J., Lee, V., Deline, V., Huang, T. G., Savoy, R., Duran, J., Sandstrom, R. L. (1989), "The Effect of Substrate Temperature on the Superconducting Properties of Tl-Ca-Ba-Cu-O Films Sputter-Deposited from Stoichiometric Oxide Targets," *Physica C* **160,** 511–516.

Lyman, T. (Ed.) (1948), *Metals Handbook,* Metals Park, OH: American Society for Metals.

Manoharan, S. S., Kumar, D., Hegde, M. S., Satyalakshami, K. M., Prasad, V., Subramanyam, S. V. (1995), "Giant Magnetoresistance in Se-Doped $La_{1-x}MnO_{3-\delta}$ Thin Films," *J. Solid State Chem.* **117,** 420–423.

Narayan, J., Holland, O. W., White, C. W., Young, R. T. (1984), "Excimer Laser Annealing of Ion-Implanted Silicon," *J. Appl. Phys.* **55,** 1125–1130.

Narayan, J., Holland, O. W. (1984), "Rapid Thermal Annealing of Ion-Implanted Semiconductors," *J. Appl. Phys.* **56,** 2913–2921.

Narayan, J., Holland, O. W., Christie, W. H., Wortman, J. J. (1985), "Rapid Thermal and Pulsed Laser Annealing of Boron Fluoride-Implanted Silicon," *J. Appl. Phys.* **57,** 2709–2716.

Shyu, S. G., Louh, R. F., Buchanan, R. C. (1988), "Annealing Effects on Properties of High-T_c Y-Ba-Cu-O Ceramics," in: M. B. Brodsky, R. C. Dynes, K. Kitazawa, H. L. Tuller (Eds.), *High Temperature Superconductors,* MRS Symposium Proceedings Vol. 99, Pittsburgh: Materials Research Society, pp. 655–658.

Singh, R. K., Narayan, J. (1989), "A Novel Method for Simulating Laser-Solid Interactions in Semiconductors and Layered Structures," *Mat. Sci. Eng. B* **3,** 217–230.

Singh, R. K., Gilbert, D. R., Viatella, J. (1996), "A Novel Method to Predict Laser-Induced, Nonlinear Thermal Effects in Semiconductors," *Mat. Sci. Eng. B* **40,** 89–95.

Unruh, W. P., Chen, Y., Abraham, M. M. (1973), "V_{Al} and V^- Centers in MgO", *Phys. Rev. Lett.* **30,** 446–449.

Further Reading

Askeland, D. R. (1989), *The Science and Engineering of Materials,* Boston: PWS-KENT Publishing Company.

Brodsky, M. B., Dynes, R. C., Kitazawa, K., Tuller, H. L. (Eds.) (1988), *High Temperature Superconductors,* MRS Symposium Proceedings Vol. 99, Pittsburgh: Materials Research Society.

Callister, W. D., Jr., (1994), *Materials Science and Engineering: An Introduction,* New York: Wiley.

Clarebrough, L. M., Hargreaves, M. E., West, G. W. (1955), "The Release of Energy during Annealing of Deformed Metals," *Proc. R. Soc. London, Ser. A* **232,** 252–270.

Fair, R. B. (Ed.) (1993), *Rapid Thermal Processing,* San Diego: Academic.

Narayan, J., Tan, T. Y. (1981), *Defects in Semiconductors,* Amsterdam: North-Holland.

Ohring, M. (1991), *The Materials Science of Thin Films,* San Diego: Academic.

Singh, R., Norton, D. P., Laude, L., Narayan, J., Cheung, J. (Eds.) (1996), *Advanced Laser Processing of Materials—Fundamentals and Applications,* MRS Symposium Proceedings Vol. 397, Pittsburgh: Materials Research Society.

THERMAL CONDUCTION IN METALLIC LIQUIDS

N. H. March, *University of Oxford, Oxford, England*

Introduction 89
1. **Direct Measurement** 90
2. **Relation to Electrical Conductivity** 90
2.1 Role of Relaxation Time 91
2.2 Current Density Related to Relaxation Time 91
2.3 Wiedemann–Franz Law: Relation between Thermal and Electrical Conductivity in a Metallic Conductor 92
3. **Kohlrausch Method of Measuring Ratio of Thermal to Electrical Conductivity** 92
3.1 Effect of Liquid–Solid Freezing Transition on Conductivities Separately and on Lorenz Number: Example of In 94
4. **Neutron Scattering and Dynamical Structure** 94
4.1 Influence of Thermal Conductivity on Dynamical Structure Factor 94
5. **Binary Liquid-Metal Alloys** 96
5.1 Role of Thermal Conduction in Neutron Scattering 96
5.2 Thermal Conduction in Relation to Ultrasonic Attenuation 97
6. **Technology of Liquid Metals and Alloys** 97
7. **Sources of Deviation from Wiedemann–Franz Law** 98
8. **Convection, Prandtl Number, and Instabilities** 99
9. **Summary** 100
Works Cited 100
Further Reading 101

INTRODUCTION

The transfer of heat from one point in a medium to another is an example of flow taking place under the influence of a "driving force," which in this case is a temperature gradient. Other transport properties will inevitably enter any treatment of thermal conduction in metallic liquids, the most important being the electrical conductivity. In this case the force is provided by the gradient of the electrical potential. We shall see below, from both theory and experiment, that there is an intimate relation between thermal conductivity λ and electrical conductivity σ, in metallic liquids such as molten Cu or Na. A third transport process that is invoked later on in this article is diffusion, the "driving force" this time being a concentration gradient.

Quantitative studies of these processes have a long history, and in particular in the case of thermal conduction, the flow of heat was shown by Fourier (1822) to be a linear function of the temperature gradient driving the flow. Ohm and Fick established analogous linear relations for electricity and for mass transport, respectively. Thus one writes a so-called phenomenological linear relationship for one dimension, x, say, as

$$J_q = -\lambda \frac{dT}{dx}, \tag{1}$$

where J_q is the heat flow per unit area per unit time, T is the temperature, and λ is the coefficient of thermal conductivity. We turn immediately to summarize the essence of one direct method of measuring the thermal conductivity.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. DIRECT MEASUREMENT

When a temperature gradient is applied to a liquid, heat transfer may take place by one or more of the following mechanisms:

1. pure heat conduction,
2. convection, which is the transfer of relatively large masses of liquid from the cold to the hot zone, or
3. radiation.

In practical setups, contribution **3** is normally small in comparison with the others and can usually be ignored. Heat transfer by convection may be comparable with, or larger than, heat transfer by conduction, and it is essential to minimize it in any determination of the thermal conductivity of liquid metals.

The experimental method used by Goldratt and Greenfield (1980) consists in measuring the temperature profile along a sample enclosed in a thin-walled stainless steel tube, which was coated with a high-emissivity paint and was *in vacuo*. The sample, enclosed within a constant-temperature environment, was heated from above and allowed to radiate freely. In the steady state, the radiation to the surrounding constant-temperature enclosure leads to a temperature gradient along the sample. The resulting temperature profile of the sample can then be measured with an infrared microscope traveling inside the vacuum along the length of the sample.

From this measured temperature profile, Goldratt and Greenfield (1978) set out a straightforward procedure to obtain the thermal conductivity $\lambda(T)$ as a function of temperature throughout the extent of the measured temperature profile. As these workers stress, this procedure provides the value of λ for each sample at its temperature T. In particular, they emphasize that they do not determine an average of $\lambda(T)$ over a range of temperature, nor need it be assumed that the temperature gradient is linear.

An important feature of the above method is that it allows an independent check to be made of the results, making possible the detection of systematic errors. By resetting the heater, one finds a different steady-state temperature profile, for which the temperature gradient at each temperature is different from that of the previous heater setting. Furthermore, any given temperature occurs at different locations along the sample for each of the two heater settings. It follows that accord between the two measurements of $\lambda(T)$, calculated from the two different heater settings, greatly reduces the possibility of systematic errors. As an example of the use of this technique, the results of Goldratt and Greenfield (1980) for liquid metal In are reported below, as an illustration of the effect of the freezing transition on especially the Lorenz number of this liquid metal.

In their particular experiment, the indium metal was 99.999% pure. Sample tubes, of 2 cm diameter and 15 cm length, were filled with molten indium and slowly cooled from below to prevent the formation of bubbles. Experimental checks were performed to ensure sample homogeneity as well as the absence of convection in the liquid.

Above, it has been pointed out that convection is a troublesome aspect in relation to direct measurement of thermal conductivity. This may be the reason why, in Cubbery (1979), if one selects liquid Ga, say, one finds a considerable spread of values at 350 K, the melting point of Ga being 302.9 K. The extremes of three values recorded have a ratio of 1.25. Because of this spread, λ has been measured directly by Okada and Ozoe (1992) in a convection enclosure. They remark that the measurement of thermal conductivity should be carried out for a shallow fluid layer. Their finding is that the thermal conductivity of Ga can be fitted by

$$\lambda_{\mathrm{Ga}} = -7.45 + 0.1256T \quad \mathrm{W/(m \cdot K)} \tag{2}$$

for $306\ \mathrm{K} \leq T \leq 312\ \mathrm{K}$. The errors $\delta\lambda$ decrease as the temperature is raised, from ±2.6 at 307.7 K to ±0.9 at 312.1 K.

2. RELATION TO ELECTRICAL CONDUCTIVITY

The electrical conductivity, loosely speaking, is a measure of the ease with which charge can be carried through the liquid metal, whereas with thermal conduction we are concerned, of course, with the transfer of thermal energy.

It turns out that one can adapt the classical kinetic theory of the heat conductivity of gases, as set out, for example, in the book by

Jeans (1940), for the present purposes. Taking this formula quite literally for a moment, one can write

$$\lambda = \tfrac{1}{3} c_v v l, \tag{3}$$

where c_v is the specific heat per unit volume, v is the appropriately averaged velocity, while l is the mean free path of the gaseous molecules.

Turning to the electrons in liquid metals, even at the melting temperature T_m the thermal energy $k_B T_m$, k_B being Boltzmann's constant, is only a small fraction of the Fermi energy E_F, which is typically some electron volts in simple metals. Thus, the specific heat per electron in the Fermi gas, which the conduction electrons in a simple liquid metal near freezing constitute, is the classical value got from the equipartition of energy, namely $(\frac{3}{2})k_B$ multiplied by the fraction of electrons near the Fermi surface, which is evidently of the order of $k_B T/E_F$. Therefore, with electron density denoted by n, the specific heat in Eq. (3) can be taken to be approximately $3nk_B^2T/2E_F$. The factors can be calculated exactly for a free-electron-gas model (see, for example, Ziman, 1960), with the result for the electronic specific heat c_{el}

$$c_{el} = \pi^2 n k_B^2 T / 2E_F. \tag{4}$$

The velocity v in the kinetic formula, by arguments essentially the same as above, can be replaced by the Fermi velocity v_F:

$$m v_F^2/2 = E_F. \tag{5}$$

Hence the electronic contribution to the thermal conductivity of the liquid metal can be written

$$\lambda_{el} = \pi^2 n k_B^2 T l / 3 m v_F. \tag{6}$$

But the relaxation time τ (see below for more detail) is in this case of degenerate electrons given by the simple formula

$$l = v_F \tau, \tag{7}$$

and hence one is led to the result

$$\lambda_{el} = \pi^2 n k_B^2 T \tau / 3m. \tag{8}$$

2.1 Role of Relaxation Time

Following the so-called Drude model, let us assume that an electron experiences a collision, resulting in an abrupt change in its velocity, with a probability per unit time that is written as $1/\tau$. This means that the probability of an electron undergoing a collision in any infinitesimal interval of time dt is dt/τ. The time τ is most frequently referred to as relaxation time, already used above, but some authors term it collision time or, occasionally, mean free time. This time τ plays a very basic role in the theory of electronic transport in metallic conductors such as molten metals under discussion here.

It then follows from the above that an electron selected at random at a chosen instant will, on average, travel for a time τ since its last collision. In the simplest form of the Drude model, the relaxation time τ is taken independent of the position or velocity of an electron. This turns out (compare Ashcroft and Mermin, 1976) to be a good assumption for many, though not for all, applications.

It is also assumed that electrons come to thermal equilibrium with their surroundings only through collisions. These collisions are taken to maintain local thermodynamic equilibrium in a very simple manner. Immediately after each collision, an electron is supposed to emerge with a velocity that is not related to its velocity immediately before the collision, but randomly directed. The time τ in a liquid metal is typically 10^{-14} to 10^{-15} s.

2.2 Current Density Related to Relaxation Time

Let us now introduce the current density $\mathbf{j}$. It is a vector parallel to the flow of charge, whose magnitude is the amount of charge per unit time crossing a unit area perpendicular to the flow.

At any point in a liquid metal, electrons, as mentioned above, are always moving in a variety of directions. If the number of electrons per unit volume is denoted by n, the electronic charge by $-e$, then if $\mathbf{v}$ is the average electronic velocity, the net current density is given by

$$\mathbf{j} = -ne\mathbf{v}. \tag{9}$$

Without the application of an electric field, electrons are just as likely to be moving in any one direction as in any other. Thus the velocity $\mathbf{v}$ will average to zero and there is no net flow of current. But when an electric field $\mathbf{F}$ is applied, there will be a mean electronic velocity directed in the opposite sense to the field, the electronic charge being negative.

The next step is to relate the velocity in the presence of the field $\mathbf{F}$ to the relaxation time τ. This can be achieved as follows. Let us focus on a typical electron at time $t = 0$. Let t be that time that has elapsed since it last experienced a collision. Its velocity at time $t = 0$ will be that, say $\mathbf{v}_0$, immediately after that collision, plus the added velocity $-e\mathbf{F}t/m$ it has acquired in the field $\mathbf{F}$ since that last collision. Since it is assumed that an electron emerges from a collision in a random direction, there will be no contribution from $\mathbf{v}_0$ to the average electronic velocity: This will only depend on the average of $-e\mathbf{F}t/m$. But the average of t is the relaxation time τ, and therefore one can write

$$\mathbf{v} = -e\mathbf{F}\tau/m, \tag{10}$$

and hence it follows from Eq. (9) that the current density $\mathbf{j}$ may be written in the form

$$\mathbf{j} = -(ne^2\tau/m)\mathbf{F}. \tag{11}$$

But the electrical conductivity σ is introduced through

$$\mathbf{j} = \sigma\mathbf{F}, \tag{12}$$

the electrical resistivity ρ being the reciprocal of σ. From Eqs. (11) and (12), one arrives at the basic formula for the electrical conductivity of a liquid metal, namely

$$\sigma = ne^2\tau/m. \tag{13}$$

2.3 Wiedemann–Franz Law: Relation between Thermal and Electrical Conductivity in a Metallic Conductor

Let us now combine Eq. (8) for the electronic contribution to the thermal conductivity of a liquid metal with Eq. (13) for the electrical conductivity. Then one can readily show that

$$\frac{\lambda_{\mathrm{el}}}{\sigma T} = \frac{\pi^2}{3}\left(\frac{k_{\mathrm{B}}}{e}\right)^2 = \text{const}, \tag{14}$$

where the constant, usually denoted by L, is called the Lorenz number and has a value of 2.45×10^{-8} W·Ω·K^2.

Before discussing the relation of the Wiedemann–Franz law (14) to experiment, we summarize below the way the ratio L involving the thermal conductivity of metals can be directly measured (see also the review by Powell, 1969).

3. KOHLRAUSCH METHOD OF MEASURING RATIO OF THERMAL TO ELECTRICAL CONDUCTIVITY

We began this account of thermal conduction with a direct measuring technique for liquid metals. But it is much simpler to measure the electrical conductivity of a liquid metal than the thermal conductivity. Hence, if the Wiedemann–Franz law were fully quantitative, it would enable the thermal conductivity to be got directly from the measurement of electrical conductivity, provided the electronic contribution dominated the total thermal conductivity. While this is most often the case, insulating liquids exhibit thermal conduction, as discussed at length in THERMAL CONDUCTIVITY OF NONMETALLIC LIQUIDS.

However, before we embark on the question of "deviations" from the Wiedemann–Franz law, we note that this relation between λ and σ has motivated measurements of the ratio directly, and to such an approach we now turn. The method goes back to Kohlrausch (1900, 1960) and has been utilized by Busch and co-workers (1972). The method was originally designed by Kohlrausch (1900) for measurements on solids. It has been employed specifically by Busch *et al.* (1972) to make measurements on both solid and liquid Ga, and also on both phases of Hg. Their approach and results form the basis for what follows.

Reference may be made to Fig. 1 to see the way in which Busch *et al.* extended the Kohlrausch method to liquids; the constriction-shaped sample depicted there schematically was used. Given the assumption that the Wiedemann–Franz law holds, the Lorenz number is given by

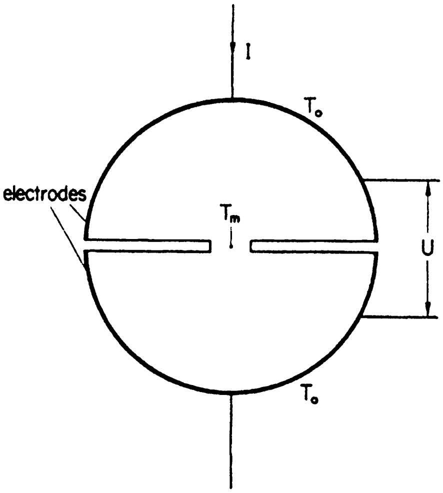

FIG. 1. Schematic representation of the constriction-shaped Ga sample used in extension of Kohlrausch method to liquid metals by Busch *et al.* (1972).

$$L = \tfrac{1}{4}\, U^2/(T_{cc}^2 - T_0^2). \qquad (15)$$

Here U is the voltage across the constriction, T_0 is the temperature, which is held constant, of the electrodes, while T_{cc} is the temperature in the center of the constriction. As already mentioned above, for the experiments on the liquid state using this technique, special attention needs to be paid to eliminate the influence of heat convection and other disturbing influences. Table 1 collects the data on the Lorenz number for solid and liquid Ga near the melting temperature $T_m = 303$ K. The material used was of high purity (99.999%), as obtained by Busch *et al.* (1972). These workers estimate that their experimental errors are less than 0.2%, and they find that their measurements show deviations from the theoretical value of the Lorenz number L in Eq. (14) of less than 1%. Subsequently, Busch *et al.* (1972) have made similar measurements to those just reported on Ga on solid and liquid mercury. These were made over the temperature range 225–350 K, the melting point being 234.8 K. Table 2 shows the measured Lorenz number L for these two phases of mercury, for the same purity as for Ga above. Busch *et al.* note that the experimental error in L is somewhat larger than for Ga for a variety of reasons.

To conclude this section, it is clear that for these two materials, the most accurate way at the time of writing to determine absolutely the thermal conductivity λ is via the Kohlrausch-type measurement of the Lorenz number, which is then to be combined with a direct measurement of the electrical conductivity σ. For the two materials Ga and Hg so far studied in this way, the theoretical estimate of the Lorenz number is verified to an accuracy of the order of the experimental error in the measurement of L. Nevertheless, there are theoretical principles that allow deviations from the theoretical estimate of the Lorenz number in Eq. (14), and we shall briefly summarize these in Sec. 7. Before leaving the topic of Lorenz number, we record that the nearly-free-electron (NFE) theory of electrical resistivity of liquid metals initiated by Krishnan and Bhatia (1945) and brought to full fruition by Ziman (1960) using pseudopotentials affords for simple *s-p* metals such as Na an almost quantitative route to calculate the electrical resistivity at the freezing point. However, it is known for the heavy alkali metals, Rb and Cs, that this theory will not work as liquid metals are taken up toward the critical point along the liquid–vapor coexistence curve (see, for example, Ascough *et al.*, 1988, for liquid metal Rb). This theory of electrical resistivity is posed in terms of the static liquid structure factor $S(q)$ and, as already mentioned, a suit-

Table 1. Lorenz number of solid and liquid Ga (near the melting point 303 K). After Busch *et al.* (1972).

Temp. (K)	L (V^2 deg^{-2})	Temp. (K)	L (V^2 deg^{-2})
293.3	2.447×10^{-8}	307.3	2.463×10^{-8}
294.7	2.441×10^{-8}	310.3	2.457×10^{-8}
302.4	2.443×10^{-8}	313.3	2.441×10^{-8}
304.2	2.440×10^{-8}	348.3	2.442×10^{-8}

Table 2. Lorenz number of solid and liquid Hg (near the melting point 234.8 K). After Busch *et al.* (1972).

Temp. (K)	L (V^2 deg^{-2})	Temp. (K)	L (V^2 deg^{-2})
227.7	2.440×10^{-8}	282.2	2.457×10^{-8}
231.5	2.452×10^{-8}	291.2	2.423×10^{-8}
234.9	2.443×10^{-8}	304.8	2.441×10^{-8}
242.4	2.424×10^{-8}	321.6	2.440×10^{-8}
273.0	2.467×10^{-8}	339.1	2.454×10^{-8}

able (screened) pseudopotential—see, for instance, Faber (1972), and also March (1990).

3.1 Effect of Liquid–Solid Freezing Transition on Conductivities Separately and on Lorenz Number: Example of In

Goldratt and Greenfield (1980) have obtained experimental data for thermal conductivity λ and Lorenz number L for metallic In on either side of the melting point. In Fig. 2 the thermal conductivity in the two phases is plotted, the vertical dashed line marking the melting point. Evidently, there is a very marked increase in the measured thermal conductivity as the liquid metal freezes, by something approaching a factor of 2. But as Fig. 3 demonstrates, there is a relatively mild variation only in the Lorenz number as one passes through the freezing point. The horizontal line in Fig. 3 shows the theoretical prediction of L in Eq. (14). To conclude this section, Table 3 records for a variety of liquid metals measured values of the Lorenz number L, compiled from Faber (1972).

4. NEUTRON SCATTERING AND DYNAMICAL STRUCTURE

So far, we have not discussed structural properties beyond reference to the static structure factor $S(q)$, which entered the NFE theory of electrical resistivity as one of the crucial ingredients (see Sec. 3). Since thermal conductivity is a transport property, it enters the equations of hydrodynamics for a liquid metal (see, for example, March, 1990), or a binary liquid-metal alloy, both of these systems being considered below. It is natural to begin with a pure liquid metal.

4.1 Influence of Thermal Conductivity on Dynamical Structure Factor

We focus below on the way that thermal conduction is reflected in the dynamical structure factor $S(q,\omega)$ of a liquid metal. The physical meaning of this quantity, introduced into the theory of metals by Van Hove (1954), is that it represents the probability that a neutron incident on the liquid will transfer momentum $\hbar q$ and energy $\hbar\omega$ to the liquid. One immediate property of $S(q,\omega)$ is that, when it is integrated over all energy transfers, the static structure factor $S(q)$ mentioned above is regained:

$$\int_{-\infty}^{\infty} S(q,\omega)d\omega = S(q). \tag{16}$$

The hydrodynamic equations have already been referred to. These, in suitable linearized form, are invoked below to set up the form of $S(q,\omega)$ in the appropriate regime in which the hydrodynamic equations have validity. These equations are, in fact, valid in the long-wavelength and low-frequency limits, or, equivalently, in the Fourier-transform

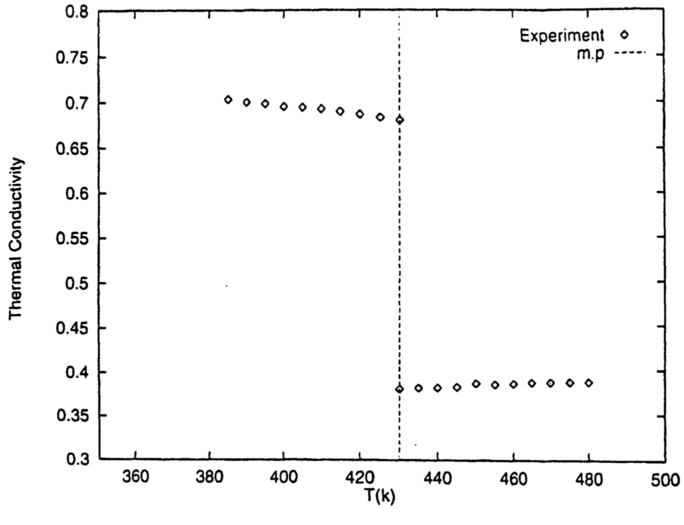

FIG. 2. Variation of thermal conductivity of indium with temperature as measured by Goldratt and Greenfield (1980).

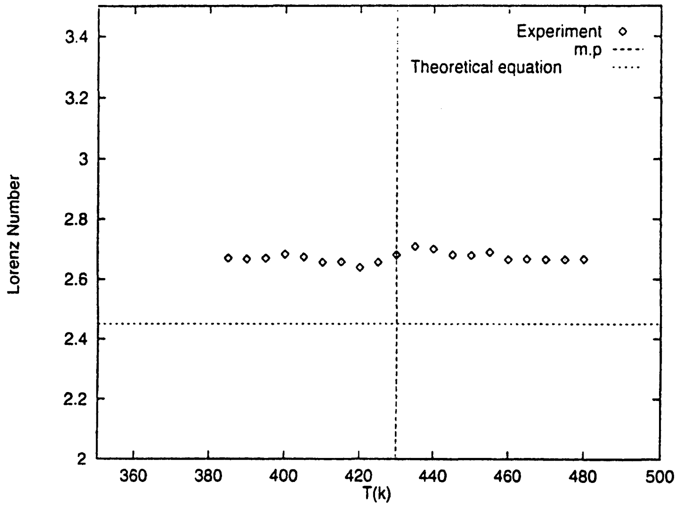

FIG. 3. Variation of Lorenz number of indium with temperature as measured by Goldratt and Greenfield (1980).

variables $\mathbf{q} \to \mathbf{r}$, $\omega \to t$, the limit of large r and long time t. Following the pioneering work of Landau and Placzek (1934), the usual procedure of such a calculation is to use the linearized hydrodynamic equations of irreversible thermodynamics to describe the relaxation in time of relevant fluctuating thermodynamic variables in the liquid; this procedure is reviewed, for example, by Mountain (1970).

Though more detail will be reported below for liquid-metal alloys, let us summarize here the main physical content emerging from such an approach. There are basically three peaks evident in $S(q,\omega)$ as a function of angular frequency ω. The central ($\omega = 0$) component (Rayleigh peak) is found to consist of the sum of two Lorentzians, the width of one of these being controlled largely by thermal conduction [compare Eq. (21) below]. There are then two "Doppler-shifted" peaks, referred to as Brillouin components, centered at frequencies $\omega = \pm C_0 q$ where C_0 is the adiabatic speed of sound given by

$$C_0 = \left(\frac{\partial P}{\partial \rho}\right)_S^{1/2}, \tag{17}$$

where S is the entropy. These peaks arise from sound propagation in opposite directions (for a given wave vector $\mathbf{q}$), and their width, having the form Γq^2, turns out to be controlled by acoustic attenuation (see also below). It is noteworthy that the ratio of the integrated intensity I_0 of the Rayleigh components in $S(q,\omega)$ to that of the doublet Brillouin side peaks, $2I_1$, is quite generally given by

$$I_0/2I_1 = \gamma - 1. \tag{18}$$

Table 3. Lorenz numbers $L \times 10^8$ in units of W Ω K^{-2} for a variety of liquid metals (Faber, 1972; see also Tankeshwar and March, 1996).

Metal	$L \times 10^8$	Metal	$L \times 10^8$	Metal	$L \times 10^8$
Li	2.6	Pb	2.4	Ta	2.4
Na	2.2	Sb	2.6	Re	1.75
K	2.1	Bi	2.5	Os	1.75
Cs	2.4	Ti	2.9	Pt	2.3
Cd	2.5	Zr	2.25	In	2.7
Hg	2.75	Hf	2.7	La	2.65
Zn	3.2	Mo	2.6	Ce	2.56
Al	2.4	W	2.5	Pr	2.89
Ga	2.07	Ru	2.45	Nd	2.27
Tl	3.2	Ir	1.95	Gd	1.83
Sn	2.9	Nb	2.6	Dy	2.34

The generalization of this result to liquid-metal alloys in which there are size differences between the two components is treated later.

There are also "non-Lorentzian" parts in $S(q,\omega)$, though these make no contribution to the integrated intensity and, usually, only a relatively small contribution to the intensity distribution. However, if $\Gamma q/C_0$ is not much less than unity, they can markedly affect the position and shape of the Brillouin peaks in $S(q,\omega)$.

5. BINARY LIQUID-METAL ALLOYS

Turning from pure metals to binary liquid alloys, we note first that Bhatia and Thornton (1970) introduced number–concentration (N-C) dynamical structure factors $S_{NN}(q,\omega)$, $S_{NC}(q,\omega)$, and $S_{CC}(q,\omega)$ to describe the structural aspects of scattering processes in such alloys.

5.1 Role of Thermal Conduction in Neutron Scattering

The aim in what follows is to discuss the hydrodynamic limit of long wavelength and low frequency of the N-C dynamical structure factors of such a liquid alloy. This follows the approach of Bhatia *et al.* (1974), who made extensive use of the earlier study of Cohen *et al.* (1971). These workers gave results for correlations between mass density and mass concentration. Let us first consider $S_{NN}(q,\omega)$. If we take the limit of a dilute alloy such that the concentration c goes to zero, then $S_{NN}(q,\omega)$ becomes the dynamical structure factor of pure liquid metal discussed above. Paralleling the discussion of $S(q,\omega)$, the Rayleigh component in $S_{NN}(q,\omega)$ is again found to consist generally of the sum of two Lorentzians. As before, the width of one Lorentzian is largely controlled by thermal conduction, the other by mutual diffusion. The two Doppler-shifted (Brillouin) components are largely as in the one-component case [see Bhatia *et al.* (1974) for fuller details of $S_{NN}(q,\omega)$ in the hydrodynamic regime]. It is worth noting that the formula for the ratio of the integrated intensity I_0 of the Rayleigh components in $S_{NN}(q,\omega)$ to that of the doublet Brillouin side peaks, $2I_1$, is modified from the result of Eq. (18). Though still proportional to $\gamma - 1$, with γ of course the specific heat ratio at the particular concentration considered in the liquid alloy, there is a multiplying factor of the form $1 + \text{const.} \times \delta^2$, where δ measures the size difference of the two components in the alloy:

$$\delta = \frac{1}{V}\left[\frac{\partial V}{\partial c}\right]_{P,T;N} = \frac{v_1 - v_2}{v}. \tag{19}$$

Here $v_\alpha = [\partial V/\partial N]_{P,T,N,\beta}(\beta \neq \alpha)$ denotes the partial molar volume, per atom, of species α, while v, the mean molar volume per atom, is given by

$$v = \left[\frac{\partial V}{\partial N}\right]_{P,T,c} = cv_1 + (1 - c)v_2. \tag{20}$$

Of course, in the dilute-alloy limit c tends to zero, whereas S_{NN} tends to the one-component $S(q,\omega)$, and S_{NC} and S_{CC} tend to zero. These two quantities are therefore crucially about the alloy at finite concentration c. Let us write out the essence of the form of S_{NC}, without going into full details (for these, see Bhatia *et al.*, 1974). It is such that

$$\begin{aligned} S_{NC}(q,\omega) = \frac{Nk_BT}{2\pi}\Bigg[&\frac{K_1q^2}{\omega^2 + X^2q^4} + \frac{K_2q^2}{\omega^2 + Y^2q^4} \\ &+ K_3\left(\frac{\omega + C_0q}{(\omega + C_0q)^2 + \Gamma^2q^4} - \frac{\omega - C_0q}{(\omega - C_0q)^2 + \Gamma^2q^4}\right)\Bigg]. \end{aligned} \tag{21}$$

In this general expression, the constants are all known in terms of macroscopic constants, both thermodynamic and irreversible transport coefficients being involved. The two Lorentzians (terms involving K_1 and K_2) primarily arise in the number–concentration structure factor $S_{NC}(q,\omega)$ respectively from diffusive processes and the coupling between this type of mode and the thermal conductive mode. If this coupling is zero, then the latter component is no longer present and the width of the remaining Lorentzian is simply Dq^2, where D is the coefficient of mutual diffusion.

The concentration–concentration structure factor is of the essence in understanding the thermodynamic properties of liquid binary alloys, when integrated over all energy transfers ω. As to its frequency dependence, this again follows from the linearized hydrodynamic equations and has the form of the sum of two Lorentzians:

$$S_{\mathrm{CC}} = \frac{Nk_{\mathrm{B}}T}{2\pi Z}\left[\frac{K_3q^2}{\omega^2 + X^2q^4} + \frac{K_4q^2}{\omega^2 + Y^2q^4}\right]. \tag{22}$$

These two components behave in a similar manner to the Lorentzians in $S_{\mathrm{NC}}(q,\omega)$. The quantity Z appearing in Eq. (22) is related to the Gibbs free energy G of the liquid alloy by

$$Z = \left(\frac{\partial^2 G}{\partial c^2}\right)_{P,T,N}. \tag{23}$$

5.2 Thermal Conduction in Relation to Ultrasonic Attenuation

The way in which thermal conduction affects sound-wave attenuation in liquid-metal alloys is considered here. Specifically, the width Γq^2 of the Brillouin components in the number–number dynamical structure factor $S_{\mathrm{NN}}(q,\omega)$ is related to the acoustic attenuation in the liquid. Explicitly, the amplitude attenuation per wavelength, α, is given by

$$\frac{\alpha}{\omega} = \frac{\pi[\zeta + (4/3)\eta]}{\rho C_0^2} + \frac{\pi(\gamma - 1)\lambda M}{\rho C_0^2 C_p} + \left[\frac{\alpha}{\omega}\right]_D, \tag{24}$$

where

$$\left[\frac{\alpha}{\omega}\right]_D = \frac{\pi MD}{Z}\left[\delta - \delta_m + \frac{Zk_T\alpha_T}{C_p}\right]^2. \tag{25}$$

The first two terms appearing in Eq. (24) come from

1. bulk (ζ) and shear (η) viscosities, and
2. attenuation caused by thermal conductivity, our main interest here,

while the last term gives the contribution due to diffusion. Bhatia *et al.* (1974) consider the limiting case of Eqs. (24) and (25) for an ideal-gas mixture. Here there is no size effect, so $\delta = 0$, the coefficient of thermal expansion $\alpha_T = 1/T$, while the isothermal compressibility $K_T = 1/P$, as follows from the ideal-gas equation of state $PV = Nk_{\mathrm{B}}T$. From the Gibbs free energy in this ideal-gas model, it is readily shown that Z is given by

$$Z = Nk_{\mathrm{B}}T/c(1 - c). \tag{26}$$

Hence, for the last term one readily finds

$$\left[\frac{\alpha}{\omega}\right]_D = \pi Dc(1 - c)\frac{\rho}{P}\left[\frac{m_2 - m_1}{m} + \frac{(\gamma - 1)\beta}{\gamma}\right]^2, \tag{27}$$

where $\beta = k_T/c(1 - c)$, the thermal diffusion factor, has been introduced. This formula has been derived previously from kinetic theory calculations and has been verified experimentally for noble-gas mixtures of various compositions. For a fuller discussion of ultrasonic absorption, and in particular the role of thermal conduction, the reader may consult the book by Bhatia (1967).

6. TECHNOLOGY OF LIQUID METALS AND ALLOYS

Some of the motivation for much research on liquid metals and their alloys has come directly from the standpoint of practical engineering utility. This is discussed at length in the book by Sittig (1956), with the focus on sodium and its alloys in the liquid phase. It is worth noting, because of the very important role of sodium, the comparison of its thermal conductivity in the liquid state with the other alkalis. Then, Na data will be presented in more detail. In units of cal/[cm^2·s·(K/cm)], the liquid-state values of thermal conductivity λ are for Li, 0.09; for Na, 0.18; for K, 0.10; and for Rb, 0.075. Sodium is seen to be almost twice as favorable as the other alkali liquid metals in this respect. A detailed survey of sodium and sodium–potassium alloy (NaK) has been presented by Jackson and his collaborators (1955). A treatment of the design and performance of liquid-metal (sodium and NaK) heat exchangers and steam generators has been given by Brooks and Rosenblatt (1953). As Sittig (1956) emphasizes, sodium and its alloys have many advantages as heat-transfer media. Besides their heat-transfer coefficients being among the highest known, they are very stable in that they do not normally polymerize, carbonize, or decompose. Table 1 in Sittig's book summarizes some of the significant properties of sodium and other high-temperature heat-transfer fluids. Rather than reproduce that quantitative detail here, we list below an extensive variety of practical engineering uses of Na, which are referred to by Sittig (1956, Chap. 5).

1. Heat-transfer medium in sodium-cooled valves for aircraft engines.
2. Heat-transfer media for control of casting-head temperature for die-casting machines.
3. Shale-oil retort—solved industrial heat-transfer problem.
4. High-temperature reaction kettles.
5. Control of glass mold temperatures.
6. Nuclear power plants. Advantages of liquid Na include, in addition to high heat-transfer coefficient, (a) fairly low neutron-absorption cross section, (b) high boiling point, (c) lack of attack on ordinary metals at high temperatures, and (d) extreme thermal stability.

Table 4 collects quantitative data from two different experimental sources on the thermal conductivity of Na in the liquid state as a function of temperature. A simple linear formula [compare Eq. (1) for liquid Ga] quoted in the footnote of Table 4 gives a quantitative fit of both sets of experimental data.

7. SOURCES OF DEVIATION FROM WIEDEMANN–FRANZ LAW

Although the measurement of Lorenz number by the Kohlrausch method, discussed in Sec. 3, shows for Ga and for Hg that deviations from the theoretical value of Eq. (14) are barely outside experimental error, there are certainly cases where this is not true in general (see, for instance, Tankeshwar and March, 1996). Thus it is of interest here to summarize briefly some theoretical work on the foundations of the Wiedemann–Franz law. This has a long history and for metals has been well reviewed by Chester and Thellung (1961) and by Ziman (1960). Though the discussions were oriented in the main toward solid metals, it is worth emphasizing at this point that the theoretical result for the Lorenz number L in Eq. (14) does not depend on such details as the density of electron states, the origin of the scattering, etc., but it is based on the assumption of elastic scattering. In liquid metals focussed on here, this is discussed, for instance, by Rice (1970). His conclusion (see also Egelstaff, 1992) was that only minor corrections to Eq. (14) for L were to be expected in molten metals from inelastic scattering of electrons.

Table 4. Thermal conductivity of liquid Na [cal/(cm^2) s (°C/cm)].

Temperature (°C)	Measured	Simple linear formula[a]
100	0.2055	0.2050
120	0.2039	0.2027
140	0.2019	0.2004
160	0.1994	0.1980
180	0.1966	0.1957
200	0.1932	0.1934
210	0.1914	0.1922
181.7	0.1984	0.1955
181.8	0.1993	0.1955
235.7	0.1888	0.1893
267.5	0.1847	0.1856
269.3	0.1847	0.1854
374.3	0.1728	0.1732
375.1	0.1730	0.1731
496.2	0.1601	0.1590
512.5	0.1587	0.1571

[a] Simple linear formula used is $\lambda = 0.2166 - 0.000116t$, with t in °C (see book by Sittig, 1956).

However, Tankeshwar and March (1996) have plotted the Lorenz number versus atomic number for some 30 metals, and their results are reproduced in Fig. 4. Though we have cautioned earlier about the difficulties of direct measurement of $\lambda(T)$, it seems inescapable that there are in many cases quite substantial deviations from the theoretical prediction (14), shown by the dashed line in Fig. 4.

Most recently, therefore, attention has been focussing, though again mainly with solid metals in mind, on the contribution at high temperatures from electron–electron scattering. There is an important review, but again on solid metals, by Kaveh and Wiser (1984) that can leave no doubt as to the importance of electron–electron scattering in solid metals at high temperatures. Certainly some of their ideas and theoretical framework are relevant to the metallic liquids under discussion. But more specifically addressed to the question of the constancy, or otherwise (see Fig. 4), of the Lorenz number is the work of Schulz and Allen (1995), which appeared as this article was nearing completion. This is a very detailed study of transport in metals in the presence of electron–electron scattering. The problem of electron–electron scattering is treated by techniques that enable the different scatter-

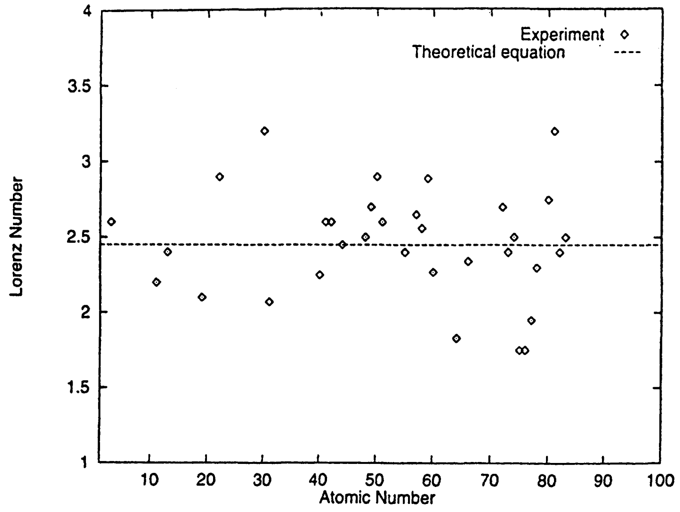

FIG. 4. Variation of experimentally determined Lorenz number (10^{-8} W K^{-2}) of metals with atomic number, after Tankeshwar and March (1996).

ing mechanisms (electron–electron, impurity, and electron–ion) to be treated on the same footing. The attack is based, as usual, on the solution of the Boltzmann transport equation (see, for example, Ziman, 1960) and use of a special set of basis functions. Results for thermal and electrical conductivity, as well as explicit results for Lorenz number, are plotted in detail by Schulz and Allen, which enables the relative importance of the different scattering mechanisms to be assessed. There is little doubt that their approach has considerable relevance for metallic liquids, but detailed applications are not available at the time of writing. However, as in the work of Kaveh and Wiser (1984), they place stress on the importance of electron–electron scattering at high temperature.

8. CONVECTION, PRANDTL NUMBER, AND INSTABILITIES

There is considerable interest at the time of writing in the continuing study of the transition to chaos. Best known is probably the Bénard instability encountered in convection. Consider a layer of fluid heated from below and held at a fixed temperature at the top. Provided the temperature is sufficiently low, heat transfer is by conduction. But when the gradient reaches a critical value, a macroscopic motion begins in the fluid. The heated parts expand and will move upward, being cooled, and then fall back again to the bottom. This motion is well regulated. Typically one observes rolls, but hexagons, seen from above, have also been noted. When the temperature is further raised, the rolls start a wavy motion along the axis—a motion that varies with time. Ultimately the orderly pattern breaks up and a chaotic motion ensues.

The basic equations governing the Bénard motion of the system are (compare Lundqvist *et al.*, 1988, p. 14)

1. the Navier–Stokes equation, in which one adds the force of gravity $\mathbf{F} = (0,0,g)$;
2. the equation of heat conduction,

$$\frac{dT}{dt} = \lambda\nabla^2 T, \tag{28}$$

and

3. the equation of continuity,

$$\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}) = 0. \tag{29}$$

One must impose the boundary conditions that $T = T_0 + \Delta T$ at the bottom and $T = T_0$ at the top. The above partial differential equations governing this problem—namely convection or Bénard instability—can be transformed into a set of ordinary differential equations for the Fourier components of the various hydrodynamical quantities. E. N. Lorenz (1963) truncated these equations, taking only the leading Fourier component into account. This type of Bénard convection experiment provides examples of pattern formation in systems far from equilibrium. One of the (various) dimensionless quantities of interest in such systems is the so-called Prandtl number, $C_p\eta/M\gamma$, which is related (see Egelstaff, 1992) to the similar ratio $M\lambda/k_B\eta$. This latter number evidently connects the thermal conductivity λ with the shear viscosity η. This latter ratio has been compared and contrasted for metallic and insulating liquids by Ascough *et al.* (1988), to whose article the interested reader is referred.

9. SUMMARY

On the basis of simple assumptions of elastic scattering and degenerate conduction electrons, the electronic contribution to the thermal conductivity of a liquid metal is related to its electrical conductivity via the Lorenz number L having the value given in Eq. (14). As Fig. 4 demonstrates, while the predicted Lorenz number for degenerate electrons represents a useful average of the available data, considerable fluctuations occur about this simple prediction of the Wiedemann–Franz law. Some discussion has been given of useful techniques for measuring the thermal conductivity of liquid metals. One example, that of liquid In taken through its freezing point, has demonstrated that for this metal there are marked changes in thermal and electrical conductivities separately at the freezing point but only a relatively mild change in Lorenz number in passing through the phase transition.

Works Cited

Ascough, J., Chapman, R. G., March, N. H. (1988), *Phys. Chem. Liq.* **18,** 253.

Ashcroft, N. W., Mermin, N. D. (1976), *Solid State Physics,* New York: Holt, Rinehart and Winston.

Bhatia, A. B. (1967), *Ultrasonic Absorption,* Oxford: Oxford Univ. Press.

Bhatia, A. B., Thornton, D. E. (1970), *Phys. Rev. B* **2,** 3004–3012.

Bhatia, A. B., Thornton, D. E., March, N. H. (1974), *Phys. Chem. Liq.* **4,** 97–111.

Brooks, R. D., Rosenblatt, A. L. (1953), *Mech. Eng.* **75,** 36368.

Busch, G., Güntherodt, H.-J., Wyssmann, P. (1972), *Phys. Lett.* **39A,** 89–90.

Busch, G., Güntherodt, H.-J., Haller, W. H., Wyssmann, P. (1972), *Phys. Lett.* **41A,** 29–30.

Chester, G. V., Thellung, A. (1961), *Proc. Phys. Soc.* **77,** 1005–1013.

Cohen, C., Sutherland, J. W. H., Deutch, J. M. (1971), *Phys. Chem. Liq.* **2,** 213–235.

Cubbery, W. H. (1979), *Liquid Metals Handbook II,* 9th ed., Metals Park, OH: American Society for Metals.

Egelstaff, P. A. (1992), *An Introduction to the Liquid State,* Oxford: Oxford Univ. Press.

Faber, T. E. (1972), *An Introduction to Liquid Metals,* Cambridge, U.K.: Cambridge Univ. Press.

Goldratt, E., Greenfield, A. J. (1978), *Rev. Sci. Instrum.* **49,** 1531–1536.

Goldratt, E., Greenfield, A. J. (1980), *J. Phys. F* **10,** L95–L99.

Jackson, C. B. (Ed.) (1955), *Liquid Metals Handbook, Sodium-NaK Supplement,* Washington, DC: U.S. GPO.

Kaveh, M., Wiser, N. (1984), *Adv. Phys.* **33,** 257–372.

Kohlrausch, F. (1900), *Ann. Phys. (Leipzig)* **1,** 132–140.

Kohlrausch, F. (1960), *Praktische Physik I,* Stuttgart: Teubner.

Krishnan, K. S., Bhatia, A. B. (1945), *Nature* **156,** 503–504.

Landau, L. D., Placzek, G. (1934), *Phys. Z. Sowjetunion* **5,** 172–177.

Lundqvist, S., March, N. H., Tosi, M. P. (Eds.) (1988), *Order and Chaos in Nonlinear Physical Systems,* New York: Plenum.

March, N. H. (1990), *Liquid Metals,* Cambridge, U.K.: Cambridge Univ. Press.

Mountain, R. D. (1970), *CRC Crit. Rev. Solid State Sci.* **1,** 5–46.

Okada, K., Ozoe, H. (1992), *ASME J. Heat Transf.* , 107–114.

Powell, R. W. (1969), *Contemp. Phys.* **10,** 579–600.

Rice, M. J. (1970), *Phys. Rev.* **B2,** 4800–4804.

Schulz, W. W., Allen, P. B. (1995), *Phys. Rev. B* **52,** 7994–8001.

Sittig, M. (1956), *Sodium: Its Manufacture, Properties and Uses,* American Chemical Society Monograph Series No. XX, New York: Reinhold.

Tankeshwar, K., March, N. H. (1996), *Phys. Chem. Liq.* (in press).

Van Hove, L. (1954), *Phys. Rev.* **95,** 249–262, 1374–1384.

Ziman, J. M. (1960), *Electrons and Phonons,* Oxford: Oxford Univ. Press.

Further Reading

Barasoni, M., Williams, R. K., McElroy, D. L. (1968), in: B. A. Peary, R. Flynn (Eds.), *Proceedings of the 7th Conference on Thermal Conductivity,* NBS Special Publication No. 302, Washington, DC: U.S. GPO.

Cubbery, W. H. (1979), *Liquid Metals Handbook II,* 9th ed., Metals Park, OH: American Society for Metals.

Jeans, J. (1940), *An Introduction to the Kinetic Theory of Gases,* Cambridge, U.K.: Cambridge Univ. Press.

Laubitz, M. J. (1970), *Phys. Rev. B* , 2252–2254.

Lielausis, O. (1975), "Liquid Metal Magnetohydrodynamics," *At. Energy Rev.* (3), 527–581.

Lorenz, E. N. (1963), *J. Atmos. Sci.* **20,** 130–141.

Lundqvist, S., March, N. H., Tosi, M. P. (Eds.) (1988), *Order and Chaos in Nonlinear Physical Systems,* New York: Plenum.

Okada, K., Ozoe, H. (1992), *ASME J. Heat Transf.* **114,** 107–114.

THERMAL CONDUCTION IN NONMETALLIC LIQUIDS*

RICHARD A. PERKINS AND DANIEL G. FRIEND, *Physical and Chemical Properties Division, National Institute of Standards and Technology, Boulder, Colorado, U.S.A.*

Introduction 103
1. **Theoretical Approaches** 105
1.1 Thermal Conductivity from a Molecular Viewpoint 105
1.2 Kinetic Theory of Thermal Conductivity 106
1.3 Enskog Theory and Its Modifications 109
1.4 Corresponding-States Approaches 110
1.5 Hard-Sphere Liquids 113
1.6 Critical-Region Behavior 113
1.7 Computer Simulation of Thermal Conductivity 115
2. **Experimental Techniques** 116
2.1 Heat Transfer Including Radiation and Convection 117
2.2 Steady-State Techniques 118
2.3 Transient Techniques 119
2.4 Periodic Methods 120
2.5 Optical Methods 121
2.6 Reference Materials and Data 121
3. **Pure Fluids** 123
3.1 Saturated Liquid 124
3.2 Compressed Liquid 125
3.3 Critical-Region Behavior 125
4. **Mixtures** 125
5. **Miscellaneous Additional Topics** 127
5.1 Cryogenic Fluids 128
5.2 Refrigeration Fluids 128
5.3 Liquid Fuels and Petrochemicals 128
5.4 Aqueous Systems 129
Glossary 129
Works Cited 130
Further Reading 131

INTRODUCTION

Materials may be conveniently classified as gases, liquids, and solids at a fixed set of conditions, that is, temperature, pressure, and composition. However, a given material can exist in each of these states under the appropriate conditions, barring chemical decomposition at elevated temperatures. On a microscopic scale, the state of a material is defined by the interactions among the molecules in the material and hence on their interatomic potentials. The solid phase is characterized by a high degree of order, symmetry, and structure in the system. This article does not consider the solid phase. The liquid and gas phases have much less structure, and the properties of these phases are increasingly dependent on collisions between molecules and correlated groups of molecules as the density of the fluid decreases.

Simple model intermolecular potentials have characteristic distance and energy parameters. For instance, the Lennard–Jones potential can describe molecules like noble gases, which are spherical and nonpolar; more complex molecular systems can be described by *effective* Lennard–Jones parameters. For a pure (model) Lennard–Jones fluid, the reduced temperature T_r is related to the temperature T and the energy parameter ϵ by $T_r = kT/\epsilon$, where k is Boltzmann's constant, while the reduced density ρ_r is related to the density ρ and the distance parameter σ by $\rho_r = 2\pi N_a \sigma^3 \rho/3$, where N_a is Avogadro's number. Alternatively, on a macroscopic level, the reduced temperature and density of a pure material can be related to the temperature ($T_r = T/T_c$) and density ($\rho_r = \rho/\rho_c$) at its gas–liquid critical point. The equilibrium and transport properties of all fluids exhibit sim-

---

*Contribution of the National Institute of Standards and Technology, not subject to U.S. copyright.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ilarity when they are plotted in terms of these reduced temperatures and densities. This similarity between the properties of all fluids in terms of reduced parameters has led to the development of several very successful corresponding-states theories, which are considered in more detail in later sections.

This similarity of fluid properties, which forms the basis of the corresponding-states principle, can be used both as a predictive tool and as a consistency check for equilibrium and transport properties. For instance, data on several pure fluids can be used to predict the properties of another pure fluid or a mixture of several components. Just as corresponding-states theory can be used to relate the properties of several fluids, thermodynamic- and transport-property theory can be used to relate properties in the liquid, vapor, and supercritical phases to microscopic intermolecular potentials. Thus, even though this article focuses on the thermal conductivity (λ) of a liquid, it is necessary to consider the behavior of thermal conductivity, thermal diffusivity (a), density (ρ), and specific heat capacity at constant pressure (C_p) in the vapor and supercritical regions.

The thermal conductivity of a material has dimensions of energy per unit time, length, and temperature, and the SI units are $\mathrm{W \cdot m^{-1} \cdot K^{-1}}$. This linear transport coefficient is defined by Fourier's law for heat conduction,

$$\mathbf{J}_q = \lim_{|\nabla T| \to 0} (-\lambda \nabla T), \tag{1}$$

where $\mathbf{J}_q$ is the heat-flux vector (in power per unit area) and T is the absolute temperature. Measurement of transport properties, such as the thermal conductivity, requires a small perturbation from equilibrium conditions to approximate the limiting zero gradient in Eq. (1). Thus, thermal conductivity is defined in the limit of zero temperature gradient, viscosity (η) is defined in the limit of zero shear gradient, and mass diffusivity (D) is defined in the limit of zero chemical-potential gradient. The requirement of a small, but finite, temperature gradient introduces several complications in experimental measurements, which must be carefully considered in the liquid phase.

The density of most fluids (water near the solid line is an exception) decreases as the temperature increases when the pressure is held constant. A temperature gradient is required to determine the thermal conductivity according to Eq. (1) in any measurement device, or heat-transfer application, and so there are also density gradients introduced in the fluid. A fluid, unlike a solid, is free to flow in response to these density gradients, and free convection may result. This can lead to measurement errors and discrepancies between experimental data sets since the definition of thermal conductivity assumes that the fluid is stationary. Secondly, the molecules in the fluid, or solid, can absorb and emit thermal radiation. Since there is a temperature gradient present, energy will be transferred by radiation from high-temperature regions of the fluid and its surroundings to lower-temperature regions. Emission and absorption of radiation occur at the molecular level at discrete wavelengths according to the spectra of energy levels available in the fluid and the absolute temperature of the fluid. For heat-transfer calculations, it is convenient to integrate these discrete levels over all wavelengths to obtain an average gray-body absorption/emission coefficient for the fluid. The radiative transfer of heat from one element of fluid at T_1 to another at T_2 is given by

$$J_q = \sigma \epsilon_g (T_1^4 - T_2^4) \approx \sigma \epsilon_g T^3 (T_1 - T_2), \tag{2}$$

where σ is the Stefan–Boltzmann constant, ϵ_g is the average gray-body emissivity, and T is the average absolute temperature. Equation (2) shows that the heat transported by thermal radiation increases in proportion to the cube of the absolute temperature. Furthermore, it is proportional to the temperature gradient in the fluid and is difficult to separate from heat transfer due to thermal conduction.

In general, the heat flux is often due to a combination of several different modes of heat transfer and may include contributions from conduction, convection, radiation, boiling, condensation, etc. To deal with this complicated situation, engineers have developed the concept of a heat-transfer coefficient (U_h) that is defined implicitly by

$$J_q = U_h (T_1 - T_2) = U_h \Delta T, \tag{3}$$

where ΔT is defined according to the situation. The parameter U_h is a lumped, macroscopic quantity that is usually correlated in terms of the thermophysical properties of the materials involved. For instance, the linearization of the Stefan–Boltzmann Eq. (2) allows the radiative heat flux to be written

$$J_q = \sigma\epsilon_g T^3(T_1 - T_2) = U_h(T)\Delta T, \tag{4}$$

so $U_h = \sigma\epsilon_g T^3$ for the case of pure radiative heat transfer. In the case of heat transfer due to free convection (buoyancy driven), the heat-transfer coefficient is correlated in terms of several characteristic dimensionless parameters of the fluid. The classical correlation for heat transfer due to free convection at a vertical surface is in terms of the Nusselt number

$$\begin{aligned} N_{\mathrm{Nu}} &= \frac{U_h L}{\lambda} = c_h N_{\mathrm{Gr}}^m N_{\mathrm{Pr}}^m \\ &= c_h\left(\frac{L^3\rho^2 g_c \beta\Delta T}{\eta^2}\right)^m \left(\frac{C_p\eta}{\lambda}\right)^m, \end{aligned} \tag{5}$$

where N_{Gr} is the Grashof number, N_{Pr} is the Prandtl number, the parameters c_h and m are correlated for a given geometry, L is the characteristic length, g_c is the gravitational constant, β is the volumetric coefficient of thermal expansion, and C_p is the isobaric specific heat of the fluid. For convection, boiling, and condensation heat transfer, ΔT is the difference between the surface temperature and the bulk fluid temperature outside of the film region of the fluid where large temperature gradients exist. The heat-transfer coefficient determined from Eq. (5), $U_h = \lambda N_{\mathrm{Nu}}/L$, can then be used in Eq. (3); effects due to free convection will dominate in a situation like this over conductive and radiative contributions. Much literature is devoted to engineering heat-transfer correlations; the interested reader will find a good introduction in the book by Bird *et al.* (1960) and a general summary in the handbook by Perry *et al.* (1984).

The dimensionless parameters used in heat-transfer correlations require accurate thermophysical property data both in the development of heat-transfer correlations and in their subsequent use. Detailed finite-element or finite-difference computer simulations of complex heat-transfer situations are possible but have the same requirements for accuracy in the underlying thermophysical property data. These required data include thermal conductivity, thermal diffusivity, viscosity, density, specific heat, radiative emissivity, phase-change enthalpy, and interfacial tension. This article concentrates on the thermal conductivity but cannot be limited to thermal conductivity. The thermal conductivity is normally measured as a function of temperature, pressure, and composition, but it is most conveniently correlated in terms of temperature, density, and composition. This requires an accurate equation of state to calculate the liquid density. The thermal diffusivity a is the key property for transient heat transfer in fluids and is defined as

$$a = \frac{\lambda}{\rho C_p} = \frac{\text{(energy transport)}}{\text{(volumetric storage of energy)}}. \tag{6}$$

Thermal diffusivity has units of m²/s, as do the binary mass diffusion coefficient D_{12} and the viscous diffusivity ν. Thus, in addition to thermal conductivity, either thermal diffusivity or the fluid density and specific heat are required to describe transient conduction through fluids.

1. THEORETICAL APPROACHES

1.1 Thermal Conductivity from a Molecular Viewpoint

The physical basis for the conduction of heat energy in terms of the molecular constituents of a fluid system can be understood from a classical point of view. The particles in a fluid possess kinetic energy because of their thermal motion, internal energy related to electronic, vibrational, and rotational excitations, and potential energy associated with intermolecular interactions or external fields. These forms of energy can be interchanged through collisional processes and can be carried by the thermal motion of the molecules. When a temperature gradient is present in a fluid system, the flow of energy re-establishes the equilibrium state defined by the fundamental laws of statistical physics. This relaxation to equilibrium in response to a perturbing temperature gradient, as distinguished from processes related to macroscopic mass flow (convection) or electromagnetic radiation, is the mechanism of thermal conduction. In this article, we confine ourselves to linear phenomena in nonreactive systems, which practically involves small departures from equilibrium and small temperature gradients. The phenomenological macroscopic definition of the thermal conductivity, Fourier's law of heat conduction, was given in Eq. (1). In this section, we will relate Eq. (1) and the value of λ to microscopic processes in a fluid system.

The simplest crude model that we consider is a fluid subjected to a steady-state one-dimensional temperature gradient in the x direction, and in which the flow of energy is assumed to be due to the molecular transport of the system's internal energy per molecule, U. Thus, in Fig. 1, we assume that at $x = -l$, a local equilibrium at temperature T_1 is achieved, so that the particle carries an energy $U(T_1)$. We further assume a mean free path of l and that there is no change in the value of $U(T)$ when a particle moves through this distance. If the average speed of a particle is $\langle v \rangle$ at the origin, and $\frac{1}{6}$ of the particles at $x = -l$ move in the positive x direction and $\frac{1}{6}$ of the particles at $x = l$ move in the negative x direction, we can write the energy flux crossing the zero line as

$$J_q = \tfrac{1}{6}\langle v \rangle n[U(T_1) - U(T_2)], \tag{7}$$

where n is the number density of particles in the fluid. A Taylor expansion of the energy about $U(T_0)$ gives

$$U(T_1) = U(T_0) - l\frac{dU(T_0)}{dx}, \tag{8}$$

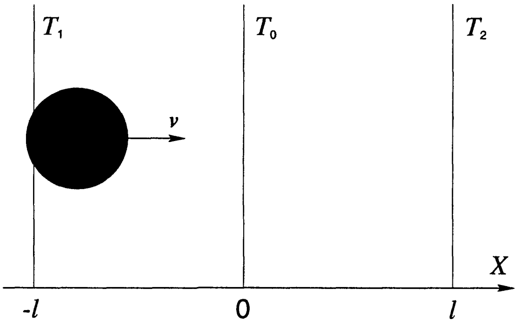

FIG. 1. Molecular transport of heat in a one-dimensional temperature gradient.

so that the energy difference in the brackets of Eq. (7) becomes $-2l dU/dx$. Without concerning ourselves, for the moment, about the additional state-variable dependence in U, we can write

$$n\frac{dU}{dx} = n\frac{dU}{dT}\frac{dT}{dx} = \rho C_v\frac{dT}{dx}, \tag{9}$$

where C_v is the molar specific heat (at constant volume) and ρ is the molar density. With this information, we can now combine Eqs. (1) and (7) to find

$$\lambda = \tfrac{1}{3}\langle v \rangle l \rho C_v. \tag{10}$$

This simple kinetic model is not unreasonable, especially in the dilute system, although we have ignored collisional transfer (in which energy is transported not through the movement of the molecules, but through their interaction) and have not dealt explicitly with the internal degrees of freedom in a molecule. Equation (10) can be evaluated in a more complete approach, and the rigorous solution for a dilute hard-sphere system gives

$$\lambda_0^{hs}(T) = \frac{25}{32}\left(\frac{kT}{\pi m}\right)^{1/2}\frac{C_v}{N_a\sigma^2}, \tag{11}$$

where σ is the hard-sphere diameter and m is the molecular mass. This expression can be considered a starting point for theoretical approaches that consider liquid-state properties from density expansions about the dilute region, that can incorporate collisional transfer and intramolecular energy processes, and that use more realistic interparticle potentials. Although Eq. (11) was derived for the dilute-gas regime, and effective hard-sphere diameters can generally be found to describe such gases, in Table 1 we show some effective liquid hard-sphere diameters derived from this formalism. As expected, effective liquid hard-sphere diameters are significantly smaller than the dilute-gas values, because we have ignored several important effects, as mentioned above.

1.2 Kinetic Theory of Thermal Conductivity

The more formal kinetic theory of thermal conductivity starts from a study of the N-particle distribution function f_N and the

Table 1. Effective hard-sphere diameters of selected liquids at the normal boiling point; from Eq. (11).

Fluid	T (K)	C_v ($\mathrm{J \cdot mol^{-1} \cdot K^{-1}}$)	λ ($\mathrm{W \cdot m^{-1} \cdot K^{-1}}$)	σ^{eff} liquid (nm)	σ^{eff} dilute (nm)
Argon	87	22.3	0.128	0.131	0.468
Water	373	67.9	0.679	0.174	0.572
R134a	247	86.5	0.104	0.294	0.858
Toluene	384	140.7	0.111	0.416	1.013
Methanol	338	72.9	0.191	0.288	0.702
Propane	231	64.0	0.129	0.275	0.850

Liouville equation, which describes its time evolution. Quantum versions of these equations, in terms of Wigner distribution functions, are also available. To solve for the thermal conductivity, equations for f_N are written in terms of coupled equations for distributions representing fewer particles. Approximations can be used to decouple these equations. The Boltzmann equation (BE) for a dilute system of structureless particles can be written in terms of the single-particle distribution $f(\mathbf{v},\mathbf{r},t)$ where $f(\mathbf{v},\mathbf{r},t) \times d^3v\, d^3r\, dt$ gives the number of particles with a velocity in the infinitesimal range about $\mathbf{v}$, at a position near $\mathbf{r}$, at a time near t. The equilibrium distribution is the Maxwell–Boltzmann expression, and collisional processes that conserve mass, momentum, and energy ensure that this equilibrium is achieved. In the presence of small linear driving forces, the single-particle distribution can be considered a perturbation of the equilibrium solution, and the BE can be solved when the interparticle potential is known. Thus, Eq. (11) can be generalized to monatomic structureless systems interacting with realistic potentials (Millat *et al.*, 1996), such as the Lennard–Jones 12-6 potential, the m-6-8 potential (Hanley and Klein, 1972), or the Aziz–Chen potential for argon (Aziz and Chen, 1977), with the classic first-order result

$$\lambda_0(T) = \frac{\lambda_0^{hs}(T)}{\Omega^{*(2,2)}(\Phi,T)}. \tag{12}$$

In this expression, $\Omega^{*(2,2)}$ is a reduced collision integral, which is a functional of the interparticle potential $\phi(r)$.

To account for the internal degrees of freedom that can affect energy flow, Eucken (1913) developed an *ad hoc* correction factor associated with this expression; Enskog, who (with Chapman) is credited with solving the dilute-gas BE, also extended the approach to dense systems of hard-sphere fluids; see Chapman and Cowling (1939) and Hirschfelder *et al.* (1954). To deal with polyatomic fluids, instead of using the Boltzmann equation, we can use a kinetic equation in which the distribution functions have additional variables, and energy exchange among internal and kinetic modes is allowed (McCourt *et al.*, 1991a; Millat *et al.*, 1996). For instance, instead of the single-particle distribution above, we can write $f(\mathbf{v},\mathbf{j},\mathbf{r},t)$ where $\mathbf{j}$ is the rotational angular momentum of a nonspherical molecule. Such kinetic equations have been developed in classical, semiclassical (by Wang Chang *et al.*, 1964), and quantum (by Waldmann, 1957 and Snider 1960) (WS) approaches, although the solutions to such equations are substantially more difficult.

To see how the approaches progress, we can write the evolution of the single-particle distribution as

$$\left[\frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla\right] f = \mathfrak{S}(f,f_1), \tag{13}$$

where the left-hand side includes the streaming term, which indicates how particle flow changes f, and the right is the collision term, which indicates the effect of interactions between particles in the specified state and particles with general coordinates $\mathbf{v}_1$, $\mathbf{r}_1$, which will be integrated over. In the Boltzmann equation, we have

$$\mathfrak{S} = \int d^3v_1 \int_0^{2\pi} d\epsilon \int_0^{\infty} b\,db|\mathbf{v}_1 - \mathbf{v}| \times \left[f(\mathbf{r},\mathbf{v}',t)f(\mathbf{r},\mathbf{v}_1',t) - f(\mathbf{r},\mathbf{v},t)f(\mathbf{r},\mathbf{v}_1,t)\right], \tag{14}$$

where collisions cause particles to enter state points with the primed velocities and leave states with the unprimed velocities. In this expression, ϵ represents the angle between the velocity vectors of the incoming particles, b is the impact parameter of the collision, and the velocities are constrained to conserve momentum and kinetic energy. Note that the BE describes collisions taking place at a point, which may be an appropriate approximation for dilute systems only. In the Waldmann–Snider equation, the collision term can be written as

$$\begin{aligned}\mathfrak{S} = \mathrm{Tr}_1 \int d^3v_1 \sum_j P_j \\ \times \left[\int d^3\epsilon a(E;\boldsymbol{\epsilon},\boldsymbol{\epsilon}')|\mathbf{v}_1' - \mathbf{v}'|f'f_1'a^\dagger(E;\boldsymbol{\epsilon},\boldsymbol{\epsilon}') \right. \\ \left. - \frac{4\pi\hbar}{m}\mathrm{Im}[a(E;\boldsymbol{\epsilon},\boldsymbol{\epsilon})ff_1]\right]P_j, \qquad (15)\end{aligned}$$

where the trace gives the quantum mechanical statistical average, P_j projects onto the specified (rotational) state, and a and $a^\dagger$ are annihilation and creation operators for energy level E with specified directions ϵ for the relevant velocity vectors. With either collision term, the kinetic equation, Eq. (13), is solved using the continuity equations for mass, momentum, and energy. The thermal conductivity is found by identifying the coefficient of the temperature gradient in an expression for the heat-flux vector.

A microscopic expression for the classical heat-flux vector of Eq. (7) can be written as

$$\begin{aligned}\mathbf{J}_q = \int d^{6N}\Gamma f_N(\Gamma,t)\left[\sum_i \mathbf{v}_i\left[\frac{mv_i^2}{2} + \frac{1}{2}\sum_j \phi_{ij}\right]\right. \\ \left. - \frac{1}{2}\sum_{ij} \mathbf{r}_{ij}\mathbf{v}_i\cdot\mathbf{F}_{ij}O_{ij}\right]\delta(\mathbf{r} - \mathbf{r}_i), \qquad (16)\end{aligned}$$

where Γ represents the full $6N$-dimensional phase space, $\mathbf{v}_i$ is the peculiar velocity of particle i (the particle velocity minus the macroscopic streaming velocity), ϕ is the interparticle potential, F is the interparticle force, and O is an operator that can incorporate certain nonlocal effects. The solution of the integro-differential kinetic equation is not trivial, but formal solutions for polyatomic gases have been obtained; see the monograph by McCourt *et al.* (1991a). Instead of the collision integral $\Omega^{*(2,2)}$ described in Eq. (12), we can write the polyatomic solution in terms of related effective cross sections (which can be associated with matrix elements of the operators in the kinetic equation) as

$$\begin{aligned}\lambda_0^{\mathrm{poly}}(T) = \lambda_0^{hs}(T)\pi\sigma^2\frac{4R}{5C_v} \\ \times \frac{\mathfrak{S}(1001) - 2r\mathfrak{S}\begin{pmatrix}1010\\1001\end{pmatrix} + r^2\mathfrak{S}(1010)}{\mathfrak{S}(1010)\mathfrak{S}(1001) - \mathfrak{S}^2\begin{pmatrix}1010\\1001\end{pmatrix}}, \\ (17)\end{aligned}$$

where R is the universal gas constant, the indices on the effective cross sections $\mathfrak{S}$ refer to matrix elements associated with elastic collisions, internal energy states, and inelastic collisions, and r^2 is the ratio of the rotational heat capacity to the translational portion of the isobaric heat capacity.

The formal result of Eq. (17) serves to motivate the Eucken correction, which provides a more empirical attempt to extend the kinetic-theory results from monatomic to more complex molecular systems. Mason and Monchick (1962) made their theoretical arguments from a more classical approach (Millat *et al.*, 1996). We can write

$$\lambda = \lambda^{\mathrm{trans}} + \lambda^{\mathrm{int}} \qquad (18)$$

to separate effects due to transport of translational energy and interchange with internal (intramolecular) energy of a molecular fluid. Instead of an expression like Eq. (11), we then use

$$\lambda_0^{\mathrm{trans}} = \epsilon_{\mathrm{trans}}\frac{C_v^{\mathrm{trans}}}{C_v}\lambda_0^{\mathrm{mon}} \qquad (19)$$

and

$$\lambda_0^{\mathrm{int}} = \epsilon_{\mathrm{int}}\frac{C_v^{\mathrm{int}}}{C_v}\lambda_0^{\mathrm{mon}}, \qquad (20)$$

where Eq. (12) gives the monatomic result λ_0^{mon}, the isochoric heat capacity has been divided into its translational and internal contributions, and ϵ is a measure of the relative efficiencies of energy transport by these mechanisms. In this notation, $\epsilon_{\mathrm{trans}}$ is often approximated as 1, ϵ_{int} is $\frac{2}{5}$, $C_v^{\mathrm{trans}} = 3R/2$, and $C_v^{\mathrm{int}} = C_v - C_v^{\mathrm{trans}}$. These approximations correspond to fast interchange between translational and rotational energy states and correspond to the original Eucken correction in the dilute limit:

$$\lambda_0 = \left(\frac{9\gamma}{10} - \frac{1}{2}\right)\lambda_0^{\text{mon}}, \tag{21}$$

where γ is the ratio of heat capacities, C_p/C_v. We can develop improved approximations from the Wang Chang–Uhlenbeck deBoer equation or from the solution to the Waldmann–Snider equation, Eq. (17), where the cross terms in the translational–rotational energy exchange are related to a diffusional process. An improved estimate for a modified Eucken correction gives $\epsilon_{\text{int}} \approx 0.528$ (Reid *et al.*, 1987), and the full Mason–Monchick calculation gives values of ϵ_{trans} and ϵ_{int} in terms of the self-diffusion coefficient, the rotational contribution to the heat capacity, the viscosity, and a rotational collision number (or a related relaxation time). The value of λ_{int} is often assumed to be independent of density, and thus the dilute-gas value discussed here is combined with a density-dependent value for λ^{trans} to calculate the thermal conductivity at arbitrary state points, including the liquid phase. Although this assumption cannot be strictly true, it is often sufficient to obtain an estimate or correlation for $\lambda_0^{\text{int}}(T)$ and use this value at all densities.

Before we conclude this long section on kinetic theory, we briefly discuss another microscopic approach to the calculation of thermal conductivity. A simple expression for conservation of energy, the heat-conduction equation, can be written, using Eq. (1), as

$$\rho C_v \frac{\partial \tilde{E}}{\partial t} = \lambda \nabla^2 \tilde{E}, \tag{22}$$

where we are examining the time rate of change of the local excess energy $\tilde{E}$ (its local value relative to a macroscopic average) due to heat flow. Fluctuation theory tells us how to evaluate deviations from an ensemble average (McQuarrie, 1976), and we can obtain the Green–Kubo result that

$$\lambda = \frac{1}{Vk_{\text{B}}T^2}\int_0^\infty \langle \mathbf{J}_q(t)\cdot\mathbf{J}_q(0)\rangle, \tag{23}$$

where an expression for the heat-flux vector has been given in Eq. (16). This time–correlation-function formalism is especially useful in determining the thermal conductivity from equilibrium computer simulations of fluid systems.

1.3 Enskog Theory and Its Modifications

In order to extend the Boltzmann equation to higher densities, we must consider at least two modifications of Eqs. (13) and (14). First, because collisional processes have a finite spatial extent, energy can cross a boundary line, such as the zero line in Fig. 1, without any particles actually crossing this boundary. This collisional transfer process can be described, in part, by a nonlocality expansion of the operator O_{ij} of Eq. (16) or by using an analogous treatment in a kinetic equation. Second, the frequency of collisions, which is implicitly incorporated into the BE from a simple statistical argument, must be modified. This is because, in a dense system, two particles whose trajectories at time t_0 ordinarily result in a collision at a time t may be shielded by the presence of a third particle, which is not considered in the dilute-system formulation of the BE. Thus, multiparticle correlation must be considered in a dense system. Enskog provided a solution to such a problem in the special case of a hard-sphere system.

The collision term of Eq. (14) becomes

$$\begin{aligned}\mathfrak{S} = \int d^3v_1 \int_0^{2\pi} d\epsilon \int_0^\infty b\,db|\mathbf{v}_1 - \mathbf{v}| \\ \times [y(\mathbf{r} + \tfrac{1}{2}\sigma\hat{\mathbf{k}})f(\mathbf{r},\mathbf{v}',t)f(\mathbf{r} + \sigma\hat{\mathbf{k}}, \mathbf{v}_1', t) \\ - y(\mathbf{r} - \tfrac{1}{2}\sigma\hat{\mathbf{k}})f(\mathbf{r},\mathbf{v},t)f(\mathbf{r} - \sigma\hat{\mathbf{k}}, \mathbf{v}_1, t)],\end{aligned} \tag{24}$$

where $y(\mathbf{r})$ is related to the pair distribution function and $\hat{\mathbf{k}}$ is the unit vector connecting the centers of the two colliding particles. The kinetic equation can then be solved with the result

$$\begin{aligned}\lambda(T,\rho) = \lambda_0(T)B(T)\rho \\ \times \left[\frac{1}{y(T,\rho)} + 1.20 + 0.755y(T,\rho)\right],\end{aligned} \tag{25}$$

where B is the second virial coefficient ($2\pi\sigma^3/3$ for hard spheres), ρ is the fluid density, λ_0 is the dilute-gas solution of Eq. (11), and y in this Enskog theory can be equated to the residual compressibility factor, $y_{\mathrm{ET}} = P/\rho RT - 1$. This hard-sphere Enskog solution can be used directly for real fluids or for more realistic potentials by using the appropriate values for B and y. However, a modification of the Enskog result (see Hirschfelder *et al.*, 1954), also proposed by Enskog, provides a better description of dense-fluid thermal conductivity. In this modification, the MET, the pressure in the expression for y is replaced by a "thermal" pressure $T(\partial P/\partial T)|_\rho$ so that the factor becomes $y_{\mathrm{MET}} = T(\partial P/\partial T)|_\rho/\rho RT - 1$. To ensure that the expression in Eq. (25) reduces to the correct value in the dilute limit, the second virial coefficient is also replaced by $B + T\,dB/dT$. For polyatomic molecules, this approach is often associated with the translational contribution to the thermal conductivity defined in Eq. (18), and the internal contribution is treated with a density-independent Eucken factor. This approach generally gives reasonable agreement with experimental data for dense systems. In Fig. 2, we illustrate the Enskog-theory results in comparison to a correlation for the thermal conductivity of argon along an isotherm.

1.4 Corresponding-States Approaches

The thermal conductivity of a pure fluid can be generally written as

$$\lambda(\rho,T) = \lambda_0^{\mathrm{trans}}(T) + \lambda_0^{\mathrm{int}}(T) + \lambda_{\mathrm{ex}}^{\mathrm{trans}}(\rho,T) + \lambda_{\mathrm{ex}}^{\mathrm{int}}(\rho,T) + \Delta\lambda_c(\rho,T), \qquad (26)$$

where the division into translational internal contributions, discussed in Eq. (18), is incorporated into both the dilute term and the excess term λ_{ex}, and the critical enhancement $\Delta\lambda_c$ is introduced. This last contribution arises from cooperative effects associated with fluctuations near the critical point of the fluid and is discussed in Sec. 1.6. The contributions of these terms are illustrated schematically in Fig. 3. In some studies, the excess thermal conductivity is further divided; for example, a density series can be used to define a first density correction, analogous to the second virial coefficient in thermodynamic studies. The independent variables Eq. (26), the temperature and density, are the fundamental determinants of the thermal conductivity in the sense that thermal motions and collisional processes play the key roles; for a mixture, the composition variable(s) can be included. Although the single temperature variable uniquely defines the fluid state and the thermal conductivity of a pure liquid in equilibrium with its vapor (saturated liquid), it is the density variation along the saturation boundary that is most important in determining thermal conductivity.

In practice, the internal contribution to the thermal conductivity is often considered to depend only on the temperature (and thus it is always equal to its dilute-gas value), so that $\lambda_{\mathrm{ex}}^{\mathrm{int}}$ vanishes; also, the translational portion of the excess term is often considered to be independent of temperature, and the critical enhancement is significant only near the

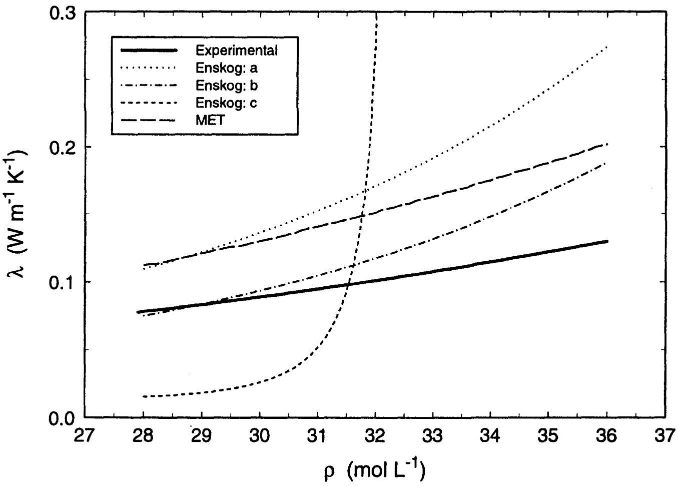

FIG. 2. Thermal conductivity of compressed liquid argon along isotherm at 125 K. Solid line is experimental correlation. Models use σ = 0.357 nm. Dotted line (a) is Enskog hard-sphere theory, dot-dashed line (b) is Enskog theory with experimental value for λ_0, short-dashed line (c) is Enskog theory with experimentally derived values of λ_0 and $y(T,\rho)$, and long-dashed line is modified Enskog theory (MET).

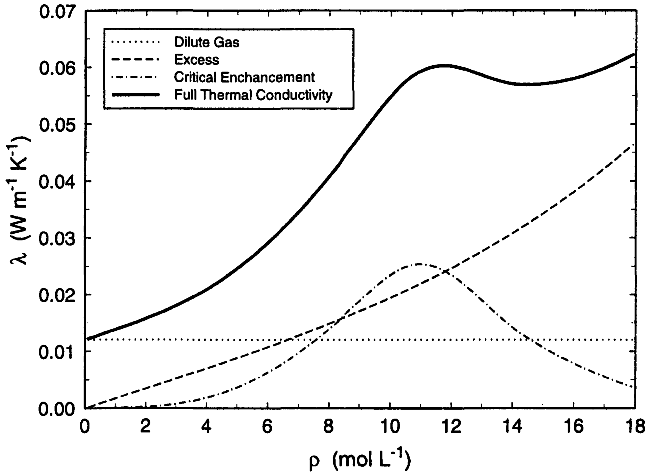

FIG. 3. Contribution of dilute gas, excess, and critical enhancement to the thermal conductivity of supercritical nitrogen at 128 K.

critical point. Although these assertions cannot be strictly true, this approximation simplifies Eq. (26) to

$$\lambda(\rho,T) \approx \lambda_0^{\text{trans}}(T) + \lambda_0^{\text{int}}(T) + \lambda_{\text{ex}}^{\text{trans}}(\rho). \tag{27}$$

The topology and amplitudes for the thermal conductivity surfaces of a wide assortment of liquids are similar. This observation has inspired the development of scaling-factor arguments that lead to corresponding-states approaches to the calculation of liquid thermal conductivity. From a molecular point of view, we can show that if the interparticle potentials of two fluids are conformal, that is, equivalent when values for an energy parameter and length parameter are properly scaled, at least the translational contributions to the thermal conductivity must also scale. This approach leads to the relationship

$$\lambda(\rho,T) = \Lambda\lambda^*_{\text{ref}}(\rho^*,T^*), \tag{28}$$

which can be applied to any or all terms of Eqs. (26) or (27). The factor Λ can be taken as $k^{1/2}\epsilon^{1/2}/m^{1/2}\sigma^2$ from kinetic-theory arguments [Eq. (11) or (23), for example], where k is the Boltzmann constant, ϵ and σ are energy and length scales (both of which can be obtained from an interparticle potential), and m is the molecular mass. Alternatively, we can include the thermodynamic $C_v(\rho,T)$ in Λ, use the viscosity (which, from kinetic theory, is directly related to the translational portion of the thermal conductivity in the dilute-gas limit), develop the length and energy scales from critical constants or thermodynamic corresponding-states approaches, or construct empirical approximations for Λ. Models in which Λ is constant can be developed, or Λ can be made a function of composition, temperature, and/or density. The relations between ρ and ρ^* and between T and T^* can again be constant (or composition-dependent) length and energy scaling factors, such as $\rho^* = \rho\rho_c/\rho_c^{\text{ref}}$ and $T^* = T_c^{\text{ref}}/TT_c$ where the subscript c indicates a critical or pseudocritical value and "ref" indicates a reference fluid, or can be density- and/or temperature-dependent transformations defined from the thermodynamic surface or from an empirical examination of thermal conductivity data. In this manner, we can relate the thermal conductivity of any liquid system (pure fluid or mixture) to that of a reference thermal-conductivity surface. The reference surface can be established from correlation of experimental data, or it can be developed from theoretical solutions of a model system.

With these corresponding-states arguments, we have formally created a mathematical mapping between the thermal-conductivity surfaces of all fluid systems. In fact, the mapping becomes simpler when the system of interest and the reference system are similar: that is, where both fluids are noble gases, nonpolar fluids, aqueous systems, etc. Because the temperature dependence of the thermal conductivity is small at high densities or effectively confined to the dilute-gas region, it is often explicitly removed from consideration to reduce the dimensionality of the problem. Even for the reference fluid, then, we can write $\lambda(\rho,T) = \Lambda^*\lambda(\rho')$. Any residual temperature dependence can be included in Λ^* and/or in the transformation between ρ and ρ' to obtain conformality of the isotherms to an arbitrary accuracy. Figure 4 illustrates this corresponding-states behavior in nearly 4200 experimental points for five compositions in the binary methane–ethane system over the temperature range from 140 to 330 K (Friend and Roder, 1987). In this illustration, all temperature dependence was incorporated in the dilute-gas term (so the residual temperature dependence is indicated in the figure), Λ^* has been set equal to 1, and the reduced density was empirically determined to be the system density divided by a simple composition-dependent reducing factor. Similarly, thermal conductivity data for a variety of liquid systems, pure fluids and mixtures, can be made to lie on a single curve in a graph of reduced thermal conductivity $\lambda^* = \lambda/\Lambda^*$ versus reduced density. There are several algorithms for achieving this corresponding-states relationship, and systematic methods for describing various classes of systems have been devised; the single curve obtained in this method can then be regressed from data or described from a theoretical description of a model system.

To make this discussion of corresponding-states calculations of thermal conductivity more explicit, we summarize an approach widely used for hydrocarbon systems, refrigeration fluids, etc. According to Eq. (26), we treat each contribution to the thermal conductivity as a separate problem. We assume that the thermodynamic surface of the fluid is known from experimental data or from an equilibrium corresponding-states algorithm. A standard thermodynamic corresponding-states model (Leach *et al.*, 1968) constructs a mapping such that the reduced residual Helmholtz energy, $(A - A^{\text{ideal gas}})/RT$, and residual compressibility factor x of the fluid of interest are (approximately) equal to that of the reference fluid when the state points are related by $\rho^* = \rho_x h_x$ and $T^* = T_x/f_x$. In this thermodynamic formalism, the equivalent substance reducing parameters, h_x and f_x, are defined by this relationship and can be related to the critical parameters or implicitly defined as temperature- and density-dependent functions depending on the details of the model. In this approach, the dilute-gas contributions are assumed to obey a corresponding-states relation in the sense that a single potential model, with parameters estimated from data or from a scheme involving the critical parameters and the acentric factor, is used to solve for the translational portion of the collision integral, Eqs. (11) and (12). An Eucken correction is used for the internal contributions, such that the value of C_v^{int} of Eq. (20) is approximated as $C_p^{\text{ideal}} - 5R/2$ and ϵ_{int} becomes a temperature-dependent function, $\epsilon_1 + \epsilon_2 T_c/T$, which can be fitted to dilute-gas data or assumed to obey a corresponding-states principle. This term is considered independent of the density. The factor Λ of Eq. (28) for the translational contribution can be taken as a temperature- and density-dependent function,

$$\Lambda = (M_{\text{ref}}/M)^{1/2} f_x^{1/2} h_x^{-2/3}, \tag{29}$$

which can be further modified to take into account shape dependences by including a multiplier based on the acentric factors of the fluids. This approach can be used for mixtures (Ely and Hanley, 1983), in which case the effective molar mass M can be a function of the mole fractions, molar masses, and equivalent substance reducing ratios of each of the components, and the scheme can be extended to include a corresponding-states approach for the critical-region enhancement.

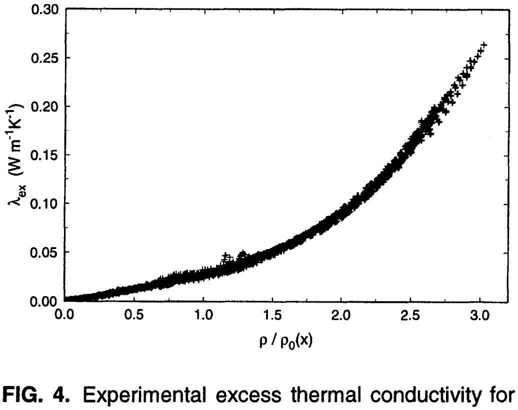

FIG. 4. Experimental excess thermal conductivity for methane–ethane mixtures as a function of reduced density.

1.5 Hard-Sphere Liquids

In Eq. (25), we have given a kinetic-theory solution for the thermal conductivity of a hard-sphere fluid, and a model for rough spheres can be shown to give a result directly proportional to this expression. This can be combined with the dilute hard-sphere gas solution, Eq. (11), with $C_v = 3R/2$. In an expression for $\lambda^{hs}/\lambda_0^{hs}$, the molecular size parameter appears explicitly only as a factor of σ^2. This motivates a factor of $\rho^{2/3}$ in the definition of Λ in Eq. (28). A corresponding-states expression based on these observations, which has been successfully applied to a variety of liquids, is

$$\lambda(\rho,T) = \hat{R}_\lambda \rho^{2/3} T^{1/2} M^{-1/2} \lambda^*(\rho^*). \tag{30}$$

In this expression, λ^* is treated as a universal curve for liquids that can be obtained from the hard-sphere result or regressed from available data. The reduced density is defined by $\rho^* = \rho/\rho_0(T)$, where ρ_0 is related to an effective hard-sphere diameter and has a weak temperature dependence. The reducing parameter $\rho_0(T)$ shows a systematic behavior for various families of fluids and can be regressed from available viscosity and self-diffusion data as well as thermal conductivity data in a consistent approach to all three transport coefficients (Assael *et al.*, 1992a). The factor $\hat{R}_\lambda$ includes contributions from the hard-sphere result as well as an empirical factor R_λ, which represents a deviation from hard-sphere behavior, or roughness factor, which can account for internal contributions to the thermal conductivity. The multiplier R_λ also shows a systematic behavior within a family of molecules and generally has values somewhat greater than 1 (Assael *et al.*, 1992a); it can also be adjusted to have temperature dependence to describe the thermal conductivity of polar liquids (Assael *et al.*, 1994, 1995; Bleazard and Teja, 1996).

1.6 Critical-Region Behavior

The critical point of a fluid system has been discussed in the article CRITICAL PHENOMENA. It has been identified as the point at which the order parameter takes on a nonzero value (from above) and a symmetrical phase is broken into a phase with lower symmetry—e.g., liquid–vapor phase separation. Several thermodynamic properties exhibit divergences at this point, and dynamical properties also show enhancements. Microscopic fluctuations are known to play a dominant role in this region, and critical slowing down becomes important for transport processes. The renormalization-group approach for thermodynamics can be extended to a dynamical renormalization group, and model systems or mode-coupling approaches can be used to understand the behavior of the thermal conductivity in this region. Observationally, the thermal conductivity exhibits an enhancement in the critical region, which is most easily observed for pure fluids along supercritical isotherms. The thermal conductivity reaches a maximum on this path, which is centered near the critical density and is nearly symmetrical around this density. The maximum approaches infinity as the temperature approaches the critical temperature, and an enhancement, although small, may be seen for some systems at temperatures up to about twice the critical temperature. In Fig. 5, we show the critical-region behavior of the thermal conductivity of carbon dioxide. In the compressed-liquid state, isotherms terminate on the saturation boundary, and so the critical density is not accessible. Nonetheless, along liquid isotherms at higher temperatures, the critical enhancement is also present and becomes dominant as the critical temperature is approached. Although it is difficult to isolate this critical-enhancement contribution to the liquid thermal conductivity at lower temperatures in order to ascertain its extent, this phenomenon must be considered in any general description of liquid-phase thermal conductivity.

A phenomenological explanation of the critical enhancement of the thermal conductivity can be developed from the relation between the thermal diffusivity and thermal conductivity given in Eq. (6): $a = \lambda/\rho C_p$. We combine this with the simple Stokes–Einstein expression for the diffusion of particles in a fluid medium, $D = kT/6\pi\eta R$, where D is the diffusion coefficient, η is the fluid viscosity, and R is the radius of the spherical particle. Because the density fluctuations have a correlation length of ξ, we consider the thermal diffusivity as a diffusive process of these

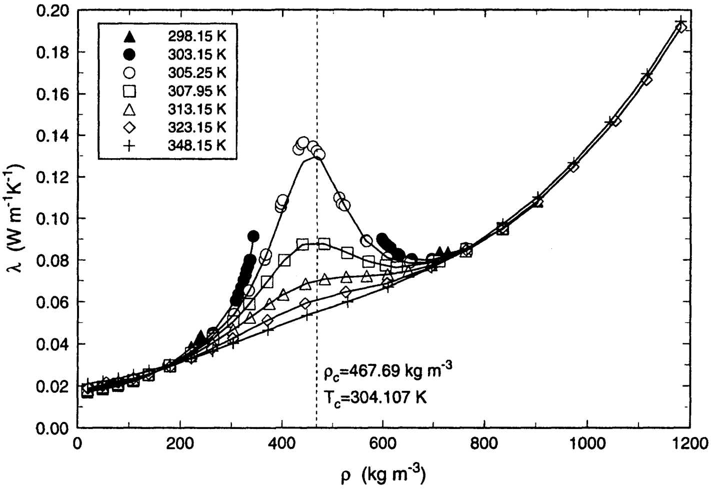

FIG. 5. Thermal conductivity of carbon dioxide in the critical region as a function of density along several isotherms. The experimental data are those of Michels *et al.* (1962). The solid curves are calculated from the correlation of Vesovic *et al.* (1990).

fluctuations that equates *a* to *D* with the result

$$\lambda = \rho k T C_p / 6\pi\eta\xi. \qquad (31)$$

The divergence of the thermal conductivity comes about, in this formalism, from the strong critical divergence of the isobaric heat capacity (with critical exponent near 1.2 along the critical isochore). Both the correlation length and the viscosity in Eq. (31) also diverge, but these critical exponents are significantly smaller. Although the asymptotic behavior of the thermal conductivity described by Eq. (31) may be confined to a very small region around the critical point and is thus difficult to achieve, the general character of the thermal conductivity surface is consistent with this asymptotic behavior.

In order to describe the effect of critical-region fluctuations throughout the range of their influence, the first four terms in Eq. (26) are often combined to form a background term $\lambda_b(\rho,T)$, so that the critical enhancement is empirically defined by $\Delta\lambda_c(\rho,T) = \lambda(\rho,T) - \lambda_b(\rho,T)$. This definition is ambiguous, and it is based on the concept that the background should be well behaved throughout a complete range of temperature and density. This implies that a description of a background thermal-conductivity surface developed from data well outside the critical region can be extrapolated into the critical region with reliable results. The critical enhancement can then be correlated either empirically or with a theoretically based crossover formalism, which attempts to bridge the range of independent variables between the asymptotic critical region of Eq. (31) to the region in which the fluctuation-induced behavior becomes negligible. The enhancement can be described empirically by a Gaussian function (Roder, 1982) such as

$$\Delta\lambda_c(T,\rho) = F(T)\exp[-A(\rho - \rho_c)^2], \qquad (32)$$

where the temperature-dependent amplitude *F* is strongly peaked or infinite at the critical temperature and *A* is an empirical width parameter. Alternatively, we can use the asymptotic form of Eq. (31) and write

$$\Delta\lambda_c = R\,\frac{\rho k T C_p}{6\pi\eta\xi}\,\tilde{\Omega}(T,\rho), \qquad (33)$$

where *R* is a universal amplitude with a value of approximately 1.02. The damping function $\tilde{\Omega}$ has been considered empirically (Sengers, 1985), often with an overall multiplicative thermodynamic factor of $(C_p - C_v)/C_p$. More recently, approximate solutions to the mode-coupling equations have become available (Olchowy and Sengers, 1988; Perkins *et al.*, 1991b). Thus $\tilde{\Omega}$ can be expressed in terms of the thermodynamic quantities, background values for the viscosity and thermal conductivity, and a single system-dependent wave-number cutoff. This cutoff represents a maximum wave number (or minimum wavelength) above which the critical fluctuations do not contribute to $\Delta\lambda_c$. The correlation length in Eq. (33) can also be estimated from thermodynamic quantities and is related to the fluid's isothermal compressibility.

This discussion of the critical-region enhancement is appropriate for pure fluids, but the problem for mixtures is more complicated. In fact, however, these approaches have been used with some success for both pure-fluid and mixture corresponding-states calculations of the thermal conductivity (Huber *et al.,* 1992). The nature of the order parameter for mixture critical phenomena depends on whether the fluid is near a vapor-liquid critical line or a liquid–liquid critical line, and the behavior of the thermal conductivity must be considered accordingly. This problem has received considerable recent attention, and much progress has been made in this area (Luettmer-Strathmann and Sengers, 1996). We conclude this section by noting that for liquid systems of both pure fluids and mixtures, explicit incorporation of the critical enhancement contribution can be ignored as long as the state point is sufficiently far from the critical region. In this case, contributions from this process may be implicitly included in an empirical description of the thermal-conductivity surface, but extreme care must be taken not to extrapolate such a surface into the critical region.

1.7 Computer Simulation of Thermal Conductivity

Advances in computer technology have made it possible to examine complex thermal conduction problems and evaluate liquid thermal conductivity by simulating the microscopic behavior of the fluid constituents. This enables us to examine in detail the assumptions underlying the theoretical and correlational approaches to thermal conductivity, including the kinetic processes, the interaction between internal and external degrees of freedom, and the sensitivity to details of the intermolecular interactions. In principle, computer simulations can complement or even supplant experimental determinations of heat-transfer processes, especially for extreme regions of temperature, pressure, or chemical reactivity where physical experiments become difficult but simulation experiments remain tractable. The equations of motion in a computer simulation are often based on Newtonian dynamics, although additional constraints can be included to ensure that the ensemble temperature remains constant, intramolecular bonds are maintained, etc. To augment the number of particles in a simulation, which cannot approach the Avogadro number, which quantifies the number of molecules in a real macroscopic system, periodic boundary conditions are often used. Additionally, to make the problem tractable, short-range potentials or cutoff length scales are used to reduce the number of force pairs that must be considered. Simulation algorithms for the calculation of thermal conductivity can be divided into two major classifications, equilibrium molecular dynamics (MD) and nonequilibrium molecular dynamics (NEMD), which we discuss briefly, in turn.

It may be intuitively difficult to see how an equilibrium simulation can determine parameters for dynamical processes such as the transport of heat energy in a temperature gradient; however, the Green–Kubo expression for the thermal conductivity, Eq. (23), expresses just this relationship. The thermal conductivity tells us the rate at which fluctuations in temperature relax to equilibrium in the context of an autocorrelation function for the heat-flux vector. In a classic simulation for thermal conductivity by Alder *et al.* (1970), an integral of the heat-flux vector was monitored for a system of 108 particles interacting through a hard-sphere repulsion. Thus the thermal conductivity was evaluated from

$$\lambda = \frac{1}{2Vk_{\mathrm{B}}T^2 s} \times \left\langle \left[\sum_{i=1}^{N} x_i(t+s)E_i(t+s) - x_i(t)E_i(t) \right]^2 \right\rangle, \tag{34}$$

where N is the number of particles and s is the time step in the simulation. Thus each particle's position x and total energy E_i (or rather each component of the position and energy) are monitored at every time, and the simple differencing and time averaging of Eq. (34) is performed to obtain the thermal conductivity. This simulation gave excellent agreement with the Enskog result of Eq. (25), and even allowed examination of the separate contributions from the kinetic, potential, and cross terms in the theoretical approach. However, questions remained concerning the long-time behavior of the autocorrelation function within a finite ensemble of particles.

More recent equilibrium MD simulations have focussed on more realistic potentials, such as the Lennard–Jones 6–12 potential, and have included simulations of liquid mixtures (Hoheisel and Vogelsang, 1988). The heat-flux vector itself, defined microscopically by an expression similar to Eq. (16), can be used directly as an alternative to Eq. (34), and various simulation ensembles have been used. Even without explicitly accounting for internal degrees of freedom or the angular dependence in the intermolecular potential, reasonable results for liquid thermal conductivity can be obtained from a simulation based on Eq. (23) using effective parameters for the particle interactions. For noble-gas fluids, *ab initio* potential surfaces based on spectroscopic data can be used within the Green–Kubo formalism in MD simulations. For supercritical liquidlike states of neon, for example, the differences between MD results and experimental thermal conductivities are less than 10% (Eggenberger *et al.*, 1992). For polyatomic fluids, potentials based on a rigid-rotor model for nitrogen, for example, have been used for dilute-gas calculations based on a Monte Carlo evaluation of the collision integrals such as in Eq. (12) (Heck and Dickinson, 1994). These nonisotropic potential surfaces for nonspherical systems are intrinsically more difficult to incorporate in MD simulations.

Nonequilibrium molecular dynamics (NEMD) is a computer-simulation technique that more closely parallels the experimental determination of transport coefficients based on their definition, such as Eq. (1). The computer "particles" respond with a steady-state nonequilibrium flux to a driving thermodynamic force, the temperature gradient. An alternative NEMD algorithm uses a fictitious heat-field vector $\mathbf{F}_q$, which corresponds to $-\nabla \ln T$ in the sense that it drives a heat-flux vector. In this case, $\mathbf{F}_q$ contributes to the time rate of change of each particle's momentum, but the thermodynamic temperature is constrained to be constant in the system. A heat-flux vector is given by

$$\mathbf{J}_q = \frac{1}{V}\sum_{i=1}^{N} \mathbf{v}_i e_i - \tfrac{1}{2}\sum_{i,j\neq i} \mathbf{r}_{ij}\mathbf{F}_{ij}\cdot\mathbf{v}_i \tag{35}$$

for an N-particle system, with e_i representing the total energy of particle i; this is a straightforward implementation of Eq. (16). The thermal-conductivity coefficient is a time average defined by

$$\lambda = \lim_{F_q \Rightarrow 0} \langle \mathbf{J}_q \cdot \mathbf{F}_q / T F_q^2 \rangle \tag{36}$$

in analogy to Eq. (1). This NEMD method has been very successful for atomic fluids and has been extended to polyatomic fluids by including rotational degrees of freedom and torque terms in the equations of motion (Hanley and Evans, 1996); potentials based on site–site interactions for a molecular fluid such as benzene have also been used (Ravi *et al.*, 1992). Resultant thermal conductivities are found that agree very well with experimental results, and the mechanisms contributing to the heat-transfer process can be studied in controlled microscopic detail (Millat *et al.*, 1996).

2. EXPERIMENTAL TECHNIQUES

Experimental techniques for the measurement of thermal conductivity can be classified according to their mode of operation. Thus, thermal-conductivity measurement techniques are generally either steady state or transient. Temperatures and heat fluxes can be either measured directly or inferred from optical measurements of refractive-index gradients or fluctuations. Some techniques measure thermal conductivity directly, while others measure thermal diffusivity, from which thermal conductivity can be calculated using Eq. (6), assuming that both density and isobaric heat capacity are known with sufficient accuracy. Some transient techniques are capable of measuring both thermal conductivity and thermal diffusivity (Perkins *et al.*, 1991a), thus providing a consistency check for both quantities if the isobaric specific heat and density are well known. Alternatively, measurements of both thermal diffusivity and thermal conductivity can be used with density data to obtain estimates of the isobaric specific heat of a fluid (Matos Lopes *et al.*, 1995). Acoustic techniques are available to measure thermal conductivity and thermal diffusivity, but are presently restricted to the gas phase and will not be discussed here.

Generally some temperature gradient is imposed on the fluid and the heat flux is measured. Thus in the general sense, it is a heat-transfer coefficient, or apparent thermal conductivity, that is measured. The principal experimental problem during thermal-conductivity determinations is to minimize heat transfer from mechanisms other than pure conduction, such as thermal radiation and convection. These parasitic heat-transfer mechanisms are always present to some extent and can become significant for certain fluids under certain temperature and density conditions in even the most carefully designed experiment. Nieto de Castro *et al.* (1986) showed a plot of the accepted value of the thermal conductivity of saturated liquid toluene as a function of the year measured. In 1920 the accepted value was about 150 $\mathrm{mW \cdot m^{-1} \cdot K^{-1}}$ at 300 K with an uncertainty of about 5%. The accepted value decreased steadily so that it was about 8–10% lower in 1968, as understanding of the effects of convection and thermal radiation on the measurements increased. The encouraging news is that the accepted value for this substance has remained relatively constant since 1968, and toluene is recommended as a standard reference fluid by the International Union of Pure and Applied Chemistry (IUPAC) (Nieto de Castro *et al.*, 1986).

Instruments to measure thermal conductivity can be classified as relative or absolute instruments. Relative instruments require a well-characterized calibrating fluid such as toluene or water. Many instruments are operated in a relative fashion because of inadequacies in their working models. It may not be adequate to calibrate with a single fluid, since measurements of another fluid may involve extrapolation of the calibration. Calibration with two fluids is preferable, especially if the two fluids have properties that lie on either side of the fluid of interest. Measurements are then based on interpolation, rather than extrapolation, between the two reference fluids. Absolute measurements require knowledge of properties such as cell geometry, temperature, pressure, voltage, and resistance. Standard reference thermal-conductivity surfaces are established from measurements with absolute instruments and are useful to check the proper operation of new and modified absolute instruments.

2.1 Heat Transfer Including Radiation and Convection

Three differential equations are required to describe combined heat and momentum transfer and must be solved simultaneously in the general case with no heat generation in the fluid. These are the continuity equation, the equation of motion, and the conservation of energy equation. The continuity equation is

$$\frac{\partial}{\partial t}\rho = -\nabla \cdot (\rho \mathbf{v}), \tag{37}$$

where ρ is the fluid density and $\mathbf{v}$ is the fluid velocity. The momentum equation or equation of motion is

$$\frac{\partial}{\partial t}\rho \mathbf{v} = -\nabla \cdot [\rho \mathbf{v}\mathbf{v}] - \nabla \cdot \boldsymbol{\tau} - \rho g_c \beta (T - \bar{T}), \tag{38}$$

where $\boldsymbol{\tau}$ is the shear stress tensor ($\boldsymbol{\tau} = -\eta \nabla \mathbf{v}$), g is the gravitational constant, and $\bar{T}$ is the average fluid temperature. Finally, the conservation of energy is given by

$$\begin{aligned}\frac{\partial}{\partial t}\rho C_p T = {} & -\nabla \cdot (\rho C_p T \mathbf{v}) - \nabla \cdot (\mathbf{J}_q \\ & - \mathbf{J}_r) - (\boldsymbol{\tau}:\nabla \mathbf{v}) + \left(\frac{\partial \ln V}{\partial \ln T}\right)_p \frac{Dp}{Dt} \\ & + \rho T \frac{DC_p}{Dt},\end{aligned} \tag{39}$$

where $\mathbf{J}_r$ is the radiant flux and the hydrodynamic derivative $D/Dt = \partial/\partial t + \mathbf{v} \cdot \nabla$. Steady state is reached when the temperature at any given position does not vary with time ($\partial T/\partial t = 0$), and transient terms, which are dependent on time t, are eliminated.

Equations (37)–(39) are simplified significantly when the thermophysical properties are constant, an accurate assumption for small temperature gradients, and when there is no pressure gradient or viscous dissipation. These equations become

$$\nabla \cdot \mathbf{v} = 0, \tag{40}$$

$$\rho \frac{\partial \mathbf{v}}{\partial t} = -\rho \nabla \cdot \mathbf{v}\mathbf{v} + \eta \nabla^2 \mathbf{v} - \rho \mathbf{g} \beta (T - \bar{T}), \quad (41)$$

and

$$\rho C_p \frac{\partial T}{\partial t} = -\rho C_p \nabla \cdot T\mathbf{v} + \lambda \nabla^2 T - \nabla \cdot \mathbf{J}_r. \quad (42)$$

Equations (40)–(42) describe processes in which heat transfer from conduction, free convection, and thermal radiation are all present in the limit of a small temperature difference. Equation (42) indicates that the energy storage is the sum of convection plus conduction plus radiation heat transfer. Equations (42) and (6) show that, in the absence of convection ($\mathbf{v} = 0$), the thermal diffusivity of the fluid can be found from a transient experiment. All three of these equations must be considered for any real thermal conduction experiment. Experiments to measure the thermal conductivity of a fluid strive to minimize the contributions from free convection and thermal radiation, while using the most simple geometries possible.

2.2 Steady-State Techniques

Steady-state techniques confine the fluid of interest between two surfaces at different temperatures in a simple, easily modeled geometry. A constant, steady-state heat flux is emitted by the hot surface, and the temperature difference between the two surfaces is recorded. The ideal model assumes that all of the heat emitted by the hot surface is absorbed by the cold surface and that heat transfer is purely by thermal conduction. This is an approximation, so that parasitic heat losses from edges and ends must be minimized and carefully accounted for, often with the aid of measurements in vacuum. Convection in the fluid can be detected and characterized by a study of the measured apparent thermal conductivity as a function of the applied power or the measured temperature difference. Pure thermal conduction has no power dependence on the assumption that the limit of the gradient of temperature approaching zero is effectively realized (although the thermal properties are themselves temperature dependent), so that the apparent thermal conductivity is nearly constant when there is no convection. Absorption and re-emission of thermal radiation by the fluid can be detected and characterized by varying the gap between the two surfaces of the measurement cell. If the contribution of thermal radiation is significant, the apparent thermal conductivity will be a function of the gap spacing, since the optical path length between the two measurement surfaces is changing. Generally, small gaps are preferred to minimize both convection and radiation corrections and errors. Additionally, the surfaces of the measurement cell should have low infrared emissivities to minimize thermal radiation contributions for both transparent and absorbing fluids.

The experimental thermal conductivity of the fluid is given by

$$\lambda = C_{\text{cell}} J_{\text{app}} / \Delta T, \quad (43)$$

where C_{cell} is the cell constant of the steady-state instrument and is defined by the geometry of the instrument. For any geometry, it is necessary to know the cell constant of the instrument to obtain the thermal conductivity of the fluid from the measured applied power J_{app} and temperature rise ΔT. The cell constant of absolute instruments can be obtained from a careful characterization of the cell dimensions and the coefficients of thermal expansion of the materials of construction. Alternatively, the cell surfaces can be used as capacitor plates, so that a vacuum measurement of the ac electrical properties of the cell provides a direct measurement of the cell constant. In the case of relative instruments, the cell constant is obtained through the use of Eq. (43) for one or more reference fluids of known thermal conductivity. The common geometries for steady-state thermal conductivity instruments are parallel plates, concentric cylinders, and heated wire in a cylinder.

The parallel-plate instrument is perhaps the most simple geometry, but is not easily used at very high pressures. Convection in parallel-plate instruments is suppressed by placing the hot plate above the cold plate; the temperature gradient cannot drive free convection. This technique has been used to get within several millikelvins of the gas–liquid critical point. The concentric-cylinder instrument has significant advantages for use at very high pressures. This geometry is normally operated so that the cylindrical axis is vertical. Thus, even though convection can be suppressed through the use of narrow gaps between the hot and cold cylinders, the concentric-cylinder technique is not the method of choice in the critical region where convection is a severe problem. Finally, the hot-wire technique is easy to set up and use from low to extremely high pressures. The hot-wire technique is generally more prone to fluid convection errors than even the concentric-cylinder technique, since the gap between the wire and the outer cylinder is larger than the gap in most concentric-cylinder instruments. Parasitic heat loses are generally minimized through the use of guard heaters at the ends of the plates or cylinders. Consult Le Niendre *et al.* (1991) for a more detailed discussion of steady-state measurement techniques.

2.3 Transient Techniques

The hot wire immersed in an infinite fluid medium is the most common geometry for transient thermal-conductivity measurements. The wires are used simultaneously as electrical heat sources and resistance thermometers. Differences between various instruments are based on the geometry of the cell and the power-forcing function that is used during the measurement. Useful forcing functions include impulse, step, and ramp power. The fundamental working equation for the transient hot-wire techniques assumes that the heat is generated from an infinitely long line source. The line heat source has no volume and no radius, and so there is no thermal lag and all of the heat is dissipated by conduction into the fluid medium. The fluid medium is assumed to be infinite, which implies that the transient thermal wave does not penetrate to the cell boundary during the time frame of the experiment. The rate at which the thermal wave penetrates into the fluid is directly related to the thermal diffusivity of the fluid. Thus, the outer-boundary radius does not need to be very large for liquids, which typically have quite small thermal diffusivities. Transient hot-wire cells are designed to approximate the line-source solutions as closely as possible with wire diameters ranging from 4 to 25 μm. The smallest diameters, 4–7 μm, are preferred for gas-phase measurements. Deviations between the actual heat transfer and the line-source solution are treated as time-dependent corrections to the experimental data such that

$$\Delta T_{\mathrm{id}}(t) = \Delta T_w(t) + \sum_i \delta T_i(t), \tag{44}$$

where ΔT_{id} is the corrected temperature rise, ΔT_w is the experimental temperature rise, and δT_i are the corrections to the temperature rise for nonideal heat transfer. The article by Assael *et al.* (1991) provides a detailed discussion of the various corrections associated with the transient hot-wire technique.

The line-source solutions are similar for the three cases of impulse-power, step-power, and ramp-power forcing and are described by Mohammadi *et al.* (1981). The ideal model for impulse-power input is

$$\Delta T(t) = \frac{q^0}{4\pi\lambda t}\exp\left(\frac{-r_w^2}{4at}\right), \tag{45}$$

where q^0 is the total heat content of the pulse, t is the time that has elapsed since the pulse, and r_w is the wire radius. For the case of a step power-forcing function, the long-time solution is

$$\Delta T(t) = \frac{q}{4\pi\lambda}E_1\left(\frac{r_w^2}{4at}\right) \approx \frac{q}{4\pi\lambda}\ln\left(\frac{4at}{r_w^2 C}\right), \tag{46}$$

where q is the power applied per unit length during the experiment, E_1 is the exponential integral, and $C = \exp(\gamma)$ is the exponential of Euler's constant. For the case of the ramp power-forcing function, the long-time solution is

$$\frac{\Delta T(t)}{t} \approx \frac{\hat{q}}{4\pi\lambda}\ln\left(\frac{4at}{r_w^2 D}\right), \tag{47}$$

where $\hat{q}$ is the slope, or time derivative, of the power applied per unit length during the experiment and $D = \exp(1 + \gamma)$. The long-time approximation in Eqs. (46) and (47) is normally valid for experimental times greater than 50–100 ms. The characteristic dimensionless time for the transient hot-wire experiment is $\tau = at/r_w^2$ with the validity of the long-time approximation extending to shorter real times for fluids with large thermal diffusivities and for cells with small hot wires. The step power-forcing function is most commonly used because of simplicity and a well-developed set of corrections to the ideal heat-transfer model. The ramp power-forcing function is less likely to drive convection in the fluid, since the temperature gradient in the fluid remains smooth and continuous throughout the fluid during an experiment. However, this technique has not been widely used. A key advantage of all the transient techniques is that both thermal conductivity and thermal diffusivity can be obtained.

Effects due to axial thermal conduction can be eliminated through the use of a compensating dual-wire cell. In the dual-wire configuration, the transient temperature rise of a short wire can be subtracted from that of the main long wire using a Wheatstone bridge. The resulting signal represents the response from a finite-length wire with a length that is equal to the difference between the long and short wire lengths. The contribution of the ends, where the axial temperature gradient is significant, is eliminated using this dual-wire technique. Alternatively, fine voltage-sensing leads, which measure the wire resistance increase due to heating, can be located at points on the hot wire that are at a sufficient distance from the end supports so that there is no axial temperature gradient present. The transient hot-wire techniques readily allow detection of convection during measurements, since a finite length of time is required at any given power to move the fluid in the presence of the temperature gradient. Convection occurs at shorter times as the applied power of the measurement is increased and is apparent because it causes deviations from the ideal conduction heat-transfer model.

Since the hot wires are directly immersed in the fluid of interest, problems arise when the fluid is highly polar and has a significant electrical conductivity. These electrical difficulties can be directly overcome by electrically insulating the hot wire. The electrical insulation should be thin to minimize corrections to the ideal line-source model. The most common systems are platinum wires insulated with polyester, anodized tantalum wires, and liquid metals in quartz capillaries. The thermal conductivity of polar fluids such as alternative refrigerants, alcohols, water, and even aqueous salt solutions can be measured using electrically insulated wires or capillaries. Additional details are available in the references at the end of this article.

2.4 Periodic Methods

Periodic methods such as forcing the heat flux using sinusoidal power (or temperature) measure the thermal diffusivity of the fluid, since the penetration distance into the fluid is a function of the period or frequency of the temperature cycle. If the average applied power is known during a periodic heating experiment, then it is possible to obtain the fluid thermal conductivity directly. The most common periodic heating method uses a wire that is driven with a low-frequency power cycle such that $Q = Q_0[1 + 2 \times \cos(\omega t)]$. The third harmonic of the heating current is proportional to the wire temperature cycling amplitude:

$$\Theta_3 = \frac{Q_0}{2\pi\lambda L} \ln\left[\frac{2}{\gamma r_w}\left(\frac{a}{2\omega}\right)^{1/2}\right], \tag{48}$$

where $\omega = 2\pi f$ is the angular frequency of the heating current and L is the wire length. This equation holds for low frequencies and fine wire diameter where

$$r_w^2\omega \leq 4\lambda/\pi C_w \rho_w \tag{49}$$

and C_w and ρ_w are the heat capacity and density of the wire. Equation (48) shows that both the thermal conductivity and the thermal diffusivity can be found if a range of frequencies is studied. A detailed discussion of periodically heated wires can be found in Shpilrain *et al.* (1975) and the monograph by Vargaftik *et al.* (1994).

Recently, Peltier drivers have been used to produce sinusoidally varying temperatures to measure the thermal diffusivity of fluids. Peltier drivers allow the fluid surfaces to be periodically cooled and warmed relative to the bulk fluid temperature. Czarnetzki and Roetzel (1994) describe periodic temperature experiments for several geometries including a circular cylinder, a sphere, a semi-infinite body, and a slab. They determined from this work that the cylinder and slab geometries produce thermal diffusivity data with the least uncertainty. No attempt was made to quantify the heat that was input and removed during these experiments, so only thermal diffusivity was determined. However, the thermal conductivity of the fluid can be obtained from these measured thermal diffusivities using Eq. (6) if the fluid density and isobaric heat capacity are available or can be estimated using an accurate equation of state.

2.5 Optical Methods

Optical methods are attractive for measuring the thermal diffusivity of transparent liquids. As with the periodic methods, the thermal conductivity of the fluid can be obtained from the measured thermal diffusivity using Eq. (6) if the fluid density and isobaric heat capacity are available or can be estimated using an accurate equation of state. Photon-correlation spectroscopy does not require a temperature gradient, and so it is particularly useful in the critical region, where the susceptibility of the fluid to convection is greatly enhanced. In Rayleigh scattering, the spectrum of the scattered light has three peaks, the Rayleigh line and the symmetric pair of Brillouin lines. In the light-scattering technique, the Rayleigh spectrum, due to fluctuations of the fluid's refractive index (or dielectric constant), is measured at a known angle relative to the incident laser beam. For a pure fluid, these entropy fluctuations are directly related to the thermal diffusivity of the fluid,

$$a = 1/q^2\tau_s, \tag{50}$$

where $q = 2k\sin(\theta/2)$, k is the incident wave number, q is the scattered wave number, θ is the scattering angle, and τ_s is the intensity autocorrelation decay time for entropy fluctuations. For a mixture, there are contributions from both entropy and concentration fluctuations. The concentration fluctuations become significant if there is a significant difference in the refractive index of the components in the mixture and provide a measure of the mass diffusion coefficient

$$D = 1/q^2\tau_c, \tag{51}$$

where τ_c is the intensity autocorrelation decay time for concentration fluctuations. Fortunately, the decay time of the concentration fluctuations is usually much longer than that for the entropy fluctuations, so that both the thermal diffusivity and the mass diffusion coefficient can be resolved in many situations.

In forced Rayleigh scattering, the thermal diffusivity of a sample is found by observing the decay time of light diffracted from a periodically heated area in the fluid. The heating is provided by an excitation laser and is localized by the interference pattern from two crossed laser beams in the fluid. The localized heating of the fluid at the antinodes of the interference pattern generates a sinusoidally periodic temperature pattern during the heating cycle. The periodic temperature pattern affects the refractive index of the fluid and acts like a diffraction grating on the probe laser beam. The excitation beam is then turned off, and the decay of the temperature pattern is monitored through the diffraction of the low-powered probe beam. The thermal diffusivity is given by

$$a = -\tfrac{1}{2}\left(\frac{P}{2\pi}\right)^2\left(\frac{d\ln I_1}{dt}\right), \tag{52}$$

where I_1 is the first-order diffracted beam intensity and P is the grating spacing. This method has been applied to extremely corrosive materials such as molten salts. A more detailed description is provided by Shaumeyer *et al.* (1991).

2.6 Reference Materials and Data

A variety of absolute and relative experimental techniques are available to measure the thermal conductivity and thermal diffusivity of liquids. There are uncertainties associated with each of these techniques, which are increased under certain conditions. For instance, thermal radiation errors in techniques that require a temperature gradient are larger at high temperatures and for fluids that moderately absorb infrared radiation. Photon-correlation spectroscopy has large uncertainty at low temperatures relative to the gas–liquid critical point, but it has significant advantages near the gas–liquid critical point since it does not require a macroscopic temperature gradient. Thus, light-scattering measurements of thermal diffusivity are complementary to transient and steady-state measurements of thermal conductivity. It is desirable to use data from a variety of absolute techniques to develop standard reference data for the thermal conductivity of liquids. These standard reference data can be used both to verify the performance of new absolute measurement techniques and to calibrate relative thermal-conductivity and thermal-diffusivity instruments. A number of reference liquids have been proposed by IUPAC collectively spanning a temperature region from 180 to 575 K at saturation pressure. These materials can be obtained in high purity and do not decompose easily at elevated temperatures.

A number of criteria are used by IUPAC to select standard reference materials for the thermal conductivity of liquids:

1. The temperature interval between the freezing point and the normal boiling point, i.e., the extent of the normal liquid range, must be large and include some thermometric fixed point.
2. The liquid should be nontoxic and noncorrosive with respect to the usual engineering materials.
3. The liquid should be obtainable at low cost and guaranteed high purity.
4. Several fluids are required to cover as wide a range of thermal conductivity as possible.
5. Reliable values of liquid densities for the saturated and compressed liquid should be available in the literature, allowing reliable extrapolation of high-pressure data to the saturation line. Additionally, it is desirable for the reference fluids to have a high viscosity to minimize uncertainty due to fluid convection and to be either totally transparent or totally opaque to thermal radiation, to minimize uncertainty due to thermal radiation.

Water and toluene are two of the most widely studied and best characterized liquids. Their thermal conductivities are very different, so they are useful for calibrations of relative instruments. However, water is a very good solvent and picks up ionic impurities, air, and carbon dioxide. The electrical conductivity of water and other highly polar liquids increases as these impurities are dissolved and can be a significant problem for many measurement techniques. Water should not be used to calibrate instruments, such as bare transient hot-wire cells, that are sensitive to electrical conduction in the sample. Likewise, care must be taken when handling hydrocarbon reference fluids such as toluene, dimethylphthalate, and *n*-heptane, since they have a tendency to absorb water from the air in open containers, particularly at lower temperatures. Dimethylphthalate has the highest viscosity of these hydrocarbon reference fluids, which makes it the least sensitive to temperature gradients that might induce convection errors.

Water has been widely studied because of its industrial importance and its significance in natural and geological phenomena. A widely accepted standard surface correlation for the thermal conductivity of water and steam was developed by the International Association for the Properties of Water and Steam (IAPWS) over the temperature range of 273–1073 K at pressures at least to 100 MPa (Sengers and Watson, 1986). IAPWS provides recommended values for the thermal conductivity of saturated liquid water from 273 to 646 K. IUPAC has recommended several reference correlations for saturated liquid water that have a smaller uncertainty but that cover a more limited temperature range. These correlations have evolved through the years as more accurate data have been reported. An early reference correlation was proposed by Nieto de Castro *et al.* (1986) and covered the temperature range from 274 to 370 K with an estimated uncertainty of 1%, and subsequently by Le Neindre (1987). A new IUPAC correlation was presented by Ramires *et al.* (1995) over the same temperature range, which reduces the estimated uncertainty to 0.7% using new data from transient hot-wire experiments using electrically insulated wires. Agreement between the IAPWS and IUPAC recommendations is within 1%, and these correlations will evolve as new data are obtained.

Toluene is also very well characterized near the saturated-liquid line using steady-state, transient, and light-scattering techniques. Two IUPAC reference correlations were proposed by Nieto de Castro *et al.* (1986) from 230 to 360 K (1%) and 189 to 360 K (1.5%) and subsequently by Le Neindre (1987). A new correlation has recently been described by Ramires *et al.* (1996) that covers the range from 189 to 553 K. The estimated uncertainty of this reference data correlation is 1% from 189 to 440 K, 1.5% from 440 to 480 K, and 2% above 480 K.

An IUPAC reference data correlation is described by Nieto de Castro *et al.* (1986) and Le Neindre (1987) for saturated liquid *n*-heptane. This correlation covers the temperature region from 191 to 365 K with an estimated uncertainty of 1.5%. This correlation was based on a limited number of data sets that had significant discrepancies between them but remains the best available recommendation for the thermal conductivity of saturated liquid *n*-heptane.

A good discussion of the advantages of dimethylphthalate as a thermal-conductivity reference fluid is provided by Poltz and Jugel (1982) along with a demonstration of the contribution of thermal radiation on thermal-conductivity measurements. An IUPAC standard reference data correlation is described by Le Neindre (1987) for saturated liquid dimethylphthalate from 273 to 493 K. No estimated uncertainty is provided for the dimethylphthalate correlation.

3. PURE FLUIDS

The thermal-conductivity surface for argon is shown in Fig. 6 as a function of temperature and density. These data were reported by Perkins *et al.* (1991b) using the transient hot-wire technique. Argon is a spherical, nonpolar, monatomic molecule and hence is a nearly ideal fluid, yet it is representative of a wide range of fluids as will be shown. The gas–liquid critical point for argon is about T_c = 150.66 K, ρ_c = 13.29 mol/L, and P_c = 4.906 MPa, and the triple point is T_t = 83.80 K, ρ_t = 35.40 mol/L, and P_t = 0.0689 MPa. The compressed-liquid region exists at pressures above the saturation pressure from the triple point to the critical-point temperature. Argon is supercritical at temperatures above T_c, but the smooth behavior along the supercritical isotherms relative to both the subcritical gas and liquid phases illustrates the need to consider all three of these regions in order to understand the liquid-phase thermal conductivity. An increase in the thermal conductivity is apparent in the supercritical isotherm nearest to the critical point. This bump is due to the thermal-conductivity critical enhancement, which is significant in an extended region around the critical point and thus affects the subcritical liquid and vapor isotherms. Even a material such as argon "gas" has a liquid phase at a low enough temperature. The reduced parameters of temperature ($T_r = T/T_c$) and density ($\rho_r = \rho/\rho_c$) relative to the gas–liquid critical point are important in the context of the phase behavior and the thermophysical properties of fluids.

The thermal conductivity of water, calculated from the IAPWS correlation (Sengers and Watson, 1986), is shown in Fig. 7. Although water is a fluid that is liquid at ambient conditions, the thermal conductivity, when it is plotted in terms of reduced temperature and density, appears to be very similar to that of argon. The gas–liquid criti-

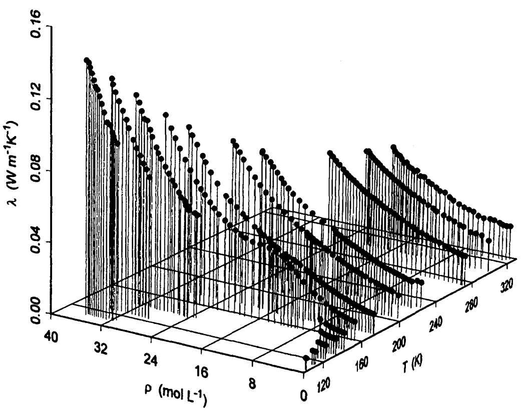

FIG. 6. Thermal-conductivity measurements for argon, reported by Perkins *et al.* (1991b).

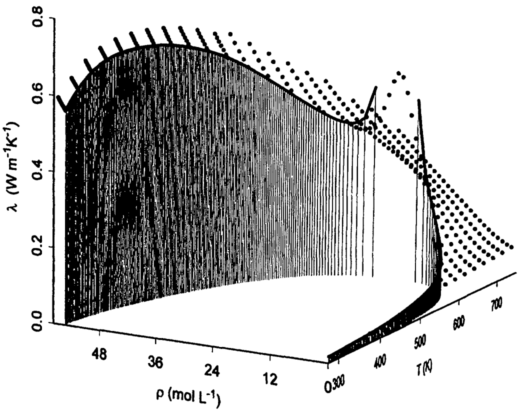

FIG. 7. Thermal-conductivity surface for water from IAPWS model (Sengers and Watson, 1986).

cal point for water is at T_c = 647.096 K, ρ_c = 17.87 mol/L, and P_c = 22.064 MPa, and the triple point is at about T_t = 273.16 K, ρ_t = 55.50 mol/L, and P_t = 611.657 Pa. The compressed-liquid region again exists at pressures above the saturation pressure between T_t and T_c. Water is very polar, with a dipole moment of about 6.14×10^{-24} C·m (1.84 debyes) and is very unusual since the solid line has a negative slope at pressures from the triple point to 209.9 MPa. This unusual behavior can also be seen in the thermal conductivity of saturated liquid water. Figure 7 shows that the thermal conductivity of saturated liquid water increases from the triple point to a maximum of about 0.68 W/(m K) near 400 K. The thermal conductivity of most liquids at saturation is a monotonically decreasing function of temperature as seen in argon. However, in water, we have a competing mechanism related to the association phenomenon; at low temperatures, the effective diameter of the fluid constituents is increased, so that the thermal conductivity is decreased, according to the structure of Eq. (11). The effect of the critical enhancement of thermal conductivity is apparent in the saturated-liquid and vapor lines as they approach the critical point in Fig. 7.

3.1 Saturated Liquid

At saturation, a liquid is in equilibrium with vapor. For a given temperature between the triple point and the critical point, this occurs at the saturation vapor pressure. The properties at saturation are a unique function of temperature for any given pure fluid. Properties such as the saturated-liquid density, saturated-vapor density, and thermal conductivity of the liquid or vapor at saturation are specified by one variable, normally either temperature or pressure. The thermal conductivity of most liquids decreases as a function of temperature at saturation. Figure 8 shows the thermal conductivity of several liquid alkanes at saturation as calculated with the NIST computer databases for thermophysical properties of Friend and Huber (1994). An upturn is present as the gas–liquid critical point is approached since the thermal conductivity is divergent at the critical point. The critical temperatures of methane, ethane, propane, and *n*-butane are near 190, 305, 370, and 425 K, respectively. The lowest temperatures shown in Fig. 8 correspond to the highest fluid densities and hence have the highest thermal conductivity for any given fluid. The curves in Fig. 8 cross each other in the temperature coordinate system, but would not if plotted in terms of saturated-liquid density. This reflects the weak dependence of liquid thermal conductivity on temperature. Measurement of the thermal conductivity directly along the saturated-liquid line is not possible since a temperature gradient is normally required during the measurement process so that the phase equilibrium will be disturbed. Normally measurements are made in the vicinity of the saturation line and extrapolated to the saturation boundary to obtain the true saturated-liquid thermal conductivities. It is possible to cross over the phase boundary into the two-phase region during a measurement when very clean surfaces or rapid experi-

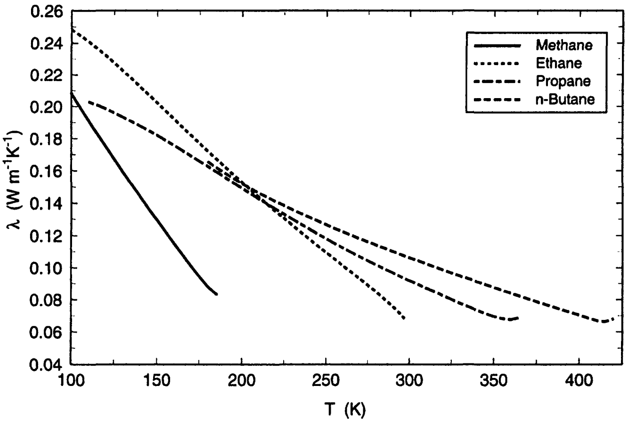

FIG. 8. Family plot of methane, ethane, propane, *n*-butane thermal conductivity calculated using the NIST database correlations.

ments suppress the nucleation of bubbles in the liquid during the measurement process. Under these conditions the fluid is not in equilibrium but is instead in a metastable state. Such data for metastable thermal conductivity have significance in industrial heat-transfer equipment.

3.2 Compressed Liquid

The thermal conductivity of a liquid that is compressed to a pressure and density greater that its saturation value is generally larger than its saturation value. Figure 9 illustrates this effect using the hard-sphere correlation of Assael *et al.* (1992b) for liquid propane at pressures to 70 MPa. This figure shows that the thermal conductivity of liquid propane is a strong function of density with a much weaker dependence on temperature. It is the temperature dependence that keeps the isotherms at different temperatures from falling on a single density-dependent curve. To study this temperature dependence it is necessary to have data over a wide range of pressure from saturation to 50 MPa or more. The curvature of the isotherms becomes more pronounced as the temperature is increased since the fluid becomes more compressible. The figure further illustrates how data along isotherms can be used to extrapolate to the thermal conductivity of the saturated liquid.

3.3 Critical-Region Behavior

Application of simple models, such as the hard-sphere theory, for the thermal conductivity of liquids must be limited to temperatures where the critical enhancement is negligible. Thus, hard-sphere theories work best at lower reduced temperatures. Figure 5 showed the behavior of the thermal conductivity of carbon dioxide in the critical region for liquid, vapor, and supercritical isotherms. Figure 10 shows the relative magnitude of the critical enhancement that is calculated for 1,1,1,2-tetrafluoroethane (R134a) along its vapor pressure locus using the correlation surface of Krauss *et al.* (1993). The critical temperature of R134a is about 374.2 K. The critical enhancement contributes more than 2% to the saturated-liquid thermal conductivity for temperatures above 300 K and exceeds 15% at temperatures above 360 K. Correlations and theories for thermal conductivity of liquids that do not consider the critical enhancement should be limited to temperatures below about $0.85T_c$. The contribution of the critical enhancement is even more significant in the vapor phase where the critical enhancement adds to a smaller background thermal conductivity as shown in Fig. 10. A discontinuity is present in the correlation of Krauss *et al.* (1993) for the saturated vapor near 360 K where the equation of state switches from a classical form ($T <$ 360 K) to a crossover form for the near-critical region.

4. MIXTURES

The topology of the phase behavior for mixtures can be very complex, because multiple liquid phases may be present, the sim-

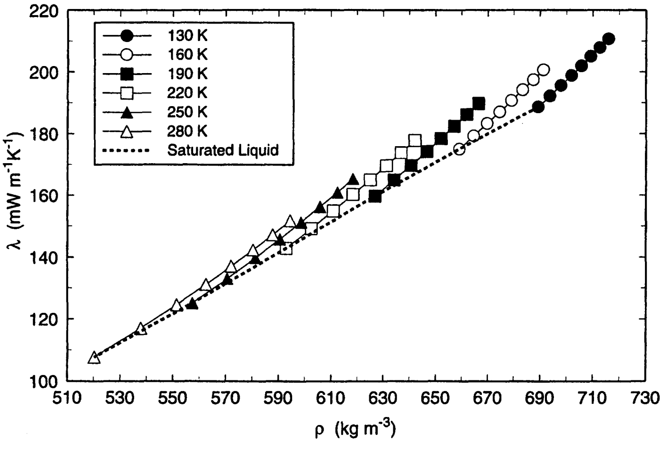

FIG. 9. Thermal conductivity of compressed liquid propane at low reduced temperatures from the hard-sphere correlation of Assael *et al.* (1992b).

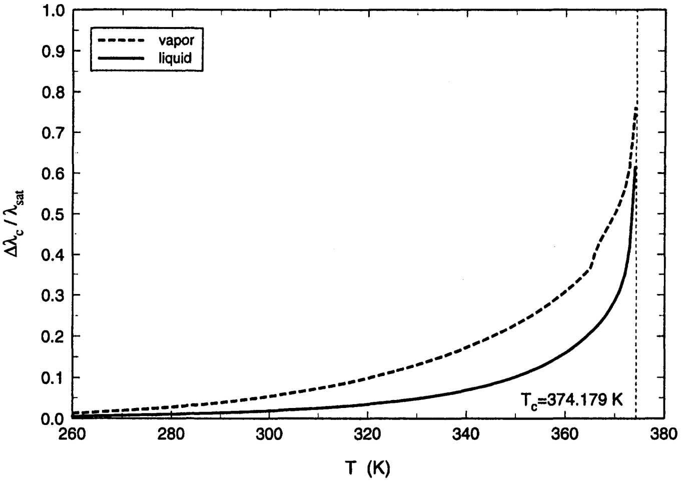

FIG. 10. Fraction of the thermal conductivity of 1,1,1,2-tetrafluoroethane (R134a) due to critical enhancement along the saturation boundaries. Calculated from Krauss *et al.* (1993) correlation.

ple melting line can be complicated by eutectic points and multiple solid phases, and the liquid–vapor critical locus can have a variety of behaviors. Nonetheless, the general behavior of the thermal conductivity of liquid mixtures is similar to that of pure liquids. As an example, Fig. 11 shows thermal conductivity data for a standard dry air, which contains 78.12% nitrogen, 20.96% oxygen, and 0.92% argon, where the composition is expressed as mole fractions. These data were obtained using the transient hot-wire technique (Perkins and Cieszkiewicz, 1991), and the thermal conductivity of air looks very similar to the thermal conductivity of argon that was shown in Fig. 6. For this standard composition, there is a triple point near 59.75 K (although the coexisting solid and vapor phases have different compositions), and the liquid–vapor coexistence boundary extends to its maximum temperature near 132.62 K (which is slightly above the critical temperature near 132.52 K).

In Fig. 12, we have shown the extensive data (and correlating surface) for the methane–ethane binary system, which was illustrated in reduced units in Fig. 4. Again, the density dependence dominates for the compressed fluid, and the mixture values of thermal conductivity interpolate smoothly between the pure-fluid values. The data for each mixture composition exhibit the char-

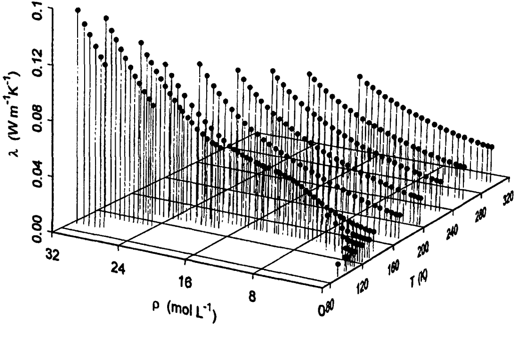

FIG. 11. Thermal-conductivity data surface for air reported by Perkins and Cieszkiewicz (1991).

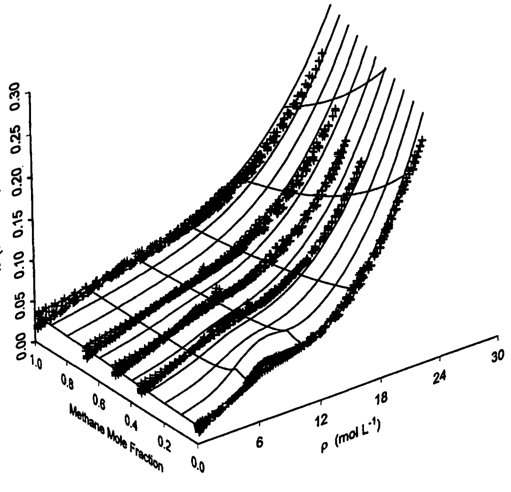

FIG. 12. Thermal-conductivity surface for methane–ethane binary mixtures reported by Friend and Roder (1987).

acteristics that have been described for pure fluids. At a fixed density, the thermal conductivity generally increases with increasing temperature. The thermal-conductivity critical enhancement is responsible for isotherms at lower temperatures crossing higher-temperature isotherms and the bumps that are centered near the composition-dependent critical density in Fig. 12.

For mixtures, such as air or methane + ethane, the liquid–vapor phase boundary for a fixed composition is defined by the dew-point locus and the bubble-point locus rather than a simple vapor-pressure curve. The bubble point gives the state at which the first bubble of vapor will form when a compressed liquid of fixed composition is brought to a lower pressure along an isotherm. Generally, the vapor will have a different composition than the liquid with which it coexists. The dew-point locus lies at lower pressures and gives the temperature and pressure where the first condensation occurs when a vapor at fixed composition is isothermally compressed. Again, the liquid composition generally differs from the composition of the original vapor. The dew-point locus and bubble-point locus can meet at the critical point, which represents the limiting state at which the vapor-phase and liquid-phase properties are identical. However, the critical point does not necessarily coincide with the maximum temperature for phase separation, which is often called the maxcondentherm temperature. When we discuss the thermal conductivity of a saturated-liquid mixture, we mean that the temperatures and pressures are along the bubble-point locus. Measurements of the thermal conductivity at the bubble point are complicated because some of the heat input can generate bubbles of gas. In Fig. 13, we show calculated values of the thermal conductivity for three binary mixtures of nitrogen and oxygen along their bubble-point curves and the analogous values for the saturated liquid of the pure components. Again, the similarity of these curves is readily apparent.

The behavior of the thermal conductivity of mixtures in the region of the critical point remains a subject of experimental and theoretical interest. In binary systems such as methane + ethane (Roder and Friend, 1985), helium-3 + helium-4 (Meyer, 1984), and carbon dioxide + ethane (Leuttmer-Strathmann and Sengers, 1996), this behavior looks similar to that seen in pure fluids. However, in contrast to pure fluids, theoretical calculations of the critical enhancement term indicate that it should have a finite value as the critical point is approached; this maximum is difficult to observe experimentally because of its close proximity (several millikelvins) to the critical point.

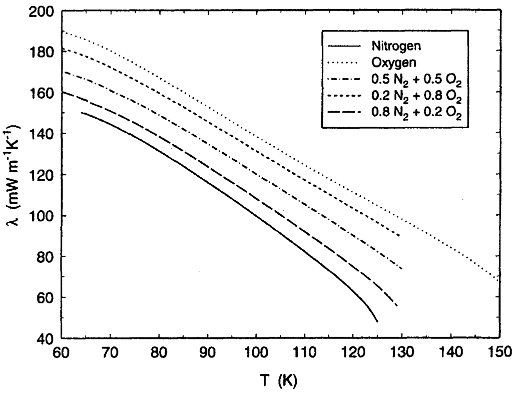

FIG. 13. Thermal conductivity of mixtures of nitrogen and oxygen along bubble-point locus.

5. MISCELLANEOUS ADDITIONAL TOPICS

The previous sections have provided the basis for approaching the wide range of literature that is available on the thermal conductivity of liquids. It is now apparent that there are difficulties in separating other modes of heat transfer from pure thermal conduction during experimental measurements. This is important to consider when assessing work in the literature where widely different values for thermal conductivity are reported for the same fluid. The need for thermal-conductivity data in engineering heat-transfer calculations has also been discussed. Theoretical tools that are available for the correlation of liquid thermal-conductivity data have been summarized. The fact that liquid thermal conductivity cannot be considered in isolation from the thermal conductivity of the vapor and supercritical gas has also been illustrated. It remains to consider some interesting cases that are important but that did not fit well in the context of the previous discussion.

5.1 Cryogenic Fluids

The discussions above have assumed that classical descriptions are adequate, and this is true for normal liquid systems. However, quantum mechanics represents a more fundamental approach to the understanding of physical phenomena, and we will very briefly discuss situations in which this may become important. For a theoretical understanding of the internal modes of energy storage in a molecular system, the quantization of available states may be useful; see, e.g., Eq. (15). From the point of view of statistical physics, quantum effects become important when the thermal de Broglie wavelength $[= h/(2\pi \times mkT)^{1/2}]$ becomes comparable to the interparticle spacing in the fluid system. From this we can obtain a dimensionless parameter Λ defined by

$$\Lambda \approx 1.474\rho^{1/3}/M^{1/2}T^{1/2},$$

where M is the usual molar mass (in g/mol), T is in kelvins, and ρ is in moles per cubic decimeter. When Λ is comparable to or greater than 1, quantum effects must be considered. For water at ambient conditions, $\Lambda \approx 0.08$, and the system is essentially classical. On the other hand, for light molecules at low temperatures, quantum effects are important. Thus, for liquid helium near 2 K, $\Lambda \approx 1.7$, and the system is, in fact, in the He II superfluid state at temperatures below the lower lambda point near 2.18 K. Heat transfer in this system is not well defined by Fourier's law, Eq. (1), and for some flows the heat-flux vector is proportional to the cube root of the temperature gradient (Srinivasan and Hofmann, 1985). Molecular hydrogen is another system in which quantum effects play an important role. In this fluid, the relative spins of the nucleon can be parallel (orthohydrogen) or antiparallel (parahydrogen), and these modifications have distinct energy states available. The "normal" population of a system of fluid hydrogen contains a 3:1 ratio of ortho-to-para modifications, but the parahydrogen molecule dominates for equilibrium systems at very low temperatures. Because the conversion rate between the two modifications is slow in the absence of a catalyst, measurements of the thermal conductivity of hydrogen can give different results depending on the history of the system. We can calculate and observe different dilute-gas values for the thermal conductivity of the two modifications, and in fact the fluid can be considered a mixture with each component having a different value for the thermal conductivity. Empirically, we can describe the thermal conductivity of the "normal" population, the equilibrium population, or the value for each pure modification (Roder, 1984). However, for the liquid phase, which exists below the critical temperature of about 33 K, these differences in thermal conductivity are small.

5.2 Refrigeration Fluids

Concerns with the environmental problems associated with chlorofluorocarbon (CFC) refrigerants have led to very active research on the thermophysical properties, including thermal conductivity, of alternative fluids such as the hydrochlorofluorocarbons (HCFCs), hydrofluorocarbons (HFCs), propane, butanes, and their mixtures. There is a rich collection of recent data and correlations for the thermal conductivity of these alternative refrigerants. For the thermal conductivity of a wide variety of alternative refrigerants and their mixtures in the liquid phase, there are two notable correlation packages. First, the hard-sphere model of Assael *et al.* (1995) provides estimates for the density, viscosity, and thermal conductivity of nine refrigerants; the model is embodied in a computer package called TRANSP. Second, there is the extended corresponding-states model of Friend and Huber (1994), which is available in REFPROP from NIST. The current NIST REFPROP computer program provides estimates or correlations for density, entropy, enthalpy, specific heat, sound speed, compressibility, phase equilibria, viscosity, and thermal conductivity of 43 fluids and mixtures of these fluids with up to five components. Both of these correlation schemes have been compared extensively to the most reliable data that are available.

5.3 Liquid Fuels and Petrochemicals

There has been a long-term interest and extensive literature on the thermal conductivity of liquids that are used as fuels or as feedstocks in petrochemical processes. These materials can range in complexity from pure fluids such as hydrogen and methane to mixtures of numerous components, where it is only possible to characterize the boiling-point range of the liquid. Other nonpolar liquids such as nitrogen, oxygen, argon, carbon monoxide, carbon dioxide, and hydrogen sulfide can be considered in this group. These materials can all be treated well with extended corresponding-states correlations or with hard-sphere correlations. Assael *et al.* (1992b) describe their hard-sphere model for the self-diffusion, viscosity, and thermal conductivity of *n*-alkanes from methane to hexadecane. This work has been extended to include mixtures of *n*-alkanes (Assael *et al.*, 1992c), of aromatic hydrocarbons (Assael *et al.*, 1992d), and of alcohols (Assael *et al.*, 1994). All of these materials, and the refrigerants described above, are included in the computer package called TRANSP developed by Assael, Dymond, and co-workers. Hard-sphere correlations for the thermal conductivity and viscosity of a wide range of polar liquids are also provided by Bleazard and Teja (1996). A total of 58 polar liquids, including alcohols, diols, disulfides, amines, carboxylic acids, alcohol-ethers, pyridines, ethanol-amines, ethanoates, and poly(ethylene glycol)s are represented by this correlation at temperatures from 293 to 493 K.

Thermal-conductivity correlations for many of these materials are contained in three NIST databases: NIST12, NIST14, and NIST SUPERTRAPP (Friend and Huber, 1994). NIST12 contains detailed information on a wide range of thermophysical properties, including liquid thermal conductivity, for pure light gases, including *n*-alkanes up to butane. NIST14 is a mixture correlation that contains many of the materials in NIST12 plus *n*-alkanes to hexane, and hydrogen sulfide. The extended corresponding-states technique used in NIST14 uses propane as a reference fluid, and exact shape factors are computed from the pure-fluid correlations for each of the components in this database. NIST14 also uses the pure-fluid correlation directly when a pure fluid is selected. The NIST SUPERTRAPP database currently has information on 116 fluids for which there is much less reliable data available. The NIST SUPERTRAPP program can estimate the properties of pure fluids and mixtures containing up to 20 components using extended corresponding states. The package does not currently contain information on alcohols but does include normal and branched alkanes to *n*-tetracosane, cyclic and aromatic hydrocarbons, alkenes, and di-alkenes. The extended corresponding-states model of NIST SUPERTRAPP has recently been extended to undefined liquid mixtures by Baltatu *et al.* (1995), and requires only the specific gravity and the mean average boiling point as input parameters to obtain estimates of the liquid thermal conductivity. A collection of liquid thermal-conductivity data and correlations can also be found in the American Institute of Chemical Engineers DIPPR Data Compilation of Pure Compound Properties.

5.4 Aqueous Systems

The thermal conductivity of water is best described by the water model of the International Association for the Properties of Water and Steam (IAPWS) and can be used as the basis of property calculations associated with processes using steam as the working fluid, such as electric power generation (Sengers and Watson, 1986). Many measurements have been reported for the thermal conductivity of aqueous salt solutions, but no global correlation scheme has been reported as for the case of refrigerants, fuels, and petrochemicals largely because of theoretical difficulties. The maximum in the saturated-liquid thermal conductivity of water near 400 K, shown in Fig. 7, provides some insight into the difficulties in predicting the thermal conductivity of mixtures containing water as a component. There is certain to be new development in the description of the thermal conductivity of aqueous mixtures because of their industrial and environmental importance.

GLOSSARY

Bubble-Point Locus: The locus of state points, characterized by temperature, pressure, etc., where the first bubble of gas begins to form in a liquid mixture of defined composition at equilibrium.

Corresponding States: When the intermolecular potentials of two fluids are conformal, such that characteristic energy (temperature) and distance (density) parameters can be properly scaled, the thermophysical properties of one fluid can be mapped to those of another.

Critical Point: In a pure fluid the gas–liquid critical point is where the vapor-pressure curve terminates. In a mixture the gas–liquid critical point is the point where the dew-point locus meets the bubble-point locus.

Dilute-Gas Thermal Conductivity: The dilute-gas thermal conductivity is defined in the limit of zero density. This is not physically attainable since heat transfer at zero density is due to molecular conduction (collisions of molecules with enclosure surfaces rather than collisions between molecules themselves), so that extrapolation from higher densities is necessary.

Excess Thermal Conductivity: The contribution to total thermal conductivity from increasing the density (pressure) of the fluid.

Hard-Sphere Model: A liquid-phase theory that accounts for properties based on collisions between hard-sphere molecules.

Heat-Transfer Coefficient: A lumped parameter that is used by engineers to account for complex heat-transfer situations with thermal conduction, fluid convection, thermal radiation, and phase changes. Temperature differences are system specific, usually between a surface and the bulk fluid.

Thermal Conductivity: A linear coefficient that is defined by Fourier's law for heat conduction. This property characterizes heat transfer through a medium that is stationary, opaque to thermal radiation, and at steady state.

Thermal-Conductivity Critical Enhancement: The thermal conductivity of a pure fluid diverges at the gas–liquid critical point with a characteristic critical exponent. The thermal conductivity is enhanced in a region of temperature and density in the vicinity of the critical point and makes a smooth transition from divergent to background values in this crossover region.

Thermal Diffusivity: A linear coefficient that characterizes the transient diffusion of heat through a medium that is stationary and opaque to thermal radiation.

Vapor-Pressure Locus: The locus of points (temperature, pressure, and densities) where vapor and liquid coexist for a pure fluid. The vapor-pressure locus terminates in the gas–liquid critical point.

Works Cited

Alder, B. J., Gass, D. M., Wainwright, T. E. (1970), *J. Chem. Phys.* **53,** 3813–3826.

Assael, M. J., Nieto de Castro, C. A. N., Roder, H. M., Wakeham, W. A. (1991), in: W. A. Wakeham, A. Nagashima, J. V. Sengers (Eds.), *Measurement of the Transport Properties of Fluids,* Oxford, U.K.: Blackwell Scientific Publications, pp. 164–194.

Assael, M. J., Dymond, J. H., Papadaki, M., Patterson, P. M. (1992a), *Fluid Phase Equilibria,* **75,** 245–255.

Assael, M. J., Dymond, J. H., Papadaki, M., Patterson, P. M. (1992b), *Int. J. Thermophys.* **13,** 269–291.

Assael, M. J., Dymond, J. H., Papadaki, M., Patterson, P. M. (1992c), *Int. J. Thermophys.* **13,** 659–669.

Assael, M. J., Dymond, J. H., Patterson, P. M. (1992d), *Int. J. Thermophys.* **13,** 895–905.

Assael, M. J., Dymond, J. H., Patterson, P. M. (1994), *Int. J. Thermophys.* **15,** 189–201.

Assael, M. J., Dymond, J. H., Polimatidou, S. K. (1995), *Int. J. Thermophys.* **16,** 761–772.

Aziz, R. A., Chen, H. H. (1977), *J. Chem. Phys.* **67,** 5719–5726.

Baltatu, M. E., Chong, R. A., Huber, M. L. (1995), in: A. Nagashima (Ed.), *The Fourth Asian Thermophysical Properties Conference,* Tokyo: Keio Univ., pp. 531–537.

Bleazard, J. G., Teja, A. S. (1996), *Ind. Eng. Chem. Res.* **35,** 2453–2459.

Chapman, S., Cowling, T. G. (1939), *The Mathematical Theory of Non-uniform Gases,* London: Cambridge Univ. Press.

Czarnetzki, W., Roetzel, W. (1994), in: T. Tong (Ed.), *Thermal Conductivity 22,* Lancaster, PA: Technomic Publishing, pp. 276–286.

Eggenberger, R., Gerber, S., Huber, H., Searles, D., Welker, M. (1992), *Mol. Phys.* **76,** 1213–1219.

Ely, J. F., Hanley, H. J. M. (1983), *I&EC Fund.* **22,** 90–97.

Eucken, A. (1913), *Z. Phys.* **14,** 324–332.

Friend, D. G., Huber, M. L. (1994), *Int. J. Thermophys.* **15,** 1279–1288.

Friend, D. G., Roder, H. M. (1987), *Int. J. Thermophys.* **8,** 13–26.

Hanley, H. J. M., Evans, D. J. (1996), in: J. Millat, J. H. Dymond, C. A. Nieto de Castro (Eds.), *Transport Properties of Fluids: Their Correlation, Prediction, and Estimation,* New York: Cambridge Univ. Press, pp. 210–225.

Hanley, H. J. M., Klein, M. (1972), *J. Phys. Chem.* **76,** 1743–1751.

Heck, E. L., Dickinson, A. S. (1994), *Mol. Phys.* **81,** 1325–1352.

Hoheisel, C., Vogelsang, R. (1988), *Comp. Phys. Rep.* **8,** 1–70.

Huber, M. L., Friend, D. G., Ely, J. F. (1992), *Fluid Phase Equil.* **80,** 249–261.

Krauss, R., Luettmer-Strathmann, J., Sengers, J. V., Stephan, K. (1993), *Int. J. Thermophys.* **14,** 951–988.

Leach, J. W., Chappelear, P. S., Leland, T. W. (1968), *A.I.Ch.E. J.* **14,** 568–576.

Le Neindre, B. (1987), in: K. N. Marsh. (Ed.), *Recommended Reference Materials for the Realization of Physicochemical Properties,* Oxford, U.K.: Blackwell Scientific Publications, pp. 322–370.

Le Niendre, B., Tufeu, R., Sirota, A. M. (1991), in: W. A. Wakeham, A. Nagashima, J. V. Sengers (Eds.), *Measurement of the Transport Properties of Fluids,* Oxford, U.K.: Blackwell Scientific Publications, pp. 113–160.

Leuttmer-Strathmann, J., Sengers, J. V. (1996), *J. Chem. Phys.* **104,** 3026–3047.

Mason, E. A., Monchick, L. (1962), *J. Chem. Phys.* **36,** 1622–1639.

Matos Lopes, M., Ramires, M. L. V., Nieto de Castro, C. A., Kraft, K., Leipertz, A., Perkins, R. A. (1995), in: A. Nagashima (Ed.), *The Fourth Asian Thermophysical Properties Conference,* Tokyo: Keio University, pp. 795–798.

McQuarrie, D. A. (1976), *Statistical Mechanics,* New York: Harper & Row.

Meyer, H. (1984), *Physica B + C* **126,** 326–327.

Michels, A., Sengers, J. V., Van der Gulik, P. S. (1962), *Physica* **28,** 1216–1237.

Mohammadi, S. S., Graboski, M. S., Sloan, E. D. (1981), *Int. J. Heat Mass Trans.* **24,** 671–683.

Nieto de Castro, C. A., Li, S. F. Y., Nagashima, A., Trengove, R. D., Wakeham, W. A. (1986), *J. Phys. Chem. Ref. Data* **15,** 1073–1086.

Olchowy, G. A., Sengers, J. V. (1988), *Phys. Rev. Lett.* **61,** 15–18.

Perkins, R. A., Cieszkiewicz, M. T. (1991), *NISTIR 3961: Experimental Thermal Conductivity, Thermal Diffusivity, and Specific Heat Values for Mixtures of Nitrogen, Oxygen, and Argon,* Boulder, CO: National Institute of Standards and Technology, pp. 1–22.

Perkins, R. A., Roder, H. M., Nieto de Castro, C. A. (1991a), *J. Res. Natl. Inst. Stand. Tech.* **96,** 247–269.

Perkins, R. A., Friend, D. G., Roder, H. M., Nieto de Castro, C. A. (1991b), *Int. J. Thermophys.* **12,** 965–984.

Poltz, H., Jugel, R. (1982), *Int. J. Heat Mass. Trans.* **25,** 1093–1101.

Ramires, M. L. V., Nieto de Castro, C. A., Nagasaka, Y., Nagashima, A., Assael, M. J., Wakeham, W. A. (1995), *J. Phys. Chem. Ref. Data* **24,** 1377–1381.

Ramires, M. L. V., Nieto de Castro, C. A., Perkins, R. A., Nagasaka, Y., Nagashima, A., Assael, M. J., Wakeham, W. A. (1997), Submitted to *J. Phys. Chem. Ref. Data.*

Ravi, P., Murad, S., Hanley, H. J. M., Evans, D. J. (1992), *Fluid Phase Equil.* **76,** 249–257.

Reid, R. C., Prausnitz, J. M., Polling, B. E. (1987), *The Properties of Gases and Liquids,* 4th ed., New York: McGraw Hill.

Roder, H. M. (1982), *J. Res. Natl. Bur. Stand. (U.S.)* **87,** 279–309.

Roder, H. M. (1984), *Int. J. Thermophys.* **5,** 323–350.

Roder, H. M., Friend, D. G. (1985), *Int. J. Thermophys.* **6,** 607–617.

Sengers, J. V. (1985), *Int. J. Thermophys.* **6,** 203–232.

Sengers, J. V., Watson, J. T. R. (1986), *J. Phys. Chem. Ref. Data* **15,** 1291–1314.

Shaumeyer, J. N., Gammon, R. W., Sengers, J. V., Nagasaka, Y. (1991), in W. A. Wakeham, A. Nagashima, J. V. Sengers (Eds.), *Measurement of the Transport Properties of Fluids,* Oxford, U.K.: Blackwell Scientific Publications, pp. 197–225.

Shpilrain, E. E., Umanskii, A. S., Gorshkov, Y. A. (1975), *High Temp.—High Press.* **7,** 361–393.

Snider, R. F. (1960), *J. Chem. Phys.* **32,** 1051–1060.

Srinivasan, R., Hofmann, A. (1985), *Cryogenics* **25,** 652–657.

Vesovic, V., Wakeham, W. A., Olchowy, G. A., Sengers, J. V., Watson, J. T. R., Millat, J. (1990), *J. Phys. Chem. Ref. Data* **19,** 763–808.

Waldmann, L. (1957), *Z. Naturforsch.* **12a,** 660–662.

Wang Chang, C. S., Uhlenbeck, G. E., de Boer, J. (1964), in: J. de Boer, G. E. Ulhenbeck, (Eds.), *Studies in Statistical Mechanics,* Amsterdam: North-Holland, Vol. 2, pp. 249–268.

Further Reading

Books

Bird, R. B., Stewart, W. E., Lightfoot, E. N. (1960), *Transport Phenomena,* New York: Wiley.

Hirschfelder, J. O., Curtiss, C. F., Bird, R. B. (1954), *Molecular Theory of Gases and Liquids,* New York: Wiley.

Marsh, K. N. (1987), *Recommended Reference Materials for the Realization of Physicochemical Properties,* Oxford, U.K.: Blackwell Scientific Publications.

McCourt, F. R. W., Beenakker, J. J. M., Kohler, W. E., Kuscer, I. (1991a), *Nonequilibrium Phenomena in Polyatomic Gases—Dilute Gases,* Oxford, U.K.: Oxford Univ. Press.

McCourt, F. R. W., Beenakker, J. J. M., Kohler, W. E., Kuscer, I. (1991b), *Nonequilibrium Phenomena in Polyatomic Gases—Cross Sections, Scattering and Rarefied Gases,* Oxford, U.K.: Oxford Univ. Press.

Millat, J., Dymond, J. H., Nieto de Castro, C. A. (Eds.) (1996), *Transport Properties of Fluids: Their Correlation, Prediction, and Estimation,* New York: Cambridge Univ. Press.

Perry, R. H., Green, D., Maloney, J. O. (1984), *Perry's Chemical Engineers' Handbook,* New York: McGraw-Hill.

Vargaftik, N. B., Filippov, L. P., Tarzimanov, A. A., Totskii, E. E. (1994), *Handbook of Thermal Conductivity of Liquids and Gases,* Boca Raton, FL: CRC.

Wakeham, W. A., Nagashima, A., Sengers, J. V. (1991), *Measurement of the Transport Properties of Fluids,* Oxford, U.K.: Blackwell Scientific Publications.

THERMAL CONDUCTION IN SOLIDS

Paul G. Klemens, *Department of Physics and Institute of Materials Science, University of Connecticut, Storrs, Connecticut, U.S.A.*

Introduction 133
1. Basic Concepts 133
2. Measurements 134
2.1 Steady-State Methods 135
2.2 Dynamic Methods 135
3. Lattice Waves 136
4. Thermal Conductivity of Insulators 137
4.1 Intrinsic Thermal Conductivity 137
4.2 Boundary Scattering 138
4.3 Defects 138
5. Electronic Thermal Conductivity 139
6. Metals 140
7. Alloys 141
8. Semimetals and Semiconductors 143
9. Superconductors 144
10. Inhomogeneous Conductors 144
11. Special Topics 145
11.1 Heat-Conducting Substrates ... 145
11.2 Amorphous Solids 146
11.3 Thermal-Barrier Coatings 146
11.4 Layered Crystals: Graphite 147
11.5 Radiation Damage 148
Glossary 148
Works Cited 149
Numerical Data 150
Further Reading 150

INTRODUCTION

The thermal conductivities of various solids show differences in both magnitude and temperature dependence. These result from different mechanisms of heat transport: lattice waves, free electrons, and sometimes internal electromagnetic radiation. The interaction and scattering processes that limit the contributions of these carriers have been studied not only to understand basic processes in solids, but also to understand thermal conductivity better, so that the available data can be interpolated and extended for technical purposes beyond the measured values.

1. BASIC CONCEPTS

A temperature difference within a solid causes heat flow. At any place the local value of the heat current per unit area **Q** is proportional to the local temperature gradient grad *T*. This linear relation or Fourier's law

$$\mathbf{Q} = -\boldsymbol{\lambda}\cdot \mathrm{grad}\, T \qquad (1)$$

defines the thermal conductivity $\boldsymbol{\lambda}$. In anisotropic solids it is a symmetric tensor. In isotropic materials, such as a polycrystalline aggregate without preferred orientation, $\boldsymbol{\lambda}$ becomes a scalar. In crystals of cubic symmetry, $\boldsymbol{\lambda}$ has only diagonal components equal to each other, and the conductivity acts like a scalar. For scalar λ the heat current is parallel to grad *T*. Polycrystalline cubic solids will be isotropic more readily than other polycrystalline materials.

In the definition of thermal conductivity it is implied that there is no transport of matter and no electric current. Hence a temperature gradient is generally accompanied by a thermoelectric field. Convection can be disregarded in solids, but infrared radiative heat transfer can occur within dielectric solids at high temperatures, provided they do not first become conducting and thus opaque.

The SI unit of thermal conductivity is the $\mathrm{W \cdot m^{-1} \cdot K^{-1}}$, though the $\mathrm{W \cdot cm^{-1} \cdot K^{-1}}$ is still frequently used. Conversion between various thermal conductivity units is given in the article Heat Transfer.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

The principal carriers of heat in solids are lattice vibrations, free electrons, and electromagnetic radiation. The first is a mechanical transfer of energy between atoms, and this lattice thermal conductivity is present in all solids. It can often be regarded as energy flow carried by lattice waves, i.e., high-frequency elastic waves. In electrically conducting solids, thermal energy is also transferred by mobile electrons, and this electronic component of thermal conduction is related to the electrical conductivity. In transparent or translucent solids there can be transport by electromagnetic radiation, which is often size dependent.

Thus the thermal conductivity of solids shows a variety of behavior that depends on which component is dominant. The range of variation is not as wide as that of electrical conductivity, since the lattice component provides a lower limit to the thermal conductivity.

One can express the thermal conductivity as a sum of contributions by different carriers (a), i.e.,

$$\lambda = \sum_a \lambda_a = \sum_a \tfrac{1}{3} C_a v_a l_a, \tag{2}$$

where C_a is the contribution of carrier a to the specific heat per unit volume, v_a its speed, and l_a its mean free path. For each carrier, λ_a can be similarly divided into contributions from subgroups (bands and energy ranges for electrons, polarization, and frequency ranges for lattice waves). The mean free path of each carrier is limited by interactions with lattice vibrations and with the other carriers, and also by scattering at crystal imperfections, such as impurities and solutes, vacancies, dislocations, and boundaries. The lower the temperature, the greater is the relative importance of imperfections.

Table 1 gives thermal conductivities of a few solids, ranging from metals and alloys to semiconductors and insulators. This illustrates differences in behavior and ranges of variability.

2. MEASUREMENTS

The thermal conductivity λ of solids can be measured by steady-state methods or can be deduced by transient measurements that yield the thermal diffusivity α, which is defined by

$$\lambda = \alpha C, \tag{3}$$

where C is the specific heat *per unit volume.* To obtain λ from α requires a knowledge of the specific heat, but this property is not nearly as sensitive to crystal imperfections as are the thermal conductivity and diffusivity. Therefore, one can use specific-heat data obtained separately, except for ceramics at high temperatures, where an electronic contribution may depend on impurities.

Thermal-diffusivity measurements have the advantage that they do not require measuring quantities of heat; only precise knowl-

Table 1. Thermal conductivity of some solids ($\mathrm{W \cdot cm^{-1} \cdot K^{-1}}$).

T (K)	20	90	300	600	1200	1500
Gold	15	3.48	3.15	3.04	2.62	
Copper	105	5.14	3.98	3.83	3.42	
Silver	51	4.60	4.27	4.05	3.6	
Nickel	8.56	1.72	0.95	0.655	0.761	0.381
Cu-30%Zn	0.300	0.664	1.22	1.59		
Silicon	49.4	10.8	1.48	0.619	0.22	0.25
SiC	22	30	4.9	2.0	0.53	0.2
Graphite ATJ	0.03	0.5	1.29	0.95	0.55	0.47
Pyrolytic graphite $\parallel$	4.2	47.3	20.0	9.3	4.4	3.4
Pyrolytic graphite $\perp$	1.1	0.46	0.095	0.04	0.021	0.015
Al_2O_3 (sapphire)	160	6.4	0.46	0.19		
Al_2O_3 (sintered)		1.5	0.36	0.16	0.07	0.06
MgO (crystal)	33	3.4	0.6			
MgO (sintered)			0.48	0.21	0.077	0.065
SiO_2 (c-axis)	7.2	0.45	0.055			
SiO_2 (fused)	0.001 53	0.0062	0.0138	0.0175	0.04	

edge of distance and time are needed. If heat input and temperature excursions are also measured, one can obtain specific heat and thermal conductivity as well. However, in the widely used flash diffusivity technique, which measures relative temperature excursions at well-defined distances and times, one only obtains diffusivities but avoids heat measurements.

2.1 Steady-State Methods

Thermal conductivity is measured using a steady heat current, produced by a heater of known power. The current is made to flow through a specimen of regular shape. The resulting temperature gradient either is uniform or has a variation depending on the geometrical symmetry of the arrangement. It is determined by two or more thermometers at defined locations. The specimens can be rods, linking a heater and a heat sink, or plates of large area, with the heat input uniformly distributed on one side, or again they may be long cylinders with a uniform axial heat source. The problem is to ensure that all the heat that is supposed to flow through the specimen actually does so, and in a uniform manner, which is not always easy. Thermometer connections and other leads can divert heat flow. Radiation and gas conduction cause leakage from the heater and, in the case of rods, also from the side walls. Gas conduction can be reduced by evacuating the surroundings.

Radiation increases rapidly with temperature. It can be disregarded only at low temperatures, below 50 K and sometimes below 100 K or even somewhat higher, depending on the conductivity and the shape of the specimen. At higher temperatures, the specimen should be shorter, but this makes it harder to define thermometer positions. Sideways radiation can be reduced by fine powder packed around the specimen. Guard cylinders, matching the temperature drop, reduce the problem but do not eliminate it.

Comparative methods have been used with success: the same heat current is made to flow through a standard and the unknown specimen; the temperature gradients in each are compared. Each conductivity range requires a separate standard.

Since thermal conductivities vary with temperature, the temperature difference between thermometers should be small compared to the absolute temperature; this requires accurate thermometry. One must also ensure that a steady-state temperature distribution and heat flow have been reached. This requires patience and judgment. Automatic data acquisition can be of great help.

There is no quick and easy way to measure thermal conductivities. Difficulties increase with increasing temperature and decreasing conductivity. Different ranges of these values require different geometries.

2.2 Dynamic Methods

The time-dependent Fourier equation for heat conduction is

$$\frac{\partial T}{\partial t} + \alpha \operatorname{div} \operatorname{grad} T = 0, \tag{4}$$

where T is temperature, t is time, and α is the thermal diffusivity. This equation governs the temperature field for a time-dependent heat input, and the decay of temperature inhomogeneities in the absence of heat input. General methods of solving Eq. (4) and special cases are given by Carslaw and Jaeger (1959). The diffusivity is found by measuring the temperature response at times well defined compared to L^2/α, where L is the distance from a heat source to the temperature sensor.

In the Ångström method, with sinusoidal heat input at one point and a sensor at another, one measures the phase difference. The characteristic time is usually of the order of 10 s, and the heating frequency is well below 1 Hz for bar-shaped specimens. For low-conductivity materials one can use long cylinders with a central heater wire, either pulsed or sinusoidally heated. If only response times and distances are measured, the diffusivity is obtained. If heat input and temperature changes are measured as well, the thermal conductivity and the specific heat are also measured.

In the flash diffusivity method, heat is deposited on one side of a specimen disc. The temperature rise on the other side is monitored as function of time. Heat may be generated by a current pulse in a thin-film resistor or deposited by a pulse of radiative energy. An absorbing layer is used for translucent specimens. An accurate measure of specimen thickness is needed, with corrections for absorption depth or for the heat conversion layer. Corrections are required for sideways conduction. On the back face, the sensor can be a contact thermometer, such as a thermocouple, or may be an infrared detector. Modern techniques of timing and rapid data acquisition have made this method feasible, and it is now widely used.

Special techniques have been developed for fibers and thin films. Oriented bundles of fibers have been used, while for single fibers one can use the deflection of an optical beam by the heated air at a point along the fiber to measure its thermal diffusivity.

For low-conductivity materials and thin films, a powerful method has been developed recently (Cahill, 1990). A thin metal strip deposited on the surface is heated by an alternating current, and its temperature is measured. The amplitude of the third harmonic of temperature, studied as function of ac frequency, yields the thermal conductivity.

Thermal-conductivity measurements have been discussed by Parrott and Stuckes (1975), by several contributors in Tye (1969), and by Ho in Touloukian *et al.* (1970). Transient techniques are treated by Kubičar (1990).

3. LATTICE WAVES

In a regular lattice of atoms linked to each other by harmonic forces, the normal modes of vibration are progressive displacement waves. The displacement $\mathbf{u}$ of the atom at site $\mathbf{x}$ at time t is, for each such wave, of the form

$$\mathbf{u}(\mathbf{x},t) = A(\mathbf{q},j)\boldsymbol{\epsilon}_j \exp(i\mathbf{q}\cdot\mathbf{x} - i2\pi ft), \qquad (5)$$

where $\mathbf{q}$ is the wave vector, f the frequency, $\boldsymbol{\epsilon}_j$ a unit vector in the direction of displacement or polarization, and A the amplitude. Each normal mode or lattice wave is specified by $\mathbf{q}$ and by the polarization, denoted by the index j. In an isotropic medium, the three polarization directions $\boldsymbol{\epsilon}$ are longitudinal (one) and transverse (two). In a crystal, these three directions are still mutually perpendicular but not simply related to the direction of $\mathbf{q}$ except for propagation in a principal direction. At low frequencies or long wavelengths these waves are the same as elastic waves in the corresponding continuum, where for any direction, frequency and wave vector are proportional to each other. Thus the phase velocity $2\pi f/q$ depends on direction of $\mathbf{q}$ but not its magnitude.

At high frequencies, when wavelengths are comparable to interatomic distances, f does not vary linearly with $\mathbf{q}$, and the three polarization directions, still mutually perpendicular, depend also on the magnitude of $\mathbf{q}$. The periodicity of the lattice limits the number of physically distinct values of $\mathbf{q}$; this number equals the number of unit cells in the crystal, while the number of polarizations for each $\mathbf{q}$ is three times the number of atoms per unit cell, so that there are three normal modes for each atom. There are three polarization branches for which $f \to 0$ as $q \to 0$; these are acoustic branches. For the others, the optical branches, there are no corresponding waves in the elastic continuum.

The vibrational energy of each mode is

$$E(\mathbf{q},j) = (2\pi f)^2 M|A(\mathbf{q},j)|^2, \qquad (6)$$

where M is the average atomic mass. The thermal energy of vibration is the sum of E over all normal modes.

According to quantum mechanics, the energy of each normal mode is not continuously variable but consists of an integral number N of quanta, each of energy hf, where h is the Planck constant. There is, in each mode, also an irreducible zero-point energy $hf/2$, which is part of the energy of formation of the crystal, and not regarded as thermal energy. The quanta of energy, hf, are the phonons; there are $N(\mathbf{q})$ phonons in each mode $\mathbf{q}$ (the polarization index j will now be implied).

Because real lattice forces have anharmonic (nonlinear) components, and because real crystals are not structurally perfect and also have external boundaries, the lattice waves are not true normal modes, so that energy is exchanged between them continuously. This interaction establishes thermal equilibrium. The average number of phonons and the average energy content of a mode of frequency f at temperature T in thermal equilibrium is

$$N^0 = [\exp(hf/kT) - 1]^{-1}, E^0 = hfN^0, \qquad (7)$$

where k is the Boltzmann constant. In the classical limit (when $T > hf/k$), E^0 approaches kT.

For thermodynamic properties, it suffices that interactions exist and establish equilibrium. To understand transport properties such as thermal conductivity one must also know the strength of these interactions.

The specific heat (per unit volume) due to lattice waves is

$$C = \int C(f)df = \int g(f)hf\frac{dN^0(f)}{dT}df, \qquad (8)$$

where $g(f)$ is the number of modes per unit volume in the frequency interval df. The lattice vibrations can be regarded as a gas of phonons, each carrying energy hf with a velocity equal to the group velocity of the wave, $\mathbf{v} = 2\pi\partial f/\partial \mathbf{q}$.

4. THERMAL CONDUCTIVITY OF INSULATORS

Except at very high temperatures, where electronic conduction or radiation may become appreciable, the thermal conductivity is due to phonons. Equation (2) can be generalized to a sum over polarizations and an integration over frequencies, so that

$$\lambda_p = \tfrac{1}{3}\sum_j \int C_j(f)v_j(f)l_j(f)df. \qquad (9)$$

Here $C_j(f)$ is the spectral contribution to the specific heat per unit volume due to the jth polarization branch, v the magnitude of the group velocity, and l the mean free path, a function of frequency and polarization. The factor $\frac{1}{3}$ applies to isotropic or cubic solids. The equation must be modified for highly anisotropic solids. Note that $l = v\tau$, where τ is a relaxation time, which describes the return to equilibrium of a perturbed distribution. The integrand thus contains the group velocity as a factor v^2; even modes of negative group velocity make a positive contribution to the conductivity. The major contribution to λ_p comes usually from the three acoustic-mode branches. Optical modes contribute significantly when two sublattices have nearly equal masses, so that the optical branch is almost a continuation of the acoustic one.

The mean free path $l(f)$ depends on the temperature and crystal perfection. It is limited by the following processes:

1. interaction between lattice modes, or phonon–phonon interactions, due to departures of the interatomic potential energies from a quadratic (harmonic) dependence on relative displacements;
2. scattering of phonons by lattice imperfections (e.g., solute atoms, other point defects, dislocations, extended defects, and inclusions), similar to scattering of elastic waves by irregularities; and
3. scattering of phonons by external boundaries, or by grain boundaries in polycrystalline aggregates.

4.1 Intrinsic Thermal Conductivity

This is governed by the anharmonic or nonlinear components of the interatomic force constants, that is, components of the potential energy of cubic or higher power in the relative displacements or strains. The cubic terms cause energy interchange between groups of three modes at a time. In the quantum mechanical description, two modes lose one phonon each, and one mode gains one phonon, or vice versa. For three interacting modes, of wave vectors $\mathbf{q}$ and $\mathbf{q}'$, transferring energy into mode $\mathbf{q}''$, there is a condition of energy or frequency conservation

$$f + f' = f''; \qquad (10)$$

and since the anharmonic linkages form a regular lattice, there is a wave-vector selection condition

$$\mathbf{q} + \mathbf{q}' = \mathbf{q}'' + \mathbf{b}. \qquad (11)$$

Here $\mathbf{b}$ is a reciprocal lattice vector of the crystal lattice. The case $\mathbf{b} = 0$ is a normal or N-process. When $\mathbf{b} \neq 0$, one speaks of a U-process (Umklapp or flip-over process), a combined three-phonon interaction and Bragg reflection. U-processes cannot occur in an elastic continuum. N-processes cannot, by themselves, restore equilibrium if the phonon distribution is biased in direction, as it is in the presence of a temperature gradient. However, they can enhance the relaxation rate in combination with other processes, by transferring energy from modes that are weakly scattered to those that experience stronger scattering.

At temperatures at or above the Debye temperature, i.e., in the equipartition regime when $E^0 = kT$ for all modes, the intrinsic mean free path becomes

$$l_i(f,T) = (16\pi)^{-1}v(\mu a^3/kT)f_m/f^2, \qquad (12)$$

where f_m is the highest frequency of the acoustic branch, a^3 the volume per atom, and μ the shear modulus. The numerical constant is somewhat uncertain and depends on the approximations made. If one uses a Debye-type approximation for $C(f)$, i.e., $C(f) \propto f^2$, and integrates Eq. (9) from 0 to f_m, the intrinsic thermal conductivity becomes

$$\lambda_i = (3/16\pi)\mu v^2 T^{-1} f_{\mathrm{D}}^{-1} N^{-2/3}, \qquad (13)$$

where f_{D} is the Debye frequency and N the number of atoms per unit cell. This expression is similar to the one first derived by Leibfried and Schlömann, which has been compared to the conductivity of many dielectric crystals by Slack (1979).

It is interesting that, since $C(f) \propto f^2$ and $l_i \propto f^{-2}$, the integrand in Eq. (9) containing $C(f)l_i(f)$ is independent of f, so that equal frequency intervals make equal contributions to the intrinsic conductivity. This contrasts with the vibrational energy content, which is dominated by the highest frequencies.

At low temperatures heat is carried mainly by low-frequency waves ($f \sim kT/h$). Intrinsic thermal resistivity is then due to U-processes, which require the prior presence of phonons near the zone boundary. As a result, the intrinsic mean free path has the form

$$1/l_i \propto f \exp(-T_c/T), \qquad (14)$$

where kT_c/h is comparable to the frequency of zone boundary modes, but may change with temperature and defect content. Typically T_c is around half the Debye temperature. The intrinsic thermal conductivity increases rapidly with decreasing temperature, until at very low temperatures the mean free path is limited by boundary scattering. Near that transition, thermal conductivities go through a pronounced maximum.

4.2 Boundary Scattering

Both external boundaries and grain boundaries limit the phonon mean free path to a value L, which is of the order of the smallest linear dimension of the crystal, or, in the case of polycrystalline aggregates, of the order of the average grain diameter. However, L is not quite independent of frequency but increases slowly with decreasing frequency. In the case of external boundaries, the diffuse component of scattering depends on roughness and increases with frequency. Thus, below the maximum, the thermal conductivity increases with crystal size, and $\lambda(T)$ increases with T, but not as rapidly as the specific heat.

Polycrystalline solids have a conductivity at low temperatures much lower than single crystals. Grain-boundary scattering is probably due to thin intergranular regions, rather than due to misorientation of grains. The thermal conductivity increases with temperature, but significantly more slowly than the specific heat. At ordinary and high temperatures, the effects of grain-boundary scattering persist, affecting phonons of lowest frequencies. For micron-sized grains the conductivity reduction is of the order of 1% to 3%, which is comparable to the precision of measurements.

4.3 Defects

Point defects (solute atoms, isotopic mixtures, vacancies, and interstitials) scatter as f^4 and affect mainly the high-frequency modes. Extended defects scatter with weaker frequency dependence. Sheetlike defects (inclusion plates and stacking faults) scatter as f^2. Sessile dislocations scatter as f, in virtue of their long-range strain field. Mobile dislocations show resonance scattering as well. Thus point defects are important at ordinary temperatures but may also be significant at temperatures near the conductivity maximum at lower temperatures, while extended defects are important at low temperatures. Three-phonon N-processes often enhance the effect of point defects; they conserve phonon momentum $h\mathbf{q}/2\pi$, but transfer excess momentum of the phonon gas from low frequencies, where point-defect scattering is weak, to modes of high frequencies, where scattering is strong. The role of N-processes has been well discussed by Berman (1976).

Point-defect scattering cross sections can be estimated if the nature of the defects is known, and with confidence for substitutional atoms differing from normal atoms in regard to mass only. In some crystals the naturally occurring isotopic mass fluctuations reduce the low-temperature thermal conductivity below its intrinsic value, the value it would have if limited only by anharmonic processes. Measurements on crystals with artificially controlled isotopic compositions have been made on germanium, lithium fluoride, diamond, and solid helium.

At ordinary and high temperatures, solute atoms of large mass difference and vacancies cause observable decreases in thermal conductivity if their concentration exceeds a few percent. The conductivity of mixed crystals has a minimum around 50% concentrations. At low concentrations c, if point-defect scattering is strong, $\lambda \propto (Tc)^{-1/2}$. Point defects reduce the contributions to the conductivity by high-frequency phonons. Additional reductions are caused by extended defects and grain boundaries, as well as by conduction electrons. These reduce the contribution of the low frequencies. The thermal conductivity of Si-Ge alloys has been extensively studied and conforms to this picture.

5. ELECTRONIC THERMAL CONDUCTIVITY

This is the dominant component in pure metals and dilute alloys, the most important component in alloys, and significant in semimetals and in semiconductors with large electron or hole concentration. In principle it is given by Eq. (2), so that

$$\lambda_e = \tfrac{1}{3} C_e v_e l_e, \tag{15}$$

where l_e is the electron mean free path and v_e an electron velocity. However, the electronic contribution to the specific heat is usually only a small part of the total, so that Eq. (15) is not useful. Also the mean free path l_e occurs as a factor in the electrical conductivity, and it is more useful to relate λ_e to the electrical conductivity σ. The Wiedemann–Franz law, extended by Lorenz, expresses λ_e as

$$\lambda_e = L\sigma T, \tag{16}$$

where T is the absolute temperature and L, the Lorenz ratio, is $L = n(k/e)^2$, where k is the Boltzmann constant, e the electronic charge, and n a numerical parameter whose value, given by theory, is discussed below. It usually ranges from 2 to 4, but can exceed that value substantially in intrinsic semiconductors. In pure metals at low temperatures, n can be smaller.

The electric current caused by a field is made up of contributions from electrons of different energies E, so that the electrical conductivity can be expressed as

$$\sigma = \int \left(-\frac{df^0}{dE}\right) s(E)dE, \tag{17}$$

where $s(E)$ is the product of the density of electron states and the mobility. The latter contains the electron mean free path $l_e(E)$. The other factor in the integrand is the derivative of the Fermi–Dirac electron distribution function in equilibrium,

$$f^0(E,\zeta,T) = \{\exp[(E - \zeta)/kT] + 1\}^{-1}, \tag{18}$$

where ζ, the Fermi energy, depends on temperature and electron density.

A temperature gradient perturbs the electron distribution, making electrons flowing down the temperature gradient slightly warmer. This produces both an electric and a heat current. The definition of thermal conductivity requires zero electric current; hence there is a thermoelectric field to oppose that current. Defining moments of the integrand of Eq. (17) by

$$M_m = \int - \left(\frac{df^0}{dE}\right) s(E) \left[\frac{E - \zeta}{kT}\right]^m dE, \tag{19}$$

one can show the Seebeck coefficient to be $S = (k/e)M_1/M_0$, and the Lorenz ratio to be

$$L = n(k/e)^2 = (k/e)^2[M_2/M_0 - (M_1/M_0)^2]. \tag{20}$$

This expression applies generally, from cases of high electron densities in metals to low densities in semiconductors, including when both electron and hole bands contribute (ambipolar diffusion). The Lorenz ratio is proportional to the second moment of the partial electrical conductivity about the mean energy of conduction (Klemens, 1969). However, this result assumes that the electron mean free path is the same for thermal as for electrical conduction.

6. METALS

In metals the electron density is high and the electron gas is highly degenerate—that is, at the Fermi energy the density of free-electron states is high—and both the density of states and the partial conductivity $s(E)$ in Eq. (19) are almost constant over an energy interval of order kT above and below ζ, which is the important energy range for electrical and thermal conduction. If $s(E)$ is constant, the ratio M_2/M_0 is determined by the width of the bell-shaped function $-df^0/dE$ in units of kT, and $M_2/M_0 = -\pi^2/3$, while $(M_1/M_0)^2$ is negligible (or S is small compared to $k/e = 86\ \mu\text{V/deg}$). This leads to the Sommerfeld (1933) value of L:

$$L_0 = (\pi^2/3)(k/e)^2 = 2.45 \times 10^{-8}\ \text{V}^2/\text{deg}^2. \tag{21}$$

Any asymmetry of $s(E)$ about ζ causes a small Seebeck coefficient S; L is below L_0 by S^2. Curvature of $s(E)$ causes a further deviation from L_0. This may be of either sign, but positive deviations are more common at high temperatures (White, 1991). The reason may be a tendency, for reasons of crystal stability, for ζ to lie preferably at a minimum in the density of states and avoid a maximum. The higher the temperature, the larger the deviations of L from L_0. Deviations up to about 15% occur above 1000 K in transition metals, where the subbands of the d electrons give structure to the density of states of the required scale. See Klemens and Williams (1986).

The thermal conductivity of metals can be understood in terms of the electrical conduction. The mean free path of electrons is limited by scattering processes, both by thermal vibrations or phonons, and also by impurities, solutes, and other lattice defects. The corresponding electrical resistivities are the temperature-dependent intrinsic resistivity $\rho_i(T)$, which tends to zero at very low temperatures, and the residual resistivity ρ_0, independent of T but depending on defect content or solutes. According to Matthiessen's rule the resistivity is additively composed of these two components, but there are deviations $\Delta\rho$, so that

$$\rho = \rho_0 + \rho_i(T) + \Delta\rho, \tag{22}$$

where $\Delta\rho$ vanishes as ρ_0 or T goes to zero.

The electronic thermal resistivity $W_e = 1/\lambda_e$ can be similarly expressed as

$$W_e = W_0 + W_i + \Delta W. \tag{23}$$

Now one expects that ρ_0 and W_0 are related by the Sommerfeld Lorenz ratio, i.e., $W_0 = \rho_0/L_0T$. The relation between ΔW and $\Delta\rho$ is not well known, since when $\Delta\rho$ is large, there is usually an appreciable lattice component of thermal conductivity. However, theoretically one expects ΔW to equal $\Delta\rho/L_0T$, at least approximately. Only at low temperatures in pure metals ($W_i > W_0$) does one find that ΔW is much larger than $\Delta\rho/L_0T$. This is expected from the transport equations.

At ordinary and moderately high temperatures, one also expects that $W_i(T) = \rho_i(T)/L_0T$, so that $\lambda_e = L_0T/\rho$ for metals, except for departures of L from L_0 at very high temperatures because $s(E)$ is not quite constant, as mentioned above. Since for most metals $\rho_i \propto T$, λ_e is only weakly temperature dependent.

However, at low temperatures $W_i(T)$ does not decrease as rapidly as $\rho_i(T)/T$ with decreasing T, and the Lorenz ratio ρ_i/TW_i is not constant for pure metals but is proportional to T^2, until a temperature is reached where W_0 is comparable to W_i. This follows from the formal solution of the Boltzmann equation (Wilson, 1952). It can also be physically explained in terms of a shorter electron mean free path for thermal conduction than for electrical conduction; see, e.g., Klemens (1956). Electrons are brought into equilibrium by electron–phonon interactions. In each interaction the energy of an electron is altered by an amount of order kT, and the direction by an angle of order T/θ, where θ is the Debye temperature. At low temperatures it takes many such interactions (of order θ^2/T^2) to reverse the direction of an electron, but only one interaction to change its energy significantly, so that the mean free path for electrical conduction exceeds that for thermal conduction by that same factor. In metals of simple band structure $\rho_i \propto T^5$ while $W_i \propto T^2$.

For a more detailed discussion of the Lorenz ratio see Klemens and Williams (1986), and for low-temperature conduction properties, Klemens (1956).

At ordinary and high temperatures the electrical and thermal conductivities of metals are not very sensitive to solute impurities. A rough but general rule states that it takes around 1 at.% of solutes to double the electrical and thermal resistivities at room temperature, with correspondingly smaller changes at higher temperatures. Impurities that are not dissolved in the matrix have little effect. At low temperatures the effect of solutes is more pronounced.

Figure 1 shows the thermal conductivity of pure copper. From room temperatures up, it is practically constant. At low temperatures it increases, consistent with $W_i \propto T^2$. Finally, depending on specimen perfection, a temperature is reached where W_0 and W_i are equal and the conductivity is a maximum. At still lower temperature, $\lambda = L_0T/\rho_0$ and varies from sample to sample. In copper and in other metals of high electrical conductivity, the lattice component is negligible.

7. ALLOYS

There is a gradual transition in the thermal conductivity from pure metals, with impurity content up to 0.1 at.%, to alloys with solute content exceeding 1%. The term W_0 in Eq. (23), which dominates only at low temperatures in metals, now becomes comparable with or larger than W_i at room temperatures, and with increasing solute content at even higher temperatures. The electronic component λ_e increases linearly with T and continues to increase, even though more slowly, at highest temperatures. Since λ_e is now smaller, the lattice component λ_p is no longer insignificant, particularly at low temperatures.

The lattice thermal conductivity at ordinary and high temperatures of metals and dilute alloys is similar to that of dielectric solids of comparable elastic properties, limited also by anharmonic interactions, so that $\lambda_p \propto T^{-1}$. However, in concentrated alloys it is reduced by the solute atoms, which scatter the high-frequency phonons, especially by solute atoms of substantial mass difference. This also reduces the temperature dependence until it approaches $\lambda_p \propto T^{-1/2}$, just as for mixed dielectric crystals. Solutes that dif-

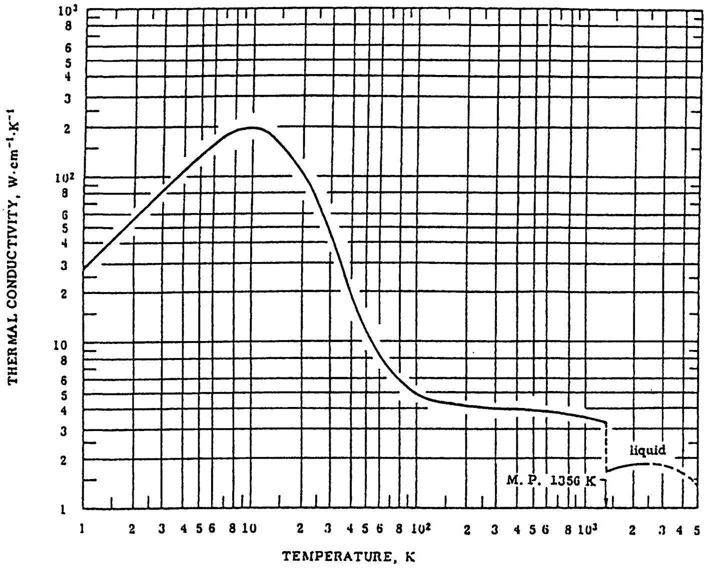

FIG. 1. Thermal conductivity of copper, "recommended" values derived from several reliable measurements for a hypothetical specimen of residual resistivity $\rho_0 = 1 \times 10^{-3}\ \mu\Omega\cdot\text{cm}$ by Touloukian *et al.* (1970).

fer substantially in size from the host have limited solubility and thus a smaller effect.

The phonon mean free path at low frequencies is limited by interactions with electrons, so that the lattice component is decreased. At low temperatures λ_p varies with T between T and T^2 and attains a maximum around 30 K. In some transition metals this maximum can be very broad. Thus λ_p is considerably lower than in dielectric solids at low temperatures and does not depend on external specimen size or grain size in polycrystalline specimens.

The thermal conductivity is thus composed of two components: λ_e varying as T at low temperatures and gradually less rapidly at higher temperatures, and λ_p, which has a maximum somewhere between 20 and 40 K. Therefore λ has a hump near that temperature. Figure 2 shows the thermal conductivity of a Cu-Zn alloy after plastic deformation and after various stages of annealing. Since $\lambda_e = L_0T/\rho$, it can be calculated from the measured electrical resistivity, and λ_p can be deduced. The lattice conductivity is reduced by plastic deformation, because phonons are scattered by dislocations. Dislocations are removed by annealing, so that λ_p increases and the "hump" becomes more pronounced. The low-temperature thermal conductivity of alloys can be more sensitive to cold work than the electrical resistivity.

While copper, silver, and gold alloys can have significant lattice components even close to room temperature, this component is smaller for aluminum alloys and negligible near room temperatures and above. For aluminum alloys the simple Wiedemann–Franz law $\lambda = L_0T/\rho(T)$ can be used to estimate the thermal conductivity, except below 70 K. Transition-metal alloys, where λ_e is smaller, sometimes have significant lattice components even at room temperatures. At low temperatures λ_p is smaller because the interaction of phonons with d-band electrons is stronger than in the simple metals (Klemens and Williams, 1986).

So far only single-phase alloys have been discussed. If two or more phases are present, the overall conductivity is an average of that of the phases, as discussed in Sec. 9. Similar considerations also govern their electrical conductivity. If all the phases are metallic, one can assume, as a first approximation, that the overall thermal and electrical conductivities are related by the Wiedemann–

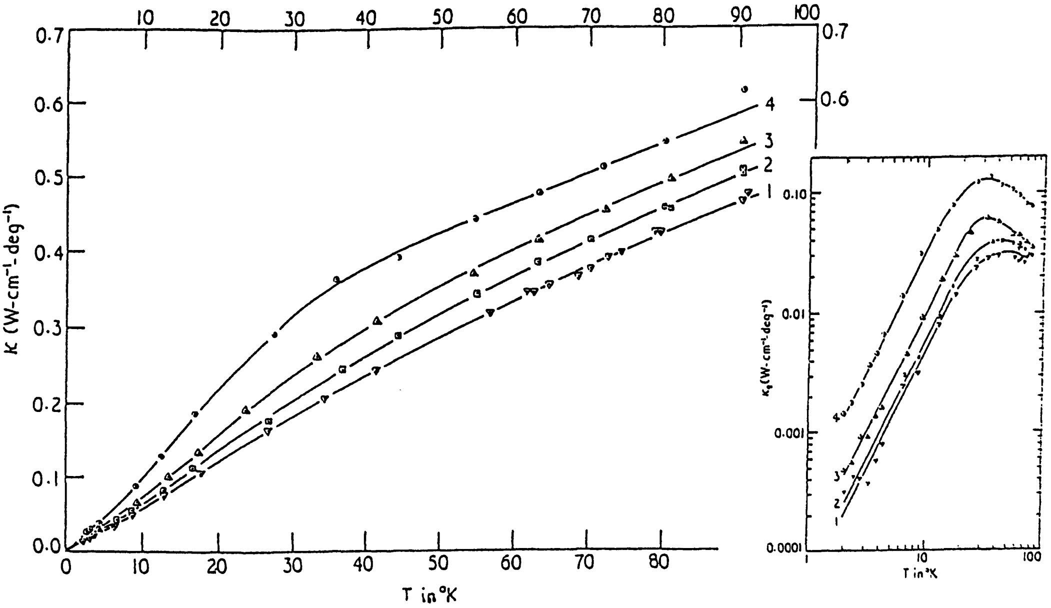

FIG. 2. Thermal conductivity of Cu-30% Zn after severe plastic deformation (curve 1), and subsequent annealings to 250 °C (curve 2), 290 °C (curve 3), and to 450 °C (curve 4). Insert shows calculated lattice components. From Kemp *et al.* (1959).

Franz law. However, if one phase is metallic and the other electrically insulating, this is no longer the case.

8. SEMIMETALS AND SEMICONDUCTORS

In semimetals the density of free electrons is less than in metals, and the electron gas is not highly degenerate at ordinary temperatures: the density of states and $s(E)$ vary over energies of order kT near ζ. Furthermore, λ_e is lower, so that λ_p is more important and, in some cases, the dominant component. At the Fermi energy the density of electron states is finite, so that there are free electrons and a finite electrical conductivity even at very low temperatures.

In semiconductors there are two bands separated by an energy gap, with the Fermi energy lying within the gap. The electron and the hole bands each contain a small number of carriers with Maxwell–Boltzmann distributions. The concentration of carriers is readily changed by donor or acceptor solute atoms. However, when these donors or acceptors provide a sufficiently large carrier concentration, either electrons or holes, the carrier concentration becomes somewhat degenerate, λ_e becomes appreciable, and the semiconductor approaches the behavior of a semimetal.

The lattice component of thermal conductivity is, in both cases, similar to that of a dielectric crystal of comparable elastic properties and size, except that phonons can also be scattered by electrons or holes. However, unlike the case in metals, only phonons of low frequency can interact with the carriers, because the partially occupied electron or hole states occupy only limited regions in wave-vector space. In semimetals this limits the wave vectors of the interacting phonons to the diameter of the Fermi surfaces, with a corresponding upper bound to their frequencies. In semiconductors this upper bound depends on temperature, and varies as $T^{1/2}$ for electron–phonon interactions within a valley. Some semiconductors contain more than one valley. Intervalley scattering reduces the mean free path of phonons in narrow frequency bands at high frequencies.

In arsenic and antimony the thermal conductivity is mainly electronic, but L departs from L_0 because $s(E)$ is not constant about ζ. The lattice conductivity shows up as a hump around 20 K in As and around 10 K in Sb. The separation into λ_e and λ_p has been made by magnetothermal conductivity measurements. Bismuth, on the other hand, has a relatively small electronic component. The thermal conductivity is essentially due to λ_p except below about 3 K, where $\lambda_e \propto T$ becomes larger than $\lambda_p \propto T^3$. At higher temperatures bismuth has the thermal conductivity of a dielectric crystal.

Graphite, with an electron and a hole band that overlap slightly, has a dominant electronic component at very low temperatures with $L = L_0$, and again at very high temperatures with $L \approx 2L_0$. Except at these extreme temperature ranges, the lattice component dominates. It also differs from other solids because of its layer structure that makes both the electrical conductivity and the thermal conductivity, including the lattice component, very anisotropic. The phonon gas has a two-dimensional character at all but low temperatures. This enhances contributions from low frequencies, and thus increases the sensitivity of the lattice thermal conductivity to extended defects (see Sec. 11.4).

The thermal conductivity of semimetals has been reviewed by Uher (1991).

In semiconductors of large band gap, such as germanium and silicon, the intrinsic electrical conductivity is small and λ_e is not important, except at high temperatures in Si. Donors and acceptors increase the electrical conductivity, so that λ_e becomes comparable to λ_p at high temperatures and carrier concentrations. The Lorenz ratio $\lambda_e/\sigma T$ depends somewhat on the energy dependence of the electron mean free path, and for single-band conduction it is related to the thermoelectric power.

In the intrinsic regime, L is increased by ambipolar diffusion. Both the electron and hole bands contribute to the conduction. This enhances M_2/M_0 in Eq. (20). Thus L increases with increasing band gap, but σ decreases, so that the ambipolar enhancement of λ_e is most important for moderately large band gaps, such as in InSb. Graphite, as mentioned above, can be regarded as an intrinsic semiconductor with zero band gap, and $L \simeq 2L_0$, though λ_e is only appreciable at very high temperatures.

The thermal conductivity of semiconductors has been reviewed by Steigmeier (1969).

9. SUPERCONDUCTORS

In the normal state, either above the transition temperature or in a sufficiently high magnetic field, the thermal conductivity is like that of a metal or alloy, with electronic and lattice components, and λ_p is usually small in pure metals. In the superconducting state, the thermal conductivity changes and may be either smaller or larger than in the normal state. This can be understood as follows: The electron gas can be divided into Cooper pairs, responsible for superconductivity, and normal electrons, reduced in number. Only the latter are disordered. Therefore:

1. The electronic component λ_e is smaller in the superconducting state than in the normal one, particularly well below the transition temperature, because Cooper pairs carry no entropy.
2. The lattice component λ_p is often increased. While the scattering of phonons by imperfections does not change, the interactions of phonons with electrons is decreased, since Cooper pairs do not scatter phonons. Those lattice waves whose phonon mean free path is normally limited significantly by the phonon–electron interactions have, in the superconducting state, a larger mean free path. At sufficiently low temperatures it is limited only by grain boundaries or by other imperfections.
3. In a mixed state, where flux is trapped in the form of flux tubes or vortices within the material and forms normal regions, the phonon mean free path within the superconducting regions can be limited by the interaction with these vortices. This additional scattering mechanism reduces λ_p in the superconducting regions. The overall thermal conductivity is that of an inhomogeneous solid composed of regions of two different conductivities. Inhomogeneous solids will be discussed in Sec. 10.

In conventional pure-metal superconductors below the transition temperature, the thermal conductivity is therefore lower than in the normal state, where the conductivity is mainly electronic. Well below the transition temperature, λ_e decreases exponentially with decreasing T, because the electronic specific heat C_e behaves likewise. At very low temperatures λ_e is so small that λ_p dominates and varies as T^3, since the mean free path is boundary limited. However, λ_p can be reduced by plastic deformation, because of dislocation scattering; the material thus behaves like an insulator, but with the plastic properties and dislocation density of a metal.

In alloys, depending on the ratio of λ_e to λ_p in the normal state, the conductivity can be either larger or smaller in the superconducting state. At very low temperatures, λ_p will again dominate in the superconducting state. Scattering of phonons by solute atoms is then rarely important.

In the high–transition-temperature superconductors, often called cuprate superconductors, the change in thermal conductivity is usually smaller. The lattice component is a significant fraction of the total in the normal state, but since the transition occurs at higher temperature, the role of electrons in limiting λ_p is smaller; thus the removal of the phonon–electron interaction has a smaller effect on λ_p. Also, the lattice conductivity is often limited by imperfections. There is, however, a possibility that λ_e increases in the superconducting state. Even though C_e decreases, it is possible that the mean free path of the electronic quasiparticles increases. This may give rise to an increase in the thermal conductivity just below the transition temperature.

Thermal conductivities of superconductors were reviewed by Klemens (1956) and Berman (1976), and that of high-T_c superconductors, by Uher (1992).

10. INHOMOGENEOUS CONDUCTORS

Many technically important solids are composites of different materials, each with a different thermal conductivity. Solids containing nonconducting inclusions or pores form a special case, as do solids with highly conducting inclusions in a low-conductivity matrix. The problem of estimating the overall thermal conductivity is mathematically the same as that of calculating the electrical conductivity and the dielectric constant of composites. In addition to composites with sharp boundaries between components, one can also have continuous variations with position of the local conductivity. These problems have not been solved in general, but there are various approximations, applicable to different cases. The literature of these mathematical treatments is widely scattered according to different applications. A useful review is by Landauer (1978).

A related problem is that of averaging elastic moduli of composites, and also thermal expansivity, though in these problems one deals with more parameters. Both of these problems can be formulated in terms of minimizing strain energy, while the conduction problem can be expressed in terms of minimizing the rate of entropy production.

An upper bound to the effective conductivity is provided by the volume average of the conductivities, while an upper bound to the effective resistivity, or a lower bound to the conductivity, is given by the volume average of the resistivities. If there is not too much contrast between local conductivities, these two bounds can be very useful. In other cases one of these bounds may be useful, while the other bound is far removed and thus useless.

A perturbation method, first given by Herring and subsequently extended (Klemens, 1991), expresses the space variations of the conductivity, or those of resistivity, in terms of Fourier components. The effective conductivity λ_{eff} can be expressed as

$$\lambda_{\mathrm{eff}} = \lambda_0 - \langle\cos^2\theta\rangle\langle\Delta\lambda^2\rangle/\lambda_0, \tag{24}$$

where λ_0 is the volume average of the conductivity, $\langle\Delta\lambda^2\rangle$ the mean square departure from that average, and $\langle\cos^2\theta\rangle$ the average square of the direction cosine of the Fourier components with respect to the direction of the temperature gradient. In the absence of a preferred direction, $\langle\cos^2\theta\rangle = \frac{1}{3}$. Similarly the effective resistivity W_{eff} can be expressed in terms of the volume average W_0, the mean square fluctuations $\langle\Delta W^2\rangle$, and the mean square of the sine of the Fourier components with respect to the heat flow direction, by means of

$$W_{\mathrm{eff}} = W_0 - \langle\sin^2\theta\rangle\langle\Delta W^2\rangle/W_0. \tag{25}$$

Without preferred direction, $\langle\sin^2\theta\rangle = \frac{2}{3}$. Depending on which fluctuation is smaller, $\langle\Delta\lambda^2\rangle/\lambda_0^2$ or $\langle\Delta W^2\rangle/W_0^2$, one or the other of the two approximations is preferable. For large contrasts and in the middle of the concentration range, perturbation methods break down.

A different approach is that of the mean-field approximation. To obtain the effective conductivity, the field pattern is calculated around and in an inclusion in a matrix of different conductivity, but the conductivity of the matrix is not that of the matrix material itself, but that of the matrix that contains inclusions—that is, the effective conductivity. This is a good approximation, provided inclusions do not interact. An extension of this approach is to consider three regions: the inclusion, the pure matrix surrounding the inclusion, and the average material. This and other approximations are discussed by Parrott and Stuckes (1975) and also by Grimvall (1986).

While there are many theories that are used to interpret conductivity measurements, there are only a few controlled experiments to test them. No approximation appears to hold over the entire composition range for large conductivity differences of the components. Percolation theories may be applicable in such cases.

11. SPECIAL TOPICS

11.1 Heat-Conducting Substrates

Solid-state electronic components are mounted on insulating substrates, which in turn are mounted on a metallic base. Increasing miniaturization, needed for fast processing and compactness, makes special demands on the ability of the substrate to remove heat by conduction. To achieve a high thermal conductivity, the substrate should be a crystalline dielectric solid of simple atomic structure, of low atomic weight, and free of defects. It need not be a single crystal: at ordinary temperatures the conductivity of a fully densified polycrystalline solid is usually within a few percent of the corresponding single crystal.

The advantage of low atomic weight and simple structure (small unit cell) follows from Eq. (13): one notes that elastic moduli do not vary as much as does the mean atomic mass, which enters into the factor v^2/f_D. The highest conductivity is that of diamond, which can now also be grown as reasonably defect-free thin films. Aluminum nitride, silicon carbide, and beryllia also have good thermal conductivities, and these may be even higher than observed so far, if they can be made free of defects and large grained.

Of course, high thermal conductivity is not the only requirement. Good heat transfer at the interface between the semiconductor circuit element and the substrate is also important, so that thermal expansion differences and acoustic impedance mismatch should be minimized.

11.2 Amorphous Solids

Thermal conduction in amorphous solids is due to phonons, except that in glass at high temperature there is also a contribution by internal radiative heat transfer. The average phonon mean free path $\langle l \rangle = 3\lambda/Cv$ [see Eq. (2)] is quite short, only of the order of a few ångströms. The high-frequency modes, which make the major contribution to the energy content, are thus not progressive waves but vibrations localized at atomic sites, with mechanical energy transfer between sites. Alternatively one can regard these modes as waves with very short mean free path, comparable to their wavelength. This mean free path is referred to as minimum mean free path: increasing the perturbation or the strength of the interaction cannot reduce it further. However, modes of lower frequency are wavelike, since for longer waves the solid approaches an elastic continuum. These waves contribute to the thermal conductivity even at ordinary temperatures. At low temperatures they make the major contribution to the thermal conductivity, so that the average mean free path increases with decreasing temperature.

At low temperatures one can understand the conductivity in terms of Eq. (9), using a Debye model for simplicity. Thus $C(f) \propto f^2$ for f of order kT/h. The disorder of the solid can be described by strong point-defect scattering. However, the mean free path for point defects l_p varies as f^{-4}, and an additional scattering process is required to prevent the integrand $C(f)l(f)df$ from diverging at low frequencies. If this process has a mean free path $l'(f) \propto f^{-n}$, then at very low temperatures $\lambda \propto T^{3-n}$, provided $n < 3$. Combining the scattering rates, so that $1/l = 1/l' + 1/l_p$, leads to a thermal-conductivity curve that increases at low temperatures with increasing T but reaches a constant value or "plateau" at higher temperatures. If now one adjusts this model so that $l(f)$ cannot decrease below a minimum value l_0, the plateau does not extend to high temperatures, but $\lambda(T)$ increases again, still in the low-temperature regime. Eventually, with increasing temperature, it approaches $\frac{1}{3}C(T)vl_0$, where $C(T)$ is the Debye specific heat per unit volume.

The low-frequency scattering mechanism that dominates at low temperatures makes $\lambda(T)$ vary as T^2, so that $l'(f) \propto 1/f$. This can be explained by a model in which an assembly of localized defects can be in two atomic configurations. In silica glasses these could be two configurations of the Si-O-Si chain. Phonon-assisted tunneling allows transitions between these configurations. The energy spacing between them defines the frequency of the phonon that is either absorbed or emitted. Because of the general disorder each defect has a different level spacing or frequency, and these frequencies are widely distributed. A Gaussian distribution much wider than kT/h would then lead to the observed frequency dependence $l' \propto 1/f$.

This model is well supported by other properties, including an excess specific heat at very low temperatures (Phillips, 1981). The nature of these two-level defect sites is not yet fully understood, nor the factors determining their concentration and frequency distribution.

11.3 Thermal-Barrier Coatings

Thermal-barrier coatings are used in aircraft and other gas-turbine engines to insulate metal components from the hot gas stream, improving component durability and engine efficiency. In aircraft engines such coatings can reduce metal temperatures by as much as 170 °C. Current research aims to reduce the thermal conductivity of the coating material, thus allowing a larger temperature drop across the coating. The material of current interest is stabilized cubic zirconia, which, in the fully dense form and without cracks, has a thermal conductivity of about 2 $W \cdot m^{-1} \cdot K^{-1}$ over a wide range of elevated temperatures. Chemical stability is, of course, equally important, and the high oxygen diffusivity of zirconia requires barrier layers, but further improvements demand a better understanding of the thermal-conduction mechanism. Thermal conduction is mainly by lattice waves; above 1000 K a radiative contribution cannot be ruled out.

The suitability of zirconia seems to depend on the fact that the cubic form is stabilized by the addition of solutes (Ca, Mg, Y, Ce), though Ca and Mg tend to migrate at high temperatures. Except perhaps for Ce, the solutes introduce vacancies in the oxygen sublattice; these make the major contribution to the scattering of high-frequency phonons. At the very highest frequencies the phonon mean free path reaches a minimum value of the order of the interatomic spacing and cannot be decreased by additional solutes, but at intermediate frequencies, increasing solute content is still effective. At lower frequencies, where point-defect scattering is weak, one must look for other processes.

Processes scattering low-frequency phonons are grain-boundary scattering, scattering at the interfaces of thin layers, and phonon interactions by two-level systems, as in amorphous solids. At low temperatures the thermal conductivity of stabilized zirconia does show similarity to that of vitreous solids, even though zirconia is crystalline. It has been suggested that some oxygen atoms that are next to vacancies can tunnel between different configurations. This limits the phonon mean free path to about 0.5 μm. If the grain size is comparable or smaller, there would be an additional reduction in thermal conductivity. A similar reduction should be obtained if the coating is deposited as a series of thin films of comparable thickness, separated by layers of different phonon velocity. These layers could also serve as oxygen-diffusion barriers. If the layer thickness or the grain size L is small enough, and the solute concentration c large enough, anharmonic interactions would not be important over the entire frequency range. The conductivity, independent of temperature, would then vary as $c^{3/4}L^{1/4}$. The required grain size or film thickness would probably have to be a few tens of nanometers.

11.4 Layered Crystals: Graphite

Some crystals are composed of atomic layers, with strong atomic bonds within each layer but much weaker bonds between atoms in adjacent layers. As a result, the phonon velocity is very anisotropic, much larger for propagation in the layer plane than normal to it. Also the layer spacing may then be larger, reducing the Brillouin zone in the normal direction. Thus energy is propagated mainly in the layer plane for all frequencies above a critical frequency f_c, which is well below the maximum frequency of the acoustic branch. For $f < f_c$, propagation is in three dimensions. In graphite, which is very anisotropic, $f_c = 4$ THz while the in-layer acoustic branches extend to 25 and 37 THz respectively. Thus above f_c the phonon gas is two-dimensional, $C(f) \propto f$, while in three dimensions $C(f) \propto f^2$. Other layered crystals may have a relatively smaller range of two-dimensional behavior. The frequency dependence of the intrinsic mean free path is not affected by the dimensionality and varies as f^{-2}. The spectral thermal conductivity is proportional to $C(f)l(f)$ and varies as $1/f$ above f_c. In contrast to three dimensions, where the intrinsic heat current is due to modes of all frequencies, now the heat is carried mainly by modes near and just above f_c. The thermal conductivity of layered crystals is thus less sensitive to point defects and more sensitive to extended defects than in isotropic crystals. At low temperatures, if grain-boundary scattering dominates, the T dependence of $\lambda(T)$ varies between T^2 and T^3.

The lattice thermal conductivity normal to the layer plane is much smaller than within it. In a polycrystalline aggregate, the thermal conductivity is isotropic if the crystal platelets are randomly oriented but very anisotropic if they lie parallel in a common plane, as in the case of pyrolytic graphite. The overall thermal conductivity is sensitive to the shape of the platelets and their arrangement, and graphite shows a wide range of thermal conductivities.

Other layered materials are not as strongly anisotropic as graphite, but similar considerations may well apply. They have not been investigated as fully as graphite has been.

11.5 Radiation Damage

Fast-particle irradiation causes atomic displacements and lattice imperfections, which reduce the thermal conductivity of insulators. Depending on the nature and energy of irradiating particles, the damage consists of individual displaced atoms and small regions of disorder. The displaced atoms cause interstitials and vacancies. Both of these point defects can move and recombine, but some of them form larger aggregates, leaving behind a residue of individual point defects.

In the case of metals and alloys, the change in the electronic thermal conductivity parallels the change in electrical resistivity through the Lorenz ratio L_0 of Eq. (21). Radiation damage in these materials can be studied more conveniently in terms of changes in electrical resistivity. There have been some studies of changes in the lattice component of thermal conductivity of alloys, but there have been many more studies of changes in the lattice thermal conductivity of insulators and of graphite.

Early studies of the thermal conductivity of irradiated quartz and sapphire at low temperatures showed reductions at temperatures below that of the maximum, which were attributed to phonon scattering by extended imperfections, while reductions above that temperature were interpreted as due to point defects. Extended imperfections have since been seen by electron microscopy in alumina and other dielectrics. The reductions at ordinary and high temperatures in alumina were not inconsistent with low-frequency phonon scattering by such extended imperfections, together with point defects, which are not seen by electron microscopy and which scatter high-frequency phonons. The required concentration of point defects roughly matches the total volume fraction of the extended imperfections.

As the nuclear fuel UO_2 is consumed or "burnt up" in a reactor, various fission products build up as solutes in the matrix, while some that have a larger size mismatch form metallic colloids. Inert-gas bubbles are also formed. The solute ions reduce the thermal conductivity of the matrix, as does increased porosity. Metallic colloids cause a smaller increase; the net effect is a decrease. The thermal conductivity of UO_2 containing solutes that match those produced by fission has been studied in Symfuel (see Glossary) without the difficulties of radioactivity. Nonstoichiometry has also been studied. It appears that oxygen reduction does not cause oxygen vacancies but metal colloids, increasing the conductivity. Excess oxygen, in the form of oxygen interstitials, scatters phonons strongly and reduces the conductivity.

There has been much work on the reduction of thermal conductivity in neutron-irradiated graphite. Probably both point defects and extended defects contribute to the decrease. Extended defects should be relatively more effective than in other solids, because the layer structure enhances the role of the low-frequency modes, as discussed in Sec. 11.4. However, the grain structure also changes, as indicated by density changes, and this makes the interpretation more difficult.

GLOSSARY

Anharmonicity: Terms in the potential energy of a lattice as function of strain or displacement beyond the quadratic or harmonic terms. The cubic anharmonic terms are the most important, being responsible for interactions exchanging energy among triplets of normal modes. They are also responsible for thermal expansion.

Boltzmann Equation: The equation that balances the rate of change of the occupation number of particles in phase space due to forces and concentration gradients against the change due to collisions or other interaction processes. In the present context it applies not to gas atoms but to electrons and to phonons.

Boundary Scattering: Scattering of carriers, either electrons or phonons, by boundaries, which may be the external boundaries of the specimen, or internal boundaries such as grain boundaries or boundaries of a thin film. These processes may limit the carrier mean free path.

Cooper Pairs: A pair of electrons of the same momentum but opposite spin, held together by an interaction force and effectively forming a particle obeying Bose–Einstein statistics and able to form a Bose–Einstein condensate responsible for superconductivity.

Electronic Component: The contribution λ_e to the thermal conductivity from the heat flow carried by free electrons.

Intrinsic Conductivity: The conductivity, electrical or thermal, of a structurally perfect solid, limited only by carrier interaction with the vibrations of the lattice. Intrinsic lattice thermal conductivity refers to the lattice component. Intrinsic resistivity is the reciprocal of that conductivity.

Lattice Component: The contribution λ_p to the thermal conductivity from the heat flow carried by lattice waves or phonons.

Lorenz Ratio: The ratio $L = \lambda_e/\sigma T$, where λ_e is the electronic component of thermal conductivity, σ is the electrical conductivity, and T is the absolute temperature. In well-conducting metals when the lattice thermal conductivity is negligible, λ_e is sometimes replaced by the total thermal conductivity λ.

Phonon: The energy content of all waves and other harmonic oscillators is quantized, being an integral number of quanta of energy hf, where f is the wave or oscillator frequency and h is the Planck constant. In the case of elastic waves or lattice waves such quanta are called phonons.

Quasiparticle: A particle moving in a medium, such as a crystal lattice, interacting strongly with that medium. The interaction changes the relation between energy and momentum, but in other respects the particle behaves like a free particle.

Residual Resistivity: The electrical resistivity of a metal or alloy at the absolute zero of temperature, where all the resistivity is due to impurities, solute atoms, or other lattice imperfections, and thermal vibrations do not contribute.

Semimetals: Metals with a smaller electron density than ordinary metals having a density of electron states per unit energy interval that varies sufficiently rapidly about the Fermi energy that this density changes appreciably over an energy interval of the order of kT. The electron gas is then no longer highly degenerate. While this definition would depend on the temperature T, the designation of semimetal is made usually at room temperature.

Symfuel: A mixture of uranium dioxide and a series of other oxides to simulate the composition of uranium dioxide nuclear fuels after a specified fraction of the uranium atoms have undergone fission, and the fission products have formed solute cations in the oxide lattice.

Thermal Diffusivity: The ratio of the thermal conductivity divided by the specific heat per unit volume. This is the quantity usually measured by transient methods.

Two-Dimensional Phonon Gas: In crystals consisting of atomic layers that are only weakly bonded to each other, the phonon gas, or the lattice vibrations, can be regarded as approximating the vibrational spectrum of a two-dimensional system over a wide range of frequencies. The most widely studied solid of this kind is graphite.

Works Cited

Berman, R. (1976), *Thermal Conduction in Solids*, Oxford, U.K.: Oxford Univ. Press.

Cahill, D. G. (1990), *Rev. Sci. Instrum.* **61,** 802–808.

Carslaw, H. S., Jaeger, J. C. (1959), *Conduction of Heat in Solids*, 3rd ed., Oxford, U.K.: Oxford Univ. Press.

Grimvall, G. (1986), *Thermophysical Properties of Materials*, Amsterdam: North-Holland.

Kemp, W. R. G., Klemens, P. G., Tainsh, R. J. (1959), *Philos. Mag.* **4,** 845–847.

Klemens, P. G. (1956), "Thermal Conduction in Solids at Low Temperatures," in: S. Flügge (Ed.), *Handbuch der Physik*, Vol. 14, Berlin: Springer, pp. 198–281.

Klemens, P. G. (1969), "Theory of Thermal Conductivity of Solids," in: R. P. Tye (Ed.), *Thermal Conductivity*, Vol. 1, London: Academic, pp. 1–68.

Klemens, P. G. (1991), *High Temperatures—High Pressures* **23,** 241–248.

Klemens, P. G., Williams, R. K. (1986), *Int. Metals Rev.* **31,** 197–215.

Kubičar, L. (1990), *Pulse Methods of Measuring Basic Thermophysical Parameters*, Amsterdam: Elsevier.

Landauer, R. (1978), in: J. C. Garland, D. P. Tanner (Eds.), *Electrical Transport and Optical Properties of Inhomogeneous Media*, AIP Conf. Proc. No. 40, New York: Am. Inst. of Physics, p. 2.

Parrott, J. E., Stuckes, A. D. (1975), *Thermal Conductivity of Solids*, London: Pion.

Phillips, W. A. (1981), *Amorphous Solids, Low Temperature Properties*, Berlin: Springer.

Slack, G. A. (1979), "Thermal Conductivity of Non-Metallic Crystals" in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, Vol. 34, New York: Academic, pp. 1–71.

Sommerfeld, A., Bethe, H. (1933), "Electron Theory of Metals," in: S. Flügge (Ed.) *Handbuch der Physik*, Vol. 24/2, Berlin: Springer.

Steigmeier, E. F. (1969), "Thermal Conductivity of Semiconducting Materials," in: R. P. Tye (Ed.), *Thermal Conductivity,* Vol. 2, London: Academic, pp. 203–251.

Touloukian, Y. S. *et al.* (1970), *Thermophysical Properties of Matter,* Vol. 1, *Metals and Alloys;* Vol. 2, *Non-metallic Solids,* New York: Plenum.

Tye, R. P. (Ed.) (1969), *Thermal Conductivity,* London: Academic.

Uher, C. (1991), in: O. Madelung, G. K. White (Eds.), *Landolt-Börnstein, Numerical Data and Functional Relationships in Science and Technology,* Group III, *Crystal and Solid State Physics,* Vol. 15c, *Thermal Conductivity of Pure Metals and Alloys,* Berlin: Springer, pp. 402–448.

Uher, C. (1992), in: D. M. Ginsberg (Ed.), *Physical Properties of High-Temperature Superconductors,* Vol. 3, Singapore: World Scientific, pp. 159–284.

Wilson, A. H. (1952), *Theory of Metals,* 2nd ed., Cambridge, U.K.: Cambridge Univ. Press.

White, G. K. (1991), in: O. Madelung, G. K. White (Eds.), *Landolt-Börnstein, Numerical Data and Functional Relationships in Science and Technology,* Group III, *Crystal and Solid State Physics,* Vol. 15c, *Thermal Conductivity of Pure Metals and Alloys,* Berlin: Springer, pp. 6–125.

Numerical Data

Hust, J. G., Giarratano, P. J. (1975), *Thermal Conductivity of Solids below 300 K,* NBS Special Publication 260-46, Washington, DC: U.S. Dept. of Commerce.

Madelung, O., White, G. K. (Eds.) (1991), *Landolt-Börnstein, Numerical Data and Functional Relationships in Science and Technology,* Group III, *Crystal and Solid State Physics,* Vol. 15c, *Thermal Conductivity of Pure Metals and Alloys,* Berlin: Springer.

Further Reading

Berman, R. (1976), *Thermal Conduction in Solids,* Oxford, U.K.: Oxford Univ. Press.

Klemens, P. G. (1958), "Thermal Conductivity and Lattice Vibrational Modes," in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics,* Vol. 7, New York: Academic, pp. 1–98.

Parrott, J. E., Stuckes, A. D. (1975), *Thermal Conductivity of Solids,* London: Pion.

Ziman, J. M. (1960), *Electrons and Phonons,* Oxford, U.K.: Oxford Univ. Press.

THERMAL DIFFUSION

See DIFFUSION AND IONIC CONDUCTION IN LIQUIDS; DIFFUSION IN THIN FILMS

THERMAL INSULATION MATERIALS

Robert H. King, Jr., *Technical Insulation Business Group, Polymers Division, Bayer Corporation, Pittsburgh, Pennsylvania, U.S.A.*

Introduction 151
1. **Benefits of Thermal Insulation** 152
2. **Types of Thermal Insulation** 152
2.1 Heat Transfer 153
2.2 Foam Insulation 153
2.2.1 Polyurethane and Polyisocyanurate Insulation ... 153
2.2.2 Styrenic Foams 153
2.2.3 Urea Formaldehyde (U-F) Foam 153
2.3 Fiberglass Insulation 153
2.4 Mineral Insulation 154
2.5 Other Common Insulation Materials 154
2.5.1 Cellulose Fiber Loose Fill 154
2.5.2 Vermiculite 154
2.6 Methods of Manufacture 154
2.6.1 Polyurethane and Polyisocyanurate Insulation ... 154
2.6.2 Styrenic Foams 154
2.6.3 Fiberglass 154
2.6.4 Mineral Insulation 155
2.6.5 Cellulose Fiber Loose Fill 155
3. **Applications** 155
3.1 Construction 155
3.1.1 Residential 155
3.1.2 Commercial 155
3.2 Walls and Ceilings 155
3.3 Appliance Insulation 156
3.3.1 Water Heaters 156
3.3.2 Refrigerators 156
3.3.3 Ovens 156
3.4 Commercial Refrigeration 156
3.4.1 Transportation 156
3.4.2 Warehousing 156
3.5 Process Industries 157
4. **Material Selection Criteria** .. 157
4.1 Thermal Insulation Values 157
4.1.1 Increasing R-Values 157
4.1.2 Efficiency 157
4.2 Advantages and Disadvantages of Various Insulation Materials 158
4.2.1 Polyurethane and Polyisocyanurate Foams 158
4.2.2 Styrenic Foams 159
4.2.3 Fiberglass 159
4.2.4 Mineral Insulation 159
4.2.5 Other Materials 159
4.3 Geographic Demands 160
4.3.1 Climatic 160
4.4 Construction Codes 160
4.4.1 Model Energy Code 160
4.4.2 ASHRAE Standard 90.2 160
4.4.3 Good Cents Program 160
5. **Environmental Considerations** 160
5.1 Energy 160
5.1.1 Conservation 161
5.1.2 Efficiency 161
5.2 CFC Elimination 161
Glossary 161
Works Cited 162
Further Reading 162

INTRODUCTION

Thermal insulation materials help to keep buildings warm when outdoor temperatures are cold and cool when outdoor temperatures are warm. They are also useful in keeping temperatures low in commercial warehouses that store cold or frozen foods, and in trucks, trains, and ships that transport cold and frozen food products. Additionally, thermal insulation is used in the process industries for fire protection, product conservation, temperature control, and energy conservation. There are no fewer than seven different types of insulation materials. Each has its own advantages and disadvantages that are dependent upon the structure in which the materials are used and the climate where they are employed. This article is a survey of information about the properties

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

and benefits of thermal insulation. Discussion will focus on the material composition of various types of thermal insulation available, how to select the best thermal insulation for a particular purpose, the advantages and disadvantages of using each kind of insulation material, international standards, and the environmental and safety ramifications of thermal insulation manufacturing and handling. Additional information on insulation materials is available from several sources listed under Further Reading. Specific facts from some references are cited throughout this article.

It should be noted that all calculations in this article are in British units only. Because the author did not have the original thickness used for the components of the walls described in this article, an accurate conversion to SI units was not possible.

1. BENEFITS OF THERMAL INSULATION

Although there are many applications of thermal insulation, the most familiar is in building construction. The initial construction of a building offers a one-time opportunity to use high-performance building materials that can maximize the structure's thermal efficiency and help reduce its overall energy consumption. Air leakage and moisture infiltration are the leading causes of energy waste within most residential and commercial facilities. High-performance, energy-efficient insulation creates a well-sealed building envelope that can cut utility costs. It can also help the owner maintain a higher market value of the building by allowing for smaller heating and cooling systems and, in some cases, qualify for rebates from utility companies and preferred interest rates from mortgage companies.

Besides energy savings, improving the thermal efficiencies of a building has environmental benefits. Reducing demand for heated or cooled air reduces the need for burning fossil fuels, which produces carbon dioxide and sulfur dioxide, the latter a recognized element in acid rain.

2. TYPES OF THERMAL INSULATION

Insulation is dead air space or a dead gas space sometimes combined with a reflective surface. The insulation materials used to maintain a particular temperature are like the clothing that allows our body temperatures to remain comfortable in every kind of weather. The characteristics of thermal insulation are determined mainly by its composition. Although there are five major types of thermal insulation, some insulations are hybrids of these types. They are flake, fibrous, granular, cellular, and reflective. Among these types of thermal insulation, there are seven major materials from which insulation is made: polyurethane and polyisocyanurate, styrenic foams, urea formaldehyde, fiberglass, minerals, cellulose fiber loose fill, and vermiculite. All forms of insulation consist of thousands of tiny pockets where dead air is encased, protected, and preserved to maintain the desired temperature within a building, vehicle, or storage vessel.

Thermal insulation is produced in several forms:

1. *Blanket thermal insulation*—a flat, flexible thermal insulation that can be placed over flat, curved, or irregular surfaces. May or may not be faced, coated, or reinforced on one or both sides.
2. *Block thermal insulation*—rigid thermal insulation preformed into rectangular units.
3. *Board thermal insulation*—rigid and semirigid thermal insulation preformed into relatively large-area rectangular units.
4. *Cement thermal insulation*—a prepared composition, in dry form, composed of granular, flaky, fibrous, or powdery materials, which when mixed with the correct proportion of water develops into a mix of plastic consistency. The mixture is then applied by pouring or blowing it into a confined area. It can also be applied by hand.
5. *Foamed thermal insulation*—liquids sprayed on or mixed together, which foam and set into cellular or semicellular organic rigid or semirigid insulation.
6. *Pipe thermal insulation*—rigid or semirigid insulation preformed for application to pipe or tubing.
7. *Reflective thermal insulation*—sheets or preformed shapes to provide spaces, in-

corporating reflective sheets that obtain their insulating resistance from their ability to reflect a large part of the radiant energy incident on them.

2.1 Heat Transfer

Heat is energy in a transient form. It is the result of molecular motion and an interacting form. Heat transfer occurs through conduction, convection, and radiation. *Conduction* is the migration of energy through matter from one atom to another, such as in the metal burner of a stove. *Convection* is the transmission of heat by movement of matter, such as air above the surface of a stove. *Radiation* is heat transfer by electromagnetic waves that does not require an intervening medium. It is often called radiant energy because it manifests itself in rays.

Insulation manages the conduction, convection, and radiation of heat. The effectiveness of an insulating material in resisting heat flow is expressed in its R-value. The higher the R-value, the more resistant a material is to heat flow. The R-value is established in laboratory tests, and is the reciprocal of the numerical value of the heat flow through a material in British thermal units (BTUs) per hour, per square foot of material, per degree Fahrenheit of temperature difference between the heat source and "cool" side of the insulation material. For example, the R-value of 1 in. of fiberglass is 3.2 h-ft^2-°F/BTU. Therefore, the R-value of a $3\frac{1}{2}$-in. blanket of fiberglass is R-11. Table 1 gives the R-values of some commonly used insulation materials.

Table 1. Thermal resistance of common insulations. Source: Society of the Plastics Industry, Inc.; Polyurethane Division.

Insulation type	Thermal resistance, R-value per inch thickness
Rigid polyurethane foam	
With impermeable facers	7.1–8.7
With permeable facers	5.6–6.2
Extruded polystyrene	5.0
Expanded polystyrene	3.85–4.35
Phenolic foam	
Closed cell	8.2
Open cell	4.4
Dry glass fiber batts	3.0
Dry mineral wool batts	3.0
Dry mineral wool loose fill	3.0
Dry cellulose fiber loose fill	3.13–3.70
Vermiculite	2.13–2.27

2.2 Foam Insulation

2.2.1 Polyurethane and Polyisocyanurate These materials can be expanded into a cellular structure of very low density. Their R-values, or ability to resist heat flow, are excellent because the blowing agent, a gas of low thermal conductivity, remains trapped in the foam. The kind of gas used varies. For years, the industry used chlorofluorocarbons (CFCs) as the blowing agents to create the foams' cellular structure and to contribute to the products' insulating efficiencies. Blowing agents commonly used today include hydrochlorofluorocarbons, hydrocarbons, and carbon dioxide.

2.2.2 Styrenic Foams These materials are made from polymerized styrene through the extrusion or expansion of polystyrene beads. In expansion, the starting material is small, dense beads containing butane, pentane, or some other volatile gas. When the beads are heated by steam, the gas expands, puffing up the beads. Additional steam heating in a mold fuses the beads into a large block of beadboard that is cut into slabs by pushing it through an array of parallel hot wires.

2.2.3 Urea Formaldehyde (U-F) Foam Generated at the installation site by mixing the U-F resin with a catalyst and detergent and, using compressed air, blowing the frothy product into a wall, U-F sets quickly. Over a period of months, air replaces the water vapor and formaldehyde in the foam's open cell structure.

2.3 Fiberglass Insulation

Fiberglass is a silicate manufactured by melting glass and then extruding it through a whirling hollow cylinder with peripheral perforations to form fibers. Steam jets then impinge on these fibers and attenuate them to a diameter as low as 0.001 in. The resulting wad of glass wool is passed through a spray of thermosetting resin to bind the fibers together and compress it to the desired density. Next, it is baked to set the binder and sliced into batts, blankets, or rigid panels. A relatively new kind of fiberglass is being used in refrigerators as well as in ovens, water heaters, and advanced car batteries. It is composed of a vacuum panel of fiberglass that is thermally crafted inside a steel foil container. Its R-value is more than six times the R-value of polyurethane foam insulations of the same thickness. Its high-temperature resistance can approach 1000 °F. Fiberglass is resilient and has little tendency to mat together unless it is saturated with water.

2.4 Mineral Insulation

Rock wool is similar to fiberglass, except that it is made by melting steel, copper, or lead slag and then spinning it. Slag wool is made from by-product slag produced in the manufacture of steel and other metals. "Rock wool" is, in fact, usually slag wool. Often, the term "mineral wool" is applied to slag wool, although mineral wool is really a generic term covering this group of insulation materials. The properties of rock wool and slag wool are identical to those of fiberglass. In the form of loose fill, they can be blown into walls and attics.

Asbestos is the only natural mineral fiber known. It has been replaced by other insulants because of the severe health threat associated with breathing asbestos fibers. Consequently, it is unusual to find asbestos in structures built during the past 30 years.

2.5 Other Common Insulation Materials

2.5.1 Cellulose Fiber Loose Fill The most common form of cellulose building insulation today is loose fill, which contains recycled paper that is shredded, then pulverized in hammer mills, and treated with a dry fire retardant such as boric acid or ammonium sulfate. In walls, cellulose may settle with time. Under no circumstances, however, should cellulose be allowed to get wet, since exposure to water washes out the fire retardant and compresses the fibers into a small mass not much different from papier-mâché. Other cellulosic insulants include cotton waste from the clothing industry, eel grass, shredded coconut shells, and shredded sugar cane.

2.5.2 Vermiculite This material is expanded mica. Chemically, it is a layered silicate rock containing about 12% water by weight. The rock is crushed and then heated quickly to about 1600 °F. This process causes the water to expand rapidly into steam and forces the layers apart into accordion-like flaky granules. Perlite is made using a nearly identical process. The difference is that the raw material is volcanic rock, which has a different crystal structure than the silicate rock used to manufacture vermiculite. Perlite's finished product is in the form of puffy white granules.

2.6 Methods of Manufacture

2.6.1 Polyurethane and Polyisocyanurate Insulation The isocyanate and blowing agent are mixed and then extruded into a laminator where the resulting slurry foams until it reaches the height where the aluminum or other facer is located. This restricts the thickness of the foam, which adheres itself to the facers. Finished foam boards are then cut to the desired length. These materials have a closed-cell foam structure. They are available as boardstock, laminated panels, or structural insulated panels with a variety of facers. Polyurethane foam can be fabricated and installed on-site by spraying, pouring, and frothing. Polyurethane foams can also be made with metal-faced panels, primarily for use in curtain wall applications.

2.6.2 Styrenic Foams These materials are made into sheets by extrusion and by molding. The extruded product has a higher R-value than the molded one. Extruded polystyrene is made by forcing a hot mixture of polystyrene, a solvent, and a gas at high pressure into a cavity at atmospheric pressure where, because pressure is abruptly reduced, the mixture expands into a foam with fine, tiny closed cells. In making molded sheets, polystyrene beads are placed in a mold and exposed to heat. This process expands the beads and fuses them into a larger block. Blocks and extruded foam are later cut into sheets for use.

2.6.3 Fiberglass This material is made by subjecting molten glass to blasts of air that blow the material into thin threads. Additional details of the manufacture of fiberglass insulation were included in Sec. 2.2.

2.6.4 Mineral Insulation Vermiculite is made from a micalike mineral consisting of aluminum-iron-magnesium silicates that, heated to high temperatures, expand and provide a wide range of densities. Rock wool is manufactured by melting steel, copper, or lead slag and spinning it. Perlite is made from natural glass, siliceous and volcanic in origin. Through rapid heating, the crushed ore particles are expanded from 4 to 20 times their original volume.

2.6.5 Cellulose Fiber Loose Fill Cellulose is manufactured from newsprint, paperboard stock, or virgin wood fiber and is treated with a fire retardant. Additional details of the manufacture of cellulose insulation were given in Sec. 2.4.1.

3. APPLICATIONS

3.1 Construction

3.1.1 Residential Insulation is an important factor in building a home and in reducing the cost of energy used in existing homes. Regardless of the situation, the amount and placement of insulation in a home helps to determine heating and cooling costs, as well as the comfort level of the house. For two reasons, sufficient attic insulation is very important for residential construction. One, heat rises and can escape in large quantities through the attic; and two, attics are easy to insulate because they are usually unfinished. Insulation in walls and floors is also effective in reducing energy costs. Modern doors and windows are designed to reduce interior energy loss or gain during heating and air conditioning seasons, respectively. In a well-insulated home, the insulation acts as a blanket that covers surfaces adjoining areas at different temperatures, such as floors over cold areas, perimeter joists, crawl spaces, and basements.

3.1.2 Commercial Like residential buildings, commercial structures need to be insulated from the elements. The process for determining which areas need to be insulated is the same as that for deciding which parts of a home must be protected from extreme heat and cold. Although many commercial structures lack attics, wood frames, and basements, careful attention must be given to using the best material for insulating the ceiling. Failure to use the right kind of insulation in the right locations of a commercial building can be more costly than failing to insulate a home properly. Depending upon the specific building design and construction site, performance needs for specific insulating systems will vary. Today, the largest contributor to energy waste in buildings is heat loss or gain through walls, roofs, and windows.

3.2 Walls and Ceilings

Most insulation materials can be used over plywood or as sheathing when the framing is adequately installed and supported. Special nailing procedures allow exterior foam sheathing and siding to accommodate thicker composite boards. Insulation boards can also be used to fill cavities between studs, as can fiberglass. Loose materials can be blown into place between studs. Foam core panels can be used for curtain wall applications. These panels are made in a variety of thicknesses and are sandwiched between facers such as steel or aluminum for maximum effectiveness.

When properly applied, polyurethane spray foams have no seams or joints, and so they can eliminate drafts and heat leaks when applied on the interior or exterior of structural walls. Such foams will adhere to most substrates that are clean and free of grease, oil, or loose scale. On the interior, these foams can be used to fill cavities between studs or can be sprayed on block walls or other wall surfaces. Rigid polyurethane foams can be poured or frothed in place on the job site. These foams can be used to fill cavities in concrete or cinder blocks, or in spaces between masonry walls. Within minutes, the foam sets to a strong, closed-cell material that fills voids and does not shrink or settle.

Polyurethane and polyisocyanurate foams are easy to install when used for roof applications. An important advantage of polyisocyanurate roofing boards and spray polyurethane foam is that they can be installed directly on metal decks, provided their fire performance meets applicable building codes. Rigid polyisocyanurate roofing boards are lightweight and can be manufactured with a variety of flexible facers, including aluminum foil, asphalt, saturated felt, organic/inorganic felts, or multilayered composites. Rigid facers such as plywood, wafer board, or gypsum board can also be used. These composite boards have exceptional structural strength.

Lightweight rigid polyurethane spray foam roofing is applied on the job site, creating a seamless insulated surface. Spray polyurethane foams adhere to almost any substrate that is sound, clean, dry, and free of loose material. They conform easily to complex surface configurations and, therefore, are particularly effective for domed roofs and for sealing around projections such as pipes, conduit, air conditioning equipment, and bracing. Polyurethane foam also adheres to itself, allowing layers to be built up to required thicknesses and to create positive drainage pitches.

3.3 Appliance Insulation

3.3.1 Water Heaters Wrapping a water tank with a 4-in. fiberglass or foam blanket is one of the easiest ways to reduce heat loss. For the typical 40-gal tank, a thermal blanket can save hundreds of dollars per year in electricity or gas bills (Gay, 1980). Such a blanket is necessary because manufacturers cannot increase tank thickness significantly, as tanks must often pass through doorways and down basement stairs during installation. Applying a 1-in. blanket of fiberglass or urethane pipe insulation to water lines close to the hot water tank can cut energy loss from those sources by up to 50%.

3.3.2 Refrigerators Next to water heaters, refrigerators are the second-highest energy users in the home. A typical 16-cubic-foot refrigerator uses about 1800 kWh/yr compared to more than 4000 kWh/yr for a 40- or 50-gal water heater. The best refrigerators are filled with about 2.5 in. of polyurethane between the inner liner and cabinet. Urethane-filled appliances use about 20% fewer kilowatt hours per year than fiberglass-insulated refrigerators manufactured before the availability of urethane.

3.3.3 Ovens A relatively new kind of fiberglass insulation has an R-value more than six times the R-value of polyurethane foam insulations. It is a vacuum panel of thermally crafted fiberglass inside a steel foil container, and it can be used in refrigerators, water heaters, and advanced automotive batteries, as well as in ovens. This type of insulation is suitable for applications where the inside temperature approaches 1000 °F (Anon., 1993).

3.4 Commercial Refrigeration

3.4.1 Transportation The frozen food industry depends on insulation for its very existence. At the processing plant, low-temperature freezers consisting of insulated enclosures make it possible for food to be frozen immediately after it is processed. Foods are then transported in refrigerated railroad cars, trucks, and ships to wholesalers and dealers. Insulation used in refrigerated vehicles usually consists of reinforced fiberglass panels with polyurethane insulation that is 5–6 in. thick and foamed in-place (Anon., 1977–1981, p. 153). Some vehicles refrigerated by carbon dioxide use polyurethane boards to minimize heat loss. Generally, the liquid carbon dioxide changes into a vapor as it reaches piping nozzles. The vapor, which is heavier than air, moves down through vents around the lading.

3.4.2 Warehousing For the past 30 years, panels made from polystyrene and polyurethane protected by metal skins have been the most efficient materials to insulate refrigerated warehouses. Panels help to ensure that high-quality products are installed at a job site because they are produced under ideal conditions in a manufacturing plant, can be erected quickly, and last a long time. In contrast, many variables can affect the quality of materials foamed in-place on a job site. Panelized construction also provides a superior vapor barrier and can be moved from one location to another if the user's needs change. Panel bow due to skin temperature variation is desirable to prevent delamination. White-colored panels reflect more heat and have subsequently lower surface temperatures than natural aluminum or galvanized or stainless steel. The lighter the color, the less the panel tends to bow and the lower the exterior surface temperature. Prior to the introduction of polystyrene and polyurethane panels, insulation consisted largely of fiberglass sandwiched between two layers of plywood. The insulation properties of these materials are inferior to those of polystyrene and polyurethane. Consequently, they are no longer used in commercial refrigeration applications.

3.5 Process Industries

In the process industries, semirigid pipe insulation is used to protect buildings and pipes from fires, as well as to minimize the heat transfer from a material in transit or in storage vessels and in pipes. Using insulation to minimize a product's heat loss is becoming more widespread as materials are loaded into tank cars and hauled to users, arriving at their destinations within 5–15 °F of their temperature at loading (Anon., 1977–81, p. 213). Using standard heat-transfer formulas, one can calculate the required insulation thickness required for specific applications.

4. MATERIAL SELECTION CRITERIA

4.1 Thermal Insulation Values

When choosing insulation materials, the most important criterion is thermal efficiency, or a material's ability to block heat flow. Thermal insulation values are expressed as R-values, or a measure of the material's resistance to heat transfer. Pockets of air slow down the flow of heat because still air is a poor heat conductor and because convection is suppressed in small pockets of air. The closed-cell foam structure of polyurethane and polyisocyanurate foams offer very effective air infiltration resistance, minimal moisture pickup, no harboring of fungi and mildew, excellent dimensional stability, no shrinkage and settling, and resistance to organic solvents.

4.1.1 Increasing R-Values The R-values of polyurethane and polyisocyanurate foams more than double as the thickness doubles. For example, if a 1-in. piece of rigid polyisocyanurate foam insulation with a permeable facer has an R-value range of 5.6 to 6.2, a 2-in.-thick piece of the same material has an R-value range of 12.1 to 13.6. These higher R-values can reduce the thickness of the insulation required.

In contrast, styrenic foams, whether extruded or expanded, have R-values of about 4.0 to 5.0, fiberglass batts around 3.0, and cellulose fiber, mineral wool, and vermiculite still lower. The R-values of all these materials increase arithmetically relative to their thicknesses. Two inches of styrenic foams, for example, has an R-value of 8.0 to 10.0, and 2 in. of fiberglass batt insulation has an R-value of 6.0.

4.1.2 Efficiency The R-values of various layers in a thermal shell can be used to calculate the total R-value, or R(tot). The total R-value is the sum of all the R-values comprising the insulating shell. In the case of a 2 × 4-in. stud wall that is fiberglass-insulated and a double-pane window, R(tot) for each would be as follows:

Wall, between the studs	R-Value
Air film (outside, in 7.5 mph wind)	0.25
1/2-in. gypsum board	0.45
3.5-in. fiberglass	11.00
5/8-in. plywood	0.80
Air film (inside)	0.70
R(tot) =	13.20

Window	R-Value
Air film (outside, in 15 mph wind)	0.15
Two 1/8-in. glass panes	0.05
4-in. air space	1.00
Air film (inside)	0.70
R(tot) =	1.90

The total R-value of a section of the thermal shell and temperatures on each side of the section can be used to estimate the amount of heat flowing through the section. This calculation does not account for transient behavior (variations with time) or for defects or inclusions in the materials of construction, which may alter the heat flow. The basic heat-flow equation is

$$q = \frac{A(t_1 - t_2)}{\mathrm{R(tot)}}$$

where q = rate of heat flow, BTU/h, A = the area of the section (ft^2), t_1 = warmer temperature (°F), t_2 = colder temperature (°F), and R(tot) = total thermal resistance [(ft^2-°F-h)/BTU].

That equation may be used to calculate the rate at which heat is flowing into the house through *one square foot* of the wall by rewriting it as $q/A = (t_1 - t_2)/R$. The following example assumes an outside temperature of 90 °F and an inside temperature of 75 °F, and uses the R-value given previously for the wall between the studs of a 2 × 4-in. stud wall with a typical fiberglass.

$$\begin{aligned} q &= \frac{1 \times (90 - 75)}{13.2}\ \text{BTU/h} \\ &= \frac{1 \times 15}{13.2}\ \text{BTU/h} \\ &= 1.14\ \text{BTU/h per square foot of wall area.} \end{aligned}$$

The same equation can also be used to calculate the rate at which heat is flowing through the window. This example assumes an outside temperature of 15 °F, an inside temperature of 70 °F, and a window that is 3 × 5 ft.

$$\begin{aligned} A &= 3\ \text{ft} \times 5\ \text{ft} = 15\ \text{ft}^2 \\ q &= \frac{15(70 - 15)}{1.9} = \frac{15 \times 55}{1.9}\ \text{BTU/h} \\ q &= 434\ \text{BTU/h or } q/A = 28.93\ \text{BTU/h per square foot of window area.} \end{aligned}$$

A similar calculation can be done for the portion of the wall that includes the stud. To complete the calculation of the total heat loss, the known areas of each section of the composite wall structure are multiplied by the heat loss per square foot of each part of the section. To calculate the total and effective R-values of an insulated 2 × 6-ft^2 stud wall with a high–R-value interior sheathing of polyisocyanurate foam, separate calculations must be performed for the framed section and the insulated section; see Table 2.

4.2 Advantages and Disadvantages of Various Insulation Materials

All insulation materials may require a vapor barrier between the material and the interior of the building. They also make it necessary for buildings to have a mechanical fresh-air exchange system that compensates for the loss of natural air flow.

4.2.1 Polyurethane and Polyisocyanurate Foams These materials provide a high strength-to-weight ratio and resist shrinkage and settling in cavities. Additionally, they resist moisture pickup, many organic solvents, fungus, and mildew. These foam insulation materials adhere to most common building materials, adapt to mold configurations for the production of complicated shapes and intricate surface detail, require no special tools for shaping, and produce seamless surfaces when applied by spray.

Disadvantages include the need for thermal protection from fire on the interior, such as a 1/2-in. gypsum wallboard, unless fire testing demonstrates that such thermal barriers are unnecessary. Polyurethane and

Table 2. Calculation of effective R-value for composite section.

	R-values in framed section	R-values in insulated section
Outside air film (7.5 mph wind)	0.25	0.25
Wooden shingles	0.85	0.85
1/2-in. plywood (R = 1.25/in.)	0.60	0.60
5-5/8-in. studs (R = 1.25/in.)	7.05	—
Fiberglass batts, R-19	—	19.00
1-in. polyisocyanurate foam	7.50	7.50
1/2-in. gypsum board	0.45	0.45
Inside air film	0.70	0.70
R(tot) =	17.40	29.35

If the framing constitutes 15% of the total wall area, then the effective R-value of the wall is

$$R(\text{eff}) = \frac{1}{0.15/17.40 + 0.85/29.35} = 27.03$$

polyisocyanurate foams also require proper precautions during foaming and spraying operations to prevent accidental ignition of the foam. When used for exterior applications, these materials need weatherproof protection from sunlight and other weather conditions. If the insulation materials are exposed continually to temperatures greater than 175 °F, physical property deterioration may occur.

4.2.2 Styrenic Foams Polystyrenes of the expanded and extruded varieties are dense products that are hard to ignite, and their strength makes them especially suitable for use on top of roofs or to face exterior basement walls. The disadvantage of these materials is that they degrade in sunlight. They should be covered in most applications to prevent degradation.

4.2.3 Fiberglass With a relatively high R-value, fiberglass is easy to work with, does not readily absorb water, is nonflammable (although its resinous binder is flammable), and does not rot. If the paper backing on fiberglass batts and blankets is impregnated with asphalt, however, the flammability is greater. Although the R-value for fiberglass batts and blankets is usually high, its R-value *per inch* is lower than that of some foams. Working with fiberglass can irritate skin, eyes, and lungs. Studies on whether fiberglass is a carcinogen are inconclusive, and precautions to prevent exposure are desirable. Individuals installing batts or loose fill should wear a hat, pants and a long-sleeved shirt, safety goggles or glasses, and a disposable dust mask to guard against any potential harm.

4.2.4 Mineral Insulation The properties, advantages, and disadvantages of rock wool and slag wool are virtually identical to those of fiberglass, and one should take similar precautions when installing such materials.

4.2.5 Other Materials Cellulose is economical and environmentally safe, and provides nearly the same insulating performance as fiberglass. Additionally, the major cost of its manufacture is paid for when cellulosic insulant materials are used for their initial purposes. The disadvantages of cellulose include its tendency to settle over time, flammability, and high rate of moisture absorption. If the insulated area of the building is accessible, however, settling can be overcome by "topping," or adding additional material to, the insulation periodically.

Urea formaldehyde (U-F) is a relatively inexpensive material that can be injected evenly into any wall cavity. Its disadvantages are that it may shrink after injection and give off lingering formaldehyde vapors. Because it is blown in on the job site, the quality of U-F foam can vary from one installation to another. Whether the foam is effective also depends upon numerous variables, including temperature and humidity on injection day, how the ingredients are mixed, age of the ingredients, and the type of injection apparatus used. Some U-F foams can experience shrinkage or break down at temperatures and moisture levels only slightly above those found in a typical house or commercial building. U-F is generally not applied to attics because direct exposure to the atmosphere and high attic temperatures accelerates drying and causes excessive shrinkage and cracking. Since formaldehyde gas is heavier than air, it can settle into the rooms below.

Vermiculite and perlite are both noncombustible and can be used in high-temperature applications, such as industrial furnaces. They can be mixed with concrete and poured into voids in cement block walls to reinforce home insulation. Vermiculite and perlite, however, have relatively low R-values.

Reflective insulation, such as aluminum foil, can serve as an insulant between a steam or hot-water radiator and an adjacent wall. In such a narrow space, there is generally insufficient room for fiberglass or another bulky insulant, but a single layer of foil reduces heat loss through that part of the wall directly behind the radiator. Similarly, a wall may be protected from the radiant energy of a wood stove by using a protective layer of shiny aluminum. The aluminum should be positioned an inch away from the wall so that it may be cooled by air currents on both sides, and it should not be paper-backed. Aluminum foil is not effective in walls because several layers of it can trap moisture and negate its intended purpose.

4.3 Geographic Demands

4.3.1 Climatic Structures in the harshest climates, such as Scandinavia and the northeast and midwest regions of the United States, require insulation with an R-value of around 50, regardless of whether they are heated by electricity, gas, oil, or a heat pump. Buildings in moderate climates such as those of Japan, the Mediterranean Sea region, and the southern and southwestern United States should use insulation with R-values between 30 and 40. Regardless of climate, thermal insulation used should have the ability to resist moisture and to withstand the effects of broad temperature variations and humidity.

The cost of thermal insulation materials *per R-value per square foot* ranges from around $US 0.01 for fiberglass batts and loose fill to $US 0.077 for polyisocyanurate and extruded styrenic foam, both in rigid board form. In between those two extremes are cellulose at $US 0.014, loose-fill mineral wool at $US 0.019, and expanded styrenic foam at $US 0.046 (Council of American Building Officials, 1992).

4.4 Construction Codes

4.4.1 Model Energy Code The Model Energy Code (MEC) was developed by the Council of American Building Officials. The MEC governs the design and construction of new buildings that are heated and cooled. It also requires that plans and specifications show in sufficient detail pertinent data about and features of the building and its systems, including design criteria, construction materials and their heat transmission rates, and the efficiency of the insulating materials. The MEC energy-efficiency performance standards are based on several criteria, including R-values of the various components and heating degree days (HDDs). Reflecting local climate, HDDs are determined by totaling over a year the temperature differences above or below 65 °F each day, and they are used in estimating fuel consumption.

4.4.2 ASHRAE Standard 90.2 Developed by the American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE), Standard 90.2 establishes specific design parameters governing the transmission of heat and cooled air through building components for the purpose of maximizing thermal efficiency in residential construction, including single-family houses, multiple-family structures of three stories or fewer above grade, mobile homes, and modular houses. ASHRAE Standard 90.2 is similar to the MEC. It tends to predict thermal performance for buildings in warm climates more accurately than for structures in colder environments.

4.4.3 Good Cents Program An energy conservation initiative developed by Southern Electric International, a division of The Southern Company, which is headquartered in Atlanta, Georgia, U.S.A., the Good Cents Program provides energy standard guidelines for builders, owners, and operators to construct energy-efficient homes and buildings. Building contractors who participate in the Good Cents Program are committed to using high levels of insulation that allows for the downsizing of heating, ventilation, and air conditioning systems, among other energy-efficient building practices. Nearly 300 utilities in 30 states participate in the Good Cents Program. If the MEC, ASHRAE Standard 90.2, and the Good Cents Program were incorporated into new residential construction only, the United States could save more than 7.2 trillion BTUs of energy each year. Similar savings would be available worldwide if all building contractors adopted these or similar approaches to energy conservation (Council of American Building Officials, 1992).

5. ENVIRONMENTAL CONSIDERATIONS

5.1 Energy

The single largest contributor to energy waste in commercial and residential construction is heat loss or gain through walls, roofs, and windows. Rigid polyurethane and polyisocyanurate foam are the most effective materials for keeping indoor temperatures at their desired levels and minimizing the use of energy to heat or cool to a comfortable temperature. All insulation materials, however, reduce heat loss or gain. To determine which ones are best for a particular application, all of the variable factors must be considered and the projected savings relative to the investment in insulation calculated.

5.1.1 Conservation More than one-third of the energy consumed in the United States is used to heat and cool residential and commercial buildings. The Model Energy Code (MEC), American Society of Heating Refrigerating and Air Conditioning Engineers (ASHRAE) Standard 90.2, and the Good Cents Program are U.S.-based efforts to help conserve energy by promoting construction that is appropriate to specific climate zones.

As mentioned in Sec. 4.4.3, if new residential construction were built to the stipulations of the MEC, the United States would reduce its energy consumption by more than 7×10^{12} BTUs in a single year. That amount of energy saved would be enough to power New York City for several years. Over a 50-yr period, emissions of carbon dioxide could be reduced by 280 million tons (Council of American Building Officials, 1992).

5.1.2 Efficiency As discussed in Secs. 2.5 and 4.1.2, polyurethane and polyisocyanurate foam insulation have the highest R-values, followed by extruded polystyrene, fiberglass, mineral wool batts and loose fill, cellulose, and vermiculite.

5.2 CFC Elimination

Chlorofluorocarbons (CFCs), a family of fluids that are extremely stable, fully halogenated, and nontoxic, have been at the center of the ozone depletion and global warming debate. CFCs had been used as blowing agents to create the cellular structure of foam insulation and contribute to the products' R-values. Regarded by many scientists and policymakers as a key weapon in the war against energy waste in the 1970s, they are now considered a major contributor to the problem of ozone depletion in the upper atmosphere.

In 1987, 30 countries and the European Community signed the Montreal Protocol, which called for a 50% reduction in CFCs by mid-1998 and a complete ban on CFCs by the year 2000. A 1993 agreement reset the complete ban to 1 January 1996 (Anon., 1994a). Manufacturers of foam insulation board have eliminated CFCs ahead of schedule, developing safer products. As of 31 December 1993, members of the Polyurethane Insulation Manufacturers Association (U.S.A.) had stopped using halogenated CFCs in the manufacture of their products.

Some of the most promising replacements for CFCs are hydrochlorofluorocarbons (HCFCs). They pose less threat to the ozone layer than CFCs because they break down before they reach the upper portions of the atmosphere. However, they still are ozone-depleting substances, and they too will be phased out. To reduce the amount of CFCs used in the manufacture of polyurethane and polyisocyanurate products, some manufacturers have mixed water into their formulations. The R-values for these products, however, go down accordingly, and these materials could lose their R-values in applications faster than their predecessors did. This phenomenon is known as *thermal drift.*

CFC-11 and CFC-12, two commonly used blowing agents in the manufacture of rigid foam and polyurethane insulation, have been replaced by HCFC-141b and HCFC 142b. These replacement blowing agents are made using pentane, a solvent that contributes to localized smog, and are about 11% as damaging to the ozone as CFC-11 (Anon., 1994b). Manufacturers of rigid foam and polyurethane insulation continue to work on a third generation of blowing agents.

Through a public–private research project involving the United States Department of Energy, Oak Ridge National Laboratory, the United States Environmental Protection Agency, Polyurethane Insulation Manufacturers Association, the National Roofing Contractors Association, and the Society of the Plastics Industry, manufacturers are working on CFC substitutes that will not harm the atmosphere. The cost, structural soundness, fire resistance, and toxicity of these new materials are remarkably similar, if not identical, to those of the substances they replaced (Brown, 1994).

GLOSSARY

Batt: A piece of insulation, of the flexible type, cut into easily handlable sizes, square or rectangular in shape, with a vapor barrier on one side, and with or without a container sheet on the other side.

British Thermal Unit (BTU): The amount of heat that will raise the temperature of 1 lb of water by 1 °F.

Chlorofluorocarbons (CFCs): A family of fluids that is stable, hydrogen free, and nontoxic, widely used until recently as blowing agents to expand the cells in the manufacture of rigid foam boards.

Conduction: The migration of heat through matter from atom to atom.

Convection: The transmission of heat by movement of bulk matter.

Heat: Energy in transient form. The result of molecular motion and interacting forces.

Polystyrene: A resin made by polymerization of styrene as the sole monomer.

R-value: The effectiveness of an insulation material in resisting heat flow. R-value is established through laboratory testing that measures the heat flow through a material in BTUs per hour, per square foot of material, and per degree of temperature difference between the heat source side and the "cool side" of a structure.

Radiation: The transmission of heat without the benefit of a medium.

Thermal Drift: The tendency of insulation to lose R-value over time. This change in R-value stems from slow leakage of the gas trapped within the material's cells.

Thermal Insulation: Material having air- or gas-filled pockets, void spaces, or heat-reflective surfaces, which, when properly applied, will retard the transfer of heat with reasonable effectiveness under ordinary conditions.

Works Cited

Anon. (1977–1981), *IARW Proceedings*, Bethesda, MD: International Association of Refrigerated Warehouses (IARW).

Anon. (1993), *Des. News* **48** (7), 28–30.

Anon. (1994a), *Building Mag.* **44** (1), 73.

Anon. (1994b), *Arch. Digest* **51** (4), 34–35.

Brown, Linda (1994), *Am. Health* **13** (9), 51.

Council of American Building Officials (1992), *Model Energy Code*, Falls Church, VA: Council of American Building Officials.

Gay, Larry (1980), *The Complete Book of Insulating*, Brattleboro, VT: Stephen Green Press, p. 47.

Further Reading

Brief Introductory Works

U.S. Department of Energy (1987), *United States Department of Energy Insulation Fact Sheet*, Washington, DC: United States Department of Energy, Assistant Secretary, Conservation and Renewable Energy.

General Texts

Fisk, D. J. (1981), *Thermal Control of Buildings*, London: Applied Science Publishers.

Ford, Charlotte (1978), *Labor and Net Energy Effects of a National Ceiling Insulation Program*, Washington, DC: United States Department of Energy, Division of Industrial Energy Conservation.

Klempner, Daniel, Frisch, Kurt C. (1991), *Handbook of Polymeric Foams and Foam Technology*, New York: Hanser Publishers.

Langley, B. (1994), *Refrigerant Management: The Recovery, Recycling, and Reclaiming of CFCs*, Albany, NY: Delmar.

McElroy, D. L. (1990), *Chlorofluorocarbon Technologies Review of Foamed-Board Insulation for Buildings*, Oak Ridge, TN: Oak Ridge National Laboratory.

Mooz, William E. (1980), *Flexible Urethane Foams and Chlorofluorocarbon Emissions*, Santa Monica, CA: Rand.

Muncey, R. W. R. (1979), *Heat Transfer Calculations for Buildings*, London: Applied Science Publishers Limited.

Powell, Frank J. (1987), *Thermal Insulation: Materials and Systems*, Philadelphia, PA: American Society for Testing and Materials.

Turner, William C., Malloy, John F. (1981), *Thermal Insulation Handbook*, New York: McGraw-Hill.

Tye, R. P. (1977), *Thermal Transmission Measurements of Insulations*, Philadelphia, PA: American Society for Testing and Materials.

Williams, Mark F. (1995), *Exterior Insulation and Finish Systems (EIFS)*, Philadelphia, PA: American Society for Testing and Materials.

Medical Studies

Bunn, W. B., III, Bender, J. R. (1993), *J. Occ. Med.* **35** (2), 101–113. The most recently completed animal inhalation studies on refractory ceramic fibers, fibrous glass, and mineral wool, conducted on rats, are discussed.

L'Abbe, Kristan A., Hoey, John R. (November 1988), *Am. J. Pub. Health* **78,** 1489–1491. Presents an assessment of the health impact of urea formaldehyde foam insulation.

THERMAL PLASMA DEVICES

J. V. R. Heberlein, *University of Minnesota, Minneapolis, Minnesota, U.S.A.*

R. E. Voshall, *Gannon University, Erie, Pennsylvania, U.S.A.*

Introduction to Thermal Plasmas 163
1. **Generation and Characterization of Thermal Plasmas** 165
1.1 Electric-Arc Discharge 165
1.1.1 Cathode Region 165
1.1.1.1 Thermionic Emission 166
1.1.1.2 Field Emission 166
1.1.1.3 Thermionic–Field or TF Emission 166
1.1.1.4 Explosive Emission 166
1.1.2 Anode Region 167
1.1.3 Arc Column 167
1.2 Radio-Frequency Induction Plasmas 168
1.3 Alternative Thermal-Plasma Generation Methods 168
1.4 Thermal-Plasma Characterization 168
1.4.1 Thermal-Plasma Diagnostics 169
1.4.1.1 Plasma Temperature 169
1.4.1.2 Electron Densities 169
1.4.1.3 Plasma Velocity 170
1.4.2 Plasma Characterization by Modeling 170
2. **Plasma Torches and Arc-Gas Heaters** 171
2.1 Plasma-Generator Device Purpose and Categorization 171
2.2 Plasma-Spray Torches 172
2.3 Plasma Torches for Gas Heating and Chemical Processing 175
2.4 Transferred-Arc Devices 176
2.5 Plasma Torches for Space Applications 178
3. **Plasma Light Sources** 179
3.1 History and Categorization of Discharge Lamps 179
3.2 High-Intensity Discharge Lamps 180
3.2.1 Some Common Characteristics 180
3.2.2 High-Pressure Mercury Lamps 182
3.2.3 Metal-Halide Lamps 183
3.2.4 High-Pressure Sodium Lamps 183
3.3 Specialty Lamps 184
3.3.1 Short-Arc Lamps 184
3.3.2 Carbon-Arc Lamps 184
4. **Current-Interruption Devices** 185
4.1 Fundamentals of Circuit Breakers 185
4.2 Low-Voltage Air Circuit Breakers 186
4.3 Oil Circuit Breaker 187
4.4 Vacuum Circuit Breaker 188
4.5 Compressed-Air and SF_6 Breakers 189
4.6 Concluding Remarks 190
Works Cited 190
Further Reading 191

INTRODUCTION TO THERMAL PLASMAS

Plasmas are partially or fully ionized gases consisting of electrons, ions, and neutral atoms and possibly molecules. Plasmas are abundant in the universe; as a matter of fact, more than 99% of the universe is plasma, making it the most abundant state of matter. But also on Earth, plasmas appear in nature in the form of lightning, or as aurora borealis in the upper atmosphere. But only in the beginning of the 19th century did investigations on laboratory plasmas begin, after electricity had become available in the laboratory through the use of galvanic cells.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

The characteristics of a plasma are described by the electron density n_e and the electron temperature T_e, and these two properties can vary over a wide range, giving plasmas a wide variety of characteristic features. Thermal plasmas occupy a relatively small region in n_e-T_e space, as shown in Fig. 1. Typical ranges for n_e in thermal plasmas are 10^{22} to 10^{25} m^{-3}, and ranges for T_e are 5000 to 50 000 K. What distinguishes "thermal plasmas" from other types of plasmas is that

1. all species have a Maxwellian velocity distribution according to the same temperature, i.e., the temperatures of the electrons, ions, and atoms and/or molecules are identical;
2. the excited energy states are populated according to a Boltzmann energy distribution; and
3. the composition of the plasma can be derived from chemical equilibrium relations, including the Saha equation for the determination of the degree of ionization (see Sec. 1.4.2).

The term "thermal plasma" is derived from the fact that these types of plasmas are closer to thermodynamic equilibrium than any other type of plasma. The listed requirements for thermal plasmas are met in a state of complete thermodynamic equilibrium; however, this state requires in addition a blackbody radiation field in the plasma, a condition that is generally not met in terrestrial plasmas. But the conditions still are met when a state of "local thermodynamic equilibrium" (LTE) prevails, i.e., when the energy transfer from one particle to another takes place over a distance that is small compared to variations in the plasma properties. This is the reason why minimum values for the electron density and the electron temperature are needed for a thermal plasma, because there must be a sufficient number of collisions with electrons having energies above the threshold energies for excitation, dissociation, and ionization.

These requirements for existence of a thermal plasma lead to the following characteristics:

1. high energy densities, typically on the order of 10^7 to 10^9 $\mathrm{J/m^3}$, largely because of the availability of ionization and frequently also dissociation energies;
2. high luminosity because of the high population densities of excited states; and
3. high electrical and thermal conductivities because of the presence of highly mobile electrons.

These characteristics present the attractions for the use of plasmas as well as the difficulties in controlling them:

- The high energy densities make plasmas useful for material treatment—any material can be melted, vaporized, dissociated

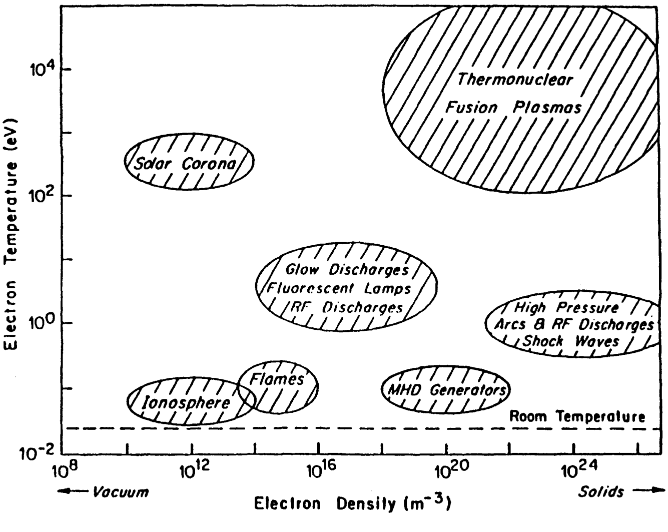

FIG. 1. Electron temperatures and electron densities for different types of plasmas (Fauchais *et al.*, 1987, with kind permission of the authors).

and even ionized with the aid of a thermal plasma, and numerous applications have evolved in the area of materials treatment starting with the establishment of the arc furnace for steel melting in the second half of the 19th century. Today, thermal-plasma technology is an indispensable and growing part in the materials-processing industry.

- The high luminosity of the thermal plasma has led to the development of plasma light sources as early as 1850, and plasma light sources have taken over illumination of the majority of the larger illuminated places.
- The high electrical and thermal conductivities and their variability with temperature as well as the high energy densities of the thermal plasma have resulted in a naturally evolving application. When an electric circuit is opened while a current is flowing in it, the circuit voltage across the opening gap results in the formation of a plasma. Cooling the plasma reduces the electrical conductivity and will result in the interruption of the circuit current.

The high energy densities of thermal plasmas require that any solid surface exposed to the plasma has to be cooled to insure its integrity. This cooling results in an energy loss from the plasma, and steady-state plasmas need an energy source for continuous generation of the plasma. The design of appropriate cooling mechanisms is a central task for the designer of thermal-plasma devices. The steep temperature gradients between the plasma and the cooled surface also result usually in nonequilibrium conditions in the boundary layer between the plasma and the surface, which have to be considered when the performance of the specific plasma device is to be predicted. On the other hand, the fact that a plasma is electrically conducting offers the use of electrical discharges as a convenient means to generate plasmas: a current flowing through the plasma dissipates sufficient energy to make up for the energy losses to the cooled environment.

This survey of thermal-plasma devices first reviews the generation and characterization of thermal plasmas. Then the three application areas in which specific characteristics of the thermal plasma are used are discussed: thermal-plasma devices for treatment of materials, for serving as light sources, and for interrupting the current in an electric circuit.

1. GENERATION AND CHARACTERIZATION OF THERMAL PLASMAS

For steady-state operation, thermal plasmas are in general generated by electric discharges, i.e., by Joule heating of a gas by the current flowing through this gas. Transient thermal plasmas can be generated by pulsed discharges (spark or arc discharges) but also by shock waves and laser pulses. Since most thermal-plasma devices operate in steady state or quasisteady state (i.e., thermal equilibrium is obtained), more emphasis is placed on the description of discharge-generated thermal plasmas. Among those, the electric arc is most frequently used for generation of thermal plasmas, with high-frequency induction discharges the second most frequently encountered method.

1.1 Electric-Arc Discharge

Among all electrical discharges, electric arcs have characteristically the highest current densities (10^3 to 10^5 A/cm^2) and the lowest electric fields (typically 5 to 50 V/cm), and among thermal plasmas they possess the highest energy density (10^7 to 10^9 J/m^3) (see SPARKS, ARCS, AND OTHER ELECTRIC DISCHARGES). Photographs of an electric-arc discharge are shown in Plate 1. The arc discharge is usually divided into three regions, in which different physical effects dominate: the cathode region, the arc column, and the anode region. Figure 2 illustrates the different regions of the arc and the potential distribution along the discharge. Each thermal-plasma device using an arc for plasma generation is characterized by the design of its cathode, its anode, and the column region.

1.1.1 Cathode Region The cathode is at a negative potential with respect to the arc discharge. It therefore attracts positive ions. Since most of the electric current is carried by electrons, the cathode must emit electrons that allow the current to flow

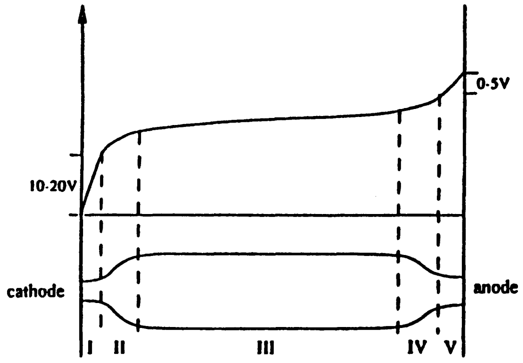

FIG. 2. Schematic of the different regions of an electric arc and the potential distribution along the arc axis. I, cathode-fall region; II, cathode boundary layer; III, arc column; IV, anode boundary layer; V, anode-fall region. Distances are not to scale.

through the sheath region in front of the cathode to the arc column. Electrons can be released from the solid cathode surface by different physical mechanisms: thermionic emission, field emission, or a hybrid of these mechanisms, TF emission, and a mechanism sometimes called explosive emission with evaporation from microscopic spots.

1.1.1.1 Thermionic Emission. This type of electron emission takes place when a material can be heated to sufficiently high temperatures (2500 to 4000 K) that enough energetic electrons exist in the solid that can escape the bonding imposed by the lattice. Usually refractory metals are required for thermionic cathodes, and the arc attachment is usually a stationary high-temperature spot on the cathode. The materials property controlling the performance of a thermionic cathode is the work function, the average energy required to extract one electron from the metal surface. The relation between the electron current density j_e, the cathode surface temperature T, and the work function Φ is given by the Richardson–Dushman equation,

$$j_e = \frac{4\pi k^2 e m_e}{h^3} T^2 \exp\left(-\frac{e\phi}{kT}\right), \tag{1}$$

where e and m_e are the electron charge and mass, respectively, h is Planck's constant, and k is the Boltzmann constant. Thermionic cathodes are characterized by relatively low erosion rates for nonoxidizing plasma gases. Tungsten and graphite are the most frequently used thermionic cathode materials, and additions of rare-earth oxides, alkaline-earth oxides, or thorium oxide are frequently used to lower the value of the work function. A lower work function leads to emission of the required number of electrons for the chosen operating current at lower cathode surface temperatures, thus reducing cathode erosion. For most operating conditions, the dominant heat input into the cathode is through the ion flux hitting the surface, and the dominant cooling mechanism is the release of electrons. Thermionic cathodes can be found in devices ranging from discharge lamps at one or a few amperes and with power dissipation of 100 to 1000 W to arc furnaces for steel production using currents of up to 100 kA and power levels up to 100 MW.

1.1.1.2 Field Emission. Emission of electrons from a cold metal surface takes place when a high electric field in front of the surface allows a substantial number of electrons to tunnel through the potential barrier separating the conduction band of the metal lattice from the surface. Fields in excess of 10^9 V/m are usually necessary, and, in general, localized field enhancement through microscopic protrusions from the surface is required. There is probably no thermal-plasma device with a cathode operating purely with field emission.

1.1.1.3 Thermionic–Field or TF Emission. This type of emission covers a wide range of operating conditions, from a reduction of the electric field required for field emission when the surface is heated or even molten, to the reduction in temperature required for thermionic emission from a refractory metal when a high electric field is present in front of the surface. Sometimes, the term TF emission is used to describe all emission mechanisms other than thermionic and field emission.

1.1.1.4 Explosive Emission. When a local surface inhomogeneity results in very high current densities, high localized heat flux or heat dissipation results in rapid evaporation of microscopic portions of the cathode metal and partial ionization of the metal vapor. This process is inherently unsteady on a very short time scale with a continuous random motion of the arc attachment spot. Application of a magnetic field can influence the motion and is used to control the arc attachment. The material evaporation can be significant at high currents, but in oxidizing atmospheres there may be less material loss than with thermionic cathodes. This type of cathode is used with plasma generators at current levels between 50 A and 10 kA and with most circuit-interruption devices.

Between the cathode surface and the arc column there is a region in which no LTE conditions exist. Immediately in front of the surface is the cathode sheath region in which over an extent of one to ten Debye lengths (i.e., the distance over which a positive charge is shielded by an electron cloud, typically a few micrometers) strong space charges exist, leading to high potential gradients. These potential gradients accelerate the ions toward the cathode surface and the electrons toward the arc column. Between the sheath region and the arc column is a thermalization region or ionization region in which the directed velocity of the electrons is randomized through collisions. While the sheath region can be described by using relations for a collisionless plasma, the conditions in the thermalization region are largely determined by diffusion fluxes. The extent of this region can be up to several millimeters; beyond this distance, LTE conditions are approached and the arc column region is reached.

1.1.2 Anode Region Since the current is mostly carried by electrons, the anode is a passive collector of the electrons and does not need to provide charge carriers. However, the arc attachment is associated with high heat fluxes, requiring cooling of the anode surface for steady-state operation. The energy transfer to the anode strongly depends on the fluid dynamic configuration of the arc. In general, heat losses from the arc to a cooled surface result in a constriction of the arc and higher electric fields and current densities that are needed to pass the current through the cooled region. This configuration is associated with high heat fluxes to the anode surface and a potential increase from the arc column to the anode (constricted attachment). If, however, a superimposed flow compresses the thermal boundary layer between the arc column and the anode leading to steep temperature and electron density gradients over a larger fluid-dynamic stagnation area, the electron-diffusion fluxes resulting from the density gradients can carry the current, and no constriction or increase in potential gradient is needed (Dinulescu and Pfender, 1980; Sanders *et al.*, 1982; Leveroni *et al.*, 1991). Plate 1 illustrates the two different situations. It should be pointed out that such an effect where blowing the hot plasma toward a surface reduces the heat flux to the surface is one of the counterintuitive features associated with electric arcs.

1.1.3 Arc Column The arc column is the region between the cathode and the anode regions and is inherently unstable. The interaction of the arc current with its own magnetic field leads to a magnification of any force interrupting the current path (magnetohydrodynamic or MHD instabilities). The consequence is that any arc operating in steady state requires stabilization. Any stabilization method is based on the fact that the thermodynamically preferred operating condition for the arc is that for which the arc voltage is a minimum (Steenbeck minimum principle) (Peters, 1956). Consequently, stabilization involves creating a condition where the arc voltage increases before the instabilities lead to arc interruption. Unstable arcs in a high-voltage circuit will of course reignite, but a steady-state condition is not possible.

The most frequently used method to stabilize the column is thermal stabilization, where any excursion of the arc from a straight path leads to strong cooling of the column and an associated voltage increase. Thermal stabilization can be obtained

1. through wall stabilization, where conduction heat flux from the arc column to a cooled wall increases strongly for an asymmetric position of the column, or
2. through convective stabilization, where a cold gas flow surrounds the arc and an asymmetric position of the column requires more gas heating and higher voltages.

A variant of this convective stabilization is the swirl stabilization where the plasma gas is introduced into the arcing device with a tangential velocity. The resulting vortex introduces centrifugal forces acting more strongly on the higher-density cold gases, creating a cold gas layer surrounding the arc. A natural stabilization mechanism is present with short arcs through the flow induced by the cathode jet action (for currents above about 50 A). A constricted cathode attachment leads to a current-density gradient in the direction from the column to the cathode, and the interaction of this current-density gradient with the self magnetic field of the arc current results in a strong plasma flow away from the cathode (Maecker, 1955). The description of the arc column by various modeling approaches is discussed in Sec. 1.4.2.

1.2 Radio-Frequency Induction Plasmas

In radio-frequency induction (RFI) devices, a high-frequency current (typically 450 kHz to 13.5 MHz) flows through a coil surrounding the plasma vessel. The current variation with time induces a voltage that in turn can sustain a current flow in a ring-shaped path through the plasma, thus generating the plasma through thermal ionization processes (see Fig. 3). As with the arc, discharge initiation requires generation of charge carriers by external means to initiate current flow, and high-frequency sparks or coronas may serve this purpose. RFI discharges do not require electrodes, but a cooled cylindrical wall surrounding the plasma is needed for discharge stabilization. The gas to be heated flows axially through the ring-shaped discharge region. Nonuniformity of the induced electric fields and skin effect lead to nonuniform heating of the plasma, resulting in a plasma region in which the hottest region can assume the shape of a doughnut. However, the temperature and density gradients are considerably smaller than those encountered in arc discharges, and the plasma volume is considerably larger for comparable power levels.

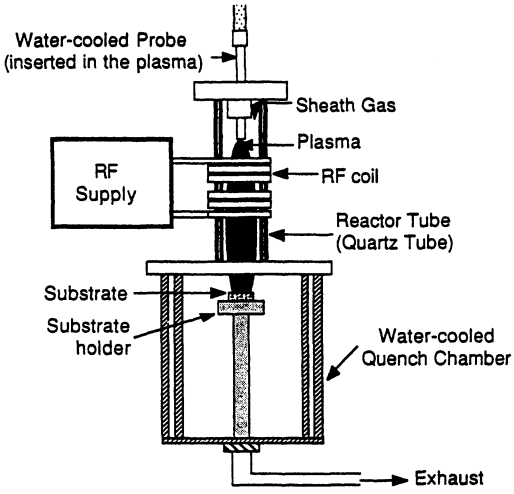
Water-cooled Probe (inserted in the plasma)
Sheath Gas
Plasma
RF Supply
RF coil
Reactor Tube (Quartz Tube)
Substrate
Substrate holder
Water-cooled Quench Chamber
Exhaust

FIG. 3. Schematic of a high-frequency induction-plasma torch.

1.3 Alternative Thermal-Plasma Generation Methods

There are several other methods with which thermal plasmas can be generated, but there are no thermal-plasma devices making use of them. They should be mentioned here only for the sake of completion. Rapid compression of a gas in a shock wave can lead to sufficient heating that a plasma is generated. One example for this occurrence is the re-entry of a spacecraft into the Earth's atmosphere. The shock wave generates a plasma surrounding much of the spacecraft, leading to a communication blackout and ablation of protective heat shields. Shock-wave–generated laboratory plasmas have been used for the measurement of plasma properties; however, the inherently transient nature of shock-wave–produced plasmas raises questions about how close to local thermal equilibrium these plasmas are.

Interaction of high-power laser or particle beams with strongly absorbing solid or liquid matter also can generate thermal plasmas. These effects have been investigated in connection with nuclear fusion research and with weapons research. Laser treatment of materials can lead to the evaporation and ionization of some of the material to be treated, leading to an enhancement or a reduction of the desired effects. Again, these effects are transient in nature, and how close the generated plasma is to LTE must be investigated from case to case.

1.4 Thermal-Plasma Characterization

The primary variables that characterize thermal plasmas are the temperature, composition, and velocity. If LTE exists, composition and density are unique functions of the temperature. However, wherever the thermal plasma interacts with solid or liquid matter—e.g., for materials processing—there are nonequilibrium regions between the thermal plasma and the material to be processed, and characterization of these regions is usually of greater interest than those in which LTE prevails. There are many methods available for characterizing well-behaved simple laboratory plasmas through description by models and by diagnostic experiments. It is more difficult to find descriptions of thermal plasmas in actual devices because there are usually three-dimensional effects, complex gas mixtures, variations with time, and nonequilibrium effects present. An exception to this rule is provided by discharge lamps, which have been modeled and diagnosed fairly accurately for quite some time.

1.4.1 Thermal-Plasma Diagnostics

1.4.1.1 Plasma Temperature. The most straightforward way to determine plasma temperatures is absolute line-intensity measurement in emission spectroscopy, since for LTE conditions a unique relationship exists between the line intensity and the temperature at a given pressure (see, for example, Lochte-Holtgreven, 1968). With a calibrated optical system consisting of an imaging system, a spectrometer, and a detection system, the intensity can be measured and the temperature calculated. One difficulty is that the measurement usually consists of a signal obtained from integration over a path along which the temperature varies because the plasma is nonuniform. Knowledge of the intensity variation along this path is needed for temperature determination, which requires multiple measurements. For a cylindrically symmetric plasma, these measurement requirements can be reduced to the measurement of the lateral intensity distribution and transforming it to the radial distribution by solving Abel's integral equation. Other methods using emission spectroscopy include the relative line-intensity method, where the intensity ratio of two or more lines can be used to derive the temperature. Line-emission spectroscopy is the preferred method for temperature determination inside an arc plasma.

An intrusive method but independent of the plasma geometry is offered by the enthalpy probe. In this calorimetric method, a small-diameter tubular probe cooled by high-pressure cooling water is inserted into the plasma stream, and a measured amount of the plasma gas is sucked into the channel and cooled by the cooling-water flow. Knowledge of the gas composition then allows the temperature to be determined (Grey and Jacobs, 1964; Capetti and Pfender, 1989).

Use of Langmuir probes in thermal plasmas is difficult, but they have been successful in determining electron temperatures close to a cooled surface in which the probe is imbedded. A method recently developed for probing thermal plasmas involves pulsing the probe with a potential positive with respect to the plasma potential and recording the probe current during the pulse time. A subsequent negative pulse attracts ions, which bombard the probe surface and clean it from contamination (Leveroni *et al.*, 1991). These measurements are particularly of interest in regions close to a surface where deviations from LTE, especially elevated electron temperatures, are to be expected.

Recent developments in high-power laser diagnostics have opened the door to temperature measurements that are truly local and that do not require the assumption of LTE. One of these methods consists of measuring the line shape of the laser light scattered in a small plasma volume and comparing it with a theoretical line shape (Snyder *et al.*, 1994). Heavy-particle and electron temperatures can be measured independently. Another laser-spectroscopic method for temperature determination, laser-induced fluorescence (LIF), consists of selectively exciting specific upper energy states with the radiation from a tunable laser, and measuring the intensity of the de-excitation emission. The intensities can be related to the lower-state population densities and therefore can yield a value for the temperature.

1.4.1.2 Electron Densities. Measuring electron densities is of interest when deviations from equilibrium values are expected. Since the dominant recombination process in thermal plasmas is three-body recombination, recombination can be relatively slow, and fast-moving or rapidly cooling plasma flows may exhibit higher electron densities than the temperatures would predict. Two emission-spectroscopic methods are used for determination of electron temperatures: the absolute continuum intensity and the Stark line-broadening method. The continuum radiation from a plasma is generated by free electrons in the charge field of ions or other electrons (bremsstrahlung), or by free electrons recombining with ions (free-bound radiation). An expression for the intensity of the continuum radiation has been derived for hydrogenic atoms with correction factors for argon, and the electron density can be derived from it (Griem, 1964). The other emission-spectroscopic method is based on line broadening by the Stark effect, i.e., the removal of degeneracies of excited-state energy levels due to the electric field from free electrons. For hydrogenic atoms, this line broadening can be calculated as a function of the electron density, and comparison of the theoretical line shape with the measured one yields the electron density.

Also laser scattering or microwave interferometry can be used for electron-density determination, and there are several laser-spectroscopic methods available for probing the composition of plasmas with several components. Most of these methods are applicable only for recombining plasma flows, because in the plasma-generation region there is usually too much background radiation from the high-temperature plasma. Also the enthalpy probe can be used to measure the composition of a recombining plasma if the gas sample is channeled to a mass spectrometer. Obviously, only stable species can be measured this way.

1.4.1.3 Plasma Velocity. The plasma velocity is usually of interest in the region outside the plasma-generation region, i.e., in the plasma jet (except for circuit-breaking devices). One of the most prominent methods for measuring velocities is laser Doppler anemometry, which can be used in a variety of different arrangements. In one of these methods, the split beams of a laser generate an interference pattern in the space to be probed, and submicron seed particles in the plasma flow scatter the laser light with a modulation imposed by the interference pattern. With knowledge of the geometry of the pattern, the particle velocity can be derived from the scatter signal.

Also the enthalpy probe can be used to measure the plasma velocity. If the temperature at a certain location in the plasma has been determined, measurement of the stagnation pressure allows derivation of the velocity. A recently developed method uses the propagation of a disturbance of the plasma to determine the plasma velocity. Such a disturbance may be caused by a short-term instability of the arc, which is carried with the plasma flow and can be detected as an intensity fluctuation propagating downstream (Coudert *et al.*, 1993).

There are numerous diagnostics that result in qualitative descriptions of thermal plasmas. Examples are schlieren photography for flow studies and high-speed videography for detection of instabilities. Computerized image analysis of these images can yield some approximate values for the parameters of interest, such as the geometry of the high-temperature core of a discharge. With the continued advances in video technology, one will probably see an increasing amount of these techniques in combination with computer analysis being applied for quantitative characterization of thermal plasmas.

1.4.2 Plasma Characterization by Modeling Of the different regions of the arc, the arc column has been fairly well described by two-dimensional models. In general, the conservation equations are used, with some approximations usually made for boundary-layer flow. Additional terms made necessary by the presence of plasmas are the magnetically induced forces in the momentum equations, and the Joule heating term $\mathbf{j} \cdot \mathbf{E}$ in the energy equation. The equations can be rather complex nonlinear partial differential equations. The nonlinearity appears because of the strong dependence of the thermodynamic and transport properties on temperature and composition, and numerical solutions of the conservation equations are required even for the simplest case of a wall-stabilized arc column where there are no axial velocity and temperature gradients (fully developed flow condition). In this case the momentum and energy conservation equations are decoupled, and the energy equation is reduced to the well-known Elenbaas–Heller equation (Pfender, 1978)

$$\frac{1}{r}\frac{d}{dr}\left(r\kappa\frac{dT}{dr}\right) + \sigma E^2 - P_R = 0, \qquad (2)$$

where κ and σ are the thermal and electrical conductivities of the plasma, respectively, E is the electrical field strength, and P_R is the power loss by radiation per unit volume. Presence of the Joule heating term requires additional equations: Ohm's law providing the relation between current density and electric field, and current conservation for dc arcs or Maxwell equations for rf plasmas. Since plasmas are by definition reacting gases (ionization reactions are always present), the perfect-gas law has to be supplemented by equations giving the composition. For LTE, minimization of Gibbs's free energy or appropriate mass-action laws are used, like the Saha equation for determining electron densities:

$$\frac{n_e n_i}{n} = \frac{2Z_i}{Z_0}\frac{(2\pi m_e kT)^{3/2}}{h^3}\exp\left(-\frac{E_i}{kT}\right), \qquad (3)$$

where n_e, n_i, and n are the densities of electrons, ions, and neutral particles, respectively, Z_i and Z_0 are the partition functions for the ions and the neutral species, k and h are Boltzmann's constant and Planck's constant, and E_i is the ionization energy. In case of nonequilibrium, in particular in recombining plasmas, a set of rate equations for the possible chemical reactions has to be used.

Plasma jets and electrode regions are more difficult to model. Plasma jets are in most plasma devices turbulent, and their behavior is to a large extent determined by entrainment of cold gas into the hot plasma jet. Two-fluid models have been recently successfully used for describing this feature (Huang *et al.*, 1995). The electrode regions are associated with strong deviations from LTE, and different approaches have to be used. In particular for the condition where the anode consists of a ring-shaped nozzle surrounding the plasma jet, the attachment is usually unstable, resulting in a constant motion of the arc root, and a three-dimensional description is necessary (Wutzke *et al.*, 1967; Paik *et al.*, 1993). In case of an arc anode attachment in a fluid-dynamic stagnation region, progress has been made in describing the different attachment modes—diffuse attachment with current flow driven by diffusion and constricted attachment with a positive anode fall (Jenista, 1995; Zhu *et al.*, 1992). Similarly for the arc cathode attachment, models exist describing the attachment with different simplifying assumptions (Hsu and Pfender, 1983; Morrow and Lowke, 1993; Zhou and Heberlein, 1994), but no model has been formulated that is generally valid for the cathode region under all conditions.

2. PLASMA TORCHES AND ARC-GAS HEATERS

2.1 Plasma-Generator Device Purpose and Categorization

The processing of materials requires the controlled change of the physical and/or chemical characteristics of a substance, such as melting or evaporation; reduction or formation of oxides, carbides, or nitrides; or transformation of materials from a bulk material to a powder or from a powder to a coating.

The principal advantages offered by the use of plasmas for the processing of materials are the following:

1. The use of electrical energy for the generation of thermal energy makes the heating of the medium independent of its oxygen potential or its combustion energy—any material can become a plasma; it also gives enhanced control over the environment.
2. The high energy densities associated with a plasma allow the processing of high-temperature materials that are stable under lower flame temperatures, and they make high quench rates possible permitting the generation of metastable compounds.

The advantages of plasmas for materials processing were recognized relatively early, and soon after electricity became widely available, arc furnaces were introduced for bulk melting of metals by W. von Siemens in 1878. In the early 1900s, a thermal-plasma reactor was developed by Birkeland in Norway for the bulk chemical synthesis of calcium nitrate at power levels extending to 250 MW. While this process has been surpassed by low-temperature nitrogen-fixation processes, the plasma process for acetylene synthesis from natural gas developed by Chemische Werke Hüls in Germany during the

1920s and 1930s has been commercial since 1938 and is presently producing acetylene with an installed capacity of 180 MW. Systematic research on stabilizing the plasma in a plasma-generation device during the 1920s and again during the 1950s has resulted in the development of numerous plasma heater designs for a number of specific applications.

Thermal-plasma devices for materials processing are generally referred to as plasma generators, plasma torches, plasma heaters or plasma furnaces, or plasma reactors. While thermal-plasma generators refers to all devices in which a thermal plasma is produced for external or internal use, the distinction between the definitions for the other terms are somewhat less clear. A *plasma torch* (sometimes also called *arc-gas heater* or *induction gas heater*) is in general a device in which a plasma is generated to heat a gas flow, which exits the device in the form of a plasma jet that is then used for various plasma processes. In this device, the function of the electrodes is only to serve as termination of the arc, and the device is frequently called *nontransferred plasma torch.* In a plasma heater or a *plasma furnace,* the flow of the plasma gas is minimized and the plasma, usually an electric arc, heats the material to be processed directly. If one of the electrodes is also the material to be processed, the device is called a *transferred-arc plasma heater.* The *plasma reactor* is the device in which the plasma promotes physical and/or chemical reactions. The plasma generator can be part of the plasma reactor, always in the case of a plasma furnace or transferred-arc plasma device, or it can be separate as in the case of some plasma torches. A further categorization method uses the cathode emission mechanism for arc plasma generators, i.e., hot-cathode (thermionic emission) generators or cold-cathode generators. Lastly, also the arc-stabilization mechanism may serve as a categorization method, such as flow- or swirl-stabilized generators or wall-stabilized plasma heaters. Each of these devices can be designed for a different range of current and voltage levels, as is shown in Fig. 4.

Obviously, there are a number of different design options for plasma-device designs, which have to be matched with the specific materials-processing requirement, such as power level and enthalpy level, oxidizing or reducing environment, amount of off-gas permitted, and length of operation between service periods. In the following, a few specific applications are discussed in detail to illuminate the matching of these requirements.

2.2 Plasma-Spray Torches

Plasma spraying is a coating process in which metallic or ceramic particles are introduced into a high-velocity plasma jet, where they are melted and accelerated toward a

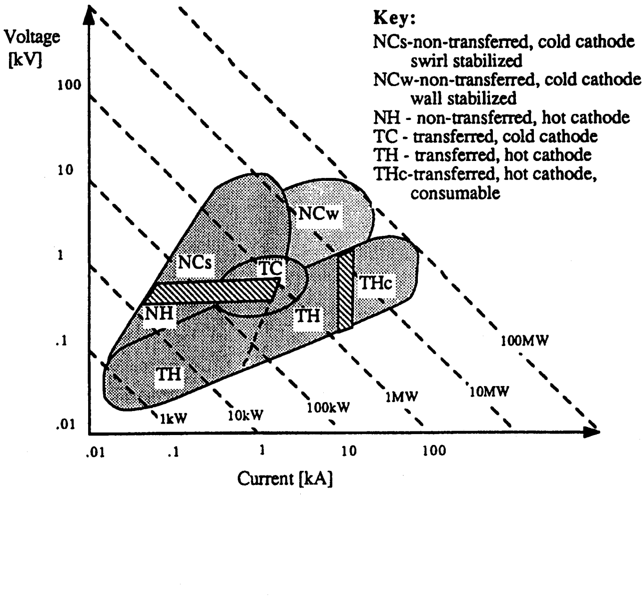

FIG. 4. Schematic of the voltage, current, and power ratings of different types of plasma torches (Heberlein, 1992).

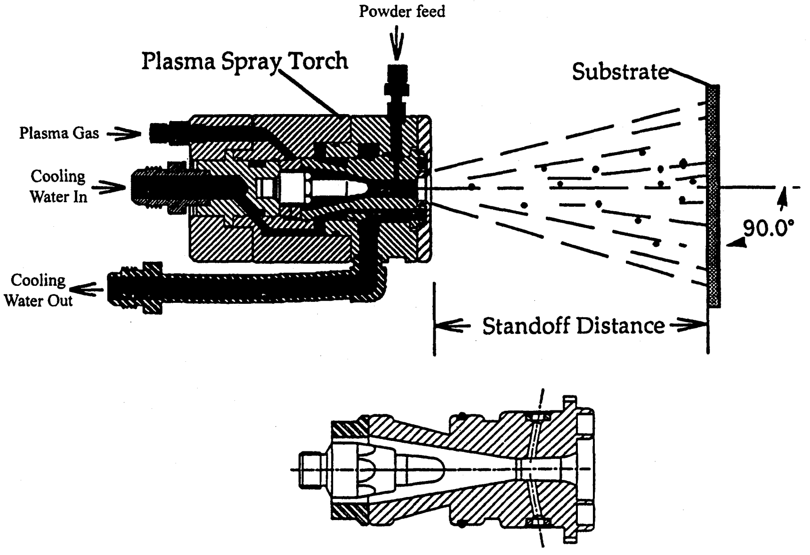

FIG. 5. Plasma-spray torch schematic (Miller SG 100, with kind permission of Praxair Thermal Spray Systems).

substrate on which the desired coating is formed. This technology has found widespread applications in the aerospace industry; e.g., the majority of the parts of a jet engine are coated to protect them against corrosion and excessive temperatures. But also in the medical implant industry and in the petrochemical industry, plasma spraying is used, and the technology is expanding into the automotive industry. Plasma spraying can be performed either in an open atmosphere (atmospheric-pressure plasma spraying or APS) or in a low-pressure chamber of 50 to 200 Torr where the plasma is expanded through a nozzle (low-pressure plasma spraying or vacuum plasma spraying, VPS). Figure 5 shows the cross section of a plasma-spray torch, with a hot cathode, usually of thoriated tungsten, and a cylindrical, water-cooled anode. The powder is injected downstream of the anode attachment point of the arc, i.e., the configuration is a nontransferred plasma heater. The arc is convection stabilized by the arc gas flow, which is frequently introduced with a swirl component at the root of the cathode rod, thus enhancing the stabilization. Typical operating conditions are 350 to 1200 A arc current, 25 to 80 V arc voltage, and 10 to 80 kW arc power. The most frequently used plasma gases are mixtures of argon with helium or hydrogen, or nitrogen. Higher-power torches are available, but their use is not as widespread. In atmospheric-pressure plasma spraying the jet is usually highly turbulent and cold air is entrained into the plasma, resulting in quite nonuniform conditions. Figure 6 shows a high-speed picture of a plasma-spray jet with particles. In this 50-ns exposure the particles have been illuminated by a laser strobe, and the image has been captured with a CCD array with an electronic shutter (Malmberg and Heberlein, 1993).

Because of the commercial importance of the plasma-spray process, the plasma-spray

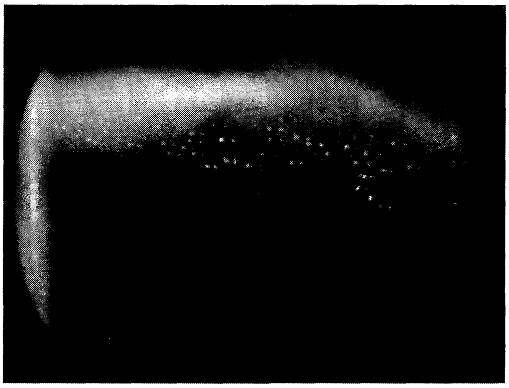

FIG. 6. High-speed videograph of a plasma-spray jet during spraying operation (exposure time 50 ns) (Malmberg and Heberlein, 1993).

jet has been well characterized through various diagnostics and modeling. Figure 7 shows the isotherms and isovelocity contours of a typical plasma-spray jet (Vardelle *et al.*, 1983); however, the values shown are time-averaged values, and significant variations may occur. Near the nozzle exit, peak velocities of 600 to 1000 m/s may be encountered in atmospheric-pressure plasma-spray jets (significantly higher in VPS), and temperatures of 12 000 to 15 000 K. The quality of the coating is predominantly determined by the particle temperature and particle velocity at the moment of impact on the substrate and by the substrate temperature. There have been several methods developed to measure particle velocity and temperature, the most elegant one developed by Fincke *et al.* (1990). In this method, two beams of a laser are used to measure particle velocity using laser anemometry and average size, and at the same time the radiation from the particle at two different wavelengths is measured and used to determine the particle temperature (two-color pyrometry).

The behavior of the arc inside the anode channel is less well characterized, partially because of the difficulty of gaining optical access and partially because of its usually unstable behavior. There is a constant motion of the arc anode attachment spot, which is desirable because it reduces anode erosion. This motion of the attachment spot is caused by the arc restrike phenomenon, in which the cold gas flow surrounding the arc drives the anode attachment downstream until an upstream restrike between the arc column and the anode wall occurs (Wutzke *et al.*, 1967), and by the plasma gas swirl. However, this motion induces fluid-dynamic instabilities in the plasma jet, causing nonuniform conditions. Some recent research efforts have concentrated on the characterization of these arc instabilities and the influences of operating conditions and different anode designs; however, no complete model exists that can describe the three-dimensional, time-dependent nature of the arc in a plasma-spray torch satisfactorily.

There are several variants of the plasma-spray process and of the plasma-spray torch for other applications. In wire-arc spraying, the coating material is introduced in the form of wires, melted by an arc struck between the two wire tips. A high-velocity stream of an atomizing gas blows the molten material in the form of fine droplets toward the substrate (see Fig. 8). In rf plasma spraying, the powder is injected into the plasma through a water-cooled injection probe in the center of the rf plasma torch. rf plasma spraying offers longer residence times of the particles in a region of moderately high temperatures, assuring complete melting of the particles. This feature may in some applications be more important than achieving the high particle velocities that can be obtained in dc plasma spraying.

Torches resembling plasma torches are also used for plasma cutting or some types of arc welding, although in these applications the arc is transferred to the workpiece. In a completely different type of application, the arc-jet thrusters used for space propulsion of

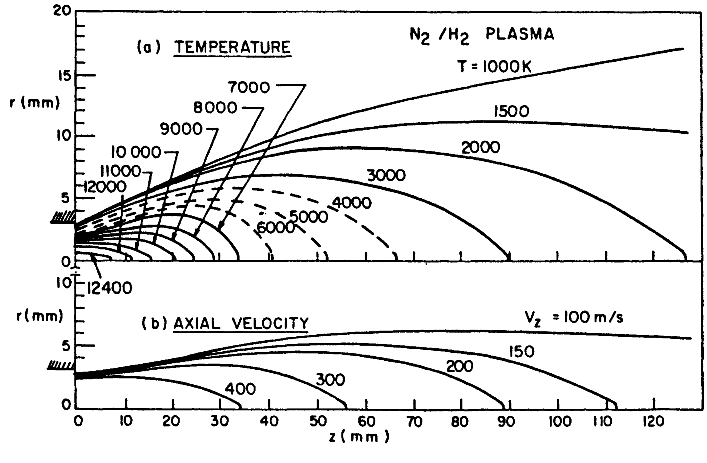

FIG. 7. Steady-state temperature and velocity contours in the jet of a plasma-spray torch (Vardelle *et al.*, 1983). The marks on the ordinate indicate the location of the nozzle wall. (With kind permission of the authors.)

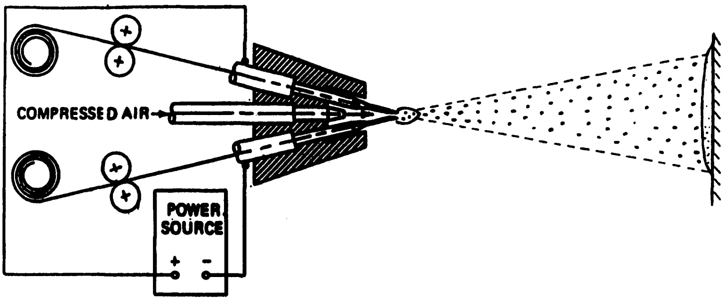

FIG. 8. Schematic of a wire-arc spray system.

satellites were derived from the same type of torch that served as the origin of the plasma-spray torch.

2.3 Plasma Torches for Gas Heating and Chemical Processing

The principal requirements for arc-gas heaters for bulk chemical processing are the efficient heating of selected gases to specific bulk enthalpy levels, and high reliability and low maintenance. The power levels used are between 0.5 and 8 MW, with most of the applications requiring 1 to 3 MW or multiple heaters at these power levels. The majority of these torches operate in the nontransferred mode, although transferred-arc operation is possible if gas heating is combined with processing of a solid material serving as electrode. These torches have usually high gas heating efficiencies of 70 to 90%, but the average gas temperatures are usually relatively low (less than 7000 K).

The majority of the gas heater designs are based on the concept developed by Hüls for the acetylene generation process. Figure 9 shows, as an example for such an arc-heating device, a plasma torch developed by Westinghouse (Wolf and Fey, 1971; Heberlein *et al.*, 1989). It consists of two tubular electrodes arranged coaxially beside each other. The electrodes are water cooled, and the arc is struck between the inside surface of one electrode and the inside surface of the other. The process gas is introduced between the electrodes with a strong swirl, though part of the process gas can be introduced axially. A current passing through a coil surrounding the electrode tubes generates a magnetic field, which leads to a rotation of the arc electrode attachment, thus increasing electrode life and improving arc heating effi-

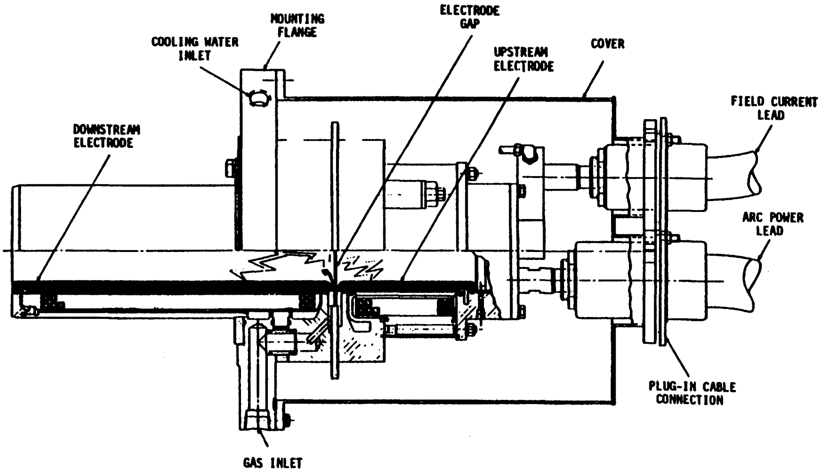

FIG. 9. Schematic of a Westinghouse arc heater. The lower half is shown as cross section. (With kind permission of Westinghouse Electric Corp.)

ciency and uniformity of the process-gas enthalpy levels.

This torch design is being used in applications ranging from air heating for metal-recovery shaft furnaces to providing process heat in a reactor for processing hazardous liquid wastes. Similar torches have been used for the commercial production of titanium dioxide pigment from arc-heated oxygen and titanium tetrachloride (George and Hare, 1983) and for the recovery of aluminum metal from the dross (Drouet, 1993).

2.4 Transferred-Arc Devices

The highest heat fluxes from an electric arc are to its electrodes. The reason for this fact is that in addition to the heat transfer by conduction, convection, and radiation, the energy transfer associated with current flow is present, i.e., for the anode, the kinetic energy of the electrons as well as the electron condensation energy (work function). In addition, usually the steepest temperature gradients are in front of the electrodes, maximizing the conduction heat flux. Heat fluxes in excess of 10 kW/cm^2 can be obtained. Figure 10 shows a typical temperature distribution in a transferred-arc device (Young *et al.*, 1983). Although this temperature field was measured at relatively low currents (200 A), the major change with increasing current is that the dimensions will increase with little change in peak temperatures.

Transferred-arc furnaces have been used extensively in the metallurgical industry, because the high heat fluxes associated with an arc anode attachment can lead to rapid melting of the anode material. As a matter of fact, the first commercial thermal-plasma device used for materials processing was the arc furnace developed by W. v. Siemens in 1878 for the melting of steel. Figure 11 shows a schematic of this device (Reid and Camacho, 1987). A graphite rod serves as a thermionic cathode while the molten metal serves as the anode. Carbon ablation from the consumable cathode can provide a reducing atmosphere. The majority of arc furnaces have been operated with ac; in particular, three-phase arc furnaces with three

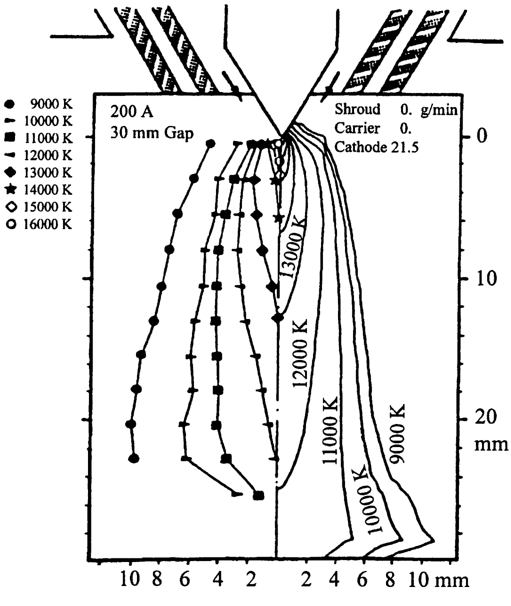

FIG. 10. Temperature distributions in a transferred arc geometry, measured (left) and calculated (right) (Young *et al.*, 1983, with kind permission of the author).

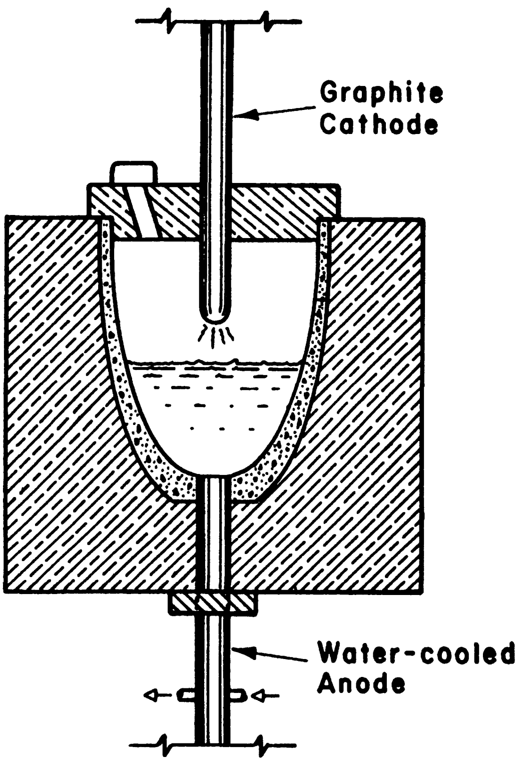

FIG. 11. Schematic of the original Siemens arc furnace (Reid and Camacho, 1987, with kind permission of the author).

electrodes each connected to one phase and the melt serving as ground are widely used. However, more efficient conversion of ac to dc has recently led to preference of dc arc furnaces because ac arc furnaces provide an inductive load with varying power dissipation resulting in very high noise levels and requiring power-factor correction between the load and the grid. In addition, dc arc furnaces allow improved process control and can reduce the graphite loss. One of the most advanced dc transferred-arc installations is the 45-MW furnace in Middleborough, South Africa, for production of ferrochrome. A schematic of this furnace developed by Mintek, S. A., is shown in Fig. 12 (Barcza *et al.*, 1989). The hollow carbon cathode allows introduction of fine raw materials through the cathode into the arc, reducing carryover of the fines with the off-gas.

While the high-power installations usually rely on consumable graphite cathodes, lower-power plasma furnaces operating at current levels below 10 kA use mostly nonconsumable cathodes usually made of tungsten with the addition of a low–work-function material for improved electron emission. Plasma heaters of this type are used increasingly for tundish heating during the continuous-casting process of steel. In this application, the molten metal flows into the tundish—a giant bath tub—from where it is cast. Because of the heat losses in the tundish, reheating of the metal at the location of the exit spout is necessary, if one wants to avoid filling the tundish with a highly superheated metal. Plasma heaters have the advantage that they can be turned on quickly when the temperature drops below a preset value and can provide a sufficient amount of heat with little gas flow. This type of heater operates typically in the power range of 1 to 3 MW, with currents rarely exceeding 3 kA.

On the lower end of transferred-arc power levels are plasma cutting torches. A schematic of such a torch is shown in Fig. 13. The arc is strongly constricted for maximizing current densities, and the metal to be cut serves as the anode. Oxygen is the preferred plasma gas, because the chemical reaction between the oxygen plasma and the metal increases the heat fluxes and therefore the cutting speed. Hafnium cathodes are used to avoid rapid cathode erosion. Torches of this type can be found in almost any sheet-metal shop, and their sizes range from 100 W to a few kilowatts.

For a few applications, transferred-arc plasma torches with cold cathodes have been developed. One example is the torch developed by Retech shown in Fig. 14 (Eschenbach, 1989). This torch, originally widely

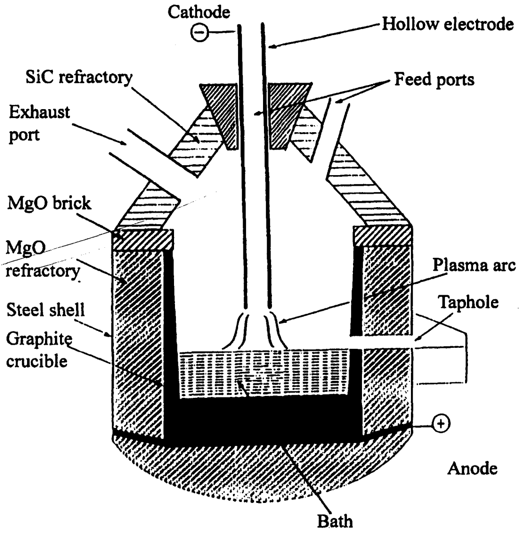

FIG. 12. Schematic of the modern plasma furnace designed by Mintek, S.A. (Barcza *et al.*, 1989, with kind permission of the author).

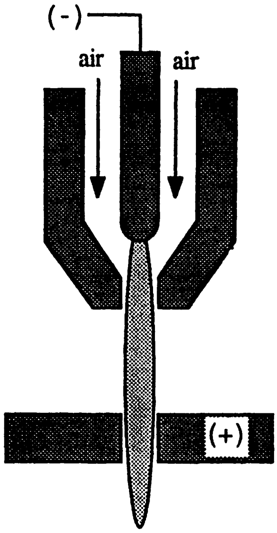

FIG. 13. Schematic of a cutting torch.

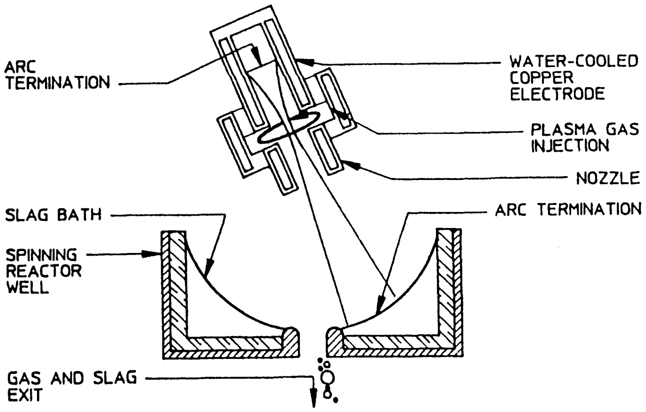
ARC TERMINATION
WATER-COOLED COPPER ELECTRODE
PLASMA GAS INJECTION
NOZZLE
ARC TERMINATION
SLAG BATH
SPINNING REACTOR WELL
GAS AND SLAG EXIT

FIG. 14. Schematic of the Retech transferred-arc plasma torch (Eschenbach, 1989, with kind permission of the author).

used for titanium remelting, has more recently been applied in waste-processing reactors. In this application, the arc is struck between the upper cylindrical electrode and the waste material contained in a rotating tub. The rotation, at approximately 50 rotations per minute, allows continuous exposure to the plasma heat until a slowing down of the rotation results in flow of the molten waste material into a crucible where a nonleachable slag is formed. This torch can operate at power levels of up to 2 MW.

2.5 Plasma Torches for Space Applications

The space program has led to the development of two types of plasma torches:

1. high-power segmented plasma heaters for testing of materials and
2. arcjet thrusters for satellite propulsion.

When a spacecraft returns into the Earth's atmosphere (or ventures into the atmosphere of another planet), the sudden change in density of the surroundings coupled with the very high velocities of the space vehicle generates a plasma surrounding the vehicle and results in extremely high heat fluxes to parts of the vehicle. To simulate these conditions and study the behavior of potential protective materials, high-enthalpy plasma facilities have been developed. Most of these high-power plasma devices are based on segmented-arc heater designs, where insulated water-cooled discs separate anode and cathode and stabilize arcs up to 3 m in length. Figure 15 shows one of these designs (Winovich and Carlson, 1980). In this design, the electrodes consist of several water-cooled rings with separate current attachments allowing for parallel current flow and reduced heat flux to each electrode section. The individual constrictor discs are arranged in modules of 30 segments, each individually cooled with high-pressure deionized water, and 13 modules comprise the entire constrictor.

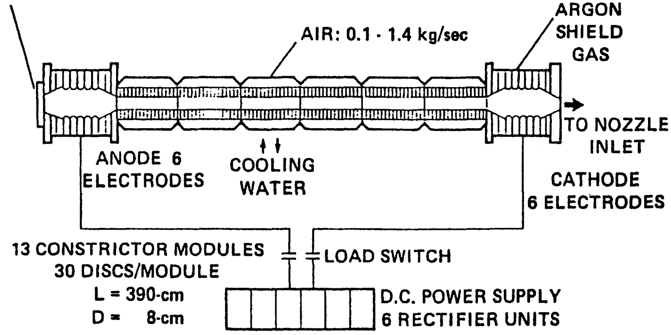
ARGON START/SHIELD GAS
AIR: 0.1 - 1.4 kg/sec
ARGON SHIELD GAS
TO NOZZLE INLET
ANODE 6 ELECTRODES
COOLING WATER
CATHODE 6 ELECTRODES
13 CONSTRICTOR MODULES
30 DISCS/MODULE
L = 390-cm
D = 8-cm
LOAD SWITCH
D.C. POWER SUPPLY
6 RECTIFIER UNITS

FIG. 15. Schematic of the NASA high-power arc heater for re-entry simulation (reprinted from Winovich and Carlson, 1980, with kind permission of Elsevier Science, Amsterdam, The Netherlands).

Typical arc currents range up to 5.4 kA, and arc voltages up to 11 kV can be obtained. The downstream electrode is followed by a nozzle in which the high-enthalpy gas stream is expanded into a low-temperature chamber in which the test specimen is mounted (e.g., a protective tile for the space shuttle). Part of the high gas temperature is transformed into hypersonic gas velocities, thus creating the conditions simulating those the spacecraft encounters during re-entry. It needs to be mentioned that gas heating efficiencies are rather low, so that this type of torch has not found applications in materials processing.

The electric-propulsion application of thermal-plasma devices makes use of the fact that electric energy can produce high-velocity gas flows with much less mass compared to combustion-based propulsion devices since no reactions between different gases are required for producing the heat. The life of a satellite is in part dependent on how long its position can be controlled by periodic propulsion thrusts. These propulsion bursts require low thrusts and last for relatively short times. Chemical propulsion requires transport of fuel and oxidant, while electric propulsion requires only one gas and the power-conditioning equipment transforming the output of the solar arrays into a high-voltage, low-current source. Arcjet thrusters are one of the electric propulsion devices having considerably better specific impulse than conventional chemical rockets, although much lower thrusts because of lower mass flows. These arcjet thrusters were originally developed in the 1960s; however, only recently have they been used in spacecraft partially because of the advances in solar-array technology. Figure 16 shows the schematic of such an arcjet thruster (Sankovic *et al.*, 1991). The design is quite similar to that of the plasma-spray torch, consisting of a thermionic rod cathode and a nozzle anode. One of the differences is that the anode nozzle consists of a tungsten piece and is entirely radiation cooled since no cooling water is available. The gas, hydrazine or hydrogen, enters the device at the root of the cathode rod, is heated by the arc in a cylindrical section of the anode, and then passes through the nozzle where it is expanded to supersonic velocities. Plate 2 shows one arcjet thruster in operation in a laboratory environment.

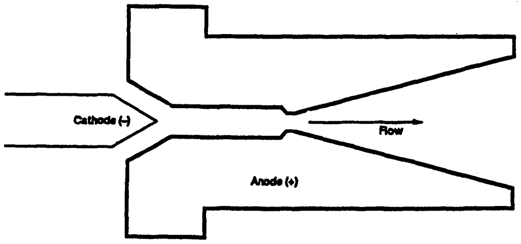

FIG. 16. Schematic of an arcjet thruster for electric space propulsion.

Recently, arcjet thrusters as well as other electric propulsion devices using nonthermal plasmas have been developed for higher-power applications for interplanetary space travel, and they are being adapted for other applications such as thermal-plasma chemical vapor deposition. As an interesting observation one might add that these arcjet thrusters and the plasma-spray torches had a common root, then were adapted by different user communities for different applications, and now start to approach each other again with the development of low-pressure plasma spraying and the use of arcjet thrusters for thermal-plasma CVD. In particular in the modeling of these devices there is a considerable overlap, and input from both user communities can be combined advantageously (George *et al.*, 1996).

3. PLASMA LIGHT SOURCES

3.1 History and Categorization of Discharge Lamps

The importance of plasmas for lighting can be illustrated by a few figures: Approximately 25% of the electrical energy consumed in the United States is used for illumination, and an increasing fraction of this amount for discharge lamps. Practically all street lighting and all lighting of large rooms uses discharge lamps because of their considerably higher efficiencies in converting electricity into light. If we compare the light output of a candle with that of a 40-W fluorescent lamp, we find that the lamp provides 250 times the light for about 1000 times the length of time (Waymouth, 1971). The importance of thermal plasma for lighting is also demonstrated by the fact that discharge lamps are probably the most investigated and most completely characterized of all thermal-plasma devices, and a considerable amount of literature is devoted to this characterization.

The high radiation intensity of the carbon arc not only has stimulated the curiosity of scientists like H. Davy, who is credited with discovering the electric arc, but also has attracted engineers who wanted to utilize this property. Already in the middle of the 19th century carbon arcs were used as light sources. Around the same time, H. Geissler discovered through spectroscopic studies of discharges that different gases emitted light at different characteristic wavelengths, creating the basis for colored discharge lights used for special effects and advertising. In the beginning of the 20th century, P. Cooper-Hewitt experimented with discharges in low-pressure environments containing mercury vapor and discovered that metal-vapor discharges were very efficient light sources. His mercury-vapor lamps were the first commercial discharge lamps, which found widespread use for industrial lighting in the early part of the century. These lamps found successors in fluorescent lamps and in high-pressure mercury lamps. In fluorescent lamps, the mercury discharge takes place in an inert buffer gas (e.g., argon), and by adjustment of the mercury-vapor pressure and the buffer gas pressure, emission of the ultraviolet radiation from the mercury resonance line at 253.7 nm can be optimized to a degree where it contains 60% of the electrical input power (Waymouth, 1971). A phosphor coating on the wall of the discharge tube absorbs this UV radiation and transforms it to light in the visible wavelength region. After the discovery of efficient phosphors in the 1930s and the development of satisfactory processing methods, fluorescent lamps became very widely used. However, the nonequilibrium low-pressure discharges in these fluorescent lamps do not consist of thermal plasmas and will not be further discussed in this article.

Three types of lamps make use of the emission of light from thermal plasmas:

1. high-intensity discharge (HID) lamps,
2. short-arc lamps, and
3. carbon-arc lamps.

The HID lamps are the most widely used illumination sources, and they can again be divided into three different groups: high-pressure mercury lamps, metal-halide lamps, and high-pressure sodium lamps. Each of these lamp types has specific advantages that render them particularly suitable for some specific application.

The performance of lamps is given by three characteristics: the efficacy, the color rendition, and the total power or total light output. The efficacy is the efficiency with which the electrical energy is transformed into light that can be detected by the human eye; i.e., the light output is weighted according to the spectral photopic response characteristics of the eye. The unit for measuring efficacy is lumens (abbreviated lm) per watt, and one watt radiated at 555 nm is equivalent to 680 lm. Accordingly, the highest efficacy would be associated with a lamp radiating only in a narrow line around 555 nm, without any loss of heat. For comparison purposes consider that the efficacy of an incandescent lamp is about 20 lm/W, and that of a fluorescent lamp is about 100 lm/W. Color rendition, on the other hand, determines how much energy is radiated in various wavelength bands allowing the light reflected from various objects to be seen in a specific color. Different tasks can require different color renditions, and specific preferences for various color renditions exist depending on people and environment. Discharge lamps exist for a wide range of color renditions. The total power or the total light output is given in watts or in lumens.

3.2 High-Intensity Discharge Lamps

3.2.1 Some Common Characteristics

The basic configuration of all HID lamps is very similar and shown in Fig. 17. The arc is contained in a sealed arc tube, which is mounted in an outer envelope made of borosilicate glass. Electrical connections are made through a brass base similar to those for incandescent lamps. The outer envelope is evacuated or filled with nitrogen at lower than atmospheric pressure. The envelope is designed to fulfill several functions:

1. to protect the arc tube from contamina-

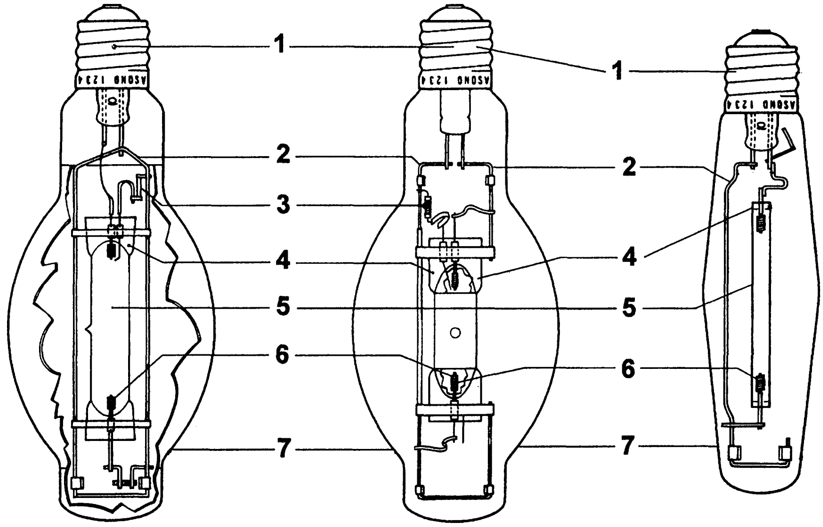

Mercury Vapor Lamp

Metal Halide Lamp

High-Pressure Sodium Lamp

FIG. 17. Construction of high-intensity discharge lamps (with kind permission of North American Philips Lighting Corp.). (1) The brass base, (2) the arc-tube mounting frame, (3) a starting resistor, (4) the arc-tube seal, (5) the arc tube, (6) the double-coiled tungsten electrodes, and (7) the borosilicate glass bulb.

tion, which would influence its operation, and to protect oxidation of the leads;

2. to protect the arc tube from asymmetric cooling during operation, which would change the character of the discharge;
3. to reduce the heat loss from the lamp by absorbing or reflecting infrared radiation from the arc tube; and
4. to absorb the harmful short-wavelength UV radiation emitted by the arc and passing through the arc-tube wall.

The arc tube contains a starting gas, typically argon or, in the case of the high-pressure sodium lamp, xenon, at a pressure of a few Torr, and the metal-containing chemical compound, which is usually in condensed form at room temperature. The choice of arc starting gas is partially determined by the use of the Penning effect: the lowest metastable energy level of a gas is, after the application of a voltage, quickly populated in equilibrium with the electron energies, and transfer of the excitation energy during a collision with an atom of a different gas or vapor with an ionization energy lower than the energy of the metastable energy level leads to efficient ionization and lowering of the breakdown voltage. The energy of the first metastable level of argon is 11.55 eV, and the ionization energy of mercury is 10.43 eV; the first metastable energy level of xenon is at 8.32 eV, and the ionization energy of sodium is 5.14 eV. After arc starting, the arc tube heats up, leading to evaporation of the metal compound accompanied by a significant increase in pressure (100 Torr to 8 atm) and light output. The low ionization energy of most metals results in relatively low power dissipation per unit length of the arc and in relatively low peak temperatures of 5000 to 6000 K.

Most HID lamps operate with alternating current, and the arc extinguishes every half cycle of the 60-Hz supply voltage. A stroboscopic effect or "flicker" is the consequence of this 120-Hz modulation of the light output. However, after initial lamp starting the gas inside the arc tube is sufficiently hot that restriking of the arc during the following half cycle is easily accomplished without special starting aids.

The arc discharge in HID lamps is stabilized by wall stabilization combined with free-convection stabilization. For vertical operation, the arc-heated gas will rise to the top of the arc tube where it will cool down somewhat and return along the arc-tube wall, thus forming free convection cells surrounding the arc column. For operation in other positions, the arc will bow and will be asymmetrically cooled by the proximity to the arc-tube wall resulting in lower efficacies. This effect is not observed with the high-pressure sodium lamp, which has a narrower arc tube and more wall stabilization.

All HID lamps have similar electrodes, shown schematically in Fig. 18. The electrodes consist of a tungsten rod onto which a tungsten coil is wound, which itself is overwound with another tungsten coil. The space between the coils is impregnated with an emission material, a rare-earth or alkaline-earth oxide combination. This emission material reduces the work function for electron emission during the current half cycle when the electrode is a cathode, thus reducing the electrode operating temperature and the material evaporation rate and prolonging lamp life.

Lamp life is usually associated with electrode effects. Evaporation of the emission material and sputtering of the tungsten leads to a darkening of the arc tube with the consequence of higher starting voltages. In particular with the high-pressure sodium lamp, the darkening of the arc-tube walls will increase the operating pressure and voltage to a degree that the voltage will exceed that supplied by the ballast. The lamp will extinguish, cool down, and reignite.

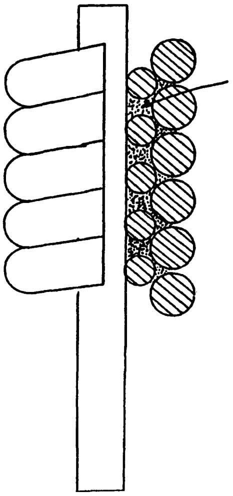

FIG. 18. Electrode design for HID lamp discharges.

Since discharges in lamps have usually falling voltage–current characteristics, i.e., the voltage decreases with increasing current, a stabilizing circuit element has to be introduced, and since most discharge lamps operate on ac, this circuit element is an inductive impedance, which is incorporated into the lamp ballast together with the transformer providing the suitable discharge operating voltage and with the starting circuit. The losses of electrical energy in the ballast have to be taken into account when comparing efficacy values, and they range from 30% for low-wattage HID lamps (50 W or less) to 7% for lamps with power ratings of 1000 W or more.

Although quite sophisticated models have been developed for description of the lamp arcs, in particular with regard to the description of the radiation transport, the Elenbaas–Heller energy equation [Eq. (4) in Sec. 1.4.2] for the fully developed wall-stabilized arc assuming optically thin radiation can give good physical insight into the arc behavior of discharge lamps, such as arc-tube heating for different fill pressures and arc currents. One must, however, bear in mind that the radiation-loss term does not contribute to the energy transport to the wall because most of the radiation penetrates the arc-tube wall. It should be noted that the Elenbaas–Heller equation was derived for describing lamp arcs.

3.2.2 High-Pressure Mercury Lamps

The arc tube of high-pressure mercury lamps consists of a fused silica ("quartz") tube with a circular cross section and pressed end seals. The end seals contain a molybdenum ribbon making the current connection to the electrode stem. A small nipple indicates the location where during the manufacturing process the tube is evacuated and filled with argon at a low pressure and with liquid mercury. The arc is established with the aid of a starting electrode, which is a tungsten rod placed close to one of the arcing electrodes requiring a lower voltage for the initial establishment of a discharge. The starting electrode is electrically connected to the opposite arcing electrode by a resistor, and once a plasma is generated in the arc tube by the initial arc, transfer of the discharge from the starting electrode to the other arcing electrode takes place. The arc results in evaporation of the mercury over a time period of a few minutes, and in a final pressure of 2 to 4 atmospheres depending on the initial amount of mercury in the lamp. During this time the arc voltage drops from approximately 200 to close to 120 V while the arc current increases from less than 1 A to the final value of between 3 and 4 A for a 400-W mercury lamp. After the lamp is turned off, a cooling-down period is necessary to reduce the pressure inside the arc tube and the voltage necessary for lamp starting.

High-pressure mercury lamps cover a wide range of powers and have efficacies of 30 to 65 lm/W. The color rendition is determined by five principal lines in the visible region of the spectrum at 404.7, 435.8, 546.1, 577, and 579 nm, indicating a dominance of green and blue. The majority of the ultraviolet resonance radiation is self-absorbed in the cooler regions near the arc-tube wall. Operation at higher pressures leads to increased broadening of the lines and some shift of the radiation to longer wavelengths; however, there remains a lack of radiation in the red region of the spectrum. To improve color rendition and at the same time make use of the considerable amount of UV radiation from the Hg discharge, the outer envelopes of some mercury lamps are coated with a red-emitting phosphor. Plate 3 shows the spectrum radiated by a high-pressure mercury lamp in comparison with other HID lamp spectra (Kaufman, 1981).

3.2.3 Metal-Halide Lamps The arc tube of the metal-halide lamp is very similar to that of the high-pressure mercury lamp, the major difference being that in addition to mercury, metal-halide compounds provide the metal vapor for the high-luminosity arc. Many metal halides have high vapor pressures at the arc-tube temperatures compared to the metals; thus the halide is readily evaporated. When the halide vapors diffuse into the high-temperature arcing region, the molecules dissociate, providing the metal atoms for the metal-vapor arc, while diffusion of the metal vapor from the high-temperature arcing region leads to recombination with the halide atoms in the cooler regions. As a consequence, one has the metal atoms inside the arc providing the high-intensity radiation, and metal-halide molecules in the colder regions close to the arc-tube wall, reducing the unwanted chemical reactions between the metal and the silica arc-tube material. Mixtures of metal halides are selected to provide a combination of high efficacy and good color rendition. For example, sodium and scandium iodides are used because sodium is one of the most easily excited atoms providing high radiation efficacy, and scandium provides emission lines in many parts of the spectrum. Similarly, a combination of sodium, thallium, and indium iodides in the arc tube provides emission lines in different parts of the visible spectrum, giving good color rendition. Efficacy values range from 75 to 125 lm/W for lamps with power ratings from 175 to 1500 W.

These desirable features have to be obtained through several design modifications compared with the simpler mercury lamp construction. To avoid condensation of the metal halides in the arc-tube ends away from the arc, a reflecting coating is applied to the arc-tube ends to increase the temperature in this region. The halides of some metals, in particular of sodium, experience electrolysis in the silica in the presence of an electric field, leading possibly to sodium diffusion through the arc-tube wall close to the current-carrying wires, as well as to shorting of the starter-electrode wire in the arc-tube end seal. Therefore, the current connections are led farther away from the arc tube, and after arc initiation the starting electrode is either shorted or disconnected from the circuit voltage by a bimetal switch.

Starting of the metal-halide lamp is similar to that of a mercury lamp, only a higher starting voltage is needed. Because the different metal halides have different vapor pressures at the different temperatures the arc tube experiences during starting, the color of the lamp changes during warmup for several minutes. Also, because of the higher arc-tube temperatures compared to that in mercury lamps, cooling down takes considerably longer, and it may take as much as 15 min before a metal-halide lamp can be restarted.

3.2.4 High-Pressure Sodium Lamps In this type of lamp, sodium is the principal light emitter, assuring the highest efficacies (exceeding 140 lm/W with high-power HPS lamps) at the expense of somewhat poorer color rendition. The arc tube contains xenon as the starting gas and sodium amalgam as the compound providing the metal vapor during operation. After warmup, the amalgam is partially evaporated, and both sodium and mercury are in the vapor phase. While the radiation properties are determined mostly by the sodium vapor, the mercury vapor serves as a buffer gas increasing the lamp pressure and with it the arc voltage and the width of the sodium line.

Since silica is attacked by sodium, fine-grained polycrystalline alumina (PCA) is used as arc-tube material. This material is highly translucent, allowing more than 90% of the radiation from the arc to be transmitted. The tube has the shape of a narrow cylinder with metal or cermet endcaps that are attached to the alumina with special frits matching the thermal-expansion characteristics of the alumina.

Even though most of the radiation stems from the resonance doublet of the sodium D lines at 590 nm, the color rendition of these lamps is rather good, because the lines are substantially collision broadened in the high-pressure environment, and a strong "self-reversal" reduces the radiation at the line center. This self-reversal is generated by the absorption of the radiation from the central portions of the arc by the vapor in the cooler outer regions. Plate 3 shows the wavelength distribution of the radiation from such a lamp (Kaufman, 1981).

Because the arc tube of the high-pressure sodium lamp has a rather small diameter, no starting electrode is incorporated, and the lamp is started with a short high-voltage pulse of 4000 to 6000 V. This high-voltage pulse is applied once during every 60-Hz ac voltage cycle until the arc is established. Full operating pressure is reached after approximately 10 min, and since the operating pressure is usually lower than with mercury or metal-halide lamps, restarting after shutoff can occur already after approximately 1 min.

3.3 Specialty Lamps

There are numerous lamps for special applications, but here only two of them are described because of their use of the high energy density of thermal plasmas.

3.3.1 Short-Arc Lamps These types of lamps derive their name from the relatively short distance between the cathode and anode of 0.3–12 mm, which leads to cathode-jet stabilization of the arc. The lamps operate with direct current and use a cathode rod with a conical tip and a slightly larger-diameter anode rod within a spherical or ellipsoidal arc tube made out of fused silica (see Fig. 19). The electrodes are usually water cooled. The most frequently used arcing media are xenon gas at several atmospheres or mercury with xenon added as starting gas. The xenon arcs have the advantage that their spectrum is rather continuous in the visible wavelength region, comparable to that of sunlight.

Short-arc lamps have been designed for power levels of up to 30 kW, and their very high luminosity makes them ideal for applications as search lights or projection lamps. The xenon arc has temperatures higher than the arcs in lamps containing metal vapor, and its appearance can be compared with the cathode region of a free-burning arc. To avoid an excessive amount of short-wavelength UV radiation escaping from the arc and leading to ozone generation around the lamp, a silica is used that cuts off transmission at 210 nm.

3.3.2 Carbon-Arc Lamps A direct-current arc burning between two carbon electrodes has been used for lighting purposes since the middle of the 19th century. Different types of carbon arcs have distinctively different characteristics. The "low-intensity carbon arc" consists of two graphite electrodes, a horizontal anode and a cathode set

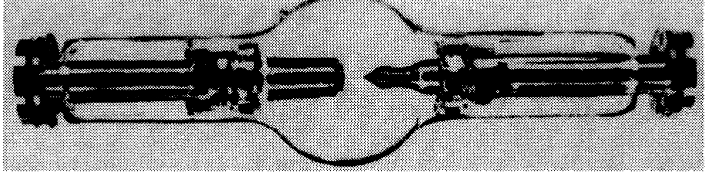

FIG. 19. Schematic of a short-arc lamp (Kaufman 1981, with kind permission of IES).

at an angle to the horizontal, the arc burning in air. The radiation from the crater at the anode tip serves as the light source. Typical currents are less than 10 A for a 9-mm-diameter anode, and typical arc voltages are around 60 V. The anode tip has a temperature of close to 4000 K, and emits essentially blackbody radiation at this temperature. The amount of radiation from the arc itself is small compared to the anode radiation.

If metal compounds are incorporated in the core of the graphite electrode rods, evaporation of these compounds introduces metal vapor into the arc leading to easy excitation of the metal atoms and high luminosity from the arc. Rare-earth metals are frequently used as additions. For increases in the current density, the metal evaporation becomes so strong that a jet from the anode appears similar to the cathode jet, resulting in an enlarged region of radiating gas as shown in Fig. 20 (high-intensity carbon arc, Beck arc).

One of the advantages of the dc carbon-arc lamp is that it provides continuous radiation whereas the lamps operating with alternating current display a 120-Hz flicker because of the periodic extinction and reignition of the arc in every half cycle of the current. Accordingly, the carbon-arc lamp finds its application in movie projection lamps where flicker may lead to elimination of some frames. One of the disadvantages is that the electrodes burn off quite rapidly and have to be continuously fed, comparable to the graphite electrodes in a transferred-arc metallurgical furnace. The carbon arc produces fine soot particles, and usually it burns in a ventilated housing.

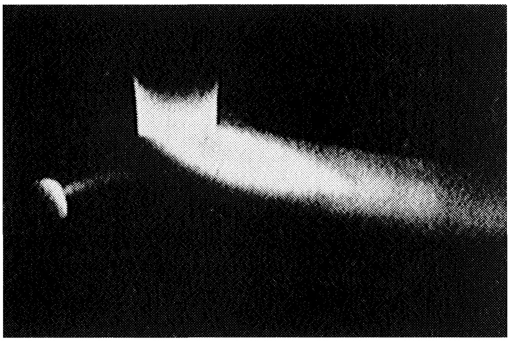

FIG. 20. Photograph of a carbon arc (Kaufman 1981, with kind permission of the IES).

4. CURRENT-INTERRUPTION DEVICES

4.1 Fundamentals of Circuit Breakers

Circuit breakers are thermal-plasma devices because current interruption in an electric power grid practically always results in the generation of an electric arc. Formation of the plasma allows the current to be interrupted because, as described in Sec. 1, current flow through a plasma is strongly dependent on the energy losses from the plasma. Strong cooling of the plasma results in a rapid decrease of the electrical conductivity by several orders of magnitude and of the current flow when the circuit voltage approaches a zero value during an ac cycle (current zero).

Current interruption is required in electrical circuits either when switching the power off to parts of the circuit, e.g., an electric installation or a motor, or when as a result of a fault (or a short circuit) a large current can flow from the power grid to the circuit section in which the fault has occurred. Circuit breakers limit the duration of the very high fault currents to a maximum of a few cycles in a 60-Hz power grid, thus minimizing the potential damage by the increased power dissipation. Depending on the location in the power grid, breakers are classified into three voltage categories:

1. low-voltage (110–1000 V) breakers are located close to the end user in the electric power grid,
2. medium-voltage (2500–38 000 V) breakers are in the distribution network, and
3. high-voltage (72–800 kV) breakers are in the transmission portion of the power grid.

During current switching, circuit breakers are used to control the making currents and interrupting currents in electrical circuits. To achieve this operation, breakers have interrupting elements actuated by an electromechanical mechanism. This element consists of a set of contacts (usually one stationary and the other movable) in a medium of gas, oil, or vacuum. For initiating currents (closing operation), assume the breaker contacts are initially separated (open position) with no current flowing through them and the circuit voltage appears across them. As the movable contact closes, the contact separation becomes progressively smaller until a voltage breakdown occurs initiating an electric arc (denoted as prestrike) that burns until the contacts actually touch. The process of ac (sinusoidal) current interruption is as follows:

1. Initially the breaker contacts are touching (closed position) and current is flowing through them.
2. The breaker mechanism is tripped causing the movable contact to separate from the stationary contact.
3. Upon contact separation, an electric arc is formed between the separating contacts and continues to burn until the ac current reaches a "natural" current zero.
4. Immediately following arc extinction at current zero, a race develops between the increasing dielectric strength of the gap between the contacts (dielectric recovery) and the increasing circuit voltage across this gap (recovery voltage). If the plasma is cooled rapidly enough so that the rate of dielectric recovery within the intercontact region is greater than the rate of the recovery voltage rise across the contacts, the arc does not reignite during the recovery period and current interruption is successful.

Circuit breakers have been used since electric power has been available. Originally, large contact separations were necessary to make the electric arcs in air sufficiently unstable that circuit interruption was possible. Since the 1930s, a significant research and development effort has been devoted to increasing arc motion control and instabilities during the early part of the circuit-interruption process in order to reduce the arcing time. New arcing media have been introduced: first oil, then SF_6 for medium- and high-voltage breakers, and vacuum for medium-voltage breakers. At present, a variety of circuit breakers is available, and we have seen a continuous decrease in arcing times and in breaker size for a given interrupting capability.

4.2 Low-Voltage Air Circuit Breakers

Low-voltage breakers (110–1000 V) are usually air magnetic breakers rated for load currents from 15 to 4000 A, and short-circuit currents from 10 to 200 kA. The principles discussed here apply to both molded-case circuit breakers and power circuit breakers. The construction of a typical low-voltage air circuit breaker is shown in Figs. 21 and 22. The breaker consists of two contacts (one stationary and one movable), a mechanism, an arc chamber, and overcurrent detection circuits. Often there are two sets of contacts—"main" contacts and "arcing" or "arc runner" contacts (Lee and Slade, 1984). During high–fault-current interruption, the contacts separate, initiating an arc between the "main" contacts. The arc transfers rapidly to the "arcing" contacts by a magnetic blowout force and into the arc chamber.

During the interruption of the load and overload currents (less than twice the load current), the breaker mechanism is tripped either manually or by a thermally triggered actuator, and the movable contact is accelerated from the stationary contact forming an electric arc of increasing length. The interaction of the self-induced magnetic field *B* with the arc current density *j* produces a magnetic force on the arc column driving it into the arc chamber plates. In the arc

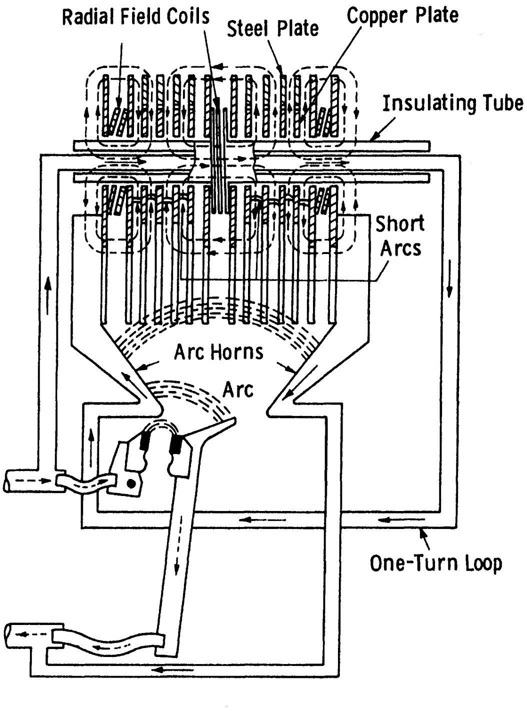

FIG. 21. Cross section of a low-voltage air circuit breaker (reprinted from Browne, 1984, p. 428, by courtesy of Marcel Dekker Inc.).

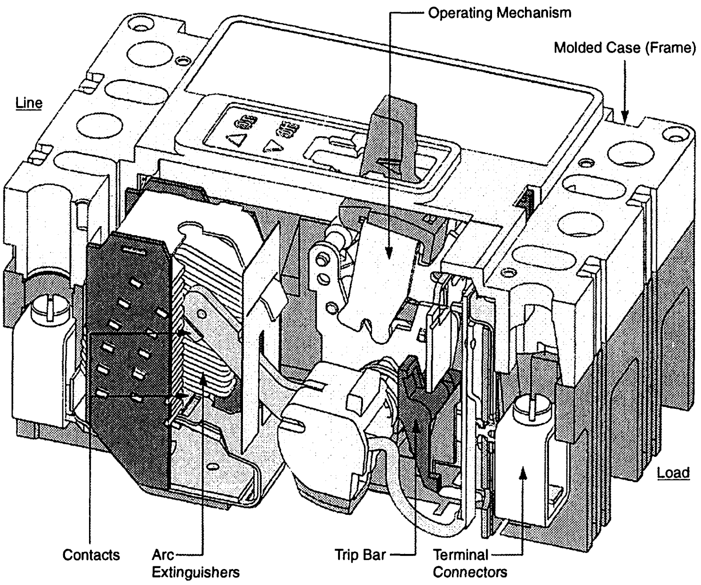

FIG. 22. Schematic of a molded-case circuit breaker (with kind permission of Cutler Hammer).

chamber, the arc divides into a series of short arcs each burning between adjacent U-shaped steel plates of the arc chamber. These short arcs are cooled by conduction to the steel plates, and an arc voltage of 25 V/gap (Lee and Slade, 1984) is generated. With 11 plates and 10 series arcs, a voltage of 300 V can be generated between the separated contacts. As the current approaches a zero value, the arc is extinguished and a recovery voltage appears across the contacts. The steel plates provide further rapid cooling of the arc plasma causing additional deionization of the hot gas in this region thus increasing the dielectric strength needed for successful current interruption.

For high–fault-current interruption, the arcing process is similar to that described except that the high arc current produces a larger-diameter and hotter arc that completely fills the arc chamber and is usually not broken into a series of shorter arcs. The plasma consists of a mixture of air, evaporated arc chute materials, and contact material vapors at about 10 000 K, with a high electrical conductivity (100 S at current crest). The cooling mechanisms for this arc are radial convection and conduction to the steel arc chamber plates, and ablation of material from the arc chamber walls. To promote high arc voltage, some breakers have means of enhancing the movable-contact opening speed with a "slot motor": This design produces an additional acceleration force on the contact arm by the interaction of the magnetic field created by the current through the contact arm with the U-shaped plates of the slot motor (see Fig. 23). As a result of the more rapid contact separation, a high arc voltage is developed quickly, thus limiting the let-through current. The current approaches zero sooner, the arcing time is shortened (e.g., from 8.3 ms or longer to 4–5 ms), and contact erosion is reduced. During the recovery period, the conductance of the arc column is decreased several orders of magnitude within microseconds by rapid cooling and deionization. Therefore, at current zero, the source voltage is low making both the rates of rise and the peak of the recovery voltage significantly lower. Thus the current-limiting effects of the thermal plasma lead to easier interruption of the fault current.

4.3 Oil Circuit Breaker

Medium-voltage (5–38 kV) oil breakers or reclosers have three sets of current interrupter elements immersed in an oil tank. Rods connected from the mechanism to the current-interrupting elements operate the contacts of the interrupter elements. Each element consists of a set of contacts. One is

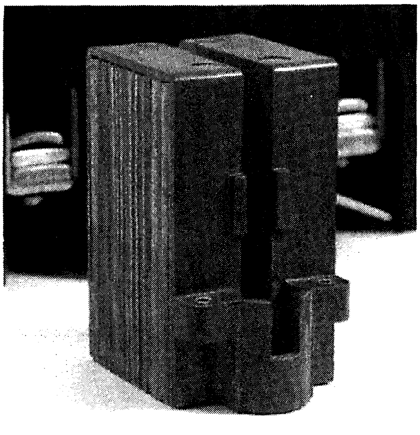

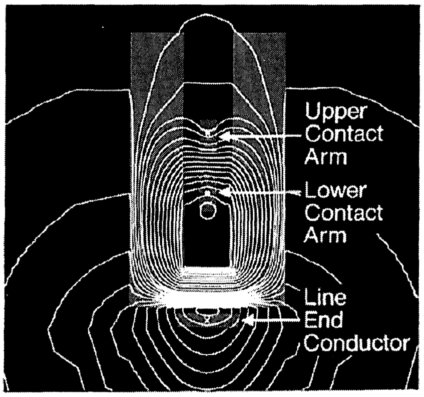

Upper
Contact
Arm
Lower
Contact
Arm
Line
End
Conductor

Magnetic Field Flux Plot-FCL Contact Arms
2000 Amps. Peak and Contacts Closed
Force Between Contact Arms = 1.9#

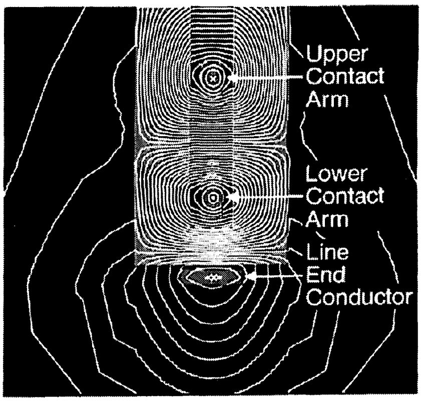

Upper
Contact
Arm
Lower
Contact
Arm
Line
End
Conductor

Magnetic Field Flux Plot-FCL Contact Arms
14000 Amps. Peak and 50% of Travel
Force Between Contact Arms = 32.8#

FIG. 23. Magnetic-field distribution in a low-voltage circuit breaker with slot motor (with kind permission of Cutler Hammer).

a stationary "tulip"-shaped contact and the other is a cylindrical movable contact inserted in the tulip-shaped contact. A spring mechanism provides the movable electrode with a motion of several inches. These contacts are surrounded by pots or deion grids used to control and cool the arc plasma. During the current-interruption process, the contacts are separated, forming a thermal plasma between them. The hot arc plasma decomposes the oil into its elements such as carbon, hydrogen, and oxygen, and the gaseous mixture fills the intercontact region. The plasma is cooled by convection and evaporation of the oil and by conduction to the deion grids. The hydrogen gas within the arc mixes with the oil producing additional cooling. The hydrogen gas is essential during the recovery period because of its excellent dielectric properties and thermal heat capacity.

4.4 Vacuum Circuit Breaker

The medium-voltage vacuum circuit breaker consists of an insulated vacuum vessel, 4 to 7 in. (10 to 18 cm) in diameter and approximately 12 in. (30.5 cm) long, made from a cylindrical section of either ceramic or glass with metal end plates attached (see Fig. 24). The vessel contains a set of contacts

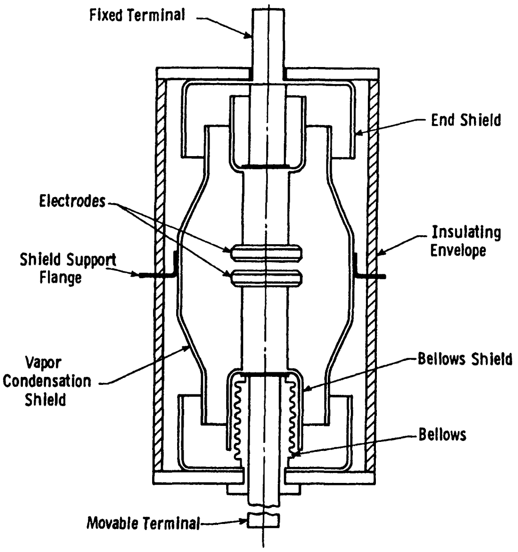

Fixed Terminal
End Shield
Electrodes
Insulating Envelope
Shield Support Flange
Vapor Condensation Shield
Bellows Shield
Bellows
Movable Terminal

FIG. 24. Schematic of a vacuum circuit breaker (reprinted from Browne, 1984, p. 457, by courtesy of Marcel Dekker Inc.).

brazed to electrode stems of which one is fixed to the end plate. A mechanism provides motion (20 mm or less) to the movable contact through the opposite end plate via a metal bellows, which serves to maintain a vacuum of 10^{-6} Torr or less. Surrounding the contacts is a metal shield that protects the envelope from deposition of arc products.

Current interruption in vacuum is similar to that in gas, except that the ionized medium is metal vapor originating from the erosion of the contacts during the arcing process. At low arc currents, this metal-vapor arc is not a thermal plasma since strong nonequilibrium conditions prevail. The cathode has a multitude of highly mobile arc spots moving rapidly over the surface, each spot supporting about 100 A with a typical current density of 10^8 A/cm^2 (Jüttner, 1981). The voltage drop of the cathode region depends on the contact material (typically 18 V for Cu). The anode surface may have a diffuse current collection without anode spots or may have a constricted anode spot causing gross erosion of the anode surface (e.g., 10^{-4} g/C). A constricted anode spot occurs when there is an insufficient number of ions in the arc column to balance the charges of the current-carrying electrons. A space charge is formed leading to local heating of the anode surface and evaporation of anode material producing the constricted arc attachment (Kimblin *et al.*, 1984).

The column region can be either diffuse or constricted. For arc currents up to several hundred amperes, the arc column is diffuse and electrons flow from the cathode to the anode with a low column voltage. At higher currents (2000 A or greater) the positive column consists of electrons and ionized metal-vapor products generated by the material evaporated from the anode and cathode. The radius of the column is determined by the outward radial pressure of the ionized and neutral species balanced by the inward magnetic constriction pressure formed by the arc current (magnetic pinching). Under these conditions, the metal-vapor arc column is nearly in thermal equilibrium radiating strongly, and the arc voltage increases with current, reaching values as high as 200 V. Cooling of the arc column is primarily by radiation to the surrounding surfaces and by thermal conduction to the electrodes. To prevent excessive heating of the anode, the contacts are designed such that current flow in part of the contact generates a radial magnetic field that interacts with the arc current causing the arc roots to move rapidly in a circular motion.

As the current decays toward a natural current zero, the constricted column becomes a diffuse arc. Close to current zero, the arc becomes unstable and suddenly extinguishes, an effect known as "current chop" with a recovery voltage across the contacts. Electrons are rapidly lost from the interelectrode region leaving an expanding ion space charge at the anode surface producing a post-arc current (several amperes) that flows for several microseconds. The recovery voltage appears across the expanding ion space-charge region. If the rate of dielectric recovery determined by the loss of ions from the intercontact region is faster than that of the rate of rise of the recovery voltage, interruption will occur.

4.5 Compressed-Air and SF$_6$ Breakers

High-voltage circuit breakers (72 to 800 kV) have continuous currents from 600 to 4000 A, and fault currents of 20–63 kA. Because of the high circuit voltage, the current-interruption medium is usually compressed air or compressed SF$_6$ because both the heat capacity and dielectric strength of these gases increase with pressure. Construction of such a breaker consists of a separate interrupting element per phase, large insulators, and a pneumatically operated mechanism (see Fig. 25). The interrupting elements are contained in a tank, which can be ungrounded (alive tank breaker) or grounded (dead tank breaker).

The arcs produced during current interruption in this breaker are different from magnetic air breakers in that

1. the arc is confined in a channel and operates at higher ambient gas pressures, and
2. the arc is extinguished primarily by high-velocity axial flow providing convective cooling.

Figure 26 shows the basic double-flow electrode arrangement for such a circuit breaker. For current interruption, the movable contact separates from the stationary, and an arc is

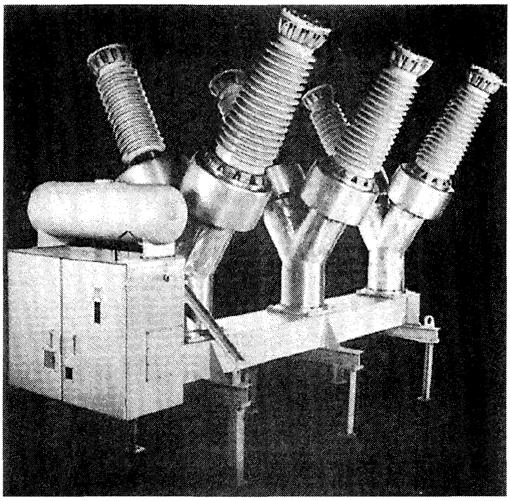

FIG. 25. Photograph of a high-voltage three-phase circuit breaker (reprinted from Browne, 1984, p. 417, by courtesy of Marcel Dekker Inc.).

initiated between the contact tips. Gas enters radially into the separated-contact upstream region and flows downstream through appropriately designed nozzles. The arc moves rapidly into the inside of the contacts, attaching to the inner contact walls. While there are a number of different high-voltage circuit breaker designs, the basic principle of cooling the arc through strong axial nozzle flows and generating instabilities due to turbulence close to current zero is the same.

Two fundamental physical models for the transient behavior of the arc column during current interruption have been developed. The first model, developed by Cassie (1953), treats the arc plasma, assumed to be in local thermodynamic equilibrium, as a hot fluid having a uniform temperature and a temperature-dependent electrical conductivity. The plasma temperature equation depends on the energy balance of the plasma. During the voltage recovery period, the second model, developed by Mayr, is used. This model assumes that the electron temperatures are greater than those of the heavy species and derives conditions for recovery depending on electron energy and recombination, ionization, and attachment processes (Browne, 1984).

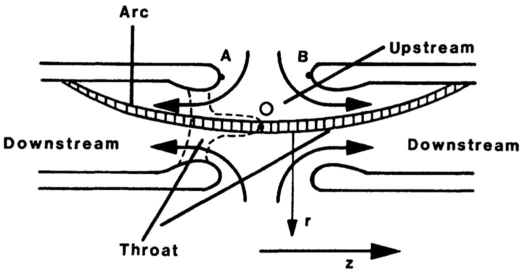

FIG. 26. Schematic of the arcing region of a double-flow high-voltage circuit breaker (reprinted from Browne, 1984, p. 158, by courtesy of Marcel Dekker Inc.).

4.6 Concluding Remarks

With circuit breakers, as with all other thermal-plasma devices, significant improvements in functional performance have been achieved through understanding the interaction of the plasma with its surroundings. In circuit breakers, motion induced by magnetic-field interaction and cooling of the plasma is used to generate the conditions for current interruption, while some of the same effects are used in other thermal-plasma devices to enhance stability of operation. Controlling the plasma using fluid dynamics, heat transfer, and magnetic fields remains the central issue for all thermal-plasma devices.

Works Cited

Barcza, N. A., Curr, T. U., Denton, G. M. (1989), in: M. Boulos (Ed.), *Proceedings of Workshop on Industrial Plasma Applications, Pugnochiuso, Italy,* Bari, Italy: Università di Bari, pp. 115–120.

Browne, T. E. (Ed.) (1984), *Circuit Interruption,* New York: Marcel Dekker.

Capetti, A., Pfender, E. (1989), *Plasma Chem. Plasma Process.* **9,** 329–341.

Cassie, A. M. (1953), "Introduction to the Theory of Circuit Interruption," in: H. Trencham (Ed.), *Circuit Breaking,* London: Butterworth, p. 46; see also CIGRE 10, Rept. 102 Paris 1939.

Coudert, J. F., Betoule, O., Planche, M. P., Fauchais, P. (1993), in: J. Harry (Ed.), *Proceedings of the 11th ISPC,* England: University of Loughborough, pp. 178–185.

Dinulescu, H. A., Pfender, E. (1980), *J. Appl. Phys.* **51,** 3149–3157.

Drouet, M. G. (1993), "High Temperature Processes for Industrial Waste Treatment and Valorisation," in: G. Bonizzoni, W. Hooke, E. Sindoni

(Eds.), *Industrial Applications of Plasma Physics,* Bologna, Italy: Societa Italiana di Fisica.

Eschenbach, R. C. (1989), in: M. Boulos (Ed.), *Proceedings of Workshop on Industrial Plasma Applications, Pugnochiuso, Italy,* Bari, Italy: Università di Bari, pp. 127–136.

Fauchais, P., Boulos, M., Pfender, E. (1987), "Physical and Thermodynamic Properties of Thermal Plasmas," in *Plasma Technology in Metallurgical Processing,* Warrendale, PA: Iron and Steel Society.

Fincke, J. R., Swank, W. D., Jeffery, C. C. (1990), *IEEE Trans. Plasma Sci.* **18,** 948–957.

George, A. P., Hare, A. L. (1983), in: M. I. Boulos, R. J. Munz (Eds.), *Proceedings of the 6th ISPC,* Montreal, Canada: Université de Sherbrooke and McGill University, pp. 144–149.

George, C., Candler, G., Young, R., Pfender, E., Heberlein, J. (1996), *Plasma Chem. Plasma Process.* **16** (1), 43S–56S.

Grey, J., Jacobs, P. F. (1964), *AIAA J.* **2** (3), 433–438.

Griem, H. R. (1964), *Plasma Spectroscopy,* New York: McGraw-Hill.

Heberlein, J. V. R. (1992), *Pure Appl. Chem.* **64** (5), 629–636.

Heberlein, J. V. R., Melilli, W. J., Dighe, S. V., Reed, W. H. (1989), in: M. I. Boulos (Ed.), *Proceedings of Workshop on Industrial Plasma Applications, Pugnochiuso, Italy,* Vol. 2, Bari, Italy: Università di Bari, pp. 1–8.

Hsu, K. C., Pfender, E. (1983), *J. Appl. Phys.* **54,** 3818–3824.

Huang, P. C., Heberlein, J., Pfender, E. (1995), *Plasma Chem. Plasma Process.* **15** (1), 25–46.

Jenista, J. (1995), "Numerical Model of the Anode Region of High-Current Electric Arcs," Master's Thesis, University of Minnesota, Minneapolis.

Jüttner, B. (1981), *J. Phys. D: Appl. Phys.* **14,** 1265–1275.

Kaufman, J. E. (Ed.) (1981), *IES Lighting Handbook, Reference Volume,* New York: Illuminating Engineering Society of North America.

Kimblin, C. W., Slade, P. G., Voshall, R. E. (1984), "Interruption in Vacuum," in: T. E. Browne (Ed.), *Circuit Interruption, Theory and Techniques,* New York: Marcel Dekker, Inc., p. 462.

Lee, A., Slade, P. G. (1984), "Molded-Case Low Voltage Circuit Breakers," in: T. E. Browne (Ed.), *Circuit Interruption, Theory and Techniques,* New York: Marcel Dekker, Inc., p. 550.

Leveroni, E., Heberlein, J. V. R., Pfender, E. (1991), in: K. Wiesemann (Ed.), *Proceedings of the 10th ISPC,* Bochum, Germany: Universität Bochum, pp. 1.3.8.1–6.

Lochte-Holtgreven, W. (Ed.) (1968), *Plasma Diagnostics,* Amsterdam: North-Holland.

Maecker, H. (1955), *Z. Phys.* **141,** 198–216.

Malmberg, S., Heberlein, J. (1993), *J. Thermal Spray Technol.* **2** (4), 339–344.

Morrow, R., Lowke, J. J. (1993), *J. Phys. D: Appl. Phys.* **26,** 634–642.

Paik, S., Huang, P. C., Heberlein, J., Pfender, E. (1993), *Plasma Chem. Plasma Process.* **13** (3), 379–397.

Peters, T. (1956), *Z. Phys.* **144,** 612–631.

Pfender, E. (1978), "Electric Arcs and Arc Gas Heaters," in: H. Oskam (Ed.), *Gaseous Electronics,* Vol. 1, New York: Academic.

Reid, K., Camacho, S. L. (1987), "Historical Review," in: J. Feinman (Ed.), *Plasma Technology in Metallurgical Processing,* Warrendale, PA: Iron and Steel Society.

Sanders, N., Etemadi, K., Hsu, K. C., Pfender, E. (1982), *J. Appl. Phys.* **53,** 4136–4145.

Snyder, S. C., Reynolds, L. D., Fincke, J. R., Lassahn, G. D., Grandy, J. D., Repetti, T. E. (1994), *Phys. Rev. E* **50,** 519–525.

Vardelle, M., Vardelle, A., Fauchais, P., Boulos, M. I. (1983), *J. AICh.E.* **29,** 236–243.

Waymouth, J. F. (1971), *Electric Discharge Lamps,* Cambridge, MA: MIT Press.

Winovich, W., Carlson, W. C. A. (1980), *ISA Trans.* **19** (2), 75–93.

Wolf, C. B., Fey, M. G. (1971), U.S. Patent No. 3,760,145.

Wutzke, S. A., Pfender, E., Eckert, E. R. G. (1967), *AIAA J.* **5** (4), 707–714.

Young, R. M., Chyou, Y. P. Fleck, E., Pfender, E. (1983), in: M. I. Boulos, R. J. Munz (Eds.), *Proceedings of the 6th ISPC,* Montreal, Canada: Université de Sherbrooke and McGill University, pp. 212–218.

Zhou, X., Heberlein, J. (1994), *Plasma Sources Sci. Technol.* **3,** 564–574.

Zhu, H., Lau, Y. C., Pfender, E. (1991), "Radio Frequency Thermal Plasma Chemical Vapor Deposition of Superconducting $Y_1Ba_2Cu_3O_{7-x}$ Films," *J. Appl. Phys.* **69,** 3404–3406.

Zhu, P., Lowke, J. J., Morrow, R. (1992), *J. Phys. D: Appl. Phys.* **25,** 1221–1230.

Further Reading

Boulos, M. I., Fauchais, P., Pfender, E. (1994), *Thermal Plasmas Fundamentals and Applications,* New York: Plenum Press.

Browne, T. E. (Ed.) (1984), *Circuit Interruption,* New York: Marcel Dekker.

Feinman, J. (Ed.) (1987), *Plasma Technology in Metallurgical Processing,* Warrendale, PA: Iron and Steel Society.

Kaufman, J. E. (Ed.) (1981), *IES Lighting Handbook, Reference Volume,* New York: Illuminating Engineering Society of North America.

Pfender, E. (1978), "Electric Arcs and Arc Gas Heaters," in: H. Oskam (Ed.), *Gaseous Electronics,* Vol. 1, New York: Academic.

Waymouth, J. F. (1971), *Electric Discharge Lamps,* Cambridge, MA: MIT Press.

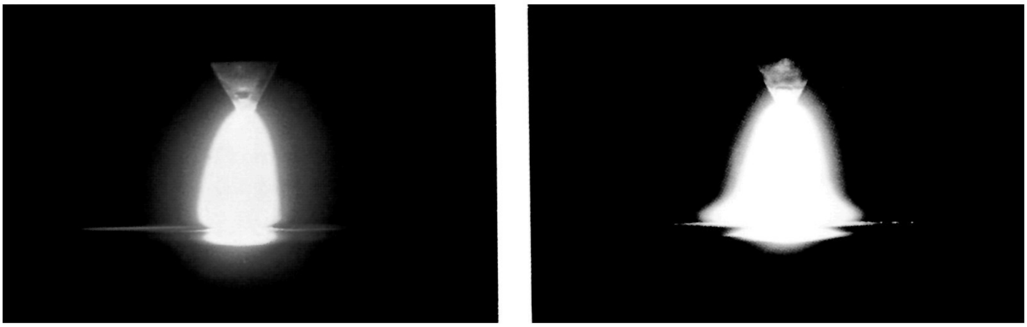

Vol. 21 – PLATE 1. Photographs of free-burning arcs showing the highly luminous high-temperature region between the rod cathode (top) and a flat anode (bottom); left, constricted anode attachment; right, diffuse anode attachment.

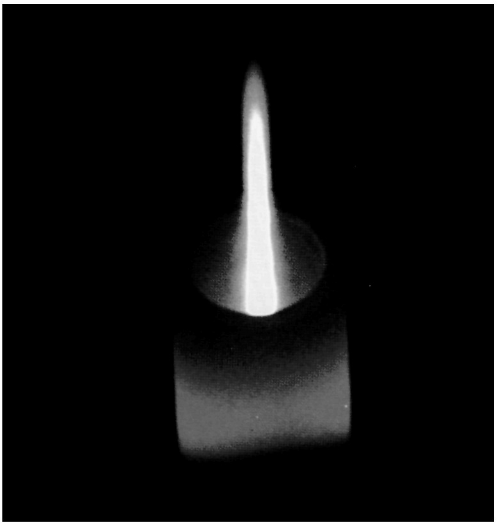

Vol. 21 – PLATE 2. Photograph of an arcjet thruster in a laboratory.

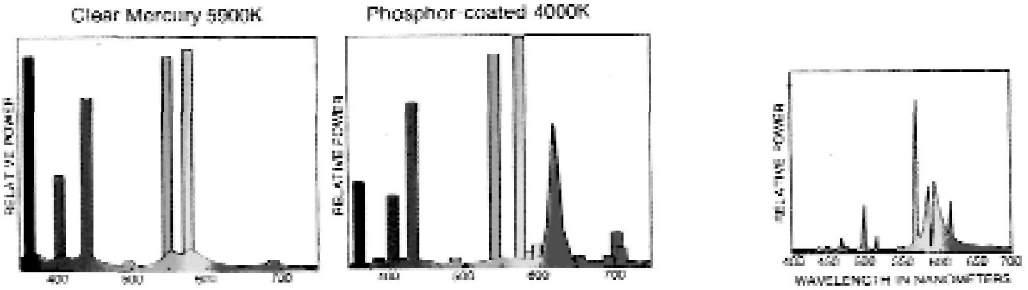

METAL HALIDE LAMPS

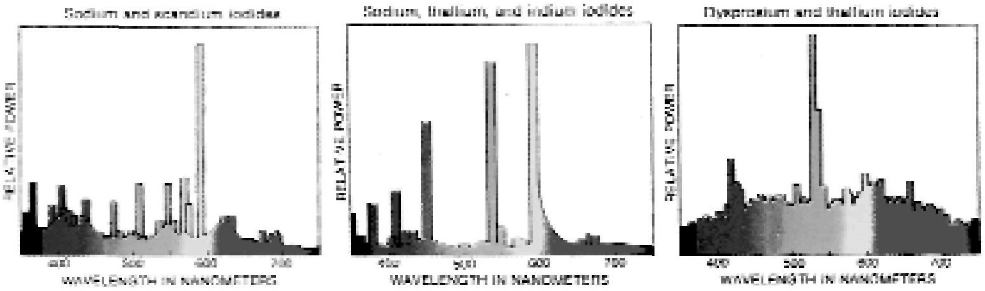

INCANDESCENT LAMPS
(INCLUDING TUNGSTEN-HALOGEN)

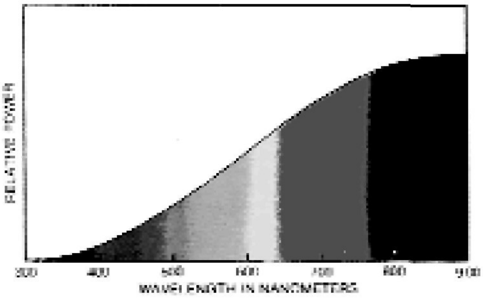

Vol. 21 – PLATE 3. HID lamp spectra (Kaufman 1981, with kind permission of the IES).

THERMAL PROCESSES*

ERIC J. COTTS, *Binghamton University, State University of New York, Binghamton, New York, U.S.A.*

DONALD G. ARCHER, *National Institute of Standards and Technology, Gaithersburg, Maryland, U.S.A.*

Introduction 193
1. **Thermodynamic Principles** 193
1.1 Free Energy, Enthalpy, Entropy, and Chemical Potential of a Single Component 194
1.2 Heat Capacity and Its Relation to Other Functions 195
2. **Measurement of Thermal Processes** 196
2.1 Calorimetry 197
2.1.1 Static Calorimeters, Adiabatic 197
2.1.2 Static and Flow Calorimeters, Isothermal 198
2.1.3 Static and Flow Calorimeters, Isoperibol 199
2.1.4 Scanning Calorimeters 199
3. **Thermal Processes** 199
3.1 Phase Changes 199
3.2 Atomic Vibrations and the Heat Capacity of a Solid 200
3.3 Electronic Effects 202
3.4 Thermal Expansion 203
3.5 Quantification of Thermal Processes 204
3.6 Defects in Materials 205
4. **Summary** 208
Glossary 208
Works Cited 208
Further Reading 208

INTRODUCTION

In the broadest sense, thermal processes of a substance or a system are those occurrences that are effected by changes in enthalpy, energy, or temperature. By this definition, all possible processes or events are essentially thermal processes. The reading of this article, the manufacture of this book, the movement of the cosmos all are examples of complex thermal processes. The melting of an ice cube or the warming of a gas in an insulated flask are simpler thermal processes.

We are not so ambitious as to attempt to describe the functioning of the universe within this article. Instead, we restrict exposition to a description of the measurement of thermal processes and some of the underlying physical behaviors that cause some observed thermal-process behaviors. In the first section we present a brief discussion of thermodynamic and statistical-mechanical relations relevant to the measurement and interpretation of thermal processes. The second section describes, in general terms, the principles of measurement of thermal properties of materials and systems. The methods described are those collectively known as calorimetry, a means to monitor thermal processes directly and therefore to understand better the states of matter. The third section contains descriptions, in physical terms, of some thermal processes appropriate to an encyclopedia concerned with *applied* physics.

1. THERMODYNAMIC PRINCIPLES

Thermal processes are essentially thermodynamic processes. Certain aspects of thermodynamics are discussed here as they relate to the measurement of thermal processes. A more complete description of

---

*Contribution of the National Institute of Standards and Technology, not subject to copyright in the United States.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

thermodynamics is found elsewhere in the Encyclopedia (see THERMODYNAMICS, EQUILIBRIUM, and THERMODYNAMICS, NONEQUILIBRIUM). Also, because the thermal properties of gases are described elsewhere in the Encyclopedia (see THERMAL PROPERTIES OF GASES), they are excepted from the present discussion.

Thermal processes can be divided into two categories. The first is the change in thermodynamic properties of a system that is undergoing chemical or nuclear reactions. Although these are thermal processes, they are considered in detail elsewhere in the Encyclopedia (see THERMOCHEMISTRY) and so are not considered here. The second category consists of changes in thermodynamic properties of an individual substance that result from a change in its temperature. For either of these categories, the systems may be open or closed to mass transport.

Measured thermodynamic properties, or thermal processes, are essentially averages over an extremely large number of atoms or molecules, each of which may have a large number of distinct energy levels available to it. As a result, the thermal properties cannot tell us the nature of the energy levels or their specific populations. The thermal properties of a system may be calculated from a hypothetical description of the energy levels of a system and their populations. The link between the energy levels, and their populations, of a large sample of individual particles and thermodynamics is made through statistical mechanics (see STATISTICAL MECHANICS, CLASSICAL). Agreement between the hypothesized physical behavior of a system and its real behavior is often tested through measurement of the thermal (or other) properties. It is possible that more than one physical description of a molecular process may yield the same value of a particular thermal property. Often, different physical models can be differentiated by a comparison of each model's calculated thermal properties with measured values over varying temperatures (or other variables). Such tests of theory comprise an important use for measurements of the thermal properties of a substance or a process involving several substances.

Another important use of measurements of thermal properties is in their contribution to the knowledge of chemical potentials of substances as functions of temperature. The value of such direct measurements is enhanced because of the limited accuracy of physical models in the prediction of thermal properties. The chemical potentials of substances determine their relative thermodynamic stability and thus the extent of reaction of chemical systems (see THERMODYNAMICS, EQUILIBRIUM, and THERMOCHEMISTRY). Because of the importance of thermal properties to the chemical potential, we give a brief description of these properties, with a perspective related to the actual measurement of thermal processes.

1.1 Free Energy, Enthalpy, Entropy, and Chemical Potential of a Single Component

We begin with the extensive functions for a single-component system, originally defined by Gibbs (1875), and now generally termed the enthalpy and the free-energy functions:

$$H = U + pV, \tag{1}$$

$$A = U - TS, \tag{2}$$

$$G = H - TS = A + pV, \tag{3}$$

where H is the enthalpy, U is the internal energy, p is pressure, T is temperature, and V is volume. The two functions G and A are the Gibbs free energy and the work (*Arbeit*) function, respectively. All of these functions are state functions, requiring that they be path independent. The natural variables of G are T and p, and for A are T and V. The function A has also been called the Helmholtz energy, in honor of von Helmholtz who realized the usefulness of this function for some applications (von Helmholtz, 1882), some years after Gibbs proposed it for a wider range of applications (Pitzer, 1995).

Of the two free-energy functions, G and A, the Gibbs energy is the most commonly used, in part because its natural variables, T and p, can be most conveniently controlled in real applications. The work function A is useful when the primary interest is the determination of the maximum amount of work that could be obtained for an isothermal process (the maximum amount of work corresponds to thermodynamic reversibility). The work function is commonly used for molecular-dynamical and other calculational simulations of the properties of a system and also is somewhat more convenient than the Gibbs energy for creating mathematical descriptions of a fluid's equation of state near its critical point. However, because of the much greater importance of the Gibbs energy in thermodynamic calculations of real processes, the Gibbs energy and its derivative functions will be emphasized here.

In thermodynamic processes, the primary interest is in the *changes* in enthalpy and energy relative to some arbitrarily chosen state of the involved substances. This state is chosen primarily on the basis of convenience. As such, different choices of standard states might be used in similar calculations for the same system or process. Energetic contributions from such things as nuclear binding energy or mass-energy equivalence do not change significantly when a substance undergoes a thermal change or a chemical reaction; thus they are not exhibited explicitly in thermodynamic calculations. The small changes that do occur in these energies are included in the thermodynamic energy or enthalpy change. Indeed, thermodynamics was applied successfully many years before the nature of these energetics was known.

Consideration of the entropy of a substance often includes the entropy value at absolute zero, 0 K. Entropy is related to the probability of population of the energy levels available to the system. At 0 K, the entropy of a perfect crystal is related to the degeneracy of the lowest-energy state of the crystal. As such, the entropy of a perfect crystal with no degeneracy of the lowest-energy state is taken to be zero at 0 K. The degeneracy of the lowest-energy state for some substances is not zero, and these substances at 0 K have a nonzero entropy, referred to as the residual or configurational entropy.

For the reasons stated above, the measured quantities are the changes of enthalpy or energy that accompany a particular process. This concept is obvious as it relates to the measurement of an enthalpy change for a reaction. Less obvious is the determination of the enthalpy content of a single component or substance relative to some arbitrary set of conditions. Conceptually this is accomplished through changing the temperature of the material and determining the change in enthalpy required to effect the change in temperature. Such a measurement is known as an enthalpy increment and is characterized by the two temperatures and the enthalpy change.

The chemical potential in a single-component system is simply the Gibbs energy per mole of substance. Thermodynamic relations for multicomponent, homogeneous phases are more complex and are described elsewhere (see THERMODYNAMICS; also Pitzer, 1995).

1.2 Heat Capacity and Its Relation to Other Functions

Often, the discussion of thermal processes will involve the temperature derivatives of either the enthalpy or the energy. These temperature derivatives are known as heat capacities. The constant-pressure heat capacity, C_p, and constant-volume heat capacity, C_V, are defined as

$$C_p = \left(\frac{\partial H}{\partial T}\right)_p, \tag{4}$$

$$C_V = \left(\frac{\partial U}{\partial T}\right)_V. \tag{5}$$

There exists a fundamental difference in the enthalpy and the heat capacity of a system that is not immediately obvious from Eqs. (4) and (5). The extensive enthalpy of a system is the sum of the individual enthalpies of the individual entities that comprise the system

$$H = \sum_i n_i H_{m,i}, \tag{6}$$

where n_i is the number of moles of the ith substance and the subscript m denotes molar. The heat capacity is more complicated; it is

$$C_p = \sum_i n_i\left(\frac{\partial H_{m,i}}{\partial T}\right)_p + \sum_i \left(\frac{\partial n_i}{\partial T}\right)_p H_{m,i}. \tag{7}$$

Thus, the heat capacity of a complex system includes a thermodynamic component resulting from the change in distribution of entities among energy levels as well the interconversion of entities—for example, the change of amounts of products and reactants in a chemical reaction.

The constant-pressure heat capacity is also related to the entropy as

$$S_{T_2} - S_{T_1} = \int_{T_1}^{T_2} C_p d \ln T, \tag{8}$$

where S_{T_2} and S_{T_1} are the entropies at T_2 and T_1.

Often, the heat capacity is not measured directly but is obtained through manipulation of other measured values. As one example, the measured enthalpy change used to effect a change of a substance from one temperature to another (the enthalpy increment) will be divided by the difference of the two temperatures, and the resulting value will be termed heat capacity. Of course, this does not give the true differential shown in Eq. (4); it is only an approximation. Various schemes have been used to try to improve the approximation, but it nonetheless remains an approximation.

The interest in heat capacities arises primarily because they provide a visual magnification of the effects of physical phenomena on measured thermodynamic properties as indicated in Fig. 1. In Fig. 1(b) the enthalpy relative to an arbitrary temperature is shown for a substance undergoing a two-state effect (i.e., a physical behavior characterized by only two distinct energy states). Shown in Fig. 1(a) is the heat capacity for the system for the same temperature range. The enthalpy change for the two-state reaction is small compared to the total enthalpy of the material and so is not readily apparent in the figure showing enthalpy. However, it does appear in the picture of the heat capacity function. This visual magnification is the primary reason why measured enthalpy increments are converted into heat capacities.

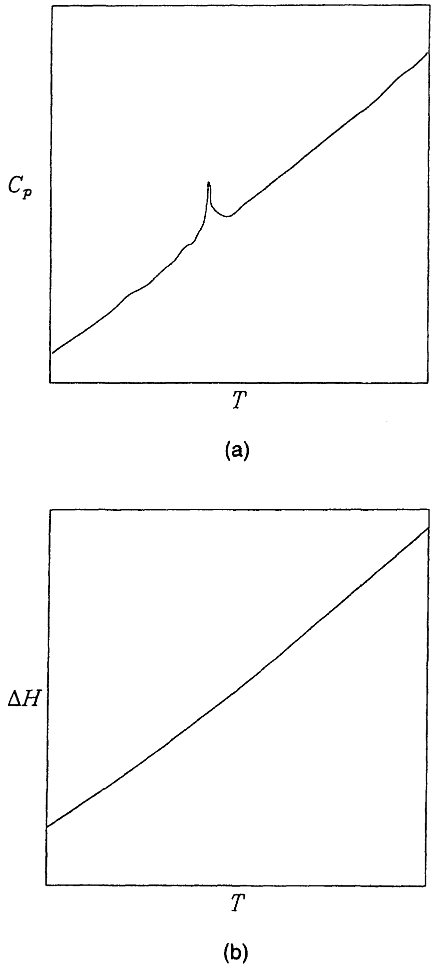

FIG. 1. Heat capacity and enthalpy functions for a substance that undergoes a tautomerization reaction. (a) Heat capacity against temperature. (b) Enthalpy against temperature; the enthalpy of the tautomerization two-state reaction is too small to be visible in the figure.

2. MEASUREMENT OF THERMAL PROCESSES

There exists a wide range of instrumentation for measurement of thermal processes, which reflects the variety of thermal processes encountered. This section provides general schemata for measurement of some thermal processes and a description of terminology associated with these instruments. The first division of classification is the type of process that is being measured. As all substances and reactions involve thermodynamic parameters, thermal processes can be quantified through measurements of heat, or energy, exchange with calorimeters. Other methods follow the change in some other property with respect to temperature; for example, differential thermogravimetric methods follow the change in mass of a sample as its temperature is changed and as a result the sample undergoes loss of a volatile component. Of all these methods the most diverse are the calorimetric methods.

2.1 Calorimetry

Calorimetry, at its simplest level, involves the measurement of the heat, or energy, exchange that accompanies some change in the sample. A calorimeter, the device that measures the heat, or energy, exchange can be considered as containing two separate thermodynamic systems, namely, the measured system and the surroundings. The measured system is comprised of the sample whose characteristics are to be determined plus any vessel or other means used to contain the sample. The measured exchange, or lack thereof, of the heat with the surroundings establishes the first set of adjectival descriptors of a calorimeter, namely, adiabatic, isothermal, and isoperibol. Occasionally, scanning is also included in this set.

Adiabatic conditions are those for which no heat or energy is exchanged between the measured system and the surroundings. Thus, under adiabatic conditions, any energy change undergone by the measured system must change its temperature, if the sample is contained either at constant pressure or at constant volume. (One exception is a substance undergoing a first-order phase change—e.g., melting.) Isothermal conditions are those in which the measured system's temperature is maintained constant, while it undergoes a change. In this case, all of the heat associated with the process is transferred between the measured system and the surroundings. Isoperibol refers to surroundings that are kept constant; usually the surroundings are maintained at a constant temperature. In this case, some portion, but not all, of the enthalpy change, or energy change, undergone by the measured system is transferred to the surroundings. All three of these conditions are limits that can be approached, but not exactly achieved, by careful design. Hence, "pseudo-adiabatic" or "quasi-adiabatic" have been used to describe imperfect adiabatic conditions in a calorimeter. Scanning refers to calorimeters in which a variable is changed continuously with respect to time. Usually this variable is temperature, but it can also be other parameters such as pressure or magnetic-field strength. Whereas the first three descriptors referred to the relation of the observed sample to its surroundings, "scanning" does not express this relation. As a result, scanning is not exclusive of the other three terms, so one can have, for example, an adiabatic scanning calorimeter.

Calorimeters can be either open or closed to the transport of mass through the measurement system. Calorimeters open to mass transport are "flow" calorimeters, whereas, those closed to mass transport are "static." Energy-mass equivalence prohibits an adiabatic flow calorimeter, in the strictest definition. The other two relations of surroundings to measured system are possible with flow calorimeters. Flow calorimeters find their importance in the measurement of thermal processes involving fluids.

Regardless of the relation of the surroundings to the measured sample, an energy equivalent of the measured system must be made. (The energy equivalent is the response of a measured variable with respect to known amounts of energy introduced into the calorimeter.) This is most often accomplished through an electrical calibration. Other determinations of the energy equivalent are also possible. For example, an energy equivalent can be realized by measuring the amount of substance undergoing a phase change, or some other thermal process, in the surroundings of a calorimeter.

Any apparent contradiction in the heating of a calorimeter that is described as adiabatic can be better understood by considering three thermodynamic systems: the measured system, a source of energy, and the surroundings to both. The measured system and the source of energy—for example, a battery that powers a resistive electrical heating circuit—are two systems both adiabatic relative to the third system, the surroundings. Heat, or energy, can be transferred from one system to the other in a controlled and measurable way, and thus one can heat an adiabatic calorimeter.

2.1.1 Static Calorimeters, Adiabatic

Adiabatic conditions are obtained through minimization of heat transfers from the measured system to the surroundings due to conduction, convection, and radiation. Heat exchanges due to conduction and convection are minimized by performing the measurements in a high vacuum and by reduction of thermal contact between the measured system and its surroundings through increased thermal resistance of electrical leads (i.e., by using very small-diameter or superconducting leads) and by the use of poor thermal conductors to support the measured system in the vacuum. The measured system is usually surrounded by a jacket whose temperature is maintained at that of the measured system to minimize radiative heat transfer. The electrical leads from the measured system are usually heat sunk to this jacket to reduce their conduction even further.

Because of the minimization of the heat transferred between the measured system and the surroundings, adiabatic calorimeters have the potential of great accuracy. They are often used in the determination of the relative enthalpy content of substances (the measurements of which are often reported in a heat-capacity form), especially for temperatures less than 400 K. The principal quantities measured in a determination of an enthalpy increment are the electrical energy used to heat the measured system and the temperature of the measured system as it is heated. Some error may arise as a result of incorrect accounting for power lost in the electrical leads of the calorimeter, creating small systematic biases in each particular calorimeter. However, this error diminishes with increasing sample size and so can be minimized in the characterization of calorimetric reference samples, which can then be used to determine the bias in a calorimeter designed for measurement of smaller amounts of samples. While the measurement of the heat or energy generated by current flow through a heating circuit is straightforward, determination of the temperature of the measured system is less so. Measurement of temperature (see TEMPERATURE) is accomplished by transfer scales. The primary temperature scales, and the subsequently transferred scales, are not perfect determinations of the true thermodynamic temperature. As such, they also may contribute to a systematic bias of measured thermal properties from the true values. Intercomparisons of enthalpy increment measurements with adiabatic calorimeters indicate that the better calorimeters have uncertainties in the range of 0.05% (Archer, 1995).

Adiabatic calorimeters have also been used to measure the energy change of a reaction occurring in either a constant-volume vessel, a common example of which is oxygen combustion calorimetry, or a constant-pressure reaction vessel such as a solution-reaction calorimeter. These are aspects of thermochemistry (*q.v.*) and will not be discussed here.

2.1.2 Static and Flow Calorimeters, Isothermal Two different techniques are used for the construction of isothermal calorimeters. In the first, the measured system undergoes a thermal process in a reservoir, or surroundings, of a substance that is undergoing a phase change. For example, one can place a system undergoing a thermal process in a reservoir containing both liquid and solid phases of a material, called the calorimetric substance. As long as both phases of the calorimetric substance coexist in the surroundings, the enthalpy transfer occurs at constant temperature. As the thermal process progresses, heat is transferred to or from the reservoir resulting in a change of the relative masses of its two phases. Determination of the mass of material converted from one phase to the other, combined with knowledge of the enthalpy of that transition, gives the change in enthalpy or energy of the measured system. These calorimeters have been used in the determination of the difference in enthalpy between two temperatures, one of which is the temperature of the calorimetric substance. These measurements are performed by establishing a constant temperature of the measured system and then rapidly transporting it into the isothermal calorimeter. This technique is capable of great accuracy. If sufficient attention is paid to detail, the limiting factor in the accuracy of this type of measurement can be determination of the temperature of the measured system before it is introduced into the calorimeter.

The second type of isothermal calorimeter is one in which the heat flux resulting from a change of the measured system passes through a measuring device between the measured system and a large thermal reservoir. The measuring devices are usually thermopiles or Peltier coolers (see THERMOELECTRIC DEVICES). It is difficult to arrange for all of the heat flux from the measured system to pass through the thermoelectric device. Compensation for the flux not transferred through the measuring device is usually accomplished by placement of a resistive heater on or near the measured system. The resistive heater is then used to calibrate the calorimeter. This calibration assumes that the same proportion of heat passes through the measurement device for both the electrical heater and the measured system. Imperfect matching of calibration and measured processes is often the limiting factor in the accuracy of these calorimeters. Several mass-flow calorimeters of this type have been constructed. Containment of fluids in small-diameter tubes that join and mix the flowing streams in the presence of an isothermal measuring device allows the determination of enthalpies of fast-reacting systems over a wide range of temperature. The accuracy of the best of these research instruments currently is on the order of 1%.

2.1.3 Static and Flow Calorimeters, Isoperibol In these calorimeters, the heat flux is between the measurement system and the constant-temperature surrounding and depends on the temperatures of both of them. In a static system, given sufficient time and in the absence of introduction of energy into the measurement system, a time-dependent heat flux between the measurement system and the surroundings exists and results in the temperature of the measurement system eventually reaching that of the surroundings. The heat flux, and its time dependence, between the measurement system and the isoperibol surroundings introduces further complications in matching a calibration energy to that of the measurement system. Yet the simplicity of creating constant-temperature surroundings, for example, by means of a constant-temperature bath or radiation shields controlled to the temperature of a large heat sink, makes these calorimeters attractive. Judicious care in the creation of a large thermal resistance between the static measurement system and the surroundings reduces the time rate of change of the heat flux between the two, which simultaneously improves matching the electrical calibration to the calorimetric response. In flow systems, constant-temperature surroundings do not cause the same time dependence of the heat flux between the measured system and surroundings. This is because the mass flow through the measurement system is equivalent to introduction of energy into the measurement system; thus a steady-state constant heat flux between measurement system and surroundings is established.

Isoperibol calorimeters are often used for determination of mixing enthalpies of fluid systems. Flow isoperibol calorimeters have been particularly useful in the determination of heat capacities of fluids.

2.1.4 Scanning Calorimeters Scanning calorimeters are those in which a process variable for the measured system is changed systematically with time, and some resultant thermal response is also monitored with respect to time. In most applications, the temperature is the altered process variable. However, pressure-scanning calorimetry has been reported, and the variation of applied magnetic or electrical fields on the measured system also constitute scanning calorimetry.

Temperature-scanning calorimetry appears in different designations: differential thermal analysis (DTA) and differential scanning calorimetry (DSC). However, the principle is similar. Typically, a small thermal resistance separates the measured system from the surroundings. The temperature of the surroundings is ramped with respect to time. As the temperature of the surroundings changes, the measured system's temperature also changes in the same direction because of the small thermal resistance. The heat flux into the measured system is derived from measurement of a thermoelectric device. Commercially available instruments are capable of measuring thermal effects for very small samples with an accuracy of a few percent; some claims of 1% accuracy with commercial instruments have been made. Although the accuracy of most of these instruments is not high compared to other types of calorimeters, they do have certain advantages. The combination of small sample size and rapid thermal response allows studies of kinetic effects, activation effects, and thermodynamically unstable systems, such as supercooled and superheated fluids.

3. THERMAL PROCESSES

3.1 Phase Changes

The variety of phase transitions observed in different systems underscores the vast scope of the study of thermal processes. Consideration of the transition between states, rather than the physics of phase equilibria, provides the focus of the study of thermal processes. These transformations are often most easily viewed by changing sample temperature.

First-order phase changes are those for which the chemical potential is continuous with respect to controlled variables such as temperature and for which the first derivatives of the chemical potential with respect to that variable are discontinuous. Under the conditions of this type of phase transition, energy is required to transform mass from one phase to the other without a change in the variables that characterize the transition. Therefore, second derivatives with respect to temperature, i.e., the heat capacities, are infinite at this transition. These phase transitions include changes in lattice structure, melting, vaporization, and sublimation.

Second-order phase changes are those for which both the chemical potential and its first derivative are continuous with respect to a variable. The second derivatives of the free energy show a discontinuity, but often are not infinite, at the transition. Higher-order thermal functions—e.g., dC_p/dT—do show an infinity. One example of a second-order phase change is the order–disorder transition that occurs in solid solutions.

Many phase transitions occur in solids without structural rearrangement of atoms. Often, these phenomena involve the electrons in the material, as for instance in magnetic transitions, superconductivity, or the excitation of electrons into conducting states. Examination of all of these effects may involve the intrinsic thermal properties of the material. Because the heat capacity visually magnifies changes in the thermal properties of a material with respect to temperature, the theoretical exposition of the effect of physical phenomena on thermal properties is often conducted in terms of heat capacities.

3.2 Atomic Vibration and the Heat Capacity of a Solid

The change in temperature of a solid near room temperature is, in principle, a simple thermal process. The associated enthalpy function or its temperature derivative, the heat capacity, reflects the motion of the atoms in the material. We first consider the case where the interatomic potential is similar for all of the atoms in a crystalline solid. Atoms in these crystalline solids arrange themselves in periodic structures and oscillate about their equilibrium positions. This vibration of atoms in these solids, for temperatures near room temperature, can be adequately approximated as a classical, harmonic oscillator, analogous to a pendulum or a mass on a spring. That is to say, the interatomic potential (sketched in Fig. 2) can be approximated with a parabola for small-amplitude displacements from equilibrium. A mass in a classical harmonic oscillator vibrates about an equilibrium position with an amplitude that depends upon the energy of the system. While the true potential experienced by atoms in solids is inherently asymmetric, the potential near to the atoms' equilibrium positions can be well approximated as harmonic. These three-dimensional oscillators then have an energy per atom of $3k_bT$, where k_b is the Boltzmann constant. The heat capacity of many solids near room temperature is found to be close to $3rNk_b$, or approximately $25 \cdot r$ J/(K mol) (where r specifies the number of atoms that constitute the stoichiometry of the material). Deviations of the values of observed heat capacities from this classical-mechanical Dulong and Petit constant provide insight into many thermal processes that occur in solids. In Fig. 3, two such deviations are illustrated that are intrinsic to all solids:

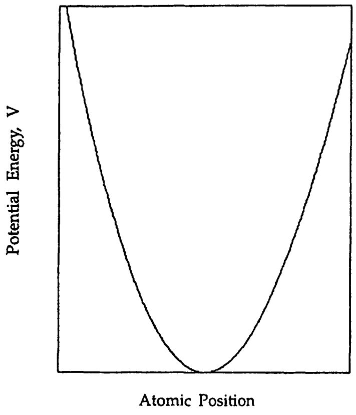

FIG. 2. Interatomic potential; potential energy against atomic position.

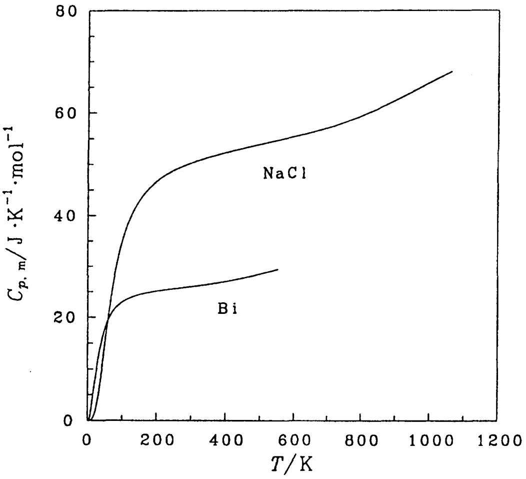

FIG. 3. Molar heat capacities of sodium chloride and bismuth.

1. Heat capacities generally approach zero as temperature approaches absolute zero, and
2. a monotonic addition to the Dulong–Petit constant is observed in the heat capacity of solids at temperatures generally near and/or above room temperature.

The observation (cf. Fig. 3) that heat capacities tend to vanish at low temperature reflects the quantum nature of harmonic oscillations of the atoms in a solid. In the quantum harmonic oscillator, instead of a continuous energy spectrum ($3k_bT$, for all $T \geq 0$), only discrete, quantized values of the vibration energy, E_n, of atoms occur. These energies are evenly spaced, with $E_n = (h/2\pi)\omega(n + 1/2)$, where n is any integer $n \geq 0$ and h is Planck's constant. This theory indicates that the vibrating atoms in a solid have a finite excitation energy, $(h/2\pi)\omega$. Therefore, as the average energy of the atoms in a solid decreases with decreasing temperature, it becomes less and less probable that an individual oscillator will be excited (i.e., have a nonzero value of n). Quantum mechanics indicates that the contribution to the heat capacity of a solid from vibrational energy goes to zero at low temperatures, as is observed. Whether these vibrations are modeled as isolated oscillators (Einstein first advanced this theory) or whether the collective vibrational nature of atoms within the crystal lattice is considered (as in the more sophisticated model of Debye, where the heat capacity is found to be proportional to the cube of the temperature at low temperatures), qualitative agreement with experimental thermal measurements is found. Quantitatively, the Einstein theory did not provide the correct thermal behavior as temperature approached 0 K because of its approximation that the frequencies of all atomic vibrations in a crystal were identical. The Debye theory predicted the correct temperature dependence of the thermal properties of atomic lattices at low and high temperatures (Dulong and Petit value of the heat capacity) but did not work as well at intermediate temperatures. Despite limitations of the Debye theory, it has the attractive quality of introducing a single parameter, the Debye characteristic temperature (Θ_D), that characterizes the thermal properties of a lattice. Most crystal lattices begin to show deviations from the Debye theory for temperatures greater than $\Theta_D/50$. Theories that treat lattice vibrations more rigorously than the Debye theory are available (see LATTICE VIBRATIONS, STATISTICS OF).

Generally it is an approximation to consider the atomic potential in solids to be harmonic, though, remarkably, deviations from harmonicity are generally small. Interatomic potentials, as functions of the interatomic distances, $V(r)$, are represented as

$$V(r) = ax^2 - bx^3 - cx^4 + \ldots, \tag{9}$$

where generally $b \ll a$. The x^3 term means that the potential is asymmetric. The first-order approximation is the harmonic potential, $V(r) = ax^2$. However, as energies increase and the amplitude of vibration of the atoms in the solid increases, the additional terms in the expansion for $V(r)$ become more important. The heat capacity becomes larger than would be expected on the basis of a harmonic oscillator; the extra contribution is approximately linear in temperature. In fact, a constant increase in the heat capacity with respect to temperature at temperatures above room temperature is a common observation. At temperatures approaching the melting point, equilibrium vacancy creation is generally believed to be the cause of a nonlinear portion of the heat capacity against temperature that has been observed for many substances (cf. Fig. 3.) Vacancy creation and annihilation are discussed below.

The contribution of atomic motion to the thermal properties of crystalline materials becomes significantly more complicated when one considers systems in which the interatomic potentials become increasingly anisotropic. Extreme anisotropy, or directionality, of interatomic potentials occurs in molecular crystalline solids such as frozen oxygen, methane, and larger organic molecules. In these materials, atomic motions are affected by both strong intramolecular potentials and weaker intermolecular potentials. Additionally, motions of groups of atoms, and of the molecule as a whole, is affected by intermolecular potentials. The total of these possible motions results in a wealth of thermal behaviors. One example is the free rotation of methane molecules within the lattice, which occurs cooperatively as the temperature increases through 20 K. This effect results in a second-order phase transition, indicated by a lambda peak in the heat capacity near 20 K, and is easily seen with calorimetric measurements, first reported by Clusius (1929), as shown in Fig. 4.

3.3 Electronic Effects

A variety of electronic effects are possible in solids. These effects are almost always accompanied by a thermal signature readily detected by calorimetry. Some electronic effects important for applied physics are described here.

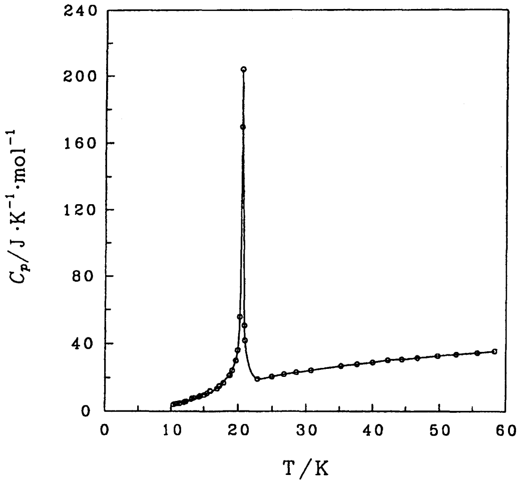

FIG. 4. Molar heat capacity of methane; measurements from Clusius (1929).

The heat capacity of electrically conducting materials includes a contribution from lattice vibrations, as outlined above, and an additional contribution from changes in electron energies. Near room temperature, only a tiny fraction of the electrons can occupy the higher energy states, resulting in a relatively small contribution to the heat capacity (see STATISTICAL MECHANICS, CLASSICAL, particularly discussion of Fermi–Dirac statistics). The heat capacity of metals is similar to that of insulators near room temperature. However, at temperatures of roughly $T < \Theta_D/50$, the intrinsic heat capacity of a conductor is given by

$$C = AT + BT^3. \tag{10}$$

As the contribution BT^3 from the lattice goes to zero with decreasing temperature, the heat capacity of the electrons, AT, becomes a more significant fraction of the total heat capacity and is best extracted from low-temperature thermal measurements. At higher temperatures, the conduction electrons also make a contribution proportional to temperature to the heat capacity. This contribution at higher temperatures occurs in combination with the similar vibrational contribution proportional to temperature that arises from the anharmonicity of the interatomic potentials. Thus, at higher temperatures, separation of these two contributions from measurements of thermal functions requires additional knowledge.

In considering the temperature dependence of physical processes, the quantity of interest is essentially the logarithm of thermodynamic temperature. In other words, a range of physical changes undergone by a solid substance from 100 to 1000 K would be comparable to a different set of physical changes that might be undergone from 10 to 100 K, from 1 to 10 K, and so on. As such, distinct variations from the behavior typified by Eq. (10) are often observed at yet lower temperatures. Thus, the low-temperature regime constitutes a vast expanse for the observation of low-energy thermal processes. Consider copper as an example. The heat capacity of copper at 1 K is 1/100 000 times its room-temperature value. Yet below this temperature there remains two additional orders of magnitude of temperature variation readily accessible by means of calorimetric measurements. For instance, in copper samples with less than 4 ppm concentrations of any magnetic impurities, a distinct anomalous contribution [$\Delta C = 3 \times 10^{-5}T^{-2}$ mJ/(K·mol)] to the heat capacity is observed below 0.2 K. This contribution to the heat capacity is as large at 0.03 K as values expected from Eq. (10) and may be due to a small ($\approx 4 \times 10^{-6}$ eV) Schottky-type splitting of an impurity electron state. A Schottky peak in the heat capacity is a distinct signature of discrete energy states and the thermal processes associated with their populations. A clear example of a Schottky peak is provided by the heat capacity of $CeS_{1.457}$ at temperatures between 0.5 and 20 K. Subtraction of the lattice and electronic contributions to the heat capacity, as estimated on the basis of Eq. (10) and the higher-temperature data, revealed a peak whose shape was consistent with a ground-state doublet with energy splitting 0.25 meV and corresponded to an entropy change of R ln2 = 5.8 J/mol K. This effect is shown in Fig. 5. The occupation of these energy levels is clearly an important thermal process in this temperature range.

Calorimetry is an effective way to study superconductivity. In fact, as epitomized by Methfessel *et al.* (1980), "Caloric measurements are the sole reliable way to ascertain the bulk superconductivity of a given compound or alloy." As the electrons in a material begin to condense into a coherent superconducting state at a critical temperature, with corresponding zero dc resistance and the expulsion of magnetic fields, the heat capacity shows a distinct jump. An example is provided in Fig. 6 for gallium. This change is consistent with a decreasing entropy associated with the ordering of the state. The contrast with the heat capacity in the absence of the superconducting transition, suppressed with a magnetic field, is apparent. These observations are important both in monitoring and in understanding the superconductivity of a material, because essential parameters such as the energy gap can be determined from such measurements.

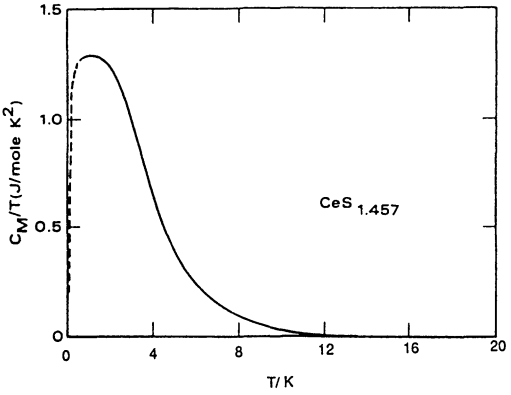

FIG. 5. The heat capacity due to a Schottky effect in $CeS_{1.457}$ (reproduced from Ho *et al.*, 1982).

There is a wealth of other electronic effects that occur with changing temperature. They involve various magnetic behaviors and low-energy electronic behaviors. The interested reader is referred to the books on these subjects listed below.

3.4 Thermal Expansion

Any change in the dimensions of a sample with a change in temperature is known as thermal expansion. A change in the volume

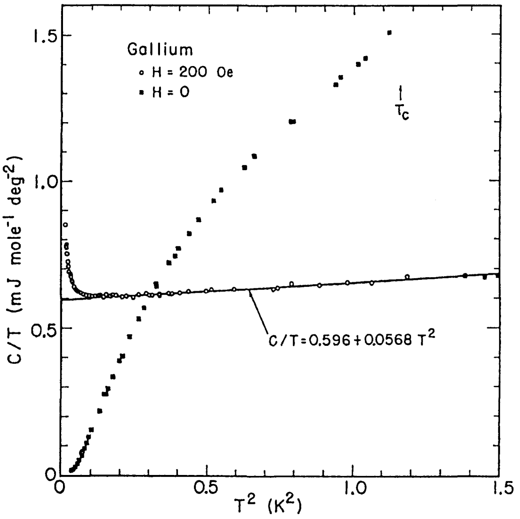

FIG. 6. The low-temperature heat capacity (divided by temperature) of gallium at two magnetic-field strengths vs the square of the temperature. Measurements of the superconductive state are the solid squares. The open circles are for the normal state (nonsuperconductive). (Reproduced from Phillips, 1964.)

of a material due to a change in temperature is a profound effect of the anharmonicity of the interatomic potentials in solids. For an infinitesimally small change in temperature of a substance, the ratio of the corresponding change in volume ΔV to the change in temperature ΔT, divided by the substance volume V, is the coefficient of thermal expansion β:

$$\beta = \frac{1}{V}\left(\frac{\Delta V}{\Delta T}\right)_p. \tag{11}$$

The subscript p indicates that the pressure is held constant. The thermal expansion of solids reflects changes in the configuration of the atoms in the material. Most substances expand with increasing temperature, reflecting the asymmetry in the interatomic potential due to the short-range atomic repulsion. As the temperature of a solid increases and the atoms in the solid vibrate with larger amplitudes, the average separation of the atoms, and the sample volume, increases.

The linear coefficient α of the thermal expansion of solids,

$$\alpha = \frac{1}{l}\left(\frac{\Delta l}{\Delta T}\right)_p, \tag{12}$$

is of practical consideration. Here l is the length of a sample, and Δl is the change in its length corresponding to an infinitesimally small change in temperature, ΔT. For isotropic solids $\beta \cong 3\alpha$. The values of α and β for solids are small. For example, common steel has a value of $\alpha = 11 \times 10^{-6}\ \mathrm{K}^{-1}$. A change in volume, $\Delta V = (l + \Delta l)^3 - l^3$, for an isotropic solid cube can be simply related to a change in length Δl with $\Delta V/V$ very close to $3\Delta l/l$, since α, and thus Δl, is small. Therefore $\beta = 3\alpha$ is an excellent approximation. Thermal expansion is important in the design of structures and devices. Of particular concern is differential thermal expansion: When two materials with different values of α are in contact, allowance must be made for the relative movements of these materials upon changes in temperature, or the relative movement of the objects must be constrained. Such design considerations are found in such very different scales as between the girders in skyscrapers and in the 60-μm space between chip and board in microelectronics packages. Examples of the differences of thermal expansion of different metals are shown in Fig. 7.

3.5 Quantification of Thermal Processes

While the subject of thermal processes clearly includes atoms in the liquid and vapor phases, many aspects may be illustrated with a continued examination of the motion of atoms in a solid. For instance, many processes are elucidated by the consideration of a relatively simple picture of atomic jumps between minima in the atomic potential energy in the solid; these jumps may well constitute diffusion. This simple picture is sketched in one dimension in Fig. 8. Models of atomic vibration, thermal expansion, defect production and annihilation, diffusion, and many aspects of phase transitions can be related to this picture. Many thermal processes can be characterized as thermally activated (or Arrhenius), such as in Eq. (14). In this case it is often assumed that the reaction rate $d(n/N)/dt$ is equal to the product of separable functions $f(n/N)$ and $g(T)$:

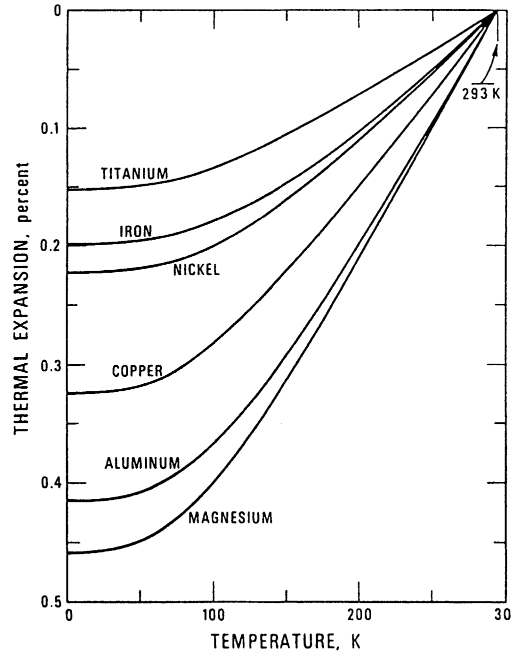

FIG. 7. Thermal expansion of some metals vs temperature (reproduced from Clark, 1983).

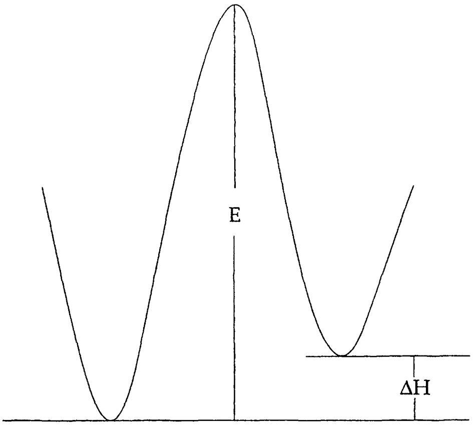

FIG. 8. Reaction-coordinate diagram showing the relation of equilibrium states of products and reactants to activation energy. The activation energy is *E*, and the thermodynamic energy or enthalpy difference between the two equilibrium states is *ΔH*. The two minima represent the equilibrium energies of products and reactants.

$$\frac{d(n/N)}{dt} = f(n/N)g(T), \tag{13}$$

where the function $g(T)$ depends on the temperature T and the function $f(n/N)$ depends on the fraction of the sample, n/N, that has reacted. The function $f(n/N)$ reflects the fundamental mechanisms of the reaction, while generally $g(T)$ is assumed to be characterized by an activation energy E_A and a constant g_0,

$$g(T) = g_0 \exp(-E_A/kT). \tag{14}$$

This formalism is used in the characterization of thermal processes as diverse as steel processing and electron drift in semiconductors. Nevertheless, the simple Arrhenius form of Eq. (14) is not always applicable. A determination of the appropriateness of Eqs. (13) and (14) for a particular reaction may be made from analysis of reaction rates obtained from calorimetric data. Additional quantitative study by measurements as diverse as electrical resistivity and diffraction of x rays and electrons, as functions of heat treatment, can yield insight into the forms of $f(n/N)$ and $g(T)$. The function $f(n/N)$ is best characterized on the basis of isothermal experiments, in order to separate out $g(T)$. The dependence of reaction rate on temperature can be deduced by comparing reaction rates at equal values of degree of conversion, n/N, as a function of temperature. In the case of reactions that can be characterized by Eqs. (13) and (14), a well-known technique advanced by Kissinger (1956) and more recently discussed by Ozsawa (1970) provides a simple identification of points of equal degree of conversion n/N for calorimetric data of different heating rates. Thus, the function $g(T)$ can be analyzed independently of $f(n/N)$. Characterization of these thermal processes is very important for many areas in applied physics, such as electronics processing, metals working, and polymer sets.

3.6 Defects in Materials

Materials processing often concerns the production or annihilation of defects in polycrystalline materials. The properties of materials depend strongly on their defect structure, grain size, and impurity and defect content. As defect structures are generally nonequilibrium, thermal processes related to these defects generally concern the agglomeration and/or annihilation of defects. Crystal defects include point defects, impurities, dislocations, and different grain-boundary structures. In the production of silicon crystals for integrated circuits, great care is expended to limit the number of most defects, while introducing a small number (e.g., 5 ppm) of special impurity defects. This is in contrast to the thermal processes associated with the production of some structural alloys, where high-temperature defect structures are quenched into the room-temperature material to increase hardness.

Crystals contain a finite density of point defects and vacancies in thermal equilibrium at temperatures above 0 K. Consideration of the concentration of these defects, and its variation with temperature, provides a better understanding of nonequilibrium defect densities, and the thermal processes associated with their annihilation and creation. The free-energy change ΔG for the formation of n vacancies in a particular amount of material, given by

$$\Delta G = n\Delta H - T\Delta S, \tag{15}$$

is minimized (condition of equilibrium) with a finite density of vacancies, where ΔH is the enthalpy of formation of a vacancy and ΔS

corresponds to the entropy change realized with the creation of n vacancies. This entropy change includes the configurational entropy associated with the degeneracy of energy states that correspond to the different possible locations of the vacancy. The equilibrium concentration, $c = n/N$, of vacancies (assuming, as is observed, that n is much less than the number of atoms in the crystal, N) is given by

$$c = A_v \exp(-\Delta H/kT), \tag{16}$$

where k is the Boltzmann constant.

The observed concentration of vacancies in solids is often greater than the equilibrium concentration given by Eq. (16), depending also on the thermal and mechanical history of the sample. The thermal process by which a crystal adjusts its vacancy concentration and distribution with a change in temperature involves long-range diffusion. These diffusion processes can be modeled by consideration of the movement of either the vacancy or an entity that borders the vacancy. For equilibrium to be realized, vacancies must be able to sample a large number of possible sites. To accomplish this, atoms must be able to move, or diffuse, through the sample. This requires a finite amount of time that depends on the sample temperature, size, atomic configuration, and chemistry. Thus it is possible to quench a nonequilibrium concentration of vacancies into a sample where the decay rate of these vacancies can be negligible at room temperature. Clearly this recovery rate is proportional to the diffusion constant D of atoms, where generally

$$D = D_0 \exp(-E_a/kT). \tag{17}$$

The activation energy E_a for diffusion will generally be greater than ΔH, because E_a reflects the highest-energy state an atom must attain in order to move (for instance E in Fig. 8) rather than the energy difference of final states reflected in ΔH.

The transformation of a polycrystalline solid with a variety of defects into a single crystal with a minimum of defects involves a number of different thermal processes, which are divided into the two general categories of recovery and recrystallization. Recovery concerns the annihilation of specific defect structures such as point defects and dislocations, while recrystallization includes the nucleation and growth of new, generally larger grains in a lower-energy configuration. Recovery processes generally have lower activation energies than those of recrystallization and thus are observed in a distinctly different, lower-temperature regime. Measurements can be made of the heat flow from a sample during recovery and recrystallization as in a scanning calorimeter. These heat flows, measured as a function of temperature, typically show two general peaks: the lower-temperature effect is associated with recovery, and the higher-temperature peak is associated with recrystallization. Individual processes, such as dislocation annihilation or defect–vacancy combination, may result in specific subpeaks. An example of heat flow vs temperature is shown in Fig. 9. The total integrated heat release provides a measure of the energy associated with the defects. Energies of up to 0.1 eV/atom have been observed, and so, clearly, this is an important thermal process.

Recovery and recrystallization are important thermal processes in the consideration of solids and materials processing. Extended defects, such as grain-boundary structures and other dislocations, are generally introduced into solids by mechanical means, such as by deformation during working, by ion bombardment, or by quenching from higher temperatures. Often a material's properties are enhanced by a given defect structure.

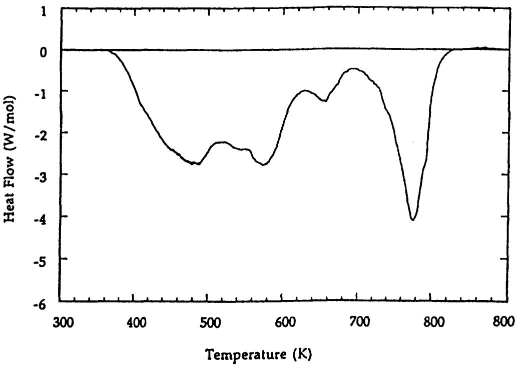

FIG. 9. Heat flow for conversion of a polycrystalline solid into a single crystal. The large heat flow centered near 520 K is due to recovery; the large heat flow centered near 775 K is due to recrystallization.

The processes of recovery and recrystallization also provide some insight into the nature of the defect structures themselves. Characterization of rate constants, including activation energies for these processes, can be related to specific atomic processes.

Another example of a thermal process, the crystallization of a metallic glass, is indicative of the relation of atomic motion to thermal processes. In Fig. 10, the heat flow from a NiZr metallic glass is plotted as a function of temperature for a constant heating-rate anneal. A broad, low-level heat release between temperatures of 450 and 600 K is observed and can be contrasted to the distinct crystallization peak at a temperature of 640 K. The lower-temperature process is structural relaxation, i.e., the progression of the glassy matrix to a denser glass with a more relaxed short-range order. This process involves atomic motion, but it does not involve the long-range collective motion of all the atoms in the glass, which would result in crystallization of the glass. A similarity with the recovery and recrystallization processes exists, in which some simple processes can occur at lower temperatures, whereas longer-range, collective rearrangement of atoms associated with recrystallization occur at higher temperatures.

A final example of a thermal process in the Ni–Zr system illustrates the importance of thermal processes in the fabrication of materials. Interdiffusion between pure, polycrystalline metals at relatively low temperatures (approximately half the pertinent melting temperatures) and over short distances can result in the formation of metastable, metallic glasses. In this example, multilayered Ni/Zr composites were produced by vapor deposition of polycrystalline layers of Ni and Zr in equal thicknesses of 45 nm. A measurement of the heat flow (Fig. 11) from such a composite reveals two separate reactions: The first, near 580 K, was related by means of structural analysis to the interdiffusion and amorphization of the original multilayers, while the second reaction, near 830 K, was correlated with the subsequent crystallization of the amorphous phase. These temperature ranges can be compared to those observed for recovery and recrystallization.

Metallic glass may form by solid-state reaction because a disparity in the mobility of the different atomic species in the diffusion couple provides a kinetic constraint on the formation of equilibrium intermetallic compounds. In the temperature range (≈400 to ≈650 K) where solid-state amorphization reactions (SSAR) are commonly observed in the Ni–Zr system, the relative diffusion coefficients of Ni and Zr in the growing amorphous phase differ by several orders of magnitude, with Ni (approximately 22% smaller atomic radius than Zr) being the faster diffusing species.

The DSC scan of Fig. 11 provides a direct measurement of the large negative enthalpy

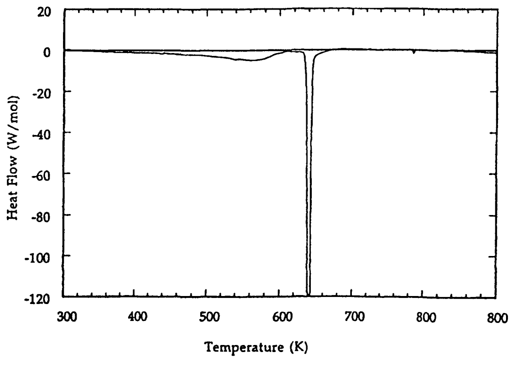

FIG. 10. Heat flow for relaxation and crystallization of a NiZr metallic glass. The lower-temperature process is due to rearrangement of the glass matrix. The higher-temperature process is due to crystallization of the glass.

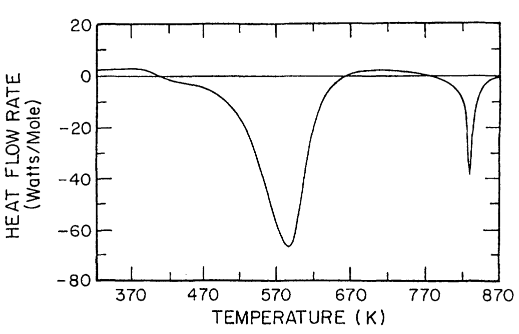

FIG. 11. The heat-flow rate as a function of temperature for a sputtered, multilayered thin film of average stoichiometry $Ni_{68}Zr_{32}$ measured at a constant scanning rate of 10 K/min. The thin film is comprised of equal-thickness (300 Å) layers of Zr and Ni with half-layer Zr caps.

THERMAL PROPERTIES OF GASES

J. H. DYMOND, *Chemistry Department, The University of Glasgow, Glasgow, U.K.*

Introduction 209
1. **Thermal Expansion** 210
2. **Heat Capacity** 210
2.1 Measurements at Constant Volume 211
2.2 Measurements at Constant Pressure 212
2.3 Ideal-Gas Values 213
2.3.1 Speed-of-Sound Measurements 213
2.3.2 Calculation 213
2.3.3 Representation 216
2.3.4 Estimation 216
2.4 Effects of Molecular Interactions 216
2.5 Critical Region 216
3. **Joule–Thomson Effect** 217
4. **Thermal Conductivity** 218
4.1 Measurements 218
4.1.1 Concentric-Cylinder Method 218
4.1.2 Parallel-Plate Method 219
4.1.3 Transient Hot-Wire Method 219
4.1.4 Experimental Results 221
4.2 Theoretical Description for Monatomics 221
4.3 Theoretical Description for Polyatomics 222
4.4 Correlations for Dense Gases 222
4.5 Critical Region 222
5. **Thermal Diffusion** 223
6. **Thermal Diffusivity** 224
7. **Thermal Transpiration** 224
Glossary 224
Works Cited 225
Further Reading 226

INTRODUCTION

Accurate knowledge of thermal properties of gases is of fundamental importance in chemical technology. In reactions involving gases, the heat capacity is an important property for the determination of the Gibbs energy change under the experimental conditions, which provides the answer to whether the reaction is thermodynamically feasible under those conditions. Also, it is essential for the accurate calculation of equilibrium constants. This article commences with a short discussion of these applications and then describes methods for the measurement of the two heat capacities, the isochoric (C_V) and isobaric (C_P) heat capacities. Values of C_P under pressure are thermodynamically related to ideal-gas values, which can be determined from speed-of-sound measurement and, importantly, can be accurately calculated from spectroscopic information using the methods of statistical thermodynamics. The success of this approach has resulted in many "experimental" values being obtained in this way. Methods are described for their temperature representation, and for their estimation using group-contribution methods. Finally, in this section, the effects of molecular interaction at higher pressures and the behavior of heat capacities in the critical region are described.

The success of a chemical process plant depends on its optimum design, and here thermal properties are required for heat-transfer considerations. Besides the heat capacity, of prime importance is the thermal conductivity. In Sec. 4, methods are described for its measurement, its calculation for low-pressure monatomic and polyatomic gases, and its correlation at high pressure. Behavior in the critical region is also described. Thermal diffusion, thermal diffusivity, and thermal transpiration are briefly discussed.

Liquefaction of gases occupied the minds of many famous scientists. One method is by application of the Joule–Thomson effect, which must be understood to appreciate why a preliminary cooling of a gas, in the case of helium to very low temperatures, is neces-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

sary before liquefaction can occur by this method.

As well as being of technological importance, these properties also have provided valuable information for the development of our detailed understanding of the interactions between molecules.

1. THERMAL EXPANSION

The isobaric coefficient of thermal expansion, or isobaric thermal expansivity, α, is the fractional change in volume with change in temperature at constant pressure,

$$\alpha = \frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_P. \tag{1}$$

For an ideal gas, the P, V, T behavior is given by the relationship

$$PV = nRT, \tag{2}$$

where n is the number of moles of gas and R is the gas constant, 8.314 J K^{-1} mol^{-1}, and the isobaric coefficient of thermal expansion is then equal to $1/T$. At 300 K, therefore, α for an ideal gas has the value 3.3×10^{-3} K^{-1}.

The behavior of a real gas differs from that of an ideal gas because of molecular interactions. At relatively large separations R, a pair of real molecules will experience a long-range attractive interaction that falls off as R^{-6}, and, at close distances, there is a steep repulsive interaction energy that falls off exponentially with distance. The simplest model that takes into account these features of short-range repulsion and long-range attraction, by approximating them as a hard-core repulsion and a very weak very long-range attraction, is that of van der Waals (1873), whose equation for P, V, T is

$$(P + a/V_m^2)(V_m - b) = RT, \tag{3}$$

where V_m is the molar volume, b is the excluded volume due to the finite size of the hard-core molecules, and the a/V_m^2 term accounts for the attractive interaction.

For a van der Waals gas, α is given by the expression

$$\alpha = \frac{R}{V_m(P - a/V_m^2 + 2ab/V_m^3)}, \tag{4}$$

which gives the ideal-gas result when $a = b = 0$ (ideal gas) or V_m becomes very large, as for a real gas at low pressures.

Historically, measurements on gas expansion were important in defining the ideal-gas temperature scale. Here, the coefficient of expansion α_e was considered as the increase in volume at constant pressure for a one-degree temperature rise as a fraction of the initial volume of gas at 0 °C. Thus,

$$V_t = V_0(1 + \alpha_e t), \tag{5}$$

where t °C is the final temperature. For real gases, α_e extrapolated to zero pressure has the same value of 1/273.15 $°C^{-1}$. By defining the unit of temperature on this scale as the kelvin, the ice point became 273.15 K (and the triple point of water was set equal to 273.16 K). Constant-pressure air thermometers, and also constant-volume thermometers that are based on the linear variation in pressure with temperature change for a fixed mass of gas, have historical interest.

As pressure increases, α_e and hence α, which equals $(t + 1/\alpha_e)^{-1}$, differ from the ideal-gas value. There is an initial decrease at high reduced temperatures, where the temperature is divided by the critical temperature, for gases such as hydrogen and helium that exhibit small attractive interactions, but for the majority of gases the effect of attractive forces is to produce an increase in α, as expected from the van der Waals equation.

α becomes infinite at the critical point. The van der Waals equation predicts that, at the critical density, α would increase as $(T - T_c)^{-1}$, where T_c is the critical temperature, but in fact it increases even more rapidly.

2. HEAT CAPACITY

Heat-capacity values are essential in thermodynamics for calculating changes in entropy S and enthalpy H between different temperatures for pure gases, and also for determining the corresponding changes in these functions in reactions of gases. In order to determine whether, thermodynamically, a reaction can occur, it is necessary to calculate the Gibbs energy change ΔG, given by $\Delta H - T\Delta S$; ΔG must be negative. The equilibrium constant K is determined from the value for the Gibbs energy change ΔG^0 when all species are in their standard state, which, for a gas, is the (hypothetical) state where ideal-gas behavior is exhibited at the standard pressure (10^5 Pa):

$$\Delta G^0 = -RT \ln K. \tag{6}$$

In chemical engineering, heat-capacity data are required in problems of energy transfer and for making energy balances. Experimentally, measurements for gases are made under conditions of constant volume (isochoric heat capacity, C_V) or constant pressure (isobaric heat capacity, C_P). These heat capacities are related to the thermodynamic functions internal energy, denoted by U, and enthalpy,

$$C_V = \left(\frac{\partial U}{\partial T}\right)_V \quad \text{and} \quad C_P = \left(\frac{\partial H}{\partial T}\right)_P. \tag{7}$$

The relationship between these heat capacities is given by

$$C_P - C_V = T\left(\frac{\partial P}{\partial T}\right)_V\left(\frac{\partial V}{\partial T}\right)_P = \frac{\alpha^2 VT}{\kappa_T}, \tag{8}$$

where α is the isobaric coefficient of thermal expansion and κ_T is the isothermal coefficient of compressibility, $-(1/V)(\partial V/\partial P)_T$.

Changes in U, H, and in entropy S when 1 mole of a real gas goes from pressure P_1 and volume V_1 at temperature T_1 to pressure P_2 and volume V_2 at temperature T_2 are as follows:

$$U(T_2,V_2) = U(T_1,V_1) + \int_{T_1,V_1}^{T_2,V_1} C_V(T,V_1)dT + \int_{T_2,V_1}^{T_2,V_2}\left[T_2\left(\frac{\partial P}{\partial T}\right)_v - P\right]dV, \tag{9}$$

$$H(T_2,P_2) = H(T_1,P_1) + \int_{T_1,P_1}^{T_2,P_1} C_P(T,P_1)dT + \int_{T_2,P_1}^{T_2,P_2}\left[V - T_2\left(\frac{\partial V}{\partial T}\right)_P\right]dP, \tag{10}$$

$$S(T_2,P_2) = S(T_1,P_1) + \int_{T_1,P_1}^{T_2,P_1} \frac{C_P(T,P_1)}{T}dT - \int_{T_2,P_1}^{T_2,P_2}\left(\frac{\partial V}{\partial T}\right)_P dP. \tag{11}$$

For U and H there is no contribution from the final term for an ideal gas; for entropy, the contribution arising from the pressure change in this case is $R \ln (P_2/P_1)$ per mole.

2.1 Measurements at Constant Volume

Heat-capacity measurements at constant volume may be made by adiabatic calorimetry. In outline, a sample of known mass (and, hence, number of moles n) is confined to a vessel. Since measurements are usually made over a range of pressure, the vessel is referred to as a "bomb." A precisely measured amount of electrical energy, Q, is applied, and the resulting temperature rise ΔT is accurately measured. In a separate experiment, the amount of energy Q_0 to heat the bomb by the same ΔT is determined, and the molar heat capacity of the gas is then obtained from the relationship

$$C_v = (Q - Q_0)/n\Delta T. \tag{12}$$

The experiment can be repeated with the same charge of gas at different temperatures to give a set of isochoric heat capacities. At the end of the run some gas can be released and further measurements made at a lower density. In this way, a series of $C_V(\rho)$ isotherms can be obtained. In each case, the number of moles may be calculated from the measured bomb volume and an accurate equation of state for the gas that gives the density.

Figure 1 gives a diagram of the apparatus of Goodwin (1961) at the former National Bureau of Standards, Boulder, Colorado (now the National Institute of Standards and Technology), recently used by Magee (1994). The 5-cm-diameter type 316 stainless steel calorimeter has a Constantan wire heater varnished directly onto the sphere. The copper case protects the heater and serves as a reference temperature for two thermopiles, 1, which control a guard ring and an adiabatic shield. Temperature control is also aided by thermocouples 2, 3, and 4. The electrical energy is given basically by the product of potential and current and the heating interval. Details of the corrections to be applied for accurate results are given by Goodwin and Weber (1969a). As an example of C_V values, Goodwin and Weber (1969b) found for oxygen that between temperatures

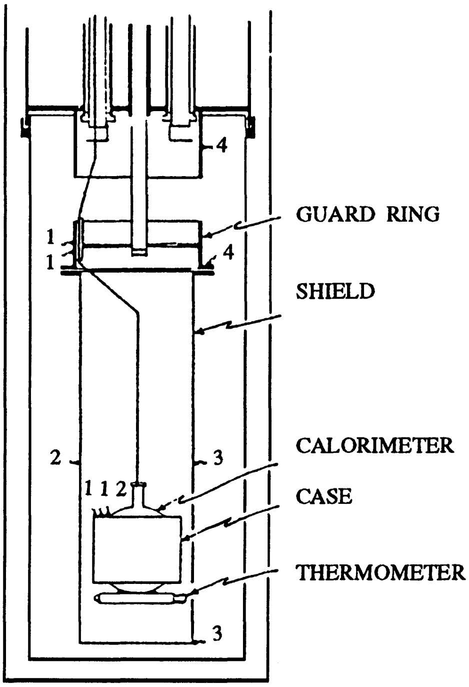

FIG. 1. Constant-volume adiabatic calorimeter, from Goodwin (1961), courtesy of National Bureau of Standards, Boulder, Colorado. 1, thermopiles; 2,3,4, thermocouples.

of 210 and 155 K, C_V showed a small increase at densities of 6.10 mol L^{-1} (22.6 to 27.8 J K^{-1} mol^{-1}) and 20.2 mol L^{-1} (24.3 to 28.6 J K^{-1} mol^{-1}). However, at 13.1 mol L^{-1}, the increase is from 23.9 to 47.9 J K^{-1} mol^{-1}. The increase of heat capacity with density is most pronounced in the neighborhood of the critical point (see Sec. 2.6 below).

Since the heat capacity of the bomb is several times that of the gas itself, Michels and Strijland (1952) adopted a differential method. Two identical calorimeters, one evacuated and one filled with gas, are heated simultaneously by an electric current. The temperature of the surroundings is made to follow that of the calorimeters automatically. The difference in temperature rise of the two calorimeters is accurately measured and the heat capacity determined from this and from the total temperature rise, the energy input, and the measured mass of gas.

2.2 Measurements at Constant Pressure

The heat capacity of a gas can be measured by adiabatic constant-flow calorimetry. The apparatus used by Wacker *et al.* (1947) is illustrated in Fig. 2. The gas enters at I and passes through the helical tube T_2, emerging into the evacuated chamber at the bottom at the bath temperature. It then passes through a series of baffles and is heated by heater H_C. The bath temperature is adjusted to follow that of the gas. After passing through more baffles to equilibrate the temperature, the gas leaves the calorimeter through throttle valve V. The rate of flow r of gas is determined by weighing the gas that passes through in a given time. The pressure of the gas before and after heating is measured by means of manometer M. The heat capacity at the mean pressure is determined from the power W applied to the heater and the measured temperature rise ΔT, according to

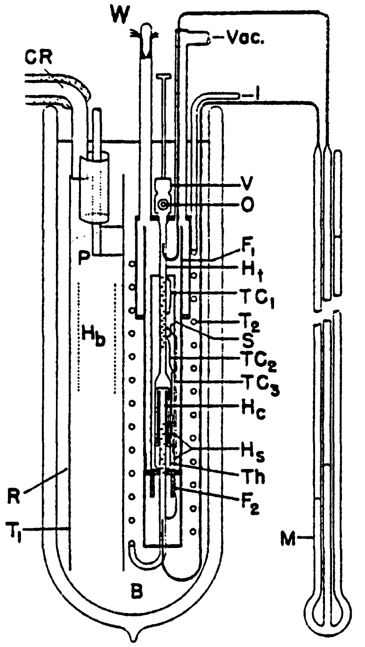

FIG. 2. Adiabatic flow calorimeter, from Scott and Mellors (1945), courtesy of National Bureau of Standards, Washington, D.C. CR, copper rod; P, propeller; H_b, H_t, H_c, H_s, electric heaters; R, copper resistance thermometer; T_1, stirrer tube; W, wax seal; Vac, vacuum line; O, outlet; F_1, F_2, flanges; TC_1, TC_2, TC_3, thermocouples; S, radiation shield; Th, metal thimble; B, constant temperature bath; M, manometer.

$$C_P = W/r(\Delta T + \delta t), \tag{13}$$

where δt is the decrease in temperature that arises as a result of Joule–Thomson cooling (see Sec. 3). It is measured by passing the gas through at the same inlet and outlet pressures without heating.

Values at different pressures can be extrapolated to zero pressure to give ideal-gas heat capacities.

2.3 Ideal-Gas Values

2.3.1 Speed-of-Sound Measurements In the absence of dispersion, the speed of sound u in a gas of molar mass M can be expressed in terms of the temperature T and molar density ρ as

$$u^2 = (RT\gamma/M)\{1 + \beta\rho + \ldots\}, \tag{14}$$

where γ is related to the ideal-gas heat capacity at constant pressure by $\gamma = C_P/(C_P - R)$, and β is termed the second acoustic virial coefficient.

The speed of sound is measured from low-order radially symmetric modes in a spherical resonator. Such modes are nondegenerate and have resonant frequencies that are insensitive to any geometric imperfections of the cavity. An apparatus used for the measurement of speed of sound in several simple gases over wide temperature ranges has been described by Ewing and Trusler (1989). Basically, the resonator is a sealed aluminum-alloy sphere of internal radius 40 mm and wall thickness 10 mm, constructed by machining and polishing two halves separately and then electron-beam welding them together. Two transducer ports, at 90° to each other, were machined in the upper half, one to take an electroacoustic source transducer and the other the detector transducer.

The speed of sound is calculated from the measured frequencies $f_{0,n}$ of the radial modes,

$$u = 2\pi a(f_{0,n} - \Sigma\Delta f)/\nu_{0,n}, \tag{15}$$

where a is the cavity radius, $\nu_{0,n}$ is an accurately known eigenvalue, and $\Sigma\Delta f$ gives the sum of correction terms to the simplified theory used. Ideal-gas heat capacities of high accuracy are derived from measurements at different pressures.

2.3.2 Calculation When simple gases are heated at constant volume, energy is taken up in molecular translation, rotation, and vibration. According to the kinetic theory of gases, the energy of 1 mole of a gas whose particles possess only translational energy is equal to $3RT/2$, and so the molar isochoric heat capacity C_V equals $3R/2$. This is in accord with the classical law of equipartition of energy, which gives the molar energy as $RT/2$ and the isochoric heat capacity as $R/2$, per squared term where, for translation, there are three squared terms since the velocity has components in three dimensions. For a diatomic molecule, in addition to translational motion, there is the possibility of rotation about two axes at right angles, and here there are two more squared terms, and the contribution to the total C_V will equal R. A diatomic molecule will also possess one mode of vibrational motion, but this is associated with kinetic and potential energy, and the contribution here is also equal to R. A linear polyatomic molecule containing N nuclei will also have a contribution R from rotational motion, and $(3N - 5)R$ from vibration. If the molecule is nonlinear, the contributions from rotation and vibration become $3RT/2$ and $(3N - 6)R$ respectively. These results are summarized in Table 1 and compared with experimental data corrected to zero pressure for some simple gases. For an ideal gas,

$$C_P/R = C_V/R + 1. \tag{16}$$

For monatomics such as argon there is excellent agreement between the classical value and experiment, but for polyatomic gases, the classical values are all higher than the experimental results at 298 K. The experimental values increase as the temperature is raised, but only in the case of bromine of the gases listed at 1000 K does the experimental value closely approximate the classical value. However, it is found that the experimental results at 298 K for nitrogen (and other diatomics such as oxygen, carbon monoxide, and nitric oxide) agree closely with the expected value without the vibrational contribution, namely, $3.5R$. Nonlinear polyatomics such as water similarly agree with the classical sum of the translational and rotational contributions plus R, namely, $4R$.

In order to understand why the heat ca-

Table 1. Comparison of molar heat capacities of some simple gases with classical theory.

Gas	Translational	Rotational	Vibrational	C_V/R	C_P/R	C_P/R (expt.) 298 K	C_P/R (expt.) 1000 K
Ar	1.5	0	0	1.5	2.5	2.500	2.500
N_2	1.5	1	1	3.5	4.5	3.518	3.933
Br_2	1.5	1	1	3.5	4.5	4.336	4.534
CO_2	1.5	1	4	6.5	7.5	4.465	6.532
H_2O	1.5	1.5	3	6	7	4.039	4.958
CH_4	1.5	1.5	9	12	13	4.286	8.635
C_2H_6	1.5	1.5	18	21	22	6.330	14.76
C_3H_8	1.5	1.5	27	30	31	8.841	21.05

pacities for molecules differ from these classical values, it is necessary to consider the results of quantum mechanics. This tells us that molecules exist in well-defined energy states, and it is the difference in energy between adjacent energy states that determines the amount of heat required to raise the temperature of the gas by one degree, and hence the heat capacity. Where the energy levels are closely spaced with respect to $k_B T$, where k_B is the Boltzmann constant, as in the case of translational energy, the input of heat will result in the excitation of a very large number of molecules and the heat capacity will be identical to the classical value. However, when the energy gap between adjacent levels is large with respect to $k_B T$, as is generally true for vibrational energy at 298 K, then few molecules get promoted to a higher level, and there is little increase in vibrational energy and hence only a small contribution to the heat capacity from vibrational motion.

The temperature dependence of the perfect-gas molar heat capacities for polyatomics can be understood by application of the methods of statistical thermodynamics, which are described in detail in texts such as Fowler and Guggenheim (1939) and Hill (1956). Just the main points are presented here. Molecules exist in different quantum states and, although it is not possible to state with certainty that a particular molecule will be in a specific state at a given moment, the probability of this occurrence can be given. The reason is that we are dealing with such a large number of molecules. According to the Boltzmann distribution, the probability $p(\epsilon_i)$ of finding a molecule in energy state ϵ_i at temperature T is given by

$$p(\epsilon_i) = \exp[-\epsilon_i/k_B T]/q, \tag{17}$$

where q, known as the molecular partition function, is obtained by summation over all energy states, $\Sigma \exp[-\epsilon_i/k_B T]$. The internal energy per mole of gas, relative to the gas where all molecules are in their ground state, is given by $N_A \Sigma p(\epsilon_i)\epsilon_i$, from which the molar heat capacity at constant volume is given by

$$C_V = 2RT\left(\frac{\partial \ln q}{\partial T}\right)_V + RT^2\left(\frac{\partial^2 \ln q}{\partial T^2}\right)_V. \tag{18}$$

To an excellent approximation, the total energy of a molecule can be considered as the sum of the translational, rotational, vibrational and electronic energies, and the molecular partition function is then given by

$$q = q_{\text{trans}} \times q_{\text{rot}} \times q_{\text{vib}} \times q_{\text{elect}}, \tag{19}$$

where q_{trans} depends on molecular mass m and volume V; h is the Planck constant:

$$q_{\text{trans}} = (2\pi m k_B T/h^2)^{3/2} V; \tag{20}$$

q_{rot} depends on the moment of inertia I and, for a diatomic molecule, is generally given by

$$q_{\text{rot}} = 8\pi^2 I k_B T/\sigma h^2, \tag{21}$$

where σ, the symmetry number, equals 1 for heteronuclear and 2 for homonuclear diatomics. The exception is where the moment of inertia is low, as for hydrogen, or the temperature is low so that the gap between adjacent energy levels is large with respect to $k_B T$. The rotational part of the molecular partition function then has to be evaluated by summation over the allowed energy states. Restrictions on the wave function for the nuclei of a diatomic molecule account for the existence of two distinct forms, the ortho and para forms, which possess different properties. Only for hydrogen and deuterium is this important in the calculation of heat capacities. See, for example, McLelland (1973).

For polyatomic molecules where rotation can occur about three mutually perpendicular axes, the expression for the rotational contribution to the partition function is

$$q_{\mathrm{rot}} = 8\pi^2(8\pi^3 I_A I_B I_C)^{1/2}(k_{\mathrm{B}}T)^{3/2}/\sigma h^3, \tag{22}$$

where the symmetry number has a value equal to the number of equivalent positions into which the molecule can be turned by rotation. For water, $\sigma = 2$, and for methane, $\sigma = 12$.

The contributions to the isochoric heat capacity from translation and rotation are $3R/2$ and R (linear polyatomic) or $3R/2$ (nonlinear polyatomic molecule), as given classically above.

The temperature dependence of the heat capacity arises from the vibrational contribution to the molecular partition function, which for a polyatomic molecule in the rigid-rotor harmonic oscillator model is given by

$$q_{\mathrm{vib}} = \prod_{i=1}^{n}\left[1 - \exp\left(-\frac{h\nu_i}{k_{\mathrm{B}}T}\right)\right]^{-1}, \tag{23}$$

where $n = 3N - 5$ for an N-atomic linear molecule and $n = 3N - 6$ for a nonlinear molecule. This has been evaluated relative to the ground energy state.

The corresponding contribution to the heat capacity is

$$C_{V,\mathrm{vib}} = \sum_{i=1}^{n} \frac{R(h\nu_i/k_{\mathrm{B}}T)^2 \exp(h\nu_i/k_{\mathrm{B}}T)}{[\exp(h\nu_i/k_{\mathrm{B}}T) - 1]^2}. \tag{24}$$

At high temperatures, when $T \gg h\nu_i/k_B$, $C_{V,\mathrm{vib}}$ has the classical value of R per vibrational mode.

The temperature dependence of $C_{V,\mathrm{vib}}$ is illustrated in Fig. 3 for a diatomic molecule ($n = 1$). The contribution at 298 K will depend upon the value of $h\nu_i/k_{\mathrm{B}}$, termed the characteristic vibrational temperature, denoted by Θ_{vib}. Values of Θ_{vib} in kelvins for some simple gases are oxygen 2239, nitrogen 3352, and carbon monoxide 3084. Thus, the contribution to the total heat capacity from vibrational motion for these gases at 298 K is very small.

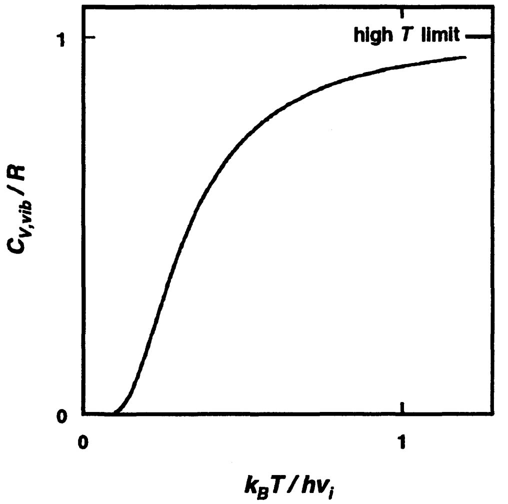

FIG. 3. Reduced temperature dependence of $C_{V,\mathrm{vib}}$ for a diatomic molecule.

When a molecule contains groups such as methyl, hydroxyl, carbonyl, or carboxy connected by a single chemical bond, then the molecule possesses a mode of torsional oscillation, or internal rotation. The evaluation of the partition function is then more complex. This contribution may be factored out of the vibrational contributions to the molecular partition function. It is then necessary to know the internal rotation energy levels, which are determined by solution of the Schrödinger equation with an appropriate expression for the potential energy. For a methyl group, a symmetric top with threefold symmetry, the potential energy function is usually given as a function of θ, the angle of internal rotation, by $\frac{1}{2}V_3(1 - \cos 3\theta)$, where V_3 is the internal rotation barrier height. Alternatively, the gas may be assumed to consist of a mixture of conformers, each corresponding to a minimum in the potential energy function for an internal rotation.

The electronic contribution to the partition function is normally equal to the degeneracy of the ground electronic energy level, since the energy gap between the ground and first excited energy level is higher than that for vibrational energies. For most molecules the ground electronic level is nondegenerate so that $q_{\text{elect}} = 1$. For some molecules, such as nitrogen (II) oxide, there is a low-lying electronically excited level, and this must be included in the summation for q_{elect}. The contribution of higher levels must also be included in calculations of heat capacities at very high temperatures.

Examples of details of these calculations, and a comparison of the calculated ideal-gas heat capacities with experimental data, are given by Chao *et al.* (1973) for ethane and propane, Chao *et al.* (1974) for chloroethanes, Chao and Zwolinski (1975) for ethylene and propylene, Chao *et al.* (1986) for C_1 to C_4 organic oxygen compounds, Chao and Hall (1984) for *o*-, *m*-, and *p*-xylenes, and Chao *et al.* (1990) for organic nitrogen compounds.

2.3.3 Representation Ideal-gas heat capacities are given in a number of compilations, including those of Gurvich *et al.* (1978), the tables produced by the Subcommittee on Technical Data, American Petroleum Institute (API) Research Project 44 (1974), the JANAF Thermochemical Tables (Stull and Prophet, 1971), and the Landolt-Börnstein tables (Schäfer and Lax, 1961). The API tables give, in addition, values for the difference between the real-gas heat capacity and the gas in its standard state for low–molecular-weight hydrocarbons up to 1500 K and a pressure of 500 atm (ethane) or 700 atm (higher alkanes).

For inorganic compounds, the compilation of Wagman *et al.* (1982) gives values of ideal-gas heat capacities at 298.15 K and 10^5 Pa. Barin and Knacke (1973) represent the temperature dependence of the heat capacity of inorganic compounds in their compilation by the equation

$$C_P = A + B \times 10^{-3}T + C \times 10^{-3}T^{-2} + D \times 10^{-6}T^2 \tag{25}$$

and give values for the parameters A, B, C, and D for a large number of substances. Mills (1974) uses this equation, with $D = 0$, to represent the heat capacities of sulfides, selenides, and tellurides.

Ideal-gas heat capacities for a number of technologically important substances have been tabulated in a series of monographs produced by the International Union of Pure and Applied Chemistry (see de Reuck and Craven, 1993, and references therein). These values are thermodynamically consistent with other equilibrium thermodynamic properties, all of which have been used to construct a reduced Helmholtz energy equation, from which values of these properties may then be calculated by mathematical manipulation.

2.3.4 Estimation Where ideal-gas heat-capacity data are not available, various estimation methods based on group contributions can be used. The method of Benson (1968) can be recommended. Numerical contributions are given for atoms or groups with specific neighboring groups, with adjustments for ring compounds. The ideal-gas heat capacity is then determined by identifying the various groups linked to each key atom and summing the corresponding contributions. Illustrative examples are given by Benson *et al.* (1969). Errors are generally less than 1%.

2.4 Effects of Molecular Interactions

Real gases have interactions between the molecules, and if these are represented by the van der Waals equation, Eq. (3), then, from the relationship between C_P and C_V given in the introduction to Sec. 2, on neglecting the term involving a^2,

$$C_P - C_V = R + 2aP/RT^2. \tag{26}$$

So, for a real gas, the difference in heat capacities is greater than for an ideal gas, and this difference becomes larger at high pressures and at low temperatures. Under given conditions, the difference is greater for substances with a higher value for the parameter a, which is a measure of the effect of attractive forces. When the equation of state of the gas is known, the difference is obtained directly from

$$C_P - C_V = T\left(\frac{\partial P}{\partial T}\right)_V\left(\frac{\partial V}{\partial T}\right)_P. \tag{27}$$

2.5 Critical Region

There is a critical anomaly in C_V, as illustrated in Fig. 4 by the results of Michels and Strijland (1952) for carbon dioxide. At the

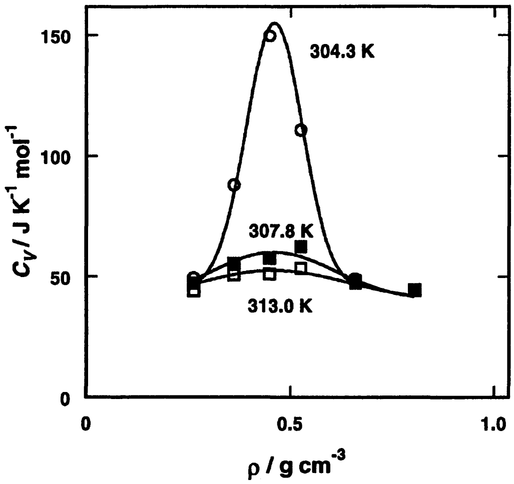

FIG. 4. Heat capacity of carbon dioxide.

critical density, C_V becomes infinite as ln $|T - T_C|$, or slightly greater. The van der Waals equation does not predict this anomaly. This divergence is significantly less than that of C_P, which diverges with the compressibility. The van der Waals equation predicts that the compressibility increases as $(T - T_C)^{-1}$, although in fact it increases more rapidly.

3. JOULE–THOMSON EFFECT

It was shown by Joule and Thomson (later Lord Kelvin) (1853) that when a gas at a constant pressure is passed through a porous plug of cotton wool or silk and emerges at a constant lower pressure, in an apparatus where no heat enters or leaves, then at room temperature the gas becomes cooler. Subsequent experiments with different porous plugs, such as earthenware, confirm this fall in temperature exhibited by most gases around room temperature, though hydrogen and helium show a heating effect under these conditions. This phenomenon is termed the Joule–Thomson effect. Its importance lies in the clear indication of the existence of intermolecular forces, and its application in the liquefaction of gases.

Under these experimental conditions, the process is isenthalpic, and the effect is quantified by introducing the Joule–Thomson coefficient μ_{JT}, defined as $(\partial T/\partial P)_H$. This can be determined from experimental measurements, when there is a cooling effect, of the quantity of electrical energy q passed through a heater immediately after the plug, or throttle, to keep the gas at the same temperature. This heat $q = \Delta H$ and division by the pressure drop ΔP gives $(\Delta H/\Delta P)_T$, which gives $(\partial H/\partial P)_T$ as $\Delta P \rightarrow 0$. μ_{JT} is obtained from the relation

$$\mu_{JT} = -C_P^{-1}\left(\frac{\partial H}{\partial P}\right)_T. \tag{28}$$

The Joule–Thomson coefficient is related to the equation of state for the gas by

$$\mu_{JT} = \left\{T\left(\frac{\partial V}{\partial T}\right)_P - V\right\}\Big/C_P, \tag{29}$$

which, for an ideal gas, gives $\mu_{JT} = 0$.

Real gases behave similarly under identical reduced conditions. At high reduced temperatures (where the temperature is divided by the critical temperature) there is a heating effect, and at low reduced temperatures there is a cooling effect. Thus, there is a temperature, known as the inversion temperature, where a real gas behaves like an ideal gas in that there is no observed temperature change. This is pressure dependent, and part of an inversion curve.

For a van der Waals gas, the Joule–Thomson coefficient is given generally in terms of the molecular parameters a and b by

$$\mu_{JT} = (2a/RT - b) \tag{30}$$

and at the inversion temperature, $T = 2a/Rb$.

When the PVT behavior is represented by the virial equation,

$$PV/RT = 1 + B/V + \ldots, \tag{31}$$

where B is the second virial coefficient, then

$$\mu_{JT} = \left\{T\left(\frac{dB}{dT}\right) - B\right\}C_P^{-1}. \tag{32}$$

The technological importance of the Joule–Thomson effect is in the liquefaction of gases. The first major applications of this technique were made by von Linde (1895) and Hampson (1895). Gas at a pressure of about 200 atm (20 MPa) streamed through a valve, or throttle, and emerged at a pressure of 20 atm (2 MPa) (von Linde) or 1 atm (0.1 MPa) (Hampson). Below its inversion temperature, the gas will cool; it then passes over the outer surface of the tube bringing the incoming gas to the expansion jet, which in turn is further cooled. After further cycles, the gas liquefies. The collection, and storage, of liquefied gas was made possible by the invention of the double-walled vacuum vessel, or Dewar flask, by Dewar (1893).

For hydrogen and helium, whose inversion temperatures are about 190 and 100 K, respectively, it is necessary to cool the gases below this temperature before further cooling and eventual liquefaction will occur by the Joule–Thomson effect.

4. THERMAL CONDUCTIVITY

Fourier's phenomenological law of heat conduction relates the heat (energy) flux Q across an area A to the temperature gradient (dT/dz) by the expression

$$Q = -\lambda A \frac{dT}{dz}, \qquad (33)$$

where λ is the thermal conductivity. Accurate values for the thermal conductivity of gases are of vital importance to chemical and mechanical engineers concerned with heat-transfer processes—for example, in heat exchangers. Dilute-gas data also provide information about the pair-interaction potential energy function.

4.1 Measurements

The methods for measurement of the thermal conductivity of a gas can conveniently be classified into two groups. In the first of these, the steady-state experiments, a constant temperature difference is maintained between two surfaces in the gas and the heat flux is measured. The instruments are usually based on parallel plates or concentric cylinders and have the advantage of simple working equations. The main difficulty in such experiments is the minimization of the effects of other mechanisms of heat transfer, such as convection and radiation. The accuracy of such methods is typically ±(1 to 3)%. The second group, the transient methods, eliminates the convective contribution by imposition of a small temperature gradient to a gas initially at equilibrium. In close approximation to the ideal model, this perturbing heat flux is normally produced by electrical dissipation in a thin, cylindrical wire. The thermal conductivity of gases can be measured with an accuracy of ±0.3% by this method. Examples of each of these types of measurement are described below.

4.1.1 Concentric-Cylinder Method

This method is based on the idealized model of two coaxial cylinders, in which a heat flux generated in the inner cylinder is conducted radially through the gas contained between the cylinders to the outer cylinder in steady-state conditions. If the external radius of the inner cylinder, termed the emitter, is r_1 and the internal radius of the outer cylinder, termed the receiver, is r_2 and the temperatures of these surfaces are respectively T_1 and T_2, then the heat flux per unit time per unit length is given by

$$\Phi = 2\pi\lambda(T_1 - T_2)/\log(r_2/r_1). \qquad (34)$$

The thermal conductivity determined from the heat flux and the temperature difference is the value at the average temperature $(T_1 + T_2)/2$.

In practice, the cylinders are not infinite in length, and heat can be transported across the ends. However, if end pieces of the same diameter as the inner cylinder are placed at either end of this cylinder, separated from it by a gap of similar magnitude to the distance between the cylinders, and maintained at the same temperature as the inner surface of the outer cylinder, the heat flux is again proportional to the temperature difference and the thermal conductivity, with a proportionality constant that depends on the geometry of the instrument. This may be determined directly from capacitance measurements to give absolute values for the thermal conductivity, or determined from measurements using a gas of known thermal conductivity, in which case the measurements are relative.

As an example of the many designs of such instruments, the coaxial cylinder cell of Le Neindre (1969) and Tufeu (1971) is illustrative. The cell was made of silver with its axis vertical. The inner cylinder was 120 mm in length, with an outer diameter of 19.6 mm. The outer cylinder had an inner diameter of 20 mm, giving an annular gap of 2 mm, an outer diameter of 50 mm, and a length of 200 mm. A heating element 112 mm in length was formed from Pt/25% Rh wire of diameter 0.3 mm, helically wound on an alumina tube of 4-mm external diameter, and set up in the central hole of the inner cylinder. The temperature difference was measured by eight Pt/10% Rh thermocouples, four in each cylinder. Temperature differences of about 2 K were measured with a precision of 0.003 K. The heat emitted from the inner cylinder was determined from the measured voltage and current intensity in the heating element. Corrections were applied for heat losses through conduction through the cylinder element, electric leads, and ceramic supports. Measurements made with this instrument were recently reported by Sousa *et al.* (1992) for HCFC-142b.

To reduce the problem of axial heat loss, the two end pieces are maintained at the same temperature as the inner cylinder, and then all the heat passes radially from one cylinder to the other. The instrument of Tanaka *et al.* (1991) is a good illustration of an instrument with two such guard cells.

4.1.2 Parallel-Plate Method In this method, the gas is confined between two horizontal circular plates, usually made of copper, with the upper one heated to a higher temperature. For an infinite gaseous layer, of thickness δ, between two parallel plates with a temperature difference ΔT, application of the Fourier law gives an average thermal conductivity. This equals the thermal conductivity at the mean temperature of the plates, assuming that the thermal conductivity varies linearly with temperature, which is normally a good approximation. In practice, the distance δ is maintained constant by the use of spacers made from a low–thermal-conductivity material, and is in the range 0.2–0.5 mm to reduce possible effects of convection. The plates are highly polished and protected against oxidation to reduce the emissivity and lower the radiation correction. Information on the magnitude of this effect can be determined by repeating the measurements with a different plate separation. Account must be taken of the finite diameter of the plates. Radial and axial heat losses can be minimized by use of guard heaters. Thermal-conductivity measurements by this method can be either relative or absolute. A brief history of early measurements has recently been given by Sirota (1991). Special mention should be made of the work of Michels *et al.* (1962), who confirmed the existence of an anomalous increase in the thermal conductivity of carbon dioxide in the critical region. That this was not due to convection, which increases heat transfer significantly in this region, was demonstrated by repeating the experiments for different plate distances and temperature differences, and showing that the results were independent of the Rayleigh number. The parallel-plate method has recently been used for measurements on polar alternative refrigerant gases by Hammerschmidt (1995). Although this method has lower accuracy than the transient hot-wire method, described in the following section, there are no electrical or chemical interaction problems to consider.

4.1.3 Transient Hot-Wire Method The advantage of transient methods over the methods described above is that convective contributions to the heat transfer can be eliminated. There are exact working equations for this technique, which leads to absolute measurements for all regions of the phase diagram except the critical region very close to the critical point. The ideal model is of an infinitely long, vertical line source of heat, with zero heat capacity and infinite thermal conductivity, immersed in an infinite isotropic gas at the same initial equilibrium temperature. A time-dependent perturbation, in the form of a stepwise heat flux q per unit length, is applied, and the temperature response of the gas is monitored. In practice, instead of a line source, a cylindrical wire of radius r_0 is used. This has a uniform temperature equal to that of the gas at distance r_0 in the ideal model, and for small values of r_0 the temperature rise $\Delta T_{id}(r_0,t)$ is given by

$$\Delta T_{id}(r_0,t) = \frac{q}{4\pi\lambda} \ln\left(\frac{4at}{r_0^2 C}\right), \qquad (35)$$

where $C = \exp(\gamma)$, where γ is Euler's constant (0.5772157 . . .), and a is the thermal diffusivity, given by $a = \lambda/\rho C_p$ and taken as constant.

The thermal conductivity of a gas may be determined from a plot of $\Delta T_{id}(r_0,t)$ vs ln t, and the thermal diffusivity can be obtained by substitution for a measured value of $\Delta T_{id}(r_0,t)$ at any given time.

In any practical instrument, a finite length of metallic wire provides the source of heat and acts as a resistance thermometer. The measured temperature rise ΔT_w equals the sum of the ideal value ΔT_{id} plus corrections δT_i. Most of these corrections were considered by Healy *et al.* (1976), and a more extensive treatment was given by Assael *et al.* (1991). Careful apparatus design will reduce these corrections to less than 0.5% of ΔT_{id}. Use of a supported wire of finite length results in the temperature field decreasing near the wire supports as a result of heat losses. To eliminate these "end effects," two wires are employed instead of one. These wires differ solely in their length, and, if they are sufficiently long, then the end effects will be the same when they are subjected to the same heating current. Measurements of the difference in resistance of the two wires at specific times will be equivalent to the resistance of a finite length of an infinitely long wire. The wires are typically made of platinum, 7 to 10 μm in diameter and with the short wire about 70 mm and the long wire 170 mm in length. The mounting of these wires is the key to accurate thermal conductivity measurement. They must be vertical, subject to constant tension, and unaffected by changes in temperature and pressure. To overcome the lateral movement resulting from temperature changes when the wire is fixed to rigid supports at either end, a beryllium-copper alloy spring has been fastened to the upper end of the wire to keep the wire under tension. Figure 5 shows the mounted wires in the transient hot-wire cell of Haran and Wakeham (1982) for measurements in gases from 77 to 520 K at pressures from 0.1 to 10 MPa.

The two cells are machined from a single stainless steel block. The supports for the platinum wires are formed from brass terminal posts set in a stainless steel cylinder in a ceramic bush with a stainless steel sleeve.

Experimentally, the change in resistance of the wires is followed over a period of about 1 s after the heating cycle, using a bridge circuit. In the method of Roder (1981), two arms are standard resistors of 100 Ω, and the resistance of the other arms, each of which contains one of the wires with the leads and ballast resistors, can be adjusted to 100 Ω. The bridge is balanced prior to the start of an experiment, the power supply is activated, and the voltage developed across the bridge is read and stored at fixed time intervals. The accuracy of the measured thermal conductivity is estimated to be better than 1%. An alternative procedure, employed by Mardolcar *et al.* (1986), uses a bridge with one arm arranged to provide a series of preset balance points as the resistance difference between the two wires increases. The times at which these balances occur provides a set of data of temperature

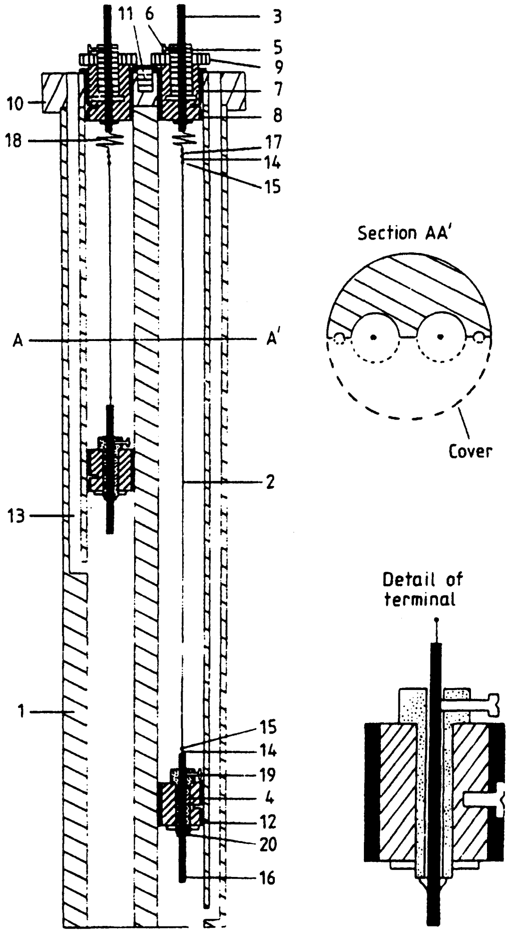

FIG. 5. Transient hot-wire cell, from Haran and Wakeham (1982), courtesy of IOP Publishing Ltd.

rise vs time. Use of a computer-controlled bridge by Charitidou *et al.* (1987) increased the number of $(\Delta T,t)$ points to 1024 for a single run of 1-s duration and improved the precision of the thermal conductivity measurements to about ±0.1% and the accuracy to ±0.3%.

4.1.4 Experimental Results Values for the thermal conductivity of a number of simple gases at 0.1 MPa are given in Table 2.

At 0.1 MPa, the thermal conductivity increases with increase in temperature, as a result of purely kinetic effects. The rate of increase diminishes as the pressure is raised until, for pressures above about 20 MPa, $(\partial\lambda/\partial T)_p$ has the opposite sign as collisional transfer predominates. This mechanism is highly dependent on density. In fact it is density that is the important variable for the theoretical interpretation of thermal-conductivity data, as is illustrated by the regular dependence illustrated in Fig. 6 for argon, with the experimental results of Michels *et al.* (1963) at 273.6 and 348.2 K. Data of Calado *et al.* (1987) for liquid argon at 140.1 K are included for comparison.

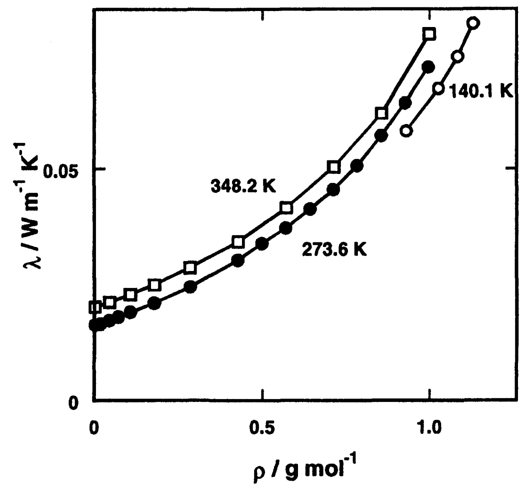

FIG. 6. Density dependence of the thermal conductivity of argon.

However, as the temperature decreases toward the critical temperature, the thermal conductivity vs density plots exhibit a maximum, as shown in Fig. 8 (Sec. 4.5) for carbon dioxide. The critical enhancement of the thermal conductivity is apparent at densities from about 0.1 to 2.1 times the critical density, decreasing in magnitude to about 1% of the total thermal conductivity at these extremes. It is greatest at the critical density and persists at temperatures up to about 1.4 times the critical temperature.

4.2 Theoretical Description for Monatomics

In the simple kinetic theory of gases (see, for example, Jeans, 1962), the thermal conductivity of a gas is related to the temperature and to the mass m and diameter σ of the molecules, considered as hard spheres, by the expression

$$\lambda_{\mathrm{HS}} = \frac{c_V}{3\pi\sigma^2}\left(\frac{8k_BT}{\pi m}\right)^{1/2}, \tag{36}$$

Table 2. Thermal conductivity data (mW m^{-1} K^{-1}) for some representative gases.

		T(K)			
Substance	Reference	200	300	400	600
	Low-pressure limiting values				
Helium	Kestin *et al.* (1984)	119.3	156.7	190.6	252.4
Argon	Kestin *et al.* (1984)	12.4	17.9	22.6	30.6
Xenon	Kestin *et al.* (1984)	3.6	5.5	7.3	10.4
Hydrogen	Assael *et al.* (1986)	131.7	186.9	230.4	
Nitrogen	Stephan *et al.* (1987)	18.7	26.0	32.3	44.0
Carbon monoxide	Millat and Wakeham (1989)		25.0	32.3	45.7
Carbon tetrafluoride	Uribe *et al.* (1990)		16.0	24.1	39.9
	0.1 MPa				
Methane	Younglove and Ely (1987)	22.5	34.3	49.1	84.1
Ethane	Younglove and Ely (1987)	11.0	21.3	35.4	70.5
Propane	Younglove and Ely (1987)		18.0	30.6	61.9
Butane	Younglove and Ely (1987)		16.4	28.4	59.1

where c_v is the molecular heat capacity at constant volume. The rigorous expression differs in the numerical factor; $2\sqrt{2/3\pi}$ is replaced by 25/32.

This theory predicts that thermal conductivity is pressure independent. Experimentally, the effect of an initial pressure increase is small; for example, for argon at 348.2 K and at 0.1 MPa, a 50-fold increase in pressure produces only a 9% increase in thermal conductivity. The effect of a temperature increase on the thermal conductivity of a gas at 0.1 MPa is correctly predicted, but experimentally the magnitude of the effect is greater than the $T^{1/2}$ dependence that is predicted. It is necessary to describe the interactions between molecules by a more realistic pair-interaction potential-energy function, $U(r)$. The thermal conductivity can then be calculated to the required order of approximation from reduced collision integrals (see Maitland *et al.*, 1981). To first-order approximation,

$$\lambda = \lambda_{HS}/\Omega^{(2,2)*}(T^*). \qquad (37)$$

Values for the reduced collision integral $\Omega^{(2,2)*}(T^*)$ have been tabulated by Maitland *et al.* (1981) for different pair-interaction potential-energy functions at different reduced temperatures, $T^* = k_BT/\epsilon$, where ϵ is the depth of the attractive potential well.

The thermal conductivity of noble gases at low pressure can be calculated from measurements of their viscosity η, since it can be shown theoretically that to a good approximation

$$\lambda = 15k_B\eta/4m. \qquad (38)$$

4.3 Theoretical Description for Polyatomics

In principle, a quantum mechanical treatment is required to account for energy exchange on collision between internal states, and between internal states and translational motion. This becomes very complicated, and it is therefore customary to use semiclassical theory, as given by Wang Chang and Uhlenbeck and independently by de Boer (see Wang Chang *et al.*, 1964). Here, each internal energy state is considered essentially as an independent chemical species, and the translational motion is treated classically. Evaluation of the formal expression for the thermal conductivity is possible by substitution of various contributions to the integrals by other experimental quantities, as shown by Monchick *et al.* (1965). In this way, the thermal conductivity is related to the viscosity to the equivalent of first-order approximation (see Maitland *et al.*, 1981).

4.4 Correlations for Dense Gases

Thermal-conductivity data for gases at temperatures and pressures where there is no significant critical enhancement fall on parallel curves when plotted as isotherms vs density, as shown in Fig. 6. The data are very satisfactorily correlated by a plot of the excess thermal conductivity $\Delta\lambda(\rho,T)$ against density, where $\Delta\lambda(\rho,T)$ is the difference between the value at density ρ and the dilute-gas value. This is illustrated by experimental results for isobutane, shown in Fig. 7. The points fall on a single curve and do not exhibit any temperature dependence.

4.5 Critical Region

The thermal conductivity of a gas at temperature T and density ρ, $\lambda(\rho,T)$, is conveniently represented as the sum of three contributions:

$$\lambda(\rho,T) = \lambda^{(0)}(T) + \Delta\lambda(\rho,T) + \Delta\lambda_c(\rho,T), \qquad (39)$$

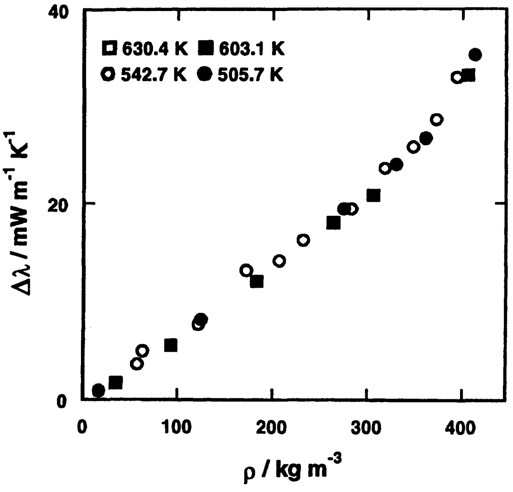

FIG. 7. Excess thermal conductivity for isobutane.

where $\lambda^{(0)}(T)$ is the dilute-gas value, obtained by extrapolation of results at different pressures to zero pressure; $\Delta\lambda(\rho,T)$ is the excess thermal conductivity; and $\Delta\lambda_c(\rho,T)$ is the critical enhancement. Large density fluctuations in the gas in the neighborhood of the critical point lead to a significant contribution from $\Delta\lambda_c(\rho,T)$. Even at temperatures up to 50% above the critical temperature, and densities up to twice the critical density, the contribution is more than 1% of the total thermal conductivity. Asymptotically close to the critical point, the thermal conductivity has a singular behavior, which is given in terms of a correlation length ξ by the equation

$$\Delta\lambda_C = \frac{R_D k_B T \rho c_P}{6\pi\eta\xi}, \tag{40}$$

where c_P is the isobaric specific heat capacity, and R_D is a universal amplitude, with a value of about 1.02. To cover the range from this limit to conditions where the critical enhancement is negligible, a crossover function is used (Sengers and Luettmer-Strathmann, 1996). This was developed from mode-coupling theory of the critical fluctuations and contains only one adjustable parameter, a microscopic cutoff distance, which has been found to have a value of about 0.2 nm for argon and ethane. The correlation length ξ lies also about 0.15–0.2 nm.

The density dependence of the thermal conductivity is illustrated in Fig. 8 for ethane along different isotherms. Agreement between the values calculated as above and the experimental measurements of Mostert *et al.* (1990), obtained with a parallel-plate apparatus, is very good, even close to the critical region where the discrepancies are greater, for here there are larger uncertainties in the density.

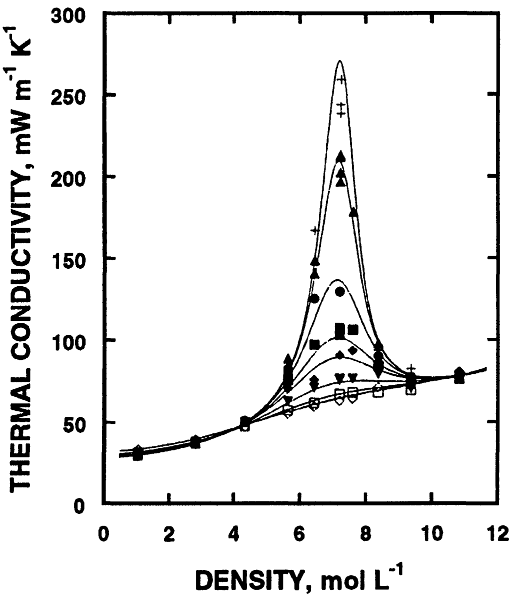

FIG. 8. The thermal conductivity of ethane. Solid line is calculated from Eq. (39). Experimental data, Mostert *et al.* (1990): cross, 305.46 K; solid triangle, 305.56 K; solid circle, 306 K; solid square, 307 K; solid diamond, 308 K; inverted solid triangle, 312 K; open square, 320 K; open diamond, 333 K. Figure from Vesovic *et al.* (1994), courtesy of Plenum Publishing Corp.

5. THERMAL DIFFUSION

Thermal diffusion is the process whereby an initially uniform mixture of gases will partially separate when a temperature gradient is applied. The extent of the separation is denoted by the thermal-diffusion factor, which can be expressed by kinetic theory in terms of quantities related to the molecular masses, and to the potential energy functions for the various binary interactions, as described by Maitland *et al.* (1981). This property is very sensitive to the form of the potential function, but accurate measurement is difficult, although in principle the method is straightforward. In its simplest form, the apparatus consists of two bulbs connected by a tube and maintained at different temperatures, T_1 and T_2. The mole fractions $x_{a,1}$ and $x_{a,2}$ of the heavier gas in each of the bulbs are determined under steady-state conditions. The thermal-diffusion factor, α_T, is obtained from the relation

$$\log\left(\frac{x_{a,1}/(1 - x_{a,1})}{x_{a,2}/(1 - x_{a,2})}\right) = \int_{\log T_2}^{\log T_1} \alpha_T d(\log T). \tag{41}$$

In practice, $x_{a,1} - x_{a,2}$ is typically about 0.05 for a temperature difference of 100 K. The first experimental confirmation of thermal diffusion in gases was obtained by Chapman and Dootson (1917), who filled their two-bulb apparatus with $H_2 + CO_2$ (and $H_2 + SO_2$) and heated one bulb, which became enriched with hydrogen by (2–3)%. This method has been widely used, with modifications for accurate analysis, since the separation is small, and a detailed description of various two-bulb devices has been given by Taylor (1991). This reference also describes the Trennschaukel and thermal-diffusion column apparatus, which are usually employed to magnify the effects.

6. THERMAL DIFFUSIVITY

The thermal diffusivity a is defined as $\lambda/\rho C_P$, that is, the ratio of the thermal conductivity λ to the product of density ρ and isobaric molar heat capacity C_P. Values for thermal diffusivity have been determined from measurements of thermal conductivity combined with calculations of density and heat capacity from accurate equations of state. They can also be determined by the technique of dynamic light scattering, which gives thermal diffusivity by analysis of the frequency spectrum of the light scattered from a gas by local statistical temperature fluctuations in the absence of any applied gradient. The signal is greater in a gas close to the critical temperature, but recently the method has been applied at temperatures further away from the critical temperature by Liepertz (1988) and by Kruppa *et al.* (1988). The principle of the method is as follows (Kraft and Leipertz, 1994). When laser light is incident on a pure fluid at equilibrium, it is scattered by hydrodynamic modes in all directions. The spectrum of the scattered light consists of three lines. There is an unshifted Rayleigh line resulting from entropy (temperature) fluctuations, and two symmetrically shifted Brillouin lines caused by pressure fluctuations. The spectral linewidths depend on the scattering vector q_{sv}, and on the thermal diffusivity, in the case of the Rayleigh line. For small observation angles Θ, q_{sv} is given in terms of the laser wavelength λ_1 by the relation

$$q_{sv} = 2\pi \sin\Theta/\lambda_1. \tag{42}$$

The Rayleigh linewidth is given by aq^2, from which the thermal diffusivity is calculated. In view of the narrowness of the lines, dynamic light scattering or photocorrelation spectroscopy is used to determine the linewidth. Full experimental details are given in references in the article by Kraft and Leipertz (1994). As an indication of the values for the property, their results for refrigerant R152 vapor along the saturation line decrease from 10.2×10^{-8} m^2 s^{-1} at 86.9 K to 0.49×10^{-8} m^2 s^{-1} at 112.7 K.

7. THERMAL TRANSPIRATION

Under the conditions for thermal-diffusion measurement (see above), where two bulbs containing a gas mixture at different temperatures are joined by a tube, there is unlikely to be any pressure difference set up. If, however, the tube is replaced by one, or more, capillary tubes or a porous membrane or plate, then a pressure difference results whether the bulbs are filled with a gas mixture or a pure gas. This phenomenon is termed thermal transpiration. It is important in the investigation of gas–surface interactions, in the determination of surface diffusion of adsorbed molecules, especially for flow through porous media, and for the evaluation of the translational part of the thermal conductivity, and other quantities required for a complete description of the thermal conductivity of polyatomic molecular gases.

A summary of the experimental methods employed to measure the pressure difference ΔP (the "thermomolecular pressure difference") and the theoretical evaluation of parameters in a kinetic theory description have been given by Millat (1991).

GLOSSARY

Calorimeter: An instrument for measuring heat changes.

Critical Temperature: The highest temperature at which a substance can exist in the liquid state.

Intermolecular Potential Energy Function: A representation of the dependence of the potential energy of a pair of molecules on their separation (and relative orientation).

Inversion Temperature: The temperature at which Joule–Thomson expansion produces no temperature change in a real gas.

Isobaric Heat Capacity: The amount of heat to raise the temperature of a substance by one degree at constant pressure.

Isochoric Heat Capacity: The amount of heat to raise the temperature of a substance by one degree at constant volume.

Joule–Thomson Effect: The temperature change that is observed when a real gas passes from a constant-pressure region through a throttle to a region of lower constant pressure.

Thermal Conductivity: The heat flux through unit area per unit temperature gradient.

Thermal Diffusion: The partial separation of an initially uniform gas mixture when it is subjected to a temperature gradient.

Thermal Diffusivity: The ratio of thermal conductivity to the product of density and molar heat capacity at constant pressure.

Thermal Transpiration: The process whereby a temperature gradient between two bulbs connected by capillaries and containing a gas, or gas mixture, results in a pressure developing.

van der Waals Equation: An equation that relates gas pressure, molar volume, and temperature and takes into account the finite size of molecules and their attraction.

Works Cited

Assael, M. J., Mixafendi, S., Wakeham, W. A. (1986), *J. Phys. Chem. Ref. Data* **15,** 1315–1322.

Assael, M. J., Nieto de Castro, C. A. N., Roder, H. M., Wakeham, W. A. (1991), in: W. A. Wakeham, A. Nagashima, J. V. Sengers (Eds.), *Measurement of the Transport Properties of Fluids,* Oxford, U.K.: Blackwell Scientific Publications, pp. 161–194.

Barin, I., Knacke, O. (1973), *Thermochemical Properties of Inorganic Substances,* Berlin: Springer Verlag.

Benson, S. W. (1968), *Thermochemical Kinetics,* New York: Wiley, Chap. 2.

Benson, S. W., Cruikshank, F. R., Golden, D. M., Haugen, G. R., O'Neal, H. E., Rodgers, A. S., Shaw, R., Walsh, R. (1969), *Chem. Rev.* **69,** 279–324.

Calado, J. C. G., Mardolcar, U. V., Nieto de Castro, C. A., Roder, H. M., Wakeham, W. A. (1987), *Physica A* **143,** 314–325.

Chao, J., Hall, K. R. (1984), *Thermochim. Acta* **72,** 323–334.

Chao, J., Zwolinski, B. J. (1975), *J. Phys. Chem. Ref. Data* **4,** 251–261.

Chao, J., Wilhoit, R. C., Zwolinski, B. J. (1973), *J. Phys. Chem. Ref. Data* **2,** 427–437.

Chao, J., Rodgers, A. S., Wilhoit, R. C., Zwolinski, B. J. (1974), *J. Phys. Chem. Ref. Data* **3,** 141–162.

Chao, J., Hall, K. R., Marsh, K. N., Wilhoit, R. C. (1986), *J. Phys. Chem. Ref. Data* **15,** 1369–1436.

Chao, J., Gadalla, N. A. M., Gammon, B. E., Marsh, K. N., Rodgers, A. S., Somayajulu, G. R., Wilhoit, R. C. (1990), *J. Phys. Chem. Ref. Data* **19,** 1547–1615.

Chapman, S., Dootson, F. W. (1917), *Philos. Mag.* **33,** 248–253.

Charitidou, E., Dix, M., Assael, M. J., Nieto de Castro, C. A., Wakeham, W. A. (1987), *Int. J. Thermophys.* **8,** 511–519.

de Reuck, K. M., Craven, R. J. B. (1993), *Methanol. International Tables of the Fluid State–12,* Oxford, U.K.: Blackwell Scientific Publications.

Dewar, J. (1893), *Proc. R. Inst.* **14,** 1–12.

Ewing, M. B., Trusler, J. P. M. (1989), *J. Chem. Phys.* **90,** 1106–1115.

Fowler, R. H., Guggenheim, E. A. (1939), *Statistical Thermodynamics,* London: Cambridge Univ. Press.

Goodwin, R. D. (1961), *J. Phys. Chem. Ref. Data* **65C,** 231–243.

Goodwin, R. D., Weber, L. A. (1969a), *J. Phys. Chem. Ref. Data* **73A,** 1–13.

Goodwin, R. D., Weber, L. A. (1969b), *J. Phys. Chem. Ref. Data* **73A,** 15–24.

Gurvich, L. V., Veyts, I. V., Medvedev, V. A. (1978), *Thermodynamic Properties of Individual Substances,* Vol. 1, Pt. 2, 3rd ed., Moscow: Nauka.

Hammerschmidt, U. (1995), *Int. J. Thermophys.* **16,** 1203–1211.

Hampson, W. (1895), British Patent No. 10165; (1896), *Nature* **55,** 485.

Haran, E. N., Wakeham, W. A. (1982), *J. Phys. E* **15,** 839–842.

Healey, J., de Groot, J. J., Kestin, J. (1976), *Physica C* **82,** 392–408.

Hill, T. L. (1956), *Statistical Mechanics,* New York: McGraw-Hill.

Jeans, J. (1962), *An Introduction to the Kinetic Theory of Gases,* Cambridge, U.K.: Cambridge Univ. Press.

Joule, J. P., Thomson, W. (1853), *Philos. Trans. R. Soc.* **143,** 357.

Kestin, J., Knierim, K. D., Mason, E. A., Najafi, B., Ro, S. T., Waldman, M. (1984), *J. Phys. Chem. Ref. Data* **13,** 229–303.

Kraft, K., Leipertz, A. (1994), *Int. J. Thermophys.* **15,** 791–802.

Kruppa, B., Jany, P., Straub, J. (1988), *Int. J. Thermophys.* **9,** 911–921.

Leipertz, A. (1988), *Int. J. Thermophys.* **9,** 897–909.

Le Neindre, B. (1969), Ph.D. Thesis, Paris University.

Magee, J. W. (1994), *Int. J. Thermophys.* **15,** 849–861.

Maitland, G. C., Rigby, M., Smith, E. B., Wakeham, W. A. (1981), *Intermolecular Forces. Their Origin and Determination,* Oxford, U.K.: Clarendon Press, Chap. 6.

Mardolcar, U. V., Nieto de Castro, C. A., Wakeham, W. A. (1986), *Int. J. Thermophys.* **7,** 259–272.

McLelland, B. J. (1973), *Statistical Thermodynamics,* New York: Wiley.

Michels, A., Sengers, J. V., van der Gulik, P. S. (1962), *Physica* **28,** 1201–1215.

Michels, A., Sengers, J. V., van der Klundert, L. J. M. (1963), *Physica* **29,** 149–160.

Michels, A., Strijland, J. (1952), *Physica* **8–9,** 613–628.

Millat, J. (1991), in: W. A. Wakeham, A. Nagashima, J. V. Sengers (Eds.), *Measurement of the Transport Properties of Fluids,* Oxford, U.K.: Blackwell Scientific Publications, pp. 375–383.

Millat, J., Wakeham, W. A. (1989), *J. Phys. Chem. Ref. Data* **18,** 565–581.

Mills, K. C. (1974), *Thermodynamic Data for Inorganic Sulphides, Selenides and Tellurides,* London: Butterworths.

Monchick, L., Pereira, A. N. G., Mason, E. A. (1965), *J. Chem. Phys.* **42,** 3241–3256.

Mostert, R., van den Berg, H. R., van der Gulik, P. S., Sengers, J. V. (1990), *J. Chem. Phys.* **92,** 5454–5462.

Roder, H. M. (1981), *J. Res. Natl. Bur. Std.* **86,** 457–493.

Schäfer, K., Lax, E. (Eds.) (1961), *Landolt-Börnstein Zahlenwerte und Funktionen aus Physik, Chemie, Astronomie, Geophysik, Technik,* 2. Band, *Eigenschaften der Materie in ihren Aggregatzuständen,* Teil 4, *Kalorische Zustandgrössen,* Berlin: Springer-Verlag.

Scott, R. B., Mellors, J. W. (1945), *J. Res. Natl. Bur. Std.* **34,** 243–254.

Sengers, J. V., Luettmer-Strathmann, J. (1996), in: J. Millat, J. H. Dymond, C. A. N. Nieto de Castro (Eds.), *Transport Properties of Fluids: Their Correlation, Prediction and Estimation,* New York: Cambridge Univ. Press, Chap. 6.

Sirota, A. M. (1991), in: W. A. Wakeham, A. Nagashima, J. V. Sengers (Eds.), *Measurement of the Transport Properties of Fluids,* Oxford, U.K.: Blackwell Scientific Publications, pp. 142–160.

Sousa, A. T., Fialho, P. S., Nieto de Castro, C. A., Tufeu, R., Le Neindre, B. (1992), *Int. J. Thermophys.* **13,** 383–389.

Stephan, K., Krauss, R., Laesecke, A. (1987), *J. Phys. Chem. Ref. Data* **16,** 993–1023.

Stull, D. R., Prophet, H. (Eds.) (1971), *JANAF Thermochemical Tables,* 2nd ed., National Standard Reference Data System—National Bureau of Standards 37, Washington, D.C.: Government Printing Office.

Subcommittee on Technical Data, American Petroleum Institute Research Project 44 (C. C. Williams, III, Chairman) (1974), *Selected Values of Properties of Hydrocarbons and Related Compounds,* Texas A&M University, College Station, TX.

Tanaka, Y., Nagata, M., Makita, T. (1991), *Int. J. Thermophys.* **12,** 949–963.

Taylor, W. L. (1991), in: W. A. Wakeham, A. Nagashima, J. V. Sengers (Eds.), *Measurement of the Transport Properties of Fluids,* Oxford, U.K.: Blackwell Scientific Publications, Chap. 10.3.

Tufeu, R. (1971), Ph.D. Thesis, Paris University.

Uribe, F. J., Mason, E. A., Kestin, J. (1990), *J. Phys. Chem. Ref. Data* **19,** 1123–1136.

van der Waals, J. D. (1873), Dissertation, Universiteit Leiden.

Vesovic, V., Wakeham, W. A., Luettmer-Strathmann, J., Sengers, J. V., Millat, J., Vogel, E., Assael, M. J. (1994), *Int. J. Thermophys.* **15,** 33–66.

von Linde, C. (1895), German Patent No. 88824; (1896), *Ann. Phys.* **57,** 328–332.

Wacker, P. F., Cheney, R. K., Scott, R. B. (1947), *J. Res. Natl. Bur. Std.* **38,** 651–659.

Wagman, D. D., Evans, W. H., Parker, V. B., Schumm, R. H., Harlow, I., Bailey, S. M., Churney, K. L., Nuttall, R. L. (1982), *The NBS Tables for Chemical Thermodynamic Properties,* J. Phys. Chem. Ref. Data, Suppl. 2.

Wang Chang, C. S., Uhlenbeck, G. E., de Boer, J. (1964), *Studies in Statistical Mechanics,* Vol. 2, Pt. C, Amsterdam: North-Holland.

Younglove, B. A., Ely, J. F. (1987), *J. Phys. Chem. Ref. Data* **16,** 577–798.

Further Reading

Dean, J. A. (1992), *Lange's Handbook of Chemistry,* 14th ed., New York: McGraw-Hill. Tables of low-pressure thermal-conductivity coefficients.

de Reuck, K. M. (1990), *Fluorine, International Tables of the Fluid State—11,* Oxford, U.K.: Blackwell Scientific Publications. Extensive tabulations of the ideal-gas and real-gas isochoric and isobaric heat capacities.

de Reuck, K. M., Angus, S., Cole, W. A., Craven, R. J. B., Wakeham, W. A. (Eds.) (1988), *Ethylene (Ethene). International Tables of the Fluid State—10,* Oxford, U.K.: Blackwell Scientific Publications. Extensive tabulations of the ideal-gas and real-gas isochoric and isobaric heat capacities.

Frankiss, S. G., Green, J. H. S. (1972), *Statistical Methods for Calculating Thermodynamic Functions,* Chemical Society Specialist reports, *Chemical Thermodynamics,* Vol. 1, London: The Chemical Society. Discussion of heat-capacity calculation from spectroscopic information, including molecules with internal rotation, molecules exhibiting inversion, and molecules with ring motion.

Lide, D. R. (Ed.) (1995), *Handbook of Physics and Chemistry,* 75th ed., Boca Raton, FL: CRC Press. Tables of low-pressure thermal-conductivity coefficients.

Mason, E. A., Malinauskas, A. P. (1983), *Gas*

Transport in Porous Media—The Dusty Gas Model, New York: Elsevier. A discussion of the "dusty gas" model used as a basis for the interpretation of thermal transpiration measurements.

Millat, J., Dymond, J. H., Nieto de Castro, C. A. N. (Eds.) (1996), *Transport Properties of Fluids: Their Correlation, Prediction and Estimation,* New York: Cambridge Univ. Press, The results of recent theoretical advances and empirical correlation and estimation methods are brought together in this work.

Partington, J. R. (1960), *An Advanced Treatise in Physical Chemistry.* Vol. 1. *Fundamental Principles, The Properties of Gases,* London: Longman. An authoritative, comprehensive description of early work (pre-1947) on the properties of gases.

Reid, R. C., Prausnitz, J. M., Poling, B. E. (1987), *The Properties of Gases and Liquids,* 4th ed., New York: McGraw-Hill. A full discussion of various group-contribution methods for the estimation of ideal-gas heat capacities, and a comparison of the results with experimental values.

Stull, D. R., Westrum, E. F., Sinke, G. C. (1969), *The Chemical Thermodynamics of Organic Compounds,* New York: Wiley. Ideal-gas heat capacities are tabulated as a function of temperature for several hundred organic compounds.

Wagner, W., de Reuck, K. M. (1987), *Oxygen. International Tables of the Fluid State—9,* Oxford, U.K.: Blackwell Scientific Publications. Extensive tabulations of the ideal-gas and real-gas isochoric and isobaric heat capacities.

Wakeham, W. A., Nagashima, A., Sengers, J. V. (Eds.) (1991), *Measurement of the Transport Properties of Fluids,* Oxford, U.K.: Blackwell Scientific Publications. An excellent description of all the important experimental techniques for transport-property measurement.

THERMIONIC EMISSION

Peter W. Hawkes, *Centre d'Elaboration des Matériaux et d'Etude des Structures et Laboratoire d'Optique Electronique du Centre National de la Recherche Scientifique, Toulouse, France*

Introduction 229
1. **Principles** 229
1.1 The Physics of Thermionic Emission 229
1.2 Plane Barriers 230
1.3 Schottky Emission 233
2. **Devices** 233
2.1 Typical Triode Electron-Gun Properties 234
2.2 Extended Schottky-Emission Guns 236
3. **Brightness** 237
3.1 Definitions 237
3.2 Coherence and radiometry 240
Glossary 242
Works Cited 243
Further Reading 243

INTRODUCTION

Electrons may be released from metals by several mechanisms: by heating, by application of a high electric field, by bombardment with photons or charged particles, or by a combination of these. Thermionic emission is provoked by heating the metal, to temperatures of the order of 2500 °C for tungsten or about 1100–1700 °C for lanthanum hexaboride, the most commonly used filament materials. The energy thus furnished to free electrons inside the metal enables them to pass over the potential barrier at the surface, which confines them to the interior at lower temperatures. In practice, an electric field is often present outside the metal. When this field is strong enough to lower the potential barrier but not sufficient to cause electrons to tunnel through it, we speak of Schottky emission.

Thermionic emission is widely used to produce the electron beams needed in cathode-ray tubes, x-ray tubes, and electron microscopes. The last are now capable of achieving resolutions of a few ångströms (1 Å = 100 pm = 0.1 nm), though for the very highest resolution, field emission (see FIELD EMISSION AND FIELD IONIZATION) is preferable. Thermionic emission is characterized by the total current produced, by the current density, and by the current density per unit solid angle, which is known as the brightness (measured in A m^{-2} sr^{-1}). This is closely analogous to, but not identical with, the radiant intensity of light sources, which is measured in W sr^{-1}. Other important characteristics of thermionic sources are the angular and energy distributions of the electrons.

1. PRINCIPLES

1.1 The Physics of Thermionic Emission

Thermionic emission is excited by heating a metal, but few metals begin to emit significantly before melting. In practice, only tungsten and certain borides, notably lanthanum hexaboride (LaB_6) and cerium hexaboride (CeB_6), are employed. The most robust is tungsten, the melting point of which is about 3380 °C; the working temperature of tungsten cathodes, T_c, is typically of the order of 2500 °C. At this temperature, tungsten slowly evaporates, limiting the cathode lifetime to between 12 and 20 h. The mean lifetime decreases rapidly with temperature, from some 200 h at 2200 °C to a mere 5 h at 2600 °C. Lanthanum hexaboride emits at a lower temperature—working values are 1100–1700 °C—thanks to its significantly lower work function ϕ_W (for LaB_6, ϕ_W = 2.6 eV, while for tungsten, ϕ_W = 4.5 eV), but a much better vacuum is required to protect the filament from bombardment by positively charged stray ions. For tungsten, a vacuum pressure

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

of 1–10 mPa is sufficient, whereas LaB_6 requires 10–100 μPa. Cerium hexaboride has been found to be superior to lanthanum hexaboride in some respects (Christman *et al.*, 1991) and is now available commercially. The work function of clean CeB_6 is very close to that of LaB_6, but the former can be much less volatile and hence have a longer lifetime. At a temperature of about 1480 °C, the material loss rate for LaB_6 is about four times that of CeB_6; at 1530 °C, the ratio has fallen to two, and at 1580 °C, the rates are about the same. In addition, CeB_6 is more resistant to contamination by carbon than LaB_6, an important property since such contamination leads to an increase in the work function and hence to deterioration of the emissive characteristics.

The theory of electron emission from metals resulting from thermal motion of the conduction electrons is based on a simple model describing the confinement of these electrons within the conductor, introduced by Sommerfeld (1928; see Sommerfeld and Bethe, 1933). The metal is treated as a box, inside which the potential is uniform, and at low temperature, the electrons fill the lowest-energy one-particle states up to a level known as the Fermi level. At higher temperatures, the energy (E) of the electrons is governed by the Fermi–Dirac distribution

$$f(E,T) = \frac{1}{1 + \exp\{(E - \phi_F)/kT\}}, \qquad (1)$$

in which k is Boltzmann's constant ($k = 1.38 \times 10^{-23}$ J/K) and ϕ_F is the Fermi energy level, at which $f(\phi_F,T) = \frac{1}{2}$ for all T (Fig. 1).

The various quantities that characterize the emission may be calculated with the aid of the function $f(E,T)$ together with a transmission coefficient $D(\mathbf{r},\mathbf{p})$; the latter denotes the ratio of the wave-mechanical intensity after and before passage through the barrier, where $\mathbf{r}$ and $\mathbf{p}$ denote position and momentum, respectively. For example, the electron charge density is given by

$$\rho(\mathbf{r}) = -\frac{2e}{h^3}\iiint f(E,T)D(\mathbf{r},\mathbf{p})d^3p \qquad (2)$$

where $-e$ is the charge on the electron ($e = 1.6 \times 10^{-19}$ C) and h is Planck's constant ($h = 6.62 \times 10^{-34}$ J·s). More important, the current-density vector $\mathbf{j}$ is given by

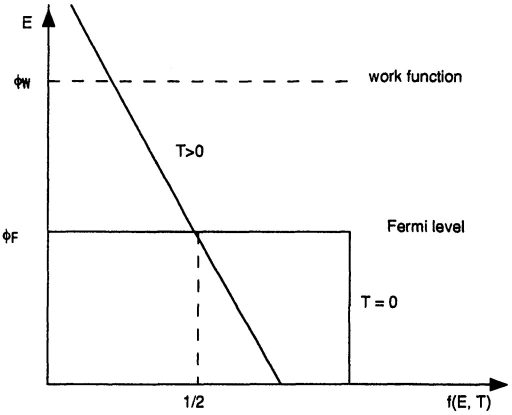

FIG. 1. The Fermi–Dirac distribution, shown as E vs $f(E,T)$ for $T = 0$ and a nonvanishing temperature. Note that $f(E,T) = \frac{1}{2}$ for all T at the Fermi energy level ϕ_F.

$$\mathbf{j}(\mathbf{r}) = \frac{2e}{mh^3}\iiint f(E,T)\mathbf{p}D(\mathbf{r},\mathbf{p})d^3p. \qquad (3)$$

We shall see in the next section how practical expressions for the quantities that describe real electron sources can be derived from these simple formulas.

1.2 Plane Barriers

The emitter in an electron gun usually consists of a hairpin-shaped filament, but only a small zone close to the tip emits significantly. It is hence a reasonable first approximation to neglect the curvature of the emissive area and consider transmission of electrons through a plane barrier. The potential barrier can then be represented by a function $V(z)$ of the coordinate perpendicular to the barrier, which we denote by z (Fig. 2). Conservation of energy then tells us that

$$\begin{aligned}(1/2m)(p_x^2 + p_y^2) &= \text{const.} = U_t,\\ (1/2m)p_z^2 + V(z) &= \text{const.} = U_n,\end{aligned} \qquad (4)$$

and of course $U_t + U_n = E = \text{const.}$ After a simple calculation, we find that the current density is given by

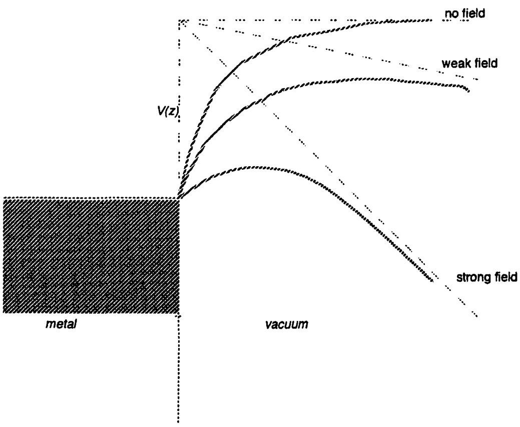

FIG. 2. Potential energy diagram for pure thermionic emission (no field), Schottky emission (weak field), and field emission (strong field). Note the lowering of the potential barrier in the weak-field case. (After Reimer, 1993, courtesy of Springer Verlag.)

$$j_z = \frac{4\pi mekT}{h^3} \times \int_{-\infty}^{\infty} \ln\left(1 + \exp\frac{\phi_F - U_n}{kT}\right) D(U_n)dU_n \qquad (5)$$

in the direction normal to the surface.

If it is permissible to neglect any electric field outside the filament, the potential function $V(z)$ will have the form of a step function:

$$V(z) = \begin{cases} -(\phi_F + \phi_W) & \text{for } z < 0, \\ 0 & \text{for } z \geq 0, \end{cases} \qquad (6)$$

in which the work function ϕ_W is a property of the material of which the cathode is composed. It is slightly temperature dependent. Electrons can escape only if $U_n \geq 0$, which implies that the transmission factor $D(U_n)$ is zero if $U_n < 0$ and unity if $U_n \geq 0$. The unit term in $f(E,T)$ is here negligible compared with the exponential. Because of the choice of zero of potential adopted in Eq. (6), the term $(E - \phi_F)$ in Eq. (1), where the origin of E was left arbitrary, becomes $(E + \phi_W)$, and we may hence write

$$f(E,T) = \exp\left(-\frac{E + \phi_W}{kT}\right). \qquad (7)$$

The current density is given by

$$j_s = A(kT)^2 \exp(-\phi_W/kT) \qquad (8)$$

in which

$$A = 4\pi me/h^3, \quad Ak^2 = 120\ \mathrm{A\,K^{-2}\,cm^{-2}}. \qquad (9)$$

Equation (8) is the Richardson–Dushman formula for the saturation current density, and we have written j_s rather than j_z to emphasize that this is an upper limit.

This formula was in fact originally obtained by Richardson (1902) on the basis of thermodynamic reasoning and the Drude model of electron conduction of 1900. Dushman (1923) later obtained the same expression by quantum-mechanical arguments and established the expression for A (apart from a factor 2, omitted since the notion of spin had not yet been introduced). The theory was later completed by Fowler (1928) and Nordheim (1929). In practice, the value of A may be appreciably lower than this theoretical value.

We can now calculate the various emission characteristics. The element of current density, d^3j, takes the form

$$d^3j = S(E,T)\cos\theta dE d\Omega \qquad (10)$$

in which the function

$$S(E,T) = \frac{1}{\pi} AE \exp\left(-\frac{E + \phi_W}{kT}\right) \qquad (11)$$

is known as the spectral radiance. The angle θ measures the direction of emission relative to the normal to the barrier (the z axis). $d\Omega$ is the element of solid angle, $d\Omega = \sin\theta d\theta d\varphi$, and φ is an azimuthal coordinate,

$$p_x = (2mE)^{1/2}\sin\theta\cos\varphi,$$
$$p_y = (2mE)^{1/2}\sin\theta\sin\varphi.$$

The total energy distribution of the current is given by integrating over all possible emission angles ($0 \leq \varphi \leq 2\pi$ and $0 \leq \theta \leq \pi/2$):

$$dj = AE\exp\left(-\frac{E + \phi_W}{kT}\right)dE, \qquad (12)$$

and a final integration over E again gives us (8). The mean energy is easily found to be

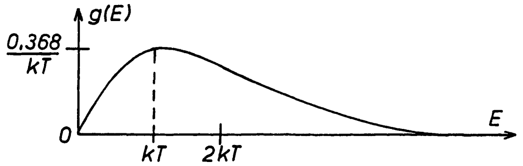

FIG. 3. The normalized current density as a function of energy. The long tail of the spectrum is striking. (After Reimer, 1993, courtesy of Springer Verlag.)

$\langle E\rangle = 2kT$ with standard deviation $\Delta E = kT\sqrt{2}$.

It is usual to define a normalized current-density distribution, $g(E)$, as follows:

$$g(E)dE = dj/j_s, \tag{13}$$

so that

$$g(E) = \frac{E}{(kT)^2}\exp\left(-\frac{E}{kT}\right). \tag{14}$$

This function is shown schematically in Fig. 3, where the most probable energy, $E_m = kT$, is indicated. The long tail of high-energy electrons is usually undesirable since for many applications, a small energy width is essential; energy-loss spectroscopy and high-resolution electron microscopy are examples of situations in which such a wide energy spread cannot be tolerated.

The full width at half-height is given by $\delta E = 2.45kT_c$; for $T_c = 2500$ °C, therefore, this energy width is about 0.5 eV. In practice, the energy width is increased by the Boersch effect (Boersch, 1954), caused by electron–electron interactions in regions where the current density is high. As a result, δE is typically of the order of 2 eV for thermionic guns. These interactions are extensively discussed by Rose and Spehr (1983) and Jansen (1991).

The angular emission distribution is obtained by integrating (10) over energy, which gives

$$\tilde{B}(\theta) = \frac{dj}{d\Omega} = \frac{j_s\cos\theta}{\pi}. \tag{15}$$

This bears a close resemblance to Lambert's cosine law for light sources. Like that law, it is obtained only if the transmission through the barrier is independent of emission direction.

The other parameter of interest in characterizing gun performance is the brightness (*Richtstrahlwert*). This quantity (B) measures the current density per unit solid angle:

$$B = j/\pi\theta_{\max}^2, \tag{16}$$

in which $\theta_{\max}$ is the largest angle of emission; B is given to a good approximation by

$$B = j_s E/\pi kT_c. \tag{17}$$

Typical values are listed in Table 1. The

Table 1. Typical values of the characteristics of thermionic and field-emission cathodes.

	Thermionic emission from tungsten (hairpin filament)	Thermionic emission from LaB$_6$	Field emission from tungsten
ϕ_W (eV)	4.5	2.7	4.5
Ak^2 (A cm^{-2} K^{-2})	75–120	30	—
js (A cm^{-2})	1–5	25–50	10^4–10^6
Maximum current emitted (μA)	10–100	40	1–10
Cathode temperature (°C)	2500	1100–1700	20 (cold field emission) 700–1500 (thermal field emission)
Brightness (A cm^{-2} sr^{-1})	50 000 at 10 keV (1–5) × 10^5 at 100 kV	3 × 10 at 15 kV	5 × 10^7–5 × 10^8 at 20 kV 2 × 10^8–2 × 10^9 at 100 kV
($dI/d\Omega$) (A sr^{-1})	10^{-2}	—	10^{-4}
Crossover diameter (μm)	20–50 (hairpin) 10–30 (point)	10–20	(5–10) × 10^{-3}
ΔE (eV)	1–2	0.5–2	0.2–0.5
Lifetime (h)	25	150–1000	>1000 (for cold field emission)
Vacuum (mPa)	1–10	10^{-2}–1	10^{-5}–10^{-4}

brightness is of great importance because it can be shown to be invariant: the brightness of an electron beam remains constant as it traverses lenses and diaphragms (stops). This invariant property is strictly true only for the generalized brightness function (see Hawkes and Kasper, 1989, Chap. 47, for full discussion, and Sec. 4 below) but can be regarded as valid for the mean brightness B. An important consequence is that, for a given brightness, the current density j is proportional to $\theta^2_{\max}$ [Eq. (16)]. In some practical applications, such as low-angle electron diffraction and the electron microscope study of magnetic specimens, the angular spread of the beam must imperatively be kept very small, with the result that the current density is low and long exposure times are necessary. For this reason, field-emission sources, which have much higher brightness, are to be preferred for such applications. Such sources are also indispensable when a very small probe is required in scanning microscopes to keep the scan time per frame within reasonable bounds. The replacement of thermionic guns by field-emission guns has opened up a new range of studies to scanning electron microscopes (see Reimer, 1985, or Joy, 1985, Tuggle *et al.*, 1985, and Orloff, 1985). The development of field-emission guns was the essential step in the development of the scanning transmission electron microscope by A. V. Crewe (Crewe, 1973).

1.3 Schottky Emission

We have assumed above that any electric field in the space outside the emissive filament is negligible. In practice, the electrons are accelerated immediately after emission to the required energy by applying the appropriate voltages to the anodes of the electron gun, as discussed below. The filament is hence usually not in field-free space, and, in order to estimate its emission characteristics realistically, the surrounding field must not be neglected. The simplest way of accounting for such fields is to replace the work function ϕ_W in the Richardson–Dushman equation (8) by a smaller value,

$$\phi_S = \phi_W - (e^3F/4\pi\epsilon_0)^{1/2}, \tag{18}$$

where F denotes the electric-field strength at the emitter surface. The emission is thereby increased by virtue of the "Schottky effect." The terms "Schottky emission" and "Schottky effect" are reserved for the current increase in the thermionic emission regime: the potential barrier is lowered and narrowed, but the electrons continue to escape thanks to their thermal energy. The electric-field strength is far too low to generate any perceptible current by tunneling. In practice, however, guns are frequently operated in an intermediate regime, in which the electric field is high enough to provoke some tunneling, though far less than in the case of true field emission. This situation is known as "extended Schottky emission." It can be modeled very simply by using a more complicated model for $D(U_n)$; instead of setting the latter equal to zero or unity, we write

$$D(U_n) = \frac{1}{1 + \exp[(\phi_m - U_n)/c]}, \tag{19}$$

in which

$$c = \frac{\mathrm{tr}(4\pi\epsilon_0 eF^3)^{1/4}}{\pi m^{1/2}} \tag{20}$$

and ϕ_m is the potential maximum at the barrier. We then find that

$$j_s = A(kT)^2 \exp\left(-\frac{\phi_s}{kT}\right)\frac{\pi q}{\sin\pi q}, \tag{21}$$

in which

$$q = c/kT \geq 0.7. \tag{22}$$

The standard deviation of the energy distribution is now

$$\Delta E = \sqrt{2}kT\,\pi q/\sin\pi q. \tag{23}$$

Extended Schottky emission is typically obtained from a zirconiated tungsten filament, ZrO/W; the absorbed zirconium oxide lowers the work function of the tungsten to about 2.7 eV, and a suitable working temperature is 1500 °C. We give more details in Sec. 2.

2. DEVICES

Thermionic emission has been used in a host of devices, some of which have declined in importance, notably the vacuum tubes or "valves" that have been replaced by transistors so completely that older equipment must often be completely replaced when a vacuum tube fails for want of the spare part. Emitters have been designed with large surface areas to provide broad beams; dispenser cathodes, in which an active material such as barium diffuses through a sintered, porous form of a base metal, typically tungsten or molybdenum, provide the high current densities required in some applications, in microwave tubes for example. The three-electrode or triode gun remains of great importance in many instruments, and we have therefore chosen to describe this example in detail here. The three electrodes are the emissive filament, or cathode; the control electrode or "grid," usually known as the wehnelt; and the anode. Such guns, especially those using LaB_6 emitters, are under further development in the expanding field of electron nanolithography.

2.1 Typical Triode Electron-Gun Properties

Conventional thermionic guns consist of three electrodes:

1. the filament, which has the form of a hairpin or a sharp point supported on a V-shaped support, through which the heating current passes;
2. the wehnelt or control electrode (Wehnelt, 1903; Wehnelt and Jentsch, 1909), the potential of which is negative relative to that of the filament; and
3. the anode, at high positive potential, which accelerates the electrons to the desired "accelerating voltage," which can reach 200 kV;

beyond this, extra electrodes are required and we enter the realm of particle accelerators. The behavior of triode guns is determined by the geometry of the three electrodes and the values of the applied voltages. A first distinction is between short-focus and telefocus guns. In the former, the electrons emitted by the filament are rapidly brought to a focus or crossover zone, whereas in telefocus guns, the filament may well protrude beyond the wehnelt, which is often conical, and the electrons are brought to a focus far from the gun (Fig. 4).

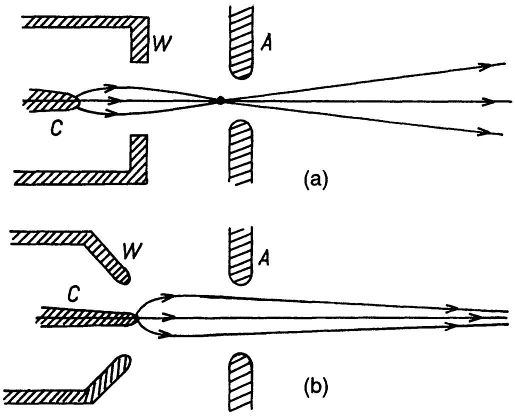

FIG. 4. Schematic diagram showing typical rays in (a) a short-focus and (b) a long-focus or Fernfokus gun. (After Hawkes and Kasper, 1989, courtesy of Academic Press.) C, cathode; W, wehnelt; A, anode.

The principal features of a short-focus gun are shown schematically in Fig. 5. Pencils of rays are seen to emerge from various points on the filament tip. They are brought to a focus in a broad region, the narrowest

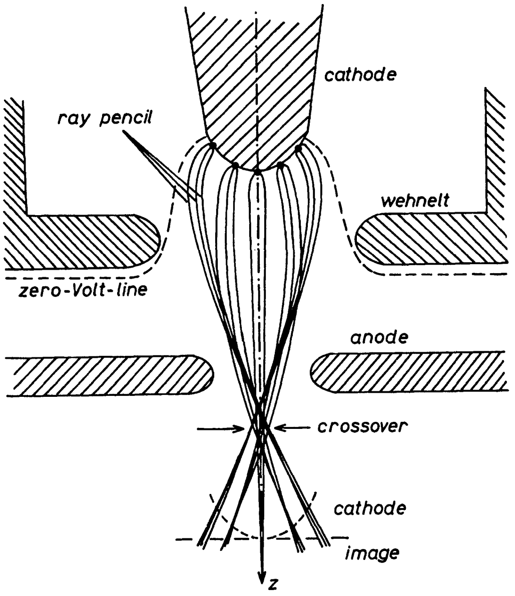

FIG. 5. Crossover formation in a conventional triode gun. The zero-volt line limits the zone that contributes to the beam. The crossover and not the cathode image is the effective source for any subsequent lenses. (After Hawkes and Kasper, 1989, courtesy of Academic Press.)

part of which is known as the crossover. An important element is the "zero-volt line," which delimits the zone on the filament surface from which electrons will be emitted. We shall see below that the position of this line moves across the filament as the wehnelt bias is altered; if it is brought to the axis, no electrons are emitted and the corresponding voltage is the cutoff bias.

It is usual to supply the bias not from an external voltage source but by introducing a large resistor (of resistance $R_W = 1 - 10$ MΩ) in the high-voltage supply (Fig. 6). In this situation, the current emitted saturates at a value determined by R_W; this saturation has often been attributed to the buildup of space charge in front of the cathode. This does indeed occur but is not the cause of the saturation, which is a consequence of the behavior of the circuit containing the large resistor; space charge has a harmful effect only at high filament temperatures.

The detailed behavior of short-focus guns may be understood qualitatively, and to a limited extent quantitatively, with the aid of a model that has been extensively studied by Lauer (reviewed in Lauer, 1982). In this, the electric field in the vicinity of the emitting area of the filament is assumed to have spherical symmetry, which is very plausible for both pointed and hairpin tips. Between the wehnelt and the anode, the electric field is taken to be uniform, at least close to the optic axis. Careful examination of this model shows that the pencils of electron rays emerging from the cathode appear to come from a virtual cathode, smaller in radius than the real tip; if one pursues the rays further back toward the center of the cathode, they are found to define a circular zone, the entrance pupil of the gun (Fig. 7). This entrance pupil is conjugate to the circle of least confusion outside the gun if the crossover is formed beyond the anode, and this exit pupil and the crossover then coincide. If, however, the crossover is formed between the wehnelt and the anode, as is often the case, the exit pupil will be virtual. In both cases, the effective source for further calculations, of illumination conditions in microscopes for example, is the crossover as seen from the space beyond the anode.

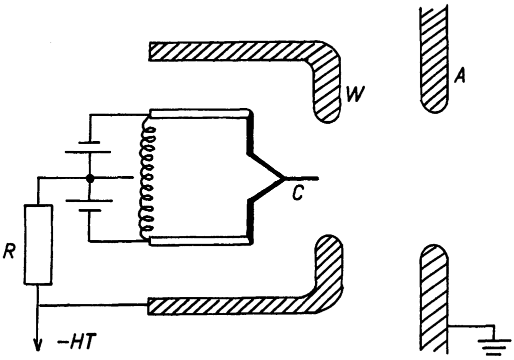

FIG. 6. The autobias circuit of a triode gun. The resistor *R*, of a few megohms resistance, provides the wehnelt bias, and the associated feedback stabilizes the emission. (After Hawkes and Kasper, 1989, courtesy of Academic Press.)

We have seen that the voltage applied to the wehnelt determines the position of the zero equipotential at the filament and hence the size of the emitting area. Figure 8 shows three situations, in which the wehnelt bias is progressively more negative from (a) to (c). The zero equipotential is seen to move toward the optic axis, and a small further increase will cut off the emission altogether. The emissive area is not the only parameter affected by the wehnelt voltage, however. As the wehnelt bias is made less negative, the structure of the beam changes in a characteristic way (Fig. 9). Initially a full beam, close to cutoff, the beam becomes hollow, in the sense that the current density is higher around the periphery than at the center. A further (positive) increase renders the beam compact again, then a particularly narrow waist is formed (the telefocus mode), after which the beam again becomes hollow (Fig. 9). Full details are to be found in Lauer (1982).

The foregoing theory is essentially paraxial, in the sense that the quantitative predictions assume that the paraxial approximation of electron optics (see CHARGED-PARTICLE OPTICS) is applicable. In fact, the electrons may be emitted relatively far from the axis and travel at steep angles to it, for the first part of their trajectories at least, and geometric aberrations are hence to be expected. We illustrate this in Fig. 10, where effects due to the failure of the paraxial approximation are illustrated.

The filament material is commonly tungsten or thoriated tungsten. At the cost of a better vacuum, higher brightness can be ob-

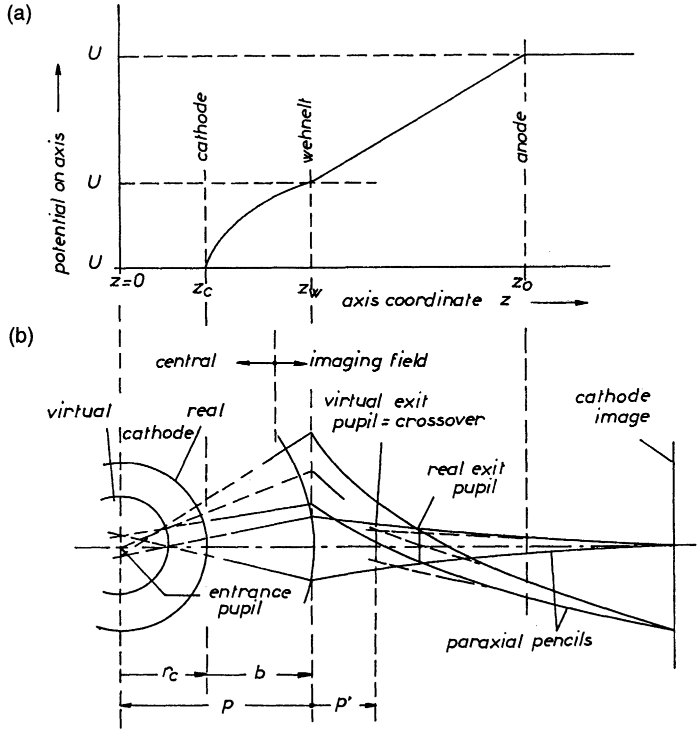

FIG. 7. Model used to study the characteristics of triode guns. The associated vocabulary is defined here. (After Hawkes and Kasper, 1989, courtesy of Academic Press.)

tained by the use of single-crystal lanthanum hexaboride cathodes. The emission characteristics are closely related to the crystal orientation, and ⟨100⟩ and ⟨346⟩ have been found suitable on both experimental and theoretical grounds for the rod axis (Table 2). The advantages of using rare-earth and alkaline-earth oxides were pointed out long ago by Lafferty (1951): the peak emission density can be high (1 A/mm^2), the lifetime can be of the order of 100 h, and the improved vacuum that is necessary is far worse than that required for field emission. The early exploitation of these oxides, LaB$_6$ in particular, which rapidly established itself as the most suitable, was hampered by the fact that indirect heating seemed to be inevitable because of the high chemical reactivity of the lanthanum hexaboride with the material on which it was supported. Once this problem had been solved (Crawford, 1979), reliable mountings emerged and direct heating became usual; some typical designs are shown schematically in Fig. 11. A modern mounting of the "mini-Vogel" type is also shown.

The emission from LaB$_6$ depends on the source geometry, of course, but above all on the shape and radius of the cathode tip. The latter is usually conical with a pointed end portion that is approximately spherical, of the order of 2–5 μm in radius. However, a better performance is obtained by forming a faceted tip (Fig. 12), the faces of which are typically (110) planes since the work function for these is low (2.6 eV); the angle of the cone is 90°. Figure 13 shows the brightness as a function of temperature for a range of tip shapes.

A significant improvement in performance can in principle be achieved by artificial limitation of the area from which electrons are emitted and by suppression of the wehnelt electrode. The calculations of Broers *et al.* (1991) suggest that the current density and brightness could be increased by more than an order of magnitude in this way.

2.2 Extended Schottky-Emission Guns

Here, the electric field at the filament is not negligible. The emissive tip protrudes beyond the wehnelt (Fig. 14) so that the anode potential exerts a direct effect on the tip. The

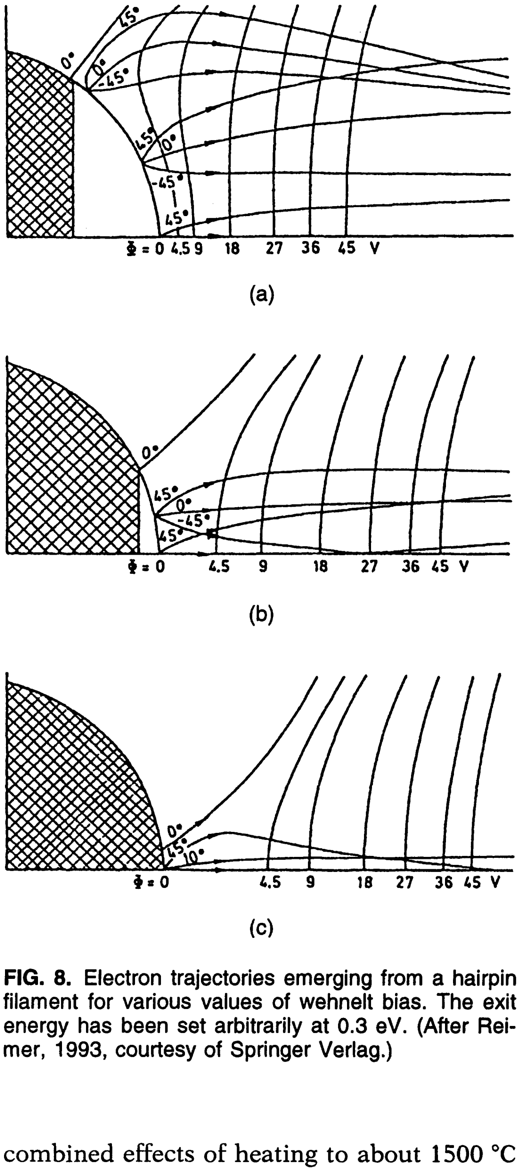
45°
0°
-45°
45°
0°
-45°
45°
Φ = 0 4.5 9 18 27 36 45 V
(a)
0°
45°
0°
-45°
45°
Φ = 0 4.5 9 18 27 36 45 V
(b)
0°
45°
10°
Φ = 0 4.5 9 18 27 36 45 V
(c)

FIG. 8. Electron trajectories emerging from a hairpin filament for various values of wehnelt bias. The exit energy has been set arbitrarily at 0.3 eV. (After Reimer, 1993, courtesy of Springer Verlag.)

combined effects of heating to about 1500 °C and the application of an electric field, of the order of 0.5–1 V/nm, cause migration of surface atoms on the filament in order to achieve minimum surface free energy. A facet forms at the tip of the filament, as shown in Fig. 14 (dashed line). The properties of this geometry have been studied experimentally and by computer modeling by Tuggle and Swanson (1985). Figure 15 shows their calculated Schottky plots of current density against (applied field)$^{1/2}$ on a logarithmic-linear scale for several values of the work function. A useful comparison of these guns, true field-emission guns, and LaB_6 guns is given in a survey by Orloff (1984).

3. BRIGHTNESS

Source brightness and source coherence are extremely important in electron microscopy. In scanning instruments, a high brightness is essential if a high-resolution image is to be generated in a reasonably short time. In the traditional scanning electron microscope (SEM), it is usually sufficient to use a lanthanum hexaboride cathode, but for scanning transmission electron microscopy (STEM) and for very high-resolution SEM, field emission is the only choice. In conventional transmission electron microscopy, the image resolution is degraded by effects caused by wavelength spread, since the chromatic aberration of electron lenses is enormous, and by the size of the source. The latter may be understood by recalling that the electron microscope image is the result of scattering in the specimen and not of absorption; electrons arriving from different points on an extended source will create images at slightly different places in the image plane, and the superimposition of all these will blur the resulting overall image. Both the wavelength spread and the source size must hence be kept very small, which is equivalent to requiring the source to be highly coherent, both temporally (negligible wavelength spread) and spatially (negligible source size). In energy-loss spectroscopy, too, it is essential to maintain the energy spread of the electron-beam incident on the specimen as narrow as possible.

3.1 Definitions

The measured brightness of electron sources is inevitably a mean brightness, $\bar{B}$:

$$\bar{B} = \Delta I / \Delta a \Delta \Omega \tag{24}$$

in which ΔI is the current flowing across an area Δa into a solid angle $\Delta\Omega$. The formal theory of brightness, on the other hand, is concerned with a function, the brightness function, defined by

$$B(x,y,z,x',y') = \lim_{\Delta a \to 0} \lim_{\Delta\Omega \to 0} \left(\frac{\Delta I}{\Delta a \Delta \Omega} \right). \tag{25}$$

Here, Δa is the element of area $\Delta x \Delta y$ in the neighborhood of a point (x,y,z), and $\Delta\Omega = \Delta x' \Delta y'$ is the element of solid angle

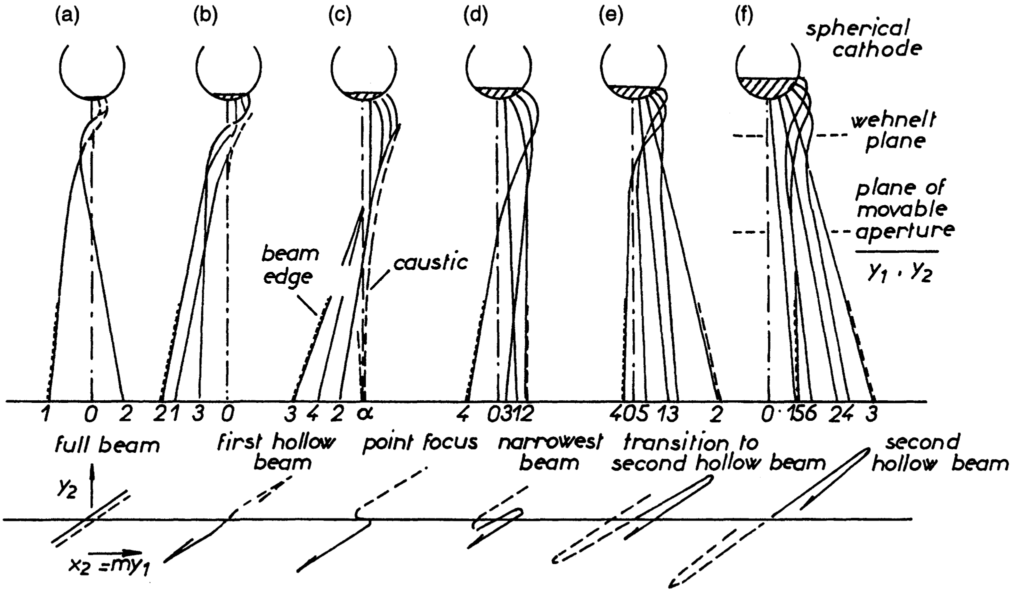
(a)
(b)
(c)
(d)
(e)
(f)
spherical cathode
wehnelt plane
plane of movable aperture
$y_1 \cdot y_2$
beam edge
caustic
1 0 2 21 3 0 3 4 2 α 4 0312 405 13 2 0·156 24 3
full beam
first hollow beam
point focus
narrowest beam
transition to second hollow beam
second hollow beam
y_2
$x_2 = my_1$

FIG. 9. Thermionic emission in a triode gun based on the paraxial approximation. Left: wehnelt bias close to cutoff; left to right: the wehnelt bias is raised and the sequence of beam structures illustrated is observed. The patterns below are emittance diagrams. (After Lauer, 1982, courtesy of Academic Press.)

around the direction characterized by the gradient (x',y'). The prime denotes differentiation with respect to z, the optic axis. This is essentially a paraxial definition, and this restriction led to the introduction of a *generalized* brightness function $B(\mathbf{r},\mathbf{s})$, in which $\mathbf{r}$ is the position vector (x,y,z) and $\mathbf{s}$ is a unit vector in an arbitrary direction. The element $\Delta\Omega$ is around the direction $\mathbf{s}$, and the area Δa is normal to $\mathbf{s}$.

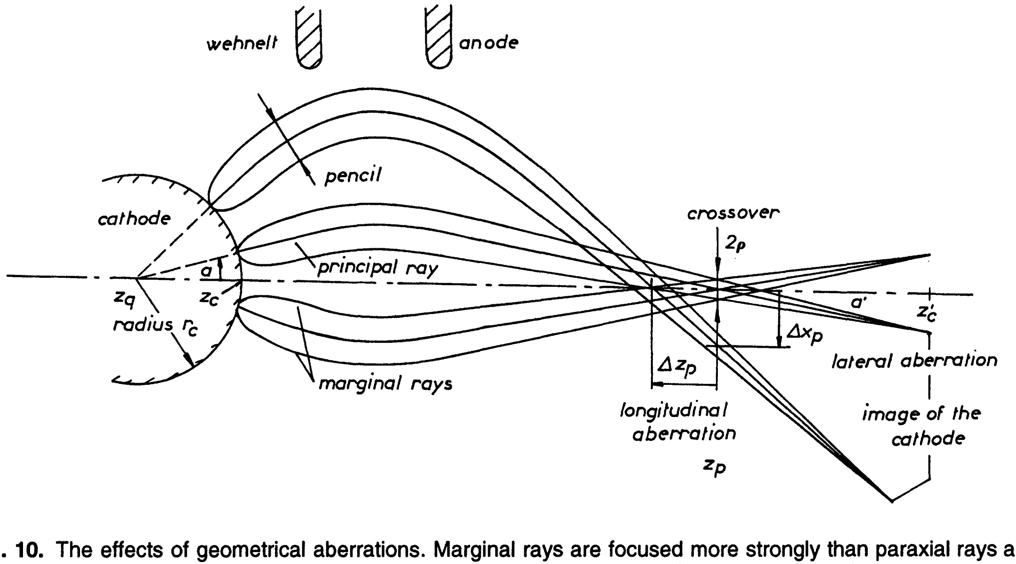
wehnelt
anode
pencil
cathode
principal ray
a
z_q
z_c
radius r_c
marginal rays
crossover
2ρ
a'
z_c'
Δx_p
lateral aberration
Δz_p
longitudinal aberration
z_p
image of the cathode

FIG. 10. The effects of geometrical aberrations. Marginal rays are focused more strongly than paraxial rays and the crossover is broadened. (After Hawkes and Kasper, 1989, courtesy of Academic Press.)

Table 2. Thermionic work function for various LaB$_6$ crystal orientations.

Orientation	ϕ_W (eV)
100	2.57
111	2.90
110	2.64
210	2.46
310	2.42
321	2.8
346	2.4

(After Gesley and Swanson, 1984, and Swanson *et al.*, 1980.)

With the aid of this function, we can obtain an explicit expression for the brightness of thermionic sources:

$$B(\mathbf{r},\mathbf{s}) = \frac{e j_s(\mathbf{r})}{\pi k T}\, \Phi_T(1 + \epsilon \Phi_T) \qquad (26)$$

in which j_s is the saturation current calculated earlier [Eq. (8)] and $\epsilon = e/2m_0c^2$. If Φ is the accelerating voltage, then Φ_T denotes

$$\Phi_T = \Phi + kT/e. \qquad (27)$$

We see that B is independent of $\mathbf{s}$, a reminder that Lambert's law applies to thermionic sources. At the cathode surface, where by convention Φ vanishes, the brightness is just j_s/π; it is clearly the maximum value of the angular distribution function $\tilde{B}(\theta)$ of Eq. (15). Beyond the anode, and hence far from the cathode, the term Φ (of the order of tens or hundreds of kilovolts) is much larger than kT/e ($\sim$0.3 V), and the brightness is proportional to the (relativistically corrected) accelerating voltage $\Phi(1 + \epsilon\Phi)$. The quantity $B/\Phi(1 + \epsilon\Phi)$ is known as the normalized brightness.

From Eq. (25), it is seen that

$$\Delta I = B(\mathbf{r},\mathbf{s};\nu)\cos\theta \Delta a \Delta \Omega. \qquad (28)$$

In this formula, we have introduced the frequency ν to emphasize that the relation is to be understood to apply to a particular wavelength or frequency; this is not of importance in traditional radiometry but will be necessary when we come to consider the relation between the radiometric quantities and coherence. It is usual to introduce the radiant intensity J and the radiant emittance E:

$$J(\mathbf{s},\nu) = \int B \cos\theta da, \qquad (29)$$

$$E(\mathbf{r},\nu) = \int B \cos\theta d\Omega, \qquad (30)$$

so that

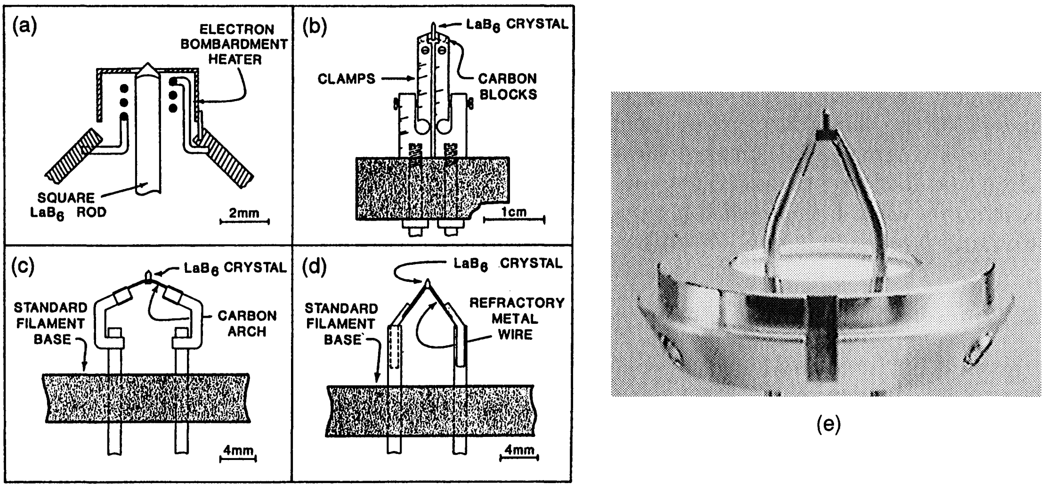

FIG. 11. Various ways of mounting LaB$_6$ tips: (a) Rod heated by electron bombardment; (b) tip clamped between carbon blocks; (c) tip mounted on a carbon arch; (d) tip mounted on a refractory metal arch; (e) a modern mini-Vogel mounting. [(a)–(d) after Crawford, 1979, courtesy of Scanning Microscopy International. (e) courtesy of FEI Company.]

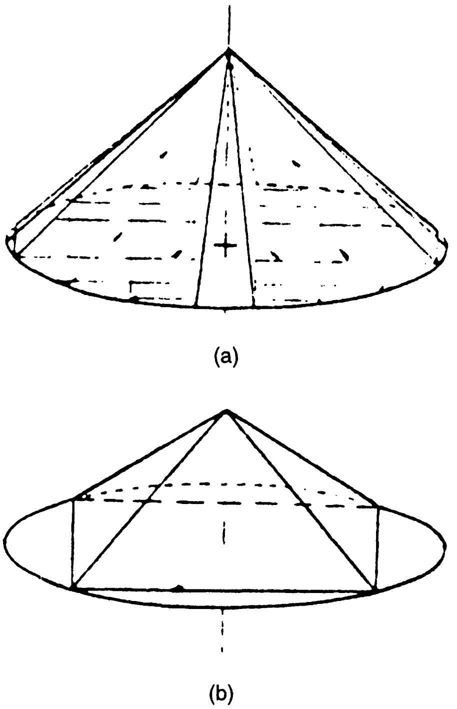

FIG. 12. The tip shape of LaB$_6$ emitters: (a) Cone-shaped tips; several crystallographic planes (with different work functions) participate in the emission. (b) Faceted tip, only a single family of planes emit, typically (110). (After Hohn, 1985, courtesy of Scanning Microscopy International.)

$$\begin{aligned} I &= \int J(\mathbf{s},\nu)d\Omega, \\ &= \int E(\mathbf{r},\nu)da. \end{aligned} \tag{31}$$

We note in passing that when B is independent of the direction $\mathbf{s}$, then I collapses to $J_0 \cos\theta$, where J_0 is a constant for a given source: the fact that Lambert's cosine law is obeyed only when B is a function only of position and frequency is now obvious.

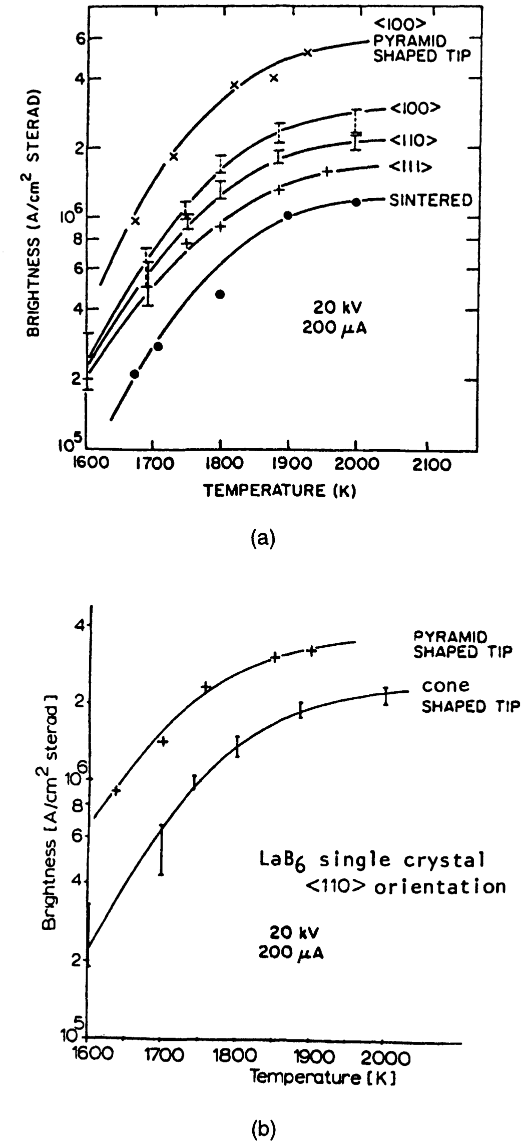

FIG. 13. Mean brightness as a function of temperature for various LaB$_6$ tips. (After Hohn, 1985, courtesy of Scanning Microscopy International.)

3.2 Coherence and Radiometry

Most of the foregoing material is based on classical mechanics, apart from the discussion of the emission process itself, where some elementary quantum mechanics was required. An important characteristic of all sources is, however, their degree of coherence, and in order to study this, we need to bring together the classical and the quantum approaches. The degree of coherence of a source, whatever its nature, is a measure of any relation between the emission from different points on the source and at different wavelengths. It is therefore expressed in terms of a correlation function, which gives a quantitative measure of this relation. It is important in practice to understand the coherence properties of the source used, in microscopy and holography in particular; in the former, the source coherence is directly re-

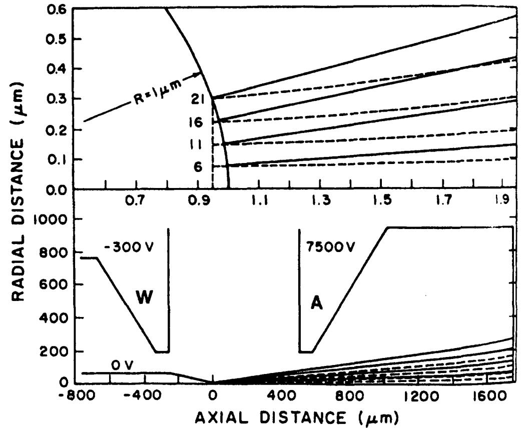

FIG. 14. Schottky emission from a ZrO/W tip. A facet forms at the tip as shown in the enlargement (above). Notice that the tip protrudes beyond the wehnelt (W) toward the anode (A). (After Tuggle and Swanson, 1985, courtesy of the American Vacuum Society.)

lated to the performance of the instrument, and in the latter, a high degree of coherence is imperative.

We shall see that coherence, brightness, and the associated radiometric quantities are intimately related. It is on this aspect of source coherence that we concentrate here.

In the practical study of coherence, it is convenient, and in almost all cases permissible, to discuss temporal coherence and spatial coherence separately. The temporal coherence of a source is determined by the range of wavelengths and hence of particle energies in the beam emitted. The spatial coherence is a measure of source size, the smallest sources being the most coherent. These are important quantities for electron microscope imaging, for the highest resolutions can be obtained only if the illumination is both temporally and spatially highly coherent. For a detailed and self-contained account of this semi-classical approach to the effect of partial coherence of electron image formation, see Chap. 66 of Hawkes and Kasper (1994). The formal theory of coherence (presented in detail in Born and Wolf, 1980, Chap. 10) is expressed in terms of the mutual coherence function, which is just the cross-correlation between the values of electron wave function (in electron optics) or of the analytic signal associated with, for example, the electric-field vector (in light optics) at two points; it is denoted by $\Gamma(\mathbf{r}_1,\mathbf{r}_2;\tau)$, in which $\mathbf{r}_1$ and $\mathbf{r}_2$ are the coordinates of two points and τ is the difference between the times of arrival of the wave fronts at some other point (see Born and Wolf, 1980, or Hawkes and Kasper, 1994, Chap. 78 for extensive explanation). The temporal Fourier transform of the mutual coherence function plays an important role in the following argument. It is essentially the (temporal) spectrum of $\Gamma(\mathbf{r}_1,\mathbf{r}_2;\tau)$. We write

$$\Gamma(\mathbf{r}_1,\mathbf{r}_2;\tau) = \int_0^\infty W(\mathbf{r}_1,\mathbf{r}_2;\nu)\exp(-2\pi i\nu\tau)d\nu \tag{32}$$

and refer to W as the cross-spectral density.

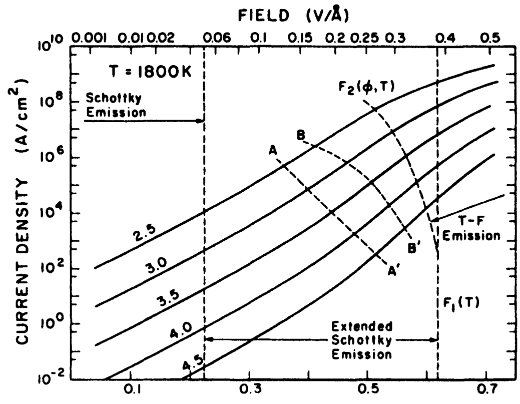

FIG. 15. Current emitted vs electric field at the tip for various regimes, for ZrO/W. (After Tuggle and Swanson, 1985, courtesy of the American Vacuum Society.)

The spatial spectrum of $\Gamma(\mathbf{r}_1,\mathbf{r}_2;\tau)$ is also of central importance. It is convenient to introduce the frequency ν rather than the time τ in this function, $A(\mathbf{s}_1,\mathbf{s}_2;\nu)$:

$$W(\mathbf{r}_1,\mathbf{r}_2;\nu) = \int_{-\infty}^{\infty} A(\mathbf{s}_1,\mathbf{s}_2;\nu)\exp\left\{\frac{2\pi i}{\lambda}(\mathbf{s}_1\cdot\mathbf{x}_1 + \sigma_1 z_1 - \mathbf{s}_2\cdot\mathbf{x}_2 - \sigma_2 z_2)\right\}d^3s_1d^3s_2, \tag{33}$$

where $\sigma^2 = 1 - \mathbf{s}\cdot\mathbf{s}$ and $\mathbf{r} = (\mathbf{x},z)$. It can be shown straightforwardly that, far from the source, the current J is given by

$$J(\mathbf{r}) = \Gamma(\mathbf{r},\mathbf{r};0) \sim \int\left(\frac{\lambda}{r}\right)^2\cos^2\theta A(\mathbf{s},\mathbf{s};\nu)d\nu \tag{34}$$

(the symbol $\sim$ denotes "tends asymptotically to").

But a change of coordinates

$$\begin{aligned}\mathbf{r}' &= \mathbf{r}_1 - \mathbf{r}_2,\\ \mathbf{r} &= \tfrac{1}{2}(\mathbf{r}_1 + \mathbf{r}_2),\end{aligned} \tag{35}$$

enables us to write

$$A(\mathbf{s},\mathbf{s};\nu) = \frac{1}{\lambda^2 \cos\theta}\int b(\mathbf{r},\mathbf{s};\nu)d^3r, \tag{36}$$

where $\mathbf{r}$ is now assumed to lie in the source plane and b denotes

$$\begin{aligned} b(\mathbf{r},\mathbf{s};\nu) = \frac{\cos\theta}{\lambda^2}\int W\left(\frac{\mathbf{r}+\mathbf{r}'}{2}, \frac{\mathbf{r}-\mathbf{r}'}{2}; \nu\right)\\ \times \exp\left(-\frac{2\pi}{\lambda}\mathbf{r}'\cdot\mathbf{s}\right)d^3r'. \end{aligned} \tag{37}$$

The far-field intensity is hence

$$I = \int_{2\pi} d\Omega \cos\theta \int b(\mathbf{r},\mathbf{s};\nu)d^3r, \tag{38}$$

and the function b plays exactly the same role in the radiometric formulas as the "incoherent" brightness B [Eq. (16)]. It seems therefore that the "generalized radiance" b, introduced tentatively by Walther (1968) and studied thoroughly by Marchand and Wolf (1974), is the partially coherent counterpart of the incoherent brightness B.

Unfortunately, the function b proves to possess certain properties incompatible with the intuitive notion of brightness. It is certainly real—b and its complex conjugate are identical—but it is not necessarily positive, even though the integral over any real source area is positive. It is as though a few electrons in a beam from a partially coherent source might be found to be traveling in the wrong direction! Worse, it may fail to vanish in the source plane outside the emitting area. Both of these unacceptable features vanish for a certain class of sources, known as *quasihomogeneous sources.* For these, the cross-spectral density $W(\mathbf{r}_1,\mathbf{r}_2;\nu)$ can be written in separated form:

$$W(\mathbf{r}_1,\mathbf{r}_2;\nu) = I\left(\frac{\mathbf{r}_1+\mathbf{r}_2}{2}, \nu\right)m(\mathbf{r}_1 - \mathbf{r}_2; \nu). \tag{39}$$

Here I is a *slow* function of its spatial argument whereas m is a *fast* function of $\mathbf{r}_1 - \mathbf{r}_2$. The brightness b now separates into two factors, one of which is essentially I at the source while the other is the (spatial) spectrum of m. The brightness is now never negative and behaves in all respects like a classical radiometric brightness.

GLOSSARY

Brightness: Quantity that measures the current density emitted by a source per unit solid angle.

Conventional Transmission Electron Microscope: Electron microscope in which a broad beam of electrons illuminates the specimen and lenses are used to enlarge the resulting image in the plane of recording or observation.

Cross-spectral Density: Temporal Fourier transform of the mutual coherence function.

Electron Gun: A structure typically consisting of three electrodes: an emitter (cathode), a control electrode (wehnelt), and an accelerating electrode (anode), capable of providing an electron beam.

Fermi Level: The maximum energy level of electrons in a metal at absolute zero temperature; it is the energy at which the Fermi–Dirac distribution reaches the value 0.5 for all temperatures.

Mutual Coherence Function: Cross-correlation function of the wave functions at two points.

Scanning Electron Microscope: Electron microscope in which the illuminating beam is focused to a small spot on a massive specimen; deflection fields cause the focused spot to explore the specimen, thereby giving rise to various signals (backscattered electron current, secondary electron current, x rays, cathodoluminescence, electron-beam–induced current), which are collected and used to form images on monitors scanned in synchronism.

Scanning Transmission Electron Microscope: Electron microscope in which the illuminating beam is focused to a small spot on a thin specimen and the transmitted beam is collected on various detectors placed downstream from the specimen; the signals obtained are displayed on monitors scanned in synchronism.

Work Function: Smallest energy required to take an electron from inside a metal to just outside it.

Works Cited

Boersch, H. (1954), *Z. Phys.* **139,** 115–146.

Born, M., Wolf, E. (1980), *Principles of Optics,* 6th ed., Oxford, U.K.: Pergamon.

Broers, A., Xia, S., Maloney, C., Zhu, X., Munro, E. (1991), *J. Vac. Sci. Technol. B* **9,** 2929–2933.

Christman, L. A., Schwind, G. A., Swanson, L. W. (1991), in: G. W. Bailey (Ed.), *Proceedings of the 49th Annual Meeting, Electron Microscopy Society of America,* San Francisco: San Francisco Press, pp. 346–347.

Crawford, C. K. (1979), *Scanning Electron Microsc.* (Pt. I), 19–30.

Crewe, A. V. (1973), "Production of Electron Probes Using a Field Emission Source," *Prog. Opt.* **11,** 223–246.

Dushman, S. (1923), *Phys. Rev.* **21,** 623–636.

Fowler, R. H. (1928), *Proc. R. Soc. London A* **117,** 549–552; **118,** 229–232.

Gesley, M. A., Swanson, L. W. (1984), *Surf. Sci.* **146,** 583–599.

Hawkes, P. W., Kasper, E. (1989), *Principles of Electron Optics,* Vols. 1 and 2, London: Academic.

Hawkes, P. W., Kasper, E. (1994), *Principles of Electron Optics,* Vol. 3, London: Academic.

Hohn, F. J. (1985), "Development and Use of High Brightness Lanthanum Hexaboride Electron Guns," *Scanning Electron Microsc.,* 1327–1338.

Jansen, G. H. (1991), "Coulomb Interactions in Particle Beams," *Adv. Electron. Electron Phys.,* Suppl. **21**.

Joy, D. C. (1985), *J. Microsc. (Oxford)* **140,** 283–292.

Lafferty, J. M. (1951), *J. Appl. Phys.* **22,** 299–309.

Lauer, R. (1982), "Characteristics of Triode Electron Guns," *Adv. Opt. Electron Microsc.* **8,** 137–206.

Marchand, E. W., Wolf, E. (1974), *J. Opt. Soc. Am.* **64,** 1219–1226.

Nordheim, L. (1929), *Phys. Z.* **30,** 177–196.

Orloff, J. (1984), in: J. J. Hren, F. A. Lenz, E. Munro, and P. B. Sewell (Eds.), *Electron Optical Systems for Microscopy, Microanalysis and Microlithography,* Chicago: Scanning Electron Microscopy, pp. 149–162.

Orloff, J. (1985), *J. Microsc. (Oxford)* **140,** 303–311.

Reimer, L. (1985), *Scanning Electron Microscopy,* Berlin: Springer.

Reimer, L. (1997), *Transmission Electron Microscopy,* 4th ed., Berlin: Springer.

Richardson, O. W. (1902), *Proc. Cambridge Philos. Soc.* **11,** 286–295.

Rose, H., Spehr, R. (1983), "Energy Broadening in High-Density Electron and Ion Beams: The Boersch Effect," *Adv. Electron. Electron Phys.,* Suppl. **13C,** 475–530.

Sommerfeld, A. (1928), *Z. Phys.* **47,** 1–32 and 43–60.

Sommerfeld, A., Bethe, H. (1933), "Elektronentheorie der Metalle," in: H. Geiger, K. Scheel (Eds.), *Handbuch der Physik,* Vol. 24, Berlin: Springer, pp. 333–622.

Swanson, L. W., Gesley, M. A., Davis, P. K. (1980), *Surf. Sci.* **107,** 263–289.

Tuggle, D. W., Swanson, L. W. (1985), *J. Vac. Sci. Technol. B* **3,** 220–223.

Tuggle, D. W., Li, J. Z., Swanson, L. W. (1985), *J. Microsc. (Oxford)* **140,** 293–301.

Walther, A. (1968), *J. Opt. Soc. Am.* **58,** 1256–1259.

Wehnelt, A. (1903), *Ber. Dtsch. Phys. Ges.* **5,** 29–34.

Wehnelt, A., Jentsch, F. (1909), *Ann. Phys. (Leipzig)* **28,** 537–552.

Further Reading

Hawkes, P. W., Kasper, E. (1989), *Principles of Electron Optics,* Vols. 1 and 2, London: Academic.

Lauer, R. (1982), "Characteristics of Triode Electron Guns," *Adv. Opt. Electron Microsc.* **8,** 137–206.

Mandel, L., Wolf, E. (1995), *Optical Coherence and Quantum Optics,* Cambridge, U.K.: Cambridge Univ. Press.

Nottingham, W. B. (1956), "Thermionic Emission," in: S. Flügge (Ed.), *Handbuch der Physik,* Vol. 21, Berlin: Springer, pp. 1–175.

Reimer, L. (1985), *Scanning Electron Microscopy,* Berlin: Springer.

Reimer, L. (1997), *Transmission Electron Microscopy,* 4th ed., Berlin: Springer.

Sommerfeld, A., Bethe, H. (1933), "Elektronentheorie der Metalle," in: H. Geiger, K. Scheel (Eds.), *Handbuch der Physik* Vol. 24, Berlin: Springer, pp. 333–622.

THERMOACOUSTIC ENGINES AND REFRIGERATORS

GREGORY W. SWIFT, *Condensed Matter and Thermal Physics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, U.S.A.*

Introduction 245
1. **Simple Examples** 245
2. **Important Length Scales** 247
3. **Dissipative Acoustic Processes** 247
4. **Lagrangian View of Thermoacoustic Processes** ... 250
4.1 Standing-Wave Refrigerator ... 251
4.2 Standing-Wave Engine 252
4.3 Traveling-Wave Systems 253
5. **Components** 253
5.1 Stacks and Regenerators 254
5.2 Resonators 255
5.3 Electroacoustic Power Transducers 256
5.4 Heat Exchangers 256
6. **Power and Efficiency** 256
7. **Methods of Computation** 258
7.1 Rott's Approach 258
7.2 Approximate Methods 260
7.3 Similitude 261
Glossary 263
Works Cited 263
Further Reading 263

INTRODUCTION

We usually think of a sound wave in a fluid (gas or liquid) as consisting of coupled pressure and displacement oscillations, but temperature oscillations also occur. Near a solid surface, the combination of these temperature oscillations with the pressure and displacement oscillations produces a rich variety of "thermoacoustic" effects.

Although thermoacoustic effects are too small to be readily noticed in everyday life, they can be harnessed to produce powerful refrigerators and heat engines, using intense acoustic standing waves in suitable geometries. Whereas typical engines and refrigerators have crankshaft-coupled pistons or rotating turbines, standing-wave thermoacoustic engines and refrigerators have no moving parts, or have flexing moving parts (as in a loudspeaker) with no sliding seals. Such thermoacoustic systems appear attractive because of their elegance, reliability, and low cost, in spite of only modest efficiency.

While standing-wave thermoacoustic systems have matured only recently, Stirling engines and refrigerators have a long, rich history. Recently, new insights have resulted from applying a thermoacoustics viewpoint to Stirling systems, treating them as traveling-wave thermoacoustic systems. Although Stirling systems are usually regarded as a separate subject, we will include some discussion of thermoacoustic aspects of Stirling systems here.

To be consistent with the article HEAT ENGINES AND REFRIGERATORS, we will adopt the convention that "heat engines" (sometimes called prime movers) are thermodynamic devices that convert heat to work and that "refrigerators" (including heat pumps) are devices that use work to transfer heat from low temperature to high temperature.

1. SIMPLE EXAMPLES

The Sondhauss tube, shown in Fig. 1(a), is the earliest reported thermoacoustic engine. Over 100 years ago, glassblowers noticed that when a small, hot glass bulb was being blown on a cool glass tubular stem, the stem tip sometimes radiated sound. Early investigators understood that the sound was generated by oscillatory thermal expansion and contraction of the air in the tube, which in turn was due to acoustic motion of the air toward and away from the hot

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

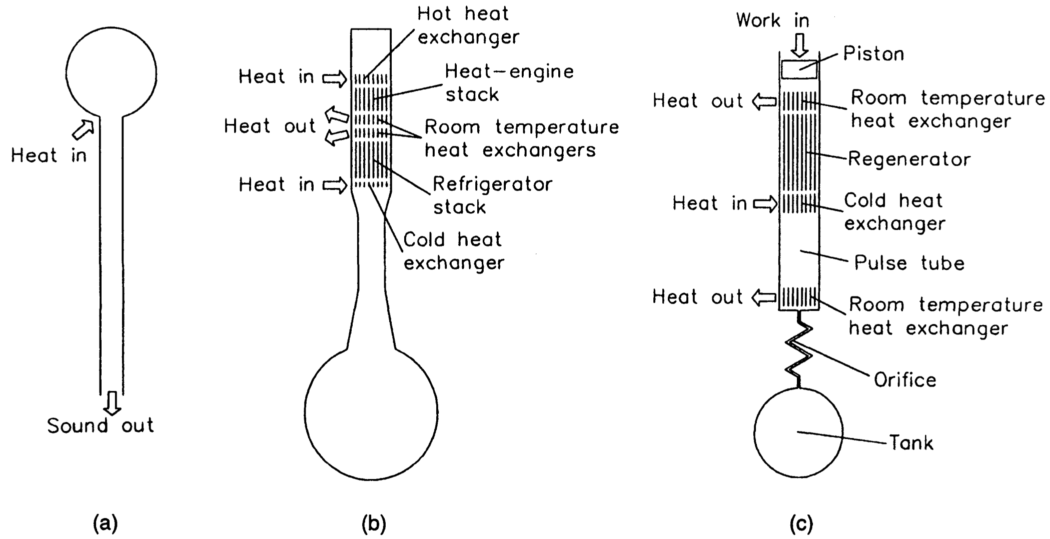
Heat in
Sound out
Hot heat exchanger
Heat in
Heat–engine stack
Heat out
Room temperature heat exchangers
Heat in
Refrigerator stack
Cold heat exchanger
Work in
Piston
Heat out
Room temperature heat exchanger
Regenerator
Heat in
Cold heat exchanger
Pulse tube
Heat out
Room temperature heat exchanger
Orifice
Tank
(a)
(b)
(c)

FIG. 1. Examples of thermoacoustic engines and refrigerators. (a) Sondhauss tube, the earliest thermoacoustic engine. (b) A heat-driven refrigerator, comprising a thermoacoustic engine and a thermoacoustic refrigerator sharing the same standing wave. (c) Orifice pulse-tube refrigerator, a traveling-wave system.

end of the tube. The Sondhauss tube was a heat engine, converting heat into mechanical work in the form of sound. It utilized a predominantly standing acoustic wave, with the bulb and stem forming a resonator.

In the 1960s it was realized by Carter *et al.* (1962) that the performance of such devices could be greatly enhanced by inclusion of proper heat-exchange components, especially what is now known as the "stack." The heat-driven thermoacoustic refrigerator shown in Fig. 1(b) illustrates these components, as well as the heat pumping/refrigerating function in thermoacoustics. The refrigerator's resonator consists of a cylindrical tube, closed at the top and opening at the bottom into a spherical bulb. Near the top is the heat-engine stack, which is an array of thin, uniformly spaced solid plates aligned parallel to the tube axis. Above the stack is the hot heat exchanger, consisting of well-spaced conducting strips; below the stack, a similar set of strips forms a room-temperature heat exchanger, cooled by circulating water. This is the heat-engine assembly, producing acoustic work from heat: When enough heat is supplied to the hot heat exchanger by an external heat source, the gas oscillates spontaneously, maintaining a high-amplitude acoustic standing wave in the gas.

Just below the heat engine in Fig. 1(b), another stack and pair of heat exchangers function as a refrigerator, which is driven by the acoustic work generated in the heat-engine stack. The refrigerator stack is identical to the engine stack, and its upper, room-temperature heat exchanger and lower, cold heat exchanger are made from copper strips. When the hot heat-exchanger temperature is high enough that the gas oscillates, the cold heat exchanger cools and absorbs an external heat load, as heat is pumped thermoacoustically up the refrigerator stack, from the cold heat exchanger to the room-temperature heat exchanger. Hence, the whole system functions as a refrigerator, with no moving parts, powered by heat delivered to the hot heat exchanger at high temperature.

Ceperley (1979, 1985) has discussed thermoacoustic engines and refrigerators using traveling waves. One example of such a refrigerator is the orifice pulse-tube refrigerator (Radebaugh, 1990) shown in Fig. 1(c). In this refrigerator, an oscillating piston functions as a source of acoustic work at one end of the refrigerator, and an orifice and tank act as a sink of acoustic work at the other end. Between them, a regenerator, heat exchangers, and an open tube are filled with gas. The regenerator and adjacent heat exchangers function just like those of a Stirling-cycle refrigerator and are the heart of this refrigerator. The oscillatory motion of the piston at the top and the orifice and tank at the bottom cause the gas to experience pressure oscillations and to move with the correct amplitude and phase relative to the pressure oscillations. The gas is thereby carried through four steps: compression, displacement downward, expansion, and displacement upward. The expansion cools the gas at the cold end of the pulse tube adiabatically, so that it can absorb heat in the cold heat exchanger when it is displaced upward. The so-called "pulse tube" itself functions as a temperature buffer space between the cold heat exchanger and the warm orifice; its length is larger than the gas displacement, so that the part of the gas that stays in the pulse tube throughout the cycle acts as a thermally insulating, compressible piston. With the hot heat exchangers anchored at room temperature, the cold heat exchanger of a pulse-tube refrigerator reaches typically 50 K, and refrigeration is produced with a coefficient of performance as high as $\frac{1}{4}$ of Carnot's.

2. IMPORTANT LENGTH SCALES

In most thermoacoustic engines and refrigerators, the acoustic oscillation is essentially a pure tone, so that velocity, pressure, and temperature essentially vary sinusoidally with time, at the oscillation frequency f. The wavelength $\lambda = a/f$, where a is the sound speed, is an important length scale, especially in the direction of the gas displacement oscillations. For example, in the heat-driven refrigerator of Fig. 1(b) above, the distance from the hot end cap at the top of the tube to the tube–bulb junction is approximately $\lambda/4$, so that the tube is essentially a quarter-wave acoustic resonator, with pressure maximum and velocity node at the top, and pressure node and velocity maximum at the entrance to the bulb. In thermoacoustic systems designed to date, 10 cm $\lesssim \lambda \lesssim$ 10 m. The lengths of stacks are much less than λ.

The amplitude x_1 of the gas displacement oscillations is a second important length scale. Typically, x_1 is everywhere much smaller than both the wavelength and the overall length of the system. Within the stack x_1 is typically smaller than the length of the stack, and in the heat exchangers x_1 is typically comparable to the length of the exchangers.

In the directions perpendicular to the displacement oscillations, the key length scale is the thermal penetration depth $\delta_\kappa = \sqrt{\kappa/\pi f \rho c_p}$, where κ is the thermal conductivity of the gas, ρ is its density, and c_p is its isobaric heat capacity per unit mass. The thermal penetration depth is approximately the distance that heat can diffuse through the gas in a time $1/\pi f$. Gas much farther than this from the nearest solid surface experiences adiabatic oscillations and will not participate in thermoacoustic effects. Thus, the spacing in stacks and heat exchangers cannot be much greater than δ_κ. In typical thermoacoustic systems, δ_κ is a fraction of a millimeter. In standing-wave thermoacoustic devices, a key geometrical requirement is that the smallest transverse dimension in the stack be roughly 1 to 3 times δ_κ; in traveling-wave systems, the smallest transverse dimension must be much smaller than δ_κ.

Closely related to the thermal penetration depth is the viscous penetration depth $\delta_\nu = \sqrt{\mu/\pi f \rho} = \sqrt{\nu/\pi f}$, which is approximately the distance over which momentum can diffuse in a time $1/\pi f$. (Here μ is viscosity and ν is kinematic viscosity.) Thus, within δ_ν of solid surfaces, viscous shear causes gradients in the oscillatory velocity and displacement. The ratio $\delta_\nu/\delta_\kappa = \sqrt{\mu c_p/\kappa}$ is the square root of the Prandtl number, a dimensionless measure of the ratio of viscous to thermal effects, and is of order 1 for ideal gases. Hence, in general, viscous effects are important whenever thermoacoustic effects are important. In some of the heuristic discussion below, we will neglect viscosity; but it must be included when quantitatively accurate results are desired.

3. DISSIPATIVE ACOUSTIC PROCESSES

To introduce the concepts and notation of thermoacoustics, we begin with a review of ordinary acoustic dissipation at solid boundaries, which arises from viscous and thermal contact. We consider sound propagating in the x direction in an ideal gas within a duct with constant cross section. We adopt the usual complex notation for time-oscillatory quantities (pressure p; temperature T; velocity components u, v, and w; density ρ; and entropy per unit mass s):

$$p = p_m + \mathrm{Re}[p_1(x)e^{i\omega t}] + \ldots, \tag{1}$$

$$u = \mathrm{Re}[u_1(x,y,z)e^{i\omega t}] + \ldots, \tag{2}$$

$$T = T_m + \mathrm{Re}[T_1(x,y,z)e^{i\omega t}] + \ldots, \tag{3}$$

$$\rho,\ s,\ \text{etc.} = \text{similar to } T, \tag{4}$$

$$v,\ w = \text{similar to } u, \tag{5}$$

with μ, κ, etc. constant. In the acoustic approximation, all the time dependence appears in the factor $e^{i\omega t}$, with $\omega = 2\pi f$. The mean values (subscript m) are real, but the small amplitudes (subscript 1) are, in general, complex to account for the time phasing of the oscillating quantities. We will assume that this expansion to lowest order in the acoustic amplitude suffices for all variables.

To develop quantitative understanding of viscous effects, we begin by finding the y and z dependence of the gas velocity, using the x component of the momentum (Navier–Stokes) equation, for which the acoustic approximation is

$$i\omega\rho_m u_1 = -\frac{dp_1}{dx} + \mu\left[\frac{\partial^2 u_1}{\partial y^2} + \frac{\partial^2 u_1}{\partial z^2}\right]. \tag{6}$$

The x derivatives of u_1 have been neglected because they are of order $1/\lambda$ and hence are much smaller than the y or z derivatives, of order $1/\delta_\nu$. Equation (6) is a partial differential equation for $u_1(y,z)$. With boundary condition $u_1 = 0$ at the solid surface, its solution is

$$u_1 = \frac{i}{\omega\rho_m}[1 - h_\nu(y,z)]\frac{dp_1}{dx}, \tag{7}$$

where $h_\nu(y,z)$ depends on the specific geometry under consideration. Below, we will need the spatial average (over y and z) of Eq. (7):

$$\langle u_1\rangle = \frac{i}{\omega\rho_m}(1 - f_\nu)\frac{dp_1}{dx}, \tag{8}$$

where f_ν is the spatial average of h_ν. Table 1 displays the functional forms (Arnott *et al.*, 1991; Swift and Keolian, 1993) of h_ν and f_ν, and Fig. 2 shows f_ν, with the hydraulic radius r_h characterizing the size of the duct. In the pin array, gas flows in the space between cylinders of radius r_i aligned with the x axis and arranged in a triangular lattice with center-to-center separation $\sqrt{2\pi}r_o/\sqrt[4]{3}$.

These functions describe how viscosity reduces the magnitude of the oscillatory velocity and shifts its phase. Gas that is much closer than δ_ν to the nearest solid surface is nearly at rest. Gas that is much farther than δ_ν from the nearest solid surface experiences essentially no viscous shear; it moves with a

Table 1. Complex function h and its spatial average f for several geometries. In the second column, the origin is at the center of symmetry of the geometry, except in the boundary-layer limit where it is at the surface. In cases with a sort of circular symmetry, $r = (y^2 + z^2)^{1/2}$. The hydraulic radius r_h is the ratio of gas volume to gas–solid contact area.

Geometry	h	f
Parallel plates, spacing $2y_0$; $r_h = y_0$	$\dfrac{\cosh[(1+i)y/\delta]}{\cosh[(1+i)y_0/\delta]}$	$\dfrac{\tanh[(1+i)y_0/\delta]}{(1+i)y_0/\delta}$
Circular pores, radius R; $r_h = R/2$	$\dfrac{J_0[(i-1)r/\delta]}{J_0[(i-1)R/\delta]}$	$\dfrac{2J_1[(i-1)R/\delta]}{J_0[(i-1)R/\delta](i-1)R/\delta}$
Pin array (triangular lattice), pin radius r_i, centers' separation $(2\pi)^{1/2}r_o/3^{1/4}$ $\simeq 1.905r_o$; $r_h = (r_o^2 - r_i^2)/2r_i$	$\simeq \dfrac{Y_1(\xi_o)J_0(\xi) - J_1(\xi_o)Y_0(\xi)}{Y_1(\xi_o)J_0(\xi_i) - J_1(\xi_o)Y_0(\xi_i)}$, where $\xi = (i-1)r/\delta$	$\simeq -\dfrac{2\xi_i}{\xi_o^2 - \xi_i^2}\dfrac{Y_1(\xi_o)J_1(\xi_i) - J_1(\xi_o)Y_1(\xi_i)}{Y_1(\xi_o)J_0(\xi_i) - J_1(\xi_o)Y_0(\xi_i)}$
Rectangular pores, $2y_0 \times 2z_0$; $r_h = y_0z_0/(y_0 + z_0)$	$1 - \dfrac{16}{\pi^2}\sum_{m,n\ \mathrm{odd}}\dfrac{\sin(m\pi y/2y_0)\sin(n\pi z/2z_0)}{mnY_{mn}}$, where $Y_{mn} = 1 - i\dfrac{\pi^2\delta^2}{8y_0^2z_0^2}(m^2z_0^2 + n^2y_0^2)$	$1 - \dfrac{64}{\pi^4}\sum_{m,n\ \mathrm{odd}}\dfrac{1}{m^2n^2Y_{mn}}$
Boundary-layer limit	$e^{-(1+i)y/\delta}$	$\dfrac{(1-i)\delta}{2r_h}$

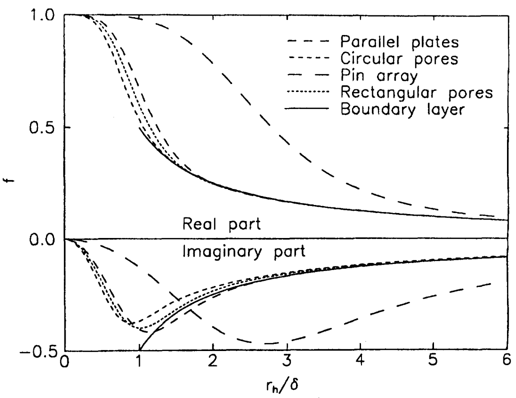

FIG. 2. Spatial-average function f for several geometries. (See text and Table 1 for a more complete description.) The rectangle here has 6:1 aspect ratio, and the pin array has $r_o/r_i = 6$. The hydraulic radius r_h is the ratio of gas volume to gas–solid contact area. The boundary-layer limit is approached at large r_h for all geometries.

velocity that is independent of y and z. Gas that is roughly δ_ν from the nearest solid surface moves with a reduced velocity and significant phase shift. The acoustic power dissipation per unit volume, which is proportional to the square of ∇u_1, is in general greatest at the solid surface, where the velocity gradients are highest.

Similarly, to find the dependence of the oscillatory temperature on y and z, we begin with the general equation of heat transfer, for which the acoustic approximation is

$$i\omega\rho_m c_p T_1 - i\omega p_1 = \kappa\left[\frac{\partial^2 T_1}{\partial y^2} + \frac{\partial^2 T_1}{\partial z^2}\right]. \tag{9}$$

In general, the solid has sufficient heat capacity and thermal conductivity to enforce $T_1 = 0$ on the gas at the solid surface. Equation (9) is a partial differential equation for $T_1(y,z)$, similar in form and boundary condition to Eq. (6) for $u_1(y,z)$. Exploiting the similarity, the solution can be written

$$T_1 = (1/\rho_m c_p)[1 - h_\kappa(y,z)]p_1, \tag{10}$$

and its spatial average can be written

$$\langle T_1\rangle = (1/\rho_m c_p)(1 - f_\kappa)p_1, \tag{11}$$

where f_κ and h_κ are the same as f_ν and h_ν, but with δ_ν replaced by the thermal penetration depth δ_κ.

Similar to the effect of viscosity discussed above, thermal contact with the solid surface reduces the magnitude and shifts the phase of the oscillatory temperature. Gas that is much closer than δ_κ to the nearest solid surface is nearly isothermal. Gas that is much farther than δ_κ from the nearest solid surface is essentially adiabatic, experiencing adiabatic temperature oscillations $T_1 = (1/\rho_m c_p)p_1$ in phase with the pressure oscillations. At approximately δ_κ from the nearest solid surface, these oscillations are reduced in magnitude and shifted in phase.

To find the time-averaged acoustic power $d\dot{W}$ *dissipated in length* dx of the duct, we write

$$\frac{d\dot{W}}{dx} = -\int \frac{d(\overline{pu})}{dx}\,dy\,dz, \tag{12}$$

where the overbar denotes time averaging and the integral is over the cross-sectional area A of the duct. Rewriting Eq. (12) in complex notation and expanding the derivative gives

$$\frac{d\dot{W}_2}{dx} = -\tfrac{1}{2}A\,\mathrm{Re}\left(\langle\tilde{u}_1\rangle\frac{dp_1}{dx} + \tilde{p}_1\frac{d\langle u_1\rangle}{dx}\right), \tag{13}$$

where the tilde denotes complex conjugation. We can obtain dp_1/dx from Eq. (8) above. To find $d\langle u_1\rangle/dx$, we will use our expression for $\langle T_1\rangle$ in the continuity equation $\partial\rho/\partial t + \nabla\cdot(\rho\mathbf{v}) = 0$, which can be averaged with respect to y and z in the acoustic approximation to obtain

$$i\omega\langle\rho_1\rangle + \rho_m\frac{d\langle u_1\rangle}{dx} = 0. \tag{14}$$

Using $d\rho = -(\rho/T)dT + (\gamma/a^2)dp$ (where γ is the usual ratio of isobaric to isochoric specific heats), we can express the spatially averaged density as $\langle\rho_1\rangle = -(\rho_m/T_m)\langle T_1\rangle + (\gamma/a^2)p_1$. Substituting this into Eq. (14), using Eq. (11) for $\langle T_1\rangle$, and eliminating c_p by means of the thermodynamic identity $\gamma - 1 = a^2/Tc_p$ yields

$$i\omega[1 + (\gamma - 1)f_\kappa]p_1 + \rho_m a^2\frac{d\langle u_1\rangle}{dx} = 0 \tag{15}$$

as an acoustic expression of the continuity equation that can be solved for $d\langle u_1\rangle/dx$.

Finally, substituting Eqs. (15) and (8) into Eq. (13), we obtain

$$\frac{d\dot{W}_2}{dx} = \frac{A\omega}{2}\left[\frac{\rho_m|\langle u_1\rangle|^2}{|1-f_\nu|^2}\,\mathrm{Im}(-f_\nu) + \frac{(\gamma-1)|p_1|^2}{\rho_m a^2}\,\mathrm{Im}(-f_\kappa)\right]. \tag{16}$$

The first term gives the viscous damping of sound, and the second term gives the less intuitively obvious thermal damping. The work thus lost from the sound wave appears as heat, at the temperature T_m. In the boundary-layer approximation, in which all dimensions of the duct are much larger than the penetration depths, Eq. (16) reduces to

$$\frac{d\dot{W}_2}{dS} = \tfrac{1}{4}\,\rho_m|\langle u_1\rangle|^2\delta_\nu\omega + \tfrac{1}{4}\frac{|p_1|^2}{\rho_m a^2}(\gamma-1)\delta_\kappa\omega, \tag{17}$$

where S is surface area of the duct. This expression is readily interpreted in terms of the acoustic energy density:

- The mean kinetic energy per unit volume $\rho|\langle u_1\rangle|^2/4$ times δ_ν is roughly the kinetic energy in the gas within a viscous penetration depth of the solid surface, per unit area of the surface; the viscous dissipation term shows that this energy is dissipated by viscous shear at an average rate ω. Intuitively, we imagine layers of gas sliding relative to one another, with viscous "friction" between the layers turning mechanical energy into heat.
- Similarly, the thermal dissipation term shows that the mean adiabatic compressive energy density $|p_1|^2/4\rho_m a^2$ stored in unit volume of the gas is dissipated at an average rate $(\gamma - 1)\omega$ within a region of thickness δ_κ near the solid surface. The extra factor $(\gamma - 1)$ appears because this thermal dissipation is proportional to the difference between the isothermal and adiabatic compressibilities, which is proportional to $\gamma - 1$. Here, the intuitive picture is more subtle than for viscous dissipation. Gas immediately adjacent to the surface experiences isothermal density and pressure oscillations, which are reversible; gas far from the surface experiences adiabatic density and pressure oscillations, which are also reversible. We imagine the gas approximately δ_κ from the surface experiencing a complex cycle of density changes in response to the pressure oscillations: first an increase in density due to quasi-adiabatic compression by the sound wave, then a further increase in density as thermal contact with the surface removes heat from the gas, then a decrease in density due to quasi-adiabatic expansion by the sound wave, and finally a further decrease in density as thermal contact with the surface delivers heat to the gas. Since this gas experiences thermal expansion at low pressure and thermal contraction at high pressure, it absorbs work from the sound wave. This gas, approximately δ_κ *from the surface,* is the most effective at absorbing work from the sound wave, whereas in the case of viscous dissipation the gas *at the surface* is most effective.

4. LAGRANGIAN VIEW OF THERMOACOUSTIC PROCESSES

In the derivation of acoustic dissipation above, we adopted the usual Eulerian point of view of fluid mechanics, focusing attention on what happens at *a given point in space* as the gas moves by. Qualitative insight into thermoacoustic processes will be gained in this section by considering the alternative, Lagrangian point of view, focusing attention on the position, pressure, volume, and temperature of *a given parcel of gas* as it moves. It is simplest here to imagine square-wave motion and pressure in the gas, with pressure and displacement in phase for standing waves and pressure and velocity in phase for traveling waves. However, real thermoacoustic systems operate sinusoidally, with a variety of phases between motion and pressure. Viscosity is ignored in this section, even though it has a significant, harmful effect on thermoacoustic efficiency. With these conceptual simplifications, the thermodynamic cycle for standing waves becomes identical to the Brayton cycle, with two reversible adiabats and two irreversible constant-pressure heat transfers, while the cycle for traveling waves becomes two isotherms and two reversible polytropic processes.

Time phasing is an important factor in

the operation of traditional heat engines and refrigerators: Pistons, valves, and sparks must have the correct relative timing in order to produce the desired thermodynamic cycle in the working gas. Thermoacoustic systems often contain no obvious moving parts to perform these functions, yet the acoustic transport of heat and the generation (or absorption) of acoustic work point to some type of timed phasing of thermodynamic processes, here achieved in a remarkably simple way.

One source of phasing in thermoacoustic systems is the presence of two thermodynamic media: gas and solid. As the gas oscillates along the solid at the acoustic frequency, it experiences oscillations in temperature. Part of the temperature oscillations comes from compression and expansion of the gas by sound pressure, and the rest is a consequence of the local temperature of the solid itself. In Sec. 3 above, the local solid temperature was constant. In this section, the temperature gradient along the stack ensures that each moving parcel of gas "sees" a *time-dependent* local solid temperature because of its motion along the temperature gradient.

In keeping with the square-wave assumption of this section, we need not retain complex notation to account for the phase of time-dependent variables. Hence, in this section only, all variables with subscript 1 are real and positive.

4.1 Standing-Wave Refrigerator

The standing-wave thermoacoustic refrigerator can be understood with reference to Fig. 3(a), where the four steps of its cycle are displayed for one typical parcel of gas. In response to the standing wave, the gas parcel is pressurized and moves up along the adjacent solid in step 1, and it is depressurized and moves down along the adjacent solid in step 3, absorbing and doing work dW''. In this picture, the pressure changes in steps 1 and 3 are adiabatic, so that the gas parcel experiences adiabatic temperature changes proportional to the pressure changes. Steps 2 and 4 allow thermal equilibration between the gas parcel and the adjacent solid. In these two steps, net work $dW - dW'$ is absorbed from the sound wave by the gas parcel, as it thermally contracts at high pressure in step 2 and thermally expands at low pressure in step 4. If the sign of dT_m/dx is as shown (with higher T_m toward the nearest pressure antinode), the net thermal effect of the parcel of gas is to move heat $dQ \approx dQ'$ a short distance along the solid, from slightly lower temperature to slightly higher temperature, during each cycle of the sound wave. Hence the system functions as a refrigerator, with acoustic work being absorbed to pump heat from a lower temperature to a higher temperature.

If $dT_m/dx = 0$, the situation is identical to that described quantitatively in Sec. 3 (but with $\mu = 0$ here). In this case, $dW - dW'$ represents the thermal damping of sound, given by the second term in Eq. (16) or (17).

There is a so-called critical temperature gradient $\nabla T_{\text{crit}} \equiv (dT_m/dx)_{\text{crit}} = p_1/\rho_m c_p x_1$ at which no heat is transferred in steps 2 and 4. This occurs when the temperature rise due to adiabatic compression in step 1, $2p_1/\rho_m c_p$, exactly matches the rise in adjacent solid temperature that the gas sees because of its motion along the solid, $2x_1 \nabla T_{\text{crit}}$. For all temperature gradients less than ∇T_{crit}, heat transfer between gas and solid has the signs shown in Fig. 3(a), with heat flow dQ from gas to solid in step 2, at the higher temperature and position, and heat flow dQ' from solid to gas in step 4, at the lower temperature and position. For such low temperature gradients it is also true that $dW > dW'$, so that the gas absorbs work from the standing wave. Hence, a standing-wave system acts as a refrigerator only if $dT_m/dx < \nabla T_{\text{crit}}$.

In Fig. 3(a), we label the parcel's distance from the solid as $\sim\delta_\kappa$ because imperfect thermal contact between gas and solid is required for the cycle to operate as described. If thermal contact between gas and solid were perfect, the gas would arrive at its uppermost position at the end of step 1 already equilibrated at the local solid temperature, so that no heat would be transferred during step 2; and similarly for steps 3 and 4. At the other extreme, if thermal contact were nonexistent, no heat could be transferred in steps 2 and 4. Thus, intermediate, *imperfect* thermal contact between gas and solid is required for the cycle to operate as described. For the sinusoidal motion and pressure of real systems, this imperfect thermal contact is achieved for gas neither much closer than δ_κ to the nearest solid nor much farther than

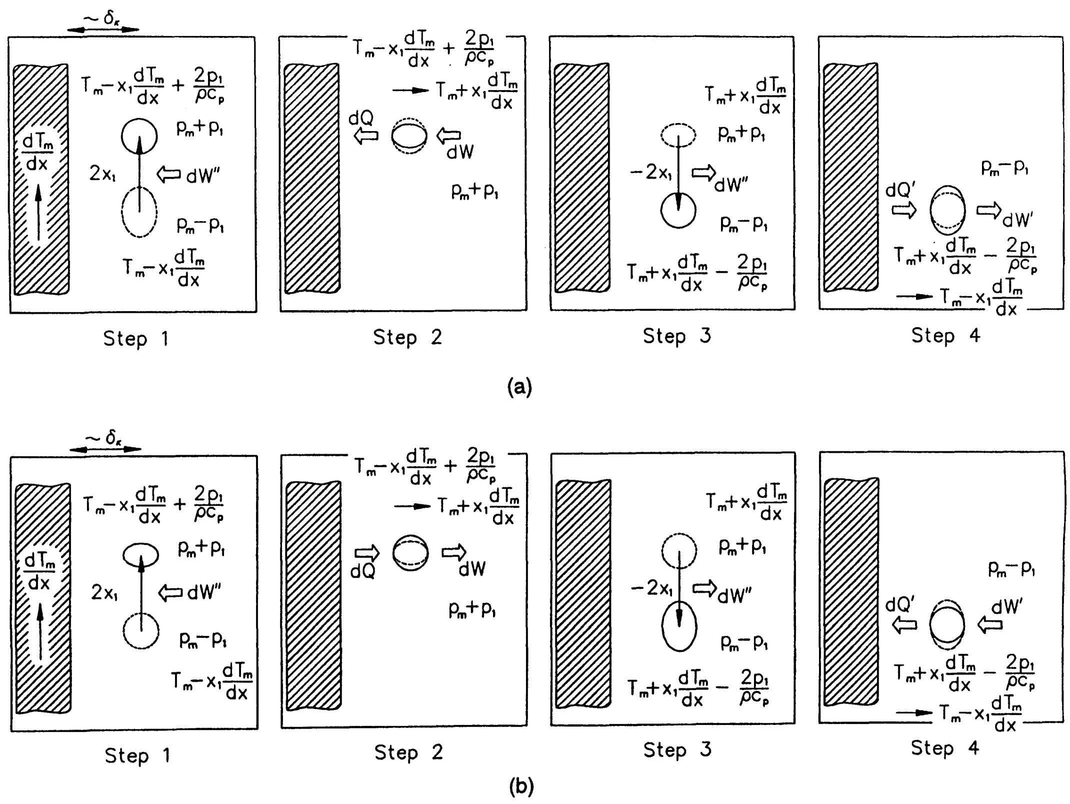

FIG. 3. (a) A typical gas parcel executing the four steps of the cycle of a standing-wave thermoacoustic refrigerator, assuming an inviscid gas and square-wave acoustic motion and pressure. In each step, the dashed and solid ellipses show, respectively, the initial and final positions and volumes of the parcel, while labels show the initial and final temperatures and pressures of that step; and broad arrows show directions of attendant heat flows dQ or dQ' (between parcel and solid) and work flows dW, dW', or dW'' (into or out of the parcel). The solid temperature at the average location of the gas parcel is T_m. Averaged over one cycle, the parcel absorbs work $dW - dW'$ from the sound wave. In Figs. 3 and 4, we have dropped our usual complex notation, so that all variables here are positive real numbers. (b) A typical gas parcel executing the four steps of the cycle of a standing-wave thermoacoustic engine, assuming an inviscid gas and square-wave acoustic motion and pressure. Averaged over one cycle, the parcel delivers work $dW - dW'$ to the sound wave.

δ_κ from the nearest solid. As suggested by the thermal dissipation term in Eq. (16), the quantitative measure of the correct geometry is $\mathrm{Im}(-f_\kappa)$. Thus, effective thermoacoustic systems contain stacks that fill the entire cross section of the resonator, with spacings of a few thermal penetration depths, as shown in Fig. 1(b); and with $\mathrm{Im}(-f_\kappa)$ as large as possible, suggesting (cf. Fig. 2) that parallel plates and pin arrays are the best stack geometries.

4.2 Standing-Wave Engine

To understand the conversion of heat to acoustic work in a standing-wave thermoacoustic engine, consider the magnified view of part of the stack shown in Fig. 3(b), which shows a typical parcel of gas at four instants of time during one cycle of the acoustic wave. The standing wave carries the parcel up and down, compressing and expanding it, with phasing such that it is at its most compressed state when at its upper position, and its most expanded state at its lower position. Here, the solid temperature gradient is greater than ∇T_{crit}, and so the directions of heat transfer in steps 2 and 4 are reversed relative to the standing-wave refrigerator. Thus, the net thermal effect of the parcel of gas shown is to move a little heat down the solid, from higher to lower temperature; again, all parcels have a similar effect, so that overall a substantial amount of heat is carried down along the solid from the hot heat exchanger to the cold heat exchanger.

Since heat flows into the gas in step 2, and out of it in step 4, the gas expands at high pressure in step 2 and contracts at low pressure in step 4. Thus, the net effect in the thermoacoustic heat engine is the production of work, at the acoustic frequency, by this and all other gas parcels. One can imagine all the gas in the stack region of a thermoacoustic engine thermally expanding and contracting because of its displacement along the temperature gradient, with the correct phasing relative to the oscillatory standing-wave pressure to do net work. In the Sondhauss tube, for example, this net work maintains the standing wave against viscous and thermal dissipation in the resonator and provides power to radiate into the room.

4.3 Traveling-Wave Systems

Ceperley (1979, 1985) introduced the idea that Stirling engines (*q.v.*) and refrigerators, and related devices such as pulse-tube refrigerators, may be regarded as traveling-wave thermoacoustic systems. This point of view is possible because the phasing between displacement and pressure in such engines and refrigerators is substantially that of traveling waves, with displacement and pressure substantially out of phase. A second difference between standing- and traveling-wave systems is the typical spacing between solid elements: In traveling-wave systems, this distance is typically $\ll\delta_\kappa$, ensuring excellent thermal contact between gas and solid at all times.

These features are illustrated for the traveling-wave refrigerator in Fig. 4. Isothermal compression in step 2 transfers heat from the gas parcel to the solid at a higher temperature, and isothermal expansion at step 4 transfers heat from solid to gas at a lower temperature. The other two steps serve to preheat and precool the gas, reversibly storing and recovering heat in the solid. The net effect of one parcel over an entire cycle is, again, to move a little heat a short distance up the solid, up the temperature gradient. Excellent thermal contact between gas and solid is required in steps 1 and 3 for effective preheating and precooling, respectively; in realistic, sinusoidal traveling-wave systems, this is achieved by maintaining transverse dimensions much smaller than a thermal penetration depth.

Traveling-wave engines are also possible in principle; from the perspective of thermoacoustics, all Stirling engines can be considered as such in the low-amplitude limit.

5. COMPONENTS

The first efficient thermoacoustic refrigerator, designed, built, and studied by Tom Hofler (1986, 1988), is shown in Fig. 5(a). It illustrates several features of many of today's standing-wave thermoacoustic devices. Figure 5(b) shows some of the data obtained

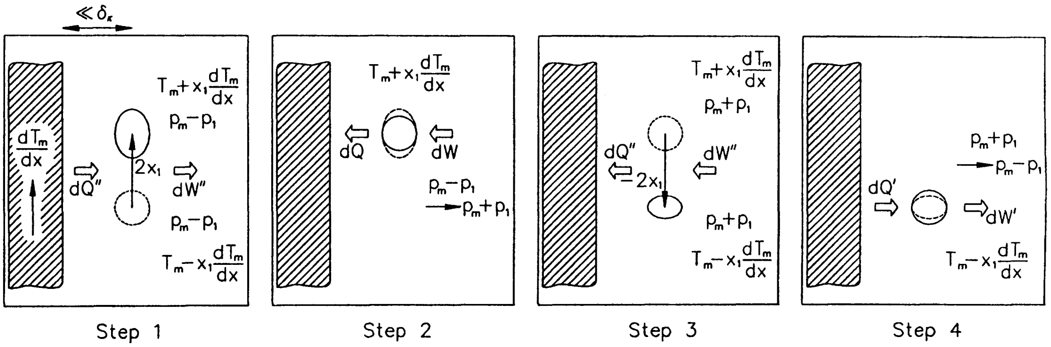

FIG. 4. A typical gas parcel executing the four steps of the cycle of a traveling-wave thermoacoustic refrigerator, assuming an inviscid gas and square-wave acoustic motion and pressure. Symbols have the same meaning as in Fig. 3.

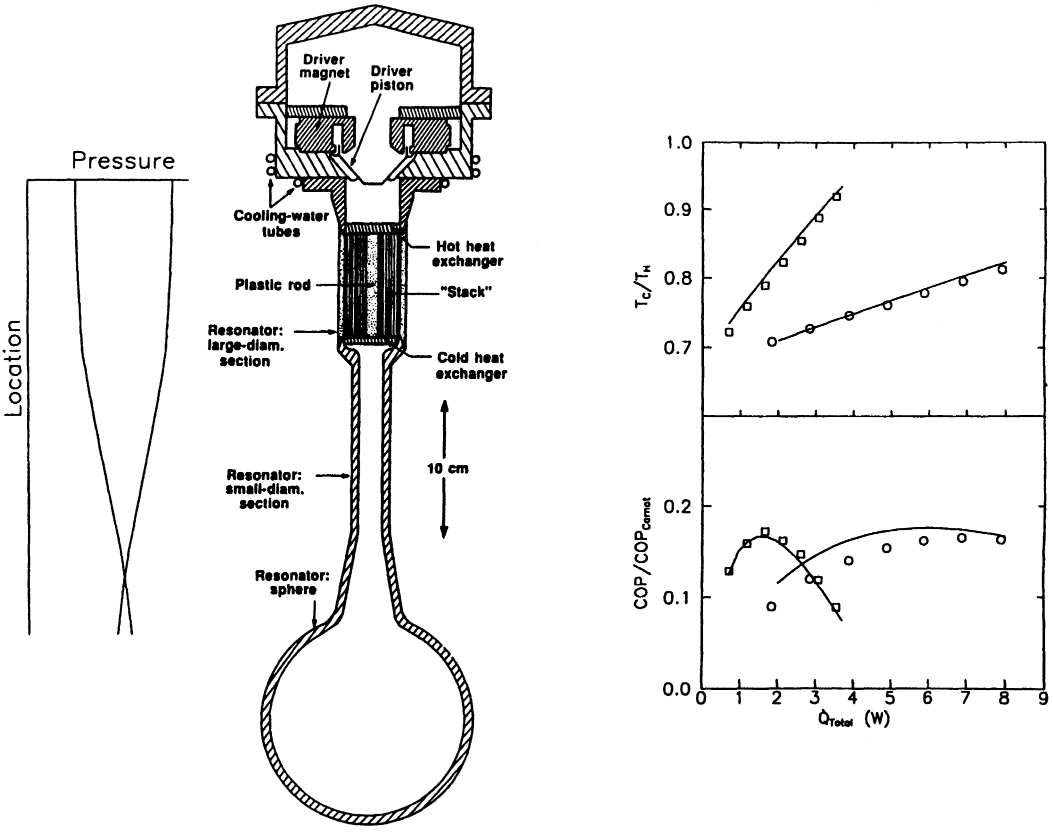

(a) (b)

FIG. 5. (a) An early thermoacoustic refrigerator and (b) some of its measured and calculated performance, with 500-Hz pressure oscillations in 10-bar helium gas and with hot heat-exchanger temperature $T_H = 300$ K. Data from Hofler (1988): squares, $|p_1|/p_m = 0.015$; circles, $|p_1|/p_m = 0.03$. Lines calculated using Ward and Swift (1994). The cold temperature T_C was measured in the cold heat exchanger. The gross cooling power $\dot{Q}_{Total}$ includes the deliberately applied load plus some small parasitic loads such as heat leak from room temperature (see Hofler, 1988, for details). COP $= \dot{Q}_{Total}/\dot{W}$, with $\dot{W}$ the acoustic power delivered to the resonator.

with this refrigerator, which had a cooling power of several watts with acoustic pressure amplitudes of 3% of the mean pressure.

5.1 Stacks and Regenerators

The heart of a thermoacoustic system is its stack (or, in the case of a traveling-wave system, its regenerator). For standing-wave systems, Fig. 2 indicates that parallel plates and pin arrays are the best stacks known, because they have the largest values of $\mathrm{Im}(-f_\kappa)$. As yet there have been no experimental studies using pin arrays, and so the parallel plates are a preferred geometry. Figure 2 suggests that the spacing of parallel plates should be about $2.2\delta_\kappa$, although viscous and system-wide effects often modify the optimum significantly.

For ease of construction and structural rigidity, parallel-plate stacks have usually been spirally wound. The stack in Fig. 5 was a plastic spiral wound around a small plastic rod, with numerous tiny, parallel plastic rods serving as spacers between layers of the spiral. Hence the stack channels were not ideal parallel plate; they were essentially rectangular channels with 13:1 aspect ratio, and with slight curvature. Spiral stacks have also been built with spacing determined by internal bumps throughout the stack or by radial ribs at the ends of the stack.

Typically stack materials are thin enough to block only 10–20% of the total cross-sectional area of the stack. Generally any stack material rigid enough to support its own weight and survive routine handling has sufficient heat capacity to enforce the $T_1 = 0$ boundary condition at the surface. [Liquid-metal thermoacoustics (Migliori and Swift, 1988) is a rare exception.] Hence the key requirement on the stack material is a low thermal conductivity, lest heat be wastefully conducted along the stack from hot to cold. Near and below room temperature, plastics have generally been used; above room temperature, stainless steel, nickel alloys, and ceramics have been used.

Other stack geometries that have been tested include metal honeycomb with hexagonal cells and ceramic honeycomb with square cells (from automotive catalytic converters).

In traveling-wave systems, for which the "stack" is known as a regenerator, the requirement that the spacing be $\ll \delta_\kappa$ makes it difficult to construct parallel plates with uniform spacing. Hence, enhanced heat-transfer surfaces such as screen beds or intricately etched foils (Yaron *et al.*, 1996) are generally used. These geometries are not discussed here, but they can be treated by thermoacoustic methods (Ward and Swift, 1994).

5.2 Resonators

A resonator is required for standing-wave thermoacoustic systems. It should satisfy many requirements, which we illustrate with Fig. 5. Most fundamentally, it should have the desired resonance frequency, and permit location of the stack at a location with the desired ratio of $|p_1|/|\langle u_1\rangle|$. The resonator of Fig. 5 is essentially a quarter-wavelength resonator [even though $p_1(x)$, shown at left, is not exactly a cosine], with the pressure antinode at the driver piston and the pressure node near the sphere. The stack is located close to the pressure antinode, where $|p_1|$ is nearly equal to its value at the pressure antinode and $|\langle u_1\rangle|$ is significantly smaller than it is at the pressure node. (In Sec. 6 we discuss this choice of location in terms of efficiency.)

A conceptually simpler quarter-wavelength resonator would comprise a uniform-diameter tube $\lambda/4$ long connected to an essentially infinite sphere. The resonator of Fig. 5 differs from this in important, advantageous ways. First, the tube consists of two sections of different diameters. The small-diameter section near the pressure node allows the tube to be significantly shorter than $\lambda/4$, reducing the overall size and weight. More subtly, this configuration also reduces the dissipation of acoustic power in the resonator, which is governed by Eq. (17). Basically, as the diameter of the lower section of the tube is reduced, its surface area decreases while $|\langle u_1\rangle|$ increases and $|p_1|$ remains about the same. For a modest reduction in diameter, as in Fig. 5, the net effect is a reduction in the total dissipation, by as much as a factor of 2.

The resonator of Fig. 5 also suppresses harmonic content in the standing wave. In this regard, a uniform-diameter resonator is extremely undesirable, because its higher resonant modes occur at frequencies that are exact integer multiples of the fundamental. In this situation, nonlinear effects drive resonances, and hence the harmonic content of the wave is high. These nonlinear effects, such as the $(\mathbf{v}\cdot\nabla)\mathbf{v}$ term in the momentum equation (neglected elsewhere in this article), generate oscillations at integer multiples of the operating frequency. The resonance frequencies of the first few modes of the resonator of Fig. 5 are not integer multiples, and so higher-mode resonances are avoided.

Choice of sphere diameter in Fig. 5 was a compromise between size/weight and dissipation: A smaller sphere experiences larger pressure oscillations and requires a longer tube to maintain the desired resonance frequency. The tapered connection between the sphere and the small-diameter section reduces the Reynolds number at the sphere opening, in order to minimize dissipation due to turbulence.

Another typical standing-wave thermoacoustic geometry is essentially half-wavelength (Garrett *et al.*, 1994), with mirror-image stacks near the pressure antinodes at each end of the resonator. The central portion of such a resonator generally has a reduced diameter, to exploit the reduced size, dissipation, and harmonic content described above. Half-wavelength resonators are sometimes bent or coiled to reduce overall size.

With the requirement that $\delta_\kappa \gg$ spacing, traveling-wave systems typically operate at lower frequencies than standing-wave systems, pushing conventional acoustic resonators to impractical lengths. Hence, traveling-wave systems often incorporate a mechanical element into the resonator, providing resonance with solid mass bouncing on a gas spring. Such "free-piston" Stirling machines are fully described elsewhere (Urieli and Berchowitz, 1984; STIRLING ENGINES).

5.3 Electroacoustic Power Transducers

To date, thermoacoustic systems have often used loudspeaker-like transducers, with permanent magnets and coils of wire in relative oscillatory motion. Piezoelectric and magnetohydrodynamic transducers are also candidates.

In Fig. 5, the loudspeaker-like driver was located at a pressure antinode of the standing wave, so that the acoustic power was delivered with high force and small displacement, easing engineering difficulties associated with the flexing portion of the driver. This location also placed it at room temperature, where Joule heat generated in the driver could be removed most efficiently.

Commercial loudspeakers are often poorly matched to thermoacoustic applications, because they are designed to drive a comparatively low acoustic impedance with large bandwidth. For better impedance matching, the loudspeaker in Fig. 5 was modified, with the addition of an aluminum cone and associated flexing metal seal to exert the speaker's force over a small area. With this modification, it converted electric power to acoustic power with approximately 20% efficiency. Further modifications to commercial loudspeakers can achieve efficiencies of over 50%. In the field of Stirling machines, linear motors and linear alternators of 90% efficiency have been demonstrated (Corey and Yarr, 1992), indicative of the performance to be expected when loudspeaker-like transducers are designed specifically for efficient conversion of power at a single frequency.

5.4 Heat Exchangers

Heat exchangers are required in thermoacoustic engines and refrigerators to supply and extract heat at the two ends of the stack or regenerator. The length of each along the direction of the acoustic oscillation is of the order of the local displacement amplitude x_1; transverse dimensions are roughly of the order of those of the adjacent stack or regenerator. Equation (16) can be used to estimate the acoustic power dissipation in the heat exchangers, although entrance effects prevent quantitatively accurate predictions.

In the system shown in Fig. 5(a) and other small thermoacoustic systems, the heat exchangers can simply consist of multiple copper strips spanning the resonator at the stack ends. Larger-diameter, high-power systems require integral pumped-fluid channels, heat pipes, or electric resistance heaters imbedded in the heat exchangers to deliver or extract heat uniformly from the stack ends. In this case, the typical geometry comprises a set of high-conductivity fins extending from pipes (Garrett *et al.*, 1994); the assembly typically has 50–75% of the cross-sectional area open to gas flow.

6. POWER AND EFFICIENCY

A thermoacoustic engine converts heat into acoustic work. More specifically, it absorbs heat Q_H from a heat source at high temperature T_H, rejects heat Q_C to a heat sink at low temperature T_C, and produces work W. The first law of thermodynamics determines that $W + Q_C = Q_H$; the second law shows that the efficiency W/Q_H is bounded above by the Carnot efficiency $1 - T_C/T_H$.

Similarly, a thermoacoustic refrigerator uses acoustic work W to absorb heat Q_C from a heat source at low temperature T_C, rejecting waste heat Q_H to a heat sink at higher temperature T_H. The first law of thermodynamics determines that $W + Q_C = Q_H$; the second law shows that the relevant efficiency, known as the coefficient of performance COP $= Q_C/W$, is bounded above by $\mathrm{COP}_{\mathrm{Carnot}} = T_C/(T_H - T_C)$.

Experience has shown that the power (here, heat per unit time) of standing-wave thermoacoustic devices can be estimated roughly by $p_m a A |p_1/p_m|^2/30$, with A the cross-sectional area of the stack. Thus, high amplitude (as a fraction of mean pressure), high mean pressure, and high sound speed increase power density. Heat flux densities of the order of 1 MW/m^2 are easily achievable in high-pressure helium gas, which has the highest sound speed of the inert gases, and so it is often used, typically at a pressure above 10 bar. Helium's high thermal conductivity, the largest of the inert gases, is an additional advantage, allowing spacings within the stack and heat exchangers to be conveniently large. In today's thermoacoustic systems, $|p_1/p_m|$ ranges from 0.03 to 0.10, chosen as a compromise between the high power density achieved at high amplitude and the high confidence in the quantitative accuracy of calculations (described in the next section) for systems operating at low amplitude.

Calculations show that standing-wave thermoacoustic systems have second-law efficiencies (i.e., efficiencies normalized by the Carnot efficiency) as high as 0.5, considering effects in the stack only. Including resonator, heat exchanger, and transducer losses reduces this efficiency to between 0.15 and 0.35 for large–cross-section systems, for which resonator losses are not a significant fraction of the total power. Small-diameter systems can have even lower efficiencies, as a result of resonator losses, auxiliary equipment (such as circulating liquid loops for heat transfer to air streams), and heat losses through insulation to ambient temperature. Traveling-wave thermoacoustic systems can have higher efficiencies.

To date, the most efficient laboratory standing-wave engine and refrigerator have had second-law efficiencies of 0.2. This is impressive for devices with no moving parts and is comparable to the efficiencies of small, inexpensive commercial equipment. It falls far short of the efficiencies of well-engineered, expensive steam turbines or large-scale vapor-compression refrigeration equipment, for which over 80% of Carnot has been achieved. If future inventions and improvements to basic understanding can improve the efficiency or raise the power density of thermoacoustic engines and refrigerators, without sacrificing their simplicity, they will find more widespread use.

The efficiency of standing-wave thermoacoustic devices falls below Carnot's efficiency because of six major sources of irreversibility, which we call inherent loss, viscous loss, conduction loss, auxiliary losses, transduction loss, and unknown loss:

1. Inherent loss arises from the heat transfer to and from each parcel of gas in the stack as illustrated in Figs. 3 and 4: Whenever heat dq is transferred across a nonzero temperature difference δT, the entropy of the universe increases (Bejan, 1982) by $dq\ \delta T/T^2$. This irreversibility is unavoidable in the standing-wave thermoacoustic process, relying as it does on *imperfect* thermal contact across the thermal penetration depth for the correct phasing between pressure oscillations and thermal expansion/contraction, and on nonzero heat transfer across that imperfect thermal contact for work production or refrigeration.
2. Viscous loss in the stack occurs because work must be expended in overcoming viscous shear forces as the gas oscillates between the plates. The viscous penetration depth is nearly as large as the thermal penetration depth, so that most of the gas in the stack experiences significant viscous shear (with the parcels in Figs. 3 and 4 actually experiencing significant shape distortions).
3. Simple conduction of heat from the hot heat exchanger to the cold heat exchanger through the stack material and the gas reduces system efficiency.
4. The first three losses occur in auxiliary parts of a thermoacoustic system. Viscous and inherent losses in the heat exchangers, and conduction loss in the portion of the resonator case surrounding the stack, are usually the most important auxiliary losses in large systems, while viscous and inherent losses on other surface areas in the resonator are important in small systems.
5. With rare exceptions such as shown in Fig. 1(b), electric power—not acoustic power—is the usual interface to external systems, so that electroacoustic transducers are needed. In principle, these can be 100% efficient; in practice, 90% appears possible, while 55% has been achieved to date in laboratory thermoacoustic systems.
6. For $|p_1/p_m|$ near 0.1, systems built to date have suffered from additional losses of unknown origin (see below).

For many designs with high power density, the first four sources of irreversibility contribute roughly equally to the inefficiency of thermoacoustic devices. Some reasons for this balance are easy to understand. First, thermoacoustic systems have inherent trade-offs between power density and efficiency, due to the inherent losses in the stack. In the inviscid picture described in Secs. 4.1 and 4.2, at the critical temperature gradient there are no inherent losses, but there is no work done or dissipated and no heat transported; the difference between dT_m/dx and ∇T_{crit} controls both power and inherent loss. Thus, for a given temperature difference across which the engine or refrigerator must operate, the choice of the length of the stack is essentially a compromise between power density and efficiency. Often this compromise is made with $dT_m/dx \sim 2\nabla T_{\text{crit}}$ for engines and $\sim \nabla T_{\text{crit}}/2$ for refrigerators. Second, stack location in the standing wave controls the trade-off between viscous losses, which are highest when the stack is near the pressure node, and conduction losses, which are highest near the pressure antinode where ∇T_{crit} is steepest. Usually this trade-off leads to a stack position much closer to the pressure antinode than to the node; the proportions shown in Fig. 5 are typical.

One way to increase efficiency is to use a mixture of helium and a heavier inert gas as the working substance to reduce viscous effects. Kinetic theory predicts that the Prandtl number $\sigma = \mu c_p/\kappa = \frac{2}{3}$ for hard-sphere monatomic gases, and indeed real monatomic gases have values very close to this (e.g., $\sigma = 0.68$ for helium at room temperature). Fortunately, mixtures of a heavy and a light monatomic gas have Prandtl numbers significantly lower than $\frac{2}{3}$. The most efficient thermoacoustic refrigerator mentioned above (Garrett *et al.*, 1993) used 89% helium, 11% xenon, with $\sigma = 0.27$, to achieve its highest efficiency of 20% of Carnot (cf. 17% of Carnot in Fig. 5, for a similar apparatus optimized for pure helium gas). However, with a sound speed less than half that of pure helium, the power density using this gas mixture was reduced.

7. METHODS OF COMPUTATION

Although attempts at theoretical understanding of these phenomena began with Lord Rayleigh 120 years ago, a quantitatively accurate theory of thermoacoustics was not developed until the 1970s, through the efforts of Nicholas Rott and co-workers (Rott, 1980). It is remarkable that such qualitatively simple classical phenomena went without a quantitatively accurate explanation until late in the 20th century.

7.1 Rott's Approach

Quantitative thermoacoustics analysis is based on Rott's wave equation and energy-flux equation. These equations result from the momentum, continuity, and energy equations in the acoustic approximation. The results are valid for arbitrary phasing between p_1 and $\langle u_1 \rangle$, i.e., for both standing-wave and traveling-wave systems.

As with simpler one-dimensional acoustic-wave equations, Rott's wave equation may be written as two coupled first-order differential equations in the oscillatory pressure $p_1(x)$ and the spatially averaged oscillatory velocity $\langle u_1(x) \rangle$. Rott's energy-flux equation may be regarded as a first-order differential equation in mean temperature $T_m(x)$, coupled to the pressure and velocity equations. In this form, displayed in Table 2, Rott's equations are suitable for numerical integration along x.

Table 2. Expressions used for numerical integrations in ideal-gas thermoacoustics. Compare to Eqs. (8) and (15), which are appropriate when $dT_m/dx = 0$.

$$\frac{dp_1}{dx} = -\frac{i\omega\rho_m}{(1 - f_\nu)} \langle u_1 \rangle$$

$$\frac{d\langle u_1 \rangle}{dx} = -\frac{i\omega}{\rho_m a^2} [1 + (\gamma - 1) f_\kappa] p_1 + \frac{(f_\kappa - f_\nu)}{(1 - f_\nu)(1 - \sigma)} \frac{1}{T_m} \frac{dT_m}{dx} \langle u_1 \rangle$$

$$\frac{dT_m}{dx} = \left\{ \frac{\dot{H}_2}{A_{\text{gas}}} - \tfrac{1}{2} \operatorname{Re}\left[p_1 \langle \tilde{u}_1 \rangle \left(1 - \frac{(f_\kappa - \tilde{f}_\nu)}{(1 + \sigma)(1 - \tilde{f}_\nu)} \right) \right] \right\} \left\{ \frac{\rho_m c_p |\langle u_1 \rangle|^2}{2\omega(1 - \sigma^2)|1 - f_\nu|^2} \operatorname{Im}(f_\kappa + \sigma \tilde{f}_\nu) - \kappa - \frac{A_{\text{solid}}}{A_{\text{gas}}} \kappa_{\text{solid}} \right\}^{-1}$$

The equations are based on the acoustic approximation, with the relevant variables written as

$$p = p_m + \mathrm{Re}[p_1(x)e^{i\omega t}], \tag{18}$$

$$u = \mathrm{Re}[u_1(x,y,z)e^{i\omega t}], \tag{19}$$

$$v, w = \text{similar to } u, \tag{20}$$

$$T = T_m(x) + \mathrm{Re}[T_1(x,y,z)e^{i\omega t}], \tag{21}$$

$$\rho, s = \text{similar to } T, \tag{22}$$

$$\mu = \mu(x), \tag{23}$$

$$\kappa, a, c_p, \text{ etc.} = \text{similar to } \mu. \tag{24}$$

We are most interested in T_m, p_1, and $\langle u_1 \rangle$.

For clarity, we consider here only the case of large solid heat capacity ($\epsilon_s = 0$ in Swift, 1988), so that the temperature of the solid material in the stack is simply $T_m(x)$, independent of time, y, and z, and we restrict ourselves to ideal gases.

The expression for dp_1/dx shown in Table 2 is identical to Eq. (8) derived in Sec. 3; the presence of nonzero dT_m/dx has no effect on the acoustic momentum equation. In effect, we regard this approximation to the momentum equation as the origin of pressure gradient in thermoacoustics: The moving gas causes the pressure gradient. If $f_\nu = 0$, the pressure gradient is entirely "inertial," but when $f_\nu \neq 0$ the presence of viscosity and the stationary boundaries adds a resistive component to the pressure gradient and also effectively changes the magnitude of the inertial contribution.

In both the continuity and energy equations an expression for the spatial average over y and z of the oscillatory temperature, $\langle T_1 \rangle$, is required. As in Sec. 3, it is derived from the general equation of heat transfer, for which the acoustic approximation is

$$\rho_m c_p\left(i\omega T_1 + u_1\frac{dT_m}{dx}\right) - i\omega p_1 = \kappa\left[\frac{\partial^2 T_1}{\partial y^2} + \frac{\partial^2 T_1}{\partial z^2}\right] \tag{25}$$

when $dT_m/dx \neq 0$. Following the same procedure as in Sec. 3, we find

$$T_1 = \frac{1}{\rho_m c_p}(1 - h_\kappa)p_1 + \frac{i}{\omega(1 - f_\nu)}\frac{dT_m}{dx}\left(1 - \frac{h_\kappa - \sigma h_\nu}{1 - \sigma}\right)\langle u_1 \rangle \tag{26}$$

and

$$\langle T_1 \rangle = \frac{1}{\rho_m c_p}(1 - f_\kappa)p_1 + \frac{i}{\omega(1 - f_\nu)}\frac{dT_m}{dx}\left(1 - \frac{f_\kappa - \sigma f_\nu}{1 - \sigma}\right)\langle u_1 \rangle \tag{27}$$

[cf. Eqs. (10) and (11)].

Using Eq. (27) in the continuity equation $i\omega\langle \rho_1 \rangle + d(\rho_m\langle u_1 \rangle)/dx = 0$ yields

$$\frac{d\langle u_1 \rangle}{dx} = -\frac{i\omega}{\rho_m a^2}[1 + (\gamma - 1)f_\kappa]p_1 + \frac{(f_\kappa - f_\nu)}{(1 - f_\nu)(1 - \sigma)}\frac{1}{T_m}\frac{dT_m}{dx}\langle u_1 \rangle, \tag{28}$$

which is also shown in Table 2. This expression finds easy physical interpretation as a thermoacoustic approximation to the continuity equation, expressing the fact that $d\langle u_1 \rangle/dx$ is nonzero if and only if there is a density oscillation in the gas between x and $x + dx$. The two terms in Eq. (28) show that such density change can be caused either by pressure or by velocity along the temperature gradient. Consider the pressure term first. If $f_\kappa = 0$, there is no thermal contact between gas and solid, and so the density oscillations are adiabatic; in this case, $\rho_m a^2$ is the correct compressibility. If $f_\kappa = 1$, the thermal contact between gas and solid is perfect, and so the gas is anchored at the local solid temperature. In this case, $\rho_m a^2/\gamma$, the isothermal compressibility, is appropriate. For intermediate thermal contact, an effective compressibility $\rho_m a^2/[1 + (\gamma - 1)f_\kappa]$, intermediate in magnitude and with nontrivial phase, describes the density oscillations in response to pressure oscillations. Next consider the velocity term, in the easily interpreted inviscid limit, with $f_\nu = 0$ and $\sigma = 0$. If $f_\kappa = 0$, there is no thermal contact between gas and solid, the velocity term is zero, and the gas density does not respond to motion along the temperature gradient. In the other extreme, if $f_\kappa = 1$, the gas is always at the local solid temperature, so that as it flows toward higher T_m its density decreases and its velocity increases. In the more interesting, intermediate regime, oscillatory motion of the gas along the temperature gradient leads to complex density oscillations.

At this point, we could combine Eqs. (8) and (28), eliminating $\langle u_1 \rangle$, to obtain a second-order differential equation in p_1, Rott's wave equation:

$$[1 + (\gamma - 1)f_\kappa]p_1 + \frac{\rho_m a^2}{\omega^2}\frac{d}{dx}\left(\frac{1 - f_\nu}{\rho_m}\frac{dp_1}{dx}\right) - \frac{a^2}{\omega^2}\frac{f_\kappa - f_\nu}{1 - \sigma}\frac{1}{T_m}\frac{dT_m}{dx}\frac{dp_1}{dx} = 0. \tag{29}$$

However, for numerical computations, it is easiest to use Eqs. (8) and (28) separately.

Finally, the energy equation yields an expression for dT_m/dx. To lowest order, the total energy flux (Rott, 1980; Swift, 1988) is approximately

$$\begin{aligned}\dot{H}_2 &= A_{\text{gas}}\langle \rho_m \overline{w_1 u_1}\rangle - (A_{\text{gas}}\kappa + A_{\text{solid}}\kappa_{\text{solid}})\frac{dT_m}{dx} \\ &= A_{\text{gas}}\rho_m c_p \langle \overline{T_1 u_1}\rangle - (A_{\text{gas}}\kappa + A_{\text{solid}}\kappa_{\text{solid}})\frac{dT_m}{dx}.\end{aligned} \tag{30}$$

Here w is the enthalpy per unit mass of the gas. Using Eqs. (7) and (26) for u_1 and T_1, respectively, and taking the spatial averages of the products in Eq. (30), we find

$$\begin{aligned}\dot{H}_2 = {} & \frac{A_{\text{gas}}}{2}\text{Re}\left[p_1\langle \tilde{u}_1\rangle\left(1 - \frac{(f_\kappa - \tilde{f}_\nu)}{(1+\sigma)(1 - \tilde{f}_\nu)}\right)\right] \\ & + \frac{A_{\text{gas}}\rho_m c_p |\langle u_1\rangle|^2}{2\omega(1 - \sigma^2)|1 - f_\nu|^2}\text{Im}(f_\kappa + \sigma\tilde{f}_\nu)\frac{dT_m}{dx} \\ & - (A_{\text{gas}}\kappa + A_{\text{solid}}\kappa_{\text{solid}})\frac{dT_m}{dx}.\end{aligned} \tag{31}$$

The time-averaged energy flux $\dot{H}_2$ *along the stack must be independent of* x, because the case surrounding the stack is assumed to be rigid and thermally insulated and the system is in time-averaged steady state. With $\dot{H}_2$ *constant, we solve Eq. (31) for* dT_m/dx as shown in Table 2, using it to predict the evolution of T_m.

Because the first two equations in Table 2 are complex, these three coupled equations represent five real coupled equations in the five variables $\text{Re}[p_1]$, $\text{Im}[p_1]$, $\text{Re}[\langle u_1\rangle]$, $\text{Im}[\langle u_1\rangle]$, and T_m, and as such are easily integrated numerically using, for example, a Runge–Kutta method (see ANALYTIC METHODS; SCIENTIFIC COMPUTING BY NUMERICAL METHODS). These results are the basis of all present quantitative analyses of standing-wave system and are equally valid for traveling-wave systems.

The curves in Fig. 5(b), in reasonable agreement with the data, were calculated using publicly available software (Ward and Swift, 1994) based on this theory. The software combines the thermoacoustic computations with similar acoustic solutions for the other portions of the resonator, using continuity of T_m, p_1, and complex volumetric velocity $A\langle u_1\rangle$ at the junctions between ducts and thermoacoustic components. The calculations shown in Fig. 5(b) had no adjustable parameters; they simply used the geometry of the apparatus and the properties of helium gas.

7.2 Approximate Methods

The thermoacoustic calculation method summarized in the previous subsection is suitable for numerical integration. In this subsection we discuss approximate methods, suitable for making estimates of standing-wave system performance on the blackboard, with a hand calculator, or in a spreadsheet. An assumption common to all such approximate methods (Swift, 1988; Atchley, 1992) is that the stack is short enough and the temperature spanned is small enough that all x- and T_m-dependent variables can be regarded as constant, including dT_m/dx itself. For example, with this approximation the pressure drop across the stack would be estimated by $\Delta p_1 = -i\omega\rho_m\langle u_1\rangle\Delta x/(1 - f_\nu)$, where Δx is the stack length, with ρ_m, $\langle u_1\rangle$, and f_ν evaluated at the stack midpoint. For further simplification, the boundary-layer limit can be used for f_ν and f_κ with little additional loss in accuracy.

An approximate expression for the acoustic power absorbed or produced in the stack can be obtained by following the same procedure as was used to derive Eq. (16), but including the $\langle u_1\rangle$ term in Eq. (28). The result is

$$\frac{d\dot{W}_2}{dx} = \frac{A_{\mathrm{gas}}\omega}{2}\left[\frac{\rho_m|\langle u_1\rangle|^2}{|1 - f_\nu|^2}\,\mathrm{Im}(-f_\nu) + \frac{(\gamma - 1)|p_1|^2}{\rho_m a^2}\,\mathrm{Im}\left(-f_\kappa + \frac{f_\kappa - f_\nu}{(1 - \sigma)(1 - f_\nu)}\,\frac{dT_m}{dx}\,\frac{i\rho_m c_p\langle u_1\rangle}{\omega p_1}\right)\right]. \tag{32}$$

The total stack power $\Delta\dot{W}_2$ can then be obtained approximately by evaluating the right-hand side using values at the stack midpoint and multiplying by the stack length. This equation shows that a modest value of dT_m/dx reduces the power dissipated in the stack and that a large enough value of dT_m/dx changes the sign of $\Delta\dot{W}_2$, so that the stack produces power. For $\mu = 0$ and standing-wave phasing, the crossover occurs at $dT_m/dx = \nabla T_{\mathrm{crit}}$.

Equations (31) and (32) can be further simplified by using either the boundary-layer limit expressions for f_κ and f_ν or exact standing-wave phasing. With both these assumptions simultaneously, simple approximate expressions result:

$$\dot{H}_2 = \frac{A_{\mathrm{gas}}}{4}\,\frac{\delta_\kappa}{r_h}\,\frac{|p_1||\langle u_1\rangle|}{(1 + \sigma)e}\left(1 + \sqrt{\sigma} - \frac{\delta_\nu}{r_h} - \Gamma\,\frac{1 + \sqrt{\sigma} + \sigma}{1 + \sqrt{\sigma}}\right) - (A_{\mathrm{gas}}\kappa + A_{\mathrm{solid}}\kappa_{\mathrm{solid}})\,\frac{dT_m}{dx}, \tag{33}$$

$$\Delta\dot{W}_2 = \frac{A_{\mathrm{gas}}\Delta x}{4r_h}\left[\frac{\rho_m|\langle u_1\rangle|^2\delta_\nu\omega}{e} + \frac{(\gamma - 1)|p_1|^2\delta_\kappa\omega}{\gamma p_m}\left(1 - \frac{\Gamma}{(1 + \sqrt{\sigma})e}\right)\right], \tag{34}$$

where

$$\Gamma = (dT_m/dx)/\nabla T_{\mathrm{crit}}, \tag{35}$$

$$\nabla T_{\mathrm{crit}} = \omega|p_1|/\rho_m c_p|\langle u_1\rangle|, \tag{36}$$

$$e = 1 - \delta_\nu/r_h + \delta_\nu^2/2r_h^2. \tag{37}$$

These expressions are easily used in spreadsheets or for hand calculations.

7.3 Similitude

The most promising route to higher power densities is by increasing $|p_1/p_m|^2$, but doing so takes us farther from the range of applicability of Rott theory, which is based on small oscillatory amplitude. Rott's assumptions include gas displacement amplitude much smaller than the length of the stack and other components, Reynolds number of the oscillations small enough to avoid turbulence, and $|p_1| \ll p_m$. Extension of thermoacoustics theory beyond these assumptions will be challenging. In the meantime, application of similitude to thermoacoustics (Olson and Swift, 1994) has shown how to organize and correlate experimental data outside the range of our current quantitative understanding and allows meaningful experimental study of scale models of thermoacoustic apparatus reminiscent of wind-tunnel models of aircraft.

The similitude analysis shows that the dimensionless variable groups in Table 3 provide a *complete* description of the gas in a thermoacoustic system. (Additional variables can be added to include transducers or solid properties if needed.) The analysis is powerful because it is true even when the thermoacoustic system is operating beyond the assumptions inherent in Rott's method. For example, it "includes" turbulence, even though the Reynolds number is not explicitly present in the table.

Figure 6, from the similitude study (Olson and Swift, 1994), illustrates both the power of similitude analysis and the amplitude dependence of the accuracy of our current un-

Table 3. Dimensionless groups relevant to thermoacoustics. The reference location, indicated by the subscript "ref," can be the room-temperature end of the stack. Of the many dimensions x_j of the apparatus, three are singled out for special attention: the overall length L, the stack area A_{ref} at the reference location, and its hydraulic radius r_h at the reference location. The constants b_μ and b_κ account for temperature dependences of μ and κ: $\mu = \mu_{\mathrm{ref}}(T/T_{\mathrm{ref}})^{b_\mu}$ and similarly for b_κ.

Dimensions	A_{ref}/L^2, r_h/L, x_j/L
Gas properties	γ, σ_{ref}, b_μ, b_κ
Thermophysical lengths	fL/a_{ref}, δ_κ/r_h
Powers	Power$/p_m a_{\mathrm{ref}} A_{\mathrm{ref}}$
Spatially dependent variables	$T(\mathbf{x},t)/T_{\mathrm{ref}}$, $p(\mathbf{x},t)/p_m$, $\mathbf{v}(\mathbf{x},t)/a_{\mathrm{ref}}$

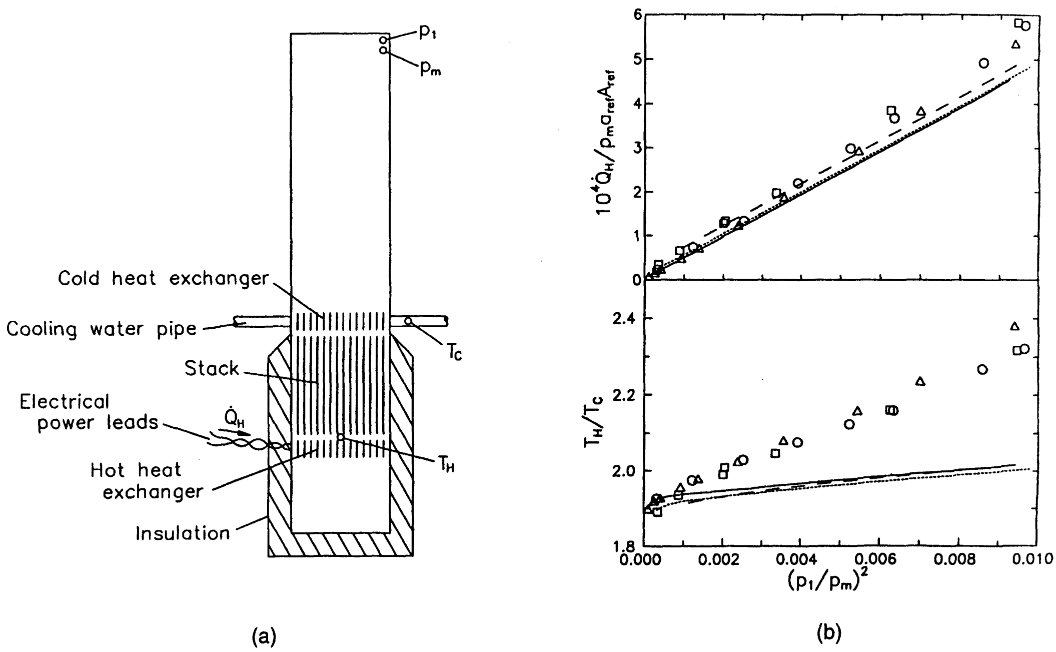

FIG. 6. A recent example of thermoacoustic measurements and calculations. (a) Schematic of apparatus, a standing-wave engine driving no load. Small circles indicate locations of sensors. (b) Heater power $\dot{Q}_H$ and hot temperature T_H as functions of $|p_1|$, all normalized. Points are measurements; lines are calculations. Triangles and solid lines: helium, 0.96 MPa. Circles and dotted lines: neon, 0.70 MPa. Squares and dashed lines: argon, 0.36 MPa.

derstanding of thermoacoustic phenomena. The data shown as points in Fig. 6(b) were obtained with the apparatus sketched in Fig. 6(a), a thermoacoustic engine driving no load. The plots can be interpreted as displaying what heater power $\dot{Q}_H$ *and hot temperature* T_H were required at the hot heat exchanger to maintain steady oscillations at a given pressure amplitude $|p_1|$.

The lines are the results of calculations (Ward and Swift, 1994) based on Rott's theory, the geometry of the apparatus, and the properties of the gases. The calculations agree well with the measurements in the limit of small $|p_1|$, as expected from the assumptions in the theory. However, at $|p_1|/p_m \sim 0.1$, measurements deviate from calculations significantly. The deviations are not surprising, in that they are of the same order as $|p_1|/p_m$, but they are disturbing from a practical point of view, because both $\dot{Q}_H$ *and* T_H deviate from Rott theory in directions that decrease the efficiency.

Three monatomic gases were used in the measurements, with different average pressures selected to make δ_κ the same for all three cases, ensuring similarity. The crux of the similitude analysis is that the only relevant thermophysical length scales are λ, δ_κ, and δ_ν, so that if these are held fixed all other dimensionless groups (such as Reynolds and Mach numbers) will also stay fixed. The data for all three gases do indeed fall along the same curves, demonstrating similarity. Although the similitude analysis does not indicate the cause of the deviations of the measurements from Rott-style calculations, it does allow elimination of some potential suspects, such as gravity-driven convection, resonator vibration, and stack flutter, which are not included in the similitude analysis and would not be expected to scale from gas to gas.

Similitude also permits the building of meaningful thermoacoustic scale models. For example, to model a full-size helium thermoacoustic engine at mean pressure p_m, we can use a half-scale model with argon at $0.79p_m$, the pressure that ensures that δ_κ in the argon is exactly half its value in the full-size helium system. Experiments with the scale model can be easier than those with the full-size system because of its smaller size and because powers are reduced by $(p_m a A)_{\mathrm{Ar}}/(p_m a A)_{\mathrm{He}} = \frac{1}{16}$.

GLOSSARY

Heat Engine: Thermodynamic device that converts heat to work.

Refrigerator: Thermodynamic device that uses work to transfer heat from low to high temperature.

Regenerator: Key heat-exchange element in traveling-wave thermoacoustic device, a porous structure with nonnegligible heat capacity.

Stack: Key heat-exchange element in standing-wave thermoacoustic device, an open structure with nonnegligible heat capacity.

Standing Wave: Wave with oscillatory pressure and oscillatory motion substantially in phase.

Thermoacoustic: With both thermodynamic and acoustic characteristics.

Traveling wave: Wave with oscillatory pressure and oscillatory motion substantially out of phase.

Works Cited

Arnott, W. P., Bass, H. E., Raspet, R. (1991), "General Formulation of Thermoacoustics for Stacks Having Arbitrarily Shaped Pore Cross Sections," *J. Acoust. Soc. Am.* **90,** 3228–3237.

Atchley, A. A. (1992), "Standing Wave Analysis of a Thermoacoustic Prime Mover below Onset of Self-oscillation," *J. Acoust. Soc. Am.* **92,** 2907–2914.

Bejan, A. (1982), *Entropy Generation through Heat and Fluid Flow,* New York: Wiley.

Carter, R. L., White, M., Steele, A. M. (1962), Private communication of Atomics International Division of North American Aviation Inc. See also Feldman, K. T. (1966), "A Study of Heat Generated Pressure Oscillations in a Closed End Pipe, Ph.D. Dissertation, Mechanical Engineering Department, University of Missouri, Columbia, MO.

Ceperley, P. H. (1979), "A Pistonless Stirling Engine—The Traveling Wave Heat Engine," *J. Acoust. Soc. Am.* **66,** 1508–1513.

Ceperley, P. H. (1985), "Gain and Efficiency of a Short Traveling Wave Heat Engine," *J. Acoust. Soc. Am.* **77,** 1239–1244.

Corey, J. A., Yarr, G. A. (1992), "HOTS to WATTS: The LPSE Linear Alternator System Reinvented," in: *Proceedings of the 27th Intersociety Energy Conversion Engineering Conference,* Warrendale, PA: Society of Automotive Engineers, pp. 5.289–5.294.

Garrett, S. L., Adeff, J. A., Hofler, T. J. (1993), "Thermoacoustic Refrigerator for Space Applications," *J. Thermophysics and Heat Transfer (AIAA)* **7,** 595–599.

Garrett, S. L., Perkins, D. K., Gopinath, A. (1994), "Thermoacoustic Refrigerator Heat Exchangers: Design, Analysis, and Fabrication," in: G. F. Hewitt (Ed.), *Heat Transfer 1994: Proceedings of the 10th International Heat Transfer Conference,* Vol. 4, Rugby, England: Institution of Chemical Engineers, pp. 375–380.

Hofler, T. J. (1986), "Thermoacoustic Refrigerator Design and Performance," Ph.D. Thesis, Physics Department, University of California, San Diego.

Hofler, T. J. (1988), "Concepts for Thermoacoustic Refrigeration and a Practical Device," in: *Proceedings of the 5th International Cryocoolers Conference, August 1988, Monterey, California,* Wright-Patterson, OH: Wright-Patterson Air Force Base, pp. 93–101.

Migliori, A., Swift, G. W. (1988), "Liquid Sodium Thermoacoustic Engine," *Appl. Phys. Lett.* **53,** 355–357.

Olson, J. R., Swift, G. W. (1994), "Similitude in Thermoacoustics," *J. Acoust. Soc. Am.* **95,** 1405–1412.

Radebaugh, R. (1990), "A Review of Pulse Tube Refrigeration," *Adv. Cryogenic Eng.* **35,** 1191–1205.

Rott, N. (1980), "Thermoacoustics," *Adv. Appl. Mech.* **20,** 135–175, and references therein.

Swift, G. W. (1988), "Thermoacoustic Engines," *J. Acoust. Soc. Am.* **84,** 1145–1180.

Swift, G. W., Keolian, R. M. (1993), "Thermoacoustics in Pin-Array Stacks," *J. Acoust. Soc. Am.* **94,** 941–943.

Urieli, I., Berchowitz, D. M. (1984), *Stirling Cycle Engine Analysis,* Bristol, U.K.: Adam Hilger.

Ward, W. C., Swift, G. W. (1994), "Design Environment for Low Amplitude Thermoacoustic Engines," *J. Acoust. Soc. Am.* **95,** 3671–3672. Fully tested software and user's guide available from Energy Science and Technology Software Center, U.S. Department of Energy, Oak Ridge, TN. For a beta-test version, contact ww@lanl.gov (Bill Ward) via Internet.

Yaron, R., Shokralla, S., Yuan, J., Bradley, P. E., Radebaugh, R. (1996), "Etched Foil Regenerator," *Adv. Cryogenic Eng.* **41,** 1339–1346.

Further Reading

Ballister, S. C., McKelvey, D. J. (1995), "Shipboard Electronics Thermoacoustic Cooler," Master's Thesis, Naval Postgraduate School, Monterey, California.

Bejan, Adrian (1996), *Entropy Generation Minimization,* New York: CRC Press.

Godshalk, K. M., Jin, C., Kwong, Y. K., Hershberg, E. L., Swift, G. W., Radebaugh, R. (1996), "Characterization of 350 Hz Thermoacoustic Driven Orifice Pulse Tube Refrigerator with Measurements of the Phase of the Mass Flow and Pressure," *Adv. Cryogenic Eng.* **41,** 1411–1418.

Swift, G. W. (1992), "Analysis and Performance of a Large Thermoacoustic Engine," *J. Acoust. Soc. Am.* **92,** 1551–1563.

Readers with Internet access can obtain computer animations of the processes described in Secs. 3 and 4 at http://rott.esa.lanl.gov/; select "educational demonstrations."

THERMOCHEMISTRY*

P. A. G. O'Hare, *Chemical Kinetics and Thermodynamics Division, Chemical Science and Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland, U.S.A.*

Introduction 265
1. **Basic Concepts** 266
1.1 Thermochemistry and the Laws of Thermodynamics 266
1.2 Second- and Third-Law Calculations 266
1.3 Hess's Law 268
2. **Standard States** 268
3. **Standard Molar Enthalpy of Formation** 269
3.1 Units of Thermochemical Quantities 269
4. **Compilations of Thermodynamic Data** 269
5. **Apparatus for Thermochemical Studies** 270
5.1 A Generic Calorimeter 270
5.2 Combustion Calorimeters 270
5.3 Synthesis Calorimeters 273
5.4 Calibration of Combustion Calorimeters 273
5.5 Solution Calorimeters 274
6. **Applications of Thermochemical Information** 275
6.1 Calorimetric Results 275
6.1.1 Combustion Calorimetry 275
6.1.2 Solution Calorimetry 276
6.1.3 A Third-Law Calculation of $\Delta_f H_m^0$ 277
Glossary 277
Works Cited 278
Further Reading 278

INTRODUCTION

The branch of experimental thermodynamics that deals with changes in enthalpy or energy (both quantities are defined in Sec. 1) associated with chemical reactions is known as thermochemistry. It has practical applications in all branches of science. Thermochemical studies provide valuable information about, for example, the energy to be expected from combustion of fossil fuels; atmospheric pollution; equilibria in geological formations; products and consequences of an accident in a nuclear reactor; chemical effects of enzymes in the physiological milieu; compositions of planetary atmospheres; and strengths of chemical bonds between atoms in molecules. Modern chemical industry is very dependent on thermochemical data for the design of efficient processes, where the concept of "heat balance" comes into play. Thermochemical calculations applied to putative chemical processes may be used to identify reactions that are feasible (but that, ultimately, could be limited by kinetic factors) and those that are not. For example, spontaneous conversion of diamond to graphite is favored thermodynamically but is precluded by kinetic limitations. Put to such uses, thermochemistry can be of great practical and economic value.

It follows, then, that the provision of high-quality thermochemical information is an important scientific pursuit in its own right. High quality implies, above all else, accuracy; a secondary index of merit of thermochemical values is precision, although it, in no way, can be considered to imply accuracy. In fact, those two qualities are quite distinct. Accuracy is related to the closeness with which a particular experimental value approaches the true one, while precision reflects the reproducibility with which experimental results are measured. Thus, a precise result does not necessarily connote an accurate one.

---

*Contribution of the National Institute of Standards and Technology, not subject to U.S. copyright.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. BASIC CONCEPTS

1.1 Thermochemistry and the Laws of Thermodynamics

The close relation of thermochemistry with thermodynamics means that the former is subject to the laws of the latter. As to their impact on thermochemistry, it can be said that, of the three laws of thermodynamics, the first is the most significant:

$$\Delta U = q + w + u. \tag{1}$$

Here, ΔU represents the change in internal energy of a system; q, the heat energy; w, the "pressure-volume" work energy; and u, any other energy absorbed or released by the system. If $u = 0$ in a particular experiment, and the volume of the system is kept constant, so that $w = 0$, then

$$\Delta U = q. \tag{2}$$

If, on the other hand, an experiment is performed in which $u = 0$ also, but the pressure is kept constant, then

$$w = -\int p dV = -p \int dV = -p \, \Delta V, \tag{3}$$

where dV denotes an infinitesimal change in the volume of the system. Thus, from Eq. (1),

$$\Delta U = q - p \, \Delta V. \tag{4}$$

It follows that

$$\Delta U + p \, \Delta V = q, \tag{5}$$

and

$$\Delta U + \Delta(pV) = q. \tag{6}$$

In other words,

$$\Delta(U + pV) = q. \tag{7}$$

But, since

$$H = U + pV \tag{8}$$

by definition (see THERMODYNAMICS, EQUILIBRIUM, then

$$\Delta H = q. \tag{9}$$

Thus, depending on experimental conditions, the measured heat is either ΔU (constant volume) or ΔH (constant pressure). Neither U nor H can be determined absolutely, and so, in practice, the difference in those properties between one state and another, denoted by ΔU or ΔH, is measured. Because most thermochemical experiments are performed at constant (atmospheric) pressure, ΔH is reported more frequently than ΔU. When one is dealing with chemical reactions, as we are in the greater part of this article, those quantities are further decorated as $\Delta_r H_m$ or $\Delta_r U_m$, where the subscript r indicates that the change in H or U refers to a chemical reaction and the subscript m denotes a molar quantity. Such changes in energy are customarily measured in a calorimeter, and we discuss several variants of that apparatus in a later section.

An immutable requirement of the first law of thermodynamics is that the initial and final states of systems under investigation must be known precisely, but the path between them need not. Thus, for example, the initial and final temperatures and pressures have to be specified. In the case of chemical reactions, the exact compositions of starting materials and products have to be known as well.

1.2 Second- and Third-Law Calculations

Although, at first sight, the thermochemistry of any chemical change should be amenable to calorimetric study, it is found in practice that many reactions do not occur, or occur very slowly, under normal calorimetric operating conditions of temperature or pressure, and, therefore, other techniques have to be employed to obtain the desired thermochemical information. Of those, two are most common: high-temperature studies of chemical equilibria, and measurements of the potentials of high-temperature electrochemical cells. Although these two approaches may be thought of as having little in common, they are both basically thermodynamic probes of the changes that occur in chemical reactions.

When heated to sufficiently high temperatures, even the most refractory compounds decompose to some extent. Often, only gaseous products are formed:

$$\mathrm{AB(s)} \rightarrow \mathrm{A(g)} + \mathrm{B(g)}. \qquad (10)$$

The equilibrium constant of reaction (10) is given by

$$K = p(\mathrm{A})p(\mathrm{B})/(p^0)^2, \qquad (11)$$

where p denotes the partial pressure of the species, and p^0 is the reference pressure (usually 101.325 or 100 kPa, but we discuss this matter in detail later) and, by convention, the pure solid phase AB is excluded from consideration.

From the second law of thermodynamics, it can be shown that

$$\Delta_r G_m^0 = -RT \ln K, \qquad (12)$$

where $\Delta_r G_m^0$ denotes the standard Gibbs free-energy change for a reaction at the temperature T. (The use of superscript 0 to denote standard state is discussed in Sec. 3.) Clearly, then, an experimental measurement of K can give directly the Gibbs free-energy change of a reaction. At this point, we need to introduce another relation:

$$\Delta_r G_m^0 = \Delta_r H_m^0 - T\,\Delta_r S_m^0, \qquad (13)$$

where $\Delta_r S_m^0$ denotes the change of standard molar entropy S_m^0 for reaction (10). Thus, by combining Eqs. (12) and (13), one obtains

$$\ln K = -\Delta_r H_m^0/RT + \Delta_r S_m^0/R. \qquad (14)$$

In practice, a plot of $\ln K$ against $1/T$ approaches linearity, with the slope of the line corresponding to $-\Delta_r H_m^0/R$. From this operation, $\Delta_r H_m^0$ is obtained at $\langle T \rangle$, the mean temperature of the measurements. This procedure is known as a "second-law" calculation. Values of $\Delta_r H_m^0$ at other temperatures are obtained by application of heat capacity corrections.

As mentioned above, Eq. (12) yields directly a value for the Gibbs free-energy change at $\langle T \rangle$ for a generic reaction

$$\mathrm{A} \rightarrow \mathrm{B} + \mathrm{C}. \qquad (15)$$

This result can be converted to the enthalpy change $\Delta_r H_m^0(T)$ at some other temperature, in practice usually 298.15 K, by a procedure known as the "third-law" method. The fine details of this approach are beyond the scope of this article. Briefly, for each species A, B, and C, numerical values are required for a thermodynamic quantity colloquially referred to as the "Gibbs free-energy function" or the "free-energy function." (Despite lengthy discussions over the past 30 years, international nomenclature bodies have failed to agree on a universally accepted name for this quantity.) For the purpose of this discussion, we shall call the function $\Phi_m^0(T)$; it is equal to $\Delta_0^T S_m^0 - \Delta_{T'}^T H_m^0/T$, where $T' = 298.15$ K. Usually $\Phi_m^0(T)$ is based on measurements of the heat capacity of the solid or liquid or, for gases, is often calculated by means of statistical thermodynamics. In the most favorable cases, reliable values of $\Phi_m^0(T)$ are available for A, B, and C in standard tabulations; however, it is not always so, and estimates often must be made. Mechanically, "third-law" calculations involve the combination of $\Delta_r G_m^0(T)$ with the algebraic sums of $\Phi_m^0(T)$ as follows:

$$\Delta_r H_m^0(298.15\ \mathrm{K}) = \Delta_r G_m^0(T) + T\{\Phi_m^0(\mathrm{B},T) + \Phi_m^0(\mathrm{C},T) - \Phi_m^0(\mathrm{A},T)\}. \qquad (16)$$

In this way, the enthalpy change at $T = 298.15$ K, or any other temperature, may be deduced for a chemical reaction studied at elevated temperatures.

A similar procedure applies to chemical reactions that take place in electrochemical cells. Necessary prerequisites for the validity of such experiments are that the presumed reaction be the true cell reaction and that the electrochemical cell be reversible. In that case,

$$\Delta_r G_m(T) = -nFE(T), \qquad (17)$$

where n denotes the stoichiometric number of electrons in the cell reaction as written, F is the Faraday constant ($F = 96485.3$ $\mathrm{C \cdot mol^{-1}}$), and $E(T)$ is the experimental emf of the cell at the temperature T. Measurements of $E(T)$, combined with a knowledge of n, lead to $\Delta_r G_m(T)$, which, in turn, becomes the starting point for the calculation of $\Delta_r H_m^0(298.15\ \mathrm{K})$ by the "third-law" method, as illustrated in Eq. (16).

1.3 Hess's Law

One statement of the first law of thermodynamics, namely, that the energy of an isolated system remains constant, leads to Hess's law and its application to thermochemical cycles.

Suppose that the chemical reaction W → Z can be broken down into the three simple steps: W → X, X → Y, and Y → Z, which is shown pictorially in Fig. 1. This cycle implies that the reaction of W to form Z is equivalent to the sum of the three consecutive steps: W to form X, X to form Y, and Y to form Z. In other words, one can arrive at Z either directly from W or by the circuitous route from W via X and Y. Hess's law makes a parallel statement to the effect that the enthalpy change in going from W to Z is equal to the sum of the enthalpy changes of the other three individual steps (enthalpy cannot be created or destroyed: it is a "state function"). Thus,

$$\Delta_r H_m^0(1) = \Delta_r H_m^0(2) + \Delta_r H_m^0(3) + \Delta_r H_m^0(4). \tag{18}$$

In thermochemical reasoning and calculations, this law is key. In practice, it means that if $\Delta_r H_m^0$ of the reaction of W to form Z cannot be measured for some reason, or simply has not been determined, it still can be calculated by summing the $\Delta_r H_m^0$'s of the other three reactions. Choice of a four-component cycle is purely arbitrary; in practice, Hess cycles are encountered that comprise many more, but sometimes fewer, component reactions, and the journey from W to Z can encompass many intermediate steps. Needless to say, if even one of the auxiliary $\Delta_r H_m^0$'s is in error, the derived quantity will likewise be in error. The reliability of the calculated $\Delta_r H_m^0(1)$ is no greater than that of any, or all, of the other $\Delta_r H_m^0$'s. Another consequence of Hess's law is that in those cases where $\Delta_r H_m^0$ is known for all components, four as in Eq. (18), then the algebraic sum of the $\Delta_r H_m^0$'s must equal zero within the combined uncertainties of the individual values. Absent that, one of the component values, or some of them or, less likely, all of them, must be in error. This requirement is a powerful diagnostic tool for experimental thermodynamicists and data evaluators alike.

$$\begin{tikzcd} W \arrow[r, "\Delta_r H_m(1)"] \arrow[d, "\Delta_r H_m(2)"'] & Z \\ X \arrow[r, "\Delta_r H_m(3)"'] & Y \arrow[u, "\Delta_r H_m(4)"'] \end{tikzcd}$$

FIG. 1. A thermochemical cycle. The molar enthalpy of reaction is denoted by $\Delta_r H_m$, and $\Delta_r H_m(1) = \Delta_r H_m(2) + \Delta_r H_m(3) + \Delta_r H_m(4)$.

2. STANDARD STATES

Because thermodynamic quantities such as H_m cannot be determined absolutely, they are measured and reported as differences ΔH_m. Clearly, a thermodynamic reference state must be designated for each chemical compound. The standard reference state of a substance at a given temperature is defined as the pure solid or liquid at a specified pressure, or pure gas at a definite fugacity (fugacity may be regarded as a type of corrected pressure; for ideal gases, fugacity and pressure are identical; for real gases, fugacity and pressure approach one another as the pressure approaches zero). For many years, the reference pressure and fugacity had been accepted as 1 atm (1.01325×10^5 Pa), and they are still used by many authors for a variety of reasons. Recently, IUPAC recommended that the standard-state pressure be changed to 1 bar (10^5 Pa), and more and more contemporary studies adhere to that recommendation. Accordingly, investigators are careful to specify the standard pressure that applies to their work. Most modern thermodynamic literature eschews the "atmosphere" and has replaced it with the SI (International System) units of pressure: Pa, kPa, or MPa. Despite those changes, the "normal boiling temperature," of a liquid, for example, still has as its basis the reference pressure 101.325 kPa.

3. STANDARD MOLAR ENTHALPY OF FORMATION

The formation reaction of each substance from the constituent elements is unique, as is the thermodynamic description thereof. For the formation of (solid) ABC from its (solid) reference-state elements,

$$A(s) + B(s) + C(s) \rightarrow ABC(s), \qquad (19)$$

the change of enthalpy has a unique value denoted by the symbol $\Delta_f H_m^0(\mathrm{ABC,s})$, called the standard molar enthalpy of formation of ABC(s). Here, the subscript f shows that the thermochemical quantity refers to a formation reaction, and the superscript 0 specifies that A, B, C, and ABC are all in their standard reference states. The temperature of reaction (19) is necessarily included, as $\Delta_f H_m^0(\mathrm{ABC},s,T)$; frequently, that temperature is 298.15 K (or 25 °C; although in many thermodynamic journals, use of °C now is discouraged). Note that it is essential to designate the physical states (s = solid; l = liquid; g = gas; cr = crystal) of the substances as, for example, in the formation of $CO_2(g)$:

$$\mathrm{C(cr,graphite)} + O_2(g) \rightarrow CO_2(g). \qquad (20)$$

Because carbon can exist as either diamond or *reference-state* graphite (Acheson spectroscopic, to be precise), $\Delta_f H_m^0(\mathrm{CO_2,g})$ must refer to the graphite modification. Note that $\Delta_f H_m^0$ is just a special case of $\Delta_r H_m^0$; thus, $\Delta_r H_m^0(20) \equiv \Delta_f H_m^0(\mathrm{CO_2,g})$.

There is another relation in thermochemistry of great practical importance, because it combines $\Delta_r H_m^0$ and $\Delta_f H_m^0$. For the reaction

$$AB(s) + CD(s) \rightarrow AC(s) + BD(s), \qquad (21)$$

we have

$$\Delta_r H_m^0(21) = \Delta_f H_m^0(\mathrm{AC}) + \Delta_f H_m^0(\mathrm{BD}) - \Delta_f H_m^0(\mathrm{AB}) - \Delta_f H_m^0(\mathrm{CD}). \qquad (22)$$

Thus, if $\Delta_r H_m^0$ is measured, and reliable values of $\Delta_f H_m^0$ for the compounds AC, BD, and AB are available from other sources, then $\Delta_f H_m^0(\mathrm{CD})$ can be deduced without measurement. If, for example, $\Delta_r H_m^0(20)$ is determined experimentally, then

$$\Delta_r H_m^0(20) = \Delta_f H_m^0(\mathrm{CO_2,g}) - \Delta_f H_m^0(\mathrm{C,cr,graphite}) - \Delta_f H_m^0(\mathrm{O_2,g}). \qquad (23)$$

Since the last two quantities in Eq. (23) are for the standard reference states and are, therefore, both zero by definition, the measured standard molar enthalpy of reaction is, in fact, the standard molar enthalpy of formation of $CO_2(g)$.

3.1 Units of Thermochemical Quantities

The fundamental unit of both heat and work is the joule, denoted by the symbol J. Most modern thermochemical quantities are expressed in terms of this unit, although the related *calorie* has not entirely disappeared from the literature (1 cal = 4.184 J). The molar mass of a compound is calculated from the atomic weights of the individual atoms. Thus, for ethene (C_2H_4), the molar mass is $[2(12.011) + 4(1.00794)]\ \mathrm{g \cdot mol^{-1}}$ or $28.054\ \mathrm{g \cdot mol^{-1}}$. Most numerical thermochemical information is given as molar quantities, in units $\mathrm{J \cdot mol^{-1}}$ or $\mathrm{kJ \cdot mol^{-1}}$. Thus, $\Delta_f H_m^0(\mathrm{H_2O,l,298.15\ K}) = -285.83\ \mathrm{kJ \cdot mol^{-1}}$.

4. COMPILATIONS OF THERMODYNAMIC DATA

For day-to-day thermochemical calculations by engineers, scientists, and other users of thermochemical information, the availability of tables of values of $\Delta_f H_m^0$ is essential, because tedious and time-consuming searches of the literature are obviated. Many such tables are accessible; some have as their bases critically evaluated information, others are simply a potpourri of values of frequently marginal reliability that are often not even internally consistent; that is, they do not satisfy Hess's law. All such compilations are available in hard copy, some in electronic versions.

A subset of "key" thermochemical values has been identified by international agreement for simple and frequently encountered molecules and aqueous ions such as O_2, H_2O, HCl, CO_2, SiO_2, $Cl^-(aq)$, and $OH^-(aq)$. In all, a total of 151 species is included. Unlike in some other compilations, numerical values, uncertainties, and the rationale for their selection are recorded (Cox *et al.*, 1989). Extensive and detailed tables of $\Delta_f H_m^0$ and $\Delta_f G_m^0$ (as well as other thermodynamic properties such as heat capacities) that deal mostly, but not exclusively, with inorganic substances are available in the *JANAF Thermochemical Tables* (Chase *et al.*, 1985) and in *Thermodynamic Properties of Individual Substances* (Gurvich *et al.*, 1978 *et seq.*; Russian-language version published as separate volumes at various times, and in an electronic form in English: Gurvich *et al.*, 1993). Sections of this publication are being translated into English, and three volumes have appeared to date (Gurvich *et al.*, 1989, 1991, 1994). Thermochemical properties of organic compounds have been critically evaluated and tabulated by Cox and Pilcher (1970), Pedley (1994), and Frenkel *et al.* (1994), who also include a PC diskette listing of tables.

5. APPARATUS FOR THERMOCHEMICAL STUDIES

The foundations of thermochemistry can been traced back to the end of the 18th century (Domalski, 1979). Shortly before the beginning of the French Revolution, Lavoisier and LaPlace were engaged in pioneering thermochemical studies. Other illustrious names associated with the early history of thermochemistry include Black, Davy, Rumford, and Andrews. The last, for example, measured the energy of combination of hydrogen and oxygen to form water.

In the preceding sections, quantities such as $\Delta_r H_m^0$ and $\Delta_f H_m^0$ figure prominently. Now, we outline how they are determined experimentally. Methods are described for the accurate measurement of $\Delta_r H_m^0$ and $\Delta_r U_m^0$. Apparatus used for that purpose are called calorimeters, and the corresponding branch of science, calorimetry. In essence, we discuss experimental equipment used to quantify the energy or enthalpy changes associated with chemical reactions.

5.1 A Generic Calorimeter

Put simply, a calorimeter for thermochemical experiments frequently consists of a vessel in which the chemical reaction under study takes place, a well-stirred water-filled container in which the vessel sits, and a thermometer. In principle, the energy or enthalpy change of the reaction raises, or lowers, the temperature of the vessel and surrounding water. The thermometer rests in the water, where it measures the resulting change of temperature, which, in turn, is related to the change of energy or enthalpy brought about by the chemical change. General reactions of substances most amenable to calorimetric study have included combustion with an oxidizing gas in a static or flow system, solution in aqueous media or inorganic melts, and direct combination of the component elements at some high temperature. Each approach has special demands and, therefore, requires a calorimeter with its own unique attributes. Several such individual systems are now described.

5.2 Combustion Calorimeters

Thermochemical studies of combustion reactions are most frequently carried out in calorimeters equipped with a combustion bomb, a typical one of which is illustrated in Fig. 2. A bomb is required to confine the high pressure of the oxidizing gas. Let us suppose that the energy of combustion of an organic substance, succinic acid, is to be determined. An accurately weighed pellet of analyzed high-purity succinic acid is placed in the combustion crucible close to the ignition wire. A small quantity of water is pipetted into the bomb, which is then closed by means of the sealing nut and charged with high-purity oxygen to a pressure of ca. 3 MPa. The bomb is then placed in the calorimeter can filled with an accurately known mass of water. This entire assembly sits within a thermostat whose temperature is usually close to 298.15 K. While the water in the can is stirred at a constant rate, the temperature-against-time profile of the calorimeter is recorded, and, when the investigator is satisfied that a steady state has been reached, a measured amount of electrical energy is passed through the ignition circuit, combustion of the succinic acid ensues, and the temperature of the calorimeter increases rapidly. After a steady state has once again been reached, another temperature-against-time profile is recorded, and the experiment is terminated. All being well, the combustion

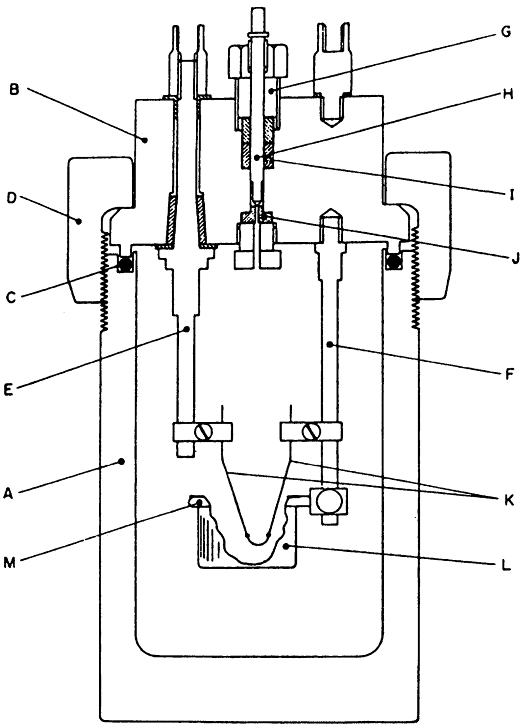

FIG. 2. A typical combustion bomb. A, bomb body; B, bomb head; C, O-ring; D, sealing nut; E, insulated ignition electrode; F, grounded ignition electrode; G, oxygen charging valve; H, valve needle; I, packing gland; J, valve seat; K, ignition wire; L, platinum crucible; M, crucible support ring.

reaction of succinic acid should have proceeded as follows:

$$C_4H_6O_4(cr) + (7/2)O_2(g) \rightarrow 4CO_2(g) + 3H_2O(l). \quad (24)$$

In the most carefully performed determinations, great care would have been expended to ensure that the succinic acid and oxygen were of the highest purity, that the stoichiometrically exact amount of CO_2 had formed, and that by-products such as soot and CO(g) were absent. Recall that, in setting up the experiment, some water was added to the bomb. That was done to ensure that, before the experiment, the gas volume would be saturated with water vapor, which, in turn, would guarantee that all the H_2O produced during the experiment formed as a liquid. In summary, all precautions are taken to make certain that the energy change measured was unequivocally for reaction (24).

Equation (24) represents a relatively straightforward chemical reaction. However, the combustion calorimetry of a molecule containing atoms other than carbon, hydrogen, and oxygen—for example, sulfur, nitrogen, or a halogen (fluorine, chlorine, bromine, or iodine)—is much more complicated. That is illustrated by a recent study by Masuda *et al.* (1994) of the thermochemistry of dimethyl sulfoxide:

$$C_2H_6SO(l) + (9/2)O_2(g) + 113H_2O(l) \rightarrow 2CO_2(g) + H_2SO_4 \cdot 115H_2O(sln). \quad (25)$$

Recall that the first law of thermodynamics dictates that the initial and final states of the system under study be defined precisely. It requires, for reaction (25), not only high-purity dimethylsulfoxide and oxygen as starting materials, but that the concentration of the H_2SO_4 product be known (it is determined by titration) and, moreover, that the H_2SO_4 be at a uniform concentration throughout the combustion bomb. A diagram of one of the first calorimeters designed specifically to fulfill those criteria is shown in Fig. 3 (Hubbard *et al.*, 1954). The equipment is designed in such a way that the bomb rotates while, at the same time, it is also turned "head over heels," thus ensuring that the water in the moving bomb washes all areas of the interior and that, after the combustion has ended, the H_2SO_4 forms a uniform solution. This diagram also illustrates a typical experimental arrangement with the calorimetric can, thermostatted environment, and thermometer.

Oxygen-bomb calorimetry has also been used successfully in studies of organometallic and inorganic compounds. Examples of such applications include the research of Shaulov *et al.* (1969) on tetramethyl germane:

$$Ge(CH_3)_4(l) + 8O_2(g) \rightarrow GeO_2(cr) + 4CO_2(g) + 6H_2O(l); \quad (26)$$

of Mah (1959) on the oxidation of tungsten dioxide:

$$WO_2(s) + (1/2)O_2(g) \rightarrow WO_3(s); \quad (27)$$

and of King and Armstrong (1964) on the combustion of aluminum carbide:

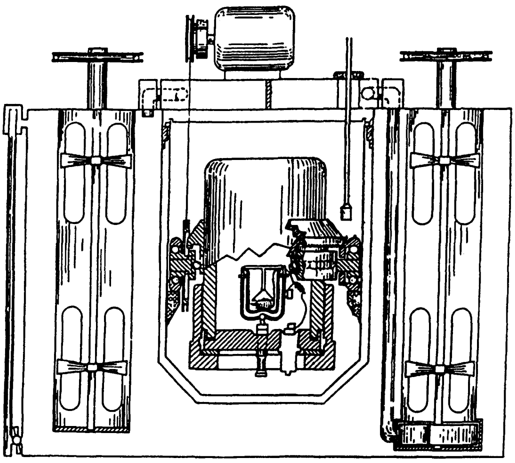

FIG. 3. Rotating-bomb calorimeter for studies of sulfur- and halogen-containing organic compounds. The bomb is similar to that shown in Fig. 2, but it can be rotated and turned head over heels inside the can to bring the aqueous combustion product to a uniform concentration. A chamber filled with stirred water maintained at a constant temperature surrounds the entire assembly.

$$Al_4C_3(s) + 6O_2(g) \rightarrow 2Al_2O_3(s) + 3CO_2(g). \tag{28}$$

Although numerous inorganic substances such as nitrides, carbides, and oxides were found to be responsive to oxygen-bomb calorimetry, it became clear about 35 years ago that many of the emerging technologically important materials could not be studied reliably by this technique. For example, chalcogenides (compounds of a metal with sulfur, selenium, or tellurium) reacted incompletely with high-pressure oxygen, yielding multiple products whose characterization and quantification required heroic efforts in analytical chemistry. In other words, definition of the end states of such systems, as required by the first law of thermodynamics, was very difficult, often nigh impossible. In a nutshell, high-pressure oxygen lacked the requisite oxidizing power, and a more powerful reagent was necessary.

Such considerations led to the invention of fluorine-bomb calorimetry, a detailed description of which was given by Hubbard (1962). Essentially, this technique substitutes fluorine for oxygen as a reagent. That change required a revamping of the traditional oxygen bomb and the incorporation of fluorine-resistant materials of construction such as Monel, lead, gold, and Teflon into the design of the new apparatus.

The great advantage of this method can be illustrated by many examples. Combustion of molybdenum disulfide, MoS_2, with oxygen forms a mixture of molybdenum oxides, sulfur oxides, and, possibly, ternary compounds formed from molybdenum, sulfur, and oxygen, whereas the reaction with fluorine,

$$MoS_2(s) + 9F_2(g) \rightarrow MoF_6(g) + 2SF_6(g), \tag{29}$$

yields well-defined, quantifiable products that require a minimum of analytical characterization. Fluorine-bomb calorimetry has been used successfully to study numerous inorganic substances. Among them are fluorides, oxides, chalcogenides, nitrides, phosphides, silicides, carbides, and borides, as well as individual materials that do not belong in the preceding categories. Developments of the technique over the first 20 years or so were traced by Leonidov and Medvedev (1978), and there is a recent survey of fluorine-bomb calorimetry (O'Hare, 1995).

Although some authors (Kolesov, 1979) have described techniques by means of which vapors of volatile organic liquids are burned in oxygen in a bomb, reactions of gaseous compounds are usually studied by the method of flame calorimetry. Rossini (1931) was a pioneer of such research, and Pilcher and his colleagues at the University of Manchester have also contributed significantly: see Pilcher and Chadwick (1967), for example. Central to this work is the custom-designed burner vessel. It is traditionally made of glass, sits in the calorimeter, and essentially takes the place of the combustion bombs described earlier. It facilitates combination of the reacting gases. A classic example of this procedure is the measurement (Rossini *et al.*, 1931) of the enthalpy of reaction of hydrogen and oxygen:

$$H_2(g) + (1/2)O_2(g) \rightarrow H_2O(l). \qquad (30)$$

Formation of many other simple compounds such as D_2O, CO, SO_2, and HCl has been studied by flame calorimetry.

Just as fluorine-bomb calorimetry was developed as an extension of conventional oxygen calorimetry, fluorine was substituted for oxygen in flame calorimetry for purposes of additional versatility. Because of the aggressive nature of fluorine, Monel, rather than glass, was the main material of construction of the burner. King and Armstrong (1968) used this technique to measure the enthalpies of reaction of $H_2(g)$ with $F_2(g)$, and of $OF_2(g)$ with $H_2(g)$.

5.3 Synthesis Calorimeters

So far, we have described only calorimeters designed for reactions that involve combustion by either gaseous oxygen or fluorine, in which the compound of interest is converted to various products. Some investigators have used an essentially opposite strategy and measured $\Delta_r H_m^0$ for the formation of a substance from its elements,

$$m\mathrm{M}(s) + n\mathrm{N}(s) \rightarrow \mathrm{M}_m\mathrm{N}_n(s), \qquad (31)$$

by means of a synthesis calorimeter.

Such reactions do not, as a rule, occur spontaneously and must usually be initiated by electrical heating or by otherwise encouraging the combination of an intimate mixture of the elements M and N in the calorimeter. The total energy released less the electrical energy input corresponds to $\Delta_r H_m^0$ of reaction (31), in other words, $\Delta_f H_m^0(\mathrm{M}_m\mathrm{N}_n)$.

Among the modern practitioners of this type of calorimetry are Maslov *et al.* (1977), who studied the formation of $MoSi_2$ and other substances, and Vidavskii and Korotkevich (1992), who introduced the pulsed laser as a source of external energy in a determination of $\Delta_f H_m^0$ of the mineral heazlewoodite, Ni_3S_2. Just as in other methods, one must ascertain that only $\mathrm{M}_m\mathrm{N}_n$ is formed; if present, extraneous products and unreacted elements must be corrected for, which oftentimes requires extended analytical procedures and can introduce significant uncertainties in the final $\Delta_f H_m^0(\mathrm{M}_m\mathrm{N}_n)$.

5.4 Calibration of Combustion Calorimeters

Hubbard *et al.* (1956) and other investigators have shown how the corrected temperature change of a calorimeter may be calculated accurately, but that is beyond the scope of this article. The question we wish to address here is how can the measured temperature rise of an experiment be correlated with the energy release that caused it? In other words, how is a calorimeter calibrated?

A few, mostly standardizing, laboratories are equipped to carry out electrical calibrations of combustion calorimeters. Basically, an accurately measured quantity of electrical energy is passed through a resistance heater inserted in the bomb, and the temperature rise of the calorimeter is determined. To calculate the energy equivalent of the calorimetric system, the quantity of electrical energy is divided by the corrected temperature change. Thus, if an input of 12 000 J caused a temperature increase of 1.2 K, the energy equivalent $\epsilon = (12\,000\ \mathrm{J})/(1.2\ \mathrm{K}) = 10\,000\ \mathrm{J \cdot K^{-1}}$. Having established ϵ, one can then know that a temperature rise of 1.23456 K in a combustion experiment, for example, corresponds to an energy release of 12345.6 J.

Most laboratories do not have such facilities and must rely instead on a chemical standard, selected as benzoic acid for combustion calorimetry, by which the unit of electrical energy is essentially transferred from the certifying to the investigator's own laboratory. Thermochemical standard benzoic acid has a certified energy of combustion when reacted with oxygen under specified conditions. Thus, a calorimetrist can calculate precisely how much energy is released by the combustion of a certain mass of the benzoic acid, and the value of ϵ is deduced as if the increase in the temperature of the calorimeter were caused by electrical heating. Thermochemical standard benzoic acid is available from the National Institute of Standards and Technology in the United States (NIST-SRM-39j), from the Russian

Scientific Research Institute of Metrology (sample K-1), and from other sources.

With the calorimeter calibrated, and especially if it is being put into service for the first time, it is important to perform "check" experiments to be sure that no systematic errors are present. Several substances are available whose energies of combustion have been determined by numerous investigators and are, therefore, known with considerable accuracy. Two of those are succinic acid, for combustions of compounds containing carbon, hydrogen, and oxygen, and thianthrene, for combustions of sulfur-containing compounds.

5.5 Solution Calorimeters

Each candidate substance for calorimetric study presents a unique set of challenges to the investigator. Among the most common is the identification of a chemical reaction that yields well-defined products in a reasonable length of time. One might form the impression from the preceding sections that combustion or synthesis calorimetry is applicable to all problems. That is not so; for example, many inorganic compounds do not react completely with high-pressure oxygen or even fluorine, nor can they be readily synthesized directly from the elements. Such substances may, however, dissolve in or combine with aqueous reagents or liquids to form well-defined end products, and, because of that, thermochemical investigations are often carried out in solution calorimeters. A survey of modern solution calorimetry and its applications was published recently (Marsh and O'Hare, 1994).

There are five essential components of a solution calorimeter: the reaction vessel that contains the solvent, a mechanism for introducing the solute into the solvent, a stirrer to promote solution or reaction and to ensure that the product solution is homogeneous, a thermometer, and an electrical calibration heater. We note in passing that almost all modern solution calorimeters, unlike their combustion counterparts, are equipped for electrical calibration.

A typical solution calorimeter is illustrated in Fig. 4. The reaction vessel A is usually constructed of glass for measurements with innocuous solutions; however, aggressive reagents such as aqueous hydrofluoric acid (used in studies of silicates, for example) require platinum, gold, or gold-plated vessels. The electrical heater B and associated equipment are used to calibrate the system. A capsule C, usually constructed of glass or plastic and having thin end windows, contains the substance to be dissolved or reacted, and is held in the tines of the stirrer D. To initiate the reaction, the stirrer is pushed downward, the spike F pierces the windows, the substance is released into the solution, and the reaction begins. The thermometer E measures the temperature of the calorimeter. A metal can G encloses the reaction vessel, and the entire assembly is placed in a thermostatted bath.

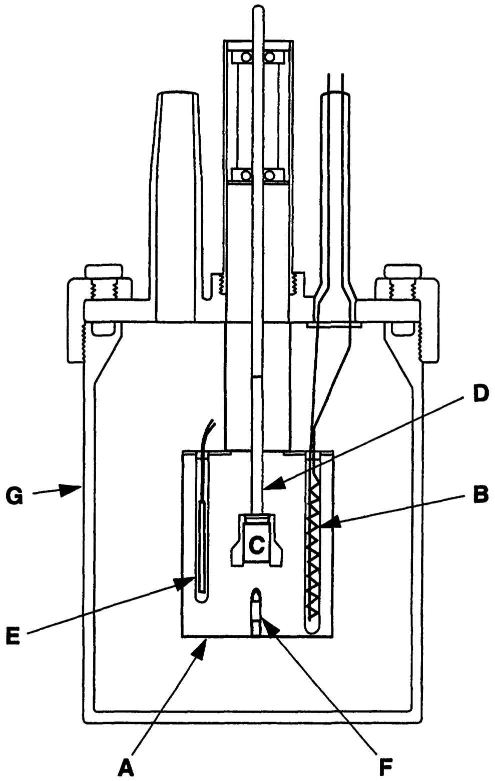

FIG. 4. Diagram of a solution calorimeter. A, reaction vessel filled with solution; B, electrical calibration heater; C, ampoule containing sample to be dissolved or reacted; D, combined stirrer and ampoule holder; E, temperature-measuring device (thermistor, thermometer); F, spike on which ampoule is broken when stirrer is pushed down, thereby introducing sample into solution; G, outer shell of reaction vessel assembly. The entire apparatus shown in this figure sits in a temperature-controlled water bath.

Although the temperature in combustion calorimetric experiments invariably increases (such reactions are called exothermic), decreases in temperature are not uncommon in solution calorimetry. For example, dissolution of crystalline NaCl in water (an endothermic process) would cause the temperature to fall. As in other kinds of calorimetry, careful characterization of the sample under study and the products of reaction are *sine qua non.*

Calorimeters of the kind just described usually are operated at, or within 50 K of, room temperature because vapors of the aqueous solvent at higher temperatures are frequently corrosive. To circumvent that restriction, solution experiments can be, and are, carried out at much higher temperatures, but novel solvents are required. A survey of high-temperature solution calorimetry was given recently by Colinet and Pasturel (1994). They listed a large number of solvents used for different applications, including the liquid metals aluminum, copper, tin, lead, and bismuth. Molten alloys such as $Mn_{0.4}Ni_{0.6}$, and the molten mixed oxides $2PbO \cdot B_2O_3$, $3Na_2O \cdot 4MoO_3$, and $LiBO_2 \cdot NaBO_2$, have also been effective. Colinet and Pasturel (1994) compiled extensive lists of compounds and alloys for which $\Delta_f H_m^0$ has been determined by high-temperature solution calorimetry.

Conventional aqueous solution calorimetry is, as we have mentioned, carried out close to room temperature. Modern advances, however, have led to the construction of equipment that accommodates operating temperatures and pressures much higher than used heretofore. Such apparatus, which revealed unexpected behavior of water as a solvent, among other phenomena, have been described by Smith-Magowan and Wood (1981), Rogers and Pitzer (1981), and Busey *et al.* (1984).

Most of the calorimetric techniques described above consume several grams of substance during a single investigation. Nowadays, however, many compounds of interest (in the pharmaceutical industry, for instance) are extremely expensive, which means that just very small quantities are available for reaction calorimetry (a destructive technique). Accordingly, numerous efforts have been made to design calorimeters capable of accurate measurements on samples with masses of approximately a few milligrams. To reach that goal, the energy equivalent had to be reduced by several factors of 10. A succinct list of about a dozen microbomb calorimeters has been compiled by Månsson (1979).

6. APPLICATIONS OF THERMOCHEMICAL INFORMATION

6.1 Calorimetric Results

Let us suppose that the calorimetric measurements have been completed and all the criteria (purity of substance and reaction, etc.) that we have discussed have been satisfied. How do we proceed from there? We give the following illustrations from the contemporary literature.

6.1.1 Combustion Calorimetry Diogo *et al.* (1993) measured the standard massic energy of combustion $\Delta_c u^0$ of Buckminsterfullerene, C_{60}, with the objective of determining its standard molar enthalpy of formation, that is, the enthalpy of reaction for

$$60C(cr,graphite) \rightarrow C_{60}(cr). \qquad (32)$$

The combustion reaction was shown to proceed entirely as follows:

$$C_{60}(cr) + 60O_2(g) \rightarrow 60CO_2(g), \qquad (33)$$

and the authors found $\Delta_c u^0 = -(35\,924 \pm 16.8)\ J \cdot g^{-1}$, where the quantity after the ± is known as the uncertainty. Although this energy of combustion is the experimental result, it is of limited value *per se* and has to be recast as a molar quantity before further calculations can be performed with it. The molar mass of C_{60} is $720.66\ g \cdot mol^{-1}$, and the molar energy of combustion $\Delta_c U_m^0 = (720.66\ g \cdot mol^{-1})(-35\,923.6 \pm 16.8)\ J \cdot g^{-1} = -(25\,888\,702 \pm 12\,107)\ J \cdot mol^{-1}$ or, more compactly, $-(25\,888.7 \pm 12.1)\ kJ \cdot mol^{-1}$. Next, the molar energy of reaction $\Delta_c U_m^0$ is changed to the molar enthalpy of reaction $\Delta_c H_m^0$ on the basis of the following equation:

$$\Delta_c H_m^0 = \Delta_c U_m^0 + \Delta\nu^g RT, \qquad (34)$$

where $\Delta\nu^g$ represents $\nu(CO_2,g) - \nu(O_2,g)$, ν

denoting the stoichiometric number. It so happens in this example that $\nu(CO_2) = \nu(O_2) = 60$; thus, $\Delta\nu^g = 0$ and, from Eq. (34), $\Delta_c H_m^0 = -(25\,888.7 \pm 12.1)\ kJ\cdot mol^{-1}$. One other piece of information is required: $\Delta_f H_m^0(CO_2,g)$ is found in the critical tabulations mentioned earlier to be $-(393.51 \pm 0.13)\ kJ\cdot mol^{-1}$. Thus, the following are available:

$$C_{60}(cr) + 60O_2(g) \rightarrow 60CO_2(g),\quad \Delta_c H_m^0 = -(25\,888.7 \pm 12.1)\ kJ\cdot mol^{-1}; \tag{35}$$

$$60C(cr,graphite) + 60O_2(g) \rightarrow 60CO_2(g),\quad \Delta_c H_m^0 = 60 \times [-(393.51 \pm 0.13)]\ kJ\cdot mol^{-1}. \tag{36}$$

By appropriate combination of Eqs. (35) and (36), one obtains

$$60C(cr,graphite) \rightarrow C_{60}(cr),\quad \Delta_f H_m^0 = (2278.1 \pm 14.4)\ kJ\cdot mol^{-1}. \tag{37}$$

Note that we have introduced the symbol $\Delta_c H_m^0$ in several places; the subscript c indicates that the thermodynamic quantity refers to a combustion reaction; in other words, $\Delta_c H_m^0$ is another special case of $\Delta_r H_m^0$. This calculation illustrates an application of Hess's law.

Fluorine-combustion calorimetry was discussed in Sec. 5.2. That technique was used (O'Hare *et al.*, 1992) to determine $\Delta_c u^0 = -(37\,943 \pm 8)\ J\cdot g^{-1}$ at $T = 298.15$ K for the following reaction:

$$S(cr,rhombohedral) + 3F_2(g) \rightarrow SF_6(g), \tag{38}$$

the rhombohedral form being the standard reference state of sulfur. Taking the molar mass of S as 32.066 $g\cdot mol^{-1}$, one calculates $\Delta_c U_m^0 = [-(37\,943 \pm 8)\ J\cdot g^{-1}] \times (32.066\ g\cdot mol^{-1}) = -(1\,216\,680 \pm 257)\ J\cdot mol^{-1} = -(1216.7 \pm 0.3)\ kJ\cdot mol^{-1}$. For reaction (38), $\Delta\nu^g = 1 - 3 = -2$; thus $\Delta\nu^g RT = -2 \times (8.314\,51\ J\cdot K^{-1}\cdot mol^{-1}) \times (298.15\ K) = -5\ kJ\cdot mol^{-1}$. Therefore, $\Delta_c H_m^0 = -(1216.7 \pm 0.3)\ kJ\cdot mol^{-1} - 5\ kJ\cdot mol^{-1} = -(1221.7 \pm 0.3)\ kJ\cdot mol^{-1}$. Since Eq. (38) also represents the formation of $SF_6(g)$ in its standard state from the elements in their standard states, $\Delta_f H_m^0(SF_6,g,298.15\ K) = -(1221.7 \pm 0.3)\ kJ\cdot mol^{-1}$.

6.1.2 Solution Calorimetry

The choice of a calorimetric method, where more than one is accessible, or seems to be, is usually dictated by the straightforwardness of the chemical reactions available. Thus, for example, a determination of $\Delta_f H_m^0(BaO,cr)$ appears, at first sight, to be a candidate for combustion calorimetry in oxygen. However, barium metal, Ba, could well form BaO_2 as well as BaO, thus presenting the kind of difficult challenge in analytical chemistry to which we alluded earlier. Furthermore, low-melting metals such as Ba tend not to react cleanly in combustion calorimetry. They melt as the oxidation proceeds, splatter over the interior surface of the bomb, and thereby cause premature quenching of the reaction. Although the analytical problems in this example would probably not be insuperable, solution calorimetry presents a much more attractive alternative, because BaO dissolves readily in dilute mineral acids.

Cordfunke *et al.* (1990) took advantage of that behavior. They performed two sets of measurements: of the enthalpies of reaction of Ba and of BaO with dilute (≈1 $mol\cdot dm^{-3}$) hydrochloric acid, HCl(aq):

$$Ba(cr) + 2HCl(aq) \rightarrow BaCl_2(aq) + H_2(g),\quad \Delta_r H_m^0 = -(535.29 \pm 1.72)\ kJ\cdot mol^{-1}; \tag{39}$$

$$BaO(cr) + 2HCl(aq) \rightarrow BaCl_2(aq) + H_2O(aq),\quad \Delta_r H_m^0 = -(273.08 \pm 0.24)\ kJ\cdot mol^{-1}. \tag{40}$$

Subtracting Eq. (40) from Eq. (39) yields

$$Ba(cr) + H_2O(l) \rightarrow BaO(cr) + H_2(g),\quad \Delta_r H_m^0 = -(262.21 \pm 1.74)\ kJ\cdot mol^{-1}. \tag{41}$$

Next, Eq. (41) and the following are combined:

$$H_2(g) + (1/2)O_2(g) \rightarrow H_2O(l),\quad \Delta_r H_m^0 = -(285.83 \pm 0.04)\ kJ\cdot mol^{-1}, \tag{42}$$

to yield

$$Ba(cr) + (1/2)O_2(g) = BaO(cr),\quad \Delta_r H_m^0 = -(548.04 \pm 1.74)\ kJ\cdot mol^{-1}. \tag{43}$$

The $\Delta_r H_m^0$ for Eq. (42), identical to $\Delta_f H_m^0(\mathrm{H_2O,l})$, was obtained from standard reference sources cited earlier. Note, also, that $\Delta_r H_m^0$ for reaction (43) is identical to $\Delta_f H_m^0(\mathrm{BaO,cr,298.15\ K})$.

6.1.3 A "Third-Law" Calculation of $\Delta_f H_m^0$ Binary silicides are usually thought of as refractory substances, but at sufficiently high temperatures, they are known to break down. A thermodynamic study of one such material, pentatungsten trisilicide ($\mathrm{W_5Si_3}$), has been described by Kematick (1993). Heated to temperatures in the vicinity of 2000 K, it decomposes as follows:

$$\mathrm{W_5Si_3(s) \rightarrow 5W(s) + 3Si(g)}. \qquad (44)$$

Kematick determined the equilibrium vapor pressure of Si(g) as a function of temperature for the preceding reaction, and from it, as outlined in Eqs. (11) and (12), deduced the relation

$$\Delta_r G_m^0/(\mathrm{kJ\cdot mol^{-1}}) = 1447.2 - 0.4077(T/\mathrm{K}). \qquad (45)$$

At the mean experimental temperature, therefore, $\Delta_r G_m^0(1900\ \mathrm{K}) = 672.6\ \mathrm{kJ\cdot mol^{-1}}$. As outlined in Eq. (16), the following relation holds for reaction (44):

$$\begin{aligned}\Delta_r H_m^0(298.15\ \mathrm{K}) = {}& \Delta_r G_m^0(1900\ \mathrm{K}) \\ &+ 1.9[5\Phi_m^0(\mathrm{W,s,1900\ K}) \\ &+ 3\Phi_m^0(\mathrm{Si,g,1900\ K}) \\ &- \Phi_m^0(\mathrm{W_5Si_3,s,1900\ K})\}.\end{aligned} \qquad (46)$$

We take, from compilations mentioned earlier, values of 58.99 $\mathrm{J\cdot K^{-1}\cdot mol^{-1}}$ for $\Phi_m^0(\mathrm{W})$; 189.52 $\mathrm{J\cdot K^{-1}\cdot mol^{-1}}$ for $\Phi_m^0(\mathrm{Si,g})$; and 431.6 $\mathrm{J\cdot K^{-1}\cdot mol^{-1}}$ for $\Phi_m^0(\mathrm{W_5Si_3})$. Therefore, $\Delta_r H_m^0 = (1493 \pm 32)\ \mathrm{kJ\cdot mol^{-1}}$ for reaction (44) at $T = 298.15$ K, or

$$\mathrm{5W(s) + 3Si(g) \rightarrow W_5Si_3(s)}, \quad \Delta_r H_m^0 = -(1493 \pm 32)\ \mathrm{kJ\cdot mol^{-1}}. \qquad (47)$$

Standard compilations give for the sublimation reaction at $T = 298.15$ K

$$\mathrm{3Si(s) \rightarrow 3Si(g)}, \quad \Delta_r H_m^0 = 3 \times (450 \pm 8)\ \mathrm{kJ\cdot mol^{-1}}. \qquad (48)$$

By combining reactions (47) and (48), one obtains

$$\mathrm{5W(s) + 3Si(s) \rightarrow W_5Si_3(s)}, \quad \Delta_r H_m^0 = -(143 \pm 40)\ \mathrm{kJ\cdot mol^{-1}}. \qquad (49)$$

In other words, $\Delta_f H_m^0(\mathrm{W_5Si_3,s,298.15\ K}) = -(143 \pm 40)\ \mathrm{kJ\cdot mol^{-1}}$. This example shows how thermodynamic properties at room temperature or, indeed, any other temperature can be deduced from high-temperature studies.

GLOSSARY

Chemical Equilibrium: The reaction A + B $\rightleftarrows$ C + D is said to be in equilibrium when the rate of the forward reaction between A and B is equal to the rate of the reverse reaction between C and D.

Electrochemical Cell: There are two types of electrochemical cell. Spontaneous chemical reactions occur within the galvanic cell, while in the electrolytic cell, reactions result from an external potential difference.

Enthalpy: The enthalpy H of a system is defined as $H = U + pV$, where U is the energy (or internal energy) and pV denotes the pressure–volume work energy.

Entropy: The difference in entropy S between the states a and b of a system is given by $\Delta S = \int_a^b dq_{\mathrm{rev}}/T$. Thus, to compute ΔS, imagine moving the system from state a to state b via a reversible path, and for each infinitesimal step, computing dq/T, the sum of which is ΔS.

Equilibrium Constant: In a reaction $a\mathrm{A(g)} + b\mathrm{B(g)} \rightarrow c\mathrm{C(g)} + d\mathrm{D(g)}$ in which all the gases are at thermodynamic equilibrium, the equilibrium constant is given by $K = p(\mathrm{C})^c p(\mathrm{D})^d / p(\mathrm{A})^a p(\mathrm{B})^b$.

Fugacity: The molar Gibbs energy of a real gas G_m is related to the standard Gibbs energy G_m^0 by $G_m = G_m^0 + RT\ln(f/p^0)$, where the fugacity is denoted by f and the standard pressure by p^0. As the pressure of a system approaches zero, the gas approaches ideal behavior, and the fugacity approaches the pressure. Thus, in the limit of zero pressure, $p = f$. Units of f and p are the same.

Molar Quantity: A thermodynamic quantity for 1 mol of substance.

Pressure–Volume Work Energy: Work performed in the expansion or compression of a substance, given by the product of the pressure and the change in volume.

Works Cited

Busey, R. H., Holmes, H. F., Mesmer, R. E. (1984), *J. Chem. Thermodyn.* **16,** 343–372.

Chase, M. W., Jr., Davies, C. A., Downey, J. R., Jr., Frurip, D. J., McDonald, R. A., Syverud, A. N. (1985), *J. Phys. Chem. Ref. Data* **14,** Suppl. No. 1.

Colinet, C., Pasturel, A. (1994), in: K. N. Marsh, P. A. G. O'Hare (Eds.), *Solution Calorimetry,* London: Blackwell, Chap. 6.

Cordfunke, E. H. P., Konings, R. J. M., Ouweltjes, W. (1990), *J. Chem. Thermodyn.* **22,** 991–996.

Cox, J. D., Pilcher, G. (1970), *Thermochemistry of Organic and Organometallic Compounds,* New York: Academic.

Cox, J. D., Wagman, D. D., Medvedev, V. A. (1989), *CODATA Key Values for Thermodynamics,* New York: Hemisphere.

Diogo, H. P., Minas da Piedade, M. E., Dennis, T. J. S., Hare, J. P., Kroto, H. W., Taylor, R., Walton, D. R. M. (1993), *J. Chem. Soc. Faraday Trans.* **89,** 3541–3544.

Domalski, E. S. (1979), *Experimental Chemical Thermodynamics,* Vol. 1, New York: Pergamon, Chap. 18, pp. 401–428.

Frenkel, M., Kabo, G. J., Marsh, K. N., Roganov, G. N., Wilhoit, R. C. (1994), *Thermodynamics of Organic Compounds in the Gas State,* Vols. I and II, College Station, TX: TRC.

Gurvich, L. V., *et al.* (1978 *et seq.*), *Termodinamicheskie Svoistva Individualnikh Veshchestv* (six volumes), Moscow: Nauka.

Gurvich, L. V., Veyts, I. V., Alcock, C. B. (1989), *Thermodynamic Properties of Individual Substances,* Vol. 1 (two parts), New York: Hemisphere.

Gurvich, L. V., Veyts, I. V., Alcock, C. B. (1991), *Thermodynamic Properties of Individual Substances,* Vol. 2 (two parts), New York: Hemisphere.

Gurvich, L. V., Iorish, V. S., Chekhovskoi, D. V., Yungman, V. S. (1993), *IVTANTHERMO—A Thermodynamic Database and Software System for the Personal Computer,* NIST Special Database 5, Boca Raton, FL: CRC.

Gurvich, L. V., Veyts, I. V., Alcock, C. B. (1994), *Thermodynamic Properties of Individual Substances,* Vol. 3 (two parts), Boca Raton, FL: CRC.

Hubbard, W. N. (1962), *Experimental Thermochemistry,* Vol. II, New York: Wiley, Chap. 6, pp. 95–127.

Hubbard, W. N., Katz, C., Waddington, G. (1954), *J. Phys. Chem.* **58,** 142–152.

Hubbard, W. N., Scott, D. W., Waddington, G. (1956), *Experimental Thermochemistry,* New York: Interscience, Chap. 5, pp. 75–128.

Kematick, R. J. (1993), *J. Alloys Compounds* **202,** 225–229.

King, R. C., Armstrong, G. T. (1964), *J. Res. Natl. Bur. Stand. (U.S.)* **68A,** 661–668.

King, R. C., Armstrong, G. T. (1968), *J. Res. Natl. Bur. Stand. (U.S.)* **72A,** 113–131.

Kolesov, V. P. (1979), *Experimental Chemical Thermodynamics,* Vol. 1, New York: Pergamon, Chap. 13, pp. 291–306.

Leonidov, V. Ya., Medvedev, V. A. (1978), *Ftornaya Kalorimetriya,* Moscow: Nauka.

Mah, A. D. (1959), *J. Am. Chem. Soc.* **81,** 1582–1583.

Månsson, M. (1979), in: S. Sunner, M. Månsson (Eds.), *Experimental Chemical Thermodynamics,* Vol. 1, New York: Pergamon, Chap. 17.

Marsh, K. N., O'Hare, P. A. G. (Eds.) (1994), *Solution Calorimetry,* London: Blackwell.

Maslov, V. M., Neganov, A. S., Borovinskaya, I. P., Merzhanov, A. G. (1977), *Fiz. Goreniya Vzryva* **14,** 73–82.

Masuda, N., Nagano, Y., Sakiyama, M. (1994), *J. Chem. Thermodyn.* **26,** 971–975.

O'Hare, P. A. G. (1995), *Thermochim. Acta* **267,** 1–13.

O'Hare, P. A. G., Susman, S., Volin, K. J., Rowland, S. C. (1992), *J. Chem. Thermodyn.* **24,** 1009–1017.

Pedley, J. B. (1994), *Thermochemical Data and Structures of Organic Compounds,* College Station, TX: TRC.

Pilcher, G., Chadwick, J. D. M. (1967), *Trans. Faraday Soc.* **63,** 2357–2361.

Rogers, P. S. Z., Pitzer, K. S. (1981), *J. Phys. Chem.* **85,** 2886–2895.

Rossini, F. D. (1931), *J. Res. Natl. Bur. Stand. (U.S.)* **6,** 1–35.

Shaulov, Yu. Kh., Fedorov, A. K., Genchel', V. G. (1969), *Zh. Fiz. Khim.* **43,** 1336–1337.

Smith-Magowan, D., Wood, R. H. (1981), *J. Chem. Thermodyn.* **13,** 1047–1073.

Vidavskii, L. M., Korotkevich, I. I. (1992), *J. Chem. Thermodyn.* **24,** 577–584.

Further Reading

Brown, H. D. (Ed.) (1969), *Biochemical Microcalorimetry,* New York: Academic.

Marsh, K. N., O'Hare, P. A. G. (Eds.) (1994), *Solution Calorimetry,* London: Blackwell.

Médard, L., Tachoire, H. (1994), *Histoire de la*

Thermochimie, Aix-en-Provence: Université de Provence.

Kubaschewski, O., Alcock, C. B., Spencer, P. J. (1993), *Materials Thermochemistry,* New York: Pergamon.

Pitzer, K. S. (1995), *Thermodynamics,* New York: McGraw-Hill.

Rossini, F. D. (Ed.) (1956), *Experimental Thermochemistry,* New York: Interscience.

Skinner, H. A. (Ed.) (1962), *Experimental Thermochemistry,* Vol. II, New York: Interscience.

Sunner, S., Månsson, M. (Eds.) (1979), *Experimental Chemical Thermodynamics,* Vol. 1, *Combustion Calorimetry,* New York: Pergamon.

THERMODYNAMICS, EQUILIBRIUM

B. Widom, *Department of Chemistry, Baker Laboratory, Cornell University, Ithaca, New York, U.S.A.*

	Introduction	281
1.	**The First Law**	281
1.1	Temperature	281
1.2	Work	282
1.3	Energy and Heat	283
2.	**The Second Law**	284
2.1	Spontaneous Processes	284
2.2	Absolute Temperature	285
2.3	Entropy	286
2.4	Entropy of an Isolated System	288
2.5	$S = S(U,V,M_1,M_2,\ldots)$	289
3.	**Free Energy**	292
3.1	Legendre Transforms; Free Energies *F* and *G*	292
3.2	Equilibrium and Convexity	292
3.3	Potentials, Fields, and Densities	293
3.4	Further Identities	295
4.	**Phase Equilibrium**	297
4.1	Phase Rule	297
4.2	Phase Diagrams	297
4.3	Clapeyron Equations	302
5.	**Dilute Systems**	304
6.	**The Third Law**	307
	Acknowledgments	309
	Glossary	309
	Works Cited	309
	Further Reading	309

INTRODUCTION

Thermodynamics is the science of heat: its transformations and the relations among the various physical quantities ("thermodynamic functions") that arise in its definition and study. It is built upon two laws of nature—the first and second laws of thermodynamics—both of universal validity so far as is now known. Additional properties of the functions that arise in the second law, beyond those properties implied by that law itself, are also recognized and are often referred to as the "third law," or, as in this article, as the "Nernst heat theorem." Sometimes one speaks also of the "zeroth" law, which is a recognition of the existence of temperature as a measure of "degree of hotness." It is an important concept but one that will here be taken as primitive—not requiring definition. Our account of thermodynamics begins with it.

1. THE FIRST LAW

1.1 Temperature

We recall that there are elementary devices—thermometers, such as the familiar mercury-in-glass thermometer—that measure degree of hotness, or temperature, and express that measure on an empirical temperature scale such as the centigrade (Celsius) or Fahrenheit scale. When two bodies of different temperatures are put in intimate thermal contact, the temperature of the warmer body falls and that of the cooler body rises until the temperatures become equal, and then they no longer change. By analogy with the flow of water between two communicating vessels, equalizing the height of the water in the two and then ceasing, one says that "heat," thought of as a caloric fluid, flows from a warmer body to a cooler one with which it is in contact, the flow ceasing only when the respective temperatures have become equal. But that is just a colloquialism; there is no such fluid. One of the accomplishments of thermodynamics is to have given an objective definition of the heat that flows from warmer to cooler bodies and causes their temperatures to become equal.

A *system* (i.e., any object of interest, everything else constituting the *surroundings* of that system) may be thermally insulated from its surroundings, so as to have no tendency to reach temperature equality with

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

them on practical time scales. Then "no heat flows" into or out of the system. Any changes that occur within the system or any processes to which the system is subject under such conditions are said to be *adiabatic.*

1.2 Work

One (i.e., the surroundings) may do work on a system or a system may do work on its surroundings. Work may be of many kinds: it may be pressure–volume work, in which a volume changes against an opposing pressure, or stress–strain work, or mass × the difference of gravitational potential between two heights, or electric field × electric polarization, or magnetic field × magnetization, etc. Of these, pressure–volume (p-V) work is representative and is what is mostly referred to here for illustrating general principles.

We adopt the convention of saying that w is the work done on a system by its surroundings and that at the same time the work done on the surroundings by the system is $-w$; so that if $w > 0$, work is in fact done on the system, while if $w < 0$, work is in fact done by the system.

In general when a system undergoes any process, as when it does work or has work done upon it or when heat flows into or out of it, its state, or condition, changes. There are conceptually important processes that are said to be *reversible.* In such a process—which is an idealization that may be closely approached in practice but not rigorously achieved—if at any moment the process were interrupted and the system were thermally and mechanically isolated from its surroundings, no further changes would occur in the system: it would be in a state of equilibrium. Thus, in undergoing a reversible process a system passes through states of equilibrium and is at every moment in such a state. It is because in undergoing reversible processes a system would be always in a state of equilibrium that such idealized processes play an important conceptual role in equilibrium thermodynamics.

(Sometimes one uses the term "quasistatic" to denote a process in which a system passes through a sequence of equilibrium states and reserves the word "reversible" for a process that can be reversed so as to return both the system and its surroundings to their initial states, but in this article we are not so formal and simply use "reversible" to mean either or both as the occasion requires.)

The work done on or by a system in a reversible process may be represented graphically. When an isotropic fluid undergoes a reversible expansion or compression, for example, its pressure p is well defined at each moment, and no distinction need be made between the internal and external pressure, so that the variation of p with V in the process may be represented by a curve in the p-V plane, as illustrated in Fig. 1. The work $-w$ done by the system in expanding from the volume V_A to the volume V_B is

$$-w = \int_{V_A}^{V_B} p dV \tag{1}$$

and is therefore the area under the curve, indicated by the shading in the figure.

This work $-w$ is the sum (integral) of the infinitesimal increments of work done during all the infinitesimal volume changes dV that occur between V_A and V_B. This infinitesimal element of reversible work is

$$đw_{\text{rev}} = -pdV; \tag{2}$$

the reason for the line drawn through the stem of the d in $đw_{\text{rev}}$ will be apparent presently.

For any process, even a reversible one, the work w that is done depends not only on the

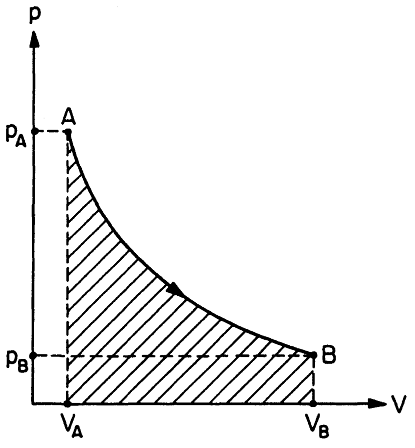

FIG. 1. Expansion from state A to state B, represented by a curve in the pressure (p)–volume (V) plane.

initial and final states of the system but also on the details of the process. For example, the same change from the system state A to state B pictured in Fig. 1 could also be brought about by an alternative reversible process represented by a different curve connecting A and B, the area under which could differ from that in Fig. 1.

A process in which a system ends in the same state in which it started is termed *cyclic.* A reversible cyclic process is represented by a closed curve in the p-V plane, as illustrated in Fig. 2. If the system traverses the cycle clockwise, the work *it does* along the upper part of its path is the area under that part, while the work *done on it* along the lower part is the area under that part; so the net work $-w$ done by the system is the area enclosed by the cycle:

$$-w = \oint p\,dV = \text{area enclosed}. \tag{3}$$

If the cycle is traversed counterclockwise, that area is the net work w done *on* the system. These relations apply even for paths of more complicated shape than that in Fig. 2, as long as they are closed curves.

1.3 Energy and Heat

An *adiabatic* process, one resulting in an adiabatic change in the state of a system, was defined at the end of Sec. 1.1. In Sec. 1.2 it was remarked that the work w done on a system depends, in general, not merely on the initial and final states of the system in the process the system undergoes but also on the details of that process. The *first law of thermodynamics* is the universally verified observation that the work w_{ad} done on a system in an adiabatic process does indeed depend *only* on the initial and final states of the system and is otherwise independent of the details of the process.

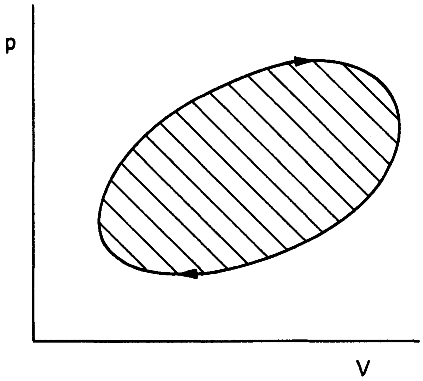

FIG. 2. A reversible cyclic process in which the cycle is traversed clockwise. The work $-w_{\mathrm{rev}}$ done by the system is the area enclosed by the curve, shown shaded.

Thus, there is a function of the system's state—we call it the system's *energy* and symbolize it by U—such that

$$w_{\mathrm{ad}} = U_{\mathrm{B}} - U_{\mathrm{A}} = \Delta U, \tag{4}$$

where U_{B} and U_{A} are the energies of the system in its final state B and initial state A, respectively. The symbol Δ means "change in"; ΔU is the change in the system's energy when the system undergoes an adiabatic process in which work w_{ad} is done on it (or when it does work $-w_{\mathrm{ad}}$). The same change in state, A → B, with, therefore, the same change ΔU in the system's energy, may be brought about by processes that are not adiabatic. The work w done on the system is then not in general equal to ΔU. The difference between w and ΔU is now the *definition* of the *heat* q absorbed by the system during the process:

$$\Delta U = q + w; \tag{5}$$

i.e., q is the difference between the work that *would* have been done on the system in changing its state from A to B *had* that change been brought about by an adiabatic process, and the work w that actually *was* done. If the process is adiabatic, then $w = w_{\mathrm{ad}}$ and $q = 0$. When $q > 0$, one says that the system has absorbed heat q, or that there has been a flow of heat q into the system from its surroundings; when $q < 0$, one says that heat q was evolved, or flowed from the system to the surroundings. With this view, Eq. (4) is a law of nature—the first law of thermodynamics—while Eq. (5) is the definition of heat.

Since w depends not only on the initial and final states of the system but also on the details of the process in which the system's state changes, while ΔU depends only on the initial and final states and not on the details of the process, q, too, must depend on the details of the process. The net work w in a process is the sum (integral) of the infinitesimal increments of work $\text{đ}w$ in each infinitesimal part of the process, where, as in Eq. (2), we put a line through the stem of the d in $\text{đ}w$; and likewise the net heat absorbed, q, is the sum of the infinitesimal increments of absorbed heat, $\text{đ}q$:

$$w = \int_A^B \text{đ}w, \qquad q = \int_A^B \text{đ}q. \tag{6}$$

The symbols đ indicate *inexact* differentials: the integrals depend not only on the initial and final states A and B but also on the details of the process. That, we saw, is true even for reversible processes, for which $\text{đ}w$ is given by Eq. (2): pdV is an *inexact* differential in the two variables p and V. This is to be contrasted with the *exact* differential dU, which is such that

$$\Delta U = U_B - U_A = \int_A^B dU \tag{7}$$

independently of the details of the path between the initial and final states A and B. The inexactnesses in $\text{đ}w$ and $\text{đ}q$ cancel in the combination

$$dU = \text{đ}w + \text{đ}q. \tag{8}$$

When the only work considered is p-V work, if the system's volume is fixed, the system is essentially mechanically isolated from its surroundings, and $w = 0$ in any process it then undergoes. Thus,

$$\Delta U = q \quad \text{(constant volume).} \tag{9}$$

We define another function of the state of the system—the *enthalpy* H—by

$$H = U + pV. \tag{10}$$

If a system is confined under a fixed pressure p then $w = -p\Delta V = -\Delta(pV)$, and so, from Eqs. (5) and (10),

$$\Delta H = q \quad \text{(constant pressure).} \tag{11}$$

2. THE SECOND LAW

2.1 Spontaneous Processes

One recognizes many processes in nature that occur spontaneously, i.e., without the intervention of an outside agency, so that they would occur even in a system that was thermally and mechanically isolated from its surroundings. One example is the spontaneous flow of heat from a warmer to a cooler body already referred to. Another is the diffusion of a substance from a region in which it is present at high concentration to a region in which its concentration is lower. Any such process is characterized by an inexorable drive to a condition of equilibrium, when the process then ceases.

In each of the processes carried out as described, no work is done by the system on its surroundings, but the same change in the system's state may be made to occur indirectly in such a way as to be a source of work. In the first example referred to above, let the two bodies ("heat reservoirs") be kept apart so that heat does not flow directly from one to the other. Then immerse in the warmer reservoir a system ("engine") consisting of a gas in a cylinder fitted with a movable piston and initially at a temperature between that of the two reservoirs (Fig. 3). The gas is allowed to expand as it warms, and the moving piston, mechanically coupled to the surroundings, may do work. The cylinder

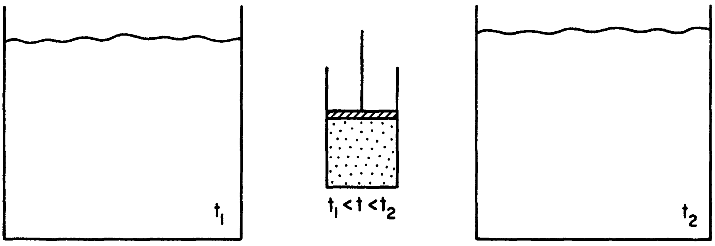

FIG. 3. Heat reservoirs at temperatures $t_1 < t_2$ and a gas confined in a cylinder at an intermediate temperature t.

is then removed from the warmer reservoir and immersed in the cooler, where the gas is allowed to contract, while the now receding piston (the sense of the coupling having in the meantime been switched) again does work. Once the cylinder and gas have been cooled to their initial temperature and the piston is back to its original position, the engine is removed from the cooler reservoir.

There has been no net change in the state of the engine; the only net changes have been the absorption of heat from the warmer reservoir, thus lowering its temperature, the conversion of some of that heat into work on the environment, and the discharge of heat into the cooler reservoir, thus warming it. The spontaneous tendency of heat to flow from high to low temperatures was taken advantage of to do work, but it was at the expense of the respective temperatures' being now closer to equality, i.e., of a spontaneous process's having run part way down.

That work can be done only at the expense of spontaneous processes' running down is a universally observed principle: the second law of thermodynamics.

2.2 Absolute Temperature

In the example in Sec. 2.1 in which the spontaneous flow of heat from high to low temperature was made to do work, it was essential, in restoring the engine to its initial state in the final stage, that it be cooled at the lower temperature. According to the second law it would have been impossible for the engine to have undergone such a cyclic process, the only result of which was the absorption of heat from the higher-temperature reservoir and its conversion to work, without some heat's having also been conducted to the lower-temperature reservoir. From this observation alone one may follow a classic route to the concepts of entropy and absolute temperature.

The argument, which may be found in standard references (see Further Reading), typically proceeds via the consideration of a particular reversible cyclic process called a *Carnot cycle,* represented in the *p*-*V* plane in Fig. 4. Adiabatic steps BC and DA alternate with isothermal steps AB at the constant temperature t_1 and CD at the constant temperature t_2. No heat is absorbed—$q = 0$—in the adiabatic steps. The heats absorbed in the isothermal steps are q_{t_1} on AB and q_{t_2} on CD, and are opposite in sign. The figure (although not the principle to be derived) assumes a system with positive coefficient of thermal expansion, in which case t_1 is a higher temperature than t_2, and in traversing the cycle clockwise, as in the figure, the system absorbs heat $q_{t_1} > 0$ at the higher temperature and evolves heat $-q_{t_2} > 0$ at the lower temperature. In traversing the cycle once in the indicated sense, the system does positive work $-w$ given by

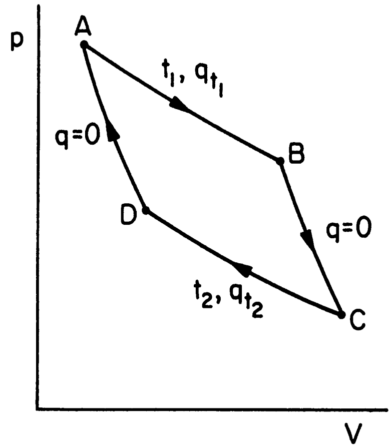

FIG. 4. A reversible Carnot cycle represented in the pressure–volume plane. Adiabatic ($q = 0$) steps BC and DA alternate with isothermal steps AB and CD at the constant temperatures t_1 and t_2, respectively. The heats q_{t_1} and q_{t_2} absorbed in the isothermal steps are opposite in sign.

$$-w = q_{t_1} + q_{t_2} = q_{t_1} - (-q_{t_2}). \tag{12}$$

In classical thermal engineering one defines the *efficiency e* of this Carnot cycle as the fraction of the heat absorbed at the higher temperature that is converted to work:

$$e = \frac{-w}{q_{t_1}} = \frac{q_{t_1} - (-q_{t_2})}{q_{t_1}} = 1 - (-q_{t_2})/q_{t_1}. \tag{13}$$

This is always less than 1 because some heat $-q_{t_2} > 0$ must be discharged at the lower temperature, and it is always greater than 0 because $-w$ and q_{t_1} are both positive. The argument then proceeds by showing that the efficiency *e* depends *only* on the two temperatures t_1 and t_2; two systems undergoing reversible Carnot cycles between the same two temperatures t_1 and t_2 must have the same efficiency. The main idea is that if it were not so, one could run the more efficient process clockwise and the less efficient one counterclockwise in such a way as to absorb heat at the higher temperature and convert all of it to work, with no net heat discharged at the lower temperature—which would violate the second law.

Let t_0 now be an arbitrary fixed reference temperature. Then by the principle just discussed, and from (13), when t_0 and any other temperature t are the two temperatures in a reversible Carnot cycle, $|q_t/q_{t0}|$ depends only on t, and so does any fixed constant multiple c of $|q_t/q_{t0}|$. Any such multiple, with arbitrary choices of t_0 and c, may be defined to be the *absolute* temperature $T(t)$ at the (empirical) temperature t:

$$T(t) = c|q_t/q_{t0}|. \tag{14}$$

For fixed t_0 and c this is then a function of t alone, independently of the system undergoing the cycle. The arbitrary constant $c = T(t_0)$ determines the sign of T and the size of the "degree" on this temperature scale.

Taking t_0 to be the triple point of water that has the isotopic composition of ocean water, and $c = 273.16$, defines the Kelvin scale of absolute temperature, the size of the degree on which is the same as that on the empirical Celsius scale. With that same t_0, the choice $c = (9/5) \times 273.16 = 491.69$ defines the Rankine scale of absolute temperature, the size of the degree on which is the same as that on the empirical Fahrenheit scale. Temperatures T_K and T_R on the Kelvin and Rankine scales are then related to temperatures $t_{°C}$ and $t_{°F}$ on the Celsius and Fahrenheit scales by the empirical rules

$$\begin{aligned} T_K &= t_{°C} + 273.15 \\ T_R &= t_{°F} + 459.68 \end{aligned} \tag{15}$$

(since the triple point of water lies 0.0075 °C or 0.0135 °F above the 1-atm freezing point).

By tradition c, hence T, is chosen positive, although it need not have been. Throughout this account the traditional choice is made without further comment. Some inequalities to be derived later would have had their sense reversed if absolute temperature had been chosen to be negative instead of positive.

2.3 Entropy

If T_1 and T_2 are the absolute temperatures in the two isothermal steps of a reversible Carnot cycle, and if q_{T_1} and q_{T_2} are the heats absorbed in those two steps, one of these being positive and one negative, then from the definition of the absolute temperature in Eq. (14),

$$q_{T_1}/T_1 + q_{T_2}/T_2 = 0. \tag{16}$$

One next considers a general reversible cycle in the p-V plane such as that in Fig. 2, now in Fig. 5. The plane may be imagined filled with infinitely closely spaced adiabats, such that for each point C on the cycle there is another point D on it accessible from C by a reversible adiabatic process, and likewise for the pair of points A and E on the cycle, close to C and D. These two adiabats may then be connected by the short isothermal segments DF and BC, as shown in the figure. BCDF is then a reversible Carnot cycle, which we allow the system to traverse in the same sense in which it traverses its original cycle.

The classical argument then proceeds by showing that in the limit in which the grid of adiabats in the p-V plane is made infinitely fine, the two ratios q_{AC}/q_{BC} and q_{DE}/q_{DF} both approach 1. But, since q_{BC} and q_{DF} are the heats absorbed in the two isothermal

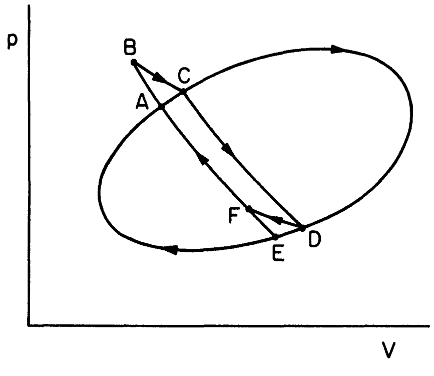

FIG. 5. A general reversible cycle on which are superimposed two closely spaced adiabats CD and EB connected by segments of isotherms BC and DF, so that BCDF is a Carnot cycle.

steps of a reversible Carnot cycle, they satisfy (16). Therefore, so do q_{AC} and q_{DE}:

$$q_{AC}/T_C + q_{DE}/T_D = 0, \tag{17}$$

where T_C is the absolute temperature of the system in the state C (or in any state infinitesimally distant from C, such as A) and T_D is that in state D.

The same relation holds for all such pairs of points C and D around the whole cycle, each such pair lying on an adiabat. Each of the corresponding terms q_{AC}/T_C is of the form $đq_{rev}/T$, where $đq_{rev}$ is the infinitesimal heat absorbed by the system in that infinitesimal part of the reversible cycle and T is the system's absolute temperature at that point. Thus, summing over the whole cycle,

$$\oint \frac{đq_{rev}}{T} = 0. \tag{18}$$

Now let $(PQ)_1$ and $(PQ)_2$ be any two distinct reversible paths connecting two states P and Q. Then form (18),

$$\underset{\text{(on path 1)}}{\int_P^Q \frac{đq_{rev}}{T}} + \underset{\text{(on path 2)}}{\int_Q^P \frac{đq_{rev}}{T}} = 0 \tag{19}$$

because following path 1 from P to Q and path 2 from Q to P constitutes a reversible cycle. Changing the direction in which a reversible path is traversed changes the sign of the heat $đq_{rev}$ absorbed along each infinitesimal element of the path, so (19) is equally well

$$\underset{\text{(on path 1)}}{\int_P^Q \frac{đq_{rev}}{T}} = \underset{\text{(on path 2)}}{\int_P^Q \frac{đq_{rev}}{T}}. \tag{20}$$

But paths 1 and 2 are *any* two reversible paths connecting states P and Q. Therefore, (20) implies that

$$\int_P^Q \frac{đq_{rev}}{T} \tag{21}$$

is independent of the path; it depends only on the initial and final states P and Q. It is therefore the change ΔS in some function S of the state of the system:

$$\Delta S = \int \frac{đq_{rev}}{T}, \tag{22}$$

where the integral extends from the initial to the final state and is independent of the path that connects them.

The state function S is the entropy. The second law of thermodynamics has thus been seen to imply the existence of two functions of the state of a system, absolute temperature and entropy, of which the former is a function only of the system's (empirical) temperature and not of any other variable defining the system's state.

Since ΔS depends only on the initial and final states of the system, its value is independent of the nature of the process by which the change in state is effected and, in particular, is independent of whether the process is reversible or not. The meaning of (22) is that the difference in entropy $\Delta S = S_Q - S_P$ between two specified equilibrium states P and Q of a system, while independent of the nature of the process P → Q, may be calculated by Eq. (22), with $đq_{rev}$ the infinitesimal heat the system *would* absorb during an infinitesimal stage of a *hypothetical reversible* process that takes it from P to Q, and T is what the system's absolute temperature *would* be in that stage of the process.

Since ΔS is independent of the path, Eq. (22) implies that $T^{-1}đq_{rev}$ is an exact differential, dS:

$$\Delta S = \int dS, \qquad dS = \frac{1}{T}\,đq_{rev}. \tag{23}$$

In mathematical terms, $1/T$ is an integrating factor for the inexact differential $đq_{rev}$. When the only work done is pressure–volume work, so that $đw_{rev} = -pdV$, we have from (8) and (23)

$$đq_{rev} = dU + pdV = TdS; \tag{24}$$

i.e., $1/T$ is an integrating factor for the inexact differential $dU + pdV$, where the pressure $p = p(U,V)$ is some function of the system's state as defined by U and V. Mathematically, a differential in two variables always has an integrating factor, and in this case it would then have been some function of U and V; but the second law goes beyond this mathematical statement to assert that the integrating factor $1/T$ is not merely some function of the system's state as defined, for example, by the two variables U and V but is more specifically a function of only one variable, the system's temperature. When there is more than one reversible work term, the inexact differential $đq_{\text{rev}}$ ($= dU - đw_{\text{rev}}$) is a differential in more than two variables, and then mathematics alone no longer guarantees the existence of an integrating factor, while thermodynamics still asserts that there is one, and that it is $1/T$, a function only of the system's temperature. The second law is a law of nature, not of mathematics.

If a process is both adiabatic and reversible, then $đq = đq_{\text{rev}} = 0$, and so $dS = 0$; i.e., such a process is isentropic. A reversible adiabat, such as BC and DA in Fig. 4 and CD and EB in Fig. 5, is an isentrope.

2.4 Entropy of an Isolated System

Let A be a constrained equilibrium state of an isolated system. An example is an isolated system—a system thermally and mechanically isolated from its surroundings—consisting of two subsystems at different temperatures but thermally insulated from each other, so that there is no heat flow between the subsystems and their respective temperatures do not equalize. Another example is a thermally and mechanically isolated mixture of chemical substances that are potentially capable of reacting with each other but only in the presence of a catalyst, which is absent, so that no reaction takes place. Let B be a state obtained from A by the relaxation of one or more of the constraints while the system as a whole is still isolated. In the first example above, the subsystems may be put in thermal contact, so that heat flows between them and their temperatures approach equality. In the second example, introduce the catalyst so that reaction occurs and the system approaches chemical equilibrium.

As these examples show, each time a constraint is removed a spontaneous (hence irreversible) process A → B occurs in this otherwise isolated system, and the system's state becomes more nearly one of unconstrained equilibrium. In Fig. 6 there is indicated schematically, by hatch marks, such a spontaneous change from state A to state B of an isolated system. Then, with the system in state B, put it in contact with its surroundings and allow it to undergo a reversible adiabatic change to a state C, and then a reversible isothermal change from C to a state D that lies on the same reversible adiabat as state A, so that the reversible adiabatic process D → A then completes a cycle (Fig. 6).

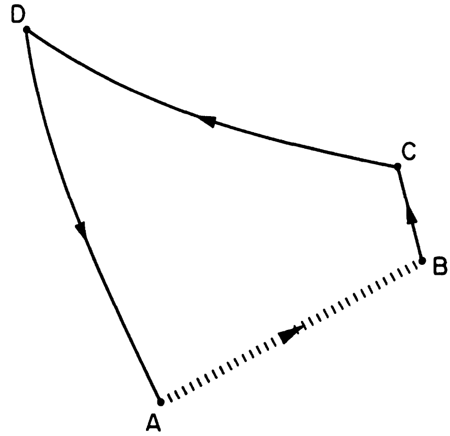

FIG. 6. A system, while isolated, undergoes an irreversible process A → B from a state of constrained equilibrium A to a less constrained or unconstrained state B; while B → C and D → A are reversible adiabatic processes and C → D is a reversible isothermal process at the temperature T.

In the original irreversible change from A to B, the system is isolated so that $q_{\text{AB}} = w_{\text{AB}} = 0$. No heat is absorbed in the adiabatic changes B → C and D → A, and so $q_{\text{BC}} = q_{\text{DA}} = 0$, and these are reversible adiabatic changes, hence isentropic, and so $\Delta S_{\text{BC}} = \Delta S_{\text{DA}} = 0$ as well. Let T be the constant absolute temperature in the reversible isothermal process C → D, so that $\Delta S_{\text{CD}} = q_{\text{CD}}/T$. This q_{CD} is the only heat absorbed or evolved by the system in the entire cycle. From this picture and principles already annunciated, one may conclude that the second law would be violated unless q_{CD} and the equivalent work $-w(=q_{\text{CD}})$ done by the system in the cycle were both negative, and that therefore the entropy change ΔS_{AB} for the process A → B were positive:

$$\Delta S_{\text{AB}} > 0. \tag{25}$$

The principle expressed in (25) is that as constraints in an isolated system are relaxed and the system then approaches equilibrium by irreversible processes occurring in it, its entropy increases. As a corollary, the condition that a process occurring in an isolated system be spontaneous is that, as a result of the process, the system's entropy increase. An isolated system at full, unconstrained equilibrium then already has its maximum possible entropy, so that $dS = 0$ for any virtual infinitesimal change that would occur in it while it remains isolated.

2.5 $S = S(U,V,M_1,M_2,\ldots)$

The volume V of a system is the sum of the volumes of its parts; V is then said to be an extensive variable. The mass M of a system is likewise the sum of the masses of its parts, and so M is also an extensive variable, and the same is true, too, of each of the separate masses M_1, $M_2, \ldots$ of the different chemical substances 1, 2, . . . of which the system is composed. The energy U is also extensive: the net heat (positive or negative) absorbed by a system from its surroundings and net work (positive or negative) done on the system by its surroundings, in any process, is the sum of the heats absorbed from the surroundings by, and work done by the surroundings on, the component subsystems, so that ΔU for the system is the sum of the ΔU's of the system's component parts. (The heat absorbed by one subsystem from another and the work done by one on another need not be considered in this arithmetic because their net contribution to the sum of the energy changes suffered by all the subsystems together is nil.) Similarly, the entropy S is extensive: $đq_{\text{rev}}$ in (22) is the sum of the infinitesimal reversible heats absorbed from (or evolved to) the surroundings by the system's component parts.

When the only work is p-V work, so that $đw_{\text{rev}} = -pdV$, a system's equilibrium state is fixed by specifying its energy, its volume, and its chemical composition. (These are the variables that would be conserved if the system were isolated; the constancy of V would be a consequence of the mechanical isolation, which would be achieved by rigid container walls, ensuring that no p-V work is done on or by the system.) The entropy, which depends on the state of the system, may then be taken to be a function of those same variables: $S = S(U,V,M_1,M_2,\ldots)$. (The composition is specified by the masses M_1, $M_2, \ldots$ of as many *independent* chemical components as there are. What one means by independent components is stated more precisely and illustrated later, in connection with the discussion of chemical equilibrium at the end of Sec. 3.3.)

If there were more than one kind of work possible, there would be additional extensive variables analogous to V required to define the state, and S would be a function of those also. Every system has its own $S(U,V,M_1,M_2,\ldots)$ expressing its equilibrium entropy as a function of its energy, volume, and chemical composition. The relation

$$S = S(U,V,M_1,M_2,\ldots) \tag{26}$$

may be said to be that system's equation of state (thus extending somewhat the more traditional use of that term).

We cannot deduce a system's equation of state from thermodynamics alone (that is the domain of statistical mechanics), but we can, from general principles, deduce an important property of the function $S(U,V,\ldots)$. Let A and B be two equilibrium states of a system Σ. Take a sample $\Sigma_{\alpha,\mathrm{A}}$ of the system Σ in its state A, this sample consisting of a fraction α of the original system in state A, so that all the extensive functions S, U, V, $M_1, \ldots$ of the sample are equal to the same fraction α of those variables in the original system. Let $\Sigma_{1-\alpha,\mathrm{B}}$, likewise, be a fractional sample of Σ in state B, the fraction now being $1 - \alpha$. Consider the composite system consisting of $\Sigma_{\alpha,\mathrm{A}}$ and $\Sigma_{1-\alpha,\mathrm{B}}$, and isolate it. If these two subsystems are not in contact, the entropy of the composite system, by extensivity, is $\alpha S(U_\mathrm{A},V_\mathrm{A},\ldots) + (1 - \alpha)S(U_\mathrm{B},V_\mathrm{B},\ldots)$, where $S(U,V,\ldots)$ is the entropy of the original system Σ as a function of its energy U, volume V, etc. If these subsystems are now put into contact with each other and the (isolated) composite system allowed to come to equilibrium, its entropy will increase, by the principle deduced in Sec. 2.4. But this equilibrated entropy is now that of the original system Σ in a new state C in which, because of the isolation of the composite system during its equilibration, $U_\mathrm{C} = \alpha U_\mathrm{A} + (1 - \alpha)U_\mathrm{B}$, etc. Thus, by the inequality just alluded to,

$$S[\alpha U_A + (1-\alpha)U_B, \ldots] > \alpha S(U_A, \ldots) + (1-\alpha)S(U_B, \ldots). \qquad (27)$$

This is true whatever the fraction α.

In mathematical terms, (27) says that $S(U,V,M_1,M_2,\ldots)$ is a *concave* (downward) function of its arguments U, V, M_1, M_2, ... At fixed V, M_1, ..., for example, the entropy $S(U)$ as a function of the remaining variable U will be as depicted in Fig. 7: for all U between U_A and U_B the curve $S = S(U)$ lies above the chord that connects U_A, S_A to U_B, S_B. The curve is pictured with positive slope at all U to be in accord with the usual convention of taking absolute temperatures $T > 0$ and with $\partial S/\partial U = 1/T$ from (24). More generally, the curve $S = S(U)$ in Fig. 7 would be a surface or hypersurface $S = S(U,V,M_1,M_2,\ldots)$ and the chord AB a corresponding plane or hyperplane intersecting it, and within the region of the intersection, the former would lie above the latter.

The downward concavity of $S(U,V,M_1, M_2,\ldots)$ requires that the eigenvalues of the symmetric matrix of second derivatives (Hessian matrix) of S,

$$\begin{pmatrix} \dfrac{\partial^2 S}{\partial U^2} & \dfrac{\partial^2 S}{\partial U \partial V} & \cdots \\ \dfrac{\partial^2 S}{\partial U \partial V} & \dfrac{\partial^2 S}{\partial V^2} & \cdots \\ \vdots & \vdots & \end{pmatrix}, \qquad (28)$$

be all negative. This implies, in particular, that the elements of the principal diagonal, $\partial^2 S/\partial U^2$, etc., are all negative:

$$\frac{\partial^2 S}{\partial U^2} < 0, \quad \frac{\partial^2 S}{\partial V^2} < 0, \quad \frac{\partial^2 S}{\partial M_1^2} < 0, \cdots. \qquad (29)$$

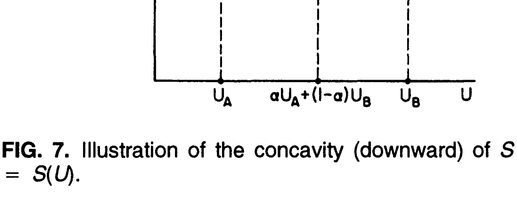

FIG. 7. Illustration of the concavity (downward) of $S = S(U)$.

While the conditions (29) are consequences of the concavity of $S(U,V,M_1,M_2,\ldots)$ they are not the totality of such consequences since the negativity of all the eigenvalues implies restrictions on the mixed second derivatives $\partial^2 S/\partial U \partial V$, etc., as well.

A necessary qualification of some of these statements should be noted. If the "states" A and B in (27) are two different phases in equilibrium (a liquid and its equilibrium vapor, for example), the inequality becomes an equality. To allow for such exceptional cases the "greater than" sign in (27) should be replaced by $\geq$, meaning "not less than." In the case of such phase equilibrium a portion of the entropy surface is a ruled surface, i.e., consists of straight line segments, at each point of which one of the eigenvalues of the Hessian matrix is then zero rather than strictly negative. Where three phases are in equilibrium, two of the eigenvalues are zero; etc. The associated eigenvectors, however, are not in general in the direction of any of the coordinate axes, so the inequalities (29) remain true.

Until now the systems considered have, implicitly, been closed systems, i.e., those in which no material substances enter or leave so that the variables M_1, M_2, ... are fixed. For such a system there need be no explicit reference to these composition variables, and one may take $S = S(U,V)$ with fixed M_1, M_2, ... understood. When such a system undergoes an infinitesimal change in its equilibrium state, the accompanying infinitesimal changes dS, dU, and dV in the entropy, energy, and volume are not independent but are related to each other by (24). We will use the conventional thermodynamic partial-derivative notation according to which $(\partial S/\partial U)_V$, for example, means the rate of change of S with U at fixed V and, implicitly, fixed M_1, M_2, ... as well. From (24) we then have the two thermodynamic identities (the first of these having already been noted)

$$\left(\frac{\partial S}{\partial U}\right)_V = \frac{1}{T}, \quad \left(\frac{\partial S}{\partial V}\right)_U = \frac{p}{T}. \qquad (30)$$

The ratio of the heat absorbed by a system in an infinitesimal process to the infinitesimal temperature rise dT that accompanies the absorption of that heat is called the *heat capacity* of the system; the greater the heat capacity the smaller the temperature rise for a given amount of heat absorbed. If the absorption of heat occurs at constant volume, this defines the constant-volume heat capacity, C_V. Since by (9) the infinitesimal heat absorbed is then dU, we have the thermodynamic identity

$$C_V = \left(\frac{\partial U}{\partial T}\right)_V . \tag{31}$$

This is clearly a result of the first law alone; the temperature occurs in (31) only as the difference dT, and so could be any empirical temperature. From the first of Eqs. (30), however, we have $(\partial S/\partial T)_V = T^{-1}(\partial U/\partial T)_V$, and so from (31),

$$C_V = T\left(\frac{\partial S}{\partial T}\right)_V . \tag{32}$$

This is a result of the second law. Note that the temperature appears in it not merely as the difference dT, which could be measured on any empirical scale, but also as the multiplying factor T, where it is explicitly the absolute temperature.

From the concavity condition $\partial^2 S/\partial U^2 < 0$, which is one of the inequalities (29), together with (31) and the first of Eqs. (30), we have $0 > (\partial T^{-1}/\partial U)_V = -(T^2 C_V)^{-1}$, and so

$$C_V > 0. \tag{33}$$

The inequality (33) is the first of several such—often called thermodynamic stability conditions—that we shall obtain. Indeed, as was seen above, the concavity of S implies more conditions than just $\partial^2 S/\partial U^2 < 0$, and there are then derivable more stability conditions than just that in (33). Some of these will be derived later by a different route.

If one now explicitly recognizes the dependence of S on chemical composition and allows that composition itself to be variable (so considering now an *open* rather than a closed system), one must add to $dS = (1/T)dU + (p/T)dV$ [Eq. (24)], which gives the change in S due to changes in U and V at fixed composition, the further contributions to dS that would arise from infinitesimal changes dM_1, dM_2, ... Thus, this fundamental differential identity now becomes

$$dS = \frac{1}{T}dU + \frac{p}{T}dV - \frac{\mu_1}{T}dM_1 - \frac{\mu_2}{T}dM_2 - \cdots , \tag{34}$$

where μ_1, μ_2, . . .—called the *chemical potentials* of the substances 1,2, . . . in the system—are by definition

$$\mu_i = -T\left(\frac{\partial S}{\partial M_i}\right)_{U,V,\text{ all } M_{j(\neq i)}} . \tag{35}$$

The partial derivative, as indicated, is with respect to the mass of substance i holding fixed the total energy of the system, the volume, and the masses M_j of all the *other* independent chemical components.

(Chemical potentials are often defined with respect to the numbers of moles or molecules rather than the masses of the several components, but then when thermodynamic functions are calculated for specific molecular models, the molecular basis is obscured by the molecular weights' not appearing explicitly in the results. With the present definition the molecular basis of such applications is manifest in the explicit appearance of the molecular weights. This important didactic point was made by Gibbs; see Gibbs, 1961.)

Since $S(U,V,M_1,M_2, \ldots)$ is an extensive function of extensive arguments, $S(\alpha U, \alpha V, \ldots) \equiv \alpha S(U,V, \ldots)$ for all α; i.e., S is homogeneous of the first degree in its extensive arguments. According to Euler's theorem on homogeneous functions, if a function $f(\alpha x, \alpha y, \ldots) \equiv \alpha^n f(x,y, \ldots)$, so that f is homogeneous of degree n in x, y, . . . , then $nf = x\partial f/\partial x + y\partial f/\partial y + \cdots$. By this theorem, with $n = 1$, it follows from (35) that

$$S = \frac{1}{T}U + \frac{p}{T}V - \frac{\mu_1}{T}M_1 - \frac{\mu_2}{T}M_2 - \cdots , \tag{36}$$

an important thermodynamic identity.

3. FREE ENERGY

3.1 Legendre Transforms; Free Energies F and G

The fundamental differential identity (34) may be rewritten

$$dU = TdS - pdV + \mu_1 dM_1 + \mu_2 dM_2 + \cdots, \tag{37}$$

and equivalently, from the definition of the enthalpy in (10),

$$dH = TdS + Vdp + \mu_1 dM_1 + \mu_2 dM_2 + \cdots. \tag{38}$$

Defining

$$F = U - TS, \qquad G = H - TS = F + pV, \tag{39}$$

one has also

$$\begin{aligned} dF &= -SdT - pdV + \mu_1 dM_1 + \mu_2 dM_2 + \cdots, \\ dG &= -SdT + Vdp + \mu_1 dM_1 + \mu_2 dM_2 + \cdots. \end{aligned} \tag{40}$$

The transformations from (37) to (38) and (40) replace one or both of V and S as independent variables by their respective conjugates $-p$ and T. Mathematically, these are Legendre transformations. The Legendre-transformed energy H is the enthalpy; the Legendre-transformed energies F and G defined by (39) are called free energies, the former that of Helmholtz and the latter that of Gibbs. Just as S, U, V, M_1, $M_2, \ldots$ are extensive quantities, so also are the transformed energies H, F, and G.

As alternatives to the definition of the chemical potential μ_i in (35), one now has, from (37), (38), and (40), four additional expressions for μ_i, the two most useful of which are those from (40),

$$\mu_i = \left(\frac{\partial F}{\partial M_i}\right)_{T,V,\ \text{all}\ M_{j(\neq i)}} = \left(\frac{\partial G}{\partial M_i}\right)_{T,p,\ \text{all}\ M_{j(\neq i)}}. \tag{41}$$

Likewise, as analogs of the identities (30), in which fixed M_1, $M_2, \ldots$ is understood, one has from (37), (38), and (40) eight additional identities, of which the four most useful are again those from (40),

$$\left(\frac{\partial F}{\partial T}\right)_V = \left(\frac{\partial G}{\partial T}\right)_p = -S, \quad \left(\frac{\partial F}{\partial V}\right)_T = -p, \quad \left(\frac{\partial G}{\partial p}\right)_T = V. \tag{42}$$

From (39) one has $G = U - TS + pV$, and so from (36), which was a consequence of the extensivity of the entropy as a function of the extensive variables U, V, M_1, $M_2, \ldots$, one has

$$G = \mu_1 M_1 + \mu_2 M_2 + \cdots. \tag{43}$$

On substituting (43) for G on the left-hand side of the second of Eqs. (40), one obtains

$$M_1 d\mu_1 + M_2 d\mu_2 + \cdots = -SdT + Vdp, \tag{44}$$

which is called the *Gibbs–Duhem equation.* It is the result of having replaced, by Legendre transformation, all the extensive independent variables S, V, M_1, $M_2, \ldots$ on the right-hand side of (37) by their conjugates T, $-p$, μ_1, $\mu_2, \cdots$. Replacing by Legendre transformation all but the volume V leads to the ("grand canonical") free energy $\Omega(T,V,\mu_1,\mu_2,\ldots) = -Vp(T,\mu_1,\mu_2,\ldots)$, just proportional to V, with $d\Omega = -pdV - SdT - M_1 d\mu_1 - \cdots$.

3.2 Equilibrium and Convexity

Imagine a very large isolated system containing in its interior a much smaller subsystem the volume V and chemical composition M_1, $M_2, \ldots$ of which are fixed. The rest of the entire, isolated system will be referred to as the heat bath or the thermostat in which the smaller subsystem is immersed. The smaller subsystem is the system of interest. It is imagined to be at first in a state of constrained equilibrium at the temperature T of the thermostat; but then the constraint is removed and the subsystem spontaneously (hence irreversibly) approaches a state of complete internal equilibrium in which its temperature is again T. The thermostat, because of its much larger size, is scarcely perturbed by this process and so may be taken to be always in a state of equilibrium at the fixed temperature T; that *defines* a heat bath or thermostat. Then any heat q (positive or negative) absorbed by the subsystem during its equilibration is lost by the thermostat reversibly, and so the entropy change suffered by the thermostat is $-q/T$, where T is its constant temperature as well as the initial and final temperature of the subsystem. The entropy change suffered by the subsystem—the system of interest—during its equilibration is some ΔS, and so, by extensivity, that of the entire system is $\Delta S - q/T$. The whole system is isolated so this net entropy change is positive: $\Delta S - q/T \geq 0$ (the = sign holding only if the subsystem was in complete internal equilibrium at the temperature T to begin with).

Since the volume V of the subsystem is fixed, the heat q it absorbs is equal to the change in its energy, ΔU, by (9), and so this inequality is $\Delta S - \Delta U/T \geq 0$; or, from the definition of F in (39), and the usual convention for the sign of T, it is

$$\Delta F \leq 0 \qquad (\text{at fixed } T, V, M_1, M_2, \ldots). \tag{45}$$

Thus, when a spontaneous process takes place in a system of fixed T, V, M_1, $M_2, \ldots$ the Helmholtz free energy F of the system decreases, and correspondingly the condition of equilibrium in such a system is that its free energy be minimal for the given T, V, M_1, $M_2, \ldots$.

By a similar argument one shows that

$$\Delta G \leq 0 \qquad (\text{at fixed } T, p, M_1, M_2, \ldots); \tag{46}$$

i.e., that at fixed T, p, M_1, $M_2, \ldots$ the Gibbs free energy G spontaneously approaches a minimum and so is minimal at equilibrium.

Just as the extremal condition on the entropy led, in Sec. 2.5, to the concavity (downward, or convexity upward) of $S(U, V, M_1, M_2, \ldots)$ as a function of its extensive arguments U, V, M_1, $M_2, \ldots$, so also do (45) and (46) require that *$F(T,V,M_1,M_2,\ldots)$ at fixed T* be a convex (downward, or concave upward) function of its *extensive* arguments V, M_1, $M_2, \ldots$, and a concave downward function of T at fixed V, $M_1, \ldots$; and that *$G(T,p,M_1,M_2,\ldots)$ at fixed T and p* be a convex (downward, or concave upward) function of its *extensive* arguments M_1, $M_2, \ldots$, and a concave downward function of T and p at fixed M_1, $M_2, \ldots$. If all but one of the extensive variables is replaced by its conjugate by Legendre transformation (as with the free energy Ω, for example, defined at the end of Sec. 3.1), the resulting function is neither concave nor convex in the remaining extensive variable but simply proportional to it.

The convexity of F and G lead to further useful stability conditions analogous to (33) for C_V. The isothermal compressibility χ is defined as the fractional rate of decrease of volume with increasing pressure at fixed temperature (and, it is understood, at fixed M_1, $M_2, \ldots$ as well),

$$\chi = -\frac{1}{V}\left(\frac{\partial V}{\partial p}\right)_T. \tag{47}$$

As the analog of the inequalities (29) (now with > instead of < because we have convexity instead of concavity), we have $\partial^2 F/\partial V^2 > 0$, etc. Then from the third of the equalities (42) and the definition of χ in (47), we have the stability condition

$$\chi > 0. \tag{48}$$

Likewise, from $\partial^2 F/\partial M_i^2 > 0$ and the first of the equalities (41),

$$\left(\frac{\partial \mu_i}{\partial M_i}\right)_{T,V,\text{ all } M_{j(\neq i)}} > 0. \tag{49}$$

Similarly, from $\partial^2 G/\partial M_i^2 > 0$ and the second of the equalities (41),

$$\left(\frac{\partial \mu_i}{\partial M_i}\right)_{T,p,\text{ all } M_{j(\neq i)}} > 0. \tag{50}$$

The inequalities (33) and (48)–(50) are only a selection of the full set derivable from the convexity–concavity conditions but are the most generally useful.

3.3 Potentials, Fields, and Densities

The functions $S(U,V,M_1,M_2,\ldots)$, $F(T,V,M_1,M_2,\ldots)$, $G(T,p,M_1,M_2,\ldots)$, and $p(T,\mu_1,\mu_2,\ldots)$, of just those specified arguments, are examples of what one may call *thermodynamic potentials*, by which is meant that from any of those functions of those arguments one may obtain all other thermodynamic functions of the system by differentiations alone, no integrations with their attendant unknown constants of integration being required. In the case of the last of those mentioned, $p(T,\mu_1,\mu_2,\ldots)$, one obtains densities—U/V, S/V, M_1/V, etc.—and their derivatives, and not the extensive functions V, U, S, $M_1, \ldots$ themselves, but that is a minor difference. [One may instead consider the extensive potential $\Omega(V,T,\mu_1,\mu_2,\ldots) \equiv -Vp(T,\mu_1,\mu_2,\ldots)$ introduced at the end of Sec. 3.1, which is then on the same footing as the others.] Those thermodynamic functions that are the coefficients of the differentials (37), (40), and (44) (the latter after division by V) are immediately the first derivatives of the thermodynamic potentials, and all other thermodynamic properties are obtained by (at most) further differentiations with respect to those same independent variables. That is not to say that one may not, for example, contemplate p as a function of T, M_1/V, $M_2/V, \ldots,$ for p is indeed determined when the values of those variables are specified; but as a function of *those* variables, p is not a thermodynamic potential as defined above.

The pressure p and the variables T, μ_1, $\mu_2, \ldots,$ the differentials of which occur in the Gibbs–Duhem equation (44), are termed *fields* (Griffiths and Wheeler, 1970). (The name derives from the subject of magnetism, where the magnetic field is the thermodynamic conjugate of the extensive magnetization.) After division by V in (44), the coefficients of those differentials are the *densities* S/V, M_1/V, etc. More generally, the term density is used to denote the ratio of any pair of extensive variables. Fields and densities are *intensive,* in the sense that their values in a homogeneous system at equilibrium are the same everywhere; they reflect the intrinsic properties of such a system and are independent of the system's size. But the fields, unlike the densities, also have uniform values in an inhomogeneous system at equilibrium. (The pressure is sometimes an exception. When the interfaces between different phases in contact are curved, the pressure has unequal values in the phases; but when the interfaces are planar, the pressure at equilibrium is the same in all the phases.) The uniformity of the fields may be derived as consequences of the conditions of equilibrium found earlier—(25), (45), or (46), with the inequalities holding as equalities—by considering the flow of heat, volume, or matter between the subsystems of a composite system. The distinction between extensive and intensive variables—when the latter are densities—is often superficial or trivial but that between fields and densities is fundamental.

From an argument similar to that by which one derives uniformity of chemical potential as a necessary condition for equilibrium, one may also derive the condition of equilibrium with respect to chemical reaction. Any balanced chemical equation is of the form

$$aA + bB + \cdots \rightleftharpoons xX + yY + \cdots, \tag{51}$$

where $A, B, \ldots, X, Y, \ldots$ are the chemical formulas of the various substances that take part in the reaction as reactants or products while $a, b, \ldots, x, y, \ldots$ are the stoichiometric coefficients giving the relative numbers of molecules (or moles) of those substances used up or formed in the reaction. (Such stoichiometry, which is central in chemistry, is an intrinsically molecular notion. It is the first occurrence of a molecular model in this account and so, in accord with a parenthetical remark made in Sec. 2.5, it will be the first instance in which the molecular weights of chemical substances appear in a fundamental formula. That will not happen again until Sec. 5). One may then derive, as the condition of equilibrium with respect to reaction (51),

$$am_A\mu_A + bm_B\mu_B + \cdots = xm_X\mu_X + ym_Y\mu_Y + \cdots, \tag{52}$$

where m_A, etc., are the molecular weights (as they are known colloquially, but more properly, molar masses) of the various substances. Note that (52) is homogeneous in the molecular weights $m_A, \ldots$ and so is independent of the molecular-weight scale—as expected, since any such scale is of relative masses only.

It is now possible to answer the question raised in Sec. 2.5 about how one counts the number of independent chemical components. Pure water, for example, is a single component. (For simplicity, we ignore the matter of isotopic composition. If that were variable, it would indeed alter the count but would not require any new principles.) Because of hydrogen bonding one may imagine there to be distinct molecular species $(H_2O)_n$

with $n = 1, 2, \ldots$, but at equilibrium one has $2H_2O \rightleftharpoons (H_2O)_2$, etc. Thus, however many species one recognizes in pure water, there is that number minus one independent equilibrium conditions (52), and there then remains only one independently variable chemical potential; hence the conclusion that pure water is a one-component system. One might choose to consider separately the H and O components of H_2O, but then the stoichiometric constraint $n_H = 2n_O$ on the relative numbers of moles or atoms of H and O again reduces the number of independent components to one. The general rule is that the number of independent chemical components is the number of species minus the number of constraints, which may be either of equilibrium or of stoichiometry.

3.4 Further Identities

Many useful identities in addition to those already derived follow from (39)–(44).' The Gibbs–Duhem equation, (44), for example, may be rewritten

$$dp = s\,dT + \rho_1 d\mu_1 + \rho_2 d\mu_2 + \cdots \tag{53}$$

where $s = S/V$ is the entropy density and $\rho_i = M_i/V$ is the mass density of component i. Then from the thermodynamic potential $p(T,\mu_1,\mu_2,\ldots)$, as anticipated earlier, one obtains

$$s = \left(\frac{\partial p}{\partial T}\right)_{\mu_1,\mu_2,\ldots}, \qquad \rho_i = \left(\frac{\partial p}{\partial \mu_i}\right)_{T,\ \text{all}\ \mu_j(\neq i)}. \tag{54}$$

From (39), (43), and (44) the Gibbs–Duhem equation may be written in still another form,

$$d\frac{p}{T} = -u\,d\frac{1}{T} + \rho_1 d\frac{\mu_1}{T} + \rho_2 d\frac{\mu_2}{T} + \cdots, \tag{55}$$

where $u = U/V$ is the energy density; so

$$u = -\left(\frac{\partial(p/T)}{\partial(1/T)}\right)_{\mu_1,\mu_2,\ldots}. \tag{56}$$

From (39) and (42) there follow the *Gibbs–Helmholtz equations*,

$$\left(\frac{\partial(F/T)}{\partial T}\right)_V = -\frac{U}{T^2}, \qquad \left(\frac{\partial(G/T)}{\partial T}\right)_p = -\frac{H}{T^2} \tag{57}$$

or, equivalently,

$$\left(\frac{\partial(F/T)}{\partial(1/T)}\right)_V = U, \qquad \left(\frac{\partial(G/T)}{\partial(1/T)}\right)_p = H. \tag{58}$$

Heat capacity was defined in Sec. 2.5, and that led to the identities (31) and (32) for the constant-volume heat capacity C_V. If the heating process that defines the heat capacity occurs at constant pressure instead of at constant volume, the ratio of heat absorbed to temperature rise is then the constant-pressure heat capacity C_p. Then from (11) and (37) one has as the analogs of (31) and (32),

$$C_p = \left(\frac{\partial H}{\partial T}\right)_p = T\left(\frac{\partial S}{\partial T}\right)_p. \tag{59}$$

Likewise, as an analog of the isothermal compressibility χ defined in (47), one defines the coefficient of thermal expansion α as the fractional rate of increase of volume with increasing temperature at fixed pressure,

$$\alpha = \frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_p. \tag{60}$$

Unlike χ, which is always positive [inequality (48)], α may be of either sign. For liquid water at 1 atm, for example, which has its maximum density at 4 °C, α is negative below 4° and positive above. Related to α and χ is the thermal pressure coefficient $(\partial p/\partial T)_V$,

$$\left(\frac{\partial p}{\partial T}\right)_V = \frac{\alpha}{\chi}, \tag{61}$$

which has the sign of α. These quantities C_V, C_p, α, χ, and $(\partial p/\partial T)_V$ occur in the further identities that follow.

From the equalities of mixed second derivatives $\partial^2 F/\partial V\partial T = \partial^2 F/\partial T\partial V$ and $\partial^2 G/\partial p\partial T = \partial^2 G/\partial T\partial p$, one obtains from (40), (61), and the definition of α in (60)

$$\left(\frac{\partial S}{\partial V}\right)_T = \left(\frac{\partial p}{\partial T}\right)_V = \frac{\alpha}{\chi}, \qquad \left(\frac{\partial S}{\partial p}\right)_T = -V\alpha. \tag{62}$$

These are *Maxwell relations*. From these and (37) and (38) one obtains the important formulas for the volume dependence of the energy and the pressure dependence of the enthalpy, at fixed temperature:

$$\left(\frac{\partial U}{\partial V}\right)_T = T\left(\frac{\partial p}{\partial T}\right)_V - p = -\left(\frac{\partial(p/T)}{\partial(1/T)}\right)_V$$
$$\left(\frac{\partial H}{\partial p}\right)_T = V(1 - T\alpha). \tag{63}$$

That T occurs multiplicatively in the formulas (63) shows that these are consequences of the second law even though the quantities U, H, V, p, and α, as well as the temperature differences implicit in the temperature derivatives, are recognizable from the first law alone.

From (32) and (62) for the temperature and volume derivatives of the entropy, it follows that in any infinitesimal change in the thermodynamic state of a system (at fixed M_1, M_2, . . .), the infinitesimal changes in S, V, and T are related by $dS = (C_V/T)dT + (\alpha/\chi)dV$. This holds, in particular, when the change occurs at constant pressure; and so, with the expression (59) that relates $(\partial S/\partial T)_p$ to C_p and the definition of α in (60),

$$C_p = C_V + TV\alpha^2/\chi. \tag{64}$$

This important identity relates the two principal heat capacities C_p and C_V. Since $\chi > 0$ we note that C_p is the greater of the two. (For liquid water at 1 atm and 4 °C, the two heat capacities are the same because then $\alpha = 0$). That T occurs multiplicatively on the right-hand side of (64) shows that this formula, too, is a consequence of the second law in spite of the fact that all the other quantities in it (including the temperature differences implicit in the derivatives defining C_V and C_p in terms of U and H) are definable from the first law alone. Since $C_p \geq C_V$ it follows *a fortiori* from the stability condition $C_V > 0$ in (33) that $C_p > 0$.

The last of the identities to be derived here relates the isentropic (i.e., adiabatic and reversible) rate of change of pressure with volume, $(\partial p/\partial V)_S$, to the isothermal rate of change, $(\partial p/\partial V)_T$; or equivalently, by the definition (47), relates the isentropic to the isothermal compressibility. From the mathematical identities

$$\left(\frac{\partial p}{\partial V}\right)_S = -\frac{\left(\frac{\partial S}{\partial V}\right)_p}{\left(\frac{\partial S}{\partial p}\right)_V} = -\frac{\left(\frac{\partial S}{\partial T}\right)_p\left(\frac{\partial T}{\partial V}\right)_p}{\left(\frac{\partial S}{\partial T}\right)_V\left(\frac{\partial T}{\partial p}\right)_V}$$
$$= \frac{\left(\frac{\partial S}{\partial T}\right)_p}{\left(\frac{\partial S}{\partial T}\right)_V}\left(\frac{\partial p}{\partial V}\right)_T \tag{65}$$

and the thermodynamic identities (32) and (59), one obtains

$$\left(\frac{\partial p}{\partial V}\right)_S = \frac{C_p}{C_V}\left(\frac{\partial p}{\partial V}\right)_T. \tag{66}$$

If the isothermal compressibility χ defined by (47) is written χ_T to distinguish it from the isentropic $\chi_S = -V^{-1}(\partial V/\partial p)_S$, then (66) is

$$\chi_S = \gamma^{-1}\chi_T, \qquad \gamma = C_p/C_V. \tag{67}$$

Since $C_p \geq C_V > 0$ we have $\gamma \geq 1$ and $\chi_S \leq \chi_T$. From (66) it follows that when isentropes and isotherms in the p-V plane intersect—for example, at A, B, C, and D in Fig. 4 (since reversible adiabats are isentropes)—the isentrope has a steeper (more negative) slope at the point of intersection than the isotherm does.

Equation (67) has an application in acoustics in a formula for the speed of sound, c. From acoustic theory one has $c^2 = dp/d\rho$, the rate of change of pressure with density in the medium in which the sound is traveling, the derivative to be evaluated for the conditions under which the sound propagates. Any practical sound frequency is so great that in the short time of one period negligible heat is transported into or out of any element of the medium of length equal to the wavelength of the sound wave, so that the sound propagation is effectively adiabatic; and at frequencies below that at which there is significant dispersion—which means, typically, below ultrasonic frequencies—the periodic expansions and compressions are effectively also reversible. Thus, over a great range of frequencies, the sound propagation is both adiabatic and reversible, hence isentropic; so $dp/d\rho$ here means $(\partial p/\partial \rho)_S$. Since $\rho = (\text{mass})/V$, we have for a fixed mass $(\partial p/\partial \rho)_S = -(V/\rho)(\partial p/\partial V)_S$. Then from (67) and the formula for c^2 quoted above,

$$c = \sqrt{\gamma/\rho\chi_T}. \tag{68}$$

For example, for dry air at 1 atm and 20 °C one has $\rho = 1.205 \times 10^{-3}$ g/cm^3, $\gamma = 1.40$, and $\chi_T = 1.00$ atm^{-1} = 9.87×10^{-7} cm s^2/g, so $c = 3.43 \times 10^4$ cm/s = 343 m/s. For liquid water at its triple point one has $\rho = 1.000$ g/cm^3, $\gamma = 1.001$, and $\chi_T = 5.089 \times 10^{-5}$ bar^{-1} = 5.089×10^{-11} cm s^2/g, so $c = 1.402 \times 10^5$ cm/s = 1402 m/s (Rowlinson and Swinton, 1982).

4. PHASE EQUILIBRIUM

4.1 Phase Rule

The (Gibbs) phase rule states how many of a system's field variables (variables such as pressure, temperature, and chemical potential; Sec. 3.3) may be independently varied when the system consists of some number C of independent chemical components (defined in Sec. 3.3) and some number P of coexisting phases. This number, f, of independently variable fields is termed the number of degrees of freedom in the system. It is the number of field variables remaining to be specified in order that all the fields then have prescribed values. It is equivalently the dimensionality of the thermodynamic field space, in which a thermodynamic state of the P-phase, C-component system is represented by a point. Because only fields but not densities are specified, the thermodynamic state so represented is one in which only the intensive properties of each of the P phases present are determined but not their amounts (i.e., not the masses or volumes of the phases).

The pressure $p^{(\alpha)}$ in each of the phases $\alpha = 1, \ldots, P$ separately is some function $p^{(\alpha)}(T^{(\alpha)}, \mu_1^{(\alpha)}, \ldots, \mu_C^{(\alpha)})$ of the temperature $T^{(\alpha)}$ and the chemical potentials $\mu_i^{(\alpha)}$ of the C components in that phase (Sec. 3.3). The arguments $T^{(\alpha)}$, $\mu_1^{(\alpha)}, \ldots, \mu_C^{(\alpha)}$ of these P functions together constitute $P(C + 1)$ variables that one would need to specify to fix the states of all P phases if the phases were independent; but if the phases are in equilibrium with each other, the temperature T and the C chemical potentials μ_i must each have the same value in every phase (Sec. 3.3); and so also must the pressure p (assuming planar interfaces). These equilibrium conditions

$$\begin{aligned} T^{(1)} &= \cdots = T^{(P)} \\ \mu_1^{(1)} &= \cdots = \mu_1^{(P)} \\ &\vdots \\ \mu_C^{(1)} &= \cdots = \mu_C^{(P)} \\ p^{(1)} &= \cdots = p^{(P)} \end{aligned} \tag{69}$$

constitute $(P - 1)(C + 2)$ constraints on the values of the $P(C + 1)$ fields $T^{(\alpha)}$, $\mu_1^{(\alpha)}, \ldots, \mu_C^{(\alpha)}$, leaving only $f = P(C + 1) - (P - 1)(C + 2)$ of them independently variable. That is the phase rule:

$$f = C - P + 2. \tag{70}$$

According to this rule a C-component, one-phase system has $C + 1$ degrees of freedom so that its states may be represented by points in a $(C + 1)$-dimensional thermodynamic space; the P-phase states constitute $(C - P + 2)$-dimensional submanifolds in that space.

When two or more phases are in equilibrium, the field variables in the system necessarily have the values given by the coordinates of a corresponding point on one of the lower-dimensional phase-equilibrium manifolds in the $(C + 1)$-dimensional field space, but the converse is not true: for reasons that will be explained in the following section, one cannot, in the laboratory, impose phase equilibrium by fixing field variables alone. It will transpire that for two-phase equilibrium one must fix at least one density, for three-phase equilibrium at least two densities, etc.

4.2 Phase Diagrams

Phase diagrams tell us what may be the equilibrium phases that are present in any given thermodynamic state. A phase diagram may be displayed in a space of fields, of densities, or of mixed fields and densities, and may be of any dimensionality $\leq C + 1$. Most often it is presented as two-dimensional, so that in a system of $C > 1$ components it is then either a section of the full $(C + 1)$-dimensional space in which $C - 1$ fields are fixed or else a projection from that space onto the plane of two of the variables.

A schematic solid (S)–liquid (L)–vapor (V) phase diagram of a one-component system in a space of the two fields p and T is shown in Fig. 8. The one-phase regions are labeled S, L, and V, showing which is the stable

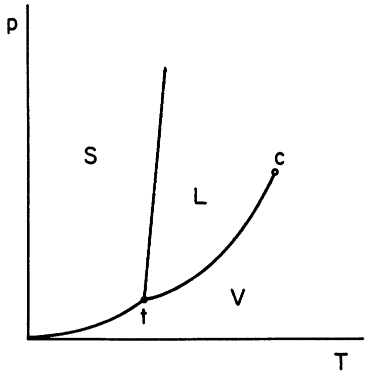

FIG. 8. *p,T* phase diagram of a one-component system with solid (S), liquid (L), and vapor (V) phases. The triple point is marked t and the liquid–vapor critical point *c*.

phase in each region. The boundaries between regions are the two-phase lines of SV, SL, and LV equilibria, and the three lines and three regions come together at the triple point, marked *t*.

The LV coexistence line is the vapor-pressure curve, giving the vapor pressure *p* of the liquid as a function of the temperature *T*. It is also the boiling-point curve, giving the temperature *T* at which the liquid boils under an external pressure *p*. The SL line is the melting-point or freezing-point line, giving the equilibrium melting point *T* of the solid or equilibrium freezing point *T* of the liquid as a function of the pressure *p*. The SV line is the vapor-pressure curve of the solid, giving its vapor pressure *p* as a function of the temperature *T* or its sublimation temperature *T* as a function of the pressure *p*.

The liquid–vapor curve ends in a critical point, marked *c*, where the previously distinct liquid and vapor phases become identical. Only a single homogeneous fluid phase is present at the critical point, but one must not then take the phase rule (70) to read $f = 1 - 1 + 2 = 2$. On the contrary, at the critical point one must take for the number of phases *P* the number of previously distinct phases, 2, and then subtract an additional 1 from the number of degrees of freedom *f* to account for the additional constraint of the two phases' being identical. Hence, $f = 1 - 2 + 2 - 1 = 0$: a critical point in a one-component system is a single, uniquely determined state. In a two-component system, where the liquid–vapor states constitute a surface in the three-dimensional field space, that surface is terminated by a whole line (curve) of critical points; etc.

When the LV curve in Fig. 8 is crossed from the liquid to the vapor side or the reverse, at any point, there is a discontinuity in the properties of the fluid: the density ρ, for example, of the liquid and vapor are different even when they are in equilibrium at the same *p* and same *T*. However, because the LV line ends in a critical point, it is possible to pass continuously, with no sudden change of properties, from what is clearly a liquid state on the liquid side of the LV line to what is clearly a vapor state on the vapor side, by following a path in the *p,T* plane that goes around the end of the LV line instead of crossing it. This is the principle of the continuity of the liquid and vapor states of Andrews and van der Waals, and implies that there is no fundamental distinction between a liquid and a vapor.

The same phases and phase equilibria that are seen in the *p,T* phase diagram in Fig. 8 for a one-component system are shown in a mixed field (*T*) and density (ρ) phase diagram in Fig. 9. What had been lines of two-phase equilibria in Fig. 8 are now regions filled with horizontal tie-lines. The values of ρ at the two ends of each tie-line are the densities of the separate coexisting phases at that temperature.

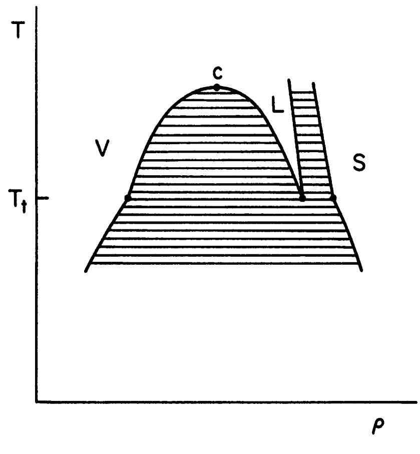

FIG. 9. *T,ρ* phase diagram of a one-component system with solid (S), liquid (L), and vapor (V) phases. The triple-point temperature is T_t; the liquid–vapor critical point is at *c*.

When two phases coexist, the average density in the system lies between that of the separate phases; a point ρ,T in any of the three two-phase regions in Fig. 9 represents that average density at the temperature T, and the exact location of that point on the tie-line determines the relative amounts of the two phases. If $\rho^{(\alpha)}$ and $\rho^{(\beta)}$ with $\rho^{(\alpha)} < \rho^{(\beta)}$ are the densities of the separate phases at some temperature T, and ρ is the mean density in the system, then the point ρ, T divides the tie-line into a segment of length $\rho - \rho^{(\alpha)}$ and one of length $\rho^{(\beta)} - \rho$. If the masses and volumes of the two phases are $M^{(\alpha)}$, $V^{(\alpha)}$ and $M^{(\beta)}$, $V^{(\beta)}$ then $\rho^{(\alpha)} = M^{(\alpha)}/V^{(\alpha)}$ and $\rho^{(\beta)} = M^{(\beta)}/V^{(\beta)}$. The total mass and volume of the system are $M = M^{(\alpha)} + M^{(\beta)} = \rho^{(\alpha)}V^{(\alpha)} + \rho^{(\beta)}V^{(\beta)}$ and $V = V^{(\alpha)} + V^{(\beta)}$, while $\rho = M/V$. From these relations it follows that

$$V^{(\alpha)}/V^{(\beta)} = (\rho^{(\beta)} - \rho)/(\rho - \rho^{(\alpha)}), \tag{71}$$

expressing the relative amounts (volumes) of the two coexisting phases in terms of the relative lengths of the two segments into which the point ρ,T divides the tie-line. Thinking of the tie-line as a lever and the point ρ,T as the position of a fulcrum on which the lever rests, with $V^{(\alpha)}$ and $V^{(\beta)}$ the "weights" at the two ends at the respective distances $\rho - \rho^{(\alpha)}$ and $\rho^{(\beta)} - \rho$ from the fulcrum, we have (71) as the condition of balance; hence the name *lever rule* for that relation. If the "density" variable had been the specific volume $v = 1/\rho$ instead of the mass density ρ, the rule would have been $M^{(\alpha)}/M^{(\beta)} = (v - v^{(\beta)})/(v^{(\alpha)} - v)$, expressing the relative amounts of the two phases as the ratio of their masses rather than of their volumes.

If in this one-component system one tried, in the laboratory, to create a two-phase state by fixing p and T (or any other pair of field variables) at values corresponding to a point on one of the two-phase lines in Fig. 8, the attempt would fail for reasons both of practice and of principle. As a practical matter it would fail because one cannot fix laboratory variables with infinite precision and so cannot force the system's state to be exactly on the coexistence line; inevitably, it would be slightly to one side or the other, in one of the one-phase regions. As a matter of principle it would fail because if only field variables and no density variable were fixed, only that phase that had the lower free energy of contact with the container walls would be present, since if both phases were present there would be both a less favorable (more positive) free energy of contact with the wall and an additional positive free energy (interfacial tension) of contact of the two phases with each other. It is seen in Fig. 9 how by specifying a field and a density instead of two fields one may now impose the coexistence of two phases. One need only specify, with practical laboratory precision, that the temperature T and overall average density ρ in the system lie somewhere in the now two-dimensional two-phase *region*, and at equilibrium there will then be two phases with an interface between them.

The same phases and phase equilibria in this one-component system are depicted again, in Fig. 10, now in the plane of two densities—the mass density ρ and energy density u in this illustration. The two-phase regions are again filled with tie-lines connecting the densities $u^{(\alpha)}$, $\rho^{(\alpha)}$ and $u^{(\beta)}$, $\rho^{(\beta)}$ of coexisting phases, while the triple point—a *point* in Fig. 8 and a *line* in Fig. 9—is now a two-dimensional *region;* it is the triangle in the middle of the figure. The coordinates of the vertices of the triangle are the energy and mass densities $(u^{(\alpha)},\rho^{(\alpha)})$, $(u^{(\beta)},\rho^{(\beta)})$, and

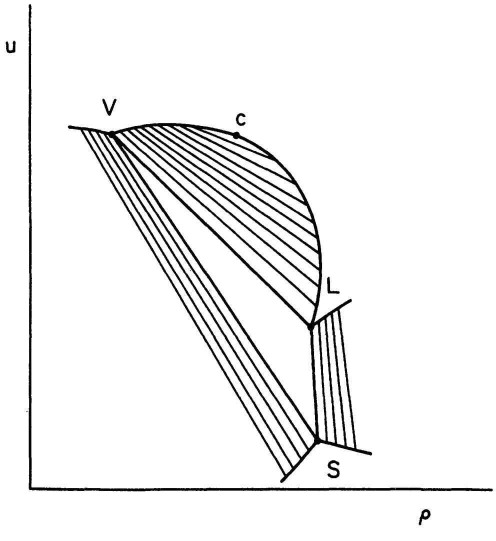

FIG. 10. u,ρ phase diagram of a one-component system with solid (S), liquid (L), and vapor (V) phases. The triangle is the region of the three-phase coexistence (triple point); the liquid–vapor critical point is at *c*.

$(u^{(\gamma)},\rho^{(\gamma)})$ of the three coexisting phases α, β, and γ. Neither by specifying two fields, as in Fig. 8, nor one field and one density, as in Fig. 9, could one have imposed three-phase coexistence; but by imposing two densities—requiring that the overall average energy and mass densities u,ρ in the system lie somewhere within the three-phase triangle in Fig. 10—the system at equilibrium would be required to consist of three coexisting phases. Wherever in the three-phase triangle the imposed u,ρ lies, i.e., as long as all three phases are present in equilibrium, the system will automatically adjust its pressure and temperature to be exactly the unique triple-point p,T in Fig. 8. The location of the point u,ρ in the three-phase triangle determines the relative amounts of the three phases by the three-phase analog of the lever rule (71):

$$\frac{V^{(\alpha)}}{V^{(\beta)}} = \frac{\begin{vmatrix} u - u^{(\beta)} & u^{(\gamma)} - u \\ \rho - \rho^{(\beta)} & \rho^{(\gamma)} - \rho \end{vmatrix}}{\begin{vmatrix} u^{(\alpha)} - u & u^{(\gamma)} - u \\ \rho^{(\alpha)} - \rho & \rho^{(\gamma)} - \rho \end{vmatrix}}, \text{ etc.} \qquad (72)$$

For a lever rule like (71) to hold when ρ is a density in a general sense, it is necessary that ρ be of the form X/Z where each of X and Z is an extensive function whose value in a *heterogeneous* system is the sum of its values in the separate phases. (U, V, and M are such extensive functions while V^2/M is not, even though in a *homogeneous* system it is proportional to the system's size.) Then (71) is more generally a formula for $Z^{(\alpha)}/Z^{(\beta)}$. This was illustrated in (71) itself with $\rho = M/V$ and subsequently with the "density" $v = V/M$. For the tie-lines in the two-phase regions of figures like Fig. 10 to be linear, i.e., for the overall average u,ρ to lie on the line connecting $u^{(\alpha)}$, $\rho^{(\alpha)}$ and $u^{(\beta)}$, $\rho^{(\beta)}$, and for the three-phase analog (72) of the lever rule (71) to hold, it is necessary that the "densities" u,ρ be of the compatible form $u = Y/Z$ and $\rho = X/Z$, where X, Y, and Z are all extensive in the specific sense defined above. The energy and mass densities $u = U/V$ and $\rho = M/V$ are such a compatible pair, as are the specific energy and volume $u' = U/M$ and $v = V/M$, while u',ρ and u,v are both incompatible pairs.

Where two phases coexist in a one-component system, as in any of the three two-phase regions in Fig. 9, the composite system's constant-pressure heat capacity C_p, compressibility χ, and coefficient of thermal expansion α are all infinite: If a small amount of heat is added to the system at fixed pressure, then a proportional amount of phase change (S → V, S → L, or L → V in the example in Fig. 9) occurs, the absorbed heat being just the enthalpy change ("latent heat") that is required to accompany the transition; and, because in the two-phase region the temperature is fixed by the pressure (Fig. 8), there is no accompanying increase in temperature; hence, $C_p = \infty$. Likewise, if for a fixed mass the volume of the two-phase system is decreased isothermally, all that happens is that some of the less dense phase transforms into some of the more dense phase, but the pressure remains unchanged; hence $\chi = \infty$. Meanwhile, $\alpha = V^{-1}(\partial V/\partial T)_p = -V^{-1}(\partial p/\partial T)_V(\partial V/\partial p)_T = \chi(\partial p/\partial T)_V$; and in the two-phase region $(\partial p/\partial T)_V$ is just dp/dT, the (nonzero) slope of the corresponding equilibrium curve in Fig. 8 at that point; so $\alpha = \infty$, too.

The constant-volume heat capacity C_V, by contrast, is finite in the two-phase region. This, too, is clear from Fig. 9: if heat is added at constant volume (fixed ρ), there is again a proportionate amount of phase change, but it is now accompanied by a proportionate temperature rise, and so C_V is finite. That C_p, χ, and α are infinite in the two-phase region is still true at the liquid–vapor critical point (point c in Fig. 9), which is the limit point of that two-phase coexistence. Remarkably, as one knows from statistical-mechanical theory and experiment, C_V is also infinite at the critical point—although it diverges to ∞ very slowly as the critical point is approached, and so at moderate distances from the critical point is not very large, as befits there having been no thermodynamic necessity for it to diverge there at all. At three-phase coexistence, on the other hand, C_V is necessarily infinite, as one may visualize in Fig. 10: adding a small amount of heat at fixed overall density ρ increases u and causes the relative amounts of the three phases to change in accordance with (72), but there is no accompanying rise in temperature, which remains fixed at that of the triple point.

It is seen in Fig. 8 that the angles between the three two-phase equilibrium lines where they meet at the triple point are all less than

180°. This illustrates the "180° rule." The $\alpha\beta$ equilibrium line in the p,T plane, including the extension of that line beyond the triple point, is defined by $\mu^{(\alpha)}(p,T) = \mu^{(\beta)}(p,T)$. If the angle occupied by the α phase, say, in the diagram, were greater than 180°, this extension of the $\alpha\beta$ line, on which $\mu^{(\alpha)} = \mu^{(\beta)}$, would lie in the α region. But since this is the region in which α is supposed to be the stable phase, one must have $\mu^{(\alpha)} < \mu^{(\beta)}$ at every point p,T in this region; so in fact the supposition that the angle occupied by the α phase could be greater than 180° is inadmissible. That is likewise true for the other two angles: the extension of any of the two-phase lines $\alpha\beta$ must lie in the region of stability of the third phase γ. On the extension of the $\alpha\beta$ equilibrium line beyond the triple point into the region of stability of the γ phase, the phases α and β may still be in equilibrium with each other, but they are then unstable with respect to γ, to which they would ultimately transform irreversibly. There are more rigorous derivations of the 180° rule that do not make reference to metastable extensions of the equilibrium lines (Wheeler, 1974; Fisher, 1990; Teubner, 1991).

The 180° rule in a two-dimensional field space such as the p,T plane of Fig. 8 has its analog in a density space such as that in Fig. 10, where it is known as Schreinemakers' rule. First, the angles formed by the tangents to the boundaries of the one-phase regions at each vertex of the three-phase triangle (the angles labeled S, L, V in Fig. 10) must each be less than 180°. Second, when the two tangent lines forming any of those angles are continued beyond their intersection, they must either both go inside the three-phase triangle (as at L in Fig. 10) or both go outside (as at V and at S in the figure) but not one inside and the other outside. There are analogous conditions for the mixed field and density diagram of Fig. 9, but these are less often referred to.

The transitions that occur across the two-phase lines in Fig. 8 are said to be of first order, and with the appropriate constraints (Figs. 9 and 10) may be associated with phase coexistence, as discussed. In Fig. 11 is a schematic version of the p,T phase diagram of ^{4}He. The regions II and I are superfluid and nonsuperfluid liquid. There is no solid–liquid–vapor triple point. The curves separating solid from liquid (I or II) and liquid from vapor in Fig. 11 are loci of ordinary first-order transitions, but the line between liquids I and II is not. The I–II transition is said to be of second order; even if densities were prescribed instead of fields, the liquid phases I and II could not be made to coexist as two separate phases with an interface between them. The I–II–V and I–II–S junctions are therefore not triple points.

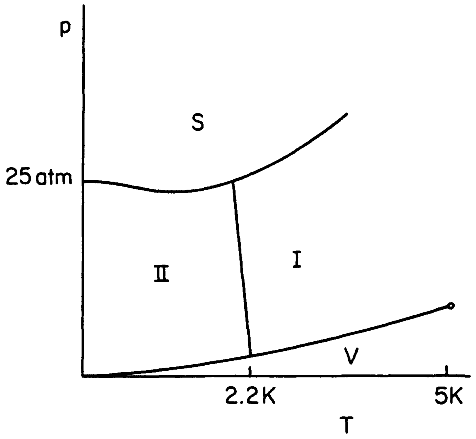

FIG. 11. Phase diagram of ^{4}He (schematic), showing solid (S), vapor (V), superfluid liquid (II), and nonsuperfluid liquid (I) regions. The boundary between I and II is a line of second-order transitions.

In general, a first-order transition is accompanied by discontinuities in densities u, ρ, etc., but at a second-order transition the densities are continuous: at the I–II boundary in Fig. 11 the two liquids have the same entropy, energy, and mass densities. The slopes of the vapor–liquid and solid–liquid phase boundaries are continuous at the respective junctions I–II–V and I–II–S: the angle between the I–V and II–V boundaries and that between the I–S and II–S boundaries are 180°, which is thus a limiting case of the 180° rule. Instead of discontinuities in the densities at the I–II transition line, there are now singularities in C_p, χ, and α that are much like the divergence in C_V at a liquid–vapor critical point that was remarked above. Because of a fancied resemblance to the letter "lambda" of a plot of C_p against T, such second-order transitions are often called lambda points; the I–II boundary in Fig. 11 is a locus of such points.

4.3 Clapeyron Equations

The Gibbs–Duhem equation in the form (53) relates to each other the infinitesimal changes in the fields p, T, $\mu_1, \ldots$ that result from an infinitesimal change in the state of a system. When two phases are in equilibrium and an infinitesimal change is made in the two-phase state, each of the phases must undergo the same infinitesimal change in its pressure because this has to have the same value in both phases after as well as before the change (assuming planar interfaces). Thus, if the two phases be called α and β, one has from (53)

$$0 = (s^{(\alpha)} - s^{(\beta)})dT + (\rho_1^{(\alpha)} - \rho_1^{(\beta)})d\mu_1 + \cdots, \tag{73}$$

relating to each other the changes dT, $d\mu_1, \ldots$ in the other field variables. In a one-component system (73) gives the slope of the two-phase equilibrium curve in the μ,T plane in terms of the differences $s^{(\alpha)} - s^{(\beta)}$ and $\rho^{(\alpha)} - \rho^{(\beta)}$ in the entropy and mass densities of the two phases:

$$\frac{d\mu}{dT} = -\frac{s^{(\alpha)} - s^{(\beta)}}{\rho^{(\alpha)} - \rho^{(\beta)}}. \tag{74}$$

The same Gibbs–Duhem equation (53) could have been rewritten to express dT in terms of dp, $d\mu_1, \ldots$ or $d\mu_1$ in terms of dp, dT, $d\mu_2, \ldots$, etc., and so in a C-component system there are $C + 2$ equations of the form of (73), each one relating the infinitesimal changes in $C + 1$ of the field variables to each other when an infinitesimal change is made in a two-phase equilibrium. Another $C + 2$ equations of the same general form as (73) follow when the Gibbs–Duhem equation is taken in the form (55) with p/T, $1/T$, $\mu_1/T, \ldots$ as the field variables. All of these, in a one-component system, reduce to a relation like (74), expressing the slope of the two-phase equilibrium curve in some two-dimensional field space in terms of the ratio of differences of "densities," with "density" understood in a general sense.

For example, if we represent specific entropy and volume by $\sigma = s/\rho$ and $v = 1/\rho$, then another form of the Gibbs–Duhem equation (53) for a one-component system is $d\mu = vdp - \sigma dT$, so that an alternative to (74), giving the slope of the equilibrium curve in the p,T plane, is

$$\frac{dp}{dT} = \frac{\sigma^{(\alpha)} - \sigma^{(\beta)}}{v^{(\alpha)} - v^{(\beta)}}. \tag{75}$$

All the relations of the form of (73) in a multicomponent system or their special cases like (74) and (75) in a one-component system may collectively be called Clapeyron equations. Of these, (75) is the most familiar and most used, but none is more fundamental than another.

The phase change $\beta \rightarrow \alpha$ at a point on the $\alpha\beta$ equilibrium curve in the p,T plane may be made to occur reversibly, with the phases in equilibrium at every stage and with p and T both fixed. The heat ("latent heat") absorbed or evolved in the process is then both the change in the system's enthalpy, ΔH (because p is constant), and T times the change in its entropy, ΔS (because the process is isothermal and reversible). Therefore $\Delta H = T\Delta S$ in the process; so the difference $\sigma^{(\alpha)} - \sigma^{(\beta)}$ in the specific entropies (entropies per unit mass) of the two phases is also $(h^{(\alpha)} - h^{(\beta)})/T$, where $h = H/M$ is specific enthalpy. Then (75) is equally well

$$\frac{dp}{dT} = \frac{h^{(\alpha)} - h^{(\beta)}}{T(v^{(\alpha)} - v^{(\beta)})}, \tag{76}$$

another frequently used form of the Clapeyron equation for a one-component system.

Heat is usually absorbed in the melting of a solid (an interesting exception being helium at very low temperatures), and always in the evaporation of a liquid or solid, and so for those transformations $\beta \rightarrow \alpha$ the differences $h^{(\alpha)} - h^{(\beta)}$ and $\sigma^{(\alpha)} - \sigma^{(\beta)}$ are both positive. The volume increases in the evaporations so that $v^{(\alpha)} - v^{(\beta)}$ is then also positive, implying positive slope dp/dT for the SV and LV equilibrium lines in Fig. 8. The difference $v^{(\alpha)} - v^{(\beta)}$ on melting may be positive or negative depending on whether the liquid or the solid is the denser. More often the solid is, and so if, as is usual, it is also of lower entropy and enthalpy, then dp/dT is again positive on the melting line, as in Fig. 8. An important exception is water, for which the density of the liquid is greater than that of the solid, and so in the analog of Fig. 8 for water the SL equilibrium line has negative rather than positive slope; and another exception, for the reason already stated, is helium at very low temperatures, as seen in the schematic Fig. 11.

The infinite values that the constant-pressure heat capacity C_p and compressibility χ have at two-phase coexistence (Sec. 4.2) are both $+\infty$, by the stability conditions $C_p > 0$ (Sec. 3.4) and $\chi > 0$ (Sec. 3.2). Then since $\chi = (\partial \ln \rho/\partial p)_T$ [from the definition of χ in (47), with $\rho = M/V$] and $C_p = (\partial H/\partial T)_p = (\partial S/\partial \ln T)_p$ [Eq. (59)], it follows that at a two-phase line in a p,T phase diagram the phase that lies in the direction of increasing T as the transition line is crossed has both the greater entropy and the greater enthalpy of the two, while the phase that lies in the direction of increasing p as the line is crossed has the greater density (smaller specific volume). These rules are consistent with the sign of dp/dT as implied by (76).

Density variables such as the entropy, energy, and mass densities $s = S/V$, $u = U/V$, $\rho_1 = M_1/V, \ldots$, the specific entropy and energy S/M and U/M, and the mass fractions M_1/M, $M_2/M, \ldots$ of the several components, where V and M are the total volume and mass of a phase, all generally have different values in coexisting phases. Those differences occur in the general Clapeyron equations like (73) or in the special cases of those equations like (74)–(76). When, exceptionally, a density variable has the same value in different phases at equilibrium, the event is called an *azeotropy*.

Each azeotropic constraint such as $\rho^{(\alpha)} = \rho^{(\beta)}$ decreases by one the number of degrees of freedom of a system, and so in a one-component system such states can only occur as isolated points on the two-phase equilibrium lines in the system's thermodynamic plane (its p-T plane, for example). Figure 12 is a schematic representation of the phase diagram of elemental carbon, showing the regions of stability of graphite (gr), diamond (dia), and liquid (liq) in the p-T plane (Bundy, 1989). The region of stability of the vapor is invisible on this scale. There is a gr–dia–liq triple point, marked t on the diagram. The liquid always has a greater entropy and enthalpy than either solid, the graphite in particular; so by the Clapeyron equation for a one-component system, (75) or (76) (or by the principle that the denser phase is the one that lies in the direction in which p increases as the two-phase line is being crossed), at any point on the gr–liq equilibrium curve at which $dp/dT > 0$, the graphite is the denser phase, whereas where $dp/dT < 0$, the liquid is the denser. Where $dp/dT = \pm\infty$ ($dT/dp = 0$), the two densities are equal: $v^{(\mathrm{gr})} = v^{(\mathrm{liq})}$. That is an azeotropy. It must thus occur at a temperature extremum ($dT = 0$) in this representation and is seen here to be at a temperature maximum, at the point marked az in the figure.

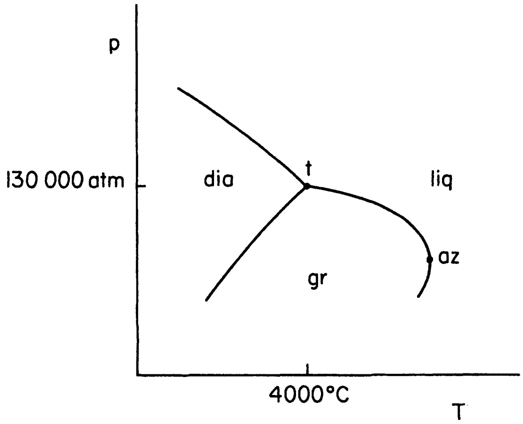

FIG. 12. Phase diagram of carbon showing regions of stability of graphite (gr), diamond (dia), and liquid (liq). The gr–dia–liq triple point is marked t; the azeotropic point on the gr–liq equilibrium curve is marked az.

Azeotropes are more commonly found in multicomponent systems because in a mixture of $C > 1$ components azeotropes occur as whole $(C - 1)$-dimensional manifolds of states instead of as isolated points. Higher-order azeotropies, where some density has equal values simultaneously in more than two coexisting phases, or where, in a two-phase equilibrium, one has $\rho_i^{(\alpha)} = \rho_i^{(\beta)}$ for more than one density i, occur on manifolds of correspondingly lower dimensionality and are therefore more difficult to achieve and are of rarer occurrence.

The *simultaneous* equalities $\rho^{(\alpha)} = \rho^{(\beta)}$ and $s^{(\alpha)} = s^{(\beta)}$ (or $u^{(\alpha)} = u^{(\beta)}$) in a *one*-component system (as at the λ-line in liquid ^{4}He; Sec. 4.2) define a second-order transition; they are not associated with the coexistence of distinct phases and so are not an azeotropy in the usual sense.

5. DILUTE SYSTEMS

Any relation among thermodynamic functions that is specific to the system and not a mere thermodynamic identity may be called an equation of state, a term that was used before for the expression of S as a function of $U, V, M_1, \ldots$ in (26). Such equations of state are determined by experiment or by the methods of statistical mechanics, which lie beyond classical thermodynamics. These equations often refer to the molecular constitution of the system. When the equation is that for a thermodynamic potential such as $S = S(U,V,M_1,M_2,\ldots)$ or $p = p(T,\mu_1,\mu_2,\ldots)$, all the thermodynamic properties of the system (in the latter case, all the intensive properties) are derivable from it (Sec. 3.3).

An example is the expression for p as a function of $T, \mu_1, \mu_2, \ldots$ for an *ideal gas,*

$$p = RT \sum_i c_i^0(T) \exp(m_i\mu_i/RT), \tag{77}$$

where m_i is the molecular weight of constituent i and $c_i^0(T)$ is some function of the temperature alone, with the dimensions of a molar concentration (moles per unit volume), characteristic of that constituent. The gas constant R is a universal constant, $R = 8.315$ J/mol·K. In molecular terms $R = N_0k$, where $N_0 = 6.022 \times 10^{23}$/mol is Avogadro's number and $k = 1.381 \times 10^{-23}$ J/K is Boltzmann's constant, but (77) itself is a relation among macroscopic quantities only. It holds accurately for real gases only when these are dilute. It is derivable as the equation of state of a gas of noninteracting molecules from statistical mechanics, which also allows the explicit calculation of the functions $c_i^0(T)$ characteristic of each species i.

From (54) and (77) one finds the density ρ_i of component i, or equivalently its molar concentration $c_i = \rho_i/m_i$,

$$c_i = c_i^0(T) \exp(m_i\mu_i/RT), \tag{78}$$

so that

$$\mu_i = \frac{RT}{m_i} \ln \frac{c_i}{c_i^0(T)}. \tag{79}$$

From (79) one sees that the chemical potential of any constituent of an ideal gas depends logarithmically on its concentration and, at any fixed T, is independent of the concentrations—indeed, independent of the presence or absence—of any other species. The molecular origin of this independence is the near absence of interactions among the molecules of a dilute gas.

From (77) and (78),

$$p = RT \sum_i c_i = nRT/V, \tag{80}$$

where $n = n_1 + n_2 + \cdots$ is the total number of moles of gas. Equation (80) is the "ideal-gas law." The pressure p as given by (80) is an example of a *colligative* property: it expresses p at any T as being proportional to the total molar concentration n/V of the gas whatever its chemical constitution. Thus, every molecule makes the same average contribution to p (at given V and T) whatever its nature. There are analogous colligative properties of dilute solutions, as will be noted later.

From (80) and the definitions (47) and (60), together with (61), one has for an ideal gas

$$\chi = \frac{1}{p}, \quad \alpha = \frac{1}{T}, \quad \left(\frac{\partial p}{\partial T}\right)_V = \frac{p}{T} = \frac{nR}{V}. \tag{81}$$

From (80) and (63) one finds $(\partial U/\partial V)_T = 0$ and $(\partial H/\partial p)_T = 0$; i.e., the energy and enthalpy of an ideal gas depend on temperature alone, and for fixed temperature are then constant, independent of the gas's state of compression. From the identities (31) and (59) the same is then also true of the heat capacities C_V and C_p; and from (64) and (81) one has for their difference

$$C_p - C_V = nR, \tag{82}$$

depending only on the total number of moles of gas and neither on the chemical species nor on the temperature.

From (80), the equation of an isotherm (T = const.) in the p-V plane of an ideal gas is pV = const. (Boyle's law). One may obtain a corresponding equation of an isentrope (reversible adiabat). From the definition of the isentropic compressibility χ_S that appears in (67), and from that equation and the first of Eqs. (81), one has as a differential equation for an isentrope in the p-V plane of an ideal gas, $d \ln p + \gamma d \ln V = 0$ with $\gamma = C_p/C_V$

now a function of the temperature alone. While the difference $C_p - C_V$ of the principal heat capacities is a constant for an ideal gas [Eq. (82)], their ratio γ, in general, depends on the temperature and so varies along the isentrope, but it is often nearly enough constant so that to a good approximation the differential equation may be integrated to yield pV^γ = const. as the explicit equation of the isentrope in the p-V plane. From (80), the corresponding equations in the T-V and p-T planes are then $TV^{\gamma-1}$ = const. and $T^\gamma/p^{\gamma-1}$ = const., respectively.

The ideal-gas law combined with the Clapeyron equation (76) leads to an important approximation for the liquid–vapor and solid–vapor equilibrium lines in Fig. 8. Let α in (76) be the vapor and β the liquid or solid. Then, except near the critical point on the liquid–vapor line, $v^{(\alpha)} \gg v^{(\beta)}$; and the vapor is typically dilute enough to be a nearly ideal gas, so that by (81), $mv^{(\alpha)} = RT/p$, the molar volume of the vapor, where m is the molecular weight of the *vapor*. Thus, from (76),

$$\frac{d\ln p}{dT} = \frac{m\Lambda}{RT^2}, \qquad \frac{d\ln p}{d(1/T)} = -\frac{m\Lambda}{R}, \tag{83}$$

where $\Lambda = h^{(\alpha)} - h^{(\beta)}$ is the heat of vaporization per unit mass. It is significant that the molecular weight m is that of the vapor, not of the liquid or solid, for the molecular weight of a condensed phase, especially of a liquid of strongly associating molecules like water, is often a meaningless concept, whereas the molecular weight of a dilutely dispersed species, where the molecules are separated by large distances and essentially noninteracting, is well defined.

Equation (83), called the Clausius–Clapeyron equation, is frequently applied to estimate p at one temperature when it is known at another and m and Λ are known: if Λ is nearly constant over the temperature range T_0 to T, then $p = p_0 \exp[-(m\Lambda/R)(1/T - 1/T_0)]$, where p and p_0 are the vapor pressures at T and T_0, respectively. This is an approximate equation of the vapor-pressure curve, p vs T, parametrized by the two parameters $p_0 \exp(m\Lambda/RT_0)$ and $m\Lambda$.

The ideal-gas law (80) is no longer accurate when the gas is not dilute. Corrections to it are often expressed as power series in the density called *virial series*. For a one-component gas one has

$$p = \frac{nRT}{V}\left[1 + B(T)\frac{n}{V} + C(T)\left(\frac{n}{V}\right)^2 + \cdots\right], \tag{84}$$

where $B(T)$, $C(T)$, . . . , which are functions of the temperature alone characteristic of the substance, are called the second, third, . . . *virial coefficients*. This reduces to the ideal-gas law in the limit of vanishing density, $n/V \to 0$.

A typical plot of the second virial coefficient B as a function of T is shown schematically in Fig. 13. B is negative at low temperatures and becomes rapidly more negative as the temperature decreases, while it is positive at high temperatures and then increases only slowly with increasing temperature. The temperature T_B at which $B(T) = 0$ is called the Boyle temperature: it is the unique temperature at which the deviation of pV/nRT from 1 is proportional to the second rather than the first power of n/V at low densities, and so the temperature at which Boyle's law pV = const. remains accurate to the highest densities. Negative values of $B(T)$ are manifestations mainly of the attractive forces that molecules exert on each other when they are separated by moderate distances, while the positive $B(T)$ are manifestations mainly of the strong intermolecular repulsions at short distances. At $T = T_B$ the intermolecular attractions that tend to decrease the pressure p and the repulsions that tend to increase it

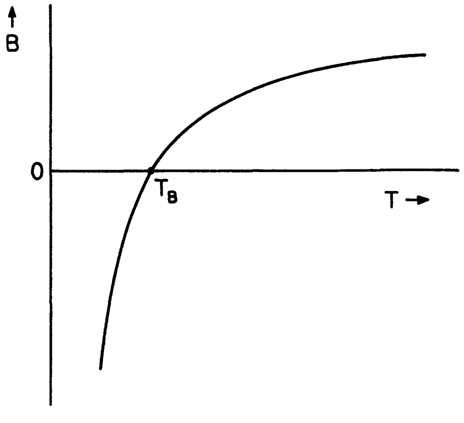

FIG. 13. Second virial coefficient $B(T)$ as a function of T, with T_B the Boyle temperature.

are roughly in balance. The Boyle temperatures of helium (^{4}He), hydrogen ($^1H_2$), and nitrogen (N_2) are 21, 110, and 327 K, respectively.

While U and H of an ideal gas depend only on the temperature and are independent of the state of compression, that is no longer true when the correction terms in (84) are significant. From the virial expansion (84) and the identities (63), with the coefficient of thermal expansion α defined in (60), one finds, to leading order in n/V,

$$\left(\frac{\partial U}{\partial V}\right)_T = RT^2\frac{dB}{dT}\left(\frac{n}{V}\right)^2 + O\left[\left(\frac{n}{V}\right)^3\right],$$
$$\left(\frac{\partial H}{\partial p}\right)_T = n\left[\frac{d(B/T)}{d(1/T)} + O\left(\frac{n}{V}\right)\right]. \tag{85}$$

When a gas expands into an evacuated, thermally insulated space (Joule expansion), no work is done or heat absorbed, $w = q = 0$, so that $\Delta U = q + w = 0$: the expansion is at constant energy. The rate of change of temperature with volume in such an expansion is then

$$\left(\frac{\partial T}{\partial V}\right)_U = -\frac{1}{C_V}\left(\frac{\partial U}{\partial V}\right)_T, \tag{86}$$

by (31). Then from the first of Eqs. (85) and the positivity of C_V, one sees that for a dilute gas the sign of this quantity is opposite that of dB/dT, hence negative (Fig. 13): the gas cools in the expansion.

When, instead, the gas is forced through a porous plug from a region of a constant high pressure p_1, where it occupies a volume V_1 initially, to a region of a constant low pressure p_2, where it occupies a volume V_2, but with the whole apparatus thermally insulated so that $q = 0$, the net work w done on the system is $p_1V_1 - p_2V_2$ and the net energy change is then $\Delta U = q + w = p_1V_1 - p_2V_2$. Then the enthalpy change is $\Delta H = \Delta U + \Delta(pV) = \Delta U + p_2V_2 - p_1V_1 = 0$. This expansion, called a Joule–Thomson expansion, is thus isenthalpic. The rate of change of temperature with pressure in such an expansion is then

$$\left(\frac{\partial T}{\partial p}\right)_H = -\frac{1}{C_p}\left(\frac{\partial H}{\partial p}\right)_T, \tag{87}$$

by the first of Eqs. (59). This quantity is called the Joule–Thomson coefficient. From the second of Eqs. (85) and the positivity of C_p, one sees that for a dilute gas its sign is opposite that of $d(B/T)/d(1/T)$, hence the same as that of $d(B/T)/dT$. From Fig. 13 one sees that since B increases only very slowly with increasing T at high temperatures, B/T vs T has positive slope for temperatures up to some T_i that is greater than T_B and negative slope for $T > T_i$. Thus, a dilute gas undergoing a Joule–Thomson expansion ($dp < 0$) cools ($dT < 0$) when $T < T_i$ and warms ($dT > 0$) when $T > T_i$. The temperature T_i is called the Joule–Thomson inversion temperature. For nitrogen $T_i = 633$ K, or about twice the Boyle temperature, which is typical.

When the intermolecular forces are of very long range, such as Coulomb forces, the deviations from ideality are no longer of the form (84). From the Debye–Hückel theory applied to a fully ionized gas (a plasma), analogous to the D–H theory of an electrolyte solution, one knows that the leading term in $pV/nRT - 1$ is proportional not to n/V, as in (84), but to $(n/V)^{1/2}$, and so the deviations from ideal-gas behavior begin at much lower densities than in gases whose molecules interact with short-range forces.

Another important class of dilute systems comprises liquid or solid mixtures in which one component (the "solvent") is present in large excess while all the other components (the "solutes") are dilutely dispersed in the solvent. When the solution is so dilute that the solute molecules interact very little with each other, we have the analog of the ideal gas, now called an ideal-dilute solution. Both from statistical-mechanical theory and from experiment one learns that the chemical potential μ_i of solute species i in such a solution is again given by (79), as in an ideal gas,

$$\mu_i = \frac{RT}{m_i}\ln\frac{c_i}{c_i^0}, \tag{88}$$

independently of the concentrations (or of the presence or absence) of any other solute species in the solution; but now c_i^0 not only depends on the chemical nature of the solute i and on the temperature, as in (79), but also depends on the nature of the solvent and on the pressure p. Equation (88) for the solute species i is called Henry's law. From (88) and the Gibbs–Duhem equation in the form

(53) one finds that the chemical potential μ_{solv} of the solvent depends on the total molar concentration $c = \Sigma_i c_i$ of all the solutes, according to

$$\mu_{\text{solv}} - \mu^0_{\text{solv}} = -RTc/\rho^0_{\text{solv}}, \tag{89}$$

where ρ^0_{solv} and μ^0_{solv} are the mass density and chemical potential of the pure solvent at the temperature T and pressure p of the solution. This is called Raoult's law.

The molar concentration c_i is conventionally called the molarity. An alternative measure of solute concentration, called molality, is the number of moles of solute per unit mass of solvent. In a dilute solution, which is almost all solvent, the ratio c/ρ^0_{solv} that appears in (89) is then the total solute molality.

All the properties of ideal-dilute solutions are derivable from (88) and (89). In particular, one may derive a group of colligative properties, which, analogously to the ideal-gas law, are properties that are proportional to the number of moles or molecules (the molar concentrations) of the dilute solute species but independent of their kind. By measuring any such colligative property one may determine the molecular weight m of the solute, since the measurement itself then leads to a knowledge of the solute's molar concentration c while the mass density ρ of the solute is known from how the solution was prepared, and then $m = \rho/c$. Since the solute species in question is dilute, its molecules, and so its molecular weight, are well defined—unlike those of the solvent. [Compare with the similar remark made in connection with the Clausius–Clapeyron equation, Eq. (83).]

The colligative properties of dilute solutions include the lowering of the solvent's vapor pressure, the elevation of the boiling point by nonvolatile solutes, and the depression of the freezing point when the solid that separates from the freezing solution is pure solid solvent.

Another colligative property of a dilute solution is its osmotic pressure. When a solution in one compartment is separated from pure solvent in another by a membrane that is freely permeable to solvent but impermeable to the solutes (which may be large polymer molecules, for example—too large to pass through the pores in the membrane), solvent will flow spontaneously from the pure solvent into the solution, to dilute it and thus tend to equalize the chemical potentials on the two sides. The process is called osmosis. The flow may be reversed by pressing on the solution with a pressure sufficiently in excess of that on the pure solvent; this is reverse osmosis. At one value of the excess pressure—termed the osmotic pressure, π—the chemical potential of the solvent will be the same on both sides of the membrane, and there will be no flow of solvent in either direction. This condition, with the dilute-solution law (89) and the known pressure dependence of the Gibbs free energy as given by (42), yields

$$\pi = RTc. \tag{90}$$

This is the law of van 't Hoff. It says that the osmotic pressure of a dilute solution of some total solute concentration is the same as the pressure of an ideal gas of that same concentration [Eq. (80)]. As with any other colligative property, as noted above, it may be used to determine the molecular weight of the dilute species. Measurement of the osmotic pressure is one of the principal methods of determination of the molecular weights of both synthetic and biological macromolecules in solution. According to (90) a solution of a polymer of molecular weight 50 000 g/mol at a concentration of 20 g/L would have an osmotic pressure at 25 °C of a hundredth of an atmosphere. That is a pressure of 10 cm of water, which is easily measurable.

6. THE THIRD LAW

The "third law of thermodynamics," unlike the first and second laws, does not introduce any new functions of state but rather asserts a universal behavior of the entropy at low temperatures. The most useful form of the third law is the Nernst "heat theorem."

Suppose a system undergoes a change in its thermodynamic state at a fixed temperature T (isothermal change) and that the accompanying entropy change is ΔS_T. The change could be a chemical reaction, a change of physical state (melting, for example, or transformation between two different crystalline forms), an isothermal expansion or compression, etc. We ask only that the system in both its initial and final states be in a condition of (complete or constrained) equilibrium so that it would be possible in principle to carry out the process reversibly, however it may actually be carried out, and so that the system has well-defined values of its thermodynamic functions both initially and finally. At the low temperatures of interest here the relevant processes are almost always those in which the system is a solid in both its initial and final states. Helium below 25 atm is an exception; see Fig. 11.

The third law of thermodynamics asserts as an empirical fact that as the temperature T at which such an isothermal process occurs approaches 0, so does the accompanying entropy change ΔS_T:

$$\lim_{T\to 0} \Delta S_T = 0. \tag{91}$$

This is the Nernst heat theorem. It says that the entropies in the initial and final states become equal at $T = 0$. This is true whether the change be merely one of physical state (crystal structure or density) or a change of chemical identity (chemical reaction). For all such states that are in principle transformable one into the other by a reversible process, physical or chemical, one may arbitrarily assign a value of the entropy in one of them at $T = 0$, and then, by (91), that will also be the value of the entropy in all the others at $T = 0$. By a generally accepted convention, that universal value of the entropy at $T = 0$ is taken to be 0. For substances that are not transformable into each other by chemical reaction that convention may be, and is, applied separately to each. Thus, one arrives at the statement that the entropies of substances in internal equilibrium, i.e., in well-defined thermodynamic states, are all 0 at $T = 0$, but it should be understood that the "0" for the value of the entropy is an arbitrary convention, not a law of nature. The law of nature is (91), which is a law about entropy changes. Only such differences are measurable physical quantities.

While the Nernst heat theorem is about entropy, it is an additional principle not derivable from the second law. It is understood from statistical mechanics to have its origin in the relative sparseness of the quantum mechanical energy levels of a macroscopic system in the neighborhood of its ground state.

An illustration of the third law may be seen in Fig. 11. According to the Clapeyron equation (75) the slope dp/dT of the solid–liquid-II equilibrium line in the p-T plane equals the ratio of entropy to volume changes that accompany that phase transition. The densities of those two equilibrium condensed phases approach limiting values, different from each other, as $T \to 0$, and so the volume difference has a nonvanishing limit; while the entropy difference, by (91), vanishes in that limit, implying that the slope dp/dT of the equilibrium line must then also vanish as $T \to 0$, as it is seen to do.

(That the slope of the vapor–liquid-II equilibrium line also vanishes at $T = 0$ is not an illustration of the third law. That slope vanishes because the specific volume of the equilibrium vapor rapidly becomes infinite as $T \to 0$. Because of the vanishing density of the vapor, the third law is not applicable to it.)

On the scale on which the entropy of a pure crystalline substance is taken to be 0 at $T = 0$, the entropy of that substance in any other physical state at any temperature and pressure may be identified with the entropy change that accompanies the transformation of the substance from a crystal at $T = 0$ to the state in question. This may be determined from measurements of heat capacities [by integrating C_p/T with respect to T at fixed p, for example, according to (59)] and the latent heats $\Delta H (= T\Delta S)$ of equilibrium phase transformations.

One may similarly obtain the enthalpy in that same state, hence also the Gibbs free energy $G = H - TS$ at that point, by (39). That, per unit mass, is the substance's chemical potential μ in that state, by (43). In this way, for example, one would have determined the c_i^0 characteristic of the substance i that appears in (79) or (88).

When the concentrations of the various species in equilibrium in a chemical reaction in a dilute system are then calculated from (52) with (79) or (88) and with the c_i^0's that are obtained from the third law of thermodynamics in the way just outlined, and are then also determined by direct measurement—by chemical analysis of equilibrium mixtures, for example—they are found to agree. These constitute important experimental verifications of the third law.

ACKNOWLEDGMENTS

The author is grateful to J. C. Wheeler for correcting some errors and both to him and to R. B. Griffiths for helpful suggestions. The preparation of this article was supported in part by the National Science Foundation and the Cornell University Materials Science Center.

GLOSSARY

Azeotropy: The circumstance of a density variable's having, exceptionally, the same value in coexisting phases.

Carnot Cycle: A reversible cycle consisting of alternating isothermal and adiabatic steps.

Colligative Property: Property proportional to the number but not the kind of molecules responsible for it.

Density (Variable): Ratio of two extensive variables.

Efficiency: Fraction of the heat absorbed at the higher temperature that is converted to work in a Carnot cycle.

Exact Differential: A quantity the integral of which depends only on the limits but not on the path of integration.

Field (Variable): Variable such as temperature or chemical potential that has a uniform value even in a heterogeneous system.

Heat Capacity: Ratio of heat absorbed to temperature rise.

180° Rule: Rule restricting the angles with which two-phase lines may meet at a triple point.

Joule Expansion: Expansion of a gas into an insulated space without its doing work (isoenergetic expansion).

Joule–Thomson Coefficient: Isenthalpic rate of change of temperature with pressure.

Joule–Thomson Expansion: Expansion of a gas through a plug from a constant high pressure to a constant low pressure (isenthalpic expansion).

Latent Heat: Enthalpy change (heat absorbed or evolved) in a phase transition.

Lever Rule: Rule giving the relative amounts (volumes, masses, etc.) of two phases in equilibrium.

Nernst Heat Theorem: The useful form of the third law of thermodynamics.

Osmotic Pressure: Excess pressure one must exert on a solution to prevent the spontaneous flow of solvent into it.

Schreinemakers' Rules: Density-space analogs of 180° rule.

Virial Series, Virial Coefficients: Expansion of the pressure in a series of powers of the density; the coefficients in that expansion.

Works Cited

Bundy, F. P. (1989), *Physica A* **156,** 169–178.

Fisher, M. E. (1990), in: D. G. Caldi, G. D. Mostow (Eds.), *Proceedings of the Gibbs Symposium,* American Mathematical Society and American Institute of Physics, pp. 39–72.

Gibbs, J. W. (1961), *The Scientific Papers of J. Willard Gibbs,* Vol. 1, *Thermodynamics,* New York: Dover, p. 434.

Griffiths, R. B., Wheeler, J. C. (1970), *Phys. Rev. A* **2,** 1047–1064.

Rowlinson, J. S., Swinton, F. L. (1982), *Liquids and Liquid Mixtures,* 3rd ed., London: Butterworths, p. 46.

Teubner, M. (1991), *J. Chem. Phys.* **95,** 5243–5248.

Wheeler, J. C. (1974), *J. Chem. Phys.* **61,** 4474–4489.

Further Reading

Callen, H. B. (1985), *Thermodynamics and an Introduction to Thermostatics,* 2nd ed., New York: Wiley.

Denbigh, K. G. (1981), *The Principles of Chemical Equilibrium,* 4th ed., Cambridge, U.K.: Cambridge Univ. Press.

Donnan, F. G., Haas, A. (Eds.) (1936), *A Commentary on the Scientific Writings of J. Willard Gibbs,* New Haven: Yale Univ. Press.

Gibbs, J. W. (1961), *The Scientific Papers of J. Willard Gibbs,* Vol. 1, *Thermodynamics,* New York: Dover.

Guggenheim, E. A. (1967), *Thermodynamics, an*

Advanced Treatment for Chemists and Physicists, 5th ed., Amsterdam: North-Holland.

Pippard, A. B. (1957), *The Elements of Classical Thermodynamics,* Cambridge, U.K.: Cambridge Univ. Press.

Planck, M. (A. Ogg, transl.) (1927), *Treatise on Thermodynamics,* 3rd ed., London: Longmans, Green.

Prigogine, I., Defay, R. (D. H. Everett, transl.) (1954), *Treatise on Thermodynamics Based on the Methods of Gibbs and De Donder,* London: Longmans, Green.

THERMODYNAMICS, NONEQUILIBRIUM

Dilip K. Kondepudi, *Department of Chemistry, Wake Forest University, Winston-Salem, North Carolina, U.S.A.*

Ilya Prigogine, *Director, International Solvay Institutes, Brussels, Belgium*

Introduction 311
1. **Formulation of Nonequilibrium Thermodynamics Based on the Assumption of Local Equilibrium** 312
1.1 Local Thermodynamic Variables 312
1.2 The Conservation of Mass 313
1.3 Conservation of Momentum and Energy 313
1.4 The Entropy-Balance Equation 315
1.5 Rational Thermodynamics 315
2. **Extended Thermodynamics** 316
2.1 The Need to Go Beyond Local Equilibrium 316
2.2 The Extended Gibbs Equation 317
2.3 Entropy Production and Nonequilibrium Entropy 318
3. **Linear Nonequilibrium Thermodynamics** 319
3.1 Linear Phenomenological Laws 319
3.2 The Onsager Reciprocal Relations 320
3.3 Thermoelectric Phenomena 320
3.3.1 The Seebeck Effect 321
3.3.2 The Peltier Effect 322
3.4 Thermogalvanomagnetic Phenomena 322
3.5 Isothermal Diffusion 323
3.6 Heat Conduction in Anisotropic Solids 324
3.7 Thermal Diffusion 324
3.8 Electrokinetic Phenomena and the Saxen Relations 325
3.9 Other Cross Effects 327
3.10 Dielectric and Magnetic Relaxation 327
3.11 Theorem of Minimum Entropy Production 327
3.11.1 Stationary States in Thermal Conduction 328
3.11.2 Stationary States in Chemical Systems 328
3.12 Entropy Production and the Stability of Stationary States 329
4. **Far-from-Equilibrium Systems** 330
4.1 General Properties of Entropy Production 330
4.2 Stability of Far-from-Equilibrium Systems 331
4.3 Linear Stability Analysis and Bifurcation of New States 332
4.4 Dissipative Structures 332
Works Cited 336
Further Reading 336

INTRODUCTION

The preceding article (Thermodynamics, Equilibrium) contains the thermodynamics of equilibrium systems. In classical thermodynamics, the changes of entropy are generally evaluated by considering idealized *reversible* processes that connect the initial and final equilibrium states. For reversible transformations, the change in entropy is $dS = dQ/T$ in which dQ is the heat exchanged and T is the absolute temperature. Since changes in entropy depend only on the initial and final states, the entropy changes that occur in nature as a result of irreversible processes can be computed using the idealized reversible processes. This approach does not relate the rate of change of entropy to *irreversible* processes such as heat conduction, diffusion, and chemical reactions. A formalism that relates entropy to irreversible processes (Prigogine, 1947, 1967) begins with the recogni-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

tion that dS is a sum of two parts (see Fig. 1):

$$dS = d_iS + d_eS, \quad d_iS \geq 0, \tag{1}$$

in which d_iS is the entropy change due to irreversible processes within the system and d_eS is the change in entropy due to exchange of matter and energy with the exterior, both occurring in a time dt. For a closed system, $d_eS = dQ/T$, where dQ is the heat exchanged with the exterior in a time dt. For idealized reversible processes $d_iS = 0$; we then have the classical result $dS = dQ/T$. The formulation of thermodynamics presented here relates d_iS explicitly to irreversible processes and encompasses both equilibrium and nonequilibrium systems.

Irreversible processes can be characterized by *thermodynamic forces,* which we shall denote by X_α, and *thermodynamic flows,* J_α. A gradient of temperature, for example, is the thermodynamic force that causes the flow of heat. Explicit expressions for the forces and flows in nonequilibrium systems are given below (see Sec 1.4). A basic quantity in nonequilibrium thermodynamics is the entropy production per unit volume, σ, which, as shown below, is a bilinear function in the forces and flows:

$$\sigma = \sum_\alpha J_\alpha X_\alpha. \tag{2}$$

Nonequilibrium systems may be classified into two regimes: the linear regime, which is characterized by a linear relationship between the forces and flows, and the nonlinear regime, in which the flows are nonlinear functions of the forces. In the linear regime, general results, such as the Onsager reciprocal relations and the theorem of minimum entropy production, are valid. In the nonlinear regime, we encounter instability and spontaneous transition to organized states called *dissipative structures.*

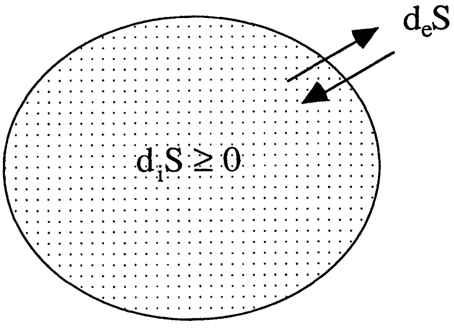

FIG. 1. The change of entropy of a system $dS = d_iS + d_eS$, in which d_iS is the change due to irreversible processes and d_eS is the change due to exchange of energy and matter with the exterior.

1. FORMULATION OF NONEQUILIBRIUM THERMODYNAMICS BASED ON THE ASSUMPTION OF LOCAL EQUILIBRIUM

1.1 Local Thermodynamic Variables

This formulation of nonequilibrium thermodynamics is based on the assumption that the temperature $T(x,t)$ in a nonequilibrium system, though not uniform, is a well-defined, measurable quantity at each point x and time t. Furthermore, it is assumed that in every elemental volume to which a definite temperature may be assigned, thermodynamic variables such as entropy and chemical potential may be defined, and that all the equilibrium relationships remain valid for these variables. It is assumed that, in spite of the existence of gradients of temperature, concentration, or other variables, the equilibrium relations between local variables remain valid. This is the *assumption of local equilibrium.* It has been shown that this assumption is valid whenever the local velocity distribution is nearly Maxwellian (Prigogine, 1949); a detailed discussion of this point may be found in Chapter IX of De Groot and Mazur (1969). Systems that are not locally in equilibrium are exceptional; for these systems, no general thermodynamic theory exists.

In accordance with the assumption of local equilibrium the following local variables are defined:

$$s = \text{entropy per unit mass,} \tag{3}$$

$$u = \text{total internal energy per unit mass,} \tag{4}$$

$$\rho_k = \text{mass per unit volume of component } k, \tag{5}$$

$$\rho = \sum_k \rho_k, \text{ mass per unit volume,} \tag{6}$$

$$\omega_k = \rho_k/\rho, \text{ the mass fraction of component } k, \tag{7}$$

$$\nu = \text{the specific volume.} \tag{8}$$

As in the case of equilibrium systems, it is assumed that s is a function of u, ν, and ω_k,

$$s = s(u,\nu,\omega_k), \tag{9}$$

and that

$$Tds = Td_is + Td_es = du + pd\nu - \sum_k \mu_k d\omega_k, \tag{10}$$

in which d_is and d_es are the entropy production and entropy exchange, respectively, p is the pressure, and μ_k is the chemical potential of component k. The total internal energy U and the total entropy S of a system of volume V are given by

$$U = \int_V \rho u dV, \tag{11}$$

$$S = \int_V \rho s dV. \tag{12}$$

The entropy production per unit volume

$$\sigma \equiv \frac{d_is}{dt}\rho \tag{13}$$

is derived incorporating the law of conservation of mass, momentum, and energy into Eq. (10).

1.2 The Conservation of Mass

The classical nonrelativistic mass is conserved in all physical and chemical processes. In any volume, the change in mass can only be due to flow through the surface enclosing the volume. In fluids, if $v(x,t)$ is the velocity at the location x at time t, the conservation of mass is expressed in the differential form

$$\frac{\partial \rho}{\partial t} = -\nabla\cdot(\rho\mathbf{v}). \tag{14}$$

For each component (or compound) k, on account of chemical reactions there is a "source" term that produces or consumes it. If the reaction rate (or velocity) of reaction j is v_j, and the flow velocity of the component is $\mathbf{v}_k$, the mass balance equation for k is

$$\frac{\partial \rho_k}{\partial t} = -\nabla\cdot(\rho_k\mathbf{v}_k) + \sum_j \nu_{jk}v_j, \tag{15}$$

in which $\nu_{jk} = M_k\nu'_{jk}$, where ν'_{jk} is the stoichiometric coefficient of component k in reaction j, and M_k is the molecular mass of k. The flow $\rho_k\mathbf{v}_k$ can be separated into the center-of-mass or barycentric velocity $\mathbf{v}$ and the flow $\mathbf{J}_k$ with respect to the center of mass, called the "diffusional flow." $\mathbf{v}$ and $\mathbf{J}_k$ are defined by

$$\mathbf{v} = \sum_k \rho_k\mathbf{v}_k/\rho, \tag{16}$$

$$\mathbf{J}_k \equiv \rho_k(\mathbf{v}_k - \mathbf{v}), \quad \sum_k \mathbf{J}_k = 0. \tag{17}$$

Substitution of (17) into (15) gives

$$\frac{\partial \rho_k}{\partial t} = -\nabla\cdot\mathbf{J}_k - \nabla\cdot(\rho_k\mathbf{v}) + \sum_j \nu_{jk}v_j. \tag{18}$$

In order to look at the time derivatives in the center-of-mass reference frame, the following "barycentric time derivative" may be introduced:

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla. \tag{19}$$

Using Eqs. (14)–(19), it can be shown that the mass balance of component k in the center-of-mass reference frame can be written as

$$\rho\frac{d\omega_k}{dt} = -\nabla\cdot\mathbf{J}_k + \sum_j \nu_{jk}v_j. \tag{20}$$

1.3 Conservation of Momentum and Energy

The rate of change of the center-of-mass momentum at a point x is given by

$$\frac{\partial \rho\mathbf{v}}{\partial t} = -\nabla\cdot(\rho\mathbf{v}\mathbf{v} + \mathsf{P}) + \sum_k \rho_k\mathbf{F}_k, \tag{21}$$

in which $\mathbf{vv}$ is a dyadic product, P is the pressure or stress tensor, and F_k is the force per unit mass acting on the component k. In terms of the Cartesian components of the vectors and the tensors,

$$\frac{\partial \rho v_\alpha}{\partial t} = -\sum_\beta \frac{\partial}{\partial x_\alpha}(\rho v_\alpha v_\beta + P_{\alpha\beta}) + \sum_k \rho_k(\mathbf{F}_k)_\alpha, \quad \alpha, \beta = 1, 2, 3. \tag{22}$$

A conservative force $\mathbf{F}_k$ is a gradient of a potential energy ψ_k:

$$\mathbf{F}_k = -\nabla\psi_k, \tag{23}$$

in which ψ_k is the interaction energy or potential energy of component k per unit mass. For static fields ψ_k is independent of time. In chemical reactions the potential energy remains unchanged. As a consequence,

$$\sum_k \psi_k \nu_{jk} = 0. \tag{24}$$

Using mass balance equation (15) and Eqs. (17) and (22)–(24), the following equation for the rate of change of the sum of kinetic and potential energy for the center-of-mass motion is obtained:

$$\frac{\partial(\mathbf{v}^2/2 + \psi)\rho}{\partial t} = -\nabla\left((\mathbf{v}^2/2 + \psi)\rho\mathbf{v} + \mathsf{P}\cdot\mathbf{v} + \sum_k \psi_k\mathbf{J}_k\right) + \mathsf{P}:\nabla\mathbf{v} - \sum_k \mathbf{J}_k\cdot\mathbf{F}_k. \tag{25}$$

In the above expression, ψ is the potential energy density defined by

$$\rho\psi \equiv \sum_k \rho_k\psi_k. \tag{26}$$

Equation (25) shows that the sum of the kinetic and potential energies is not conserved. This loss is compensated by a corresponding change in the total internal energy. Only when we include the total internal energy u do we have conservation of energy. The total energy per unit mass, e, may be defined by

$$e = \tfrac{1}{2}\mathbf{v}^2 + \psi + u, \tag{27}$$

in which the potential energy ψ is not considered a part of the internal energy. (One may also define u such that the potential energy is a part of the internal energy). The conservation of energy can now be written as

$$\frac{\partial \rho e}{\partial t} + \nabla\cdot\mathbf{J}_e = 0, \tag{28}$$

in which $\mathbf{J}_e$ is the energy current density. A more explicit expression for $\mathbf{J}_e$ is

$$\mathbf{J}_e = \rho e\mathbf{v} + \mathsf{P}\cdot\mathbf{v} + \sum_k \psi_k\mathbf{J}_k + \mathbf{J}_u. \tag{29}$$

In this expression the term $\mathsf{P}\cdot\mathbf{v}$ represents the energy flow due to mechanical work, while the term $\Sigma_k\psi_k\mathbf{J}_k$ represents the change in potential energy due to the flows. The term $\mathbf{J}_u$ is the flow of internal energy or, as it is sometimes referred to, the "heat flow." Using (25), (27), and (29) in (28), one obtains the following balance equation for the internal energy:

$$\frac{\partial \rho u}{\partial t} + \nabla\cdot(\rho u\mathbf{v} + \mathbf{J}_u) = -\mathsf{P}:\nabla\mathbf{v} + \sum_k \mathbf{J}_k\cdot\mathbf{F}_k. \tag{30}$$

Using (19), the balance equation for the internal energy in the center-of-mass reference frame can be written as

$$\rho\frac{du}{dt} + \nabla\cdot\mathbf{J}_u = -\mathsf{P}:\nabla\mathbf{v} + \sum_k \mathbf{J}_k\cdot\mathbf{F}_k. \tag{31}$$

For thermodynamic considerations, it is necessary to separate the hydrostatic part of the pressure tensor P through the use of the relation

$$\mathsf{P} = p\mathsf{I} + \mathsf{p}, \tag{32}$$

in which p is the hydrostatic pressure, I is the unit tensor, and p is the dissipative part of the pressure tensor because it contains viscous forces. By virtue of (14) and (19), it can be seen that

$$\rho\frac{d\nu}{dt} = \nabla\cdot\mathbf{v}, \tag{33}$$

in which $\nu = 1/\rho$ is the specific volume. Using (33) and (32) in (31), the following balance equation for the internal energy u in the center-of-mass reference frame is obtained:

$$\frac{du}{dt} + v\nabla\cdot\mathbf{J}_u = -p\frac{dv}{dt} - v\mathsf{p}{:}\nabla\mathbf{v} + v\sum_k \mathbf{J}_k\cdot\mathbf{F}_k. \tag{34}$$

1.4 The Entropy-Balance Equation

For the entropy-balance equation, the starting point is the assumption of local equilibrium as expressed in Eq. (10), which is assumed to be valid in the center-of-mass reference frame. For the time derivative of the specific entropy, we have

$$T\frac{ds}{dt} = \frac{du}{dt} + p\frac{dv}{dt} - \sum_k \mu_k \frac{d\omega_k}{dt}. \tag{35}$$

The equations of balance of internal energy, (34), and mass, (20), can now be used in (35) to obtain the following entropy-balance equation:

$$\begin{aligned}\rho\frac{ds}{dt} + \nabla\cdot\left(\frac{\mathbf{J}_u - \sum_k \mu_k\mathbf{J}_k}{T}\right) &= \mathbf{J}_u\cdot\nabla\left(\frac{1}{T}\right)\\ &- \sum_k \mathbf{J}_k\cdot\nabla\left(\frac{\mu_k}{T}\right) + \sum_k\frac{\mathbf{F}_k\cdot\mathbf{J}_k}{T} - \frac{\mathsf{p}{:}\nabla\mathbf{v}}{T} + \sum_j\frac{A_j v_j}{T},\end{aligned} \tag{36}$$

in which A_j is the chemical affinity of the reaction j, defined by

$$A_j = -\sum_k \nu_{jk}\mu_k. \tag{37}$$

Chemical affinity of a reaction is a fundamental concept introduced by De Donder and Van Rysselberghe (1936; see also Prigogine and Defay, 1967) to describe the thermodynamic driving force of chemical reaction. Equation (36) enables us to identify an entropy current $\mathbf{J}_S$ and an entropy production per unit volume σ as

$$\mathbf{J}_S \equiv \frac{1}{T}\left(\mathbf{J}_u - \sum_k \mu_k\mathbf{J}_k\right), \tag{38}$$

$$\begin{aligned}\sigma &\equiv \mathbf{J}_u\cdot\nabla\left(\frac{1}{T}\right) - \sum_k \mathbf{J}_k\cdot\nabla\left(\frac{\mu_k}{T}\right)\\ &+ \sum_k \mathbf{J}_k\cdot\frac{\mathbf{F}_k}{T} - \mathsf{p}{:}\frac{\nabla\mathbf{v}}{T} + \sum_j v_j\frac{A_j}{T}\\ &= \sum_\alpha J_\alpha X_\alpha \geq 0,\end{aligned} \tag{39}$$

in which we have emphasized the positivity of σ in accordance with the second law. Each of the terms in (39) is a product of a thermodynamic flow caused by a thermodynamic force: the gradient of $1/T$ causes the "heat flow" $\mathbf{J}_u$, the gradient of μ_k/T causes the matter flow $\mathbf{J}_k$, and affinities A_j cause the chemical reactions j with velocities v_j. The form of σ is bilinear in the thermodynamic forces and flows, as expressed in (2). Table 1 lists the thermodynamic forces and the corresponding flows. Alternative ways of expressing the entropy production and "heat flow" $\mathbf{J}_u$ can be found in De Groot and Mazur (1969) or Glansdorff and Prigogine (1971). The total entropy production of the system is given by

$$\frac{d_iS}{dt} = \int_V \sigma dV \geq 0. \tag{40}$$

1.5 Rational Thermodynamics

Some authors (Truesdell, 1969) have argued that the traditionally used Clausius inequality $TdS \geq \delta Q$ and other thermodynamic statements are lacking in mathematical rigor. For example, according to Truesdell (1969, p. 3), the reader "is told that dS is a differential, but not of what variable S is a function; that δQ is a small quantity not generally a differential; he is expected to believe not only that one differential can be bigger than another, but even that a differential can be bigger than something that is not a differential." In order to make thermodynamic more rigorous or "rational," an axiomatic approach is taken (Coleman and Noll, 1963; Coleman, 1964; Truesdell, 1969; Noll, 1974). This approach also formally includes systems with memory for which the values of the de-

Table 1. Table of thermodynamic forces and flows.

	Force X_α	Flow $\mathbf{J}_\alpha$
Heat conduction	$\nabla(1/T)$	Heat flow $\mathbf{J}_u$
Diffusion	$-\nabla(\mu_k/T)$	Diffusion current $\mathbf{J}_k$
Conservative force	$-\nabla\psi_k/T = F_k/T$	Current $\mathbf{J}_k$
Chemical reactions	A_j/T	Velocity of reaction v_j
Viscous dissipation	$-\nabla\mathbf{v}$	Stress p

pendent variables depend on the history of the system; i.e., the values of the dependent variables at time t are functionals of the independent variables for all times less than $t' < t$.

In rational thermodynamics, s and T are taken as primitive concepts. No physical or operational interpretation is given to these variables; in other words, how these quantities are related to experimentally measured variables is not made clear at the outset. It is assumed that the variables of the system are in general memory dependent. Furthermore, functions such as s are assumed to be dependent not only the variables such as T and ν but also on the gradient of T and the velocity of the fluid. Thus, for example,

$$s = s(\nu, T, \mathbf{v}, \nabla T). \tag{41}$$

Similarly, other quantities such as the heat flux $\mathbf{J}_q$ are also functionals of the same variables. Such formal relations that express one set of variables as functions of the other are called *constitutive equations*. Forces and energy fluxes are assumed to be such that the balance of momentum and energy are always valid. The conservation of energy (34) includes an additional term ρr due to radiation. The force terms $\mathbf{F}_k$ are assumed to be such that the conservation of momentum is satisfied. Unlike the formulation in Sec. 1.3, rational thermodynamics does not introduce the forces as known functions; the momentum-balance equation is taken to be the definitions of the forces.

The other axioms on which rational thermodynamics is based are the following:

- The constitutive equations should be such that the Clausius–Duhem inequality, stated as

$$\frac{d}{dt}\int_V \rho s dV \geq -\int_a \frac{\mathbf{J}_q \cdot d\mathbf{a}}{T} + \int_V \frac{\rho r}{T} dV, \tag{42}$$

  holds, in which $d\mathbf{a}$ is the area-element vector and $\mathbf{a}$ is the area enclosing the volume V. The second law is incorporated into the theory using this condition or other local forms that can be derived from this basic condition.
- The *principle of equipresence* holds, which states that all independent variables are equally present in the functions that specify the dependent variables except when the Clausius–Duhem inequality is violated.
- The *principle of material frame-indifference* holds, according to which all constitutive equations must be independent of the frame of reference, which undergoes arbitrary time-dependent translations and rotations. That is, equations for quantities such as heat flux should be the same for every frame of reference including arbitrary rotations.

Other more detailed aspects of the formalism can be found in Truesdell (1969).

With the above axioms, it is shown that the specific entropy s and energy u are local functions of ν and T, in agreement with the assumption of local equilibrium. In this formulation the local validity of the Gibbs equation is derived from the axioms, not assumed. Many of the results obtained by the theory that assumes local equilibrium are also obtained in rational thermodynamics, though with more emphasis on mathematical rigor than on applications.

Rational thermodynamics is mostly of formal interest, and its application has been beset with many computational difficulties. The most serious difficulty is the principle of material frame-indifference, which is found not to be valid in kinetic theory of gases, turbulence, rheology, and other subjects (Jou, 1993).

2. EXTENDED THERMODYNAMICS

2.1 The Need to Go Beyond Local Equilibrium

When a system is not in thermodynamic equilibrium, there are flows such as heat flow or diffusion. In general, every part is associated not only with the usual variables such as energy density, specific volume, pressure, and temperature but also with flows, such as heat flow. In view of this, it is natural to ask if quantities such as the entropy density s depend on the flows as well. Indeed, nonequilibrium statistical mechanics does lend support to such a notion. Local-equilibrium formulation assumes that such dependence can be ignored.

Another aspect of local-equilibrium theory is that linear phenomenological laws that can be obtained from the expression for the entropy production, such as the Fourier law of heat conduction and Fick's law of diffusion, imply infinite propagation velocity for heat and matter, which are clearly unphysical. As an example, let us consider heat conduction. According to Fourier's law, the heat flow $\mathbf{J}_q = -\kappa\nabla T$ (κ = heat conductivity) is assumed to be established instantaneously with the temperature gradient. In reality, this is not true because the heat flow $\mathbf{J}_q$ has a relaxation time during which it evolves to the value given by Fourier's law. This relaxation time, τ_q, is very small, of the order of 10^{-13} s (Jou, 1993), but not zero. From a thermodynamic point of view, the relaxation of these flows is itself an entropy-producing process, and hence the entropy depends on the value of these flows. Thus, in dealing with rapid processes in which relaxation times of the flows cannot be ignored, the local-equilibrium approach is inadequate. However, a very large class of phenomena, discussed in later sections of this article, can be described using the local-equilibrium thermodynamics.

The more recently developed *extended thermodynamics* is a formalism that goes beyond the assumption of local equilibrium and finds the explicit dependence of entropy on the flows. It is better suited for describing fast processes and viscoelastic phenomena in which the relaxation time of the flows cannot be ignored. Because of space limitations, we give only a brief outline of this theory. An excellent description of extended thermodynamics and its applications may be found in Jou (1993) and Müller and Ruggeri (1993).

2.2 The Extended Gibbs Equation

In presenting the basic formalism of extended thermodynamics, for simplicity we consider only the heat flow and assume that there are no gradients of velocity. The specific entropy s is then a function of u, ν, and the heat flow $\mathbf{J}_q$:

$$s = s(u,\nu,\mathbf{J}_q), \tag{43}$$

and

$$ds = \left(\frac{\partial s}{\partial u}\right)du + \left(\frac{\partial s}{\partial \nu}\right)d\nu + \left(\frac{\partial s}{\partial \mathbf{J}_q}\right)\cdot d\mathbf{J}_q. \tag{44}$$

A nonequilibrium absolute temperature θ and pressure π may now be defined as

$$\frac{1}{\theta} = \frac{\partial s}{\partial u}, \quad \frac{\pi}{\theta} = \frac{\partial s}{\partial \nu}. \tag{45}$$

To make a connection with the temperature $T(u,\nu)$ and pressure $p(u,\nu)$ of the local equilibrium theory, it is assumed that to the leading order the temperature θ is equal to T, i.e.,

$$1/\theta = 1/T(u,\nu) + \text{higher-order terms involving } \mathbf{J}_q. \tag{46}$$

Similarly,

$$\frac{\pi}{\theta} = \frac{p(u,\nu)}{T(u,\nu)} + \text{higher-order terms involving } \mathbf{J}_q. \tag{47}$$

To relate the formalism to relaxation times of the heat flow more directly, the derivative of s with respect to $\mathbf{J}_q$ is written in the form

$$\frac{\partial s}{\partial \mathbf{J}_q} = -\frac{\nu}{T}\boldsymbol{\alpha}(u,\nu,\mathbf{J}_q), \tag{48}$$

where α is a function to be determined. To the first order in the flow $\mathbf{J}_q$, the function α may be written as

$$\boldsymbol{\alpha} = \alpha_0(u,\nu)\mathbf{J}_q, \tag{49}$$

where α_0 is a function of u and ν, as indicated. To the leading order in $\mathbf{J}_q$, the differential of the specific entropy can now be written as

$$ds = \frac{1}{T}du + \frac{p}{T}d\nu - \frac{\nu\alpha_0}{T}\mathbf{J}_q\cdot d\mathbf{J}_q. \tag{50}$$

This extension of the Gibbs equation is the basic equation of extended thermodynamics. The success and usefulness of this equation depend on the specification of α_0 and its experimental determination. As shown below,

$$\alpha_0 = \tau_q/\kappa T, \tag{51}$$

where τ_q is the relaxation time for the heat flow and κ is the heat conductivity.

The inclusion of viscous effects into the expression (50) can be similarly done. The result is the expression

$$ds = \frac{1}{T}du + \frac{p}{T}dv - \frac{v\alpha_0}{T}\mathbf{J}_q \cdot d\mathbf{J}_q - \frac{v\alpha_p}{T}p^v dp^v - \frac{v\alpha_p'}{T}\mathsf{p}^v{:}d\mathsf{p}^v, \qquad (52)$$

in which the dissipative part of the pressure tensor p, defined by (32), has been expressed as

$$\mathsf{p} = p^v\mathsf{I} + \mathsf{p}^v, \quad p^v = \tfrac{1}{3}Tr(\mathsf{p}). \qquad (53)$$

The tensor p^v is the traceless part of the dissipative pressure tensor p. As in the case of α_0, the terms α_p and α_p' can be related to the relaxation times of the viscous flows (Jou, 1993):

$$\alpha_p = \tau_\zeta/\zeta, \quad \alpha_p' = \tau_\eta/\eta, \qquad (54)$$

in which ζ and η are the bulk and shear viscosities and τ_ζ and τ_η are the corresponding relaxation times.

2.3 Entropy Production and Nonequilibrium Entropy

From expression (50) for the differential of the specific entropy, the following expression for the time derivative of s can be obtained using the energy conservation equation (34) when the only flow is the heat flow:

$$\rho\frac{ds}{dt} = \frac{1}{T}\nabla \cdot \mathbf{J}_q - \frac{\alpha_0}{T}\mathbf{J}_q \cdot \frac{d\mathbf{J}_q}{dt}, \qquad (55)$$

in which we have used the density $\rho = 1/v$, and the barycentric time derivative (19). Also, since the energy flow is entirely the heat flow, $\mathbf{J}_q$ has been used in the place of $\mathbf{J}_u$. As was done for the case of local equilibrium, if this equation can be cast in the form

$$\rho\frac{ds}{dt} + \nabla \cdot \mathbf{J}_s = \sigma, \qquad (56)$$

the entropy production σ and the entropy current or entropy flow $\mathbf{J}_s$ can be identified. In the above case, one can easily see that

$$\mathbf{J}_s = \frac{\mathbf{J}_q}{T} \quad \text{and} \quad \sigma = \mathbf{J}_q \cdot \left[\nabla\frac{1}{T} - \frac{\alpha_0}{T}\frac{d\mathbf{J}_q}{dt}\right]. \qquad (57)$$

As can be seen, the expression for σ is a product of the flow $\mathbf{J}_q$ and the force $\nabla(1/T) - (\alpha_0/T)(d\mathbf{J}_q/dt)$. Comparing it with (39), we see an additional term containing the time derivative of the heat flow. The thermodynamic flow is driven by the thermodynamic force, and it vanishes when the force vanishes. Thus, to the first order, i.e., the linear approximation, we may write

$$\mathbf{J}_q = L_{qq}\left[\nabla\frac{1}{T} - \frac{\alpha_0}{T}\frac{d\mathbf{J}_q}{dt}\right]. \qquad (58)$$

This equation is essentially the same as the following Cattaneo equation (Jou, 1993; Müller and Ruggeri, 1993) for heat conduction, which is derived using kinetic theory of gases and which takes the relaxation time of the heat current into account:

$$\mathbf{J}_q = -\tau_q\frac{d\mathbf{J}_q}{dt} - \kappa\nabla T. \qquad (59)$$

Comparing (58) and (59) we see that

$$L_{qq} = \kappa T^2, \quad \alpha_0 = \tau_q/\kappa T. \qquad (60)$$

Thus the coefficients α_0 and L_{qq} can be related to measurable quantities such as κ and τ_q. Using the relation (60) for α_0 in (50), one can now obtain a more explicit expression for ds:

$$ds = \frac{du}{T} + \frac{p}{T}dv - \frac{\tau_q v}{\kappa T}\mathbf{J}_q \cdot d\mathbf{J}_q. \qquad (61)$$

When the heat current vanishes, the specific entropy corresponds to the equilibrium entropy. Thus, if we integrate for the equilibrium state of $\mathbf{J}_q = 0$ to $\mathbf{J}_q$, we obtain

$$s(u,v,\mathbf{J}_q) = s(u,v)_{\text{eq}} - \frac{\tau_q v}{2\kappa T}\mathbf{J}_q \cdot \mathbf{J}_q. \qquad (62)$$

This expression shows that because of the flow the system has slightly smaller entropy than the equilibrium entropy. Expressions (57) and (62) can be generalized to include other thermodynamic flows and forces (Jou, 1993; Müller and Ruggeri, 1993).

As in the case of local equilibrium theory, in extended thermodynamics the second law is stated as $\sigma > 0$. The specific entropy is assumed to be an additive, convex function of all the independent variables.

3. LINEAR NONEQUILIBRIUM THERMODYNAMICS

The thermodynamic flows are a function of the thermodynamic forces. At equilibrium, the forces and the flows vanish. Close to equilibrium, the flows are linear functions of the forces. This is the regime of linear nonequilibrium thermodynamics. General results valid in this regime are summarized in this section. When a system is driven far from equilibrium, the flows become nonlinear functions of the forces. In the nonlinear regime, instability, self-organization, and complex behavior can arise. Nonlinear conditions can most easily be achieved in chemical systems.

3.1 Linear Phenomenological Laws

In the linear regime, the thermodynamic flows J_α are linear functions of forces X_α:

$$J_\alpha = \sum_\alpha L_{\alpha\beta} X_\beta. \tag{63}$$

The coefficients $L_{\alpha\beta}$ are called the *phenomenological coefficients*. Equation (63) implies that the forces and their corresponding flows do not act independently; they are coupled to each other. As is clear from Table 1, the forces X_β have various tensorial characters. In the case of a chemical reaction, the thermodynamic force is a scalar; for the case of heat flow, it is a vector; and for viscous dissipation, it is a tensor of the second rank. Symmetry considerations restrict the possible coupling between forces and flows. In isotropic systems, forces and flows of different tensorial character do not couple. This symmetry principle, sometimes referred to as the Curie principle (Prigogine, 1947, 1967; De Groot and Mazur, 1969), implies, for example, that a chemical affinity, which is a scalar, could not be coupled to heat flow, which is a vector. Thus, the corresponding phenomenological coefficients $L_{\alpha\beta}$ vanish. On the other hand, the coupling of diffusion to heat flow or the coupling of heat flow to electrical current is possible. These cross effects are the thermal-diffusion and thermoelectric effects, respectively.

In the linear regime, the entropy production σ takes the quadratic form

$$\sigma = \sum_\alpha J_\alpha X_\alpha = \sum_{\alpha,\beta} L_{\alpha\beta} X_\alpha X_\beta \geq 0. \tag{64}$$

For the above quadratic form, the positivity of entropy production leads to the following conditions on $L_{\alpha\beta}$:

$$L_{\alpha\alpha} > 0 \quad \text{and} \quad 4L_{\alpha\alpha}L_{\beta\beta} \geq (L_{\alpha\beta} + L_{\beta\alpha})^2. \tag{65}$$

Long-known empirical laws relating flows and the corresponding causes have been unified by nonequilibrium thermodynamics. Some simple examples of well-known empirical linear laws are Fourier's law of heat conduction,

$$\mathbf{J}_q = -\kappa \nabla T(x), \tag{66}$$

Fick's law of diffusion,

$$\mathbf{J}_k = -\rho D_k \nabla \omega_k(x), \tag{67}$$

and Ohm's law,

$$\mathbf{I} = \mathbf{E}/\rho_e, \tag{68}$$

in which κ is the heat conductivity, D_k is the diffusion coefficient, ω_k is the mass fraction of k [cf. Eq. (7)], $\mathbf{I}$ is the electric current density, $\mathbf{E}$ is the electric field, and ρ_e is the specific resistance. These laws remain nearly linear even for very large gradients. The relation between the empirical constants such as κ in (66) and the corresponding phenomenological coefficient L_{qq} can be obtained by equating the phenomenological heat current with the one according to (66):

$$\mathbf{J}_q = L_{qq} \nabla(1/T) = -\kappa \nabla T, \tag{69}$$

from which it follows that

$$\kappa = L_{qq}/T^2. \tag{70}$$

For diffusion in a multicomponent system, the currents are not independent because $\Sigma_k \mathbf{J}_k = 0$ [cf. (17)]. For dilute solutions in which the isothermal diffusion of a minor component is considered, we may assume that the center of mass is essentially fixed to the stationary major component. In this case, the relation between the phenomenological coefficient L_{kk} and the diffusion coefficient D_k is (see Sec. 3.5)

$$D_k = \left(\frac{\partial \mu_k}{\partial \omega_k}\right) \frac{L_{kk}}{\rho T}, \tag{71}$$

in which μ_k is the chemical potential and ω_k is the mass fraction. Similarly, for electrical conduction at constant temperature (see Sec. 3.3),

$$1/\rho_e = L_{ee}/T, \tag{72}$$

in which ρ_e is the specific resistance. For chemical systems, however, the linear laws are valid only for small perturbations in the concentrations from the equilibrium values. If R_{fj} and R_{rj} are the forward and reverse reaction rates of an elementary reaction, the reaction velocity is

$$v_j = R_{fj} - R_{rj}. \tag{73}$$

The relation between the velocity and the affinity can be shown to be (Prigogine, 1967)

$$v_j = R_{fj}[1 - \exp(-A_j/RT)], \tag{74}$$

in which R is the gas constant. Equation (74) shows that only when the affinities are much smaller than RT is there a linear relation between the velocity and the affinity. Affinities of the order of 100 $\mathrm{kJ \cdot mol^{-1}}$ are commonly encountered. At room temperature, since RT is about 2.5 $\mathrm{kJ \cdot mol^{-1}}$, chemical systems are generally not in the linear regime of thermodynamics.

3.2 The Onsager Reciprocal Relations

For the phenomenological coefficients $L_{\alpha\beta}$, based on the assumption of microscopic reversibility, Lars Onsager (Onsager, 1931a, 1931b) proposed the *reciprocal relations*

$$L_{\alpha\beta} = L_{\beta\alpha}. \tag{75}$$

The theoretical basis of these relations requires careful considerations (Casimir, 1945; Wigner, 1954; Truesdell, 1969; Kreuzer, 1981) that are beyond the scope of this article. Examples of cross effects and the reciprocal relations are discussed in Sec. 3.3–3.9. In the presence of a magnetic field **B**, the reciprocal relations take the form

$$L_{\alpha\beta}(\mathbf{B}) = L_{\beta\alpha}(-\mathbf{B}). \tag{76}$$

As we have noted above [cf. (17)], in the center-of-mass reference frame the flows are not all independent. However, the reciprocal relations are valid even when the flows are not all independent. In general, if there exists a linear relation between the n flows,

$$\sum_{\alpha=1}^{n} y_\alpha J_\alpha = 0, \tag{77}$$

then the expression for the entropy production can be expressed in terms of the $n - 1$ independent flows:

$$\sigma = \sum_{\alpha=1}^{n} J_\alpha X_\alpha = \sum_{\alpha=1}^{n-1} J_\alpha \left(X_\alpha - \frac{y_i}{y_n} X_n\right). \tag{78}$$

For these $n - 1$ flows, we have the new linear phenomenological relations

$$J_\alpha = \sum_{\alpha=1}^{n-1} \tilde{L}_{\alpha\beta} \left(X_\beta - \frac{y_i}{y_n} X_n\right). \tag{79}$$

Given that $L_{\alpha\beta} = L_{\beta\alpha}$, it can be shown that the reciprocal relations are also valid for $\tilde{L}_{\alpha\beta}$, i.e.,

$$\tilde{L}_{\alpha\beta} = \tilde{L}_{\beta\alpha}. \tag{80}$$

In addition to a linear relation between the thermodynamic flows, if there is also a linear relation between the thermodynamic forces, then the reciprocal relations are not generally valid for any set of phenomenological coefficients $L_{\alpha\beta}$. However, it is always possible to choose a set of flows and forces for which the reciprocal relations are valid. Examples of cross effects and reciprocal relation are discussed in the following sections.

3.3 Thermoelectric Phenomena

As a first illustration of the above theory, we consider thermoelectric effects that involve the flow of heat $\mathbf{J}_q$ and electric current $\mathbf{I}_e$ in conducting wires (in which the subscript e indicates that the flow corresponds to the flow of electrons). The entropy production per unit volume due to these two irreversible processes and the linear phenomenological laws associated with it are

$$\sigma = \mathbf{J}_q \cdot \nabla(1/T) + \mathbf{I}_e \cdot \mathbf{E}/T, \tag{81}$$

$$\mathbf{J}_q = L_{qq}\nabla(1/T) + L_{qe}\mathbf{E}/T, \tag{82}$$

$$\mathbf{I}_e = L_{ee}\mathbf{E}/T + L_{eq}\nabla(1/T). \tag{83}$$

In the above equations, **E** is the electric field. For a one-dimensional system, such as a conducting wire, the vectorial aspect of $\mathbf{J}_q$ and $\mathbf{I}_e$ is unimportant and they may be treated as scalars. To relate the coefficients L_{qq} and L_{ee} with the heat conductivity κ and resistance R, we can write Eqs. (82) and (83) in a one-dimensional system as

$$J_q = -\frac{1}{T^2}L_{qq}\frac{\partial T}{\partial x} + L_{qe}\frac{E}{T}, \tag{84}$$

$$I_e = L_{ee}\frac{E}{T} - \frac{1}{T^2}L_{eq}\frac{\partial T}{\partial x}. \tag{85}$$

To find the relation between L_{ee} and the resistance R, we note that the emf $V = -\Delta\phi = \int_0^l E dx$ in which l is the length of the system. The current I_e is a constant in the entire system. At uniform temperature ($\partial T/\partial x = 0$), the current is entirely due to electrical potential difference. Integrating (85) over the length of the system we obtain

$$L_{ee} = \frac{T}{(R/l)} = \frac{T}{r}, \tag{86}$$

in which R is the total resistance of the wire and r is the resistance per unit length. More generally,

$$L_{ee} = T/\rho_e, \tag{87}$$

in which ρ_e is the specific resistance. When we consider a one-dimensional system, ρ_e is replaced by r, resistance per unit length. We have already seen that $L_{qq} = T^2\kappa$ [cf. (70)].

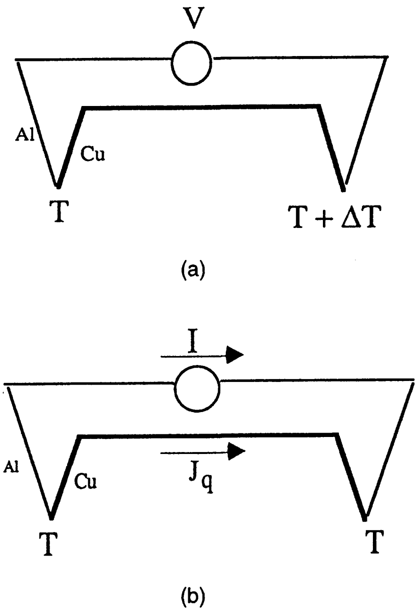

FIG. 2. The thermoelectric effects, examples of "cross effects" between thermodynamic forces and flows. (a) In the Seebeck effect, two dissimilar metal wires are joined and the junctions are maintained at different temperatures. (b) In the Peltier effect, the two junctions are maintained at the same temperature, and a current is passed through the system. The current flow drives a heat flow J_q from one junction to the other.

3.3.1 The Seebeck Effect The cross coefficients L_{qe} and L_{eq} can also be related to experimentally measured quantities. In the Seebeck effect (see Fig. 2), a temperature difference between two junctions of dissimilar metals produces an emf. This emf is measured at zero current. Setting $I_e = 0$ in Eq. (85) we obtain

$$0 = L_{ee}ET - L_{eq}\frac{\partial T}{\partial x}. \tag{88}$$

This equation may now be integrated to obtain a relation between the temperature difference ΔT and the emf generated due to this temperature difference, $\Delta\phi = -\int E dx$. In doing this integration, we shall assume that the total variation ΔT is small and make the approximation $\int TE dx \approx T\int E dx = -T\Delta\phi$. This gives us the relation

$$L_{eq} = -L_{ee}T\left(\frac{\Delta\phi}{\Delta T}\right)_{I=0}. \tag{89}$$

Experimentally, the ratio $-(\Delta\phi/\Delta T)_{I=0}$, called *thermoelectric power*, is measured. Using (87) and (89), the coefficient L_{eq} can be related to the measured quantities.

3.3.2 The Peltier Effect In the Peltier effect, the two junctions are maintained at a constant temperature while a current I is passed through the system (see Fig. 2). This causes a flow of heat from one junction to another. The two junctions are maintained at the same temperature by supplying heat to one junction and removing heat from the other junction—thus maintaining a steady heat flow J_q. Under these conditions, the ratio

$$\Pi = J_q/I, \tag{90}$$

which can be measured, is called the *Peltier heat.* The cross coefficient L_{qe} can be related to the Peltier heat as follows. Since there is no temperature difference between the two junctions, $\partial T/\partial x = 0$, and Eqs. (84) and (85) become

$$J_q = L_{qe}E/T, \tag{91}$$

$$I_e = L_{ee}E/T. \tag{92}$$

Dividing one equation by the other and using Eqs. (90) and (86) we obtain

$$L_{qe} = \Pi L_{ee} = \Pi\frac{T}{(R/l)} = \Pi\frac{T}{r}. \tag{93}$$

In this manner, all the phenomenological coefficients L_{qe} and L_{eq} can be related to the experimental parameters of the cross effects.

According to the Onsager reciprocal relations, one must find $L_{qe} = L_{eq}$. Upon using Eq. (89) for L_{eq} and Eq. (93) for L_{qe}, we find

$$-L_{ee}T\left(\frac{\Delta\phi}{\Delta T}\right) = \Pi L_{ee}, \quad \text{or} \quad -\left(\frac{\Delta\phi}{\Delta T}\right) = \frac{\Pi}{T}. \tag{94}$$

Extensive data verifying the reciprocal relations may be found in Miller (1960, 1974).

3.4 Thermogalvanomagnetic Phenomena

Thermogalvanomagnetic phenomena deal with the cross effects between the flows of heat and electric current in the presence of a magnetic field. An example is the Hall effect (see Fig. 3) in which a current flow in the x direction in the presence of a magnetic field in the z direction results in the generation of a voltage in the y direction. The entropy production due to thermal and electric currents in a system under mechanical equilibrium is [cf. Eq. (81)]

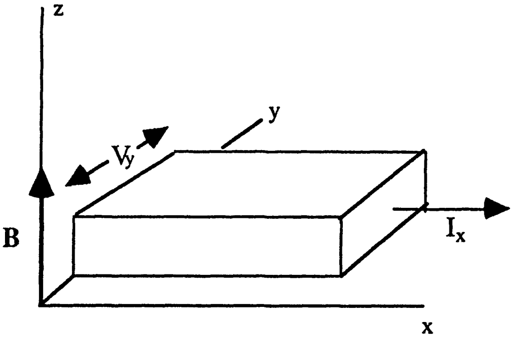

FIG. 3. The Hall effect, an example of a galvanomagnetic effect. If an electric current I_x is made to flow in the x direction in the presence of a magnetic field B in the z direction, a voltage V_y is generated in the y direction.

$$\sigma = \mathbf{J}_q\cdot\nabla(1/T) + \mathbf{I}_e\cdot\mathbf{E}/T. \tag{95}$$

In this case the vectorial character of the current is essential to the analysis. The phenomenological laws are

$$\mathbf{I}_e = \mathsf{L}_{ee}(\mathbf{B})\cdot\mathbf{E} + \mathsf{L}_{eq}(\mathbf{B})\cdot\nabla(1/T), \tag{96}$$

$$\mathbf{J}_q = \mathsf{L}_{qe}(\mathbf{B})\cdot\mathbf{E} + \mathsf{L}_{qq}(\mathbf{B})\cdot\nabla(1/T). \tag{97}$$

In the presence of a magnetic field $\mathbf{B}$, the coefficients are tensors of the second rank instead of scalars, and, as shown explicitly, the elements of L_{ee}, L_{qe}, etc., depend on the magnetic field. In this case, the Onsager reciprocal relations take the form

$$\mathsf{L}_{ee}(\mathbf{B}) = \mathsf{L}_{ee}^{T}(-\mathbf{B}), \tag{98}$$

$$\mathsf{L}_{qq}(\mathbf{B}) = \mathsf{L}_{qq}^{T}(-\mathbf{B}), \tag{99}$$

$$\mathsf{L}_{eq}(\mathbf{B}) = \mathsf{L}_{qe}^{T}(-\mathbf{B}), \tag{100}$$

where the superscript T stands for the transpose of the tensor, in which the rows and columns of the matrix representation are interchanged. In the absence of thermal gradients, the phenomenological coefficients are the Hall coefficients. In terms of the elements of the tensors L, the reciprocal relations (100), for example, take the form

$$[\mathsf{L}_{eq}(\mathbf{B})]_{ij} = [\mathsf{L}_{qe}(-\mathbf{B})]_{ji}. \tag{101}$$

3.5 Isothermal Diffusion

For isothermal diffusion, the entropy production is

$$\sigma = -\sum_k \mathbf{J}_k \cdot \nabla(\mu_k/T), \tag{102}$$

in which $\mathbf{J}_k$ is the matter current and μ_k is the chemical potential of the species k. Under isothermal conditions the associated linear laws are

$$\mathbf{J}_i = -\sum_k \frac{L_{ik}}{T} \nabla\mu_k. \tag{103}$$

For simultaneous diffusion of several species (under isothermal conditions), Fick's law takes the form

$$\mathbf{J}_i = -\sum_k \rho D_{ik} \nabla\omega_k(\mathbf{x}). \tag{104}$$

Since the diffusion flows in a two-component system are not independent, to verify Onsager's reciprocal relations we need at least three components. For three components the entropy production per unit volume is

$$\sigma = -\frac{\mathbf{J}_1}{T}\cdot\nabla\mu_1 - \frac{\mathbf{J}_2}{T}\cdot\nabla\mu_2 - \frac{\mathbf{J}_3}{T}\cdot\nabla\mu_3, \tag{105}$$

in which [cf. (17)]

$$\mathbf{J}_1 + \mathbf{J}_2 + \mathbf{J}_3 = 0. \tag{106}$$

The Gibbs–Duhem equation gives the following relation between the gradients of μ_k:

$$\omega_1(\nabla\mu_1)_{p,T} + \omega_2(\nabla\mu_2)_{p,T} + \omega_3(\nabla\mu_3)_{p,T} = 0. \tag{107}$$

If $\mathbf{J}_3$ and μ_3 are the variables for the solvent while $\mathbf{J}_1$, μ_1 and $\mathbf{J}_2$, μ_2 are the variables of two solutes whose diffusion cross effects are of interest, then, by use of (106) and (107), $\mathbf{J}_3$ and μ_3 can be eliminated from the expression for the entropy production. The entropy production can then be written in terms of only the variables $\mathbf{J}_1$, μ_1 and $\mathbf{J}_2$, μ_2 of the two solutes:

$$\sigma = \mathbf{X}_1\cdot\mathbf{J}_1 + \mathbf{X}_2\cdot\mathbf{J}_2, \tag{108}$$

in which the thermodynamic forces $\mathbf{X}_1$ and $\mathbf{X}_2$ are

$$\mathbf{X}_1 = -\frac{1}{T}\left[\nabla\mu_1 + \frac{\omega_1}{\omega_3}\nabla\mu_1 + \frac{\omega_2}{\omega_3}\nabla\mu_2\right], \tag{109}$$

and

$$\mathbf{X}_2 = -\frac{1}{T}\left[\nabla\mu_2 + \frac{\omega_2}{\omega_3}\nabla\mu_2 + \frac{\omega_1}{\omega_3}\nabla\mu_1\right]. \tag{110}$$

The associated phenomenological laws then take the form

$$\mathbf{J}_1 = L_{11}\mathbf{X}_1 + L_{12}\mathbf{X}_2, \tag{111}$$

$$\mathbf{J}_2 = L_{12}\mathbf{X}_1 + L_{22}\mathbf{X}_2. \tag{112}$$

To verify the reciprocal relations, we must now relate L_{ik} and the experimentally measured diffusion coefficients D_{ik} of Fick's law:

$$\mathbf{J}_1 = -\rho D_{11}\nabla\omega_1 - \rho D_{12}\nabla\omega_2, \tag{113}$$

$$\mathbf{J}_2 = -\rho D_{21}\nabla\omega_1 - \rho D_{22}\nabla\omega_2. \tag{114}$$

The gradient of the chemical potential μ_1 may be written as

$$\nabla\mu_1 = \sum_k \left(\frac{\partial\mu_1}{\partial\omega_k}\right)\nabla\omega_k. \tag{115}$$

A similar relation can be written for the gradient of μ_2. By use of these relations in (109) and (110) and substituting them in (111) and (112), the flows $\mathbf{J}_k$ can be written in terms of the gradients of ω_k. On comparison of these with Fick's law (113) and (114), the following relations between the diffusion coefficients and the linear Onsager coefficients can be obtained:

$$L_{11} = \frac{dD_{11} - bD_{12}}{ad - bc} \qquad L_{12} = \frac{aD_{12} - cD_{11}}{ad - bc}, \tag{116}$$

$$L_{21} = \frac{aD_{21} - bD_{22}}{ad - bc} \qquad L_{22} = \frac{aD_{22} - bD_{21}}{ad - bc}, \tag{117}$$

in which

$$a = \left(1 + \frac{\omega_1}{\omega_3}\right)\left(\frac{\partial\mu_1}{\partial\omega_1}\right) + \frac{\omega_2}{\omega_3}\left(\frac{\partial\mu_2}{\partial\omega_1}\right),$$
$$b = \left(1 + \frac{\omega_2}{\omega_3}\right)\left(\frac{\partial\mu_2}{\partial\omega_1}\right) + \frac{\omega_2}{\omega_3}\left(\frac{\partial\mu_1}{\partial\omega_1}\right), \tag{118}$$

$$c = \left(1 + \frac{\omega_1}{\omega_3}\right)\left(\frac{\partial\mu_1}{\partial\omega_2}\right) + \frac{\omega_2}{\omega_3}\left(\frac{\partial\mu_2}{\partial\omega_2}\right),$$
$$d = \left(1 + \frac{\omega_2}{\omega_3}\right)\left(\frac{\partial\mu_2}{\partial\omega_2}\right) + \frac{\omega_2}{\omega_3}\left(\frac{\partial\mu_1}{\partial\omega_2}\right). \tag{119}$$

From these relations it is easy to see that reciprocal relations $L_{12} = L_{21}$ imply that

$$aD_{12} + bD_{22} = cD_{11} + dD_{21}. \tag{120}$$

Experimental data for several three-component systems may be found in Miller (1960), Sugawara *et al.* (1977), and Spera and Trial (1993).

3.6 Heat Conduction in Anisotropic Solids

In an anisotropic solid, the flow of heat $\mathbf{J}_q$ may not be in the direction of the temperature gradient. The entropy production is

$$\sigma = \sum_{i=1}^{3} J_{qi} \frac{\partial}{\partial x_i}\left(\frac{1}{T}\right), \tag{121}$$

in which x_i are the Cartesian coordinates. The phenomenological laws for this system are

$$J_{qi} = \sum_k L_{ik} \frac{\partial}{\partial x_k}\left(\frac{1}{T}\right) = \sum_k \left(\frac{-L_{ik}}{T^2}\right)\frac{\partial T}{\partial x_k}. \tag{122}$$

For anisotropic solids the heat conductivity κ is a tensor of the second rank. The empirical Fourier law of heat conduction then is written as

$$J_{qi} = -\sum_k \kappa_{ik} \frac{\partial T}{\partial x_k}. \tag{123}$$

Comparison of (122) and (123) leads to

$$L_{ik} = T^2 \kappa_{ik}. \tag{124}$$

Reciprocal relations $L_{ik} = L_{ki}$ then imply that

$$\kappa_{ik} = \kappa_{ki}; \tag{125}$$

i.e., the heat conductivity is a symmetric tensor. However, for many solids the symmetry of the crystal structures itself implies that (125) be valid. Hence experimental verification of (125) would not be a confirmation of the reciprocal relations. On the other hand, for solids with trigonal (C_3,C_{3i}), tetragonal (C_4,S_4,C_{4h}), or hexagonal (C_6,C_{3h},C_{6h}) symmetry the crystal symmetries imply

$$\kappa_{12} = -\kappa_{21}. \tag{126}$$

If the reciprocal relations are valid, then

$$\kappa_{12} = \kappa_{21} = 0. \tag{127}$$

One method of experimental verification of this relation is due to Voigt and Curie as shown in Fig. 4. Another method may be found in Miller (1960, 1974).

3.7 Thermal Diffusion

The interaction between heat and matter flows results in two effects: the Soret effect and the Dufour effect. In the Soret effect, heat flow causes a flow of matter. In the Dufour effect, concentration gradients drive a heat flow. The effects and the reciprocal relations in this context can be obtained by writing the entropy production due to diffusion and heat flow. In this case, it is more convenient to write the entropy production using the "heat flow" defined as

$$\mathbf{J}_q \equiv \mathbf{J}_u - \sum_{k=1}^{n} h_k \mathbf{J}_k, \tag{128}$$

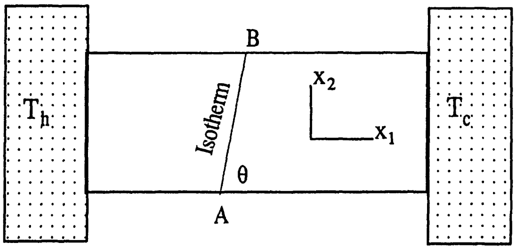

FIG. 4. The method of Curie–Voigt to verify the reciprocal relations for anisotropic heat conduction. An anisotropic solid whose crystal structure implies $\kappa_1 = -\kappa_2$ is placed in contact with two heat reservoirs of temperatures T_h and T_c. If the reciprocal relations are valid, then $\kappa_{12} = \kappa_{21} = 0$. If this is true, then the isotherms should be perpendicular to the direction x_1; i.e., θ should be 90°.

in which h_k is the partial specific enthalpy of the component k. In terms of $\mathbf{J}_q$ the entropy production may be written as

$$\sigma = \mathbf{J}_q \cdot \nabla\left(\frac{1}{T}\right) - \sum_{k=1}^{n-1} \mathbf{J}_k \cdot \frac{[\nabla(\mu_k - \mu_n)]_{T,p}}{T}, \tag{129}$$

in which Eq. (17) has been used to eliminate the flow $\mathbf{J}_n$. The corresponding chemical potential can be eliminated by use of the Gibbs–Duhem relation at constant p and T:

$$\sum_{k=1}^{n} \omega_k(\nabla\mu_k)_{p,T} = 0. \tag{130}$$

In terms of the $n - 1$ independent $\mathbf{J}_k$ and μ_k, the entropy production becomes

$$\sigma = \mathbf{J}_q \cdot \nabla\left(\frac{1}{T}\right) - \sum_{k,j=1}^{n-1} \mathbf{J}_k \cdot \frac{[\delta_{kj} + \omega_j/\omega_n](\nabla\mu_k)_{T,p}}{T}, \tag{131}$$

in which δ_{ij} is the Kronecker delta. The forces and flows can now be identified as

$$\mathbf{J}_q = L_{qq}\nabla\left(\frac{1}{T}\right) - \sum_{k,j=1}^{n-1} L_{qk} \frac{[\delta_{kj} + \omega_j/\omega_n](\nabla\mu_k)_{T,p}}{T}, \tag{132}$$

$$\mathbf{J}_k = L_{kq}\nabla\left(\frac{1}{T}\right) - \sum_{j,m=1}^{n-1} L_{kj} \frac{[\delta_{jm} + \omega_m/\omega_n](\nabla\mu_j)_{T,p}}{T}. \tag{133}$$

As was done in Sec. 3.5, to relate these equations to the empirical laws such as Fick's law, we may use

$$(\nabla\mu_k)_{p,T} = \sum_{j=1}^{n-1} \left(\frac{\partial\mu_k}{\partial\omega_j}\right)_{p,T} \nabla\omega_j, \tag{134}$$

in which ω_k are the mass fractions. The reciprocal relations in this case are

$$L_{qk} = L_{kq} \quad \text{and} \quad L_{jk} = L_{kj}. \tag{135}$$

If we consider a two-component system, the empirically measured constants such as the heat conductivity κ and the Dufour coefficient D'' are related to the phenomenological coefficients through the following (De Groot and Mazur 1969): heat conductivity,

$$\kappa = L_{qq}/T^2; \tag{136}$$

Dufour coefficient,

$$D'' = L_{q1}/\rho\omega_1\omega_2T^2; \tag{137}$$

Soret coefficient,

$$D' = L_{1q}/\rho\omega_1\omega_2T^2; \tag{138}$$

and diffusion coefficient,

$$D = L_{11}\left(\frac{\partial\mu_1}{\partial\omega_1}\right)_{p,T}(\rho\omega_2T)^{-1}. \tag{139}$$

[Note that if the mass fraction $\omega_2 \approx 1$, Eq. (139) reduces to Eq. (71).] The reciprocal relation $L_{q1} = L_{1q}$ implies that $D' = D''$.

3.8 Electrokinetic Phenomena and the Saxen Relations

Electrokinetic phenomena are due to the coupling between electrical current and matter flow. Consider two chambers, I and II, separated by a porous wall. We assume that the temperature and the concentrations are uniform throughout the system. If a voltage V is applied between the two chambers [see Fig. 5(a)], a current will flow until a pressure difference Δp is established at the steady state. This pressure difference is called the *electro-osmotic pressure.* Conversely, if a fluid flow J from one chamber to another is achieved by a piston, an electric current I, called the *streaming current,* flows through the electrodes. As before, the thermodynamic description of these effects begins with the expression for the entropy production under the conditions specified above. In this case we essentially have a discontinuous system in which there are no gradients but differences such as $\mu^{\mathrm{I}} - \mu^{\mathrm{II}}$. For discontinuous systems, the entropy production per unit volume σ is replaced by the total entropy production d_iS/dt. Furthermore, the entropy produced by the flow from chamber I to chamber II may be formally thought of as from a chemical reaction for which the difference in the electrochemical potential becomes the affinity (Prigogine, 1967). Thus we have

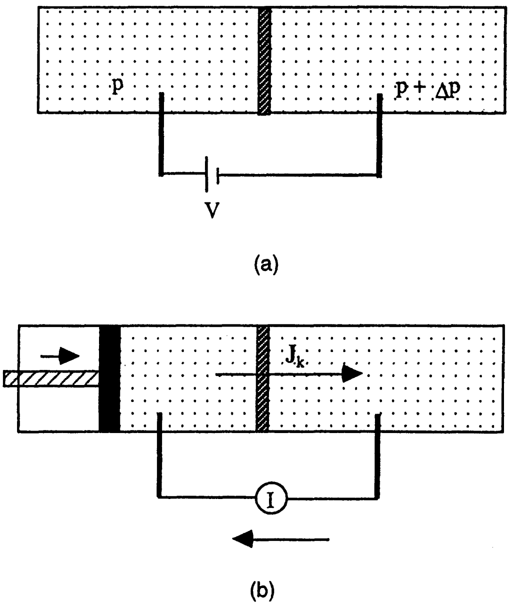

FIG. 5. Electrokinetic phenomena. Two chambers containing electrolytes are separated by a porous wall or capillary. (a) An applied potential V generates a pressure difference Δp called the *electro-osmotic pressure.* (b) If the fluid flows from one chamber to the other by action of a piston, there is generated an electrical current I called the *streaming current.*

$$\frac{d_iS}{dt} = \sum_k \frac{A_k}{T}\frac{d\xi_k}{dt}, \tag{140}$$

in which

$$A_k = (\mu_k^{\mathrm{I}} + z_kF\varphi^{\mathrm{I}}) - (\mu_k^{\mathrm{II}} + z_kF\varphi^{\mathrm{II}}), \tag{141}$$

$$d\xi_k = -dn_k^{\mathrm{I}} = dn_k^{\mathrm{II}}. \tag{142}$$

In the above equations z_k is the ion number of the component k, F is the Faraday constant, and φ is the electrical potential. For a relatively small difference in the pressure between the two chambers, since $\partial\mu_k/\partial p = \nu_k$, the partial specific volume, we may write

$$\mu_k^{\mathrm{I}} - \mu_k^{\mathrm{II}} = \nu_k\Delta p. \tag{143}$$

Equation (140) may now be written as

$$\frac{d_iS}{dt} = \frac{1}{T}\sum_k\left(-\nu_k\frac{dn_k^{\mathrm{I}}}{dt}\right)\Delta p + \frac{1}{T}\sum_k(-I_k)\Delta\varphi, \tag{144}$$

in which $\Delta\varphi = \varphi^{\mathrm{I}} - \varphi^{\mathrm{II}}$ and $I_k = z_kF dn_k^{\mathrm{I}}/dt$, the electric current due to the flow of component k. On combining of all the matter flow terms and the ion flow terms, (144) can now be written in the compact form

$$\frac{d_iS}{dt} = \frac{J\Delta p}{T} + \frac{I\Delta\varphi}{T}, \tag{145}$$

where

$$J = -\sum_k \nu_k\frac{dn_k^{\mathrm{I}}}{dt} \tag{146}$$

is the "volume flow" and

$$I = -\sum_k I_k \tag{147}$$

is the electric current. The phenomenological equations that follow from (145) are

$$I = L_{11}\Delta\varphi/T + L_{12}\Delta p/T \tag{148}$$

$$J = L_{21}\Delta\varphi/T + L_{22}\Delta p/T, \tag{149}$$

the reciprocal relations being

$$L_{12} = L_{21}. \tag{150}$$

Experimentally, the following quantities can be measured: the streaming potential,

$$\left(\frac{\Delta\varphi}{\Delta p}\right)_{I=0} = -\frac{L_{12}}{L_{11}}; \tag{151}$$

the electro-osmosis,

$$\left(\frac{J}{I}\right)_{\Delta p=0} = \frac{L_{21}}{L_{11}}; \tag{152}$$

the electro-osmotic pressure,

$$\left(\frac{\Delta p}{\Delta\varphi}\right)_{J=0} = -\frac{L_{21}}{L_{22}}; \tag{153}$$

and the streaming current,

$$\left(\frac{I}{J}\right)_{\Delta\varphi=0} = \frac{L_{12}}{L_{22}}. \tag{154}$$

As a consequence of the reciprocal relations $L_{12} = L_{21}$, we see from (151)–(154) that

$$\left(\frac{\Delta\varphi}{\Delta p}\right)_{I=0} = -\left(\frac{J}{I}\right)_{\Delta p=0}, \qquad (155)$$

$$\left(\frac{\Delta p}{\Delta\varphi}\right)_{J=0} = -\left(\frac{I}{J}\right)_{\Delta\varphi=0}. \qquad (156)$$

These two relations, called the Saxen relations, were obtained originally by kinetic considerations for particular systems, but by virtue of the formalism of nonequilibrium thermodynamics, we see their general validity.

3.9 Other Cross Effects

A number of other cross effects, such as thermomolecular pressure, thermomechanical effect, viscous flow in magnetic fields, and electrochemical phenomena, exist for which the above formalism can be applied. Description of these phenomena can be found in Miller (1960, 1974); Katchalsky (1965); Prigogine (1967); De Groot and Mazur (1969); Kreuzer (1981); Haase (1990); and Jou (1993). In the case of chemical reactions, the reciprocal relations imply the generally accepted principle of detailed balance. According to the principle of detailed balance, at equilibrium, every chemical reaction is balanced by the corresponding reverse reaction. In all cases that have been experimentally studied, Onsager's reciprocal relations were verified within experimental uncertainty.

3.10 Dielectric and Magnetic Relaxation

The formalism presented in Sec. 1 can also be used to describe dielectric and magnetic relaxation. When an electric or magnetic field is applied to the system, the electric and magnetic dipole moments relax to their equilibrium value. This irreversible process is accompanied by the production of entropy. Let us consider a simple example of a system at rest in which the only irreversible process is the relaxation of the electric dipole moment. For magnetic relaxation the magnetic dipole moment takes the place of the electric dipole moment. By using the methods similar to those used in Sec. 1, it can be shown (De Groot and Mazur, 1969) that the entropy production associated with electric relaxation is given by

$$\sigma = -\frac{1}{T}\frac{\partial \mathbf{P}}{\partial t}\cdot[\mathbf{E}_{\mathrm{eq}} - \mathbf{E}(t)] \geq 0, \qquad (157)$$

in which $\mathbf{P}$ is the dipole moment per unit volume, $\mathbf{E}_{\mathrm{eq}}$ is the electric field at equilibrium, and $\mathbf{E}(t)$ is the instantaneous electric field. For this case, the linear phenomenological relation between flow and the force is

$$\frac{\partial \mathbf{P}}{\partial t} = -\frac{L}{T}[\mathbf{E}_{\mathrm{eq}} - \mathbf{E}(t)]. \qquad (158)$$

If the electrical susceptibility is χ_e, then $\mathbf{P} = \chi_e \mathbf{E}_{\mathrm{eq}}$, and the above equation can be written as

$$\frac{\partial \mathbf{P}}{\partial t} = -\frac{L}{T\chi_e}[\mathbf{P} - \chi_e \mathbf{E}(t)]. \qquad (159)$$

For a static field $\mathbf{E}(t) = \mathbf{E}_0$, the solution of the above equation is

$$\mathbf{P}(t) = \chi_e \mathbf{E}_0(1 - e^{-t/\tau}), \qquad (160)$$

in which the relaxation time is $\tau = \chi_e T/L$. For a time-dependent $\mathbf{E}(t)$, the Fourier transform of (159) gives

$$\tilde{\mathbf{P}}(\omega) = \tilde{\chi}_e(\omega)\tilde{\mathbf{E}}(\omega), \qquad \tilde{\chi}_e(\omega) = \chi_e/(1 - i\omega\tau), \qquad (161)$$

in which the quantities with a tilde denote the Fourier transforms. The inverse Fourier transform of (161) gives $\mathbf{P}(t)$, the polarization as function of time.

3.11 Theorem of Minimum Entropy Production

A system maintained in the linear nonequilibrium regime reaches a steady state. A simple example is a system in contact with two thermal reservoirs of different temperatures (Fig. 6). For systems in the linear regime, where the linear phenomenological laws and the Onsager reciprocal relations are valid, the following general theorem due to Prigogine (Prigogine 1947, 1967) characterizes the stationary states:

> In the linear regime, the total internal entropy production, d_iS/dt reaches a minimum value at the nonequilibrium stationary state.

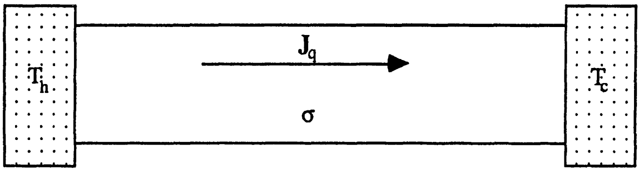

FIG. 6. A conducting solid in contact with two heat reservoirs at temperatures T_h (hot) and T_c (cold). The temperature of such a solid reaches a stationary state in which the total entropy production is minimized.

This theorem may be thought of as an extension of the situation at equilibrium at which $d_iS/dt = 0$: the entropy production, unable to reach zero because of nonequilibrium constraints, reaches the minimum possible value. The following examples illustrate this theorem.

3.11.1 Stationary States in Thermal Conduction Let us consider a system of length L in contact with a hot thermal reservoir with a temperature T_h (hot) at one end and a cold thermal reservoir with a temperature T_c (cold) at the other (see Fig. 6). We assume that the conduction of heat is the only irreversible process. The entropy production per unit volume is given by

$$\sigma = \mathbf{J}_q \cdot \nabla(1/T). \tag{162}$$

In the linear regime, we have the phenomenological law $\mathbf{J}_q = L_{qq}\nabla(1/T)$. The total entropy production P is given by

$$P \equiv \frac{d_iS}{dt} = \int_V L_{qq}\left(\nabla \frac{1}{T}\right)^2 dV. \tag{163}$$

According to the theorem of minimum entropy production, the temperature distribution at the stationary state corresponds to the one that minimizes P. Close to equilibrium, we may assume that L_{qq} is nearly constant throughout the system. Using the calculus of variation, one could obtain the following Euler–Lagrange equation whose solution minimizes P:

$$\nabla \cdot [\nabla(1/T)] = 0 \quad \text{or} \quad \nabla \cdot \mathbf{J}_q = 0. \tag{164}$$

That Eq. (164) represents the stationary state can be confirmed as follows. For the flow of heat the conservation of energy gives

$$\rho \frac{\partial u}{\partial t} = -\nabla \cdot \mathbf{J}_q. \tag{165}$$

The change in the energy due to heat flow is $du = c_V dT$, in which c_V is the specific heat. Hence (165) may be written as an equation for the temperature T:

$$\rho c_V \frac{\partial T}{\partial t} = -\nabla \cdot \mathbf{J}_q. \tag{166}$$

We can now see that (164) and (166) imply that the state that minimizes P by satisfying (164) is also a stationary state for which $\partial T/\partial t = 0$.

3.11.2 Stationary States in Chemical Systems The open near-equilibrium chemical system that we consider is shown in Fig. 7. The chemical reaction scheme is

$$A \overset{(1)}{\rightleftharpoons} X \overset{(2)}{\rightleftharpoons} B. \tag{167}$$

We assume that suitable flows of A and B keep the chemical potentials μ_A and μ_B fixed. This in turn implies that the sum of the affinities of reactions (1) and (2), $A_1 + A_2 \equiv \bar{A}$, is also fixed:

$$\begin{aligned} A_1 + A_2 &= (\mu_A - \mu_X) + (\mu_X - \mu_B) \\ &= \mu_A - \mu_B \equiv \bar{A}. \end{aligned} \tag{168}$$

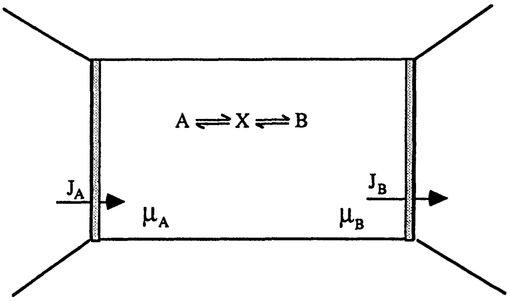

FIG. 7. An open chemical system in which the chemical potentials μ_A and μ_B are maintained at a given nonequilibrium value by an inflow of the component A and an outflow of the component B. In this system the concentration of X is maintained at a nonequilibrium value.

Let v_1 and v_2 be the velocities of the two reactions in (167). We shall now obtain the condition that v_1 and v_2 must satisfy to minimize the entropy production. The entropy production per unit volume for this system (which we assume is homogeneous), can be written as

$$\frac{1}{V}P = \frac{1}{V}\frac{d_iS}{dt} = \sigma = \frac{A_1}{T}v_1 + \frac{A_2}{T}v_2 = \frac{A_1}{T}v_1 + \frac{(\bar{A} - A_1)}{T}v_2, \tag{169}$$

in which V is the system volume and in which we have included the constraint (168) explicitly. The value of the chemical potential (or concentration) of X determines the value of A_1 and hence the value of the entropy production (169). In the linear regime, since the two reactions are independent, we have

$$\begin{aligned} v_1 &= L_{11}A_1, \\ v_2 &= L_{22}A_2 = L_{22}(\bar{A} - A_1), \end{aligned} \tag{170}$$

in which we have used (168). Substituting (170) into (169) we obtain σ as a function of A_1:

$$\sigma(A_1) = L_{11}\frac{A_1^2}{T} + L_{22}\frac{(\bar{A} - A_1)^2}{T}. \tag{171}$$

This function reaches its minimum value when

$$\frac{\partial\sigma(A_1)}{\partial A_1} = \frac{L_{11}}{T}2A_1 - \frac{L_{22}}{T}2(\bar{A} - A_1) = 0, \tag{172}$$

i.e.,

$$L_{11}A_1 - L_{22}A_2 = v_1 - v_2 = 0. \tag{173}$$

It is easy to see that the condition $v_1 = v_2$ corresponds to the stationary state. In the reaction scheme (167), since the concentrations of A and B are fixed, X is the only independent variable. Since reaction 1 produces X and reaction 2 consumes X, the kinetic equation for the concentration $[X]$ of X is

$$\frac{d[X]}{dt} = v_1 - v_2. \tag{174}$$

At the stationary state $d[X]/dt = 0$, which is the same as the state that minimizes the entropy production.

3.12 Entropy Production and the Stability of Stationary States

In the previous section we have seen that the stationary states in the linear regime are also states that minimize the entropy production. The *stability* of these stationary states can also be studied using the entropy production P. First we note that the approach to equilibrium can be described not only as a steady increase in entropy to its maximum value but also as a steady decrease in entropy production to zero. The stability of the equilibrium state is guaranteed by the fact that any perturbation from equilibrium is countered by irreversible processes that drive the system back to equilibrium. This fact may also be seen as the system being driven by irreversible processes to a state in which the entropy production vanishes. It is this later approach that naturally extends to the linear regime close to equilibrium.

The stationary states near equilibrium have a constant positive value for the entropy production P that is also a minimum. To establish the stability of this state, we must show that for all perturbations from the stationary state

$$\frac{dP}{dt} < 0. \tag{175}$$

The validity of this inequality depends on the fact that quantities such as specific heat c_v are positive. For example, in the case of the heat conduction discussed in Sec. 3.10.1, the time derivative of P given by (163) is

$$\frac{dP}{dt} = \int_V L_{qq}2\left(\nabla\frac{1}{T}\right)\cdot\left(\nabla\frac{\partial}{\partial t}\frac{1}{T}\right)dV. \tag{176}$$

By use of the phenomenological relation $\mathbf{J}_q = L_{qq}\nabla(1/T)$ and the identity $\nabla\cdot(\mathbf{A}f) = f(\nabla\cdot\mathbf{A}) + \mathbf{A}\cdot(\nabla f)$, Eq. (176) may be written as

$$\frac{dP}{dt} = 2\int_V \nabla\cdot\left(\mathbf{J}_q\frac{\partial}{\partial t}\frac{1}{T}\right)dV - \int_V\left(\frac{\partial}{\partial t}\frac{1}{T}\right)\nabla\cdot\mathbf{J}_q dV. \tag{177}$$

By Gauss's theorem the first integral can be expressed as a surface integral. Since the boundary conditions are time independent, on the boundary $\partial(1/T)/\partial t = 0$, and the first integral vanishes. By using Eq. (166) in the second integral of Eq. (177) we find

$$\frac{dP}{dt} = -2\int_V \frac{\rho c_v}{T^2}\left(\frac{\partial T}{\partial t}\right)^2 dV. \tag{178}$$

Since $c_v > 0$, it follows that $dP/dt < 0$. In fact, this is generally true for all stationary states in the linear regime in which the linear phenomenological relations are valid. The general situation is summarized in Fig. 8.

4. FAR-FROM-EQUILIBRIUM SYSTEMS

By increasing the flow of matter and energy, a system can be driven farther and farther from the state of equilibrium. In contrast to equilibrium and near-equilibrium states, there is no general principle that characterizes far-from-equilibrium states or guarantees their stability. Indeed, these states can become unstable and evolve to new states. The new states thus reached can be complex and highly organized. In fact, much of the complexity and organized behavior we see in nature occurs under far-from-equilibrium conditions. The complex behavior is a consequence of irreversible processes such as chemical reactions and diffusion. In this section we present an overview of entropy production, stability, and self-organization under far-from-equilibrium conditions.

4.1 General Properties of Entropy Production

As we have seen, in the linear regime the total entropy production $P = \int_V \sigma dV$ reaches its minimum value for the stationary states and $dP/dt < 0$. In the far-from-equilibrium nonlinear regime, there are no such general principles that characterize stationary states. There are, however, some general statements that can be made regarding the total entropy production P. Since $P = \int_V \sigma dV = \int_V \Sigma_\alpha X_\alpha J_\alpha dV$,

$$\frac{dP}{dt} = \int_V \left(\sum_\alpha \frac{\partial X_\alpha}{\partial t} J_\alpha\right) dV + \int_V \left(\sum_\alpha X_\alpha \frac{\partial J_\alpha}{\partial t}\right) dV \equiv \frac{d_X P}{dt} + \frac{d_J P}{dt}, \tag{179}$$

in which $d_X P/dt$ is the first integral, which contains the time derivative of the forces X_α, and $d_J P/dt$ is the second integral, which contains the time derivative of the flows J_α. We state without proof two general properties that have been proved by Glansdorff and Prigogine (1971):

1. In the linear regime,

$$\frac{d_X P}{dt} = \frac{d_J P}{dt}. \tag{180}$$

2. Under all conditions,

$$\frac{d_X P}{dt} < 0. \tag{181}$$

Equations (179) and (180) give the earlier result that in the linear regime $dP/dt < 0$. Though Eqs. (180) and (181) are among the

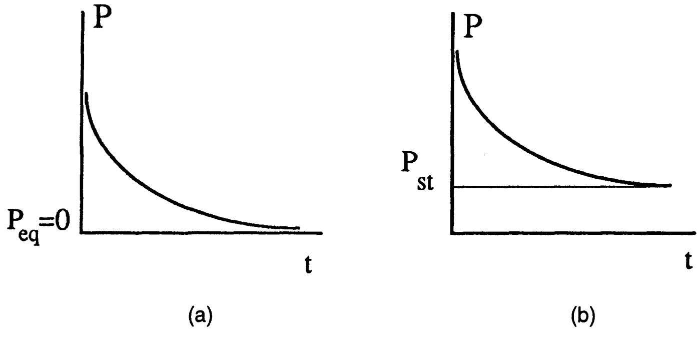

(a) (b)

FIG. 8. The time variation of the entropy production $P = d_iS/dt$ for equilibrium and near-equilibrium states. (a) For a fluctuation from the equilibrium state, the initial nonzero value of P decreases to its equilibrium value of zero. (b) In the linear regime, a fluctuation from a nonequilibrium stationary state in which $P = P_{st}$ increases P; the subsequent time development of P drives the system back to the stationary state.

most general result in thermodynamics, they do not provide a means to specify nonequilibrium states or their stability.

4.2 Stability of Far-from-Equilibrium Systems

A general condition for the stability of any state is the existence of a Lyapunov functional. A functional is a rule by which a number can be associated with a state. Let y be a perturbation from a state Y_0 whose stability is of interest. A Lyapunov functional $\Lambda(y)$ associated with a state Y_0 has the following properties:

1. For all perturbations y,

$$\Lambda(y) > 0. \tag{182}$$

2. For all times t,

$$\frac{d\Lambda(y)}{dt} \le 0. \tag{183}$$

If such a Λ exists, then Y_0 is stable. In the linear regime, we see that the total entropy production P is a Lyapunov functional.

Under all conditions, a functional that always has a definite sign is the second variation of entropy $\delta^2 S$, which is defined as follows (Glansdorff and Prigogine, 1971). The change in entropy ΔS due to a perturbation δx_k from a stationary state x_{sk} can be written as

$$\begin{aligned}\Delta S &= \int \left[\sum_k \left(\frac{\partial(\rho s)}{\partial x_k}\right)_{xk=xsk} \delta x_k\right] dV \\ &\quad + \frac{1}{2}\int \left[\sum_{ij} \left(\frac{\partial^2 \rho s}{\partial x_i \partial x_j}\right)_{xk=xsk} \delta x_i \delta x_j\right] dV \\ &= \delta S + \tfrac{1}{2}\,\delta^2 S,\end{aligned} \tag{184}$$

in which δS is the first integral and $\delta^2 S$ is the second integral. Based on the validity of local equilibrium, it has been shown (Glansdorff and Prigogine, 1971) that

$$\begin{aligned}\delta^2 S = &-\int \frac{\rho}{T}\left[\frac{c_V}{T}(\delta T)^2 + \frac{\rho}{\chi}(\delta \nu)_c^2 \right. \\ &\left. + \sum_{i,j}\left(\frac{\partial \mu_k}{\partial \omega_j}\right)_{T,p} \delta\omega_i \delta\omega_j\right] dV \le 0,\end{aligned} \tag{185}$$

in which

$$(\delta\nu)_c = \left(\frac{\partial \nu}{\partial T}\right)_{p,\omega k} \delta T + \left(\frac{\partial \nu}{\partial p}\right)_{T,\omega k} \delta p \tag{186}$$

and

$$\chi = -\frac{1}{\nu}\left(\frac{\partial \nu}{\partial p}\right)_{T,ck}. \tag{187}$$

Thus $-\delta^2 S$ is a functional that satisfies one of the conditions (182) for a Lyapunov functional. In addition, if $d(-\delta^2 S)/dt \le 0$, we have a Lyapunov functional. Now it has been shown (Glansdorff and Prigogine, 1971) that the derivative of $-\delta^2 S$ has the general form

$$\frac{d(-\delta^2 S)}{dt} = -\sum_\alpha \delta X_\alpha \delta J_\alpha, \tag{188}$$

in which δX_α and δJ_α are the perturbations of the forces and flows from the stationary state whose stability is of interest. Thus the general *sufficient* condition for the stability of a nonequilibrium stationary state is

$$\sum_\alpha \delta X_\alpha \delta J_\alpha \ge 0. \tag{189}$$

The quantity $\Sigma_\alpha \delta X_\alpha \delta J_\alpha$ is called the *excess entropy production.* One must note that (189) is sufficient but not a necessary condition for stability. When (189) is not satisfied, i.e., $\Sigma_\alpha \delta J_\alpha \delta F_\alpha \le 0$, the nonequilibrium steady state *may* become unstable. As a consequence of the instability, a perturbation may grow, driving the system to a new state.

In chemical systems, violation of the condition (189) occurs in autocatalytic systems. As an example we may consider the autocatalytic reaction

$$2X + Y \underset{k_r}{\overset{k_f}{\rightleftharpoons}} 3X, \tag{190}$$

which appears in a reaction scheme called the "Brusselator" that we consider below. For this reaction, let a nonequilibrium stationary state have the concentrations $[X]_s$ and $[Y]_s$ and consider a perturbation δX. Using the forward and reverse rates $\rho_f = k_f[X]^2[Y]$ and $\rho_r = k_r[X]^3$ in the general expressions for A and v, one can calculate the expression for the excess entropy production and obtain

$$\sum_{\alpha} \delta A_{\alpha} \delta v_{\alpha} = -R(2k_f[X]_s[Y]_s - 3k_r[X]_s^2)\frac{(\delta X)^2}{[X]_s}. \tag{191}$$

The excess entropy production (191) can become negative, particularly if $k_r \ll k_f$. Hence the stability is no longer assured and the steady state *may* become unstable.

Instability often leads to organized states as summarized in Fig. 9. In it, the value of the parameter Δ is a measure of the distance from equilibrium. For each value of Δ, the system will relax to a stationary state, denoted by X_s. The equilibrium state corresponds to $\Delta = 0$. X_s is a continuous extension of the equilibrium state, and it is called the *thermodynamic branch.* As long as the condition (189) is satisfied, the thermodynamic branch is stable. If (189) is violated, then the thermodynamic branch may become unstable.

4.3 Linear Stability Analysis and Bifurcation of New States

The above-mentioned instability of the thermodynamic branch often gives rise to organized states. These states, which are created and maintained by dissipative processes, are called *dissipative structures.* The onset of instability as the system is forced away from equilibrium is analyzed by means of linear stability theory. Consider a system whose state is represented by variables x_k. Let λ be a parameter representing the value of a constraint that keeps the system away from equilibrium. We wish to study the behavior of the system as λ changes. The conservation laws and other empirical laws for the rates of chemical reaction and heat conduction give us a set of equations for the time variation of x_k, which we represent by

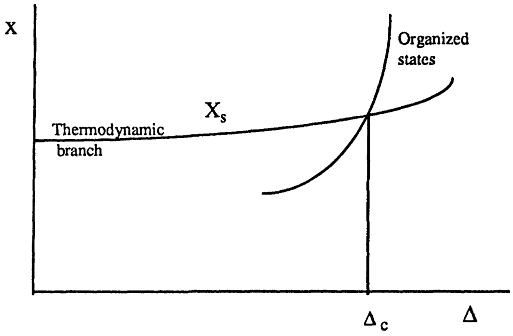

FIG. 9. Here each value of X represents a state of the system. The distance from equilibrium is represented by the parameter Δ. When Δ is small, the system is in a near-equilibrium state, which is an extrapolation of the equilibrium state; this family of states is called the *thermodynamic branch.* In some systems, such as those that have autocatalysis, when Δ reaches a critical value Δ_c, the state belonging to the thermodynamic branch becomes unstable. When this happens, the system makes a transition to a new branch of organized states.

$$\frac{dx_i}{dt} = F_i(x_1 \ldots x_n; \lambda). \tag{192}$$

We assume that there is a stationary state x_{sk} whose stability is of interest. Since x_{sk} is a stationary state, we have

$$F_k(x_{s1}, x_{s2}, \ldots, x_{sn}, \lambda) = 0. \tag{193}$$

The stability of the state x_{sk} depends on the eigenvalues of the Jacobian matrix M whose elements are

$$M_{ij} \equiv \left(\frac{\partial F_i}{\partial x_j}\right)_{x_k = x_{sk}}. \tag{194}$$

[If Eq. (192) is a partial differential equation, then a linear partial differential operator M is defined.] If the real parts of all the eigenvalues of M are negative, the stationary state x_{sk} is stable. As λ changes, a negative real eigenvalue or a pair of complex eigenvalues with negative real parts may cross the imaginary axis. One or more eigenvalues crossing the imaginary axis defines the onset of instability. At this point, new solutions to Eq. (192) become available. The new states are the dissipative structures. A detailed description of the onset of instability and the appearance "bifurcation" of new solution (which may be found in Nicolis and Prigogine, 1977) is beyond the scope of this article.

4.4 Dissipative Structures

In general, dissipative structures have lower entropy compared to that of a corresponding equilibrium system. As an illustration of a nonequilibrium transition to a dissipative structure, let us consider a few examples.

The first is a simple example related to a fundamental aspect of biochemistry. It is well known that the building blocks of life, viz., amino acids, DNA, and RNA, have a fundamental asymmetry: of the two possible mirror-image molecular structures, only one appears in nature. Molecules that are not identical to their mirror image are said to be *chiral,* and the two forms or *enantiomers* are termed L (levo) and D (dextro). Amino acids are almost all of the L form, and the ribose in DNA and RNA is of the D form. This is remarkable because chemical reactions that produce chiral molecules from achiral molecules produce equal numbers of L and D molecules. The "organized state" that contains predominantly one of the two chiral forms is an example of a dissipative structure that arises under far-from-equilibrium conditions.

Through a variation of a model reaction devised by Frank (1953) (see Fig. 10), we can see how a set of chemical reactions that produce mirror image forms X_L and X_D of a molecule X, with equal preference, can nevertheless make a transition to a state dominated by either X_L or X_D. The reaction scheme we consider, which includes two autocatalytic reactions, is shown in the upper part of Fig. 10. In a closed or an isolated system, the set of reactants S, T, X_L, X_D, and P will reach the state of equilibrium, which is a *symmetric state,* i.e., $[X_L] = [X_D]$. The symmetric state corresponds to the thermodynamic branch. However, if we consider an open system into which S and T are pumped and P is removed, then on account of the instability of the thermodynamic branch created by the autocatalysis, the system can break the chiral symmetry and make a transition to an *asymmetric state* in which $[X_L] \neq [X_D]$ (For details see Kondepudi and Nelson, 1984, 1985).

The onset of instability and transition to a state of broken symmetry may be analyzed by defining a parameter $\lambda = [S][T]$. When λ is greater than λ_c, the symmetric state (the thermodynamic branch) becomes unstable. The symmetric and asymmetric states can be associated with a variable $\alpha = [X_L] - [X_D]$. The transition from a symmetric state to an asymmetric state can now be described by

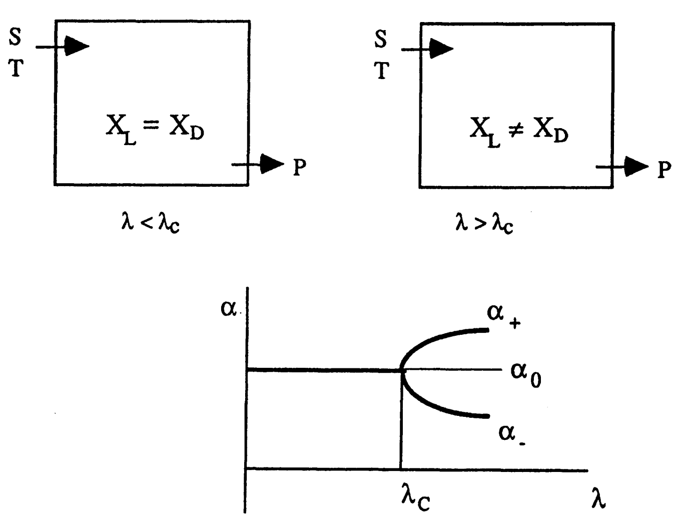

FIG. 10. A reaction scheme showing some general features of transitions to dissipative structures. The autocatalytic reaction scheme in which X_L and X_D are produced with equal preference in an open system leads to a dissipative structure in which $X_L \neq X_D$.

the variables λ and α. When the transition to an asymmetric state occurs, α can be positive or negative. Thus there are two possible asymmetric states to which the system can make a transition. The variables α and λ obey a general equation of the form

$$\frac{d\alpha}{dt} = -A\alpha^3 + B(\lambda - \lambda_c)\alpha, \tag{195}$$

in which A and B are constants. The form of this equation is independent of the details of the particular chemical scheme that breaks the chiral symmetry; only the numerical values of A and B depend on the details of the kinetics. The stationary states of α, obtained from (195) by setting $d\alpha/dt = 0$, are shown in Fig. 10. They are given by:

$$\alpha = 0 \quad \text{and} \quad \alpha = \pm[(B/A)(\lambda - \lambda_c)]^{1/2}. \tag{196}$$

When $\lambda > \lambda_c$, the state $\alpha = 0$ becomes unstable and the system evolves to one of the two states $\alpha = \pm[(B/A)(\lambda - \lambda_c)]^{1/2}$. The instability described by Eq. (195) and the associated steady-state diagram is generally valid for any nonequilibrium system that breaks a twofold symmetry. In each of these systems, parameters α and λ can be defined that, in the vicinity of the critical point λ_c, obey an equation of the form (195).

The next example we consider illustrates oscillations and spatial structures in chemical systems. For this purpose, we use a model reaction scheme, which has acquired the name "Brusselator." Proposed by Prigogine and Lefever in 1968 (Prigogine and Lefever, 1968; Nicolis and Prigogine, 1977), this model has been used extensively as a prototype to gain an understanding of the various aspects of chemical dissipative structures. Since then, several other models have been studied in great detail both experimentally and theoretically (Nicolis and Prigogine, 1977; Vidal and Pacault, 1981; Field and Burger, 1985; Baras and Walgraef, 1992). Many of the general features discovered theoretically have been experimentally realized in the most extensively studied Belusov–Zhabotinsky (B-Z) reaction (Belusov, 1958; Zhabotinsky, 1974) for which a kinetic model was proposed by Field and Noyes (1974).

The reaction scheme of the Brusselator is shown in Fig. 11. We assume that the concentrations of the reactants A and B are maintained at a desired nonequilibrium value through appropriate flows. The products D and E are removed as they are formed. All the reverse reactions are assumed to be slow enough that they could be neglected. The rates of reaction are assumed to obey mass-action kinetics. For example, the rate of the first reaction in the scheme is $k_1[A]$, that of the second reaction is $k_2[B][X]$, and so on. Since all other concentrations are fixed, the state of the system is determined by the concentrations $[X]$ and $[Y]$. Transport of X and Y occurs through diffusion. With these assumptions, it is straightforward to write the kinetic equations for the concentrations of X and Y. For mathematical convenience, one can redefine "normalized" concentrations (Nicolis and Prigogine, 1977) in such a way that all the rate constants are equal to unity. These normalized concentrations, which are denoted by X, Y, etc., obey the following equations, in which D_x and D_y are the diffusion coefficients:

$$\frac{\partial X}{\partial t} = A - (B + 1)X + X^2Y + D_x\nabla^2X, \tag{197}$$

$$\frac{\partial Y}{\partial t} = BX - X^2Y + D_y\nabla^2Y. \tag{198}$$

A detailed analysis of this (and many other models) may be found in Nicolis and Prigogine (1977). Here we state some of the main features. For appropriate values of the diffusion coefficients and of concentrations A and B, the concentrations of X and Y may exhibit homogeneous oscillations, form stationary spatial patterns, or form time-dependent patterns that are propagating waves. Almost every feature observed experimentally appears in this system under appropriate conditions.

When the two diffusion coefficients are nearly equal, for a fixed value of A, as the value of B increases, the thermodynamic branch becomes unstable. The reason for this instability is the autocatalytic reaction $2X + Y \rightarrow 3X$. When D_x is nearly equal to D_y, the condition for the instability is

$$B > 1 + A^2 + (\pi^2m^2/l^2)(D_x + D_y). \tag{199}$$

This condition is derived by linear stability

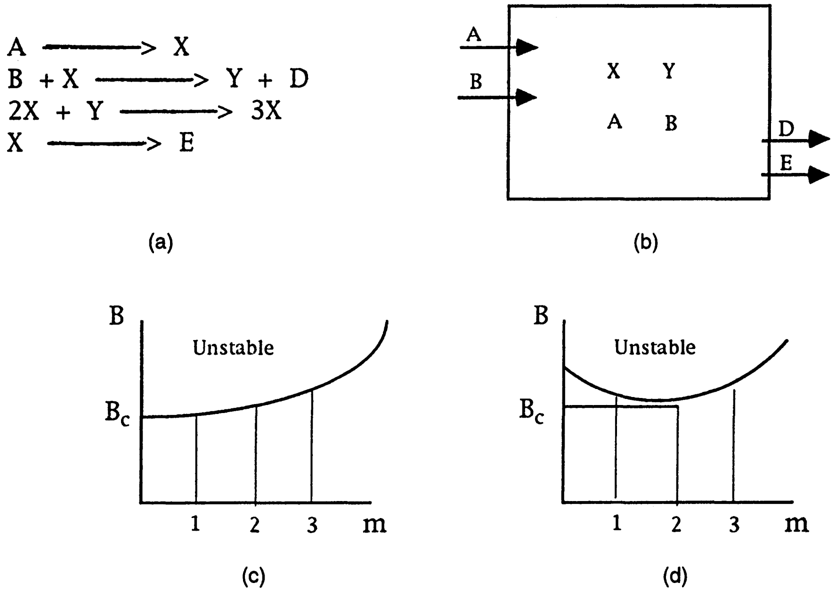

(a) (b)

(c) (d)

FIG. 11. (a) Reaction scheme of the Brusselator. (b) The conditions under which dissipative structures are formed are indicated. The concentrations of reactants A and B are maintained at constant values through an inflow, and the products D and E are removed. Under these conditions, the concentrations of X and Y may oscillate in time or form spatial patterns, depending on the values of the diffusion coefficients D_X and D_Y of X and Y. (c) Stability diagram showing the value of B for which modes of wave number m become unstable when D_X and D_Y are nearly equal. The instability occurs when $B = B_c$ at $m = 0$, leading to homogeneous oscillations. (d) Stability diagram showing the value of B for which modes of wave number m become unstable when the difference between D_X and D_Y is large. In this case when $B = B_c$, the corresponding mode $m = 2$ becomes unstable, leading to the formation of a spatial pattern.

analysis. In this equation, l is the size of the system, and m is an integer that corresponds to the wave number of the spatial mode that becomes unstable. Clearly, the smallest value of B that will satisfy this inequality corresponds to $m = 0$, which is the homogeneous mode. The value of B at which the mode $m = 0$ becomes unstable is shown graphically in Fig. 11(c). Here we see that when B reaches the critical value $B_c = 1 + A^2$, the homogeneous mode becomes unstable. This instability leads to the growth of homogeneous oscillation of X and Y. The amplitude of these oscillations depends on the value of $B - B_c$.

Spatial patterns can arise when the difference between D_x and D_y is large. The precise condition for this instability of the thermodynamic branch is given by the following expression (Nicolis and Prigogine, 1977):

$$B > 1 + \frac{D_x}{D_y}A^2 + \frac{A^2}{D_y\pi^2 m^2}l^2 + \frac{D_x\pi^2 m^2}{l^2}. \tag{200}$$

In this case, the minimum value of B that will satisfy this condition does not correspond to $m = 0$, but to a nonzero value of m. The minimum value of B that satisfies (200) defines the onset of instability. It can be shown (Nicolis and Prigogine, 1977) that

$$B_{\min} = \left[1 + A\left(\frac{D_x}{D_y}\right)^{1/2}\right]^2 \quad \text{when}$$
$$m = m_{\min} = \frac{Al^2}{\pi^2(D_1D_2)^{1/2}}. \tag{201}$$

If $B > B_{\min}$, all the integer values of m that satisfy (200) represent unstable modes. This is shown graphically in Fig. 11(d). Each of the unstable modes can grow. Similarly, under appropriate conditions, the Brusselator also shows that propagating waves may arise.

Though there are no general thermodynamic results for far-from-equilibrium systems, many interesting observations have been made by several people such as John Ross, Cathy Hunt, Paul Hunt, and others. A representative list of these works has been included in the list for further reading. In addition, there has been a lot of work in the statistical theory of nonequilibrium thermodynamics. This work constitutes a subject in itself and requires a separate treatment for a full description. Books and other comprehensive works on this subject are also listed under Further Reading.

Works Cited

Baras, F., Walgraef, D. (Eds.) (1992), Special issue on "Nonequilibrium Chemical Dynamics: From Experiment to Microscopic Simulation," Physica A **188.**

Belusov, B. B. (1958), *Sb. Ref. Radiats. Med. Moscow,* 145–146.

Casimir, H. G. B. (1945), "On Onsager's Principle of Microscopic Reversibility," *Rev. Mod. Phys.* **17,** 343–350.

Coleman, B. D. (1964), "Thermodynamics of Materials with Memory," *Arch. Rat. Mech. Anal.* **17,** 1–46.

Coleman, B. D., Noll, W. (1963), "The Thermodynamics of Elastic Materials with Heat Conduction and Viscosity," *Arch. Rat. Mech. Anal.* **13,** 167–178.

De Donder, T., Van Rysselberghe, P. (1936), *Affinity,* Menlo Park, CA: Stanford Univ. Press.

De Groot, S. R., Mazur, P. (1969), *Non-Equilibrium Thermodynamics,* Amsterdam: North-Holland.

Field, J., Noyes, R. M. (1974), *J. Chem. Phys.* **60,** 1877–1884.

Field, R. J., Burger, M. (Eds.) (1985), *Oscillations and Traveling Waves in Chemical Systems,* New York: Wiley.

Frank, F. C. (1953), *Biochem. Biophys. Acta* **11,** 459–463.

Glansdorff, P., Prigogine, I. (1971), *Thermodynamics of Structure Stability and Fluctuations,* New York: Wiley-Interscience.

Haase, R. (1990), *Thermodynamics of Irreversible Processes,* New York: Dover.

Jou, D. (1993), *Extended Irreversible Thermodynamics,* New York: Springer.

Katchalsky, A. (1965), *Nonequilibrium Thermodynamics and Biophysics,* Cambridge, MA: Harvard Univ. Press.

Kondepudi, D. K., Nelson, G. W. (1984), "Chiral Symmetry Breaking States and Their Sensitivity in Nonequilibrium Chemical Systems," *Physica* **125A,** 465–496.

Kondepudi, D. K., Nelson, G. W. (1985), "Weak Neutral Currents and the Origin of Biomolecular Chirality," *Nature* **314,** 438–441.

Kreuzer, H. J. (1981), *Nonequilibrium Thermodynamics and Its Statistical Foundations,* Oxford, U.K.: Clarendon Press.

Miller, D. G. (1960), "Thermodynamics of Irreversible Processes," *Chem. Rev.* **60,** 15–37.

Miller, D. G. (1974), "The Onsager Relation: Experimental Evidence," *Foundation of Continuum Thermodynamics,* London: Macmillan.

Müller, I., Ruggeri, T. (1993), *Extended Thermodynamics,* New York: Springer-Verlag.

Nicolis, G., Prigogine, I. (1977), *Self-Organization in Nonequilibrium Systems,* New York: Wiley-Interscience.

Noll, W. (1974), *Foundations of Mechanics and Thermodynamics,* Berlin: Springer.

Onsager, L. (1931), "Reciprocal Relations in Irreversible Processes I," *Phys. Rev.* **37,** 405–426.

Onsager, L. (1931), "Reciprocal Relations in Irreversible Processes II," *Phys. Rev.* **38,** 2265–2279.

Prigogine, I. (1947), *Etude Thermodynamique des Processus Irreversibles,* Liege: Desoer.

Prigogine, I. (1949), "Le Domaine de Validité de la Thermodynamique des Phenoménes Irreversibles," *Physica* **15,** 272–284.

Prigogine, I. (1967), *Introduction to Thermodynamics of Irreversible Processes,* New York: Wiley.

Prigogine, I., Defay, R. (1967), *Chemical Thermodynamics,* London: Longmans.

Prigogine, I., Lefever, R. (1968), *J. Chem. Phys.* **48,** 1695–1700.

Spera, F. J., Trial, A. (1993), "Verification of the Onsager Reciprocal Relations in Molten Silicate Solution," *Science* **259,** 204–206.

Sugawara, H., Nagata, K., Goto, K. (1977), *Metall. Trans. B.* **8,** 605–615.

Truesdell, C. (1969), *Rational Thermodynamics,* New York: McGraw-Hill.

Vidal, C., Pacault, A. (Eds.) (1981), *Non-Linear Phenomenon in Chemical Dynamics,* Berlin: Springer.

Wigner, E. P. (1954), "Derivations of Onsager's Reciprocal Relations," *J. Chem. Phys.* **22,** 1912–1915.

Zhabotinsky, A. M. (1974), *Self-Oscillating Concentrations,* Moscow: Nauka.

Further Reading

Coleman, B. D., Truesdell, C. (1960), "On the Reciprocal Relations of Onsager," *J. Chem. Phys.* **33,** 28–31.

Eckart, C. (1940), "The Thermodynamics of Irreversible Processes I: The Simple Fluid," *Phys. Rev.* **58,** 267–269.

Eckart, C. (1940), "The Thermodynamics of Irreversible Processes II: Fluid Mixtures," *Phys. Rev.* **58,** 269–275.

Eckart, C. (1940), "The Thermodynamics of Irreversible Processes III: Relativistic Theory of Simple Fluids," *Phys. Rev.* **58,** 919–924.

Epstein, I., Kustin, K., De Kepper, P., Orban, M. (1983), *Sci. Am.* **248** (3), 112–118.

Epstein, I. R. (1989), *J. Chem. Ed.* **69**, 191.

Eu, B. C. (1992), *Kinetic Theory of Irreversible Thermodynamics*, New York: Wiley.

Forland, K. S. (1988), *Irreversible Thermodynamics: Theory and Applications*, New York: Wiley.

Gray, P., Scott, S. K. (1990), *Chemical Oscillations and Instabilities*, Oxford, U.K.: Clarendon Press.

Haase, R. (1990), *Thermodynamics of Irreversible Processes*, New York: Dover.

Hansen, N. F., Ross, J. (1996), "Lyapunov Functions and Relative Stability in Reaction-Diffusion Systems with Multiple Stationary States," *J. Phys. Chem.* **100**, 8040–8043.

Hasse, R. (1969), *Thermodynamics of Irreversible Processes*, Reading, MA: Addison-Wesley.

Hemptinne, X. d. (1992), *Non-Equilibrium Statistical Thermodynamics: Applied to Fluid Dynamics and Laser Physics*, River Edge, NJ: World Scientific.

Jou, D. (1996), *Extended Irreversible Thermodynamics*, New York: Springer.

Keizer, J. (1987), *Statistical Thermodynamics of Nonequilibrium Processes*, New York: Springer-Verlag.

Klimontovich, I. L. (1995), *Statistical Theory of Open Systems*, London: Kluwer.

Kocks, P., Ross, J., Björkman, O. (1995), "Thermodynamic Efficiency and Resonance of Photosynthesis in a C3 Plant," *J. Phys. Chem.* **99**, 16483–16489.

Kreuzer, H. J. (1981), *Nonequilibrium Thermodynamics and Its Statistical Foundations*, Oxford, U.K.: Clarendon Press.

Kuiken, G. D. C. (1994), *Thermodynamics of Irreversible Processes*, New York: Wiley.

Lavenda, B. H. (1978), *Thermodynamics of Irreversible Processes*, New York: Wiley.

Lavenda, B. H. (1985), *Nonequilibrium Statistical Thermodynamics*, New York: Wiley.

Manneville, P. (1990), *Dissipative Structures and Weak Turbulence*, San Diego: Academic.

Muschik, W. (1990), *Aspects of Non-Equilibrium Thermodynamics: Six Lectures on Fundamentals and Methods*, Singapore: World Scientific.

Nicolis, G. (1989), Physics Far from Equilibrium and Self-organization, *New Physics*, Cambridge, U.K.: Cambridge Univ. Press.

Nicolis, G., Dewel, G., Turner, J. W. (Eds.) (1981), *Order and Fluctuations in Equilibrium and Nonequilibrium Statistical Mechanics*, New York: Wiley.

Ross, J., Harding, R. H., Wolff, A. N., Xiolin, C. (1991), "The Relation of Fluxes and Forces to Work in Nonequilibrium Systems," *J. Chem. Phys.* **95**, 5206–5211.

Ross, J., Hunt, K. L. C., Hunt, P. M. (1992), "Thermodynamic and Stochastic Theory for Nonequilibrium Systems with Multiple Reactive Intermediates: The Concept and Role of Excess Work," *J. Chem. Phys.* **96**, 618–629.

Salamon, P., Sieniutycz, S. (Eds.) (1992), *Extended Thermodynamic Systems*, New York: Taylor and Francis.

Samohyl, I. (1987), *Thermodynamics of Irreversible Processes in Fluid Mixtures: Approached by Rational Thermodynamics*, Leipzig: B. G. Teubner.

Stratonovich, R. L. (1992), *Nonlinear Nonequilibrium Thermodynamics*, New York: Springer-Verlag.

Stratonovich, R. L. (1994), *Nonlinear Nonequilibrium Thermodynamics II: Advanced Theory*, New York: Springer.

Tisza, L. (1966), *Generalized Thermodynamics*, Cambridge, MA: MIT Press.

Tykodi, R. (1967), *Thermodynamics of Steady States*, New York: Macmillan Co.

Valanis, K. (1972), *Irreversible Thermodynamics of Continuous Media: Internal Variable Theory*, Wien, New York: Springer-Verlag.

Woods, L. C. (1985), *The Thermodynamics of Fluid Systems*, Oxford, U.K.: Clarendon Press.

Zubarev, D. N. (1974), *Nonequilibrium Statistical Thermodynamics*, New York: Consultants Bureau.

THERMOELECTRIC DEVICES

DONALD T. MORELLI, *Physics and Physical Chemistry Department, General Motors Research and Development Center, Warren, Michigan, U.S.A.*

	Introduction	339
1.	**Thermoelectric Properties of Materials**	340
1.1	Macroscopic Properties	340
1.1.1	Seebeck and Peltier Effects	340
1.1.2	Electrical Resistivity	341
1.1.3	Thermal Conductivity	341
1.2	Underlying Microscopic Properties	342
1.2.1	Energy Gap and Doping	342
1.2.2	Effective Mass and Effective Mobility	344
1.2.3	Phonon Conduction and Phonon Scattering	344
2.	**Device Operation and Design**	345
2.1	Efficiency of a Thermoelectric Generator	345
2.2	Coefficient of Performance of a Thermoelectric Cooler	347
2.3	Some General Design Considerations	348
2.3.1	Thermoelectric Material	348
2.3.2	Contact Metallization and Electrical Connections	348
2.3.3	Heat Exchangers	349
3.	**Examples of Thermoelectric Generators and Coolers**	349
3.1	Radioisotope Thermoelectric Generator for Interplanetary Spacecraft Propulsion	349
3.2	Bi-Te–Based Coolers for Electronic and Optical Applications	350
4.	**Potential New Materials for Thermoelectric Conversion Devices**	351
	Glossary	353
	Works Cited	353
	Further Reading	353

INTRODUCTION

In 1823 when Thomas Seebeck discovered the effect that now bears his name, most certainly he did not realize that, a century and a half later, an entire industry would develop from the seed of his simple experiments. But indeed, his observation of the occurrence of a potential difference in a circuit composed of two dissimilar metals with junctions at different temperatures turned out to be the basis for the thermoelectric generator that today provides on-board power for interplanetary space probes. And when barely a decade later Jean Peltier discovered the inverse effect of the cooling of a junction between two dissimilar metals upon the passage of electrical current, the foundation was formed for what is today a worldwide industry of thermoelectric cooling with applications in microelectronics, defense, and recreational industries. While the laws of thermodynamics deem that all energy ultimately evolves into heat and all systems move toward increasing entropy, the concepts of converting heat to electricity and using electricity to cool act like localized, temporal rebellions against time's arrow and continue to intrigue man today.

With increasing environmental concerns over the deleterious effects of waste heat and the harmful nature of compressed gases and chemical refrigerants, thermoelectric devices have received renewed attention in recent years. The idea of getting "free" electrical power from waste heat or obtaining cooling power from a solid-state device is at first glance very appealing from both environmental and economic points of view. However, many serious challenges remain to be met before large-scale thermoelectric refrigeration and power generation become a reality: Can new more efficient materials be developed? Can device design be improved and adapted to new applications? Are the

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

thermoelectric materials themselves environmentally "friendly"? And, perhaps most importantly, can a new generation of thermoelectric devices for energy conversion on a large scale be made economically?

While this review is not aimed at answering all of the above questions, we must always keep these challenges in mind when discussing both realized and potential applications of thermoelectric devices. Current developments in the field of thermoelectrics are unfolding along two main fronts:

1. incremental improvements in existing thermoelectric materials and device design, and
2. research into the development of new materials (or old materials in new geometries) with enhanced thermoelectric properties.

This review first gives a brief introduction to the physical properties of thermoelectric materials, then discusses the theory of device operation, provides several examples of thermoelectric coolers and generators using state-of-the-art materials, and finally concludes with some comments on new materials for thermoelectric refrigeration and power generation.

1. THERMOELECTRIC PROPERTIES OF MATERIALS

1.1 Macroscopic Properties

It is shown later that the ability of a material to act as a thermoelectric depends on three fundamental transport properties: the thermoelectric power S, the electrical conductivity σ, and the thermal conductivity κ. Quantitatively, we define the *thermoelectric figure of merit* Z:

$$Z = S^2\sigma/\kappa. \qquad (1)$$

The efficiency of thermoelectric devices is directly determined by Z and thus by the fundamental transport properties S, σ, and κ. The "electrical" part of this relationship, namely, the numerator $S^2\sigma$, is termed the power factor P of the material. Since the dimensions of Z are inverse temperature, it is convenient to define a dimensionless quantity ZT, the product of the figure of merit and the absolute temperature. We begin our discussion with a general overview of these three fundamental transport properties in various types of materials.

1.1.1 Seebeck and Peltier Effects The observations of Seebeck in 1823 and Peltier in 1834 were based on experiments in which two different metallic materials, A and B, were connected together in a circuit. The *Seebeck effect* (Fig. 1) is the occurrence of a potential difference V_{AB} in the circuit under conditions of zero net current flow when the ends of the two metal strips are at a temperature T_0 and the junction between them is at a different temperature $T_0 + \Delta T$. The magnitude of the voltage is proportional to the temperature difference ΔT and additionally depends on the type of materials of which the circuit is comprised. The *relative Seebeck coefficient* S_{AB} of the circuit shown in Fig. 1 is given by

$$S_{AB} = \lim_{\Delta T \to 0} \left(\frac{V_{AB}}{\Delta T}\right). \qquad (2)$$

While S_{AB} is a property of the couple A and B, it is possible to define absolute Seebeck coefficients S_A and S_B for each of the materials A and B. The relative Seebeck coefficient is then just the difference between the absolute coefficients:

$$S_{AB} = S_A - S_B. \qquad (3)$$

The Seebeck effect is the means by which

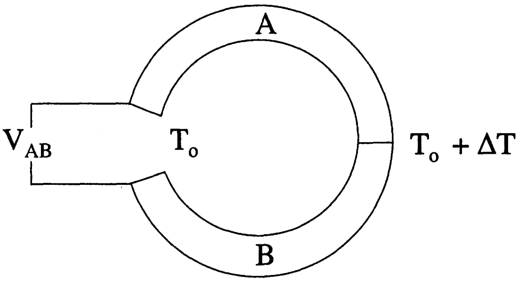

FIG. 1. Thermoelectric circuit composed by joining two dissimilar metals A and B. When the ends of the metal strips are at a temperature T_0 and the junction is at a different temperature $T_0 + \Delta T$, a thermoelectric voltage V_{AB} is developed under open-circuit conditions.

thermoelectric generators convert heat to electricity. Clearly, the larger the Seebeck coefficient, the larger the voltage that can be generated per temperature interval. The Seebeck coefficient, or thermoelectric power, can exhibit a wide range of values. In superconductors, S is rigorously zero, while in good metals it is only a few microvolts per kelvin at room temperature. Doped semiconductors, on the other hand, can have Seebeck coefficients of several hundred microvolts per kelvin.

In the *Peltier effect,* when the two dissimilar metals A and B are joined in a closed circuit and a current I is allowed to flow (Fig. 2), an amount of heat Q is absorbed per unit time at the junction between the two strips and evolved at their ends. The *relative Peltier coefficient* π_{AB} of the circuit in Fig. 2 is defined as the ratio of heat absorbed (or evolved) to the current:

$$\pi_{AB} = Q/I. \tag{4}$$

It can be shown using thermodynamic arguments that the Peltier coefficient and Seebeck coefficient are related to one another by the Kelvin relation

$$S_{AB} = \pi_{AB}/T. \tag{5}$$

As in the case of the Seebeck coefficient, it is possible to define absolute Peltier coefficients π_A and π_B for each arm of the thermoelectric circuit of Fig. 2, such that

$$\pi_{AB} = \pi_A - \pi_B. \tag{6}$$

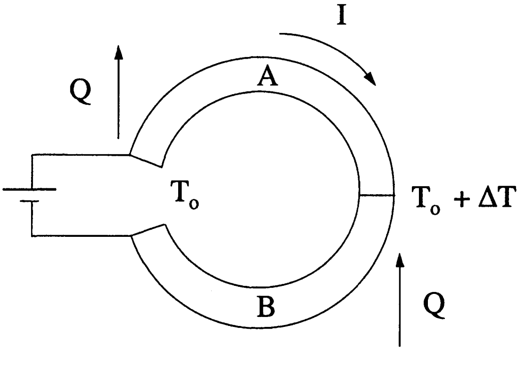

FIG. 2. Thermoelectric circuit consisting of two dissimilar metals A and B joined together. When a current is passed through the circuit, an amount of heat Q is absorbed at the junction between the metal strips and liberated at their ends.

1.1.2 Electrical Resistivity While the Seebeck coefficient of certain materials, such as wide-band-gap semiconductors and polymers, can become very large, these systems are not good thermoelectric materials since they also have low electrical conductivity. The electrical conductivity of a material is related to the number density of charge carriers in the system, designated by n, and their mobility, μ, according to

$$\sigma = ne\mu. \tag{7}$$

In good metals, the electrical conductivity is large because of large n; however, S and hence P in these systems are usually very small. In undoped wide-band-gap semiconductors, the electrical conductivity is small because of small n, and P can never be large. For thermoelectric applications, the intermediate case is of interest, namely, moderately doped semiconductors. Such systems have carrier densities smaller than in metals but can have large mobilities, so that the electrical conductivity can be reasonably large. At the same time, since the carrier density is not large, the thermoelectric power also can be fairly substantial. The precise conditions under which one obtains a large power factor are determined by the microscopic physical properties of the system such as energy band structure, carrier effective mass, and carrier mobility and are discussed in detail in the next section. However, we can already note from our discussion of the macroscopic properties of materials that large power factors will be found among the narrow- to intermediate-band-gap semiconductors.

1.1.3 Thermal Conductivity The thermal conductivity of a solid material is made up of two parts: the thermal conductivity of the charge-carrier system, κ_c, and the thermal conductivity of the lattice, or phonon system, κ_p, the total thermal conductivity being the sum of the two components. Evidently for thermoelectric materials we want to have κ as small as possible. In metallic systems, κ_c is the dominant contribution to the total thermal conductivity. The Wiedemann–Franz law (see METALS AND ALLOYS, CONDUCTIVITY IN) provides a relationship between the carrier electrical conductivity and the carrier thermal conductivity:

$$\kappa_c = LT\sigma, \tag{8}$$

where L is the Lorenz number. When the scattering of charge carriers is elastic, the Lorenz number takes on the Sommerfeld value $L_0 = 2.45 \times 10^{-8}\ \mathrm{V}^2\ \mathrm{K}^{-2}$. If the scattering is partially inelastic, the Lorenz number is depressed below L_0 and the Wiedemann–Franz law (using the value of L_0) provides an upper limit on the charge-carrier thermal conductivity. For most metals, $\kappa_c \gg \kappa_p$ and, by use of the Wiedemann–Franz law, the figure of merit reduces to

$$ZT = \frac{S^2\sigma T}{\kappa_c + \kappa_p} \approx \frac{S^2}{L_0}. \tag{9}$$

Thus to obtain $ZT = 1$ (the approximate value of state-of-the-art materials), one requires $S \approx 157\ \mu\mathrm{V}\ \mathrm{K}^{-1}$. Such a high thermopower has never been obtained on a metallic system, although some alloys made from *d*- and *f*-band metals do exhibit thermopowers of several tens of microvolts per kelvin. So once again we turn our attention to semiconducting systems for thermoelectric device applications. In this case the phonon contribution to the total thermal conductivity can be substantial and in many circumstances is the dominant mode of heat transport in these solids. The phonon thermal conductivity can span a surprisingly large range of values, from on the order of $10^{-2}\ \mathrm{W\ cm^{-1}\ K^{-1}}$ for mixed crystals to as high as $25\ \mathrm{W\ cm^{-1}\ K^{-1}}$ for diamond. The lattice thermal conductivity depends inversely on both the atom masses and the primitive-unit-cell complexity; diamond has a very high thermal conductivity because carbon is light and the unit cell contains only one atom. In order to obtain the low intrinsic thermal conductivity necessary for a thermoelectric material, one needs to work with compounds made up of heavy elements. For semiconductors this traditionally has meant working with elements in the lower right corner of the periodic table, such as Bi, Pb, Te, and Se. Thus some of the best thermoelectrics are Bi_2Te_3 and PbTe.

Of all the physical properties that determine the figure of merit, the thermal conductivity is the one most amenable to manipulation in the laboratory. A great deal of energy has been spent on reducing the thermal conductivity of thermoelectric materials by virtually every means possible—for example, using mixed crystals, substitutional point defects, radiation-induced damage, and large-scale inclusions, to name just a few techniques. When this methodology is combined with the restriction that the carrier mobility in a thermoelectric should be as large as possible, the engineering of thermoelectric materials can be summarized in a single sentence: How can I impede the phonons as much as possible while not upsetting charge-carrier transport? While at first it might seem that this is an impossible endeavor, an important difference in phonon and electron spectra does in fact make it possible to achieve such a condition: and that is that while the wavelength of a typical heat-carrying phonon is only a few lattice spacings, the deBroglie wavelength of a charge carrier is many times larger. Thus it is possible to "engineer" structures in which defects interact strongly with phonons but essentially do not perturb the Bloch waves.

This completes our discussion of the macroscopic properties of thermoelectric materials. We have seen that a good thermoelectric material will have a large Seebeck coefficient, large electrical conductivity, and small thermal conductivity, and we have identified semiconducting compounds of heavy constituent-atom masses as the most likely materials for thermoelectrics. In the next section we look in detail at the microscopic properties of such materials with the object of understanding at a fundamental level how these properties can be achieved.

1.2 Underlying Microscopic Properties

The relevant transport properties of thermoelectric materials, namely, the Seebeck coefficient and electrical and thermal conductivities, depend on a variety of microscopic parameters of the system, including the electronic structure and electron- and phonon-scattering processes. We begin our discussion by introducing the concepts of energy bands, the energy gap, and doping.

1.2.1 Energy Gap and Doping A semiconductor is a material that possesses a forbidden energy gap between the valence band

and the conduction band; see Fig. 3. In an *intrinsic* semiconductor, the topmost valence band is completely filled with electrons and the lowest-lying conduction band, lying above the top of the valence band by an energy E_g, is empty [Fig. 3(a)]. Thermal excitation across the gap produces electrons in the conduction band, leaving behind an equal number of empty states, or holes, in the valence band. The holes act as current carriers of positive charge $+e$. In an extrinsic semiconductor, impurity atoms (called *donors*) produce electrons in an energy level close to the bottom of the conduction band [Fig. 3(b)]. Because in this situation the energy difference ΔE between the donor level and the conduction band is relatively small, these electrons can be excited into the conduction band at fairly low temperatures, and an *n-type* semiconductor is obtained. Analogously, impurity atoms of another class, called *acceptors*, produce energy levels in the forbidden zone of the semiconductor but near the top of the valence band [Fig. 3(c)]. Here electrons can be promoted into states in the impurity level, thereby leaving mobile holes in the valence band that are then capable of carrying current; in this situation a *p-type* semiconductor is obtained.

In an intrinsic material, conduction proceeds via both electrons and holes (see INSULATORS AND SEMICONDUCTORS, CONDUCTIVITY IN). The total electrical conductivity in such a circumstance is the sum of the partial conductivities of electrons and holes:

$$\sigma = \sigma_e + \sigma_h = ne\mu + pe\nu. \tag{10}$$

Here n and μ are the electron concentration and mobility and p and ν are the hole concentration and mobility, respectively. The total Seebeck coefficient of an intrinsic material is given by

$$S = \frac{S_e\sigma_e + S_h\sigma_h}{\sigma_e + \sigma_h}. \tag{11}$$

Because S_e is negative and S_h is positive, the total Seebeck coefficient is smaller in magnitude than either of the partial Seebeck coefficients. For this reason, it is not desirable to use intrinsic semiconductors for thermoelectric applications. Instead it is preferable to dope the material n or p type in order to take full advantage of the thermoelectric properties of a single carrier type. Even in such an extrinsic semiconductor, intrinsic conduction will take place at a high enough temperature, and the thermoelectric properties will be degraded. As a rule of thumb,

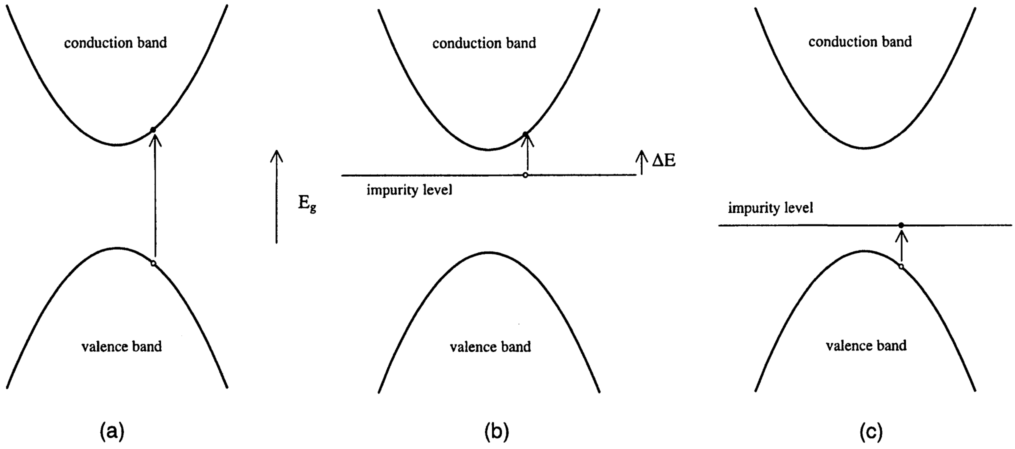

FIG. 3. (a) Energy-band structure of an intrinsic semiconductor. The empty conduction band is at an energy E_g above the full valence band. Thermal excitation of an electron across the gap can occur, leaving an empty state, or hole, in the valence band. (b) Band structure of a semiconductor doped *n* type. An impurity level is situated at an energy $\Delta E < E_g$ below the bottom of the conduction band. Electrons from this level can be easily promoted into the conduction band. (c) Band structure of a semiconductor doped *p* type. An impurity level is situated at an energy $\Delta E < E_g$ above the top of the valence band. Electrons from the valence band can be promoted easily to this level, creating a mobile hole.

one usually chooses a material such that the intrinsic energy gap is greater than a few times $k_B T$ in order to avoid intrinsic conduction at the operating temperature T.

The extent to which energy levels within the conduction band (or valence band in a p-type semiconductor) are occupied is determined by the Fermi level E_F. The probability that a state is occupied is determined by the difference between that state's energy and E_F. When E_F lies deep within the gap, the material is termed nondegenerate, and Boltzmann statistics apply. When E_F lies high up in the conduction band, the material is called degenerate, and the properties are similar to those of a metal. For thermoelectric applications, enough carriers are required to maintain good electrical conductivity, and we are interested in the intermediate case of partial degeneracy, in which case E_F is just below, at, or slightly above the band edge. In this case it is necessary to use Fermi–Dirac statistics in order to determine the probability of occupation of a state.

The Fermi level of an extrinsic semiconductor is adjusted by the degree of doping. In the partial or complete degeneracy cases, the carrier concentration is given by

$$n = \frac{\sqrt{2}}{\pi^2 \hbar^3}(m^* k_B T)^{3/2} F_{1/2} \tag{12}$$

where $\hbar = h/2\pi$ (with h = Planck's constant), m^* is the carrier *effective mass,* and

$$F_i = \int_0^\infty \frac{x^i dx}{1 + \exp(x - \mu^*)}, \tag{13}$$

where $\mu^* = E_F/k_B T$ is the reduced Fermi potential. The Seebeck coefficient, on the other hand, is given by

$$S = \frac{\pm k}{e}\left[\frac{(\frac{5}{2} + \lambda)F_{3/2+\lambda}}{(\frac{3}{2} + \lambda)F_{1/2+\lambda}} - \mu^*\right], \tag{14}$$

where λ is determined by the energy dependence of the electron-scattering time τ:

$$\tau \propto E^\lambda. \tag{15}$$

For scattering of electrons by acoustic phonons, $\lambda = -\frac{1}{2}$, whereas for scattering of electrons by ionized impurities, $\lambda = \frac{3}{2}$. In many cases, it is not possible to assign a unique value to λ.

1.2.2 Effective Mass and Effective Mobility As can be seen from Eq. (14), the Seebeck coefficient for a given type of electron scattering is determined only by the Fermi level and increases in magnitude with decreasing μ^*. On the other hand, n (and thus the conductivity) decreases with decreasing μ^* but also depends on the electron effective mass m^*. In order to maintain high electrical conductivity a relatively high carrier mobility is required. Clearly a large power factor $S^2\sigma$ is obtained by optimization of the carrier concentration by adjusting E_F by doping. It has been shown (Slack, 1995) that the power factor is largest for materials having a large *effective mobility* U*:*

$$U = \mu(m^*/m_0)^{3/2}. \tag{16}$$

Thus we seek semiconductors with both large mobility and large effective mass. These two criteria are to some extent mutually exclusive, since a large electron mass normally implies low mobility, and vice versa. However, U does in fact reach a maximum for certain types of materials, and these indeed have been found to be good thermoelectrics.

1.2.3 Phonon Conduction and Phonon Scattering As we have seen in Sec. 1.1.3, for most thermoelectric materials the main component of heat conduction is via lattice vibrations, or phonons, and we seek to reduce the lattice thermal conductivity as much as possible in order to obtain a large figure of merit and corresponding large device efficiency. It is desirable to begin with compounds whose intrinsic thermal conductivity (that is, the thermal conductivity determined by phonon–phonon scattering) is already quite low. Slack (1979) has shown that the intrinsic thermal conductivity of pure crystals depends on the average atomic mass M, Debye temperature θ, Grüneisen parameter γ, and number of atoms in the primitive unit cell N. Thus PbTe ($M = 167.4$, $\theta = 127$ K, $N = 2$) has a thermal conductivity at the Debye temperature of $\kappa_\theta = 0.068$ W cm^{-1} K^{-1}, while diamond ($M = 12$, $\theta = 1800$ K, $N = 1$) has $\kappa_\theta \approx 2$ W cm^{-1} K^{-1}. This simple criterion can thus be used to judge which materials might possess a low thermal conductivity.

Even though materials such as PbTe and Bi_2Te_3 have low intrinsic thermal conductivity, it is necessary to reduce their thermal conductivity still further in order to enhance Z to the point that these compounds and others can be technologically useful thermoelectrics. This is done by introducing defects or some type of disorder into the crystals. The subject of phonon–defect scattering and the resultant reduction in thermal conductivity is a very rich one of its own accord (see PHONONS IN CRYSTAL LATTICES), and several excellent reviews and monographs on the subject are available; see, for instance, Klemens (1958) and Berman (1976). We simply note here that the thermoelectrician has at his or her disposal several experimental methods of reducing the lattice thermal conductivity without substantially altering the power factor of a given thermoelectric material. We will illustrate this important technique with a couple of examples.

Probably the most popular and effective method of reducing the thermal conductivity of a compound is to make a mixed crystal. In this technique one of the constituent atoms of the compound is replaced by an element from the same column of the periodic table; as an example, one can start with PbTe and partially substitute Se for Te to make the mixed crystal $PbTe_{1-x}Se_x$. Simple valence-electron arguments indicate that in many cases the resulting electronic structure of the mixed crystal will not differ greatly from that of the pure compound, and thus the electrical properties will be quite similar. On the other hand, the resulting mass difference is capable of producing strong phonon scattering, and the thermal conductivity can be reduced over a fairly large temperature range.

A second example of an effective means of reducing the thermal conductivity is given in the Si/20% Ge mixed crystal system. This thermoelectric material was developed for a high-temperature generator application. In this system scattering of high-frequency phonons is achieved by the mass fluctuation scattering of the Ge impurity atoms, and low-frequency phonons are scattered by grain boundaries and electrons. However, a large range of intermediate-frequency phonons that carry a substantial amount of heat at high temperature are relatively unaffected by these scattering mechanisms. In order to decrease the mean free path of this part of the phonon spectrum, inert particulates of approximate size of 50 Å can be placed inside of the SiGe grains. Experiments on hot pressed samples containing BN, B_4C, and Y_2O_3 scattering centers indeed showed a remarkable reduction in thermal conductivity by as much as 50% relative to standard SiGe alloys. The theoretical predictions and subsequent experiments on this system showed the strong predictive power of lattice thermal conductivity models based on frequency-dependent scattering rates.

This ends our discussion of the macroscopic and microscopic properties of thermoelectric materials. We have seen how the desired properties of high Seebeck coefficient, large electrical conductivity, and small thermal conductivity are related to underlying physical properties such as Fermi level, effective mass, effective mobility, atomic masses, unit-cell complexity, and phonon-scattering mechanisms. On this basis we have identified the most likely high-Z thermoelectric materials to be high-mobility, large-effective-mass semiconductors with heavy constituent atom masses and large unit cells. Now that an understanding of thermoelectric materials is in hand, we look briefly at the theory of thermoelectric device operation.

2. DEVICE OPERATION AND DESIGN

2.1 Efficiency of a Thermoelectric Generator

Consider the thermocouple consisting of one n-type element and one p-type element, both of length L and cross-sectional area A, as shown in Fig. 4. The two elements are connected together at one end by a metal strip, and across the other end is placed a load of resistance R_L. The hot end of this couple is at temperature T_H and the cold end at temperature T_C. For simplicity we assume that the thermoelectric properties of the thermoelements are identical so that $S_n = S_p = S$, $\kappa_n = \kappa_p = \kappa$, and $\rho_n = \rho_p = \rho$, where S, κ, and ρ are the Seebeck coefficient, thermal conductivity, and electrical resistivity of the elements. The general case of differing

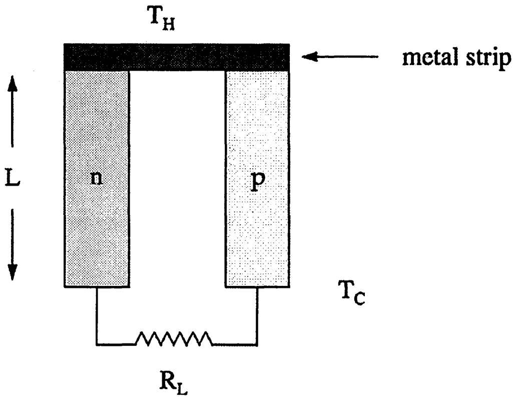

FIG. 4. Thermoelectric generator consisting of one *n*-type element and one *p*-type element, each of length *L* and cross-sectional area *A* (not shown), joined by a metal strip. With a temperature difference $T_H - T_C$ across the couple, a current will flow in the load resistance R_L.

thermoelectric properties for the two legs is more detailed than we wish to include here but may be found in Kerr (1960). Let R and K equal the combined electrical resistance and thermal conductance of the couple, respectively:

$$R = 2\rho L/A \quad \text{and} \quad K = 2\kappa A/L. \tag{17}$$

The relative Seebeck coefficient of the couple is $2S$.

The arrangement shown in Fig. 4 is a thermoelectric generator because the temperature difference $T_H - T_C = \Delta T$ across the thermoelements causes a current I to flow in the load resistance R_L. The efficiency of this generator is defined as the ratio of electrical power P delivered to the load to the energy consumed from the heat source, Q_s. Consider now the conservation of energy at the hot junction. The heat delivered by the source Q_s is balanced by the heat conducted down the arms of the thermocouple Q_c, and the Peltier heat developed at the hot junction Q_p due to the flow of current. In addition, of the Joule heat I^2R developed in the arms of the thermocouple, one-half is returned to the hot junction. Thus

$$Q_s = Q_p + Q_c - \tfrac{1}{2} I^2 R. \tag{18}$$

Now $Q_p = 2SIT_H$ and $Q_c = K(T_H - T_C)$. In addition, the current in the circuit is the thermoelectric voltage divided by the total circuit resistance:

$$I = \frac{2S(T_H - T_C)}{R_L + R}. \tag{19}$$

The generator efficiency ϵ can thus be written

$$\epsilon = \frac{I^2 R_L}{Q_p + Q_c - I^2 R}, \tag{20}$$

$$\epsilon = \frac{T_H - T_C}{T_H} \times \frac{m/(m+1)}{1 + KR(m+1)/4S^2 T_H - (T_H - T_C)/2T_H(m+1)}, \tag{21}$$

where $m = R_L/R$.

The maximum power delivered to the load occurs when the load resistance is equal to the generator internal resistance, i.e., when $m = 1$. Under this condition the efficiency is

$$\epsilon = \frac{\Delta T}{T_H} \frac{1}{2 + 4/ZT_H - \Delta T/2T_H}, \tag{22}$$

where $Z = 4S^2/KR = S^2/\kappa\rho$ is the figure of merit of one of the arms of the thermocouple. Figure 5 shows the efficiency under maximum power output conditions as a function of the figure of merit and temperature difference ΔT.

It is interesting to note that the efficiency under maximum power transfer to the load is not the maximum efficiency $\epsilon_{\max}$. This latter quantity can be determined by differentiating ϵ with respect to R_L (i.e., m) and setting the result equal to zero. This condition yields the value for the load resistance at maximum efficiency:

$$(R_L)_{\max} = R\sqrt{1 + ZT_{\text{ave}}}, \tag{23}$$

where $T_{\text{ave}} = (T_H + T_C)/2$. Thus the load resistance required for maximum generator efficiency is always greater than that required for maximum power output. Substituting the value of R_L/R from Eq. (23) into Eq. (21) yields an expression for the maximum generator efficiency:

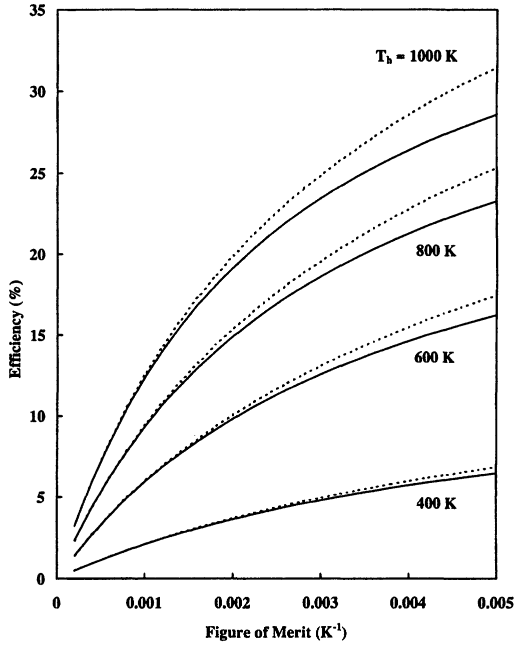

FIG. 5. Efficiency of a thermoelectric generator as a function of the figure of merit for several hot-side temperatures, assuming a cold-side temperature of 300 K. Solid lines: under conditions of maximum power transfer; dashed lines: under conditions of maximum efficiency.

$$\epsilon = \frac{\Delta T}{T_H} \frac{\sqrt{1 + ZT_{\text{ave}}} - 1}{\sqrt{1 + ZT_{\text{ave}}} + T_C/T_H}. \tag{24}$$

Thus we see that the efficiency of a thermoelectric generator cannot exceed the Carnot efficiency given by $\Delta T/T_H$. The generator efficiency increases with increasing values of Z and for larger Carnot efficiency. Figure 5 shows the maximum efficiency as a function of both Z and the temperature difference ΔT. For state-of-the-art thermoelectrics today ϵ does not exceed 10%. Recent theoretical models (Slack, 1995) have suggested that actual Z values may be augmented by at most a factor of 2–4. With such materials the thermoelectric efficiency could be increased to more than 20%.

2.2 Coefficient of Performance of a Thermoelectric Cooler

For the case of thermoelectric cooling we consider the circuit shown in Fig. 6. Here a dc power supply is connected in the circuit at T_H while the thermocouple legs are joined at the source (at temperature T_C) from which heat is to be removed. Considering conservation of energy at the source, an amount of heat Q_p is removed from the source by Peltier cooling at the junction, but an amount of heat Q_c flows into the source due to conduction down the arms of the thermocouple. In addition, of the Joule heat I^2R developed in the arms of the thermocouple, one-half is returned to the cold junction. Thus the total rate of heat removal Q_r from the cold junction is given by

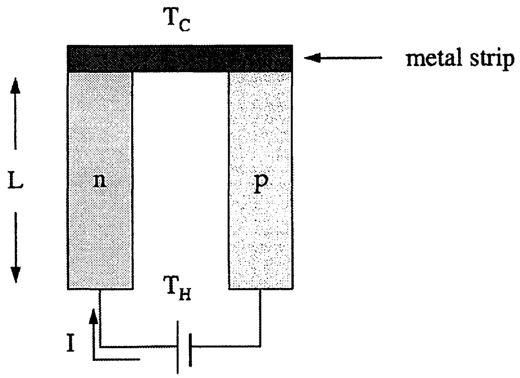

FIG. 6. Thermoelectric cooler consisting of one n-type element and one p-type element, each of length L and cross-sectional area A (not shown), joined by a metal strip. With current flowing in the circuit, heat is liberated at the hot junction and absorbed at the cold junction, producing cooling.

$$Q_r = Q_p - Q_c - \tfrac{1}{2} I^2 R. \tag{25}$$

Under the same notation and assumptions as in the case of the thermoelectric generator, the rate of heat conduction down through the thermocouple legs to the source is $Q_c = K(T_H - T_C)$, and the rate of Peltier cooling at the cold junction is $Q_p = 2SIT_C$. Thus

$$Q_r = 2SIT_C - K(T_H - T_C) - \tfrac{1}{2} I^2 R, \tag{26}$$

and the temperature difference achieved is

$$T_H - T_C = \Delta T = 2SI - \tfrac{1}{2} I^2 R - Q_r. \tag{27}$$

The ability of a refrigerator to cool is given by the coefficient of performance (COP), which is the ratio of the rate of heat removal Q_r to the amount of power consumed, P:

$$\mathrm{COP} = Q_r/P. \tag{28}$$

The power consumed is the sum of the Joule heat generated in the arms of the thermocouple and the power required to overcome the thermal EMF:

$$P = I^2R + 2IS\Delta T. \tag{29}$$

Substitution of (25) and (28) into (27) yields

$$\mathrm{COP} = \frac{2SIT_c - K(T_H - T_C) - \frac{1}{2}I^2R}{I^2R + 2IS\Delta T}. \tag{30}$$

Thus both the temperature difference attainable (27) and the COP (30) depend on the current in the circuit. The maximum temperature difference occurs when $Q_r = 0$ and for a current given by the condition $d(\Delta T)/dI = 0$, which yields

$$I(\max \Delta T) = 2ST_c/R. \tag{31}$$

The corresponding maximum temperature difference is

$$\Delta T_{\max} = \tfrac{1}{2}ZT_C^2. \tag{32}$$

The maximum COP, on the other hand, occurs for a current given by the condition $d(\mathrm{COP})/DI = 0$, which gives

$$I(\max \mathrm{COP}) = \frac{2S\Delta T}{[\sqrt{1 + \frac{1}{2}Z(T_H + T_C)} - 1]R}. \tag{33}$$

The corresponding maximum COP is

$$(\mathrm{COP})_{\max} = \frac{T_C}{\Delta T}\,\frac{\sqrt{1 + ZT_{\mathrm{ave}}} - T_H/T_C}{\sqrt{1 + ZT_{\mathrm{ave}}} + 1}. \tag{34}$$

2.3 Some General Design Considerations

There are many different designs of thermoelectric devices for specific applications of this technology. However, all of these designs share several common features that are required for operation, including *n*- and *p*-type thermoelectric elements; contact metallization; electrical connections to supply current in the case of a cooling device or to draw off current in the case of a generator; and heat exchangers on both the hot and cold sides to maintain the maximum possible temperature difference. The overall efficiency can depend strongly on these factors, and several key points should be kept in mind when preparing to implement them in a device.

2.3.1 Thermoelectric Material Typically the figure of merit of a particular thermoelectric material is a rather strong function of temperature, and one wishes to choose a material the figure of merit of which is optimized for the temperature range of the particular application and which is chemically stable at the operating temperature. For thermoelectric coolers operating near room temperature and for thermoelectric generators operating below 300 °C, the material of choice is Bi_2Te_3. Above about 200 °C the figure of merit of Bi_2Te_3-based materials begins to decline as intrinsic conduction becomes important, and for temperatures between about 400 and 700 °C PbTe-based compounds have the highest figure of merit. For temperatures above 700 °C the best materials today are SiGe compounds.

2.3.2 Contact Metallization and Electrical Connections A critical issue in the successful implementation of thermoelectric technology is the ability to make low-resistance electrical contact to the thermoelectric elements. Because of the relatively low resistivity of thermoelectric materials (a necessity for their operation in order that Joule heating effects be negligible), it is essential that the electrical contacts to the elements have low resistance as well. Sizable contact resistances can act against the Peltier cooling in a thermoelectric refrigerator and can limit the power delivered to the load in a thermoelectric generator, and every effort should be made to minimize these effects. What must be kept in mind is that the thermoelectric elements are semiconductors, and making good low-resistance contacts is a nontrivial problem involving a large amount of both solid-state physics and engineering and has been the subject of a good deal of literature in its own right (e.g., Williams, 1984). For our purposes we simply point out that a contact between a semiconductor and a metal may be rectifying rather than Ohmic. Rectifying contacts can become highly resistive depending on conditions such as current magnitude and direction, and temperature, and thus must be avoided. In addition to the nature of the contact itself, two additional contact problems can arise, particularly in the case of thermoelectric devices operating at high temperature; these are *diffusion* and *thermal expansion mismatch*. Certain metals can have very high rates of diffusion into the bulk of the semiconductor elements, resulting in not only degradation of the contact but also an undesirable alteration of the properties of the thermoelectric material itself. A classic example is the influence of copper on bismuth telluride; plating copper onto *p*-type bismuth telluride even at room temperature can transform the material into an *n*-type semiconductor (Goldsmid, 1960). The problem of thermal expansion mismatch is also a delicate one, as a metallic contact with an expansion coefficient greatly different from that of the thermoelement may crack, peel, or delaminate under temperature excursions. Taken together, the problems of nature of the contact, diffusion, and thermal expansion can present a difficult challenge depending on the thermoelectric material and the temperature range of interest. Frequently one finds that, for instance, a material that makes a good Ohmic contact also has a high diffusion coefficient, or a metal that has a good thermal expansion match makes a rectifying contact. These situations can be dealt with by using multiple layers of metals to provide a diffusion barrier, thermal match, and Ohmic contact at the same time.

2.3.3 Heat Exchangers The purpose of heat exchangers is to bring the temperature of the hot side of the thermoelectric device as close as possible to that of the heat sink (or source) and the temperature of the cold side of the thermoelectric device as close as possible to the cold sink. This means that the heat exchanger should have a large surface area and be made of a material with high thermal conductivity. Because of its low density and high thermal conductivity, aluminum is usually chosen as the exchanger material, and in order to maximize surface area for a given volume, some type of fin arrangement is employed. A related issue is the requirement that the heat exchanger obviously must be in good thermal contact with, but electrically insulated from, the thermoelectric elements. A popular method of achieving this is to use strips of mica or alumina in between the exchangers and the elements, but the added thermal resistance of such a layer will produce a non-negligible temperature difference between the element and the heat exchanger. With the wide variety of technologies available today (such as sputtering, vapor deposition, and thermal spray, to name a few) for producing thin layers of insulating films, it should be possible to provide high-quality electrically insulating, thermally conducting layers that minimize the temperature differences between the elements and exchangers.

3. EXAMPLES OF THERMOELECTRIC GENERATORS AND COOLERS

3.1 Radioisotope Thermoelectric Generator for Interplanetary Spacecraft Propulsion

Probably the most well-known application of thermoelectric power generation is the use of a radioisotope thermoelectric generator (RTG) to provide electrical power for interplanetary spacecraft such as Voyager, Galileo, and Ulysses. These generators were developed by the Jet Propulsion Laboratory over the last quarter-century. Of primary concern was the spacecraft missions' requirements of long life and high reliability, rather than generator efficiency or overall system cost, thus making thermoelectrics the technology of choice for this application.

The source of thermal energy for the RTG is the heat generated during the radioactive decay of a plutonium-238 heat source. In order to produce enough electrical power for the spacecraft (about 250 W per RTG), 4300 W of thermal power is required. The active thermoelectric elements are Si/20% Ge alloys doped *p* type and *n* type with boron and phosphorus, respectively. Degradation of the electrical power output as a function of time for the RTG is due principally to the decay of the radioactive source and increase in thermoelement resistivity due to dopant diffusion. The output electrical power for these generators is typically 90, 80, and 75% of the original output power after 4, 10, and 16 years, respectively. Because of the high operating temperature of the RTG (hot-side temperatures in excess of 1000 °C), several materials issues related to diffusion, sublimation, oxidation, and metallization had to be solved. The successful implementation of these power sources in long-term interplanetary space missions represents the greatest testament to the reliability and long life of thermoelectric devices.

3.2 Bi-Te–Based Coolers for Electronic and Optical Applications

Compared to the extreme conditions to which an RTG is exposed in outer space, the operational conditions of thermoelectric coolers are considerably less stringent. To begin with, the temperature range of operation (typically −50 to 100 °C) minimizes and eliminates several difficulties. Problems associated with thermal-expansion mismatch, contact resistance, and dopant diffusion are far less severe. That is not to say that these issues can be neglected. In the operation of a cooler, the heat rejected from the device must be removed using a heat sink. If this is not done, the hot side of the cooler can overheat, and some or all of the problems listed above can become an issue. Single-stage coolers are commercially available with cooling powers of up to almost 70 W and maximum temperature differences of nearly 70 °C. A typical thermoelectric cooler is depicted in Fig. 7. Several hundred p- and n-type elements are arranged in an electrical series connection and placed between two ceramic plates. The thermoelements are thus connected in parallel thermally; one of the plates is the hot side, and the other, the cold side. The electrical connection between neighboring elements is a small copper strip soldered to each element. The ceramic plates serve the purpose of providing a structurally strong member for mounting the device to a heat sink on the hot side and the part to be cooled on the cold side, while at the same time providing electrical insulation between the elements and these entities. It is desirable for the ceramic to have a low thermal resistance in order to minimize any temperature difference across it. Thin sheets of adamantine compounds such as Al_2O_3 and BeO are the materials of choice. If the surface area of the cooler is not too large, the outer surface of the ceramics can be metallized so that they can be soldered to the heat and cold sinks. For large surface areas, thermal-expansion mismatch can be a problem, and it is usually desirable to make thermal contact by means of a thermally conducting grease or compression. A very nice feature of thermoelectric coolers is that the temperature difference obtained and the cooling power are functions of the device current, thus allowing precise temperature control using a feedback loop.

Thermoelectric devices have been used to great advantage for cooling solid-state lasers and infrared detectors, to name just two applications. In addition, they have been employed in small-scale volume-cooling applications such as small refrigerators, picnic baskets, and electronic-equipment enclosures. They are not currently used in applications requiring cooling powers in excess of a couple of hundred watts, because the coefficient of performance of present-day thermoelectric coolers is smaller than that of standard vapor-compression air conditioning units in this cooling-power range. However, with the development of new materials of enhanced figure of merit, the COP of thermoelectrics will become comparable to that of standard units even for higher–cooling-

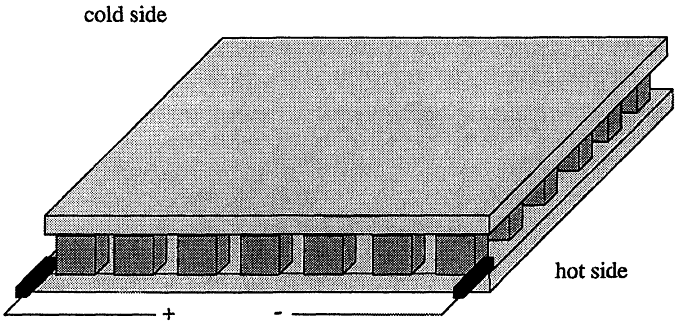

FIG. 7. Typical thermoelectric cooler. Several hundred p- and n-type elements are joined together in an electrical series, thermal parallel arrangement and sandwiched between high–thermal-conductivity electrical insulators.

power applications such as room or vehicle air conditioning. Assuming that such materials are found and the COP of TE coolers becomes competitive, whether such large-scale application occurs will depend on many highly variable factors. Certainly the all-solid-state operation, high reliability, long life, low noise, and low vibration levels of TE coolers are strong advantages. But several questions remain: Can device design be improved and adapted to new applications? Is there enough raw material available to satisfy the needs of the industry? Are the thermoelectric materials themselves environmentally friendly, easy to manufacture, and recyclable? Can thermoelectric cooling on a large scale really be done economically? These are serious issues that will have to be addressed somewhere down the road. But the first step and the real key to opening up large markets for thermoelectric devices is the *development of new materials with enhanced properties.* An area of intense research in the early 1960s, thermoelectric materials research has been done on a very small scale over the last 30 years. However, our knowledge today of the minimum thermal conductivity achievable in materials together with the development of new materials with unusual and exotic transport properties has renewed hope that thermoelectric materials with figures of merit substantially larger than that of Bi_2Te_3 can exist. Within the last two years, activity on both the theoretical and experimental fronts has increased dramatically. This review concludes with a short description of some potentially interesting materials systems that are just beginning to give us the hint that better and more efficient thermoelectric materials may be just around the corner.

4. POTENTIAL NEW MATERIALS FOR THERMOELECTRIC CONVERSION DEVICES

In the past 50 years, hundreds of compounds have been investigated with the hopes of finding a material with thermoelectric properties significantly better than Bi_2Te_3, with decidedly negative results. Thus one may rightly ask the question, If we have failed at this endeavor in the past, why should we hope to succeed now? There are several reasons why there is hope today that better thermoelectrics might someday be developed. In the last 20 years we have witnessed explosive growth in the field of materials research. New material systems, such as the two-dimensional electron gas and heavy-fermion, intermediate-valence, hybridization-gap, and correlated-electron systems, have been discovered. Innovative ways of making materials, including molecular-beam epitaxy and sputtering, have been developed. And, not least of all, a deeper understanding of the physical processes underlying the observed properties of materials has been developed. One potential route to high Z is to take advantage of unusual structures, geometries, or effects in order to produce outstanding thermoelectric properties. Some examples would include using quantum confinement to alter in a positive way the density of states of traditional thermoelectric materials such as PbTe and BiSb; fabricating unusual "rattling atom" structures such as filled skutterudites or clathrate compounds to produce solids with good electrical properties but very low thermal conductivity; taking advantage of strong phonon-scattering processes such as $4f$-electron scattering in rare earth-compounds or charge-transfer type scattering in mixed-valence systems. These are some of the several areas of research that have become very active in the last two years. Since the purpose of this article is to review the physics and technology of thermoelectric *devices*, it is not our desire to give an exhaustive review of thermoelectric *materials*. On the other hand, we would be remiss to not include a brief discussion of some potential new thermoelectric materials, since the development of such materials can potentially revolutionize the thermoelectric device industry.

In order to give the reader a flavor for the type of research currently being carried out on advanced thermoelectric materials, we will highlight a couple of specific examples. The first is the so-called skutterudite compounds that have emerged as a fertile area to search for high-Z thermoelectrics. These little-known compounds, apparently semiconductors, are of the form AB_3 where $A = Ir$, Rh, or Co and $B = Sb$, As, or P. One reason it was suggested that these materials may be interesting is that the unit cell is large and contains many (16) heavy atoms, one of the criteria for low lattice thermal conductivity. In addition, p-type samples can exhibit mobilities of nearly 10 000 $\mathrm{cm^2\ V^{-1}\ s^{-1}}$, with effective masses of approximately $0.1m_0$, leading to quite large effective mobility. It turns out that the thermal conductivity, while small, is still too large (by about a factor of 5–10) to make these materials effective thermoelectrics. Mixed crystals can be made that decrease the thermal conductivity and increase Z, but no systems have been found to date with ZT in excess of about 0.6. What makes the skutterudite system so fascinating, however, is an unusually large void in the unit cell. This void can be filled with a rare-earth (RE) atom, creating structures of the form $\mathrm{REFe_4B_{12}}$. Such "filled skutterudite" structures have been known for several years. An important feature of the filled systems is that the radius of the rare-earth atom is significantly smaller than the radius of the void in which it resides. The rare earth "rattles" in a cage of Sb atoms, or in other words, exhibits a soft-phonon mode. It has been predicted and experimentally verified that similar rattling atom systems, such as clathrates, exhibit extremely low, almost glasslike thermal conductivity. Thus it might be possible to create filled skutterudite compounds that exhibit the desirable electrical properties of the binary but with strongly depressed thermal conductivity, thus leading to high Z. Slack (1995) has appropriately termed such structures "phonon-glass electron-crystal," or PGEC, systems. The first demonstration of a strong decrease in the thermal conductivity of a skutterudite upon filling the void was given by Morelli and Meisner (1995). They studied the transport properties of the compound $\mathrm{CeFe_4Sb_{12}}$ and showed that the thermal conductivity was about a factor of 10 lower than that of $\mathrm{CoSb_3}$ and $\mathrm{IrSb_3}$. Unfortunately, the filling of the void with a trivalent rare-earth atom changes the band structure significantly: the filled compounds become metallic, and the power factor decreases substantially. An active area of research today is creating filled skutterudite compounds in which the three electrons donated by the rare-earth atom are compensated by the presence of three electron-deficient atoms on the Sb site, e.g., $\mathrm{NdIr_4Sn_3Sb_9}$. These or similar structures may maintain the high power factor of binary skutterudites while also exhibiting the glassy thermal conductivity of the filled compounds.

Another potential area of interest is in semiconductors in which one of the constituent elements is a rare earth. The $4f$ electrons of the rare-earth atom are subject to the crystalline electric field due to its neighbors. Because of the interaction of the electron spin with the field, the ground-state energy levels of these electrons become split. The splittings of the $4f$-electron levels in the rare-earth atoms are much smaller (typically less than 100 meV) than those in the d-shell electrons in transition metals because the latter, being the outermost electrons, are much more strongly influenced by the crystal field. However, it will be noted that the splitting in the rare-earth case is on the order of the energy of a typical phonon. This means that a phonon of the proper energy can be absorbed by an electron in the lowest lying sublevel in order to promote the electron to the next highest level. When this electron falls back down to the lowest level, a phonon will be re-emitted incoherently; thus this process is a phonon-scattering process. Here we have a means by which phonons can be strongly scattered but which in principle should not strongly affect the conduction electrons in the material. In other words, one can use the magnetic scattering of phonons by crystal-field-split $4f$ electron levels to decrease the phonon mean free path while leaving the electrons alone, a potential route to high Z. One possible area of research would be to search for ternary semiconducting rare-earth–containing compounds that have energy-level splittings corresponding to the energy of the dominant heat-carrying phonons for a particular temperature range.

As a final example we make mention of heavy-fermion, intermediate-valence, and Kondo systems. These structures possess a wide variety of interesting and exotic properties including superconductivity, magnetism, and very heavy electron masses, to name just a few. These features all arise from the proximity of the $4f$-electron energy levels of the rare-earth or actinide atom to the Fermi energy. This same feature leads to anomalously large peaks in the thermoelectric power. In fact, such systems are known to have thermoelectric power factors much larger than even the best thermoelectric materials, but they unfortunately are almost always metals with high thermal conductivity. The question arises as to whether such "heavy-electron" behavior can occur in systems of lower carrier density. This is an emerging area that provides great potential for finding very high-Z materials. If the large power factors of intermediate-valence and heavy-fermion type systems can be combined with the low thermal conductivity of a complex semiconducting structure, there is a good chance of finding substantially larger values of Z.

GLOSSARY

Acceptor: An impurity atom in a semiconductor that produces an energy level close to the top of the valence band. Electrons from the valence band can be easily promoted into this level, producing empty states, or holes in the valence band.

Coefficient of Performance: The ratio of the heat removed by a thermoelectric cooler to the amount of electrical power consumed.

Contact Resistance: The electrical resistance at the junction between a metal and a semiconductor.

Diffusion: The migration of one species or element into another.

Electrical Conductivity: A measure of the ability of a material to conduct electrical charge.

Energy Gap: The energy difference between the completely filled topmost valence band and the totally empty lowest lying conduction band in a semiconductor.

Figure of Merit (Z): The quantity that determines the efficiency of a thermoelectric generator and the coefficient of performance of a thermoelectric cooler; it is related to the Seebeck coefficient S, electrical conductivity σ, and thermal conductivity κ of a material according to $Z = S^2\sigma/\kappa$.

Heat Exchanger: A large-surface-area metallic member of a generator (cooler), the purpose of which is to maintain the temperature of the hot side of the device as close as possible to that of the heat source (sink) and the cold side of the device as close as possible to the cold sink (source).

Mobility: A measure of the ability of a charge carrier to move through a crystal lattice.

Peltier Effect: The liberation of heat at one junction and the absorption of heat at the second junction when current is allowed to flow in a circuit composed of two dissimilar metals.

Phonon: A quantum of vibrational energy in a crystalline material.

Power Factor: The product of the square of the Seebeck coefficient and the electrical conductivity.

Seebeck Coefficient, or Thermoelectric Power: The ratio of the open-circuit electric field to the temperature gradient in a material.

Seebeck Effect: The occurrence of a potential difference in a circuit consisting of two dissimilar metals under conditions of zero net current flow when the two junctions are held at different temperatures.

Thermal Conductivity: A measure of the ability of a material to conduct heat.

Thermal Expansion Mismatch: The difference in the thermal expansion coefficient of two materials joined together.

Thermoelectric Efficiency: The ratio of electrical power produced by a thermoelectric generator to the amount of thermal power absorbed from the heat source.

Works Cited

Berman, R. (1976), *Heat Conduction in Solids*, Oxford, U.K.: Oxford Univ. Press.

Goldsmid, H. J. (1960), *Applications of Thermoelectricity*, New York: Wiley, p. 59.

Kerr, D. L. (1960), in: I. B. Cadoff, E. Miller (Eds.), *Thermoelectric Materials and Devices*, New York: Reinhold, p. 227.

Klemens, P. G. (1958), in: F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, Vol. 7, New York: Academic, p. 1.

Morelli, D. T., Meisner, G. P. (1995), *J. Appl. Phys.* **77**, 3777.

Slack, G. A. (1979), in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, Vol. 34, New York: Academic, p. 1.

Slack, G. A. (1995), in: D. M. Rowe (Ed.), *Thermoelectric Handbook*, Boca Raton, FL: CRC Press, p. 407.

Williams, R. E. (1984), *Gallium Arsenide Processing Techniques*, Dedham, MA: Artech House, Inc.

Further Reading

Cadoff, I. B., Miller, E. (Eds.) (1960), *Thermoelectric Materials and Devices*, New York: Reinhold Publishing.

Goldsmid, H. J. (1960), *Applications of Thermoelectricity,* New York: Wiley, p. 59.

Goldsmid, H. J. (1986), *Electronic Refrigeration,* London: Pion Limited.

Ioffe, A. F. (1957), *Semiconductor Elements and Thermoelectric Cooling,* London: Infosearch Limited.

Slack, G. A. (1995), in: D. M. Rowe (Ed.), *Thermoelectric Handbook,* Boca Raton, FL: CRC Press, p. 407.

THERMOLUMINESCENCE

STEPHEN W. S. MCKEEVER, *Department of Physics, Oklahoma State University, Stillwater, Oklahoma, U.S.A.*

Introduction 355
1. **Basic Physical Principles** 357
1.1 Phosphorescence and Thermoluminescence 357
1.2 Phenomenology: Simple Models 358
1.3 More Advanced Models 359
1.4 Analytical Methods 361
1.4.1 General 361
1.4.2 Peak-Shape Methods 362
1.4.3 Peak-Position Methods 362
1.4.4 Curve-Fitting Methods 362
1.5 Dose Dependence 363
1.5.1 Gamma Dose 363
1.5.2 LET Dependence 365
2. **Instrumentation** 365
2.1 Basic Principles 365
2.2 Heating System 365
2.2.1 Resistance Heating 365
2.2.2 Hot-Gas Heating 366
2.2.3 Optical Heating 366
2.3 Light-Detection System 366
2.4 Emission Spectra 366
3. **Applications** 367
3.1 Dosimetry 367
3.1.1 General Principles 367
3.1.2 Personnel Dosimetry 367
3.1.3 Environmental Dosimetry 367
3.1.4 Medical Dosimetry 368
3.1.5 Retrospective Dosimetry 368
3.1.6 Materials 368
3.2 Dating 369
3.2.1 General Principles 369
3.2.2 Pottery Dating 370
3.2.3 Sediment Dating 370
3.3 Other Applications 371
Acronyms 371
Glossary 371
Works Cited 372
Further Reading 372

INTRODUCTION

> Ninthly, I found that holding it a while near the Flame of a Candle, (from which yet I was careful to avert my Eyes) and being immediately remov'd into the Dark, it disclosed some faint Glimmering, but inferiour to that, it was wont to acquire by Rubbing. And afterward holding it near a Fire that had but little Flame, I found the Stone to be rather less than more excited, than it had been by the Candle.
>
> Tenthly, I likewise indeavour'd to make it Shine, by holding it a pretty while in a very Dark place, over a thick piece of Iron, that was well Heated, but not to that Degree as to be Visibly so. And though at length I found that by this way also, the stone acquired some Glimmering, yet it was less than by either of the other ways above mention'd.
>
> Sir Robert Boyle, 1664

Thus is written the first scientific description of a phenomenon that would later become popularly known as *thermoluminescence,* although the modern word would not appear in print for over two centuries (Wiedemann and Schmidt, 1895). The above quotation comes from Boyle's discourse to the Royal Society, *Experiments and Considerations Touching Colours* (Boyle, 1664), and is included alongside a discussion of a related phenomenon, now known as *triboluminescence,* which is the stimulation of luminescence from an object by fracture of the material. The advance of more than 300 years has resulted in a quantification and understanding of thermoluminescence that, although still far from complete, has enabled

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

it to be used to great effect in the fields of radiation dosimetry, archaeological dating, and many more.

Stated in language less archaic than Boyle's, thermoluminescence (TL) is the emission of light from an insulator (or semiconductor) when the object is heated, following an initial irradiation of the sample. This latter component (i.e., radiation absorption) is an essential feature of the TL phenomenon. Thermoluminescence is one of a family of processes collectively known as thermally stimulated phenomena. Also included in the family are thermally stimulated conductivity (TSC), thermally stimulated capacitance (TSCap), thermally stimulated polarization (TSPC) and depolarization currents [TSDC, including ionic thermocurrents (ITC)], deep-level transient spectroscopy (DLTS), thermogravimetry (TG), and several more (Chen and Kirsh, 1981). Each of these phenomena may be described by two fundamental stages: *stage I*, the perturbation of the system from equilibrium into a metastable state; and *stage II*, the thermally stimulated relaxation of the system back to equilibrium. In each of the above techniques one monitors the nonisothermal change of a particular property of the material (e.g., luminescence, conductivity, capacitance, etc.) as the system returns to equilibrium during stage II.

In the case of TL, stage I necessitates the absorption of energy, normally from ionizing radiation, in order to perturb the material into a nonequilibrium state. Energy storage occurs through the processes of electron–hole pair production, exciton creation, and/or direct displacement damage, followed by charge localization (trapping) at defects within the host lattice of the irradiated material.

During stage II, the stored energy is released as the temperature of the sample is increased, and a fraction of the released energy is in the form of luminescence. The luminescence originates from electron–hole recombination or vacancy–interstitial recombination. In either case, electrons undergo de-excitation from metastable excited states to the ground state, thereby restoring equilibrium. The simplest manner in which this can be achieved is via electron–hole recombination. For example, we may assume that electrons are the thermally freed charges (although the discussion could just as well proceed on the basis that it is holes that are thermally liberated). The process is initiated when phonon coupling between the trapped electron and the lattice results in the absorption of an amount of thermal energy by the electron. The probability per second that this amount of energy is enough to release the trapped electron from its localized state (i.e., to overcome the energy barrier E and undergo a transition into the conduction band) is given by

$$p(T) = s(T)\exp\{-E/kT\}, \qquad (1)$$

where $s(T)$ is a weakly temperature-dependent term known as "the frequency factor" and is related to the local lattice vibrational frequency and the entropy change associated with the charge release. Thus, once the temperature is high enough, the electron will be released into the conduction band and will then be free either to retrap (at either the same or different localized states) or to recombine with trapped holes. The recombination process results in the emission of phonons (nonradiative recombination) or photons (radiative recombination). It is the latter that are monitored when recording TL emission. Thus, we have the thermally stimulated return of the system from its metastable state to equilibrium, with a portion of the excess energy being liberated as light. It should be noted that, for the mechanism described, TSC should also be observed during TL emission since the freed electrons spend time in the conduction band before recombination takes place. However, it is often observed that TL is not accompanied by TSC, suggesting recombination processes that do not involve the delocalized bands. Nevertheless, the concept of thermally stimulated excitation of trapped charge into an excited state from which recombination with charges of the opposite sign can occur is the main concept involved in TL, whether or not the recombination process proceeds via the delocalized bands.

The general principles and mechanisms of TL production are described in the following sections, along with the basic equations that allow a mathematical description of the phenomenon. This is followed by an outline of the principles of instrumentation and descriptions of some of the major applications of the process.

1. BASIC PHYSICAL PRINCIPLES

1.1 Phosphorescence and Thermoluminescence

Luminescence is the emission of light from a material following the initial absorption of energy from an external source—e.g., ultraviolet or high-energy radiation. The emission can be categorized as either fluorescence or phosphorescence, depending upon the characteristic lifetime τ_L between absorption of the excitation energy and emission of the luminescence. The distinction between the two processes is not always clear and is perhaps most easily made using the temperature dependence of τ_L. Consider a ground-state energy level g and an excited-state energy level e, illustrated in Fig. 1(a). Fluorescence is the emission of light that follows the excitation of an electron from g to e and its subsequent return to level g. On the other hand, if the return to the ground state is delayed by a transition into and out of a metastable level m, then delays between excitation and emission can result [Fig. 1(b)], and in this case the process is known as phosphorescence. If the transition into level m occurs at a temperature T, where the energy E of separation between m and e is such that $E \geq$ several kT (k is Boltzmann's constant), then the electron is likely to reside in m for a considerable period. In this case, assuming a Boltzmann distribution of energies, the probability p per unit time for thermal excitation from the trap is exponentially dependent upon temperature according to Eq. (1), and the time between excitation and final relaxation back to the ground state can be considerably delayed by the residence of the electron in the metastable state. The temperature dependence described by Eq. (1) provides a means for distinguishing between the weakly temperature-dependent process of fluorescence and the strongly temperature-dependent process of phosphorescence.

If the metastable level is an electron trap, at an energy E_t below the conduction band, the activation energy for phosphorescence $E = E_t$. For typical numerical values of $E_t = 1.5$ eV and $s = 10^{12}\ \mathrm{s}^{-1}$, the calculation of p at 298 K using Eq. (1) indicates that, for all practical situations, the trap would never thermally release its trapped electron, and, as a result, phosphorescence would never be observed at this temperature. However, if the temperature were to be raised, a temperature may be reached at which p is high enough to ensure that electron release occurs and luminescence is observed. For example, at $T = 300$ K and using the above values for E and s, one can calculate a lifetime $\tau = 1/p$ of 5.06×10^5 years; in contrast, at 500 K, $\tau = 0.36$ s. Thus, consider raising the temperature of the system at some arbitrary rate $\beta = dT/dt$. As T rises, so p increases, producing enhanced luminescence emission. Unlike phosphorescence, which is normally assumed to occur at a fixed temperature, this emission is stimulated in a nonisothermal situation and is thus termed *thermally stimulated luminescence* (TSL), or, more popularly, *thermoluminescence* (TL). As the temperature continues to rise, so the TL intensity increases until such time as the population of trapped electrons in the metastable state is sufficiently depleted, at which point the TL intensity decreases with further increase in temperature. The result is the

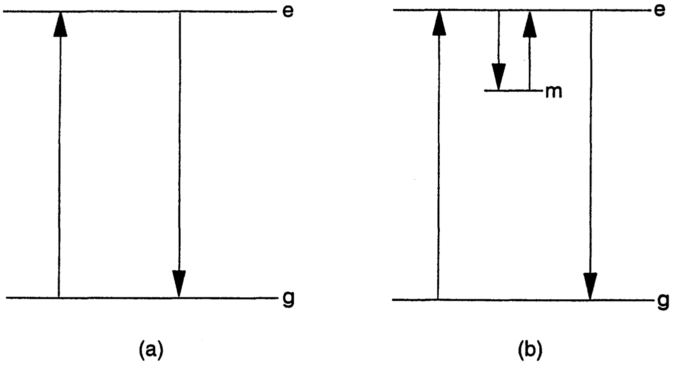

FIG. 1. (a) The ground-state (g) and excited-state (e) energy levels associated with fluorescence emission. (b) The transitions involved in phosphorescence showing the metastable level (m) that causes a temperature-dependent delay between excitation and emission.

characteristic TL peak (the "glow peak"), for which the temperature at the peak maximum is related to the trap depth E_t, the frequency factor s, and the rate of warming β.

1.2 Phenomenology: Simple Models

Using the simple concepts outlined above, Randall and Wilkins (1945) were the first to attempt a mathematical description of the TL glow-peak shape. The phenomenological model used by these authors is based on the band model of an insulator or semiconductor and is illustrated schematically in Fig. 2. Here we see a metastable localized level (level 1) at an energy E_t below the conduction band and a deep level (level 2) assumed to reside below the system's Fermi level (E_F). Prior to irradiation (at equilibrium), level 1 is assumed to be empty, and level 2 is assumed to be full. During irradiation [Fig. 2(a)], electrons are excited across the material's band gap and become trapped at level 1. Likewise, holes become localized at level 2. The result is a nonequilibrium excess of electrons above E_F, and a deficiency of electrons below E_F. During heating of the sample [Fig. 2(b)], thermal excitation of electrons out of level 1 into level 2 occurs, with transport occurring via the conduction band. Recombination of electrons and holes results in the emission of light (TL).

An important feature of the Randall–Wilkins view of TL emission is that the probability of recombination of the released electron with the trapped hole is greater than the probability of retrapping of the released electron. This leads to the concept that the rate of release is proportional to the trapped charge concentration, and is thus termed a "first-order" reaction. The release rate of the trapped electron is thus described by

$$\frac{dn}{dt} = -ns\exp\left\{-\frac{E_t}{kT}\right\}, \tag{2}$$

where n is the trapped electron concentration. [Here the explicit temperature dependence of s from Eq. (1) is dropped.] From this Randall and Wilkins developed their famous equation:

$$I(T) = \phi n_0 s\exp\left\{-\frac{E_t}{kT}\right\} \times \exp\left[-\frac{s}{\beta}\int_{T_0}^{T}\exp\left(-\frac{E_t}{k\theta}\right)d\theta\right]. \tag{3}$$

Here, ϕ is the efficiency of luminescence production from electron–hole recombination and takes into account the fact that not all recombination events lead to a detected photon, n_0 is the initial concentration of trapped charge at the starting temperature T_0, and θ is a dummy variable representing temperature.

A modification of this view was presented by Garlick and Gibson (1948), who used the same one-trap, one-center model of Fig. 2, but who included the concept of significant retrapping of the released charges. This leads to the rate of the reaction being proportional to the square of the trapped charge concentration, and, thus, we have a "second-order" reaction. Here we have

$$\frac{dn}{dt} = -n^2 s'\exp\left\{-\frac{E_t}{kT}\right\}, \tag{4}$$

where $s' = s/N$, and N is the total concentra-

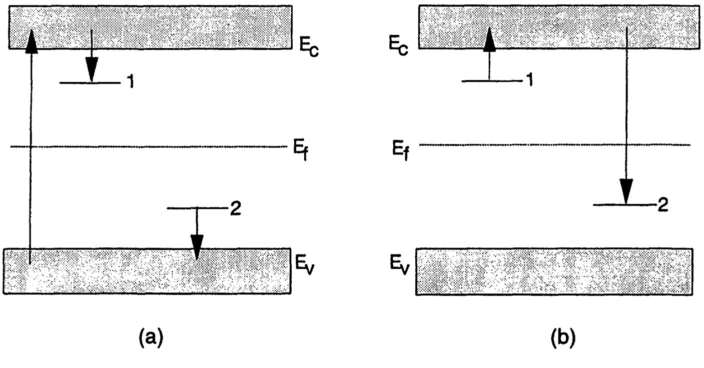

FIG. 2. Simple Randall–Wilkins model for TL. (a) Transitions during irradiation; (b) transitions during heating. E_c and E_v are the conduction- and valence-band edges; E_F is the Fermi level; 1 and 2 are the deep localized states involved in the TL process. The arrows indicate electron transitions.

tion of available electron traps (i.e., $n = N$ corresponds to all available traps full). This leads to the Garlick–Gibson equation for TL under second-order kinetics, namely

$$I(T) = \frac{\phi n_0^2 s' \exp\{-E_t kT\}}{\left[1 + \frac{n_0 s'}{\beta}\int_{T_0}^{T} \exp\left\{-\frac{E_t}{k\theta}\right\} d\theta\right]^2}. \qquad (5)$$

Examples of first- and second-order TL glow peaks are illustrated in Fig. 3. Most theories and descriptions of TL use either one or other of these simple, two-level models, although several empiricists have introduced the idea of "general order," for which the equivalent expression for TL is

$$I(T) = \frac{\phi n_0 s'' \exp(-E_t/kT\}}{\left[1 + \frac{(b-1)s''}{\beta}\int_{T_0}^{T} \exp\left\{-\frac{E_t}{k\theta}\right\} d\theta\right]^{b/(b-1)}}, \qquad (6)$$

where b is neither 1 nor 2 and $s'' = s' n_0^{b-1}$.

1.3 More Advanced Models

The choice of a model for the TL process is not trivial, since it places restrictions upon the number of TL phenomena that one can describe. Since at any given temperature one might expect traps that are "shallow" enough (i.e., E_t small compared with kT so that they are empty of trapped charge) as well as traps that are "deep" enough for the charge to be stable at that temperature, several authors have preferred a more realistic model to that illustrated in Fig. 2, consisting of an active electron level (AT), a set of shallow electron traps (ST), a set of deep, thermally disconnected electron traps (DET), and a set of thermally disconnected hole traps (DHT), as shown in Fig. 4. The term "thermally disconnected" refers to the thermal stability of the traps. If, during TL production, the charge trapped in deeper centers is stable, the deeper traps are said to be "thermally disconnected." The shallow traps (ST) play no role in governing the shape of the TL peak due to AT.

Over the temperature region of interest (the "active" temperature range), electrons from AT are released, and the freed electrons recombine with holes trapped at DHT, with the subsequent emission of thermoluminescence, of intensity $I(T)$. Defining n_c to be the concentration of free electrons, we may write a set of rate equations describing the traffic of charge between the various energy levels, thus:

$$\frac{dn}{dt} = -ns \exp\left\{-\frac{E_t}{kT}\right\} + n_c(N - n)A \qquad (7)$$

and

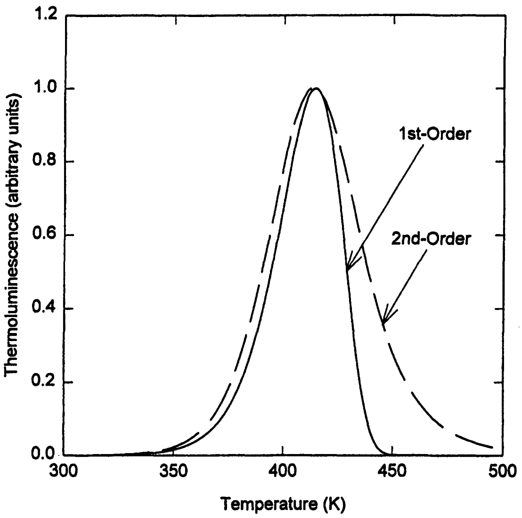

FIG. 3. Normalized first- and second-order TL glow peaks calculated with $E_t = 1.0$ eV, $s = 10^{11}$ s^{-1}, and $\beta = 1.0$ °C s^{-1}. For the second-order peak $n_0/N = 1$.

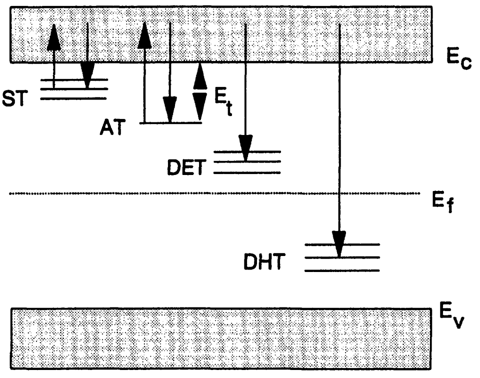

FIG. 4. More advanced model for TL (from Lewandowski *et al.*, 1994). ST, shallow traps; AT, active traps; DET, deep, thermally disconnected electron traps; DHT, deep, thermally disconnected hole traps. The arrows indicate electron transitions during heating. E_c, conduction-band edge; E_F, Fermi level; E_v, valence-band edge.

$$\frac{dn_c}{dt} = -\frac{dn}{dt} - \frac{n_c}{\tau_r}, \tag{8}$$

where N is the total concentration of traps, of which n are occupied; A is the trapping transition constant ($m^3\ s^{-1}$); and τ_r is the recombination lifetime [equal to $(n_h A_r)^{-1}$, where n_h is the concentration of holes trapped at the recombination sites and A_r is the recombination transition constant]. The first term on the right-hand side of Eq. (7) represents the rate of thermal excitation of the trapped electrons from the trap, whereas the second term represents the rate of retrapping. Similarly, the second term in Eq. (8) is the rate of recombination of free electrons with trapped holes. Approximate solutions to these equations normally proceed via the introduction of the "quasiequilibrium" (QE) approximation, which states that $dn_c/dt \ll n_c/\tau_r$, or $dn/dt = -n_c/\tau_r$. Since the intensity of the thermoluminescence emission is given by $I(T) = \phi n_c/\tau_r$, then

$$I(T) = \frac{\phi n s \exp(-E_t/kT)}{\tau_r[1/\tau_r + (N - n)A]}. \tag{9}$$

With the inclusion of the "slow-retrapping" or "first-order" assumption that $\tau_r^{-1} \gg (N - n)A$, we have, after integration, the Randall–Wilkins equation given in Eq. (3). Similarly, with the assumptions $A = A_r$ and $n = n_h$, one gets the Garlick–Gibson expression, given in Eq. (5).

The importance of understanding the shape of a TL glow peak and the parameters that govern it lies in the interpretation of changes to the glow curve that often result from, say, an increase in absorbed dose, or from various thermal treatments (including preirradiation or postirradiation storage at room temperature). One sometimes finds complex alterations in the glow-curve shape that are difficult to understand without an appreciation of the shape of the individual glow peaks that make up the complete glow curve. However, before one can proceed to analyze an experimental TL glow curve, it is necessary to verify whether or not the varied assumptions used in the derivation of the simple equations are in fact justified. The most important of these is the QE approximation ($dn_c/dt \ll n_c/\tau_r$). To examine the validity of this assumption, it is useful to define a function $Q(T)$ thus:

$$Q(T)\frac{n_c}{\tau_r} = -\frac{dn}{dt}, \tag{10}$$

where $Q(T) = 1$ signifies QE. If QE is not sustained during TL emission, then, for first-order kinetics, the actual glow peak shape is the same as that described in Eq. (3), but divided by the function $Q(T)$. Modified equations that exactly describe the TL peak, even when the QE assumption is not valid, have been developed (Lewandowski *et al.*, 1994). In a wide variety of cases, however, $Q(T) \approx 1$ over the active temperature range, and thus the need for this more exact equation is weak in these circumstances. In particular, only (negligibly) small errors in E_t and s result from the use of Eq. (3) in most cases of first-order kinetics. However, similar conclusions *cannot* be reached for non-first-order kinetics, and indeed the dual assumptions of second-order kinetics *and* QE are questionable.

The second assumption of importance is that of kinetic order. A feature of first-order TL peaks is that they are characterized by an invariance of the position of the peak as a function of initial trap population—viz. Eq. (3). Thus, this property can be tested by following the variation in the peak position as a function of irradiation time (i.e., absorbed dose). Any shifts in the peak with absorbed dose should lead one to suspect non-first-order kinetics. An empirical estimate of the kinetic order was devised by Chen (1969) who used a geometrical factor $\mu_g = \delta/\omega$, where ω is the full width of the peak at its half height (in °C) and δ is the half-width on the high-temperature side of the peak (in °C). Theoretical first-order peaks were shown to have $\mu_g = 0.42$; second-order peaks gave $\mu_g = 0.52$. Intermediate values of μ_g correspond to intermediate kinetics for which, unfortunately, physical meaning is lost.

The simple models for charge trapping and detrapping can also be examined experimentally. In the model described in Fig. 4 it was explicitly assumed that only the electrons are thermally unstable over the "active" temperature range; no allowance was made for the thermal release of holes from the recombination centers. If this in fact were not the case, then the TL intensity would exhibit thermal quenching (i.e., the efficiency ϕ would be temperature dependent, decreasing with increasing temperature). One can search for this in a number of ways—for example, by monitoring the luminescence emission during irradiation (known as radioluminescence) as a function of irradiation temperature, or by performing the TL measurement at a wide variety of heating rates, such that the temperature of the TL peak is shifted to higher temperature for higher heating rates, where loss of TL would be observed if thermal quenching was occurring. Although not a unique cause of thermal quenching, one should not rule out simultaneous hole release as a potential difficulty if this phenomenon is observed. In these cases, analysis of the TL peak will reveal a value for E_t that cannot accurately be interpreted as the trap depth depicted in Fig. 4.

Other possibilities for a temperature-dependent luminescence efficiency exist. The model of Fig. 4 allows for the recombination of electrons at only one recombination center. In real materials several such centers may exist, each with a temperature-dependent recombination probability. If all photons were detected with equal efficiency, independent of their wavelength, the existence of multiple recombination sites would not be a problem. However, modern detectors exhibit a varying detection efficiency as a function of wavelength such that changes to the glow-curve shape will occur as the temperature increases, and monochromatic detection at fixed wavelengths will yield TL peaks in different positions, depending upon the wavelength chosen. Such effects are most easily accounted for by measuring simultaneously the TL intensity as a function of temperature and as a function of wavelength. An example of such a glow curve—known as a "3D" plot—is shown in Fig. 5. Such data reveal much information regarding the luminescence sites in the TL material. In this case, emission due to transitions of the 4*f* electrons in Dy^{3+} impurity ions charge compensated by fluorine interstitial ions in C_{4v} symmetry are observed at 480 nm ($^4F_{9/2}$–$^6H_{15/2}$), 575 nm ($^4F_{9/2}$–$^6H_{13/2}$), 660 nm ($^4F_{9/2}$–$^6H_{11/2}$), and 750 nm ($^4F_{9/2}$–$^6H_{9/2}$). The emission near 550 nm is probably due to Er^{3+}. The emission at 375 nm for the TL signal at 300 °C is due to Ce^{3+} charge compensated by O^{2-} ions in C_{3v} symmetry, while the doublet at 324 nm and 340 nm near 100 °C is due to the split ground states of Ce^{3+} ions (i.e., $^2D_{3/2}$–$^2F_{5/2}$ and $^2D_{3/2}$–$^2F_{7/2}$) charge compensated by fluorine interstitial ions in C_{4v} symmetry.

1.4 Analytical Methods

1.4.1 General Before analysis of a TL glow curve can be attempted, one practical difficulty that has to be overcome is that real glow curves rarely consist of isolated glow peaks. In addition there may be many decay routes that are nonradiative, yet influence the overall TL performance. These in turn give rise to problems of two kinds. First, the presence of more than one glow peak and/or more than one recombination pathway points immediately to the possibility that the traps will interact with each other. That is, charge detrapped from one level will either

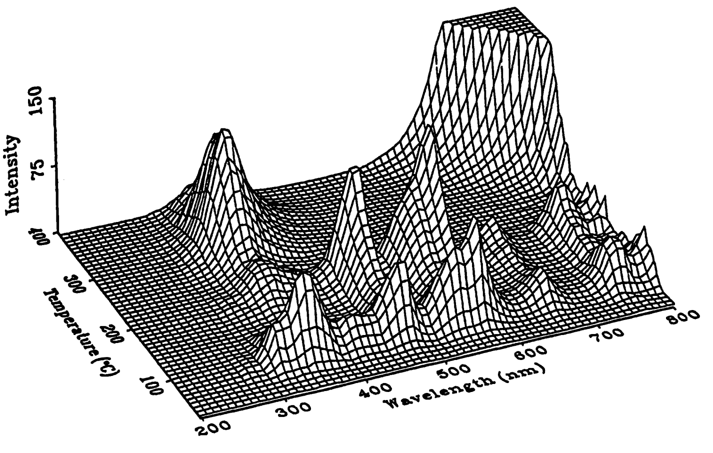

FIG. 5. A "3D" glow curve of TL versus wavelength versus temperature for a sample of Brazilian fluorite, mixed with sodium chloride as a binder. Reproduced from McKeever *et al.* (1995) with kind permission of Nuclear Technology Publishing, Bristol, U.K.

recombine at the radiative recombination site, recombine at the nonradiative recombination site, or be retrapped in deeper traps. The models applied to analyze these cases, however, rarely take into account interactions of this type and treat the glow peaks as isolated entities. However, for first-order kinetics, in which there is little retrapping, problems due to interactive kinetics are minimal, and the glow curve can be considered to be the simple superposition of individual first-order peaks. Peaks that exhibit non-first-order kinetics, however, cannot be treated in isolation.

The second difficulty is a more practical one. The observed glow peaks are often overlapping to the extent that clean peaks, necessary for analysis, are seldom obtained. The usual course is to adopt a "thermal cleaning" program in which the sample is heated to an intermediate temperature that removes the low-temperature, interfering signals. Ideally, subsequent reheating of the sample will then reveal a clean, higher-temperature peak. Several variants of these cleaning processes exist (see McKeever, 1985). Each is reasonably useful in arriving at clean, rising portions of the TL signals, but rarely can a complete, clean peak be isolated, making application of some analytical methods difficult.

There have been many methods of analysis devised in order to extract parameter values (E and s in particular) from real glow curves, and there are certainly too many to review in this article. For more detailed discussions of these methods, we refer the reader to the books by McKeever (1985) and, especially, Chen and Kirsh (1981). Here we mention just three of the most popular methods.

1.4.2 Peak-Shape Methods Chen's analysis of the kinetic order of a TL peak using the geometrical factor $\mu_g = \delta/\omega$ (Chen, 1969) led to several expressions for the trap depth E_t based on the shape of the peak. With the definition of δ and ω as above, and with the additional definition of τ as the half-width on the low-temperature side of the peak, Chen arrived at the following general expression for E_t:

$$E_t = c_\gamma\left(\frac{kT_m^2}{\gamma}\right) - b_\gamma(2kT_m), \qquad (11)$$

where γ is either δ, ω, or τ. The values of the constants c_γ and b_γ depend upon whether the kinetics of the peak are first or second order and are tabulated in Table 1.

Modifications to the above values for c_γ and b_γ are necessary if the kinetics are neither first nor second order, or if the frequency factor (assumed to be constant in the above treatment) is temperature dependent.

1.4.3 Peak-Position Methods A second family of techniques utilizes the change in the peak position as a function of the heating rate β used during glow-curve readout. Differentiation of Eq. (2) and equating the result to zero at $T = T_m$ leads to

$$\frac{\beta E_t}{kT_m^2} = s\exp\left\{-\frac{E_t}{kT_m}\right\}, \qquad (12)$$

from which we see that a plot of $\ln(T_m^2/\beta)$ against $1/T_m$ will yield a straight line of slope E_t/k and intercept $\ln(E_t/sk)$. This method is useful because it relies only upon an accurate assessment of the shift in the peak as a function of heating rate β and does not require clean peaks. Too much overlap with neighboring peaks, however, can give rise to false peak positions.

Although calculated for QE and first-order kinetics, the analysis can be applied to nonequilibrium, non-first-order cases with little additional error. A linear heating function (β = constant) is not necessary; in this case the value of β at $T = T_m$ is all that is required.

1.4.4 Curve-Fitting Methods Perhaps the first of the methods that can be designated as "curve fitting" is that of Garlick and Gibson (1948), who developed the so-called "initial rise" technique. Here one relies on the fact that at low temperature, just as the trap starts to empty (i.e., when n is approximately constant), all TL equations can be described by

Table 1. Coefficients for the calculation of E_t using Chen's method.

	First order			Second order		
	ω	τ	δ	ω	τ	δ
c_γ	2.52	1.51	0.976	3.54	1.81	1.71
b_γ	1.0	1.58	0	1.0	2.0	0

$$I(T) = \text{constant} \times \exp\{-E_t/kT\}. \tag{13}$$

Thus a plot of $\ln(I)$ versus $1/kT$ for this initial-rise part of the curve will reveal a straight line of slope $-E_t/k$, from which E_t can be calculated. The frequency factor s can then be calculated from the position of the peak maximum, T_m, by use of Eq. (12).

Equation (13) applies for *all* kinetic orders and is insensitive to the QE approximation, making it a valuable analytical method. Its accurate use, however, relies on a clean rising part of the TL peak being available—fortunately, something that can be achieved with thermal cleaning techniques.

In the initial-rise analysis one fits the rising portion of the TL curve by Eq. (13). In doing so, however, one "throws away" significant amounts of data from the rest of the glow peak. This is even more true for the peak-shape and peak-position methods in which only one or two datum points are used in the analysis. Better still would be a method that utilized all of the available data—i.e., one in which the whole peak was fitted by a predetermined peak shape. The overlap of peaks is handled in this treatment by making initial guesses at the number of peaks present, and fitting the complete data set by the superposition of several peaks using nonlinear least-squares fitting. The normal procedure is to minimize the error between the experimental glow curve and the sum of the theoretical peaks, using E_t, s, and peak size as fitting parameters. An example of a LiF:Mg,Ti glow curve analyzed in this way is shown in Fig. 6.

1.5 Dose Dependence

A characteristic of thermoluminescence is that it requires the initial absorption of energy from ionizing radiation. One might expect, therefore, that a straightforward relationship between the initial absorbed dose of irradiation and the subsequent TL signal will be observed. This relationship is now examined, for both highly penetrating radiation and weakly penetrating radiation.

1.5.1 Gamma Dose The *dose response* $F(D)$ is defined as the functional dependence of the intensity of the measured TL signal upon the absorbed dose. The ideal material would have a linear dose response over a wide dose range; however, most materials used in practical dosimetry display a variety of nonlinear effects, as shown in Fig. 7. In particular, one often finds that the response of a TL material is linear, then supralinear, then sublinear as the dose is increased. The normalized dose-response function (or supralinearity index) $f(D)$ is defined as follows:

$$f(D) = \frac{F(D)/D}{F(D_1)/D_1}, \tag{14}$$

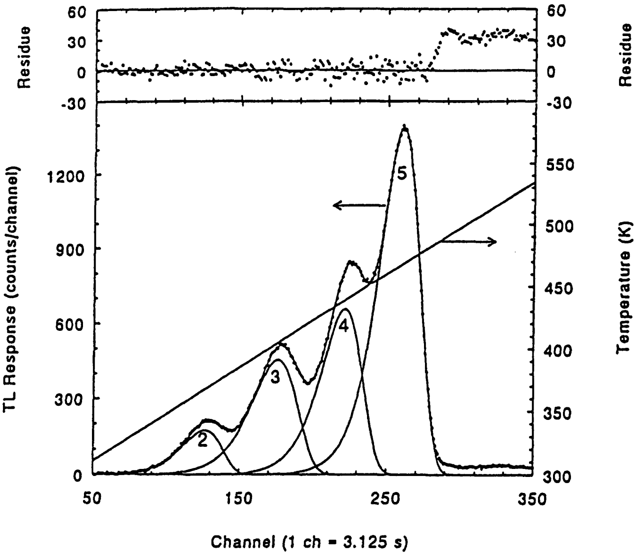

FIG. 6. Example fit of an experimental glow curve for LiF:Mg,Ti. Reproduced from Bos *et al.* (1992) with kind permission of the authors and the Institute of Physics Publishing Ltd., Bristol, U.K.

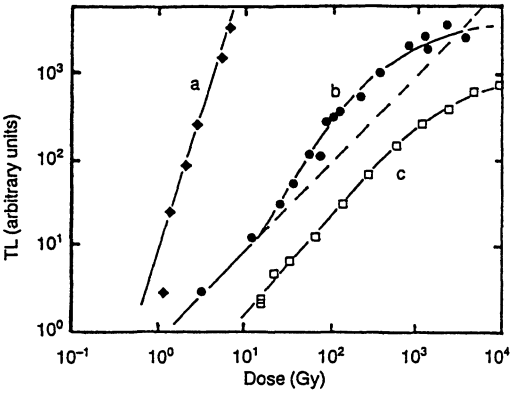

FIG. 7. The dose response of three TLD materials: (a) SiO_2, (b) LiF:Mg,Ti, and (c) CaF_2:Mn. Reproduced from McKeever *et al.* (1995) with kind permission from Nuclear Technology Publishing, Ashford, U.K.

where $F(D)$ is the dose response at a dose D and D_1 is a low dose at which the dose response is linear. Thus, our ideal material would satisfy $f(D) = 1$ over a wide dose range. In many thermoluminescence dosimeter (TLD) materials, however, $f(D) = 1$ is found only over a narrow dose range, up to a few grays. Supralinearity, defined as $f(D) > 1$, is commonly observed, while sublinearity $[f(D) < 1]$ is most often observed during the approach to saturation. These features are illustrated in Fig. 7.

The actual value of $f(D)$ at any given dose is governed by a wide range of material and external factors, the latter including the type and energy of the radiation used and the detailed procedure used to heat the sample during glow-curve measurement. Of particular importance is the observation that $f(D)$ depends upon the LET [linear energy transfer, dE/dx (eV/m)] or ionization density of the radiation field. The LET dictates the rate at which the energy E of the incident particle is lost as a function of distance x into the material. High-LET radiation loses its energy quickly, leaving regions of high ionization density within the material. Heavy charged particles (e.g., alpha particles) deposit their energy in densely ionized particle tracks. Thus, $f(D)$ values, obtained at about 200 Gy, for LiF:Mg, Ti have been observed as high as ~3.5 for ^{60}Co irradiation, but only ~1.5 for 20-kV$_p$ x-ray irradiation, and ~1.0 for alpha-particle irradiation (McKeever *et al.*, 1995). Not only is the extent of the supralinearity dependent upon LET, but the dose at which the onset of supralinearity becomes evident is also LET dependent.

Important material factors include the nature of the host material itself. For example, in Fig. 7 we see that the growth of TL with dose in quartz is highly supralinear $[f(D) > 2]$ while for CaF_2:Mn it is almost linear $[f(D) = 1]$. LiF:Mg,Ti, however, displays a linear-supralinear-sublinear growth. In LiF:Mg,Ti, decreasing the Mg content enhances supralinearity, while increasing the OH impurity content decreases supralinearity. All these factors indicate a dependence of $f(D)$ on the nature of the host material and its defect structure.

Many theories have been forwarded to explain the cause of supralinearity in TL materials—particularly LiF:Mg-type materials. Conveniently, they can be classified as models in which the critical processes occur during the absorption of radiation and models in which the critical processes occur during heating (i.e., during TL readout). However, most evidence clearly indicates that the critical mechanism is that of competition during the heating stage of TL readout. During heating, electrons released from traps may either recombine with trapped holes to produce TL or be retrapped in deeper traps, which act as competing centers. At low doses, where the distance between traps and recombination sites is large, the probability that a freed charge can find a recombination site without falling into a competing trap is low. Thus the response of the dosimeter is reduced in this low-dose region. At higher doses, however, not only is the mean distance between traps and recombination sites decreased, but the number of competing centers is also reduced. (Note that a competing trap is an *empty* trap; once filled, it no longer acts as a competitor.) Thus, at higher doses the response of the material is enhanced and more TL/unit dose is the result. Eventually, at very high doses, the material saturates. Such a material has $f(D) > 1$ over the entire dose range, up to saturation; an example is SiO_2, shown in Fig. 7. [Note that the data are plotted on a log–log plot, such that $f(D) = 1$ produces a slope = 1. Supralinearity shows as a slope > 1.]

If the material did not have any competing traps (no competition), the sensitivity (i.e., TL/unit dose) would be higher at low doses and linearity would be observed up to saturation (an example is CaF_2:Mn, shown in Fig. 7). Thus we see that supralinearity should correctly be considered to be an under-response at low doses, rather than an over-response at high doses.

With some materials one often finds a linear region *before* the supralinear region (see LiF:Mg,Ti in Fig. 7). To explain this it is necessary to invoke spatial localization between the traps and the recombination centers. Spatial association of this type produces a linear term at low doses, before one observes the nonlinear effect at higher doses.

1.5.2 LET Dependence With high-energy gamma irradiation (low LET) one may safely view the distribution of the filled traps and recombination sites as an "isotropic sea." However, for heavy–charged-particle (HCP) irradiation (high LET), we have a nonisotropic distribution of densely ionized regions along linear tracks. Calculations indicate that the dose is deposited within a few tens of nanometers of the track center for high-LET particles. Thus, there are regions *within* the tracks of high concentrations of trapped charges and very few competitors, and regions *between* the tracks of very few trapped charges and high numbers of competitors. For high enough ionization densities the mean free path of a freed electron inside the track is less than the track radius, and recombination within the track, leading to TL, is highly probable. Thus, the dose-response curve is linear, although it saturates early since all the available sites within the track volume quickly become saturated. For lower-LET particles, there is a probability that some electrons will escape from the track core without recombining and thus enter the intertrack region. Here the probability of finding a competitor is high, and the electron is likely to be lost from the TL process. However, as the dose is increased the probability of reaching an adjacent track also increases since the intertrack distance decreases. The escaped electron may now recombine with a hole in an adjacent track and so produce a TL photon. Under these circumstances, the TL output (per unit dose) will also increase, and thus supralinearity will occur. Thus, the lower the LET, the greater the degree of supralinearity. It is clear that as the LET of the HCP increases, the tracks become more localized, and consequently the dose levels at which charge exchange between tracks can take place also increase (track interaction). In this way the observed LET dependence of the value of $f(D)$, and the dose at which the onset of supralinearity occurs, are explained.

2. INSTRUMENTATION

2.1 Basic Principles

Although the instrumentation required to monitor TL has come a long way since Sir Robert Boyle's day, the simplicity of the necessary equipment is such that an accurate and reliable system can be set up for little expenditure. The basic requirements are an oven, a reliable temperature controller, a photodetector, and a data-storage medium. Additional design criteria are required if TL below room temperature is monitored, or if the spectrum of the TL emission is measured.

2.2 Heating System

The heart of any TL instrument is the heating system. Three different types of heating method can be found in widespread use. The objectives are reproducible and controllable heating profiles. Ovens and sample chambers are most often filled with inert gas during heating to enhance thermal contact between the heater and sample, and to reduce spurious emissions of light from chemical effects.

2.2.1 Resistance Heating Here one passes high current through a thin, low-impedance load, such as a tantalum, nickel-chrome, or platinum alloy strip (the planchet). The sample (powder or solid chip) is placed in direct contact with the planchet, the temperature of which is monitored via thermocouples. Alternative designs involve the use of wire coils in contact with more massive metal blocks, or in direct contact with the sample itself. (An example of the latter is the U.S. Navy's "bulb dosimeter" in which the sample is directly deposited onto the coil using adhesive and the whole encased in a glass bulb. Electrical leads protruding through the glass envelope enable electrical contact to be made.) The low thermal inertia of thin planchets or coils enables fine temperature control and rapid heating and cooling. The range of heating rates achievable with resistance-heating techniques is from fractions of a degree per minute to several tens of degrees per second.

2.2.2 Hot-Gas Heating Rapid heating of solid samples, with precision temperature control, can be achieved via the use of hot gas (usually nitrogen) as a heat-exchange medium. A jet of the hot gas is directed via nozzles, at a typical flow rate of a few liters per minute, onto the sample and extracted from the sample chamber at the same rate, thus preventing heating of the sample surroundings. Uniform heating of the sample can be achieved.

2.2.3 Optical Heating Some early, commercial TL systems utilized the absorption of infrared radiation from a tungsten lamp, filtered to remove unwanted visible and ultraviolet wavelengths. The absorption of the infrared energy by the sample resulted in a controllable rise in temperature. Only small, thin samples of low thermal mass can be used effectively with such systems. A modern development of this concept is the use of infrared (CO_2) cw lasers to heat either the sample or a substrate to which the sample has been intimately bonded. The active samples are thin (<100 μm; 10–20 μg), and extremely high heating rates ($\beta \geq 10^4$ C°/s) can be achieved with short (10–1000 ms) laser pulses (at powers of a few watts). Such high heating rates result in a correspondingly high signal-to-noise ratio because of the combined effect of an enhancement of the TL output, according to Eq. (3), (5), or (6), and the discrimination against the noise counts from the detector. Potential limitations against the use of such high rates of heating stem from "thermal quenching" of the luminescence efficiency. As the heating rate increases, the temperature at which the TL is emitted also increases. Some phosphors lose luminescence efficiency with temperature increase, and potential gains in output intensity are offset by the quenching phenomenon. Furthermore, for all materials the incandescence background also increases with temperature, sometimes to the point of intolerance.

2.3 Light-Detection System

The low-intensity light emitted during a TL experiment is normally detected with photomultiplier tubes (PMTs). Either dc current or photon-counting modes may be used, the former at high signal levels, and the latter at low signal levels. The key to high detectability is to match the energy response of the PMT photocathode to the wavelength of the emitted light. Most wide–band-gap materials emit in the visible, particularly blue, part of the spectrum, and thus bialkali photocathodes are well matched to such signals. Longer-wavelength emitters, however, require the use of extended bialkali or GaAs photocathodes. High-gain (10^6–10^7), low–dark-current (nanoamperes) PMTs are essential. Dark currents may be reduced by the use of Peltier-cooled PMT housings to limit thermionic emission from the photocathode.

2.4 Emission Spectra

An examination of the schematic energy-band diagram of Fig. 2 suggests that only one emission wavelength will be observed during TL readout since only one type of radiative recombination center is present. However, real materials often contain several types of recombination site, each of which may be represented by its own characteristic emission wavelength. Conventional measurements of TL glow curves use detectors that are sensitive over a broad range of wavelengths and make no effort to discriminate between the possible different emission colors. Detection through broadband filters, centered in the visible, is the norm. Unfortunately, in doing this, much potential information is lost regarding the emission processes. In addition, one can easily arrive at false conclusions regarding the kinetics of the TL process, and, furthermore, one may misinterpret observed changes in, for example, glow-curve shape and intrinsic sample sensitivity following thermal annealing.

Nevertheless, TL recorded through broadband filters is still the most popular recording method, and an additional dimension of information regarding the TL process is often lost. Some early attempts at measurement of spectral data were made using narrow-band interference filters, but more recently greater detail has been obtained by measuring the TL intensity as a function of temperature and wavelength simultaneously. Such isometric measurements lead to plots of intensity versus temperature versus wavelength ("3D" plots). Various spectrometers have been described in the literature to achieve this, including simple rotating filter packs and rapid-scanning monochromators. However, most information is obtained by multiplexing the data and dispersing the signal over a defined wavelength range, and measuring the intensity over each wavelength interval separately but simultaneously. Modern techniques include Fourier-transform spectroscopy or, more popularly, use of position-sensitive photodetectors. In the latter, the light is dispersed over the wavelength range of interest (using a grating or prism spectrograph) and the dispersed light is detected over the active area of a device such as a CCD camera, diode array, or position-sensitive PMT. Such measurements are now quite common. The utility of this detection mode has been reviewed by Townsend and Kirsh (1989). An example of a "3D" glow curve is shown in Fig. 5.

3. APPLICATIONS

3.1 Dosimetry

One of the earliest suggested applications of TL was in the detection and measurement of absorbed radiation—*radiation dosimetry.* Dosimetry applications fall into several general categories, including the detection of the absorbed dose to people and the environment.

3.1.1 General Principles The basic premise in the use of TL in dosimetry is that the intensity of the TL can be related, in a straightforward way, to the absorbed radiation dose. In an ideal system, the TL signal would be directly proportional to the absorbed dose over a wide dose range. The dose response could be calibrated and unknown doses determined by simply measuring the TL intensity resulting from the exposure; thereby, a useful dosimetric method is presented. In fact, although complications are abundant, this simple concept can be considered a reality. The sensitivity and wide dynamic range of TL in this application are unrivaled. The method is routinely used to measure doses as small as those delivered by the natural environment over a period of days, or doses as high as those delivered inside nuclear reactors—a dose difference of a factor of approximately 10^{11}. In what follows we describe briefly the range of uses of TL in radiation dosimetry, and in doing so we introduce the term thermoluminescence dosimeter, or TLD.

3.1.2 Personnel Dosimetry The primary objective of personnel dosimetry is the monitoring of the radiation dose delivered to personnel during routine occupational exposure. Examples include reactor workers, hospital radiotherapy technicians, or naval personnel on nuclear-powered vessels. In addition to routine monitoring, the field of personnel dosimetry also includes the determination of absorbed doses caused by accidental exposure to radiation. The overall desired aim is to limit an individual's total exposure to the maximum permissible dose values recommended by the regulatory agencies.

Broadly, personnel dosimetry covers the following:

1. Extremity dosimetry—determination of the maximum dose equivalent to any part of the hands, arms or feet.
2. Whole-body dosimetry—determination of the dose equivalent absorbed at a depth of 400–1000 mg cm^{-2} in human tissue below the surface of the body and/or the dose equivalent in critical organs. Primary interest is in penetrating radiation [gamma rays, x rays (>15 keV), and neutrons].
3. Tissue dosimetry (i.e., skin dose)—determination of the dose equivalent absorbed at a depth of 5–10 mg cm^{-2}. Here one is interested in nonpenetrating radiation (beta particles, and <15-keV x rays).

In all such applications the primary requirement of a TLD material is that of tissue equivalency. Energy independence and high sensitivity are also desirable.

3.1.3 Environmental Dosimetry The continuous monitoring of radiation released to the environment has become a major concern for industrialized nations, and TLD usage is an important factor in this type of activity. Since exposure levels are low (with equivalent doses of, typically, 10^{-2} mGy), long exposure times are required, and, thus, long-term stability of the TLD becomes vitally important, along with extreme sensitivity. Gamma and neutron emitters are the main radiation sources of interest from such potential sources as escape of gaseous radionuclides from nuclear power stations during day-to-day operations, low-level waste disposal, nuclear fuels reprocessing, nuclear power-station accidents, and others. Throughout the United States and Europe, TLD systems are in place near nuclear installations for the purpose of monitoring radiation levels, above the natural background, that can be linked with the operation of these facilities. In addition, since the advent of manned space flight, there has been a growing interest in space dosimetry. One obvious cause for this interest is the increased exposure to harmful radiations that people (astronauts) can expect in this environment. Additional concern lies in the prolonged exposure of electronic components, where the decreasing scale of the active regions in these devices (a few square microns in cross section) and the shift from analog to digital circuitry means that device failure due to discrete radiation events (interactions with single particles, known as single-event upsets) is becoming a serious concern. The radiation sources are galactic cosmic rays, of which the main component is protons (1–1000 MeV), and HCPs from the solar wind. TLDs have been used on several recent manned space flights for the purpose of determining the dose from such penetrating radiation.

3.1.4 Medical Dosimetry The small size and high sensitivity of TLD materials has long been exploited in clinical studies since small TLD samples can be inserted into appropriate openings on the human body before exposing the patient to ionizing radiations during diagnosis and/or therapy. The exposed TLDs are then retrieved and analyzed. In this way physicians are able to determine actual doses delivered to critical internal organs during these procedures and from such information are able to prescribe necessary additional treatments. Such utility is hardly possible with any other form of radiation dosimeter.

The two areas of use for clinical radiation exposure of humans are diagnostic radiology (e.g., x-ray exposure in mammography, dentistry, and general health screening) and radiotherapy (primarily cancer therapy of various types). Radiation types include x rays (as low as 10 keV), gamma rays (from ^{137}Cs or ^{60}Co), beta particles (up to 40 MeV), heavy charged particles, and neutrons. Doses range from 10^{-5} to 1 Gy in radiology and up to 10 Gy in radiotherapy. Dose estimation accuracy of better than ±3% is aimed for since errors greater than this can adversely affect the outcome of the treatment.

Clearly, the foremost requirement of a TLD material to be used in this fashion is tissue equivalency. High sensitivity is desirable so that TLD sizes can be kept as small as possible for *in vivo* measurements. Nontoxicity is also a required property. Furthermore, since the doses used can be quite high, a linear dose response over a wide dose range is an advantageous property.

3.1.5 Retrospective Dosimetry Worldwide, interest is growing in the development of new and improved methods for retrospectively assessing the radiation dose in accidentally contaminated areas. For a number of years TL techniques have been used with some success in this application. Examples include Chernobyl (Byelorussia and Ukraine) and the Techa River (Russia), along with similar efforts at Hiroshima and Nagasaki (Japan) and the Nevada bomb test site (United States). The thrust has been to use commonly available materials, located at the actual sites, that may be used as local dosimeters. Example TL materials that are proving to be useful in this context are pottery, tile, porcelain, and other ceramic objects. From these materials quartz and feldspar can be extracted, and the absorbed dose to these components is determined by TL dosimetry techniques.

3.1.6 Materials Obvious desirable properties of TLD materials include a linear response to absorbed dose and an adequate sensitivity. As discussed in Sec. 1.5, the dose response of most TL materials is far from linear, displaying a variety of nonlinear effects depending upon the dose range and the irradiation type. However, at sufficiently low doses most popular TLD materials display an approximately linear response, along with a sufficient sensitivity, to enable their use in several of the fields discussed above.

Tissue equivalency is an important consideration for those materials to be used in personnel or medical dosimetry. For photon radiation the important material factor is the so-called effective atomic number, Z_{eff}, which dictates the amount of energy absorbed by the material in a given radiation field. For human tissue $Z_{\mathrm{eff}} = 7.4$, and thus any material with an effective atomic number around this value will mimic human tissue and is known as "tissue equivalent." The higher the value of Z_{eff} and the lower the energy of incident radiation photons, the larger the TL response to a given dose. Thus, non-tissue-equivalent materials tend to exhibit an "over response" at lower photon energies. Normally, one defines the energy response of a TLD material with reference to the response at a gamma photon energy of 1.25 MeV. Thus, CaF_2 ($Z_{\mathrm{eff}} = 16.3$) has an over response of 3 at an energy of 10^{-1} MeV, while LiF ($Z_{\mathrm{eff}} = 8.14$) has almost no over response at the same energy and is approximately tissue equivalent.

TLD materials should exhibit high stability of the signal over exposure and analysis time scales, and preferably should be insensitive to light exposure, chemical reaction, and humidity.

The most popular material for dosimetry applications is LiF doped with Mg and Ti. This tissue-equivalent material has found application in all personnel and medical applications. CaF_2 (doped with Mn, Dy, or Tm impurities) is a popular environmental dosimeter, and some dosimetry services still use natural fluorite, even for personnel dosimetry. Other popular materials include $CaSO_4$ (Mn, Dy, or Tm doped) and $Li_2B_4O_7$ doped with Mn or Cu. Recent developments include new, ultrasensitive forms of LiF (LiF:Mg,Cu,P) and sapphire (Al_2O_3:C).

3.2 Dating

3.2.1 General Principles The accumulated dose absorbed by ceramic artifacts over their archaeological or geological lifetime can be appreciable, and this dose lends itself to determination using TL. The materials of interest are ceramics containing luminescent minerals—particularly quartz and feldspar—that, when heated after irradiation exposure, emit TL proportional to the time of their exposure. The radiation originates from cosmic rays and from gamma, beta, and alpha irradiation from the local surroundings (due to traces of uranium, thorium, and potassium). The "natural" TL signal is thus related to the age of the specimen by

$$\text{Age} = \text{Natural TL}/[(\text{TL per unit dose}) \times (\text{natural dose rate})]. \qquad (15)$$

Thus, age assessment consists of measurement of the natural TL, calibrating the TL signal from the material to determine the TL/unit dose, and measurement of the natural dose rate in the location of the find. Strictly, the "age" being determined is the time since the TL signal was last reset to zero. Thus a "zeroing" event—such as high-temperature heating (e.g., in the manufacture of pottery) or optical bleaching (e.g., during the deposition of sediments)—must have occurred in order for the method to be applicable. Otherwise, a TL signal related to the geological age of the component mineral will be determined instead.

In principle, a typical experiment would proceed by determining the natural dose (or "equivalent dose") by either

1. removing the natural TL signal (i.e., by the act of TL measurement itself) and subsequently calibrating the sample by exposing it to known doses of radiation in the laboratory from which the natural dose can be estimated (the "regeneration method"), or
2. adding a laboratory dose on top of the natural dose and using an addition method, analogous to many chemical analysis techniques; here, by using several equivalent samples, a plot of TL versus added dose can be constructed from which the natural dose may be estimated by extrapolation to zero TL (the "additive dose method").

Figure 8 illustrates both procedures. Since the act of reading the TL signal often alters the sensitivity of the material to subsequent radiation doses, the additive dose method is often preferred. Note that the calibration doses administered in the laboratory are usually from gamma or beta sources. Thus, since natural irradiation consists of gamma,

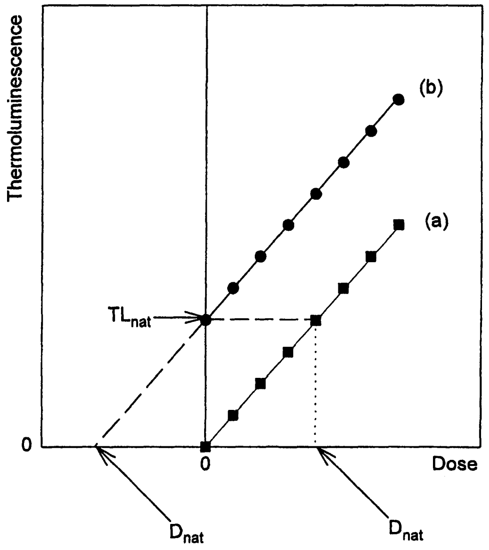

FIG. 8. Schematic illustration of the (a) "regeneration" and (b) "additive dose" methods of TL dating. T_{nat}, natural TL intensity; D_{nat}, estimated natural dose.

beta, and alpha irradiation, one has to account for the different TL production efficiencies of gamma and beta radiation, compared with alpha radiation, and determine the gamma, beta, and alpha equivalent doses separately. Alternatively, one can select large grains for which the alpha TL component is only found in the surface regions. Removal of these regions by acid etching leaves a sample from which the TL signal is due to cosmic, gamma, and beta doses exclusively.

Once the equivalent dose has been determined, one has to estimate the rate at which the sample has been irradiated in antiquity. Perhaps the most straightforward way to do this is to emplace conventional synthetic TL dosimeters in the location of the find. By leaving the dosimeter *in situ* for a number of weeks or months, one can then determine the natural absorbed dose integrated over this period, and from this, establish the dose rate.

3.2.2 Pottery Dating Pottery consists of a baked clay matrix containing small (up to a millimeter diameter) inclusions of quartz, feldspar, and minor amounts of other minerals. The discovery of TL from ancient pottery (Grögler *et al.*, 1960) led the way to the dating of pottery using the relationship in Eq. (15). Since these early observations, several methodologies have been suggested to arrive at a date for the pottery manufacture, and these can be found described in detail in several texts (e.g., Aitken, 1974, 1985).

One such method is the "inclusion method," wherein large quartz grains (~100 μm) are selected and, by etching away the alpha-irradiated surface, the age of the grains since the firing of the pottery (the "zeroing" event) can be determined from an assessment of the absorbed dose due to beta, gamma, and cosmic ray irradiation. Alternatively, in the "fine-grain method" examination of the fine grains (<10 μm) leads to an age estimation by accounting for the accumulated alpha dose, and the sensitivity of TL production of alphas compared to the other radiation types.

Each of the above methods relies on the accumulation of trapped charge as a function of irradiation time leading to the growth of the TL signal with age. A procedure that utilizes an entirely different principle is the so-called "predose method" in which the sensitivity of the sample (i.e., the propensity of the material to yield a TL signal for a fixed absorbed dose) is related to the irradiation age of the specimen. Here one monitors the TL for a small "test" dose, before subjecting the material to elevated temperatures, whereupon administration of a second test dose is found to yield a larger TL signal. The increase in the TL signal for the same test dose (i.e., the increase in the sensitivity) is related to the prior dose absorbed in nature and is activated by the heating of the sample between test doses. Calibration of the sensitivity increase provides a method for age determination.

Although many procedural difficulties are encountered, TL dating has proven to be a useful and versatile dating procedure for non-carbon-bearing artifacts and provides age estimates that overlap the more well-known ^{14}C-dating method.

3.2.3 Sediment Dating The principle of sediment dating—*vis-à-vis* determination of the dose absorbed by grains of quartz or feldspar—is the same as that described above for pottery dating. The difference lies in the mechanism presumed to be responsible for the initial zeroing of the TL signal. Whereas with pottery dating this event is the heating of the specimen during firing of the pot, with sediment dating the event is presumed to be the optical emptying of the traps by exposure to sunlight during deposition of the sediment. During deposition, be it in air or under water, the sediments are exposed to varying intensities of sunlight. Experimental tests have verified that exposure to light in this fashion empties a subset of the charge traps responsible for part of the TL signal and therefore resets these TL levels to zero. Subsequent burial of the sediment by additional layers shields the material from further light exposure and allows the buildup of the TL signal to occur as a result of the absorption of radiation from natural sources. Thus, when the material is excavated some time later, the TL signal is proportional to the burial time—i.e., the age of the sediment.

The technique was first demonstrated for deep-sea sedimentary cores wherein the TL level was found to increase with depth, corresponding to older deposits (Wintle and Huntley, 1979). More recent developments of sedimentary dating include the use of light to stimulate the luminescence signal, rather than heat. Thus, instead of TL, one monitors optically stimulated luminescence (OSL). This development has come to be called "optical dating" (Huntley *et al.*, 1985).

3.3 Other Applications

Age determination and radiation dosimetry are the two most extensive practical applications of TL. However, one must not overlook the use of TL in solid-state physics, where it is used as a tool for detecting the presence of defects and for establishing such parameters as the trap depths and capture cross sections, along with information regarding the dynamics of the various charge recombination kinetics.

Outside of these applications TL has found use in both terrestrial and extraterrestrial geology for mineral identification and for determining the classification and the irradiation and thermal histories of meteorites (Sears, 1993), source identification for various minerals, radioactive ore, and oil and gas well prospecting, and other, more esoteric uses. Descriptions of these and other applications can be found in several texts (e.g., McDougall, 1968; McKeever, 1985).

ACRONYMS

CCD: Charge-coupled device.

cw: Continuous wave (for lasers).

dc: Direct current.

DLTS: Deep-level transient spectroscopy.

DSC: Differential scanning calorimetry.

HCP: Heavy charged particle.

ITC: Ionic thermocurrents.

LET: Linear energy transfer.

OSL: Optically stimulated luminescence.

PMT: Photomultiplier tube.

QE: Quasiequilibrium.

TG: Thermogravimetry.

TL: Thermoluminescence.

TLD: Thermoluminescence dosimeter.

TSC: Thermally stimulated conductivity.

TSCap: Thermally stimulated capacitance.

TSDC: Thermally stimulated depolarization currents.

TSPC: Thermally stimulated polarization currents.

GLOSSARY

Dose (unit – Gray, Gy): The amount of energy absorbed by the target per unit mass of target. 1 Gy = 1 J/kg.

Effective Atomic Number, Z_{eff}: The effective atomic number of a compound is calculated from the atomic number of the constituent elements and the law of mixtures for the compound.

Fluorescence: *Luminescence* characterized by a decay over short-time intervals after removal of the excitation. Prompt luminescence due to electron de-excitation from an excited state to a ground state.

Luminescence: A light emission process that involves transitions between electronic energy states.

Optically Stimulated Luminescence: *Luminescence* emission stimulated by illuminating a previously irradiated sample. Optical energy excites charge from traps, initiates recombination, and causes *luminescence.* Unlike *photoluminescence,* the light only stimulates the recombination processes. The initial energy source is the radiation.

Phosphorescence: *Luminescence* characterized by a decay over long times after removal of the excitation. The distinction between *fluorescence* and *phosphorescence* based only on the characteristic time response is ill-defined, however, and somewhat arbitrary. If the process involves the trapping of charge in metastable states before thermal stimulation to the ground state, the process is clearly phosphorescence.

Photoluminescence: *Luminescence* excited by the absorption of energy from light.

Radiation Dosimetry: The measurement of absorbed radiation *dose.*

Thermoluminescence: *Luminescence* emission stimulated by heating a previously irradiated sample. Thermal energy excites charge from traps, initiates recombination, and causes luminescence. The heat only stimulates the recombination processes. The initial energy source is the radiation.

Works Cited

Aitken, M. J. (1974), *Physics and Archaeology,* Oxford, U.K.: Oxford Univ. Press.

Aitken, M. J. (1985), *Thermoluminescence Dating,* London: Academic.

Bos, A. J. J., Vijverberg, R. N. M., Piters, T. M., McKeever, S. W. S. (1992), *J. Phys. D: Appl. Phys.* **25,** 1249–1257.

Boyle, R. (1664), *Experiments and Considerations Touching Colours,* London: The Royal Society.

Chen, R. (1969), *J. Electrochem. Soc.* **116,** 1254–1258.

Chen, R., Kirsh, Y. (1981), *Analysis of Thermally Stimulated Processes,* Oxford, U.K.: Pergamon Press.

Garlick, G. F. J., Gibson, A. F. (1948), *Proc. Phys. Soc. London A* **60,** 574–590.

Grögler, N., Houtermans, F. G., Stäuffer, H. (1960), *Helv. Phys. Acta* **33,** 595–601.

Huntley, D. J., Godfrey-Smith, D. I., Thewalt, M. L. W. (1985), *Nature* **313,** 105–107.

Lewandowski, A. C., Markey, B. G., McKeever, S. W. S. (1994), *Phys. Rev. B* **49,** 8029–8047.

McDougall, D. J. (Ed.) (1968), *Thermoluminescence of Geological Materials,* London and New York: Academic.

McKeever, S. W. S. (1985), *Thermoluminescence of Solids,* Cambridge, U.K.: Cambridge Univ. Press.

McKeever, S. W. S., Moscovitch, M., Townsend, P. D. (1995), *Thermoluminescence Materials: Properties and Uses,* Ashford, U.K.: Nuclear Technology Publishing.

Randall, J. T., Wilkins, M. H. F. (1945), *Proc. R. Soc. London A* **184,** 366–389.

Sears, D. W. G. (1993), in: D. R. Vij, (Ed.), *Thermoluminescence Materials,* Englewood Cliffs, NJ: Prentice Hall.

Townsend, P. D., Kirsh, Y. (1989), *Contemp. Phys.* **30,** 337–354.

Wiedemann, E., Schmidt, G. C. (1895), *Ann. Phys. Chem. Neue Folge* **54,** 604–611.

Wintle, A. G., Huntley, D. J. (1979), *Nature* **279,** 710–712.

Further Reading

Aitken, M. J. (1974), *Physics and Archaeology,* Oxford, U.K.: Oxford Univ. Press.

Aitken, M. J. (1985), *Thermoluminescence Dating,* London: Academic.

Becker, K. (1973), *Solid State Dosimetry,* Boca Raton, FL: CRC Press.

Bräunlich, P. (Ed.) (1979), *Thermally Stimulated Relaxation in Solids,* Topics in Applied Physics Vol. 37, Berlin: Springer-Verlag.

Cameron, J. R., Suntharalingham, N., Kenney, G. N. (1968), *Thermoluminescence Dosimetry,* Madison, WI: Univ. of Wisconsin Press.

Chen, R., Kirsh, Y. (1981), *Analysis of Thermally Stimulated Processes,* Oxford, U.K.: Pergamon Press.

Horowitz, Y. S. (Ed.) (1983), *Thermoluminescence and Thermoluminescent Dosimetry,* Vols. 1–3, Boca Raton, FL: CRC Press.

McDougall, D. J. (Ed.) (1968), *Thermoluminescence of Geological Materials,* London and New York: Academic.

McKeever, S. W. S. (1985), *Thermoluminescence of Solids,* Cambridge, U.K.: Cambridge Univ. Press.

McKeever, S. W. S., Moscovitch, M., Townsend, P. D. (1995), *Thermoluminescence Materials: Properties and Uses,* Ashford, U.K.: Nuclear Technology Publishing.

Oberhofer, M., Scharmann, A. (1981), *Applied Thermoluminescence Dosimetry,* Bristol, U.K.: Adam Hilger.

THIN FILMS, ELECTRICAL PROPERTIES

Noriyoshi Yamauchi, *NTT Laboratories, Shinjuku, Tokyo, Japan*

Introduction 373
1. **Preparation of Thin Films** 373
2. **Measurement Techniques of Thin-Film Electrical Properties** 374
2.1 Resistivity Measurement 374
2.2 Hall-Mobility Measurement 375
3. **Electrical Properties of Thin Films** 376
3.1 Conductor Thin Films 376
3.1.1 Islandlike Thin Films 376
3.1.2 Continuous Thin Films 377
3.1.3 Effects of Defects and Contamination 378
3.2 Semiconductor Thin Films 378
3.2.1 Amorphous Semiconductor Thin Films 378
3.2.2 Polycrystalline Semiconductor Thin Films 379
3.3 Insulator Thin Films 380
4. **Applications of Thin-Film Electrical Properties** 382
4.1 Very Large-Scale Integration (VLSI) 383
4.2 Thin-Film Transistors 383
5. **Conclusion** 385
Glossary 385
Works Cited 385
Further Reading 386

INTRODUCTION

A thin film is defined as a solid-state material, deposited on a substrate by a controllable method with uniform thickness of less than 10 μm or so. Numerous materials including almost all the elements can be prepared in the form of a thin film. The crystallographic status of the thin films ranges from amorphous through polycrystalline to crystalline depending on the thin-film materials, substrates, and deposition conditions. The electrical conductivity of thin films ranges from insulating through semiconducting to conducting and superconducting depending on the materials as well as the structure of the thin film.

This article reviews the electrical properties of thin films, focusing on the basic principles of conductor, semiconductor, and insulator thin films. Superconductors are discussed in SUPERCONDUCTING DEVICES; SUPERCONDUCTIVITY, HIGH-TEMPERATURE; and SUPERCONDUCTIVITY, LOW-TEMPERATURE. We discuss amorphous and polycrystalline thin films in the article. Crystalline thin films are reviewed in CRYSTAL GROWTH.

Thin films are modeled as a medium confined between a pair of parallel surfaces. The electrical properties of conductor and semiconductor thin films are greatly affected by the existence of the surfaces and the defects. We discuss the effects of surfaces and defects in conductor thin films, amorphous semiconductor thin films, and polycrystalline semiconductor thin films. In insulator thin films, carrier transport in the direction perpendicular to the surface occurs under sufficiently high electric field. Current mechanisms in insulator thin films are also discussed.

Thin films are the essential components in many industrial applications. Typical examples are microdevices including very large-scale integration (VLSI). Driven by the strong demands, research and development have been conducted actively in industry and academia. In this article, we discuss the applications of thin films taking VLSI and thin-film transistors (TFT) as examples.

1. PREPARATION OF THIN FILMS

Various technologies have been developed to deposit high-quality thin films. There are three basic types of deposition methods: evaporation, sputtering, and chemical vapor

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

deposition (CVD). Each method is reviewed in detail in the articles CVD (CHEMICAL VAPOR DEPOSITION); SPUTTERING; and THIN-FILM DEPOSITION. There are two kinds of CVD methods; one is thermal CVD, which uses thermal decomposition of material gases, and the other is plasma CVD, which uses plasma to enhance the deposition rate. Plasma CVD allows thin-film deposition at lower temperatures than thermal CVD. This category includes the glow discharge method in which material gas is decomposed by discharging.

Another way to form thin films is thermal oxidation or nitridation of silicon. Silicon substrate can be oxidized or nitrided at high temperatures, typically 900 °C or higher. Chemical reactions of silicon with the ambient gases O_2 or N_2 occur at the surface to form thin films of SiO_2 or Si_3N_4. Thermally grown SiO_2 thin films are used as gate oxides of metal-oxide-semiconductor (MOS) transistors. They are also used as local-oxidation-of-silicon (LOCOS) thin films to isolate transistors in MOS devices. Thermally grown Si_3N_4 thin films are used in nonvolatile memory devices. In general, thermal oxidation and nitridation provide not only low-resistivity insulator thin films but also much lower trap-state density at the interface with the underlying silicon. This is the reason that this method is used for gate-insulator formation in MOS devices, in which insulator quality is crucial to achieve high device performance.

The sputtering method covers the widest range of thin-film materials among the three deposition methods mentioned above, from metals through semiconductors to insulators. Use of the evaporation method is mostly limited to metals with relatively low melting points. The CVD method requires appropriate material gases, and this limits the thin-film materials that can be deposited by this method. The electrical properties of thin films depend on deposition methods as well as the materials and deposition conditions. Thus it is not easy to generalize the relations between electronic properties and deposition methods. The most appropriate deposition method is chosen after examining experimental results in many cases. Throughput in the deposition method is also a factor to be considered in manufacturing.

Thin films are often processed after the deposition to improve or change their properties. A typical process is annealing, including furnace annealing and laser annealing. Defects in thin films can be reduced by an annealing process. By an annealing process with proper conditions, amorphous thin films can be transformed to polycrystalline thin films. In polycrystalline thin films, defects can be drastically reduced by an annealing process, which results in a lower resistivity in conductor and semiconductor thin films.

2. MEASUREMENT TECHNIQUES OF THIN-FILM ELECTRICAL PROPERTIES

Carrier concentration and carrier mobility are two basic properties of conductor and semiconductor thin films. In this section we discuss the measurement methods for resistivity and carrier mobility. Carrier concentration is derived from mobility and resistivity.

2.1 Resistivity Measurement

Here we define a thin-film pattern delineated on an insulating substrate with a length L, a width W, and a thickness d as shown in Fig. 1. When a voltage V is applied between a and b in Fig. 1, the current I that flows along the direction of L is

$$I = V/R \tag{1}$$

according to Ohm's law. R in Eq. (1) is the resistance we define.

The resistance R of the pattern in Fig. 1 along L is proportional to L and inversely proportional to the cross section of the pat-

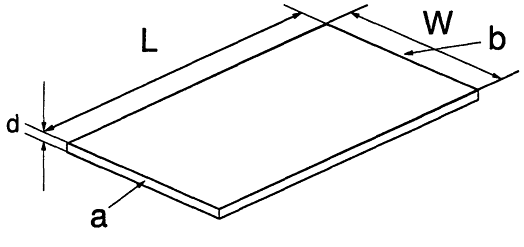

FIG. 1. Schematic view of a thin-film pattern with length *L*, width *W*, and thickness *d*.

tern, i.e., Wd. Therefore, R is expressed using the formula.

$$R = \rho L/Wd. \tag{2}$$

The proportionality constant ρ is defined as the resistivity. As we discuss later, ρ is dependent on d when the thickness is relatively thin, typically thinner than 200 nm. We also define the conductivity σ as

$$\sigma = 1/\rho. \tag{3}$$

The four-probe method is effective in order to measure the resistance accurately. A measurement structure for this method is schematically shown in Fig. 2. When a voltage V_{ad} is applied between point a and point d through the probe electrodes, the current I between point a and point d can be measured by using an ampere meter. The voltage difference between point b and point c, V_{bc}, on the other hand, is measured using a voltmeter. In the structure shown in Fig. 2, there exists resistance called contact resistance at the interface between the thin-film material and the electrode material due to the work-function difference and impurities. Here we denote such a resistance as $R_{contact}$.

The relations between R_{ad}, R_{bc}, $R_{contact}$, V_{ad}, V_{bc}, and I are expressed by $V_{ad} = (R_{ad} + 2R_{contact})I$, $V_{bc} = (R_{bc} + 2R_{contact})I$, and $R_{ad}/R_{bc} = L_{ad}/L_{bc}$. From these relations, R_{ad} and $R_{contact}$ are obtained independently as follows:

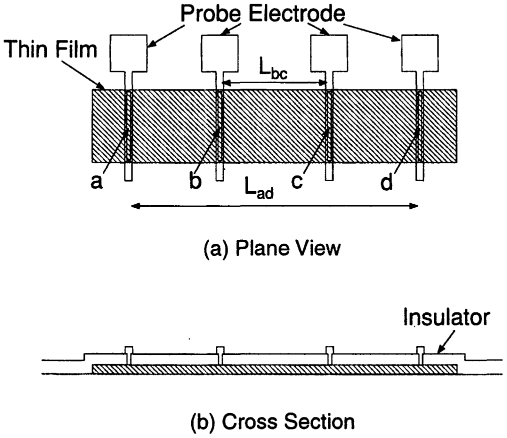

(a) Plane View

(b) Cross Section

FIG. 2. A structure for four-probe resistance measurement.

$$R_{ad} = (V_{ad} - V_{bc})L_{ad}/(L_{ad} - L_{bc})I, \tag{4}$$

$$R_{contact} = (V_{bc}L_{ad} - V_{ad}L_{bc})/2(L_{ad} - L_{bc})I. \tag{5}$$

The resistivity of the thin film is obtained by normalizing R_{ad} with the length, the width, and the thickness using Eq. (2).

Another convenient method to measure the resistance of thin films is to form a Wheatstone bridge circuit as shown in Fig. 3 (see ELECTRICAL INSTRUMENTATION). The bridge circuit has four resistors. Two of them are fixed resistors with known resistance (R_1 and R_2 in Fig. 3). The other two resistors are a variable resistor R_3 and the thin-film specimen (R_x) of which the length, the width, and the thickness are known. A voltage is applied between a and b in Fig. 3 and the current between c and d is measured. In the measurement, R_3 is varied to make the current between c and d zero. When the current is zero, the resistance of the thin-film specimen is determined as

$$R_x = R_3(R_2/R_1). \tag{6}$$

The resistivity of the thin film is derived from R_x, the length, the width, and the thickness using Eq. (2).

2.2 Hall-Mobility Measurement

By making use of the Hall effect, carrier concentrations of the conductor and semiconductor thin films can be measured (see MICROSENSORS). Figure 4 shows a schematic configuration of the Hall effect. When a

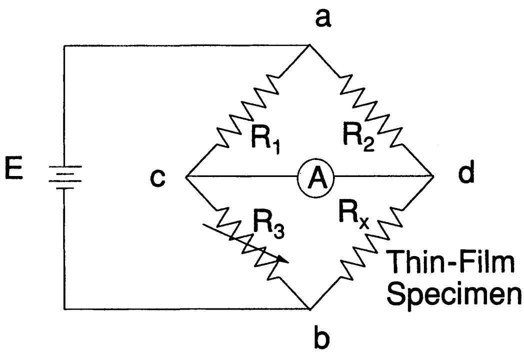

FIG. 3. Wheatstone bridge circuit for resistance measurement.

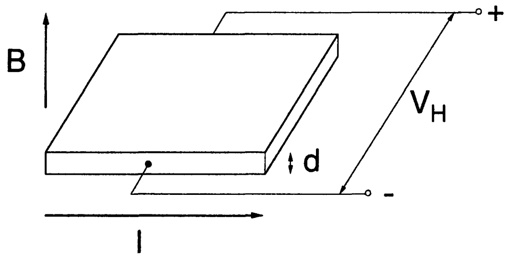

FIG. 4. Schematic representation of the Hall effect.

magnetic field B is applied perpendicularly to a thin film in which the current I is flowing, a voltage V_H is produced perpendicular to the directions of both the current I and the field B. The voltage V_H is proportional to the current I and the magnetic field B and inversely proportional to the thickness d as follows:

$$V_H = R_H IB/d, \tag{7}$$

where R_H is called Hall coefficient. R_H is negative when the carriers in the thin film are electrons and positive when the carriers are holes.

For an n-type semiconductor, the Hall coefficient is written as

$$R_H = -1/n_e q, \tag{8}$$

where n_e is the electron concentration and q is the electron charge. For a p-type semiconductor, R_H is written as

$$R_H = 1/n_h q, \tag{9}$$

where n_h is the hole concentration. The carrier concentration n (either n_e or n_h) is derived from Eqs. (8) and (9) as

$$n = 1/q|R_H|. \tag{10}$$

The conductivity of the thin film σ is related to the Hall mobility μ_H as

$$\mu_H = \sigma|R_H| = |R_H|/\rho. \tag{11}$$

From this relation, the Hall mobility is derived from the Hall coefficient and resistivity.

The Hall mobility reflects the conductive properties of the bulk of the thin films. This is in contrast to the field-effect mobility, which reflects the surface or interface properties of thin films.

3. ELECTRICAL PROPERTIES OF THIN FILMS

In this section, we discuss electrical properties of conductor, semiconductor, and insulator thin films. Conductivity in metals and alloys is discussed in detail in METALS AND ALLOYS, CONDUCTIVITY IN. Here, we focus on electrical properties that are unique in thin films.

Thin films are modeled as a medium with numerous defects confined between a pair of parallel surfaces. The unique properties of thin films that are different from those of bulk materials originate from the effects of surfaces and defects in conductor and semiconductor thin films.

3.1 Conductor Thin Films

3.1.1 Islandlike Thin Films At an early stage of conductor thin-film deposition, typically when the thickness is less than a few nanometers or so, thin films consist of islandlike structures separated from one another. In such thin films, it has been experimentally known that the resistivity is much higher than in bulk, of the order of 10^{10} to 10^{13} Ω cm. It has also been known that the thermal coefficient is negative and the thermal-activation type conductivity is observed, as given by

$$\sigma = \sigma_0 \exp(-E_a/kT), \tag{12}$$

where E_a is the activation energy, k is the Boltzmann constant, and T is the temperature. The activation energy E_a is of the order of −0.1 eV or less. It decreases as the thickness increases. Figure 5 shows experimental results on the temperature dependence of gold conductivity (Kazmerski and Rachine, 1975).

Although several theoretical models have been proposed to analyze the electrical conductivity of the islandlike conductor thin films, it has not been possible yet to explain every aspect of the observed phenomena. Possible electron transport mechanisms are thermal emission and tunneling. The path of

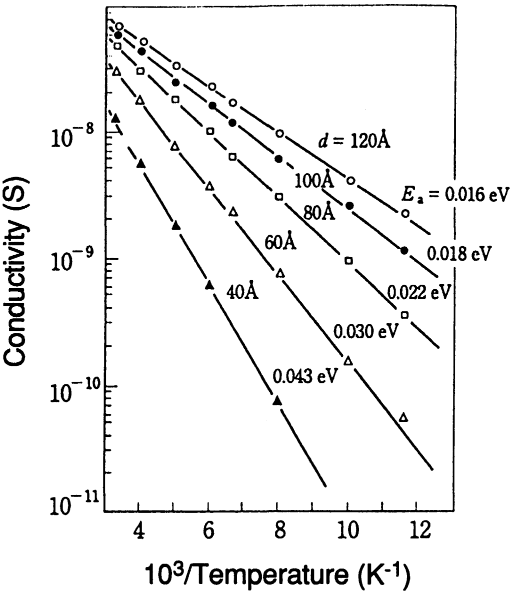

FIG. 5. Temperature dependence of conductivities in gold thin films.

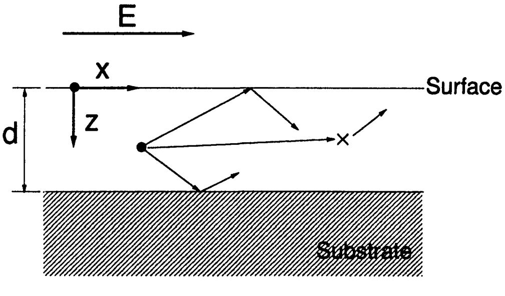

FIG. 6. Schematic electron scattering in thin film.

the electron transport can be free space (vacuum) and/or the surface of the substrate.

3.1.2 Continuous Thin Films When the thin-film thickness is thicker than a few nanometers or so, the film becomes continuous. Even in continuous thin films, the properties are different from those of the bulk materials. It has been known that conductivity, Hall coefficient, and thermal coefficient depend on the film thickness. This has been attributed to the effects of surfaces. It has also been known that the above characteristics still differ from those of bulk materials even when their change with the film thickness saturates as the thickness is increased. This is obviously due to the structural defects incorporated in the film during deposition.

The surface effects are prominent when the film thickness is comparable with or smaller than the mean free path of electrons. Figure 6 shows a schematic of the electron transport mechanism in a conductor thin film. Fuchs (1938) and Sondheimer (1950) formulated the effects of scattering at surfaces on the resistivity of conductor thin films. Electron transport characteristics can be derived by solving the Boltzmann equation for the conduction-electron distribution function $f(r,v)$,

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\frac{\partial f}{\partial \mathbf{r}} + \mathbf{F}\cdot\frac{\partial f}{\partial \mathbf{v}} = \left(\frac{\partial f}{\partial t}\right)_{\text{collision}}, \tag{13}$$

with appropriate boundary conditions. In Eq. (13), F represents the external field, and the right-hand term represents the change due to relaxation. When we assume that the distribution f_0 changes to $f_0 + f_1$ and the relaxation time is t, Eq. (13) becomes

$$v_z\frac{\partial f_1}{\partial z} + \frac{f_1}{2} = \frac{eE}{m_0}\frac{\partial f_0}{\partial v_x}, \tag{14}$$

and the boundary condition is

$$f_1(z = 0) = f_1(z = d) = 0. \tag{15}$$

We define here the probability of an electron being elastically scattered at the surface as p $(0 < p < 1)$. By solving Eq. (15), the ratio of a thin-film resistivity ρ to the resistivity of a virtual thin film with infinite thickness ρ_{inf} can be expressed as follows (Kimbara, 1987):

$$\rho/\rho_{\text{inf}} = 1 + \tfrac{3}{8}(1 - p)l_{\text{inf}}/d, \tag{16}$$

where l_{inf} is the mean free path of electrons in a thin film with infinite thickness and d is the thin-film thickness. If $p = 1$, the ratio ρ/ρ_{inf} is 1, and there is no surface effect.

Figure 7 shows the relation between ρ/ρ_{inf} and d/l_{inf} obtained in Fuchs–Sondheimer theory (Kamiyama, 1983). It shows that the normalized resistivity increases drastically as the normalized thickness d/l_{inf} becomes smaller than 1, except in the case in which $p = 1$.

It has been experimentally shown that p ranges from 0.5 to 1 in epitaxial and single-

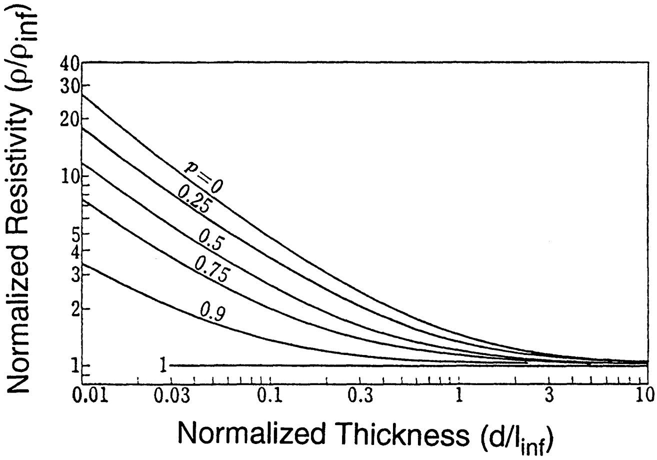

FIG. 7. Plots of normalized resistivity as a function of normalized thin-film thickness as the results of Fuchs–Sondheimer theory.

crystalline thin films. In polycrystalline thin films, in general, p is close to zero.

3.1.3 Effects of Defects and Contamination If a material consists of an infinite and perfect lattice, there is no obstacle to electron passage. The deviation from that perfect situation causes the scattering of electrons (see METALS AND ALLOYS, CONDUCTIVITY IN). The existence of islandlike structures in thin films of less than a few nanometers thickness and the surfaces in continuous thin films are examples of the deviations that cause electron scattering.

Defects in thin films are also major factors that cause electron scattering. Thin films contain various kinds of defects such as point defects, dislocations, stacking faults, and grain boundaries. They are incorporated in the films at deposition. Among these defects, the point defects are known to have the greatest effects on resistivity in conductor thin films. Typical point-defect concentration is 0.1 to a few at.% in metal thin films. The increase in resistivity in metals is typically in the range of (1 to 4) $\times 10^{-6}$ $\Omega \cdot$cm per 1% increase of the defects (see METALS AND ALLOYS, CONDUCTIVITY IN).

In case of polycrystalline thin films, grain boundaries work as a kind of defect, as we discuss later.

In addition to defects, impurities in thin films cause an increase in resistivity. Some of the impurities are incorporated in thin films during deposition. Other impurities enter thin films after deposition, by adsorption. Oxygen adsorbs on metal thin films easily. In most metals, with the exception of gold, oxygen adsorption and subsequent oxidation of the surface increase the resistivity of thin films.

The increase in resistivity due to oxidation is quite harmful in insuring reliability of the thin-film electrical properties. To prevent this, a passivation technique can be used. In this technique, a stable insulator thin film such as SiO_2 and Si_3N_4 is deposited on top of thin films that are vulnerable to exposure to the air.

3.2 Semiconductor Thin Films

3.2.1 Amorphous Semiconductor Thin Films Amorphous materials lack translational symmetry in atomic configuration. This means that there is no long-range (ten atomic spacings or so) order in amorphous materials (see AMORPHOUS MATERIALS, STRUCTURE OF). It should be noted, however, that amorphous materials do maintain relatively good short-range order. Ioffe and Regel (1960) predicted that this short-range order results in a band gap in amorphous materials even when there is no long-range order. The prediction was later proved by experiments.

Amorphous semiconductors can be categorized roughly into chalcogenide semiconductors and tetrahedrally bonded semiconductors.

Chalcogenide semiconductors are materials with chemical formulas containing one or more atoms of S, Se, and/or Te. The atoms in chalcogenide semiconductors form one-dimensional or two-dimensional networks. These networks are attracted to each other by weak Van der Waals force. There are many kinds of chalcogenide semiconductors, including *a*-Se, *a*-As_2S_3, and *a*-As-S-Te-Ge. (Here, *a*- denotes amorphous.)

The electrical properties of amorphous chalcogenide thin films show some similarities to crystalline semiconductors in that the conduction is by electrons and is activation type. The conductivity of chalcogenide thin films can be expressed as

$$\sigma = \sigma_0 \exp(-E_a/kT). \tag{17}$$

In most chalcogenide amorphous semiconductors, the conduction type is weak *p*-type. They are not sensitive to the type of dopants. Thus it is not usually possible to control conduction type, *p*-type or *n*-type, by doping impurities (Kamiyama, 1983).

Another type of amorphous semiconductor thin films is tetrahedrally bonded thin film. Typical materials of this type are *a*-Si and *a*-Ge. (See SILICON, AMORPHOUS.)

a-Si thin films can be deposited by evaporation and sputtering. In such *a*-Si thin films, electronic conduction occurs through long-range hopping near the Fermi level. The conductivity is expressed as follows (Mott and Davis, 1979):

$$\sigma = \sigma_0 \exp\left\{-B\left(\frac{a^3}{kTN(E_F)}\right)^{1/4}\right\}, \tag{18}$$

where k is the Boltzmann constant, $N(E_F)$ is the electron state density at the Fermi energy E_F, and a and B are constants.

It is possible to incorporate hydrogen atoms into *a*-Si thin films by using the glow-discharge or plasma-CVD method of deposition. Here we denote hydrogenated amorphous silicon thin films as *a*-Si:H.

In *a*-Si:H, the energy state density in the band gap is significantly decreased because hydrogen atoms fill most of the dangling bonds. Thus the conductivity shows an activation type as follows:

$$\sigma = \sigma_0 \exp(-E_a/kT), \tag{19}$$

where E_a is the energy difference between E_F and the conduction-band edge energy E_c or valence-band edge energy E_v.

In *a*-Si:H thin films, *p*-type or *n*-type dopants can be incorporated during deposition. Spear and Le Comber (1975) demonstrated that phosphorus and boron can be doped into *a*-Si:H thin films using phosphine (PH_3) and diborane (B_2H_6) in glow-discharge deposition. It is well known that this discovery triggered the rush in basic and applied research in *a*-Si:H thin films.

Figure 8 shows the effect of phosphorus and boron doping on the conductivity and Fermi-level position. The conductivity changes by up to ten orders of magnitude depending on the impurity concentration. The change in Fermi level position ranges up to 1.4 eV.

With the control of *n*-type and *p*-type conduction as well as the wide range of change in conductivity, *a*-Si:H thin films have been pursued for many applications. Solar cells (see SOLAR ENERGY) and thin-film transistors are examples of very successful applications (Kanicki, 1993).

3.2.2 Polycrystalline Semiconductor Thin Films

A polycrystalline thin film consists of crystalline grains that vary in size and crystallographic orientation. In a crystalline grain, atoms configure a lattice like crystalline materials. Between adjacent grains, there exist grain boundaries where defects due to misorientation between different grains concentrate. These grain boundaries cause unique characteristics in polycrystalline semiconductor thin films. Figure 9 shows a schematic structure, charge distribution, and potential barrier height in polycrystalline semiconductor thin films (Seto, 1975).

Here, we take *p*-type polycrystalline silicon thin films as an example and analyze their electrical properties after Seto's theory (Seto, 1975). We discuss two cases, when $LN < N_t$ and when $LN > N_t$, where L is the grain size, N is the impurity concentration, and N_t is the trap state density at the grain boundary.

When LN is less than N_t, the *p*-type impurities in the grain are all ionized. A number of trap states equal to the number of ionized impurities are filled with positive charge at the grain boundary. The potential barrier V_B at the grain boundary in this case is expressed as

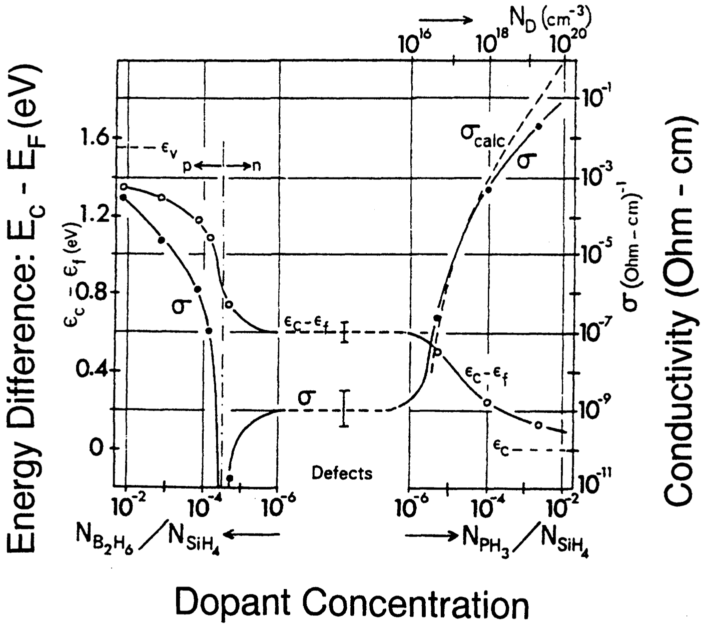

FIG. 8. Conductivity and energy difference ($E_c - E_F$) in *a*-Si:H thin films as functions of doping concentration.

$$V_B = qL^2N/8\epsilon, \tag{20}$$

where q is the electron charge and ϵ is the dielectric constant of silicon.

When LN is greater than N_t, the number of trapped carriers per unit area of grain boundaries remains constant at the value N_t. The potential-barrier height V_B is

$$V_B = qN_t^2/8\epsilon N. \tag{21}$$

From the analysis of thermionic emission at the grain boundaries (Kamins, 1988), the conductivity in polycrystalline silicon thin films is obtained as

$$\sigma = (q^2nv_cL/kT)\exp(-qV_B/kt), \tag{22}$$

where n is the free-carrier density and v_c is the collection velocity defined as $v_c = 2kT/\pi m$.

3.3 Insulator Thin Films

In insulator thin films, carrier transport in the direction perpendicular to the surface occurs when the electric field is sufficiently high. When the electric field is increased further, the thin film breaks down eventually.

In the discussion of carrier transport in insulators, we must first identify the origin of the carrier and the transport mechanism. The origin of the carrier is either from outside the insulator or from inside the insulator. We denote here the number of electrons originating in the insulator thin film as n and the number of electrons injected from outside the insulator thin film as n_e. When the electric field is relatively low, n_e is much lower than n. Therefore, the number of electrons in the thin film is approximated as n and the current is expressed as

$$J = qn\mu E. \tag{23}$$

In this case, the current is proportional to the electric field and the relation is Ohmic.

When the electric field is increased and n_e is not negligible, several carrier transport mechanisms may occur (Sze, 1981).

Schottky emission is a thermionic emission across the metal–insulator interface. The current J is expressed as a function of temperature T and the electric field E as

$$J \sim A^*T^2\exp\left[\frac{-q(\phi_B - \sqrt{qE/4\pi\epsilon i})}{kT}\right], \tag{24}$$

where A^* is the effective Richardson constant, ϕ_B is the barrier height between the insulator thin film and the metal electrode, E

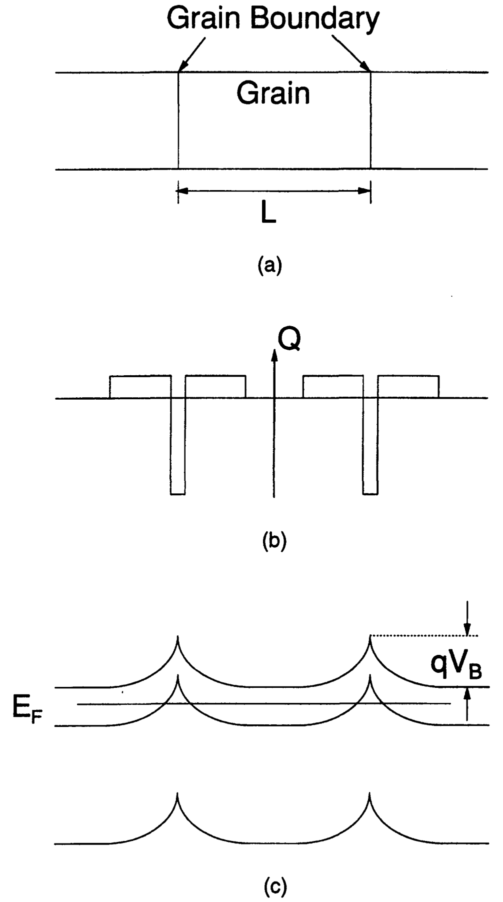

FIG. 9. (a) Schematic structure, (b) charge distribution, and (c) potential height in a polycrystalline semiconductor thin film.

is the electric field, and ϵ_i is the insulator permitivity.

Frenkel–Poole emission is due to electric-field–enhanced thermal excitation of trapped electrons into the conduction band. The current is expressed as

$$J \sim E \exp\left[\frac{-q(\phi_B - \sqrt{qE/\pi\epsilon_i})}{kT}\right]. \tag{25}$$

For trap states with Coulomb potentials, the expression is virtually identical to that of the Schottky emission.

Tunnel emission is caused by field ionization of trapped electrons into the conduction band or by electron tunneling from the metal's Fermi energy into the metal conduction band. The current is expressed in the form

$$J \sim E^2 \exp\left[-\frac{4\sqrt{2m^*}(q\phi_B)^{3/2}}{3q\hbar E}\right], \tag{26}$$

where m^* is the effective mass of the electron and $\hbar$ is the Dirac–Plank constant.

Space-charge-limited current results from carrier injection into the insulator thin film where no compensating charge is present. The current for the unipolar trap-free case is proportional to the square of the applied voltage.

For a given insulator, each conduction mechanism may dominate in a certain temperature and voltage range. However, two or more mechanisms may coexist, making the observed phenomena apparently complex.

Figure 10 shows a plot of current as a function of electric field in a Au-Si_3N_4-Si capacitor structure (Sze, 1981). The +(Au) and −(Au) show the curves with positive and negative voltages to the gold electrode, respectively. There are two distinct regions. Under high electric field, the current in-

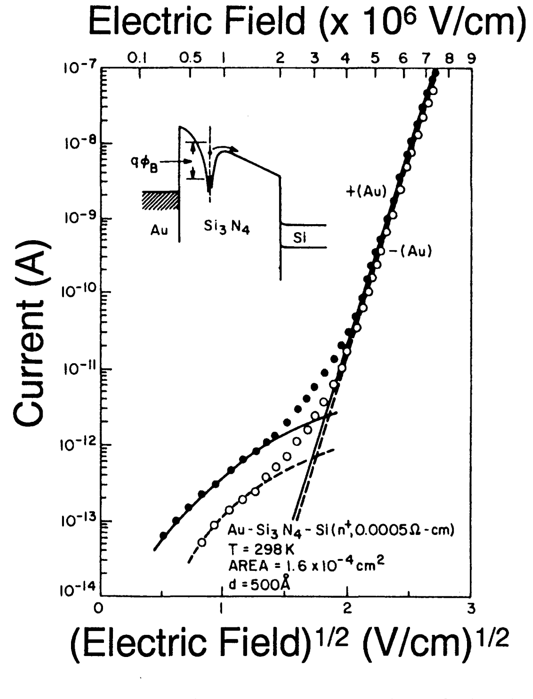

FIG. 10. A plot of current as a function of electric field in a Au-Si_3N_4-Si structure.

creases exponentially with the square root of the electric field, showing Frenkel–Poole type emission. At lower electric field, the characteristics are Ohmic. When the electric field is relatively low, the number of electrons in the insulator is much less than the number of electrons originating in the insulator. In this case, space charge does not affect the current, and the current is proportional to the electric field. This explains the Ohmic relation in Figure 10. At higher electric field, electrons are injected into the insulator, creating space charge.

The maximum dielectric strength, which is also called breakdown field, is an important property of insulator thin films. The essential breakdown processes in insulators are thermal breakdown and electronic breakdown. At relatively low temperatures, the breakdown is due to an electronic process in which the breakdown is directly caused by the electric field. The breakdown field is nearly independent of temperature in this case. At higher temperatures, thermal breakdown occurs. In this case, the breakdown is caused by the local heat in the thin film.

Figures 11(a) and 11(b) show the relation between the breakdown field and the insulator thickness in SiO_2 thin films in two different thickness ranges (Sze, 1981). When the thickness is above 100 nm the breakdown field is constant at 9×10^6 V/cm. It should be noted that the thinner the film, the higher the breakdown field. For very thin SiO_2 films, thinner than 10 nm, the breakdown field approaches 3×10^7 V/cm.

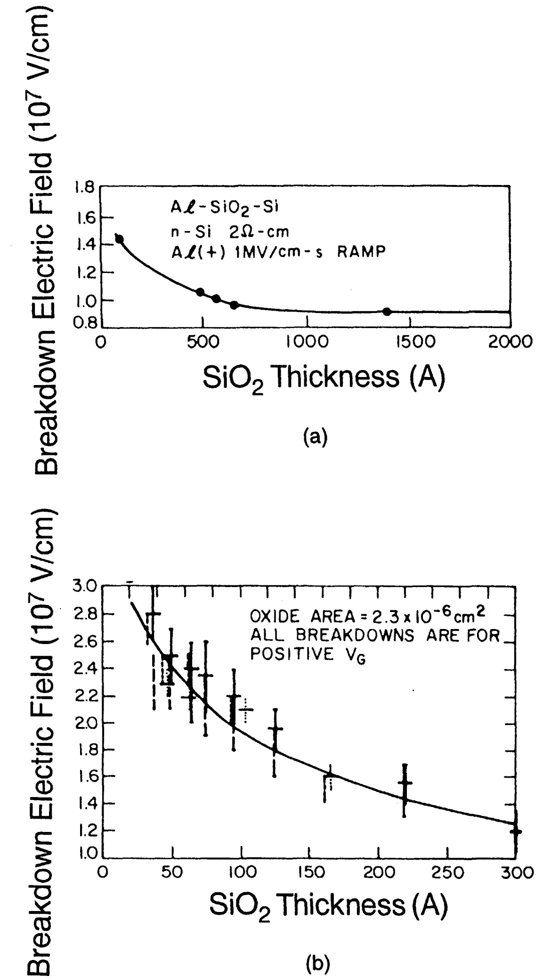

FIG. 11. Relation between breakdown electric field and insulator (SiO_2) thickness in an Al-SiO_2-Si structure.

4. APPLICATIONS OF THIN-FILM ELECTRICAL PROPERTIES

Thin films are used in various industrial products including microdevices such as very large-scale integrated circuits, thin-film transistors, and micromechanical devices to name a few. The major advantages of thin films for use in industrial products are as follows. First, they can be deposited on solid-state substrates of virtually all kinds with adequate control and throughput. Stacking one thin film on top of another is also easily achieved. Second, the control of thin films in terms of thickness and uniformity is excellent in general. This results in high reliability of products using thin films. Third, they can be patterned using microlithography technology (see MICROLITHOGRAPHY) with feature sizes down to 1 μm or less. Such fine patterning makes extremely high-density integration of active and passive devices possible.

In this section, we review the applications of thin films taking very large-scale integrated circuits and thin-film transistors as appropriate examples. In these applications, high-quality thin films of conductor, semiconductor, and insulator are fully utilized to achieve required performance with reliability (see VLSI—INTEGRATED CIRCUITS; MICROELECTRONICS).

4.1 VERY LARGE-SCALE INTEGRATION (VLSI)

The integration of active devices (transistors and diodes) and passive devices (capacitors and resistors) on semiconductor substrates, typically silicon and gallium arsenide, has achieved the revolution in electronics providing high-performance microprocessors (see MICROPROCESSORS) and high-capacity memory chips. In VLSIs, thin films are used as conductors to interconnect each device electrically and as insulators to isolate different levels of conductor patterns.

Figure 12 shows a schematic cross section of a typical CMOS dynamic random-access memory. Note that all the components, except the silicon substrate, are thin films.

Now we discuss the roles of the thin films following the fabrication process. A relatively thick ($\sim 0.5\ \mu$m) SiO_2 thin film is thermally grown first to isolate each MOS transistor. After n-type well formation using the ion implantation technique, a gate insulator SiO_2 thin film is grown by thermal oxidation. Very high dielectric strength ($>10^7$ V/cm) and high-quality interface between the SiO_2 and silicon substrate surface are achieved. A phosphorus-doped polycrystalline-silicon thin film is deposited by the CVD method and pattern delineated to form gate electrodes. This polycrystalline silicon thin film is highly doped with phosphorus ($\sim 10^{19}/cm^3$) so that it works as a conductor. Subsequently, p^+ (highly doped p-type) and n^+ (highly doped n-type) regions are formed using boron and phosphorus ion implantation. These regions will become the source and drain of each MOS transistor. To complete the integration, each device must be interconnected. For this purpose, a SiO_2 thin film is deposited by CVD or sputtering as an interlayer insulator. Phosphorus is usually doped into this SiO_2 thin film. The phosphorus captures alkaline ions in the device structure that may cause instability in MOS transistor characteristics. Next, via holes are opened to expose the source and drain regions as well as the doped polycrystalline silicon thin films. An Al thin film is then deposited by sputtering followed by pattern delineation to complete the device structure.

Since steps are formed as the fabrication process goes, step coverage of thin films is required. With optimized deposition, this requirement has been met in VLSI manufacturing.

4.2 Thin-Film Transistors

Thin film transistors (TFTs) are defined as field-effect transistors built using a semiconductor thin film that is originally deposited on an insulating substrate (Weimer, 1962). Various semiconductor materials have been studied for use in TFTs including CdS, CdSe, Te, InSb, PbTe, Se, Ge, SiGe, and Si (Yamauchi and Reif, 1994). Among these materials, only two types of TFTs that use silicon thin films, either amorphous or polycrystalline, have found practical applications.

Figure 13 shows cross sections of the three most common TFT structures: the coplanar type in Fig. 13(a), the staggered type in Fig. 13(b), and the inverted staggered

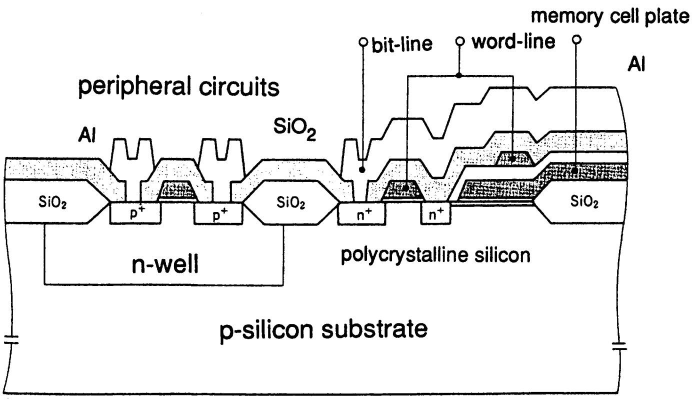

FIG. 12. Cross section of a typical CMOS dynamic random-access memory integrated circuit.

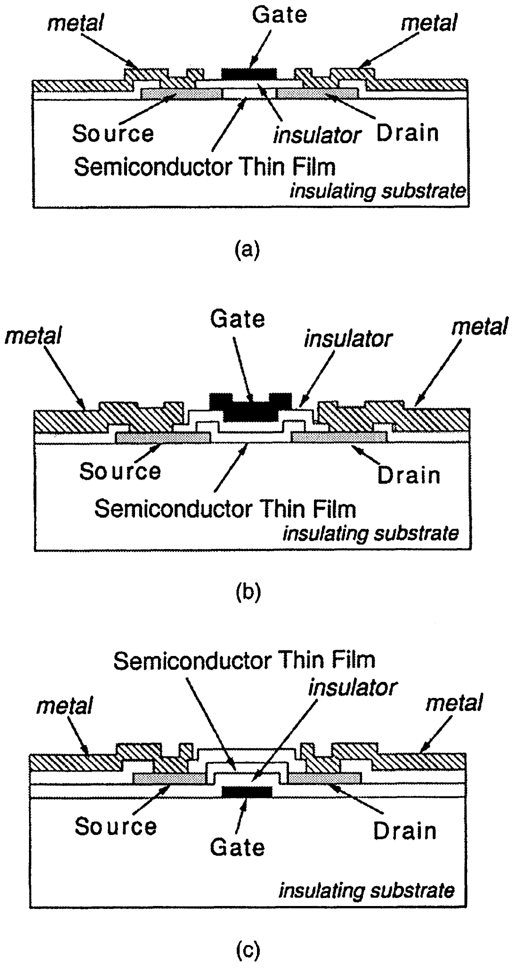

FIG. 13. Schematic cross section of thin-film transistors.

structure in Fig. 13(c). As shown in Fig. 13, all types of thin films, conductor, semiconductor, and insulator, are used in TFTs. Like a crystalline silicon MOS transistor, a TFT consists of source and drain regions that are doped with impurities, and a gate electrode. The gate bias, applied through the gate insulator, controls the current that runs between source and drain. Either *n*-channel or *p*-channel TFTs can be made by selecting *n*-type impurities (phosphorus or arsenic) or *p*-type impurities (boron), respectively. This makes CMOS circuits possible by using both types of TFTs.

The most important feature of the TFTs is that they are built on an insulating substrate, typically glass or fused quartz. This allows optical transparency of the device structure that is essential in most of their applications. Moreover, it allows use of much larger substrate sizes than those possible with crystalline semiconductor substrates. TFTs have been successfully used in active-matrix–addressed liquid-crystal displays, in which each pixel is addressed by a TFT (see DISPLAY TECHNOLOGY).

The electrical properties of semiconductor thin films and the semiconductor–gate insulator interface are the two main factors that determine TFT performance. Efforts have been made to achieve high-quality semiconductor thin films by optimizing the deposition conditions in both amorphous and polycrystalline silicon TFTs.

Field-effect mobilities achieved in amorphous-silicon TFTs are in the range of 0.5–1 $\mathrm{cm^2/V \cdot s}$. These mobilities are two to three orders of magnitude lower than that of crystalline silicon MOS transistors. However, such mobilities in amorphous silicon TFTs are sufficient for current active-matrix display applications.

In polycrystalline silicon TFTs, on the other hand, field-effect mobilities over 50 $\mathrm{cm^2/V \cdot s}$ have been achieved, enabling high-speed performance in TFT integrated circuits. In laser-annealed polycrystalline silicon TFTs, a mobility of 382 $\mathrm{cm^2/V \cdot s}$ has been reported (Serikawa *et al.*, 1989). This mobility is about half of the mobility in crystalline-silicon MOS transistors. Such high mobilities in polycrystalline silicon TFTs have been achieved mainly by reducing the trap state density at grain boundaries.

Levinson *et al.* (1982) formulated an expression for the drain current of a polycrystalline-silicon TFT using a carrier-transport model at the grain boundary first given by Seto (1925). The following equation shows the drain current I_D as a function of the gate voltage V_G and the drain-source voltage V_D:

$$I_D = \mu_{GB}(W/L)(C_{OX})V_G V_D \exp(-V_B/kT), \qquad (27)$$

where μ_{GB} is the carrier mobility inside the grain, W is the channel width, L is the channel length, C_{OX} is the gate capacitance, and V_B is the potential barrier height at the grain boundary [Levinson *et al.*, 1982, Eq. (21)].

Equation (27) shows that the drain current is determined by the mobility inside a grain and the grain-boundary trap density

N_t. This is a natural consequence of the premise that the grain boundaries are the dominant factor that distinguishes polycrystalline silicon TFTs from crystalline-silicon MOS transistors.

5. CONCLUSION

The electrical properties of thin films are greatly affected by the existence of the surfaces and defects. In conductor thin films, the carrier transport is affected by the scattering at the surfaces especially when the thickness is smaller than the mean free path of an electron. In amorphous-semiconductor thin films, the carrier transport is dominated by the defects and the energy-state distribution originating from the lack of long-range order. In hydrogenated amorphous-silicon thin films, the electrical conduction can be controlled by doping impurities. In the case of polycrystalline semiconductor thin films, the potential barrier at the grain boundary plays a dominant role. Carrier transport in insulator thin films under high electric field occurs through the energy states in the band gap. There are a few current mechanisms in insulator thin films. One or more of the mechanisms dominate depending on the temperature and electric field.

Thin-film technologies have been the basis of microdevices that achieved the electronics revolution in the latter part of the twentieth century. There is no doubt that they will continue to be the key technologies in the future. There seem to be two major goals in thin-film applications. One is to further scale down the existing microdevices to achieve not only even higher density integration but superior performance. The other is to create novel microdevices with new functionalities. To these ends, control of electrical properties in thinner and more finely patterned thin films, and in thin films of new materials, must be studied taking full advantages of state-of-the-art deposition methods and related science and technologies.

GLOSSARY

Annealing: A process to provide energy (usually thermally) to a material to improve its quality by reducing defects.

Chalcogenide: A group of materials that contain one or more atoms of S, Se, or Te.

Complementary Metal-Oxide–Semiconductor (CMOS) Circuit: A circuit consisting of *p*-type and *n*-type MOS transistors in which low power consumption performance is achieved due to complementary operation of serially connected *p*- and *n*-type transistors.

Field-Effect Mobility: A mobility measured in a field-effect transistor structure reflecting the conduction at the insulator–semiconductor interface.

Grain Boundary: The boundary between adjacent grains, where defects concentrate.

Hall Effect: A phenomenon in which an electric field arises when a magnetic field is applied perpendicular to a current flow in a material.

Ion Implantation: A method to incorporate atoms into a solid-state material by accelerating ions by high electric field.

Metal-Oxide–Semiconductor (MOS) Transistor: A field-effect transistor with a stacked structure of metal-oxide insulator and semiconductor.

Passivation: A technique to protect a thin film that is vulnerable to exposure to air by covering the thin films using a stable insulator thin film.

Thin-Film Transistor (TFT): A field-effect transistor built using a semiconductor thin film that is originally deposited on an insulator substrate.

Very Large-Scale Integration (VLSI): Circuits composed of transistors, capacitors, and resistors integrated on a semiconductor (usually silicon) chip.

Works Cited

Fuchs, K. (1938), *Proc. Cambridge Philos. Soc.* **34,** 100.

Ioffe, A. F., Regel, A. R. (1960), *Prog. Semicond.* **4,** 239–246.

Kamins, T. (1988), *Polycrystalline Silicon for Integrated Circuit Applications,* Boston: Kluwer Academic Publishers.

Kamiyama, M. (Ed.) (1983), *Thin Film Handbook,* Tokyo: Ohm Corporation.

Kanicki, J. (Ed.) (1993), *Amorphous & Microcrystalline Semiconductor Devices—Optoelectronic Devices,* Boston: Artech House.

Kazmerski, L. L., Rachine, D. M. (1975), *J. Appl. Phys.* **46,** 791–798.

Kimbara, A. (1987), *Basic Technologies of Thins*

Films (in Japanese), Tokyo: University of Tokyo Press.

Levinson, J., Shepherd, F. R., Scanlon, P. J., Westwood, W. D., Este, G., Rider, M. (1982), *J. Appl. Phys.* **53,** 1193–1202.

Mott, N. F., Davis, E. A. (1979), *Electronic Processes in Non-Crystalline Materials,* 2nd ed., Oxford, U.K.: Clarendon Press.

Serikawa, T., Shirai, S., Okamoto, A., Suyama, S. (1989), *Jpn. J. Appl. Phys.* **28,** L1871–L1873.

Seto, J. Y. W. (1975), *J. Appl. Phys.* **46,** 5247–5252.

Sondheimer, E. H. (1950), *Phys. Rev.* **80,** 401–406.

Spear, W. E., Le Comber, P. G. (1975), *Solid State Commun.* **17,** 1193–1196.

Sze, M. S. (1981), *Physics of Semiconductor Devices,* New York: Wiley.

Weimer, P. K. (1962), *Proc. IRE* **50,** 1462–1469.

Yamauchi, N., Reif, R. (1994), *J. Appl. Phys.* **75,** 3235–3257.

Further Reading

Brody, T. P. (1984), *IEEE Trans. Electron Devices* **ED-31,** 1614–1628.

Eckertova, L. (1986), *Physics of Thin Films,* New York: Plenum Publishing Co.

Kamins, T. (1988), *Polycrystalline Silicon for Integrated Circuit Applications,* Boston: Kluwer Academic Publishers.

Kanicki, J. (Ed.) (1993), *Amorphous & Microcrystalline Semiconductor Devices-Optoelectronic Devices,* Boston: Artech House.

THIN FILMS, NONELECTRONIC PROPERTIES

H.-U. Habermeier, *Max-Planck-Institut für Festkörperforschung, Stuttgart, Germany*

Introduction 387
1. **Mechanical Properties of Thin Films** 388
1.1 General Comments 388
1.2 Hardness of Thin Films 388
1.3 Internal Stress of Thin Films 390
1.3.1 Thin Discontinuous Films 390
1.3.2 Epitaxial Single-Crystal–Type Thin Films 390
1.3.3 Continuous Polycrystalline Films 391
1.4 Stress Relaxation 391
1.5 Adhesion of Thin Films 392
1.6 Fracture Mechanisms and Thin-Film Delamination 393
2. **Optical Properties of Thin Films** 394
2.1 Basics of Thin-Film Optics 394
2.2 Optical Constants of Thin-Film Materials 396
2.3 Multilayer Optical Systems 397
3. **Analysis of Nonelectronic Properties of Thin Films** 398
3.1 Analysis of Structural and Chemical Composition 398
3.2 Analysis of Mechanical Properties 400
3.2.1 Stress Measurements 400
3.2.2 Measurement of Thin-Film Hardness 401
3.2.3 Measurement of Thin-Film Adhesion 402
3.3 Analysis of Optical Properties 402
3.4 Surface Analysis 402
3.4.1 Diffraction Methods 402
3.4.2 Scattering Methods 403
3.4.3 Scanning Probe Microscopy 404
Glossary 405
Works Cited 406
Further Reading 407

INTRODUCTION

In general, the properties of any solid-state material, either bulk or thin film, are determined by the arrangement of the constituent atoms, the nature of the chemical bonds between the constituents, and the elementary excitations occurring in the system. Consequently, the distinction between electronic and nonelectronic properties is somewhat artificial since the electronic structure of the material is one of the fundamental ingredients determining the physical properties of matter. However, based on the generally accepted terminology, the following distinction is used as a guideline: electronic properties—all properties that are determined by charge transfer within the system; nonelectronic properties—all properties not dominated by charge transfer.

According to this definition, this article reviews the basic phenomena of mechanical and optical properties of thin films, and treats the analysis of structure and chemical composition as well as the analysis of mechanical and optical properties. Since film properties are substantially affected by surface effects, a brief survey of relevant surface analytical techniques is included, irrespective of the fact that they are not used for thin-film analysis exclusively. To deduce the nonelectronic properties of thin films from first principles determining the atom arrangement and to treat the specific relationship between the microstructure and the physical properties deserves other articles in this volume of the Encyclopedia, and the reader is referred to them (especially the chapters dealing with electron structures in Volume 5 and the article THIN FILMS, ELECTRONIC PROPERTIES in this volume). Consequently, a rigorous treatment of the substrate precon-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ditioning, thin-film nucleation, growth mechanisms, and microstructure evolution is not included here. Furthermore, not covered in this article are nonelectronic thin-film phenomena that are determined by charge transfer such as mechanical stability changes due to electromigration, dielectric properties, and thermal-transport properties.

In science and technology, the mechanical and optical properties of thin films play a vital role. The mechanical properties such as hardness, wear, durability, adhesion, and cracking of thin films are major topics to be considered in all applications associated with microelectronics and microsystem technology. Hard, wear-resistant and protective coatings are used in improving the lifetime of metal-forming tools, cutting tools, and all areas where tribological properties of materials have to be improved—e.g., in bearings and gears.

Thin-film optics is a topic of invaluable importance. The application of thin films as antireflecting coatings in everyday consumer products is obvious. All sorts of optical devices have been improved or even made possible by the application of thin films and thin-film multilayers. In recent years the application of thin-film technology suitable to fabricate optical devices for x-ray optics has experienced a breakthrough. Optical systems such as monochromators, mirrors, and focusing mirrors are available now.

There is a myriad of diagnostic tools suitable for thin-film analysis. The general trend, however, is the development of element-specific, nondestructive techniques with high lateral resolution. Systems based conceptually on the scanning tunneling microscope (near-field characterization probes) (see TUNNELING MICROSCOPY AND MICROANALYSIS) with a lateral and vertical resolution in the atomic range will be indispensable tools for thin-film analysis in the years to come.

1. MECHANICAL PROPERTIES OF THIN FILMS

1.1 General Comments

The mechanical properties of thin films—similarly to the optical, magnetic, and electrical properties—depend mainly on the material-specific intrinsic characteristics as determined by the type of chemical interactions of the atoms involved, resulting in the chemical bonding, crystal structure, and electronic band structure. Superimposed are extrinsic features arising from film growth and microstructure evolution. For a fundamental description of the mechanical properties of thin films, it is therefore mandatory to correlate the observed behavior with their microstructure. The microstructure is well known to depend on substrate selection, substrate preconditioning, and the deposition method itself (classified, e.g., by the kinetic energy of impinging atoms and/or molecules) as well as deposition rate, substrate temperature, and residual gas composition and pressure. Adsorption, desorption, and diffusion of the particles impinging at the substrate site are the fundamental processes to be controlled. Adhesion, e.g., is determined by nucleation as well as by the early stages of film growth. Other properties like hardness, internal stress, and fracture depend on the subsequent growth and resultant crystallographic type (amorphous, polycrystalline, textured, and single crystal) of the film.

A rigorous treatment of the mechanical properties of thin films therefore has to be based on an atomistic description of the arrangement of atoms as determined by the film-forming processes and the interaction of the atoms involved. Instead of this rigorous treatment, in this section a more pragmatic concept is used, describing especially those mechanical properties of thin films that are relevant for applications, such as hardness, stress, adhesion, fracture, and delamination. For simplicity reasons atomistic bonding concepts are used to pave the way for a fundamental understanding of thin-film hardness. For the description of thin-film stress as well as adhesion, fracture, and delamination processes based on stress issues, the atomic structure is mostly ignored, and the properties are explained in terms of continuum mechanics and thermodynamics.

1.2 Hardness of Thin Films

Hardness of thin films is defined by the ratio of the peak load and the projected contact area in a microindenter experiment (see details in Sec. 3.4.2). The properties of materials depend on the bond within the solid-state matter; mechanically hard materials including thin films, therefore, can be classified by their bonding types:

1. metallic hard materials,
2. hard materials with covalent bonding, and
3. hard materials with heteropolar bonding (ionic character).

Since the classification of bond holds only in extremes, a sharp separation is not always possible.

Metallic hard thin films are the carbides, nitrides, silicides, and borides of group-IIIA–VIA transition metals as well as some of the group-VIIA and group-VIIIA transition metals that form interstitial solid-solution structures. Typical values for the Vickers hardness (HV) are 1500–3000 HV (as a reference, diamond has 8000 HV). These films have low ductility, a high coefficient of thermal expansion, and good electrical conductivity. To this category belongs WC, the first artificially synthesized hard material. From the application point of view, TiN and TiAlN are the most widely used thin films of this category, especially in the cutting tool industry, tribological (bearings and gears), as well as decorative applications (e.g., substitution for gold). In TiAlN an increasing Al content causes a more wear-resistant coating; commonly the composition $Ti_{0.5}Al_{0.5}N$ is used. The main advantage of TiAlN is its superior performance at elevated temperatures due to the enhanced oxidation temperature (700 °C compared to 500 °C for TiN).

Among the covalent hard materials—diamond as well as the carbides and nitrides of group-III and group-IV elements B, Si, Al—cubic BN as well as B_4C are important as the second and third hardest materials known. *c*-BN coatings, however, are difficult to synthesize because of the competition of the mechanically soft hexagonal modification *h*-BN in the deposition process. Plasma chemical vapor deposition (PCVD) as well as pulsed-laser deposition (PLD) are the techniques of choice to prepare these films. On account of the unique properties of diamond (hardness, thermal conductivity, chemical inertness, wide energy band gap, dielectric constant), thin films are of great technological importance. Thin films are mainly deposited using plasma-assisted or hot-filament–assisted CVD.

Heteropolar hard thin films are represented by the oxides; they have the advantage of high strength at high temperatures and—naturally—no oxidation problems. Furthermore, they are interesting as diffusion barriers and thermal barrier coatings.

Table 1 compiles a selection of the mechanical properties of materials from these three categories. For details in thin-film

Table 1. Properties of different hard materials.

Phase	Density (g/cm^3)	Hardness (HV/100)	Young's modulus (kN/mm^2)	Resistivity (Ω cm)	Thermal expansion ($10^{-6}/K$)
			Metallic hard materials		
TiN	5.40	21	590	25	9.4
VC	5.41	29	430	59	7.3
VN	6.11	15.6	460	85	9.2
NbN	8.43	14	480	58	10.1
WC	15.72	23.5	720	17	3.8
LaB_6	4.73	25.3	400	15	6.4
			Covalent-bonded hard materials		
B_4C	2.52	30–40	441	5.10^5	4.5
c-BN	3.48	50	660	10^{18}	—
C (diamond)	3.52	80	910	10^{20}	1.0
SiC	2.43	26	480	10^5	5.3
AlN	3.26	12.3	350	10^{15}	5.7
			Heteropolar-bonded materials		
Al_2O_3	3.98	21	400	10^{20}	8.4
TiO_2	4.25	11	205	—	9.0
ZrO_2	5.76	12	190	10^{16}	11
BeO	3.03	15	390	10^{23}	9.0

preparation of hard and protective materials, the reader is referred to a recent review by Knotek and Schrey (1995).

1.3 Internal Stress of Thin Films

The formation of heterophase boundaries such as film/substrate or film A/film B as well as homophase boundaries such as grain, twin, and domain boundaries are the main sources for internal stress occurring in thin films. In layered structures generally two types of internal residual stress occur whose physical origins are quite different: thermal stress and intrinsic stress.

Thermal stress is due to the constraint imposed by the film–substrate bonding and is produced by the thermal mismatch between layers of different thermal expansion coefficients, α. In the approximation for an isotropic thin film on a thick substrate, biaxial stress conditions are present and, provided there is no change in microstructure and no phase transition between deposition temperature and measurement temperature, the total thermal strain ϵ_{therm} is simply

$$\epsilon_{\text{therm}} = \int (\alpha_{\text{film}} - \alpha_{\text{sub}})dT = \Delta\alpha\Delta T, \qquad (1)$$

where $\Delta\alpha$ corresponds to the difference in the linear expansion coefficients of film and substrate, and ΔT to the difference of deposition temperature and measurement temperature. Combined with Hooke's law, we get for the thermal stress

$$\sigma_{\text{therm}} = \frac{E_f}{1 - \nu_f}\Delta\alpha\Delta T, \qquad (2)$$

where E_f corresponds to the Young's modulus and ν_f to the Poisson ratio of the film. Typical values for α (in units of 10^{-6}/°C) are 8 for glass, 10–20 for many metals, and 30–40 for alkali halides. If ΔT and $\Delta\alpha$ are positive, a tensile stress arises; for negative ΔT and positive $\Delta\alpha$ compressive stress is found.

The second contribution, the intrinsic stress, summarizes all contributions to the total residual stress that are not of thermal origin. They depend on a variety of experimental parameters such as film material, structure, thickness, deposition technique, deposition rate, deposition temperature, cooling rate, etc. Generally, several of these parameters are not measured quantitatively in conventional deposition experiments; hence tables for data of intrinsic stress for various combinations of substrate/film materials have to be treated with caution. For deposition experiments performed at room temperature, typical values for the intrinsic stress range from 10^9 Pa (iron on mica) to nearly zero (tin on glass). A variety of mechanisms produce intrinsic stress in different types of films.

1.3.1 Thin Discontinuous Films

Surface tensions are the main source, and, with increasing coverage (coalescence of islands), compressive stress can change sign to tensile stress. Since $\sigma \sim 1/t$, only very thin films show appreciable stress $>10^8$ Pa.

1.3.2 Epitaxial Single-Crystal–Type Thin Films

Differences in the substrate and film lattice parameters are the main reasons for intrinsic stress, which may be either tensile or compressive. When the elastic energy of a coherently distorted lattice is larger than the energy required to create misfit dislocations, a dislocated incoherent interface is formed. Figure 1 compares the lattice planes of film and substrate in cases with perfect lattice-matched epitaxy with a coherently strained lattice-mismatched interface and a relaxed one. In Fig. 1(a) an epitaxial film is grown upon a substrate with the same crystal structure and lattice parameter; in principle, in this case there is no driving force leading to elastic energy accumulation and thus dislocation formation. In Fig. 1(b) an epitaxial

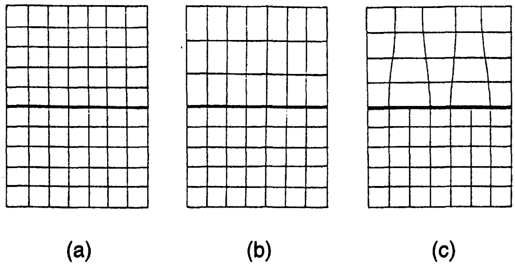

FIG. 1. Schematic representation of the lattice planes of film (top) and substrate (bottom) in (a) a lattice-matched heteroepitaxy, (b) coherently strained lattice-mismatched heteroepitaxy, and (c) relaxed lattice-mismatched heteroepitaxy. At a critical thickness configuration (b) converts to configuration (c).

film is grown upon a substrate with different lattice parameter. For thin layers, up to a critical thickness t_c, the difference in lattice parameter can be accommodated by a biaxial stress in the interfacial plane causing a distortion along the growth normal. If the film thickness exceeds t_c, this elastic strain is relaxed by the formation of misfit dislocations [Fig. 1(c)]. In an energy-balance consideration, the elastic energy is calculated as a function of lattice mismatch, film thickness, and dislocation density. By minimizing this energy, the dislocation density can be calculated as shown by Hull and Bean (1992).

1.3.3 Continuous Polycrystalline Films

A general thermodynamic principle states that atomic disorder is usually accompanied by an increase in volume, and any relaxation to a more ordered state causes a decrease in volume. This principle can be used as a guideline to judge whether stress is tensile or compressive. Tensile stress is expected in films whose deposition temperature is larger than room temperature. This holds especially for thin films deposited via vacuum evaporation. In the presence of impurities, however, this tensile stress can convert into compressive stress. On the other hand, sputtered thin films are usually found in a state of compressive stress. Atom bombardment (peening), whether from energetic depositing atoms or deliberate ion bombardment, creates this compressive stress. Models based on the theory of radiation damage in solids can explain the development of compressive stress. A detailed analysis of stress in sputtered films shows that the normalized momentum $P^* = \vartheta\sqrt{ME}$ (ϑ = flux of energetic particles/atomic flux, M = atomic mass, E = energy of impinging atoms) is the appropriate scaling factor. In Fig. 2 an idealized stress–momentum curve is given, depicting the variation of the intrinsic stress with P^* for any sputtering or ion-assisted deposition process. At low P^*, porous films are deposited unable to support high tensile stress; as the momentum increases, densification occurs as voids collapse, and the tensile stress reaches a maximum. At moderately high momentum, the tensile stress decreases, as compensating compressive stress becomes operative caused by the atomic bombardment mechanisms. When the microstructure becomes fully densified, a conversion to compressive stress takes place that scales with $\sqrt{E}$ as expected theoretically by the forward sputtering cascade model. In the high-momentum region, plastic flow is commonly observed as the yield strength of the film is exceeded, and thus the compressive stress is reduced. Problems associated with internal stress formation have been reviewed by Windischmann (1992).

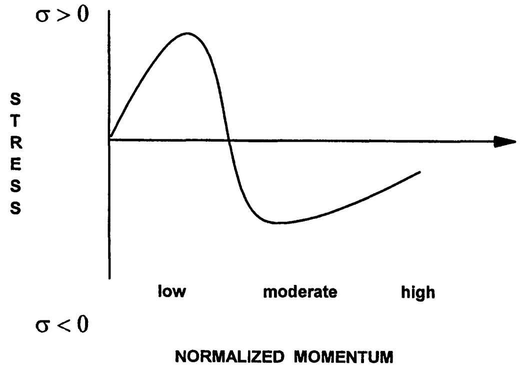

FIG. 2. Idealized intrinsic stress of a film vs normalized momentum P^* for sputter deposition (after Windischmann, 1992). Details are explained in the text.

Furthermore, recrystallization occurring upon cooling from deposition temperature to room temperature in polycrystalline thin films can give rise to microstructural changes, e.g., of grain size D_0 to $D > D_0$; and thus a strain

$$\epsilon = (\Delta - \alpha)(1/D_0 - 1/D) \qquad (3)$$

occurs with $\Delta - \alpha$ being the grain-boundary relaxation length (values around 0.1 nm are typical). Volume changes due to phase transitions and polymorphous transitions are further examples for thickness-independent mechanisms for stress generation.

1.4 Stress Relaxation

The residual stress in thin films, either extrinsic or intrinsic, can be quite large; typical numbers are 0.1–1 GPa. A stress of this magnitude corresponds to about 1% strain in many materials and is comparable to the yield strength of most bulk materials. When film peeling does not occur (shear stress < adhesion for peeling), several mechanisms will lead to stress relaxation: dislocation generation and gliding, dislocation climbing, and diffusion creep. For low temperatures, $T < 0.4T_m$ where T_m is the melting temperature, and small values for stress, $\sigma \leq \sigma_c$ (σ_c: critical shear stress), there is an elastic distortion, only. Dislocation gliding occurs for $T < 0.4T_m$, when

$$\sigma \geq 2\mu b/t_f = \sigma_c \tag{4}$$

(μ, shear modulus; b, Burgers vector; t_f, film thickness). Typical values for σ_c are 4×10^8 Pa for lead and 10^9 Pa for Ta thin films of thickness 100 nm. For $\sigma \leq \sigma_c$ and temperatures around $\frac{1}{2}T_m$, dislocation climbing (the dislocation moves perpendicular to the glide plane) is the dominant process, which is associated with point-defect diffusion, especially the diffusion of vacancies.

For high temperatures, $T > 0.4T_m$, and small values for σ, diffusion creep occurs. The rate $\nabla_t \epsilon$ is proportional to σ and depends on the temperature in a complicated way. A classical review of diffusion creep is given by Gibbs (1966). To get an overview of the different mechanisms as a function of temperature and acting stress, the introduction of deformation maps has been proven to be a practical tool. Murakami (1984) reviews the essential features of deformation maps.

1.5 Adhesion of Thin Films

For the adhesion of dissimilar materials with atomically flat surfaces, a variety of mechanisms can be responsible, ranging from covalent bonding to van der Waals attraction, and can include capillary adhesion, electrostatic attraction and mechanical interlocking.

Practically, the adhesion is generally determined by the properties of an interfacial region, which can be structurally as well as chemically different from the bulk phases involved. The properties of these interfacial layers can be modified by physical—e.g., diffusion processes—or chemical—chemical reactions—methods or artificially introduced intermediate layers. In cases of film/substrate couples that have no bulk chemical affinity (e.g., metals on ceramics), a stable bonded interface must be created that will intrinsically link the materials involved by formation of chemical complexes.

Quantitatively, the energy of adhesion is defined by the work W_{ad} required to separate a unit area of the bonded surfaces:

$$W_{\text{ad}} = \gamma_f + \gamma_s - \gamma^*, \tag{5}$$

where γ_f and γ_s correspond to the surface free energy of the film and the substrate, respectively, and γ^* to the interface free energy.

Another quantity that characterizes the strength of bonding across the interface is the theoretical strength of the interface, $\sigma_{\text{max}} = f_c/A$, where f_c is the load required to remove the film of area A. In an idealized experiment, the displacement δ and the load σ are measured [c.f. Fig. 3a] during the uniform separation of two bonded materials. The form of the plot of applied stress vs displacement is represented schematically in

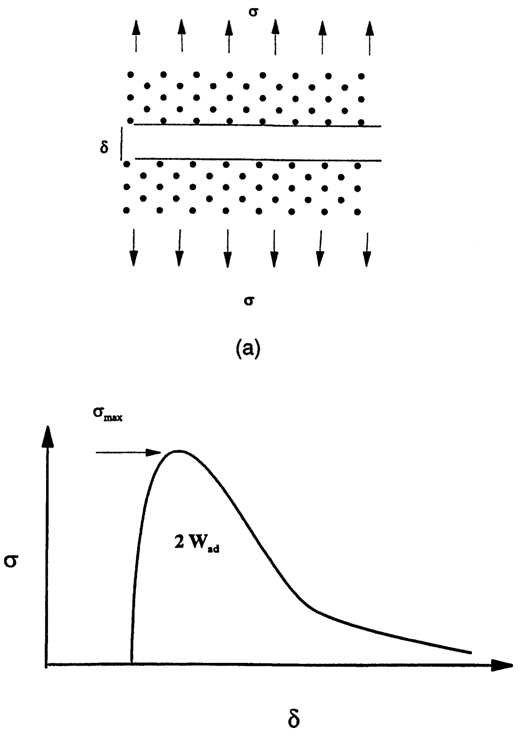

(a)

(b)

FIG. 3. Idealized stress-displacement plot for the definition of W_{ad} and σ_{max}. (a) Uniform separation of an interface by an applied stress results in (b) a stress/displacement plot, defining the theoretical strength of the interface σ_{max} and the thermodynamic work of adhesion $2W_{\text{ad}}$. (After Thouless, 1994)

Fig. 3b; here the area under the curve gives the thermodynamic work of adhesion, and the peak stress, $\sigma_{\max}$, gives the theoretical strength of the interface.

The strategy to improve adhesion derived from Eq. (5) is to lower the surface energy of the interlayer in order to maximize W_{ad}. An example for the implementation of a reactant layer to accomplish adhesion is the commonly used technique of bonding metals to glass via a Cr/Cr-oxide intermediate layer. Often, simply disorder introduced at a crystallographically perfectly terminated substrate surface is sufficient to induce new chemical bond configurations. Plasma modification of polymer surfaces is a standard technique for adhesion improvement. Other techniques to create well controlled, low-energy, chemically stable interfacial layers consist in the different methods of using energetic ions. Table 2 summarizes several different strategies for the enhancement of film adhesion as suggested by Baglin (1994). Presputtering in vacuum using inert species such as Ar will clean the surface from contaminants. Furthermore, this procedure will break up surface-bond chains from a polymer substrate or the terminating layer of a metal oxide and thus activate surface bonding sites. Ion-assisted deposition may be used to bombard the receiving surface with energetic particles. The increased kinetic activity at the receiving surface causes removal and/or separation of surface contaminations in addition to the formation of denser films. Generally, dense films have reduced intrinsic stress, which in turn lowers stored stress energy and thus improves adhesion. Ion implantation is a method for adding reactive species in a controlled way even in the vicinity of an already formed film/substrate interface. Modifications in a graded way are accessible using the dependence of the straggling range of ions on the kinetic energy of the impinging ions. In ion-beam stitching techniques, energetic, inert ions pass through a film and the interface and into the substrate. Each ion creates a sheathlike collision cascade within the solid. During the short period of high atomic mobility, a substantial rearrangement of the atoms within the cascade can occur, resulting in substantial layer interdiffusion. Adhesion is improved by the reconfiguration of the interface structure, interface roughening, and contaminant dispersion.

1.6 Fracture Mechanisms and Thin-Film Delamination

Cracking and, as a result, thin-film delamination occurs when the elastic energy of a film of thickness t and unit area equals a characteristic energy, W_f, required to form a free surface. In a descriptive way the formation of a dislocation network can be imagined that leads to weakening the adhesion and results finally in delamination. The energy balance

$$E^*\epsilon^2 t = W_f \tag{6}$$

(E^*, average Young's modulus; ϵ, total strain) results in a "phase diagram" for the different phenomena as a function of film thickness (cf. Fig. 4). In this figure, the phase boundaries are given in a stress/thickness representation, separating the areas where the different mechanisms are occurring. It is worth noting that the strain scales with $\epsilon \sim \sqrt{1/t}$ as long as γ can be regarded to be thickness independent. The empirical rule that the me-

Table 2. Mechanisms for ion-beam–assisted enhancement of thin-film adhesion (e, effective; ve, very effective; —, noneffective).

Adhesion enhancement mechanism	Substrate presputtering in vacuum	Ion-asissted deposition	Ions implanted at interface	Ion-beam stitching
Remove/disperse contamination	ve	ve	—	e
Activate substrate bonding sites	ve	ve	—	—
Reconfigure interface structure	—	—	e	e
Tailor elemental abundance	e	e	e	—
Add reactive species		—	ve	—
Roughen disperse interface	e	—	e	e
Reduce film stress	—	ve	—	—

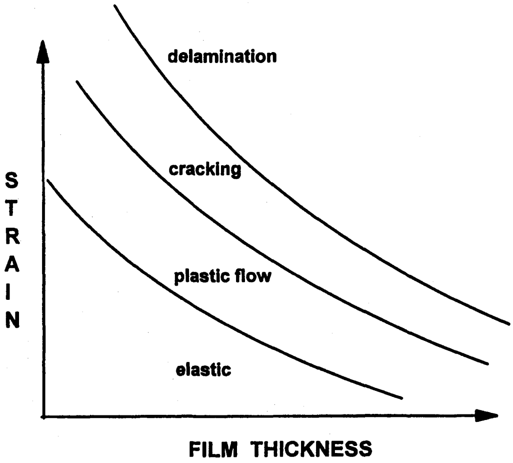

FIG. 4. Schematic "phase diagram" separating the areas of elastic, plastic, cracking, and delamination processes in thin films.

chanical stability of thin films decreases with increasing film thickness is based on this.

The typical mechanism leading to delamination is crack growth by breaking of bonds in a highly strained region of the material. The driving force for crack growth, F_c, is

$$F_c = (\partial/\partial A)(W_{ex} - U), \qquad (7)$$

with W_{ex} being the work done by external forces and U the elastic energy stored in a unit volume.

Delamination occurs independent of whether the stress is compressive or tensile. Compressive stress causes buckling-driven delamination; tensile stress results in a bending momentum around the neutral axis acting at the interface of film and substrate. A comprehensive treatment of problems associated with thin-film adhesion, friction, and delamination has been published by Thouless (1994).

2. OPTICAL PROPERTIES OF THIN FILMS

2.1 Basics of Thin-Film Optics

The fundamental properties investigated by thin-film optics are reflection R, transmission T, and absorption A of light at various wavelengths λ and angles of the incident beam, Φ_0. The phenomenological description of reflection, transmission, and consequently absorption (through the law of energy conservation) is given by macroscopic electrodynamics based on Maxwell's equations. For details the reader is referred to the corresponding articles of this Encyclopedia.

The macroscopic optical properties of an isotropic medium conventionally are characterized by the complex refractive index

$$N = n - ik, \qquad (8)$$

where the real part n usually is called refractive index and the imaginary part k is known as the extinction coefficient. Using the description with the refractive index, the rigorous treatment of calculating reflection, transmission, and absorption based on macroscopic electromagnetic theory is reduced to the determination of the steady-state amplitudes of the electrical- and magnetic-field vectors at the interfaces arising from an incident light wave of certain characteristics.

For parallel-sided, nonabsorbent thin films of refractive index n_1 supported by a substrate with refractive index n_2 and incident beam from a medium (e.g., air) with n_0 at angle Φ_0, the reflectance and transmittance are readily obtained by summing multiple reflected beams (cf. Fig. 5):

$$\begin{aligned} R &= \frac{r_1^2 + r_2^2 + 2r_1r_2\cos\delta}{1 + r_1^2 + r_2^2 + 2r_1r_2\cos\delta}, \\ T &= \frac{n_2 t_1^2 t_2^2}{n_0(r_1^2 + r_2^2 + 2r_1r_2\cos\delta)}; \\ \delta &= 2\pi\nu n_1 t_1 \cos\Phi. \end{aligned} \qquad (9)$$

Here, r_1, r_2, t_1, and t_2 correspond to the Fresnel reflection and transmission coefficients (for the electric vector either perpendicular or parallel to the plane of incidence) for the air/film (index 1) and the film/substrate (index 2) interfaces, respectively. δ is the phase difference between successive beams.

These general expressions show an oscillatory variation with film thickness (see Fig. 6). Written in terms of the refractive index, they are rather complicated; for the special case of normal incidence and nonabsorbent films and $n_0 < n_1 < n_2$, however, these expressions reduce to

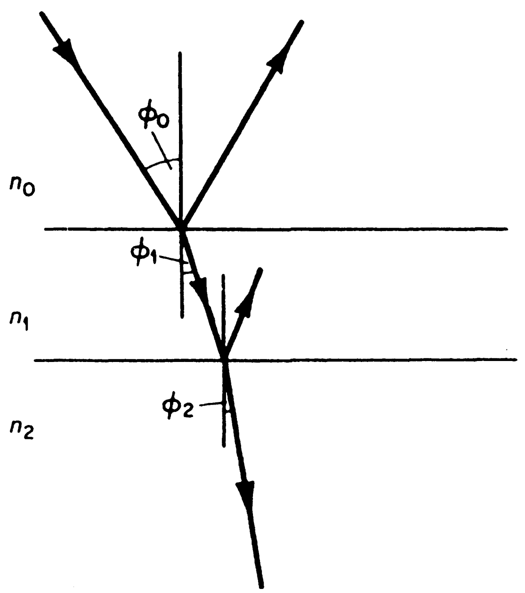

FIG. 5. Ray diagram of a light beam incident at angle Φ_0 on a plane, parallel-sided film of thickness t, and refraction index n_1 supported by a substrate of refraction index n_2.

$$R_{\max} = [(n_2 - n_0)/(n_2 + n_0)]^2 \quad \text{and}$$

$$R_{\min} = [(n_1^2 - n_0 n_2)/(n_1^2 + n_0 n_2)]. \qquad (10)$$

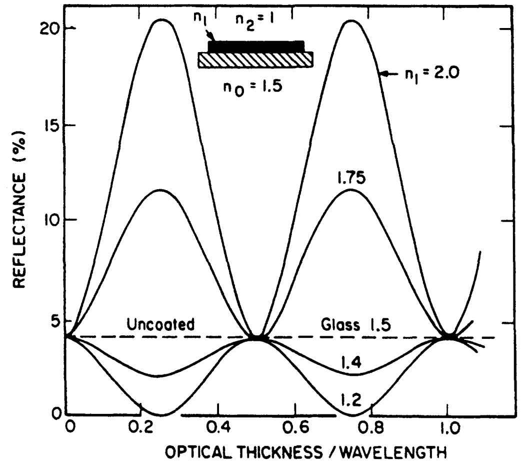

FIG. 6. Theoretical variation of reflectance with normalized film thickness for films with different refractive indices on a glass substrate with n_0 = 1.5.

The evaluation of Eqs. (9), which are valid for any type of film and substrate media, is cumbersome especially when stacks with alternating absorbing and nonabsorbent films are treated. For nonnormal incidence the s and p components of nonpolarized light have different amplitudes, and at a certain angle of incidence (Brewster's angle, ϕ_B) unpolarized light is reflected plane polarized:

$$\tan\phi_B = n_0/n_1. \qquad (11)$$

In case of absorbing thin films, both the amplitude and the phase of originally linearly polarized light change, and consequently the incident light beam is reflected elliptically polarized (see Fig. 7). The phase change of E_s relative to E_p, Δ (E_s and E_p being the field components of the electrical-field vector parallel and perpendicular to the plane of incidence, respectively), and the amplitude change Ψ are related to the sample complex reflection coefficient for light polarized parallel (R_p) and perpendicular (R_s) to the plane of incidence:

$$\tan\Psi \exp(i\Delta) = R_p/R_s = \rho(\lambda,\Phi). \qquad (12)$$

For a bare substrate with no overlayer and a perfectly smooth surface, R_p and R_s are simply the Fresnel reflection coefficients and ρ is readily converted into the dielectric function of the substrate. If the substrate is covered by a film and the same procedure is done, the pseudodielectric function $\langle\epsilon(\omega)\rangle$ is obtained,

$$\langle\epsilon\rangle = \sin^2\Phi\left[1 + \frac{1-\rho}{1+\pi}\tan^2\Phi\right]. \qquad (13)$$

The ratio R_p/R_s includes the refractive indices of a film (or stack of films), the corresponding thickness, angle of incidence, and light wavelength. The complete mathematical treatment of the electromagnetic theory of reflection and transmission in general cases is given in textbooks of optics, e.g., by Born and Wolf (1972).

The discussion of optical properties of absorbing materials is similar; however, the refractive index n has to be replaced by the complex refractive index N. It can readily be seen that the computation of reflectance and transmittance in the case of complex refractive index is not simple, and the reader is referred to the corresponding literature—e.g., Levi (1968).

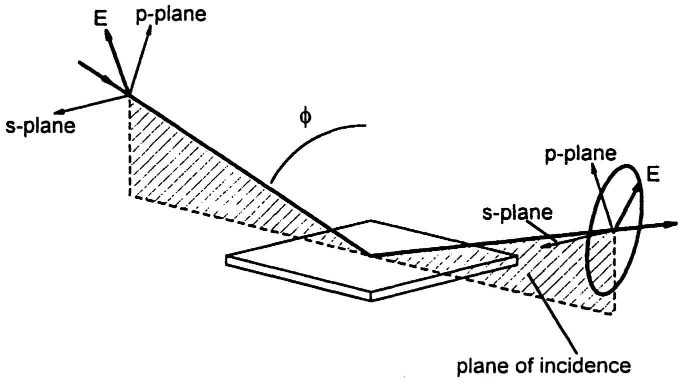

FIG. 7. Change in polarization state of a linearly polarized light beam incident on a sample at oblique angle Φ_0 (basic principle of an ellipsometer).

2.2 Optical Constants of Thin-Film Materials

The dielectric susceptibility $\chi(\omega)$ and the dielectric function $\epsilon(\omega)$ are macroscopic average quantities. Microscopically, these quantities are derived from the summation over the local ones; thus, the average functions $\chi(\omega)$ and $\epsilon(\omega)$ depend on their atomistic counterparts and the way in which atoms are arranged and bonded with the crystal. Consequently, the microstructure, by determining the local arrangement of atoms, influences the dielectric function and thus the real and imaginary parts of the refractive index. The microstructure, in turn, is strongly related to the deposition method and the corresponding deposition parameters. Condensing species deposited by thermal evaporation usually have insufficient kinetic energy to migrate into equilibrium positions after sticking at the substrate or the growing film surface. This causes a reduced mechanical density of the film as well as reduced real part and enhanced imaginary part of the refraction index. Heating the substrate provides additional thermal energy to enhance the diffusivity for surface migration of the evaporant, resulting in more dense films. Figure 8 gives an example showing the dispersion curves of sapphire thin films deposited at different substrate temperatures T_s. The increase of T_s results in an increase of n; the bulk value ($n = 1.76$ for $\lambda = 0.5$ μm), however, is not reached. Thermally evaporated coatings usually show a columnar microstructure with discrete columns of material extending from the substrate/film interface to the film surface. The columns are separated by voids, which can absorb water vapor when the deposition system is vented, and especially in higher humidity atmosphere, these voids may be filled with water molecules in an unpredictable manner. Reactions due to the enlarged inner free surface affect not only the optical properties but also the chemical stability of the films. Most of the materials used for optical coatings are compounds; many of them are either susceptible for preferential sputtering of the anion or decompose when thermally evaporated. Both effects result in nonstoichiometric deposits. Precise stoichiometry control, however, is mandatory to achieve low-loss optical coatings. Compared to thermal or electron-beam evaporation, the sputtering technique provides more dense films mainly because of the higher kinetic energy of the

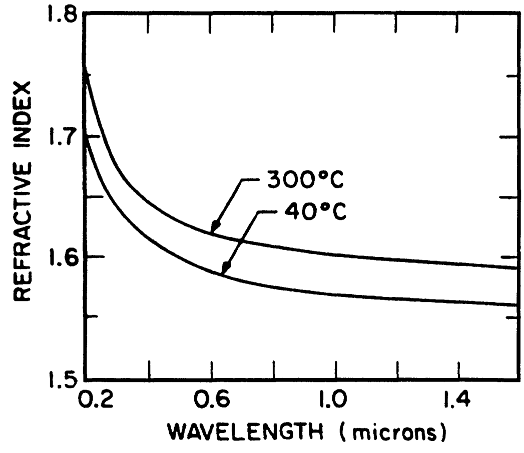

FIG. 8. Refractive index of Al_2O_3 thin films evaporated at different deposition temperatures, as a function of wavelength.

Table 3. Refractive indices for representive materials for optical coatings.

Material	Refractive index	Wavelength for measurement [μm]	Spectral range for transparency [μm]
CaF_2	1.24	0.546	0.15–15
CeF_3	1.65	0.55	
Si	3.3	2.0	>1.0
Ge	4.4	2.0	1.7–20
InAs	3.4	3.8	>7.3
InSb	4.0	7.3	>7.3
KBr	1.56	0.589	0.6–30
MgF_2	1.35	2.0	0.23–10
Na_3AlF_6 (cryolyte)	1.36	0.57	0.2–10
NaCl	1.54	0.589	0.2–>15
SiO_2	1.44	0.55	0.2–5
SiO	1.55–2.0	0.55	0.5–8
ZnS	2.35	0.55	0.4–14

adatoms (1–10 eV) impinging on the substrate and growing film. These films, however, suffer from compressive stress depending on the deposition parameters used (see Sec. 1.3). To circumvent these problems, ion-beam–assisted evaporation is a widely used technique for industrial applications; here, the thermally evaporated material is exposed at the substrate side to an energetic ion bombardment. Increased packing density, reduced permeability to water, and changes in the film crystal structure result from ion-beam bombardment. Typical ion energies range from 30 to 500 eV. In the case of oxide thin films an oxygen ion beam not only provides the kinetic energy required to form thin films with a density close to the bulk; it also furnishes activated oxygen to achieve stoichiometric thin films. A further parameter affecting the film properties is the ion-beam current density. There is a critical value for achieving a maximum in the refractive index; ion-beam current densities larger than the critical value cause film degradation due to nonstoichiometry, creation of closed isolated voids, or incorporation of gas bubbles into the film. As a result of the dependence of the refractive index on the actual deposition process as well as on the deposition parameters, the data collected in Table 3 therefore can only be treated as representative figures.

2.3 Multilayer Optical Systems

The derivation of the expressions of reflectance and transmittance of multilayer systems is rather straightforward. The usual electromagnetic boundary conditions are applied at each interface, and the amplitudes of the waves in any layer are written in terms of the adjacent layers; the successive operations reveal the amplitude of the waves emerging from either side, expressed in terms of the incident wave. There are several methods for the mathematical treatment to solve this problem, and the reader is referred to the special literature, e.g., Born and Wolf (1972).

Typical examples for multilayer optical systems include antireflecting coatings, high-reflectance nonmetallic mirrors, interference filters, and bandpass and blocking coatings. High-reflectance nonmetallic mirrors consist of a stack of layers with equal optical thickness and alternating refractive index. For the wavelength λ_0 and layer thickness $\lambda_0/4$, all multiple-reflected beams are in phase for perpendicular incidence. At λ_0 the system has a high reflectance and a flat top for a large number of layers; Fig. 9 represents an

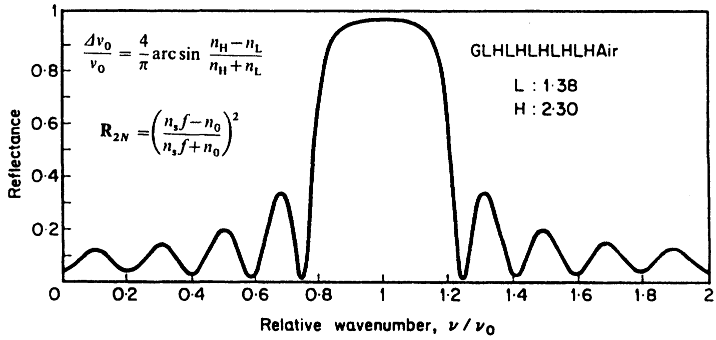

FIG. 9. Reflectance band of a quarter-wave stack on glass (G) with n = 1.51. The low-index material (L) has a refraction index n = 1.38, the high-index material, n = 2.3. The formulas (left upper part) give the width of the reflection band ($\Delta\nu_0/\nu_0$) and the reflectance at ν_0. In the right upper part the stacking sequence is given.

example of a quarter-wave stack with ten layers. The width of the reflection band and the reflectance at λ_0 are determined by the refractive indices of the individual layers and the number of layers according to the equations given in the inset in Fig. 9.

Antireflecting coatings, on the other hand, consist in the application of at least one single layer of optical thickness $\lambda_0/4$ and refractive index $\sqrt{n_s}$, which reduces the reflectance of a surface of index n_s to zero at λ_0. To achieve a low reflection over an extended range of wavelengths, suitable stacks of dielectric layers have to be used. An example for a four-layer antireflecting system is given in Fig. 10. The construction parameters of such antireflection coatings depend on the refractive indices of the substrate and the ambient medium as well as on the width of the spectral region over which the reflection is to be reduced. Dobrowolski and Sullivan (1996) review design methods that yield optimum solutions to the antireflection problem for a given pair of coating materials and for a given overall thickness of the multilayer.

In the past decade a particular class of optical multilayers has been developed especially for soft x rays (100 eV–3 keV, i.e., λ = 12.4–0.41 nm) and extreme UV (15–100 eV, i.e., λ = 80–12.4 nm). These systems consist of a periodic layered microstructure (superlattice) of materials with significantly different scattering powers for x rays. Model calculations using a Fresnel multilayer computational code allowing the optimization of multilayer structure and composition for maximum reflectivity show, for normal incidence, peak reflectivities in the range of 40–80% for λ = 0.6–10 nm. The best material pairs for these purposes will be those that form smooth layers with compositionally abrupt interfaces and have high optical contrast with low absorption in the frequency range required. Selection criteria are therefore not only the optical constants but also the ability of the materials involved to form continuous smooth layers, low interdiffusivity, low chemical reactivity, and low chemical affinity to the residual gases of the deposition system (especially hydrogen, oxygen, water vapor, and carbon oxides). Material combinations typically used for soft x-ray mirrors include Mo, Ru, Rh, or W as high–electron-density layers and C, Si, B, or Be as low–electron-density layers, depending on the wavelength used. Practically, the achieved peak reflectivities for normal incidence are lower than the calculated ones but still as high as 66% for λ = 13.4 nm in Si/Mo multilayers. The degradation of the peak reflectivity is mainly due to the fact that superlattices are not in a thermodynamic equilibrium, and thus the structure is sensitive to preparation methods and growth conditions. Interfacial roughness between the layers, compositionally graded interfaces, scatter in the period of the superlattice, and problems with substrate surface flatness are the main reasons for the degradation of the peak reflectivity. The physics, fabrication, and applications of these types of multilayered structures have been reviewed by Spiller (1994).

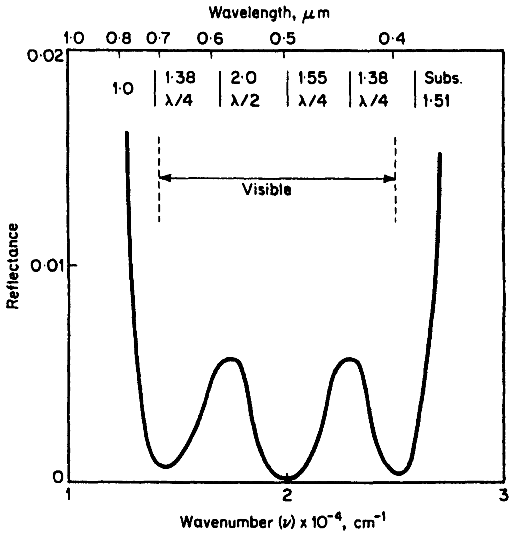

FIG. 10. Four-layer antireflecting system on glass with n = 1.51. The notation in the upper part indicates the composition of the stack with films of different refractive indices.

3. ANALYSIS OF NONELECTRONIC PROPERTIES OF THIN FILMS

3.1 Analysis of Structural and Chemical Composition

Standard x-ray diffraction analysis of thin films has the advantage of relative ease and convenience and makes possible the accurate determination of film properties such as orientation, phases present, and lattice parameter. In the commonly used Bragg–Brentano configuration (Θ–2Θ scan), the Bragg condition of diffraction,

$$n\lambda = 2d\sin\Theta, \tag{14}$$

gives the position of the diffraction peaks when the angle between the incident beam and the film plane is scanned (λ is the x-ray wavelength, n the order of reflection, d the lattice spacing, and Θ the angle of incidence). The diffraction pattern represents the average through the film, and usually the simultaneous display of the diffraction patterns of the film and the substrate is observed. Texture analysis is accomplished when in addition to the Bragg angle the polar angles ω and Φ are scanned; the mosaic spread (or tilt) of domains in the films can be investigated by scanning the angle ω around $\omega = 0$, while Θ is set to Bragg conditions. The rotation of the film domains around the surface normal (Φ scan with Θ in Bragg conditions) gives the in-plane orientation. Furthermore, a quantitative analysis of the full width at half maximum of the diffraction peaks allows the determination of the mean grain size $\langle s\rangle$ of thin film according to the Scherrer formula

$$\langle s\rangle = \lambda/[\Delta(2\Theta)\cos\Theta], \tag{15}$$

with $\Delta(2\Theta)$ being the angular broadening of the diffraction peaks. X-ray analysis has been refined recently by Fullerton *et al.* (1992) to such extents that detailed structural information like disorder, chemical interdiffusion, and lattice strain—even in the case of superlattices—can be revealed, when the diffraction intensities are measured quantitatively and compared with model calculations.

Complementary to x-ray diffraction techniques, Raman scattering (inelastic photon–phonon scattering) has been proven to be a powerful diagnostic tool, which can provide answers to questions such as

1. state, quality, and phase of thin films, crystallinity, polycrystallinity, microstructures, etc.;
2. residual strains following material treatment (implantation, thermal or laser annealing, polishing, etching) in thin films, heterostructures, and superlattices.

The power of light-scattering spectroscopies for thin-film analysis results from the rather stringent selection rules that arise from the conservation of the wave vector in the scattering process. The probing depth depends on the penetration depth of the light used, and structural as well as compositional peculiarities can be investigated even for thin films with thickness below 100 nm as demonstrated by Zhang *et al.* (1996). Details, like the presence of foreign phases, film orientation, degree of epitaxy, and in some cases details of the degree of oxygenation in thin-film oxides, can readily be revealed from Raman spectra. As an example for a quantitative determination of the degree of epitaxy, Fig. 11 compares the x-ray (103) pole figures with the Raman spectra of an epitaxially grown $YBa_2Cu_3O_{7-\delta}$ thin film with two different settings of the polarizer and analyzer of the Raman-scattering experiment. The x-ray pole figure—basically a three-dimensional representation of the scattered x-ray intensity for a given Bragg angle Θ while scanning the polar angles ω and Φ—demonstrates the fourfold symmetry in the pole density, suggesting in-plane epitaxy. From the intensity ratio of the Cu–O in-plane vibration (at 340 cm^{-1}) for the two settings, the degree of epitaxy of the film is determined to be 0.91, i.e., more than 90% of the grains are aligned with respect to the substrate (in this case strontium titanate) as demonstrated by Thomsen *et al.* (1992).

Several methods are applicable for chemical analysis. The most simple consist in removing the thin film by a solvent and analyzing the solution by chemical techniques. Emission spectroscopy is one of the methods capable of quantitative elemental analysis.

Nondestructive techniques like electron microprobe analysis (EMA), electron spectroscopy for chemical analysis (ESCA), or Rutherford backscattering (RBS) are favored, especially since RBS is capable of quantitative analysis of composition, thickness, and depth profile of thin films. RBS is the energy analysis of ions backscattered from the surface. Ions with low mass (e.g., He^+) are accelerated toward the sample and are backscattered with an energy E_1 at an angle Φ_0 resulting from the energy and momentum

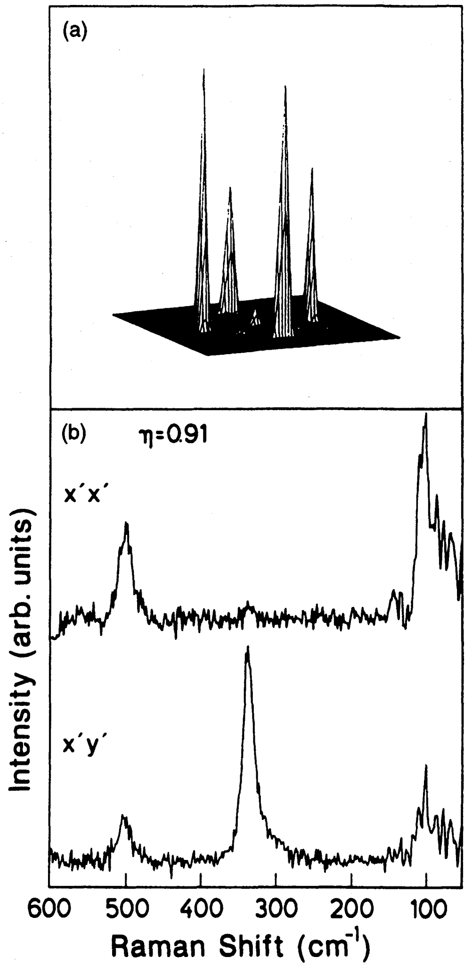

FIG. 11. X-ray (103) pole figure and Raman spectra of an epitaxially grown Y-Ba-Cu-O thin film measured in two different polarization configurations to determine the degree of epitaxy.

conservation laws for scattering. The concept of close-impact scattering collisions implies that it is independent of features of the electron shell and the chemical bonding; thus RBS is not sensitive to the electron configuration and bonding states. This simple scattering concept allows RBS to sort out the various masses of a sample and thus determines the sample composition. Applications include compositional analysis, depth profiling, and interdiffusion studies. A rigorous treatment of RBS is given by Chu *et al.* (1978). Recently, Chu and Langouche (1993) gave a review on quantitative RBS analysis in thin films including channeling phenomena to analyze lattice disorder as well as the location of impurity atoms.

Electron microprobe analysis is based on the detection and measurement of characteristic x rays of material excited by energetic electrons. The essential feature is the localized excitation of a small volume of the sample with a finely focused electron beam and the detection of the x rays as demonstrated in Fig. 12. When suitable standards are available, even a quantitative analysis can be carried out. Feldman and Mayer (1986) give a comprehensive introduction into the details of microprobe analysis.

3.2 Analysis of Mechanical Properties

3.2.1 Stress Measurements There are three categories of measurements used in practice for stress measurements:

1. measurement of substrate bending,
2. mechanical resonance techniques, and
3. x-ray or electron-diffraction techniques.

In static methods based on substrate bending, the thin film of thickness t is deposited on a substrate in the form of a rectangular strip or a square plate. When the substrate of length L and thickness d is fixed at one end (cantilever beam), the stress in the film causes a strain in the substrate, and a displacement of the free end, δ, occurs that is measured by interferometric, electromechanical, or capacitive techniques. The stress is calculated from the displacement, and for cases where the film thickness is much less than the substrate thickness, the stress σ is determined by the Stoney equation

$$\sigma = E_s d^2 \delta / 3(1 - \nu) L^2 t, \qquad (16)$$

where E_s is the Youngs modulus for the substrate and ν the Poisson's ratio for the substrate.

Mechanical resonance techniques rely on the fact that the resonant frequency of a membrane is a function of both its mass density and its stress, which together determine the speed of sound in the membrane. The resonant frequency f of a film/substrate combination is given by

$$f = \frac{C}{R} \left\{ \frac{\sigma_f t + \sigma_s d}{\rho_f t + \rho_s d} \right\}^{1/2}. \qquad (17)$$

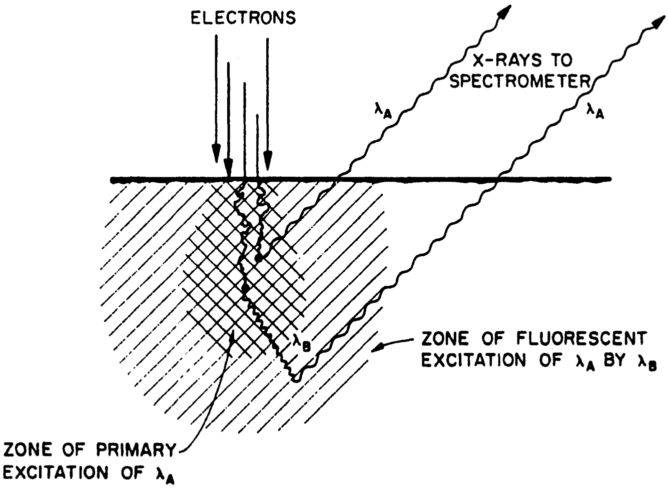
ELECTRONS
X-RAYS TO SPECTROMETER
λ_A
λ_A
λ_B
ZONE OF FLUORESCENT EXCITATION OF λ_A BY λ_B
ZONE OF PRIMARY EXCITATION OF λ_A

FIG. 12. Schematic illustration of the generation of x rays by electron microprobe analysis indicating the different generation spheres for primary and fluorescent x-ray excitation.

Here, C is a constant related to the mode of vibration, R is the radius of the circular membrane, and ρ_i the mass densities of film and substrate, respectively.

Electron-diffraction and x-ray diffraction methods can be used to measure changes in lattice spacing and hence stresses in thin films. In cases where the crystallite sizes are less than 100 nm, the diffraction lines are broadened as described by the Scherrer formula [see Eq. (15)], and a distinction between crystallite size broadening and stress-induced broadening is complicated. Hoffman (1966) has published an extensive review of the literature dealing with stress measurement, and the various methods for macrostress and microstress determination are examined critically. A more recent review by Clemens and Bain (1992) especially treats problems associated with stress measurements by x rays.

3.2.2 Measurement of Thin-Film Hardness Hardness of a material is defined as the resistance to deformation and is usually measured as the permanent deformation of a surface by a specifically shaped indenter under a given load. Usually micro- or nanoindentation is used; here, a diamond indenter is driven into and removed from a film while the load on and displacements of the indenter are measured. Figure 13(a) shows schematically the physical processes involved. As the indenter is driven into the film, both elastic and plastic deformations occur. The plastic deformation results in a permanent impression with the contact

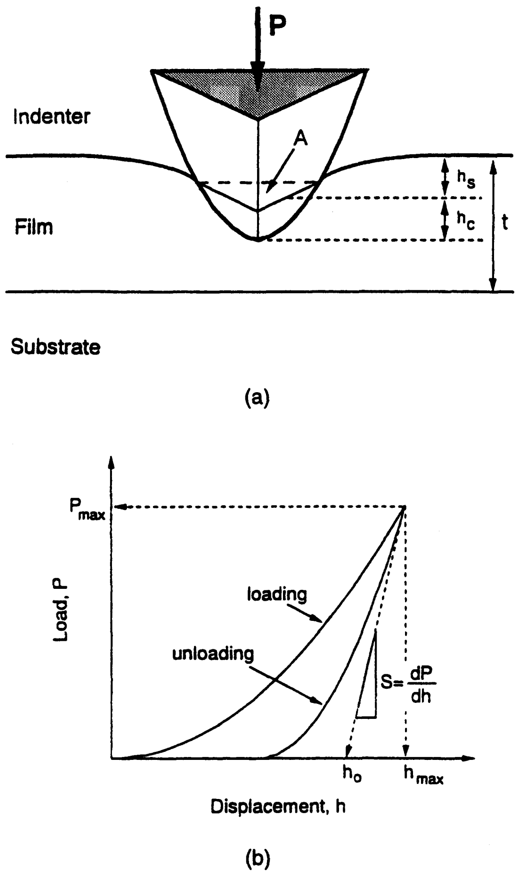
P
Indenter
A
h_s
h_c
t
Film
Substrate
(a)
P_{max}
Load, P
loading
unloading
$S=\frac{dP}{dh}$
h_0
h_{max}
Displacement, h
(b)

FIG. 13. Determination of thin-film hardness by nanoindentation. (a) Physical processes involved in an indentation experiment (for explanations see the text) and (b) indentation typical load/displacement curve defining P_{max} and h_0.

depth h_c. The elastic deformation during the exertion of the load causes the bending of the film surface at the edge of the contact by h_s. In a load–displacement curve [see Fig. 13(b)] the relevant quantities $P_{\max}$ and $h_c = h_0$ are revealed. This allows the determination of the hardness as defined by

$$H = P_{\max}/A. \tag{18}$$

The contact area A is determined by evaluating the area function $A = F(h_c)$ at the depth h_c. A recent review by Pharr and Oliver (1992) highlights physical and practical problems with nanoindentation.

3.2.3 Measurement of Thin-Film Adhesion On a microscopic level the removal of a film from the substrate consists in breaking of the bonds of individual atoms of the film and the substrate. The macroscopic adhesion is therefore a summation of the individual forces. There are a large number of adhesion tests available, as reviewed by Mittal (1987), ranging from tape methods—classical "Scotch-tape test"—to tensile pull tests using wires bonded to the film to be tested, scratch tests, abrasive, and deceleration methods in an ultracentrifuge. Whereas the simple Scotch-tape test reveals only qualitative information, especially for poor adhesion, the tensile pull test allows a more quantitative access to the adhesion forces. A good survey of test methods and their limitations is given in the article of Campbell (1970); Mattox (1994) recently gave a comprehensive review of adhesion measurements.

3.3 Analysis of Optical Properties

Optical probes for analyzing film properties [interferometry (see INTERFEROMETERS AND INTERFEROMETRY), reflectometry, ellipsometry (*q.v.*), and light scattering] have the advantage of directly measuring the complete film/substrate system. Additionally, they are often used as *in situ* real-time analytical probes during the deposition process.

Ellipsometry, e.g., is widely used to characterize the optical properties correlated with the pseudodielectric function $\langle\epsilon\rangle$ (cf. Sec. 2.1). The components of the complex refractive index, n and k, as well as the film thickness can be determined from the measurements of the ellipticity. Furthermore, in spectroscopic ellipsometry where $\langle\epsilon(\omega)\rangle$ is measured, the dispersion of $n = n(\omega)$ as well as $k = k(\omega)$ is accessible. In practice, problems arising from inhomogeneous thin films, rough surfaces, or film stacks on a substrate can be solved by the variation of film parameters and minimizing the deviations from the measured ellipsometric curves. A comprehensive recent review of *in situ* ellipsometry is given by Irene and Woollam (1995) as well as by Snyder (1995).

In contrast to spectroscopic ellipsometry for thin-film analysis, other techniques like Fourier-transform infrared spectroscopy (FTIR), photoluminescence (PL), and Raman spectroscopy play a less dominant role. In FTIR the measured absorption sensitively depends on the presence or absence of impurities in the thin film. A classical example is oxygen in silicon, where the 9-μm band absorption is the fingerprint of interstitial oxygen in the material. In the case of diamond-like amorphous carbon (DLC) films, whose hardness and electrical insulation are believed to depend on the ratio of sp^2 to sp^3 bonding structures, FTIR is an extremely valuable tool to determine the absorption lines of the corresponding stretching bonds.

Photoluminescence is a simple nondestructive technique for the evaluation of the optical properties of semiconductors. Here, electrons are photoexcited from the ground state in the valence band to the conduction band, and a combination of thermalization and radiative recombination occurs, revealing details of the band structure as well as the position of defect levels in the band gap—as long as the transitions are radiative. Impurity-level determination and impurity identification are the most prominent applications of this technique, as reviewed by Nishino *et al.* (1985).

3.4 Surface Analysis

3.4.1 Diffraction Methods In addition to the x-ray diffraction techniques already covered in Sec. 3.1, special cases of x-ray diffraction such as grazing-incidence x-ray diffraction (GIXRD) are used for surface-specific analysis. The techniques based on GIXRD require a bright x-ray source and, usually, access to a synchrotron facility.

Electron-diffraction techniques, on the other hand, can be used as *in situ* real-time analytical tools and do play a much larger role in thin-film analysis. Low-energy electron diffraction (LEED) and reflection high-energy electron diffraction (RHEED) require UHV environment. In LEED electrons of energy 30–150 eV impinge on the sample at normal incidence; the backward-diffracted electrons provide information on the atomic geometry of the top layers of the film. In many ways LEED is analogous to the classical XRD with the addition of high surface sensitivity. It is important to recognize that LEED patterns reveal the periodicity of atoms at the surface as well as the overall symmetry of the surface and not the detailed atomic position. Changes in periodicity of the atom arrangement at the surface result in changes of the diffraction pattern that are easily observable and can readily be interpreted in terms of the new two-dimensional symmetry.

RHEED, on the other hand, is a grazing-incidence reflection technique with a primary electron beam of 10–40 keV. The diffracted pattern is formed on a screen opposite to the electron gun. It is worth mentioning that the probing depths of the two techniques, RHEED and LEED, are quite comparable. This is because the energy E_p normal to the surface to be analyzed is $E_p = E_0 \sin\phi$, with ϕ being the angle of grazing incidence. In addition to the detection of the atomic configuration of the crystalline surfaces, RHEED has the advantage of being able to provide information on surface disorder, defects, and surface topography including roughness, surface facets, and step arrays. Since RHEED is sensitive to topography, it can be used to distinguish between three-dimensional island clustering and two-dimensional layer-by-layer growth in an *in situ* experiment. In the past years RHEED has been the indispensable diagnostic tool for *in situ* deposition control for atomic layer-by-layer molecular-beam epitaxy for semiconductor superlattices, high-temperature superconductors, and colossal magnetoresistance materials. Superstructure modulation, occurrence of twinning, and time-dynamic features such as structural phase transitions can be revealed by RHEED. Bozovic and Eckstein (1995) recently reviewed the application capabilities of RHEED for the analysis of growing films of complex oxides. Lagally and Savage (1993) treat the possibilities and limitations of quantitative electron diffraction by thin films.

3.4.2 Scattering Methods In contrast to the diffraction methods, where the wave nature of particles is applied, in scattering methods, the particle character is used. Some inconsistencies—especially in texture analysis and determination of degree of epitaxy—in comparisons of different methods—e.g., x-ray diffraction and Raman scattering—can be explained by this. Methods sensitive to probing the structure and composition of the surface of a film are based on the interaction of a particle beam (photons, electrons, ions) with the material to be analyzed and the detection of energies and intensities of the emitted beam. The penetration depth of the incident beam together with the escape depth of the emitted particles determines the surface sensitivity. The energy provides the signature of the scattering atoms; the intensity gives the amount of atoms probed. The concept of physics used in these techniques is the inner–electron-shell vacancy formation by irradiation [x-ray photons for x-ray photoemission spectroscopy and x-ray fluorescence (XRF) spectroscopy, electrons for Auger emission spectroscopy (AES) and electron microprobe analysis]. The excited atoms release their energy either by radiative transitions with the emission of x rays (XRF, EMA) or by nonradiative transitions with the emission of electrons (AES). In general, these techniques are extended probe-beam techniques; however, several of them are available as scanning-beam techniques capable of spatial resolutions down to the micron size. There is an impressive multitude of techniques available for thin-film surface analysis. Nearly all combinations of particles are used, and a myriad of acronyms can be found in the literature. The main methods generally used in thin-film analysis can be classified into two categories:

1. identical particles for incident and emergent beam, and
2. incident beam is different from scattered beam.

In Table 4 the most popular of these are listed, indicating the processes involved and giving some average values for probing depth

Table 4. List of volume-averaging techniques for surface analysis, indicating the effective probing depth and surface monolayer sensitivity.

Method	Particle in	Particle out	Acronym	Effective probing depth	Monolayer sensitivity
Auger-electron spectroscopy	Electron	Electron	AES	2 nm	10^{-1}
Rutherford backscattering	Ion (α-particle)	Ion (α-particle)	RBS	~10 nm	10^{-4}
X-ray fluorescence spectroscopy	X ray	X ray	XRF	~1000 nm	none
Photoelectron spectroscopy	X ray	Electron	XPS	~3 nm	10^{-1}
Electron microprobe analysis	Electron	X ray	EMA	~1000 nm	none
Secondary-ion mass spectroscopy	Ion	Target ion	SIMS	~1 nm	10^{-5}

and surface sensitivity. For thin-film analysis AES, Rutherford backscattering (RBS) (see Sec. 3.1), and secondary-ion mass spectroscopy (SIMS) are the most widely used techniques.

In AES the Auger-electron transitions appear as small features superimposed on a background of secondary electrons; consequently, derivative techniques are used to generate a function $dN(E)/dE = f(E_0)$, with E_0 being the energy of the impinging electron beam. Typical values for E_0 are several keV; the detection limit for AES is around 0.1 at.%. AES is widely used for studying surface reactions, surface contaminations, and depth profiles. For this purpose a sputtering gun has to be implemented into the system for layer sectioning. One of the advantages of AES is its sensitivity to low-mass atoms, e.g., carbon, nitrogen, and oxygen, which are typical contaminants at surfaces and interfaces.

Secondary-ion mass spectroscopy is based on the erosion of the sample by energetic ion bombardment, and the relative abundances of the sputtered species give a direct measure of the composition of the layer removed. SIMS (as well as its advanced counterpart secondary neutral mass spectroscopy, SNMS) is used for surface and depth concentration analysis of thin films and superlattices. The problem in all depth-profiling techniques using film sectioning by sputtering is the accurate determination of the depth axis, which is usually done by measurements of the resulting crater depth and converting a time scale to a depth scale. It is beyond the scope of this article to review these methods critically in further details; the reader is referred to the book of Feldman and Mayer (1986) and special monographs like the article of McGuire (1994).

3.4.3 Scanning Probe Microscopy The methods mentioned above cannot detect surface inhomogeneities over atomic distances. Scanning probe microscopy based on the scanning tunneling microscope (STM) invented by Binnig *et al.* (1982) closes the gap to atomic resolution. An atomically sharp tip biased with respect to the sample is positioned in a proximity distance of less than 1 nm from the surface of the sample. Electrons from the tip can tunnel into empty states of the sample, and a tunneling current is generated, increasing exponentially with decreasing distance. This exponential dependence is controlled by piezoelectric transducers; a three-dimensional set of piezoelectric transducers allows the corresponding movements with a resolution of around 0.1 nm. This method allows a two-dimensional representation of the sample surface with atomic resolution. The method relies on the fact that electrons can tunnel into empty states in the film to be analyzed. This is fulfilled in the cases of metals and most semiconductors, but not for insulators. For surface analysis of insulators with a resolution close to the atomistic, repulsive, and/or attractive forces exerted from the film to the tip are used. This type of instrument is called atomic-force microscope (AFM) since its functional principle makes use of van der Waals forces between tip and film. This concept is realized when the tip is fixed on a cantilever, whose deflection upon scanning is measured either by an STM—this was the original version of Binnig, Quate, and Rohrer—or by optical systems using deflection techniques. Tremendous progress has been made in developing other noncontact near-field surface characterization techniques in the past ten years.

For thin-film analysis the superior capabilities of scanning tunneling microscopies have been recently added to the more volume-averaging techniques discussed in the previous sections. These microscopies extend the resolution range beyond high-resolution scanning electron microscopy while adding the capability of obtaining three-dimensional atomic or unit-cell details without extensive and destructive sample preparation required, e.g., for transmission electron microscopy (TEM).

For film growth and surface morphology as well as tribology and hardness/friction studies of thin films, STM analysis is a vital tool because of its combination of extremely high vertical resolution and local nonaveraged details. In fundamental studies of film growth (e.g., high-temperature superconductors, ferroelectric as well as colossal-magnetoresistance thin films), STM and AFM analysis of the substrate surface and the early stages of film deposition play a crucial role in closing the gap between fundamental understanding at an atomic scale and macroscopic properties. Raistrick and Hawley (1994) review recent results of studies of the microstructure of Y-Ba-Cu-O thin films deposited on various oxide substrates and relate the nucleation and growth process to film properties. Special growth-related features, known as growth spirals—a terrace-type layered structure with unit cell height—have been demonstrated to affect the physical properties (e.g., transport, flux pinning) of the films dramatically. Recently, Haage *et al.* (1996) have demonstrated a substrate-mediated defect tailoring in Y-Ba-Cu-O thin films by a careful STM investigation of the relation of substrate preconditioning and film growth.

Scanning probe techniques for thin-film analysis will be one of the indispensable diagnostic tools in the future to bridge the gap between fundamental understanding of substrate conditioning, thin-film nucleation, and film growth on an atomic scale and macroscopic thin-film properties.

GLOSSARY

Adhesion: Ability of a film to stick mechanically on top of a substrate. The thermodynamic work of adhesion, defined as the energy to separate unit area of the two bonded surfaces, and the theoretical strength of the interface (maximum stress in a stress/displacement curve) are used for a quantitative description of adhesion.

Antireflecting Coatings: Reduction of the optical reflection of a substrate (e.g., glass) by depositing one or several dielectric thin films with alternating high and low refractive index materials.

Deformation Map: Stress/temperature diagram identifying the different regions where a particular strain relaxation mechanism (e.g., grain-boundary diffusional creep, dislocation climbing, dislocation gliding) is dominant.

Delamination: Peeling of thin films due to localized breaking of bonds in highly stressed regions of the film/substrate interface; typically at the tip of a propagating crack.

Dielectric Function: In cases where the tensor nature of the complex, second-rank dielectric tensor $\epsilon_{ij}(\omega)$—which couples the electrical-field vector to the Fourier transform of the electric-displacement vector—is not important, $\epsilon_{ij}(\omega)$ reduces to $\epsilon(\omega)$; $\epsilon(\omega)$ is known as dielectric function.

Epitaxial Thin Films: Single-crystal–type thin films grown on a single-crystal substrate with crystallographic relations between substrate and film.

Hardness: Ratio of the peak load and the projected contact area in a microindenter experiment. The measurement method consists of the continuous recording of the indentation load and displacement in a complete cycle of loading and unloading.

Intrinsic Stress: Summarizes the mechanical stress in thin films not due to the difference of thermal expansion coefficients of film and substrate. In general they have a microstructural origin. Examples are interfacial stress in heteroepitaxy, stress due to phase transitions occurring between deposition temperature and room temperature.

Microstructure: Structure of a material at the atomic level including compositional ordering, chemical bonding, presence/absence of impurity atoms as well as point (vacancies, interstitials), line (dislocations), and planar (grain, tilt, twist boundaries) defects.

Polycrystalline Thin Films: Thin films with a random orientation of the crystallites.

Pseudodielectric Function: Dielectric function of a complex, film-covered surface. In contrast to the dielectric function, the pseudodielectric function includes effects due to differences in film density, surface roughness, composition, and microstructure.

Raman Scattering: Inelastic scattering of photons by elementary excitations such as optical phonons, magnons, or plasmons.

Refractive Index: Complex quantity used to describe the macroscopic optical properties of an isotropic medium. The real part is usually referred to as the refractive index; the imaginary part is usually called extinction coefficient.

Textured Thin Films: Films where one crystal axis of the grains show a crystallographic relation with the substrate, while the others are random.

Thermal Stress: Mechanical stress in thin films arising during variations of temperature due to the difference in thermal expansion coefficients of film and substrate.

X-ray Optical Components: Development of optical components (mirrors, lenses) for extreme UV and soft x rays using material combinations with high and low electron density.

Works Cited

Baglin, J. E. E. (1994), *IBM J. Res. Dev.* **38** (4), 413–422.

Binnig, G., Rohrer, H., Berber, Ch., Weibel, E. (1982), *Phys. Rev. Lett.* **49,** 52–55.

Born, M., Wolf, E. (1972), *Principles of Optics,* New York: Pergamon Press.

Bozovic, I., Eckstein, J. N. (1995), *Bull. MRS* **20** (5), 32–38.

Campbell, D. S. (1970), in: L. I. Maissel, R. Glang (Eds.), *Handbook of Thin Film Technology,* New York: McGraw Hill, Chap. 12.

Chu, W. K., Langouche, G. (1993), *Bull. MRS* **18** (1), 32–40.

Chu, W. K., Mayer, J. W., Nicolet, M. A. (1978), *Backscattering Spectrometry,* New York: Academic.

Clemens, B. M., Bain, J. A. (1992), *Bull. MRS* **17** (7), 46–51.

DiNardo, N. J. (1994), *Nanoscale Characterization of Surfaces and Interfaces,* Weinheim: VCH.

Dobrowolski, J. A., Sullivan, B. T. (1996), *Appl. Opt.* **35,** 4993–4997.

Feldman, L. C., Mayer, J. W. (1986), *Fundamentals of Surface and Thin Film Analysis,* Amsterdam: North-Holland.

Fullerton, E. E., Schuller, I. K., Bruynseraede, Y. (1992), *Bull. MRS* **17** (12), 24–32.

Gibbs, G. B. (1966), *Philos. Mag.* **13,** 1203.

Haage, T., Li, J. Q., Leibold, B., Cardona, M., Zegenhagen, J., Habermeier, H.-U., Forkl, A., Joos, Ch., Warthmann, R., Kronmüller, H. (1996), *Solid State Commun.* **99,** 553–557.

Hoffmann, R. W. (1966), *Physics of Thin Films,* Vol. 3, New York: Academic.

Hull, P., Bean, J. C. (1992), *CRC Crit. Rev. Solid State Mater. Sci.* **17,** 507–546.

Irene, E. A., Woollam, J. A. (1995), *Bull. MRS* **20** (5), 24–28.

Knotek, O., Schrey, A. (1995), in: D. A. Glocker (Ed.), *Handbook of Thin Film Process Technology,* Bristol: Institute of Physics Publishing, Chap. XI.

Lagally, M. G., Savage, D. E. (1993), *Bull. MRS* **18** (1), 24–31.

Levi, L. (1968), *Applied Optics,* New York: Wiley.

Mattox, D. M. (1994), in: F. Bunshah (Ed.), *Handbook of Deposition Technologies for Films and Coatings,* Park Ridge, NJ: Noyes Publications, Chap. 6.

McGuire, G. E. (1994), in: F. Bunshah (Ed.), *Handbook of Deposition Technologies for Films and Coatings,* Park Ridge, NJ: Noyes Publications, Chap. 15.

Mittal, K. L. (1987), *Adhesion Sci. Technol.* **1,** 247.

Murakami, M. (1984), *CRC Crit. Rev. Solid State Mater. Sci.* **11,** 317.

Nishino, T., Nakayama, H., Hamakawa, Y. (1985), in: J. Nishizawa (Ed.), *Japan Annual Review in Electronics, Computers and Telecommunications: Semiconductor Technologies,* New York: North-Holland, Vol. 8, p. 295.

Pharr, G. M., Oliver, W. C. (1992), *Bull. MRS* **17** (7), 28–33.

Raistrick, I. D., Hawley, M. (1994), in: S. L. Shinde, D. A. Rudman (Eds.), *Interfaces in High T_c Superconducting Systems,* New York: Springer-Verlag, Chap. 2, pp. 28–70.

Snyder, P. (1995), in: D. A. Glocker (Ed.), *Handbook of Thin Film Process Technology,* Bristol: Institute of Physics Publishing, Chap. D2.3.

Spiller, E. (1994), *Soft X-Ray Optics,* Bellingham, WA: SPIE Optical Engineering Press.

Thomsen, C., Wegerer, R., Habermeier, H.-U., Cardona, M. (1992), *Solid State Comm.* **81,** 589–593.

Thouless, M. D. (1994), *IBM J. Res. Dev.* **38,** 367–378.

Windischmann, H. (1992), *CRC Crit. Rev. Solid State Mater. Sci.* **17,** 547–596.

Zhang, P. X., Haage, T., Habermeier, H.-U., Ruf, T., Cardona, M. (1996), *J. Appl. Phys.* **80,** 2935–2938.

Further Reading

Brief Introductory Works

Eckertova, L. (1977), *Physics of Thin Films,* New York and London: Plenum Press. A simple introduction to thin-film preparation, thin-film structure, and thin-film properties.

General Texts

Bunshah, R. F. (1994), *Handbook of Deposition Technologies for Films and Coatings,* 2nd ed., Park Ridge, NJ: Noyes Publications. An up-to-date overview of deposition technologies, nucleation, film growth and microstructure evolution, thin-film properties, and thin-film analysis.

Glocker, D. A. (1995), *Handbook of Thin Film Process Technology,* Bristol and Philadelphia: Institute of Physics Publishing. An excellent source for details in thin-film preparation and characterization.

Maissel, L. I., Glang, R. (1970), *Handbook of Thin Film Technology,* New York: McGraw-Hill. The classical handbook of thin-film technology, including chapters describing properties of thin films.

Mechanical Properties

Baglin, J. E. E. (1991), in: L.-H. Lee (Ed.), *Fundamentals of Adhesion,* New York: Plenum Press, Chap. 13.

Campbell, D. S. (1970), in: L. Maissel, R. Glang (Eds.), *Handbook of Thin Film Technology,* New York: McGraw-Hill.

Gille, G. (1985), in: E. Kaldis (Ed.), *Strength of Thin Films and Coatings,* Current Topics in Materials Science Vol. 12, Amsterdam: North-Holland Publishing, Chap. 7.

Hoffmann, R. W. (1966), *Physics of Thin Films,* Vol. 3, New York: Academic.

Mittal, K. L. (1987), "Selected Bibliography on Adhesion Measurement of Films and Coatings," *J. Adhesion Sci. Technol.* **1,** 247–259.

Optical Properties

Flory, F. (1995), *Thin Films for Optical Systems,* New York: Dekker.

Hummel, R. E., Günther, K. H. (Eds.) (1995), *Handbook of Optical Properties Series,* Vol. 1, Boca Raton, FL: CRC.

McLeod, H. A. (1986), *Thin Film Optical Filters,* Bristol: Hilger.

Pulker, H. K. (1987), *Coatings on Glass,* 3rd ed., Amsterdam: Elsevier.

Thelen, A. (1989), *Design of Optical Interference Coatings,* New York: McGraw-Hill.

Thin-Film Analysis

Aspnes, D. E. (1985), in: E. D. Palik (Ed.), *Handbook of Optical Properties of Solids,* Orlando, FL: Academic.

Chen, C. J. (1993), *Introduction to Scanning Tunneling Microscopy,* Oxford, U.K.: Oxford Univ. Press.

DiNardo, N. J. (1994), *Nanoscale Characterization of Surfaces and Interfaces,* Weinheim: VCH.

Drevillon, B., Yakovlev, V. (1994), "Optical Characterisation of Real Surfaces and Thin Films," in: K. Vedam (Ed.), *Physics of Thin Films,* Vol. 19, Orlando, FL: Academic.

Feldman, L. C., Mayer, J. W. (1986), *Fundamentals of Surface and Thin Film Analysis,* Amsterdam: North-Holland.

Ghosh, P. K. (1983), *Introduction to Photoelectron Spectroscopy,* New York: Wiley-Interscience Publishers.

Oechsner, H. (Ed.) (1984), *Thin Film and Depth Profile Analysis,* New York: Springer-Verlag.

Relevant Journals

Thin Solid Films
Surface Science
Journal of Applied Physics
Journal of Vacuum Science and Technology

THIN-FILM DEPOSITION

JOHN R. ARTHUR, *Oregon State University, Corvallis, Oregon, U.S.A.*

	Introduction	409
1.	**Essentials of Vacuum Technology**	410
2.	**Techniques for Deposition of thin Films**	412
3.	**Physical Methods of Deposition**	413
3.1	Vacuum Evaporation	413
3.2	Sputter Deposition	414
3.3	Ion Plating	416
3.4	Ion-Beam Deposition	416
3.5	Liquid-Phase Deposition—Langmuir–Blodgett Films	416
4.	**Chemical Methods**	417
4.1	Chemical Vapor Deposition	418
4.2	Plasma-Enhanced CVD	418
5.	**Epitaxial Deposition**	419
5.1	Molecular-Beam Epitaxy	420
5.2	Vapor-Phase Epitaxy	423
5.3	Metal-Organic CVD	424
5.4	Liquid-Phase Epitaxy	424
5.5	Solid-Phase Epitaxy	425
5.6	Atomic-Layer Epitaxy	425
6.	**Surface Science and Thin-Film Deposition**	425
	Glossary	426
	Works Cited	427
	Further Reading	428

INTRODUCTION

Thin-film deposition is the process of placing a thin layer of one substance on another in order to modify the properties of the second material—e.g., to increase hardness—or to change electrical conduction or to modify the optical properties of the underlying material referred to as the substrate. The film itself may also have useful properties of electrical conduction, optical reflectivity or absorption, or corrosion resistance. Thin films have become a significant part of microelectronic circuitry by virtue of their small size, i.e., their nearly two-dimensional nature, and, of course, this permits the fabrication of very small device structures, or microcircuits. However, thin films have many features other than size that make them useful in electronic applications; viz., a large variety of methods for deposition exist that can be quite efficient in the usage of material. Also, many of these deposition methods utilize a vacuum environment, which allows very high purity to be obtained in the resulting film; and finally, it is fairly straightforward to deposit multiple layers of different materials to produce composite films with properties not possessed by any of the components alone. In some very important cases, the high purity of the deposition environment combined with cleaning of the substrate permits the films to be deposited with a single-crystalline structure that "fits" on the crystal structure of the substrate, a process called "epitaxial growth."

We begin by defining more specifically what is meant by a thin film. In electronic applications distinction is made between "thin" and "thick" films. A thin film is a layer of material that is typically less than a few micrometers in thickness, while a thick film could be as much as a millimeter in thickness. The principal distinction, however, lies in the deposition process. Thin films are deposited by the addition of atoms or molecules to the substrate, while thick films are deposited as particles, usually contained in a liquid suspending medium. Thick films are typically applied by painting or spraying, which might be followed by a firing process to decompose or fuse the binder material. In this discussion we consider only thin films. Thin films are most often applied in a closed chamber, frequently under vacuum conditions, either by a physical process such as thermal evaporation of the film material or by a chemical process in which a vapor species is chemically or thermally decomposed

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

on the surface to be covered. We also consider some methods of thin-film deposition from liquid solutions.

Films may be amorphous in structure or display varying amounts of atomic order, at least locally. They are often composed of microcrystals and in some cases may consist of very large (in two dimensions) single-crystal grains. Films with crystalline order may have a structural relationship with the substrate on which they are deposited. A polycrystalline film that shows an ordering of the microcrystals with respect to the crystalline structure of the substrate is often called an ordered polycrystalline film as opposed to a randomly ordered film. For many electronic applications where the film may be carrying an electrical current, the scattering of carriers at grain boundaries may be intolerable, and consequently it is necessary to minimize the scattering centers by constraining the film to grow as a large, thin single crystal.

In semiconductor devices, thin films are used to form the conducting lines connecting individual devices, as well as the contact pads to which are bonded the wires that connect the circuit to the packaging structure. They are also used to form insulating layers to isolate conducting films from the underlying device structure. In many cases, part or all of the active device may be composed of a single crystal of semiconductor material deposited epitaxially on the underlying supporting substrate, in order to provide semiconductor material with properties not available from the bulk substrate crystal. In brief, thin-film technology is an essential component of microcircuit fabrication, and advances in circuit complexity have progressed in step with advances in our understanding of how to control the deposition and properties of thin films. See Smith (1995) and Campbell (1996) for a more complete discussion.

Thin films have been extremely important in developing an understanding of surfaces and interfaces. Because of the essentially two-dimensional nature of films, many of the properties are dominated by the proximity of interfaces. Consequently, films have provided experimental access to changes in the surface environment. Perhaps one of the most dramatic demonstrations of thin-film science has come from the electronic properties of "quantum wells," i.e., the properties of thin films that have one or more dimensions smaller than the wavelength of the electrons in the film. These "low-dimensional structures" show clearly the transition from a continuum of energy states to discrete, quantized electronic states in the optical and transport properties (Allan *et al.*, 1986).

Much of the recent scientific interest in films has come from the increasingly sophisticated and powerful array of analytical techniques for examining the atomic structure of thin films. Electron, ion, and photon spectroscopic techniques, discussed by Feldman and Mayer (1986), identify the chemical composition and bonding state of atoms in films with extremely high spatial resolution, while the new tunneling probe microscopies display the atomic microstructure. These techniques have provided a much better understanding of the nature of thin films and have played a crucial role in optimizing the methods for their deposition.

The present discussion is limited to an overview of methods for depositing thin films, with an emphasis on those used in the fabrication of electronic devices. The properties of films will be considered only as they are influenced by the method of deposition. Readers wishing more detailed discussion are encouraged to consider the suggestions in the Further Reading list appended to this chapter.

1. ESSENTIALS OF VACUUM TECHNOLOGY

Before considering deposition *per se,* we discuss some elementary aspects of vacuum technology, since the vacuum environment is essential to most methods of deposition (O'Hanlon, 1989). Except in a few notable instances that are discussed, high-purity films are deposited in vacuum, and, most generally, in ultrahigh vacuum (UHV), which we define as a pressure lower than 1×10^{-7} Torr. The reasons for this are evident from a discussion of elementary gas kinetics.

Since a typical solid contains $\sim 10^{22}$ atoms/cm^3, a surface has an atomic density of $\sim 3 \times 10^{14}$ atoms/cm^2, i.e., the volume density raised to the $\frac{2}{3}$ power. It can be shown from fundamental gas-kinetic theory that the number n of gas atoms impinging on unit area of surface in unit time is

$$\frac{dn}{dt} = \frac{P}{\sqrt{2\pi mkT}}, \tag{1}$$

where P is the gas pressure, m is the atomic mass, k is the Boltzmann constant, and T is the absolute temperature. Equation (1), the Knudsen equation, is extremely important in thin-film physics, because it permits a comparison of the rate of deposition with the rate of contamination from the residual gas in a vacuum system. For example, if the vacuum pressure P is measured in Torr and if the atomic mass m is converted to molecular weight M in grams, then Eq. (1) becomes

$$\frac{dn}{dt} = 3.5 \times 10^{22} \frac{P}{\sqrt{MT}} \quad \mathrm{cm}^{-2}\,\mathrm{s}^{-1}. \tag{1a}$$

For a typical residual gas molecular weight of 40 g and a temperature of 25°C, the rate of gas arrival is $3.2 \times 10^{20} P(\mathrm{Torr})$, and for a pressure of 10^{-6} Torr the arrival rate is $3.2 \times 10^{14}\ \mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. In other words, the arrival rate at 10^{-6} Torr is just about equal to the number of atoms in a square centimeter of surface. Thus it takes only a second to completely cover an initially clean surface with residual gas, assuming all of the arriving atoms or molecules remain on the surface. The pressure of 10^{-6} Torr is a fairly representative ultimate vacuum attained in typical diffusion-pumped systems that are opened frequently to insert substrates for film deposition—for example, bell-jar systems used for thermal evaporation of metal films for circuit metallization. The cautionary point is that the rate of deposition must be orders of magnitude faster than the contamination rate in order to avoid impure films. This requires either better vacuum or high rates of supply from the film source.

In order to improve the vacuum, we need to consider what limits the vacuum in a system. The rate of change of pressure in a continuously pumped chamber is given by

$$V\frac{dP}{dt} = -PS + G \quad \mathrm{L\ Torr/s}, \tag{2}$$

where P is the pressure in Torr, V is the volume of the chamber, S is defined as the speed of the pumping system, and G is the generation rate of gas from various sources within the system. When G is small compared to P, e.g., at the beginning of the pump-down cycle, Eq. (2) predicts that the pressure will decrease exponentially with time with a time constant V/S. The ultimate pressure limit is reached when the rate of removal of gas, PS, is equal to the generation of gas, G, and the ultimate pressure is

$$P = G/S. \tag{3}$$

Obviously, reducing the generation rate G and increasing the pumping speed S will both act to lower the ultimate pressure obtainable.

The need for improved vacuum in thin-film technology has been a major driving force in the development of improved pumping systems. Today there are many choices of pumping schemes, including diffusion pumps, ion pumps, turbomolecular pumps, and cryopumps, for achieving vacuum less than 10^{-6} Torr (see VACUUM TECHNOLOGY AND INSTRUMENTATION). Cryopumps in particular feature very high pumping speeds, since they are based on the principle of providing a large cold surface that effectively condenses all arriving gas molecules. One application of cryopumps is in the coating of automobile bumpers with chromium or aluminum films. In this instance, the vacuum system is designed for very high throughput of material—essentially a moving assembly line—and large cryopumps provide the extremely high pumping speed to deal with the heavy gas load. Each type of pumping technology has advantages and disadvantages, and the reasons for the possible choices are discussed in detail in references in the Further Reading list.

The other approach to improved vacuum lies in reducing the gas generation rate. There are two ways in which to accomplish this. The straightforward approach is to raise the temperature of the entire vacuum chamber for a prolonged "bakeout" to desorb gas from all the interior surfaces of the vacuum system, and this is essential to achieve pressures in the 10^{-10}–10^{-11} range. Once the internal walls of a vacuum system are clean, it is possible to insert and remove substrates without recontaminating the system by the addition of a vacuum interlock system in which an intermediate chamber is used to load, pump down, and outgas the substrates prior to insertion in the hard vacuum. This approach requires additional hardware, valves, and pumps but is increasingly useful in the electronic industry where rapid throughput and high vacuum standards are both necessary (Hablanian, 1990).

The question of the vacuum necessary to avoid contamination of thin films from the residual gas depends to a great extent on what level of contamination can be tolerated. For instance, an Al film used for metallization of an integrated circuit might be evaporated onto a substrate in a vacuum of 10^{-6} Torr at a rate of 0.33 μ/min, which is about 20 atom layers per second. From the discussion above, we might expect that as much as 5% impurity might be found in the film, since about one atom layer of impurity per second comes from the residual gas. Actually, since much of the residual gas in the deposition system would consist of fairly inert species such as N_2 that do not adsorb readily, it is highly unlikely that the impurity content would be this great, but even if it were, this would not seriously affect the quality of the film, since the metallic conduction would not be drastically impaired. On the other hand, deposition of a semiconductor film under similar conditions would be impossible because that level of contamination would completely overwhelm the desired semiconducting properties of the film. We know that a good semiconducting film should have an impurity level $<10^{-6}$. To decrease the impurity level by a factor of 5×10^4 means a similar reduction in background pressure to less than 10^{-10} Torr. Thus while high vacuum may be adequate for the deposition of metals, UHV will be necessary for semiconductors.

2. TECHNIQUES FOR DEPOSITION OF THIN FILMS (Ohring, 1992; Campbell, 1996)

Methods for the deposition of thin films can be conveniently divided into two classes: physical deposition techniques, in which atoms or molecules of the source material are vaporized by physical processes such as thermal evaporation or by bombardment from an energetic beam of electrons, photons, or ions; and chemical deposition techniques, which rely on a chemical reaction involving one or more vapor species incident on a substrate to produce a film of the desired composition. Physical deposition is typically carried out in vacuum to avoid contamination of the film by the ambient atmosphere, while chemical deposition often utilizes an inert carrier gas, and may even be carried out at atmospheric pressure.

The steps of film deposition are as follows:

1. Generation of a supply of atoms or molecules from a source, which may be solid, liquid, or vapor. This is accomplished by supplying energy to the source in the form of heat or incident energetic particles or from chemical reaction.
2. Transport of the constituent atoms or molecules to the substrate. From a technological point of view, this is a crucial part of the deposition, because the manner of transport will determine the uniformity of coverage of the substrate. In integrated-circuit fabrication, uniformity of thin films becomes much more difficult to achieve as the size of the substrate wafer increases. The deposition geometry that provided 2% uniformity across a 100-mm wafer is not adequate for the new 300-mm wafers just coming into production.
3. Deposition of atoms or molecules on the substrate. The condition of the substrate surface is the key issue in this phase of the process. If the substrate surface is contaminated with a fraction of a monolayer of impurity atoms, the surface mobility of the arriving film atoms may be reduced to an extent that the crystal orientation of the film bears little relationship to that of the substrate, and the film is polycrystalline. Conversely, if the substrate surface is essentially clean, the arriving film atoms are likely to adsorb in sites related to the substrate orientation, and the film may be oriented or even epitaxial.
4. Postgrowth treatment. In this phase of the process, a prolonged annealing of the film at elevated temperatures may yield a film with superior properties as a result of crystal grain growth in the film.
5. Postgrowth analysis. Modern film analytical techniques have been critical in improving the deposition parameters to in-

crease the quality of the film, in terms of either structure or purity, or both.

3. PHYSICAL METHODS OF DEPOSITION

Physical deposition is a process for transferring atoms or molecules from a source to a substrate that relies on physical methods to produce the vapor species. We consider some typical physical deposition processes along with their advantages and disadvantages. For the time being, we will not be concerned with the ordering of the deposited film, since for a large number of applications, a polycrystalline film is entirely adequate and generally is much simpler and more inexpensive to produce.

3.1 Vacuum Evaporation

Vacuum evaporation is possible for a large range of materials, particularly metals, except for the refractory metals with very low vapor pressures. By heating the material source, such as a crucible containing an Al charge, the vapor pressure of the charge is raised to the point that sublimation or evaporation occurs, depending on whether the charge is solid or molten. The relevant physical constraints are that since the deposition usually occurs in a conventionally pumped vacuum chamber where the base pressure is ~10^{-6} Torr, and since the deposition rate should be large compared to the rate of contamination from the background, the vapor pressure must be large enough to provide a sufficient flux at the substrate; i.e., the pressure of the source material at the substrate must be ~10^{-5} Torr. (We note that in fact there is not an equilibrium pressure of atoms at the substrate but rather a unidirectional flux; however, for simplicity we describe the flux in terms of an effective pressure.) It can easily be shown by geometrical argument that the effective pressure at a substrate a distance L from a source of area A is related to the vapor pressure at the source P by

$$P_{\text{eff}} = PA/4\pi L^2; \tag{4}$$

consequently, for a 1-cm^2 source 10 cm from a substrate, the pressure at the substrate is reduced by a factor of more than 1000. Thus the source temperature should be high enough to provide a vapor pressure of $>10^{-2}$ Torr. From data on the vapor pressure of Al given by Honig (1969), this would require an evaporation temperature of ~1200 °C.

While the discussion above suggests that rather high deposition rates are required to obtain adequate purity of metal films, in actual fact a metal such as Al is highly reactive, and the film deposition occurs over a large area of the vacuum chamber. The reactive Al film acts as a "getter pump" and quickly reduces the reactive gas concentration in the system. Consequently, adequately pure films can be obtained at lower growth rates; however, for convenience it is desirable to have growth rates of ~0.1 to 1 μm/min.

The rate of deposition can be obtained from Eqs. (1) and (4); and using the fact that the increase in thickness per unit area, R_t, is just the rate of mass increase divided by the density ρ leads to

$$R_t = \sqrt{\frac{M}{2\pi k\rho^2}}\left(\frac{P_c}{\sqrt{T}}\right)\frac{A}{4\pi L^2}, \tag{5}$$

where P_e is the equilibrium vapor pressure at the temperature T and M is the molecular weight. Written in this way, it is evident that the first term is material dependent, the second term is temperature dependent, and the third term is geometry dependent. From Eqs. (1a) and (4), with vapor pressure in Torr, this simplifies to

$$R_t = 4.62 \times 10^{-3}\frac{\sqrt{M}}{\rho}\frac{P_e}{\sqrt{T}}\frac{A}{L^2} \text{ cm/s}. \tag{5a}$$

For Al at 1200 °C with a vapor pressure of 10^{-2} in the geometry described above, this results in a deposition rate of 5×10^{-5} cm/s or 3 μm/min. The example is somewhat unrealistic because the distance between source and substrate would be significantly larger in a production evaporator, with a consequent reduction in deposition rate.

Uniformity of film thickness is an important issue in the production of integrated circuits. There are two physical limitations to consider. The first limitation is indicated in Eq. (4), namely, that for a point source of atoms, the flux decreases as the square of the distance. The second limitation is that for a source with a finite area, the flux decreases as the cosine of the angle from the normal to the source; i.e., the substrate "sees" a smaller area of source. In a technology in which many wafers are coated simultaneously, the configuration of source and substrates must be arranged so that all the wafers receive the same deposition flux. This can be accomplished by placing the substrates and the source crucible at positions corresponding to points on the circumference of a sphere, the "planetary configuration." In this configuration the substrates that are closer to the source are also located at increasing angle from the normal to the source, so that the distance and angle effects exactly compensate. Rotation of each substrate then insures that the average flux across the wafer is uniform. Runyan and Bean (1990) provide a good discussion of the methods for uniform deposition.

As the diameter of Si wafers increases, the thickness uniformity of films across a wafer becomes the important issue. This may be improved by rotation of the wafer during deposition and by using multiple deposition sources. However, some effort must be taken to determine empirically the exact angular distribution of atom flux from the sources.

A further issue of considerable importance is the step coverage. An integrated circuit is typically etched into mesas and steps, and the metal interconnects must pass over these features without significant reduction in thickness. Without wafer rotation during deposition, shadowing at step edges will cause thinning. Even more of a problem is the coating of holes with high aspect ratio (defined as the depth-to-width ratio). Shadowing is unavoidable in these structures, and the only way to achieve sufficiently uniform coverage is to rely on surface migration of the deposited atoms to regions of lower coverage. As a result, heating of the substrate during deposition is beneficial to improving the coverage; however, for high–aspect-ratio features it is necessary to use other deposition techniques more suitable for filling holes, as is discussed below.

Some metals have such low vapor pressures that they are difficult to evaporate from a resistively heated crucible. Electron-beam evaporators utilize the energy of a beam of high-energy electrons to heat a small portion of a metallic charge held in a water-cooled hearth to a high temperature. In effect, the molten charge is contained in an unmolten container of the same material, a phenomenon sometimes referred to as "skull melting." While this does solve the crucible problem for these metals, nevertheless, the thermal energy radiated from the hot center of the charge can be excessive.

3.2 Sputter Deposition (Vossen and Cuomo, 1978)

In order to evaporate a material it is necessary to heat it to a temperature where it has a significant vapor pressure. Many materials of interest for electronic films have low enough vapor pressures that it is not practical to attempt to evaporate them, as discussed above. For example, the radiant power emitted by a hot source can cause problems with organic films such as the photoresist films used in lithography. Furthermore, increasing the distance between source and substrate is not useful because both the emitted radiant thermal flux and the deposition flux decrease as the inverse square of the distance.

An alternative method of deposition that avoids these problems is the ejection of atoms or molecules from a target bombarded by energetic ions, a process known as sputtering (*q.v.*) (Anderson and Bay, 1981). In the simplest form, sputtering is obtained by generating a beam of inert-gas ions, accelerating the beam to energies of a few keV, and directing the beam at a target composed of the material to be deposited. Atoms or molecules from the target are ejected onto the substrate (Wehner, 1956, 1957). Sputtering can also be accomplished in a diode configuration where the target is biased negatively to attract the positive ions from a glow discharge. Figure 1 shows these two configurations.

Sputtering has a number of advantages:

1. Nearly any thermally stable material can be sputtered, including compounds or alloys where one component may have a much larger vapor pressure than the other. While the evaporated flux from such a material would have a composition very different from the source be-

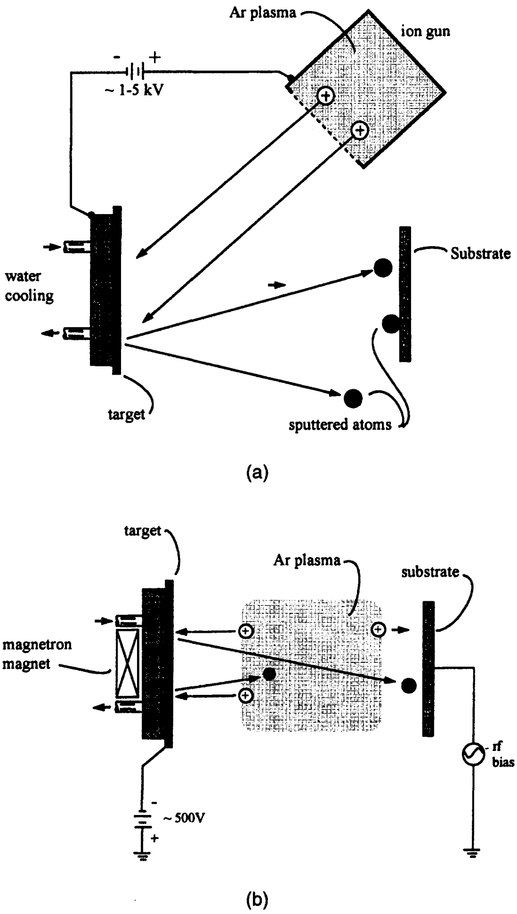

FIG. 1. System configurations for sputter deposition of thin films. (a) Direct ion-beam bombardment configuration. (b) Magnetron diode configuration.

cause of the disparate vapor pressures of the constituents, the sputtered flux from this material after a surface equilibration time will have the same composition as the source target, as shown by Westwood (1989) for example. A simple explanation for this arises from the fact that the sputtering yield from a surface depends only on the immediate surface composition, while the vapor pressures are much more a function of the bulk composition.

2. While the target is heated by the incident ion beam, the power radiated to the substrate is much less than from a hot evaporating source.
3. The arriving atoms or molecules typically have higher kinetic energies than evaporated atoms, which improves step coverage; furthermore, in some methods of sputtering, there may be resputtering of film material on the substrate, which aids in improving step coverage.
4. Particularly for refractory metals or oxides, where evaporation rates are quite low, sputtering can produce reasonable deposition rates.
5. Sputtered films tend to have a finer grain size than do evaporated films.
6. Sputtering targets are readily fabricated by hot-pressed powder sintering; consequently, the composition can be easily controlled.
7. Ion bombardment of the substrate often enhances adhesion of the film.

There are also some disadvantages to sputter deposition of films compared to evaporation:

1. Sputtering requires an inert gas, which must be highly purified to avoid contamination of the film.
2. Sputtered films nearly always contain inclusions of the sputtering gas.
3. There may be ion-bombardment damage of the substrate.
4. Impurities from the fixtures within the system are also sputtered and may be incorporated into the film.

A variety of sputtering configurations are possible, depending on the application, ranging from the ion source–target–substrate configuration discussed above to a parallel-plate configuration of target and substrate where the ions are formed in a glow discharge, as shown in Fig. 1. Most common in manufacturing because of the greater charge density and consequent faster sputtering rate is the planar magnetron, in which the plasma is confined by a magnetic field parallel to the target. Older systems may use a cylindrical magnetron geometry.

A further modification often used to deposit insulating films is called reactive-ion sputtering (Molarius and Orpana, 1991). For example, if a Ta target is sputtered in a mixture of inert gas and N_2, a film comprised of Ta and/or Ta_2N and TaN is produced (Tsai *et al.*, 1992). The exact composition of the film depends reproducibly on the gas composition; thus this technique can be used to produce films of varying electrical resistivity. Reactive sputtering is commonly used to deposit oxides, carbides, and nitrides. It is particularly useful in the deposition of the oxynitrides of Ti, Ta, and Si, where the controlled addition of both oxygen and nitrogen can be used to minimize the tensile stress in the insulating film.

3.3 Ion Plating

Various hybrid techniques combining the best features of both evaporation and sputtering are also common. In one of these, called ion plating by Mattox (1973), a flux of high-energy ions bombards the substrate at the same time that deposition by evaporation is taking place. The ion bombardment assists in cleaning the substrate surface and aids in improving adhesion; at the same time, the "throwing power" or coverage over nonplanar features on the substrate is improved because the evaporating flux is scattered by the inert-gas ambient. The major use of ion plating has been to coat steel and other metals with very hard films for abrasion resistance.

3.4 Ion-Beam Deposition (Harper, 1978)

While less commonly used in manufacturing of integrated circuits, the direct deposition of ion beams onto a substrate has some very useful applications. By ionizing the material to be deposited, focusing the ions into a fine beam, and then directing the beam electrostatically to a desired location, it is possible to deposit small amounts of film material at very precise locations, i.e., to carry out direct "writing" of features. A useful characteristic of this technique is that as the ion-beam energy is increased, the operational mode changes from deposition to ion-beam etching. (At still higher energies, the incident ions become implanted into the target.) Focused–ion-beam systems using field-enhanced ionization from sharpened tips combine the versatility of localized etching or deposition with the optical simplicity of a high-brightness point source of ions. These systems are most generally used for the repair of photomasks or in prototype IC failure analysis. In some cases the ion beam is used to decompose a precursor vapor containing the metal to be deposited as described by Puretz and Swanson (1992).

A different form of ion deposition known as ionized-cluster beam deposition (ICB) has been studied for a number of years by Takagi (1987). In this process an evaporation cell with a small aperture is used to generate a flux of atom clusters, reportedly consisting of as many as 100 metal atoms each. The atom clusters are given a negative charge by a beam of electrons and are accelerated toward the substrate, with the result that each cluster has significant kinetic energy on arrival. It is suggested that this normal kinetic energy is converted to translational kinetic energy on impact, with the result of greater surface mobility and the creation of nucleation sites. While the concept of ICB is interesting, there is much that is not yet clear about the process, including the actual number and size of the proposed clusters, and it has not yet achieved much general acceptance.

3.5 Liquid-Phase Deposition—Langmuir–Blodgett Films (Ulman, 1991; Roberts, 1990)

A very different type of physical deposition takes place when a substrate is dipped into a liquid, usually aqueous, that contains an organic film on the surface. The substrate will adsorb a layer of the film as it is lowered into the solution and a second layer as it is withdrawn. Organic layers that are transferred in this manner from the liquid–gas interface to a solid surface are described as Langmuir–Blodgett (LB) films. The organic molecules that are used in this type of deposition contain two types of functional groups, with one end of the hydrocarbon chain containing an acid or alcohol group, which is soluble in water (hydrophilic), and the other end containing insoluble hydrocarbon groups (hydrophobic). As a result, the molecules form a film on the surface of water (a Langmuir film) with the hydrophilic end at the water side and the hydrophobic end at the air side. Pockels showed that such a film could be compressed to form a continuous monolayer on the liquid (Pockels, 1891; Rayleigh, 1899). Blodgett (1935) subsequently showed that the film adheres to a solid surface passing through the air–water interface.

A unique feature of LB films is the ability to form an ordered structure on a solid surface from a noncrystalline material. When a substrate is passed from the water phase through the film to the air phase (emersion), the hydrophilic group will bond to the solid surface. When the coated substrate is then passed back into the water phase (immersion), the hydrophobic tails in the surface film will bond to the hydrophobic tails exposed on the substrate. This process can be repeated to add additional layers in a head-to-head, tail-to-tail configuration. This type of deposition is described as Y-type. However, if a hydrophobic surface—clean Si is hydrophobic—is passed from air to water, the hydrophobic tails will bond to the surface. The apparatus can be designed to remove the substrate from an uncovered part of the water and reimmerse it in the covered part of the water, thus creating a sequence of head-to-tail layers on the substrate. This method is described as X-type. The significant features are first, that the mode of deposition is easily controllable; second, that the film thickness is precisely controllable in steps of the length of the molecule; and finally, that the X-type depositions are noncentrosymmetric, which has important implications for nonlinear optics applications, as discussed by Gaines (1966). Figure 2 shows the process of Y-type deposition.

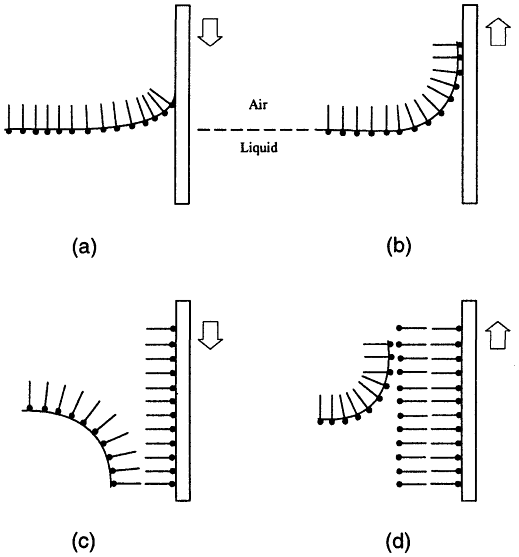

FIG. 2. Deposition of monolayers by the Langmuir–Blodgett method: (a) first immersion, (b) first withdrawal, (c) second immersion, and (d) second withdrawal.

While LB films have a number of potential applications for semiconductor devices, at the present time they are not used in conventional IC manufacturing because they compete with much more mature device systems and processing tools. Questions of long-term reliability have yet to be answered. Applications for which LB films seem to have significant potential include the following:

1. Photoresist technology. LB films can be rapidly and uniformly deposited over large wafer areas at much smaller thicknesses than can conventional spin-on photoresist polymers. Particularly as wafer diameters increase and lithographic features decrease, this could solve problems in photoresist uniformity and resolution.
2. Sensor applications. LB films can incorporate functional groups on the hydrophobic end that react with specific organic or bio-organic molecules. Such films could be incorporated into insulating gates on field-effect transistors or onto quartz crystal oscillators for sensitive measurement of mass change.
3. Flat-panel display applications. LB films have been used in electroluminescent film devices and offer the ability to tailor insulator thickness for optimum performance.
4. Optical devices. A wide range of nonlinear optical materials can be conveniently deposited on appropriate semiconductor substrates by LB techniques. It is easy to envision second-harmonic–generating films on light-emitting diodes and lasers.

The potential applications and recent experiments in this rapidly expanding field are discussed in the book edited by Roberts (1990).

4. CHEMICAL METHODS

Chemical methods of deposition involve a chemical reaction between vapor species containing the element or elements comprising the film. The reaction takes place on a suitable substrate, which may be involved in the reaction. In many cases the substrate acts primarily as a catalyst to mediate the energy involved in the reaction. The energy to initiate the reaction may be thermal or may be provided by rf or optical radiation. Chemical deposition can take place in vacuum or in an ambient consisting of the reactant gases and an inert carrier gas.

4.1 Chemical Vapor Deposition

Chemical vapor deposition (CVD) of thin films occurs when the reacting gas species come in contact with a heated substrate that catalyzes the reaction [see CVD (CHEMICAL VAPOR DEPOSITION)]. The substrate may supply the heat used to pyrolyze the reactant(s), or the entire reaction chamber may be heated. For example, the pyrolysis of silane can be used to deposit Si films (Sherman, 1987):

$$SiH_4(\text{gas}) \leftrightarrow Si(s) + 2H_2(\text{gas}). \tag{6}$$

In fact, this reaction is commonly used in LPCVD (low-pressure CVD) systems operating at ~600 °C and 1–10 Torr to deposit polycrystalline Si used for IC interconnects (Adams, 1985). Deposition rates of some hundreds of angstroms/minute are typical. To make the polysilicon conducting, gases such as diborane, phosphine, or arsine are added to dope the film with B, P, or As, respectively.

Other important CVD reactions are

$$SiH_4(\text{gas})\ O_2 \rightarrow SiO_2(s) + 2H_2 \tag{7}$$

and

$$3SiH_4 + 4NH_3 \rightarrow Si_3N_4(s) + 12H_2 \tag{8}$$

for the deposition of dielectric films.

Refractory metals are difficult to evaporate and are more readily deposited by a CVD process (Schmitz, 1992):

$$WF_6 + 3H_2 \rightarrow W + 6HF. \tag{9}$$

Since CVD reactions are normally carried out at pressures in the range 0.5 to 10 Torr, gas-flow dynamics play a significant role in determining reaction rates and uniformity of the resulting films. A large variety of chamber designs have been used to minimize turbulence and to facilitate the removal of undesirable reaction products (Kern, 1989). We have more to say about CVD under the topic of epitaxial growth. Figure 3 shows some generic reactor schematics.

4.2 Plasma-Enhanced CVD (Lucovsky *et al.*, 1986)

Plasma-enhanced CVD (PECVD) is used extensively to deposit a large variety of films, especially insulators on integrated circuits. Silicon nitride is an important dielectric material used extensively in IC fabrication. At later stages of processing, the circuits are not able to withstand high temperatures. By using a rf plasma to aid in the decomposition, reactions such as that in Eq. (8) can take place with no additional heating of the substrate. Heating by the plasma alone will typically keep the substrate below 300 °C. With the addition of small amounts of N_2O or H_2O, the composition of the dielectric Si_3N_4 can be modified to an oxynitride to minimize the stress in the film.

A recent related approach is to supply energy to the gas mixture by electron cyclotron-resonance (ECR) plasmas, which has the desirable effect of gaining efficient ionization of the source gas at lower pressures, typically 10^{-5}–10^{-3} Torr. This allows efficient deposition with lower power input to the substrate (Herak and Thomson, 1990).

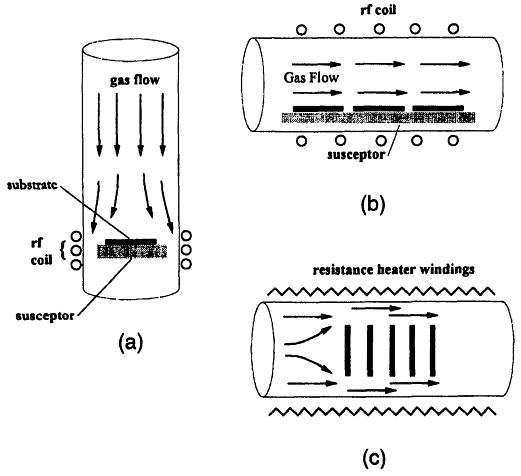

FIG. 3. Generic CVD reactor types: (a) axisymmetric with flow impingement normal to surface, (b) tube furnace with flow parallel to wafer surface, and (c) tube furnace with axial flow.

Finally, decomposition of the reactant gases can be accelerated by use of a focused laser beam, either by a pyrolytic process, in which the substrate is heated by the laser beam, or by a photolytic process, in which the reactants are directly dissociated by energetic photons. This technique is useful for direct "writing" of deposits—e.g., metal lines on prototype circuits. Because direct writing is a slow process, this is not yet a common technique in industry for IC production.

5. EPITAXIAL DEPOSITION

Epitaxy describes the extended single-crystal growth of a film on a crystalline substrate. In early work, the formation and ripening of crystalline nuclei on various substrates such as rock salt was observed and was described as epitaxial growth by Royer (1928); however, such films were highly imperfect on account of defects such as twins formed when the nuclei coalesced into a continuous film. Modern epitaxial films can be extremely perfect depending on the quality of the substrate on which they are grown. For example, growth on whole wafers can be carried out such that the lateral dimensions of the single-crystal film are a few inches while the thickness is only a few monolayers! Epitaxial layers are extremely important in semiconductor devices, because they can be produced with greater purity than bulk material and can provide carrier or potential profiles that are not possible to obtain by other technologies.

Two types of epitaxial growth can be distinguished: *homoepitaxy*, in which the film is grown upon a crystalline substrate of the same material, and *heteroepitaxy*, in which the film and substrate are different. Homoepitaxy has the obvious advantage that the substrate and film have the same lattice constant; provided that the substrate surface is initially clean of contamination, it can serve as a template to facilitate growth of the film as a single crystal. Homoepitaxy is most useful for semiconductor-on-semiconductor growth, in which the film has different properties from the substrate because of a different doping impurity content. For example, the film might have a much lower doping concentration than the substrate. This is not possible to achieve in a diffusion process without carrier compensation, since diffusion must be driven by a larger concentration at the surface.

Heteroepitaxy almost invariably involves mismatch between film and substrate. However, a clean, ordered substrate provides a strongly orienting template for deposited atoms even though this may introduce strain in the film. Various film structures occur depending on the amount of lattice mismatch and the resulting interfacial strain. If the mismatch is small, <0.2%, epitaxial growth can occur without serious difficulty. For larger lattice mismatch, the growth will initially often be "pseudomorphic," in which the film grows with the lattice constant of the substrate, resulting in strain in the plane of the interface. As a pseudomorphic film increases in thickness, the strain energy increases to a point at which misfit dislocations can be generated, leading to a relaxation of strain, with many defects introduced into the film. Frequently, strain relaxation causes the flat, planar epitaxial growth to become rough on an atomic scale. For lattice mismatch >5%, the growth is likely to be relaxed or even polycrystalline (Freund, 1992); however, some growth schemes have been used to accomodate larger mismatch. Figure 4 shows a diagram of the types of epitaxial configurations ranging from lattice matched [Fig. 4(a)] to relaxed mismatched [4(b)] to pseudomorphic mismatched [4(c)]. The pseudomorphic growth [4(c)] also demonstrates the uniaxial compression or expansion in the strained film that occurs normal to the growth direction.

As film thickness increases, so does the total strain energy in a strained film. At

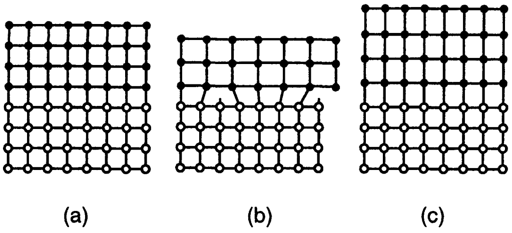

FIG. 4. Schematic cross section of epitaxial growth with varying mismatch: (a) commensurate growth with lattice matching, (b) strain-relaxed incommensurate growth, and (c) pseudomorphic growth with uniaxial expansion normal to growth interface.

some thickness, this energy is reduced by the spontaneous formation of misfit dislocations, as in Fig. 4(b). Calculations and experimental measurements have established the limits for strained growth as a function of film thickness. Figure 5 shows limits of defect-free growth in the Si–Ge system measured by Bean (1986). The solid line is based on calculations originally by Matthews (1975).

The modern key to epitaxial growth has been the clean, ordered substrate on which the film is grown. In the previous sections on deposition, we have avoided much discussion of substrate cleaning, and indeed, evaporated or sputtered films are usually deposited on oxidized semiconductor surfaces that have no long-range order. Here we define "cleaning" as, first, the removal of damage by polishing and etching, followed by removal of *all* impurities that influence the surface ordering. The combination of substrate processing under UHV conditions and the inclusion of surface-analysis tools in growth apparatus has made it possible to prepare and characterize atomically clean substrates, and to preserve the pristine state until growth can occur (e.g., see Fenner *et al.*, 1989).

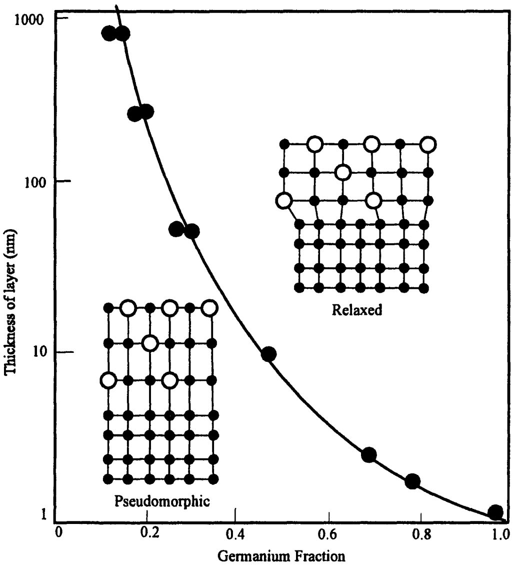

FIG. 5. Experimentally determined thickness limits for defect-free strained-layer epitaxy of Ge_xSi_{1-x} growth on Si. The mismatch is proportional to the Ge fraction, with a maximum of 4.2% at $x = 1$.

5.1 Molecular-Beam Epitaxy (Cho and Arthur, 1975)

Molecular-beam epitaxy (*q.v.*) (MBE) is conceptually the simplest epitaxial process and is also arguably the best understood. MBE occurs by the interaction of beams of atoms and molecules with a clean substrate surface contained in UHV and is discussed more thoroughly in another article in this Encyclopedia; nevertheless, since MBE is a prototypical epitaxial process, we describe it briefly here. Typically the beams are generated by the thermal evaporation of the constituent elements contained in refractory crucibles so arranged that the beams can be turned on or off by opening or closing shutters in the line of sight to the substrate. This allows the growth of a specific material to be initiated or terminated extremely rapidly. The UHV environment prevents contamination of the surface when growth is not taking place and also allows a variety of analytical tools to be available in the vacuum chamber for surface analysis, beam-flux measurement, and film-structure studies. This *in situ* analysis capability has been a major factor in developing an understanding of the process.

Two features of MBE make it especially powerful in the growth of heterostructure devices consisting of thin layers of different semiconductor materials. The UHV environment protects the quality of the growing surface so that contamination does not occur even at low growth temperatures and growth rates. This allows the use of growth temperatures low enough to minimize bulk diffusion processes and thus preserves the abruptness of heterointerfaces grown in the structure. The ability to turn both constituent and dopant beams on and off abruptly with the shutters also facilitates the growth of layered structures where composition or doping level change abruptly across a heterointerface. See Fig. 6 for a schematic view of the MBE growth chamber.

We have mentioned the need for atomically clean substrates for good epitaxial growth. The incorporation of sensitive surface tools such as the Auger-electron spectrometer (AES) in the MBE system makes it possible to determine when the desired state of cleanliness has been achieved. Each substrate material is likely to require a different method of preparation to allow cleaning in

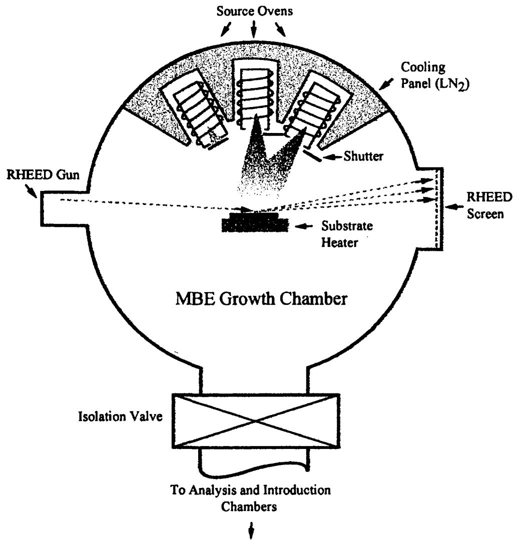

FIG. 6. Schematic top view of an MBE growth chamber showing the source ovens, substrate and heater, RHEED gun and screen, and mass spectrometer.

the growth system. Typically, these include washing in a sequence of etching solutions applied before introducing the substrate into vacuum, followed by a heating cycle to evaporate the residual oxide and to order the clean surface.

Nearly all commercial MBE systems today include equipment for displaying a RHEED (for reflection high-energy electron diffraction) pattern of the substrate to provide an indication in real time of the ordering of the surface. The RHEED pattern thus not only shows when the surface oxide is removed (since it is amorphous) but also shows the improvement in the surface ordering that occurs with the subsequent annealing (Cohen *et al.*, 1986). Clean semiconductor surfaces are reconstructed into a geometric configuration that minimizes the energy of electrons in localized bonds at the surface; this reconstruction is evident in the RHEED patterns as diffraction features positioned between the bulk diffraction spots. The presence of these "fractional order features" provides a qualitative measure of the long-range ordering of the surface (Chadi, 1989).

The most important usage of RHEED information, however, has come from the inference of the mechanism of film growth obtained from the time dependence of the intensity of the diffraction features. On a properly cleaned and annealed (therefore flat) substrate, as growth occurs, the intensity of the diffraction features oscillates in time. Joyce (1985, 1990) has suggested that this is strong evidence for a growth mechanism in which the growth of each atom layer is largely completed before the next begins, i.e., a layer-by-layer growth mechanism (see Frank and van der Merwe, 1949). See Fig. 7 for a schematic explanation of this observation. This growth mechanism is also supported by cross-sectional high-resolution TEM images of the interface between, say, Si and a Si–Ge alloy, which show the interface to be essentially atomically abrupt. Thus it appears that under the best conditions, MBE growth produces nearly ideal interfaces by virtue of the simple 2D layer growth mechanism originally proposed by Burton *et al.* (1951) (BCF). The ideas proposed by BCF have been substantiated with amazing agreement by elegant studies of the MBE growth of Si (Mo *et al.*, 1990) and GaAs (Biegelsen *et al.*, 1990) using scanning tunneling microprobe (STM) analysis.

MBE was initially used for the growth of III–V compound films, specifically GaAs and GaP, where the constituent elements have drastically different vapor pressures; the intent was to solve the vapor-pressure incompatibility with the use of separately controlled sources. In fact, it soon became apparent that the reaction between the metallic and nonmetallic source beams was nearly self-correcting. The vapor pressure of the group-V element was much lower over the stoichiometric compound than over an excess of nonmetal adsorbed on the substrate surface. Hence, a stoichiometric film was obtained by maintaining a slight excess of nonmetal beam flux over the metallic beam flux. The excess group-V element simply re-evaporated. It quickly became apparent that MBE offered a major advantage by virtue of the UHV environment that made possible the inclusion of surface analysis tools such as RHEED for real-time analysis of the growth process. At first it seemed that the semiconductor growth systems were unique in the layer-by-layer growth process that the RHEED analysis indicated. However, in subsequent studies of a wide range of film materials, from semiconductors to metals, RHEED studies indicate that layer-

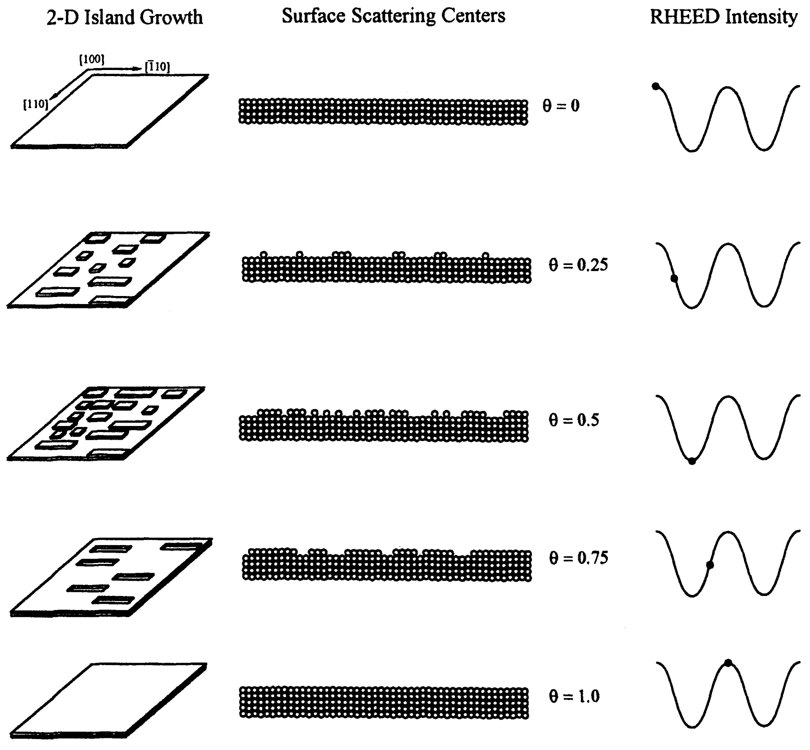

FIG. 7. Real-space representation of the growth of one monolayer, with the corresponding RHEED signal intensity.

by-layer growth is common on clean, low-index crystal faces.

One of the important consequences of the 2D growth is the ability to produce structures consisting of one or more very thin layers of one semiconductor material surrounded by layers of a different material. For device purposes, it is frequently of most interest that the center layer have a band gap smaller than the band gap of the cladding layers. If the thickness of the narrow-band gap center well is much less than the wavelength of electrons in the material, in the range from a few to tens of atom layers, the electron energy levels are discrete in the well, i.e., "quantized," and the structure is described as a quantum well. A structure composed of a periodic sequence of quantum wells is a superlattice. Quantum wells and superlattices are of great interest because the electronic and optical properties are quite different from these of bulk semiconductor material. Many of the theoretical predictions of solid-state physics have been precisely confirmed in these devices; e.g., tunneling of electrons can occur between quantum wells separated by a thin potential barrier of a few hundred angstroms thickness.

Much of the early work on superlattices and quantum wells was based on the material system GaAs-$Ga_xAl_{1-x}As$ because the Al content in the alloy increases the band gap, and yet the two materials have nearly the same lattice constant. The MBE process, at least in principle, should make possible the growth of selected ternary and quaternary III–V alloys, since the composition of the film can be controlled by the relative beam fluxes. Figure 8 shows the lattice constant vs band gap for III–V binary materials and their ternary alloys. It is evident that the lattice constant varies widely for the various alloys. If one is restricted to binary compound substrates, the mismatch between ternary film and binary substrate can be substantial. Thus the issue of growing strained layers is extremely important in order to have access

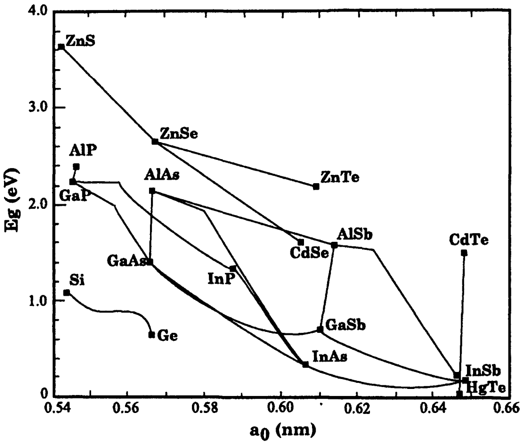

FIG. 8. Relationship between band gap and lattice constant for alloys in the InGaAsP and AlGaAsP systems. Also shown is the wavelength of light associated with the band gap energy.

to the full range of materials and properties offered by the III–V alloy family, as well as other families, notably, Si–Ge.

5.2 Vapor-Phase Epitaxy (Houng, 1992)

Vapor-phase epitaxy (VPE) and the related process, MOCVD, discussed below are chemical deposition processes as opposed to the physical process of MBE. VPE has evolved from the CVD of Si on Si substrates. The chloride process has been extensively used in industry:

$$SiCl_4(gas) + 2H_2(gas) \leftrightarrow Si(solid) + 4HCl(gas). \qquad (10)$$

This reaction is typically carried out at ~1200 °C and is reversible, which means that if the carrier gas—N_2, for example—that initially enters the chamber contains some HCl, the Si surface will be etched to clean the substrate prior to growth. VPE has been used from the early days of Si device processing to prepare thin layers of silicon with properties superior to those of the bulk substrate material; a typical application has been to grow lightly doped Si on a heavily doped substrate.

CVD is typically carried out at pressures in the range from a few Torr to atmospheric pressure. This has the advantage of simplifying the apparatus and increasing the growth rate; however, the complex fluid flow of reactant gases at these pressures is difficult to model. Furthermore, a large number of intermediate molecular species are actually involved in the reaction, as shown in Fig. 9. Thus CVD is mechanistically more complex than the UHV growth processes. An excellent discussion of these issues is given by Hess *et al.* (1985) and is also covered in CVD (CHEMICAL VAPOR DEPOSITION).

Meyerson (1986, 1990) has developed a UHV CVD process that has enabled the growth and fabrication of Si–Ge heterostructure devices with exceptional performance. These devices depend on the electronic properties of Si–Ge alloy quantum wells contained within Si barriers. This process illustrates that with care, it is possible to grow state-of-the-art films in a system designed for production.

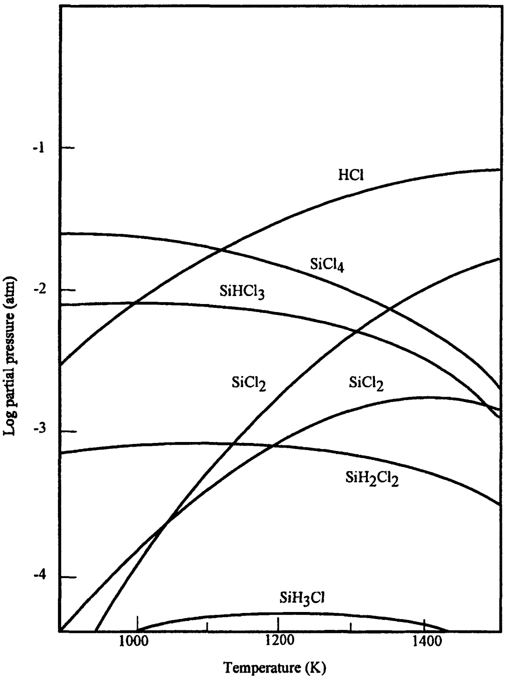

FIG. 9. Equilibrium partial pressures in the Si–Cl–H system at 1 atm and a Cl-to-H ratio of 0.06. The figure indicates the chemical complexity of CVD growth.

5.3 Metal-Organic CVD (Stringfellow, 1989)

Metal-organic CVD (MOCVD) is a process for the growth of III–V semiconductors that combines some of the best features of MBE and CVD. Rather than relying on an elemental source to provide the metal constituent, MOCVD utilizes a volatile organic compound to provide a vapor source that does not require the high-temperature cell used in MBE. Typically, organometallic compounds such as trimethyl gallium are used. The volatility of these species reduces the thermal load in the vacuum system and provides larger fluxes for faster growth; and, since the metal organic is introduced from a source outside the vacuum chamber, there is no need to open the system for replenishing the supply of metal. MOCVD also has made control of the composition of ternary and quaternary materials easier, since it is easier to control the partial pressures of several metal-organic sources than to control the temperature of elemental sources. A variation of MOCVD uses vapors of nonmetal compounds as well; for example, AsH_3 has been used as a source of As in the growth of GaAs (Panish, 1980). With beam sources that rarely need replenishing and an interlock system that allows substrate entry and exit from the growth chamber without impairing the vacuum, the MOCVD system can operate for very long periods without down time.

While MOCVD was first used in the late 1960s with systems operating at atmospheric pressure that were essentially identical to CVD systems of the period, present systems are much more closely related to MBE systems and operate at reduced pressure, or even at UHV in a molecular-beam mode. It has been possible to include RHEED display in UHV growth systems, and the observations of RHEED oscillations during growth are nearly identical to those observed in MBE and suggest a similar layer-by-layer growth mechanism.

Since hydrocarbons are normally to be avoided in UHV systems because of the difficulty in removing C contamination from clean surfaces, it is not surprising that early MOCVD films were rather heavily contaminated with C. It has since been found, however, that the C contamination can be essentially eliminated by adjustment of the group-V beam intensity. As a result, MOCVD produces very high quality III–V films, with the advantage of high throughput due to the faster growth rates possible.

5.4 Liquid-Phase Epitaxy (Nakajima, 1985)

Liquid-phase epitaxy (LPE) involves the growth of an epitaxial film on a substrate that is placed in contact with a saturated solution of the film material. In some cases the solvent may be one of the film constituents; e.g., GaAs films are grown from a saturated solution of As in molten Ga. When the solution is cooled very slightly it becomes supersaturated and growth occurs on the substrate. LPE is essentially an equilibrium process, where knowledge of the thermodynamic parameters allows a prediction of the composition of the film. Cleaning of the substrate surface is accomplished by raising the temperature of the solution briefly to make it unsaturated with respect to the substrate material causing some etching prior to growth. The advantages of LPE are the following:

1. Simplicity. The growth chamber is typically glass, and no complex pumping or valving is involved since growth occurs at or near atmospheric pressure in an atmosphere of purified H_2.
2. Near-equilibrium growth. Since diffusion rates in solution are rapid, LPE growth is rapid, >0.1 μm/min; consequently, fairly thick films can easily be deposited.
3. Films grown from a liquid-metal solution can be extremely pure because the distribution ratio of impurities between the solid (film) phase and the reactive liquid is low; in effect the liquid "getters" impurities, and the resulting films are quite pure, even though not grown in a clean vacuum environment.

However, there are enough disadvantages that LPE is not much used industrially:

1. Only small substrates can be used. Whole-wafer growth is not normally possible, and this limitation keeps LPE a research technique.
2. The composition of the film depends on the concentration of material in the liquid solution. Typically small volumes of solution are used, and the concentration of

film material may be small enough so that composition of the film changes as the liquid is depleted of solute.

3. Control of film thickness is difficult since the growth rate is an extremely sensitive function of solution temperature; as a result, interfaces between layers can be rough.

LPE played a very significant role in the development of III–V devices in the 1960s, and it is still used for research applications because of the high optoelectronic quality of the layers produced; however, the film uniformity and thickness are not sufficiently under control for modern device fabrication.

5.5 Solid-Phase Epitaxy

Solid-phase epitaxy describes the regrowth of a polycrystalline or amorphous film on a crystalline substrate at an elevated temperature. The most common example of solid-phase epitaxy is the regrowth of a semiconductor that has been ion implanted to an extent that the surface has been amorphized by the ion flux. Annealing causes regrowth from the unimplanted region outward. Other instances can involve the use of scanned laser beams or other optical sources to carry out a rapid annealing of the surface of a wafer on which a noncrystalline layer has been deposited. In most cases, solid-phase epitaxy is a homoepitaxial process, to recrystallize the surface of a semiconductor.

5.6 Atomic-Layer Epitaxy (Suntola, 1989)

Atomic-layer epitaxy (ALE) is a variation of MBE but with some unique modifications. In the II–VI family of semiconductors, e.g., ZnS, the vapor pressure of each of the pure elements is considerably higher than the partial pressure of each over the compound. ALE makes use of this fact to obtain stoichiometric growth by exposing the heated substrate alternatively to beams or vapors of the constituent elements. The substrate is heated to a temperature hot enough to evaporate the elements but not hot enough to vaporize the compound. Thus an exact monolayer of each element is adsorbed during each half cycle, sufficient to react with the monolayer below. ALE has been shown to produce very good-quality II–VI films, particularly for luminescent devices. However, the ALE process is often utilized in nonepitaxial film growth because it provides excellent conformal growth; i.e., the coverage over steps and in wells is not dependent on line-of-sight beam flux from a source, which can cause shadowing effects. Rather, the higher pressures of vapors used in the process penetrate into holes on nonplanar surfaces, and yet the stoichiometry of the film is preserved by the physics of the process.

6. SURFACE SCIENCE AND THIN-FILM DEPOSITION

The importance of the relationship between our understanding of surfaces and the technology of thin-film deposition cannot be overstated. The tools for studying surfaces developed during the past 25 years have played a crucial role in moving from polycrystalline films to epitaxial semiconductors that are more perfect than their substrates. In particular, the spectroscopic methods for determining surface composition—e.g., AES, XPS, SIMS—and the structural characterization tools, RHEED and now the STM, have opened the door to an understanding of how the various deposition methods influence the films that they produce (Czanderna, 1975). Figure 10 is an STM image of the stepped

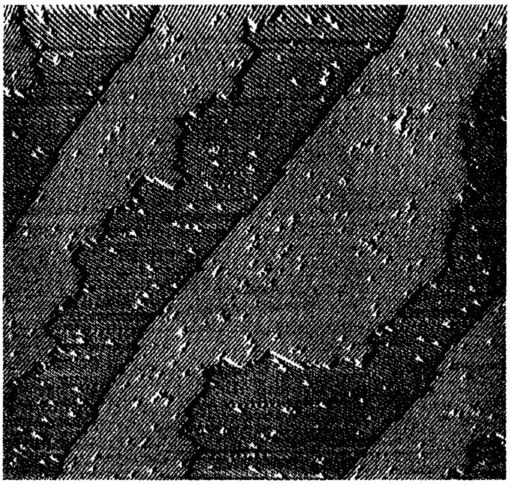

FIG. 10. STM image of stepped Si surface after MBE growth. Surface is oriented slightly off the [100] direction (~1°). Note the alternating irregular and smooth step edges.

surface of a Si layer that has been grown by MBE on a Si substrate oriented slightly off the [100] direction. The striking feature is how closely the actual surface, with its steps and nearly uniform terraces, resembles the idealized model proposed by Burton, Cabrera, and Frank some 45 years ago (Burton *et al.*, 1951).

GLOSSARY

AES: Auger-electron spectroscopy.

Aspect Ratio: The ratio of vertical dimension to lateral dimension of a feature on an integrated circuit.

Chemical Deposition: The process of depositing a thin film involving a chemical reaction.

Cryopump: A vacuum pump that condenses gas molecules on cryogenically cooled walls. Has a high pumping speed for condensable gas species and can be used to below 10^{-10} Torr.

Diffusion Pump: A vacuum pump that entrains gas in a vapor stream derived from the boiling pump fluid. Normally used down to $\sim 10^{-7}$ Torr, although lower pressures can be achieved.

Electroluminescent Film: A semiconducting film structure that emits visible light when excited by an electric field, typically from an alternating voltage.

Epitaxy: The growth of a thin film with an ordered relationship to the substrate.

Evaporation: The process of transferring material by the vaporization of a source at high temperature to a substrate at lower temperature.

Heteroepitaxy: Epitaxial growth of a film on a substrate of dissimilar material.

Homoepitaxy: Epitaxial growth of a film on a substrate of the same material.

Hydrophilic: Attracted to water; describing that portion of an organic molecule that is soluble in aqueous solution.

Hydrophobic: Describing the hydrocarbon tail of an organic molecule that is insoluble in aqueous solution.

Ion Pump: A vacuum pump that ionizes gas molecules and drives them into the pump electrodes. Useful down to $\sim 10^{-11}$ Torr.

Lattice Mismatch: A measure of the difference between the lattice constants of film and substrate. Defined as $f = [a_0(\text{sub}) - a_0(\text{film})]/a_0(\text{film})$, where a_0 represents the lattice constant of film or substrate.

Light-Emitting Diode: A semiconducting diode structure that emits band-gap light when electrically driven in the forward direction.

Low-Dimensional Structure: A thin film containing one or more layers that produce quantum wells with carrier confinement in one, two, or three dimensions; i.e., having one or more dimensions of the order of the electron wavelength.

Mesa Etching: Chemical etching during the processing of integrated circuits to leave a mesa structure on the semiconductor wafer. Typically done to electrically isolate a single device in the mesa.

Metallization: The process of covering a substrate with a metallic thin film.

Monolayer: A thin film that is exactly one molecule thick. For most materials this amounts to about 3×10^{14} molecules/cm^2.

Photolytic: Dissociated by the energy of incident photons.

Photoresist Film: An organic film that is light sensitive and that is used to define a region on an integrated circuit using photolithography.

Physical Deposition: The process of depositing a thin film by physical means, such as evaporation.

Polysilicon: A nearly glassy form of polycrystalline silicon with very small microcrystals.

Pseudomorphic Growth: Heteroepitaxial growth of a film on a substrate in which the film assumes the lattice constant of the substrate.

Pyrolytic: Dissociated by a thermal process.

Quantum Well: A thin film containing a region that confines carriers in a potential well whose dimensions in at least one direction are comparable to the wavelength of an electron or hole in the region.

Reactive-Ion Sputtering: Sputtering in the presence of a gas that reacts chemically with the sputtered particles.

RHEED: Reflection high-energy electron diffraction.

SIMS: Secondary-ion mass spectroscopy.

Sputtering: The process of using high-energy ions to remove material from a target. In many cases, the material removed is transferred to a substrate; however, in some cases sputtering is used simply for etching the target.

STM: Scanning tunneling microscopy.

Substrate: The solid material on which a thin film is deposited. It may be crystalline or not.

Superlattice: A film structure consisting of a periodic stacking of dissimilar films, with spatial dimensions of the order of the electron mean free path.

TEM: Transmission electron microscopy.

Torr: A unit of pressure equal to 1 mm Hg or to 133.3 Pa.

UHV (ultrahigh vacuum): A vacuum in which the pressure is considerably less than 10^{-7} Torr.

XPS: X-ray photoelectron spectroscopy.

Works Cited

Adams, A. C. (1985), "Dielectric and Polysilicon Film Deposition," in: S. M. Sze, Ed., *VLSI Technology,* New York: McGraw-Hill.

Allan, G., Bastard, G., Boccara, N., Lannoo, M., Voos, M. (Eds.) (1986), *Heterojunctions and Semiconductor Superlattices,* Berlin: Springer-Verlag.

Anderson, H. H., Bay, H. L. (1981), in: I. R. Behrisch (Ed.) *Sputtering by Particle Ion Bombardment,* Berlin: Springer-Verlag.

Bean, J. C. (1986), *Phys. Today* **39** (10), 36–42.

Biegelsen, D. K., Bringans, R. D., Northrup, J. E., Swartz, L.-E. (1990), *Phys. Rev. B* **41,** 5701–5706.

Blodgett, K. B. J. (1935), *J. Am. Chem. Soc.* **57,** 1007–1022.

Burton, W. K., Cabrera, N., Frank, F. C. (1951), *Philos. Trans. R. Soc. London* **A243,** 299–358.

Campbell, Stephen A. (1996), *The Science and Engineering of Microelectronic Fabrication,* New York: Oxford Univ. Press.

Chadi, D. J. (1989), *Ultramicroscopy* **31,** 1–9.

Cho, A. Y., Arthur, J. R. (1975), *Prog. Solid-State Chem.* **10,** 157–191.

Cohen, P. I., Pukite, P. R., Van Hove, J. M., Lent, C. S. (1986), *J. Vac. Sci. Technol. A* **4,** 1251–1258.

Czanderna, A. W. (Ed.) (1975), *Methods of Surface Analysis,* Amsterdam: Elsevier.

Feldman, L. C., Mayer, J. W. (1986), *Fundamentals of Surface and Thin Film Analysis,* Amsterdam: North-Holland.

Fenner, D. B., Biegelsen, D. K., Bringans, R. D. (1989), *J. Appl. Phys.* **66,** 419–424.

Frank, F. C., van der Merwe, J. H. (1949), *Proc. R. Soc.* **A198,** 205–225.

Freund, L. B. (1992), *Scripta Met. Mater.* **27,** 669–674.

Gaines, G. L., Jr. (1966), *Insoluble Monolayers at Liquid Gas Interfaces,* New York: Wiley.

Hablanian, M. H. (1990), *High Vacuum Techniques: A Practical Guide,* New York: Marcel Dekker.

Harper, J. M. E. (1978), "Ion Beam Deposition," in: J. L. Vossen, W. Kern (Eds.), *Thin Film Processes,* New York: Academic, p. 175.

Herak, T. V., Thomson, D. J. (1990), *J. Appl. Phys.* **67,** 6347–6352.

Hess, D. W., Jensen, K. F., Anderson, T. J. (1985), "Chemical Vapor Deposition: A Chemical Engineering Perspective," *Rev. Chem. Eng.* **3-2,** 97–147.

Honig, R. E. (1969), *RCA Rev.* **30,** 285–305.

Houng, Y.-M. (1992), "Chemical Beam Epitaxy," *Crit. Rev. Solid State Mater. Sci.* **17,** 277–290.

Joyce, B. A. (1985), *Rep. Prog. Phys.* **48,** 1637–1697.

Joyce, B. A. (1990), *J. Crystal Growth* **99,** 9–36.

Kern, W. (1989), in: R. A. Levy (Ed.), *Microelectronic Materials and Processes,* Dordrecht: Kluwer Academic.

Lucovsky, G., Richard, P. D., Tsu, D. V., Lin, S. Y., Markunas, R. J. (1986), *J. Vac. Sci. Technol. A* **4,** 681–688.

Matthews, J. W. (1975), "Coherent Interfaces and Misfit Dislocations," in: J. W. Matthews (Ed.), *Epitaxial Growth Part B,* New York: Academic.

Mattox, D. M. (1973), *J. Vac. Sci. Technol.* **10,** 47–52.

Meyerson, B. S. (1986), *Appl. Phys. Lett.* **48,** 797–799.

Meyerson, B. S. (1990), *IBM J. Res. Dev.* **34,** 806–815.

Mo, Y.-W., Kariotis, R., Swartzentruber, B. S., Webb, M. B., Lagally, M. G. (1990), *J. Vac. Sci. Technol. A* **8,** 201–206; *ibid B* **8,** 232–236.

Molarius, J. M., Orpana, M. (1991), in: S. Coffa, F. Priolo, E. Rimini, J. M. Poate (Eds.), *Issues in Semiconductor Materials and Processing Technologies,* Dordrecht, Kluwer.

Nakajima, K. (1985), "The Liquid-Phase Epitaxial Growth of InGaAsP," in: W. T. Tsang (Ed.), *Semiconductors and Semimetals,* Vol. 22, New York: Academic.

O'Hanlon, J. F. (1989), *User's Guide to Vacuum Technology,* 2nd Ed., New York: Wiley.

Ohring, Milton (1992), *The Materials Science of Thin Films,* Boston: Academic.

Panish, M. B. (1980), *J. Electrochem. Soc.* **127,** 2729–2733.

Pockels, A. (1891), *Nature (London),* **43,** 437–439.

Puretz, J., Swanson, L. W. (1992), *J. Vac. Sci. Technol.* **B10,** 2695–2698.

Lord Rayleigh (1899), *Philos. Mag.* **48,** 321.

Roberts, Gareth (Ed.) (1990), *Langmuir-Blodgett Films,* New York: Plenum Press.

Royer, L. (1928), *Bull. Soc. Fr. Mineral Cristallogr.* **51,** 7–11.

Runyan, W. R., Bean, K. E. (1990), *Semiconductor Integrated Circuit Processing Technology,* Reading, MA: Addison-Wesley.

Schmitz, J. E. J. (1992), *Chemical Vapor Deposition of Tungsten and Tungsten Silicides,* Park Ridge, NJ: Noyes.

Sherman, A. (1987), *Chemical Vapor Deposition for Microelectronics: Principles, Technology, and Applications,* Park Ridge, NJ: Noyes.

Smith, Donald L. (1995), *Thin-Film Deposition, Principles and Practice,* New York: McGraw-Hill.

Stringfellow, G. B. (1989), *Organometallic Vapor-Phase Epitaxy,* Boston: Academic.

Suntola, T. (1989), "Atomic Layer Epitaxy," *Mater. Sci. Rept.* **4,** 261–312.

Takagi, T. (1987), in: M. H. Francombe, J. L. Vossen (Eds.), *Physics of Thin Films,* Vol. 13, New York: Academic.

Tsai, W., Fair, J., Hodul, D. J. (1992), *J. Electrochem. Soc.* **139,** 2004–2007.

Ulman, Abraham (1991), *An Introduction to Ultrathin Organic Films from Langmuir–Blodgett to Self-Assembly,* Boston: Academic.

Vossen, J. L., Cuomo, J. J. (1978), in: J. L. Vossen, W. Kern (Eds.), *Thin Film Processes,* New York: Academic.

Wehner, G. K. (1956), *Phys. Rev.* **102,** 690–704.

Wehner, G. K. (1957), *Phys. Rev.* **108,** 35–45.

Westwood, W. D. (1989), in: R. A. Levy (Ed.), *Microelectronic Materials and Processes,* Dordrecht: Kluwer Academic.

Further Reading

Thin Film Deposition

Bunshah, R. F. (1994), *Handbook of Deposition Technologies for Films and Coatings,* 2nd ed., Park Ridge, NJ: Noyes.

Fraser, D. B. (1978), in: J. L. Vossen, W. Kern (Eds.), *Thin Film Processes,* New York: Academic.

Glang, R. (1970), in: L. I. Maissel, R. Glang (Eds.), *Handbook of Thin Film Technology,* New York: McGraw-Hill.

Glocker, D. A. (1995), *Handbook of Thin Film Process Technology,* Philadelphia: Institute of Physics Publishing.

Holland, L. (1956), *Vacuum Deposition of Thin Films,* New York: Wiley.

Kern, W., Ban, V. S. (1978), in: J. L. Vossen, W. Kern (Eds.), *Thin Film Processes,* New York: Academic.

Maissel, L. M., Francombe, M. H. (1973), *An Introduction to Thin Films,* New York: Gordon and Breach.

Matsuo, S., Kiuchi, M. (1982), in: *Proceedings of the Symposium on VLSI Science Technology,* Electrochemical Society, Pennington, NJ.

Matthews, J. W. (Ed.) (1975), *Epitaxial Growth Part A,* New York: Academic.

Thornton, J. A., Penfold, A. S. (1978), in: J. L. Vossen, W. Kern (Eds.), *Thin Film Processes,* New York: Academic.

Vossen, J. L., Kern, W. (Eds.) (1978), *Thin Film Processes,* New York: Academic.

Vacuum Technology

Delchar, T. A. (1993), *Vacuum Physics and Techniques,* London: Chapman and Hall.

Dushman, S. (1962), *Scientific Foundations of Vacuum Techniques,* New York: Wiley.

Roth, A. (1976), *Vacuum Technology,* Amsterdam: North-Holland.

Molecular Beam Epitaxy

Cho, Alfred Y. (1994), *Molecular Beam Epitaxy,* Woodbury, NY: American Institute of Physics.

Herman, M. A., Sitter, H. (1989), *Molecular Beam Epitaxy,* Berlin: Springer-Verlag.

Panish, M. B., Temkin, H. (1993), *Gas Source Molecular Beam Epitaxy,* Berlin: Springer-Verlag.

Parker, E. H. C. (1985), *The Technology and Physics of Molecular Beam Epitaxy,* New York: Plenum Press.

Tsao, J. Y. (1992), *Materials Fundamentals of Molecular Beam Epitaxy,* Boston: Academic.

Surface Science and Characterization

Briggs, D., Seah, M. P. (1984), *Practical Surface Analysis by Auger and Photoelectron Spectroscopy,* New York: North-Holland.

Brundle, C. R., Evans, Jr., C. A., Wilson, S. (Eds.) (1992), *Encyclopedia of Materials Characterization: Surfaces, Interfaces, Thin Films,* Boston: Butterworth-Heinemann.

Chen, C. J. (1993), *Introduction to Scanning Tunneling Microscopy,* New York: Oxford Univ. Press.

Howie, A., Valdre, U. (Eds.) (1988), *Surface and Interface Characterization by Electron Optical Methods,* New York: Plenum Press.

Nizzoli, F., Rieder, K. -H., Willis, R. F. (Eds.) (1985), *Dynamical Phenomena at Surfaces, Interfaces and Superlattices,* Berlin: Springer-Verlag.

THIN-FILM INTEGRATED FERROELECTRICS

L. ERIC CROSS AND S. TROLIER-MCKINSTRY, *Intercollege Materials Research Laboratory, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.*

Introduction 429
1. **Needs for Ferroelectric Films on Silicon** 429
1.1 DRAM Applications 429
1.2 FRAM Applications 430
1.3 IR Thermal Imaging 430
1.4 Mechanics 431
1.5 Packaging 431
1.6 Microchemistry 431
2. **Ferroelectric Material Systems** 431
2.1 Properties of Bulk Materials 431
2.2 Symmetry Considerations 432
2.3 Domains and Polarization Switching 433
2.4 Transport Properties 434
2.5 Perovskite-Structure Systems 435
2.6 Bismuth Oxide Layer Structures 438
2.7 Alternative Systems 439
2.8 Relaxor Ferroelectric Systems 439
3. **Ferroelectric Film Preparation** 440
3.1 Boundary Conditions for Films on Silicon 440
3.2 Deposition Methods 441
3.3 Vapor-Phase Techniques 441
3.4 Liquid-Phase Techniques 441
4. **Characterization of Film Microstructure** 442
4.1 Introduction 442
4.2 Direct Methods 442
4.3 Optical Methods 443
5. **Properties of Films on Silicon** 443
5.1 High-Permittivity Films (DRAM) 443
5.2 Hysteretic Films (FRAM) 444
5.3 Pyroelectric Films 446
5.4 High–Energy-Density Transducers 446
6. **Conclusions** 447
Glossary 448
Works Cited 448
Further Reading 451

INTRODUCTION

The recent explosion of research on ferroelectric thin films has led to significant advances in the processing and properties of these materials. A major driving force for this research is the advantages that accrue from integrating ferroelectric elements with integrated circuits (ICs) (especially silicon-based electronics). In particular, ferroelectric films can be optimized for high dielectric constants, hysteretic polarization behavior, and pyroelectric, electromechanical, or electro-optic properties. In this article, the potential uses of ferroelectric films, the structure and properties of ferroelectrics, the deposition methods for films, the properties of ferroelectric films on silicon, and the current status of research around the world are reviewed.

1. NEEDS FOR FERROELECTRIC FILMS ON SILICON

1.1 DRAM Applications

One crucial need for next-generation high-density memory circuits is the development of high–dielectric-constant, low-leakage films that are compatible with silicon circuitry for ultracompact charge storage. Until recently, the standard capacitors used in integrated circuitry were either SiO_2 or silicon oxynitride thin films. However, the dielectric constants of both of these materials are quite

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

low ($\epsilon_r \sim 4$ and ~ 9, respectively). Thus, as the linewidths of active devices have decreased, it has been hard to achieve a comparable reduction in capacitor size, even through the use of extremely thin layers or trench geometries. The situation is particularly severe in dynamic random-access memories (DRAM). Consequently, it is becoming imperative to integrate higher–dielectric-constant materials with Si circuitry. As a first attempt at this, several companies have switched to Ta_2O_5 ($\epsilon_r \sim 18$–25), but for extremely high integration levels, an even higher dielectric constant will eventually be required. For this purpose, considerable research is now being directed at development of low-loss paraelectric thin films.

1.2 FRAM Applications

The possibility of developing a high-density nonvolatile ferroelectric random-access memory (FRAM), in which the polarization state of the ferroelectric thin film is used to store information, has been one of the major drivers for the explosion of effort in ferroelectric thin films. Because intrinsically the switching in ferroelectric films is extremely fast (<100 ps) (Larsen, 1995), a ferroelectric-based memory might ultimately challenge existing technologies for computer applications. In the short term, however, FRAMs are becoming competitive with other nonvolatile information storage systems on silicon such as EEPROM (electrically erasable programmable read-only memory). Potential large-volume markets include the information storage systems on "smart cards."

1.3 IR Thermal Imaging

In ferroelectric crystals and films the strong temperature dependence of the spontaneous electric polarization leads to the possibility of very strong pyroelectric effects, as the pyroelectric coefficient p is defined by

$$p = \frac{\partial P_s}{\partial T}, \tag{1}$$

where P_s is the polarization and T is the temperature.

For a thin sheet of crystal or ceramic, poled normal to the sheet and blackened on one surface, absorbed long-wavelength infrared radiation will change the temperature, and the pyroelectric current may be used to monitor the radiation level. Such current-mode detectors can be very fast and are usable as power meters for pulsed CO_2 lasers working at 10 μm. If the radiation is chopped at a frequency $\omega \sim 30$ Hz, and the element is thin and very well thermally isolated, the thermal sensitivity is given by

$$\frac{\Delta T}{W} = \frac{\eta}{(g^2 + \omega^2 c^2 t^2)^{1/2}}, \tag{2}$$

where ΔT is the temperature rise for incident IR energy W, a fraction η of which is absorbed by the detector chip; g is the effective thermal conductance to the environment; c is the volume specific heat; and t is the detector thickness. For high sensitivity the cardinal importance of low g and very small t for any given frame frequency ω is clearly apparent.

In isolated point detectors it is easy to provide excellent thermal isolation by mounting the element on an ultrathin polymer "drum head," and very inexpensive, high-sensitivity thermal IR detectors made from a wide range of ferroelectric materials are commonly used in industry.

For imaging, where the individual element is down to a pixel size of ~25 μm, the conflicting requirements of electrical interconnection and refined thermal isolation provide challenging problems. Present systems use hybrid technology in which the detector plane is a very thin sheet of poled ferroelectric ceramic, bump bonded to the silicon integrated interrogation circuit. With an $f1$ lens system, such imaging systems give resolution of 0.07 K in the object plane (Kulwicki *et al.*, 1992).

Currently, there is a strong push to move to integrated systems using microelectromechanical technology to form air-bridge or diaphragm thermal isolation on the silicon chip itself (Shorrocks *et al.*, 1995; Kohli *et al.*, 1995; Watton, 1994). Major problems concern the conflicting processing needs of ferroelectric films and silicon microcircuits, but the prize would be a system that could be fitted in automobiles to provide drivers with useful night vision.

1.4 Mechanics

Piezoelectricity and electrostriction entail a coupling between electrical and mechanical energy in materials. That is, if a stress is applied to the material, it responds with an electrical polarization. Conversely, if an electric field is applied, the material undergoes a strain. Piezoelectric materials find widespread use in both sensor and actuator applications. A significant research effort is now being directed toward miniaturization of piezoelectric devices for microelectromechanical systems (MEMS). For these applications, the capability of processing thin-film piezoelectrics such as $Pb(Zr_xTi_{1-x})O_3$ (PZT) on platinum-coated silicon allows for the integration of the active sensing or actuator elements directly with on-chip driving or receiving electronics (including all of the necessary amplification). This can significantly increase the sensitivity of devices such as microphones and hydrophone detectors (Kim, 1990; Kim and Muller, 1987). From the standpoint of actuation, it is clear that as devices are miniaturized, piezoelectric devices permit comparatively higher energy densities, compared to other means of generating displacement. Thus, as active elements are integrated, electromechanical devices become progressively more attractive.

1.5 Packaging

Ferroelectrics are widely used as high–volumetric-efficiency capacitors in the bulk form. Thus, it is not surprising that there is interest in developing thin-film ferroelectrics as decoupling capacitors for integrated-circuit applications (Schwartz *et al.*, 1994). In particular, in some applications it is possible to replace chip capacitors with higher-density thin-film integrated capacitors in order to decrease circuit noise. For example, Matsushita has reported use of $Ba_{1-x}Sr_xTiO_3$ films in GaAs MMICs for portable telephones and in conjunction with CCDs for VCRs and camcorders (Shimada *et al.*, 1995). Required properties for films used in these applications are a high dielectric constant at the use frequency, a low dielectric loss, and a high resistivity.

1.6 Microchemistry

A specialized area of MEMS that is receiving rapidly increasing interest is that associated with chemistry and biochemistry on a chip. For a number of important liquid-phase chemical reactions, there are accessible regions of the boundary conditions where the reactions can become unstable and uncontrollable. By splitting the reactants into very small reaction chambers that can be replicated by inexpensive semiconductor technology and where sensing and control circuits may be on the same chip, control time constants are drastically shortened ensuring better control, but even in the extreme case of a "runaway," the volumes involved ensure that the result is not catastrophic (Jensen, 1996).

For many biological problems it is essential to work with microliter volumes and below, so that the micropumps and microvalves accessible in MEMS become essential components. DNA sequencing by electrophoretic capillary separation is just one example where widespread use in the criminal justice system is pressing toward massive markets (Sonnenschein, 1996).

In the microchemical–biochemical systems the piezoelectric/electrostrictive components are primarily for mechanical actuation (pumps and valves), pressure flow, and temperature sensing.

2. FERROELECTRIC MATERIAL SYSTEMS

2.1 Properties of Bulk Materials

Ferroelectricity is a property of crystals; there are no ferroelectric glasses or liquids. The appearance of a polarization vector in a crystal will certainly modify the point symmetry and create a new phase. If the resulting phase is to be ferroelectric, there must be more than one equivalent equilibrium orientation for the polar vector, and it must be switchable between states by a realizable electric field.

The electrical consequence of the reorientable spontaneous electric polarization is the appearance of dielectric hysteresis in the relation between dielectric polarization and applied electric field at higher field levels

(Fig. 1). This behavior is strongly reminiscent of the ferromagnetic hysteresis between intensity of magnetization and magnetic field and led to the adoption of the name ferroelectric for these materials. From the hysteresis loop, the terms remanent polarization and coercive field are defined (Fig. 1) also by analogy with the magnetic systems. As with magnets, many ferroelectrics exhibit a phase change to a nonferroelectric (paraelectric) form at higher temperature, and the transition point is defined as the Curie temperature (T_c). Similarly above T_c one of the diagonal elements of the permittivity tensor often follows a Curie–Weiss relation of the form

$$\epsilon_{ii} = c/(T - \theta), \qquad (3)$$

where θ is the Curie–Weiss temperature. However, in the dielectric case, the Curie constant c is 5 to 7 orders of magnitude larger than in the corresponding magnetic case, indicating the much "softer" ordering (weaker interaction) in the ferroelectric and leading to very high values of dielectric permittivity in the paraelectric phase above T_c. Also, unlike magnetism, the remanent polarization of a ferroelectric does not give rise to a strong permanent electric field in the surrounding medium, as the bound dipolar charge at the surface is always eventually compensated by the free charges that are available in the electrical case.

2.2 Symmetry Considerations

All point symmetries other than 1, which has no symmetry to lose, may be prototypic (paraelectric) phases out of which ferroelectric species may occur. A complete cataloging of all possibilities has been given by Shuvalov (1970) and Aizu (1969), and compact notations have been presented for describing the state shifts. The tabulation for the cubic point groups is given in Table 1, adapted from Shuvalov. Taking, for example, one of the common species from point group *m*3*m*, the symbol

$$m3m\ (3)\ \mathrm{D4F}\ 4mm \qquad (4)$$

means that prototype *m*3*m* has a ferroelectric state (F) that will have three (3) equivalent orientations (six vector directions) which are definite (D) along the fourfold axis (4) giving rise to six domain states, each in tetragonal point symmetry 4 *mm*.

A consequence of the general symmetry

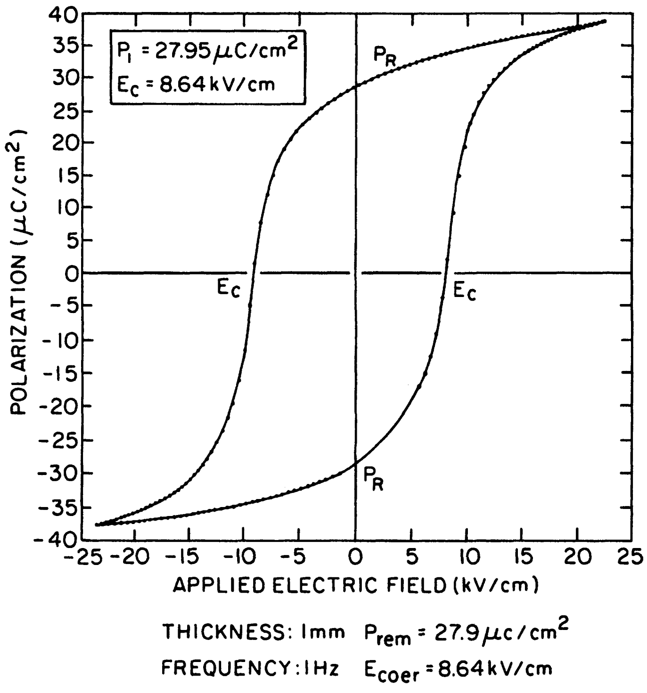

FIG. 1. Typical hysteresis between electric polarization and electric field in a ferroelectric material in its ferroelectric phase. P_r, remanent polarization. E_c, coercive field.

Table 1. Ferroelectric phases accessible from cubic point symmetries, showing the orientations and symmetries of resulting ferroelectric species.

Kind of transition—kind of ferroelectric phase	Corresponding simple form of initial phase	Order of transition
$m3m$ (3) D4F $4mm$ ‖	Hexahedron (cube)	II
$m3m$ (4) D3F $3mm$ ‖	Octahedron	II
$m3m$ (6) D2F $mm2$ ‖	Rhombic dodecahedron	I
$m3m$ (12) A4F m ‖	Tetrahexahedron	I
$m3m$ (12) A*2F m ‖	Trigonal trioctahedron	I
$m3m$ (12) A**2F m ‖	Tetragonal trioctahedron	I
$m3m$ (24) A1F 1 ‖	Hexoctahedron	I
$\bar{4}3m$ (3) D$\bar{4}$F $mm2$ ‖	Hexahedron (cube)	I
$\bar{4}3m$ (4/2) D3F $3m$	Tetrahedron	I
$\bar{4}3m$ (12/2) A*mF m	Tetragonal tritetrahedron	I
$\bar{4}3m$ (12/2) A**mF m	Trigonal tritetrahedron	I
$\bar{4}3m$ (24/2) A1F 1	Hexatetrahedron	I
432 (3) D4F 4 ‖	Hexahedron (cube)	II
432 (4) D3F 3 ‖	Octahedron	II
432 (6) D2F 2 ‖	Rhombic dodecahedron	I
432 (24/2) A1F 1	Pentagonal trioctahedron	I
$m3$ (3) DmF $mm2$ ‖	Hexahedron (cube)	II
$m3$ (4) D3F 3 ‖	Octahedron	II
$m3$ (6) AmF m ‖	Pentagonal dodecahedron	I
$m3$ (12) A1F 1 ‖	Didodecahedron	I
23 (2) D2F 2 ‖	Hexahedron (cube)	I
23 (4/2) D3F 3	Tetrahedron	I
23 (12/2) A1F 1	Pentagonal tritetrahedron	I

arguments was the realization that there could be ferroelectric species in which the polar vector is reorientable but not reversible. An example is in the cubic group $\bar{4}3m$, i.e., $\bar{4}3m$ (12/2) AmF m. Again $\bar{4}3m$ is the prototype group. In the ferroelectric species there are 12 vector directions for ferroelectric polarization; (12/2) indicates that the orientation is redirectable but not reversible; A indicates arbitrary, but in the plane of the mirror (m), and the ferroelectric phase has monoclinic (m) symmetry.

In Shuvalov's table the corresponding simple form is the solid surface that would be formed by the normals to the full family of polar directions. In the "nature of the phase change" column, Roman II indicates that the Curie transition may be either first or second order thermodynamically, Roman I indicates that the change must be first order (discontinuous). The small symbols at the end of the letter descriptors indicate whether the sublattices in the different domain states are parallel (‖) or canted.

It should be stressed that the symmetry classification is a tabulation of the possible; i.e., all actual ferroelectrics of any symmetry must fall within this grouping. However, whether any structure of a given symmetry is actually prototypic depends on details of its atomic architecture, which are not controlled by the crystal symmetry.

2.3 Domains and Polarization Switching

Only simple proper ferroelectrics from centric prototype (paraelectric) symmetries have so far been applied to thin films, and discussion of domains and switching is confined to such crystals. In these proper ferroelectrics, the appearance of polarization leads to significant shape change driven through the phenomenon of electrostriction; i.e., the strain induced from the prototype $x_{ij(s)}$ is given by

$$x_{ij(s)} = Q_{klij} P_{k(s)} P_{l(s)}, \qquad (5)$$

where Q_{klij} are the symmetry-permitted electrostriction constants and $P_{k(s)}P_{l(s)}$ are components of the spontaneous polarization vector in the ferroelectric domain. Einstein summation convection is assumed, and the expression is expanded for repeated subscripts to get the full family of equations.

Clearly for these centric prototypes, domain inversion will not change the shape; however, for non-180° changes, the domains are strain distinct, and acceptable planes joining such domains (domain walls) will be regulated by elastic as well as electric boundary conditions. The spontaneous strains $x_{ij(s)}$ are quite large, of the order of 1% in $BaTiO_3$ and from 2% to 6% in tetragonal PZTs, so that the elastic constraints are strong. A tabulation of strain-preferred walls has been given by Fousek and Janovec (1969).

For $BaTiO_3$ and PZT the phenomenological Gibbs free energy is known, and it is possible to calculate the field level required for intrinsic coherent polarization switching within the single domain. Such calculations yield values orders of magnitude larger than the observed coercive field. Thus it is generally agreed that in bulk crystals the switching is incoherent and occurs by the motion of domain walls.

For $BaTiO_3$ 180° switching has been extensively studied and has shown to be by nucleation and growth of new domains, preferentially against existing walls, and not by a true sidewise motion of the wall. The presence of non-180° walls in general appears to raise the effective coercive field, but shape-changing ferroelastic switching has been much less studied.

In the randomly axed polycrystal both 180° and ferroelastic shape-changing switching must be involved, and the process varies widely between different systems. For $BaTiO_3$ less than 30% of the single-domain spontaneous polarization is remnant, and the response is a strong function of grain size. For PZT compositions close to the morphotropic phase boundary at the 52:48 Zr:Ti ratios, good "square loop" switching is observed. From elementary electrostatic considerations, Arlt (1996) has shown recently that grain-by-grain switching in a ceramic is impossible and that the switching nucleus must be a larger entity. Given the fact that thin films often have appreciably different grain sizes and stress states than are characteristic of bulk ceramics, it is possible that the switching process in films will differ markedly from that observed in bulk materials.

2.4 Transport Properties

Comprehensive reviews of the electrical conduction mechanisms in ferroelectric materials have been given by Smyth (1994) and Waser (1994). As has been described there, the principal charge carriers in the perovskite oxides are electrons, holes, and oxygen vacancies. In most of the perovskites, nominally undoped materials contain an excess of acceptors from lower-valent impurities, and so act as acceptor-doped compounds. This is especially true for the lead-based materials, where lead vacancies also act as acceptor sites. These materials are insulating at room temperature on account of the existence of hole traps (Warren *et al.,* 1994). It is important to note, though, that either variations in the film stoichiometry or preparing films without sufficient oxygen present changes the defect chemistry and can lead to unacceptably high conductivity levels.

Upon application of a dc voltage to a ferroelectric capacitor, the capacitor passes a polarization current. Eventually, this transient dies away, and the current decays to the value determined by the leakage through the capacitor. At still longer times, if the applied dc field is high enough, the sample may undergo resistivity degradation and ultimately time-dependent dielectric breakdown (TDDB) (Chen *et al.,* 1994).

As has been shown by Waser and coworkers, the current initially follows the Curie–Von Schweidler law (Waser and Klee, 1992):

$$J \propto t^{-m}, \tag{6}$$

where J is the current density, t is time, and m is approximately 0.5 (actually 0.4 to 1) (Dietz and Waser, 1995). It then flattens out as the true leakage value is reached. It is important to note that the transition to the steady-state current can be quite slow ($>10^3$ s) (Chen *et al.,* 1994). Techniques for measuring this properly have been described by Dietz and Waser (1995). In particular, it has been demonstrated that stepped changes in the applied voltage (with appropriately long wait times between steps), where both charging and discharging currents are measured, are preferable to ramped voltage experiments in determining the true leakage currents. This is especially true at low temperatures (i.e., 25 °C).

2.5 Perovskite-Structure Systems

The simple cubic perovskite ABO_3 structure is the prototype structure for many useful ferroelectric materials. Fundamentally, it consists of large 12-coordinated cations such as Na, K, Rb, Ca, Sr, Ba, Pb, etc. at the corners of a unit cell, with a smaller cation such as Ti, Sn, Zr, Nb, Ta, or W at the unit cell center octahedrally coordinated to the anions at the face centers (see Fig. 2). These octahedral units form a three-dimensional array via corner linking. The resulting structure has a high symmetry (the point group symmetry is *m*3*m*). This high symmetry is extremely useful, as it implies that when the material distorts to form a ferroelectric phase, there are many equivalent differently oriented domain state polarization directions. As a result, it is possible in the randomly axed polycrystal to thread the polarization vector through the structure without major discontinuities. Consequently, large remanent polarizations can be developed, as well as substantial piezoelectric and electro-optic coefficients.

Among the technologically important materials possessing this crystal structure are the solid solutions $BaTiO_3$:$SrTiO_3$ (BST), $PbZrO_3$:$PbTiO_3$ (PZT), and $KTaO_3$:$KNbO_3$ (KTN). $BaTiO_3$ itself shows three ferroelectric phases as a function of temperature. At temperatures above ~135 °C, the octahedra are undistorted, and the material possesses symmetry group *m*3*m*. On cooling through the Curie point, however, the Ti^{4+} ion displaces closer to one of the adjacent oxygen along the [001] direction. This results in an elongation along the *c* axis, contraction along the *a* and *b* axes, and, most importantly, development of a spontaneous polarization along the ferroelectric *c* axis. This tetragonal structure is stable to ~5 °C, at which point the Ti^{4+} ion shifts again, along one of the [110] directions of the original cubic unit cell. This leads to an orthorhombic structure with the polarization oriented along the direction of Ti^{4+} displacement. Finally, at ~ -90 °C, the TiO_6^{8-} octahedra distort again, with the Ti^{4+} ion (and hence P_s) displacing along one of the original [111] directions. Each of these symmetry changes corresponds to a first-order displacive transition (see Fig. 3).

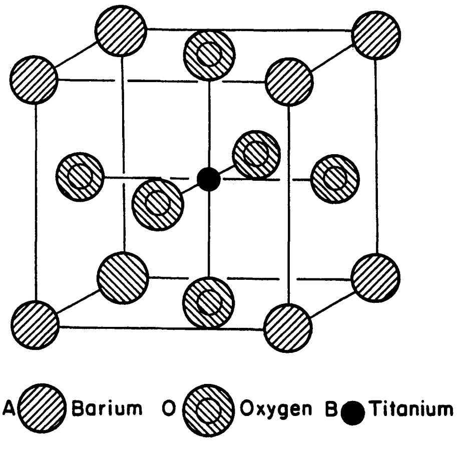

FIG. 2. Cubic form of the perovskite structure using barium titanate as a typical example.

Associated with each of the transitions in $BaTiO_3$ is a maximum in the low-field dielectric permittivity as shown in Fig. 4. Fundamentally, each transition marks a temperature where two phases are similar in energy. On application of an electric field, then, the material can be distorted between the two different polytypes. This, in turn, leads to high polarizability near transition temperatures. The high dielectric permittivity even between phase transitions in ferroelectric $BaTiO_3$ is tied to the presence of the spontaneous polarization and to the existence of domain walls.

$BaTiO_3$ and $SrTiO_3$ form a complete substitutional solid solution in which the ferroelectric–paraelectric transition temperature of the compound decreases nearly linearly with Sr composition. As a result, it is possible to tailor the transition temperature as well as the dielectric properties. For DRAM applications, paraelectric BST compositions are being widely investigated because of the combination of relatively high dielectric permittivity with low dielectric loss.

The most important of the piezoelectric ceramics are found in the lead zirconate–lead titanate solid-solutions series. PZT shows a temperature-independent morphotropic phase boundary between tetragonally and rhombohedrally distorted ferroelectric

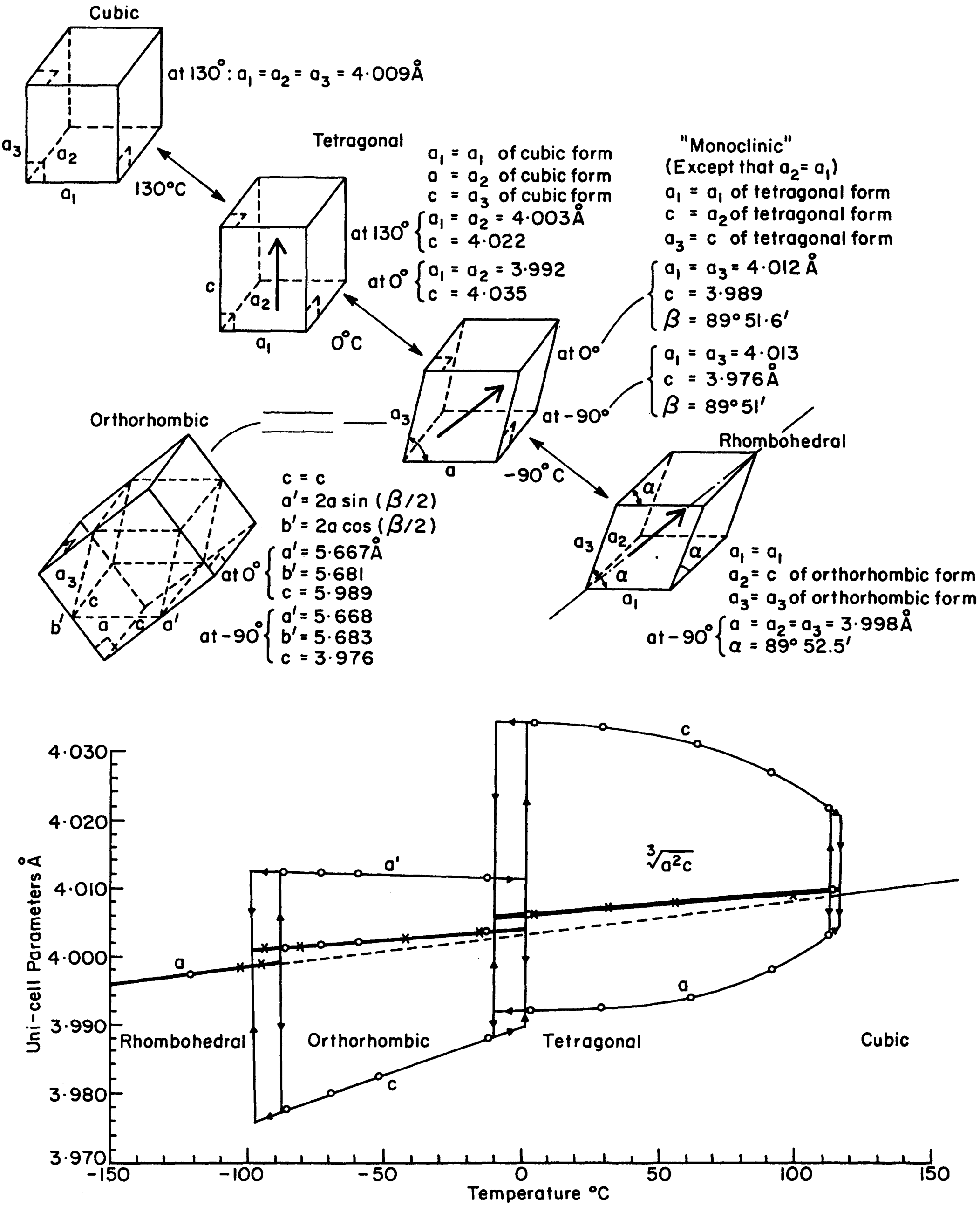
Cubic
at 130°: a1 = a2 = a3 = 4·009Å
a3
a2
a1
130°C
Tetragonal
a1 = a1 of cubic form
a = a2 of cubic form
c = a3 of cubic form
at 130° {a1 = a2 = 4·003Å, c = 4·022}
at 0° {a1 = a2 = 3·992, c = 4·035}
c
a2
a1
0°C
"Monoclinic"
(Except that a2 = a1)
a1 = a1 of tetragonal form
c = a2 of tetragonal form
a3 = c of tetragonal form
{a1 = a3 = 4·012 Å, c = 3·989, β = 89°51·6'}
at 0°
{a1 = a3 = 4·013, c = 3·976Å, β = 89°51'}
at -90°
Orthorhombic
a3
a
-90°C
c = c
a' = 2a sin (β/2)
b' = 2a cos (β/2)
at 0° {a' = 5·667Å, b' = 5·681, c = 5·989}
at -90° {a' = 5·668, b' = 5·683, c = 3·976}
a3
c
b'
a
c
a'
Rhombohedral
α
a3
a2
α
a1
a1 = a1
a2 = c of orthorhombic form
a3 = a3 of orthorhombic form
at -90° {a = a2 = a3 = 3·998Å, α = 89° 52.5'}
Uni-cell Parameters Å
4·030
4·020
4·010
4·000
3·990
3·980
3·970
c
a'
∛(a²c)
a
a
c
Rhombohedral
Orthorhombic
Tetragonal
Cubic
-150
-100
-50
0
50
100
150
Temperature °C

FIG. 3. Temperature dependence of the crystal structure, orientation of spontaneous polarization, and consequent electrostrictive strain in barium titanate.

phases at the Zr:Ti ratio of 52:48 as seen in Fig. 5. As a result of the large number of potential orientations for the spontaneous polarization, this material shows efficient poling and large enhancements in the dielectric and piezoelectric properties at this composition.

Potassium niobate, $KNbO_3$, is a monovalent/pentavalent analog of $BaTiO_3$ in essentially the same perovskite structure. It goes

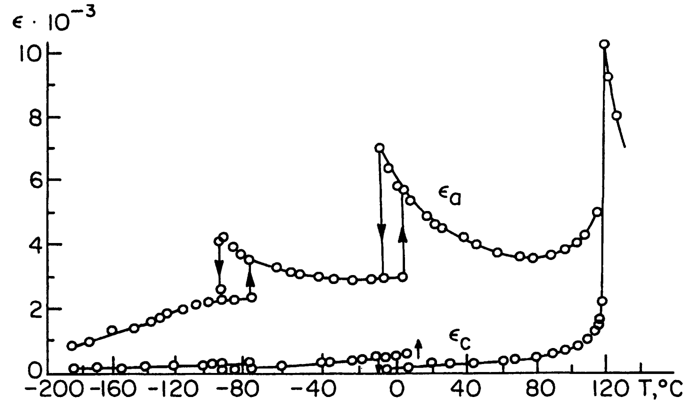

FIG. 4. Dielectric permittivity in weak electric field for a Remeika-grown c-grown single crystal of $BaTiO_3$. Note that for temperatures below ~ −90 °C the data reflect the anisotropic strain in these crystals and do not follow the symmetry of a perfect crystal.

through the same sequence of phase changes associated with the onset and reorientation of spontaneous electric polarization. In $KNbO_3$ the Curie point is at 420 °C, the tetragonal ⇄ orthorhombic transition is at 220 °C, and the orthorhombic ⇄ rhombohedral phase change is at −10 °C. Replacing Nb with Ta rapidly lowers T_c, and the sequence of change is similar to that induced in $BaTiO_3$ by Sr substitution.

At low concentration of Nb in $KTaO_3$ an interesting new phenomenon occurs, associated with the formation of small droplets of polar phase within the high-K tantalate matrix leading to massive dielectric maxima ($K \sim 100\,000$) with strong relaxation character, behavior in many ways compatible to the relaxor ferroelectrics discussed in Sec. 2.4.

Sodium niobate, like lead zirconate, is antiferroelectric at room temperature, and as in lead zirconate, the antipolar state can be destabilized under high–electric-field switch-

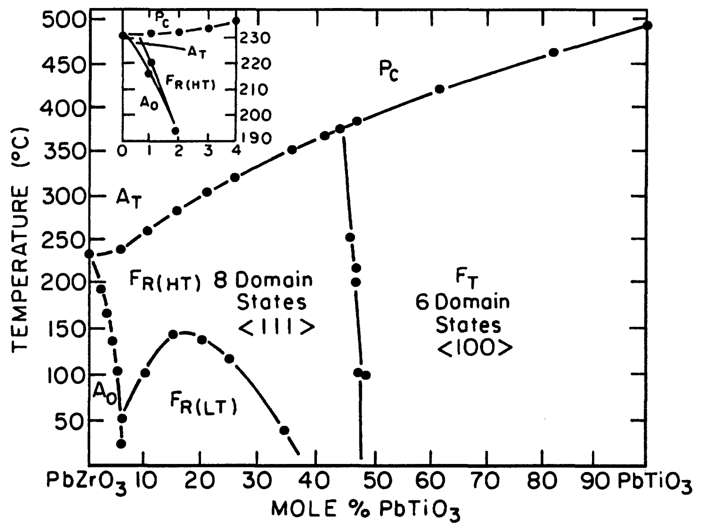

FIG. 5. Phase diagram for the $PbZrO_3$:$PbTiO_3$ solid-solution ferroelectric system. Note the almost vertical (morphotropic) phase boundary at the Zr/Ti 52/48 composition separating tetragonal and rhombohedral ferroelectric forms. P_c is the paraelectric prototypic cubic state. F_T is a tetragonal ferroelectric phase with six alternative domain states polarized along the ⟨100⟩ family of directions of the unit cube. F_R is a rhombohedral ferroelectric state with eight equivalent polar directions in its domain structure with polar vectors along equivalent ⟨111⟩ family of directions. The rhombohedral form has a simple high-temperature phase in which oxygen octahedra are in orthogonal array, and a low-temperature form in which the oxygen octahedra are slightly rotated about the ⟨111⟩ family of directions. A_0 is an orthorhombic antiferroelectric phase in which opposing polarizations occur on a unit cell scale.

Inset: An enlargement of the higher-temperature high-zirconate compositions showing the probable presence of a small insert of an antiferroelectric tetragonal phase (A_T).

ing to a strong ferroelectric form (Cross and Nicholson, 1955).

2.6 Bismuth Oxide Layer Structures

Particularly in the area of FRAMs, the family of bismuth oxide layer-structure ferroelectrics is also important. The bismuth oxide layer structures are made up of sheets of perovskite-like corner-linked oxygen octahedra separated by $Bi_2O_2^{2+}$ layers (Fig. 6) (Dorrian *et al.*, 1971). The prototypic symmetry is 4/*mmm* with the $Bi_2O_2^{2+}$ sheets normal to the *c* axis interleaved by from one to five perovskite-like layers. In the bulk crys-

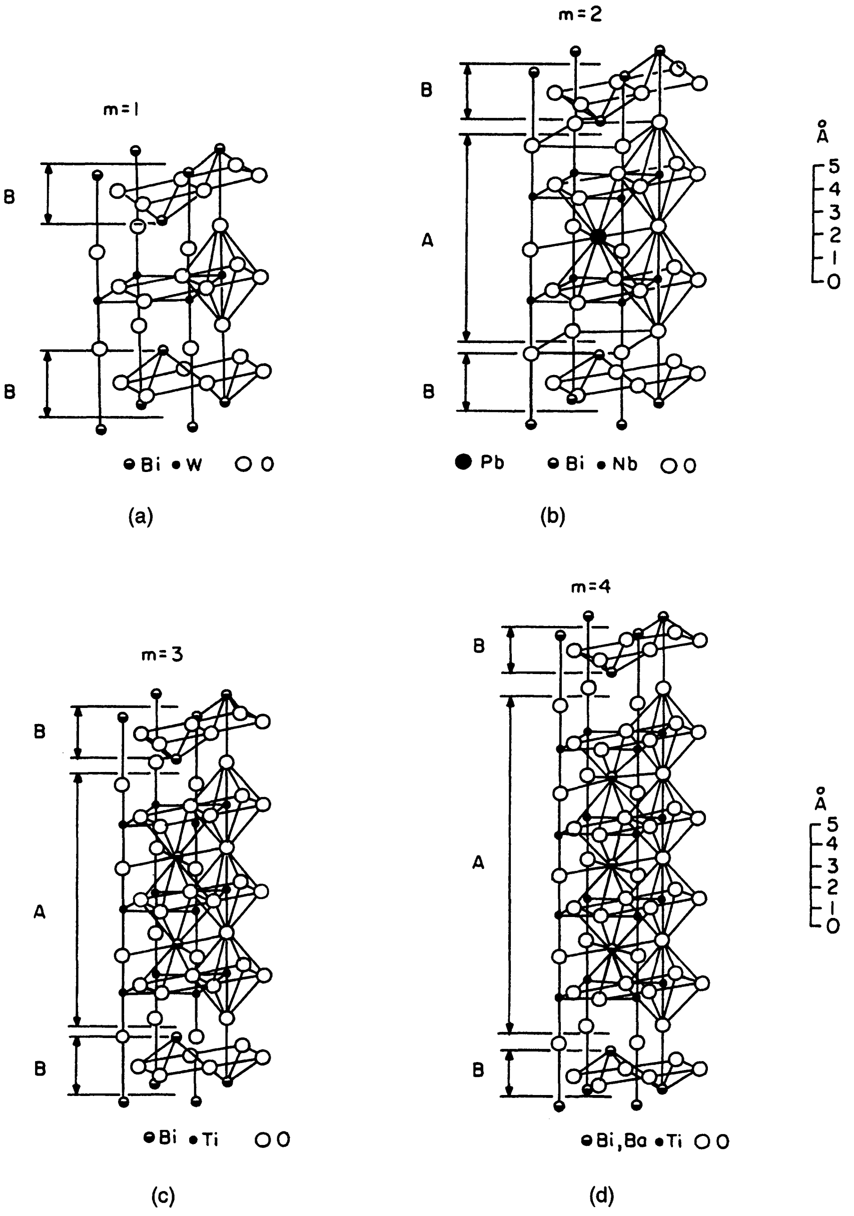

FIG. 6. Crystal structures for 1-, 2-, 3-, and 4-layer bismuth oxide layer-structure ferroelectrics. The figure shows one-half of the unit cell of (a) Bi_2WO_6, (b) $PbBi_2Nb_2O_9$, (c) $Bi_4Ti_3O_{12}$, (d) $BaBi_4Ti_4O_{15}$ (after Aurivillius). "A" denotes perovskite-like layers; "B" denotes $(Bi_2O_2)^{2+}$ sheets.

tals there are major differences both in structure and in properties between compositions with odd (1,3,5) and even (2,4) numbers of interleaving perovskite layers. Even-layer structures retain a mirror normal to the fourfold axis in the ferroelectric phases so that the polarization vector is confined to the plane of the perovskite sheet (Newnham *et al.,* 1971), while in the one- and three-layer structures the mirror is lost, and in $Bi_4Ti_3O_{12}$ the polarization tilts out of the perovskite plane. Similarly, in properties odd-layer compounds have a single phase change on cooling, leading to a prominent dielectric anomaly. Two- and four-layer structures go through two phase changes, only one of which carries the major dielectric anomaly (Diaz-Colon, 1973).

To date, the best studied thin films have been in the $Bi_4Ti_3O_{12}$ and $SrBi_2Ta_2O_9$ compositions. The latter is particularly important, as it appears that when fabricated with simple platinum electrodes it shows lower coercivity and much lower fatigue on repeated cycling than comparable PZT films. It is difficult to see why a randomly axed polycrystal film made up from crystallites with only four in-plane domain states should have low coercivity and quite high remanent polarization, and one suspects that the processing and consequent microstructure of the film may significantly modify the bulk symmetry and properties.

2.7 Alternative Systems

In addition to these compounds there are alternative normal ferroelectric systems that are, for the most part, somewhat less well studied in thin-film form. Examples include $LiNbO_3$ and its tantalate analog, the tungsten bronze systems, and some of the improper ferroelectrics. $LiNbO_3$ is widely used as a bulk single crystal for nonlinear optical applications and as a high-temperature piezoelectric (its ferroelectric transition temperature $T_c \approx 1200$ °C). Thin films are amenable to wave-guiding applications and so have been pursued for optical switches and second-harmonic–generation applications. In thin films, as in bulk crystal growth, maintaining the correct stoichiometry is problematic. The tungsten bronze family of ferroelectrics is based on a relatively complex crystal structure in which oxygen octahedra are the primary structural units (Lines and Glass, 1977). The multiple possibilities for cation disordering in these compounds, coupled with the large solid-solution regions, lead to widely tailorable dielectric, electro-optic, and electromechanical properties. Epitaxial $Sr_{5-x}Ba_xNb_{10}O_{30}$ (SBN) films have been grown on single-crystal substrates with different Sr:Ba ratios (Trivedi *et al.,* 1996). Among improper ferroelectrics such as $Gd_2(MoO_4)_3$, the primary order parameter is not the polarization, so that the ferroelectricity is a secondary consequence of a zone-boundary structural mode (Dvorak, 1971). A study of thin films of this material might be rather useful, given the strong effect that the elastic clamping of the film to the substrate is expected to have on the properties of most ferroelectrics. In $Gd_2(MoO_4)_3$, where the elastic effects should be much more important than the electrical boundary conditions, it may be easier to isolate the importance of the film stress state.

There are ferroelectric polymers of which the odd-number nylons and polyvinylidene difluoride (PVDF) are much the most widely researched. PVDF and its copolymer with polytetrafluoethylene (PTFE) is commercially produced in thick film sheet ($\tau > 10$ μm) and finds useful piezoelectric and pyroelectric application. The material is in fact a composite between a ferroelectric crystalline form and a flexible amorphous component. The polarization levels are more than an order of magnitude less than in PZTs, but the permittivity is also low so that voltage coefficients of piezoelectric and pyroelectric response are quite attractive for large-area, low-frequency sensors.

There are also large families of ferroelectric liquid crystals, which find major application in display technologies. Since, however, they are liquids, they take the form of the containment vessel and have not been widely explored in thin-film form.

2.8 Relaxor Ferroelectric Systems

In contrast to normal ferroelectrics, relaxor ferroelectrics demonstrate a diffuse phase transition and dispersion in the low-frequency dielectric properties. The broad temperature response, in conjunction with the high dielectric constants that can be achieved in bulk ceramics, make such materials attractive as relatively temperature-stable high–volume-efficiency capacitors. To date, thin films based on relaxor ferroelectric compositions such as lead magnesium niobate (PMN) or lead scandium tantalate (PST) show lower dielectric constants than do bulk materials, though they are still clearly high enough to warrant investigation as alternative capacitor compositions. Additional areas where thin-film relaxor ferroelectrics would be attractive include pyroelectric sensors and actuators based on either electrostrictive or biased piezoelectric response. Among the compositions available, solid solutions based on the PMN–$PbTiO_3$ and PST–$PbTiO_3$ systems look particularly good.

3. FERROELECTRIC FILM PREPARATION

3.1 Boundary Conditions for Films on Silicon

Most of the interesting ferroelectrics for integration on silicon come from the perovskite and the layered perovskite families. Unfortunately, in most cases when deposited directly on silicon, these materials react at the temperatures required for processing crystalline films. The lead-based compounds, in particular, form a lead silicate glass phase when processed directly on silicon or SiO_2, which leaches lead from the ferroelectric and acts as a low–dielectric-constant barrier layer between the ferroelectric and the semiconductor. As a result, reaction-blocking chemical barrier layers between the ferroelectric and the Si are required. Development of appropriate barrier-layer and electrode schemes is currently an area of intense research. This is particularly important in memory elements where the device capacitor is placed directly over the contact plug (often poly-Si). For such a geometry, a barrier layer is required to prevent the plug material from oxidizing or reacting with the dielectric during the capacitor processing.

There are also processing temperature restrictions that must be considered in integrating ferroelectric elements with Si-based semiconductor circuitry. All of the interesting ferroelectrics are multiple-cation compositions. While this is not inherently a limitation, the fact that many of the interesting compositions contain at least one volatile component, either Pb or Bi, limits the upper temperature to which the film can be exposed before appreciable changes in the film chemistry occur. As a result, the ferroelectric elements are typically introduced into the semiconductor after the majority of the CMOS processing has been completed. This, in turn, places restrictions on the temperature excursions allowed during the ferroelectric film processing, so that the silicon circuitry itself is not affected. This time–temperature limitation is expected to become progressively more restrictive as linewidths in the circuitry continue to fall. In order to minimize the thermal budget associated with the preparation of the ferroelectric element, rapid thermal processing (RTP) is preferred over conventional annealing for *ex situ* crystallized films. For films deposited at elevated temperatures so that they can be prepared in the desired crystalline form, routes that lower the deposition temperature are particularly attractive.

An additional parameter that must be considered, particularly for thick-film deposition, is the thermal expansion coefficient mismatch between the ferroelectric and the substrate. Since the thermal expansion coefficient of PZT does not match that of the underlying Si substrate, appreciable stresses develop during cooling from typical processing temperatures of 650 to 750 °C. Thus, particularly in thick films, the resulting stresses are large enough to cause film cracking and delamination. While this is unlikely to be a serious limitation in capacitor applications, for piezoelectric devices where large electrically induced displacements are required, larger thicknesses are necessary.

Some of the most important issues to be resolved in integrating ferroelectric (or paraelectric) capacitors into silicon-based circuitry are

1. development of appropriate electrode materials,
2. development of IC-compatible processing (including conformal coverage of nonplanar surfaces as well as all of the necessary etch chemistries), and
3. maintaining the required electrical properties through all of the processing steps required to form the device.

To date, the most important electrodes

have been Ti/Pt and RuO_2. Both retain good conductivity after firing in an oxidizing atmosphere at the temperatures required for crystallization of the capacitor. They do, however, suffer from problems. Pt, in particular, frequently undergoes significant stress relief during heating (which is often accompanied by hillocking) (Bruchhaus *et al.*, 1992) along with the migration of O and Ti through the Pt layer. PZT films electroded with Pt show much higher levels of fatigue than the same films on RuO_2. In addition, no good dry-etch chemistry has been developed for Pt. RuO_2, on the other hand, leads to improved fatigue behavior but degraded leakage current values in PZT. It also has a lower conductivity than Pt, which can lead to degradation in the high-frequency response of ferroelectric capacitors. Moreover, many of the volatile Ru compounds are toxic. Additional electrode materials that have been investigated include Pd, Ir, IrO_2, $SrRuO_3$, $CaRuO_3$, $La_{0.5}Sr_{0.5}CoO_3$, $YBa_2Cu_3O_{7-x}$, and TiN.

From an equipment standpoint, since the high volatility of Pb leads to appreciable diffusion during annealing steps, it will be necessary to develop secondary, contaminated, processing lines for the ferroelectric film processing in order to avoid contaminating the CMOS fabrication equipment. This is a significant additional expense.

3.2 Deposition Methods

There are a wide variety of deposition methods that have been used to prepare ferroelectric thin films, including chemical methods such as sol-gel or metal-organic decomposition (MOD), sputtering, laser ablation, metal-organic chemical-vapor deposition (MOCVD), reactive evaporation, molecular-beam epitaxy (MBE), and liquid-phase epitaxy. Factors that may dictate the techniques chosen for a specific composition and application include

- availability of appropriate precursors,
- uniformity of coverage,
- particulate generation,
- ease of composition control,
- growth temperature limitations,
- desired residual stress level,
- required thickness and deposition rate,
- step coverage, and
- compatibility with additional processing (especially semiconductor processing).

Since the literature is by now extensive enough that it is not possible to give a comprehensive review of each of the techniques, a short review is given in the following section. More complete descriptions are contained elsewhere (Myers and Kingon, 1991; Kingon *et al.*, 1992; Myers *et al.*, 1993; Tuttle *et al.*, 1994; Desu *et al.*, 1995). It is important that all of the methods described can be used to produce high-quality films, once problems associated with film composition and film microstructure are overcome.

3.3 Vapor-Phase Techniques

There are a host of vapor-phase techniques for preparing ferroelectric films, which differ primarily in the method for generating the vapor phase. Multiple texts are available that discuss the equipment requirements and deposition mechanisms for these methods. One factor that distinguishes vapor-phase production of ferroelectric films from deposition of other materials is the fact that most ferroelectrics are multiple-cation compounds that show large degradation in properties when made nonstoichiometric. This is especially problematic because many of the useful ferroelectric compounds contain cations with vastly different sputtering yields and vapor pressures (especially Bi and Pb relative to Ti, Zr, or Ta.). Consequently, in many of the vapor-phase deposition techniques, maintenance of the film composition is nontrivial. Of the vapor-phase techniques, Table 2 summarizes the advantages and disadvantages associated with each.

3.4 Liquid-Phase Techniques

Both sol-gel and MOD techniques rely on introducing an organometallic precursor into an appropriate solvent system. This liquid is then either spun, misted, sprayed, or dip-coated onto the substrate. Following solvent evaporation, the organic phase must be removed through a pyrolysis step and the film crystallized. While good results have been obtained with furnace annealing, a consensus is developing that RTP is preferable for integrated applications, given the vastly lower thermal budget. In principle, the stoi-

Table 2. Advantages and disadvantages of vapor-phase thin-film deposition techniques.

Method	Advantages	Disadvantages
Magnetron sputtering	High deposition rate possible. IC-compatible. Good compositional flexibility.	Target chemistry changes in reactive gas environment. Preferential sputtering/resputtering of multicomponent materials.
Ion-beam sputtering	Independent control of ion current density and ion energy. Atom/ion arrival ratios controllable.	Low deposition rates.
Laser ablation	Good target/film compositional fidelity. High deposition rate.	Boulder production. Poor thickness uniformity.
MOCVD	Excellent step coverage. Large area coverage possible. High deposition rate possible.	Appropriate precursors required. Precursor toxicity.
MBE	Layer-by-layer control.	Low throughput. Low coverage areas.

chiometry of the resulting film should be identical to that of the initial solution. It is clear, however, that compositional heterogeneities can arise during gelation, pyrolysis, or crystallization, any of which may lead to some segregation of cations from an originally homogeneous precursor. Procedures for solution preparation of high-quality lead zirconate titanate, as well as other perovskite films, are well documented for a variety of solvent–precursor systems (Tuttle and Schwartz, 1996). Similarly, several reports on the deposition of $Sr_2BiTa_2O_9$ films from solution precursors are available (Huffman, 1995). Some of the significant advantages of the sol-gel techniques are good compositional tailorability, even in multiple-cation systems, large area coverage, and the capability of building up films several microns thick (Chen *et al.*, 1996). In terms of integrated ferroelectric films, sol-gel materials have been used for most of the published work on MEMS. Debits include the fact that the liquid-phase techniques do not permit conformal coverage, and they are not compatible with high-density integrated-circuit processing.

4. CHARACTERIZATION OF FILM MICROSTRUCTURE

4.1 Introduction

Demonstration of good crystallinity in thin films is not, in and of itself, a guarantee of acceptable film properties. Equally important are the microstructure and the defect structure of the films. The role of defect structure has been discussed already. In this section, the use of microstructural characterization as an essential complement to structural and electrical characterization is described. Numerous articles describing microstructure–property relations in ferroelectric films are available in the literature (Tuttle *et al.*, 1995; Barlingay and Dey, 1992; Reaney *et al.*, 1994). Given the wide variety of data available, only generic features are covered here.

4.2 Direct Methods

Key microstructural features that can significantly affect the electrical properties include grain size, crystallite orientation, phase purity, density, and, in some cases, the depth profile of each of these parameters. In particular, the importance of phase purity and crystallographic orientation have been well described in numerous texts (Jaffe *et al.*, 1971; Herbert, 1985). Structure-sensitive probes such as x-ray diffraction and selected-area diffraction can be used to determine the phase content and to investigate the global and local crystallographic orientation, respectively. However, x-ray diffraction measurements alone are not a conclusive demonstration of single-phase films. Especially in the lead- and bismuth-based compounds, second phases can be present to levels as high as 15% before they appear in most x-ray patterns. High-resolution SEM, or TEM, particularly in conjunction with local composition probes, can be used to identify second phases such as PbO-rich pockets, or pyrochlore surface layers in $PbTiO_3$ and PZT films (Tuttle *et al.*, 1996; Reaney *et al.*, 1994; Fox *et al.*, 1992). Similarly, second-phase porosity can often be identified. Such microscopy studies have been instrumental in identifying the nucleation and growth processes in ferroelectric films. Similar studies are now coming out on the compound $SrBi_2Ta_2O_9$ (Scott *et al.*, 1996).

In interpreting the film properties, a knowledge of the grain size is also important. Electron microscopy and, to a lesser extent, atomic-force microscopy (AFM) can be utilized to determine the film microstructure, including the primary grain size, directly. In general, it cannot be assumed that grains in thin films are isometric, and so measurement in cross section is required. It is also important to note that lower-resolution SEM studies often identify cluster boundaries, which may be considerably larger than the primary grain size in the material.

TEM analysis also permits investigation of the domain structure in films. Several studies have demonstrated that PZT thin films tend to show relatively high domain densities, at least once the grains are above a certain size (Reaney *et al.*, 1994; Voight *et al.*, 1993). This is consistent with the observation in bulk PZT ceramics that the domain size drops with the grain size (Kim, 1994). It is not clear, however, that these twin walls are mobile in thin films (Tuttle and Schwartz, 1996; Trolier-McKinstry *et al.*, 1995a). Consequently, it may not be possible to rely, in thin films, on the extrinsic contributions to the dielectric and piezoelectric properties that are available in bulk materials.

4.3 Optical Methods

Optical methods, including wave-guiding experiments (Potter *et al.*, 1993), light scattering (Sinclair *et al.*, 1995), and spectroscopic ellipsometry (Trolier-McKinstry *et al.*, 1993; Trolier-McKinstry *et al.*, 1995b) can also be used to depth profile dielectric films nondestructively. These methods can be quite sensitive to small inhomogeneities in the density of the film through the film thickness. Such information is useful in electrical applications since introduction of a low-density (and hence low dielectric permittivity) layer leads to an appreciable decrease in the effective dielectric constant, piezoelectric coefficient, and electro-optic coefficient of the film. For optical applications, such as wave guiding, identification and elimination of low-density layers is essential, as the associated refractive index discontinuities lead to unacceptable optical scattering (Fork *et al.*, 1996). Similarly, scattering centers such as rough surfaces, grain boundaries, and domain walls can significantly decrease the distance over which light can be guided. The importance of these factors is best determined by optical measurements, as a microstructure-based technique will not necessarily be a complete predictor of the scattering losses.

5. PROPERTIES OF FILMS ON SILICON

5.1 High-Permittivity Films (DRAMs)

In the area of memory applications, the drive to incorporate paraelectric materials is predicated on the need for dielectrics that permit development of gigabit memories. A good review of the requirements for memory dielectrics is given by Jones *et al.* (1995). For DRAM applications, information is kept by storing charge on a capacitor and periodically reading and resetting it (typically, around ten times per second). In such a configuration, it is imperative that the charge be retained without excessive leakage during a cycle time, and it must be large enough that α-particle events do not change the logic state. This necessitates a capacitor of at least 50–75 fF (Jones *et al.*, 1995).

Until this point, SiO_2 and SiO_2/Si_3N_4 have been used as the capacitor material. However, as memory density increases beyond 256 Mbit, it is difficult to achieve sufficient capacitance per element with these materials. Ta_2O_5 films, with their higher relative permittivity, permit higher densities and may be adequate for 1-Gbit DRAM. Considerable simplification of the capacitor structure would, however, be facilitated by materials of higher ϵ_r. Potential candidates include $SrTiO_3$, $Ba_xSr_{1-x}TiO_3$, or paraelectric La-doped PZT (Jones *et al.*, 1995).

The primary materials parameters that must be optimized for DRAM applications are a high storage density, a low leakage current, little resistance degradation, and suppression of time-dependent dielectric breakdown (TDDB). Paraelectric compositions are preferred in these applications. Most dielectric films for DRAM applications show somewhat lower dielectric constants than their bulk counterparts, especially when prepared as layers <100 nm in thickness. However, several groups have demonstrated that it is possible to prepare films that meet the capacitance density, leakage, and breakdown strengths required for DRAMs. A listing of some of the reported efforts in DRAM manufacturing is given in Table 3. A discussion of some of the important processing integration issues is given below in the section on FRAM materials, since many of the issues are identical.

5.2 Hysteretic Films (FRAM)

In the design of ferroelectric nonvolatile memories, the polarization state of the film is used to store the ones and zeros of the binary code. In principle, such memories are nonvolatile since the remanent polarization state of the ferroelectric is not lost when power is cut off. Low-density FRAMs are attractive for applications such as smart cards and drop-in EEPROM replacements, while high-density FRAM could replace some Si-based memory in computing applications. FRAMs can be written quite rapidly, and so they offer significant advantages over other nonvolatile memories such as EEPROM. In practice, several barriers need to be overcome before the full benefits of a high-density, nonvolatile memory based on ferroelectric films will be viable. These include

1. fatigue, i.e., loss of switchable polarization associated with repeated polarization reversals over the lifetime of the device;
2. poor retention of the written polarization direction;
3. imprint, or development of a preferred polarization state; and
4. development of IC-compatible processing.

Significant progress has been made in each of these areas over the past few years.

At present, the two best-developed candidates for this application are PZT and $SrBi_2Ta_2O_9$. For FRAM applications, PZT is attractive because of its large switchable polarization ($2P_r$ around 45–55 $\mu C/cm^2$) and the lower temperatures required for crystallization. However, PZT films deposited on Pt electrodes suffer significant fatigue in the magnitude of the switchable polarization after the polarization has been reversed on the order of 10^7 times. This can lead to significant difficulties in distinguishing the memory state. While for a few applications this is acceptable, in computer memories it would be preferable to have 10^{12}–10^{15} polarization reversal cycles. Several years ago, it was demonstrated that fatigue could be eliminated in PZT films by using conductive oxide electrodes such as RuO_2, IrO_2, or LSCO (Au-

Table 3. Some efforts in DRAM production.

Company	Device	Dielectric/Ferroelectric	Reference
Matsushita	4-, 6-Mbit DRAM	BST	Shimada *et al.*, 1995
Matsushita	256 kbit	$SrBi_2Ta_2O_9$	Itoh *et al.*, 1995
NEC	Gbit-compatible DRAM	$SrTiO_3$	Lesaicherre *et al.*, 1995
NEC	1-Mbit FRAM	$SrBi_2Ta_2O_9$	Scott *et al.*, 1996
Mitsubishi	4-, 9-Mbit DRAM Gbit-compatible	BST	Itoh *et al.*, 1995
GEC-Marconi	4-kBit FRAM silicon-on-sapphire	PZT	Bland *et al.*, 1995
GEC-Marconi	Infrared imaging	PZT	Shorrocks *et al.*, 1995
Ramtron	64-kbit shadow RAM	PZT	Bondurant, 1990
U.S. DRAM Consortium	ULSI DRAM	BST	Kingon *et al.*, 1996
Samsung	64-kbit FRAM	PZT	Chung *et al.*, 1996
Samsung	USLI DRAM	BST	Hwang *et al.*, 1995
Motorola/Symetrix	FRAM	$SrBi_2Ta_2O_9$	Melnick *et al.*, 1995
SHARP	FRAM	$Bi_4Ti_3O_{12}$	Kijima *et al.*, 1996
Texas Instruments	Infrared imaging	PZT	Udayakumar, 1996
Northern Telecom	FRAM	PZT	Chivukuha *et al.*, 1995
FRAM	PZT		Wakamura *et al.*, 1995

ciello and Ramesh, 1996; Nakamura *et al.*, 1994; Ramesh *et al.*, 1993). In some cases these electrodes also minimize imprint problems (Nakamura *et al.*, 1994). An alternative choice is to switch to one of the Bi-layer structure ferroelectrics like $SrBi_2Ta_2O_9$. While in polycrystalline films, the reversible polarization in this system is smaller than in the PZTs ($2P_r < 22$ $\mu C/cm^2$), they also show negligible fatigue out to 10^{12} cycles, even when deposited on Pt electrodes (Jones and Desu, 1996).

In addition to the materials requirements, there are many processing issues that must be addressed in order for paraelectric or ferroelectric films to be integrated into DRAM or FRAM. First, most groups are reporting memory capacitor structures similar to that shown in Fig. 7. Since the capacitor must be deposited over a 0.25–0.75 μm step in order to increase the element capacitance while maintaining a small feature size, a deposition process that enables conformal coverage is essential. For such applications, MOCVD is the method of choice. Second, each of the materials utilized must be patternable at the required feature size, which will probably require reactive ion etching (RIE). Third, the dielectric film properties must be retained throughout the entire processing cycle. As has been discussed elsewhere (Jones and Desu, 1996), this is a significant challenge. Especially problematic is the RIE (Pan *et al.*, 1994), which can result in degradation of the ferroelectric's hysteresis loop (for FRAMs). Additional degradation can be induced when the interlevel dielectric is deposited over the patterned capacitor. Moreover, any processing step that involves significant exposure to hydrogen can lead to reduction of the ferroelectric layer, with a consequent increase in the conductivity, and a decrease in the switchable polarization. A more complete review of these effects is given by Archard *et al.* (1995).

Research in both DRAM and FRAM memories is being conducted in Japan, the United States, Europe, and Korea. Table 3 summarizes recent publications in these areas.

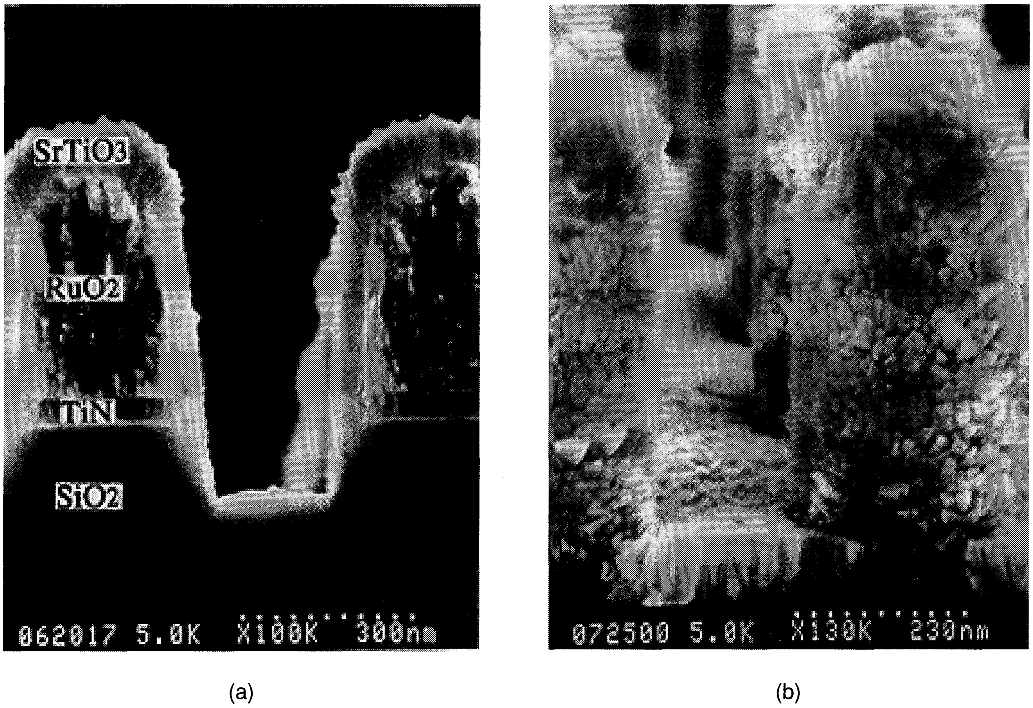

(a) (b)

FIG. 7. SEM micrographs showing (a) the conformal step coverage of 1000-Å-thick $SrTiO_3$ films over RuO_2/TiN electrodes and (b) a high-magnification image of the MOCVD $SrTiO_3$ films. Pictures taken from Lesaicherre et al. (1995).

5.3 Pyroelectric Films

As described in Sect. 1.3, thin-film pyroelectrics offer significant advantages in terms of integration with integrated circuitry and low thermal time constants relative to bulk materials (Shorrocks *et al.*, 1995). For pyroelectric applications, $PbTiO_3$, $(Pb,La)TiO_3$, $(Pb,Ca)TiO_3$, and PZT have been the most widely studied film compositions. Interest in the $PbTiO_3$-based compounds stems from its combination of a large polarization (around 80 $\mu C/cm^2$) and hence the potential for a large pyroelectric coefficient, and its relatively modest dielectric permittivity. Both factors are necessary for a substantial figure of merit (FOM). Ca or La doping serves the dual purpose of bringing the ferroelectric transition temperature closer to room temperature and increasing the electrical resistivity of the film. Taking p/ϵ_r as an approximate FOM for a pyroelectric detector, PZT films have been demonstrated with FOMs of ~0.7 $\mu C/(m^2\ K)$ (Shorrocks *et al.*, 1995; Kohli *et al.*, 1995), while $PbTiO_3$-based films have been reported with FOMs as high as 2.9 $\mu C/(m^2\ K)$ (Kamada *et al.*, 1995; Seifert *et al.*, 1996).

In terms of thermal isolation, two families of device geometries have been pursued: diaphragm structures prepared by bulk Si micromachining and air-bridge arrays (Shorrocks *et al.*, 1995; Polla, 1995) for higher-resolution imaging. Thin-film $(Pb,La)TiO_3$ pyroelectric sensors are now being used commercially in some Matsushita air-conditioning units so that the ventilation can be automatically adjusted on the basis of the number of heat sources in the room (Kamada *et al.*, 1995). There is considerable room in this field for optimization of detector properties based on both the film composition and microstructure as well as new device designs.

5.4 High–Energy-Density Transducers

The primary requirements for a thin-film transducer are that it demonstrate an appreciable controllable displacement and that it have a high energy density. Piezoelectric and electrostrictive materials are attractive from both standpoints, as large electrically induced displacements can be achieved. Moreover, on account of the high dielectric constant of most ferroelectrics, they have a significant advantage in energy density over electrostatic actuators. In addition, they show much higher piezoelectric activity as well as higher piezoelectric coupling coefficients than ZnO. For many years, ZnO has been the predominant piezoelectric transducer for thin-film applications, largely because oriented thin films could be sputtered reproducibly. However, the piezoelectric activity of ZnO is modest (typically 5–10 pm/V). As a consequence, in applications where large displacements and efficient electromechanical coupling are required, ferroelectric thin films (with d_{33} = 50–300 pm/V) offer significant advantages.

Among the ferroic candidates for actuator films, three distinct mechanisms can be utilized to generate a strain in the film: piezoelectricity, electrostriction, and phase switching. Fundamentally, the piezoelectric effect is the linear relationship between the applied electric field and the resultant displacement. Electrostriction, on the other hand, implies proportionality between the square of the electric field and the shape change, and is observed when piezoelectricity is absent. In practice, both of these relations are applicable over a limited electric-field range. For example, when a large alternating electric field is applied to excite piezoelectric displacement, hysteresis in the strain is observed due to domain-wall motion. While electrostrictive materials show much lower levels of hysteresis in the strain, it has been found that the strain is properly proportional to the square of the polarization rather than the square of the electric field at high field levels. Phase-switching compositions, on the other hand, rely on the electrically induced transition from an antiferroelectric phase to a ferroelectric phase. Since the volumes of the unit cells differ between the two, an appreciable surface displacement can be generated.

Thin-film actuators based on each of these mechanisms have been demonstrated over the past several years. Among the piezoelectric compositions, PZT-based materials have been the best studied. In most cases, PZT thin films show smaller effective piezoelectric coefficients than do bulk ceramics of the same composition (i.e., d_{33} ~ 70–140 pm/V for undoped PZT films where undoped bulk ceramic PZT shows a d_{33} of ~220 pm/V). Nonetheless, this represents a significant step up from ZnO-based devices. Comparable properties have been demonstrated by several groups, including some in the United States (Polla, 1995; Udayakumar *et al.*, 1992; Li *et al.*, 1994), Europe (Muralt *et al.*, 1995), and Japan (Fujii and Watanabe, 1996). In many cases, the piezoelectric coefficients of PZT thin films can be increased if the ac driving field is superimposed over a dc bias electric field (Udayakumar *et al.*, 1992; Muralt *et al.*, 1995). Thus far, a conclusive demonstration of soft PZT properties in thin films has not been reported in the literature, possibly as a result of either the small grain size or the stress state in films, either of which can restrict the domain-wall mobility (Shepard *et al.*, 1996; Damjanovich and Demartin, 1996). However, this may be obviated, at least in part, by the fact that thin films typically have high breakdown strengths (~1 MV/cm), and so they can be driven harder than a comparable bulk actuator.

Over the last few years, several types of MEMS actuators using ferroelectric films have been demonstrated, including micromotors (Udayakumar *et al.*, 1992; Muralt *et al.*, 1995; Flynn *et al.*, 1992), pumps (Polla, 1995), acoustic imaging devices (Bernstein *et al.*, 1996), AFM components (Fujii and Watanabe, 1996), and meanderline actuators (Polla and Francis, 1996). Outstanding issues in the field include the effect of the film stress state and the film microstructure on the piezoelectric coefficients, the aging of the properties, and the long-term reliability of the structures for large-amplitude devices.

Considerably less data are available in the literature on the use of electrostrictive compositions for thin-film actuators. PMN:$PbTiO_3$ electrostrictive compounds show very large strain levels in bulk ceramics without the hysteretic response that is characteristic of PZT under large drive levels. If comparable properties can be achieved in thin films, such materials would be attractive for precise positioning components.

In contrast, there have been several reports on the properties of antiferroelectric–ferroelectric phase-switching materials for thin-film actuators, with the majority of the activity being centered in the United States (Gaskey *et al.*, 1995; Tani *et al.*, 1994; Yamakawa *et al.*, 1996). Square double hysteresis loops with well-defined coercive fields have been prepared from lead zirconate and doped lead zirconate compositions (especially La- and Sn-doped materials). Of particular note is the fact that strain levels of 0.3% have been generated in films, which exceeds that available in piezoelectrically actuated materials. Such films are especially attractive in digital displacement applications where well-defined on and off states are required.

Many MEMS devices require larger surface displacements than can be generated in thin films. In such applications, thicker films are required. Preparation of thick films entails considerable care in the processing in order to prevent film cracking or delamination. Several groups have demonstrated thick piezoelectric films, largely prepared by wet-chemical methods (i.e., sol-gel or MOD). In particular, PZT films of 12 μm in thickness have been prepared on Pt-coated Si (Chen *et al.*, 1996), and up to 10 μm in thickness on ZrO_2 substrates (Brooks *et al.*, 1996).

6. CONCLUSIONS

Over the past few years, it has been clearly demonstrated that ferroelectric thin films have tremendous potential for integration with silicon-based electronics. Significant strides in appropriate materials selection, film deposition and characterization, and IC-compatible processing have been made. In particular, prototype devices have been demonstrated using ferroic films in DRAMs, FRAMs, decoupling capacitors, pyroelectric sensors, and piezoelectric or phase change transducers. Both low-density FRAMs and pyroelectric sensors are now being produced commercially.

It is clear, though, that much additional work is required in order to realize fully the benefits of ferroelectric thin films in integrated applications. For the high-density memory applications, where there are tremendous potential advantages to ferroic film incorporation, significant work still needs to be done to optimize the film properties to promote both short-term and long-term device reliability, facilitate low-temperature deposition, and improve compatibility of both the electrode and the capacitor with silicon electronics processing. In the areas of pyroelectric and piezoelectric devices, only modest efforts have been expended to date in the development of film composition and film integration. Vendors capable of supplying a range of thin-film compositions for prototyping would also be an important boost in terms of demonstrating the viability of the technology.

GLOSSARY

Aging: The continuous change of properties in a ferroelectric as a function of time under zero external stimulus and constant temperature.

Curie Temperature: The temperature that is the boundary between ferroelectric and paraelectric phases in a ferroelectric crystal.

Curie–Weiss Temperature: If the dielectric stiffness in the paraelectric phase is of the form $x = x_0(T - \theta)$, i.e. following a Curie–Weiss law, it is the Curie–Weiss temperature θ where the extrapolated stiffness would go to zero.

Fatigue: The loss of switchable polarization as a function of the number of switching cycles between "saturated" polarization states.

Imprint: Development of a favored remanent polar state under asymmetric pulsed switching conditions.

Morphotropic Phase Boundary (MPB): A phase boundary in the phase diagram of a ferroelectric solid solution system that is nearly vertical from the composition axis; i.e., a composition close to the boundary will remain close over a wide temperature interval.

Works Cited

Aizu, K. (1969), "Possible Species of 'Ferroelastic' Crystals and of Simultaneously Ferroelectric and Ferroelastic Crystals," *J. Phys. Soc. Jpn.* **27,** 387.

Archard, H., Mace, H., Peccoud, L. (1995), "Device Processing and Integration of Ferroelectric Thin Films for Memory Applications," *Microelec. Eng.* **29,** 19.

Arlt, G. (1996), in: C. A. Paz de Araujo (Chm.), *8th International Symposium on Integrated Ferroelectrics, Tempe, Arizona,* New York: Gordon and Breach.

Auciello, O., Ramesh, R. (Eds.) (1996), Special Issue on Electroceramic Thin Films, Part I, Processing, *Mater. Res. Bull.* **21** (6), 21–58.

Barlingay, C. K., Dey, S. K. (1992), "Observation of Sol-Gel Solid Phase Epitaxial Growth of Ferroelectric $Pb(Nb,Zr,Ti)O_3$ Thin Films on Sapphire," *Appl. Phys. Lett.* **61,** 1278.

Bernstein, J., Houston, K., Niles, L., Li, K., Chen, H., Cross, L. E., Udayakumar, K. (1996), "Integrated Ferroelectric Monomorph Transducers for Acoustic Imaging," in: C. A. Paz de Araujo (Chm.), *8th International Symposium on Integrated Ferroelectrics, Tempe, Arizona,* New York: Gordon and Breach.

Bland, T., Ohbi, J., Patel, A., Kirby, P., Robinson, M., Kerr, J. (1995), "Integration of Sol-Gel PZT with Silicon-on-Sapphire CMOS Circuitry," *Microelec. Eng.* **29,** 29.

Bondurant, D. (1990), "Ferroelectronic RAM Memory Family for Critical Data Storage," *Ferroelectrics* **112,** 273.

Brooks, K. G., Kohli, M., Taylor, D. V., Maeder, T., Reaney, I., Kholkin, A., Muralt, P., Setter, N. (1996), in: B. Kulwicki (Ed.), *Proceedings of the International Symposium on Applications of Ferroelectrics (ISAF 96),* New Brunswick, NJ: IEEE.

Bruchhaus, R., Pitzer, D., Scheithaurer, U., Hoesler, W. (1992), "Investigation of Pt Bottom Electrodes for 'In-Situ' Deposited $Pb(Zr,Ti)O_3$ (PZT) Thin Films," in: A. I. Kingon, E. R. Myers, B. A. Tuttle (Eds.), *Ferroelectric Thin Films II,* MRS Proceedings Vol. 243, Pittsburgh: Materials Research Society, p. 355.

Chen, H. D., Udayakumar, K. R., Gaskey, C. J., Cross, L. E. (1996), "Fabrication and Electrical Properties of Lead Zirconate Titanate Thick Films," *J. Am. Ceram. Soc.* **79,** 2189.

Chen, X., Kingon, A. I., Al-Shareef, H., Bellur, K. R. (1994), "Electrical Transport and Dielectric Breakdown in $Pb(Zr,Ti)O_3$ Thin Films," *Ferroelectrics* **151,** 133.

Chivukula, V., Ilowski, J., Emesh, I., McDonald, D., Leung, P., Sayer, M. (1995), "Dielectric Properties of Ferroelectric Thin Films in the Frequency Range of mHz–GHz," *Integ. Ferroelec.* **10,** 247.

Chung, I., Kim, C. J., Chung, C. W., Yoo, I. K. (1996), "Integration of Ferroelectric Capacitors Using Multilayered Electrode," in: B. Kulwicki (Ed.), *Proceedings of the International Symposium on Applications of Ferroelectrics (ISAF 96),* New Brunswick, NJ: IEEE.

Cross, L. E., Nicholson, B. J. (1955), "The Optical and Electrical Properties of $NaNbO_3$ Single Crystals," *Philos. Mag.* **46,** 453.

Damjanovich, D., Demartin, M. (1996), "Dependence of the Direct Piezoelectric Effect in Coarse and Fine Grain Barium Titanate Ceramics on Dynamic and Static Pressure," *Appl. Phys. Lett.* **68,** 3046.

Desu, S. B., Ramesh, R., Tuttle, B. A., Jones, R. E., Yoo, I.-K. (1995), *Ferroelectric Thin Films V,* MRS Proceedings Vol. 433, Pittsburgh: MRS.

Diaz-Colon, F. A. (1973), "Dielectric and Electro-Optic Properties of Mixed Bismuth Oxide Layer Structure Compounds," Ph.D. Thesis, The Pennsylvania State University, University Park, PA.

Dietz, G. W., Waser, R. (1995), "How to Analyze Relaxation and Leakage Currents of Dielectric Thin Films: Simulation of Voltage-Step and Voltage-Ramp Techniques," *Integ. Ferroelec.* **9,** 317.

Dvorak, V. (1971), "The Origin of the Structural Phase Transition in $Gd(MoO_4)_3$," *Phys. Status Solidi* **B45,** 147.

Dorrian, J. F., Newnham, R. E., Smith, D. K., Kay, M. I. (1971), "Crystal Structures of $Bi_4Ti_3O_{12}$," *Ferroelectrics* 3 (1), 17–27.

Flynn, A. M., Tavrow, L. S., Bart, S. F., Brooks, R. A., Ehrlich, D. J., Udayakumar, K. R., Cross, L. E. (1992), "Piezoelectric Micromotors for Microrobots," *J. Microelectromech. Syst.* **1,** 44.

Fork, D. K., Armani-Leplingard, F., Kingston, J. J. (1996), "Application of Electroceramic Thin Films to Optical Waveguide Devices," *MRS Bull.* **21** (7), 53.

Fousek, J., Janovec, V. (1969), "The Orientation of Domain Walls in Twinned Ferroelectric Crystals," *J. Appl. Phys.* **40,** 135.

Fox, G. R., Krupanidhi, S. B., More, K. L., Allard, L. F. (1992), "Composition/Structure/Property Relations of Multi-Ion-Beam Reactive Sputtered Lead Lanthanum Titanate Thin Films: Part I, Composition and Structure Analysis," *J. Mater. Res.* **7,** 3039.

Fujii, T., Watanabe, S. (1996), "Feedback Positioning Cantilever Using a Lead Zirconate Titanate Thin Film for Force Microscopy Observations of Micropattern," *Appl. Phys. Lett.* **68,** 467.

Gaskey, C. J., Udayakumar, K. R., Chen, H. D., Cross, L. E. (1995), "'Square' Hysteresis Loops in Phase-Switching Nb-Doped Lead Zirconate Stannate Titanate Thin Films," *J. Mater. Res.* **10,** 2764.

Herbert, J. M. (1985), *Ceramic Dielectrics and Capacitors,* New York: Gordon and Breach.

Huffman, M. (1995), "Liquid Source Misted Chemical Deposition (LMSCD)—A Critical Review," *Integ. Ferroelec.* **10,** 39.

Hwang, C. S., Park, S. O., Kang, C. S., Kang, H. K., Lee, S. I., Lee, M. Y. (1995), "Deposition of Extremely Thin $(Ba,Sr)TiO_3$ Thin Films for Ultra-Large-Scale Integrated Dynamic Random Access Memory Application," *Appl. Phys. Lett.* **67,** 2819.

Itoh, H., Kashihara, K., Okudaira, T., Tsunemine, Y., Ohno, Y., Nishimura, T., Horikawa, T., Shibano, T. (1995), "Integration of BST Thin Film for DRAM Fabrication," *Integ. Ferroelec.* **11,** 101.

Jacobsen, S. C., Ramsey, M. (1994), "Microchip Capillary Electrophoresis with an Integrated Post Column Detector," *Anal. Chem.* **66,** 3472.

Jaffe, B., Cook, W. R., Jr., Jaffe, H. (1971), *Piezoelectric Ceramics,* Marietta, OH: R. A. N. Publishers.

Jensen, K. (1996), private communication.

Jones, R. E., Jr., Desu, S. B. (1996), "Process Integration for Nonvolatile Ferroelectric Memory Fabrication," *Mater. Res. Bull.* **21** (6), 55.

Jones, R. E., Jr., Zurcher, P., Chu, P., Taylor, D. J., Lii, Y. T., Jiang, B., Maniar, P. D., Gillespie, S. J. (1995), "Memory Applications Based on Ferroelectric and High Permittivity Dielectric Thin Films," *Microelec. Eng.* **29,** 3.

Kamada, T., Takayama, R., Fujii, S., Degushi, T., Hirao, T. (1995), "Pyroelectric Infrared Sensors Made of La-Modified $PbTiO_3$ Thin Films and Their Applications," *Integ. Ferroelec.* **11,** 15.

Kijima, T., Nagata, M. M., Matsunaga, H. (1996), "Ultra Thin Fatigue Free Bi_4Ti_3-O_{12} Films for Nonvolatile Ferroelectric Memories," *Jpn. J. Appl. Phys.* **35** (Pt. 2B), 1246.

Kim, E. S. (1990), "Integrated Microphone with CMOS Circuits on a Single Chip," Ph.D. Thesis, University of California, Berkeley.

Kim, E. S., Muller, R. S. (1987), "IC-Processed Piezoelectric Microphone," *IEEE Electron. Dev. Lett.* **EDL-8,** 467.

Kim, N. (1994), "Grain Size Effect on the Dielectric and Piezoelectric Properties in Compositions which are Near the Morphotropic Phase Boundary of Lead Zirconate-Titanate Based Ceramics," Ph.D. Thesis, The Pennsylvania State University, University Park, PA.

Kingon, A. I., Myers, E. R., Tuttle, B. (1992), *Ferroelectric Thin Films II,* MRS Proceedings Vol. 243, Pittsburgh: MRS.

Kingon, A. I., Streiffer, S. K., Basceri, C., Summerfeldt, S. R. (1996), "High-Permittivity Perovskite Thin Films for Dynamic Random Access Memories," *MRS Bull.* **21** (7), 46.

Kohli, M. (1996), private communication.

Kohli, M., Huang, Y., Maeder, T., Wuetherich, C., Bell, A., Muralt, P., Setter, N., Ryser, P., Forster, M. (1995), "Processing and Properties of Thin Film Pyroelectric Devices," *Microelec. Eng.* **29,** 93.

Kulwicki, B. M., Amin, A., Beratan, H., Hanson, C. M. (1992), "Pyroelectric Imaging," in: A. Sarari, A. Kingon, G. Haertling (Eds.), *8th International Symposium on Applications of Ferroelectrics (ISAF 92),* New Brunswick, NJ: IEEE.

Larsen, P. (1995), private communication.

Lesaicherre, P. Y., Tamamichi, S., Takemura, K., Yamagushi, H., Tokoshiki, K., Miyasaka, Y., Yoshida, M., Ono, H. (1995), "A Gbit-Scale DRAM Stacked Capacitor with ECR MOCVD $SrTiO_3$ over RIE Patterned RuO_2/TiN Storage Nodes," *Integ. Ferroelec.* **11,** 81.

Li, J. F., Viehland, D. D., Tani, T., Lakeman, C. D. E., Payne, D. A. (1994), "Piezoelectric Properties of Sol-Gel Derived Ferroelectric and Antiferroelectric Thin Layers," *J. Appl. Phys.* **75,** 442.

Lines, M. E., Glass, A. M. (1977), *Principles and Applications of Ferroelectrics and Related Materials,* Oxford: Clarendon.

Melnick, B. M., Gregory, J., Paz de Araujo, C. A. (1995), "Characterization of an n-channel 1T-1C Nonvolatile Memory Cell Using Ferroelectric $SrBi_2Ta_2O_9$ as the Capacitor Dielectric," *Integ. Ferroelec.* **11,** 145.

Muralt, P., Kohli, M., Maeder, T., Kholkin, A., Brooks, K., Setter, N., Luthier, R. (1995), "Fabrication and Characterization of PZT-Thin Film Vibrators for Micromotors," *Sensors Actuators* **A48,** 157.

Myers, E. R., Kingon, A. I. (1991), *Ferroelectric Thin Films I,* MRS Proceedings Vol. 200, Pittsburgh: MRS.

Myers, E. R., Tuttle, B. A., Desu, S. B., Larson, P. K. (1993), *Ferroelectric Thin Films III,* MRS Proceedings Vol. 310, Pittsburgh: MRS.

Nakamura, T., Nakao, Y., Kamisama, A., Takasu, H. (1994), "Preparation of Pb(Zr,Ti)O_3 Thin Films on Ir and IrO_2 Electrodes," *Jpn. J. Appl. Phys.* **33** (Pt. 1), 5211.

Newnham, R. E., Wolfe, R. W., Dorrian, J. F. (1971), "Structural Basis of Ferroelectricity in the Bismuth Titanate Family," *Mater. Res. Bull.* **6,** 1029.

Pan, W., Desu, S. B., Yoo, I. K., Vijay, D. P. (1994), "Reactive Ion Etching of $PbZr_{1-x}Ti_xO_3$ and RuO_2 Films by Environmentally Safe Gases," *J. Mater. Res.* **9,** 2976.

Polla, D. L. (1995), "Microelectromechanical Systems Based on Ferroelectric Thin Films," *Microelec. Eng.* **29,** 51.

Polla, D. L., Francis, L. F. (1996), "Ferroelectric Thin Films in Microelectromechanical Systems Applications," *MRS Bull.* **21** (7), 59–65.

Potter, B. G., Jr., Sinclair, M. B., Dimos, D. (1993), "Electrooptical Characterization of Pb(Zr,Ti)O_3 Thin Films by Waveguide Refractometry," *Appl. Phys. Lett.* **63,** 2180.

Ramesh, R., Gilchrist, H., Sands, T., Keramidas, V. G., Haakenaasen, R., Fork, D. K. (1993), *Appl. Phys. Lett.* **63,** 3592.

Reaney, I. M., Brooks, K., Klissurska, R., Pawlacyk, C., Setter, N. (1994), "Use of Transmission Electron Microscopy for the Characterization of Rapid Thermally Annealed, Solution-Gel, Lead Zirconate Titanate Films," *J. Am. Ceram. Soc.* **77,** 1209.

Schwartz, R. W., Dimos, D., Lockwood, S. J., Torres, V. M. (1994), "Preparation and Properties of Sol-Gel Derived PZT Thin Films for Decoupling Capacitor Applications," *Integ. Ferroelec.* **4,** 165.

Scott, J. F., Ross, F. M., Paz de Araujo, C. A., Scott, M. C., Huffman, M. (1996), "Structure and Device Characteristics of $SrBi_2Ta_2O_9$-Based Nonvolatile Random-Access Memories," *MRS Bull.* **21** (7), 33.

Seifert, A., Muralt, P., Setter, N. (1996), private communication.

Shepard, J. F., Jr., Trolie-McKinstry, S., Hendrickson, M. A., Zeto, R. (1996), "Properties of PZT Thin Films as a Function of In-Plane Biaxial Stress," in: B. Kulwicki (Ed.), *Proceedings of the International Symposium on Applications of Ferroelectrics (ISAF 96,)* New Brunswick, NJ: IEEE.

Shimada, Y., Nagano, Y., Fujii, E., Azuma, M., Uemoto, Y., Sumi, T., Judai, Y., Hayashi, S., Moriwaki, N., Nakane, J., Otsuki, T., Paz de Araujo, C. A., McMillan, L. D. (1995), "Integration Technology of Ferroelectrics and the Performance of Integrated Ferroelectrics," *Integ. Ferroelec.* **11,** 229.

Shorrocks, N. M., Patel, A., Walker, M. J., Parsons, A. D. (1995), "Integrated Thin Film PZT Pyroelectric Detector Arrays," *Microelec. Eng.* **29,** 59.

Shuvalov, L. A. (1970), "Symmetry Aspects of Ferroelectricity," *J. Phys. Soc. Jpn.* **38,** 285.

Sinclair, M. B., Dimos, D., Potter, B. G., Jr., Schwartz, R. W., Buchheit, C. D. (1995), "Light Scattering from Sol-Gel Pb(Zr,Ti)O_3 Thin Films: Surface Versus Volume Scattering," *Integ. Ferroelec.* **11,** 25.

Smyth, D. M. (1994), "Ionic Transport in Ferroelectrics," *Ferroelectrics* **151,** 115.

Sonnenschein, B. (1996), "Gene Sequencers Seek to Boost the Spread of their Technology," *R&D Magazine* **38** (2), 41.

Tani, T., Li, J. F., Viehland, D., Payne, D. A. (1994), "Antiferroelectric-Ferroelectric Switching and Induced Strains for Sol-Gel Derived Lead Zirconate Thin Layers," *J. Appl. Phys.* **75,** 3017.

Trivedi, D., Tayabati, P., Tabat, M. (1996), "Measurement of Large Electro-Optic Coefficients in Thin Films of Strontium Barium Niobate ($Sr_{0.6}Ba_{0.4}Nb_2O_6$)," *Appl. Phys. Lett.* **68,** 3227.

Trolier-McKinstry, S., Hu, H., Krupanidhi, S. B., Chindaudom, P., Vedam, K., Newnham, R. E. (1993), "Spectroscopic Ellipsometry Studies on Ion Beam Sputter Deposited Pb(Zr,Ti)O_3 Films on Sapphire and Pt-Coated Silicon Substrates," *Thin Sol. Films* **230,** 15.

Trolier-McKinstry, S., Randall, C. A., Maria, J. P., Theis, C., Schlom, D. G., Shepard, J., Jr., Yamakawa, K. (1995a), "Size Effects and Domains in Ferroelectric Thin Film Actuators," in: S. B. Desu, R. Ramesh, B. A. Tuttle, R. E. Jones, I.-K. Yoo (Eds.), *Ferroelectric Thin Films V,* MRS Proceedings Vol. 433, Pittsburgh: MRS.

Trolier-McKinstry, S., Chen, J., Vedam, K., Newnham, R. E. (1995b), "*In situ* Annealing Studies of Sol-Gel Ferroelectric Thin Films by Spectroscopic Ellipsometry," *J. Am. Ceram. Soc.* **78,** 1907.

Tuttle, B. A., Schwartz, R. W. (1996), "Solution Deposition of Ferroelectric Thin Films," *Mater. Res. Bull.* **21** (6), 49.

Tuttle, B. A., Desu, S. B., Ramesh, R., Shiosaki, T. (1994), *Ferroelectric Thin Films IV,* MRS Proceedings Vol. 361, Pittsburgh: MRS.

Tuttle, B. A., Garino, J. J., Voight, J. A., Headley, T. J., Dimos, D., Eatough, M. (1995), in: O. Auciello, R. Waser (Eds.), *Science and Technology of Electroceramic Thin Films,* Boston: Kluwer Academic.

Udayakumar, K. R. (1996), private communication.

Udayakumar, K. R., Chen, J., Flynn, A. M., Brooks, K. G., Cross, L. E., Ehrlich, D. J. (1992), "Piezoelectric Thin Film Ultrasonic Micromotors," in: A. I. Kingon, E. R. Myers, B. Tuttle (Eds.), *Ferroelectric Thin Films II,* MRS Proceedings Vol. 243, Pittsburgh: MRS.

Voight, J. A., Tuttle, B. A., Headley, T. J., Eatough, M. O., Lamppa, D. L., Goodnow, D. (1993), "Oriented Lead Zirconate Titanate Thin Films: Characterization of Film Crystallization," in: E. R. Myers, B. A. Tuttle, S. B. Desu, P. K. Larson (Eds.), *Ferroelectric Thin Films III,* MRS Proceedings Vol. 310, Pittsburgh: MRS.

Wakamura, T., Nakao, Y., Kimasawa, A., Takasu, H. (1995), "PZT on Ir Doped Pt: No Fatigue," *Integ. Ferroelec.* **11,** 161,

Warren, W. L., Robertson, J., Dimos, D. B., Tuttle, B. A., Smyth, D. M. (1994), "Transient Hole Traps in PZT," *Ferroelectrics* **153,** 303.

Waser, R. (1994), "Charge Transport in Perovskite-Type Titanates: Space Charge Effects in Ceramics and Films," *Ferroelectrics* **151,** 125.

Waser, R., Klee, M. (1992), "Theory of Conduction and Breakdown in Perovskite Thin Films," *Integ. Ferroelec.* **2,** 23.

Watton, R. (1994), "Ferroelectric Materials and IR Bolometer Arrays: From Hybrid Arrays Towards Integration," *Integ. Ferroelec.* **4,** 175.

Wolfe, R. W., Newnham, R. E., Smith, D. K., Kay, M. I. (1971), "Crystal Structure of Bi_3TiNbO_9," *Ferroelectrics* **3,** 1.

Yamakawa, K., Trolier-McKinstry, S., Dougherty, J. P. (1996), "Electrical Properties and Phase Transformations in Antiferroelectric Lead Zirconate Thin Films," in: B. Kulwicki (Ed.), *Proceedings of the International Symposium on Applications of Ferroelectrics,* New Brunswick, NJ: IEEE.

Further Reading

General Information on Ferroelectricity

Setter, N., Colla, E. L. (1993), *Ferroelectric Ceramics. Tutorial Reviews, Theory, Processing and Applications,* Boston: Birkhauser.

Xu, Yuhuan (1991), *Ferroelectric Materials and Their Applications,* Amsterdam: North-Holland.

Practical Applications of Ferroelectrics

Krupanidhi, S., Kurtz, S. K. (1990), *7th International Symposium on Applications of Ferroelectrics (ISAF 90),* New Brunswick, NJ: IEEE.

Pandey, R. K., Liu, M., Safari, A. (1994), *9th International Symposium on Applications of Ferroelectrics (ISAF 94),* New Brunswick, NJ: IEEE.

Safari, A., Kingon, A., Haertling, G. (1992), *8th International Symposium on Applications of Ferroelectrics (ISAF 92),* New Brunswick, NJ: IEEE.

Specific to Ferroelectric Thin Films

Desu, S. B., Ramesh, R., Tuttle, B. A., Jones, R. E., Yoo, I. K. (Eds.) (1996), *Ferroelectric Thin Films V,* MRS Proceedings Vol. 433, Pittsburgh: MRS.

Kingon, A. I., Myers, E. R., Tuttle, B. A. (Eds.) (1992), *Ferroelectric Thin Films II,* MRS Proceedings Vol. 243, Pittsburgh: MRS.

Myers, E. R., Kingon, A. I. (Eds.) (1990), *Ferroelectric Thin Films I,* MRS Proceedings Vol. 200, Pittsburgh: MRS.

Paz de Araujo, C. A., Scott, J. F., Taylor, G. W. (1996), *Ferroelectric Thin Films: Synthesis and Basic Properties,* New York: Gordon and Breach.

Tuttle, B. A., Myers, E. R., Desu, S. B., Larsen, P. K. (1993), *Ferroelectric Thin Films III,* MRS Proceedings Vol. 310, Pittsburgh: MRS.

Tuttle, B. A., Desu, S. B., Ramesh, R., Shiosaki, T. (Eds.) (1994), *Ferroelectric Thin Films IV,* MRS Proceedings Vol. 361, Pittsburgh: MRS.

TIDAL AND WAVE ENERGY

RICHARD J. SEYMOUR, *Texas A&M University, College Station, Texas, U.S.A., and the University of California–San Diego, La Jolla, California, U.S.A.*

Introduction 453
1. **Tidal Energy** 455
1.1 Physical Principles of Tides 455
1.2 Global Distribution of Tidal Energy 457
1.3 Utilization of the Tidal Resource 458
1.3.1 Types of Power-Extraction Systems 458
1.3.2 Tidal-Power Installations 459
1.3.3 Environmental Effects 459
1.3.4 Economics 460
2. **Wave Energy** 460
2.1 Physical Principles of Wind-Generated Ocean Waves 460
2.1.1 Global Distribution of Wave Energy 460
2.1.2 Spectral and Directional Distribution of Wave Energy 462
2.2 Utilization of the Wave Resource 462
2.2.1 Types of Power-Extraction Systems 462
2.2.2 Wave-Power Installations 464
2.2.3 Environmental Effects 465
2.2.4 Economics 465
Glossary 465
Works Cited 466
Further Reading 466

INTRODUCTION

Tides and wind-driven waves in the ocean are major repositories of renewable energy at global scales. Astronomically forced tides acquire their energy from gravitational attraction forces produced by the moon and sun. Tides are simply a particular class of ocean waves. The remainder of the ocean waves are dominated energetically by wind-generated types. The energy source for these waves is heat from the sun, which drives the variations in the distributions of pressure within the Earth's atmosphere that result in winds. Tides and, to a lesser extent, other ocean waves are responsible for currents that are locally very important in the transport of water masses, suspended and dissolved material distribution, and the erosion and accretion of shorelines and the seafloor near the ocean boundaries. Currents driven by winds and by density differences such as those caused by varying water temperatures and salinities are principally responsible for the mixing and transport phenomena within ocean basins. However, very near the shoreline, the movements caused by tides and waves are often the dominant mechanisms.

Energy in any system of gravity waves, tides included, is divided between potential and kinetic forms. Linear wave theory provides that there is equipartion between kinetic and potential energy in surface gravity waves of low steepness in deep water. The assumption of equipartition has been extended, in practice, to all waves in all depths of water. Wind waves are inherently random in amplitude and with a broad spread of directions such that estimates of their energy density are averaged both spatially and temporally to provide useful engineering values. Tides, conversely, are very predictable and change sufficiently slowly that meaningful information is contained in the discrete elevation history.

Attempts have been made to estimate the global level of energy in tides and wind waves (Isaacs and Seymour, 1973; Isaacs and Schmitt, 1980) with assumed values on the order of 3 TW for each resource. Power has been extracted from both sources in a limited number of applications representing only a minuscule drain on the overall resources. Tidal power, perhaps because of its predictability and accessibility in sheltered locations, was the first to be utilized in pre-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

vious centuries through waterwheel-driven mills. In the last half of this century, a few sites have been exploited to produce electricity from tides (Warnock and Clark, 1992), and a significant effort has been undertaken internationally to develop a similar capability from wave power (McCormick, 1981; Carmichael and Falnes, 1992).

Tides have been observed and predicted by mariners for many centuries, and their correlation with the moon phases has been understood (Darwin, 1911). The earliest records of continuous observations date from the thirteenth century at London Bridge on the Thames. Newton in the Principia explained the astronomical tide-generating force in 1687 and provided a basis for the number of tidal cycles per day. Lord Kelvin, in about 1870, was the first to develop the concept of the tides as a sum of harmonic components. This led directly to the modern methodology for tide prediction. The frequencies of the components are developed from the detailed knowledge of the motions of the sun and the moon relative to the Earth. Measurements are made at a specific site, typically hourly for 18.6 years to encompass the longest astronomical period, and the phases and amplitudes of the tidal responses are determined for the designated frequencies. The coefficients can then be applied to future values of the astronomical forcing functions and summed to make predictions of the tides at that location. Tides at locations between observation points are estimated with lesser accuracy by interpolation and, in some cases, by measurement over shorter periods of time.

Laplace (Lamb, 1932) developed a dynamical theory of tides through a set of shallow-water wave equations for a fluid layer on a spherical Earth, derived from the conservation of mass and momentum. These form the basis for modern tide theory, later enhanced by a number of theoreticians to include consideration of density stratification, Coriolis forces, varying depth, and the effects of solid-Earth tides. Realistic solutions to the modified Laplace equations have been obtained numerically for global tides. A comprehensive review of the development of modern theory is found in Hendershott (1981).

The understanding that winds act on the surface of oceans and lakes to produce waves has been available for centuries. However, theoretical work on these interactions between wind and waves has been relatively recent, occurring entirely within the twentieth century. The initial efforts were directed at attempts to understand the influence of the ocean on the boundary layer of the atmosphere. Charnock (1955) first proposed that the roughness length over the sea surface should vary as the wind speed. Drag coefficients have been determined empirically by a number of investigators (Large and Pond, 1981; Wu, 1982). Although weakly dependent upon wind speed, drag coefficients were found to vary significantly. Attempts followed to factor in wave characteristics to explain the variation in the drag coefficient (Janssen, 1989). None of this work, however, provided any significant insights into the mechanisms whereby the wind stress transmitted energy and momentum to the waves.

During World War II, an effort was begun to develop empirical models for estimating the wave-generation process from knowledge of the wind fields over the ocean. Bretschneider (1952) extended the wartime work of Sverdrup and Munk (1947) in developing semiempirical models relating the wind speed, duration, and fetch—the distance over which the wind acts—with the height and characteristic period of the waves. Pierson *et al.* (1953) provided a method for predicting a spectrum of wave energy, an improvement over the two-parameter height-period characterization. More recently the JONSWAP group (Hasselmann *et al.*, 1973) developed a refined spectrum from an extensive wave-generation observation program in the North Sea.

Waves that are still increasing in size as a result of the wind are referred to as *sea;* those that travel beyond the generation area are called *swell.* Snodgrass *et al.* (1966) showed that swell could propagate from Antarctica to the western coast of North America without significant loss of energy. Because swell obeys a dispersion relationship such that energy in deep water propagates at a speed inversely proportional to frequency, the swell is dispersed over a greater area over time and the height diminishes as the energy is spread. In many locations, swell dominates the local wave climate. In the Pacific and the South Atlantic, very energetic swell with characteristic heights of more than 3 m is not uncommon.

Significant research programs have been undertaken internationally, particularly in the 1970s, to develop methodologies for extracting energy from waves. Although extensive laboratory testing and some field deployments have been accomplished, only a limited number of installations have reached the commercial stage. Although environmentally benign and a truly renewable resource, wave energy is inherently of low quality as a power source because of its extreme variability, and the mechanisms are subject to extreme loads, stress corrosion, and fatigue.

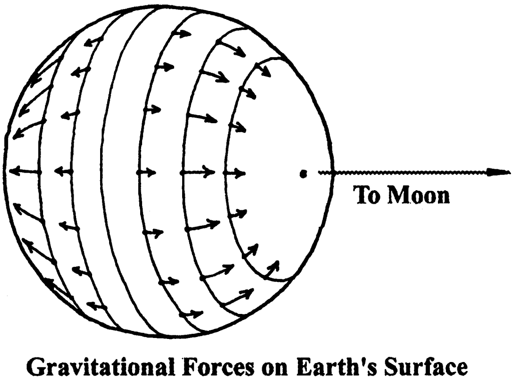

FIG. 1. Tangential components at the Earth's surface resulting from the gravitational attraction of the moon. Similar, but smaller, forces are produced by the sun.

1. TIDAL ENERGY

1.1 Physical Principles of Tides

The astronomical forcing of tides results from the gravitational attraction of the moon and the sun. The rotation of the Earth and the perturbations of the orbits of the Earth about the sun and of the moon about the Earth result in time-varying forces when viewed from any spot on the Earth's surface. The direct attraction of the moon is the principal component, that of the sun second, both with periods on the order of half a day. The variations produce attraction components of lesser degree with periods ranging from about a day to more than a decade. Typically, as many as 30 periodic components are employed to represent the complete astronomical attraction force. At a particular spot on the Earth's surface, each periodic component produces a small horizontal oscillatory force with a unique amplitude and phase. The distribution of these forces for a single component is illustrated in Fig. 1.

Lord Kelvin (Lamb, 1932) showed a solution for the Laplace shallow-water wave equations on a rotating sphere that produces a trapped wave at the shoreline. For waves traveling around an ocean basin in an counterclockwise direction in the Northern hemisphere, these Kelvin waves have maximum amplitude at the shoreline, decaying exponentially with offshore distance. For waves of infinitesimal height, the wave velocities are everywhere zero normal to the shoreline. On a sloping shore with real waves, a small normal component is required to conserve mass, such that the orbits of Kelvin waves are highly flattened ellipses, as shown in Fig. 2, with the long axis parallel to the shoreline. The Kelvin wave provides a capability to form resonant wave systems in ocean basins where an integer number of waves can fit into a rotary pattern with a wave crest passing a point about every 12 or 24 h. The resonant pattern corresponding to the principal lunar forcing component for a portion of the North Atlantic is shown in Fig. 3. The apparent center of rotation of the system, which is

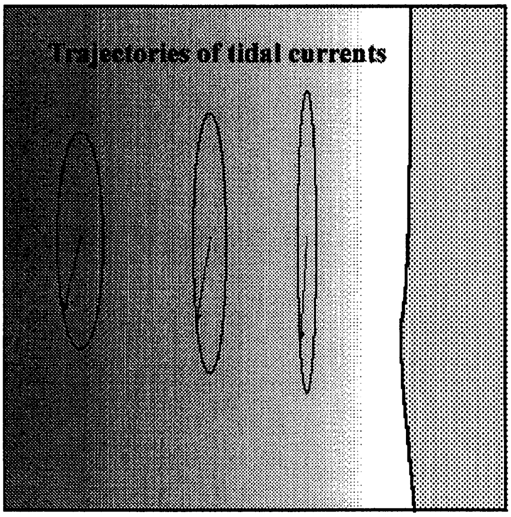

FIG. 2. The velocities produced by Kelvin waves on the continental shelf follow orbits that are flattened ellipses, particularly in shallow water. The dominant tidal currents along relatively straight coastlines are therefore largely parallel to the shoreline.

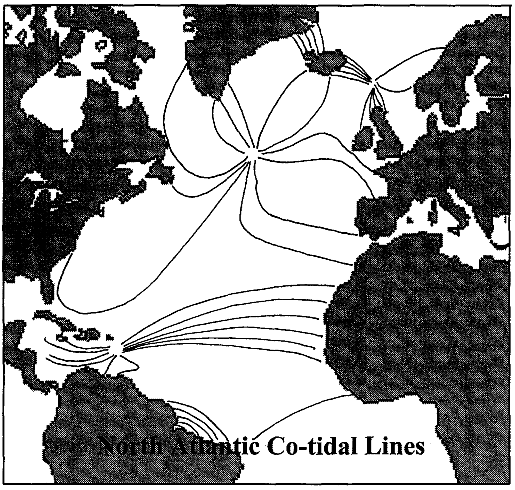

FIG. 3. The co-tidal lines, or regions with the same tidal phase, for a single component of the tide in the North Atlantic. Thus, the lines can be interpreted as the location of the crests of the Kelvin waves at one instant in time. The rotary system, which moves counterclockwise in the Northern Hemisphere, is clearly apparent.

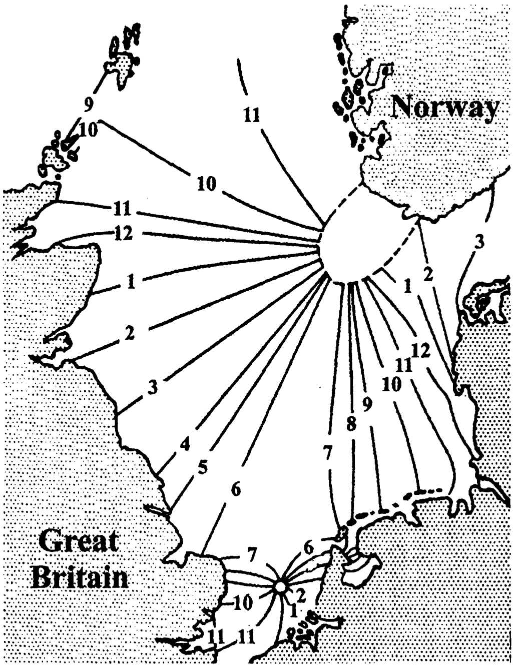

FIG. 4. An illustration of the interaction between systems of rotary tidal waves in the North Sea, with relative times of high tide in hours after the moon passes over Greenwich. Only a single tide component is shown. The smaller system in the channel propagates slowly because of the reduced water depth and is therefore able to remain in phase with the larger North Sea system.

a point of disappearingly small tidal excursion, is called an amphidromic point. The propagation speed of these trapped coastal waves is determined by the water depth, as given by

$$c = \sqrt{gh}, \tag{1}$$

where c = celerity of the tidal wave, g = gravitational constant, and h = water depth.

Because slower waves in shallow water basins have proportionately shorter wavelengths, the overall size of these shallow systems is smaller than those in deep ocean basins. This allows neighboring systems of rotary resonances to mesh constructively. An example of these interactions in the vicinity of the British Isles is shown in Fig. 4.

In many locations, long narrow embayments, such as the Adriatic, the Bay of Fundy, and the Gulf of California, open to large seas or ocean basins. A Kelvin wave beginning at the open end of such a marginal sea and traversing the right-hand shoreline will reflect from the end and continue down the opposite shore. The effect of the interactions between the incoming and outgoing wave is to form what appears to be a rotary system of waves with one or more amphidromic points near the center of the bay. In the three examples mentioned, the amplification that occurs from shoaling and narrowing in the shallower water at the head of the sea results in tidal heights that are many times the magnitude that occurs at the open end where tide characteristics must match those of the ocean basin. For example, the Adriatic Sea has very substantial tides at its northern end, in the vicinity of Trieste, but matches the nearly tideless Mediterranean at its southern opening.

The result is that any water body of the right shape, size, and depth to support a resonant wave system with a period close to 12 or 24 h will exhibit a recognizable tide, provided that the resonant system can exist in synchrony with the resonant tidal waves in neighboring and connected bodies of water. If these conditions are not met, the tides in that basin or region will be negligible as, for example, in the Baltic Sea.

Tides, as gravity waves, are variations about mean sea level. If worldwide sea level changes by tens of meters, as it has over geological time scales, the pattern of tides will change radically as well. This is caused by the detuning effect of changing both the water depth (which changes the speed of propagation of the tidal waves) and the size and shape of the water body substantially, affecting the ability to establish resonant wave systems at the appropriate periods. Temporary changes in sea level can result from wind stress. Hurricanes and typhoons can raise sea level by several meters along local sections of wide and shallow continental shelves. Global changes in climate, such as the El Niño–Southern Oscillation event that alters wind patterns over the Pacific, can result in changes of tens of centimeters along the boundaries of whole ocean basins for intervals on the order of a year. Astronomical tides persist as variations about these temporarily modified sea levels, although, in general, the depth and geometry changes are either too local or too small in amplitude to result in significant changes to astronomical tides.

In shallow basins such as the Patagonian shelf and the Bering Sea, the friction of tidal currents can reach significant levels (Hendershott, 1981). Energy is conserved within the sun-moon-earth system by a gradual slowing of the Earth's rotational speed, resulting in a relative increase in the length of the Earth day. Angular momentum is conserved by an appropriate increase in the distance between the moon and the Earth. Hendershott (1981) provides a comprehensive review of the evolution of the physical oceanography of tides.

1.2 Global Distribution of Tidal Energy

Rotary systems of Kelvin waves account for most of the tidal energy in the large ocean basins of the world, as shown in Fig. 5. Although difficult to convey with a single diagram, the tides are the sum of many rotary wave systems that are largely linearly independent. With a few exceptions, the tidal ranges are limited to about 2 m or less, resulting in energy densities too small to be of economic interest. Along the Patagonian coast and the French coast of the English Channel, for example, tidal waves do achieve substantial ranges (4 to 6 m). For the most part, however, such large tidal ranges are limited to a few scattered locales around the world where marginal seas, embayments,

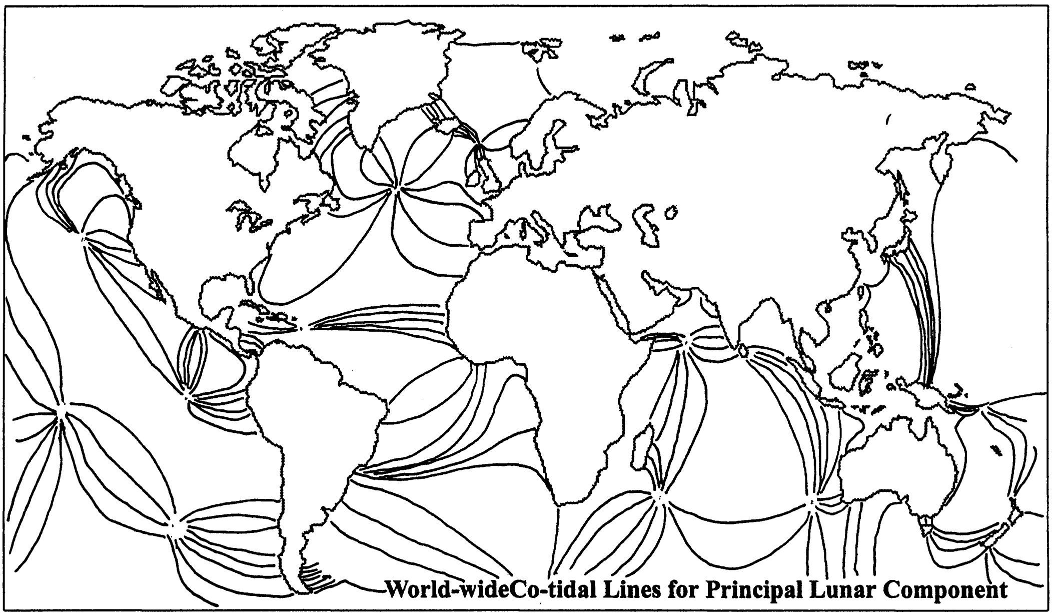

FIG. 5. One component of the world tidal system is shown to illustrate the interactions between ocean basins. The world tide is seen to be dominated by rotary systems of Kelvin waves.

and estuaries provide sites for amplification. Rivers at the heads of these locations may feature substantial reversing bores that can convey tidal energy up to tens of kilometers inland. Ippen and Harlemann (1966) review the tidal dynamics in estuaries. Figure 6 indicates where tidal ranges reach what might be economically attractive size.

1.3 Utilization of the Tidal Resource

Tidal power has been employed for mechanical energy in small, local applications for centuries, analogous to the use of windmills, to power machinery of various types. Early in the twentieth century there was strong interest in generating electrical power in regions without access to conventional hydropower. Studies of electrical power plants in the Bay of Fundy go back at least as far as the 1920s. For the reasons discussed in the following sections, the number of actual installations has remained small and is not expected to grow substantially.

1.3.1 Types of Power-Extraction Systems Existing and planned tidal-power installations follow a similar plan. A relatively small basin adjacent to a larger body of water with a substantial tidal range is selected for modification. A dam (also referred to regionally as a dyke or a barrage) is constructed to close off the basin. Openings are provided in the dam wall so that controlled flows can occur between the basin and the outer body. This system can be arranged for single-effect or double-effect power extraction. In a single-effect system, the water is allowed to flow relatively unimpeded through sluice gates into the basin on the rising tide. As the tide outside the basin begins to fall, these gates are closed, and when the head between the basin and the falling level outside reaches about half of its maximum potential, the outflow is allowed to pass through one or more low-head turbines, which drive generators and produce electricity. In a double-effect system, the turbines are capable of operating with flows in either direction, similar to reversing the propellers on an aircraft, and so generate on both rising and falling tide stages. The generators can be operated as motors, allowing the turbines to be used as pumps so that the system can provide for pumped storage to assist in load matching when the tide cycles are fav-

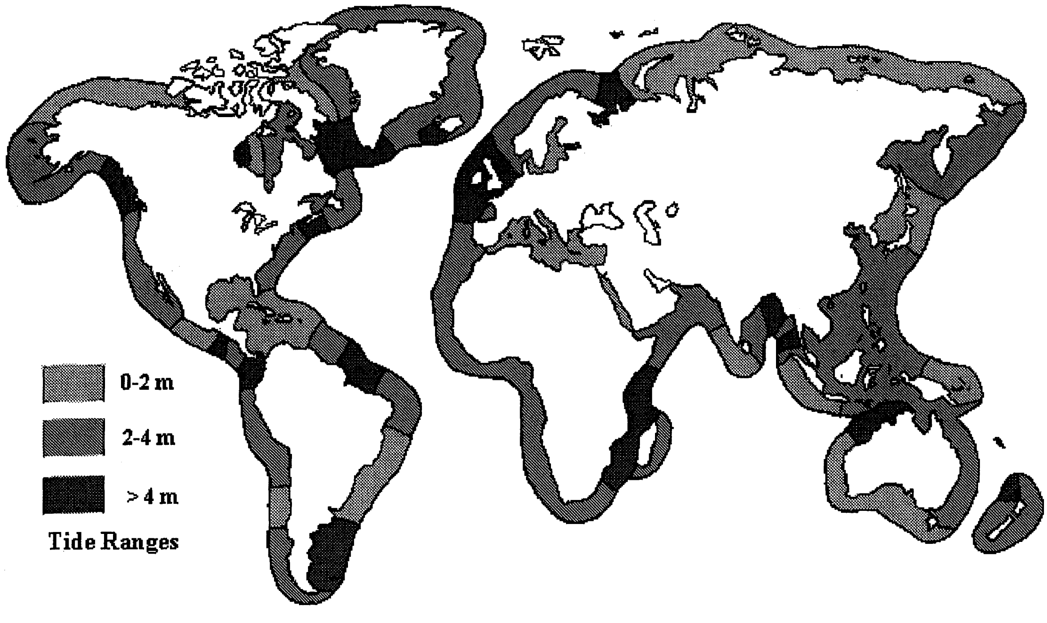

FIG. 6. Tidal ranges vary significantly globally. The figure indicates broad categories of ranges. The areas with darkest shading are macrotidal ranges (>4 m), the intermediate shading defines the mesotidal ranges (2–4 m), and the lighter shading defines the microtidal ranges (0–2 m).

orable. In double-effect systems, the turbine blades can be fully feathered to assist the sluice gates in allowing flow into the basin. These turbines operate at very low heads (4–15 m), contrasted with the hundreds of meters available in many hydroelectric systems, and are therefore expensive and inefficient by comparison. The development of bulb-type submerged generators, coaxial with the turbines as shown in Fig. 7, has made the modern tidal-power installations feasible. A recent development has allowed even more efficient flow configurations by mounting the generator in a ring outside of the turbine blades, such that the flow is entirely inside of the generator rotor. A review of the state of the art is found in Warnock and Clark (1992).

1.3.2 Tidal-Power Installations The four existing major tidal-power installations in the world were all constructed in the last half of the twentieth century. The largest is at La Rance on the Brittany coast of France. It contains 24 generators rated at 10 MW each and operates with a 8.5-m tidal range. The turbines are double effect, and the units can be used for pumped storage. The second largest is at Annapolis, Nova Scotia, Canada. Although many sites have been studied on the Bay of Fundy, this is the only plant yet constructed. It took advantage of an existing causeway to provide a basin. The single generator is of the concentric type, rated at 20 MW, and the turbine runner is 7.6 m in diameter. It is considered a prototype for modern tidal-power installations. The third largest is at Jiangxia, Zhejiang Province, China. Again, an existing dam intended to reclaim marshland was employed. Five bulb-type and one concentric generators were installed for a total capacity of 3.9 MW. The mean tidal range is 5 m. The bulb units are double effect. The remaining unit is at Kislaya Guba on the Barents Sea coast of Russia. It is a modest unit with a single (400-kW) bulb turbine but was intended to be the prototype of a 10 000-MW installation at Tugur Bay on the Sea of Okhotsk and a 15 000-MW installation at Mezen Bay on the White Sea. Neither of these larger units was ever constructed.

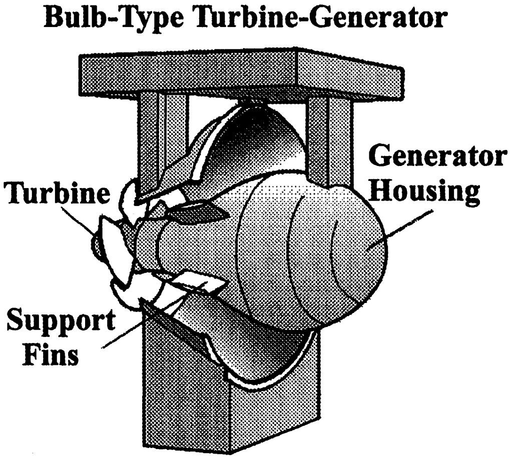

FIG. 7. The general configuration of a bulb-type, low-head turbine generator used in modern tidal-energy installations. The generator is mounted in the bulbous housing and is directly coupled to the turbine blades.

During the last two decades of the twentieth century, a number of major sites have been investigated. These are shown in Table 1. Of these, only the two in the United Kingdom, the Severn, and the Mersey Estuaries showed sufficient economic potential to proceed into detailed engineering design. Both of these sites are considered as candidates, but no construction decisions have been made.

1.3.3 Environmental Effects Concern for adverse environmental impacts appears to have been as important as weak economic justification in delaying or canceling proposed tidal-power projects. Two of the existing major plants involved minor changes to barrages constructed years earlier before the international environmental movement coalesced. In this manner, they avoided being assessed by both the major cost of the barrage and the responsibility for disturbances

Table 1. Proposed sites for major tidal-power installations.

Site	Proposed capacity (MW)
Cook Inlet, Alaska, U.S.A.	1440
Garolim Bay, Korea	400
Gulf of Kachchh, India	600
Site B9, Bay of Fundy, Canada	5338
Severn Estuary, United Kingdom	8640
Mersey Estuary, United Kingdom	700

to wetlands. Salinity and sedimentation patterns can be substantially altered by the modifications to the flow characteristics, which results in stress to the unique assemblages of plants and animals that colonize the large intertidal regimes. The impediments to migratory fish stocks and the fish mortality associated with passage through the turbines have resulted in substantial objections by commercial and sports fishermen to proposed installations. Numerical modeling of the potential effects of large Bay of Fundy projects has resulted in predictions of measurable modification of the ocean tides along the northern coast of the United States (Greenberg, 1984). Although these predicted changes are small, they could, in principle, result in more frequent closing of Logan Airport in Boston Harbor because of water on the runway. The liability from such an occurrence is one of the factors cited in reducing the economic attractiveness of Bay of Fundy power installations. Tidal-power generating stations have a beneficial environmental effect in minimizing the production of greenhouse gases. Because of the difficulty in predicting the influence of the installation on the natural production of carbon dioxide, it is not possible to state that this power source results in no atmospheric pollution. Daborn (1976) reviews the overall environmental impacts in the Bay of Fundy.

1.3.4 Economics The barrages and the generating equipment in tidal plants have exceptionally long working lives, which tends in part to offset the very high installed cost. Further, the load factor is relatively low so that the capital cost per unit of power generated is further increased. With relatively low maintenance and no fuel costs, the amortization of the capital cost and interest on the investment constitute 85–90% of the total energy cost for tidal-power schemes (Shaw, 1992). The value of the energy produced depends to a large degree on the local market. Even with pumped-storage potential, it is never possible to match the plant output with typical demand cycles because the dominant tidal period is slightly greater than 12 h. This degrades the value of the energy when compared with sources that can provide true load-matching capabilities. The recent history of economic analyses suggests that only in those cases where the cost of the barrage can be partially or fully borne by some other application, such as a bridge connecting otherwise remote shorelines, can tidal power be competitive with thermal power plants. A review of the factors influencing tidal-power economics is found in Shaw (1992).

2. WAVE ENERGY

2.1 Physical Principles of Wind-Generated Ocean Waves

The transfer of energy and momentum from wind to waves has been studied extensively over most of this century, but no general theory has yet been agreed upon. Jeffreys (1925) proposed that waves were generated by wind stress acting normal to the sea surface. Sverdrup and Munk (1947) added the energy transfer due to the tangential component of stress. A series of basically empirical models, beginning during World War II, has resulted in a useful ability to predict wave growth from a knowledge of the wind field. However, the precise physical mechanisms involved have yet to be explained. Presently, considerable effort is centered upon considering the state of development of the waves (Donelan *et al.*, 1993). In all of this work, wind stress on the surface has been treated as a scalar aligned with the mean wind direction. Recently, Rieder *et al.* (1994) have shown that the direction of the wind stress is strongly influenced by the direction of pre-existing swell waves that are not aligned with the wind.

The models for wave generation are based upon the general equation

$$\frac{\partial E}{\partial t} + v_j \frac{\partial E}{\partial x_j} = S_i + S_n + S_d, \tag{2}$$

where E = wave energy, v_j = group velocity along coordinate x_j, S_i = atmospheric input source term, S_n = nonlinear wave–wave interaction source term, and S_d = dissipation term. The effort by investigators, utilizing field observations and theoretical considerations, has been to develop relationships for the three source terms in equation (2).

2.1.1 Global Distribution of Wave Energy Three global-scale features dominate the distribution of wave energy on a climatic basis. The first is the existence of the trade-wind bands in the tropics on either side of the equator. Most of the year, the wind is comparatively strong, steady, and unidirectional (trade winds, blowing toward the west) at these latitudes resulting in one of the most consistent wave climates anywhere on the globe, particularly in the Pacific where the wind can act on the waves for a very long distance. Because, under these circumstances, the waves develop toward a spectrum with energy peaked at a frequency that reaches a propagation speed equal to the wind speed, the rate of input of additional energy decreases sharply. Further, slowly occurring nonlinear interactions among wave frequencies result in the creation of waves of lower frequency, which run ahead of the waves in the generation area. This results in a very constant appearance of low to moderate swell in the tropical Pacific. Although not as energetic as some of the waves produced by higher-latitude winds, the constancy of the resource is attractive for certain energy-extraction scenarios. The second important global-scale effect is the very energetic storm sequences that are found in higher latitudes (generally toward the east) during the winter season for either hemisphere. This results in a seasonality to the wave climate that introduces problems in energy extraction, even though the resource is large averaged over the year. The third feature is the particular geometry of the continental land masses in the southern hemisphere. The band of ocean surrounding Antarctica is the only place on Earth where there is a continuous fetch. That is, most of the energy in the waves generated by the strong winds blowing out of the west is never dissipated by striking a shoreline, but it is continuously maintained at the highest possible energy density as it travels endlessly around the world. The result is the highest waves and the highest average wave-energy densities of any place on Earth. No serious plan has been made to use this resource both because of its extreme nature and because of the relatively lower power demands in the Southern Hemisphere. A generalized representation of the global average wave-energy densities is shown in Fig. 8. These represent deep-water values. There is a marked reduction in energy density as the waves approach shore. The power associated with

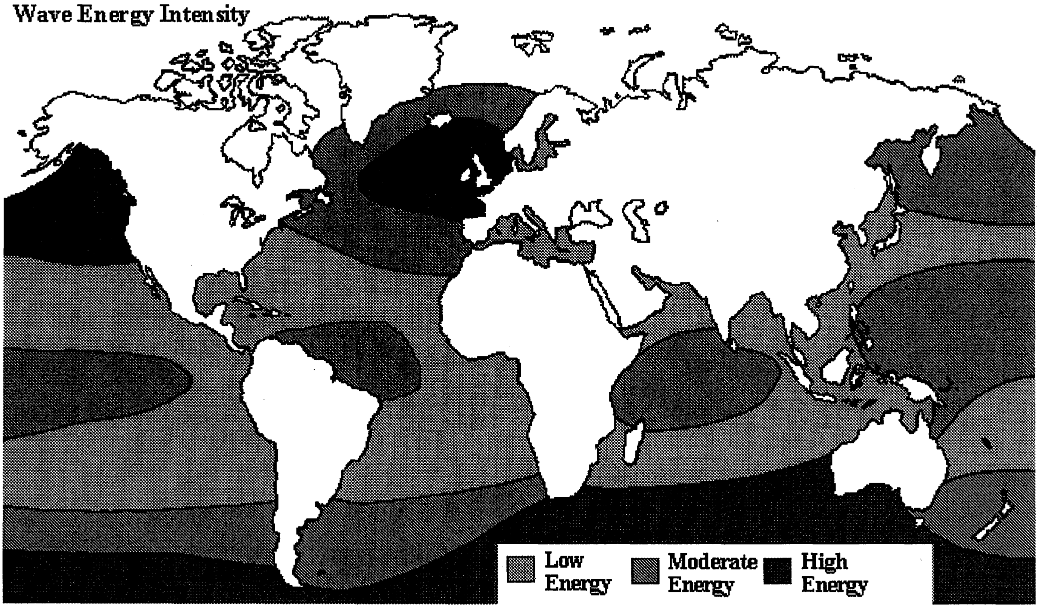

FIG. 8. A generalized view of the global distribution of relative average wave energy. Darker shading indicates a more energetic wave climate. The North Sea, the area in which the greatest research and development activities have been undertaken, is a high-energy region.

ocean waves can be calculated through the equation

$$P = (\rho g^2/64\pi)H_s^2T_m, \tag{3}$$

where P = wave power per unit of wave front in deep water, ρ = density of seawater, g = acceleration of gravity, H_s = significant wave height, given by four standard deviations of the surface elevation variation, and T_m = period associated with the maximum energy in the wave spectrum.

2.1.2 Spectral and Directional Distribution of Wave Energy Wind imparts energy to waves in all directions with the energy distributed roughly proportional to $\cos^n(\Theta/2)$, where Θ is measured relative to the mean wind direction. The exponent n increases with decreasing frequency so that higher-frequency waves are more broadly spread in direction. The generation mechanism results in wave energy being broadly spread in frequency as well. The spectrum of wave energy at the point of generation is shown generically in Fig. 9. The energy is sharply peaked, falling off nearly vertically at lower frequencies and approximately proportional to f^{-5} at frequencies f higher than the peak. As previously discussed, when the waves propagate away from the generation area, or if the wind diminishes, the spectrum narrows in frequency because of dispersion. Further, the directional spreading narrows as well as the wave field diverges. As a result, incident wave energy varies from the fully spread sea within the generating area to swell that is essentially unidirectional and very sharply peaked in frequency. This imposes design restrictions on wave-power systems, which must be capable of effective operation in a wide range of conditions.

2.2 Utilization of the Wave Resource

2.2.1 Types of Power-Extraction Systems There have been a very large number of systems proposed for harnessing the energy in waves. Internationally, the patents granted in this area number in the hundreds. However, it is possible to classify almost all of the systems into three broad types. The first, the tapered-channel system, in which energy is converted into hydrodynamic head behind a low dam, has many of the characteristics of the tidal-power installations. This system is shown schematically in Fig. 10. The second broad category is the caisson type in which a bottom-resting chamber with

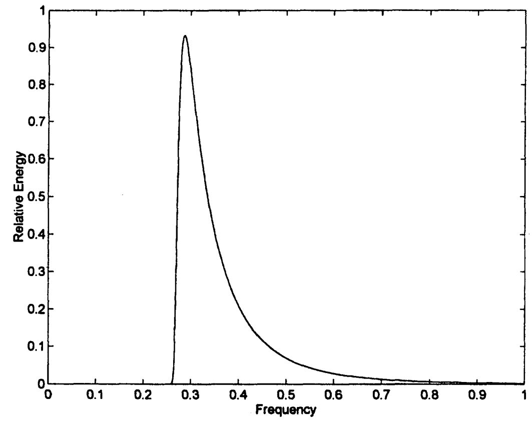

FIG. 9. A generalized spectrum of wind wave energy in a fully developed sea, i.e., one where the waves are in equilibrium with the wind. Coordinate units are arbitrary. The wave components at the peak of the spectrum are moving at a phase speed approximately equal to the wind speed. The energy falls off rapidly at either side of the peak.

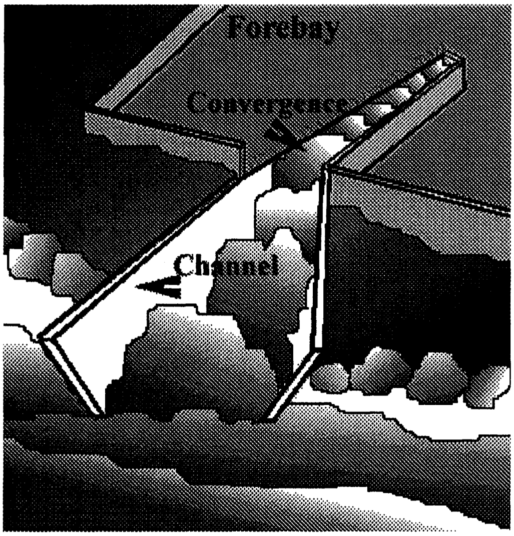

FIG. 10. A schematic representation of the tapered-channel wave-energy extraction system. The waves build rapidly in height as they traverse a converging channel and then spill over a collector system into a forebay that maintains the level above the local sea level. Energy is extracted through low-head turbines similar to those employed in tidal-power plants.

one or more openings below the surface captures the wave energy through the pressure oscillations in an air column above the water or by mechanically driving a float or flap. A typical pneumatic caisson type is shown in Fig. 11. The third type encompasses the floating systems in which the heaving or swaying motion of the hull is converted into mechanical energy. Carmichael and Falnes (1992) and Hagerman (1992) provide details of the various mechanisms.

Waves approaching shallow water tend to increase in crest height until they become unstable at a depth of the same order as their height and break, resulting in the conversion of most or all of the energy into small-scale turbulence and heat. If the shoaling is very rapid, a wave may not have time to break and will uprush onto the shoreline to substantially above its crest height. If a wave enters a channel with converging walls, the energy in the wave is trapped between the narrowing channel boundaries and increases in magnitude until it becomes unstable and breaks. The first type of extraction device, the tapered channel, utilizes both of these effects and also borrows technology from tidal plants. A natural basin is selected facing a shoreline with water of sufficient depth so that typical waves do not break before reaching it. The basin is sealed off with a low dam, similar to that employed in tidal basins. An artificial channel, or an artificially enhanced natural channel, is constructed that rapidly decreases in depth and narrows in width approaching the dam. The resulting uprush of the moderate to large waves is sufficient to carry a substantial quantity of water over the top of the dam, trapping it in the basin. Power is extracted by a low-head turbine exhausting to a sheltered area not subject to wave impact (Menlum, 1991). The height differential between the basin and mean sea level can be maintained at about the 3-m range along a coast with a robust wave climate. Although this is a smaller head than is typically found to be economically attractive for tide-energy extraction, the power generation is continuous so long as the wave intensity persists. Wave persistence is typically very much longer than the fraction of a tide cycle where tidal power can be extracted. Further, the head is constant over this period so that turbines can be optimized for a fixed flow rate.

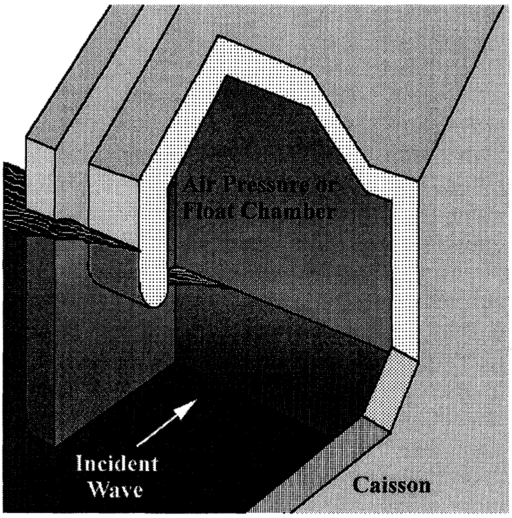

FIG. 11. A cross section through a generic pneumatic-caisson type of wave-energy system. The rising and falling water level inside the caisson chamber produces oscillating air pressure, or may be used to lift a float or other mechanical device.

The caisson-type devices have been the subject of significant theoretical and laboratory research (Mei, 1976; Falnes, 1980). In the most common configuration, a capture chamber is constructed so that the incident wave energy enters the front of the chamber but is trapped by side and rear walls. This results in a significant amplification as the wave is reflected. The chamber shape and the entrance configuration have been tested in a number of configurations in an attempt to achieve broad-band resonance with the incident waves. If the top of the chamber is closed, air trapped above the rising wave will be strongly compressed, while a decompression will occur on the falling portion of the cycle. Typically, an air turbine is arranged above the chamber and is driven by the flow of air out of and into the chamber in response to the water elevation changes. Valves have been used to rectify the air flow (Masuda, 1987) and self-rectifying turbines developed that turn in a single direction regardless of the direction of flow (Gato and Falcao, 1988). Direct mechanical conversion has been attempted in a modification to the basic caisson approach that utilizes oscillating flaps driven directly by the water motions. The pneumatic caisson provides a distinct advantage in that no moving parts are in direct contact with seawater. Because severe storm waves typically are an order of magnitude more energetic than the commonly occurring waves, material fatigue, stress corrosion, and outright structural failure are constant threats to machinery directly impacted by waves. In the case of the Kvaener pneumatic caisson discussed below, the caisson was constructed into a nearly vertical rock face, and the entire system was lost because of fracturing of the rock foundation during a severe storm.

In general, the first two types are shore-based or are constructed in relatively shallow water close to shore. The third type of system, of which there are almost unlimited variations, can be employed offshore in moderately deep water. In these applications, the hull of the device floats at the surface, tethered such that it remains in location, and responds to the wave forcing by some combination of heaving (vertical motion), swaying (horizontal motion), or pitching (tilting). The motion of the hull is constrained by attachment to an anchoring system, by inertia, or by making the hull large compared to the dimensions of the waves, thus summing the various forces along the wave to produce a nearly stationary platform. One of the simpler forms (Fig. 12) uses the vertical excursions of the hull to activate a pump. The pumps have included conventional piston types as well as flexible hoses that contract when put in tension and decrease in internal volume. The hose provides the anchoring cable in this instance. Researchers in the United Kingdom, Denmark, and Sweden have been particularly active in research on this general class of energy-extraction devices. A survey of these developments is contained in Ross (1995). Although these devices can, in principle, be shown to be more efficient in trapping incident wave energy than the previous two classes, they suffer from inherently high installation and maintenance costs because of having moving parts in contact with seawater and being subject to millions of fatigue cycles per year. The reliability of a submarine power cable subjected to repeated flexure cycling has never been established over the anticipated service lifetimes. Any of the wave-power systems can be vulnerable to total destruction by waves exceeding those that it was designed to resist. However, the nearshore systems can be somewhat less sensitive to this catastrophic failure mode if sited in areas where shallow water offshore limits the height of the wave that can impact it.

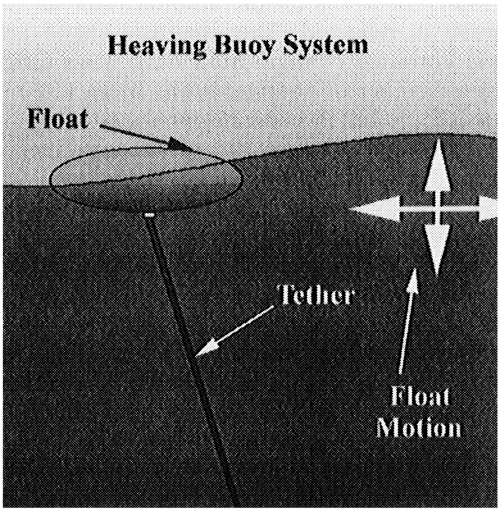

FIG. 12. A schematic arrangement of a heaving-buoy system, in which the vertical motions of a moored float are utilized to extract power from waves. The motion of the mooring line can drive a pump or other mechanical system.

2.2.2 Wave-Power Installations Only a few installations intended to supply energy from waves for public use have ever been attempted, in spite of the large research and development programs undertaken by several nations. Two of these were constructed along the Atlantic coast of Norway during the late 1980s. The tapered-channel system at Toftestallen, with a capacity of 0.35 MW, remains the only commercial wave-power system in operation in the world at this time. The Kvaerner pneumatic caisson, whose failure was discussed previously, was constructed nearby with a capacity of 0.5 MW, although its capacity factor was only 15%. This compares to 25% for the Toftestallen tapered-channel installation (Hagerman, 1992). Japan has led in the total number of installations, including a floating-caisson type system in a shiplike hull (Carmichael and Falnes, 1992), but all of them have been developmental in nature and none has survived. A 150-kW pilot plant of the caisson type has been installed and operated off the southwest coast of India. The British attempted a large installation of a caisson system with 2-MW capacity, called the OSPREY. It sank while being towed to its installation site in 1995.

2.2.3 Environmental Effects All of the installations to this time have been at such a small scale that they have had no discernible effect on the environment. If large-scale facilities were to be constructed in the future, and these power plants blocked a substantial fraction of the incident wave energy, some significant modification of the local environment is possible. For example, if, in an extreme case, the wave-energy extraction systems were sited close to shore and were to function as a breakwater and provide relatively calm water in its lee, then the transport of sediment along the shoreline would be reduced. As is the case with conventional offshore breakwaters, sand could be deposited in the lee forming a spit. In the process, this sand would be denied to downcoast beaches and substantial erosion could occur. If the wave-power station were sufficiently far away from the shoreline, diffraction would tend to eliminate the wave shadow. At an offshore distance several times the shore-parallel length of the wave-power system, the shadowing effect would be significantly reduced by diffraction. Because considerable freedom may exist in the selection of the location for an installation within a broad area of suitable resource, it could be feasible to avoid environmentally significant bottom resources such as reefs or rock outcrops that support unique biotic assemblages. It is generally assumed that wave-energy systems will be among the most environmentally benign power sources.

2.2.4 Economics Hagerman (1992), in an optimistic assessment of the economics of wave-energy utilization, comes to the conclusion that wave power is more economical than any other form of ocean energy, except the largest tidal plants, such as those presently under study in the United Kingdom. This finding includes systems employing the temperature difference between warm surface water and cold deep water (Johnson, 1992), as well as tidal energy. He also finds that tapered-channel type wave-power systems are less costly than conventional coal-fired plants when the plant size is less than 30 MW. Unless a significant economic penalty is assessed in the future for atmospheric pollution from fossil fuel plants, it would appear that wave power will be relegated to small, niche markets, such as isolated islands, where fossil fuel is relatively more expensive to obtain. This study, which uses a standardized cost estimation methodology in an attempt to obtain a fair comparison of a wide variety of systems, also concludes that the tapered-channel systems will produce energy at roughly one-third to one-half the cost of the caisson or floating types. Because of the extreme variability in the resource, wave power will necessarily have a lower economic value than power that can be generated to match demand. Some improvement in this condition can be made through power-storage devices, such as conventional pumped storage and the storage of compressed air. The latter mechanism has particular attraction in the ocean where air can be stored in simple open-bottom containers, using hydrostatic pressure of the surrounding water to maintain the necessary pressure. Other solutions to the variable resource include the manufacture of storable products, such as hydrogen (Isaacs and Seymour, 1973) and the direct desalinization of seawater by reverse osmosis (Hagerman, 1992). The reverse-osmosis application is well adapted to heaving systems that directly drive pumps.

GLOSSARY

Amphidromic Point: In a system of rotary tidal waves, the apparent axis of rotation, usually with negligible tide elevations.

Capacity Factor: Similar to **Load Factor** as defined below. It is usually applied to wave-power systems, which must be designed with the capacity to withstand an occasional very large wave. It refers to the fraction of the power rate associated with this ultimate capacity that is achieved on average.

Characteristic Period: In a wave field with a broad frequency spectrum, a single period (reciprocal of frequency) chosen to represent the central or significant frequency, such as the period at which the energy in the spectrum is at a maximum.

Dispersion Relationship: Surface gravity waves in water travel at a speed inversely proportional to frequency such that broad-band wave fields tend to disperse into narrow-band fields given enough time and distance.

Fetch: The distance over water, measured in the direction of the wind, that waves are subject to energy input from a wind field.

Load Factor: In a power-generation plant, the fraction of the maximum power output rate achieved on average.

Marginal Sea: A large body of water, usually semienclosed, with free communication to an ocean basin.

Works Cited

Bretschneider, C. L. (1952), *Trans. Am. Geophys. Union* **33,** 381–389.

Carmichael, A. D., Falnes, J. (1992), in: R. J. Seymour (Ed.), *Ocean Energy Recovery: The State of the Art,* New York: American Society of Civil Engineers, pp. 182–212.

Charnock, H. (1955), *Quart. J. Roy. Meteor. Soc.* **81,** 639–640.

Daborn, G. R. (Ed.) (1976), *Proceedings of a Workshop on Environmental Implications of Fundy Tidal Power,* Wolfville, Nova Scotia: Acadia University Institute.

Darwin, G. H. (1911), *The Tides and Kindred Phenomena in the Solar System,* London: John Murray.

Donelan, M. A., Dobson, F. W., Smith, S. D., Anderson, R. J. (1993), *J. Phys. Oceanog.* **23,** 2143–2149.

Falnes, J. (1980), *Appl. Ocean Res.* **2** (2), 75–80.

Gato, L. M. C., Falcao, A. F. O. (1988), *Int. J. Mech. Sci.* **31,** 383–395.

Greenberg, D. A. (1984), *The Effects of Tidal Power Development on the Physical Oceanography of the Bay of Fundy and the Gulf of Maine,* Canadian Technical Report of Fisheries and Aquatic Science.

Hagerman, G. (1992), in: R. J. Seymour (Ed.), *Ocean Energy Recovery: The State of the Art,* New York: American Society of Civil Engineers, pp. 213–254.

Hasselmann, K. *et al.* (1973), *Dtsch. Hydrogr. Z. Erganzungsheft Reihe A,* 12.

Hendershott, M. C. (1981), in: B. A. Warren, C. Wunch (Eds.), *Evolution of Physical Oceanography,* Cambridge, MA: MIT Press, pp. 292–341.

Ippen, A. T., Harlemann, D. R. F. (1966), in: A. T. Ippen (Ed.), *Estuary and Coastline Hydrodynamics,* Engineering Societies' Monographs, New York: McGraw-Hill, Chap. 10.

Isaacs, J. D., Schmitt, W. R. (1980), *Science* **207,** 265–273.

Isaacs, J. D., Seymour, R. J. (1973), *Int. J. Environ. Studies* **4,** 201–205.

Janssen, P. A. E. M. (1989), *J. Phys. Oceanogr.* **19,** 745–754.

Jeffreys, H. (1925), *Proc. R. Soc. London A* **107,** 189–206.

Johnson, F. A. (1992), in: R. J. Seymour (Ed.), *Ocean Energy Recovery: The State of the Art,* New York: American Society of Civil Engineers, pp. 70–108.

Lamb, H. (1932), *Hydrodynamics,* New York: Dover.

Large, W. G., Pond, S. (1981), *J. Phys. Oceanogr.* **11,** 324–336.

Masuda, Y. (1987), in: M. E. McCormick, Y. C. Kim (Eds.), *Utilization of Ocean Waves,* New York: American Society of Civil Engineers, pp. 1–33.

McCormick, M. E. (1981), *Ocean Wave Energy Conversion,* New York: Wiley.

Mei, C. C. (1976), *J. Ship Res.* **20** (2), 63–66.

Menlum, E. (1991), *Oceans '91 Proceedings,* New York: Institute of Electrical and Electronic Engineers, pp. 535–538.

Pierson, W. J., Neumann, G., James, R. W. (1995), U.S. Navy Hydrographic Office Publication No. 603, Washington, D.C., U.S. GPO.

Rieder, K. F., Smith, J. A., Weller, R. A. (1994), *J. Geophys. Res.* **91,** 22589–22596.

Ross, D. (1995), *Power from the Waves,* Oxford, U.K.: Oxford Univ. Press.

Shaw, T. L. (1992), in: R. J. Seymour (Ed.), *Ocean Energy Recovery: The State of the Art,* New York: American Society of Civil Engineers, pp. 34–41.

Snodgrass, F. E., Groves, G. W., Hasselmann, K., Miller, G. R., Munk, W. H., Powers, W. H. (1996), *Philos. Trans. R. Soc. London A* **259,** 431–497.

Sverdrup, H. U., Munk, W. H. (1947), U.S. Navy Hydrographic Office Publication No. 601, Washington, D.C., U.S. GPO.

Warnock, J. G., Clark, R. H. (1992), in: R. J. Seymour (Ed.), *Ocean Energy Recovery: The State of the Art,* New York: American Society of Civil Engineers, pp. 4–33.

Wu, J. (1982), *J. Geophys. Res.* **87,** 9704–9706.

Further Reading

Hendershott, M. C. (1981), in: B. A. Warren, C. Wunch (Eds.), *Evolution of Physical Oceanography,* Cambridge, MA: MIT Press, pp. 292–341.

Kinsman, B. (1965), *Wind Waves,* Englewood Cliff, NJ: Prentice-Hall.

Ross, D. (1995), *Power from the Waves,* Oxford, U.K.: Oxford Univ. Press.

Seymour, R. J. (Ed.) (1992), *Ocean Energy Recovery: The State of the Art,* New York: American Society of Civil Engineers.

Wiegel, R. L. (1964), *Oceanographical Engineering,* Englewood Cliff, NJ: Prentice-Hall.

TIME AND FREQUENCY, MEASUREMENT OF

MASA-KATSU FUJIMOTO, *Space-time Astronomy Section, National Astronomical Observatory, Mitaka, Tokyo, Japan*

	Introduction	467
1.	**Time Scales**	468
1.1	Universal Time	468
1.2	Atomic Time	469
1.2.1	International Atomic Time	469
1.2.2	Coordinated Universal Time	469
1.3	Dynamical Time Scales	469
2.	**Frequency Generators and Clocks**	470
2.1	Quartz-Crystal Oscillators	470
2.2	Atomic Frequency Standards	470
2.2.1	Cesium-Beam Frequency Standards	470
2.2.2	Other Frequency Generators	471
2.3	Frequency Stability of Generators	472
2.4	Laser: Optical and Near-Infrared Frequency Generators	472
3.	**Measurement of Frequency and Time**	473
3.1	Measurements by Means of Electronic Counters	473
3.2	Phase Measurements	474
3.3	Frequency Chains	474
3.4	Optical-Frequency Synthesis	476
4.	**Remote Comparisons**	477
4.1	Portable Clocks	478
4.2	Radio Time Signals	478
4.3	GPS	478
4.4	Computer Clock Synchronizations by Network Time Protocol	479
5.	**Applications of Precise Time and Frequency Measurements**	480
5.1	Experimental Relativity	480
5.1.1	Test of Local Position Invariance	480
5.1.2	Clock Synchronizations	480
5.2	Pulsar Timing Observations	481
5.3	Metrological Standards	481
5.3.1	Standard of Length	481
5.3.2	Voltage Standard	482
	Works Cited	482
	Further Reading	483

INTRODUCTION

The measurement of time and frequency is the most precise measurement we can make at present. The accuracy and stability of standard frequency generators and clocks have been improved by many orders of magnitude since the operation of the first cesium-beam frequency standard in the mid-1950s. Rapid progress has been made in the measurement technologies as well. The frequency range as well as the precision of the measurement are still growing, a result mainly of the progress in optics including nonlinear optics materials, tunable semiconductor lasers, stable solid-state lasers, high-quality mirrors, laser cooling of atoms, and so on. Such progress has expanded the applications of precise time and frequency measurement. One example is the standard of length, which is now derived from the standard of time. The standard voltage is also defined in terms of frequency, the inverse of time.

Time is a continuous dimension in which events occur. We can allot a set of real numbers to it in various ways, but in due order. There is no absolute rule for the allotment. Periodic phenomena such as the rotation of the Earth, the orbital motion of the Earth, oscillation of a pendulum, vibrational oscillation of a quartz crystal, or oscillation of radiation from an atom have been used to define the unit of duration. By fixing the unit

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

of duration as well as the epoch of origin (zero), the allotment is automatically done. Thus, each event has a corresponding real number, which is the reading of time at the epoch of the event. The difference between two numbers is the reading of time interval between the corresponding two events. The frequency of a periodic phenomenon is defined to be the inverse of the time interval at which events repeat.

Graduation of time based on the periodicity of a physical event has been confronted with difficulties, because general relativity predicts that the period of the same physical event is invariant only in local inertial frames, but can change from place to place, and because the amount of variation may be large enough to be recognized by accurate clocks. The proper time (and the proper frequency) associated with a clock can maintain the periodicity, but it serves only along the world line of the clock. The definition of time scales and that of clock synchronizations must be revised within the framework of relativity.

Remote comparisons of time play an important role in the international coordination of time scales. The Global Positioning System (GPS) has been conveniently used for this purpose. Intercontinental clocks are compared every day with an accuracy of 10 ns, which requires relativistic effects to be taken into account in the data processing. In this way, the realization of international time scales as well as the international comparisons of clocks are performed usually on the basis of general relativity. On the other hand, general relativity has been tested by accurate clocks and precise comparisons.

1. TIME SCALES

When any event is described by its position and its time, three dimensional space coordinates and one dimensional time coordinate are used as a reference system of coordinates. The time coordinate differs from the other space ones by its unidirectional nature: time is passing from past to future irreversibly.

To establish a system of time coordinate, one must define two quantities: the unit of duration and the epoch, or the origin of the time. An example of a naively defined one is the time used in Japan in the Edo era, in which the unit of duration in day time was the period from sunrise to sunset and the one in night time was the period from sunset to the next sunrise. Each period was equally divided into six parts to give a unit time interval. This time scale was not uniform because the interval between sunrise and sunset changes in a period of year and as a function of latitude of the location. The origin of time was also different from place to place because the times of sunrise and sunset depend on the location. However, such nonuniformity and differences of time did not matter at that era, since human activities were limited to a small area and the speed of transportation was slow enough to neglect the time difference.

The Industrial Revolution changed the situation. Human activities were expanded to worldwide scales by the technological developments such as railroads and telecommunications. The different time systems of different countries caused much trouble and confusion. The unification of longitudes and times was desired not only for scientific purposes but also for the purposes of navigation, commerce, and international communications.

1.1 Universal Time

The International Meridian Conference held at Washington, D.C., U.S.A., in 1884 decided on the adoption of the unique determination of longitude based on the Greenwich prime meridian. The adoption of Universal Time (UT) was also decided, which is the mean solar time of Greenwich (Greenwich Mean Time, GMT). Following the decision, many countries adopted the standard longitude to determine the standard time, which differs from UT by a (usually) integral number of hours. In UT and the standard times, the unit of duration is the mean solar day, which is the rotational period of the Earth with respect to the "fictitious mean Sun," a fiducial point on the celestial equator. Variations in the rotation of the Earth were discovered more than 50 years ago through comparisons with quartz-crystal clocks. UT is no longer a uniform time scale, but it still remains the basis of public time in the form of Coordinated Universal Time

(UTC). (Note: initialisms such as UTC are based on the French forms of the names.)

1.2 Atomic Time

Atomic time scales are integrated time scales that are produced by accumulating standard time intervals. The unit of duration is the International System (SI) second that is defined as follows: "The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom."

1.2.1 International Atomic Time International Atomic Time (TAI) is a coordinated time maintained by the International Bureau of Weights and Measures (BIPM) according to international conventions. TAI is realized in two steps. First, data from all available high-precision atomic clocks, of which are now more than two hundred distributed worldwide, are combined using an algorithm to form a free-running data-controlled time scale. It is compared in duration with primary frequency standards, the accuracy of each of which is carefully evaluated. Corrections are applied to maintain the unit of duration for TAI as close as possible to the definition of the SI second.

The increasing accuracy of clocks and advances in precise time comparisons require us to treat TAI as one of the four-dimensional space-time coordinates consistently with the framework of relativity. Apart from its error, each atomic clock gives an ideal atomic time at the location of the clock, which is called a proper time in terms of relativity. Different (ideal) atomic clocks located at different places show different time readings in general. For example, an atomic clock located on the top of a mountain (1000 m above sea level) has a clock rate higher by amount of 10^{-13} than other atomic clocks at rest at sea level. In order to establish TAI with an accuracy better than 10^{-15}, it is needed to know the location of each high-precision atomic clock with an error of height less than 10 m. The precise locations of clocks are also needed for precise clock comparisons using electromagnetic signals. Comparisons with a precision of 10 ns require the estimation of the travel path length of signals with an error less than 3 m.

In the refined definition of TAI, the unit of duration is one SI second realized as a proper time interval of an ideal atomic clock fixed on the geoid (the equipotential surface of the Earth's gravity potential that most nearly coincides with mean sea level), and TAI is a time coordinate of the geocentric coordinate system by which proper times of any fixed or mobile clocks near the geoid are related to TAI within the accuracy of TAI itself. Propagation of any electromagnetic signals in the vicinity of the Earth is also described in terms of TAI and the space coordinates of the system (see Sec. 5.1.2).

1.2.2 Coordinated Universal Time UTC is a time scale that is approximate to UT but is generated from TAI. UTC differs from TAI only by an integer number of seconds. Therefore the duration of time is the same as that of TAI. UTC is maintained within 0.90 s of UT by introducing a 1-s step (leap second) when necessary, usually at the end of June or December. Standard Time and broadcast time of most countries are now based on the UTC.

1.3 Dynamical Time Scales

Every physical time scale is a dynamical time scale in some sense. It is usually based on the periodicity of a dynamical process. For example, UT is related to the rotational motion of the Earth, which has a period determined by the dynamics concerned. However, because of many dissipative processes that are not well modeled yet, changes of UT are not predicted by any dynamical model to the required accuracy. UT is not categorized into dynamical time scales for this reason.

The orbital motion of the Earth in the solar system is described to good accuracy by dynamical theories such as Newtonian mechanics and Einstein's theory of general relativity. When the variations in rotational motion of the Earth became significant, the orbital motion of the Earth was adopted in determining the unit of time. The observable configuration of the solar system is given as a function of the time argument in the ephemeris as a solution of the equations of motion in Newtonian mechanics. The observed configuration at some instant gives the corresponding time argument in the ephemeris, Ephemeris Time (ET), which is a dynamical time scale. The second was defined as the fraction 1/31 556 925.9747 of the tropical year for 0 January 1900 (i.e., 31 December 1899) at 12 h Ephemeris Time. This definition was adopted by the 11th General Conference on Weights and Measures (CGPM) in 1960 and was replaced in 1967, after only 7 years, by the new one based on the quantum transition of Cs-133 atoms.

The concept of dynamical time scales has an increasing importance, because of the increasing accuracy in atomic time scales. As mentioned in Sec. 1.2, the description of any physical event in four-dimensional space-time is required to correspond to the accuracy of the measurements. A dynamical time scale in the framework of general relativity is considered as a realization of coordinate time, which is one of the coordinates in space-time. Dynamical time scales used in astronomy and their relations to TAI have been very important in the field of astronomy and metrology (Guinot, 1989; Seidelmann and Fukushima, 1992).

2. FREQUENCY GENERATORS AND CLOCKS

2.1 Quartz-Crystal Oscillators

A quartz crystal, a single crystal of SiO_2, is one of the best macroscopic resonators for stable frequency generators. It has very low internal friction losses and low temperature coefficients of elasticity and expansion. The piezoelectric effect of this material is used to excite mechanical vibration by putting an oscillating electric field on the crystal. An electric circuit coupled to the mechanical resonator forms a positive-feedback amplifier and generates an oscillating signal. Quartz-crystal oscillators are used as stable frequency generators in various fields of application such as satellite communications, digital telecommunications, wrist watches, TV and radio receivers, local oscillators used in atomic clocks, and so on. Recent progress in understanding and design of resonators as well as in understanding the influence of electric circuits has led to significant improvement in their operation. Reduced aging, better short-term stability, low thermal transients, and small environmental dependence have been obtained and have led to advantages in design of atomic frequency standards (Bessen, 1989).

A precise frequency control of the oscillator, commonly required in many applications, is attained by a voltage-controlled crystal oscillator (VCXO), in which a capacitance in the equivalent circuit of the oscillator is changed by an external voltage control.

2.2 Atomic Frequency Standards

Atomic frequency standards are based on the principles of quantum mechanics. The frequency of the electromagnetic radiation emitted or absorbed by an atom or molecule during a change of its quantum state between two discrete energy levels is related only to the energy difference between the two levels. The energy levels for unperturbed atoms or molecules are regarded as constants, and thus the frequency of the radiation has a unique value.

Atomic frequency standards are categorized by their mode of operation into two types, namely, active and passive. Figure 1 shows the principles of atomic frequency standards in each type. The active mode of operation uses maser or laser action with the stimulated emission of radiation. In the passive mode of operation, an ensemble of particles undergoing the desired quantum transition interacts with synthesized radiation. The transition occurs at maximum probability if the frequency of the radiation exactly coincides with the proper frequency of the transition. A VCXO is ordinarily used as a slave oscillator, and its frequency is controlled in a feedback system to maintain the coincidence.

2.2.1 Cesium-Beam Frequency Standards Cesium-beam frequency standards typically operate in a passive mode. Figure 2 shows the operation principle of the cesium-beam frequency standard. Since the current definition of the second is based on the hyperfine transition of the cesium atom, as mentioned before, cesium-beam frequency standards play an important role in generating atomic time scales as well as in realizing the unit of time, the second. Commercially available frequency standards are distributed worldwide and contribute significantly to TAI. Laboratory-made primary frequency standards are working in several time-keep-

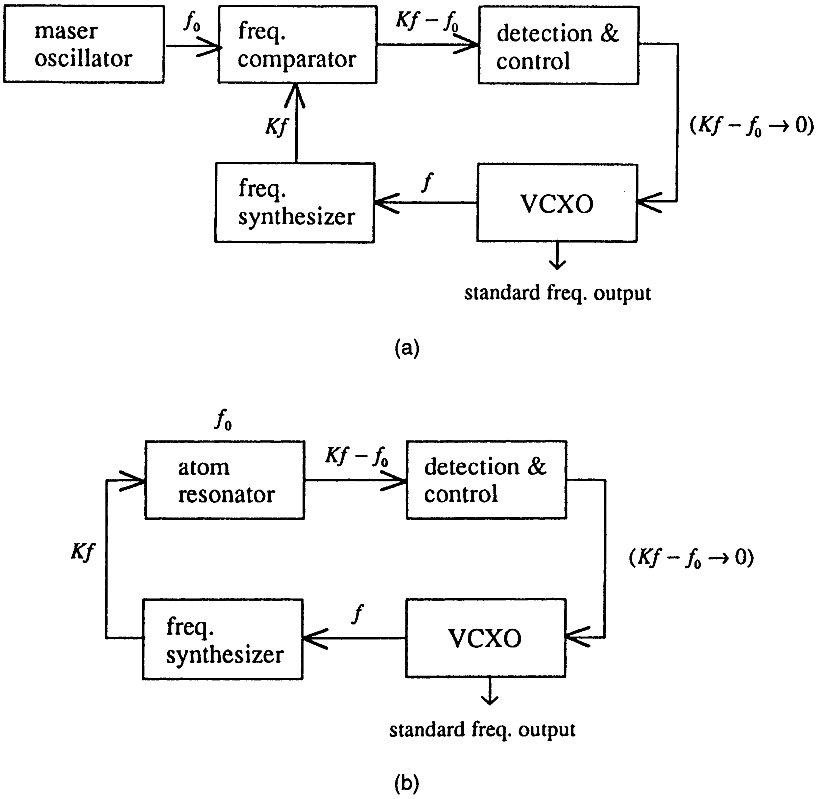
maser oscillator
f_0
freq. comparator
$Kf - f_0$
detection & control
Kf
$(Kf - f_0 \rightarrow 0)$
freq. synthesizer
f
VCXO
standard freq. output
(a)
f_0
atom resonator
$Kf - f_0$
detection & control
Kf
$(Kf - f_0 \rightarrow 0)$
freq. synthesizer
f
VCXO
standard freq. output
(b)

FIG. 1. Schematic diagram of atomic frequency standards: (a) active and (b) passive.

ing laboratories to realize the second. New primary cesium standards have been studied and developed intensively that use optical pumping of the desired energy level of the atom.

2.2.2 Other Frequency Generators

Atomic frequency standards based on the hyperfine transition in atomic hydrogen typically use the active maser mode of operation. Figure 3 shows the principle of the hydrogen-

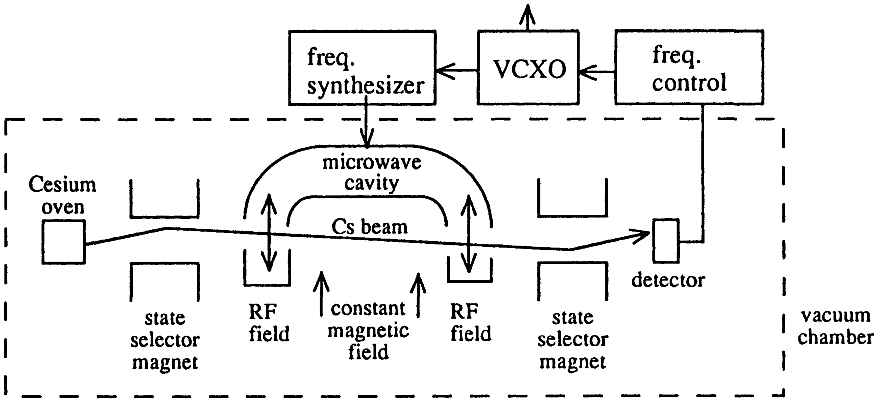
freq. synthesizer
VCXO
freq. control
Cesium oven
microwave cavity
Cs beam
detector
state selector magnet
RF field
constant magnetic field
RF field
state selector magnet
vacuum chamber

FIG. 2. Principle of the cesium-beam frequency standard.

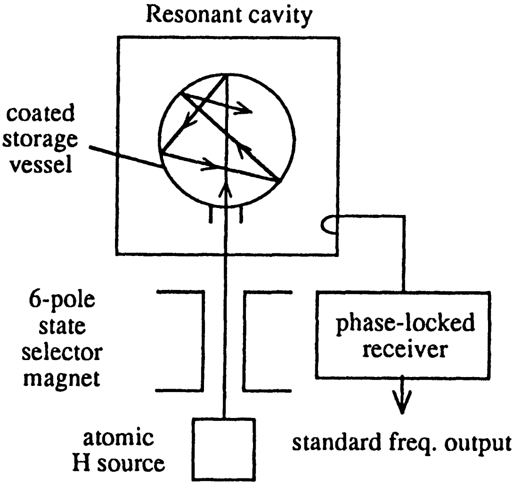

FIG. 3. Principle of the hydrogen-maser frequency standard.

maser standard. Hydrogen masers are very stable for averaging times from 1 to above 10^4 s. They are used as local oscillators for very long-baseline interferometry, as well as clocks and reference frequency generators.

Rubidium clocks are relatively low-cost clocks and have better stability than quartz-crystal oscillators for measurement times >100 s. They are usually subject to greater frequency drift than hydrogen masers or cesium clocks.

Several new-type atomic frequency generators such as Mg-beam standards, trapped Hg ion standards, and Cs fountain standards have been proposed and are under development. Among them, a first prototype of a cesium-atom fountain has been realized and has worked as a frequency standard to indicate a stability of the order of 2×10^{-15} for durations of a few days (Clairon *et al.*, 1995).

A superconducting cavity can be used as a frequency discriminator in stabilizing a generator. Such a generator is called a superconducting-cavity–stabilized oscillator (SCSO) and has one of the best stabilities for averaging times below several thousand seconds; however, it has no absolute frequency reference.

2.3 Frequency Stability of Generators

Frequency standards and clocks are subject to random as well as systematic variations of their frequency or rate. A nearly sinusoidal signal voltage generated by a frequency standard is written as

$$V(t) = [V_0 + e(t)]\sin[2\pi f_0 t + \phi(t)]. \tag{1}$$

Here V_0 is the nominal value of the amplitude and $e(t)$ is the amplitude fluctuation, which we assume negligible in most cases. f_0 is the nominal value of the frequency, and $\phi(t)$ is the phase fluctuation. The instantaneous frequency of the signal is given by

$$\begin{aligned} f(t) &= f_0 + \frac{1}{2\pi}\frac{d\phi}{dt} \\ &= f_0\{1 + dx(t)/dt\} = f_0\{1 + y(t)\}, \end{aligned} \tag{2}$$

where two useful quantities are introduced: $x(t)$ is the instantaneous "phase-time" error and $y(t)$ is the instantaneous normalized frequency error (Kartaschoff, 1978). The instantaneous frequency and thus $y(t)$ cannot be measured, but the average normalized frequency error is measured instead; it is given by

$$\langle y\rangle(t,\tau) = \frac{x(t+\tau) - x(t)}{\tau} = \frac{1}{\tau}\int_t^{t+\tau} y(t)dt. \tag{3}$$

As a statistical measure of the frequency stability, popularly and conveniently used is the Allan variance, which is defined by

$$\sigma^2(\tau) = \frac{1}{2N}\sum_{k=1}^{N}(y_k - y_{k-1})^2,$$

where y_k and y_{k-1} represent successive samples of $\langle y\rangle$. Figure 4 shows the frequency stabilities of typical frequency standards and clocks in terms of the Allan variance as a function of the averaging time interval τ (Allan, 1989).

2.4 Laser: Optical and Near-Infrared Frequency Generators

A variety of lasers have been developed since the first operation of a ruby laser in 1960. The frequency of laser emission is predominantly governed by the resonant frequency of its optical cavity combined with the active medium inside the cavity in which the laser action occurs. Careful design of the cavity may reduce the frequency noise. An

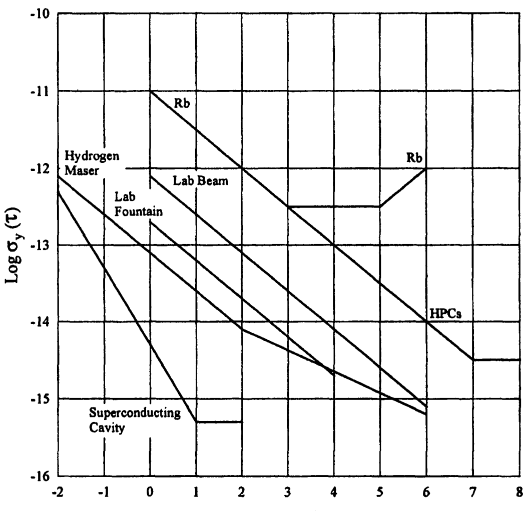

FIG. 4. Fractional frequency stability $\sigma_y(\tau)$ as a function of averaging time τ for some precision frequency references. Rb denotes a commercial rubidium gas-cell frequency standard. HPCs denotes a high-performance commercial cesium-beam frequency standard. Lab Beam denotes laboratory cesium-beam standards. Lab Fountain denotes laboratory cesium-fountain frequency standards. The curve labeled Hydrogen Maser represents active cavity-tuned hydrogen masers; that labeled Superconducting Cavity represents a superconducting microwave resonator.

example is a monolithic isolated single-mode end-pumping ring laser (MISER), which has a spectral linewidth less than 10 kHz under free-running conditions, because the cavity is monolithically integrated with the laser crystal. A laser frequency can be further stabilized to a high-finesse Fabry–Pérot cavity, and a linewidth of less than 1 Hz has been attained (Day *et al.*, 1992; Uehara and Ueda, 1993). Very stable Nd:YAG lasers of this type have been studied as powerful coherent light sources for a very long-baseline laser interferometer for the detection of gravitational waves (Giazotto, 1989; Fujimoto, 1992). Because the resonant frequency of a cavity is geometrically determined, the absolute frequency is not well determined. An accurate laser frequency is obtained by its stabilization to a certain spectral line of an atom or molecule. An example is a CH_4-stabilized He-Ne laser, where the inverted Lamb dip in the saturated absorption of CH_4 at 3.39 μm is used as the frequency reference because of its Doppler-free nature. Hyperfine-structure components in the absorption line of I_2 at 633 nm are also widely used to stabilize a He-Ne laser in the visible region. These stabilized He-Ne lasers are of importance in the realization of the meter (see Secs. 3.3 and 5.3.1).

3. MEASUREMENT OF FREQUENCY AND TIME

3.1 Measurements by Means of Electronic Counters

Electronic counters are widely used for frequency, period, and time-interval measurements. The basic block diagram of such an instrument is shown in Fig. 5. The input circuits process the wave form of input signals so as to make the events unambiguously countable. The gate starts the counting operation when an "open gate" signal is received from the control unit, and stops on arrival of a "close gate" signal. All timing operations are controlled by the time base. A stable quartz-crystal oscillator is usually used as a frequency source for the time base, from which the timing signals with a required time interval are generated by frequency dividers.

In frequency measurements, the number of events is counted in a time interval τ and is divided by τ to give the mean frequency averaged over the time interval τ.

In period measurements, event signals and timing signals interchange their roles: an event signal opens the gate to start counting the timing signals, and the succeeding event signal closes the gate to stop the counting (Fig. 6).

Time-interval measurements differ from period measurements only in the gate signals: the "open gate" and "close gate" signals are produced separately by the two different input signals (Fig. 7).

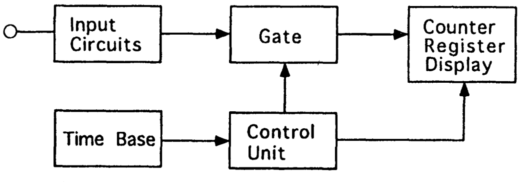

FIG. 5. Direct frequency measurement.

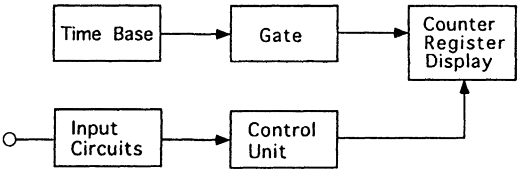
Time Base
Gate
Counter Register Display
Input Circuits
Control Unit

FIG. 6. Period measurement.

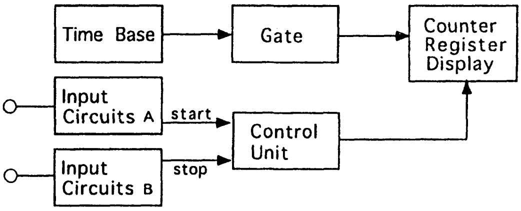
Time Base
Gate
Counter Register Display
Input Circuits A
start
Control Unit
Input Circuits B
stop

FIG. 7. Time-interval measurement.

The leading error in the measurements comes from the ±1 count ambiguity, which is usually larger than the other errors such as the trigger errors in the gate operation and the time-base accuracy. This ambiguity corresponds to a normalized resolution of $1/f\tau$ in a direct frequency measurement with averaging time τ. In a period measurement with averaging time τ, the normalized resolution amounts to $1/f_0\tau$, where $1/f_0$ is one period of the time base corresponding to the ambiguity. If the measured frequency is higher than the frequency of the time base, the frequency measurement gives better resolution than the period measurement, but in the case of a lower-frequency measurement, the period measurement is better than the frequency measurement. In order to increase the resolution of frequency measurement with limited averaging time, a beat frequency produced by mixing the measured signal with a reference signal is measured by its period. By this heterodyne method, the resolution is improved by a ratio of the signal frequency to the beat frequency.

3.2 Phase Measurements

If the signal frequency is almost equal to the reference one in the heterodyne method, the beat frequency is almost zero and is difficult to measure in a limited time. This is the case when two very stable oscillators such as atomic frequency standards are compared in frequency and time. Phase measurement is useful in this case. Phase measurement is in principle a time-interval measurement between the zero crossings of two signals. Figure 8 shows a block diagram of the so-called dual-mixer time-difference (DMTD) measurement, as an example of phase measurements. Two mixers and a transfer oscillator produce two beat signals from the two oscillators to be compared. The time difference between positive zero crossings of the two beat signals is measured by a time-interval counter. Resolution of the measurement is enhanced by the ratio of the signal frequency to the beat frequency in the same way as in the heterodyne method. By this method, very high time resolution, better than 1 ps, is obtained (Allan, 1975).

3.3 Frequency Chains

Even the best electronic counter can measure frequency only up to about 100 GHz, which is three or four orders of magnitude lower than optical frequencies. The precise measurement of near-infrared or optical frequencies has been carried out by using a skillful method, a frequency chain. Figure 9 shows a historical example of the chain in

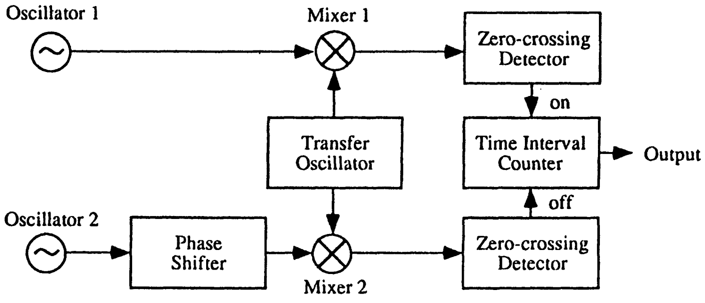
Oscillator 1
Mixer 1
Zero-crossing Detector
on
Transfer Oscillator
Time Interval Counter
Output
off
Oscillator 2
Phase Shifter
Zero-crossing Detector
Mixer 2

FIG. 8. Dual-mixer phase-time difference measurement.

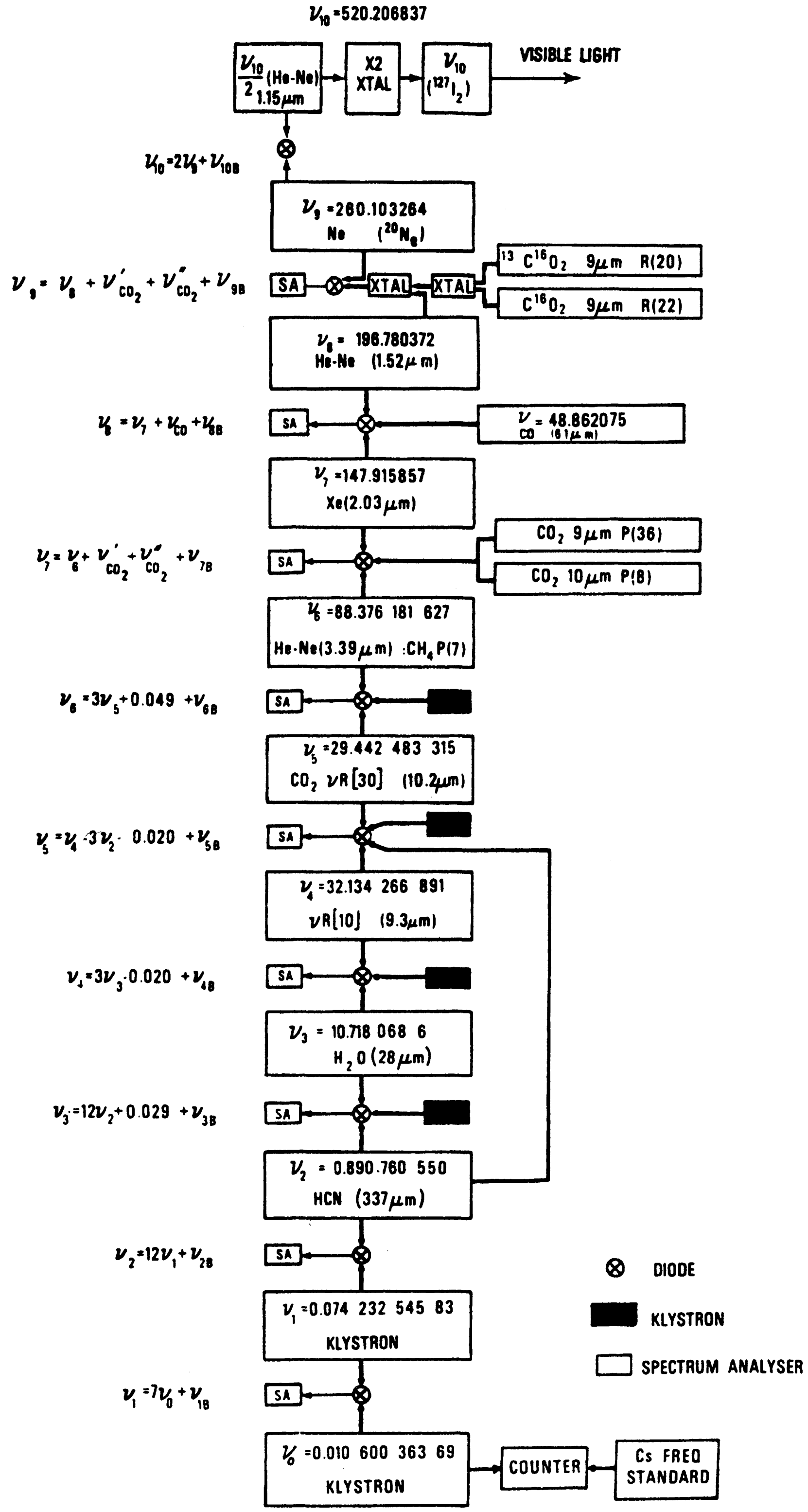
ν_{10}=520.206837
$\frac{\nu_{10}}{2}$ (He-Ne) 1.15μm
X2 XTAL
ν_{10} ($^{127}I_2$)
VISIBLE LIGHT
$\nu_{10}=2\nu_9+\nu_{10B}$
ν_9=260.103264 Ne ($^{20}N_e$)
$\nu_9=\nu_8+\nu'_{CO_2}+\nu''_{CO_2}+\nu_{9B}$
SA
XTAL
XTAL
$^{13}C^{16}O_2$ 9μm R(20)
$C^{16}O_2$ 9μm R(22)
ν_8= 196.780372 He-Ne (1.52μm)
$\nu_8=\nu_7+\nu_{CO}+\nu_{8B}$
SA
ν= 48.862075 CO (6.1μm)
ν_7=147.915857 Xe(2.03μm)
$\nu_7=\nu_6+\nu'_{CO_2}+\nu''_{CO_2}+\nu_{7B}$
SA
CO_2 9μm P(36)
CO_2 10μm P(8)
ν_6=88.376 181 627 He-Ne(3.39μm) :CH_4P(7)
$\nu_6=3\nu_5$+0.049 $+\nu_{6B}$
SA
ν_5=29.442 483 315 CO_2 νR[30] (10.2μm)
$\nu_5=\nu_4-3\nu_2$- 0.020 $+\nu_{5B}$
SA
ν_4=32.134 266 891 νR[10] (9.3μm)
$\nu_4=3\nu_3$-0.020 $+\nu_{4B}$
SA
ν_3 = 10.718 068 6 H_2O (28μm)
$\nu_3=12\nu_2$+0.029 $+\nu_{3B}$
SA
ν_2 = 0.890 760 550 HCN (337μm)
$\nu_2=12\nu_1+\nu_{2B}$
SA
ν_1=0.074 232 545 83 KLYSTRON
$\nu_1=7\nu_0+\nu_{1B}$
SA
ν_0=0.010 600 363 69 KLYSTRON
COUNTER
Cs FREQ STANDARD
⊗ DIODE
■ KLYSTRON
☐ SPECTRUM ANALYSER

FIG. 9. Stabilized laser frequency synthesis chain. All frequencies are in terahertz.

frequency measurement that produced the first measurement of a visible frequency (Baird *et al.*, 1979) and of a CH_4-stabilized He-Ne laser in the near-infrared region (Evenson *et al.*, 1973). The frequency ratio between the visible radiation at 520 THz and a klystron at 10.6 GHz is about 5×10^4. This huge gap was bridged by introducing several intermediate-frequency generators including a HCN laser, a H_2O laser, six CO_2 lasers, three He–Ne lasers, a Xe laser, and a CO laser. Each frequency difference between the neighboring generators up to the 1.52-μm He-Ne laser was measured by heterodyne method, in which a point-contact W-Ni diode played an important role by generating harmonics of each signal and by mixing them to produce a beat frequency within the measurable frequency range. The product of the CH_4 frequency by the interferometric measurement of the wavelength gave the absolute value of the light velocity, which was afterward adopted as a defining constant in the definition of the meter (see Sec. 5.3.1). The direct frequency measurement of the iodine-stabilized He-Ne laser at 473 THz (633 nm) was the basis for redefinition of the meter because its wavelength could be calculated very accurately from its frequency. Then it could be used to realize the meter with a significant increase in accuracy over the krypton method (Jennings *et al.*, 1983). Frequency chains consist of a series of oscillators at intermediate frequencies between optical and microwave regions of the electromagnetic spectrum. The intermediate frequency generators must have good frequency stability and sufficient power for harmonic generation. They can be phase-locked to each other in the chain by feedback using a number of harmonic mixers and interpolation oscillators. Consequently, each frequency chain has been designed and constructed for exclusive use at a particular frequency, and its operation has required elaborate work. A recent measurement of the Ca intercombination line, by use of a phase-coherent optical-frequency chain from the Cs atomic clock to the visible, gives the optical frequency of the line with an uncertainty less than 10^{-12} in normalized frequency. The chain is composed of ten laser oscillators, eight microwave oscillators, and a quartz oscillator (Schnatz *et al.*, 1996).

3.4 Optical-Frequency Synthesis

The optical or near-infrared frequencies that are connected to the microwave-frequency standards by chains are so coarsely distributed in frequency that any one of them is not always close enough to the frequency to be measured. There remains a large frequency difference that cannot be measured by means of electronic counters. Several approaches to measure large optical-frequency differences have been proposed. One method is a coherent bisection of optical-frequency intervals by using a nonlinear phase-locking scheme (Telle *et al.*, 1990; Wynands *et al.*, 1992). As shown in Fig. 10, the second-harmonic frequency of the third laser, $2f_3$, is compared with the sum frequency of the two lasers, $f_1 + f_2$, generated in a nonlinear crystal, and the frequency of the third laser is controlled by an optical phase-locked loop such that $2f_3 = f_1 + f_2$, or $f_3 = (f_1 + f_2)/2$. Then we obtain $f_3 - f_1 = (f_2 - f_1)/2$. Cascading of n bisection stages generates a difference frequency $f_n - f_1 = (f_2 - f_1)/2n$. After 14–16 stages, an optical-frequency difference is reduced to one in the microwave region. Optical-frequency division using an optical parametric oscillator (OPO) has also been proposed (Wong, 1990). Figure 11 shows how an OPO acts as a frequency divider. The input pump frequency f_p is divided into two output frequencies, a signal f_1 and an idler f_2. The output difference frequency $\delta = f_1 - f_2$ is phase-locked to a microwave reference frequency.

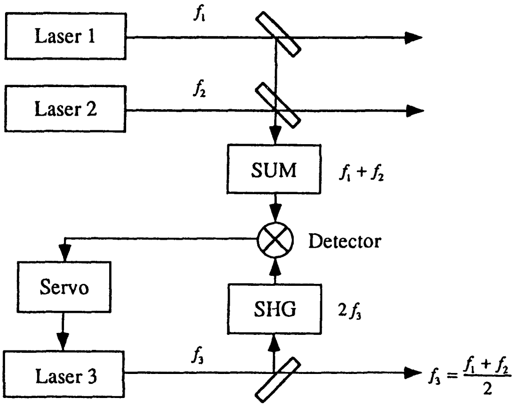

FIG. 10. Optical difference-frequency divider based on the nonlinear phase-locking scheme.

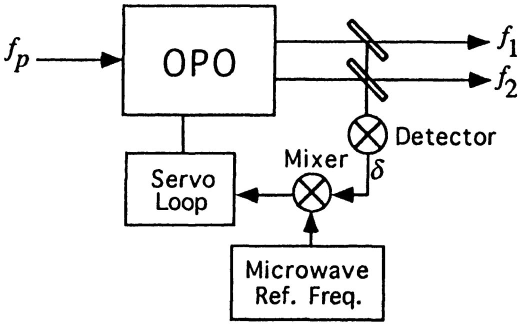

FIG. 11. Optical-frequency divider using an optical parametric oscillator.

Then we obtain $f_1 = (f_p + \delta)/2$ and $f_2 = (f_p - \delta)/2$. As shown in Fig. 12, the optical-frequency differences can be measured in the microwave-frequency region if we use the OPO dividers in series for frequency down-conversion and if we use a pair of the *n*th-stage OPO dividers for frequency comparison.

A different method of optical-frequency synthesis is to use an electro-optic modulator in an optical cavity to generate higher-order sidebands spanning the difference frequency of more than a few THz at the frequency interval of a few GHz (Kourogi *et al.*, 1992). The comb of frequencies can be used as a ruler that has graduations limited by wavelength dispersion in the cavity. The frequency difference between the nearest modulation sideband and the optical frequency to be measured is obtained by measuring the heterodyne frequency.

4. REMOTE COMPARISONS

International coordination of time-scale determination requires precise time and frequency comparisons among clocks or frequency standards in remote locations. The remote comparisons also provide a means of disseminating the time and the second to many sites around the world. Any remote comparison consists at least of two direct comparisons. As pointed out by Einstein (1905) in his discussion on simultaneity of physical events, clocks can be compared only at the same place and at the same instant. Therefore only a direct comparison has a sense. In order to make an indirect comparison between two distant places, an intermediate clock is required to move from one site to another and to be compared directly with each clock in each site. Not only a real operating clock but also time signals are used as the intermediary in remote comparisons. As shown in Fig. 13, an essential part of re-

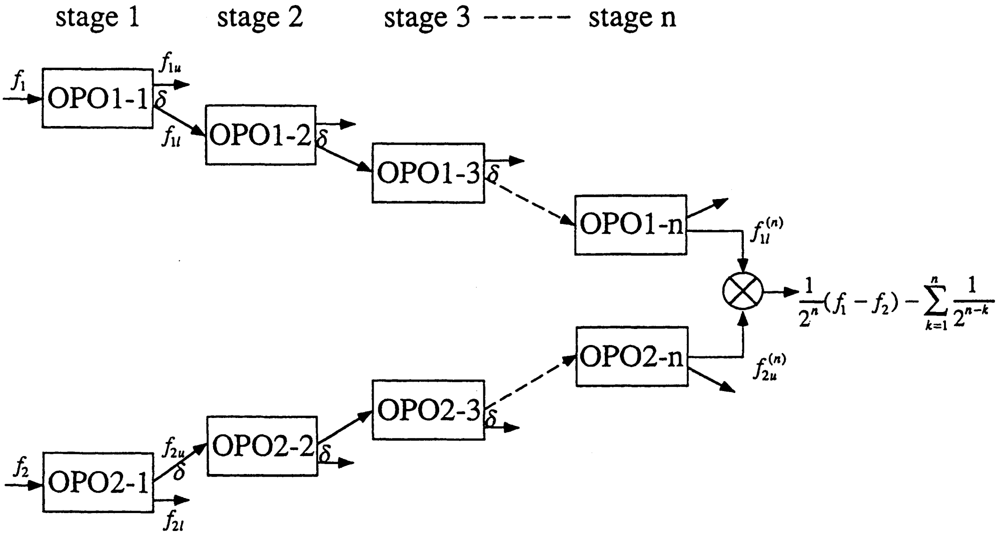

FIG. 12. Serial configuration of *n* OPOs for frequency division (only upper or lower of the two chains), and parallel configuration of *n* stages of two OPOs for frequency comparison.

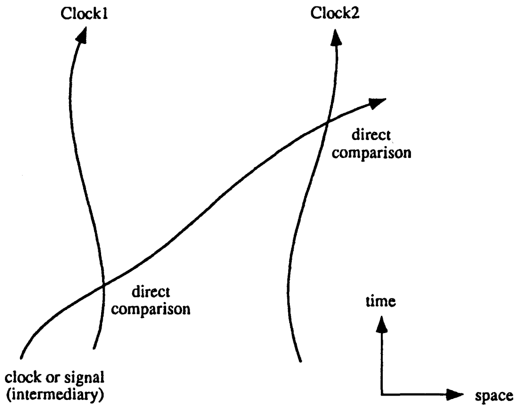

FIG. 13. World-line scheme describing remote comparison of clocks.

mote comparisons is described in terms of relativity by three world lines, intersections of which express the events of direct comparison.

4.1 Portable Clocks

A naive remote comparison of clocks is made by use of a portable clock. A portable clock keeps in operation during a trip. The clock is adjusted and kept as close to the frequency of the reference clock at one site as possible and then carried to another site and compared there. Usually a round trip is made to compare the clock again with the reference clock. By this procedure, the clock error accumulated during the trip is evaluated, which serves as a measure of confidence about the errors in this comparison. For accurate comparisons, relativistic effects must be taken into account (Hafele and Keating, 1972): the effect of gravitational field on the clock rate difference (about 1 part in 10^{13} per kilometer change in altitude), the effect of transportation velocity of the clock relative to ground-based clocks, and so on. The clock rate changes by about 1 part in 10^{13} when it is transported by airplane with speed of 700 km/h. Time comparisons by portable atomic clocks were the most reliable and accurate method of international and domestic clock comparisons before the common-view method of GPS time comparison was widely adopted in about 1985.

4.2 Radio Time Signals

Time comparisons by the use of time signals require the knowledge of the propagation delay. The accuracy of the comparison is often limited by the uncertainties in the propagation delay. Two-way methods, in which the time signals are exchanged simultaneously back and forth along the same path, can eliminate the common delay in the data processing. These methods, using artificial satellites, can give the accuracy of time comparisons at the nanosecond level. However, in the two-way methods complex equipment is required in order to transmit and receive the time signals. One-way methods are more popularly used. Time signals superimposed on transmitted radio waves are widely used as simple and cost-effective means of distributing time. Many countries have standard frequency transmissions with a time code. These high-frequency time signals are subject to delay variations caused by ionospheric disturbances and can provide an accuracy of only 1 ms. One mode of operation is to receive the same signal at two or more sites. By exchanging the data, the difference of readings of the clocks is obtained if the differential delay of propagation is known. The signals of TV, LORAN-C, GPS, and quasars [measured with very long-baseline interferometry (VLBI)] are used for accurate time comparisons by this method. Among them, the GPS has popularized rapidly and has a sufficient accuracy for most users. Therefore only the GPS time comparison is further described in the next section.

4.3 GPS

Global Positioning System is a satellite positioning system developed by the Department of Defense of the United States, which includes 24 satellites on 12-sidereal-hour orbits. The positioning is based on range measurements. Each satellite emits a pseudorandom noise-coded signal, the timing of which is controlled by the onboard atomic clock, the satellite clock. The received signal is correlated with the same code generated by a GPS receiver, the timing of which is controlled by the local clock. The tracking of the satellite signal gives the time difference between the time of the signal reception, registered by the local clock, and the time of emission counted by the satellite clock. This difference contains the actual propagation time interval from the satellite to the receiver antenna, the time offset between the satellite and the local clocks, and the receiver delay including instrumental and cable delays. The positioning requires at least four range measurements for different satellites. Four unknown parameters, three space coordinates of the antenna position and one time offset of the local clock, can be solved for from the four relations given by the four measurements. The other parameters including the relations are position coordinates of each satellite, propagation delays due to the refraction, and time difference between each satellite clock and a common time reference in the system called GPS time. All these are computed from the ephemeris, model parameters for the refraction, and extrapolated values of time offset for each satellite clock, which are all transmitted by the satellites. If the position of the antenna is already determined with good precision, a single measurement of a satellite gives the time offset between the local clock and GPS time. Precise time comparisons between two remote clocks are made by exchanging the value of the offset (local clock minus GPS time) with each other and by eliminating GPS time. When the tracking of the same satellite is carried out simultaneously at both sites, the error in estimating GPS time from the satellite clock cancels completely, and the error in the satellite position also cancels to some extent. This common-view method of GPS time comparison is widely used for the international coordination of time scales (Allan *et al.*, 1985). Accuracy in time comparison by this method reaches about 10 ns on daily averages. Apart from the precise time comparisons, GPS receivers have been popularized very quickly mainly for the application of automobile and small-boat navigation. Inexpensive and downsized receivers are commercially available today, and some of them generate a second pulse synchronized to the UTC that is useful to obtain the time as accurate as a few microseconds at any location.

4.4 Computer Clock Synchronizations by Network Time Protocol

The network time protocol (NTP) is a method to synchronize clocks of UNIX-based computers via computer networks (Mills, 1992). NTP servers that supply the time information at the request of a client are ranked into many strata. The server in stratum 1 has a clock that is synchronized to UTC by means of direct connection with an atomic clock or reception of the GPS time signals. When packets are exchanged between the server and the client, each packet is time-marked by the transmitting computer clock, and its reception time is registered by the receiving computer clock. If the propagation delays in the network are equal in the two directions, the clock difference between the server and the client and the propagation delay between them are estimated from the four readings of the clocks (Fig. 14). The accuracy of the clock synchronization by NTP is less than 1 ms, mainly limited by the fluctuations in propagation delay, but is attainable accurate to the tens of microseconds level.

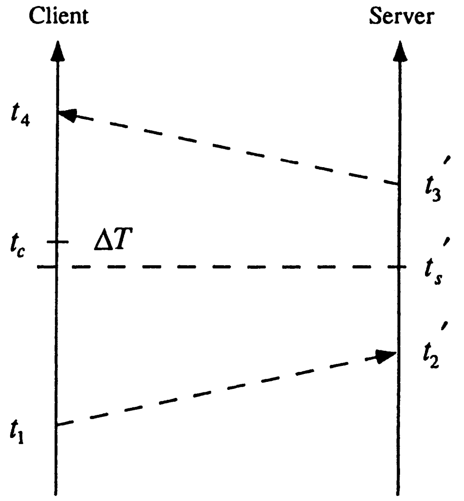

$$\Delta T = t_c - t_s' = \frac{t_1 + t_4}{2} - \frac{t_2' + t_3'}{2}$$

FIG. 14. Schematic diagram of signals used in clock synchronization by network time protocol.

5. APPLICATIONS OF PRECISE TIME AND FREQUENCY MEASUREMENTS

5.1 Experimental Relativity

5.1.1 Test of Local Position Invariance

Local position invariance (LPI) is a part of Einstein's principle of equivalence, which plays an essential role in constructing the relativistic theory of gravitation. It asserts that the outcomes of experiments made in local free-falling frames do not vary with time and position. Gravitational redshift experiments test the existence of spatial dependence. If LPI is valid, a gravitational redshift is universal, independent of the nature of the clocks being measured, and is given for weak gravitational fields by $df/f = dU/c^2$, where f is the nominal clock frequency, U is the Newtonian gravitational potential, and c is the speed of light (Will, 1981). A hydrogen maser, with frequency stability better than 10^{-14} over 100-s averaging time, was flown on a rocket to an altitude of about 10 000 km, and its frequency was compared to a similar clock on the ground (Vessot and Levine, 1979). In this experiment, the gravitational redshift on the clock was verified with 0.02% uncertainty. It is the most precise experimental result on this effect until now. A null gravitational redshift experiment that compares two different types of clocks in the same laboratory has been performed (Turneaure *et al.*, 1983). The rates of two H maser clocks were compared with those of three superconducting-cavity–stabilized oscillator clocks over a 10-d period. During this period the solar gravitational potential at the laboratory changed sinusoidally with a 24-h period by 3×10^{-13} because of the Earth rotation, and also changed approximately linearly at 3×10^{-12}/d because of the Earth's orbital motion. An upper limit on the relative frequency variation of 1.7% of the external potential was set. It was a direct test of LPI. If every type of clock varied with position in the same manner, there would be no operational way to detect the violation of LPI, since clock rates are always measured relative to a chosen standard clock.

Temporal dependence of the outcomes of local experiments is the other key test of LPI. The constancy of the fundamental nongravitational constants over cosmological time scales was studied in a wide range of observations such as comparisons of spectral emission lines in distant galaxies and quasars (Wolfe *et al.*, 1976; Pagel, 1977), measurements of isotopic abundances of elements in the solar system (Shlyakhter, 1976), laboratory comparisons of atomic clocks (Turneaure and Stein, 1976), and so on.

It is not an obvious fact that the time argument appearing in the dynamical equations of motion coincides with the one used in quantum mechanics. The question remains to be verified by all possible experiments whether an atomic time scale, which is an accumulation of the SI second defined by a quantum mechanical transition, is exactly equal to a proper time that is a concept of space-time physics, such as special and general relativity. As discussed in Sect. 1.2, TAI is a realization of a coordinate time in a geocentric system of coordinates, which is established on the assumption that Einstein's principle of equivalence is valid. Long-term observations of dynamical behavior in the solar system or in extraterrestrial stellar systems such as a spinning neutron star (pulsar) system, by using TAI as a time reference and by comparing with the theoretically expected configurations, can provide an opportunity to check the validity of the assumption.

5.1.2 Clock Synchronizations

In relativity, the notion of simultaneity is not given *a priori* so that a conventional choice of a definition has to be made. A widely used definition is that of coordinate simultaneity and corresponding coordinate synchronization (Allan and Ashby, 1986; Klioner, 1992). Two events are said to be coordinate simultaneous when they have the same value of time coordinate in some specified reference system of coordinates, whatever their space coordinates are. The synchronization of two clocks by this definition is entirely dependent on the chosen reference system.

Taking the GPS as an example, it adopts four-dimensional space-time coordinates to describe the motion of the satellites, behavior of the satellite clocks, and propagation of the radio signals. The time coordinate is GPS time, and the space coordinates are those of the World Geodetic System (WGS), which is a geocentric coordinate system with uniform rotation. Ideal atomic clocks operating on satellites run at different rates with respect to GPS time, because they move with different relative velocities and are submitted to different gravitational fields. The mean rate difference between the ideal clock on the satellite and that on the rotating geoid is 4.43×10^{-10}. Taking this into account, the real atomic clocks carried by the satellites are all tuned to a frequency shifted lower by 4.45×10^{-10}, before being launched. This makes the satellite clocks easy to be synchronized with GPS time. Each satellite moves in a nearly circular orbit with an eccentricity less than 0.01. But because of the small eccentricity, the reading of the satellite clock differs from GPS time. The difference varies sinusoidally with a 12-sidereal-hour period and with an amplitude of 25 ns at most. Propagation delay of signals from the satellite to the antenna of the receiving site is calculated from the space coordinates of satellite and those of the antenna. The calculation includes an effect on light propagation in a rotating reference system, the Sagnac effect, which depends on the configuration of the satellite and the antenna and amounts to more than 100 ns in some cases. Since the accuracy of GPS time comparisons is less than 10 ns, all these relativistic effects are taken into account every day in the international time comparisons (Allan *et al.*, 1985).

Synchronizations of much better clocks with much more precise comparison methods, which are expected in the near future, can be made by taking all the relativistic terms up to 10^{-18} into account (Wolf and Petit, 1995).

5.2 Pulsar Timing Observations

Pulsars are rapidly rotating neutron stars emitting beams of electromagnetic radiation. Their observed periods range from 1 ms to a few seconds and have been found to be very stable except for occasional irregularities. If we regard a pulsar as a stable clock that emits pulses regularly, then timing observations of the pulsar, yielding a series of measurement of the pulse arrival time by the observer's clock, is a kind of the remote clock comparison, which may be called an interstellar clock comparison. In this comparison, the propagation delay involves the following: the geometrical change of distance between the pulsar and the observation site originating from a relative acceleration due to the differential rotation of our Galaxy (Damour and Taylor, 1991), orbital motion of the pulsar if it forms a binary system, and the orbital and rotational motion of the Earth; change in propagation speed of the electromagnetic wave originating from the fluctuation of dispersion measures in the circumstellar, interstellar, and interplanetary plasmas; and Shapiro delays (Shapiro, 1964) in the gravitational fields of a possible companion star, possible gravitational lensing bodies, and the solar system bodies, as well as the effect of gravitational waves possibly crossing the propagation path, and change of relativistic time difference between the terrestrial time scale and the time scale referred to the solar-system barycenter. Timing observations of the binary pulsar PSR 1913+16 provide a successful test for the existence of gravitational radiation through the precise determination of the orbital motion of the pulsar (Taylor *et al.*, 1979). The pulsar PSR 1937+21 having a period of 1.6 ms shows very small irregularities in the period. Because of the intrinsic increase of the period, which is about 10 times larger than the deceleration of the Earth's rotation, the pulsar alone cannot give a good measure of time. However, the pulsar has probably smaller random instabilities than the best atomic time scales for averaging times longer than one year, and use of the pulsar may improve the long-term stability of atomic time scales (Davis *et al.*, 1985). The timing observation of this pulsar has also set an upper limit on the energy density of stochastic gravitational waves that may have been generated, in the early stage of our Universe, by vibrations of cosmic strings, collisions of cosmic bubbles or giant black holes, or other unknown processes.

5.3 Metrological Standards

5.3.1 Standard of Length The speed of light, $c = 299\,792\,458$ m/s, consists of a number, 299 792 458, a unit of length, 1 m, and a unit of time, 1 s. The speed of light was determined by measuring the wavelength and the frequency of electromagnetic radiation. The relation $c = f\lambda$ gives the value of the speed of light. Limitations on accuracy of the measurement due to the definition of the length forced us to face the problem of the definition of the unit of length. The uncertainty in the measured number for the speed of light was always limited by the uncertainty of wavelength, which comes from the uncertainty of the unit of length.

This problem was solved by introducing a revolutionary new redefinition of the scale of length. The new definition of the meter, adopted by the 17th General Conference on Weights and Measures (CGPM) in 1983, is as follows: "The meter is the length of the path travelled by light in vacuum during a time interval of 1/299 792 458 of a second." In this way, the meter is defined in terms of the second because frequency can be measured much more accurately than wavelength. The numerical value of the speed of light, 299 792 458, is exactly fixed in this new definition.

Since relativity requires that the speed of light is an invariant physical quantity and that time multiplied by the light velocity is a metrical component of four-dimensional space-time along with the space components, it seems quite natural from the relativistic point of view to define a single metrical unit, such as the second.

The meter is realized practically by interferometric techniques using optical or infrared radiation. In order to meet the definition, it is necessary to compare the frequency of radiations in the optical and infrared region with one in microwave region with sufficiently high accuracy. The development of frequency measurement in the optical or near-infrared region is still a necessary spectroscopic activity to provide frequency and wavelength standards in all parts of the electromagnetic spectrum (see Secs. 3.3 and 3.4).

5.3.2 Voltage Standard In the ac Josephson effect, the frequency of oscillations generated by a voltage bias V applied to a weak link of superconductor is given by $f = (2e/h)V$. This relation is used to define the volt in terms of frequency with the defining constant, $2e/h$ = 483 597.9 GHz/V (Taylor and Witt, 1989; see also CONSTANTS, FUNDAMENTAL). 1- and 10-V Josephson voltage standards are now commercially available, and 1-V standards in the national metrological laboratories have been compared by means of a transportable apparatus with uncertainty below 3×10^{-10} in each case.

Works Cited

Allan, D. W. (1975), in *Proceedings of the 29th Annual Symposium on Frequency Control,* Fort Monmouth, NJ: US Army Electronics Command, p. 404.

Allan, D. W. (1989), in: A. De Marchi (Ed.), *Frequency Standards and Metrology,* Heidelberg: Springer, p. 29.

Allan, D. W., Ashby, N. (1986), in: J. Kovalevsky, V. A. Brumberg (Eds.), *Relativity in Celestial Mechanics and Astronomy,* Dordrecht: Reidel, p. 299.

Allan, D. W., Davis, D. D., Weiss, M., Clements, A., Guinot, B., Granveaud, M., Dorenwendt, K., Fischer, B., Hetzel, P., Aoki, S., Fujimoto, M.-K., Charron, L., Ashby, N. (1985), *IEEE Trans. Instrum. Meas.* **IM-34,** 118–125.

Baird, K. M., Evenson, K. M., Hanes, G. R., Jennings, D. A., Petersen, F. R. (1979), *Opt. Lett.* **4,** 263–264.

Bessen, R. J. (1989), in: A. De Marchi (Ed.), *Frequency Standards and Metrology,* Heidelberg: Springer, p. 157.

Clairon, A., Laurent, P., Santarelli, G., Ghezali, S., Lea, S. N., Bahoura, M. (1995), *IEEE Trans. Instrum. Meas.* **IM-44,** 128–131.

Damour, T., Taylor, J. H. (1991), *Astrophys. J.* **366,** 501–511.

Davis, M. M., Taylor, J. H., Weisberg, J. M., Backer, D. C. (1985), *Nature* **315,** 547–550.

Day, T., Gustafson, E. K., Byer, R. L. (1992), *IEEE J. Quantum Electron.* **28,** 1106–1116.

Einstein, A. (1905), *Ann. Phys. (Leipzig)* **17,** 891–921.

Evenson, K. M., Wells, J. S., Peterson, F. R., Danielson, B. L., Day, G. W. (1973), *Appl. Phys. Lett.* **22,** 192–195.

Fujimoto, M.-K. (1992), in: H. Sato, T. Nakamura, R. Ruffini (Eds.), *Proceedings of the 6th M. Grossmann Meeting on General Relativity,* Singapore: World Scientific, p. 1182.

Giazotto, A. (1989), *Phys. Rep.* **182,** 365–424.

Guinot, B. (1989), in: J. Kovalevsky, I. I. Mueller, B. Kolaczek (Eds.), *Reference Frames,* Dordrecht: Kluwer, p. 351.

Hafele, J. C., Keating, R. E. (1972), *Science* **177,** 166–170.

Jennings, D. A., Pollock, C. R., Petersen, F. R., Drullinger, R. E., Evenson, K. M., Wells, J. S., Hall, J. S., Layer, H. P. (1983), *Opt. Lett.* **8,** 136–138.

Kartaschoff, P. (1978), *Frequency and Time,* London: Academic.

Klioner, S. A. (1992), *Celestial Mech. Dyn. Astron.* **53,** 81–109.

Kourogi, M., Nakagawa, K., Ohtsu, M. (1992), in *International Conference on Quantum Electronics. Technical Digest,* Vienna: Institut für Nachrichtentechnik der Technischen Universität Wien, p. 110.

Mills, D. L. (1992), *Network Time Protocol (Version 3) Specification, Implementation and Analysis,* Request for Comments No. 1305, available through ftp.

Pagel, B. E. J. (1977), *Mon. Not. R. Astron. Soc.* **179,** 81–85.

Schnatz, H., Lipphardt, B., Helmcke, J., Riehle, F., Zinner, G. (1996), *Phys. Rev. Lett.* **76,** 18–21.

Seidelmann, P. K., Fukushima, T. (1992), *Astron. Astrophys.* **265,** 833–838.

Shapiro, I. I. (1964), *Phys. Rev. Lett.* **13,** 789–791.

Shlyakhter, A. I. (1976), *Nature* **264,** 340.

Taylor, B. N., Witt, T. J. (1989), *Metrologia* **26,** 47–62.

Taylor, J. H., Fowler, L. A., McCulloch, P. M. (1979), *Nature* **277,** 437–440.

Telle, H. R., Meschede, D., Hansch, T. W. (1990), *Opt. Lett.* **15,** 532–534.

Turneaure, J. P., Stein, S. R. (1976), in: J. H. Sanders, A. H. Wapstra (Eds.), *Atomic Masses and Fundamental Constants,* Vol. 5, New York: Plenum, p. 636.

Turneaure, J. P., Will, C. W., Farrell, B. F., Mattison, E. M., Vessot, R. C. (1983), *Phys. Rev D* **27,** 1705–1714.

Uehara, N., Ueda, K. (1993), *Opt. Lett.* **18,** 505–507.

Vessot, R. F. C., Levine, M. W. (1979), *Gen. Rel. Grav.* **10,** 181–204.

Will, C. M. (1981), *Theory and Experiment in Gravitational Physics,* Cambridge U.K.: Cambridge Univ. Press.

Wolf, P., Petit, G. (1995), *Astron. Astrophys.* **304,** 653–661.

Wolfe, A. M., Brown, R. L., Roberts, M. S. (1976), *Phys. Rev. Lett.* **37,** 179–181.

Wong, N. C. (1990), *Opt. Lett.* **15,** 1129–1131.

Wynands, R., Mukai, T., Hansch, T. W. (1992), *Opt. Lett.* **17,** 1749–1751.

Further Reading

De Marchi, A. (Ed.) (1989), *Frequency Standards and Metrology,* Heidelberg: Springer-Verlag.

Drever, R. W. P., Hough, J., Munley, A. J., Lee, S.-A., Spero, R., Whitcomb, S. E., Ward, H., Ford, G. M., Hereld, M., Robertson, N. A., Kerr, I., Pugh, J. R., Newton, G. P., Meers, B., Brooks III, E. D., Gursel, Y. (1983), in: P. Meystre, M. O. Scully (Eds.), *Quantum Optics, Experimental Gravity, and Measurement Theory,* New York: Plenum, p. 503.

Fukushima, T. (1989), in: J. Kovalevsky, I. I. Mueller, B. Kolaczek (Eds.), *Reference Frames,* Dordrecht: Kluwer, p. 417.

Guinot, B. (1989), in: J. Kovalevsky, I. I. Mueller, B. Kolaczek (Eds.), *Reference Frames,* Dordrecht: Kluwer, p. 379.

Jespersen, J., Jane, F.-R. (1982), *From Sundials To Atomic Clocks,* New York: Dover.

Ohtsu, M., Nakagawa, K. (1991), in: Y. Yamamoto (Ed.), *Coherence, Amplification, and Quantum Effects in Semiconductor Lasers,* New York: Wiley, p. 137.

Seidelmann, P. K., Guinot, B., Doggett, L. E. (1992), in: P. K. Seidelmann (Ed.), *Explanatory Supplement to the Astronomical Almanac,* California: University Science Books, Chap. 2, p. 39.

Quinn, T. J. (1993/94), *Metrologia* **30,** 523–541.

Sullivan, D. B., Allan, D. W., Howe, D. A., Walls, F. L. (1990), *Characterization of Clocks and Oscillators,* NIST Technical Note No. 1337, Washington: U.S. GPO.

TOMOGRAPHY

See ACOUSTICAL TOMOGRAPHY; OPTICAL TOMOGRAPHY; X-RAY IMAGING, MEDICAL: TOMOGRAPHY AND RADIOGRAPHY

TOMOGRAPHY, POSITRON-EMISSION

Darrell D. Burckhardt, *The Johns Hopkins University School of Medicine, Baltimore, Maryland, U.S.A.*

	Introduction	485
1.	**Basic Principles**	486
1.1	Radioactive Decay and Ionizing Radiation	486
1.2	Positron Decay	486
1.3	Photon Interactions with Matter	488
2.	**Acquisition Devices**	489
2.1	Imaging Systems	489
2.1.1	Anger Camera	489
2.1.2	Multihead Camera Systems	490
2.1.3	Ring Systems	490
2.2	Annihilation-Photon Detection and Localization	490
2.2.1	Scintillation Detectors	490
2.2.2	Coincidence Electronics	491
2.3	Resolution	492
2.3.1	Spatial Resolution	492
2.3.1.1	Transverse Resolution	492
2.3.1.2	Axial Resolution	494
2.3.2	Temporal Resolution	495
2.4	Sampling	495
2.5	Sensitivity	496
2.6	Coincidence Events	496
2.6.1	True Coincidence Events	497
2.6.2	Random Coincidence Events	497
2.6.3	Scatter Coincidence Events	498
2.7	Dead Time	498
2.8	Uniformity	498
3.	**Tomographic Reconstruction**	499
3.1	Radon Transform	499
3.2	Fourier Reconstruction	500
3.3	Filtered Backprojection	500
3.4	Corrections	501
3.4.1	Normalization	501
3.4.2	Random-Event Correction	502
3.4.3	Scatter-Event Correction	502
3.4.4	Attenuation Correction	503
3.4.5	Dead-Time Correction	504
3.4.6	Calibration	505
4.	**Image Characterization**	505
4.1	Accuracy	505
4.2	Precision	506
5.	**Image Quality**	506
5.1	Contrast	507
5.2	Resolution	507
5.3	Statistical Noise	508
5.4	Signal-to-Noise Ratio in PET	511
6.	**Diagnostic Quality**	513
6.1	Detectability	513
6.2	Information Content	513
6.3	Enhancement	514
6.4	Interpretation	515
7.	**Image Processing**	515
7.1	Arithmetic Operations	515
7.2	Transformations	516
7.3	Filters	516
7.3.1	Classification	516
7.3.2	Derivation	519
8.	**Example Studies and Images**	520
	Glossary	523
	Works Cited	524
	Further Reading	524

INTRODUCTION

Positron-emission tomography (PET) is the noninvasive external detection of positron-emitting radionuclides distributed according to physiological and biochemical function in a living subject. The goal of PET is to obtain knowledge about normal and pathological states in humans and to apply this knowledge to the research, diagnosis, prognosis, treatment, and monitoring of disease. In its most general use, PET refers to the process of producing an image of the distribution of radionuclide in a slice

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

through a living subject. The steps involved in the PET process are

1. acquisition of radionuclide distribution data,
2. reconstruction of quantitative images of the radionuclide distribution,
3. extraction of quantitative measurements of the radiopharmaceutical distribution, and
4. interpretation of results related to functional or parametric information.

The most attractive advantage of PET is that common isotopes within biologically active molecules can be labeled with positron-emitting isotopes with little mass differences (i.e., ^{11}C for ^{12}C and ^{18}F for ^{1}H). The biochemical behavior of these radiolabeled molecules or traceable compounds is known, and they can be introduced into the body in small amounts without affecting the body's normal processes. PET thus provides quantitative measurements of the distribution of labeled biologic and synthetic compounds or drugs.

In Sec. 1, a review of the basic principles of positron decay provides an understanding of the advantages of positron-emitting radionuclides for detection and localization in the living human body. The incorporation of these physical advantages, through annihilation-photon coincidence detection, into commercially available positron scanners is described in Sec. 2. In addition, the limitations of quantification in PET are discussed both in the context of acquisition devices and in regard to their effect on the interpretation of PET results.

Inherently, the interpretation of PET images requires an understanding of the sources of errors, the accuracy of correction methods, and knowledge of the applied image-processing methods producing the resultant images. In Sec. 3, methods of image reconstruction from projections and techniques for correcting the factors affecting the accuracy of PET images are given. Sections 5 and 6 describe quantitative accuracy, image quality, and diagnostic accuracy. A basic understanding of digital image processing as it applies to PET is given in Sec. 7.

Section 8 discusses clinical PET applications and provides some example studies. A list of Further Reading is provided for the interested reader.

1. BASIC PRINCIPLES

1.1 Radioactive Decay and Ionizing Radiation

Radioactive decay is a nuclear transformation process involving the emission of particles and/or photons and the release of nuclear energy. The emissions can interact with atoms, nuclei, and electrons of surrounding matter, whereby energy is imparted or transferred to the surrounding material, causing *ionization* and *excitation.* When referring to emissions from radioactive decay, the term *ionizing radiation* is often used, since ionization is the primary event.

1.2 Positron Decay

Nuclei containing excess protons undergo spontaneous nuclear transformations leading to a more stable condition. To reach greater stability, these nuclei reduce proton content by either *electron capture* or *positron emission.* Electron capture occurs when an inner orbital electron combines with a nuclear proton to form a neutron and a small massless, chargeless particle called a *neutrino.* The overall result is a decrease of the atomic number by one and the emission of characteristic x rays as a consequence of the rearrangement of orbital electrons. Positron emission results when a nuclear proton is transformed into a nuclear neutron, resulting in the ejection of a positively charged electron or *positron* and a massless neutrino from the nucleus. The positron travels a short distance, depending on its kinetic energy and the nature of the material in which it is contained, and combines with an electron in an annihilation event. Because of the laws of conservation of energy and momentum, two 511-keV photons are emitted simultaneously from the site of annihilation at approximately 180° ± 0.25° to each other (see Fig. 1). The noncollinearty of the annihilation photons results because the positron or electron may not be completely at rest upon annihilation.

Since positron emission decreases the atomic number by one, the number of orbital electrons must also decrease by one as soon as the positron is ejected from the nucleus. The atomic mass of the daughter nuclide must be two electron masses smaller

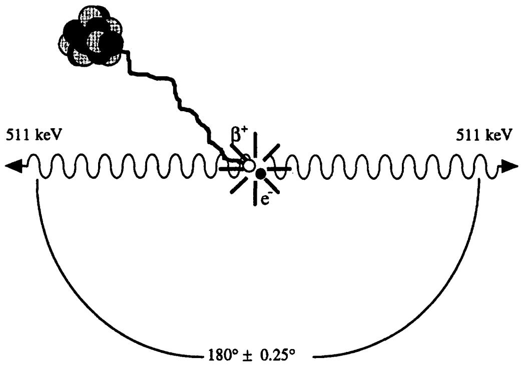

FIG. 1. Diagram of positron emission and annihilation event.

than the parent radionuclide. Since this mass difference eventually manifests itself as energy, a minimum of 1.02 MeV is required for positron emission. Since there is no energy requirement for electron capture, proton-rich nuclei reach stability by these different but competing modes, each having its own characteristic emissions. As the energy difference between parent and daughter nuclei approaches the 1.02-MeV requirement, the probability of positron emission decreases, while the probability of electron capture increases. Positron emission occurs frequently in lighter elements, while electron capture is favored in heavier elements. The transformation of a single unstable atom to a more stable atom is irreversible and is called a *disintegration* or *decay event.*

All unstable nuclei in a sample decay randomly. The number of unstable nuclei, N, and therefore of disintegrations, decreases continuously with time. The rate of decay or the disintegrations per unit time is called the *radioactivity, A*, of a sample.

$$A = -\frac{dN}{dt}. \tag{1}$$

Radioactivity is proportional to the number of unstable nuclei present in a given sample and is given by

$$A = \lambda N, \tag{2}$$

where λ is the *decay constant* representing the constant fraction of activity or number of nuclei disintegrating per unit time interval. The mathematical expression for N at time t_1, $N(t_1)$, is given as

$$N(t_1) = N(t_0)e^{-\lambda(t_1-t_0)}, \tag{3}$$

and since A is proportional to N,

$$A(t_1) = A(t_0)e^{-\lambda(t_1-t_0)}. \tag{4}$$

Thus, for given values of λ and $\Delta t = t_1 - t_0$, the quantity $e^{-\lambda\Delta t}$ is called the *decay factor.* The reciprocal of the decay factor, $e^{\lambda\Delta t}$, can be applied to $A(t_1)$ to obtain $A(t_0)$, the activity *decay corrected* to reference time t_0.

The *half-life*, $T_{1/2}$, of a radionuclide is defined as the time required for an amount of radionuclide to decay to one-half of its original activity and is given as

$$T_{1/2} = \ln(2)/\lambda. \tag{5}$$

Each positron-emitting radionuclide has its own unchangeable characteristics as shown in Table 1.

Nuclear decay is a statistical process in which the probability of a nuclear disintegration event is independent of all other events. A single disintegration can occur at any time. With this in mind, $N(t_1)$ given in Eq. (3) must be sufficiently large to allow for statistical averaging when using the decay constant to determine the radioactivity of a sample at any other time. Because of nuclear decay, multiple measurements are often impractical, and a single measurement is made as an estimate of the true value of the radioactive sample. The probability of obtaining the true value from a single radioactivity measurement, N, follows the *Poisson* distribution and is related to the variance σ^2 of the distribution. For the Poisson distribution,

Table 1. Physical properties of positron-emitting radionuclides of medical interest.

Isotope	Half-life (min)	Maximum positron energy (MeV)	Positron emission (%)
^{11}C	20.30	0.96	99.8
^{13}N	9.97	1.19	100.0
^{15}O	2.03	1.70	100.0
^{13}F	109.80	0.64	97.0
^{68}Ga	68.30	1.90	88.0
^{82}Rb	1.27	3.35	96.0

the variance is equal to the true value, but since we cannot know the true value exactly, we use N as an approximation. One measure of the statistical reliability of N is the percent coefficient of variation, defined as

$$\%\mathrm{CV} = 100\% \cdot \frac{\sigma}{N} = 100\% \cdot \frac{\sqrt{N}}{N} = \frac{100\%}{\sqrt{N}}, \qquad (6)$$

which indicates that the uncertainty in a single measurement decreases with increasing number of counts.

1.3 Photon Interactions with Matter

When photons traverse matter, some pass through without any alteration and some lose all or part of their energy through a series of interactions with the atoms, nuclei, or electrons of the material. Photon interactions occur in several different ways that result in high-energy electrons and secondary photons of lesser energy. The two principal mechanisms by which annihilation photons interact with matter are the *photoelectric* and *Compton scattering* processes. In the photoelectric process, a photon is completely absorbed by an atom, through the transference of all its energy to an orbital electron. This photoelectron is ejected from the atom with kinetic energy E_{pe} equal to the energy absorbed minus the binding energy of the photoelectron ($E_{pe} = E_0 - E_B$). Compton scattering is modeled as an elastic collision between a photon and an orbital electron. The incident photon is deflected into a different direction through a scattering angle with reduced energy, E_{sc}. The relationship between the scattering angle θ and the energy of the incident photon, E_0, is given by

$$E_{sc} = \frac{E_0}{1 + (E_0/m_0c^2)(1 - \cos\theta)}, \qquad (7)$$

where $m_0c^2 \approx 511$ keV is the rest energy of an electron.

The kinetic energy of the ejected recoil electron, E_{re}, is equal to the energy absorbed by the atom minus energy of the scattered photon ($E_{re} = E_0 - E_{sc}$). From Eq. (7), the amount of energy transferred from the photon to the recoil electron does not depend on any property of the absorbing material; it depends only on the incident photon energy and scattering angle. The angular distribution of the scattered photon is dependent upon the incident photon energy E_0. For 511-keV photons, the scattering angle is usually small (i.e., the photon scatters in the forward direction), resulting in Compton-scattered photons near the energy of the incident photons.

The probability that a photon interaction will occur depends on the energy of the photon and the characteristics of absorbing or attenuating material. For 511-keV photons in body tissue, the most probable interaction is Compton scattering. In order to examine the quantitative aspects of photon attenuation, consider Fig. 2, in which a beam of collimated (parallel) monoenergetic photons is incident on an absorber of thickness Δx. As the beam with incident intensity I_0 passes through the absorber, some of the photons are completely absorbed, some are scattered, and some are transmitted through the absorbing material with intensity I. The probability that a photon will be transmitted through the absorber is given by

$$I/I_0 = e^{-\mu x}, \qquad (8)$$

where μ is the linear attenuation coefficient of the absorber for the energy of the incident radiation. For a positron-emitting radionuclide surrounded by photon-attenuating media (see Fig. 3), the probability that either of

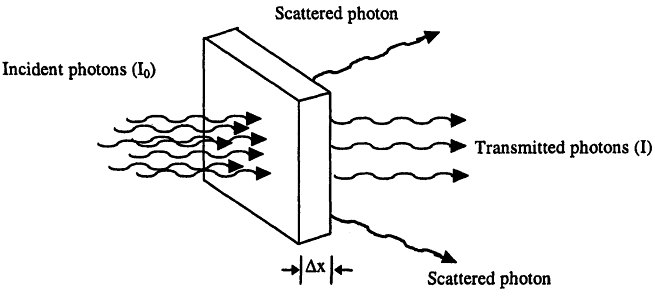

FIG. 2. Monoenergetic photon beam transmitted and scattered through an absorbing medium.

the two annihilation photons will exit the body without interaction is given by the product of each photon's probability of escape:

$$I/I_0 = e^{-\mu x_1}e^{-\mu x_2} = e^{-\mu(x_1+x_2)} = e^{-\mu l}, \qquad (9)$$

where x_1 and x_2 are the respective distance the two photons travel through the absorbing material (as shown in Fig. 3) and l is the total absorber thickness ($l = x_1 + x_2$).

The attenuation of an annihilation photon pair, thus, is dependent on the total distance l along the path traveled and on the absorbing material's linear attenuation coefficient μ, and is independent of the position of the annihilation event. This is an important fact in estimating the absolute radioactivity concentration in an object, which is confounded by photon attenuation, as discussed later.

2. ACQUISITION DEVICES

Because of their high energy, annihilation photons are not well suited for detection by conventional nuclear-medicine imaging devices. Devices for the imaging of positron emitters typically use special annihilation coincidence-detection electronics and detectors that efficiently absorb high-energy photons. However, conventional imaging devices using special collimators and dual-head coincidence detectors are becoming popular, as the use of positron-emitting radiopharmaceuticals becomes more attractive in the clinical setting. Even so, this discussion will emphasize dedicated tomographic devices.

Tomographic imaging devices differ from the majority of nuclear-medicine imaging approaches, in that the data acquisition and image reconstruction are based on rigorous mathematical algorithms. The data collected consist of directional views of the radioactivity distributed within an object, taken at many angles around the object. The value at any point of a particular view represents the contribution of all photon emissions detected along lines extending through the object projecting onto that point. These acquired projections are the raw input data for the computed methods of tomographic reconstruction of images of the radioactivity distribution.

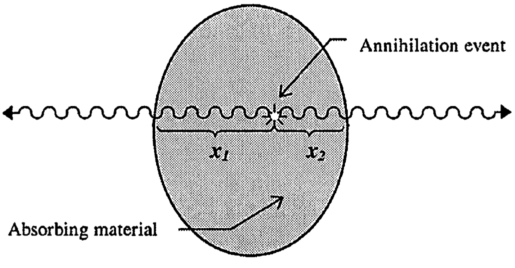

FIG. 3. Annihilation photons travel respective distances of x_1 and x_2 through an absorbing material.

2.1 Imaging Systems

Imaging systems consist primarily of an acquisition device and its electronics, a computer, and a workstation. The workstation is the interface between the operator and the computer, and it controls the acquisition device and performs post-acquisition data processing such as tomographic reconstruction. The major difference among most imaging systems is in the acquisition device and its electronics, reflecting the specific application involved and thus the type of processing required.

2.1.1 Anger Camera Anger cameras use heavy-metal absorbers as collimators, placed in front of position-sensitive detectors, to discriminate against photons traveling along undesirable directions toward the detector. Because of the attractiveness of positron-emitting radiopharmaceuticals, Anger cameras have been fitted with high-energy (511 keV) collimators. Since most of the potentially useful photons are removed by the absorptive collimation, and because the detectors employed are usually designed for much lower-energy (100–200 keV) single-photon–emitting radionuclides, the performance of these systems is poor in comparison with true PET imaging systems.

Planar images obtained from the Anger camera are projection images. The radioactivity in the organ of interest can be obscured by activity either in front of or behind the organ, and the photons originating from the organ are attenuated by the surrounding tissue. These problems are overcome by rotating the camera to acquire multiple projections and correcting the attenuation in a tomographic reconstruction (described later). This method of radionuclide imaging is called single-photon–emission computed tomography or SPECT.

2.1.2 Multihead Camera Systems As with the Anger camera, the attractiveness and availability of positron-emitting radiopharmaceuticals have spawned a great interest in development of relatively inexpensive imaging devices for clinical and research use. These systems, like the more complex dedicated tomographic systems, employ "electronic collimation" of position-sensitive detectors through coincidence circuitry. However, because of their simple design, they do not have the ability of simultaneous detection of photon emissions at all angles about the object, as do ring systems. In order to obtain tomographic information, these detector systems must undergo some kind of motion (i.e., rotation and/or translation) and are useful for measuring radionuclide distributions that do not change rapidly in time. When kept stationary, these systems are utilized for determining rapid rates of uptake or elimination in studies that do not require tomographic reconstruction for radionuclide localization and quantitation.

2.1.3 Ring Systems The majority of PET scanners in use today utilize a multiple-ring design; these represent dedicated tomographic devices. Typically classified by ring diameter into head or whole-body scanners, almost all commercial ring PET scanners operate on the same basic principle of a circumferential array of block detectors (described later) that provide both high detection efficiency and rapid data acquisition. Recent advances in technology have provided three-dimensional (3D) imaging in addition to the more common two-dimensional (2D) or transverse-slice imaging.

The physical configuration of most commercially available PET systems consists of one or more circular rings of detectors. The field of view (FOV) is determined by scanner and detector geometry and the electronic configuration of detector pairs. If the detector size is constant, then the number of rings determines the axial FOV. The transaxial FOV is dependent upon the ring diameter and the selection of coincidence detector pairs. Conventional transverse-slice (2D) imaging devices are designed to maximize the detection of annihilation photons within a transaxial plane or transverse slice while minimizing the detection of out-of-plane activity. The former is achieved by the fact that each detector in a ring can be in coincidence with any other detector in the ring. The latter is accomplished by separating rings by thin heavy-metal shielding called *septa.*

Modern PET scanners have retractable septa, allowing scanners to be operated in a fully 3D mode. By eliminating septa from the FOV and electronically allowing each detector to be in coincidence with all other detectors in the scanner, maximum detection of annihilation photons is achieved.

2.2 Annihilation-Photon Detection and Localization

Radiation detection is based on the interaction of photons with matter. A photon is detected through the ionization effects of the secondary electrons produced by its deposition of energy within a detector. As previously stated, the probability of photon–detector interactions is dependent on both photon energy and the physical properties of the detector. Since the annihilation-photon energy is high, the stopping power of the detector must also be sufficiently high.

The localization of a positron emitter in an object is based on the fact that annihilation photons are emitted simultaneously at approximately 180° to each other. This property allows the coincidence detection of the two annihilation photons by a pair of detectors placed on opposite sides of the object, as shown in Fig. 4. This *electronic collimation* of annihilation photons is a more efficient method than conventional *absorptive collimation* in the detection and subsequent localization of annihilation photons.

2.2.1 Scintillation Detectors Scanners in nuclear medicine are based on the use of scintillation detectors. Scintillation detectors are crystalline material containing small amounts of impurity. When ionizing radiation deposits energy within a detector, a free secondary electron is produced in the crystal lattice. This electron moves through the lattice until it encounters an impurity and becomes trapped. The trapping atom is in an excited state and will relieve this excitation by rapidly dissipating energy through the emission of visible or ultraviolet light. Any

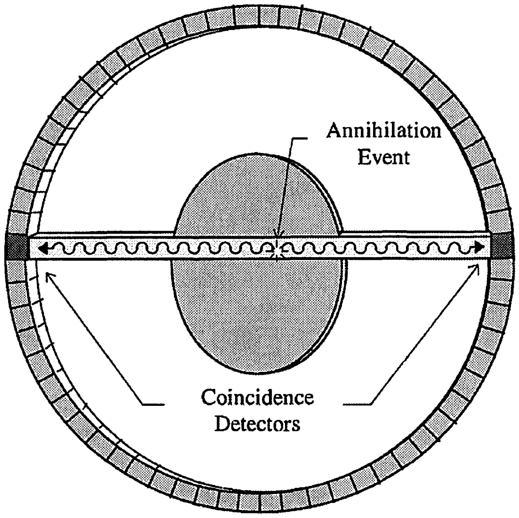

FIG. 4. The coincidence detection of annihilation photons arriving in opposing detectors.

energy not emitted as light will be dissipated eventually as heat. Since the energy levels of the excited impurity are different from those of the main substance, there is a high probability that visible light will escape the crystal. Each photon of visible light represents a free electron produced by the breaking of a valence bond by a definite amount of energy. The total amount of visible light thus produced is proportional to the total energy deposited in the detector crystal.

Important characteristics of scintillators are photon yield or the average number of light photons per keV of ionizing radiation absorbed (the higher the better), density and atomic number (stopping power; the higher the better), the decay of the burst of visible light (decay time; the shorter the better), emission of light in the visible energy range (matched to the characteristics of the photocathode of the photomultiplier tube), little or no self-absorption of visible light, and low cost. Many different scintillators have been used in nuclear medicine, but thallium-activated sodium iodide, NaI(Tl), has been the most common. Although NaI(Tl) is an efficient scintillator, it is fragile and hygroscopic, requiring special handling, temperature, and sealing to prevent fractures and discoloration, which impair visible light transmission through the crystal. The first commercial PET scanners used NaI(Tl), but most modern PET scanners use a block concept of bismuth germanate ($Bi_3Ge_4O_{12}$) or BGO block detectors (described in the next section). BGO has become the scintillator of choice because of its high density and atomic number, which indicates that high stopping power is the most important detector characteristic (see Table 2). However, because it has lower light output when compared with NaI(Tl), BGO has poorer energy resolution and timing characteristics. Lutetium orthosilicate [$Lu_2SiO_5(Ce)$] or LSO is a new scintillator material that, like BGO, has high stopping power but also has significantly better light output and shorter decay time.

Although scintillators are efficient photon detectors, they are inefficient energy converters. Through the use of a photomultiplier tube, PMT, optically coupled to the scintillation detectors, weak light signals striking the PMT photocathode are converted to a pulse of electrical current that is proportional to the intensity of the light signals. The amount of electrical current produced at the PMT anode is, thus, also proportional to the amount of energy deposited in the crystal. The electrical pulse of the PMT is used by electronic components that determine the location of photon–detector interaction, the amount of energy deposited in the detector, and the time when the interaction took place.

2.2.2 Coincidence Electronics As mentioned above, most commercial PET scanners are designed with BGO block detectors in coincidence. A block detector is one in which an array of small detector crystals (scintillators) are interfaced to a small num-

Table 2. Comparison of NaI(Tl) and BGO scintillator properties.

	NaI(Tl)	BGO	LSO
Density (g/cm^3)	3.67	7.13	7.4
Effective atomic number	50	74	172
Decay time (ns)	230	300	40
Photon yield (per keV)	40	4.8	30
Energy resolution (FWHM %)	7	15	11
Hygroscopic	Yes	No	No
Emission wavelength (nm)	415	480	420

ber of PMTs to generate the location of the detector–photon interaction. The encoding or addressing scheme used to identify the crystal in which an annihilation photon was absorbed, commonly referred to as Anger logic, is obtained by comparing the relative output from individual PMTs within a block (see Fig. 5). The total output of all PMT pulses in a block is used to determine the amount of energy deposited and the time when the photon was detected in the block. Once produced, the block pulse is compared to preset upper and lower thresholds. If the pulse amplitude falls between the upper and lower voltages established, the energy requirements for the event are fulfilled and the timing and positioning information are retained.

The coincidence logic circuit receives the timing and position information and with the timing information checks for the occurrence of the simultaneous detection of another event in an opposing detector. A coincidence event is recorded only if pairs of events in opposing detectors occur within a narrow interval of time determined by the resolving time of the detectors and the electronics used for energy discrimination and position determination. After a coincidence event is detected it is stored in computer memory at a location corresponding to the particular line of response (LOR) between the coincidence-detector pair. In some systems, events not found in coincidence, called *singles,* are also stored in computer memory according to detector position.

Timing resolution is thought of as the capability of the PET system to determine the time of photon–detector interaction as in time-of-flight acquisition, where temporal information is used to improve the accuracy of image reconstruction. In order to use time-of-flight information, very fast detectors (i.e., short decay time) and electronics must be used. BGO is not sufficiently fast, and most current scanners do not use this additional time-of-flight information.

2.3 Resolution

Resolution is the process of separating or clarifying the parts or components of an object. In PET, resolution is a measure of a scanner's ability to localize the positron emitter both in time and in space.

2.3.1 Spatial Resolution The spatial resolution of a PET scanner determines its visual and quantitative capabilities. Typical measurements of spatial resolution include *transverse* and *axial* resolution. Spatial resolution is limited physically by the kinetic energy of the positron and the distance it travels from the nucleus prior to its annihilation, the deviation from exact collinearity of the annihilation photons, the dimensions of the detectors, and the distance between opposing detector pairs.

2.3.1.1 Transverse Resolution. Transverse or in-plane resolution is characterized by the response of the detector system to a

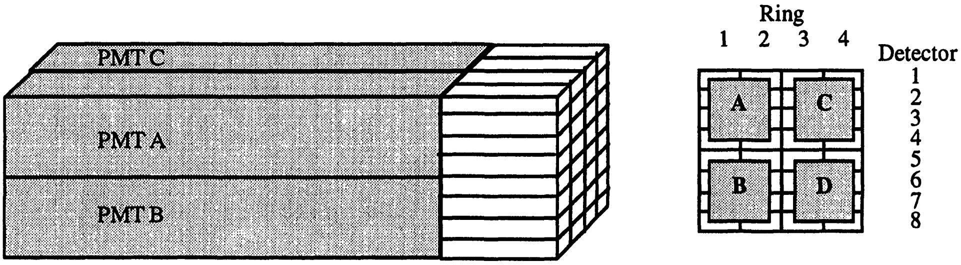

$$\text{Ring} = \frac{(A+B)-(C+D)}{A+B+C+D}$$

$$\text{Detector} = \frac{(A+C)-(B+D)}{A+B+C+D}$$

FIG. 5. Block-detector array and Anger logic for crystal identification.

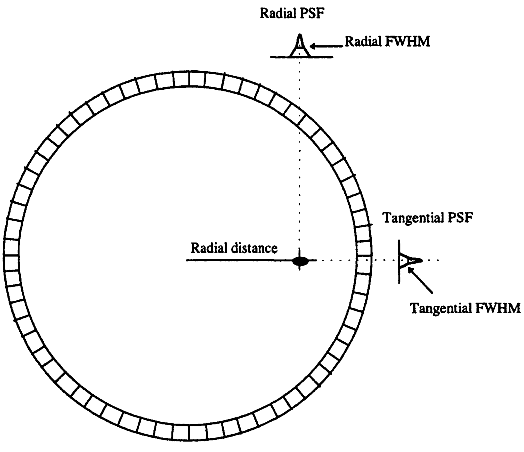

FIG. 6. Transverse resolution determined at a radial distance from the PET scanner FOV center in both the radial and tangential point-spread functions (PSFs).

point source or impulse of radioactivity. From an image of a point source, a count profile through the center of the source, called a point-spread function (PSF) or impulse response function, is obtained along lines through the point in both radial and tangential (perpendicular to the radial) directions, as shown in Fig. 6. The most common measures of transverse spatial resolution are the full width at half maximum (FWHM) and the full width at tenth maximum (FWTM) of the radial and tangential PSFs at various radial distances from the transverse FOV center.

Spatial resolution is the distance that two point sources must be apart from each other in order to be distinguished as separate objects. Shown in Fig. 7 are the contributions of two individual point-source profiles at various distances of separation to the resultant line profile through both point sources.

If a radioactivity distribution can be conceptualized as a linear combination of point sources, then the image of the radioactive object is the linear combination of scaled point-spread functions superimposed on one another. The object's profile thus becomes a blurred version of the actual profile, and the image is an inaccurate reproduction of the actual object radioactivity distribution, as shown in Fig. 8. In essence, the image is the convolution of the object with the scanner's PSF.

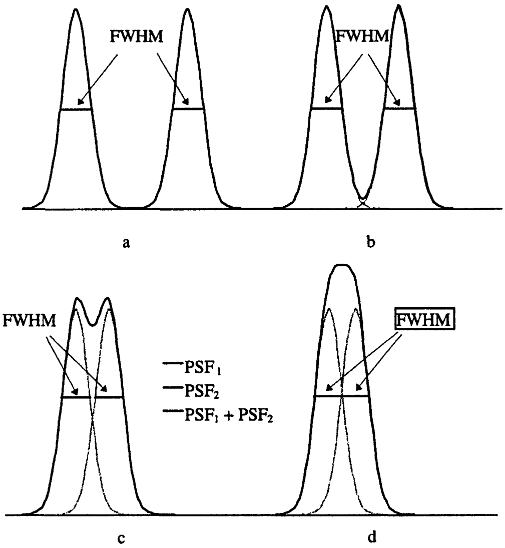

FIG. 7. The effects of imperfect spatial resolution on the discrimination of two impulse response functions. (a)–(c) As the distance between two sources approaches the FWHM, the ability to distinguish one point source from the other becomes more difficult until (d) they no longer appear as two separate objects.

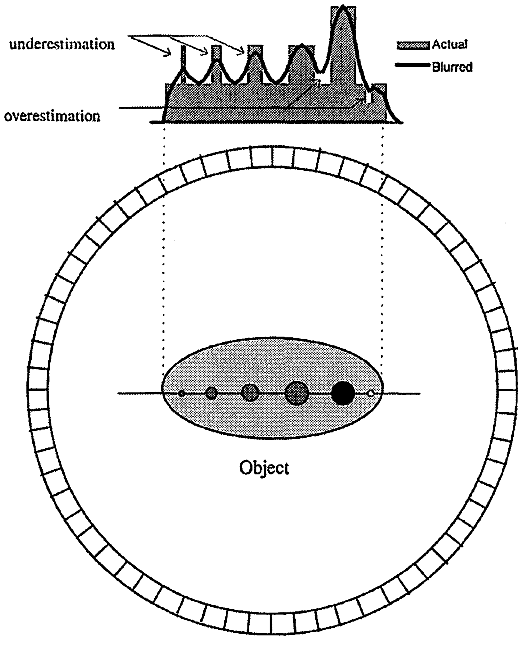

FIG. 8. The profile of an actual object (shaded) and blurred (line) by a PSF.

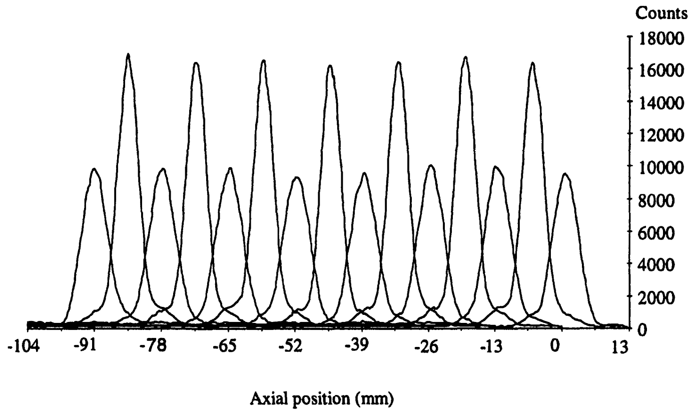

FIG. 9. Axial response function obtained at the center of the FOV in a 15-slice PET scanner by moving a point source in 0.5-mm steps in the axial direction.

Because of the limited spatial resolution of the PET system, as objects become smaller the detected object radionuclide concentration is reduced. As shown in Fig. 8, all objects in the image can become "contaminated" with radioactivity from surrounding objects, making the delineation of the radionuclide concentrations between different objects difficult.

2.3.1.2 Axial Resolution. Axial resolution is analogous to transverse resolution in that an impulse response along the axial direction is obtained from a stationary point source. This parameter is only appropriate if the axial sampling (determined by the axial dimension of the detectors) is sufficient, as discussed below. What is often indicated as axial resolution is more appropriately termed the *axial response function.* The axial response function is obtained, for each scanner plane, from the coincidence counts of a point source of radioactivity as it passes in the axial direction through the scanner FOV in small increments at various radial positions, as shown in Fig. 9. Parameters used to describe the axial response function are the FWHM, FWTM, and the center-to-center slice separation for a specific radial distance from the center of the FOV.

The axial response functions indicate that a source has its greatest contribution to the PET scanner response (highest count recovery) at the axial center of each plane and that interplane (cross) slices have higher sensitivity than do in-plane (direct) slices (see Fig. 10). Transverse and axial resolution define the smallest volume that can be resolved by the PET scanner. The total amount of radioactivity detected from this volume determines the volume's observed radionuclide concentration. For example, the amount of radioactivity from a single small object partially occupying this volume, in the absence of any other source of radioactivity in the volume, can only be resolved as originating over the entire volume; thus the object's measured radionuclide concentration is less than its true radionuclide concentration. In addition, the presence of other sources in the volume (i.e., from adjacent objects) also contribute or spill over radioactivity into the object. These *partial-volume* and *spillover* effects reduce both qualitative and quantitative accuracy and are most noticeable for small objects and the edges of all objects. Partial-volume effects become less significant as ob-

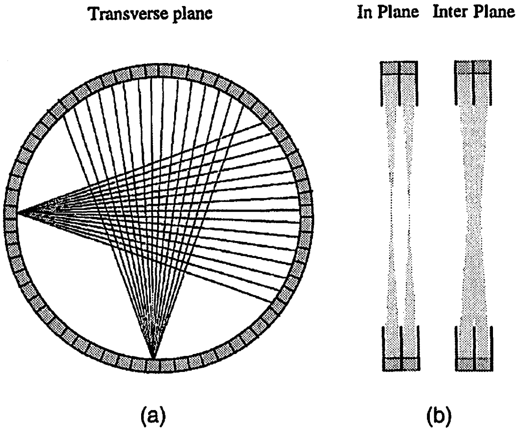

FIG. 10. Sampling patterns of detectors in a multiple circumferential array (ring) configuration.

ject size increases relative to the spatial resolution in all directions, but spillover effects (and blurring) remain along edges defining the boundaries between structures.

2.3.2 Temporal Resolution Temporal resolution is the ability of the PET scanner to depict the changing distribution of coincidence events in time. For this purpose, PET scanners may be operated in either histogram or list mode. In histogram mode (by far the most common method of acquisition), all coincidence events detected along a particular LOR over a specified interval of time are accumulated at a corresponding location in computer memory. For each new acquisition frame, the information accumulated must be transferred, stored, and cleared. The temporal sampling rate is thus dependent upon the amount of memory required and limited by the detection electronics described in the previous sections. In general, the longer the accumulation time interval or frame, the higher the number of events collected, which in turn provides a statistically better (less noisy) estimate of the radionuclide concentration. Alternatively, the PET scanner can be operated in list mode, where coincidence events are stored sequentially as they are detected with their position and timing information. This information can be used for retrospective reconstruction of variable-duration image frames. The ability to reflect accurately the temporal changes in radionuclide concentration is important for the determination of pharmacological rate constants obtained by tracer kinetic modeling of image-derived measurements of physiological processes. Generally, image acquisition protocols consist of a number of short scans followed by scans increasing in duration. The short scans are required to measure the initial rapid change in tissue tracer concentration. Although noisy, more initial short measurements are often more desirable than fewer, less noisy or longer measurements in characterizing the radionuclide concentration as a function of time.

2.4 Sampling

With arrays of discrete detectors, conventional PET scanners measure radionuclide distributions at discrete points. These points are thus separated by distances equal to the detector's dimensions. As discussed above, one of the limitations imposed on the spatial resolution of a PET system is detector size. In earlier PET systems, with large detectors, the activity distribution was grossly undersampled, and as a result, images suffered from poor resolution and artifacts. One common method to increase sampling was to employ some sort of gantry motion (i.e., wobble or translation). With the use of small high-density detectors, the detector spacing or the linear sampling interval is smaller, reducing the need for additional gantry movement to improve spatial resolution. Since PET scanners using circumferential arrays of block detectors are operated in coincidence with multiple detectors on the opposing side of the array and in coincidence with detectors in other rings (Fig. 10), the continuous radioactivity distribution is sampled or viewed from many discrete angles around an object (*angular sampling*) simultaneously. This configuration, however, does not provide uniform sampling, but is sufficient (not optimal) to provide quantitatively accurate images of the radionuclide distribution.

Expressed differently, information is lost in the sampling process, and the solution becomes one of establishing sampling conditions that allow the radioactivity distribution to be recovered accurately. The Fourier transform and the convolution theorem provide the tools for an in-depth study, as is addressed below. For now, it suffices to recognize that complete recovery of the continuous radioactivity distribution from discrete samples requires a linear sampling interval whose spacing across each projection, d, satisfies

$$d = 1/2\nu_{\max}, \tag{10}$$

where $\nu_{\max}$, known as the Nyquist frequency, is the maximum spatial-frequency component of the continuous radioactivity distribution. If the continuous radioactivity distribution sampled at interval d happens to be limited to frequencies less than $\nu_{\max}$, then the distribution is completely determined by its discrete samples. If Eq. (10) is not satisfied, then the frequency (information) content of the radionuclide distribution above $\nu_{\max}$ is erroneously contained in frequencies below $\nu_{\max}$. This effect, referred to as *aliasing*, prevents the radionuclide distribution from being recovered fully and leads to reduced contrast and image artifacts. In practice, the sampling interval is limited by finite detector width, and the frequency spectrum is examined to determine whether aliasing effects are minimized (i.e., the Fourier transform approaches zero as the frequency approaches $\nu_{\max}$). With circular arrays of coincidence detectors, the number of view angles, w, is equal to the number of detectors over an arc of 180° around the object:

$$w = \frac{n}{2} = \frac{2\pi r}{2d} = \frac{\pi D_{\text{ring}}}{2d}, \tag{11}$$

where n is the total number of detectors in the ring, r is the gantry radius equal to $\frac{1}{2}$ of the ring diameter D_{ring}, and d, again, is the linear sampling distance. The angular sampling interval should be equal to the linear sampling about the periphery of an object located within the FOV diameter, D_{FOV}, having the same requirements of the linear sampling interval. These sampling requirements are both met when D_{FOV} is equal to $D_{\text{ring}}/2$.

The size of the detectors and the scanner geometry thus determine the sampling of the radioactivity distribution. The effects of insufficient transverse sampling in PET systems are a suppression of high-frequency information (reduction of spatial resolution) and aliasing. Another major problem with stationary PET systems employing a circular array of detectors is the fact that transverse sampling is irregular, producing spatially variant resolution. This is the reason why spatial resolution is usually measured both tangentially and radially, and at different radial positions, as described above.

The axial sampling or interplane spacing between rings is also important for multiple-ring systems. Since two rings sample the interplane space, the sensitivity is better; however, the interplane axial response function varies to a greater degree in comparison with in-plane resolution (see Fig. 10).

2.5 Sensitivity

Because of the multiple sampling paths for each detector, the sensitivity of a PET system is greatly increased in comparison with single sampling path systems like SPECT (single-photon emission computed tomography, which is based on the Anger camera; see above). Sensitivity is the counting rate (c/s) produced per unit radioactivity concentration ($\mu\text{Ci}\cdot\text{mL}^{-1}$) for a given distribution. Sensitivity is expressed as the counting rate of a ring or plane for an activity concentration of a pure positron emitter uniformly distributed in a 20-cm-diameter phantom. The sensitivity of a ring PET system is dependent on the *geometric* efficiency of the system and the *intrinsic* efficiency of the detector crystals.

The total geometric efficiency for a circumferential array is the sum of individual detector geometric efficiencies g given by

$$g = S/4\pi r^2, \tag{12}$$

where S is the surface area of the detector face and r is the ring diameter. Since g decreases according to $1/r^2$ and the total number of detectors increases in proportion to r, the total geometric efficiency and therefore the total coincidence counting rate decrease in proportion to $1/r$ for a source located at the ring's center.

Geometric efficiency is thus defined as the fraction of photons emitted by a source that actually strike detectors. The *intrinsic efficiency* of the detectors, primarily determined by the detector material and size, is defined as the fraction of photons striking the detector that deposit energy above the threshold of acceptance for the detectors.

2.6 Coincidence Events

When a photon deposits its energy in a detector, an electronic pulse is produced whose amplitude is proportional to the energy deposited. Most PET scanners employ two discriminators for energy verification. The first is the timing discriminator, which prevents noise pulses and very low-energy events from triggering timing event pulses. Once the pulse amplitude exceeds a low-energy threshold, a timing event is generated corresponding to the instant the detector was hit by a photon. The timing pulse is delayed pending verification of the initial pulse by a second energy discriminator set for the lowest and highest acceptable photon energies. This timing pulse has a finite width τ, called the *coincidence resolving time.* Timing pulses from opposing detectors that overlap within ± τ of each other are considered in coincidence. A *coincidence event* is thus defined as two photons depositing sufficient energy in opposing detectors within the *coincidence timing window* of 2τ (i.e., $\pm\tau$).

The width of the timing pulse is chosen such that most of the signal from photons emitted from the same annihilation event fall within the timing window of the coincidence circuitry for maximum sensitivity. Because of this finite pulse width τ, there is a chance that photons from unrelated annihilation events will accidentally or randomly be detected in coincidence. In addition, energy discriminators form a window that the photon energy deposited in the detector must fall within. The lower setting is used to discriminate against the detection of scattered photons; however, for 511-keV photons, scatter is in the forward direction, imparting little energy to recoil electrons, resulting in photon energies at or near 511 keV. The upper setting is used to discriminate against additional energy deposited by multiple photons striking the detector at the same time (*multiple coincidences*). Thus, for a detector pair the total coincidence counting rate, R_{total}, is given by

$$R_{\text{total}} = R_{\text{true}} + R_{\text{scatter}} + R_{\text{random}}, \qquad (13)$$

where R_{true}, R_{scatter}, and R_{random} are the true, scatter, and random counting rates, respectively. Figure 11 graphically depicts the events producing these counting rates in a coincidence-detector pair.

2.6.1 True Coincidence Events True coincidence events occur when two unscattered photons from the same annihilation event are detected in coincidence in opposing detectors. The true coincidence counting rate R_{true} is given by

$$R_{\text{true}} = E\epsilon_1\epsilon_2 g e^{-\mu l}, \qquad (14)$$

where E is the annihilation-photon pair emission rate, ϵ_1 and ϵ_2 are the intrinsic efficiencies of each detector in the coincidence pair, g is the geometric efficiency, μ is the linear attenuation coefficient of the absorbing medium, and l is the length or distance traversed by the annihilation photon pair through the absorbing medium [see Eq. (9)].

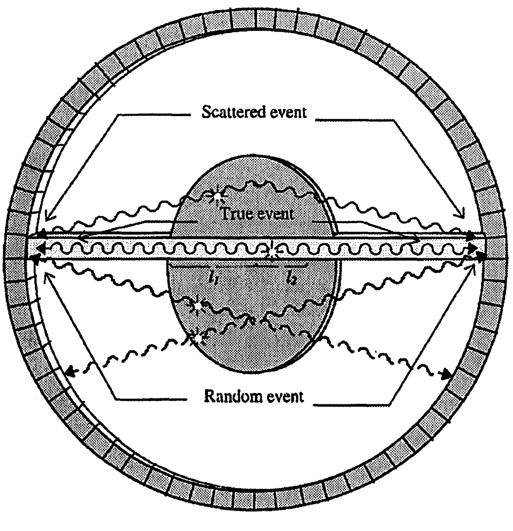

FIG. 11. Coincidence detection of annihilation events. True events occur within the volume defined by the detector pair and electronic collimation. Scatter coincidences involve events occurring outside the sensitive volume in which one or both annihilation photons arrive simultaneously at opposing detectors. Random events occur by chance detection of photons resulting from unrelated annihilation events.

The attenuation of photons reduces the total number of true coincidence events detected.

2.6.2 Random Coincidence Events The random or accidental counting rate R_{random} is given by

$$R_{\text{random}} = 2\tau R_{\text{singles1}} R_{\text{singles2}}, \qquad (15)$$

where R_{singles1} and R_{singles2} are the singles counting rates from detector 1 and detector 2 of the coincidence pair, respectively, and τ is the width of the pulse. The singles counting rate R_{singles} for a source located midway between two coincidence detector pairs is given by

$$R_{\text{singles}} = 2E\epsilon g e^{-\mu l/2}, \qquad (16)$$

where $l/2$ appears in the equation to account for the distance traveled by a single photon in the absorbing material and the factor 2 accounts for each photon produced by each emission in E. Since g enters into Eq. (15) twice, R_{random} decreases as $1/r^4$ for a detector pair. However, the total geometric efficiency decreases as $1/r$ for a circumferential array of detectors, and thus the total random counting rate decreases as $1/r^2$.

2.6.3 Scatter Coincidence Events Compton scattering is the most probable interaction for annihilation photons in human tissue. Scattered coincidence events that occur when one or both photons originating from the same annihilation event scatter and are detected in coincidence result in erroneous positioning of annihilation events. Scatter also results in undetected events for photons scattered out of the sensitive volume between two detectors, thereby reducing (attenuating) the true coincidence counting rate. In addition to scatter in the object, in high-resolution PET employing narrow detectors, photons scatter from one detector into another, resulting in loss of spatial resolution by broadening of the radial PSF as it moves from the center of the FOV outward (Fig. 6) and in the tangential PSF mainly below the FWTM (Fig. 12). Because scatter depends on both the detector and the object geometry, it is impossible to represent scatter by a definite expression. However, the contribution of scatter to total coincidence counting rate for a source located midway between two coincidence detector pairs can be expressed as

$$R_{\text{scatter}} = KE\epsilon^2 g^2, \tag{17}$$

where K is a constant that accounts for the scatter absorption and angular distribution, detector and object geometry, and all other aspects of scatter.

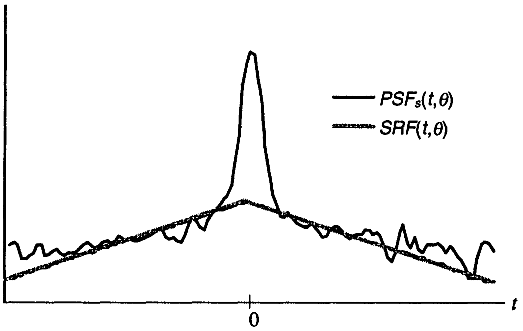

FIG. 12. Point-spread function in a scattering medium.

2.7 Dead Time

Dead time is the time during which the scanner cannot detect a new event because it is processing previous events. Electronic pulses produced by radiation detectors have a finite time duration or width as previously described. If an event occurs during the dead time of a preceding event, two pulses overlap. When this overlap occurs in the detector, the energy-sensitive information of the output pulse is distorted, and the second pulse introduces its own additional dead time. Shifted or overlapped detector pulse amplitudes may fall outside the energy discriminator window, resulting in the loss of both events and extending the dead time to subsequent events (*paralyzable* dead time). Dead time in some components of the electronic circuitry may result in the loss of only the second event by limiting the counting rate to one event per dead-time interval (*nonparalyzable* dead time). In PET, each detector is assigned to coincidence circuitries with many other detectors. The coincidence detection of three or more single events results in the rejection of all events (multiple-coincidence rejection circuitry) even though two of the events may have produced a true coincidence event. Because of dead-time losses, the observed counting rate is less than the true counting rate. Dead-time losses can be minimized by lowering the amount of radioactivity in the FOV or shortening the dead time through a reduction of pulse duration and integration times.

Dead time depends on the distribution of radioactivity in the scanner FOV, the difference between arrival times of the two annihilation photons at opposing detectors, the detector recovery time, the light integration time, the position discriminator circuitry, the energy discriminator circuitry, the coincidence-circuitry resolving time, and the time required to transfer data from the electronic circuitry to the computer.

2.8 Uniformity

A PET scanner consists of a large number of separate detectors with both electronic and other manufactured components. Inherently, the electronic pulses produced by each PMT in a detector block are different, as is the efficiency of each individual detector crystal. The efficiencies of detector pairs are also nonuniform on account of the geometry of the system. In ring-type PET systems, coincidence detector pairs toward the edge of the FOV have significantly smaller *solid angle* of detection (the angle between the axis of the detector and the line defined by the two detectors) than that at the FOV center. Not only does the effective width of the detector become narrower, but the probability of photon penetration into a neighboring detector becomes greater as a result of the reduced effective depth of the detector crystal. The reduction in surface area reduces the geometric efficiency [Eq. (12)], while the increased photon penetration in adjacent detectors results in asymmetry and broadening of the detector PSF. These effects become greater with increasing detector displacement from both the transverse and the axial FOV center. However, for a particular FOV, the magnitude of the nonuniformity decreases with increasing ring diameter. While proper system design and component optimization are effective in reducing nonuniformities, it is still necessary to balance the individual detector efficiency and the overall efficiency of the entire system to avoid serious artifacts and provide the best statistical quality in the images produced.

3. TOMOGRAPHIC RECONSTRUCTION

A tomogram is an image of a cross-sectional slice of an object. The formation, processing, and interpretation of PET images is similar to that found in other scientific disciplines (i.e., radio-astronomy or electron microscopy) and other medical applications [i.e., x-ray computed tomography (CT)]. In these applications, some property of the object is obtained by indirect measurements of radiations reflected from, transmitted through, or emitted by the object. For devices such as PET, these measurements are estimates of line integrals of the radionuclide distribution or the projection of the object's radionuclide distribution onto a particular angular view. These projections are related to the process under study but are often not readily suitable for interpretation. Tomographic image reconstruction of the acquired projection data facilitates the interpretation of the process or property under investigation through the formation of images representing the distribution of radionuclide concentration.

3.1 Radon Transform

The problem described as image reconstruction from projections, defined in this section, will be in the context of a function $f(x,y)$ on a cross-sectional plane through a three-dimensional object. The elements or members of a projection are the summation of all events along a line of response (LOR) between detector pairs called a Radon transform, given by the line integral

$$R(t,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) \times \delta(x\cos\theta + y\sin\theta - t)\,dx\,dy, \qquad (18)$$

where δ is the Dirac delta function.

As shown in Fig. 13, a parallel projection is formed by combining a set of line integrals for a constant θ. Similarly, a fan beam projection can be formed from all line integrals meeting at point t.

Thus, the inversion of all Radon transforms will produce an image of $f(x,y)$ (the object distribution).

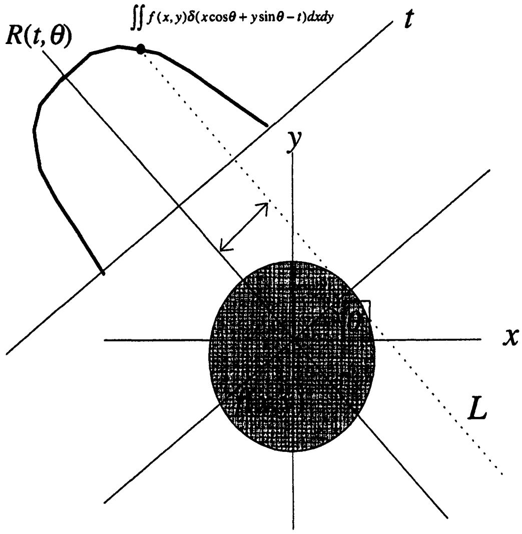

FIG. 13. Parallel projection obtained for object $f(x,y)$ at angle θ is the collection of all line integrals parallel to line $L(t,\theta)$ (i.e., lines having the same θ).

3.2 Fourier Reconstruction

The inversion of the Radon transform can be achieved in many ways. One such method is based on the *projection, Fourier,* or *central slice* theorem, which states that the coefficients of the two-dimensional (2D) Fourier transform of the object along a line through the origin at angle θ are the coefficients of the one-dimensional (1D) Fourier transform of the Radon transform (projection) obtained at angle θ. Mathematically, the theorem is described as follows.

The 2D Fourier transform $F(u,v)$ of the object distribution $f(x,y)$ is

$$F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-i2\pi(ux+vy)}dxdy, \qquad (19)$$

and the inverse Fourier Transform is given by

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(u,v)e^{i2\pi(ux+vy)}dudv, \qquad (20)$$

where u and v are dimensions of spatial frequency. If we introduce the polar coordinates

$$x = t\cos\theta,\quad y = t\sin\theta,\quad u = \omega\cos\theta,$$
$$v = \omega\sin\theta, \qquad (21)$$

Eq. (19) becomes

$$\begin{aligned} F(\omega\cos\theta,\omega\sin\theta) &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) \right.\\ &\quad \times \delta(x\cos\theta + y\sin\theta - t)\\ &\quad \left.\times e^{-i2\pi t\omega}dxdy\right]dt\\ &= \int_{-\infty}^{\infty} R(t,\theta)e^{-i2\pi t\omega}dt. \qquad (22)\end{aligned}$$

Thus, by obtaining the 1D Fourier transforms of the Radon transforms, forming parallel projections, interpolating them to appropriate locations to provide a 2D array of Fourier coefficients, and applying the inverse 2D Fourier transform to this coefficient array, an image of the object can be obtained. This method is called *Fourier reconstruction.*

3.3 Filtered Backprojection

Another, more common reconstruction method, called filtered backprojection, is derived from Eq. (20) where polar coordinates are again used, as follows:

$$\begin{aligned} f(x,y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(u,v)e^{i2\pi(ux+vy)}dudv,\\ &= \int_{0}^{\pi}\int_{0}^{\infty} |\omega|F(u,v)e^{i2\pi\omega(x\cos\theta+y\sin\theta)}d\omega d\theta,\\ &= \int_{0}^{\pi}\int_{0}^{\infty} |\omega|\left[\int_{-\infty}^{\infty} R(t,\theta)e^{-i2\pi t\omega}dt\right]\\ &\quad \times e^{i2\pi\omega(x\cos\theta+y\sin\theta)}d\omega d\theta, \qquad (23)\end{aligned}$$

where the integration can be thought of as a filtering followed by a backprojection of filtered projections, $R_f(t,\theta)$. Filtering is performed for each projection and is given by

$$R_f(t,\theta) = \int_{0}^{\infty} |\omega|\left[\int_{-\infty}^{\infty} R(\tau,\theta)e^{-i2\pi t\tau}d\tau\right]e^{i2\pi\omega t}d\omega \qquad (24)$$

and is described as the inverse 1D Fourier transform of the Fourier product, $|\omega|$ (known as the ramp filter because it increases 1 to 1 with increasing frequency), and the 1D Fourier transform of the Radon transform at angle θ. The backprojection is also performed at each angle and is given as

$$\begin{aligned} f(x,y) &= \int_{0}^{\pi} R_f(t,\theta)d\theta,\\ &= \int_{0}^{\pi} R_f(x\cos\theta + y\sin\theta,\theta)d\theta,\\ &= \int_{0}^{\pi}\int_{-\infty}^{\infty} R_f(t,\theta)\delta(x\cos\theta + y\sin\theta - t)\\ &\quad \times dtd\theta. \qquad (25)\end{aligned}$$

It is worth noting that the filtering portion of filtered backprojection can be performed equivalently in the spatial domain and that filtering after backprojection is also possible. These two latter techniques, however, are seldom used in practice.

The accuracy of Fourier and filtered backprojection reconstruction methods is affected by bandwidth limiting and interpolation. Bandwidth limiting is imposed because of the discrete sampling of the projections at finite intervals equal to the width of the detectors as discussed above. The main effect of bandwidth limiting is to limit spatial resolution. For PET scanners employing a circumferential array of detectors, interpolation of projections is required for the digital filtering prior to backprojection reconstruction. For Fourier reconstruction, interpolation is required to form the two-dimensional array of Fourier coefficients prior to applying the inverse 2D Fourier transform to obtain an image. It must also be emphasized that for computer implementation of these techniques, integration is replaced by discrete summation in the equations above. The degree of accuracy, with the exception of the loss in high-frequency information imposed by sampling and interpolation, is not further degraded by the use of discrete summation.

There are several problems with the Radon inversion approach. First, the assumptions that the data are noise free and that the detector response is spatially invariant are, obviously, untrue. Second, the resultant images contain negative numbers that, although are mathematically feasible, are not physically possible.

Since PET measurements are based on Poisson counting statistics, a reasonable approach to image reconstruction is to apply a more complete statistical model. Several reconstruction algorithms have been based on this approach. From a historical perspective (of PET), an iterative expectation-maximization (EM) approach is the oldest of these algorithms. Because each iteration requires computation time roughly equal to that of filtered backprojection and hundreds of iterations are often required, EM algorithms have not been widely accepted for routine clinical use. However, practical computation times are now being achieved by accelerating EM algorithms through the introduction of Bayesian "priors" (penalty functions) and noise-reduction techniques to reduce the convergence of the EM algorithm to the maximum-likelihood reconstruction.

3.4 Corrections

One of the main advantages of PET is its ability to provide a quantitative image of the radionuclide distribution in an object or subject. Often these images, in turn, are used for functional parameter estimation. The techniques used to estimate functional parameters are critically dependent upon the quantitative accuracy of the raw data. PET reconstruction techniques assume that each datum point $R(t,\theta)$ is a line integral that represents all activity between two coincidence detectors. Inherent in this assumption is that the spaces between detectors and their respective responses and efficiencies are the same. However, detector sensitivity and resolution are not uniform, and events are lost on account of attenuation and dead time. Random and scatter coincidence events are assigned to erroneous lines of response. All these factors reduce accuracy in PET. Many of these aspects have been addressed through improved system designs, materials, and electronics, but inevitably design tradeoffs lead to compromises in order to reduce costs and maximize detection efficiency and accuracy. Because of these hardware compromises, software correction schemes, incorporated in all PET scanners, are essential for quantitative PET.

3.4.1 Normalization The efficiencies of the many detector pairs in a PET system are nonuniform because of differences in the materials and electronics and the geometry of the system. While it is mechanically possible to adjust or tune the electronic components so that the PMT signals are optimized before they enter the energy discrimination, timing, and positioning circuitry, the individual detector efficiencies remain nonuniform. Normalization is the application of small correction factors to data obtained from individual detector pairs to produce consistent counting rates per unit activity in all detector combinations. These correction factors are produced from a data set obtained with an activity source that has minimum scattering material, low activity to avoid counting-rate effects (i.e., random coincidence events and dead-time losses), and a uniform distribution. The majority of circular PET scanners employ a small-diameter rotating-rod source that over time facilitates the detection of many events in all coincidence-detector pairs. Correction factors that normalize individual detector efficiencies relative to other detectors in the system are calculated and stored. Thereafter, the normalization correction factors are applied to acquired data prior to reconstruction.

Failure to correct for differences and changes in detector performance can result in serious artifacts and reduced statistical quality of images. For these reasons normalization should be performed at least monthly or as needed depending upon the stability of the particular system.

3.4.2 Random-Event Correction If left uncorrected, random or accidental coincidence events produce a diffuse background of activity leading to reduced image contrast. Randoms can be effectively reduced by decreasing the coincidence timing window [Eq. (15)]; however, if the timing window is too short, true coincidence events may be lost. In addition, larger system diameters have higher true-to-random event ratios on the basis of efficiency relationships derived above. Usually, random events are measured indirectly and subtracted from the total acquired coincidence events prior to reconstruction. Random events can be measured by delaying signals produced in one detector of a coincidence pair by several times the coincidence timing window and passing these with the nondelayed signals of the other detector through a second coincidence circuit. The delay introduced cannot produce any true coincidence events, and so the resulting coincidences in this second circuit can only be due to unrelated random events. Random events determined in this manner can easily be subtracted on line. A second method is to measure the singles counting rate of individual detectors, determine the random events using Eq. (15), and subtract them from the true coincidence events off line.

3.4.3 Scatter-Event Correction Scattered coincidence events, like random coincidences, are indistinguishable from true coincidences and result in misplaced events, adding a slowly spatially varying (low spatial frequency) background to an image. Photon scattering occurs within the object, in material surrounding the PET system, and in the detector itself. Unlike random coincidences, there is no technique for measuring scatter events electronically, and system design becomes very important in reducing photon scattering. In PET, scatter is reduced by the use of septa between rings to shield them from scattered photons originating within the object but outside a given ring. Septa are also placed at the ends of the scanner to remove scattered photons originating outside the scanner (e.g., body or walls). For 3D imaging techniques, the septa between rings are retracted, producing an increased detection of scattered coincidences. Within a given ring, photon scattering in detectors at the edge of the FOV (because of the large solid angle) reduces spatial resolution by broadening the FWTM of the PSF and results in a loss of image contrast. Reducing the solid angle, however, reduces the detection of true coincidence events, resulting in poorer sensitivity. Scattered events are also reduced by raising the low-energy discriminator since scattered photons have lower energy than unscattered photons. However, most scattered photons have energy near that of unscattered photons. Since the energy resolution of BGO detectors, employed in most PET scanners, is 50 to 100 keV, raising the energy threshold is inadequate to remove the majority of scattered events. The scatter fraction, the ratio of scattered events to total coincidences, is reduced with increased ring diameter. The degradation of image contrast, resolution, and quantitative accuracy due to scattered photons depends on the distribution of radioactivity, the PET system design and operation, and the scattering media. Since there is no exact method for the correction of scattered events in software, the approach to minimize the effects of scattered photons is initially through system design and then by careful study design and image analysis.

There are several methods for estimating and removing scattered events (Bergstrom *et al.*, 1983; Logan and Bernstein, 1983; Bendiriem *et al.*, 1987; Bentourkia *et al.*, 1995). The majority of these do not fully correct for the spatial dependences of scattered photons (subtle, high-frequency errors) but are effective in reducing the majority of large (gross, low-frequency) errors. Based on Eq. (13), the observed projection data $P(t,\theta)$ is the sum of true, scattered, and random events. After random events are removed, $P(t,\theta)$ is described as

$$P(t,\theta) = E(t,\theta) + S(t,\theta), \tag{26}$$

where $E(t,\theta)$ and $S(t,\theta)$ are respectively the contributions of true and scattered events. Scattered events, $S(t,\theta)$, are most often modeled as a convolution of $E(t,\theta)$ with a scatter response function $\mathrm{SRF}(t,\theta)$:

$$S(t,\theta) = E(t,\theta)^{*}\mathrm{SRF}(t,\theta), \tag{27}$$

where the asterisk denotes convolution. Substituting Eq. (27) into Eq. (26) gives

$$\begin{aligned} P(t,\theta) &= E(t,\theta) + E(t,\theta)*\mathrm{SRF}(t,\theta), \\ &= E(t,\theta)*[\delta(t,\theta) + \mathrm{SRF}(t,\theta)], \end{aligned} \tag{28}$$

where $\delta(t,\theta)$ is the Dirac delta function. Substituting Eq. (28) into Eq. (27) and rearranging; we obtain

$$E(t,\theta) = P(t,\theta) - E(t,\theta)*\mathrm{SRF}(t,\theta). \tag{29}$$

This convolution-subtraction approach requires knowledge of $E(t,\theta)$, which obviously is that which we desire to know. It has been demonstrated (Bergstrom *et al.*, 1983) that Eq. (29) can be written as

$$E(t,\theta) = P(t,\theta) - P(t,\theta)*\mathrm{SRF}'(t,\theta), \tag{30}$$

where $\mathrm{SRF}'(t,\theta)$ is determined experimentally for a specific object producing $P(t,\theta)$. The standard approach is to estimate the $\mathrm{SRF}'(t,\theta)$ of the system from a line or point source placed at different locations in a scattering medium, typically a 20-cm cylindrical phantom. Since $E(t,\theta)$ for complex objects is considered to be the superpositioning of scaled system-response or point-spread functions $\mathrm{PSF}(t,\theta)$, a projection obtained for a point source in a scattering medium, $\mathrm{PSF}_s(t,\theta)$, can be used to estimate $\mathrm{SRF}'(t,\theta)$ by fitting $\mathrm{SRF}'(t,\theta)$ expressed as two monoexponential functions to data on both sides of the $\mathrm{PSF}_s(t,\theta)$ peak extrapolated to intersect inside the peak at the peak position, as shown in Fig. 12.

The standard approach is to apply Eq. (30) to determine $E(t,\theta)$. Another, less common, approach to correct $P(t,\theta)$ is by the repositioning or deconvolution of scattered events. On replacing $\delta(t,\theta)$ with the system $\mathrm{PSF}(t,\theta)$ obtained in air (without scatter), Eq. (28) becomes

$$P(t,\theta) = E(t,\theta)*\mathrm{PSF}_s(t,\theta), \tag{31}$$

where $\mathrm{PSF}_s(t,\theta)$ is the point-spread function obtained in a scattering medium. Using Fourier transforms (FTs) and the convolution theorem, and solving for $E(t,\theta)$, we have

$$E(t,\theta) = \mathrm{FT}^{-1}\left[\frac{\mathrm{FT}[P(t,\theta)]}{\mathrm{FT}[\mathrm{PSF}_s(t,\theta)]}\right]. \tag{32}$$

The accuracy of scatter correction methods are measured by scanning a 20-cm-diameter cylindrical phantom containing a uniform water distribution of radioactivity, *hot,* in which a second cylinder containing only water, *cold,* is offset from the center. Measurements of radioactivity are obtained for the hot and cold regions of an image reconstructed with all corrections. The percent scatter fraction (%SF) determined as

$$\%\mathrm{SF} = 100 \times \frac{\mathrm{cold}}{\mathrm{hot}} \tag{33}$$

should be 0% for scatter correction to be considered to have removed the majority of scattered events. Other measurements such as the FWHM, FWTM, and the full width at hundredth maximum of the $\mathrm{PSF}_s(t,\theta)$ and measurements of other more complex phantoms can help determine the effectiveness of the scatter correction with respect to the spatial variation of scatter. Scatter correction is usually performed after normalization and random subtraction and must be performed prior to attenuation correction and image reconstruction.

3.4.4 Attenuation Correction One of the attractive features of PET is that the attenuation of annihilation photons is independent of position between coincidence detectors [Eq. (9)]. If the attenuating tissue is homogenous, the detected annihilation photon intensity for true coincidences, $E(t,\theta)$, is given as

$$E(t,\theta) = A(t,\theta)e^{-\mu l}, \tag{34}$$

where $A(t,\theta)$ is the annihilation photon activity emitted in the direction of the coincidence detectors or the line integral along $L(t,\theta)$ in the Radon domain (see Fig. 13),

$$\begin{aligned} A(t,\theta) = T\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(x,y) \\ \times\ \delta(x\cos\theta + y\sin\theta - t)dxdy \end{aligned} \tag{35}$$

for the spatial distribution of positron radioactivity, $A(x,y)$, sampled over time T. The attenuation correction of $E(t,\theta)$ is achieved by solving Eq. (34) for $A(t,\theta)$:

$$A(t,\theta) = E(t,\theta)e^{\mu l}. \tag{36}$$

The attenuation correction factor, $\mathrm{ACF}(t,\theta)$, is

$$\mathrm{ACF}(t,\theta) = e^{-\mu l}, \tag{37}$$

where l is the total length or distance traveled by both photons through the attenuating media having linear attenuation coefficient μ. The assumption of homogenous tissue and, therefore, constant μ is reasonably valid for areas such as the brain and lower abdomen that have relatively constant attenuation coefficients. The $\mathrm{ACF}(t,\theta)$ given in Eq. (37) can be calculated by obtaining the path length l from the intersection of $L(t,\theta)$ and the perimeter of $A(x,y)$ provided the shape of $A(x,y)$ is known. The shape of $A(x,y)$ can be determined as a geometric (ellipse) or irregular shape fitted around the cross-sectional contour of an emission image of the body.

For areas of inhomogenous tissue distributions (i.e., areas containing significant amounts of air and bone), a significant error in quantifying attenuation can occur and the $\mathrm{ACF}(t,\theta)$ must be determined from the distribution of attenuation coefficients $\mu(t,\theta)$ along $L(t,\theta)$ as

$$\mathrm{ACF}(t,\theta) = \exp\left[\int_{L(t,\theta)} \mu(t,\theta)dl\right]. \tag{38}$$

Although accurately described, the above calculation of the $\mathrm{ACF}(t,\theta)$ is difficult to perform accurately because the varying tissue shape, type, and thickness are usually crudely approximated on the basis of a single additional measurement. A more accurate attenuation correction can, however, be obtained from two additional measurements. Because of the independence of source position along $L(t,\theta)$ in PET (i.e., between two coincidence detectors), a positron source external to the body that transmits annihilation photons through the body has the same probability of attenuation as an internal source that emits annihilation photons. So, without any attenuation material in the FOV, and neglecting the absorption of photons in air, an annihilation-photon source rotated around the PET FOV that transmits an intensity $I_0(t,\theta)$ along any projection $L(t,\theta)$ produces a blank scan $B(t,\theta)$ given as

$$B(t,\theta) = I_0(t,\theta). \tag{39}$$

When an attenuating object is placed in the FOV, a transmission scan $T(t,\theta)$ of the object results with the following relationship:

$$T(t,\theta) = I_0(t,\theta)\exp\left[-\int_{L(t,\theta)} \mu(t,\theta)dl\right]. \tag{40}$$

The attenuation correction factor, $\mathrm{ACF}(t,\theta)$, is easily determined by the division of Eq. (39) by Eq. (40):

$$\begin{aligned}\mathrm{ACF}(t,\theta) &= \frac{B(t,\theta)}{T(t,\theta)} = \frac{I_0(t,\theta)}{I_0(t,\theta)\exp\left[-\int_{L(t,\theta)} \mu(t,\theta)dl\right]}\\ &= \exp\left[\int_{L(t,\theta)} \mu(t,\theta)dl\right]\end{aligned} \tag{41}$$

on the assumption that both blank and transmission scans were obtained for the same duration, using the same positron source intensity $I_0(t,\theta)$.

Attenuation correction factors in PET can thus be obtained by either direct measurement, calculated approximations, or a combination of calculated and measured methods. As described by Eq. (36), the attenuation correction is performed prior to reconstruction in the projection space or Radon domain.

3.4.5 Dead-Time Correction In PET, contributions to system dead time vary with counting rate and source distribution. In PET systems employing block detectors, individual detector dead-time losses due to pulse pileup, detector recovery time, energy discrimination, and position encoding are affected by the singles counting rate of the entire block. In addition to the singles dead time, there are losses due to the overlap of pulses in the coincidence circuitry. Because of the complexity of various designs, the spatially varying radionuclide distributions of different objects, and the different types of studies (i.e., blank, transmission, and emission), each PET scanner configuration requires its own individual dead-time correction scheme.

Average dead-time loss as a function of counting rate for a particular configuration can be determined by obtaining a series of measurements of the particular radionuclide distribution. Using the physical half-life of the radionuclide, the true higher counting rates can be calculated from low-counting-rate measurements that fall with the half-life. The difference between the calculated and the measured counting rates is an estimate of the dead-time losses.

The accuracy of the dead-time correction is often determined by obtaining a series of measurements for a particular radionuclide distribution as a function of time. Typically a cylindrical phantom, uniformly filled with a known concentration of a short-lived radioisotope, is scanned repeatedly over several half-lives until dead time is negligible. To determine the range for which the dead-time correction can be considered adequate, dead-time corrected measurements are compared with expected measurements obtained by using the known radionuclide concentration and the physical half-life of the radioisotope. The highest dead-time corrected measurement of radionuclide concentration that is within ±5% of the true radionuclide concentration in a cylindrical phantom and the counting rate at which dead-time losses reach 50% are often used as a measure of PET system performance. However, the radionuclide distribution in a cylindrical phantom may not represent that found in an *in vivo* PET study, and more realistic configurations should be used to assess the accuracy of the dead-time corrections. Nevertheless, these performance parameters are useful in judging the effect of dead time on arbitrary radionuclide distributions and are important when determining the amount of radioactivity that can be administered in studies in which the activity in the scanner FOV is expected to be high.

3.4.6 Calibration In PET, local concentration of positron emitter is the quantity measured. Since only a fraction of the radiation emitted is detected, calibration of the efficiency of the PET system is required to produce measurements in standard terms of radioactivity concentration (e.g., $\text{nCi}\cdot\text{mL}^{-1}$ or $\text{Bq}\cdot\text{mL}^{-1}$). Calibration requires a PET measurement of a known concentration of positron emitter that closely matches the radionuclide distribution normally scanned. For brain studies, a 20-cm-diameter cylindrical phantom can be used. The phantom is filled uniformly with a concentration of radioactivity that is determined independently of the PET system with a calibrated radiation detector such as a well counter or dose calibrator. Since reconstructed image pixel values are based on detected coincidence counts over an interval of time, the efficiency is given in units of activity concentration [i.e., PET $\text{counts}\cdot\text{s}^{-1}\cdot\text{pixel}^{-1}\cdot(\text{nCi}\cdot\text{mL}^{-1})^{-1}$]. The calibration factors are used to scale image pixel value to units of radioactivity concentration.

4. IMAGE CHARACTERIZATION

Positron-emission tomography provides a quantitative representation of the distribution of a radiopharmaceutical in the form of a digital image. Each image element or *pixel* value is an estimate of the radionuclide concentration found in the corresponding location in the object. In PET, information can be obtained through the subjective and nonquantitative visual examination of images, as in other medical-imaging modalities, and also through the absolute quantification of region-of-interest (ROI) measurements of local radionuclide concentration. In order to characterize the capability of a PET image to depict true measures of radionuclide concentration, we must have estimates of precision and accuracy (Budinger *et al.*, 1984; Hoffman *et al.*, 1984).

For simplicity and ease of illustration, the examples following will be depicted in one dimension, while the mathematical equations will be presented in two dimensions.

4.1 Accuracy

In PET, *accuracy* is a measure of the difference between the image and the true radionuclide distribution it estimates. Since in *in vivo* experiments the truth is unknown, accuracy is difficult to measure and thus must be determined with images of test objects for which the radionuclide distribution is known independent of the PET scanner. It is important that phantoms (test objects) used to evaluate accuracy are similar to the distributions seen in *in vivo* imaging. It is well known that radionuclide concentration measurements obtained for objects whose dimension (in any direction) is less than twice the spatial resolution of the PET scanner are dependent on the object's size and position in the PET scanner.

The primary sources of inaccuracy in PET are spatial resolution and partial-volume effects. The quantitative accuracy is determined by the spatial resolution in that a system with high resolution will more accurately place events into their position of true origin, while a system with low resolution will spread events outside their point of origin. Other sources of inaccuracy include photon attenuation, random and scatter coincidence events, dead time, and system calibration. The most common measure of accuracy in PET is the recovery coefficient RC, given as

$$\mathrm{RC} = \frac{\text{Measured}}{\text{True}}, \tag{42}$$

where Measured is the image measurement (usually the maximum ROI pixel value) and True is the actual radionuclide concentration of a radioactive object (Hoffman *et al.*, 1979). RC is usually reported as a function of object diameter and FOV position for the maximum pixel obtained from an image of a radioactive (hot) spheres in a nonradioactive background and is often referred to as the hot-spot recovery coefficient, HSRC. For objects such as radioactive (hot) spheres in a background of activity, a contrast recovery coefficient CRC, defined as

$$\mathrm{CRC} = \frac{\text{Measured} - \text{Background}}{\text{True} - \text{Background}}, \tag{43}$$

is used where Measured and True are the same as given above and Background is the activity concentration surrounding the hot object (Kessler *et al.*, 1984). For nonradioactive (cold) objects in a background of activity, a cold-spot recovery coefficient CSRC, defined as

$$\mathrm{CSRC} = \frac{\text{Minimum}}{\text{Background}}, \tag{44}$$

is used, where Minimum is the smallest pixel value obtained from an image of the cold spot.

The RCs calculated above use the minimum and maximum values from a single pixel in a given region of a cold and hot object, respectively. In routine ROI analysis, the average value obtained within the ROI is used to reduce the effect of image noise on a measurement. However, because of the partial-volume effects of spatial resolution and the size of the object, too large a ROI can result in loss of recovery by reducing the measured pixel value [i.e., Measured in the numerator of Eq. (42)]. The distance by which an object is displaced from the axial center of a transverse plane also has the effect of reducing recovery.

4.2 Precision

Precision is the ability to reproduce a measurement. The precision of a PET image is affected primarily by the statistical (Poisson) variations inherent in nuclear decay and instrumentational variations in the detection of photons. These statistical variations are propagated through the image reconstruction process and can be amplified by the reconstruction filter. Ideally, the precision or random error in reconstructed images could be obtained from the variance of repeated measurements of the same test object. In practice, obtaining measurement replicas that differ only by random noise is difficult because of nuclear decay where the variation in a given measurements increases with decreasing radioactivity. Thus, precision measurements are often obtained from a single ROI (local) measurement of test object (Alpert *et al.*, 1982) as the percentage coefficient-of-variation, %CV, defined as

$$\%\mathrm{CV} = (100\%)\sigma/\mu, \tag{45}$$

where σ and μ are the standard deviation and mean radionuclide ROI concentration.

5. IMAGE QUALITY

The usefulness of PET images for diagnostic interpretation and quantitative assessment is limited primarily by finite spatial resolution and statistical noise. Improvements in instrumentation and radiopharmaceuticals have led to improvements in image quality. Image quality is difficult to define because it is highly subjective and heavily dependent on the given application. The physical factors that affect an image can serve as an objective basis for evaluating and maintaining the control of image quality.

Contrast, resolution, and *noise* describe physical aspects of an image that are not independent of one another in that improvement in one may cause serious deterioration in the other.

In PET, tomographic reconstruction preserves high object contrast within the limitations of spatial resolution governed by the detector and the reconstruction filter, photon scattering, and attenuation. Although quantitative accuracy is very important for the extraction of functional information from a PET image, much of the diagnostic information is obtained from the visual detection of physiological abnormalities. The ability to identify small changes in the patterns of radionuclide distribution reliably depends also upon the statistical noise (error) inherent in the image data.

5.1 Contrast

Contrast is the difference in image intensity between areas in the object having different radionuclide concentrations. In general, an important part of PET imaging is designing radiopharmaceuticals to have the highest uptake in a specific target structure of an object. Proper radiopharmaceutical design is crucial for maximizing object contrast and is beyond the scope of this text. Instead, we will discuss the physical and instrumentation aspects of image contrast. Image contrast C is defined as

$$C = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}, \qquad (46)$$

where $I_{\max}$ and $I_{\min}$ are the maximum and minimum intensities (radionuclide concentrations) in adjacent areas of an object. C ranges between 0 (no contrast) and 1 (maximum contrast). Image contrast is affected by the presence of background radioactivity, as shown in Fig. 14. In the presence of a constant background intensity I_b, Eq. (46) can be rewritten as

$$\begin{aligned} C &= \frac{(I_{\max} + I_b) - (I_{\min} + I_b)}{(I_{\max} + I_b) + (I_{\min} + I_b)}, \\ &= \frac{I_{\max} - I_{\min}}{(I_{\max} + I_{\min}) + 2I_b}, \end{aligned} \qquad (47)$$

where the maximum contrast (i.e., $I_{\min} = 0$) is decreased by the presence of $2I_b$ in the denominator.

In PET, scatter and random coincidences have the effect of adding an almost constant background if not corrected, substantially reducing image contrast and affecting image accuracy, as demonstrated in Fig. 14.

5.2 Resolution

Image resolution is the degree of discernible detail in an image. Image resolution, which depends on the sampling of the object (i.e., detector size and geometry), is also affected by the cutoff frequency of the reconstruction filter, the dimensions of the image pixels, and object motion. As mentioned in Sec. 2.3, the most common measure of spatial resolution is the FWHM of the PSF obtained from an image acquired with the best possible sampling and reconstructed with a

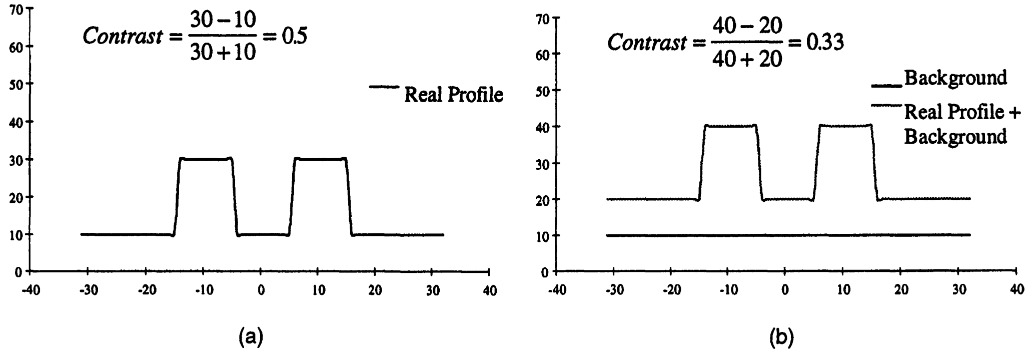

FIG. 14. The effect of adding a constant background on contrast.

filter with the highest cutoff frequency into an image display matrix having the finest element dimension.

Another, more detailed measure of resolution is the modulation transfer function (MTF), or the frequency response of the PET system. The MTF, obtained as the Fourier transform of the PSF (Fig. 15), quantitatively describes how well coarse or gradually varying changes and fine detail including small object recovery and edges of the object are reproduced in an image. An MTF value of unity indicates that a spatial frequency component of an object is accurately represented in the image.

Figure 15 demonstrates that both high- and low-resolution MTFs have values near unity at low spatial frequencies and roll off to zero at higher frequencies. The higher-resolution MTF decreases to zero much more slowly, showing that fine detail is preserved to a greater extent than in a system with a lower-resolution MTF. MTF values near zero indicate that the corresponding frequency in the image is more likely to be composed of statistical variation (noise) than true signal.

Illustrated in Fig. 16 is the reduction in image contrast and edge sharpness due to the smearing of data from one region into an adjacent region in the image. This effect of finite spatial resolution on image contrast is characteristic of all PET scanners and is equivalent to a convolution of the object distribution with a smoothing function (i.e., the PSF). However, PET systems with high MTF values that extend to higher spatial frequencies produce images with higher contrast and sharper detail than systems having lower MTF values.

5.3 Statistical Noise

In PET, noise is the statistical error in an image that is related to the number of coincidence events detected. For individual pixel values in tomographic images, noise is not just a simple matter of counting statistics, as described in Sec. 5.4. For now, we recognize that the smoothness of an image is determined by the number of true coincidence events detected. In the previous sections, we demonstrated that image contrast is a function of both object contrast and spatial resolution in statistically smooth images (i.e., object contrast is reduced by the modulation transfer function and the reconstruction filter). In this section, we demonstrate that the number of detected events required to preserve object contrast is reduced with improved resolution.

The low counts in the profiles of Fig. 17 were obtained by superimposing random, count-dependent noise on the profile plots of Fig. 16, and the higher-count profiles were obtained by scaling the low-count plots of Fig. 16. The calculation of random noise is the same. The high and low levels of noise correspond to fewer and greater number of

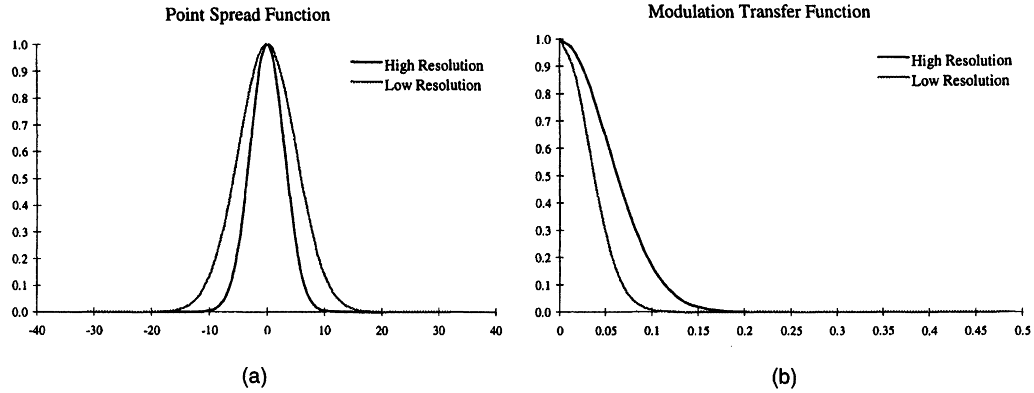

(a) (b)

FIG. 15. (a) PSFs and their (b) corresponding MTFs. A poorer-resolution system will spread or blur acquired events to a greater degree than a high-resolution system. The FWHM of the PSF is a measure of this spread. MTF values quantitate how well a spatial frequency is reproduced in an image. Note that the highest frequency that can be preserved in a PET scan is 1/(2 × sampling distance), and the highest frequency that can be represented by an image is 1/(2 × pixel dimension) or 0.5 cycle per pixel.

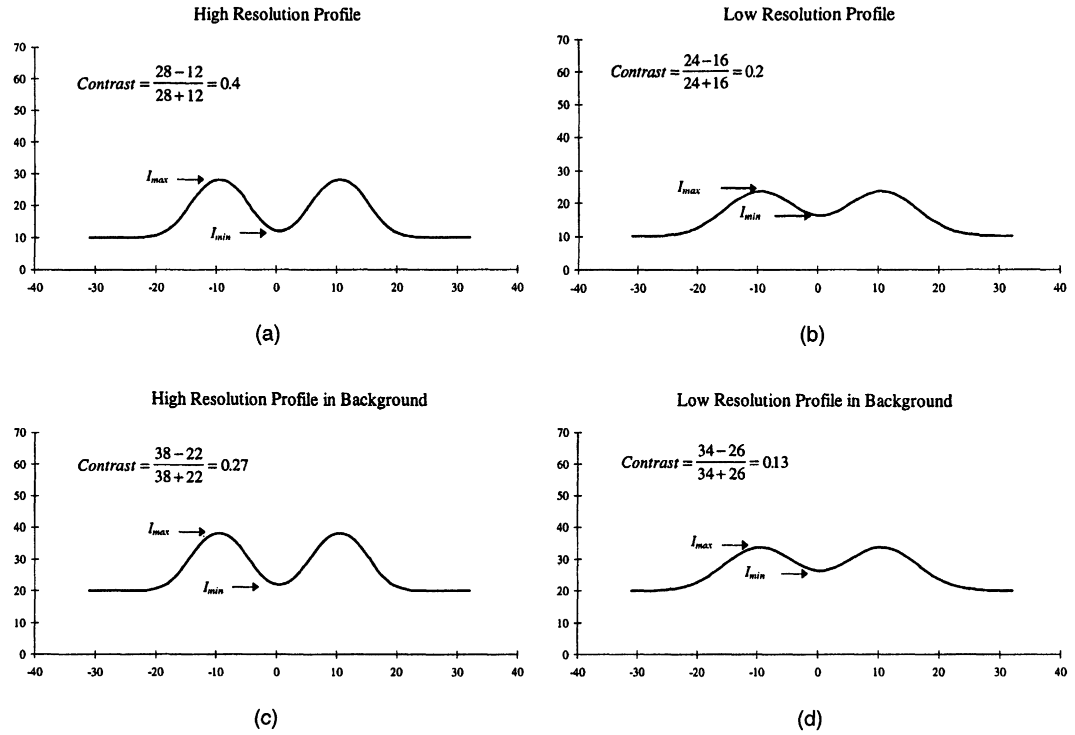

FIG. 16. Effect of reduced spatial resolution on contrast (a), (b) in an object and (c), (d) for the same object in background.

total counts, respectively. For profiles that have higher counts (lower noise) and higher spatial resolution, Figs. 17(a) and 17(b), contrast is preserved. For high counts and low spatial resolution, Figs. 17(c) and 17(d), contrast, as expected, is reduced primarily by spatial resolution. By virtue of higher spatial resolution, the activity is more correctly placed in Figs. 17(e) and 17(f), and contrast is better preserved, in comparison to Figs. 17(g) and 17(h), where contrast is virtually nonexistent because of high noise and low resolution.

Statistical noise is related to the number of coincidence events acquired, which is determined by the activity of the object, the sensitivity of the PET system, and the measurement (acquisition) time.

We have addressed the role of resolution, noise, and contrast and their effects on image quality through computer simulations in this and previous sections. We now introduce as an objective measure of image quality the root mean square error e_{rms}, which is given by

$$e_{\mathrm{rms}} = \left[\frac{1}{M}\sum_{i=0}^{M-1}[I_i - O_i]^2\right]^{1/2}, \tag{48}$$

where M is the number of image measurements (pixels), I is the observed image radionuclide distribution degraded by spatial resolution and noise, and O is the actual object radionuclide distribution. A related objective criterion is the root mean square signal-to-noise ratio ($\mathrm{S/N_{rms}}$) of the image data, which is given by

$$\mathrm{S/N_{rms}} = \left(\sum_{i=0}^{M-1} I_i^2\right)\left(\sum_{i=0}^{M-1}[I_i - O_i]^2\right)^{-1}, \tag{49}$$

where I and O are the image and object distributions, respectively. Table 3 illustrates the significance of the loss of quantitative measures of image quality due to degradation effects of resolution, statistical noise, and background. e_{rms} and $\mathrm{S/N_{rms}}$ are measures of global or overall statistical precision and accuracy, respectively, and contrast is a measure of local image quality.

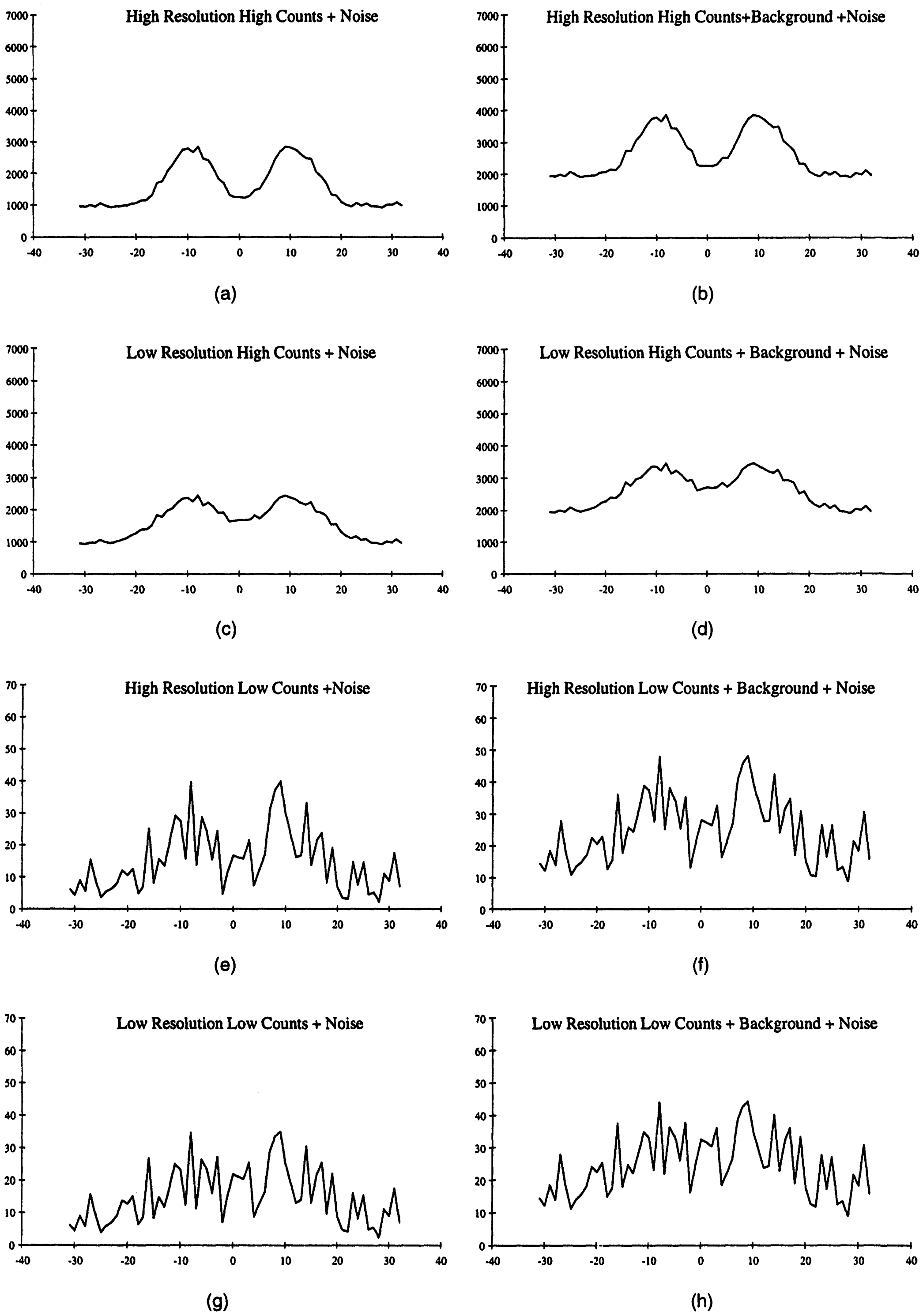

FIG. 17. The combined effects of statistical noise, background, and spatial resolution on an object profile.

Table 3. Measured parameters of objective image quality for the computer simulations given in Figs. 14, 16, and 17. Contrast measure C_0 is obtained from noise-free simulations (Figs. 14 and 16) for the given resolution and background.

Infinite resolution	e_{rms}	S/N$_{\mathrm{rms}}$	N	Contrast	S/N = $C_0N^{1/2}$
High counts (no noise)	1852.12	∞	104 000	0.50	161.25
High counts + BG (no noise)	2756.04	∞	168 000	0.33	136.63
High counts + noise	1843.69	817.69	103 255	0.55	160.67
High counts + BG + noise	2743.41	1131.74	166 998	0.37	136.22
Low counts (no noise)	0.00	∞	1 040	0.50	16.12
Low counts + BG (no noise)	0.00	∞	1 680	0.33	13.66
Low counts + noise	6.51	8.41	965	0.91	15.54
Low counts + BG + noise	7.13	14.88	1 587	0.70	13.28
High resolution					
High counts (no noise)	1737.77	17.50	104 000	0.40	129.43
High counts + BG (no noise)	2680.53	41.62	168 000	0.27	109.67
High counts + noise	1729.08	16.92	103 244	0.39	128.96
High counts + BG + noise	2667.66	39.72	166 991	0.26	109.34
Low counts (no noise)	4.20	17.50	1 040	0.40	12.94
Low counts + BG (no noise)	4.20	41.62	1 680	0.27	10.97
Low counts + noise	7.80	5.21	964	0.78	12.46
Low counts + BG + noise	8.26	10.49	1 587	0.57	10.66
Low resolution					
High counts (no noise)	1681.75	8.22	104 000	0.18	59.17
High counts + BG (no noise)	2644.56	20.33	168 000	0.12	50.13
High counts + noise	1673.16	8.03	103 244	0.20	58.95
High counts + BG + noise	2631.77	19.71	166 993	0.14	49.98
Low counts (no noise)	5.93	8.22	1 040	0.18	5.92
Low counts + BG (no noise)	5.93	20.33	1 680	0.12	5.01
Low counts + noise	8.87	3.79	964	0.67	5.70
Low counts + BG + noise	9.28	8.08	1 586	0.46	4.87

5.4 Signal-to-Noise Ratio in PET

The signal, in PET, is the detected true coincidence events that form an image, and the noise is anything that deviates from the signal (i.e., random and scatter coincidences and statistical variation associated with the detection of photons arising from nuclear decay). From Eqs. (14)–(17), we know that the geometric and intrinsic efficiencies of the detectors determine the number of true coincidence events and the contribution to total coincidences by random and scatter coincidence events. On the basis of these equations, the improvements in total geometric efficiency can be achieved by reducing the system diameter and increasing the number of lines of response (2D vs 3D) by removing septa. However, either of these considerations has the effect of increasing randoms and scatter to a greater extent than trues; thus, the ratio of trues to random plus scatter is increased. To this extent the noise-equivalent counting rate,

$$\mathrm{NEC} = \frac{R_{\mathrm{true}}}{1 + R_{\mathrm{scatter}}/R_{\mathrm{true}} + R_{\mathrm{random}}/R_{\mathrm{true}}}, \tag{50}$$

which defines the effective true coincidence counting rate by taking into account random and scatter coincidences, can be used to assess the image quality. This measurement is not sensitive to local image variations and is only considered a global measurement, which is usually obtained for a uniform cylinder at different levels of activity.

When employing a uniform cylinder to study the signal-to-noise ratio, S/N can be expressed as

$$\mathrm{S/N} = N^{1/2}/1.2n^{3/4}, \tag{51}$$

where N is the number of counts in the image and n is the number of resolution cells (pixels) in the image containing activity (Budinger *et al.*, 1978). For the case of the uniform phantom, use of the above expression masks the effect of resolution on S/N since there is little high-frequency structure in the object. In fact, using Eq. (51) to study S/N leads (incorrectly) to the conclusion that a twofold increase in spatial resolution requires an eightfold increase in the total number of image counts to realize the same S/N when structured objects are considered (Muehllehner *et al.*, 1985; Hoffman and Phelps, 1986). As demonstrated in Table 3, when total noise e_{rms} is unchanged, improved resolution increases image contrast (local signal) and higher resolution also provides improved contrast and $\mathrm{S/N_{rms}}$, when fewer counts are present. Thus, the conclusions that apply to a large uniform distribution do not necessarily apply to structured objects. A general relationship for S/N that incorporates object contrast as a measure of signal and thereby accounts for the spatial resolution of the imaging device and uses the fact that total image noise is proportional to $N^{1/2}$ [Eq. (51)] and independent of imaging device (Muehllehner *et al.*, 1985) is given by

$$\mathrm{S/N} \propto C_0 N^{1/2}, \tag{52}$$

where N again is the total image counts and C_0 is contrast as given in Eq. (46) in the absence of noise. S/N, thus derived, is given in Table 3.

With respect to tomographic imaging and conventional filtered backprojection reconstruction, we summarize the relationship between signal, noise, S/N, resolution, and contrast as follows: In tomographic image reconstruction, true coincidence events measured through many pixels are superimposed onto a single image pixel by the reconstruction process. In the absence of noise, the signal contributed to a particular pixel's value from all other pixels is eliminated by means of the ramp reconstruction filter. However, statistical errors in detected coincidence events, instrumentational errors, and the extensive data processing involved in the corrections for scatter, randoms, and attenuation in projections are often amplified by the ramp filter and are propagated to the pixels of the resultant image during reconstruction. A complete derivation of the noise contribution of each coincidence measurement is well beyond the scope of this text, and the interested reader is referred to the suggested readings. The number of true coincidence events acquired (signal) and therefore the resultant image statistical (Poisson) noise are directly related to the system's detection efficiency. The spatial resolution and, therefore, the local signal are related to the width and the sampling distances of the system's detectors. The local signal is also determined by the object's contrast and its position in the detector system.

Improvements in image noise thus translate into a requirement for higher-sensitivity PET scanners. However, through the use of smaller-diameter systems, larger detectors, or the retraction of septa, sensitivity is increased, but only at the expense of reduced spatial resolution. In addition to increasing the total number of true coincidences, these design changes also increase the number of scatter and random coincidence events (background) and further degrade spatial resolution and resolution uniformity (Hoffman and Phelps, 1986). In high-resolution PET, only increases in intrinsic detector efficiency can increase the total number of true coincidence events relative to the number of scatter and random coincidences. This is the main reason why high detector stopping power has become one of the most attractive features in the development of modern high-resolution PET scanners.

The signal and noise characteristics in high-resolution PET can be best understood in terms of the spatial frequency response of the detectors and the spatial frequency amplification or gain of the reconstruction filter in relation to statistical noise. As detector resolution is improved, more high-frequency signal (if present) is preserved to a greater extent, and at these high frequencies, S/N is increased. During reconstruction, the reconstruction filter amplifies the amplitude of a spatial frequency component of the data according to its coefficient. The spatial frequency components of the data are eliminated above a cutoff frequency (often well below the Nyquist sampling rate). In the case where very little or no signal is present beyond the cutoff frequency of the reconstruction filter, high-frequency noise is adequately suppressed. However, when high-frequency signal is present, as is the usual case in a highly structured object such as the brain or heart, the use of a higher reconstruction-filter cutoff frequency can increase signal but also increase noise. Thus, the ability to change a reconstruction filter's cutoff allows for improved S/N usually at the expense of spatial resolution. In other words, one can produce the highest resolution by selectively amplifying the high-frequency signal and reduce noise by suppressing high frequencies where noise dominates.

6. DIAGNOSTIC QUALITY

In a clinical setting, PET images are often interpreted subjectively (i.e., "by eye") in diagnostic decision making. The visual recognition (detection) of a pattern, structure, or defect (object) is dependent on the observer's ability to delineate or distinguish one object from another. The high-frequency information of the object necessary for border definition is dependent on the spatial resolution of the system (i.e., blurring and spillover effects) and is contaminated with excessive high-frequency noise. Noise has the effect of masking the presence of objects (producing false-negative interpretations) and also can lead to the detection of apparent "objects" that are not actually present (false positives). As described above, physical measures of image quality can be useful for comparing different imaging system or techniques and establishing quality-control procedures and can provide useful estimates of minimum detectable object size and contrast. However, in most cases, the diagnostic accuracy of detection is best estimated by experiments using human observers whose interpretations are verified by actual clinical studies. The time required for a physician to extract useful information from an image and the accuracy of his decision can be improved with image-enhancement techniques. Diagnostic significance and quality of an enhancement technique can only be based on the interpretation of many actual or phantom studies similar to those found in a clinical situation.

6.1 Detectability

Numerous studies have shown that

1. for reliable visual detection of an object, at a given image contrast, a sufficient number of true coincidence events must be acquired to reduce e_{rms} and increase S/N;
2. high spatial resolution improves visual discrimination by removing the blurring between adjacent structures having different radionuclide concentrations; and
3. high object contrast obtained through proper radiopharmaceutical design improves the recognition of anatomical structures having differing physiological functions.

Although contrast, resolution, e_{rms}, and S/N are objective criteria of image quality, the true object distribution is often unknown prior to clinical interpretation (except in the use of phantoms for evaluation of a system or technique). Consequently, measuring an image's diagnostic quality by the subjective evaluations of human observers is appropriate. This is usually accomplished by showing a set of images to a cross section of viewers (receivers) and averaging their evaluations (operating characteristics). These receiver operating characteristic (ROC) evaluations include determination of the presence or absence of a disease, structure, or defect, and its location, and uses a rating scale that represents the observer's confidence in his decision. In clinical ROC studies, the absolute truth is obtained from actual clinical studies. Side-by-side comparisons of images and ranking them with a scale from best to worst can also be useful when evaluating imaging systems or techniques. In either case, subjective evaluation of the diagnostic quality of the image may result in a *usable* rating for an image, while objective evaluation of image quality might have the opposite rating and vice versa.

6.2 Information Content

The ultimate goal of PET is to provide greater accuracy and to minimize both subjectivity and bias through the representation of relevant image information for either nonquantitative visual examination or quantitative analysis. The determination of the functional state of a biologic process or structure depends on the interpretation of the physiological meaning of the information contained in an image. Our ability to take advantage of the functional image requires accurate methods of data acquisition, presentation, analysis, and recognition of the limits of these approaches in practical applications.

A practical example is in determining the statistical limitations of dynamic tomographic imaging given a scanner with fixed sensitivity and detector resolution. Two important questions of interest that determine the total number of counts in an image are these:

1. How much activity is required to provide an image with high statistical accuracy?
2. What temporal sampling frequency is required to measure the dynamic changes in the radiopharmaceutical distribution?

The amount of radioactivity is limited by the dose that can safely be administered to the subject and the counting-rate characteristics of the scanner. The duration or interval of scanning is limited by the physical half-life of the radionuclide, the biological half-life of the radiopharmaceutical (effective half-life), and other impediments such as patient motion and comfort.

Knowledge of the minimum activity required to detect or quantitate the activity concentration in a structure and of the amount of activity in the FOV above which there is no statistical benefit in acquiring more counts can aid in determining the amount of activity and the scan duration. Thus, the sensitivity of the PET system; the amount of radioactivity, size, and location of the object in the FOV; and the duration of the scan determine the number of image counts. The effects of acquiring more counts are reduced image noise and improved image quality and quantitative precision. The spatial resolution with respect to the object's size and location determines the quantitative accuracy.

In PET, information is represented by a digital image. For a digital image, its matrix size and pixel dimension determine the image FOV and detail, and its pixel depth (numerical range) determines the precision with which the number of events at a given location (pixel) can be represented. If the number of image counts is fixed, the image matrix size and pixel dimension determine the number of events per pixel and therefore the statistical noise at any pixel. The pixel dimension also determines the amount of detail in an image up to the limits of spatial resolution determined by the reconstruction filter and the detector. As a rule of thumb, the pixel dimension should be on the order of FWHM/3 to avoid loss of spatial resolution so that all information acquired will be preserved. Most systems offer pixel depths of 8 bits (256 levels) or 16 bits (65 536 levels) of dynamic range or precision with which the information (counts/pixel) is represented. However, if total information in the image has 3% e_{rms}, then there are about 32 statistically distinct levels of information represented by the image, regardless of pixel depth.

Because of the physical and biological properties of the radiopharmaceutical, images often have low counts, and statistical noise can become overwhelming. Usually, overall image quality can be improved by simple processing, but the poor quantitative information remains. In these instances, if possible, repeating scans under the identical conditions and then combining the scans can effectively reduce the amount of statistical noise and increase the information content.

6.3 Enhancement

In the previous sections, we stated the importance of the imaging protocol and the digital processing in obtaining and representing useful information degraded by statistical noise and spatial resolution. We have also stated that the reconstruction filter cutoff frequency can be adjusted to trade off image resolution and noise. In addition, other image-processing techniques, which we describe in a later section, can be used to remove unnecessary and unwanted information from the image while retaining and often enhancing relevant image information. The combinations of contrast enhancement and histogram equalization or the use of thresholding and histogram specification can have a considerable effect on the image and increase or decrease the dynamic range with which the image is presented. The selection of color scale, medium, display, brightness, hue, saturation, and the physical size with which to view the image can often simplify the identification of an object. Simple subtraction of one image from another obtained under identical scanning conditions but different physiological conditions can also be used to monitor therapy or the progression of a disease.

6.4 Interpretation

PET studies produce images from which both quantitative and nonquantitative data can be extracted. In diagnostic situations where the purpose is to determine the presence or absence of a defect or disease, images are usually visually interpreted. Quantitative data in the form of radiopharmaceutical uptake and functional parameters can be used to supplement the more subjective visual interpretation. In either quantitative or nonquantitative imaging studies, interpretation depends on the human perception of images, curves, or numbers and understanding of the underlying physiologic process or function under investigation.

Quantitative data are most often obtained from images by sampling (measuring) the values represented by a region of interest (ROI) in an image. The numeric value of an ROI can be that of a minimum, maximum, average, or summation of pixel values that represents the activity concentration or some functional or parametric value in a volume of tissue (e.g., brain, heart, or kidney). Because of partial-volume effects, some interpretation of the object is required when obtaining values with ROIs. If the object sampled is a "hot" object in a cold background and is smaller than three times the FWHM of the system, then a large ROI placed at least 1 FWHM outside of the object can be used to estimate the total activity of the object. In the same situation, if an ROI is placed within 1 FWHM of the border of the object, it can be used to estimate the activity concentration of the object. In the event that the object and the background both contain activity, the situation is more difficult to interpret.

In many studies, we are not only concerned with determining the radiopharmaceutical concentration in the structure but also the kinetic properties that determine its uptake and elimination from the structure. In such cases, a dynamic (temporal) series of images is obtained. The length of time and the duration of temporal sampling intervals and their effect on image noise have already been addressed. Although the data present in the ROI may be of sufficient precision and accuracy to permit a numerical analysis of the data, an image may not be of sufficient quality to provide sufficient information for ROI placement. Therefore, to analyze dynamic data one places an ROI either on a coregistered anatomic image or on a functional PET image that contains the most detail and then computes a time-activity curve (TAC) (activity concentration vs time) for that ROI using all the data in the dynamic sequence. A TAC can then be used with a mathematical model that describes the physiological process of the structure or the pharmaokinetic parameters of the radiopharmaceutical to extract further information.

7. IMAGE PROCESSING

The radionuclide distribution estimate obtained by computer reconstruction techniques is represented in the form of a digital image. This allows the application of a variety of digital image-processing techniques. Image-processing techniques are used to maximize and extract information and remove unwanted information through the use of arithmetic operations, transformations, and filters applied to pixels, ROIs, TACs, and images.

7.1 Arithmetic Operations

The arithmetic operations of addition, subtraction, multiplication, and division are carried out pixel by pixel on images. One of the most common operations is image summation to reduce statistical noise. As the number of images added together increases, signal increases and the variability decreases (precision increases). In practice the images must be registered to avoid blurring. Image subtraction can be used either to view or to quantitate the change of the radionuclide distribution between two images. Often images are not obtained under exactly the same conditions, and a scalar multiplication or division can be applied to normalize the data. Some examples are the normalization to injected activity or to average activity concentration. In addition the division of a subtraction image by one of the images and the scalar multiplication of each pixel by 100 can give a percentage increase or decrease relative to the image divisor. The use of point processing or pixel arithmetic operations has numerous important applications.

7.2 Transformations

Geometric translation, rotation, and scaling are used to form a new image by the mapping or converting of the original image onto a new coordinate space and are the basic transformations used to align or register one image or image volume to another. By resampling, reslicing, and reorienting an image along the planes of another data set, image surface and volume rendering employing sophisticated feature detection and warping algorithms can be used for simultaneous multimodality display. In many clinical situations (e.g., surgical planning and diagnosis), a functional PET image is combined for display purposes with structural images, such as x-ray computed tomography or magnetic resonance imaging, for precise anatomical localization of the radiopharmaceutical; and in image quantification, anatomical images are preferred for ROI placement. Anatomical images of a subject used for the correction of partial volume effects must also be coregistered with the corresponding functional PET image. Stereotaxic transformation of PET image to a standardized coordinate space of anatomic structures allows for both intersubject and intrasubject comparisons and has been widely used in functional brain mapping to interpret and communicate results.

7.3 Filters

In addition to pixel-by-pixel processing and image transformation, digital filters can be used to extract additional quantitative information and improve the statistical quality of the image. A filter is a mathematical operation applied to an image $f(x,y)$ to form a new image $f'(x,y)$ that is in some sense better, for a specific purpose, than the original image. Noise reduction, edge detection, and contrast enhancement are a few of the various possible operations provided by digital filtering. The most common method of filtering is the changing of a pixel value based on itself and its neighboring pixels. In this spatial application, the filter is an $N \times N$ matrix of coefficients, $h(x,y)$, centered over a pixel in the original image where each pixel under the filter is multiplied by the corresponding coefficient and the summation of the results is placed at the location of the central pixel in the new image. The spatial filter is then applied to the next pixel, and the process is repeated until all pixel locations have been filtered. This process, known as convolution, is given as

$$f'(x,y) = \sum_{x'=-N/2}^{N/2} \sum_{y'=-N/2}^{N/2} h(x',y') \times f(x-x',y-y'). \tag{53}$$

The application of the spatial filter at each pixel location can be computationally expensive.

As described above, an image or function in the spatial domain can be equivalently represented in the spatial frequency domain through use of the Fourier transform [Eq. (19)]. By the convolution theorem, which states that convolution in the spatial domain is the exact equivalence of multiplication in the frequency domain and vice versa, we can perform the filtering operation in the frequency domain, given by

$$F'(u,v) = H(u,v)F(u,v), \tag{54}$$

where $H(u,v)$, $F(u,v)$, and $F'(u,v)$ are the Fourier-transform representations of $h(x,y)$, $f(x,y)$, and $f'(x,y)$, respectively, and u and v are spatial frequency. The Fourier filtering approach is straightforward: We simply obtain the Fourier transform of the image, apply the Fourier filter using Eq. (54), and inverse Fourier transform the result to obtain the filtered image. Figure 18 shows the spatial and frequency-domain representations of a blurred function with different levels of noise added.

7.3.1 Classification The two methods of filtering can produce identical results and differ only in their method of implementation. Spatial filters operate on image data in a local region surrounding each point while frequency (Fourier) filters operate on the entire data. However, digital filters are best described and evaluated by the effects they have on the spatial-frequency components of the data. In fact, we have applied this concept to our discussions of the modulating effects of blurring and statistical fluctuations dominating high-frequency components of the data. We now characterize some digital filters and the effect they have on the fre-

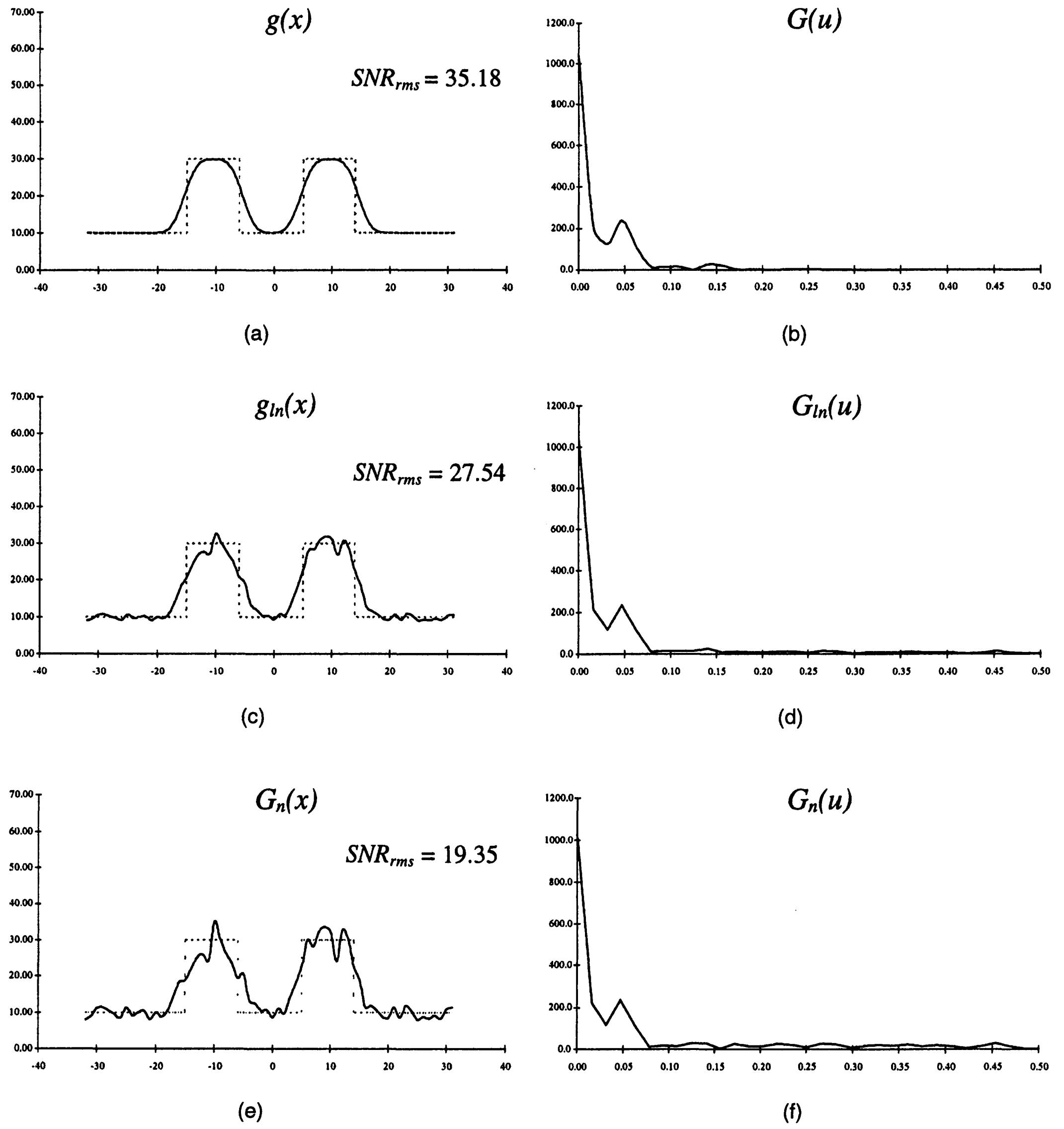

(a) (b)

(c) (d)

(e) (f)

FIG. 18. The spatial and Fourier-domain representations of a blurred function $g(x)$ superimposed with (a), (b) no, (c), (d) low, and (e), (f) high count-dependent (Poisson) noise. The dashed line represents the true object distribution.

quency components of data affected by resolution and noise.

Figure 19(a) shows a filter that passes low frequencies unaltered by multiplying the amplitude of the corresponding frequency components of the data by 1.0. The filter then rolls off, gradually reducing the amplitude of higher-frequency components of the data to zero. This low-pass "smoothing" filter has the effect of reducing high-frequency noise at the expense of spatial resolution (high-frequency signal). The filter shown in Fig. 19(b) is called the inverse filter because its coefficients are the inverse of the modulation transfer function (MTF^{-1}). Spatial resolution can be improved through the use of the inverse filter that is matched to the spatial-resolution degradation characteristics of the system. However, the inverse filter amplifies high-frequency noise as well as signal and

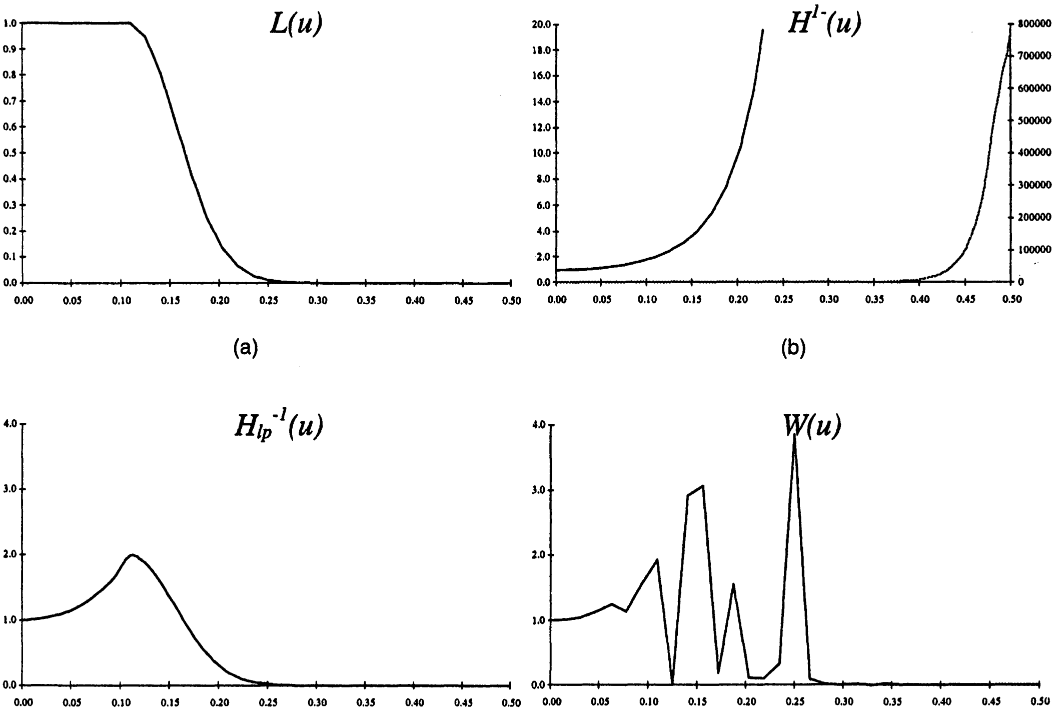

FIG. 19. The Fourier representation of (a) Low-pass, *L*(*u*), (b) inverse, $H^1(u)$, (c) inverse with a Gaussian rolloff, $H_{lp}^{-1}(u)$, and (d) Wiener, *W*(*u*), filters.

often renders the resultant image unusable. Figure 19(c) is the combination of an inverse and a low pass filter. This filter has the ability simultaneously to enhance contrast by amplifying the more reliable low- and intermediate-frequency components and to suppress noise by attenuating higher-frequency components and eventually eliminating the highest-frequency components dominated by noise. The filter shown in Fig. 19(d) is a frequency-selective filter known as the Wiener filter. Like the combination filter shown in Fig. 19(c), the Wiener filter has the ability to amplify the reliable low-frequency data and attenuate high-frequency components of noise. The Wiener filter, however, uses both the characteristics of the imaging system and the degraded image to form its frequency response.

By examination of the spatial-domain filters in Fig. 20, it is difficult to judge what a particular filter will produce. Therefore, our understanding of what these filters can produce must be developed by their effect on the frequency components of the data. However, as reflected by the MTF (see Fig. 15) and demonstrated by Fig. 18, the image information content (signal) at low spatial frequencies is greater than at high spatial frequencies where noise dominates. Therefore, the benefit of low-pass and other filters that employ a high-frequency cutoff is to effect an overall improvement in image S/N by the reduction of noise. The benefit of restoration filters that employ the inverse MTF is to remove the effects of imperfect spatial resolution by amplifying the signal at the more reliable low and intermediate frequencies and to suppress or eliminate noise at higher frequencies.

The simplest method of determining the coefficients of a spatial-domain filter is to specify a desired frequency response and then take the inverse Fourier transform of the filter, as shown in Fig. 20.

In Fig. 21, we demonstrate the application of the four filters described above to the noise-contaminated simulation data given in

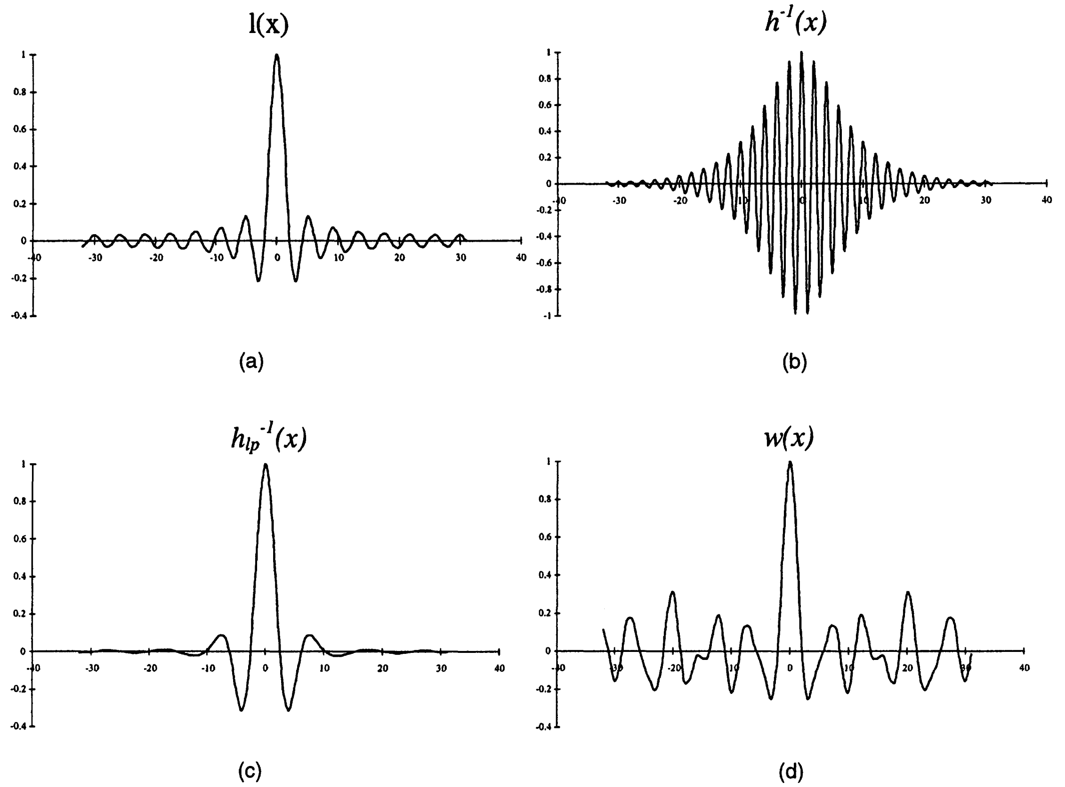

FIG. 20. The spatial representation of (a) low-pass, $l(x)$, (b) inverse, $h^{-1}(x)$, (c) inverse with a Gaussian rolloff, $h_{lp}^{-1}(x)$, and (d) Wiener filters $l(x)$, shown in Fig. 19.

Fig. 18(c). Figure 22 depicts the result of the first three applied to the noise-free simulated data [Fig. 18(b)]. Figure 22(d) shows the application of the Wiener filter to a low-noise simulation [Fig. 18(c) since its derivation defaults to the inverse filter in the absence of noise, as is discussed in the following section]. With respect to S/N_{rms}, all filters perform pretty much the same in the presence of noise [Figs. 21(a) through 21(d)] with of course the exception of the inverse filter [Fig. 21(b)]. From Fig. 22, we can deduce that the inverse filter is the best solution in the case of noise-free data; while the low-pass and modified inverse filter perform satisfactorily, we can see that their effect at high frequencies is to increase blurring as expected; and, also as expected, the Wiener filter provides a comparable restoration even in the presence of noise.

7.3.2 Derivation Carefully designed digital filters can be used to selectively enhance certain features and suppress others by considering the characteristics of the imaging device, the acquired data, and the information to be extracted. The principal objective of enhancement techniques is to process the data for some specific purpose (problem oriented); on the other hand, restoration techniques attempt to reconstruct an image that has been degraded, by modeling the process, and then applying the inverse of the modeled process to the data (model oriented). This requires knowledge of the transfer characteristics of the system (i.e., the PSF or MTF), the object, and the noise.

As we have previously stated, the resultant PET image $g(x,y)$ is a linear combination of superimposed PSFs $h(x,y)$ scaled according to the object distribution $f(x,y)$. This process can be described by the convolution of the object distribution $f(x,y)$ and the system impulse response $h(x,y)$, given by

$$g(x,y) = h(x,y)*f(x,y). \qquad (55)$$

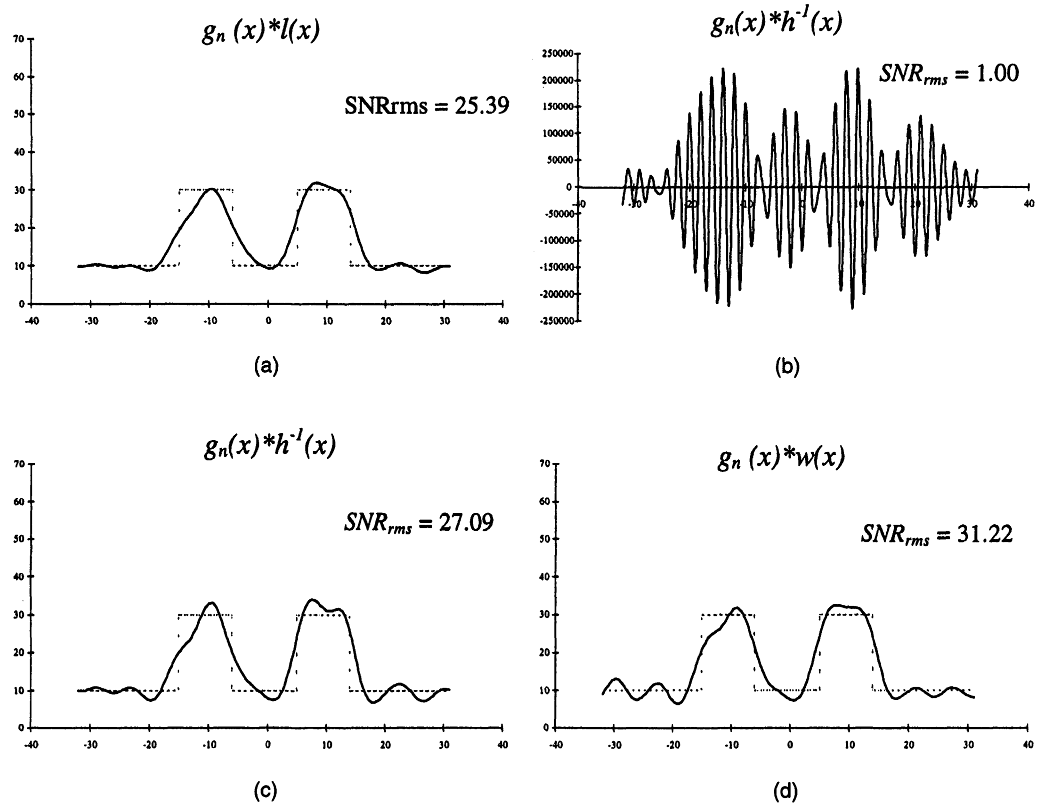

FIG. 21. Noisy simulation data filtered with the Fourier filters in Fig. 19.

where the asterisk denotes the operation of convolution, and its Fourier representation

$$G(u,v) = H(u,v)F(u,v). \tag{56}$$

The restoration solution (estimate) of the true signal $F(u,v)$ can be obtained by solving Eq. (56) for $F(u,v)$ to obtain

$$F(u,v) = H^{-1}(u,v)G(u,v). \tag{57}$$

This solution is the process known as inverse filtering. The method often does not work because $g(x,y)$ cannot be known precisely. Equation (55) is actually expressed as

$$g(x,y) = h(x,y)*f(x,y) + n(x,y). \tag{58}$$

where the asterisk again denotes the operation of convolution and $n(x,y)$ is statistical noise, and its Fourier representation is thus given by

$$G(u,v) = H(u,v)F(u,v) + N(u,v). \tag{59}$$

The optimum solution (filter), which minimizes mean squared error between the original distribution and the filtered image, is the Wiener filter, $W(u,v)$, given as

$$W(u,v) = H^{-1}(u,v)\frac{H^2(u,v)}{H^2(u,v) + N^2(u,v)/F^2(u,v)}. \tag{60}$$

8. EXAMPLE STUDIES AND IMAGES

Regional cerebral blood flow, rCBF, can be measured with $H_2{}^{15}O$. Because of the short physical half-life of ^{15}O, $H_2{}^{15}O$ measurements can be made between 10 and 15 min apart in the same subject during the same session. rCBF measured with $H_2{}^{15}O$ is currently the most popular method for PET activation studies. Within a few minutes after a bolus intravenous injection of $H_2{}^{15}O$,

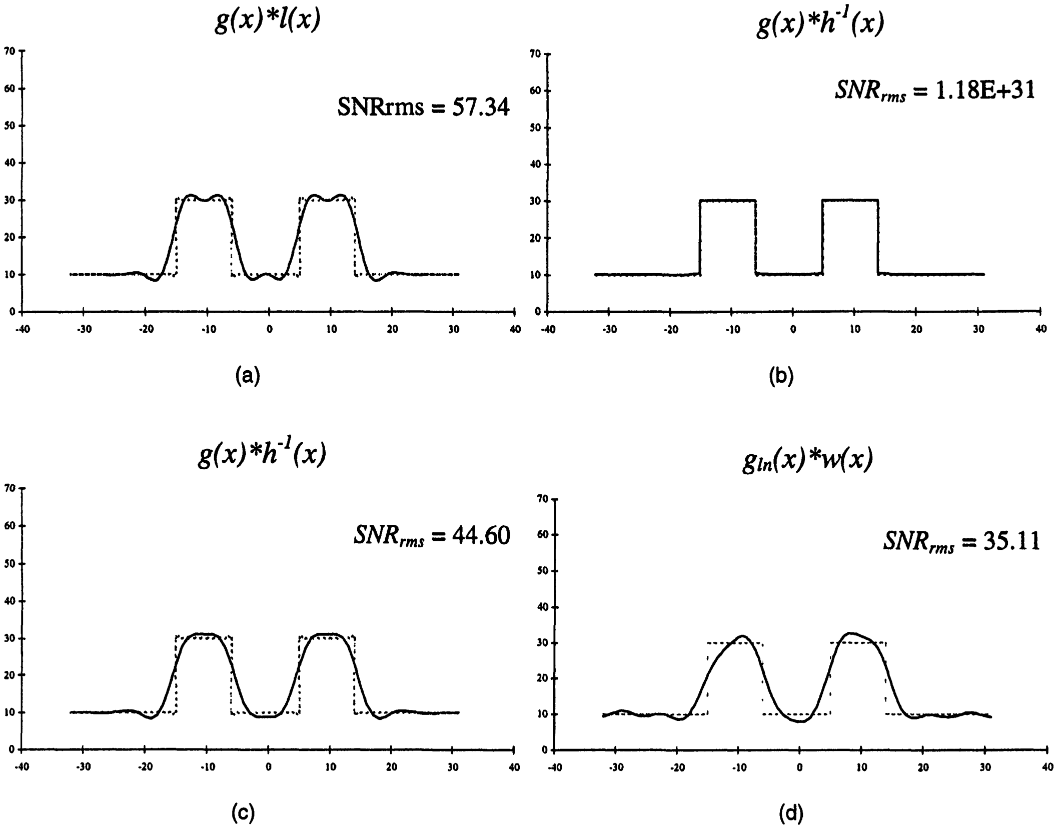

FIG. 22. Noise-free simulation data filtered with the Fourier filters of Figs. 19(a), 19(b), and 19(c). (d) The result of applying the Wiener filter to low-noise data.

quantitative rCBF measurements can be obtained. Absolute quantification requires measurement of the arterial time-activity curve, TAC (input function), but is not necessary for activation studies because rCBF and local tissue activity concentration is nearly linear. The distribution of $H_2{}^{15}O$ in human brain is shown in Fig. 23.

$H_2{}^{15}O$ is also ideal for quantitative region myocardial flow measurements because it is completely taken up by the myocardium (100% extraction) and it is not affected by metabolism. The extraction of $H_2{}^{15}O$ remains linear in the myocardium, even at high myocardial flow rates (unlike rCBF), and thus the distribution of tracer can be used to accurately reflect region perfusion.

The measurement of region cerebral metabolic rate of glucose, rCMRglu, with [^{18}F]-2-fluoro-2-deoxyglucose, FDG, is a well-established tool for diagnostic use and for the investigation of human disorders. The long half-life of ^{18}F makes it difficult to perform more than a single study in a day. The typical static FDG study requires a single scan after a 45- to 60-min uptake phase post injection and can be performed with and without arterial blood sampling depending on the need for absolute quantitation of rCMRglu or glucose uptake, respectively. Dynamic scanning, though rare, can also be used to determine rCMRglu. The distribution of FDG in human brain is shown in Fig. 24.

^{18}FDG is widely used to evaluate myocardial viability. Ischemic, but viable, myocardium utilizes glucose in preference to other substrates and will therefore exhibit increased uptake of ^{18}FDG. The determination of ischemic, but viable, myocardium is made only in relationship to perfusion. When regional myocardial ^{18}FDG uptake is increased compared with regional myocardial blood

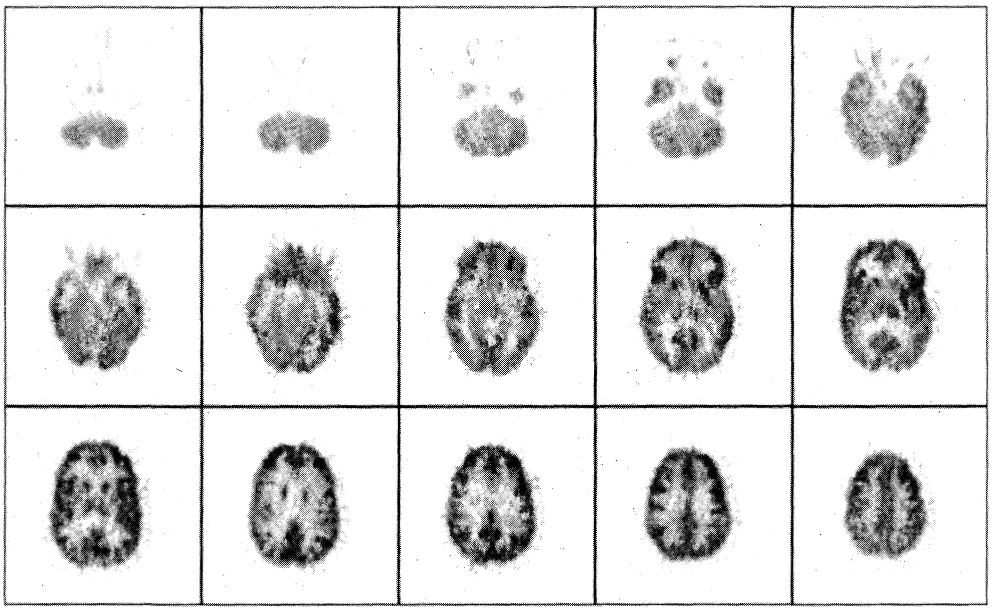

FIG. 23. Fifteen-slice PET image of rCBF acquired over 60 s after the intravenous injection of $H_2{}^{15}O$.

flow, the perfusion-metabolism mismatch is highly indicative of myocardial viability.

Using ^{18}FDG imaging, malignant tumor cells can be detected because of their enhanced glycolytic rate under anaerobic conditions as compared with nonmalignant tissue cells. Metabolic imaging of the brain using ^{18}FDG has also provided evidence for altered metabolic patterns in schizophrenia, Alzheimer's disease, and patients with a va-

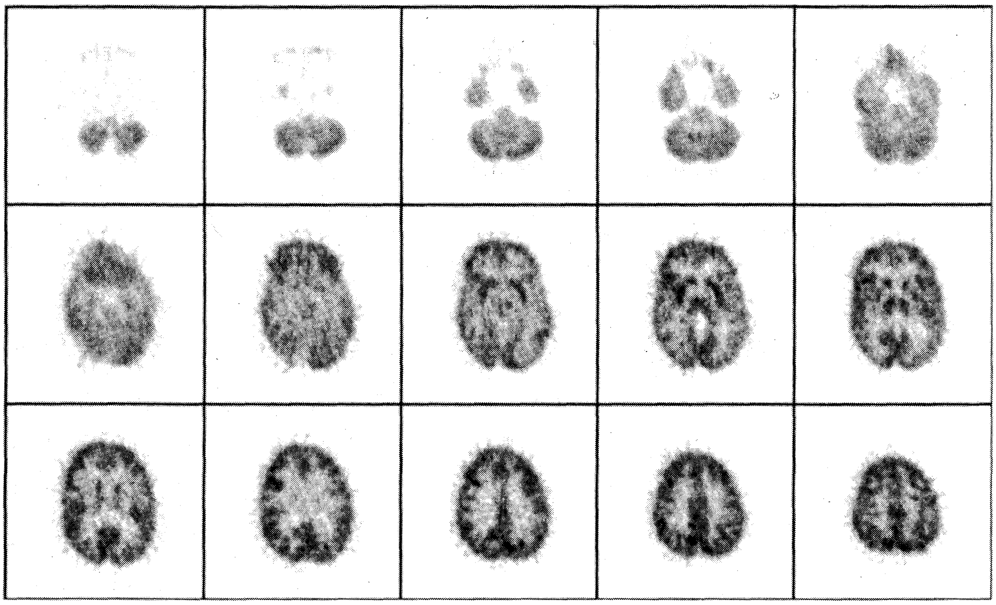

FIG. 24. Fifteen-slice 20-min PET image of rCMRglu starting 40 min after an intravenous injection of FDG.

riety of neuropsychiatric and neurodegenerative disorders.

Neurodegenerative disorders result from the death and or alteration of neurons. With PET it is possible to examine the many factors associated with neuroreceptor binding of PET ligands. For quantitative *in vivo* studies that measure receptor densities and affinities, the appropriate model of the pharmacokinetics of the administered neuroreceptor binding ligand, the time course of activity at the binding site, and an input function are required for the computation of the specific binding parameter. There are numerous receptors, receptor subtypes, and receptor binding ligands that can be investigated with PET. Shown in Fig. 25 is a dynamic series of human brain scans of the distribution of ^{11}C-Raclopride. The differences in image quality in dynamic scanning as compared with the static scans given in the previous examples arise from the differences in the number of acquired counts in each image. Images later in the dynamic study have fewer counts on account of temporal sampling requirements for pharmacokinetic modeling and physical decay.

GLOSSARY

Accidental Coincidence: The random overlap of two unrelated events within the coincidence resolving time of the detector.

Annihilation Photons: Two nearly collinear 511-keV photons resulting from the interaction of a positron and an electron in which their mass is converted to electromagnetic energy.

Coincidence Detection: The detection of two photons in two detectors within a small interval of time (typically less than 15 ns).

Dead Time: The time during which the scanner cannot respond to new events because of the processing of previous events.

FWHM: Full width at half maximum of a point-spread function. Usually used as a measure of spatial resolution.

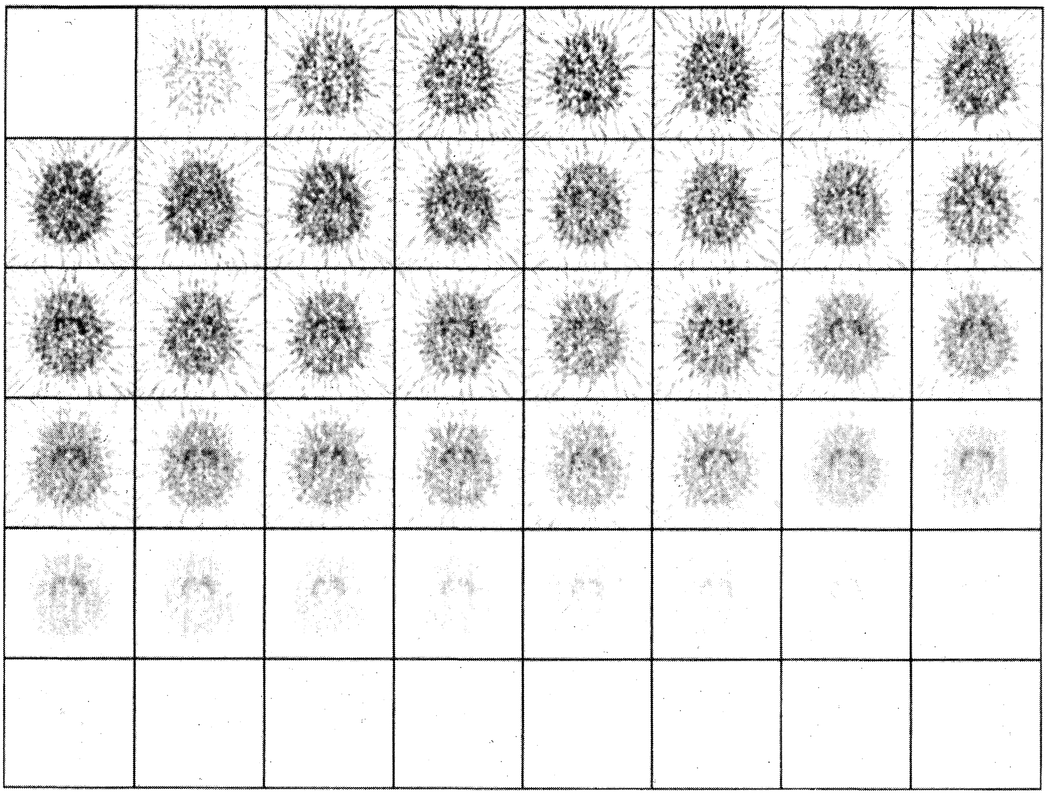

FIG. 25. A single PET slice of human brain over 60-min post-injection of ^{11}C-Raclopride.

Partial-Volume Effect: The underestimation or overestimation of radionuclide concentration in a structure that does not completely fill the sensitive volume of the resolution element of the system.

Point-Spread Function: The count profile of a line through an image of a point source of radioactivity.

Random Coincidence: See **Accidental Coincidence.**

Scatter Coincidence: A coincidence event in which one or both the photons originating from the same annihilation event have undergone Compton scattering.

Tracer: A substance used to trace or follow a chemical compound or process without perturbing the process being studied.

Works Cited

Alpert, N. M., Chesler, J. A., Correia, J. A., Ackerman, R. H., Chang, J. Y., Finklestein, S., Davis, S. M., Brownell, G. L., Taveras, J. M. (1982), *IEEE Trans. Med. Imag.* **2,** 142–146.

Bendiriem, B., Wong, W. H., Michael, C., Adler, S., Mullani, N. (1987), *J. Nucl. Med.* **28,** 681.

Bentourkia, M., Msaki, P., Cadorett, J., Lecomte, R. (1995), *J. Nucl. Med.* **36,** 121–130.

Bergstrom, M., Eriksson, L., Bohm, C., Blomqvist, G., Litton, J. (1983), *J. Comput. Assist. Tomogr.* **7,** 42–50.

Budinger, T. F., Derenzo, S. E., Greenberg, Gullberg, G. T., Huesman, R. H. (1978), *J. Nucl. Med.* **19,** 309–315.

Budinger, T. F., Derenzo, S. E., Huesman, R. H. (1984), *Ann. Neuro.* **15,** S35–S43.

Hoffman, E. J., Phelps, M. E. (1986), "Positron Emission Tomography: Principles and Quantitation," in: M. E. Phelps, J. C. Mazziotta, H. R. Schelbert, (Eds), *Positron Emission Tomography and Autoradiography: Principles and Application for the Brain and Heart,* New York: Raven Press, pp. 237–286.

Hoffman, E. J., Huang, S. C., Phelps, M. E. (1979), *J. Comput. Assist. Tomogr.* 299–308.

Hoffman, E. J., van der Stee, M., Ricci, A. R., Phelps, M. E. (1984), *Ann. Neuro.* **15,** S25–S34.

Kessler, R. M., Ellis, J. R., Eden, M. (1984), *J. Comput. Assist. Tomogr.* **8,** 514–522.

Logan, J., Bernstein, H. J. (1983), *J. Comput. Assist. Tomogr.* **7,** 316–320.

Muehllehner, G. (1985), *Phys. Med. Biol.* **30,** 163–173.

Further Reading

Gonzalez, R. C., Woods, R. E. (1992), *Digital Image Processing,* Reading, MA: Addison-Wesley.

Myers, R., Cunningham, V., Bailey, D., Jones, T. (Eds.) (1996), *Quantification of Brain Function Using PET,* New York: Academic.

Phelps, M. E., Mazziotta, J. C., Schelbert, H. R. (Eds.) (1986), *Positron Emission Tomography and Autoradiography: Principles and Application for the Brain and Heart,* New York: Raven Press.

Sorenson, J. A., Phelps, M. E. (1987), *Physics in Nuclear Medicine,* New York: Harcourt Brace Jovanovich.

Wagner, H. N. Jr., Szabo, Z., Buchanan, J. W. (Eds.) (1995), *Principles of Nuclear Medicine,* Philadelphia, PA: W.B. Saunders Company.

TOPOGRAPHY, GEOGRAPHICAL

KEITH C. CLARKE, *Department of Geography, University of California, Santa Barbara, California, U.S.A.*

Introduction and Definitional Constraints 525
1. **Geodetic Models and the Earth's Reference Ellipsoid** 525
1.1 The Datum and Sea-Level Definitions 526
1.2 Structure of Topographic Form 527
2. **Topographic Measurement** .. 527
2.1 The Digital-Elevation Model ... 528
2.2 Cartographic Depiction of Topography 529
2.3 Empirical Findings 530
3. **Hydrological and Skeletal Terrain Topology** 530
Glossary 530
Works Cited 531
Further Reading 531

INTRODUCTION AND DEFINITIONAL CONSTRAINTS

Definitions of topography vary in scope immensely. Those with a basis in topographic mapping define topography as the configuration of the surface of the Earth, including its relief, the position of its streams, roads, and cities, and other geographic features. A single feature such as a river or a valley is termed a topographic feature, and these features are subdivided into hypsography (the relief features), hydrography (the water and drainage features), culture (the man-made features), and vegetation. These terms are somewhat specific to the land-surface topography, that is, the surface elevation of the features above sea level on Earth. More recent and far broader definitions do not restrict the topographic surface to the land, neither do they restrict it to a tangible surface form. Thus in oceanography and geophysics, the ocean floor's surface depicted by the measurement of bathymetry or water depth is called topography, as are not only the solid surfaces of other planetary bodies but also the distribution of continuously measurable or field variables over large areas such as salinity in the ocean or barometric pressure in the atmosphere. In the broader definition, the tendency is to consider topography as the spatial expression of a field variable in geographic space as a single-valued function, usually the outer or maximum measured value.

Terminology associated with topography includes terrain surface, land surface, and sometimes landscape for land topography, and bathymetry or ocean floor for the oceans. Maps showing topography are called topographic maps, though the means for the graphic or cartographic depiction of the surface varies substantially from map to map. In general, the lowest surface elevation within a specified area, measured with respect to some arbitrary Earth model and datum, subtracted from the highest is referred to as the relief. Thus the Earth's land topographic relief to the nearest meter is 8846 m.

The measurement of topography implies determining the surface elevation. The elevation is determined opposite to the surface "down" vector, which can vary since the Earth's gravity varies from place to place. Elevation is measured not from the Earth's center, but from a preset level known as a datum. Often, sea level forms the local datum, but precise elevation determination must account for the exact three-dimensional geometry of the Earth.

1. GEODETIC MODELS AND THE EARTH'S REFERENCE ELLIPSOID

Over time, the Earth has been considered as having three different three-dimensional

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

forms. The simplest, and still suitable for small-scale global mapping, is the sphere. While our planet has considerable complexity of internal structure, active tectonism, and some remarkable rifts, mountains, and ocean trenches, nevertheless the total surface relief is minor with respect to the Earth's size, a mere 0.1% of the equatorial circumference. The difference between polar and equatorial radii is also only about 0.33%, and as a result the sphere is a good first approximation of the Earth's shape for global mapping at generalized scales. In the early eighteenth century, Isaac Newton argued that the Earth should be flatter at the poles than the equator, contrary to early French estimates that had indicated the Earth to be egg-shaped, with the greatest curvature at the poles. The French Academy of Sciences sent expeditions to Peru and Lapland starting in 1735, eventually validating Newton's hypothesis and beginning a series of 26 meridional measurements to estimate the Earth's dimensions between 1799 and 1951 (Snyder, 1987).

The Earth's shape was determined to be that of an ellipse rotated about its shorter or minor axis, a figure known as an oblate ellipsoid or oblate spheroid. The ratio of the lengths of the two semiaxes of the ellipse of revolution, subtracted from 1, is termed the flattening (Fig. 1). All flattening estimates are about one part in 300. Some highly influential measurements of the lengths of the semiaxes and the flattening are those of Alexander Ross Clarke in 1860 and 1866, John Fillmore Hayford in 1909, and Clarke and Helmert in 1911. The Clarke 1866 ellipsoid used the measurements of 6 378 206.4 m for the equatorial radius and 6 356 583.8 m for the polar radius, giving a flattening of 1/294.9787. Hayford's measurement used the rounded values of 1/297 and a semimajor axis of exactly 6 378 388 m. These values were adopted by the International Union of Geodesy and Geophysics (IUGG) in 1924, and termed the International or Hayford Ellipsoid. Since mapping of the United States was already under way, the Clarke 1866 values were used for the United States. Mapping efforts worldwide have used more than a dozen different ellipsoids; for example, many African countries use the Clarke 1860 ellipsoid.

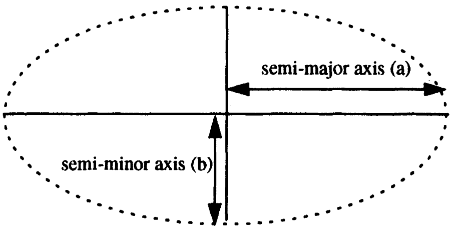

$$f = 1 - \frac{b}{a}$$

For the WGS84 (World Geodetic System)
a = 6,378,137 meters
b = 6,356,752.3 meters so f = 1/298.257223563

FIG. 1. The Earth's ellipsoid and its parameters.

Ellipsoids determined by measurement within one continent reflect minor differences in the Earth's local shape. The satellite era has allowed calculations based on the entire Earth. Two such measurements are the World Geodetic System measurement of 1972 (WGS72) and the Geodetic Reference System measurement of 1980. The latter (GRS80) was adopted by the IUGG and with other measurements has been the basis for mapping in the United States as the ellipsoid for the North American Datum of 1983 (NAD83) and the Department of Defense's WGS84. WGS84 is commonly used with the Global Positioning System (GPS) and assumes an equatorial radius of 6 378 137 m and a flattening of 1/298.257 223 563. As a result, most features on topographic maps have "shifted" as much as 300 m horizontally, though less vertically, from the older NAD27 datum based on the Clarke 1866 ellipsoid.

1.1 The Datum and Sea-Level Definitions

The third model of the Earth's shape uses the geoid, that is, the precise surface deviation from a global ellipsoid at each topographic elevation point, measured to include local gravitational field variations. With many ellipsoids, this can give several topographic surfaces. Figure 2 shows a local configuration of the measurements. A datum is defined as a numerical or geometrical quantity or set of quantities that may serve as a reference or base for other quantities. The datum associated with an ellipsoid becomes locally a reference level for measurement of

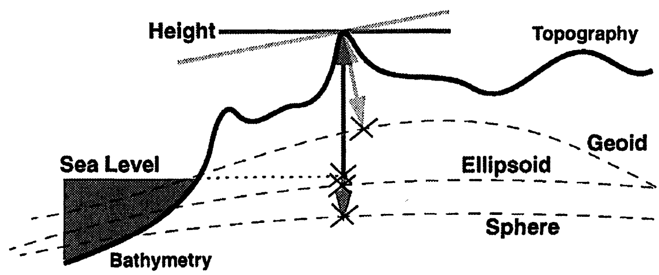

FIG. 2. Topographic elevations defined with reference to a sphere, ellipsoid, geoid, or local sea level are all different.

topographic elevation. This may be close to the sea level, but a datum can trend above and below this level, which itself varies by the tides, by the season, and over the long term for various reasons, such as sea-level change. The datum of mean sea level is common on nautical charts.

Detailed topographic mapping proceeds by extending measurements outward from known points. In much of the world, the level of geodetic precision possible with the GPS is superior to most previous measurements and has led to a resurveying of the control points used as the skeleton for topographic mapping. Even hand-held portable GPS receivers can be used to determine position and elevation locally, though accuracy varies. On most maps showing topography, therefore, the minimum requirement is to list the datum, the ellipsoid, and the measured elevation. In practice, points are extended locally from control points or bench marks with previously measured accurate elevations by traditional surveying methods.

1.2 Structure of Topographic Form

The land surface and ocean floor surface topographies at the local scale contain a set of features with common names determined mostly by their relief, absolute elevation, and shape. At the broadest scale below the planetary, landforms are the result of tectonic forces shaping the movement and position of solid continental plates floating on a more liquid mantle. Major topographic features are the result of plate collisions, dynamics at plate boundaries, hot spots or holes in the plates, and time. Gravity, volcanism, and other surface and ocean processes continuously modify the topography, so that large-scale features are constantly in flux. Landform features include mountains, hills, plateaus, lowlands, depressions, rift valleys, and mesas. Each of these can be caused by several means, including volcanism, folding, faulting, mass wasting (soil creep, slumping, and debris flow), or erosion and deposition.

Topographic features dominated by gravity include cliffs, scree slopes, alluvial fans, and lava flows. Gravity dominates in the deep ocean; waves and currents move material to form some bathymetric landforms, while chemical solution or coral growth forms others. Those landforms dominated by water erosion are valleys, channels, and gullies. Ice is a major topographic erosional feature determinant, forming hanging valleys, aretes, and U-shaped valleys. Water deposition forms deltas, flood plains, levees, and marshland, while ice deposition forms drumlins, eskers, and moraine. Wind erosion can sculpt wind caves and arches, while wind deposition can form dunes and large soil deposits known as loess. In addition, some types of geology have characteristic landforms such as the karst or sink-hole topography of chemical solution erosion on limestone. Most land features can be flooded as oceans rise or land falls and retain some of their form underwater, such as the Hudson Canyon off of New York. Many features are specific to the oceans, including the mid-ocean ridges, the deep trenches, and the continental shelves.

2. TOPOGRAPHIC MEASUREMENT

The measurement of topographic features dates from the first methods of determining height on maps by triangulation. The Romans could determine line of sight horizontally, and they must have used the clinometer to determine vertical angles. Along with the length of a line on the map, the vertical angle allows vertical elevation difference calculation and, therefore, relative elevations to be computed. This technique reached a higher degree of sophistication with the invention in the sixteenth century of the engineer's transit and later the theodolite, which used circular angle scales and pivoted telescopes to measure both horizontal and vertical angles precisely. During the nineteenth century, two additional methods became available. The altimeter is a precision barometer that allows the user to compute elevation from the relationship between air pressure and temperature and altitude above sea level. The precision or spirit level allowed horizontal lines to be extended over long distances while measuring elevation increments on calibrated poles. Most early elevation mapping used one or more of these methods, for example, the Great Topographical Survey of India in the nineteenth century.

After World War II, the science of photogrammetry reached the point where air photographs could be used to compute relief. This was done by taking two photographs from different viewpoints, say two passes over a target on different but parallel flight paths, with a substantial overlap in the images. The same objects are viewable, but with different displacements on the two photographs. Measurements from the photographs along with the plane's altitude and the focal length of the camera then allow elevations to be calculated at points. Stereophotogrammetry has reached a great degree of sophistication, and the method is now used extensively in engineering, surveying, and military intelligence. Digital soft-copy photogrammetry, with computer software taking over the tasks of matching the images and computing the elevations, has replaced a first generation of mechanical devices for these measurements. Instruments now fly on aircraft and spacecraft, and stereo cameras are even used at close range for detailed mapping—for example, of buildings and monuments, and inaccessible locations.

Several new technologies have changed topographic mapping recently. The GPS, with a system of 21 operational satellites in known orbits along tracks, each transmitting an encoded time signal, has made direct measurement of position and elevation relatively simple and possible at extreme levels of precision, such as submeter accuracy (Herring, 1996). In addition, active remote sensing in the radar part of the spectrum has proven a versatile method for terrain mapping. With use of synthetic-aperture radar from two passes together, the interference pattern can be used to calculate relief, in some cases to the millimeter level of accuracy. Pioneered over the ocean surface by SeaSat during the 1970s, radar interferometry has been accomplished for the Earth from the SIR-C, ERS, JERS, and Radarsat instruments, by Galileo for Venus, and by Magellan for Jupiter (Plate 1). In addition, many scientific satellites now use laser altimeters to measure their track altitude, allowing precision elevation measurement along a profile on the ground.

2.1 The Digital-Elevation Model

A digital-elevation model (DEM) is a data set containing topographic elevations at regularly spaced points on the ground. Known also as a digital terrain model and a topographic grid, these data have become broadly available for use in mapping and scientific analyses. A typical DEM might consist of an array of elevations in a grid of 30-m ground spacing, forming an M by N array covering a rectangular ground area. In some cases, the spatial units are geographic coordinates of latitude and longitude, so that spacing depends upon latitude and the map projection. The DEM and the alternative triangulated irregular network, in which the elevations are irregularly spaced and linked in a data structure with their nearest neighbors in a Delaunay triangulation, are the two major means by which elevation data are made available for topographic mapping and analysis. A Delaunay triangulation is a subdivision of an area into triangles such that the circumcircle of every triangle does not contain any vertex in the triangulation, and which is optimal in that narrow triangles are eliminated.

Worldwide coverage of bathymetry and land topography by DEM is available at 5 min of arc from the National Geophysical Data Center, and at about 1 km from the United States Geological Survey. Otherwise, coverage is national, and a complete survey is available on the World Wide Web at http://www.geo.ed.ac.uk/home/ded.html.

The United States has a particularly rich set of DEMs available. At the 1:250 000 map scale, and with a DEM resolution of 3 arc sec, the whole country is mapped in arrays covering 1° by 1° of arc. About half of the nation is available at the 30-m resolution. Examples of the two data sets are shown in Plate 2.

Sources of DEM data vary. Some comes from digitized contour maps, but the majority comes from the stereo models used to rectify large-scale maps to planar geometry, particularly the 1:24 000 topographic maps and the 1:12 000 Orthophoto quadrangles. Bathymetry has come from the detailed sounding data available from side-scan sonars. Accurate topography is often maintained as a military secret, usable for targeting in particular the cruise missile, and for navigating submarines. Detailed global bathymetric data released by the U.S. Navy in 1995 is an example.

The DEM has many advantages for the topographic analyst. Using the array as the basis for calculations, computation of surface aspect and slope is simple, and image-processing methods can be used for the detection of features and for further operations. The recent Geographic Information System technology (see CARTOGRAPHY) has been especially suitable for these operation.

2.2 Cartographic Depiction of Topography

Since the early origins of cartography, depiction of the land surface features has been a challenge for map makers. The oldest maps show mountains as rounded "mole hills" or triangular shapes in rows or chains as they might appear from within a valley. After the fifteenth century the symbols were aligned with the line of sight and included outlines of cone-shaped and craggy pointed mountains differentiated by size, with shadowing suggested with lines with an apparent illumination. In the sixteenth century, innovative terrain depictions by Leonardo da Vinci influenced cartography toward more natural map symbology for hills and mountains. A three-dimensional appearance was achieved through slope lines and the use of shadow hachures suitable for printing using the wood-cut method. Later, copper engraving allowed for an increase of topographic content and far finer line detail.

France's national triangulation completed in 1774 by Cassini was accompanied by advances in both geodetic and topographic survey and copper-engraving map printing. A transition to a planimetric map view of topography took place near the turn of the nineteenth century. Areas of slope were depicted with the level areas left blank and hachures drawn following the rule that the greater the slope and ruggedness, the denser and darker were the lines. In 1799 the Saxon military topographer Johann Georg Lehmann developed a system called slope hachuring, with lines drawn in the direction of slope and arranged so that their end points began and ended on what were assumed to be contour lines with regular vertical intervals. Slope hachuring became the standard means of showing terrain through the nineteenth century and up until the middle of the twentieth century with apparent directional illumination (Imhof, 1982).

An alternative method for terrain depiction was invented by Edmund Halley, although earlier versions exist such as an isobath map by Pieter Buuinss in 1584. Originally used for magnetic declination, or deviation between true and magnetic north, the technique is generically called the isoline method, but for topography it is called contour mapping for elevation or isobaths for depths. Contour lines are imaginary lines that represent the intersections that arise from horizontally slicing up landforms into equal vertical intervals. Contour lines are measured up from a base datum, while the lines measured down from the datum are called depth contours or isobaths. The vertical height difference between contours is called the contour interval. Contours have an immense advantage over other methods because they record and convey the geometric form of the terrain including numerical elevations and the angles and directions of slopes, though their interpretation can take considerable skill.

The development of lithographic printing, including color printing, in 1796 and its application to printing maps in 1826 changed the ability of maps to depict topography. Hachures were changed to brown while other elements such as contours or different land surfaces could be printed in various colors. Hachures evolved into shade tones and areas printed in continuous color. During the latter part of the nineteenth century, hypsometric tints and naturalistic and symbolic landscape tints were first used. While some national mapping efforts in this era incorporated elegant shading methods, for example, in Switzerland, in most countries including the United States the contour became the standard topographic depiction method. Various attempts have been made to draw contours in such a way that they appear three dimensional, including methods by Tanaka and Yoeli (Imhof, 1982).

Block diagrams are a pictorial method that used to be used primarily for the portrayal of geologic and geophysical relationships by showing a portion of the Earth's surface and crust from an oblique viewing angle. The origins of the block diagram were eighteenth century geological reports illustrated with cross sections. Pictorial sketching was added to the section to add realism and relate the surface expression to subsurface geology. The gridded perspective map is another popular use of the block diagram for displaying surface terrain, varying viewing parameters such as azimuth, viewing height, zenith angle, and view point (Plate 3).

2.3 Empirical Findings

Many earth scientists have sought to describe topography quantitatively. Early efforts used descriptive parameters of elevation data, stream network branching and the terrain skeleton, and spectral analysis (Warntz, 1975; Horton, 1945; Strahler, 1964; Pike and Rozema, 1975). More recent techniques have used methods of statistics as adjusted for topographic information, called geostatistics (Journel and Huijbregts, 1978), and in particular a statistical relationship between elevation variance and measurement-point separation called the semivariogram that uses regionalized variable theory. For surface elevation Z at points u and at a separation h,

$$2\gamma(h) = \mathrm{Var}\{Z(u + h) - Z(u)\}.$$

Topography has been found usually to conform to a simple type of the variogram, with elevation semivariance (γ) increasing to a constant value (the sill) at a particular distance (the range). Several models of this relationship have been proposed and are used in the statistical interpolation method known as Kriging (Burrough, 1986).

Alternative models of topography have used the fractal theory of Mandelbrot (Klinkenberg and Clarke, 1992). In the fractal model, variance is attributed to every level of aggregation within the terrain model, from the finest scale to the size of the whole landscape. Although the fractal model has not been effective at explaining all topography, it is an excellent model of some terrain and produces useful topographic simulations, in many cases hard to distinguish from reality.

3. HYDROLOGICAL AND SKELETAL TERRAIN TOPOLOGY

A terrain skeleton is the network of ridges and valleys partitioned by break points and extrema of slope in a piece of topography. Elements on the skeleton are termed "very important points" and include peaks, the bottoms of depressions, ridge and break lines, and break points along the drainage system. The terrain skeleton is identified in order to perform further analysis, such as drainage patterns or hillshading. Lines that divide up topography are called skeletal lines, structure lines, or break lines and have been used to improve the visual interpretation of hill and mountain shapes and to increase the fidelity of DEMs. The terrain skeleton has been used as a constructional aid in the representation of terrain in both manual and automated methods. Topographic skeleton features can be automatically extracted from a digital elevation model, from digitized photogrammetric data, or from topographic maps.

Geomorphologists have conducted considerable research on topographic skeletons as a way of syntactically modeling terrain. Significant features in the skeleton reduce to pits, peaks, passes, ravines (vertical surfaces), and ridges (Fig. 3). These locations define also critical points in modeling the behavior of surface water in hydrologic modeling (Band, 1986; Douglas, 1986; Mark, 1975; Woldenberg, 1969).

GLOSSARY

Bathymetry: The topography of the ocean floor, including a surface of depths representing measurements of distances from the

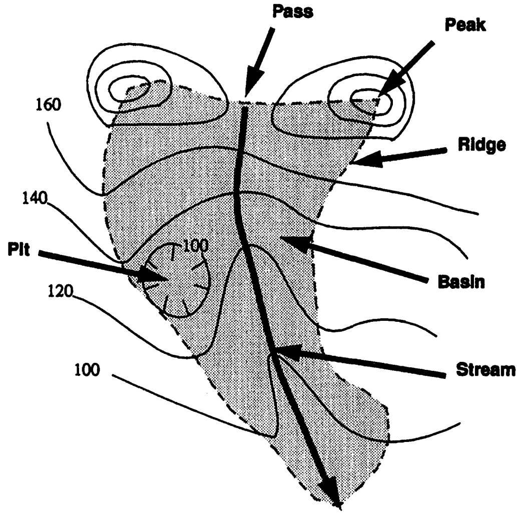

FIG. 3. Features of the terrain skeleton.

surface defined by the datum to the upper surface of the ocean floor.

Contour: A line on a map joining points of equal topographic elevation.

Datum: A reference level used to measure surface elevation, often sea level or a model of the Earth's shape such as an ellipsoid.

Digital Elevation Model: A data set containing topographic elevation measurements, usually in the form of a grid and for an *M* by *N* rectangular array.

Geoid: A more complex Earth shape model, including measured deviations from an ellipsoid.

Oblate Ellipsoid: A model of the Earth's surface shape, the locus of an ellipse rotated in space about its shorter axis.

Relief: The highest minus the lowest topographic elevation within a defined region.

Terrain Skeleton: The network of surface extrema, ridge, and stream and break lines that define a syntax for topography and partition the surface into homogeneous facets.

Topography: The spatial expression of a field variable in geographic space as a single-valued function, usually the outer or maximum measured value.

Works Cited

Band, L. E. (1986), "Topographic Partitioning of Watersheds with Digital Elevation Models," *Water Resources Res.* **22,** 15–24.

Burrough, P. A. (1986), *Principles of Geographical Information Systems for Land Resources Assessment,* Oxford: Clarendon Press.

Douglas, David H. (1986), "Experiments to Locate Ridges and Channels to Create a New Type of Digital Elevation Model," *Cartographica* **23** (4), 29–61.

Herring, T. (1996), "The Global Positioning System," *Sci. Am.* **274** (2), 44–50.

Horton, R. E. (1945), "Erosional Development of Streams and Their Drainage Basins: Hydrophysical Approach to Quantitative Geomorphology," *Geol. Soc. Am. Bull.* **56,** 275–370.

Imhof, E. (1982), in: H. J. Steward (ed.), *Cartographic Relief Presentation,* New York: de Gruyter.

Journel, A., Huijbregts, C. J. (1978), *Mining Geostatistics,* New York: Academic.

Klinkenberg, B., Clarke, K. C. (1992), "Exploring the Fractal Mountains," in: I. Palaz, S. K. Sengupta (Eds.), *Automated Pattern Analysis in Petroleum Exploration,* New York: Springer-Verlag, pp. 201–212.

Mark, D. M. (1975), "Computer Analysis of Topography: A Comparison of Terrain Storage Methods," *Geogr. Ann.* **57A,** 179–188.

Pike, R. J., Rozema, W. J. (1975), "Spectral Analysis of Landforms," *Ann. Assoc. Am. Geog.* **65,** 499–516.

Snyder, J. P. (1987), *Map Projections—A Working Manual,* U.S. Geological Survey Professional Paper 1395, Washington, D.C., U.S. GPO.

Strahler, A. N. (1964), "Quantitative Geomorphology of Drainage Basins and Channel Networks," in: Ven Te Chow (Ed.), *Handbook of Applied Hydrology: Compendium of Water Resources Technology,* New York: McGraw-Hill, pp. 39–76.

Warntz, W. (1975), "Stream Ordering and Contour Mapping," *J. Hydrol.* **25,** 209–227.

Woldenberg, M. J. (1969), "Spatial Order in Fluvial Systems: Horton's Laws Derived from Mixed Hexagonal Hierarchies of Drainage Basin Areas," *Geol. Soc. Am. Bull.* **80,** 97–112.

Further Reading

ASCE/ACSM/ASPRS (1994), *Glossary of the Mapping Sciences.* Washington, D.C.: American Society of Civil Engineers, American Congress on Surveying and Mapping, and the American Society for Photogrammetry and Remote Sensing.

Clarke, K. C. (1995), *Analytical and Computer Cartography,* Englewood Cliffs, NJ: Prentice Hall.

Hall, S. S. (1992), *Mapping the Next Millennium: The Discovery of New Geographies,* New York: Random House.

Harley, J. B., Woodward, D. (Eds.) (1987–), *The History of Cartography,* Chicago: University of Chicago Press (multi-volume history).

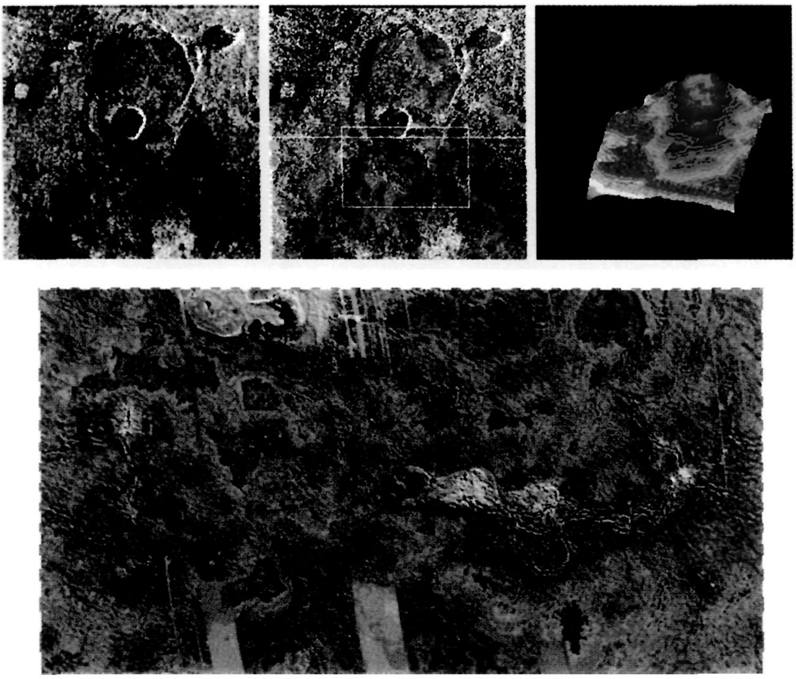

Vol. 21 – PLATE 1. Top: Topography of Kilauea, Hawaii, imaged by SIR-C (Shuttle Imaging Radar) on the Space Shuttle (Source: Jet Propulsion Laboratory). Bottom: Topographic surface of Venus imaged by Magellan. Source: Jet Propulsion Laboratory, Pasadena, CA.

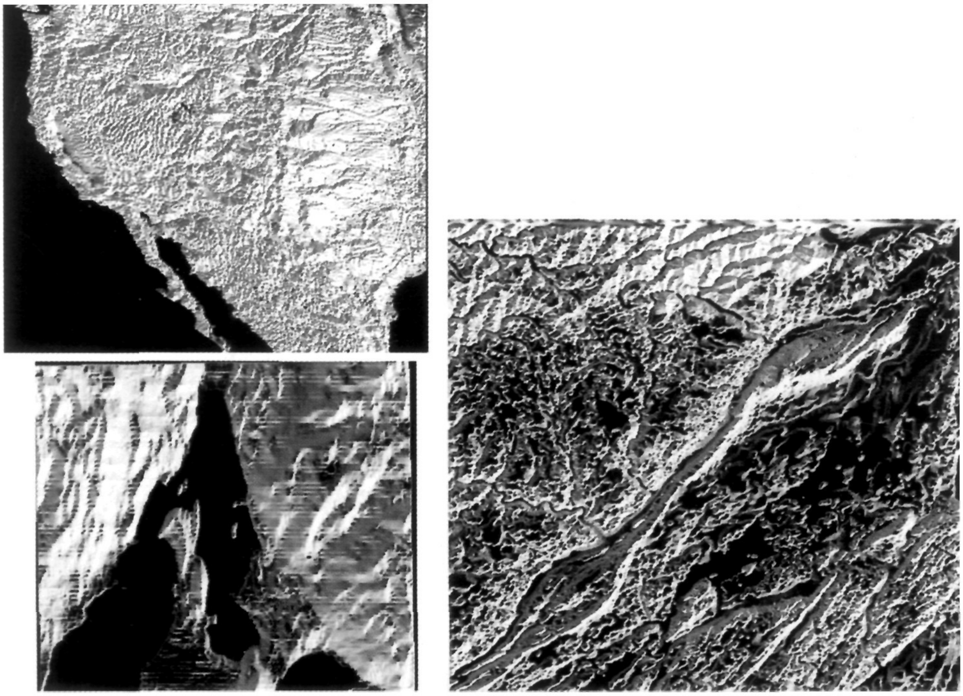

Vol. 21 – PLATE 2. Hill-shaded digital-elevation models. Upper left: Western United States from the ETOPO5 data (National Geophysical Data Center), resolution 5 arc min of latitude and longitude, land only shown; right: Scranton, PA, East 1° block, resolution 3 arc sec of latitude and longitude; lower left: McCall, ID, seven-and-a-half minute quadrangle at 30-m ground resolution.

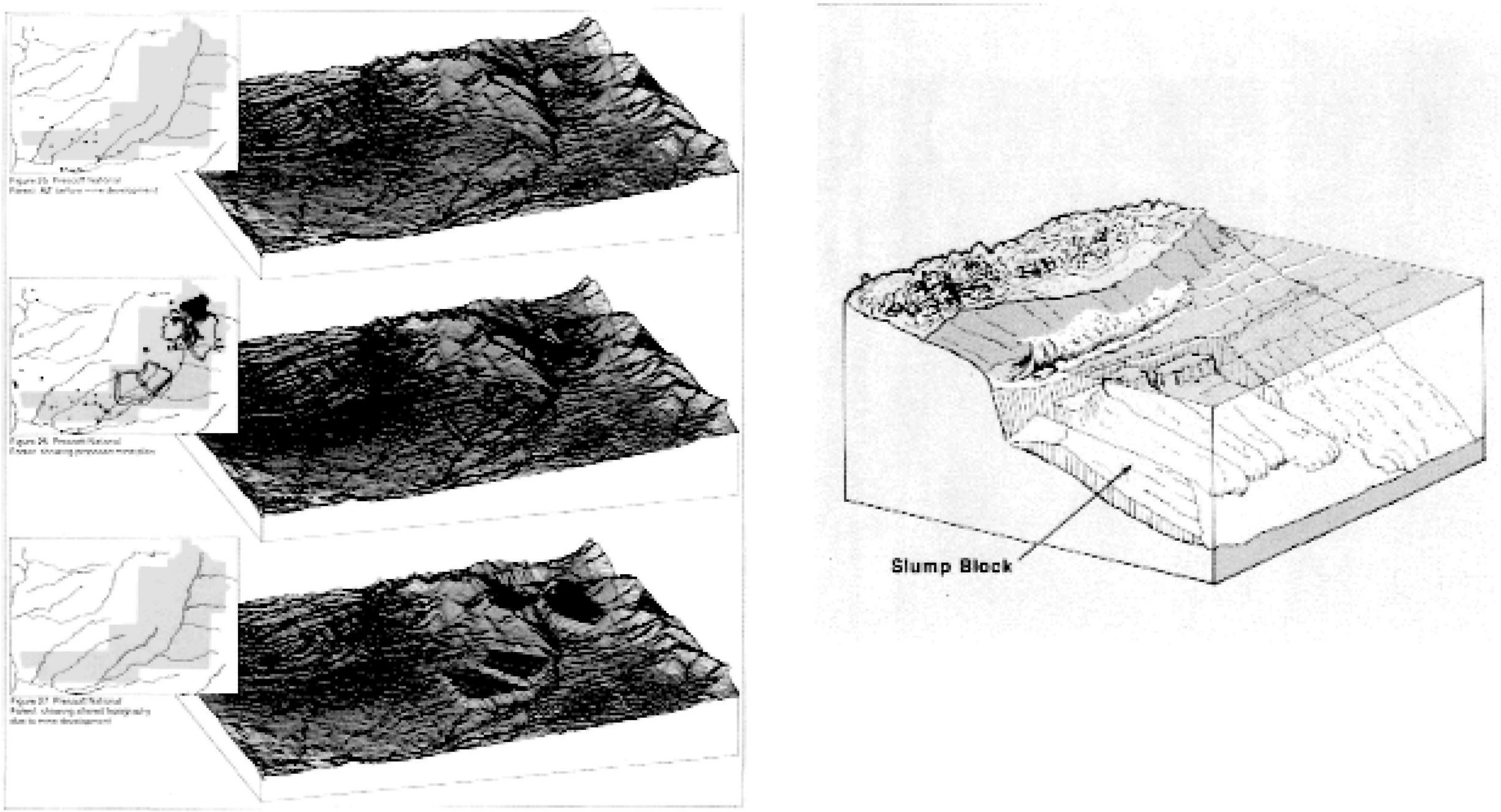

Vol. 21 – PLATE 3. Right: Block diagram. Source, Alpha Tau Rho, U.S. Geological Survey. Left: Gridded or „Fishnet“ topographic perspective view. Source, U.S. Geological Survey.

TOPOGRAPHY, X-RAY

MICHAEL DUDLEY, *Department of Materials Science & Engineering, State University of New York at Stony Brook, Stony Brook, New York, U.S.A.*

Introduction 533
1. **Contrast and Resolution** 533
2. **Conventional Radiation Techniques** 535
3. **Synchrotron Radiation Techniques** 536
4. **Mechanisms of Contrast Formation** 538
5. **Examples of Applications of X-Ray Topography** 539
5.1 Dislocation Analysis 539
5.1.1 Determination of Line Direction 539
5.1.2 Determination of Burgers-Vector Direction 540
5.1.3 Determination of Burgers-Vector Sense and Magnitude 540
5.2 *In Situ* Studies of Crystals Subjected to External Stimuli 543
6. **Comparison with Transmission Electron Microscopy** 544
Glossary 546
Works Cited 547
Further Reading 547

INTRODUCTION

X-ray topography is the generic name for a family of x-ray–diffraction imaging techniques capable of providing information on the character and distribution of crystallographic defects in large, nearly perfect single-crystal materials. The lesser used, more complete name of x-ray–diffraction topography is a little more informative in that it indicates that the technique is concerned with the topography of the internal diffracting planes, i.e., local changes in spacing and rotations of these planes, rather than with external surface topography, which perhaps may be misconstrued from the shortened version of the name. An x-ray topograph is a two-dimensional image obtained by projecting the distribution of diffracted intensity in an area-filling diffracted beam, produced by a low-divergence area-filling x-ray beam incident on a single crystal set at the Bragg angle, onto a two-dimensional detector (usually high-resolution x-ray film). Local differences in diffracting power between the distorted regions around defects inside the crystal and the more perfect surrounding regions lead to local intensity variations on the topograph, which are known as x-ray topographic contrast and which constitute "images" of the defects. Since no magnification is involved, the nature of the projection imaging process enables one to correlate the relative position of an image of a defect observed on an x-ray topograph with its location inside the crystal.

Quantitative information on the character of the defect (for example, the line direction and Burgers vector of a dislocation) can be obtained by detailed interpretation of the variations in contrast obtained under different diffraction conditions. Such interpretation requires an understanding of the mechanisms of contrast formation. These mechanisms are sensitive functions of the diffraction conditions and are discussed in a separate section.

1. CONTRAST AND RESOLUTION

The real information available in an x-ray topograph is manifested in the form of contrast within the projected diffraction spot. This topographic contrast arises from differences in the intensity diffracted as a function of spatial coordinate inside the crystal. Diffracted intensity is a sensitive function of local crystal perfection. For example, under

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

the correct diffraction conditions, highly mosaic regions of crystal (i.e., those regions comprising small subgrains that are slightly tilted with respect to the perfect crystal) will diffract kinematically, whereas more nearly perfect regions of crystal will diffract dynamically. Kinematical diffraction is characterized by an absence of the primary extinction effects that are the essence of dynamical diffraction. These primary extinction effects, created by the interference between diffracted waves and incident waves inside the crystal, lead to an overall reduction in diffracted intensity for a perfect crystal over the mosaic crystal. Conversely, there is an increase in diffracted intensity as the degree of mosaicity increases, eventually reaching the limit of the "ideally imperfect crystal" (White, 1950). The difference between the intensities diffracted from one region of crystal that diffracts kinematically and another that diffracts dynamically is one of the ways that dislocations can be rendered visible in topography (see Sec. 5). However, the situation is not as simple as this. Even for a crystal that diffracts dynamically, the diffracted intensity is a sensitive function of the local distortion in the crystal. In addition, the type of contrast that arises from a particular type of distortion is also a function of the absorption conditions in the crystal. In Sec. 5 a brief review of contrast mechanisms that arise in topography is presented, with particular attention being paid to those mechanisms that give rise to dislocation images.

It is important to point out that the contrast from individual defects will only be clearly discernible if the spatial resolution is adequate. There is no magnification involved in topography, because no lenses (in the normal sense of the word) exist for x rays, and spatial resolution is controlled solely by geometry. The geometrical factors arise as follows. The Bragg law is defined as

$$\lambda = 2d \sin\theta, \qquad (1)$$

where λ is the x-ray wavelength, d is the spacing of the diffracting planes, and θ is the angle between the incident rays and the diffracting planes, otherwise known as the "Bragg" angle. When this equation is satisfied, possible incident and diffracted beam directions lie, diametrically opposed, on the surface of a cone with semiapex angle 90° $- \theta$, the axis of which is the active reciprocal lattice vector, **g** (i.e., the normal to the diffracting planes). This cone is known as the Bragg cone. Therefore, if an x-ray source has a finite size, which it always does, the possibility arises that at a given point in the crystal the diffracting planes will receive radiation, at the correct Bragg angle, from several different points located at different positions on the source (which produce rays lying on the surface of the Bragg cone). The locus of these points on the source is the arc defined by the intersection between the operative Bragg cone and the source surface. This acceptance of rays with, for example, a finite vertical divergence imparts a finite vertical divergence to the diffracted beam that emanates from the point of interest in the crystal, and hence gives rise to a blurring effect in the resultant "image" of the point. In fact the diffracted rays emanating from the point of interest in the crystal will lie on the arc defined by the intersection between the same Bragg cone and the detector plane (the length of this arc will, of course, be defined by the divergence angle of the accepted rays from the arc on the source). This is illustrated in Fig. 1, which shows how an image of a point **P** in the crystal, recorded with rays that emanate from points lying on the arc **abc** on the source, is spread over the arc **a′b′c′** on the detector. Therefore, the angle subtended by the longest dimension of the source in the direction perpendicular to the plane of incidence (defined by the incident and diffracted beam directions, and the active reciprocal lattice vector) is the crucial parameter that should be minimized. The amount of blurring is proportional to the distance between the crystal and detector, and can be approximately written as

$$R = SD/C, \qquad (2)$$

where R is the blurring (or the effective resolution), S is the maximum source dimension in the direction perpendicular to the plane of incidence, D is the specimen–film distance, and C is the source–specimen distance. In the plane of incidence, diffraction occurs over a finite range of angles determined by acceptance angle of the crystal in that plane, or the rocking-curve width. This finite acceptance angle means that a given point in the crystal receives radiation from a

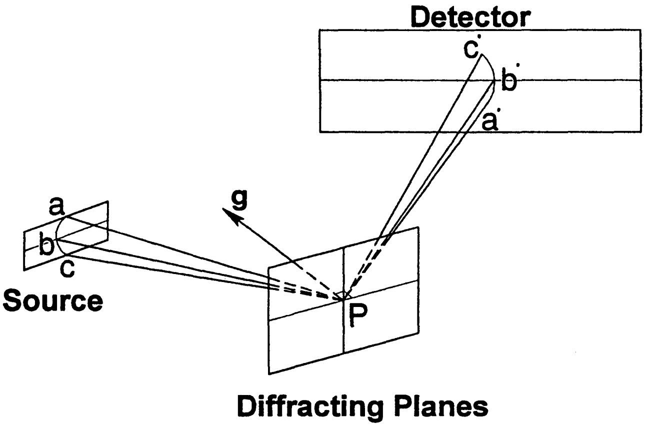

FIG. 1. Schematic diagram showing the effect of source dimension on resolution. **g** is the active reciprocal lattice vector. The angle **bPg**, which is the semiapex angle of the Bragg cone, is 90° − θ, where θ is the Bragg angle.

finite width on the source (for the case of a horizontal plane of incidence). This width can be determined by back projecting the fan of incident rays accepted at a point in the crystal to the source itself. Again a finite divergence, this time in the horizontal plane, can be attributed to the diffracted beam, and again a blurring effect will be evident. However, as can be verified by simple calculation, the limiting factor in determining spatial resolution is the finite source height (i.e., the source dimension perpendicular to the incidence plane), rather than the finite effective source width (i.e., the dimension parallel to the incidence plane).

Since the spatial resolution is proportional to the distance between the crystal and detector, this distance should be minimized. However, the optimum spatial resolution of the photographic detectors typically utilized in topography is limited, by photoelectron tracking between adjacent grains in the emulsion, to around 1 μm. The specimen–film distance is usually set to yield a calculated spatial resolution that approximately coincides with this. It should be noted that while this spatial resolution is greatly inferior to that of electron microscopy, it can be more than sufficient for the high-quality crystals of many materials that can currently be grown.

2. CONVENTIONAL RADIATION TECHNIQUES

The first techniques to be developed were designed to make use of characteristic radiation from conventional x-ray generators. These include the Berg-Barrett, Barth-Hosemann, double-crystal, and Lang techniques (for a review see Bowen and Hall, 1975). The most commonly used of these techniques is the transmission Lang technique. In this technique the crystal to be studied is set on a goniometer with the reciprocal lattice vector **g** of the intended diffracting planes lying parallel to the plane in which diffraction is to occur (usually the horizontal plane), and is then oriented for diffraction of a $K_{\alpha 1}$ characteristic line (usually from a fine-focus source, say ≈100 μm in dimension) from these planes. The resulting diffracted beam is recorded on high-resolution x-ray film. The angular resolution required to resolve this characteristic line (i.e., to separate the $K_{\alpha 1}/K_{\alpha 2}$ doublet) is usually achieved by collimation of the incident beam via an appropriately placed slit (typically a 100-μm vertical slit placed ≈50 cm from the source). This clearly results in an incident beam that has been restricted to a thin vertical "sheet" of x rays. Images recorded in this stationary configuration are known as section topographs, since under the correct conditions one can discern a cross-sectional view of the defect content of the crystal. In order to get large volume coverage, it becomes necessary to traverse the ensemble of crystal and detector in front of the incident beam. (It should be noted that the combination of drastic collimation and the necessity for translation usually leads to long exposure times, anywhere from minutes to hours, or even days). The image produced, which is essentially a two-dimensional projection of the diffracted intensity from the irradiated volume of crystal, constitutes a Lang or projection x-ray topograph. The spatial resolution attainable in the Lang technique is typically around 2 μm per centimeter of specimen–film distance.

3. SYNCHROTRON RADIATION TECHNIQUES

The advent of dedicated sources of synchrotron radiation has enabled the development of a new realm of topography, synchrotron topography. Synchrotron radiation has numerous advantages over conventional radiation when applied in topography. These advantages derive from the high brightness, tunability, and natural collimation of synchrotron radiation.

One of the most important synchrotron topographic techniques developed is white-radiation topography (Tuomi *et al.*, 1974; Miltat, 1980). This technique is basically analogous to the Laue transmission technique, except with greatly enhanced capabilities that derive from the aforementioned natural collimation and high intensity of the synchrotron beam. The natural collimation (typically $\approx 2 \times 10^{-4}$ rad in the vertical plane coupled with an acceptance angle of typically a few milliradians in the horizontal plane) allows the use of very long beamlines to maximize the area of the beam delivered at the sample without incurring significant losses in the total intensity originally available at the tangent point. This situation contrasts with that which would hold for the isotropic emission from a conventional source. The large beam area delivered at the sample location allows studies to be carried out on relatively large-scale single crystals, and crystals as large as 6 or even 12 in. (~13 or even ~25 cm) in diameter can be imaged by using precision translation stages similar to those used in the Lang technique (although exposure times are much shorter). Long beamlines also lead to small angles subtended by the source at points in the specimen, which in turn leads to excellent geometrical resolution capabilities. For example, at the Stony Brook Topography Station, on beam line X-19C at the National Synchrotron Light Source (NSLS), the optimum spatial resolution that is theoretically attainable is ≈0.04 μm per centimeter of specimen–film distance. Thus, in summary, we have in a synchrotron source a large-area x-ray beam, with inherent low divergence, and with excellent geometrical resolution capability.

Consequently, if a single crystal is oriented in this beam, and the diffracted beams recorded on a photographic detector, each diffraction spot on the resultant Laue pattern will constitute a map of the diffracting power from a particular set of planes as a function of position in the crystal with excellent point-to-point resolution (typically of the order of less than 1 μm). In other words, each diffraction spot will be an x-ray topograph.

The excellent geometrical resolution capability has another important consequence, since it relaxes to a certain extent the requirement of having very small specimen–film distances (<1 cm) in order to achieve good resolution. Thus, crystals can be surrounded with elaborate environmental chambers, necessitating considerable increases in specimen–detector distances (which would lead to intolerable resolution losses in conventional systems), without significant loss of resolution. The high intensity over the wide spectral range of the radiation emitted has several important advantages. Not only are exposure times necessary to record a topograph drastically reduced from the order of days on conventional systems to seconds on a synchrotron, but since we have a white beam, a multiplicity of images is recorded simultaneously. This leads to a great enhancement in the rate of data acquisition. The multiplicity of images also enables extensive characterization of strain fields present in the crystal. For example, instantaneous dislocation Burgers-vector analysis can become possible by comparison of dislocation images obtained on several different Laue spots (see Fig. 3, later). Similarly, lattice rotation in the specimen can be characterized by analysis of the asterism (the spreading of a Laue spot) observed on several different Laue spots. In addition, the good signal-to-noise ratios obtained associated with the high intensity also open up the possibility of direct imaging of topographs. Thus it becomes possible to conduct truly dynamic, quasi real-time studies of crystals subjected to some kind of external stimulus such as applied fields, applied stress, heating, cooling, etc.

Another important advantage of the white-beam topographic technique over conventional topographic techniques is its enhanced tolerance of lattice distortion. The wide spectral range available in the synchrotron beam allows crystals that either exhibit a uniform range of lattice orientation (for example, are uniformly bent by a small amount) or contain several regions of distinctly different orientation (for example, contain subgrains, grains, or twinned regions) or regions of different lattice parameter (for example, contain more than one phase or polytype) to be imaged in a single exposure. Analysis of Laue spot shape or Laue spot asterism (the deviation from the shape expected from an undistorted crystal) enables quantitative analysis of lattice rotation. Simultaneous measurement of the variation in diffracted wavelength as a function of position in the Laue spot—for example, using a solid-state detector—also enables determination of any variation in *d* spacing of the particular diffracting planes. For those crystals that contain several regions of distinctly different orientation, so-called orientation contrast becomes evident whereby the two neighboring regions of crystal separated by the boundary (twin boundary or grain boundary, for example) give rise to diffracted beams that travel in different directions in space and to image shifts on the detector. Analysis of the directions of these diffracted beams from the measured image shifts can enable the orientation relationships between the two regions of crystal to be established, leading to, for example, determination of twin laws. An example of a diffraction pattern recorded from a ZnTe single crystal containing twins is shown in Fig. 2. The surface of the crystal is (111). Detailed analysis of the orientation relationships between the segments of the various diffraction spots enables the twin operation to be defined as a 180° rotation about [1$\bar{1}$1].

By comparison, in the conventional topographic case, the maximum tolerable range of misorientation is defined by the convolution of characteristic linewidth with the beam divergence. Since both of these quantities are small (typical linewidths are of the order of 10^{-3} Å, equivalent to an angular divergence of approximately 100 arc sec, and typically, beam divergences must be less than approximately 20 arc sec, in order to obtain the necessary angular resolution in the plane of incidence), misorientations greater than a few arc seconds lead to the situation where only part of the crystal fulfills the diffracting condition at a given time. In other words, Bragg contours are produced. These contours delineate those regions of the crystal that are in the diffracting condition from those that are not. Since their presence means that there are regions not set for diffraction, effectively there are blind spots in the crystal. Clearly this is a very undesirable situation if dynamic-type studies in large single crystals are to be conducted. Similar problems are encountered in monochromatic synchrotron radiation topography.

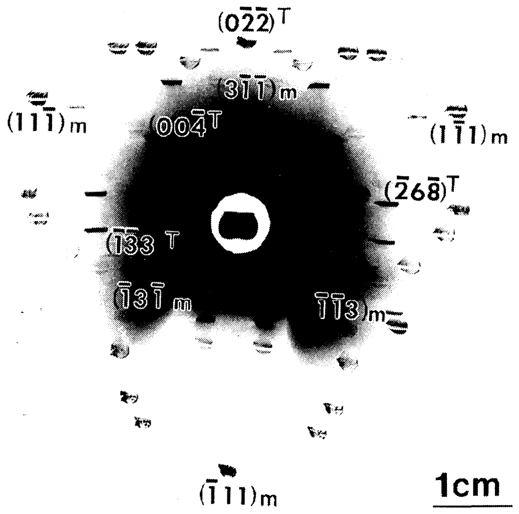

FIG. 2. Synchrotron white-beam diffraction pattern, recorded in transmission geometry (specimen–film distance 10 cm), from a twinned zinc telluride crystal. The pattern is partially indexed, with subscripts **m** and superscripts **T** indicating reflections from matrix and twin regions, respectively.

Contrast mechanisms operative in white-beam topography are largely similar to those that are operative in Lang topography (see Sec. 5), with some notable subtle differences (Hart, 1975). Some of these differences are due to the fact that in white-beam topography one can have the added complication of harmonic contamination (the presence of several orders of diffraction in a single diffraction spot), which may lead to a situation where more than one contrast mechanism operates in a single topograph. Harmonic contamination can, however, be avoided by judicious choice of diffraction geometry. In the following section a review of basic contrast mechanisms, pertinent to both Lang and white-beam topography, is presented.

4. MECHANISMS OF CONTRAST FORMATION

Apart from those defects that separate domains having different orientation, so that the domains themselves give rise to orientation contrast of the kind illustrated in Fig. 2 (simply arising from regions of differing orientation giving rise to diffracted beams that travel in different directions), the kinds of defects that can be imaged in x-ray topography are basically restricted to those that give rise to measurable differences in diffracted intensity. These include those giving rise to localized strains—for example, dislocations and precipitates—and those consisting of boundaries that either have a localized strain field associated with them or give rise to phase shifts in the x-ray wave fields propagating in the crystal—for example, stacking faults, rotational twin boundaries, mirror twin boundaries, subgrain and grain boundaries, and growth-sector boundaries. Among the defects not observable by x-ray topography are point defects.

For those defects having a strain field associated with them, the mechanism of contrast formation is a sensitive function of the diffraction conditions. It is customary to classify such contrast mechanisms according to the absorption regime within which the experiment is being conducted. Absorption conditions are usually defined by determining the product of the linear absorption coefficient μ and the thickness of crystal, t, traversed by the x-ray beam, or simply μt. Generally three regimes of absorption are identified: low, intermediate, and high, corresponding to μt values of <1–2, 3–9, and ≥ 10, respectively. To appreciate contrast formation mechanisms in the intermediate- and high-absorption regimes, an understanding of the dynamical theory of x-ray diffraction is needed. Unfortunately, a detailed treatment of dynamical theory is beyond the scope of this article. Therefore, discussions of contrast formation mechanisms will be restricted to the low absorption regime. Detailed discussion of the other regimes can be found in Tanner (1976).

Topographic contrast mechanisms under low-absorption conditions are by far the simplest to understand, and it is for this reason that topographers often strive to work within this regime, preferring to avoid the more complicated dynamical effects that can arise under higher-absorption conditions. The main mechanism of contrast formation under these conditions is related to the presence or absence of primary extinction effects. As briefly mentioned earlier, in regions of crystal that diffract dynamically, primary extinction effects are operative, whereas in regions that diffract kinematically these effects are absent. The phase shifts associated with primary extinction effects lead to an overall reduction in diffracted intensity, so that the absence of primary extinction effectively leads to enhanced diffracted intensities. Consequently, if the mode of diffraction—dynamical or kinematical—varies as a function of location in the crystal, contrast should become evident according to these differences in diffracted intensity.

By way of illustration consider the distortion field around a dislocation in an otherwise perfect single crystal. It has been shown (Authier, 1967, 1970) that there is a region of crystal that surrounds the dislocation line where the distortion field is such that dynamical diffraction, along with its associated extinction effects, is no longer operative. Rather, the region behaves like a miniature mosaic crystal that is embedded in the parent crystal and that diffracts kinematically. This effect can be quantified as follows. When a crystal diffracts dynamically, it does so over a finite range of incidence angles δ, known as the perfect-crystal rocking-curve width. This range is determined by the particular diffraction conditions and is typically of the order of a few seconds of arc. Usually the divergence of the incident beam is somewhat greater than this, which means that typically only a small proportion of a given incident beam will actually undergo diffraction, with most of the incident beam passing straight through the crystal, and simply undergoing normal photoelectric absorption. However, if the undiffracted portion of the incident beam should encounter distorted regions present inside the crystal, it becomes possible that these regions be set at the correct orientation for diffraction of this portion of the beam, provided that their misorientation, or what the x-ray beam detects as a misorientation (see next paragraph), is not greater than the incident-beam divergence.

Basically, in diffraction of a monochromatic x-ray beam, there are two ways a distortion field can influence the local diffracting conditions: it can either change the spacing of the diffracting planes or change the orientation of the diffracting planes, or both. Both of these translate, as far as the incident beam is concerned, into a misorientation, as illustrated in the differential form of Braggs law,

$$\Delta\theta = \frac{\delta d}{d}\tan\theta \pm \omega, \tag{3}$$

where ω is the contribution of lattice rotation to the misorientation.

It is usual to define this misorientation with respect to some reference orientation, such as the condition for exact fulfillment of the Bragg condition. This latter misorientation is referred to as the deviation from the exact Bragg condition, $\Delta\theta$. Generally an incident beam will not initially be set exactly in fulfillment of the Bragg condition, i.e., it will start out in the perfect region of crystal with a finite deviation. Subsequently, a distortion field may give rise to a further deviation, and it is then customary to define an "effective misorientation," $\delta(\Delta\theta)$ (Authier, 1967), which is simply the change in the deviation from exact Bragg condition that is encountered by the incident beam as it traverses the crystal. If this effective misorientation is larger than the perfect-crystal rocking-curve width, yet smaller than the effective beam divergence, then the distorted region will give rise to a new diffracted beam. Further, if the distorted region is small in size, as is the case for the volume of interest surrounding the dislocation, then the region itself should diffract kinematically. In addition, since the distorted region is misoriented with respect to the parent crystal, the resulting diffracted beam will not be at the correct orientation with respect to the relevant perfect-crystal planes to undergo the primary extinction effects that are otherwise operative. As a result the diffracted beam that emanates from the distorted volume of crystal will only undergo normal photoelectric absorption on passing through the crystal and will not suffer the effective enhanced absorption associated with extinction effects to which the diffracted beams from the perfect region of crystal are subjected. Hence there is an enhancement of diffracted intensity from this volume compared to the rest of the crystal, or topographic contrast. This is known as direct or kinematical image formation. Note that in thicker crystals (μt > 3–4) the diffracted beam from the distorted region will not survive the photoelectric absorption in the crystal, and thus the effect disappears.

It should be noted that in the white-beam technique, contributions from slightly different wavelengths have to be taken into account in considering direct image formation. This requires a modification of Authier's effective misorientation formula to take energy dispersion into account. This leads to the definition of a different parameter, the "effective deviation," which can then be used to predict the size and shape of direct dislocation images on white-beam topographs (for details see Dudley *et al.*, 1995). The requirement that the divergence of the incident beam be greater than the inherent rocking-curve width of the crystal for direct image formation to be possible may not be fulfilled in some monochromatic techniques (conventional or synchrotron) where the incident beam is of very low divergence. In such cases, no direct image formation is expected, even under low μt conditions.

5. EXAMPLES OF APPLICATIONS OF X-RAY TOPOGRAPHY

5.1 Dislocation Analysis

Topography, both synchrotron and conventional, is well suited for analysis of low densities ($<10^6$ cm^{-2}) of dislocations in crystals. The restriction to low densities arises from the fact that topographic dislocation images can be anywhere from around 5 to around 15 μm wide, so that greater densities would lead to image overlap and therefore loss of information.

5.1.1 Determination of Line Direction For dislocations created by slip processes, knowledge of the line direction as well as detailed information on the Burgers vector of the dislocations is required to assign the active slip system fully. Knowledge of both line direction and Burgers vector of crystal-growth–induced dislocations is also very important in understanding their origin and for the development of strategies to reduce the density of such dislocations. The projected directions of direct images of growth dislocations have also been used very successfully to compare with line-energy calculations designed to determine why particular line directions are preferred by such dislocations in crystals (for review see Klapper, 1980).

The line direction of a dislocation can be obtained by analyzing its direction of projection on two or more topographs recorded with different reciprocal lattice vectors. The use of analytical geometry enables the line direction to be determined either directly from the measured direction of projection (see Yuan and Dudley, 1992) or indirectly by comparing calculated projected directions of expected dislocation line directions for the material of interest with the measured projection directions (see Miltat and Dudley, 1980). Such analysis is most readily carried out on direct images of dislocations, since they are the most well defined, although similar analysis can be performed on dynamical or intermediary images.

5.1.2 Determination of Burgers-Vector Direction For sufficiently low dislocation densities, standard Burgers-vector analysis, which enables the determination of the direction of the Burgers vector, is readily carried out in the low-absorption regime, using the criterion $\mathbf{g}\cdot\mathbf{b} = 0$ for invisibility of screw dislocations, and the combination of $\mathbf{g}\cdot\mathbf{b} = 0$ and $\mathbf{g}\cdot\mathbf{b}\times\mathbf{l} = 0$ criteria for invisibility of edge or mixed dislocations (where $\mathbf{b}$ is the dislocation Burgers vector, and $\mathbf{l}$ is the dislocation line direction). These criteria are also used in the analysis of transmission electron microscopy images of dislocations. An example of dislocation analysis in a single crystal of silicon is presented in Fig. 3. The dislocations shown in these figures form slip bands and were introduced as a result of bending stresses experienced during rapid thermal annealing. The surface orientation of the crystal is (111), and the dislocations shown lie in the $\bar{1}11$ slip plane. Three kinds of dislocation segment are visible in Fig. 3(a) ($\mathbf{g} = 11\bar{1}$, $\lambda = 0.68$ Å), labeled **a**, **b**, and **c**. Line-direction analysis (Miltat and Dudley, 1980) shows that segment **a** has line direction [110], while segments **b** and **c** have line directions $[0\bar{1}1]$ and [101], respectively. Segment **a** is clearly visible on the $11\bar{1}$ reflection [Fig. 3(a)], but invisible on the $\bar{1}11$ and $1\bar{1}1$ reflections [Figs. 3(b) and 3(c), respectively]. Application of the criterion $\mathbf{g}\cdot\mathbf{b} = 0$ to possible $\frac{1}{2}\langle 110\rangle$ type Burgers vectors, which lie on the $\bar{1}11$ slip plane, shows that segment **a** has Burgers vector $\frac{1}{2}[110]$, so that it is pure screw. Segments **b** and **c** also show weak contrast on the $\bar{1}11$ and $1\bar{1}1$ reflections and are also clearly part of the same dislocation loop as segments of kind **a**, so that they also must have Burgers vector $\frac{1}{2}[110]$, and are thus of 60° character (i.e., the Burgers vector makes a 60° angle with the respective line directions). Minimum contrast for the 60° segments is obtained on reflections for which the criteria $\mathbf{g}\cdot\mathbf{b} = 0$ and $\mathbf{g}\cdot\mathbf{b}\times\mathbf{l} = 0$ are both satisfied, which in this case corresponds to $\mathbf{g} = 2\bar{2}4$. Note that the incomplete disappearance of segments **a** in Figs. 3(b) and 3(c) is due to a modification of the dislocation strain field at its intersection with the crystal surface due to the effects of surface relaxation (Wu *et al.*, 1992).

5.1.3 Determination of Burgers-Vector Sense and Magnitude The determination of the actual sense and magnitude of the Burgers vector of a dislocation once the Burgers-vector direction has been determined by $\mathbf{g}\cdot\mathbf{b}$ analysis requires more detailed analysis. Chikawa (1965) developed conventional-radiation, divergent-beam techniques for the determination of the sense of both edge and screw dislocations through measurement of the sense of tilt of lattice planes surrounding the dislocation core. For the case of screw dislocations, Mardix *et al.* (1971) subsequently further developed Chikawa's divergent-beam technique to enable the magnitude of this tilt to be measured as a function of distance from the core. This could then be fitted to the corresponding theoretical expression, enabling the Burgers-vector magnitude to be determined. Si *et al.* (1996) and Dudley *et al.* (1997) recently developed two methods, one section and one projection, that are analogous to those of Chikawa and of Mardix *et al.*, but that make use of synchrotron white radiation. In these techniques the lattice tilt surrounding dislo-

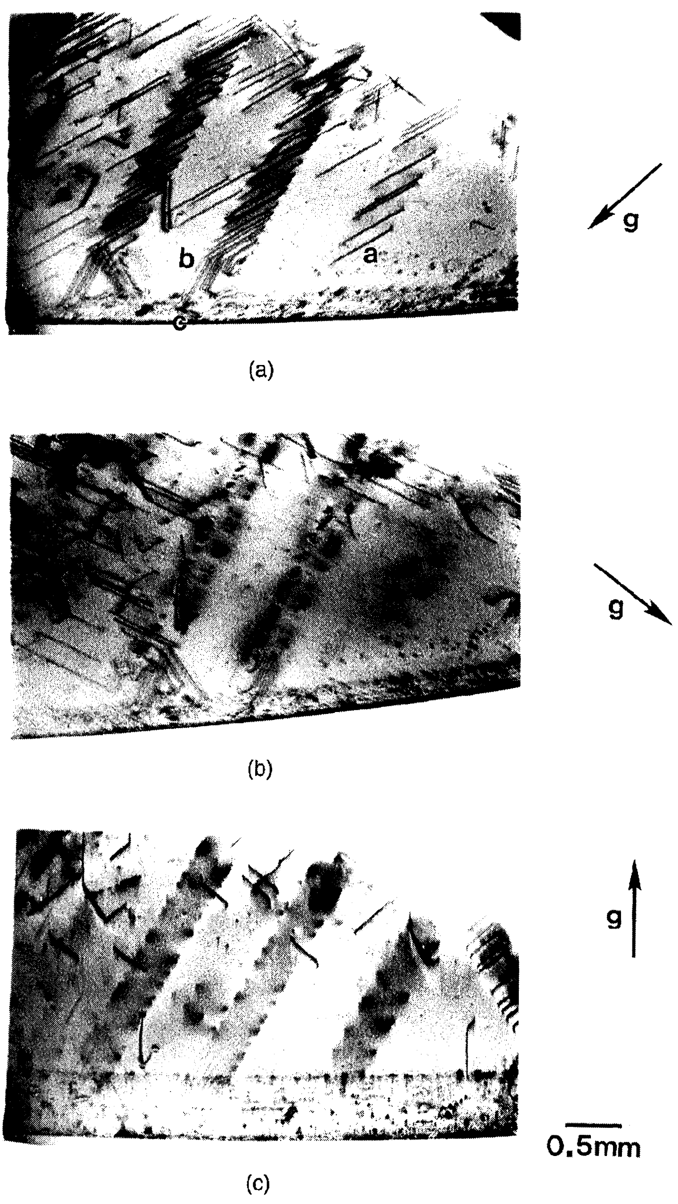

FIG. 3. Synchrotron white-beam topographs, taken from a single diffraction pattern recorded in transmission geometry (specimen–film distance 10 cm), from a silicon single crystal containing slip dislocations: (a) **g** = $11\bar{1}$, λ = 0.68 Å; (b) **g** = $\bar{1}11$, λ = 0.68 Å; (c) **g** = $1\bar{1}1$, λ = 0.68 Å. Note the disappearance of segments **a**, and the weak contrast from segments **b** and **c**, in (b) and (c).

cations of mainly screw character was measured with an accuracy that was significantly improved on account of the relaxation of the requirement for short specimen–film distances that is inevitable in conventional-radiation techniques.

A recent study of individual superscrew dislocations in 6H-SiC using synchrotron white-beam x-ray topography serves well to illustrate the principles of the two techniques. Let us consider first the projection technique, which enables the magnitude of the Burgers vector to be determined from the magnitude of the tilt of the lattice planes on both sides of the dislocation core. This simple method can be illustrated as follows: The lattice tilt ϕ around the screw dislocation, at distance r from the dislocation core, can be described by $\phi = b/2\pi r$, where b is the magnitude of the Burgers vector. In projection topographs ($\mathbf{g} = 0006$, $\lambda = 0.89$ Å) such as those shown in Figs. 4(a) and 4(b), this lattice tilt manifests itself in the form of a mutual longitudinal shift S of the two parallel components of the bimodal dislocation images (black lines), which is observable at the outcrop of the dislocations on the as-grown surface. By changing the sample-to-film distance D_{sf}, ϕ can be determined from the slope k of the linear plot of S versus D_{sf} (where $k = 4\phi$), which is shown in Fig. 4(c). Since the superscrew dislocations in this material always have groups of basal-plane dislocations attached to them, and these basal-plane dislocations have edge components, the bimodal image width of the superscrew dislocations diverges as D_{sf} increases. In order to obtain the value of r corresponding to the measured tilt angle ϕ, the width of the bimodal image is plotted as a function of D_{sf} as shown in Fig. 4(d). The image width W_0 at $D_{sf} = 0$, obtained by extrapolation, corresponds to $2r$. This enables the magnitude of the Burgers vector to be determined from the equation $b = \pi W_0 k/4$. For the dislocation shown in Figs. 4(a)–4(d), the Burgers-vector magnitude was found to be 52.6 Å, corresponding to (3–4)c (where c is the lattice constant along [0001] axis, $c = 15.17$ Å), and the sense is left-handed.

As for the section technique, this method consists of illuminating the sample by a strip of highly collimated white beam (less than 20 μm across). In this case, section topographs were recorded in symmetric transmission, i.e., using Bragg planes perpendicular to the thin crystal platelet, but that were also perpendicular to the screw dislocation.

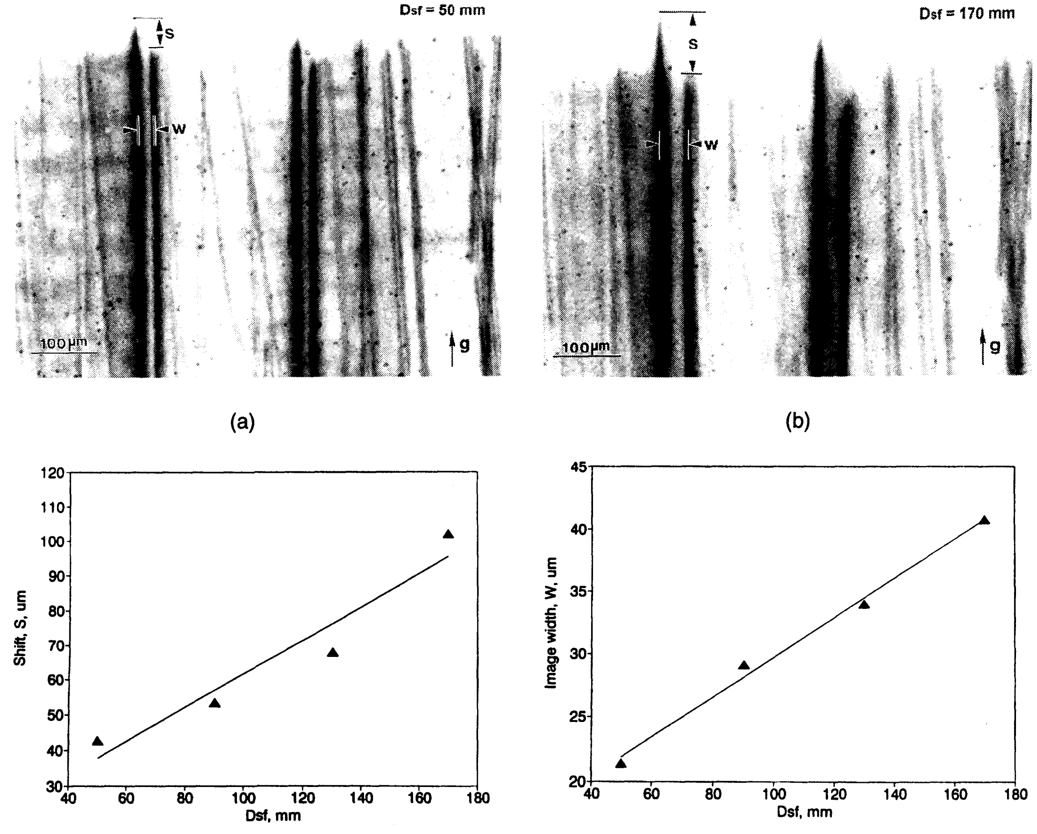

(a) (b)

(c) (d)

FIG. 4. The projection topographic determination of the Burgers vector of superscrew dislocations in 6H-SiC: (a) projection topograph, $\mathbf{g} = 0006$, $\lambda = 0.89$ Å, $D_{sf} = 5$ cm; (b) projection topograph, $\mathbf{g} = 0006$, $\lambda = 0.89$ Å, $D_{sf} = 17$ cm; (c) bimodal image shift S vs D_{sf}; and (d) bimodal image width W vs D_{sf}.

In such cases the lattice tilts are easily derived from the linear displacements of Bragg-diffracted beams. In principle, the lattice tilt ϕ around the screw dislocation can be described as $\phi = b/2\pi r$, where b is the magnitude of the Burgers vector, r is the distance from the dislocation core. As to the white-beam diffraction geometry, the deviation of Bragg angle $\Delta\theta_B$ would be 2ϕ. Assuming the incident beam is strictly parallel, the displacement Y recorded in the x-ray film at sample-to-film distance D_{sf} would be $Y \approx D_{sf}\Delta\theta_B$ ($\Delta\theta_B$ is very small) $= 2\ D_{sf}\phi$, thus $b = \pi Yr/D_{sf}$.

Figure 5(a) shows the lattice tilt ϕ around a screw dislocation, and Fig. 5(b) shows a white-beam section topograph of a screw dislocation, D_{sf} = 16 cm, **g** = 0006, λ = 0.89 Å. From the hyperbolas drawn on the pictures, the magnitude of the Burgers vector is 62.6 Å (upper left lobe), and 45.0 Å (lower right lobe), which again corresponds to $3c$ to $4c$ (c is the lattice constant along the [0001] axis, c = 15.17 Å).

5.2 In Situ Studies of Crystals Subjected to External Stimuli

The high intensity of synchrotron radiation sources and the subsequently short exposure times (coupled with high signal-to-noise ratios) in synchrotron white-beam x-ray topography open up the possibility of conducting such experiments in a dynamic fashion. For example, images can be recorded sequentially while crystals are being subjected to some kind of external stimulus such as heat, magnetic field, electric field, etc. Alternatively, such images can be recorded in a quasi real-time fashion, using, for example, charge-coupled device (CCD) detectors (see CHARGE-COUPLED DEVICES). In either case, information can be gathered *in situ* while the crystals are undergoing, for example, plastic deformation (see, for example, Dudley *et al.*, 1984; Wu and Dudley, 1993; Liu *et al.*, 1993, 1995a, 1995b), chemical reactions (for review, see Dudley *et al.*, 1991), or phase transitions (see, for example, Dudley and Yao, 1993). Figures 6 and 7(a)–7(e) show an example of images recorded both before and during plastic deformation, in this case induced by compressive loading of an ice bicrystal. The dislocation structures originally present, which are generated at facets $\mathbf{F_i}$, are seen to be modified, in particular via the generation and propagation of $[11\bar{2}0]$ line direction screw dislocation segments [see Liu *et al.* 1995a,b for details]. Figures 8(a)–8(d) and 9(a)–9(d) show an example of images recorded as a crystal are cycled through a phase transition, in this case in a $LaAlO_3$ crystal, which undergoes a second-order rhombohedral-to-cubic phase transition on heating through 450°C. The microstructure of the room-temperature rhom-

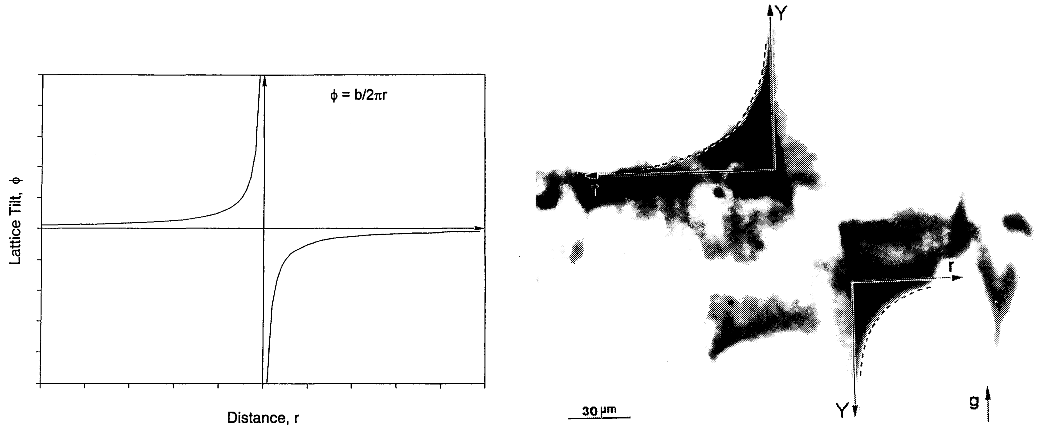

(a) (b)

FIG. 5. The section topographic determination of the Burgers vector of superscrew dislocations in 6H-SiC. (a) Schematic diagram of the lattice tilt around a screw dislocation, and (b) a section topograph of a superscrew dislocation, **g** = 0006, λ = 0.89 Å, D_{sf} = 16 cm.

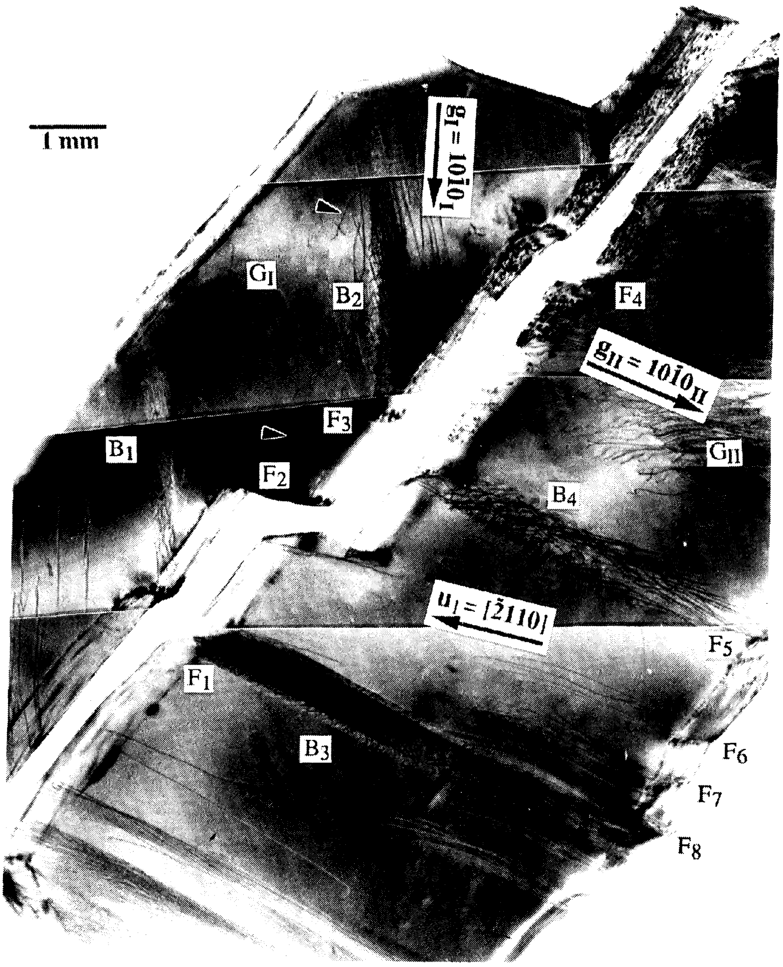

FIG. 6. Synchrotron white-beam transmission topograph recorded from an ice bicrystal prior to applied compressive deformation. Note the dislocation structures that have been generated from facets F_I.

bohedral phase comprises a preponderance of twins. These twins are seen to disappear gradually as the rhombohedral structure gradually distorts to the high-temperature cubic structure [see Dudley and Yao (1993) for details].

6. COMPARISON WITH TRANSMISSION ELECTRON MICROSCOPY

X-ray topographic imaging is in many ways complementary to transmission electron microscopy (TEM) imaging. The principles of dark-field TEM are indeed close to those of x-ray topography. Generally speaking, both x-ray topography and TEM rely on an understanding of dynamical diffraction theory for the interpretation of contrast. Both are capable of detailed analysis of defect structures in crystals. X-ray topography, on the one hand, is capable of analyzing defects, nondestructively, in very large single crystals [wafers as large as 12 in. (25 cm) in diameter can be routinely examined, with thicknesses ranging from a few hundred microns to greater than 1 mm, depending on the x-ray absorption, for transmission geometry, and unlimited thicknesses for reflection geometry]. TEM, on the other hand, can only examine much smaller-dimension sections of either polycrystals or single crystals. Thicknesses of less than one thousand angstroms are typically required to achieve electron

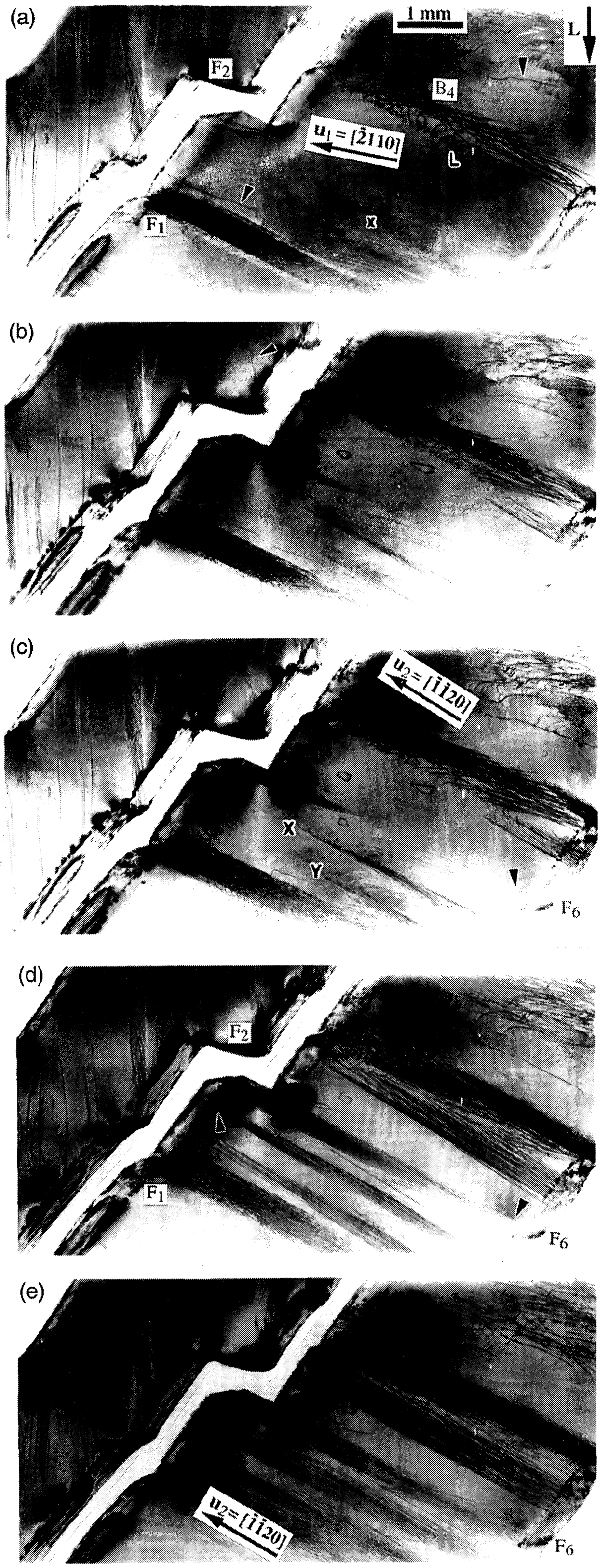
(a)
1 mm
L
F2
B4
$u_1 = [\bar{2}110]$
L
F1
X
(b)
(c)
$u_2 = [\bar{1}\bar{1}20]$
X
Y
F6
(d)
F2
F1
F6
(e)
$u_2 = [\bar{1}\bar{1}20]$
F6

FIG. 7. Sequence of images recorded from the central region of the crystal shown in Fig. 6 showing dislocation structures generated during *in situ* compressive deformation: (a) following loading of 0.1 MPa for 5 min; (b) after further loading of 0.2 MPa for 10 min; (c) recorded 5 min later without loading; (d) after further loading of 0.2 MPa for 20 min; (e) after further loading of 0.2 MPa for 25 min. Of particular note are the dislocations generated at the facet $\mathbf{F_7}$ [visible on Fig. 6, but to the lower right of Figs. 7(a)–7(e), just out of the field of view] by grain-boundary sliding. For details see Liu *et al.* (1995).

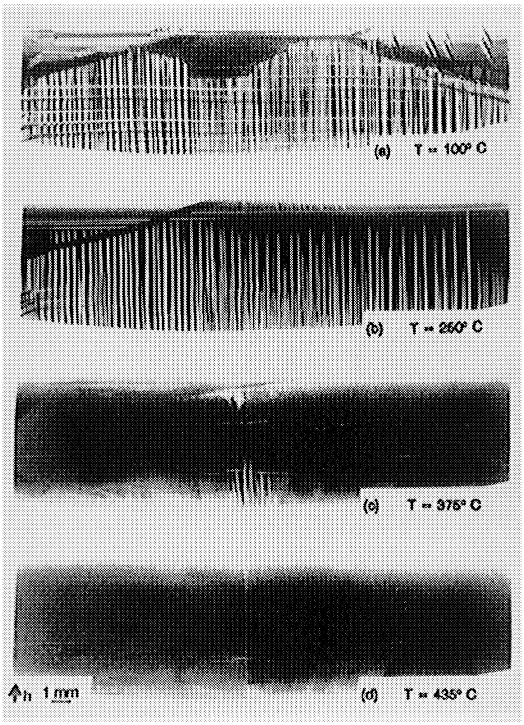

FIG. 8. Series of synchrotron white-beam x-ray reflection topographs recorded with increasing temperature from an $LAlO_3$ crystal as it undergoes a rhombohedral-to-cubic phase transition. Note the gradual disappearance of the twins, which are visible via orientation contrast.

transparency, while the fields of view can range from a few millimeters at low magnification to tens of angstroms at high magnification.

The relatively broad dislocation image width in x-ray topography (typically 5–10 μm), which derives from the high strain sensitivity of the technique, means that image overlap occurs for dislocation densities in the vicinity of 10^6 cm^{-2}. On the other hand, the very narrow dislocation image width in TEM (typically around 5 nm) means that dislocations cannot be resolved at low densities. This also means that dislocation densities in excess of 10^9 cm^{-2} are required to make TEM analysis realistically feasible.

GLOSSARY

Direct Dislocation Image: Dislocation image whose contrast arises from the difference in diffracted intensity between the distorted region immediately surrounding the dislocation, which diffracts kinematically, and the surrounding perfect regions of crystal, which diffract dynamically.

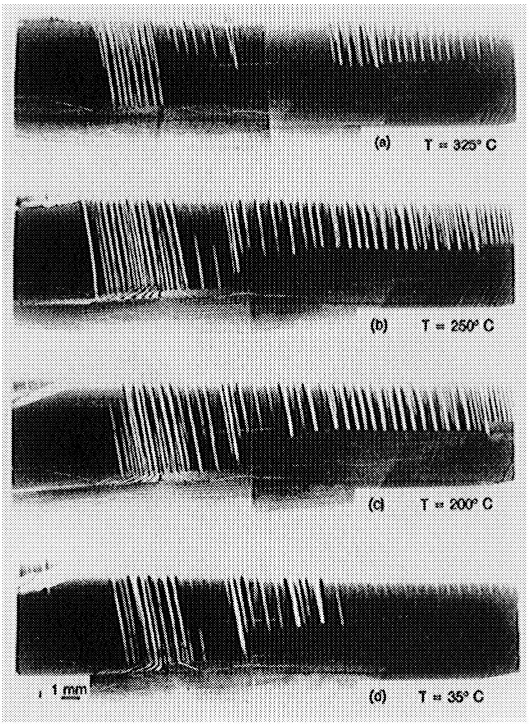

FIG. 9. Series similar to Fig. 8, recorded with decreasing temperature. For details see Dudley and Yao (1993).

Dynamical Diffraction: The kind of x-ray diffraction produced by nearly perfect crystals wherein primary extinction effects are present.

Kinematical Diffraction: The kind of x-ray diffraction produced by ideally imperfect crystals or regions of crystal wherein primary extinction effects are absent.

Mosaic Crystal: Crystal made up of small subgrains that are mutually misoriented by a small angle.

Mosaicity: The range of misorientations present in a mosaic crystal.

Orientation Contrast: Contrast resulting from the overlap or separation of parts of an x-ray topographic image arising from misorientations between adjacent regions of crystal.

Primary Extinction: Phenomenon encountered during diffraction of x rays by a nearly perfect crystal whereby interference between scattered and incident rays inside the crystal leads to an overall reduction in diffracted intensity.

X-Ray Topographic Contrast: Local intensity differences within an x-ray topographic image.

Works Cited

Authier, A. (1967), *Adv. X-ray Anal.* **10,** 9–31.

Authier, A. (1970), in: S. Amelinckx, R. Gevers, G. Remaut, J. Van Landuyt (Eds.), *Modern Diffraction and Imaging Techniques in Materials Science,* Amsterdam: North-Holland, pp. 481–520.

Bowen, D. K., Hall, C. R. (1975), *Microscopy of Materials,* London: Macmillan.

Chikawa, J.-I. (1965), *J. Appl. Phys.* **36,** 3496–3502.

Dudley, M., Yao, G.-D. (1993), *J. Phys. D: Appl. Phys.* **26,** 120–125.

Dudley, M., Miltat, J., Bowen, D. K. (1984), *Philos. Mag. A* **50,** 487–504.

Dudley, M., Sherwood, J. N., Bloor, D. (1991), *Proc. R. Soc. London A* **434,** 243–261.

Dudley, M., Wang, S., Huang, W., Carter, C. H., Jr., Tsvetkov, V. F., Fazi, C. (1995), *J. Phys. D: Appl. Phys.* **28,** A63–A68.

Dudley, M., Si, W., Wang, S., Carter, C. H., Jr., Glass, R., Tsvetkov, V. F. (1997), *Nuovo Cimento* **19,** in press.

Hart, M. (1975), *J. Appl. Crystallogr.* **8,** 436–444.

Klapper, H. (1980), in: B. K. Tanner, D. K. Bowen (Eds.), *Characterization of Crystal Growth Defects by X-ray Methods.* NATO Advanced Study Institutes Series B, Physics, No. 63, New York: Plenum, pp. 133–160.

Liu, F., Baker, I., Dudley, M. (1993), *Philos. Mag. A* **67,** 1261–1276.

Liu, F., Baker, I., Dudley, M. (1995a), *Philos. Mag. A* **71,** 15–42.

Liu, F., Baker, I., Dudley, M. (1995b), *Philos. Mag. A* **71,** 1–14.

Mardix, S., Lang, A. R., Blech, I. (1971), *Philos. Mag.* **24,** 683–693.

Miltat, J. (1980), in: B. K. Tanner, D. K. Bowen (Eds.), *Characterization of Crystal Growth Defects by X-ray Methods.* NATO Advanced Study Institutes Series B, Physics, No. 63, New York: Plenum, pp. 401–420.

Miltat, J., Dudley, M. (1980), *J. Appl. Crystallogr.* **13,** 555–562.

Si, W., Dudley, M., Carter, C. H., Jr., Glass, R., Tsvetkov, V. F. (1996) in: L. Terminello, S. Mini, D. L. Perry, H. Ade (Eds.), *Applications of Synchrotron Radiation to Materials Science,* Materials Research Society Symposium Proceedings Vol. 437, Pittsburgh: Materials Research Society, pp. 129–134.

Tanner, B. K. (1976), *X-ray Diffraction Topography,* Oxford: Pergamon.

Tuomi, T., Naukkarinen, K., Rabe, P. (1974), *Phys. Status Solidi A* **25,** 93–106.

White, J. E. (1950), *J. Appl. Phys.* **21,** 855–859.

Wu., J., Dudley, M. (1993), in: *Applications of Synchrotron Radiation Techniques to Materials Science,* Materials Research Society Symposium Proceedings, Vol. 307, Pittsburgh: Materials Research Society, pp. 231–236.

Wu, J., Fanning, T., Dudley, M., Shastry, V., Anderson, P. (1992), in: S. Ashok, J. Chevallier, K. Sumino, E. Weber (Eds.), *Defect Engineering in Semiconductor Growth, Processing and Device Technology,* Materials Research Society Symposium Proceedings Vol. 262, Pittsburgh: Materials Research Society, pp. 265–270.

Yuan, D., Dudley, M. (1992), *Mol. Cryst. Liq. Cryst.* **211,** 51–58.

Further Reading

Authier, A. (1967), *Adv. X-ray Anal.* **10,** 9–31.

Authier, A. (1970), in: S. Amelinckx, R. Gevers, G. Remaut, J. Van Landuyt (Eds.), *Modern Diffraction and Imaging Techniques in Materials Science,* Amsterdam: North-Holland, pp. 481–520.

Bowen, D. K., Hall, C. R. (1975), *Microscopy of Materials,* London: Macmillan.

Miltat, J. (1980), in: B. K. Tanner, D. K. Bowen (Eds.), *Characterization of Crystal Growth Defects by X-ray Methods,* NATO Advanced Study Institutes Series B, Physics, No. 63, New York: Plenum, pp. 401–420.

Tanner, B. K. (1976), *X-ray Diffraction Topography,* Oxford: Pergamon.

TOPOLOGICAL PHASE EFFECTS

J. M. Robbins, *Basic Research Institute in the Mathematical Sciences, Hewlett-Packard Laboratories Bristol, and Department of Mathematics, University of Bristol, Bristol, United Kingdom*

Introduction 549
1. **Basic Account of the Geometric Phase** 550
1.1 Quantum Adiabatic Theorem 550
1.2 The Geometric Phase 551
1.3 Two-State Systems 552
1.4 Degeneracies 553
1.5 Time-Reversal Symmetry 554
1.6 Experimental Observation 554
1.7 Geometric Magnetism 555
2. **Theoretical Developments** 556
2.1 Geometric Phase as Holonomy 556
2.1.1 Fiber Bundles 557
2.1.2 Connections and Curvature 557
2.1.3 Geometric Phase Revisited 558
2.1.4 Chern Classes 558
2.2 Degenerate Eigenstates and Non-Abelian Geometric Phases 559
2.3 Nonadiabatic Evolution 561
2.4 Geometric Metric Tensor 563
2.5 Beyond Adiabaticity 564
2.5.1 Corrections to the Geometric Phase 564
2.5.2 The Geometric Amplitude 565
3. **Classical Systems** 566
3.1 The Hannay Angle 566
3.1.1 The Classical Adiabatic Theorem 566
3.1.2 Basic Account of the Hannay Angle 567
3.1.3 Examples 568
3.1.4 Semiclassical Limit 568
3.1.5 Classical Reaction Forces 569
3.2 Symmetry and Kinematic Geometric Phases 569
3.2.1 Falling Cats and Swimming Paramecia 570
3.2.2 Geometric Phases and Symmetry Reduction 571
3.2.3 Optimization and Control 572
4. **Optics** 573
4.1 Geometric Phase of Coiled Light 573
4.2 Pancharatnam's Phase 574
5. **Molecular and Condensed-Matter Physics** 576
5.1 Pseudorotation in Triatomic Molecules 576
5.2 Nuclear Magnetic Resonance 577
5.3 The Quantum Hall Effect 579
Glossary 582
Works Cited 582
Further Reading 584

INTRODUCTION

In an article published in 1984, Sir Michael Berry (Berry, 1984) found that a quantum system undergoing adiabatic evolution acquires a phase factor of purely geometrical origin. This discovery, now called the Berry phase or the geometric phase, initiated a tremendous surge of research in a range of disciplines across physics. Berry and subsequent workers quickly established connections to a number of apparently disparate phenomena, including among others the Aharonov–Bohm effect (Aharonov and Bohm, 1959), its molecular physics analog (Mead and Truhlar, 1979), polarization optics (Pancharatnam, 1956), the quantum Hall effect (Thouless *et al.*, 1982), and the Foucault pendulum (Eco, 1989). Beyond providing a unified description of these phenomena, the geometric phase led to predictions of new effects, subsequently observed experimentally, and a number of substantial extensions and generalizations. Among these, the discovery by Hannay (1985) of an analogous effect in classical mechanics, the Han-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

nay angle, stimulated much new research in purely classical physics.

The mathematical phenomenon that underlies the geometric phase is holonomy. Holonomy describes transformations in a quantity induced by cyclic changes in the variables that control it. An example from geometry is the rotation of vectors parallel transported along closed paths on a curved surface. The mathematics of holonomy and curvature has long played a role in physics through general relativity and quantum field theory. With Berry's work came the appreciation of a broader scope of application. Whenever a physical system is divided into parts and one attempts to describe a subsystem in isolation, the influence of the excluded part of the system is manifested through geometric-phase effects. The pervasiveness throughout physics of this reductive procedure accounts for the ubiquity of the geometric phase. Introducing a collection of major articles in the field, Shapere and Wilczek (1989a) wrote, "We believe that the concept of a geometric phase, repeating the history of the group concept, will eventually find so many realizations and applications in physics that it will repay study for its own sake, and become part of the *lingua franca.*"

An introductory account of the geometric phase is given in Sec. 1. This is followed in Sec. 2 by a discussion of several theoretical developments. Applications to classical systems, optics, and molecular and condensed-matter physics are given in Secs. 3–5. The topics chosen for discussion are meant to be representative rather than comprehensive. Among several omissions, we mention the theory of fractional statistics, which is covered extensively in the volume by Shapere and Wilczek (1989a), and applications to elementary-particle physics, which are reviewed by Aitchison (1987). The historical development of the subject is discussed by Berry (1990a).

1. BASIC ACCOUNT OF THE GEOMETRIC PHASE

The geometric phase was discovered by Sir Michael Berry in 1983 in the course of a critical re-examination of the adiabatic theorem in quantum mechanics. This theorem says, in essence, that if a quantum system is changed sufficiently slowly, it does not undergo transitions. Thus a system initially in an eigenstate of a slowly changing Hamiltonian remains in the evolved eigenstate. Berry considered the overall phase accumulated by the changing eigenstate and found an unexpected contribution. This extra phase is of a geometrical nature, in that it does not depend on the rate at which the adiabatic process is carried out.

One way that a slowly varying Hamiltonian can arise physically is when the system under consideration interacts with another, much slower system. This leads to the consideration of the quite general problem of two coupled systems evolving on different time scales. In this context, the geometric phase describes an effect on the fast system produced by the slow one. As action in physics is accompanied by reaction, there is a reciprocal effect on the slow system generated by the fast. This takes the form of an effective vector potential in the slow Hamiltonian; the associated Lorentz-type force is called geometric magnetism. The same geometry underlies both geometric magnetism and the geometric phase.

This geometrical structure is intimately connected to degeneracies in the energy spectrum. A typical Hamiltonian has a nondegenerate spectrum (excepting systematic degeneracies due to symmetry, which are discussed in Sec. 2.2). Varying a single parameter (e.g., time itself in the case of a time-dependent Hamiltonian) is unlikely to produce any degeneracies; typically two or three parameters (the number depends on whether time-reversal symmetry is present or not) must be adjusted independently in order to find one. Nevertheless, the presence of degeneracies in Hamiltonians that are, in some sense, near the Hamiltonians of the adiabatic process are largely and sometimes wholly responsible for the geometric phase. This connection underlies several applications, including the quantum Hall effect (Sec. 5.3).

1.1 Quantum Adiabatic Theorem

Adiabatic invariants are quantities that are nearly conserved when the environment of an isolated system is slowly changed (slowly means in comparison with the system's internal dynamics). Adiabatic theorems provide estimates as to how accurately this conservation is respected. In quantum mechanics, the adiabatic invariants are the probabilities $P_n = |\langle n|\psi\rangle|^2$ to be in energy eigenstates. These are constant for stationary Hamiltonians but vary if H is time dependent. The quantum adiabatic theorem states that as dH/dt goes to zero, so do the probabilities of transitions between eigenstates (the rate depends on how smooth dH/dt is), so that the P_n are nearly constant. Therefore, within the adiabatic approximation, dynamics under a slowly varying Hamiltonian $H(t)$ is essentially no more complicated than under a stationary one. In particular, if $|\psi\rangle$ begins in an eigenstate, it remains in one (nonadiabatic transitions are discussed in Sec. 2.5), and its evolution is determined up to a phase factor.

We shall now examine this phase. Suppose that at $t = 0$, $|\psi\rangle$ is in the nth eigenstate $|n\rangle$ (assumed to be nondegenerate) of H. For its subsequent evolution we may write

$$|\psi(t)\rangle \approx \exp\left(-\frac{i}{\hbar}\int_0^t E_n(\tau)d\tau\right)|n(t)\rangle. \tag{1}$$

In so doing we have explicitly identified part of the overall phase, the time integral of the frequency $E_n/\hbar$, which is called the *dynamical phase.* It is just the phase one would expect for a solution of the Schrödinger equation. Any remaining phase is contained implicitly in the eigenstate $|n(t)\rangle$. This phase is determined by requiring that $\langle n|(i\hbar\partial_t - H)|\psi\rangle$ vanish [it would automatically if $|\psi(t)\rangle$ were an exact solution of the Schrödinger equation]. This requirement leads to the condition

$$\langle n|\dot{n}\rangle = 0, \tag{2}$$

which determines the phase of $|n(t)\rangle$ given an initial choice at $t = 0$.

1.2 The Geometric Phase

What Berry discovered is that condition (2) yields a path-dependent phase factor. This phase factor is particularly interesting when the Hamiltonian is cycled, i.e., made to return to itself after some time. Then it depends only on the cycle itself, and not, in contrast to the dynamical phase, on the rate at which the Hamiltonian is cycled. For this reason it is called, as well as the Berry phase, the *geometric phase.*

To discuss this further, it helps to formulate the problem slightly differently. Suppose the time dependence of the Hamiltonian is expressed through its dependence on some external parameters $\mathbf{R} = (X,Y,Z)$, which are made to vary slowly in time along a path $\mathbf{R}(t)$. Then the eigenstates $|n(\mathbf{R})\rangle$ depend on $\mathbf{R}$ as well, and (assuming no degeneracies) are determined for each $\mathbf{R}$ up to an overall phase. Let us fix this phase so that it varies smoothly but otherwise arbitrarily with $\mathbf{R}$. By fixing the phase independently of the dynamics, we may no longer satisfy the condition (2) along every path $\mathbf{R}(t)$. We therefore allow for an additional phase factor $\exp[i\gamma_n(t)]$ in the adiabatic *Ansatz* (1), which becomes

$$|\psi(t)\rangle \approx \exp\left(-\frac{i}{\hbar}\int_0^t E_n(\mathbf{R}(\tau))d\tau\right) \times \exp[i\gamma_n(t)]|n(\mathbf{R}(t))\rangle. \tag{3}$$

The condition $\langle n|(i\hbar\partial_t - H)|\psi\rangle = 0$ now yields an evolution equation for γ_n. This may be written in the form $\dot{\gamma}_n(t) = -\mathbf{A}_n(\mathbf{R}(t))\cdot\dot{\mathbf{R}}(t)$, where $\mathbf{A}_n(\mathbf{R})$ is a vector field in parameter space given by

$$\mathbf{A}_n(\mathbf{R}) = \mathrm{Im}\langle n(\mathbf{R})|\nabla n(\mathbf{R})\rangle, \tag{4}$$

or equivalently, in terms of eigenfunctions $\psi_n(\mathbf{r},\mathbf{R}) = \langle \mathbf{r}|n(\mathbf{R})\rangle$, by

$$\mathbf{A}_n(\mathbf{R}) = \mathrm{Im}\int \psi_n^*(\mathbf{r},\mathbf{R})\nabla\psi_n(\mathbf{r},\mathbf{R})d^3r$$

(here ∇ acts on the parameters $\mathbf{R}$).

If the Hamiltonian is cycled around a closed curve C in parameter space, then $|\psi(t)\rangle$ acquires a geometric phase

$$\gamma_n(C) = -\int_0^T \mathbf{A}_n\cdot\dot{\mathbf{R}}dt = -\oint_C \mathbf{A}_n\cdot d\mathbf{R}. \tag{5}$$

Its expression as a line integral reveals its geometrical character; the geometric phase depends only on the cycle C, and not on the manner in which the cycle is traversed in time. Using Stokes's theorem, γ_n can be expressed as

$$\gamma_n(C) = -\int_S \mathbf{V}_n(\mathbf{R})\cdot d\mathbf{S}, \tag{6}$$

the flux of the vector field $\mathbf{V}_n = \nabla \times \mathbf{A}_n$ through a surface S in parameter space bounded by C. $\mathbf{V}_n$ is given by

$$\mathbf{V}_n = \mathrm{Im}\langle \nabla n | \times | \nabla n \rangle. \tag{7}$$

From (6) it is evident that only cycles that enclose nonzero area can have nontrivial geometric phases. For example, $\gamma_n(C)$ vanishes for cycles consisting of a path and its retrace.

It is easy to see that the geometric phase does not depend on the choice of phase for $|n(\mathbf{R})\rangle$. Indeed, under the change of phase

$$|n(\mathbf{R})\rangle \rightarrow \exp[i\chi(\mathbf{R})]|n(\mathbf{R})\rangle, \tag{8}$$

$\mathbf{A}_n$ transforms according to $\mathbf{A}_n(\mathbf{R}) \rightarrow \mathbf{A}_n(\mathbf{R}) + \nabla\chi(\mathbf{R})$; this shift in $\mathbf{A}_n$ by a perfect gradient leaves its integral around a closed loop, $\gamma_n(C)$, and its curl, $\mathbf{V}_n$, unchanged. Thus, the geometric phase is an intrinsic property of the family of eigenstates $|n(\mathbf{R})\rangle$.

It is useful to draw an analogy with electromagnetism, in which $\mathbf{A}_n$ corresponds to the vector potential and $\mathbf{V}_n$ to the magnetic field. The geometric phase γ_n corresponds to a magnetic flux. The change of phase (8) corresponds to a gauge transformation, which shifts the vector potential by a gradient and leaves the magnetic field unchanged.

Substituting the first-order perturbation expansion

$$|\nabla n\rangle = \sum_{m \neq n} \frac{\langle m|\nabla H|n\rangle}{(E_n - E_m)} |m\rangle + \langle n|\nabla n\rangle|n\rangle \tag{9}$$

into (7) leads to a useful formula,

$$\mathbf{V}_n = \mathrm{Im} \sum_{m \neq n} \frac{\langle n|\nabla H|m\rangle \times \langle m|\nabla H|n\rangle}{(E_n - E_m)^2}. \tag{10}$$

In this form, the independence of $\mathbf{V}_n$ on the eigenstate phases is manifest. Another formula with this property is

$$\mathbf{V}_n = \mathrm{Im}\,\mathrm{Tr}(\rho_n \nabla \rho_n \times \nabla \rho_n), \tag{11}$$

where $\rho_n = |n\rangle\langle n|$ is the density operator. It too is useful in applications.

One of the applications discussed by Berry (1984) is to the Aharonov–Bohm effect (Aharonov and Bohm, 1959). The Aharonov–Bohm effect describes the dependence of quantum mechanics on electromagnetic fields in regions of space that are physically inaccessible. This dependence is mediated through the vector and scalar potentials, whose relation to the electric and magnetic fields is nonlocal. The most celebrated example concerns an electron moving in the vicinity of an impenetrable perfect solenoid of infinite extent. Since outside the solenoid the magnetic field is zero, the classical mechanics of the electron is insensitive to any magnetic flux inside the solenoid. But the quantum-mechanical behavior of the electron does depend on the interior flux, as Aharonov and Bohm showed by calculating its scattering cross section (a prediction subsequently experimentally confirmed by Chambers, 1960).

The Aharonov–Bohm effect can be understood in terms of interference between de Broglie waves passing on opposite sides of the solenoid. The phase difference between the two waves includes the integral

$$\gamma_{\mathrm{AB}} = -\frac{e}{\hbar c} \oint_C \mathbf{A}^{(s)}(\mathbf{R}) \cdot d\mathbf{R} \tag{12}$$

of the vector potential $\mathbf{A}^{(s)}$ of the solenoid around a loop C enclosing it. [In (12), c is the speed of light and $-e$ is the electron charge.] The Aharonov–Bohm phase γ_{AB} is just 2π times the magnetic flux in the solenoid measured in units of the flux quantum hc/e.

It may be regarded as a geometric phase in the following manner. We imagine the electron confined to a large box, centered at $\mathbf{R}$, which is not penetrated by the solenoid. The eigenstates of the electron $|n(\mathbf{R})\rangle$ depend parametrically on $\mathbf{R}$; specifically, changes in $\mathbf{R}$ induce gauge transformations in the eigenfunctions. One finds that $\mathrm{Im}\langle n(\mathbf{R})|\nabla n(\mathbf{R})\rangle = (e/\hbar c)\mathbf{A}^{(s)}(\mathbf{R})$, so that in this case the geometric vector potential (4) coincides with the physical vector potential of the solenoid. On being transported (not necessarily adiabatically in this case) around the solenoid, the electron eigenstates acquire a geometric phase, which is precisely the Aharonov–Bohm phase.

1.3 Two-State Systems

Two-state systems provide a canonical example of the geometric phase. The requisite calculations are easily carried out, and the main features of the general case are represented. The parametrized family of Hamiltonians can be written in the form

$$H(\mathbf{R}) = E(\mathbf{R})\mathbf{I} + \mathbf{F}(\mathbf{R})\cdot\boldsymbol{\sigma}, \tag{13}$$

where **I** is the two-by-two identity matrix and $\boldsymbol{\sigma}$ are the Pauli spin matrices. The term $E(\mathbf{R})\mathbf{I}$ shifts the energy by a constant and affects neither the eigenstates nor their geometric phases. Both of these are determined by the field $\mathbf{F}(\mathbf{R})$.

An illustrative example is a spin-$\frac{1}{2}$ particle in a uniform magnetic field $\mathbf{B}(\mathbf{R}) = \mathbf{R}$. The parameters **R** are just the components of the magnetic field, so that $\mathbf{F}(\mathbf{R}) = \mu\mathbf{R}$, where $-\mu$ is the magnetic moment. The eigenstates

$$|+(\mathbf{R})\rangle = \begin{pmatrix} \cos\theta/2 \\ e^{i\phi}\sin\theta/2 \end{pmatrix}, \quad |-(\mathbf{R})\rangle = \begin{pmatrix} -\sin\theta/2 \\ e^{i\phi}\cos\theta/2 \end{pmatrix} \tag{14}$$

depend on the polar angles (θ,ϕ) of **R**, and the energy levels $E_\pm(\mathbf{R}) = \pm\mu R$ on its magnitude. Straightforward calculation using (4) and (7) yields

$$\mathbf{A}_\pm(\mathbf{R}) = \pm\frac{1}{2}\frac{1-\cos\theta}{R\sin\theta}\hat{\phi}, \quad \mathbf{V}_\pm(\mathbf{R}) = \pm\frac{1}{2}\frac{\mathbf{R}}{R^3}. \tag{15}$$

Thus, $\mathbf{A}_\pm$ and $\mathbf{V}_\pm$ are respectively the vector potential and magnetic field of a magnetic monopole of strength $\pm 2\pi$ at the origin. Under the influence of a slowly varying magnetic field $\mathbf{R}(t)$ taken through a cycle C, a spin initially in an "up" state, say, returns to itself with a dynamical phase $-(\mu/\hbar)\int_0^T R dt$ and a geometric phase $\gamma_+(C) = -\int_S \mathbf{V}_+\cdot d\mathbf{S}$. Gauss's law implies that

$$\gamma_+(C) = -\tfrac{1}{2} \times \text{solid angle subtended by } C. \tag{16}$$

This spin-$\frac{1}{2}$ geometric phase has been observed in neutron spin-rotation experiments by Bitter and Dubbers (1987).

A special case is when C is confined to a plane containing the origin. Then its solid angle is 2π times the number of times m the origin is enclosed, and the corresponding phase factor is $(-1)^m$. This corresponds to the well-known sign change of spin-$\frac{1}{2}$ particles under rotations through 2π.

It is straightforward to generalize these results to the case of arbitrary $\mathbf{F}(\mathbf{R})$ in (13). The flux field is then given by

$$\mathbf{V}_\pm(\mathbf{R}) = \pm\tfrac{1}{2}\epsilon_{ijk}F_i\nabla F_j \times \nabla F_k/F^3 \tag{17}$$

(ϵ_{ijk} is the antisymmetric Levi–Cevita symbol), and it has monopole sources at points where $\mathbf{F}(\mathbf{R})$ vanishes. The monopoles have strength $\pm 2\pi\sigma$, where σ is the sign of the (assumed nonvanishing) determinant $\det(\partial F_i/\partial R_j)$ evaluated where $\mathbf{F} = 0$.

As is familiar from Dirac's analysis of magnetic monopoles (Dirac, 1931), attached to the monopole there is necessarily a string of singularities of the vector potential. The position of the string depends on the choice of gauge, and Dirac's quantization condition for electric charge follows from demanding that all gauges be physically indistinguishable. In the context of the two-state geometric phase, the monopole strings of $\mathbf{A}_\pm(\mathbf{R})$ lie along discontinuities in the phases of $|\pm(\mathbf{R})\rangle$. When the eigenstates are chosen as in (14), these discontinuities lie along the negative z axis. By varying the phase of the eigenstates [as in (8)], the string can be made to pass in and out of a given cycle C. As it does, the geometric phase $\gamma_\pm(C)$ changes by $\pm 2\pi$. Therefore, the geometric phase factor $\exp[i\gamma_\pm(C)]$ remains unchanged, and there are no observable consequences.

1.4 Degeneracies

In the preceding example of a spin-$\frac{1}{2}$ particle in a magnetic field, the flux fields $\mathbf{V}_\pm(\mathbf{R})$ are generated by point sources of strength $\pm 2\pi$ at the origin $\mathbf{R} = 0$. What makes the origin special is that it is the only point in parameter space where the energy levels are degenerate. This relationship between the geometric phase and energy level degeneracies turns out to be quite general.

In the absence of symmetries, a typical Hamiltonian in a parametrized family has a nondegenerate spectrum. However, at certain points in parameter space, belonging to the so-called *degenerate set,* two (or more) energy levels become degenerate. In the absence of time-reversal symmetry (the time-reversal symmetric case is discussed in the following section), the dimension of the degenerate set is typically three less than the dimension of parameter space itself (von Neumann and

Wigner, 1929). Thus, for the three-parameter families $H(\mathbf{R})$ we have been considering, degeneracies occur at isolated points. We have already seen this to be the case in the spin-$\frac{1}{2}$ example. (Note that there, the magnetic field is responsible for breaking time-reversal symmetry.)

More generally, let us consider the behavior of the flux field $V_n(\mathbf{R})$ in the neighborhood of a point of degeneracy $\mathbf{R}_*$ between two eigenstates, say the nth and $(n - 1)$st, of an arbitrary family of Hamiltonians. As $E_n(\mathbf{R}) - E_{n-1}(\mathbf{R})$ is small near $\mathbf{R}_*$, perturbation theory [cf. (9)] implies that the dominant component of $|\nabla n\rangle$ lies along the ket $|n - 1\rangle$. Therefore, for the purposes of approximating $\mathbf{V}_n$ (which is defined in terms of $|\nabla n\rangle$), the Hamiltonian can be restricted to the nearly degenerate subspace spanned by the states $|1\rangle \overset{\text{def}}{=} |n(\mathbf{R}_*)\rangle$ and $|2\rangle \overset{\text{def}}{=} |n - 1(\mathbf{R}_*)\rangle$. One obtains in this way a two-state system of the form (13), where the field $\mathbf{F}(\mathbf{R})$ is given by

$$F_x = \mathrm{Re}\langle 1|H(\mathbf{R})|2\rangle, \quad F_y = -\mathrm{Im}\langle 1|H(\mathbf{R})|2\rangle,$$
$$F_z = \frac{\langle 1|H(\mathbf{R})|1\rangle - \langle 2|H(\mathbf{R})|2\rangle}{2}. \tag{18}$$

In the neighborhood of $\mathbf{R}_*$, $\mathbf{V}_n(\mathbf{R})$ is given approximately by (17). There it has a single monopole source at the point of degeneracy $\mathbf{R}_*$ of strength $\pm 2\pi$. We may apply a similar analysis in the neighborhood of each degeneracy of $|n(\mathbf{R})\rangle$ to obtain

$$\nabla \cdot \mathbf{V}_n = \sum_{\mathbf{R}_*} \pm 2\pi\delta^3(\mathbf{R} - \mathbf{R}_*). \tag{19}$$

Thus, the flux field for the nth eigenstate has monopole sources of strength $\pm 2\pi$ at points where $|n\rangle$ is degenerate. [The sign of the monopole depends on whether the degeneracy is with the state above or below, and also on the sign of the determinant det $(\partial F_i(\mathbf{R}_*)/\partial R_j)$.] $\mathbf{V}_n$ may in addition have a purely divergenceless contribution.

It follows that the flux of $\mathbf{V}_n$ through a closed surface in parameter space is 2π times the number of degeneracies, counted with their signs, contained within.

1.5 Time-Reversal Symmetry

The presence of time-reversal symmetry imposes constraints on the geometric phase. These are strongest when the eigenstates themselves are time-reversal symmetric. We consider such cases first. Then the wave functions $\psi_n(\mathbf{r},\mathbf{R}) = \langle \mathbf{r}|n(\mathbf{R})\rangle$ can be chosen to be real, and the geometric phase factor is necessarily a sign ± 1, so that γ_n is an integer multiple of π. In this case, the geometric phase depends only on the topology of the closed curve, not on its geometry. As C is deformed, unless $\gamma_n(C)$ changes discontinuously, it necessarily remains constant. It can change (discontinuously) when C passes through degeneracies (points $\mathbf{R}_*$ where $|n\rangle$ is degenerate). For systems with time-reversal symmetry, the dimension of this degenerate set is typically two less than the dimension of parameter space (von Neumann and Wigner, 1929), so that it consists of lines in three-dimensional parameter space or points in two-dimensional parameter space. The geometric phase $\gamma_n(C)$ is π times the number of degeneracies enclosed by C, counted with the appropriate sign. An important consequence is that the flux field $\mathbf{V}_n$ has delta-function singularities at the energy-level degeneracies and is zero elsewhere.

The example of a spin $\frac{1}{2}$ in a magnetic field (cf. Sec. 1.3) provides an illustration if we fix Y (the y component of the magnetic field) to be zero. Then the spin Hamiltonian $\frac{1}{2}\mu\mathbf{B}\cdot\boldsymbol{\sigma}$ is real symmetric rather than complex Hermitian, and we can take its eigenvectors to be real. The parameter space is the XZ plane, and the solid angles subtended by curves in this plane are multiples of 2π. The geometric phase $\gamma_\pm(C)$ is just $\pm\pi$ times the number of times C encircles the origin. Another example concerns the electronic states of triatomic molecules X_3 (cf. Sec. 5.1).

In cases where the eigenstates are themselves not time-reversal symmetric, and instead occur in pairs related by time reversal, the constraints on the geometric phases are weaker. They are no longer necessarily multiples of π (i.e., the phase factors are no longer necessarily real). However, geometric phases for states related by time reversal necessarily have opposite signs.

1.6 Experimental Observation

Since the physical properties of a quantum system do not depend on its overall phase, one could wonder whether the geometric phase is observable. Numerous experiments have shown that it is.

Most direct experimental studies of the geometric phase belong to one of three types. The first is the "one state, two Hamiltonian" experiment. An ensemble of states initially in an energy eigenstate is coherently divided into two streams. The first evolves under a constant Hamiltonian, while the Hamiltonian driving the second is cyclically varied. The two streams are then recombined and allowed to interfere, and their relative phase determined. The relative phase includes both dynamical and geometric contributions, and to extract the geometric component the dynamical phase must be either made to vanish or else subtracted off explicitly. Examples include the optical polarization experiments discussed in Sec. 4.2.

The second type is the "two state, one Hamiltonian" experiment. Here an initial state $|\psi_i\rangle$ is prepared in a superposition $\alpha_i|m\rangle + \beta_i|n\rangle$ of energy eigenstates. The Hamiltonian governing the evolution is taken through an adiabatic cycle. The final state $|\psi_f\rangle = \alpha_f|m\rangle + \beta_f|n\rangle$ is physically distinguishable from the initial one, as the relative phase of the superposed eigenstates has changed. Part of this (observable) change in the relative phase involves the difference $\gamma_m - \gamma_n$ in geometric phases. The observation of optical activity in coiled fibers by Tomita and Chiao (1986) (see Sec. 4.1) provides an example of this second type.

In the third type of experiment, the adiabatic cycle is performed not once but repeatedly, with period T. With each cycle, $|n\rangle$ acquires a phase $\phi_n = -\int_0^T E_n(\mathbf{R}(t))dt/\hbar + \gamma_n$, so that there appears in its mean angular frequency $\omega_n = \phi_n/T$ a geometric shift $\Delta\omega_n = \gamma_n/T$. This frequency shift can be detected in spectroscopic measurements and is the basis for nuclear-magnetic-resonance studies of the geometric phase (cf. Sec. 5.2).

1.7 Geometric Magnetism

So far we have been regarding the evolution of the parameters $\mathbf{R}$ as externally prescribed. Now let us suppose the parameters of the Hamiltonian $H(\mathbf{R})$ are dynamical variables in their own right. For example, they could be the coordinates of a heavier (and therefore slower) system to which the original one, the fast system, is coupled. For definiteness, we take the Hamiltonian of the coupled system to be of the form $\mathbf{P}^2/2M + H(\mathbf{R})$, where $\mathbf{P}$ is the momentum of the slow system, and the dependence on the fast coordinates and momenta is left implicit in $H(\mathbf{R})$. Both systems are to be treated quantum mechanically. The geometric phase describes an effect on the fast system due to the slow. There is a corresponding influence on the slow system due to the fast.

This is the natural setting for the Born–Oppenheimer approximation, in which the fast and slow systems are the electrons and nuclei of a molecule. (For consistency with the preceding discussion, we denote the nuclear coordinates by a single vector $\mathbf{R}$; it is easy to adapt the notation to accommodate more than one nucleus.) By assuming that the electronic state adjusts rapidly to changes in the nuclear configuration, one is led to look for approximate eigenfunctions of the form $\Psi_n(\mathbf{r},\mathbf{R}) = \phi_n(\mathbf{R})\psi_n(\mathbf{r},\mathbf{R})$, where the electronic wave function $\psi_n(\mathbf{r},\mathbf{R})$ is an eigenstate with energy $E_n(\mathbf{R})$ of the electronic Hamiltonian $H(\mathbf{R})$, in which the nuclear coordinates appear as parameters. The nuclear wave function $\phi_n(\mathbf{R})$ is an eigenstate of the nuclear Hamiltonian $-\hbar^2\nabla^2/2M + E_n(\mathbf{R})$; the electronic energy plays the role of a potential. The Born–Oppenheimer approximation consists of neglecting the relatively slow dependence of the electronic wave function $\psi_n(\mathbf{r},\mathbf{R})$ on the nuclear coordinates. [It is essentially the adiabatic approximation (1) adapted to coupled systems.]

This was the context for the seminal work of Mead and Truhlar (1979). The primary objective of this work, which anticipated certain aspects of the geometric phase, was to derive corrections to the Born–Oppenheimer approximation. Such corrections appear when the momentum operator $\mathbf{P} = -i\hbar\nabla$ is applied to the putative eigenstate $\phi_n(\mathbf{R}) \times |n(\mathbf{R})\rangle$. The result, $-i\hbar(\nabla\phi_n|n\rangle + \phi_n|\nabla n\rangle)$, is no longer proportional to the electronic eigenstate $|n(\mathbf{R})\rangle$. However, if we neglect electronic components orthogonal to $|n\rangle$, we obtain $-i\hbar(\nabla\phi_n + \langle n|\nabla n\rangle)|n\rangle$, which is. The effect on the nuclear Schrödinger equation is to redefine the momentum operator,

$$\mathbf{P}\phi \overset{\text{def}}{=} (-i\hbar\nabla + \mathbf{A}_n)\phi, \tag{20}$$

where $\mathbf{A}_n(\mathbf{R})$ is given by (4). Thus, the field whose line integral gives the electronic geometric phase appears as a vector potential in the nuclear Hamiltonian. There it corresponds to an effective magnetic field $\mathbf{V}_n = \nabla \times \mathbf{A}_n$ that modifies the nuclear dynamics. This is called *geometric magnetism.*

The situation described above is quite general. Whenever there are two coupled systems evolving on different time scales, we can obtain an approximate description of the slow system by averaging over the fast motion. There appears in the slow dynamics an effective vector potential and associated magnetic field. The vector potential and magnetic field are precisely the fields whose line integrals and fluxes describe the geometric phase. (In cases where the coupling involves not only the slow coordinates but the slow momenta as well, this description has to be modified somewhat.)

Geometric magnetism is responsible for a variety of phenomena. Mead and Truhlar (1979) analyzed its effects in nuclear rotation–vibration spectra, which have since been observed experimentally (cf. Sec. 5.1). These effects can be nontrivial even when time-reversal symmetry is present (cf. Sec. 1.5), because of the change in sign of the electronic wave function $\psi_n(\mathbf{r},\mathbf{R})$ when the nuclear coordinates $\mathbf{R}$ are carried around a degenerate configuration. Like the magnetic field of the Aharonov–Bohm effect (Sec. 1.2), the flux field $\mathbf{V}_n(\mathbf{R})$ in this case vanishes everywhere except at these degenerate configurations, where it is singular. Mead (1980) has called this the "molecular Aharonov–Bohm effect."

Geometric magnetism survives in the classical limit. When this limit is applied to the slow system alone, it leads to geometric phase modifications of the semiclassical (Bohr–Sommerfield) quantization conditions. These are discussed by Kuratsuji and Iida (1985), Wilkinson (1984), and Littlejohn and Flynn (1992).

The situation where both the fast and slow systems are fully classical is discussed in Sec. 3.1.5. Here let us mention a simple but compelling illustration discussed by Li and Mead (1992). Let us consider the Born–Oppenheimer approximation applied to an atom rather than a molecule. The electronic energy levels $E_n(\mathbf{R})$ depend on the atom's center of mass $\mathbf{R}$, and appear as potentials in the equations of motion $M\ddot{\mathbf{R}} = -\nabla E_n + \mathbf{F}(\mathbf{R})$, along with any external force $\mathbf{F}(\mathbf{R})$. In the absence of external potentials, the energy levels are independent of $\mathbf{R}$, and $\mathbf{F} \equiv 0$; the center of mass behaves as a free particle. Now suppose the atom moves in a uniform magnetic field $\mathbf{B}$. The energy levels are still $\mathbf{R}$ independent (by translation invariance), but there now appears in the center-of-mass dynamics a Lorentz force, so that $M\ddot{\mathbf{R}} = \mathbf{F} = +Ze\dot{\mathbf{R}} \times \mathbf{B}$. The predictions of this analysis are spectacularly unphysical; walking through the Earth's magnetic field, you would experience a vertical force sufficient to hurl you into the air. What is missing from this analysis is the geometric magnetism from the electrons; the field $\mathbf{V}_n(\mathbf{R})$ introduces into the center-of-mass dynamics a second Lorentz force that either exactly or nearly (in case the electronic eigenstate has a nonzero magnetic moment) cancels the first.

2. THEORETICAL DEVELOPMENTS

A number of theoretical developments of the geometric phase are described. These include its mathematical setting in terms of holonomy under parallel transport, as well as several generalizations that remove the assumptions of nondegeneracy and adiabaticity.

2.1 Geometric Phase as Holonomy

Objects rigidly transported around a closed path in a curved space can return with a different orientation—this is *holonomy.* Here is a simple (and surely physical) example. Hold your right arm outstretched before you with your palm facing down. Keeping your wrist rigid, swing your arm into your chest, then raise it clockwise in front of you, and finally swing it away downward until it is outstretched before you again. Your arm has returned to its original position, but your hand has turned clockwise through 90°.

The mathematical setting for this phenomenon is the theory of connections on fiber bundles. What follows in Secs. 2.1.1–2.1.4 is an informal account of the relation between this theory and the geometric phase. No attempt is made at mathematical precision; the intention is to convey a sense of the principal ideas involved. Two references on fiber bundles oriented toward physicists are Eguchi *et al.* (1980) and Nash and Sen (1983).

2.1.1 Fiber Bundles A *fiber bundle E* is a space, or manifold, composed of two smaller ones, the *base manifold M* and the *fiber F.* The fiber bundle is constructed by attaching a copy of the fiber to each point of the base in a particular way. The simplest case, a *trivial* fiber bundle, is where E is just the Cartesian product $M \times F$ of the base with the fiber. A general fiber bundle cannot be so expressed. While it is always possible to partition it into smaller pieces that are themselves Cartesian products $M_\alpha \times F$ of subsets of the base with the fiber, these pieces may fit together in an interesting and nontrivial way, which reflects the topological properties of the bundle.

The configuration space E of an outstretched arm can be described as a fiber bundle. In this case, the base manifold M is the unit sphere, whose polar coordinates (θ,ϕ) determine the arm's direction. The fiber F is the unit circle, whose single angle coordinate χ describes the orientation of the hand relative to some reference position. In this case E is not simply the Cartesian product $M \times F$. This is because it is not possible to assign, in a continuous way, a reference hand position for every arm direction. This is a consequence of the "hairy ball theorem," i.e., the fact that it is impossible to comb flat a hairy ball without introducing singularities (e.g., crowns, cowlicks, bald patches).

As in the arm example, it is often the case in physical applications that the fiber describes internal degrees of freedom. Usually the fiber has some additional mathematical structure as well. If it is a real or complex vector space (describing spin degrees of freedom, for example), then E is called a *real* or *complex vector bundle.* If the vector space is one dimensional, E is called a *line bundle* (real or complex). Or else, the fiber might be a group (for example, the two- or three-dimensional rotation group, describing the axial or spatial orientation of a rigid body). Then E is called a *principal bundle.*

2.1.2 Connections and Curvature To a given path P on the base there correspond innumerable paths in the fiber bundle that project down to it; a motion in the base leaves the fiber motion undetermined. A *connection* is a rule for associating to P a particular path $\hat{P}$ through a given point in the bundle, the *lifted path* (see Fig. 1). A point moving along the lifted path $\hat{P}$ is said to be *parallel transported.* The connection rule is expressed in differential form by making the fiber velocity a linear function of the base velocity; this function can also depend on the base and fiber positions. For vector bundles, the dependence on fiber position should be linear as well. There are analogous restrictions in the case of principal bundles. In physical applications, a particular connection is often suggested by kinematical or dynamical considerations, e.g., moving your arm without unnecessarily twisting your wrist.

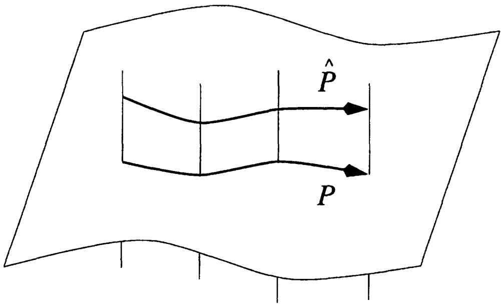

FIG. 1. A path P on the base manifold and its lift $\hat{P}$ in the fiber bundle.

Let us consider the lift $\hat{C}$ of a closed path C on the base (see Fig. 2). Its end points necessarily lie in the same fiber, but they need not coincide. If they do not, the difference between the end points is called the *holonomy* of C. In the outstretched-arm example, C is the spherical right triangle described by your arm, and its holonomy is the $\pi/2$ turn

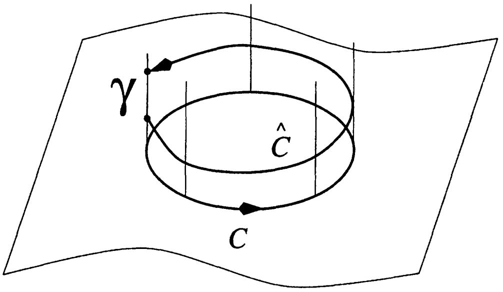

FIG. 2. A closed path C on the base manifold and its lift $\hat{C}$ in the fiber bundle. The difference between the end points of $\hat{C}$, a displacement within a single fiber, is the holonomy of C.

of your hand that results from traversing it in the manner prescribed. In case there is holonomy for arbitrarily small closed paths, the connection is said to have *curvature.* Curvature describes the infinitesimal holonomy acquired around an infinitesimal closed path, and may be expressed in terms of derivatives (a generalized curl) of the connection.

Let us consider a very simple example. We take the base M to be three-dimensional Euclidean space $\{\mathbf{R} \in \mathbb{R}^3\}$, the fiber F to be one-dimensional complex vector space $\{z \in \mathbb{C}\}$, and the bundle E to be their Cartesian product $\{(\mathbf{R},z) \in \mathbb{R}^3 \times \mathbb{C}\}$. Thus E is a trivial complex line bundle. A connection specifies the fiber velocity $\dot{z}$ as a function of the fiber position z, the base position $\mathbf{R}$, and the base velocity $\dot{\mathbf{R}}$; the dependence on z and $\dot{\mathbf{R}}$ is linear. The general form of such a relation is given by $\dot{z} = [\mathbf{W}(\mathbf{R})\cdot\dot{\mathbf{R}}]z$, where $\mathbf{W}(\mathbf{R})$ is a complex vector field. $\mathbf{W}$ is called the connection form, or simply the connection.

Given a motion $\mathbf{R}(t)$ in the base, its lifted motion in the fiber is given by $z(t) = \exp(\int_{\mathbf{R}(t)}\mathbf{W}\cdot d\mathbf{R})z_0$, where z_0 denotes the initial position in the fiber of the lifted path. If $\mathbf{R}(t)$ describes a closed path, then its holonomy is the scalar factor $\exp\Gamma$, where $\Gamma = \oint_{\mathbf{R}(t)}\mathbf{W}(\mathbf{R})\cdot d\mathbf{R}$. Using Stokes's theorem, we may express Γ as the flux of a vector field $\mathbf{V} = \nabla \times \mathbf{W}$ through a surface S bounded by C. $\mathbf{V}$ is the curvature of the connection $\mathbf{W}$.

Of particular relevance to the geometric phase is the case in which the connection preserves the length $|z(t)|$ of vectors along the lifted path. Then the connection is said to be *Hermitian.* The connection form $\mathbf{W}$ is purely imaginary and may be written as $i\mathbf{A}$, where $\mathbf{A}(\mathbf{R})$ is real.

2.1.3 The Geometric Phase Revisited

Simon (1983) observed that the Berry phase finds its natural mathematical expression as the holonomy of a connection on a fiber bundle. The fiber bundles in question are the complex line bundles $E^{(n)}$ associated with nondegenerate eigenstates of a quantum Hamiltonian $H(\mathbf{R})$. The base manifold is the parameter space (we continue to assume this to be $\mathbb{R}^3$), and the fibers are vectors $|\psi\rangle$ in Hilbert space that satisfy the Schrödinger equation $[H(\mathbf{R}) - E_n(\mathbf{R})]|\psi\rangle = 0$; these are just the unnormalized nth eigenstates, related to each other by a complex scalar.

Given an eigenstate $|\psi\rangle$ of $H(\mathbf{R})$, a connection determines an eigenstate $|\psi\rangle + |d\psi\rangle$ of the nearby Hamiltonian $H(\mathbf{R} + d\mathbf{R})$ by fixing its norm and phase. On physical grounds it is sensible to require that the connection conserve probability, i.e., that $|\psi\rangle + |d\psi\rangle$ and $|\psi\rangle$ both have unit norm. This implies that $\mathrm{Re}\langle\psi|d\psi\rangle = 0$. The simplest rule for fixing the phase of $|\psi\rangle + |d\psi\rangle$ is to take $\mathrm{Im}\langle\psi|d\psi\rangle = 0$ as well. This determines completely a connection on $E^{(n)}$, which is defined by the relation

$$\langle\psi|d\psi\rangle = 0. \tag{21}$$

Under parallel transport around a closed path C, an eigenstate $|\psi\rangle$ acquires a phase factor $\exp(i\gamma_n)$. This phase factor, the holonomy of C, is precisely the geometric phase factor. This follows by noting that the connection (21) is just the condition (2) derived from adiabatic evolution.

There is an exact mathematical analogy between the outstretched arm and a variant of the spin example of Sec. 1.3. If a spin-1 (instead of a spin-$\frac{1}{2}$) particle, initially in its "down" state, is driven by a magnetic field that is slowly cycled around a closed path C, then its geometric phase γ_1 is precisely the turn in the hand of an arm describing the same circuit. Both are equal to the solid angle subtended by C.

2.1.4 Chern Classes

Some fiber bundles are trivial Cartesian products $M \times F$, and others are not. Topology lies at the root of this distinction, and *characteristic classes* make this distinction precise. They describe an intrinsic twistedness in fiber bundles that cannot be combed away. Bundles whose characteristic classes are the same can be smoothly transformed into one another; bundles whose characteristic classes are different cannot.

For complex vector bundles, the characteristic classes are called *Chern classes.* The number of Chern classes is equal to the dimension of the fiber. The rth Chern class associates an integer $c_r(N)$, the *Chern number,* to every $2r$-dimensional, boundaryless subspace N of the base manifold M. Thus, the first Chern class assigns an integer $c_1(S)$ to every closed surface S in M. There are various ways to calculate Chern numbers. Among these, there are formulas in terms of integrals over the subspace N of the curvature and functions thereof.

An illustrative example is the celebrated Gauss–Bonnet theorem from classical differential geometry. The theorem states that the integral $\int_S K dS$ of the Gaussian curvature K over a closed one-sided surface S is equal to 2π times an integer χ, the Euler characteristic. The Euler characteristic is a topological feature of the surface. It is related by the formula $\chi = 2 - 2g$ to the genus g, the number of handles on S when it is regarded through a topologist's eyes as a many-handled sphere. The Gauss–Bonnet theorem can be verified immediately for the unit sphere ($K = 1$, $g = 0$). What is remarkable is the implication that $\int_S K dS = 4\pi$ for any surface smoothly deformable to a sphere, but $\int_S K dS \neq 4\pi$ for any surface (such as a torus) that is not.

The Gauss–Bonnet theorem can be formulated in terms of fiber bundles. The bundle consists of the surface S (the base) together with its tangent vectors at each of its points (the fibers). A connection is defined through the Riemannian rule for parallel transport on a curved surface. By identifying two-dimensional real tangent vectors with one-dimensional complex ones, the fiber bundle may be regarded as a complex line bundle, and its first Chern number $c_1(S)$ turns out to be $1/2\pi$ times the integral $\int_S K dS$ of the Gaussian curvature.

Returning to considerations of the geometric phase, let us consider the complex line bundles $E^{(n)}$ of quantum eigenstates, discussed in the preceeding section. Each has a single Chern class $c_1^{(n)}$ that assigns an integer to every closed surface S in parameter space. This integer is given by

$$c_1^{(n)}(S) = \frac{1}{2\pi}\oint_S \mathbf{V}_n \cdot d\mathbf{S} = \frac{1}{2\pi}\int_B \nabla \cdot \mathbf{V}_n d^3R, \tag{22}$$

where in the second equality we have used the divergence theorem; B is the three-dimensional region bounded by S. From (19), the last expression is precisely the number of energy-level degeneracies [points $\mathbf{R}_*$ where $E_n(\mathbf{R}_*) = E_{n\pm1}(\mathbf{R}_*)$] contained in S, counted with their appropriate sign (see Sec. 1.4). In terms of the electromagnetic analogy (Sec. 1.2), the Chern number is the number of magnetic monopoles in S counted with their charge. Thus, energy-level degeneracies manifest themselves in the nontrivial topology of the eigenstate line bundles $E^{(n)}$.

2.2 Degenerate Eigenstates and Non-Abelian Geometric Phases

If a system possesses a sufficient degree of symmetry, its energy levels will be degenerate. Two examples are the angular-momentum multiplets of rotationally symmetric systems, and Kramers degeneracy in systems with time-reversal symmetry and an odd number of fermions. If the symmetry persists throughout the parametrized family $H(\mathbf{R})$, then so do these degeneracies. Under a slowly varying Hamiltonian $H(\mathbf{R}(t))$, the adiabatic theorem implies that transitions to states outside a degenerate multiplet are small. However, transitions to states within the multiplet need not be, no matter how slow the variation (transitions between degenerate states cost no energy).

Wilczek and Zee (1984) generalized Berry's analysis to degenerate eigenstates. They showed that under adiabatic cycling around a closed circuit C, such states need not return to themselves up to a phase, but instead can evolve into new states within the multiplet. The transformation from initial to final states is generated by a unitary matrix $U(C)$, called the *non-Abelian geometric phase.* Non-Abelian geometric phases can be interpreted as holonomy, but of a more general sort than was considered in Sec. 2.1.3. There is a corresponding generalization of geometric magnetism.

The analysis, which runs parallel to the nondegenerate case, proceeds as follows. Let $E(\mathbf{R})$ denote an r-fold degenerate energy level whose degenerate eigenspace is spanned by an orthonormal basis $|j(\mathbf{R})\rangle$, $j = 1, 2, \ldots, r$. We consider the evolution under a slowly changing Hamiltonian $H(\mathbf{R}(t))$ of states $|\psi_j(t)\rangle$ initially equal to $|j(\mathbf{R}(0))\rangle$. Neglecting transitions to states outside the multiplet, we look for approximate solutions of the form

$$|\psi_j(t)\rangle = \exp\left(-\frac{i}{\hbar}\int_0^t E(\mathbf{R}(\tau))d\tau\right) \times \sum_{k=1}^{r} U_{jk}(t)|k(\mathbf{R}(t))\rangle, \qquad U_{jk}(0) = \delta_{jk}, \tag{23}$$

i.e., a superposition of evolving degenerate eigenstates multiplied by an overall dynamical phase factor. For the inner products $\langle \psi_k(t)|\psi_j(t)\rangle$ to be conserved in time, the $r \times r$ matrix $\boldsymbol{U}(t)$ must be unitary. Its evolution is determined by requiring $\langle k(\mathbf{R}(t))|(i\hbar\partial_t - H)|\psi_j(t)\rangle$ to vanish for each k (it would automatically if $\psi_j(t)$ were an exact solution of the Schrödinger equation). This requirement leads to the equation of motion

$$\dot{\boldsymbol{U}}(t) = -i\boldsymbol{U}(t)\mathsf{A}(\mathbf{R}(t))\cdot\dot{\mathbf{R}}(t) = -i\boldsymbol{U}(t)\boldsymbol{A}(t), \tag{24}$$

where $\mathsf{A}(\mathbf{R}) = (\boldsymbol{A}^{(1)},\boldsymbol{A}^{(2)},\boldsymbol{A}^{(3)})(\mathbf{R})$ is a vector of $r \times r$ Hermitian matrices whose matrix elements are given by

$$A_{jk}^{(l)}(\mathbf{R}) = -i\langle k(\mathbf{R})|\nabla_l j(\mathbf{R})\rangle, \tag{25}$$

and $\boldsymbol{A}(t) = \mathsf{A}(\mathbf{R}(t))\cdot\dot{\mathbf{R}}(t)$.

The solution of (24) is given by the time-ordered product

$$\begin{aligned}\boldsymbol{U}(t) &= T\exp\left(-i\int_0^t \boldsymbol{A}(\tau)d\tau\right)\\ &= \boldsymbol{I} - i\int_{0<\tau<t} \boldsymbol{A}(\tau)d\tau\\ &\quad - \iint_{0<\tau<\tau'<t} \boldsymbol{A}(\tau)\boldsymbol{A}(\tau')d\tau d\tau' + \cdots,\end{aligned} \tag{26}$$

where $\boldsymbol{I}$ is the $r \times r$ identity matrix. The time ordering indicates that in expanding the exponential of the integral, all matrices are to be ordered with later times to the right. If the matrices $A(\tau)$ at different times commute with each other, then the ordering of the factors is irrelevant, and the right-hand side of (26) simplifies to

$$\begin{aligned}&\boldsymbol{I} - i\int_{0<\tau<t} \boldsymbol{A}(\tau)d\tau - \left[\int_{0<\tau<t} \boldsymbol{A}(\tau)d\tau\right]^2\Big/2! + \cdots\\ &\quad = \exp\left(-i\int_0^t \boldsymbol{A}(\tau)d\tau\right),\end{aligned}$$

i.e., simply the exponentiated integral. In particular, if $r = 1$, then $A(\tau)$ is a scalar, and $U(t)$ is a phase factor.

Suppose the Hamiltonian is taken around a closed cycle C. Then the relation between initial and final states is described, apart from the dynamical phase factor, by the unitary matrix

$$\boldsymbol{U}(C) = T\exp\left(-i\oint_C \mathsf{A}(\mathbf{R})\cdot d\mathbf{R}\right). \tag{27}$$

$\boldsymbol{U}(C)$ is the non-Abelian geometric phase. The terminology derives from the fact that the factors $\mathsf{A}(\mathbf{R})\cdot d\mathbf{R}$ in the time-ordered product do not in general commute. They do, however, in the nondegenerate case $r = 1$; then $U(C)$ is just the ordinary (Abelian) geometric phase factor $\exp[i\gamma(C)]$. Like the ordinary geometric phase, $U(C)$ depends only on the path C and not on the rate at which it is traversed.

For infinitesimal circuits, $U(C)$ is of the form $\boldsymbol{I} + i\Gamma(C)$, where $\Gamma(C)$ is an infinitesimal Hermitian matrix given by (minus) the flux of the covariant curl of A,

$$\mathsf{V} = \nabla \times \mathsf{A} - i\mathsf{A} \times \mathsf{A}, \tag{28}$$

through the infinitesimal area bounded by C. (Note that because the components of A are matrices, the cross product $\mathsf{A} \times \mathsf{A}$ does not vanish, just as the cross product $\boldsymbol{\sigma} \times \boldsymbol{\sigma} = i\boldsymbol{\sigma}$ of Pauli matrices does not vanish.) This result does not extend to finite circuits, though, as there is no non-Abelian version of Stokes's theorem.

Like the ordinary geometric phase, $U(C)$ may be interpreted as holonomy, but now on a complex r-dimensional vector bundle rather than a line bundle. The base manifold is parameter space ($\mathbb{R}^3$ in the examples considered here), and the fibers are the degenerate eigenspaces spanned by $|j(\mathbf{R})\rangle$. The matrix-valued vector field $\mathsf{A}(\mathbf{R})$ is the connection form, which describes how states in the fibers are parallel transported along paths $\mathbf{R}(t)$ in the base. $\mathsf{V}(\mathbf{R})$ is the associated curvature form.

The non-Abelian geometric phase does not depend on the choice of basis $|j(\mathbf{R})\rangle$ in any intrinsic way. Under a unitary change of basis

$$|j(\mathbf{R})\rangle \rightarrow |\bar{j}(\mathbf{R})\rangle = \sum_{k=1}^{r} W_{jk}(\mathbf{R})|k(\mathbf{R})\rangle, \tag{29}$$

the non-Abelian version of the gauge transformation (8), $\boldsymbol{U}(C)$ transforms according to $\boldsymbol{U}(C) \rightarrow \boldsymbol{W}\boldsymbol{U}(C)\boldsymbol{W}^{\dagger}$. This unitary conjugation represents simply the change of basis applied to $\boldsymbol{U}(C)$ and leaves its observable properties unchanged. The fields A(**R**) and V(**R**) transform according to

$$\mathsf{A} \rightarrow W\mathsf{A}W^{\dagger} - i\nabla W W^{\dagger}, \quad \mathsf{V} \rightarrow W\mathsf{V}W^{\dagger} \qquad (30)$$

under (29). These transformation properties characterize the potential and field strength of Yang–Mills field theories, generalizing the electromagnetic analogy with the geometric phase (cf. Sec. 1.2) to the non-Abelian case.

Given a particular closed circuit C, one can choose the basis $|j(\mathbf{R})\rangle$ so as to diagonalize $U(C)$. Then under parallel transport around C, each basis state returns to itself up to a phase factor. Under certain conditions, one can find a basis that diagonalizes every non-Abelian geometric phase, regardless of C. In such cases, the degenerate and nondegenerate cases are essentially the same. In general, though, there is no choice of basis that renders every $U(C)$ diagonal. What distinguishes the general case is that the generators of the symmetries responsible for the degeneracies depend on parameters themselves. A particular example, the Λ-doubling of electron doublets in diatomic molecules, is discussed by Moody *et al.* (1986).

The discussion of geometric magnetism in the Born–Oppenheimer approximation (cf. Sec. 1.7) can also be generalized to the degenerate case. As before, we regard the parameters **R** as the coordinates of a slow system to which the original one, the fast system, is coupled. The Hamiltonian is given by $P^2/2M + H(\mathbf{R})$, where **P** is the slow momentum, and the dependence on the fast coordinates and momenta is left implicit in $H(\mathbf{R})$. We look for approximate eigenfunctions of the coupled system in which the fast system belongs to a degenerate multiplet of $H(\mathbf{R})$. These are of the form $\Psi(\mathbf{r},\mathbf{R}) = \Sigma_{j=1}^{r}\phi_j(\mathbf{R})\psi_j(\mathbf{r},\mathbf{R})$, where $\psi_j(\mathbf{r},\mathbf{R}) = \langle \mathbf{r}|\, j(\mathbf{R})\rangle$ are the fast eigenfunctions (which depend parametrically on the slow variables **R**). Substituting this *Ansatz* into the stationary Schrödinger equation and making approximations similar to those in the nondegenerate case (the details are omitted), one sees that the slow system is described by a multicomponent wave function $\phi(\mathbf{R}) = (\phi_1(\mathbf{R}), \ldots, \phi_r(\mathbf{R}))$ that is an eigenfunction of the matrix-valued Hamiltonian

$$\boldsymbol{H} = (\mathbf{P}\boldsymbol{I} - \mathsf{A})^2/2M + E(\mathbf{R})\boldsymbol{I}. \qquad (31)$$

Here A(**R**) is the matrix-valued vector potential of (25) and $\boldsymbol{I}$ is the $r \times r$ identity matrix. Thus, coupling to a degenerate fast eigenstate introduces a generalized form of geometric magnetism, in which the slow Hamiltonian acquires spin degrees of freedom that are coupled to a Yang–Mills-like field.

Finally, there is a connection between the non-Abelian geometric phase and higher-order degeneracies, i.e., points in parameter space where two or more degenerate levels coalesce. These points act as generalized monopole sources for the Yang–Mills field V.

2.3 Nonadiabatic Evolution

Geometric phases appear in contexts other than adiabatic evolution. An important generalization was found by Aharonov and Anandan (1987), who showed that geometric phases accompany any cyclic evolution, not necessarily adiabatic, of a quantum state. Subsequently, Samuel and Bhandari (1988) showed that geometric phases can arise in evolutions that are nonunitary, not cyclic, and not even continuous. What underlies these generalizations is a natural prescription for the parallel transport of quantum phase factors.

The setting for the Aharonov–Anandan phase is quite general. A state $|\psi(t)\rangle$ evolves according to the Schrödinger equation $[i\hbar\partial_t - H(t)]|\psi\rangle = 0$ under a (usually) time-dependent Hamiltonian $H(t)$. After a time T the state is presumed to return to itself up to a phase factor, so that $|\psi(T)\rangle = \exp(i\beta)|\psi(0)\rangle$. In the trivial case where H is time independent and $|\psi\rangle$ is an eigenstate, β is simply the dynamical phase $-ET/\hbar$. Guided by this simple case, we are led to subtract from β a generalized dynamical phase, namely, the time integral of the expectation value of energy,

$$\beta_d = -\frac{1}{\hbar}\int_0^T \langle\psi|H|\psi\rangle dt = \mathrm{Im}\int_0^T \langle\psi|\dot{\psi}\rangle dt. \qquad (32)$$

The phase that remains, $\beta_g \overset{\mathrm{def}}{=} \beta - \beta_d$, is given by

$$\beta_g = \beta - \beta_d = \arg\langle\psi(0)|\psi(T)\rangle - \mathrm{Im}\int_0^T \langle\psi|\dot{\psi}\rangle dt. \qquad (33)$$

What makes the decomposition $\beta = \beta_d +$

β_g of the total phase interesting is that β_g is, in a sense to be explained, a geometric phase. First, β_g depends only on the path through Hilbert space described by $|\psi(t)\rangle$ and not on the rate at which this path is traversed. Indeed, the state $|\tilde{\psi}(t)\rangle = |\psi(\tau(t))\rangle$ traverses the same path but, because of the time reparametrization $t \to \tau(t)$, at a different rate. However, it is easy to verify that its geometric phase β_g, as given by (33), is the same as that of $|\psi(t)\rangle$, even though its dynamical phase is different. [Note that $|\tilde{\psi}(t)\rangle$ might not even satisfy the Schrödinger equation.] Second, β_g does not depend on the overall phase of $|\psi(t)\rangle$. Indeed, if in (33) $|\psi(t)\rangle$ is replaced by $|\tilde{\psi}(t)\rangle = \exp[i\phi(t)]|\psi(t)\rangle$, a simple calculation shows that β_g is unchanged.

As $|\psi(t)\rangle$ is periodic up to a phase, the density matrix $\rho_\psi(t) = |\psi(t)\rangle\langle\psi(t)|$ is strictly periodic, and over the time interval $0 < t < T$ describes a closed curve C in the space of density matrices. What the preceding discussion shows is that β_g is a geometric property of C. It does not depend on the rate at which C is traversed, and it can be computed with (33), substituting any state $|\tilde{\psi}(t)\rangle$ whose density matrix $|\tilde{\psi}(t)\rangle\langle\tilde{\psi}(t)|$ is equal to $\rho_\psi(t)$.

The geometric phase for adiabatic cycles can be obtained as a special case of (33). Within the adiabatic approximation, a solution to the Schrödinger equation for a slowly cycled Hamiltonian $H(\mathbf{R}(t))$ is given by $|\psi(t)\rangle = \exp[i\beta(t)]|n(\mathbf{R}(t))\rangle$ [cf. (3)]. From (32), the dynamical phase β_d is given by $-(1/\hbar)\int_0^T E_n(\mathbf{R}(t))dt$. The geometric phase is easily computed from (33) if $|\tilde{\psi}(t)\rangle = |n(\mathbf{R}(t))\rangle$ is substituted for $|\psi(t)\rangle$ (this is permissible, since the two states differ by an overall phase). Since $\arg\langle n(\mathbf{R}(0))|n(\mathbf{R}(T))\rangle = 0$, β_g is given by $-\mathrm{Im} \int_0^T\langle\tilde{\psi}|\dot{\tilde{\psi}}\rangle dt = -\mathrm{Im}\oint\langle n(\mathbf{R})|\nabla n(\mathbf{R})\rangle\cdot d\mathbf{R}$, in agreement with (5).

What underlies this general setting for the geometric phase is once more the mathematics of connections and fiber bundles. However, now the entire Hilbert space is to be regarded as a fiber bundle! The base manifold $\bar{\mathcal{H}}$ consists of all pure-state density matrices ρ, normalized so that $\mathrm{Tr}\rho = 1$. (This is sometimes called the *projective* Hilbert space.) The fiber attached to ρ consists of all state vectors $|\psi\rangle$ whose density matrices $|\psi\rangle\langle\psi|$ are equal to ρ. (Thus, states in the same fiber differ only by a phase factor.) To an infinitesimal displacement $d\rho$ in the base, there corresponds displacements $|d\psi\rangle$ in the fiber satisfying $|\psi + d\psi\rangle\langle\psi + d\psi| = \rho + d\rho$. With a little algebra, this reduces to $|d\psi\rangle - \langle\psi|d\psi\rangle|\psi\rangle = d\rho|\psi\rangle$. This equation does not determine $|d\psi\rangle$ uniquely; to any solution $|d\psi\rangle$ can be added an infinitesimal imaginary component $id\alpha|\psi\rangle$ along $|\psi\rangle$. A connection fixes this component and determines a particular $|d\psi\rangle$ for a given $d\rho$. A natural choice is to require that

$$\langle\psi|d\psi\rangle = 0. \tag{34}$$

This connection is precisely the condition (2) found by Berry for adiabatic cycles, but now generalized to arbitrary cycles in Hilbert space. Its holonomy is precisely the geometric phase β_g given by (33).

This setting admits further generalizations (Samuel and Bhandari, 1988). For example, one can calculate a geometric phase for a nonunitary evolution $|\phi(t)\rangle$, for which $\langle\phi|\phi\rangle$ is no longer constant, simply by evaluating (33) with $|\psi(t)\rangle \overset{\text{def}}{=} |\phi(t)\rangle/\langle\phi|\phi\rangle^{1/2}$.

In fact, one can calculate a geometric phase for an open path $|\psi(t)\rangle$, where $|\psi(T)\rangle$ and $|\psi(0)\rangle$ are no longer proportional. The formula (33) for β_g can be computed for open paths, and the result does not depend on the parametrization in time of $|\psi(t)\rangle$, and remains unchanged by the change of phase $|\psi(t)\rangle \to \exp[i\phi(t)]|\psi(t)\rangle$. Further insight into the interpretation of the geometric phase for open paths is given in Sec. 2.4.

There is also a discrete version of the formula (33). Given a sequence of states $|\psi_1\rangle$, $|\psi_2\rangle, \ldots |\psi_N\rangle$, one can associate a phase according to the formula

$$\beta_g = \arg\langle\psi_1|\psi_2\rangle\langle\psi_2|\psi_3\rangle \ldots \langle\psi_N|\psi_1\rangle. \tag{35}$$

One sees easily that β_g does not depend on the overall phases of the states $|\psi_j\rangle$. If there are just two states, then $\beta_g = \arg|\langle\psi_1|\psi_2\rangle| = 0$ identically. But if $N > 2$, the phase β_g is typically nonvanishing. Note that $|\psi_N\rangle$ need not be proportional to $|\psi_1\rangle$; the sequence of states need not close on itself.

The formula (35) is the discrete analog of (33), and defines a geometric phase for discontinuous evolutions. Thus, a geometric phase can be defined for processes that include measurements and wave-function collapse. A version of formula (35) appears in

Pancharatnam's study of optical polarization dynamics, in which the $|\psi_j\rangle$ represent polarization states. Pancharatnam's phase is discussed in Sec. 4.2.

2.4 Geometric Metric Tensor

How close are two quantum states $|\psi\rangle$ and $|\phi\rangle$? The usual quantum mechanical measure of distance is the Hilbert-space norm $\langle\psi - \phi|\psi - \phi\rangle^{1/2}$ of their difference. There is another, introduced by Provost and Vallee (1980), namely, the quantity $s^2(\psi,\phi) \overset{\text{def}}{=} 1 - |\langle\psi|\phi\rangle|^2/\langle\psi|\psi\rangle\langle\phi|\phi\rangle$. Clearly, this vanishes if the states are the same. What distinguishes it from the Hilbert-space norm is that it is unchanged if either $|\psi\rangle$ or $|\phi\rangle$ is multiplied by a complex scalar. In fact, what this quantity really describes is the distance between their normalized density matrices $\rho_\psi = |\psi\rangle\langle\psi|/\langle\psi|\psi\rangle$ and $\rho_\phi = |\phi\rangle\langle\phi|/\langle\phi|\phi\rangle$, and indeed can be alternatively expressed as $1 - \text{Tr}\rho_\psi\rho_\phi$. For infinitesimally displaced states $|\phi\rangle = |\psi + d\psi\rangle$, a little algebra gives

$$\begin{aligned} ds^2(d\psi) &\overset{\text{def}}{=} s^2(\psi, \psi + d\psi) = 1 - \text{Tr}(\rho_\psi\rho_{\psi+d\psi}) \\ &= \tfrac{1}{2}\,\text{Tr}(d\rho_\psi^2) = \frac{\langle d\psi|I - \rho_\psi|d\psi\rangle}{\langle\psi|\psi\rangle}, \qquad (36) \end{aligned}$$

where $d\rho_\psi = \rho_{\psi+d\psi} - \rho_\psi$ and I is the identity operator. Equation (36) defines a metric on the space of normalized density matrices $\bar{\mathcal{H}}$ (or equivalently, the projective Hilbert space).

The quantum metric enters into the analysis of the geometric phase in a number of interesting ways. Below we describe two aspects: a deeper understanding of the geometric phase for open paths and physically significant corrections to the Born–Oppenheimer approximation.

The quantum metric determines geodesics, i.e., curves $\rho(s)$ that minimize, or at least render stationary, the distance between two given density matrices in $\bar{\mathcal{H}}$. These are most easily described in terms of their corresponding state vectors; it may be shown that if $|\psi(s)\rangle$ satisfies the differential equation $d^2|\psi\rangle/ds^2 + |\psi\rangle = 0$, and if the vectors $|\psi(0)\rangle$ and $|\psi'(0)\rangle$ are normalized and orthogonal to one another, then the curve $\rho_\psi(s) = |\psi(s)\rangle\langle\psi(s)|$ is a geodesic with respect to the metric (36), and the parameter s along the curve is the arc length. [The solutions $|\psi(s)\rangle$ are of the form $\cos s|\psi(0)\rangle + \sin s|\psi'(0)\rangle$.]

A simple illustration is provided by two-state systems. In this case, $\bar{\mathcal{H}}$ may be identified with the unit sphere (in the context of polarization optics, this is the Poincaré sphere—see Sec. 4.2); each density matrix ρ may be expressed in the form $(I + \mathbf{\hat{e}}\cdot\boldsymbol{\sigma})/2$, where $\mathbf{\hat{e}}$ is a unit vector and $\boldsymbol{\sigma}$ are the Pauli matrices. The quantum metric (36) reduces to the usual metric on the sphere (up to a scalar factor), and the geodesics are the arcs of great circles.

As discussed in Sec. 2.3, geometric phases can be defined via (33) for open as well as closed paths in $\bar{\mathcal{H}}$. The two cases may be related as follows. An open path P may be closed by joining its end points with a geodesic. The resulting closed path, it turns out, has the same geometric phase as P. The geometric phases for continuous and discontinuous evolutions, (33) and (35), may be similarly related; given a closed sequence of density matrices $\rho_1, \rho_2, \ldots \rho_N, \rho_1$, we can construct a closed polygonal path $\rho(s)$ in $\bar{\mathcal{H}}$ by connecting its points with geodesics. $\rho(s)$ turns out to have the same geometric phase as the discrete sequence ρ_j.

The quantum metric also plays a role in the Born–Oppenheimer description of the dynamics of a slow system coupled to a fast one (cf. Sec. 1.7). There it appears in the slow Hamiltonian as an additional scalar potential $\Phi_n(\mathbf{R}) = \Sigma_\alpha ds^2|\nabla_\alpha n(\mathbf{R})\rangle/2M$, which produces a "geometric electric" force. The scalar potential appears at a higher order of adiabatic expansion than the vector potential $A_n(\mathbf{R})$; with appropriate scalings, the geometric magnetic force $\dot{\mathbf{R}} \times \mathbf{V}_n$ is of the order of the square root $\sqrt{m/M}$ of the mass ratio of the fast and slow systems, whereas the geometric electric force $-\nabla\Phi$ is of the order m/M. However, the electric force can become significant in the vicinity of degeneracies $\mathbf{R}_*$ [i.e., points where the fast energy level $E_n(\mathbf{R})$ becomes degenerate]. There the force is repulsive and so tends to steer the slow dynamics away from points where the Born–Oppenheimer approximation would break down. A discussion of the scalar potential can be found in the reviews of Berry (1989) and Jackiw (1988). Additional terms in the slow dynamics of comparable order to geometric electricity have also been found by Weigert and Littlejohn (1993).

2.5 Beyond Adiabaticity

2.5.1 Corrections to the Geometric Phase For the adiabatically cycled eigenstates described by Eq. (3), the dynamical and geometric phases are but the first two terms in an asymptotic expansion for the total phase. An elegant iterative scheme for evaluating higher-order corrections was developed by Berry (1987c). These corrections to the geometric phase turn out to be geometric phases themselves. The full asymptotic series diverges in a manner that reveals the breakdown of the adiabatic approximation.

The method is illustrated by the example of a spin-$\frac{1}{2}$ particle in a magnetic field (cf. Sec. 1.3). We consider a time-dependent Hamiltonian $H_0(\tau) = \mu\mathbf{B}_0(\tau)\cdot\boldsymbol{\sigma}$, cyclic over the infinite time interval $[-\infty,\infty]$ with $\mathbf{B}_0(\pm\infty) = B_*\hat{\mathbf{z}}$. $\mathbf{B}_0(\tau)$ is assumed to be analytic, and we assume as well that $B_0(\tau) \geq B_{\text{min}}$ for all τ, so that the Hamiltonian does not pass through any degeneracies. $\tau = \epsilon t$ is a dimensionless time, and ϵ a small parameter (with dimensions of frequency) describing the slow rate of change of the Hamiltonian. (The corresponding dimensionless small parameter is $\delta = \epsilon\hbar/\mu B_{\text{min}}$.) Let us consider the evolution of a spin state $|\psi(t)\rangle$ with initial condition $|\psi(-\infty)\rangle = |-\rangle$. The analysis is facilitated by transforming to a rotating frame in which the direction $\hat{\mathbf{b}}_0$ of the magnetic field is fixed (say, for definiteness, along $\hat{\mathbf{z}}$). Such a frame is described by a rotation matrix $\mathcal{R}_1(\tau)$ satisfying

$$\hat{\mathbf{b}}_0(\tau) = \mathcal{R}_1(\tau)\cdot\hat{\mathbf{z}}. \tag{37}$$

This condition does not determine $\mathcal{R}_1(\tau)$ uniquely, but only up to a rotation (applied from the left) about $\hat{\mathbf{b}}_0$. Differentiating with respect to τ (this is denoted by an overdot), we obtain a condition

$$\dot{\hat{\mathbf{b}}}_0(\tau) = \boldsymbol{\omega}_1(\tau)\times\hat{\mathbf{b}}_0(\tau) \tag{38}$$

on the angular velocity $\boldsymbol{\omega}_1$ of the frame, where $\omega_{1i} = \frac{1}{2}\Sigma_{jkl}\epsilon_{ijk}\mathcal{R}_{1kl}\dot{\mathcal{R}}_{1jl}$ (ϵ_{ijk} is the antisymmetric Levi–Cevita symbol). (38) does not determine $\boldsymbol{\omega}_1$ uniquely, but only up to its component along $\hat{\mathbf{b}}_0$. We fix this indeterminacy (and the corresponding indeterminacy in $\mathcal{R}_1$) by requiring that

$$\boldsymbol{\omega}_1\cdot\hat{\mathbf{b}}_0 = 0. \tag{39}$$

(39) is a condition for parallel transport, analogous to the requirement of moving your arm without unnecessarily twisting your wrist (cf. Sec. 2.1).

In the rotating frame, the Hamiltonian $H_1(\tau)$ is given by $\mu\mathbf{B}_1(\tau)\cdot\boldsymbol{\sigma}$, where

$$\mathbf{B}_1(\tau) = B_0(\tau)\hat{\mathbf{z}} - \mathbf{B}_{L1}(\tau) \tag{40}$$

is the effective magnetic field. $\mathbf{B}_{L1}(\tau) = \epsilon\hbar/\mu\mathcal{R}_1^{-1}(\tau)\cdot\boldsymbol{\omega}_1(\tau)$ is the Larmor field induced in the rotating frame (the magnetic analog of the Coriolis force). As the rotation is slow, the Larmor field is small (of order ϵ), and in the lowest-order adiabatic approximation (cf. Secs. 1.1 and 1.2), it is simply neglected. Then $H_1(\tau)$ is diagonal, and the time-dependent Schrödinger equation is easily solved. The geometric phase $\gamma_-(C_0)$, half the solid angle of the cycle C_0 described by $\mathbf{B}_0(\tau)$, appears as a direct consequence of the parallel transport condition (39) when the solution is transformed back to the fixed frame.

If the Larmor field is not neglected, then $\mathbf{B}_1(\tau)$ may be regarded as a slowly varying magnetic field to which the above procedure may be applied as before. That is, we transform to a second rotating frame $\mathcal{R}_2(\tau)$ in which $\hat{\mathbf{b}}_1(\tau)$ is fixed along z. The frame is uniquely determined by imposing the parallel-transport condition $\boldsymbol{\omega}_2\cdot\hat{\mathbf{b}}_1 = 0$, where $\boldsymbol{\omega}_2$ is the angular velocity of the frame $\mathcal{R}_2$. Note that because $\hat{\mathbf{b}}_1$ differs from $\hat{\mathbf{z}}$ by terms of order ϵ, the angular velocity $\boldsymbol{\omega}_2$ is of order ϵ. In this second frame, $H_2(\tau) = \mu\mathbf{B}_2(\tau)\cdot\boldsymbol{\sigma}$, where the effective magnetic field $\mathbf{B}_2(\tau)$ is given by $B_1(\tau)\hat{\mathbf{z}} - \mathbf{B}_{L2}(\tau)$, with $\mathbf{B}_{L2}(\tau) = (\epsilon\hbar/\mu)\mathcal{R}_2^{-1}(\tau)\cdot\boldsymbol{\omega}_2(\tau)$. The new Larmor field $\mathbf{B}_{L2}$ is of order ϵ^2.

It is easy to see that this procedure can be iterated; we can transform through successive rotating frames $\mathcal{R}_{n+1}(\tau)$ to compensate for the Larmor fields $\mathbf{B}_{Ln}(\tau)$ induced in the preceding ones. The geometric phase and its corrections emerge when the sequence of rotations is unraveled and the solution transformed back to the fixed frame. One obtains a formal series

$$\gamma_-(C_0) + \sum_{k=1}^{\infty}\gamma_-(C_k), \tag{41}$$

in which the corrections to the geometric phase are themselves geometric phases; $\gamma_-(C_k)$ is half the solid angle of the cycle C_k described by $\mathbf{B}_k(\tau)$.

It turns out that the series (41) is divergent. Although successive cycles C_k initially decrease in extent by a factor of ϵ, asymptotically the terms $\gamma_-(C_k)$ behave as $\epsilon^k k!$, a form characteristic of divergent asymptotic series (Dingle, 1973). An optimal approximation is obtained by truncating the series at its smallest term $k \sim 1/\epsilon$, for which $\epsilon^k k! \sim \exp(-1/\epsilon)$ is exponentially small.

That the series (41) diverges can be understood on general grounds. If, on the contrary, the series converged, along with the infinite sequence of rotating frames, then upon transforming back to the fixed frame, we would obtain an exact solution $|\psi(t)\rangle$ of the Schrödinger equation that would begin and end, at $t = -\infty$ and $t = +\infty$, respectively, in the spin-down state $|-\rangle$. This would imply a vanishing transition probability $\langle +|\psi(+\infty)\rangle$. But it is known that, with the exception of some special Hamiltonians, the transition probability is not zero, but instead [for $\mathbf{B}_0(\tau)$ analytic] is exponentially small, of order $\exp(-1/\epsilon)$ (Dykhne, 1962). These exponentially small transitions (the subject of the following section) signal the ultimate failure of the adiabatic approximation and cannot be captured by the iterative scheme. It follows that this scheme must inevitably diverge.

2.5.2 The Geometric Amplitude Adiabatic evolution is an approximation. For typical slowly varying Hamiltonians $H(\tau)$, where $\tau = \epsilon t$, no matter how small the rate of change ϵ, there are transitions between evolving eigenstates. The transition probability vanishes as ϵ goes to zero, and if H depends analytically on τ, it does so exponentially rapidly (faster than any power of ϵ).

In a study of adiabatic evolution in two-state systems, Berry (1990b) found that this transition probability may contain a geometric factor. The setting is similar to that of Sec. 2.5.1. We consider a two-state Hamiltonian $H(\tau) = \mathbf{R}(\tau)\cdot\boldsymbol{\sigma}$, where $\mathbf{R}(\tau)$ is analytic in the dimensionless time $\tau = \epsilon t$ and is nonvanishing for real τ. As $\tau \to \pm\infty$, $\mathbf{R}(\tau)$ approaches limiting values $\mathbf{R}_{\pm *}$, which need not coincide, so that $\mathbf{R}(\tau)$ need not be closed. From (3), the adiabatic solution for the up state is given by $|\psi(t)\rangle = \exp[i\delta_+(t) + i\gamma_+(t)] \times |+(\mathbf{R}(\tau))\rangle$, where the eigenstates $|\pm(\mathbf{R})\rangle$ are given by (14) and the evolving dynamical and geometric phases $\delta_+(t)$ and $\gamma_+(t)$ are given explicitly by

$$\delta_+(t) = -\frac{1}{\epsilon\hbar}\int_0^\tau R(\tau)d\tau,$$

$$\gamma_+(t) = -\tfrac{1}{2}\int_0^\tau (1-\cos\theta)\dot{\phi}d\tau. \qquad (42)$$

Here (R,θ,ϕ) denote the polar coordinates of $\mathbf{R}(\tau)$, and the overdot denotes the derivative with respect to τ. We have used (15) for the vector potential $\mathbf{A}_+(\mathbf{R})$.

Nonadiabatic transitions may be attributed to roots of $R(\tau)$ in the complex τ plane. The dominant contribution to the transition amplitude comes from the root τ_c closest to the real axis. [We assume that τ_c is a simple zero of $R^2(\tau)$, so that it is a square-root branch point of $R(\tau)$.] Davis and Pechukas (1977) showed that the transition amplitude may be obtained by analytically continuing the adiabatic solution around the branch point τ_c. [Under this continuation, the up and down states $|+(\mathbf{R}(\tau))\rangle$ and $|-(\mathbf{R}(\tau))\rangle$ are interchanged.] The transition probability $P_{+-} = |\langle -(\mathbf{R}(\infty))|\psi(\infty)\rangle|^2$ is given by a product of factors $\exp(-2|\Delta_+|)\exp(2\Gamma_+)$, whose exponents are obtained from the analytical continuation of the dynamical and geometric phases. Explicitly, we have that

$$\Delta_+(t) = \frac{2}{\hbar\epsilon}\,\mathrm{Im}\int_0^{\tau_c} R(\tau)d\tau,$$

$$\Gamma_+ = -\tfrac{1}{4}\,\mathrm{Im}\oint \cos\theta\, d\phi. \qquad (43)$$

In the expression for Γ_+, the contour is taken around the cut joining τ_c and its complex conjugate τ_c^*.

The factor $\exp(\Gamma_+)$ is the geometric amplitude. It depends only on the cycle followed by $\mathbf{R}(\tau)$, not on the slowness parameter ϵ nor on $\hbar$. It can be shown to vanish if $\mathbf{R}(\tau)$ lies in a plane, or if $\mathbf{R}(\tau)$ can be rotated into itself about an axis through the origin. Thus it vanishes for the well-known Landau–Zener model $\mathbf{R}(\tau) = (a,a,A\tau)$ (Zener, 1932), for which $\mathbf{R}(\tau)$ describes a vertical line. A simple example with nonzero geometric amplitude is $\mathbf{R}(\tau) = (a\cos(\omega\tau^2), a\sin(\omega\tau^2), A\tau)$. It describes a helix on the cylinder of radius a, which winds for $\tau < 0$ and unwinds for τ

> 0. For this curve Γ_+ is given by $-a^2\omega \operatorname{sgn} A/A^2$.

The geometric amplitude was observed by Zwanziger *et al.* (1991) in a nuclear-magnetic-resonance experiment (cf. Sec. 5.2). A time-dependent spin Hamiltonian with the requisite properties was generated by applying a modulated rf magnetic field to a sample of spin-$\frac{1}{2}$ nuclei, and the induced magnetization was observed in a rotating frame. The logarithm of the transition probability P_{+-} was plotted as a function of the slowness parameter ϵ. Theory predicts that this plot should be linear (because of the dynamical amplitude) with a nonzero intercept (due to the geometric amplitude). This behavior was confirmed in the experiment, with good quantitative agreement between the observed and predicted values of Γ_+.

3. CLASSICAL SYSTEMS

Following Berry's discovery of the geometric phase, Hannay (1985) found an analogous phenomenon in classical mechanics. Adiabatically cycled classical oscillators undergo a shift in their phase of oscillation. Part of this shift is simply the time integral of the frequency. But there appears in addition a purely geometrical contribution, now known as the Hannay angle. The Hannay angle includes and generalizes such well-known phenomena as the precession of the Foucault pendulum. In many cases it may be directly observed. The Hannay angle emerges as the classical ($\hbar \to 0$) limit of the Berry phase (in cases where chaos is absent from the classical dynamics). In the context of coupled systems, it leads to classical geometric magnetism in the slow dynamics.

Shapere and Wilczek (1989b–1989d) studied phenomena in classical kinematics similar to the Hannay angle. A simple example is the manner in which falling cats and platform divers change their spatial orientations through a cycle of internal motions. Provided their (conserved) angular momentum is zero, the change in their orientation does not depend on the rate at which the internal motions are performed. Classical geometric phases of this type have been investigated systematically by Marsden, Montgomery, and co-workers (Marsden *et al.*, 1991; Marsden, 1992) as part of a general theory of Hamiltonian dynamics with symmetry.

3.1 The Hannay Angle

3.1.1 The Classical Adiabatic Theorem

Imagine a swinging pendulum slowly shortened to half its original length. During this process, is anything conserved, either exactly or approximately? This question is addressed by the classical adiabatic theorem. While energy is not conserved (work is being done), there is an approximately conserved quantity, namely, the classical adiabatic invariant, or the *action.*

The classical adiabatic theorem is most simply described for systems of one degree of freedom, and so we concentrate on this case. The dynamics in the two-dimensional phase plane is described by Hamilton's equations $\dot{q} = \partial_p H$, $\dot{p} = -\partial_q H$. The classical Hamiltonian $H(q,p,\mathbf{R})$ is taken to depend on a vector $\mathbf{R}$ of parameters (like the quantum Hamiltonians of Sec. 1). If these are held fixed, then energy is conserved, and orbits lie on contours of H. Let us assume that the motion is bounded, so that the energy contours are closed. The orbits go around them periodically with oscillation frequency $\omega = (\partial_E I)^{-1}$, where

$$I(E,\mathbf{R}) = \frac{1}{2\pi}\int_{H(q,p,\mathbf{R})\leq E} dq\,dp \tag{44}$$

is the action. Apart from the factor of 2π, the action is just the area enclosed by the orbit of energy E. If the parameters vary in time, then energy is no longer conserved. However, if the variation is slow compared to the oscillation frequency ω, then the action (44) is nearly conserved, and the variation of the energy is determined implicitly by

$$I(E(t),\mathbf{R}(t)) = I_0 \overset{\text{def}}{=} I(E(0),\mathbf{R}(0)). \tag{45}$$

This is the classical adiabatic theorem.

In higher dimensions, the classical adiabatic theorem generalizes provided the dynamics is either *integrable* or *ergodic.* Integrable systems are characterized by a high degree of symmetry. They possess N independent constants of the motion (where N is the number of degrees of freedom), which are in involution (that is, the Poisson bracket between any pair of them vanishes). Typical orbits in an integrable system execute periodic or quasiperiodic motion on N-dimensional tori embedded in the $2N$-dimensional phase space, and the equations of motion may be solved more or less explicitly. In sharp contrast stand ergodic systems, for which energy is the only constant of the motion. Typical orbits cover the $(2N - 1)$-dimensional energy surface uniformly, so that time averages and microcanonical ensemble averages coincide. The adiabatic theorems in higher dimensions are rather more difficult to state and considerably harder to prove. The difficulty is related to the fact that the adiabatic invariants are not so well conserved as in the one-dimensional case (Lochak and Meunier, 1988).

3.1.2 Basic Account of the Hannay Angle The classical adiabatic theorem (45) determines the time evolution of the energy of a one-dimensional oscillator under a slowly changing Hamiltonian. Hannay (1985), motivated by Berry's analysis of the quantum geometric phase, asked a more refined question: What is the evolution of the phase of the oscillator?

To frame this question precisely we introduce canonical coordinates, the *action–angle variables* (θ,I). These are related to (q,p) by a canonical transformation $q = q(\theta,I,\mathbf{R})$, $p = p(\theta,I,\mathbf{R})$ that depends on parameters. The action I, the generalized momentum, determines the energy of the oscillator according to (44). The angle θ, the generalized coordinate, describes the phase of the oscillator, and varies periodically around the orbit in the phase plane with period 2π. Expressed in terms of action–angle variables, the Hamiltonian $H(I,\mathbf{R})$ is independent of θ, so that the canonical equations of motion, $\dot{\theta} = \partial_I H$, $\dot{I} = -\partial_\theta H$, imply that I is constant while θ evolves at a constant frequency $\omega \overset{\text{def}}{=} \partial_I H$.

Now suppose the parameters $\mathbf{R}$ of the Hamiltonian are taken slowly around a closed cycle C, returning to their original values $\mathbf{R}(0)$ after a time T. Since the action is nearly conserved, a trajectory returns approximately to the orbit, or contour of $H(I,\mathbf{R}(0))$, on which it started. Of course, it will be displaced by an angle $\Delta\theta$ from its initial position. Hannay obtained the two dominant contributions to $\Delta\theta$ (further corrections are of order $1/T$). The first is the *dynamical angle* $\theta_d = \int_0^T \omega(I,\mathbf{R}(t))dt$. It is simply the time integral of the frequency, taking into account its slow variation with parameters. The second, unexpected, contribution,

$$\psi_g(I,C) = -\partial_I \oint_C \mathbf{A}(I,\mathbf{R})\cdot d\mathbf{R}, \tag{46}$$

is the *geometric angle*, or the *Hannay angle*. It is expressed in terms of the line integral around C of a vector field $\mathbf{A}(I,\mathbf{R})$ given by

$$\mathbf{A}(I,\mathbf{R}) = \frac{1}{2\pi}\int_0^{2\pi} p(\theta,I,\mathbf{R})\nabla q(\theta,I,\mathbf{R})d\theta. \tag{47}$$

The Hannay angle is geometric in the same sense as is the geometric phase; it depends only on the parameter cycle C and not on the rate at which it is traversed. Using Stokes's theorem, the Hannay angle can also be expressed as the surface integral $-\partial_I\int_S \mathbf{V}(I,\mathbf{R})\cdot d\mathbf{S}$, where S is a surface bounded by C and $\mathbf{V} = \nabla \times \mathbf{A}$. From (47),

$$\mathbf{V}(I,\mathbf{R}) = \frac{1}{2\pi}\int_0^{2\pi} \nabla p(\theta,I,\mathbf{R}) \times \nabla q(\theta,I,\mathbf{R})d\theta. \tag{48}$$

As with the geometric phase, the presence of time-reversal symmetry (cf. Sec. 1.5) imposes constraints on the Hannay angle. In particular, if the orbits themselves are time-reversal symmetric (for all values of parameters), then the Hannay angle vanishes identically.

Like the geometric phase, the Hannay angle may be interpreted as the holonomy of a connection on a fiber bundle (cf. Sec. 2.1.3). The parameters $\mathbf{R}$ constitute the base manifold, and the fibers attached to them are orbits of the Hamiltonian $H(q,p,\mathbf{R})$ with given action I. These orbits are parameterized by the angle θ, and the connection $\partial_I\mathbf{A}(I,\mathbf{R})$ describes how θ changes under parallel transport.

The Hannay angle can be detected by determining the initial and final angles of an adiabatically cycled oscillator, and subtracting from their difference the dynamical angle θ_d. As θ_d depends only on the parameter cycle $\mathbf{R}(t)$, it can be computed without solving the equations of motion explicitly. In this way, Hannay's analysis, together with the adiabatic theorem (45), provides a complete asymptotic solution of the equations of motion for slowly varying Hamiltonians in one dimension. The analysis generalizes directly to integrable systems in higher dimensions, where there are as many Hannay angles as there are degrees of freedom. However, detecting them can be a more delicate matter, as the adiabatic theorem and corrections thereto become more complicated in higher dimensions (Golin *et al.*, 1989). For ergodic systems, there is no general analog of the Hannay angle, though a related holonomy phenomenon emerges in some special cases (Robbins, 1994; Jarzynski, 1995; Schroer, 1996).

3.1.3 Examples A bead moving freely around a planar wire loop provides a simple example of the Hannay angle. Suppose the bead has unit mass, and the loop has circumference C, area A, and is otherwise arbitrary in shape. While the system as a whole has time-reversal symmetry, the individual orbits do not (they are either clockwise or counterclockwise). The action I is just the velocity v, and the angle $\theta = 2\pi s/C$ is proportional to the distance s along the wire. Suppose the loop is slowly rotated once around an interior point over a time T, and let Δs denote the total distance the bead has traveled during this time. The dynamical contribution s_d (which is $C/2\pi$ times the dynamical angle θ_d) is just vT (v is the adiabatic invariant). The geometric contribution s_d may be computed to be $-4\pi A/C^2$. It accounts for the distance the bead has slipped as the wire turned. For a fixed circumference C, $-s_d$ assumes its maximal value, C, for a circular loop; in this case, the bead simply slips a full revolution. For thin, elongated loops, s_d approaches zero; the only slippage occurs at the hairpin turns.

The second example is the family of harmonic oscillators $H = (Z - X)p^2/2 + Ypq + (Z + X)q^2/2$, where the term in qp breaks time-reversal symmetry. For $X^2 + Y^2 < Z^2$, the motion is bounded and oscillatory, and we consider adiabatic cycles confined to this region of parameter space. The flux field $\mathbf{V}$ of (48) may be computed to be

$$\mathbf{V}(I,\mathbf{R}) = -\frac{I}{2}\frac{\mathbf{R}}{(Z^2 - X^2 - Y^2)^{3/2}}. \qquad (49)$$

Note that because $\mathbf{V}$ is linear in I, the Hannay angle θ_g is independent of I. θ_g may be described geometrically as follows. We regard the parameter space as three-dimensional Minkowski space, in which X and Y are spacelike coordinates, Z is timelike, and the region $X^2 + Y^2 < Z^2$ is the interior of the light cone. Then $\theta_g(C)$ is the area of the projection of the closed curve C onto the unit mass shell $Z^2 - X^2 - Y^2 = 1$.

The precession of a Foucault pendulum is yet another example of a Hannay angle. Consider a spherical pendulum initially set to oscillate in a plane. During the course of a day, the axis of gravity (whose direction represents the parameters of the pendulum) is carried once around a line of latitude α. At the day's end, the plane of oscillation has turned through an angle $2\pi(1 - \cos\alpha)$. This is just the solid angle described by the axis of gravity and may be computed as a Hannay angle using (46). Indeed, the phenomenon described by the Hannay angle is more general. If one turned the Earth about its center adiabatically (but otherwise arbitrarily) before returning it to its original orientation, the plane of oscillation of the spherical pendulum would still precess through an angle equal to the solid angle described by the Earth's axis.

The spherical pendulum is a two-degree-of-freedom system, integrable because of its axial symmetry. Therefore, it has two Hannay angles, one each for its polar and azimuthal degrees of freedom. For planar oscillations, however, only the azimuthal Hannay angle, which describes the Foucault precession, is nonzero.

Let us briefly mention two final examples drawn from plasma physics and astronomy, respectively. Littlejohn (1988) found a Hannay angle in the cyclotron oscillations of charged particles in strong magnetic fields. Berry and Morgan (1996) (see also Golin *et al.*, 1989) compute "the Hannay angle of the world," an annual displacement of the Earth along its orbit of approximately -150 m (or -2×10^{-4} arc sec) caused by the adiabatic and periodic variation of Jupiter's gravitational field.

3.1.4 Semiclassical Limit Following Hannay's work, Berry (1985) showed that the Hannay angle emerges in the semiclassical limit of the geometric phase. The precise relation between the two is given by

$$\gamma_{n+1}(C) - \gamma_n(C) \xrightarrow{\hbar\to 0, n\to\infty} -\theta_g(I,C), \tag{50}$$

in which the classical action is related to the quantum number by the semiclassical quantization condition $I = n\hbar$. Expression (50) conforms to the usual relation between quantum phases and classical angles, the most familiar instance being the relation $(E_{n+1} - E_n)/\hbar \to \omega$ between energy levels and frequencies, whose time integral connects the dynamical phase and the dynamical angle. The result (50) is derived by substituting the WBK approximation for the eigenstates $|n(\mathbf{R})\rangle$ into formula (4) for the quantum vector potential $\mathbf{A}_n$. One obtains thereby expression (47) for the classical vector potential. This analysis extends to integrable systems in more than one dimension, for which semiclassical eigenfunctions may be similarly determined.

On the other hand, for systems whose classical dynamics is chaotic, the quantum geometric phase has no classical limit in general. However, the vector field $\mathbf{V}_n$, when averaged over quantum number, does have a classical limit (Robbins and Berry, 1992), which turns out to play a role in the classical Born–Oppenheimer approximation (cf. Sec. 3.1.5). The classical limit

$$\begin{aligned}\mathbf{V}(E,\mathbf{R}) = {} & \frac{1}{2\partial_E \Omega(E,\mathbf{R})} \\ & \times \partial_E\left[\partial_E \Omega(E,\mathbf{R}) \int_0^\infty \langle \nabla H_t(\mathbf{q},\mathbf{p},\mathbf{R}) \right. \\ & \left. \times \nabla H(\mathbf{q},\mathbf{p},\mathbf{R})\rangle_E dt\right]\end{aligned} \tag{51}$$

is expressed in terms of the time integral of an antisymmetric correlation function of the parameter gradient of the Hamiltonian, $\nabla H(\mathbf{q},\mathbf{p},\mathbf{R})$. $\Omega(E,\mathbf{R}) = \int_{H(\mathbf{q},\mathbf{p},\mathbf{R})\leq E} d^N q d^N p$ denotes the phase volume inside the energy shell $H = E$, $\langle\cdot\rangle_E$ denotes the normalized microcanonical average, and ∇H_t denotes the function ∇H evolved for a time t with the classical dynamics. In one dimension, (51) is equivalent to (48).

3.1.5 Classical Reaction Forces As in Sec. 1.7, we suppose that the parameters $\mathbf{R}$ are no longer externally prescribed, but instead are the coordinates of a slow system to which the original one, the one-dimensional oscillator of Sec. 3.1.1, is coupled. Both systems are governed by classical mechanics. We take their Hamiltonian to be $\mathbf{P}^2/2M + H(q,p,\mathbf{R})$, where $\mathbf{P}$ is the slow momentum. The Hannay angle describes a holonomy in the fast dynamics induced by the slow dynamics. As in the quantum case, there is a reciprocal effect on the slow dynamics induced by the fast dynamics. This effect is found by averaging over the fast motion. According to the adiabatic theorem, the action of the fast system is nearly constant. In the lowest-order approximation, we treat the action as strictly constant. Then the energy $H(I,\mathbf{R})$ acts as a potential in the slow dynamics, whose effective Hamiltonian is given by $\mathbf{P}^2/2M + H(I,\mathbf{R})$. This is the classical analog of the Born–Oppenheimer approximation.

First-order corrections account for the fluctuations of the action about its mean. These generate an effective vector potential in the slow Hamiltonian, so that $\mathbf{P} \to \mathbf{P} - \mathbf{A}(I,\mathbf{R})$. The vector potential is precisely the field (47) whose line integral gives the Hannay angle. This is classical geometric magnetism.

One example of geometric magnetism is the precession of a symmetric top (Berry and Robbins, 1993a). In this case, the top's axis is the slow system, and the rotation about this axis is the fast system. Another example from plasma physics is the drift of the guiding center of electrons in strong, slowly varying magnetic fields (Littlejohn, 1988). The guiding center, i.e., the center of the instantaneous Larmor orbit, constitutes the slow degrees of freedom (it remains fixed in constant fields), and the motion about the guiding center, the fast degrees of freedom.

Geometric magnetism generalizes to higher-dimensional fast systems, both integrable and chaotic. In the latter case, geometric magnetism is partnered by an additional dissipative force, deterministic friction (Wilkinson, 1990); the slow dynamics is no longer conservative, as the chaotic motion of the fast system acts as an effective heat bath. The magnetic and dissipative forces appear as the antisymmetric and symmetric parts of a linear-response tensor (Berry and Robbins, 1993b; Jarzynski 1993).

3.2 Symmetry and Kinematic Geometric Phases

Kinematic geometric phases arise in systems with symmetry. They describe the evolution in degrees of freedom upon which the symmetry acts (e.g., orientational degrees of freedom in case of rotational symmetry). Their structure depends only on how the symmetry is realized and not on the details of the dynamics. The geometric character of their evolution is exactly enforced by the equations of motion and does not depend on assumptions of adiabaticity. In the following sections, two different treatments of kinematic geometric phases are described.

3.2.1 Falling Cats and Swimming Paramecia Even with nothing to push against, bodies in free fall, such as cats and platform divers, can change their orientation through internal motions. Perhaps counterintuitive at first, this phenomenon is a direct consequence of conservation of angular momentum and can be calculated as such. The main difficulty in doing so lies in the definition of the body's orientation. This problem does not arise for rigid bodies, whose every configuration is related to a chosen reference by a unique rotation. But for a body that can bend and twist and stretch, i.e., that has internal degrees of freedom, it is not clear which configurations are "upright" and which ones are "tilted."

There are various prescriptions for determining orientation. One is to refer to the orientation of the principal axes of inertia. Another is through the Eckart frame, a common convention in molecular physics. Both prescriptions have particular advantages; neither is well defined in all cases.

In fact, the structure of the solution emerges most clearly if the definition of orientation is made more or less arbitrarily. This is the point of view taken by Shapere and Wilczek (1989b). For simplicity, suppose the body consists of a finite number N of masses m_α at positions $\mathbf{r}_\alpha$. No assumption is made concerning their dynamics, except that they obey Newton's laws and that there are no external forces. By choosing a suitable reference frame, we may assume that the center of mass $\Sigma_\alpha m_\alpha \mathbf{r}_\alpha / \Sigma_\alpha m_\alpha$ begins and remains at the origin.

By some means, certain configurations $(\mathbf{R}_1, \ldots, \mathbf{R}_N)$ are designated to be "upright." This designation must be consistent, in that if two configurations are related by a rotation, then only one of them can be upright. It must also be complete, so that every configuration can be made upright by applying a rotation. If both these conditions hold, then every configuration $(\mathbf{r}_1, \ldots, \mathbf{r}_N)$ can be obtained by applying a unique rotation $\mathcal{R}$ to a unique upright configuration $(\mathbf{R}_1, \ldots, \mathbf{R}_N)$. The body's motion,

$$\mathbf{r}_\alpha(t) = \mathcal{R}(t) \cdot \mathbf{R}_\alpha(t), \qquad \alpha = 1, \ldots, N, \tag{52}$$

is then determined by the evolution of its orientation $\mathcal{R}(t)$ and its internal dynamics $\mathbf{R}_1(t), \ldots, \mathbf{R}_N(t)$. In the language of rigid-body dynamics, $\mathbf{R}_\alpha(t)$ describes the motion in the body frame, and $\mathbf{r}_\alpha(t)$, the motion in the space frame.

An object fixed in the space frame appears to rotate with angular velocity $-\boldsymbol{\Omega}(t)$ in the body frame. $\boldsymbol{\Omega}(t)$ is defined by the relation

$$\dot{\mathcal{R}} \cdot \mathbf{X} = \mathcal{R} \cdot (\boldsymbol{\Omega} \times \mathbf{X}), \tag{53}$$

which can be made to hold for arbitrary vectors $\mathbf{X}$.

Regardless of the details of the dynamics, the total angular momentum $\mathbf{l} = \Sigma_\alpha m_\alpha \mathbf{r}_\alpha \times \mathbf{v}_\alpha$ in the space frame must be conserved. Conservation of angular momentum leads to the relation

$$\boldsymbol{\Omega}(t) = \boldsymbol{I}^{-1} \cdot \left(\mathbf{L} - \sum_\alpha m_\alpha \mathbf{R}_\alpha \times \mathbf{V}_\alpha \right) \tag{54}$$

between the angular velocity $\boldsymbol{\Omega}$ and the angular momentum $\mathbf{L} = \mathcal{R}^{-1} \cdot \mathbf{l}$, velocities $\mathbf{V}_\alpha = \dot{\mathbf{R}}_\alpha$, and inertia tensor $I_{ij} = \Sigma_\alpha m_\alpha (R_\alpha^2 \delta_{ij} - R_{\alpha i} R_{\alpha j})$ in the body frame. Once the internal dynamics $\mathbf{R}_\alpha(t)$ is known, the relations (53) and (54) may be used to determine the orientation $\mathcal{R}(t)$.

The case of zero angular momentum ($\mathbf{l} = \mathbf{L} = 0$) is particularly interesting. Then one can show that the orientation $\mathcal{R}(t)$ does not depend on the rate of the internal dynamics. [One can, in effect, multiply through by dt in (53) and (54).] In particular, if the internal motion is periodic along a cycle C, then after one period it generates an overall rotation $\mathcal{R}_g = \mathcal{R}(T)\mathcal{R}(0)^{-1}$ of the body. The rotation $\mathcal{R}_g$ depends only on the cycle C, and not on the rate at which it is traversed. It is a purely geometrical quantity, a non-Abelian analog (rotations do not commute) of the Hannay angle. It is the means by which a cat dropped upside down manages to land on its feet.

Like the other geometric phases considered so far, this can be described in terms of holonomy on a fiber bundle. The fiber bundle is the configuration space of the masses m_α, constrained to have their center of mass at the origin. The base manifold consists of the distinct internal configurations, no two of which can be made to coincide by a rotation. (Mathematically, this is the quotient of the configuration space by the rotation group.) It can be parametrized by the upright configurations, provided these are always well defined. The fiber consists of the different possible orientations of a given internal configuration and may be parametrized by the rotation group. A connection determines an infinitesimal rotation $\delta\mathcal{R}$ for an infinitesimal change of internal configuration $\delta\mathbf{R}_\alpha$, and the physically relevant connection is determined by conservation of angular momentum. Parallel transport around a closed cycle C in the base yields the holonomy $\mathcal{R}_g$. The curvature of this connection, which describes the net rotation after an infinitesimal cyclic change of shape, may be computed with a formula analogous to (28).

The designation of upright configurations is arbitrary, and one can always choose them differently. It can be verified that the holonomies $\mathcal{R}_g$ do not depend intrinsically on this choice and that the various quantities that describe the orientation transform appropriately when a different choice is made. In particular, the curvature and connection transform according to equations similar to (30).

A more complex example studied by Shapere and Wilczek (1989c) concerns swimming through viscous fluids at low Reynolds number. High viscosity and low velocity means that bodies remain at rest unless acted upon by a force—a realization of Aristotelian mechanics! Bodies locomote by executing a series of internal motions, i.e., swimming strokes. Conservation of linear and angular momentum determine the consequent changes in position and orientation, but now the motion of the fluid must be taken into account. Solving the fluid-dynamical problem leads to a relation similar to (54) for both the linear and angular velocity in terms of rates of change of shape. Applications in biology include locomotion of microorganisms.

3.2.2 Geometric Phases and Reduction

Kinematic geometric phases can be framed in terms of the general theory of symmetry in Hamiltonian systems, called *reduction.* This formulation, extensively developed by Marsden, Montgomery, and co-workers (Marsden *et al.*, 1991; Marsden, 1992), has led to a unified perspective on a variety of geometric-phase phenomena as well as a number of interesting examples.

Dynamics is simplified by the presence of symmetry, and Hamiltonian dynamics especially so. A symmetry always reduces the dimension of a dynamical system by one, but in the Hamiltonian case, it reduces it by two. Indeed, when there is sufficient symmetry, the system is integrable, and the dynamics can be solved analytically (cf. Sec. 3.1.1).

The simplest symmetry is translation invariance in an ignorable coordinate q_N. Hamilton's equation $\dot{p}_N = -\partial H/\partial q_N = 0$ (or, more generally, Noether's theorem) implies that the conjugate momentum p_N is conserved. Fixing its value $p_N = p_{N0}$ leads to a Hamiltonian $H(q_1, \ldots, q_{N-1}, p_1, \ldots, p_{N-1}, p_{N0})$ in two fewer variables. Once their dynamics is obtained, the evolution of the ignorable coordinate may be determined by integrating Hamilton's equation $\dot{q}_N = \partial_{p_N} H(q_1, \ldots, q_{N-1}, p_1, \ldots, p_{N-1}, p_{N0})$. The theory of reduction extends this analysis to general symmetries. As in the simple example, the analysis consists of two stages. The first is reduction, wherein the conserved momenta are fixed, the ignorable coordinates ignored, and Hamilton's equations expressed in terms of a reduced set of coordinates. In all but the simplest cases, finding these coordinates is a nontrivial task (Jacobi's "elimination of the node" in the rotationally symmetric N-body problem is a representative example). The second state is reconstruction, wherein the evolution of the ignorable coordinates is determined from the reduced dynamics. It is in the reconstruction stage that kinematic geometric phases can appear. It turns out that the equations of motion for the ignorable coordinates can be separated into dynamical and geometric terms. The geometric term depends only on the path described by the reduced dynamics, and not on the rate at which it is traversed. The dynamical term, by contrast, does depend on the rate. Moreover, the form of the geometrical term depends only on the realization of the symmetry, and not on the details of the Hamiltonian.

In the example discussed in the preceding section, the flexible body composed of N point masses, the ignorable coordinates correspond to $\mathcal{R}$, the body's orientation; the conserved momentum to the space-fixed angular momentum $\mathbf{l}$; and the internal coordinates $\mathbf{R}_\alpha(t)$ to the reduced system. The evolution of $\mathcal{R}(t)$ is obtained by integrating the linear system (53), in which angular velocity $\boldsymbol{\Omega}(t)$ is given in terms of the reduced dynamics by (54). Equation (54) is an example of a reconstruction formula divided into its dynamical and geometric parts. The first term, $\boldsymbol{I}^{-1}\cdot\mathbf{L}$, is the dynamical contribution. The second, $\boldsymbol{I}^{-1}\cdot(\Sigma_\alpha m_\alpha \mathbf{R}_\alpha \times \mathbf{V}_\alpha)$, is the geometric contribution. As discussed above, its integral depends only on the path C described by the internal dynamics $\mathbf{R}_\alpha(t)$.

Montgomery (1991) found an even simpler example in the dynamics of a free rigid body. Its configuration (with its center of mass fixed at the origin) is described by a rotation $\mathcal{R}$, for which it will be convenient to use the Euler-angle parametrization, $\mathcal{R} = \mathcal{R}_z(\phi)\mathcal{R}_y(\theta)\mathcal{R}_z(\psi)$. The space-fixed angular momentum $\mathbf{l}$ is a constant of the motion, and without loss of generality we can assume it lies in the z direction. The body-fixed angular momentum $\mathbf{L}(t) = \mathcal{R}(t)\cdot\mathbf{l}$ has polar coordinates $(\theta(t),\phi(t))$ on the sphere of constant radius $L = |\mathbf{L}| = |\mathbf{l}|$. Its dynamics is determined by the Euler equations $\dot{\mathbf{L}} = \boldsymbol{\Omega} \times \mathbf{L}$, where $\boldsymbol{\Omega} = \boldsymbol{I}^{-1}\cdot\mathbf{L}$ is the angular velocity and $\boldsymbol{I}$ is the inertia tensor. The Euler equations describe the reduced dynamics, in which $\mathbf{L}$, or equivalently θ and ϕ, are the reduced coordinates. The ignorable coordinate is the third Euler angle ψ.

Under the Euler equations, $\mathbf{L}$ executes a periodic motion with period T. The fact that $\mathbf{L}(T) = \mathbf{L}(0)$ implies that the rigid body returns after one period to its original orientation, up to a rotation about the direction of $\mathbf{L}(T)$. The amount of rotation is just $\Delta\psi = \psi(T) - \psi(0)$, the shift in the ignorable coordinate ψ.

With a little algebra, the equation of motion for ψ may be written in the form

$$\dot{\psi} = 2E/L + \cos\theta\dot{\phi}, \qquad (55)$$

where $E = \frac{1}{2}\mathbf{L}\cdot\boldsymbol{I}^{-1}\cdot\mathbf{L}$ is the conserved rotational energy. Equation (55) is another example of a reconstruction formula divided into dynamical and geometric contributions. In this case, the dynamical contribution integrates to $2ET/L$, whereas the geometric contribution gives (modulo 2π) the solid angle $\oint_C \cos\theta d\phi$ described by the body-fixed angular momentum $\mathbf{L}(t)$ over a single period.

Another example, due to Alber and Marsden (1992), concerns phase shifts in soliton collisions. A fundamental property of certain nonlinear partial differential equations, such as the Korteweg-de Vries equation $U_t + 6UU_x + U_{xxx} = 0$, is the existence of *solitons.* These are wave forms $U(x - vt)$ that propagate at a fixed velocity v while maintaining their shape, the nonlinearity in the wave dynamics just balancing the dispersion. When two solitons collide, they emerge after a complicated but transient period of interaction essentially intact. The only signature of their collision is a shift Δx in their position $x \pm \Delta x - vt$ (+ for the faster soliton, − for the slower one). This shift, which may be determined from an asymptotic analysis of the exact solutions, can be interpreted as a geometric phase. The same analysis applies to a variety of integrable nonlinear partial differential equations, including the nonlinear Schrödinger equation and the Klein–Gordon equation.

3.2.3 Optimization and Control

The study of kinematic geometric phases suggests a problem in optimization. What is the most efficient internal motion for producing a given holonomy? As a specific example, consider a satellite orbiting the Earth with a fixed orientation, so that its angular momentum about its center of mass is zero. It is desired to change its orientation using as little energy as possible. The obvious method of applying external torques (supplied by jets attached to satellite, for example) has the drawback that any uncompensated force or torque leaves the satellite in a state of translation or rotation. The alternative is to execute a sequence of internal motions (e.g., flywheels driven by motors); in the manner described above, these generate a change in orientation without altering the total linear or angular momentum.

The problem of finding the most efficient cycle with a given holonomy leads to a constrained variational problem in the internal degrees of freedom, which can be formulated in terms of control theory. The solution depends very much on how efficiency is measured. In certain circumstances, it may be advantageous to make some sizable internal motions initially in order to achieve a configuration for which the geometric phase curvature form (cf. Sec. 2.1.2) is large; then much smaller subsequent motions can produce relatively large effects. In other situations, it may be necessary to restrict to small motions about some equilibrium configuration; in this case, the variational problem may be analyzed in terms of infinitesimal circuits. The solutions often have a physical interpretation in terms of motion in a magnetic, or more generally a Yang–Mills, field. Shapere and Wilczek (1989d) find efficient swimming strokes through viscous media in two and three dimensions under various conditions. Marsden (1992) describes a variety of other examples.

4. OPTICS

Optics has provided fertile ground for studying the geometric phase. A comprehensive survey of experimental activity is given by Bhandari (1997). Most investigations have concentrated on polarization effects. The polarization of a beam of monochromatic, coherent light behaves like a quantum mechanical spin, and under cyclic evolutions (and noncyclic ones as well) acquires a geometric phase similar to that of a spinning particle in a magnetic field (cf. Sec. 1.3). There are various ways to change polarization. One is by varying the direction of propagation, as the polarization vector must remain transverse (Tomita and Chaio, 1986); in this case the polarization has a spin-1, or vectorial, character. Another is through optical activity induced by anisotropic media, in which case the polarization has a spin-$\frac{1}{2}$ character (Pancharatnam, 1956; Bhandari and Samuel, 1988; Bhandari, 1988).

A beam of light passing through a cycle of polarization states may be recombined with a second beam whose polarization has remained unchanged. Their total phase difference (dynamical + geometric) can be determined, for example, from their interference pattern. In certain cases, the dynamical phase shift can be made to vanish. In others, it can be explicitly computed and subtracted from the total phase. In either case, one obtains a measurement of the geometric phase.

4.1 Geometric Phase of Coiled Light

In a straight, single-mode optical fiber of circular cross section, there is no preferred state of polarization; whatever the polarization state of the source, it is preserved along the fiber. This is no longer so if the fiber is made to bend; then the polarization must change in order to remain transverse to the direction of propagation $\hat{\mathbf{t}}$. Provided the radius of curvature $R = \|d\hat{\mathbf{t}}/ds\|^{-1}$ is large compared to the wavelength λ (here s measures distance along the fiber), the polarization vector $\hat{\mathbf{d}}(s)$ of the electric displacement $\mathbf{D}(s,\rho) = \operatorname{Re} e^{i(2\pi s/\lambda - \omega t)} f(\rho)\hat{\mathbf{d}}(s)$ [here $f(\rho)$ is the radial modal profile] evolves along the fiber according to

$$\frac{d\hat{\mathbf{d}}}{ds} = -(\hat{\mathbf{t}}\cdot\hat{\mathbf{d}})\hat{\mathbf{t}}. \qquad (56)$$

The corrections to this approximate solution of Maxwell's equations are of order λ/R.

For a general elliptical polarization, $\hat{\mathbf{d}}$ is complex. Within the adiabatic approximation, both its real and imaginary parts $\hat{\mathbf{d}}_r$ and $\hat{\mathbf{d}}_i$ remain orthogonal to the direction of propagation $\hat{\mathbf{t}}$, and each satisfies (56) separately. As a function of s, $\hat{\mathbf{t}}(s)$ describes a curve C on the unit sphere, and the vectors $\hat{\mathbf{d}}_r(s)$ and $\hat{\mathbf{d}}_i(s)$ are tangent to the sphere at $\hat{\mathbf{t}}(s)$. Equation (56) is the condition for their parallel transport. (Infinitesimal parallel transport of a tangent vector $\hat{\mathbf{d}}$ from $\hat{\mathbf{t}}$ to $\hat{\mathbf{t}} + d\hat{\mathbf{t}}$ is accomplished by first rigidly transporting $\hat{\mathbf{d}}$ to $\hat{\mathbf{t}} + d\hat{\mathbf{t}}$ and then projecting the translated vector onto the plane tangent to $\hat{\mathbf{t}} + d\hat{\mathbf{t}}$.) If the curve C is made to close on itself, the vectors $\hat{\mathbf{d}}_r$ and $\hat{\mathbf{d}}_i$ return rotated through an angle θ_g equal to the solid angle subtended by C. θ_g is a geometric phase, depending only on the shape of C.

Tomita and Chiao (1986) observed this effect by introducing linearly polarized light from a He-Ne laser into a length of helically wound optical fiber whose initial and final directions coincided. The direction of linear polarization of the emergent beam was measured and found to have turned through an angle equal to the solid angle subtended by the cycle of directions. The experiment was performed with helices of both fixed and variable pitch, as well as for fibers constrained to lie in the xy plane. In this last case, the cycle of directions lies on the equator of the unit sphere, and the subtended solid angle is a multiple of 2π. As predicted, the direction of linear polarization remained unchanged.

The original analysis of the Tomita–Chiao experiment was based on quantum-mechanical considerations, treating a photon propagating in the fiber as a spin-1 particle whose helicity is constrained to be ± 1. (Classically, this corresponds to left- and right-circularly polarized light.) Just as a spin with nonzero magnetic moment tracks the direction of a slowly changing magnetic field (cf. Sec. 1.3), the helicity of the photon tracks the direction of propagation $\hat{\mathbf{t}}$ and acquires a geometric phase equal to the solid angle described by $\hat{\mathbf{t}}$. By decomposing a linear polarization into a superposition of left- and right-circular polarizations, the observed rotation of linear polarization may be recovered.

A systematic derivation requires a more detailed analysis, most easily carried out within the context of classical (rather than quantum) electromagnetic theory. Indeed, the essential features of such an analysis are contained in the work of Rytov (1938) and Vladimirsky (1941), who derived and interpreted the parallel-transport law (56) in the short wavelength, or geometrical optics, limit (see Berry, 1990a, for a discussion). Strictly speaking, geometrical optics is not appropriate for single-mode fibers, in which the transverse wavelength is comparable to the fiber diameter. Berry (1987a) has carried out a classical analysis using the full Maxwell equations, and it is the conclusions of this analysis that are described above.

Kitano *et al.* (1987) performed an experiment related to Tomita and Chiao's, in which the cyclic change in the direction of propagation was accomplished nonadiabatically through a series of discrete reflections from mirrors. These reflections reverse helicity instead of conserving it, and the calculation of the geometric phase must be appropriately modified.

4.2 Pancharatnam's Phase

When light passes through a homogeneous medium, its direction of propagation remains constant. However, if the medium is optically anisotropic, the polarization can change. Such changes in polarization are accompanied by a geometric phase, discovered by Pancharatnam in 1956. (See Berry, 1987b, for a contemporary discussion of Pancharatnam's work.) For definiteness, let us consider plane electromagnetic waves propagating in the z direction with fixed wave number k and frequency ω. The electric field $\mathbf{E}(\mathbf{r},t) = \mathrm{Re}\sqrt{I}e^{i(kz-\omega t)}|A\rangle$ is described by its intensity I_A and normalized polarization vector $|A\rangle = (A_x,A_y)$, whose complex components determine the phase and relative intensity along x and y. (Dirac notation is used for complex two-vectors to highlight an analogy with two-state systems, introduced below.)

Given two such waves, what is their relative phase? Unless their polarization states $|A\rangle$ and $|B\rangle$ happen to be proportional, it is not clear that the question is a sensible one. Pancharatnam found a natural and physically relevant prescription for their relative phase by considering the intensity I_{A+B} of the superposition, $\sqrt{I_A}|A\rangle + \sqrt{I_B}|B\rangle$. This intensity is given by

$$I_{A+B} = I_A + I_B + 2\sqrt{I_A I_B}\,\mathrm{Re}\langle A|B\rangle. \tag{57}$$

If the constituent intensities I_A and I_B are fixed, the superposed intensity depends only on the interference term and is maximized when $\langle A|B\rangle = 1$. This is Pancharatnam's criterion for the two polarization states $|A\rangle$ and $|B\rangle$ to be in phase. In general, their relative phase is defined to be $\arg\langle A|B\rangle$ (this is well defined unless $|A\rangle$ and $|B\rangle$ are orthogonal). It is the amount by which the phase of $|A\rangle$ must be retarded relative to the phase of $|B\rangle$ in order that the intensity of their superposition be a maximum.

Being in phase is not a transitive relation; if $|A\rangle$ is in phase with $|B\rangle$ and $|B\rangle$ is in phase with $|C\rangle$, then, in general, $|C\rangle$ is not in phase with $|A\rangle$. The relative phase $\arg\langle C|A\rangle$ has a simple geometrical interpretation, which we now describe. To every polarization state $|A\rangle$ may be associated a vector $\mathbf{e}_A$ on the unit sphere, namely, the unit vector $\langle A|\boldsymbol{\sigma}|A\rangle$ (here $\boldsymbol{\sigma}$ denotes the Pauli matrices) rotated through $\pi/2$ about the y axis. This is how polarization is represented, up to an overall phase factor, on the *Poincaré sphere.* The y rotation, a matter of convention, puts the left and right circular polarizations at the north and south poles, respectively, of the Poincaré sphere, and the linear polarizations at the equator. (In more mathematical language, the Poincaré sphere is the projective Hilbert space for two-state systems—cf. Sec. 2.3.)

One can show that the relative phase $\arg\langle C|A\rangle$ is equal to minus half the solid angle subtended by the spherical triangle $\mathbf{e}_A\mathbf{e}_B\mathbf{e}_C$ on the Poincaré sphere, which is constructed by joining $\mathbf{e}_A$, $\mathbf{e}_B$, and $\mathbf{e}_C$ with arcs of great circles. A direct demonstration (Pancharatnam's) uses formulas from spherical trigonometry. Another way is to realize that a sequence of polarization states $|\psi(t)\rangle$ parallel-transported according to (34) along great circles on the Poincaré sphere are, according to Pancharatnam's criterion, in phase with each other. Thus $\arg\langle C|A\rangle$ is precisely the geometric phase accumulated under parallel transport around $\mathbf{e}_A\mathbf{e}_B\mathbf{e}_C$ (for which the half-solid-angle formula is obtained in Sec. 1.3).

To establish a connection between these geometrical considerations and the actual physics, some analysis of the polarization dynamics is required. Pancharatnam considered systems of *ideal polarizers,* or *analyzers,* which completely transmit a polarization state $|A\rangle$ and absorb the state $|A_\perp\rangle$ orthogonal to it in such a way that the incident and transmitted states are always in phase. Ideal polarizers are very nearly approximated by very thin polarizing plates. (Mathematically, they are represented by the Hermitian projection operator $|A\rangle\langle A|$.) Therefore, under a polarization cycle generated by ideal polarizers, the initial and final polarizations differ by Pancharatnam's phase.

Polarization cycles can also be generated by *ideal retarders;* these are perfectly transparent devices that introduce a relative phase difference between two given orthogonal polarization states $|A\rangle$ and $|A_\perp\rangle$. A particular example is a quarter-wave plate, for which the orthogonal states are linearly polarized and the phase difference is $\pi/2$. [Mathematically, an ideal retarder is represented by $\exp(-i\pi\mathbf{e}_A\cdot\boldsymbol{\sigma}/4)$, a two-dimensional unitary matrix with determinant 1.] Polarization cycles generated by ideal retarders produce a dynamical phase in addition to Pancharatnam's phase. The dynamical phase is easily calculated, however.

Polarization cycles can also arise through optical activity in a birefringent gyrotropic medium. Such a medium induces a relative phase between orthogonal polarization states whose orientation is determined by the local dielectric tensor. When the variation in the medium is small over a wavelength, the polarization of a monochromatic beam tracks the eigenstates of the dielectric tensor. Berry (1987b) applied an adiabatic analysis to Maxwell's equations and calculated the phase change over a polarization cycle generated by a spatially periodic medium. The analysis is formally the same as for a spin-$\frac{1}{2}$ particle in a slowly changing magnetic field (Sec. 1.3), and there appears a dynamical phase and a geometric phase; the geometric phase in this case is just Pancharatnam's.

Observations of Pancharatnam's phase are based on two-beam interferometry. Such experiments were initiated by Pancharatnam himself, though current versions are greatly improved with the availability of laser sources. Schematically, a plane wave with polarization $|A\rangle$ is split into two coherent beams, which propagate along the arms of an interferometer. Beam 1 propagates freely, while beam 2 passes through a sequence of optical elements that induce a polarization cycle. The beams are then recombined, and their total relative phase is determined from their interference. Differences in dynamical phases are determined and subtracted, and the geometric phase subsequently is measured.

Bhandari (1988) used laser interferometry methods to measure changes in the geometric phase, rather than the geometric phase itself. These are readily generated (while keeping the relative dynamical phases fixed) by continuously rotating a half-wave plate in the polarization cycle.

Kwiat and Chiao (1991) observed Pancharatnam's phase in single-photon interference experiments using a source of entangled two-photon states. One photon, the idler, was passed through a narrow-band energy filter to a detector. Observation of this photon fixed the energy of its partner, which was made to pass through an interferometer. Detection rates varied in accordance with the intensity, which depends directly on Pancharatnam's phase. Various coincidence detectors were used to confirm that genuine single-photon interference was being observed.

Berry and Klein (1996) performed a modern version of Pancharatnam's experiment, determining the relative phase by direct measurement of photographed interference fringes. By increasing the number of optical components in the polarization cycle, they were also able to investigate the relation between discrete and continuous geometric phases.

5. MOLECULAR AND CONDENSED-MATTER PHYSICS

Several applications of the geometric phase to molecular spectroscopy (Sec. 5.1), nuclear magnetic resonance (Sec. 5.2), and the quantum Hall effect (Sec. 5.3) are discussed. Among a number of applications that have been omitted, we mention Λ-doubling of rotational levels in diatomic molecules (Moody *et al.*, 1986), and atomic and reactive molecular scattering (Mead and Truhlar, 1979; Zygelman, 1987; Mead, 1992; Kuppermann and Wu, 1993).

5.1 Pseudorotation in Triatomic Molecules

Ordinarily, the angular degrees of freedom of a spinless system have integer quantum numbers. This is because, ordinarily, the angular momentum L_ϕ conjugate to an angle ϕ is taken to be $-i\hbar\partial_\phi$; its eigenfunctions $\exp(i\nu\phi)$ are single-valued only if ν is an integer. The Aharonov–Bohm effect (cf. Sec. 1.2) provides an exception. In the presence of a magnetic flux line on the z axis, with flux $\alpha hc/e$, the azimuthal angular momentum of an electron is given by $-i\hbar\partial_\phi + \alpha$, the constant term α coming from the magnetic vector potential $\mathbf{A}^{(s)}(\rho,\phi,z) = \alpha/\rho\hat{\phi}$. Therefore, the eigenvalues of L_ϕ are no longer integral multiples of $\hbar$ but are given instead by $(m + \alpha)\hbar$, where m is an integer.

As discussed in Sec. 1.7, there emerges in the Born–Oppenheimer approximation a molecular analog of the Aharonov–Bohm effect. For a given electronic state $|n\rangle$, the nuclear configuration space is threaded by a magnetic flux concentrated at points where $|n\rangle$ is degenerate. This flux appears in the nuclear Hamiltonian as a geometric vector potential (4). As discussed by Mead (1992), a triatomic molecule X_3 composed of identical atoms (e.g., alkali metal trimers such as Na_3) provides a simple class of examples. A striking consequence of the molecular Aharonov–Bohm effect is the appearance of anomalous half-integer quantum numbers. Their existence has been confirmed experimentally (Delacrétez *et al.*, 1986). First anticipated by Longuet-Higgens *et al.* (1958), this phenomenon is most naturally understood in terms of the geometric phase.

The nuclear configuration space for a triatomic molecule has nine degrees of freedom. Six of these correspond to overall translational and rotational motion, which can be ignored for the purposes of this discussion. That leaves three internal degrees of freedom, which describe vibrations. The internal configuration space, which we denote by Q, may be parameterized by the distances R_{12}, R_{23}, and R_{13} between the nuclei. It may be thought of as the space of noncongruent triangles (the R_{ij}'s determine the lengths of the triangle's sides) whose vertices are distinguishable.

In the Born–Oppenheimer approximation, the electronic energy levels and eigenstates are regarded as functions of the internuclear distances R_{ij}. Assuming that the molecular Hamiltonian has time-reversal symmetry, we expect degeneracies of the electronic states to lie along one-dimensional curves in Q (cf. Sec. 1.5).

It turns out that some of these degeneracies can be determined by symmetry considerations alone. For equilateral configurations $R_{12} = R_{23} = R_{13} = a$, the nuclear configuration is invariant under the dihedral group D_3 of symmetries of the equilateral triangle. Therefore, the electronic eigenstates $|n(a,a,a)\rangle$ must transform according to an irreducible representation, or irrep, of D_3 (Tinkham, 1964). The group D_3 has two one-dimensional irreps, associated with states that are either invariant or else change sign under the symmetry operations, and a two-dimensional irrep, associated with a degenerate pair of states that transform into linear superpositions of each other under the symmetry operations. Therefore, if the state $|p(a,a,a)\rangle$ transforms according to the two-dimensional irrep, then the family of states $|p(R_{12},R_{23},R_{13})\rangle$ to which it belongs must be degenerate along the curve $R_{12} = R_{23} = R_{13} = a$ (with $a > 0$) in Q. Under parallel transport around this curve [cf. (2)], $|p(R_{12}, R_{23},R_{13})\rangle$ changes sign.

One might expect the energy $E_p(R_{12}, R_{23},R_{13})$ of this state to assume its minimum in an equilateral configuration. The Jahn–Teller effect (Jahn and Teller, 1937) shows this is not the case. It turns out to be energetically favorable for the molecule to assume an asymmetrical configuration, so that for the p-electronic state, the ground vibrational state of the molecule is distorted.

To analyze the vibrational motion, it is useful to introduce cylindrical-like coordinates, (ρ,ϕ,z), in Q. The vertical coordinate z is given by $R_{12}^2 + R_{23}^2 + R_{13}^2$ and determines the length scale; oscillations in z describe uniform dilations and contractions of the nuclei. The radial coordinate ρ is given by $\rho^2/4 = z^2 - 3(R_{23}^2R_{13}^2 + R_{12}^2R_{13}^2 + R_{12}^2R_{23}^2)$ and describes the degree of distortion. For collinear configurations, $\rho = z$; for equilateral configurations, $\rho = 0$. Thus, it is along the z axis (where $\rho = 0$) that the electronic state $|p(\rho,\phi,z)\rangle$ is degenerate. The angle ϕ is given by $\tan^{-1}[\sqrt{3}(R_{23}^2 - R_{13}^2)/(2R_{12}^2 - R_{23}^2 - R_{13}^2)]$ and describes the direction of the distortion. Displacements in ϕ are called *pseudorotations.* A 2π pseudorotation returns the nuclei to their original positions, while a $2\pi/3$ rotation induces a cyclic permutation among them.

A systematic calculation of the vibrational Hamiltonian H_{vib}, described in (Mead, 1992), requires an explicit expression for the electronic energy $E_p(\rho,\phi,z)$ along with a careful derivation of the kinetic energy in terms of the (ρ,ϕ,z) coordinates. However, for this discussion, an approximate model treatment will suffice. The potential is taken to be harmonic in z and ρ and independent of ϕ, and the kinetic energy to be of the standard form. Thus

$$H_{\text{vib}} = \frac{p_\rho^2}{2M} + \frac{L_\phi^2}{2M\rho^2} + \frac{p_z^2}{2M} + \tfrac{1}{2}M\omega_\rho^2(\rho - \rho_0)^2 + \tfrac{1}{2}M\omega_z^2(z - z_0)^2, \tag{58}$$

where (ρ_0,z_0) denotes the asymmetrical configuration that minimizes E_p. The vibrational energy levels are characterized by three quantum numbers j, m, and k, for the respective degrees of freedom ρ, ϕ, and z, and are given approximately by

$$E(j,m,k) = (j + \tfrac{1}{2})\hbar\omega_r + \frac{(m\hbar)^2}{2I} + (k + \tfrac{1}{2})\hbar\omega_z, \tag{59}$$

where the moment of inertia I about the z axis is taken to be $M\rho_0^2$. (The approximation consists of ignoring the dependence of I on the radial quantum number j).

The presence of the electronic degeneracy along the z axis introduces a geometric vector potential $\mathbf{A}_p = \alpha\hat{\boldsymbol{\phi}}/\rho$ into the nuclear Hamiltonian, which accounts for the sign change of the electronic eigenstate under parallel transport. (In the chosen coordinate system, $\alpha = \frac{3}{2}$.) Its effect is to replace the pseudoangular momentum operator $L_\phi = -i\hbar\partial_\phi$ by $-i\hbar\partial_\phi + \frac{3}{2}\hbar$, in analogy to what occurs in the magnetic Aharonov–Bohm Hamiltonian. Thus, the pseudorotational quantum number m in (59) is half-integral.

The half-integral quantization of pseudorotation was confirmed in a systematic spectral analysis of the vibrational spectrum of Na_3 by Delacrétez *et al.* (1986). A sequence of rovibrational levels corresponding to radial and pseudorotational motion were identified. It was found that the observed spectrum could be fitted by (59) only by assigning half-integer quantum numbers to the pseudoangular momentum. Such an assignment yields excellent agreement between observed and calculated energy levels.

5.2 Nuclear Magnetic Resonance

An rf magnetic field applied to polarized spins generates transitions between their stationary states. When the field is in resonance with the precession frequency of the spins, the transition probabilities can become large, even though the rf field may be weak. Transitions are manifested in oscillations of the spin magnetic moment, whose amplitude and frequency can be measured with great accuracy. This is the basis for nuclear spin-resonance experiments. In nuclear magnetic resonance (NMR), the nuclear spins are polarized by a static magnetic field. In nuclear quadrupole resonance (NQR), the spins are polarized by an electric field gradient (in the ambient crystalline potential, for example) that couples to the spin through the nuclear quadrupole moment.

Moody *et al.* (1986) and Cina (1986) pointed out that cyclic variations in the rf field would generate geometric phases in the nuclear spins, which could be detected as shifts in the frequency spectrum of the nuclear magnetic moment (this is an example of the third type of experiment described in

Sec. 1.6). Pursuing this suggestion, several groups have carried out NMR studies of the geometric phase, including its adiabatic (Tycko, 1987; Suter *et al.*, 1987), nonadiabatic (Suter *et al.*, 1988), and non-Abelian (Zwanziger *et al.*, 1990b) versions. These experiments not only have confirmed theoretical predictions but also have brought to light interesting new phenomena in NMR spectroscopy.

In an NMR experiment, the nuclear spins (let $s = \frac{1}{2}$, say) are initially polarized by a strong static magnetic field $B_z\hat{\mathbf{z}}$. The Hamiltonian H_0 is given by $\frac{1}{2}\hbar\gamma B_z\sigma_z$, where γ is the gyromagnetic ratio, and the eigenstates $|\pm\rangle$ have energies $\pm\hbar\omega_0/2$, where $\omega_0 = \gamma B_z$ is the Larmor frequency. For nonstationary states, the magnetic moment $\mathbf{M}(t) = \hbar\gamma\langle\psi(t)|\boldsymbol{\sigma}|\psi(t)\rangle/2$ precesses about $\hat{\mathbf{z}}$ at the Larmor frequency, so that its frequency spectrum has a single peak at ω_0.

A weak, circularly polarized rf field, $\mathbf{B}_{\text{rf}}(t) = B_x \cos\omega t\hat{\mathbf{x}} + B_y \sin\omega t\hat{\mathbf{y}}$, is then applied. (Weak means that $B_{\text{rf}} \ll B_z$.) The subsequent dynamics is most easily analyzed in a rotating frame turning about $\hat{\mathbf{z}}$ with angular frequency ω. There the effective magnetic field $\mathbf{B}_{\text{rot}}$ is static, and the Hamiltonian is given by

$$H_{\text{rot}} = \tfrac{1}{2}\gamma\hbar\mathbf{B}_{\text{rot}}\cdot\sigma,$$
$$\mathbf{B}_{\text{rot}} = B_x\hat{\mathbf{x}} + B_y\hat{\mathbf{y}} + (B_z - \omega/\gamma)\hat{\mathbf{z}}. \tag{60}$$

The spin eigenstates $|\pm\rangle_{\text{rot}}$ in the rotating frame are polarized along $\mathbf{B}_{\text{rot}}$ with energies $\pm\hbar\omega_{\text{rot}}/2$, and the magnetic moment $\mathbf{M}_{\text{rot}}(t)$ precesses about $\mathbf{B}_{\text{rot}}$ with frequency ω_{rot}, where $\omega_{\text{rot}} = \gamma B_{\text{rot}}$. Near resonance, where $\omega \approx \omega_0$, the axis of precession is nearly perpendicular to $\hat{\mathbf{z}}$ (even though $B_{\text{rf}} \ll B_z$).

The magnetic moment $\mathbf{M}(t)$ in the laboratory frame is given by $\mathbf{M}_{\text{rot}}(t)$ rotated back through an angle $-\omega t$ about $\hat{\mathbf{z}}$. Thus, the effect of the rf field is to superpose a slow longitudinal nutation (of frequency ω_{rot}) on the fast Larmor precession of $\mathbf{M}(t)$ about $\hat{\mathbf{z}}$ (of frequency ω_0). In consequence, the frequency spectrum of the magnetic moment has additional peaks at $\omega_0 \pm \omega_{\text{rot}}$.

The Hamiltonian H_{rot} of (60) depends parametrically on the components of the rf and static magnetic fields, B_x, B_y, and B_z, and vanishes for $B_x = B_y = 0$ and $B_z = \omega/\gamma$. If B_x, B_y, and B_z are slowly modulated through a cycle C, with a period T much greater than the precession period $2\pi/\omega_{\text{rot}}$ in the rotating frame, then the eigenstates $|\pm\rangle_{\text{rot}}$ of H_{rot} evolve adiabatically and acquire with each modulation cycle a geometric phase $\gamma_\pm(C)$ given by half the solid angle α described by C with respect to $(0,0,\omega/\gamma)$ (cf. Sec. 1.3). These geometric phases produce shifts $\mp\frac{1}{2}\hbar\alpha/T$ in the energies $\pm\frac{1}{2}\hbar\omega_{\text{rot}}$ of the unmodulated Hamiltonian H_{rot}, and in turn shift the precession frequency of the magnetic moment from ω_{rot} to $\omega_{\text{rot}} - \alpha/T$. (This shift is the magnetic analog of the daily precession of the Foucault pendulum discussed in Sec. 3.1.3.)

These frequency shifts were observed by Suter *et al.* (1987) in the NMR response of proton spins in a water/acetone sample. The adiabatic cycle C was produced by placing the magnetization detector in a frame rotating with angular frequency ω_{det} about $\hat{\mathbf{z}}$. In the detector frame, the effective magnetic field turns about $\hat{\mathbf{z}}$ with angular frequency $\delta = \omega - \omega_{\text{det}}$; for small enough δ, the spins adiabatically track the turning field. A variation of this experiment was subsequently performed by Suter *et al.* (1988), in which the geometric phase was observed as a rotation in the spin-echoed magnetic moment. This technique enabled observation of the individual geometric phases $\gamma_\pm$ rather than their difference $\gamma_+ - \gamma_-$. The spins were also taken through nonadiabatic cycles, thereby incorporating the Aharonov–Anandan phase (cf. Sec. 2.3).

In a nuclear quadrupole resonance experiment, the spin polarization is induced by an inhomogeneous electric field rather than a static magnetic field. Gradients in the electric field couple to spin through the nuclear electric quadrupole moment. The effective spin Hamiltonian is of the form $H_Q = \hbar\omega_Q(\mathbf{S}\cdot\hat{\mathbf{n}})^2$, where $\mathbf{S}$ denotes the dimensionless spin operators and $\hat{\mathbf{n}}$ is the symmetry axis of the electric-field-gradient tensor (assumed here to be axially symmetric). The eigenstates $|m(\hat{\mathbf{n}})\rangle$ are polarized along $\hat{\mathbf{n}}$ with energies $\hbar\omega_Q m^2$. Note that these are doubly degenerate for $m \neq 0$.

Geometric phase effects in the NQR spectrum of ^{35}Cl (for which $s = \frac{3}{2}$) were studied by Tycko (1987). In the experiment, a coherent superposition of eigenstates of H_Q was created at $t = 0$ by an rf pulse; the induced magnetic moment $\mathbf{M}(t)$ was then observed and spectrally analyzed. The spectrum of H_Q

consists of two degenerate doublets with energies $\frac{9}{4}\hbar\omega_Q$ and $\frac{1}{4}\hbar\omega_Q$ (corresponding to $m = \pm\frac{3}{2}$ and $m = \pm\frac{1}{2}$, respectively), so that the magnetic moment spectrum has a single peak at frequency $2\omega_Q$. Adiabatic cycling of H_Q was achieved by rotating a crystalline sample of $NaClO_3$ with angular frequency $\omega_R \ll \omega_Q$ about an axis inclined to the symmetry direction $\hat{\mathbf{n}}$. With each rotation, the spin eigenstates acquire geometric phases β_m (these may be computed explicitly), which shift their energies by $\hbar\beta_m\omega_R/2\pi$ and lift the double degeneracies. For the particular geometry realized in Tycko's experiment, theory predicts that the single peak in the magnetic moment spectrum is split into three peaks at frequencies $2\omega_Q$ and $2\omega_Q \pm \sqrt{3}\omega_R/\pi$. Precisely this splitting was observed in the experiment.

Since H_Q has degenerate energy levels, one might expect non-Abelian geometric phases (cf. Sec. 2.2) to appear. Tycko (1987) and Zee (1988) showed that they do not; the experiment can be analyzed completely in terms of the ordinary (Abelian) geometric phase. One can compute the non-Abelian vector potentials $\mathsf{A}_{\pm 3/2}$ and $\mathsf{A}_{\pm 1/2}$ of (25) explicitly (there is one for each degenerate doublet of H_Q); they are matrix-valued and are proportional to the matrix elements of the spin $\mathbf{S}$ taken between states in the degenerate doublets. For $m = \pm\frac{3}{2}$, the off-diagonal matrix elements vanish identically ($\mathbf{S}$ has nonzero matrix elements only between states $|m\rangle$ and $|m \pm 1\rangle$), and the diagonal matrix elements yield the ordinary geometric phases $\beta_{\pm 3/2}$. For $m = \pm\frac{1}{2}$, the off-diagonal elements $\langle \pm\frac{1}{2}|\mathbf{S}| \mp \frac{1}{2}\rangle$ do not vanish. However, for the adiabatic cycles used in the experiment, in which $\hat{\mathbf{n}}$ is turned about a fixed axis, a basis for the $m = \pm\frac{1}{2}$ doublet can be found that diagonalizes $\mathsf{A}_{\pm 1/2}$. During the cycle there are no transitions between these basis vectors and each acquires an ordinary geometric phase $\beta_{\pm 1/2}$.

Zee (1988) showed that more general adiabatic cycles of $\hat{\mathbf{n}}$ generate genuine non-Abelian geometric phases in the NQR spectrum. Zwanziger *et al.* (1990b) performed an experiment to observe these non-Abelian effects. The more complicated adiabatic cycles were generated by mounting the $NaClO_3$ crystal on a double rotor. Careful analysis predicts that the NQR resonance should be split into five peaks instead of three. This is precisely what was observed, with excellent quantitative agreement between theory and experiment.

5.3 The Quantum Hall Effect

Charged particles in crossed uniform electric and magnetic fields experience a net drift along the direction $\mathbf{E} \times \mathbf{B}$ normal to both—this is the basis for the *Hall effect*. The Hall effect concerns specifically the behavior of electrons in a planar conductor to which a perpendicular magnetic field $\mathbf{B} = B\hat{\mathbf{z}}$ and a tangential electric field $\mathbf{E} = E\hat{\mathbf{x}}$ are applied. The resulting current density has a transverse component $j_y = \sigma_H E$. σ_H is the *Hall conductance*. Using classical theory, one obtains for large magnetic fields that

$$\sigma_{\mathrm{H}} \approx nec/B, \tag{61}$$

where n is the two-dimensional electron density.

[Alternatively, one can consider conductors of finite transverse extent, e.g., long thin wires. Then the current density is necessarily longitudinal, whereas the electric field has a transverse component. In this case it is easier to compute the *Hall resistance* $\rho_{\mathrm{H}} = E_y/j$. Classical theory gives $\rho_{\mathrm{H}} = B/nec$ for all magnetic fields, not just large ones. The Hall conductance and resistance are the off-diagonal elements of the conductance and resistivity tensors σ and $\rho = \sigma^{-1}$, and the relation between them, $\sigma_{\mathrm{H}} = \rho_{\mathrm{H}}/(\rho_{\mathrm{O}}^2 + \rho_{\mathrm{H}}^2)$, depends on the longitudinal or Ohmic resistivity ρ_{O}.]

For large magnetic fields and at low temperatures, the classical prediction (61) fails; what is actually observed is a spectacular phenomenon, the quantum Hall effect. As B is increased, instead of falling off monotonically as $1/B$, the Hall conductance goes through a series of plateaus, where its value is constant, punctuated by intervals of monotonic decrease. The constant values assumed on the plateaus are sample independent and satisfy with extremely high accuracy (about one part in ten million) the quantization condition

$$\sigma_{\mathrm{H}} = re^2/hc. \tag{62}$$

As B increases, the coefficient r takes on integer values and decreases with B in unit steps until it reaches 1. This is the regime of the integer quantum Hall effect (von Klitzing *et al.*, 1980). As B is increased still further, one enters the regime of the fractional quantum Hall effect (Tsui *et al.*, 1982); r takes on rational values (less than 1) with odd denominators.

The theory of the quantum Hall effect begins with the quantum mechanics of two-dimensional electrons in a magnetic field. The one-electron stationary states are the Landau levels $E_N = \hbar\omega_c(N + \frac{1}{2})$, $N = 0, 1, 2, \ldots$, where $\omega_c = eB/mc$ is the classical cyclotron frequency. The Landau levels are highly degenerate; each can accommodate B/Φ_0 electrons per unit area ($\Phi_0 = hc/e$ is the quantum of magnetic flux). In real materials, the crystalline potential, impurities, and electron–electron interactions all serve to break this degeneracy. However, for sufficiently large magnetic fields, there is negligible mixing between the Landau levels, so that each Landau level is split by the perturbations into B/Φ_0 sublevels per unit area.

If the electron density in the sample is n, then the ground state has the first $\nu = n\Phi_0/B$ Landau levels filled. (Thus, the larger the field B, the fewer occupied Landau levels.) If ν has a fractional part, the uppermost level is only partially filled. The integer quantum Hall effect can now be explained by making two assumptions:

1. only a small proportion of sublevels in each Landau level can carry current, and
2. each filled Landau level contributes precisely e^2/hc to the Hall conductance.

Together, these assumptions imply that a nearly empty Landau level contributes nothing to the Hall conductance, as the occupied sublevels are unlikely to be the current-carrying ones, whereas, conversely, a nearly full Landau level contributes e^2/hc. From these qualitative considerations follows $\sigma_H = [\nu]e^2/hc$ ($[\nu]$ denotes the integer nearest ν). This is precisely the relation (62), with $r = [\nu]$.

The first assumption can be justified on the basis of localization theory—most of the sublevels are localized, whereas only extended states can carry current. For the second assumption, Thouless, Kohmoto, Nightingale, and den Nijs (TKN^2) found, in 1982, a topological explanation (Thouless *et al.*, 1982). In retrospect, their analysis can be seen as an application of the theory of the geometric phase. Indeed, this connection was drawn as early as 1983 by Simon (1983).

In outline, the analysis of TKN^2 is as follows. We consider an electron in two dimensions in a uniform magnetic field $B\hat{\mathbf{z}}$ and periodic crystalline potential $V(x,y)$, with periods a and b in x and y. Because the vector potential $\mathbf{A} = (-By/2, Bx/2)$ is not periodic, neither is the Hamiltonian $mv^2/2 + V(x,y)$, in which $\mathbf{v} = (\mathbf{p} + e\mathbf{A}/c)/m$ is the velocity. However, the Hamiltonian *is* invariant under the magnetic translation operators $T_x = \exp(-imv_x a/\hbar)$ and $T_y = \exp(-imv_y b/\hbar)$, which shift x and y by a and b while leaving the velocity (rather than the momentum) unchanged.

While the magnetic translation operators commute with the Hamiltonian, they do not necessarily commute with each other. In general, $T_xT_y = \exp(2\pi iBab/\Phi_0)T_yT_x$. However, if we take the flux Bab through a unit cell to be a rational multiple p/q of the magnetic flux quantum Φ_0, then $(T_x)^q$—a translation by q lattice spacings in x—and T_y do commute, and Bloch's theorem applies. In this case, the energy levels form a band structure, whose bands $\epsilon_n(\mathbf{k})$ are periodic over the first Brillouin zone $[0,2\pi/qa] \times [0,2\pi/b]$. The eigenfunctions $\psi_{n,\mathbf{k}}(\mathbf{r})$ are Bloch waves $\exp(i\mathbf{k}\cdot\mathbf{r})u_{n,\mathbf{k}}(\mathbf{r})$, where $u_{n,\mathbf{k}}(\mathbf{r})$ is an eigenfunction of the Hamiltonian

$$H(\mathbf{K}) = (1/2m)(-i\hbar\nabla + e\mathbf{A}/c + \hbar\mathbf{k})^2 + V(x,y), \tag{63}$$

with energy $\epsilon_n(\mathbf{k})$, which is invariant under the magnetic translations $(T_x)^q$ and T_y. For convenience, we sometimes denote the indices $n,\mathbf{k}$ by a single label α.

The Hall conductance can now be calculated perturbatively. An applied electric field $E\hat{\mathbf{x}}$ generates perturbed eigenstates $|u'_\alpha\rangle = |u_\alpha\rangle + |\delta u_\alpha\rangle$, which to lowest order are given by $\langle u_\beta|\delta u_\alpha\rangle = \langle u_\beta| - eEx|u_\alpha\rangle/(\epsilon_\alpha - \epsilon_\beta)$, $\beta \neq \alpha$. The electric field induces a transverse current density $\langle u'_\alpha|\, j_y|u'_\alpha\rangle$, where $\mathbf{j} = e(\mathbf{v} + \hbar\mathbf{k}/m)/A_0$ and A_0 is the area of the sample. From (63), the current may also be written as $\mathbf{j} = e\nabla_{\mathbf{k}}H/A_0$. The Hall conductance σ_H is given by the ratio of the total induced current density $\Sigma_\alpha\langle u'_\alpha|\, j_y|u'_\alpha\rangle$ to the electric-field strength E, where the sum is taken over the occupied levels; e.g., for the ground state, the levels below the Fermi energy ϵ_F. Making use of the identity $[x,H] = i\nabla_{k_x}H$, we obtain the following explicit expression:

$$\sigma_{\mathrm{H}} = \frac{e^2}{A_0\hbar c} \sum_{\epsilon(\alpha)<\epsilon_F} \sum_{\beta\neq\alpha} \mathrm{Im} \times \frac{\langle u_\alpha|\nabla_{\mathbf{k}}H|u_\beta\rangle \times \langle u_\beta|\nabla_{\mathbf{k}}H|u_\alpha\rangle}{(\epsilon_\alpha - \epsilon_\beta)^2}. \tag{64}$$

Equation (64) is the Kubo formula for the Hall conductance of the ground state.

Suppose we regard $H(\mathbf{k})$ as a family of Hamiltonians parametrized by the Bloch vector $\mathbf{k}$. Under parallel transport around a closed cycle C, the eigenstate $|u_{n,\mathbf{k}}\rangle$ acquires a geometric phase $\gamma_{n,\mathbf{k}}(C)$, which may be expressed (cf. Sec. 1.2) as the integral of the flux field $V_n(\mathbf{k}) = \mathrm{Im}\langle\nabla_{\mathbf{k}}u_{n,k}| \times |\nabla_{\mathbf{k}}u_{n,k}\rangle$ over the interior of C. Among the various expressions for $V_n(\mathbf{k})$ given in Sec. 1.2, we note that (10), which involves matrix elements of the gradient of the Hamiltonian with respect to the parameters $\mathbf{k}$, is contained precisely in the formula (64) for the Hall conductance. This allows us to rewrite (64) as $\sigma_{\mathrm{H}} = (e^2/A_0\hbar c)\Sigma_n\Sigma_k V_n(\mathbf{k})$. For samples large compared with the unit cell, the sum $\Sigma_{\mathbf{k}}$ over Bloch vectors may be replaced by an integral $(A_0/4\pi^2)\int d^2k$ over the first Brillouin zone T_k. Thus we obtain

$$\sigma_{\mathrm{H}} = \frac{e^2}{hc}\sum_n \sigma_n, \tag{65}$$

where

$$\sigma_n = \frac{1}{2\pi}\int_{T_k} V_n(\mathbf{k})d^2k \tag{66}$$

is the Hall conductance (in units of e^2/hc) of the nth band.

The central result of TKN^2 is that σ_n is an integer. The argument is as follows. Since the energy levels $\epsilon_n(\mathbf{k})$ are periodic over the Brillouin zone, the Bloch states $|\psi_{n,\mathbf{k}}\rangle$ are periodic up to an overall phase factor. In particular, the Bloch states on the corners of the first Brillouin zone, i.e., $\mathbf{k} = 0$, $\mathbf{A}$, $\mathbf{A}+\mathbf{B}$, and $\mathbf{B}$ [here $\mathbf{A} = (2\pi/qa,0)$ and $\mathbf{B} = (0,2\pi/b)$ denote the reciprocal lattice vectors], differ only by a phase. Thus $|\psi_{n,\mathbf{A}}\rangle = e^{i\alpha}|\psi_{n,0}\rangle$, $|\psi_{n,\mathbf{A}+\mathbf{B}}\rangle = e^{i\beta}|\psi_{n,\mathbf{A}}\rangle$, $|\psi_{n,\mathbf{B}}\rangle = e^{i\gamma}|\psi_{n,\mathbf{A}+\mathbf{B}}\rangle$, and $|\psi_{n,0}\rangle = e^{i\delta}|\psi_{n,\mathbf{B}}\rangle$. Together these relations imply that $|\psi_{n,0}\rangle = e^{i(\alpha+\beta+\gamma+\delta)}|\psi_{n,0}\rangle$, and it follows that $\alpha + \beta + \gamma + \delta$ is necessarily an integer multiple m of 2π. Applying Stokes's theorem to (66), we obtain $\sigma_n = (1/2\pi)\oint\mathrm{Im}\langle u_{n,\mathbf{k}}|\nabla_{\mathbf{k}}u_{n,\mathbf{k}}\rangle\cdot d\mathbf{k}$, where the integral is taken around the boundary of the first Brillouin zone. A little calculation shows that this integral gives precisely the accumulated phase change $(\alpha + \beta + \gamma + \delta)/2\pi$ around the first Brillouin zone T_k, i.e., the integer m. In the language of fiber bundles (cf. Sec. 2.1.4), the Bloch states $|\psi_{n,\mathbf{k}}\rangle$ constitute a complex line bundle over T_k with curvature $V_n(\mathbf{k})$, and σ_n is the first Chern number of this line bundle.

For large magnetic fields, the bands $\epsilon_n(\mathbf{k})$ coalesce into clusters of p nearly degenerate sublevels, each cluster associated with one of the Landau levels of the pure magnetic Hamiltonian. It can be shown (the argument is omitted here) that the sum of the conductances σ_n over states in a given Landau level is equal to 1. This completes the topological explanation for the quantization of Hall conductance. TKN^2 also calculate the individual σ_n in the limit of a weak periodic potential; these are found to satisfy certain Diaphantine equations in p and q.

TKN^2 was the first of a number of important applications of geometric-phase methods to the study of the quantum Hall effect. Several subsequent works address the fractional effect, where electron–electron interactions come into play. Niu *et al.* (1985) show how the results of TKN^2 can be generalized to accommodate them. Arovas *et al.* (1984) calculate properties of the elementary excitations of Laughlin's many-body fractional-quantum-Hall ground state (Laughlin, 1983) using geometric-phase arguments. Concerning the integer case, Arovas *et al.* (1988) establish a topological criterion for localization in terms of Chern numbers. Wilkinson (1984) incorporates the geometric phase into semiclassical calculations of the single-particle energy spectrum.

Finally, we mention studies by Avron and co-workers (Avron *et al.*, 1988; Avron, 1995) of electron transport in quantum systems. While not specifically concerned with the quantum Hall effect, they elucidate essentially similar topological mechanisms. One class of models consists of planar networks of one-dimensional wires whose closed loops C_i are threaded by magnetic fluxes ϕ_i normal to the plane. The physical properties of the network are Φ_0 periodic in the fluxes (fluxes differing by an integral number of flux quanta are related by a gauge transformation). Fixing all but two of the fluxes, the stationary states $|\psi_n(\phi_i,\phi_j)\rangle$ constitute a complex line bundle over the two-torus of fluxes (the analog of the first Brillouin zone). The associated Chern number, the integral of the curvature, has the following physical interpretation: it represents the average with respect to ϕ_i of the number of electrons transported around C_i as ϕ_j is increased adiabatically by one flux quantum.

GLOSSARY

Dynamical Phase: The time integral of the energy of an evolving quantum state divided by Planck's constant $\hbar$.

Geometric Phase: A phase factor accumulated by a cycled quantum eigenstate that depends only on the geometry of its evolution.

Geometric Magnetism: Effective Lorentz-type force that appears in the dynamics of a slow system coupled to a fast one after the fast degrees of freedom have been averaged away.

Hannay Angle: Displacement in the angle variables of adiabatically cycled integrable systems of purely geometric origin.

Holonomy: Change in phase induced by parallel transport of a quantum state around a closed curve.

Integrable System: A Hamiltonian system with as many constants of the motion in involution as there are degrees of freedom.

Parallel Transport: A prescription to determine the overall phase of an evolving quantum state.

Reduction: General theory of Hamiltonian dynamics in the presence of symmetry.

Works Cited

Aharonov, Y., Anandan, J. (1987), *Phys. Rev. Lett.* **58,** 145–148.

Aharonov, Y., Bohm, D. (1959), *Phys. Rev.* **115,** 484–491.

Aitchison, I. J. R. (1987), *Acta Phys. Polonica* **B18,** 207–235.

Alber, M. S., Marsden, J. E. (1992), *Comm. Math. Phys.* **149,** 217–240.

Arovas, D. P., Schrieffer, J. R., Wilczek, F. (1984), *Phys. Rev. Lett.* **53,** 722–723.

Arovas, D. P., Bhatt, R. N., Haldane, F. D. M., Littlewood, P. B., Rammal, R. (1988), *Phys. Rev. Lett.* **60,** 619–622.

Avron, J. E. (1995), in: E. Akkermans *et al.* (Eds.), *Mesoscopic Quantum Physics, Proceedings of the Les Houches Summer School, Session LXI,* Amsterdam: Elsevier Science B.V., pp. 741–791.

Avron, J. E., Raveh, A., Zur, B. (1988), *Rev. Mod. Phys.* **60,** 873–915.

Berry, M. V. (1984), *Proc. R. Soc. A* **392,** 45–57.

Berry, M. V. (1985), *J. Phys. A: Math. Gen.* **18,** 15–27.

Berry, M. V. (1987a), *Nature (London)* **326,** 277–278.

Berry, M. V. (1987b), *J. Mod. Opt.* **34,** 1401–1407.

Berry, M. V. (1987c), *Proc. R. Soc. A* **414,** 31–46.

Berry, M. V. (1989), in: A. Shapere, F. Wilczek (Eds.), *Geometric Phases in Physics,* Singapore: World Scientific, pp. 7–28.

Berry, M. V. (1990a), *Phys. Today* **43** (12), 34–40.

Berry, M. V. (1990b), *Proc. R. Soc. A* **430,** 405–411.

Berry, M. V., Robbins, J. M. (1993a), *Proc. R. Soc. A* **442,** 641–658.

Berry, M. V., Robbins, J. M. (1993b), *Proc. R. Soc. A* **442,** 659–672.

Berry, M. V., Klein, S. (1996), *J. Mod. Opt.* **43,** 165–180.

Berry, M. V., Morgan, M. A. (1996), *Nonlinearity* **9,** 787–799.

Bhandari, R. (1988), *Phys. Lett. A* **133,** 1–3.

Bhandari, R. (1997), *Phys. Rep.* **281,** 2–64.

Bhandari, R., Samuel, J. (1988), *Phys. Rev. Lett.* **60,** 1211–1214.

Bitter, T., Dubbers, D. (1987), *Phys. Rev. Lett.* **59,** 251–254.

Chambers, R. G. (1960), *Phys. Rev. Lett.* **5,** 3–5.

Cina, J. A. (1986), *Chem. Phys. Lett.* **132,** 393–395.

Delacrétez, G., Grant, E. R., Whetten, R. L., Wöste, L., Zwanziger, J. W. (1986), *Phys. Rev. Lett.* **59,** 161–164.

Dingle, R. B. (1973), *Asymptotic Expansions: Their Derivation and Interpretation,* New York and London: Academic.

Dirac, P. A. M. (1931), *Proc. R. Soc. A* **133,** 60.

Dykhne, A. M. (1962), *Zh. Eksp. Teor. Fiz.* **41,** 1324–1327 [translation: *Sov. Phys. JETP* **14,** 941–943].

Eco, U. (1989), *Foucault's Pendulum,* New York: Harcourt Brace Jovanovich.

Eguchi, T., Gilkey, P. B., Hanson, A. J. (1980), *Phys. Rep.* **66,** 213–393.

Golin, S., Knauf, A., Marmi, S. (1989), *Comm. Math. Phys.* **110,** 33–49.

Hannay, J. H. (1985), *J. Phys. A: Math. Gen.* **18,** 221–230.

Jackiw, R. (1988), *Int. J. Mod. Phys.* **A3,** 285–297.

Jahn, H. A., Teller, E. (1937), *Proc. R. Soc. A* **161,** 220–235.

Jarzynski, C. (1993), *Phys. Rev. Lett.* **71,** 839–842.

Jarzynski, C. (1995), *Phys. Rev. Lett.* **74,** 1732–1735.

Kitano, M., Yabuzuaki, T., Ogawa, T. (1987), *Phys. Rev. Lett.* **58,** 523.

Kuppermann, A., Wu, Y. S. M. (1993), *Chem. Phys. Lett.* **205,** 577–586.

Kuratsuji, H., Iida, S. (1985), *Prog. Theor. Phys.* **74,** 439–445.

Kwiat, P. G., Chiao, R. Y. (1991), *Phys. Rev. Lett.* **66,** 88–91.

Littlejohn, R. G. (1988), *Phys. Rev. A* **38,** 6034–6045.

Littlejohn, R. G., Flynn, W. G. (1992), *Phys. Rev. A* **44,** 5239–5256.

Laughlin, R. B. (1983), *Phys. Rev. Lett.* **50,** 1395–1398.

Li Yin, Mead, C. A. (1992), *Theor. Chim. Acta* **82,** 397–406.

Lochak, P., Meunier, C. (1988), *Multiphase Averaging for Classical Systems,* New York: Spinger.

Longuet-Higgins, H. C., Öpik, U., Pryce, M. H. L., Sack, R. A. (1958), *Proc. R. Soc.* **244,** 1–12.

Marsden, J. E. (1992), *Lectures on Mechanics,* Cambridge, U.K.: Cambridge Univ. Press.

Marsden, J. E., O'Reilly, O. M., Wicklin, F. J., Zombro, B. W. (1991), *Nonlin. Sci. Today* **1,** 4–11, 14–21.

Mead, C. A. (1980), *Chem. Phys.* **49,** 23–40.

Mead, C. A. (1992), *Rev. Mod. Phys.* **64,** 51–85.

Mead, C. A., Truhlar, D. G. (1979), *J. Chem. Phys.* **70,** 2284–2296.

Montgomery, R. (1991), *Am. J. Phys.* **59,** 394–398.

Moody, J., Shapere, A., Wilczek, F. (1986), *Phys. Rev. Lett.* **56,** 893–896.

Nash, C., Sen, S. (1983), *Topology and Geometry for Physicists,* New York: Academic.

Niu, Q., Thouless, D. J., Wu, Y. S. (1985), *Phys. Rev. B* **31,** 3372–3377.

Pancharatnam, S. (1956), *Proc. Ind. Acad. Sci. A* **44,** 247.

Provost, J., Vallee, G. (1980), *Comm. Math. Phys.* **76,** 289–301.

Robbins, J. M. (1994), *J. Phys. A: Math. Gen.* **27,** 1179–1189.

Robbins, J. M., Berry, M. V. (1992), *Proc. R. Soc. A* **436,** 631–661.

Rytov, S. M. (1938), *Dokl. Akad. Nauk. USSR* **18,** 263; reprinted in B. Markovsky, V. I. Vinitsky (Eds.) (1989), *Topological Phases in Quantum Theory,* Singapore: World Scientific, p. 6.

Samuel, J., Bhandari, R. (1988), *Phys. Rev. Lett.* **60,** 2339–2342.

Schroer, C. G. (1996), *J. Phys. A: Math. Gen.* **29,** 3289–3297.

Shapere, A., Wilczek, F. (Eds.) (1989a), *Geometric Phases in Physics,* Singapore: World Scientific.

Shapere, A., Wilczek, F. (1989b), *Am. J. Phys.* **57,** 514–518.

Shapere, A., Wilczek, F. (1989c), *J. Fluid. Mech.* **198,** 557–585.

Shapere, A., Wilczek, F. (1989d), *J. Fluid. Mech.* **198,** 587–599.

Simon, B. (1983), *Phys. Rev. Lett.* **51,** 2167–2170.

Suter, D., Chingas, G. C., Harris, R. A., Pines, A. (1987), *Mol. Phys.* **61,** 1327–1340.

Suter, D., Mueller, K. T., Pines, A. (1988), *Phys. Rev. Lett.* **60,** 221–223.

Thouless, D. J., Kohmoto, M., Nightingale, M. P., den Nijs, M. (1982), *Phys. Rev. Lett.* **49,** 405–408.

Tinkham, M. (1964), *Group Theory and Quantum Mechanics,* New York: McGraw-Hill.

Tsui, D. C., Stormer, H. L., Gossard, A. C. (1982), *Phys. Rev. Lett.* **48,** 1559.

Tomita, A., Chiao, R. (1986), *Phys. Rev. Lett.* **57,** 937–940.

Tycko, R. (1987), *Phys. Rev. Lett.* **58,** 2281–2284.

Vladimirsky, V. V. (1941), *Dokl. Akad. Nauk. USSR* **21,** 222; reprinted in B. Markovsky, V. I. Vinitsky (Eds.) (1989), *Topological Phases in Quantum Theory,* Singapore: World Scientific, p. 11.

von Klitzing, K., Dorda, G., Pepper, M. (1980), *Phys. Rev. Lett.* **45,** 494.

von Neumann, J., Wigner, E. P. (1929), *Z. Phys.* **30,** 467–470.

Weigert, S., Littlejohn, R. G. (1993), *Phys. Rev. A* **47,** 3506–3512.

Wilczek, F., Zee, A. (1984), *Phys. Rev. Lett.* **52,** 141–144.

Wilkinson, M. (1984), *J. Phys. A: Math. Gen.* **17,** 3459–3476.

Wilkinson, M. (1990), *J. Phys. A: Math. Gen.* **23,** 3603–3611.

Zee, A. (1988), *Phys. Rev. A* **38,** 1–6.

Zener, C. (1932), *Proc. R. Soc. A* **137,** 696–702.

Zwanziger, J. W., Koenig, M., Pines, A. (1990a), *Annu. Rev. Phys. Chem.* **41,** 601–646.

Zwanziger, J. W., Koenig, M., Pines, A. (1990b), *Phys. Rev. A* **42,** 3107–3110.

Zwanziger, J. W., Rucker, S. P., Chingas, G. C. (1991), *Phys. Rev. A* **43,** 3232–3240.

Zygelman, B. (1987), *Phys. Lett. A* **125,** 476–481.

Further Reading

Aitchison, I. J. R. (1987), *Acta Phys. Polonica* **B18,** 207–235.

Avron, J. E. (1995), in: E. Akkermans *et al.* (Eds.), *Mesoscopic Quantum Physics, Proceedings of the Les Houches Summer School, Session LXI,* Amsterdam: Elsevier Science B.V., pp. 741–791.

Berry, M. V. (1989), in: A. Shapere, F. Wilczek (Eds.), *Geometric Phases in Physics,* Singapore: World Scientific, pp. 7–28.

Berry, M. V. (1989), *Sci. Am.* **256** (12), 34–40.

Berry, M. V. (1990a), *Phys. Today* December, 34–40.

Berry, M. V. (1990), in: R. G. Chambers *et al.* (Eds.), *Sir Charles Frank 80th Birthday Festschrift,* Adam Hilger, Bristol.

Berry, M. V. (1990), in: M. Bregoal *et al.* (Eds.), *Anomalies, Phases, Defects,* Naples: Bibliopolis, pp. 125–181.

Bhandari, R. (1997), *Phys. Rep.* **281,** 2–64.

Bohm, A. (1996), *Quantum Mechanics: Foundations and Applications,* 3rd ed., New York: Springer-Verlag.

Jackiw, R. (1988), *Int. J. Mod. Phys.* **A3,** 285–297.

Markovski, B., Vinitsky, S. I. (Eds.) (1989), *Topological Phases in Quantum Theory,* Singapore: World Scientific.

Marsden, J. E., O'Reilly, O. M., Wicklin, F. J., Zombro, B. W. (1991), *Nonlin. Sci. Today* **1,** 4–11, 14–21.

Marsden, J. E. (1992), *Lectures on Mechanics,* Cambridge, U.K.: Cambridge Univ. Press.

Mead, C. A. (1992), *Rev. Mod. Phys.* **64,** 51–85.

Shapere, A., Wilczek, F. (Eds.) (1989), *Geometric Phases in Physics,* Singapore: World Scientific.

Shapere, A., Wilczek, F. (1989), *Am. J. Phys.* **57,** 514–518.

Vinitsky, V. I., Derbov, V. N., Dubovik, V. N., Markovski, B. L., Stepanovskii, Y. P. (1990), *Usp. Fiz. Nauk* **160** (6), 1–49 [translation (1992): *Sov. Phys. Usp.* **33,** 403–429].

Zwanziger, J. W., Koenig, M., Pines, A. (1990), *Annu. Rev. Phys. Chem.* **41,** 601–646.

TOPOLOGY

S. P. SMITH, *Department of Mathematics and Computer Science, California State University, Hayward, Hayward, California, U.S.A.*

	Introduction	1
1.	**Point-Set Topology**	4
1.1	Basic Concepts	4
1.2	Continuity	6
1.3	Connectedness and Compactness	6
2.	**Algebraic Topology**	8
2.1	Some Basic Tools	8
2.2	Homotopy	10
2.3	Homology	12
2.4	Cohomology	15
3.	**Differential Topology**	15
3.1	Manifolds and Bundles	16
3.2	Vector Fields and the Poincaré–Hopf Index Theorem	19
3.3	Differential Forms and de Rham Cohomology	19
3.4	Morse Theory	20
4.	**A Brief Guide to Further Reading**	21
4.1	Point-Set Topology	22
4.2	Algebraic Topology	22
4.3	Differential Topology	22
4.4	Physical Applications	22
	Glossary	22
	Further Reading	24

INTRODUCTION

Topology is the study of properties that are invariant under continuous deformation. If an object A can be continuously deformed into an object B, and B can be continuously deformed into A, then A and B are considered topologically equivalent. For example, a two-dimensional circular disk A is equivalent to a square B since each may be continuously deformed into each other. Such a disk is not, however, equivalent to the circle that is its boundary, since there is no way to deform a disk to a circle without introducing a discontinuity, i.e., a hole or tear, in the disk. Topology is sometimes called "rubber sheet geometry." It ignores metric aspects of a space such as distance, angle, and area, and concentrates on the aspects associated with the relative position of its points. The word "topology" comes from the Greek words "topos" and "logos" and means "analysis of place"; in some older literature, topology is called *analysis situs.*

"Cut-and-paste" techniques are common methods used to construct spaces in topology. For example, imagine a square sheet of completely flexible and stretchable material sitting in the plane with the corner points labeled A, B, C, and D starting with the upper left-hand corner and moving clockwise as in Fig. 1. Stretching the square out lengthwise and gluing the top edge AB to the bottom edge DC, with A attaching to D and B attaching to C, gives an orientable space that is a portion of a cylinder; it has two sides, and its boundary is equivalent to two circles. In Fig. 1 sets of arrows are placed on identified sides to indicate the orientation used in gluing. If the edge AB is given a 180° twist before the gluing so that A attaches to C and B attaches to D, then the resulting surface is the *Möbius strip,* which is nonorientable: it has only one side, and its boundary is equivalent to a single circle. Walking once around the central circle of a Möbius strip, a traveler returns to the same point but in an upside-down position. In the same way, the usual *torus T* is obtained from the above square by gluing the side AB to the side DC and the side BC to the side AD. The first gluing produces the cylinder as before, which is then stretched out and bent around for the second gluing. The torus is equivalent to the surface of a doughnut or the surface of a ball with a single hole drilled through it. The *Klein bottle K* is obtained by gluing AB to DC as before, but giving BC a 180° twist before gluing it to AD so that B attaches to D, and C attaches to A. The *real projective plane* $\mathbf{R}P^2$

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

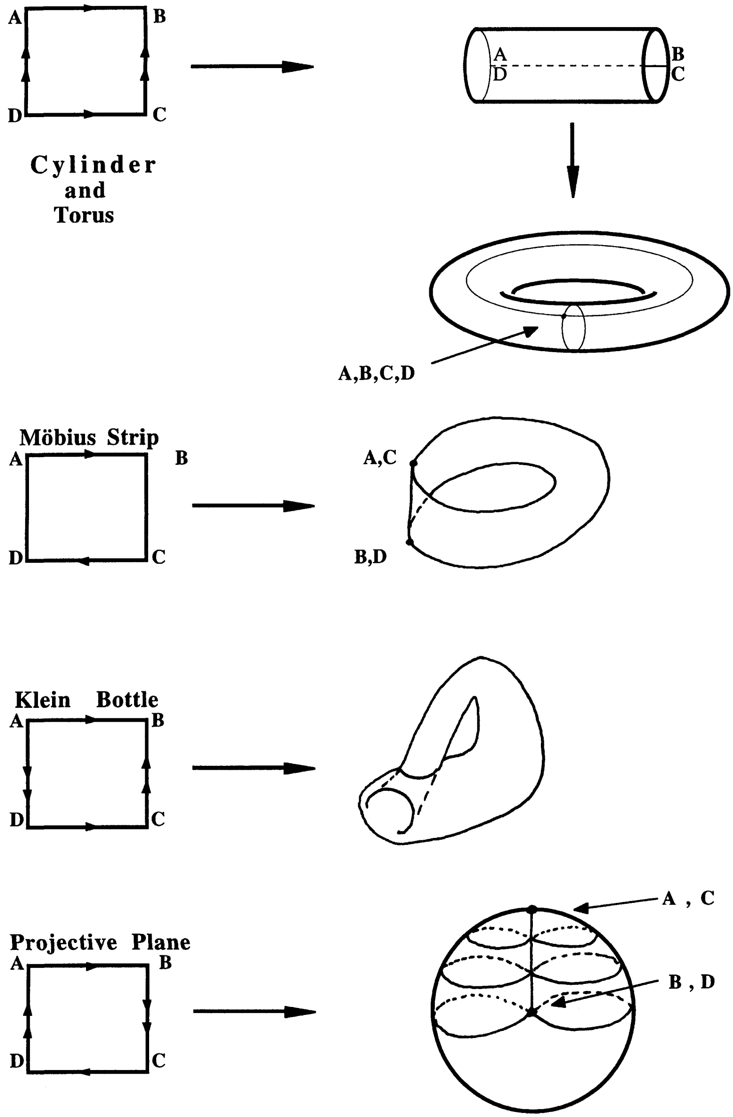

FIG. 1. Cut-and-paste methods for standard two-dimensional spaces.

results from twisting both AB and BC 180° before gluing, so that AB glues to CD and CB glues to AD with A attaching to C and B attaching to D. Of course, after the gluings for the torus, Klein bottle, and real projective plane, the points A, B, C, and D become a single point. The Klein bottle and real projective plane are nonorientable surfaces, and also both of these surfaces have self intersections when considered as subsets of $\mathbf{R}^3 = \{(x_1,x_2,x_3)|x_1,x_2,x_3 \in \mathbf{R}\}$. The above construction of $\mathbf{R}P^2$ can be viewed as taking a two-dimensional disk and gluing each point of its boundary to its antipodal point on the boundary. $\mathbf{R}P^2$ is the set of lines through the origin in $\mathbf{R}^3$. In general, *n-dimensional real projective space* $\mathbf{R}P^n$ is the set of all one-dimensional subspaces (i.e., lines through the origin) in $\mathbf{R}^{n+1}$, and the *n-dimensional complex projective space* $\mathbf{C}P^n$ is the set of all one-dimensional subspaces in $\mathbf{C}^{n+1}$. The dimensions in these projective spaces are taken with respect to the underlying scalars; thus $\mathbf{C}^3$ is three-dimensional as a complex manifold but six-dimensional as a real manifold. The ordinary sphere S^2 can also be obtained from the square ABCD by simply collapsing the boundary of the square to a point. If the

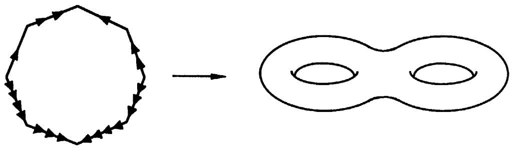

FIG. 2. Gluing the indicated sides produces the two-holed torus.

sides of a regular octagon are identified as in Fig. 2, the result is a two-holed torus.

After the spaces $\mathbf{R}^n$, the spheres are the most commonly encountered spaces in topology and are best considered simply as the set of points in $\mathbf{R}^{n+1}$ that are at unit distance from the origin; i.e., the *n-dimensional sphere* is $S^n = \{x \in \mathbf{R}^{n+1} \mid \|x\| = 1\}$. Each n-sphere S^n is the boundary of the unit $(n + 1)$*-dimensional ball* $B^{n+1} = \{x \in \mathbf{R}^{n+1} \mid \|x\| \leq 1\}$. Notice that the 0-sphere S^0 consists of two points and is therefore the only sphere that is not connected. The *g-holed torus* T_g is the two-dimensional space equivalent to the surface of a three-dimensional ball with g nonintersecting holes drilled through it. Using the above spaces, large numbers of examples can be constructed using additional operations such as Cartesian products and connected sums. Given spaces M and N, the *Cartesian product* $M \times N$ is defined to be $\{(p,q) | p \in M \ \&\ q \in N\}$. If M and N have the same dimension n, the *connected sum* $M \# N$ is constructed by removing an n-dimensional ball from each of M and N, and then gluing the resulting spaces together along the boundaries of the removed balls. Thus, the Cartesian product of two circles $S^1 \times S^1$ is the two-dimensional torus T, and the connected sum $T_g \# T_k$ of a g-holed torus and a k-holed torus is a $(g + k)$-holed torus.

Important isolated problems of a topological nature were treated by Descartes and Euler (the Descartes–Euler characteristic for polyhedra and the Königsberg bridge problem), and Riemann emphasized the importance of topological properties in his studies of complex analysis and algebraic functions. The key property of a Riemann surface is its genus, which counts the number of "holes" associated to the surface. Thus a two-dimensional sphere has genus 0, and a two-dimensional torus such as the surface of a doughnut has genus 1. With Cantor's development of set theory, mathematics entered a new era. More attention was paid to foundational questions, and a higher level of rigor was demanded. The real-number system had been carefully developed by Dedekind and Cantor in a constructive way, and mathematicians saw an opportunity to place all of mathematics on an axiomatic foundation similar to Euclid's *Elements*. The central areas in this new development were topology and abstract algebra. A very large part of twentieth century mathematics is associated with this program.

This development seems at first to be very remote from physics and the sciences. It is more like mathematical housecleaning: making more precise definitions, giving clearer statements and proofs of theorems, finding the weakest hypotheses that give a result, determining whether a property is metric or topological in nature, etc. However, it is important to note that most of the key questions arose from analysis of tools used in physical applications. Cantor's set theory came from his study of the convergence of Fourier series; Sophus Lie developed the theory of continuous groups to make the connection between symmetries and methods of solution of differential equations; Poincaré's deep study of the question of the "stability of the solar system" led him to realize that complicated systems were very difficult to treat quantitatively and that chaotic behavior was possible. In studying periodic orbits in celestial mechanics, Poincaré formulated a key difficulty as a topological conjecture, Poincaré's last geometric theorem, which was proved by G. D. Birkhoff. It was clear at this point that topology was essential to the qualitative analysis of dynamical systems. This was one of the major motivating forces behind the initial development of topology and remains so. Further, the development of general relativity and the associated area of differential geometry brought into focus the key relation between local and global aspects of spaces. To illustrate: Suppose that the given space is a two-dimensional sphere in three-space. Locally—for example, in the Northern hemisphere—there exist an infinite number of harmonic functions, i.e., solutions of the Laplace equation $\Delta f = 0$, even if the value of f is prescribed at the North Pole. However, globally only one such harmonic function exists; it must be a constant function. The topology of the sphere forces each function that is harmonic on the entire sphere to be a constant function. Similarly, it is possible to construct a unit-length tangent-vector field on the Northern hemisphere, but globally such a vector field is impossible. However, the two-dimensional torus does admit such a global unit-length tangent-vector field, and this torus is the only orientable surface without boundary with this property. This result follows from the famous Poincaré–Hopf index theorem discussed in Sec. 3.2. Topology plays an essential role in quantifying these global aspects of spaces and is therefore essential in global analysis and boundary-value problems. It is important to note that in applications, topology is usually combined with other areas of mathematics such as algebra and differential geometry.

This article discusses the three most prominent branches of topology: point-set topology, algebraic topology, and differential topology. These areas are considered to be closest to applications in physics.

1. POINT-SET TOPOLOGY

Point-set topology, also called general topology, provides the basic foundation for mathematical analysis, the portion of mathematics most directly connected with the calculus and the theory of differential and integral equations. The key concepts are continuity and convergence. The subject assumes only basic set theory. Point-set topology begins with topological spaces and continuous functions between such spaces. The basic properties of such spaces and functions are studied, and then gradually new concepts such as connectedness and compactness are added and examined. This step-by-step development leads to a clearer understanding of the role each hypothesis plays in a given result and often points the way to important extensions and generalizations. For example, the extreme-value theorem may be stated in a simple case as follows: A real-valued function that is continuous on a closed bounded interval $[a,b] \subset \mathbf{R}$ attains its maximum and minimum values at some points of the interval. Further analysis of this theorem shows that two basic ideas, continuity and compactness, are essential hypotheses for this theorem. The more general theorem is then this: a real-valued function that is continuous on a compact topological space attains its maximum and minimum values at some points of that space.

A very concise summary of essential definitions and results follows. This list is not comprehensive; it contains only the most commonly encountered definitions and results of point-set topology. These are particularly important for a clear understanding of functional analysis as used in physics. For example, point-set topology is basic to the theory of Fourier transforms, Sobolev spaces, finite-element methods, and the theory of distributions such as the Dirac delta function. The theory of distributions was a major motivation for the development of the theory of topological vector spaces.

Throughout this article, "iff" means "if and only if" and frequently indicates that a definition is being given.

1.1 Basic Concepts

A collection T of subsets of a set X is a *topology* on X iff

1. the empty set $\emptyset$ and the set X are in T;
2. T is closed under arbitrary unions; i.e., given any collection of subsets of X that are in T, their union is in T;
3. T is closed under finite intersections; i.e., given any finite collection of subsets of X that are in T, their intersection is in T.

A *topological space* (X,T) is a set X with a topology T on X. It is common to abbreviate (X,T) to X and *topological space* to *space*. If several topological spaces are involved, T may be replaced by T_X or another symbol such as S for clarity.

To illustrate, let $X = \{a,b,c\}$ be a set with three elements; then $T_1 = \{\emptyset,X\}$ and $T_2 = \{\emptyset,\{a\},\{b\},\{a,b\},X\}$ are topologies on X, but $T_3 = \{\emptyset,\{a\},\{b\},X\}$ is not since condition 2 is not satisfied. In general, for any set X, there are two standard topologies that can always be defined on X: the *discrete topology* consisting of the collection of all subsets of X, and the *indiscrete* or *trivial topology* that consists of only $\emptyset$ and X itself.

If (X,T) is a topological space, then the elements of T are called the *open sets* of the space. A subset A of a topological space (X,T) is *closed* iff its complement $X - A$ is open.

In the above example (X,T_2), $\{c\}$ is a closed set since its complement $\{a,b\}$ is open. Notice that open and closed are not opposites; a set can be both open and closed (e.g., X in T_1) or neither open nor closed (e.g., $\{a\}$ in T_1). The above topological spaces illustrate the definition clearly but are too artificial to be of much use in physics. A much more interesting example is given by $X = \mathbf{R}^n$, the n-dimensional vector space of all n-tuples of real numbers, with the *metric topology.* This is one of the most important topological spaces, and it is worth studying carefully. To define this topology some preliminary ideas are needed. For each pair of points a and b in $\mathbf{R}^n$, let $d(a,b)$ be the usual Euclidean distance from a to b. The *open ball* $B_r(p)$ centered at p and of radius $r > 0$ in $\mathbf{R}^n$ is defined to be $\{x \in \mathbf{R}^n | d(x,p) < r\}$, i.e., the set of all points strictly within distance r of p. Using these open balls the metric topology on $\mathbf{R}^n$ is defined as follows: A subset A is *open* in $\mathbf{R}^n$ iff for each $p \in A$, $B_r(p) \subset A$ for some $r > 0$. Intuitively, a subset A is open iff there is "space in all directions" in A around each of its points. This "space in all directions" is essential for the definition of key analytic concepts such as continuity and differentiation that require that the value of a function at a point be compared with values at all nearby points. Applying the definition is usually simple on an intuitive level. Thus, in the real line $\mathbf{R}$ the "open" interval $(0,1)$ is an open set, the "closed" interval $[-1,2]$ is a closed set since its complement, $(-\infty,-1) \cup (2,\infty)$, is an open set, and the half-open interval $(-3,5]$ is neither open nor closed.

In general, a metric topology may be defined on any set that has an appropriate notion of distance, i.e., a *metric,* defined on it. Often the metric is given by a more elaborate structure on the set such as an inner product or a norm. Inner products, norms, and metrics are formally defined as follows: Let V be a vector space over F where F is either the real or the complex number field, and let M be a set. Then $\langle , \rangle : V \times V \to F$ is an *inner product* on V iff

$\langle A,B \rangle = \overline{\langle B,A \rangle}$ for all $A, B, \in V$,
$\langle A, bB + cC \rangle = b\langle A,B \rangle + c\langle A,C \rangle$ for all $A, B, C \in V$ and $b, c \in F$, and
$A \in V$ and $A \neq 0$ imply that $\langle A,A \rangle > 0$.

An *inner product space* $(V,\langle , \rangle)$ is a vector space V together with an inner product $\langle , \rangle$ defined on it.

Next, $\|\cdot\|: V \to \mathbf{R}$ is a *norm* on V iff

$\|aA\| = |a|\|A\|$ for all $A \in V$ and $a \in F$,
$\|A + B\| \leq \|A\| + \|B\|$ for all $A, B \in V$, and
$A \in V$ and $A \neq 0$ imply that $\|A\| > 0$.

A *normed space* $(V,\|\cdot\|)$ is a vector space V together with a norm $\|\cdot\|$ defined on it.

Lastly, $d: M \times M \to \mathbf{R}$ is a *metric* on M iff

$d(x,y) = d(y,x)$ for all $x, y \in M$,
$d(x,z) \leq d(x,y) + d(y,z)$ for all $x, y, z \in M$, and
$d(x,y) \geq 0$ for all $x, y \in M$ and $d(x,y) = 0$ iff $x = y$.

A *metric space* (M,d) is a set M together with a metric d defined on it. Each of these structures gives the vector space V a natural metric topology since each inner product $\langle , \rangle$ induces a norm $\|\cdot\|$, which in turn induces a metric via the rules $\|A\| = \sqrt{\langle A,A \rangle}$ and $d(A,B) = \|A - B\|$.

If (X,T) is a topological space, and A is a subset of X, then A inherits a topology S from X as follows: B is *open in A* iff $B = A \cap U$ for some open set U in X. Briefly, this is written $B \in S$ iff $B = A \cap U$ for some $U \in T$. This topology is called the *relative topology* on A. It is also called the *induced* or *subspace topology* on A. Thus in the metric topology on $X = \mathbf{R}$ above, the interval $(1,2]$ is not open in X, but it is open in $A = [0,2]$ since $(1,2] = A \cap U$, with $U = (1,3)$ for example. A subset $N \subset X$ is a *neighborhood* of a point $p \in X$ iff $p \in U \subset N$ for some open set $U \in T$. If, in addition, N is an open set, then N is called an *open neighborhood* of p.

Let A be a subset of a topological space (X,T). $p \in X$ is an *interior point* of A iff $p \in U \subset A$ for some $U \in T$, i.e., A is a neighborhood of p. $p \in X$ is an *exterior point* of A iff p is an interior point of $X - A$. $p \in X$ is a *closure* or *limit point* of A iff each neighborhood of p contains at least one point of A. $p \in X$ is a *boundary point* of A iff p is a closure point of both A and $X - A$. $p \in X$ is an *accumulation point* of A iff each neighborhood of p contains at least one point of A other than p. $p \in A$ is an *isolated point* of A iff p is not an accumulation point of A. The *interior* of A is the set of all interior points of A and is denoted int$A$. It is also the union of all the open subsets of A and is the largest open set contained in A. The *closure* of A is the set of all closure points of A and is denoted cl$A$. It is also the intersection of all the closed subsets that contain A and is the smallest closed set containing A. The *boundary* of A is the set of all boundary points of A and is denoted bd$A$. It is equal to cl$A$ − int$A$. In many cases these concepts behave as the names indicate. Thus, for the metric topology on **R**, the set $A = [0,1)$ has the open interval $(0,1)$ as its interior, the closed interval $[0,1]$ as its closure, and the set $\{0,1\}$ consisting of the two "boundary points" as its boundary. However, these simple definitions contain some subtleties. For example, the subset of all rational numbers in $[0,1]$ has interior equal to the empty set, and both the closure and the boundary are equal to the closed interval $[0,1]$. The notations A°, $\bar{A}$, and ∂A are frequently used for int$A$, cl$A$, and bd$A$, respectively. A is *dense* in B iff $B \subset \mathrm{cl}A$. For example, the set **Q** of rational numbers is dense in the set **R** of real numbers, and the set of polynomials with rational coefficients is dense in the space $C^0([0,1],\mathbf{R})$ of all continuous real-valued functions on $[0,1]$ with the sup norm $\|f\| = \max\{|f(x)| \mid x \in [0,1]\}$. This notion of denseness is essential in approximation theory.

A topological space (X,T) is *Hausdorff* iff for each pair of points a and b there exist disjoint open sets U and V such that $a \in U$ and $b \in V$. Hausdorff spaces are also called T_2 spaces. Hausdorff is the most common of the several *separation* conditions that may be imposed on a topological space. All metric spaces are Hausdorff spaces.

The above definitions give the flavor of point-set topology. It is abstract and very set theoretic. It is important to keep in mind some of the primary examples that are discussed in the following such as surfaces in $\mathbf{R}^3$ with the induced metric topologies, topological and differentiable manifolds, and especially function spaces such as Hilbert and Banach spaces.

1.2 Continuity

The δ-ϵ definition for continuity given in calculus is as follows: a function $f{:}\mathbf{R} \to \mathbf{R}$ is *continuous at the point p* iff for each $\epsilon > 0$ there exists a $\delta_\epsilon > 0$ such that $|x - p| < \delta_\epsilon$ implies that $|f(x) - f(p)| < \epsilon$. The function is said to be *continuous on* $A \subset \mathbf{R}$ iff f is continuous at each point of A. Intuitively, continuous functions are allowed to stretch or contract portions of their domains, but no breaks or tears are permitted. The above definition is easily generalized to functions between metric spaces by simply replacing the absolute values by the appropriate distances, but it does not apply to more general topological spaces in which no such distance is given. The following definition generalizes the notion of continuity to topological spaces and is equivalent to the δ-ϵ definition above in the metric-space case. A function $f{:}(X,T_X) \to (Y,T_Y)$ is *continuous on X* iff $f^{-1}(U)$ is open for each $U \subset Y$; i.e., inverse images of open sets are open. Here the inverse image $f^{-1}(U)$ is defined to be $\{x \in X \mid f(x) \in U\}$; inverse images are defined for all functions, invertible or otherwise. The above function f is *continuous on* $A \subset X$ iff f restricted to A with the relative topology is continuous on A. It is common to abbreviate *continuous function* to *map*. Two topological spaces (X,T_X) and (Y,T_Y) are considered to be *topologically equivalent* or *homeomorphic* iff there exists a bijective function $f{:}(X,T_X) \to (Y,T_Y)$ that is continuous on X and that has an inverse f^{-1} that is continuous on Y. Such a function f is called a *homeomorphism*. For the purposes of topology, homeomorphic spaces are considered to be the same and must have identical topological properties.

1.3 Connectedness and Compactness

Compactness, connectedness, and path connectedness are among the most important topological properties and occur very frequently in topological analysis. Let A be a subset of a topological space (X,T_X). A collection of open subsets of X is an *open cover* of A iff A is contained in their union. A *finite subcover* of such an open cover is any finite subcollection that also contains A in its union. $A \subset X$ is *compact* iff each open cover of A has a finite subcover. For example, the open cover of the real line **R** given by $\{(n - 1, n + 1) \mid n \text{ is an integer}\}$ has no finite subcover; so **R** is not compact. Open covers are often denoted $\{U_i\}_{i \in I}$, where each U_i is an open set in the cover and I is an index set. $A \subset X$ is *connected* iff there do not exist open sets U and V in X with $A \subset U \cup V$, $A \cap U \neq \emptyset$, $A \cap V \neq \emptyset$, and $A \cap U \cap V = \emptyset$. Equivalently, $A \subset X$ is *connected* iff the only subsets of A that are both open and closed in the relative topology of A are $\emptyset$ and A. $A \subset X$ is *path connected* iff for each pair of points p and q in A there is a continuous function $\gamma:[0,1] \to A$ such that $\gamma(0) = p$ and $\gamma(1) = q$. Such a continuous γ is called a *path* in A from p to q. Path-connected spaces are necessarily connected, but not vice versa. If $f:(X,T_X) \to (Y,T_Y)$ is a function continuous on X and $A \subset X$ is compact, then the image $f(A)$ is also compact; similarly, if A is connected (resp. path connected), then $f(A)$ is also connected (resp. path connected). Two important theorems result from these concepts. Continuity plus compactness gives the extreme-value theorem: If $f:\mathrm{X} \to \mathbf{R}$ is continuous on X, and X is compact, then f must attain its maximum and minimum values at some points of X. Continuity plus connectedness gives the intermediate-value theorem: If $f:\mathrm{X} \to \mathbf{R}$ is continuous on X, and X is connected, and $f(a) < z < f(b)$ for some a and b in X and some real number z, then there is a $c \in X$ with $f(c) = z$. Two simple consequences of the intermediate-value theorem are the simple fixed-point theorem: Each continuous function $f:[0,1] \to [0,1]$ must have a fixed point, i.e., there is an $x \in [0,1]$ such that $f(x) = x$; and the heated-ring theorem: Given any continuous function $f:S^1 \to \mathbf{R}$, there exists at least one pair of antipodal points $x,x' = -x$ such that $f(x) = f(x')$. If S^1 is considered as the unit circle in the complex plane, and f is taken to be the temperature at each point, then this means that $f(x) = f(-x)$ for some $x \in S^1$; i.e., there are always at least two opposite points with the same temperature.

Continuity and compactness also combine to produce uniform continuity. A function $f:X \to \mathbf{R}$ is *uniformly continuous* on a metric space (X,d) iff for each $\epsilon > 0$ there is a $\delta_\epsilon > 0$ such that $x,y \in X$ and $d(x,y) < \delta_\epsilon$ imply $|f(x) - f(y)| < \epsilon$. The key point here is that δ_ϵ depends on ϵ but not on the points x and y involved. The important result is then that if $f:X \to \mathbf{R}$ is continuous on X, and X is compact, then f is uniformly continuous on X.

Compactness has several equivalent formulations, each appropriate to various applications. The Heine–Borel theorem says that $A \subset \mathbf{R}^n$ is compact iff A is closed and bounded. Here A is *bounded* iff A is contained in some sufficiently large open ball $B_r(p)$. This criterion is extremely easy to use, but it fails for most function spaces, which are usually infinite dimensional. For these more general cases, further conditions are usually required. As an illustration, consider the space $C^0([0,1],\mathbf{R})$ given in Sec. 1.1. The Arzela–Ascoli theorem says: $A \subset C^0([0,1],\mathbf{R})$ is compact iff A is closed, bounded, and equicontinuous. Here A is *equicontinuous* iff for each $\epsilon > 0$ there is a $\delta_\epsilon > 0$ such that $x,y \in [0,1]$ and $|x - y| < \delta_\epsilon$ imply $|f(x) - f(y)| < \epsilon$ for all $f \in A$. The essential part of this definition is that the same δ_ϵ works for all $f \in A$. Equicontinuity extends the notion of uniform continuity and is therefore sometimes called *uniform uniform continuity.*

A sequence $x_1, x_2, x_3, \cdots$ of points in a topological space X *converges* to $p \in X$ iff for each neighborhood U of p there exists an integer N_U such that $x_k \in U$ for all $k \geq N_U$. To make this definition useful, it is necessary to require that the space X be Hausdorff; otherwise a sequence may converge to two distinct points. For example, in the space $X = \{a,b\}$ with the indiscrete topology $\{\emptyset,X\}$, the sequence $a, b, a, b, a, b, a, \cdots$ converges to both a and b. This pathological behavior does not occur in Hausdorff spaces. A sequence $x_1, x_2, x_3, \cdots$ in a metric space M is a *Cauchy sequence* iff for each $\epsilon > 0$ there exists an integer N_ϵ such that $d(x_m,x_n) < \epsilon$ for all $m,n \geq N_\epsilon$. A metric space M is *complete* iff each Cauchy sequence in M converges to a point of M. A metric space M is *totally bounded* iff for each $\epsilon > 0$ there exists a finite collection of open balls of radius ϵ that contains M in their union; i.e., they form a finite open cover of M. A complete normed space is called a *Banach space.* A complete inner-product space is called a *Hilbert space.*

Four equivalent criteria for compactness are given as follows: A is compact iff

1. each open cover of A has a finite subcover (topological version);
2. each sequence in A has a subsequence that converges to a point of A (*sequential compactness* version);
3. A is complete and totally bounded (metric-space version); or
4. each infinite subset of A has an accumulation point in A (*Bolzano–Weierstrass property*).

The following properties are used frequently. Each closed subset of a compact topological space is compact. Each compact subset of a Hausdorff topological space is closed. The Tychonoff theorem states that the Cartesian product of an arbitrary number of compact topological spaces is compact.

An essential tool in the piecing together of local objects to form a global object is a *partition of unity.* Some definitions are needed first. A topological space X is *paracompact* iff X is Hausdorff and each open cover of X has a locally finite refinement. A *locally finite refinement* of an open cover $\{U_i\}_{i\in I}$ is an open cover $\{V_j\}_{j\in J}$ such that each V_j is contained in some U_{i_j} and each point of X is in at most a finite number of the V_j's. Given an open cover $\{U_i\}_{i\in I}$ of the space X, a family of continuous functions $\varphi_i{:}X \to [0,1]$ is a *partition of unity* subordinate to the cover $\{U_i\}_{i\in I}$ iff $\mathrm{supp}(\varphi_i) \subset U_i$ for each $i \in I$, and for each $p \in X$, $\varphi_i(p) = 0$ for all but a finite number of indices i and $\Sigma_{i\in I}\varphi_i(p) = 1$. Here $\mathrm{supp}(\varphi)$ denotes the *support* of $\varphi{:}X \to \mathbf{R}$ and is defined to be the closure of the subset of X on which φ is nonzero, i.e., $\mathrm{cl}\{p \in X|\varphi(p) \neq 0\}$. The basic theorem in this regard is this: Each paracompact topological space admits a partition of unity. This result is important for the theory of manifolds (see Sec. 3.1). It implies that each paracompact manifold can be given a Riemannian metric (see GEOMETRICAL METHODS, Sec. 2.3). Essentially, the coordinate charts on the manifold give induced metrics on parts of the manifold, which are then pieced together via an associated partition of unity.

2. ALGEBRAIC TOPOLOGY

The key idea in algebraic topology is to assign algebraic invariants to topological spaces in such a way that homeomorphic spaces have the same algebraic invariants. It is hoped that the algebraic information given by groups, rings, vector spaces, and homomorphisms is then easier to analyze. For example, if two spaces have different invariants, then they cannot be equivalent. This gives a method for distinguishing many spaces.

Physically, the most direct motivation for algebraic topology comes from potential theory. The fundamental problem of electrostatics is to find the equilibrium charge-density distribution on the surface of a charged conductor. This difficult problem is replaced by the simpler and in many cases more relevant problem of the determination of the electric field $\mathbf{E}$ produced by such a charge distribution. This field problem is in turn reduced to a function problem via the introduction of a potential function φ. The equation div $\mathbf{E} = 0$ then becomes the Laplace equation, $\Delta\varphi = 0$, with solution given by the line integral $\varphi(X) = \int_\gamma \mathbf{E}\cdot d\mathbf{l}$, where γ is any smooth path from a fixed point A to the variable point X. The path γ is also required to lie inside the region exterior to the conductor. The point A is usually taken to be a point at infinity or some point assumed to have zero potential. If the field $\mathbf{E}$ is given by $\mathbf{E} = (P,Q,R)$, and the line element $d\mathbf{l}$ is denoted by (dx,dy,dz), then $\varphi(X) = \int_\gamma Pdx + Qdy + Rdz$ expresses the potential as the integral of the differential one-form $\alpha = Pdx + Qdy + Rdz$ over a path from A to X. Further, the condition curl $\mathbf{E} = 0$ implies that the one-form α is closed (i.e., $d\alpha = 0$), and therefore the Stokes theorem implies that the line integral above will give the same value on any smooth path γ' connecting A to X that can be continuously deformed into γ without crossing the surface of the conductor. In such a case the paths γ and γ' are said to be *homotopic* to each other within the region exterior to the conductor. Two examples are useful here. If the conductor is a solid ball, then each pair of paths from arbitrary points A and X and lying in the region exterior to the ball are homotopic to each other in that region. Such a region is called *simply connected.* An equivalent form of this condition is that each closed path (*loop*) beginning and ending at a given point A should be continuously deformable within the region to the point A. For a nonsimply connected example, consider a solid torus (doughnut-shaped) conductor; then a loop at A that passes once through the hole in the torus is not deformable to the point A. This results, in general, in multivalued potential functions.

2.1 Some Basic Tools

Helmholtz's theorem also requires simply connected regions. A vector field $\mathbf{V}$ with curl $\mathbf{V} = 0$ in a simply connected region can be expressed as a gradient of appropriate potential functions, and a vector field $\mathbf{V}$ with div $\mathbf{V} = 0$ in a simply connected region can be expressed as a curl of appropriate vector potentials. These basic results are simplified and generalized by the theory of differential forms. For simplicity consider the case of $\mathbf{R}^3$ with the usual Cartesian coordinate system. A *0-form* f is simply a smooth real-valued function on $\mathbf{R}^3$, such as $f(x,y,z) = x^2y + \sin(xyz^2)$. A *1-form* α has the form $\alpha = Pdx + Qdy + Rdz$, and a *2-form* β has the form $\beta = Pdy \wedge dz + Qdz \wedge dx + Rdx \wedge dy$, where P, Q, and R are 0-forms. Finally, a *3-form* γ has the form $\gamma = fdx \wedge dy \wedge dz$, where f is a 0-form. The *wedge product* symbol $\wedge$ is used to emphasize that this multiplication is skew commutative; i.e., $dy \wedge dx = -dx \wedge dy$ and, in particular, $dx \wedge dx = 0$; $dy \wedge dz \wedge dy = 0$, etc. As a consequence, all differential forms of degree higher than the dimension of the underlying space are zero. The properties of these forms are based on those of determinants where interchanging any two distinct rows reverses the sign of the determinant. Determinants measure signed lengths, areas, volumes, and hypervolumes; differential forms measure infinitesimal signed lengths, areas, volumes, and hypervolumes and, therefore, appear as integrands in integrals giving such quantities. The wedge product generalizes the cross product to spaces of arbitrary dimension. The key operator in differential forms is the exterior derivative d. Letting E^p denote the real vector space of all p-forms on $\mathbf{R}^3$, this derivative maps each p-form α to a $(p + 1)$-form $d\alpha$, i.e., $d{:}E^p \to E^{p+1}$. Thus, for each 0-form f, $df = f_xdx + f_ydy + f_zdz$, where the subscripts denote partial derivatives; for each 1-form $\alpha = Pdx + Qdy + Rdz$, $d\alpha = (R_y - Q_z)dy \wedge dz + (P_z - R_x)dz \wedge dx + (Q_x - P_y)dx \wedge dy$; for each 2-form $\beta = Pdy \wedge dz + Qdz \wedge dx + Rdx \wedge dy$, $d\beta = (P_x + Q_y + R_z)dx \wedge dy \wedge dz$; and $d\gamma = 0$ for each 3-form γ. A key property of this derivative is $d^2 = 0$; i.e., $d(d\alpha) = 0$ for each p-form α. A p-form α is *closed* iff $d\alpha = 0$, and α is *exact* iff $\alpha = d\varphi$ for some $(p - 1)$-form φ. The $d^2 = 0$ property implies that each exact form is also a closed form. If a vector field $\mathbf{X} = (P,Q,R)$ is identified with the 1-form α above, then curl $\mathbf{X}$ corresponds to $d\alpha$; and if $\mathbf{X}$ is identified with the 2-form β above, then div $\mathbf{X}$ corresponds to $d\beta$. The basic vector analysis identities curl $\nabla\mathbf{X} = 0$ and div curl $\mathbf{X} = 0$ are special cases of $d^2 = 0$. In physics, the vector fields corresponding to 1-forms are called "polar" vector fields, and the vector fields corresponding to 2-forms are called "axial" vector fields. Using this formalism, classic formulas such as

the fundamental theorem of calculus:

$$\int_a^b f'(x)dx = f(b) - f(a),$$

Green's theorem:

$$\iint_G (Q_x - P_y)dx \wedge dy = \int_{\partial G} Pdx + Qdy,$$

the Stokes theorem:

$$\iint_S (\text{curl}\,\mathbf{X}\cdot\mathbf{n}dS) = \int_{\partial S} \mathbf{X}\cdot d\mathbf{s}, \text{ and}$$

the divergence theorem:

$$\iiint_D \text{div}\,\mathbf{X}dV = \iint_{\partial D} \mathbf{X}\cdot\mathbf{n}dS$$

are special cases of the "generalized Stokes theorem," which has the form

$$\iint \cdots \int_M d\alpha = \int \cdots \int_{\partial M} \alpha$$

for each n-dimensional orientable manifold M with boundary ∂M and $(n - 1)$-form α on M. Thus, integration of the right sort of n-form, namely, an exterior derivative, over a given region reduces to an integral over the boundary of the region. In a sense this results from the fact that whatever enters or leaves a region must cross the boundary. The differential-forms formalism and the above theorems are very directly connected to basic techniques of potential theory and are therefore relevant to electromagnetism and in particular to Maxwell's equations.

The constructions in algebraic topology are designed to measure the topological complexity of a given space relative to the above considerations. For example, if T is a cylindrical tube with boundary ∂T given by two circles C_1 and C_2 with proper orientations, and α is a given closed 1-form on M, then $\int_{C_1}\alpha = \int_{C_2}\alpha$. Thus, C_2 is equivalent to C_1 in these circumstances. Two such spaces that form the boundary of a third space of one higher dimension are said to be *homologous* and are considered to be equivalent in *homology*. For example, any two disjoint circles on the two-dimensional sphere S^2 form the boundary of an annular region on S^2, and so the homology group (as defined in Sec. 2.3) $H_1(S^2)$ of S^2 is 0. The *de Rham cohomology* of M (see Sec. 3.3) is defined directly in terms of the differential-forms structure as the quotient of the vector space of closed p-forms on M by the vector space of exact p-forms on M for each p and therefore measures the extent to which each closed form can be represented as an exact form. In the more general case, cohomology is defined by taking the appropriate notion of a *dual* of the homology of the space.

Some of the most famous theorems provable via algebraic topological methods are the following (let n be a positive integer, B^n be the unit ball in $\mathbf{R}^n$, and S^n be the unit sphere in $\mathbf{R}^{n+1}$):

Jordan–Brouwer separation theorem: Each compact connected hypersurface X in $\mathbf{R}^n$ divides $\mathbf{R}^n$ into two open sets: the "inside" of X and the "outside" of X. The special case with X a simple closed curve in $\mathbf{R}^2$ is called the Jordan curve theorem. A *hypersurface* in $\mathbf{R}^n$ is an $(n - 1)$-dimensional submanifold of $\mathbf{R}^n$.

No-retraction theorem: There is no continuous map from B^n to its boundary S^{n-1}.

Brouwer fixed-point theorem: If $f:B^n \to B^n$ is continuous, then f has a fixed point; i.e., there exists a point $p \in B^n$ such that $f(p) = p$.

Borsuk–Ulam theorem: If $f:S^n \to \mathbf{R}^n$ is continuous, then $f(x) = f(-x)$ for some $x \in S^n$; i.e., f takes the same value on at least one pair of antipodal points.

Ham-sandwich theorem: Let A, B, C be three bounded subsets of $\mathbf{R}^3$ each of which has a volume; then there exists a plane that bisects each of the three sets. (This theorem generalizes to the case of n bounded sets with hypervolume and guarantees the existence of a hyperplane in $\mathbf{R}^n$ that bisects each of them.)

To illustrate via a physical example: Imagine a fluid flowing inside a torus-shaped container in such a way that each point of the fluid is moving counterclockwise around the torus as in Fig. 3; then the Brouwer fixed-point theorem implies that this flow must have a closed (periodic) orbit. Here the B^n involved is a disk placed perpendicular to the central circle of the torus, and the map f associates to each point on this disk the return point obtained by flowing once around the torus. This application is close to one of Poincaré's original motivations for topology. He wanted to prove the existence of periodic orbits in phase-space flows associated with celestial mechanics. The intercepting space is called a *Poincaré section.*

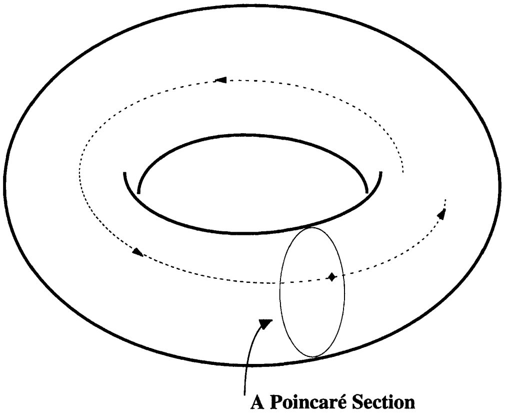

FIG. 3. Placing a Poincaré section in a fluid flow in a torus.

2.2 Homotopy

Intuitively, two paths from point P to point Q and lying in a region R are *homotopic* iff each can be continuously deformed into the other without leaving the region. This notion clearly depends on the region involved. For example, consider the paths γ_1 = the upper semicircle and γ_2 = the lower semicircle connecting the points $P = (-1,0)$ and $Q = (1,0)$ in the plane $\mathbf{R}^2$. These are homotopic in the region $\mathbf{R}^2$, but they are not homotopic in the punctured plane, $\mathbf{R}^2 - \{(0,0)\}$. Thus, if the one-form α has no singularities on or inside the unit circle, then $\int_{\gamma_1}\alpha = \int_{\gamma_2}\alpha$; but for the one-form

$$\alpha = \frac{-ydx + xdy}{x^2 + y^2},$$

which has a singular point at the origin, these two line integrals have distinct values. The *fundamental group* measures the complexity of a space in this regard. In Fig. 4 the underlying space is a torus, and the paths γ_1 and γ_2 are homotopic, but neither path is homotopic to the path γ_3.

Let I be the unit interval [0,1]. A *path*

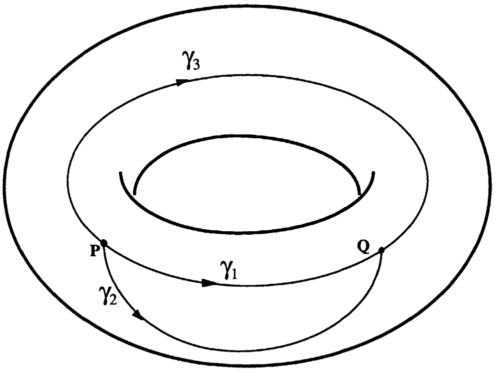

FIG. 4. Path γ_1 is homotopic to γ_2, but not to γ_3.

from P to Q in a topological space M is a continuous function $f{:}I \rightarrow M$ with $f(0) = P$ and $f(1) = Q$. Two paths f and g from P to Q in M are *homotopic* iff there exists a continuous function $F{:}I \times I \rightarrow M$ such that $F(0,t) = P$, $F(1,t) = Q$, $F(s,0) = f(s)$, and $F(s,1) = g(s)$ for all s and t in I. Thus F maps the unit square $I \times I$ into M so that all points on the left edge go to P, all points on the right edge go to Q, the bottom edge is mapped to the path f, and the top edge is mapped to the path g. This is illustrated in Fig. 5. This definition makes precise the notion of "continuous deformation" of f to g in M. Notice that all paths are reparametrized if necessary so as to have domain I. To define the fundamental group of M, a single point x_0 in M is fixed, and only *loops*, i.e., paths with initial and terminal points at x_0, are considered. Two loops are considered *equivalent* iff each can be deformed into the other within M. In particular, a loop is *trivial* iff it can be deformed to the basepoint x_0; i.e., it is homotopic to the constant path at x_0. The *fundamental group* of M with *basepoint* x_0 is defined to be the set of equivalence classes of such loops at x_0. If M is path connected, then the fundamental group is independent of the choice of basepoint. The group product operation $*$ in the fundamental group is defined as follows: Given two equivalence classes $[f]$ and $[g]$ of loops at x_0, pick representative loops, f and g, from each class; reparametrize f to have domain $[0,1/2]$ and g to have domain $[1/2,1]$, and define a new loop, denoted $f \cdot g$, which is just f followed by g; then the product is defined by $[f] * [g] = [f \cdot g]$. This operation is independent of the choices made and gives the set of equivalence classes a group structure. If M is path connected and the fundamental group of M is 0, i.e., the group has only one element, then M is said to be *simply connected*. The fundamental group of M is also called the *Poincaré group* or *first homotopy group* of M and is denoted $\pi_1(M,x_0)$ or just $\pi_1(M)$ if M is path connected. Simple examples are $\pi_1(\mathbf{R}^n) = 0$ for all n, $\pi_1(S^1) = \mathbf{Z}$ for the circle S^1, and $\pi_1(S^n) = 0$ for all spheres S^n with $n \geq 2$. The real projective plane has fundamental group $\mathbf{Z}_2$ as does the rotation group $R(3) = \mathrm{SO}(3)$. To see this, consider $R(3)$ as a solid ball of radius π in $\mathbf{R}^3$ with each point on its boundary identified to its antipodal point; i.e., each rotation in $\mathbf{R}^3$ with an axis corresponds to a point in the ball along that axis at a distance given by the angle of rotation about that axis. The identity rotation corresponds to the center of the ball, and each

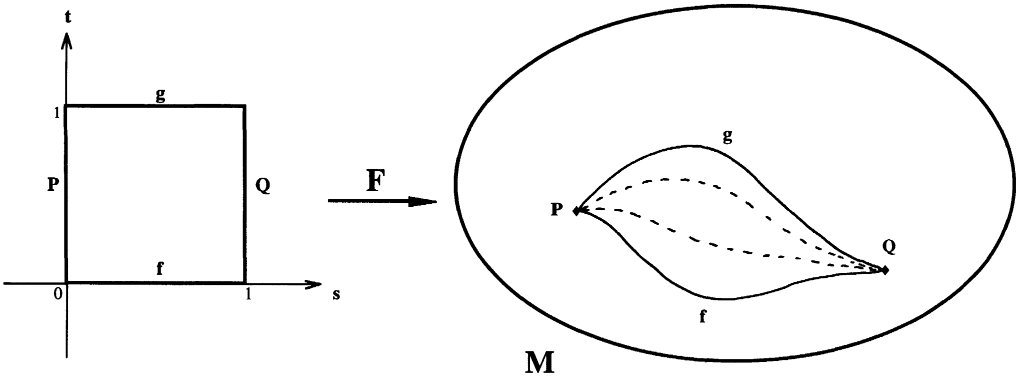

FIG. 5. A homotopy F between paths f and g.

point on the boundary is a rotation through π radians and corresponds to a rotation through $-\pi$ radians, which is the antipodal point. The loop starting at the origin, going upward to the North pole, which is identified with the South pole, and then continuing upward to the origin is not homotopic to a point; yet the same loop traversed twice is deformable to a point. This is the topological property of $R(3)$ that is associated to "spin" and "spinor" representations in quantum mechanics.

Each loop with basepoint x_0 can be considered as a map of the unit circle $S^1 \subset \mathbf{R}^2$ into M that sends (1,0) to the point x_0. $\pi_1(M)$ is then the set of equivalence classes of such maps. In an analogous fashion, higher homotopy groups $\pi_n(M)$ are defined to be the sets of equivalence classes of maps of n-spheres S^n into M, which map a chosen fixed point of S^n to the basepoint x_0. These groups detect higher-dimensional "holes" in the given space M. For example, let M be a solid ball of radius 3 with a concentric inner ball of radius 1 removed. M is simply connected, and so $\pi_1(M) = 0$. However, $\pi_2(M) \neq 0$ since no two-sphere that encloses the inner hole can be collapsed to a point inside of M. In general, $\pi_1(M)$ is non-Abelian. However, the higher-dimensional counterparts $\pi_n(M)$ are Abelian groups for $n \geq 2$.

One of the most important unsolved problems in topology is the Poincaré conjecture. Poincaré asked if a compact three-dimensional space with trivial fundamental group had to be homeomorphic to the three-sphere. Higher-dimensional versions of this conjecture have been proved to be true by Smale for all dimensions greater than 4 and by Freedman in dimension 4, yet the original case remains unresolved. This problem continues to play a very important role in the development of topology.

2.3 Homology

There are several ways to define homology groups for a given topological space M. Examples include combinatorial, simplicial, cellular, and singular homology. Each method has some advantages, and for the majority of cases the resulting groups are identical. Simplicial homology was very popular in the early development of algebraic topology in part because it is easy to visualize spaces as constructed step by step with standard pieces such as points, line segments, triangles, and tetrahedra that are called simplices and are more precisely defined later. Thus, any triangulated surface could be approximated by its associated simplicial complex. For example, a two-sphere is homeomorphic to a (hollow) tetrahedron. Using such simplicial approximations, the homology groups can be computed via linear algebra. Unfortunately, the computations are unwieldy for complexes with large numbers of simplices. Singular homology is more frequently encountered and is the only case treated here. Let $\mathbf{Z}$ be the integers, $\mathbf{R}$ be the real numbers, and $\mathbf{C}$ be the complex numbers.

The *standard p-simplex* Δ_p is defined by

$$\Delta_p = \{(t_0,t_1,\cdots,t_p) \in \mathbf{R}^{p+1} | t_0 + t_1 + \cdots + t_p = 1 \text{ and each } t_i \geq 0\}.$$

See Fig. 6. The *vertices* of Δ_p are $e_0 = (1,0,0,\cdots,0)$, $e_1 = (0,1,0,\cdots,0)$, $\cdots$, $e_p = (0,0,0,\cdots,1)$, which form the standard or-

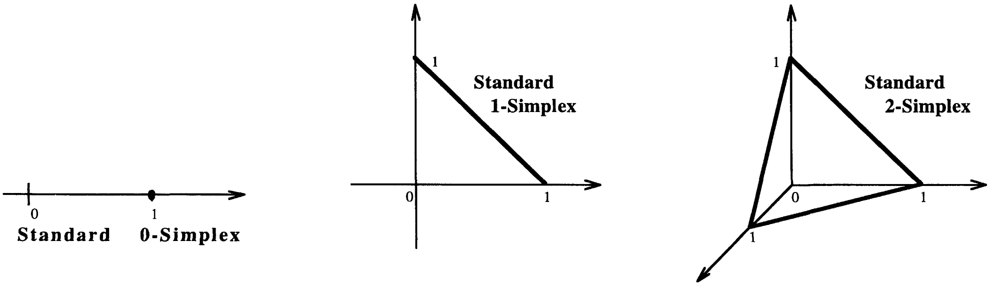

FIG. 6. Standard zero-, one-, and two-dimensional simplices.

thonormal basis in $\mathbf{R}^{p+1}$. Thus Δ_1 is the one-dimensional line segment from (1,0) to (0,1) in $\mathbf{R}^2$, and Δ_2 is the two-dimensional triangle with vertices at (1,0,0), (0,1,0), and (0,0,1) in $\mathbf{R}^3$. $[e_0,e_1,\cdots,e_p]$ is a convenient notation for Δ_p. A p-simplex with its vertices listed in such an order is called an *oriented p-simplex.* Two oriented p-simplices are *equivalent* iff one can be obtained from the other by an even permutation of the vertices, e.g., $[e_2,e_0,e_1]$ and $[e_0,e_1,e_2]$ are equivalent oriented standard 2-simplices.

A *singular p-simplex* in a topological space X is simply a continuous function $f:\Delta_p \rightarrow X$. Thus the points of X may be thought of as the images of singular 0-simplices, and paths in X correspond to the images of singular 1-simplices. The word *singular* is used to indicate that the dimension of the image may be less than the dimension p of the domain Δ_p. So points can in fact be the images of 1-simplices, 2-simplices, etc., and the image of any p-simplex can be regarded as the image of a q-simplex if q is larger than p. The boundary elements of each p-simplex break up into lower-dimensional simplices called *faces,* e.g., the triangle Δ_2 has three one-dimensional faces given by its edges and three zero-dimensional faces given by its vertices. For each $0 \leq i \leq p$, the ith *face* of the singular p-simplex f is the singular $(p-1)$-simplex $F_i(f):\Delta_{p-1} \rightarrow X$ defined by $F_i(f)(t_0,t_1,\cdots,t_{p-1}) = f(t_0,t_1,\cdots,t_{i-1},0,t_i,\cdots,t_{p-1})$. The *boundary operator* ∂_p is defined on each singular p-simplex f by $\partial_p f = F_0(f) - F_1(f) + \cdots + (-1)^p F_p(f)$. This agrees with the notion of boundary used in the Stokes theorem. To illustrate, let $f:\Delta_2 \rightarrow X$ be a singular 2-simplex, so that f maps the triangle $[e_0,e_1,e_2]$ in $\mathbf{R}^3$ to X. The boundary $\partial_2 f$ is then given by the combination of three 1-simplices: $\partial_2 f = F_0(f) - F_1(f) + F_2(f)$, where $F_0(f)$ corresponds to the map f restricted to side $[e_1,e_2]$, $F_1(f)$ corresponds to the map f restricted to side $[e_0,e_2]$, and $F_2(f)$ corresponds to the map f restricted to side $[e_0,e_1]$. The negative sign attached to $F_1(f)$ indicates that the orientation is reversed, so that $\partial_2 f$ simply traverses the boundary of f in the expected manner. See Fig. 7.

As seen above, the boundary operators give sums of simplices. The notion of a *p-chain* is introduced to define these correctly as well as to discuss more complicated subsets of X and to develop the algebraic structure of the theory. The pth *singular chain group* $C_p(X)$ of X is defined to be the free Abelian group generated by the set of all singular p-simplices in X; i.e., c is a *singular p-chain* in X iff $c = n_1 f_1 + n_2 f_2 + \cdots + n_k f_k$ for some integers $n_1, n_2, \cdots, n_k$ and singular p-simplices $f_1, f_2, \cdots, f_k$. So a singular p-chain is simply a finite integer-weighted sum of singular p-simplices. Intuitively, imagine such a sum as a collection of parametrized regions over which an integration is performed. An integer weight k indicates that the value of the integral over that region is counted k times; a negative weight indicates that the integral is taken with the opposite

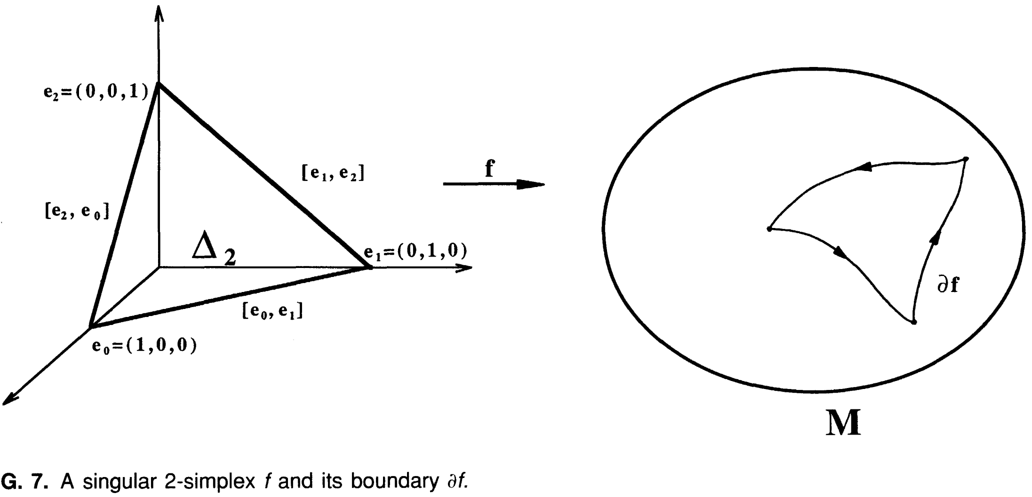

FIG. 7. A singular 2-simplex f and its boundary ∂f.

orientation, and therefore the sign of the value is reversed. The boundary operator extends to such singular p-chains by $\partial_p c = n_1\partial_p f_1 + n_2\partial_p f_2 + \cdots + n_k\partial_p f_k$. This gives a sequence of boundary operators $\partial_p: C_p(X) \to C_{p-1}(X)$ defined for each $p \geq 1$. ∂_0 is defined by $\partial_0 c = 0$ for each singular 0-chain c. These operators satisfy $\partial_p \circ \partial_{p+1} = 0$ for all $p \geq 0$; i.e., the boundary of a boundary is always 0 where by convention 0 represents the special degenerate singular chain whose image is the empty set. This important condition is usually abbreviated to $\partial^2 = 0$.

A singular p-chain $z \in C_p(X)$ is called a *p-cycle* in X iff $\partial_p z = 0$. $b \in C_p(X)$ is called a *p-boundary* in X iff $b = \partial_p c$ for some $c \in C_{p-1}(X)$. The group of p-cycles in X is denoted by $Z_p(X)$, and the group of p-boundaries in X is denoted by $B_p(X)$. By convention, $Z_0(X)$ is equal to $C_0(X)$. The condition $\partial^2 = 0$ implies that each p-boundary is also a p-cycle, and so $B_p(X) \subset Z_p(X)$ for each p. For each $p \geq 0$ the pth *singular homology group with integer coefficients* of the topological space X is defined to be the factor group $H_p(X,\mathbf{Z}) = Z_p(X)/B_p(X)$. Thus a sequence of Abelian groups is assigned to each space X. "Integer coefficients" refers to the integer weights in the chain groups; other coefficient groups may also be used.

Here are a few famous examples. Assume $n \geq 1$.

For n-dimensional real spaces $\mathbf{R}^n$,

$H_p(\mathbf{R}^n,\mathbf{Z}) = \mathbf{Z}$ if $p = 0$, and 0 otherwise.

For the n-dimensional spheres S^n,

$H_p(S^n,\mathbf{Z}) = \mathbf{Z}$ if $p = 0$ or n, and 0 otherwise.

For the real projective plane, $\mathbf{R}P^2$:

$H_p(\mathbf{R}P^2,\mathbf{Z}) = \mathbf{Z}$ if $p = 0$, $\mathbf{Z}_2$ if $p = 1$, and 0 otherwise.

Here $\mathbf{Z}_2$ is the Abelian group with exactly two elements. In general, $\mathbf{Z}_n$ denotes the cyclic group with exactly n elements. For the two-dimensional Klein bottle K:

$H_p(K,\mathbf{Z}) = \mathbf{Z}$ if $p = 0$, $\mathbf{Z} \oplus \mathbf{Z}_2$ if $p = 1$, and 0 otherwise.

For the two-dimensional torus $T = S^1 \times S^1$:

$H_p(T,\mathbf{Z}) = \mathbf{Z}$ if $p = 0$ or 2, $\mathbf{Z}^2 = \mathbf{Z} \oplus \mathbf{Z}$ if $p = 1$, and 0 otherwise.

For the two-dimensional g-holed torus T_g:

$H_p(T_g,\mathbf{Z}) = \mathbf{Z}$ if $p = 0$ or 2, $\mathbf{Z}^{2g}$ if $p = 1$, and 0 otherwise.

Here $\mathbf{Z}^{2g}$ denotes the direct sum of $2g$ copies of $\mathbf{Z}$. The nonnegative integer g is called the *genus* of T_g. For the n-dimensional tori $T^n = S^1 \times \cdots \times S^1$:

$H_p(T^n,\mathbf{Z}) = Z^{C(n,p)}$ for all $0 \leq p \leq n$, and 0 otherwise, where $C(n,p)$ is the binomial coefficient.

For the n-dimensional real projective spaces, $\mathbf{R}P^n$:

$H_p(\mathbf{R}P^n,\mathbf{Z}) = \mathbf{Z}$ if $p = 0$ or n with n odd, $\mathbf{Z}_2$ if p is odd and $0 < p < n$, and 0 otherwise.

For the $2n$-dimensional complex projective spaces, $\mathbf{C}P^n$:

$H_p(\mathbf{C}P^n,\mathbf{Z}) = \mathbf{Z}$ if $p = 0, 2, 4, \cdots, 2n$, and 0 otherwise.

In most cases of interest such as compact manifolds, the above homology groups are *finitely generated Abelian groups;* i.e., they can be put into the *cyclic normal form*

$$\mathbf{Z}^m \oplus \mathbf{Z}_{t_1} \oplus \mathbf{Z}_{t_2} \oplus \cdots \oplus \mathbf{Z}_{t_k}$$

for some integer $m \geq 0$ and integers $t_1, t_2, \cdots, t_k$ such that each $t_i > 1$ and t_i divides t_{i+1} for each $1 \leq i < k$. Here $\mathbf{Z}^m$ denotes the direct sum of m copies of $\mathbf{Z}$, and $\mathbf{Z}_n$ denotes the cyclic group of order n.

For $H_p(X,\mathbf{Z})$ in the above form, the integer m is called the pth *Betti number*, β_p, of X, and $t_1, t_2, \cdots, t_k$ are called the pth *torsion coefficients* of X, and the number k of such torsion coefficients is denoted τ_p. The *Euler–Poincaré characteristic* $\chi(X)$ is given by the alternating sum of the Betti numbers of X:

$$\chi(M) = \sum_{p \geq 0} (-1)^p \beta_p.$$

The homology groups, Betti numbers, torsion coefficients, and Euler–Poincaré characteristic are topological invariants of the space X; i.e., these invariants are the same for homeomorphic spaces. However, these invariants may also be the same for some nonhomeomorphic spaces such as $\mathbf{R}$ and $\mathbf{R}^2$. So these invariants are generally more useful in proving spaces are not homeomorphic than in proving them homeomorphic. If M and N are compact manifolds of dimension m and n, respectively, then $\chi(M \times N) = \chi(M) \times \chi(N)$, and if $m = n$, then $\chi(M \# N) = \chi(M) + \chi(N) - [1 + (-1)^n]$, where $M \times N$ is the Cartesian product and $M \# N$ is the connected sum of M and N.

In the 1920s Lefschetz defined the *relative homology* groups of a *pair* (X,A) of topological spaces with $A \subset X$. Intuitively, this corresponds to the homology of the space obtained from X by collapsing A to a point. Formally, the group of *relative p-cycles mod A* is $Z_p(X,A) = \{c \in C_p(X) | \partial_p c \in C_{p-1}(A)\}$, and the group of *relative p-boundaries mod A* is $B_p(X,A) = \{c \in C_p(X) | c - c' \in B_p(X)$ for some $c' \in C_p(A)\}$. The pth *relative singular homology group of the pair* (X,A) is then defined to be the factor group $H_p(X,A) = Z_p(X,A)/B_p(X,A)$ for each $p \geq 0$. One very important consequence of relative homology is the *long exact sequence in homology.* For each pair (X,A) two inclusion maps of pairs are naturally defined: $i{:}(A,\emptyset) \to (X,\emptyset)$ and $j{:}(X,\emptyset) \to (X,A)$. These induce natural maps i_* and j_* in homology, and combining these with appropriate induced boundary operators ∂_* yields the long exact homology sequence:

$$\cdots \to H_{p+1}(X,A) \xrightarrow{\partial_*} H_p(A) \xrightarrow{i_*} H_p(X)$$
$$\xrightarrow{j_*} H_p(X,A) \xrightarrow{\partial_*} H_{p-1}(A) \to \cdots$$

Such a sequence $\cdots \to G \xrightarrow{f} H \xrightarrow{g} K \to \cdots$ of groups connected by group homomorphisms is *exact* at the group H iff the image of f equals the kernel of g; i.e., $f(G) = \{a \in H | g(a) = e_K\}$, where e_K is the identity element in the group K. Thus, the long exact sequence in homology is exact at each group in the sequence. Further, each continuous map $f{:}(X,A) \to (Y,B)$ between pairs induces natural homomorphisms between their associated long exact sequences. This structure is very important in the theoretical development and application of algebraic topology in general. Similar long exact sequences also exist for homotopy and cohomology. To illustrate briefly a typical use of the structure of exact sequences, consider such a sequence

$$\cdots \to F \xrightarrow{f} G \xrightarrow{g} H \xrightarrow{h} K \to \cdots.$$

If the group F is the zero group, then the map g must be injective; if the group K is the zero group, then the map g must be surjective; thus if both F and K are zero groups, then the map g is an isomorphism, and the groups G and H are isomorphic.

2.4 Cohomology

Cohomology groups are additional algebraic structures associated to a given space. Like homology, there are various methods for defining and computing cohomology such as simplicial, cellular, singular, Čech, and de Rham. Singular cohomology proceeds by constructing dual objects called cochain groups and their associated coboundary operators and performing quotient group operation as in homology. As usual, for the spaces typically found in physics, such as manifolds, all of these methods produce essentially the same results. The cohomology groups of a space are dual to the homology groups in a sense similar to that of the duality between Dirac's bra and ket vectors, or between contravariant and covariant tensors. The most commonly encountered cohomology in physics is the de Rham cohomology associated with the differential forms on a manifold, and this is discussed in Sec. 3.

3. DIFFERENTIAL TOPOLOGY

Differential topology is the area of topology most directly associated with physics. The basic objects are smooth manifolds, and the functions of interest are smooth maps between such manifolds. Vector fields, differential forms, differential operators, and integration are naturally the key tools. The subject is very closely tied to differential geometry and global analysis, and these subjects usually appear together in applications. This section has much in common with the article GEOMETRICAL METHODS. This section treats manifolds, bundles, vector fields and differential forms, de Rham cohomology, and Morse theory. Manifolds are the natural answer to the question: What sort of spaces should be the domains and ranges of functions? It is clear that higher-dimensional spaces are needed for problems involving several variables such as dynamical systems. Also spaces with curvature are necessary to treat surfaces in Euclidean spaces as well as general relativity. Coordinate systems are required in order to apply the techniques of calculus and analysis. These allow problems in the space to be transferred to subsets of $\mathbf{R}^n$, analyzed, and then mapped back to the original space. Manifolds are spaces designed with these objectives in mind. Bundles, more precisely called fiber bundles, are structures built on manifolds. They answer the question: Where are objects such as vector fields, differential forms, and general tensor fields located? Perhaps the most familiar example in physics is phase space. Here the underlying "base" manifold is configuration space: the set of allowable positions or generalized positions for the system. A simple planar pendulum, for example, has the circle S^1 as its configuration space and phase space given by an infinite cylinder. This infinite cylinder is the "cotangent" bundle of S^1. The double planar pendulum has configuration space given by a two-dimensional torus T, and its phase space is the cotangent bundle of T, which is four-dimensional. Two of these dimensions give the position of the pendulum bobs, and the remaining two give their momenta. If the generalized momenta are replaced by generalized velocities, then the appropriate bundle is the tangent bundle of the configuration space M. Velocities are given by vector fields on M and are therefore located in the tangent bundle, whereas momenta are given by 1-forms on M and so are located in the cotangent bundle.

The topology of a manifold M places restrictions on the behavior of the functions, vector fields, and differential forms defined on M. For example, each smooth tangent vector field on a 2-sphere must be zero at some point. Intuitively, at each point in time there must be some point on the Earth at which the (horizontal component of the) wind velocity is zero. The Poincaré–Hopf index theorem makes this relation precise. Similarly, the theorems of de Rham and Hodge use de Rham cohomology to quantify the topological restrictions on the differential forms on manifolds. One of most beautiful areas of differential topology is Morse theory, which Marston Morse developed in connection with his study of geodesics on manifolds. The original motivation for this work came from the work of Poincaré and Birkhoff in dynamical systems where the motion of the system is described by geodesics. Morse theory quantifies the topological restrictions on the critical points of smooth nondegenerate functions. For example, the theory guarantees that each smooth nondegenerate function on a g-holed torus must have at least $2g + 2$ critical points.

Characteristic classes are specific cohomology classes assigned to bundles. They were designed to answer two questions: (1) When is a bundle trivial, i.e., a simple Cartesian product of the base manifold and the fiber? and (2) When are two bundles equivalent? An important area of application is the Atiyah–Singer index theorem that quantifies the number of solutions of certain operators on manifolds. Although mentioned briefly in this article, these topics require more extensive machinery than is reasonable in this discussion. See Sec. 4 for appropriate references.

3.1 Manifolds and Bundles

The notion of a manifold is a generalization of that of a surface in $\mathbf{R}^3$ such as a two-dimensional sphere. It is not possible to cover the sphere with a single coordinate system; at least two are needed. If each coordinate system is viewed as an observer, then in the regions in which their coordinate systems overlap, they should be able to compare data, i.e., effectively transform each coordinate system into that of any overlapping system. Since not only positions but also velocities and accelerations should be transformable, the changes of coordinates are required to be *smooth;* i.e., the functions and all their derivatives of all orders are required to be continuous. Typical examples are the configuration and phase spaces used in classical mechanics. These are usually higher-dimensional manifolds.

An *n-dimensional differentiable manifold M* is a locally Euclidean topological space.

More precisely, M has a covering by *coordinate charts,* which are homeomorphisms $\varphi_i{:}U_i \to \varphi_i(U) \subset \mathbf{R}^n$ for each $i \in I$ such that each U_i is an open subset of M, $M \subset U_{i\in I}U_i$, and on the overlap between the images of each pair of charts, the function $\varphi_j \circ \varphi_i^{-1}$: $\varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j)$ is a smooth function. See Fig. 8. Coordinate charts are also called coordinate systems, local charts, or just charts. On a manifold, calculus is done by transferring the process back to $\mathbf{R}^n$ via the coordinate charts, performing the appropriate calculus operations, and then transferring the result back to the manifold. For example, a function $f{:}M \to N$ between manifolds M of dimension m and N of dimension n is smooth iff $\psi \circ f \circ \varphi^{-1}$ is smooth as a map from an open set of $\mathbf{R}^m$ to an open set of $\mathbf{R}^n$ for each pair of charts φ on M and ψ on N. A smooth function $\varphi{:}M \to N$ with a smooth inverse is called a *diffeomorphism.* A manifold M is *orientable* iff a covering by coordinate charts $\{\varphi_i\}_{i\in I}$ can be chosen for M in such a way that the determinants of the Jacobian matrices of the overlap maps $\varphi_j \circ \varphi_i^{-1}$ are positive for each $i,j \in I$. Manifolds for which such a covering does not exist are called *nonorientable.* Among two-dimensional surfaces, the sphere and g-holed tori are orientable, whereas the Klein bottle and real projective plane are nonorientable. Very frequently, manifolds are assumed to be Hausdorff or even paracompact topological spaces, and this is the case for almost all cases typically encountered. If, in the above definition of differentiable manifold, $\mathbf{R}^n$ is replaced by $\mathbf{C}^n$ and the overlap maps $\varphi_j \circ \varphi_i^{-1}$ are required to be complex analytic, then the space M is called an *n-dimensional complex manifold,* and the tools of complex analysis are appropriate. *Manifolds with boundary* are defined by allowing two kinds of coordinate charts: the usual manifold style charts as given originally, plus special charts adapted for boundary points as follows. Let $\mathbf{R}^n_+ = \{(x_1,x_2, \cdots,x_n)|x_n \geq 0\}$ be the standard *upper half-space* in $\mathbf{R}^n$ with the induced topology. Then each point of a *manifold with boundary* is required to be homeomorphic to an open subset of either $\mathbf{R}^n$ or $\mathbf{R}^n_+$, and where charts overlap, their transition functions must be

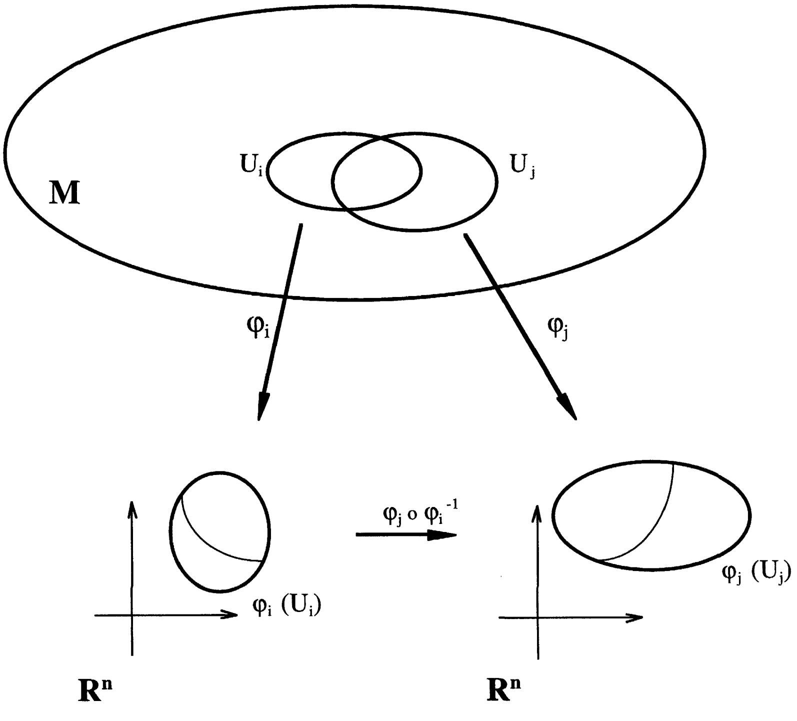

FIG. 8. Manifold charts ϕ_i and ϕ_j must have smooth transition functions $\phi_j \circ \phi_i^{-1}$ on overlaps $\phi_i(U_i \cap U_j)$.

smooth. In Fig. 9 the manifold with boundary is a torus with a disk removed at the top. Each point such as P on the boundary requires a half-space chart as illustrated by ψ. The points such as Q in the interior of M have usual manifold-style charts. The boundary of M is denoted ∂M. Typical examples are the Möbius strip and the upper hemisphere of the 2-sphere, each of which has boundary homeomorphic to the circle S^1, and the closed ball in $\mathbf{R}^n$ with boundary given by the $(n - 1)$-sphere S^{n-1}. In discussions involving manifolds both with and without boundary, manifolds without boundary are often called *closed manifolds.*

Although the initial motivation for manifolds came from spaces such as surfaces sitting in Euclidean spaces, the definition of a manifold does not assume this, and so it is an interesting question whether such abstractly given manifolds can be considered as embedded in $\mathbf{R}^n$ for n sufficiently large. The Whitney embedding theorem guarantees that this is the case, i.e., each paracompact manifold of dimension $n \geq 1$ has an embedding into $\mathbf{R}^{2n}$. Thus the Klein bottle always intersects itself when considered as a subset of $\mathbf{R}^3$, but it can be placed in $\mathbf{R}^4$ without such a self intersection.

Manifolds are the natural choice for the complicated geometrical structures that occur in physics. In addition to the configuration and phase spaces mentioned above, Minkowski spacetime, the universe of general relativity if singularities are ignored, and all Lie groups such as the rotation groups, Lorentz groups, unitary groups, and Poincaré groups are manifolds. However, structures such as vector fields, differential forms, and other tensors defined on such spaces require more general constructions, namely, bundles built on such manifolds. Bundles (see GEOMETRICAL METHODS) are special manifolds that are generalizations of Cartesian products. $\mathbf{R}^2$ is the Cartesian product of $\mathbf{R}$ and $\mathbf{R}$ and so has a global product structure. In a bundle only a local product structure is required. The most commonly used bundles on a manifold M are the tangent bundle TM, the cotangent bundle T^*M, the exterior product bundles $\Lambda^p M$, and more generally the tensor bundles $T^r_s M$. These are vector bundles; i.e., the fibers are vector spaces and the sections of these bundles give the vector fields, differential forms, and tensors on M. A *section* of a bundle is simply a smooth map that assigns to each point of the base manifold an element of the fiber over that point. Thus, a section of the tangent bundle of M is a (tangent) vector field on M. The two basic formulations of classical mechanics, Lagrangian and Hamiltonian, are associated to the bundles TM and T^*M, respectively. For example, if M is the configuration space of the system, then the phase space is the cotangent bundle T^*M, and the Hamiltonian is a map $H:T^*M \to \mathbf{R}$. In gauge theories vector bundles are replaced by the

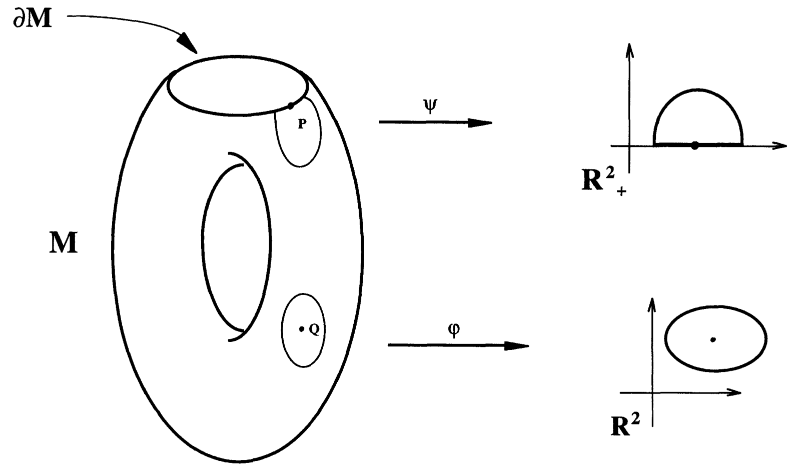

FIG. 9. Manifolds with boundary require "half-space charts" ψ for boundary points such as P.

more general principal fiber bundles in which each fiber is a Lie group appropriate to the symmetries of the fields involved. Characteristic classes are special cohomology classes giving topological information about bundles. These can be used, for example, to determine whether a given bundle has a global product structure, or to distinguish various bundles since bundles with different characteristic classes must be different. The Euler class, Stiefel–Whitney classes, Pontryagin classes, and Chern classes are standard characteristic classes. Secondary characteristic classes such as the Chern–Simons classes also play a role in gauge theories and mechanics.

3.2 Vector Fields and the Poincaré–Hopf Index Theorem

Viewed in terms of the above considerations, a (tangent) vector field on a manifold M is a section of the tangent bundle of M. Each point of M at which a vector field X is zero is called a *zero* of X, and the set of such zeros is denoted Z_X. If a zero p of X is isolated, i.e., p is the only zero of X in some neighborhood of p, then an integer *index* $i_X(p)$ can be assigned to p. Consider first the vector field X in $\mathbf{R}^2$ defined by $X(x,y) = (x^2 - y^2, 2xy)$. X has the origin, $p = (0,0)$, as its only zero. Let C be the circle of radius 1 centered at p, and consider the normalized unit vector field $U = X/\|X\|$ restricted to C. As C is traversed once counterclockwise, the tip of U performs exactly two counterclockwise rotations about its origin; therefore the index of X is 2. This X is not the field of a dipole, but it does have the same flow lines. Similarly, $X(x,y) = (x^3 - 3xy^2, 3x^2y - y^3)$ has the same flow lines as a quadrupole and index 3. The vector field $X(x,y) = (x,-y)$ has the flow lines of a saddle and index -1. For the case of a general planar vector field $X(x,y) = (P(x,y),Q(x,y))$ with an isolated zero at the origin p, the index may be computed by the formula

$$i_X(p) = \frac{1}{2\pi}\int_C \frac{PdQ - QdP}{P^2 + Q^2},$$

where C is a circle centered at p with radius small enough so that p is the only zero of X inside C, and C is traversed counterclockwise. In the n-dimensional case, the index is defined to be the degree of the map of a small $(n - 1)$-dimensional sphere S^{n-1}, centered at p and containing no other zero of X, to itself produced by appropriately normalizing X restricted to this S^{n-1}. This index is always an integer and is independent of the choice of radius of the sphere provided p is the only zero of X interior to that sphere.

The Poincaré–Hopf index theorem states that if X is a vector field having only isolated zeros on a compact orientable manifold M of dimension n, then

$$\sum_{p\in Z_X} i_X(p) = \chi(M).$$

Thus, among the orientable surfaces, the sum of the indices of the zeros of such a vector field must be 2 for the sphere and $2 - 2g$ for the g-holed torus. In particular, only the usual torus can admit a nonzero (tangent) vector field.

3.3 Differential Forms and de Rham Cohomology

It is hard to overemphasize the importance of differential forms in differential geometry and topology. Although they are just completely alternating covariant tensor fields on manifolds and are therefore part of the general theory of tensors, the simplicity of their structure and operations plus the direct connection with the underlying geometry of the space have given them a very special place in these subjects. Examples of forms in $\mathbf{R}^3$ were given in Sec. 2.1 above. See GEOMETRICAL METHODS for the more general case as well as definitions of the exterior derivative operator d, the Hodge star operator $*$, and the codifferential operator δ, which is the adjoint of d with respect to the inner product $(,)$ defined by $(\alpha,\beta) = \int_M \alpha \wedge *\beta$ for each pair of p-forms α and β on M. Here E^pM denotes the vector space of all p-forms on M. A Riemannian or pseudo-Riemannian metric is required for the definition of both $*$ and δ. If M has dimension n, then $d{:}E^pM \to E^{p+1}M$, $*{:}E^pM \to E^{n-p}M$, and $\delta{:}E^pM \to E^{p-1}M$ for each p. The $d^2 = 0$ condition again holds, i.e., $d(d\alpha) = 0$ for all $\alpha \in E^pM$, as well as $\delta^2 = 0$. The Laplace–Beltrami operator $\Delta{:}E^pM \to E^pM$ is then defined by $\Delta = d\delta + \delta d$. This generalizes the usual Laplace operator $\Delta = \nabla^2$, which is the divergence of the gradient on functions. The Laplace operator certainly has a claim to being the most important and the most studied operator in mathematical physics. It occurs in the Laplace equation, Poisson equation, heat equation, wave equation, wave-equation forms of Maxwell equations, and Schrödinger equation among others. The underlying reason is that it measures the deviation from equilibrium. So it is very natural to seek a generalization of this operator. However, it is important to be careful about the sign of the operator. The usual conventions for the Hodge star operator give the negative of the divergence of the gradient. This agrees with Maxwell's convention; he called this Δf the *concentration* of f. As a result of this convention the spectrum (set of eigenvalues) of Δ on a compact manifold is nonnegative. Δ is then a nonnegative, elliptic, self-adjoint, second-order operator on each E^pM. Also, Δ acting on one-forms corresponds to the negative of the usual Laplacian, $\nabla \operatorname{div} X - \operatorname{curl} \operatorname{curl} X$, on vector fields X as used in vector analysis. It is important to notice that since a metric is needed to define the Hodge star operator $*$, and therefore also the codifferential operator δ, the Laplace–Beltrami operator involves the geometry of the manifold as well as its topology.

Again, $\alpha \in E^pM$ is *closed* iff $d\alpha = 0$, and $\alpha \in E^pM$ is *exact* iff $\alpha = d\varphi$ for some $\varphi \in E^{p-1}M$. Setting $Z^p(M)$ equal to the set of closed p-forms, and $B^p(M)$ equal to the set of exact p-forms, $d^2 = 0$ implies that $B^p(M) \subset Z^p(M)$, and the *de Rham cohomology* vector spaces are defined by $H^p_{\mathrm{dR}}(M,\mathbf{R}) = Z^p(M)/B^p(M)$ for each p. The *Poincaré lemma* says that each p-form that is closed in a neighborhood equivalent to an open ball is exact in that neighborhood; the de Rham cohomology measures the extent to which this holds globally on the manifold M. The de Rham theorem says that this cohomology agrees with that derived via the more topological methods such as singular and Ĉech. Further, on each compact orientable manifold M of dimension n, the Hodge star operator induces an isomorphism called *Poincaré duality* between $H^p_{dR}(M,\mathbf{R})$ and $H^{n-p}_{dR}(M,\mathbf{R})$ for each p. The Hodge theorem gives an orthogonal decomposition $E^pM = \mathcal{H}^pM \oplus dE^{p-1}M \oplus \delta E^{p+1}M$, where $\mathcal{H}^pM$ denotes the set $\{\alpha \in E^pM | \Delta\alpha = 0\}$ of *harmonic* p-forms, $dE^{p-1}M$ is the set of *exact* p-forms, and $\delta E^{p+1}M$ is the set of *coexact* p-forms on M. Further, the dimension of the space of harmonic p-forms is $\beta_p(M)$, the pth Betti number of M. The Hodge theorem is a generalization of Helmholtz's theorem that represents vector fields in $\mathbf{R}^3$ as sums of curls and gradients. More sophisticated global theorems such as the Hirzebruch–Riemann–Roch theorem and the general Atiyah–Singer index theorem utilize the theory of characteristic classes.

3.4 Morse Theory

Morse theory relates the critical-point structure of a smooth function on a manifold to the topology of the manifold. Given a smooth function $f:A \to \mathbf{R}$ where A is an open subset of $\mathbf{R}^n$, a point $p \in A$ is a *critical point* of f iff

$$\frac{\partial f}{\partial x_i}(p) = 0$$

for each $i = 1, 2, \cdots n$; i.e., the gradient of f is 0 at p. Such a critical point of f is *nondegenerate* iff the determinant of the Hessian matrix $Hf(p)$ is nonzero. The *Hessian* $Hf(p)$ is the symmetric matrix of second derivatives of f evaluated at p; i.e.

$$(Hf(p))_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(p).$$

The *index* k of a critical point p is the number of negative eigenvalues of $Hf(p)$. For example, $f(x) = 3x^5 - 5x^3 + 1$ has a degenerate critical point at 0, and nondegenerate critical points at -1 and 1 with indices 1 and 0, respectively. The Rosenholtz function $f(x,y) = 4x^2e^y - 2x^4 - e^{4y}$ has exactly two critical points, located at $(-1,0)$ and $(1,0)$, both nondegenerate and of index 2. This remarkable function has two local maxima and no other critical points; its graph has two mountains but no saddle points. Index-0 critical points correspond to local minima, and critical points of maximal index (equal to the dimension of the domain of the function) correspond to local maxima; the critical points of intermediate index behave like saddle points.

The Morse lemma states that for each nondegenerate critical point p of index k of f

there is a neighborhood U of p and a set of coordinates $\{t_1,t_2,\cdots,t_n\}$ in U such that f can be expressed as $f(t_1,t_2,\cdots,t_n) = f(p) - t_1^2 - t_2^2 - \cdots - t_k^2 + t_{k+1}^2 + \cdots + t_n^2$ in U. So f has a *normal form* in some neighborhood of each of its nondegenerate critical points. This implies that each nondegenerate critical point is isolated. These local notions easily extend to smooth real-valued maps on manifolds, although it is worth noting that in general a Riemannian metric is needed to define the Hessian at noncritical points. A standard example is the height function f defined on a two-dimensional torus M placed on end on the x-y plane in $\mathbf{R}^3$ as in Fig. 10. The value $f(p)$ is just the height (the z coordinate) of the point p above the $z = 0$ plane. This function has four critical points located at a, b, c, and d, all nondegenerate: the bottom point a with index 0, the two saddle points b and c each of index 1, and the top point d with index 2. For the torus laid flat on the plane, the height function has two circles of critical points at top and bottom, and each of these critical points is degenerate; however, the slightest tilting of the torus from this position is sufficient to give a height function with all nondegenerate critical points. A smooth function $f{:}M \to \mathbf{R}$ is a *Morse function* on M iff each critical point of f is nondegenerate. The Morse functions are dense in the set of smooth functions on M in the following sense: Given any smooth function $f{:}M \to \mathbf{R}$, there exist arbitrarily nearby Morse functions.

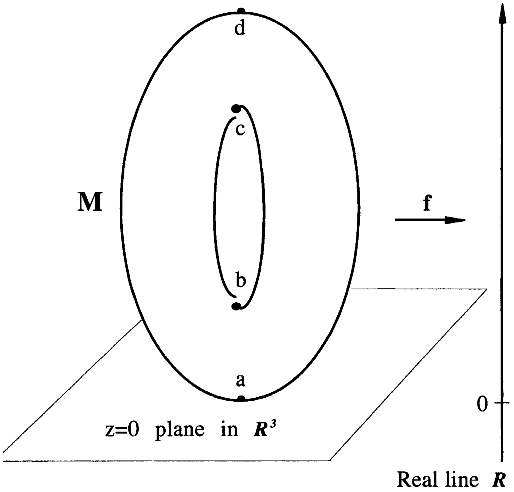

FIG. 10. The height function f has critical points at a, b, c, and d. The minimum at a has index 0, the saddle points at b and c have index 1, and the maximum at d has index 2.

Morse, following the lead of his advisor G. D. Birkhoff, found a sequence of inequalities giving topological restrictions on the number c_k of critical points of index k for a general Morse function f on a compact manifold M of dimension n. Letting β_p denote the pth Betti number of M, the *Morse inequalities* are

$$\begin{aligned} c_0 &\geq \beta_0 \\ c_1 - c_0 &\geq \beta_1 - \beta_0 \\ c_2 - c_1 + c_0 &\geq \beta_2 - \beta_1 + \beta_0 \\ &\vdots \\ \sum_{0\leq i\leq p} (-1)^{p-i} c_i &\geq \sum_{0\leq i\leq p} (-1)^{p-i}\beta_i \\ &\vdots \\ \sum_{0\leq i\leq n} (-1)^{n-i} c_i &= \sum_{0\leq i\leq n} (-1)^{n-i}\beta_i. \end{aligned}$$

Notice that the last relation is in fact always an equality. Thus the alternating sum of the *critical numbers* c_p of f equals the Euler–Poincaré characteristic $\chi(M)$. These relations immediately imply $c_p \geq \beta_p$ for each p. These inequalities were extended by Pitcher to include the influence of the torsion coefficients of M as follows: For each p,

$$c_p \geq \beta_p + \tau_p + \tau_{p-1},$$

where τ_p denotes the number of torsion coefficients in the cyclic normal form of the pth homology group $H_p(M,\mathbf{Z})$, as discussed in Sec. 2.3. As a simple application, consider a surface M sitting in $\mathbf{R}^3$. For almost all points p not on M, the distance function from p to each of the points of M is a Morse function on M (the other points are called *focal points*), and the critical points of this function correspond to lines from p to M that are perpendicular to M. Thus the Morse inequalities imply that the number of normals to M from any nonfocal point p is greater than or equal to the sum of the Betti numbers of M.

4. A BRIEF GUIDE TO FURTHER READING

Topology has a vast literature with texts ranging from very elementary to very specialized. This section is designed to select a relatively small number of readily available books in each of the areas covered in this article. These books are usually intended to be used in courses and therefore are rather condensed. More physics-oriented texts such as Felsager (1983), Nash and Sen (1983), von Westenholz (1981), Eguchi *et al.* (1980), Nash (1991), Göckeler and Schücker (1987), and Nakahara (1990) are demanding texts, but they usually keep physical applications in the forefront. Browsing through these texts for an initial orientation is strongly recommended.

4.1 Point-Set Topology

Many of the classic texts on point-set topology were designed to provide a foundation for analysis. Good examples are Simmons (1963), Wilansky (1970), and Kelley (1955). Texts by Munkres (1975), Sieradski (1992), Jänich (1984), Dugundgi (1966), and Hocking and Young (1961) also include discussion of algebraic-topological topics such as homotopy and homology. Most algebraic-topology texts include brief discussions of point-set topology topics. Armstrong (1983) and Kinsey (1993) are good examples at an undergraduate level. Most functional-analysis texts also contain relevant material; see, for example, Naylor and Sell (1982), Kreyszig (1978), Reed and Simon (1980), and Lang (1993).

4.2 Algebraic Topology

Seifert and Threlfall (1980), Veblen (1931), Cairns (1961), and Wallace (1968) are older texts that treat simplicial homology. Armstrong (1983), Kinsey (1993), and Jänich (1984) are good undergraduate-level introductions. Munkres (1984), Greenberg and Harper (1981), Rotman (1988), Massey (1991), Dold (1980), and Spanier (1981) are popular graduate-level texts. Hilton and Wiley (1960), Mac Lane (1963), and Vick (1994) treat homology theory, and Hu (1959) is a basic early reference for homotopy theory.

4.3 Differential Topology

Guillemin and Pollack (1974), Milnor (1965), Dodson and Poston (1991), and Wallace (1968) are good introductions. Bröcker and Jänich (1982), Hirsch (1976), Kosinski (1993), Lang (1995), and Warner (1983) are popular graduate-level texts. Springer (1981) and Weyl (1955) are classic introductions to Riemann surface theory. Steenrod (1951), Chern (1979), Kobayashi and Nomizu (1963, 1969), Gilkey (1995), Bott and Tu (1982), Milnor and Stasheff (1974), Hirzebruch (1966), and Husemoller (1994) are standard references for bundles and characteristic classes. Milnor (1963), Hirsch (1976), Bott (1994), and Kosinski (1993) discuss Morse theory. Eguchi *et al.* (1980), Gilkey (1995), Nash (1991), and Nakahara (1990) are good references for the Atiyah–Singer index theorem. Choquet-Bruhat *et al.* (1982, 1989) is a good general reference that covers a wide range of topics in differential topology and geometry particularly relevant for physics. Chern (1979), Wells (1980), and Hübsch (1992) treat complex manifolds.

4.4 Physical Applications

Flanders (1963) is the classic reference for differential forms. Schutz (1980) and von Westenholz (1981) are recommended for more advanced applications. Felsager (1983), Nash and Sen (1983), Eguchi *et al.* (1980), Nash (1991), Göckeler and Schücker (1987), and Nakahara (1990) are highly recommended for physicists. Birkhoff (1927) is a classic text in dynamical systems; more recent mathematically oriented texts are Arnol'd (1989), Abraham and Marsden (1978), and Abraham *et al.* (1988) for classical mechanics and dynamical systems, and Atiyah (1990), Green *et al.* (1987), and Hübsch (1992) for relations to string theories.

GLOSSARY

Accumulation Point (see Sec. 1.1)
Balls (see Introduction)
Banach Space (see Sec. 1.3)
Basepoint (see Sec. 2.2)
Betti Numbers (see Sec. 2.3)
Bolzano–Weierstrass Property (see Sec. 1.3)
Boundary (see Sec. 1.1)
Boundary Operator (see Sec. 2.3)
Boundary Point (see Sec. 1.1)
Bounded Subset (see Sec. 1.3)
Bundle (see Sec. 3.1)

Cartesian Product (see Introduction)
Cauchy Sequence (see Sec. 1.3)
Chain (see Sec. 2.3)
Chain Group (see Sec. 2.3)
Charts (see Sec. 3.1)
Closed Differential Form (see Secs. 2.2 and 3.3)
Closed Manifold (see Sec. 3.1)
Closed Set (see Sec. 1.1)
Closure (see Sec. 1.1)
Closure Point (see Sec. 1.1)
Coexact Differential Form (see Sec. 3.3)
Compact Space (see Sec. 1.3)
Complete Metric Space (see Sec. 1.3)
Complex Manifold (see Sec. 3.1)
Complex Projective Spaces (see Introduction)
Connected Space (see Sec. 1.3)
Connected Sum (see Introduction)
Continuity (see Sec. 1.2)
Convergent Sequence (see Sec. 1.3)
Coordinate Charts (see Sec. 3.1)
Critical Numbers of a Map (see Sec. 3.4)
Critical Point (see Sec. 3.4)
Cyclic Normal Form (see Sec. 2.3)
De Rham Cohomology (see Sec. 3.3)
Dense (see Sec. 1.1)
Diffeomorphism (see Sec. 3.1)
Differential Form (see Secs. 2.1 and 3.3)
Equicontinuity (see Sec. 1.3)
Euler–Poincaré Characteristic (see Sec. 2.3)
Exact Differential Form (see Secs. 2.1 and 3.3)
Exact Sequence (see Sec. 2.3)
Exterior Point (see Sec. 1.1)
Finite Subcover (see Sec. 1.3)
Finitely Generated Abelian Group (see Sec. 2.3)
First Homotopy Group (see Sec. 2.2)
Focal Points (see Sec. 3.4)
Fundamental Group (see Sec. 2.2)
***g*-Holed Torus** (see Introduction)
Harmonic Differential Form (see Sec. 3.3)
Hausdorff Space (see Sec. 1.1)
Hessian (see Sec. 3.4)
Hilbert Space (see Sec. 1.3)
Homeomorphic (see Sec. 1.2)
Homeomorphism (see Sec. 1.2)
Homologous (see Sec. 2.1)
Homology Group (see Sec. 2.3)
Homotopic Paths (see Secs. 2.1 and 2.2)
Hypersurface (see Sec. 2.1)
Index of a Zero of a Vector Field (see Sec. 3.2)
Inner Product (see Sec. 1.1)
Inner Product Space (see Sec. 1.1)
Interior (see Sec. 1.1)
Interior Point (see Sec. 1.1)
Isolated Point (see Sec. 1.1)
Klein Bottle (see Introduction)
Laplace–Beltrami Operator (see Sec. 3.3)
Limit Point (see Sec. 1.1)
Locally Finite Refinement (see Sec. 1.3)
Long Exact Sequence (see Sec. 2.3)
Loop (see Sec. 2.2)
Manifold (see Sec. 3.1)
Map (see Sec. 1.2)
Metric Space (see Sec. 1.1)
Metric Topology (see Sec. 1.1)
Möbius Strip (see Introduction)
Morse Function (see Sec. 3.4)
Neighborhood (see Sec. 1.1)
Nondegenerate Critical Point (see Sec. 3.4)
Nonorientable Manifold (see Sec. 3.1)
Norm (see Sec. 1.1)
Normed Space (see Sec. 1.1)
Open Ball (see Sec. 1.1)
Open Cover (see Sec. 1.3)
Open Neighborhood (see Sec. 1.1)
Open Set (see Sec. 1.1)
Orientable Manifold (see Sec. 3.1)
Oriented Simplex (see Sec. 2.3)
Pair of Topological Spaces (see Sec. 2.3)
Paracompact (see Sec. 1.3)
Partition of Unity (see Sec. 1.3)
Path (see Sec. 1.3)
Path Connected Space (see Sec. 1.3)
Poincaré Duality (see Sec. 3.3)
Poincaré Group (see Sec. 2.2)
Poincaré Section (see Sec. 2.1)
Real Projective Spaces (see Introduction)
Relative Homology (see Sec. 2.3)
Relative *p*-Boundary (see Sec. 2.3)
Relative *p*-Cycle (see Sec. 2.3)
Relative Topology (see Sec. 1.1)
Separation Conditions (see Sec. 1.1)
Sequential Compactness (see Sec. 1.3)
Simplex (see Sec. 2.3)
Simply Connected (see Secs. 2 and 2.2)
Singular Homology (see Sec. 2.3)
Singular Simplex (see Sec. 2.3)
Smooth Map (see Sec. 3.1)
Space (see Sec. 1.1)
Spheres (see Introduction)
Subspace Topology (see Sec. 1.1)
Sup Norm (see Sec. 1.1)

Support (see Sec. 1.3)
Topological Space (see Sec. 1.1)
Topologically Equivalent (see Sec. 1.2)
Topology (see Sec. 1.1)
Torsion Coefficients (see Sec. 2.3)
Torus (see Introduction)
Totally Bounded Metric Space (see Sec. 1.3)
Trivial Loop (see Sec. 2.2)
Vector Field (see Sec. 3.2)
Zero of a Vector Field (see Sec. 3.2)

Further Reading

Abraham, R., Marsden, J. (1978), *Foundations of Mechanics,* 2nd ed., New York: Addison-Wesley.

Abraham, R., Marsden, J., Ratiu, T. (1988), *Manifolds, Tensor Analysis, and Applications,* 2nd ed., New York: Springer-Verlag.

Armstrong, M. (1983), *Basic Topology,* New York: Springer-Verlag.

Arnol'd, V. (1989), *Mathematical Methods of Classical Mechanics,* 2nd ed., New York: Springer-Verlag.

Atiyah, M. (1990), *The Geometry and Physics of Knots,* Cambridge, U.K.: Cambridge Univ. Press.

Azcárraga, J. de, Izquierdo, J. (1996), *Lie Groups, Lie Algebras, Cohomology and Some Applications in Physics,* Cambridge, U.K.: Cambridge Univ. Press.

Birkhoff, G. (1927), *Dynamical Systems,* New York: American Mathematical Society.

Bott, R. (1994), *Collected Papers,* 4 volumes, Basel: Birkhäuser Verlag.

Bott, R., Tu, L. (1982), *Differential Forms in Algebraic Topology,* New York: Springer-Verlag.

Bredon, G. (1993), *Topology and Geometry,* New York: Springer-Verlag.

Bröcker, T., Jänich, K. (1982), *Introduction to Differential Topology,* Cambridge, U.K.: Cambridge Univ. Press.

Cairns, S. (1961), *Introductory Topology,* New York: Ronald Press Company.

Chern, S.-S. (1979), *Complex Manifolds without Potential Theory,* 2nd ed., New York: Springer-Verlag.

Choquet-Bruhat, Y., DeWitt-Morette, C., Dillard-Bleick, M. (1982, 1989), *Analysis, Manifolds, and Physics,* 2 volumes, Amsterdam: North-Holland.

Dodson, C., Poston, T. (1991), *Tensor Geometry,* New York: Springer-Verlag.

Dold, A. (1980), *Lectures on Algebraic Topology,* 2nd ed., New York: Springer-Verlag.

Dugundji, J. (1966), *Topology,* Boston: Allyn and Bacon.

Eguchi, T., Gilkey, P., Hanson, A. (1980), "Gravitation, Gauge Theories and Differential Geometry," *Phys. Rep.* **66,** 213–393.

Felsager, B. (1983), *Geometry, Particles and Fields,* 2nd ed., Odense, Denmark: Odense Univ. Press.

Flanders, H. (1963), *Differential Forms,* New York: Academic.

Fulton, W. (1995), *Algebraic Topology,* New York: Springer-Verlag.

Gilkey, P. (1995), *Invariance Theory, the Heat Equation, and the Atiyah–Singer Index Theorem,* 2nd ed., Boca Raton, FL: CRC Press.

Göckeler, M., Schücker, T. (1987), *Differential Geometry, Gauge Theories, and Gravity,* Cambridge, U.K.: Cambridge Univ. Press.

Green, M., Schwarz, J., Witten, E. (1987), *Superstring Theory,* 2 volumes, Cambridge, U.K.: Cambridge Univ. Press.

Greenberg, M., Harper, J. (1981), *Algebraic Topology, A First Course,* New York: Benjamin-Cummings.

Guillemin, V., Pollack, A. (1974), *Differential Topology,* Englewood Cliffs, NJ: Prentice-Hall.

Hilton, P., Wylie, S. (1960), *Homology Theory,* Cambridge, U.K.: Cambridge Univ. Press.

Hirsch, M. (1976), *Differential Topology,* New York: Springer-Verlag.

Hirzebruch, F. (1966), *Topological Methods in Algebraic Geometry,* 3rd ed., New York: Springer-Verlag.

Hocking, J., Young, G. (1961), *Topology,* New York: Addison-Wesley.

Hu, S. (1959), *Homotopy Theory,* New York: Academic.

Hübsch, T. (1992), *Calabi-Yau Manifolds, A Bestiary for Physicists,* Singapore: World Scientific.

Husemoller, D. (1994), *Fibre Bundles,* 3rd ed., New York: Springer-Verlag.

Jänich, K. (1984), *Topology,* New York: Springer-Verlag.

Kelley, J. (1955), *General Topology,* New York: Van Nostrand.

Kinsey, L. (1993), *Topology of Surfaces,* New York: Springer-Verlag.

Kobayashi, S., Nomizu, K. (1963, 1969), *Foundations of Differential Geometry,* 2 volumes, New York: Wiley-Interscience.

Kosinski, A. (1993), *Differential Manifolds,* New York: Academic.

Kreyszig, E. (1978), *Introductory Functional Analysis with Applications,* New York: Wiley.

Lang, S. (1993), *Real and Functional Analysis,* New York: Springer-Verlag.

Lang, S. (1995), *Differential and Riemannian Manifolds,* New York: Springer-Verlag.

Mac Lane, S. (1963), *Homology,* New York: Springer-Verlag.

Massey, W. (1991), *A Basic Course in Algebraic Topology,* New York: Springer-Verlag.

Milnor, J. (1963), *Morse Theory,* Princeton: Princeton Univ. Press.

Milnor, J. (1965), *Topology from the Differentiable Viewpoint,* Charlottesville, Va: The University Press of Virginia.

Milnor, J., Stasheff, J. (1974), *Characteristic Classes,* Princeton, NJ: Princeton Univ. Press.

Munkres, J. (1975), *Topology, A First Course,* Englewood Cliffs, NJ: Prentice-Hall.

Munkres, J. (1984), *Elements of Algebraic Topology,* New York: Addison-Wesley.

Nakahara, M. (1990), *Geometry, Topology and Physics,* Bristol: Institute of Physics Publishing.

Nash, C. (1991), *Differential Topology and Quantum Field Theory,* New York: Academic.

Nash, C., Sen, S. (1983), *Topology and Geometry for Physicists,* New York: Academic.

Naylor, A., Sell, G. (1982), *Linear Operator Theory in Engineering and Science,* New York: Springer-Verlag.

Reed, M., Simon, B. (1980), *Methods of Modern Mathematical Physics,* Vol. 1, rev. and enl. ed., New York: Academic.

Rotman, J. (1988), *An Introduction to Algebraic Topology,* New York: Springer-Verlag.

Schlichenmaier, M. (1989), *An Introduction to Riemann Surfaces, Algebraic Curves and Moduli Spaces,* New York: Springer-Verlag.

Schutz, B. (1980), *Geometrical Methods of Mathematical Physics,* Cambridge, U.K.: Cambridge Univ. Press.

Seifert, H., Threlfall, W. (1980), *A Textbook of Topology,* New York: Academic.

Sieradski, A. (1992), *An Introduction to Topology and Homotopy,* Boston: PWS-Kent.

Simmons, G. (1963), *Introduction to Topology and Modern Analysis,* New York: McGraw-Hill.

Spanier, E. (1981), *Algebraic Topology,* New York: Springer-Verlag.

Springer, G. (1981), *Introduction to Riemann Surfaces,* 2nd ed., New York: Chelsea Publishing Company.

Steenrod, N. (1951), *The Topology of Fibre Bundles,* Princeton: Princeton Univ. Press.

Stillwell, J. (1993), *Classical Topology and Combinatorial Group Theory,* 2nd ed., New York: Springer-Verlag.

Veblen, O. (1931), *Analysis Situs,* 2nd ed., Providence, RI: American Mathematical Society.

Vick, J. (1994), *Homology Theory,* 2nd ed., New York: Springer-Verlag.

von Westenholz, C. (1981), *Differential Forms in Mathematical Physics,* Amsterdam: North-Holland.

Wallace, A. (1957), *An Introduction to Algebraic Topology,* New York: Pergamon.

Wallace, A. (1968), *Differential Topology, First Steps,* New York: Benjamin.

Warner, F. (1983), *Foundations of Differentiable Manifolds and Lie Groups,* New York: Springer-Verlag.

Wells, R. (1980), *Differential Analysis on Complex Manifolds,* New York: Springer-Verlag.

Weyl, H. (1955), *The Concept of a Riemann Surface,* 3rd ed., New York: Addison-Wesley.

Wilansky, A. (1970), *Topology for Analysis,* Waltham, MA: Ginn and Company.

TRACKING PARTICLE DETECTORS

See DETECTORS, PARTICLE, TRACKING

TRANSDUCERS FOR MECHANICAL MEASUREMENT

Patrick L. Walter, *Department of Engineering, Texas Christian University, Fort Worth, Texas, U.S.A.*

Introduction 27
1. **Mechanical Displacement Measurement** 28
1.1 Variable Resistance 28
1.2 Variable Inductance 29
1.3 Variable Capacitance 29
1.4 Piezoelectricity 30
2. **Flexural Devices** 30
2.1 Bourdon Tubes 31
2.2 Cantilever Beams 31
2.3 Clamped Diaphragms 31
3. **Dynamic Model for Flexural Transducers** 31
3.1 Force, Pressure, Acceleration Transducers 32
3.2 Applications 32
4. **Measurement Planning Checklist** 33
Glossary 33
Works Cited 33
Further Reading 33

INTRODUCTION

Measurement involves the transfer of information about a process and, as a by-product, the transfer of energy from a process (Stein, 1992). Measurements are made by transducers. A transducer is a component in a measurement system that converts a physical quantity such as displacement, force, pressure, or acceleration into a proportional electrical quantity. A transducer can be modeled as having two input ports and one output port (Stein, 1992). At each input port, two direct, temporal, or spatial functions of energy exist. Their product possesses units of energy, energy rate, energy flux, etc. One input is desired while the second is undesired. For example, the inputs to a force transducer are force and the resultant displacement of the force collecting element. It would be desired to measure force using an infinitely stiff transducer. Since infinitely stiff force transducers do not physically exist, the transducer's input stiffness or impedance (force/displacement) should be as high as possible compatible with producing an adequate electrical signal output. Any deflection of the force transducer modifies the process being measured by absorbing energy (force times displacement) from the process. Similarly, the inputs to a pressure transducer are pressure and the accompanying volumetric change of the pressure collector. The volumetric change of the pressure collector modifies the pressure being measured through a similar energy (pressure times volume) absorption process. A pressure transducer's input impedance (pressure/volume) should be as high as possible.

A measurand is the physical or chemical process to be measured. A measurement system is a chain of components that link the measurand to the recorded data in a proportional manner. The transducer is the first component in the system. At subsequent component interfaces, it is desired to maintain the output impedance of the preceding component as small as possible and the input impedance of the next component as high as possible. For electrical components, a high input impedance (voltage/current) minimizes the current drain and thus the energy transfer from the preceding component. In addition to high input impedances and low output impedances required in all measurement system components, supplemental requirements exist for components used to measure time-varying phenomena. These requirements include a linear component input/output relationship to the measurand as well as flat amplitude and linear phase-frequency response characteristics.

Aside from the transducer, any measure-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ment system can be considered to be comprised of three additional subsystems.

Transmitting system—The transmitting system typically consists of some or all of the following components: cable, amplifier, subcarrier oscillator, filter, analog-to-digital (A/D) converter, transmitter, and antenna.

Receiving system—The receiving system typically consists of some or all of the following devices: antenna, preamplifier, multicoupler, receiver, tape recorder, discriminator, decommutator, digital-to-analog (D/A) converter, and output filter.

Data output and display system—The data output and display system typically consists of some or all of the following devices: oscilloscope, analog meter, digital meter, graphic display, and digital printer. Either these devices may be connected directly to the output of the receiving system or a computer may process the data from the receiving system before display.

The field of mechanical measurements encompasses or is synergistic with topics such as vacuum, temperature, acoustic, and flow measurements; telecommunications; calibration techniques; electrical instrumentation; and error analysis. These topics are described in other articles in the *Encyclopedia of Applied Physics*.

1. MECHANICAL DISPLACEMENT MEASUREMENT

The measurement of the basic quantity of linear displacement allows derived quantities such as length, width, thickness, position, level, erosion, wear, surface quality, strain, or vibration to be determined. Similarly, the measurement of angular displacement allows derived quantities such as attitude, angle of incidence, angle of turn, angle of flow, or angular vibration to be determined. Sensing techniques used to measure mechanical displacement include variable resistance, variable inductance, variable capacitance, and piezoelectric effect.

1.1 Variable Resistance

Common resistance-based transducers used to measure displacement include potentiometers and strain gages. The simplest displacement-measuring transducer is a linear or rotary potentiometer excited by a dc voltage source. It may be configured to produce an output voltage directly proportional to displacement. The resistance element is either a wire-wound or film potentiometer. The potentiometer's wiper arm is attached through a linkage to respond to motion.

Potentiometers are used for large displacements, typically operating over a range of 2–250 mm. The resolution of the wire element ranges from 0.05 to 0.5% of full scale. Linearity is typically 1.0% of full scale. Resistance values range from 100 to 1000 Ω and output impedance varies with wiper position. A potentiometric displacement transducer usually requires no signal conditioning because of its ability to produce a large output voltage. This output voltage is limited only by the supply voltage applied and the power dissipation capabilities of the potentiometer. Potentiometers are susceptible to shock and vibration. Their sliding wiper arm can become contaminated by dirt or corrosion. Wear due to friction associated with this wiper arm typically limits their life to 10^6 to 10^7 cycles.

Resistance strain gages are most commonly available in either foil or semiconductor form. Foil strain gages are essentially small printed circuits. A 0.003- to 0.005-mm-thick foil grid is manufactured from a process that includes a master gage pattern, photographic reduction of the pattern, application of a light-sensitive emulsion to the foil, exposure to the photographic plate, a development process, and chemical etching. The grid is then bonded to a backing. Alternately, semiconductor strain gages begin their fabrication process with a single high-purity silicon crystal. The crystal is sliced into wafers, and atoms such as phosphorus (*n* type) or boron (*p* type) are doped into the material at high temperature (>1000 °C) to lower its resistivity. Each wafer is then masked and etched to produce a suitable grid pattern that is bonded to a backing. The strain-induced change in resistance of either the foil or the semiconductor grid results in an electrical signal output when the strain gage is configured into a Wheatstone bridge circuit (Walter, 1990).

Strain gages can withstand only relatively small deflections. A strain-gage–based displacement transducer would consist of an

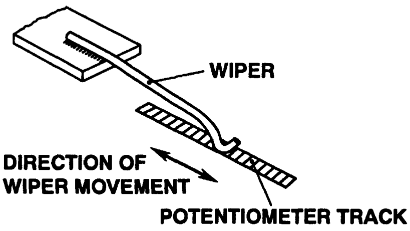

FIG. 1. Potentiometer.

elastic element with one to four attached strain gages. Walter (1990) provides additional description of resistance strain gages, their associated circuity, and their application in various transducers. Strain-gage transducers possess typical outputs from 30 to 500 mV and linearities from fractional percentage to 2% of full scale. Figures 1 and 2 show a potentiometer and a resistance strain gage, respectively.

1.2 Variable Inductance

A change in displacement can be electrically related to a change in inductance through a change in coil geometry, change in reluctance of a magnetic path by a varying air gap, change of permeability of a core material (magnetostrictive material), change in mutual coupling between two coils, or change in coupling between a primary and a secondary coil as in a transformer. A linear variable differential transformer (LVDT) illustrates the use of this last inductive coupling effect to measure displacement.

In an LVDT, three coils of conductive wire are wrapped along a hollow insulated form. A ferromagnetic core can move within the form along the common axis of the coils in either a positive or a negative direction. An ac voltage applied to the primary (center) coil induces voltages of the same frequency in each secondary coil. If the core is centered along the length of the three coils, and the two secondary coils are connected together in series opposition, the net voltage output from the LVDT is zero. A demodulator (ac signal-conditioning component producing a phase-sensitive dc output) can be interfaced to the series-connected output of the secondary coils to produce a proportional positive or negative output voltage as the core is displaced along the center line of the coils. LVDTs are typically used for displacements of 2 to 150 mm. The upper frequency displacement limit is a function of the demodulator but typically extends to tens or hundreds of hertz. LVDTs exhibit no wear, friction, or hysteresis, produce a continuous output, and are relatively insensitive to cross-axis motion. Dependent on their demodulator circuitry, they typically provide 1 to 10 V full-scale output. Their input–output is electrically isolated, and their operating temperature range can extend from −270 to 600 °C. Figure 3 provides a schematic of an LVDT. Herceg (1976) provides detailed information on LVDT design, selection, and application.

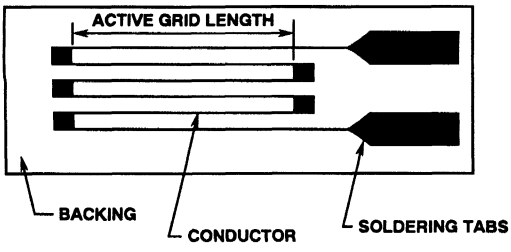

FIG. 2. Foil resistance strain gage.

1.3 Variable Capacitance

A capacitance displacement transducer can be designed by measuring the capacitance variation due to a change in distance

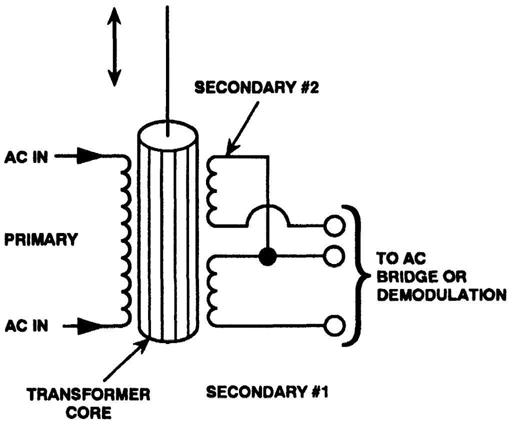

FIG. 3. Linear variable differential transformer (LVDT).

between its electrodes, keeping the electrodes fixed and measuring the change in capacitance due to a dielectric of varying permittivity being displaced between the electrodes, or changing the area of the electrodes in a manner that is related to displacement (e.g., by a rolling action of one electrode). Capacitive transducers known as Inductosyns are also used to measure linear or angular position. In an inductosyn, one capacitor plate moves over a second, which is configured in a grid or comb-type pattern, resulting in a pulse-type output. The pulse count can be a measure of position if grid spacing is known. Accuracies of 0.0025 mm can be obtained for linear displacements.

The simplest capacitance transducer takes advantage of a relative displacement between its electrodes. One capacitor plate of fixed area senses motion of either a metallic target plate or, if the target is an insulator, a metallic coated surface. The sensitivity is inversely proportional to the distance to be measured. Because of this relationship, capacitance transducers are well suited to measure small displacements. Typical displacement ranges extend from 3 to 6 mm. Variable-capacitance transducers, like variable-inductance transducers, are typically used in ac-powered bridge circuits with accompanying demodulated outputs ranging from 1 to 10 V. Figure 4 provides an illustration of a capacitance displacement transducer.

1.4 Piezoelectricity

Piezoelectric displacement transducers take advantage of the ability of a material to generate an electrical charge as a result of a mechanical stress that changes its shape. The basis for a piezoelectric material is absence of a center of charge symmetry. Thirty-two crystal forms occur in nature, of which 20 are piezoelectric. Quartz is a naturally occurring piezoelectric. Ceramics composed of ferroelectric crystals can be pressed, fired, electroded, and subjected to a high dc voltage to produce a material that will behave roughly like a single-domain piezoelectric crystal. Lead zirconate titanate is such a ceramic. The output impedance of any piezoelectric device is almost purely capacitive, and therefore, piezoelectric transducers will not produce a dc response to a static displacement. Their static response decays according to the circuit *RC* time constant. This time constant is determined by the transducer capacitance and the input impedance of its subsequent signal-conditioning component.

Piezoelectric transducers are extremely stiff and rugged, operate over a wide (10 000:1) dynamic range, and provide exceptionally high-frequency response. They can be made small in size and can operate over temperature ranges of −250 to +760 °C. As noted, they can perform only dynamic measurements. They require special cables and electronics and their high output impedance is susceptible to contamination.

2. FLEXURAL DEVICES

Force, pressure, and acceleration transducers can be constructed by having a displacement transducer sense the motion of a flexural element. This motion must be proportional to the quantity being measured.

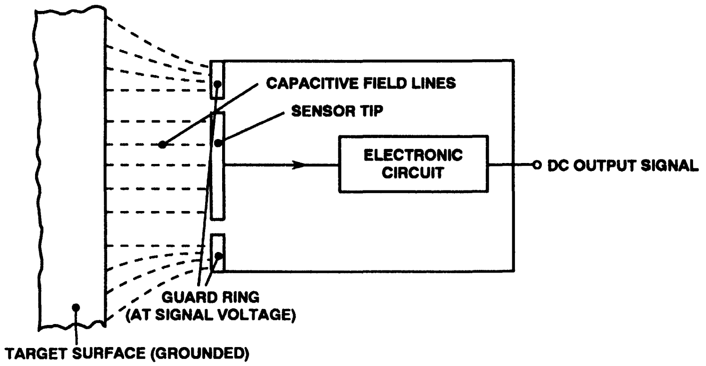

FIG. 4. Capacitance displacement transducer.

Potentiometric, inductive, and capacitive displacement transducers can directly measure this motion. The same is true for piezoelectric transducers if the motion is dynamic. Resistance strain gages take advantage of the strain pattern on the surface of the flexure. Among the gamut of flexural elements are Bourdon tubes, cantilever beams, and clamped diaphragms.

2.1 Bourdon Tubes

Bourdon tubes are one of the most widely used flexures for sensing pressure. The original patent for this device was granted to Eugene Bourdon in 1852. Bourdon tubes are hollow tubes that are twisted or curved along their length. The application of pressure deforms the tube wall that, depending on tube shape, causes it to untwist or unwind. Motion of the tube is measured and related to its internal pressure. Bourdon tubes can be integrated into transducers to achieve extremely high accuracies and have been manufactured from perfectly elastic materials such as quartz. Transducers employing Bourdon tubes tend to be physically large and easily damaged by environmental inputs such as acceleration. In addition, the tubes themselves afford poor frequency response to time-varying pressure.

2.2 Cantilever Beams

Cantilever beams are routinely used as the flexural element in transducers used to measure force, pressure, and acceleration. Strain near the clamped end of the beam can be measured, as can displacement of the free end, to correlate to force or torque applied to the free beam end, dynamic pressure associated with fluid flow acting over the beam surface, acceleration of the base to which the beam is mounted, etc. The compliance of a cantilever beam is defined as

$$y/F = L^3/3EI, \tag{1}$$

where y is the deflection of the free beam end, F is the force applied to this end, L is the beam length, E is the modulus of elasticity of the beam material, and I is the beam area moment of inertia. A compliant fixture will result in a transducer with a large electrical signal output. The above equation indicates that a compliant fixture can be achieved by a long, thin, narrow beam of low-modulus material. The penalty attached to such a design in application could be a transducer that is bulky, displays undesirable response to physical inputs orthogonal to its sensing direction, and has poor dynamic response.

2.3 Clamped Diaphragms

Clamped diaphragms are flexures used to transform strain or displacement into a measurement proportional to applied pressure. A small, flat, circular diaphragm can be made simply and can be placed flush against surfaces whose flow dynamics are being studied. This type of diaphragm is typically made to deflect in accord with theory associated with clamped circular plates. Corrugated diaphragms provide extensibility over a greater linear operating range than do flat diaphragms. A catenary diaphragm consists of a flexurally weak seal diaphragm bearing against a thin cylinder whose motion is measured. The compliance of a flat, circular diaphragm is defined as

$$y/P = 3R^4(1 - \nu^2)/16t^3E, \tag{2}$$

where y is the deflection of the center of the diaphragm, P is the applied pressure, R is the diaphragm radius, ν is Poisson's ratio, t is the diaphragm thickness, and E is the modulus of elasticity of the diaphragm material. Somewhat analogous to the cantilever beam, a compliant diaphragm would have a large radius, be thin, and be made of a low-modulus material. Equation (2) holds for deflections no greater than t.

3. DYNAMIC MODEL FOR FLEXURAL TRANSDUCERS

Flexural transducers can be dynamically modeled through their first resonant frequency as in Fig. 5. This single-degree-of-freedom system is characterized by a linear second-order differential equation with constant mass (m), viscous damping (c), and stiffness (k) coefficients. The governing equation for the system in Fig. 5 is

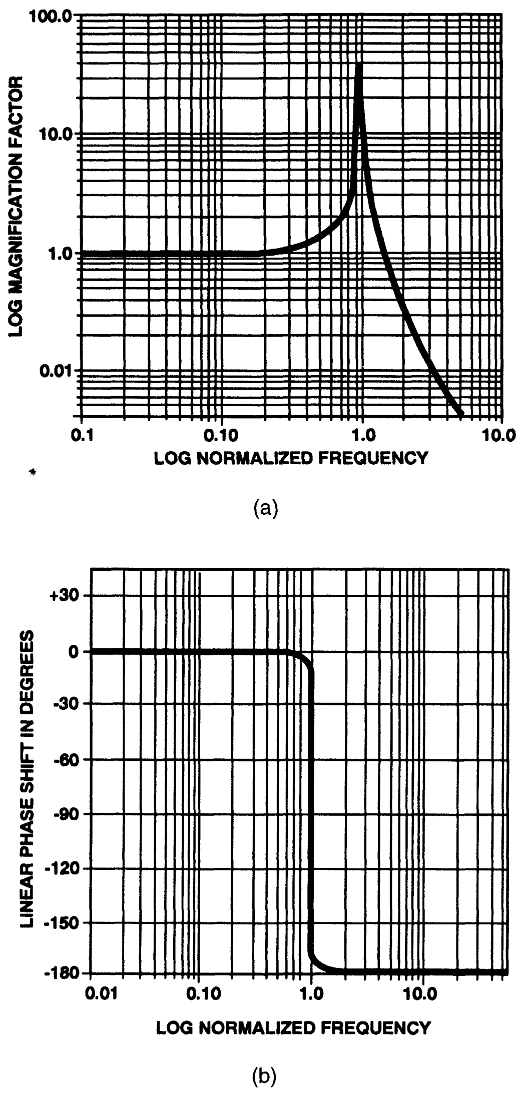

FIG. 5. Dynamic model for flexural transducers: (a) amplitude-frequency response; (b) phase-frequency response.

$$mx'' + cx' + kx = f(t), \qquad (3)$$

where the prime denotes a time derivative of the dependent variable x and $f(t)$ is the time-varying forcing function. Figure 5 represents a system viscously damped to 0.01 of critical, typical for transducers that are not intentionally damped. The maximum gain in Fig. 5(a) is 50:1, and the response is constant (flat) within 4% of the 0-Hz value to one-fifth of the maximum gain or resonant frequency. The phase in Fig. 5(b) is linear over this same range of frequencies. Thus, the criteria for measuring time-varying phenomena presented in the Introduction of this article are satisfied.

3.1 Force, Pressure, Acceleration Transducers

Force transducers measuring impact forces satisfy the model of Fig. 5 where the transducer's reciprocal compliance is the stiffness in Eq. (3), and the mass is the resultant summation of the effective mass of the transducer and the mass of the impacting object. Flush-diaphragm pressure transducers also satisfy the model of Fig. 5 where the reciprocal compliance of the diaphragm is the stiffness in Eq. (3) and its effective mass the mass term. Recess mounting of the transducer's diaphragm degrades transducer dynamic performance. Rigidly mounted accelerometers containing beam or other flexural elements satisfy the model of Fig. 5 in an analogous fashion as does the pressure transducer diaphragm. The significant difference in Eq. (3) when applied to force and pressure transducers versus accelerometers is that for force and pressure transducers x represents the absolute displacement of the mass while for accelerometers x represents the relative displacement between the mass and the accelerometer base. As mentioned in Sec. 2, this displacement can be measured with displacement transducers and related to the measurands of force, pressure, or acceleration as in Fig. 5.

3.2 Applications

Force transducers can be used in weighing or pressing applications where only their static response is important. Their dynamic response becomes important in applications such as dynamic force control, impact studies, and machine dynamics. Static pressure measurements are important in applications such as proof testing, monitoring of stationary flow processes, and liquid-level determination. Dynamic pressures are encountered in applications that include ballistics, explosions, cold-gas shock tubes, and combustion processes. Representative static accelerations can be found associated with centrifuges, rocket motor burns, and powered aircraft flights. Time-varying or dynamic accelerations are represented in vibration testing, mechanical shock testing, crash environments, and setback accelerations in cannons.

4. MEASUREMENT PLANNING CHECKLIST

Optimizing the mechanical measurement of displacement, either directly or as it relates to force, pressure, or acceleration stimuli, requires a knowledge of the physics of both the transducer and the process being measured. In adverse environments, it also requires an experienced measurements practioner. For the less experienced practioner, a checklist can provide guidance to assist in instrumentation system design. Consideration should be given to the following items:

1. Transducer type
2. Transducer interface to the measurand
   A. Special fixturing
3. Number of data channels
   A. Multiplexing desirability
   B. Location of signal conditioning
      1. Cable type
4. Signal conditioning type
5. Data filtering
6. Recording technique
   A. Time correlation requirements
7. Grounding and shielding
8. Calibration techniques
   A. Transducer calibration
   B. Measurement system checkout
   C. Field calibration
9. Data validation technique(s)
10. Data reduction technique(s)
11. Data presentation format(s)
12. Measurement system documentation

The included list of Further Reading contains detailed information about the various elements of this checklist to allow the successful design of measurement systems.

GLOSSARY

Measurand: The physical or chemical process to be measured.

Measurement: The transfer of information about a process and, as a by-product, the transfer of information from a process.

Transducer: A component in the measurement system that converts a measurand into a proportional electrical quantity.

Works Cited

Herceg, E. E. (1976), *Schaevitz Handbook of Measurement and Control,* Pennsauken, NJ: Schaevitz Engineering.

Stein, P. K. (1992), *The Unified Approach to the Engineering of Measurement Systems,* ISBN 1-881472-00-0, Phoenix, AZ: Stein Engineering Services, Inc.

Walter, P. L. (1990), in C. L. Nachtigal (Ed.), *Instrumentation and Controls—Fundamentals and Applications,* New York: Wiley, Chap. 8, pp. 265–306.

Further Reading

Dally, J. W., Riley, W. F., McConnell, K. G. (1984), *Instrumentation for Engineering Measurement,* New York: Wiley.

Doeblin, E. O. (1975), *Measurement System Application and Design,* New York: McGraw-Hill.

Nachtigal, C. L. (Ed.), (1990), *Instrumentation and Controls—Fundamentals and Applications,* New York: Wiley.

Stein, P. K. (1992), *The Unified Approach to the Engineering of Measurement Systems,* ISBN 1-881472-00-0, Phoenix, AZ: Stein Engineering Services, Inc.

TRANSDUCERS FOR TEMPERATURE, PRESSURE, AND FLOW

DAVID E. HOLCOMB AND TIMOTHY E. MCKNIGHT, *Oak Ridge National Laboratory, Instrumentation and Controls Division, Oak Ridge, Tennessee, U.S.A.*

Introduction 36
1. **Temperature Transducers** 36
1.1 Physics of Temperature 36
1.2 Temperature Scales 37
1.3 Thermocouples 37
1.4 Resistance Thermometers 38
1.4.1 Qualitative Sketch of Electrical-Conductivity Variance with Temperature in Metals 38
1.4.2 Qualitative Sketch of Electrical-Conductivity Variance with Temperature in Semiconductors 39
1.4.3 Resistance Temperature Detectors 40
1.4.4 Thermistors 40
1.5 Radiation Thermometers 41
1.5.1 Emission Physics 41
1.5.2 Types of Radiation Thermometers 42
1.5.3 Radiation Detectors 42
1.6 Special-Application Thermometers 42
1.6.1 Semiconductor-Junction (Diode) Thermometers 42
1.6.2 Liquid-in-Glass Thermometers 43
1.6.3 Resonant-Frequency Thermometers 43
1.6.4 Acoustic Thermometers 43
1.6.5 Noise Thermometers 44
1.6.6 Fiber-Optic–Coupled Phosphor Thermometers 45
1.6.7 Bimetallic Thermometers 45
1.6.8 Vapor-Pressure Thermometers 45
1.6.9 Fizeau-Cavity Thermometers 45
1.6.10 Distributed Fiber-Optic Thermometers 45
2. **General Concepts of Pressure and Pressure Instrumentation** 47
2.1 Physical Nature of Pressure 47
2.2 Manometers 47
2.3 Elastic-Member Pressure Transducers 48
2.4 Piezoresistive Transducers 49
2.5 Linear-Variable Differential-Transformer (LVDT) Transducers 49
2.6 Capacitive Deformation Transducers 49
2.7 Piezoelectric Transducers 49
2.8 Optical Deformation Measurement 50
2.9 Resonance Pressure Transducers 50
2.10 Hall-Effect Transducers 51
2.11 High-Pressure Ruby and Alexandrite Fluorescence 51
2.12 Low-Pressure Thermal-Conductivity Gauge 51
2.13 Low-Pressure Ionization Gauge 52
3. **General Concepts of Flow and Flow Instrumentation** 52
3.1 Nature of Flow and the Conservation Laws 52
3.2 Fluid Properties 53
3.3 Fully Developed Flow 54
3.4 Flowmeter Classifications 54
3.5 Inferential Flowmeters 54
3.5.1 Differential Pressure Flowmeters 54
3.5.2 The Orifice Plate 55
3.5.3 The Venturi and Nozzle 56
3.5.4 The Pitot and Pitot-Static Tube 56
3.5.5 The Laminar Flow Element 56
3.6 Variable-Area Flowmeters 57
3.6.1 The Rotameter 57
3.6.2 The Spring-Loaded Variable-Aperture Flowmeter 57
3.6.3 The Drag-Body or “Target” Flowmeter 58
3.6.4 The Vortex-Shedding Flowmeter 59

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

3.6.5 The Coanda, or Fluidic, Flowmeter 60
3.7 Direct-Measurement Flowmeters 60
3.7.1 The Positive-Displacement Flowmeter 60
3.7.2 The Turbine Flowmeter 60
3.8 Energy-Additive Flowmeters ... 60
3.8.1 The Magnetic Flowmeter 60
3.8.2 The Thermal Mass Flowmeter and Hot-Wire Anemometer 61
3.8.3 The Tracer-Dilution and Cross-Correlation Flowmeter .. 62
3.8.4 The Coriolis Flowmeter 62
3.8.5 The Ultrasonic Time-of-Flight and Doppler or Laser-Doppler Flowmeters 63
3.8.6 Nuclear-Magnetic-Resonance Flowmeters 63
Works Cited 64
Further Reading 65

INTRODUCTION

This article provides a conceptual overview of the physical principles underlying the operation of the common sensors for temperature, pressure, and flow. The article is not intended as a measurement engineering text and consequently does not address implementation issues. Important topics for measurement system application such as time and frequency response, cross-sensitivity, nonlinearity, response drift, measured-system disturbance, signal transmission, and noise are only adumbrated. For more information on measurement system design and application, the reader is referred to any of the several recent, excellent texts on the subject (see Further Reading). This article is neither mathematically rigorous nor exhaustive in scope, seeking instead to provide insight into the operational principles of sensors that are commonly employed.

The terms sensor, sensor element, transducer, and measurement are key to any description of the operation of measurement systems, yet these common terms are frequently confused. For the purpose of this article, the terms are defined as follows:

- transducer—an object that converts one form of energy into another;
- sensor element—the set of transducers used to convert energy from the measured form to a form usable to transfer information;
- sensor—an object that converts energy from a measured form into a form usable to transfer information; and
- measurement—the process of quantifying a physical process or event or the result of such a process.

A diaphragm-deflection–based pressure measurement system is illustrative of these terms. In this system the diaphragm whose front surface is exposed to the pressure being measured is the transducer. The transducer converts energy due to applied pressure into strain energy in the diaphragm. The strain energy in the diaphragm is in turn converted into an electrical signal by means of a strain gage attached to the diaphragm and a voltage source. The diaphragm and strain gauge together constitute the sensor element. The electrical signal is the form of energy used to transfer information about the quantity of pressure. The sensor comprises the sensor element and the voltage source.

1. TEMPERATURE TRANSDUCERS

1.1 Physics of Temperature

Temperature is a physical state parameter that represents the mean translational kinetic energy of the constituent elements (atoms, molecules, and electrons) of a thermal system. Temperature and heat are intimately related quantities. Temperature represents the potential for heat flow, while heat is the form of energy transferred when two systems at different temperatures are brought into thermal contact. Heat transfer occurs by one or more of the following methods:

- *conduction,* the diffusion of heat (without bulk atomic motion) from regions of higher to lower temperature;
- *convection,* the increased transfer of heat in the presence of a temperature gradient, due to bulk atomic motion; and
- *radiation,* the emission of heat in the form of electromagnetic energy due to accelera-

tion of the constituent elements of a thermal system in turn resulting from the interaction of the constituent elements because of their kinetic energy.

Many monographs exist on temperature and thermometry. The interested reader is directed to Quinn (1990), Kerlin and Shepard (1982), and Ricolfi and Scholz (1990) for more in-depth coverage.

1.2 Temperature Scales

Although in principle a temperature scale can be developed on the basis of statistical mechanics, relating temperature to mean kinetic energies (which are dependent only on mass, length, and time standards for their description), elemental kinetic energies are not generally measurable at present. The SI unit for temperature is the kelvin, which is defined as 1/273.16 of the temperature interval between absolute zero and the triple point of water. Two practical temperature scales, which remain in common use, were developed before the understanding of the thermodynamic basis of temperature. The Celsius temperature scale assigns 0 °C to the ice point (the water–ice equilibrium temperature at one atmosphere pressure) and 100 °C to water's boiling point at one atmosphere pressure. A one-degree Celsius change is equal to a one-kelvin change, and 0 °C is equal to 273.15 K. The Fahrenheit temperature scale originally assigned 0 °F to the lowest temperature that could be achieved by a mixture of ice, water, and sel-ammoniac and 96 °F to the blood temperature of a healthy man (Fraden, 1997, p. 99). The Fahrenheit scale was later amended, assigning 32 °F to the ice point and 212 °F to the boiling point of water.

1.3 Thermocouples

The operation of a thermocouple can be qualitatively understood using a simple free-electron–gas model of the conduction process. A more rigorous description of thermoelectricity is provided in Pollock (1985) and Blatt *et al.* (1976). Park (1993) provides detailed information on the application of thermocouples to temperature measurement. For a conductor in a thermal gradient, the electrons at the hot end have more kinetic energy than those at the cold end. The increased motion of the electrons results in a net diffusion of electrons from hot to cold. The skewed distribution of electrons produces an electric potential gradient opposing the thermal diffusion. The detailed mechanics of the diffusion are intimately involved with how the electrons scatter both among themselves and with the lattice atoms. The relation of the mean free path of electrons in a conductor and their kinetic energy is quite complex. The effect of point defects in the lattice structure (grain boundaries, interstitial impurities, etc.) has a strong influence on the quantitative value of the local diffusion parameters. The dependence of electron diffusion on lattice parameters results in thermally induced electric potentials being substantially different for different materials. It is these differences that are exploited to create thermocouple temperature sensors.

A thermocouple consists of two electrical conductors with different thermoelectric coefficients that are electrically connected at one end, isolated along their length, and connected to an electrical potential measurement circuit at the other end (see Fig. 1). Note that since all conductors exhibit thermoelectric potential (including those in the measurement circuit), the measurement ends of both thermoelements should be at the same temperature.

The change in electric potential in a material with temperature is quantified as the Seebeck coefficient, $S(T)$. The voltage developed along a thermocouple circuit (ΔV) can be expressed as the integral of $S(T)$ along the circuit:

$$\Delta V = \int S(T)dl, \tag{1}$$

where dl represents differential length along the thermoelements and the integral is performed from one end to the other of the thermocouple circuit.

Almost any set of conductors can be used as a thermocouple. However, eight sets of thermoelements with advantageous properties (large voltage response, lattice stability, durability, etc.) have received *letter designation* (type J, K, S, R, B, E, N, and T) and are commonly available commercially (ASTM, 1993).

Thermocouples are among the most du-

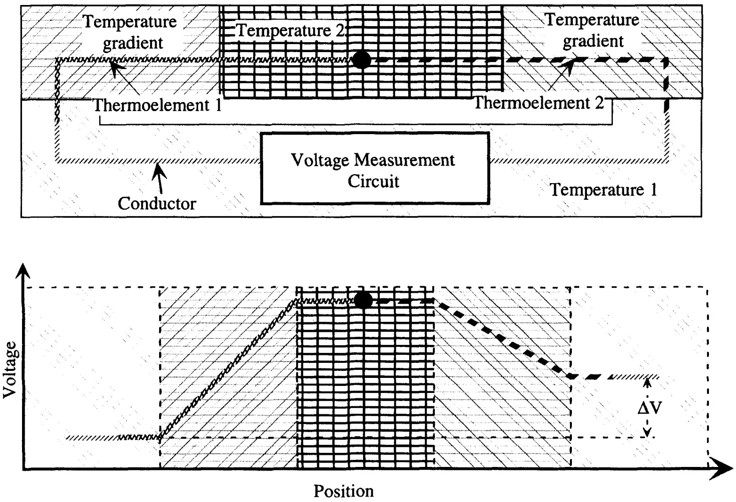

FIG. 1. Thermocouple voltage development.

rable and widely used temperature sensors available. Although accuracies of 0.25% or better are achievable using standards-laboratory practices, thermocouples, as commonly used in industrial practice, typically have inaccuracies of ~0.75% of reading. Anything that effects the mean free path of the electrons in the thermoelement lattice (impurity diffusion, work hardening, phase changes, etc.) affects the thermoelectric potential buildup. At higher temperatures, electrical shunting between the elements also becomes a concern.

A key concept in understanding thermocouple operation is that thermocouple voltage is developed over regions with temperature gradients and *not* across the thermoelement junction. Since thermocouple voltage builds up over a temperature gradient and degradation in thermocouple response also is highly dependent on the temperature gradient, thermocouple drift cannot be calibrated out without locating the thermocouple in the same temperature gradient where it is to be used.

1.4 Resistance Thermometers

Resistance thermometers function by measuring the change in electrical resistance of a transducer with temperature. Both metals and semiconductors are commonly used as resistance thermometer elements. The conventional conceptual models for electrical conduction in metallic conductors and semiconductors are substantially different. For a formal introduction to electrical conduction in solids, the reader is referred to introductory solid-state theory texts such as Kittel (1986). Quinn (1990, p. 208 *et seq.*) also provides a well-reasoned overview of conductivity variation with temperature.

1.4.1 Qualitative Sketch of Electrical-Conductivity Variance with Temperature in Metals The basic conceptual model for electrical conductivity in metals is that of free-electron waves propagating through a periodic lattice. Conductivity is determined by how frequently the electrons collide, the energetics of those collisions, and how rapidly the electrons accelerate back to drift velocity following collisions. The electron collisions can be separated into collisions with lattice phonons, with point defects in the lattice, and with other electrons. Collisions with lattice imperfections are largely independent of temperature.

At room temperatures, the mean free path for electron–electron collisions is at least an order of magnitude more than that of electron–phonon collisions, and so electron–phonon collisions are the dominant source of electrical resistance temperature variation in metals near room temperature. Near room temperature the concentration of thermal lattice phonons is proportional to temperature. Hence metal electrical resistivity is roughly proportional to temperature near room temperature.

At very low temperatures (liquid helium

range) thermal phonon scattering in metals competes with electron–electron scattering, and lattice imperfection scattering. Bloch obtained an analytic result for electron–phonon scattering at low temperatures in metal that resistivity varies as T^5 (T is absolute temperature). At these temperatures, electron–electron collision mean free paths can be several centimeters (inter-conduction-electron spacing is ~2 Å). These large free paths [due mostly to the low thermal energy of the conduction electrons and the Pauli requirement to have unfilled orbitals with energies suitable for in-scattering (Kittel, 1986, p. 280)] result in much greater conductivity for pure metals at low temperatures as compared to room temperature (conductivity ratios of 10^5 are common). For transition metals (those with partially filled *d* orbitals such as platinum and nickel), the situation is further complicated by large-angle electron–electron scattering. This effect becomes dominant below about 10 K in transition metals, producing a T^2 dependence in electrical resistivity.

Transition metals, because of their partially filled band structure, have significant variations in the density of electron states that are thermally populable. This is generally expressed as a further T^2 modification (which can be positive or negative) in the simple T dependence of the electrical resistance of metals near room temperature.

Above room temperature, the electrical resistance is modified by an additional T^2 term due to the thermal expansion of the lattice. The expansion causes the interatomic spacing to increase, the stiffness of the interatomic bonding therefore to decrease, and the amplitude of atomic vibration to increase, which, in turn, causes atomic nuclei to become more effective electron-scattering centers.

At still higher temperatures, lattice vacancies begin to increase as thermal vibrations become comparable to lattice bonding energies, resulting in an exponential increase in electrical resistivity with the exponent value given by the ratio of the energy required to form a vacancy (~1.5 eV for platinum) and kT (k is Boltzmann's constant).

1.4.2 Qualitative Sketch of Electrical-Conductivity Variance with Temperature in Semiconductors The distinction between semiconductors and conductors is that the allowed electron energy bands of semiconductors exhibit a gap, which can be thermally bridged at temperatures below lattice atom migration temperatures. The semiconductor upper energy band (unpopulated at low temperatures) is called the conduction band while the lower energy band (completely filled at low temperatures) is the valence band. In semiconductors, the electron energy level that has a 50% occupancy probability (the Fermi level) resides within the energy band gap at low temperatures (see Fig. 2). In contrast, conductor Fermi levels lie within an energy band at all temperatures. For more complete introduction to semiconductor solids and their thermoelectrical properties see, for example, Streetman (1980) or Kittel (1986).

Electrical conduction in semiconductors takes place by a combination of migration of electrons in the conduction band and countermigration of electron vacancies, termed holes, in the valence band. Electrical resistivity of semiconductors is strongly dependent on the electron population of the conduction band and the hole population of the valence band.

The semiconductors used in temperature measurements are doped. That is, they have impurity atoms substitutionally included within the lattice. These impurity atoms either have an additional valence electron as compared to the bulk lattice, which is available to donate to the conduction process, or have one fewer valence electron and can serve as a sink for lattice electrons in the valence band. Conceptually these dopants either raise (in the case of donor atoms) or lower (in the case of acceptor atoms) the lattice Fermi level such that much lower temperatures are required to place electrons in the conduction band or holes in the valence band.

Four major conduction temperature regimes can be identified in semiconductors. In the highest temperature regime, the conductivity is dominated by electrons thermally excited from the valence band to the conduction band. Here the resistance falls off roughly exponentially with increasing temperature. In the next lower temperature region, all of the donor and acceptor atoms have been ionized, but valence- to conduction-band thermal excitation occurs rarely. In this region, the conductivity depends on

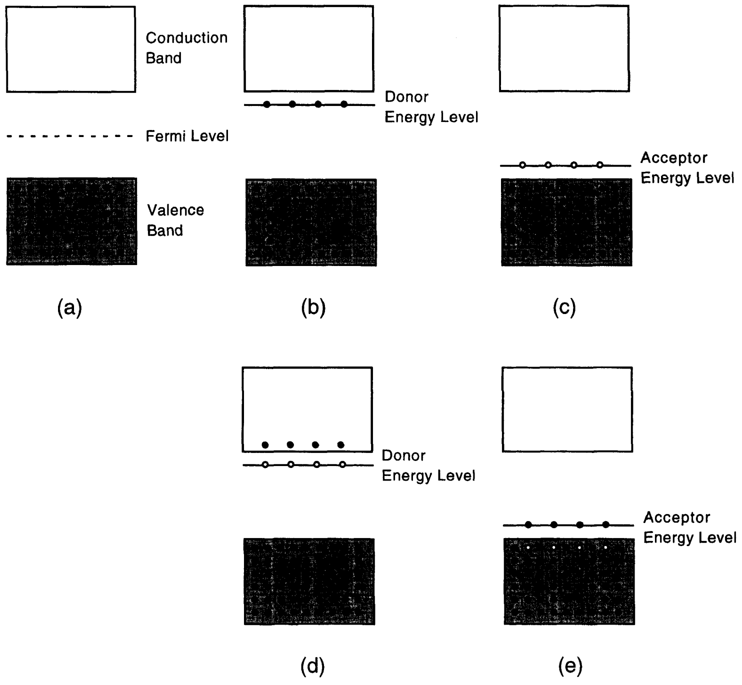

FIG. 2. Semiconductor band population diagram. (a) Semiconductor band structure, (b) donor energy-level population at 0 K, (c) acceptor energy-level population at 0 K, (d) donor energy-level population at 50 K, and (e) acceptor energy-level population at 50 K.

temperature only through the mobility of the electrons and holes, which decreases (because of increased scattering) with increasing temperature. Hence the resistivity has a positive temperature coefficient. In the next lower temperature region, ionization of the donor and acceptor atoms is only partial and conductivity is dominated by the ionization fraction of the dopant atoms. In this region, the resistivity of the semiconductor falls roughly exponentially with increasing temperature as the dopant atoms are thermally ionized. In the lowest temperature region, conduction takes place by electrons jumping from one dopant atom to the next. This is known as impurity conduction, and in this temperature region conductivity exponentially increases with increasing temperature.

1.4.3 Resistance Temperature Detectors Platinum resistance temperature detectors (RTDs) are among the most common of all general-purpose temperature sensors. Platinum resistance elements are commonly used from ~0.5 K (alloyed with cobalt) to ~1400 K. Several different versions of platinum RTDs are available for different applications. One limitation of metal RTDs is that the thermal expansion coefficient of metals is significantly greater than that of available insulation support structures. Therefore, for the most accurate implementations, to avoid strain-induced resistance changes minimally supported platinum wires are used. For industrial-grade applications, both thick films and thin films of platinum deposited on insulator substrates are commonly employed. Wire-wound platinum elements fully supported by powdered insulation are also commonly used industrially (see Fig. 3). Capsule-type platinum RTDs are used to make highly accurate temperature measurements from ~10 to ~500 K. Long-stem platinum RTDs are used for temperatures from ~50 to ~1400 K. Platinum RTDs are available with ice-point resistance ranging from ~0.25 Ω to ~5 kΩ with the dominant fraction having ice-point resistances of either 25 or 100 Ω.

1.4.4 Thermistors Thermistors (literally temperature-sensitive resistors) are the semiconductor form of resistance temperature detectors. One type of thermistor is the doped monocrystalline germanium resistor, which is used for accurate temperature measurements from ~1 to 30 K. A large class of thermistors are composed of mixed oxides of transition metals such as titanium, iron,

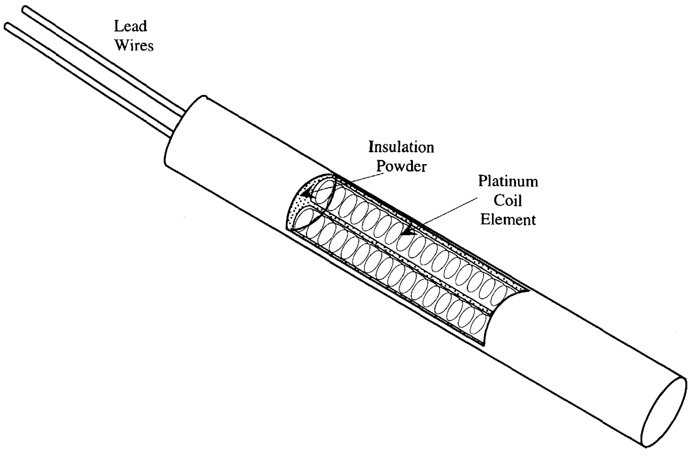

FIG. 3. Ceramic-insulated wire-wound industrial platinum resistance thermometer element, cutaway view.

nickel, copper, and cobalt. This composition of thermistor is referred to as a negative–temperature-coefficient (NTC) thermistor because of its resistivity decrease with increasing temperature. NTC thermistors are most commonly formed by mixing powdered oxide with a liquid binder to form a slurry that is applied to parallel wire strands. The assembly is then dried and the oxide sintered in a controlled atmosphere. The resultant beads are then overcoated with glass or epoxy for environmental protection. Chip-type thermistors (fabricated using a tape-casting process) are becoming more common as thermistors are implemented in surface-mount form for circuit-board temperature monitoring.

Positive–temperature-coefficient thermistors are most commonly made from perovskite-type oxides of barium and strontium titanate. Above the Curie point of the composite material, the resistance of the sensor rapidly rises with temperature over a few-degree temperature range (typically more than 100%/°C). The anomalous resistance-to-temperature relationship of these materials has been explained (Sachse, 1975) on the basis of the sensitivity of the activation energy (amount of energy required to cause a dopant atom to release its charge carrier) of impurity centers in the semiconductor to the material relative dielectric constant. The relative dielectric constant of most semiconductor materials is roughly constant with temperature. However, in ferroelectric materials, such as $BaTiO_3$, the relative dielectric constant varies strongly with temperature near their Curie temperature. An additional feature of this conductivity–temperature relationship is that the steep increase in resistivity with temperature is not a bulk material property but is caused by barrier layers developed at the grain boundaries. The electric barrier results from surface states acting as trapping locations for the conduction electrons released by dissociating dopant atoms.

The basic characteristics of thermistors is that they are very small and vibration resistant, and have a very large resistance change over a small temperature window. Mixed-oxide type thermistors are commonly used for high-precision temperature measurement over a limited temperature range such as in medical applications, while the positive–temperature-coefficient thermistors are most commonly used as temperature limit switches. Mixed-oxide thermistors are most commonly used with resistances between 1 MΩ and 100 Ω. On account of variations in the manufacturing process, mixed-oxide form thermistors often exhibit a significant resistance offset and are not yet as easily interchangeable as RTD elements.

1.5 Radiation Thermometers

1.5.1 Emission Physics Radiation thermometers function by measuring some fraction of the radiant energy emitted by a body as a consequence of its temperature. The amount and wavelength of radiation emitted (spectral radiance) by a blackbody (an ideal object that absorbs all incident radiation) was first derived by Planck from quantum mechanics (Liboff, 1980):

$$L_b(\lambda,T) = \frac{2hc^2}{\lambda^5[\exp(hc/\lambda kT) - 1]}, \tag{2}$$

where h is Planck's constant, k is Boltzmann's constant, T is absolute temperature, c is the speed of light in a vacuum, λ represents wavelength, and the subscript b on L denotes blackbody conditions. Several key points about thermal radiation are provided by Planck's law. For all wavelengths, the amount of radiation emitted increases with temperature. As temperature increases, relatively more radiation is emitted at progressively shorter wavelengths. Integrating Planck's law over wavelength and all possible directions yields the total *exitance* (M),

$$M_b = \frac{2k^4\pi^5}{15c^2h^3}T^4 = \sigma T^4. \tag{3}$$

This formula is known as the Stefan–Boltzmann law, and it underlies all of radiation thermometry. Planck's and Stefan–Boltzmann's laws apply without modification only to ideal blackbodies.

The surface radiative property *emissivity* (ϵ) is defined as the ratio of the radiation emitted by the surface to the radiation emitted by a blackbody at the same wavelength in the same direction. In general, the emissivity depends on wavelength, temperature, surface conditions (roughness, fouling, oxidation), angle of emission, and state of polarization. The largest uncertainty in most radiation thermometry applications is due to uncertainty in the value for the emissivity. For a detailed description of the theory and practice of radiation thermometry, the interested reader is referred to DeWitt and Nutter (1988).

1.5.2 Types of Radiation Thermometers Total-radiation thermometers measure radiant energy over as wide as possible a wave band. These devices are typically used in general-purpose, low-temperature instruments. Single–wave-band thermometers measure the incident radiation in a single narrow wave band. The wave band measured is selected so that the rate of change of radiant energy with temperature is as high as possible (2–3%/°C) in the designed operating range (655 nm is a common wavelength choice). Ratio thermometers work on the principle of comparing the energy in multiple wave bands. Their advantage is that their response is independent of target size, optical path, and emissivity (provided that the optical absorption and emissivity are wavelength independent).

1.5.3 Radiation Detectors Radiation detectors are classified as thermal or photon detectors. Silicon and germanium photodiodes are examples of photon-sensitive detectors. They directly convert incident photons within their response wave bands to electrical signals. Thermal detectors are designed to collect all incident radiation and respond uniformly to all incident wavelengths. A thermal detector's sensor temperature is designed to rise until an equilibrium is achieved with heat losses to the surroundings. The sensor temperature is then measured with a thermopile, resistance thermometer, or pyroelectric detector. A thermopile is a series connection of thermocouples with their sensing junctions receiving the incoming radiation while the cold junctions are shielded.

Pyroelectric detectors are composed of ferroelectric ceramics. Each grain in the crystal behaves as a miniature electric dipole. The dipoles are oriented in the sensor, giving the crystal a residual polarization (and surface charge). The degree of polarization (and the surface charge) decreases with increasing temperature as increased thermal energy randomizes the atomic oscillations. This decreased surface charge results in a decrease in the potential developed in a capacitor placed along the polarization direction of the crystal. Connection to any measurement circuit allows a flow of charge that ultimately neutralizes the polarization effect. Thus pyroelectric detectors cannot directly measure steady radiation. Beam choppers, which periodically block the incident radiation, are employed to overcome this limitation (upper frequency limit typically <1 MHz) and allow the measurement of the rate of temperature rise in a pyroelectric detector.

1.6 Special-Application Thermometers

1.6.1 Semiconductor-Junction (Diode) Thermometers Semiconductor-junction–based temperature sensors function on the basis of the temperature sensitivity of the voltage–current characteristics of *pn* semiconductor junctions. This type of sensor can be monolithically integrated with other electronic components and has consequently won widespread acceptance in the area of instrument temperature monitoring. Silicon-based diode temperature sensors function from about 1 to 525 K. Forward-biased diode characteristics are such that practically no current flows below the voltage necessary for charge carriers to overcome the junction potential. Above this voltage, the current increases approximately exponentially with increasing voltage-to-temperature ratios. Diode-based temperature sensors are often implemented using a constant-current source connected in series with the diode, since in this configuration the forward voltage varies approximately linearly (~ -2 mV/K) with increasing temperature ($T > \sim 20$ K).

1.6.2 Liquid-in-Glass Thermometers Liquid-in-glass thermometers function on the basis of the differential thermal expansion of the liquid and glass container material (see Fig. 4) and are typically optimized to measure the temperature of a fluid. Different styles of liquid-in-glass thermometers are available based upon the immersion depth into the fluid. *Partial immersion* thermometers are inserted to a fixed point marked on the stem. This is the least accurate implementation, since the liquid and glass temperatures above the surface may be significantly different from that of the immersed portion. This affects their differential thermal expansion and consequently directly affects the measured value. *Total immersion* thermometers are designed to be immersed until the liquid in the column is totally immersed in the fluid. *Complete immersion* thermometers are designed for use completely submerged in the fluid.

The dominant cause of shifts in thermometer response are variations in the glass bulb volume due to glass structural changes. The major requirement for good thermometric glass is that its shape and rigidity remain stable on heating and with thermal cycling. Since glass is not a crystal structure, plastic flow can occur with heating. For accurate temperature measurement, the effects of pressure on liquid volume must also be accounted for. The volume of the glass bulb is equivalent to several thousand degrees temperature shift in the stem; thus even small changes in the bulb volume affects the temperature reading (0.1 °C per atmosphere is a typical correction factor).

1.6.3 Resonant-Frequency Thermometers Quartz crystals form the sensing elements of most resonant-frequency thermometers. Ricolfi and Scholtz (1990, p. 270 *et seq.*) give a thorough review of the operation of quartz resonator temperature sensors. Quartz crystals are piezoelectric—mechanical deformations in the crystal cause a change in the electric polarization. As the crystal polarization changes, the surface charge density changes—this allows for easy excitation and detection of its vibrations. Different crystal resonant frequencies are produced depending on the orientation of the crystallographic axes and the vibration. Along some of the crystallographic axes, the variation in resonance frequency with temperature is very small (for temperatures less than 250 °C). Along other axes, the resonant frequency varies strongly with temperature (up to 90 ppm/K). Quartz thermometers are often realized by beating the resonant-frequency signals obtained along different crystallographic axes and monitoring the resultant heterodyne signal.

1.6.4 Acoustic Thermometers Acoustic thermometers function by measuring the propagation time of sound waves through a

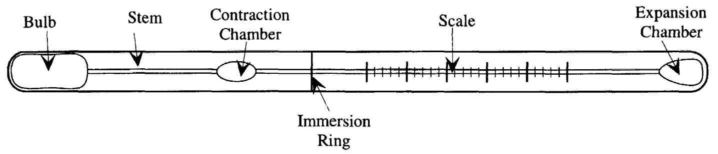

FIG. 4. Liquid-in-glass thermometer principle features. The immersion line is provided only on partial-immersion thermometers. The expansion chamber is provided to prevent excess pressure from being developed in gas-filled thermometers when the thermometer is used near the top of its range.

medium. Lynnworth (1989) provides more details about the operation of acoustic thermometers. Two classes of acoustic thermometers are conventionally specified—those in which the sound is transmitted through a separate element in the monitored environment and those in which the environment itself propagates the sound. Acoustic thermometers are available that are designed to function with gaseous, liquid, and solid media. The velocity of sound in a medium is commonly measured by one of three techniques. The velocity can be determined through simple geometry by measuring the times of arrival of sound emitted from a single source at two different receivers at known locations. Alternatively, a fixed-frequency wave can be launched from a single source. The phase relationship of the source signal and that recorded at the sensor can be used to determine wave velocity. The resonant frequencies of acoustic resonators (e.g., a flute) also depend on gas temperature. A typical application of media-as-sensor implementation is in measuring the flame temperature in power-plant type fireboxes. In these systems a series of acoustic transmitter/receivers rings the fireball. Each individual sound path gives an average temperature along the line. Using multiple lines through the fireball, a temperature profile can be constructed.

Foreign-medium–based acoustic thermometers function by measuring the time delay between the arrival of a reflected pulse received from a discontinuity in a thin wire waveguide and another discontinuity (see Fig. 5). The time of flight is determined by the length of the sensor between the discontinuities and the velocity of sound in the sensor material (thermal-expansion length changes are typically negligible).

1.6.5 Noise Thermometers Johnson noise is a fundamental representation of temperature—it is the vibration of the electronic field surrounding atoms as they thermally vibrate. Johnson-noise thermometry technology has recently been reviewed (White *et al.*, 1996). As an absolute measurement, Johnson-noise thermometry is inherently drift free. Moreover, Johnson-noise is insensitive to the material condition of the sensor and consequently is immune to contamination and thermomechanical response shifts.

The mean squared noise voltage, in a frequency interval df, appearing across the ends of any resistor is given by (Ricolfi and Scholz, 1990, p. 228)

$$d\overline{V^2} = \frac{4hf\,\mathrm{Re}(Z(f))df}{e^{hf/kT} - 1}, \tag{4}$$

where h is Planck's constant, k is Boltzmann's constant, Re(Z) is the real part of the complex impedance Z, and f is the frequency. An important consequence of this relationship is that Johnson noise provides a means of equating temperature to fundamental physical variables (length, time, and mass) through a voltage and an impedance measurement.

A wide variety of measurement circuits have been devised to minimize the effects of electronic noise and to avoid needing to measure the sensor resistance. Simultaneous measurement of the noise current, which is inversely proportional to Re(Z), and noise voltage allows cancellation of the resistance in the determination of noise power, which is a fundamental representation of temperature.

The major limitation of Johnson-noise thermometry lies in the very low levels of signals produced and the difficulty in acquir-

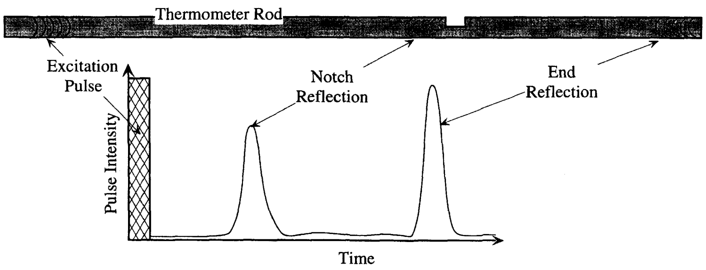

FIG. 5. Acoustic notch thermometer.

ing and processing them. Both radiatively coupled electronic noise and vibrationally induced noise degrade thermally produced noise's relative signal level. For a 300-K measurement using a 100-Ω resistor and a 100-kHz frequency band, the root of the mean squared noise voltage is approximately 4×10^{-7} V.

1.6.6 Fiber-Optic–Coupled Phosphor Thermometers The sensing element in a major class of phosphor thermometers is composed of rare-earth atoms substitutionally included in a refractory lattice (Y_2O_3:Eu or LaF_3:Gd, for example). A more detailed description of phosphor thermometry is given in Grattan and Zhang (1995). The sequence of events underlying rare-earth phosphor thermometry is as follows. First, electrons in the rare-earth atom are excited out of their ground state. Since rare-earth atoms have incomplete, widely spaced inner-electron shells, some of the ground-state energy inner electrons can be pumped into an excited state ($4f \rightarrow 5d$ transition) using UV/deep blue light. Following this, the electrons deexcite either radiatively or nonradiatively. The radiative transition probability is governed by the quantum-mechanical selection rules for inner-electron transitions and as such is essentially independent of temperature. In contrast, nonradiative transitions are mediated by phonons. As such they are governed by the phonon density (a fundamental temperature indication) and the allowed optical-phonon modes (function of the short-range lattice order). Since only radiative transitions can be easily monitored, the effect of the increasing probability of nonradiative transitions with increasing phonon population (temperature) is to create an apparent radiative decay-time variation with temperature.

The variation in fluorescence decay time can be exploited to determine temperature by monitoring the time relationship between the exciting energy and the emitted fluorescence. This is typically done either by exciting the phosphor with a pulsed laser and monitoring the exponential phosphor decay lifetime, or by sinusoidally exciting the phosphor and monitoring the phase lag of the resultant fluorescence.

1.6.7 Bimetallic Thermometers A bimetallic temperature sensor takes advantage of the differential coefficients of thermal expansion between different metals. When two metal strips are bonded together and heated, the differential expansion with temperature will cause the joined strips to bend. The most common implementation of this effect is in household thermostat temperature indicators. In these devices, a bimetallic strip is wound into a spiral, one end of which is fixed and the other attached to a temperature-control switch and an indicator needle that points to a temperature scale.

1.6.8 Vapor-Pressure Thermometers Constant-volume vapor-pressure thermometry relies upon the change in vapor pressure of a gas with temperature as the basis for measurement. The ideal-gas law relates temperature, pressure, and volume for ideal gases. Corrections to the ideal-gas law are required for use with nonideal gases at common temperatures and pressures. Constant-volume vapor-pressure thermometers can be used as absolute thermometers at low pressures. Since gas performance only becomes truly ideal at zero pressure, a set of measurements is typically made while reducing gas pressure to allow extrapolation to zero pressure. Constant-volume gas thermometers used in this fashion have been used in defining the thermodynamic temperature scale.

1.6.9 Fizeau-Cavity Thermometers Fizeau-cavity–based thermometers are another optically based temperature measurement scheme (Fig. 6). Fizeau cavities are a low-finesse (sensitivity) implementation of Fabry-Pérot etalons, which are described in Sec. 2.8. The transmittance and reflectance of Fizeau cavities are sensitive to the spacing of the reflectors and the wavelength of the incident light. Thus, thermal expansion of the material between the reflectors yields an optical signal that is a function of temperature (the "expansion" is typically dominated by an optical expansion due to change in index of refraction as opposed to a change in physical size). To avoid signal variation due to nonthermal effects, multiple-wavelength ratiometric techniques are incorporated into the signal detection and processing.

1.6.10 Distributed Fiber-Optic Thermometers One form of distributed fiber-optic temperature measurement relies on temperature-dependent change in the relative amounts of Stokes and anti-Stokes scattering

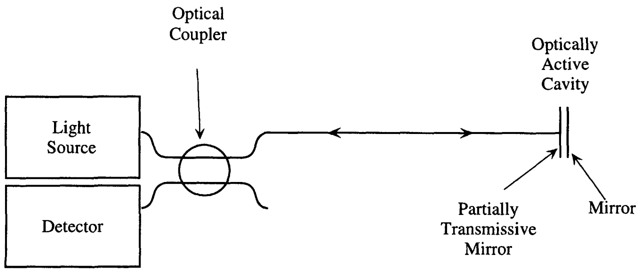
Optical Coupler
Optically Active Cavity
Light Source
Detector
Partially Transmissive Mirror
Mirror

(Mirror and cavity dimensions have been exaggerated for clarity)

FIG. 6. Fizeau-cavity–based temperature sensor.

from an incident light pulse (Udd, 1991). Stokes and anti-Stokes scattering are forms of Raman scattering in which the incident photon interacts with a phonon of the material lattice in which the photon is propagating (Fig. 7). In Stokes scattering, the propagating photon interacts with the material lattice, emitting a phonon and downshifting in energy. In anti-Stokes scattering the photon absorbs a phonon and is thereby upshifted in energy. The number of phonons in the material (and hence the ratio of Stokes to anti-Stokes scattering) is a function of temperature. This form of sensing is implemented by injecting a narrow–wavelength-band optical pulse into a fiber at a known time and measuring the subsequent Stokes and anti-Stokes light scattered back down the fiber. Since both the time of injection of the optical pulse and the index of refraction

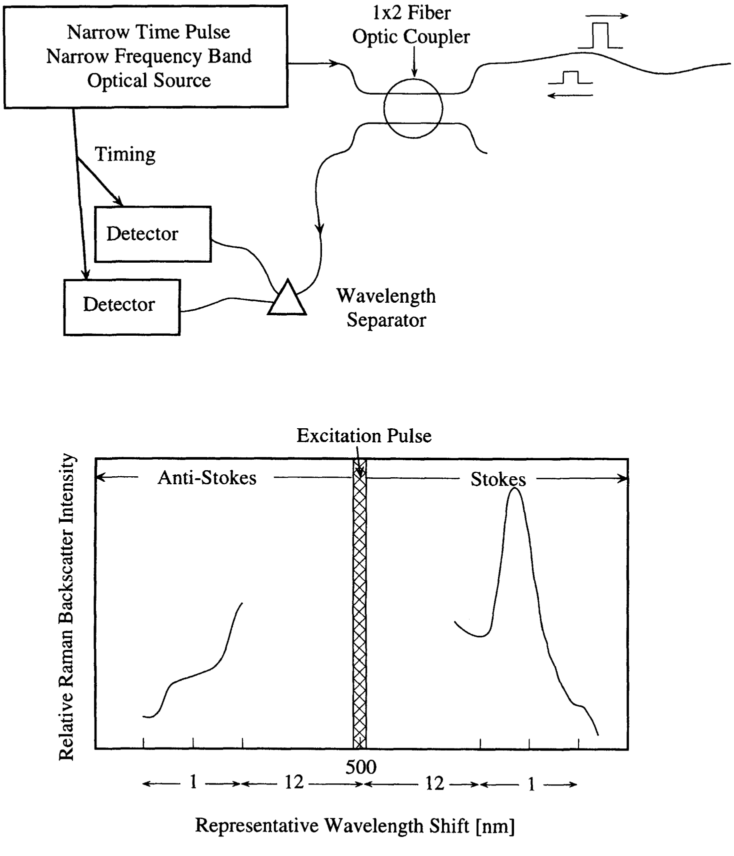
Narrow Time Pulse Narrow Frequency Band Optical Source
1x2 Fiber Optic Coupler
Timing
Detector
Detector
Wavelength Separator
Excitation Pulse
Anti-Stokes
Stokes
Relative Raman Backscatter Intensity
500
1
12
12
1
Representative Wavelength Shift [nm]

FIG. 7. Raman-scattering–based distributed temperature sensor. The plot displays a typical reflected Stokes and anti-Stokes wavelength shift and pulse shape for a narrow-band (500 nm) excitation pulse (using a nonlinear abscissa).

of the optical fiber are known, the time at which the backscattered light arrives at the fiber end is a measure of the location of the scattering. Correspondingly, the intensity ratio of the Stokes to anti-Stokes component of the backscattering is a measure of the temperature. The technique of obtaining spatial information about reflections from the time interval between pulse injection and reflection return is referred to as *optical time-domain reflectometry* (OTDR).

Other forms of OTDR-based distributed fiber-optic temperature have also been demonstrated. In one system a conventional fiber-optic core is coated with a polymer cladding containing scattering centers. The polymer cladding has a temperature-dependent refractive index. Therefore, both the optical power present in the cladding and the amount of backscattered light vary with temperature. This sensor has been reported to be able to pick out hot spots of <10 cm in length and to have an accuracy of ±2 °C from 0 to 60 °C and ±5 °C from 60 to 150 °C (Boiarski and McGinniss, 1993).

2. GENERAL CONCEPTS OF PRESSURE AND PRESSURE INSTRUMENTATION

2.1 Physical Nature of Pressure

Pressure is the force exerted normal to a unit surface area,

$$P = \frac{dF}{dA}. \tag{5}$$

At a given position, pressure is directionally independent and is also unaffected by the geometries of the surfaces upon which it acts. For gases, pressure may be evaluated as a measure of the total kinetic energy of the molecules, as presented from the ideal-gas law (Fraden, 1997)

$$P = 2U/3V = NRT, \tag{6}$$

where U is the kinetic energy of the gas, and P, V, N, R, and T are pressure, volume, number of molecules per unit volume, specific gas constant, and absolute temperature, respectively.

Pressure measurement is referred to as either absolute, gauge, or differential. Absolute pressure indicates pressure referenced to total vacuum, as opposed to gauge, which references to ambient pressure conditions. Differential pressure is, as the name implies, a differential measurement of pressure, often used to quantify upstream and downstream flow conditions, or a pressure difference between control volumes. Common pressure units include the N/m^2, or pascal (Pa), pounds per square inch (PSI), bar (10^6 dyn/cm^2), and various gravimetric bases: mm mercury (Torr), inches mercury, and inches of water column.

2.2 Manometers

Manometers indicate pressure through the balancing of forces on a fluid column within a fixed cavity. The most common manometer is the mercury barometer, consisting of a column of mercury contained in a tube, sealed at one end and open to an exposed reservoir at the other. The ambient pressure P upon the open reservoir imparts a distributed force that displaces the mercury column to a height h at which the weight of the column balances the pressure on the reservoir:

$$P = \rho g h. \tag{7}$$

The measurement is dependent upon the density of the mercury, ρ, and upon the local gravity constant g. The sealed end of the mercury barometer creates a vacuum reference condition making the barometer an absolute pressure measurement. Gauge pressures may be measured in the same manner, with the top of the liquid column connected to a constant reference pressure. Under this condition, the force balance upon the liquid column is the summation of the reference pressure at the top of the liquid column and the weight of the column against the process pressure at the reservoir. Multitube manometers perform the same balance, but with multiple process-pressure tubes connected in parallel to a reference pressure reservoir. Other manometer configurations include spring and inverted-bell manometers, in which the process-pressure force and weight of displaced liquid balance with spring loading and the weight of the bell housing, respectively.

The performance of a liquid-column manometer is influenced primarily by the density of the reservoir fluid. Range of use may be extended by using fluids of higher density, at the cost of resolution of the liquid-column displacement measurement. To improve resolution, the column may be inclined rather than vertical, such that a large horizontal component may be used to relate to the smaller vertical displacement of the fluid.

2.3 Elastic-Member Pressure Transducers

Elastic-member pressure transducers indicate pressure through elastic deformation of a transducer subjected to pressure-induced force. The resultant mechanical deformation may be used to measure pressure through mechanical amplification or may be converted into an electrical signal.

The C-Bourdon tube is a "C"- or horseshoe-shaped tube of elliptical cross section, sealed at one end, and open to process pressures at the other (Fig. 8). The pressure acting within the tube generates a uniform force distributed across the inner surface of the tube. Because of the longer length of and resultant additional force upon the convex outer tube wall, the internal pressure acts to straighten the entire assembly. A force balance results where the internal pressure forces equal the spring force of the elastically deformed C. A mechanical coupling to the closed end of the Bourdon tube may be used to indicate the internal process pressure. The nonlinearity of the system, resulting from the spring force of the C tube and its displacement, may also be addressed by implementing an additional force-balancing component onto the structure. Typically, a variable force member is coupled to the sealed end of the tube, restricting the motion of the C and reducing the nonlinear spring force of the system.

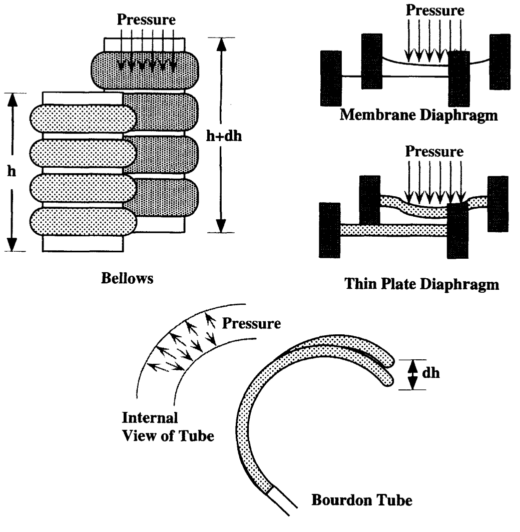

FIG. 8. Traditional pressure measurements rely on the elastic deformation of a mechanical member, such as a diaphragm, bellows, or Bourdon-tube element.

Other implementations of the Bourdon-tube technique include the spiral and helical Bourdon elements. The former is a series connection of many sequential C tubes, in a spiral configuration with only the final element sealed off; the latter is simply the helical analog. As a result of the longer internal surface area, these tubes are more sensitive and have higher resolution. They may thus be used for much lower process pressures.

Diaphragm sensors indicate pressure through elastic deformation of a diaphragm, which may be in the form of a thin membrane, a plate, or a capsule. The deflection y of the plate diaphragm is dependent upon the pressure P, the thickness of the plate t, the radial position r with respect to the radius a, and the Poisson's ratio ν and Young's modulus E of the material:

$$y = \frac{3}{16} P \frac{(1 - \nu^2)}{Et^3} (a^2 - r^2)^2. \tag{8}$$

The diaphragm sensor may be used in an absolute-pressure mode, where one side of the diaphragm is evacuated, or the sensor may be used to indicate the differential pressure between sides of the element. Sensor sensitivity is increased by increasing the area of the diaphragm, providing a larger surface area upon which the pressure may exert force, and through selection of materials. Nonlinear deformation characteristics may be minimized by careful shaping of the diaphragm element, including *corrugation* of the surface. A force-balancing technique may also be used to alleviate nonlinearity by restricting deformation of the diaphragm with an alternative balancing element, such as a force bar.

The bellows sensor also indicates pressure through elastic deformation of a closed vessel, behaving as a series of interconnected capsule diaphragm elements. The bellows chamber inflates under pressure, with design features that cause preferential deformation in the axial direction of the bellows. The length of axial deformation indicates pressure, or the chamber may be force-balanced with an additional mechanical member to reduce nonlinearity. Increasing the overall length of the bellows increases the sensitivity of the measurement.

2.4 Piezoresistive Transducers

Measurement of deformation of a diaphragm, or other elastic elements of a pressure-sensing system, is traditionally performed using piezoresistors, materials with bulk resistivity that is dependent upon applied mechanical stress (Bau *et al.*, 1994, pp. 146–172). The bulk resistivity, i.e., the number of charge carriers and the mobility of these carriers, within a piezoresistor is influenced by the extent of lattice strain. This strain may result from unbonded contact with a pressure element, such as a piezoresistive wire stretched between the diaphragm and a fixed wall of the sensor; or bonded, where an entire surface of the piezoresistive element experiences the pressure-induced mechanical stress. The latter case is often employed in silicon-based sensor technology, as silicon provides a strong substrate, with high Young's modulus, and piezoresistive properties dependent upon impurity dopant levels. The effects of temperature-dependent resistance changes, a common concern for all piezoresistive transducers, may also be mitigated in silicon by employing extremely small feature sizes, increasing transducer isothermality and facilitating temperature compensation. Also, the silicon dopant levels may be tailored to decrease overall temperature sensitivity.

2.5 Linear-Variable Differential-Transformer (LVDT) Transducers

LVDTs are commonly employed to measure the displacement of an elastically deflected pressure member. Used predominantly for linear displacement measurements, the LVDT indicates the displacement of a transformer core coupled to the deformed pressure element. The transformer features dual secondary windings that are connected in series but are wound to produce opposing voltages. Under null displacement conditions, when the core is centered within the transformer windings, a zero indication results. Repositioning of the core to either side of the null position, however, creates a voltage output that is linearly proportional to the displacement.

2.6 Capacitive Deformation Transducers

Displacement of an elastically deformed member may also be measured capacitively. A capacitive diaphragm assembly is the simplest of capacitive-type pressure members, where the diaphragm *itself* acts as one plate of either a dual- or a multiplate capacitor. Capacitance between the plates is dependent upon the deflection y, the distance between the plates, d, the plate cross-sectional area A, the permittivity of free space, ϵ_0, and the dielectric constant of the material between the plates, ϵ_r:

$$C = \epsilon_r \epsilon_0 A/(d - y). \qquad (9)$$

In multiplate capacitive diaphragms, the gap between plates may be filled with gas, for high diaphragm displacement and sensitivity. A solid or liquid dielectric material may also be used to minimize dielectric spatial variability and to allow higher rangeability of the measurement by providing stiffness to the deformable member.

2.7 Piezoelectric Transducers

Piezoelectrics are materials that mechanically deform when subjected to voltage and, conversely, generate a surface charge differential when subjected to stress through electrical polarization. The piezoelectric effect is dependent upon lattice geometry, and the orientation of applied stress. For pressure sensing, the transverse piezoelectric effect, in which voltage is generated in a direction transverse to the strain, is generally used. The magnitude of voltage response U of a strained, $\Delta l/l$, piezoelectric is governed by the transverse piezoelectric modulus d, the mechanical modulus along the axis of deformation s^E (for a constant electric field E), the transverse thickness t, and the transverse permittivity of the material ϵ^T (for a constant stress, T). Bau *et al.* (1994, pp. 616–619) give a more detailed description of piezoelectric voltage generation:

$$U = \frac{d}{d^2 - s^E \epsilon^T} t \frac{\Delta l}{l} . \qquad (10)$$

Piezoelectrics typically feature high-frequency response and are often selected for dynamic pressure measurement and generation. On account of their high-frequency response, environmental vibration isolation or compensation of the piezo crystal may be necessary for improved signal-to-noise ratio. Temperature compensation is also often required, because of temperature sensitivity of the piezoelectric materials. As the piezoeffect is dependent on crystal geometry and the orientation of loading, crystal geometry can often be tailored to decrease temperature sensitivity of the mechanical system.

2.8 Optical Deformation Measurement

A variety of optical techniques may be applied for the measurement of elastic member deformation. Both Udd (1991) and Krohn (1992) provide extensive material on optical implementation of pressure measurements. Optical reflectance may be used to measure the deformation of a reflective diaphragm using a fixed-position source and detector. In this system, pressure deforms the diaphragm altering the amount of reflected light coupled back to the detector. The variation in reflected light intensity at the detector is used to indicate pressure.

Various types of optical interferometers may also be employed to measure the displacement of an elastic member. These techniques rely upon the linear superposition of electromagnetic waves, typically coherent light, experienced when the incident wave form interacts with reflected wave fronts from a surface. A Fabry-Pérot etalon is a planar reflective interferometric device conducive to micromachined construction of pressure elements. The etalon is comprised of two planar, parallel reflectors with transmission and reflectivity dependent upon the wavelength of incident radiation and the separation of the mirror surfaces. Elastic deformation of one, or both, surface due to pressure will alter the mirror separation, changing the transmission of the overall cavity. A change in reflected, or transmitted, intensity may thus be used to provide a measurement of pressure.

2.9 Resonance Pressure Transducers

In contrast to the aforementioned deflection-based measurement techniques, resonance transducers measure pressure by observing the pressure-dependent resonant frequency of a vibrating mechanical member. The pressure dependence of these structures, which may be micromachined cantilevers or conventionally scaled elements such as Bourdon tubes, may be evaluated by viewing the structure as a spring-mass-damper system. The effect of each of these components upon the pressure-dependent resonance differs depending on the measurement conditions. Under near-ambient pressure conditions within a gaseous environment, the pressure dependence of the resonant element is dominated by the *mass* component of this system. Oscillatory motion of the resonant structure will carry with it a small, additional mass of surrounding air in consequence of the no-slip condition resultant from fluid viscosity (see Sec. 3.2). Over a limited range and provided the amount of additional mass is much less than the mass of the resonant structure, the resonant frequency will be linearly proportional to the pressure of the system, through the resultant fluid density (Bau *et al.*, 1994, p. 223). Under vacuum conditions, the gaseous environment becomes rarefied, resulting in a longer mean free path for fluid molecular motion. Within this regime, resonant frequency is dominated by pressure dependence upon molecular momentum transfer. Collisions with the resonant element are governed by the Maxwell–Boltzmann velocity distribution, resulting in a linear dependence upon pressure (Bau *et al.*, 1994, p. 224). Under high pressures, and within liquid media, fluid viscosity dominates the pressure dependence of resonant elements, altering system response through its *damping* component. Regardless of the pressure-dependence mechanism of a resonant structure, temperature will also have a considerable impact on the structure's resonant frequency. Temperature dependence is dominated by materials issues, i.e., modulus of elasticity, resulting in changes in the spring component of the system.

2.10 Hall-Effect Transducers

The Hall effect describes the generation of an electric field, $\mathbf{E} = \mathbf{j} \times \mathbf{B}$, across a conductor due to the drift velocity imparted to the electrons of a current **j** by a magnetic field **B** (Kittel, 1986). The magnitude of the electric field is dependent upon the magnitude of the current and magnetic field, and the Hall sensitivity of the material. The Hall sensitivity is dependent upon the charge mobility and resistivity of a material and has been studied for several group III–V semiconductors, in particular, gallium arsenide. As these latter properties are determined by lattice structure, Hall sensitivity is dependent upon pressure. The sensitivity of gallium arsenide has been shown to feature highly linear dependence upon pressure from ambient to approximately 70 MPa (Selim, 1995).

2.11 High-Pressure Ruby and Alexandrite Fluorescence

Extremely high-pressure measurement, >1 MPa, or measurements within harsh chemical environments cannot be performed using traditional elastic-member deformation techniques, because of materials limitations under the extreme conditions. The pressure-dependent fluorescence of some refractory phosphor materials, such as ruby and alexandrite, may be used to measure these intense pressures under extreme conditions. Short-wavelength (ultraviolet to visible) photons will cause these materials to fluoresce. The Stokes form of Raman scattering will cause photon emission at a longer wavelength. For example, ruby, when excited with green or blue light, will downshift the frequency and emit red light. Both the wavelength and persistence, or lifetime, of the fluorescent emission are strongly pressure and temperature dependent, as a result of altering transition probabilities within the crystal lattice. Over a fairly wide range (1 MPa–43 GPa), the lifetime and wavelength shift of the emission are both linearly dependent upon pressure (Grattan and Zhang, 1995). Compensation for temperature variation may be performed by simultaneously measuring fluorescence lifetime, which is highly temperature dependent (Hirschfeld and Haugen, 1988).

2.12 Low-Pressure Thermal-Conductivity Gauge

Below approximately 10^{-4} mm mercury, or Torr, direct indication of pressure through the mechanical deformation of an elastic member is not practical (O'Hanlon, 1989). Rarefaction of the fluid medium, and the resultant increase in mean-free-path length, decreases the occurrence of molecular collisions necessary for diaphragm deflection. This same decrease in collisions, however, may be employed for pressure detection by observing the resultant decrease in thermal conductivity of the fluid medium. Provided the mean free path is sufficiently large, such that the actual geometry of a device is much less than the mean free path, the loss of thermal energy from a heated wire within a rarefied volume will be strongly dependent upon the pressure (Fig. 9). This dependence results from the increased probability of

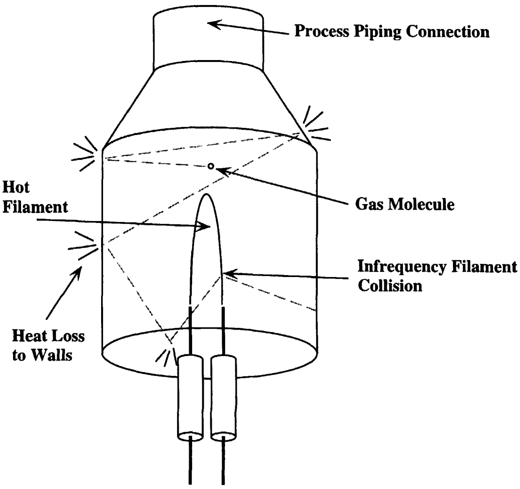

FIG. 9. Thermal conductivity vacuum gauges measure thermal losses from a heated filament caused by infrequent collisions of molecules within the rarefied atmosphere. Provided the mean free path of molecular motion is much greater than gauge dimensions, molecules achieve isothermality with wall boundaries, and the heat transfer is primarily dependent upon molecular density, and thus pressure.

molecules, acting as thermal carriers, interacting with volume boundaries as opposed to the heated wire. As the mean free path increases, fewer molecules will interact with the wire. Those that do will have a high probability of being at the boundary temperature, as the majority of collisions are with the solid boundary. Lower pressure limits exist with thermal-conductivity techniques because of radiative heat losses from the wire, and conductive losses through the wire and wire supports.

2.13 Low-Pressure Ionization Gauge

Lower vacuum pressure measurement may be performed by ionization gauges, which ionize gas molecules through electron impact and measure the particle density of the generated ionic current. Provided temperature remains constant within the fixed volume of the ion gauge, particle density is proportional to pressure through the ideal-gas law. Traditionally, ion gauges are classified as either hot- or cold-cathode. The former depends on thermionic emission of electrons from a hot filament that are accelerated by an electric potential to a velocity sufficient to ionize any gas molecules within their path. This collision produces positive ions, which are preferentially collected by a negatively biased electrode, allowing the evaluation of positive-ion density. The electrons are captured on the acceleration-inducing, positively charged grid. Below 10^{-6} Torr, electron collision with this grid generates x rays, which create a photocurrent that is also collected upon the negatively biased ion collector, resulting in aberrant readings (O'Hanlon, 1989). The Bayard–Alpert gauge (Fig. 10) addresses this problem by protecting a fine-wired positive-ion collector within the electron charge collector. The minimized cross-sectional area limits interaction of the positive-ion collector with emitted x rays from the electron-collection grid, and allows operation to 10^{-8} Torr.

The cold cathode employs a grounded cathode and a wire anode at a high bias voltage (thousands of volts). A permanent magnet of 0.1 to 0.2 T surrounds the sensor tube. The interaction of the magnetic fields with the electron current within the tube causes the electrons to follow an increased-path-length, spiral trajectory before colliding with the anode. The increased path length increases the probability of molecular impact during flight and enhances the ionization efficiency of the gauge.

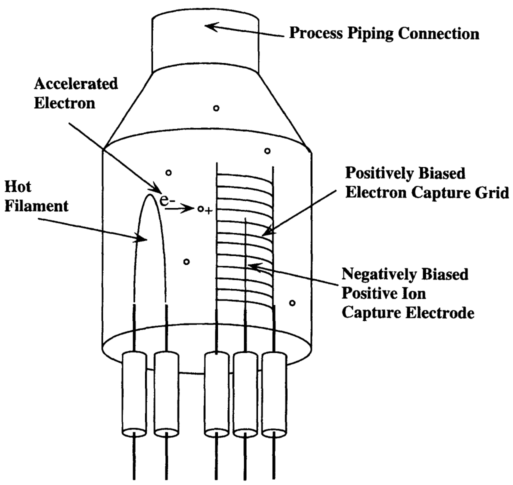
Process Piping Connection
Accelerated Electron
Hot Filament
e-
o+
Positively Biased Electron Capture Grid
Negatively Biased Positive Ion Capture Electrode

FIG. 10. The Bayard–Alpert hot-cathode gauge accelerates thermionic electrons emitted from a heated filament into rarefied gas molecules, creating positive-ion flow. Preferential positive-ion collection is performed by a thin, negatively biased electrode, minimizing the effects of x-ray generation at the electron-capture grid.

3. GENERAL CONCEPTS OF FLOW AND FLOW INSTRUMENTATION

3.1 Nature of Flow and the Conservation Laws

Fluid flow processes are most commonly described using laws of conservation of mass and energy, wherein the total mass and energy within a closed system must remain constant. Within this light, fluid systems are easy to visualize, as they are typically well-bounded control volumes, with discrete input and output features, i.e., closed systems such as pipes and ducts, or bounded systems such as weirs and flumes. Mathematically, the conservation of mass, or the *continuity statement* as applied to fluid flow, may be expressed as

$$-\frac{\partial\rho}{\partial t} = \frac{\partial}{\partial x}(\rho v_x) + \frac{\partial}{\partial y}(\rho v_y) + \frac{\partial}{\partial z}(\rho v_z), \qquad (11)$$

where the rate of accumulation of mass in a system, or the change in density (ρ) with respect to time, must equal the total mass flux into the system, represented here in rectangular coordinate components as the partial differential of the product of the density and respective components of velocity **v**. For steady-state conditions, where the rate of mass accumulation into the system is zero, the continuity statement simplifies further to state that the mass flow into the system must equal the mass flow out. Many velocimetric styles of flowmeters apply this principle to derive total volumetric flow rate $\dot{m}/\rho$ by measuring velocity v through a known cross-sectional area A, where

$$\dot{m}_{\text{in}} = \dot{m}_{\text{out}} = \rho A v = \text{const.} \qquad (12)$$

Conservation of energy applies the same principles to total energy of the system, which is expressed in several *mechanical* forms for flow problems; kinetic, potential, and static energy. Kinetic energy is the energy due to velocity of the mass in the system. Potential energy is the energy of the mass in the system capable of performing work against gravity. Static energy is the energy capable of performing work against a pressure gradient. In the 18th century, Bernoulli expressed the law of conservation of energy with respect to these three terms respectively as

$$\bar{v}^2/2g + z + P/\rho = \text{const.}, \qquad (13)$$

where kinetic energy is represented as average flow velocity $\bar{v}$, potential energy as the height z of fluid above a reference plane acted upon by gravity g, and the static energy as the static pressure P of the system. The principles of Bernoulli's law have been applied to create a wide array of flowmeter styles, including orifice plates, venturis, pitot tubes, and both variable-area rotameters and spring-loaded apertures.

Bernoulli's formulation of the law of conservation of energy applies only to closed systems that perform no work. In reality, no flow system operates without consuming some energy through work, typically in the form of frictionally generated heat. The amount of energy lost can range from major, as from frictional losses with solid boundaries, or minor, as from losses created by acceleration of the flow in fittings, contractions, etc.

Many monographs exist on fluid transport and flow properties. The interested reader is directed to Bird *et al.* (1960), Crane (1976), or Fox and McDonald (1978) for more in-depth coverage.

3.2 Fluid Properties

Flow situations are classified predominantly by the physical properties associated with the flowing fluid. Of these, density and viscosity are the two most influential. Flows in which density variations may be assumed negligible are termed *incompressible* and include most liquid flows and slurries. Gas and vapor flows, however, are considered *compressible*, and density variations play a significant role in fluid behavior. An important result of fluid density, in both cases, is that the density of a moving fluid creates an inertial force, due to the momentum imparted by moving particles, which may be employed to indicate fluid flow.

Viscosity (μ) represents the amount of shear force imparted to a fluid due to a velocity gradient within a flow stream, with units of mass/(length × time). For most fluid systems, viscosity gives rise to the "no-slip" condition, which indicates that a fluid in contact with a solid boundary will have a velocity equal to that of the boundary. A velocity gradient within the flow stream is formed dependent on distance from that solid boundary. This spatial variation in fluid velocity is termed the *flow profile.*

A useful classification of flow systems arises from a comparison of density-dependent fluid inertial forces to viscosity-dependent viscous forces. This relationship is indicated through a dimensionless parameter, the Reynolds number, in terms of a characteristic dimension L of the flow, such as a pipe diameter or duct width:

$$\text{Re} = \rho v L/\mu. \qquad (14)$$

Evaluation of the Reynolds number allows any fluid system to be categorized according to the ratio of these forces. Although no definite values exist because of variability of systems, flows with Reynolds number of less than ~2000 are dominated by viscous forces and are termed *laminar,* indicating that mixing across fluid layers parallel to the direction of flow is insignificant. For Reynolds numbers above ~4000, flow is considered turbulent, with significant inertial forces of molecular collisions causing mixing between adjacent fluid layers superimposed upon the mean motion of flow.

3.3 Fully Developed Flow

Many measurement systems infer a bulk fluid condition from a measurement of a small, localized portion of the flow. Development of the flow profile is an important condition that must be accounted for in these flowmetering systems. By reason of both inertial and viscous effects, a flow profile does not immediately change when boundary conditions along the flow path change. Flow-path irregularities, curves, restrictions, expansions, etc., will cause velocity profile changes within flow streams, but the momentum and viscosity of the flow will cause those changes to be propagated at different rates through the flow field. *Fully developed flow* describes flow fields for which the velocity profile no longer changes as the flow moves downstream. For most engineering flow conditions, 10–20 pipe diameters of straight, uniform flow path are required prior to a flow measurement to insure that flow has fully developed (ASME, 1988).

3.4 Flowmeter Classifications

Flowmeters may be classified by the general method in which flow information is extracted from the fluid system. They may be *inferential,* in which a physically measurable parameter of the flow stream other than flow or velocity is used to indicate the flow rate. Velocity may also be *directly* measured. Alternatively, *energy additive* flowmeters inject additional energy into the flow stream to create a useful, measurable parameter.

Inferential flowmeters are, by far, the most widely used form of flowmeter. Common types of this flowmeter include the orifice plate, venturi, sonic nozzle, and pitot/pitot-static tube, which all infer flow rate from a differential pressure measurement imparted by an *inertial* change in flow velocity. Laminar flow elements also measure flow velocity from an inferential, *frictionally* generated differential pressure basis. The variable-area flowmeters, including the rotameter and spring-loaded variable aperture, as well as the drag-body or target flowmeter, vortex-shedding flowmeter, and fluidic or Coanda flowmeter also use an inferential basis to measure flow.

Direct-measurement flowmeters include the positive-displacement flowmeter and the turbine, or paddle, flowmeter, which directly measure volumetric flow rate and velocity, respectively.

Energy-additive flowmeters include the magnetic flowmeters, thermal mass flowmeters, tracer-dilution methods, and Coriolis, ultrasonic, and nuclear-magnetic-resonance flowmeters. Each of these devices imparts additional energy into the system—thermal, acoustic, electric, etc.—and monitors the effects of the flow upon that energy.

3.5 Inferential Flowmeters

3.5.1 Differential Pressure Flowmeters

The most widely applied style of flowmeter within industry is the differential pressure flowmeter, which applies Bernoulli's principle by creating a measurable pressure change within a flow system through variation of flow kinetic energy. Many varieties of the differential pressure, or DP, flowmeter exist, most of which create a localized pressure change in the flow by restricting the flow path (see Fig. 11). Conservation of mass entails that, with the cross-sectional area of the flow restricted, the flow velocity through the restricted area will increase. For incompressible flows, this velocity increase may be expressed by a rearrangement of the continuity equation,

$$v_{\text{restriction}} = A_{\text{flow}} v_{\text{flow}} / A_{\text{resriction}}. \tag{15}$$

For compressible flows, the local variation in fluid density imparted by the restriction complicates the situation; however, the overall end result is still a localized kinetic energy increase. Provided no elevation changes occur within the flow path, Bernoulli's principle requires that the overall system energy must remain constant, resulting in a concurrent decrease in static energy. The static-energy decrease is expressed as a localized drop in pressure across the restriction. The

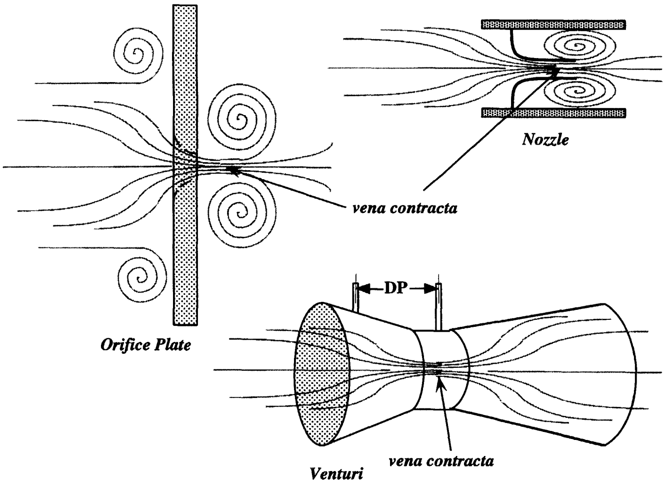

FIG. 11. Differential-pressure flowmeters restrict the flow path, causing a localized increase in velocity, and a concurrent decrease in static pressure of the system. A square-root relationship between pressure drop and flow velocity results.

magnitude of the pressure drop is often a convenient parameter to measure and may be used to infer the velocity of the fluid, and thus the flow rate.

3.5.2 The Orifice Plate The most common DP flowmeter in use today is the orifice plate, which exemplifies the simplicity of differential-pressure flow measurement. The orifice plate is typically a flat metal plate that is inserted perpendicular to a flow stream. It is simple to install and maintain, and may be interchanged for different flow situations by inserting an orifice plate with a differently sized opening.

The relationship of volumetric flow rate Q to pressure drop across the orifice is typically presented as

$$Q = CA\sqrt{\Delta P/\rho}, \tag{16}$$

which is merely an algebraic rearrangement of Bernoulli's principle. The coefficient C is conventionally termed the discharge coefficient and accounts for *permanent* energy losses caused by the restriction. Often the discharge coefficient is represented as two separate factors: C_c, the contraction coefficient, and C_v, the velocity coefficient. The contraction coefficient arises from incomplete development of the flow within the short confines of the sharp-edged orifice. Rather, the actual location of maximum flow contraction, or *vena contracta,* occurs downstream from the orifice. This further contraction reduces the virtual cross-sectional area of the flowmeter, necessitating the use of a contraction correction. The velocity coefficient reflects nonideal, reduced velocity conditions within the *vena contracta* due to frictional energy losses. As both of these loss mechanisms are density and viscosity dependent, the overall discharge coefficient is functionally dependent upon the Reynolds number of the flow, allowing the orifice plate to be conveniently implemented on a variety of fluids and flow conditions. Typical orifice styles, with sharp edges and resultant abrupt changes in cross-sectional area, provide more predictable discharge coefficients at higher Reynolds numbers as a result of a more randomized distribution of nonaxial velocities. Orifice design may vary to improve this predictability for lower Reynolds number flows, with smooth-contoured orifices allowing good results at Reynolds numbers as low as 250 (Spitzer, 1991). Other orifice design modifications, such as nonconcentric orientation of the flow opening, may be implemented to allow better performance in slurries, dirty gases, and two-phase flow scenarios. Design modifications, such as rounding of orifice edges, may also be employed to decrease the permanent pressure loss, caused by intermolecular frictional losses of nonaxial fluid collisions within the restriction and by eddy formation upstream and downstream of the insertion. On account of these factors, permanent pressure loss is significant in orifice plates.

Actual flow measurement is performed by a differential pressure sensor across the orifice. Placement of pressure taps is important, as the velocity profile, and concurrent pressure change, form gradually as the flow profile develops. Caution is also necessary to insure symmetry of the flow in the direction of flow, as singular pressure taps require flow symmetry to reflect bulk flow conditions correctly.

3.5.3 The Venturi and Nozzle The variation of orifice discharge coefficient with lower Reynolds number, and the tendency of the orifice plates to become fouled by eddy-deposited accumulation, promotes the use of the venturi and the flow nozzle as more accurate and reliable DP producers. A venturi is a flow restriction device designed such that the solid boundary of the restriction better matches the actual flow profile. Venturis feature longer, smooth-curved restrictions, with a pressure tap located at the position of maximum contraction, or throat, of the device. By shaping the solid boundary in the profile of the actual flow stream, the throat of the device and the *vena contracta* of the flow coincide, effectively reducing the contraction coefficient C_c to zero. The end result is that the venturi is much more predictable and repeatable than the orifice plate for widely ranging flow conditions. Further, the smooth flow profile of the device minimizes frictional abrasion and the accumulation of deposits within the meter, increasing reliability. The reduction of frictional losses also promotes improved pressure recovery in the flow stream.

Flow nozzles feature a smooth entrance, lengthened similar to a venturi, and a sharp exit, characteristic of an orifice plate. The frictional losses of the sudden radial expansion at the exit cause high permanent pressure losses, as with the orifice, but the entry pressure loss is typically lower than in a similar-sized orifice, because of the smooth entrance. Thus, the overall unrecoverable pressure drop across the entire system is lower than with an orifice plate.

For higher accuracy and repeatability, venturi meters and nozzle flowmeters may employ multiple pressure taps axisymmetrically through the solid boundary. The multiple taps are channeled to the individual upstream and downstream ports of the differential pressure device, and effectively integrate these individual pressure conditions. Asymmetric variations are thus averaged out and bulk flow conditions are better inferred.

3.5.4 The Pitot and Pitot-Static Tube The pitot and pitot-static flowmeters infer fluid velocity by generating a highly localized velocity stagnation pressure at the tip of a tube placed normal to a flow stream. Per Bernoulli's principle, all kinetic energy at this impact point is converted to static energy, in the form of a pressure increase. A pressure tap in communication with the impact port will indicate total system pressure, including the prevailing pressure of the moving fluid, in addition to the pressure created by velocity stagnation. By a subsequent measurement of the prevailing pressure, the velocity stagnation pressure may be isolated to indicate velocity of the flow. In a pitot-static tube, an annular pressure port is provided within the pitot tube assembly for measurement of the static pressure component of the total system pressure. If this provision is not available, as in a pure pitot tube, an additional pressure port must be provided to determine the static pressure within the system.

Because of the highly localized nature of a stagnation velocity measurement, the pitot and pitot-static tube flowmeters are commonly used to map flow profiles. By positioning the measurement point of the pitot tube at various insertion depths into the flow stream, localized fluid velocities may be determined and an accurate flow profile may be obtained. Considerable care must be taken, however, to insure proper positioning of the pitot assembly within the flow, as slight misalignment of the tube within the flow may cause significant measurement error due to incomplete stagnation of flow at the impact port. To reduce this error mechanism, pitot tube assemblies often employ symmetrical pressure ports along the measurement tube. These ports are connected to differential pressure indicators to indicate unbalanced velocity distribution between the ports, indicative of misalignment of the probe with respect to the flow.

3.5.5 The Laminar Flow Element The previous DP-producing flowmeters all apply Bernoulli's principle to create a measurable static pressure change in the flow, dependent upon changing fluid velocities, while intentionally minimizing frictional energy losses of the system. In conventional terms of fluid dynamics, these devices have harnessed measurability from the *minor* energy losses of flow systems attributable to pressure changes due to sudden expansions and contractions of the flow stream, while minimizing *major* energy losses attributable to frictional losses with the solid boundary. Bernoulli's principle indicates that velocity and flow-rate measurements from this basis will always be governed by a "square-root" relationship with the pressure drop across the measurement system. In contrast to this, the laminar flow element is considered a linear device.

The laminar flow element is a DP-producing flowmeter that is typically composed of many small-diameter flow paths bundled into a single assembly that is inserted into a flow path. The purpose of the multiple small-diameter flow paths is to force the flow into the laminar regime by reduction of the radial path length of turbulent, i.e., nonaxial, flow. Laminar pipe flow subscribes to a linear velocity/pressure-drop relationship, as formulated for internal pipe flow by Hagen and Poiseuille:

$$\nu = (\Delta P)\pi r^2/8\mu L, \qquad (17)$$

where the differential pressure drop is caused by frictional energy losses of shearing fluid layers along a pipe of length L and radius r. Although predominated by linear viscous losses, some inertial losses will be generated at the entrance and exit of a laminar flow element, where the flow stream is accelerated into and out of the laminarizing restrictions of the assembly. These losses will cause the laminar flow element to operate with some nonlinearity, due to the square-root relationship of pressure drop with these unavoidable velocity changes. For very low–Reynolds-number flows (i.e., <200), either density or velocity is sufficiently low that the inertial effects are minor and can usually be neglected.

3.6 Variable-Area Flowmeters

The variable-area flowmeters operate on the same principles as differential pressure flowmeters, but they employ a changing cross-sectional flow area, as opposed to a pressure measurement, to indicate flow rate. The two most common variable-area flowmeters are the rotameter and the spring-loaded variable aperture, although many variations exist on the common themes of these two styles.

3.6.1 The Rotameter The rotameter is most traditionally employed as a demarcated, tapered glass tube, containing an indicating float that is free to move vertically within the tube in response to the rate of fluid flow moving vertically through the assembly (see Fig. 12). The tapering of the tube and the fixed diameter of the float provide for a variable annular cross-sectional area for the flow, dependent upon the position of the float within the tube. The variable area, in turn, allows for a variable velocity and resultant pressure drop across the float as dictated by conservation of mass and Bernoulli's principle. The indication of flow rate thus results from the vertical positioning of the float, which occurs where all forces upon the float, including gravity, buoyancy, viscous drag, and differential pressure, become balanced. A mechanical free-body diagram assists in visualizing the required force balance.

The weight of the float is fixed and dependent upon the float material selected. The buoyancy force is caused by the displacement of fluid by the suspended float and is usually a relatively minor force component as typical float materials are very dense, i.e., stainless steel or sapphire. The viscous force upon the float is attributable to the average shear on the surface of the float and is thus dependent upon the velocity of flow past the float, viscosity of the flow, and float geometry. A large proportion of rotameter floats are spherical balls, for which viscous drag may be evaluated using Stokes's drag relationships (Bird *et al.*, 1960). To increase application of rotameters through a range of fluids, float geometry is often tailored to reduce these viscous drag forces. These float models feature very sharp downstream flow edges that serve as sharp-edged orifices, minimizing frictional viscous losses while creating significant inertial forces as velocity increases through the resulting annular orifice.

3.6.2 The Spring-Loaded Variable-Aperture Flowmeter The spring-loaded variable-aperture flowmeter operates on the same principles as the rotameter, with fluid

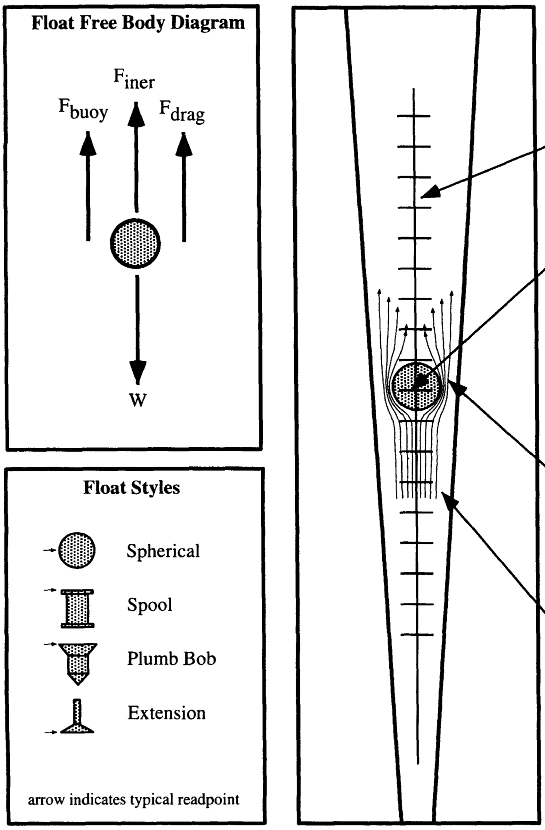
Float Free Body Diagram
Finer
Fbuoy
Fdrag
W
Float Styles
Spherical
Spool
Plumb Bob
Extension
arrow indicates typical readpoint
Scale Indication
Float
Variable Area Annular Orifice
Flow Streamlines

FIG. 12. The rotameter indicates flow by inertial and viscous displacement of a vertically oriented float. Displacement may be measured in a variety of ways including visually, magnetically, optically, etc.

forces opposing a spring force in addition to the weight of a vertically oriented float. A plunger restricts fluid flow through a sharp-edged orifice, creating an inertially driven pressure differential across the plunger. This pressure differential exerts a force upon the plunger, which is opposed by a spring force dependent upon the position of the plunger. The balance of forces upon the assembly results in steady-state positioning of the plunger, resulting in a pressure drop linearly proportional to the flow rate. Horizontal orientation removes the effects of weight and buoyancy on the plunger. System behavior is thus dominated by the opposition of inertial forces with spring forces, with the latter increasing rangeability of the device over other DP producers.

3.6.3 The Drag-Body or "Target" Flowmeter In the event that fluid properties restrict either the use of moving parts, as in the variable-area flowmeters, or the presence of pressure taps for differential-pressure flowmeters, the drag-body or "target" flowmeter may provide an acceptable alternative (see Fig. 13). The drag body operates on the same principles of inertially driven differential pressure forces and viscous drag forces

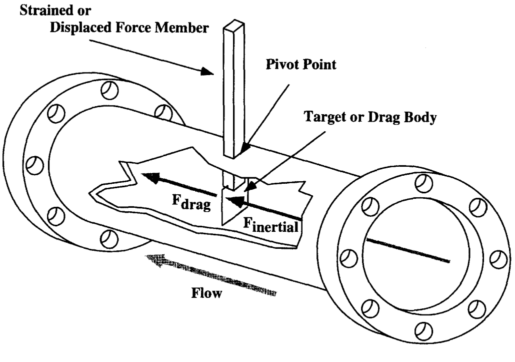
Strained or Displaced Force Member
Pivot Point
Target or Drag Body
Fdrag
Finertial
Flow

FIG. 13. Target, or drag-body, flowmeters indicate flow through the inertial and viscous displacement or straining of a cantilevered element placed into the flowstream.

but employs a mechanically constrained target to provide measurement of fluid flow. This target acts as the traditional flow obstruction, restricting flow cross-sectional area and causing localized velocity increases. These increases cause the Bernoulli-explained differential-pressure drivers, which exert force upon the target. Additionally, fluid flow around the target exerts viscous drag forces. These forces upon the constrained target may be measured using strain measurements, or the relatively small changes in target position with response to the flow forces may be mechanically amplified for direct, mechanical indication of the flow. Rangeability of the device may be altered by simply changing the size or shape of the target element.

3.6.4 The Vortex-Shedding Flowmeter

The vortex-shedding flowmeter, or bluff-body flowmeter (Fig. 14), is another device that may be used when flow conditions prohibit pressure taps or moving parts. If a bluff body, or strut, is placed perpendicular to a flow stream, the localized velocity distortion around the bluff body will cause an unstable wake behind the obstruction. Eddy currents will be created that grow into vortices and are shed from each side of the bluff body in a repeatable, oscillatory nature. Opposite rotation direction will result for eddies from opposing sides of the bluff body by reason of the "no-slip" interface on each side with the solid boundary. Oscillation occurs because of the localized pressure increase of swirling eddy separation from the body. The frequency of oscillation f is linearly dependent of the velocity of the flow stream and the characteristic diameter d of the bluff body:

$$f = St\, v/d. \tag{18}$$

The Strouhal number St is a dimensionless constant of proportionality for this frequency–velocity relationship and is dependent on the Reynolds number of the flow with respect to the characteristic diameter of the strut. Over a wide range of Reynolds numbers (with respect to bluff-body dimensions) within the turbulent regime, the Strouhal number is essentially constant. Below this, the Strouhal number decreases with decreasing Reynolds number. Within the low laminar regime, vortex formation does not occur as viscous forces tend to reduce eddy formation, and pressure fluctuations around the bluff body are not sufficient to toggle the eddy wake. The specific Reynolds-number cutoffs are highly dependent upon bluff-body geometry (Spitzer, 1991, p. 298). However, above this cutoff value, the vortex-shedding flowmeter can be widely applied to a large range of flow conditions, including steam, clean liquids, and gas.

By measuring the frequency of oscillatory vortex generation, an indication of flow rate may be obtained. Many different methods are available for vortex detection. A widely applied method is to measure the strain induced upon a cantilevered strut, which acts like the previously discussed drag element of a target flowmeter. The shedded vortices create oscillatory drag upon the strut, which may be measured as strain in the strut's cantilevered suspension. Piezoelectric elements may also be employed to generate voltage potential due to the oscillatory strain of the strut. The vortices may also be counted downstream of the strut by various ultrasonic methods. Vortices shed from opposing sides of the bluff body will have opposite rotational spins. Vortex counting may make use of this by timing the transversal speed of an acoustic pulse. During the first half of a clockwise spinning vortex and last half of a counterclockwise vortex, transversal speed will be increased through the summation of mean sound velocity and the vortex angular velocity. During the latter half of a clock-

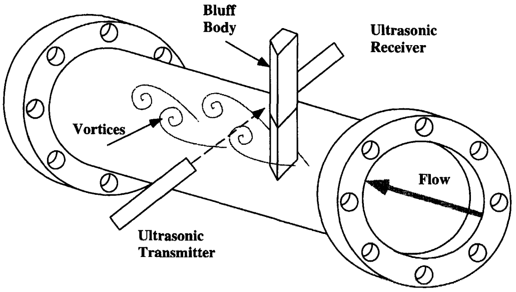

FIG. 14. A bluff body, obstructing fluid flow, will create a bistable wake of vortices at a frequency linearly proportional to flow velocity. This vortex flowmeter counts shed vortices by measuring the propagation rate of an ultrasonic signal across the flowstream. Clockwise spinning vortices shed from one side of the bluff body will increase signal velocity, counterclockwise vortices will decrease signal velocity.

wise, and front half of a counter-clockwise spinning vortex, transversal speed will be decreased by the negative effect of the vortex angular velocity. Thus, the opposing vortices may be counted by differentiating between the transversal time of the acoustic pulse. Another counting method is to rely on the acoustic focusing and defocusing effects of the fluid angular velocity upon a cross-stream transducer/receiver pair.

3.6.5 The Coanda, or Fluidic, Flowmeter The Coanda flowmeter operates on the same principles as the vortex-shedding flowmeter, by monitoring the same oscillatory nature of a bistable fluidic *jet,* as opposed to the vortex *wake.* In the near presence of a solid boundary, a fluid jet will attach itself to the solid boundary according to no-slip principles, and the associated effects of boundary-layer fluid velocity. A Coanda flowmeter applies this phenomena by allowing two separate fluid/boundary jet paths, with fluidic feedback from each path to toggle the flow from one path to the other. Under similar Reynolds-number limitations as a vortex meter, the toggling frequency of the bistable jet path is linearly related to the fluid velocity through the Strouhal number.

The frequency of fluid jet toggling may be measured in a variety of ways. The pressure within a single fluid path may be monitored, or optical methods may also be employed if the fluid is optically dense. Another technique is to employ a self-heating temperature element, which is cooled by the flow during the presence of the jet through one path, and reheats when the flow is in the other path.

3.7 Direct-Measurement Flowmeters

3.7.1 The Positive-Displacement Flowmeter Positive-displacement techniques split the flow into known-volume collection chambers, which are counted, or *totalized,* to indicate flow. The oldest form of this style of meter is the reciprocating piston, in which two separate piston chambers are alternately filled with fluid. The reciprocating motion of the displaced pistons may be mechanically counted, or end-position switches may be employed to count fill cycles of the chambers. Other chamber configurations may employ a rotary vane of equally sized volumes, sets of rotating lobes, or a nutating disk. Each method requires a means of counting chamber fill cycles to totalize the flow. Counting methods may be as simple as a mechanical counter, or may employ optical or electrical proximity techniques, most commonly in the form of a variable-reluctance, i.e., magnetic, pick-up. The latter are often used to totalize rotary vane passes, or end-positioning of linear pistons.

3.7.2 The Turbine Flowmeter The turbine flowmeter is very similar to a rotary-vane positive-displacement flowmeter, but does not require known-volume collection chambers. Rather, vanes of a turbine placed perpendicular to the flow stream will be forced to turn with an angular velocity that is linearly proportional to the flow velocity of the stream. Like the positive-displacement flowmeter, vane rotation may be counted using a variable-reluctance pick-up or other proximity-style detection methods, with the resultant frequency indicative of flow. Above a specific flow rate, at which frictional losses of the turbine system become negligible, turbine meters perform with a very stable frequency-to-flow-rate relationship.

3.8 Energy-Additive Flowmeters

3.8.1 The Magnetic Flowmeter For electrically conductive fluids, Faraday's law of induction may be used to measure volumetric flow rate magnetically (see Fig. 15) (Spitzer, 1991). Faraday determined that a changing magnetic field could induce a voltage proportional to the time rate of change of the current that produced the magnetic field. Applied analogously to flow, the relative average motion $\bar{v}$ of an electrically conductive fluid flowing perpendicular through a magnetic field of density B over a path length L will generate a voltage E proportional to the volumetric flow rate by a constant c:

$$E = cBL\bar{v}. \tag{19}$$

The generated voltage is influenced by the total cross-sectional area of the flow experiencing the magnetic field, and the measurement is thus independent of velocity profile of the flow and the parameters involved with

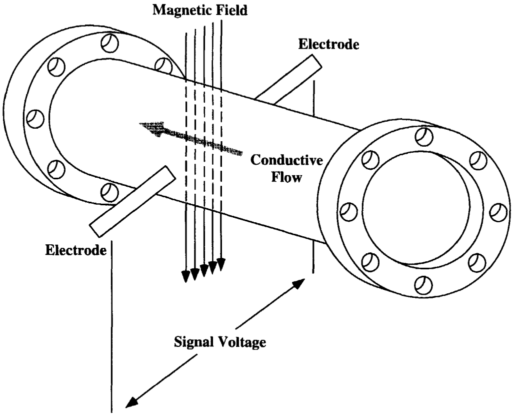

FIG. 15. For conductive fluids, magnetic flowmeters apply Faraday's law, indicating flow through the generation of an electric potential at right angles to a magnetic flux perpendicular to the direction of flow.

this profile, i.e., density and viscosity. The magnetic flowmeter thus provides a measurement independent of Reynolds number of the flow, and of typical flow-measurement restraints such as straight pipe lengths upstream of the measurement area. Further, this flowmeter is inherently nonintrusive, as the magnetic field and subsequent voltage measurement can be generated and sensed external to the flow system. Magnetic flowmeters therefore feature small pressure losses, limited only by "straight path" frictional losses of the piping system.

3.8.2 The Thermal Mass Flowmeter and Hot-Wire Anemometer Another flow technique that may be used noninvasively is the thermal mass flowmeter. Used extensively in semiconductor manufacturing, these flowmeters indicate mass flow rate by observing the effects of heat energy added to the flow stream as governed by the equation of heat transfer (Tison, 1996),

$$Q = \dot{m}C_p\Delta T. \tag{20}$$

A typical application of this technique is to wrap a thin-walled, thermally conductive flow path with heating elements, which serve in dual capacity as temperature sensors. Under no-flow conditions, the thermal profile of the fluid between these heaters will achieve a steady-state symmetric profile as depicted in Fig. 16. Under flowing conditions, however, this thermal profile will become skewed, as heat is carried downstream with the flow. Under laminar flow conditions, this skewing is linearly related to flow rate, and the temperature difference between the heating elements may be used to indicate mass flow rate, dependent primarily upon the heat capacity of the fluid. To maintain laminar flow, the measurement flow path is typically a small capillary tube capable of handling only low flow rates (~cubic centimeters per minute). For higher flows, an appropriately sized laminar bypass flow path may be used in parallel to the sensing flow path to increase rangeability of the device. Provided both paths remain laminar, and minimum minor pressure losses occur between them, the flow splitting between the paths will remain constant, and the sensor flow indication will be

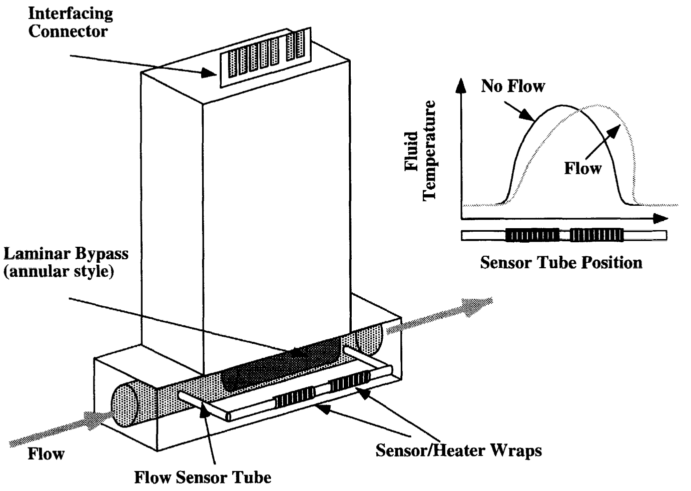

FIG. 16. The thermal mass flowmeter measures flow by detecting the skewing of a thermal profile due to convective transport by the flow stream. This configuration uses a pair of dual-capacity heaters/sensors to heat and measure the average upstream and downstream fluid temperatures. A bypass is used to increase rangeability of the flow.

linearly proportional to total flow of the system.

Localized, as opposed to bulk-fluid, mass flow rate may be obtained using a hot-wire anemometer. This flowmeter employs the same principles as the thermal mass flowmeter by typically measuring the heat loss from a heated wire maintained at constant temperature, and the temperature of the flow stream with an unheated wire. The heated wire, serving in dual capacity as a thermometric element, will lose an amount of the heat to the flow depending upon the Reynolds number and heat-transfer characteristics of the fluid, and the mass flow rate. Typically very fast, because of the low thermal mass of the wire element, these devices are often used for profiling flows.

3.8.3 The Tracer-Dilution and Cross-Correlation Flowmeter The tracer-dilution flowmeter is based on mass conservation methods and depends on measuring the dilution of a known amount of tracer added to the flow. Dilution techniques require adequate mixing of tracers, and thus are typically employed for high-velocity, highly turbulent flows. Tracer materials include chemical salts, fluorescent dyes, and radioactive materials, in addition to other materials that lend themselves to easy measurement within the flow system. Administration of tracers may be performed either at a constant rate of injection or as a periodic injection of a slug of material. Downstream dilution measurement techniques include fluid electrical conductivity, scintillation detection, and optical-based methods such as fluorometry, transmissivity, etc.

Provided the tracer input quantity is known, tracer dilution techniques do not require knowledge of the geometry of the flow path. Flow is determined by measuring the dilution of the *known* amount of tracer. For applications where the flow geometry is well characterized, tracers may also be applied in a cross-correlation technique, where the administration and measurement of a tracer is performed within a known path length. The cross-correlation measurement under these conditions provides a time-of-flight indication of flow rate, which, with appropriate geometry information, may be used to indicate volumetric flow rate.

3.8.4 The Coriolis Flowmeter Flowmeters in this category apply inertial Coriolis forces to measure flow. Coriolis force is produced by introducing an oscillating acceleration, about an axis, to a pipe containing a flowing mass of fluid (see Fig. 17) (Bau *et al.*, 1994). The elemental Coriolis force is quantified as two times the elemental mass times the cross product of the angular velocity and the velocity of the fluid perpendicular to the axis of rotation,

$$df_c = 2dm\boldsymbol{\omega} \times \boldsymbol{v}. \tag{21}$$

A common implementation of Coriolis mass-flow measurement is the vibrated u-tube, in which flow is directed through a horseshoe-shaped pipe of low stiffness. At the apex of the horseshoe, the tube is oscil-

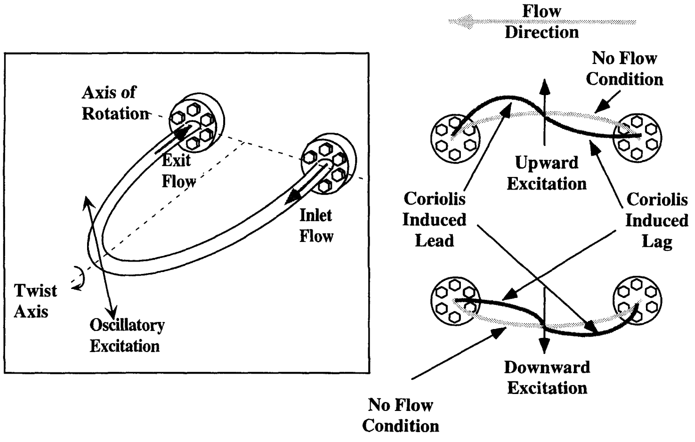

FIG. 17. The U-tube Coriolis flowmeter excites the flow path in an oscillatory pattern. Velocity perpendicular to the axis of rotation causes a torsional force about the twist axis.

lated (about an axis passing through the tips of the horseshoe) at a natural frequency. This oscillation equates to an angular velocity about the axis of rotation. Along the inlet and outlet legs of the horseshoe, a torsional force is produced as the fluid flow lags the pipe wall motion. This force is of opposite direction and equal magnitude in the inlet and outlet legs of the horseshoe. Through displacement measurement of the magnitude of the oscillating twist in the horseshoe, the mass flow rate of the fluid may be obtained. Often, to balance the effects of the torsional forces of the u-tube on process piping, a dual u-tube flow path will be used, with the tubes oscillated at 180° out of phase. The resultant torsional forces of the tubes then destructively cancel one another, resulting in a net balance of forces at the process piping.

3.8.5 The Ultrasonic Time-of-Flight and Doppler or Laser-Doppler Flowmeters The ultrasonic flowmeter monitors the effects of a flow stream on either the time of flight or the scattering of an acoustically propagated signal (Bau *et al.*, 1994). In the former case, the *time-of-flight,* or *transit-time,* flowmeter measures the transit time of an acoustic signal between an upstream and downstream pair of transceivers. These transceivers are set on opposite sides of the flow path from one another and are carefully adjusted such that the acoustic signal travels at the fixed angle between them. The propagation speed of the signal will be the vector sum of the temperature-dependent speed of sound through the medium, plus the additive (for downstream propagation) or subtractive (for upstream) effect of the moving fluid. Dependence upon fluid speed of sound may be eliminated by beating signals from upstream-to-downstream and downstream-to-upstream transceivers. The beat difference between upstream and downstream propagation is dependent only on fluid velocity, the fixed angle between transducers, and the linear path length between them. Time-of-flight measurements are indicative of average fluid velocities, as the propagation rate of the signal will be influenced by the fluid velocity profile throughout the path length between transceivers. Multiple-path transceivers may be employed to indicate bulk volumetric flow better by averaging the measurement across many different acoustic paths.

The Doppler flowmeter is another propagation method for determining average flow rate. Inhomogeneous inclusions within a flow stream will cause reflection of a wave front propagating through the medium (see Fig. 18). The relative motion between the inhomogeneities and the wave front will create a Doppler shift in the frequency of the reflected signal. Provided the inclusions are traveling at the fluid velocity, this Doppler shift is linearly proportional to the velocity of the fluid. Within liquids, entrained air and dirt may act as reflectors and an ultrasonic transceiver may be employed to measure their Doppler shift. In gases, where acoustic propagation is diminished, lasers may be employed, with entrained solids such as dust or condensing water vapor acting to the same effect. The latter are often termed laser-Doppler velocimeters or laser-Doppler anemometers.

3.8.6 Nuclear-Magnetic-Resonance Flowmeters Nuclear-magnetic-resonance (NMR) flowmeters provide a noncontact method for measuring fluid flow of materials with a nuclear magnetic moment (odd number of protons). NMR flowmeters apply a radio-frequency (rf) field to sense resonance interaction between an externally applied magnetic field and the magnetic moments of the material being measured. Hydrogen has the largest nuclear magnetic moment of any element, and thus NMR flow measurements are particularly well suited to hydrogenous materials. One particularly advantageous situation for NMR flowmeters is the imaging of blood flow within intact, functioning human hearts.

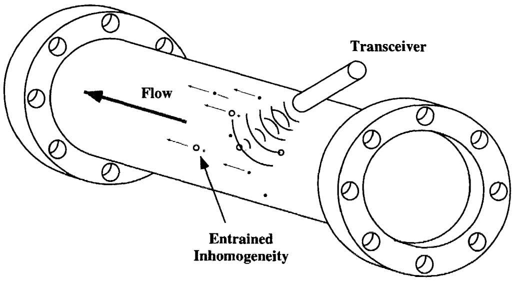

FIG. 18. Ultrasonic Doppler flowmeters measure flow velocity by observing the Doppler frequency shift imparted to a propagated wave form by the relative motion of entrained inhomogeneities in the flow stream.

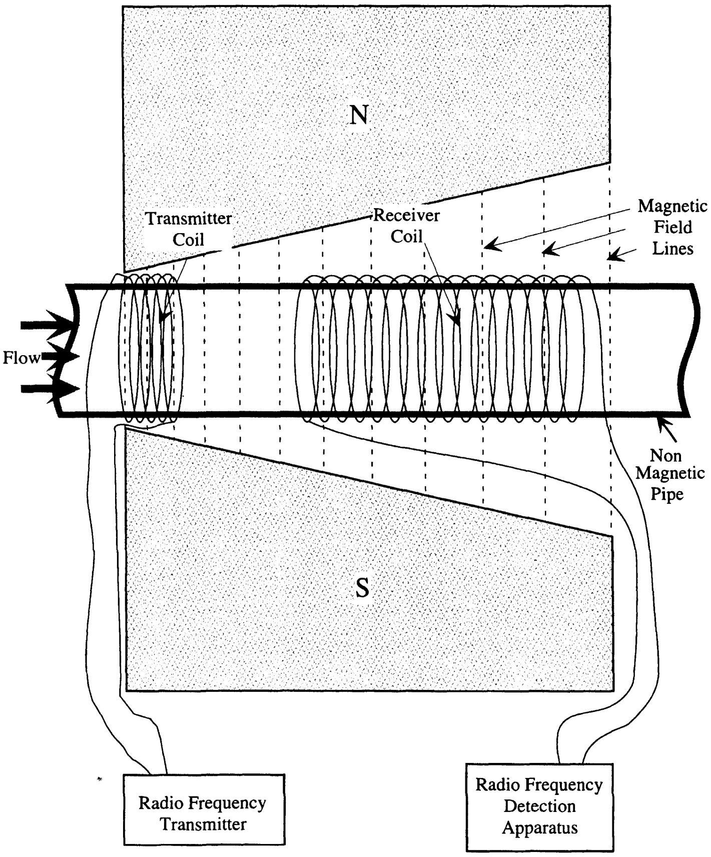

FIG. 19. Gradient-field–type nuclear-magnetic-resonance flowmeter.

Several different implementations of NMR flowmeters have been developed. An approximate grouping scheme for these methods is one-coil inflow and outflow, two-coil labeled time-of-flight, and field-gradient techniques. The field-gradient technique is selected here to illustrate the concepts underlying NMR flowmeter technology (see Fig. 19, derived from King and Riewerts, 1985).

In this implementation, the nuclear spins of the flowing fluid material are aligned by a transverse magnetic field. A linear gradient magnetic field along the flow is simultaneously applied. rf pulses at the material's magnetic resonance frequency are briefly applied to rotate the magnetic moment relative to the externally applied static magnetic field. In the gradient field, only a narrow slice of the flowing material has its resonant frequency attuned to the rf pulse. After the termination of the rf pulse, the subatomic particles realign to the external magnetic field and radiate electromagnetic fields that induce measurable voltages in an rf sensing coil. As the excited slice of the material moves into a field region of different magnetic intensity (flows downstream), the frequency of the signal emitted by the particles, as they realign with the magnetic field, changes proportionately. The signal intensity indicates the number of particles realigning with the external magnetic field. The combination of signal time delay (relative to the applied rf pulse) and signal frequency indicates the particle velocity.

Works Cited

American Society of Mechanical Engineers (1988), *Method for Establishing Installation Effects on Flowmeters,* ASME MFC-10M, Fairfield. CT: ASME.

American Society for Testing and Materials (1993), *Temperature Electromotive Force (EMF) Tables for Standardized Thermocouples,* 14.03, ASTM E230, West Conshohocken, PA: ASTM.

Bau, H. H., de Rooij, N. F., Kloeck, B. (Eds.)

(1994), *Mechanical Sensors,* Vol. 7 of *Sensors, a Comprehensive Survey,* New York: VCH Publishers.

Bird, R. B, Stewart, W. E., Lightfoot, E. N. (1960), *Transport Phenomena,* New York: Wiley.

Blatt, F. J., Schroeder, P. A., Foiles, C. L., Greig, D. (1976), *Thermoelectric Power in Metals,* New York: Plenum.

Boiarski, A. A., McGinniss, V. D. (1993), *Fiber Optic Distributed Temperature Sensor Demonstration,* EPRI TR-101950, Palo Alto, CA: Electric Power Research Institute

Crane Co. (1976), *Flow of Fluids through Valves, Fittings, and Pipe,* Technical Paper No. 410, Chicago: Crane Co.

DeWitt, D. P., Nutter, G. D. (1988), *Theory & Practice of Radiation Thermometry,* New York: Wiley.

Fox, R. W., McDonald, A. T. (1978), *Introduction to Fluid Mechanics,* New York: Wiley.

Fraden, J. (1997), *Handbook of Modern Sensors—Physics, Designs, and Applications,* 2nd ed., Woodbury, NY: AIP Press.

Grattan, K. T. V., Zhang, Z. Y. (1995), *Fiber Optic Fluorescence Thermometry,* London: Chapman and Hall, pp. 222–225.

Hirschfeld, T. B., Haugen, G. B. (1988), *Method and Apparatus for Simultaneously Measuring Temperature and Pressure,* U.S. Patent No. 4768 886.

Kerlin T. W., Shepard, R. L. (1982), *Industrial Temperature Measurement,* Research Triangle Park, NC: Instrument Society of America.

King, J. D., Riewerts E. S. (1985), *Method and Apparatus for Measuring Flow in a Pipe or Conduit,* U.S. Patent No. 4536 711.

Kittel, C. (1986), *Introduction to Solid State Physics,* 6th ed., New York: Wiley.

Krohn, D. A. (1992), *Fiber Optic Sensors,* 2nd ed., Research Triangle Park, NC: Instrument Society of America.

Liboff, R. L. (1980), *Introductory Quantum Mechanics,* Oakland, CA: Holden Day, pp. 585–589.

Lynnworth, L. C. (1989), *Ultrasonic Measurement for Process Control,* Boston: Academic.

O'Hanlon, J. F. (1989), *A User's Guide to Vacuum Technology,* New York: Wiley.

Park, R. M. (Ed.) (1993), *Manual for the Use of Thermocouples in Temperature Measurement,* MNL-12, West Conshohocken, PA: ASTM.

Pollock, D. D. (1985), *Thermoelectricity, Theory, Themometry, Tool,* ASTM Special Publication 852, West Conshohocken, PA: ASTM.

Quinn, T. J. (1990), *Temperature,* 2nd ed., Boston: Academic.

Ricolfi, T., Scholtz, J. (Eds.) (1990), *Thermal Sensors,* Vol. 4 of *Sensors, a Comprehensive Survey,* New York: VCH Publishers.

Sachse, H. B. (1975), *Semiconductor Temperature Sensors and Their Applications,* New York: Wiley.

Selim, C. Z. (1995), "The GaAs Hall Element as a Pressure Transducer," *Rev. Sci. Instrum.* **66,** 4996–4998.

Spitzer, D. W. (1991), *Flow Measurement, Practical Guides for Measurement and Control,* Research Triangle Park, NC: Instrument Society of America.

Streetman, B. G. (1980), *Solid State Electronic Devices,* 2nd ed., Englewood Cliffs, NJ: Prentice-Hall.

Tison, S. A. (1996), "Accurate Flow Measurement in Vacuum Processing," *Solid-State Technol.* **39** (10), 73, 74, 78, 80, 83, 85.

Udd, E. (Ed.) (1991), *Fiber Optic Sensors,* New York: Wiley, pp. 333–334.

White, D. R., Galleano, R., Actis, A., Brixy, H., Degroot, M., Dubbeldam, J., Reesink A. L., Edler, F., Sakurai, H., Shepard, R. L., Gallop, J. C. (1996), "The Status of Johnson Noise Thermometry," *Metrologia* **33,** 325–335.

Further Reading

Benedict, R. P. (1984), *Fundamentals of Temperature, Pressure, and Flow Measurements,* 3rd ed., New York: Wiley.

Brignell, J., White, N. (1994), *Intelligent Sensor Systems,* Bristol, U.K.: IOP Publishing.

Doebelin, E. O. (1990), *Measurement Systems—Application and Design,* 4th ed., New York: McGraw-Hill.

Finkelstein, L., Grattan, K. T. V. (Eds.) (1994), *Concise Encyclopedia of Measurement & Instrumentation,* Oxford: Pergamon Press.

Fraden, J. (1997), *Handbook of Modern Sensors—Physics, Designs, and Applications,* 2nd ed., Woodbury, NJ: AIP Press.

Göpel, W., Hesse, J., Zemel, J. N. (Series Eds.), (1989–1991), *Sensors—A Comprehensive Survey,* New York: VCH Publishers.

Soloman, S. (1994), *Sensors and Control Systems in Manufacturing,* New York: McGraw-Hill.

TRANSDUCERS, UNDERWATER ACOUSTIC

W. JACK HUGHES, *Applied Research Laboratory, Pennsylvania State University, State College, Pennsylvania, U.S.A.*

	Introduction	67
1.	**Operating Environment**	68
2.	**Materials**	68
2.1	Piezoelectric Materials	68
2.2	Piezoelectric Ceramics	69
2.3	Composite Ceramics	70
2.4	Piezoelectric Polymers	70
2.5	Magnetostrictive Materials	70
3.	**Testing of Underwater Transducers**	70
3.1	Calibration Formulas	71
3.2	Calculated Parameters	73
3.3	Cavitation	74
4.	**Low-Frequency Transducers (10–5000 Hz)**	75
4.1	Flexural Disk	75
4.2	Flextensional Transducer (300–3000 Hz)	76
4.3	Free-Flooded Ring	77
4.4	Mechanical Transducer	77
4.5	Electromagnetic, Moving-Coil Transducers	77
4.6	Resonant Bubble Transducer	78
4.7	Helmholtz Resonator	78
4.8	Hydroacoustic Projector	78
4.9	Spherical-Shell Transducer	78
4.10	Impulsive Transducer	78
5.	**Longitudinal Vibrator Transducers (3 kHz to >100 MHz)**	79
5.1	Tonpilz Transducers	79
5.2	Thickness Vibrators	81
5.3	Shaped Ceramic Transducers	82
6.	**Other Transducers**	83
	Works Cited	84
	Further Reading	84

INTRODUCTION

A transducer is a device that converts energy from one form to another. An electroacoustic transducer transforms electrical energy into acoustic energy, or vice versa. An electromechanical transducer transforms electrical energy to mechanical energy or movement, which then may transmit energy. These are the energy transformations that will be involved in the transducers mentioned in this article. Understanding and design of transducers is multidisciplinary and involves a knowledge of materials, physics, acoustics, electrical engineering, mechanical engineering, and chemistry. Fortunately, there are many guidelines that can be followed in which the beginner can understand the various disciplines involved. Some people call the successful fabrication of transducers a "black art." This is an overworked term; transducer design and fabrication follow a well thought-out path, but do require a broad knowledge in the sciences and engineering. They also require learning how acoustics interacts with different materials and their acoustic impedances. Baffles, housings, and resonant structures typically surround the transducer and can have a major impact on its performance. Just like in most fields, the "bugs" need to be "worked out" from the design for the transducer to be an integral part of a measuring system.

Landmarks in the history of electroacoustics were Bell's invention of the telephone (1876) and Edison's work on the carbon microphone and the phonograph. Langevin and others used steel-quartz-steel resonator plates, which form a composite underwater projector transducer for the detection of submarines. Equivalent circuits that describe the action of piezoelectric resonators were developed by Van Dyke (1928) and Mason (1934, 1948), and good introductions can be found in Wilson (1985), Stansfield (1990), Hunt (1982), Berlincourt *et al.* (1964), Kinsler *et al.* (1982), and Bobber (1988). The discovery of

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

the large piezoelectric effects in ceramics (e.g., barium titanate, 1946) led to the development of the technology of modern piezoelectric ceramics, and to the improvement in the performance of electromechanical transducers.

There are many different types of transducers that can be discussed. A good list of the different types of transducers used in the field of acoustics can be seen in the *Journal of the Acoustical Society of America* index under the listing 43.38, "Transduction; Acoustical devices for the generation and reproduction of sound." The transducer types from this list that are generally used in underwater acoustics are magnetostrictive transducers, electromagnetic and electrodynamic transducers, piezoelectric and ferroelectric transducers; transducer arrays, acoustic interaction effects in arrays, hydroacoustic and hydraulic transducers, and impulse transducers. Acoustic transducers of various types must satisfy the application needs over a frequency range of seven decades. In Fig. 1 is shown a convenient comparison of the large acoustic frequency range, the application, and devices and materials. This article is devoted mostly to describing underwater transducers, how they are tested, and how they operate. An excellent introduction to underwater electroacoustics can be found in Wilson (1985) and in Kinsler *et al.* (1982).

1. OPERATING ENVIRONMENT

Operating a transducer in an underwater environment is dramatically different from operating in air, and significantly affects how the transducer is designed and the types of materials used in the design. To gather an understanding of the difference, the relative pressure for a given particle displacement in air and water is now determined. The ratio of acoustic pressure p to particle velocity v is equal to the specific acoustic impedance of the medium ρc, where ρ is the density and c is the sound velocity. The particle displacement is given by

$$\chi = v/\omega = (p/\omega)\rho c, \tag{1}$$

where $\omega = 2\pi f$ is the angular frequency and f is the frequency. The acoustic impedance ρc for water is 1.5×10^6 Rayls (Rayl = kg/m^2·sec) and for air is 415 Rayls. For a given pressure and frequency, the ratio of the particle velocities is equal to the ratio of the medium impedances; hence, 3600 times more motion is needed in air than in water to create the same pressure level. From this it is obvious that a transducer in air must be much more compliant than in water (e.g., a compliant loudspeaker versus a stiff ceramic).

2. MATERIALS

Materials typically used in an underwater environment undergo a length or volume change when an alternating electric or magnetic field is applied to the material. They can be generally grouped into piezoelectrics, ceramics, composites, polymers, and magnetostrictive materials.

2.1 Piezoelectric Materials

The piezoelectric effect arises because of the asymmetries in the crystal structure, which create an electric dipole moment in the crystal lattice that is sensitive to both elastic strain and applied electric field.

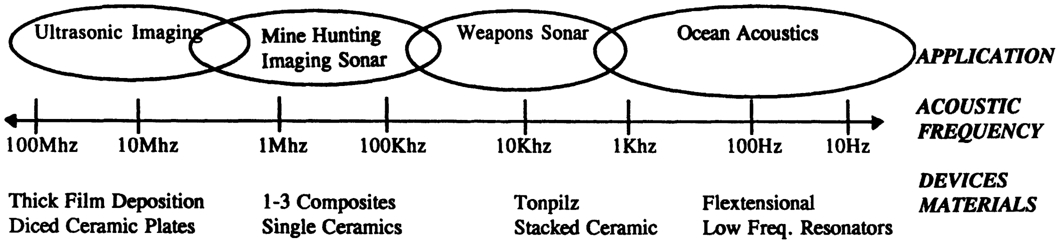

FIG. 1. Acoustic frequency range for transducers.

Quartz is one of the earliest materials discovered with the piezoelectric effect. It is rarely used for underwater projectors; however, its very low internal losses and very high mechanical *Q* lead to the use of quartz in stable resonant circuits in the communication and electronics fields. Rochelle salt, an extremely active piezoelectric material, was used for early underwater projectors and hydrophones, and as phonograph pickups and microphones. It is very sensitive to moisture, which can dissolve the crystal, and hence was replaced by more environmentally stable materials for underwater transducers. Ammonium dihydrogen phosphate (ADP) replaced Rochelle salt in the 1940s. Modern transducers use piezoelectric ceramics, which are sintered polycrystalline materials polarized by large electrical fields while they are cooling from temperatures near or above the Curie point (typically 200–400 °C), a characteristic temperature above which a ceramic element shows no sign of piezoelectric activity. In practice, the operating temperature should be limited to about half of the Curie point. Ceramics can have their properties tailored to fit specific applications. In addition to the ceramics, piezorubber and piezoelectric polymers such as polyvinylidene fluoride (PVDF) have come into use. Also, a matrix material called "1–3 composites" (because of the orientation direction of its small PZT rods) is coming into use for large transducer systems.

Other piezoelectric materials that are sometimes used are these: Lithium sulfate ($LiSO_4 \cdot H_2O$) is used for hydrophones; tourmaline has been used in the calibration of hydrostatic pressure devices and in the measurement of blast wave pressures; lead metaniobate ($PbNb_2O_6$) has a very high Curie temperature and a low mechanical *Q*, and is used for deep submergence hydrophones; lead titanate ($PbTiO_3$) has a large hydrostatic piezoelectric constant, which allows this material to be used as a volume expander. A volume expander means that all surfaces of the material expand in phase, allowing simple hydrophone mounting materials to be used.

2.2 Piezoelectric Ceramics

The most widely used ceramics for underwater transducers are made from lead zirconate titanate (PZT). The materials used for ceramics are polycrystalline, which have Curie temperatures between about 200 and 400 °C. Ceramic materials are quite variable in their properties, are difficult to maintain consistency between batches and manufacturers, and have nonlinear properties, and their properties change with age. Even with these problems, ceramics are the predominant type of piezoelectric material used today. A good description of the properties of piezoelectric materials can be found in Wilson (1985). The technology of barium titanate (B) and lead zirconate titanate (PZT) has been relatively stable so that their physical properties can be listed with some degree of confidence. A list can be obtained from any manufacturer or from many of the references given above.

Because the U.S. Navy procures ceramics for their sonar systems, a standard was developed for obtaining reproducibility of transducer characteristics from one manufacturer to another. The military specification is known as MIL-STD-1376(SHIPS), and classifies ceramics in this way:

Classification. The ceramic shall be of the following types:

Type I—Hard lead zirconate-titanate, such as PZT-4, with a Curie temperature equal to or greater than 310 °C.

Type II—Soft lead zirconate-titanate, such as PZT-5, with a Curie temperature equal to or greater than 330 °C.

Type III—Very hard lead zirconate-titanate, such as PZT-8, with a Curie temperature equal to or greater than 290 °C.

Type IV—Barium titanate with additives and a Curie temperature equal to or greater than 100 °C.

Type V—A high dielectric constant material similar to type II.

Type VI—An extremely high dielectric constant material, such as PZT-5H, similar to type II.

The type I compositions are hard materials developed for high-power projectors. Type I has a high electric field, high coupling coefficient, and low dielectric loss properties under high drive fields. The low dielectric loss is important so that heating effects are minimal. Type III materials provide high stability for high drive levels and very low dielectric losses. Types II, V, and VI are soft materials developed for hydrophones and have high dielectric constants with high sensitivity, but also have high dielectric losses. Type VI has an extremely high dielectric constant and is often used in very high-frequency medical transducers.

The words hard and soft refer to how difficult it is to depolarize the ceramic. Unless the manufacturing process is well controlled, the values for a particular ceramic may vary as much as 10%. A solution to the varying properties from different batches and manufactures is to purchase all of the ceramic that will be needed, plus spares, from a single batch.

The piezoelectric constant *d* expresses the ratio of the field developed along an axis to the applied stress. It quantifies the movement (motion) of the ceramic when a voltage is applied. A high value of *d* is looked for in ceramic used for a high-power projector. The piezoelectric constant *g* expresses the ratio of strain developed along an axis to the applied field. It quantifies the output voltage of the ceramic when a motion, or pressure, is applied. The product of the two piezoelectric strain coefficients, *g* and *d*, is a very useful measure of "goodness" per unit volume of a piezoelectric material when minimization of the hydrophone internal noise is a major design criterion. Lists of piezoelectric properties can be obtained from any manufacturer or in Wilson (1985).

2.3 Composite Ceramics

Considerable effort has been directed to the development of new composite configurations of the PZT ceramics with nonactive materials. Piezorubber is composed of ceramic particles imbedded into a rubber matrix. It is very flexible, is used as hydrophones, and is called a 0-3 composite material. The 0 means that the ceramic particles do not touch one another, and the 3 means that it is used in the thickness mode. Another new material is composed of many small parallel ceramic rods imbedded vertically in a matrix material. These are called 1-3 composites because the ceramic is connected in one direction, and it is used in the thickness direction. This 1-3 composite is available in large sheets and can be used as a projector as well as a hydrophone. Large-area sensors can be used because of the minimal planar vibration modes and the capability to electroplate the surfaces. This material is finding many applications in underwater transducers.

2.4 Piezoelectric Polymers

Large piezoelectric effects are present in the polymer material polyvinylidene fluoride (PVDF or PVF_2). When the material is stretched along one axis and polarized, it becomes a very useful sensor. It is available in sheets with electrodes on the surfaces. For hydrophone uses, thicknesses of the order of 500 μm are used. PVDF has the advantages of flexibility, an impedance close to that of water, and a high hydrophone sensitivity relative to thickness. It is a poor transmitter and can be used only up to a temperature of about 100 °C. PVDF can be used where a very thin sensor is needed and where complex shapes can be etched into its surface. There are many other applications where a continuous sensor material that does not have any significant planar vibration modes can be used.

2.5 Magnetostrictive Materials

The magnetostrictive effect occurs when a magnetic field is applied to a material and causes a strain that is a nonlinear function of the applied field. The most important magnetostrictive materials are nickel and Terfenol-D. Nickel was used in the 1930s and 1940s in the construction of sonar projectors and hydrophones. Its use diminished because of its low electromechanical coupling coefficient, which relates the amount of kinetic energy available in the material to its total energy. Recently, the development of Terfenol-D by Clark (1980) has made available a high electromechanical coupling coefficient and a high electromagnetic strain. This material appears to have a chance to replace ceramics in some sonar transducers, especially in low-frequency projectors. A combination of Terfenol-D and ceramics in a single transducer has some interesting properties as described by Butler *et al.* (1990).

3. TESTING OF UNDERWATER TRANSDUCERS

To be able to use transducers properly, certain characteristics of the transducer and arrays of transducers must be measured and

Table 1. Calibration formulas.

Calibration test	Abbr.	Eq. No.	Formula	Units
Free-field voltage sensitivity	FFVS, M	2	$M_s - E_s + E_x$ $E_x - E_t - \mathrm{TVR}_s + 20 \log D$	dB re 1 V/μPa
Transmitting voltage response	TVR, S	3	$E_s - M_s - E_t + 20 \log D$	dB re μPa/V @ 1 m
Transmitting current response	TCR	6	$E_s - M_s - I_t + 20 \log D$ $\mathrm{TVR} + 20 \log(\lvert Z \rvert)$	dB re μPa/A @ 1 m
Transmitting power response	TWR	7 8 14	$E_s - M_s + 20 \log D - 10 \log(e_t i_t \cos \phi)$ $\mathrm{TVR} + 10 \log(R_p)$ $170.8 + \mathrm{DI} + 10 \log(\mu)$	dB re μPa/W @ 1 m
Source level	SL	12 13 15	$\mathrm{TWR} + 10 \log (P_t)$ $\mathrm{TVR} + E_t$ $170.8 + 10 \log (P_t) + \mathrm{DI} + 10 \log(\mu)$	dB re μPa @ 1 m
Equivalent hydrophone thermal noise	SPL_n	24	$10 \log(R) - M - 198$	dB re 1 V/μPa
Equivalent parallel resistance of projector	R_p	9	$\lvert Z \rvert/\cos \phi$	Ohms
Efficiency	μ		$10 \log(P_t/P_0)$	
Beamwidth (−3 dB)	BW (−3 dB)	16 18	Circular piston $60\ \lambda/d$ Continuous line $50\ \lambda/d$	Degrees
	BW (−10 dB)	20	1.8 BW (−3 dB)	
	BW (null)	21	2.3 BW (−3 dB)	
Directivity index	DI	17 19	Circular piston $10 \log(4\pi \times \text{area}/\lambda^2)$ Continuous line $10 \log(2L/\lambda)$	
Cavitation threshold intensity	I_c	27	$0.15[P_c(0) + h/33]^2$	W/cm^2
Cavitation threshold pressure at zero depth	$P_c(0)$	26	$0.000\,25f^2 + 0.045f + 1$	Atmospheres

understood. Equally important is knowing the various tests that can be performed on transducers and how to calculate the transducer parameters; see ANSI (1972). The term "transducer" as presented here implies that the device can be used either as a projector (transmitter) or as a hydrophone (receiver). When the term "projector" or "hydrophone" is mentioned, the characteristic being discussed will only apply to that type of device. For example, the directivity pattern of a transducer will be the same whether it is used as a projector or as a hydrophone. Efficiency is only associated with a projector. For quick reference, a tabulation of calibration formulas is presented in Table 1, and a definition of variables is listed in Table 2.

3.1 Calibration Formulas

The *free-field voltage sensitivity* (FFVS) of a hydrophone (or receiver) is the open-circuit electrical output of the hydrophone per

Table 2. Definition of variables used in Table 1.

E_s	Measured open-circuit voltage of standard hydrophone (dB)
E_x	Measured open-circuit voltage of test hydrophone (dB)
E_t	Transmit voltage (dB)
e_t	Transmit voltage (V)
M_s	FFVS of standard hydrophone (dB)
D	Separation distance (m)
I_t	Transmit current (dB)
i_t	Transmit current (A)
ϕ	Voltage/current phase angle (deg)
DI	Directivity index (dB)
P_t	Input power (W)
P_0	Output acoustic power (W)
R	Equivalent series resistance of hydrophone (Ω)
d	Diameter of piston (m)
L	Length of line source (m)
λ	Wavelength in water (m)
h	Depth
f	Frequency (kHz)

unit free-field pressure. The symbol M is typically used and was derived from the word "microphone," the air equivalent of a hydrophone. The open-circuit voltage measurement is made with a preamplifier with a high input impedance. The free-field pressure is what would exist at a given location without the presence of reflections. The easiest method to obtain the FFVS is by comparing the unknown hydrophone with a standard hydrophone when placed into the same acoustic field (usually a pure tone burst of sufficient time). A standard hydrophone will typically have a table of decibel signal level versus frequency. The signal level received by the hydrophone is determined by measuring the level within a time-gated portion of the signal that corresponds to the direct path. This way, interference due to reflected pulses are ignored, and a high signal-to-noise ratio is obtained. The FFVS, M, is given by

$$\text{FFVS(dB)} = M_x\text{(dB)} = M_s - E_s + E_x, \tag{2}$$

where M_s is the FFVS (in dB) of the standard hydrophone, E_s is the measured open-circuit voltage (in dB) of the standard hydrophone, and E_x is the measured open-circuit voltage (in dB) of the hydrophone being calibrated. The subscript x denotes the unknown or test hydrophone, and the subscript s the standard hydrophone. All units are decibels referenced to 1 V per micropascal (dB re 1 V/μPa). Sometimes this is called the receive voltage sensitivity (RVS).

The *transmitting voltage response* (TVR) of a projector is the transmitted free-field pressure produced for a 1-V signal applied across the electrical terminals of the transducer and referenced to a distance of 1 m. The symbol S is commonly used and stands for "speaker." The measurement is made by monitoring the output voltage of a standard hydrophone placed in the acoustic far field (defined below) of the projector. The TVR, S, is given by

$$\begin{aligned}\text{TVR(dB)} &= S_x\text{(dB)}\\ &= E_s - M_s - E_t + 20 \log D,\end{aligned} \tag{3}$$

where E_s is the measured open-circuit voltage (in dB) of the standard hydrophone, M_s is the FFVS (in dB) of the standard hydrophone, E_t is the projector transmit voltage (in dB), and D is the separation distance in meters. The units are decibels referenced to 1 micropascal per volt at 1 meter (dB re μPa/V @ 1 m).

The *far-field distance* x from a square or circular transducer is conservatively defined as

$$x = \text{area}/\lambda, \tag{4}$$

where λ is the wavelength in water. The far-field distance x from a line or thin cylinder is conservatively defined as

$$x = L^2/\lambda, \tag{5}$$

where L is the line length.

The *transmitting current response* (TCR) of a projector is the transmitted free-field pressure produced for a 1-A signal applied to the electrical terminals of a projector and referenced to a distance of 1 m. The measurement is made by monitoring the output voltage of a standard hydrophone placed in the far field of the projector. The TCR is given by

$$\text{TCR}_x\text{(dB)} = E_s - M_s - I_t + 20 \log D, \tag{6}$$

where E_s is the measured open-circuit voltage (in dB) of the standard hydrophone, M_s is the FFVS (in dB) of the standard hydrophone, I_t is the projector transmit current (in dB), and D is the separation distance in meters. The units are decibels referenced to 1 micropascal per ampere at 1 meter (dB re μPa/A @ 1 m).

The *transmitting power response* (TWR) of a projector is the transmitted free-field pressure produced for a signal of 1 W applied to the electrical terminals of a transducer and referenced to a distance of 1 m. The TWR is given by

$$\begin{aligned}\text{TWR} &= E_s - M_s + 20 \log D\\ &\quad - 10 \log(e_t i_t \cos \phi),\end{aligned} \tag{7}$$

where E_s is the measured open-circuit voltage (in dB) of the standard hydrophone, M_s is the FFVS (in dB) of the standard hydrophone, D is the separation distance in meters, e_t is the projector transmit voltage, i_t is the projector transmit current, and ϕ is the voltage/current phase angle. The units are decibels referenced to 1 micropascal per watt at 1 meter (dB re μPa/W @ 1 m).

Using Eqs. (3) and (7), the TWR can be written as a function of TVR as

$$\mathrm{TWR} = \mathrm{TVR} + 10\log(R_p), \tag{8}$$

where R_p is the equivalent parallel resistance of the projector and is given by

$$R_p = |z|/\cos\phi. \tag{9}$$

It is evident that either a standard hydrophone or a calibrated projector is required to obtain the transmit or receive response of an unknown transducer. Hydrophones that have calibration response data can be purchased, or can be sent to a calibration facility.

An absolute method to calibrate transducers without the need of a standard is called a *reciprocity* calibration. The reciprocity calibration allows an absolute calibration without the need of a transducer with a known response (see ACOUSTICAL INSTRUMENTS). It requires three transducers, one of which is a projector, one a hydrophone, and one which is reciprocal. A reciprocal transducer has the ratio of its receiving sensitivity M (when acting as a hydrophone) to its transmitting current response S (when acting as a projector) equal to a constant called the *reciprocity parameter, J*:

$$J = \frac{M_e}{S_i} = \frac{\mathrm{FFVS}}{\mathrm{TCR}}. \tag{10}$$

For spherical waves,

$$J_s(\mathrm{dB}) = 20\log(2/\rho c) = 354 + 20\log f, \tag{11}$$

where $\rho = 1000$ kg/m^3 is the density of water, $c = 1500$ m/s is the sound velocity, and f is the frequency. All units are MKS. Fortunately, most ceramic transducers are reciprocal. Details of this calibration method will not be presented here, as there are several excellent discussions available; see Albers (1965) or Bobber (1988).

The phase response of a projector is not of concern in most calibrations. When measuring phase response, the best reference is a standard, omnidirectional hydrophone having a constant sensitivity and operating well below its resonance frequency. In this region the phase of its transduction is constant and well suited as a reference. Comparison phase measurements are obtained by reference to the phase of the drive voltage.

The *source level* (SL) of a projector is the transmitted free-field pressure produced by a signal of the applied power or applied voltage to the electrical terminals of a projector and referenced to a distance of 1 m. The SL is given by

$$\mathrm{SL(dB)} = \mathrm{TWR} + 10\log P_t, \tag{12}$$

$$\mathrm{SL(dB)} = \mathrm{TVR} + E_t, \tag{13}$$

where P_t is the projector input power (in watts) and E_t is the projector input voltage (in dB). The units are decibels referenced to 1 micropascal at 1 meter (dB re μPa @ 1 m).

The power delivered to a projector is dissipated in either the electrical or the mechanical resistance within the transducer, or it is radiated into the water. The *efficiency* (η) of a projector is the ratio of the total acoustic output power to the total electrical input power. The TWR, Eq. (7), and SL, Eq. (12), can also be written as functions of efficiency,

$$\mathrm{TWR(dB)} = 170.8 + \mathrm{DI} + 10\log\eta, \tag{14}$$

$$\mathrm{SL(dB)} = 170.8 + 10\log P_t + \mathrm{DI} + 10\log\eta, \tag{15}$$

where P_t is the input power, DI is the directivity index (see Sec. 3.2), and the quantity 170.8 dB accounts for the ρc of the medium and the reference of 1 μ Pa. The directivity index is very difficult to measure because it involves the total power radiated and requires the measurement of directivity patterns in several planes (Sec. 3.2 shows the calculation of DI). The DI needs to be accurate to 0.1 dB for a good efficiency calculation, which for most projectors ranges from 50% to 80%.

3.2 Calculated Parameters

One of the major parameters of a transducer or transducer array is the directivity pattern (often called beam pattern), which is the response level of the transducer as a function of the incident angle of the sound. A transducer will have the same directivity pattern whether operating as a projector or a hydrophone. It is a function of the aperture and frequency and is used to calculate the power and efficiency parameters of the transducer. A high directivity indicates that the acoustic power of a projector is concentrated into a small region, and for a hydrophone high directivity allows high imaging resolution and reduces the influence of ocean isotropic noise. The *directivity index,* DI, of a transducer is a very important parameter that quantifies the sharpness, or directionality, of the pattern and is closely associated with the beamwidth. Approximate equations (Stansfield, 1990) for the half-power (−3 dB) beamwidth BW and directivity index DI of a circular piston in an infinite baffle are

$$\mathrm{BW} \approx 60\lambda/d, \tag{16}$$

$$\mathrm{DI} = 10\log(4\pi \times \mathrm{area}/\lambda^2). \tag{17}$$

The approximate equations for the half-power (−3 dB) beamwidth BW and directivity index DI of a continuous line are

$$\mathrm{BW} \approx 50\lambda/L, \tag{18}$$

$$\mathrm{DI} = 10\log(2L/\lambda). \tag{19}$$

The beamwidth of a square radiating surface uses Eq. (18), where L will be the side of the square, and Eq. (17) for the DI. Parameters for rectangular and elliptical radiating surfaces can be found in Albers (1965) or Stansfield (1990). The beamwidths at the −10-dB level and between the first nulls can be obtained by using a multiplying factor:

$$\mathrm{BW}(-10\ \mathrm{dB}) \approx 1.8\mathrm{BW}(-3\ \mathrm{dB}), \tag{20}$$

$$\mathrm{BW(btwn.\ nulls)} \approx 2.3\mathrm{BW}(-3\ \mathrm{dB}). \tag{21}$$

These equations apply for dimensions greater than one wavelength of sound in water.

The *equivalent thermal-noise pressure level* is important in determining the lowest signal that can be developed in a hydrophone. This noise level of the hydrophone can be associated with an equivalent sound-pressure level in the ocean. The open-circuit voltage E_n developed by thermal agitation in a 1-Hz band is called the Johnson noise,

$$E_n = \sqrt{4KTR}, \tag{22}$$

where K is the Boltzmann gas constant (1.38×10^{-23} J/K), T is the absolute temperature in kelvins, and R is the equivalent series resistance of the electrical impedance of the transducer in ohms (in general, this is a function of frequency). The equivalent noise sound-pressure level, SPL_n, is

$$\mathrm{SPL}_n = E_n/M, \tag{23}$$

and written in decibel form for a temperature of 15 °C is

$$\mathrm{SPL(dB)} = 10\log(R) - M - 198. \tag{24}$$

For example, a 1000-Ω impedance hydrophone with a FFVS of −200 dB re 1 V/μPa will have an equivalent thermal-noise sound-pressure level of 20 dB re 1 V/μPa.

3.3 Cavitation

The onset of cavitation occurs when the motion of the radiating face creates a partial vacuum that is greater than the vapor pressure of the water. The cavitation threshold is the point where the power applied to the projector is sufficient to start the formation of bubbles at its face. An empirical curve (Urick, 1975) that shows the frequency dependence of the cavitation threshold P_c, using continuous-wave (cw) data in fresh water at atmospheric pressure, is

$$P_c(0) = 0.00025f^2 + 0.045f + 1 \quad \text{(atmospheres)}, \tag{25}$$

where $P_c(0)$ is the cavitation threshold at zero depth and f is the frequency in kHz. This is approximately valid up to frequencies of a few hundred kilohertz. As the depth of operation is increased, the cavitation threshold increases as

$$P_c(h) = P_c(0) + h/10 \quad \text{(atmospheres)}, \tag{26}$$

where $P_c(h)$ is the cavitation threshold at depth h and h is the depth in meters. The cavitation threshold can be approximately expressed as occurring when the acoustic intensity at the radiating face of the transducer is equal to

$$I_c = 0.15[P_c(0) + h/10]^2\ \mathrm{W/cm^2}. \tag{27}$$

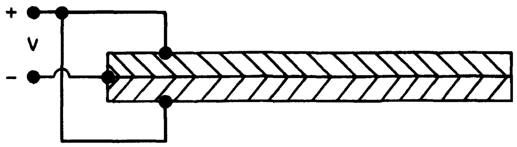

FIG. 2. Schematic of a basic bilaminar flexural disk transducer (Wilson, 1985).

In practice the intensity can be 2 to 3 times higher before cavitation will affect the source level. A value of 0.3 is used by Wilson (1985). The intensity at the surface of a transducer or array can be calculated by multiplying the acoustic output power by the area of the transducer. The cavitation threshold is sensitive to temperature, dissolved-air content, frequency, acoustic pulse length, depth, and acoustic hot spots on the transducer face. Quantifying the onset of cavitation is difficult and, if important, should be determined experimentally. See Urick (1975) and Stansfield (1990) for more information.

4. LOW-FREQUENCY TRANSDUCERS (10–5000 Hz)

The design of transducers and arrays capable of radiating power at frequencies less than about 1 kHz has the basic requirement of producing large displacements efficiently. The size and weight of the transducer becomes an important issue at these low frequencies, since the source level depends upon the amount of volume of water that is moved. Radiation loading, which determines the amount of sound radiated, is also typically small because the size of the transducer is small relative to a wavelength. At 1 kHz the wavelength of sound in water is 5 ft; at 100 Hz it is 50 ft. Fortunately, for most low-frequency active transducers, the output power requirements are not as high as those for transducers operating at higher frequencies. To design high-power transducers at low frequencies in a moderate size, they need to be less stiff than a ceramic thickness vibrator. To accomplish this more compliant transducer design, use is made of a mechanical amplification such as a lever. This provides a mechanical impedance transformer resulting in a lower stiffness. One technique for doing this is to use flexural motion of plates and shells that have a lower stiffness and a lower mechanical resonance frequency for the same overall dimensions of the structure. Other methods involve the use of a gas-filled enclosure as the stiffness element of electromagnetic or hydraulic systems. The general characteristics of these designs are discussed in the following.

4.1 Flexural Disk

An example of a transducer that involves flexure is the flexural plate transducer; see Fig. 2. The basic principal of a flexural disk is that if a force is applied to one face of a thin metal disk, it will be displaced at the center, or flex, putting one side into compression and the other side into tension. A bilaminar disk will have two piezoelectric ceramic disks bonded together. An electrical field across the ceramics will result in an expansion in the ceramic disk that has the field and the polarization in the same direction, and a compression in the ceramic that has the field and the polarization in the opposite direction. The result is to cause a bending of the pair of disks converting a small displacement in the ceramic into a larger displacement of the surface radiating into water. For more efficiency, a double bilaminar disk transducer can be used, which has two faces radiating in opposite directions to act as a volume expander, Fig. 3. Basic design pa-

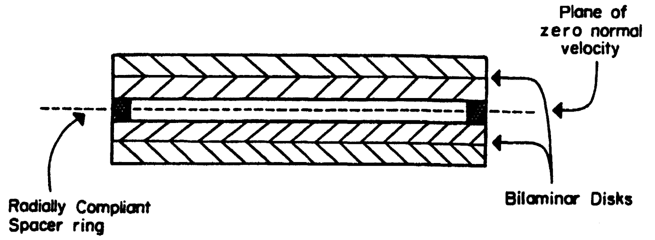

FIG. 3. Double bilaminar disk (Wilson, 1985).

rameters were derived by Woollett (1960) and can serve as further reading.

A metal plate can be inserted between the two ceramic disks, a trilaminar design, or one of the ceramic disks can be replaced with a metal plate. Typically, the ceramic is placed on the outside of the metal so that under a hydrostatic load the ceramic will be in compression so that it will not fracture. Limitations on hydrostatic pressure can be overcome by providing air pressure compensation. This, however, adds complexity and affects performance.

Flexural disk transducers are good as small, modest-power sources at low frequencies. Mechanical values of Q typically range between 10 and 15. They are relatively low cost because of their simple construction, and hence will be an economical and compact solution for not too demanding an application. Flexural disks are extensively used as low-frequency hydrophones because of their low cost and high sensitivity.

4.2 Flextensional Transducer (300–3000 Hz)

This device also utilizes flexure motion, but instead of using the ceramic itself in flexural motion, it uses a mechanical gain to transform the thickness-mode vibration in the ceramic into a larger displacement. The basic features of the flextensional transducer are illustrated schematically in Fig. 4. A stack of ceramics operates in the thickness mode and is connected to a surrounding shell. When the stack expands, the major axis of the shell moves outward, and the minor axis moves inward. This produces a large volume displacement, transforming the high-impedance, small motion of the stack into a lower-impedance, larger movement of the shell. Several different types of design use this mechanism and are divided into classes of flextensional transducers. Flexural transducers primarily are designed to resonate in the frequency range of 300 to 3000 Hz, and generally have higher power outputs than the flexural disks, although at a higher cost (Stansfield, 1990).

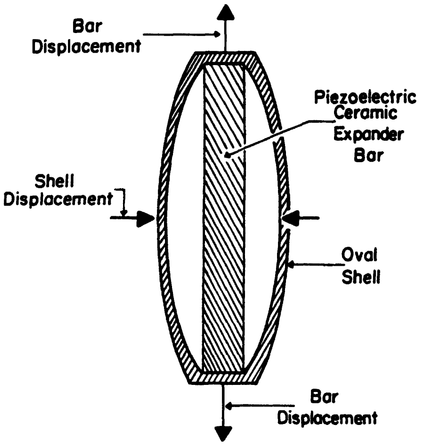

FIG. 4. Schematic of a flextensional transducer (Wilson, 1985).

A Class I flextensional transducer is a three-dimensional structure created by rotating about an axis along the stack; see Fig. 5. If it is rotated about an axis normal to the stack, a ring-shell projector is obtained, a Class V flextensional. The ceramic in a Class V flextensional is also extended in the third dimension and forms either a disk, which drives the shell by radial stress in the disk, or a ring around the circumference joining the two half shells.

Another configuration is an elliptical cylinder generated by extending the figure out of the plane of the sketch and capping it on both top and bottom. This is called a Class IV flextensional transducer and has been studied more thoroughly than other classes. Design characteristics have been calculated by analytical and by finite-element analysis; however, the design depends on a number of parameters and does not lead to any simple design equations. The resonance frequency and the maximum hydrostatic pressure that an elliptical shell can withstand depend on the major and minor axis, wall thickness, and material properties of the shell, with the stack having a lesser influence. A rule of thumb is that eccentricity should not be greater than about 3. The use of a glass-reinforced plastic shell helps achieve a low resonance frequency and wide bandwidth.

The size of flextensional transducers is generally much less than a wavelength in water at their resonance frequency; hence, they tend to have an omnidirectional directivity pattern in the plane perpendicular to their axis. Values of mechanical Q down to 2 are achievable. Compared to a flexural disk of the same size, the Q of a flextensional

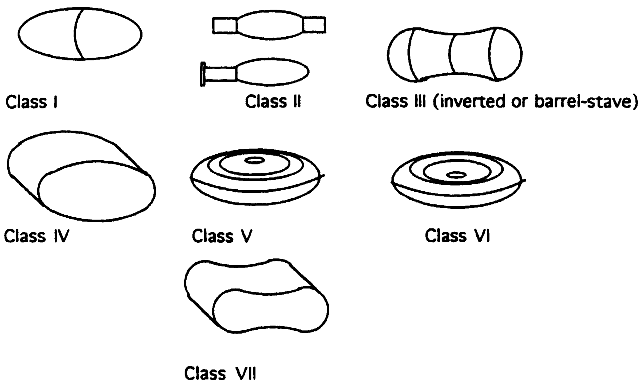

FIG. 5. Flextensional transducer classes (McCollum *et al.*, 1992).

transducer can be lower by a factor of 2–4. Also, a flextensional can be smaller than a flexural disk when hydrostatic pressure is the limiting requirement.

A compressive prestress must be applied to the stack in order to achieve high power output. It is usually accomplished by squeezing the minor axis, which expands the major axis, then inserting the stack into the shell. The relaxation of the shell applies the proper preload needed for the stack. It must also be realized that the hydrostatic pressure, which applies force to the shell, will also compress the minor axis and reduce the preload on the stack. The preload must be sufficiently designed so that hydrostatic pressure does not adversely affect its performance. Finite-element analysis is needed to detail the static stresses in the shell under hydrostatic pressures. Material fatigue during high-power operation must also be taken into account.

In summary, flextensional transducers constitute useful sources within the band of about 300–3000 Hz. They can provide significant power from a compact size, can operate at high hydrostatic pressures, and have wide bandwidths, especially when compared to other designs such as the air-backed flexural disk. The cost, however, is higher. The design process needs considerable care, using finite-element analysis to account for the complex interactions between parameters.

4.3 Free-Flooded Ring

A cylindrical ring that is open to the water (a free-flooded ring) can be used for low-frequency projectors that are independent of hydrostatic pressure. This transducer operates in the ring, or circumferential, mode of vibration. The radiation from the inner and end surfaces of the ring oppose that from the outer surface and can lead to partial cancellation of the radiation. It is possible, however, to design the height of the ring such that cancellation effects are not too great. Because it can be free flooded, no pressure compensation mechanisms are required. Since the resonance frequency is dependent on the diameter of the ring, a resonance of 1000 Hz would result in a ring of about 1 m diameter, which is impractical. The free-flooded ring can be typically used for frequencies above about 4000 Hz.

4.4 Mechanical Transducer

In the frequency range of 1–50 Hz an electric motor can be used to drive pistons to sufficiently large amplitudes of vibration. The frequency can be changed by varying the speed of the motor; however, it is not easy to change the amplitude of the oscillation, which limits the versatility of this approach. They are very useful when high power is required at frequencies below about 20 Hz.

4.5 Electromagnetic, Moving-Coil Transducers

Moving-coil electromagnetic transducers can be considered at audio frequencies. They are used as loudspeakers in air because large displacements are easily generated. In water they can be used as low-power, wide-bandwidth sources from 40 Hz to 20 kHz. The J9 transducer is a standard projector that has a moving coil mounted on a very compliant suspension system. It is necessary to balance the air pressure within the transducer to the external water pressure by means of an air-filled collapsible rubber bag. Beyond a depth of about 25 m an air compensation device is needed. The maximum power for the J9 is 20 W at frequencies above 200 Hz, but the efficiency is less than 1%. This device requires careful handling so that the suspension is not damaged. Also, the calibrated response may vary with drive level. It should be calibrated every time it is used.

4.6 Resonant Bubble Transducer

A gas-filled bag can be placed at the face of a moving-coil transducer to increase the radiation loading and increase the efficiency near resonance. The broadband low efficiency of a moving-coil transducer is transformed into a narrower-bandwidth, higher-efficiency transducer. The compliance in the bubble, or bag, provides a compact low-frequency source of modest power, narrow bandwidth, and a resonance frequency that varies with depth.

4.7 Helmholtz Resonator

The Helmholtz resonator is another resonant device that can be used as a low-frequency source. This resonator uses a rigid enclosed volume with a short neck of specified length and cross-sectional area. The outer end of the neck is exposed to the medium. The compliance of the cavity volume resonates with the mass of the fluid in the neck to form a simple mass–spring system. A transducer is placed in the wall of the cavity to excite the resonance. The resonance frequency is not dependent on density but is proportional to the sound velocity of the medium. Therefore, the resonance frequency of a Helmholtz resonator in water is about four times higher than in air. The mechanical Q_M is typically well above 20. The best way to reduce the resonance frequency is to maximize the cavity volume to a limit where the dimensions remain less than $\lambda/4$. One method of exciting the cavity is to place a flexural disk at one end of the cavity; another is to use ceramic rings in the wall of a cylindrical cavity. Cavitation can occur within the cavity at high source levels.

The Helmholtz resonator is a potential solution for a compact low-frequency projector that can operate down to great depths. In practice, to achieve rigid walls in a water environment requires thick walls, hence extra weight, and the narrow bandwidth and low efficiency have limited their usefulness. The use of this device, some specific design parameters, and their effects on the acoustic performance are addressed by Henriquez and Young (1980) and Kinsler *et al.* (1982).

4.8 Hydroacoustic Projector

The hydroacoustic projector, developed by Bouyoucos (1975), is a device that uses hydraulic fluid at high pressure to drive a flexing disk or a piston and is capable of large amplitudes of motion. A high-pressure hydraulic fluid flow is controlled by an electromechanical shuttle valve that causes the pressure oscillations in the acoustic drive cavity. This, in turn, exerts pressure oscillations on the flexural disks, which then radiate into the water. The electromechanical valve is typically a ceramic stack and provides control for the transducer. Power is derived from the high-pressure hydraulic fluid. The flexural disks, plus the various components, are designed to resonate over a fractional bandwidth of about 20%. It is an effective high-power source at low frequencies where weight and size are not critical. It is best used at frequencies below a few kilohertz and is not suitable for use as a hydrophone.

4.9 Spherical-Shell Transducer

A spherical shell composed of piezoelectric ceramic is another design for achieving a low-resonance frequency with a reasonable value of *Q*. One approach is to fabricate a large sphere with a number of triangular-shaped ceramic plates. These plates are cemented together into a mosaic to form a spherical shape, which radiates an omnidirectional pattern. A 30-in. diameter sphere has been fabricated with a usable frequency range from 0.5 to 2.0 kHz and a minimum efficiency of 75%.

4.10 Impulsive Transducer

Explosive sources are the most common impulsive source. Urick (1975) gives a good description of their use as underwater sound sources. These are commonly small charges of explosive materials ranging from a few grains to a few pounds. High velocity of detonation explosives such as TNT (>15 000 ft/s) are preferred over other materials such as black powder (0.1–1 ft/s), which has an explosion process called deflagration, or burning. The detonation introduces a shock wave normally followed by a series of pressure "bubble" pulses. Only one or two bubble oscillations are of practical significance. About 40% of the total chemical energy is radiated as acoustic energy. Most of the acoustic energy is concentrated between 5 and 1000 Hz. Explosive charges are easily portable and yield a short, omnidirectional, high-power, broadband pulse. The disadvantage is that explosives cannot be made repeatable, give a high background of reverberation, and cannot be coded for sonar applications.

The *sonar thumper* is an impulsive source generated by the repulsion of an aluminum plate by a large coil imbedded in a plastic matrix; see Albers (1965). A pulse of current in the coil induces eddy currents in the aluminum plate, causing it to be repelled by the coil. A large pulse of short duration (a capacitive discharge) needs to be passed through the coil. Although the pressure pulses are smaller than using an explosive, the thumper can be triggered and synchronized with a recorder system.

An *underwater spark source* acts as a small impulsive source but can be repeated for a number of pulses. A capacitive discharge across a gap between a pair of electrodes in the water generates the acoustic noise. It is a very simple transducer but has the disadvantage that the condition of the electrodes changes with use, which varies the pressure level of successive sparks. Peak pressures of 240 dB re 1 μPa at 1 m are possible. An *air gun* impulse source is created when high-pressure air is momentarily vented to the surrounding water to produce an air-filled cavity, which then collapses and radiates a transient pulse.

5. LONGITUDINAL VIBRATOR TRANSDUCERS (3 kHz to >100 MHz)

Longitudinal vibrators represent the standard transducer type used for a broad frequency range from about 3 kHz to above 100 MHz. *Tonpilz* transducers, which are a composite of ceramic and "head" and "tail" masses, predominate up to a frequency of about 100 kHz. Near the lower frequency of 3 kHz, the length and mass of the transducer becomes very large and other types become useful. Near 100 kHz the tonpilz transducer becomes very small, and it is more convenient simply to use the thickness vibration mode of a ceramic. The frequency range of 5–50 kHz is mainly used for weapon and detection sonars. From about 50 to 600 kHz, transducers are used for mine hunting, ocean bottom profiling, and fish finders. Frequencies from 600 kHz to 3 MHz are used for in-water high-resolution imaging arrays. Above about 3 MHz, transducers are used for medical ultrasound imaging and nondestructive testing. Ultrasonic microscopes operate up to the gigahertz frequency range.

5.1 Tonpilz Transducers

The tonpilz transducer is a longitudinal vibrator and gets its name from a composition of two German words, "sound" and "mushroom" or "singing mushroom." It is named, as can be seen, from its shape as shown in Fig. 6. These transducers are operated near resonance and typically are combined into arrays of many elements. Since arrays are often electrically phased to steer their directivity patterns to angles away from the normal to the plane of the array, the size of the radiating face needs to be about half of a wavelength in water. The tonpilz transducer has head and tail masses and a central ceramic (or magnetostrictive) motor section that are bonded together and put into compression by a central bolt. This creates a mass-spring-mass system that is a very efficient vibrator and radiator. The face of the head couples into water, and the tail has as light an acoustic load (such as air) as possible. A large radiated sound power can be achieved by this design.

There are four basic approaches to the tonpilz transducer design. The first approach is mentioned by Wilson (1985), who treats the transducer as a mass-spring-mass system. A second approach uses equivalent circuits or four-port networks as discussed by many authors. This is the most common method of tonpilz transducer design and requires electrical circuit programs or special codes to handle the complex, coupled equations. A third design approach is from Mc-

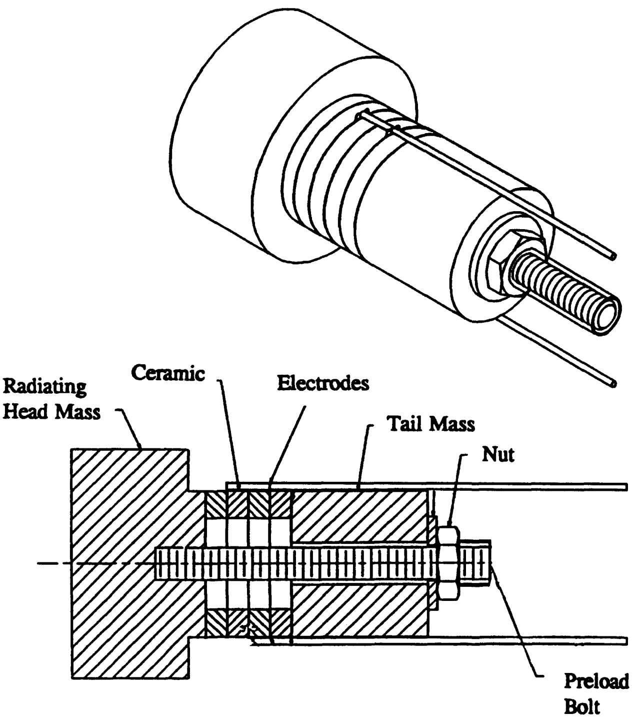

FIG. 6. Schematic of a tonpilz longitudinal vibrator.

Cammon and Thomson (1980), who treat the tonpilz transducer from an energy and waveguide basis, and a fourth approach uses finite-element analysis, which gives a very detailed vibration characterization, but is very time-consuming. Often the design is made by using a combination of the second and fourth approaches to generate an approximate design (within 10%) and then to detail the design. The first approach is not used very often but is more illustrative.

A tonpilz transducer can be used for frequencies from about 3 kHz to about 100 kHz. At frequencies above about 100 kHz, a thickness vibrator can be used. At 100 kHz a ceramic thickness vibrator made of a PZT-4 rod would have a length of 2 cm, and it would be less expensive. A tonpilz transducer, being a composite element, is used to reduce the overall length, make the stresses in the active material more uniform, make use of the piezoelectric material more effective, and reduce the mechanical Q. The radiating face typically has a larger area than the ceramic in order to increase the power output for a given strain in the ceramic and to obtain a better impedance match between water and the active transducer material. The head/ceramic area ratio is limited to about 3–10 by the need to avoid flexural resonances in the tonpilz head and to have a sufficient volume of ceramic.

It is typically assumed that all components of the element are small compared to a wavelength. The desire is to have the element face dimension be half of a wavelength in water so that multiple-element arrays can be phased for beam steering. The head mass should be kept light by using a low-density metal, such as aluminum or magnesium, and a minimal thickness determined by setting its flexural resonance at about twice the fundamental resonance. The ceramic stack design must consider the number of ceramics, dimensions, compliance, electric field limits, power-handling capability, electrical impedance, electromechanical coupling coefficient, and cost. For most cases the ceramic used for transducers that both transmit and receive is PZT-4 (Navy type I). If high acoustic fields are required, then projectors use PZT-8 (Navy type III) to minimize high-power heating effects. Designs typically use multiple stacked rings of 4 to 6 ceramics for low-impedance, high-power applications where the transducer must also receive. An even number of rings is used so that they can be connected in parallel with alternating polarity and electric fields. The maximum usable electrical field limit of ceramics is often a limit for the operation of the transducer. For PZT-4 (Navy type I), the maximum field is about 2400 V/cm (6 V/mil), and for PZT-8 (Navy type III) it is about 4000 V/cm (10 V/mil). The maximum voltage that can be applied to a transducer is equal to the field limit times the thickness of a single ceramic. For more information on the power-handling capabilities of ceramics, see Berlincourt *et al.* (1964).

To maximize the displacement of the head, the tail mass should be made as large as possible and have a mass 3 to 5 times that of the head. Typically a high-density metal such as tungsten, stainless steel, or brass is used. The in-water mechanical Q will typically have a value of 5 to 7; however, when assembled into an array, the Q will drop to a value of 2 to 5. A bolt is typically used between the head and tail to place a compressive load of 2000 to 6000 lb/in.^{2} on the ceramic so that high drive levels do not place the ceramics into tension. Placing the ceramics into tension may result in creating microcracks that will lead to failure.

The tonpilz transducer is capable of very high source levels, has an efficiency of (80–90)%, and is typically formed into arrays, which can form patterns capable of being steered to different angles. These arrays are used for imaging and locating objects. The cavitation threshold for the transducer or the array is calculated using Eq. (25). The intensity times the array face area gives a cavitation limited maximum power.

5.2 Thickness Vibrators

At frequencies higher than 100 kHz, solid ceramic shapes are typically used. These transducers can become a significant portion of a wavelength in size, and it is often necessary to account for vibration modes set up in the ceramic. The mathematics of analysis may become more complex, but the design itself has less flexibility than the low-frequency designs. It is at these frequencies that the broad-area transducers such as PVDF and the 1-3 PZT composites are becoming popular. These newer materials have greatly reduced or eliminated unwanted structural vibrations, allowing single transducers with a size of many wavelengths.

A thickness vibrator is a thin plate in which both the electrical field and the polarization are in the thickness direction. A *length expander* is a long thin bar with the electrical field parallel to the length. The polarization is also in the same direction as the electric field. To determine the fundamental resonance frequency for these vibration modes as well as for others, the equations are presented in Table 3 as a function of the *frequency constant* and the controlling dimension of the vibrator. This is a very useful table for determining the ceramic dimensions needed for a particular design. Fish finders, mine-hunting transducers, bottom profilers, ultrasonic medical transducers, and nondestructive testing transducers are typically fabricated by starting with a ceramic plate, dicing it into small elements, and then filling the gaps with a polymer. These are, in essence, 1–3 composite materials that have a specific design and application. Complex shape transducers using injection-molded 1–3 composite materials have been fabricated by the author using electroplating techniques. Complete arrays and large-area projector/hydrophone transducers are possible because of the mode decoupling achieved in a 1–3 composite.

When an air-backed transducer is excited by a pulse, little energy is lost and it rings at its resonance frequency. This impedance mismatch at the back of the transducer is desirable for continuous-wave operation; however, it is very undesirable for medical ultrasonic transducers. Lateral resolution in ultrasonic imaging is dependent on the transducer reproducing a pulse with little ringing present. Ringing can be reduced by using absorptive backings such as a tungsten-loaded epoxy. The penalty with using an absorptive backing to reduce ringing is a reduction in the sensitivity. A more current approach is to apply matching layers of suitable thickness (approximately quarter wavelength) and characteristic impedance between the ceramic and the medium. These layers permit efficient power transfer from the ceramic to water (the human body) without much loss in sensitivity. Typically a light acoustic backing will also be used. The use of matching layers to reduce ringing also results in a broader bandwidth. See Shung (1992) for more detail.

Table 3. Vibration modes for common transducer configurations.

					Frequency constant in kHz-m (kHz-in)			
Shape and electrode position	Vibration mode	Type of coupling	Direction of polarization	Controlling dimension	Type I PZT-4	Type III PZT-8	Type IV $BaTiO_3$	Type VI PZT-5H
Long, thin bar or strip	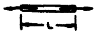 Longitudinal mode Length varies	k_{31}	↓	L	1.60 (63)	1.75 (69)	2.34 (92)	1.41 (56)
Plate or Thin disc	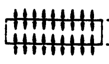 Thickness mode Thickness varies	k_{33}	↓	T	2.03 (80)	2.13 (84)	2.87 (113)	1.75 (69)
Thin disc	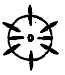 Radial mode Radius varies	k_p	↓	D	2.16 (85)	2.36 (93)	3.18 (125)	1.98 (78)
Thin walled tube	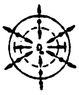 Radial mode Mean diameter varies	k_{31}	↖↑↗ ↙↓↘	D	1.02 (40)	1.07 (42)	1.47 (58)	0.89 (35)
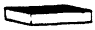 Shear plate	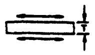 Thickness shear mode rectangular cross-section is distorted into a rhombus	k_{15}	←	T	1.32 (52)	1.33 (57)	1.59 (62)	1.10 (43)
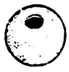 Thin walled sphere	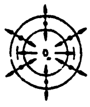 Radial mode Mean diameter varies	k_p	↖↑↗ ↙↓↘	D	1.72 (68)	1.83 (72)	2.46 (97)	1.47 (58)
Hemisphere	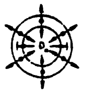 Radial mode Mean diameter varies	k_p	↖↑↗ ↙↓↘	D	2.08 (82)	2.26 (89)	3.02 (119)	1.85 (73)
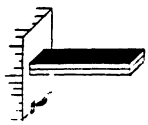 Bender transducer	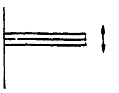 Flexural mode Free end moves at right angle to beam axis	k_{31}	↓ or ↑ ↑ ↓	$\frac{L^2}{t}$	0.51 (20)	0.56 (22)	0.74 (29)	0.46 (18)

5.3 Shaped Ceramic Transducers

Ceramic piezoelectric materials can be fabricated into a wide variety of shapes and sizes, and they lend themselves to designs for specific applications that would be difficult for single ceramic pieces. The shapes shown in Table 3 are an example.

Spherical transducers are typically the choice for omnidirectional patterns. Complete omnidirectionality can be obtained by using a spherical transducer that has a diameter smaller than a wavelength. One of the problems of a sphere is attaching leads to the inner electrode. For this reason two hemispheres are used, with the wires being attached prior to bonding the two pieces together. Small *cylindrical transducers* will also produce omnidirectional patterns and are often easier to fabricate. When used as a hydrophone, the cylinder dimensions are determined by the frequency range over which a flat response is desired. The upper limit of the flat range is slightly below the lowest resonance frequency. When operating as a projector, the fundamental resonance mode will produce the peak acoustic pressure level. There are three fundamental modes of vibration: length mode (a half-wavelength resonator), radial mode (the breathing mode, calculated using the mean diameter), and wall-thickness mode. Some coupling is present between these various modes of vibration. The coupling between the length and radial modes becomes very strong when the length is approximately three times the mean radius, and results in distorted patterns and frequency responses. This ratio is a design "forbidden zone" and should be avoided. See Hueter and Bolt (1955) for more detail. The cylindrical ceramic typically has its ends capped to shield the inner cylindrical surface from the acoustic medium. Since the outside and inside surfaces are electroded, and the polarization is radial, the electrical impedance is low. Also, the cylinder has a large radiating area in relation to the cross section of the ceramic, resulting in the mechanical impedance of the element being much lower than a solid ceramic of the same dimension. This permits a better acoustic match between the ceramic and the water.

A *bimorph* is a bender transducer (see Table 3), which flexes and has the capability for handling larger motions and smaller forces than single piezoelectric plates. This transducer is fabricated by bonding two shear-mode plates or two longitudinal-mode plates together in such a manner that when a voltage is applied, the plates will deform in opposite directions, resulting in a bending or twisting action. One example is the low-frequency flextensional transducer discussed earlier. The bimorph is typically a high-frequency transducer operating in air and is often used for ac and dc relays, audio alarms, tweeters, and micromanipulators. A shear-mode transducer has been used in shear accelerometers and for generating surface waves on structures.

6. OTHER TRANSDUCERS

Parametric transducers are used to create very narrow beamwidths at low frequencies. They rely on the nonlinear properties of water at high source levels. When two frequencies are generated at high amplitudes, intermodulation products are produced in the nonlinear medium such that the column of fluid in front of the transducer acts as a thin line array radiating in the direction of propagation. This characteristic is equivalent to an end-fire acoustic array. The low-frequency difference beam, called the secondary frequency component, has nearly the same beamwidth as at the primary frequency. The narrow beamwidth is achieved by use of a transducer that has a much smaller aperture than a conventional transducer. Even though the narrow beamwidth is very attractive, the parametric array is not typically used because the process is very inefficient.

In *fiber optic sensors* the variation in pressure in acoustic waves affects the optical properties of thin fiber optic cables. Compared to other hydrophones, it has high sensitivity and is insensitive to electromagnetic interference. In almost all of the standard configurations the sensing of the acoustic wave is accomplished in an optical detector that is sensitive to phase differences between light that has traveled a distance in fiber exposed to the sound and light that has traveled about the same distance but was not exposed to the sound. Mechanical noise pickup is a problem.

Chemical acoustic sensors result from phenomena in electrolytes that permit sensing acoustic pressure by measuring electrochemical potential. An acoustic wave that strikes an electrode where gas is being liberated by electrolysis causes an alternating component of the electrical potential. In another phenomenon, when a layer of electrolyte-filled nonconducting fiber surrounds an electrode, an electrokinetic effect in the fiber changes the electrode potential relative to the surrounding electrolyte. A microphone for low-frequency sounds called the solion infrasonic microphone uses a chemical reaction that depends on the flow of ions through an acoustic resistance placed in the electrolyte between the electrodes.

Works Cited

Albers, V. M. (1965), *Underwater Acoustics Handbook—II,* University Park: The Pennsylvania State Univ. Press.

ANSI (1972), *Procedures for Calibration of Underwater Electroacoustic Transducers,* New York: American National Standards Institute.

Berlincourt, D. A., Curran, D. R., Jaffe, H. (1964), "Piezoelectric and Piezomagnetic Materials and Their Function in Transducers," in: Warren P. Mason (Ed.), *Physical Acoustics,* Vol. IA, New York: Academic, Chap. 3, p. 250.

Bobber, R. J. (1988), *Underwater Electroacoustic Measurements,* Los Altos, CA: Peninsula Publishing.

Bouyoucos, J. V. (1975), "Hydroacoustic Transduction," *J. Acoust. Soc. Am.* **57,** 1341–1351.

Butler, J. L., Butler, S. C., Clark, A. E. (1990), "Unidirectional Magnetostrictive/Piezoelectric Hybrid Transducer," *J. Acoust. Soc. Am.* **88,** 7–11.

Clark, A. E. (1980), "Magnetostrictive Rare Earth-Fe_2 Compounds," in: E. P. Wohlfarth (Ed.), *Ferromagnetic Materials,* Vol. 1, Amsterdam: North-Holland Chap. 7.

Henriquez, T. A., Young, A. M. (1980), "The Helmholtz Resonator as a High-Power Deep-Submergence Source for Frequencies below 500 Hz," *J. Acoust. Soc. Am.* **65,** 1556–1558.

Hueter, T. F., Bolt, R. H. (1955), *Sonics,* New York: Wiley.

Hunt, Frederick V. (1982), *Electroacoustics,* Cambridge, MA: Harvard Univ. Press. Reprinted by the Acoustical Society of America, New York.

Kinsler, L. E., Frey, A. R., Coppens, A. B., Sanders, J. V. (1982), *Fundamentals of Acoustics,* 3rd ed., New York: Wiley.

Mason, Warren P. (1934), "Electrical Wave Filters Employing Quartz Crystals as Elements," *Bell Syst. Tech. J.* **13,** 405–452.

Mason, Warren P. (1948), *Electromechanical Transducers and Wave Filters,* 2nd ed, New York: Van Nostrand.

McCammon, D. F., Thomson, W. Jr., (1980), "The Design of Tonpilz Piezoelectric Transducers Using Nonlinear Goal Programming," *J. Acoust. Soc. Am.* **68,** 754–757.

McCollum, D. F., Hamonic, B. F., Wilson, O. B. (1992), *Transducers for Sonics and Ultrasonics, Proceedings of the Third International Workshop,* Lancaster, PA: Technomic Publishing Co., Inc.

Shung, K. K., Smith, M. B., Tsui, B. M. W. (1992), *Principles of Medical Imaging,* New York: Academic Press, Inc.

Stansfield, D. (1990), *Underwater Electroacoustic Transducers,* Bath Univ. Press and Institute of Acoustics.

Urick, R. J. (1975), *Principles of Underwater Sound,* 2nd ed., New York: McGraw-Hill.

Van Dyke, K. S. (1928), "The Piezoelectric Resonator and Its Equivalent Network," *Proc. I.R.E.* **16,** 742–764.

Wilson, O. B. (1985), *An Introduction to the Theory and Design of Sonar Transducers,* Washington DC: U.S. Government Printing Office.

Woollett, R. S. (1960), Theory of the Piezoelectric Flexural Disk Transducer with Applications to Underwater Sound, USL Research Report No. 490, U.S. Navy Underwater Sound Laboratory, New London CT.

Further Reading

Beranek, Leo L. (1954), *Acoustics,* New York: McGraw-Hill.

Bouyoucos, J. V. (1975), "Hydroacoustic Transduction," *J. Acoust. Soc. Am.* **57,** 1341–1351.

Brigham, G., Glass, B. (1980), "Present Status in Flextensional Transducer Technology," *J. Acoust. Soc. Am.* **64,** 1046–1052.

Butler, J. L., Cipolla, J. R., Brown, W. D. (1981), "Radiating Head Flexure and Its Effect on Transducer Performance," *J. Acoust. Soc. Am.* **70,** 500–503.

Cady, W. G. (1964), *Piezoelectricity,* Rev. Ed., Vol. 1, New York: Dover.

Firestone, F. A. (1933), "A New Analogy between Mechanical and Electrical Systems," *J. Acoust. Soc. Am.* **4,** 249–267.

Groves, Ivor V. (Ed.) (1981), *Acoustic Transducers,* Benchmark Papers in Acoustics No. 14, Stroudsburg, PA: Hutchinson Ross.

Kinsler, L. E., Frey, A. R., Coppens, A. B., Sanders, J. V. *Fundamentals of Acoustics,* 3rd ed., New York: Wiley.

Martin, G. E. (1964), "On the Theory of Segmented Electromechanical Systems," *J. Acoust. Soc. Am.* **36,** 1366–1370.

Mason, Warren P. (1981), "Piezoelectricity, Its History and Applications," *J. Acoust. Soc. Am.* **70,** 1561–1566.

Olson, Harry F. (1943), *Dynamic Analogies,* New York: Van Nostrand.

Sims, C. S. (1963), "Bubble Transducer for Radiating High-Power Low-Frequency Sound in Water," *J. Acoust. Soc. Am.* **32,** 1305–1308.

Woollett, R. S. (1977), *Underwater Helmholtz Resonator Transducers: General Design Principles,* NUSC Technical Rept. 5633, Naval Underseas Systems Center, New London, CT.

TRANSFORMERS, ELECTRONIC

Enrico Santi, *TESLAco, Irvine, California, U.S.A.*

	Introduction	85
1.	**Theory of Operation**	86
1.1	The Ideal Transformer	86
1.2	Transformers with Internal Energy Storage	87
1.2.1	The Case of a One-Winding Inductor	87
1.2.2	The Case of a Two-Winding Transformer	88
1.3	Transformers with Nonperfect Magnetic Coupling	89
1.4	Transformers with Power Losses	89
2.	**Design Considerations**	90
2.1	Ferromagnetism	90
2.1.1	From Material Characteristics to Inductance Characteristic	91
2.1.1.1	Physical Scaling	91
2.1.1.2	Number-of-Turns Scaling	91
2.1.2	Core Loss	91
2.1.2.1	Hysteresis Loss	91
2.1.2.2	Eddy-Current Loss	92
2.1.3	Ferromagnetic Distortion	92
2.1.3.1	Low Source Impedance	92
2.1.3.2	High Source Impedance	93
2.1.4	Introduction of an Air Gap	93
2.2	Magnetic Materials	94
2.2.1	Metallic Alloys	95
2.2.1.1	Iron–Silicon Alloys	95
2.2.1.2	Iron–Nickel Alloys	95
2.2.1.3	Iron–Cobalt Alloys	95
2.2.2	Ferrites	95
2.2.3	Powdered Cores	95
2.2.4	Amorphous Alloys	95
2.3	Leakage Inductances	96
2.4	Conduction Loss: Skin and Proximity Effects	96
2.5	Parasitic Capacitances	96
2.6	Other Considerations	97
3.	**Wide-Band Transformers**	97
3.1	Transformer Equivalent Models	98
3.2	Midband Frequency Response	99
3.3	Low-Frequency Response	99
3.3.1	Low-Frequency Distortion	100
3.4	High-Frequency Response	100
3.4.1	Model with Leakage Inductance	100
3.4.2	Model with Leakage Inductance and Shunt Capacitance	101
3.5	Special Types of Wide-Band Transformers	102
4.	**Pulse Transformers**	102
4.1	The Rise Period	102
4.2	The Top Period	103
4.3	The Off Period	103
4.4	Design Issues	104
5.	**Electronic Power Transformers**	105
5.1	Transformers in Switching Converters	106
5.1.1	Forward Converter	106
5.1.2	Flyback Converter	107
5.1.3	Isolated Ćuk Converter	107
5.2	Requirements for Electronic Power Transformers	107
5.3	Coupled Inductors and Zero Current Ripple	109
6.	**Magnetic Amplifiers**	109
6.1	Simple Magnetic Amplifier	109
6.2	Other Magnetic Amplifier Circuits	111
	Glossary	111
	Works Cited	111
	Further Reading	112

INTRODUCTION

Transformers are passive devices used to transform voltage and current. They are used in countless applications and come in a variety of forms and sizes. However, the principles of operation are the same for all of them.

The voltage and current scaling property of transformers is used for impedance

digital Encyclopedia of Applied Physics

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

matching between a source and a load. Moreover, they provide electrical insulation between input and output, which allows ground shifting and polarity inversion.

Two classes of transformers can be identified: power transformers and electronic transformers. The term power transformer is normally used for transformers operating over a narrow frequency range (50–60 Hz) and used in distribution systems for electric utilities. The term electronic transformers is basically used for all other types of transformers. Examples of electronic transformers are audio and wide-band transformers, pulse transformers, magnetic amplifiers, and transformers used in high-frequency switching power supplies.

Many electronic components come as off-the-shelf standardized parts, but only a limited selection of off-the-shelf transformers is available. The problem is that transformers need to be specifically tailored and optimized for specific applications. As a result, transformers tend to be expensive.

Another characteristic of transformers is that they tend to be bulky, especially in comparison with other miniaturized electronic components. For this reason in many applications designers try to do without them if possible.

In Sec. 1 the theory of operation of transformers is reviewed starting from the ideal transformer and then introducing various nonidealities. In Sec. 2 various issues related to transformers are discussed: ferromagnetism and ferromagnetic materials, core and copper losses, leakage inductances, and parasitic capacitances. In Sec. 3 the wide-band transformer is analyzed in the frequency domain, whereas in Sec. 4 the pulse transformer is analyzed in the time domain. Sec. 5 discusses transformers in high-frequency switching power supplies. Finally, Sec. 6 deals with magnetic amplifiers.

1. THEORY OF OPERATION

The case of an ideal transformer is discussed first, and the impedance scaling law of transformers is derived. An ideal transformer satisfies three assumptions: no internal energy storage, perfect magnetic coupling between the windings, and no power loss. These assumptions are relaxed one by one in subsequent sections, and the relative equivalent models are presented.

1.1 The Ideal Transformer

In order to understand the basics of transformers, the case of an ideal transformer is considered here. The physical system is shown in Fig. 1(a), and the circuit representation of an ideal transformer is shown in Fig. 1(b). The following assumptions are made:

1. There is perfect magnetic coupling between primary and secondary windings.
2. There is no instantaneous energy storage in the transformer.
3. The transformer is lossless.

Real-life transformers only approximately satisfy these assumptions, but the transformer designer usually strives to come as close as possible to satisfying them.

Assumption 1 states that the same flux ϕ couples the two windings. Application of Faraday's law gives

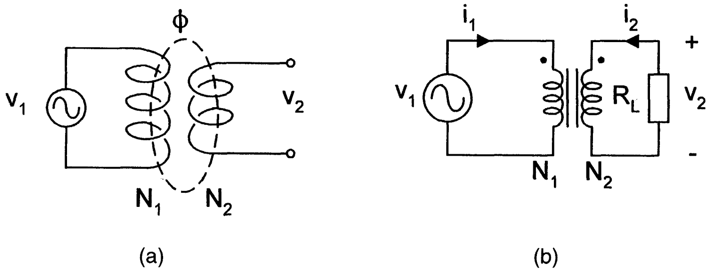

FIG. 1. Ideal transformer: (a) physical system and (b) circuit representation.

$$\frac{d\phi}{dt} = \frac{v_1}{N_1} = \frac{v_2}{N_2}. \tag{1}$$

From Eq. (1),

$$v_1/v_2 = N_1/N_2. \tag{2}$$

This is the *voltage scaling law* of transformers, which says that the voltage in the various windings is proportional to the number of turns. In the circuit representation of Fig. 1(b), the voltage polarity in the various windings is given by the *dot convention:* a positive voltage applied to the marked terminal of one winding induces a positive voltage on the corresponding marked terminal of the other winding.

From assumptions 2 and 3 it follows that there is no net power flowing into the transformer at any time, and therefore the power balance requires that

$$v_1 i_1 + v_2 i_2 = 0. \tag{3}$$

Combining Eqs. (2) and (3) gives

$$i_1/i_2 = -N_2/N_1. \tag{4}$$

This is the *current scaling law* of transformers.

Dividing (2) by (4) we get the *impedance scaling law* of transformers

$$\frac{v_1/i_1}{-v_2/i_2} = \frac{R_L'}{R_L} = \left(\frac{N_1}{N_2}\right)^2. \tag{5}$$

The secondary load impedance R_L appears at the primary as an impedance

$$R_L' = (N_1/N_2)^2 R_L. \tag{6}$$

This is a very useful property of transformers. In electronics, it frequently happens that a source that would operate most efficiently or with least distortion having as a load an impedance R_L' must be connected to a load of impedance R_L. In these cases a transformer can be inserted between source and load to provide the desired impedance scaling.

1.2 Transformers with Internal Energy Storage

With respect to the ideal transformer of Sec. 1.1, the assumption of no energy storage in the transformer is now relaxed. As a result, a magnetizing inductance, representing the energy stored in the transformer, is defined.

1.2.1 The Case of a One-Winding Inductor In order to describe the basic operation of a transformer, we start with only one winding with N_1 turns wound around a high-permeability magnetic core, as shown in Fig. 2. Current i_m flowing in the winding generates a magnetic flux ϕ that closes in the magnetic core. The relationship between voltage and flux is given by Faraday's law

$$v_1 = N_1\frac{d\phi}{dt}. \tag{7}$$

Ampère's law

$$I = \oint \mathbf{H}\cdot d\mathbf{l} \tag{8}$$

states that the integral of the magnetic field **H** along a closed path is equal to the enclosed current *I*.

Ampère's law is now applied to the mean magnetic path inside the core shown in Fig. 2. The path encloses the N_1 turns, and therefore the enclosed current is $N_1 i_m$. This quantity is called the *ampere-turns* linked by the magnetic flux path. It is assumed that the magnetic field **H** is of constant amplitude, and it is parallel to the integration path at all points as in Fig. 2. Under these assumptions, Eq. (8) becomes

$$Hl = N_1 i_m, \tag{9}$$

where l is the length of the mean magnetic path. This equation gives the relationship be-

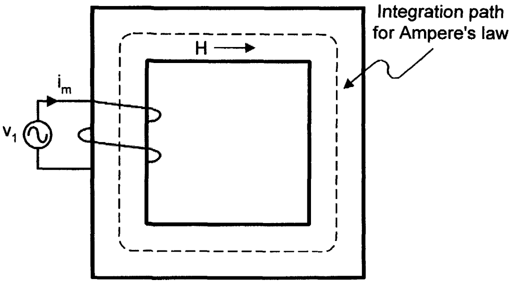

FIG. 2. Magnetic core with one winding.

tween magnetic field H and exciting current i_m.

The relationship between magnetic field H and magnetic induction B is in general nonlinear for the high-permeability materials normally used for transformer cores, and it is discussed in Sec. 2.1 on ferromagnetic materials. For now a linear relationship

$$B = \mu H \tag{10}$$

is used, where μ is the material *permeability*. Assuming that the magnetic induction is constant throughout the core cross section S, flux ϕ in the core is

$$\phi = BS. \tag{11}$$

The left-hand side of Eq. (9) is multiplied and divided by the constant μS. Taking into account Eqs. (10) and (11), we get

$$\phi \mathcal{R}_c = N_1 i_m \tag{12}$$

where $\mathcal{R}_c$ is the *reluctance* of the magnetic flux path in the core and is defined as

$$\mathcal{R}_c = l/\mu S. \tag{13}$$

Both sides of Eq. (12) are now multiplied by N_1 and, after rearranging, we have

$$N_1 \phi = L_m i_m, \tag{14}$$

with the *magnetizing inductance* L_m defined as

$$L_m = N_1^2/\mathcal{R}_c. \tag{15}$$

Equation (14) states that *flux linkage*, i.e., flux multiplied by the number of turns linking the flux, is proportional to the current that generates it, and magnetizing inductance L_m is flux linkage per unit current. The current i_m that generates the magnetic flux ϕ is called the *magnetizing current*.

1.2.2 The Case of a Two-Winding Transformer A second open-circuited winding with N_2 turns is now added, and the transformer shown in Fig. 3(a) is obtained. The current i_1 in the first winding is unchanged. If the two windings are perfectly coupled, flux ϕ generated by the primary winding links the secondary winding also, and Eqs. (1) and (2), derived in the case of an ideal transformer, hold.

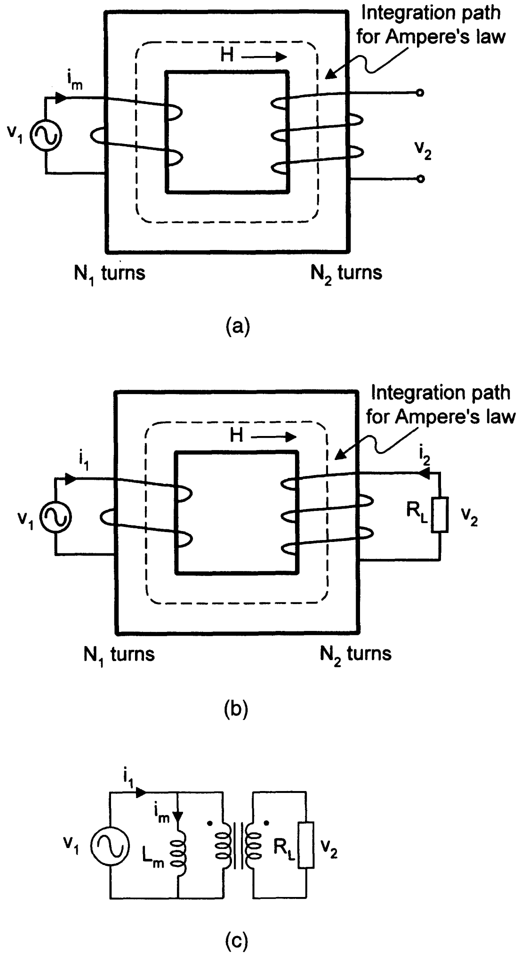

FIG. 3. Magnetic core with two windings: (a) no load, (b) load R_L, and (c) equivalent circuit.

Suppose now that we connect a load R_L on the secondary of the transformer as in Fig. 3(b). Neglecting second-order effects (like, for example, resistive voltage drops in the windings), the primary and secondary voltages v_1 and v_2 are unchanged. Equations (1) and (2) still hold. Current i_2 is dictated by the load impedance:

$$i_2 = -v_2/R_L. \tag{16}$$

Ampère's law can be applied to the path shown in Fig. 3(b). Notice that this path encloses both N_1 primary turns and N_2 secondary turns:

$$Hl = N_1 i_1 + N_2 i_2. \tag{17}$$

Following the same steps as in Sec. 1.2.1, we multiply and divide the left-hand side by μS, and we get

$$\phi \mathfrak{R}_c = N_1 i_1 + N_2 i_2. \tag{18}$$

This equation is compared with Eq. (12) derived in the case of a one-winding inductor. Flux ϕ in the core, which is determined by the winding voltage, is unchanged, and therefore the left-hand sides of the two equations are identical. Equating the right-hand sides gives

$$i_1 = i_m - (N_2/N_1) i_2. \tag{19}$$

This equation can be interpreted in the following way: the primary current i_1 has two components; one is the magnetizing current i_m, which generates the flux in the core, and the other component is the reflected load current. This current cancels the ampere-turns generated by current i_2 and does not generate any flux in the core, as can be seen substituting (19) into (18).

Equation (19) should be compared with Eq. (4) for the ideal transformer. The difference is the term i_m, related to the energy stored in the transformer.

The circuit model is shown in Fig. 3(c). Magnetizing inductance L_m defined in (14) appears in parallel with the ideal transformer.

1.3 Transformers with Nonperfect Magnetic Coupling

At this point the assumption of perfect magnetic coupling between primary and secondary is relaxed.

As shown in Fig. 4(a), the flux coupled with each winding can be divided into two components: coupled flux ϕ and leakage fluxes ϕ_{l1} and ϕ_{l2},

$$\phi_1 = \phi_{l1} + \phi, \tag{20}$$

$$\phi_2 = \phi_{l2} + \phi. \tag{21}$$

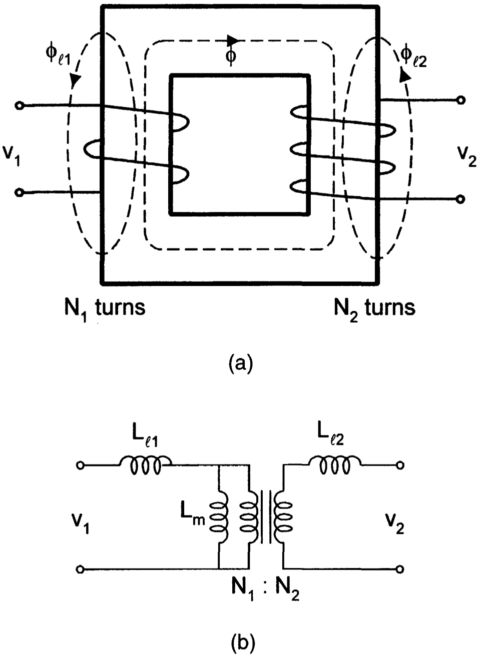

FIG. 4. Transformer with leakage fluxes: (a) physical system and (b) equivalent circuit.

The reluctances of the leakage flux paths of ϕ_{l1} and ϕ_{l2} are called $\mathfrak{R}_{l1}$ and $\mathfrak{R}_{l2}$, respectively. The corresponding leakage inductances, defined in the same way as (14), are

$$L_{l1} = N_1^2/\mathfrak{R}_{l1}, \tag{22}$$

$$L_{l2} = N_2^2/\mathfrak{R}_{l2}. \tag{23}$$

The equivalent model is shown in Fig. 4(b).

Notice that the value of magnetizing inductance L_m is a function of nonlinear core-material characteristics, and therefore it is not constant. On the other hand, leakage inductances represent flux in air, which is a linear medium, and therefore they are approximately constant. For a more complete discussion of leakage, refer to M.I.T. (1943).

1.4 Transformers with Power Losses

Next the assumption of no loss in the transformer is relaxed. In a real-life transformer there are various loss mechanisms:

- Conduction losses. These are Ohmic losses in the windings. Current distribution in the conductors is frequency dependent (skin and proximity effects) and so are

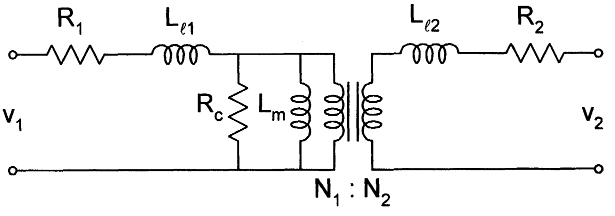

FIG. 5. Low-frequency equivalent model of a transformer.

conduction losses. These losses are modeled by series resistances R_1 and R_2.

- Core losses. ac fluxes in ferromagnetic materials cause losses, which are a nonlinear function of flux swing, flux bias, excitation frequency, and previous magnetization history. These losses are usually modeled by a nonlinear resistor R_c in parallel with magnetizing inductance L_m.

The equivalent model is shown in Fig. 5. This model is valid at low frequencies. At high frequencies, transformer parasitic capacitances must be included also, as explained in Sec. 2.5.

2. DESIGN CONSIDERATIONS

2.1 Ferromagnetism

In the derivation of magnetizing inductance L_m of a transformer in Sec. 1.2, a linear relationship between magnetic field H and magnetic induction B was assumed [Eq. (10)]. In this section the nonlinear characteristics of ferromagnetic materials are discussed (see Smit, 1971).

A typical magnetization curve of a ferromagnetic material (see Charap, 1974) is shown in Fig. 6. If H is varied from a large negative value to a large positive value and back again, the *B-H* curve describes a loop, called *hysteresis curve,* which is transversed in the direction indicated by the arrows in Fig. 6. The following observations can be made.

Material permeability is higher than free-space permeability. *Relative permeability* is defined as the ratio of material permeability to free-space permeability. Relative permeabilities of magnetic materials vary from 10–100 for powdered cores to 100 000 for nickel alloys. The high relative permeability of ferromagnetic materials is the main reason why they are used for transformers. Their effect is to reduce reluctance $\mathcal{R}_c$ of the flux path in the core and boost magnetizing inductance L_m, as can be seen from Eqs. (13) and (15).

The magnetization tends to lag the magnetic field. When H increases from a negative to a positive value, B takes a given set of values, and, when H decreases, B takes a different set of values, higher than before. In particular, when H decreases to zero, B still has a positive value B_r called *remanent magnetization.* In order to bring the magnetization to zero, a negative magnetic field H_c called *coercive force* must be applied. B_r and H_c are shown in Fig. 6.

The shape of the hysteresis loop changes with the frequency of excitation. In particular, the hysteresis loop becomes wider at higher frequencies, as shown in Fig. 7.

The *B-H* characteristic exhibits magnetic saturation. The slope of the *B-H* curve at a given point is called *differential permeability.* For low values of magnetic field, differential permeability is high, but for higher values it decreases, and it eventually becomes equal to the permeability of free space. A family of hysteresis curves can be obtained for different values of ac excitation at a given frequency. Such a family of curves is shown in

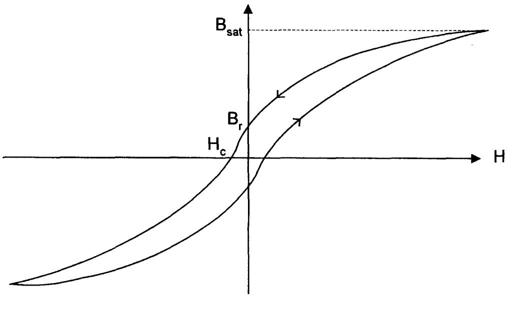

FIG. 6. *B-H* hysteresis characteristic.

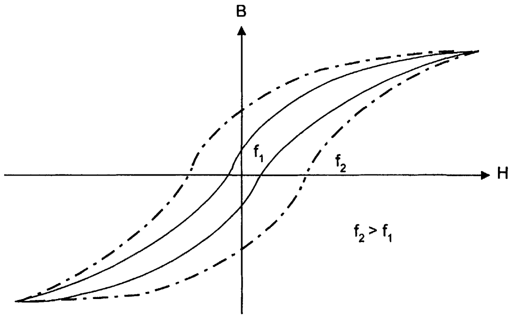

FIG. 7. *B-H* loop widens with frequency.

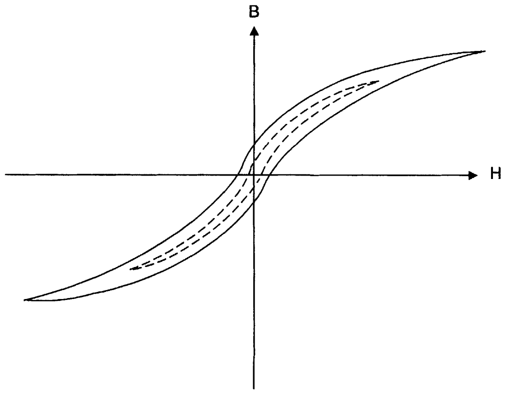

FIG. 8. Family of *B-H* loops for various amplitudes of excitation.

Fig. 8. Magnetic saturation gives an upper limit to the usable flux swing in the core.

2.1.1 From Material Characteristics to Inductance Characteristic Suppose we have a transformer built using a ferromagnetic material with a characteristic of the type shown in Fig. 9(a). It is instructive to relate the material characteristic (*B-H* loop) to the circuit characteristics ($N_1\phi$-i_m curve) of the transformer (see Watson, 1980).

The process involves two steps: physical scaling and number-of-turns scaling.

2.1.1.1 Physical Scaling. We start from the *B-H* curve. Slope in this curve represents permeability. The x axis is scaled by the magnetic path length l, so that it represents ampere-turns (9). The y axis is scaled by core cross section S, so that it represents flux ϕ in the core. The resulting characteristic is shown in Fig. 9(b). The slope represents the inverse of reluctance, $\mathcal{R}_c^{-1}$ (also called *permeance*).

2.1.1.2 Number-of-Turns Scaling. The y axis is now scaled by number of turns N_1, so that it represents flux linkage $\lambda = N_1\phi$. The x axis is scaled by $1/N_1$, so that it represents current i_m. The slope of the resulting characteristic of Fig. 9(c) is magnetizing inductance L_m. Notice how it changes as we move along the hysteresis loop.

2.1.2 Core Loss Core loss is the power dissipated in the core. Two different loss mechanisms can be identified: hysteresis loss (see Chen, 1978; Tan *et al.*, 1995) and eddy-current loss.

2.1.2.1 Hysteresis Loss. Let us consider a winding of N_1 turns wound on a ferromagnetic core (e.g., the primary of a trans-

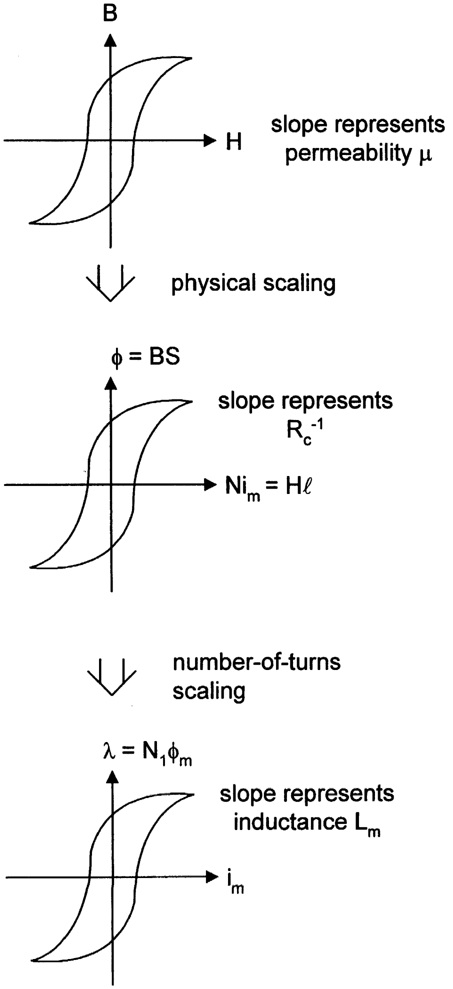
B
H
slope represents permeability μ
physical scaling
φ = BS
slope represents R_c^{-1}
$Ni_m = H\ell$
number-of-turns scaling
$\lambda = N_1\phi_m$
slope represents inductance L_m
i_m

FIG. 9. From material characteristic to inductance characteristic.

former with no secondary load) as in Sec. 1.2.1. The energy delivered to the core in an interval Δt is

$$J = \int v_1 i_m dt. \tag{24}$$

Substituting Eqs. (7), (9), and (11) yields

$$J = \int N_1 \frac{d\phi}{dt}\frac{Hl}{N_1} dt, \tag{25}$$

$$= Sl \int H\, dB. \tag{26}$$

The quantity Sl represents the core volume, and therefore the integral represents the energy delivered to the core per unit volume.

Now we want to find the energy lost during one period of ac excitation, i.e., during one transverse of the hysteresis loop of Fig. 10.

The differential energy per unit volume delivered to the core while the magnetic field has an incremental increase dB (path 1–2) is represented by the sum of the shaded and hatched areas of Fig. 10. The hatched portion of this energy is returned to the external circuit when, later during the ac period, magnetic field decreases by dB (path 3–4), and it represents reactive energy exchanged back and forth between core and external circuit. The shaded area in Fig. 10 represents energy dissipated in the core. Notice that in the case of a single-valued B-H characteristic with no hysteresis there is no such loss, and all energy delivered to the core is then returned.

From the previous discussion it follows that the energy lost in one period of ac excitation is equal to the area of the hysteresis loop. The power loss is

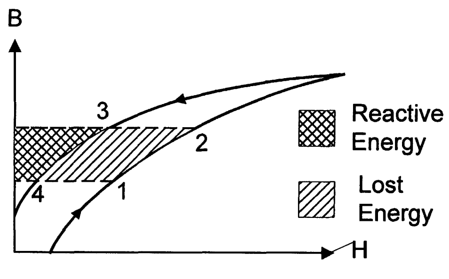

FIG. 10. Hysteresis loss calculation from *B-H* loop.

$$P_{\text{loss}} = Vf \oint H\, dB, \tag{27}$$

where V is the core volume, f is the excitation frequency, and the integral is calculated along the hysteresis loop. As shown in Fig. 7, the hysteresis loop widens with frequency, and therefore hysteresis loss increases more than proportionally to frequency.

2.1.2.2 Eddy-Current Loss. Eddy-current losses are resistive losses due to circulating currents in a conductive magnetic core. The varying flux in the core induces currents in it in the same way as currents are induced in the secondary of a transformer. These currents tend to be large because metals and other alloys used for magnetic cores have rather low resistivity. A solution to mitigate this problem is to reduce the path in the core available for current flow by using core laminations insulated from each other and stacked together. To be effective, the lamination plane should be normal to the induced current direction.

Losses in metals become unacceptable for radio and microwave frequencies. Ferrites, which have much higher resistivity, are used at these frequencies. Ferromagnetic materials are discussed in Sec. 2.2.

2.1.3 Ferromagnetic Distortion

Ferromagnetic distortion is the distortion in the current and voltage wave forms of a transformer due to the nonlinear core characteristic. This distortion is usually undesirable. Given the nonlinear B-H loop characteristic, if the applied voltage is sinusoidal, the magnetizing current will be distorted, and, conversely, if the applied current is sinusoidal, the flux in the core and the voltage induced in the secondary will be distorted. The actual voltage and current wave forms depend on the external circuit as well as on the core characteristic.

2.1.3.1 Low Source Impedance. In the case of low source impedance, the applied voltage is approximately sinusoidal. Figure 11 shows the salient wave forms. The flux is sinusoidal, 90° delayed with respect to the voltage (7). If the hysteresis curve is linear, the current is sinusoidal in phase with the flux. Otherwise, the relationship between current and flux is given by the ϕ-i characteristic of Fig. 9(b) obtained from the hysteresis characteristic. The current is distorted

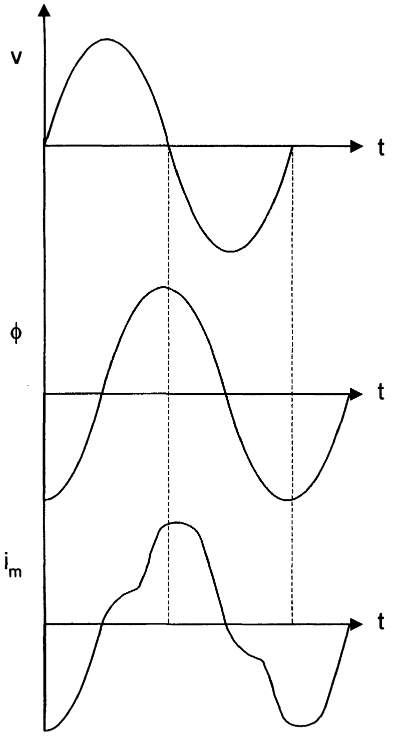

FIG. 11. Ferromagnetic distortion.

as shown in Fig. 11. However, if the core flux is sinusoidal, the voltage induced on the secondary is also sinusoidal. It may seem that the distorted magnetizing current is of no consequence, but in reality this current will cause harmonic voltages on the transformer winding resistances and any other parasitic resistances, so that the voltage on the transformer primary will have harmonic components and the induced secondary voltage will be distorted. It should be stressed that the current shape depends on the hysteresis loop shape. For example, a square hysteresis loop gives approximately a square-wave current in phase with the voltage.

2.1.3.2 High Source Impedance. If the source impedance is large, the external circuit acts as a current source, and the source current is sinusoidal. In this case the flux will be distorted and so will be the induced voltage.

2.1.4 Introduction of an Air Gap In transformers discussed up to this point, the high-permeability core forms a closed high-permeability path for the magnetizing flux. As a result, high values of magnetizing inductance can be achieved. This is usually regarded as a desirable property, since the transformer better approximates the behavior of the ideal transformer of Sec. 1.1, which has infinite magnetizing inductance.

However, there are two problems:

- Distortion. The magnetizing inductance is nonlinear on account of the nonlinear *B-H* loop, and this causes distortion, especially at low frequencies, as explained in Sec. 2.1.3.
- Saturation. Even a small amount of dc current in one of the windings may cause core saturation.

To solve these problems, an air gap can be introduced in the magnetic flux path. The result is a reduction of the value of magnetizing inductance, an increase of dc current capability of the transformer, and a stabilization of the value of magnetizing inductance, which is now dominated by gap reluctance. Figures 12(a) and 12(b) show the idealized transformer characteristic for the ungapped and gapped cores, respectively. The core characteristic is sheared as an effect of gap.

A transformer with an air gap can be analyzed following the steps of Sec. 1.2.1. When an air gap is introduced, the flux in the core flows through the gap. Maxwell's equation states that

$$\oint_S \mathbf{B}\cdot d\mathbf{S} = 0, \tag{28}$$

which means that the magnetic field does not have sources or sinks. As a result, induction B in the core and in the gap is the same, as shown in Fig. 13. It follows that

$$B_c = B_g, \tag{29}$$

$$\mu H_c = \mu_0 H_g, \tag{30}$$

where μ and μ_0 are the permeabilities of the core and of free space, respectively. H_c and H_g, the magnetic field in the core and in the gap, respectively, are assumed constant throughout the cross section S.

Referring to Fig. 13(a), we can apply Ampère's law (8):

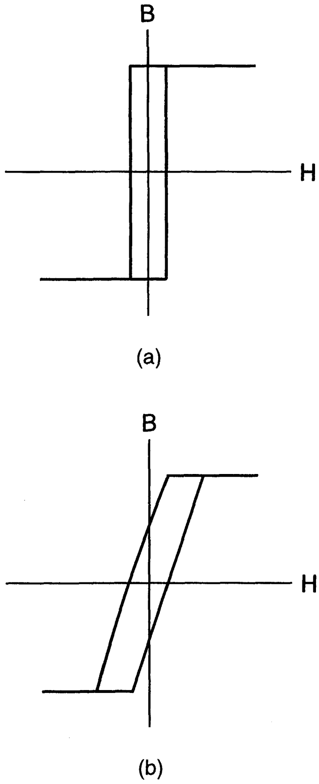

(a)

(b)

FIG. 12. Idealized transformer characteristic (a) for ungapped core and (b) for gapped core.

$$H_c l + H_g l_g = N_1 i_m, \tag{31}$$

where l is the length of the magnetic flux path in the core and l_g the length of the magnetic flux path in the gap, i.e., the gap size. The first term of (31) is multiplied and divided by μS and the second term by $\mu_0 S$.

The result is

$$\phi(\mathcal{R}_c + \mathcal{R}_g) = N_1 i_m, \tag{32}$$

where $\mathcal{R}_c$ is the core reluctance defined in (13) and $\mathcal{R}_g$ is the gap reluctance defined as

$$\mathcal{R}_g = l_g/\mu_0 S. \tag{33}$$

Usually $\mathcal{R}_g \gg \mathcal{R}_c$, and therefore gap reluctance dominates. Multiplying Eq. (32) by N_1 and rearranging, we get

$$\lambda = N_1 \phi = L_m i_m \tag{34}$$

with

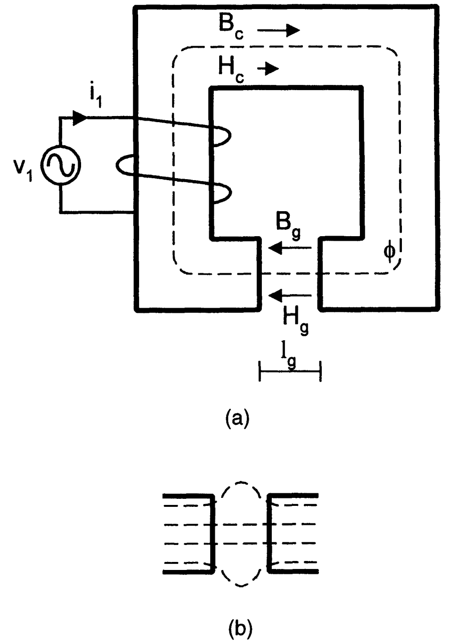

(a)

(b)

FIG. 13. (a) Introduction of an air gap in the core and (b) detail of fringing flux at gap.

$$L_m = N_1^2/(\mathcal{R}_g + \mathcal{R}_c). \tag{35}$$

Magnetizing inductance L_m in the presence of the gap is reduced from the ungapped case of (14).

The calculation of gap reluctance $\mathcal{R}_g$ described above does not take into account the effect of *fringing flux.* The flux lines crossing the gap tend to spread out as shown in Fig. 13(b). As a result the effective cross section of the gap is bigger than the core cross section, and the measured inductance is bigger than the value one would expect based on (33) and (35). This effect is particularly significant for big gaps.

2.2 Magnetic Materials

Different magnetic materials with different properties are available. The following material characteristics are of interest:

- Shape of the *B-H* loop. Magnetic materials are characterized as hard or soft materials. Hard materials, which have high remanent magnetization B_r and coercive force H_c, are good for permanent magnets because they have high flux density with no excitation (remanent flux density) and they are hard to demagnetize (high coercive force). Soft materials, which have low coercive force, are good for transformers and inductors, because they have less hysteresis losses.

- Resistivity. This is important for eddy-current losses in the core. See Sec. 2.1.2.
- Saturation flux density. Materials with higher saturation flux density are desirable because they allow a reduction of the magnetic core physical size. However, full utilization of the available flux swing is often prevented by excessive core losses, especially at higher frequencies.
- Curie temperature. This is the temperature at which the material loses its ferromagnetic properties. It gives an upper bound to the operating temperature of a transformer.
- Cost.

2.2.1 Metallic Alloys The main ferromagnetic materials are iron, nickel, cobalt, and their alloys. These materials are used mostly at power frequencies. They are fairly good conductors, and therefore their use at higher frequencies is limited by eddy-current losses.

2.2.1.1 Iron–Silicon Alloys. Silicon steel is widely used for low- and intermediate-frequency applications. Addition of silicon (from 0.5% to 4.5%) to steel reduces both hysteresis and eddy-current losses.

Silicon steel is sold in laminations of various thicknesses, usually insulated, to reduce eddy currents. Silicon steel can be obtained in *grain-oriented* form: the material is cold rolled to obtain the easy magnetization direction parallel to the rolling direction. As a result it exhibits very low core losses.

2.2.1.2 Iron–Nickel Alloys. These alloys have a 45% to 80% nickel content. They are used for low-field applications where high permeability is required. These types of applications are common at higher frequencies, where core losses become a limiting factor. Extremely square-loop materials can also be obtained with proper processing.

2.2.1.3 Iron–Cobalt Alloys. Iron–cobalt alloys are used when a high magnetic saturation is required. Supermendur is an iron–cobalt alloy with 2% vanadium, 49% iron, and 49% cobalt. Grain orientation is obtained by annealing the material under a magnetic field. Supermendur has a high saturation flux density (2.2 T), and it has the highest permeability at elevated flux levels (100 000 relative permeability at 2 T) and the highest remanent magnetization of any alloy. These materials are extremely expensive.

2.2.2 Ferrites Ferrites (see Snelling, 1969, 1983) are nonmetallic ferromagnetic oxides that are used from the high audio range well into the megahertz range. The main advantage of these materials is the high electrical resistivity (10^6 to 10^{10} Ω cm), normally 10^6 times that of metallic magnetic alloys, which makes them attractive at higher frequencies. Different ferrite materials with different compositions are optimized for a certain range of operating frequencies. Ferrites are limited by their low Curie temperature and by their low saturation flux density (0.3–0.5 T), which decreases with temperature. Material permeability, which ranges from few hundreds to 10 000, is also lower than the permeability of metallic magnetic alloys.

2.2.3 Powdered Cores These materials are used for higher-frequency applications into the 200-kHz range. Powdered ferromagnetic materials are compacted with an insulating binder, thereby significantly reducing eddy-current losses. This process effectively builds a "distributed gap" in the material, which lowers the relative permeability to values between 3 and 100, depending on the mix used. The advantage is that the permeability is fairly stable in value at high dc current levels, and the material exhibits a soft (not abrupt) saturation characteristic.

2.2.4 Amorphous Alloys Also called magnetic metallic glasses (see Raskin and Davis, 1981; Boll and Warlimont, 1981; Graham and Egami, 1980), these amorphous alloys, recently (1970–1980) introduced, have significantly lower losses than comparable crystalline alloys, which makes them attractive at high frequencies. They represent an alternative to ferrites with the advantage of higher saturation flux densities (1.6 T). They are alloys of iron, boron, and silicon, which are quenched so rapidly that crystal formation typical of other magnetic alloys is prevented. Disadvantages are that the material is very brittle and very strain sensitive, and that only a limited set of shapes are available on account of manufacturing difficulties.

2.3 Leakage Inductances

As explained in Sec. 1.3, leakage flux in a transformer is flux generated by one winding that does not couple (or only partially couples) with the other winding. This is usually undesirable. For example, it reduces the high-frequency bandwidth of wide-band transformers and pulse transformers.

In order to reduce leakage, windings are physically placed as close as possible. Factors that limit how close the windings are placed are insulation thickness, the need to limit capacitive coupling between layers, and cost of manufacturing.

A very effective approach to reduce leakage and improve coupling is to split the windings into sections and then to sandwich them. This process is called *sectionalization* or *interleaving*. The case of a secondary split into two with interleaved primary is shown in Fig. 14.

Another way to reduce leakage, frequently used especially with toroidal cores operated at high frequencies, is to use *bifilar windings*, i.e., wind primary and secondary together around the core.

Leakage inductances can be calculated by calculating the energy stored in the leakage magnetic field between and through the windings. As frequency increases, leakage inductance decreases because the leakage magnetic field is confined to the surface of the conductors by the skin effect (see Dowell, 1966).

2.4 Conduction Loss: Skin and Proximity Effects

Conduction losses in high-frequency transformer windings are higher than one would expect on the basis of dc resistance measurements. This is due to nonuniform current distributions inside the windings caused by the varying magnetic field produced within the conductor by its own current and by the currents in neighboring conductors. These phenomena are often called *skin effect* and *proximity effect*.

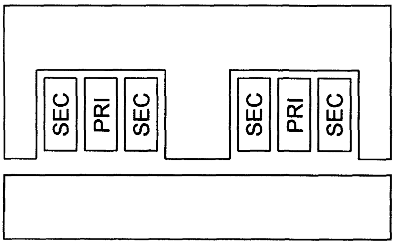

FIG. 14. Transformer with interleaved windings.

Skin effect can be studied analytically for some simple geometries—for example, in the case of a semi-infinite conductor or a round conductor (see Ramo *et al.*, 1984). The result is that the electric and magnetic fields decay exponentially as they penetrate the conductor. The depth of penetration or *skin depth* δ is found to be

$$\delta = 1/\sqrt{\pi f \mu \sigma}, \tag{36}$$

where f is the excitation frequency, μ is the permeability, and σ is the conductivity.

A way to visualize skin effect is to say that the magnetic field induces a current in the conductor by Faraday's law, which tends to oppose the variation of flux. These eddy currents limit the penetration of the magnetic field to a region near the surface of the conductor.

Proximity effect is analyzed in Dowell (1966) and Vandelac and Ziogas (1987). Another approach to the study of proximity effects is to use finite-element simulations (see Goldberg *et al.*, 1989a,b).

2.5 Parasitic Capacitances

When a transformer is powered, different electric field gradients arise almost everywhere. A number of different capacitances can be identified as in Fig. 15:

- between turns,
- between layers,
- between windings, and
- between the core and the windings.

Since the voltages change from point to point, distributed capacitances should be considered. For analysis purposes, these capacitances are usually lumped together into equivalent capacitances.

Usually the capacitances between adjacent layers of a winding and the capacitances between different windings (or, more precisely, the capacitances between adjacent layers of

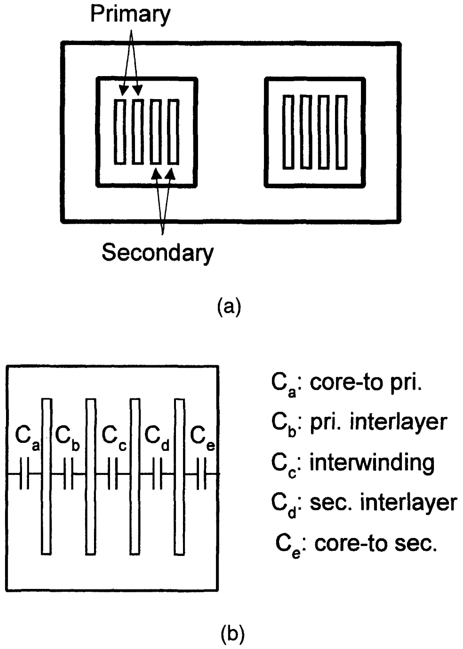

FIG. 15. (a) Example of transformer and (b) detail of windings showing parasitic capacitances.

different windings) are the most relevant ones.

Let us consider two adjacent layers AB and CD (which may or may not belong to the same winding) as in Fig. 16(a). During transformer operation, the voltage of a layer at a given moment changes linearly from point to point. Because of this, the equivalent capacitance is different from the static capacitance C_0 of Fig. 16(b), and it can be calculated from the energy stored between the layers. It can be easily shown that the equivalent capacitance referred to a terminal with voltage V is

$$C_{\text{eq}} = \frac{C_0(V_{\text{AC}}^2 + V_{\text{AC}}V_{\text{BD}} + V_{\text{BD}}^2)}{3V^2}. \tag{37}$$

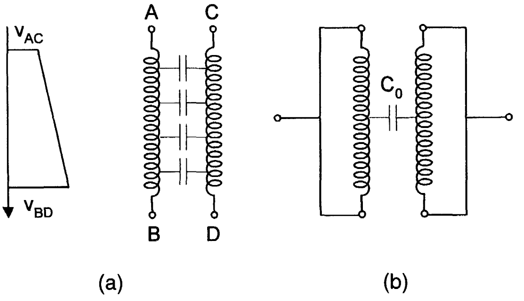

FIG. 16. (a) Calculation of parasitic capacitance. (b) dc measurement of interwinding capacitance does not give correct result.

The different capacitances can be calculated and then combined together.

A Π network of lumped capacitances C_1, C_2, and C_{12} is frequently used for modeling purposes, as shown in Fig. 17(a). These capacitances have an effect on the high-frequency behavior of the transformer. In order to operate a transformer at higher frequencies, it is necessary to try to reduce them. A way to eliminate undesirable capacitive couplings is to use a *screen,* i.e., a grounded conductive layer placed between the two layers of windings where no capacitive coupling is desired.

2.6 Other Considerations

In the design of transformers, various additional considerations are of importance: insulation, thermal management, magnetic shielding, and mechanical construction. Refer to Flanagan (1993), Grossner (1983), and Lee *et al.* (1988) for further discussion of these issues. For more discussion on design procedures, refer to McLyman (1978, 1981), Ray and Sartori (1971), and Ohri *et al.* (1976).

3. WIDE-BAND TRANSFORMERS

A wide-band transformer (see Lord, 1950; Flanagan, 1993; Grossner, 1983; Lee *et al.*, 1988) is a transformer specifically designed to transmit signals with a wide frequency spectrum accurately and without distortion. Wide-band transformers are analyzed in the frequency domain, and the desired performance characteristics are also specified in the frequency domain. They are used to transmit complex wave forms like voice and music or a train of pulses.

The requirement of undistorted transmission of signals puts a limit to the operating frequency range of wide-band transformers. A typical frequency response is shown in Fig. 18. In the midband frequency range, the am-

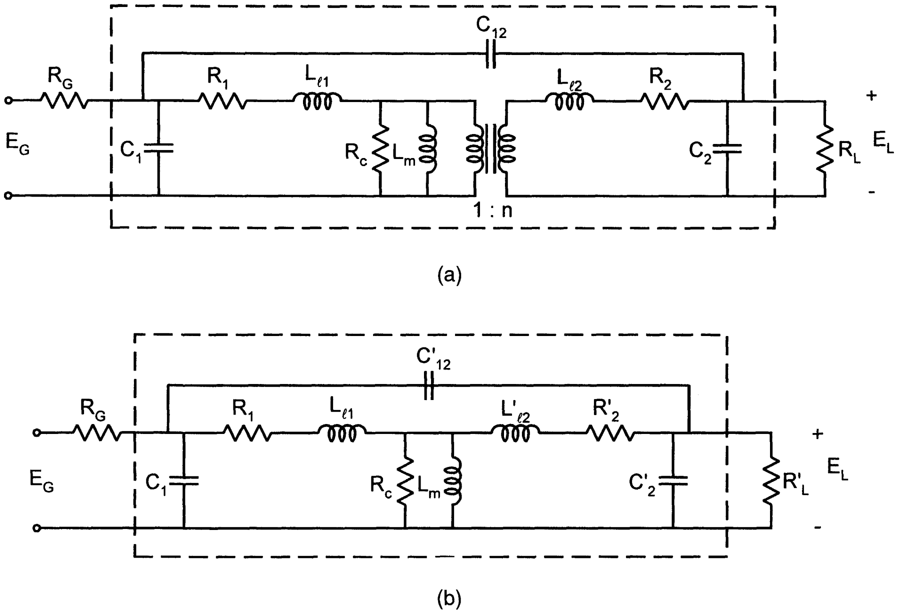
C12
RG
R1
Lℓ1
Lℓ2
R2
+
EG
C1
Rc
Lm
C2
RL
EL
-
1 : n
(a)
C'12
RG
R1
Lℓ1
L'ℓ2
R'2
+
EG
C1
Rc
Lm
C'2
R'L
EL
-
(b)

FIG. 17. (a) Complete transformer model and (b) complete model with secondary reflected to primary.

plitude and phase are constant, which means that the signal will be transmitted without distortion. In the low- and high-frequency ranges, the frequency response amplitude decreases, and the signal is distorted.

3.1 Transformer Equivalent Models

Equivalent models are used to study the behavior of wide-band transformers. The problem in the modeling of a wide-band

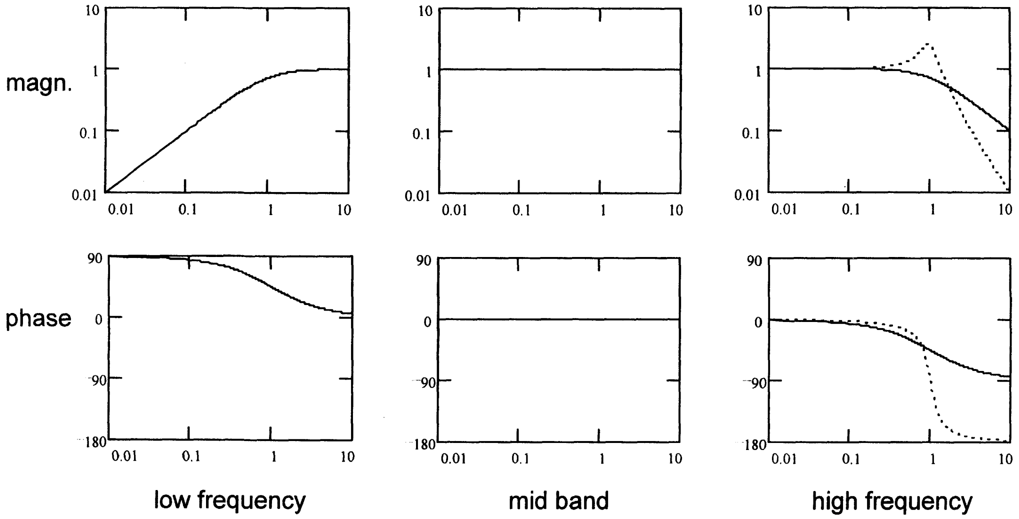

FIG. 18. Wide-band transformer frequency response.

transformer is in finding a model complete enough to describe accurately the behavior of the transformer over the desired frequency range and yet simple enough to lend itself to analysis and design.

In order to study the performance of a transformer in a given electric circuit, the characteristics of the electric circuit itself (source impedance and load impedance) must be taken into account. In the following, source and load impedances are assumed to be purely resistive. In the equivalent circuit of Fig. 17(a), R_G is the source resistance and R_L is the load resistance.

A reasonably complete equivalent model for a wide-band transformer is shown in Fig. 17(a). This model consists of an ideal 1:n transformer, magnetizing inductance L_m, resistance R_c that models core loss, primary and secondary leakage inductances L_{l1} and L_{l2}, primary and secondary winding resistances R_1 and R_2, and the Π network of lumped capacitances C_1, C_2, and C_{12} that model the distributed capacitances described in Sec. 2.5.

Usually the turns ratio is chosen to match the load to the source, i.e.,

$$n = \sqrt{R_L/R_G}. \qquad (38)$$

This ensures maximum power transfer and also prevents reflections from transmission-line effects. Half of the power is delivered to the load, and the other half is dissipated in the source resistance R_G.

A first simplifying step is to "refer" the entire circuit to the primary, eliminating the 1:n ideal transformer. The result is shown in Fig. 17(b), where

$$L'_{l2} = L_{l2}/n^2, \qquad (39)$$

$$R'_2 = R_2/n^2, \qquad (40)$$

$$R'_L = R_L/n^2. \qquad (41)$$

Capacitances C_1, C_2, and C_{12} become (see M.I.T., 1943, pp. 396–398)

$$C'_1 = C_1 + C_{12}(1 - n), \qquad (42)$$

$$C'_2 = n^2C_2 + nC_{12}(n - 1), \qquad (43)$$

$$C'_{12} = nC_{12}. \qquad (44)$$

For analysis purposes, the transformer models of Figs. 17(a) and 17(b) are assumed to be linear and all circuit elements to be constant. It should be kept in mind that this is an approximation: magnetizing inductance is nonlinear (see Sec. 2.1.1), core losses are frequency dependent, winding resistances change with frequency as an effect of skin and proximity effects, and distributed capacitances are only approximated by lumped capacitances.

The equivalent circuit of Fig. 17(b) is fairly complicated. It is customary to study separately the behavior of the wide-band transformer in the midband frequency, low-frequency, and high-frequency ranges. Appropriate approximations simplify the analysis in each case. The starting point of the following analysis is the equivalent circuit reflected to the primary side of Fig. 17(b). For simplicity, the apices will be omitted in the notation, e.g., reflected load resistance R'_L becomes R_L.

3.2 Midband Frequency Response

Midband frequency is defined as the frequency range where all reactive elements can be neglected: shunting reactive elements have high impedance, and series reactive elements have low impedance. Winding resistances R_1 and R'_2 can be included in the generator or load resistances. The midband equivalent circuit is shown in Fig. 19. The ratio of output to input voltage is

$$\frac{E_L}{E_G} = \frac{R_L}{R_L + R_G}. \qquad (45)$$

3.3 Low-Frequency Response

As frequency decreases, the impedance of the magnetizing inductance decreases. A significant magnetizing current flows in L_m and causes a voltage drop on the source resis-

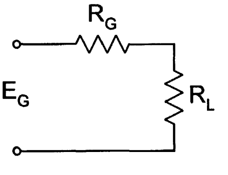

FIG. 19. Midband frequency equivalent model.

tance. As a result, the output voltage decreases.

The equivalent circuit of Fig. 20(a) is used. Application of Thévenin's theorem gives the circuit of Fig. 20(b). Resistance R_{par} is defined as

$$R_{\text{par}} = R_G R_L/(R_G + R_L). \tag{46}$$

The ratio of output to input voltage is

$$\frac{E_L}{E_G} = \frac{R_L}{R_L + R_G}\frac{j\omega L_m}{j\omega L_m + R_{\text{par}}}. \tag{47}$$

This equation can be rewritten as

$$\frac{E_L}{E_G} = \frac{R_L}{R_L + R_G}\frac{1}{1 + \omega_L/j\omega}, \tag{48}$$

where

$$\omega_L = R_{\text{par}}/L_m. \tag{49}$$

The first term of Eq. (48) represents the midband attenuation and the second term the low-frequency pole due to the magnetizing inductance. From Eq. (49) it follows that to extend the transformer operation to lower frequencies it is necessary either to decrease R_{par} or to increase L_m. Since usually the transformer designer does not have control over the value of R_{par}, the only option is to increase L_m, which requires an increase in the physical size of the transformer. Therefore, the desired low-frequency performance dictates transformer size.

3.3.1 Low-Frequency Distortion Distortion due to nonlinear behavior of the system is not modeled by the linear equivalent circuit of Fig. 20. Two distinct causes of distortion can be identified:

1. Ferromagnetic distortion. It is due to nonlinear L_m, and it is discussed in Sec. 2.1.3. It is a problem at low frequency in particular, because, for an ac signal of constant amplitude, the ac flux increases as frequency decreases. At some point core saturation ensues. Usually distortion becomes unacceptable before saturation occurs.
2. Generator distortion. A generator of limited volt-ampere capabilities or with nonlinear source impedance can cause distortion.

3.4 High-Frequency Response

At high frequency the impedance of shunt magnetizing inductance L_m becomes large, and it can be neglected. Two factors limit the high-frequency response: impedance of series leakage inductance increases, and it acts as a voltage divider with load impedance; the impedance of shunt capacitances decreases, so that more current is drawn from the source and flows through the source resistance causing the output voltage to drop.

The equivalent circuit at high frequency is shown in Fig. 21(a). Leakage inductance L_l is the combination of primary and secondary leakage inductances L_{l1} and L_{l2}. This circuit is fairly complicated to analyze. Two special cases will be considered here.

3.4.1 Model with Leakage Inductance If capacitances are neglected (usually a valid approximation for audio transformers, i.e., transformer with an upper frequency below 20 kHz), the circuit of Fig. 21(b) can be studied. The voltage ratio is

$$\frac{E_L}{E_G} = \frac{R_L}{R_L + R_G}\frac{1}{1 + j\omega/\omega_{H1}} \tag{50}$$

with

$$\omega_{H1} = R_{\text{ser}}/L_l, \tag{51}$$

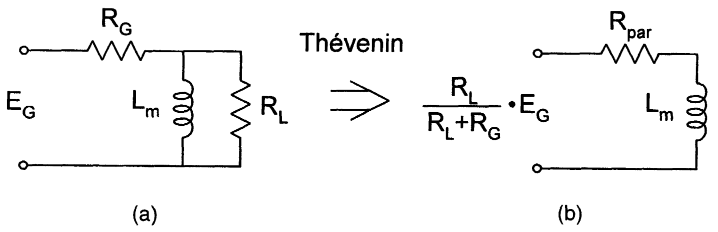

FIG. 20. Low-frequency equivalent model.

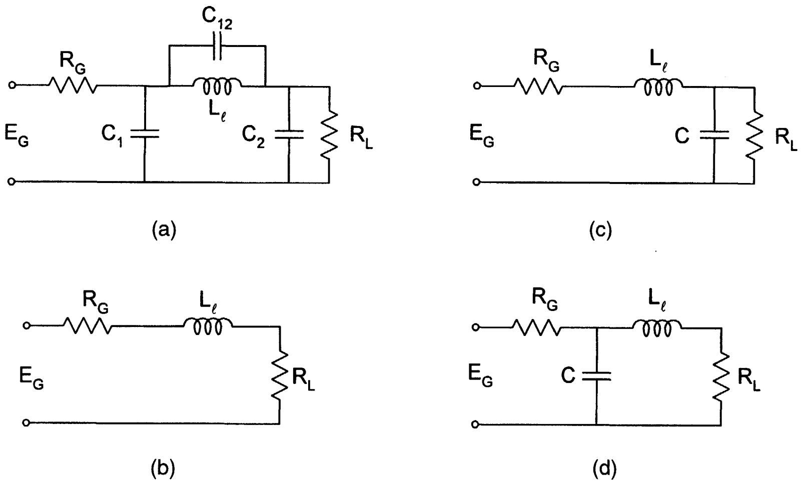

FIG. 21. High-frequency equivalent models: (a) complete model, (b) model with leakage inductance only, (c) step-up transformer, and (d) step-down transformer.

$$R_{\text{ser}} = R_G + R_L. \tag{52}$$

The frequency response has a pole at frequency ω_{H1}.

The ratio of the highest to lowest operating frequencies, a figure of merit for transformers, is given by the ratio of (51) to (49),

$$\frac{\omega_{H1}}{\omega_L} = \frac{R_{\text{ser}}}{R_{\text{par}}}\frac{L_m}{L_l} \tag{53}$$

and is proportional to the ratio of magnetizing to leakage inductance. The frequency range can be increased by any means that increases the magnetizing inductance without a corresponding increase of leakage inductance—for example, using high-permeability core materials. On the other hand, increasing number of turns does not help: both inductances L_m and L_l increase, and the operating frequency band moves to lower frequencies.

3.4.2 Model with Leakage Inductance and Shunt Capacitance In wide-band transformers operating beyond the audio-frequency range, shunt capacitances C_1 and C_2 usually cannot be neglected. It is frequently possible to include only one of these capacitances. Since the shunting effect is more significant on the high-impedance side of the transformer, the capacitance of the side with higher number of turns is usually retained. The case of a step-up transformer with secondary capacitance C is considered [Fig. 21(c)]. The output-to-input voltage ratio is

$$\frac{E_L}{E_G} = \frac{R_L/(1 + j\omega CR_L)}{R_G + j\omega L_l + R_L/(1 + j\omega CR_L)}, \tag{54}$$

and, after rearranging,

$$\frac{E_L}{E_G} = \frac{R_L}{R_L + R_G}\left[1 + \frac{j\omega}{\omega_{H2}}\left(\frac{K_1K_2 + 1}{K_1 + K_2}\right) - \left(\frac{\omega}{\omega_{H2}}\right)^2 \frac{K_2}{K_1 + K_2}\right]^{-1} \tag{55}$$

with

$$\omega_{H2} = 1/\sqrt{L_lC}, \tag{56}$$

$$Z = \sqrt{L_l/C}, \tag{57}$$

$$K_1 = R_G/Z, \tag{58}$$

$$K_2 = R_L/Z; \tag{59}$$

ω_{H2} is the corner frequency, Z is the transformer characteristic impedance, and K_1 and K_2 are dimensionless parameters. Notice that when $K_1 = K_2$ load and source resistances are matched. If $K_1 = K_2 = 1$, the transformer characteristic impedance is also matched to the source and load.

3.5 Special Types of Wide-Band Transformers

The wide-band transformer is normally used to perform basic functions such as matching a generator to its load. In this section, two special types of wide-band transformers are briefly described: hybrid transformers and transmission-line (balun) transformers.

Hybrid transformers are multiport transformers used to combine and split signals. They are common in communications. For more information, refer to Sartori (1968).

Transmission-line transformers (also called *balun transformers*) are used at high frequencies when dc isolation is not required. Usually the high-frequency operation of transformers is limited by resonance of distributed capacitance with leakage inductance. Balun transformers (see Ruthroff, 1959) are built as transmission lines with the windings arranged so that interwinding capacitance becomes part of the characteristic impedance of the line. The transmission line is usually realized with bifilar winding wound around a toroid. Very broad bandwidths can be obtained, as high as 0.1 to 1000 MHz within ±1 dB.

4. PULSE TRANSFORMERS

As the name implies, pulse transformers (see Lord, 1971; Glasoe and Lebacqz, 1948) are transformers specifically designed to transmit pulses. The term pulse indicates a square wave with low repetition rate.

A typical application of pulse transformers is as an interface between a pulse generator and a magnetron oscillator in a radar system. The pulse transformer provides matching of the pulse generator to the load for maximum power transfer, electrical insulation, and, if needed, polarity reversal. In this type of application the power delivered during the pulse can be very large. Pulse transformers have been built capable of delivering pulses of amplitude in the tens of kilovolts and power in the tens of megawatts. Pulse transformers are also used in pulse-type blocking oscillators and solid-state trigger circuits for firing SCRs and other power semiconductors.

Pulse transformers are specified and studied in the time domain. Figure 22 illustrates the output of a pulse transformer for a rectangular input pulse. Some commonly used terms describing features of the pulse are shown in this figure. Three separate periods can be identified: a rise period in which the voltage ramps up, a top period in which the voltage is constant, and an off period in which the voltage goes back to zero. Ideally, the pulse should be faithfully transmitted by the transformer. The three periods are now separately considered and different equivalent circuits identified.

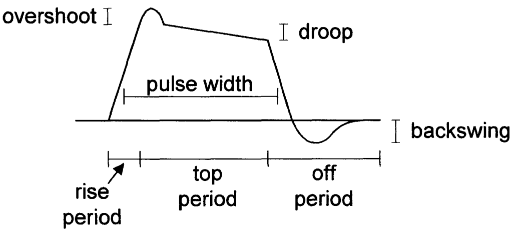

FIG. 22. Pulse response of transformer.

4.1 The Rise Period

The equivalent circuit is shown in Fig. 23. This circuit is identical to the high-frequency model of Fig. 21(c), because in this period the transformer is excited with high-frequency components. The pulse rise is the response of the system to a step of amplitude E_G applied at the input.

The Laplace transform expression of the step response is [see (54)]

$$\frac{E_L(s)}{E_G} = \frac{R_L}{R_L + R_G}\left\{s\left[1 + s\left(\frac{R_G}{L_l} + \frac{1}{R_L C}\right) + s^2 L_l C \frac{R_L}{R_L + R_G}\right]\right\}^{-1}. \tag{60}$$

In the second-order term in the denominator

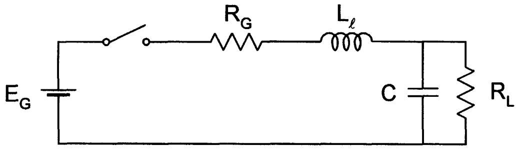

FIG. 23. Rise interval equivalent circuit.

of (60), a natural frequency ω_R and a damping factor ζ_R can be identified:

$$\omega_R = \sqrt{\frac{R_L + R_G}{R_L}} \cdot \frac{1}{\sqrt{L_l C}}, \tag{61}$$

$$\zeta_R = \frac{R_G R_L C + L_l}{2\sqrt{R_L L_l C (R_G + R_L)}}. \tag{62}$$

Equation (60) becomes

$$\frac{E_L(s)}{E_G} = \frac{R_L}{R_L + R_G} \cdot \frac{1}{s[1 + 2\zeta_R s/\omega_R + (s/\omega_R)^2]}. \tag{63}$$

The normalized response

$$\frac{e_L(t)}{E_G} \frac{R_L + R_G}{R_L}$$

is shown in Fig. 24 as a function of normalized time

$$T = (\omega_R/2\pi)t \tag{64}$$

with damping ζ as a parameter. From (64) it is clear that a large value of ω_R guarantees a fast response. That requires keeping leakage inductance L_l and capacitance C as small as possible. Notice that for low damping the response is faster but it is oscillatory, whereas for high damping it is monotonic but slower. Usually a big overshoot is undesirable, and the transformer is designed to be critically damped or slightly underdamped.

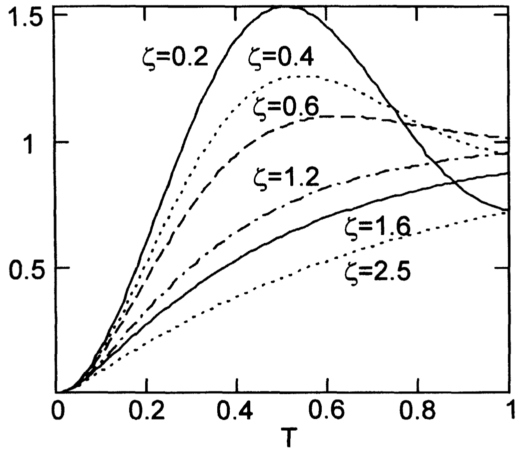

FIG. 24. Normalized rise period response as a function of damping ζ.

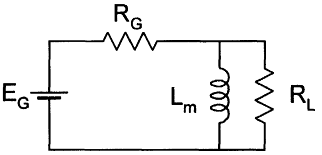

FIG. 25. Top period equivalent circuit.

4.2 The Top Period

The top of the pulse is studied using the equivalent circuit of Fig. 25, which is identical to the low-frequency equivalent circuit of Fig. 20. The top of the pulse depends on the value of magnetizing inductance. The output voltage has a droop, which becomes more pronounced the smaller the magnetizing inductance.

The response is

$$\frac{e_L(t)}{E_G} = \frac{R_L}{R_L + R_G} \exp(-\omega_L t) \tag{65}$$

with

$$\omega_L = R_{\text{par}}/L_m. \tag{66}$$

4.3 The Off Period

At the end of the pulse the supply voltage source E_G is usually opened (disconnected). The equivalent circuit of Fig. 26 is used for the analysis. The initial conditions are

$$V_C(0) = E_G, \tag{67}$$

$$I_L(0) = E_G \tau / L_m, \tag{68}$$

where E_0 is the voltage at the top of the pulse, neglecting the droop for simplicity, and τ is the pulse duration.

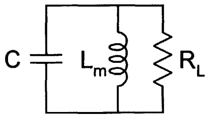

FIG. 26. Off period equivalent circuit.

The Laplace-transform expression of the response is

$$E_L(s) = \frac{sL_mCV_C(0) - L_mI_L(0)}{1 + sL_m/R_L + s^2L_mC}. \tag{69}$$

We can define the quantities

$$\omega_0 = 1/\sqrt{L_mC}, \tag{70}$$

$$Z = \sqrt{L_m/C}, \tag{71}$$

$$D = Z/2R_L, \tag{72}$$

$$T = (\omega_0/2\pi)t, \tag{73}$$

$$F = \omega_0\tau, \tag{74}$$

where ω_0 is the natural frequency of the LC circuit, Z the characteristic impedance, D the damping factor, T the normalized time, and F the initial current factor. Equation (69) becomes

$$\frac{E_L(s)}{E_G} = \frac{(1/\omega_0)(s/\omega_0 - F)}{1 + 2Ds/\omega_0 + s^2/\omega_0^2}. \tag{75}$$

The response is a function of two parameters, the initial current factor F and the damping factor D. Two families of curves, for $F = 0$ and $F = 1$, are shown in Fig. 27. A certain amount of backswing is unavoidable. A more serious problem is that if the system exhibits oscillatory behavior (low value of damping D), the backswing will become positive again on the first oscillation. This may be erroneously interpreted as a second pulse.

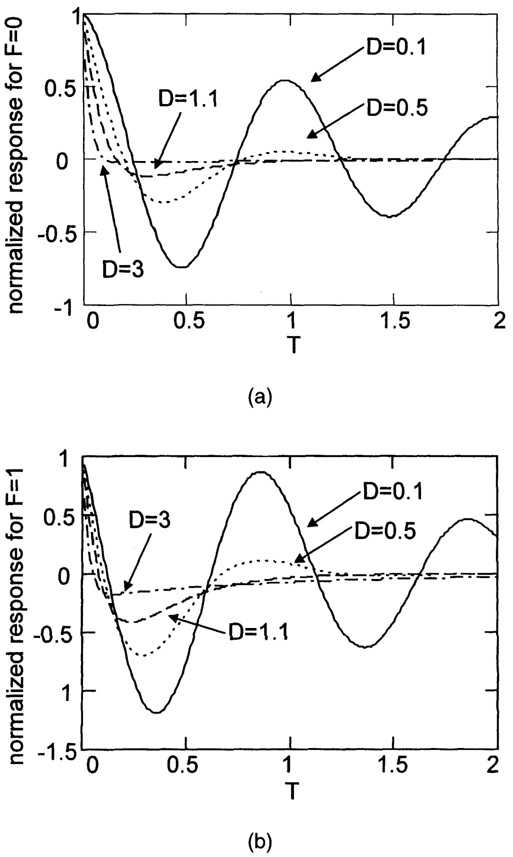

(a)

(b)

FIG. 27. Off period response of the transformer for (a) $F = 0$ and for (b) $F = 1$.

4.4 Design Issues

Here are some comments relevant to the design of pulse transformers:

1. *Leading-edge response.* In order to have a fast leading-edge response, it is critical to make the corner frequency ω_R as high as possible, which means to make L_l and C small. This requires careful choice of winding geometry and coil arrangements.
2. *Top-period response.* The performance of the pulse transformer during the top period depends on magnetizing inductance L_m. The maximum acceptable droop determines the minimum value of L_m. It is desirable to use high-permeability core materials, which allow large values of L_m with relatively few turns. In this way L_l and C can be small, and a satisfactory leading-edge response can be obtained.
3. *Off-period response.* The backswing response is usually of less concern to the pulse transformer designer. If needed, external circuits can be added to improve it.

It is interesting to look at the *B-H* loop of a pulse transformer as in Fig. 28. In this figure both the dc and the pulse loop are shown. The *pulse permeability* μ_p, given by the slope of a line connecting the apices of the pulse loop, is smaller than the dc permeability of the material. The area of a pulse loop represents core losses per unit volume, as explained in Sec. 2.1.2.

If the core is ungapped as in Fig. 28(a), at

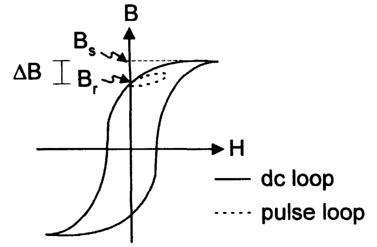

ungapped core
no resetting

(a)

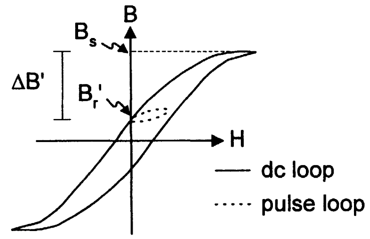

gapped core
no resetting

(b)

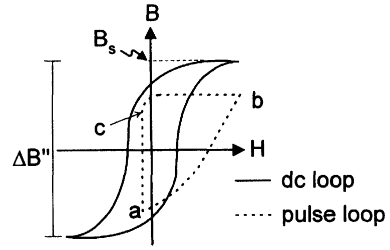

ungapped core
with core reset

(c)

FIG. 28. *B-H* loop of pulse transformer in various cases.

the end of each pulse the magnetization returns to the dc residual magnetization B_r. The available flux swing ΔB is limited by core saturation flux density B_s and is

$$\Delta B = B_s - B_r. \tag{76}$$

If the core is gapped as in Fig. 28(b), the sheared loop has much lower residual magnetization B_r' and the available flux swing $\Delta B'$ is correspondingly increased. Another way to increase flux swing capability is to "reset" the core at the end of each pulse, i.e., to force a negative current in one of the windings (possibly an auxiliary winding) with the effect of moving the magnetization to a negative value. Figure 28(c) shows a pulse loop with reset using a constant negative current (vertical line c–a).

5. ELECTRONIC POWER TRANSFORMERS

For simplicity, only the case of high-frequency dc-to-dc converters (see Middlebrook and Ćuk, 1983; Kassakian *et al.*, 1991) is considered here. Other types of switching converters (ac-to-dc, dc-to-ac, and ac-to-ac) can usually be considered as modifications of dc-to-dc converters.

Three nonisolated converters are shown in Fig. 29. In all cases, transistor Q is operated as a controllable switch: It is turned on during interval t_{ON} and turned off during interval t_{OFF}. The "freewheel diode" D acts as a complementary switch and turns on during the OFF period to give the inductor current a path to flow. The voltage gain of these converters can be derived under the assumption

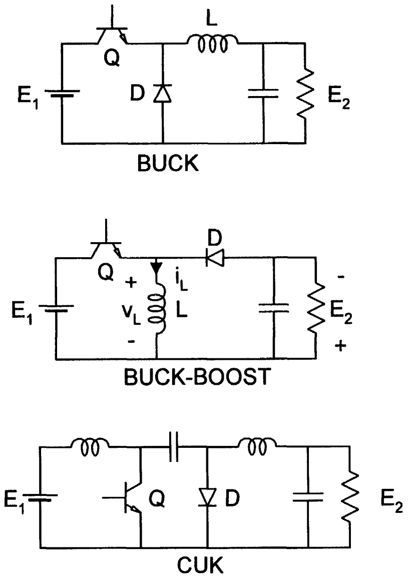

FIG. 29. Nonisolated converters.

of steady-state operation. For example, in the case of the buck-boost converter of Fig. 29(b), inductor voltage v_L is equal to the input voltage E_1 during the ON interval and to the output voltage E_2 during the OFF interval. Since in steady state the average inductor voltage during a period must be zero (otherwise inductor current would grow without limits), the voltage balance during a period is

$$E_1 t_{\mathrm{ON}} = E_2 t_{\mathrm{OFF}}. \tag{77}$$

From this equation, the voltage gain of the converter is

$$E_2/E_1 = t_{\mathrm{ON}}/t_{\mathrm{OFF}}. \tag{78}$$

This shows that output voltage can be controlled by varying the ratio of ON interval to OFF interval.

Operating wave forms are shown in Fig. 30. The switch voltage v_Q is a square wave, and the inductor current is a triangular wave.

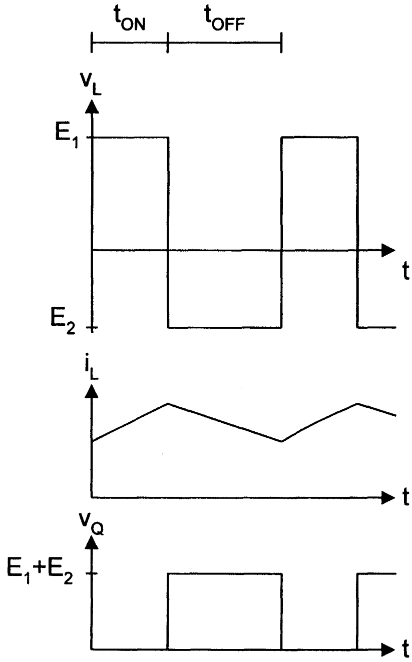

FIG. 30. Operating wave forms of buck-boost converter.

5.1 Transformers in Switching Converters

Transformers are introduced in dc-to-dc converters for two reasons: to provide electrical insulation and to reduce component stresses that result when input–output voltage conversion ratio is far from unity.

The converters of Fig. 29 can be modified to accommodate an insulation transformer as shown in Figs. 31–33. It is interesting to examine these converters, because the insulation transformers operate in three different regions of the *B-H* loop.

5.1.1 Forward Converter The forward converter shown in Fig. 31(a) is a modifica-

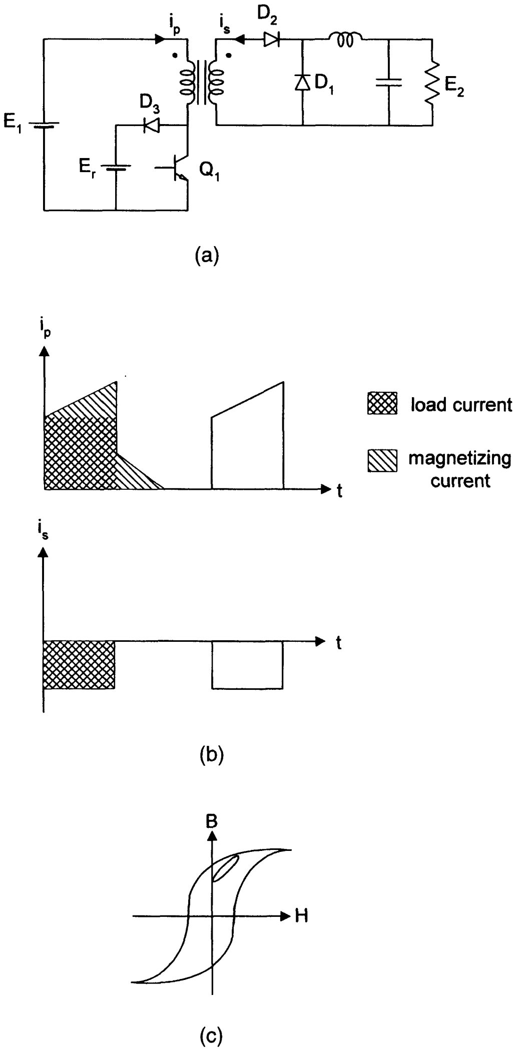

(a)

(b)

(c)

FIG. 31. Forward converter.

tion of the buck converter of Fig. 29(a). Additional elements are added to allow *resetting* of the transformer. During the ON interval, switch Q conducts, and primary voltage is equal to E_1. In order to prevent core saturation, a negative voltage must be applied during the OFF interval, so that the average transformer voltage is zero (transformer reset). Diode D_2 is needed to allow the reset: Without it, conduction of diode D_1 during the OFF interval would short the secondary and make reset impossible. During the OFF interval diode D_3 conducts, and auxiliary voltage E_r resets the transformer. Figure 31(b) shows primary and secondary currents distinguishing between magnetizing and load components. The B-H loop of Fig. 31(c) shows that the converter operates with excitation in one direction only. The resetting circuit brings magnetic field H back to zero at the end of the cycle. A small gap is usually inserted in the transformer to reduce remanent magnetization.

5.1.2 Flyback Converter The flyback converter of Fig. 32 is a straightforward modification of the buck-boost converter. As can be seen from the current wave forms, energy is stored in the transformer during the ON interval and delivered to the load during the OFF interval. The transformer is required to store significant energy, and all the current in the transformer is magnetizing current. There is no load current component since primary and secondary never conduct at the same time. The flyback transformer has a significant gap to increase current capability: The transformer operates along a minor B-H loop in the first quadrant as shown in Fig. 32(c).

5.1.3 Isolated Ćuk Converter In the isolated Ćuk converter of Fig. 33(a), the transformer is a true ac transformer. Capacitors C_1 and C_2 block any dc voltage. The magnetizing and load components are shown in Fig. 33(b). The load current is a square wave, and triangular components represent the magnetizing current. The B-H loop shown in Fig. 33(c) is centered around the origin, and it has no dc bias as in the other two cases.

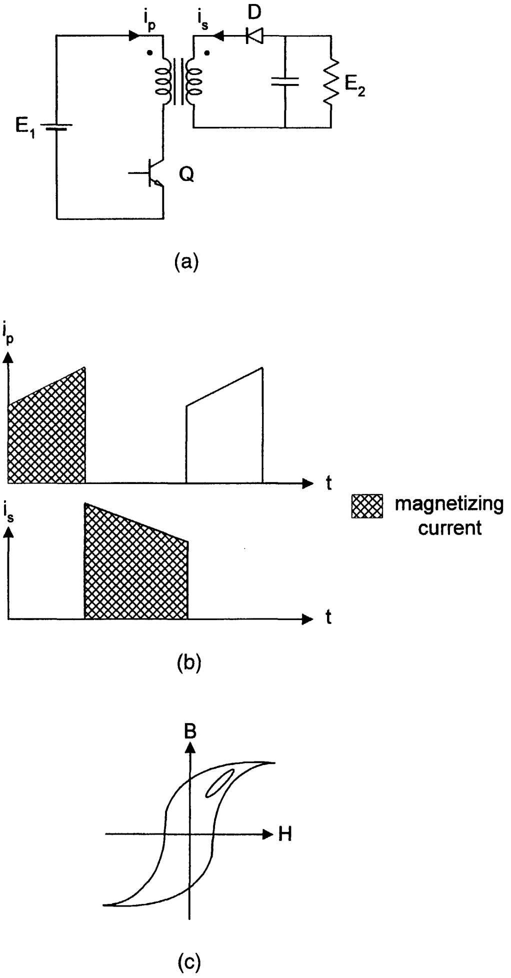

FIG. 32. Flyback converter.

5.2 Requirements for Electronic Power Transformers

The overriding requirements of switching converters are high efficiency and small size. In recent times switching converters have been operated at higher and higher frequencies in order to reduce the reactive component sizes. Commercial converters usually operate at frequencies between 30 and 300 kHz, with some converters operating at 1 MHz or more.

Transformers in switching converters are excited by a train of voltage pulses at the switching frequency. The techniques used in the study of pulse transformers may be used in this case as well (see Lee *et al.*, 1988). In the following, specific requirements of trans-

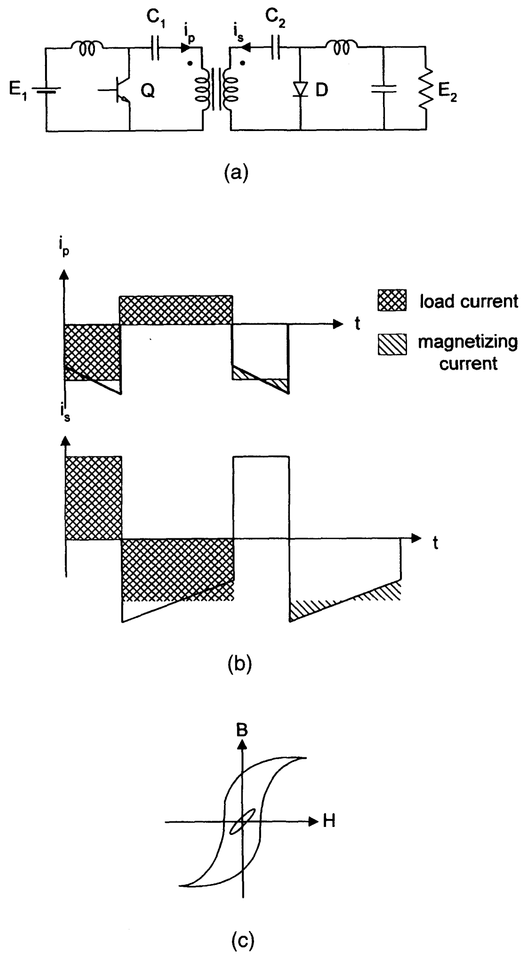

FIG. 33. Isolated Ćuk converter.

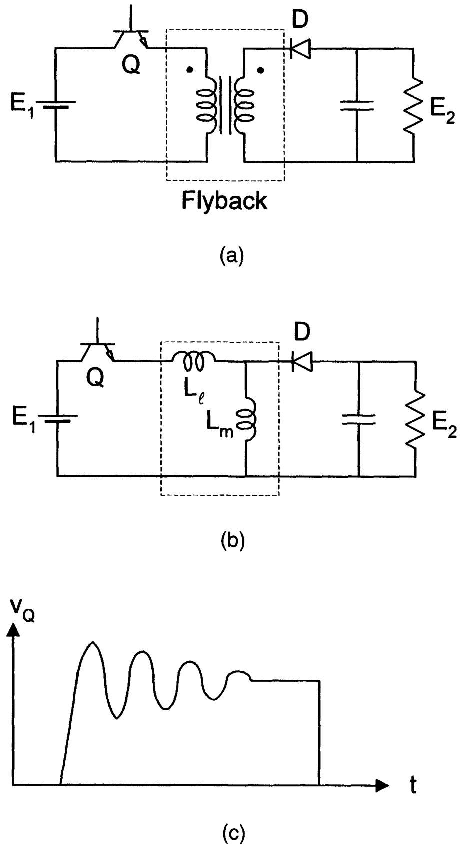

FIG. 34. (a) Flyback converter, (b) flyback converter with transformer equivalent model, and (c) switch wave form with overshoot.

formers in switching converters are discussed.

1. *Magnetizing inductance.* The value of magnetizing inductance is dictated by the switching frequency and by the desired amount of magnetizing current.
2. *Leakage inductance.* Leakage inductance should be reduced as much as possible, because it causes overshoot and ringing of the switch voltages. To illustrate this, the case of a buck-boost and its isolated counterpart, the flyback converter, is considered. In the nonisolated case, the main switch wave form in a properly designed converter is not very different from the idealized wave form of Fig. 30. The case of the flyback converter is shown in Fig. 34(a). Introducing an equivalent circuit for the transformer consisting of magnetizing inductance L_m and leakage inductance L_l, we obtain the circuit of Fig. 34(b). The main switch wave form is shown in Fig. 34(c). When the switch turns off, the leakage inductance resonates with all the parasitic capacitances of the circuit, such as switch capacitance and transformer capacitance. Since the converter is designed for high efficiency, the oscillation is lightly damped. Usually, a *snubber* circuit is added to damp this oscillation. Significant energy is dissipated in these circuits unless extra circuitry is added to recover the snubber energy. In any case, the penalty for leakage inductance is either lower efficiency or added cost. For this reason, power electronics transformers are built for low leakage, with extensive interleaving. Safety requirements of high-voltage insu-

lation between primary and secondary sometimes make it difficult to achieve the desired low leakage.

3. *Transformer capacitance.* Primary-to-secondary transformer capacitance should be kept low for EMI (electromagnetic interference) reasons. During transformer operation, this capacitance acts as a source of common mode noise (see Nave, 1991). Shielding is frequently necessary to meet EMI emission requirements.
4. *Core material.* For high-frequency operation the core material of choice is ferrite for its low eddy-current losses. Hysteresis core losses, however, can still be significant, especially at higher frequencies. The ac flux swing is frequently limited by core losses rather than by saturation.
5. *Copper losses.* The windings must be sized to carry the load currents with reasonably low losses. Skin and proximity effects can increase copper losses significantly.

5.3 Coupled Inductors and Zero Current Ripple

In certain converters, separate inductors can be combined together in transformerlike structures called *coupled inductors* in the case of two inductors and *integrated magnetics* in the case of three or more inductors. In the case of a Ćuk converter it can be shown that the two inductors have identical voltage wave forms, and they can therefore be combined in a coupled inductor structure, usually with significant reduction in size. If the coupled inductors meet a certain zero current ripple condition (having to do with leakage and turns ratio), the current ripple can be steered into one winding, leaving the other one virtually ripple free. The situation is depicted in Fig. 35. The output current is pure dc, and the filtering capacitor C can be reduced with savings in size and cost. For an explanation of the zero ripple phenomenon, refer to Ćuk (1978), Polivka and Ćuk (1983), Zhang and Ćuk (1986), and Santi and Ćuk (1993, 1994).

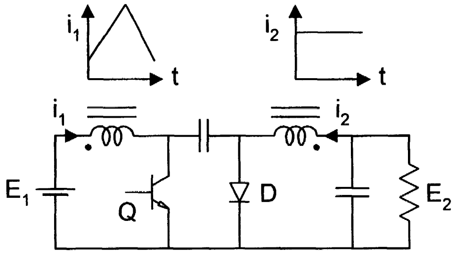

FIG. 35. Coupled-inductor Ćuk converter with zero current ripple at the output.

6. MAGNETIC AMPLIFIERS

Saturation in magnetic cores, usually regarded as a drawback, is put to use in a number of electronic circuits. Amplifiers with saturable reactive elements are called *magnetic amplifiers.* In this section one example of magnetic amplifiers is given. For more information refer to Frost-Smith (1966), Storm (1955), and Fink and Christiansen (1989).

The main idea is as follows. In an ungapped transformer a small amount of dc current can cause the core to saturate. Suppose one winding of a transformer is interposed between an ac source and a load. The power delivered to the load will be almost zero if the core is not saturated, because the transformer magnetizing impedance is very high. On the other hand, if the core is saturated, the magnetizing impedance is small and the ac source is effectively connected to the load. In summary, the saturable transformer acts as a switch, connecting or disconnecting the ac source to the load. The switch is controlled by a small amount of dc power applied to a second winding and used to saturate the core. Power amplification is possible in such an arrangement.

6.1 Simple Magnetic Amplifier

A so-called *simple magnetic amplifier* is shown in Fig. 36(a). In this circuit voltage E_{AC} is the unregulated supply voltage. The voltage applied to load R_L is controlled by dc control voltage E_c. The circuit uses two identical saturable transformers A and B. The secondaries are connected in series with the ac source E_{AC} and the load R_L. The primaries are connected in series with opposite polarity to the control voltage source E_c. This arrangement decouples the dc and the ac circuits. If only one transformer were used, ac voltages would be induced in the dc winding, and they would be effectively shorted by the dc voltage source E_c.

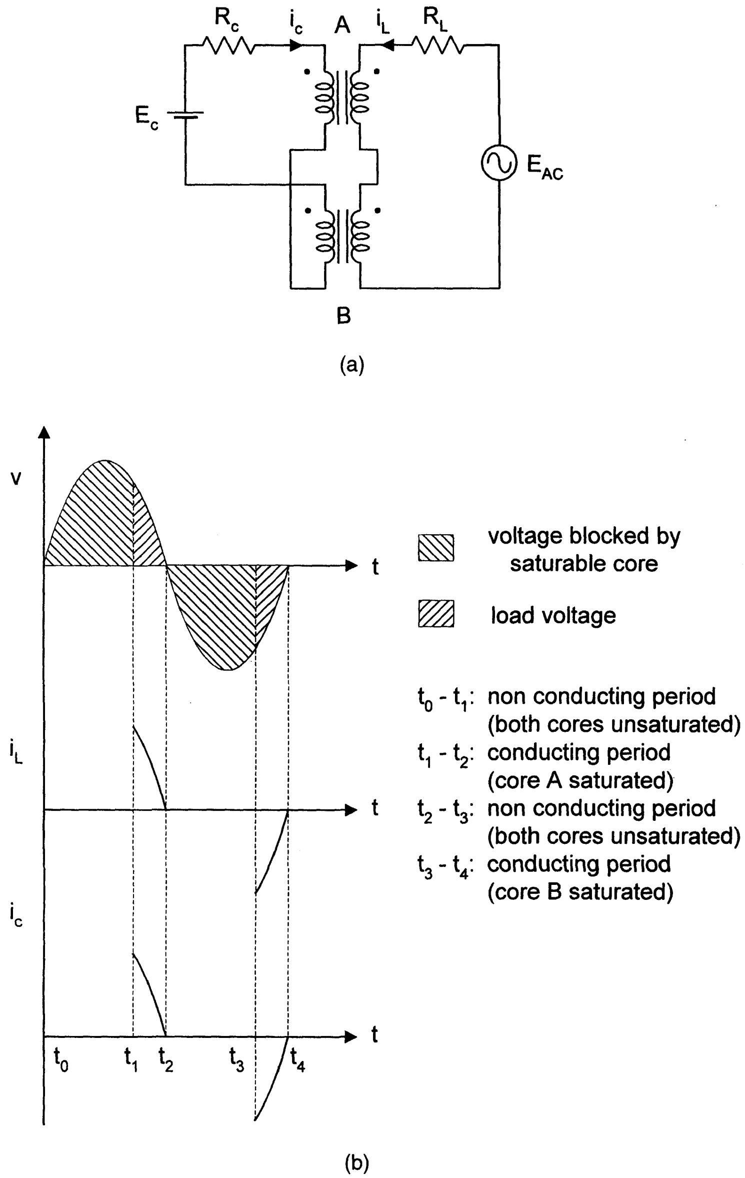
R_C
i_C
A
i_L
R_L
E_C
E_AC
B
(a)
V
t
i_L
t
i_C
t
t_0
t_1
t_2
t_3
t_4
▧ voltage blocked by saturable core
▨ load voltage
t_0 - t_1: non conducting period (both cores unsaturated)
t_1 - t_2: conducting period (core A saturated)
t_2 - t_3: non conducting period (both cores unsaturated)
t_3 - t_4: conducting period (core B saturated)
(b)

FIG. 36. Simple magnetic amplifier: (a) electrical circuit and (b) wave forms.

If current i_c is zero, the two transformers are designed not to saturate during the ac period. The high impedance of the transformer prevents current from flowing through the circuit, except for a small magnetizing current. The load voltage is practically zero.

Let us assume now that a positive current i_c is applied. Wave forms are shown in Fig. 36(b). This current adds to the magnetizing current in transformer A and subtracts from the magnetizing current in transformer B according to the current conventions of Fig. 36(a). At some instant t_1 during the positive semiperiod, transformer A saturates and ac supply voltage E_{AC} is applied to the load. During interval $t_1 - t_2$, current i_L is dictated by load R_L. Notice that transformer B is not saturated since the control current i_c demagnetizes it. If resistor R_c is small, current i_c is dictated by current i_L in the secondary of transformer B. In the case of 1:1 turns ratio, current i_c is identical to current i_L. During the other semiperiod the roles of transformers A and B are interchanged. In short, the magnetic amplifier provides control of the output voltage by controlling the saturation of the saturable transformer.

6.2 Other Magnetic Amplifier Circuits

Many different magnetic amplifier circuits are given in the literature. In order to reduce the dc control power, so-called *self-saturated magnetic amplifiers* saturate with zero control current delivering full voltage to the load. The control current unsaturates the core for part of the period reducing the output voltage. Another circuit (see Ramey, 1951a,b) generates an output voltage proportional to the dc control voltage. A circuit similar to the one in Fig. 36 can be used to measure a dc current (see Geyger, 1957, pp. 313–314). Magnetic amplifiers are also used in switching power converter applications (see Brkovic *et al.,* 1993; Vollin *et al.,* 1993).

GLOSSARY

Magnetic Amplifier: An amplifier that uses saturable reactive elements.

Pulse Transformer: A transformer specifically designed to transmit pulses.

Wide-Band Transformer: Transformer designed to transmit with minimal distortion a wide-band signal.

Works Cited

Boll, R., Warlimont, H. (1981), "Applications of Amorphous Magnetic Materials in Electronics," *IEEE Trans. Magn.* **MAG-17,** 3053–3058.

Brkovic, M., Pietkiewicz, A., Ćuk, S. M. (1993), "Novel Soft-Switching Converter with Magnetic Amplifiers," in: *Proceedings of the IECON '93—19th International Conference on Industrial Electronics, Control and Instrumentation,* Piscataway, NJ: IEEE, pp. 830–835.

Charap, S. (1974), "Magnetic Hysteresis Model," *IEEE Trans. Magn.* **MAG-10,** 1091–1096.

Chen, D. (1978), "Comparisons of High Frequency Core Losses: A Sinusoidal Voltage and a Square Wave Voltage," in: *PESC 78 Record. 9th Annual IEEE Power Electronics Specialists Conference,* Piscataway, NJ: IEEE, pp. 237–241.

Ćuk, S. M. (1978), "Switching Dc-to-Dc Converter with Zero Input or Output Current Ripple," in: *Proceedings of the IEEE Industry Applications Society Annual Meeting,* New York: IEEE, pp. 1131–1146.

Dowell, P. L. (1966), "Effects of Eddy Currents in Transformer Windings," *Proc. IEE* **113,** 1387–1394.

Fink, D., Christiansen, D. (1989), *Electronics Engineers' Handbook,* 3rd ed., New York: McGraw-Hill.

Flanagan, W. M. (1993), *Handbook of Transformer Design & Applications,* 2nd ed., New York: McGraw-Hill.

Frost-Smith, E. H. (1966), *The Theory and Design of Magnetic Amplifiers,* London: Chapman & Hall.

Geyger, W. A. (1957), *Magnetic-Amplifier Circuits,* 2nd ed., New York: McGraw-Hill.

Glasoe, G., Lebacqz, J. (1948), *Pulse Generators,* New York: McGraw-Hill.

Goldberg, A. F., Kassakian, J. G., Schlecht, M. F. (1989a), "Finite-Element Analysis of Copper Loss in 1–10-MHz Transformers," *IEEE Trans. Power Electron.* **4,** 157–167.

Goldberg, A. F., Kassakian, J. G., Schlecht, M. F. (1989b), "Issues Related to 1–10-MHz Transformer Design," *IEEE Trans. Power Electron.* **4,** 113–123.

Graham, C. D., Egami, T. (1980), "FeBC Alloys for Use in Power Transformers," *J. Magn. Magn. Mater.* **19,** 185–186.

Grossner, N. R. (1983), *Transformers for Electronic Circuits,* 2nd ed. New York: McGraw-Hill.

Kassakian, J. G., Schlecht, M. F., Verghese, G. C. (1991), *Principles of Power Electronics,* Reading, MA: Addison-Wesley.

Lee, R., Wilson, L., Carter, C. E. (1988), *Electronic Transformers and Circuits,* 3rd ed. New York: Wiley.

Lord, H. W. (1950), "The Design of Broad-Band Transformers for Linear Electronic Circuits," *Trans. Am. IEE* **69,** 1005–1010.

Lord, H. W. (1971), "Pulse Transformers," *IEEE Trans. Magn.* **MAG-7,** 17–28.

M.I.T. Electrical Engineering Staff (1943), *Magnetic Circuits and Transformers,* New York: Wiley.

McLyman, W. T. (1978), *Transformer and Inductor Design Handbook,* New York: Marcel Dekker.

McLyman, W. T. (1981), *Magnetic Core Selection for Transformers and Inductors,* New York: Marcel Dekker.

Middlebrook, R. D., Ćuk, S. M. (Eds.) (1983), *Advances in Switched-Mode Power Conversion,* Vols. 1-3, Pasadena, CA: TESLAco.

Nave, M. J. (1991), *Power Line Filter Design for Switched-Mode Power Supplies,* New York: Van Nostrand Reinhold.

Ohri, A., Wilson, T., Owen, H. (1976), "Design of Air Gapped Magnetic Core Inductors for Superimposed Direct and Alternating Currents," *IEEE Trans. Magn.* **MAG-12,** 564–574.

Polivka, W. M., Ćuk, S. M. (1983), "Analysis of Integrated Magnetics to Eliminate Current Ripple in: Switching Converters," in R. D. Middlebrook, S. M. Ćuk (Eds.), *Advances in Switched-Mode Power Conversion,* Vol. 3, Pasadena, CA: TESLAco, pp. 239–264.

Ramey, R. A. (1951a), "On the Mechanics of Magnetic Amplifier Operation," *Trans. Am. IEE* **70,** 1214.

Ramey, R. A. (1951b), "On the Control of Magnetic Amplifiers," *Trans. Am. IEE* **70,** 2124.

Ramo, S., Whinnery, J. R., Van Duzer, T. (1984), *Fields and Waves in Communication Electronics,* 2nd ed. New York: Wiley.

Raskin, D., Davis, L. (1981), "Metallic Glasses: A Magnetic Alternative," *IEEE Spectrum* **18** (11), 28–33.

Ray, R., Sartori, E. (1971), "Computer Design of an Inductor Carrying Direct Current," *IEEE Trans. Magn.* **MAG-7,** 453–455.

Ruthroff, C. L. (1959), "Some Broad-Band Transformers," *Proc. IRE* **47,** 1337–1342.

Santi, E., Ćuk, S. M. (1993), "Accurate Leakage Models of Gapped Magnetic Circuits," in: *APEC '93—Eighth Annual Applied Power Electronics Conference and Exposition. Conference Proceedings,* Piscataway, NJ: IEEE.

Santi, E., Ćuk, S. M. (1994), "Comparison and Design of Three Coupled Inductor Structures," in: *Proceedings of the IECON '94—20th International Conference on Industrial Electronics, Control and Instrumentation,* Piscataway, NJ: IEEE, pp. 262–267.

Sartori, E. (1968), "Hybrid Transformers," *IEEE Trans. Parts, Materials and Packaging* **PMP-4,** 59–66.

Smit, J. (1971), *Magnetic Properties of Materials,* New York: McGraw-Hill.

Snelling, E. C. (1969) *Soft Ferrites,* London: Iliffe Books.

Snelling, E. C. (1983), *Ferrites for Inductors and Transformers,* New York: Wiley.

Storm, H. (1955), *Magnetic Amplifiers,* New York: Wiley.

Tan, F. D., Vollin, J., Ćuk, S. M. (1995), "A Practical Approach for Magnetic Core-Loss Characterization," *IEEE Trans. Power Electron.* **10,** 124–130.

Vandelac, J., Ziogas, P. D. (1987), "A Novel Approach for Minimizing High-Frequency Transformer Copper Losses," in: *PESC '87 Record. 18th Annual IEEE Power Electronics Specialists Conference,* Piscataway, NJ: IEEE, pp. 355–367.

Vollin, J. Tan, F. D., Ćuk, S. M. (1993), "Magnetic Regulator Modeling," in: *APEC '93—Eighth Annual Applied Power Electronics Conference and Exposition. Conference Proceedings,* Piscataway, NJ: IEEE, pp. 604–611.

Watson, J. (1980), *Applications of Magnetism,* New York: Wiley.

Zhang, Z., Ćuk, S. M. (1986), "Coupled-Inductor Analysis and Design," in: *PESC 86 Record. 17th Annual IEEE Power Electronics Specialists Conference,* Piscataway, NJ: IEEE, pp. 655–665.

Further Reading

Bennet, E., Larson, S. C. (1940), "Effective Resistance of Alternating Currents of Multilayer Windings," *Trans. Am. IEE* **59,** 1010–1017.

Brauer, J. (1975), "Simple Equations for the Magnetization and Reluctivity Curves of Steel," *IEEE Trans. Magn.* **MAG-11,** 81.

Kelley, A. W., Edwards, S. W., Rhode, J. P., Baran, M. E. (1995), "Transformer Derating for Harmonic Currents: A Wideband Measurement Approach for Energized Transformers," in: *Proceedings of the IEEE Industry Applications Society Conference,* New York: IEEE, pp. 840–847.

Niemela, V. A. *et al.* (1989), "Calculating the Short Circuit Impedances of a Multiwinding Transformer from its Geometry," in: *PESC 89 Record. 20th Annual IEEE Power Electronics Specialists Conference,* Piscataway, NJ: IEEE, pp. 373–385.

Perry, M. P. (1979), "Multiple Layer Series Connected Winding Design for Minimal Losses," *IEEE Trans. Power Apparatus and Systems* **PAS-98,** 116–123.

Urling, A. M., Niemela, V. A., Skutt, G. R., Wilson, T. G. (1989), "Characterizing High Frequency Effects in Transformer Windings—A Guide to Several Significant Articles," in: *PESC 89 Record. 20th Annual IEEE Power Electronics Specialists Conference,* Piscataway, NJ: IEEE, pp. 607–617.

TRANSFORMERS, POWER

Enrico Levi, *Department of Electrical Engineering, Polytechnic University, Brooklyn, New York, U.S.A.*

Introduction 113
1. **Construction** 114
1.1 Core 114
1.2 Windings 114
1.3 Tank 116
1.4 Insulating Oils and Liquids 116
2. **Elementary Theory** 116
3. **Design Guidelines** 118
4. **Thermal Characteristics and Cooling** 120
4.1 Conduction 121
4.2 Convection 121
4.3 Radiation 121
5. **Polyphase Connections** 121
5.1 Three-Phase Transformations 121
5.2 Two- to Three-Phase Transformations 121
5.3 Zig-zag Connection 122
5.4 Three- to Six-Phase Transformations 123
5.5 Load Tap Changing 123
5.6 Autotransformer 123
5.7 Parallel Operation 124
6. **Harmonics and Noise** 124
7. **Special Transformers** 125
7.1 Voltage Regulators 125
7.2 Constant-Current Transformers 125
7.3 Instrument Transformers 125
7.4 Reactors or Choke Coils 125
Glossary 126
Works Cited 126
Further Reading 126

INTRODUCTION

The invention of the transformer in the late 1870s made possible for the first time the transmission of electrical energy over large distances. Among the first applications of the transformer, one can mention the Paris World Exhibition, the All-Russian Industrial Exhibition, the lighting installation at Great Barrington, Massachusetts, and the transmission of polyphase power from Lauffen to Frankfurt, Germany. Various claims have been made for the invention itself. Among the pioneers who should be mentioned are P. N. Yablochkov, I. F. Usagin, M. O. Dolivo-Dobrovolsky, W. Stanley, N. Tesla, G. Ferraris, C. E. L. Brown, and W. Boveri, as previously reported by Kostenko and Piotrovsky (1974).

The *transformer* is a device for transferring electrical energy from an alternating circuit to another having different voltage or current levels. A schematic representation of a transformer is shown in Fig. 1. It consists essentially of two or more independent electric circuits coupled by a common magnetic circuit. The electric circuits are made up of windings of insulated magnet wires and the magnetic circuit of steel laminations. The winding connected to the supply side is called *primary* and is denoted by the subscript 1; those connected to the receiver circuits are called *secondaries* and are denoted by the subscript 2. The laminations are stacked to form the *core.*

When the primary of the transformer is connected to the supply mains but its secondaries are open, only a small current referred to as the *magnetizing current* flows through the winding. This current establishes an alternating *magnetic flux* Φ through the core. This flux, in turn, establishes a voltage, or *counter emf,* across the terminals of the primary. The counter emf equals in magnitude the voltage V_1 of the mains less a small voltage drop due to the impedance of the primary. The alternating flux also establishes a voltage V_2 across the terminals of each secondary that is equal to the primary voltage

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

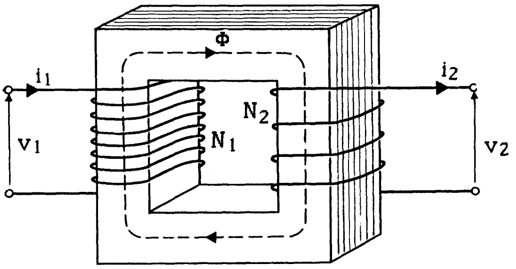

FIG. 1. Schematic representation of a transformer (adapted from Slemon and Straughen, 1980).

multiplied by the *ratio* of the number of *turns N* in the two windings.

When the terminals of a secondary circuit are connected to an impedance or *load,* a current *I* will flow that will tend to demagnetize the core. To compensate for this effect and maintain the flux at the level required by the voltage impressed on the primary, additional current is drawn from the supply mains. The ratio of the magnitudes of the primary and secondary currents is inversely proportional to the turns ratio in their windings.

This article deals with the transformers that constitute the main elements of an electric power supply system. At the source is an alternator, mostly a three-phase synchronous generator, whose rating varies usually from 30 to 1700 MVA at a voltage level up to 25 kV. Its terminals are connected directly to a *unit transformer* whose function is to reduce the current by stepping up the voltage, in order to minimize the Joule losses in the transmission lines. Transmission-line voltages are approaching today the megavolt range. At the receiving end of these lines there are substations in which the voltage level is reduced to subtransmission levels (69–230 kV) by what are properly called *power transformers.* The subtransmission circuits spread out from these substations to others located at the load centers, where other power transformers further reduce the voltage level (2.3–69 kV). Finally distribution lines reach industrial loads or residential areas, where *distribution transformers* (up to 200 kVA) reduce the voltage to consumer levels, in U.S. practice 120/240 V when single phase, or 120/208 V, or 240 V, or 480 V, or even higher when three phase. Three-phase power may be transformed either by a *bank* of *single-phase transformers,* or by a single *three-phase transformer.*

This article presents the elements of the construction, theory, and design of power transformers and, for completeness sake, also some special transformers, such as voltage and current regulators, instrument transformers, and reactors or choke coils.

1. CONSTRUCTION

1.1 Core

The core of a transformer is usually made of silicon sheet-steel laminations. For 60-Hz transformers, a 3% to 4.5% silicon steel sheet, 0.012 to 0.014 in. (0.3 to 0.35 mm) thick is used; for low frequency, 25 Hz, a 2.5% to 3.0% silicon steel sheet up to 0.019 in. (0.48 mm) thick is often used. Silicon steel is used because of its nonaging properties and low eddy-current and hysteresis losses. The magnetic steel is cold rolled and grain oriented, so that the rolling direction is the same as that of the flux lines. Much lower losses obtain with ferromagnetic amorphous metals, because of reduced eddy currents due to reduced conductivity. Since they come in the form of ribbons, usually the core is not made up as a stack of laminations, but as a roll of wound ribbon. There are three types of construction in common use: core, shell, and distributed-core type. In the core type each leg carries half of the primary and half of the secondary winding, as shown in Fig. 2 for a single-phase transformer and in Fig. 3 for a three-phase one. In the shell type the primary and secondary windings are located on an interior leg and are surrounded by exterior legs, as shown in Fig. 4 for a single-phase transformer and in Figs. 5(a) and 5(b) for a three-phase one. The cross section of the exterior legs and yokes is about half of that of the interior ones, because they carry about half of the flux. In the distributed-core type the core is built of a few parts, as shown in Fig. 6 for four parts.

1.2 Windings

The windings of transformers are built up of turns connected in series or in parallel. The material of the conductors is either cop-

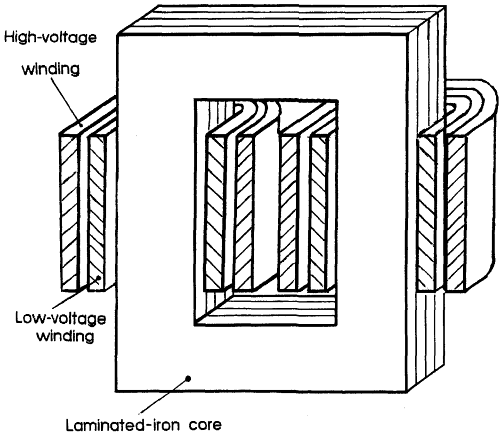
High-voltage winding
Low-voltage winding
Laminated-iron core

FIG. 2. Core-type, single-phase transformer.

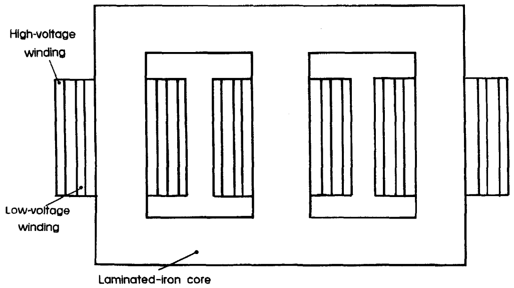
High-voltage winding
Low-voltage winding
Laminated-iron core

FIG. 3. Core-type, three-phase transformer.

per or aluminum. They must be insulated with materials having high dielectric and mechanical strength and not soluble in hot transformer oil. They are usually insulated with enamel, cotton, or treated paper held in place by bands of cotton thread. To avoid large eddy-current losses, large conductors must be built up of several small, separately insulated conductors connected in parallel. When the parallel conductors are wound with a uniform orientation to one another, a

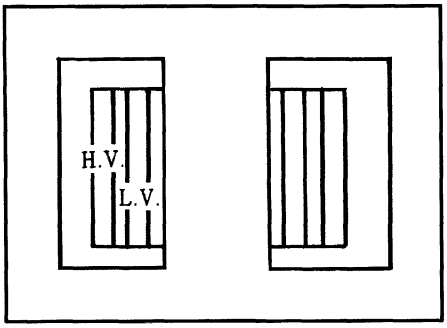

FIG. 4. Shell-type, single-phase transformer.

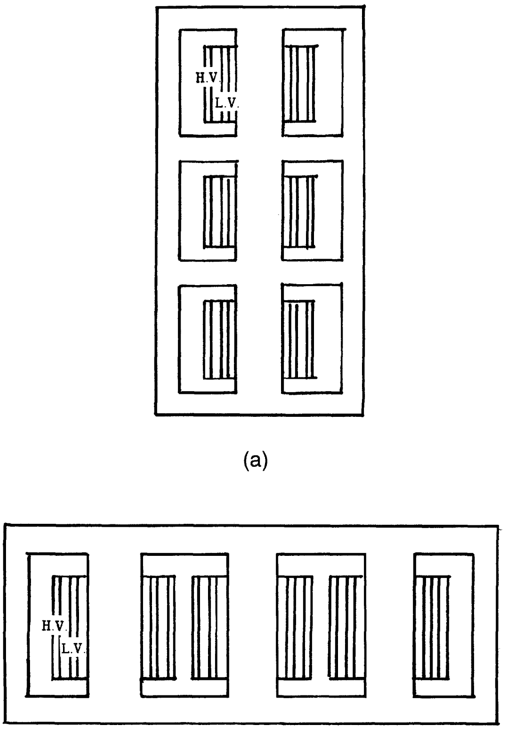

(a)

(b)

FIG. 5. Shell-type, three-phase transformers: (a) single-cored and (b) three-cored.

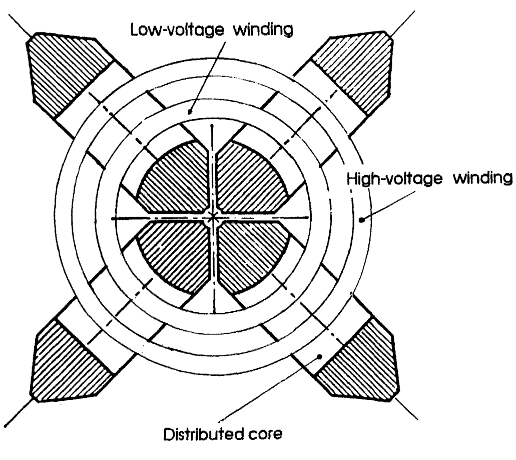
Low-voltage winding
High-voltage winding
Distributed core

FIG. 6. Distributed core for a transformer.

circulating current is induced, because the conductors do not link the same amount of magnetic flux. This circulating current can be eliminated by transposing the conductors with respect to one another. Coils with high layer voltage require heavy insulation between layers. Paper and empire cloth are generally used for layer insulation. The insulation between core and windings generally consists of molded cylindrical or rectangular tubes. The windings of transformers may be arranged concentrically with reference to one another with the low-voltage winding on the inside, sometimes referred to as a "barrel" winding, or they may be arranged in groups of high- and low-voltage coil disks stacked alternately one above the other, sometimes referred to as "pancake" winding.

1.3 Tank

Transformers are cooled by one of the following methods:

1. *Natural-air-cooled,* in which the natural circulation of the surrounding air is relied upon to carry away the heat generated by the losses. A sheet metal enclosure protects the windings from mechanical injury.
2. *Natural-oil-cooled,* in which the transformer is immersed in transformer oil that carries the heat to the walls of the containing tank and sometimes to a heat exchanger, by natural convection.
3. *Oil-and-water-cooled,* in which the transformer is immersed in transformer oil and cooling water is circulated through coils of pipe placed near the top of the tank, but under the surface of the oil.
4. *Cooled by forced circulation of the oil,* in which the oil is drawn from the tank, passed through a heat exchanger on the outside of the tank, and returned to the bottom of the tank.
5. *Air-blast-cooled,* in which a continuous stream of cool air is forced through the core and windings.

The welded steel tank is used almost exclusively. In the larger-capacity transformers, the radiating surface is increased by welding cooling vanes to the tank, or by using corrugated-sheet steel. Terminals of transformers connected to overhead lines or open bus work usually consist of porcelain bushings. Oil-filled terminal chambers for cable potheads are frequently used for direct connection to underground cables.

1.4 Insulating Oils and Liquids

The function of insulating liquids is to provide electrical insulation and heat transfer, and the viscosity selected is often a compromise between the two. Other considerations are flammability, solidification at low temperatures, and pressure development at high temperatures in sealed systems. Low cost, high dielectric strength, and ability to recover after dielectric overstress make mineral oil refined from crude the most used transformer insulating material. Chlorinated aromatic-Askarel liquids have somewhat higher dielectric strength, but have limited ability to recover after dielectric overstress. Some transformer liquids that in the past have been widely used are not environmentally sound, because they develop poisonous polychlorinated biphenyls (PCB's), and are being phased out.

2. ELEMENTARY THEORY

The transformer is a straightforward application of the laws of the electromagnetic field. Faraday's law in integral form expresses the electromotive force as

$$\text{emf} = \int_l \mathbf{E} \cdot d\mathbf{l} = -\frac{d\phi}{dt}, \tag{1}$$

where $\mathbf{E}$ is the electric field intensity, $d\mathbf{l}$ is an element of length, t is time, and Φ is the magnetic flux, which, assuming a uniform density B over the cross section A of the core, is

$$\phi = BA. \tag{2}$$

On the assumption that the winding has N turns and that the voltage impressed on its terminals is sinusoidal with radian frequency ω and frequency f, the rms value of the voltage is

$$V = (1/\sqrt{2})\omega N \Phi_m = 4.44 f N B_m A, \tag{3}$$

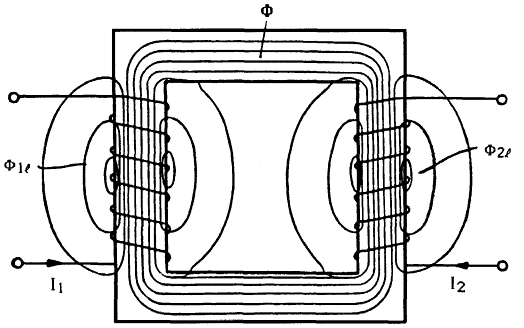

FIG. 7. Schematic representation of leakage flux.

where the subscript m stands for the peak value.

Ampère's law in integral form expresses the magnetomotive force as

$$\text{mmf} = \int_l \mathbf{H}\cdot d\mathbf{l} = \int_S \mathbf{J}\cdot d\mathbf{s} = \sum I, \tag{4}$$

where **H** is the magnetic field intensity, **J** is the current density, and ΣI is the total current enclosed in the magnetic circuit.

Figure 7 represents the flux paths in a single-phase transformer. Under ideal conditions all the turns of the windings completely encircle a magnetic flux Φ traveling in a magnetic circuit of uniform cross section and infinite permeability. It follows from Eq. (3) that the ratio of the voltages in the primary (1) and in the secondary (2) windings equals the turns ratio, or $V_1/V_2 = N_1/N_2$ and that H and, therefore, the mmf are zero, so that, according to Eq. (4), $I_1/I_2 = -N_2/N_1$. In practice, however, the permeability of the iron is not infinite, and part of the flux closes outside the iron, as indicated schematically by Φ_{1l} and Φ_{2l} in Fig. 7. The part of the flux that is not common to both windings can be represented by a leakage inductance. In terms of lumped parameters, R for winding resistance, L for leakage inductance, and M for mutual inductance, the governing equations can be written as

$$\begin{aligned} v_1 &= R_1 i_1 + \frac{d}{dt}[(L + M)i_1] + \frac{d}{dt}(Mi_2), \\ v_2 &= R_2 i_2 + \frac{d}{dt}[(L + M)i_2] + \frac{d}{dt}(Mi_1), \end{aligned} \tag{5}$$

where the lower case denotes instantaneous value.

These equations can be represented by an equivalent circuit when the secondary quantities are referred to the primary, so that $v_2' = (N_1/N_2)v_1$, $i_2' = (N_2/N_1)i_2$, $R_2' = (N_1/N_2)^2R_2$, and $L_2' = (N_1/N_2)^2L_2$. The equivalent circuit per phase is shown in Fig. 8. A resistance R_m has been added in parallel with M to account, in an approximate way, for the core losses. The magnetizing current i_μ is practically equal to the no-load current i_0. When the transformer has more than one secondary, these can be represented by attaching additional branches to the upper node. Both Eq. (5) and the equivalent circuit apply, in general, to both transients and steady state. Taking into account the sinusoidal variation with time of v and i in steady state and introducing phasor notation, one arrives at the diagram of Fig. 9. One advantage of this diagram is that it puts in evidence the phase relations among the variables. Also it puts in evidence the variation of the secondary voltage with the magnitude and phase of the

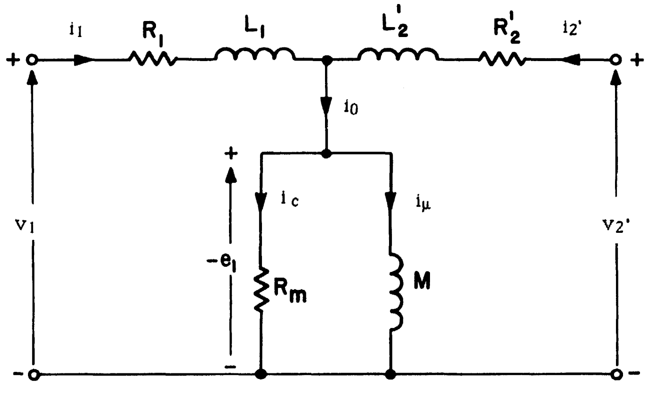

FIG. 8. Equivalent circuit per phase.

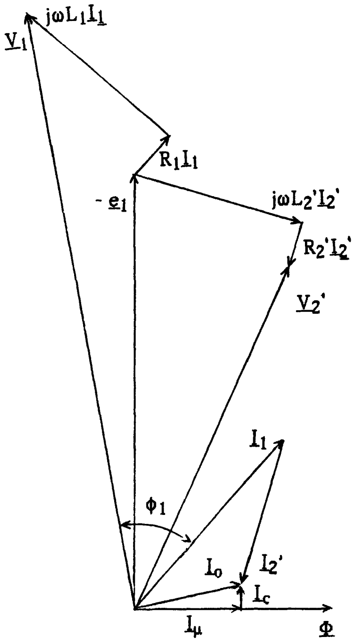

FIG. 9. Phasor diagram.

load. The locus of the secondary voltage when only the phase of the load changes and the small magnetizing current is neglected is known as the Kapp diagram and is shown in Fig. 10. It appears that a possibly dangerous overvoltage, known as the "Ferranti rise," can occur when the load is capacitive. This may happen at no load when the transformer feeds a long transmission line, especially if this is underground. In contrast, the secondary voltage is reduced when the load is inductive. A particular case of inductive load is short circuit, because when V_2' is reduced to zero, the current is limited by the impedance of the windings, which is predominantly reactive. For this reason, in the design stage, the leakage reactances are made larger the larger the VA capacity of the transformer, and may reach over 20% of the rated load impedance in the MVA range. In general the variation of the secondary voltage with load is expressed by the percentage *regulation*

$$\epsilon = 100(V_{2\ \text{no-load}} - V_{2\ \text{rated}})/V_{2\ \text{rated}}, \qquad (6)$$

which may reach 10% in large-size transformers. High leakage reactances, while reducing the magnitude of the short-circuit currents, increase the regulation at reactive power factors.

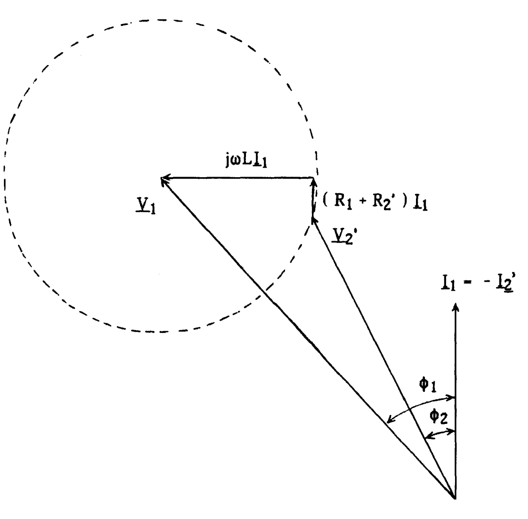

FIG. 10. Kapp's diagram.

3. DESIGN GUIDELINES

The starting point for the design is a set of performance specifications, which usually conform with standards issued by organizations such as the Institute of Electrical and Electronic Engineers (IEEE), American National Standards Institute (ANSI), and National Electrical Manufacturers Association (NEMA) in the U.S., VDE in Germany, and the International Electrotechnical Committee. Transformers may be designed for minimum first-investment cost without regard to the losses, or for minimum total losses and, thus, minimum running expenses. There is a lower limit to the cost, because heating limits the current density in the conductors, and the flux density in the iron core is limited by magnetic saturation. The lowest-cost transformer is generally the one in which the costs of iron and copper are approximately equal. With given current density in the conductors and magnetic flux density in the core, the transformer will have the maximum efficiency at full load when the losses in the copper are equally partitioned between the primary and the secondary and the loss in the iron equals the total loss in the copper. However, if the load is light during the greater part of the day, as in the case of lighting, the core loss should be as small as possible. The losses of a conductor in watts per cubic meter are given by

$$p_{\mathrm{co}} = J^2/\gamma, \tag{7}$$

where J, the current density, ranges from 1.5×10^6 A/m^2, in transformers up to about 50 kVA, to 4×10^6 A/m^2 in large oil-cooled transformers, and γ, the conductivity at 75 °C, is 4.7×10^7 S/m for copper and 2.67×10^7 S/m for aluminum. The losses in the iron core in watts per cubic meter can be expressed as

$$p_{\mathrm{Fe}} = afB_m^2 + b(\Delta f B_m)^2, \tag{8}$$

where B_m is the peak value of B, f is the frequency, Δ is the thickness of the laminations in millimeters, a is about 114, and b is about 3.8. Here the first term accounts for the hysteresis loss, and the second, for the eddy-current loss.

The mechanical construction of a transformer is determined in large measure by the forces of repulsion between the primary and the secondary windings that occur during short circuit. Taking into account the maximum value of the instantaneous dc component, the initial short-circuit current is about $2.8V_1/\omega(L_1 + L_2')$, and the force in newtons exerted on a conductor of length l and carrying the current I is IBl.

The number of turns in a winding is given, according to Faraday's law Eq. (3), by

$$N = V/4.44fB_mA, \tag{9}$$

where B_m can be selected near the knee of the magnetization curve of Fig. 11. The cross section of the core A can be determined by the following considerations. The apparent power of a single-phase transformer is

$$VA = V_1I_1 = (4.44fN_1B_mA)(J_1s_{1\,\mathrm{co}}), \tag{10}$$

where $s_{1\,\mathrm{co}}$ is the cross section of the primary conductor. If we assume that the volumes of the primary and secondary windings are equal, their total volume is

$$\mathcal{V}_{\mathrm{co}} = 2N_1s_{1\,\mathrm{co}}l_{\mathrm{co}} \tag{11}$$

where l_{co} is the average length of one turn in both windings. Assuming that the core and yoke have the same cross section, and denoting by l_{Fe} the total length of the magnetic circuit, the volume of the iron is

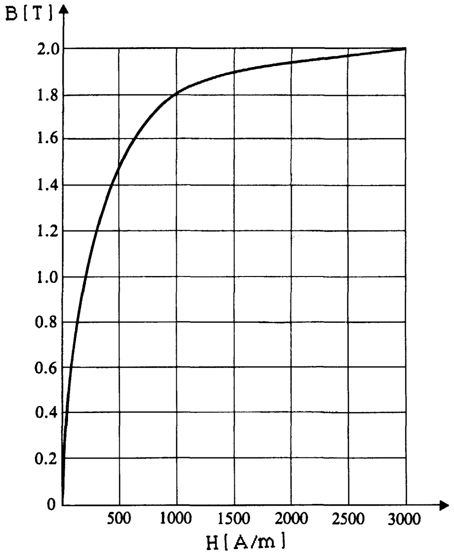

FIG. 11. Magnetization curve of grain-oriented steel.

$$\mathcal{V}_{\mathrm{Fe}} = Al_{\mathrm{Fe}}. \tag{12}$$

Multiplying Eq. (10) by Eq. (12), one obtains

$$VA\mathcal{V}_{\mathrm{Fe}} = 4.44fN_1B_mA^2J_1s_{1\,\mathrm{co}}l_{\mathrm{Fe}}. \tag{13}$$

Making use of Eq. (12) and introducing the total losses P_{co} and P_{Fe}, one obtains for the cross section of the core

$$A = \sqrt{\frac{VA}{4.44}}\sqrt{\frac{P_{\mathrm{Fe}}}{P_{\mathrm{co}}}}\sqrt{\frac{p_{\mathrm{co}}}{p_{\mathrm{Fe}}}}\sqrt{\frac{2}{fB_mJ}}\sqrt{\frac{l_{\mathrm{co}}}{l_{\mathrm{Fe}}}}. \tag{14}$$

This equation is not limited to single-phase transformers but also holds for polyphase ones. The ratio of the lengths $l_{\mathrm{co}}/l_{\mathrm{Fe}}$ is approximately 0.57 for single-phase transformers and 0.38 for three-phase ones.

The height of the core and, therefore, of the window is mainly a function of the type of cooling, because it determines the surface for the transfer to the cooling medium of the heat generated by the core and conductor losses. It is expressed in terms of the linear current density K as

$$h = (N_1I_1 + N_2I_2)/K, \tag{15}$$

where K is approximately equal to 90 $(\mathrm{kVA})^{0.15}$ in air-cooled transformers and 520 $(\mathrm{MVA})^{0.24}$ in oil-cooled ones, kVA or MVA being the rating of the transformer.

The last main dimension of the transformer, the width of the window, is determined by the cross sections of the conductors and of the insulation, whose thickness increases with the voltage, and can be expressed in terms of a filling coefficient f_{co}, which represents the ratio of the cross section of the conductors to the total cross section of the window. The filling coefficient is approximately equal to 0.13 $(\mathrm{kVA})^{0.22}$ for air-cooled transformers and 0.12 $(\mathrm{MVA})^{0.13}$ for oil-cooled ones.

Once the main dimensions of the transformer have been established, one can proceed to determine the parameters of the equivalent circuit. The resistance of a winding is

$$R = Nl_{\mathrm{co}}/s_{\mathrm{co}}\gamma. \tag{16}$$

Physically it is not possible to distinguish between the primary and secondary leakage inductances, and so in the equivalent circuit of Fig. 8, half of the total inductance L is assigned to L_1 and half to L_2'. If we denote by a_1 and a_2 the thicknesses of the primary and secondary and by b the separation between them, the inductance can be expressed approximately as

$$L = 4\pi \times 10^{-7} \frac{fl_{\mathrm{co}}N_1^2}{h}\left(b + \frac{a_1 + a_2}{6}\right) \tag{17}$$

for a concentric winding and, if c denotes the number of disks in the primary winding,

$$L = 4\pi \times 10^{-7} \frac{fl_{\mathrm{co}}N_1^2}{ch}\left(b + \frac{a_1 + a_2}{6}\right) \tag{18}$$

in a stacked-disk winding.

According to Ampère's law, Eq. (4), the magnetizing current is given by

$$I_\mu = \frac{\int_l \mathbf{H} \cdot d\mathbf{l}}{\sqrt{2}N_1} = \frac{(\mathrm{AT})}{\sqrt{2}N_1}, \tag{19}$$

where the numerator, also called the total ampere turns (AT), can be determined with the help of Fig. 11 with B determined by the continuity of the magnetic flux. An approximate value of the mutual or magnetizing inductance M can then be derived with the help of the equivalent circuit of Fig. 8 as

$$M = \frac{V_1}{\omega I_\mu} \approx \frac{\sqrt{2}N_1^2}{(\mathrm{AT})} BA. \tag{20}$$

An approximate value of the magnetizing resistance R_m can be obtained in terms of the total iron losses P_{Fe} as

$$R_m = (mV_1^2)/P_{\mathrm{Fe}}, \tag{21}$$

where m is the number of phases.

4. THERMAL CHARACTERISTICS AND COOLING

If all the dimensions of a transformer are increased in the same ratio, all surfaces, including the cooling surfaces, and all the cross sections are increased as the square of the ratio. For given densities, the flux and the mmf increase as the square of the ratio, and, therefore, the apparent power increases as the fourth power. Since all losses increase with the volume and, therefore, only with the cube of the ratio, it follows that the efficiency increases with the power. However, since the losses increase faster than the cooling surface, more effective cooling methods must be used. Natural air cooling is adequate only for sizes up to a few kVA. For larger sizes oil cooling is needed.

The temperature rise affects the lifespan of the insulation in two ways: It accelerates chemical reactions, such as oxidation, and it subjects the insulation to mechanical stresses due to the differential elongation between adjacent materials. The first process is the dominant cause of failure in organic materials. The second process is the dominant cause of failure in insulation known as class H, which includes materials characterized by poor abrasion properties, such as polyimides. The lifetime of the insulation, which is commonly expressed in the form of Arrhenius relations, decays exponentially with the temperature rise and, in the range of practical interest, is reduced by about one-half with every increase in temperature of about 10 K. This consideration has resulted in peak allowable temperatures of 110–120 °C.

The thermal characteristics of transformers involve

1. dissipation of the heat resulting from the iron and the copper losses,
2. variations in the kVA rating with temperature rise in steady load conditions,
3. temperature rise under transient conditions (heating up and cooling down),
4. effects of certain factors such as altitude, color of tank, and sunshine under service conditions, and
5. the effect of the temperature on the lifetime.

Three modes of heat transfer are involved in the dissipation of heat in a transformer: conduction, convection, and radiation.

4.1 Conduction

Thermal transfer through solids is governed by Fourier's law according to which the temperature rise in one dimension is

$$\theta = ql/\lambda, \tag{22}$$

where q is the heat flow per unit area, l is the length of the path, and λ, the thermal conductivity of the medium, is equal to 345 W/m·K in copper and 207 W/m·K in aluminum.

4.2 Convection

Convection is the thermal transport from a solid surface by a fluid in motion and can be related to the heat transfer coefficient h as

$$\theta = q/h. \tag{23}$$

The flow of the coolant may be laminar or turbulent, depending on the speed of the fluid. Turbulent flow is much more effective for cooling purposes than laminar. The heat transfer coefficient ranges from 40 $\mathrm{W/m^2 \cdot K}$ with air cooling to 200 $\mathrm{W/m^2 \cdot K}$ with oil cooling.

4.3 Radiation

Radiation is the process by which heat is transferred by means of electromagnetic waves according to the laws of optics. Since wave propagation can occur in vacuum, it is most important for cooling in outer space applications. It is less important in terrestrial applications, in which convection is dominant. The power radiated per unit surface is proportional to the difference between the fourth powers of the absolute temperatures of the hot body and of the ambient, according to the Stefan–Boltzmann law. In transformers the temperature rise is small and so is the radiated power. It is, therefore, possible to make a linear approximation and lump the radiated power together with the heat transferred by convection. This is done by augmenting h in Eq. (23) by a factor 1.5.

5. POLYPHASE CONNECTIONS

5.1 Three-Phase Transformations

The delta–delta (Δ-Δ), star (or wye)-delta (Y-Δ), delta-star (or wye), (Δ-Y), and star–star or wye–wye (Y-Y) connections are the most common; they are illustrated in Fig. 12. The first two are used in step-up transformers for moderate voltages. The star-delta connection has the advantage of providing a grounding point in the low-voltage system that does not shift with unbalanced load and has the additional advantage of being free of third-harmonic currents and voltages (see Sec. 6). The delta–delta connection has the advantage of permitting operation in "V" in case of fault in one of the phases. The delta-star step-up and the star-delta step-down connections are most commonly used for high-voltage transmission, because they provide a point at which the system can be grounded either directly or through an impedance to damp oscillations and to limit line-to-ground fault currents. The star–star connection must be provided with a delta tertiary, in order to provide a path for the third-harmonic component of current.

5.2 Two- to Three-Phase Transformations

The one most commonly used is the Scott connection shown in Fig. 13. It uses two single-phase transformers that are electrically but not magnetically connected. Across the lines B and C is the so-called *main* transformer that has a 50% tap. The *auxiliary* or *teaser* transformer has an 86.6% tap and is

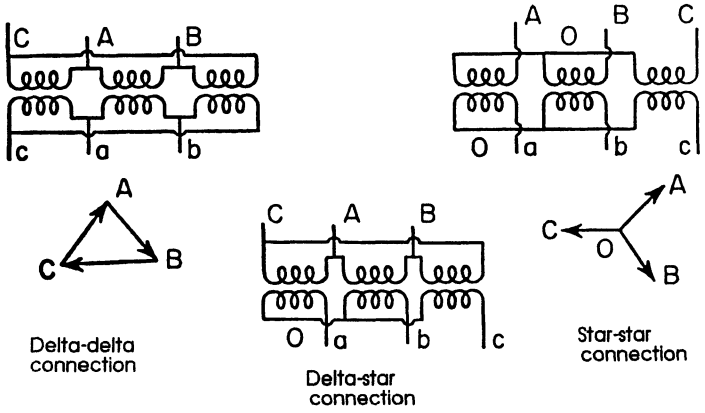

FIG. 12. Common three-phase transformer connections.

connected between the other line terminal, *A* in this case, and the center tap on the main transformer. If the two transformers have the same number of turns, the voltage per turn and, therefore, the flux is the same. A neutral point could also be provided on the three-phase side, and from the geometry of the triangle, this tap would have to be located at 33% from *D*. The windings on the three-phase side of the transformers must be designed to carry 115.5% of the current that the single-phase ratings of the transformers indicate. Alternatively, if the additional extra copper is not provided, the rating of the bank should be reduced to 86.6% of the combined single-phase ratings. This can be deduced from Fig. 13. Assuming for simplicity that $N_1 = N_2 = N$ and neglecting the magnetizing current, one obtains for the teaser transformer

$$0.866N\underline{I}_A + N\underline{I}_a = 0 \quad \text{or}\ \underline{I}_A = 1.15\underline{I}_a, \tag{24}$$

where the underline denotes a phasor, the capital subscript denotes the three-phase side, and the lower-case subscript the two-phase side. For the main transformer one has

$$\underline{I}_B N/2 - \underline{I}_C N/2 - \underline{I}_b N = 0 \quad \text{or}$$
$$\underline{I}_b = (\underline{I}_B - \underline{I}_C)/2, \tag{25}$$

and letting $\underline{I}_C = -\underline{I}_A - \underline{I}_B$, one obtains

$$\underline{I}_B = -\underline{I}_A/2 + \underline{I}_b \ \text{and similarly}$$
$$\underline{I}_C = -\underline{I}_A/2 - \underline{I}_b \tag{26}$$

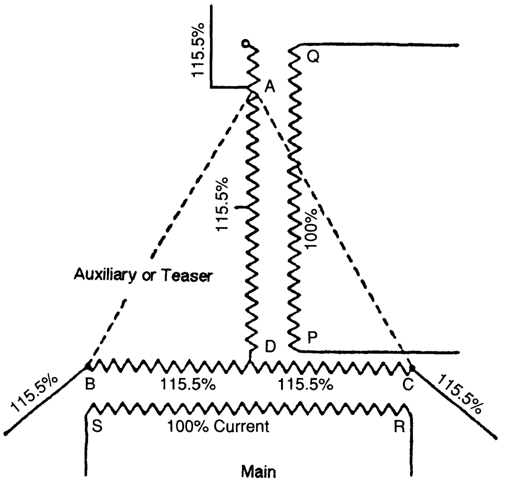

FIG. 13. Scott connection (adapted from Knowlton, 1941).

5.3 Zig-zag Connection

It is possible to obtain an arbitrary desired phase shift with a polyphase transformer if the primary and/or the secondary are divided into a number of sections. The zig-zag connection may suppress third harmonics in the line-to-neutral and line-to-line voltages. As shown in Fig. 14, half sections are connected in opposition, thus giving a larger fundamental voltage, but causing the co-phasal third harmonics to cancel out. The fundamental line-to-neutral voltage is only 0.86 times the arithmetic sum of the two section voltages, and this means that additional turns and copper are required. However, a better distribution of the load is obtained on the primary side when the secondary load is unbalanced.

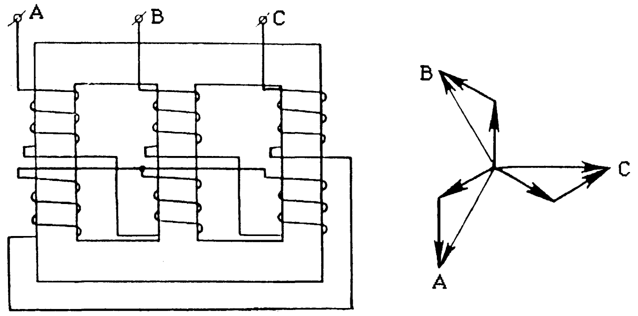

FIG. 14. Zig-zag connection.

5.4 Three- to Six-Phase Transformations

Twelve-pulse rectifiers must be supplied by six phases. These can be implemented either by connecting the center taps of the secondary windings or by having two secondary windings connected in double *Y* with a 180° phase difference between one *Y* and the other, or with one *Y*-*Y* connection and one *Δ*-*Y* connection.

5.5 Load Tap Changing

To permit some adjustment of the voltage or to maintain the secondary voltage against supply or load variations, tappings are brought out to terminals so that the number of turns on one winding, usually the high-voltage winding, can be changed by ±2.5% to 5%. In order to change the turns ratio without interfering with the load, the current is shunted from one tap to the next with the help of bridging resistances, or reactors. The underlying principle is illustrated in Fig. 15 with the current-carrying portions drawn in black. The procedure is to transfer the line current first from contact 1 to contacts 2 and 3 and resistances R_1 and R_2 and next to contact 4. Tap changing is implemented by a special motor. Alternatively, the mechanical contacts may be replaced by semiconductors. Typically, one uses two silicon-controlled rectifiers (SCRs) connected in antiparallel. For voltages above about 6 kV, a series connection of SCRs is required. The advantages are as follows: no pitting of contacts, switchover in a few microseconds exactly at the zero crossing of the current, no voltage drop during the switchover, and unlimited number of successive switchings that with resistances is limited by heating. Tap changing with SCRs is used especially in electric locomotives and is appropriate also for furnace or rectifier transformers.

5.6 Autotransformer

This uses a single winding with part of the winding common to both the primary and secondary. When the ratio of transformation approaches unity and most of the

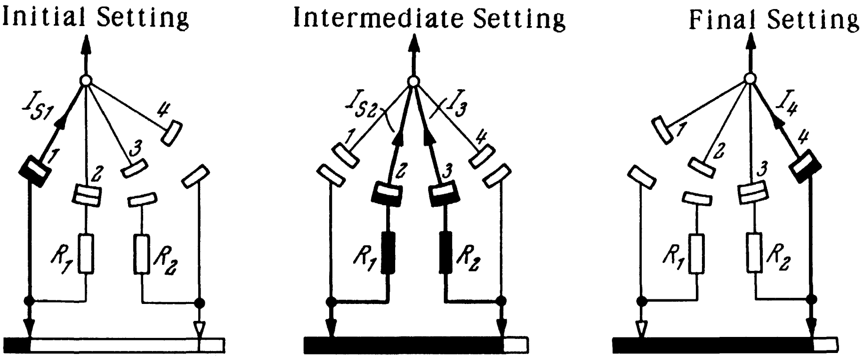

FIG. 15. Load tap changer (adapted from Boedefeld-Sequenz, 1971).

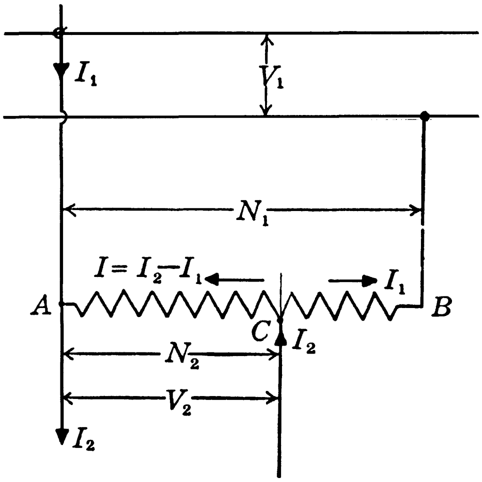

FIG. 16. Schematic diagram of autotransformer.

winding is common, the autotransformer offers great advantages with regard to regulation, reduced weight and size per kVA, lower cost, higher efficiency, and lower magnetizing current. However, the autotransformer has the disadvantage that it does not provide electrical isolation between the primary and the secondary circuits. Figure 16 shows a schematic diagram. As in a two-winding transformer, the core flux is determined by the impressed voltage V_1, the number of primary turns N_1, and the line frequency f. The common part of the winding carries a current that is the difference between I_1 and I_2. One of the most common applications of the autotransformer is the "Variac," "Regavolt," or "Regulac." These have a toroidal core and winding, with a thin carbon brush making contact with the winding in any desired position, as shown in Fig. 17. The output voltage can be varied between zero and about 120% of the input voltage.

5.7 Parallel Operation

Two or more transformers are often operated in parallel to deliver power to a load over a common busbar. In order to avoid local flow of current circulating in the paralleled windings, the following conditions must be satisfied in single-phase transformers:

1. proper connection with regard to polarity,
2. identical voltage levels,
3. equal turn ratios,
4. ratio of the winding impedances inversely proportional to the current ratio, and
5. ratio of the winding resistances equal to the ratio of the reactances.

In addition, in three-phase transformers, the phase angles must be identical.

6. HARMONICS AND NOISE

Harmonic currents are sources of losses, voltage distortion, electromagnetic and acoustic noise, and telephone interference. They are becoming a major concern of the utilities due to the ever-increasing diffusion of nonlinear loads containing solid-state switching elements. One must distinguish between sources that are intrinsic in the transformer itself and sources in the load. Starting with the first, the magnetization curve of Fig. 11 is nonlinear on account of iron saturation. Therefore, if it is desired to keep B, and therefore Φ and V, sinusoidal, the magnetizing current must contain harmonic components of which the third and its multiples, the triplen, are the dominant ones. Then, as was already mentioned in the previous section, the transformer must be provided with at least one delta winding in which these current components can flow freely. The associated losses, however, cause

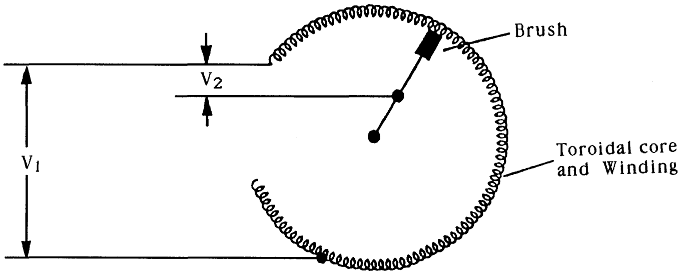

FIG. 17. Schematic diagram of "Variac" (adapted from Hindmarsh, 1965).

additional heating, which must be taken into account in the design stage.

The harmonics that originate in nonlinear loads, such as fluorescent lighting, computers, and adjustable-speed drives, must be kept under control, if needed, by imposing penalties on the consumers when they do not comply with standards, such as IEC 1000 or IEEE 51a. Limits on the total harmonic distortion (THD), which is the ratio of the root mean square value of the harmonics to the amplitude of the fundamental component, and on the telephone influence factor (TIF), which is given by a weighting curve of the relative interfering effects, are set by standards, such as those of the ANSI-IEEE and of the European Community. When harmonics are present, the associated losses are unavoidable and must be taken into account in the rating of the transformer. In transformers that feed rectifiers the current may contain a dc component that may drive the iron core into saturation.

The transformer also has an intrinsic source of noise that is due to magnetostriction of the iron laminations. This consists of an elongation in the direction of the flux. The noise, whose dominant frequency is twice that of the supply, can be attenuated by avoiding mechanical resonances or by attaching dampers to the external walls of the tank.

Another source of noise is the auxiliary cooling equipment, such as fans, blowers, and pumps.

7. SPECIAL TRANSFORMERS

7.1 Voltage Regulators

These are used when the power transformer is not adequately provided with taps. In this case the taps and switches are placed in the winding of a separate autotransformer.

7.2 Constant-Current Transformers

These are used mainly for street lighting when the lamps are connected in series, and they maintain the current within ±1/2%, regardless the number of lamps connected in series. In these transformers the primary and secondary windings are stacked one on top of the other with the primary stationary and the secondary free to move vertically. Counterweights are used to effect a balance between the weight of the secondary and the force of repulsion between the two windings. Any increase in current increases the repulsion force, thereby increasing the leakage reactance and, thus, decreasing the secondary voltage to the value required to produce the rated current.

7.3 Instrument Transformers

These are divided into two classes: potential (PT) and current (CT) transformers. Their principle of operation is essentially the same as that of the power transformer. They are built with high-quality iron operating at very low flux densities so that I_0 is very small. The PT is designed with an impedance as small as possible, in order to keep the voltage ratio practically constant. The CT often has only one turn formed by passing the line conductor through the window of the iron core. The CT's secondary must never be left open-circuited, because, without the secondary ampere turns, the primary current would become entirely a magnetizing current I_μ, which would saturate the core and induce dangerous overvoltages in the secondary.

Each instrument transformer serves two purposes:

1. to insulate the high-voltage circuit from the measuring one, in order to protect the measuring apparatus and the operator, and
2. to make possible the measurement of high voltages with low-voltage voltmeters (usually 115 V), or large currents with low-current ammeters (usually 5 A).

These transformers are also used to operate relays, solenoids, and circuit breakers, or other controlling devices.

7.4 Reactors or Choke Coils

Their main applications are for limiting the fault current, when connected in series with feeder lines or with neutral ground connections, and for supplying the charging current, when connected in shunt with underground cables. They store energy in magnetic form, and their VA rating is, therefore.

$$N = 2\pi f L I^2. \qquad (27)$$

Their structure may be air-cored or iron-cored. However, even in the latter case the energy is stored outside the iron core in so-called air gaps distributed inside the coil. The function of the steel is to concentrate the energy storage in small volumes of high flux densities and thus reduce the coil size and the electrical losses. The coils must be designed to withstand the strong radial stresses that develop during short circuits. The inductance of iron-cored reactors can be controlled by introducing a dc current component and thus driving the iron into saturation during an adjustable part of the half cycles.

GLOSSARY

Coil: A series of turns of electric wire.

Conduction: Transmission of heat through a medium.

Convection: The circulating motion that occurs in a fluid at a nonuniform temperature due to the variation of its density and the action of gravity.

Core: The part of the magnetic circuit that carries most of the flux, usually made of a stack of ferromagnetic laminations.

Harmonic: A component frequency that is an integral multiple of the fundamental frequency.

Noise: Acoustic or electromagnetic radiation that involves random changes in frequency or amplitude.

Primary: The electric circuit connected to the power supply.

Radiation: Emission of energy in the form of waves.

Reactor: A device used to introduce reactance into an electrical circuit.

Secondary: The electric circuit connected to the user.

Tank: The metallic envelope of a transformer.

Transformer: A device employing mutual inductance to convert variations of current in a primary electric circuit into variations of voltage and current in a secondary circuit.

Turn: A single round of wire.

Winding: A series of coils wound around the core.

Works Cited

Boedefeld, Th., Sequenz, H. (1971), *Elektrische Maschinen,* Wien: Springer-Verlag.

Hindmarsh, J. (1965), *Electrical Machines and Their Applications,* Oxford, U.K.: Pergamon Press.

Knowlton, A. E. (Ed.) (1941), *Standard Handbook for Electrical Engineers,* New York: McGraw-Hill, p. 575.

Kostenko, M., Piotrovsky, L. (1974), *Electrical Machines,* Moscow: Foreign Languages Publishing House, vol. 1.

Slemon, G. R., Straughen, A. (1980). *Electric Machines,* Reading, MA: Addison-Wesley.

Further Reading

ANSI-IEEE (1994), *Standard C57* (1-55937-271-0), New York: IEEE Press.

Arnold-La-Cour, E, (1927), *Die Wechselstromtechnik,* Berlin: Springer-Verlag.

Boedefeld, Th., Sequenz, H. (1971), *Elektrische Maschinen,* Wien: Springer-Verlag.

Del Toro, V. (1985), *Electric Machines and Power Systems,* Englewood Cliffs, NJ: Prentice-Hall.

Hindmarsh, J. (1965), *Electrical Machines and Their Applications,* Oxford, U.K.: Pergamon Press.

Judd, F. F., Kressler, D. R. (1977), "Design Optimization of Small Low-Frequency Power Transformers," *IEEE Trans. Magn.* **MAG13,** 1058–1069.

Kuhlmann, J. H. (1950), *Design of Electrical Apparatus,* New York: Wiley.

Liwschitz, M. (1934), *Die Elektrische Maschinen,* Leipzig and Berlin: Teubner.

McPherson, G. (1981). *An Introduction to Electrical Machines and Transformers,* New York: Wiley.

Richter, R. (1950). *Elektrische Maschinen,* Berlin: Springer-Verlag.

Sankaran, C. (1992), *Introduction to Transformers,* New York: IEEE Press.

Slemon, G. R., Straughen, A. (1980), *Electric Machines,* Reading, MA: Addison-Wesley.

Stigant, S. A., Franklin, A. C. (1973), *The J & P Transformer Book,* London: Newness-Butterworths.

TRANSFORMS, MATHEMATICAL

See FOURIER AND OTHER MATHEMATICAL TRANSFORMS

TRANSISTORS, BIPOLAR

Richard C. Barker, *Departments of Electrical Engineering and Applied Physics and Yale Center for Microelectronic Materials and Structures, Yale University, New Haven, Connecticut, U.S.A.*

Introduction 127
1. **Semiconductor Basics** 128
1.1 Energy Bands, Densities of States 128
1.2 Fermi–Dirac Occupation Probability 128
1.3 Carrier Concentrations 129
2. ***pn*-Junction Fundamentals** 130
2.1 Junction Potential and Carrier Concentrations 130
2.2 Voltage Control of Minority-Carrier Boundary Values 131
2.3 The Continuity Equation and Carrier Transport 132
2.4 Diffusion Equation for Minority Carriers 133
2.5 The *pn* Junction as a Generation/Recombination Device 135
3. **Bipolar Transistor Physics** 136
3.1 Minority-Carrier Flow, and Current Gain 137
3.2 Common-Base and Common-Emitter Characteristics 140
3.3 The BJT as a Generation/Recombination Device 141
3.4 Other Related Structures 142
4. **Materials Science, Technology and Application** 143
4.1 Design Objectives 143
4.2 Epitaxial Silicon/Germanium BJT 144
4.3 Polysilicon Emitter 145
Works Cited 145
Further Reading 146

INTRODUCTION

The seeds of the electronics revolution were planted in the early part of the industrial revolution by Lee Deforest, Yale Class of 1899, when he invented the vacuum tube triode amplifier in 1906. He immediately applied it to radio, with which it was his ambition to "bring culture to millions." But radio found a major market and financial sponsor during World War I. At the beginning of World War II, television had been on display at the 1939 World Fair in New York, and again a major drive developed to create all sorts of military radio and other complex electronic systems, which by 1945 were being produced with "peanut tubes" the size of your thumb, packed together so closely that the boxes containing them were hot as a light bulb. Subsequent scaling down of electronic circuits soon ran into a power-density limit. And the power was being used only to drive electrons out of a metal cathode, already brilliantly designed to minimize the work function, and not to perform any of the electronic functions themselves. To get around this problem, studies were already under way to find ways of controlling electron flow inside materials without the use of heaters.

At the Bell Telephone Laboratories materials science had produced high-quality single germanium crystals, with growth-controlled group III or V inpurities. While looking for electric-field effects on the surface of a polished crystal, theorist John Bardeen and experimentalist Walter Brattain discovered in December 1947 that the current between two closely spaced points, an emitter and a collector, on the surface of a germanium base could be controlled by a current between the emitter and the base and, further, that the controlled current was larger than the current controlling it. The bipolar junction transistor (BJT) was born.

It had taken about 50 years to get the hot filament out of the electronic amplifying valve. Dissidents from Bell Laboratories then moved to Texas Instruments to initiate a silicon technology, leading to the second 50

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

years it has taken to complete the electronics revolution. The outcome has ushered in another revolution, the information revolution. What DeForest has taught us, and he himself stressed, is not how to build the vacuum-tube triode, but what one can accomplish with an electronic amplifying device. The message was not lost on David Sarnoff and others who have used it, some for selfish gain but all, unwittingly, to build a truly global community.

1. SEMICONDUCTOR BASICS

The semiconductor concepts and relationships vital to BJT device physics are summarized in Fig. 1. Quantities referenced therein are defined in the context of the following discussion. Descriptive terms are defined in the context of the next section. The physics leading up to these fundamentals can be found in SEMICONDUCTORS, ELEMENTAL—ELECTRONIC PROPERTIES; SEMICONDUCTORS, ELEMENTAL—MATERIAL PROPERTIES; Shockley (1939, 1950), and texts that deal with semiconductor physics (e.g., Kittel, 1996, or Sze, 1981). The essentials are as follows.

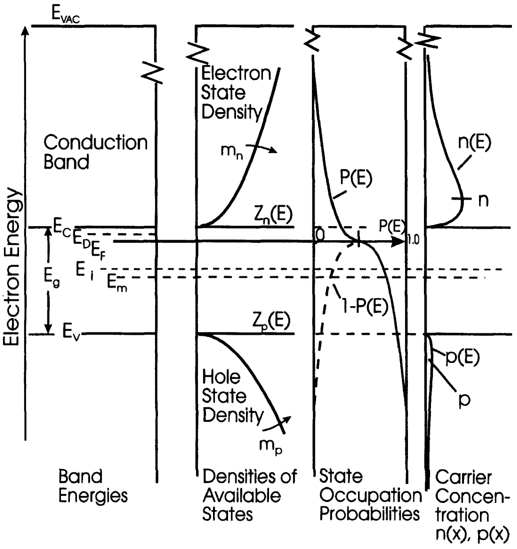

FIG. 1. Determination of electron concentration in terms of energy bands, densities of available states, and the Fermi–Dirac distribution function.

1.1 Energy Bands, Densities of States

A semiconductor energy-band diagram for a uniformly doped n-type material in thermal equilibrium is shown in Fig. 1. Unless otherwise stated, the material under discussion is crystalline silicon. Doping the (group IV) silicon up to about 0.1% with group V "donor" atoms, such as phosphorus (P) or arsenic (As), makes available up to one electron per atom that substitutes for silicon in its tetrahedral covalent bonding configuration, producing n-type silicon (n-Si). Doping with group III "acceptor" atoms makes available one hole per substitutional atom, producing p-type silicon (p-Si).

The four panels of Fig. 1, with energy levels aligned, show for n-Si

1. the valence band and its upper edge, E_V; the conduction band and its lower edge, E_C; the energy gap E_g between these levels; the midgap energy E_m; the electron "donor" energy level E_D; the "Fermi energy" E_F; and the "intrinsic" Fermi energy E_i;
2. a parabolic curve representing the quantum mechanically determined density of *available* electron states, $Z_n(E)$, in the (nearly empty) conduction band, and the corresponding density of available vacancy (hole) states, $Z_p(E)$ in the valence band;
3. the temperature-dependent Fermi–Dirac distribution function $P(E)$, centered on E_F, which gives the probability of any state of energy E in each band being occupied by an *electron*, and (dashed) $1 - P(E)$, the corresponding probability of any state of energy E being unoccupied by a *hole* in the valence band; and
4. the product of the available state densities in **2** and their occupation probabilities in **3**, which gives the energy distributions of state-occupying electrons in the conduction band, holes in the valence band, and their integrals n and p in units of numbers per cm^3.

1.2 Fermi–Dirac Occupation Probability

The Fermi–Dirac *electron* occupation probably for these states is given by the well-known Fermi–Dirac distribution function,

$$P(E) = \{\exp[(E - E_F)/kT] + 1\}^{-1}. \quad (1)$$

The product of $P(E)$ and the density-of-states function yields the densities of both holes and electrons with reasonable accuracy so long as the Fermi energy E_F remains within the energy gap. The Fermi energy is known as the "chemical potential" of the material, because a single electron added to the system in equilibrium resides (statistically) at the energy $E = E_F$. Joining two materials with two Fermi energies (measured from E_{vac}) produces a flow of electrons from the material with the higher Fermi energy to that with the lower. Removing a single electron from a material requires the energy $\phi = E_{vac} - E_F$, defined as the *work function.* The energy given up by an external electron at rest when dropping down to the conduction-band edge E_C is called the *electron affinity* χ, about 3.25 eV (electron volts) in silicon.

1.3 Carrier Concentrations

The energy gap E_g, an interband ionization energy, is about 1.1 eV. It takes much less energy to ionize a donor atom such as P or As, about 0.040 eV (40 meV), than to "ionize the valence band," i.e., to create an electron–hole (*e-h*) pair. Therefore, the donor energy level E_D shown in Fig. 1 lies 40 meV below the conduction-band edge. (Note that the figure is not drawn to scale.) The same can be said for ionizing group III atoms such as boron (B) and aluminum (Al), each of which "accepts" an electron from the valence band to complete the four silicon bonds required for it to "substitute" for silicon, thus creating a positively charged valence-band vacancy, or "hole," which is mobile in *p*-Si. Thus, an acceptor energy level E_A would be placed above the valence-band energy E_V by its ionization energy. A figure comparable to Fig. 1 for *p*-Si, and with the same acceptor ionization energy as that for the donor shown, can be constructed by simply rotating Fig. 1 about the midgap energy E_m.

In pristine silicon, with no dopants, the electron and hole densities are equal, $n = p$, since there are no other sources for either. The Fermi energy for this *intrinsic* semiconductor is written E_i and lies slightly above E_m in Fig. 1, because $Z_p(E)$ is larger than $Z_n(E)$ at every E, and when $E_F = E_i$, n must equal p.

The density of *ionized* donors is N_D and of *ionized* acceptors is N_A in the corresponding *n*-Si and *p*-Si. Thus in *n*-Si at equilibrium, with donor concentration N_D, $n_{n0} = N_D + p_{n0}$ with $p_{n0} \ll N_D$, and therefore $pn \approx p_{n0}N_D$. Correspondingly in *p*-Si, with acceptor concentration N_A, $p_{p0} = N_A + n_{p0}$ with $n_{p0} \ll N_A$ and $pn \approx n_{p0}N_A$. Thus, in equilibrium, taking $pn = n_i^2$, the minority-carrier concentrations are approximately

$$\begin{aligned} &\text{in } n\text{-Si}, \quad p_{n0} = n_i^2/N_D; \\ &\text{in } p\text{-Si}, \quad n_{p0} = n_i^2/N_A. \end{aligned} \tag{2}$$

For a sense of numbers, in silicon at room temperature, $n_i \approx 1.5 \times 10^{10}$ cm^{-3}, and $n_i^2 = 2.25 \times 10^{20}$ cm^{-3}. The atomic density is 5×10^{22} cm^{-3}. A normal doping level might be $N_D = 10^{18}$ cm^{-3}, or one dopant atom for 5×10^4 silicon atoms. On the assumption that all dopant atoms are substituted for silicon and are all ionized, the *majority-carrier* (electron) density, taking $n_i = 10^{10}$ cm^{-3} for simplicity, is

$$n_{n0} = N_D = 10^{18} \text{ electrons cm}^{-3},$$

and the *minority-carrier* (hole) density is

$$p_{n0} = 10^{20}/10^{18}, = 100 \text{ holes cm}^{-3}!$$

The ratio is $n_{n0}/p_{n0} = 10^{16}$, yet it is the diffusion of minority carriers that makes things happen in the bipolar junction transistor, which is also known as the "diffusion transistor."

In thermal equilibrium, *e-h* pairs are generated at a thermal generation rate G_{th}, usually involving defect and impurity states near midgap, without which G_{th} would be much smaller. This rate is matched by a recombination rate rnp, where r is a recombination rate constant. The rate is dependent on the simple product pn because the recombination process is assumed to be statistically independent. Given G_{th} and r, therefore, the product np must be independent of any specific value of n or p, established by local doping concentrations. One such situation is the intrinsic semiconductor where $n = p$ defined as n_i. Thus at fixed temperature and constitution of material $pn = n_i^2$ for any p and n.

Regardless of the shape of the density-of-available-states functions, the exponential character of the Fermi–Dirac function makes the integrated carrier concentrations n and p exponentially dependent on the displacement of E_F from E_C and E_V, respectively. This accounts for the analogy between the carrier concentrations in a potential gradient and an ideal gas, properties that are vital to the functioning of all junction and field-effect devices.

The electron or hole volume concentration n or p at any point in the crystal is given by a $P(E)$-weighted integral of the electron (hole) density of available states as indicated above. The results of these integrals, which are worked out in Kittel (1996) and Sze (1981), are

$$n = 2\left(\frac{2\pi m_n \kappa T}{h^2}\right)^{3/2} \exp\left(-\frac{E_C - E_F}{\kappa T}\right), \tag{3}$$

$$p = 2\left(\frac{2\pi m_p \kappa T}{h^2}\right)^{3/2} \exp\left(-\frac{E_F - E_C}{\kappa T}\right). \tag{4}$$

Putting $E_F = E_C$ for n and $E_F = E_V$ for p gives the special values $n = N_C$ and $p = N_V$, the upper-limit values for the electron and hole densities predicted by this model. In this terminology, the np product is found to be, with $E_g = E_C - E_V$,

$$n_i^2 = np = N_C N_V \exp(-E_g/\kappa T). \tag{5}$$

2. *pn*-JUNCTION FUNDAMENTALS

For simplicity, all of the basic ideas to be developed here are based on a one-dimensional model, i.e., an infinite plane structure with the carriers flowing normal to the plane. Uniform doping concentrations are used for analytical simplicity, except where doping gradients are essential to device design. Translation to realistic two- and three-dimensional structures and doping concentrations follows without conceptual difficulty, but requires complex computer modeling for quantitative analysis.

2.1 Junction Potential and Carrier Concentrations

Two moderately doped n-type and p-type silicon crystals have the same electron affinity χ but different work functions Φ as shown in Fig. 2. The work-function difference, $\Delta\Phi$, is equal to the Fermi energy difference between the two isolated crystals, typically about 1 V for moderately doped silicon.

When the two crystals are brought together, the key energies at the surfaces are joined as shown by the arrows in Fig. 2. Equilibrium of the new system is established by the flow of electrons from the n-Si to the p-Si, and the flow of holes in the opposite direction, with near annihilation of both by recombination. The resulting positive and negative fixed (ionized dopant) charge on either side of the junction, Fig. 3(a), creates an electric field [Fig. 3(b)] that grows until the potential-energy barrier V_0 (c) for both holes and electrons equals the original $\Delta\Phi$.

At this final stage, the Fermi energy is constant throughout the composite system, which is by definition the equilibrium state. The conduction-band edge, representing the rest energy of an electron in the system, is recognized as a plot of the electron potential energy through the junction. Divided by the electron charge $-q$, the conduction-band energy becomes the *electrostatic potential* function $\phi(x)$ through the junction region, except for an arbitrary zero point, as shown in Fig. 3(d), which shows this equilibrium potential. A similar argument identifies the valence-band edge as the potential energy of a hole, with energy measured *downward.* Therefore the potential $\phi(x)$ is a common potential representing the potential energy for both holes

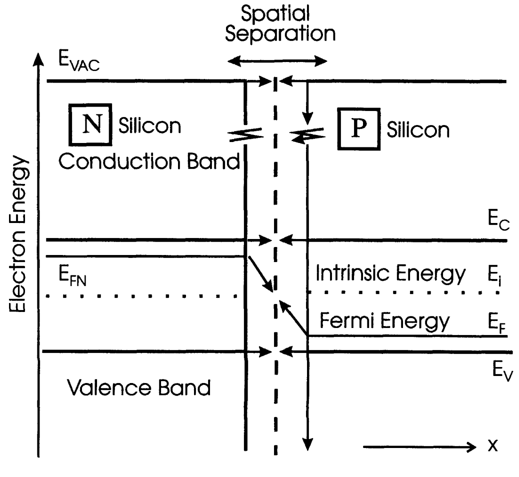

FIG. 2. Joining n-type and p-type semiconductors at a junction.

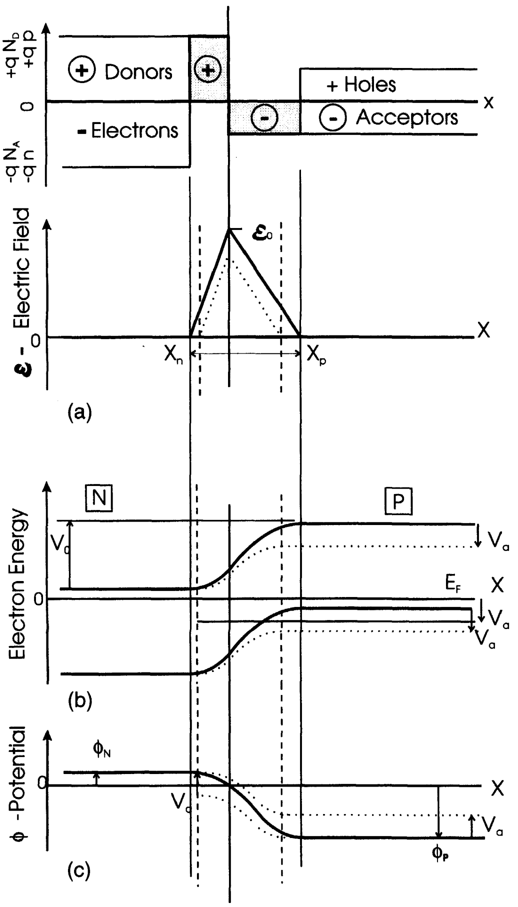

FIG. 3. (a) Uniform dopant, and depleted carrier concentrations around *pn* junction with corresponding electric field in equilibrium; (b) corresponding conduction- and valence-band edges; (c) junction potential. Dotted curves: shifts with forward bias applied, with two potentials in (c).

and electrons in the region where it is defined. It can easily be shown that

$$\phi(x) = E_F - E_i(x). \tag{6}$$

At equilibrium, the difference between the extreme values $\phi_N - \phi_P = V_0$, a potential barrier sometimes referred to as the "built-in voltage." Also shown in Fig. 3(c) is an example of the two *quasi-equilibrium* potentials ϕ_n and ϕ_p, separated by a "forward" applied voltage V. These determine both carrier concentrations in forward bias through the depletion region, in quasi equilibrium with their sources, $+qN_D$ and $-qN_A$.

It is convenient to take the zero point for the potential at the point where $n = p = n_i$, the *intrinsic* carrier concentration, defined in Eq. (5). With reference to the intrinsic point, the concentrations for electrons and holes throughout the space-charge region in equilibrium are written in terms of the potential difference measured to that point, behaving like the densities of ideal gases:

$$n(x) = n_i \exp[+q\phi(x)/\kappa T], \tag{7}$$

$$p(x) = n_i \exp[-q\phi(x)/\kappa T]. \tag{8}$$

These densities typically vary by factors of 10^{12} to 10^{16} from one side to the other. The potential *energies* of the holes and electrons increase to the left and to the right, respectively, at any point in the common electric field $\epsilon(x)$. The proper potential function and the charge distribution are found by solving Poisson's equation simultaneously with the equations for $n(x)$ and $p(x)$, Eqs. (7) and (8). This is a complex process even for the simplified model under discussion here. Finding the approximation to the potential is greatly simplified, however, if one adopts the "depletion approximation" (Shockley, 1949). In this approximation, it is assumed that all of the free electrons are removed from the depleted *n*-Si layer back a length X_n and all the holes are removed from the *p*-Si layer back a length $X_p = (N_D/N_A)X_n$, a "space-charge layer" that is overall electrically neutral. Having found a potential by this depletion approximation, the potential can be used to make a second approximation to the carrier concentration, inconsistent with the depletion approximation, of course, but remarkably useful for the purpose of physical reasoning. The results yield figures similar to Fig. 3(c), with the further assumption that doping concentrations are constant (but not equal) and the host semiconductor is the same on each side of the junction.

2.2 Voltage Control of Minority-Carrier Boundary Values

Given this model (Shockley, 1949), with the electron concentration on the *n* side of the depletion layer given as $n_{n0} = N_D$, and $p_{p0} = N_A$, the electron concentration at the *p*-Si boundary and the hole concentration at the n-Si boundary of the depletion layer are, respectively,

$$n_{p0} = N_D \exp(-qV_0/\kappa T),$$
$$p_{n0} = N_A \exp(-qV_0/\kappa T). \tag{9}$$

With a positive battery voltage V applied to the p-Si with respect to the n-Si, the chemical potential between the n-Si and the p-Si that has to be balanced by the space-charge potential difference becomes $V_0 - V$. This requires that the integral of the electric field $\epsilon(x)$ (Fig. 3) be reduced by positive V, requiring that the junction width be correspondingly reduced. The flow of majority carriers into the two sides of the junction corresponding to this inward boundary displacement permits one to calculate a "junction" capacitance per unit area as the dielectric constant divided by the thickness of the depletion layer (Sze, 1981).

It is assumed that in this process, the majority-carrier concentrations at the n-Si and p-Si boundaries, even as they move, are unchanged when V is applied, and that the carrier concentrations within the junction remain in equilibrium with their sources on opposite boundaries. Thus, with a voltage applied there are two quasi-equilibrium potential functions, obtained by integrating the common electrostatic field within the junction, one anchored at ϕ_N on the n-Si side and the other at ϕ_P on the p-Si side, separated by the applied voltage V, which determine $n(x)$ and $p(x)$ within the junction, including the boundary values for the minority-carrier concentrations in the nondepleted, "charge-neutral" regions on the opposite sides, as they do in equilibrium.

Correspondingly, each of these boundary concentrations is increased by the factor $\exp(qV/\kappa T)$. At the p-Si boundary, the minority-carrier electron concentration becomes, in equilibrium,

$$\begin{aligned} n_p(0) &= N_D \exp[-q(V_0 - V)/\kappa T] \\ &= n_{p0} \exp[+qV/\kappa T], \end{aligned} \tag{10}$$

and the corresponding *excess* electron concentration n' at the boundary is just

$$n'(0) = n_p(0) - n_{p0}.$$

Analogous relations establish the minority-carrier hole concentration at the n-Si boundary. The *excess minority-carrier* concentrations that appear at the boundaries of the charge-neutral n-Si and p-Si regions are therefore, on the p-Si side,

$$p'(0) = p_{n0}\{\exp(qV/\kappa T) - 1\}, \tag{11}$$

and on the n-Si side,

$$n'(0) = n_{p0}\{\exp(qV/\kappa T) - 1\}. \tag{12}$$

These famous exponential relationships (Shockley, 1949) are found over and over again in semiconductor device physics (Sze, 1981).

2.3 The Continuity Equation and Carrier Transport

The junction potential through the space-charge layer establishes carrier *concentrations* within the space-charge region and at its boundaries where they become minority-carrier boundary conditions in the charge-neutral n-Si and p-Si regions. The continuity equation, which is applicable everywhere, is used in this model to determine how the carriers flow inside these quasineutral regions.

For a given carrier type, the continuity equation simply sums up the mechanisms by which the concentration of that carrier type can increase with time. For electrons and holes independently, continuity at any point in space and time is expressed as follows:

$$\frac{\partial n}{\partial t} = G_{\mathrm{Th}} + g_0 - rnp - \nabla \cdot \boldsymbol{F}_n, \tag{13}$$

$$\frac{\partial p}{\partial t} = G_{\mathrm{Th}} + g_0 - rnp - \nabla \cdot \boldsymbol{F}_p, \tag{14}$$

where G_{Th} is the thermal generation rate of electron–hole (e-h) pairs, g_0 is the sum of all other e-h pair generation rates, r is a recombination rate constant, and $\boldsymbol{F}_n$ and $\boldsymbol{F}_p$ are the flow flux density vectors for the indicated carriers.

In equilibrium, all terms are zero except for the thermal generation and the recombination terms:

$$\text{Equilibrium: } 0 = G_{\mathrm{Th}} - rnp, \tag{15}$$

$$np = G_{\mathrm{Th}}/r = n_i^2. \tag{16}$$

It is assumed for simplicity that G_{Th} and r are constants throughout the system. Thus, the product np is constant everywhere. Therefore in the quasineutral p-Si with majority-carrier density $p_{p0} = N_A$, the recombination rate, consisting of four terms (with the x dependence removed to reduce clutter), is

$$r(np) = rn_i^2 + r[N_A n' + (n_i^2/N_A)p' + n'p'].$$

The first term must balance G_{Th} by Eq. (16). Next, the high conductivity and approximate charge neutrality of this p-Si region require $n'(x) \approx p'(x)$, with the result that the second term is much larger than the third term. With rN_A, the "minority-carrier lifetime," defined as the constant τ_n on dimensional grounds, the recombination rate becomes, in one dimension,

$$rnp(x) = n'(x)/\tau_n + rn'^2(x). \tag{17}$$

The nonlinear term is one that arises under conditions of "high injection," which is important at relatively high forward voltage. The condition for "low injection," for which this term can be neglected, is therefore $n'(0) \ll N_A$. Solution of Eq. (13) for an initial uniform $n'(x)$ shows that $n'(t)$ decays exponentially with the time constant τ_n. Thus the term "minority-carrier lifetime."

For the present purpose we take g_0 (the basis for phototransistors, solar cells, video cameras, avalanche detectors, etc.) to be a constant to indicate its presence in some cases.

Anywhere in the pn-junction system, the electron and hole carrier-flow flux densities F_n and F_p, which go into Eq. (13), are accounted for by an electric-field "drift" or "conductivity" term and the simplest diffusion term:

$$F_n(x) = -n(x)\mu_n\epsilon(x) + D_n \frac{\partial}{\partial x} n(x), \tag{18}$$

$$F_p(x) = p(x)\mu_p\epsilon(x) + D_p \frac{\partial}{\partial x} p(x). \tag{19}$$

These equations are most useful in describing carrier flow outside the depletion region.

2.4 Diffusion Equation for Minority Carriers

It may be shown that at normal, uniform doping levels and low injection, the nonlinear, ϵ-driven majority-carrier flow is much less than the carrier-density–gradient diffusion flow. Using this convenient assumption, putting the pieces back into the continuity equation for the excess minority electron concentration, and assuming that the spatial and time-dependent derivatives of $n_{p0}(x,t)$ are zero yields, in one dimension,

$$\frac{\partial n'}{\partial t} = -\frac{n'(x)}{\tau_n} + D_n \frac{\partial^2}{\partial x^2} n'(x) + g_0. \tag{20}$$

This is the famous time-dependent diffusion equation, which is used for modeling device dynamics. It is simply a specialized statement of the original continuity equation under the conditions stated. The time-independent diffusion equation (TIDE), which describes steady-state behavior, is

$$0 = -\frac{n'(x)}{\tau_n} + D_n \frac{\partial^2}{\partial x^2} n'(x) + g_0. \tag{21}$$

In cases where the above sequence of assumptions cannot be made, the field components in Eqs. (18) and (19) must be retained (Kroemer, 1985; Sze, 1981), and other consequences of alloying and dopant gradients, discussed in Sec. 4, must be used to correct the illustrative treatment of this section.

The solution of the above TIDE, in one Cartesian coordinate, is of the form

$$n'(x) = A^+ \exp(+x/L_n) + A^- \exp(-x/L_n) + g_0, \tag{22}$$

in which $L_n^2 = D_n\tau_n$ is the minority-carrier "diffusion length," the average distance that a diffusing excess carrier travels in the neutral p-Si region before it recombines with a majority hole.

The coefficients A^+ and A^- are determined by two boundary conditions on $n'(x)$. In the "long junction," where the p-Si extends far beyond L_n, the physical boundary condition at $x = \infty$ (or $x \gg L_n$) requires that $A^+ = 0$ and $A^- = n'(0)$. Thus, in the neutral p-Si region, and correspondingly in the neutral n-Si region, the excess minority-carrier concentrations are, using x_p and x_n as defined in Fig. 4 for symmetry,

$$n'(x_p) = [n'(0) - g_0\tau_n]\exp(-x_p/L_n), \tag{23}$$

$$p'(x_n) = [p'(0) - g_0\tau_p]\exp(-x_n/L_p). \tag{24}$$

These excess-carrier concentrations are electrically compensated by majority-carrier concentrations that accumulate during the transients between steady states by majority-carrier drift in the field created by the excess minority carriers. The result is a compensating hole and electron population in a "diffusion capacitance" each side of the junction in which the charges of opposite sign are separated by statistical improbability of their recombining instead of an insulator. Each capacitance per unit junction area is given by the excess-carrier charge per volt, which can be calculated by inspection of Fig. 4. These two "diffusion capacitors" are in parallel with the junction capacitor defined in Sec. 2.2. All are voltage dependent, a very useful property especially in reverse bias.

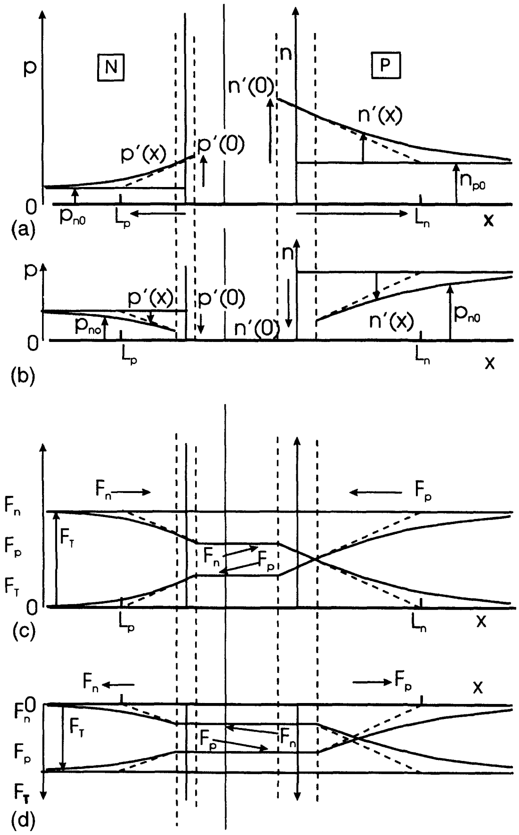

FIG. 4. (a) Minority-carrier concentrations near junction in forward bias; (b) same for reverse bias; (c) electron and hole flux densities near junction in forward bias; (d) same for reverse bias.

The corresponding minority-carrier flow rates are defined in terms of the carrier flux densities, using the diffusion component of Eq. (18) and its equivalent for holes on the p-Si and n-Si sides of the junction, respectively:

$$F_n(x_p) = D_n\frac{[n'(0) - g_0\tau_n]}{L_n}\exp\left(-\frac{x_p}{L_n}\right), \tag{25}$$

$$F_p(x_n) = D_p\frac{[p'(0) - g_0\tau_p]}{L_p}\exp\left(-\frac{x_n}{L_p}\right). \tag{26}$$

The above two minority-carrier concentrations and current densities for a long junction for forward and reverse biases illustrate the role of optical generation rates in the pn junction. For example, a voltage input controls $F_n(x_p)$ through $n'(0)$. Optical generation g_0 controls $F_n(x_p)$ through $g_0\tau_n$. But these two input quantities drive the electron flux density in opposite directions. The first is the normal semiconductor rectifier, and the second is the pn-junction solar cell. Having made this point, we set g_0 equal to zero for simplicity until it is needed.

The above electron and hole flux densities are plotted in Fig. 4. In the depletion layer, which is exaggerated here, the carrier flux densities are shown constant, meaning that $G_{\text{Th}} - rnp$ in this layer is assumed to be zero.

Along a common x coordinate the two minority-carrier flow components above have opposite signs. Electrons flow to the right, holes to the left. The corresponding carrier flux densities are multiplied by charges of opposite sign. Thus, the two current components are in the same direction, in this case to the left. If we assume, for simplicity, that $N_D \gg N_A$, then the electron current density dominates. With a diode area A, the electron current density in the $-x$ direction is

$$I = I_0\{\exp(qV/\kappa T) - 1\}, \tag{27}$$

$$I_0 = qA(D_n/L_n)n_{p0}. \tag{28}$$

This expression is plotted in Fig. 5. In gen-

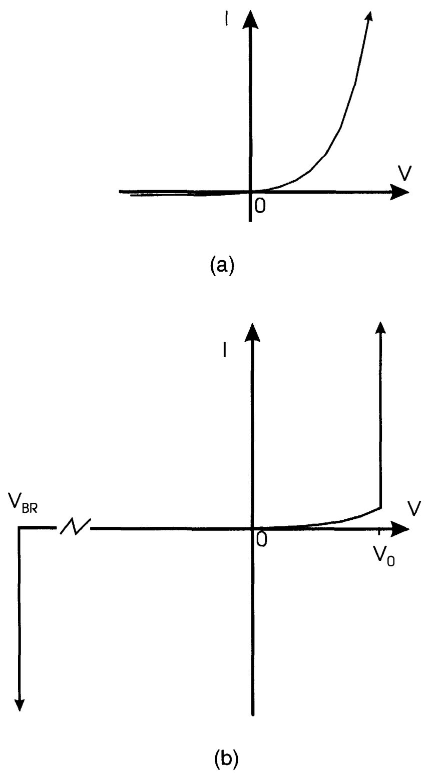

FIG. 5. (a) "Ideal" electron current vs voltage for $p^{++}n$ junction. (b) Observed *IV* characteristics, with the "ideal" component reduced to show the practical "reverse-breakdown" and "flat-band" breakdown components.

eral, the hole component would simply add to I_0.

It should be emphasized that in this model, the details of recombination and carrier transport in the depletion region have been bypassed. The carrier concentrations obtained by quasi-equilibrium potentials as outlined simply supply boundary conditions, or "injection" concentrations at the boundaries. However, the carrier flux densities can always be described in terms of field- and gradient-driven components even if the transport model is not as simple as Eq. (8) and its hole equivalent. In equilibrium, each carrier population at each position in the same field, but in its own gradient, adjusts its own concentration and gradient for zero net flow, and the product is n_i^2 even though n_i^2 may be a function of position. In forward bias, the field is reduced, and the diffusion flow overbalances the field-driven component. In reverse bias the unbalance reverses sign and favors the field over the diffusion flow.

When a forward voltage approaching V_0 is applied, the "low injection" approximation fails, and the analysis above fails. When the voltage reaches V_0, the space-charge layer disappears, the potential barrier, which controlled the minority-carrier boundary values, disappears, and the recombination process still required to maintain junction current is confined to a continuously thinning layer. At $V = V_0$ the current rises abruptly with V, giving rise to the rectifying property that has so many practical uses. Compared with normal operating forward currents, the current $I(V)$ expressed in Eq. (27), represented by Fig. 5(a), is essentially negligible, although the concepts are vital in the BJT. Commonly encountered I–V characteristics of "diode rectifiers" are shown in Fig 5(b).

In reverse bias, the field ϵ_0 at the center of the junction increases until the energy gained between inelastic collisions is large enough (about 1.8 eV) to cause impact ionization. When the *e-h* pair creation rate per collision exceeds unity, there is again a sharp rise in the device current. Under controlled conditions, this phenomenon can be used to advantage, such as multiplication in the collector of a transistor or as a very efficient photon detector (Sze, 1981, Chap. 13). Many useful functions can be performed by designing *pn*-junction devices with a variety of doping levels and gradients, and operating in various voltage ranges (Sze, 1991; Streetman, 1995).

2.5 The *pn* Junction as a Generation/Recombination Device

It should be clear that in *forward bias*, electrons flow into the junction area from the *n*-Si region, but no electrons flow out of the *p*-Si region. Correspondingly, holes that flow in from the *p*-Si region do not flow out through the *n*-Si region. Since the charge carried by each carrier type is the same except for the sign, all electrons and holes flowing into the junction *recombine* in the diffusion regions on either side of the junction as well as in the junction depletion region, as shown schematically in Fig. 6. Each of the recombination processes is represented by a circled *R* in the figure. At the

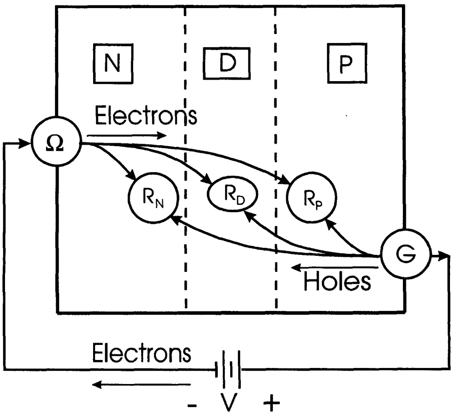

FIG. 6. *pn*-junction diode recombination (*R*) and generation (*G*) centers in forward bias.

boundary between the *p*-Si and the metal contact, a generation process, equal to the sum of all of the recombination process, is shown as a circled *G*. The junction of the *n*-Si with the metal contact is assumed to be Ohmic and is represented by a circled *Ω*. The recombination processes generate heat in most cases and generate photons in others, such as LEDs and lasers. The generation processes absorb heat, in some designs used for electronic cooling. Reversing the voltage polarity changes all the directions of flow and interchanges all the *R* and *G* centers. In Fig. 7 an n^+p "short diode," with the *n*-Si side very heavily doped, is shown. On the *p*-Si side, at a distance *W*, much shorter than L_n, from the depletion layer, the doping concentration N_A increases abruptly by, say, a hundredfold. The solution to the diffusion equation within the length *W* from the depletion edge is approximately a straight line. The electron flow at the boundary is therefore

$$F_n(0) = D_n n'(0)/W, \tag{29}$$

and the recombination rate within *W* is

$$R_{P1} = n'(0)W/2\tau_n. \tag{30}$$

Since $W \ll L_n$, the flow due to the gradient almost all flows through the layer *W* into the heavily doped region, where it is consumed by the recombination rate R_{P2}. Therefore the ratio of the two recombination rates is approximately

$$R_{P2}/R_{P1} = 2(L_n/W)^2. \tag{31}$$

Such "short diodes" store relatively little diffusion charge and therefore have a relatively short response time. Further, for a given forward bias, the current flow is increased by a factor of L_n/W over the long diode, in both directions.

In the BJT we separate these two current components and use the small one to control the other.

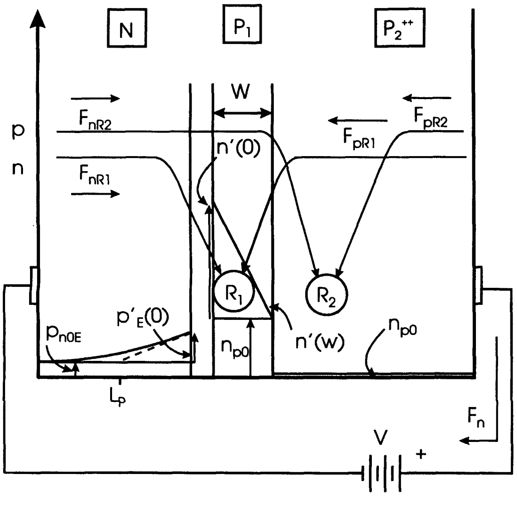

FIG. 7. Modification of *pn* junction of Fig. 7 with light and heavy doping on the *p*-Si side.

3. BIPOLAR TRANSISTOR PHYSICS

The original "semiconductor triode" consisted of a block of *n*-type germanium (the base), with two closely spaced wire "points" on the surface, as shown in Fig. 8. By pass-

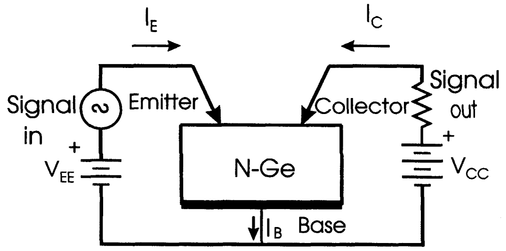

FIG. 8. The original Bardeen and Brattain (1948) germanium point-contact type-A transistor.

ing electrons at high density from the wire "points" into the base, atoms from the metal wire (tungsten or phosphor bronze) were diffused into the germanium base, forming a hemispherical *p*-Si shell under each contact. Figure 9(a) shows the popular schematic representation of the *pnp* and the *npn* transistors. Figure 9(b) shows the *pnp* and *npn* transistors arranged in the two-port circuit configurations corresponding to common-base and common-emitter connections. A readable history of the technology development during the early years of the transistor and its antecedent semiconductor devices can be found, along with copious references, in the valuable book by Shive (1959). Probably the first major compilation of transistor technology was published about this time by Western Electric (1958).

The fundamental physics of the planar BJT follows simply from the *pn*-junction physics discussed in Sec. 2, on the basis of uniform doping in the emitter, base, and collector. This is not a quantitatively adequate representation of the most advanced high-frequency structures in use today, but the concepts are fundamental to all transistors, and the mathematical formulation is more tractable than the computer simulations now in common use. The earliest comprehensive texts (Gärtner, 1960; Lindmayer and Wrigley, 1965) treat the essential physics in a more cumbersome way and have a heavy emphasis on circuit models. There are a great many books that have made it to the marketplace in recent times, some quite readable (e.g., Streetman, 1995). Other fairly recent texts deal with transistors and modern technology in more complex detail (Warner and Grung, 1983; Muller and Kamins, 1986; Wolf 1986, 1990, 1993).

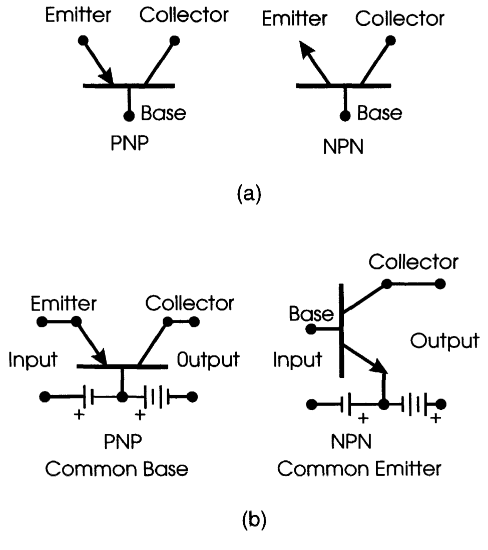

FIG. 9. (a) Standard symbols for *pnp* and *npn* transistors. (b) Two-port connections for common-base and common-emitter configuration.

3.1 Minority-Carrier Current Flow, and Current Gain

Figure 10 shows an illustrative minority-carrier profile of a planar *npn* transistor under three conditions. The thin *p*-Si region sandwiched between the two *n*-Si regions is the functional equivalent of Bardeen and Brattain's "base" as shown in Fig. 8. As such, it is similar to the npp^{++} structure of Fig. 7, except that the minority-carrier concentration at the end of the thin *p* layer is voltage controlled by a second junction rather than the transition to a heavily doped p^{++} layer.

The base is much thinner than the minority-carrier diffusion length in the material, L_n

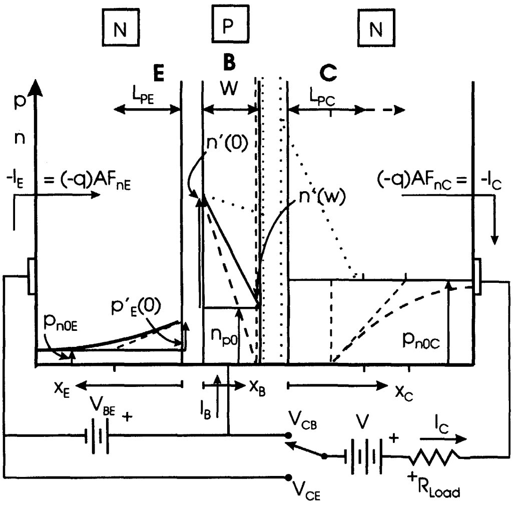

FIG. 10. Schematic diagram of equilibrium minority carrier connections and voltage dependent boundary conditions for the *npn* bipolar transistor for forward biased emitter-base junction, zero-biased, $V_{CB} = 0$, (solid); and both forward (dotted) and reverse (dashed) biased collector-base junction. Switch converts from common-base to common-emitter configuration. Transistor area is A.

= $\sqrt{d_n\tau_n}$. The two n-Si regions differ in that the emitter is typically doped much *greater* than the base, to keep the diffusion capacitance and recombination rate low on that side of the junction. The collector on the right is typically doped much *less* (i.e., much higher minority-carrier concentration) than the base so that when the reverse-bias voltage across the collector–base junction varies, most of the accompanying variation in the depletion-layer thickness takes place in the collector. This reduces the negative feedback resulting from normal signal variations in the load current. The same doping ratio puts most of the depletion layer between collector and base and also permits higher operating voltage before the depletion layer in the base is pushed all the way through to the emitter, a condition known as base "punchthrough."

Consistent with the above, a high doping concentration in the base compensates the geometrical resistance increase, due to the thin base layer, through which the base-emitter current flows, perpendicular to the currents in the plane of the cross section of Fig. 10. The low-resistivity subcollector, which returns the transistor current, is not shown.

The essential physics of gain in the *npn* transistor is illustrated by biasing the electrodes as shown in Fig. 10, with the switch in the common base (CB) position, and both R_{Load} and $V_{\mathrm{CB}} = 0$. The emitter–base (EB) junction is forward biased, V_{EB}, as shown in Fig. 4. The EB space charge region narrows slightly on both sides, but we eliminate this detail in this and related figures. The excess minority-carrier boundary conditions shown, $p'_E(0)$ and $n'(0)$, rise exponentially with V_{BE}. Thus, at only 0.1 V, these concentrations are over 50 times the equilibrium concentrations. On the scale of $n'(0)$ the equilibrium concentrations n_{p0} and p_{n0C} would be difficult to see, and on the scale of p_{n0}, n_{p0} would be difficult to see. To cover the full range, therefore, it is normal to represent concentrations on a logarithmic scale. Nevertheless, to describe minority-carrier diffusion, carrier concentrations are measured on linear scales, which accounts for schematic diagrams such as Fig. 10, which, even though they are quantitatively unrealistic, are convenient for discussing the physics. In particular, to illustrate the physics of current gain, it is convenient to assume that the excess carrier concentration on each side of the CB junction is zero, which is the condition $V_{\mathrm{CB}} = 0$ illustrated by solid lines in Fig. 10.

The excess-carrier concentrations on each side of the EB junction are established by the respective equilibrium minority-carrier concentrations and the voltage V_{EB}. On the emitter side of the junction, as demonstrated in connection with Fig. 4, and with

$$p_{n0E} = n_i^2/N_{PE},$$

we have

$$p'_E(0) = p_{n0E}\{\exp(qV_{\mathrm{EB}}/\kappa T) - 1\}, \tag{32}$$

and at the boundary on the other side of the junction in the p-Si base,

$$n'(0) = n_{p0}\{\exp(qV_{\mathrm{EB}}/\kappa T) - 1\}. \tag{33}$$

The excess minority-carrier concentration in the base, a decaying exponential solution of the diffusion equation, Eq. (23), is almost a straight line, because the diffusion length L_n is much greater than the width of the neutral base region, W. Actually, the solution sags slightly below the straight-line approximation shown, giving a small positive second derivative, which by the continuity equation tracks the local recombination rate, expressed by Eq. (21). The gradient of the straight line is used to calculate the diffusion flow of electrons from emitter to collector, and the integral of the line divided by τ_n gives the essential base recombination rate required to support this flow. Their ratio gives the highest possible current gain, the "ideal β" of the transistor. In summary, the following relations are written in terms of the conventional characteristic lengths and times, and also the corresponding materials parameters in order to emphasize their equivalence.

Base diffusion rate (cm^{-2} K^{-1}):

$$\begin{aligned} F_{nD(\mathrm{EC})} &= D_n n'(0)/W \\ &= D_n \frac{n_i^2}{N_A}\frac{1}{W}\left\{\exp\left(\frac{qV}{\kappa T}\right) - 1\right\}. \end{aligned} \tag{34}$$

Base recombination rate (cm^{-2} K^{-1}):

$$\begin{aligned} R_B &= n'(0)W/2\tau_n \\ &= \tfrac{1}{2} r_n n_i^2 W\{\exp(qV/\kappa T) - 1\}. \end{aligned} \tag{35}$$

Ratio = ideal β:

$$\beta_{\text{ideal}} = I_{\text{collector}}/I_{\text{base}} = 2(L_n/W)^2 = 2D_n/(r_B N_A)W^2. \tag{36}$$

This base-limited current gain is the ideal. Recombination in the base is the minimum price to pay for the diffusion flow through the base. In the early transistors with a thick base, it was the dominant price.

Another price to pay is the corresponding diffusion flow from the base into the emitter, established by the same V_{EB}. However, although V_{EB} (with N_{DE}) establishes the minority-carrier boundary value adjacent to the base, the carrier flow and ultimate recombination that takes place in the emitter depends on the emitter structural details, discussed in Sec. 5.2. In the simplest example of the long emitter, the gradient is determined by L_{pE}, i.e., by the electron diffusivity D_{pE}, and the minority-carrier (hole) lifetime τ_{pE}. The excess hole concentration at the emitter side of the junction is

$$p'(x_E) = p'(0)\exp(-x_E/L_{pE}). \tag{37}$$

The corresponding hole diffusion flux density ("hole injection") from base into emitter is

$$F_{pD(E)}(0) = -D_p \left|\frac{\partial p'(x_E)}{\partial x}\right|_{x_E=0} \tag{38}$$

regardless of where this flow is consumed by recombination. Continuity for the long, uniformly doped emitter requires that the diffusion rate at the boundary be equal to the recombination rate in the emitter:

$$F_{pD(E)}(0) = D_{pE}\frac{p_E'(0)}{L_{pE}} = \frac{p_E'(0)L_{pE}}{\tau_{pE}} = \left(\frac{D_p r}{N_{DE}}\right)^{1/2} n_i^2\left\{\exp\left(\frac{qV}{\kappa T}\right) - 1\right\}. \tag{39}$$

This flow rate, regardless of its description, is another component of the total majority-carrier flow rate into the base and adds another term to the denominator in the expression for β. The "actual" β becomes, allowing for different values of n_i^2 in the base and emitter,

$$\beta_{\text{actual}} = \frac{\dfrac{D_n}{W}}{\dfrac{W}{2\tau_n} + \left(\dfrac{n_{iE}}{n_{iB}}\right)^2 \dfrac{N_A}{N_{DE}}\dfrac{L_{pE}}{\tau_{pE}}} = \frac{\dfrac{D_n}{W}}{\dfrac{W}{2} r_B N_A + \left(\dfrac{n_{iE}}{n_{iB}}\right)^2 \dfrac{N_A}{N_{DE}}\sqrt{D_{pE} r_E N_{DE}^+}}. \tag{40}$$

An equivalent transit time may be determined using the above relations. The base diffusion flow rate is equated to the average minority-carrier concentration in the base times the effective velocity of the average carrier density, a numerically equivalent, but physically different, *drift* conductivity process. From this, one can calculate the *effective diffusion-transit velocity,* $\mathcal{V}_{\text{eff}}$:

$$D_p\frac{p'(0) - p'(W)}{W} = \frac{p'(0) + p'(W)}{2}\mathcal{V}_{\text{eff}},$$

from which, for $p'(W) = 0$, is just

$$\mathcal{V}_{\text{eff}} = (2/W)D_p. \tag{41}$$

But by definition, the effective diffusion transit velocity is related to the mean-carrier transit time τ_t by

$$\tau_t = W/\mathcal{V}_{\text{eff}}. \tag{42}$$

The number of carriers that make it through the base between recombination events is the current gain of the device:

$$\frac{\tau_p}{\tau_t} = 2\frac{D_p\tau_p}{W^2} = 2\left(\frac{L_p}{W}\right)^2, \tag{43}$$

which is in agreement with the ideal β, Eq. (36), not the actual β, because it depends upon the base transport process only. Since the transport mechanism of the useful current in the device is by diffusion of minority carriers through the base, many authors refer to this generic structure as the "diffusion transistor." In designs where substantial electric fields are built in, a "drift" component may be added in the "high-injection" range, which complicates the details of the analysis. Whatever the transport mechanism, it is the minority-carrier flow in the base that carries the useful device current *between regions in which they are majority carriers.*

Modern technology has produced manufacturable BJTs with high current gains and very high cutoff frequencies. The implications are that the base is very thin—1,000 Å or less—with little recombination taking place there; that the minority-carrier concentration in the emitter is extraordinarily low; that the recombination rate constant r is abnormally high, and the corresponding n_i^2 is low; that most of the recombination is taking place in the emitter; and that the minority-carrier transport in the base is significantly assisted by the electric field. The details, which are difficult to confirm experimentally, are discussed further in Sec. 4.

3.2 Common-Base and Common-Emitter Characteristics

The above discussion is based on the *common-base* (CB) configuration, in which the two junction voltages that control the two minority-carrier boundary values in the base are specified independently for conceptual simplicity. Whatever the circuit configuration, the collector/emitter current ratio, α, is a little less than one. In that configuration, however, the low-voltage emitter-circuit *current* essentially replicates itself in the high-voltage collector circuit with a consequent power gain. However, the *common-emitter* (CE) configuration is the one most commonly used in communications, control, and computer circuits because the current gain is much higher, namely, the β derived above. In addition, the "output" current I_C is approximately linearly related to the "input" current I_B over a large part of the normal operating range.

The two (*npn*) circuit configurations, CB and CE, are shown in one diagram in Fig. 10, to emphasize their relationship. Minority-carrier concentrations in the base and collector for three cases are illustrated by solid lines for the case $V_{\mathrm{CB}} = 0$; dashed lines for the normal reverse-biased CB junction; and dotted lines for the forward-biased CB junction required to shut off the collector current. Given zero net recombination in the BC depletion layer, minority electrons and majority holes diffuse into the quasineutral collector where they recombine to give zero net collector current.

The three-terminal properties of the two circuit configurations are shown in Figs. 11(a) and 11(b). For the CB configuration, with the $V_{\mathrm{CB}} = 0$ switch setting in Fig. 10, increasing V_{BE} raises $n'(0)$ in the base exponentially as indicated above in Eq. (33), and the electron diffusion through the base increases proportionally, neglecting base width change. This is shown schematically in Fig. 11(a), where F_{nC} has been converted to iI_C, and F_{nE} has been converted to $-I_E$. The emitter current is shown on a perpendicular fold as a function of V_{EB}. It is then projected back on the common I_B-I_C axis, although $I_C = \alpha I_B$, with α typically close to 1.0. The two currents are shown equal along this axis for convenience. The I_B curve is multiplied by α

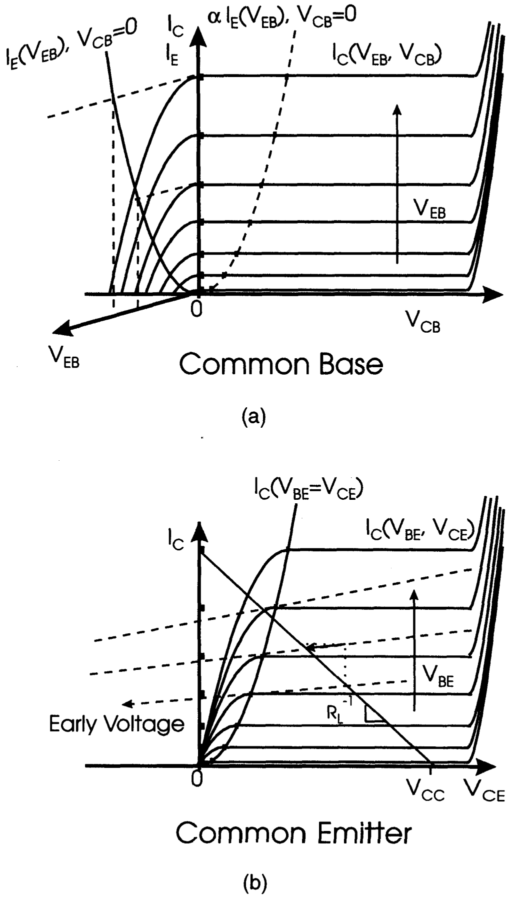

(a)

(b)

FIG. 11. Transistor current–voltage characteristics for (a) common-base and (b) common-emitter connections, with I_C vs $V_{\mathrm{CB}} = 0$ superimposed for both cases.

and swung around on the $I_C(V_{EB},V_{CB})$ coordinates.

With increasing V_{CB} above zero, the collector current shown as horizontal lines in Fig. 11(a) slopes upward as shown by the dotted lines in Fig. 11(b). This increase of current, known as the "Early effect," is due to a depression of $n'(W)$ at the collector boundary of the base and to a corresponding narrowing of the base, as shown in Fig. 12, both of which increase the gradient of the minority-carrier concentration in the base and add generation processes either side of the CB junction, which are illustrated in Sec. 3.3. The Early voltage for a given bias is the point where the corresponding dotted line intersects the negative V_{CE} axis.

Reversing the polarity of V_{CB} moves the characteristics into the northwest quadrant of Fig. 11(a). With given V_{BE} and $n'(0)$, $n'(W)$ is raised, reducing the minority-carrier gradient in the base eventually to zero, at which point the collector carrier flow is reduced to zero.

In the CE structure, V_{BE} is still an independent variable, but the voltage supply V_{CB} is reconnected from collector to emitter, and is labeled V_{CE}. In this case, to keep $V_{CB} = 0$ while raising V_{BE} requires maintaining V_{CE} equal to V_{BE}. In effect, the F_{nC} axis of Fig. 11(a) bends over and becomes congruent with the dotted exponential curve in that figure, and the shutoff curves to the left of that boundary all correspond to a forward-biased CB junction, as in Fig. 11. All turnoff curves go through zero together, because both junctions have the same forward bias when emitter and collector are tied together.

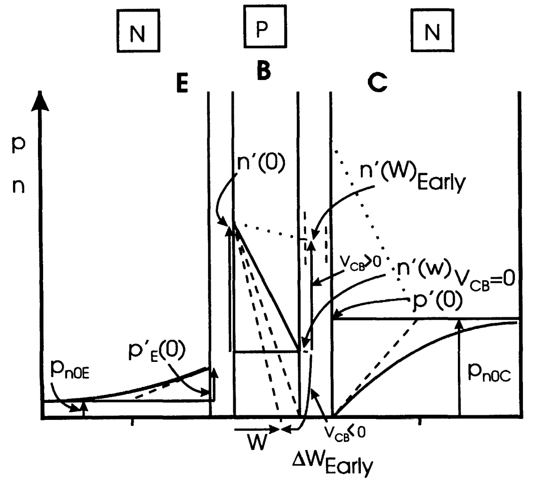

FIG. 12. Effect of collector reverse-bias voltage on base narrowing and ultimate punchthrough preceded by sharp current rise, and the forward "Early voltage" bias required to bring I_c to zero.

The *I–V* characteristics in this figure are voltage parametrized for convenience of discussing the device physics, and are thus exponentially spaced. Using input current as the parameter would result in linear spacing of the horizontal, "saturation" current lines. Since both I_B and I_C are exponential functions of V_{BE}, it is common practice to plot them on logarithmic plots of V_{BE}. The transistor properties approach the ideal if the two curves run parallel to each other, in which case the current gain is constant over the range. Such curves, which are shown as exponentials in Fig. 11, are known as Gummel plots, and are found in almost every report on transistor research.

Operating conditions in which the CB junction becomes forward biased shut off the diffusion escape of minority carriers in the base when trying to switch off the transistor with the BE junction. Discharging the stored charge in the base depends then on recombination of the excess minority carriers in the base, which depends on the minority-carrier lifetime in the base, τ_n. This limitation can sometimes be avoided by means of the "Schottky clamp," which for the *npn* structure is an extension of the aluminum metal connection to the base or a platinum deposit that overlaps the collector. The aluminum makes Ohmic contact to the *n*-Si base, but makes a Schottky-barrier diode with the *n*-Si collector. The resulting Schottky diode, the equivalent of an abrupt one-sided *pn* junction, is polarized the same as the base-collector junction it parallels. But the Schottky-barrier diode is much more conductive than the base-collector junction, which prevents any significant forward voltage from appearing across either. In the normal polarity, both junctions are reverse biased, and the Schottky-barrier diode can be considered to be an open circuit.

3.3 The BJT as a Generation/Recombination Device

A helpful exercise to elucidate the fundamentals of BJT operation is to carry out the generation- and recombination-rates accounting in detail for specific examples.

As indicated in Sec. 3, each of the generation (*G*) and recombination (*R*) processes essential or peripheral to BJT operation involves equal and opposite electron and hole flows into (*R*) or out of (*G*) these centers. Each such center is therefore associated with a continuous charge-flow loop.

Figure 13 shows all the closed flow loops for a *pnp* BJT under normal bias conditions, for a long emitter. Loops passing through the base are grouped all together for the common-emitter connection and in pairs for the common-base connection illustrated in Fig. 11. Alteration of device parameters to give "abnormal" bias conditions and/or a "short, inhomogeneous, composite" emitter does not alter this picture. Such changes affect only the values associated with each of the *R* or *G* processes. All electron-flow components into the collector are *R* processes that correspond to *G* processes elsewhere, as identified in the figure. *R* and *G* processes interchange when external voltages reverse. The base components are all electrons with continuity through Ohmic contacts. From the physical description of each of these processes, simple mathematical expressions for the characteristics of the BJT under all conditions can be written down.

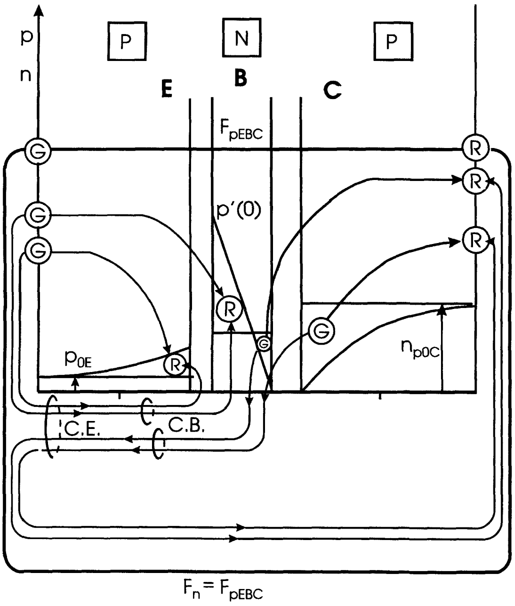

FIG. 13. Normal BJT bias conditions, showing continuity in terms of generation and recombination centers and all closed charge-flow loops.

The "useful current density" flow in this structure is the outer loop, $F_{pD(EC)}$, in which holes generated in the metal–emitter interface recombine again at the collector–metal interface unchanged in flow rate, which is controlled by diffusion transport through the base as minority carriers. All of the other closed flow components exist solely to support this hole-diffusion flow component. The reader should repeat Fig. 12 for the *npn* transistor to establish fully this vital concept.

3.4 Other Related Structures

There are many other application-inspired device structures that have been devised. Most of them are covered in Sze (1981), which in addition is an excellent index to the literature. All of them build on concepts developed in the preceding sections. Power semiconductor device design is in a class by itself, but there are three important structures that deserve mention here because they relate closely to the material developed in this article: the drift transistor, the *npnp* switch, and the photodiode and transistor. First we treat the photodiode and transistor.

The description of carrier transport in the base in Sec. 3 was limited to minority-carrier diffusion in the uniformly doped base. It is possible to achieve nearly uniform doping by the epitaxial growth technology referred to in Sec. 4, but traditional diffusion and implantation processes result in strongly graded, nearly Gaussian dopant profiles. The result is a strong rate of majority-carrier density *decrease* in the base between emitter and collector variation of majority-carrier concentration through the base, by as much as 100/1. The diffusion of majority carriers toward the collector in the overall charge-neutral base comes to an equilibrium value at every point, as it does in the depletion of the *pn* junction, when a large enough electric field, created by uncompensated charged dopant atoms, develops to balance it, Eq. (18) or Eq. (19). This same field drives *minority* carriers toward the collector, balancing, of course, the equilibrium minority-carrier *diffusion* in the "wrong" direction. When minority carriers are injected from the emitter, diffusion is incremented in the direction of the collector, as expected, but the large electric "drift" field (say, 20 meV over 100 nm = 2000 V/cm) contributes significantly to their velocity toward the collector, decreasing the net base transit time. The details of this "drift transistor" transport process are discussed in Kroemer (1985) and Sze (1981).

The BJT operates by means of a forward bias V_{EB} raising the minority-carrier concentration at one end of the base while holding the other end down with a reverse V_{CB}, and pouring minority carriers into the collector by diffusion. The same carrier gradient in the base can be achieved by removing the emitter and allowing the minority-carrier concentration to rise with photon absorption, the result of the $g_0\tau_n$ term in Eq. (24). That is essentially the photodiode, and the *pn*-junction solar cell in raw form (Sze, 1981, Chap. 4).

The phototransistor adds positive feedback to the photodiode above. The minority electrons generated in the lightly doped *p*-type base would normally flow to the collector, originating from the metal–semiconductor diode *R* contact. Instead, the *P* contact floats, and the photogenerated holes are drawn into the base of a subsidiary transistor at the edge of the structure. The emitter of that transistor is positively biased with respect to the collector, and so an electron diffusion from emitter to collector flows in response, and in parallel with the photogenerated electron flow from the photodiode.

A second device structure of special interest is the *npnp* switch. As noted in Sec. 3.3, the *G* center in the collector provides a minority-carrier diffusion flow into the transistor base, which reduces the base current required to support the base diffusion electron flow. Normally this quantity is essentially fixed because the junction is in reverse-bias saturation, as shown in Fig. 11. That component of diffusion current from collector back into the base could be increased, however, by *raising* the other end of the minority-carrier distribution in the collector.

The obvious way of doing this in the case of the photodiode was photogenerated minority carriers, the g_0 process. In the transistor, the means was to add a junction. In the present case, one simply adds a junction after the collector in Fig. 11. The last three layers, then, form a *pnp* transistor, the input of which is the original diffusion flow of electrons, which places a negative charge in the collector, and the output of which is the forward-bias flow of minority holes through the (original) collector into the (original) base, which constitutes positive feedback. This so-called "hook collector" (Lindmayer and Wrigley, 1965) was one of the earliest transistor innovations, along with the "drift transistor." If the gain is greater than one, this device will be unstable and be switchable from one state to the other. The result is the *npnp* switch, which led to the silicon-controlled rectifier and the thyristor, key components in power-control technology. These are discussed in TRANSISTORS, POWER.

4. MATERIALS SCIENCE, TECHNOLOGY AND APPLICATION

From the beginning, the transistor—the BJT, field-effect transistor (FET), or any of its derivatives—has been a multidisciplinary effort. All disciplines involved have contributed to the engineering of smaller, faster, lower-power, more reliable devices, driving and being driven by systems objectives. The participating disciplines themselves have been enriched, in some cases completely reconstituted, and in others created altogether.

Astonishing developments announced in the past few years rest on a succession of ideas and exploratory studies embedded in this evolutionary process, in analogy with the precursor studies that led to the first transistor, except that the driving forces have been much more sophisticated design objectives. In particular, the market pressures for "personal assistant" electronics, starting with the pocket calculator and moving up to the organizers, notebooks, the cellular telephone, pagers, and the global positioning system, implied that the final stage would be the integration of all these units by means of radio communications, from the very short to the very long range, directly or though cellular links that extend seamlessly the wired internet.

4.1 Design Objectives

To accomplish this, it would be necessary to include dynamic random-access memory (DRAM), microprocessor, analog, nonvolatile memory, and radio-frequency circuits all on the same chip, and all manufactured by compatible processes, materials, and temperature/time profiles. The result would be the low-voltage, low-power "mixed signal" technology that forms the centerpiece of the enormous range of applications that microelectronics is only beginning to fill.

One of the key challenges was electronics that could operate in the communications band that extends up to 80 GHz. The range above about 2 GHz has been the province of gallium arsenide technology, a difficult technology to manage, and incompatible with silicon technology. The challenge, then, was to displace the gallium arsenide technology with one that could perform equivalent functions within the realm of developing silicon technology. Developments already taking place in silicon technology to achieve higher-speed, lower-power, and higher-yield digital circuits and memory have now made this possible.

The key to success was the evolution of a new concept of semiconductor device manufacturing: building the transistor *on top of* rather than *down inside of* the silicon wafer. In the traditional methodology, the BJT was made by means of a sequence of doped-glass deposition and thermal, diffusive drive-in, or alternative dopant-ion implantation steps, followed by an anneal at temperatures high enough to activate the dopants. The deepest of these doped layers (taking an *npn* transistor for example) must be a low-resistance *n*-Si buried "subcollector" layer below the collector for carrying the collected current back to the surface. Second comes the *n*-Si collector, at a much lower doping level, then the *p*-Si base at a higher doping level, and the *n*-Si at a still higher level, to minimize base-emitter injection current. All of these processes involve pushing the dopants through the silicon and through each other, involving some uncontrollable crystal-induced channeling, anisotropic diffusion, and attendant damage. These result in nonuniformity of the semiconductor and transport parameters, and often reliability problems. Even in the best case, the complementary error function and Gaussian doping profiles determined by the physics of the deposition and implantation properties are not those that the designer would wish to choose (Harame *et al.*, 1995).

4.2 Epitaxial Silicon/Germanium BJT

The latest strategy, at least for high-speed microelectronics, is (for the *npn* transistor) to grow the *n*-Si collector epitaxially on a heavily doped *n*-Si "subcollector," and, further, to design and implement accurately a variety of heterostructures (Kroemer, 1957, 1982, 1985). The reactor is a UHV/CVD system using silane for the silicon, phosphene and diborane for *n*- and *p*-type doping, and germane for growing silicon-germanium layers, all in a diluting hydrogen carrier gas. The subcollector, used to carry the current back to the surface through a low-resistance path, is first grown epitaxially on a lightly doped *p*-Si epitaxial layer, which can be used for a complimentary MOS transistor. This layer is first grown on a heavily doped, conductive *p*-Si wafer substrate. Such a layered structure also sets the stage for complete device isolation by means of surrounding trenches.

At the designed collector thickness, the phosphene is replaced by a programmed concentration of diborane, which provides a silicon base of controlled dopant-concentration profile. Simultaneously, more or less, a controlled flow of germane is abruptly added to the stream, growing a SiGe alloy base, starting typically at about 10% Ge. At the end of the approximately 100 nm of epitaxial base growth, the germane flow, which has been decreasing linearly toward zero, and the diborane, which has been rising toward its peak concentration, are terminated, and a moderately doped *n*-Si emitter "cap" of about 10 nm thickness is grown. On top of this is deposited a highly doped arsenic polysilicon contact that is annealed by rapid thermal anneal (RTA) enough to activate the arsenic dopants, and just about penetrate the "cap" by diffusion (Li *et al.*, 1987; Harame *et al.*, 1995; Crabbé *et al.*, 1993). There are many details to be understood about the effects of the profiles and their overlaps at the junctions, and even the option of placing an undoped silicon layer in the middle of each layer. The discussions in the literature are not easy to follow. However, there is no question that SiGe-BICMOS technology for telecommunications above 50 GHz has been established (Kermarrec *et al.*, 1994). Device designs up to 75 GHz, and perhaps higher, have been reported (for reference, 10 GHz corresponds to the popular 3-cm radar frequency) (Patton *et al.*, 1990; Kermarrec *et al.*, 1994).

The use of UHV/CVD has made it possible to grow the structures described with atomic-layer precision and repeatability. In such structures, current gain and current density have been increased, the Early voltage, Fig. 12, has been increased, junction capacitances have been reduced, frequency response has been increased, and collector breakdown voltage has been maintained. A significant advantage has been gained by means of the graded Si/Ge epitaxial base. Although the details are beyond the scope of this article, and some may not yet be fully understood, it can be said that the two salient properties of the structure derived from the graded Si/Ge base—the effect of the high germanium content near the collector, and the doping profile in the very thin base—are these:

1. The sharp discontinuity in germanium content near the collector, from zero to about 10% at the CB junction, sets the strain in the thin Si/Ge base to the thick-film–determined value. Estimates of the consequent reduction in the band gap E_g are between 30 and 75 meV. As illustrated in Sec. 3.4, this would result in an increase of n_i^2 of between 3 and 20. The high doping concentration at the base also produces a band-gap lowering, and a consequent additional increase in n_i^2, offsetting at least partially the decrease normally associated with high doping concentrations (Klaassen *et al.,* 1992).
2. Since the base layer is very thin, the recombination of base current in the base itself should be very small, leaving the emitter recombination rate to control the gain of the transistor, as indicated in the denominator of Eq. (41). The thin base, however, requires compensatingly high doping to maintain a reasonably high value of the punch-through collector voltage. The high doping level at the emitter interface presents a low resistance for the supply of base–emitter injection current, and with the low diffusion capacitance of the emitter, a low *RC* time constant.

4.3 Polysilicon Emitter

Perhaps the most interesting result of this new technology is the unique interface properties and dramatic current gain increase associated with the "Ohmic" polycrystalline silicon contact to emitter, which becomes a part of the emitter itself. In spite of many publications, there remains some doubt about the role played by the polysilicon, but no doubt about the factor of 7 or larger gain that results. It has been demonstrated that the high arsenic doping of the polysilicon diffuses out into the epitaxial, approximately 10-nm "cap" during the RTA process that activates the arsenic. Heavy doping of the emitter and consequent low n_i^2 guarantee low emitter charge storage and diffusion capacitance, but do not alone guarantee the observed low diffusion injection into the emitter, Eq. (38), which has been supposed to account for increased current gain (Patton *et al.,* 1988; Eltoukhy and Roulston, 1982; Ashburn and Soerowirdjo, 1981; Graul *et al.,* 1976).

In the manufacturing environment, the epitaxial growth step takes place at relatively low temperatures, which is less than optimum for defect-free epitaxy. Therefore, this technology, which is also much more costly and has fewer transistors per functional unit than microprocessors or memory, seems to be best suited for just the technology best suited for it.

Works Cited

Abstreiter, G., Brugger, H., Wolf, T., Jorke, H., Herzog, H.-J. (1985), *Phys. Rev. Lett.* **47,** 1333.

Ashburn, P., Soerowirdjo, B. (1981), *Solid-State Electron.* **24,** 475–476.

Bardeen, J., Brattain, W. H. (1948), *Phys. Rev.* **74,** 230–233.

Braustein, R., Moore, A. R., Herman, F. (1958), *Phys. Rev.* **109,** 695–710.

Crabbé, E. F., Comfort, J. H., Cressler, J. D., Sun, J. Y.-C., Stork, J. M. C. (1993), *IEEE Electron. Dev. Lett.* **14,** 478–480.

Eltoukhy, A. A., Roulston, D. J. (1982), *IEEE Trans. Electron. Dev.* **ED-29,** 1862–1869.

Gärtner, W. W. (1960), *Transistors: Principles, Design, and Applications,* Princeton, NJ: Van Nostrand.

Graul, J., Glasl, A., Murrmann, H. (1976), *IEEE J. Solid-State Circuits* **11,** 491–495.

Harame, D. L., Comfort, J. H., Cressler, J. D., Crabbe, E. F., Sun, Y.-C., Meyerson, B. S., Tice, T. (1995), *IEEE Trans. Electron. Dev.* **40,** 455–468.

Kermarrec, C., Tewksbury, T., Dawe, G., Baines, R., Meyerson, B., Harame, D., Gilbert, M. (1994), in: *Proceedings of the 1994 Bipolar/BICMOS Circuits & Technology Meeting, October 10–11,* *1994, Minneapolis, MN,* Piscataway, NJ: IEEE, pp. 155–163.

Kittel, C. (1996), *Introduction to Solid State Physics,* 7th ed., New York: Wiley.

Klaassen, D. B. M., Slotbloom, J. W., De Graaff, H. C. (1992), *Solid-State Electron.* **35,** 125–129.

Kroemer, H. (1957), *Proc. IRE* **45,** 1535–1537.

Kroemer, H. (1982), *Proc. IEEE* **70,** 13–25.

Kroemer, H. (1985), *Solid-State Electronics* **28,** 1101–1103.

Lang, D. V., People, R., Bean, J. C., Sergant, A. M. (1985), *Appl. Phys. Lett.* **47,** 1333–1335.

Li, G. P., Ning, T. H., Chuang, C. T., Ketchen, M. B., Tang, D. D., Maur, H. (1987), *IEEE Trans. Electron. Dev.* **34,** 2246–2254.

Lindmayer, J., Wrigley, C. Y. (1965), *Fundamentals of Semiconductor Devices,* Princeton, NJ: Van Nostrand.

Muller, R. S., Kamins, T. I. (1986), *Device Electronics for Integrated Circuits,* 2nd ed. New York: Wiley.

Patton, G. L., Iyer, S. S., Delage, S. L., Tiwari, S., Stork, J. M. C. (1988), *IEEE Electron. Dev. Lett.* **9,** 1988.

Patton, G. L. *et al.* (1990), in: *Proceedings of the IEEE Symposium on VLSI Technology,* Piscataway, NJ: IEEE, pp. 90–91.

People, R. (1986), *IEEE J. Quantum Electron.* **22,** 1696–1710.

Shive, J. N. (1959), *Semiconductor Devices,* New York: Van Nostrand.

Shockley, W. (1939), *Bell Syst. Tech. J.* **18,** 645–723.

Shockley, W. (1949), *Bell Syst. Tech. J.* **28,** 435–489.

Shockley, W. (1950), *Electrons and Holes in Semiconductors,* New York: Van Nostrand.

Streetman, B. G. (1995), *Solid State Electronic Devices,* Englewood Cliffs, NJ: Prentice-Hall.

Warner, R. M., Grung, B. L. (1983), *Transistors for the Integrated Circuit Engineer,* New York: Wiley Interscience.

Western Electric Co. (1958), *Transistor Technology,* Vols. 1–3, Princeton, NJ: Van Nostrand.

Wolf, S. (1986, 1990, 1993), *Silicon Processing for the ULSI Era,* Vols. 1–3, Sunset Beach, CA: Lattice Press.

Sze, S. M. (1981), *Physics of Semiconductor Devices,* 2nd ed., New York: Wiley.

Further Reading

The following works constitute an exhaustive resource for further exploration of items published before 1995.

Gärtner, W. W. (1960), *Transistors: Principles, Design, and Applications,* Princeton, NJ: Van Nostrand.

Harame, D. L., Comfort, J. H., Cressler, J. D., Crabbe, E. F., Sun, Y.-C., Meyerson, B. S., Tice, T. (1995), *IEEE Trans. Electron. Dev.* **40,** 455–468.

Kittel, C. (1996), *Introduction to Solid State Physics,* 7th ed., New York: Wiley.

Li, G. P., Ning, T. H., Chuang, C. T., Ketchen, M. B., Tang, D. D., Maur, H. (1987), *IEEE Trans. Electron. Dev.* **34,** 2246–2254.

Lindmayer, J., Wrigley, C. J. (1965), *Fundamentals of Semiconductor Devices,* Princeton, NJ: Van Nostrand.

Western Electric Co. (1958), *Transistor Technology,* Vols. 1–3, Princeton, NJ: Van Nostrand.

The following works provide treatment of developments since 1995.

IEEE International Electron Devices Meeting, Technical Digest, issued annually since 1982.

IEEE Transactions on Electron Devices, since 1992.

IEEE Electron Device Letters, since 1992.

Iwai, H., Selvakumar, C. R., Shenai, K. (Eds.) (1995), Special issue on Bipolar BiCMOS/CMOS Devices and Technologies, *IEEE Trans. Electron. Dev.* **42** (3).

TRANSISTORS, FIELD-EFFECT

JOHN FARICELLI, *Digital Equipment Corporation, Hudson, Massachusetts, U.S.A.*

	Introduction	147
1.	**Junction Field-Effect Transistor (JFET)**	148
2.	**Metal-Semiconductor Field-Effect Transistor (MESFET)**	150
3.	**Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET)**	152
3.1	Introduction	152
3.2	Basic MOSFET Operation	152
3.3	Overview of the MOSFET Fabrication Process	153
3.4	Physical Model of a MOSFET	154
3.4.1	MOS Capacitor	154
3.4.2	Threshold Voltage	155
3.4.3	Linear Bias Region	156
3.4.4	Saturation Bias Region	157
3.4.5	Subthreshold Bias Region	158
3.4.6	Breakdown Phenomena	158
3.4.7	Temperature Effects	159
3.5	Small-Device ("Short-Channel") Effects	159
3.5.1	Threshold Voltage	159
3.5.2	Velocity Saturation	160
3.5.3	Hot Carriers and Reliability	160
3.6	MOSFET Variants	161
3.6.1	Silicon-on-Insulator (SOI) MOSFET	161
3.6.2	Nonvolatile Memory	162
4.	**Advantages and Disadvantages of FETs**	162
	Glossary	163
	Works Cited	164
	Further Reading	164

INTRODUCTION

The field-effect transistor (FET) is an electronic device for modifying the flow of current in a material across two terminals, under the control of an electric field induced by a voltage applied at a third terminal. The use of the field effect in an electronic device was first described in the 1930s (Lilienfeld, 1930; Heil, 1935). Shockley (1952) published the first analysis of a practical electronic device based on the field effect. Figure 1 shows a diagram of the basic concept behind the field-effect transistor. Two terminals are attached to a block of conducting material (which in all further discussion we will assume to be semiconducting). Current flow is induced in the material by applying a voltage across the terminals, which are identified as the source and drain. Current flow from source to drain is altered by an applied transverse electric field from a third terminal, called the gate. The gate, or field plate, is separated from the material in which the current flows by a poorly conducting or nonconductive region. Thus, the input resistance of the device is high, and the device is considered to be voltage controlled, as opposed to current controlled as in a bipolar transistor. As the voltage on the gate is increased, the transverse field increases, and the current flow is modified. Depending on the polarity of the applied field, the current flow may be enhanced or reduced. Practical application of the field-effect transistor had to wait until the growth and processing of semiconductors was sufficiently reliable and clean so that the applied electric field was not overcome by other effects, such as uncontrolled charges along the surface of the semiconductor region. Engineering control of semiconductor devices has grown tremendously since the original concept of the field-effect transistor, such that as of the writing of this article (mid 1990s), it is commonplace to fabricate integrated circuits containing several million FETs on a single integrated circuit, with minimum feature sizes in the range of 0.50–0.35 μm.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

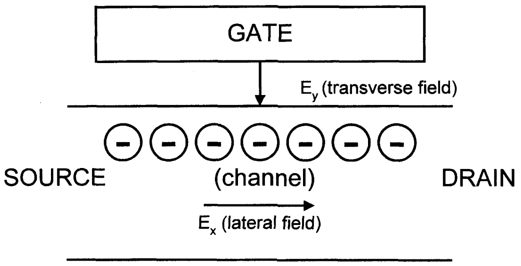

FIG. 1. Cross section of a generic field-effect transistor.

1. JUNCTION FIELD-EFFECT TRANSISTOR (JFET)

The junction field-effect transistor (JFET) was realized early on (Dacey and Ross, 1953), since it was relatively easy to fabricate, but is now only used in specialized applications. Figure 2 shows the IEEE standard circuit schematic symbols for a JFET (IEEE, 1975). Figure 3 shows a schematic cross section of a JFET. The body consists of lightly doped *n*-type or *p*-type semiconductor, typically with a dopant concentration on the order of 10^{17} atoms/cm^3. A JFET with an *n*-type body is denoted an *n*-channel device, while a JFET with a *p*-type body is denoted a *p*-channel device. Highly doped regions of the same doping type as the body are then created, typically by solid-state diffusion of dopants at high temperatures, to form the source and drain contacts. Highly doped regions are required at the contacts to ensure that an Ohmic (nonrectifying) contact is formed between the semiconductor and any subsequent metal used for supplying voltage to the device. A *p-n* junction is formed at the surface of the semiconductor, between the source and drain. In consequence of a balance of diffusion of free carriers and electrostatic attraction between the carriers and their ion cores, a region relatively free of carriers (the so-called "depletion region") is formed under the *p-n* junction. The extent of the depletion region can be modified by applying a bias voltage to the *p-n* junction. For an *n*-channel device, with a *p*-type gate region, negative (reverse) bias is applied to increase the depletion region under the junction. For *p*-channel devices, the opposite sign bias is applied.

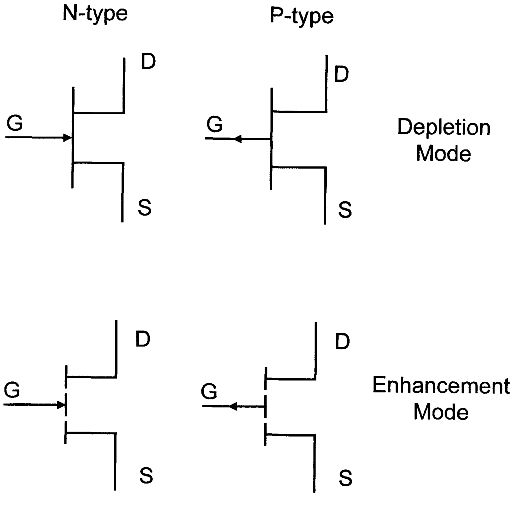

FIG. 2. IEEE standard schematic symbols for JFETs.

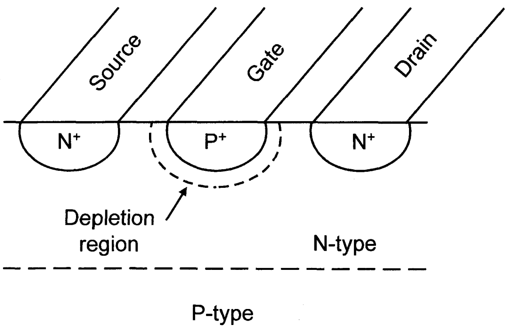

FIG. 3. Cross section of a junction field-effect transistor.

Consider the case where a very small voltage is applied to the drain, so that the potential from source to drain is essentially uniform [Fig. 4(a)]. For pedagogical reasons, we consider the device to be constructed so that there is a symmetry plane parallel to the channel. The device is assumed to be uniform in the third (width) dimension. We assume that the potential reference is at the source contact; thus all voltages are referred to the source, e.g., the gate-to-source voltage V_{gs}. As the reverse bias is increased on the gate, the current flow between the source and drain is constricted, and the resistance of the device increases. At some point, enough bias is applied so that the channel is completely constricted, and very little current flows between the source and drain. Un-

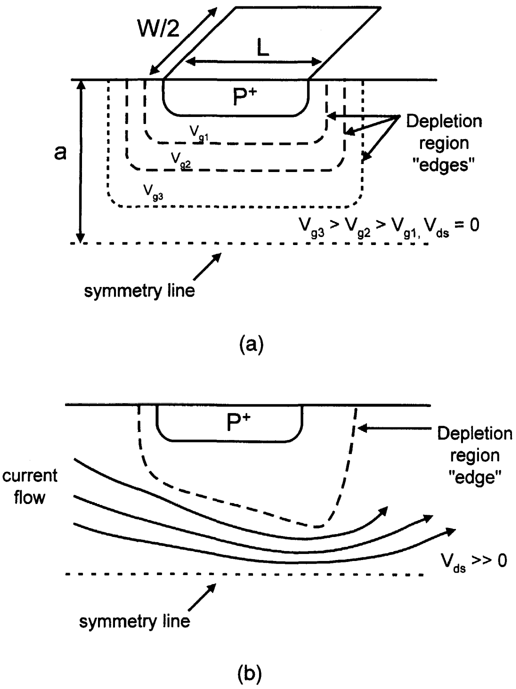

FIG. 4. Cross section of a JFET showing the depletion region under small drain bias (linear region) and under high drain bias (saturation region).

der these conditions, the device acts like a voltage-controlled variable resistance. This is the *linear* region of device operation, so named because the device acts like a linear resistance under these bias conditions. The voltage at which the device passes little current is called the *pinch-off voltage,* V_p. Furthermore, since this device operates by *depleting* the channel of carriers, it is called a depletion-mode FET. The equation for the drain current in the linear region of an n-channel depletion mode JFET is given by (Sze, 1981)

$$I_{ds} = I_p[3V_{ds}/V_p - 22[(V_{ds} + V_{gs} + V_{bi})^{3/2} - (V_{gs} + V_{bi})^{3/2}]/V_p^{3/2}] \quad (1)$$

The built-in potential of the junction V_{bi} is given by

$$V_{bi} = (kT/q)\ln(N_d/n_i), \quad (2)$$

where k is Boltzmann's constant, T is the absolute temperature, q is the electronic charge, N_d is the doping concentration in the channel, and n_i is the concentration of carriers in undoped semiconductor (intrinsic concentration). The pinch-off voltage V_p is given by

$$V_p = qN_d a^2/2\epsilon_s, \quad (3)$$

where a is the thickness of the JFET channel and ϵ_s is the permittivity of the semiconductor. The pinch-off current I_p is given by

$$I_p = W\mu q^2 N_d^2 a^3/6\epsilon_s L, \quad (4)$$

where W is the width of the device, μ is the carrier mobility, and L is the channel length.

Now consider the case where bias is applied to the drain terminal, with the gate at a zero bias condition [Fig. 4(b)]. As the drain bias increases, the current flow through the device increases, since the device is in the linear operating region. The drain voltage is dropped across the channel region. As the drain voltage increases, the gate-to-channel potential near the drain side of the p-n junction becomes increasingly reverse biased. This, in turn, causes the depletion region under the p-n junction to become larger near the drain end of the device. The effective channel resistance now increases, because the amount of nondepleted semiconductor in the channel is less. Eventually, the drain bias becomes so large that the depletion region from the p-n junction "touches" the depletion region from its mirror image. At this point, the incremental resistance of the channel to further increases in drain voltage is very high, and the so-called "pinch-off" voltage is reached. In reality, the channel never can pinch off entirely, as then no current would be able to flow in the device, and there would be no current to induce the voltage drop along the channel of the FET. Instead, the carrier density drops to a low value, and the drain current tends to remain constant as further drain bias is applied. This region of operation is thus called the *saturation region.* The equation for the drain current in the saturation region is given by (Sze, 1981)

$$I_{ds,sat} = I_p\left[1 - 3\left(\frac{V_{gs} + V_{bi}}{V_p}\right) + 2\left(\frac{V_{gs} + V_{bi}}{V_p}\right)^{3/2}\right]. \quad (5)$$

In the derivation of this equation, it was assumed that the carrier mobility is a constant independent of dopant atom density or electric field and that the potential drops gradually along the channel so that the Poisson's equation,

$$\frac{dE_x}{dx} + \frac{dE_y}{dy} = \frac{\rho}{\epsilon_s} \qquad (6)$$

(where E is the electric field and ρ is the space charge), can be simplified by dropping the term due to dE_x/dx. These assumptions will be revisited later.

JFETs were among the first FETs to be manufactured because they could be built using semiconductor processing steps that were originally developed for bipolar transistors. The modern use of JFETs is limited, since other types of FETs, notably the MOSFET, are superior in most respects.

2. METAL-SEMICONDUCTOR FIELD-EFFECT TRANSISTOR (MESFET)

The metal-semiconductor field-effect transistor (MESFET) is similar to the JFET, with the exception that the *p-n* control junction is replaced by a rectifying metal-semiconductor (Schottky) contact. It was first proposed by Mead (1966). It is typically used for fabricating FETs in semiconductors other than silicon, such as GaAs or InP. JFETs are not easily fabricated in compound semiconductors, because *p-n* junctions are difficult to form because of lack of suitable *p*-type dopant atom species, while metal-oxide-semiconductor FETs (MOSFETs) are difficult to build because of the lack of a suitable native oxide to insulate the gate from the semiconductor. Silicon MESFETs can also be manufactured, although they are seldom used commercially. Since the Schottky-barrier junction can only be operated in reverse bias without causing current to flow into the gate, MESFETs are typically operated as depletion-mode devices.

A cross section of a MESFET is shown in Fig. 5. Most MESFETs built in compound semiconductors start with a semi-insulating substrate. A thin layer of *n*-type semiconductor is grown, using epitaxy techniques (such as liquid-phase epitaxy or molecular-beam epitaxy). The gate metal is deposited and then etched to pattern the gate region. Source and drain junctions are formed by bombarding the semiconductor with accelerated *n*-type dopant ions (ion implantation). The channel region must be prevented from having source and drain dopants introduced; in addition, some space must be maintained between the gate and the source and drain regions. This can be done, for example, with a masking material over the gate. A short, high-temperature annealing (e.g., 700–800 °C) is used to diffuse the dopant atoms onto substitutional sites in the semiconductor lattice, thus allowing them to become electrically active. Since the gate dimensions of MESFETs are commonly well below 1 μm, lithographic techniques that are highly controllable, such as electron-beam lithography, are often used. Finally, Ohmic contacts to the source and drain regions are formed, usually by depositing a series of thin metal layers.

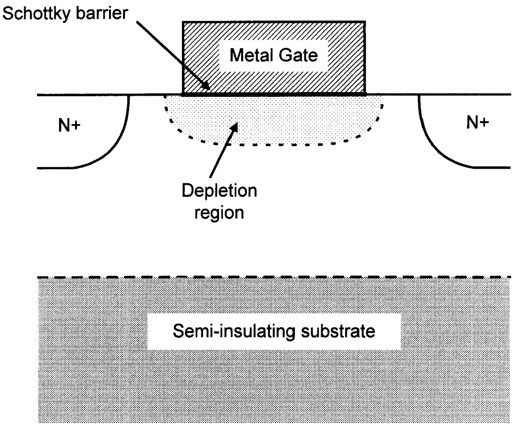

FIG. 5. Cross section of a metal-semiconductor FET (MESFET).

The physical operation of MESFETs is essentially similar to the JFET, with the built-in voltage V_{bi} of the JFET *p-n* junction replaced by the Schottky-barrier potential. Two important physical considerations must be added to the simple model presented in the earlier section: velocity saturation and nonequilibrium transport effects. In the previous formulas for the drain current, the carrier mobility was assumed to be constant. In reality, the mobility of electrons and holes in semiconductors is dependent on both the density of dopant atoms and the electric field. In the channel, the dopant density is relatively uniform. Although extra dopants may be incorporated to modify the pinch-off voltage, the channel dopant density is approximately uniform to an order of magnitude. Thus, the mobility can be adjusted by using various formulas fitted to experimental measurements, such as Arora *et al.* (1982). The effect of electric field along the direction of the current flow, however, cannot be ignored. At gate lengths typical of practice (below 1 μm), the electric field along the channel direction induced by the drain voltage becomes sufficiently high that the carriers can achieve high enough energy to emit optical phonons, which effectively drain off excess energy to the underlying lattice (Jacoboni *et al.,* 1977). This causes the carrier velocity to saturate at approximately 10^7 cm/s. For GaAs, velocity saturation can occur at fields as low as 15–20 kV/cm. Figure 6 plots the carrier velocity vs lateral electric field for electrons and holes in silicon, and for electrons in GaAs. The derivative of these curves is the mobility. The nonmonotonic behavior of the electron velocity in GaAs is due to the complicated nature of the electronic-band structure of this compound semiconductor, where electrons can occupy states with different effective masses, depending on their energy. Velocity saturation reduces the amount of drain current, compared to what is predicted by Eq. (5). The drain current is now increased linearly with gate voltage, as opposed to the square-law dependence of Eq. (5).

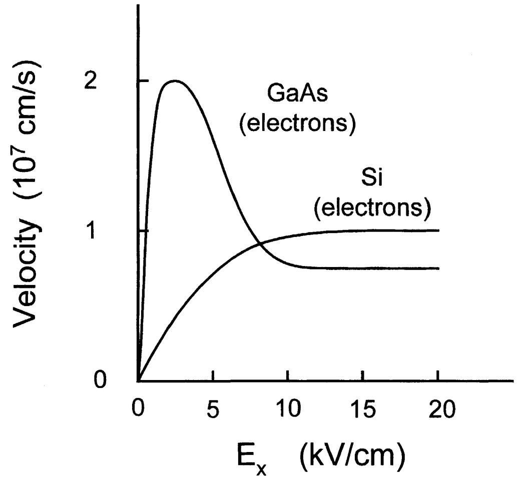

FIG. 6. Carrier velocity vs lateral electric field for electrons and holes in silicon, and electrons in GaAs (after *Physics of Semiconductor Devices,* 2nd ed., S. M. Sze, Copyright © 1981, John Wiley & Sons, Inc. Reprinted with permission of John Wiley & Sons, Inc.).

At very short channel lengths, the electric field at the drain end of the channel becomes sharply peaked, and the electrons may not have time (or distance) to achieve equilibrium between the energy gain from the electric field and energy-loss mechanisms, such as optical-phonon emission. This nonequilibrium state can allow the carrier velocity to exceed temporarily the equilibrium saturation velocity ("velocity overshoot"). Such effects are difficult to quantify in analytical expressions, and sophisticated numerical analysis programs are needed to predict device behavior under such conditions.

Common applications for MESFETs are in high-frequency (microwave) communication electronics and in high-speed digital computation elements. MESFETs are highly competitive for communication applications, because the higher mobility of compound semiconductors compared to silicon allows them to pass more current for a given gate length. Higher mobility also reduces the resistance of the source and drain areas, thus reducing thermal noise, an important factor for low-noise amplifiers. The semi-insulating substrate also provides less parasitic capacitance, further improving performance. Compound semiconductors are also compatible with optoelectronic devices. Multiple MESFETs can be joined in parallel to form power microwave devices. In digital-circuit applications, MESFETs are used in very high-speed computational circuits, although their density and cost are not yet competitive with silicon MOSFETs. Drawbacks to using compound semiconductors include lower thermal conductivity (poorer dissipation of heat), lack of a complementary device (only *n*-channel devices), higher costs for starting materials, less mature manufacturing, and lower levels of integration.

A variant of the MESFET suitable for fast switching or communication applications is the high–electron-mobility transistor (HEMT), in which a superlattice structure is formed by layers of different compound semiconductors. The channel of the device is formed of a lightly doped region, sandwiched between regions of different materials that are highly doped. The material properties of the sandwich are such that electrons in the highly doped regions "fall" into the lightly doped channel. This *real-space transfer* of carriers results in a channel that is conductive, yet with few ion cores to induce scattering from ionized impurities. This results in a device with high carrier mobility, at least at low transverse fields.

3. METAL-OXIDE-SEMICONDUCTOR FIELD-EFFECT TRANSISTOR (MOSFET)

3.1 Introduction

The metal-oxide-semiconductor field-effect transistor (MOSFET) is the workhorse device of modern microelectronics. Microprocessors using 5–10 million MOSFETs, operating at frequencies of several hundred megahertz, and with costs in the range of a few hundred to a few thousand U.S. dollars, are typical for the mid-1990s. The enormous growth in personal computing, telecommunications, consumer electronics, and other areas, most using integrated circuits containing MOSFETs as their main active element, has led the global semiconductor industry to invest billions of dollars in semiconductor fabrication facilities around the world.

3.2 Basic MOSFET Operation

MOSFETs come in two varieties, depending on whether electrons or holes are used to conduct the current between the source and drain. When electrons are used, the MOSFET is called an n-channel device; when holes are used, it is called a p-channel device. n-channel devices are divided into two types: enhancement mode, or normally off [Fig. 7(a)], and depletion mode, or normally on [Fig. 7(b)]. The source and drain of an n-channel MOSFET consist of heavily doped (n^+) areas, separated by the channel region. The gate, which is typically made of polysilicon, is insulated from the channel by a thin layer of silicon dioxide. The gate material is very highly doped, so that it acts essentially like a metal.

In an enhancement-mode device in which no voltage is applied to the gate electrode, there is no path for electrons to flow from the source to the drain; hence the device is in the "off" state. When a positive voltage is applied to the gate, electrons are attracted to the channel region from the pool of electrons in the source and drain regions. When a sufficiently positive voltage is applied, electrons form a path for current to flow from source to drain. This "enhanced" conductivity region (hence "enhancement" mode) is called the *inversion layer,* since the field effect has caused the conductivity of the channel region to change from p-type to n-type. The inversion layer is exceedingly thin, on the order of 1–10 nm.

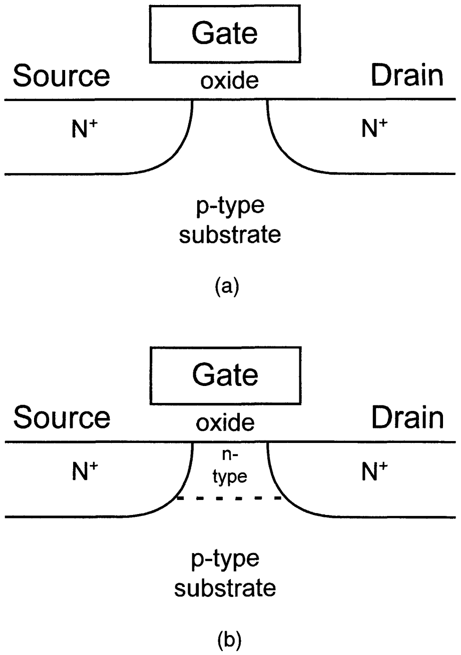

FIG. 7. Cross-section of n-channel MOSFETs. (a) enhancement mode, (b) depletion mode.

In a depletion-mode device, the channel region consists of lightly doped n-type material, on the order of 10^{17} dopant atoms/cm^3. Since there is a conductive path between the source and drain, this device is normally on, even when no voltage is applied to the gate. When negative voltage is applied to the gate, electrons are repelled from the channel, and a region depleted of electrons forms under the gate (hence "depletion mode"). With sufficiently large negative bias on the gate, all the electrons will be depleted from the channel, and the conductive path between the source and drain is disrupted. The device is now turned "off."

Figure 8 shows a cross section of both enhancement- and depletion-mode *p*-channel devices. The operation of *p*-channel devices is exactly the same as an *n*-channel device, except that the opposite sign voltages must be applied to the gate to attract or repel holes. In addition, the voltage applied to the drain is also negative, in order to get holes to flow from source to drain.

Figure 9 shows the IEEE standard circuit-schematic symbols for all four types of MOSFETs. Note that the MOSFET is a four-terminal device. Besides the source, drain, and gate, there is a terminal that allows direct electrical contact to the channel region. This is denoted the bulk or substrate contact. In analyzing MOSFET operation, the source contact is usually considered to be the potential reference; hence the applied voltages are gate to source (V_{gs}), drain to source (V_{ds}), and bulk to source (V_{bs}).

3.3 Overview of the MOSFET Fabrication Process

Figure 10 shows a simplified sequence of steps used to create a MOSFET. The actual fabrication sequence for a modern IC manu-

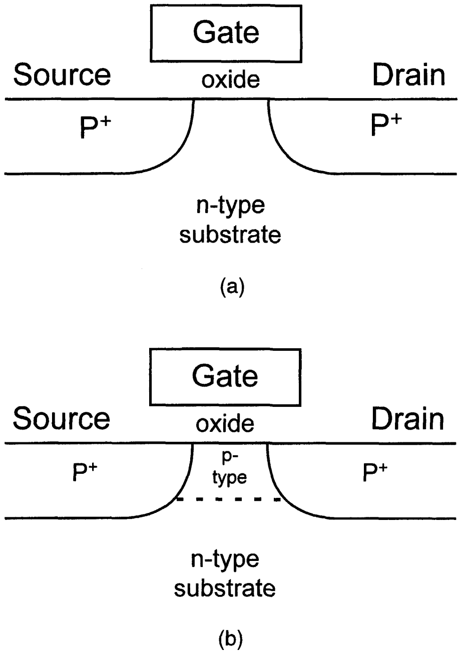
Gate
Source
oxide
Drain
P+
P+
n-type substrate
(a)
Gate
Source
oxide
Drain
P+
p-type
P+
n-type substrate
(b)

FIG. 8. Cross section of *p*-channel MOSFETs. (a) enhancement mode, (b) depletion mode.

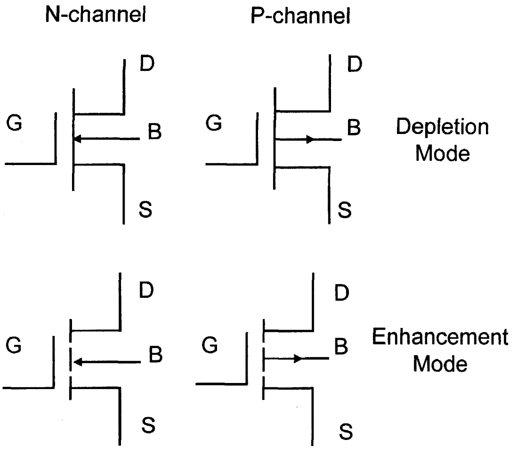
N-channel
P-channel
D
G
B
S
D
G
B
S
Depletion Mode
D
G
B
S
D
G
B
S
Enhancement Mode

FIG. 9. IEEE standard schematic symbols for MOSFETs.

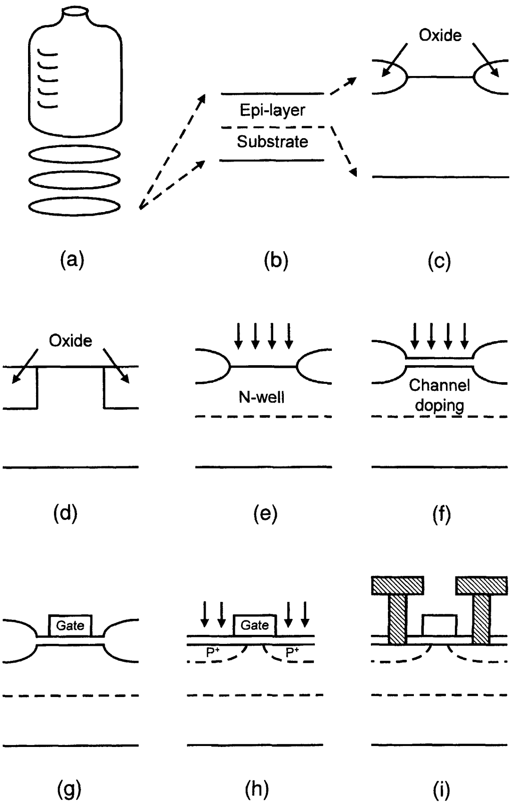
Oxide
Epi-layer
Substrate
(a)
(b)
(c)
Oxide
N-well
Channel doping
(d)
(e)
(f)
Gate
Gate
P+
P+
(g)
(h)
(i)

FIG. 10. Manufacturing sequence for fabricating an *p*-channel MOSFET.

facturing process with both n- and p-channel devices on the same chip can consist of hundreds of discrete steps. Integrated-circuit manufacturers typically purchase silicon wafers from a vendor who refines the silicon from raw materials into large-diameter (e.g., 200 mm or larger) single-crystal boules [10(a)]. The boules are then sliced into thin wafers and are typically p type. The wafers may be grown lightly doped, or they may be grown highly doped, with a thin layer of lightly doped silicon grown on top, via epitaxial growth [10(b)]. The wafers are then processed to define regions where the MOSFETs will be formed. In between these regions, the MOSFETs must be electrically isolated from one another. This is typically done by creating regions of silicon dioxide between MOSFETs, either by treating selected regions of the silicon to an oxidizing atmosphere in a furnace (LOCal Oxidation of Silicon, or LOCOS) [10(c)] or by etching a trench in the silicon substrate and depositing oxide in the trench [10(d)]. If both n- and p-channel devices are to be formed on the same wafer, a region of n-type doping ("well" or "tub") is formed to hold the p-channel devices [10(e)]. In some processes, both n- and p-type wells are formed. Dopant atoms are introduced into the silicon wafer by the ion-implantation process, in which an electrostatic field is used to accelerate ionized dopant atoms, which are then embedded into the silicon wafer. By integrating the charge collected by the wafer, the number of dopant atoms per unit area (dose) can be monitored to high precision. The doping in the channel regions is adjusted to create enhancement- or depletion-mode transistors [10(f)]. The polysilicon gate material is then deposited and etched [10(g)]. The source and drain are then formed, again by ion implantation [10(h)]. The polysilicon gate acts as a masking layer, preventing the channel from being doped during source and drain formation. This *self-aligned* process is key to scaling MOSFET devices to submicron dimensions, since the source and drain are automatically aligned to the gate edges. In many processes, a refractory metal, such as Ti or Co, is deposited over the source, drain, and gate areas, and a metal-silicon compound is formed (silicide). This is done to reduce the resistance of these regions. Finally, a series of metal and silicon dioxide layers are deposited to allow connections to the MOSFET and to connect MOSFETs together [10(i)].

3.4 Physical Model of a MOSFET

3.4.1 MOS Capacitor Much of the basic physics in the MOSFET is contained in a simpler device, the MOS capacitor. Consider a region of lightly doped p-type silicon, covered by a thin layer of silicon dioxide, and capped with a layer of metal. The silicon region is connected to ground potential. For pedagogical purposes, consider the difference in work function between the metal and the semiconductor to be zero (the work function is the voltage necessary to raise an electron from the material's Fermi level to the vacuum level). Since the metal is electrically isolated from the silicon, a capacitor is formed. Any applied voltage on the metal induces a surface charge on the metal, which must be matched by an opposite charge on the silicon side. Figure 11 shows the energy-band diagram of the metal-oxide-silicon sandwich under various applied biases to the metal. With zero applied bias, there is no distortion of the energy bands, and the MOS capacitor is in the so-called *flat-band* condition. When a negative voltage is applied to the metal, the metal's energy bands are raised upward (energy $= -qV$). Part of the applied voltage drops across the oxide region, and the rest drops across the silicon region [Fig. 11(a)]. Holes, supplied from the large pool in the silicon bulk, are attracted to the negative potential and accumulate at the silicon/silicon dioxide interface. Thus, the MOS diode is said to be in *accumulation* bias. When positive voltage is applied to the metal, the energy bands are bent down [Fig. 11(b)]. The positive voltage repels holes from the silicon/silicon dioxide interface, and the holes are absorbed at the ground contact. The MOS capacitor is said to be in *depletion* bias, since the silicon region near the interface is depleted of holes. At higher positive voltage [Fig. 11(c)], electrons generated by random thermal excitation in the silicon bulk are attracted to the interface. At this point, the MOS capacitor is in *inversion* bias.

An analogous situation occurs in n-type MOS capacitors, with positive bias inducing accumulation of electrons and negative bias

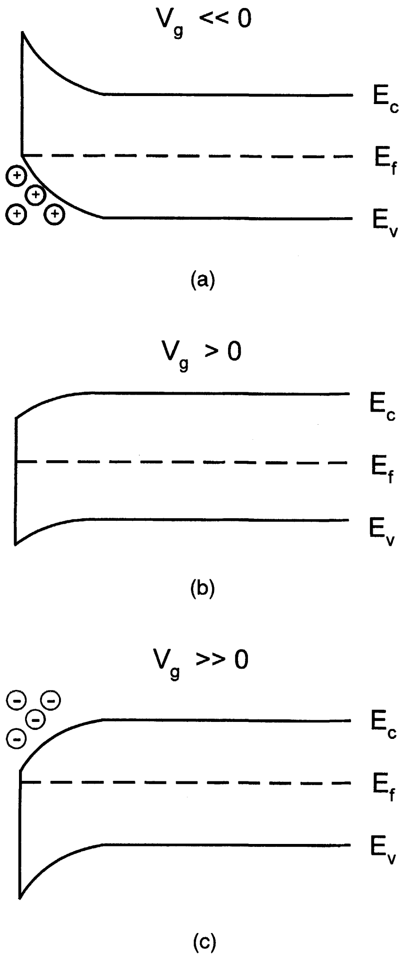

FIG. 11. Energy-band diagrams for a MOS capacitor under (a) accumulation, (b) depletion, (c) inversion bias.

inducing depletion of electrons and finally hole inversion.

3.4.2 Threshold Voltage One can express the change in electron and hole density with applied bias quantitatively by introducing the formulas for the electron and hole concentration as a function of the electrostatic potential ψ (defined as the potential at the midgap of the semiconductor and rigidly shifting with the edges of the semiconductor bands) (Sze, 1981):

$$n = n_0 \exp(q\psi/kT), \tag{7}$$

$$p = p_0 \exp(-q\psi/kT), \tag{8}$$

where n_0 and p_0 are the equilibrium electron and hole concentrations, q is the electronic charge, k is Boltzmann's constant, and T is absolute temperate. Again, consider a p-type MOS diode. As the metal, or gate, is positively biased, the hole concentration in the silicon decreases. This leaves behind the ionized dopant atoms, which serve to terminate field lines emanating from the gate. Since the p-type region is relatively lightly doped (e.g., 10^{16}–10^{17} cm^{-3}), a relatively large region of silicon must be depleted of charge to balance the surface charge induced on the gate metal. As the gate bias becomes more positive, the electron concentration becomes appreciable. At some point, the electron concentration becomes high enough to balance any additional charge induced on the metal gate. In fact, since the electron concentration increases exponentially, the potential at the silicon surface becomes *pinned* and need vary only slightly to provide all the charge needed to balance the induced charge on the metal gate. At this point, the MOS diode is said to cross its *threshold voltage,* in which the inversion channel is strongly formed. Unfortunately, there is no definitive definition for the threshold voltage. One common definition of threshold voltage is when the surface potential reaches twice the Fermi potential $\phi_b = q/kT \log(N_a/n_i)$, where N_a is the doping concentration in the silicon and n_i is the intrinsic carrier concentration. This leads to the following expression for the threshold voltage (Sze, 1981):

$$V_{th} = \sqrt{2q\epsilon_{Si}N_a(2\phi_b)}/C_i + 2\phi_b, \tag{9}$$

where V_{th} is the threshold voltage, ϵ_{Si} is the silicon permittivity, and C_i is the capacitance of the oxide layer. The first term represents the voltage dropped in the silicon bulk during the depletion phase, and the second term, the voltage dropped in establishing strong inversion. In practical MOSFETs, a third term is added due to differences in the work function between the gate material and the silicon bulk, and any fixed charges at the Si/SiO$_2$ interface.

When a MOS capacitor is transformed into a MOS transistor by the addition of a source and drain, the threshold voltage denotes the voltage at which the conducting channel forms. By varying the gate material and/or the doping concentration or type (p

or n), one can create MOSFETs with both positive and negative threshold voltages. This allows the formation of enhancement- or depletion-mode transistors.

3.4.3 Linear Bias Region We now consider the situation where source and drain contacts have been added to the MOS capacitor, so that a proper MOSFET has been formed. By convention, the source contact is considered the reference contact for all voltages and currents. In the following discussion, all equations are given for n-channel devices. The equations for p-channel devices are essentially the same as for n-channel, with the signs of the voltages changed. Figure 12(a) shows a cross section of an n-channel MOSFET under small drain bias. Using the gradual channel approximation discussed in Sec. 2 on MESFETs, an expression for the drain current, I_{ds}, in an n-channel MOSFET in the linear region ($V_{ds} \ll V_{gs} - V_{th}$) is given by (Sze, 1981)

$$I_{ds} = (W/L)\mu_n C_i(V_{gs} - V_{th})V_{ds}, \qquad (10)$$

where W is the gate width, L is the gate length, μ_n is the electron mobility, V_{gs} is the gate-to-source voltage, and V_{ds} is the drain-to-source voltage. The threshold voltage V_{th} is given by an expression similar to that given previously for the MOS capacitor, with the addition of a term that accounts for any bias applied to the substrate, V_{bs} (Sze, 1981):

$$V_{th} = 2\phi_b + \sqrt{2q\epsilon_{Si}N_a(2\phi_b + V_{bs})}/C_i. \qquad (11)$$

This bias region is called the "linear" region because the drain current I_{ds} is linearly proportional to the drain voltage V_{ds}. In this bias regime, the MOSFET acts like a voltage-controlled resistance. Figure 13 shows a plot of I_{ds} vs V_{ds}, with V_{gs} as a parameter. Note that for low V_{ds}, the drain current is linearly proportional to V_{ds}.

It should be noted that the mobility used in the expression for I_{ds} is not the bulk mobility of electrons (or holes) in silicon, but a mobility that accounts for additional scattering due to the proximity of the Si/SiO$_2$ interface and two-dimensional gas effects due to the quantum well formed by the sharp band bending near the interface. This inversion-layer mobility is a function of the applied

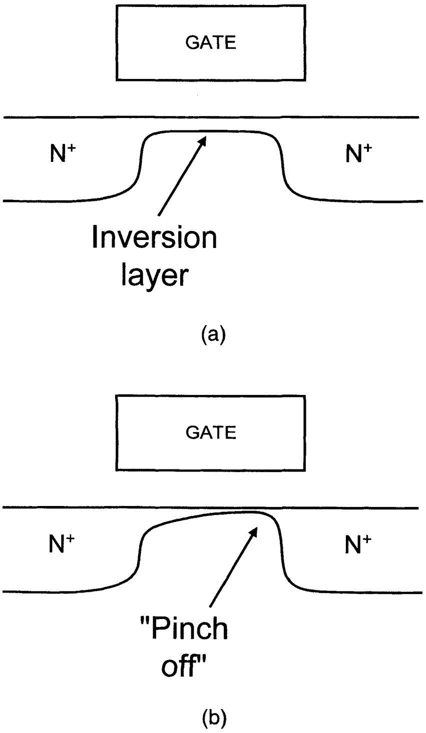

FIG. 12. Cross section of an n-channel enhancement-mode MOSFET under small drain bias (linear region) and high drain bias (saturation region).

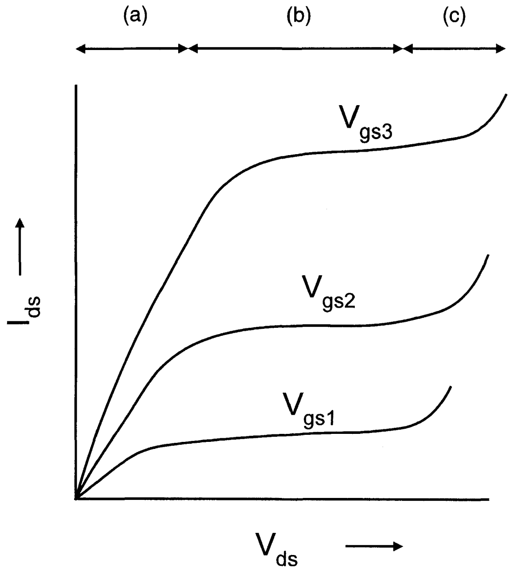

FIG. 13. Drain-current vs drain-voltage characteristics for an n-channel enhancement-mode MOSFET, with gate voltage as a parameter. (a) Linear region, (b) saturation region, (c) breakdown.

transverse field in the silicon; hence the mobility is a function of the gate voltage. If the measured mobility is plotted against the effective transverse field (defined as the electric field averaged over the electron distribution in the inversion layer), one can discern a so-called "universal" mobility curve, which is largely independent of channel doping concentration, as long as the effective field is sufficiently large enough to form a strongly inverted channel. Figure 14 shows a plot of the inversion mobility vs effective field (Sabinis and Clements, 1979).

3.4.4 Saturation Bias Region As the drain voltage is increased, the voltage drop along the MOSFET channel length can no longer be neglected. The gate-to-channel potential, which gives rise to the inversion layer, decreases as one approaches the drain, as shown in Fig. 12(b). This leads to a flattening of the I_{ds} characteristic, as given by the following expression for long n-channel devices (Sze, 1981):

$$I_{ds} = (W/L)\mu_n C_i\left\{\left(V_{gs} - 2\phi_b - \frac{V_{ds}}{2}\right)V_{ds} - \frac{2}{3} \times \frac{\sqrt{2q\epsilon_{Si}N_a}}{C_i}\left[(V_{ds} + 2\phi_b)^{3/2} - (2\phi_b)^{3/2}\right]\right\}. \quad (12)$$

As the drain voltage increases further, the gate-to-channel potential at some point near the drain is reduced sufficiently that an inversion channel can no longer be maintained. This is the so-called "pinch-off" point, similar to that in MESFETs. The electron concentration in the pinch-off region is greatly reduced compared to the inversion layer but is sufficient to conduct the drain current. At this point, further increases in drain voltage have only a small effect on I_{ds}. The expression for the drain current in saturation is (Sze, 1981)

$$I_{ds,sat} = (W/L)m\mu_n C_i(V_{gs} - V_{th})^2, \quad (13)$$

where m is a function of doping, with a value on the order of $\frac{1}{2}$. Note that the drain current in saturation varies with the square of the gate drive $V_{gs} - V_{th}$. As is discussed later, in modern devices with submicron gate lengths, velocity saturation reduces this to a linear dependence on gate drive. Note that in practice, the drain current does increase slightly as the drain bias increases, as shown in Fig. 13. In long-channel devices, this can be attributed to a small increase in the pinch-off region as the drain bias increases ("channel length modulation").

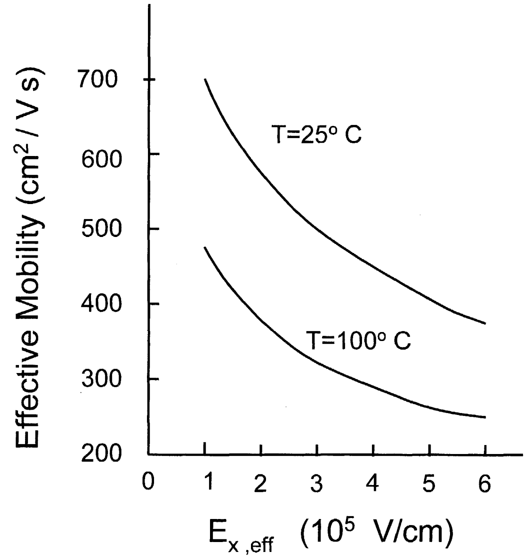

FIG. 14. Effective electron inversion layer mobility vs effective transverse field (Sabinis and Clements, 1979) (after *Physics of Semiconductor Devices*, 2nd ed., S. M. Sze, Copyright © 1981, John Wiley & Sons, Inc. Reprinted with permission of John Wiley & Sons, Inc.).

The saturation region is an important bias region for both analog and digital circuits that use MOSFETs. In analog amplification circuits, the MOSFET is biased in saturation, with the signal to be amplified placed on the gate. The amount of amplification, or *transconductance*, is given by differentiating the expression for I_{ds} with respect to V_{gs}:

$$g_m = 2m(W/L)\mu_n C_i(V_{gs} - V_{th}). \quad (14)$$

Again, note that for modern devices with velocity saturation, the transconductance will be almost independent of the gate drive.

In digital applications, the MOSFET is typically paired in series with another MOSFET, either an n-type depletion mode device, to form NMOS logic, or a p-type inversion-mode device, to form complementary MOS logic (CMOS); see Fig. 15. These transistor pairs form an inverter, in which a high bias

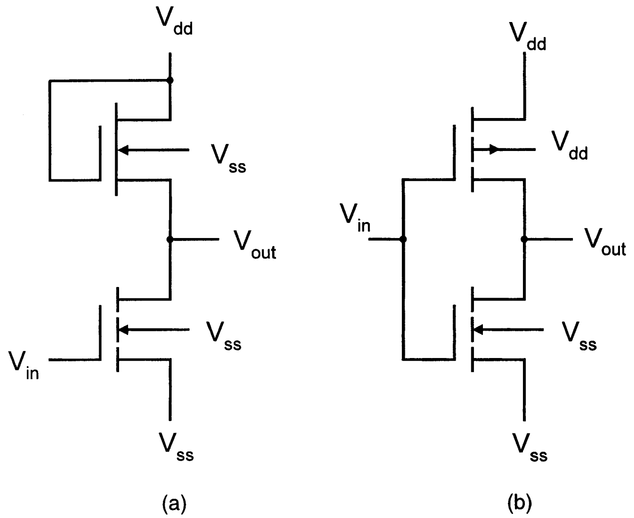

FIG. 15. MOS transistors configured as digital logic inverter gates. (a) NMOS logic (*n*-channel enhancement mode plus *n*-channel depletion mode), (b) CMOS logic (*n*-channel enhancement mode plus *p*-channel enhancement mode).

(logic "1") applied to the input of the inverter (gate) causes the output (drain) to switch to a low bias (logic "0"). The output of the inverter typically feeds other circuitry or interconnect lines, which are capacitive loads, and must be charged up or down. The amount of current available during the switching cycle for this charging is determined by the saturation-region current. Thus, a key parameter in increasing the speed of digital circuitry is the saturation current. The major lever to increasing the saturation current is reducing the dimensions of the gate length; hence the drive toward smaller and smaller dimensions.

3.4.5 Subthreshold Bias Region The third bias region of importance in the MOSFET is the subthreshold region, which may be loosely defined as the bias where the surface potential is between one and two Fermi potentials, ϕ_b. This region is also called "weak inversion." In this region, the drain current is formed by diffusion of carriers between the source and drain, rather than Ohmic conduction. In this bias regime, the drain current is exponentially dependent on the gate voltage (Fig. 16). In digital applications, a sharp subthreshold behavior is desirable, since this provides a more ideal "switch." In addition, poor subthreshold characteristics lead to higher currents in the "off" state, which lead to higher current and power consumption. It can also upset certain types of digital logic by allowing stored charge on circuit nodes to leak away. A measure of the "goodness" of the subthreshold characteristic is the amount of gate voltage required to increase the drain current by a decade, denoted the subthreshold swing (measured in mV/decade).

3.4.6 Breakdown Phenomena At sufficiently high drain biases, the electric field at the drain junction of the MOSFET can impart enough energy to carriers near the drain

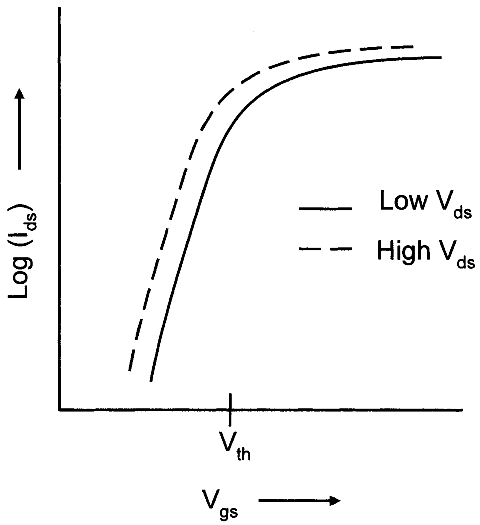

FIG. 16. Logarithm of drain current vs gate voltage for an *n*-channel MOSFET. Solid line: small drain bias. Dashed line: large drain bias.

that they can emit that energy as an electron–hole pair (the impact ionization process; Chynoweth, 1958). One can think of this crudely as an electron colliding with a lattice atom, and ejecting an electron, leaving behind a hole. The electrons are swept into the drain contact, and the holes are emitted into the bulk (assuming an *n*-channel device). As the drain voltage increases, more energy is available to impart to carriers, and the impact ionization rate increases. At high enough drain voltage, this can cause considerable current to flow between the drain and substrate, and may permanently damage the drain junction by localized heating. In short-channel MOSFETs, another type of breakdown phenomena may occur earlier. The generated holes that are swept into the substrate can cause a localized voltage drop between the channel region and the deep bulk. This drop is such that it causes the source junction to become forward biased. This causes more electrons to be injected from the source, increasing the drain current. Larger drain currents increase the amount of carriers available for impact ionization, thus providing more carriers to participate in impact ionization. This positive feedback mechanism can eventually cause the drain current to increase sharply and possibly cause catastrophic failure.

3.4.7 Temperature Effects The main temperature effect in the MOSFET is the change in carrier mobility vs temperature. The inversion layer mobility has a T^{-2} power relation (Sabinis and Clements, 1979). The saturation velocity has a slight dependence on temperature, ranging from 10^7 cm/s at 20 °C to approximately 9×10^6 cm/s at 100 °C for electrons (Canali *et al.*, 1975; Cooper and Nelson, 1981). The saturation velocity for holes is slightly lower (8×10^6 cm/s at 20 °C). Both these effects lead to lower drain currents at higher temperatures. The threshold voltage decreases with increasing temperature, on the order of a few mV/°C. The subthreshold swing is directly proportional to temperature. Temperature effects are important to consider in practical designs, since even with heat sinks or active cooling, the actual operating temperature may be 50 to 75 °C higher than the ambient temperature.

3.5 Small-Device ("Short-Channel") Effects

During the initial development of MOSFETs, the channel length was many microns long, and the device could be designed using the simple formulas presented in the previous sections. These formulas assume that the source and drain regions are sufficiently far from the channel so as not to affect the device behavior. In addition, the channel length was assumed long enough that the current flow in the device could be considered one-dimensional and that the gradual-channel approximation was valid. As MOSFET channel lengths became smaller, and especially in the regime of channel lengths below 1 μm, these assumptions were violated. The following section discusses some of the important effects that modify the behavior of small-dimension MOSFETs.

3.5.1 Threshold Voltage In a long-channel enhancement-mode MOSFET, the threshold voltage is achieved solely by the applied gate potential, which alters the surface potential so that the inversion layer of electrons or holes is formed. Now consider a short-channel MOSFET, where the source and drain junctions are in close proximity to the channel, as in Fig. 17. The source and drain junctions are surrounded by regions depleted of carriers, even at zero source or

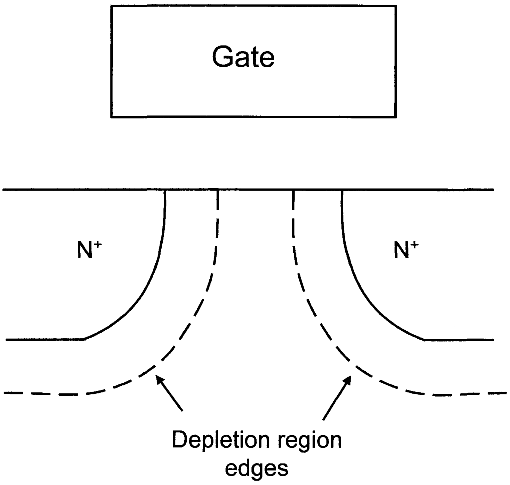

FIG. 17. Cross section of a short-channel MOSFET, showing encroachment of source and drain depletion regions into the channel.

drain bias. When the device length becomes smaller, these depletion regions intrude into the channel, and "predeplete" it. Since part of the channel is already depleted at zero gate bias, it takes less applied voltage to form the inversion layer, and thus the threshold voltage is lowered. This "roll-off" of the threshold voltage is worsened as the channel length decreases. In fact, at some point, the gate loses control over the channel, and the device is considered to be in *punch-through.* A related problem is the so-called drain-induced barrier lowering (DIBL). This effect is the lowering of the device threshold when the drain is biased in saturation. It is an extension of the short-channel roll-off previously discussed. At high drain biases, the depletion region around the drain junction extends farther into the channel. This "predepletes" more of the channel and lowers the threshold voltage further.

Because these effects lower the device threshold voltage, these two effects also increase the amount of current passed by the device when in the "off" state. Recall that in the subthreshold region, the drain current is logarithmically dependent on $V_g - V_{th}$. Even a small decrease in the absolute value of the threshold voltage will cause a large increase in the "off" current. The off current is a critical concern in digital electronics. Besides increasing the amount of current, and hence power, used by devices when they are off, excessive amounts of off current can even cause certain types of circuits to fail to operate properly. Therefore, control of device threshold is a major concern of MOSFET designer engineers.

3.5.2 Velocity Saturation A significant deviation in MOSFET behavior in the saturation bias regime occurs at short gate lengths as a result of carrier velocity saturation. This effect becomes important for micron- and submicron-length devices. As mentioned in the previous discussion on MESFETs in Sec. 2, velocity saturation is the failure of carriers to maintain a linearly increasing velocity with applied electric field and is due to the emission of optical phonons, which are very efficient at transferring energy from carriers to the lattice. Since velocity saturation occurs at high electric fields, which implies high V_{ds}, this effect is noticed in the saturation bias regime. Besides delivering less drain current than expected, velocity saturation also tends to make the saturation current linearly dependent on gate voltage, as opposed to a square-law dependence. It should be noted that the performance of short-channel MOSFETs is not limited entirely by velocity saturation. The carrier mobility at low drain biases is also still important in determining the drain current in saturation. In fact, *p*-channel devices show less velocity-saturation effects than *n*-channel devices (for a given gate length) because the mobility of holes is considerably less than that of electrons (Fig. 6). The higher the low-field mobility, the "faster" the carriers reach their saturated velocity, and the sooner velocity-saturation effects are seen in the drain current.

3.5.3 Hot Carriers and Reliability Another issue that must be considered in short-channel MOSFET design is the effect that high fields have on long-term reliability. We have seen that high electric fields in the MOSFET channel cause the carrier velocity to saturate. These fields are not uniform but are sharply peaked at the drain junction. The carriers gain energy from this field, and some fraction of the high-energy tail of the distribution ("hot" carriers) can acquire enough energy to cause a variety of effects. One effect is impact ionization, where an energetic carrier interacts with the lattice and transfers its energy into creating an electron–hole pair. The newly created electron adds to the drain current, while the hole is swept down into the substrate and adds to current flowing out of the substrate. If a carrier has even more energy, it can surmount the barrier between the silicon and the silicon dioxide and be collected as gate current. These hot carriers can cause damage to the gate oxide in a number of ways. For example, they can induce charges in the oxide, or create traps at the Si/SiO_2 interface that can be occupied by electrons or holes, depending on the available carriers in the channel (Fig. 18). If enough damage occurs to the oxide, it can cause shifts in the threshold voltage or other effects that degrade device performance over time. Typically, the gate currents are very small and thus difficult to measure; however, the substrate current, a close cousin of the gate current, can be used as a proxy to monitor gate current. Accelerated

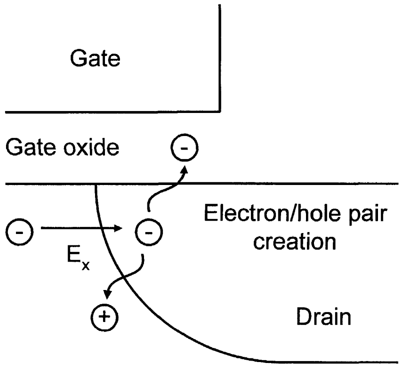

FIG. 18. MOSFET degradation due to hot-carrier injection into the gate oxide.

testing at higher voltages is also used to look at effects that might take years to occur under normal operating conditions. The main tool that MOSFET designers use to reduce hot-carrier effects is modifying the dopant concentration at the source and drain junctions to try to equalize the voltage drop across a larger area, thus reducing the electric field that drives hot-carrier generation. MOSFET designers also have the option of using modified gate oxides (oxynitrides) that are more robust to hot-carrier damage but tend to produce MOSFETs with lower mobility, because of larger amounts of fixed charge at the silicon–oxide interface. The trend toward lower operating voltages (3.3 V and below) reduces the lateral electric fields in the device, although this is counteracted by the continuous trend toward smaller channel lengths, which raise them.

3.6 MOSFET Variants

3.6.1 Silicon-on-Insulator (SOI) MOSFET There are several advantages to fabricating MOSFETs on an insulating substrate, rather than a conducting substrate as is commonly done in industrial practice. Figure 19 shows a cross section of such a device. Earlier SOI MOSFETs were fabricated using thin silicon films grown on completely insulating material that had a close match to the lattice constant of silicon, such as sapphire. The resulting films were prone to defects, and the mobility in the films was not comparable to that in bulk silicon. Recently, interest has grown in using SOI material grown by the silicon implanted oxide (SIMOX) process, where normal silicon wafers are implanted with a high dose of oxygen to form a buried insulating layer. These wafers have better electrical quality than previous silicon on insulators, although maintaining control over the thickness of the upper silicon film is still problematic. The main advantages of SOI over bulk silicon are the reduction in parasitic capacitances under the source and drain junctions due to the source and drain areas butting against the insulating layer, improved resistance to operational upset due to high-energy radiation (important for space applications, for example), immunity to catastrophic interaction between n- and p-channel devices when used together in an integrated circuit (CMOS latch-up), and improved subthreshold characteristics. Disadvantages include less controllable manufacturing, due to difficulties in control-

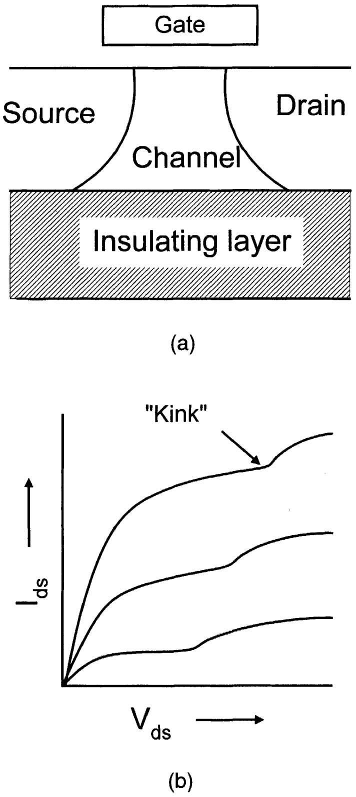

FIG. 19. Cross section of a silicon-on-insulator (SOI) MOSFET. Inset: drain-current vs drain-voltage characteristics, showing the "kink" effect.

ling the silicon film thickness (resulting in changes in threshold voltage), less ability to dissipate heat generated by the device due to the buried oxide, higher wafer costs, and the lack of a contact to the body of the MOSFET. Without this contact, the body of the MOSFET is not tied firmly to any potential and can "float." In particular, any carriers generated by impact ionization at the drain cannot be easily collected and cause the body of the MOSFET to rise to a higher potential, causing a "kink" in the MOSFETs I_{ds}–V_{ds} characteristics (see inset, Fig. 19). Many of these problems are technological, rather than absolute, and may be resolved in the future. SOI MOSFETs remain an area of intense research.

3.6.2 Nonvolatile Memory It is possible to produce MOSFETs whose threshold voltage can be altered electrically by injecting charge onto a gate that is not connected to any other potential ("floating"). In particular, enough charge can be stored on the gate so that the sign of the threshold voltage changes, and the device turns from normally off to normally on. With proper design, the charge can remain on the floating gate for years. Such transistors can be used as a memory device. There are many possible configurations of nonvolatile memory cells. Figure 20 shows one possible configuration, using two levels of polysilicon. The lower level is the floating gate and is separated from the MOS channel by a thin oxide. The second polysilicon layer (control gate) is used to program the device. A large positive voltage is applied to the control gate, which induces a large electric field between the floating gate and the MOS channel. The bands of the silicon substrate are bent sufficiently to allow electron tunneling between the channel and the floating gate, thus storing electrons on the gate. These electrons will make it more difficult to form an inversion channel in the MOSFET, on account of electrostatic repulsion, thus increasing the threshold voltage. The device is now turned off. The floating gate can be discharged by applying a voltage of the opposite sign to the control gate, or, in some designs, by shining ultraviolet light through a window on the device package.

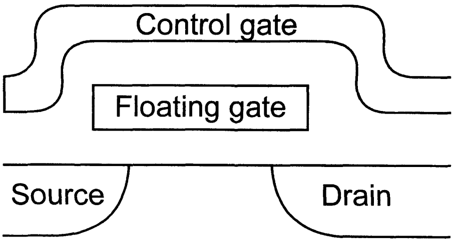

FIG. 20. Cross section of nonvolatile memory, using MOSFET technology.

The obvious benefit of nonvolatile memory is that the memory state is retained, even if the power supply for the device is turned off, unlike traditional semiconductor memory chips. In addition, such memories require less power than traditional semiconductor memory, an important attribute for low-power and portable electronics. At the time of this writing, some industry observers suggest that nonvolatile memory chips may rival traditional semiconductor memory chips in the future.

4. ADVANTAGES AND DISADVANTAGES OF FETs

The two main families of active semiconductor devices are the bipolar transistor family and the field-effect transistor family. The following is a brief comparison of the two:

- Field-effect transistors can be formed in many types of semiconducting materials, since only one type of carrier, usually *n* type, is required. Many compound semiconducting materials, such as GaAs, can only form *n*-type material. Bipolar transistors require both *n*- and *p*-type regions.
- Field-effect transistors can take advantage of advances in lithography to scale their gate dimensions. Since the FET current is proportional to *W/L*, smaller gate lengths improve performance. The current passed by bipolar transistors is proportional to their area ($W \times L$), so scaling dimensions alone do not improve device performance. Bipolar transistors must also rely on advanced processing to create thin layers of different type semiconductor (e.g., base region) to improve performance.
- Bipolar transistors can generally deliver more current for a given area than field-effect transistors, since their current scales

with their area. To generate large currents, field-effect transistors must increase their width, while keeping their length small.

- Field-effect transistors can generally operate at higher and lower temperature ranges than bipolar transistors and do not suffer from thermal runaway effects as bipolar transistors do.
- *n*- and *p*-type silicon MOSFETs can be combined in complimentary digital circuits (CMOS) such that current flows (and power is consumed) only when the circuit is switching. This can lead to lower power consumption than logic circuits implemented with bipolar transistors.
- Bipolar transistors have certain advantages in analog circuits. They have greater linearity than field-effect transistors and can have higher gain in amplifying circuits.
- Bipolar transistors are generally more complex devices than MOSFETs, have more parasitic elements, and are harder to model for computer-aided design.
- Although bipolar transistors have been traditionally used in analog circuitry, MOSFETs can be used to implement a variety of important analog circuits, such as switched capacitor filters and analog-to-digital converters, in an inexpensive and efficient manner.

GLOSSARY

Analog Circuit: An electronic circuit that performs a useful function by representing a quantity, e.g., speech, by an infinitely variable electronic quantity, such as a voltage or current. Examples of analog circuits are amplifiers, modulators and demodulators, and oscillators.

Bipolar Transistor: A three-terminal semiconductor electronic device that operates by modulating the current flow of carriers between two regions of like dopant type (termed the emitter and the collector) separated by a third region of opposite dopant type (denoted the base). The emitter-to-collector current is modulated by the injection of current into the base region. The bipolar transistor relies on the interaction of both carrier types (electrons and holes) to operate.

Compound Semiconductor: A semiconductor that is based on more than one elemental atom. Examples of compound semiconductors are GaAs, InP, and AlGaInAs.

Digital Circuits: An electronic circuit that performs a useful function by representing a quantity, e.g., speech, by a discrete set of electronic quantities, such as voltage or current. Most typical digital circuits have two levels (a binary system), in which one voltage or current level represents a 0, and another a 1.

Dopant Atom: An atom added to a semiconductor crystal that is able to modify the conductivity of the semiconductor by donating an electron (donor atom) or accepting an electron and thus creating a hole (acceptor atom).

Effective Mass: An approximation used to describe the mass of an electron or hole when moving in a crystal lattice. The mass of the carrier appears to change because of the interaction between the carrier and the periodic potential imposed by the crystal lattice.

Energy Bands: A description of the allowed energy states that electrons and holes may have in a crystal.

Heat Sink: A mechanical device that is used to remove generated heat from a semiconductor device. Most heat sinks are made of highly conductive metal, constructed so that they have a high surface area to promote heat transfer.

Hole: The absence of an electron in a material's filled valence band. The motion of a hole can be thought of as the motion of an electron from the sea of electrons in the valence band into the gap. Holes are positively charged. Although holes and electrons are often thought of as identical, except for charge, this is only an approximation. A hole's properties are controlled by the valence band(s), which can be very different from an electron's conduction band(s).

Intrinsic Carrier Concentration (n_i): The number of holes or electrons that are present in a semiconductor through thermal generation only, i.e., in a pure crystal.

Inversion Layer: A thin conductive layer that forms when the surface potential of a MOS device is raised (or lowered) so that a layer of electrons (holes) is formed at the semiconductor–insulator interface.

Inverter (logic gate): The basic binary digital circuit, which transforms one binary state at the input to the other binary state at the output.

Ion Implantation: A technology that takes an ionized impurity atom and accelerates it by means of an electric field and bombards it into a semiconductor target. This technique is used to introduce dopant atoms into the semiconductor. It is used in all modern IC processing because of its ability to control the amount of dopant introduced, as well as its depth in the semiconductor.

Lithography: A technology for transferring patterns from a predrawn mask onto the surface of a semiconductor wafer. Typically, the pattern is formed in a layer of photosensitive material (photoresist) that changes chemically when selectively exposed to light or electron bombardment. The chemically altered photoresist is either hardened or made easier to remove.

Mobility: The proportionality factor between an applied electric field and the resulting velocity of an electron or hole. The mobility is not a constant but depends on the number of dopant atoms, the temperature, and even the electric field itself.

Optical Phonon: A mechanical vibration quantum in a crystal that has energy in the range of an electron volt.

Schottky Barrier: A contact between a metal and a lightly doped semiconductor, such that a rectifying (non-Ohmic) contact is formed. A rectifying contact readily passes current in one direction but blocks current flow in the opposite direction.

Semiconductor: In a semiconductor, a gap of approximately 1 eV occurs between the energy states of the tightly bound, or valence, electrons and the outer, conduction electrons. Thermal energy can excite a limited number of electrons from the valence band into the conduction band, thus producing a poorly (semi)conducting material.

Superlattice: A semiconductor structure that is formed by repeated layers of different composition, such that the electronic properties are altered for a useful purpose.

Works Cited

Arora, N. D., Hauser, J. R., Roulston, D. J. (1982), "Electron and Hole Mobilities in Silicon as a Function of Concentration and Temperature," *IEEE Trans. Electron Dev.* **ED-29,** 292–295.

Canali, C., Manji, G., Minder, R., Ottaviani, G. (1975), "Electron and Hole Drift Velocity Measurements in Silicon and Their Empirical Relation to Electric Field and Temperature," *IEEE Trans. Electron Dev.* **ED-22,** 1045–1047.

Chynoweth, A. G. (1958), "Impact Ionization Rates for Electrons and Holes in Silicon," *Phys. Rev.* **109,** 1537–1540.

Cooper, J. A., Nelson, D. F. (1981), "Measurement of the High Field Drift Velocity of Electrons in Inversion Layers on Silicon," *IEEE Electron Dev. Lett.* **EDL-2,** 171–173.

Dacey, C., Ross, I. M. (1953), "Unipolar Field Effect Transistor," *Proc. IRE.* **41,** 970–979.

Heil, O. (1935), British patent No. 439,457.

IEEE (1975), *Standard Graphic Symbols for Electrical and Electronics Diagrams* (STD 315-1975), New York: IEEE.

Jacoboni, C., Canali, C., Ottaviani, G., Quaranta, A. A. (1977), "A Review of Some Charge Transport Properties of Silicon," *Solid State Electron.* **20,** 77–89.

Lilienfeld, J. E. (1930), U.S. patent No. 1,745,175.

Mead, C. A. (1966), "Schottky Barrier Gate Field Effect Transistor," *Proc. IEEE* **54,** 307–308.

Sabinis, A. G., Clements, J. T. (1979), "Characterization of the Electron Mobility in the Inverted ⟨100⟩ Si Surface," *IEEE Tech Digest,* Int. Electron Devices Meeting, p. 18.

Shockley, W. (1952), "A Unipolar Field Effect Transistor," *Proc. IRE* **40,** 1365–1376.

Sze, S. M. (1981), *Physics of Semiconductor Devices,* New York: Wiley.

Further Reading

Nicollian, E. H., Brews, J. R. (1982), *MOS Physics and Technology,* New York: Wiley.

Pierrot, R. F. (1983), *Modular Series on Solid State Devices,* Volume 4, *Field Effect Devices,* Reading, MA: Addison-Wesley.

Streetman, B. (1980), *Solid State Electronic Devices,* Englewood Cliffs, NJ: Prentice Hall.

Sze, S. M. (1981), *Physics of Semiconductor Devices,* New York: Wiley.

Sze, S. M. (Ed) (1983), *VLSI Technology,* New York: Wiley.

Tsividis, Y. P. (1987), *Operation and Modeling of the MOS Transistor,* New York: McGraw Hill.

TRANSISTORS, MICROWAVE

YOUNG-KAI CHEN, *High Speed Electronics Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey, U.S.A.*

Introduction 165
1. **Equivalent-Circuit Models and Figures of Merit** 166
1.1 Small-Signal Transistors 166
1.2 Microwave Low-Noise Transistors 167
1.3 Microwave Power Transistors 168
2. **Microwave Bipolar Transistors** 169
2.1 Homojunction Bipolar Transistors 169
2.1.1 Device Operation 169
2.1.2 Device Models 169
2.1.3 Noise Characteristics 170
2.1.4 Power Characteristics 171
2.2 Heterojunction Devices 172
3. **Microwave Field-Effect Transistors** 173
3.1 Equivalent Circuit and Its Frequency Response 173
3.2 Noise Characteristics 174
3.3 Microwave Metal-Oxide-Semiconductor Field-Effect Transistors (MOSFETs) 175
3.4 Microwave Schottky Metal-Semiconductor Field-Effect Transistors (MESFETs) 176
3.5 High–Electron-Mobility Transistors (HEMTs) 177
4. **Monolithic Microwave Integrated Circuits (MMICs)** 179
4.1 Semiconductor Substrate for MMICs 180
4.2 Matching Networks 181
4.3 Passive Lump Elements 181
4.4 Passive Distributed Elements 182
4.5 Computer-Aided Design of MMICs 183
Glossary 183
Works Cited 185
Further Reading 185

INTRODUCTION

Microwave is any electromagnetic wave with a frequency in a range commonly regarded as extending from 300 MHz to 300 GHz, covering the spectrum lying between the radio frequency and the far infrared. The part of the spectrum higher than 30 GHz is also regarded as the millimeter-wave band, as the corresponding wavelength in free space is comparable to or shorter than a millimeter. Microwave is widely used to carry voice and data signals over the air and into deep space. Beginning in the World War II era, the microwave spectrum has also been segmented and labeled into many sub-bands: *L*-band (1.12–1.70 GHz), *R*-band (1.70–2.60 GHz), *S*-band (2.60–3.95 GHz), *H*-band (3.95–5.85 GHz), *C*-band (5.85–8.20 GHz), *X*-band (8.20–12.4 GHz), *Ku*-band (12.4–18.0 GHz), *K*-band (18.0–26.5 GHz), *Ka*-band (26.5–40.0 GHz), *Q*-band (33.0–50.5 GHz), *U*-band (40.0–60.0 GHz), *V*-band (50.0–75.0 GHz), *E*-band (60.0–90.0 GHz), *W*-band (75.0–110.0 GHz), *F*-band (90.0–140.0 GHz), *D*-band (110.0–170.0 GHz), and *G*-band (140.0–220.0 GHz).

Microwave transistors are widely used in many instrumentation and wireless communication systems: they provide the essential function of generating, amplifying, and modulating current source at microwave frequency. Various materials such as semiconductors, superconductors, or metal-based systems have been experimented with to realize microwave transistors. Nevertheless, almost all of today's microwave transistor functions are implemented on semiconductors for any practical applications. This article first reviews figures of merit of basic microwave applications and then surveys a few key semiconductor microwave transistor technologies based on these factors.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. EQUIVALENT-CIRCUIT MODELS AND FIGURES OF MERIT

In order to link the physical transistor behaviors to the microwave characteristics, the equivalent-circuit model and a few figures of merit are introduced in this section for several applications.

1.1 Small-Signal Transistors

For small-signal excitations, the complex current–voltage characteristics of a microwave transistor can be simplified and linearized for a given dc bias condition. For example, a *p-n* junction can be modeled as a capacitor and a resistor in parallel for microwave signals, where its resistance value represents the slope of the diode's exponential voltage–current characteristics and its capacitance value reflects both the diffusion capacitance and the capacitance from the depletion layer under a steady-state bias condition. By exercising this technique to simplify physical carrier-transport equations, a set of equivalent-circuit models assembled from a set of simple linear lump elements such as capacitors, resistors, and dependent current sources would describe the small-signal microwave characteristics of the intrinsic part of a microwave transistor adequately.

To describe accurately the transistor frequency response at microwave frequencies, the contribution from the extrinsic parasitic elements in the microwave transistor, such as interelectrode capacitors, contact resistors, and bonding-wire inductance, is significant and should be included in the equivalent circuit. Unlike the equivalent-circuit elements of the intrinsic part of the transistor, the value of these extrinsic parasitic elements is independent of the operating bias of the transistor. At higher end of the microwave band, for example, in the millimeter-wave bands, the physical dimensions of a microwave transistor and its interconnection leads would be as large as a fraction of the wavelength, and distributed effects should be incorporated into the equivalent-circuit model. Because of this distributed nature at microwave frequencies, both the amplitude and the phase of the instantaneous voltage and current wave forms of the transistor are highly sensitive to the external impedance presented to the transistor terminals, and it is very difficult to determine them precisely. Therefore, it is common to use the incident and reflected power wave forms and scattering matrices from a known termination reference impedance to characterize the transistor at microwave frequency rather than using the instantaneous current and voltage wave forms at dc or low frequency. Nevertheless, the instantaneous current and voltage can be derived from the scattering parameters if the terminating impedance at the transistor terminals is known (Vendelin *et al.*, 1990).

Once the equivalent circuit of the transistor is determined and verified by experimental data, the frequency response can be predicted by circuit theory for any external termination conditions. The amplifier gain of the transistor usually rolls off at high frequency because of the long transit delays of the carriers and the charging time of capacitive elements. When the gain is reduced to unity, the corresponding frequency is called the cutoff frequency of a transistor. Figure 1 shows the frequency dependence of commonly used transistor gain characteristics. Hybrid current gain (h_{21}) is defined as the output-to-input current gain of the transistor, when the output of the transistor is short-circuited. The short-circuit current gain cutoff frequency, f_T, is widely used to monitor the intrinsic transport properties of the charge carriers. Unilateral power gain (U) is defined as the maximum power gain

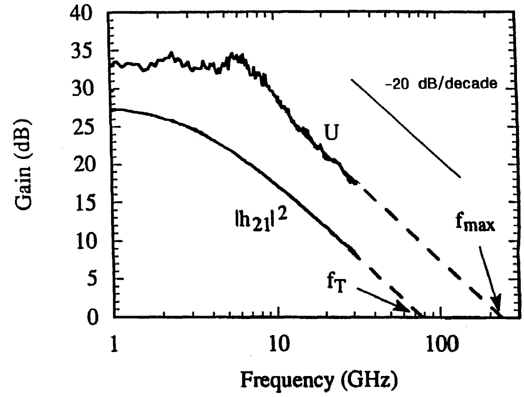

FIG. 1. Frequency dependence of the unilateral power gain, U, and current gain, h_{21}, of a microwave transistor. The gain rolls off at a slope of −20 dB/decade under the influence of a dominant time constant.

of the transistor when both input and output of the transistor are connected to an impedance conjugate matched to the internal impedance of the transistor and the internal feedback in the transistor is nullified by an external feedback circuit. The *U* represents the highest power gain a transistor can produce theoretically for a given frequency and bias condition. When *U* rolls off to unity, the corresponding frequency is defined as the maximum unilateral power-gain cutoff frequency, $f_{\max}$, which is the highest frequency at which a transistor can still produce gain, usually in an oscillator circuit.

In most cases, the frequency response of a microwave transistor is limited by a dominant time constant. This exhibits a dominant pole in the frequency response, and the linear gain of the transistor reduces at a rate of -6 dB/octave or -20 dB/decade with the rise of frequency. For a transistor with an $f_{\max}$ of 100 GHz, we may expect a maximum gain of 20 dB at 10 GHz [i.e., $20 \log_{10}(f_{\max}/f)$]. The actual amplifier gain will be less than 20 dB because of the mismatches between external impedance and the input and output impedances of the transistor. The mismatches on the external impedance are necessary in many cases to enhance the amplifier noise figure, to obtain a broad and flat frequency response, or to avoid instability of the amplifier.

1.2 Microwave Low-Noise Transistors

Many microwave amplifiers are designed to boost very weak signals. However, a microwave transistor also generates noises and adds it to the amplified signal. The noise figure, *F*, is a measure of the output signal quality, defined as the ratio of the input signal-to-noise ratio to that of the output:

$$F = \frac{(S/N)_i}{(S/N)_0}. \tag{1}$$

There are many internal noise sources in a microwave transistor such as low-frequency generation–recombination noise, broad-band current shot noise, current fluctuation noise, and thermal noise. The output noise of a transistor is the ensemble average of the amplified input noise plus this internally generated noise. These internal noise sources, depending on their physical origins, are added to the output of the transistor through a web of equivalent-circuit elements.

From the quantum-mechanical fluctuation and dissipation theory, the internally generated noise will be minimized once these noise sources can be dissipated through an optimal external impedance. For a given set of frequency and bias conditions, the noise figure of a transistor reaches its minimum value $F_{\min}$ when its input is connected to an optimal admittance Y_{opt} to dissipate the internally generated noise power effectively. Given this input termination of Y_{opt} and a conjugate matched-output matching network, the maximum available gain of this transistor is called associated gain, G_A. Together with the equivalent internal noise resistor R_n, $F_{\min}$, G_A, and Y_{opt} completely describe the noise and gain characteristics of a microwave transistor. The gain can be increased by connecting another value of admittance *Y* to the transistor. However, the noise figure of this new amplifier will be degraded from its lowest value to

$$F = F_{\min} + [R_n/\mathrm{Re}(Y)]|Y - Y_{\text{opt}}|^2. \tag{2}$$

To obtain a desirable signal level, cascading of multiple stages of transistors is necessary. The gain and noise figure of a multiple-stage amplifier is related to that of each amplifying stage by

$$G = G_1G_2G_3 \ldots \tag{3}$$

and

$$F = F_1 + (F_2 - 1)/G_1 + (F_3 - 1)/G_1G_2 + \ldots, \tag{4}$$

where F_i's and G_i's are the noise figure and associated gain of each amplifying stage, respectively. It is important to note that both the noise figure and the associated gain of the first couple of stages dominate the overall amplifier performance. A low-noise microwave transistor located at the forefront of the amplifier should have a very low noise figure and a very high power-gain cutoff frequency to produce enough gain margin at the operating frequency. The excess gain is needed to compensate the loss from the impedance mismatch Y^*_{opt} presented at the input of the transistor for $F_{\min}$ and to discount the noise contribution from the following stages. For many satellite microwave receivers operating at a few gigahertz frequency, it is common to use low-noise transistors with cutoff frequency more than 100 GHz so that the transistor in the first stage will have the lowest noise figure while it still provides plenty of gain so as to minimize the noise contributions from the rest of the stages.

1.3 Microwave Power Transistors

In many applications, microwave transistors are utilized to handle large signal swings. For example, in the mixer circuit, a transistor is driven into the nonlinear operation regime such that it mixes two input signals to produce desirable higher-order harmonic contents at the output. For power-amplifier applications, the signal level is also strong enough that the characteristics of a microwave transistor can no longer be described by its small-signal model, and a complete description of the transistor device model should be used.

Power amplification is a very common application for microwave transistors. Many figures of merit of power amplifiers, such as saturated output power, power-added efficiency, and available power gain, are linked to the transistor characteristics. Power transistors are biased in one of three modes: class A, B, or C. The class A operation offers the best linearity, while the class B and C configurations provide better peak power and efficiency.

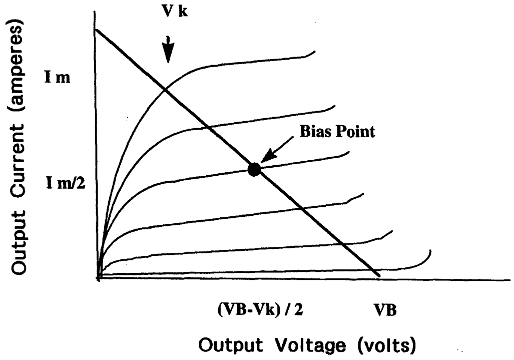

FIG. 2. Bias point of a microwave transistor for class-A linear amplification. The knee voltage V_K, breakdown voltage V_{BK}, and maximum current I_m of the transistor determine the maximum output power of the transistor.

As shown in Fig. 2, the bias point for class A operation is chosen in the middle of the peak voltage and current swing such that the peak output current and voltage will not reach the nonlinear region of the transistor. The maximum power level and maximum power-added efficiency are approximated by

$$\text{Maximum Output Power } (P_{\max}) = (V_{BK} - V_K)I_{\max}/8 \quad (5)$$

$$\sim V_{BK}I_{\max}/8, \quad (6)$$

and

$$\text{Power-Added Efficiency (PAE)} = P_{\max}/(P_{dc} + P_{\max}/G) \quad (7)$$

$$\sim 1/(2 + 1/G) < 50\% \quad (8)$$

with the voltage and current of the transistor biased at $(V_{BK} - V_K)/2$ and $I_{\max}/2$, respectively. V_K is the output knee voltage at the maximum output current $I_{\max}$, V_{BK} is the output breakdown voltage, P_{dc} is the dc power consumption, and G is the associated microwave gain of the transistor amplifier. In order to provide the desired current and voltage at the output, the resistance of the output load line is approximately $V_{BK}/I_{\max}$, which is usually much less than the output resistance of the transistor for the optimal matching for gain. Assuming a mismatch loss of α to the maximum available gain of the transistor at an operating frequency f, the associated gain and the power-added efficiency can be rewritten as

$$\text{Associated Gain} \sim \alpha f_{\max}/f, \quad (9)$$

$$\text{Maximum Power-Added Efficiency} = 1/(2 + f/\alpha f_{\max}). \quad (10)$$

For high output power, high gain, and good efficiency, it is clear that we need transistors with large V_{BK}, $I_{\max}$, and $f_{\max}$.

For class B and C operations, the transistor is biased near the cutoff region at $V_{BK}/2$ in order to obtain a full peak current swing of $I_{\max}$. It relies on the external filter circuit to remove the nonlinear distortion. For example, the maximum saturated power and power-added efficiency of an ideal class B power amplifier can be expressed as

$$P_{\max,B} = (V_{BK}/2 - V_K)I_{\max}/2 \quad (11)$$

$$P_{\mathrm{dc,B}} = (\pi/4)(V_{\mathrm{BK}}/2 - V_{\mathrm{K}})I_{\mathrm{max}} \qquad (12)$$

$$\begin{aligned}\mathrm{PAE}_{\mathrm{max,B}} &= P_{\mathrm{max,B}}/P_{\mathrm{dc}} \\ &< 2/\pi = 63.6\%. \qquad (13)\end{aligned}$$

The maximum power-added efficiency of a class C power amplifier is even higher when the amplifier is driven more into the saturation regime. The value will be close to 100% when the output voltage and current wave forms look like a square wave with one at its peak and the other one at zero. However, in practical applications, a significant filtering network will be needed at the output of this nonlinear amplifier in order to remove higher-order harmonics and intermodulation products.

2. MICROWAVE BIPOLAR TRANSISTORS

Bipolar transistors are widely used to generate, amplify, and mix microwave signals because of their unique noise characteristics and current-handling capability. In this section, we review the microwave characteristics of both homojunction and heterojunction bipolar transistors. The general operation principles and dc characteristics of bipolar transistors can be found in other articles in this Encyclopedia, such as Transistors, Bipolar.

2.1 Homojunction Bipolar Transistors

2.1.1 Device Operation Homojunction bipolar transistors are configured in either *p-n-p* or *n-p-n* doped multilayers as emitter (E), base (B), and collector (C). The microwave signals are carried by the minority carriers such as holes in the *n*-doped base of *p-n-p* transistors or electrons in the *p*-doped base of *n-p-n* transistors, respectively. These minority carriers are injected from the emitter side of the forward-biased E-B junction into the base region. These minority carriers then drift and diffuse through the heavily doped base region. Some minority carriers will recombine with the majority carriers in the base and E-B junction as part of the base current. Those base minority carriers reaching the reverse-biased B-C junction will be accelerated by the high electric field and collected by the collector as part of the collector current.

The intrinsic frequency response of the bipolar transistor is determined by the total transit delay from the injection of carriers in the emitter till the collection of these carriers in the collector. This intrinsic delay consists of the charging delay of the E-B junction, the transit delay in the base and B-C space-charge region, and finally the charging delay of the B-C junction. The base doping level needs to be high to reduce the resistance of base current in the thin base.

2.1.2 Device Models The equivalent circuit of a bipolar transistor is shown in Fig. 3 when the bipolar transistor is operated

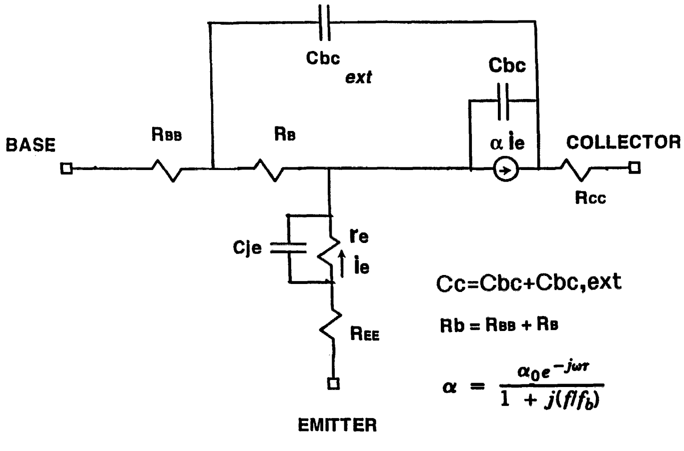

FIG. 3. Equivalent circuit of a microwave bipolar transistor.

in the active mode with a forward-biased E-B junction and a reverse-biased B-C junction. The total transit delay τ_T from the emitter to the collector is the sum of the emitter charging time (τ_E), base transit time (τ_B), collector transit time in the B-C space-charge region (τ_{SCC}), and collector charging time (τ_C), where

$$\tau_E = (R_E + R_B + Rcc + R_{EE})C_{be}, \tag{14}$$

$$\tau_C = (R_E + R_B + Rcc + R_{EE})C_{bc}, \tag{15}$$

$$\tau_B = W_B^2/2D, \tag{16}$$

$$\tau_{SCC} = t_C/2V_c, \tag{17}$$

and W_B, t_C, D, and v_c are the thickness of the neutral base, the thickness of the collector space-charge region, the diffusion constant of minority carriers in the base, and the carrier velocity in collector, respectively.

From this intrinsic delay, many figures of merit such as cutoff frequencies can be derived. The common-emitter current-gain cutoff frequency f_T and the maximum power-gain cutoff frequency $f_{\max}$ can be expressed as

$$f_T = 1/2\pi\tau_T \tag{18}$$

$$\begin{aligned} &= 1/2\pi[C_{be}(R_E + R_B + Rcc + R_{EE}) \\ &\quad + C_{bc}(R_E + R_B + Rcc + R_{EE}) \\ &\quad + \tau_B + \tau_{SCC}] \end{aligned} \tag{19}$$

and

$$f_{\max} = \sqrt{\frac{f_T}{8\pi C_{bc}R_B}}. \tag{20}$$

Both f_T and $f_{\max}$ are dominated by the emitter charging time $R_E C_{be}$ at low operating collector current, while they are limited by the transit time at high current. The power-gain cutoff frequency $f_{\max}$ is also limited by the external parasitic elements such as R_B and C_{bc}. For microwave applications, special attention in the IC processing has to be addressed on how to reduce the base resistance and collector capacitance, and these transistors are usually operated at high current density to reduce the emitter charging time. Figure 4 shows the cross-section diagram of a state-of-the-art silicon bipolar transistor with a super self-aligned process (Ugajin and Fossum, 1994). A super self-aligned process is needed to reduce the parasitic base resistance and collector junction capacitance.

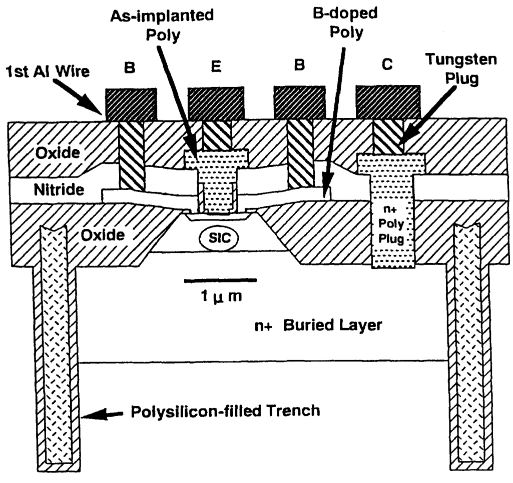

SIC = Selectively Ion-implanted Collector

FIG. 4. Cross section of a high-performance silicon bipolar junction transistor fabricated with self-aligned implantation technology. The process is carefully designed to minimize base resistance and collector feedback capacitance to achieve high cutoff frequency.

2.1.3 Noise Characteristics There are many noise sources in a bipolar transistor. The dominant noise sources are generation–recombination noise in the emitter space-charge region, shot-noise sources in the emitter junction and collector junctions, thermal (or Johnson) noise from the base resistance, recombination noise in the extrinsic base region, multiplication noise in the high-field collector space-charge region, etc. These noise sources can be introduced into an equivalent circuit such as Fig. 3. The collective terminal noise characteristics of a bipolar transistor result from the interplay of these sources with the rest of the device equivalent circuit.

The frequency dependence of the spectral density of the collector noise current is illustrated in Fig. 5. The noise-current spectrum shows three regions:

1. $1/f$ rolloff characteristics at low frequency,
2. a flat noise floor over a very broad bandwidth, and
3. a steep rise ($\sim f^2$) at high frequency.

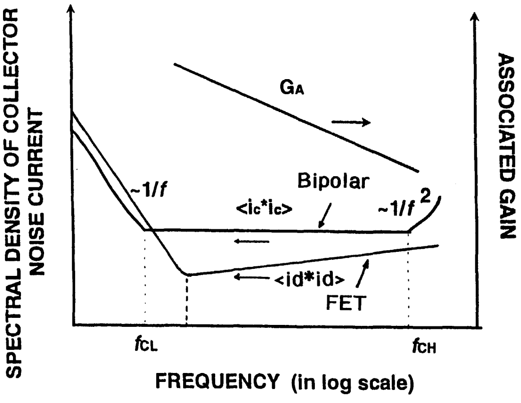

FIG. 5. Illustration of current-noise power spectrum of a bipolar transistor and a FET.

The shot noise, thermal noise, and multiplication noise in the collector depletion region make up the broad-band noise floor. The generation and recombination noise in the emitter-base depletion region and the surface states dominates the low-frequency noise spectrum. At very high frequency, those two shot-noise sources in the emitter junction and collector junction are no longer correlated by the transit delay, and total noise current increases substantially.

The low-frequency noise is very important to many applications, such as high-precision analog-to-digital converters, ultrastable microwave oscillators, etc. The low-frequency current fluctuations in the oscillating transistor will be up-converted to the microwave frequency, and the stability of the oscillating frequency (or the resultant phase noises) will significantly deteriorate. In the field-effect transistors (FETs) based on lateral carrier transportation parallel to the surface of the semiconductor, the generation and recombination current from the buffer layer and substrate is the major source of low-frequency current noise. As a result, the low-frequency noise-current density is high in FETs. Devices based on vertical carrier transport, such as emitter-up bipolar junction transistors (BJTs) or heterojunction bipolar transistors (HBTs), are less sensitive to the electronic defects and imperfections in the buffer layer and substrate material. The dominant carrier-transport mechanism in the emitter and base of bipolar transistors is effectively shielded from noise sources in the substrate by the high electric field in the collector as well as the highly doped and quasineutral n^+ region of the subcollector contacting layer.

The high-frequency noise of bipolar transistors is dominated by the broad-band thermal and shot noises. Therefore, the minimum noise figure can be calculated from the macroscopic equivalent-circuit model using the approximation from Hawkins (1977):

$$F_{\min} = \frac{a(R_b + R_{\text{opt}})}{r_e} + \frac{(1 + f^2/f_F^2)}{\alpha_0}, \tag{21}$$

where

$$R_{\text{opt}} = \left[(R_b^2 - X_{\text{opt}}^2) + \frac{(1 + f^2/f_F^2)r_e(2R_b + r_e)}{a\alpha_0}\right]^{1/2}, \tag{22}$$

$$X_{\text{opt}} = 2\pi f C_{je} r_e^2 (1 + f^2/f_F^2)/a\alpha_0, \tag{23}$$

and

$$a = [(1 + f^2/f_F^2)(1 + f^2/f_e^2) - \alpha_0]/\alpha_0, \tag{24}$$

with $f_e = 1/2\pi\tau_e$, $f_F = 1/2\pi\tau_F$, and dc common-emitter current gain $\beta_0 = \alpha_0/(1 - \alpha_0)$. Here we replace the base transit-time cutoff frequency f_b by the total transit-time cutoff frequency f_F.

An optimal microwave low-noise bipolar transistor requires very low base resistance and very small forward transit delay. Low base resistance and high associated gain can be realized with a submicron emitter stripe width and high base doping density. The shot noise would be significantly reduced by operating the transistor at a small collector current. The microwave gain is preserved by the high collector current density in the small active area.

2.1.4 Power Characteristics Bipolar transistors have been a very popular choice for microwave power applications because of their large current-driving capability and thin vertical structure for high-frequency performance. The maximum power that can be delivered by a bipolar transistor is limited by its voltage swing and current swing. The output voltage swing is dominated by the breakdown voltage of the collector junction. In order to obtain a high collector breakdown voltage, a very thick and lightly doped collector depletion region is used. However, the lightly doped collector depletion region also limits the space-charge current density and, therefore, the maximum current that can be driven through the transistor. This collector-current density places a limit on the charging time constant of the emitter–base junction and the cutoff frequency of the bipolar transistor. For a silicon bipolar transistor, the product of the current-gain cutoff frequency f_T and the collector breakdown voltage VB_{CEO} is a constant typically less than 120 GHz-V (Nanba *et al.,* 1993). The low cutoff frequency of a microwave transistor means a low power gain and poor power-added efficiency. The microwave gain is degraded even further when ballast emitter resistors are added to the emitter to improve the linearity and thermal stability.

In order to deliver high current, many active emitters are connected in parallel to form a power transistor. Because of the nonuniform temperature distribution and the potential drops along the interconnection lines, the base–emitter junction voltage may be different by a few multiples of thermal voltage kT/q from one part of the transistor to another. This slight variation in the base potential in one cell will amplify the collector current of this cell exponentially. The increase of the collector current heats this weak cell even hotter and leads to thermal runaway. This is the major reliability problem of power bipolar transistors. In order to desensitize the variations in the turn-on potential and thermal distribution of massive parallel emitter–base junctions in a power transistor, a ballast resistor is usually added in series with each emitter. This ballast resistor also acts as a negative feedback to linearize the output collector current such that the output current is no longer responding to the input voltage in an exponential way (Yamada *et al.,* 1996).

2.2 Heterojunction Devices

Kromer (1982) first suggested that the high-frequency performance of transistors may be improved by replacing the heavily doped homojunction emitter with a lightly doped heterojunction emitter made of a material with the same lattice constant but with a higher band-gap energy than the base material. The high band-gap energy of the emitter of a heterojunction bipolar transistor (HBT) effectively suppresses the back-injected carriers from the base to preserve the dc current gain. Therefore, a lightly doped emitter can be used to lower the base–emitter junction capacitance, and the high-frequency power gain is improved by the high base doping. With the advances of the modern epitaxial growth technologies, such as molecular-beam epitaxy (MBE), organometallic vapor-phase epitaxy (OMVPE), and low-pressure chemical vapor deposition (LPCVD), high-frequency microwave HBTs with more than 100-GHz cutoff frequency have been realized using many III–V compound semiconductor materials (Ishibashi and Yamauchi, 1988; Chen *et al.,* 1989) as well as traditional silicon materials with a strained narrow–band-gap base with a small germanium composition (Crabbe *et al.,* 1993). The heavily doped base also alleviated some problems in designing traditional homojunction bipolar transistors such as base width modulation from the collector bias [a.k.a. Early effect (Early, 1952)] and gain suppression at high injection level.

Most microwave HBTs are implemented with one heterojunction and typically as *n-p-n* configuration to utilize the high mobility and high velocity of electrons. A large valence-band offset between the emitter and base (ΔE_V) is desirable to suppress the back-injected holes from the heavily doped base. The dc current gain β can be expressed as

$$\beta \approx \frac{N_{D,E}}{N_{A,B}} \exp\left(\frac{\Delta E_v}{kT}\right) \tag{25}$$

where $N_{D,E}$ and $N_{A,B}$ are the donor density in the emitter and the acceptor density in the base, respectively. It can be seen that high current gain and a high emitter-to-base doping ratio is possible even with a small valence-band discontinuity.

Double heterojunction HBTs are also used to enhance the breakdown voltage for high-power applications by employing wide–band-gap material at both emitter and collector. However, care should be taken to minimize the conduction-band spike of the base–collector heterojunction, which may block the transport of base minority carriers into the collector depletion region. Table 1 lists some

Table 1. Commonly used heterostructures for microwave heterojunction bipolar transistors (HBTs).

Substrate	Emitter	Base
Silicon	Si	$Si_{(1-x)}Ge_x$
GaAs	AlGaAs	GaAs
	InGaP	$In_xGa_{1-x}As$
	AlGaP	InGaP
InP	AlInAs	InGaAs
	InP	
	InGaP	
GaSb	AlSb	InAs
SiC	AlGaN	SiC
Sapphire	GaN	GaInN

of the popular material selections to build HBTs.

Despite the difference of carrier-transport dynamics between the HBT and BJT, the ac equivalent-circuit model and the frequency response are dominated by the carrier transit time as well as charging-time delay from the junction and contact parasitic. Both the microwave frequency response and the noise properties of HBTs are limited mainly by the base resistance and base–collector junction capacitance. By reducing the base thickness for short carrier transit time and increasing its doping concentration to reduce the excess base resistance, microwave HBTs with greater than 100-GHz frequency response and excellent microwave noise characteristics are realized in many material systems.

3. MICROWAVE FIELD-EFFECT TRANSISTORS

The general operation principles of various field-effect transistors (FETs), such as MOSFETs, MESFETs, and HEMTs, are described in detailed in other articles in this Encyclopedia and will not be repeated here. Critical device parameters affecting microwave operation are examined in this section.

3.1 Equivalent Circuit and Its Frequency Response

In general, the current transport in the FET is by the majority carrier. The carriers are injected from the source contact into the conduction channel underneath the gate electrode. After the carriers are accelerated by the electric field in the channel, they are collected by the drain contact. The weak ac signal is typically applied to the gate electrode, and then amplified by the channel conductance modulation through the source-to-drain current.

The high-frequency small-signal equivalent-circuit model of a FET is depicted in Fig. 6. The high-frequency response of a FET is limited by the carrier transit time between the source and drain underneath the gate field:

$$\tau = 1/2\pi f_T = L_g/\langle v_s \rangle = 1/[g_m/2\pi(C_{gs} + C_{gd})], \tag{26}$$

where f_T is the current-gain cutoff frequency, L_g is the effective gate length, $\langle v_s \rangle$ is the average carrier velocity in the channel, g_m is the transconductance, and C_{gs} and C_{gd} are the gate-to-source and gate-to-drain capacitance, respectively. The maximum power-gain cutoff frequency f_{max} is limited by the parasitic resistive and capacitive elements:

$$f_{max} = f_T/[4(R_g + R_s + R_{in})/R_{ds} + 4\pi f_T C_{gd}(R_g + R_s + R_{in})]^{1/2}, \tag{27}$$

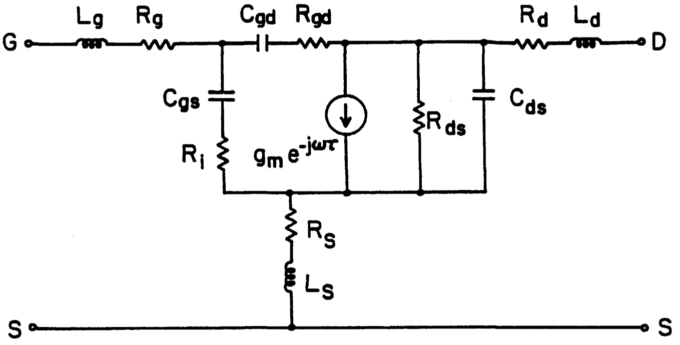

FIG. 6. Equivalent circuit of a microwave field-effect transistor.

where R_g, R_s, R_{in}, and R_{ds} are the gate resistance, source resistance, channel resistance, and source–drain output resistance, respectively, and C_{gd} is the gate–drain capacitance. For devices fabricated on a highly conductive substrate, the frequency response would be further degraded by the parasitic capacitance from the substrate.

To design a good microwave FET, it is necessary to increase the high-frequency gain by reducing the gate length as well as parasitic resistance and capacitance. However, the gate resistance increases with the reduction of gate length. The effective R_{gs} can be reduced by increasing the gate cross section using a T-shaped structure, which has a small footprint and large conductive overlay. A multiple-finger structure with shorter unit gate width per finger would also reduce the total gate resistance. To reduce the source resistance (R_s), one uses a self-aligned structure that places an ion-implanted source next to the gate as in metal-oxide-semiconductor field-effect transistors (MOSFETs) and Schottky metal-semiconductor field-effect transistors (MESFETs), or utilizes a high-mobility material as in a high–electron-mobility transistor (HEMT) structure.

With the high electric field forming under the short gate electrode, the hot-electron injection into the substrate will increase the transit delay and reduce the output resistance R_{ds}. The increase of output conductance from hot carriers will reduce the high-frequency power gain and the maximum power-gain cutoff frequency f_{max}. A higher barrier layer between the active channel and the substrate is needed to confine these hot carriers inside the channel using either a homojunction or a heterojunction. By taking all these into consideration, today's state-of-the-art microwave transistors with 0.1-μm T-shaped metal gate have demonstrated very high power-gain cutoff frequency f_{max}, up to 600 GHz using InP substrate (Smith *et al.*, 1995).

3.2 Noise Characteristics

The FET usually offers the best microwave noise performance of all types of microwave transistors. The physical reason behind this selection is that the intrinsic noise generators inside a FET are the mobility fluctuation (or velocity fluctuation) of the channel current and the thermal (or Johnson) noise from contact electrodes. For the same amount of operating current, the noise power from a resistive source is half that produced by the shot noise through a *p-n* junction. However, from the quantum-mechanical noise generation and dissipation theorem, these intrinsic noise reservoirs should be discharged through an optimal drainage impedance presented at the gate electrode at the operation frequency. Otherwise, when a nonoptimal termination impedance is connected to the gate electrode, the excess gate noise would be amplified by the gain mechanism (i.e., the transconductance) in the output drain current. Unfortunately, as shown in Fig. 7, the optimal input impedance for the lowest noise figure is not the same as the best impedance to obtain the highest gain. The selection of terminal impedance at the input is a trade-off between the noise figure and gain.

The minimum noise figure of a FET in a common-source configuration can be expressed by the widely used empirical Fukui's equation (Fukui, 1966):

$$F_{\min} = 1 + K_f 2\pi f C_{gs} \sqrt{\frac{(R_g + R_s + R_i)}{g_m}}, \qquad (28)$$

where C_{gs}, g_m, R_g, R_s, and R_i are the gate-to-source capacitance, transconductance, gate resistance, source contact resistance, and channel resistance, respectively. The constant

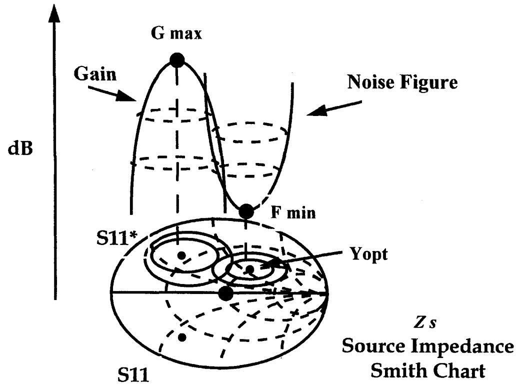

FIG. 7. Illustration of the influence of the input termination impedance on the microwave gain and noise figure of a transistor. Usually, the optimal impedance for gain is not the same as the optimal impedance for the lowest noise figure.

K_f is Fukui's constant, which has a value between 1 and 3. The value of Fukui's constant reflects the electron temperature of the drain current. It depends on the device structure as well as the carrier-transport dynamics in the channel.

The minimum-noise figure F_{min} is also a function of frequency and bias current, which is low at lower drain current and voltage. To build a low-noise microwave FET, it is important to keep the gate length short for small C_{gs}, to minimize the parasitic resistance for lower thermal noise, and to select a channel material and device structure to increase g_m and lower the electron temperature. In general, K_f is approximately 3 for the silicon *n*-MOSFET, 2.2 for a GaAs *n*-MESFET, 1.7 for GaAs HEMTs, and 1.1 for an InP *n*-HEMT. Nevertheless, the K_f factor varies quite a bit from device to device even with the same material system depending on the device structure and bias condition.

The low-frequency noise fluctuations in a FET channel are much worse than in the junction bipolar transistor. The large amount of low-frequency fluctuation, which is well below tens of megahertz, comes from the trapping and detrapping process from the deep states in substrate and surface, usually in the $1/f$ or Lorentzian form. This low-frequency noise current would significantly deteriorate the frequency stability, or the phase noise performance, of a microwave FET oscillator. To reduce the low-frequency noise current, a better surface passivation layer and better buffer layer (Hughes *et al.*, 1987), such as a superlattice or a *p-n* junction, should be introduced to isolate the FET channel from these noise sources.

3.3 Microwave Metal-Oxide-Semiconductor Field-Effect Transistors (MOSFETs)

Most of the high-speed MOS processes for digital integrated circuits use polycrystalline silicon gate electrode for logic gates with very small gate width of tens of microns. However, a large gate periphery, in the order of hundreds of microns, is needed to produce enough transconductance for most microwave applications. The large gate resistance from polysilicon material degrades the noise and power gain significantly. For microwave applications, low-resistance silicide and metal films are used for the gate electrode material.

The cross section of a typical microwave MOSFET is shown in Fig. 8. The lightly doped drain contact region, L_r, determines the breakdown voltage of the transistor. The source is connected to the substrate ground plane through a deep P^+-implanted sinker to reduce the inductance of the source contact. The turnon knee voltage at a high drain current is determined by the gate length, L_g. For applications requiring high power-added efficiency or running at low supply voltage, a submicron gate is needed to reduce this turnon voltage (Yoshida and Katsueda, 1995).

Because of the low mobility associated with a silicon MOSFET, a large gate perimeter is needed to obtain high voltage gain and large power density. The large transistor dimension suffers high microwave loss from interconnection metallization and large cross talk between devices fabricated on a highly conductive silicon substrate. The use of silicon-on-insulator (SOI) technology, where a very thin silicon epitaxial layer is formed on top of an insulating dielectric film or substrate such as sapphire, greatly reduces these parasitics. Submicron-gate microwave CMOS/SOI FETs have been demonstrated with excellent noise characteristics (de la Houssaye *et al.*, 1995). However, the defect density of the silicon epilayer on the insulators is still high (in the order of 10^6/cm^2 today) for practical integrated-circuit applications.

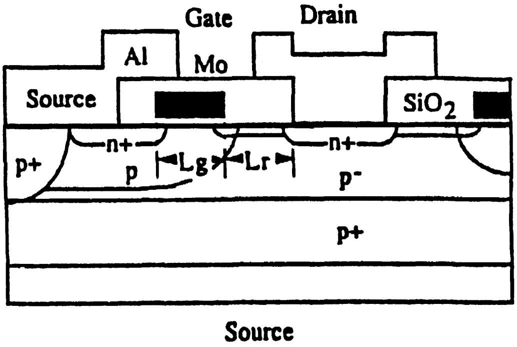

FIG. 8. Cross section of a microwave silicon MOSFET.

3.4 Microwave Schottky Metal-Semiconductor Field-Effect Transistors (MESFETs)

Since the 1970s, MESFETs (metal-semiconductor field-effect transistors) have dominated microwave transistors beyond 1 GHz because they are implemented with the high electron mobility and saturation velocity of the GaAs compound semiconductor. It is difficult to form a stable insulating layer on the GaAs surface to build GaAs MOSFETs or MISFETs. Instead, a Schottky junction is used to control the channel current by placing the metal electrode directly on a thin, doped semiconductor layer. The conductance of the channel is then modulated by the depth of depletion region formed underneath the Schottky metal. As more reverse bias is applied to the gate, eventually the whole channel is depleted and the transistor is turned off. The frequency response of a MESFET is determined, to the first order, by the single pole formed by this depletion capacitance C_{gs} and the transconductance g_m, and its rolloff frequency is $f_T = g_m/2\pi C_{gs}$. For MESFET transistors designed for high frequency, it is important to keep a small C_{gs} by shrinking the gate length and to obtain high dc transconductance by utilizing a thin and highly conductive semiconductor with high mobility and saturation velocity. The overall frequency response of the MESFETs would be further degraded by other parasitic elements, such as gate resistance, gate–drain capacitance, substrate capacitance, etc.

For most MESFETs made of compound semiconductors using GaAs or InGaAs, the channel current is limited by the saturated electron velocity rather than the low-field mobility as in silicon. In these GaAs FETs, electrons are injected from the source contact and accelerated under low electrical field in the channel toward the gate depletion region in the central valley of the conduction band with a high mobility. These electrons are then modulated by the gate potential and accelerated by the high electric field along the boundary of the gate–drain depletion region. However, as electrons acquire more energy, most of them are no longer confined in the high-mobility main valley (Γ valley) in the conduction band; instead, they are scattered into low-mobility satellite valleys (L and X valleys). This happens under a small drain bias, typically less than a volt. These electrons continue to drift toward the drain contact with the saturated velocity.

The threshold voltage V_T, where the channel current is cut off, is

$$V_T = \phi_B - \frac{1}{\epsilon}\iint \rho(x)d^2x \tag{29}$$

with a maximum channel current I_{max} of

$$I(V_{GS}) = qZv_s \int_{VT}^{VGS} \rho(x)dx, \tag{30}$$

where the $\rho(x)$ is the channel doping profile, ϕ_B is the Schottky barrier height, Z is the total gate width, and v_s is the saturated electron velocity. The channel doping, $\rho(x)$, can be introduced either by directly implanting the dopants into the semi-insulating substrate or by using epitaxial growth techniques. The threshold voltage can be made either positive for enhancement-mode devices or negative for depletion-mode devices by adjusting the doping and thickness of the channel material. However, significant gate current and even catastrophic damage will occur when the gate electrode is forward biased beyond the turnon voltage of the Schottky diode.

Figure 9 shows a typical GaAs MESFET structure with a double-recessed channel region for high-power applications. A short gate length and low parasitic source and channel resistance are needed to produce high power gain at microwave frequency. The metal gate is usually placed asymmetrically between the source and drain, which is

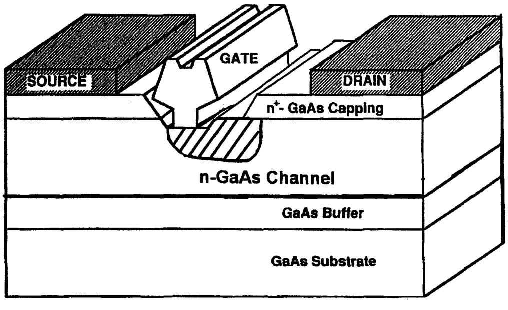

FIG. 9. Cross section of a microwave GaAs MESFET.

closer to the source to reduce the source resistance and away from the drain to obtain higher breakdown voltage. The source–drain breakdown voltage BV_{DS} equals $BV_{\mathrm{GS}} - V_{\mathrm{T}}$, where BV_{GS} is the gate–drain breakdown voltage and V_{T} is the threshold voltage. To increase the breakdown voltage, a low-doped semiconductor is placed under the gate metal, and a fully depleted and well-passivated surface layer is placed on top of the channel between the gate and drain contacts.

The transconductance $g_{\mathrm{m}} = \Delta I_{\mathrm{DS}}/\Delta V_{\mathrm{GS}}$ usually increases with the gate voltage. The nonuniform transconductance will affect the linearity of a high-power amplifier under large signal modulation. Special doping profiles, such as exponential or pulse shape, can be prepared by modern epitaxial growth techniques to obtain a constant transconductance over a wide bias range.

Combining the self-aligned implantation and a T-shaped metal gate, both the source resistance R_{s} and gate resistance R_{g} of a MESFET with a submicron gate length can be made very small. The MESFET structure will have a current-gain cutoff frequency f_{T} as high as that of a HEMT structure assuming the same gate length and saturated carrier velocity in the channel (Moll *et al.*, 1994) This also means that an optimized GaAs MESFET structure would deliver the similar minimum noise figure as a HEMT structure with the same gate length, gate resistance, source resistance, and transconductance, from Fukui's noise equation (Feng and Laskar, 1993).

However, for the same noise performance, the power gain of a GaAs HEMT or the power-gain cutoff frequency f_{max} is usually higher than that of a GaAs MESFET of the same gate length. The output resistance of a HEMT is also higher than that of a MESFET with similar gate length. This leads to a lower overall noise figure of a cascaded multistage amplifier where high gain in the first and second stages is essential to reduce effectively the noise generated from the later stages as shown in Eq. (4). This is the reason why HEMTs are widely used in commercial multistage microwave low-noise amplifiers to provide high gain with lower noise figure and fewer stages.

The output resistance of a short-channel MESFET is lowered by the injection of high-energy carriers into the buffer and substrate from the doped channel. For MESFETs with short gate length, a significant amount of electrons are injected into the substrate by the high electric field under the gate. The hot carriers in the substrate have a long transit delay, which is dephased from the carriers staying in the channel, and decreases the output resistance of the transistor. The low output resistance or high output conductance would reduce the microwave gain and increase the drain noise current. To reduce the output conductance of the short-channel MESFETs, the thickness of the doped channel should be reduced and a high–potential-barrier buffer layer such as a heterojunction or *p-n* junction is usually added between channel and substrate. With the reduced channel thickness, a higher channel doping density is needed to maintain a high transconductance. This may present a practical problem in fabricating an ion-implanted structure with gate length below 0.3 μm. It is difficult to control the thin depth and high dosage under a low implantation energy. Therefore, most short-channel MESFETs utilize an epitaxial structure with a thin, highly doped channel and a heterojunction energy barrier at the channel–substrate interface. This is usually done by incorporating a "pseudomorphic" or "lattice-strained" structure with a small fraction of indium added into the channel. The compressive strain energy as well as the indium-rich alloy effectively increases both the energy separation of the high-mobility center valley and low-mobility satellite valleys of the channel and the channel–substrate interface barrier. This results in a higher saturated carrier velocity and better output conductance, which have already been built into the quantum-well HEMT structure.

3.5 High-Electron-Mobility Transistors (HEMTs)

Esaki and Tsu (1969) first suggested the concept of utilizing a modulation doping technique to separate electrons spatially from their ionized parent donors by an abrupt heterojunction interface. Electrons from the donors in the wide-band-gap materials are transferred and confined in a two-dimensional potential well in the narrow-band-gap material forming a two-dimensional electron gas (2DEG). In the absence of scattering by the ionized donors, these electrons have high mobility and large carrier density confined within a thin layer. With the advancement of molecular-beam epitaxy (MBE) technology, this concept was first demonstrated by Stromer *et al.* (1979). In 1980, the first HEMT was demonstrated by Mimura *et al.* (1980). Combining the high electron mobility and the thin quantum-well structure for better charge confinement in a scaled FET, the HEMT has dominated low-noise applications from 10 to 100 GHz. When HEMTs are cooled down to cryogenic temperature, their microwave low-noise performance rivals that of microwave masers. Today, HEMTs have been used extensively in radars, satellite receivers, and high-speed optical-fiber communication systems.

Figure 10 shows a typical HEMT structure, and Fig. 11 shows the conduction-band diagram underneath the metal gate. By placing a metal gate on the wide-gap AlGaAs supply layer, the Schottky barrier depletes the doped AlGaAs and transfers the electrons into the GaAs quantum well to form the 2DEG. Other features of the HEMT active structure are the thin undoped AlGaAs "spacer" layer between the highly doped AlGaAs supply layer and the undoped GaAs channel, and the heavily doped GaAs "cap" region above the AlGaAs layers. The spacer layer further separates the 2DEG from ionized donors for high mobility at the cost of total charge transferred to the GaAs. The heavily doped GaAs cap layer facilitates formation of Ohmic contacts to the device.

For a typical heterointerface such as AlGaAs/GaAs, the conduction-band offset is about 0.3 eV and generally only two lower subbands are populated at room temperature. When the AlGaAs layer is totally depleted, the density of the 2DEG can be controlled by the gate bias as

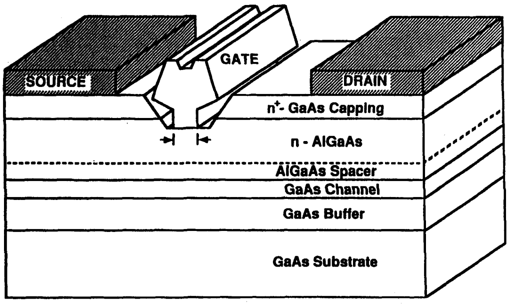

FIG. 10. Cross section of an AlGaAs/GaAs HEMT.

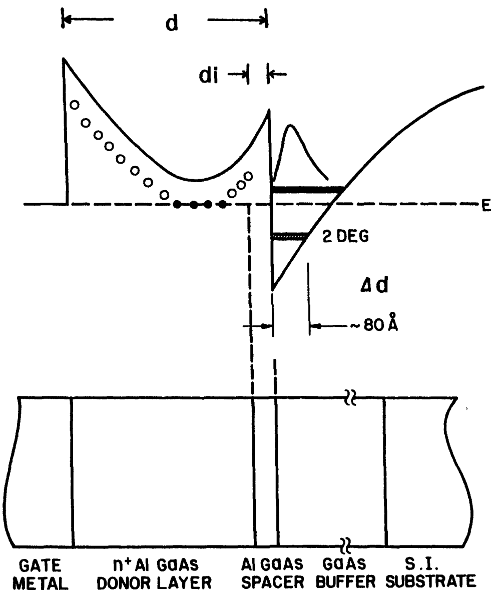

FIG. 11. Conduction-band diagram of an AlGaAs/GaAs HEMT at thermal equilibrium.

$$n_s = C_0(V_G - V_T)/q, \qquad (31)$$

where $C_0 = \epsilon\epsilon_0/(d + \Delta d)$ is the depletion capacitance between that gate and the 2DEG, $d + \Delta d$ is the average distance between the gate and the 2DEG, Vg is the gate bias, and V_T is the threshold voltage (Delagebeaudeuf and Linh, 1982). As the gate potential becomes more positive, the thickness of the depletion region is reduced, and eventually the low-mobility AlGaAs supply layer will start conducting and stop transferring electrons over the heterointerface. The maximum 2DEG density n_{s0} that can be provided by a heterointerface is approximated by

$$n_{s0} = \epsilon\epsilon_0(\Delta E_c - E_{F0})/(\alpha + q^2 d_i) \qquad (32)$$

through a self-consistent Poisson solver, where ΔE_c is the discontinuity of the conduction-band edge at the heterointerface, $\epsilon\epsilon_0$ is the specific dielectric constant of the supply layer, E_{F0} is the quasi-Fermi level in the 2DEG, d_i is the thickness of the spacer layer, and α is a constant.

The lateral-current transport mechanism in a HEMT from the source to drain can be modeled as a MESFET with two conductive layers, the low-mobility doped wide–band-gap supply layer and the high-mobility 2DEG in the undoped narrow-gap quantum well. It is important to design a HEMT structure such that all the lateral current flow is conducted by the high-mobility 2DEG layer. This is achieved by using a thin, heavily doped supply layer, which will be depleted by both the Schottky barrier potential and the potential spike from the conduction-band discontinuity. Under this condition, the maximum transconductance g_{m0} and maximum 2DEG channel current $I_{m,2DEG}$ are given by

$$g_{m0} = Zv_sC_0, \tag{33}$$

$$I_{m,2DEG} = qn_{s0}v_sZ, \tag{34}$$

where the average saturated velocity v_s is used for a HEMT with a total gate width of Z.

Because the carriers are traveling at the saturated velocity, the current-gain cutoff frequency is limited to $v_s/2\pi L_g$, which is similar to that of a MESFET of the same gate length. However, the microwave output conductance improved tremendously with the quantum confinement, and the high-frequency gain, or power-gain cutoff frequency, f_{max}, is higher than that of the conventional MESFET structure. The high mobility, thin supply layer, and quantum-confined layer structure are translated into low source resistance, high transconductance, and high output resistance for low-noise, high-gain, submicron-gate HEMTs. The processing steps for HEMTs are greatly simplified by the use of precise epitaxial growth techniques such as molecular-beam epitaxy (MBE) with an atomic-scale accuracy.

To enhance the current-handling capability of a microwave HEMT, it is important to select a heterostructure with a high conduction-band discontinuity and a narrow-gap material with high saturation velocity. A high conduction-band discontinuity helps suppress the back injection, or the real-space transfer, of the energetic 2DEG carriers into the supply layer by the electric field in the channel. The conduction-band discontinuity can be increased by introducing a lattice-strained (or pseudomorphic) quantum well at the heterointerface, such as adding a small fraction of indium into the GaAs quantum well. However, there is a limit on how much lattice mismatch the substrate material may tolerate, which is governed by the critical thickness of the strained epilayer, before the epilayer becomes unstable and relaxed. Another approach to increase the current-driving capacity is to incorporate multiple heterointerfaces. Special attention is needed in designing these structures to minimize the parallel conduction channels in the low-mobility supply layers.

To achieve good microwave performance, good process practice has to be exercised to reduce the gate resistance and source contact resistance. Usually, a T-shaped gate metal profile is fabricated with a submicron-gate footprint on the semiconductor and a thick and wide top-metal layer on the top to reduce the gate resistance. The source contact resistance is obtained by using a very heavily doped and sometimes lattice-strained narrow-gap cap layer. For higher gate-to-drain breakdown voltage, a pulsed-doped supply layer replaces the uniformly doped supply layer by an undoped wide-gap layer under the Schottky gate and a very highly doped thin supply layer in a few atomic layers.

Operating these HEMTs at cryogenic temperature enhances the mobility of the 2DEG and decreases the source resistance R_s by reducing the polar optical-phonon scattering. The noise figure is then improved significantly by the reduction of Johnson noise and by the increase of transconductance. The carrier confinement of the 2DEG is also improved at low temperature. However, the deep donors in the wide-gap supply layer may be frozen out at cryogenic temperature, and a LED is usually needed to activate these donors. Some emerging new material systems, such as AlInAs/InGaAs *n*-channel HEMTs on the InP substrate and Si/SiGe *p*-channel HEMTS on the silicon substrate, have demonstrated excellent microwave performance. Table 2 shows semiconductor material systems commonly used to fabricate HEMTs.

4. MONOLITHIC MICROWAVE INTEGRATED CIRCUITS (MMICs)

Mehal and Wacker from Texas Instruments demonstrated the first GaAs MMIC in 1968 (Mehal and Wacker, 1968). By the mid-

Table 2. Commonly used heterostructures for microwave high electron mobility transistors (HEMTs).

Substrate	Supply layer	2DEG
Silicon	Si	$Si_{(1-x)}Ge_x$
GaAs	AlGaAs	GaAs
	InGaP	$In_xGa_{1-x}As$
	AlGaP	
	ZnSe	
InP	AlInAs	InGaAs
	InP	
	InGaP	
GaSb	AlSb	InAs
SiC	AlGaN	GaInN
Sapphire	GaN	

1970s, a great advancement of GaAs MESFET technology had been achieved, and researchers began to build MMICs. The MMIC is a microwave-integrated circuit that is built on a single semiconductor substrate and composed of both active microwave transistors and passive circuit elements such as capacitors, inductors, transmission-line elements, and bias and interconnection lines. During the cold war era through the end of 1980s, the early drive on MMICs was mainly by the instrumentation and military applications, which demand high performance, small size, and weight reduction. Not until the early 1990s were the real advantages of MMICs, in terms of low cost, high performance, small size, high reliability, reproducibility, and tolerance to environment changes, realized through mass consumer applications. Today, MMICs have been used to implement applications from simple microwave-radio building blocks such as low-noise amplifiers, oscillators, and power amplifiers to multifunctional subsystems such as integrated-frequency down-converters that integrate low-noise amplifiers, local oscillators, mixers, and intermediate-frequency amplifier all on a single chip. Many GaAs MMICs and silicon MMICs can be found in many popular items such as cordless phones, wireless mobile phones, and direct-broadcast satellite (DBS) receivers.

Figure 12 shows a microphotograph of a Lucent Technologies' broad-band dc–20 GHz 4-stage distributed amplifier using 0.25-μm HEMTs as active elements and on-chip integrated resistors, inductors, and capacitors to form passive low-loss matching networks.

4.1 Semiconductor Substrate for MMICs

Many semiconductor materials, such as gallium arsenide (GaAs), silicon (Si), and indium phosphide (InP) substrate, which are

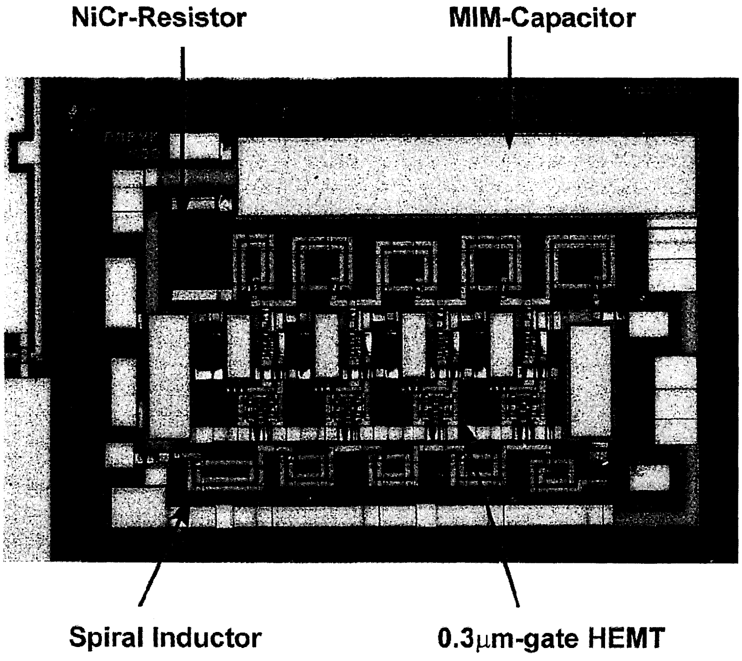

FIG. 12. Photomicrograph of a dc–15-GHz HEMT traveling-wave amplifier integrated circuit. Active HEMTs, metal-insulator-metal (MIM) capacitors, spiral inductors, and thin-film resistors are fabricated on the same chip. The dimensions of the circuit are 1 × 1.5 mm^2.

used to fabricate active microwave transistors, can be used to integrate passive elements as well. However, the Ohmic loss and dielectric loss of the substrate may impede the performance of low-loss transmission lines and high-quality inductors at microwave frequency. MMICs with operating frequency more than 100 GHz have been realized on semi-insulating GaAs and InP substrates. However, the high substrate leakage limits the operating frequency of silicon MMICs below 5 GHz using conventional highly doped substrates and less than 50 GHz on high-resistivity bulk silicon substrates or silicon-on-insulator substrates.

To control the characteristic impedance of the transmission lines on the substrate and improve the heat dissipation, the substrates are usually thinned down to a thickness of 50 to 100 μm with the back side metallized as heat sink and ground plane. To reduce the inductive interconnections to ground terminals, via holes are formed by selectively etching holes through the substrate and backfilled by metals to connect the top electrodes to the ground plane on the back side of the wafer.

4.2 Matching Networks

The overall circuit response of a MMIC is determined by both the microwave transistors and other circuit elements. As discussed in the previous sections, the gain of a microwave transistor decreases at high frequency. The noise figure and output microwave power of a microwave transistor are enhanced when specific input and output impedance are presented to its terminals. The optimal value of the impedance that a transistor would like to see may be quite different from the impedance provided by a microwave source and the load. Matching networks are introduced to transform the input impedance of a transistor, with no or little loss, to the source impedance of a generator, and its output impedance to a specific load (maybe the input of the next transistor). The frequency-dependent gain of the transistor is equalized by using a combination of high-pass and low-pass networks throughout the circuit. The matching network is also critical to eliminate the out-of-band instability and spurious response of the circuits by lowering the gain and impedance level outside the signal bandwidth with a frequency-selective filter.

The matching network may be implemented with both active microwave transistors and passive elements. The transistor is a very good broad-band impedance transformer up to about one-tenth of its cutoff frequency. When operated near its cutoff frequency, the frequency of this active transformer is loaded by many parasitic elements and is very difficult to handle. The use of the active matching network is further limited by the excessive noise generation and dc power consumption of transistors.

Most of the matching networks at microwave frequency are realized with passive circuit elements. For microwave frequencies below 10 GHz, low-loss passive lump elements such as spiral inductors, parallel-plate metal-insulator-metal (MIM) capacitors, and resistors can be used when their physical dimensions are much smaller than a quarter of a wavelength and the distributed effect is not dominating. Beyond 10 GHz, the use of distributed transmission-line circuit elements is necessary.

4.3 Passive Lump Elements

The quality factor (Q), that is, the ratio of the microwave energy stored to the energy dissipated in the network, directly affects the bandwidth, noise, and power performance of the MMIC. It is very important to select high-Q passive elements to implement the input, output, and interstage matching networks. For the lower microwave frequencies, less than 15 GHz, matching networks may be implemented using lumped passive elements such as capacitors, inductors, and resistors. Beyond 15 GHz, it is very difficult to utilize these lump elements because of the self-resonance, caused by many parasitics from the distributed effects and substrate ground plane.

Metal–insulator–metal (MIM) capacitors are commonly realized in the MMIC process for by-pass or coupling purposes. They use two metal layers separated by a thin dielectric material, typically silicon dioxide (SiO_s) or silicon nitride (Si_3N_4) with a thickness about 200 nm. The Q factor of a capacitor is

$$Q_C = 1/2\pi fRC, \qquad (35)$$

where f is the operating frequency, C is the capacitance, and R is the loss of the dielectric and metal plates. The use of MIM capacitors at high frequency is limited by the self-resonant frequency of about 12 GHz. At high frequency, the distributed effects are present on top and bottom electrodes. The self-resonant frequency is determined by these inductive elements and the distributed parallel-plate capacitance. For frequency beyond 10 GHz, capacitors are realized with interdigitated fingers or a small gap between two transmission lines.

Inductors are realized by multiturn spirals or rings. The inductance is determined by the self-inductance of each line segment and the mutual inductance among line segments. The Q factor of an inductor is

$$Q_L = 2\pi f L/R \tag{36}$$

where f is the operating frequency, L is the inductance, and R is the loss from metal lines and substrate. To build a low-loss inductor, it is essential to reduce the substrate loss and metal loss. The substrate loss may be reduced by using high-resistivity substrates, while the metal loss is dominated by the width, length, and conductivity of metal lines. At low microwave frequency, around 1 GHz, the metal loss is dominated by the skin depth of the metal, typically about 1 μm for gold and aluminum. It is important to deposit a very thick metal layer, at least three times the skin depth, to minimize the conductor loss for inductors and MIM capacitors at low microwave frequency. The self-resonant frequency of a spiral inductor is limited by the capacitive elements from the ground plane usually to 15 GHz on a semi-insulating GaAs substrate or 5 GHz on a doped bulk silicon substrate.

The resistors in a MMIC are realized by using a deposited thin resistive film such as nickel-chrome (NiCr) and tantalum nitride (TaN) or the doped semiconductor layer. The resistance R is determined by

$$R = \rho L/W, \tag{37}$$

where ρ is the resistivity of the film, L is the length, and W is the width. The deposited resistors are accurate and insensitive to temperature variation but low in resistivity and occupying a large chip area. The semiconductor film provides higher resistivity, but it is sensitive to temperature variations and parasitic junction capacitance from the buffer layer or substrate.

4.4 Passive Distributed Elements

Both inductive and capacitive elements at microwave frequency are realized using short transmission-line sections. It is also easy to implement any impedance matrices by adjusting length (L), width (W), and the characteristic impedance (Z_0) of the integrated transmission lines. Open, short, and impedance transformation can easily be implemented with the help of quarter-wavelength transmission lines. Transmission lines are very stable over the temperature variations, and their loss is very low. Disadvantages of the transmission-line matching network are its large size, frequency sensitivity, and narrow bandwidth.

Commonly used transmission lines for MMICs are microstrip (MS) and coplanar transmission lines. Both of them have very low propagation loss and frequency dispersion on a semi-insulating substrate, with cutoff frequency beyond 100 GHz. The microstrip transmission line consists of a top metal line on a substrate with a back-side ground plane. The characteristic impedance of a microstrip transmission line is determined by the width of the stripe and the thickness and dielectric constant of the substrate. The dependence of characteristic impedance on the width of a microstrip line on a GaAs substrate is illustrated in Fig. 13. The use of microstrip lines also requires substrate via holes to provide low-inductance ground connections to the circuit elements on the top of the wafer.

The coplanar waveguide (CPW) transmission line is a true TEM waveguide that consists of a central signal metal line separated by a gap from two adjacent ground planes on top of a semi-insulating semiconductor substrate. The characteristic impedance of the CPW lines is controlled by the width of the central conductor, the width of the gap, the substrate material, and the distance of the back-side ground plane. Usually, MMICs using CPW matching networks may avoid via holes and thinning of the substrate, which add extra processing cost. Higher characteristics can be realized on CPW lines than on

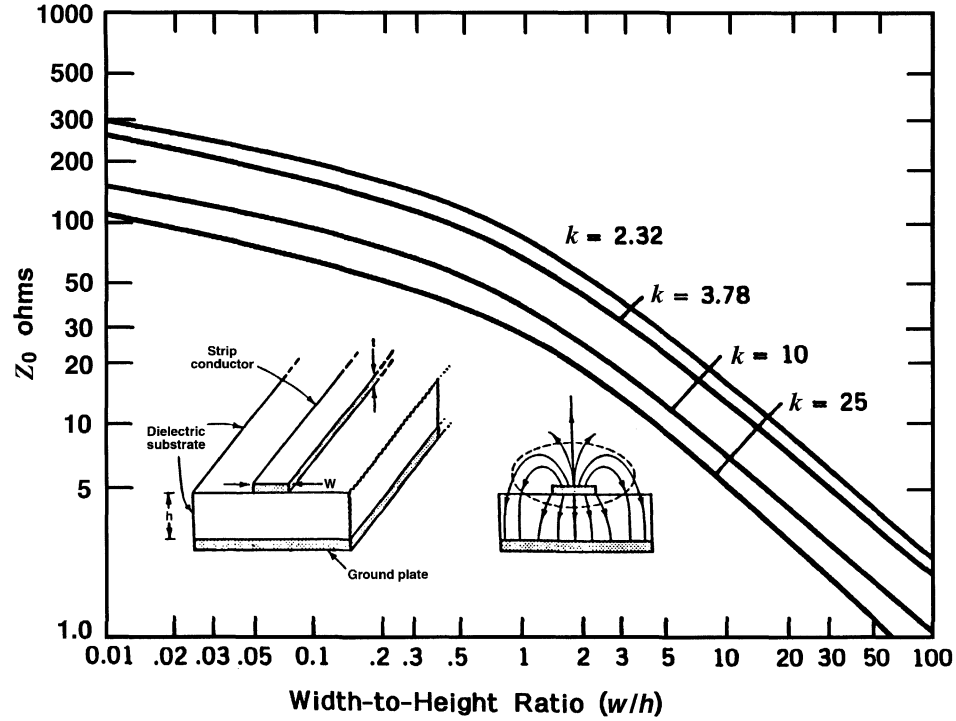

FIG. 13. Characteristic impedance of microstrip transmission line as a function of the width/height ratio, *w*/*h*, and effective dielectric constant, *k*.

MS lines. However, the propagation loss is slightly higher for CPW because of the strong coupling effects on the edges of the central conductor to ground planes. Additional connections between two ground planes are also needed along the transmission line to avoid possible unbalanced potentials between them.

4.5 Computer-Aided Design of MMICs

As the dimensions of circuit elements integrated on a MMIC are comparable to a fraction of the wavelength, many fringing fields will be produced near the intersections where different transmission lines, bias lines, and devices are connected together. These small field disruptions may affect the frequency response of matching networks and have to be accommodated by transmission-line design parameters. This is usually done by using a two-dimensional (2D) or three-dimensional (3D) finite-element electromagnetic (EM) wave computer simulator to predicate their behavior. Many high-performance circuits also require taking the large signal-time transient into the design process with the passive elements.

Thanks to the investments during the cold-war era in the late 1980s, many of these sophisticated time- and frequency-domain tools such as SPICE for time domain, harmonic balance simulators for frequency domain, and 3D EM simulator for complete MMIC layout are readily available and highly integrated into other circuit simulators. The signal integrity, interference, and cross talk of the microwave signals on the chip and between MMICs may now be analyzed and optimized. These CAD tools greatly eliminate much of the mystery of microwave and millimeter-wave design and shorten the design cycles significantly. These tools help realizing many sophisticated, multifunctional microwave integrated circuits and subsystems with high performance and low production cost.

GLOSSARY

Ballast Resistor: The resistor placed in series with each emitter when massive parallel emitter-base junctions are assembled to build a power bipolar transistor. This resistor is needed to desensitize the variations in the turn-on potential and thermal distribution among transistor cells.

Coplanar Waveguide Transmission Lines (CPW): A transmission-line structure that is a true TEM (transverse electrical and magnetic wave) waveguide, consisting of a central metal signal line separated by a gap from two adjacent ground planes on top of a semi-insulating substrate.

Equivalent-Circuit Model: The dynamics of physical carrier-transport properties of a transistor is simplified and modeled by a set of simple linear lump elements such as resistors, capacitors, and dependent current and voltage sources.

Heterojunction Bipolar Transistor (HBT): A bipolar transistor utilizing an emitter composed of a lighter doped and wider band-gap material than that in the base. A high current gain is realized by using the heterojunction barrier at the emitter–base interface to block the back-injected majority carriers from the base and subsequently to reduce the recombination current.

High–Electron-Mobility Transistor (HEMT): A transistor structure that utilizes a heterostructure to accumulate two-dimensional electron gas in the lightly doped high-mobility, narrow–band-gap channel and locate ionized donors from the highly doped wide–band-gap layer. The separation of electrons and ionized donors results in high electron mobility and low access resistance.

Johnson Noise: See **Thermal Noise.**

Matching Network: A circuit network that matches the impedance between its source and its load. It may be implemented with either active or passive elements.

Maximum Unilateral Power-Gain Cutoff Frequency (f_{max}): The frequency at which the unilateral power gain U decreases to unity, when both input and output of the transistor are matched and internal feedback is also nullified by an external tuning element. This is also known as the **maximum oscillation frequency** at which a microwave oscillator circuit can operate using this transistor.

Metal-Insulator-Metal (MIM) Capacitor: An integrated capacitor structure on a semiconductor that uses two metal layers separated by a thin dielectric material, usually silicon dioxide (SiO_2) or silicon nitride (Si_3N_4).

Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET): A transistor structure that utilizes the electrical field between the metal gate electrode and the semiconductor surface, separated by a thin oxide layer, to modulate the conductivity of the semiconductor channel.

Metal-Semiconductor Field-Effect Transistor (MESFET): A transistor structure that utilizes the Schottky barrier of the metal–semiconductor interface to modulate the conductivity of the semiconductor channel.

Microstrip Transmission Lines: A transmission-line structure that is a quasi-TEM (transverse electrical and magnetic wave) waveguide, consisting of a metal signal line on top of a semi-insulating substrate with a back-side ground plane.

Microwave: Any electromagnetic wave in a frequency range commonly regarded as extending from 300 MHz to 300 GHz, covering the spectrum lying between the radio frequency and the far infrared.

Millimeter Wave: The part of the spectrum higher than 30 GHz is regarded as the millimeter-wave band, as the corresponding wavelength in the free space is comparable or shorter than a millimeter.

Minimum Noise Figure (F_{min}): The noise figure of a transistor reaches its minimum value F_{min} when its input is connected to an optimal admittance set by the quantum-mechanical fluctuation and dissipation theory.

Molecular-Beam Epitaxy (MBE): A nonequilibrium epitaxial growth technique utilizing molecular beams in a high-vacuum chamber to form thin films lattice-matched on single-crystal semiconductor substrates.

Monolithic Microwave Integrated Circuit (MMIC): A microwave circuit that integrates both active and passive circuit elements on a single semiconductor substrate.

Noise Figure (F): A measure of the output signal quality, defined as the ratio of the input signal-to-noise ratio to that of the output.

Organometallic Vapor-Phase Epitaxy (OMVPE): An equilibrium epitaxial growth technique that utilizes atoms from organometallic vapors in a high-temperature reactor to form thin films lattice-matched on single-crystal semiconductor substrates.

Quantum-Mechanical Fluctuation and Dissipation Theory: A theory stating that the internally generated noise will be minimized once these noise sources can be dissipated through an optimal external impedance.

Scattering Parameters: A representation of a two-port network that utilizes incident and reflected power wave forms from a known termination reference impedance to characterize the transistor at microwave frequency.

Shot Noise: A broad-band noise source produced by the random scattering of carriers passing through a semiconductor junction. Its magnitude is proportional to the current. The upper frequency limit of the shot noise is determined by the mean scattering time of the charged carrier in the depletion region.

Short-Circuit Current-Gain Cutoff Frequency (f_T): The highest frequency at which the current gain of a microwave transistor can reach unity with a short-circuit load.

Thermal (or Johnson) Noise: A broad-band noise generated by a resistive element. Its magnitude depends on the operating temperature and the value of the resistance.

Two-Dimensional Electronic Gas (2DEG): Electrons that are confined in a two-dimensional quantum well to form an electron gas. The quantum well can be formed by using a thin-surface inversion layer on a metal–oxide–semiconductor interface or a heterointerface between two semiconductors with different band-gap energy.

Unilateral Power Gain (U): The maximum power gain that a transistor can deliver at a given frequency, when the input and output of the transistor are connected to conjugate-matched impedance and the internal feedback in the transistor is also nullified by an external feedback circuit for maximum power transfer.

Works Cited

Chen, Y. K., Nottenburg, R., Panish, M. B., Hamm, R. A., Humphrey, D. A. (1989), *IEEE Electron Dev. Lett.* **10,** 267–269.

Crabbe, E., Meyerson, B., Harame, D., Stork, J. *et al.* (1993), *IEEE Trans. Electron Dev.* **40,** 2100–2101.

de la Houssaye, P. R., Chang, C. E., Offord, B., Imthurn, G. *et al.* (1995), *IEEE Electron Dev. Lett.* **16,** 289–292.

Delagebeaudeuf, D., Linh, N. (1982), *IEEE Trans. Electron Dev.* **29,** 955–960.

Early, J. M. (1952), *Proc. IRE* **40,** 1401–1406.

Esaki, L., Tsu, R. (1969), *IBM Res. Rep.* RC-2418.

Feng, M., Laskar, J. (1993), *IEEE Trans. Electron Dev.* **40,** 9–17.

Fukui, H. (1966), *IEEE Trans. Electron Dev.* **13,** 329–341.

Hawkins, R. J. (1977), *Solid-State Electron.* **20,** 191–196.

Hughes, B., Fernandez, N., Gladstone, J. (1987), *IEEE Trans. Electron Dev.* **34,** 733–741.

Ishibashi, T., Yamauchi, Y. (1988), *IEEE Trans. Electron Dev.* **35,** 401–404.

Kromer, H. (1982), *Proc. IEEE* **70,** 13–25.

Mehal, E., Wacker, R. (1968), *IEEE Microwave Theory Tech.* **MTT-16,** 451–454.

Mimura, T., Hiyamizu, S., Fujii, T., Nanbu, K. (1980), *Jpn. J. Appl. Phys.* **19,** L. 225–227.

Moll, N., Feng, M., Laskar, J. (1994), *IEEE Trans. Electron Dev.* **41,** 115–118.

Nanba, M., Uchino, T., Kondo, M. Nakamura, T. *et al.* (1993), *IEEE Trans. Electron Dev.* **40,** 1563–1565.

Smith, P. M., Liu, S.-M.J., Kao, M.-Y., Ho, P. *et al.* (1995), *IEEE Microwave Guided Wave Lett.* **5,** 230–232.

Stromer, H., Dingle, R., Gossard, A., Wiegman, W., Sturge, M. (1979), *J. Vac. Technol.* **16,** 1517–1519.

Ugajin, M., Fossum, J. G. (1994), *IEEE Trans. Electron Dev.* **10,** 1796–1800.

Vendelin, G. D., Pavio, A. M., Rohde, U. L. (1990), *Microwave Circuit Design Using Linear and Nonlinear Techniques,* New York: Wiley, ISBN 0-471-60276-0, Chap. 1.

Yamada, H., Ohara, S., Iwai, T., Yamaguchi, Y. *et al.* (1996), *IEEE Trans. Microwave Theory Tech.* **44,** 2398–2402.

Yoshida, I., Katsueda, M. (1995), *IEICE Trans. Electron.* **E78-C,** 979–983.

Further Reading

Chang, M. F. (Ed.) (1996), *Current Trends in Heterojunction Bipolar Transistors,* Singapore: World Scientific Publishing Co. Pte. Ltd., ISBN 981-02-2097-9 (on heterojunction bipolar transistors).

Gonzalez, Guillermo (1984), *Microwave Transistor Amplifiers Analysis and Design,* New Jersey: Prentice-Hall Inc., ISBN 0-13-581646-7 (on the design of microwave circuit design and matching network).

Ikeda, H., Ashikawa, K., Urita (1980), "Power MOSFETs for Medium-Wave and Short-Wave Transmitters," *IEEE Trans. Electron Dev.* **ED-27** 330–334 (on microwave MOSFETs).

Kromer, H. K. (1982), "Heterostructure Bipolar Transistors and Integrated Circuits," *Proc. IEEE* **70,** 13–25 (on heterojunction bipolar transistors).

Mitsui, Y. (1988), *MMIC-Monolithic Microwave Integrated Circuits,* New York: Gordon and Breach Science Publishers, ISBN 2-88124-286-3 (on monolithic microwave integrated circuits).

Morkoc, H., Unlu, H., Ji, G. (1991), *Principles and Technology of MODFETs,* Chichester, U.K.: Wiley, ISBN 0-471-92932-8 (on high–electron-mobility transistors).

Nguyen, L. D., Larson, L. E., Mishra, U. K. (1992), "Ultra-High-Speed Modulation-Doped Field Effect Transistors: A Tutorial Review," *Proc. IEEE* **80,** 494–518 (on high–electron-mobility transistors).

Pozar, David (1990), *Microwave Engineering,* Reading, MA: Addison-Wesley Publishing Company, ISBN 0-201-50418-9 (on scattering parameters, passive distributed microwave components, and microwave engineering).

Pucel, Robert (Ed.) (1985), *Monolithic Microwave Integrated Circuits,* New York: IEEE Press, ISBN 0-87942-192-4 (on monolithic microwave integrated circuits).

Shah, Jagdeep (1992), *Hot Carriers in Semiconductor Nanostructures,* Boston: Academic, ISBN 0-12-638140-2 (on hot-carrier transport dynamics in semiconductors).

Sze, Simon (1981), *Physics of Semiconductor Devices,* New York: Wiley, ISBN 0-471-05661-8 (general reference on semiconductor devices).

Wang, Shyh (1989), *Fundamentals of Semiconductor Theory and Device Physics,* New Jersey: Prentice-Hall Inc., ISBN 0-13-344409-0 (general reference on semiconductor devices).

Yoshida, I., Katsueda, M. (1994), "Highly Effi-

TRANSISTORS, POWER

B. Jayant Baliga, *Power Semiconductor Research Center, North Carolina State University, Centennial Campus, Raleigh, North Carolina, U.S.A.*

Introduction 187
1. **Ideal Power-Transistor Characteristics** 187
2. **The Bipolar Power Transistor** 188
2.1 Current Gain 189
2.2 Voltage Saturation Region 190
2.3 Emitter-Current Crowding 190
2.4 Turnoff 191
2.5 Trends 191
3. **The Power Metal-Oxide-Semiconductor Field-Effect Transistor** 192
3.1 Basic Structure and Operation 192
3.2 Output Characteristics 193
3.3 Threshold Voltage 194
3.4 On-State Resistance 194
3.5 Input Capacitance 196
3.6 Cell Topology 197
3.7 Trends 197
4. **The Insulated-Gate Bipolar Transistor** 197
4.1 Device Structure and Operation 197
4.2 Blocking Characteristics 199
4.3 On-State Characteristics 199
4.4 Parasitic Thyristor Latchup 200
4.5 Switching Characteristics 201
4.6 Trade-Off Curve 201
4.7 Trends 202
5. **Conclusions** 202
Further Reading 203

INTRODUCTION

Power electronics is an essential part of a large variety of consumer and industrial systems. At lower power levels, it is used in power supplies (for computers and telecommunications), in lighting, in displays, and in automotive electronics. At medium power levels, it is used in motor control (for heating, ventilating, and air conditioning), in factory automation (or robotics), and in electric vehicle drives. At high power levels, it is used in traction (electric streetcars/trams and locomotives) and in electricity distribution (high-voltage dc transmission). In order to regulate the power flow, all the power electronic circuits must utilize a power switch. Power switches based upon semiconductor technology are recognized as the enabling technology with advances in their performance dictating improvements in the efficiency, size, and weight of all power-electronics–based systems. Power transistors are the predominant switch technology used in low- and medium-power applications. Their current- and voltage-handling capability (called ratings) is as yet insufficient for satisfying the requirements of the high-power applications, which rely on thyristors at present.

1. IDEAL POWER-TRANSISTOR CHARACTERISTICS

An *ideal* power transistor must be able to control the flow of power from a source to a load without internal power dissipation. The load may be inductive (such as motors, solenoids, etc.), resistive (such as heaters and lamp filaments), or capacitive (such as transducers and display elements). Most often, the power to the load is controlled by switching the device on a periodic basis to generate pulses of current flowing through the device. In this case, there must be no voltage drop across the device when it is conducting current in the on state, and there must be no current flowing through the device when it is supporting voltage in its off state. An *ideal* device must also be able to switch between these states with no power dissipation. This requires an instantaneous change in the voltage and current during the

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

switching transients. In the ideal case with 100% efficiency, the output power must be controlled with zero input power. In addition, the device should be able to control an arbitrarily high power level, which implies high voltage- and current-handling capability.

The output characteristics of an ideal three-terminal power switch are shown in Fig. 1. For general applicability, the ideal switch must be able to block and conduct current for both directions of current flow under the control of a signal applied to its gate electrode. This requires operation of the device in both the first quadrant and third quadrant of the *I–V* characteristics. In either quadrant, the device must be capable of supporting a very high voltage with zero current flow. These states are referred to as *forward blocking mode* and the *reverse blocking mode.* When switched to the conducting state, the device must be capable of carrying a very high current with zero voltage drop. This is referred to as the *on state.* In addition, it is desirable for the device to regulate the output current under the guidance of the gate signal. This requires a gate-controlled current saturation characteristic as shown in the figure. In order to maintain zero power losses in the input circuit, the gate control signal can be a current applied with zero voltage drop in the control circuit or a voltage signal with zero current flow in the control circuit. It has been found that it is easier to approach the ideal control case with the latter option.

Power transistors are usually designed to operate only in the first quadrant. They are not expected either to support or to carry substantial current in the fourth quadrant. They are used in power circuits with a dc power bus to supply the power to loads. The output characteristics of power transistors that are commercially available vary greatly from those outlined for an ideal device. The characteristics of specific devices are discussed in this article. In general, all the devices can support only a limited voltage and with a finite *leakage current* when blocking current flow. This result is power dissipation during the blocking mode, which increases rapidly with increasing temperature. This can lead to a thermal runaway problem unless the heat sink is designed to maintain the temperature below defined limits. Further, all the transistors exhibit a finite voltage drop during on-state current flow, leading to power loss. Because of the nonzero switching times for the devices, substantial power losses can also occur during the turn-on and turnoff transients, leading to large power losses at higher operating frequencies. Further, there can be substantial power dissipation in the input control circuits for some devices, which must be accounted for when optimizing the overall power system performance.

2. BIPOLAR POWER TRANSISTORS

The basic structure for a bipolar power transistor (see TRANSISTORS, BIPOLAR) is shown in Fig. 2. The bipolar power transis-

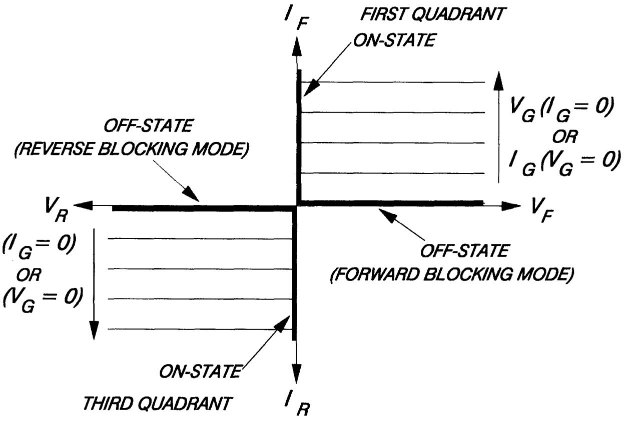

FIG. 1. Characteristics of an ideal power transistor. V_G, gate voltage; I_G, gate current; V_F (V_R), forward (reverse) voltage; I_F (I_R), forward (reverse) current.

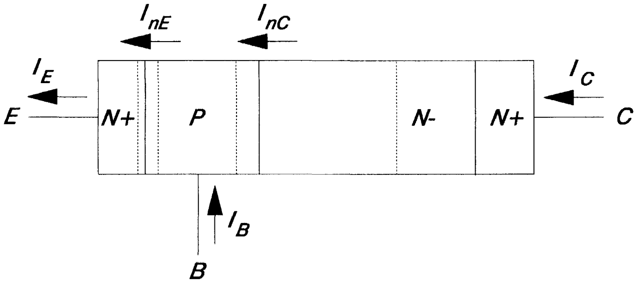

FIG. 2. Bipolar power transistor structure with its internal current components. *E*, emitter; *B*, base; *C*, collector; I_n, electron component of current.

tor is fundamentally a current-controlled three-terminal switch. It is usually operated in the common-emitter configuration with the collector current controlled by the magnitude and polarity of the input base current. For ease of control, it is desirable to operate the device with small base currents to control large collector currents. This allows operation of the device with a high power gain, which is of course also dependent upon the ratio of the collector and gate operating voltages. In some power circuits, it becomes necessary to operate the bipolar transistor in the common-base configuration as well. This operation results in a current gain of less than unity.

2.1 Current Gain

In both the common-base and common-emitter configurations, the collector current flow is controlled by the application of a voltage between the base and emitter terminals so as to forward bias the emitter–base junction. For the case of the *n-p-n* transistor, the forward bias results in the injection of electrons from the n^+ emitter into the *p* base region, and the injection of holes from the *p* base region into the emitter region. The emitter current is comprised of both of these components. However, the portion of the emitter current due to hole injection into the n^+ emitter region does not contribute to the collector current. The electrons injected into the *p* base region diffuse through it and are collected at the reverse-biased base–collector junction. Because of the finite minority-carrier lifetime in the *p* base region, some of the injected electrons are lost by recombination.

The current transport between the emitter and collector for an *n-p-n* transistor in the common-base configuration is shown in Fig. 2 with depletion boundaries indicated by the dotted lines and the internal electron current components indicated by the arrows. The emitter–base junction is assumed to be under forward bias while the collector–base junction is assumed to be under reverse bias. The current I_{nE} is the emitter-current component due to electron injection at the emitter–base junction. The current I_{nC} is the electron current at the collector–base junction. Using these current components, the common-base current gain can be written as

$$\alpha = \frac{\delta I_{\mathrm{C}}}{\delta I_{\mathrm{E}}} = \frac{\delta I_{\mathrm{nE}}}{\delta I_{\mathrm{E}}}\frac{\delta I_{\mathrm{nC}}}{\delta I_{\mathrm{nE}}}\frac{\delta I_{\mathrm{C}}}{\delta I_{\mathrm{nC}}}. \tag{1}$$

From this expression, it can be seen that the change in electron current flow between the emitter and collector can be partitioned into three portions. The first term, referred to as the *emitter injection efficiency* or *gamma*, is the fraction of the total emitter current that comprises the electron current. As the name implies, it is a measure of the ability of the emitter to inject electrons into the base region in an efficient manner. Since the emitter current due to hole injection into the emitter region does not contribute to the collector current, the injection efficiency term is always less than unity. In general, a higher injection efficiency can be obtained by increasing doping concentration of the n^+ emitter region relative to the *p* base region. However, it has been found that an optimum emitter doping is required to maximize the injection efficiency because of phenomena associated with high doping in semiconductors.

The second term in Eq. (1), referred to as the *base transport factor*, is a measure of the ability for electrons that are injected into the base from the emitter to reach the collector–base junction. In the presence of a finite recombination in the base region, some electrons will be lost during transport through the base region. Consequently, the base transport factor is always less than unity. A large base transport factor can be obtained by reducing the width of the p base region with high minority-carrier lifetime.

The third term in Eq. (1), referred to as the *collector efficiency,* is a measure of the ability to transport electrons through the collector region. In the case of a reverse-biased collector–base junction, a strong electric field is established within a depletion region at this junction. The electrons transported through the base region are swept out by this electric field into the collector region. At collector biases well below the avalanche breakdown voltage of the collector–base junction, this process occurs without loss of electrons. Consequently, the collector efficiency can be assumed to be equal to unity.

In the case of a power bipolar transistor, it is necessary to consider both the emitter injection efficiency and the base transport factor in analyzing the current gain. This is because, unlike signal transistors, power transistors must be designed with relatively large base thickness to prevent reach-through breakdown. Further, as a result of the high operating current densities in power transistors, high-level injection effects occur in the p base region that lead to a reduction in the current gain with increasing current density. This is referred to as the *Rittner effect.* Further, the effective base width becomes greater than the metallurgical base width at high current densities on account of an alteration in the electric-field distribution across the base–collector junction. This is called the *Kirk effect.* A detailed treatment of these effects is beyond the scope of this article. Their combined effect on the current gain, illustrated in Fig. 3, leads to a current gain of less than 10 at typical operating current densities.

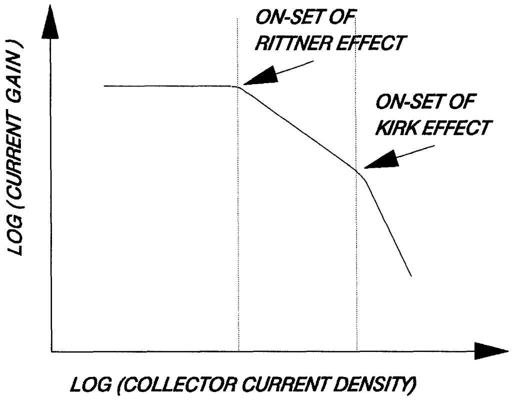

FIG. 3. Decrease in the current gain of a bipolar power transistor at higher current densities.

2.2 Voltage Saturation Region

The on-state portion of the current–voltage characteristic of the bipolar transistor is also referred to as the voltage saturation region. In this region of operation, the collector bias becomes less than the base drive voltage, which is about 1 V. Since both the emitter–base junction and the collector–base junction are then forward biased, the on-state voltage drop across the bipolar transistor in the voltage saturation region is less than the junction potential (typically 0.7 V). This is an advantage from the point of view of reducing the power dissipation in the on state. When the collector bias is sufficiently small, the base–collector junction operates with a sufficient forward bias to produce high-level injection (defined as a condition where the injected minority-carrier density exceeds the doping concentration) of carriers throughout the drift region. This regime of operation is called the *voltage saturation region.* As the collector bias is increased at any fixed base drive current, the forward bias across the base–collector junction is reduced. This results in a decrease in the injection of carriers into the drift region. Under these conditions, only a portion of the drift region near the base–collector junction operates under high-level injection while the rest of the drift region is not modulated. This regime of operation is called the *voltage quasisaturation region.* The different operating regions are illustrated in Fig. 4 at relatively small collector voltages close to the on-state voltage drop.

2.3 Emitter-Current Crowding

In the previous sections, the current flow through the bipolar power transistor was assumed to occur uniformly through its entire

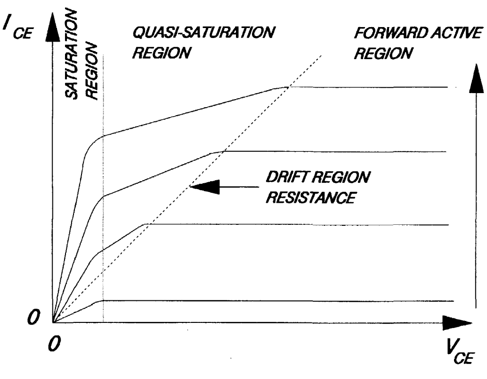

FIG. 4. Output characteristics of a bipolar power transistor. I_{CE}, collector–emitter current; V_{CE}, collector–emitter voltage.

emitter area. In an actual power transistor, the emitter and base contacts must be interdigitated as illustrated in Fig. 5. When the transistor is biased into its on state, a base drive current is applied at the base terminal. This current flows across the forward-biased emitter–base junction resulting in the injection of electrons from the emitter. Because of the interdigitation of the contacts, in order to forward bias the emitter–base junction at point A, the base current must flow through the base resistance R_{B}. This produces a voltage drop between points A and B, resulting in the forward bias across the emitter–base junction at point A being lower than that at point B. Consequently, the emitter injection at point A is smaller than that at point B, and the emitter current distribution is not uniform across the width of the emitter n^+ diffusion. Thus, it can be concluded that the current-carrying capability of the power bipolar transistor does not scale with increase in the emitter area. Instead, as a result of current crowding at the edges of the emitter, the current scales with increasing emitter periphery. Consequently, it is common practice to interdigitate the emitter and base contacts into a finger geometry. In spite of the need for interdigitation, the photolithography design rules for the fabrication of power bipolar transistors is not particularly demanding with state-of-the-art process technology because the emitter width (W_{E}) is relatively large (hundreds of microns in size).

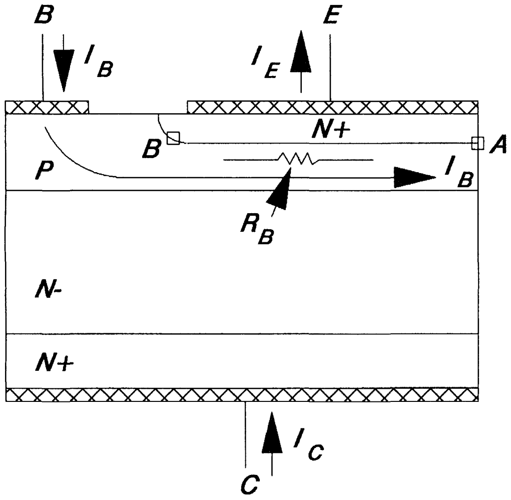

FIG. 5. Interdigitation of emitter and base contacts in a bipolar power transistor. R_{B}, base-region resistance to lateral current flow.

2.4 Turnoff

A high concentration of electrons and holes is present within the base and collector regions of the bipolar power transistor during the on state. This stored charge must be removed during the turnoff process to restore the device to its blocking state. One method for achieving this goal is to simply open circuit the base terminal, i.e., to stop providing the base drive current needed to keep the bipolar transistor in its on state. The stored charge will then decay by recombination at a rate dictated by the lifetime in the base and collector regions. Since this process is relatively slow and produces turnoff times that are too long for applications, it is preferable to apply a reverse bias to the base–emitter junction to enhance the turnoff process. The reverse base current flow extracts a significant fraction of the stored charge, which enables faster turnoff. However, this makes the control circuit for the bipolar power transistor complex and expensive because the reverse base current is typically one-fifth of the collector current.

2.5 Trends

The bipolar power transistor has been extensively used in power electronic systems since its initial commercialization in the

1950s. For these applications, devices have been developed with breakdown voltages up to 1500 V and with current-handling capability of several hundred amperes. However, the power bipolar transistor is a current-controlled device with a relatively low current gain (<10). Its application has required development of complex and expensive gate-drive circuits. For these reasons, the bipolar power transistor has been replaced by the power metal-oxide-semiconductor field-effect transistor (MOSFET) in applications where the operating voltages are below 200 V and by the insulated gate bipolar transistor (IGBT) in applications where the operating voltages range from 200 to 1500 V. The improvement in system performance achieved by the use of the MOS-gated devices in the 1980s has led to a lack of demand for further developments in bipolar power-transistor technology.

3. POWER METAL-OXIDE-SEMICONDUCTOR FIELD-EFFECT TRANSISTORS

In order to address the performance limitations experienced with power bipolar transistors, the power MOSFET was developed in the 1970s. These devices evolved from MOS integrated-circuit technology. In the power MOSFET (see TRANSISTORS, FIELD-EFFECT), the control signal is applied to a metal gate electrode that is separated from the semiconductor surface by an intervening insulator (typically silicon dioxide). The control signal required is essentially a bias voltage with no significant steady-state gate-current flow in either the on state or the off state. Even during the switching of the devices between these states, the gate current is small at typical operating frequencies (less than 100 kHz) because it serves only to charge and discharge the input gate capacitance. The high input impedance is a primary feature of the power MOSFET that greatly simplifies its gate-drive circuitry.

In contrast to the power bipolar transistor, the power MOSFET is a unipolar device. Current conduction occurs via transport of majority carriers in the drift region without the presence of minority-carrier injection required for bipolar transistor operation. No delays are observed due to storage or recombination of minority carriers in power MOSFETs during turnoff. Their inherent switching speed is orders of magnitude faster than for bipolar transistors. This feature is particularly attractive in circuits operating at high frequencies where switching power losses are dominant. Power MOSFETs have also been found to display an excellent safe operating area (i.e., they can withstand the simultaneous application of high current and voltage, for a short duration, without undergoing destructive failure). Further, these devices can be easily paralleled because the forward voltage drop of power MOSFETs increases with increasing temperature. This feature promotes an even current distribution between paralleled devices. These characteristics of power MOSFETs make them important candidates for many applications. They are being used in audio/rf circuits and in high-frequency inverters such as for switch-mode power supplies. Other applications for these devices are for lamp ballasts and motor control circuits.

3.1 Basic Structure and Operation

A cross section of the double-diffused MOSFET (DMOSFET) structure is shown in Fig. 6. This DMOS structure is fabricated by means of planar diffusion technology with a refractory gate, such as polysilicon, as a mask. In these devices, the *p* base region and

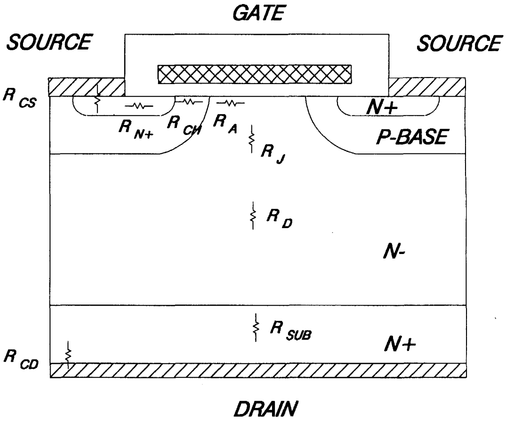

FIG. 6. Cross section of the DMOSFET structure with its internal resistances. R_{CH}, resistance of channel region; R_J, resistance of JFET region; R_{SUB}, resistance of substrate.

the n^+ source regions are diffused through a common window defined by the edge of the polysilicon gate. The name for this device is derived from this *double-diffusion process*. The p base region is driven in deeper than the n^+ source. The difference in the lateral diffusion between the p base and n^+ source regions defines the surface channel region.

In the power MOSFET structure, the p-n junction between the p base region and the n drift region provides the forward blocking capability. Note that the p base region is connected to the source metal by a break in the n^+ source diffusion. This is important to establishing a fixed potential to the p base region during device operation. If the gate electrode is externally shorted to the source, the surface of the p base region under the gate (i.e., the channel region) remains unmodulated at a carrier concentration determined by the doping level. When a positive drain voltage is now applied, it reverse biases the p-base/n-drift-region junction. This junction supports the drain voltage by the extension of a depletion layer on both sides. On account of the higher doping level of the p base region, the depletion layer extends primarily into the n drift region. A higher drain-blocking voltage capability requires a lower drift-region doping and a larger width, which results in a higher resistance within the device.

In order to carry current from drain to source in the power MOSFET, it is essential to form a conductive path extending between the n^+ source regions and the n drift region. This can be accomplished by applying a positive bias to the gate electrode. The gate bias modulates the conductivity of the channel region by the strong electric field created normal to the semiconductor surface through the oxide layer. The gate-induced electric field attracts electrons to the surface of the p base region under the gate. This field strength is sufficient to create a surface electron concentration that overcomes the p base doping. The resulting surface electron layer in the channel provides a conductive path between the n^+ source regions and the drift region. The application of a positive drain voltage now results in current flow between drain and source via the n drift region and the channel. This current flow is controlled by the resistance of these regions. Note that the current flow occurs solely by transport of majority carriers (electrons for n-channel devices) along a resistive path comprising the channel and drift regions. No minority-carrier transport is involved for the power MOSFET during current conduction in the on state.

In order to switch the power MOSFET into the off state, the gate bias voltage must be reduced to zero by externally shorting the gate electrode to the source electrode. When the gate voltage is removed, the electrons are no longer attracted to the channel and the conductive path from drain to source is broken. The power MOSFET then switches rapidly from the on state to the off state without any delays arising from minority-carrier storage and recombination as experienced in bipolar devices. The turnoff time is controlled by the rate of removal of the charge on the gate electrode because this charge determines the conductivity of the channel. Turnoff times of under 100 ns can be achieved with a moderate gate-drive current flow arising from the discharge of the input gate capacitance of the device.

3.2 Output Characteristics

The power MOSFET is capable of blocking voltage only in one quadrant. For n-channel structures, the devices are operated with a positive voltage applied to the drain. When the gate electrode is shorted to the source, the device can support a large drain voltage across the p-base/n-drift-region junction. The forward blocking capability is shown in Fig. 7 by the lowest trace. Although a finite leakage current is illustrated in the figure, it is very small except at high operating temperatures. The maximum forward blocking voltage is determined by the avalanche breakdown voltage of the p-base/n-drift-layer junction.

When a positive gate bias is applied, the channel becomes conductive. At low drain voltages, the current flow is essentially resistive with the "on" resistance determined by a combination of the channel and drift-region resistances. The channel resistance decreases with increasing gate bias while the drift-region resistance remains constant. The total "on" resistance decreases with increasing gate bias until it approaches a constant value. At large gate-bias voltages, the channel resistance becomes small compared with the

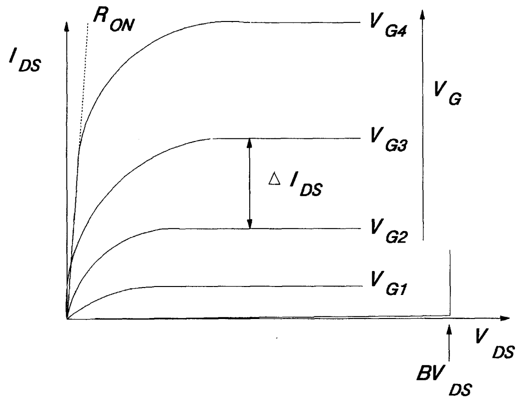

FIG. 7. Output characteristics of the power MOSFET. I_{DS}, drain-to-source current; V_{DS}, drain-to-source voltage.

drift-region resistance and the device "on" resistance becomes independent of gate bias. The "on" resistance is an important power MOSFET parameter. It is a measure of the current-handling capability of the device because it determines the power dissipation during current conduction. The "on" resistance is defined as the slope of the output characteristic in the linear region at low drain voltages. At high drain voltages, the resistance of the power MOSFET increases. Ultimately, the current saturates at high drain voltages as shown in Fig. 7. The current saturation in power MOSFETs can be used to provide a current-limiting function in power circuits as long as the power dissipation in the devices is kept within reasonable limits.

An important device parameter for power MOSFETs is the transconductance. It is defined by $g_m = \Delta I_{DS}/\Delta V_{GS}$ where $\Delta V_{GS} = V_{G3} - V_{G2}$ as illustrated in Fig. 7. A large transconductance is desirable to obtain a high current-handling capability with low gate-drive voltage, and for achieving high-frequency response. In the saturated-current region of operation, the output characteristics are controlled by the gate-induced channel characteristics. The transconductance is, therefore, determined by the design of the channel and gate structure.

3.3 Threshold Voltage

The voltage on the gate electrode at which strong inversion begins to occur in the MOS structure is an important design parameter for power MOSFETs because it determines the minimum gate bias required to induce an n-type conductance in the channel. This voltage is called the *threshold voltage.* For proper device operation, its value cannot be too large or too small. If the threshold voltage is large, a high gate-bias voltage will be needed to turn on the power MOSFET. This imposes problems with the design of the gate-drive circuitry. It is also important that the threshold voltage not be too low. On account of the existence of charge in the gate oxide, it is possible for the threshold voltage to be negative for *n*-channel power MOSFETs. This is an unacceptable condition because a conductive channel will now exist at zero gate-bias voltage; i.e., the devices will exhibit normally "on" characteristics. Even if the threshold voltage is above zero for an *n*-channel power MOSFET, its value should not be too low because the device can then be inadvertently triggered into conduction either by noise signals at the gate terminal or by the gate voltage being pulled up during high-speed switching. Typical power MOSFET threshold voltages are designed to range between 2 and 3 V.

3.4 On-State Resistance

Current flow in a power MOSFET during forward conduction is achieved by the application of a positive gate-bias voltage for *n*-channel devices in order to create a conductive path across the *p* base region underneath the gate. The current flow is then limited by the total resistance between the source and drain. This resistance consists of many components, as illustrated in the DMOS cross section in Fig. 6, which combine to determine the on-state voltage drop when the device is carrying current. The resistances of the n^+ emitter (R_{n+}) and substrate (R_{SUB}) regions are generally negligible for high-voltage power MOSFETs that have high drift-region resistance. They become quite important when the drift and channel resistances become small as in the case of low–breakdown-voltage (<100 V) devices. The channel (R_{CH}) and accumulation-layer (R_{A}) resistances are determined by the conductivity of the thin surface layer induced by the gate bias. These resistances are a function of the charge in the surface layer and the electron mobility near the surface. In addition to these resistances, the drift layer contributes two more components to the total "on" resistance. The portion of the drift region that comes to the upper surface between the cells contributes a resistance R_J that is enhanced at higher drain voltages by the pinch-off action of depletion layers extending from adjacent *p* base regions. This phenomenon has been termed the *JFET action*. Finally, the main body of the drift region contributes a large series resistance (R_D) especially for high-voltage devices. Only the channel and drift-region resistances will be treated in more detail in this article.

3.4.1 Channel Resistance The channel resistance can be derived as a function of the gate and drain voltages by using the *gradual channel approximation:*

$$I_D = \frac{\mu_{ns} C_{ox} Z}{2L} [2(V_G - V_T)V_D - V_D^2], \qquad (2)$$

where I_D is drain current, μ_{ns} is mobility of electrons in the inversion layer, C_{ox} is specific capacitance of the oxide layer, L is the channel length, V_G is the gate bias, V_T is the threshold voltage, V_D is the drain voltage, and Z is the channel width. According to this equation, when the drain voltage is small, the drain current increases linearly with drain voltage. In this linear region, the channel resistance R_{CH} is given by

$$R_{CH} = L/Z\mu_{ns}C_{ox}(V_G - V_T). \qquad (3)$$

As the drain voltage and current increase, the second term in Eq. (2) becomes increasingly important and causes the drain current to saturate. Physically, this corresponds to a reduction in the channel inversion-layer charge near the drain with increasing drain voltage. Ultimately, the inversion layer charge at the drain end of the channel becomes zero and the drain current saturates. The saturated drain current is given by

$$I_{DS} = \frac{\mu_{ns} C_{ox} Z}{2L} (V_G - V_T)^2. \qquad (4)$$

The saturated drain current is an important parameter because it determines the maximum current that the channel will support. The transconductance of the device in the saturated current region of operation can be obtained by differentiating Eq. (4) with respect to the gate voltage:

$$g_{ms} = \frac{dI_D}{dV_G} = \mu_{ns} C_{ox} \frac{Z}{L} (V_G - V_T). \qquad (5)$$

3.4.2 Ideal Specific "On" Resistance Consider the ideal case where the resistances of the n^+ emitter, n^+ substrate, *n*-channel region, accumulation region, and JFET region are negligible. The specific "on" resistance of the power MOSFET will then be determined by the drift region alone. In addition, if it is assumed that the current flows uniformly through the drift region without current-spreading effects, the resistance of the drift region is referred to as the *ideal specific "on" resistance* for the power MOSFET. This is the resistance of a drift region with the doping concentration and thickness required to support the desired breakdown voltage. For *n*-channel devices,

$$R_{on,sp} = \frac{W_D}{q\mu_n N_D} = 5.93 \times 10^{-9}(BV_{PP})^{2.5}, \qquad (6)$$

where W_D is the width of the drift region, μ_n is the mobility of electrons in the bulk, N_D is the dopant concentration in the drift region, and BV_{PP} is the breakdown voltage of a parallel-plane junction; while for *p*-channel devices,

$$R_{on,sp} = \frac{W_D}{q\mu_p N_A} = 1.63 \times 10^{-8}(BV_{PP})^{2.5}, \qquad (7)$$

where μ_p is the mobility of holes in the bulk. The ideal specific "on" resistance of the *p*-channel MOSFET is about 2.5 times larger than that for the *n*-channel MOSFET as a result of the lower mobility for holes in silicon when compared with electrons. For this reason, most commercial power MOSFETs are *n*-channel devices.

It is possible to approach the ideal specific "on" resistance for DMOSFETs with higher breakdown voltages when compared to those with lower breakdown voltages. This is illustrated in Fig. 8, where the specific "on" resistance is plotted as a function of the breakdown voltage. The ideal specific "on" resistance is also shown in the figure for comparison. The state-of-the-art line for devices of 1981 vintage is representative of a

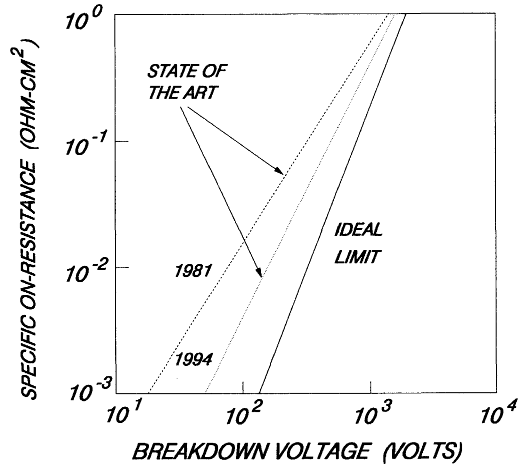

FIG. 8. Increase in the "on" resistance of a power MOSFET with increasing breakdown voltage.

process design rule that requires a 16-μm polysilicon window. With advancements in lithographic tools used to fabricate these devices, it has been possible to shrink the polysilicon window size. From the plots in Fig. 8, it can be seen that silicon power MOSFETs with high breakdown voltages approach the ideal performance limit, while improvements in the MOSFET structure are required for approaching ideal performance for the devices with lower breakdown voltages.

3.5 Input Capacitance

The power MOSFET is inherently capable of operation at high frequencies because of the absence of minority-carrier transport. Two limits to operation at high frequencies are set by the transit time across the drift region and the rate of charging of the input gate capacitance. In a power MOSFET, the input capacitance is relatively large on account of the large active area required to carry substantial drain currents. Consequently, the frequency response is usually limited by the charging and discharging of the input capacitance. To analyze the frequency response, it is necessary to evaluate each of the components of the input capacitance arising in the power MOSFET structure. Consider the conventional DMOS structure shown in Fig. 9 with the gate electrode extending between adjacent DMOS cells. The input gate-to-source capacitance for this structure contains several components:

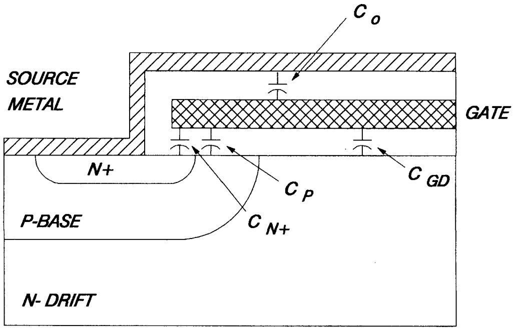

FIG. 9. Capacitances in the power DMOSFET structure.

1. the capacitance C_{n+} arising from the overlap of the gate electrode over the N^+ source region,
2. the capacitance C_p arising from the MOS structure created by the gate electrode over the p base region, and
3. the capacitance C_0 arising from running the source metal over the gate electrode.

The total gate-to-source capacitance is given by

$$C_{\mathrm{GS}} = C_{n+} + C_p + C_0. \tag{8}$$

The gate-to-drain capacitance varies with gate and drain voltage. It has a high value during the "on" state because the surface of the drift region is under accumulation. As the drain voltage increases and the device supports high voltages, the gate-to-drain capacitance decreases. Because of the amplification of the gate-to-drain capacitance by the Miller effect, this capacitance can severely reduce the frequency response. It is therefore important to reduce its magnitude. The gate-to-drain overlap capacitance can be drastically reduced by altering the gate structure by eliminating the gate overlap over the drift region. However, it is important to note that eliminating the gate-to-drain overlap creates a high electric field at the edge of the gate. This can reduce the cell breakdown voltage. In addition, it adversely affects the "on" resistance through an increase in the resistance between the channel and the drift region because an accumulation layer is no longer formed over a portion of the surface between the base regions.

3.6 Cell Topology

In previous discussions of power MOSFETs, cross sections of devices were shown without consideration of the cell layout on the surface. The cell windows that are commonly used for the design of power MOSFETs have linear, square, circular, hexagonal, or the atomic-lattice-layout (A-L-L) shapes. These windows can be located in either a square cell or a hexagonal cell pattern. The impact of the layout of these cells upon the resistance has been analyzed under the assumption that they all have the same drift-region doping concentration. It has been found that the "on" resistances of all cellular designs are equal if the size of the cell window and the ratio of the area of the cell window to the total cell area are kept the same. However, it has been found that the drift-region doping concentration must be adjusted on the basis of the electric field crowding within the cell. When this is taken into account, the A-L-L design has been found to be superior to the other designs. The A-L-L design also has a lower drain overlap capacitance, which is favorable for operation at high frequencies.

3.7 Trends

Power MOSFETs have the following attributes:

1. a voltage-controlled characteristic with low input gate power,
2. a stable negative temperature coefficient for the "on" resistance,
3. a wide safe operating area, and
4. a high switching speed.

Because of these features, when first introduced, the devices were expected to replace power bipolar transistors in many applications. This has occurred for systems with low operating voltages. Examples are switch-mode power supplies, computer peripherals, and automotive electronics. This displacement has not occurred for systems operating at high voltages on account of the high "on" resistance of power MOSFETs designed to support voltages above 200 V. With rapid progress along the learning curve, the manufacturing cost of power MOSFETs now rivals that for the bipolar transistors. Since the "on" resistance of the silicon power MOSFETs is already close to the ideal value, progress in increasing their power-handling capability can be achieved only by increasing the area of the device. This is expected to occur at a pace dictated by improvements in process technology that lead to a reduction in defect density. Recently, it has been theoretically shown that replacement of silicon with silicon carbide would lead to a 200-fold reduction in the specific "on" resistance. This offers a tremendous opportunity for improvement of device characteristics if the technological problems with manufacturing silicon carbide devices can be overcome.

4. INSULATED-GATE BIPOLAR TRANSISTORS

The development of the insulated-gate bipolar transistor (IGBT) was based upon a innovative approach that combines the physics of bipolar current conduction with the physics of MOS gated current control within the same semiconductor region. These devices have replaced bipolar power transistors in medium-power applications, where the blocking voltages lie between 300 and 1500 V.

4.1 Device Structure and Operation

A cross section of the IGBT is shown in Fig. 10 based upon the DMOS process. In this structure, current flow cannot occur

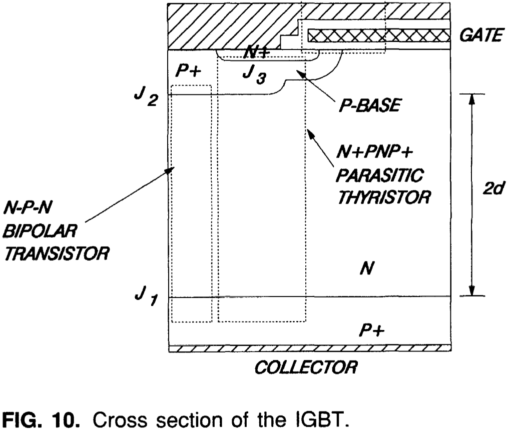

FIG. 10. Cross section of the IGBT.

when a negative voltage is applied to the collector with respect to the emitter because the lower junction (J_1) will become reverse biased. This provides the device with its reverse blocking capability. In this mode of operation, the depletion region extends in the *n* drift region. When a positive voltage is applied to the collector with the gate shorted to the emitter, the upper junction (J_2) becomes reverse biased, and the device operates in its forward blocking mode. In this mode of operation, the voltage is supported by a depletion region formed in the *n* drift region. The forward and reverse blocking capabilities of the IGBT cell structure shown in Fig. 10 are approximately equal because they are determined by the thickness and resistivity of the same *n* drift layer.

If a positive gate bias is applied of sufficient magnitude to invert the surface of the *p* base region under the gate when the device is in its forward blocking mode, it can be switched into its on state. In the forward conducting state, electrons flow from the n^+ emitter to the *n* drift region. This provides the base drive current for the vertical *p-n-p* transistor in the IGBT structure. Since the emitter junction (J_1) for this bipolar transistor is forward biased, the p^+ region injects holes into the *n* base region. (It is worth pointing out here that the IGBT terminals have been named from the point of view of replacing a power bipolar transistor with the IGBT in circuits. Unfortunately, this nomenclature is in conflict with the internal physics of operation of the device. Thus, the junction J_1 acts as an emitter of the internal *p-n-p* transistor even though the p^+ region is connected to the collector terminal.) When the positive bias on the collector terminal of the IGBT is increased, the injected hole concentration increases until it exceeds the background doping level of the *n* drift region.

If the inversion-layer conductivity is reduced by application of a gate bias close to the threshold voltage, a significant voltage drop occurs across this region due to the electron current flow in a manner similar to that observed in conventional MOSFETs. When this voltage drop becomes comparable to the difference between the gate bias and the threshold voltage, the channel becomes pinched off. At this point, the electron current saturates. Since this limits the base drive current for the *p-n-p* transistor, the hole current flowing through this path is also limited. Consequently, the device operates with current saturation in its active region with a gate-controlled output current. These characteristics of the IGBT are shown in Fig. 11.

In order to switch the IGBT from its on state to the off state, it is necessary to discharge the gate by shorting it to the emitter. In the absence of a gate voltage, the inversion region at the surface of the *p* base under the gate cannot be sustained. Removal of the gate bias cuts off the supply of electrons to the *n* drift region and initiates the turnoff process. Because of the presence of the high concentration of minority carriers injected into the *n* drift region during forward conduction, the turnoff does not occur abruptly. At first, an abrupt reduction in the anode current is observed because the electron current via the channel is terminated. The collector current then decays gradually with a characteristic time constant determined by the minority-carrier lifetime.

The IGBT structure contains a parasitic *p-n-p-n* thyristor structure between the collector and the emitter terminals as indicated in Fig. 10. If this thyristor latches up, the current can no longer be controlled by the MOS gate. It is important to design the device in such a manner that this thyristor action will be suppressed. This can be achieved by preventing the injection of electrons from the n^+ emitter region into the *p* base during

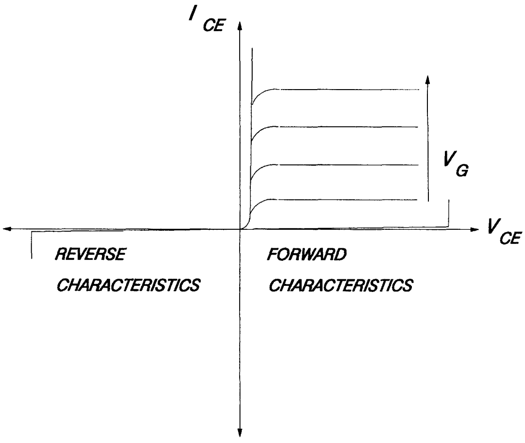

FIG. 11. Output characteristics of the symmetrical IGBT structure.

device operation. If the resistance to hole current flow within the p base region is sufficiently small, the thyristor will not be turned on because of the low gain of the upper n-p-n transistor. The equivalent circuit for the IGBT then consists of a MOSFET driving the base of the p-n-p transistor in a Darlington configuration.

4.2 Blocking Characteristics

The forward and reverse blocking voltage capability of the IGBT defines one of its important limits in performance for applications. When a negative collector voltage is applied, a large voltage can be supported by the IGBT structure shown in Fig. 11 because junction J_1 becomes reverse biased. When junction J_1 becomes reverse biased, its depletion layer extends primarily into the lightly doped n drift region. It is important to note that the breakdown voltage during reverse blocking is determined by an open-base transistor formed between the p^+ collector, the n drift region, and the p base region. This structure is prone to reach-through breakdown if the n drift region is too lightly doped. As a general guideline, the width of the n drift region is chosen so that its thickness is equal to the depletion width at the maximum operating voltage plus one diffusion length. In dc circuit applications, the IGBT is not required to support reverse voltage. This offers the opportunity to reconfigure the device structure to optimize the forward conduction characteristics for a given forward blocking capability without consideration for the reverse blocking capability. The doping profile and electric-field distribution for such an asymmetrical IGBT are shown in Fig. 12. In the asymmetrical IGBT structure, the uniformly doped n drift region of the symmetrical IGBT is replaced by a two-layer n drift region. This alters the electric-field distribution as illustrated on the right-hand side of the figure. If the critical electric field for breakdown is assumed to be independent of the n drift-region doping level and the n drift-region doping in layer 1 is very low, the electric-field distribution changes from the triangular case in the symmetrical IGBT to a rectangular case in the asymmetrical structure. Under these circumstances, if the n buffer-layer thickness (d_2) is assumed to be approximately equal to one diffusion length (L_p), the forward blocking capability of the asymmetrical device will be twice as high as for the symmetrical device.

4.3 On-State Characteristics

The analysis of the forward conduction characteristics can be performed by treating the device as a p-i-n rectifier connected in series with a MOSFET. The voltage drop across the p-i-n rectifier ($V_{\mathrm{F},pin}$) is related to its forward conduction current density ($J_{\mathrm{F},pin}$) by

$$V_{\mathrm{F},pin} = \frac{2kT}{q} \ln\left[\frac{J_{\mathrm{C}}d}{2qD_{\mathrm{a}}n_{\mathrm{i}}F(d/L_{\mathrm{a}})}\right], \tag{9}$$

where J_{C} is the collector current density, d is the half thickness of the drift region, D_{a} is the ambipolar diffusion coefficient, n_{i} is the intrinsic carrier current, L_{a} is the ambipolar diffusion length, and $F(d/L_{\mathrm{a}})$ is a complex function. The MOSFET current is given by

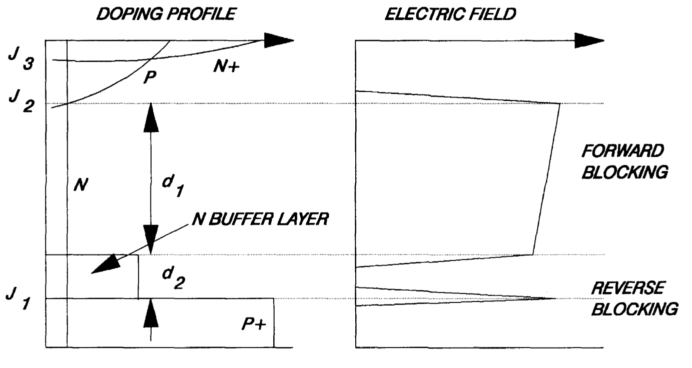

FIG. 12. Doping profile and electric-field distribution within the asymmetrical IGBT structure.

$$I_C = \frac{\mu_{ns} C_{ox} Z}{2L_{ch}} [2(V_G - V_T)V_{F,MOS} - V_{F,MOS}^2], \tag{10}$$

where $V_{F,MOS}$ is the voltage drop across the MOSFET portion of the IGBT and L_{CH} is the MOSFET channel length. In the forward conduction mode, sufficient gate voltage is applied that the forward voltage drop across the device is low. Under these conditions, $V_{F,MOS} \ll V_G - V_T$, and the MOSFET section of the devices is operating in its linear region. The voltage drop across the MOSFET section is, then, given by

$$V_{F,MOS} = \frac{I_C L_{CH}}{\mu_{ns} C_{ox} Z (V_G - V_T)}. \tag{11}$$

On the basis of this model for the IGBT, the forward voltage drop across the IGBT is the sum of the voltage drops across the MOSFET and the *p-i-n* rectifier:

$$V_F = \frac{2kT}{q} \ln\left[\frac{I_C d}{2qWZD_a n_i F(d/L_a)}\right] + \frac{I_C L_{CH}}{\mu_{ns} C_{ox} Z (V_G - V_T)}. \tag{12}$$

Several important conclusions about IGBT performance can be derived from the *p-i-n* rectifier/MOSFET model. First, it becomes apparent that there will be a knee in the output conduction characteristic (as shown in Fig. 11) below which very little current flow will occur, but the current density will rise very rapidly (exponentially) thereafter. This behavior has been experimentally verified in devices resulting in operating on-state current density for the IGBT being substantially greater than for the devices. This indicates that a smaller chip size can be used, resulting in reduced cost for the IGBT.

Using the *p-i-n* rectifier/MOSFET model, it is also possible to derive the IGBT characteristics under current saturation. When the gate bias is reduced to a value close to the threshold voltage such that the voltage drop across the MOSFET channel becomes significant, the current becomes limited by the MOSFET. As in the case of a MOSFET, the IGBT collector current will saturate as shown in Fig. 11 at a value given by

$$I_{C,sat} = \frac{\mu_{ns} C_{ox} Z}{2L_{CH}} (V_G - V_T)^2. \tag{13}$$

The transconductance of the IGBT will then be identical to that for the MOSFET with the same cell size and channel length. In actual devices, the transconductance for the IGBT has been found to be larger than that for the MOSFET with the same cell size and channel length. The reason for this is an additional component of current flowing through the *p-n-p* transistor, which was not accounted for in the *p-i-n*/MOSFET model.

4.4 Parasitic Thyristor Latchup

The maximum operating current for the IGBT is limited by the presence of the parasitic thyristor within the structure. As shown in Fig. 10, this thyristor is formed between the n^+ source region of the MOSFET acting as the cathode of the thyristor, the *p* base region, the *n* drift region, and the p^+ collector region. During operation of the IGBT in the on state, electrons are supplied via the MOSFET channel and holes injected from junction J_1 are collected at junction J_2. The hole current flows to the emitter terminal via the resistance of the *p* base region. This current flow produces a voltage drop that forward biases the junction J_3. At normal operating current levels for the IGBT, this voltage drop can be made much smaller than that for a forward-biased diode (V_{bi}) by making the shunting resistance small. Under these conditions, the current gain of the *n-p-n* transistor is very small, and the thyristor cannot latch up. However, when the on-state current density is increased, the forward bias on junction J_3 can become sufficiently large to increase the current gain of the *n-p-n* transistor. If the sum of the current gains of the *n-p-n* and *p-n-p* transistors exceeds unity, the thyristor can latch up. The collector current can now flow directly to the emitter terminal, bypassing the MOSFET channel. This implies that the IGBT current is no longer controlled by the gate bias.

From the above discussion, it can be concluded that it is essential to suppress the turnon of the parasitic thyristor in order to maintain gate control over the collector current. This can be achieved by reducing the gain of either the *n-p-n* transistor, the *p-n-p* transistor, or both transistors. Since the on-state current flows partially via the *p-n-p* transistor, a reduction in its gain produces an increase in the on-state voltage drop. It is therefore preferable to reduce the gain of the *n-p-n* transistor. However, methods employed to improve the turnoff time for the IGBT tend to reduce the gain of the *p-n-p* transistor, which then leads to an increase in the current density at which the parasitic thyristor will latch up. In consequence of successful suppression of the latchup of the parasitic thyristor, IGBTs have become commercially available with performance suitable for many applications.

4.5 Switching Characteristics

An important characteristic of the IGBT is its MOS-gate–controlled turnoff capability. Since the current flowing through the MOS channel controls the output characteristics, the collector current flow can be interrupted by removing the gate-drive voltage. To perform the transition from the "on" state to the forward blocking state, the gate is connected to the emitter terminal via an external circuit that allows discharging the gate capacitance. When the gate voltage falls below the threshold voltage of the MOS gate structure, the channel inversion layer can no longer exist. At this point, the electron current supplied by the channel ceases. If the gate turnoff is performed using a low external resistance in the gate-drive circuit so as to reduce the gate voltage abruptly to zero, the collector current will drop abruptly because the channel current is suddenly discontinued. Even after this occurs, the collector current continues to flow because the hole current (I_h) does not cease abruptly. The high concentration of minority carriers stored in the *n* drift region during the on state supports the hole-current flow. As the minority-carrier density decays through recombination, it leads to a gradual reduction in the collector current. This current flow is sometimes referred to as a *current tail.*

Typical examples of the IGBT voltage and current wave forms during device turnoff are provided in Fig. 13 for gate-controlled turnoff. In the case of a resistive load, the collector voltage rises during turnoff at the same rate at which the collector current decays. In contrast, in the case of an inductive load, the collector voltage abruptly rises to the supply voltage when the gate-drive voltage is reduced to zero. In either case, it is important to note that a large power dissipation occurs during the time that the collector current decays to zero because of the relatively high voltage across the IGBT during this time. It is therefore important to minimize the duration of the current tail by reducing the minority-carrier lifetime. The minority-carrier lifetime can be reduced by performing electron irradiation, directed perpendicular to the wafer surface, after device fabrication to introduce recombination centers. A reduction in the turnoff time from over 20 μs to less than 200 ns can be achieved by electron irradiation, making the IGBT suitable for applications operating over a wide frequency range.

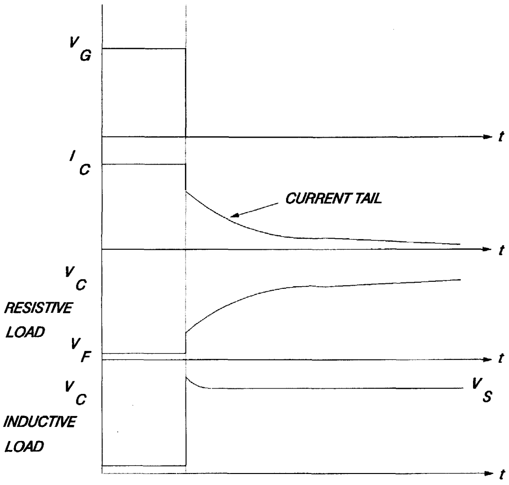

FIG. 13. Switching wave forms for the IGBT.

4.6 Trade-Off Curve

The results described above demonstrate that electron irradiation can be utilized to control the switching speed of IGBTs. However, this increase in the switching speed is accompanied by an increase in the forward voltage drop. Since a short gate turnoff time is desirable in order to reduce switching losses and a low forward voltage drop is desirable in order to reduce the conduction losses, it becomes necessary to perform a trade-off between these device characteristics. This can be most conveniently performed by using a plot of the forward voltage drop versus the gate-controlled turnoff time as shown in Fig. 14 for the symmetrical and asymmetrical IGBT structures.

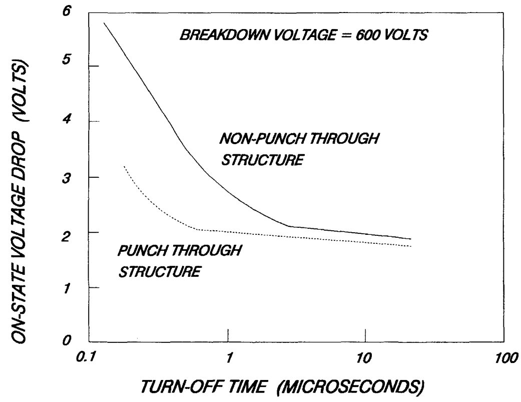

FIG. 14. Power loss trade-off curves for the IGBT.

Depending upon the application, the appropriate device characteristics can be selected by choosing the irradiation dose. In the case of circuits operating at low frequencies with large duty cycles, where the conduction losses dominate the switching losses, IGBTs with turnoff times in the range of 5–20 μs would be the best. An example of this is line-operated phase-control circuits. For higher-frequency circuits with shorter duty cycles, where the switching losses are comparable to the conduction losses, IGBTs with turnoff times in the range of 0.5–2.0 μs would be appropriate. An example of this type of application is in ac motor drives operating at frequencies ranging from 1 to 10 kHz. For high-frequency circuits, the switching losses would become dominant and IGBTs with gate turnoff times ranging from 100 to 500 ns would be necessary. An example of such circuits is in uninterruptible power supplies (UPS) operating at between 20 and 100 kHz. Thus, the IGBTs can be tailored to match the power switching requirements of a broad range of applications.

It is worth pointing out that the trade-off curve for the asymmetric IGBT structure has been found to be superior to that for the symmetric IGBT structure as shown in the figure. This is because of the smaller thickness of the drift region in the asymmetrical structure for obtaining the same forward blocking voltage as discussed earlier. The smaller drift-region thickness results in a lower on-state voltage drop for the same lifetime in the drift region. It has also been found that the trade-off curve can be improved upon by increasing the doping concentration in the buffer layer. This results in a reduction in the injection efficiency of the lower junction (J_1), which reduces the current gain of the *p-n-p* transistor.

4.7 Trends

The IGBT was the first commercially successful device based upon combining the physics of MOS-gate control with bipolar current conduction. The MOS gate provides an easy interface that can be addressed using low-cost integrated circuits. The bipolar current conduction makes high-current, high-voltage devices available to circuit designers at a relatively low cost. Further, the IGBT structure is very similar to that for the power MOSFET from a fabrication standpoint. This has made its manufacturing relatively easy immediately after conception and the suppression of the parasitic thyristor. After commercial introduction of IGBTs in the mid 1980s, the ratings of devices have grown at a very rapid pace because of the ability to scale up both the current and blocking voltage ratings. In addition, the differences in cost for production of these MOS-technology–based devices and the bipolar power transistor have diminished. This has made the IGBT very attractive from an applications point of view, resulting in rapid displacement of power bipolar transistors. Some of the applications for which the IGBT is used are adjustable-speed motor control for air conditioning, numerical controls for factory automation and robotics, and appliance controls. The device has also been chosen for the development of the electric car. As the ratings of the devices are scaled up, it is anticipated that the devices may find application in traction (electric street cars) as well.

6. CONCLUSIONS

An ideal power transistor must be able to operate with no internal power dissipation while controlling large amounts of power, and it should operate with infinite power gain. Historically, the first transistors developed for power electronics applications were the bipolar power transistors commercially introduced in the 1960s. Although refined and used for many applications, they had the drawback of requiring large input current to control the output current flow. This poor power gain resulted in complex and expensive control circuits. In the 1970s, the first power MOSFETs were introduced. Because of their very high input impedance, they have replaced the bipolar power transistors in applications operating at low (<100 V) voltages. In the 1980s, the IGBT was introduced for medium-voltage (up to 1500 V) applications. In addition to its high input impedance, the IGBT has the feature that it can control much greater power for the same chip size than the power MOSFET. As a consequence, the IGBT has replaced the bipolar transistor in medium-power applications. In order to replace the gate-turnoff thyristors used at present for high-power applications, manufacturers are now developing IGBTs with blocking voltages above 2000 V.

Further Reading

Baliga, B. J. (1995), *Power Semiconductor Devices*, Boston: PWS Publishing Company.

Ghandhi, S. K. (1977), *Semiconductor Power Devices*, New York: Wiley.

Grant, D. A., Gowar, J. (1989), *Power MOSFETs*, New York: Wiley.

Roulston, D. J. (1990), *Bipolar Semiconductor Devices*, New York: McGraw Hill.

TRANSPARENT OPTICAL MATERIALS

MARVIN J. WEBER, *Lawrence Livermore National Laboratory, University of California, Livermore, California, U.S.A.*

	Introduction	205
1.	**Fundamental Optical Phenomena**	206
1.1	Absorption	206
1.2	Refraction	207
1.3	Reflection	208
1.4	Scattering	209
1.5	Elasto-optic Properties	209
1.6	Thermo-optic Properties	210
1.7	Magneto-optic Properties	210
1.8	Nonlinear Optical Properties	211
1.9	Luminescence and Stimulated Emission	211
1.10	Optical Damage	212
2.	**Optical Glasses**	212
2.1	Classification	213
2.2	Properties	214
2.2.1	Optical Properties	214
2.2.2	Thermal Properties	215
2.2.3	Mechanical Properties	216
2.2.4	Chemical Durability	216
2.3	Quality and Forms	216
2.4	Special Glasses and Applications	217
2.4.1	Abnormal-Dispersion Glass	217
2.4.2	Gradient-Index Glass	217
2.4.3	Mirror-Substrate Glass	217
2.4.4	Optical-Filter Glass	217
2.4.5	Nonlinear Optical Glass	218
2.4.6	Laser Glass	218
2.4.7	Faraday-Rotator Glass	219
2.4.8	Acousto-optic Glass	219
2.4.9	Radiation-Detection Glass	220
2.4.10	Radiation-Shielding Glass	220
2.4.11	Photosensitive Glass	220
3.	**Optical Plastics**	221
3.1	Optical and Physical Properties	221
3.2	Applications	222
4.	**Optical Crystals**	223
4.1	Optical Properties	223
4.2	Physical Properties	224
4.3	Applications	225
4.3.1	Optical Components	225
4.3.2	Electro-optic Crystals	225
4.4	Magneto-optic Crystals	226
4.5	Acousto-optic Crystals	226
4.6	Laser Crystals	227
4.7	Scintillator Crystals	227
5.	**Optical Liquids**	227
5.1	Optical Properties	227
5.2	Physical Properties	228
5.3	Chemical Properties	228
5.4	Applications	229
5.5	Liquid Crystals	229
	Glossary	230
	Works Cited	230
	Further Reading	232

INTRODUCTION

Transparent optical materials encompass a broad class of materials—organic and inorganic, crystalline and amorphous, solids and liquids. The optical range of the electromagnetic spectrum is also broad—from the infrared to the ultraviolet. In this article, we consider the optical wavelength range extending from ~100 nm on the short-wavelength side, limited by the fundamental band gap of solids, to ~25 μm in the infrared, beyond which strong absorption by molecular vibrations limits the useful transmission of the material. The materials described are bulk transmitting optical materials of the types mentioned above; metals that reflect light, photodetector materials, and thin-film optical coatings are therefore not included.

We begin with a brief review of fundamental optical phenomena relevant to understanding the properties and applications of optical materials. Included are both passive properties (absorption, reflection, refraction, and polarization) and special properties such as electro-optic, magneto-optic, and elasto-optic effects, luminescence and stimulated emission, and other photosensitive processes

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

in materials. A more extensive discussion of optical properties of solids is presented in OPTICAL PROPERTIES OF SOLIDS.

Optical glasses are considered first because they are by far the most widely used optical materials. They are well-developed and well-characterized materials that are commercially available in a variety of chemical compositions, in high quality and large sizes, and with reproducible optical properties. Glasses are the quintessential optical material and epitomize an engineered material. Organic polymers constitute another class of noncrystalline optical materials, having a somewhat different and narrower range of properties than for glasses. Optical liquids are included in a separate section. Optical glasses, plastics, and liquids are optically isotropic materials. This simplifies the discussion of their optical properties and contributes to their general usefulness.

Optical crystals encompass a wide range of chemical compositions and atomic and molecular structures. Many cubic crystals exist that have isotropic optical properties and other desirable physical properties that render them attractive for many applications. Crystals of lower symmetry can, in addition, exhibit a variety of direction-dependent, anisotropic optical effects including polarization, double refraction, optical activity, and linear and nonlinear electro-optic phenomena. These features have made possible a large number of special and novel applications.

Both crystalline and amorphous optical materials can have multiple uses both as passive transmitting and absorbing materials and as active materials. In the latter case they may contain ions or molecular groups, either as a basic constituent or by doping, that are used to produce laser action, scintillation for detecting high-energy radiation, paramagnetic and ferromagnetic Faraday rotation, and photochromic and other photosensitive properties.

Prompted by the development of lasers and nonlinear optics, the fields of optics and electro-optics have expanded greatly. Concurrent research and development of optical materials has led to the availability of many different transmitting materials too numerous to tabulate within the scope of this article. Instead, for each class of optical material, fundamental optical properties, principal applications, and representative examples of the more useful materials are given. Other important physical and chemical properties and their variation with temperature and other ambient conditions are also covered. References to books and handbooks where more extensive discussion and tabulation of optical materials and their properties can be found are given in the list of Further Reading at the end of the article.

1. FUNDAMENTAL OPTICAL PHENOMENA

The interaction of light with a solid can be specified by various optical response functions. To describe the amplitude and phase of the response, these functions are generally frequency-dependent complex quantities. The complex dielectric constant ϵ, for example, relates the amplitude of an optical electric field E of frequency ω to the polarization wave $\mathbf{P}(\omega)$ induced in the material and is given in terms of the complex index of refraction $\mathbf{n}(\omega)$ by $\epsilon(\omega) = \mathbf{n}^2(\omega)$. The index of refraction, in turn, can be related to the amplitude and phase of the reflectivity. For each of these complex quantities the real and imaginary components are related by Kramers–Kronig integral equations (Zallen, 1983). Thus if one of the components is known over a sufficiently large frequency range, the other component can be obtained from an integral transform.

1.1 Absorption

The strong absorption of optical materials in the infrared (IR) is associated with the interaction of light with molecular vibrations giving rise to multiphonon absorption processes; in the visible and ultraviolet (UV), absorption is associated with electronic transitions between the valence band and the conduction band or exciton levels. The positions of both the long- and short-wavelength absorption edges depend on the chemical composition and the interatomic and intermolecular bonding of the material. Oxide materials typically have a broad transmission window in the visible and near infrared; halide materials, especially fluorides, can have a more extended range of transparency. The region of transparency is shifted into the mid-infrared for more covalently bonded halides and chalcogenide materials and into the ultraviolet for more ionic fluoride materials.

The intensity I of light transmitted through a material of path length l is given by the Lambert–Beer law as

$$I = I_0 \exp(-\alpha l), \tag{1}$$

where I_0 is the incident intensity and α is the absorption coefficient. The fundamental absorption edge at short wavelengths can be described by an absorption coefficient exponentially dependent on the frequency ω (Urbach's rule). The long-wavelength absorption edge arises from multiphonon excitation of overtone and combination bands of fundamental vibrational modes of the solid and can also be fitted by an exponential dependence on energy with an average optical phonon frequency and a material constant characterizing the vibrational anharmonicity.

The range of absorption coefficients for crystals and glasses over the entire optical spectrum covers many orders of magnitude, from 10^6 cm^{-1} in the UV to less than 10^{-4} cm^{-1} in the near IR. This is illustrated for a soda-lime-silicate glass in Fig. 1. The extremely low absorption in the region between the UV and IR absorption edges is obtained only in ultrapure materials. Absorption in this region is frequently dominated by the presence of trace quantities (parts per billion) of iron and other transition- or post–transition-group impurities in the raw materials, by color centers due to vacancies or electron traps, and by higher-order harmonics of the lattice or glass network vibrations. Impurities may also mask the true UV absorption edge. Water is frequently present in materials in the form of OH^- groups. This gives rise to a strong fundamental absorption near 3 μm and weaker overtone and combination absorption bands throughout the near infrared. For most optical materials, an absorption coefficient of $<10^{-2}$ cm^{-1} in the wavelength region of interest is sufficient. For laser materials, absorption due to impurities is usually kept to less than 10^{-3} cm^{-1} at the lasing wavelength.

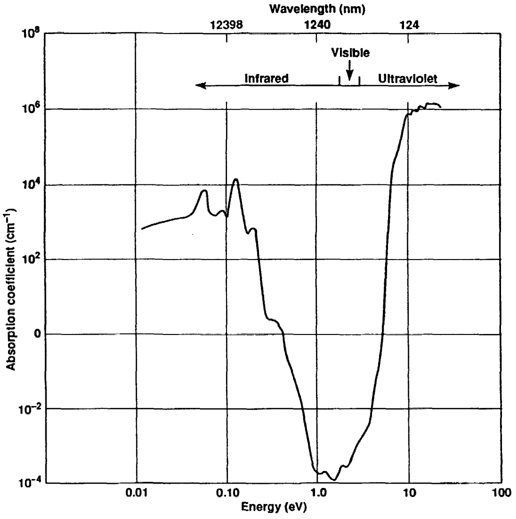

FIG. 1. Absorption spectrum of a soda-lime-silicate glass (Bagley *et al.*, 1976)

In many materials the absorption is polarization dependent. Dichroic materials (e.g., tourmaline) are materials in which light polarized in one direction is more strongly absorbed than light polarized at right angles to that direction.

1.2 Refraction

The index of refraction n for monochromatic radiation is the ratio of the velocity of light in vacuum to the velocity of light in the medium, that is, $n = c/v$. For an absorbing dielectric medium, the index of refraction at wavelength λ is complex and given by

$$\mathbf{n} = n - ik, \tag{2}$$

where the extinction coefficient k is related to the absorption coefficient by $k = \alpha\lambda/4\pi$. In vacuum the propagation velocity is the same for all wavelengths, but in a transmitting medium the velocity and refractive index vary with wavelength and direction of propagation. In regions far removed from absorption, where $k \approx 0$, the wavelength dependence of the refractive index is frequently expressed by the Sellmeier relationship

$$n^2(\lambda) - 1 = \sum_{i=1}^{m} \frac{A_i \lambda^2}{\lambda^2 - \lambda_i^2}, \tag{3}$$

where A_i and λ_i are the effective strength and mean wavelength of the ith absorption band and the summation is over all infrared (vibrational) and ultraviolet (electronic) bands.

The refractive index n and the dispersion $(dn/d\lambda)_\lambda$ are prime optical properties of materials.

For glasses and for crystals with cubic symmetry, the propagation of light is isotropic with a single index of refraction. Other symmetry classes of crystals are optically anisotropic. The propagation of light in these materials depends on the direction of wave propagation relative to the crystal axes and is treated using tensor calculus. The refractive index has a directional dependence, and Eq. (3) is extended to yield three indices of refraction n_1, n_2, n_3. These can be described by an index ellipsoid (optical indicatrix). A light wave propagating in an arbitrary direction can be decomposed into two waves having orthogonal polarizations and different velocities and thereby exhibiting double refraction. These waves have the same propagation velocity only in specific directions called optic axes. A uniaxial crystal has one such axis; a biaxial crystal has two. In birefringent materials, one polarization has an isotropic propagation velocity and obeys Snell's law (the ordinary wave); the orthogonal polarization has an anisotropic propagation velocity and does not obey Snell's law (extraordinary wave) (Guenther, 1990).

Materials that exhibit optical activity rotate the plane of polarization naturally, in contrast to rotation by electro-optic or magneto-optic effects that require an applied electric or magnetic field. Light traversing such a material of length l has its plane of polarization rotated through an angle

$$\theta = (\pi l/\lambda)(n_- - n_+), \tag{4}$$

where n_- and n_+ are the refractive indices for left- and right-circularly polarized light. For dextrorotatory materials $n_- > n_+$, and the right-circularly polarized light travels faster than the left; the opposite condition holds for levorotatory materials. Solids and liquids having molecules with no center of symmetry and no plane of symmetry exhibit optical activity.

1.3 Reflection

When a flux of light Φ traversing a medium of refractive index n' is incident on a material of index n, it is partially reflected and partially refracted as shown in Fig 2. By Snell's laws, $\theta_1' = \theta_1$ on reflection and $n \sin\theta_2 = n' \sin\theta_1$ on refraction. At Brewster's angle θ_p, given by $n'/n = \tan\theta_p$, only the component of light polarized perpendicular to the plane of incidence is reflected, and the intensity is plane polarized; at other angles of incidence the reflected light has mixed polarization. In general, the dependence of the reflectivity on wavelength and angle of incidence (Fresnel laws) is complex, especially near an absorption band and for anisotropic materials (Born and Wolf, 1980). The index of refraction and the absorption coefficient of a material can be derived from reflectivity data integrated over a broad range of wavelengths beyond the IR and UV absorption edges using a Kramers–Kronig analysis.

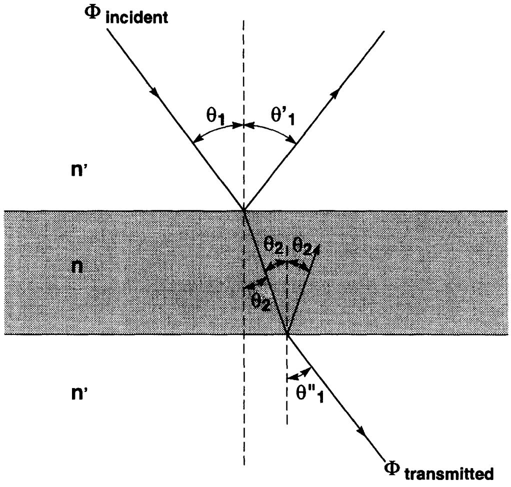

FIG. 2. Reflection, refraction, and transmission of light at material boundaries ($n' < n$).

The fraction of light of a given wavelength reflected by a material in air ($n' \approx 1$) for normal incidence is given by the Fresnel reflectivity

$$r = \frac{(n-1)^2 + k^2}{(n+1)^2 + k^2}. \tag{5}$$

In the nonabsorbing transparent region where $k \approx 0$, this reduces to $r = 4\%$ for a material of refractive index $n = 1.5$. The reflection of normal-incidence, monochromatic radiation when multiple reflections at both entrance and exit surfaces are considered is given by the reflection factor

$$R = \frac{2n}{n^2 + 1}. \qquad (6)$$

Fresnel reflective losses become large for materials and at wavelengths where n is large. Above the fundamental absorption edge, materials are highly reflective.

More generally, the reflectance depends on the angle of incidence and whether light is incident from the more dense or less dense medium. In addition to spectral reflection, the transmitted flux may also be reduced by diffuse reflection from irregular entrance and exit surfaces and imperfections in the interior of the material.

1.4 Scattering

The flux transmitted through an optical material is reduced from the incident flux by a combination of losses due to absorption, reflection, and scattering. Light scattering may be due to extrinsic causes or it may be intrinsic in nature. The former include discrete scattering centers such as bubbles, inclusions, impurities, or flaws introduced during production. The latter includes Rayleigh scattering caused by stationary density, temperature, and compositional fluctuations and Brillouin scattering caused by propagating fluctuations in the dielectric constant. Colloidal particles that diffract light—the Tyndall effect—also contribute to light scattering.

Scattering processes may be distinguished by the size and absorptivity of the active centers (Van de Hulst, 1957). Rayleigh scattering is caused by scattering centers that are small compared to the wavelength of light and exhibits a characteristic λ^{-4} dependence on wavelength. As the particle size increases, scattering occurs increasingly in the forward direction and exhibits a λ^{-2} dependence (Rayleigh–Gan scattering). Mie scattering occurs when the scattering centers have a size comparable to the wavelength of light. Scattering in optical materials begins to be visible for centers $\gtrsim 10$ nm in size. For large numbers of centers of dimensions ~100 nm, the material appears cloudy. For high densities of centers having dimensions $\gtrsim 1000$ nm, the material is usually opaque.

For most bulk optical materials, losses due to extrinsic imperfections and impurity absorptions are usually larger than intrinsic scattering losses. Light-scattering losses in single-component glasses such as SiO_2 are small and limited by microscopic density fluctuations associated with the random molecular structure. Light-scattering losses in multicomponent glasses and mixtures are generally larger because of additional concentration fluctuations (Schroeder, 1977). For optical fibers, scattering is the major loss mechanism when absorption is successfully reduced.

1.5 Elasto-optic Properties

Materials normally considered to be isotropic (cubic crystals, glasses, and plastics) become anisotropic when subject to mechanical stress or strain. Photoelastic properties describe the effects of elastic deformation on the refractive index and result in distortions of the optical wave front; acousto-optics are dynamic strain effects. The piezo-optic and elasto-optic coefficients determine the effects of stress and strain, respectively, and their number depends on the symmetry of the material. In general, a material may possess 36 independent piezo-optic or elasto-optic constants (tensor components) (Nye, 1976). Because of the high sensitivity of optical properties to material strain, the presence of strain can lower the apparent point-group symmetry of a crystal.

Because glass is isotropic, only two independent coefficients are needed to characterize the piezo-optic or elasto-optic properties. The change in refraction index Δn caused by a change in stress σ for light polarized parallel or perpendicular to the line of stress is given by

$$\Delta n = \Delta n_{\parallel} - \Delta n_{\perp} = \frac{n^3}{2}(q_{11} - q_{12})\sigma, \qquad (7)$$

where the q's are the piezo-optic constants. Corresponding relations involving elasto-optic constants p (Pockels coefficients) describe changes in refractive index and the state of strain in the glass. Mechanical stress in an optical material, either developed internally during production, applied externally, or as a consequence of fluctuating temperature, changes the refractive index. The resulting change in optical path length S for a material of length l caused by stress birefringence is $\Delta S = \Delta nl = Kl\sigma$, where from Eq. (7) the stress-optical coefficient K is defined by

$$K = \frac{n^3}{2}(q_{12} - q_{11}). \tag{8}$$

Stress birefringence varies with both temperature and wavelength.

1.6 Thermo-optic Properties

Absorption of intense light in a material can lead to the formation of temperature gradients and stress in the material. The optical path length of light passing through a material varies with temperature as a result of thermal expansion and of changes in the refractive index and is given by

$$dS/dT = (n - 1)\alpha + dn/dT, \tag{9}$$

where α is the linear coefficient of thermal expansion. For most crystals and glasses $|dn/dT| \leq 10^{-5}$/K and increases at shorter wavelengths. By adjusting the composition of a glass so that the change in path length due to thermal expansion is compensated by a negative value of dn/dT, the thermo-optic coefficient can be made positive, negative, or zero (Izumitani, 1986). In principle, with the proper combination of glass properties, no optical distortions would appear in the presence of a nonuniform temperature distribution. Even if this ideal athermal glass is not achieved, thermo-optic distortions in optical glass can be reduced by compositional tailoring. Glass ceramics have also been tailored to be athermal over a wide temperature range, but with poorer transmission.

1.7 Magneto-optic Properties

Optical rotation occurs in a medium when the refractive indices n_+ and n_- for right- and left-circularly polarized light are unequal. The microscopic origin of magnetic-optical rotation is the inequality that is created by the splitting of the ground- or excited-state energy levels due to an applied magnetic field or magnetization (Zeiger and Pratt, 1973). When the magnetic vector is parallel to the light propagation vector (Faraday effect), the angle of magnetic-optical rotation is given by Eq. (4). The sense of rotation is positive if the rotation is in the direction of the positive current in the coil that produces the magnetic field. In contrast to natural rotation, which depends on the direction of wave propagation, Faraday rotation depends on the direction of the magnetic field. When the magnetization vector is perpendicular to the light propagation vector (Cotton–Mouton effect), the magnetic linear birefringence is much smaller and of less technological importance. For the Faraday effect, the rotation of the plane of polarization of light is given by

$$\theta_F = Vl\mathbf{H}\cdot\boldsymbol{\epsilon}, \tag{10}$$

where $\mathbf{H}$ is the strength of the applied field, $\boldsymbol{\epsilon}$ is a unit vector in the direction of light propagation, and the constant of proportionality V is the Verdet constant. In the small-field limit, the Verdet constant is a function of ω, T, and n and may be of diamagnetic, paramagnetic, or ferromagnetic origin. The classical formula for the Verdet constant of diamagnetic materials, due to Becquerel, is

$$V_{\text{dia}} = -\frac{e\lambda}{2mc^2}\frac{dn}{d\lambda}. \tag{11}$$

This expression holds for most materials if a multiplicative factor γ, the magneto-optic anomaly, is introduced. This factor varies with the nature of the bonding in the material and ranges from near unity for predominantly ionic bonding to 0.28 for strongly covalently bonded diamond. The Verdet constant of diamagnetic materials exhibits only a small temperature dependence and in the long-wavelength limit (far from the absorption edge) varies as $1/\lambda^2$.

When paramagnetic ions are present in a material, Faraday rotation is dependent on the number of paramagnetic ions and the frequency of electronic transitions of the ions. Magnetic-optical rotation occurs for ions from various transition-metal groups (Buckingham and Stephens, 1966). In the long-wavelength limit, V_{para} is negative, again varies as $1/\lambda^2$, and, except for extremely low temperatures, is inversely proportional to temperature. All transparent materials exhibit diamagnetic Faraday rotation; however, when paramagnetic ions are present in sufficient quantities, paramagnetic Faraday rotation can dominate. Because the diamagnetic and paramagnetic effects produce opposite rotations, the sign of the net Verdet constant of a material can change with paramagnetic ion content.

The magneto-optic Kerr effect observed when light is reflected from the surface of a solid has the same physical origin as the Faraday effect.

1.8 Nonlinear Optical Properties

The polarization **P** induced in a medium by an external optical electric field E of frequency ω can be expressed by the power series

$$\begin{aligned}\mathbf{P}(\omega) = \mathbf{P}^{(1)} + \mathbf{P}^{(2)} + \mathbf{P}^{(3)} + \ldots = \chi^{(1)}(\omega)E(\omega) \\ + \chi^{(2)}(\omega = \omega_1 + \omega_2)E(\omega_1)E(\omega_2) \\ + \chi^{(3)}(\omega = \omega_1 + \omega_2 + \omega_3)E(\omega_1) \\ \times E(\omega_2)E(\omega_3) + \ldots, \qquad (12)\end{aligned}$$

where the complex dielectric susceptibilities $\chi^{(n)}$ are tensors of rank $(n + 1)$ and are related to the microscopic (electronic and nuclear) structure of the material (Bloembergen, 1965; Shen, 1984). Classical or linear optics is concerned with the first-order term $\chi^{(1)}$. The linear refractive index n_0 and the linear absorption coefficient k are proportional to the real and imaginary parts of $\chi^{(1)}$. In regions of no absorption, $n^2 - 1 = 4\pi\chi^{(1)}$.

Second-harmonic generation, the linear electro-optic (Pockels) effect, and sum- and difference-frequency mixing are described by the second-order nonlinear susceptibility $\chi^{(2)}$ in Eq. (12). In the first process an optical signal is generated at twice the frequency of the incident optical field, that is, $\omega_1 = \omega_2$ and $\omega = 2\omega_1$. In the second process, the refractive indices for the ordinary and extraordinary waves become different on account of the birefringence induced by the application of the strong dc field. The second-order nonlinearity term $\chi^{(2)}$ is important for those materials having point-group symmetry with no center of inversion. Materials that exhibit a linear electro-optic effect are also piezoelectric. The strain produced by the piezoelectric effect modifies the index of refraction, giving rise to the photoelastic or strain-optic effect.

When a material has inversion symmetry (glasses, plastics, liquids, and gases), the Pockels effect is absent and the observable effect of an applied electric field E on the refractive index is the dc Kerr effect, given by

$$\Delta n = BE^2\lambda, \qquad (13)$$

where the Kerr constant B is proportional to the real component of $\chi^{(3)}$. Thus in an applied field an isotropic material can exhibit birefringence due to this quadratic effect. The crystal symmetry determines the nonzero terms in the nonlinear coefficient tensor.

Other third-order nonlinear processes involving $\chi^{(3)}$ are third-harmonic generation, two-photon absorption, Raman and Brillouin scattering, self-focusing, self-phase modulation, optical-phase conjugation, and a large number of four-wave mixing processes (Hellwarth, 1977). The optical (ac) Kerr effect is a four-wave mixing process involving optical-frequency fields. Through the optical Kerr effect, the refractive index becomes intensity-dependent and is given by

$$n = n_0 + n_2 I, \qquad (14)$$

where n_0 is the linear index of refraction, I is the time-averaged intensity of the incident optical beam, and n_2 is the nonlinear refractive-index coefficient, derived from the real part of $\chi^{(3)}$. If a high-intensity beam propagating through a material has a nonuniform spatial profile, the spatial variation of the induced refractive index will lead to self-focusing (or -defocusing depending on the sign of n_2) (Bliss *et al.,* 1974). The intensity-dependent refractive index also causes the phase at the peak of an optical pulse to be retarded with respect to its leading or trailing edges. In addition, certain solutions of the nonlinear wave equation, called solitons, propagate without any change in shape (Mollenauer *et al.,* 1980).

The imaginary part of $\chi^{(3)}$ contributes to two-photon absorption. Raman and Brillouin scattering also arise from the imaginary part of $\chi^{(3)}$ and involve interactions with optical and acoustic vibrations of the material.

1.9 Luminescence and Stimulated Emission

Many forms of luminescence occur in optical materials: photoluminescence (excited by optical radiation), electroluminescence (*q.v.*) (excited by electrons in an applied electric field), radioluminescence (excited by ionizing radiation), thermally induced luminescence (from trapping centers), and cathodoluminescence (*q.v.*) (electron-beam excitation) (Blasse and Grabmeier, 1994). Luminescence may be due to a variety of ionic and molecular species (activator ions or molecules) listed in Table 1. Photoluminescence (*q.v.*) can arise from the presence of various transition-metal ions (principally iron-group and lanthanide-series elements having partially filled electron shells), filled-shell and post–transition-group elements, semiconductor clusters, molecular complexes, organic dyes, radiative electron–hole recombination, excitons, and defect or color centers. Fluorescence lifetimes range from nanoseconds or less to milliseconds or more depending on the degree of forbiddenness of the radiative transition probability. While many of the general features of the luminescence from a given species are similar whether the host is an amorphous or crystalline material, site-to-site variations in the local environments in the former result in a distribution of energy levels (inhomogeneous broadening) and of radiative and nonradiative transition probabilities (non-single-exponential excited-state decays).

Stimulated emission (laser action) requires an inverted ion population of the excited species (Siegman, 1994) and has been achieved for a large number of ions and transitions at wavelengths ranging from the far infrared to soft x rays (Weber, 1991a). Solid-state lasers incorporate many different transition-metal ions, organic dyes, and color centers in inorganic and organic crystalline and amorphous hosts. Lasing is produced principally by optical pumping. In semiconductor-junction laser materials, a large electron current creates excited electron and hole states and induces laser action (Yariv, 1989).

1.10 Optical Damage

The extremely large optical electric fields associated with the high peak powers and fluences of lasers can lead to dielectric breakdown and damage of reflecting and transmitting optical components (Wood, 1986). Because of the intensity-dependent index of refraction, a light beam having a nonuniform spatial profile will self-focus as it traverses a glass (Bliss *et al.*, 1974). If uncontrolled, self-focusing can lead to catastrophic damage in the form of bubbles or tracking along the light path. Self-focusing is minimized by using optical materials with small nonlinear refractive-index coefficients n_2 and by controlling spatial nonuniformities in the beam profile. Interaction of intense light with structural or chemical imperfections at the surface leads to damage in the form of pitting. Laser-induced damage usually arises from both extrinsic and intrinsic material properties. The damage threshold is usually dependent on the optical wavelength and pulse duration.

2. OPTICAL GLASSES

Glasses constitute a broad class of optical materials including oxides, halides, and chalcogenides (Kreidl, 1983). Their chemical

Table 1. Luminescent species in solids.

Species	Transition	Examples
Transition-metal ions	*nd*–*nd*	Mn^{2+}, Ni^{2+}, . . ; Ti^{3+}, Cr^{3+}, . . ,
Lanthanide/actinide ions	*nf*–*nf*	Pr^{3+}, Nd^{3+}, . . , Yb^{3+}
	nf–$(n+1)d$	Ce^{3+}, Eu^{2+}, Yb^{2+}
Filled-shell ions	nd^{10}–$nd^{9}(n+1)s$	Cu^{+}, Ag^{+}, Zn^{2+}, Ga^{3+}, Sb^{5+}
Post–transition-group ions	ns^2–$nsnp$	Sn^{2+}, Sb^{3+}, Tl^{+}, Pb^{2+}, Bi^{3+}
Molecular complexes (d^0 ions)	Charge transfer	VO_4^{3-}, WO_4^{2-}, TaO_4^{3-}
Color centers	$2p$–$1s$	LiF (F_2^{+}), KCl [F_A(II)]
Band-to-band	Electron-hole recombination	GaAs, (Ga,In)N
Donor-acceptor pair	Electron-hole recombination	GaP:S,Si; ZnS:Ag,Cl
Excitons	Electron-hole recombination	Alkali halides, ZnS, CdS, GaP
Core (hole)–valence (electron)	*np*(cation)–*n′p*(anion)	BaF_2, CsF, CsCl, RbF, $KMgF_3$
Organic molecules	Singlet–singlet; triplet–singlet	Anthracene, stilbene, rhodamine

composition consists of a glass network former such as silicate, phosphate, borate, etc. and numerous network-modifier cations added to improve durability, ease of glass forming, or other physical and chemical properties. With this compositional versatility, it is possible to tailor the chemical composition to optimize various desired properties. In addition to homogeneous glass, two-phase glasses, such as borosilicates (Pyrex), and glass ceramics (e.g., Zerodur), which contain fine crystals dispersed in a glass matrix, are used as optical materials. Glass ceramics can exhibit good optical transparency if there is a good match of the refractive index of the crystalline and glassy phases or if the dimensions of the crystals are small compared to the wavelength of light.

2.1 Classification

Commercially available glasses for optical components and ophthalmic use are principally colorless oxide glasses. They provide a wide range of refraction, dispersion, specific gravity, and other optical, thermal, and mechanical properties. Optical glasses are characterized and designated by their refractive index and dispersion. The most common measure is the refractive index at the wavelength of the He *d* line (587.6 nm) or the Na *D* line (589.3 nm). The difference in the refractive indices at the hydrogen *F* (486.1 nm) and *C* (656.3 nm) lines, $n_F - n_C$, is called the average or principal dispersion. The ratio $(n_F - n_C)/(n_d - 1)$ is called the relative dispersion; the reciprocal of this quality is the Abbe number ν_d. A six-digit number is used to specific optical glasses, where the first three digits designate the refractive index n_d with the preceding "1" omitted and the last three digits designate the Abbe number ν_d with the decimal point omitted. Thus a borosilicate glass BK 7 having an $n_d = 1.51680$ and $\nu_d = 64.17$ has a designation 517 642. This six-digit refraction index–Abbe number code is the most universally useful way of designating optical glasses used in standard applications in which precise n and ν values are not required.

Glasses having $n_d < 1.60$, $\nu_d > 55$ are called "crown" (K) glass; other glasses having $n_d > 1.60$, $\nu_d < 50$ are called "flint" (F). These letters, plus others, are usually contained in the manufacturer's designation of optical glasses. Designations vary with the country of origin. "Light" or "dense" indicate the relative amounts of heavy-metal oxides such as PbO or La_2O_3. Designations for various glass types are given in Table 2; the first designation is that used for Schott glasses (Schott, 1982), and designations used by other manufacturers are given in parentheses (Weber, 1991b). Only a few nonoxide optical glasses such as heavy-metal fluorides and chalcogenides are available commercially. These are specified by the manufacturer's name or by the composition if it is simple. Heavy-metal fluoride glasses are usually designated by their chemical composition, e.g., ZBLAN for a ZrF_4-BaF_2-LaF_3-AlF_3-NaF fluoride glass.

Table 2. Designations and types of commercial optical glasses. Schott designations are given; other designations are in parentheses.

Designation	Type	Designation	Type
BaF (BF, FB)	Barium flint	LaSK	Dense lanthanum crown
BaK (BaC, LBC)	Barium crown	LF (FL)	Light flint
BaLF (LBC, BCL)	Light barium flint	LLF (BLF, FEL)	Extra light flint
BaSF (DBF, FBD)	Dense barium flint	PK (PC)	Phosphate crown
BK (BSC)	Borosilicate crown	PSK (DPC, PCD)	Dense phosphate crown
F (DF, FD)	Flint	SF (EDF, FDS)	Dense flint
FK (FC)	Fluorocrown	SFS	Special dense flint
FP (FK)	Fluorophosphate	SK (DBC, BCD)	Dense barium crown
FZ	Fluorozirconate	SSK (EDBC, BCDD)	Extra dense barium crown
K (C)	Crown	TaF	Tantalum flint
KF (CF, CHD)	Crown flint	TaK	Tantalum crown
KzFS (ADF)	Dense short flint	TaSF	Dense tantalum flint
LaF (LaFL)	Lanthanum flint	TiF	Titanium flint
LaK (LaC, LaCL)	Lanthanum crown	TiK	Titanium crown
LaSF	Dense lanthanum flint	TiSF (FF)	Dense titanium flint

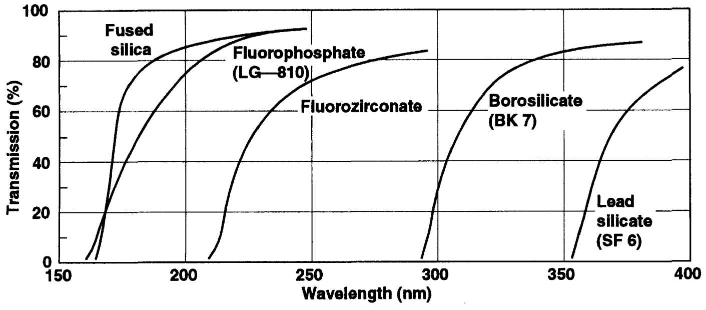

FIG. 3. Ultraviolet absorption edge of representative optical glasses. Sample thicknesses: 5 mm except for SiO_2 and LG 810, which are 2 mm. (From Cook and Mader, 1980; Schott, 1982).

Amorphous silica glass, SiO_2, is called fused quartz if made by melting crystalline quartz and fused silica if prepared synthetically by chemical vapor deposition starting with $SiCl_4$. The latter process can yield material of high purity with respect to metal impurities and OH^-. Various grades of silica are available that differ with respect to their UV and IR transmission properties and susceptibility to radiation damage.

2.2 Properties

2.2.1 Optical Properties Glasses exist that are optically transparent in the vacuum ultraviolet (VUV) and others in the mid-infrared. Figures 3 and 4 illustrate the variation of the ultraviolet (see also Sigel, 1977) and infrared absorption edges of representative optical glasses. The optical transmission in these figures depends on the wavelength dependence of the absorption, reflection, and scattering of the glass. The transmission of optical glasses is frequently given in terms of the internal transmission T_i after correction for reflective losses; that is, $T_i = T/R$, where R is the reflection factor given by Eq. (6). The deviation of T_i from 100% is a measure of absorption due to impurities and scattering due to defects. The internal transmittance is usually reported at a number of standard wavelengths from the ultraviolet to the infrared.

Optical glasses cover a general range of refractive indices $n_d = 1.4$ to 2.0 and reciprocal dispersions $\nu_d = 20$ to 90. These are almost exclusively oxide glasses and are shown in an n_d-ν_d plot in Fig. 5 where the lines divide the main compositional types in Table 2. Optical-glass companies supply such a figure with markers scattered throughout this parameter space indicating available glasses. The range of n_d and ν_d values shown in Fig. 5 is larger than that given in the usual glass catalog so as to include low-index, low-dispersion fluoride glasses (Baldwin *et al.*, 1981). Amorphous SiO_2 and BeF_2 are the extrema for oxide and halide glasses (many fluoroberyllate glasses transmit well into the vacuum ultraviolet). Many higher-index chalcogenide glasses cannot be located in an n_d-ν_d plot because the absorption edge extends into the visible and it is not always possible

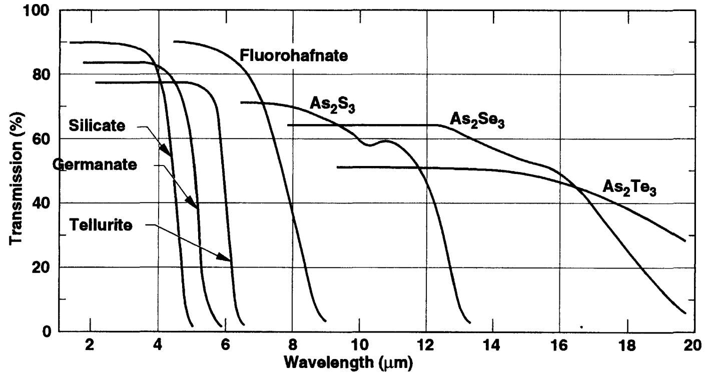

FIG. 4. Infrared absorption edge of representative optical glasses. Sample thicknesses: 2 mm. (From Dumbaugh, 1985.)

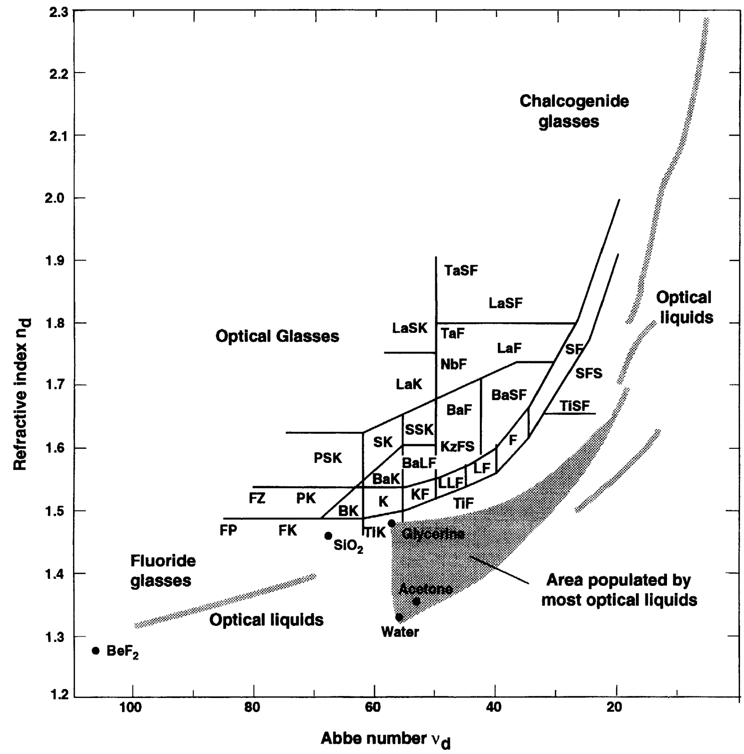

FIG. 5. Location of the optical glasses in Table 1 and optical liquids in a refractive index–Abbe number diagram.

to measure ν_d. For these infrared materials, plots are made using a reciprocal dispersion based on measurements of refractive index at longer wavelengths.

To interpolate values of the refractive index at the other wavelengths, glass manufacturers use an approximate dispersion formula derived from a power-series expansion of Eq. (3) of the form

$$n^2 = A_0 + A_1\lambda^2 + A_2\lambda^{-2} + A_3\lambda^{-4} + A_4\lambda^{-6} + A_5\lambda^{-8}, \qquad (15)$$

where the constants A_i are determined from a least-squares fit of the measured values (Schott, 1982). Using this equation, refractive indices in the wavelength range 370–1000 nm can be calculated to an accuracy of $\pm 5 \times 10^{-6}$ or better.

Various relative partial dispersions $P_{xy} = (n_x - n_y)/(n_F - n_C)$ are defined for other wavelengths x and y. The relative partial dispersion of most glasses obeys a linear relationship on ν_d of the form

$$P_{x,y} \approx a_{xy} + b_{xy}\nu_d, \qquad (16)$$

where a and b are constants. Because it is not possible to correct for second-order chromatic aberrations using so-called "normal" glasses that satisfy Eq. (16), optical-glass catalogs also list deviations of the relative partial dispersions from the normal for glasses covering a wide range of ν_d values.

2.2.2 Thermal Properties The temperature range in which a glass transforms from its solid state into a "plastic" state is called the transformation region. The viscosity of the melt at the transformation temperature T_g is approximately 10^{13} poise. A low T_g is important for molded optical elements as are the tendency of the glass to stick to the mold and the flatness of the viscosity-temperature dependence.

The coefficient of thermal expansion α of glass ranges from near zero for special low expansion glasses such as titania-doped SiO_2 and tailored glass ceramics to values greater than 20×10^{-6}/K. For optical glasses α ranges from about $(4 \text{ to } 16) \times 10^{-6}$/K. The thermal expansion coefficient increases with increasing temperature, exhibiting a nonlinear increase up to about room temperature, followed by an approximately linear range until the glass begins to exhibit plastic behavior, and then a rapid increase with increasing structural mobility in the glass.

The thermal coefficient of refractive index depends on the wavelength, temperature, and pressure. Both positive and negative values of dn/dT occur. The thermal conductivity κ of optical glasses ranges from about 0.5 to 1.5 W/m K, being high for silica and low for fluoride glasses and oxide glasses containing large quantities of heavy elements such as Pb, Ta, Ba, and La. The thermal conductivity of glass increases with temperature but only slightly above 300 K and is generally much smaller than that for crystals. Values of T_g, α, κ, specific heat c_p, and the hardness H (discussed below) are included in Table 3.

2.2.3 Mechanical Properties The mechanical response of a glass to an applied force is described by various moduli. Optical-glass catalogs usually list moduli such as Young's modulus E (extension in tension) and the modulus of rigidity or shear G, which are important for thermal and mechanical stress determinations. The stress-optical coefficient K varies with glass type and wavelength. It is usually positive, although it can become negative for silicate glasses having a high lead content (so-called Pockels glasses). Values of K generally range from $-2 < K < 4$ TPa^{-1} for oxide glasses to $-40 < K < 20$ TPa^{-1} for chalcogenide glasses.

The hardness H of a glass is an indicator of its vulnerability to surface damage and the ease with which it can be polished. It is usually measured from the indentation of Knoop or Vickers penetrators. Values (Knoop) for oxide glasses range from ~250 for high-lead-content glasses to >600 for lanthanum crown glasses. Representative values for a variety of glass types are included in Table 2. The Knoop hardness generally correlates with the Young's modulus.

2.2.4 Chemical Durability An important consideration for many optical glasses is their chemical reactivity with slurries during cutting and polishing of components such as lenses, windows, and prisms and with its environment where it may be subject to chemical attack by water, water vapor, other gases, acids, etc. in the atmosphere. No simple test and parameter is sufficient to characterize chemical reactivity and corrosion under all conditions. Thus many terms and tests are used by glass manufacturers to rank their glasses with respect to their resistance to acids, straining, climate, weathering, etc. This information is given in optical-glass catalogs.

2.3 Quality and Forms

Several hundred optical-glass types are available commercially. These are multicomponent glasses whose chemical compositions have been tailored to achieve the range of refractive indices and other physical properties discussed above. By careful batching, processing, and annealing, optical glasses can be made with very reproducible properties. The refractive index is controlled by the rate of annealing in the transformation range. Tolerances for n_d of ± 0.001–0.002 and for ν_d of $\pm 0.8\%$ are standard. The homogeneity of the refractive index is important for imaging systems; for selected melts or selected blanks, homogeneities of $\pm 10^{-5}$–10^{-7} are possible.

Table 3. Properties of common optical glasses (temperature range: 20–300 °C).

Glass type	ρ (g/cm^3)	H (Knoop)	α (10^{-6} K^{-1})	κ (W/m K)	c_p (J/g K)	T_g (°C)
Fused silica (SiO_2)	2.20	635	0.55	1.38	0.75	1175
Fluorophosphate (FK 54—487 865)	3.18	390	14.6	1.06	0.71	410
Fluorosilicate (FK 5—487 704)	2.45	520	9.2	0.92	0.81	464
Fluorozirconate (ZBLA)	4.61	240	18.7	~0.6	0.52	310
Borosilicate (BK 7—517 642)	2.51	610	7.1	1.11	0.85	557
Barium silicate (SK 14—603 606)	3.44	600	6.0	0.85	0.64	644
Alkali lead silicate (F 2—620 364)	3.61	420	8.2	0.78	0.56	438
Lanthanum borate (LaF 2—744 448)	4.34	560	8.1	0.67	0.48	658
Lead silicate (SF 6—805 254)	5.18	370	8.1	0.67	0.39	426
Arsenic trisulfide (As_2S_2)	3.20	180	26.1	0.17	0.47	163
Zerodur (Schott)	2.53	630	0.5	1.64	0.82	—
Pyrex (Corning 7740)	2.23	418	3.2	1.13	1.05	287

Residual stress birefringence varies with glass size and type. For some glasses values of ≤20 nm/cm are achievable for large (~1 m) diameter blanks. For smaller pieces (~100 mm), values of ≤4 nm/cm are possible by special annealing. Most glasses can be produced virtually striae and inclusion (bubble) free.

Optical glass is available in many forms including

1. slabs and blocks with two or more machined and polished faces used for test purposes,
2. strips, rods, and rolled glass with unfinished surfaces,
3. semi-finished and pressings of some required form, and
4. simple gobs.

Glass fibers are produced by controlled drawing from a melt or preform with diameters of <10 to >100 μm. Doped silica fibers are the most commonly used for optical applications (see FIBER OPTICS; OPTICAL COMMUNICATIONS), but fibers can be prepared from many different single- or multiple-component oxide and fluoride glasses. To reduce losses, a core glass is clad with another glass or plastic of lower refractive index so that light being transmitted is totally reflected internally. For ultralow-loss glass fibers for optical communications, losses of less than 1 dB/km have been achieved in the near infrared corresponding to an absorption coefficient of $\sim 10^{-6}$ cm^{-1}.

2.4 Special Glasses and Applications

2.4.1 Abnormal-Dispersion Glass The partial dispersions of optical glass are important in the design of compound lenses. By combining a positive lens of a crown glass with a negative lens of a flint glass, chromatic aberration can be avoided at two wavelengths. Correction for second-order chromatic aberration (secondary spectrum) is accomplished using glasses with equal partial dispersions for different Abbe values. Abnormal-dispersion glasses depart from the "normal line" and the linear relationship in Eq. (16). These glasses are produced by tailoring the chemical composition of the glass so as to alter the relative strength and positions of the ultraviolet and infrared absorption bands of the constituents and thereby change the dispersion.

2.4.2 Gradient-Index Glass A conventional glass lens consists of a homogenous material with a single index of refraction. Glass elements can also be manufactured in which the index of refraction varies continuously within the material. These are called gradient-index or GRIN materials. The refractive-index gradient may be axial (varying along the optical axis), radial (varying outwardly from the optical axis), or spherical (varying symmetrically from a point) (Moore, 1980). For imaging systems, gradient-index glass provides an alternative to the generation of complex aspheric lens surfaces to control aberrations. Gradient-index rods and fibers are used in micro-optics, imaging optics, and optical-fiber communication systems. Techniques for creating refractive-index variations include neutron irradiation, chemical vapor deposition of layers of glass of slightly different refractive indices, ion exchange by immersing the glass in an appropriate solution, ion stuffing in porous glass, and sol-gel methods.

2.4.3 Mirror-Substrate Glass For front surface mirrors where retention of the optical figure is important, the thermal stability of the environment is a major consideration. In a temperature-controlled environment, low-cost borosilicate glasses can be used for mirror substrates. If, however, the substrate is subject to temperature variations, materials with the lowest possible coefficient of thermal expansion are needed. Silica containing small quantities of titanium dioxide or other heavy-metal oxides and glass ceramics can be produced that are hard materials with extremely small, near-zero coefficients of thermal expansion ($\alpha \leq 0.1 \times 10^{-6}$/K). Glass ceramics (such as Zerodur) are solid solutions in which the very small or even negative thermal expansion coefficient of the crystalline phase compensates for the positive expansion coefficient of the remaining glass matrix. Glass-ceramic mirror blanks with diameters of more than 8 m have been produced for telescope mirrors.

2.4.4 Optical-Filter Glass Transmitting filter glasses change the spectral properties of optical radiation and provide selective filtering for scientific and technological applications, for color enhancement in ophthalmic lenses, and for contrast enhancement in displays (see OPTICAL FILTERS). Filters are categorized as band-passing and band-blocking filters, short-wavelength and long-wavelength cutoff filters, and neutral-density and color-conversion filters. The transmission characteristics are determined by a combination of the base glass and ion doping or colloidal coloration. The UV and IR cutoff wavelengths vary greatly depending on the glass composition; thus the composition can be tailored to achieve selected transmission characteristics. Other cutoff and bandpass filters are produced by doping with transition metals, rare earths, and uranium (Cook and Stokowski, 1986). The fundamental absorption edge of semiconductor-doped glasses depends on the size and absorption characteristics of the colloidal particles. By careful heat treatment and processing of Cd(S,Se)- and CdTe-doped silicate and borosilicate glasses, a series of long-pass filters can be manufactured with sharp cutoff wavelengths varying over a wide spectral range in the visible and near infrared.

Optical-filter glass, like other optical glasses, is manufactured in high optical quality, free of striae and inclusions, and in a variety of forms. Transmission is carefully controlled. Filter-glass catalogs present extensive data on transmittance; data on refractive index and other thermal and mechanical properties and chemical durability are also listed. Under prolonged exposure to high-intensity ultraviolet radiation, some filter glass may solarize and exhibit reduced transmission, particularly at shorter wavelengths. Some ionically or colloidally colored filter glasses also exhibit *photoluminescence* (Turner, 1973).

2.4.5 Nonlinear Optical Glass Optical nonlinearities in glass arising from the ac Kerr effect are of interest in two extremes: in high-power lasers where glass with a small nonlinear refractive index n_2 is desired to reduce self-focusing and optical damage, and in all optical switching devices where glass with a large n_2 is desired to minimize the optical power requirements (Vogel *et al.*, 1991). The magnitude of the optical Kerr effect is known for a wide range of undoped and doped glasses (Borrelli and Hall, 1991). Glasses with small values of linear refractive index and dispersion generally have small n_2 values (Adair *et al.*, 1987). The smallest linear and nonlinear refractive indices for glass occur for fluoride-containing glasses (Weber *et al.*, 1978).

For all-optical switching devices, materials are sought having a large n_2 and a small absorption loss coefficient at the operating wavelength. Very large resonant optical nonlinearities are observed in semiconductor-doped glasses (glasses doped with I–VII, II–VI, and III–V compounds). The nonlinear behavior appears as intensity-dependent changes in absorption and refractive index with response times of $<10^{-11}$ s originating from the generation of a short-lived free-electron plasma and band-filling effects. Organic dye molecules have intense singlet–singlet transitions and have been added to low–melting-point glasses such as boric-acid glass and lead-tin fluorophosphate glass (Tompkin *et al.*, 1987) without decomposition. Saturation of the ground-state absorption of the dye produces optical nonlinearities with a very large effective $\chi^{(3)}$. Because the response time is governed by the relaxation time of the triplet state of the dye, recovery of the saturated nonlinearity is very slow compared to that of nonresonant electronic nonlinearities. Glasses containing very small metallic particles also exhibit an efficient, fast optical Kerr effect (Vogel *et al.*, 1991).

The dc Kerr constants for glasses are in the order of or less than 2×10^{-14} m/V^2 (Borrelli and Hall, 1991). The largest values are found for glasses containing large quantities of heavy, polarizable ions such as Nb, Ta, Tl, Pb, and Bi. Although glass is macroscopically isotropic and therefore it is not expected to exhibit second-harmonic generation due to the second-order nonlinear susceptibility $\chi^{(2)}$, several photoinduced nonlinear optical effects are observed after a glass has been exposed to light for periods of time varying from minutes to hours (see Vogel *et al.*, 1991, and references cited therein). These include refractive-index gratings and second-harmonic generation in fibers. These effects have been reported primarily for Ge-P- or Ge-doped silica fibers.

2.4.6 Laser Glass A large number of different lasing ions (principally trivalent lanthanides) and host glasses (oxides and fluorides) have been used for glass lasers (Hall and Weber, 1991). The forms used for laser glasses include bulk, fiber, and planar waveguide. The spectral range of glass lasers extends from ~0.4 to 3.9 μm. The spectroscopic properties important for laser action vary with the chemical composition of the host. Although a large number of experimental laser glasses have been reported in the literature, commercial laser glasses are almost exclusively silicates and phosphates. Fluorophosphate glasses have been developed for applications requiring low n_2 values. Phosphate glasses generally have narrower linewidths and, for ultraphosphates, have less fluorescence concentration quenching than for silicate glasses; thus they are more appropriate for applications requiring high gain and efficient energy extraction (Marion and Weber, 1991).

Laser glass shares with optical glass the requirements of refractive-index homogeneity (variations $\leq 10^{-6}$), high purity (absorption coefficient at the lasing wavelength of $\leq 10^{-3}$ cm^{-1}), freedom from bubbles and striae, low birefringence ($\leq$5 nm/cm), good chemical durability, and capacity of being finished to high optical tolerances (λ/20). High-quality phosphate laser glasses satisfying these requirements have been produced with dimensions greater than 0.5 m. Athermal phosphate glasses have been developed in which the change of optical path length with temperature is near zero (Izumitani, 1986).

2.4.7 Faraday-Rotator Glass The Faraday effect has been utilized in the construction of fast optical switches, modulators, circulators, isolators, and sensors for magnetic fields and electric currents. Diamagnetic and paramagnetic glasses have been used in bulk and fiber forms for these applications. The largest diamagnetic Verdet constants are found for high-index, high-dispersion glasses located in the upper right-hand corner of Fig. 5. Chalcogenide glasses have very high dispersion and large Verdet constants and are appropriate for use in the infrared. The temperature dependence of V_{dia} is very small, which makes diamagnetic glasses attractive for high-stability Faraday-effect sensors.

The magnitude and the wavelength and temperature dependences of paramagnetic glasses vary depending upon the choice of the paramagnetic ion. For lanthanide-series ions Ce^{3+}, Tb^{3+}, and Eu^{2+} have large regions of transparency in the visible-near infrared region and large paramagnetic Verdet constants. Because V_{dia} and V_{para} have opposite signs, the best host glasses are those having low dispersion (therefore small V_{dia}) and capable of incorporating large quantities of the paramagnetic ion. Verdet constants for several diamagnetic and paramagnetic glasses are given in Table 4.

2.4.8 Acousto-optic Glass Acoustic waves create a time-varying refractive-index grating in a material via the photoelastic effect. The grating spacing is equal to the acoustic wavelength; the grating depth is determined by the drive power of the transducer. A light beam traversing the medium is deflected by the grating. This phenomenon provides fast deflection or modulation of light and is used in display technology. An acousto-optic material, in addition to having low losses at the acoustic and optical wavelengths, should also have a large index of refraction and small sound velocity. These include materials containing heavy cations such as lead silicate or tellurite glasses and chalcogenide glasses. If a large bandwidth is desired, materials should have a large sound velocity. In Table 5 the properties of glasses are compared to those of other acousto-optic materials. Glasses are attractive for many applications because they are isotropic and can be prepared in large sizes of excellent optical quality.

Table 4. Verdet constants of diamagnetic and paramagnetic materials at 633 nm and 300 K.

Crystals	V (rad/T m)	Glasses	V (rad/T m)	Liquids	V (rad/T m)
NaCl	8.5	Fused silica (SiO_2)	3.7	Water (H_2O)	3.35
$Bi_4Ge_3O_{12}$	30.1	Lead silicate (SF 59)	29	Ethanol (C_2H_5)H)	2.90
ZnS	83.4	Arsenic trisulfide (As_2S_3)	74	Benzene (C_6H_6)	7.39
CeF_3	−114	Cerium phosphate (FR-4)	−31	Carbon disulfide (CS_2)	10.4
$Tb_3Al_5O_{12}$	−180	Terbium borosilicate (FR-5)	−71	Arsenic trichloride ($AsCl_3$)	12

Table 5. Properties of acousto-optic materials.

Material	Transmission range (μm)	Density (g/cm^3)	Refractive index (633 nm)	Sound velocity[a] (km/sec)
Liquids:				
Water (H_2O)	0.3–0.9	1.0	1.33	1.5
Glasses:				
Fused silica (SiO_2)	0.2–4.5	2.2	1.46	6.0
Lead silicate (SF 59)	0.46–2.5	6.3	1.95	3.2
Tellurite (AOT 5)	0.47–2.7	—	2.09	3.4
Arsenic trisulfide (As_2S_3)	0.6–11	3.2	2.61	2.6
Crystals:				
$LiNbO_3$	0.40–4.5	4.7	2.20–2.29	6.6
TeO_2	0.35–5	6.0	2.26	4.2 (L) 0.62 (S)
$PbMoO_4$	0.4–5.9	6.9	2.62	3.7
GaAs	1–15	5.3	3.4 (1.9 μm)	5.5 (L) 3.3 (S)
Ge	2.0–20	5.3	4.0 (2.8 μm)	5.5

[a] L, longitudinal wave; S, shear wave.

2.4.9 Radiation-Detection Glass Glass is used in two forms for detection of high-energy radiation and particles: as a pure glass for the generation of Cherenkov radiation by energetic particles and as a doped glass for scintillation caused by x rays, gamma rays, charged particles, and neutrons. Since *Cherenkov radiation* is intense in the blue-violet, for the radiation to be detected the glass must maintain good transparency in this spectral region. Large blocks of lead-silicate optical glass having a high refractive index, $n \geq 1.7$, are used in Cherenkov counters capable of detecting particles traveling at 0.6 of the speed of light in vacuum.

High-energy charged particles and ionizing radiation are detected by scintillation from glass activated by a fluorescing species such as Ce^{3+}. Absorption of ionizing radiation leads to the creation of large numbers of electron–hole pairs, subsequent excitation of the activator, and fluorescence, which is detected. Glasses containing ^6Li, which has a large neutron-capture cross section, are used for detecting neutrons. Scintillation glasses are available in bulk form for energy calorimetry and in fiber form for particle tracking. The light yield of glass scintillators is usually much less than that of the best crystalline scintillators because of low transfer efficiency from the host to the activator.

2.4.10 Radiation-Shielding Glass Glass is used for remote viewing of radiation environments such as x-ray facilities, hot cells in nuclear research, irradiation facilities for food preservation, and medical sterilization and therapeutic installations. In these applications the optical glass must both provide adequate shielding and not discolor during exposure to the radiation. High–lead-content glasses are used for radiation-shielding glass. These glasses have about one-half the absorption of lead of comparable thickness.

2.4.11 Photosensitive Glass Photosensitive glasses are a special class of glasses whose optical properties are changed by optical/thermal treatment. Among these are photochromic glasses that change color under ultraviolet or short-wavelength light, then revert back to the original state when the light is reduced. In ophthalmic and architectural applications, the transmission of such glasses in the 400–700-nm region can be reduced by more than 50% when exposed to bright light. Photochromism is observed in strongly reduced alkali silicate and other multicomponent silicate and borosilicate glasses, in glasses containing suspensions of crystallites, and in glasses colored by silver halides. Many different base glasses can be used; the choice of composition, especially the halide, and thermal treatment influence the photochromic properties. The mechanisms involved in photochromism are complex and include treatment of precipitation kinetics and charge-carrier diffusion (Araujo, 1977). Most glasses are photosensitive in the negative sense that they become colored (so-

Table 6. Properties of common optical plastics.

Polymer (trade names)	n_D	ν_D	Density (g/cm^3)	Relative haze	Hue (coloration)
Polyurethane (Isoplast)	1.65	—	1.2	Light	Colorless
Polysulfone (Udel)	1.63	22.5	1.24	Light	Yellow
Polystyrene (styrene, Dylene, Styron, Lustrex)	1.589	31	1.06	Low	Colorless
Polycarbonate (Lexan, Merlon, Calibre)	1.585	30	1.2	Slight	Light straw
Styrene acrylonitrile (Lustran, Tyril, SAN)	1.571	35.3	1.07	Slight	Light straw
Polystyrene-comethylmethacrylate (NAS)	1.564–1.535	35–41	1.09–1.13	Slight	Colorless
Allyd diglycol carbonate (CR-39, ADC)	1.498	59.3	1.32	Low	Light straw
Polymethylmethacrylate (acrylic, Lucite, Plexiglas)	1.491	57.4	1.19	Low	Light straw
Fluoropolymer (Teflon AF)	1.29–1.31	—	—	Clear	Colorless

larize) after exposure to ultraviolet or ionizing radiation. To prevent solarization in optically pumped laser glass and in glass used for UV lamps, small quantities ($\lesssim 1\%$) of polyvalent ions such as cerium, antimony, and molybdenum are added to the glass.

3. OPTICAL PLASTICS

Many clear and colored plastics are of interest for optical applications. Optical plastics differ from optical glasses in several ways that make them uniquely suited for certain components. Their advantages are that they are less costly, lightweight, and not fragile, and they can be formed into complex shapes using simple fabrication techniques; their disadvantages are that they have a low tolerance for heat and abrasion, are subject to haze (internal scattering) and solarization (radiation-induced coloration), and have a narrower range of optical properties than other solid materials. Because plastics are normally formed in molds, nonspherical (aspheric) surfaces and nonrotationally symmetric surfaces are nearly as easy to produce as spherical surfaces. Although not every polymer exhibits the high clarity and absorption-free visible transmission associated with acrylic plastics, some may have other properties such as solvent resistance, thermal stability, or moldability that make them useful in particular optical applications.

3.1 Optical and Physical Properties

A number of the optical properties of some common plastics are summarized in Table 6. The refractive indices of plastics are in the general range from 1.3 to 1.7. The location of several optical plastics are shown in a refractive index–Abbe number diagram in Fig. 6 and compared to the larger, more extensive range of optical glasses. The maximum useful operating temperature for most plastics is ≤100 °C. The temperature dependence of the refractive index is large, in the

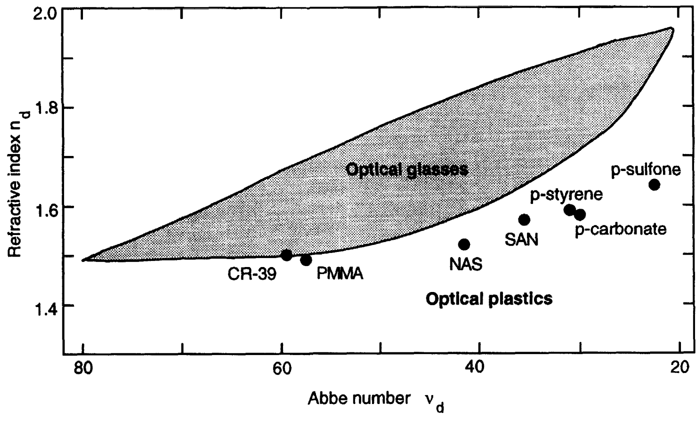

FIG. 6. Location of optical glasses and optical plastics in a refractive index–Abbe number diagram.

order of (1–2) × 10^{-4}/K for glassy polymers compared to ~3 × 10^{-6}/K for a typical optical glass. The coefficient of linear expansion with temperature is also large: (6–7) × 10^{-5}/K for a typical polymer compared to (5–10) × 10^{-6}/K for a typical glass. With such large temperature dependences, thermal changes must be considered in any optical design involving plastics. A laser beam traversing a plastic may also exhibit "blooming" as a result of thermally induced refractive-index changes. Because the temperature dependence of the refractive index is negative for polymers, it is possible to design thermally compensated glass/polymer optical doublets.

Transmission spectra of common optical plastics are given in Fig. 7. The useful transmittance range for most polymers is from about 380 to 1000 nm. The short-wavelength limit for 50% transmission through a 3-mm path length is approximately 380 nm for polystyrene, 340 nm for polycarbonate, 280 nm for ultraviolet acrylic, and below 250 nm for fluoropolymers (where some or all of the hydrogen atoms are replaced by fluorine atoms). Most of the near-infrared region above 1000 nm is absorbed by C–H vibrational overtone and combination bands, except for transmission windows around 1300 and 1500 nm. Optical properties can also change with water content and the humidity of the environment.

Transmission losses are process and thickness dependent; therefore, haze and coloration are described by relative values. Most high–refractive-index plastics may have some intrinsic yellowness, but other plastics are clear with good transmission throughout the visible and into the near-ultraviolet. Haze levels well below 0.5% for 5-mm thicknesses of plastics are readily possible. For some materials the haze can be reduced at the sacrifice of some other properties. Lower haze is often achievable in casting materials (e.g., cast polymethylmethacrylate, PMMA) since these can be filtered more easily than molding compounds. Plastics are marketed either with or without chemical additives to prevent or reduce yellowing or the development of haze. Some clear resins are supplied with a blue "toner" as an optical brightener; the color additive imparts a "water clear" appearance by compensating for the slight yellow coloration of some plastics. Plastic fibers for optical transmission may have typical losses of about 400–500 dB/km, but fibers have been produced with losses of ~13 dB/km.

3.2 Applications

Stryene, polycarbonates, and acrylics are thermoplastics. Styrenes are low-cost, high-clarity, high-index materials; polycarbonates have high heat and impact strength but poor scratch resistance; acrylics have high scratch resistance and good optical transmission and hence are used for ophthalmic lenses. Thermoplastics (which can be reheated to a viscous state and reformed) are generally processed by injection or compression molding. Very high-volume or large flat pieces may be extruded onto rollers. Thermosetting plastics and epoxy are usually cast and are able to be formed only once. The resins are usually received from the vendor in monomer form or in a partially polymerized state. The material

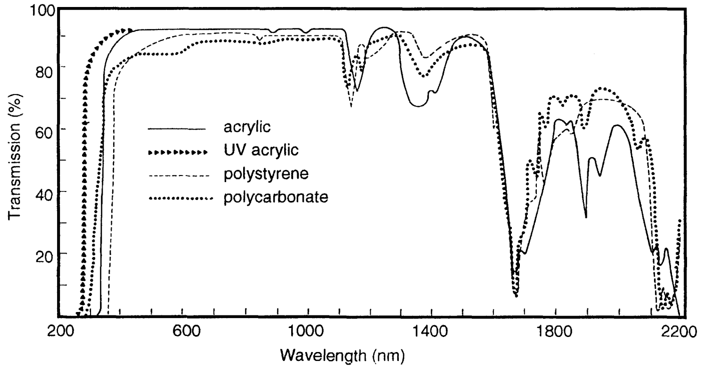

FIG. 7. Transmission spectra of optical plastics. Sample thickness: 3.2 mm. (From Welham, 1979.)

is then cast or heated by compression or transfer molding to finish the polymerization. The classical injection-molding process typically leaves some inhomogeneities in the material due to shear and cooling stresses. Cast thermosets exhibit fewer of these defects. Acrylic (PMMA), (poly)stryene, and the (poly)urethanes can be obtained in monomer form for casting.

Small-volume or prototype optical parts are often machined from cast or molded and annealed blanks. With diamond-turning technology, a number of polymers can be machined into optics such that polishing is no longer necessary. Reflective, antireflective, and dichroic coatings have been successfully applied to many optical plastics.

Dichroic polarizers are produced by stretching sheets of plastic or rubbing the surface of a plastic in one direction, thereby aligning long-chain polymeric molecules that make up polyvinyl alcohol. The layer is then treated with polymeric iodine or other appropriate dye. These are useful polarizers in the visible and have maximum-to-minimum transmission ratios of 10^{-3}–10^{-5}.

Pigments and dyes are used to provide transparent colored optics. Dye molecules can also be incorporated into plastics and used for optically pumped, tunable dye lasers. The useful lifetime of such materials depends on the resistance of the dye to photodegradation and varies with the dye and the host.

Plastics used for scintillator materials are produced by adding various chromophors such as *p*-terphenyl to polystyrene or polyvinyltoluene host materials (Knoll, 1989). Absorption of radiation by the host material results in fluorescence that is transferred radiatively or nonradiatively to the chromophore, which then emits radiation at a longer wavelength. Plastics are susceptible to radiation-induced coloration in the near-ultraviolet and visible. Since this absorption reduces scintillation light output, one or more wavelength-shifting dyes are frequently added, which absorb radiation from the first dye and re-emit it at a longer wavelength that is less subject to absorption. Plastic scintillators have relatively large light outputs, but because their density and effective atomic number are low, they are not efficient for detecting gamma rays but are used in large volumes for charged-particle detectors.

4. OPTICAL CRYSTALS

Crystals exist with isotropic properties comparable to amorphous materials; other crystals have anisotropic optical properties that make additional optical phenomena possible such as polarization and phase matching in wave mixing (Yariv and Yeh, 1984). Although crystals have a fixed chemical composition and therefore their composition cannot be tailored to achieve properties as in the case of glasses, an extremely wide range of synthetically produced crystalline materials are available with physical properties that in most cases exceed those available with glasses. Hundreds of optical crystals have been synthesized using a variety of techniques including flux growth, flame fusion (Verneuil), Bridgman, pulling (Czochralski), floating zone, heat exchanger method, chemical vapor deposition (CVD), and hydrothermal (Musikant, 1985). The obtainable sizes may be limited in some cases, but many crystals of dimensions of several centimeters or more (approaching 1 m) can be prepared. Single-crystal fibers are prepared using a laser-heated–pedestal growth technique (Yen, 1995). Several cubic materials are prepared in large sizes as polycrystalline materials by hot pressing (e.g., MgO, NaCl, MgF_2, CaF_2, ZnS, ZnSe, CdTe). The cost of crystalline materials depends on the availability and purity of the starting material, the cost of growing or processing the crystal to the size required, and final fabrication and surface finish of the material required for the application of interest. The plethora of crystalline materials is available not only for transmitting optics but as hosts for solid-state laser and nonlinear-optics applications.

4.1 Optical Properties

The fundamental band gaps of crystals range from >10 eV for some insulating fluoride materials to ~0.1 eV for some semiconductors. The dependence of transmission on chemical composition and bonding is similar to that of the glasses as shown in Figs. 3 and 4; however, several crystals transmit further into the UV than do the glasses in Fig. 3. The refractive indices of crystals also extend over the same general range of the n vs ν_d space in Fig. 5 and have varying dependences on both wavelength and temperature.

Table 7. Properties of crystals having cubic symmetry. Refractive index measured at 0.6 μm unless noted otherwise.

Material	Transmission range (μm)	Refractive index	Density (g/cm^3)	Hardness (kg/mm^2)	Melting point (K)	Thermal expansion (10^{-6}/K)	Therm. cond. (W/m K)
Oxides							
MgO	0.16–9	1.74	3.58	690	3070	10.6	59
$MgAl_2O_4$	0.21–6	1.72	3.58	1140	2300	7.0	14
Y_2O_3	0.22–8.6	1.92	5.01	875	2650	6.6	27
$Y_3Al_5O_{12}$	0.2–6	1.82	4.56	1350	2220	7.7	13
$SrTiO_3$	0.4–7	2.39	5.12	595	2350	8.3	11
Halides							
LiF	0.12–8	1.39	2.64	110	1140	34.4	11
CaF_2	0.13–12	1.43	3.18	158	1630	18.9	10
NaCl	0.19–20	1.53	2.16	18	1070	41.1	6.5
KBr	0.25–30	1.56	2.75	7	1000	38.5	4.8
CsI	0.25–60	1.78	4.51	—	894	48.6	1.1
Chalcogenides							
ZnS	0.4–14	2.35	4.09	17	2100	6.8	27
ZnSe	0.5–20	2.59	5.42	137	1790	7.1	19
CdTe	0.9–30	2.81 (10.6 μm)	6.20	56	1320	5.0	6.3
Other crystals							
C (diamond)	0.22–2.5, 6–100	2.42	3.51	5,700–10,400	3770	1.2	660
Si	1.2–7.0	3.48 (1.5 μm)	2.33	1150	1690	2.6	160
Ge	1.8–23	4.05 (2.8 μm)	5.35	800	1210	5.7	59
BN	0.2–6	2.12	3.48	4600	3500	3.5	~1300
InP	0.9–14	3.43	4.8	430	1320	4.5	70
GaAs	0.9–15	3.35 (1.9 μm)	5.32	720	1240	5.0	52

The temperature variations of n are in the general range of 10^{-5}/K but may be an order of magnitude larger or smaller. The range of transmission and refractive index of representative cubic crystals are given in Table 7. As noted earlier, whereas cubic crystals have a single index of refraction, uniaxial crystals have two (n_0,n_e), and biaxial crystals have three (n_1,n_2,n_3). Examples of the magnitude of the birefringence of crystals are tabulated in Table 8. The dependence of the refractive index on the direction of vibration of plane-polarized light traveling through a crystal can be described by an ellipsoidal surface, the optical indicatrix. The orientation of the indicatrix is related to the symmetry directions of the crystal.

Crystalline materials exhibit a metallic-like reflection at approximately twice the infrared cutoff wavelength. This reststrahten reflection is associated with fundamental resonant vibrations of the lattice.

For a discussion of nonlinear optical properties, see OPTICS, NONLINEAR.

4.2 Physical Properties

Optical crystals are generally superior to glasses and plastics with respect to their thermal and mechanical properties. Thermal conductivities of crystals may be one or more orders of magnitude greater than for glasses. The thermal conductivity of crystals is usually smaller the more atoms there are in the unit cell. Crystals exhibit a large range of hardnesses and fracture strengths; again

Table 8. Birefringence of common optical crystals at 0.60 μm and 300 K.

Crystal	Transmission (μm)	n_e–n_0	Crystal	Transmission (μm)	n_e–n_0
Al_2O_3 (corundum)	0.19–5.1	−0.008	MgF_2	0.13–7.7	0.012
$CaCO_3$ (calcite)	0.21–3.3	−0.172	SiO_2 (quartz)	0.16–4.0	0.009
CdS	0.51–14.8	−0.018	TiO_2 (rutile)	0.42–4.0	0.287
$LiNbO_3$	0.35–4.2	−0.092	YVO_4	0.35–4.8	0.023
KH_2PO_4 (KDP)	0.20–1.5	−0.041	$ZrSiO_4$ (zircon)	0.4–4.0	0.055

optical crystals exist that are greatly superior to glasses in these respects. Some crystals, however, have a tendency to cleave. As in the case of the refractive index, many physical properties of crystals are anisotropic (thermal conductivity, thermal expansion, and hardness). Melting points, hardnesses, thermal conductivities, and thermal expansion coefficients of several cubic crystals are included in Table 7.

4.3 Applications

4.3.1 Optical Components The availability of large, uniform crystals of high purity make them attractive for use as transmitting optical components such as windows, lenses, polarizers, transmission gratings, and selective reflectors. Cubic crystals, such as those listed in Table 7, or uniaxial crystals, with light propagating along the optic axis, are used for window materials and lenses covering a wide range of wavelengths—from the VUV through the IR. Crystals are especially useful in high-power optical applications where their toughness and thermal shock resistance are superior to those of glass. Alumina (Al_2O_3, corundum, and sapphire) is available in many sizes and forms and is attractive for many applications because of its hardness (1370 kg/mm^2), high melting temperature (2300 K), transmission (0.19–5.1 μm), and good thermal properties ($\alpha = 5.6 \times 10^{-6}$/K and $\kappa = 33$–35 W/m K).

The birefringence properties of crystals are utilized for various types of polarizers. A number of commonly used birefringent materials are listed in Table 8. Birefringent crystals are available for use throughout the visible and infrared regions.

4.3.2 Electro-optic Crystals Electro-optic crystals are noncubic and lack a center of symmetry. Of the 32 point-group symmetries, 11 are centrosymmetric and exhibit only third-order electro-optic effects (dc Kerr effect). The properties of a large number of crystals have been investigated; some of the more commonly used electro-optic crystals are listed in Table 9. Their ranges of transparency extend from the VUV to the mid-IR. Materials such as BBO, LBO, KTP, and KDP are useful in the UV–visible region; proustite, ZnSe, $ZnGeP_2$, and CuCl are useful in the infrared region.

In addition to inorganic crystals, many organic molecular crystals exhibiting large nonlinear coefficients have been investigated (see, for example, Chemla and Zyss, 1987). In many instances, obtaining molecular crystals of sufficient size, optical quality, and photochemical stability has been a limitation to their usefulness. Urea [$CO(NH_2)_2$], however, has been used successfully in nonlinear applications.

The Pockels effect can be used to control the intensity and phase of propagating radiation and therefore has many applications for light modulators and switches. The effect is characterized by tensor coefficients r_{ij}, where $i = 1, 2, 3$ and $j = 1$–6. The number of nonzero Pockels coefficients depends on the crystal symmetry class and is tabulated in various books (Nye, 1976). The number of wavelengths of retardation is given by $n_0^3 r_{63} V/\lambda$, where V is the applied voltage. Optical properties and the half-wave voltage required to produce a phase shift of π for an electric field applied parallel or transverse to the direction of light propagation are tabulated for several representative crystals in Table 10. Lithium niobate is widely used as an electro-optic light modulator.

A key to utilization of the second-order nonlinearity for optical harmonic generation is phase matching to compensate for material dispersion (Yariv and Yeh, 1984). For efficient energy transfer from the fundamental

Table 9. Electro-optic crystals and regions of optical transparency.

Crystal	Transmission (nm)	Crystal	Transmission (nm)
Ag_3AsS_3 (proustite)	0.63–12.5	$KTa_{0.65}Nb_{0.35}O_3$ (KTN)	0.39–6
β-BaB_2O_4 (BBO)	0.19–2.5	$KTiOPO_4$ (KTP)	0.35–4.5
$Ba_2NaNb_5O_{15}$	0.4–5	LiB_3O_5 (LBO)	0.16–2.6
$BaTiO_3$	0.4–7	$LiNbO_3$	0.35–4.2
CuCl	0.15–15	SiO_2 (quartz)	0.16–4.0
KH_2PO_4 (KDP)	0.20–1.5	ZnSe	0.51–19
$KNbO_3$	0.4–5.5	$ZnGeP_2$	0.67–12

Table 10. Half-wave voltage for electro-optic crystals where the electric field is applied longitudinally (L) or transverse (T) to the direction of light propagation.

Crystal	Transmission (μm)	n_0	n_e	λ (μm)	$V_{\lambda/2}$ (kV)
Uniaxial symmetry					
KH_2PO_4	0.20–1.5	1.512	1.470	0.546	7.65 (L) 17.8 (T)
$NH_4H_2PO_4$	0.19–1.7	1.526	1.481	0.546	9.0 (L) 7.65 (T)
$LiNbO_3$	0.35–4.2	2.297	2.208	0.60	2.9 (T)
$LiTaO_3$	0.4–5	2.183	2.188	0.60	2.7 (T)
$Ba_2NaNb_5O_{15}$	0.4–5	2.32	2.22	0.633	4.55 (T)
Cubic symmetry					
$Bi_4Ge_3O_{12}$	0.35–6	2.07		0.543	14.8 (T)
$KTa_{0.65}Nb_{0.35}O_3$	0.39–6	2.29		0.633	0.58 (T)
GaAs	0.9–15	3.42 (1.9 μm)		0.546	4.5 (T)
ZnSe	0.51–19	2.65		0.546	7.1 (T)
CuCl	0.4–20	2.00		0.546	6.2 (L,T)

to the second-harmonic wave, the phase velocities of the two waves must match. Because the refractive indices of birefringent crystals depend on the direction and polarization of the light, crystal orientations are selected to achieve phase matching for the two frequencies. The general condition for phase matching for second-order nonlinear processes, in which two photons of different frequencies and momenta **k** either add or subtract to create a third photon, is

$$\mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_3. \tag{17}$$

So-called type I phase matching, where $2\mathbf{k}_1^{(o)} = \mathbf{k}_2^{(e)}$, is illustrated in Fig. 8.

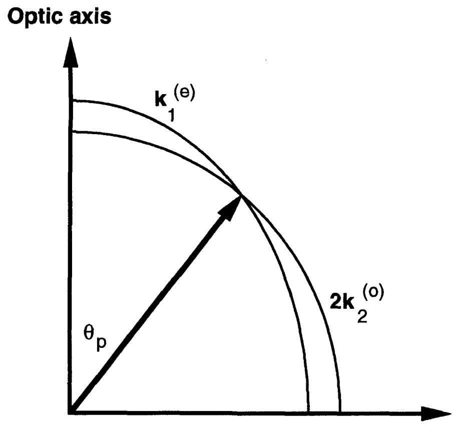

FIG. 8. Principle of phase matching in birefringent materials for second-harmonic generation.

4.4 Magneto-optic Crystals

All transmitting crystals exhibit diamagnetic Faraday rotation and, as noted for glasses, this rotation is largest for high-index, high-dispersion materials. Paramagnetic Faraday rotation depends on the presence of paramagnetic ions. Since these may be present as a constituent of the crystal rather than as a dopant, their concentration can be greater than is possible in most glasses, thus providing large Verdet constants. Only cubic or uniaxial crystals are usable for Faraday rotation. The diamagnetic and paramagnetic Verdet constants of several crystals are listed in Table 4.

The magneto-optic effect in ferromagnetic, ferrimagnetic, and antiferromagnetic materials can be much larger than in diamagnetic and paramagnetic crystals because of spin ordering (Deeter *et al.*, 1995). Rare-earth transition-metal alloys (e.g., $Y_3Fe_5O_{12}$ with trivalent rare earths or bismuth substituted for Y) are available commercially for applications at wavelengths ≥ 1 μm. Dilute magnetic semiconductors, such as $Cd_{1-x}Mn_xTe$, also exhibit large Faraday effects in the near-IR. The effect is strongly dependent on composition and wavelength.

4.5 Acousto-optic Crystals

The acousto-optic effect in a crystal can affect the direction, polarization, phase, frequency, and amplitude of an optical beam. Beam deflectors and intensity modulators are produced by generating an acoustic wave in the material with a piezoelectric driver in the frequency range 10^5–10^9 Hz. The sound wave may be oriented parallel (longitudinal wave) or normal (shear wave) to the light beam. By varying the applied radiofrequency, acousto-optic diffraction can also be utilized for tunable optical filters. Examples of acousto-optic crystals and their properties are given in Table 5.

4.6 Laser Crystals

Laser action has been demonstrated in hundreds of insulating crystals of many different symmetries and chemical compositions and at wavelengths extending from 0.17 to 7 μm (Kaminskii, 1990; Weber, 1991a). The lasing ions have been predominantly lanthanide and transition-metal ions (see LASERS, SOLID STATE). All trivalent rare earths and several divalent rare earths (Eu, Dy, and Tm) have been reported to lase in crystals. Among the most prominent host crystals for trivalent rare earths are $Y_3Al_5O_{12}$, $YAlO_3$, YVO_4, and $LiYF_4$; for divalent rare earths and transition metals, MgF_2 and CaF_2; and for iron-group elements (Ti^{3+}, V^{2+}, Cr^{3+}, Cr^{4+}, Co^{2+}, Ni^{2+}), Al_2O_3, $MgSiO_4$, and $LiCaAlF_6$ or $LiSrAlF_6$. Because the emission lines in crystals are narrower than in glasses, lower thresholds for lasing and higher optical gains are possible for crystalline lasers.

Many color-center lasers use alkali halide crystals as host materials because they are optically isotropic and transparent in the visible to near-IR. The most commonly used crystals are LiF, NaCl, KCl, and RbCl. Color-center laser action has also been reported in diamond, sapphire, and alkaline earth crystals (Mollenauer, 1991).

Semiconductor and heterojunction lasers have used binary and ternary III–V and II–VI compounds and lead salts as the active medium (Kressel and Butler, 1987).

4.7 Scintillator Crystals

Scintillator crystals are widely used for the detection of ionizing radiation in a variety of applications including high energy and nuclear physics, astrophysics, geophysical exploration, medical imaging, security inspection, and industry (Knoll, 1989). The emission may arise from the introduction of activator ions [NaI(Tl), CsI(Tl), Lu_2SiO_5(Ce)] or from intrinsic ions in the materials ($Bi_4Ge_3O_{12}$ and CeF_3), charge transfer transitions of molecular groups ($CdWO_4$ and $PbWO_4$), excitons (CsI), or core-valence transitions (BaF_2,CsF). No single scintillator crystal is optimum for all applications. Major considerations for the selection of scintillator materials include light yield and associated energy resolution, high density and stopping power to minimize the amount of material required, decay time and afterglow, emission wavelength to match photodetectors, stability with respect to environmental conditions, and radiation damage. For applications utilizing large volumes of scintillator materials, such as in detectors for high-energy physics or medical imaging tomography (where several thousand crystals and volumes of materials measured in cubic meters may be required), production and cost become major considerations. Dense crystals with good UV transmission such as PbF_2 are used for Cherenkov detectors.

5. OPTICAL LIQUIDS

Although many pure fluids may have useful optical properties, they may also have other undesirable properties. For a specific application a liquid tailored for one or more special properties is usually required. Commercial optical liquids consist of standard or custom formulations of one or more organic oils, solids, or polymers to give consistent specified properties (Sacher and Sacher, 1995). The selection of an optical liquid frequently entails more considerations than in selecting an optical solid. For example, in addition to transmittance, refractive index, and dispersion, there may be a need for stability, low toxicity, and system compatibility. Liquids that are homogeneous solutions can be considered to be an isotropic medium and their properties treated with the same concepts and formulas used to describe the properties of glasses.

5.1 Optical Properties

The most important properties for optical applications are usually the transparency of the liquid and its refractive index and dispersion. Liquids with high transparency are available from below 250 nm in the ultraviolet to greater than 2.5 μm in the infrared. The useful transmission in the infrared is dependent upon the absence of strong molecular absorption bands. The refractive indices of many liquids are in the range 1.4–1.6, but liquids exist covering the range from 1.3 to 2.1. Fluorocarbon formulations are used to achieve values in the range 1.3–1.4. The higher-index optical liquids ($n > 1.7$) may contain methylene iodide, arsenic tribromide, arsenic disulfide, sulfur, and selenium. Low–melting-point solids that must be warmed to become liquid extend the range to 2.3. The general ranges of index and dispersion of commercially available optical liquids are included in Fig. 5. A number of optical liquids, including aqueous solutions with strong hydrogen bonds, have abnormal dispersions much greater than those obtainable for optical glasses (Sigler, 1990).

Optical liquids have a high degree of internal transmission but are also subject to scattering losses, especially those having high molecular weights. Fluorescence is sometimes present, which may be detrimental for some optical applications. The high-index liquids ($n > 1.6$) exhibit a yellow to brown coloration that becomes more pronounced in the higher-index liquids. The refractive index of standard commercial optical liquids may vary with temperature on the order of several hundreds of parts per million per °C. The change is negative; therefore if significant absorption of light occurs in a liquid, it will be heated, and the refractive index will be lowered. Because of this large change, the index can be tuned by controlling the temperature. If heating is present, liquids may have to be stirred to achieve a homogeneous medium and avoid temperature gradients.

5.2 Physical Properties

The viscosity of optical liquids (measured in centistokes) covers a wide range from 1 for water, tens and hundreds for various oils, to >500 for glycerin, molasses, and honey. High viscosity is advantageous for holding two optical components together or to bridge a gap and reduces proneness to spillage; low-viscosity liquids, on the other hand, take less time to reach a uniform temperature, are easier to mix and pump, are less likely to entrain air bubbles, and are usually easier to clean off of substances. Density can be important if the suspension of particles or the formation of layers of optical liquids is required. Surface tension (cohesion) and adhesion are other considerations for some applications.

Boiling and flash points of optical liquids are usually chosen to be well above the intended working temperature, whereas the freezing point should be below the working temperature. Mixtures of liquids will not have a distinct freezing temperature, instead the liquid increases in viscosity with decreasing temperature until a transparent glasslike liquid is formed. The pour point is a measure of its useful liquid range, although it may still have satisfactory clarity and optical properties.

Whereas evaporation does not affect the index of pure materials, when the liquid is formed from mixtures of liquids with different refractive indices, the effective index may change with time as a consequence of evaporation. This effect can be minimized by using liquids with very low and/or matching vapor pressures. The specific heat may be a consideration if the liquid also serves as a circulating coolant. Thermal expansion, thermal conductivity, sound velocity, compressibility, etc. are generally of less importance. Optical liquids composed of organic oils are good insulators.

5.3 Chemical Properties

The mutual solubility of optical liquids allows one to blend them for desired optical and physical properties, such as refractive index, dispersion, and viscosity. Aqueous-based optical liquids can be removed by water and, sometimes, by ethanol or acetone. Organic-based optical liquids require specific solvents, ethyl ether, acetone, xylene, and naphtha being the most commonly used.

As noted in Sec. 5.1, many of the higher-index optical liquids ($n > 1.7$) contain substances that are toxic and corrosive; therefore suitable protective clothing and handling devices may be mandatory when using them. Odoriferous optical liquids may require ventilation. Polychlorinated biphenyl materials and chlorofluorocarbons have generally been eliminated from the formulation of optical liquids. Optical liquids may be affected by what they come in contact with, e.g., plastics, rubbers, and metals. Information about the compatibility of specific optical liquids with common materials, their toxicity, and other hazards associated with their use are usually available from the supplier.

Some optical liquids deteriorate with time because of their inherent chemical instability; inhibitors can be used to reduce these effects. Absorption of light may also cause discoloration. Extremes of temperature are to be avoided: high temperature can lead to evaporation, boiling, or decomposition; low temperatures can lead to high viscosity, crystallization, or clouding,

5.4 Applications

Optical liquids have a broad range of applications in diverse fields. One of the main uses of optical liquids is the determination of the refractive index of transparent or translucent solids by immersing them in optical liquids of different known refractive indices until the solid is observed to disappear in the liquid that matches its refractive index. This is usually done on a microscopic scale under a microscope. Using immersion liquids, surface features may be made to disappear so that interior features such as bubbles, flaws, and striae can be inspected. Stress and strain effects are examined using polariscopic immersion techniques.

Another major use for optical liquids is to reduce interface reflective losses between two optical components. This is achieved by replacing the air between the components with an optical liquid formulated with a refractive index matching one of the two components or having a value between the two. Optical lenses can be formed by filling hollow lenses with a suitable index-of-refraction liquid. Because of their abnormal dispersion properties, liquids have been used in combination with glasses for color correcting optical elements.

Optical liquids or gels (aliphatic hydrocarbons and gelling agents) are used for coupling optical fibers to reduce or eliminate internal reflections and for mode stripping, i.e., as a coating to remove signal propagating within the fiber cladding. They are also used in liquid-core fibers.

Solutions of hundreds of aromatic compounds and organic dyes in various solvents have been used for tunable lasers in the range 0.3–1.0 μm. Aprotic liquids have also been used as hosts for rare-earth laser ions, but the general corrosiveness and toxicity of these materials have limited their usefulness.

Other uses of optical liquids include microscopic studies of solids by dispersion straining, calibration of refractometers and other optical instruments, and as a mounting medium. As noted earlier, liquids can exhibit optical activity with specific rotations of >10 deg/mm. Faraday rotation is also observed in liquids; representative values of Verdet constants are included in Table 4. Many liquids have large Kerr constants.

5.5 Liquid Crystals

Liquid crystals are materials that flow like a liquid but exhibit anisotropic properties characteristic of a crystal (DeGennes, 1974). Most liquid crystals are organic substances whose general molecular structure is represented by elongated rod- or cigar-shaped groups of atoms tens of angstroms in length and a few angstroms wide. Compounds that exhibit liquid crystalline properties include aromatic esters, azoxy compounds, biphenyls, cyclohexylcyclohexanes, phenylcyclohexanes, and Schiff's bases. The ability to tailor their optical properties makes liquid crystals attractive for applications from the ultraviolet to the infrared and involving both small and large apertures.

Three types of liquid crystals are

1. smectic, where molecules are arranged in layers and the materials have high viscosity,
2. nematic, which are less highly ordered and where molecules are spontaneously aligned with their long axes parallel, and
3. cholesteric, which have long chains of carbon and hydrogen atoms with protruding side groups of atoms that destroy their rodlike appearance.

Cholesteric liquid crystals have a layered structure with molecules within the layer ordered as in a nematic. The alignment of molecules in adjacent layers traces out a helical path through the material. Cholesterics are often referred to as chiral nematics. The optical axis of a cholesteric is the helix axis. Their viscosity is low and the optical anisotropies are large. Schiff-base chiral smectic materials exhibit ferroelectric behavior.

The polarization-altering properties of liquid crystals are used extensively in nonemissive optical displays and medical thermography. Nematic liquids have an optical birefringence of 0.04–0.40 and when aligned in the homogeneous orientation can serve as a wave plate. Cholesterics have negative birefringence. The ability to adjust the cholesteric pitch by varying the temperature or composition provides selective reflectivity for polarizer, filter, mirror, and optical isolator applications. The electro-optic effect provides control of birefringence and dynamic scattering in laser applications. Nonlinear optical field-induced effects in nematic and cholesteric compounds can be large.

GLOSSARY

Acousto-optic Effect: The modification of the refractive index of a material resulting from periodic strain field (acoustic wave). It can be used to vary the direction, polarization, phase, frequency, and amplitude of an optical beam.

Activator Ion: A constituent or dopant in a material that has energy levels appropriate for luminescence, stimulated emission, scintillation, magnetic-optic, or other optical effects.

Anisotropic: Description of a material exhibiting different optical properties along different axes of propagation or for different polarizations of a light wave.

Birefringence: A characteristic of anisotropic crystals in which two light waves having the same propagation vector but different linear or circular polarizations experience different indices of refraction.

Damage Threshold: The lowest optical power or energy density at which physical damage to the surface or bulk of a material occurs.

Dichroism: For isotropic materials, the selective reflection or transmission of light as a function of wavelength independent of the direction of vibration; for anisotropic materials, the selective absorption of light vibrating in a particular plane relative to the crystalline axes.

Dispersion: The variation of the index of refraction of a material as a function of the wavelength of electromagnetic radiation.

Elasto-optic Effect: The modification of the optical properties of a solid when subject to mechanical strain.

Electro-optic Effect: The modification of the optical properties of a solid induced by an external electric field (see **Pockels Effect** and **Kerr Effect**).

Index of Refraction: A complex quantity: the real part defines the speed of light in a medium; the imaginary part is the extinction coefficient.

Isotropic: Description of a material exhibiting the same optical properties in any direction.

Kerr Effect: Quadratic electro-optic effect; it is observed in all solids.

Magneto-optic Effect: Rotation of the polarization direction of light waves due to an applied magnetic field (e.g., Faraday effect).

Optical Activity: An effect in materials in which right- and left-handed circularly polarized light experience different indices of refraction.

Photochromism: The reversible change in the absorption spectrum of a material when irradiated with a given wavelength of light.

Pockels Effect: Linear electro-optic effect; observed in noncentrosymmetric crystals.

Polarization: The restriction of the vibrations of light wave to a single plane. The polarization vector is always in a plane at right angles to the beam direction. The polarization direction can be random (unpolarized), constant (plane polarized), or due to two coherent polarization components at right angles (elliptically polarized).

Striae: Transparent imperfections in optical glass, often in the form of approximately straight, parallel streaks, having a slightly different refractive index than that of the remaining glass.

Thermoplastic: Polymeric material that can be remolded again and again, unlike thermoset plastics that cannot be remolded.

Works Cited

Adair, R., Chase, L. L., Payne, S. A. (1987), *J. Opt. Soc. Am.* **B4,** 875–881.

Araujo, R. J. (1977), in: M. Tomozawa, R. H.

Doremus (Eds.), *Treatise on Materials Science and Technology,* Vol. 12, New York: Academic, pp. 91–122.

Bagley, B. G., Vogel, E. M., French, W. R., Pasteur, G. A., Gan, J. N., Tauc, J. (1976), *J. Non-Cryst. Solids* **22,** 423–436.

Baldwin, C. M., Almeida, R. M., Mackenzie, J. D. (1981), *J. Non-Cryst. Solids* **43,** 309–344.

Blasse, G., Grabmeier, B. C. (1994), *Luminescent Materials,* Berlin: Springer-Verlag.

Bliss, E. S., Speck, D. R., Holzrichter, J. F., Erkkila, J. H., Glass, A. J. (1974), *Appl. Phys. Lett.* **25,** 448–490.

Bloembergen, N. (1965), *Nonlinear Optics,* New York: W. A. Benjamin.

Born, M., Wolf, E. (1980), *Principles of Optics,* 6th ed., Oxford: Pergamon Press.

Borrelli, N., Hall, D. W. (1991), in: N. J. Kreidl, D. R. Uhlmann (Eds.), *Optical Properties of Glass,* Cincinnati: American Ceramic Society, p. 87.

Buckingham, A. D., Stephens, P. J. (1966), *Ann. Rev. Phys. Chem.* **17,** 399–432.

Chemla, D. S., Zyss, J. (Eds.) (1987), *Nonlinear Optical Properties of Organic Molecules and Crystals,* Vols. 1 and 2, New York: Academic.

Cook, L. M., Mader, K. H. (1980), *J. Am. Ceram. Soc.* **65,** 597–601.

Cook, L. M., Stokowski, S. E. (1986), in: M. J. Weber (Ed.), *Handbook of Laser Science and Technology,* Vol. IV, Boca Raton, FL: CRC, pp. 93–183.

Deeter, M. N., Day, G. W., Rose, A. H. (1995), in: M. J. Weber (Ed.), *Handbook of Laser Science and Technology, Supplement 2: Optical Materials,* Boca Raton, FL: CRC, pp. 367–402.

DeGennes, P., Prost, J. (1993), *The Physics of Liquid Crystals,* New York: Oxford: Univ. Press.

Dumbaugh, W. H. (1985), *Opt. Eng.* **24,** 257–262.

Guenther, R. (1990), *Modern Optics,* New York: Wiley.

Hall, D. W., Weber, M. J. (1991), in: M. J. Weber (Ed.), *Handbook of Laser Science and Technology, Supplement 1: Lasers,* Boca Raton, FL: CRC, pp. 137–158.

Hellwarth, R. W. (1977), *Prog. Quant. Electr.* **5,** 1–68.

Izumitani, T. S. (1986), *Optical Glass,* New York: American Institute of Physics.

Kaminskii, A. A. (1990), *Laser Crystals,* 2nd ed., Berlin: Springer-Verlag.

Knoll, G. F. (1989), *Radiation Detection and Measurement,* 2nd ed., New York: Wiley, pp. 215–250.

Kreidl, N. J. (1983), in: D. R. Uhlmann, N. J. Kreidl (Eds.), *Glass: Science and Technology,* Vol. 1, New York: Academic, pp. 105–299.

Kressel, H., Butler, J. K. (1987), *Semiconductor Lasers and Heterojunction LEDs,* New York: Academic.

Marion, J. E., Weber, M. J. (1991), *J. Eur. Sol. State Inorg. Chem.* **28,** 271–287.

Mollenauer, L. F. (1991), in: M. J. Weber, (Ed.), *Handbook of Laser Science and Technology, Supplement 1: Lasers,* Boca Raton, FL: CRC, pp. 1101–125.

Mollenauer, L. F., Stolen, R. H., Gordon, J. P. (1980), *Phys. Rev. Lett.* **45,** 1095–1098.

Moore, D. T. (1980), *Appl. Opt.* **19,** 1035–1038.

Musikant, S. (1985), *Optical Materials,* New York: Marcel Dekker, pp. 128–147.

Nye, J. F. (1976), *Physical Properties of Crystals,* Oxford: Clarendon.

Sacher, R., Sacher, W. (1995), in: M. J. Weber (Ed.), *Handbook of Laser Science and Technology, Supplement 2: Optical Materials,* Boca Raton, FL: CRC, pp. 97–112.

Schott Catalog (1982), Mainz: Schott.

Schroeder, J. (1977), in: M. Tomozawa, R. H. Doremus (Eds.), *Treatise on Materials Science and Technology,* Vol. 12, New York: Academic, pp. 157–222.

Shen, Y. R. (1984), *The Principles of Nonlinear Optics,* New York: Wiley.

Siegman, A. E. (1994), *Lasers,* Palo Alto: University Science Books.

Sigel, Jr., G. H. (1977), in: M. Tomozawa, R. H. Doremus (Eds.), *Treatise on Materials Science and Technology,* Vol. 12, New York: Academic, pp. 5–89.

Sigler, R. (1990), *Appl. Opt.* **29,** 2451–2459.

Thompkin, W. R., Boyd, R. W., Hall, D. W., Tick, P. A. (1987), *J. Opt. Sci. Am.* **B4,** 1030–1034.

Turner, W. H. (1973), *Appl. Opt.* **12,** 480–486.

Van de Hulst, H. C. (1957), *Light Scattering by Small Particles,* New York: Wiley.

Vogel, E. M., Weber, M. J., Krol, D. M. (1991), *Phys. Chem. Glass* **32,** 231–254.

Weber, M. J. (1991a), *Handbook of Laser Science and Technology, Supplement 1: Lasers,* Boca Raton, FL: CRC.

Weber, M. J. (1991b), in: J. Zarzycki (Ed.), *Glasses and Amorphous Materials,* Materials Science and Technology Vol. 9, Weinheim: VCH, pp. 619–664.

Weber, M. J., Milan, D., Smith, W. L. (1978), *Opt. Eng.* **17,** 463–469.

Wedham, B. (1979), in: R. R. Shannon, J. C. Wyant (Eds.), *Applied Optics and Optical Engineering,* Vol. VII, San Diego: Academic, pp. 79–96.

Wood, R. M. (1986), *Laser Damage in Optical Materials,* Boston, Hilger.

Yariv, A. (1989), *Quantum Electronics,* New York: Wiley.

Yariv, A., Yeh, P. (1984), *Optical Waves in Crystals,* New York: Wiley.

Yen, W. M. (1995), in: F. Agullo-Lopez (Ed.), *Materials for Optoelectronics,* Singapore: World Scientific.

Zallen, R. (1983), *The Physics of Amorphous Solids,* New York: Wiley.

Zeiger, H. J., Pratt, G. W. (1973), *Magnetic Interaction in Solids,* Oxford: Clarendon Press.

Further Reading

Bamford, C. R. (1977), *Color Generation and Control in Glass,* Amsterdam: Elsevier.

Bass, M. (Ed.) (1995), *Handbook of Optics,* New York: McGraw-Hill.

Boyd, R. W. (1992), *Nonlinear Optics,* San Diego: Academic.

Chemla, D. S., Zyss, J. (Eds.), *Nonlinear Optical Properties of Organic Molecules and Crystals,* Vols. 1 and 2, New York: Academic.

Efimov, A. M. (1995), *Optical Constants of Inorganic Glasses,* Boca Raton, FL: CRC.

Fanderlik, I. (1983), *Optical Properties of Glass,* Glass Science and Technology, Vol. 5, Amsterdam: Elsevier.

Izumitani, T. S. (1986), *Optical Glass,* New York: American Institute of Physics.

Jacobs, S. D. (1992), *Liquid Crystals for Optics,* Vol. MS46, Bellingham, WA: SPIE.

Kaminskii, A. A. (1990), *Laser Crystals,* 2nd ed., Berlin: Springer-Verlag.

Klocek, P. (Ed.) (1991), *Handbook of Infrared Optical Materials,* New York: Marcel Dekker.

Kokorina, V. F. (1996), *Glasses for Infrared Optics,* Boca Raton, FL: CRC.

Kriedl, N., Uhlmann, D. (Eds.) (1991), *Optical Properties of Glass,* Cincinnati: American Ceramic Society.

Morey, G. W. (1954), *Properties of Glass,* 2nd ed., New York: Reinhold.

Musikant, S. (1985), *Optical Materials,* New York: Marcel Dekker.

Nye, J. F. (1976), *Physical Properties of Crystals,* Oxford: Clarendon.

Palik, E. D. (Ed.) (1985), *Handbook of Optical Constants of Solids,* New York: Academic; (1991), *Handbook of Optical Constants of Solids,* II, New York: Academic.

Weber, M. J. (Ed.) (1986), *Handbook of Laser Science and Technology,* Vols. 3, 4, and 5, *Optical Materials,* Boca Raton, FL: CRC.

Weber, M. J. (Ed.) (1995), *Handbook of Laser Science and Technology, Supplement 2: Optical Materials,* Boca Raton, FL: CRC.

Wooten, F. (1972), *Optical Properties of Solids,* New York: Academic.

Yariv, A., Yeh, P. (1984), *Optical Waves in Crystals,* New York: Wiley.

Yeh, P. (1993), *Introduction to Photorefractive Optics,* New York: Wiley-Interscience.

Zallen, R. (1983), *The Physics of Amorphous Solids,* New York: Wiley.

TRANSPORT, ELECTRONIC

See ELECTRONIC TRANSPORT IN SOLIDS; TRANSPORT PROPERTIES

TRANSPORT BY MAGNETIC LEVITATION

G. T. Danby and J. R. Powell, *Brookhaven National Laboratory, Upton, New York, U.S.A.*

Introduction—Maglev and Its History 233
1. **Basic Concepts for Magnetic Levitation** 235
1.1 Ferromagnetic Attraction 235
1.2 Permanent-Magnet Repulsion 236
1.3 Inductive Repulsion Using ac Currents 236
1.4 Superconducting Inductive Repulsion 237
2. **Maglev Systems Concepts and Performance** 237
2.1 Superconducting Electromagnetic Suspensions 237
2.2 Conducting-Sheet Suspension 239
2.3 Simple Loop Suspension 240
2.4 Null-Flux Suspension 241
2.5 Stability of Magnetic Suspensions 242
2.6 Magnetic and Aerodynamic Drag 243
2.7 Magnetic Propulsion 245
2.8 Electromagnetic Suspensions 247
3. **Maglev Development and Prototype Demonstration** 247
3.1 Japanese Superconducting Maglev 247
3.2 Japanese Electromagnetic Maglev 249
3.3 German Electromagnetic and Superconducting Maglev 249
3.4 U.S. Maglev Studies 250
4. **Future Application of Maglev** 253
4.1 Energy, Environmental, Economic, and Societal Benefits 253
4.2 Engineering Developments Needed for Large-Scale Implementation of Maglev 256
4.3 Environmental and Safety Aspects 258
4.4 Broader Applications of Maglev 259
Glossary 259
Works Cited 260
Further Reading 261

INTRODUCTION—MAGLEV AND ITS HISTORY

Maglev (magnetic levitation) is a new form of high-speed ground transport that offers major improvements in speed, mobility, and efficiency, along with substantial reductions in cost and effect on the environment. Unlike existing ground transport, where wheeled vehicles travel along roads or rails, in Maglev the vehicles are magnetically levitated and propelled at high speed along a dedicated guideway. Because the Maglev vehicles do not contact the guideway, the drag and wear that accompany mechanical friction are eliminated.

This article reviews the historical development of Maglev, starting with the concepts of the early 1900s and continuing through the present prototype development programs in Japan and Germany. The four basic concepts for Maglev suspensions are described—an attractive-force suspension, based on the attraction of electromagnets on a vehicle to an iron rail guideway, and three different repulsive-force suspensions, based on the magnetic repulsion between

1. permanent magnets on the vehicle and opposing magnets on the guideway,
2. ac current loops on the vehicle and the induced opposing currents in a normal aluminum sheet guideway, and
3. superconducting dc magnets on a moving vehicle and the induced opposing cur-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

rents in a normal aluminum loop of sheet guideway.

Development programs have demonstrated technical feasibility of Maglev. However, engineering improvements, particularly with regard to capital cost, are required before Maglev can be implemented on a larger scale. New approaches for achieving these engineering improvements are discussed.

It has been known for more than a century that the very strong forces exerted by electromagnetism could be used for magnetic levitation. Goddard, the renowned rocket pioneer, and others discussed magnetic levitation of vehicles. Before the First World War, Bachelet (1912) proposed levitation of a vehicle utilizing eddy-current repulsion between ac coils and a conducting metal plate, and Graemiger (1911) proposed levitation by the attractive force between electromagnets on a vehicle and an overhead steel rail. These ideas were not practical at the time. The ac method was very lossy and power intensive. The electromagnet–iron-rail suspension was unstable, and feedback control systems were not then adequate for a practical transport system. Magnetically levitated transport, now popularly termed Maglev, lay dormant for half a century.

The first practical concepts for Maglev were proposed in the 1960s. In 1966, Powell and Danby (1967) proposed placing dc superconducting magnets on a vehicle to interact with passive short-circuited normal-metal loops in the guideway. The magnets on the moving vehicle automatically induced currents by transformer action in the guideway loops underneath the vehicle as it passed by. The resultant strong lifting and restoring forces—both vertical and horizontal—generated by the magnetic interaction between the magnets on the vehicle and the induced currents in the guideway loops enabled inherently stable suspensions with large clearances, typically in the range of 15–20 cm.

Subsequently, Powell and Danby (1970, 1971) proposed the "null-flux" suspension geometries. These can magnetically levitate very heavy loads while maintaining very low I^2R resistive losses in the guideway. They also invented the Linear Synchronous Motor (LSM), which uses an actively powered ac winding in the guideway to propel the vehicle with high motor efficiency (Powell and Danby, 1969). The velocity of the vehicle is precisely controlled by the frequency of the ac current wave that interacts with magnets on the vehicle. The combination of stable, phase-locked propulsion combined with inherent and passively stable levitation forces enables very reliable and safe high-speed transport.

Danby *et al.* (1974) also proposed and analyzed systems in which the superconducting vehicle magnets were attracted to iron located in the guideway. By combining the iron lift forces with null-flux stabilization loops, which provide overall inherent strong stability, extremely heavy vehicles can be levitated with very low power consumption.

Powell and Danby's work stimulated a broad-based field of Maglev studies and research projects. In 1967 and 1968, following discussions with Powell and Danby, Stanford Research Institute (SRI) and Rockwell International studied magnetic levitation of very high-speed rocket sleds in collaboration with Sandia Corporation (Williams, 1969; Guderjahn and Wipf, 1969). As an alternative to loop guideways SRI proposed a continuous, shaped aluminum steel guideway, which they felt would be lower in capital cost. The higher drag power necessitated by the sheet guideway was deemed acceptable.

In 1969, Coffey *et al.* (1969, 1973) described a sheet guideway for vehicle levitation, and in 1970 they built a 500-ft-long test guideway. Reitz and Davis (1970) also studied a continuous sheet guideway in 1969. In 1970, Kolm and Thornton (1972, 1973) designed and subsequently constructed a 1/25 scale model of a system called the Magneplane. This was based on a conducting-sheet trough-shaped guideway. It differed from other sheet concepts in that vehicles could roll in a coordinated bank in the trough guideway, with the capability of operating at different bank angles if vehicle speed were altered. The model was propelled by a Linear Synchronous Motor.

In 1975 the Department of Transportation decided not to pursue the development of high-speed ground transport systems for the United States. The elimination of federal funding also discouraged private investment. This policy continued until 1990. Maglev programs continued in other countries, particularly Japan and Germany. They constructed full-scale demonstration systems that are discussed later (Miyata *et al.*, 1993; Wiescholek *et al.*, 1993).

Interest in Maglev has recently revived in the United States where a huge potential market exists. Maglev could greatly improve mobility, increase energy efficiency, and reduce pollution and reliance on oil imports. This revival was sparked by the request in 1988 by U.S. Senator Daniel Patrick Moynihan for a Maglev Task Force study. The findings of the Task Force (Powell and Danby, 1989), which was organized and chaired by the authors, laid out the benefits to the United States of a large-scale Maglev transport system. Its recommendations for a U.S. Maglev development program were incorporated into the 1990 Intermodal Surface Transport Energy Efficiency Act (ISTEA).

1. BASIC CONCEPTS FOR MAGNETIC LEVITATION

1.1 Ferromagnetic Attraction

The strong force of attraction of ferromagnetic material to a source of magnetic induction is a familiar phenomenon. At its saturation induction of 2.1 T an area of only 48 cm^2 (7 in.^{2}) on an iron pole face exerts a force of 1 t (tonne). Even at 1.0 T, to avoid saturation due to edge effects, etc., only 190 cm^2 (30 in.^{2}) is needed to suspend 1 t. These large forces can readily support heavy vehicles.

The force of attraction is very unstable, however. If the source of induction is an electromagnet of constant ampere turns, the magnetic field in the air gap between two iron pole faces is inversely proportional to the thickness of the air gap, assuming that the iron is not saturated and that edge effects can be ignored. The magnetic force is then proportional to the square of the field strength. For example, with an air gap of 0.5 cm, the attractive force between the iron pole faces is four times as great as for an air gap of 1.0 cm.

This instability is familiar from the attraction of a small permanent magnet to a steel refrigerator door. The force strongly increases as the distance decreases. Even a weak permanent magnet generates a force stronger than gravity when it is close to the surface of the door; however, this force diminishes rapidly with distance from the surface.

In order for ferromagnetic attractive forces to be useful for Maglev, the force attracting the vehicle electromagnets to the iron rails above them must be servo-controlled to prevent a crash. With a sufficiently fast-acting and sensitive control system, the vehicle can be levitated with a quasiconstant air-gap thickness between the vehicle magnets and the iron rails. The vehicle can oscillate about its equilibrium position because of inevitable imperfections in guideway position, wind forces, etc., but it will not strike the guideway as long as the control system functions adequately.

Figure 1 illustrates how electronic feedback can be used to levitate objects stably, using the inherently unstable magnetic force of attraction. The magnitude and rate of change of the thickness of the air gap (in both vertical and horizontal directions, though only the vertical is shown in Fig. 1) are continuously sensed. A fast electronic control system (millisecond response time) rapidly and appropriately changes the magnet ampere turns, altering the attractive force between the electromagnets on the vehicle and the iron rails in the guideway above, so as to maintain the air gap at the design value.

This technique has long been used for magnetically suspended high-speed centrifuges (Polgren, 1966). The same approach is used in the German Transrapid Maglev System (Wiescholek *et al.*, 1993). Laminated iron rails are located on each side of the guideway, with sets of vehicle electromagnets positioned under the rails. The upward

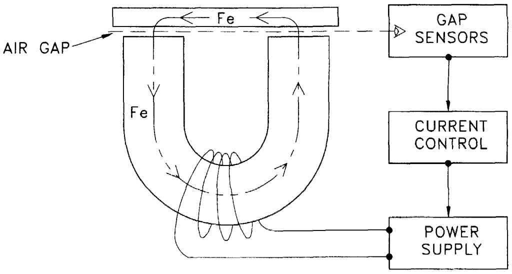

FIG. 1. Magnetic suspension based on servo control of the attractive force between an electromagnet and a ferromagnetic sheet.

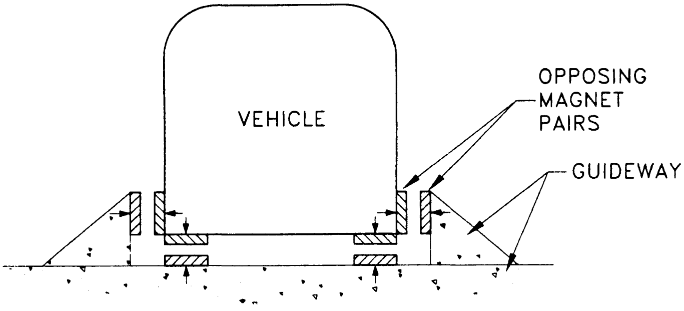

FIG. 2. Magnetic suspension based on the repulsive force between permanent magnets on a vehicle and guideway.

magnetic force then suspends the vehicle. Other suitably positioned horizontally oriented magnets are electronically controlled to provide lateral stability.

In order that electromagnetic Maglev vehicles operate safely, the sensing and electronic control systems must be extremely reliable and accurate. The iron magnets, servoed reactive power supplies, and backup power sources substantially increase vehicle weight. The maximum practical air gap is approximately 1 cm; larger gaps result in excessive I^2R power requirements for the vehicle magnets.

1.2 Permanent-Magnet Repulsion

Levitation by mutual repulsion between permanent magnets with opposed magnetic fields is a familiar phenomenon. Pairs of repelling magnets oriented for repulsion in both the vertical and horizontal directions can produce stable levitation, and lightweight toy vehicles using weak ferrite magnets have been marketed based on this approach. The suspension (Fig. 2) is inherently stable because the force of repulsion increases as the gap becomes smaller.

To date, Maglev systems based on this approach have not appeared practical because of the large cost of guideways constructed with permanent magnets. Forces 10 times greater than those obtained with ferrite magnets are possible with neodynium-iron-boron or samarium-cobalt magnets. Unfortunately, their cost is almost 100 times greater than for ferrite. The weight of the magnets required on the vehicle is also high.

In principle, electrodynamic repulsion levitation is possible using vehicle-mounted permanent magnets to interact with passive normal-metal guideway loops or a conducting-sheet guideway. However, the heavy weight, low field strength, and high cost of the necessary permanent magnets do not result in a practical transport system.

1.3 Inductive Repulsion Using ac Currents

Figure 3 illustrates the levitation of an ac current coil by a repulsive force generated by the magnetic interaction between it and eddy currents that are induced in a conducting sheet underneath. Assuming 60-Hz ac in the current loop, an aluminum sheet of 2.5 cm (1 in.) in thickness behaves diamagnetically. The resulting force on the ac current loop equals the force that would be produced between the real ac current loop and an image loop of opposite polarity located at the distance *d* below the sheet. The vehicle loop will be suspended at the height *d* where the repulsive force equals its weight. This suspension is inherently stable vertically, since the vertical force increases as *d* decreases. Horizontal stability can be provided by bending the edges of the sheet upward to make a

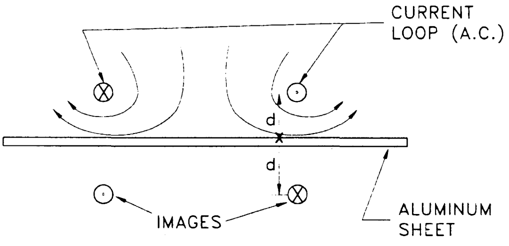

FIG. 3. Magnetic suspension based on the repulsive force between a coil carrying alternating current (ac) and a conducting sheet.

trough shape. The geometry in Fig. 3 could be inverted if desired; that is, the aluminum sheet could be located on the bottom of the vehicle, and the ac current loops, on the guideway.

The ac eddy-current repulsion concept was proposed by Bachelet (1912). Although feasible in principle, in practice to suspend useful loads the concept requires impractically large powers, both reactive and resistive. A Maglev vehicle weighing 40 metric tons, for example, would have resistive power requirements (counting both the losses in the ac coils on the vehicle and in the sheet guideway) on the order of 10 MW.

1.4 Superconducting Inductive Repulsion

Powell and Danby (1967) proposed that practical, large-clearance inductive repulsion suspensions could be provided by strong, light-weight superconducting dc magnets mounted on a moving Maglev vehicle. As the vehicle moves along a guideway, its superconducting magnets induce opposing currents in the normal-metal (e.g., aluminum) loops or sheets located beneath. The repulsive magnetic forces between the superconducting magnets and the induced currents in the guideway levitate the vehicle. The suspension is inherently stable in the vertical direction, since the repulsive forces monotonically increase as the distance between the vehicle magnets and the loops or sheets on the guideway decreases. The suspension can be designed to be inherently and strongly stable in the other directions—lateral, pitch, yaw, and roll—while continuing to move freely along the guideway.

Since the superconducting dc magnets alternate in magnetic polarity along the vehicle, the guideway beneath experiences the electromagnetic equivalent of the ac current loop described in the previous section. The equivalent ac frequency is given by dividing the vehicle speed by the magnet pole pitch (the pole pitch is twice the length of a magnet of a given polarity). For example, at 500 km/h (300 mph) and a pole pitch of 2.2 m, the effective ac frequency is 60 Hz. It is desirable that the effective ac frequency be relatively high, so as to minimize the L/R decay of induced currents in the guideway, and to allow the vehicle to "lift off" the guideway at low speed. This can be achieved by using short magnets and alternating their polarity.

Unlike the three previous suspension concepts, which can levitate vehicles at zero speed, vehicles with superconducting dc magnets must move along the guideway to levitate. Using magnet pole pitches on the order of 2 m, vehicles can lift off at only 30 km/h (20 mph). Auxiliary wheels or other devices would be used at lower speeds, primarily when entering or leaving a station.

The superconducting magnets can be positioned either along the full length of the Maglev vehicle, at the front and back ends only, or at intermediate locations along the vehicle, depending on design. In general, full-length placement results in a more efficient, structurally simpler vehicle. However, end or intermediate placement may be desirable to minimize the magnitude of the stray magnetic fields in the passenger compartment. In this connection, the use of short magnets of alternating magnetic polarity also helps to reduce stray magnetic fields.

2. MAGLEV SYSTEMS CONCEPTS AND PERFORMANCE

2.1 Superconducting Electrodynamic Suspensions

In superconducting electrodynamic Maglev, superconducting (SC) magnets on the moving vehicle induce currents in ambient-temperature metal loops (or sheets) in the guideway beneath. These induced currents levitate and stabilize the vehicle. Levitation and stabilization are automatic and inherent, requiring neither active control nor power from the vehicle or the guideway. The only requirement is that the vehicle be moving. Stability is sufficiently strong that occasional external forces—winds, curves, grades, etc.—cannot make the vehicle contact the guideway, so that safety is inherently assured under all conditions. External forces and guideway variations may cause a magnetically levitated vehicle to oscillate slightly about its equilibrium suspension point, however. Depending on amplitude, these oscillations could be objectionable from the standpoint of passenger ride comfort. If necessary, they can be countered and damped by active control systems (e.g., magnetic or aerodynamic), or by passive mechanical devices (e.g., shock absorbers), or by a combination of active and passive methods.

There are many ways to design for a superconducting Maglev system based on the fundamental idea of the inductive "magnetic spring." In this concept, as the gap between a SC magnet on the vehicle and the guideway loop or sheet beneath it *decreases*, the induced current in the guideway *increases*. The increased current, together with the smaller distance, results in stronger magnetic repulsive force. Conversely, as the gap *increases*, the magnetic repulsive force *decreases*. Consequently, there is a particular gap at which the upward magnetic repulsive force exactly counteracts the weight of the vehicle. This gap is termed the "equilibrium suspension height."

For displacements upward (gap increases) or downward (gap decreases) the magnetic suspension behaves as if there is a "magnetic spring" that pulls the vehicle back toward its equilibrium suspension point. The location of the equilibrium point is determined by the designer, according to the choice of geometry, SC magnet strength, etc. The "magnetic spring" can be stiff or soft, linear or nonlinear, damped or undamped, lossy or nonlossy, etc., depending on design.

Figure 4 illustrates in cross section three types of guideway configurations:

1. the simple loop (Powell and Danby, 1967),
2. the null-flux loop circuit (Powell and Danby, 1970), and
3. the conducting sheet (Coffey *et al.*, 1969).

The three guideway configuration options for superconducting Maglev—simple loops, null-flux loop circuits, and conducting sheets—are described in more detail below, and in the following sections their performance is analyzed. There are important and fundamental differences among the three options that affect performance. First, in the conducting-sheet guideway, the integrated induced current in the sheet must approximately equal the effective dc current in the loop above it, since the sheet diamagnetically excludes magnetic flux from the space below it. As a result of these large currents in the sheet and their consequent I^2R losses, there is a strong magnetic drag force on the vehicle. In the simple and null-flux loop guideways, on the other hand, the guideway current can be made a small fraction of that in the superconducting dc magnets. This greatly reduces the I^2R losses in the guideway and results in a much smaller magnetic drag force on the vehicle.

Second, for loop guideways the superconductive magnets on the vehicle have substantially higher ampere turns than those for sheet guideways; however, they are still far below any technical limit and, since dc superconducting magnets are lossless, do not require any additional electrical power.

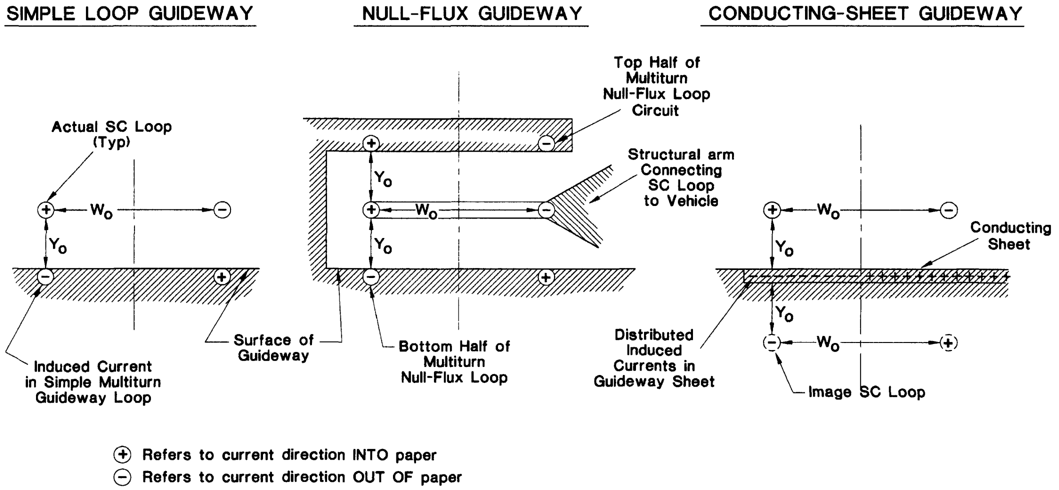
SIMPLE LOOP GUIDEWAY
NULL-FLUX GUIDEWAY
CONDUCTING-SHEET GUIDEWAY
Actual SC Loop (Typ)
W₀
Y₀
Induced Current in Simple Multiturn Guideway Loop
Surface of Guideway
Top Half of Multiturn Null-Flux Loop Circuit
Structural arm Connecting SC Loop to Vehicle
Bottom Half of Multiturn Null-Flux Loop
Distributed Induced Currents in Guideway Sheet
Conducting Sheet
Image SC Loop
⊕ Refers to current direction INTO paper
⊖ Refers to current direction OUT OF paper

FIG. 4. Guideway configurations that provide magnetic suspension by inductive interaction with superconducting (SC) magnets on a moving vehicle.

Third, in the conducting-sheet guideway, stability is relatively "soft"—that is, the restoring force per unit displacement from equilibrium is inherently low and can only be slightly adjusted in the design process. In loop guideways, by contrast, stability can be readily adjusted by design. Stability can be made "soft" or "stiff"—that is, the restoring force per unit displacement is either small or large—depending on the performance level desired. Finally, lateral stability in the conducting-sheet guideway is provided by having large induced currents flow all the time in the side wings of the sheet, even when the vehicle is centered on the guideway. These substantially increase the magnetic drag on the vehicle. In the null-flux guideway, in contrast, current does not flow in the lateral stability loops until the vehicle departs from its equilibrium position. The magnitude of this stabilizing current is proportional to the displacement from equilibrium—low for small displacement, greater for large displacement.

To simplify analysis of these guideway configurations, the SC loops on the vehicle are assumed to be long compared to the equilibrium suspension height, so that 3D effects caused by loop ends can be neglected. In addition, the velocity is assumed high enough that L/R decay of the induced currents is negligible. In practice, at high speed, the L/R decay effect is small. For a pole pitch of 3 m, for example, the transit time of a given polarity magnet past a point on the guideway is approximately 10 ms, for a vehicle speed of 500 km/h. For loop-type guideways, the L/R time constant for an optimized loop design is on the order of several hundred milliseconds. Under these conditions, the L/R decay has only a minor effect on the operating behavior of the suspension. The suspension clearance will decrease slightly—i.e., the change in clearance is only about 1% of the nominal clearance—and the I^2R losses will slightly increase, by a few percent above their value at infinite speed. The results based on the above simplifying assumptions may thus differ by a few percent from nonidealized analyses. However, the scaling relationships are valid and show the overall behavior of each configuration in relatively simple analytical form.

For each type of configuration, the moving SC loop on the vehicle induces oppositely directed currents in the ambient-temperature aluminum guideway loops or sheets. The magnitude of the induced guideway current, I_G, is some fraction f_G of the current in the SC loop, I_{SC}, such that

$$I_G = f_G I_{SC}. \tag{1}$$

The value of f_G depends on the suspension height, type of guideway, and specific design details, such as the geometric dimensions, loop inductance, etc. The magnetic repulsive force between the SC loop and the guideway for a suspension height y_0, per unit length of guideway, is then

$$F_m = \frac{k_G I_G I_{SC}}{y_0} = \frac{k_G f_G I_{SC}^2}{y_0} = \frac{K_G I_{SC}^2}{y_0}. \tag{2}$$

The values of the constants k_G and K_G depend on the type of suspension and design details. Note that K_G will in general increase as y_0 decreases because the magnitude of the guideway current will increase.

2.2 Conducting-Sheet Suspension

The solution for the sheet guideway is determined as follows. The magnetic force $(F_m)_{y_0}^S$ between the SC loop and the sheet guideway equals the force between the actual SC loop and an image loop at the distance y_0 below the sheet and is given by

$$\begin{aligned}(F_m)_{y_0}^S &= \left(\frac{\mu_0}{2\pi}\right)\frac{I_{SC}^2}{y_0}\left[1 - \frac{4y_0^2}{W_0^2 + 4y_0^2}\right]\\ &= (K_G)_{y_0}^S\left(\frac{I_{SC}^2}{y_0}\right) = mg,\end{aligned} \tag{3}$$

where W_0 is the width of the superconducting loop and mg is the suspended weight. For the case where $W_0 = 3y_0$, for example, the value of the constant $(K_G)_{y_0}^S$ is given by ($\mu_0 = 4\pi \times 10^{-7}$ H/m)

$$(K_G)_{y_0}^S \sim \frac{9}{13}\left(\frac{\mu_0}{2\pi}\right).$$

For a displacement Δy from y_0 (i.e., $y = y_0 + \Delta y$), there is a net restoring force that acts to restore the SC loop to its equilibrium suspension point. The magnitude of the restoring force, $(\Delta F_m)_y^S$, normalized to the suspended weight, is

$$\frac{(\Delta F_{\mathrm{m}})_{y}^{S}}{(F_{\mathrm{m}})_{y_0}^{S}} = \left(\frac{y_0}{y}\right)\left[\frac{(K_{\mathrm{G}})_{y}^{S}}{(K_{\mathrm{G}})_{y_0}^{S}}\right] - 1. \tag{4}$$

For an upward displacement ($\Delta y > 0$) the restoring force is downward; for a downward displacement ($\Delta y < 0$) the restoring force is upward. Figure 5 shows the variation of restoring force with displacement for the conducting-sheet suspension (chain curve). The restoring force is strongly nonlinear and relatively low for large upward displacement—that is, the "magnetic spring" is "soft" in this region.

An additional feature of the sheet suspension is that the design current in the superconducting loop is automatically determined by the choice of suspended weight and y_0, as evident from Eq. (3), and cannot be independently varied. (I_{SC} actually varies logarithmically with W_0, but the variation is small.)

2.3 Simple Loop Suspension

The magnetic lift force for the simple loop suspension (Fig. 4) is given by

$$\begin{aligned}(F_{\mathrm{m}})_{y_0}^{L} &= 2\left(\frac{\mu_0}{2\pi}\right)^2 \frac{I_{\mathrm{SC}}^2}{y_0(1+\lambda)L_{\mathrm{G}}^{\mathrm{L}}} \\ &\quad \times \left\{\left[1 - \frac{y_0^2}{W_0^2 + y_0^2}\right]\left[\ln\frac{(W_0^2 + y_0^2)}{y_0^2}\right]\right\} \\ &= (K_{\mathrm{G}})_{y_0}^{L} I_{\mathrm{SC}}^2 / y_0, \end{aligned} \tag{5}$$

where $L_{\mathrm{G}}^{\mathrm{L}}$ is the self-inductance of the guideway loop and λ is any additional (normalized) external inductance that is incorporated into the loop circuit. As with the sheet guideway, if the SC loop is displaced up or down from the equilibrium suspension height y_0, there is a magnetic restoring force acting to push it back to y_0, where the normalized restoring force, $(\Delta F_{\mathrm{m}})_{y}^{L}/(F_{\mathrm{m}})_{y_0}^{L}$, is given by

$$\frac{(\Delta F_{\mathrm{m}})_{y}^{L}}{(F_{\mathrm{m}})_{y_0}^{L}} = \left(\frac{y_0}{y}\right)\left(\frac{(K_{\mathrm{G}})_{y}^{L}}{(K_{\mathrm{G}})_{y_0}^{L}} - 1\right). \tag{6}$$

Figure 5 shows the variation of restoring force with displacement for the simple loop suspension (solid curve), for the case where $W_0 = 3y_0$. The restoring force for the loop suspension is more nearly linear with displacement, but the "magnetic spring" is somewhat softer, than that for the sheet suspension.

In contrast to the sheet suspension, in the simple loop suspension the current in the SC loop can be independently selected even when suspension height y_0 and weight mg are fixed, by adding external inductance to the guideway loop circuit. As the total inductance of the loop circuit is increased, the design value of I_{SC} has to be correspondingly increased in order to keep the lift force constant. As seen from Eq. (5), I_{SC} will scale as $(1 + \lambda)^{1/2}$. The restoring force curve is not affected by the addition of external inductance—that is, the "spring" stiffness remains

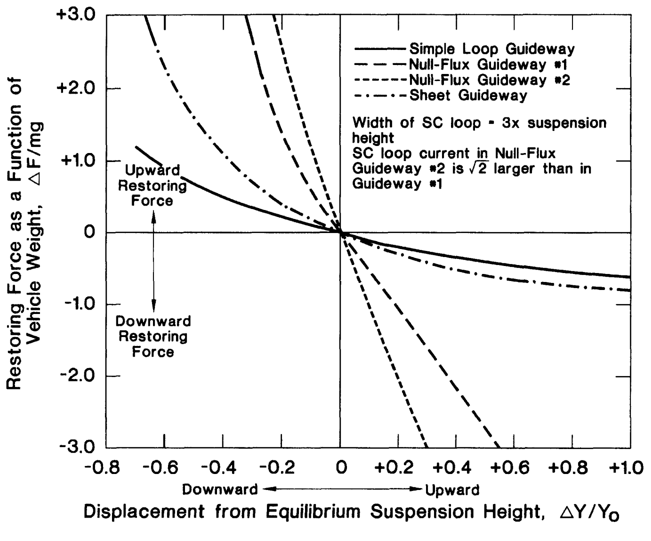

FIG. 5. Restoring forces for loop and sheet magnetic suspensions as a function of displacement from equilibrium.

constant—but the I^2R losses in the guideway loop decrease, since its current drops as the design value of I_{SC} increases. This results in an improved magnetic lift-to-drag ratio, as discussed later.

2.4 Null-Flux Suspension

The null-flux suspension behaves in a different manner than the sheet and simple loop suspensions. The restoring force is almost linear with displacement, and the magnitude of the "spring constant" is readily adjustable—that is, one can choose whether the suspension is stiff or soft.

There are many null-flux suspension geometries that can be designed. The configuration shown in Fig. 4 is illustrative of the principles involved and can be readily compared with the other two types of suspensions. The null-flux circuit shown here consists of two multiturn loops, one above the vehicle SC loop and the other below. The two loops are oppositely wound or cross connected, so that the *net* magnetic flux from the SC loop through the complete circuit is zero when the SC loop is centered between the upper and lower guideway loops. Since the net flux is zero, no current is then induced in the null-flux circuit as the SC loop moves past it.

When the SC loop is not at the center plane of symmetry between the upper and lower guideway loops, there is a net flux through the guideway loop circuit. An induced current then flows in the circuit, in such direction that the magnetic interaction between the guideway loops and the SC loop always acts so as to push the SC back toward the center plane of symmetry. If the SC loop were weightless, its equilibrium suspension location would be at the center plane of symmetry. However, since the SC loop has to support the vehicle, the SC loop only reaches equilibrium after it drops below the center to $y^* = y_0^*$, where y^* is the distance from the center plane of symmetry. The induced current in the guideway loop circuit then becomes large enough to support the vehicle. For $y^* < 0$, the magnetic force $(F_m)^{NF}_{y^*}$ on the SC loop is upward and results from the interaction of the two SC line currents with the four line currents in the guideway loop circuit:

$$(F_m)^{NF}_{y^*} = 2\left(\frac{\mu_0}{2\pi}\right) I_{SC}(i_G)_{y^*}\left[\frac{1}{y_0 + y^*} + \frac{1}{y_0 - y^*} - \frac{y_0 + y^*}{W_0^2 + (y_0 + y^*)^2} - \frac{y_0 - y^*}{W_0^2 + (y_0 - y^*)^2}\right] \tag{7}$$

with the induced current $(i_G)_y^*$ in the guideway loop circuit given by

$$(i_G)_{y^*} = \frac{(\phi_{net})_{y^*}}{L_G^{NF}} = \left(\frac{\mu_0}{2\pi}\right)\frac{I_{SC}}{L_G^{NF}}\left\{\ln\left[\frac{W_0^2 + (y_0 + y^*)^2}{(y_0 + y^*)^2}\right] - \ln\left[\frac{W_0^2 + (y_0 - y^*)^2}{(y_0 - y^*)^2}\right]\right\}, \tag{8}$$

where L_G^{NF} is the self-inductance of the guideway loop circuit, including mutual inductive coupling between the top and bottom loops, and $(\phi_{net})_y^*$ is the net flux through the guideway loop circuit.

The net force (magnetic plus gravity) is then

$$(F_{net})^{NF}_{y^*} = (F_m)^{NF}_{y^*} - mg, \tag{9}$$

where mg is the suspended weight. At $y^* = y_0^*$, $(F_m)^{NF}_{y^*} = mg$, and the net force equals zero. If the SC loop is above y_0^* ($y^* > y_0^*$), the net force acts downward; if the SC loop is below y_0^* ($y^* < y_0^*$), the net force acts upward.

The parameters y_0^* and I_{SC} are related to the suspended weight by Eqs. (7)–(9). Selecting a value for y_0^* determines I_{SC}, and vice versa. The stiffness of the null-flux suspension is thus independent of the equilibrium suspension height y_0, in contrast to the simple loop and sheet guideways, and can be selected to have whatever value is desired. Figure 5 illustrates how the restoring force varies with displacement $\Delta y/y_0$ for the null-flux suspension for two design variations. The first variation has a y_0^*/y_0 value of 0.2, while in the second, stiffer suspension, y_0^*/y_0 equals 0.1. The SC loop current in the second suspension is a factor of 2 larger than that in the first suspension.

Figure 6 shows examples of other null-flux geometries. The cross-coupled lateral loop geometry was used in the JR Maglev system at Miyazaki, and the vertical figure-

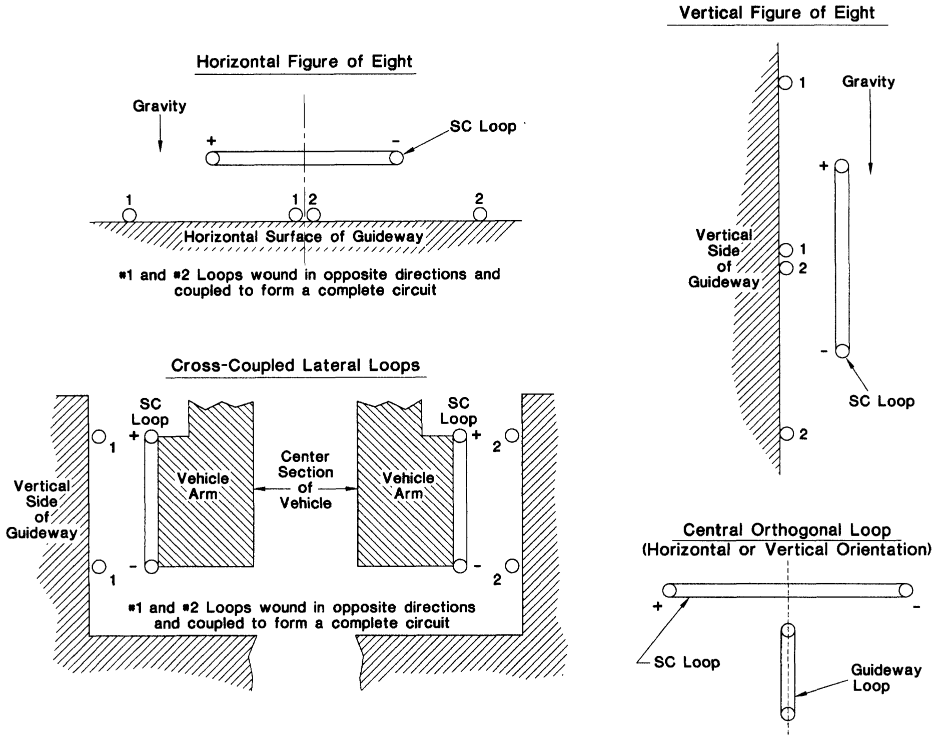
Horizontal Figure of Eight
Gravity
+
-
SC Loop
1
1
2
2
Horizontal Surface of Guideway
#1 and #2 Loops wound in opposite directions and coupled to form a complete circuit
Vertical Figure of Eight
1
Gravity
+
Vertical Side of Guideway
1
2
-
SC Loop
2
Cross-Coupled Lateral Loops
SC Loop
+
1
SC Loop
+
2
Vertical Side of Guideway
Vehicle Arm
Center Section of Vehicle
Vehicle Arm
-
1
-
2
#1 and #2 Loops wound in opposite directions and coupled to form a complete circuit
Central Orthogonal Loop (Horizontal or Vertical Orientation)
+
-
SC Loop
Guideway Loop

FIG. 6. Examples of null-flux guideway configurations used to magnetically levitate and/or stabilize vehicles.

of-eight geometry will be used in the new JR Maglev system under construction in Yamanishi Prefecture. Many other geometries are possible. The choice of geometry depends on many factors—the shape of the guideway, whether high-speed switching is used, fringing-field constraints, etc.

2.5 Stability of Magnetic Suspensions

Besides providing stable levitation, the guideway must also be configured so that vehicles are stable against displacement in the lateral, pitch, yaw, and roll directions. For the conducting-sheet guideway, roll and pitch stability is provided by having two lines of SC magnets on the vehicle, one at each side, and magnets at the front and rear ends. Lateral and yaw stability is provided by shaping the guideway sheet into a trough. The trough can be a U-shaped channel or part of a circle, as in the Magneplane design (Kolm and Thornton, 1972). For loop guideways, stability is provided by two guideway loop arrangements (one for stable levitation and the second for stable lateral guidance) and two lines of SC magnets, one at each side of the vehicle. The discrete nature of the sequential loops along the guideway automatically provides pitch, yaw, and roll stability. If the vehicle rolls to the left, for example, due to an external force, the left line of guideway loops will have a stronger upward force, and the right side a weaker, countering the roll. A nose-down pitch is countered by the guideway loops at the front of the vehicle having a stronger upward force than those at the back, countering the pitch. Similarly, yaw displacements are countered by the individual guideway loops acting to return the vehicle to its equilibrium position.

The vehicle is thus held in a five-dimensional "magnetic well." Any displacement in the vertical, lateral, pitch, yaw, or roll directions is immediately and strongly countered by a magnetic restoring force or moment. In the sixth dimension, the suspension is neutrally stable, and the vehicle is free to move along the guideway, with only a slight retarding force caused by *L*/*R* decay of the induced currents in the sheet or loop guideway. The accompanying I^2R losses equal the thrust power that must be supplied to the propulsion system to overcome the magnetic drag.

The magnetic well can be designed to be very deep with null-flux suspensions, so that no conceivable external force can make the vehicle contact or leave the guideway. Typically, an external force of several times the vehicle weight is required for such an event to occur. These forces are far greater than those due to winds, curves, etc. The large margin also provides suspension redundancy, so that a large fraction of the SC magnets on the vehicle could fail without loss of levitation.

A slow application of external force (i.e., a quasisteady wind or curve) on the vehicle simply shifts its position slightly in the magnetic well. When the external force is removed, the vehicle then returns to its normal equilibrium point. However, a fast-pulsed external force, such as a sudden wind gust or an out-of-line guideway loop, may cause the vehicle to oscillate spatially about its equilibrium suspension point. Depending on the direction of this external force or forces, the oscillation can be in the vertical, lateral, pitch, yaw, or roll direction or any combination. Because the guideway loops have low electrical resistance and the vehicle oscillation frequency is much lower than the effective frequency at which the alternating SC magnets pass the guideway loops, vehicle oscillations are damped very weakly by loop resistance.

However, vehicle oscillations can be effectively countered and damped by active and/or passive control techniques, including variable pulsed magnets, secondary mechanical and/or hydraulic suspensions and shock absorbers/dampers, variable aerodynamic surfaces, etc. used separately or in combination. Good ride quality can then be maintained even when strong external impulse forces act on a vehicle.

2.6 Magnetic and Aerodynamic Drag

A high magnetic lift/drag ratio is very important for Maglev systems, in order to minimize energy consumption and propulsion power. The I^2R losses, P_{TOT}, in the guideway are affected by the magnitude of the guideway currents, guideway geometry, amount of conductor material, and its specific resistance. For a loop-type guideway, the I^2R losses are, summing over the N guideway loops that interact with the SC magnets in the vehicle,

$$P_{\mathrm{TOT}} = \sum_{j=1}^{j=N} (i_{\mathrm{GL}})_j^2 \frac{\rho_j l_j}{A_j}, \tag{10}$$

where $(i_{\mathrm{GL}})_j$ is the single-turn equivalent current in the jth guideway loop for levitation and/or stabilization, ρ_j the specific resistance of the conductor (2.8×10^{-6} Ω cm for aluminum), A_j the total conductor cross section of the multiturn loop, and l_j the single-turn equivalent winding length of the loop.

The equivalent magnetic drag force $(F_D)_{\mathrm{MAG}}$ on a vehicle moving at velocity V_0 is then

$$(F_D)_{\mathrm{MAG}} = P_{\mathrm{TOT}}/V_0, \tag{11}$$

and the corresponding magnetic lift/drag ratio, $(L/D)_{\mathrm{MAG}}$, for a vehicle of weight F_L is

$$\left(\frac{L}{D}\right)_{\mathrm{MAG}} = \frac{F_L}{(F_D)_{\mathrm{MAG}}} = \frac{F_L V_0}{P_{\mathrm{TOT}}}. \tag{12}$$

Since the I^2R losses are constant at speeds where the guideway loops behave inductively, rather than resistively, for constant conductor cross section the magnetic lift/drag ratio increases linearly with velocity.

This is in contrast to aerodynamic drag, where the equivalent aerodynamic lift/drag ratio, $(L/D)_{\mathrm{AERO}}$, for the vehicle decreases with increasing speed, as $(\text{velocity})^{-2}$:

$$\left(\frac{L}{D}\right)_{\mathrm{AERO}} = \frac{F_L}{(F_D)_{\mathrm{AERO}}} = \frac{2F_L}{\rho_{\mathrm{AIR}} C_D A_{\mathrm{F}} V_0^2}. \tag{13}$$

Here, ρ_{AIR} is the density of air, A_{F} the frontal area of the Maglev vehicle, and C_D the effective drag coefficient based on frontal area. The value of lift force F_L in Eq. (13) is not the true aerodynamic lift, but instead is the weight of the magnetically levitated vehicle. The value of $(L/D)_{\mathrm{AERO}}$ is thus a convenient way to compare aerodynamic drag to magnetic drag on a Maglev vehicle. By suitably shaping the Maglev vehicle, a modest aerodynamic lift could be generated to assist the magnetic lift at high speeds. However, the aerodynamic lift will be relatively small and could result in a greater drag force.

Figure 7 compares the magnetic lift/drag ratio for a vehicle with its aerodynamic lift/drag ratio. The magnetic lift/drag ratio is based on the NF-2 null-flux suspension illustrated in Fig. 4, with a conductor cross-sectional area of 10 cm^2. For the short sections of guideway near stations where vehicle speed is less than 150 km/h (95 mph), conductor cross section becomes greater to maintain the magnetic lift/drag ratio at 50/1. Maglev vehicles would reach 150 km/h after 20 s and a distance of 400 m, and achieve cruise speed (500 km/h) after 70 s and 5 km.

The aerodynamic lift/drag ratio shown in Fig. 7 is based on a 40-metric-ton vehicle with a frontal area of 10 m^2 and an effective drag coefficient of 0.2, representative values for Maglev systems.

The combined lift/drag ratio, $(L/D)_{\mathrm{COMB}}$, representing the sum of the two drag effects, is given by

$$\left(\frac{L}{D}\right)_{\mathrm{COMB}} = \frac{F_L}{(F_D)_{\mathrm{MAG}} + (F_D)_{\mathrm{AERO}}} = \frac{(L/D)_{\mathrm{AERO}}}{1 + (L/D)_{\mathrm{AERO}}/(L/D)_{\mathrm{MAG}}}. \tag{14}$$

Figure 7 shows that the combined lift/drag ratio is dominated by aerodynamic drag at cruise speed and by magnetic drag at low speed, i.e., near stations. At cruise speed the combined lift/drag ratio is approximately 15, corresponding to a drive power of 3.7 MW. With an 80% efficient linear synchronous propulsion, the actual drive power would be 4.6 MW. Electric energy consumption is less than 0.1 kW h per passenger kilometer of travel for a 100-passenger vehicle. At lower speeds, lift/drag ratio and energy efficiency increase. At 400 km/h, for example, energy consumption is only 0.05 kW h per passenger kilometer. Simple loop guideways also achieve high magnetic lift/drag ratios if external inductance is incorporated in the loop circuits to reduce guideway current.

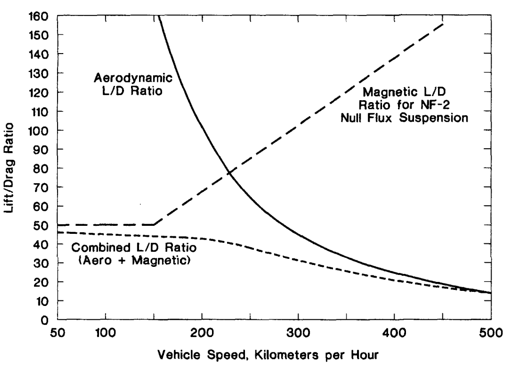

FIG. 7. Vehicle lift/drag ratios as a function of vehicle speed.

Conducting-sheet guideways have a low magnetic lift/drag ratio, since the guideway currents are large and inherently comparable to those in the superconducting magnets on the vehicle. At 500 km/h, the magnetic lift/drag ratio for conducting-sheet suspensions is in the range of 10–20, depending on design and the amount of lateral stability. Lift/drag ratio for conducting-sheet suspensions scales as (velocity)$^{-1/2}$, a result of the skin depth changing with speed. At cruising speeds of 500 km/h, the magnetic drag on a Maglev vehicle using a conducting-sheet suspension is comparable to the air drag. As a result, total drag is approximately double that of air drag, with a corresponding increase in propulsion power.

At speeds much above 500 km/h, the propulsion power to overcome air drag scales as (velocity)^3, and becomes impractically large. Moreover, the radius of curvature constraints on the horizontal and vertical straightness of the guideway also limit the maximum speed of Maglev vehicles. At 500 km/h, for example, the minimum horizontal radius of curvature for the guideway is 3 km. This assumes a bank angle of 24.5° and lateral acceleration of 1 m/s^2, which appears acceptable for passenger comfort (Powell and Danby, 1991). The radius of curvature scales as (velocity)^2, so that speeds much above 500 km/h require extremely straight and flat guideways. Accordingly, the practical limit for Maglev vehicles in the atmosphere is approximately 500 km/h. Airplanes travel faster, but at high altitudes where the air density is much less.

Approximately 70% of the proposed Tokyo–Osaka Maglev route is in deep tunnels through the mountains of central Japan. The tunnels will operate at atmospheric pressure; however, air drag is greater in enclosed tunnels (Maeda *et al.*, 1989) and may necessitate speeds less than 500 km/h. Studies of Maglev operating in low-pressure tunnels (Rather,

1992) indicate that at pressures of a few Torr, speeds of 2000 km/h are possible with very little propulsive power. Maglev technology for ultrahigh speed is similar to that for Maglev in the atmosphere; however, the low-pressure tunnels, and in particular, reducing their cost, requires long-term development. Ultrahigh-speed Maglev systems probably will not operate until well into the 21st century.

2.7 Magnetic Propulsion

The third important measure of Maglev performance, in addition to the "magnetic spring" behavior and lift/drag ratio, is propulsive effectiveness. The Linear Synchronous Motor (LSM) is used because it does not require mechanical contact. Figure 8 shows an idealized LSM system for superconducting Maglev.

The guideway incorporates an LSM drive winding that carries a small ac current. This ac current magnetically interacts with a SC loop on the vehicle, producing a thrust on it. The ends of the SC loop magnetically interact with the vertical field component of the LSM drive winding to create a forward thrust force on the vehicle. An equal and opposite magnetic reaction force on the ac LSM loop is simultaneously produced by the interaction of its current with the vertical field component from the SC loop.

The ac current in the LSM winding advances with a phase speed $V_{\mathrm{LSM}} = fz_p$, where z_p is the pole pitch of the alternating-polarity SC loops on the vehicle—i.e., z_p is twice the length of an individual SC loop—and f is the frequency of the ac current fed to the LSM drive winding. In Fig. 8(a), the SC loop on the vehicle is in a position where it magnetically interacts with the ac winding to propel the vehicle forward; in Fig. 8(b), the loop has advanced a distance 0.5 z_p in the time $(2f)^{-1}$ at the speed fz_p, and continues to receive a forward thrust.

In order to maintain phasing, the LSM drive windings and SC loops have the same pole pitch; however, the frequency of the ac current fed to the guideway can be changed to alter vehicle speed. Vehicles accelerate after leaving a station so that they can rejoin the main guideway. ac frequency monotonically increases until cruise speed is reached. When a vehicle decelerates, ac frequency de-

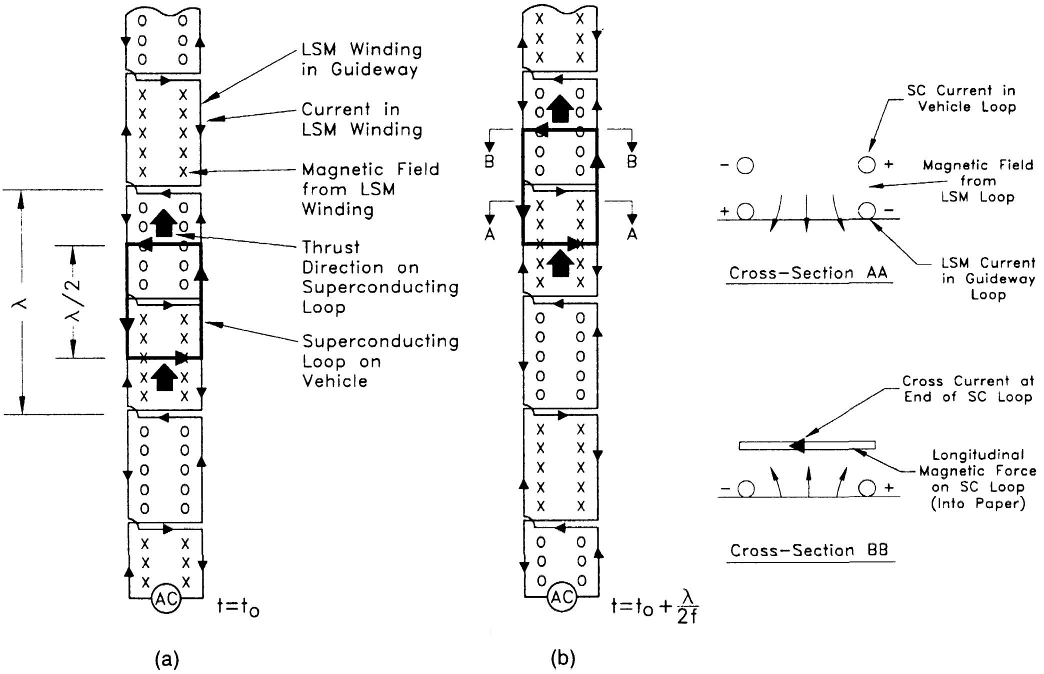
LSM Winding in Guideway
Current in LSM Winding
Magnetic Field from LSM Winding
Thrust Direction on Superconducting Loop
Superconducting Loop on Vehicle
λ
λ/2
AC
t=t$_0$
B
A
AC
t=t$_0$+$\frac{\lambda}{2f}$
SC Current in Vehicle Loop
Magnetic Field from LSM Loop
LSM Current in Guideway Loop
Cross-Section AA
Cross Current at End of SC Loop
Longitudinal Magnetic Force on SC Loop (Into Paper)
Cross-Section BB
(a)
(b)

FIG. 8. Operating configuration for the linear synchronous motor (LSM).

creases. When the ac wave runs in a braking mode, it feeds the kinetic energy of the vehicle back into the grid. Conversely, when the vehicle accelerates, it extracts energy from the grid.

Solid-state SCRs or other devices would control the frequency- and time-dependent LSM power flow to the guideway in the acceleration and deceleration sections. At cruise speed, the ac frequency is constant. Whether there should be variable ac drive-frequency capability in the cruise section of the guideway is an unresolved issue. There are arguments for this capability; it is easier to restart stalled vehicles on the guideway after grid power is restored following a large-scale blackout, and vehicles can operate at lower speeds if there is an emergency condition on the guideway. Providing this capability will increase guideway cost, however.

The loop and the conducting-sheet guideways differ with regard to LSM propulsion. In loop guideways, both simple loop and null-flux, the LSM system uses the same SC magnets that levitate and stabilize the vehicle. In contrast, conducting-sheet guideways cannot use the same SC magnets since the ac LSM currents would generate parasitic eddy-current losses in the sheet. Instead, there is an additional and separate line of SC magnets on the vehicle for LSM propulsion. In loop-type guideways, the loops are arranged so that they do not magnetically couple to the LSM drive windings.

Minimizing LSM winding current minimizes the I^2R losses and cost. LSM current is minimized by using strong SC magnets, efficiently coupling the SC magnets to the LSM windings, maximizing SC magnet coverage along the vehicle, and making the pitch of the SC magnet poles small.

The null-flux suspension enables very efficient LSM propulsion because of the high–field-strength magnets and excellent coupling. Figure 9 shows the thrust power on a 500-km/h Maglev vehicle with a NF-2 suspension, as a function of LSM current. Three cases are compared: case 1, 100% magnet fractional coverage along the vehicle and a pole pitch of 2 m; case 2, 50% coverage and pole pitch of 2 m; and case 3, 50% coverage and a pole pitch of 4 m.

The total magnetic thrust $(F_T)_{MAG}$ on the vehicle is the summation of the longitudinal magnetic forces acting on its SC magnets:

$$(F_T)_{MAG} = (N_{LI})(N_{LO})(f_C)\left(\frac{l_{vehicle}}{l_{pole}}\right) \times (2)\left(\frac{\mu_0}{2\pi}\right) I_{SC} \ln\left(\frac{W_0^2 + y_0^2}{y_0^2}\right), \quad (15)$$

where N_{LI} is the number of lines of SC magnets along the vehicle (generally, $N_{LI} = 2$); N_{LO} is the number of guideway loops per SC magnet ($N_{LO} = 2$ for the NF-2 suspension); f_C is the fraction of the vehicle length that has SC magnets; $l_{vehicle}$ is the total vehicle length; and l_{pole} is the length of a SC loop on the vehicle ($= \frac{1}{2}$ of pole pitch).

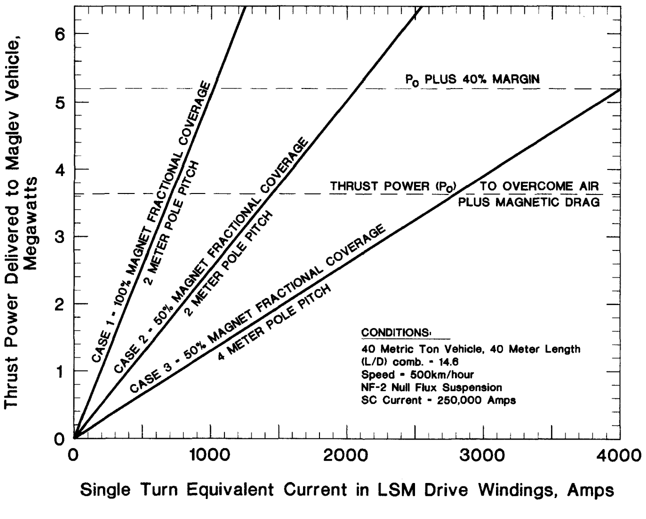

FIG. 9. Thrust power delivered by the linear synchronous motor (LSM) as a function of LSM current and SC magnet coverage and pole pitch.

With 100% coverage and a pole pitch of 2 m (case 1), the LSM drive requires only 1000 A to overcome air and magnetic drag on a Maglev vehicle at 500 km/h, with a margin of 40% for occasional head winds and upgrades. The phase angle of the SC magnets relative to the current wave in the LSM windings shifts if drag force changes, as does delivered thrust power, but vehicle speed remains constant. Clearly, it is advantageous to have strong SC magnets, as complete a coverage as possible, and short pole pitch.

Low-current LSM designs, i.e., ~1000 A, allow a 1–2-km-long block to be energized with high LSM motor efficiency (80% or more), because the I^2R losses are small. High-current LSM designs require much shorter block lengths; this increases the number and cost of controls, switches, and transformers along the guideway. Optimizing LSM design involves many trade-offs; in general, however, low-current designs appear preferable.

2.8 Electromagnetic Suspensions

Electromagnetic Maglev suspensions are very different from superconducting suspensions. They are inherently unstable and require active servo control of magnet current to prevent contact of the vehicle with the guideway. When the vehicle moves upward, the air gap between its electromagnets and the iron rails on the underside of the guideway decreases. If magnet excitation current remained constant and the iron in the magnetic circuit did not saturate, the magnetic field strength and the upward attractive force would increase. To first order, the attractive force F_Z scales with the air-gap dimension Z as

$$F_Z/mg = (Z_0/Z)^2, \tag{16}$$

where Z_0 is the gap dimension at which the attractive force equals vehicle weight. The net upward acceleration on the vehicle increases rapidly with decreasing Z; at $Z/Z_0 = 0.5$, for example, it is $3g$ if the magnet current is constant. At some point, saturation in the iron limits the net upward acceleration; the vehicle would still hit the guideway, however.

Using room-temperature electromagnets, such as those on the Transrapid and HSST Systems, Z_0 must be small, approximately 1 cm, to keep magnet power reasonable. As a consequence, small changes in gap along a guideway that result from normal construction variations or from differential settling after construction pose challenges for the magnet control system.

As the vehicle moves at high speed along the guideway, small changes in the position of the iron rail appears to the vehicle as a fluctuating thickness of the air gap. A change of 1 mm in gap thickness corresponds to a $0.2g$ net acceleration on the vehicle, with the acceleration upward if the gap thickness decreases, and downward if it increases. Because the vertical movements are very small, the magnet control system must respond very quickly. The necessary response time is on the order of 10^{-2} s.

Safe operation of electromagnetic suspensions has been demonstrated in full-scale tests of the Transrapid and HSST systems. Because of the small clearance, guideways for electromagnetic Maglev must be built and maintained to much tighter tolerances—approximately a factor of 10—than those for superconducting Maglev. Tests of the JR superconducting Maglev system, at Miyazaki, for example, have shown that high-speed operation is not affected when individual guideway loops are displaced by several centimeters to simulate the effects of an earthquake. Such displacements are not acceptable for electromagnetic Maglev. The much tighter tolerances for construction and maintenance of electromagnetic Maglev increase guideway cost substantially.

3. MAGLEV DEVELOPMENT AND PROTOTYPE DEMONSTRATION

Maglev development has proceeded in various countries during the last three decades. The Japanese and German programs have been by far the largest, both in their total funding and in the demonstration of operating prototypes. Each has spent over 10^9 dollars on the development of Maglev.

3.1 Japanese Superconducting Maglev

In 1970 the Japanese National Railway (JNR) started major development of superconducting Maglev, under the technical leadership of Kyotani (1975). They operated test vehicles in 1972 that demonstrated both levitation and LSM propulsion. In 1977 JNR began operating a large (10-t) vehicle, the ML-500, on a 7-km test track at Miyazaki. This and subsequent vehicles were based on the superconducting electrodynamic inventions of Powell and Danby.

Lift was generated by two lines of sequential multiturn horizontal loops of aluminum wire placed along the guideway. These loops interacted with eight superconducting levitation magnets on the underside of the vehicle. At the center vertical slab of the inverted-T–shaped guideway, additional aluminum loops were vertically mounted and cross connected. These provided lateral stability by null-flux interaction with eight additional superconducting magnets on the vehicle. These were vertically located on the sides of the slot in the vehicle that surrounded the vertical slab portion of the inverted-T guideway. Aluminum LSM drive windings on the guideway interacted with the superconducting magnets on the vehicle to propel it. Wheels supported the vehicle at low speed and at rest. Speeds well above 500 km/h were demonstrated.

The inverted-T guideway was modified to a U-shaped channel, and a new MLU series of vehicles were tested beginning in 1980. Coupled cars (up to 3) were operated, with capability for passengers. Speeds to 400 km/h with two cars and passengers were demonstrated. The eight superconducting magnet loops on the vehicle were vertically oriented, four on each side. Aluminum loops in the guideway underneath the vehicle provided lift. Horizontal stability and motor winding loops were located on the side walls of the U-shaped guideway. In later designs the vehicle magnets are placed on bogies at its ends, remote from the passengers, to reduce their exposure to stray magnetic field.

Figure 10 shows the planned Tokyo-to-Osaka route, which runs through the mountainous center of Japan. Most of the route will be located in deep tunnels. The first portion, a 43-km section in Yamanishi Prefecture, is now under construction with tunnels, switching, and stations. The Yamanishi track will provide a realistic full-scale demonstration of commercial system capability. The vehicles use a U-shaped guideway. Vertically oriented figure-of-eight null-flux loops on the sidewalls of the guideway provide combined horizontal stability and vertical levitation. The loops are compatible with the superimposed LSM drive windings. The guideway loops and windings interact with two rows of vertically oriented superconducting planar loops on each side of the vehicle bogies. Each 22-m-long vehicle weighs 18 t and carries 44 passengers. Trains with up to 14 coupled vehicles are planned for the commercial system.

JNR has carried out experiments in which vehicles moved past guideway loops that had

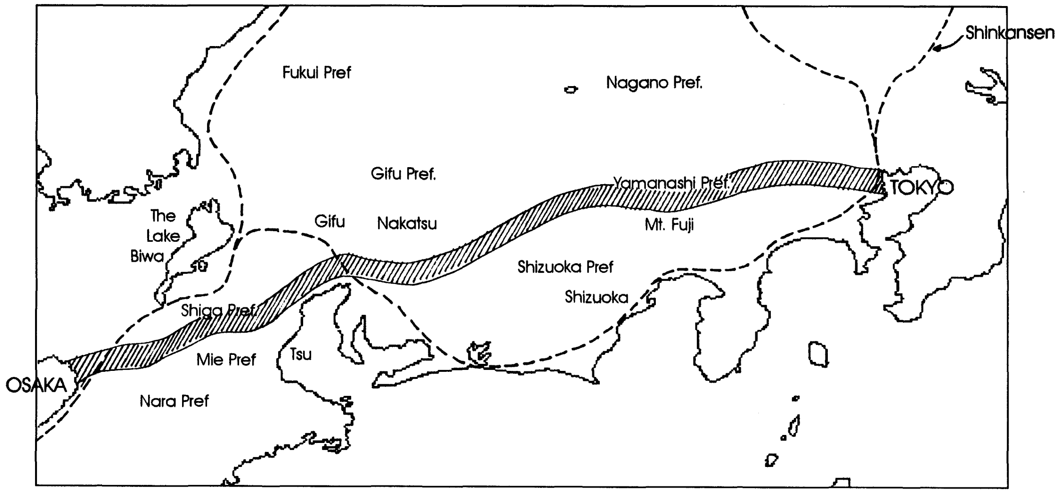

FIG. 10. Tokyo–Osaka route proposed for Japanese superconducting Maglev system.

large displacements from their normal design position to simulate the effects of earthquakes. A conventional rail train would have derailed upon encountering such large displacements. The Maglev vehicle did not experience any problems when it rode through the displaced sections. This unique safety capability, plus the advantages of low noise, minimum maintenance, and steep grade-climbing ability, make electrodynamic Maglev very attractive for a country like Japan that is organized for rail travel. Figure 11 shows a JNR demonstration vehicle.

3.2 Japanese Electromagnetic Maglev

Japan Air Lines in about 1970 initiated study of an electromagnetic Maglev system suitable for travel from Tokyo to the Narita airport. In 1973 they began construction of a prototype that attained 300 km/h in 1978. Their demonstration vehicle HSST-03 carried very large numbers of passengers at expositions in Japan and Canada. This small-gap (1 cm) system, similar in concept to the German Transrapid, is intended for moderate speeds up to 300 km/h. Such speeds do not necessarily require LSM propulsion but can use an onboard propulsion system (Suzuki *et al.*, 1987).

3.3 German Electromagnetic and Superconducting Maglev

Germany began development of both superconducting Maglev and servo-controlled electromagnetic Maglev in 1970. Their superconducting Maglev program culminated in 1976 with the test of a 12-m-long vehicle. The vehicle weighed 17 t, and it had a 10-cm gap along the guideway. It operated at speeds up to 230 km/h on a 280-m-diam circular track in Erlangen (Miller, 1987). The vehicle was propelled by a linear induction motor, which was separate from the levitation system.

In 1978 Germany discontinued work on superconducting Maglev and concentrated on the development of the electromagnetic Transrapid, because it was judged closer to commercial application. The 12-t TR02 attained 163 km/h on a 1-km track. The next step was the TR04 vehicle that weighed 18 t and carried 12 passengers at 250 km/h on a 3-km test track in the late 1970s. In 1979 a 70-passenger Transrapid demonstration vehicle at the Hamburg Fair carried passengers on a 1-km track. In the early 1980s a 32-km test track with loops at each end and a 12-km straight section was constructed at Emsland. The guideway on the end loops has a 12° bank angle.

The TR06 prototype vehicle consisted of two connected sections each 27 m long and 3.65 m wide. Total weight was 102 t without passengers or cargo. TR06 had a 20-t payload capacity and carried 196 passengers. A Linear Synchronous Motor propelled the vehicle, with guideway coils reacting against the levitation magnets. TR06 attained a maximum speed of 400 km/h (Wiescholek *et al.*, 1993). The latest design is the TR07 (Miller, 1987). It is slightly smaller in cross section and lighter than the TR06 (90 t for two sec-

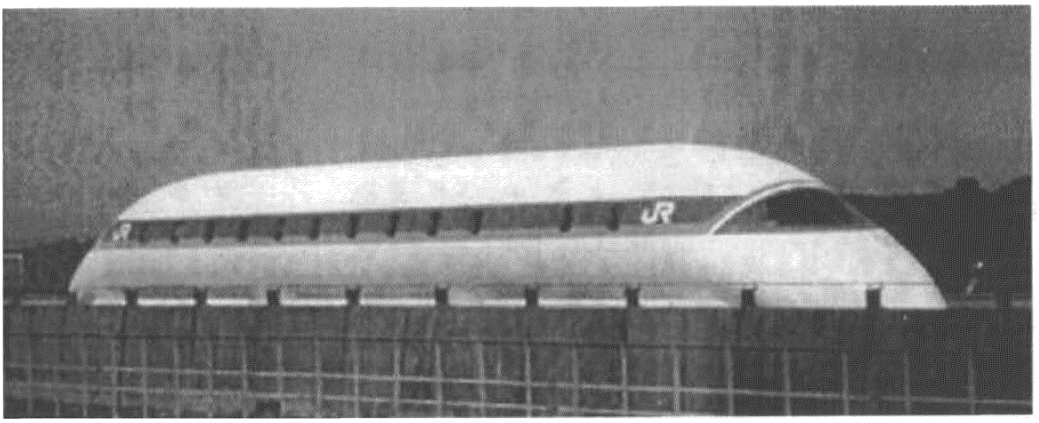

FIG. 11. Prototype JR superconducting Maglev vehicle.

tions as compared to 102 t). The air gap between the vehicle magnets and the guideway iron rails is 0.8 cm. Maximum design speed is 500 km/h.

Transrapid requires accurate position sensors and fast-acting servo control of the power fed to the vehicle electromagnets. For safety in case of guideway power loss, onboard batteries are employed. These are charged by a linear generator as the vehicle moves, picking up power from the active guideway. The electromagnets are attached to the vehicle by pneumatic springs and linkages. Figure 12 shows the Transrapid vehicle on its guideway. The vehicle rides above the T-shaped guideway shown in the photograph. The arms at the sides of the vehicle curl around under the guideway and hold the servo-controlled electromagnets, which are attracted upward to the two iron rails in the underside of the flat-topped guideway.

A mechanical switch that can transfer vehicles from one guideway to another has been demonstrated. Maximum vehicle speed on these switch sections is limited to 200 km/h. The switch section, which is 30 m in length, is pushed sideways when it is desired to transfer the vehicle from one guideway line to the other. The guideway must be constructed and maintained to very high tolerances because of the small (0.8 cm) air gap. Devices to help automate guideway fabrication and installation have been developed. In Germany, plans have been announced for a Hamburg-to-Berlin route.

3.4 U.S. Maglev Studies

After funding for U.S. Maglev development stopped in the early 1970s, individual nonfunded efforts continued on a sporadic basis; but these were limited to conceptual designs and analyses. In the early 1990s the United States resumed funding of Maglev at a low level under the National Maglev Initiative (NMI), following the favorable report by the U.S. Senate Task Force on Maglev (Powell and Danby, 1991). The NMI program was restricted to conceptual studies of generic components like guideways and power systems and system concept definition (SCD). Four SCD conceptual studies were completed, which are described below. Funding of $\$900 \times 10^6$ for Maglev prototype development was authorized as part of the Intermodal Surface Transportation Efficiency Act (ISTEA), but not appropriated, so that the United States had not yet (1994) undertaken the development of Maglev hardware.

The four SCD designs were primarily based on modifications of the existing Japanese and German Maglev programs and the U.S. efforts of two decades earlier. The Foster–Miller (1993) SCD design (Fig. 13) is a modification of the Japanese superconducting Maglev system to achieve a lighter-weight, more open guideway structure. Salient features of the design that are similar to those for the Japanese system include the placement of the superconducting magnets at the front and rear of the vehicle [Fig. 13(a)], to minimize the magnitude of stray fields in the passenger compartment, and the

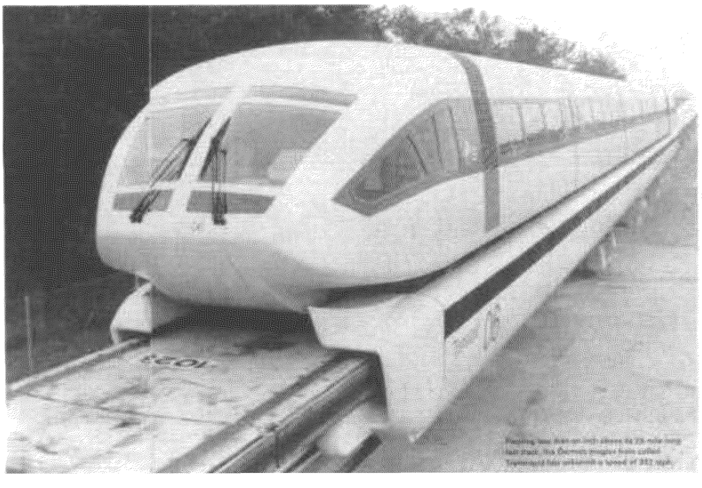

FIG. 12. Prototype Transrapid Electromagnetic Maglev vehicle.

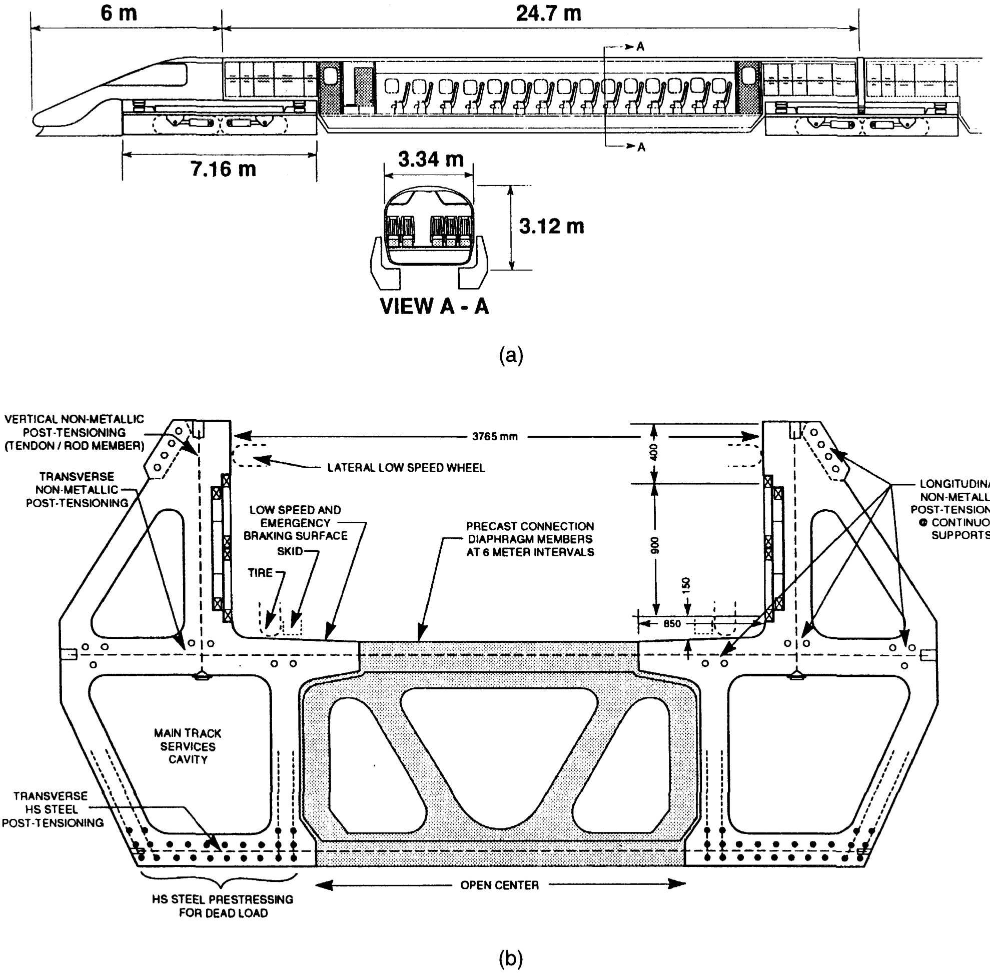
6 m
24.7 m
A
A
7.16 m
3.34 m
3.12 m
VIEW A - A
(a)
VERTICAL NON-METALLIC POST-TENSIONING (TENDON / ROD MEMBER)
TRANSVERSE NON-METALLIC POST-TENSIONING
3765 mm
LATERAL LOW SPEED WHEEL
LOW SPEED AND EMERGENCY BRAKING SURFACE
SKID
TIRE
PRECAST CONNECTION DIAPHRAGM MEMBERS AT 6 METER INTERVALS
400
900
150
850
LONGITUDINAL NON-METALLIC POST-TENSIONING
● CONTINUOUS SUPPORTS
MAIN TRACK SERVICES CAVITY
TRANSVERSE HS STEEL POST-TENSIONING
HS STEEL PRESTRESSING FOR DEAD LOAD
OPEN CENTER
(b)

FIG. 13. Superconducting Maglev design concept proposed by the Foster–Miller Corporation.

use of the sidewall for the null-flux levitation/stability loops and the LSM drive windings [Fig. 13(b)]. The guideway structure is opened up to reduce weight and to inhibit the buildup of snow and ice in cold climates.

The Grumman SCD design (Proise *et al.*, 1993), shown in Fig. 14, is a modification of the German Transrapid Electromagnetic Maglev System. It uses superconducting windings to energize the vehicle electromagnets instead of the normal room-temperature copper windings used on Transrapid. This allows the gap between the vehicle magnets and the iron rails on the underside of the guideway to be considerably larger than that of Transrapid, i.e., 4 cm, compared to 0.8 cm, while keeping the weight of the vehicle magnets within reliable limits. However, the suspension is still inherently unstable, and an active servo system must control the attractive force exerted by the vehicle electromagnets in order to prevent contact with the guideway. In the Grumman concept the superconducting windings on the vehicle magnets supply most of the attractive levitation force, while servo-controlled copper auxiliary windings with adjustable current capability are used to vary the attractive force as required.

The Magneplane SCD design, proposed by Magneplane International (1992), retains the principal features of the original 1972 Magneplane concept (Kolm and Thornton, 1972), as illustrated in Fig. 15. The two arcs of the

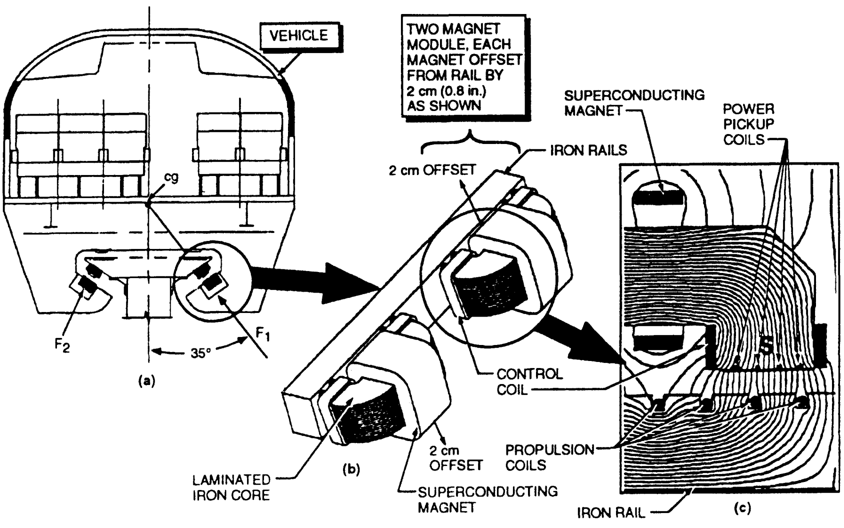
VEHICLE
TWO MAGNET MODULE, EACH MAGNET OFFSET FROM RAIL BY 2 cm (0.8 in.) AS SHOWN
SUPERCONDUCTING MAGNET
POWER PICKUP COILS
IRON RAILS
2 cm OFFSET
cg
F2
F1
35°
(a)
CONTROL COIL
LAMINATED IRON CORE
(b)
2 cm OFFSET
PROPULSION COILS
SUPERCONDUCTING MAGNET
IRON RAIL
(c)

FIG. 14. Electromagnetic Maglev design concept using superconducting magnet windings proposed by the Grumman Corporation.

conducting-sheet guideway are trough shaped, to allow the vehicle to bank when rounding a curve. Bank angle can be controlled so that the fixed guideway is compatible with coordinated turns at different vehicle speeds, without using a variable-tilt secondary suspension on the vehicle. A second set of superconducting magnets along the center of the vehicle is used for LSM propulsion; these are separate from those used for levitation. Separation of levitation and propulsion magnets is necessary with a conducting-sheet guideway, as discussed previously in this article.

The Bechtel–MIT (1992) SCD design has a beam-type guideway that is straddled by the

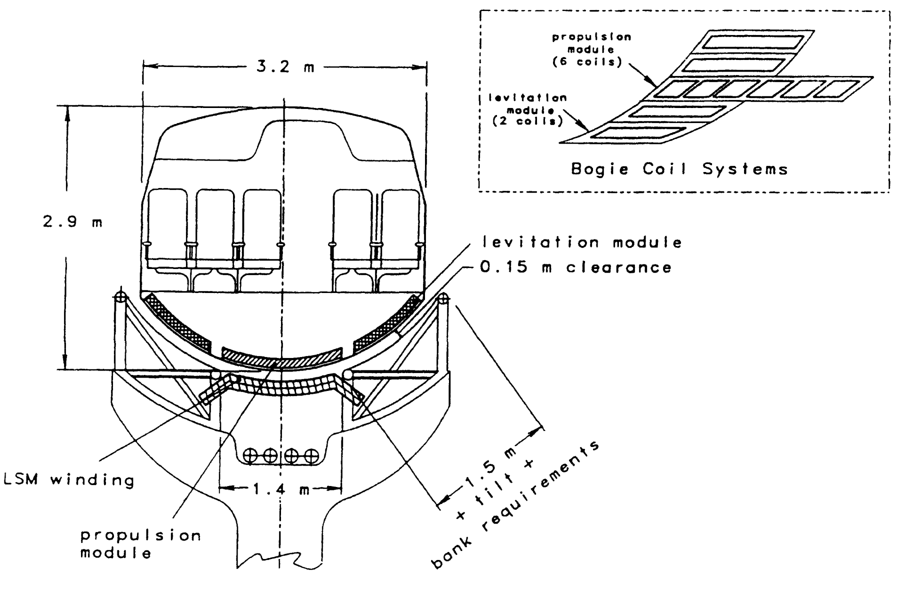
propulsion module (6 coils)
levitation module (2 coils)
Bogie Coil Systems
3.2 m
2.9 m
levitation module
0.15 m clearance
1.5 m + tilt + bank requirements
LSM winding
1.4 m
propulsion module

FIG. 15. Superconducting Maglev design concept proposed by the Magneplane Corporation.

vehicle, as illustrated in Fig. 16. A null-flux–type configuration is used to levitate and stabilize the vehicle, both vertically and horizontally. The vehicle superconducting magnets are arranged in an equivalent vertical figure-of-8 winding pattern. When the winding is at the central symmetry plane for the ladder-type aluminum guideway (a sequentially connected set of simple dipole loops), the net flux through the rectangular guideway loops is zero, and no current flows in them. When the vehicle is displaced upward or downward from the symmetry plane, the net flux through the guideway loops induces current in them that acts to push the vehicle back to the central symmetry plane. The Japanese and Bechtel–MIT designs are magnetically equivalent, with the first design using simple dipole superconducting vehicle magnets and figure-of-8 guideway loops, while the second design uses figure-of-8 superconducting magnets and simple dipole guideway loops.

4. FUTURE APPLICATION OF MAGLEV

The technical feasibility of Maglev as a transportation system has been demonstrated by the Japanese superconducting and German electromagnetic Maglev programs. Both systems are at the commercial prototype stage of development, and both plan to build first-generation operational routes.

However, the economic practicality of Maglev as a future transportation system has yet to be demonstrated. The present projected guideway costs for both types of Maglev systems are in the range of $\$20 \times 10^6$ per kilometer of two-way guideway, depending on the design and estimators. Engineering improvements are needed to reduce these costs significantly, to a level of $\$10 \times 10^6$/km or less, and to enhance the capability of Maglev vehicles to carry freight and to easily switch to off-line stations. These improvements,which appear achievable, are discussed in the next section of this article. Fundamental breakthroughs do not appear necessary for the successful large-scale application of Maglev—just the normal evolution of established basic concepts and technology.

4.1 Energy, Environmental, Economic, and Societal Benefits

Maglev can yield many important benefits. First, it is much more energy efficient than present transport. Figure 17 compares primary energy usage for Maglev, autos, and

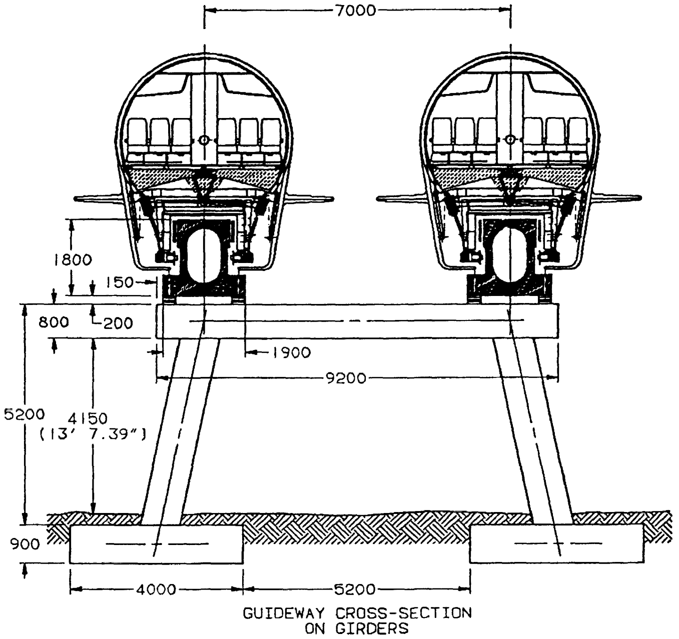

FIG. 16. Superconducting Maglev design concept proposed by MIT/Bechtel.

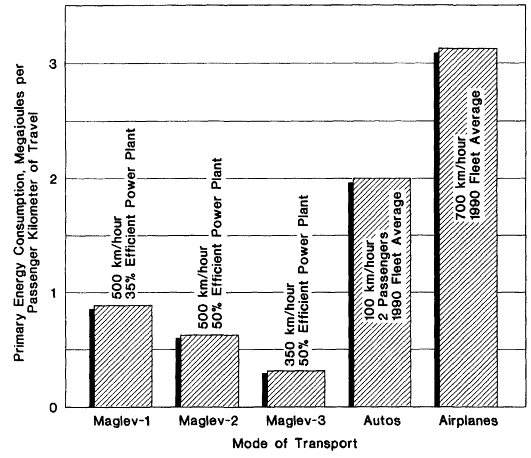

FIG. 17. Comparison of primary energy usage per passenger kilometer of travel for Maglev with other modes of transport.

airplanes, per passenger kilometer of travel. For Maglev, primary energy is defined as the thermal energy into the power plant that generates electricity; for autos or airplanes, it is the heating value of the fuel used. Current power plants operate at an average thermal efficiency of 35% (thermal energy to electricity). Per passenger kilometer of travel, Maglev at 500 km/h (~300 mph) uses 50% of the primary energy consumed by 100-km/h autos, and only 25% of that for 700-km/h airplanes. With the new more efficient (~50%) combined-cycle power plants that are beginning to come on line, Maglev primary energy consumption is only 35% of autos and 15% of airplanes. By operating at somewhat lower speed, e.g., 350 km/h (210 mph), Maglev primary energy consumption can be reduced to 18% of autos and 8% of airplanes.

The much lower energy usage, plus the fact that electricity usually comes from nonpetroleum sources—that is, coal, hydro, and nuclear—makes Maglev very effective for reducing oil consumption. Transport accounts for approximately 50% of total U.S. oil consumption; implementing Maglev on a large scale would substantially reduce oil usage, lowering oil imports—now over 50% of total consumption—and helping balance of trade problems.

The transport sector is also the major contributor to air pollution, principally by autos and trucks. Maglev, because it uses much less primary energy, and because its electricity comes from central power plants where emissions can be controlled, will greatly reduce air pollution.

Figure 18 compares pollution emissions for Maglev and autos, per billion passenger kilometers of travel. In terms of emissions that are harmful to health (hydrocarbons, NO_x, sulfur dioxide, carbon monoxide), Maglev pollution is far less than autos. In terms of CO_2 emissions and global warming, Maglev is better, by approximately a factor of 3, than autos.

The economic benefits of Maglev appear to be very large. The United States currently spends about $\$10^{12}$ (one trillion) annually on auto, truck, airplane, and rail transport. Of this total, over 50% is spent in intercity transport—$\$2.5 \times 10^{11}$/y on autos, $\$1.6 \times 10^{11}$/y on trucks, $\$6 \times 10^{10}$/y on air, and the balance on bus and rail (primarily bulk freight).

The projected operating costs for Maglev, autos, trucks, and airplanes are illustrated in Fig. 19. These operating costs do not include capital costs associated with the construction of facilities, such as airports, highways, and guideways. In general, Maglev has much smaller operating costs. This results from Maglev's lower energy usage and equipment cost per unit of travel. To exploit these lower operating costs, Maglev routes need a large volume of traffic so that its expensive guideway can be paid for by revenues.

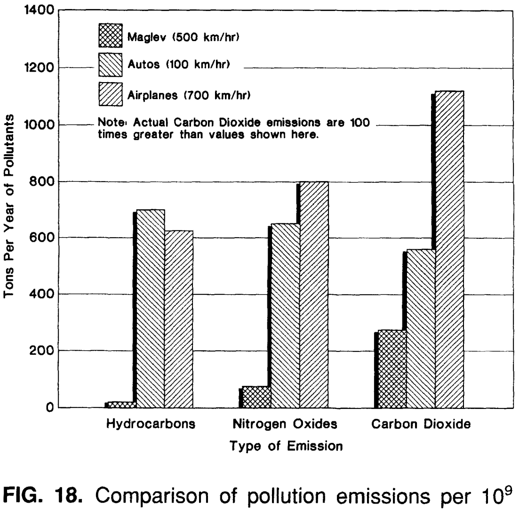

Note: Actual Carbon Dioxide emissions are 100 times greater than values shown here.

FIG. 18. Comparison of pollution emissions per 10^9 km of passenger travel for Maglev with other modes of transport.

TRANSPORT MODE	REVENUE MODE	
	PASSENGERS (Cents/Passenger km)	FREIGHT (Cents/Ton km)
MAGLEV	2	4.5
AUTOS (2 PASSENGERS)	10	-
TRUCKS	-	20
AIRPLANES	10	100

FIG. 19. Comparison of Maglev operating costs with those of other modes of transport.

An illustrative National Maglev Network for the United States is shown in Fig. 20. This example is based on the use of the rights of way along portions of the Interstate Highway System and is drawn so as to connect the major metropolitan areas in the United States. It has a total length of 28 000 km and 300 stations. The network directly serves 170 million persons (70% of U.S. population), who live within 25 km of a station. Over 90% (245 million) of the population live in states with Maglev routes. Capital investment for the network is approximately $\$200 \times 10^9$, including stations. This is less than the replacement cost in current dollars for the Interstate highways.

Figure 21 shows the payback time for this illustrative Maglev Network as a function of the fraction of the U.S. transport market that it captures. If the network carries 20% of intercity passenger and truck freight, for example, its investment could be recovered in only 3 years. Most infrastructure projects like highways, dams, and airports typically never recover their investment, or if they do, they take many years to do so.

Because the operating costs of Maglev transport will be much lower than those of competing modes, and because Maglev routes can carry extremely high traffic loads, a Maglev Network will always be able to carry goods and people at lower cost than other modes. Thus Maglev should ultimately capture a large share of the transport market, probably well above 20%.

Finally, many societal benefits can be expected from Maglev. The death rate on U.S. highways is 45 000 persons per year, with additional hundreds of thousands of major injuries. Many more thousands die or suffer

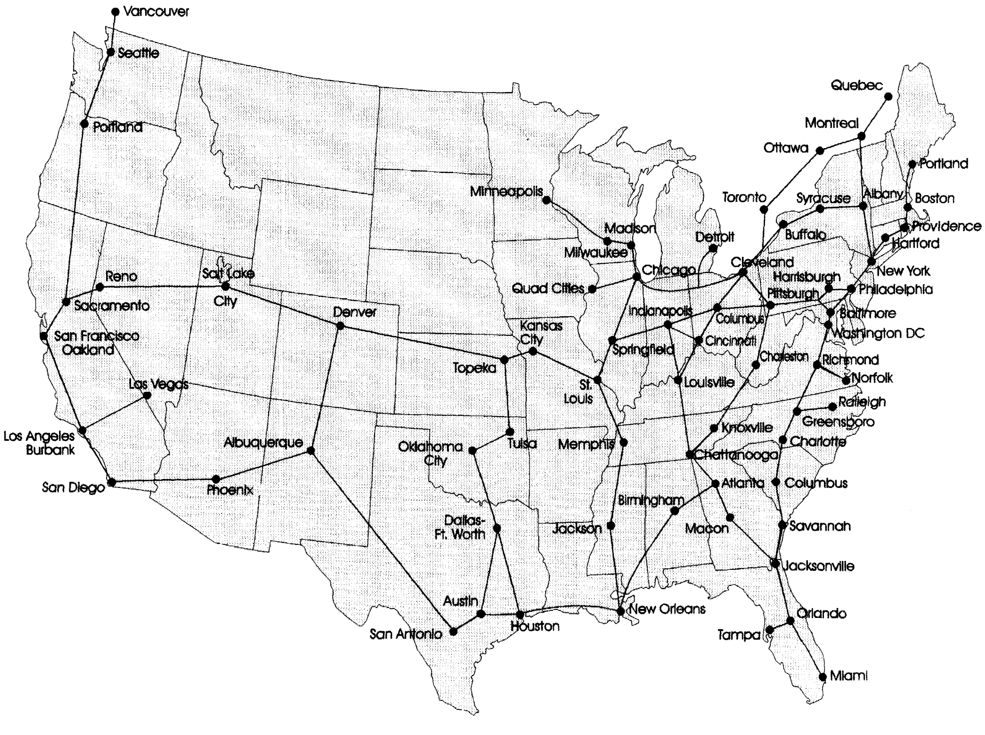
Vancouver
Seattle
Portland
Reno
Sacramento
San Francisco
Oakland
Salt Lake City
Denver
Las Vegas
Los Angeles
Burbank
San Diego
Phoenix
Albuquerque
Topeka
Kansas City
Oklahoma City
Tulsa
Dallas-Ft. Worth
Austin
San Antonio
Houston
Minneapolis
Madison
Milwaukee
Quad Cities
Chicago
Springfield
St. Louis
Memphis
Jackson
Birmingham
New Orleans
Indianapolis
Louisville
Cincinnati
Columbus
Detroit
Cleveland
Toronto
Ottawa
Montreal
Quebec
Syracuse
Albany
Portland
Boston
Providence
Hartford
Buffalo
Harrisburgh
Pittsburgh
New York
Philadelphia
Baltimore
Washington DC
Charleston
Richmond
Norfolk
Raleigh
Greensboro
Charlotte
Knoxville
Chattanooga
Atlanta
Columbus
Macon
Savannah
Jacksonville
Orlando
Tampa
Miami

FIG. 20. A potential national Maglev network connecting major U.S. population centers.

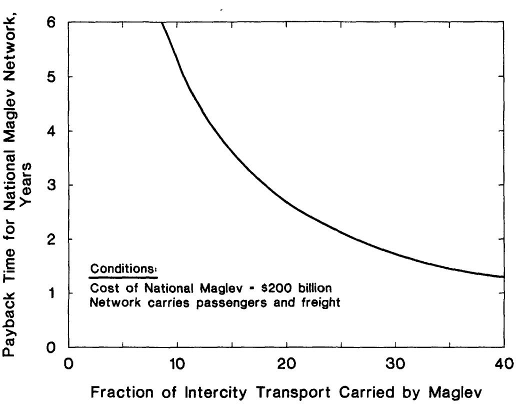

FIG. 21. Payback time for a U.S. national Maglev network as a function of intercity transport carried on the network.

serious health problems as a result of pollution from autos and trucks. A large fraction of these deaths, injuries, and health problems would be eliminated if Maglev becomes a major mode of transport.

Another important societal benefit of Maglev is enhanced mobility—a result of faster travel speed and the elimination of delays due to congestion and bad weather. Improved mobility enables better industrial competitiveness through just-in-time manufacturing. It also enables people to live in smaller, more attractive population centers that are physically distant, but closely linked in time.

4.2 Engineering Developments Needed for Large-Scale Implementation of Maglev

To be implemented on a large scale, Maglev needs further development to reduce costs, increase potential markets, simplify and speed up the construction of guideways, and increase capability to serve effectively a wide network of many stations. These engineering developments are particularly needed for the United States, because of its many dispersed population centers.

No fundamental breakthroughs appear necessary. The three principal engineering developments needed are

1. lower-cost, easily erected guideways;
2. capability to carry truck-type freight; and
3. high-speed switching to off-line stations.

Lower guideway costs can be achieved by minimizing the amounts of material, and by using prefabricated beams and piers, rather than field construction. Figure 22 shows a narrow-beam guideway design using prefabricated, reinforced concrete beams and piers. The projected cost for this two-way guideway is $\$7 \times 10^6$/km, a factor of 3 lower than present costs. In this design, beams are 20 m long and weigh 15 metric tons; pier height is 10 m, with a weight of 15 metric tons.

The beams and piers would be mass produced at factories and shipped by road or rail to the construction site. To minimize traffic disruptions, the beams and piers would be transported along already completed sections of guideway, and erected on preinstalled footings. This eliminates the delays and detours normally associated with heavy construction. Guideways could then be installed on existing rights of way alongside highways, rail lines, and power lines, without major tie-ups.

Obtaining new rights of way for transport projects is very difficult in the United States. Maglev, unlike high-speed rail, can use existing rights of way because its elevated guideways will pass over existing structures, and readily adapt to changes in grade and curves. However, disruptions to existing facilities and traffic flow during construction must be minimal.

Capability to carry truck-type freight is very important, particularly in the United States, where intercity truck freight outlays are three times larger than those for air passengers. Maglev can be a "land ferry," transporting freight containers and trailers (Fig. 23) many hundreds of miles at lower cost and much faster than by highway or rail. Transporting truck-type freight across the United States in a few hours would facilitate just-in-time manufacturing. Maglev systems, with high magnetic lift-to-drag ratios and heavy lift capability, can efficiently transport freight long distances. A Maglev freight vehicle could carry a 40-metric-ton payload, with a gross loaded weight of 60 tons and a magnetic lift-to-drag ratio on the order of 100 to 1.

The third engineering development is the ability to switch Maglev vehicles at high speed from a main line guideway into off-

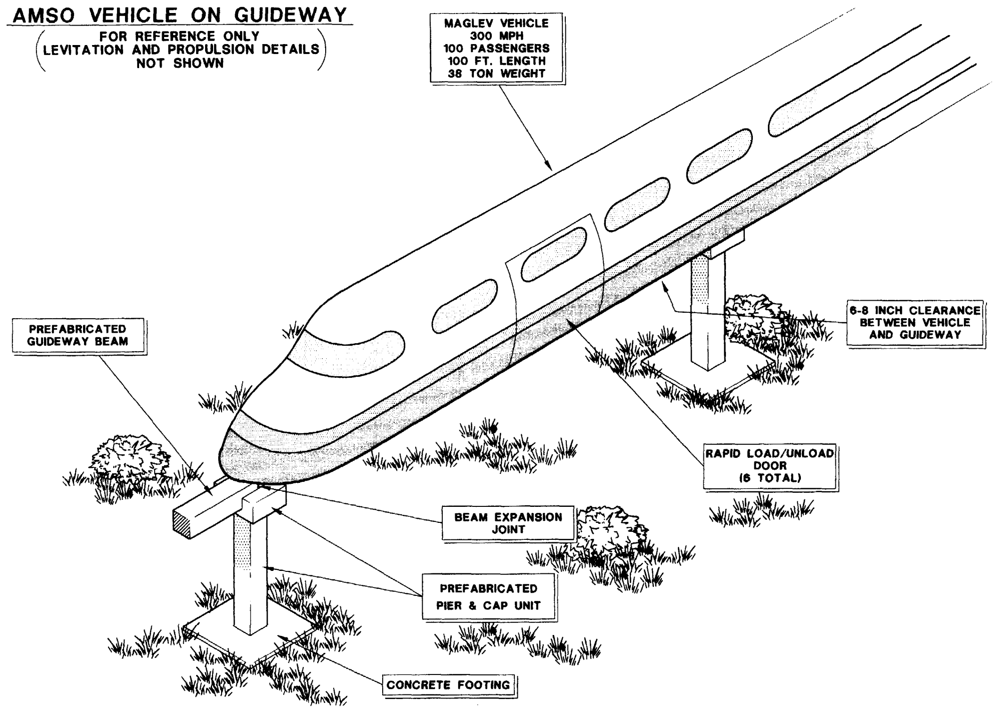
AMSO VEHICLE ON GUIDEWAY
FOR REFERENCE ONLY
LEVITATION AND PROPULSION DETAILS NOT SHOWN
MAGLEV VEHICLE
300 MPH
100 PASSENGERS
100 FT. LENGTH
38 TON WEIGHT
PREFABRICATED GUIDEWAY BEAM
6-8 INCH CLEARANCE BETWEEN VEHICLE AND GUIDEWAY
RAPID LOAD/UNLOAD DOOR (6 TOTAL)
BEAM EXPANSION JOINT
PREFABRICATED PIER & CAP UNIT
CONCRETE FOOTING

FIG. 22. Maglev systems based on a low-cost prefabricated narrow-beam guideway.

line stations, and then switch them back onto the guideway after loading. If Maglev vehicles stopped at every station on the route, stations would have to be spaced at intervals of 100 km or more to maintain reasonable average speed. Such a wide spacing would reduce accessibility and revenues, and lessen public support for Maglev.

Present Maglev vehicles either cannot switch to a secondary guideway at all or rely on low-speed mechanical switches. Superconducting Maglev vehicles should be able to switch electronically from one guideway to another at high speed, but this has not yet been demonstrated. With high-speed switching, "skip-stop" Maglev vehicles can bypass

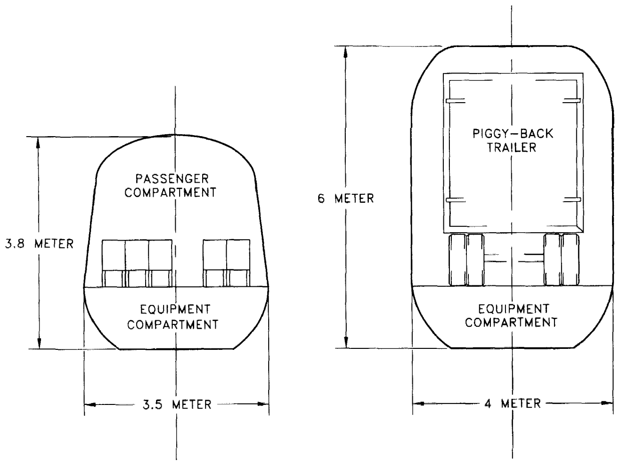
PASSENGER COMPARTMENT
EQUIPMENT COMPARTMENT
3.8 METER
3.5 METER
6 METER
PIGGY-BACK TRAILER
EQUIPMENT COMPARTMENT
4 METER

FIG. 23. Comparison of Maglev vehicles designed for passenger transport with those designed for freight transport.

stations at high speed, stopping only where necessary (Fig. 24). Vehicles would switch from the high-speed guideway onto a secondary guideway leading to the station, decelerating in approximately 3 km. After unloading/loading, vehicles would then accelerate back to full speed and rejoin the main guideway. Many stations could then be located on the network, without penalizing speed.

The issue of high-temperature superconductors is sometimes raised in the belief that they are needed for Maglev. In fact, present niobium–titanium superconductors and liquid-helium cryogenic systems are technically and economically practical for Maglev. The vehicle superconducting magnets, cryostats, and closed-cycle refrigerators account for only 1% of system capital cost. Ninety percent of the cost is due to the guideway and propulsion system, with the remainder for stations, vehicle bodies, etc. A useful illustration is provided by the Superconducting Super Collider (SSC) Project, which was recently cancelled. The SSC magnets were far more technically demanding than those required for Maglev vehicles, because of their very stringent requirements on field quality, maximum field, conductor movement, cycling capability, radiation resistance, heat leaks, etc. Even with these very challenging requirements, the cost of the magnets in the 100 km SSC ring was projected to be on the order of $\$3 \times 10^9$. The same length of magnets in Maglev vehicles would transport 30% of U.S. intercity passenger traffic. Even if Maglev magnets cost as much as SSC magnets, which is unlikely, their amortized cost would be 0.1 cent per passenger mile, a trivial amount.

4.3 Environmental and Safety Aspects

High-speed rail transport is exceptionally safe. In over 20 years of travel, the Japanese Shinkansen system has never had a fatal accident, even though it carries hundreds of thousands of passengers each day. Maglev should be even safer than high-speed rail: First, the guideways will be elevated, reducing access. Second, the Maglev vehicles will float 10–20 cm above the guideway, reducing hazards from small objects on the guideway. Third, track damage and wear will be negligible compared to high-speed rail, because the vehicles do not contact the guideway, and because the magnetic loadings are distributed under the vehicle rather than concentrated at wheels. The magnetic levitation system will use highly redundant and independent magnets. The probability of enough magnets failing that levitation is infinitesimally small—less than 10^{-10} per year. Even if levitation were to fail, a backup mechani-

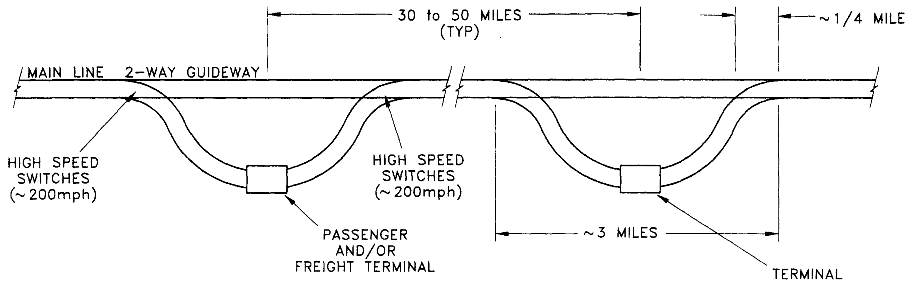

WHY OFF–LINE STATIONS?

- STATION SPACING WILL BE REALTIVELY CLOSE
  - MAXIMIZE PASSENGER/FREIGHT REVENUES
  - POLITICAL REASONS
- ON–LINE STATIONS SERIOUSLY DEGRADE AVERAGE SPEED
- MAXIMIZE GUIDEWAY TRAFFIC CAPACITY

WHAT ARE REQUIREMENTS FOR OFF–LINE CAPABILITY?

- HIGH SPEED SWITCHING (~200 mph)
- COMPUTER CONTROL OF TRAFFIC WITH PLAN–AHEAD CAPABILITY
- PRESET LOAD/UNLOAD INTERVALS FOR MAGLEV VEHICLES

FIG. 24. Schematic view of Maglev guideway system designed for high-speed switching to off-line stations.

cal skid system would take over and bring the vehicle to a safe stop. Vehicle levitation does not require external power but is inherently generated by motion. If the LSM propulsion were to cease, the levitated vehicle would simply coast until its speed dropped to approximately 30 km/h and then settle on the guideway. LSM propulsion inherently keeps fixed spacing between all vehicles regardless of local variations in force due to winds, grades, etc. Vehicles cannot gain on or collide with one another.

Maglev is pollution free and quiet, and has minimal impact on its surroundings, compared to other modes of ground transport. A two-way guideway, for example, can carry more passengers and freight than a 12-lane highway. Recently, there has been concern and dispute about the possible health effects of weak ac electromagnetic fields near power lines. It is important to distinguish between dc and ac magnetic fields. The fringing fields from the superconducting magnets on a Maglev vehicle are dc, not ac. So far, the health effects of weak dc magnetic fields have not been of concern. To ease potential concern of the riding public, the weak fringing fields in the passenger compartment can be reduced to the natural ambient magnetic field on Earth (approximately 0.5 G), using both active (distributed electrical currents) and passive (ferromagnetic materials) shielding. Without shielding, the fringing field is about 10 G. Such shielding would also screen out weak ac fields from the LSM propulsion system and fluctuations in the induced guideway currents. These fluctuations are caused by time variations of inductive coupling between the SC magnets and the guideway loops. Maglev can reduce fringe magnetic fields in the passenger compartment to whatever level is required by future standards. Weak magnetic fields are omnipresent in human society, however, and are found in subways, rail trains, work places, homes, city streets, etc. Minimizing weak magnetic fields in these areas would be much more difficult than in Maglev vehicles.

4.4 Broader Applications of Maglev

Maglev with its unique technical capabilities is attractive for easing greatly the increasing problems of suburban traffic congestion and pollution. Rapid acceleration and banking on curves is practical, limited only by passenger comfort. Lightweight economical guideway structures can be sited to coexist with the rights of way of existing transportation systems or along utility rights of way in a very flexible and adaptable manner. At the intermediate speeds required for urban transit, for example, less than 200 km/h, Maglev can be essentially noise free as well as locally pollution free.

The ability to electromagnetically switch safely and swiftly at full operating speeds without mechanical moving parts permits a high density of single vehicles to operate at high speed on a network with a large number of stations as required for convenient suburban transport. Highly flexible and individualized programming of even small numbers of passengers to a specific station can be carried out while the great majority of passengers pass by at full operating speed. In fact, this "skip-stop" fleet can include some smaller vehicles. At rush hours vehicles can be combined or "entrained" if required by ultrahigh volume to some stations.

The technical characteristics of Maglev enabling highly flexible and adaptable commuter service dovetail strongly with those that permit "land ferry" transport of freight and autos, conveniently interfaced with our existing highway-dominated transport system.

The authors expect that when Maglev comes fully of age it will have three major components:

1. 500-km/h, energy-efficient transport of people and freight over large distances;
2. urban and suburban Maglev transport with a large commuter component to complement existing networks; and
3. ultrahigh-speed transcontinental or intercontinental transport in evacuated tubes with extremely low energy consumption and with speeds of several thousand km/h.

GLOSSARY

Magnetic Spring: The magnetic restoring force that develops to push a vehicle back to its equilibrium suspension point if an external force acts to displace the vehicle from equilibrium.

Magnetic Well: A five-dimensional potential well in which the vehicle is held with the normal equilibrium suspension point at the bottom of the well. External energy input is required to make the vehicle climb up in its potential well. The five dimensions are vertical, lateral, pitch, yaw, and roll. The vehicle is unconstrained in the sixth dimension—that is, along the guideway.

Null Flux: An arrangement of loops in the guideway such that the *net* flux through the complete loop circuit is zero when the superconducting magnet on the vehicle is at the center of symmetry of the loop circuit. If the vehicle magnet moves away from the center of symmetry, a net flux and current develop in the circuit that act to push the vehicle magnet back to the center of symmetry.

Pole Pitch: The center-to-center distance between magnets of like polarity along the vehicle. The vehicle magnets alternate in polarity so that the pole pitch is twice the center-to-center distance between sequential magnets.

Superconducting Maglev: Magnetic levitation based on the repulsive force between superconducting magnets on a moving vehicle and the currents induced in conducting loops or sheets located on the guideway underneath.

Works Cited

Bachelet, E. (1912), "Foucault and Eddy Currents Put to Service," *Engineer* **114,** 420.

Bechtel-MIT, Inc. (1992), "Maglev System Concept Definition," Summary Report to National Maglev Initiative, Contact DTFR 53-92-C-00003.

Coffey, H., Chilton, F., Barbee, T. J. (1969), *Appl. Phys.* **40,** 2161.

Coffey, H. T., Colton, J. D., Mahrer, K. D. (1973), "Study of a Magnetically Levitated Vehicle," Stanford Research Institute Report FRA-RT-73-24.

Danby, G. T., Jackson, J. W., Powell, J. R. (1974), "Force Calculations for Hybrid (Ferro-Null-flux) Low-Drag Systems," *IEEE Trans. Magn.* **MAG-10,** 443–446.

Foster–Miller, Inc. (1993), "Maglev System Concept Definition: Executive Summary," DOT/FRA/ORD-92/01.

Graemiger, B. (1911), "Electromagnetic Suspension Devices," Austrian Patent 71,662.

Guderjahn, C. A., Wipf, S. L., Fink, H. J., Boom, R. W., MacKenzie, K. E., Williams, D., Downey, T. (1969), *J. Appl. Phys.* **40,** 2133–2140.

Kolm, H. H., Thornton, R. D. (1972), "Magneplane: Guided Electromagnetic Flight," in: *Proceedings of the 1972 Applied Superconductivity Conference,* New York: IEEE.

Kolm, H. H., Thornton, R. D. (1973), "Electromagnetic Flight," *Sci. Am.* **229** (4), 17–25.

Kyotani, Y. (1975), *Cryogenics* **15,** 372–376.

Maeda, T., *et al.* (1989), "Aerodynamic Drag of Shinkansen Electric Cars," Quarterly Report of RTRI (Railway Technical Research Institute) (Japan) Vol. 30 (1), p. 48.

Magneplane International (1992), "System Concept Definition Report," Summary Report to National Maglev Initiative.

Miller, L. (1987), "Transrapid 06 II—Performance and Characteristics," ibid, p. 155.

Miyata, S., Matsuda, K., Minemoto, M. (1993), "Status Quo of Superconducting Maglev System in Japan," in: *13th International Conference on Magnetically Levitated Systems and Linear Drives,* Argonne National Laboratory Report No. ANL/ES/Proc-79713, CONF-930550, pp. 7–9.

Polgren, G. R. (1966), *New Applications of Modern Magnets,* London: McDonald & Co.

Powell, J. R., Danby, G. T. (1967), *Mech. Eng.* **89,** 30.

Powell, J. R., Danby, G. T. (1969), "Magnetically Suspended Trains: The Application of Superconductors to High Speed Transport," *Cryogenics and Industrial Gases* **4,** 19.

Powell, J. R., Danby, G. T. (1970), "Dynamically Stable Cryogenic Magnetic Suspensions for Vehicles in Very High Velocity Transport Systems," in: *Recent Advances in Engineering Science,* Vol. 5, pt. 1, New York: Gordon and Breach, pp. 159–182.

Powell, J. R., Danby, G. T. (1971), *Cryogenics* **11,** 171.

Powell, J. R., Danby, G. T. (Eds.) (1989), "Benefits of Magnetically Levitated High Speed Transportation for the United States," Maglev Technology Advisory Committee (MTAC). Report to the U.S. Senate Committee on Environment and Public Works, Vol. 1, Executive Summary.

Powell, J. R., Danby, G. T. (Ed.) (1991), Technical Report to the U.S. Senate Committee on Environment and Public Works, Vol. 2, Bethpage, NY: Grumman Corp.

Proise, M., *et al.* (1993), "System Concept Definition of the Grumman Superconducting Electromagnetic (EMS) Maglev Design," in: *13th International Conference on Magnetically Levitated Systems and Linear Drives,* Argonne National Laboratory Report No. ANL/ES/Proc-79713, CONF-930550, p. 111.

Rather, J. (1992), "The New Millenium Transport System," personal communication.

Reitz, J. R., Davis, L. C. (1970), *Appl. Phys.* **41,** 2067.

Suzuki, S., Murai, M., Kawashima, M. (1987), "Operational Experience of HSST-03 System at Expo '85 and Expo '86," in: *Proceedings of the In-*

ternational Conference on Maglev and Linear Drives, New York: IEEE, p. 129.

Wiescholek, V., Mayer, W., Rogg, D. (1993), "High Speed Magnetic Levitation Train Transrapid," Planning of the Development Program until 1995 and Projects of Utilization in the Federal Republic of Germany, ibid, p. 22.

Williams, D. (1969), Report SC-DC-69-1611, Sandia Corporation.

Further Reading

Atherton, D. L., *et al.* (1977), The Canadian High Speed Magnetically Levitated Vehicle System, Canadian Institute of Guided Ground Transport Report 77-12, Queens University, Kingston, Ontario, Canada.

Atherton, D. L., Belanger, P. R., Burke, P. E., Dawson, G. E., Eastham, A. R., Hayes, W. F., Ooi, B. T., Silvester, P., Slemon, G. R. (1978), "The Canadian High Speed Magnetically Levitated System," *Can. Elec. Eng. J.* **3** (2), 3–25.

Federal Railroad Administration (1990), "Assessment of the Potential for Magnetic Levitation Transportation Systems in the United States," Report Supplement to Congress.

Foner, S., Schwartz, B. (Eds.) (1973), *Superconducting Machines and Devices,* New York and London: Plenum Press.

IEEE (Eds.) (1993), *13th International Conference on Magnetically Levitated Systems and Linear Drives,* Argonne National Laboratory Report No. ANL/ES/Proc-79713, CONF-930550.

Johnson, L. R., Rote, D. M., Hull, J. R., Coffey, H. T., Daley, J. G., Giese, R. F. (1989), "Maglev Vehicle and Superconductor Technology: Integration of High Speed Ground Transportation into the Air Travel System," ANL/CNSV-67.

Kolm, H. H., Thornton, R. D. (1973), "Electromagnetic Flight," *Sci. Am.* **229** (4), 17–25.

Rhodes, R. G., Mulhall, B. E. (1981), *Magnetic Levitation for Rail Transport,* New York: Oxford Univ. Press.

TRANSPORT PHENOMENA

HENRIK SMITH, *Ørsted Laboratory, University of Copenhagen, Copenhagen, Denmark*

Introduction 263
1. **The Boltzmann Equation** 264
1.1 The Collision Integral 265
1.2 The *H* Theorem 265
1.3 Conservation Laws 266
1.4 Linearization 267
1.5 Methods of Solution 268
1.6 The Relaxation-Time Approximation 268
1.6.1 Thermal Conductivity 269
1.6.2 Viscosity 269
1.6.3 Electrons in Metals 269
1.7 Mean Free Paths 270
2. **Classical Gases** 270
2.1 Transport Coefficients 271
2.1.1 Heat Conductivity 271
2.1.2 Viscosity 271
2.1.3 Quantum Effects 272
2.2 Knudsen Flow 272
3. **Plasmas** 273
3.1 Collective Oscillations 273
3.2 Dielectric Screening 275
4. **Metals and Semiconductors** 276
4.1 Normal Metals 276
4.2 Superconductors 277
4.3 Semiconductors 277
5. **Insulators** 278
5.1 Heat Conductivity 278
5.2 Second Sound 279
6. **Quantum Fluids** 279
6.1 Fermi Liquids 279
6.1.1 The Landau–Boltzmann Equation 279
7. **The Validity of the Boltzmann Equation** 280
Glossary 281
Works Cited 281
Further Reading 281

INTRODUCTION

Transport phenomena pervade all of physics. A major part of the current research on metals and semiconductors involves the conduction of electricity under a variety of conditions. The theoretical description of transport processes in condensed matter is carried out at many different levels of sophistication, ranging from the use of pure phenomenology to the application of quantum-field-theoretical methods. Modern transport theory is rooted in the kinetic theory of the nineteenth century, founded by Clausius, Maxwell, Boltzmann, and others. Historically, kinetic theory deals with nonequilibrium phenomena in dilute gases, in which the range of interaction between the molecules is much less than their average separation. Then the interaction between the particles may be considered to be essentially instantaneous, with the particle motion consisting of straight paths that change direction as a result of each instantaneous scattering process. An external field may cause the paths to be curved, but this does not change the fundamental separation between free-particle motion and collision.

In this article we discuss the use of the Boltzmann equation for describing transport phenomena such as electrical or thermal conduction, including collective motion such as sound or plasma oscillations. The theory is semiclassical in the sense that it involves a single-particle distribution function that depends on both position and momentum variables. Despite this limitation the Boltzmann approach to transport has been very fruitful. It can account for a variety of scattering processes, including those that involve emission or absorption of other particles such as electron–phonon scattering in metals and semiconductors. The transport of neutrons in nuclear reactors is described by a simplified Boltzmann equation, obtained by exploiting the fact that the neutron–neutron collision rate is much smaller than the rate of neutrons colliding with other particles. Strongly interacting particles or systems exhibiting a

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

condensate—such as electrons in superconductors—may also be described by a Boltzmann equation, suitably modified to allow for the effects of interaction.

The first section discusses the derivation of the Boltzmann equation and summarizes its most important properties. It describes the relaxation-time approximation and different methods for obtaining solutions to the general integro-differential equation. The following sections discuss the use of the kinetic equation for different physical systems. Since these include plasmas, metals, semiconductors, and superconductors, it is of course impossible to discuss their transport properties in detail. Instead, the attention is confined to the important features that must be considered when the Boltzmann equation is applied to systems other than the classical gas for which it was invented.

1. THE BOLTZMANN EQUATION

For a classical gas of particles the distribution function $f(\mathbf{r},\mathbf{p},t)$ is defined such that $f(\mathbf{r},\mathbf{p},t)d^3r\, d^3p/(2\pi\hbar)^3$ is the mean number of particles at time t in the phase-space element $d^3r\, d^3p$ of the six-dimensional phase space of a single particle, with $\mathbf{r}$ being the position vector of the particle and $\mathbf{p}$ its canonical momentum. The factor $(2\pi\hbar)^3$ serves to make f dimensionless, equal to the mean occupation number per quantum state for particles in a definite internal state.

The Boltzmann equation is a continuity equation with source terms in the six-dimensional phase space. In its general form it is an integro-differential equation for the distribution function f. If collisions are neglected, the rate of increase of f in a given phase-space volume element must equal the net number of particles that stream from outside to inside this phase-space volume element. This implies the continuity equation

$$\frac{\partial f}{\partial t} + \frac{\partial}{\partial x_\mu}(v_\mu f) = 0. \tag{1}$$

Here $v_\mu = (\dot{\mathbf{r}},\dot{\mathbf{p}}) = (\mathbf{v},\dot{\mathbf{p}})$ is the six-dimensional velocity, the dot denoting differentiation with respect to time, while the six-dimensional gradient operator is $\partial/\partial x_\mu = (\partial/\partial\mathbf{r}, \partial/\partial\mathbf{p})$. The balance expressed by (1) neglects, however, the effects of collisions, which are considered to be instantaneous. Such collisions move particles from other regions of the phase space *into* the one under consideration, or they move particles *away* from this region. It is therefore necessary to add a source term to the right-hand side of (1):

$$\frac{\partial f}{\partial t} + \frac{\partial}{\partial x_\mu}(v_\mu f) = \left(\frac{\partial f}{\partial t}\right)_{\text{coll}}. \tag{2}$$

The source term represents the rate of change of the distribution function due to collisions. Equation (2) is a generalized conservation law. It expresses that the number of particles in a given phase-space volume element changes

1. as a result of the motion of the particles under the influence of the fields and
2. as a result of collisions between the particles.

The left-hand side of (2) may be simplified, when the motion of the particles is determined by Hamilton's equations

$$\dot{\mathbf{r}} = \frac{\partial H}{\partial \mathbf{p}}, \quad \dot{\mathbf{p}} = -\frac{\partial H}{\partial \mathbf{r}}, \tag{3}$$

which imply that $\partial v_\mu/\partial x_\mu = 0$. This means that (2) may be written in the simpler form

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\frac{\partial f}{\partial \mathbf{r}} + \dot{\mathbf{p}}\cdot\frac{\partial f}{\partial \mathbf{p}} = \left(\frac{\partial f}{\partial t}\right)_{\text{coll}}. \tag{4}$$

The momentum $\mathbf{p}$ in (4) is the canonical momentum $\mathbf{p}_c$. In the presence of a magnetic field, it is convenient to transform to $(\mathbf{r},\mathbf{p}_k)$ as phase-space coordinates instead of $(\mathbf{r},\mathbf{p}_c)$, where $\mathbf{p}_k$ is the kinetic momentum given by $\mathbf{p}_k = \mathbf{p}_c - q\mathbf{A}$, with q being the charge of the particle. The reason is that the kinetic momentum is related in a simple way to the force according to

$$\dot{\mathbf{p}}_k = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \tag{5}$$

where $\mathbf{B} = \nabla \times \mathbf{A}$. Now (4) is readily seen to be form-invariant under this transformation: For fixed $\mathbf{r}$ and t one has $d^3p_c = d^3p_k$, which implies that the distribution function is invariant. Furthermore, the left-hand side of (4) is the substantial derivative df/dt, which may be expressed in either set of coordinates. With $\mathbf{p}$ interpreted as $\mathbf{p}_k$, we can thus use (4) together with the equation of motion (5).

1.1 The Collision Integral

The form of the collision term $(\partial f/\partial t)_{\text{coll}}$ depends on which particular scattering processes are being considered. We shall here discuss two specific examples:

1. elastic scattering of particles from fixed impurities and
2. collisions between the particles themselves.

For the following, it is convenient to introduce the abbreviation $d\tau = d^3p/(2\pi\hbar)^3$.

1. In the case of elastic impurity scattering the collision integral is determined by processes that move particles from the phase-space element $d\tau$ to the phase-space element $d\tau'$ as well as the inverse processes that move particles from $d\tau'$ to $d\tau$. Let us first consider scattering processes from $d\tau$ to $d\tau'$: The statistical description of the effect of collisions implies that the rate of change of $f(\mathbf{p})$ due to such processes is proportional to $f(\mathbf{p})$ as well as $d\tau'$. The rate of change is therefore written as $-d\tau' w(\mathbf{p}',\mathbf{p})f(\mathbf{p})$, where $d\tau' w(\mathbf{p}',\mathbf{p})dt$ is the probability of scattering of a particle with momentum $\mathbf{p}$ to the element $d\tau'$ during the time interval dt. With a similar expression for the rate of change of $f(\mathbf{p})$ due to the inverse process, scattering from $d\tau'$ to $\mathbf{p}$, the collision integral becomes

$$\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} = \int d\tau'[w(\mathbf{p},\mathbf{p}')f(\mathbf{p}') - w(\mathbf{p}',\mathbf{p})f(\mathbf{p})]$$
$$= -\int d\tau' w(\mathbf{p},\mathbf{p}')[f(\mathbf{p}) - f(\mathbf{p}')]. \qquad (6)$$

The second equality in (6) holds true when the collision probability is invariant under time reversal and spatial inversion, which together imply the symmetry $w(\mathbf{p}',\mathbf{p}) = w(\mathbf{p},\mathbf{p}')$.

2. Next we consider the case of particle–particle scattering. The probability that a particle of momentum $\mathbf{p}$ will be scattered during the time interval dt by a particle in the element $d\tau_1$, resulting in the particles ending up in $d\tau'$ and $d\tau_1'$, is $d\tau_1 d\tau' d\tau_1' \times w(\mathbf{p}',\mathbf{p}_1';\mathbf{p},\mathbf{p}_1)f(\mathbf{p}_1)dt$, since it is proportional to the number density of particles in $d\tau_1$, $f(\mathbf{p}_1)d\tau_1$. When the reverse process is taken into account, the collision integral is seen to be

$$\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} = \int d\tau_1 d\tau' d\tau_1' w(\mathbf{p},\mathbf{p}_1;\mathbf{p}',\mathbf{p}_1')$$
$$\times [f(\mathbf{p}')f(\mathbf{p}_1') - f(\mathbf{p})f(\mathbf{p}_1)], \qquad (7)$$

provided the collision probability is invariant under time reversal and spatial inversion.

It is clear that the Boltzmann approach is not limited to these two examples but applies equally well to systems containing different kinds of particles such as two-component mixtures. One important example is the transport of neutrons in a reactor, the second component being the atoms of the moderator. Since the number density of neutrons—as well as the range of their mutual interaction—is many orders of magnitude smaller than that of the atoms, the collisions between the neutrons may be neglected and the atoms assumed to be in equilibrium. This results in a collision integral of a form similar to (6), as discussed in detail by Cercignani (1988).

Our discussion has so far relied on the properties of a dilute classical gas, in which the range of interaction is much less than the interparticle distance. It is possible to generalize the classical Boltzmann equation in a number of ways. Apart from changes from Maxwell–Boltzmann (equilibrium) distributions to Fermi–Dirac or Bose–Einstein distributions, one may consider collision processes that involve emission or absorption of other elementary excitations such as phonons or magnons. In superfluids and superconductors the collision integral contains terms that correspond to the creation or annihilation of particles from a condensate. The validity of the Boltzmann equation is not restricted to dilute systems, but it may be applied to the elementary excitations in a dense and strongly interacting system.

1.2 The *H* Theorem

The development in time of the distribution function $f(\mathbf{r},\mathbf{p},t)$ is governed by the Boltzmann equation. The famous H theorem, derived by Boltzmann, demonstrates that this development in time is consistent with the law of increasing entropy for a closed macroscopic system of particles that is left to itself in a nonequilibrium state.

Let us consider the quantity $\phi(\mathbf{p},f)$, which may be any function of the momentum variable $\mathbf{p}$ and the distribution function $f(\mathbf{r},\mathbf{p},t)$. We may associate with it a space- and time-dependent density $\rho_\phi(\mathbf{r},t)$ by the definition

$$\rho_\phi(\mathbf{r},t) = \int d\tau \phi(\mathbf{p},f) f(\mathbf{r},\mathbf{p},t). \tag{8}$$

It is then a straightforward consequence of the form of the collision integral (7) as well as the symmetries obeyed by w that the rate of change of ρ_ϕ due to collisions is given by

$$\left(\frac{\partial \rho_\phi}{\partial t}\right)_{\text{coll}} = -\tfrac{1}{4}\int d\tau d\tau_1 d\tau' d\tau_1' w(\mathbf{p},\mathbf{p}_1;\mathbf{p}',\mathbf{p}_1') \times [\bar{\phi} + \bar{\phi}_1 - \bar{\phi}' - \bar{\phi}_1'](ff_1 - f'f_1'), \tag{9}$$

where $\bar{\phi} = \phi + f\partial\phi/\partial f$. Let us choose ϕ to be the natural logarithm of the distribution function, $\phi = \ln f$. Then we have $\bar{\phi} = \phi + 1$ and

$$\left(\frac{\partial}{\partial t}\int d\tau f \ln f\right)_{\text{coll}} = -\tfrac{1}{4}\int d\tau d\tau_1 d\tau' d\tau_1' w(\mathbf{p},\mathbf{p}_1;\mathbf{p}',\mathbf{p}_1') \times (ff_1 - f'f_1') \ln \frac{ff_1}{f'f_1'}. \tag{10}$$

Since w is positive and $(x - y)\ln(x/y)$ is positive or zero, the right-hand side of (10) is never positive. The quantity $\int d\tau f \ln f$ is named H, and the result (10) establishes the H theorem: that the H production due to collisions is never positive.

In order to elucidate the physical significance of this result, let us consider the entropy in a general nonequilibrium state specified by the distribution function $f(\mathbf{r},\mathbf{p},t)$. The entropy density s is obtained as the Boltzmann constant k times the logarithm of the statistical weight or

$$s = -k\int d\tau [f\ln f \mp (1 \pm f)\ln(1 \pm f)], \tag{11}$$

where the upper (lower) sign refers to bosons (fermions). The expression (11) for the entropy density follows from combinatorial arguments and hence applies to equilibrium as well as nonequilibrium situations. For a dilute gas, $f \ll 1$, and one obtains

$$s = k\int d\tau (f - f\ln f). \tag{12}$$

This shows that H equals minus the entropy density divided by the Boltzmann constant, apart from the first term in (12) involving the number density of particles, $\int d\tau f$. When the entropy density is integrated over the entire volume of the system, the first term in (12) yields k times the number of particles, which is constant and may therefore be disregarded. This proves the H theorem, and one concludes that the statistical description of the collision processes implies that the collisions cannot reduce the total entropy.

1.3 Conservation Laws

We shall now turn to a discussion of the significance of the conservation laws that apply to particle–particle scattering. In (9) we have obtained a general expression for the rate of change of the density ρ_ϕ due to collisions. If ϕ does not depend on the distribution function, we have $\bar{\phi} = \phi$. It then follows from (9) that $\partial\rho_\phi/\partial t$ is zero, if

$$\phi + \phi_1 = \phi' + \phi_1'. \tag{13}$$

A quantity satisfying (13) is called a collision invariant. Since momentum is conserved in a collision between two particles, $\phi = p_x, p_y, p_z$ are three collision invariants. In addition the kinetic energy $E = p^2/2m$ is conserved as well as the particle number. Thus $\mathbf{p}$, E, and 1 are the five collision invariants, corresponding to the conservation of momentum, energy, and particle number. The angular momentum is also a collision invariant, but this is important only for rotating systems.

The identification of the collision invariants allows one to determine the equilibrium distribution function, since in equilibrium the entropy production must be zero. This means that the right-hand side of (10) must vanish, yielding the condition $ff_1 = f'f_1'$. Therefore,

$$\ln f + \ln f_1 = \ln f' + \ln f_1', \tag{14}$$

which shows that $\ln f$ belongs to the class of functions satisfying (13). It follows that ln f must be a linear combination of the collision invariants, that is,

$$\ln f = A + \mathbf{B}\cdot\mathbf{p} + CE, \qquad (15)$$

where A, $\mathbf{B}$, and C are constants. The values of these constants are known from equilibrium statistical mechanics, $A = \mu/kT$, $\mathbf{B} = \mathbf{u}/kT$, $C = -1/kT$, where μ is the chemical potential, T the temperature, and $\mathbf{u}$ the drift velocity. We have thus established the form of the equilibrium distribution function,

$$f = \exp\{(-E + \mathbf{u}\cdot\mathbf{p} + \mu)/kT\}. \qquad (16)$$

In the fermion case, when f is not small, (16) is replaced by

$$f(1 - f) = \exp\{(-E + \mathbf{u}\cdot\mathbf{p} + \mu)/kT\}, \qquad (17)$$

which yields the Fermi–Dirac equilibrium distribution

$$f = \frac{1}{e^{(E-\mathbf{u}\cdot\mathbf{p}-\mu)/kT} + 1}. \qquad (18)$$

1.4 Linearization

In its general form the Boltzmann equation is a nonlinear, integro-differential equation. In many cases, it is sufficient to consider a linearized version, when the deviation from a local equilibrium distribution function $f^0(\mathbf{p})$ is small.

Let us consider a homogeneous system, in which the distribution function $f(\mathbf{p})$ decays toward its equilibrium value $f^0(\mathbf{p})$. In the absence of external forces, f is changed by collisions only. The Boltzmann equation is linearized by writing $f = f^0 + f^0\psi$ and keeping only terms of order ψ. One gets

$$\frac{\partial\psi}{\partial t} = L\psi, \qquad (19)$$

where

$$L\psi = -\int d\tau_1 d\tau' d\tau_1' w(\mathbf{p},\mathbf{p}_1;\mathbf{p}',\mathbf{p}_1') \times [\psi + \psi_1 - \psi' - \psi_1'] f^0(\mathbf{p}_1). \qquad (20)$$

We have used the equilibrium condition $f^0(\mathbf{p})f^0(\mathbf{p}_1) = f^0(\mathbf{p}')f^0(\mathbf{p}_1')$. The linear differential equation (19) admits solutions of the form

$$\psi = e^{-\lambda t}g_\lambda(\mathbf{p}), \qquad Lg_\lambda = -\lambda g_\lambda. \qquad (21)$$

The constant λ determines the decay in time associated with the eigenfunction g_λ of the L operator. Because of the symmetry relations obeyed by w under interchange of its momentum variables, the linear operator L is symmetric in the sense that

$$(g_1,Lg_2) = (g_2,Lg_1). \qquad (22)$$

Here we have defined a scalar product by

$$(g_1,g_2) = \int d\tau f^0(\mathbf{p})g_1(\mathbf{p})g_2(\mathbf{p}). \qquad (23)$$

The operator L is negative semidefinite, which means that its eigenvalues are all negative or zero. This is seen by writing $(g_\lambda,Lg_\lambda) = -\lambda(g_\lambda,g_\lambda)$ and using the symmetries and positivity of w.

It is often of interest to linearize the Boltzmann equation in an external field, a temperature gradient, or the like, in order to determine the linear response of the system to such a disturbance. One then linearizes the Boltzmann equation with respect to the "driving gradient," which could be the temperature gradient or a spatial derivative of the drift velocity. The function ψ is assumed to be linear in the driving gradient, and the left-hand side of the Boltzmann equation may then be approximated by replacing the distribution function f with the local equilibrium distribution f^0 and expanding it to first order in the gradients. This results in a Boltzmann equation that is an inhomogeneous, linear integral equation.

As an example of this linearization procedure let us consider heat conduction in a dilute classical gas due to the presence of a temperature gradient ∇T, assuming the absence of convection ($\mathbf{u} = 0$). In the stationary case the left-hand side of the Boltzmann equation becomes

$$\mathbf{v}\cdot\frac{\partial f^0}{\partial\mathbf{r}} = -\frac{\partial f^0}{\partial E}\mathbf{v}\cdot\left(\frac{E-\mu}{T}\nabla T + \nabla\mu\right). \qquad (24)$$

In the convectionless case under consideration the pressure is independent of position. The gradient of the chemical potential is therefore proportional to the temperature gradient [$\nabla\mu = (\partial\mu/\partial T)_p \nabla T$], and (24) takes the form

$$\mathbf{v}\cdot\frac{\partial f^0}{\partial \mathbf{r}} = -\frac{\partial f^0}{\partial E}\mathbf{v}\cdot\nabla T\frac{E-h}{T}, \tag{25}$$

where h is the enthalpy per particle, given by

$$h = \mu - T\left(\frac{\partial\mu}{\partial T}\right)_p. \tag{26}$$

We shall use this linearized form of the driving term below to calculate the heat conductivity of a classical gas.

1.5 Methods of Solution

Since the linear operator L is symmetric and negative semidefinite, it is possible to derive a variational principle that applies to the linearized Boltzmann equation. Let us define the positive semidefinite operator H by

$$L = -H. \tag{27}$$

The linear, inhomogeneous Boltzmann equation has thus the form

$$X = H\psi, \tag{28}$$

where X is the inhomogeneous driving term. In the case of electrical conduction, X will be proportional to the component of the particle velocity along the electric field. The reason is that the $\dot{\mathbf{p}}\cdot\partial f/\partial\mathbf{p}$ term in (4), when linearized in the driving force $q\mathbf{E}$, becomes $q\mathbf{E}\cdot\partial f^0/\partial\mathbf{p} = q(\mathbf{E}\cdot\partial\epsilon/\partial\mathbf{p})\partial f^0/\partial\epsilon$. The current density is given by an integral over $\mathbf{p}$ of the velocity $\partial\epsilon/\partial\mathbf{p}$ weighted by the nonequilibrium distribution $f - f^0$, and the current density (and hence the conductivity) is therefore proportional to the scalar product (X,ψ). This is a common feature, which is not limited to electrical conduction: the transport coefficient may in general be related to the scalar product (X,ψ). We now derive a bound on (X,ψ) using the Schwarz inequality

$$(\psi,H\psi)(U,HU) \geq (U,H\psi)^2, \tag{29}$$

where U is an arbitrary trial function. Since $X = H\psi$, it follows that

$$(X,\psi) \geq \frac{(U,X)^2}{(U,HU)}. \tag{30}$$

One often uses the driving term X itself as a trial function. Then $(X,X)^2/(X,HX)$ becomes a lower bound on the conductivity. The determination of the lower bound may be improved by choosing U as a linear combination of a particular set of basis functions and solving the resulting matrix equation for the coefficients.

It is also possible to solve the integral equation directly, by the use of numerical methods. If the integral equation effectively involves only one variable, such numerical calculations may be relatively simple. If account is taken of the complexities of a real metal, such as anisotropy of the Fermi surface and phonon spectra, a straightforward integration may become prohibitively difficult. It should also be emphasized that the complexities of a real metal are in general not known well enough to justify the effort involved in obtaining detailed numerical solutions to the Boltzmann equation, starting from the complete collision integral.

In a few cases it is possible to determine the eigenfunctions of the collision operator and use the method of eigenfunction expansion to obtain exact solutions. Provided the eigenfunctions of L form a complete, orthonormal set, we can write down the general solution to the Boltzmann equation (19) in the form

$$\psi = \sum_i A_i g_i(\mathbf{p}) e^{-\lambda_i t}. \tag{31}$$

If a particular eigenvalue λ_i is zero, corresponding to a collision invariant, the coefficient A_i in the expansion must be zero, since ψ is assumed to vanish as $t \to \infty$. If the eigenvalues and eigenfunctions are known, the transport coefficient may then be obtained directly. This is true only in special cases. A famous example is Maxwell's r^{-5} molecules, that is, molecules repelling each other by a force inversely proportional to the fifth power of their mutual distance.

1.6 The Relaxation-Time Approximation

Since the linearized Boltzmann equation is an inhomogeneous integral equation, which may be difficult to solve analytically or numerically, it may be useful to treat the collision operator in a simple relaxation-time approximation. This amounts to replacing the spectrum of different relaxation times by a single time τ. The resulting equation has the form

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\frac{\partial f}{\partial \mathbf{r}} + \dot{\mathbf{p}}\cdot\frac{\partial f}{\partial \mathbf{p}} = -\frac{f - f^0}{\tau}. \tag{32}$$

One usually assumes that τ is independent of $\mathbf{p}$, while the local equilibrium function f^0 may depend on $\mathbf{r}$ and t. Below we give two examples of the use of the relaxation-time approximation.

1.6.1 Thermal Conductivity A stationary temperature gradient ∇T in a gas of particles causes a heat current density given by $\mathbf{j}_{\text{th}} = -\kappa\nabla T$, where κ is the thermal conductivity. The thermal conductivity is obtained by solving the linearized Boltzmann equation. The driving term was obtained for a classical gas in (25) above. Treating the collisions in the relaxation-time approximation, we thus obtain

$$-\frac{\partial f^0}{\partial E}\frac{E - h}{T}\mathbf{v}\cdot\nabla T = -\frac{f - f^0}{\tau}. \tag{33}$$

The heat-current density is identical to the energy-current density given by

$$\mathbf{j}_\epsilon = \int d\tau E\mathbf{v}f, \tag{34}$$

since we are considering the case when there is no convection and hence no macroscopic motion of the gas. For a monatomic gas one obtains, after carrying out the integration over energy and angle in momentum space,

$$\kappa = nc_p\tau\frac{kT}{m}, \tag{35}$$

where n is the particle density and $c_p = 5k/2 = h/T$ is the specific heat per particle at constant pressure.

The use of the relaxation-time approximation is not, of course, limited to classical gases. For a degenerate gas of fermions, one finds that (35) is replaced by

$$\kappa = \tfrac{1}{3}nc_V v_{\text{F}}^2\tau, \tag{36}$$

where the Fermi velocity v_{F} is related to the Fermi momentum—and hence the particle density—by $v_{\text{F}} = \hbar k_{\text{F}}/m$, with $n = k_{\text{F}}^3/3\pi^2$, while c_V is the specific heat per particle at constant volume.

1.6.2 Viscosity According to hydrodynamics, a spatially varying flow velocity $u_x(y)$ causes a momentum-current density Π_{xy} given by $\Pi_{xy} = -\eta\partial u_x/\partial y$, where η is the (shear) viscosity. The viscosity of a classical gas is determined by linearizing the Boltzmann equation in terms of the spatial derivative of the flow velocity. With the flow velocity along the x axis and depending only on y, the Boltzmann equation becomes

$$-p_x v_y\frac{\partial f^0}{\partial E}\frac{\partial u_x}{\partial y} = -\frac{f - f^0}{\tau}. \tag{37}$$

From this the momentum current density Π_{xy} and hence the shear viscosity are obtained as

$$\Pi_{xy} = \int d\tau p_x v_y f = -\int d\tau\frac{v_x^2 p_y^2}{kT}f^0\tau\frac{\partial u_x}{\partial_y}, \tag{38}$$

resulting in the viscosity

$$\eta = n\tau kT. \tag{39}$$

Note that $\eta/nm = \tau kT/m$, while $\kappa/nc_p = \tau kT/m$. This means that the Prandtl number, which is defined as $N_{\text{Pr}} = \eta c_p/m\kappa$, is 1 within the relaxation-time approximation, contrary to the measured values which for monatomic gases are very close to $\frac{2}{3}$.

For a degenerate gas of fermions one finds that (39) is replaced by

$$\eta = \tfrac{1}{5}nmv_{\text{F}}^2\tau, \tag{40}$$

which shows that the temperature T in the classical case is replaced by a numerical factor times the Fermi temperature $mv_{\text{F}}^2/2k$.

1.6.3 Electrons in Metals The relaxation-time approximation is widely used in metals and semiconductors. Its formal validity is, however, restricted to situations where

1. the energy bands are spherical,
2. the scattering is elastic, and
3. the collision probability only depends on

the angle between the incoming and outgoing momentum.

As an example we mention the case when the electrons in a metal are scattered by fixed impurities. The linearized Boltzmann equation for electrical conduction then becomes [with $g(\mathbf{k}) = f - f^0$]

$$-e\mathbf{E}\cdot\mathbf{v}\frac{\partial f^0}{\partial \epsilon} = -\int \frac{d\mathbf{k}'}{(2\pi)^3} w(\mathbf{k},\mathbf{k}')[g(\mathbf{k}) - g(\mathbf{k}')]. \tag{41}$$

Here the energy of an electron is $\epsilon(\mathbf{k})$, and its velocity $\mathbf{v} = (1/\hbar)\partial\epsilon(\mathbf{k})/\partial\mathbf{k}$.

Under isotropic conditions it is possible to replace the right-hand side of (41) by a simple relaxation-time expression for the collision integral by virtue of the identity

$$\cos\theta_{k'} = \cos\theta \cos\theta_k + \sin\theta \sin\theta_k \cos(\phi_k - \phi), \tag{42}$$

where ϕ and θ are the polar angles of $\mathbf{k}'$ relative to the direction of $\mathbf{k}$, while ϕ_k and θ_k denote polar angles relative to the direction of $\mathbf{E}$. This is a consequence of (41) having its solution $g(\mathbf{k})$ proportional to $\cos\theta_k$, as may be verified by insertion and use of (42): since the term involving $\cos(\phi_k - \phi)$ vanishes upon integration over ϕ (as w is assumed to depend only on θ, not on ϕ), one has

$$\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} = -\frac{1}{\tau}(f - f^0), \tag{43}$$

where the relaxation time is given by

$$\frac{1}{\tau} = \int \frac{d\mathbf{k}'}{(2\pi)^3} w(\mathbf{k},\mathbf{k}')(1 - \cos\theta). \tag{44}$$

The appearance of $\cos\theta$ in this expression for τ derives from the fact that the driving term in the case of electrical (or thermal) conduction involves the component of the particle velocity along the electric field (or temperature gradient).

Despite the restrictions mentioned above on the validity of the relaxation-time approximation, it remains widely used. This is due to its simplicity and the physical insight it provides under conditions when one is not interested in the accurate value of a given transport coefficient or uses τ as a fitting parameter entering other physical quantities. The use of the relaxation-time approximation may also be justified in cases when the departure from the equilibrium distribution is localized to a small region of **k** space.

1.7 Mean Free Paths

The *mean free path* is a measure of the average distance that the particles travel between consecutive collisions. It is obtained from the relaxation time by multiplying τ by a characteristic velocity v of the carriers, $l = v\tau$. In the classical gas v is a thermal velocity of order $(kT/m)^{1/2}$, while in the degenerate case v should be identified with the Fermi velocity.

The mean free path may be estimated from the relation $l \simeq 1/n\sigma$, where n is the number density of scatterers, while σ is the total scattering cross section for a single scatterer. For a gas of hard spheres with diameter d one finds $\sigma = \pi d^2$ and hence $\tau \simeq 1/vnd^2$. Thus τ depends on temperature according to $\tau \propto T^{-1/2}$, while the mean free path is temperature independent.

Different transport properties are generally characterized by different mean free paths or relaxation times. One may write a transport coefficient such as the electrical conductivity σ of a metal in the form suggested by the Drude model, $\sigma = ne^2\tau_\sigma/m$, while the thermal conductivity may be written as $\kappa = nc_V v_F^2 \tau_\kappa/3$. The two relaxation times τ_σ and τ_κ need not, however, be equal in magnitude and may even differ in their temperature dependence, as in the case of electrons scattering against phonons. For a simple Debye model with spherical energy bands one finds $\tau_\sigma \propto T^{-5}$, while $\tau_\kappa \propto T^{-3}$, when T is much less than the Debye temperature. The physical reason for this difference is that small-angle scattering, caused by long-wavelength phonons, is much less effective in degrading an electrical than a thermal current.

2. CLASSICAL GASES

Kinetic theory was developed in the latter half of the 19th century with the aim of explaining the observed magnitude and temperature dependence of transport coefficients such as viscosity or thermal conductivity of air. The contributions by Maxwell and Boltzmann laid the foundation for the kinetic theory of gases and its generalizations to other dilute systems. Via solutions to the appropriate Boltzmann equation (obtained by variational or other methods), the calculated transport coefficients are related directly to the intermolecular potential. A transport coefficient such as the shear viscosity of monatomic gases may be measured and calculated with great accuracy, resulting in excellent agreement between experimental and theoretical values, as discussed below and—in much more detail—by Smith and Jensen (1989).

When the mean free path becomes comparable to or larger than other dimensions characterizing the flow, such as the diameter of a tube or a hole in a container, one enters a regime in which hydrodynamics no longer holds. Kinetic theory, however, retains its validity, but the solution of the Boltzmann equation must incorporate boundary conditions that describe how the particles are scattered by the wall.

2.1 Transport Coefficients

In the following we consider only monatomic gases, on account of the simplicity of their interatomic potential U, which may be quite accurately represented by the Lennard–Jones potential

$$U(r) = 4\epsilon[(\sigma/r)^{12} - (\sigma/r)^{6}], \tag{45}$$

with ϵ and σ positive constants. For all but the lightest atoms, it is possible with such a two-parameter model to obtain good fits to the magnitude and temperature dependence of the transport coefficients, using values of the parameters ϵ and σ that are consistent with other experiments.

2.1.1 Heat Conductivity In the previous section we found the thermal conductivity within the relaxation-time approximation to be $\kappa = nc_pkT\tau/m$, while the corresponding value of the shear viscosity was found to be $\eta = nkT\tau$. This does not imply that these transport coefficients are proportional to the absolute temperature, since τ is an undetermined parameter. Furthermore the measured Prandtl number $N_{\mathrm{Pr}} = \eta c_p/m\kappa$ differs considerably from the value 1 predicted by the relaxation-time approximation. It is therefore necessary to abandon the relaxation-time approximation and return to the linearized integral equation.

Let us for simplicity assume that the molecules may be represented by hard spheres of diameter d. It is then not difficult to show, using as variational solution to the Boltzmann equation a trial function proportional to X [cf. (30) above], that κ is given by the lower bound

$$\kappa \geq \frac{75k}{64d^2}\sqrt{\frac{kT}{m\pi}}. \tag{46}$$

While (46) only constitutes a lower bound on κ, subsequent improvements of the trial function (obtained by linear combinations that include the second, third, and fourth eigenfunctions belonging to the exactly solvable case of "Maxwellian molecules" with interatomic forces proportional to r^{-5}, r being their mutual separation) yield only relative corrections $\Delta\kappa/\kappa$ of order 2.3×10^{-2}, 2.1×10^{-3}, and 3×10^{-4}, respectively. The procedure thus converges very rapidly, and the inequality in (46) may be considered an equality with an accuracy of 2–3%.

While it is simpler to carry out the calculation of the thermal conductivity for an interatomic potential corresponding to hard spheres, the use of the Lennard–Jones expression (45) involves no conceptual changes, only somewhat more laborious expressions. It is a general feature of the variational method of solution that the result obtained from the simplest trial function proportional to the driving term X is quite accurate, the reason being that a first-order error in the choice of trial function yields a second-order error in the calculated value of the transport coefficient.

2.1.2 Viscosity Let us now calculate the viscosity η within the same model where the molecules are represented by hard spheres of diameter d. With a trial function proportional to the driving term X appropriate to viscosity, one finds

$$\eta \geq \frac{5}{16d^2}\sqrt{\frac{mkT}{\pi}}. \tag{47}$$

As before, (47) constitutes a lower bound on η, but subsequent improvements of the trial function yield again only corrections at the 1% level. The Prandtl number $\eta c_p/m\kappa$ is seen from (46) and (47) to be $\frac{2}{3}$, in excellent agreement with the measured value for the noble gases, which is found to be 0.666 for both Ne and Ar at room temperature.

The value of $\frac{2}{3}$ for the Prandtl number is always obtained within the first variational approximation, with a trial function proportional to X, irrespective of the form of the interatomic potential. Its agreement with experiment is therefore not a stringent test of the model potential or the validity of kinetic theory. To obtain this one must compare the calculated temperature dependence of the transport coefficients with the measured one. The temperature dependence predicted by (47) is not the one observed experimentally: the observed viscosity is proportional neither to T (as predicted for Maxwellian molecules with the interatomic forces proportional to r^{-5}) nor to $\sqrt{T}$ as predicted by the hard-sphere model according to (47). The observed temperature dependence is intermediate between these simple power laws. To illustrate the level of agreement between theory and experiment, we quote measured (calculated) values of the shear viscosity of Ne at a few selected temperatures, in units of 10^{-6} Pa·s: 11.98 (11.77) at $T = 80$ K, 23.76 (23.79) at $T = 200$ K, and 30.97 (30.65) at $T = 290$ K. The calculated values in parenthesis were obtained from a potential of the form (45). This level of agreement by far surpasses that which can be achieved when kinetic theory is applied to the transport properties of metals or semiconductors.

Other transport coefficients such as the second viscosity ζ or the self-diffusion coefficient D may be obtained in a similar fashion. One finds that the ratio of the kinematic viscosity $\nu = \eta/nm$ to the self-diffusion coefficient D (the so-called Schmidt number) depends somewhat on the assumed intermolecular potential. For hard spheres the first variational approximation yields a Schmidt number of 0.833, while one obtains the exact value 0.645 for atoms repelling each other by forces proportional to r^{-5} ("Maxwellian molecules").

2.1.3 Quantum Effects A gas is nondegenerate if the mean interparticle distance is much greater than the thermal deBroglie wavelength λ_{th} given by $\lambda_{\text{th}} = h/\sqrt{2\pi mkT}$. However, quantum effects may be important if λ_{th} is comparable with the "size" of the particles, specified, for instance, by the parameter σ in the potential (45). In such cases the scattering cross section must be calculated according to the laws of quantum mechanics, by determining the phase shifts associated with each angular momentum component of the incident wave. It is also necessary to allow for the indistinguishability of the molecules undergoing collisions. A quantitative measure of the importance of quantum effects is furnished by the deBoer parameter $h/\sigma\sqrt{m\epsilon}$, which is the square root of the ratio of the zero-point energy of a particle confined to a region of size σ and the well depth ϵ. For He this parameter is 2.60, for Ne 0.54, and for Ar 0.17.

2.2 Knudsen Flow

When a gas of particles is very dilute, the mean free path l between collisions may become comparable with characteristic lengths such as the diameter d of a tube through which the gas is flowing. Hydrodynamics then ceases to be applicable. When l/d is somewhat less than unity, it is possible to treat the leading correction to hydrodynamics by introducing a slip boundary condition in the hydrodynamic equations, relating the extrapolated flow velocity u_0 at the surface to the velocity gradient near the surface according to $u_0 = \zeta\partial u/\partial x$, where the x axis points normal to the boundary. The length ζ is called the slip length. It is of the same order of magnitude as the bulk mean free path, provided the boundary scattering is diffuse. If the boundary scattering is partly specular, the slip length is increased accordingly.

The *Knudsen number* is the ratio of the mean free path to a characteristic dimension. When it is large compared to unity, the concept of a slip length loses its meaning, and one enters the regime of rarefied gas dynamics important for vacuum technology and space research. Under these circumstances the Boltzmann equation must be solved for the geometry in question, with boundary conditions specifying the scattering properties of the walls of the container in which the gas is enclosed.

3. PLASMAS

In the previous section we considered transport phenomena in gases consisting of neutral particles interacting via forces that fall off rapidly with the distance between particles. If the gas is ionized, a very different situation occurs because of the $1/r$ dependence of the Coulomb interaction between the charged particles. A completely ionized gas is called a plasma. We here consider a two-component plasma consisting of electrons with charge $-e$ and ions with charge ze. The characteristic plasma densities and plasma temperatures span an enormous range, from those of a fusion plasma to an electron or hole plasma in a semiconductor. The unifying feature of these different physical systems is the role of the Coulomb interaction and the screening properties of the mobile charge carriers.

The screening properties of the plasma and its collective oscillations are obtained by solving the Maxwell equations together with the collisionless Boltzmann equation for the electrons and for the ions (often called the Vlasov equations). The conductivity is obtained from the Boltzmann equation and used in the Maxwell equations for the electromagnetic fields associated with the charge carriers. The Coulomb interaction between the particles is thus included in the collisionless Boltzmann equation via the self-consistent fields that determine the motion of the charged particles.

The reader may find a thorough treatment of the properties of a two-component plasma in the textbook by Lifshitz and Pitaevskii (1981).

3.1 Collective Oscillations

The charge and current induced by an external electrical field that varies in space and time according to $\exp(i\mathbf{q}\cdot\mathbf{r} - i\omega t)$ may be described in terms of a permittivity tensor $\epsilon_{ij}(\mathbf{q},\omega)$, which relates the displacement field **D** to the electric field **E** by

$$D_i = \epsilon_0\epsilon_{ij}E_j. \tag{48}$$

By inserting plane-wave solutions in the Maxwell equations and using the fact that $\mathbf{B} = \mu_0\mathbf{H}$, one arrives readily at the condition for propagation of electromagnetic waves,

$$\det\left(q_iq_j - q^2\delta_{ij} + \frac{\omega^2}{c^2}\epsilon_{ij}\right) = 0. \tag{49}$$

The condition (49) is obtained by eliminating the magnetic field from the Maxwell equations and inserting (48), using $1/\mu_0\epsilon_0 = c^2$. In empty space ($\epsilon_{ij} = \delta_{ij}$) it implies that $\omega^2 = c^2q^2$, corresponding to the propagation of electromagnetic waves with the electric and magnetic fields perpendicular to each other and to the direction of propagation.

In a medium such as a plasma, the dielectric tensor may have both a real and an imaginary part. Then (49) can only be satisfied by a complex ω, assuming that we have chosen the wave vector **q** to be real. The real part of ω yields the angular frequency of the wave, while the imaginary part determines its damping.

The dielectric tensor may be written in terms of the conductivity,

$$\epsilon_{ij} = \delta_{ij} - (1/i\omega\epsilon_0)\sigma_{ij}, \tag{50}$$

since the induced current density **j** is included with the displacement field in the Maxwell equation. The current density **j** is given by the distribution function according to

$$\mathbf{j} = \sum_{\alpha=e,i}\int d\mathbf{v}e_\alpha\mathbf{v}_\alpha f_\alpha. \tag{51}$$

We have introduced an index α, which indicates whether the corresponding quantity involves the electrons (e) or the ions (i). The quantity e_α is the charge of the particle in question ($e_e = -e$, where e is the elementary charge).

In order to linearize the collisionless Boltzmann equation in the electric field strength E we write $f = f^0 + \delta f$, with δf proportional to E, and obtain

$$-i\omega\delta f_\alpha + i\mathbf{q}\cdot\mathbf{v}_\alpha\delta f_\alpha + \frac{e_\alpha}{m_\alpha}\mathbf{v}_\alpha \times \mathbf{B}_0\cdot\frac{\partial}{\partial\mathbf{v}_\alpha}\delta f_\alpha = -\frac{e_\alpha}{m_\alpha}\mathbf{E}\cdot\frac{\partial f^0_\alpha}{\partial\mathbf{v}_\alpha}. \tag{52}$$

We have assumed the presence of a static magnetic field $\mathbf{B}_0$. The oscillatory part does not contribute, since the associated Lorentz force, which is proportional to E, is perpendicular to the velocity. For the same reason

(52) cannot be linearized in the static magnetic field, since this would make $\mathbf{B}_0$ vanish altogether from the equation.

We shall illustrate the use of the Boltzmann equation (52) by a few examples. The procedure is the following: First we determine σ_{ij} from the kinetic equations for the distribution functions of the electrons and ions. Then ϵ_{ij} is obtained from (50) and inserted into (49). The collective modes are found by setting the determinant of (49) equal to zero. The discussion may be divided into four cases, according to whether the wave vector $\mathbf{q}$ and the static magnetic field $\mathbf{B}_0$ are taken to be zero or nonzero in (52).

(a) $\mathbf{q} = 0$, $\mathbf{B}_0 = 0$. When we set $\mathbf{q} = 0$ in the Boltzmann equation, the dielectric tensor in (49) becomes independent of $\mathbf{q}$. Since the electrons are much lighter than the ions, we may neglect the contribution of the latter. The conductivity tensor is then given by the familiar Drude expression

$$\sigma_{ij} = -\frac{ne^2}{mi\omega}\delta_{ij}, \tag{53}$$

where n is the electron density. The corresponding solutions to (49) are given by $\omega^2 = c^2q^2 + \omega_{pe}^2$, if $\mathbf{E}$ is perpendicular to $\mathbf{q}$, and $\omega^2 = \omega_{pe}^2$, if $\mathbf{E}$ is parallel to $\mathbf{q}$. Here the plasma frequency $\omega_{p\alpha}$ is defined by $\omega_{p\alpha}^2 = n_\alpha e_\alpha^2/m_\alpha\epsilon_0$. The ion plasma frequency is much less than the electron plasma frequency because of the smallness of the ratio m_e/m_i.

(b) $\mathbf{q} \neq 0$, $\mathbf{B}_0 = 0$. If we keep $\mathbf{q}$ in the Boltzmann equation, we may determine the dispersion and damping of a plasma wave, neglecting the ion motion. We assume that the electrons and ions in equilibrium are described by classical distribution functions characterized by the electron temperature T_e and the ion temperature T_i, respectively. By expanding the conductivity to second order in q one finds the dispersion of the plasma wave given by

$$\omega^2 = \omega_{pe}^2 + \frac{3kT_e}{m}q^2, \tag{54}$$

while the damping is

$$\mathrm{Im}\,\frac{\omega}{\omega_{pe}} = -\frac{\pi^{1/2}}{8^{1/2}}\frac{e^2nm^{1/2}\omega}{\epsilon_0q^3(kT_e)^{3/2}}\exp\left(-\frac{m\omega^2}{2kT_eq^2}\right). \tag{55}$$

This *Landau damping* is reversible, caused by the transfer of energy from the electromagnetic wave to the motion of the individual particles.

We may also consider the motion of ions and electrons together, treating the dielectric screening due to electrons in the static limit, and that due to the ions in the dynamic limit. In the longitudinal case ($\mathbf{q}$ parallel to the electric field, which we take to be along the z axis), the vanishing of the determinant of (49) implies that $\epsilon_{zz} \equiv \epsilon_\parallel = 0$. The longitudinal part of the dielectric tensor then becomes

$$\epsilon_\parallel \simeq 1 - \frac{\omega_{pi}^2}{\omega^2}\left(1 + \frac{3q^2kT_i}{\omega^2m_i}\right) + \frac{\omega_{pe}^2m}{q^2kT_e}. \tag{56}$$

In the long-wavelength limit $q \to 0$, the condition $\epsilon_\parallel = 0$ yields the sound velocity of the ion-acoustic wave,

$$\omega^2 = q^2\frac{kT_e + 3kT_i}{m_i}. \tag{57}$$

The damping of the wave is found by taking into account the imaginary part of $\epsilon_\parallel$.

(c) $\mathbf{q} = 0$, $\mathbf{B}_0 \neq 0$. By solving the first-order differential equation (52), using cylindrical coordinates in velocity space, we find the dielectric tensor

$$\epsilon_{xx} = \epsilon_{yy} \equiv \epsilon_\perp = 1 - \sum_\alpha \frac{\omega_{p\alpha}^2}{\omega^2 - \omega_{c\alpha}^2}, \tag{58}$$

while

$$\epsilon_{xy} = -\epsilon_{yx} \equiv i\epsilon_a = i\sum_\alpha \frac{\omega_{p\alpha}^2\omega_{c\alpha}}{\omega(\omega^2 - \omega_{c\alpha}^2)}. \tag{59}$$

and

$$\epsilon_{zz} \equiv \epsilon_\parallel = 1 - \sum_\alpha \frac{\omega_{p\alpha}^2}{\omega^2}, \tag{60}$$

with $\omega_{c\alpha} = e_\alpha B_0/m_\alpha$ denoting the cyclotron frequency for electrons ($\alpha = e$) or ions ($\alpha = i$).

We return to the solution of the determinantal equation (49) and choose $\mathbf{q}$ in the xz plane, with θ denoting the angle between $\mathbf{q}$ and $\mathbf{B}_0$,

$$\mathbf{q} = (q\sin\theta, 0, q\cos\theta). \tag{61}$$

From the dielectric tensor given above together with (49) we find a quadratic equation in q^2c^2/ω^2 for the determination of ω as a function of $\mathbf{q}$:

$$(\epsilon_\perp \sin^2\theta + \epsilon_\| \cos^2\theta)\left(\frac{q^2c^2}{\omega^2}\right)^2 - [\epsilon_\perp\epsilon_\|(1 + \cos^2\theta) + (\epsilon_\perp^2 - \epsilon_a^2)\sin^2\theta]\frac{q^2c^2}{\omega^2} + (\epsilon_\perp^2 - \epsilon_a^2)\epsilon_\| = 0. \tag{62}$$

From this we get for a given θ a solution $q(\omega)$, which determines $\omega(\mathbf{q})$, since $\epsilon_\|$, $\epsilon_\perp$, and ϵ_a are given by (60), (58), and (59), respectively. Those frequencies that make the coefficient of $(q^2c^2/\omega^2)^2$ vanish are called *plasma resonances*. They evidently correspond to solutions for which $qc/\omega \to \infty$.

Let us examine whether there are oscillations with a linear dependence of ω on q for $q \to 0$. In the limit of low frequencies and with $m_e \ll m_i$, we have $\epsilon_\perp \simeq 1 + \omega_{pi}^2/\omega_{ci}^2$ and $\epsilon_\| \simeq -\omega_{pe}^2/\omega^2$. Since ϵ_a is proportional to ω in the limit $\omega \to 0$, we can neglect it, and we get two solutions, ω_1 and ω_2. One of the solutions is the magnetosonic wave, which has a frequency given by

$$\omega = \omega_1 = v_A q(1 + v_A^2/c^2)^{-1/2}. \tag{63}$$

Here we have introduced the Alfvén velocity v_A by the definition

$$v_A^2 = c^2\omega_{ci}^2/\omega_{pi}^2. \tag{64}$$

The other solution is the Alfvén wave

$$\omega = \omega_2 = v_A q|\cos\theta|(1 + v_A^2/c^2)^{-1/2}. \tag{65}$$

Another type of solution to (62) are the *whistlers* that propagate at frequencies intermediate between the ion and the electron cyclotron frequency with the dispersion relation $\omega^2 = c^4q^4\cos^2\theta\omega_{ce}^2/\omega_{pe}^4$. In solids they are named *helicon waves*.

(d) $\mathbf{q} \neq 0$, $\mathbf{B}_0 \neq 0$. When both the magnetic field and the wave vector are different from zero, the Boltzmann equation may be solved by a Fourier expansion, based on the eigenfunctions for the differential operator associated with the magnetic field. The significance of the additional term $i\mathbf{q}\cdot\mathbf{v}\delta f$ in the Boltzmann equation may be seen by considering the contribution of the electrons to the dielectric tensor when $\omega \simeq \omega_c$, provided both q_x and q_z are sufficiently small. In this case one finds that the singular behavior $\epsilon_{xx} - 1 \propto (\omega - \omega_c)^{-1}$ given by (58) is replaced by $\epsilon_{xx} - 1 \propto (\omega - \omega_c)$, provided $(\omega - \omega_c)^2 \ll kTq_z^2/m$. In addition ϵ_{xx} acquires an imaginary part.

3.2 Dielectric Screening

In the static limit ($\omega = 0$) the longitudinal dielectric function $\epsilon_\|$ obtained from the Boltzmann equation for a classical electron gas at temperature T, with neglect of the ion motion, is

$$\epsilon_\| = 1 + \frac{ne^2}{\epsilon_0 kTq^2} = 1 + \frac{q_D^2}{q^2}, \tag{66}$$

where we have introduced the Debye screening wave vector $q_D = (ne^2/\epsilon_0 kT)^{1/2}$. For degenerate electrons one obtains a similar result,

$$\epsilon_\| = 1 + \frac{3ne^2}{2\epsilon_0\epsilon_F q^2} = 1 + \frac{k_s^2}{q^2}, \tag{67}$$

where ϵ_F is the Fermi energy. Here k_s is called the Thomas–Fermi screening wave vector, and it is seen to be given by $k_s^2 = 4k_F/a_0$, where a_0 is the Bohr radius. This form of the dielectric function causes the infinite-range Coulomb interaction in a plasma of freely moving charged particles to be screened over distances of order q_D^{-1} or k_s^{-1}, depending on whether the plasma is nondegenerate or degenerate. The screening effect may be understood by dividing the Fourier transform of the unscreened Coulomb interaction, which is proportional to q^{-2}, by the dielectric function $\epsilon_\|$. This results in the replacement of q^{-2} by $1/(q^2 + k_s^2)$ (in the degenerate case), which after Fourier transformation yields a real-space potential proportional to $e^{-k_s r}/r$ of range k_s^{-1}. The magnitude of the screening length depends on the density of the plasma and—in the nondegenerate case—on the temperature. For a fusion plasma at a temperature $T = 10^8$ K and density $n = 10^{20}$ m^{-3} it is $q_D^{-1} = 7 \times 10^{-5}$ m. In metals the screening length is of the same order of magnitude as the mean interelectronic distance.

4. METALS AND SEMICONDUCTORS

Since the transport properties of metals and semiconductors have been covered in other articles (see ELECTRONIC TRANSPORT IN SOLIDS; METALS AND ALLOYS, CONDUCTIVITY IN; INSULATORS AND SEMICONDUCTORS, CONDUCTIVITY IN), we limit ourselves to a brief discussion of the use of the Boltzmann equation for the description of transport phenomena in normal and superconducting metals and in semiconductors. In doing so we merely point out the novel features that must be taken into account when using the semiclassical Boltzmann equation for such systems. While the Boltzmann equation is applicable to clean superconductors under conditions of slow variations in time and space, there exists a much more general framework for treating nonequilibrium properties of superconductors. This is the *quasiclassical method* pioneered by Eilenberger (1968) and Larkin and Ovchinnikov (1968). A review of the quasiclassical approach and the Keldysh method (Keldysh, 1964) of deriving kinetic equations has been given by Rammer and Smith (1986).

4.1 Normal Metals

The conduction electrons in a metal form a highly degenerate Fermi system of particles, which move in the periodic potential due to the ions. It is therefore essential to take quantum mechanics into account in describing their motion. In quantum mechanics it has strictly speaking no meaning to introduce a distribution function $f(\mathbf{r},\mathbf{p},t)$ that depends on the position $\mathbf{r}$ and the momentum $\mathbf{p}$ of a particle, since this would contradict the Heisenberg uncertainty relations. We may however take over the classical concept of a distribution function by interpreting particles in terms of wave packets centered around the position $\mathbf{r}$ and momentum $\mathbf{p}$ with corresponding uncertainties that satisfy the Heisenberg relations $\Delta x \Delta p_x \geq \hbar/2$ etc. With such an interpretation the classical concept of a distribution function may be taken over in quantum mechanics, provided the applied fields vary sufficiently slowly. The periodic potential, however, varies rapidly on the scale given by the extent of wave packets in real space. It is therefore necessary to treat the periodic potential according to quantum mechanics, while the external fields are treated classically. This semiclassical description of electron dynamics in a periodic lattice is obtained from the equation of motion (5) by replacing the kinetic momentum $\mathbf{p}_k$ by $\hbar\mathbf{k}$, where $\mathbf{k}$ is the Bloch vector, and identifying the velocity $\mathbf{v}$ with the group velocity $(1/\hbar)\partial\epsilon/\partial\mathbf{k}$, where $\epsilon(\mathbf{k})$ is the Bloch energy.

The Boltzmann equation for the distribution function $f(\mathbf{r},\mathbf{k},t)$ then assumes the usual form [cf. (4)], but its analysis is complicated by the fact that the energy $\epsilon(\mathbf{k})$ in general depends on the direction of $\mathbf{k}$ as well as its magnitude.

In simple cases the transport properties of a metal bear close resemblance to those of a plasma. As an example, the frequency-dependent conductivity of a metal is given by (53), if the frequency is sufficiently high that collisions may be neglected and if the energy bands are spherical. When the anisotropy is taken into account, one obtains for a cubic metal in the collisionless limit a result similar to (53), but with m replaced by an optical mass, which is inversely proportional to the velocity averaged over the Fermi surface. Similarly, the Hall effect at high magnetic fields may be obtained from the static limit ($\omega = 0$) of the conductivity tensor resulting in (58)–(60). The ion-acoustic waves given by (57) have their counterpart in the phonons of the metal, once the electrons are treated as a degenerate gas, resulting in the electron temperature in (57) being replaced by the Fermi temperature times a numerical constant.

In general, however, the transport properties of metals are much harder to analyze, because of the complexities introduced by the band structure. The collision processes limiting the electronic transport include scattering against impurities and other defects as well as electron–phonon scattering. Because of the periodic potential, the quasimomentum $\hbar\mathbf{k}$ is only conserved in collisions between electrons and phonons up to the addition of a reciprocal lattice vector. In performing realistic calculations of electrical and thermal conductivity, it is often necessary to include such *Umklapp* processes characterized by a nonzero reciprocal lattice vector.

Since the conduction electrons in a metal form a degenerate gas with a Fermi temperature that is typically a hundred times larger than room temperature, the electrical and thermal conductivities reflect the properties of the electronic states at the Fermi surface. The calculation of the thermoelectric properties of a metal requires one to take into account also the *variation* of the density of states, the group velocity, and the collision probability in the vicinity of the Fermi surface. The electrical current set up by a temperature gradient ∇T may be canceled by applying an electric field $\mathbf{E} = Q\nabla T$. The *Seebeck* coefficient Q may in a free-electron model (with energy-independent relaxation time) be shown to be equal to the entropy per particle divided by the electronic charge, but for metals with anisotropic band structure, with electrons being scattered by phonons and impurities, such thermoelectric coefficients are very difficult to calculate reliably.

4.2 Superconductors

The nonequilibrium properties of superconductors differ in several fundamental ways from those of normal metals. The most striking difference is the infinite conductivity characterizing the superconducting state. The thermal conductivity is finite, but its size and temperature dependence differ from that of normal metals, as a result of the existence of an energy gap for the excitations in the superconducting state. The thermoelectric properties differ qualitatively from those of normal metals, on account of the infinite electrical conductivity. The presence of a condensate means that the number of excitations is not a conserved quantity in a collision. This gives rise to novel relaxation phenomena that have no counterpart in normal metals. Also the magnitude of the energy gap depends on the distribution of excitations and may relax under nonequilibrium conditions.

Compared to the normal metallic state the most important novel features are thus

1. the change in the dispersion relation for the excitations due to the existence of an energy gap and
2. the coherence factors that relate the normal-state excitations to those of a superconductor.

It should also be emphasized that the conditions for the validity of the Boltzmann equation are much more stringent than in a metal, because of the existence of an energy gap Δ that is much less than the Fermi energy. The superconducting state is characterized by a coherence length of order $\hbar v_F/\Delta$, which is much longer than the interatomic distance. It is therefore necessary to restrict spatial variations to scales longer than the coherence length and temporal variations to scales longer than $\hbar/\Delta$. There are other conditions related to the phenomenon of *pair-breaking*, restricting, e.g., the inelastic scattering rate to being much less than $\Delta/\hbar$ in order that a Boltzmann description is applicable.

4.3 Semiconductors

Semiconductors are characterized by a conductivity that at room temperature is much smaller than that of metals, but much larger than that of typical insulators. In addition the conductivity of a semiconductor increases with increasing temperature, in contrast to the usual behavior observed in metals. When the temperature approaches zero, the conductivity of a semiconductor becomes vanishingly small. The presence of an energy gap for the excitation of particles (or holes) is the reason for this very different behavior of metals and semiconductors. As a result the density of current carriers is much lower in a semiconductor than in a metal, and the charge carriers may therefore usually be considered to be a nondegenerate gas. For this reason, transport in semiconductors bears some resemblance to transport in classical gases. The unique features of transport in semiconductors derive from the existence of two different kinds of charge carriers, electrons and holes. Certain aspects represent a simplification compared to metals. Thus it is often possible to use a quadratic relation between the energy and the Bloch wave vector for the states of relevance to the transport process. In both metals and semiconductors the important scattering processes limiting the transport of electricity are electron–impurity and electron–phonon scattering. Also electron–electron interaction plays a role insofar as it leads to screening of the electron–impurity and electron–phonon interaction.

In semiconductors an energy-dependent

relaxation time $\tau(\epsilon)$ is often used to account for the effect of collisions with impurities and phonons. The specific form of $\tau(\epsilon)$ depends on whether the scattering is due to, e.g., ionized impurities or deformation-potential coupling to acoustic phonons. As a result of the energy dependence of the relaxation time, even the simplest model based on a quadratic dispersion relation yields transport properties that differ qualitatively from those of the free-electron model in metals, by exhibiting, e.g., a finite magnetoresistance. For metals, the ratio between the thermal conductivity and T times the electrical conductivity is given by the universal Lorenz number $\pi^2k^2/3e^2$, provided elastic scattering dominates. In semiconductors the corresponding ratio is affected by the specific energy dependence of the relaxation time, because the charge carriers in equilibrium are described by a Maxwell–Boltzmann rather than a Fermi–Dirac distribution.

5. INSULATORS

The transport properties of insulators are dominated by the phonons, on account of the absence of free charge carriers. The heat conductivity is limited by the mutual interaction between phonons and by their scattering from defects and boundaries. The existence of independent normal modes, characterized by a wave vector $\mathbf{q}$ and frequency $\omega(\mathbf{q})$, depends on the validity of the harmonic approximation to the lattice Hamiltonian. Each normal mode is quantized as an ordinary harmonic oscillator, its excited states being specified by the number of vibrational quanta, phonons, of energy $\hbar\omega(\mathbf{q})$. If anharmonic forces are taken into account, the normal modes become coupled, but they still form a useful basis for the description of thermodynamic and transport properties of insulating crystals.

5.1 Heat Conductivity

While the thermodynamic properties of an insulating crystal are reasonably well described by treating the solid as a collection of independent phonons, the interaction between phonons is decisive in determining the transport of heat. In the absence of phonon–phonon interactions (as well as other scattering processes such as those due to impurities or boundaries), the thermal conductivity of an insulating crystal would be infinite, since nothing prevents the free flow of the phonon gas. So it is necessary to take into account anharmonicity, not as a correction, but as the dominant effect that limits the conduction of heat. The anharmonic terms in the lattice Hamiltonian give rise to both three- and four-phonon processes, since the number of phonons is not conserved in a collision. As a result of the periodicity of the lattice, the total quasimomentum in a scattering process is conserved up to the addition of a reciprocal lattice vector $\mathbf{G}$. The *normal* processes are characterized by $\mathbf{G} = 0$, while *Umklapp* processes have $\mathbf{G} \neq 0$. Only Umklapp processes may cause a heat current to decay toward zero, since the total quasimomentum is conserved by normal collisions.

The phonon spectra are anisotropic in the sense that the frequency $\omega(\mathbf{q})$ depends on the direction of $\mathbf{q}$, not just on its magnitude, in accordance with the symmetry of the underlying crystal lattice. It is therefore difficult to carry out realistic calculations of the phonon heat conductivity starting from the Boltzmann equation for the phonon distribution $N_{\mathbf{q}}$. In certain simple cases such as scattering by isotopic impurities or by boundaries, it is possible to obtain explicit answers that may be compared with experiment.

If a crystal contains isotopic impurities that are randomly distributed on the lattice sites, the elastic scattering of phonons from the isotopic impurities gives rise to an additional thermal resistance, apart from the contribution due to phonon–phonon scattering. The transition probability may be obtained from the change in kinetic energy $\Delta E = \Delta M \delta\dot{R}_i^2/2$ due to a single impurity situated at the lattice site $\mathbf{R}_i$. Here $\delta\dot{R}_i$ denotes the velocity of the isotopic impurity, with δR_i being the displacement from the lattice site. The difference in mass between the impurity and the host atom is ΔM. By introducing phonon coordinates it is then possible to determine the matrix element for scattering from $\mathbf{q}$ to $\mathbf{q}'$. The transition probability entering the Boltzmann equation is given within the Born approximation by the square of the matrix element for the scattering from a single impurity, multiplied by the number of impurities N_{imp}.

Another situation in which it is possible to derive simple expressions for the transport coefficients is thermal conduction at very low temperatures when the mean free path is so long that the important scattering mechanism is that due to the boundaries of the crystal. Under these conditions one can use expressions analogous to those derived for free molecular flow in gases. For a circular cylinder with diameter d one finds that the thermal conductivity κ is given by $\kappa = Csd/3$, where s is the sound velocity (assumed to be isotropic) and C the heat capacity per unit volume, in good agreement with experiment.

5.2 Second Sound

The phonon gas in insulators may support collective oscillations that are analogous to sound waves in an ordinary gas. Since the number density of phonons in equilibrium is determined by the temperature, the collective mode in the phonon gas is an oscillation in temperature. At sufficiently long wavelength the acoustical phonons have the same dispersion relation as the sound modes given by the equations of elasticity. The collective motion can therefore be thought of as a number-density wave in the gas of sound quanta, a "second sound." The collective temperature wave in solids only exists if the quasimomentum-destroying Umklapp processes are sufficiently *rare.* The existence of second sound in solids requires that the rate of quasimomentum-destroying processes is much less than the frequency of the collective mode, while at the same time the frequency should be much less than the rate of normal scattering processes. For this reason, second sound has only been observed in a few solids, namely, solid ^{3}He, solid ^{4}He, NaF, and Bi.

6. QUANTUM FLUIDS

With suitable modifications the Boltzmann equation may be used for the description of transport properties of quantum fluids. The two classic examples of quantum fluids are ^{3}He, which is a Fermi liquid, and ^{4}He, which is a Bose liquid. We here restrict ourselves to the discussion of Fermi liquids. For a general discussion of the concepts of Fermi-liquid theory, the reader may consult Pines and Nozières (1989).

6.1 Fermi Liquids

In liquid ^{3}He the interaction between the atoms is sufficiently strong that the specific heat at low temperatures is enhanced to a value several times that of the noninteracting system. The compressibility is strongly reduced by the repulsive interaction, yielding a sound velocity that is considerably larger than the Fermi velocity. The theory of Fermi liquids is due to Landau (1956), who assumed that the low-lying excited states, the elementary excitations, of an interacting Fermi system could be obtained by the gradual turning-on of the interaction between the particles in an ideal Fermi gas, with a one-to-one correspondence between the states of the ideal gas and those of the interacting system. These elementary excitations behave as quasiparticles in the sense that they possess a definite energy and momentum. The quasiparticle energy is a *functional* of the quasiparticle distribution function, implying that the energy of a quasiparticle depends on the occupation of the other quasiparticle states.

6.1.1 The Landau–Boltzmann Equation Formally, the energy $\epsilon_{\mathbf{p}\sigma}$ of a quasiparticle with momentum $\mathbf{p}$ and spin σ is introduced as the functional derivative of the total energy with respect to the quasiparticle distribution $f_{\mathbf{p}\sigma}$. The second functional derivative of the total energy yields the quasiparticle interaction function $f_{\mathbf{p}\sigma\mathbf{p}'\sigma'}$. The energy may thus be written as

$$\epsilon_{\mathbf{p}\sigma} = \epsilon^0_{\mathbf{p}\sigma} + \sum_{\sigma'} \int d\tau' f_{\mathbf{p}\sigma\mathbf{p}'\sigma'} \delta f_{\mathbf{p}'\sigma'}, \tag{68}$$

where $\delta f_{\mathbf{p}\sigma}$ is a small change in the distribution function. The energy ϵ^0 represents the increase in total energy when a single quasiparticle is added to an otherwise undisturbed system.

As the basis for his derivation of a generalized Boltzmann equation, Landau assumed that the energy $\epsilon_{p\sigma}(\mathbf{r},t)$ could be considered a quasiparticle Hamiltonian, with the quasiparticle velocity $\mathbf{v}_{\mathbf{p}\sigma}(\mathbf{r},t)$ given by

$$\dot{\mathbf{r}} = \mathbf{v}_{\mathbf{p}\sigma}(\mathbf{r},t) = \frac{\partial \epsilon_{\mathbf{p}\sigma}}{\partial \mathbf{p}}, \tag{69}$$

while the time rate of change of the momentum is

$$\dot{\mathbf{p}} = -\frac{\partial \epsilon_{\mathbf{p}\sigma}}{\partial \mathbf{r}}. \tag{70}$$

Because of the Hamiltonian form of these equations of motion, the Boltzmann equation becomes

$$\frac{\partial f_{\mathbf{p}\sigma}}{\partial t} + \frac{\partial \epsilon_{\mathbf{p}\sigma}}{\partial \mathbf{p}} . \frac{\partial f_{\mathbf{p}\sigma}}{\partial \mathbf{r}} - \frac{\partial \epsilon_{\mathbf{p}\sigma}}{\partial \mathbf{r}} . \frac{\partial f_{\mathbf{p}\sigma}}{\partial \mathbf{p}} = \left(\frac{\partial f_{\mathbf{p}\sigma}}{\partial t}\right)_{\text{coll}}. \tag{71}$$

Although this has the appearance of a conventional Boltzmann equation, it should be recalled that $\epsilon_{\mathbf{p}\sigma}$ is a functional of the distribution function and that $-\partial\epsilon_{\mathbf{p}\sigma}/\partial\mathbf{r}$ therefore includes interaction terms, once (68) is inserted. Such interaction terms are responsible for the unusual collective modes that may arise in a Fermi liquid such as *zero sound*. When (71) is linearized in the deviation from local equilibrium (the equilibrium Fermi function with $\epsilon_{\mathbf{p}\sigma}$ inserted), one obtains a transport equation with very different properties from the usual Boltzmann equation. For instance, one finds a soundlike mode in the collisionless regime, where the frequency ω is large compared to a typical collision rate $1/\tau$. The velocity of this mode differs from that of ordinary sound, which propagates in the opposite limit $\omega \ll 1/\tau$. The transition between ordinary (first) sound and zero sound is observed experimentally by studying sound propagation at a fixed frequency and cooling the liquid to lower temperatures, since a typical collision time varies as the inverse temperature squared, $1/\tau \propto T^2$. At the temperature corresponding to $\omega\tau \simeq 1$, the velocity of sound is found to increase (by a few percent), while the temperature dependence of the attenuation coefficient changes from that characteristic of hydrodynamic behavior (proportional to T^{-2}) to that characterizing the collisionless regime (proportional to T^2).

The $T = 0$ ground state of the Fermi liquid is a filled Fermi sea with all states with momentum $\mathbf{p}$ of magnitude less than p_F occupied. The Fermi momentum p_F is given by the same relation as in the noninteracting system, $p_F = \hbar(3\pi^2 n)^{1/3}$, where n is the number density of ^{3}He atoms. The fact that the typical collision rate is proportional to T^2 is a general property of a Fermi liquid. As a result the characteristic times entering the transport coefficients are all inversely proportional to T^2, albeit of different magnitude. As an example we quote expressions for the thermal conductivity κ and the viscosity η, which may be written as $\kappa = C_V v_F^2 \tau_\kappa/3$ and $\eta = m^* n v_F^2 \tau_\eta/5$, respectively. Here v_F is the Fermi velocity, defined by the expansion $\epsilon^0 = \mu + p_F(p - p_F)/m^*$, where μ is the chemical potential. The quantity C_V is the specific heat per unit volume, equal to m^*/m times its value in the interacting system. The times τ_η and τ_κ can be measured, since m^* is known from measurements of the specific heat. One finds at melting-curve pressure, $p = 34 \times 10^5$ Pa, that $\tau_\kappa T^2 = 2.2 \times 10^{-13}$ s K^2 and $\tau_\eta T^2 = 6.6 \times 10^{-13}$ s K^2. The calculated values, obtained from solutions to the Boltzmann equation and scattering amplitudes based on known values of the Fermi-liquid parameters, differ by less than 15% from such experimental results.

7. THE VALIDITY OF THE BOLTZMANN EQUATION

Although the Boltzmann equation was originally introduced for the purpose of describing transport phenomena in dilute classical gases, it has proved a successful approach to the description of the transport properties of both degenerate and strongly interacting systems. The limit of validity of the Boltzmann equation has been a subject of debate for many decades. The fact that agreement with experiment may be lacking for a particular system may be due to a fundamental limitation of the approach, but it may also reflect a lack of detailed knowledge of the system and the important collision processes that limit the transport. There exists no simple universal answer to the question: how good is the Boltzmann equation? Its validity depends on the particular physical system under consideration, including the nature of the different scattering processes that particles may undergo, and on external conditions such as the presence of electric or magnetic fields.

It is generally agreed that the Boltzmann description is valid for classical gases, in which the range of interaction R is much less than the average interparticle separation $n^{-1/3}$, where n is the particle density. The need for including quantum mechanics in the calculation of scattering cross sections entering the collision probability depends on whether the thermal de Broglie wavelength λ_{th} is so large that it is comparable to R. If this is not the case, classical mechanics will suffice. The validity of the Boltzmann equation for the description of transport in metals has been a much debated question. As regards impurity scattering, for electrons moving in three dimensions, the *Landau* criterion must be satisfied, $\hbar/\tau \ll \epsilon_F$. It is not, however, necessary that $\hbar/\tau$ is small compared to kT. In two dimensions, weak localization occurs, and the Boltzmann description is invalid at zero temperature no matter how weak the impurity scattering is. If only electron–phonon scattering is present in a metal, mass renormalization effects such as those entering the electronic specific heat are absent from transport coefficients such as electrical and thermal conductivity, but *not* from the thermoelectric coefficients. The additional presence of electron–impurity and electron–electron scattering complicates the issue, which must be addressed through microscopic calculations based on diagrammatic perturbation theory. Mesoscopic effects, in which the wave nature of the electron states plays an essential role, lie generally outside the scope of the Boltzmann equation.

There exist a variety of methods for deriving a Boltzmann equation starting from a Hamiltonian. Such methods have been used extensively for superconductors, in which the presence of an energy gap puts stronger limits of validity on the use of a Boltzmann equation. These more general methods supplement the approach we have described above in the sense that they may be more general but also more difficult to use. The Boltzmann approach to transport should be viewed as a useful tool—but not the only one—allowing a unified description of transport phenomena in gases, liquids, and solids.

GLOSSARY

Distribution Function: A measure of the mean number of particles in a volume element of the six-dimensional phase space of a single particle. For a classical gas in equilibrium the distribution function is given by the Maxwell distribution, and is thus proportional to $\exp(-mv^2/2kT)$, where v is the magnitude of the particle velocity and m the particle mass.

Phase Space: A six-dimensional space, which may be defined by the position vector $\mathbf{r}$ of a particle and its canonical momentum $\mathbf{p}$. In the case of an electron moving in the periodic potential of a solid, the phase space is defined by its position vector $\mathbf{r}$ and the Bloch vector $\mathbf{k}$.

Relaxation Time: A measure of the mean time between collisions. It may depend on energy and temperature and, in solids, on the direction of the Bloch vector $\mathbf{k}$. The transport time entering the transport coefficient is a related concept, but may differ in magnitude and temperature dependence from (typical) relaxation times.

Transport Coefficient: The coefficient relating the driving force to the current density. In the case of electric conduction the current density $\mathbf{j}$ equals $\sigma\mathbf{E}$, where $\mathbf{E}$ is the electric field strength and σ the transport coefficient, which is independent of the electric field under conditions of linear response.

Works Cited

Cercignani, C. (1988), *The Boltzmann Equation and Its Applications,* New York: Springer.

Eilenberger, G. (1968), *Z. Phys.* **214,** 195–213.

Keldysh, L. V. (1964), *Zh. Eksp. Teor. Fiz.* **47,** 1515–1527 [(1965), *Sov. Phys. JETP* **20,** 1018–1026].

Landau, L. D. (1956), *Zh. Eksp. Teor. Fiz.* **30,** 1058–1064 (*Sov. Phys. JETP* **3,** 920–925).

Larkin, A. I., Ovchinnikov, Yu. N. (1968), *Zh. Eksp. Teor. Fiz.* **55,** 2262–2272 [(1969), *Sov. Phys. JETP* **28,** 1200–1205].

Lifshitz, E. M., Pitaevskii, L. P. (1981), *Physical Kinetics,* New York: Pergamon Press.

Pines, D., Nozières, P. (1989), *The Theory of Quantum Liquids,* Reading, MA: Addison-Wesley.

Rammer, J., Smith, H. (1986), *Rev. Mod. Phys.* **58,** 323–359.

Smith, H., Jensen, H. Højgaard (1989), *Transport Phenomena,* Oxford, U.K.: Oxford Univ. Press.

Further Reading

Ashcroft, N. W., Mermin, N. D. (1976), *Solid State Physics,* New York: Holt, Rinehart and Winston.

Cercignani, C. (1975), *Theory and Applications* *of the Boltzmann Equation,* Edinburgh: Scottish Academic Press.

Kadanoff, L., Baym, G. (1989), *Quantum Statistical Mechanics,* Reading, MA: Addison-Wesley.

Lee, P. A., Ramakrishnan, T. V. (1985), *Rev. Mod. Phys.* **57,** 287–337.

Ziman, J. (1960), *Electrons and Phonons,* Oxford, U.K.: Oxford University Press.

TRAPPED ATOMS

See ATOMIC COOLING AND TRAPPING

TRIBOLOGY

Reizo Kaneko, *NTT Advanced Technology Corp., Musashino-shi, Tokyo, Japan*

Yuji Enomoto, *MITI Mechanical Engineering Laboratory, Tsukuba-shi, Ibaraki, Japan*

Shigeyuki Mori, *Department of Applied Chemistry and Molecular Science, Iwate University, Morioka-shi, Iwate, Japan*

Introduction 283
1. **Tribophysics** 284
1.1 General Aspects 284
1.2 Contact Mechanism 285
1.3 Friction 286
1.4 Wear 286
1.5 Lubrication 287
1.6 Tribology at Atomic and Molecular Scales 288
2. **Tribochemistry** 289
2.1 General Aspects 289
2.2 Chemical Interaction with Surfaces 290
2.2.1 Chemisorption on Oxide-Covered Surfaces 290
2.2.2 Chemisorption on Nascent Surfaces 290
2.2.3 Tribochemical Reactions 290
2.3 Lubrication Behavior 291
2.3.1 Lubrication with Adsorbed Layers 291
2.3.2 Lubrication with Reacted Layers 291
2.3.3 Chemomechanical Effect 291
2.3.4 Degradation of Lubricants 291
2.3.5 Effects of Additives 292
3. **Tribomaterials** 292
3.1 General Aspects 292
3.2 Liquid Lubricants 292
3.3 Solid Lubricants 293
3.4 Wear-Resistant Sliding Materials 294
3.5 Coating and Surface Modifications 294
4. **Summary** 295
Glossary 295
Works Cited 296
Further Reading 296

INTRODUCTION

Tribology is defined as "the science and technology of interacting surfaces in relative motion and of related subjects and practices." This word was coined by a British committee in 1966 from the word "tribos," which means "rubbing" in classical Greek.

A typical example of relative motion is sliding. The lateral force due to sliding is friction; the process of a surface being worn down by friction is wear; and the action of applying oil or some other substance in order to decrease friction and wear is lubrication. Decreasing the friction enables a machine to run more easily and reduces its energy consumption. Decreasing wear extends a machine's working life. On the other hand, brakes and friction-drive mechanisms, such as traction-drive transmissions and ultrasonic motors, make use of frictional forces. Tribology is a key technology for mechanical systems, because almost all machines have sliding mechanisms.

Sliding mechanisms have a very long history. Wheels having bearings were invented about 5000 years ago. The ancient Egyptians used liquid lubricants to slide sledges loaded with huge stones or statues. However, scientific studies of friction began much more recently than these ancient empirical applications.

In the 17th century, Amonton discovered the two basic laws of friction: The frictional force is proportional to the load and independent of the area of the sliding surfaces. These laws were verified in the 18th century by Coulomb and are now known as the Amonton–Coulomb laws. However, Leonardo da Vinci had already foreseen these laws through his scientific experiments in the 15th century. There were two different theo-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ries about the origin of friction. Coulomb concluded that friction was due to the interlocking of surface asperities. On the other hand, at about the same time, Desaguliers proposed that friction was due to molecular adhesion between the surfaces. It took about 200 years to establish the scientific theory of friction. The interlocking theory cannot explain the energy dissipation calculated from the friction force. From the end of the 19th century to the beginning of the 20th century, researchers including Hardy and Tomlinson systematized the adhesion theory. However, this theory requires a friction force proportional to the contact area, which conflicts with experimental results that the friction force is independent of the contact area. In about 1940, several researchers including Holm, Bowden, and Tabor found that the real contact area is very small compared with the geometrical contact area, and Bowden and Tabor explained the empirical Amonton–Coulomb laws using deformations of contacting asperities and adhesion of the real contact areas. This theory is now generally accepted. On the other hand, molecular dynamics studies in the 1990s have led to a new interlocking theory of friction being evaluated as an atomistic friction mechanism without wear.

In the 19th century, Tower showed that a rotating journal bearing generates high pressure in the oil film between the bearing and the spindle. This hydrodynamic pressure sustains the bearing load. Reynolds formulated an equation expressing the pressure in 1886. This lubrication equation is called the Reynolds equation. Around the turn of the 20th century, Stribeck experimentally studied the influence of variables such as load and velocity on friction. A curve of friction versus a combination of such variables is known as a Stribeck curve (see Sec. 1.5).

When lubricated sliding surfaces are so close to each other that surface asperities and the average gap are of the same order, some parts of the surfaces may be covered by very thin lubricant films (a monolayer or a few layers of lubricant molecules). In the 1920s, Hardy studied this regime, called boundary lubrication. He pointed out that the chemical constitution of the lubricant is important rather than its viscosity. Since then, chemically active lubricant additives have been developed for boundary lubrication.

The scientific study of wear started in the 1930s. In 1953, Archard found the empirical rule that the wear volume is proportional to the load and the sliding distance, and inversely proportional to the hardness. At about the same time, Burwell classified the physical mechanisms of wear. However, there are many kinds of interactions between sliding surfaces, such as mechanical, chemical, thermal, and fluidic interactions. The effects of these interactions depend on the sliding conditions and materials, and they affect each other. Furthermore, wear is made more difficult to understand by wear processes being irreversible. Although a lot of engineering data on wear has been accumulated, scientific explanations are still insufficient.

In 1938, Siebel suggested that joint research was needed by physicists, chemists, metallurgists, theorists, designers, and engineers because of the complexity of wear processes. This suggestion is still important for tribology today. In many cases, metals have been used for sliding surfaces, but new materials, such as ceramics and hard coatings, are beginning to be used widely. For progress in materials, we need much more scientific knowledge to understand friction, lubrication, and wear processes and to develop materials and machines having higher tribological performance. Not only engineering data but also scientific understanding of phenomena are needed for tribology. Thus, tribology ranges from basic physics and chemistry to machine development—it is an interdisciplinary research area.

1. TRIBOPHYSICS

1.1 General Aspects

Tribophysics depends on the dynamics of continuum bodies, and several important mechanisms have been clarified, such as deformation, friction, lubrication, and wear. Another approach is from physics at atomic and molecular scales. This approach is to reveal fundamental physical information that is important for understanding the origin of friction and wear and learning the limits of continuum-body dynamics.

1.2 Contact Mechanism

Practical surfaces are not flat but have asperities. The real contact area is small. Even with lightly loaded surfaces, the pressure at the points of contact is high, and these contact regions are responsible for the friction, wear, and interaction between the solid surfaces.

Hertz analyzed the contact of elastic bodies at the end of the 19th century. The process and results of this analysis can be found in many books on the deformation theory of materials (for instance, Timochenko and Goodier, 1951). When two spheres having radii R_1 and R_2 make contact with each other as shown in Fig. 1 under a load W, then if W is not too large, the deformation is elastic. The contact circle of radius a, the indentation distance δ, and the maximum pressure p_0 are shown in Table 1, where E is the elastic modulus, ν is Poisson's ratio, and W is the load.

Concerning the stresses in the contact region, we find the following:

Table 1. Results of Hertz contact equation between two spheres. $\Theta_1 = (1 - \nu_i^2)/E_i$, $\gamma_i = 1/R_i$, $p_0 = 1.5p_{\text{mean}}$; $i = 1, 2$.

a	$\left(\frac{3W}{4}\frac{\Theta_1 + \Theta_2}{\gamma_1 + \gamma_2}\right)^{1/3}$
δ	$\left[\frac{9W^2}{16}(\Theta_1 + \Theta_2)^2(\gamma_1 + \gamma_2)\right]^{1/3}$
p_0	$\frac{3}{2\pi}\left[W\left(\frac{3}{4}\frac{\Theta_1 + \Theta_2}{\gamma_1 + \gamma_2}\right)^{-2}\right]^{1/3}$

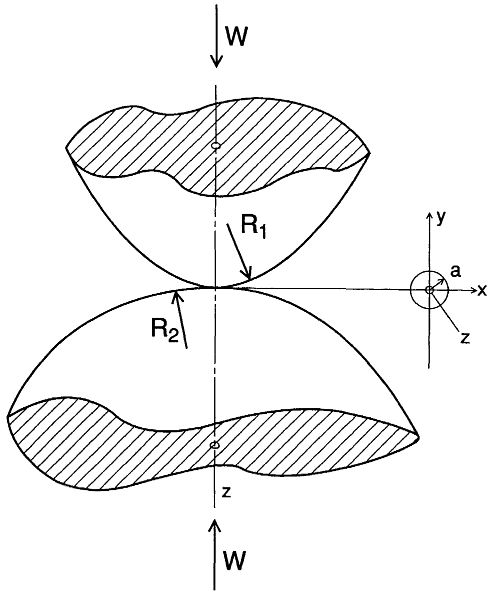

FIG. 1. Contact between two spheres.

1. Maximum compression pressure occurs at the center of the contact and is equal to the maximum pressure.
2. A brittle material is broken by the maximum tensile stress. In this case, the maximum tensile stress $(\sigma_r)_{\max}$ occurs around the contact circle and is expressed by

$$(\sigma_r)_{\max} = \tfrac{1}{3}(1 - 2\nu)p_0. \tag{1}$$

3. A plastic material tends to deform plastically at the maximum shear stress. In this case, the maximum shear stress occurs inside the material, a short distance from the contact surface. If Poisson's ratio ν is 0.3, the main shear stress is equal to $0.31P_0$ at $z = 0.47a$.
4. When friction acts on the contact area, it affects the values of the deformation and the stresses.

As the load increases, the pressure increases and exceeds the elastic limit. At this stage, plastic deformation occurs at the point of maximum shear stress in a plastic material. However, the other points still remain in a state of elastic deformation. When the load is increased further, the region of plasticity grows rapidly, and a stage is reached in which all the material around the contact region flows plastically.

Contact between practical surfaces occurs at many points. The real contact area is very small compared with the nominal contact area, and it varies with the deformation state—plastic or elastic. Greenwood and Williamson (1966) proposed the following equation for a plasticity index ψ involving the properties of the geometries of the surfaces in contact:

$$\psi = (E'/H)(\delta_a/r_m)^{1/2}$$
$$\frac{1}{E'} = \frac{1 - \nu_1^2}{E_1} + \frac{1 - \nu_2^2}{E_2}, \tag{2}$$

where H is the hardness, δ_a is the standard deviation of the distribution of asperity summits, and r_m is the average radius of the contact points. Considering the ratio of the areas of plastic contact and elastic contact, the state of deformation is estimated (Whitehouse and Archard, 1970) as follows:

1. Almost all contact points are elastic at $\psi < 0.6$.
2. Almost all contact points are plastic at $\psi > 3$.

Although the Hertz equations indicate that a zero load produces zero deformation, some surface interaction force still remains. Johnson *et al.* (1971) expressed this effect with the derivatives of the surface energy, showing a finite contact area even at zero load.

A convenient way to determine the real contact area is to measure the electrical resistance between the touching surfaces. Furthermore, electrical contacts are often used in practice in electrical relays and sliding brushes. A large electric current produces heat at the contact point and sometimes causes the contact point material to melt (Holm, 1967).

1.3 Friction

A primary factor in dry friction is the shear force needed to break the adhering points created by contact between two solid surfaces. The shear force (friction force) F and the friction coefficient μ are expressed by the following equation:

$$\mu = F/W = A_r\tau/W = \tau/p_m. \tag{3}$$

Here, A_r is the sum of the real contact areas, τ is the shear strength of adhering points, W is the load, and p_m is the yield stress.

When plastic deformation occurs at a contact area of clean metallic surfaces and adhesive junctions are formed, the contact area increases because of the combined action of tangential force and normal force. This phenomenon is called junction growth. When dry friction occurs between clean metal surfaces, the friction coefficient is very high (up to 100 in extreme cases). Buckley (1981) measured friction coefficients of several transition metals in vacuum conditions and showed that with an increase in the *d*-bond character defined by Pauling's resonating-valence-bond theory of metals (Pauling, 1949), there is a decrease in the friction coefficient; that is, the greater the degree of self-bonding (of a metal), the less the bonding across the interface.

Because the adhesive force is reduced by adsorption of gases on surfaces, the friction coefficients in air are lower than those in vacuum. Sliding surfaces covered with a low–shear-strength material or an inactive material exhibit low frictional coefficients. A lubricant layer decreases friction considerably (see Sec. 1.5).

Most of the work of friction is converted into heat. Even with moderate speeds and loads, the surface temperature often becomes very high, which changes the surface conditions.

1.4 Wear

Wear processes are classified into four types:

1. adhesive wear,
2. abrasive wear,
3. chemical wear, and
4. fatigue wear.

Adhesive wear is the result of breaking the adhering points between surfaces. Abrasive wear is a kind of cutting of a sliding surface by sharp hard points. In many cases, the wear rate depends on this type of wear. Chemical wear results from removal of the corrosion layer by environmental gases and/or liquids adsorbed on the solid surface. The most common type of corrosion is oxidation. Corrosion is accelerated by the repeated mechanical interactions of sliding. Fatigue wear is surface destruction caused by cyclic stress changes during sliding. The delamination theory explains that cracks are nucleated and propagated below the surface with repeated loading, and these cracks finally shear to the surface (Suh, 1977).

In general, wear is driven by the load, and the wear rate is proportional to the load in light-load conditions. Sometimes, there is a critical load and critical sliding speed, which indicates the transition load from a small wear rate (mild wear) to a heavy wear rate (severe wear). Figure 2 shows a wear map (Ashby and Lim, 1990). This map shows that wear processes contain many physical and chemical factors. The temperature increase

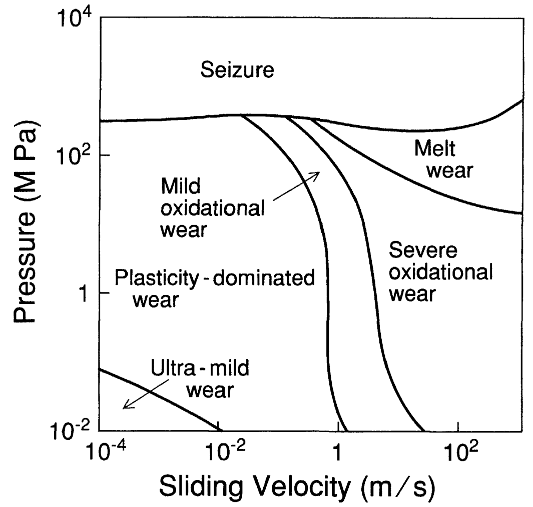

FIG. 2. Example of a wear-mechanism map. Pin-on-disk wear tests for a steel surface (Ashby and Lim, 1990).

and temperature gradients influence the material properties and activate chemical reactions. There are two temperature rises: a rise in bulk temperature and a rise in local temperature or "flash temperature." When the bulk temperature exceeds the melting point, the surface wears down rapidly by melt wear. Severe oxidational wear occurs in conditions that are not severe enough to cause melting; a high flash temperature still appears at conditions insufficient to make the bulk temperature high, and the real contact areas are hot enough to oxidize. With a low sliding velocity and low pressure, there is no significant temperature rise of any sort. In that case, plasticity-dominated wear occurs. With a very low sliding velocity and very low pressure, the plastic deformation no longer dominates. This region is still unclear. It corresponds to the range from molecular or atomic level wear to "zero wear." The wear process on a practical surface is complicated, because many factors such as adhesion between materials, chemical interactions, heat effects, and long-term changes in the surfaces affect wear processes. Our understanding of wear mechanisms is still empirical and unclear.

1.5 Lubrication

Lubrication is a process that reduces friction and wear of sliding surfaces with a low-shear material film or a fluid film. Solids with a lamellar structure, such as graphite and molybdenum disulfide, have low friction. Surfaces having low surface energy, such as a fluorinated surface, also have low friction. Solid-film lubrication is essentially a state of dry friction. Solid lubricants are introduced in Sec. 3.3.

The different types of fluid-film lubrication with oil are explained by the Stribeck curve shown in Fig. 3. This curve shows the change in the coefficient of friction as a function of the nondimensional parameter $\eta V/p$, where η is the viscosity of the lubricant, V is the sliding velocity, and p is the nominal pressure on the sliding surface. There are three modes: hydrodynamic lubrication, mixed lubrication, and boundary lubrication.

When a thick liquid or gas film exists between solid surfaces and the fluid pressure of the film sustains the load, friction force is generated by the shear force.

A fluid film in a converging wedge between sliding surfaces develops hydrodynamic pressures. These pressures can be expressed with the Reynolds equation. A generalized Reynolds equation was derived by Dowson (1962) from the Navier–Stokes

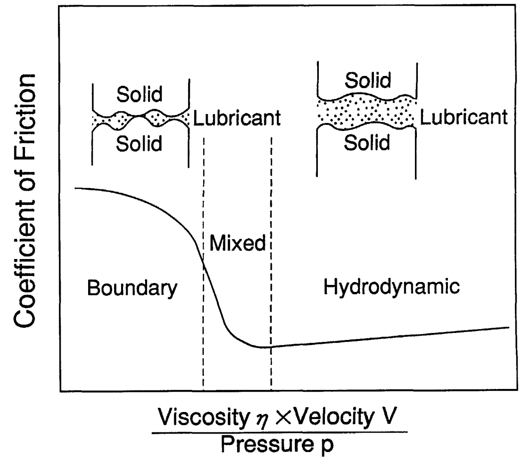

FIG. 3. Stribeck curve. Under purely hydrodynamic lubrication, the coefficient of friction is low and proportional to $\eta V/p$. Under boundary lubrication, the coefficient of friction depends mainly on the properties of lubricant molecules adsorbed on solid surfaces. When the lubricant film is broken, solid-to-solid contact occurs and the coefficient of friction becomes large. Mixed lubrication is the transitional region from hydrodynamic lubrication to boundary lubrication.

equation expressing the dynamics of continuum flow, using the following assumptions:

1. The thickness of the fluid film is very small compared with the dimension of the sliding surface.
2. There is no variation in pressure across the fluid film.
3. The flow is laminar.
4. No external forces act on the film.
5. Fluid inertia is small compared with the viscous shear.
6. There is no slip at the sliding surfaces.
7. The velocity gradient across the fluid film is dominant.

Adding the assumptions that there are no variations in the fluid density ρ or coefficient of viscosity η across the fluid film and that the surfaces slide in the x direction with speeds of V_1 and V_2, the Reynolds equation is expressed as follows, referring to Fig. 4:

$$\frac{\partial}{\partial x}\left(\frac{\rho h^3}{\eta}\frac{\partial p}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\rho h^3}{\eta}\frac{\partial p}{\partial y}\right) = 12\frac{\partial(\rho h)}{\partial t} + 6\frac{\partial}{\partial x}[\rho h(V_1 + V_2)]. \quad (4)$$

In hydrodynamic lubrication mode, the sliding surfaces are completely separated, and there is no wear. The friction coefficient is proportional to the parameter $\eta V/p$.

When very high hydrodynamic pressure is generated locally across a small gap, the surface is elastically deformed, and this affects the hydrodynamic behavior. Lubricant properties are sometimes changed by high pressure. This condition is called elastohydrodynamic lubrication (EHL).

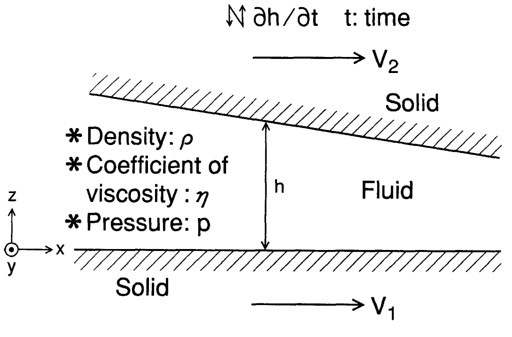

FIG. 4. Fluid film between solid surfaces. Here, p is the pressure in the film and V_1 and V_2 are sliding speeds.

When sliding surfaces are so close to each other that surface asperities are dominant, hydrodynamic effects are not sufficient to support a load. However, very thin lubricant films (a monolayer or a few layers of lubricant molecules) on asperities still sustain part of the load. This is mixed lubrication. When the sliding speed is slow, the load is supported by very thin lubricant films instead of by the hydrodynamic pressure. In many cases, several areas of the very thin lubricant films are broken by the high pressures, and solid-to-solid contact and wear occur. The friction coefficient becomes higher and independent of the viscosity of the lubricant and the sliding velocity. This is boundary lubrication. In boundary lubrication, the physical and chemical properties of a lubricant molecule are more important than the coefficient of viscosity in continuum flow.

1.6 Tribology at Atomic and Molecular Scales

The ultimate goal will be to understand tribological phenomena with physics at atomic and molecular scales. Sliding parts with very light loads are beginning to be used in magnetic recording and micromechanical systems, which require wear rates that are almost zero. This field is called "microtribology," and it needs more engineering data from atomic and molecular level studies. Thus tribology at atomic and molecular scales is important for both science and technology.

Computer simulation based on molecular dynamics can be used to trace the movement of each atom in contacting or sliding surfaces. Landman *et al.* (1990) showed atomistic mechanisms of adhesion and indentation. Hirano and Shinjo (1990) studied the origin of friction, and they found the existence of a sliding condition that generates no frictional force. This state of vanishing friction is called "superlubricity."

Scanning probe methods, such as a scanning tunneling microscope (STM) and an atomic force microscope (AFM), are used for atomic-level observation of surfaces. An AFM can also be used as a nano-hardness tester and nano-wear tester. Kaneko and Hamada (1993) reported the microwear process of polymers. For example, a polycarbonate surface first formed a plateaulike upheaval several nanometers high, then formed projections, and finally was worn down. The wear marks of the first and second stages reflect the plastic deformation that led to destruction.

Using an AFM, Mayer *et al.* (1992) observed wearless friction on surfaces covered with a Langmuir–Blodgett film. The initial stages of wear were also observed as a displacement of groups of molecules. A surface force apparatus (SFA) (Israelachvili, 1992) is used to observe the behavior of a molecularly thin film between smooth solid surfaces. One of the main achievements of the SFA has been to measure oscillatory solvation or structural forces. An ultrathin film of simple spherical molecules such as cyclohexane shows high friction with stick-slip motion, where, in the sticking state, the film becomes essentially solidlike. On the other hand, longer-chained and branched molecules give low friction and smoother sliding (Israelachvili *et al.*, 1990).

Surfactant molecules, tightly bound to the smooth surfaces, show stick-slip motion not only between a "static" and a "kinetic" state but also between a static or kinetic state and some "superkinetic" state that exhibits ultralow friction (Yoshizawa *et al.*, 1993). In the superkinetic regime, the surfactant chains become shear-aligned or combed into a configuration that results in very low friction.

One of the most important parameters in molecular gas dynamics is the Knudsen number, which is defined as $\mathrm{Kn} = \lambda/h$, where λ is the molecular mean free path and h is the characteristic length of the flow. When the value of Kn is larger than 0.1, flows in the gas film cannot be described by a continuum flow model. Fukui and Kaneko (1988) introduced a Reynolds-type equation for an arbitrary Knudsen number. This equation is called the molecular gas-film lubrication (MGL) equation. The MGL equation was derived from a linearized Boltzmann equation and contains a term expressing pressure flow for arbitrary Knudsen number. The conventional Reynolds equation based on a continuum flow model underestimates the pressure flow rate for a high Knudsen number. In fact, the molecular gas effects decrease the pressure of a gas film. These effects cannot be ignored for gas-film lubrication mechanisms having a very small spacing, such as a flying-head slider for magnetic recording that flies with a spacing of 100 nm or less.

2. TRIBOCHEMISTRY

2.1 General Aspects

Liquid lubricants are a mixture of a base oil and additives, which are mainly composed of organic compounds. Under boundary lubricating conditions, the lubricating performance of liquid lubricants is closely related to their chemical behavior. Details of lubricant components are shown in Sec. 3.2.

Tribochemistry includes chemical reactions and chemical phenomena such as chemisorption under tribological processes. Tribological behavior is strongly affected by tribochemical reactions—for example, the formation of lubricating surface films from lubricant additives, chemical or corrosive wear, and degradation of lubricants as shown in Fig. 5.

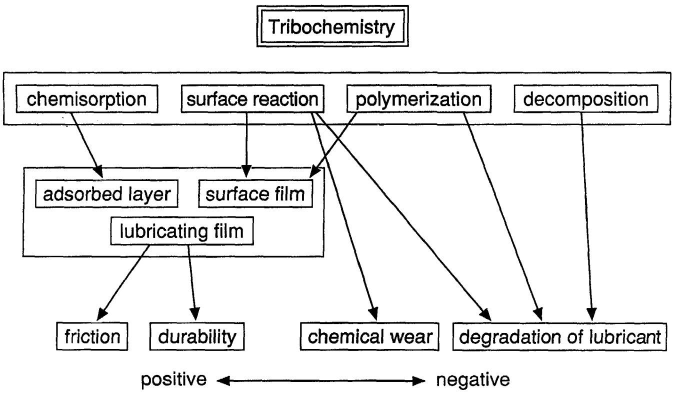

FIG. 5. Relationship between tribological performance and tribochemistry.

Tribochemical phenomena occur under the effect of mechanical contacts. Chemical reactions are enhanced by the mechanical effect at sliding contacts. There are two main causes for tribochemical reactions:

1. temperature and pressure rise at sliding contacts and
2. enhanced activity of mechanically contacting surfaces as shown in Fig. 6.

In general, surfaces are stabilized chemically by oxides, hydroxides, adsorbed water, and organic contaminants on the surfaces. When the surface layers are removed by mechanical contact, active sites such as surface defects are introduced on the surface.

2.2 Chemical Interaction with Surfaces

Adsorbed layers and reaction products from lubricant additives act as a lubricating film. Frictional properties in boundary lubrication are controlled by the surface coverage of the lubricant film to prevent direct solid–solid contact. The lubricity of the surface films depends on the chemical interaction between lubricant molecules and surfaces. The interaction is estimated from the chemisorption energy.

2.2.1 Chemisorption on Oxide-Covered Surfaces Lubricant components chemisorb on oxide or hydroxide layers on metals that are formed at atmospheric pressure. The chemisorption energy can be estimated from the heat of adsorption as well as from the heat of immersion. Compounds having a polar functional group, such as a carboxyl or amine group, chemisorb more easily than nonpolar compounds by a polar–polar interaction between the functional group and the oxidized surface. For example, the heat of immersion of water on silica or alumina is more than five times that of *n*-hexane (Wade and Hackerman, 1964). The heat of adsorption on iron oxide powder is ranked as follows: mineral oil (hydrocarbon) < cetyl alcohol < stearic acid = stearyl amine (Groszek, 1962). Fatty acids interact strongly with metal oxides on surfaces, resulting in the formation of stable surface films, sometimes a soap, and the surface films show good lubricity. Polar compounds, such as fatty acids having a higher heat of adsorption on iron oxide surface, exhibit a lower wear rate for steel (Groszek, 1962).

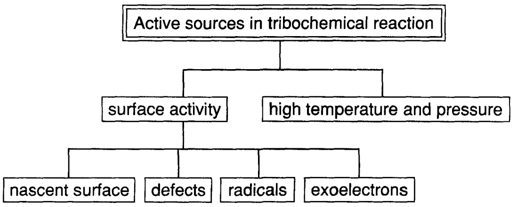
Active sources in tribochemical reaction
surface activity
high temperature and pressure
nascent surface
defects
radicals
exoelectrons

FIG. 6. Causes of tribochemical activation.

2.2.2 Chemisorption on Nascent Surfaces The surface layers on metals are removed by mechanical contact, resulting in the formation of nascent metal surfaces. Lubrication behavior under severe conditions is affected by chemical reactions on the nascent surfaces, which are chemically active because of dangling bonds on the surface. Evaporated metal films have been used as a model of the nascent surface. Since α-olefins chemisorb on evaporated aluminum films, they have been used as a metal-working fluid for aluminum (Owens *et al.*, 1966). Decomposition of carboxylic acid was also reported using evaporated iron film (Morecroft, 1971).

The nascent metal surface formed by scratching proved to be more active than evaporated films. On the other hand, the chemisorption behavior of model lubricant compounds on the nascent steel surfaces is different from that on metal oxide surfaces and closely depends on their molecular structure (Mori and Imaizumi, 1988). Nonpolar compounds such as aromatic compounds, olefins, and organic sulfides chemisorb on transition metals more easily than polar compounds such as carboxylic acids and amines that chemisorb easily on metal oxide surfaces. Saturated hydrocarbons do not chemisorb on nascent metal surfaces.

2.2.3 Tribochemical Reactions Under severe sliding conditions, the temperature at a real contact may exceed several hundred degrees centigrade (Bowden and Thomas, 1954). This high temperature at the contact is called a flash temperature or hot spot. Under such conditions, lubrication films are formed by tribochemical reaction between lubricant additives and materials. Chemical or corrosive wear is accelerated by tribochemical reaction between materials and surrounding gases and liquid chemicals such as oxygen, sulfur dioxide, and lubricant additives. Lubricant degradation is also accelerated by tribochemical activation.

Mechanical action alters the surface structure and composition, resulting in a change in chemical activity of surfaces (Buckley, 1974). The enhanced chemical activity originates from nascent surfaces of metals and ceramics and radical sites at ruptured covalent bonds of polymers and ceramics. For example, lubricant degradation and polymerization of aromatic compounds have been reported under the chemical effect of metallic surfaces (Goldblatt, 1971). The radical sites initiate some reactions such as oxidation, decomposition, and polymerization of organic compounds (Kurokawa *et al.*, 1978; Schrader *et al.*, 1969).

Electrons, known as exoelectrons, are emitted from the mechanically activated surfaces (Kramer, 1957). The emission intensity increases with stimulation by both thermal and optical energy. Since the emitted electrons have an energy of more than 1 eV, they have been identified as one of the active sources for tribochemical reactions—for example, the formation of soap from carboxylic acid (Smith and Fort, 1958) and the reaction of lubricant additives (Kajdas, 1987).

2.3 Lubrication Behavior

Under boundary lubrication conditions, lubricating behavior is controlled by surface films such as adsorbed layers and reacted layers. Chemisorbed organic layers prevent adhesion between two solid surfaces, which reduces the friction force and wear. Reacted layers are often composed of soft substances such as metal sulfides and phosphates that prevent adhesive wear.

2.3.1 Lubrication with Adsorbed Layers Adatoms such as oxygen, sulfur, and chlorine on metal surfaces reduce adhesion and friction between two metal surfaces (Ferrante, 1987; Wheeler, 1976). Organic monomolecular films of fatty acids, amines, or alcohols are also lubricating (Cottington *et al.*, 1958). Such lubricating layers of these compounds can also be formed by adsorption from their solution in a base oil. These compounds chemisorb by chemical interaction between their functional groups (COOH, NH_2, or OH) and metal oxide surfaces. The higher the carbon number of the alkyl chain in the organic compound, the lower the friction and the longer the durability (Cottington *et al.*, 1958). Multilayers give greater durability to the lubricant film than monolayers. The lubricity disappears at elevated temperatures (100–300 °C depending on the compounds), because of desorption of the organic compound. When surfaces are covered with solid organic layers, friction increases sharply above the melting point of the surface compounds (Rabinowicz and Tabor, 1951).

2.3.2 Lubrication with Reacted Layers Organic deposits, known as friction polymers, are formed at sliding contacts from lubricant components and organic vapors in the surrounding gas. The polymerized matter acts as a lubricant film (Fein and Kreuz, 1965). At elevated temperature, organic compounds decompose and carbonaceous matter is formed on the surface; this acts as a solid lubricant. Oxide layers on metal surfaces reduce the wear rate and prevent metal–metal contact. Such layers are reformed by the oxidation of nascent surfaces by atmospheric oxygen and water. Lubrication with reacted layers from additives is shown in Sec. 2.3.5.

2.3.3 Chemomechanical Effect The mechanical properties of materials are altered by the surface effects of chemicals. The mechanical strength of metals such as tin and ionic crystals is reduced by chemisorption of polar compounds (known as the Rehbinder effect). The ductility of alkali halides such as potassium chloride is reduced by the presence of water on the solid surface (Joffe effect). On the other hand, the surface hardness of metals such as cadmium is increased by the presence of an oxide layer on the surface (Roscoe effect). Since tribological behavior depends on the mechanical properties of the materials, these effects directly affect the tribological behavior. For example, friction and wear of magnesia and alumina are affected by the adsorption of *n*-alcohols (Westwood *et al.*, 1981).

2.3.4 Degradation of Lubricants Lubricating oils degrade mainly through thermal oxidation. The reaction is accelerated at elevated temperature under sliding conditions and is catalyzed by metal surfaces and metal ions formed by the oxidation of metals. Degraded oil contains metal ions, alcohols, fatty acids, soaps, and esters as oxidation products. The viscosity of oils decreases after decomposition. On the other hand, some of a lubricant oil is polymerized during the oxidation, resulting in an increase in its viscosity. Since the oxidation occurs through a radical transfer mechanism, radical scavengers such as amines are used as antioxidation agents. The role of nascent surfaces and exoelectrons in the initiation of degradation has also been pointed out (see Sec. 2.2.3).

2.3.5 Effects of Additives Adsorption and reaction layers on materials play an important role in reducing friction and wear as discussed above. Several types of additives are added to mineral or synthetic base oils. Oiliness agents having a polar group such as a carboxyl group chemisorb on the oxide surface on metals. Extreme-pressure additives containing such elements as sulfur, phosphorus, and chlorine react with material surfaces under severe conditions, resulting in the formation of a soft layer that acts as a lubricating solid film to prevent metal–metal contact (Sakurai and Sato, 1966). In general, these additives have a long-chain alkyl group in order to dissolve in base oils. Lubricating performance is affected by the surface coverage of additives that increases with the concentrations of additives. Chemical or corrosive wear occurs at higher concentrations of additives that attack materials chemically. When additives react with materials readily, wear is also accelerated chemically. Therefore, there is an optimum concentration or reactivity of additives for a lubricant mixture as shown in Fig. 7.

3. TRIBOMATERIALS

3.1 General Aspects

"Tribomaterials" include sliding materials facing each other and coatings of solid or liquid films between the sliding materials. They are required to reduce, or in some cases to control, adhesion, friction, and wear of mechanical components. During sliding, they are subjected to mechanical, thermal, and chemical attack depending on the operational conditions and environments. Friction and wear of tribomaterials are therefore very complicated phenomena involving various mechanisms and factors. Suitable selection of tribomaterials depending on the operating conditions plays an important role in controlling the life of mechanical components.

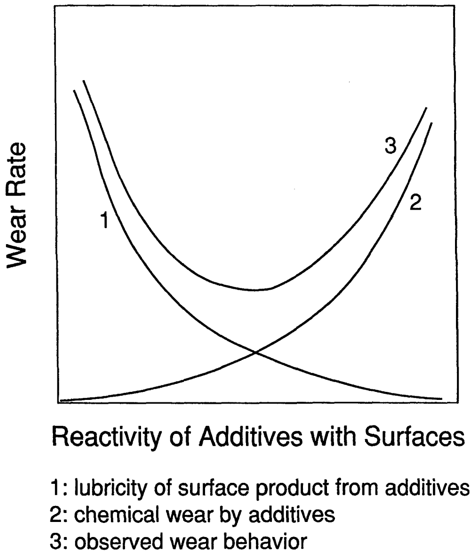

FIG. 7. Relationship between wear rate and additive reactivity with surfaces.

3.2 Liquid Lubricants

Liquid lubricants reduce or control friction and wear by separating sliding solids with a liquid layer. They also act as heat carriers to remove heat generated by sliding. Liquid lubricants are composed of a base oil and additives.

Base oils are classified as mineral or synthetic oils. Mineral oils, a mixture of paraffinic, naphthenic, and aromatic hydrocarbons, are produced from petroleum by distilling and refining. Several types of synthetic oils can be used to improve tribological performance such as viscosity index, thermal and oxidative stability, and volatility. Typical synthetic oils are hydrocarbons (olefin oligomers and arylalkanes), esters (diesters and polyol ester), halogenated organic compounds (perfluorinated ethers and chlorinated compounds), phosphonates, polyglycols, polyethers (polyphenyl ethers), and silicon derivatives. The tribological performance of synthetic oils closely depends on their molecular structure.

Lubricant additives are mainly composed of organic compounds having different functional groups. Oiliness agents having a polar functional group such as a carboxylic group chemisorb on material surfaces and reduce friction and wear. Carboxylic acids such as stearic acid, esters, and soaps are used as oiliness agents. Extreme-pressure (e.p.) additives having such elements as phosphorus, sulfur, and chlorine react with materials to form lubricating films. Organic sulfides, phosphates, and chlorides—for example, dibenzyl disulfide, tricresyl phosphate (TCP), and chlorinated paraffine—are used as e.p. additives. A typical e.p. additive is zinc dialkyl dithiophosphate. Viscosity-index improvers, high-molecular-weight polymers such as polymethacrylate (PMA), improve viscosity at elevated temperature. Oxidation inhibitors such as hindered phenol inhibit a radical transfer in the oxidation of oils. Sometimes solid lubricants are added to liquid lubricants.

Greases are composed of a base oil and fine solid particles called a thickener. A typical thickener is soap such as lithium stearate that forms a fibrous micelle. Urea compounds, fine polymer powders, and surface-treated clay powders are also used as thickeners. The base oil is held in the three-dimensional structure of the thickener. Since grease has a solidlike structure that is easily softened by mechanical shear, it has several advantages: It is easy to supply without maintenance for a long time, and it protects sliding contacts from water, dust, and corrosion.

3.3 Solid Lubricants

Solid lubricants are inorganic and organic compounds of low shear strength. Typical inorganic lubricants are molybdenum disulfide, tungsten disulfide, graphite, and carbon fluoride, which all have a hexagonal lamellar structure; lead oxide (PbO) and barium fluoride/calcium fluoride eutectic, BaF_2/CaF_2, which are nonlaminar solids; and soft metals, such as lead, silver, and gold. Typical organic lubricants are fats, soaps, waxes, and polymers, such as polytetrafluoroethylene (PTFE) and polyimide. These solid lubricants are utilized, normally in the form of coatings or self-lubricating composites, at elevated temperatures, say above 300 °C, where liquid lubricants thermally degrade (Sliney, 1982).

Graphite and molybdenum disulfide, which are the most widely used solid lubricants, have different tribological responses to the environment. Contaminants in air—water and hydrocarbons—adsorb well on molybdenum disulfide and increase the bonding strength between two adjacent lattice layers. Thus molybdenum disulfide lubricates better in the absence of these contaminants, such as in a vacuum or in dry inert gases. Conversely, graphite has better lubricative properties in ambient air than in vacuum or in dry air.

Transition-metal dichalcogenides, such as niobium disulfide and tantalum disulfide, are not themselves lubricative, even though they have a layered lattice crystal structure, but they show better lubricity after intercalation of guest elements, such as silver atoms, between adjacent lattices (Hironaka, 1992).

PTFE appears in the "Guiness Book of Records" as the solid with the lowest friction coefficient (0.04).

The temperature dependences of the friction of typical solid lubricants are shown in Fig. 8. Molybdenum disulfide, graphite, and PTFE lose their lubricative properties through decomposition or oxidation degradation. On the other hand CaF_2/BaF_2 and boron oxide (B_2O_3) acquire lubricative properties at elevated temperatures as a result of softening or melting.

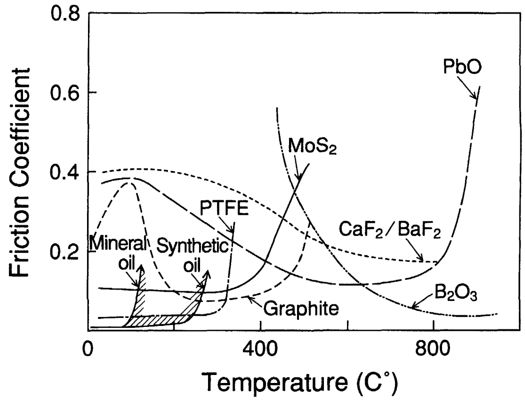

FIG. 8. Friction coefficient of lubricants as a function of temperature in air.

3.4 Wear-Resistant Sliding Materials

Wear generally occurs at real contacts between the sliding surfaces, where the stress is concentrated. Since plastic contact promotes wear, a tribological design is required to ensure that surfaces in contact are elastic. This requirement is achieved by reducing the plasticity index ψ in Eq. (2). The E'/H in the quantity ψ for pure metals is about 200, whereas for inorganic solids having covalent bonds, it is about 10; conventional materials have values between 10 and 200. Generally, the harder the material, the lower the value of E/H. Harder materials, however, are generally brittle, which leads to destructive wear. Wear-resistant tribomaterials are thus designed as multiphase solids or composites, combining hard solid phases in a ductile solid matrix of metals or polymers.

One example is a high-speed tool steel having a heat-treated martensite phase in which hard carbide phases such as $M_{23}C_6$, M_6C, or MC (M is W, Cr, or V) are precipitated. Nonferrous superalloys have been developed for wear resistance at high temperatures: for example, Triboloy® is essentially a nickel-based superalloy containing very hard intermetallic compounds (Laves phase). Stellite® is a Co–Cr based superalloy containing chromium and tungsten carbides.

Many polymers are used for mechanical components: polyamide (PA; Nylon®), polyoxymethylene (POM), polycarbonate (PC), modified polyphenylene oxide (PPO), polybutylene terephthalate (PBT), and polytetrafluoroethylene (PTFE, Teflon®). These are known as engineering plastics, whereas polyimide (PI), polyether ether ketone (PEEK), polyphenylene sulfide (PPS), and liquid crystalline polymer are known as super engineering plastics. They are superior to engineering plastics in heat resistance and other mechanical properties. PA, POM, and PBT are self-lubricative and wear resistant. PTFE is self-lubricative, but has high wear. The wear resistance can be improved by using PTFE composites, containing various fillers such as bronze mesh or carbon fiber.

Ceramics, such as silicon nitride, silicon carbide, alumina, and zirconia, are hard, low–thermal-expansion, noncorrosive, and heat-resistant materials. They therefore have potential as tribomaterials for use in extreme conditions and hostile environments, where components made of metals and polymers cannot operate. Some species in the environment form films on sliding surfaces of these ceramics as a result of tribochemical reactions, as described in Sec. 2. For instance, the friction and wear of silicon nitride were reduced by additives such as phosphorus-containing compounds, sulfonates, and glycol compounds in a paraffin-based oil but were increased by chlorinated compound additives (Gates and Hsu, 1991). The former additives are thought to form a lubricative film in the contact region. Silicon nitride also reacts with water, methanol, and ethanol to form tribochemical films of SiO_2 and $Si(OH)_x$ in water (Tomizawa and Fischer, 1986), SiO_2, amines, and polymeric compounds in n-alcohols (Hibi and Enomoto, 1989). In these fluids, hydrodynamic lubrication sets in above a certain sliding velocity.

Diamond, the hardest material, and diamond coating have a low coefficient of friction of between 0.05 and 0.15 in air. Water is effective for lubricating diamond on diamond, but oil is not (Feng and Field, 1991).

3.5 Coating and Surface Modifications

Coating and surface modifications have been widely used in tribological applications. There are two ways to reduce sliding wear. One is to reduce the stress concentration in the contact region. This can be achieved by using soft, more or less lubricative solids for coating materials. Another way is to enhance the resistance against deformation and fracture in the contact region, which can be done by using hard materials as an overlayer (Ura *et al.*, 1989).

Typical hardnesses of conventional coating materials and surfaces modified for tribological use are shown in Fig. 9. The processing temperatures fall into three main groups: at around 1000 °C, thermodynamic chemical reactions proceed on the substrate, and adhesion between film or modified layer and substrate is high. At around 400–500 °C are ion plating and plasma sputtering. Adhesion strength can be enhanced by ion cleaning a substrate by sputtering prior to coating. For other coatings at process temperatures below about 500 °C, substrates are usually sand-blasted for cleaning and roughening to enhance the adhesion of overlayers.

Optimum coating thicknesses of triboma-

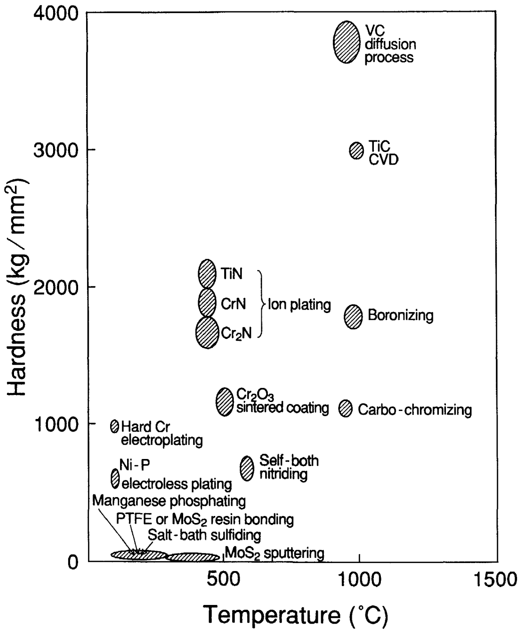

FIG. 9. Hardness of coatings as a function of substrate temperature for coatings.

terials will vary depending on whether the material is soft or hard. For soft layers, the basic tribological concept is that vertical loads are supported by the hardness of the substrate materials and soft layers lower the shear resistance. Therefore, the thinner the soft layers, the better the tribological response should be. An initial thickness of 3–15 μm is common. On the other hand, a hard coating or overlayer is designed to support the load. In the first approximation, the thickness of the hard layer is selected to be greater than the depth of the maximum shear stress, which depends on the operational conditions.

4. SUMMARY

Tribology is a key technology for mechanical systems and also an important area of surface science and material science, because science is important for understanding the origin of friction and wear. Thus, tribology ranges from basic physics and chemistry to mechanical engineering.

In this article, fundamentals of tribology were overviewed in three areas: tribophysics, tribochemistry, and tribomaterials. Tribophysics involves deformation, adhesion, friction, and wear mechanisms, which are described by the dynamics of continuum bodies. Tribochemistry involves chemical reactions on sliding surfaces. And many kinds of tribomaterials for sliding surfaces have been developed. These tribological knowledge have been applied to design many kinds of machine elements.

However, tribological phenomena are very complicated. All these areas (i.e., tribophysics, tribochemistry, and tribomaterials) should be understood by analyzing sliding mechanisms of surfaces. Not only engineering data but also scientific understanding are needed for further advance in tribology.

GLOSSARY

Boundary Lubrication: Lubrication without a liquid layer generating hydrodynamic pressure to sustain the load. The friction is affected by the physicochemical interaction at the solid-lubricant-solid interface.

Chemisorption: Adsorption that occurs through a chemical interaction between molecules (adsorbates) and surfaces. The heat of chemisorption is more than 100 kJ/mole.

***d*-Bond Character:** The amount of *d*-electrons that contribute to metal bonding in the *dsp* hybrid orbital. The greater the *d*-bond character of a structure, the more stable it is (see Pauling, 1949).

Exoelectrons: Electrons emitted from activated surfaces by mechanical treatments, irradiation of high-energy particles, and so on. One of the active sources for tribochemical reactions.

Extreme-Pressure Additives: Lubricant additives containing phosphorus, sulfur, and/or chlorine, and some metals such as zinc and molybdenum, to form surface-lubricating films by reaction with materials.

Flash Temperature: High temperature generated at mechanical contact exceeding several hundred degrees centigrade.

Nascent Surface: Chemically active surfaces formed by the removal of surface films of oxides and organic contaminants at mechanical contact.

Oiliness Agent: Surface-active agents that have long alkyl chains and polar groups, such as a carboxyl group, and that adsorb on the material surface and reduce friction and wear.

Poisson's Ratio: The ratio of the relative lateral contraction to the relative longitudinal extension. For most substances, this value lies between 0.2 and 0.4.

Viscosity Index: Viscosity decreases with increasing temperature. The viscosity index is a parameter for expressing the temperature dependence of viscosity. The higher the index, the lower the dependence.

Works Cited

Ashby, M. F., Lim, S. C. (1990), *Scripta Metall. Mater.* **24,** 805–810.

Bowden, F. P., Thomas, P. H. (1954), *Proc. R. Soc.* **A223,** 29–40.

Buckley, D. H. (1974), *ASLE Trans.* **17,** 36–43.

Buckley, D. H. (1981), *Surface Effects in Adhesion, Friction, Wear, and Lubrication,* Amsterdam: Elsevier, Chap. 6.

Cottington, R. L., Shafrin, E. G., Zisman, W. A. (1958), *J. Phys. Chem.* **62,** 513–518.

Dowson, D. (1962), *J. Mech. Sci.* **4,** 159–170.

Fein, R. S., Kreuz, K. L. (1965), *ASLE Trans.* **8,** 29–38.

Feng, Z., Field, J. E. (1991), *Surface Coatings Technol.* **47,** 631–645.

Ferrante, J. (1987), NASA TM100198.

Fukui, S., Kaneko, R. (1988), *ASME J. Tribology* **110,** 253–262.

Gates, R. S., Hsu, S. M. (1991), *Tribology Trans.* **34,** 417–425.

Goldblatt, I. L. (1971), *Ind. Eng. Chem., Prod. Res. Develop.* **10,** 270–278.

Greenwood, J. A., Williamson, J. B. P. (1966), *Proc. R. Soc.* **A295,** 300–319.

Groszek, A. J. (1962), *ASLE Trans.* **5,** 105–113.

Hibi, Y., Enomoto, Y. (1989), *Wear* **133,** 133–145.

Hirano, M., Shinjo, K. (1990), *Phys. Rev. B* **47,** 11837–11851.

Hironaka, S. (1992), *Jpn. J. Tribology* **37,** 549–556.

Holm, R. (1967), *Electric Contacts,* New York: Springer-Verlag.

Israelachvili, J. (1992), *Intermolecular and Surface Forces,* 2nd ed., London: Academic, Chap. 10.

Israelachvili, J., McGuiggan, P., Gee, M., Homola, A., Robbin, M., Thompson, P. (1990), *J. Phys.: Condens. Matter* **2,** SA89–SA98.

Johnson, K. L., Kendall, K., Roberts, A. D. (1971), *Proc. R. Soc. London* **A324,** 301–313.

Kajdas, C. (1987), *Wear* **116,** 167–180.

Kaneko, R., Hamada, E. (1993), *Wear* **162–164,** 370–377.

Kramer, J. (1957), *Acta Phys. Austriaca* **10,** 327–338.

Kurokawa, N., Sakaguchi, M., Sohma, J. (1978), *Polymer J.* **10,** 93–98.

Landman, U., Luedtke, W. D., Burnham, N. A., Colton, R. J. (1990), *Science* **248,** 454–461.

Mayer, E., Overney, R., Brodbeck, D., Howald, L., Luthi, R., Frommer, J., Guntherrodt, H.-J. (1992), *Phys. Rev. Lett.* **69,** 1777–1780.

Morecroft, D. W. (1971), *Wear* **18,** 333–339.

Mori, S., Imaizumi, Y. (1988), *STLE Trans.* **31,** 449–453.

Owens, R. S., St Pierre, L. E., Klint, R. V. (1966), *Wear* **9,** 160–168.

Pauling, L. (1949), *Proc. R. Soc.* **A196,** 343–362.

Rabinowicz, E., Tabor, D. (1951), *Proc. R. Soc.* **A208,** 455–475.

Sakurai, T., Sato, K. (1966). *ASLE Trans.* **9,** 77–87.

Schrader, R., Wissing, R., Kubsch, H. (1969), *Z. Anorg. Allg. Chem.* **365,** 191–198.

Sliney, H. E. (1982), *Tribology Int.* **15,** 303–315.

Smith, H. A., Fort, T. Jr. (1958), *J. Phys. Chem.* **62,** 519–527.

Suh, N. P. (1977), *Wear* **44,** 1–16.

Timoshenko, S., Goodier, J. N. (1951), *Theory of Elasticity,* 2nd ed., New York: McGraw-Hill, Chap. 13.

Tomizawa, H., Fischer, T. E. (1986), *ASLE Trans.* **30,** 41–46.

Ura, A., Enomoto, Y., Kimura, Y., Nishimura, M. (1989), *Jpn. J. Tribology* **34,** 517–523.

Wade, W. H., Hackerman, N. (1964), *Adv. Chem. Ser.,* no. 43, 222–231.

Westwood, A. R. C., Ahearn, J. S., Mills, J. J. (1981), *Colloids Surf.* **2,** 1–35.

Wheeler, D. R. (1976), *J. Appl. Phys.* **47,** 1123–1130.

Whitehouse, D. J., Archard, J. F. (1970), *Proc. R. Soc. London* **A316,** 97–121.

Yoshizawa, H., Chen, Y.-L., Israelachvili, J. (1993), *Wear* **168,** 161–166.

Further Reading

Bhushan, B. (Ed.) (1994), *Handbook of Micro/Nanotribology,* Boca Raton, FL: CRC Press.

Bowden, F. P., Tabor, D. (1950), *The Friction and Lubrication of Solids,* Oxford, U.K.: Clarendon Press.

Bowden, F. P., Tabor, D. (1964), *The Friction and Lubrication of Solids,* Pt. 2, Oxford, U.K.: Clarendon Press.

Czichos, H. (1978), *Tribology,* Amsterdam: Elsevier Scientific Publishing Company.

Dorison, A., Ludema, K. C. (1985), *Mechanics and Chemistry in Lubrication,* Amsterdam: Elsevier.

Gross, W. A. (1962), *Gas Film Lubrication,* New York: Wiley.

Heinicke, G. (1984), *Tribochemistry,* München: Carl Hanser.

Peterson, H. B., Winner, W. O. (Ed.) (1980), *Wear Control Handbook,* New York: ASME.

Pinkus, O., Sternlict, B. (1961), *Theory of Hydrodynamic Lubrication,* New York: McGraw-Hill.

Rabinowicz, E. (1965), *Friction and Wear of Materials,* New York: Wiley.

TUNABLE DYE LASERS

ANDREW DIENES, *Department of Electrical and Computer Engineering, University of California, Davis, California, U.S.A.*

DIEGO R. YANKELEVICH, *Centro de Investigación Científica y de Educación Superior de Ensenada, Ensenada, México*

Introduction 299
1. **Basic Dye-Laser Principles** .. 299
2. **cw Dye Lasers** 304
2.1 Theory 304
2.2 Practical cw Dye Lasers 307
2.2.1 The Laser Resonator 307
2.2.2 The Flowing-Dye Cell and Triplet Quenching 308
2.2.3 Mirror Transmission 309
2.2.4 Performance Limitations 310
2.2.5 Dyes and Pump Sources 310
2.2.6 Bandwidth Control and Tuning 312
2.3 Mode-Locked cw Dye Lasers .. 315
3. **Pulsed Dye Lasers** 318
3.1 Flashlamp-Pumped Dye Lasers 318
3.1.1 Theoretical Description and Discussion 319
3.1.2 Practical Performance Limitations of Flashlamp Systems 323
3.2 Pulsed-Laser–Pumped Dye Lasers 323
3.3 Tuning and Bandwidth Reduction 327
4. **Application of Dye Lasers** ... 330
Acknowledgements 332
Works Cited 332
Further Reading 334

INTRODUCTION

Dye lasers are among the most widely used tunable coherent sources of light from the near UV to the near IR. Although at certain specific wavelength ranges other sources have eclipsed their use, they continue to play an important role because of their availability in many possible configurations at any wavelength and at various pulse durations and output power levels. In addition, the relative simplicity and affordability (of most of the available configurations) and, most importantly, true tunability assures their continuing importance for some time to come. In this article the physical and theoretical principles of dye lasers are given and some of the most important types of systems in use today are described. It is not a comprehensive theory or a complete survey; the aim is to give an understanding of the basics and some knowledge of the practical aspects of the most important types of tunable dye laser systems. Additional, related topics can be found in the Further Reading.

1. BASIC DYE-LASER PRINCIPLES

The active materials in dye lasers are various organic dyes, which can have widely different chemical compositions. However, the basic photophysical principles are essentially the same for all of them. A description based on rate equations and stimulated absorption and emission cross sections is commonly used, and the relevant parameters have somewhat similar values. The radiative properties of laser dye solutions depend on both the dye molecular structure and the properties of the solvent. There is extensive literature (Shank, 1975; Maeda, 1984; Duarte and Hillman, 1990; Schäfer, 1990) on dye lasers and on the spectroscopic and chemical properties of dyes. No detailed description of these is required for a basic understanding of the dye laser, and only a brief summary of the relevant properties is given here.

Dye molecules are large, with many internal degrees of freedom. There are also very frequent collisions with the solvent molecules. Figure 1(a) shows the basic energy-level scheme of a typical laser dye in liquid

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

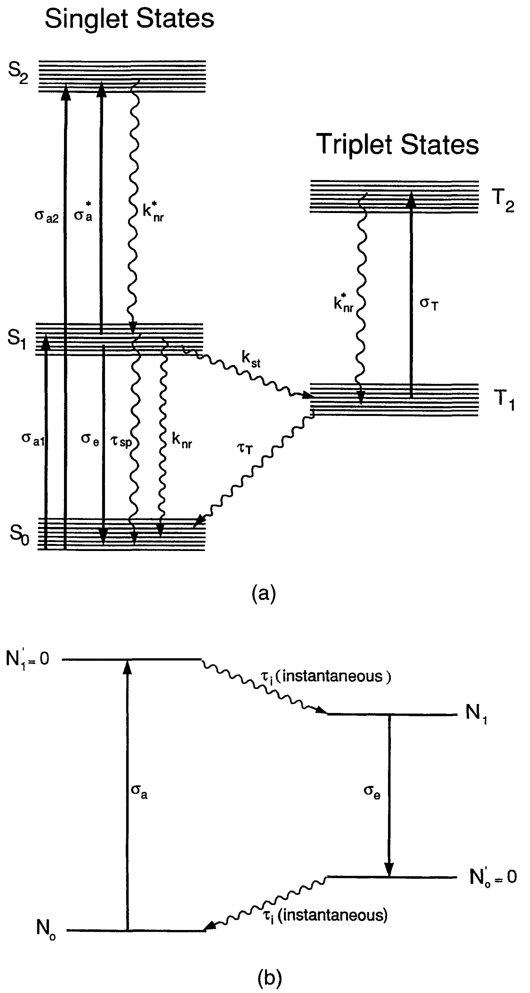

FIG. 1. (a) Schematic of the energy levels of a typical laser dye showing singlet and triplet manifolds, spontaneous rates (lifetimes), and stimulated cross sections. The nonradiative rates shown as k_{nr}^* are so fast as to quench radiative emission from those levels. (b) An equivalent four-level model of the two lowest singlet states (S_0,S_1) in (a). The primed levels have zero population.

solution form. There are two main energy manifolds: singlets and triplets. Within each electronic state there is a broad-band quasi-continuum of vibrational–rotational energy levels. The main absorption band is that corresponding to the ground-state singlet (S_0) to first excited singlet (S_1); there is also a shorter-wavelength and somewhat weaker band from S_0 to S_2. Laser action takes place between the first excited singlet (S_1) and the ground-state singlet (S_0) bands. Only the S_1 band is radiative, because a molecule in any of the higher singlet levels undergoes rapid nonradiative decay to S_1 completely quenching any emission. Nonradiative decay from S_1 to S_0 and intersystem crossing between singlet and triplet states also take place, but for laser dyes the radiative process is dominant. In the triplet energy manifold, the lowest state (T_1) is metastable while the higher-energy states are also nonradiative. As a result of the nonradiative quenching of all except the S_0, S_1, and T_1 states, only these three bands have non-negligible populations in a dye laser. Another important process is the essentially instantaneous intraband thermalization that takes place within the vibrational–rotational continuum bands. As a consequence, the total population within any band is always in quasi thermal equilibrium. The strengths of the various spontaneous (radiative and nonradiative) processes are given as lifetimes (τ_{sp}, τ_T, τ_i, etc.) or rates (k_{ST}, k_{nr}, etc.) as shown on the diagram. The total lifetime of state S_1 is τ_1, where $1/\tau_1 = 1/\tau_{sp} + k_{nr} + k_{ST}$, and is typically of the order of a few nanoseconds. Optical pumping, to achieve an effective population inversion, is usually in the main singlet absorption band. However, since the S_2 state decays very rapidly (and nonradiatively) to S_1, the secondary (S_0 to S_2) absorption band can also be used to populate the S_1 state.

The singlet manifold may be modeled as an effective four-level homogeneously broadened system, as shown in Fig. 1(b). Homogeneous means that the entire population of a given band (level) participates in the stimulated transitions from that level. The broad emission and absorption line shapes are always determined experimentally. The probability of a stimulated process is $I\sigma$, where I is photon intensity (in s^{-1} cm^{-2}) and σ is process cross section. Figure 2 shows the cross sections (σ_e, σ_a, σ_T) vs wavelength for one of the best laser dyes, rhodamine 6G (R6G), and the inset gives approximate values of some of the lifetimes and rates under conditions found in a typical laser. The main absorption-band and the emission-band curves are approximately mirror images of each other. This arises from a superposition, with Boltzmann probability distribution, of Lorentzian-shaped transitions of various oscillator strengths. The peak of the emission is energy downshifted from that of the ab-

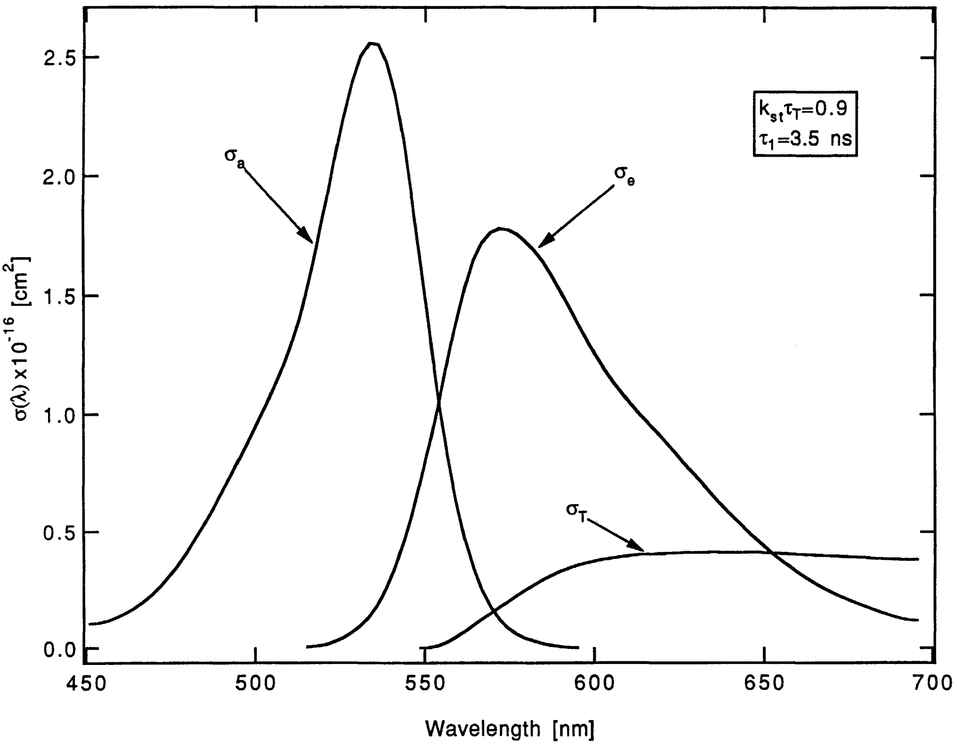

FIG. 2. Singlet absorption and emission and triplet absorption cross sections [defined by Eqs. (1) and (4)] vs wavelength for the dye R6G. The inset box shows values of the other parameters used in the calculations of Fig. 8.

sorption on account of the Franck-Condon principle. In some dyes this shift may be so large that the emission and absorption curves have only a very small overlap. The emission cross section can be obtained from the spontaneous fluorescence line shape $E(\lambda)$ (given vs the wavelength λ) by

$$\sigma_e = \lambda^4 E(\lambda)/8\pi\tau_1 c n^2, \tag{1}$$

where

$$\int_0^\infty E(\lambda)d\lambda = \phi, \tag{2}$$

n is the index of refraction, and c is the speed of light. The quantity $\phi = \tau_1/\tau_{sp}$ is the quantum efficiency of the fluorescence, and also equals the ratio of the areas under the emission and absorption curves. This parameter must obviously be as close to unity as possible for a laser dye. The nonradiative processes, i.e., k_{nr}, k_{ST}, and τ_T, are dependent on the molecular structure of the dye, on the dye concentration, and on impurities (or deliberate additives) present in the solution. For example, in R6G for concentrations larger than ~0.5×10^{-3} mol/L, τ_1 gradually decreases from the low-concentration limit because of increasing nonradiative decay (concentration quenching). It should also be noted that a value of ϕ close to unity is necessary for, but does not always guarantee, lasing. This is because the excited-state (S_1 to S_2) absorption (σ_a^* in Fig. 1) can overlap the emission and can prevent lasing. In dyes suitable for lasing, this absorption lies outside the emission band. On the other hand, the triplet absorption always at least partially overlaps the emission and is therefore an important loss factor that is discussed later. It is obvious from Figs. 1 and 2 that even if only a rather small fraction of the total molecules is in the first singlet level, a dye has an effective population inversion and therefore optical gain in the long-wavelength region of the emission. Additionally, because the emission spectrum is very broad, the gain bandwidth is wide. The resulting tunability, however, comes at a price. The high stimulated-emission probability (large σ_e) over a wide band also means that the lifetime of the lasing level is very short, of the order of a few nanoseconds. The consequence of this is that high pumping intensities are required to obtain and maintain a population in level S_1.

A very large number of dyes have been made to lase by various optical pumping schemes. However, for the purposes of practical applications, a much smaller number are used. Dye lasers may be divided into two broad categories by the method of pumping: cw and pulse pumped. The first is always pumped by a laser, while the second may be either flashlight or laser pumped. The pumping and lasing configurations are very different for these two groups. Nevertheless, the same simple basic gain equation applies to all of them. In what follows, the subscripts 0, 1, and T refer to the corresponding levels in Fig. 1(a), and the subscript l on a cross section indicates its value at the lasing wavelength, while the subscript p refers to the pump and its wavelength. The differential gain for a forward-traveling laser beam I_l^+ is

$$g = \frac{1}{I_l^+}\frac{dI_l^+}{dz} = \sigma_e N_1 - \sigma_T N_T - \sigma_{al} N_0, \qquad (3)$$

where N_i are population densities (in cm^{-3}). The total dye molecular density is $N = N_0 + N_1 + N_T$. Figure 3 shows the gain vs wavelength for R6G, assuming $N = 3 \times 10^{17}$ cm^{-3}, and various fractions in the three different levels. The most important point to observe is that, because of the displacement of the singlet absorption and emission spectra (Franck-Condon shift), gain exists in a broad spectral region even when the population of the first excited singlet state is much smaller than that of the ground state. The characteristic sharp dropoff of the gain (to loss) on the short-wavelength side is due to the singlet absorption, which partially overlaps the emission spectrum. In some dyes this gain cutoff may be considerably less sharp than seen in this example, in consequence of a larger Franck-Condon shift and a more gradual decrease of the emission. The second important characteristic of these curves is that, because the triplet absorption overlaps the singlet emission, the gain depends very strongly on the population of the lowest triplet state. The gain is drastically reduced and lasing will be prevented if the triplet state population is of the same order as that of the excited singlet. Unfortunately, the lowest triplet state is metastable, i.e., its decay rate τ_T is slow, and with time the triplet-state population tends to accumulate. In the descriptions of the various types of dye lasers, it is shown how this triplet absorption problem is partially solved.

The design and operation of all dye lasers can be best understood by studying the dynamics of the gain through standard rate

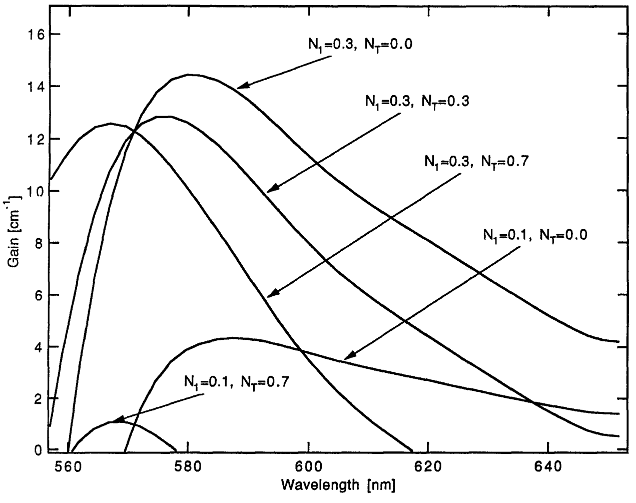

FIG. 3. Differential gain vs wavelength for the dye R6G, with various fractions of the total population in the excited singlet and in the triplet states.

equations for the population densities N_i. The two basic geometrical configurations used are collinear (or nearly collinear) pumping and transverse pumping, corresponding to the pump light being nearly in line with or perpendicular to the laser beam in the cavity, respectively. For either of these configurations, at some arbitrary point inside the laser medium (within the cavity) the population-rate equations may be written as

$$\frac{dN_1}{dt} = N_0\sigma_{ap}I_p - \frac{N_1}{\tau_1} - N_1\sigma_e I_l + N_0\sigma_{al}I_l, \quad (4)$$

$$\frac{dN_T}{dt} = k_{ST}N_1 - \frac{N_T}{\tau_T}. \quad (5)$$

Here, I_l is the *total* laser photon intensity (in a linear cavity, $I_l = I_l^+ + I_l^-$, two opposite traveling waves, while $I_l = I_l^+$ in a unidirectional ring cavity), and I_p is the pump photon intensity, each defined according to their respective directions of propagation in the dye medium. The dye-laser photon density also obeys time-dependent rate equations. However, for many dye lasers the cavity decay rate is considerably shorter than the pulse duration and the buildup time of the oscillations can be neglected. In any case, the main characteristics of most dye lasers can be understood without considering cavity dynamics. (An exception is discussed briefly in Sec. 3.2). Equation (4) is valid for laser pumping, but it is also approximately correct for a broad-band pump source (such as a flashlamp), provided that the absorption cross section used is a weighted average over the bandwidth of the pump light. Exact analytical solutions of Eqs. (3)–(5) are only possible for cw dye lasers, which are axially pumped. However, much can be learned for all dye lasers from examination of these rate equations. For the case of cw or long-pulse pumping, the solution of Eq. (5) is

$$N_T = k_{ST}\tau_T N_1, \quad (6)$$

which when substituted into (3) gives

$$g = \sigma_e f_T N_1 - \sigma_{al}N_0, \quad (7a)$$

where

$$f_T = 1 - k_{ST}\tau_T\sigma_T/\sigma_e. \quad (7b)$$

Even for wavelengths for which σ_{al} is zero, cw lasing is prevented if $f_T < 0$. The only remedy is triplet quenching, i.e., decreasing the triplet population by whatever means possible. Since it is very difficult to reduce the intersystem rate (k_{ST}), to minimize N_T the lifetime of the triplet (τ_T) must be decreased. Molecular oxygen, O_2, is a well-known triplet quencher. However, O_2 tends also to increase k_{ST}, and its concentration needs to be optimized. It has been found that for several good laser dyes, including R6G, atmospheric air has a nearly optimum O_2 concentration. Another method of triplet quenching is by the addition of certain organic additives to the dye solution. By contrast, it has also been shown that lasing is completely prevented if the dye solution is deoxygenated by bubbling pure nitrogen through it: τ_T increased to the order of milliseconds, resulting in a negative value of f_T for all wavelengths. For the case of pulsed dye lasers, accumulation of molecules in the triplet state depends on the pump pulse duration. For pump pulses of the order of or somewhat longer than the characteristic lifetimes ($1/k_{ST}$ and τ_T), the triplet population builds up in time, and therefore the factor f_T becomes a time-dependent quantity. Obviously, its value is always smaller than the cw steady-state value given by Eq. (7b). In other words, another solution for the triplet problem is to pump with short pulses. Pumping with light pulses of duration $\Delta\tau_p \ll 1/k_{ST}$ prevents significant accumulation in the triplet state. Another important point that can be learned from the above rate equations is that dye lasers require high pump intensities. From (4), just below threshold (i.e., no lasing), at wavelengths where singlet absorption is negligible, for the cw (or quasi cw) case,

$$N_1 = N_0 I_p \sigma_{ap}\tau_1. \quad (8)$$

At the other extreme, for very short pump pulses, the maximum value of N_1 is still given by (8), but with τ_1 replaced by $\Delta\tau_p$. In either case, as a result of the short lifetime of the lasing level, the required pump intensity is high. For example, to maintain $N_1/N_0 = 0.1$ in a R6G dye solution, pumping at $\lambda = 514$ nm requires $I_p = 1.8 \times 10^{23}$ photons s^{-1} cm^{-2}, or 6.9×10^4 W/cm^2.

A laser oscillator is made by placing the gain medium in a resonant cavity and coupling out some of the circulating coherent oscillation through one of the mirrors. The coupled-out fraction acts as a loss and, together with other losses inside the cavity, determines the cavity Q factor. The spectral characteristics of the output are determined by the properties both of the gain medium and of the cavity. A laser cavity has a large number of resonant frequencies that are separated by $\Delta f = c/2L$, where $2L$ is the total effective round-trip distance in the cavity. There are typically of the order of 10^6 or more resonant modes within the gain bandwidth. An important issue for any type of dye laser is control and tuning of the frequency of the laser oscillation. This is examined in separate sections for cw and pulsed lasers.

2. cw DYE LASERS

Probably the most important and most widely used dye lasers today are cw dye lasers. Although they require expensive high-power cw sources for pumping, and although their configuration can be very complicated, no other optical sources provide a comparable combination of tunability, resolution, and power. They cover a spectral range from ~365 to ~1000 nm, with more than 100 mW of power output possible through this range (more in the yellow–red region). This is adequate for most spectroscopic and other research applications. The available resolution, while not easy to obtain, is also very impressive. A true single-mode cw dye laser has a line width of ~0.5–1 MHz. A typical tuning range of $\sim 0.5 \times 10^{14}$ Hz available from a *single* dye thus can contain up to $\sim 10^8$ spectral resolution elements (linewidths), and a multiple-dye cw laser tuning range can contain more than ten times that. By use of standard frequency doubling and mixing techniques in nonlinear crystals, the tuning range has been extended down to 260 nm. Additionally, the broad gain bandwidth of a single dye makes possible ultrashort pulse generation by mode locking (see ULTRASHORT PULSE PHENOMENA), resulting in the shortest optical pulses available today (Fork *et al.*, 1987).

The high pump intensity required to achieve lasing necessitates tight focusing of the pump light into a very small active volume that in turn dictates a special cavity design and a flowing-dye cell. Although they may appear quite different, the physical configurations of nearly all the cw dye lasers used today are derived from the same basic folded-cavity design (Kohn *et al.*, 1971; Dienes *et al.*, 1972). Two commonly used versions, a three-mirror standing-wave cavity and a ring laser, are shown in Figs. 4 and 5. Both configurations include various frequency-tuning and bandwidth-control elements, which are discussed in Sec. 2.2.5. There are good reasons for the folded-cavity design. A brief simplified theoretical description greatly aids in the understanding of the basic operating principles and of the issues important for the design and operation of this device. With minor modifications, this simple theory can also be used to describe semiquantitatively the operation of most flashlamp pumped dye lasers (Sec. 3).

2.1 Theory

In what follows, it is assumed that the dye is pumped by a cw laser beam that is collinear with the lasing radiation as shown in a simplified schematic in Fig. 6. Since the transitions are homogeneous, the laser radiation can be considered to be monochromatic at some arbitrary wavelength within the emission band. It is assumed here that the pump and laser beams are plane waves

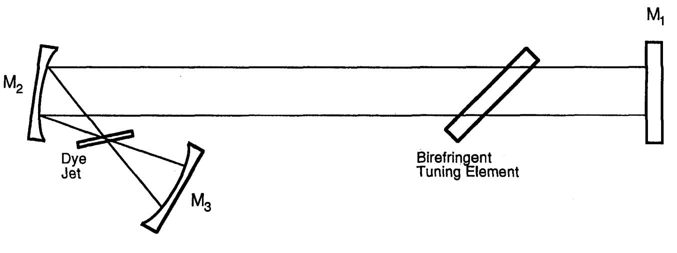

FIG. 4. A standing-wave–type folded three-mirror cavity for cw dye lasers. The various elements and their functions are described in the text.

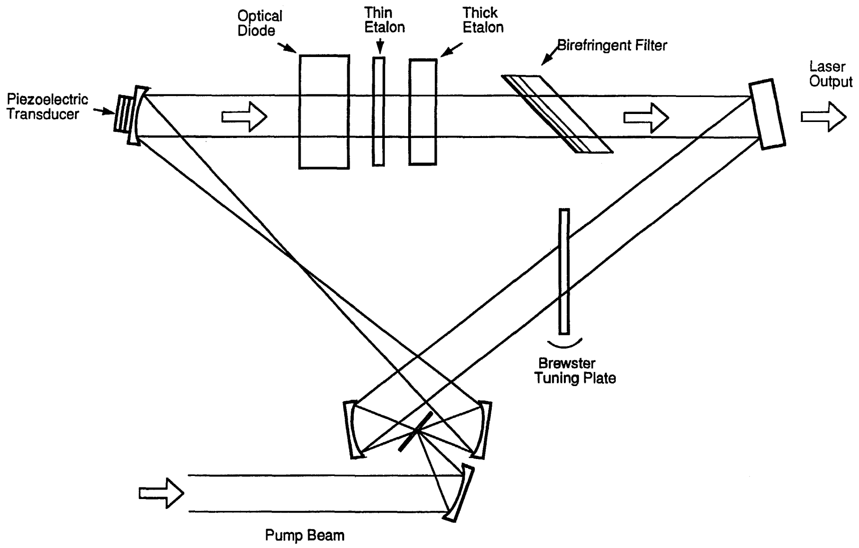

FIG. 5. A unidirectional traveling-wave ring cavity for cw dye laser, showing several frequency-control elements, which are described in the text.

with finite and equal areas. Thus in the rate equations (3)–(5) all the intensities and population densities are functions only of the longitudinal variable (z) in the medium. Solving (4) and (5) in the steady state, integrating over the length of the path (L) in the dye medium, and using the approximation that the total laser intensity I_l is constant with z, one obtains the total single-pass saturated gain $G = I_l^+(z = L)/I_l^+(z = 0)$,

$$\ln G = \sigma_e f_T \tau_1 \frac{1}{1 + I_l/I_s} I_{p0} a. \tag{9}$$

Here $I_s = 1/\sigma_e \tau_1$; the parameter a is the fraction of the pump beam absorbed in the dye cell, $I_{p0} = P_{p0}/A$, where P_{p0} is the input pump power and A is the cross-sectional area of the pump beam in the dye. Since the cavity Q is high, the required gain is small. Thus, N_1 is a small fraction of the total population, and therefore N_0 can be approximated as constant and equal to N. Then,

$$a = 1 - \exp(-\sigma_{ap} N L). \tag{10}$$

Equation (9) shows that the shorter the lifetime, the more pump intensity is required. The dye concentration (N), the pump absorption coefficient (σ_{ap}), and the cell thickness (L) only enter through the parameter a. In cw dye lasers, these quantities are adjusted so as to give $a \cong 0.9$ in a cell length that is determined by other factors, as is seen later.

Threshold is reached when the small-signal ($I_l \to 0$) gain equals the cavity loss. The gain is inversely proportional to the cross-sectional area of the excitation. Using parameters for R6G and $a = 0.9$, for a typical 15% round-trip gain ($G_0 = \sqrt{1.15} = 1.072$) at $\lambda_1 = 600$ nm, one finds a required pump intensity $I_p = 0.87 \times 10^5$ W/cm^2 at $\lambda_p = 514.5$ nm. This requirement dictates tight focusing of the pump beam and restricts possible pump sources. High gains cannot

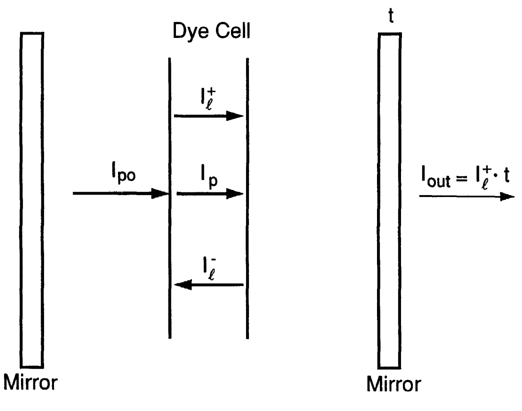

FIG. 6. Simplified schematic of the pumping and lasing geometry for the cw dye-laser theory.

be achieved, and thus the typical coupling loss is in the range of 5–20%. The first pump sources used, and still widely used, are cw ion lasers, argon and krypton. (Another, recently developed pump source is discussed in Sec. 2.2.4.) For a circular beam waist of radius ~20 μm, the above threshold corresponds to a pump power $P_p = 280$ mW. Since the confocal length corresponding to such a beam waist is ~1 mm, the dye cell must be shorter than that. Although early research used cells made of fused silica or glass, today practical cells are flowing dye jets with typical thickness of around 100–200 μm. For the above example, the dye molecular density N for 90% absorption in a 150-μm-thick cell is $1 \times 10^{18}/\text{cm}^3$, a concentration of 1.6×10^{-3} mol/L. At this concentration, increased nonradiative quenching begins to reduce both τ_1 and σ_e somewhat. Additional problems, such as solubility and aggregation of the dye molecules, also pose problems, requiring corrective measures in the form of special additives, such as surfactants to help disperse the molecules.

When pumping is done above threshold, the dye laser intensity builds up inside the cavity. The steady-state population N_1 is decreased by the stimulated emissions so that the gain saturates to the cavity loss (i.e., to the threshold gain). The population of the ground state (N_0) is thus clamped at its threshold value, and regardless of how strong the pumping is, the absorption of the pump does not decrease by depletion of the ground state but remains the *same* value as at threshold (Teschke *et al.*, 1976a; Chan, 1979). The output intensity is obtained by equating the round-trip gain to the round-trip loss and using $I_{\text{out}} = tI_l^+ \approx tI_l/2$ (where t is the transmission of the output mirror). The result, for the main part of the gain region where $\sigma_{al} = 0$, is

$$P_{\text{out}} = \eta[P_{p0} - P_{\text{th}}], \tag{11}$$

where the slope efficiency and threshold power are

$$\eta = \frac{\nu_l}{\nu_p} f_T a \frac{t}{t + l}, \tag{12a}$$

$$P_{\text{th}} = Ah\nu_p \frac{t + l}{af_T\sigma_e\tau_1}. \tag{12b}$$

Here, $1 - t - l = 1/G_0$ is the total round-trip cavity loss (l being the fractional residual loss other than the coupling-out loss). Typically, $t + l$ in actual systems is less than ~0.1. This equation shows a simple linear dependence of the output power on the pump. The threshold pump power depends linearly on the cavity losses and on the area, and inversely on the fraction of the pump power absorbed by the dye (a), on the emission cross section, and on the triplet loss factor (f_T). The slope efficiency, on the other hand, only depends on a, f_T, and the cavity losses, and of course on the energy ratio of the pump and laser photons.

This theory can be made more rigorous by including the effects of the singlet absorption losses (which are nonzero at the short-wavelength end of the emission band) and by considering the pump and laser beams to be fundamental Gaussian beams of generally different radii. Following the methods of Pike (1971) and others (Chan, 1979), it can be shown that Eq. (11) remains valid with

$$\eta = \frac{m}{\sqrt{1 + m^2}} \frac{\nu_l}{\nu_p} f_T a \frac{t}{t + l + 2s(1 - f_T)}, \tag{13a}$$

$$P_{\text{th}} = \frac{m + 1}{2} \pi w_p^2 h\nu_p \frac{t + l + 2s}{af_T\sigma_e\tau_1}. \tag{13b}$$

Here, s is the singlet absorption loss given by $s = -\ln(1 - a)[\sigma_{al}/\sigma_{ap}]$, w_p is the pump-beam radius, and $m = w_l^2/w_p^2$, the ratio of the areas. Singlet absorption adds to the losses in the cavity; the threshold pump power rapidly increases where σ_{al} is significant and lasing cuts off. As for the effects of unequal Gaussian pump and lasing beam sizes, we see that decreasing the pump-beam area relative to the lasing area increases the threshold but also increases the slope efficiency. For strongly pumped lasers a value of m between 1.5 and 2 is best, i.e., pump-beam area slightly smaller than the dye-laser-beam area. For the R6G example, pumped by the 514.5-nm Ar laser line, with $t = 0.057$, $l = 0.02$, $w_p = 20$ μm, and $m = 1.5$, at $\lambda_1 = 590$ nm we have $P_{\text{th}} = 1.3$ W and $\eta = 0.36$. Experimentally, similar values were achieved early in the history of the cw dye laser (Dienes *et al.*, 1972). A multiple-wavelength pump laser is often used in practice, since this gives more available pump power. This does not substantially alter the above theory. Although this theory was derived for a linear cavity, it is also valid for ring dye lasers, provided that the factor of 2 multiplying the singlet loss s is omitted (since the laser beam traverses the dye only once).

2.2 Practical cw Dye Lasers

The above theory gives a good description of cw dye lasers. In actual lasers there are some additional effects, which result in somewhat lower efficiencies than predicted by this theory. They are discussed briefly in Sec. 2.2.4. In practice, the parameters of most dyes are not known with sufficient accuracy to apply the theory quantitatively, and design procedures are usually carried out in a trial-and-error fashion (Yarborough, 1974; Johnston *et al.*, 1982). In this article the theory is used for a kind of "reverse engineering," through which the principles underlying the design and operation of currently used systems are explained. There are three types of cw dye lasers in general use today:

1. linear-cavity, moderately wide-bandwidth laser;
2. narrow-linewidth unidirectional ring laser (with single-mode operation possible); and
3. colliding-pulse mode-locked dye laser.

The design of these systems is now examined, beginning with consideration of the laser resonator.

2.2.1 The Laser Resonator The required very small mode area can be achieved by using a small cavity with concentric radii of curvature (R_i). This is how the first cw dye laser was made (Peterson *et al.*, 1970). However, for practical lasers a longer resonator (Kohn *et al.*, 1971; Dienes *et al.*, 1972) is desirable, with a well-collimated beam that either allows placement of elements for frequency tuning and control or allows an additional focal spot for passive mode locking (Ippen *et al.*, 1972) etc. The folded cavities shown in Figs. 4 and 5 fulfill these requirements. The three types of cw dye lasers listed above all use these, or variants of such cavities. The critical properties of this type of cavity can be understood by simple ray tracing in a three-mirror cavity as shown in Fig. 7(a). The parallel rays from mirror M_1 are focused to the focal point of mirror M_2. This is where the dye cell is placed (at Brewster's angle to minimize the internal losses). If this point is also the center of curvature of mirror M_3, then the diverging rays are returned upon themselves. Thus, for $R_3 = R_2/2 = f$, the folded arm has total length $d = R_2/2 + R_3 + \delta \cong R_2$, where δ is a small adjustment parameter of the order of a few microns. The cell is located in the middle of this folded arm. The length of the long arm d_1 is not critical but should be much longer than d, and $R_1 = \infty$ (flat mirror) or a large value. For a ring cavity R_2 and R_3 must be made equal when the cell is located in the middle (see Fig. 5) as is also obvious from ray tracing. Additional folded focusing arms

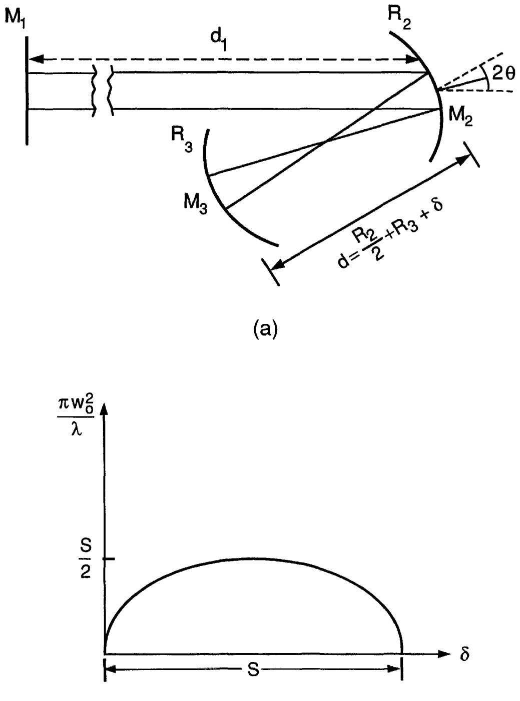

(a)

(b)

FIG. 7. (a) Simplified schematic of the three-mirror folded cavity showing mirror distances, and geometrical-optics ray tracing for the laser beam in the focusing fold and in the long collimation arm. (b) Minimum beam radius w_0 (in units of the confocal distance) vs the adjustment distance δ for the three-mirror laser cavity shown in (a).

can also be easily added (Ippen *et al.*, 1972) if needed. To find the actual mode parameters and the stability range of the resonator, standard Gaussian beam analysis is used (Kogelnik, 1966; Kogelnik and Li, 1966). The main results for the basic three-mirror version are given below.

The Gaussian "waist" size vs the position of the mirror M_3 (δ) is shown in Fig. 7(b). The stability range is $S = f^2/d_1$, and in the middle of this range the beam waist area is given by $\pi w_0^2/\lambda = S/2 = f^2/2d_1$. For example, letting $R_1 = \infty$, $R_2 = 10$ cm ($f = 5$ cm), $R_3 = 5$ cm, and $d_1 = 100$ cm yields $w_0 = 15$ μm, which is a typical value needed for lasing. In contrast, the beam in the long arm has a minimum beam radius of $w_0 \sim 0.6$ mm (on the flat mirror), a well-collimated beam. The above calculated beam waist radii are in free space. Inside the dye cell, after refraction into the Brewster-angle cell, the beam becomes elliptical and the area πw_p^2 is larger by a factor n (index of refraction of the cell). Increasing R_2 increases the waist area and will increase the threshold pump power since the pump beam must approximately overlap this waist. Consequently, on account of threshold requirements, R_2 cannot be much larger than 10 cm. On the other hand, practical geometrical considerations, as well as beam aberrations (coma and astigmatism), restrict R_2 to values larger than ~5 cm. All practical cw dye lasers use R_2 (and R_3) between 5 and 10 cm in the folded arm around the dye cell. Exact collinearity of the pump and laser beams is not essential since the very thin dye cell allows excellent overlap even with a substantial angle between the two beams. Independent focusing with a lens or a mirror is most commonly used, with the focal length about the same as or slightly smaller than that of the folding mirror, as shown in Fig. 4.

An additional feature of the folded, tightly focused cavity is that the tilting of the folding mirror introduces an astigmatism between the tangential (in the plane of the diagram) and sagittal (perpendicular) planes, due to the different effective focal lengths of the folding mirror M_3 in these two planes (Kogelnik *et al.*, 1972). This may greatly reduce the stability of the resonator by displacing the stability regions of the two planes with respect to each other, and it also results in elliptical-shaped beams. Interestingly, the Brewster-angle cell in the focal region also causes an astigmatism due to the different effective cell lengths in the two planes. These two astigmatisms are of opposite signs, allowing compensation by adjusting the cell thickness or the folding angle. Detailed analysis has been given in the literature (Kogelnik *et al.*, 1972). Because all practical cw dye lasers use thin free-flowing dye cells, full compensation is not feasible (e.g., for $L = 150$ μm and $R = 10$ cm, the theory shows that $2\theta = 4.5^0$ is needed). Thus, current systems either have some residual astigmatism or use added compensation such as an extra Brewster plate near the dye cell or inside an additional fold (Johnston *et al.*, 1982). Astigmatism may also be present in the pump beam if a tilted focusing mirror is used instead of a lens. For all of the above reasons, in an actual dye laser the beam areas and the parameter m are only approximately known.

2.2.2 The Flowing-Dye Cell and Triplet Quenching Continuous-wave operation of the laser causes continuous power dissipation. Much of this takes place in the active dye volume through internal nonradiative processes, resulting in watts of heat dissipated in the active volume of $\sim 1 \times 10^{-7}$ cm^3. To carry this heat away, a flowing-dye cell is needed, together with a small heat exchanger in the flow system. Although early designs used cells with glass windows (Kohn *et al.*, 1971; Dienes *et al.*, 1972), all modern systems use free-flowing dye jets (Runge and Rosenberg, 1972; Shank *et al.*, 1973) whose advantages of simplicity, high-speed flow, and long lifetime outweigh the disadvantage of some low-frequency noise introduced by the physical motions and local ripples of the jet. A jet nozzle simply consists of a flattened stainless steel tubing. Since the nozzle opening is very narrow, filtration of the flowing dye is needed in order to avoid clogging. Ethylene glycol has a high enough viscosity to produce a smooth-surfaced ribbon and is usually used as the solvent, sometimes mixed with water or other organic solvents and certain additives (to improve solubility and reduce aggregation). Using a centrifugal pump to produce a pressure of ~241 kN/m^2 (35 psi), an approximately rectangular ribbon of ~3-mm width and ~100–200-μm thickness is typically produced with a flow speed of up to

10 m/s. Even at this high flow speed, at high pump-laser powers beam-profile distortions (thermal blooming) (Teschke *et al.*, 1976b; Johnston *et al.*, 1982) will occur, limiting both the beam quality and the dye-laser efficiency. The dye concentration is adjusted to give around 90% absorption of the pump. For multiple-line pumping, some pump wavelengths may be weakly absorbed, and the net pump absorption may be somewhat lower.

The high-speed flow and circulation of the dye partially solve one of the other important requirements for cw lasing: quenching (i.e., shortening) the lifetime of the metastable T_1 state (Shank *et al.*, 1973; Yarborough, 1974; Schäfer, 1990). This is accomplished mostly by the aeration of the solution in the flow–circulation system but also by actually moving the molecules in the T_1 state out of the active volume. The combination of these two effects shortens the effective triplet lifetime and gives a triplet loss factor f_T sufficiently close to unity for cw operation in some dyes. In spite of this, the triplet absorption remains a limiting effect, and only certain dyes can be cw lased. The additive cyclo-octatraene (COT) can further decrease the triplet lifetime and improve laser efficiency (Yarborough, 1974; Johnston *et al.*, 1982). However, since this chemical is unstable and smells bad, its use is avoided whenever possible. For R6G, with aeration but without using COT, typically $k_{ST}\tau_T = 0.9$. This results in an f_T that is positive over a large range, decreasing from a value of 0.65 at 560 nm to zero at 650 nm.

2.2.3 Mirror Transmission The output mirror transmission (t) must be carefully designed for optimum lasing efficiency for each laser dye. For a given pump power there is an optimum value of transmission (t_{opt}) at each wavelength. This is found from Eqs. (11) and (13) as

$$t_{\text{opt}} = \sqrt{[l + s(1 - f_T)][g_0 - sf_T]} - [l + s(1 - f_T)], \tag{14}$$

where $g_0 = \ln G_0$ is the small-signal gain. For $\sigma_{al} = 0$, this reduces to $t_{\text{opt}} = \sqrt{lg_0} - l$, in agreement with the well-known general theory of homogeneously broadened lasers (Rigrod, 1963) and with experiments (Shank *et al.*, 1973). In practice, while a certain amount of tailoring of the mirror transmission is possible, compromises must be made. In Fig. 8 the calculated power output of a R6G dye laser vs wavelength is shown, with a mirror transmission that is constant at 5.7% and with mirror transmission optimized at each wavelength, together with the t_{opt} used. With the optimized mirror, improved output is seen on the long-wavelength side. The transmission is largest at 580 nm where the linear gain is the largest and decreases at both shorter and longer wavelengths. However, on the short-wavelength

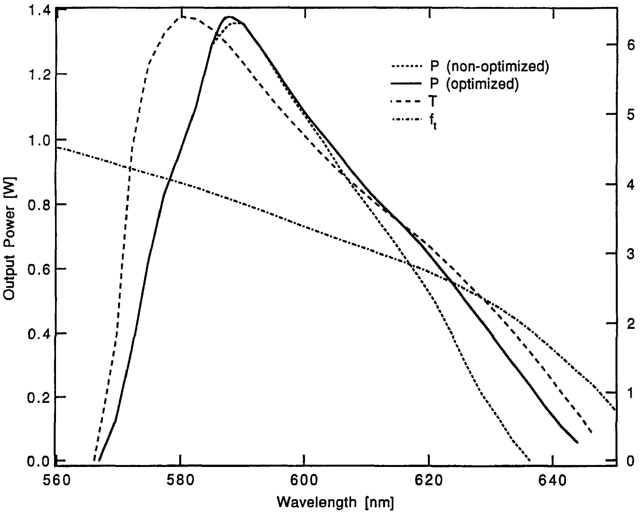

FIG. 8. Power (in watts) vs wavelength calculated for a R6G cw dye laser in a linear cavity, using the data shown in Fig. 2. The triplet reduction factor f_T (unitless) is also shown. Additional parameters: pump-beam area = 1.25×10^{-5} cm^2 (w_p = 20 μm), pump to laser beam area ratio μ = 1.5, residual loss l = 2%. The output mirror transmission t is optimized at each wavelength for one output power curve and is constant for the other. The optimized mirror transmittance vs wavelength is also shown (in %, scale on right).

side the absorption dominates the total loss, and optimizing the mirror transmission does not result in improvement. The important point for output mirror design is to optimize the transmission *on the long-wavelength side* of the peak gain. In these calculations the function f_T was calculated using $k_{ST}\tau_T = 0.9$, and is also shown on the graph. According to the theory given above, at any wavelength the output is a linear function of the input power. In reality, however, this is only true up to some value of pump intensity where thermal effects begin to be significant.

2.2.4 Performance Limitations There are factors that limit the performance of cw dye lasers but were not included in the simple theory.

1. As already mentioned, at high pumping powers, a thermal lens is created in the pumped region. As discussed in the literature (Johnston and Sasnett, 1992), this lens is astigmatic as a result of the flow and cannot be compensated. For relatively low-power systems this problem is not serious and results in only minor degradation of performance. For high powers, however, the thermal defocusing lens also becomes highly irregular and unstable, resulting in severe deterioration of the lasing beam. The laser must always be operated below this limit. For example, typically under normal laboratory conditions, 1–2 W of output can be obtained from R6G. The record powers reported in the literature can only be obtained with special solvent mixtures and additives (Johnston *et al.*, 1982) and are not easily accomplished.
2. Aggregation of the dye molecules occurs for many dyes at the concentrations needed to absorb a high fraction of the pump in the thin dye jet. This problem can be partially solved by the addition of a small amount of a surfactant such as Ammonyx LO (Peterson *et al.*, 1970; Johnston *et al.*, 1982), which, however, introduces some problems such as foaming.
3. Residual aberrations (astigmatism and coma) in both the pump and laser beams increase the effective areas and reduce the beam overlap, thus increasing threshold and decreasing slope efficiency. These effects also prevent the use of focusing mirrors (for both pump and cavity) with $R < 5$ cm.
4. Transverse modes may be partially present in the pump and may also appear in the dye laser beam.

On account of the combination of the above effects, actual threshold powers and slope efficiencies are somewhat lower than calculated by the theory, although with careful optimization nearly ideal performance *can* be approached.

2.2.5 Dyes and Pump Sources Only a small number of all the dyes that have been made to lase are suitable for cw operation. Both the UV and the IR ranges are limited by the availability of pump sources. Infrared-emitting dyes suffer from instability problems, whereas the blue- and UV-emitting dyes have much smaller emission cross sections [note the λ^4 dependence in Eq. (1)] and also tend to have large triplet absorption together with difficulty of quenching the triplet population. In the previous sections the R6G dye laser was used as an example, since the parameters of this dye are quite well known. This was the first cw dye laser and is still one of the best. Pumped with a powerful 24-W multiple-line argon laser, and with special solvent mix to alleviate thermal problems, a record high multimode power output of 6.5 W and single-mode power of 5.6 W has been obtained (Johnston *et al.*, 1982). Today, many cw laser dyes are available from the near UV to the near IR, and the performance of the best ones approaches, and in some respect surpasses, that of R6G. A comprehensive listing of dyes is not given here, but the tuning range of some of the best performers with their common abbreviations are shown in Fig. 9. These dyes, made especially for laser applications, are commercially available from various chemical companies. The overlapping lasing ranges cover the wavelength range from ~360 to ~1000 nm. Optimized systems using these dyes have been reported (Yarborough, 1974; Jarret and Young, 1979; Johnston *et al.*, 1982; Johnston, 1988). The output power depends on the power of the pump source, as well as on many details of the specific design. In Fig. 9, the pump source and specific line (or group of lines) needed is indicated for each dye. The commonly available pump lasers and their useful strong lines are listed in

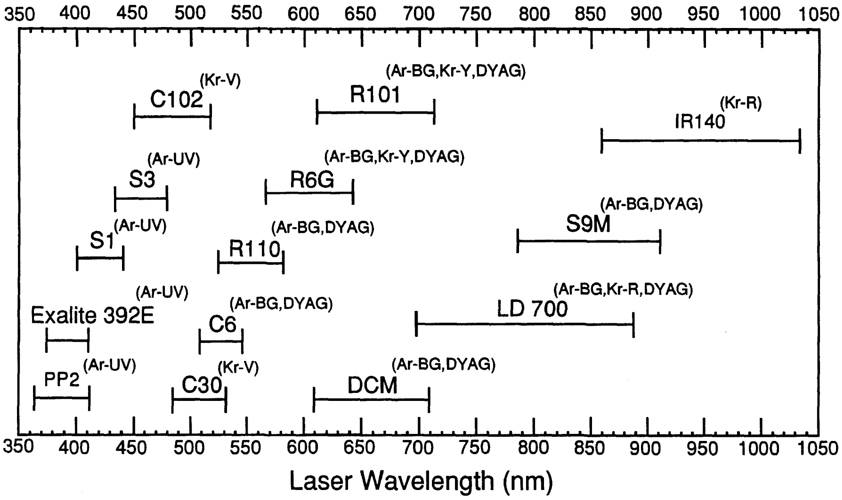

FIG. 9. The tuning ranges of some of the best cw laser dyes. For each dye the possible pump laser and pumping lines are also shown with the codes from Table 1. (C = coumarine, R = rhodamine; other abbreviations for the dye names are the standard ones used by the commercial catalogs).

Table 1. The obvious requirement is that the absorption band must overlap with one or more of these lines. A regular Ar^+ laser has the two familiar strong blue and green lines. For blue- and ultraviolet-emitting dye lasers, an UV-capable argon-ion laser is needed (which can also provide the blue-green lines with a change of mirrors). The longest-wavelength dyes can only be pumped with the red lines of the krypton laser. This laser also has a group of violet lines useful for pumping some blue-green laser dyes, and yellow-green lines of reasonable power that can be used as the excitation source for some red-lasing dyes. Thus, while a single ion laser obviously is not capable of pumping the entire range of laser dyes, a large fraction can be covered. Recent developments in all-solid-state laser sources made available a useful and practical new pump source: the diode-laser–pumped Nd:vanadate ($Nd{:}YVO_4$) laser whose output is efficiently doubled in a nonlinear crystal to provide several watts of power at $\lambda = 0.532$ μm wavelength. This system is an excellent pump source for several dyes emitting in the range 550–800 nm. Its biggest advantage is that it has much smaller electrical power and cooling requirements than ion lasers. It also has less random noise and superior beam characteristics.

Table 1. Lasers and laser lines for pumping cw dye lasers.

Laser type	Wavelength of line	Notes	Code for Fig. 11
Argon or UV-capable argon	476.5		
	488.0	a	
	496.5		Ar-BG
	501.7		
	514.5	a	
UV-capable argon	300.3–335.8	b	
	333.6–363.8		Ar-UV
	363.8–385.8	b	
Krypton	647.1–799.3	c	Kr-R
	520.8–568.2	c	Kr-Y
	337.5–422.6	c	Kr-V
Doubled $Nd{:}YVO_4$	532		DY

[a] These two lines are much stronger than the others and contribute most of the pump power.

[b] These groups of lines can be obtained only at very high tube current (e.g., 65 A in a Coherent Innova Model 100UVE). Change of mirrors is also needed.

[c] Change of mirrors is needed for each of these groups of lines.

Dye lasers between 530 and 930 nm provide the highest output powers, and they are the easiest to design, set up, and operate. This is due in part to the high emission cross section and low triplet absorption of these dyes (high σ_e, and f_T close to unity) and also to the use of the two blue-green (480 and 514.5 nm), high-power argon-ion laser lines as pump. An example is given in Fig. 10. The

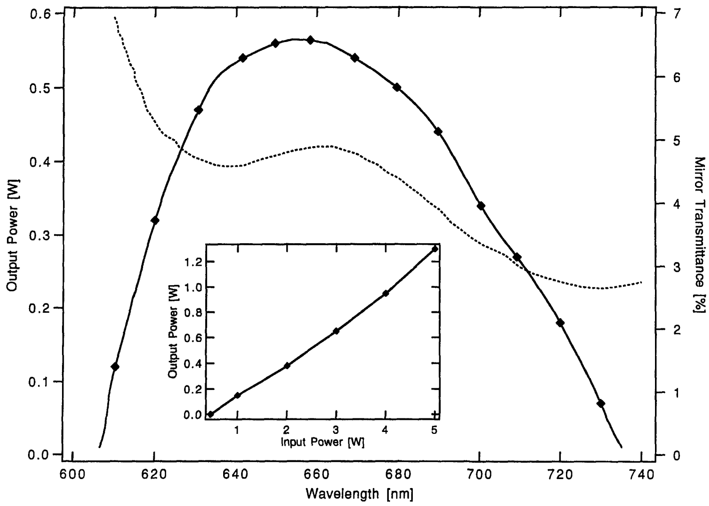

FIG. 10. Experimentally measured output power (solid line) and output mirror transmittance (dotted line) vs wavelength for the dye DCM in a ring cavity. The mirror transmission is approximately optimized. Inset: Input vs output power at the peak output wavelength. Data provided by T. Johnston, Coherent Inc.

output power and approximately optimized mirror transmission vs wavelength, as well as output vs input power at the peak emission wavelength, are shown. The dye is DCM, lasing in a single longitudinal mode, ring configuration (Johnston, 1988), similar to the one shown schematically in Fig. 5. The behavior is similar to that described earlier by the theory. The threshold pump power of 400 mW at the peak lasing wavelength is smaller than that calculated for the R6G example. This indicates a higher effective emission cross section, or a longer singlet lifetime, or a combination of both of these effects. The peak-wavelength slope efficiency is 0.78 times the value calculated for R6G. Some of this difference is due to the 1.1 times longer wavelength of the DCM emission and some to the smaller output-mirror transmission. The output-mirror transmission increases toward the peak on the long-wavelength side, as required by the theory, but unlike the calculated optimum mirror, the transmission continues to increase on the short-wavelength side because of design practicalities. This departure from the theoretical optimum makes no practical difference, as was noted in Sec. 2.2.3.

The blue- and UV-emitting dyes have lower values of the effective emission cross section ($f_T\sigma_e$), and the available pump power in the UV is also much less. As an example, the results reported for the shortest-wavelength dye PP2 (Johnston, 1988) are examined briefly. This dye was pumped with a large Ar^+ laser (Coherent Innova 100 UVE) operated at a high discharge current of 65 A. A total of 3.4 W was obtained in 6 UV lines, 300.3–335.8 nm, and was focused with a mirror to $w_p \sim 25$ μm in the dye jet, where 84% of this pump power was absorbed. The ring cavity had a folding-mirror radius R = 7.5 cm. With an output mirror transmission of 1.9%, the lasing range was 364–408 nm with a maximum output of 255 mW at 383 nm. At this wavelength the slope efficiency was ~11% and can be semiquantitatively explained by the theory as follows. The given values of $a = 0.84$, $t = 0.019$, and estimated values of $m = 1.5$, $\nu_l/\nu_p = 0.84$, $l = 0.025$, $s = 0$ give a slope efficiency of 11% if the value of the triplet factor $f_T = 0.44$, which is within the known range for blue-emitting dyes.

2.2.6 Bandwidth Control and Tuning

One of the most important characteristics of dye lasers is their very wide gain bandwidth. To utilize this large bandwidth properly, one must be able to tune the wavelength of oscillations and also to control the bandwidth of the output. For many applications, a bandwidth of ≈0.5 nm (≈500 GHz), or even larger, is quite sufficient, while some spectroscopic experiments require a bandwidth as low as ~1 MHz. To achieve this bandwidth, oscillation must be restricted to a single cavity mode (since cavity-mode spacing is typically ~150 MHz). Frequency stabilization may also be required. In this article we discuss tuning and bandwidth, but not frequency stabilization.

Dyes are spectrally homogeneous, which means that all the lasing frequencies compete for the same excited singlet population, i.e., for the same gain. This would imply, according to well-known fundamental laser theories, that the strongest lasing mode suppresses the others, and single-mode operation results. However, in practice this is not the case. Without any wavelength control, a standing-wave (linear cavity) cw dye laser operates at the peak of the gain with a bandwidth that is typically as broad as ~1–2 nm (1–2 THz). This is due to an effect known as spatial hole burning. For each of the cavity resonance modes, the two oppositely traveling beams form a standing-wave pattern inside the cavity, with the adjacent patterns having numbers of nodes differing by one. As a result, the different laser modes do not interact with the same molecules but saturate (or "burn down") different spatial locations. Mode competition is thus greatly diminished, and many laser modes oscillate. Another way to describe this effect is that oscillation at one mode is able to extract the available energy only at the antinodes, and the laser will oscillate in so many modes that an overall uniform intensity is established with maximum energy extraction. This assumption was built into the theory, since standing-wave patterns of the laser intensity were ignored. It is possible to force a linear-cavity cw dye laser into single-mode operation by placing a hierarchy of ever narrower-bandwidth filters inside the cavity. However, the output power of the laser is considerably diminished because only about half of the actual inverted medium is utilized. In practice, the single-mode output of such a laser is typically about 40% of that obtained in broad-band operation (Johnston *et al.*, 1982). This problem can be overcome by using a ring configuration (Jarret and Young, 1979; Johnston *et al.*, 1982) and forcing the laser radiation to travel unidirectionally. In such a laser the resonant modes are pure traveling waves, all modes compete for the same populations, and single-mode operation results. An element commonly called an optical diode is used to force a ring dye laser into unidirectional operation. This element is lossless for a laser beam propagating in one direction but provides a loss in the other direction. A differential loss of about 0.5% is achieved by rotating the polarization of the laser beam a few degrees for one direction of travel only. For this beam, a small loss is experienced at the various Brewster-angle surfaces of the other cavity elements that follow the "diode." Polarization rotation for only one direction of propagation is accomplished by a series combination of two well-known elements: a birefringent quartz half-wave plate rotator and a Faraday-effect rotator. The rotations due to the above two elements cancel for one direction of propagation and add for the other. Details of the design of an optical diode can be found in the literature (Johnston and Proffit, 1980). On account of mechanical instabilities, the output of a ring-type laser will still be broad band because the frequency of lasing will rapidly change among many modes. But it is possible to narrow this bandwidth by the addition of tunable-frequency selective filters to a much narrower bandwidth and even to a single mode without significant loss of output. The selectivity of these elements inside the dye-laser cavity is much stronger than would be in normal extracavity use because the spectrally homogeneous dye laser is very sensitive to small variations of the net gain (Klein *et al.*, 1973). A net gain reduction as low as 0.5–1% on a propagation direction or on a group of modes is sufficient to suppress the lasing.

One of the most obvious frequency-selective element is a diffraction grating. But its insertion loss is far too high to be used inside a cw dye-laser cavity. Diffraction gratings *are* important for the tuning of pulsed-laser–pumped dye lasers, which have high gain. While many different types of tunable wavelength-selective filters have been successfully used in cw dye lasers, the simplest, cheapest, and most reliable are the birefringent (Lyot) filter and the Fabry-Pérot etalon. A birefringent filter consists of several quartz plates, with thicknesses related by integral multiples, cut with the optic axis parallel to the surface, and tilted at the Brewster angle relative to the direction of propagation. The frequency selectivity results from the birefringence of the plates combined with the polarization-dependent loss from the Brewster-angle surfaces. Each of the plates acts as a wavelength-dependent wave plate that changes the polarization of the light passing through it. The filter is tuned by rotating it about the normal (while maintaining the Brewster-angle orientation) and thereby changing the amount of birefringence (Bloom, 1974; Teschke and Holtom, 1974). In general the polarization exiting each plate is elliptical, and loss is experienced at the Brewster surfaces. At one wavelength, however, the light exiting each plate is linearly (p) polarized and thus experiences no loss. This is the center of the main pass band. Typical birefringent elements for a dye laser have one to three plates. The main pass band is about 1.5 THz wide for a three-plate Lyot filter.

A single glass plate, placed at near-normal incidence and coated on each side for a desired reflectivity, is a suitable Fabry-Pérot etalon for the cw dye laser. The transmittance of such a parallel-plate interferometer is

$$T(\delta) = \left[1 + \frac{4R}{(1-R)^2}\sin^2(\tfrac{1}{2}\delta)\right]^{-1}, \qquad (15)$$

where $\delta = (4\pi nL/\lambda)\cos\theta$. Here, R is the power reflectance of the etalon's surfaces, L is its thickness, n is its index of refraction, and θ is the angle between the etalon normal and the inside direction of propagation. This angle is kept small to minimize "walkoff" losses. The finesse F, which is the ratio of the free spectral range (mode spacing) to the half-power bandwidth, is given by $F = \pi\sqrt{R}(1-R)^{-1}$. The center frequency of an etalon can be tuned by changing either the length L or the angle θ. A length change of one-half wavelength tunes the frequency through one free spectral range.

A hierarchy of three frequency-selective elements is typically needed to achieve single-mode operation of a cw dye laser. The resolving powers must be chosen so that one (coarser) element is able to select a single resolution element of the next (finer) component, as shown in the diagrams of Fig. 11. The element with the coarsest resolution is usually a birefringent filter with half-power bandwidth of about 1.5 THz. This is usually followed by two etalons. The first of these is a single glass plate with a free spectral range of around 200 GHz, tunable by rotation. The second is a thicker, air-spaced etalon made of two parallel glass surfaces between Brewster-cut glass prisms, with free spectral range of about 10 GHz and tunable by changing the spacing. The etalons are coated for a low finesse of about 3. By virtue of the sensitivity of the laser to small variations of the net gain, the last element is able to select a single one of the cavity modes (typically spaced about 150 MHz). The bandwidth of this single-mode laser is now determined by the mechanical stability of the optical elements and of the gain. Unless specially engineered and stabilized for true single-mode operation, some mode hopping will take place, resulting in a bandwidth that may be of the order of 1 GHz. Feedback-stabilized, true single-mode cw dye lasers have bandwidth (determined by the jitter of the single-mode frequency) typically of the order of ~1 MHz, but stability and bandwidth down to the kilohertz range has also been achieved.

It is easy to understand from the above

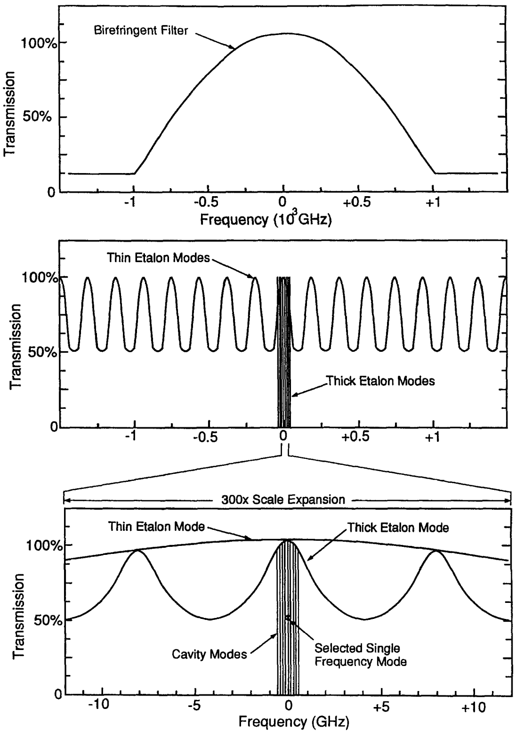

FIG. 11. Schematic of a hierarchic three-element frequency-selection and tuning scheme for a cw ring dye laser (after Duarte, 1990). The passband of each finer selective element is tuned to the center of the next coarser element. The sensitivity of the dye laser to small gain variations results in lasing in only one passband of each element. The final thick etalon selects a single cavity mode.

discussion why two designs are most often used for tunable cw dye lasers. One is a relatively simple, compact three-mirror standing-wave cavity laser with a single wide-band, mechanically controlled tuning element. The width of the broad-band output depends on the tuning element. A single-plate birefringent filter results in ~400 GHz bandwidth, but ~40 GHz can be achieved with a three-plate filter. The second common design is a unidirectional ring laser with a hierarchy of frequency-selective elements, resulting in narrow bandwidth and capable even of single-mode operation. For either of these designs one must be able both to tune and to track the operating frequency. This is a relatively easy task for the broad-band, linear-geometry laser, since the reproducibility needs to be only as accurate as the bandwidth. Tuning is accomplished by mechanically rotating the birefringent filter. The wavelength of lasing is approximately calibrated and can be externally measured. For the much narrower-bandwidth ring laser, sophisticated electromechanical design is required. The coarsest adjustment is made with the birefringent filter, usually purely mechanically. The next stage of tuning is by rotation of the thin etalon using a galvo-motor. The next coarser adjustment is made by piezoelectric tuning of the length of the thick etalon. The finest tuning adjustment, within a portion of the thick-etalon bandwidth, is made by changing the effective cavity length. This is usually done in two stages: by piezoelectric translation of one cavity mirror and by slight rotation of a thick Brewster-angle plate (or double Brewster plates). Tracking the oscillating frequency requires simultaneous and precise control of several elements. Description of a tracking system design is beyond the scope of this article. Typically, the continuous scanning range is limited to about 30 GHz. Computer-controlled tuning and tracking over the entire tuning range of a given dye has been developed, and such systems are commercially available.

Finally, it should be mentioned that injection locking has also been used to achieve bandwidth-limited performance. In this method a precise, narrow-band but very weak signal (from another source) is injected into the cavity of a more powerful oscillator. Because of the sensitivity of the oscillator to small perturbations, its output locks to the injected signal with narrow bandwidth. This method is useful and straightforward but is limited by the availability of the injection signal.

2.3 Mode-Locked cw Dye Lasers

The two commonly used tunable cw dye-laser systems described in the previous sections both aim for narrow bandwidth and wide tunability. Another very important utilization of the very wide gain bandwidth of dye lasers is for ultrashort pulse generation. In this case, lasing takes place over as wide a bandwidth as possible, with all frequencies brought into perfect coherence, to give a train of ultrashort pulses. Very little tunability is therefore possible for a given dye. Full description of these lasers properly belongs to a separate article (see ULTRASHORT PULSE PHENOMENA). However, no discussion of dye lasers would be complete without at least a brief description of these systems, which played such an important role in the development of femtosecond pulse generation.

The gain bandwidth of a typical dye would support pulses as short as 5–10 fs, but pulse-broadening mechanisms, such as dispersion, limit durations to typically an order of magnitude longer. A train of 50-fs pulses at 10-ns repetition rate (the round-trip time of the laser cavity) is composed of 2×10^5 individual modes, all having identical phase. To achieve this coherence, a strong pulse-shortening mechanism is necessary. Although the first reported mode-locked cw dye laser (Dienes *et al.*, 1971) used active modulation, and synchronous pumping has also been successfully used, passive mode locking with saturable dyes is the only method that results in pulses less than 300 fs in duration. The pumping and laser cavity configurations are simple variations of the two used for narrow-band cw dye lasers (Figs. 4 and 5). Another folded arm is added for tight focusing into the saturable absorber, and, in place of the frequency tuning and restricting elements, dispersion compensation elements are added. The shortest pulses were obtained in the ring configuration shown in Fig. 12, where two sets of pulses circulating in opposite direction "collide" in the saturable absorber, increasing the pulse-shortening action. Pulses as short as 27 fs have been reported (Valdmanis and

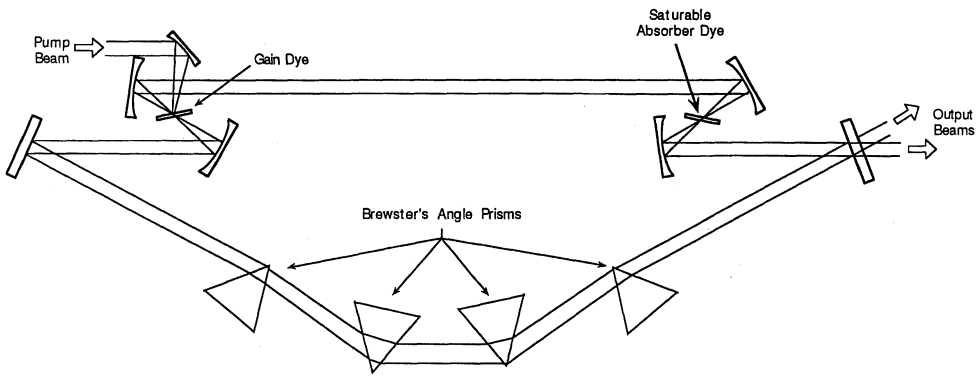

FIG. 12. Schematic diagram of a colliding-pulse mode-locked cw dye laser with a dispersion-compensating four-prism sequence included in the cavity.

Fork, 1986) from such a setup, but more commonly pulses around 50 fs long are obtained. The dynamics of pulse shortening in this laser are now examined briefly.

One can distinguish three stages in the evolution of the mode-locked pulses. In the first stage, a noise burst in each direction of circulation, around some picoseconds long, is selected by the gating action of the saturable absorber acting alone. The duration of the noise burst is limited by the recovery time of the absorber. In the second stage, two coherent single pulses (one in each direction) are formed by the combined pulse-shortening action of saturation in both (absorber and gain) dyes, together with the dispersion-compensating effect of the four-prism sequence. Figure 13(a) shows semiquantitatively the dynamics of the pulse shortening in this, most critical, stage. The center of the pulse has gain while both leading and trailing edges experience loss. This is the result of a rather delicate interplay of the two saturations, of their recovery, and of the linear loss. At the leading edge, the loss is higher than the initial gain because the gain has not fully recovered to the small-signal value after one round trip. On the other hand, the loss experienced by the trailing edge is due to the linear loss that remains after the absorber is fully saturated, while at the central portion of the pulse there is net gain due to the faster saturation of the absorption. It can be deduced from this diagram that for effective pulse shortening

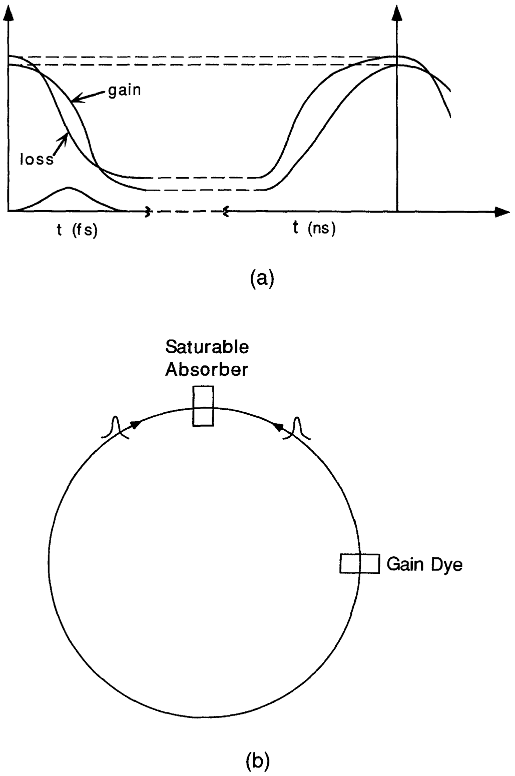

FIG. 13. (a) Pulse-shortening dynamics in passively mode-locked cw dye lasers. The loss includes that of the saturable dye, as well as the fixed losses. The saturation of the gain and of the absorption takes place on a femtosecond time scale during the passage of the pulse. The recovery is on a nanosecond time scale. (b) Schematic of the optimum placement of the gain dye and of the saturable absorber in the colliding-pulse ring configuration.

1. absorber saturation must be stronger than gain saturation; and
2. the absorber must fully recover by the time the next pulse arrives, but the gain must not fully recover.

These conditions *can* be met in a linear configuration but are easier in the ring configuration where the oppositely traveling pulses arrive simultaneously ("collide") in the absorber. Additionally, the oppositely traveling pulses should experience the same initial (not fully recovered) gain. For these reasons the pulses not only must collide in the absorber but also must anticollide (i.e., arrive with maximum separation) in the gain medium. This condition determines the location of the dyes and the corresponding focal points in the ring cavity to be as is schematically shown in Fig. 13(b). These multiple-folded ring cavities have rather complicated stability regions, depending on the dimensions of the various arms and radii of curvature of the mirrors. Detailed Gaussian-beam analysis has been done to optimize thedesign (Li *et al.,* 1981; Magni *et al.,* 1992).

The pulse shortening described above is opposed by pulse broadening due to second- and higher-order dispersion. Dispersion results in the addition of a spectral-domain phase $\Phi(\omega)$ to the oscillation spectrum (Siegman, 1986). This phase is nonlinear in the frequency ω. In a dye cavity this is the result of the reflections from the broad-band multilayer mirrors as well as of propagation through glass or fused-silica elements. Second-order or group-velocity dispersion is the most important and results in a symmetric broadening of a pulse. While pulses as short as or even shorter than ~100 fs have been achieved without dispersion compensation, the shortest pulse durations are achieved only with a prism compensator in the cavity. The four-prism sequence introduces adjustable second-order dispersion and allows optimization of the net dispersion experienced by the pulse in a round-trip. This optimum is not exactly zero value because in the third stage of pulse formation—which is reached only for pulses shorter than ~100 fs—new nonlinear effects become significant, and optimal pulse shortening requires a small amount of negative group-velocity dispersion to be present. In this third stage of pulse formation, self phase modulation by the nonlinear index n_2 in the absorber (and to a lesser extent in the gain dye) in combination with negative second-order dispersion results in a discrete soliton formation mechanism (Valdmanis and Fork, 1986). Positive frequency sweep (chirp) on the carrier due to the self phase modulation is canceled by a negative chirp due to dispersion. The pulse shortens until a balance is reached, corresponding to a fundamental soliton (a pulse of fixed shape and intensity) circulating in the cavity. This last stage is necessary if the shortest possible pulses (less than ~50 fs) are to be reached but requires very critical adjustment of cavity and other parameters.

In principle, the center wavelength of the mode-locked femtosecond pulses should be tunable within the available gain bandwidth because even the shortest pulses achieved have a spectral width that is only a fraction of that of the gain. However, in reality all mode-locked dye lasers operate only at one fixed wavelength. This is due to the small range allowed on a variety of parameters. The required combination for achieving good mode locking and femtosecond pulses can usually only be obtained near one wavelength with a limited tuning range. For the same reason, not all cw laser dyes have been successfully mode locked for femtosecond pulse operation. The best results have been achieved with R6G as the gain dye and DODCI as the saturable absorber, at the central wavelength of ~630 nm. This was the dye combination in the first passively mode-locked cw dye laser (Ippen *et al.,* 1972) which used a nonoptimum linear configuration and resulted in 1.5-ps pulses without any dispersion compensation. The invention of the colliding-pulse mode-locked ring configuration, together with dispersion compensation, gradually shortened the pulse duration to eventually 27-fs pulses (Valdmanis and Fork, 1986). Until the recent development of the Kerr-lens mode-locked Ti:sapphire laser, this was the shortest laser pulse from any system directly without external compression. Other dye combinations have resulted in cw mode-locked pulses in the femtosecond range at several additional wavelengths, but pulse durations are somewhat longer and mode locking is more difficult. A list of cw mode-locked dye lasers with lasing wavelength ranges and best reported pulse lengths is given in Table 2. Additional gain–absorber dye combinations have been mode locked using pumping by mode-locked lasers (hybrid mode locking). A description of this technique and a table of results are given in the literature (Duarte and Hillman, 1990).

3. PULSED DYE LASERS

Because dye-laser excitation requires high photon flux, cw pump sources are both limited in number and expensive. Additionally, on account of limitations of the pump sources, difficulties in quenching the accumulation of molecules in the triplet state, and thermal problems, relatively few dyes can lase in cw mode. On the other hand, pulse excitation allows excitation of a larger number of dyes using a variety of sources. To pulse-excite dye lasers, two different photon sources have been utilized: either a pulsed laser at a wavelength within one of the singlet absorption bands of the dye or a flashlamp. Historically, the first dye lasers were demonstrated (Schäfer *et al.*, 1967; Sorokin and Lankard, 1967) using pulsed pumping with giant-pulse ruby lasers and with flashlamps. Today a wide variety of pulsed dye lasers exist, either sold commercially or constructed in the laboratory. In this section both methods of excitation are discussed, and various pulsed dye-laser configurations are described. Methods of tuning and reduction of the dye-laser bandwidth are also presented.

3.1 Flashlamp-Pumped Dye Lasers

A simple, practical, and relatively inexpensive method of dye-laser pumping involves the use of the broad-band visible and UV emission that results after an electrical discharge takes place in xenon, i.e., flashlamp pumping. Flashlamp pumping allows a highly symmetric excitation of a dye solution and can yield large output laser energies. The active volume of flashlamp-pumped dye lasers is relatively large. The dye-cell length ranges from ~15 cm to over 50 cm, and the diameter from a few millimeters up to ~2 cm. In all flashlamp excitation schemes the absorption of the pump is, by necessity, transverse to the lasing direction. The dye concentration is adjusted so as to absorb in a distance about the diameter of the dye cuvette, resulting in the absorption of a fairly high fraction (~0.8–0.9) of that spectral portion of the flashlamp output that has wavelengths overlapping the singlet absorption band. The two main types of excitation use linear and coaxial flash lamps. As shown in Fig. 14(a), linear excitation usually utilizes two or more flashlamps symmetrically arranged around the active region (the dye cuvette) and a reflective surface around the perimeter. Single or double elliptical reflectors may be used to improve the dye excitation. Coaxial excitation, shown in Fig. 14(b), yields a concentric dye illumination and very uniform pump intensity distribution resulting in large output energies (Baltakov *et al.*, 1972). A third configuration involves the use of a single helical flashlamp, with the dye cuvette placed at the axis and a cylindrical re-

Table 2. Passive mode-locking dye combinations and results (after Duarte and Hillman, 1990).

Gain dye	Saturable absorber[a]	Wavelength range (nm)	Shortest pulse width (fs)	Notes
Coumarin 102	DOCI	493–502	93	Ring cavity
Coumarin 6H	DOCI	492–507	110	Ring cavity
Rhodamine 110	HICI	553–570	70	Ring cavity
Rhodamine 6G	DODCI	550–570	27	Ring cavity
Rhodamine B	DQTCI	616–658	220	Linear cavity
Rhodamine 6G/SRh 101	DQTCI	652–682	120	Linear cavity
DCM	DOTCI	655–673	680	Linear cavity
Rhodamine 700	DOTCI + DCI	727–740	350	Linear cavity
Rhodamine 700/DCM	DDI	742–761	110	Ring cavity

---

[a] The acronyms for the dyes are the standard ones used in the literature and in the commercial dye catalogs. Full chemical names are rarely if ever used but can be found in the literature (e.g., DODCI = diethyl-oxa-dicarbocyanine-iodide).

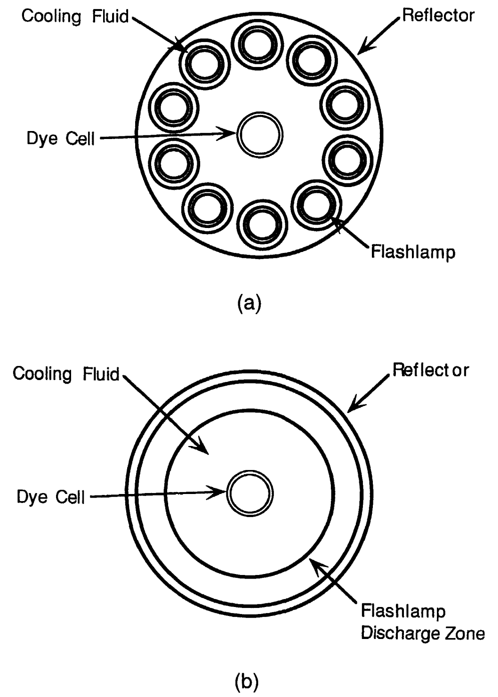

FIG. 14. (a) Cross-sectional diagram of a linear flashlamp excitation scheme using multiple flashlamps. (b) Cross-sectional diagram of a coaxial flashlamp excitation scheme. The discharge takes place longitudinally in the hollow-cylinder flashlamp.

flector outside. Flashlamps are fairly efficient in transforming electrical to light energy, and both peak output power and total output energy are high. The detailed spectral characteristics of the light output depend on the excitation current. The output is always very broad band, and only a fraction (typically ~10% or less) of this output falls in the right wavelength region to be absorbed by the main singlet absorption band of the dye. This greatly limits overall efficiency, and it also poses problems of dissipating the heat that results from the energy in the unused wavelength ranges. In spite of these problems flashlamp pumping is simple and inexpensive to implement. It is also one of the schemes that can be "home built" in almost any laboratory. Although today the use of flashlamp pumping is no longer extensive, its study, besides the obvious historical importance, greatly helps with the general understanding of all dye lasers.

3.1.1 Theoretical Description and Discussion With flashlamp excitation, no exact analytical theory can be given. However, by making a few simple approximations, useful expressions *can* be derived. In fact, with some minor but important modifications, much of the theory given in Sec. 2.1 can be used to understand the operational principles, as well as the limitations, of these devices. The duration of the flashlamp pulse is a very important parameter, and its effects are discussed below. Figure 15 shows a schematic of the excitation electronics, based on the simple principle of discharging the energy stored in a large capacitor through the ionized gas of the tube. The discharge is triggered by a spark gap. The pulse length depends on the size of the capacitor and on the impedance of the discharge (and therefore on the size of the flashlamp). Several small flashlamps together can give the same energy as a single large one, but in a much shorter pulse. Dye-laser flashlamps have durations ranging from less than 100 ns to tens and sometimes even hundreds of microseconds.

Although the pumping geometry is very different from the cw-excitation case, with absorption of the pump taking place transversely to the propagation of the laser pulse, the rate equations (for an arbitrary point in

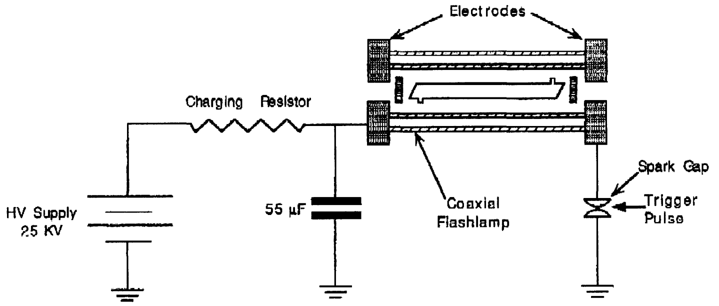

FIG. 15. Schematic of the overall configuration and of the electrical excitation of flashlamps for pumping dye lasers.

the dye medium) presented in Sec. 2 are still valid for any flashlamp-pumped dye laser. One modification is that the pump absorption coefficient σ_{ap} must be a weighted average over the entire singlet absorption band of the particular dye, because the flashlamp is a broad-band source, covering wavelengths from the infrared to the ultraviolet. Since the lifetime of the singlet (τ_1) is much shorter than the pump-pulse duration, quasi steady-state conditions can be assumed to exist for the singlet absorption pumping; i.e., the population N_1 essentially follows the pump pulse, as per Eq. (8). This is not the case, however, for the triplet-state population N_T. Just as in cw systems, triplet absorption is also a severe problem in flashlamp-pumped dye lasers, since pump-pulse durations tend to be longer than $1/k_{ST}$, while τ_T is also on the same order or longer. Because of the time-dependent pump intensity, the exact nature of the resultant triplet absorption is more complicated than for the cw case. This problem has been investigated in considerable detail (Maeda, 1984; Schäfer, 1990), and it is well known that the values of k_{ST} and τ_T to a large extent determine the duration of an optimum pump pulse and that of the resulting dye-laser pulse.

If long pulses (microseconds or longer) are desired, quasi-cw conditions apply in the triplet rate equation as well, and consequently one is restricted to using those dyes in which $k_{ST}\tau_T$ can be quenched to a small enough value to give f_T close to unity (i.e., those that are capable of cw operation). However, dyes in which the large value of $k_{ST}\tau_T$ cannot be quenched still can be made to lase if pumped with sufficiently short pulses. It is easy to show that for these dyes the rise time of the pump pulse (τ_{pr}) must be short (equal to or shorter than $1/k_{ST}$). Then, lasing can start before triplet accumulation has taken place. But lasing will be cut off by the buildup of the triplet population, which is governed by the crossover rate k_{ST} and the triplet decay rate τ_T. These effects together determine the optimum pump-pulse duration. Detailed modeling of the power output can be made by numerically solving the rate equations. Here, some simple but useful results that aid a basic understanding are given. First, below–lasing-threshold conditions are considered, assuming that for the given dye the value of the triplet factor f_T [given by Eq. (7b)] is negative, i.e., no cw or quasi-cw lasing can take place. For N_1, quasi-cw conditions can still be assumed, giving

$$N_1 = N_0\sigma_{ap}\tau_1 I_p(t). \tag{16}$$

For a fast-rise-time flashlamp ($\tau_{pr} \ll 1/k_{ST}$ and τ_T) a time-dependent triplet factor $f_T(t)$ takes the place of f_T. During the buildup of the intense pump pulse, assuming $\tau_T \gg 1/k_{ST}$,

$$N_T = k_{ST}\int_0^t N_1 dt, \tag{17}$$

and thus the triplet reduction factor becomes during this rise time

$$\begin{aligned} f_T(t) &= 1 - \frac{N_T}{N_1}\frac{\sigma_T}{\sigma_e} = 1 - \frac{k_{ST}\int_0^t N_1 dt}{N_1}\frac{\sigma_T}{\sigma_e} \\ &= 1 - \frac{k_{ST}\int_0^t I_p dt}{I_p}\frac{\sigma_T}{\sigma_e}. \end{aligned} \tag{18}$$

For $\tau_{pr} \ll \tau_T$ this is much smaller than f_T, the cw triplet reduction factor. Thus even with $f_T < 0$, the value of $f_T(t)$ can remain positive, and lasing can be achieved if the rise time is fast. However, the triplet population continues to build up and limits the lasing. If the pump intensity is approximately constant after the fast rise, then the value of $f_T(t)$ will decay toward the cw value exponentially, with time constant τ_T, and lasing will be terminated when $f_T(t) = 0$. It is evident that an optimum pump pulse should also terminate around this time. On the other hand, if the steady-state value of the triplet factor f_T is larger than zero, fast pump-pulse rise time is not needed to reach lasing threshold, and the lasing will follow the pump pulse. These points are further illustrated graphically in Fig. 16, which shows an *exact* solution for the factor $f_T(t)$ and for the gain $g(t)$ for a long pump pulse with an exponential rise time τ_{pr}, for $f_T = -0.3$, and for various ratios of τ_{pr}/τ_T. It can be clearly seen from these curves that it is wasteful and even detrimental (because of the resultant heat generated) to use too long a pump pulse. A partial list of dyes that do not lase when cw long-pulse pumped but can be lased by fast excitation is given in Sec. 3.2.

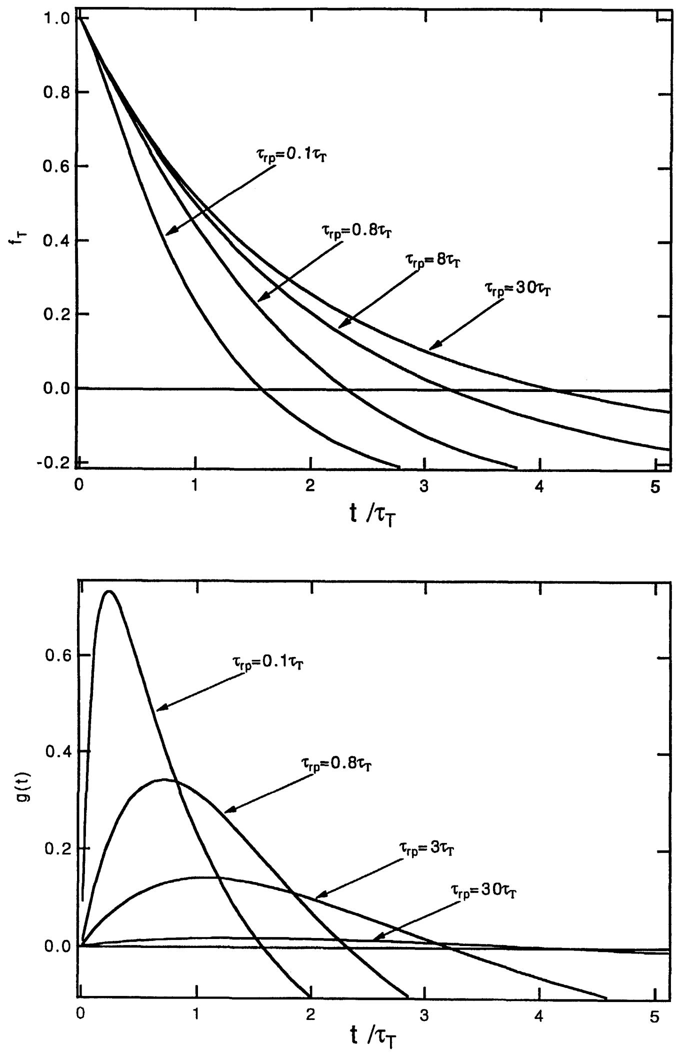

FIG. 16. Calculated time dependence of the triplet reduction factor $f_T(t)$ and of the incremental gain $g(t)$ for a long pump pulse with shape given by $I_p(t) = 1 - e^{-t/\tau_{pr}}$.

The above equations are valid at an arbitrary point in the dye-laser medium. To describe fully the behavior of a flashlamp-pumped dye laser and to find its power (or energy) output, the full rate equations, under lasing conditions, must be integrated over the entire active volume. It is not possible to do this as accurately as was the case for cw excitation. Nevertheless, a useful approximate description *can* be given. In this approximation it is assumed that the laser is well above threshold and that the cavity decay rate is much shorter than the pulse duration, and thus cavity dynamics can be ignored. In spite of the fact that both the pumping and cavity configurations look very different from the cw dye laser, and the excitation is time dependent, a great deal of similarity exists in the descriptions. This is because quasi steady-state conditions exist with regard to the singlet absorption excitation. The necessary modifications of the theoretical expressions derived in Sec. 2.1 are due to the different pumping geometry, to the broad-band nature of the flashlamp photon output, and to the possible lack of a steady-state triplet population.

The method described in Sec. 2.3 is used, with the new assumption that the excitation is uniform in the longitudinal direction (along the lasing axis) and that all quantities are averaged in the transverse cross section of the active volume. It is also assumed that the absorption length is approximately equal to the diameter of the dye cell. This, combined with a reasonably uniform transverse distribution of the photon flux around the surface of the dye cuvette, results in an approximately uniform excitation over the cross section. For a simplified derivation, a dye cuvette of length L with a square cross section D is assumed with a uniform photon flux input on one rectangular ($L \times D$) side of the cuvette. This incident photon flux is subject to an average absorption cross section σ_{ap} with an approximately constant ground-state population (N). The fraction of the incident flux absorbed is $a = 1 - \exp(-\sigma_{ap}ND)$. In the same manner as in Sec. 2.1, we integrate the population densities in the transverse dimension and then solve for the gain and the power output. It is found that the gain is proportional to the absorbed pump power divided by the cross-sectional area of the active medium (perpendicular to the lasing direction). For the flashlamp-pumped dye laser this area is simply the cross section of the dye cuvette. The resulting expression for the output power is identical to that given in Eq. (11) for the cw dye laser, with only the following slight modifications for slope efficiency and threshold:

$$\eta = \frac{\nu_l}{\nu_p} f_V f_T a \frac{t}{t+l}, \tag{19}$$

$$P_{\text{th}} = Ah\nu_p \frac{t+l}{f_T \sigma_e \tau_1 a}. \tag{20}$$

Here, the extra factor f_V represents the fraction of the dye cuvette volume that is filled by the dye-laser mode. The area A is the cross section of the dye cuvette, and the pump frequency ν_p is now value-averaged over the dye singlet absorption band. For pump pulses longer than the triplet decay rate (τ_T), the triplet factor f_T is the same as the cw value [Eq. (7b)], which must be positive. For those cases when this factor is negative, no lasing occurs with long-pulse pumping. For lasing with short-pulse pumping, the equations remain approximately valid provided that f_T is replaced by the time-dependent factor $f_T(t)$ (which can be estimated; see Fig. 16). Such results for the power output, while approximate, correctly describe the behavior of flashlamp-pumped dye lasers. They are useful not so much to calculate the output of the laser but to help understand and discuss their most important characteristics not yet examined in the preceeding paragraphs. Some of the particular points that can be made are these:

1. In order to input most of the incoherent pump light into the dye cell and to insure good uniformity of the absorbed power over the transverse cross section, the dye cuvette must typically have diameter from a few millimeters up to ~1 cm. It also follows that the required dye concentration must be rather low, of the order of 10^{-4} to 10^{-5} mol/L, in contrast with the cw dye laser where the absorption depth is around 100 μm using typical concentrations of ~10^{-3} mol/L. If the concentration is too high, absorption will take place in a thin layer near the dye-cell walls resulting in an unsuitable annular gain region. Because of the relatively low concentration, there are no dye aggregation or concentration quenching problems.
2. The pump powers required are high because the dye transverse cross section is of the order of millimeters or even a few centimeters. By comparison, in cw dye lasers the active region (defined by the focused pump beam) is ~10–30 μm diameter. In Sec. 2.1 it was calculated that, for R6G with the pump beam focused to a diameter of 20 μm, a pump power of ~280 mW is needed to obtain a threshold round-trip gain of 15%. For a flashlamp-pumped system, assuming a dye cuvette diameter of 6 mm, and assuming that the triplet decay is quenched to the same value as in a cw dye laser, the threshold pump power to achieve the same gain is ~10^5 × higher, or ~25 kW of power absorbed in the dye.
3. The power (or energy) output does not directly depend on the dye volume and concentration. Instead, for dye lasers the power output increases linearly with the pump power and is limited in practice by the power of the pump lamp itself and by the practical problem of getting rid of the excess heat. A larger flashlamp will, of course, mean larger pump power and necessitates a proportionally larger dye cuvette as well as larger reflector and housing with cooling system.

In addition to the obvious photon-frequency conversion loss (ν_l/ν_p), the fraction of pump light absorbed (a), and the reduction factor by triplet absorption [$f_T(t)$], the efficiency of the dye laser is determined by the ratio of the output-coupling loss to the total cavity losses, $t/(l + t)$, in similarity with cw systems. Typically, this last factor is smaller than in a cw dye laser, on account of the somewhat higher cavity losses. The overall electrical power to laser light efficiency is also given by Eqs. (11) and (12) if the factor a is modified to include the fraction of the electrical power input that is converted into light with wavelengths within the absorption band of the particular dye. This fraction is typically around a few percent. Thus, if the flashlamp-pumped dye laser is properly designed to have an output with duration approximately the same as that of the pump (either by strong triplet quenching or by using a fast-rising and short pump pulse), then one can estimate an overall energy efficiency as follows. Assuming the typical values $a = 0.04$, $\langle\langle f_T\rangle\rangle = 0.4$ (average value), $f_V = 0.8$, $\nu_p/\nu_l = 0.7$, and $t/(l + t) = 0.6$ gives a slope efficiency of about 0.5%. The threshold electrical energy for a 2-μs pulse is around 30 J. These estimated values are representative of the best (cw and long-pulse lasing) dyes with triplet quenching. For low–repetition-rate excitation (typically in the tens of hertz), actual laboratory or commercial systems indeed provide performance in this range for the best dyes (such as R6G and others; see Sec. 2.2.4) and also show nearly linear dependence of the output on input energy. For example, a commercially available high-energy system (by Cynoshure Inc.) provides output pulses in the 1–30-J range for electrical input energies of 250–8000 J and pulse durations of 0.5–4 μs for the most efficient (cw) dyes. For those dyes that can be made to lase only by fast pulses, the output efficiency is considerably lower. Fast flashlamps are also much smaller, providing considerably lower output energies. Today, pulsed-laser pumping is used exclusively for fast excitation. This is described in Sec. 3.2.

3.1.2 Practical Performance Limitations of Flashlamp Systems There are some additional performance-limiting effects that are not covered by the theory. It is well known that increasing the repetition rate greatly degrades performance. Both linear and coaxial flashlamps radiate large amounts of thermal (infrared) energy. Most of the unabsorbed visible and UV light is also eventually converted to heat. All this excess energy must be dissipated by liquid cooling. There are various designs for practical cooling systems, but they all rely on circulating, water-based heat exchangers. In addition, the dye itself is circulated by a separate system to provide both cooling and easy replacement of dye molecules lost to long-term deterioration. Up to a certain maximum repetition rate, the cooling system is very effective in carrying away the excess heat, and the output power increases linearly with the repetition rate. However, above a certain repetition rate, first some gradual decrease of the slope efficiency begins, followed by degradation of performance for still higher rates. As an example, while average powers ~30 W have been reported (Jethwa and Schäfer, 1974) at a repetition rate of 30 Hz, the average output power does not increase linearly for higher repetition rates. Thermal blooming as well as actual deterioration of the dye is one of the most important practical limiting effects in high-power, high-energy, flashlamp-pumped dye lasers, and it is difficult to obtain efficient operation at high (>100 Hz) repetition rates. Another important point in practical designs has to do with dye deterioration due to UV light absorption. Spectral studies of the emission produced by flashlamps indicate that a substantial portion of the photon flux peaks around 200 nm. Such emission gives rise to excitation to higher singlet states and can result in the photodegradation of the visible-emitting dyes. For this reason UV filters (Calkins *et al.*, 1982) (Pyrex or glass) have been used in the dye cell, or UV absorbers (Knyazev *et al.*, 1983) have been incorporated into the liquid refrigerant.

3.2 Pulse-Laser–Pumped Dye Lasers

While flashlamp-pumped lasers were historically very important and are still in use, today pulse-laser–pumped dye lasers are much more important. To select an appropriate pump for a particular dye (or group of dyes), the main consideration is the same as it was for the cw systems: the location of the singlet absorption bands. For best efficiency, the pumping should take place in the main (lowest) singlet absorption band (S_0–S_1 transition). Copper-vapor and pulsed xenon lasers emit blue-green lines and are suitable for pumping dyes that lase in the yellow to red spectral region. The frequency-doubled Nd:YAG laser is another suitable pump source. Dyes emitting in the infrared spectral region and wavelength longer than 1.2 μm can be pumped with the fundamental of Nd:YAG lasers. For UV and blue-green emitting dyes, excimer or nitrogen lasers with their intense UV output are excellent pump sources. These last two lasers can also be used to pump yellow-red emitting dyes. This is because the UV output of these two lasers falls into a higher-order singlet absorption band (usually S_0–S_2 transition; see Fig. 1) of these dyes. While the efficiency is obviously reduced by the considerably larger photon energy downshift, and there may also arise some problems with UV-induced photodegradation, such pumping has proven to be both easy and practical. Most of the mentioned pump sources will yield dye-laser pulses a few nanoseconds in duration and provide highly energetic pulses in the 100–500-kW range. It is obvious then from the discussion presented in Sec. 3.1.1 that, because of their short-pulse durations, these types of short-duration pump sources allow the lasing of a number of dyes that cannot be made to lase in cw or long-pulse mode.

For some of the pulsed pump sources (e.g., Xe, YAG, copper-vapor lasers) longitudinal excitation of the dye is possible. The Xe laser can operate with microsecond-long pulses, and most of the discussion on cw dye lasers applies. Pumping with a Xe laser has been used to construct an inexpensive long-pulsed and mode-locked femtosecond dye-laser system (Ansari *et al.*, 1985). Copper-vapor pumping also has been used for long-pulse operation, and the characteristics of such dye lasers have been investigated (Singh *et al.*, 1993). Longitudinal excitation is the preferred method of pumping for making dye-laser amplifiers. Until the advent of the mode-locked Ti:sapphire laser, the mode-locked cw dye laser was the only source of sub-100-fs pulses. Because these pulses are rather low in energy, amplification is necessary. Both Nd:YAG and copper-vapor lasers have been used to pump pulsed dye-laser amplifiers for femtosecond pulses (Fork *et al.*, 1987; Knox, 1988). In spite of its temperamental behavior, the copper-vapor laser has for many years been the workhorse pump for amplifying mode-locked dye-laser pulses. By dye-laser–based amplification of such pulses, fiber-based chirping and compression, and further dye-laser–based amplification, ~10 fs pulses with energies in the 100 mJ range have been achieved (Boyer *et al.*, 1991). Today, Ti:sapphire-laser–based all solid-state femtosecond oscillator amplifiers are replacing most of these femtosecond dye-laser systems. A similar, dye-laser–based amplifier has been recently demonstrated for the amplification of sub-100-fs pulses from a mode-locked cw Ti:sapphire laser, for the purpose of producing UV ultrashort pulses (Lucia and Reilly, 1995). The output of one dye laser can also be used to pump another longer-wavelength dye laser, usually (but not exclusively) in longitudinal configuration.

An alternative, and very often used, method of pumping for dye-laser oscillators is the transverse pumping configuration. This method has the advantage of being easy to implement and circumvents the sometimes problematic interaction of the pump beam with the optical elements of the dye-laser oscillator cavity in longitudinal excitation geometry. Additionally, output pulse energies can be high. One disadvantage of the transverse pumping geometry is the uneven spatial excitation of the active volume. Because this pumping scheme is very widely used, the pump sources are described in more detail. Most commonly used pump sources are various excimer lasers, and (historically earlier, but less often today) the N_2 laser. These are gaseous lasers, typically excited by a long transverse electrical discharge. Their gains are very high, and instead of a cavity, they usually have only one end mirror. Thus, the output is essentially a two-pass–amplified (and gain-narrowed) spontaneous emission. The output spectrum is rather broad and incoherent, but narrow enough to allow all the emitted light to be efficiently absorbed by a dye. As a result, much less of the light energy is converted into heat than in flashlamp pumping. N_2 lasers emit a number of closely spaced lines near 337 nm, corresponding to transitions between the C and B bands of the triplet manifold of the molecule. They typically provide peak powers of ~100 kW with pulse lengths ~10 ns (i.e., ~1 mJ energy) at repetition rates around 100 Hz. Excimers are molecules that exist only in the excited state (the upper level of the laser transition). Those used for dye pumping are excimers of rare-gas halides, which are formed in a pulsed electrical discharge. Depending on the gas mixture employed, these lasers emit with high intensity in the ultraviolet at several wavelengths from 193 to 353 nm. Specific examples of halides used in excimer lasers, and the wavelengths emitted, are KrF, 248 nm; XeCl, 308 nm; XeF, 351 nm. Typically, excimer lasers provide pulse lengths in the 10–30-ns range with repetition rates from less than 1 Hz to ~100 Hz. The pulse output energy of commercially available systems is in the 10- to 1000-mJ range, with the KrF being the most energetic. Thus the peak power is even higher than that of the N_2 laser. As a consequence of the transverse discharge geometry, the output beams of both N_2 and excimer lasers are of rectangular cross section. While this characteristic may be inconvenient for many uses, it is very well suited for transverse pumping a dye cell of the same length as the longer dimension of the output-beam rectangle. The basic configuration is shown in Fig. 17. The rectangular pump beam is focused by either a cylindrical or a spherical lens, resulting in a narrow and intense strip of light, which propagates through the dye cell in a direction orthogonal to the dye-laser beam. It is absorbed by the dye, resulting in a transversely nonuniform inverted region. UV light from the second or third harmonic of a pulsed solid-state (usually Nd:YAG) laser can also be utilized for transverse pumping. The basic configuration is the same, with some minor differences in the setup due to the circular output beam of these pump lasers. The less reliable copper-vapor laser has also been used for transverse pumping (Dixit *et al.*, 1997). Good beam quality and low divergence were achieved by using a special diffraction-filtered resonator. Today, however, the use of this pump would only be justified if no other pump source is available.

Early work with transversely laser-pumped dye lasers mainly utilized the N_2 laser (Meyer *et al.*, 1970; Shank *et al.*, 1970) and made only crude attempts at tuning, using a grating in place of one of the mirrors, as shown in the schematic of the basic setup (Fig. 17). A cylindrical lens of focal length 10–20 cm was used to focus the pump light into the dye cell. This resulted in a long pumped region with a small transverse cross section, which allowed easy lasing of many dyes. However, the beam quality was very poor. A more sophisticated N_2-laser–pumped, transverse excitation scheme was implemented by Hänsch (1972) and is described in Sec. 3.3, where tuning and bandwidth narrowing for pulse-pumped dye lasers are discussed. The transverse pumping approach is also readily applicable to excimer lasers (Duarte and Piper, 1980). Furthermore, this simple method of pumping has also been used for the excitation of amplifier stages where improved excitation has been demonstrated using two counterpropagating collinear pump beams (Klick *et al.*, 1985).

To understand in more detail the characteristics of transverse laser-pumped systems, one can still use the rate equations given in Sec. 1. However, the quasi-cw approximation used for the dynamics of the excited singlet population is no longer valid because the pump-pulse rise time and duration approach the singlet decay rate τ_1. The complete rate

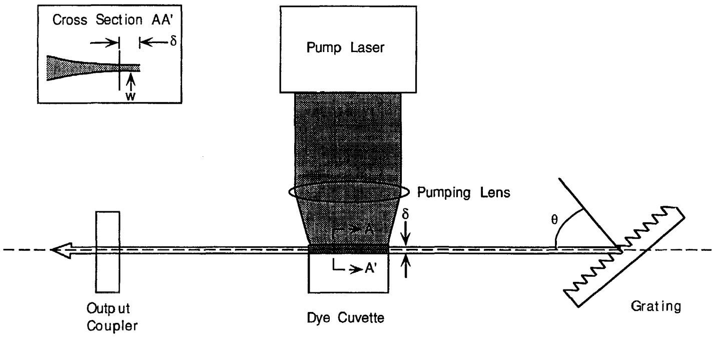

FIG. 17. Schematic of transverse pumping by N_2 or excimer lasers. The pump laser output is rectangular with the longer dimension shown. The focusing lens can be either cylindrical or spherical. The dye active region has length L and a cross-sectional area (A-A′) defined by the focused thickness of the pump beam (w) and the absorption depth in the dye (δ). The simple tuning arrangement of a diffraction grating (in the Littrow configuration) replacing one mirror is also shown.

equations must be solved *with* cavity risetime dynamics also included. Nevertheless, some important qualitative conclusions can still be reached by examination of the previously derived results. In particular, although for these systems Eqs. (8) and (9) are no longer valid, the gain is still inversely proportional to the cross-sectional area (A) of the active volume (perpendicular to the lasing axis). As shown in Fig. 17, this cross section is defined in one dimension by the focusing of the pump beam (to a width w) and in the other dimension by the absorption depth (δ) in the dye. The dye concentration is adjusted to give a short absorption depth, typically on the order of ~ 100 μm. The focused width of the pump beam (w) typically ranges from 0.1 to 1 mm. Thus, the cross-sectional area A is very small, only about one order of magnitude larger than in cw dye lasers, but the pump power is many orders of magnitude larger. The resulting gain is very high, 10–50 dB, with the higher values typical for good visible-wavelength dyes and the lower ones for the harder to lase blue and UV dyes. As a result, strong amplified spontaneous emission (ASE) is often observed without any mirrors. This ASE is normally undesirable because it is rather incoherent with a very wide bandwidth, and it will saturate the gain of the desired cavity oscillations. However, the ASE can be used to measure the gain (Dienes, 1975) and offers a convenient means of comparing various laser dyes whose parameters are not well known. Because of the high gain, cavity oscillation readily occurs even from the uncoated end surfaces of the dye cell, which must be tilted well away from the normal to avoid this occurring. The very high gain also allows placement of highly lossy elements (such as, for example, a diffraction grating) within the cavity, which is not possible with cw or flashlamp-pumped dye lasers. The use of gratings for frequency control is discussed in more detail in Sec. 3.3.

A phenomenon that occurs only in these short-pulse–pumped lasers is partial or full relaxation oscillations (also called self-Q-switching). This phenomenon is normally observed in solid-state lasers that have very long lifetimes (τ_1) that preclude equilibration of the population inversion. Dyes have very *short* lifetimes, and in all the dye lasers discussed so far, cw or quasi-cw conditions exist for the excited singlet population. Consequently, time-dependent cavity rate equations were not needed to describe their behavior. Here, however, the rise times of excimer (or nitrogen) laser pulses are *shorter* than τ_1; the pumping intensity is very high as discussed above; and the cavity Q is typically very low, i.e., the cavity lifetime is also short. As a result, relaxation oscillations can occur, whereby the population in the excited singlet level (N_1) rapidly builds up, resulting in very high gain. Intense stimulated emission rapidly depletes the population to near or below threshold, and lasing diminishes or stops (the lasing intensity can change rapidly when the cavity lifetime is fast). The process will repeat as long as the strong pump light is on. To describe this behavior quantitatively, time-dependent rate equations for the laser emission in the cavity need to be added to the dye rate equations (4) and (5). Detailed solutions of the complete rate equations have been made (Schäfer, 1990) for simulation of these lasers and yield the type of behavior qualitatively described above. Experimentally, both partial and complete relaxation oscillations have been observed (Lin and Shank, 1974) corresponding to the pulse shapes shown in Fig. 18. Subpicosecond pulses have also been produced by quenching all but the first burst (Lin and Shank, 1974). The same type of dynamics is responsible for the generation of ultrashort (picosecond duration) pulses in short-cavity dye lasers (Bor, 1980) (usually axially pumped). In these lasers, the cavity decay time is very fast because of the short cavity and the resulting short round-trip time. When pumped with nanosecond- or subnanosecond-dura-

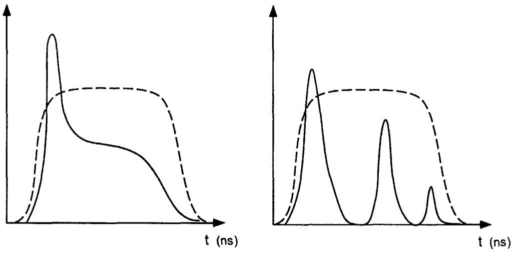

FIG. 18. Typical output pulse shapes that can be seen in N_2-laser–pumped dye lasers at high pump powers, showing partial and complete relaxation oscillations. Dashed line is pump pulse.

tion laser pulses, their output pulses can be as short ~15 ps as a result of the above described self-Q-switching mechanism.

An interesting and unique use of excimer laser pumping was for the generation of high-power tunable ultrashort pulses (Szatmári, 1992). This was accomplished by successive pulse shortening in a multistage cascaded dye-laser oscillator-amplifier system, also containing saturable absorbers. The pump light for all the active media was derived from one XeCl excimer laser in successively split-off beams. Through several stages of cascaded lasers combined with saturable absorption, highly effective pulse shortening takes place. The first three stages consisted of a transversely excimer-pumped, quenched (see paragraph above) dye laser (wavelength 340 nm, pulse length 0.2 ns), which axially pumped a short-cavity dye laser (365 nm, 15 ps). The output was passed through a gated saturable absorber. The resultant (9 ps) pulses transversely pumped the master oscillator, which was another short-cavity dye laser of the distributed-feedback type (see end of Sec. 3.3) producing tunable 0.5-ps pulses. These were then further amplified by a two-stage amplifier (both transversely excimer pumped) with still another saturable absorber for reducing the ASE. The resulting pulses were of 500-fs duration with 0.1-mJ energy; these were then frequency doubled and further amplified in an excimer laser amplifier to result in terawatt-power UV ultrashort pulses. The overall setup is rather complicated, and the reader should consult the reference. Although this system has now been superseded by even more powerful and shorter pulse Ti:sapphire-laser–based systems, it was an important high-power ultrashort-pulse source and provides a very good illustration of the complex systems possible with laser-pumped dye lasers. Both flashlamp and laser-pulse pumping allows the lasing of a large number of dyes (in addition to the efficient dyes used for cw lasing). In particular, dyes down to wavelengths as short as 310 nm have been made to lase, and the range has also been pushed into the infrared to 1.8 μm (Elsaesser and Kaiser, 1992). Table 3 gives a partial list of those dyes that are available commercially (e.g., Exciton Inc). More complete lists can be found in the Further Reading. Flashlamp excitation can also be used for a number of these dyes, but many require the small, fast-rise-time lamps. Laser excitation is very easy to implement but is more expensive. One of the dyes (4-methylumbelliferone) provides an extraordinarily wide tuning range from a single solution (Dienes *et al.*, 1973). No dyes lasing at wavelengths longer than 1.3 μm are included as these are not readily available, are rather unstable, and are seldom used.

Finally, in this section we should mention dye lasers utilizing for gain media dyes dissolved not in liquids, but in solid matrices. This is a relatively recent development and has obvious advantages of ruggedness and compactness. Silica sol-gel and polymeric matrices have both been employed as hosts, for dyes lasing in the blue and near-UV range (Brouwer *et al.*, 1995; Costella *et al.*, 1996; Chao *et al.*, 1996). These dyes are considerably more stable than those lasing in the longer visible and infrared wavelengths. However, even these shorter-wavelength solid-matrix dye lasers have a limited lifetime, much shorter than lasers based on inorganic crystals. Typically their lifetime ranges from a few thousand to several times ten thousand pulses. The simplicity of preparation and the very low cost compensate for the relatively short lifetime, and these simple lasers may find some useful niche applications.

3.3 Tuning and Bandwidth Reduction

Tuning and linewidth reduction for cw dye lasers have already been described in detail in Sec. 2. The intrinsic emission bandwidth of a pulsed dye laser employing a cavity formed by broad-band mirrors is even broader than under cw conditions and can span tens of nanometers. In general, the same principles for linewidth reduction and tuning apply for the pulsed configuration as for the cw case. An important difference is that in the transverse laser-pumped configurations, the gain is very high because of the small cross section of the active region and the high absorbed pump power. As a result, tuning and frequency-control elements with rather high insertion losses can be placed into transversely laser-pumped dye-laser cavities. A single frequency-selective element can offer tuning with a fairly broad bandwidth. A hierarchy of elements is required

Table 3. Pulsed laser dyes.

Dye	Lasing max.[a] (nm)	Wavelength range (nm)	Best laser pump (nm)
DMT	332	311-353	KrF(248)
p-terphenyl	340	335-346	XeCl(308), N_2(337)
TMQ	356	344-366	XeCl(308)
BPBP-365	361	354-385	XeCl(308)
Exalite 389	387	375-396	XeCl(308), N_2(337)
PBBO	396	378-440	N_2(337)
Exalite 400E	397	387-427	XeCl(308)
Stilbene 420	425	405-467	XeCl(308), N_2(337)
Coumarin 450	451	430-482	XeCl(308), N_2(337)
4-methyl-umbelliferone	Note b	370-600	N_2(337)
LD 470	470	450-495	XeCl(308)
Coumarin 480	480	454-523	N_2(337)
Coumarin 481	505	473-555	XeCl(308)
Coumarin 540A	544	517-612	XeCl(308), N_2(337)
Disodium fluorescein	555	525-575	N_2(337)
Rhodamine 590	574	563-635	XeCl(308), N_2(337)
Rhodamine 610	600	588-644	XeCl(308), N_2(337)
Rhodamine 640	618	608-668	XeCl(308), N_2(337)
Cresyl Violet 670	682	651-706	XeCl(308), N_2(337)
Nile Blue 690	710	688-783	XeCl(308), N_2(337)
Oxazine 720	702	660-725	XeCl(308), N_2(337)
Oxazine 750	767	734-807	XeCl(308), N_2(337)
HITC	835	805-908	XeCl(308), N_2(337)
DOTC	791	752-839	N_2(337)
IR 125	915	885-947	XeCl(308), N_2(337)
IR 140	964	906-968	N_2(337)
DNTPC-P	1124	1102-1148	NdYAG(1060)
IR-26	Solvent dependent	1150-1320	NdYAG(1060)

[a] Lasing maxima are approximate and are solvent, concentration, etc. dependent.
[b] See Dienes *et al.* (1973).

for narrow linewidth and precise frequency control. A simple, single tuning-element configuration that can tune the laser and also considerably reduce the laser linewidth consists of a grating mirror resonator (shown in Fig. 17). Early in the history of dye lasers, the use of this configuration enabled Soffer and MacFarland (1967) to demonstrate a linewidth of 0.06 nm from a frequency-doubled ruby-laser–pumped R6G dye laser. The basic principle can be understood by examining Fig. 17, where it is seen that the mirror reflector used at one end of a typical cavity has been replaced by a diffraction grating in the so-called Littrow configuration. In such a configuration a diffraction grating has the property that (for a given order) an incident beam will be reflected back exactly along the direction of incidence, provided

$$2d \cos q = m\lambda, \tag{21}$$

where d is the ruling distance, q is the angle between the incident ray direction and its projection on the grating surface, λ is the optical wavelength, and m is the order of diffraction. It is evident that the laser will operate at the wavelength that satisfies the above condition. The frequency discrimination of this setup depends on the divergence of the beam incident on the grating, as well as on the number of lines illuminated.

A more sophisticated approach uses the concept of beam expansion in the cavity. By expanding the beam, the divergence is decreased, and the number of grooves being illuminated at the grating is increased. The single-pass resolution $\Delta\lambda$ is given by

$$\Delta\lambda = \Delta\theta\left[M\frac{\partial\phi}{\partial\lambda}\right]^{-1} = \Delta\theta\left[M\frac{2\tan\phi}{\lambda}\right]^{-1}, \tag{22}$$

where $\Delta\theta$ is the beam divergence, $\partial\phi/\partial\lambda = 2 \times \tan\phi/\lambda$ is the total dispersion provided by the grating, and M is the beam magnification factor in the cavity, which improves the resolution by a factor M. Beam magnification is especially necessary in the transverse-pumped configuration. Because only a thin region near the inner wall of the dye cell is excited by the pump light (see Fig. 17), diffraction causes a substantial angular spread of the emerging light. Since the resonant condition depends on the incident angle and wavelength, and because only a small area of the grating would be illuminated by the beam, this would severely limit the resolution obtainable with the grating. Telescopic beam expanders (Hänsch, 1972) and prism beam expanders (Duarte and Piper, 1980) have both been used, but the latter is being used almost exclusively today. Telescopic expanders provide a beam magnification of 20 to 50 times and have a length that depends on the focal length of the lenses utilized. The intracavity lenses must be antireflection coated to minimize Fresnel losses and must add elements that compensate for aberrations. Such a system was first implemented by Hänsch (1972) and resulted in a very narrow output bandwidth. Using an inverted telescope with lenses of focal length 8.5 and 185 mm and a high-order echelle grating in a cavity ~40 cm long, an output linewidth of ~3 pm (or about 2.5 GHz at the R6G wavelengths) was achieved. Further reduction of the bandwidth to about 300 MHz resulted from the addition of a Fabry-Pérot etalon with a finesse of 20, made of coated solid fused silica. The wavelength was coarsely tuned by changing the grating angle and finely tuned by adjusting the etalon angle. An important characteristic of this laser was that, since the cavity was relatively long, only a few cavity round-trip passes could take place during the duration of the pump pulse. Therefore, no discrete axial-mode structure existed in this dye laser. Instead, the output was a continuum of frequencies with bandwidth about the same as that of the narrowest wavelength-selector element. As a result, the wavelength could be tuned smoothly without complicated feedback control of the cavity length as is necessary in cw dye lasers (see Sec. 2.2.5).

Today, transverse laser-pumped dye lasers with prism expanders have replaced the above system for current usage. The basic design considerations are very similar, with the main difference being in the geometric arrangement of the prism expander. A typical setup is shown in Fig. 19 and is essentially the same as first reported by Duarte and Piper (1980), but with pumping by excimer lasers or by the doubled output of Nd:YAG laser now most common. Two right-angled prisms are utilized together with a high-order echelle grating in the Littrow configuration. Although a large expansion factor (*M*) can be obtained even with a single prism, this would necessitate an angle of incidence very close to 90°, resulting in high losses and therefore reduced power output and a high fraction of undesirable broad-band amplified spontaneous emission (ASE). Using two prisms, the same expansion factor is obtained at smaller incidence angles, resulting in better efficiency with equally good resolution. The fraction of the undesirable ASE is also greatly reduced. The dye-laser output is taken from a broad-band coupling mirror (*M*1), which has high transmission (~90%) due to the high laser gain. With only the grating and the prisms in the cavity, a linewidth around 0.8 pm can be achieved (Duarte and Piper, 1980). Further reduction, as noted before, can be obtained by additional hierarchic frequency selectors (e.g., etalons). It should be noted that, for this type of arrangement, the dispersion of the prisms adds to the main contribution from

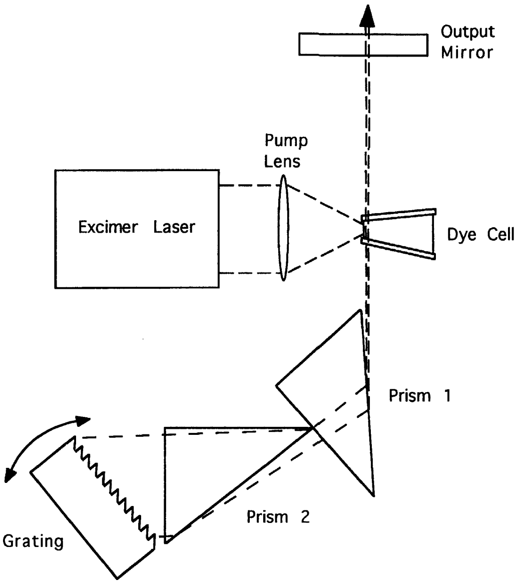

FIG. 19. Typical narrow-band excimer-laser–pumped dye laser showing a two-prism beam expander and a diffraction grating for bandwidth narrowing and wavelength tuning (after Duarte and Piper, 1980).

the grating and must be included as a small correction in the calculation of the linewidth. Detailed theories of prism beam expanders have been reported (Duarte and Piper, 1980, 1983).

Longitudinally laser-pumped configurations are tuned and bandwidth narrowed by similar methods, except that beam expansion is usually not necessary. Recently, a diffraction grating and one intracavity etalon have been used to control precisely the output of a high-power, Nd:YAG-laser–pumped pulsed R6G dye laser (Bhatia and Keto, 1996). An absolute precision of ±0.4 pm has been demonstrated in high-efficiency (40%), high-energy (100-mJ output pulses) operation. Birefringent filters can also be used for tuning of laser-pumped dye lasers. Because of the high gain, up to four plates can be used to achieve very narrow-band output (Zagone and Hetherington, 1996). Because of the broad gain bandwidth of most dyes and because it is possible to mix several dyes in a single solution, multiple-wavelength output can also be readily achieved in laser-pumped pulsed dye lasers. The high gain easily allows the sharing of the dye gain medium by two or more cavities with independent tuning elements (Khare *et al.*, 1995; Ko *et al.*, 1996).

As a result of the much larger cross section of the active volume, flashlamp-pumped dye lasers have much lower gain than transverse laser-pumped systems. Therefore, intracavity gratings cannot be used for tuning and linewidth reduction. However, all the tuning elements suitable for cw dye lasers (described in Sec. 2.2.5) can be used, with similar results. In general, one is satisfied with a moderate linewidth in these systems, and there is normally no attempt to obtain single-mode operation. An alternative linewidth-reduction method used in the early days involved a flashlamp-pumped system that incorporated intracavity isosceles prisms in conjunction with a tuning mirror (Strome and Webb, 1971). (The same principle, with a single prism, was also used in one of the first tunable cw dye lasers; Dienes *et al.*, 1972). The prism method also opens the possibility of ring-type systems. Schäfer and Müller (1971) incorporated six 60° prisms into a ring cavity and obtained a linewidth of less than 0.05 nm with a 40-J electrical-input-energy pumping. An advantage of such a multiple-prism ring cavity is that it obviates the need for mirrors with broad-band dielectric reflective coatings. Aspects of linewidth reduction and tuning using prisms are covered in several references (Marowsky, 1973; Yamagishi and Inaba, 1976). Today, practical flashlamp-pumped lasers use single or multiple prisms as well as Lyot filters and etalons to achieve tuning and output bandwidth control. For long-pulse flashlamp-pumped systems, very narrow-band and even single-mode operation has been achieved (Mandl *et al.*, 1995) by essentially the same methods as are used for cw dye lasers.

Finally, the method of distributed feedback (DFB) should be briefly described. This method, which now is most widely used in semiconductor diode lasers, was first demonstrated in an N_2-laser–pumped dye laser (Kogelnik and Shank, 1971). To achieve distributed feedback, a gain and/or index grating is made along the length of the gain medium to provide frequency-selective feedback, which is governed by the well-known Bragg condition. When implemented in a transverse-pumped dye laser, the grating results either from pumping with two beams at an angle (Kogelnik and Shank, 1971; Bor, 1980), which makes an interference pattern in the total pump intensity, or from imaging a coarse transmission grating in the path of the pump beam onto the dye cell. The laser output can be from moderate to narrow width and can be tuned by changing the grating period. Another advantage of this method is that the mirrorless cavity has a decay rate that is pump-intensity dependent, which favors short-pulse formation (Bor, 1980). This type of tuning arrangement was used in the master oscillator of the terawatt system described in Sec. 3.2. Recently, a system with two independently tunable DFB dye lasers, pumped by the second and third harmonics of a picosecond Nd:YAG laser, has been reported (Mueller, 1996). The ~12-ps output pulses were continuously tunable, with computer control, from 400 to 750 nm.

4. APPLICATIONS OF DYE LASERS

The enormous continuously covered wavelength range, the great range of available powers, and the many different kinds of outputs ranging from cw single mode with a few-megahertz bandwidth to a continuous train of *femtosecond* pulses (cw mode locked), as well as to pulse-pumped with durations from many microseconds to subnanosecond at a variety of repetition rates, allow a countless number of possible applications for dye lasers. Detailed descriptions of the many applications are properly given elsewhere. Here only a few important ones are mentioned with some examples described in more detail. Dye-laser applications include the industrial and the medical and biological areas, where a particular wavelength range and moderate energies may be needed, but there are no narrow-bandwidth or short pulse-duration requirements, to experiments in isotope separation requiring narrow linewidth and precise control, together with high cw power. The most obvious and probably the most important applications, however, are in spectroscopy, in laser manipulation of atoms and molecules, and in measurements of various chemical reaction dynamics. Various pulsed-laser–pumped systems, such as the ones described in Sec. 3.3, are today the cornerstone of molecular reaction-dynamics studies, and they are ubiquitous in chemistry departments of research universities. An important subset of reaction dynamics and spectroscopy is ultrafast spectroscopy (*q.v.*), which involves measurements of processes in liquid and solid-state materials on a femtosecond time scale. The wide bandwidth and tunability of the dye laser is essential in all these applications. It should also be noted that all pulse-laser–pumped as well as mode-locked dye lasers can readily be frequency doubled in nonlinear crystals, further extending into the UV the available wavelengths. Although today solid-state lasers and tunable systems based on them, such as optical parametric oscillators (OPO's), cw, pulsed, and cw mode-locked, are rapidly making inroads into the domain of dye-laser applications, dye lasers remain widely used because no other source has demonstrated the flexibility and the availability of all wavelengths within the very large range from the UV to the infrared.

The following historically important application to high-resolution spectroscopy is a good illustration of the possibilities in this field. The narrow-band N_2-laser–pumped system of Hänsch described briefly in Sec. 3.3 was used to perform a precise measurement of the Rydberg constant (Hänsch *et al.*, 1974) by saturation spectroscopy of the red Balmer line in hydrogen and deuterium. To narrow further the spectral width of the source laser, a scanning Fabry-Pérot interferometer was added externally. This narrowed the bandwidth to 7 MHz, and also allowed precise tuning of the frequency. Saturation spectroscopy (Hänsch *et al.*, 1972) is a powerful cw (or long-pulse) pump-probe technique to find accurately the line center of Doppler-broadened transitions. In the version used for this experiment two counterpropagating laser beams are used to probe the absorption. Moving atoms "see" the two beams Doppler shifted in frequency in opposite directions. Consequently, only atoms with zero velocity along the propagation axis experience interaction with both beams. The strong saturating beam bleaches (saturates) the absorption for the weak probe, and thus a saturation signal is obtained at the true line center. The bandwidth of this signal is the pressure-broadened homogeneous linewidth, which is very narrow in a low-pressure discharge, providing great accuracy in locating the exact resonance. To measure the Rydberg constant, the absolute wavelength of the dye laser for maximum saturation signal was determined by comparison with a separate, stabilized He-Ne laser (locked to a hyperfine component of an $^{129}I_2$ line). The final result was a more precise determination of this fundamental constant to an accuracy of 4 parts in 10^9. Such a resolution attests to the great flexibility and utility of the dye laser as a tunable source. Although now a quarter century old, this *tour de force* experiment still stands as an example of the power of applicability of the dye laser. Similar fundamental measurements by precision spectroscopy are still being performed today using dye lasers. For example, the frequency-doubled output of a modified commercial cw dye laser (Coherent CR 699) at 486 nm was stabilized to a high-finesse Fabry-Pérot cavity, and was used for narrow-linewidth two-photon spectroscopy of the 1*S*–2*S* transition at 243 nm in atomic hydrogen and in deuterium (Leibfried *et al.*, 1995). By directly comparing the frequencies of this transition in the two different isotopes, the isotope shift was deduced and information about the nuclear structure of deuterium was extracted.

Countless experiments on molecular reaction dynamics have been performed using the nanosecond-range pulses from laser-pumped tunable dye lasers, such as the prism-tuned excimer-pumped system described in Sec. 3.3. An interesting subset is the study of excited-state reactions. Atoms or molecules can be prepared in an excited state in sufficient density to allow easy observation of excited-state reactions. As one of many possible examples, the reaction of two-photon–excited Xe and Kr atoms with molecular hydrogen has been investigated using this type of dye laser, which was frequency doubled into the UV (Kawasaki *et al.*, 1992). It is noteworthy that the pulses of this laser are powerful enough that even after frequency doubling sufficient power was available for two-photon excitation.

In contrast to the above-described experiments, ultrashort-pulse spectroscopy uses mode-locked laser pulses for pump-probe exploration of spectrally dependent events on a very short time scale. Tunability is minimal, as a large fraction of the available gain bandwidth is lasing. The shortest optical pulses reported to date (~6 fs) were derived from a mode-locked cw dye laser by external pulse compression (Fork *et al.*, 1987). In the spectral domain, such a pulse corresponds to a coherent (perfectly in phase) superposition of frequencies in a band as wide as the entire (S_0–S_1) absorption band of a typical dye. In a pump-and-probe experiment with such pulses (Fragnito *et al.*, 1989), the entire vibrational band of the S_1 level of the dye malachite green was excited by a pump pulse. The absorption changes were then probed with probe pulses from the same source with ~6-fs resolution. The measured three-dimensional surface of absorption vs wavelength and time showed very distinct oscillations with a time period of 60 fs. These were attributed to vibrational distortions of the ring structure that were coherently excited by the femtosecond pulse. These oscillations persisted for a time longer than 1 ps corresponding to the vibrational dephasing time. It is interesting to note that the continuing presence of these oscillations at the "long" delay time of 1 ps demonstrates that excitation was done on a time scale three orders of magnitude shorter! Other recent ultrashort-pulse spectroscopic applications of mode-locked dye lasers include investigations of fast chemical reactions. Among these are studies of unimolecular dissociation (Rosker *et al.*, 1988), the observation of short-lived intermediate states in bimolecular reactions (Scherer *et al.*, 1987), and investigations of the ultrafast dynamics of isomerization reactions (Banares *et al.*, 1992).

Although the above brief descriptions do no more than give a glimpse into the many applications possible with dye lasers, they do serve to illustrate the tremendous range of possibilities opened up by the family of different dye lasers available to researchers today.

ACKNOWLEDGMENTS

The authors gratefully acknowledge helpful discussions with Timothy Johnston of Coherent Inc., who also provided the data of Fig. 11, and with Joseph Day of Cynoshure Inc., who provided information and data on flashlamp-pulsed dye lasers.

Works Cited

Ansari, H., Dienes, A., Whinnery, J. R. (1985), *Opt. Lett.* **10,** 19–21.

Baltakov, F. N., Barikhin, B. A., Kornilov, V. G., Mikhnov, S. A., Rubinov, A. N., Sukhanov, L. V. (1972), *Zh. Tekh. Fiz.* **42,** 1459–1461 [translation (1973), *Sov. Phys.—Tech. Phys.* **17,** 1161–1163].

Banares, L., Neikal, A. A., Zewail, A. H. (1992), *J. Phys. Chem.* **96,** 4127–4130.

Bhatia, P. S., Keto, J. W. (1996), *Appl. Opt.* **35,** 4152–4158.

Bloom, A. L. (1974), *J. Opt. Soc. Am.* **64,** 447–452.

Bor, Z. (1980), *IEEE J. Quantum Electron.* **QE-16,** 517–522.

Boyer, G., Franco, M., Schamberiet, J. P., Migus, A., Antonetti, A., Georges, P., Salin, F., Brun, A. (1991), *Appl. Phys. Lett.* **58,** 1119–1121.

Brouwer, H.-J., Krasnikov, V. V., Hilberer, A., Wildeman, J. (1995), *Appl. Phys. Lett.* **66,** 3404–3406.

Calkins, J. E., Colley, E., Hazle (1982), *Opt. Commun.* **42,** 275–280.

Chan, I. M. (1979), *A Theory of cw Dye Laser Characteristics,* Davis, CA: University of California, Davis, unpublished M.Sc. thesis.

Chao Ye, King-Shun Lam, Kin-Pong Chik, Lo, D. (1996), *Appl. Phys. Lett.* **69,** 3800–3802.

Costela, A., Garcia-Moreno, I., Figuera, J. M., Amat-Guerri, F. (1996), *Opt. Comm.* **130,** 44–50.

Dienes, A. (1975), *Appl. Phys.* **7,** 135–139.

Dienes, A., Ippen, E. P., Shank, C. V. (1971), *Appl. Phys. Lett.* **19,** 258–260.

Dienes, A., Ippen, E. P., Shank, C. V. (1972), *IEEE J. Quantum Electron.* **QE-8,** 177–178.

Dienes, A., Shank, C. V., Kohn, R. L. (1973), *IEEE J. Quantum Electron.* **QE-9,** 833–843.

Dixit, S. K., Daulatabad, S. R., Shukla, P. K., Bhatnagar, R. (1997), *Opt. Comm.* **134,** 149–154.

Duarte, F. J., Hillman, L. W. (1990), in: F. J. Duarte (Ed.), *Dye Laser Principles, with Applications,* Boston: Academic.

Duarte, F. J., Piper, J. A. (1980), *Opt. Commun.* **35,** 100–104.

Duarte, F. J., Piper, J. A. (1983), *Opt. Commun.* **43,** 303–307.

Elsaesser, T., Kaiser, W. (1992), in: M. Stuke (Ed.), *Dye Lasers: 25 Years,* Topics in Applied Physics Vol. 70, New York: Springer-Verlag, pp. 95–109.

Fork, R. L., Brito-Cruz, C. H., Becker, P. C., Shank, C. V. (1987), *Opt. Lett.* **12,** 483–485.

Fragnito, H. L., Bigot, J. Y., Becker, P. C., Shank, C. V. (1989), *Chem. Phys. Lett.* **160,** 101–104.

Hänsch, T. W. (1972), *Appl. Opt.* **11,** 895–898.

Hänsch, T. W., Shahin, I. S., Schawlow, A. L. (1972), *Nature, Phys. Sci.* **235,** 63–68.

Hänsch, T. W., Nayfeh, M. H., Lee, S. A., Curry, S. M., Shahin, I. S. (1974), *Phys. Rev. Lett.* **32,** 1336–1340.

Ippen, E. P., Shank, C. V., Dienes, A. (1972), *Appl. Phys. Lett.* **21,** 348–350.

Jarret, S. M., Young, J. F. (1979), *Opt. Lett.* **4,** 176–178.

Jethwa, J., Schäfer, F. P. (1974), *Appl. Phys.* **4,** 299–305.

Johnston, T. (1988), *Opt. Commun.* **69,** 147–152.

Johnston, T. F., Proffit, W. (1980), *IEEE J. Quantum Electron.* **QE-16,** 483–488.

Johnston, T. F., Sasnett, M. W. (Eds.) (1992), *The Effect of Pump Mode Quality on the Mode Quality of the cw Dye Laser.* Laser Energy Distribution Profiles: Measurement and Applications. Boston: SPIE.

Johnston, T. F. J., Brady, R. H., Proffitt, W. (1982), *Appl. Opt.* **21,** 2307-2316.

Kawasaki, M., Matsumi, Y., Chattopadhyay, A., Shafer, N., Satyapal, S., Tasaki, S., Yi, W., Bersohn, R. (1992), *J. Phys. Chem.* **96,** 7010-7013.

Khare, R., Daulatabad, S. R., Vora, H. S., Bhatnagar, R. (1995), *Opt. Comm.* **114,** 275–279.

Klein, M. B., Shank, C. V., Dienes, A. (1973), *Opt. Commun.* **7,** 178–180.

Klick, D., Akerman, M. A., Tsuda, H., Supurovic, D. (Eds.) (1985), *High Power Broadband Dye Laser for Manufacturing Applications.* International Conference on Lasers. McLean, VA: STS Press.

Knox, W. H. (1988), *IEEE J. Quantum Electron.* **24,** 388–397.

Knyazev, B. A., Lebedev, S. V., Fokin, E. P. (1983), *Sov. J. Quantum Electron.* **13,** 146–150.

Ko, Do-Kyeong, Lim, Gwon, Kim, Sung Ho, Lee, Jongmin (1996), *Appl. Opt.* **35,** 1995–1998.

Kogelnik, H. (1966), "Modes In Optical Resonators," in: A. K. Levine (Ed.), *Lasers: A Series of Advances,* New York: Marcel Dekker.

Kogelnik, H., Li, T. (1966), *Appl. Opt.* **5,** 1550–1567.

Kogelnik, H., Shank, C. V. (1971), *Appl. Phys. Lett.* **18,** 152–153.

Kogelnik, H., Ippen, E. P., Dienes, A., Shank, C. V. (1972), *IEEE J. Quantum Electron.* **QE-8,** 702–704.

Kohn, R. L., Shank, C. V., Ippen, E. P., Dienes, A. (1971), *Opt. Commun.* **3,** 177–178.

Leibfried, L., Geiger, H., Huber, A., König, W., Pachucki, K., Prevedelli, M., Udem, T., Walz, J., Weitz, M., Zimmermann, C., Hänsch, T. W. (1995), in: M. Inguscio, M. Allegrini, A. Sasso (Eds.), *XII International Conference on Laser Spectroscopy,* Capri, Italy: World Scientific.

Li, K. K., Dienes, A., Whinnery, J. R. (1981), *Appl. Opt.* **20,** 407–411.

Lin, C., Shank, C. V. (1974), *Appl. Phys. Lett.* **26,** 389–391.

Lucia, J. P., Reilly, J. P. (1995), *Opt. Comm.* **114,** 125–130.

Maeda, M. (1984), *Laser Dyes, Properties of Organic Compounds for Dye Lasers,* Tokyo: Academic.

Magni, V., De Silvestri, S., Cybo-Ottone, A., Nisoli, O., Svelto, O. (1992), "Novel Resonator Design for Femtosecond Lasers," in: M. Stuke (Ed.), *Dye Lasers: 25 Years,* Topics in Applied Physics Vol. 70, New York: Springer-Verlag.

Mandl, A., Klimek, D. E., Chou, H. P., Litzenberger, L. (1995), *IEEE J. Quantum Electron.* **QE-31,** 346–351.

Marowsky, G. (1973), *IEEE J. Quantum Electron.* **QE-9,** 245–255.

Meyer, J. A., Johnson, C. L., Kierstead, R. D., Shorma, R. D. (1970), *Appl. Phys. Lett.* **16,** 3–6.

Mueller, A. (1996), *Appl. Phys. B* **63,** 443–450.

Peterson, O. G., Tuccio, S. A., Snavely, B. B. (1970), *Appl. Phys. Lett.* **17,** 245–247.

Pike, H. A. (1971), *CW Dye Lasers,* Ph.D. Dissertation, University of Rochester.

Rigrod, W. W. (1963), *J. Appl. Phys.* **34,** 2602–2609.

Rosker, M. J., Dantus, M., Zewail, A. H. (1988), *J. Chem. Phys.* **89,** 6113–6127.

Runge, P. K., Rosenberg, R. (1972), *IEEE J. Quantum Electron.* **QE-8,** 910–911.

Schäfer, F. P. (1990), *Dye Lasers,* Berlin: Springer-Verlag.

Schäfer, F. P., Müller, H. (1971), *Opt. Commun.* **2,** 407–411.

Schäfer, F. P., Schmidt, W., Marth, K. (1967), *Phys. Lett.* **24A,** 280–284.

Scherer, N. F., Khundar, L. R., Bernstein, R. B., Zewail, A. H. (1987), *J. Chem. Phys.* **87,** 1451–1453.

Shank, C. V. (1975), *Rev. Mod. Phys.* **47,** 649–657.

Shank, C. V., Dienes, A., Trozzolo, A. M., Myer, J. A. (1970), *Appl. Phys. Lett.* **16,** 405–407.

Shank, C. V., Edighoffer, J., Dienes, A., Ippen, E. P. (1973), *Opt. Commun.* **7,** 176–177.

Siegman, A. (1986), *Lasers,* Mill Valley, CA: University Science Books.

Singh, S., Dasgupta, K., Thattey, S. S., Kumar, S. (1993), *Opt. Comm.* **97,** 367–372.

Soffer, B. H., McFarland, B. B. (1967), *Appl. Phys. Lett.* **10,** 266–269.

Sorokin, P. P., Lankard, J. R. (1967), *IBM J. Res. Dev.* **10,** 162–167.

Strome, F. C., Webb, J. P. (1971), *Appl. Opt.* **10,** 1348–1353.

Szatmári, S. (1992), in: M. Stuke (Ed.), *Dye Lasers: 25 Years,* Topics in Applied Physics Vol. 70. New York: Springer-Verlag, Chap. 9.

Teschke, O., Holtom, G. (1974), *IEEE J. Quantum Electron.* **QE-10,** 577–579.

Teschke, O., Dienes, A., Whinnery, J. R. (1976a), *IEEE J. Quantum Electron.* **QE-12,** 383–395.

Teschke, O., Whinnery, J. R., Dienes, A. (1976b), *IEEE J. Quantum Electron.* **QE-12,** 513–515.

Valdmanis, J. A., Fork, R. L. (1986), *IEEE J. Quantum Electron.* **QE-22,** 1986.

Yamagishi, A., Inaba, H. (1976), *Opt. Commun.* **16,** 223.

Yarborough, J. M. (1974), *Appl. Phys. Lett.* **24,** 629–630.

Zagone, R. L., Hetherington, W. M. III (1996), *Appl. Opt.* **35,** 624–628.

Further Reading

Diels, J., Rudolph, W. (1996), *Ultrashort Pulse Laser Phenomena,* San Diego: Academic.

Duarte, F. J. (Ed.) (1995a), *Tunable Lasers Handbook,* Optics and Photonics Series, San Diego: Academic.

Duarte, F. J. (Ed.) (1995b), *Tunable Laser Applications,* New York: Marcel Dekker.

Hillman, L. W. (1990), in: F. J. Duarte (Ed.), *Dye Laser Principles with Applications,* Boston: Academic.

Maeda, M. (1984), *Laser Dyes, Properties of Organic Compounds for Dye Lasers,* Tokyo: Academic.

Mollenauer, L. F., White, J. C. (Eds.) (1987), *Tunable Lasers,* Topics in Applied Physics Vol. 59, Berlin, New York: Springer-Verlag.

Schäfer, F. P. (1990), *Dye Lasers,* Berlin: Springer-Verlag.

Siegman, A. (1986), *Lasers,* Mill Valley, CA: University Science Books.

TUNNEL DIODES

A. SEABAUGH AND R. LAKE, *Corporate Research and Development, Texas Instruments Incorporated, Dallas, Texas, U.S.A.*

Introduction 335
1. **Tunnel-Device Physics** 336
1.1 Resonant-Tunneling Diode 337
1.2 Esaki Diode 340
1.3 Valley Current 343
1.4 Speed 343
2. **Fabrication** 344
2.1 Tunnel-Junction Formation 344
2.2 Device Processing and Integration 345
2.3 Manufacturing Issues 346
2.3.1 Uniformity 346
2.3.2 Reproducibility 346
2.3.3 Reliability 347
3. **Device Properties** 347
3.1 Esaki Diode 347
3.1.1 Group IV 347
3.1.2 Group III–V 351
3.2 Resonant-Tunneling Diode 351
3.2.1 Group IV 351
3.2.2 Group III–V 352
3.2.3 Group II–VI 352
3.2.4 Mixed Crystalline 352
3.2.5 Amorphous Resonant Tunneling 352
3.2.6 Hole Resonant Tunneling 352
3.3 Resonant Interband Tunnel Diode 353
4. **Applications** 353
4.1 Triggering, Pulse Forming, and Sampling 353
4.2 Memory 354
4.3 Analog-to-Digital Conversion 354
4.4 Logic 355
4.5 Oscillators 355
4.6 Tunnel Photodiodes 355
5. **Summary and Prospects** 355
Glossary 355
Works Cited 356
Further Reading 359

INTRODUCTION

The tunnel diode was first discovered by Leo Esaki in the course of his work on bipolar transistors (Esaki, 1958, 1976). Others about this same time (1957) were on the verge of the same discovery (Pearson, 1957) but fell a measurement short of seeing the distinctive *N*-shaped current–voltage (*I–V*) characteristic of the tunnel diode (see Fig. 1). Since Esaki's original work, which lead to a Nobel prize in physics in 1973, the tunnel diode has been produced in most of the common binary and compound semiconductors, and in numerous heterostructure devices. Interest in the device for electronic applications arose almost immediately upon its discovery because of its remarkable multivalued *I–V* characteristic, which features both negative differential resistance and an ultrahigh-speed transient response. This uniqueness and versatility has retained commercial and academic interest even to the present; however, only discrete devices are commercially available. Device and circuit development is, for the most part, limited to research laboratories.

Current flow in the tunnel diode is governed by the principles of quantum-mechanical particle tunneling, a direct consequence of the fact that matter has both a wave and a particle aspect. (The term *diode* arises from its original discovery in semiconductor *p-n* junctions, but the tunnel diode bears little resemblance to the common *p-n* diode.) In the particular case of a semiconductor device, the particle aspect of electron transport is most often used to describe current flow. To engineer electron transport in a device, potential-energy barriers are commonly formed by the introduction of *p-n* junctions, oxide–semiconductor interfaces, or heterojunctions (junctions formed between dissimilar materials, e.g., AlGaAs/GaAs). An electron can cross over these barriers if it has total energy exceeding the energy barrier height; otherwise, classically, it is constrained to flow alongside the barrier. Ac-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

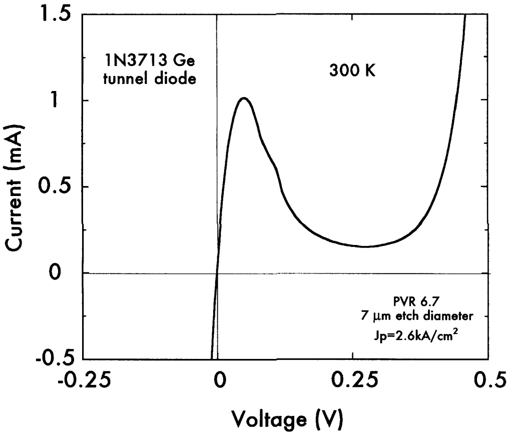

FIG. 1. Measured current–voltage characteristic for a commercial germanium Esaki tunnel diode.

cording to quantum theory, however, a particle such as the electron actually penetrates into any potential barrier for some distance, similar to the way light slightly penetrates an opaque metal surface. As a consequence of this wave penetration property, if a potential barrier is thin relative to the electron's Bloch wavelength, the particle can sometimes slip freely through the barrier without paying the energy costs to surmount it. This ghostlike passage through a barrier is referred to as tunneling.

In the original tunnel diode, also known as the Esaki diode, the barrier through which tunneling occurs is the space-charge depletion region of a *p-n* junction. Potential barriers can also be formed by heterojunctions. These junctions are usually produced by an epitaxial crystal growth process from semiconductors with similar crystallinity, but with different band-gap energies. At the heterojunction, a potential-energy discontinuity results, creating an abrupt potential barrier.

There are many devices that are referred to as tunnel diodes, which can refer to any two-terminal device where the charge transport can be attributed to tunneling; this includes semiconductor p^+n^+ Esaki diodes, resonant-tunneling diodes (RTD), resonant interband tunnel diodes (RITD), metal/insulator/semiconductor tunnel diodes, and metal/insulator/metal tunnel diodes. To limit the scope of this article we focus on devices that exhibit the characteristic *N*-shaped current–voltage behavior of the original Esaki diode. This limitation still covers the greatest body of work on tunnel diodes.

The RTD was first demonstrated by Chang *et al.* (1974), and, subsequently, the RITD was demonstrated by Söderström *et al.* (1989) and Luo *et al.* (1989). In addition to these semiconductor embodiments, the RTD has also been formed in metal– and semimetal–semiconductor heterojunctions (Suemasu *et al.*, 1992; Brehmer *et al.*, 1995) and amorphous semiconductor heterojunctions (Hirose *et al.*, 1977; Miyazaki *et al.*, 1987).

In this article we cover both the physics and technology of the Esaki and heterostructure tunneling diodes. We have surveyed and compared the properties of tunnel diodes across a wide variety of materials including semiconductor groups IV, III–V, and II–VI, and mixed-material structures. Only a few applications for the tunnel diode exist today, but the tunnel diode's high intrinsic switching speed, the fastest of any semiconductor device, and the opportunities for ultrafast and low-power circuits continue to attract attention. Applications and technology prospects are the subjects of the final sections.

1. TUNNEL-DEVICE PHYSICS

The two broadest classes of tunnel diodes that exhibit the *N*-shaped characteristic of Fig. 1 are the unipolar, intraband RTD (Mizuta and Tanoue, 1995; Brown, 1994; RESONANT TUNNELING) and the bipolar interband Esaki diode (Sze, 1981). At first glance, these two classes of structures appear to share little in common. The physics of electron transport through the RTD is based on resonant tunneling through a double-barrier structure analogous to a Fabry-Pérot cavity, whereas the physics of electron transport through the Esaki diode is based on single-barrier tunneling across the energy-band gap at a *p-n* junction.

A unifying perspective for these devices emerges if one considers the conservation laws that they share, namely, conservation of total energy and conservation of transverse momentum (*transverse* refers to the direction perpendicular to the current flow; *longitudinal* refers to the direction parallel to the current flow). We begin by considering an ideal structure, free of scattering from impurities, alloy disorder, interface roughness, and phonons. Later, we will consider the case with scattering, where an incident electron can transfer into a state of different total energy and transverse momentum.

In both intraband and interband devices, the electron transport is typically in the same direction as the crystal growth, generally (100), and the devices have wide cross-sectional area relative to the Bloch wavelength. Since, in the ideal structure, there is no force acting in the transverse direction, transverse momentum is conserved. While devices of small cross-sectional area are built in research laboratories (Reed *et al.*, 1988; Mizuta and Tanoue, 1995), we consider only wide–cross-sectional area devices, since they are of current technological interest. Similar energy and momentum conservation arguments can be used to understand the effects of small–cross-sectional area devices, in which the transverse energies are quantized (Liu and Aers, 1989; Bagwell *et al.*, 1990).

Discussions of tunnel-device current tend to focus on the calculation of the quantum-mechanical transmission coefficient. The widely used approaches are discretizing the potential into a series of steps and then connecting the wave functions at each step using a transmission matrix (Mizuta and Tanoue, 1995), a direct matrix solution of the tight-binding localized-orbital Hamiltonian (Frensley, 1994; Ting and McGill, 1995), a recursive Green's-function solution (Støvneng and Hauge, 1991), a transfer-Hamiltonian approach (Mizuta and Tanoue, 1995; Payne, 1986; Duke, 1969), or a Wentzel-Kramers-Brillouin (WKB) approximation (Sze, 1981). We instead focus on the implications of energy and momentum conservation, quoting the form of the transmission coefficients as needed.

The characteristics of most interest for these devices are the peak current density and the peak-to-valley current ratio (PVR). For high-speed applications such as logic, high current density is required for fast charging and discharging of the device and circuit capacitances. For low-power operation such as in memory, low current density is needed to minimize the static power dissipation. We develop here the underlying physics for device design for high and low peak current-density devices, and the factors controlling the PVR.

1.1 Resonant-Tunneling Diode

Analysis of the RTD begins with Fig. 2, which displays the conduction-band energy–position and energy–momentum relationships. The double potential barriers constituting the RTD are shown in the energy–position (E_x–x) diagram at the right of Fig. 2. The conduction-band profile is translationally invariant in the transverse (yz) plane. The quasi-Fermi levels in the left contact, E_F, and right contact, $E_F - eV$, are split by the applied voltage V. The horizontal line between the barriers represents the energy $E_{x,\text{res}}$, of the resonant state. Just as in a high–electron-mobility transistor, the resonant energy represents the bottom of a two-dimensional electron gas (2DEG). It is the total energy E of the state with zero transverse momentum. As bias is applied, the resonant level is pulled down to the Fermi level of the emitter. At this point, when $E_{x,\text{res}} = E_F$, the current turns on. As the bias is increased, the resonant level $E_{x,\text{res}}$ is pulled deeper into the incident flux from the Fermi sea of the emitter.

The parabola, just left of center in Fig. 2, represents the longitudinal energy dispersion, $E_x = \hbar^2 k_x^2/2m^*$, where $\hbar$ is Planck's constant ($h/2\pi$, $h = 4.135 \times 10^{-15}$ eV·s), k_x is the electron wave vector in the longitudinal direction, and m^* is the electron's effective mass in the crystal, which is typically 10 to 100 times less than the free-electron mass $m_0 = 9.109 \times 10^{-31}$ kg. The resonant energy $E_{x,\text{res}}$ corresponds to a resonant incident momentum $\hbar k_{x,\text{res}}$.

The Fermi sphere shown in Fig. 2 represents the relationship between the electron momentum in the transverse and longitudinal directions at a constant energy given by the Fermi energy. The indicated disk represents the transverse momentum states with the appropriate energy for resonance, $E_{x,\text{res}}$. The states represented by this disk are the ones that contribute to the current (Iogansen, 1965; Luryi, 1985). The current is obtained by multiplying the incident current flux per unit longitudinal energy E_x by the intrinsic width of the resonance, Γ_i, and by the number of transverse states in the shaded disk. The incident current flux per unit longitudinal energy is $2env$, where e is the electron charge ($e = 1.602 \times 10^{-19}$ C), n is the electron density per unit length and

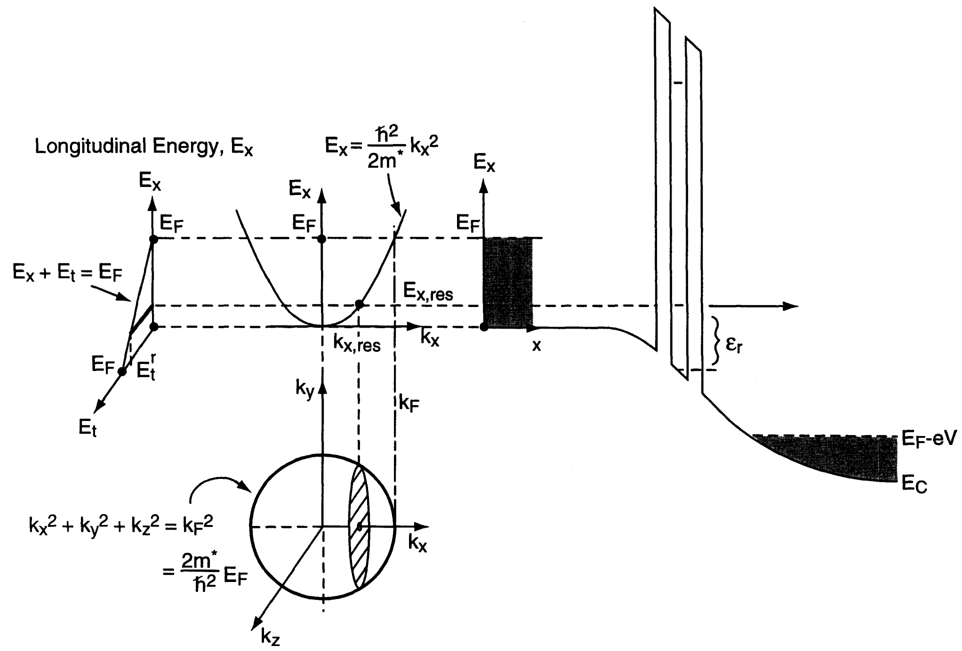

FIG. 2. Conduction-band profile of an RTD. $E_{x,\text{res}}$ is the total energy of the resonant state with zero transverse momentum. The parabola in the left contact (the emitter) represents the longitudinal energy dispersion. The Fermi sphere represents the momentum states in the emitter. The resonant states lie in the shaded circle. The triangle represents the same states in terms of energy. The resonant states lie on the horizontal line. The number of states on the line is the 2D density of states times the length of the line, $(m^*/2\pi\hbar^2)(E_f - \epsilon_r)$. The illustration was conceived by P. van der Wagt, Texas Instruments.

energy, $n = (2\pi\partial E_x/\partial k_x)^{-1} = m^*/2\pi\hbar^2 k_x$, $v = \partial E_x/\partial\hbar k_x = \hbar k_x/m^*$ is the velocity, and the factor of 2 is for spin degeneracy. Multiplying these factors together yields the incident current flux per unit longitudinal energy as $2e/h$. The number of transverse states is $1/(2\pi)^2$ times the area of the shaded disk. The area is $\pi(k_F^2 - k_x^2) = \pi(2m^*/\hbar^2)(E_F - E_{x,\text{res}})$. Multiplying the terms together gives the number of transverse states resonantly tunneling:

$$\frac{\text{number of transverse states}}{\text{unit area}} = \frac{m^*}{2\pi\hbar^2}(E_F - E_{x,\text{res}}). \tag{1}$$

Putting this all together, we get a resonant current flowing of

$$I_{\text{res}} = \frac{2e}{h}\Gamma_i \frac{m^*}{2\pi\hbar^2}(E_F - E_{x,\text{res}}),$$
$$E_c < E_{x,\text{res}} \leq E_F. \tag{2}$$

For a symmetric device, this simple derivation based on state counting differs by only a factor of $\pi/2$ from a rigorous derivation (Lake *et al.*, 1993). Note that the entire quantum mechanics of the problem is contained in Γ_i, the intrinsic resonance width. It is determined by the transmissivity of the left and right barriers (Lake *et al.*, 1993). The exact expression is

$$I_{\text{res}} = \frac{2e}{\hbar}\frac{\Gamma_L\Gamma_R}{\Gamma_L + \Gamma_R}\frac{m^*}{2\pi\hbar^2}(E_F - E_{x,\text{res}}),$$
$$E_c \leq E_{x,\text{res}} \leq E_F, \tag{3}$$

where Γ_L and Γ_R are the energy widths associated with tunneling out of the left and right barriers, respectively. The intrinsic resonance width resulting from tunneling through the barriers is $\Gamma_i = \Gamma_L + \Gamma_R$. The factor Γ_L is proportional to $v_x e^{-2\kappa d}$, where v_x is the incident longitudinal velocity, d is the left barrier width, and κ is the decay constant in the left barrier, $\kappa = \sqrt{2m^*(V_B - E_x)}/\hbar$. V_B is the potential of the left barrier. Γ_R is defined in the same way with respect to the right barrier. Equation (3) shows that the current density is exponentially sensitive to the barrier height and width.

Since the resonant energy tends to fall linearly with the applied bias, I_{res} increases linearly with bias. Once the resonant energy falls below the conduction band of the emitter, electrons can no longer tunnel while conserving both total energy and transverse momentum, and the current must flow through scattering channels (phonons, impurities, alloy, and interface disorder) or by thermionic emission.

This picture of resonant tunneling, to this point, mixes both energy and momentum. If we convert the momentum into energy, then the Fermi half-sphere for $k_x > 0$ becomes the Fermi triangle shown at the left of Fig. 2; $E_x = \hbar^2 k_x^2/2m^*$ and $E_t = \hbar^2(k_y^2 + k_z^2)/2m^*$. The shaded disk of transverse momentum is collapsed into the horizontal line of transverse energy. The number of resonant transverse states is then $(m^*/2\pi\hbar^2)E_{t,\mathrm{res}} = (m^*/2\pi\hbar^2)(E_{\mathrm{F}} - E_{x,\mathrm{res}})$, where $m^*/2\pi\hbar^2$ is the two-dimensional density of states (Beenakker and Van Houten, 1991). This is a useful way of viewing the source function.

The foregoing discussion works well for the unipolar, intraband RTDs. It is only entirely accurate when the Hamiltonian is separable. For three-dimensional emitters that contain a continuum of states in E_x, the differences between this idealized picture and the actual picture tend to get lost in the integration over energy. For two-dimensional emitters where electrons are being injected from a quantized state in the emitter, the picture can break down (Boykin *et al.*, 1995).

The above discussion has served to elucidate the common picture that one finds in the literature. However, we need to alter this picture to one that can also be used to analyze the interband, bipolar Esaki-type tunnel diodes. This picture is shown in Fig. 3. It is the one discussed by Ohno *et al.* (1990). The horizontal lines in the emitter represent the continuum of E_x states. Each energy E_x serves as the bottom of a two-dimensional subband represented by the parabolic dispersion. Note well that in contrast to Fig. 2, the parabolas in Fig. 3 represent the dispersion of the transverse energy E_t with respect to the transverse momentum $k_t = \sqrt{k_y^2 + k_z^2}$. The vertical axis on this plot is now *total energy E* as opposed to *longitudinal energy* E_x in Fig. 2. For an electron to tunnel, it must conserve total energy and transverse momentum. Figure 3 illustrates what this implies. All of the electrons in the emitter parabola that align with the well parabola can resonantly tunnel. The number of states in that parabola is directly read off to be $(m^*/2\pi\hbar^2)(E_{\mathrm{F}} - E_{x,\mathrm{res}})$, resulting in the same current I_{res} as before.

The energy of the resonance, ϵ_r, is determined by the transmissivity of the barriers and the width of the well. As a first approximation, the resonant level can be estimated by a particle-in-a-box calculation, $\epsilon_r = (\hbar^2/2m^*)(n\pi/w)^2$, where n is an integer equal to 1 for the first resonance, 2 for the second resonance, and so on; w is the width of the well, and ϵ_r in eV is measured from the bottom of the conduction band in the well (see

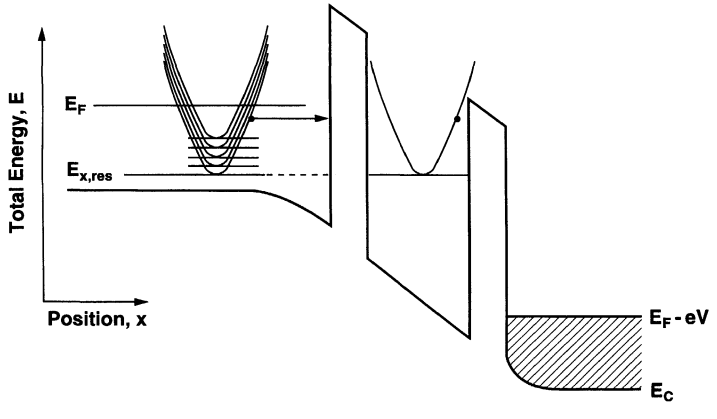

FIG. 3. A third picture of the incident states. Each longitudinal energy E_x is the bottom of a two-dimensional (2D) subband of transverse-momentum states. The resonant state in the well is also the bottom of a 2D subband of transverse-momentum states. The parabolas show the dispersion relations of the 2D subbands. The states in the emitter parabola that align with the parabola in the well can resonantly tunnel. This picture can be used for the interband tunnel devices.

Fig. 2). The well-width dependence of the resonant energy allows for good flexibility in engineering the voltage position of the peak current.

1.2 Esaki Diode

The band diagram of the traditional Esaki diode is shown under forward bias in Fig. 4(a). As bias is applied, electrons flow from the filled emitter states in the conduction band on the left to empty states in the valence band on the right. As bias is increased, the conduction band of the emitter is eventually raised above the valence band of the collector, electrons can no longer tunnel into a valence-band state while conserving both total energy and transverse momentum, and the current is reduced to a minimum determined by thermionic emission and scattering. As in Fig. 3, each longitudinal energy E_x serves as the bottom of a transverse subband. However, now the subbands are inverted from one another. For the two subbands shown, only those electrons can tunnel that lie in states on the intersecting ring shown in Fig. 4(b).

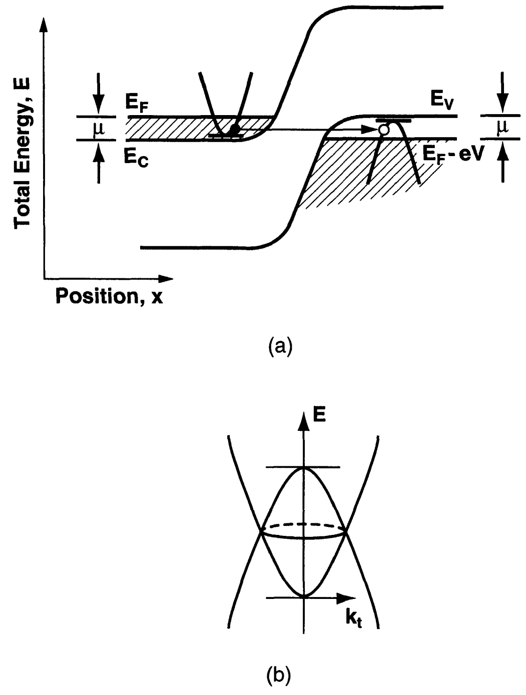

FIG. 4. (a) The band diagram of a traditional Esaki diode under forward bias. Each longitudinal energy E_x serves as the bottom of a 2D subband of transverse states; however, now the subbands are inverted from one another. (b) For the two subbands shown in (a), only those states that lie on the intersecting ring can tunnel conserving both total energy and transverse momentum.

To focus on the effect of energy and momentum conservation, we consider a direct-band-gap material and calculate the zero-temperature current assuming that the transmission coefficient is constant, the electron and hole effective mass are equal, and the doping is such that $\mu = E_F - E_c = E_v - (E_F - eV)$. The general expression for the current is

$$I = \frac{2e}{h}\int dE\,\frac{m^*}{2\pi\hbar^2}\int dE_t T(E,E_t) \times [f(E - E_F) - f(E - E_F + eV)], \qquad (4)$$

where $T(E,E_t)$ is the transmission probability, f is the Fermi factor $f(E) = [\exp(E/k_BT) + 1]^{-1}$, T is temperature, and k_B is Boltzmann's constant ($k_B = 8.617 \times 10^{-5}$ eV/K). At zero temperature, the Fermi factors are step functions that determine the limits on the integration over total energy; e.g., $f(E - E_F) = \theta(E_F - E)$. The total energy E is integrated between the two Fermi levels, E_F and $E_F - eV$. For an electron to tunnel from the left to the right in Fig. 4, its initial and final longitudinal energy, E_x, must lie above the conduction band on the left and below the valence band on the right, respectively. The limits on the integration over transverse energy, $E_t = (\hbar^2/2m^*)(k_y^2 + k_z^2)$, are understood by noting that $E_x = E - E_t$ in the conduction band and $E_x = E + E_t$ in the valence band. For a bias V such that $0 \le eV \le \mu$, the current is proportional to the integral

$$I \propto \left[\int_0^{eV/2} dE \int_0^{E-E_C} dE_t + \int_{eV/2}^{eV} dE \int_0^{E_V-E} dE_t\right] = eV\left(\mu - \frac{3}{4}eV\right), \quad 0 \le eV \le \mu, \qquad (5)$$

where we have taken the quasi-Fermi level of the right contact as the reference energy, i.e., $E_F - eV = 0$, so that $E_c = eV - \mu$ and $E_v = \mu$. Taking the derivative of Eq. (5) with respect to voltage, we see that the current peaks when $eV = \frac{2}{3}\mu$ (Sze, 1981). For a bias V such that $\mu \le eV \le 2\mu$, the current is proportional to the integral

$$I \propto \left[\int_{E_C}^{eV/2} dE \int_0^{E-E_C} dE_t + \int_{eV/2}^{E_V} dE \int_0^{E_V-E} dE_t\right]$$
$$= (\mu - eV/2)^2, \quad \mu \le eV \le 2\mu. \tag{6}$$

Since $E_C > E_F - eV$ and $E_F > E_V$, the total energy is now integrated only between E_C and E_V.

A plot of I vs V from Eqs. (5) and (6) is shown in Fig. 5. The turnoff of the ideal Esaki diode is gradual, determined by the decreasing number of states available for tunneling as the bias is increased. The turnoff of the ideal RTD is more abrupt, determined by the resonance width.

The tunneling transmission probability is usually approximated by the expression for Zener tunneling under a uniform electric field (Sze, 1981),

$$T = \exp\left(-\frac{\pi m^{*1/2} E_G^{3/2}}{2\sqrt{2}\hbar e\mathscr{E}}\right) \exp\left(-\frac{2E_t}{\bar{E}}\right), \tag{7}$$

where E_G is the band gap, $\mathscr{E}$ is the electric field, E_t is the transverse kinetic energy, and

$$\bar{E} = \frac{4\sqrt{2}\hbar e\mathscr{E}}{3\pi m^{*1/2} E_G^{1/2}}. \tag{8}$$

It is interesting to note that the transmission probability in Eq. (7) is only a function of the transverse kinetic energy and independent of the longitudinal energy E_x. Under a uniform electric field, the barrier presented by the band gap is the same for any E_x.

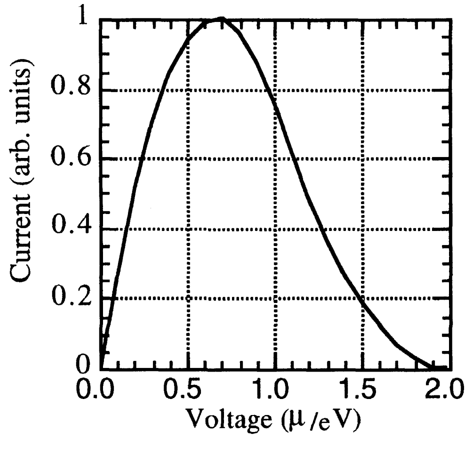

FIG. 5. Esaki-diode current–voltage characteristic calculated from Eqs. (5) and (6).

From the perspective of a device designer, Eqs. (7) and (8) teach that there is little flexibility in designing homojunction Esaki diodes for a specific current density. The electric field is set by the built-in potential of the p-n junction, which is determined by the magnitude and spatial abruptness of the doping transition. The band gap is fixed for a given material. Once the electric field is determined, only the device area can be used to modify the current density.

Recently, heterostructure Esaki diodes have been built (Tsai *et al.*, 1994; Day *et al.*, 1993); see Fig. 6. In these devices, the tunneling occurs between an electron 2DEG in the n^+ $In_{0.53}Ga_{0.47}As$ notch and the light-hole 2DEG in the p^+ $In_{0.53}Ga_{0.47}As$ notch. For these devices, the two subbands shown in Fig. 6 are the only two subbands that contribute to the current. (Actually, there is a heavy-hole subband that gives a small secondary peak in the I–V curve.) These structures are resonant triple-barrier devices (Iogansen, 1965; Malov *et al.*, 1988; Xu *et al.*, 1992; Payne, 1985), in which the resonant states on the intersecting ring of Fig. 4(b) coherently couple and split into bonding and antibonding states. For an alignment as shown in Fig. 4(b), in which the transverse energy of the electron and hole state is much greater than either the coherent splitting re-

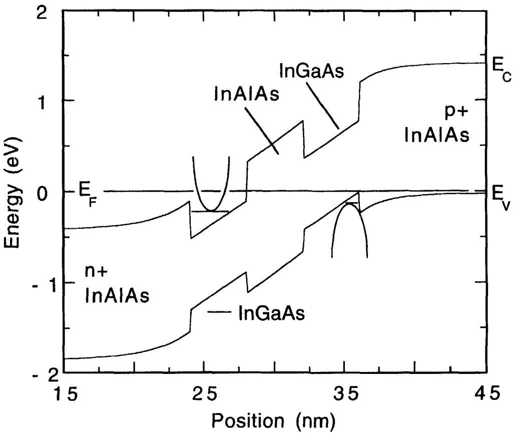

FIG. 6. Computed energy-band diagram for the published resonant interband tunnel diode of Tsai *et al.* (1994) exhibiting a peak-to-valley current ratio of 144 at room temperature. The parabolas represent the two-dimensional electron and hole subbands. (Computed using the heterojunction modeling program *Bandprof* of W. R. Frensley.)

sulting from the coupling or the intrinsic broadening resulting from leakage into the emitter and collector contacts, the number of tunneling states becomes independent of bias. There is no theoretical work explaining the shape of the *I–V* characteristics shown by Tsai *et al.* (1994) and Day *et al.* (1993). The *I–V* characteristics of the optimized devices that show the highest peak-to-valley ratio (see Fig. 1 of Tsai *et al.*, 1994, and Fig. 2 of Day *et al.*, 1993) exhibit the effects of quantum well charging discussed by Malov *et al.* (1988).

For interband tunneling in indirect-gap semiconductors such as silicon, only those conduction electrons in the longitudinal X valleys (on the k_x axis) can directly tunnel into the valence-band Γ valley while conserving both transverse momentum and total energy (see Fig. 7). The potential $V(x)$ breaks the crystal symmetry along the x axis so that k_x is no longer a good quantum number. k_y and k_z, however, must be conserved in the transition from the X valley of the conduction band to the Γ valley of the valence band. The transverse momentum in the longitudinal X valleys lies close to the k_x axis in the same region as the transverse momentum of the Γ valley valence-band states. The electrons in the transverse X valleys can only tunnel by gaining or losing the large transverse momentum difference between the transverse X valleys and the Γ valley. These electrons can obtain this momentum difference by picking up a Fourier component of a scattering potential. For a scattering potential to have a large momentum Fourier component, it must be well localized in space. The electrons in the transverse X valleys tunnel with the assistance of scattering from localized states in the band gap and from phonons. Theoretically predicted *I–V*'s resulting from scattering assisted tunneling are shown in Fig. 7 of Sze (1981).

There is one final resonant tunneling device, which is a hybrid of the conventional intraband RTD shown in Fig. 2 and the Esaki diode. It is the resonant interband tunneling diode (RITD) shown in Fig. 8 (Söderstöm *et al.*, 1989). In this device, electrons tunnel from the InAs conduction band on the left, through a light-hole 2DEG in the GaSb well, and back into the InAs conduction band on the right. When enough bias is applied, the hole 2DEG is pushed below the conduction band of the emitter, and electrons can no longer tunnel while conserving total energy.

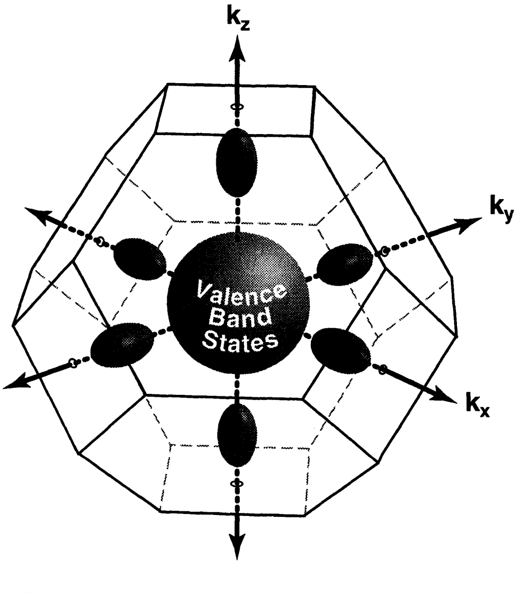

FIG. 7. For an indirect-gap semiconductor such as silicon, only the electrons in the longitudinal X valley (on the k_x axis) can directly tunnel into a valence-band state in the Γ valley while conserving both total energy and transverse momentum. The electrons in the transverse X valleys tunnel with the assistance of scattering from localized states in the band gap and from phonons.

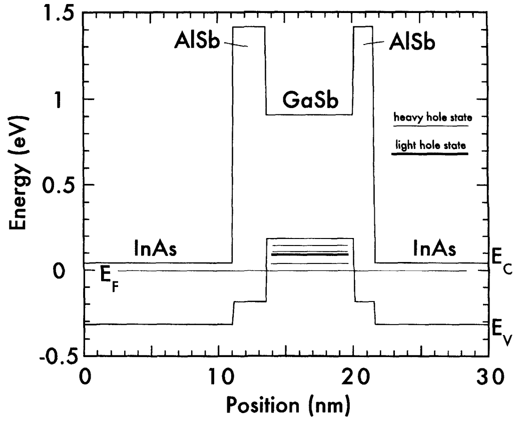

FIG. 8. Computed energy-band diagram for the published resonant interband tunnel diode of Söderström *et al.* (1989), exhibiting a peak-to-valley current density of 20 at room temperature. (Computed using the heterojunction modeling program *Bandprof* of W. R. Frensley.)

1.3 Valley Current

The valley current and turnon after the valley current for both RTDs and Esaki diodes is determined by the scattering from impurities, disorder, phonons, and other electrons, and thermionic emission. The scattering-assisted current is referred to as the *excess* current (Chynoweth *et al.*, 1961; Sze, 1981; Brown, 1994). For GaAs/Al$_x$Ga$_{1-x}$As RTDs, the room-temperature valley current is dominated by thermionic emission through the first excited state in the well or X states in the barriers (Brown, 1994). For the best In$_{0.53}$ Ga$_{0.47}$As/AlAs/InAs RTDs, the room-temperature valley current appears to be determined by scattering as opposed to thermionic emission (Smet *et al.*, 1992). Theoretical calculations comparing the effects in RTDs of acoustic phonon, polar-optical phonon, alloy, and interface roughness scattering indicate that the dominant scattering mechanisms of the four are polar optical phonon and interface roughness (Chevoir and Vinter, 1993; Roblin and Liou, 1993). The effect of electron–electron scattering and scattering from remote ionized donors in the contacts has not yet been theoretically investigated. Experimental evidence indicates that dopant segregation from the highly doped contacts to the undoped device region may strongly affect the PVR of RTDs (Cunningham *et al.*, 1993).

The notable difference between an RTD and an Esaki diode is that both elastic and inelastic scattering can contribute to the valley current of an RTD whereas only inelastic scattering can contribute to the valley current of an Esaki diode. For an RTD, Fig. 3 shows that there is always a high–transverse-energy state in the well into which an incident electron can scatter while conserving total energy. For an Esaki diode, Fig. 4(a) shows that for a bias greater than μ, there are no valence states into which an electron can elastically scatter. The excess current in Esaki diodes is attributed to localized states in the band gap combined with phonon scattering. If the thermionic emission is negligible, an Esaki-type device should, in principle, be capable of achieving a higher peak-to-valley current ratio than an RTD. In fact, the highest room-temperature peak-to-valley current ratio (PVR) of 144 has been observed in a heterojunction Esaki diode (Tsai *et al.*, 1994). This is about a factor of 3 above the best RTD (Smet *et al.*, 1992).

1.4 Speed

The speed of a tunnel diode can be characterized by the time it takes to switch from a stable bias point near the peak voltage (V_P) to a final voltage-stable bias point (V_F) beyond the peak; see Fig. 9. This will depend on the capacitance of the tunnel diode and on the change in charge between the initial and final states. If the capacitance is assumed to be approximately constant, then the change in charge can be written $\Delta Q = Q_F - Q_P = C(V_F - V_P)$, where Q_F and Q_P are the final and initial charge states and C is the capacitance. The time required to change the charge on the capacitor will depend on the current available for charging, $I_S(V) - I_{TD}(V)$, where $I_S(V)$ is the current that can be provided by the voltage source and $I_{TD}(V)$ is the I–V characteristic for the tunnel diode:

$$\tau = \int_{V_P}^{V_F} \frac{C(V_F - V_P)}{I_S(V) - I_{TD}(V)}\, dV. \tag{9}$$

The relation $I_S(V) - I_{TD}(V)$ is indicated by

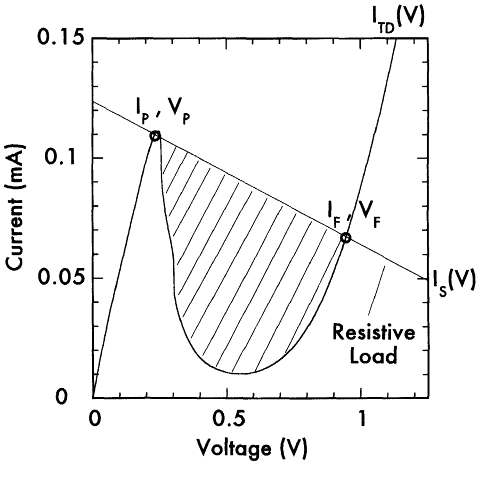

FIG. 9. Measured current–voltage characteristic for an AlAs/InGaAs double-barrier resonant tunneling diode (emitter area of 1 μm^2, wafer R7556). Superimposed on the experimental curve is the resistive load line $I_S(V)$ resulting in two stable bias points; the lined region denotes the current available for charging the device capacitance.

the lined region in Fig. 9. If we assume, after Diamond *et al.* (1989), that the capacitance is roughly constant and that the available current can be bounded by $I_s(V) \leq I_P$ and $I_{TD} \geq I_V$ (where I_P and I_V are the diode peak and valley currents, respectively), then a lower bound on Eq. (9) can be written:

$$\tau = \frac{C(V_F - V_P)}{I_P - I_V} = \left[\frac{C}{I_P}\right]\frac{V_F - V_P}{1 - I_V/I_P}. \tag{10}$$

From this relation, the switching speed is determined by the voltage swing, by the peak-to-valley current ratio I_P/I_V, and by an intrinsic factor I_P/C, called the speed index. While peak-to-valley current ratio is important, increasing PVR beyond 3 does not speed up the device significantly. Since for switching applications the speed index is the primary factor governing speed, we provide a survey of this factor across all semiconductor tunnel diodes in Sec. 3. The speed index of the resonant tunneling diodes is generally greater than that of the Esaki tunnel diode, since high current density and low capacitance are difficult to achieve simultaneously in the Esaki diode.

The capacitance of the tunnel diode is given by $C = \epsilon A/d$, where d is the thickness of the depletion layer. For uniform doping density this depletion-layer thickness varies with the square root of applied bias for both intraband and interband diodes. For the Esaki diode, the capacitance C is

$$\frac{C}{A} = \sqrt{\frac{q\epsilon N_A N_D}{2(V_{\text{bi}} - V)(N_A + N_D)}}, \tag{11}$$

where A is the junction area, q is the fundamental charge, N_A and N_D are the acceptor and donor atom densities, respectively, V_{bi} is the junction built-in voltage, and V is the applied voltage.

For the RTD, a unipolar device, the capacitance is inversely proportional to the combined length L_{RTD} of the accumulation region in the emitter, the depletion region in the collector, and the barrier–well–barrier region in the center (Brown, 1994). Like the Esaki diode, the capacitance depends inversely on the square root of the applied bias, but with the opposite bias dependence. In the Esaki diode the capacitance is generally increasing as the bias increases toward the peak voltage, while in the RTD the capacitance is decreasing with applied bias. Since the Esaki diode is very heavily doped, the length of the depletion region is much less than the length L_{RTD} of the RTD, and therefore Esaki diodes typically have larger capacitances than RTDs. In terms of both capacitance and current the RTD appears to be superior for high speed.

One should note, however, that this simple picture of the RTD capacitance decreasing monotonically with bias describes only the first-order behavior. Near resonance the capacitance peaks significantly in RTDs (T. Wei, *et al.*, 1993; Jo *et al.*, 1994). This contributes to the significantly lower measured versus theoretical speed indices shown later in Table 2.

2. FABRICATION

In 1969, Esaki tunnel diodes could be obtained commercially from as many as twelve suppliers (Baasch, 1969), but there are only a few companies that still produce them (Si and Ge). Resonant tunneling diodes are not commercially produced but can be obtained from many university and industrial laboratories. Much of the Esaki-diode fabrication experience from the 1960s and 1970s is lost, but there remains a number of excellent technology papers and patents that reveal the beautiful work of the early Esaki-diode development. Notably, for Ge-alloy tunnel diodes, see the article by Davis and Gibbons (1967) and the U.S. patent of Tiemann (1966).

2.1 Tunnel-Junction Formation

The Esaki diode requires the formation of an abrupt, heavily doped p^+n^+ junction with doping densities exceeding 10^{19} cm^{-3} on both sides of the junction. Commercial Esaki diodes are typically formed by an alloy process in which a metal containing the desired dopant is alloyed into a high-conductivity semiconductor substrate as illustrated in Fig. 10. Alloying has advantages for forming abrupt doping junctions, since generally, both the temperature and alloy time can be kept lower and shorter, respectively, than the corresponding epitaxial growth process. In addition, the process of alloying also serves

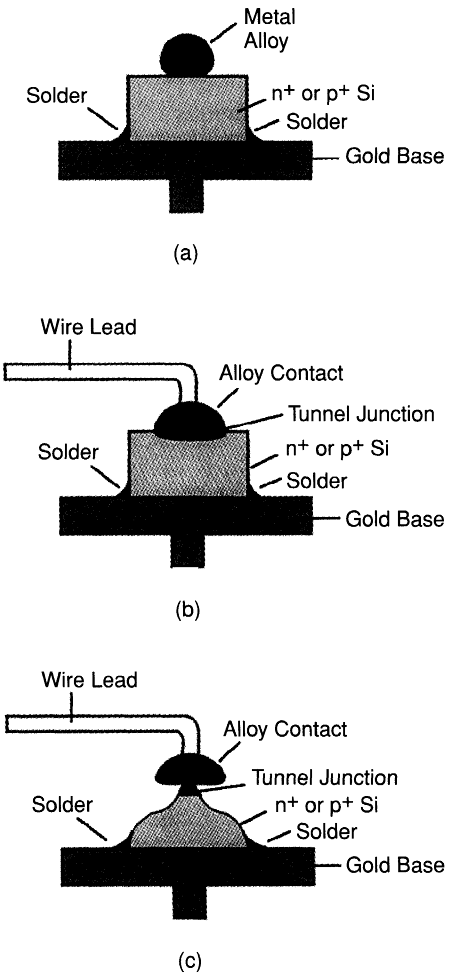

FIG. 10. Fabrication steps for an alloyed Esaki diode: (a) the high-conductivity semiconductor base material is soldered to a device header and the metal alloy containing the appropriate dopant is placed on the semiconductor, (b) the tunnel junction is formed by alloying and a wire attached, and (c) the tunnel junction is etched to obtain the desired current density.

to form the diode contact, making for a simple process, with low contact resistance. In addition to being simple, the alloy process is inexpensive relative to the cost of epitaxial-growth equipment.

Resonant tunneling diodes and other heterojunction tunnel diodes such as the heterojunction Esaki diode and the resonant interband tunnel diode are now primarily grown by molecular-beam epitaxy (MBE) (*q.v.*) (Matyi, 1994). Molecular-beam epitaxy is an ultrahigh-vacuum growth process in which the elemental semiconductors and dopants of the desired structure are evaporated at carefully calibrated and controlled rates onto a heated semiconductor substrate. The growth rate is typically about 1 monolayer of semiconductor per second (2–3 Å/s), which allows semiconductors to be engineered with layer-to-layer control of the crystal composition and doping. By MBE, semiconductor heterojunctions can be grown, one on top of another, to form the abrupt potential discontinuities needed for resonant tunneling and heterostructure tunnel diodes.

2.2 Device Processing and Integration

While considerable effort has been invested in the development of batch semiconductor processes for Ge and Si Esaki diodes, commercial tunnel diodes today are produced individually. Two commercial Esaki-diode configurations for high-speed applications are shown in Fig. 11, but the process of Fig. 10 is also still used. These processes begin by mounting the high-conductivity substrate material on a device header. A metal consisting of or containing the desired dopant is then arranged on the semiconductor. Example metals are (Watson, 1969), for a p^+ layer on n^+ Ge, In/Ga (95:5); for Si, Al/B; and for GaAs for an n^+ layer on a p^+ substrate, Sn works well. Alloy temperatures and times depend on the metal but are typically less than 500 °C and less than about 1 min.

In the alloying step the tunnel junction and contact metal are simultaneously formed. The peak current density, peak voltage, and peak-to-valley current ratio are all determined by the abruptness of the alloy junction and the absolute impurity densities. The absolute value of the current, however, can be controlled (reduced) after the junction is formed by reducing the area of the junction; this is typically achieved by an electrochemical etch.

In contrast to the Esaki diode, resonant tunneling diodes and resonant interband tunnel diodes are typically formed by a batch process such as is shown schematically in Fig. 12. This process consists of the steps of epitaxial growth, contact metallization, reactive ion etching or wet chemical etching to define the device area, mesa isolation, and

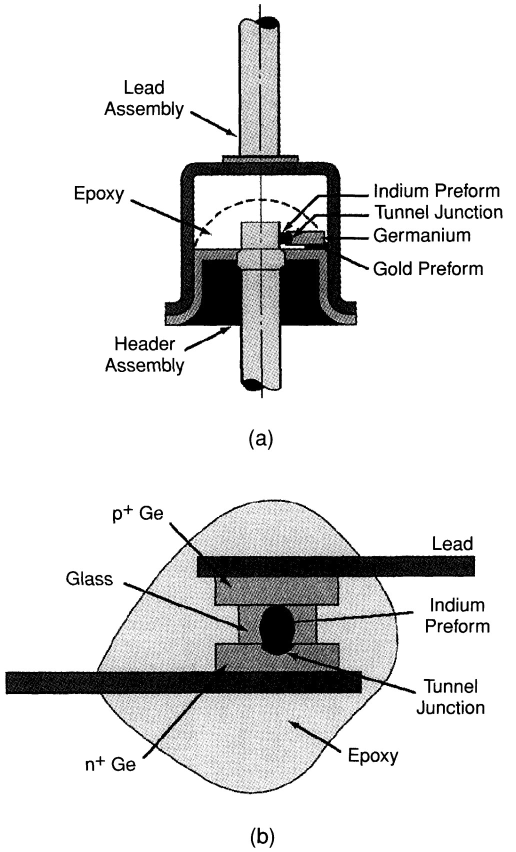
Lead Assembly
Epoxy
Indium Preform
Tunnel Junction
Germanium
Gold Preform
Header Assembly
(a)
p+ Ge
Glass
Lead
Indium Preform
Tunnel Junction
n+ Ge
Epoxy
(b)

FIG. 11. Two commercial Esaki-diode configurations: (a) low-speed package, (b) planar high-speed package. (Courtesy of W. Dawson, Germanium Power Devices.)

second contact metallization. In such a batch process, transistors can be formed side by side or connected vertically (Seabaugh *et al.*, 1993), and additional components such as diodes, resistors, and capacitors can be integrated.

2.3 Manufacturing Issues

The Esaki tunnel diode is a well-established commercial product with more than 20 years of market experience. For the resonant tunneling diode and heterojunction tunnel diodes, currently being explored for cofabrication with mainstream integrated circuits, questions about uniformity, reproducibility, and reliability are still being addressed.

2.3.1 Unformity The uniformity of RTD layer growth is primarily determined by the configuration of the MBE growth chamber, and most critically by the arrangement of the sources with respect to the substrate. For resonant tunneling diodes consisting of AlAs/InGaAs/AlAs double barriers, Moise *et al.* (1995a,b) have shown three-sigma variations in peak current, peak voltage, and peak-to-valley current ratio [from edge to edge of a 2-in. (5-cm) wafer] of 7.2%, 5.2%, and 4.9%, respectively. This represents better than one monolayer thickness control, as approximately 50% variation in the current is expected in this RTD for a single monolayer variation in the tunnel-barrier thickness. An example of the good local uniformity of RTDs obtained by MBE is shown in Fig. 13.

2.3.2 Reproducibility Because the current in tunnel diodes is dependent exponentially on tunnel barrier thickness, run-to-run reproducibility of the junction formation is more difficult to achieve than device-to-device uniformity across a wafer. To date, run-to-run reproducibility of peak current (the

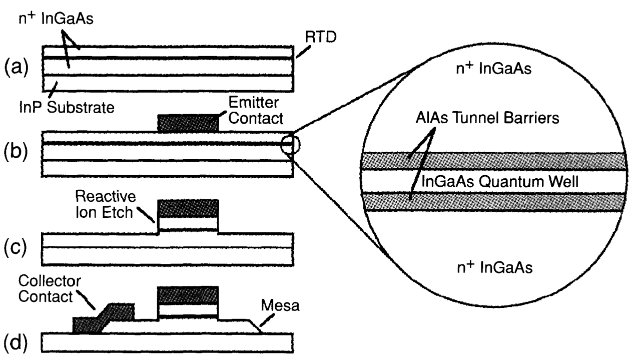
n+ InGaAs
RTD
(a)
InP Substrate
Emitter Contact
(b)
Reactive Ion Etch
(c)
Collector Contact
Mesa
(d)
n+ InGaAs
AlAs Tunnel Barriers
InGaAs Quantum Well
n+ InGaAs

FIG. 12. Batch process for the formation of AlAs/InGaAs resonant tunneling diodes: (a) starting epitaxial structure with vertical dimension drawn to scale (total epitaxial layer thickness 750 nm), (b) emitter metallization, (c) reactive ion etch to define the device active area, (d) mesa etching to isolate the device from adjacent devices followed by collector metal deposition. The inset gives an expanded view of the double-barrier RTD with typical dimensions of 2/4/2 nm for an AlAs/InGaAs/AlAs structure.

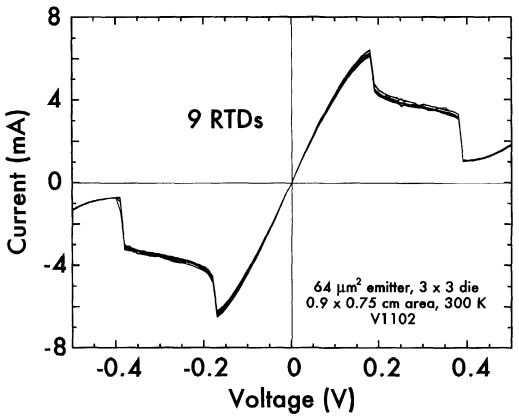

FIG. 13. Uniformity measurements of AlAs/InGaAs double-barrier resonant tunneling diodes after Moise *et al.* (1995).

most sensitive parameter) in resonant tunneling diodes is held within approximately ±12%. Further improvements in calibration procedures and control technology are expected to reduce this variation.

2.3.3 Reliability Esaki-diode reliability is well established for commercial and space use. Mean time before failure (MTBF) exceeding 100 000 hours (more than 11 years) has been reported by McDermott (1981) for tunnel-diode amplifiers in satellite communication systems. Again, reliability of RTDs is not fully qualified. Miura *et al.* (1990) have shown good endurance of $AlAs/In_{0.53}Ga_{0.47}As$ RTD oscillators; continuous operation has been reported for times exceeding 1000 hours at 90°C for oscillation frequencies between 0.6 and 1 GHz. dc bias stress tests at 80 °C have been performed on AlAs/InGaAs RTDs on InP (Seabaugh *et al.*, 1997); more than 10 400 hours (>1 year) have been logged without increase of the peak or valley current.

3. DEVICE PROPERTIES

In this section, we provide a survey and comparison of the device properties of Esaki, resonant interband, and resonant tunneling diodes across the explored range of semiconductor materials. Three tables are provided: Table 1 compares the p^+n^+ Esaki diodes; Table 2, the resonant interband tunnel diodes; and Table 3, the resonant tunneling diodes. The tables predominantly cover materials that exhibit room-temperature negative differential resistance or that give the best peak-to-valley current ratio for that particular substrate. The tables also list the key materials parameters for each structure. The p^+n^+ Esaki-diode parameters are fairly simple to envision; the RITD heterojunction alignments are not always obvious. Figures 6 and 8 show the energy-band diagrams for two of the entries in Table 2. These figures give the computed energy-band diagrams for the highest-PVR (144) tunnel diode and for the highest-PVR interband *n-p-n* tunnel diode, 20. For Table 3, an explanatory Fig. 14 is provided to define the listed parameters.

From the tables one can see that the highest speeds are obtained for the resonant tunneling diodes, and the highest peak-to-valley current ratios are obtained for the resonant interband tunnel diodes. While not shown in the table, speed index in alloy Ge diodes as high as 50 has been reported by Davis and Gibbons (1967).

3.1 Esaki Diode

Esaki diodes have been formed on a wide variety of common semiconductor substrates. In this section, we review the properties of these devices according to the columns in the periodic table from which the semiconductors are formed.

3.1.1 Group IV Silicon and germanium Esaki diodes are the only tunnel diodes commercially available today. Of these, the highest speed and peak-to-valley current ratio are obtained in the Ge diodes with switching speeds as fast as 22 ps obtained in packaged configurations such as the planar configuration shown in Fig. 11(b). The best speed index is also obtained from alloy Ge diodes with values at least as high as 50 reported by Davis and Gibbons (1967).

Esaki diodes formed of silicon carbide (SiC) on 6-H SiC substrates have shown the highest operating temperature of any tunnel diode (Dmitriev *et al.*, 1985). At room temperature the PVR is only about 1.4; however, the negative–differential-resistance region is well maintained to temperatures at least as high as 450 °C.

Table 1. Comparison of Esaki tunnel diodes operated at room temperature. Abbreviations: N_A, acceptor density; N_D, donor density; J_P, peak current density; V_P, peak voltage; PVR, peak-to-valley current ratio; *C*, junction capacitance per unit area, LPE, liquid-phase epitaxy; MBE, molecular-beam epitaxy; MOCVD, metalorganic chemical vapor deposition; MOMBE, metalorganic molecular-beam epitaxy; and RITD, resonant interband tunnel diode.

Materials	Growth method	Substrate	(10^{18} cm^{-3})		J_P (A/cm^2)	V_P (V)	PVR	C[a] (fF/μm^2)	Speed index[a] (mA/pF)	References
			N_A	N_D						
Ge	LPE	Ge	238	45	16 000	0.125	9.3[c]	15.4[b]	10.4[b]	Meyerhofer *et al.* (1962)
Ge (TD276A)	Alloy	Ge	—	—	—	0.18	8.3	—	27.5[b]	Germanium Power Devices Corp (1996)
Si	Alloy	Si	23	3	2.4	0.06	2.3	4.6	0.005	Logan (1963)
Si	Alloy	Si	200	—	1 000	—	3.8	8.3[b]	1.2[b]	Franks *et al.* (1965)
Si (1N2930)	Alloy	Si	—	—	—	0.08	3.2	—	—	Microsemi Corp. (1996)
InAs	Alloy	InAs	—	4.8	1 190	0.13	1.9	—	—	Kleinknecht (1961)
InSb	Alloy	InSb	—	0.7	—	0.06	1	—	—	Batdorf *et al.* (1960)
GaSb	Electroforming	GaSb	—	—	50 000	0.3	15	—	—	Burrus (1961)
GaAs	Alloy	PolyGaAs	—	—	—	0.1	25	—	—	Holonyak and Lesk (1960)
GaAs	Alloy	GaAs	10	50	—	0.15	12	—	—	Richard *et al.* (1993)
p^+GaAs p^+In$_{0.1}$Ga$_{0.9}$As (10 nm) n^+GaAs	MOCVD	GaAs	10	100	1 300	0.08	21	7.8	1.7	Richard *et al.* (1993)
Al$_{0.11}$Ga$_{0.89}$As	LPE	GaAs	8	8	1.3	0.12	1.2	4.9	0.003	Bedair (1979)
GaAs$_{0.9}$Sb$_{0.1}$	LPE	GaAs	10	10	7.5	0.38	1.4	5.4	0.014	Timmons and Bedair (1980)
InP	Electroforming	InP	5	—	—	0.2	5	—	—	Burrus (1962)
In$_{0.53}$Ga$_{0.47}$As	LPE	InP	20	20	793	0.31	16	10.8	0.73	Shen *et al.* (1993)
In$_{0.53}$Ga$_{0.47}$As	MOMBE	InP	30	30	9 200	0.12	16	11.8	7.8	Cohen *et al.* (1995)
In$_{0.53}$Ga$_{0.47}$As	MOMBE	InP	100	100	93 200	0.66	3.8	30.0	31.0	Cohen *et al.* (1995)

---

[a] Unless indicated, the capacitance and speed-index values are computed from the electrostatic band diagram at a bias of V_P (using the heterostructure device simulator *BandProf* of W. R. Frensley).

[b] Measured (for this device the computed capacitance and speed index are 19 and 8.4, respectively).

[c] 4.2 K.

Table 2. Comparison of resonant interband tunnel diodes operated at room temperature. Abbreviations: N_A, acceptor density; N_D, donor density; J_P, peak current density; V_P, peak voltage; PVR, peak-to-valley ratio; C, junction capacitance per unit area; LPE, liquid-phase epitaxy; MBE, molecular-beam epitaxy; MOCVD, metalorganic chemical vapor deposition; MOMBE, metalorganic molecular-beam epitaxy; and RITD, resonant interband tunnel diode.

Materials	Growth method	Substrate	(10^{18} cm^{-3})		J_P A/cm^2	V_P (V)	PVR	C^a fF/μm^2	Speed index^a mA/pF	References
			N_A	N_D						
n-InAs AlSb/p-GaSb/AlSb (1.1/5.4/1.1 nm) *n*-InAs	MBE	GaAs	—	—	32 000	0.12	6.0	2.7	118.5	Chow *et al.* (1996)
n-InAs AlSb/p-GaSb/AlSb (2.5/6.5/1.5 nm) *n*-InAs	MBE	GaAs	0.02	0.02	35	0.04	20	2.1	0.17	Söderström *et al.* (1989); see Fig. 8
p^+ GaSb AlSb/GaSb/AlSb (1.5/8/1.5 nm) InAs/AlSb (1.5/10 nm) n^+InAs	MBE	GaAs	5	1	7 200	0.34	10	11.1	6.5	Yang *et al.* (1990)
p^+ $In_{0.52}Al_{0.48}As$ $In_{0.53}Ga_{0.47}As/In_{0.52}Al_{0.48}As/$ $In_{0.53}Ga_{0.47}As$ (4/2/4 nm) $n^+In_{0.52}Al_{0.48}As$	MBE	InP	30	30	163	0.84	144	7.0	0.23	Tsai *et al.* (1994); see Fig. 6

---

[a] Unless indicated, the capacitance and speed index values are computed from the electrostatic band diagram at a bias of V_P (using the heterostructure device simulator *BandProf* of W. R. Frensley).

Table 3. Comparison of electron resonant tunneling diodes operated at room temperature. Abbreviations: N_E and N_C, doping density in the emitter and collector regions adjacent to the double barrier, respectively; J_P, peak current density; V_P, peak voltage; PVR, peak-to-valley current ratio; and C, junction capacitance per unit area. See Fig. 14 for definitions of TB, QW, emitter, collector, and spacer.

Tunnel barrier (TB)	Quantum well (QW)	Substrate	Thicknesses TB/QW/TB (nm)	Spacer (nm)	(10^{18} cm^{-3}) N_E	N_C	J_P (A/cm^2)	V_P (V)	PVR	Capacitance[a] (fF/μm^2)	Speed index[a] (mA/pF)	Reference
Si	$Si_{0.7}Ge_{0.3}$	Si	7.5/5/7.5	15	1	1	400	0.5	1.2	—	—	Ismail *et al.* (1991)
AlAs	GaAs	Si	2/5.6/2	100	1	1	1 630	1.9	2.9	0.87	19	Kan *et al.* (1988)
AlAs	$In_{0.33}Ga_{0.67}As$	GaAs	1.7/5.1/1.7	4.2	2	2	125 000	0.6	4.7	4.2	298	Kapre *et al.* (1991)
$Al_{0.6}Ga_{0.4}As$	$GaAs/In_{0.25}Ga_{0.75}As/GaAs$	GaAs	3/1.5/1.5/1.5/3	2.5	0.2	0.2	40 000	0.7	4.8	1.6	250	Lippens *et al.* (1991)
AlAs[b]/$Al_{0.2}Ga_{0.8}As$	GaAs	GaAs	1.7/5/1.7/1.1	5	0.1	0.1	31 000	0.47	6	1.4	221	Reddy *et al.* (1990)
AlAs	$In_{0.1}Ga_{0.9}As$	GaAs	1.7/4/1.7	10[c]	0.2	0.2	50 000	0.7	7	1.5	333	Lheurette *et al.* (1992)
$Al_{0.7}Ga_{0.3}As$	GaAs	GaAs	3.4/4/3.4	4[d]	1	1	10 000	2.0	7.2	2.0	500	Brugger *et al.* (1991)
AlSb	InAs	GaAs	1.5/6.4/1.5	10	2	2	370 000	1.21	3.2	3.1	1190	Söderström *et al.* (1991a)
AlSb	InAs	GaAs	1.8/7.5/1.8	10[e]	2	0.1	330 000	1.23	3.3	1.8	1830[f]	Özbay *et al.* (1993)
AlSb	InAs	GaAs	3.0/7.5/3.0	10/25	2	2	490 000	1.25	2.2	2.3	2130	Chow *et al.* (1993)
$Al_{0.5}In_{0.5}Sb$	InSb	GaAs	2.2/11/2.2	5	0.02	0.02	36 000	0.3	1.4	2.6	138	Söderström *et al.* (1991b)
AlAs	$In_{0.53}Ga_{0.47}As$	InP	1.7/5/1.7	10	0.1	1	175 000	1.3	5.5	2.4	729	Mounaix *et al.* (1995)
AlAs	$In_{0.53}Ga_{0.47}As$	InP	2.3/4.1/2.3	1.5	1	0.1	76 000	2.65	25	0.61	1250	Chow *et al.* (1992)
AlAs	$In_{0.53}Ga_{0.47}As$	InP	2.0/4.1/2.0	1.5	1	0.1	310 000	1.69	6	0.78	3970	Chow *et al.* (1992)
AlAs	$In_{0.53}Ga_{0.47}As$	InP	1.2/4.7/1.2	1.5	0.2	0.2	460 000	0.67	4	1.7	2710	Broekaert and Fonstad (1990)
AlAs	$In_{0.53}Ga_{0.47}As$	InP	1.4/4.1/1.4	1.4/3[g]	1	2	450 000	1.75	3.9	3.1	1450[h]	Shimizu *et al.* (1995)
AlAs	$In_{0.53}Ga_{0.47}As/InAs/In_{0.53}Ga_{0.47}As$	InP	3.2/0.9/1.8/0.9/2.6	1.8	0.2	0.2	5 800	1.35	51.6	1.2	48	Smet *et al.* (1992)
$In_{0.52}Al_{0.48}As$	$In_{0.53}Ga_{0.47}As$	InP	4.7/3.2/4.7	1.5	1	1	48 000	0.68	5.5	3.2	150	Sugiyama *et al.* (1988)
$AlAs_{0.56}Sb_{0.44}$	$In_{0.53}Ga_{0.47}As$	InP	2.9/4.4/2.9	0	0.1	0.1	2 370	2	15	0.70	34	Muto and Inata (1994)
$Hg_{0.2}Cd_{0.8}Te$	HgTe	CdTe	3.8/6.6/3.8	0	0.5	0.5	1 500	0.5	1.4	—	—	Reed *et al.* (1986)

---

[a] Unless indicated, the listed capacitance values are computed from the electrostatic band diagram (at a bias of V_P), using the heterostructure device simulator *BandProf* of W. R. Frensley.

[b] Asymmetric barriers, emitter-side barrier AlAs, collector-side barrier $AlAs/Al_{0.2}Ga_{0.8}As$.

[c] 5-nm $In_{0.1}Ga_{0.9}As$ nearest the AlAs barrier, followed by 5-nm GaAs.

[d] $In_{0.15}Ga_{0.85}As$ prewell on the emitter side.

[e] Collector-side undoped spacer only.

[f] Measured extrinsic speed index, 500 mV in 1.7 ps, is 294 mA/pF.

[g] Refers to emitter/collector spacer layers respectively.

[h] Measured extrinsic speed index, 630 mV in 2.2 ps, is 286 mA/pF.

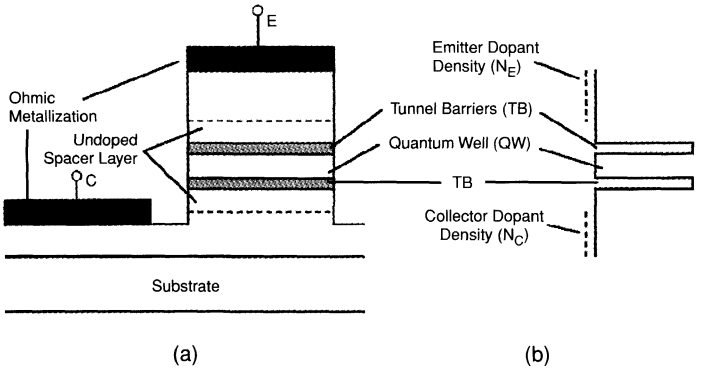

FIG. 14. Definition of terms for the resonant tunneling diode comparison Table 3: (a) schematic device cross section, (b) energy-band diagram.

3.1.2 Group III–V Peak-to-valley current ratios as high as 30 have been reported for p^+n^+ GaAs alloy tunnel diodes (Holonyak, 1960). High PVR (21) has also been obtained in the GaAs/$In_{0.1}Ga_{0.9}As$/GaAs Esaki diode of Richard *et al.* (1993), where the InGaAs layer is inserted at the *p-n* junction. For a given doping density, this technique of introducing a narrow–band-gap material at the tunneling junction is an effective way of increasing the peak tunneling current density. The wider–band-gap semiconductor, GaAs, on either side of the InGaAs provides a large band-gap blocking in the off-resonance condition, needed to keep the valley current low.

3.2 Resonant-Tunneling Diode

The resonant-tunneling diode has a wider design flexibility than the Esaki or resonant interband tunnel diode. Because the resonance condition does not depend on the overlap of valence- and conduction-band states, the resonant peak and valley voltages can be designed over a wide voltage range, from a few tens of millivolts to several volts. The highest peak voltage reported for the Esaki diode is 66 mV (for the InGaAs diode of Cohen *et al.,* 1995), whereas for RTDs the peak voltage can readily exceed 1 V.

The RTD can also be engineered for multiple current peaks by the introduction of multiple quantum wells. Shown in Fig. 15 is the measured *I–V* characteristic (Mikkelson *et al.,* 1994), of a triple-barrier, double-well RTD that has been designed for two resonance peaks. This design flexibility is not available in Esaki diodes and in RITDs; engineering of multiple current peaks with good PVR is difficult on account of the required close spacings of the resonant states.

3.2.1 Group IV There is little prior art on silicon-based electron resonant tunneling diodes. Ismail *et al.*'s (1991) demonstration of resonant tunneling in the $Si_{0.7}Ge_{0.3}$/Si/$Si_{0.7}Ge_{0.3}$ RTD (grown on a relaxed $Si_{0.7}Ge_{0.3}$ buffer layer) stands as the best demonstration of group-IV electron resonant tunneling on silicon. The room-temperature peak current density in these diodes was about 400 A/cm^2 with a PVR of 1.2. High peak-to-valley current ratios are not expected in this system

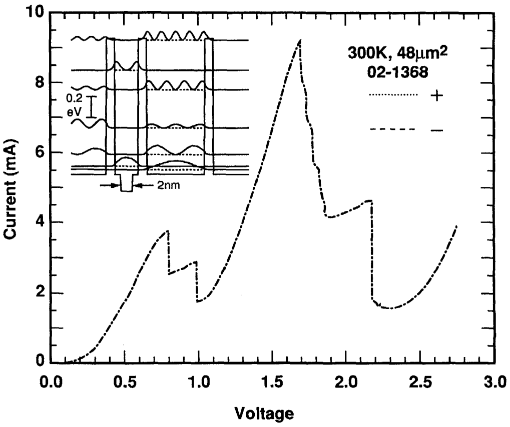

FIG. 15. *Room-temperature measured current–voltage* characteristics at room temperature for a triple-barrier, double-well resonant tunneling diode. Inset: computed energy-band diagram, transmission resonances, and wave functions. The dotted and dashed lines correspond to positive and negative voltage ramps, respectively; the two curves are essentially indistinguishable, showing a hysteresis-free behavior.

because of the low conduction-band offsets (less than 500 meV).

3.2.2 Group III–V The highest room-temperature peak-to-valley current ratios and highest current densities in the RTD have been obtained on the AlAs/InGaAs RTD on InP. Peak-to-valley current ratio of 52 by Smet *et al.* (1992) and peak current density of 460 kA/cm^2 by Broekaert and Fonstad (1990), both obtained on n^+ InP substrates, are the best results to date. As a result of this high current density, the fastest RTD switching speeds have also been measured in the III–V RTDs. For the AlAs/InGaAs RTD with 450 kA/cm^2, Shimizu *et al.* (1995) report a switching time of 1.5 ps. Comparable speed, 1.7 ps, has been reported by Özbay *et al.* (1993) for the InAs/AlSb RTD with peak current density of 330 kA/cm^2.

3.2.3 Group II–VI For the II–VIs, the only exploration of double-barrier resonant tunneling is for the HgCdTe/HgTe system by Reed *et al.* (1986). In this case, HgCdTe forms the barrier material, while HgTe is a zero–band-gap semimetal that makes up the quantum well. Characterization of this structure showed PVR as high as 1.4 at room temperature.

3.2.4 Mixed Crystalline Several heteroepitaxial systems utilizing crystalline metal and semimetal/insulator structures, and systems in which barrier and well materials come from different groups of the periodic table, have been demonstrated. These have met with limited success, most yielding tunneling diode characteristics with negative differential resistance only at low temperature. For example, Han *et al.* (1991) successfully grew the MnTe/InSb/MnTe RTD by molecular-beam epitaxy with peak-to-valley current ratio of about 1.7 at 77 K; the poor performance is attributed to scattering at the interfaces and edge leakage on the device sidewalls.

Much better results have been obtained in the $CaF_2/CoSi_2$ triple-barrier double-well resonant tunneling diode of Suemasu *et al.* (1992). In these crystalline structures grown by molecular-beam epitaxy, a peak-to-valley current ratio as high as 2 at room temperature has been obtained. Because the emitter is a metal, there is no band gap to block transmission through the quantum-well states after the states are moved below the Fermi level in the emitter. Because of the second quantum well, the peak current corresponds to the resonant *alignment* of quantum well states in adjacent wells; when the states depart from alignment at post-peak biases, the tunneling current is reduced.

Another of the mixed-crystalline RTDs to be explored is the semimetal/semiconductor AlAs/ErAs/AlAs system, which is grown on GaAs and uses n-GaAs as the contacting layers. Remarkably, the ErAs has a rock-salt cubic crystal structure, but it can be grown and fitted to the zinc-blende GaAs and AlAs crystals. Room-temperature negative differential resistance has been reported (Brehmer *et al.*, 1995). Improvements in these systems can be expected with improvements in crystal-growth processes.

3.2.5 Amorphous Resonant Tunneling Shortly after the demonstration of the first RTD by Chang *et al.* (1974), Hirose *et al.* (1977) reported resonant tunneling in the amorphous $SiO_2/Si/SiO_2$ structure grown by thermal oxidation in O_2 and with electron-beam deposition of amorphous silicon. Weak negative differential resistance was reported at room temperature. Since then, Miyazaki *et al.* (1987) have reported resonant tunneling in amorphous $Si_3N_4/Si/Si_3N_4$ double barriers, but without NDR. Fang *et al.* (1994) have reported a PVR of 9.4 at room temperature in an amorphous SiC/Si/SiC structure deposited on glass. The origin of the NDR in the amorphous SiC/Si structures is not clear since the materials contain many traps. It is conceivable that the NDR is produced by resonances through defect states.

3.2.6 Hole Resonant Tunneling Resonant tunneling diodes can be fabricated in the valence band as well as the conduction band. Because of technological interest in Si/Si_xGe_{1-x} heterostructures and because the band-gap difference between the Si and Si_xGe_{1-x} occurs predominantly in the valence band, this system has been the most widely studied. In this case the carrier transport is by holes and the particle tunneling due to hole resonant tunneling. The Si_xGe_{1-x} well can be strained, grown on relaxed Si (Liu *et al.*, 1988), or the Si barriers can be strained, grown on relaxed Si_xGe_{1-x} (Gennser *et al.*, 1990). Several works have experimentally and theoretically investigated the effect of confinement on the light-hole and heavy-hole states and the resonant tunneling peaks associated with each state (Shen *et al.*, 1992; Fu *et al.*, 1991). However, negative differential resistance has not been observed at room temperature. Theoretical calculations predict accurately the peak currents but underestimate the valley currents by several orders of magnitude (Shen *et al.*, 1992). There is no existing explanation for the large valley currents. At a temperature of 77 K, all reported peak-to-valley ratios are less than 2.0.

3.3 Resonant Interband Tunnel Diode

The best PVR (144) at room temperature of any tunnel diode is obtained in the structure of Tsai *et al.* (1994) shown in Fig. 6. Heterostructure *npn* interband tunnel structures such as that shown in Fig. 8 by Söderström *et al.* (1989) have also produced high room-temperature PVR (20). The heterostructure interband devices based on the GaSb/AlSb/InAs materials system offer a wide design flexibility (Collins *et al.*, 1991) with the low effective mass of InAs, the availability of both binary and ternary alloys, and the wide range of suitable band offsets. Most work on these materials has been done on GaAs substrates, because of the lower substrate cost and technological desirability of the GaAs substrate. Valley current in the interband tunnel devices such as the GaSb/AlSb/InAs heterostructures can be dominated by the parasitic thermionic emission of holes.

4. APPLICATIONS

Despite being the fastest semiconductor switching device, at this writing the tunnel diode is not widely used. Discrete Ge and Si Esaki diodes can be purchased from two U.S. companies, and the current size of the markets for these devices is only a few thousand components per year. The diodes are used in picosecond pulse generators, in microwave detectors in the 0.5–18-GHz band, and in miscellaneous special applications. Esaki tunnel junctions are also used in cascade solar cells to provide an Ohmic interconnection between epitaxially integrated III–V solar cells, but in this case the negative differential resistance is not important.

In the 1960s and 1970s tunnel-diode applications were believed to have broad market potential including high-speed and low-power signal processing, logic circuits, microwave detectors, amplifiers, oscillators, and pressure sensors. However, transistor technology eventually surpassed the speed of the alloy Esaki-diode technology, and the incompatibility of the fabrication technology with conventional integrated-circuit processing lines never allowed a full exploration of tunnel-diode circuits.

In the last decade, heterostructure technology has been applied to the tunnel diode, and a broader array of tunnel diodes has been developed. Unlike the early Ge Esaki diode, these devices are batch fabricated and integrable with conventional transistors and IC components.

4.1 Triggering, Pulse Forming, and Sampling

For triggering and pulse forming, the tunnel diode is used for high-speed switching, competing against the Schottky-barrier diode (Andrews, 1986). The basic triggering circuit topology* is shown in Fig. 16. When the circuit is switched to "on," the tunnel diode is biased to a point near its peak voltage by the constant-current source. With this bias an

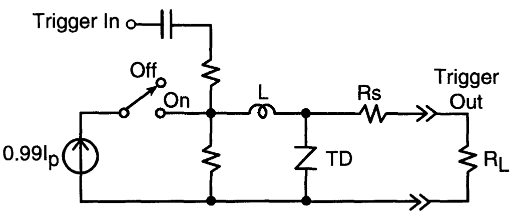

FIG. 16. Tunnel-diode trigger circuit consisting of a current source providing a bias of 99% of the peak current (I_p) to the tunnel diode (TD), a "Trigger In" signal that provides the additional current used to cause the TD to switch from low-voltage to high-voltage states, thereby generating the "Trigger Out" signal into the external load, R_L. The on/off switch is used to reset the TD from its high-voltage state to its low-voltage state. (Courtesy of J. Andrews, Picosecond Pulse Labs.)

---

*While there exists an IEEE standard symbol for the Esaki diode (Esaki Tunnel Diode Task Group, 1965), there is no industry standard for the newer resonant tunneling diodes.

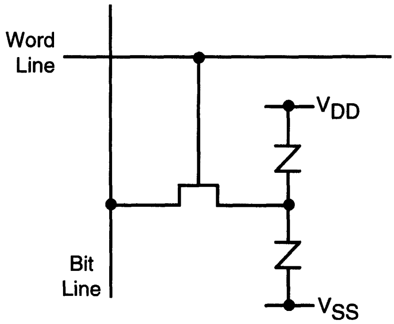

FIG. 17. Basic tunnel-diode memory cell consisting of a pass transistor and a pair of tunnel diodes. Power supplies, V_{DD} and V_{SS}, provide the appropriate biases to the tunnel-diode pair; the voltage difference, $V_{DD} - V_{SS}$, is typically less than 1 V.

input pulse at the "Trigger In" port provides sufficient current to exceed the peak current of the tunnel diode. The voltage across the tunnel diode then increases abruptly to a post-resonance voltage corresponding to the near-peak current bias. Using Esaki diodes, switching times for this transition (10% to 90%) can be less than 20 ps for a 230-mV signal into a 50-Ω load.

A monolithic RTD sampling circuit has been reported by Miura *et al.* (1990) with switching speeds of about 10 ps and estimated sampling frequency of 33 GHz. The fastest RTDs, measured in a similar configuration by Shimizu *et al.* (1995), have speeds of 1.5 ps.

4.2 Memory

The Esaki tunnel-diode "twin" conceived by Goto *et al.* (1960) has been revisited in recent years with resonant tunneling diodes and considered for cache memory and high-speed shift registers. The basic bistable cell consists of the series combination of two RTDs where one RTD forms the load of the other. The RTD pair forms a bistable element, shown in Fig. 17, which can be written to and read from "via" gates attached at the midpoint of the "twin" junction.

Extending the "twin" concept, it has been shown by Seabaugh *et al.* (1992) that multivalued RTD memory can be constructed by the series combination of resonant tunneling diodes epitaxially integrated into a single device stack. Shown in Fig. 18 is the measured *I–V* characteristic of an 8-RTD stack used in a nine-state memory circuit.

An ultralow-power RTD/heterojunction field-effect transistor static random-access memory based on InP substrates has been demonstrated by van der Wagt *et al.* (1996). Over 200 times lower power is obtained over the best previous GaAs-based memory cell; in the same cell configuration, ultralow-power tristate memory has also been demonstrated. In this circuit, which can have the same topology as Fig. 17, the RTD currents are comparable to the leakage currents in the transistor, and the storage node capacitance is supplied by the high intrinsic capacitance of the bistable RTD pair. Read and write operations occur as in dynamic random access memory (DRAM), but the cell is static on account of the RTD pair bistability.

4.3 Analog-to-Digital Conversion

It has been shown by S.-J. Wei *et al.* (1993) and Potter *et al.* (1993) that a compact folded analog-to-digital converter (ADC) can be constructed from the combination of

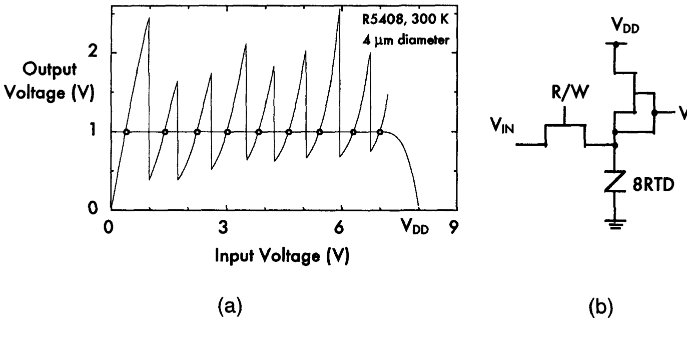

(a)

(b)

FIG. 18. Nine-state resonant tunneling diode memory cell: (a) nine-state RTD current–voltage characteristic with the field-effect-transistor load line superimposed (the nine circles denote the stable operating points of the RTD/FET combination); (b) schematic circuit diagram. The symbol labeled 8RTD represents the series integration of eight resonant tunneling diodes. Input and output ports are labeled V_{IN} and V_O, respectively, the single power supply voltage is labeled V_{DD}, and the read/write pass transistor is labeled R/W.

stacked resonant tunneling diodes and transistors. A four-bit, 2-GSps monolithic flash converter containing RTDs has also been demonstrated (Broekaert *et al.*, 1997). These converters show promise for extending ADC performance to microwave input bandwidths with low bit resolutions (four to six bits).

4.4 Logic

For logic, the most promising circuits combine the tunnel diode with transistors (Chow, 1964; Capasso *et al.*, 1990; Seabaugh *et al.*, 1993), since circuits with the tunnel diodes alone cannot meet the fan-out requirements of general-purpose logic. In recent years, pipelined approaches have been proposed for ultrafast arithmetic circuits by Mohan *et al.* (1994), and concepts for superpass gates composed of resonant-tunneling-diode/transistor combinations have been proposed by Deng *et al.* (1995). Compact multivalued adder circuits have been described by Micheel *et al.* (1993) and are now being developed.

4.5 Oscillators

When the resonant tunneling diode is biased in the negative–differential-resistance portion of its *I–V* characteristic, the circuit will oscillate. These oscillations are commonly observed when making dc measurements through the NDR region. Power is generated by these oscillations and can be utilized as signal sources to extremely high frequencies. Oscillation frequency up to 712 GHz in resonant tunneling diodes has been demonstrated by Brown (1994). The design principles for tunnel-diode oscillators are well described in the book by Chow (1964).

4.6 Tunnel Photodiodes

The use of the Esaki tunnel diode as a light-triggered switch has been explored by Allen *et al.* (1969) and recently for the resonant tunneling diode by Moise *et al.* (1995b, 1997) and others. In this application the tunnel diode is biased to near its current peak with a resistive load, which provides two stable bias points in the unilluminated condition. On illumination with a subnanosecond pulse with a few milliwatts of peak power, the diode switches from a low-voltage state to a high-voltage state. To reset the circuit, the diode bias is momentarily clocked to a low bias. Cutoff for such a circuit is expected to exceed 10 GHz with a photosensitivity of at least 300 mA/W (Allen *et al.*, 1969).

The light triggering mechanism is fundamentally different for the two devices. In the Esaki diode, the incident light generates electron–hole pairs in the depletion region of the *p-n* junction. This contributes a photocurrent that causes the switching from a low-voltage to high-voltage state. In contrast to the Esaki diode, photogenerated electrons and holes in the RTD are separated and confined at the tunnel barrier interfaces, where they act to screen the electric field and alter the resonant tunneling peak voltage, thereby triggering the bistability.

5. SUMMARY AND PROSPECTS

Continued improvements in crystal-growth technology, particularly in molecular-beam epitaxy, have increased the possibilities for design and construction of resonant tunneling and Esaki tunnel diodes. The device possibilities are expanding with the development of silicon-based heterojunction technology. A long-standing impediment to the use of tunnel diodes has been the inability of the device to integrate with conventional transistor technology. This limitation is now being challenged by the emergence of III–V resonant tunneling-diode/transistor integrated circuits. If the performance improvements long anticipated for an integrated tunnel-diode/transistor technology are realized, one can expect tunnel diodes to become as commonplace as transistors in the not-too-distant future.

GLOSSARY

Epitaxy: The process of growing a crystalline material upon a crystalline substrate. Often the material grown has a different chemical composition from that of the underlying material, but with a similar crystal structure.

Esaki Diode: A heavily doped n^+p^+ junction diode wherein electrons in the conduction band of the n^+ region can tunnel into hole states in the valence band of the p^+ region.

Excess Current: The measured current in excess of the thermionic and resonant or interband tunneling current in the tunnel diode. This current can be accounted for by mechanisms such as impurity-assisted tunneling, interface-roughness scattering, excited-state tunneling, phonon-assisted tunneling, or edge leakage.

Heteroepitaxy: The growth of layered crystalline materials where the layers differ in chemical composition but have similar crystalline structure, such as AlAs on GaAs.

Heterojunction: A junction formed between dissimilar semiconductors.

Heterostructure: A combination of heterojunctions typically constructed to enhance some device property.

Homojunction: A semiconductor junction consisting of *n*- and *p*-type regions with the same semiconductor on either side of the junction.

Hysteresis: In the tunnel diode, a term referring to a current–voltage characteristic that differs depending on the direction (increasing or decreasing) of the voltage or current sweep and the previous bias state of the device.

Interband: Between two bands such as the conduction and valence band.

Intraband: Within a single band such as the conduction band (or valence band).

Linear Region: The voltage region of a resonant tunneling or Esaki diode before the peak in the current–voltage characteristic, where the current–voltage relation is nearly linear.

Longitudinal: Refers to a direction parallel to the current flow. In heteroepitaxial structures, this direction is the same as the crystal growth direction and is perpendicular to the heteroepitaxial layers.

Longitudinal Energy: Total energy minus transverse kinetic energy, $E - (p_y^2 + p_z^2)/2m^*$.

Negative Differential Resistance (NDR): The region of the *I–V* curve immediately following the peak current, in which the current decreases with increase in applied voltage.

Peak Current: A local maximum in the current–voltage characteristic of the tunnel diode, at which the derivative of the current with respect to the voltage becomes zero.

Peak-to-Valley Current Ratio (PVR): The ratio of the peak current to the valley current of the tunnel diode, often used as a figure of merit.

Resonant Tunneling: Tunneling enhanced by constructive interference resulting in a resonant buildup of the wavefunction within a cavity defined by tunnel barriers. The optical analogy for semiconductor resonant tunneling is the Fabry-Pérot cavity.

Total Energy, *E*: The potential energy plus the kinetic energy.

Transmission Coefficient: Ratio of the probabilities of the transmitted wave and the incident wave.

Transverse: Refers to a direction perpendicular to the current flow. In heteroepitaxial structures, this direction is perpendicular to the crystal growth direction and is parallel to the heteroepitaxial layers.

Tunnel Diode: A two-terminal device whose charge transport is governed by the principles of quantum-mechanical tunneling.

Tunneling: The act of passing through a classically forbidden region such as a potential barrier.

Valley Current: A local minimum in the current–voltage characteristic of the tunnel diode at which the derivative of the current with respect to the voltage becomes zero.

Works Cited

Allen, L. B., Rice, R. R., Mathews, R. F. (1969), *Appl. Phys. Lett.* **15,** 414–417.

Andrews, J. R. (1986), in: E. K. Miller (Ed.), *Time Domain Measurements in Electromagnetics,* New York: Van Nostrand Reinhold, p. 95.

Baasch, T. L. (1969), *Electron. Prod.* **12** (6), 22–29.

Bagwell, P. F., Broekaert, T. P. E., Orlando, T. P., Fonstad, C. G. (1990), *J. Appl. Phys.* **68,** 4634–4646.

Batdorf, R. L., Dacey, G. C., Wallace, R. L., Walsh, D. J. (1960), *J. Appl. Phys.* **31** 613–614.

Bedair, S. (1979), *J. Appl. Phys.* **50,** 7267–7268.

Beenakker, C. W. J., Van Houten, H. (1991), in: H. Ehrenreich, D. Turnbull (Eds.), *Solid State Physics Advances in Research and Applications,* Vol. 44, New York: Academic, p. 16.

Boykin, T. B., Carnahan, R. E., Martin, K. P. (1995), *Phys. Rev. B* **51,** 2273–2281.

Brehmer, D. E., Zhang, K., Schwarz, Ch. J., Chau, S.-P., Allen, S. J., Ibbetson, J. P., Zhang, J. P., Palmstrøm, C. J., Wilkens, B. (1995), *Appl. Phys. Lett.* **67,** 1268–1270.

Broekaert, T. P. E., Fonstad, C. G. (1990), *J. Appl. Phys.* **68,** 4310–4312.

Broekaert, T. P. E., Brar, B., van der Wagt, J. P. A., Seabaugh, A. C., Moise, T. S., Morris, F. J., Beam, E., III, Frazier, G. A. (1997), in: *Proceedings of the IEEE GaAs Integrated-Circuit Symposium,* to be published.

Brown, E. R. (1994), in: N. G. Einspruch, W. R. Frensley (Eds.), *Heterostructures and Quantum Devices,* Orlando, FL: Academic, pp. 305–350.

Brugger, H., Meiners, U. Wölk, Marten A., Rossmanith, M., Klitzing, K. v., Sauer, R. (1991), *Microelectron. Eng.* **15,** 663–666.

Burrus, C. A. (1961), *Proc. IRE* **49,** 1101.

Burrus, C. A. (1962), *Solid-State Electron.* **5,** 357–358.

Capasso, F., Sen, S., Beltram, F. (1990), in: S. M. Sze, (Ed.), *High Speed Semiconductor Devices,* New York: Wiley, p. 465.

Chang, L. L., Esaki, L., Tsu, R. (1974), *Appl. Phys. Lett.* **24,** 593–595.

Chevoir, F., Vinter, B. (1993), *Phys. Rev. B* **47,** 7260–7273.

Chow, D. H., Schulman, J. N., Özbay, E., Bloom, D. M. (1992), *Appl. Phys. Lett.* **61,** 1685–1687.

Chow, D. H., Schulman, J. N., Özbay E., Bloom, D. M. (1993), in: C. W. Tu, D. C. Houghton, R. T. Tung (Eds.), *Semiconductor Heterostructures for Photonic and Electronic Applications,* Materials Research Society Symposium Proceedings Vol. 281, Pittsburgh: Materials Research Society, pp. 269–274.

Chow, D. H., Dunlap, H. L., Williamson, W., III, Enquist, S., Gilbert, B. K., Subramaniam, S., Lei, P.-M., Bernstein, G. H. (1996), *IEEE Electron Dev. Lett.* **17,** 69–71.

Chow, W. F. (1964), *Principles of Tunnel Diode Circuits,* Melbourne, FL: Krieger Publishing.

Chynoweth, A. G., Felmann, W. L., Logan, R. A. (1961), *Phys. Rev.* **121,** 684–694.

Cohen, G. M., Ritter, D., Cytermann, C. (1995), *Electron. Lett.* **31,** 1511–1512.

Collins, D. A., Chow, D. H., Yu, E. T., Ting, D. Z. Y., Söderström J. R., Rajakarunanayake, Y., McGill, T. C. (1991), in: L. L. Chang, E. E. Mendez, C. Tejedor (Eds.), *Resonant Tunneling in Semiconductors: Physics and Applications,* New York: Plenum Press, p. 515.

Cunningham, J. E., Goldman, V. J., Su, B., Jan, W. Y. (1993), *J. Cryst. Growth* **127,** 695–699.

Davis, R. E., Gibbons, G. (1967), *Solid-State Electron.* **10,** 461–472.

Day, D. J., Yang, R. Q., Lu, J., Xu, J. M. (1993), *J. Appl. Phys.* **73,** 1542–1544.

Deng, X., Hanyu, T., Kameyama, M. (1995), *IEE Proc. Circuits Dev. Syst.* **142,** 299–306.

Diamond, S. K., Özbay, E., Rodwell, M. J., Bloom, D. M. (1989), *Appl. Phys. Lett.* **54,** 153–155.

Dmitriev, V. A., Ivanov, P. A., Strel'chuk, A. M., Syrkin, A. L., Popov, I. V., Chelnokov, V. E. (1985), *Pis'ma Zh. Tekh. Fiz.* **11,** 976–978 (in Russian).

Duke, C. B. (1969), *Tunneling in Solids,* New York: Academic.

Esaki, L. (1958), *Phys. Rev.* **109,** 603–604.

Esaki, L. (1976), *IEEE Trans. Electron Dev.* **23,** 644–647.

Esaki Tunnel Diode Task Group (1965), *IEEE Trans. Electron Dev.* **12,** 374–386.

Fang, Y. K., Chen, K.-H., Liu, C.-R., Hwang, J.-D., Wu, K.-S., Liou, W.-R. (1994), *IEEE J. Quan. Electron.* **30,** 2293–2296.

Franks, V. M., Hulme, K. F., Morgan, J. R. (1965), *Solid-State Electron,* **8,** 343–344.

Frensley, W. R. (1994), in: N. G. Einspruch, W. R. Frensley (Eds.) *Heterostructures and Quantum Devices,* Orlando, FL: Academic, p. 273.

Fu, Y., Chen, Q., Willander, M. (1991), *J. Appl. Phys.* **70,** 7468–7473.

Gennser, U., Kesan, V. P., Iyer, S. S., Bucelot, T. J., Yang, E. S. (1990), *J. Vac. Sci. Technol.* **8,** 210–213.

Goto, E., Murata, K., Nakazawa, K., Nakagawa, K., Moto-oka, T., Matsuoka, Y., Ishibashi, Y., Ishida, H., Soma, T., Wada, E. (1960), *IRE Trans. Electron. Comp.* **EC-9,** 25–29.

Han, J., Durbin, S. M., Gunshor, R. L., Kobayashi, M., Menke, D. R., Pelekanos, N., Hagerott, M., Nurmikko, A. V., Nakamura, Y., Otsuka, N. (1991), *J. Cryst. Growth* **111,** 767–771.

Hirose, M., Morita, M., Osaka, Y. (1977), *Jpn. J. Appl. Phys.* **16,** Suppl. 16–1, 561–564.

Holonyak, N., Jr., Lesk, I. A. (1960), *Proc. IRE* **48,** 1406–1409.

Iogansen, L. V. (1965), *Usp. Fiz. Nauk* **86,** 175–179 (*Sov. Phys. Uspekhi* **8,** 413–416).

Ismail, K., Meyerson, B. S., Wang, P. J. (1991), *Appl. Phys. Lett.* **59,** 973–975.

Jo, J., Li, H. S., Chen, Y. W., Wang, K. L. (1994), *Appl. Phys. Lett.* **64,** 2276–2278.

Kan, S. C., Morkoç, H., Yariv, A. (1988), *Appl. Phys. Lett.* **52,** 2250–2251.

Kapre, R. M., Madhukar, A., Guha, S. (1991), *Appl. Phys. Lett.* **58,** 2255–2257.

Kleinknecht, H. P. (1961), *Solid-State Electron.* **2,** 133–142.

Lake, R., Klimeck, G., Datta, S. (1993), *Phys. Rev. B* **47,** 6427–6438. See Eq. (4.15) written for a single spin channel and a single transverse momentum.

Lheurette, E., Grimbert, B., François, M., Tilmant, P., Lippens, D., Nagle, J., Vinter, B. (1992), *Electron. Lett.* **28,** 937–938.

Lippens, D., Barbier, E., Mounaix, P. (1991), *IEEE Electron Dev. Lett.* **12,** 114–116.

Liu, H. C., Aers, G. C. (1989), *J. Appl. Phys.* **65,** 4908–4914.

Liu, H. C., Landheer, D., Buchanan, M.,

Houghton, D. C. (1988), *Appl. Phys. Lett.* **52,** 1809–1811.

Luo, L. F., Beresford, R., Wang, W. I. (1989), *Appl. Phys. Lett.* **55,** 2023–2025.

Luryi, S. (1985), *Appl. Phys. Lett.* **47,** 490–492.

Malov, V. V., Lyalina, E. V., Iogansen, L. V. (1988), *Pis'ma Zh. Tekh. Fiz.* **14,** 1019–1024 (*Sov. Tech. Phys. Lett.* **14,** 450–452).

Matyi, R. J. (1994), in: N. G. Einspruch, W. R. Frensley (Eds.), *Heterostructures and Quantum Devices,* Orlando, FL: Academic, 25–61.

McDermott, J. (1981), *EDN* **26** (12), 57–64.

Meyerhofer, D., Brown, G. A., Sommers, H. S., Jr. (1962), *Phys. Rev.* **126,** 1329–1341.

Micheel, L. J., Taddiken, A. H., Seabaugh, A. C. (1993), *Proceedings of the 23rd International Symposium on Multiple Valued Logic,* Los Alamitos, CA: IEEE Computer Society Press, pp. 164–169.

Mikkelson, C. H., Seabaugh, A. C., Beam, E. A., Luscombe, J. H., Frazier, G. A. (1994), *IEEE Trans. Electron Dev.* **41,** 132–137.

Miura, A., Yakihara, T., Uchida, S., Oka, S., Kobayashi, S., Kamada, H., Dobashi, M. (1990), *IEEE Trans. Microwave Theor. Tech.* **38,** 1980–1985.

Miyazaki, S., Ihara, Y., Hirose, M. (1987), *Phys. Rev. Lett.* **59,** 125–127.

Mizuta, H., Tanoue, T. (1995), *The Physics and Applications of Resonant Tunnelling Diodes,* Cambridge, U.K.: Cambridge Univ. Press.

Mohan, S., Mazumder, P., Haddad, G. I. (1994), *IEE Proc. Comput. Digit. Tech.* **141,** 104–110.

Moise, T. S., Kao, Y.-C., Katz, A. J., Broekaert, T. P. E., Celii, F. G. (1995a), *J. Appl. Phys.* **78,** 6305–6317.

Moise, T. S., Kao, Y.-C., Garrett, L. D., Campbell, J. C. (1995b), *Appl. Phys. Lett.* **66,** 1104–1106.

Moise, T. S., Kao, Y.-C., Goldsmith, C. L., Schow, C. L., Campbell, J. C. (1997), *IEEE Photonics Technol. Lett.* **9,** 803–805.

Mounaix, P., Lheurette, E., Mollot, F., Lippens, D. (1995), *Electron. Lett.* **31,** 1508–1510.

Muto, S., Inata, T. (1994), *Semicond. Sci. Technol.* **9,** 1157–1170.

Ohno, H., Mendez, E. E., Wang, W. I. (1990), *Appl. Phys. Lett.* **56,** 1793–1795.

Özbay, E., Bloom, D. M., Chow, D. H., Schulman, J. N. (1993), *IEEE Electron Dev. Lett.* **14,** 400–402.

Payne, M. C. (1985), *J. Phys. C: Solid-State Phys.* **18,** L879–L883.

Payne, M. C. (1986), *J. Phys. C: Solid-State Phys.* **19,** 1145–1155.

Pearson, G. L. (1957), *Bell Laboratories Memorandum—Case 38139-104,* November 1, 1957. In this memorandum, Pearson describes a series of electrical measurements he made on a set of *p-n* diodes differing in the resistivity of the starting substrate. He observed that the diodes fabricated on the low-resistivity materials (<0.005 Ω-cm) resulted in a "rectification of the opposite sign" compared to the diodes on higher resistivity materials. In this memo, A. G. Chynoweth is credited with the correct interpretation of this phenomenon, "Zener emission" or tunneling.

Potter, R. C., Fathimulla, A., Hier, H., Shupe, D., Abrahams, J., Eichner, C., Jones, D., Wei, S.-J., Lin, H. C. (1993), in: *Proceedings of the Fifth International Conference on InP and Related Materials (post deadline paper),* New York: IEEE, pp. 37–40.

Reddy, V. K., Tsao, A. J., Neikirk, D. P. (1990), *Electron. Lett.* **26,** 1742–1744.

Reed, M. A., Koestner, R. J., Goodwin, M. W. (1986), *Appl. Phys. Lett.* **49,** 1293–1295.

Reed, M. A., Randall, R. J., Aggarwal, R. J., Matyi, R. J., Moore, T. M., Wetsel, A. E. (1988), *Phys. Rev. Lett.* **60,** 535–537.

Richard, R. A., Chen, E. I., Sugg, A. R., Höfler, G. E., Holonyak, N., Jr. (1993), *Appl. Phys. Lett.* **63,** 3613–3615.

Roblin, P., Liou, W. (1993), *Phys. Rev. B* **47,** 2146–2161.

Seabaugh, A. C., Aldert, R. K. (1996), unpublished.

Seabaugh, A. C., Kao, Y. C., Yuan, H. T. (1992), *IEEE Electron Dev. Lett.* **13,** 479–481.

Seabaugh, A. C., Beam, E. A., III, Taddiken, A. H., Randall, J. N., Kao, Y.-C. (1993), *IEEE Electron Dev. Lett.* **14,** 472–474.

Seabaugh, A., Brar, B., Broekaert, T., Frazier, G., Morris, F., van der Wagt, P., Beam III, E. (1997), in: *Proceedings of the IEEE GaAs Integrated Circuits Symposium,* to be published.

Shen, C. C., Chang, P. T., Choi, K. Y. (1993), *Appl. Phys. A* **56,** 153–155.

Shen, G. D., Xu, D. X., Willander, M., Hansson, G. V. (1992), *J. Appl. Phys.* **71,** 3365–3369.

Shimizu, N., Nagatsuma, T., Shinagawa, M., Waho, T. (1995a), *IEEE Electron Dev. Lett.* **16,** 262–264.

Shimizu, N., Nagatsuma, T., Shinagawa, M., Waho, T., Yaita, M., Yamamoto, M. (1995b), *Electron. Lett.* **31,** 1695–1697.

Smet, J. H., Broekaert, T. P. E., Fonstad, C. G. (1992), *J. Appl. Phys.* **71,** 2475–2477.

Söderström, J. R., Chow, D. H., McGill, M. C. (1989), *Appl. Phys. Lett.* **55,** 1094–1096.

Söderström, J. R., Brown, E. R., Parker, C. D., Mahoney, L. J., Yao, J. Y., Andersson, T. G., McGill, T. C. (1991a), *Appl. Phys. Lett.* **58,** 275–277.

Söderström, J. R., Yao, J. Y., Andersson, T. G. (1991b), *Appl. Phys. Lett.* **58,** 709–710.

Støvneng, J., Hauge, E. H. (1991), *Phys. Rev. B* **44,** 13582–13594.

Suemasu, T., Watanabe, M., Asada, M., Suzuki, N. (1992), *Electron. Lett.* **28,** 1432–1433.

Sugiyama, Y., Inata, T., Muto, S., Nakata, Y., Hiyamizu, S. (1988), *Appl. Phys. Lett.* **52,** 314–316.

Sze, S. (1981), *Physics of Semiconductor Devices,* 2nd ed., New York: Wiley, pp. 513–565.

Tiemann, J. J. (1966), U.S. Patent No. 3,234,064.

Timmons, M., Bedair, S. (1980), *J. Appl. Phys.* **52,** 1134–1135.

Ting, D. Z.-Y., McGill, T. C. (1995), in: D. K. Ferry, H. L. Grubin, C. Jacoboni, A. Jauho (Eds.), *Quantum Transport in Ultrasmall Devices,* New York: Plenum Press, pp. 417–436.

Tsai, H. H., Su, Y. K., Lin, H. H., Wang, R. L., Lee, T. L. (1994), *IEEE Electron Dev. Lett.* **15,** 357–359.

van der Wagt, J. P. A., Seabaugh, A. C., Beam, E. A., III (1996), *Technical Digest IEDM,* 425–428.

Watson, H. A. (1969), *Microwave Semiconductor Devices and Their Circuit Applications,* New York: McGraw Hill, Table 13.1, p. 415.

Wei, S.-J., Lin, H. C., Potter, R. C., Shupe, D. (1993), *IEEE J. Solid-State Circ.* **28,** 697–700.

Wei, T., Stapleton, S., Berolo, E. (1993), *J. Appl. Phys.* **73,** 829–834.

Xu, J. M., MacDonald, A. G., Iogansen, L. V., Day, D. J., Sweeny, M. (1992), *Semicond. Sci. Technol.* **7,** 1097–1102.

Yang, L., Chen, J. F., Cho, A. F. (1990), *Electron. Lett.* **26,** 1277–1279.

Further Reading

Brown, E. R. (1994), in: N. G. Einspruch, W. R. Frensley (Eds.), *Heterostructures and Quantum Devices,* Orlando, FL: Academic, pp. 305–350.

Lee, M. A., Easter, B., Bell, H. A. (1967), *Tunnel Diodes,* London: Chapman and Hall.

Mizuta, H., Tanoue, T. (1995), *The Physics and Applications of Resonant Tunnelling Diodes,* Cambridge, U.K.: Cambridge Univ. Press.

Okean, H. C. (1971), "Tunnel Diodes," in: R. K. Willardson, A. C. Beer (Eds.), *Semiconductors and Semimetals,* Vol. 7, *Applications and Devices,* part B, New York: Academic, pp. 473–624.

Seabaugh, A. C., Reed, M. A. (1994), in: N. G. Einspruch, W. R. Frensley (Eds.), *Heterostructures and Quantum Devices,* Orlando, FL: Academic, pp. 351–383.

TUNNELING

See RESONANT TUNNELING

TUNNELING MICROSCOPY AND SPECTROSCOPY

L. J. WHITMAN, *Naval Research Laboratory, Washington, D.C., U.S.A.*

	Introduction	361
1.	**Microscope Components**	362
1.1	Scanner	363
1.2	Coarse Approach	363
1.3	Vibration Isolation	365
1.4	Control Electronics	365
1.5	Tip Preparation	366
2.	**Principles of Imaging and Spectroscopy**	366
2.1	Tunneling Physics	366
2.2	Density-of-States Effects	367
2.3	Imaging Modes	367
2.4	Tunneling Spectroscopy	368
3.	**Applications**	369
3.1	Structure of Crystal Surfaces	369
3.1.1	Metals	369
3.1.2	Semiconductors	370
3.1.3	Oxides	372
3.2	Adsorption and Film Growth	372
3.2.1	Atom-Resolved Surface Chemistry	373
3.2.2	Thin-Film Morphology	374
3.3	Electrochemistry	376
3.4	Surface Electronic Structure	376
3.4.1	Conductance versus Structure	376
3.4.2	Chemically Selective Imaging	376
3.5	Bulk Electronic Structure	378
3.6	Atom Manipulation and Nanolithography	379
3.6.1	Electron-Stimulated Patterning	379
3.6.2	Positioning Single Atoms	380
4.	**Related Techniques**	380
	Acknowledgment	382
	Glossary	382
	Works Cited	382
	Further Reading	383

INTRODUCTION

Scanning tunneling microscopy (STM) is a laboratory technique capable of obtaining atomic-scale resolution images of surfaces. It can also be used for tunneling spectroscopy measurements that reveal a variety of surface electronic properties with similar spatial resolution. Moreover, under appropriately controlled conditions an STM can be used to move individual surface atoms and molecules with atomic-scale precision. For such a powerful instrument, the basic operating principles are remarkably simple. As illustrated in Fig. 1, a sharp metal tip is brought within 1 nm of the surface of an electrically conductive sample, and a few volts are applied between the two, enabling electrons to tunnel quantum mechanically across the tip–surface gap. The resulting tunnel current, typically on the order of 1 nA, depends exponentially on the gap: A change of 0.1 nm causes the current to change by a factor of ~10, giving STM atomic-scale sensitivity. If the gap is regulated to keep the current constant while the tip is scanned laterally, adjacent profiles of the surface topography can be assembled into an image via computer. As shown in Fig. 2, images of the surface topography can then be displayed in gray scale (with high points brighter) or artificially rendered as a 3D surface (usually with exaggerated height).

STM has its origins in the "topografiner" developed in the early 1970s (Young *et al.*, 1972), that included most of the elements of an STM but operated with a larger tip-to-surface gap (>1 nm, at which distance electron transport occurs via field emission). Deficiencies in both the mechanical and electrical systems at that time limited the resolution to a few nanometers vertically and ~0.5 μm laterally. These problems were overcome about ten years later by Gerd Bin-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

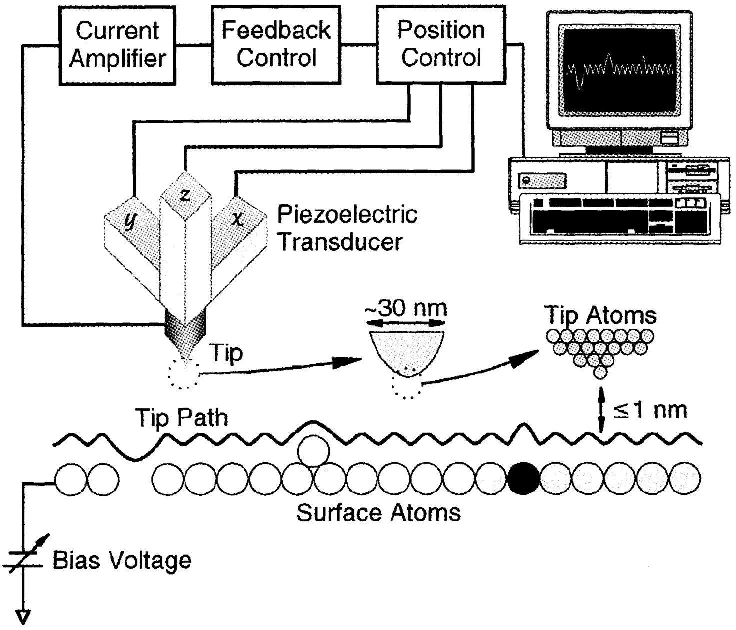

FIG. 1. The basic elements of scanning tunneling microscopy.

nig and Heinrich Rohrer at the IBM Rüschlikon laboratory, who succeeded in creating an instrument with stable vacuum tunneling and precision scanning capabilities—the conditions required for atomic resolution imaging—for which they were awarded the 1986 Nobel prize in physics (Binnig *et al.*, 1982a,b). STM has revolutionized the study of surfaces and is rapidly becoming a required tool in every surface-characterization laboratory. In addition, it has led to the development of a host of related techniques, collectively known as scanning-probe microscopies, and has spawned an industry aimed at providing reliable, easy-to-use commercial instruments.

Given the many comprehensive books and review articles about STM that are available (see Further Reading), only the experimental details and theoretical concepts essential for a general understanding of the operation and application of STM are presented here. The technical components of the microscope are discussed first, followed by the theoretical principles of operation. Applications are then presented that illustrate the capabilities and limitations of STM under optimal conditions. Lastly, some related techniques are discussed.

1. MICROSCOPE COMPONENTS

To achieve atomic-scale resolution, the mechanical and electrical components of an STM must enable the tip to be positioned within approximately 1 nm of the surface and then controlled both vertically and laterally with a precision of <0.01 nm. This precision requires low-vibration mechanical systems combined with low-noise electrical circuitry. Myriad ways have been developed to achieve the required mechanical stability; the more common elements are illustrated in Fig. 3. The STM must include a coarse approach mechanism, such as the "inchworm" motor shown, capable of bringing the tip into tunneling range with the sample. Once in range, a scanning transducer controls the vertical and lateral position of the tip over the sample. The sample and scanner are mounted in a rigid framework to minimize the effect of mechanical resonances. A vibration isolation system, often based on a magnetically damped spring suspension, is usually required to reduce transmission of external vibrations. Photographs of a complete ultrahigh-vacuum (UHV) -based system is shown in Fig. 4, including the UHV system required to maintain a pristine sample surface.

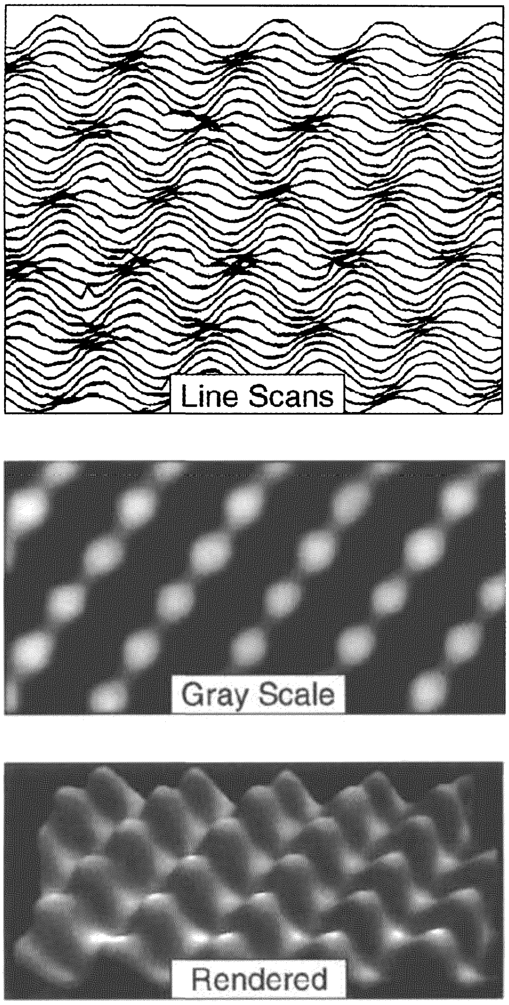

FIG. 2. Three common ways of presenting an STM image: as a set of line scans; in gray scale, with higher points brighter; and as an artificially rendered surface.

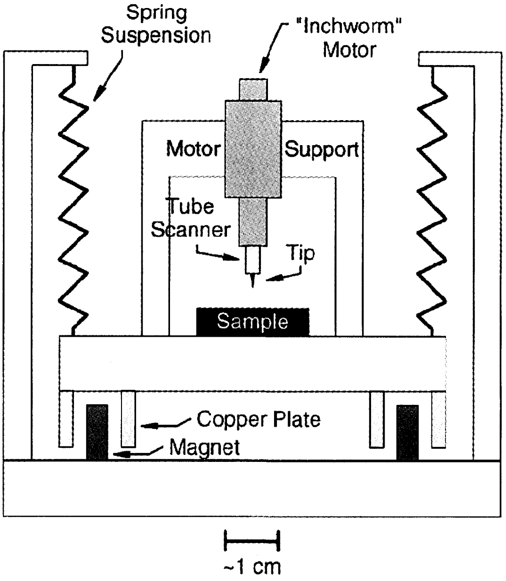

FIG. 3. Common elements of an STM, including a tip mounted on a piezoelectric tube scanner, a coarse-approach mechanism (in this case an "inchworm" motor), and a damped vibration-isolation system.

1.1 Scanner

The position transducer is the central element of the STM and is almost universally composed of the piezoelectric ceramic PZT [Pb(Zr,Ti)O$_3$]. This material expands or contracts approximately linearly with the voltage applied across it (see PIEZOELECTRIC DEVICES). Although early STM designs used a tripod of three orthogonal bars to control motion in three dimensions, a tube geometry is more common now on account of its inherent rigidity. The design of a typical tube scanner is shown in Fig. 5 (Binnig and Smith, 1986). A thin-walled tube of PZT is coated with metal on the inner and outer surfaces, with the outer surface divided into four equal quadrants by narrow metal-free regions. The inner surface is then used as the electrode for the vertical displacement (z), and the opposing pairs of outer electrodes are used for lateral motion (x and y). In this geometry a voltage applied between the inner and outer electrodes causes the tube to change length. Opposite voltages applied across opposing outer electrodes will induce one side of the tube to contract and the other to expand, causing a net lateral displacement (along with a small vertical displacement that can usually be ignored). For example, a 15-mm-long, 2.5-mm-radius, 0.75-mm-thick tube of a common commercial PZT material driven by voltages between $+150$ and -150 V would provide an overall scan range of 1.0 μm vertically and 2.8 μm laterally.

1.2 Coarse Approach

The tip–surface gap can usually be reduced to <0.1 mm via optical inspection. The basic requirement of the subsequent

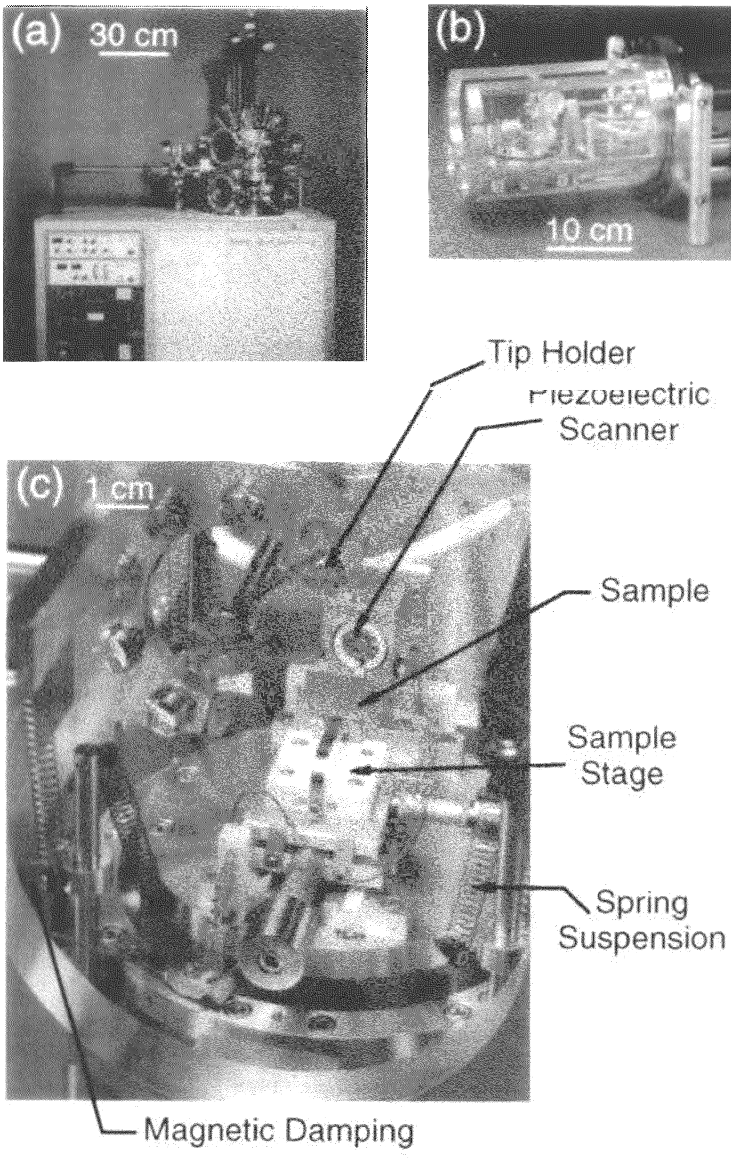

FIG. 4. Photographs of a commercially available STM (AutoProbe® VP, Park Scientific Instruments, Sunnyvale, CA). (a) Complete ultrahigh-vacuum system. (b) STM mounted on vacuum flange. (c) Close-up of STM showing the tip holder being transferred to the scanner.

coarse approach is that the remaining gap be reduced in steps less than the *z*-deflection range of the piezotube. This requirement allows a step to occur with the piezotube retracted, followed by a controlled expansion toward the surface (in a search for tunneling), without crashing the tip into the surface. Myriad ways have been developed to achieve this result. Two common techniques, the "inchworm" motor and the inertial slider, are depicted in Fig. 6. In an inchworm motor [Fig. 6(a)], a shaft is clamped inside a piezotube by additional piezoelectric material. If the position of the outer tube is fixed at one point, an inchwormlike series of clamping/unclamping and extension/retraction events will move the shaft. For simplicity, a tube scanner and tip can be mounted directly onto the end of an inchworm shaft (as in Fig. 3). Alternatively, an inchwormlike

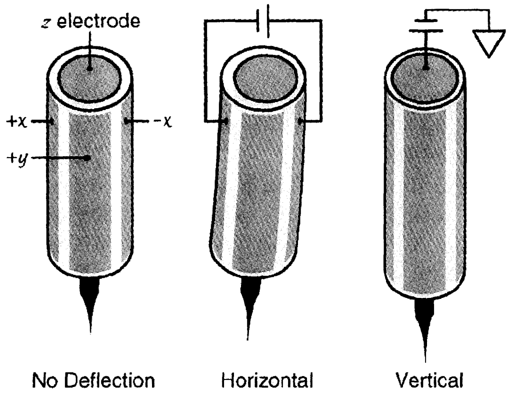

FIG. 5. Operation of a piezoelectric tube scanner with a four-quadrant outer electrode for horizontal motion and single inner electrode for vertical motion. A horizontal deflection occurs when a voltage is applied across opposing outer electrodes. A vertical deflection occurs when a voltage is applied to the inner electrode with respect to the four outer electrodes.

device can be used to push/pull the sample or scanner along a sliding mechanism.

Inertial motion uses asymmetrical acceleration to induce asymmetrical stick-slip motion of a mass. For example [Fig. 6(b)], if the transducer is first contracted slowly enough to prevent sliding, the mass will move with it; if the transducer is then expanded rapidly enough to induce sliding between the mass and its contact with the transducer, the mass will remain in place during the expansion, thereby effecting a net movement toward the anchor point. A variety of implementations of this strategy have been developed, with the tip, scanner, or sample stage acting as the mass.

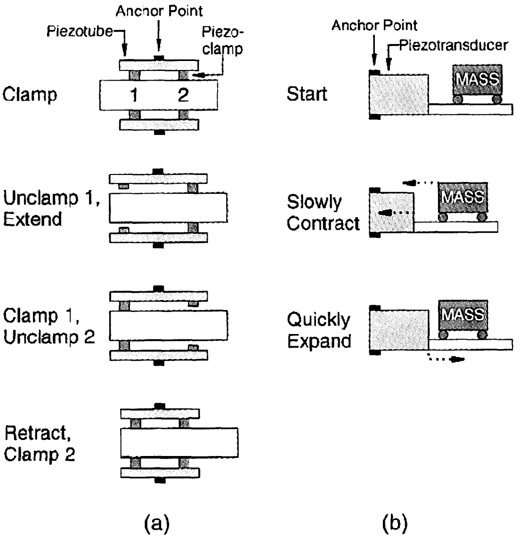

FIG. 6. Operating principles for two common coarse-approach mechanisms. (a) An "inchworm" motor employs a piezotube and piezoclamps to move an inner shaft by a series of clamping/unclamping and extension/retraction events. (b) Inertial motion of a free mass can be achieved by asymmetrical acceleration.

1.3 Vibration Isolation

The amplitude of any vibrations between the tip and surface must be reduced as much as possible to achieve optimum resolution. To achieve atomic-scale resolution the noise must be well below 0.01 nm within the frequency response of the tunneling control system (typically ~10 kHz). Because the amplitude of floor vibrations and other acoustic noise in this frequency range typically ranges from 0.1 to 1 μm, this is a considerable challenge. The transmission of these vibrations can be reduced by making the components of the STM as rigid as possible, especially those in the mechanical loop between the tip and sample, and mounting the scanner-plus-sample assembly on a vibration-isolation stage. There are two basic approaches to vibration isolation. The first involves suspending the assembly from one or more sets of springs with resonant frequencies of a few hertz. Because a spring suspension has little inherent damping (it will easily oscillate at its resonant frequency), external damping is usually added in the form of permanent magnets suspended between copper plates (see Fig. 3). So called eddy-current damping works by electromagnetically inhibiting the relative motion between the magnets and the high-conductivity metal plates. The second approach to vibration isolation is to mount the instrument on a stack of heavy metal plates separated by rubberlike spacers. Although this approach is generally less effective than a spring system, it is easier to build and simpler to operate.

1.4 Control Electronics

The essential elements of the control electronics, as included in Fig. 1, are a current amplifier; a feedback-control system; a position-control system, including digital-to-analog converters (DAC), analog-to-digital converters (ADC), and high-voltage amplifiers to control the scanner; and a computer for overall program control and image display. The tunneling current is typically in the range of 0.1–10 nA, and to keep it constant via feedback it is usually converted to a reference voltage ~1 V; this requires a current-to-voltage amplifier with a gain of 10^8–10^{10} V/A and noise below the equivalent of 10 pA (readily achievable with commercial electronic components). Because the current varies exponentially with the tunneling gap, an amplifier that converts the input signal to its logarithm is usually employed between the current-to-voltage amplifier and the feedback in order to linearize the system response. Proportional and/or integral feedback is used to generate a voltage to control the height of the tip above the surface. Because this voltage represents the surface topography, it is sent to the ADC for storage and display. The main function of the position-control system is to convert the computer-generated low-voltage DAC signals into the appropriate high-voltage signals for control of the scanner position. To record a topography image, the tip is scanned back and forth across a surface area, taking small steps perpendicular to each scan line, while periodically recording the tip height. The result is a two-dimensional array of height values that can be displayed as an image (see Fig. 2).

1.5 Tip Preparation

The tunneling tip is the most critical component of the STM, as it ultimately determines the resolution and overall image quality. Both the shape of the tip and the types of atoms on the end affect the images; in the ideal case, tunneling occurs through a single atom at the apex of a microscopically sharp metallic tip. As the number of atoms on the apex of the tip increases, the resolution degrades, although atomic resolution can still be achieved with a cluster of a few atoms. If the surface being studied is very flat (on the nanometer scale), only the atomic-scale shape of the tip apex is important, because only those atoms near the end will contribute to tunneling. In contrast, on rougher surfaces the recorded topography may be a noticeable convolution of the surface topography and the microscopic tip shape. Because the tunneling current depends not only on the gap but on the electronic structure of the tip as well (see Sec. 2), a nonmetallic atom or cluster on the tip apex may also affect the image.

The most common materials used for tips are tungsten and platinum-iridium-alloy wires. Although good resolution can sometimes be obtained using a crudely cut wire—when nature conveniently leaves a single atom or small atom cluster on the jagged apex of the cut—more reproducible results are obtained with tips electrochemically etched to a sharp point. Pt-Ir is often used for tunneling in air because of its resistance to oxidation; W is preferred for UHV applications, in part because of the knowledge gained about its preparation for use in field-ion microscopy (FIM; see FIELD ION MICROSCOPY). W tips can be easily etched to a microscopic radius of <10 nm and, fortunately, it is not uncommon for there to be an atomic-scale cluster on the end as required for atomic-scale resolution.

2. PRINCIPLES OF IMAGING AND SPECTROSCOPY

The inherent resolution of STM is a direct consequence of the quantum-mechanical tunneling between the tip and surface. This current depends both on the width and height of the tunneling barrier, and on the electronic structure of the tip and surface. By measuring the interdependence of the various tunneling parameters, i.e., bias voltage, tunneling current, and tunneling gap, images and spectra of a variety of surface electronic properties can be recorded in addition to topography. In this section, an overview of the essential physics of these processes is provided.

2.1 Tunneling Physics

Figure 7(a) illustrates an idealized potential-energy diagram for the electrons in a metal sample and tip a distance s apart, with work functions ϕ_s and ϕ_t, respectively, and a positive bias voltage V applied to the sample. In a classical picture, the work function, which is the energy required to move an electron at the surface from the Fermi level

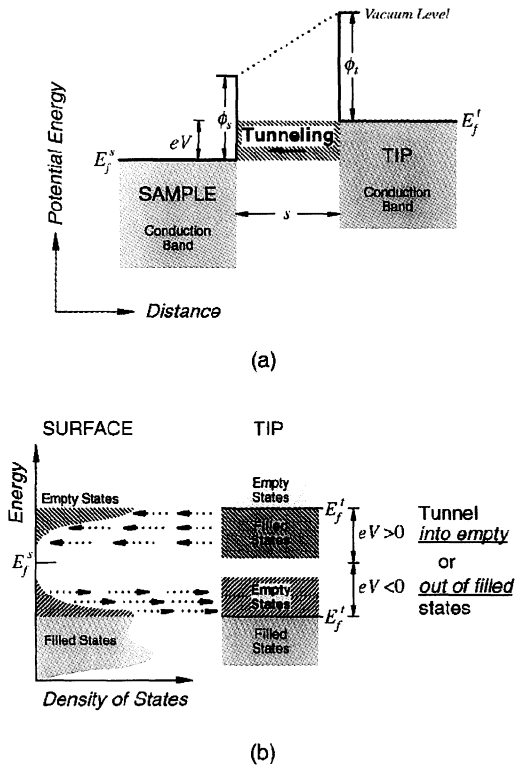

FIG. 7. (a) Schematic of quantum-mechanical tunneling between a metal tip and metal surface a distance s apart. The tunneling barrier is determined by the work functions of the surface and tip, ϕ_s and ϕ_t, respectively, and the bias voltage applied between the two, V. For the case shown, surface positive with respect to the tip, electrons tunnel from bands below the Fermi level of the tip (E_F^t) to those above the Fermi level of the surface (E_F^s). (b) Detailed schematic of the role of the density of states near E_F on the tunneling process both into and out of empty and filled states, respectively, on a semiconductor surface.

(E_F) to the vacuum, prevents electron transport between the two electrodes. Quantum-mechanical tunneling, however, can occur as a result of the overlap of electron wave functions that span the gap. If the potential-energy barrier is treated as a simple trapezoid, the probability for an electron on the tip at energy E with respect to E_F^t to tunnel to the surface can be approximated as

$$T(s,E,eV) \approx \exp\left\{-2s\left[\frac{2m}{\hbar^2}\left(\bar{\phi} - E - \frac{eV}{2}\right)\right]^{1/2}\right\}, \tag{1}$$

where $\bar{\phi}$ is the average work function of the tip and sample, m is the electron mass, e is the electron charge, and $\hbar$ is Planck's constant. For small bias voltages, the probability of tunneling for tip electrons near E_F^t can be further approximated as

$$T(s) \approx \exp(-2\kappa s), \tag{2}$$

where κ is the vacuum decay constant, $(2m\bar{\phi}/\hbar^2)^{1/2}$, about 10 nm^{-1} for typical work functions.

2.2 Density-of-States Effects

The net tunneling current depends on the total number of sample and tip–surface electronic states that overlap within the energy range for which tunneling can occur, as schematically shown in Fig. 7(b). When a positive bias voltage is applied to the sample (or conversely, a negative voltage is applied to the tip), electrons in occupied states on the tip can elastically tunnel into the energy-equivalent empty states in the conduction band on the sample. Similarly, a negative sample bias allows electrons occupying filled states on the surface to tunnel into empty states on the tip. The probability for each of these tunneling events is given by Eq. (1). Mathematically, the net tunneling current can be approximated as

$$I \propto \int_{-eV}^{0} \rho_t(E)\rho_s(eV + E)T(E,eV)dE, \tag{3}$$

where ρ_s and ρ_t are the energy-dependent densities of states for the sample and tip, respectively. For metals (including the tip), ρ is approximately constant near the Fermi level, with states being uniformly filled below E_F and empty above. On semiconductor surfaces ρ may be a strongly varying function of energy (possibly including energies where it is near zero as depicted in Fig. 7), thereby dominating the voltage dependence of the tunneling current.

2.3 Imaging Modes

The primary operational mode of STM is *constant-current imaging,* where the height of the tip is recorded as it is rastered over the surface while the tunneling current is kept constant. When tunneling occurs from sample to tip, the result is a *filled-state* topographic image; tunneling from tip to sample provides an *empty-state* image. On a perfectly crystalline surface the corrugation of the topography from atom to atom (i.e., peak-to-valley height) typically ranges from 0.001 to 0.01 nm on metals and 0.01 to 0.1 nm on semiconductors. If the density of states is relatively uniform across a surface, and modulated only by the location of surface atoms (as is usually the case for metals), the topography will represent a simple contour of constant height of the tip above the surface atoms. Frequently, however, $\rho_s(E)$ varies across the surface from atom to atom, and at some energies it may not even be a maximum over the atom cores. In such cases the images will look different at different bias voltages (the topographic maxima may even move), making interpretation difficult. The density-of-states dependence [Eq. (3)] leads to a central tenet of STM: *Constant-current topography measures (at best) the surface of constant integrated density of states, a surface that is generally bias dependent and may not correlate with the positions of the surface atoms.* This tenet also reveals another limitation of STM: *It is not directly sensitive to the chemical identity of surface atoms.* Fortunately, by complementing topographic images with spectroscopic information (as described below) and the results of other surface characterization techniques, it is often possible to infer the chemical composition of a surface. Furthermore, these limitations chiefly affect atomic-resolution studies of single-crystal surfaces and are not usually of consequence when topography on the nanometer scale is of primary interest.

A second common method of acquiring STM images is known as *current imaging.* In this mode, images are acquired by recording the tunneling current while the tip is rastered at constant height. The result is an image of the variation of the integrated density of states at a fixed height above the surface atoms. This method can only be used on very flat surfaces where the roughness is less than the tunneling gap but has the advantage of being very fast because only minimal feedback is required.

Another standard data acquisition mode is *multi-bias imaging,* where multiple constant-current topographs are recorded simultaneously with different bias voltages. Such images, typically recorded by repeating each scan line at multiple biases before advancing to the next line, reveal the spatial and energy dependence of the integrated density of states at the surface and are often useful for interpreting atomic-scale features. A number of examples can be found in Sec. 3.1.

2.4 Tunneling Spectroscopy

By measuring the dependence of the tunneling current on the tip–surface gap or the bias voltage, spectra containing information about the work function or density of states on the surface can be obtained. Furthermore, because tunneling only occurs over an area <1 nm^2 on the surface, such spectra reveal these properties on an atomic scale. The average work function can be extracted by measuring I as a function of s at fixed V. From Eqs. (2) and (3) we find that

$$\bar{\phi} \approx \frac{\hbar^2}{8m}\left(\frac{d\ln I}{ds}\right)^2. \tag{4}$$

The I-s measurements, sometimes referred to as barrier-height spectra, can be performed periodically over the surface to create an image of the average work function.

Scanning tunneling spectroscopy most often refers to a second type of measurement: the dependence of I on V at fixed s. This measurement is usually made by temporarily disabling the feedback, holding the tip in place, and ramping V while measuring I. I-V spectra reveal features of the local density of states on the surface, as expressed in Eq. (3). Because the exponential dependence of the tunneling probability on V [Eq. (2)] may conceal features in the spectra associated with variations of the density of states, the differential conductance, dI/dV, is often displayed instead of I. If we neglect any role of the tip density of states,

$$dI/dV \propto \rho_s(eV)T(eV). \tag{5}$$

The effects of $T(eV)$ can be further compensated for by normalizing dI/dV to the total conductance, I/V, revealing most clearly the features associated with the surface density of states:

$$\frac{dI/dV}{I/V} \propto \rho_s(eV). \tag{6}$$

I-V spectra can be made at every point in a topographic image to create a set of spectroscopy images that show the spatial variation of the density of states as a function of voltage. The interpretation of specific *I-V* spectral features is usually made by comparison with analogous photoemission and inverse photoemission spectra (see PHOTOEMISSION AND PHOTOELECTRON SPECTRA).

3. APPLICATIONS

One of the more remarkable aspects of STM is the wide range of environments in which it can operate: a stable tunnel current can be maintained in almost any nonconducting medium, including air, liquid, or vacuum. STM operation is also relatively forgiving when it comes to sample preparation: the main requirement is that the sample conduct ~1 nA at ~1 V. This flexibility allows a wide range of applications. Because of the inherent surface sensitivity of STM, it is most widely applied in the field of surface science—the study of the structural, electronic, and chemical properties of surfaces, interfaces, and thin films—a field important to a wide range of technologies, including catalysis, semiconductor device fabrication, electrochemistry, and chemical sensors. The technique is especially useful for elucidating the properties of nanometer-sized surface structures. The operating flexibility of STM combined with the ability to acquire spectroscopic images has also led to its application in the study of novel electronic properties of materials, such as charge-density waves and superconductivity. Finally, the close proximity of the tip to the surface enables one to modify surfaces with atomic-scale precision.

Although STM images and spectra can be recorded under a variety of conditions, the sensitivity of the tunneling current to any perturbation, such as the adsorption of a contaminant atom or molecule on the tip or sample, or mechanical vibration, is such that the best results are obtained with carefully prepared tips and surfaces within the controlled environment of a UHV chamber, electrochemical cell, or helium-filled cryogenic refrigerator. With the aim of best illustrating the unique microscopic and spectroscopic capabilities of STM, most of the applications chosen for this section were obtained under such optimal conditions.

3.1 Structure of Crystal Surfaces

The surface of a crystal can be described by its "Miller indices" (*hkl*). For a cubic crystal, the indices define a vector perpendicular to the surface in terms of the axes of the crystal lattice, $h\hat{\mathbf{x}} + k\hat{\mathbf{y}} + l\hat{\mathbf{z}}$. For example, a (111) surface is perpendicular to the long diagonal of the cube. On account of the loss of three-dimensional periodicity on a surface, the exposed atoms on most crystal faces move from their bulk lattice positions, a rearrangement referred to as surface reconstruction. The reconstruction is described by additional notation, $(m \times n)$, that refers to the size of the *unit cell,* which is the periodic structural unit on the surface, in relation to that of the bulk. For example, a (2×1) reconstruction on a cubic (001) surface would have a unit cell twice as large in the x direction and of equal size in y as on the bulk-terminated surface.

The ability to obtain real-space, atomic-resolution images of single-crystal surfaces has revolutionized the study of surface structure, particularly on surfaces with complex unit cells or point defects. This advance has had a major impact in surface science because determining the structure of clean surfaces is the first step to understanding their electronic and chemical properties. Note that whereas on metal surfaces the interpretation of STM topography is relatively straightforward, on other surfaces it is often necessary to complement STM results with other surface characterization techniques or theoretical calculations to assign topographic features properly to surface structures.

3.1.1 Metals One of the more dramatic illustrations of surface reconstruction is the so-called "herringbone" $(23 \times \sqrt{3})$ structure on Au(111), images of which are shown in Fig. 8. STM was instrumental in the determination of the complex structure of this large-scale reconstruction (Wöll *et al.,* 1989), which entails a contraction of the lattice at the surface such that there are 23 atoms for every 22 bulklike lattice sites. This contraction creates a herringbone-like superstructure of partial dislocation ridges [raised areas in Fig. 8(a)] where surface atoms

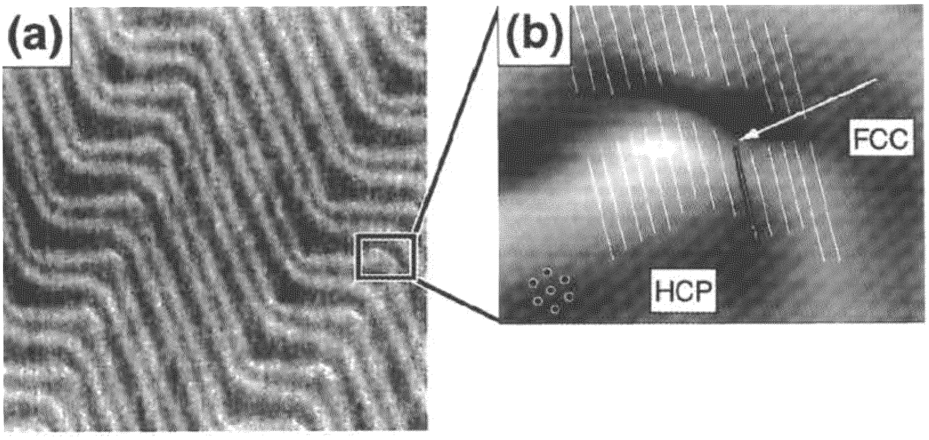

FIG. 8. Filled-state, constant-current STM images of the $(23 \times \sqrt{3})$ reconstruction on Au(111) in UHV. (a) 54 × 54 nm view showing the "herringbone" pattern (0.15-V bias, 0.5-nA current, 0.02-nm height range) (Altman and Colton, 1992, reprinted with permission of Elsevier Science—NL). (b) Atomic-resolution close-up, 6.6 × 5.2 nm, of the bulged herringbone corner where the transition between FCC and HCP stacking occurs (0.9 V, 0.5 nA) (Stroscio *et al.,* 1991). The approximate locations of some surface atoms are marked with dots; the change in height from one to the next is about 3 pm. White lines follow rows of surface atoms across the herringbone corner, highlighting the additional row on the HCP side (indicated by the arrow).

switch back and forth between their bulk positions [face-centered cubic (fcc)] and hexagonal close-packed (hcp) lattice sites. At each "knee" in the superstructure there is a surface dislocation (where the 23rd row of atoms ends), as seen with atomic resolution in Fig. 8(b). This example highlights the extreme sensitivity of STM: the surface atoms, less than 0.3 nm apart, are visible as peaks in the topography only 3 pm high.

3.1.2 Semiconductors Because of the covalent nature of the chemical bonds in a semiconductor, states within a few electron-volts of the Fermi level tend to be more spatially localized than on metal surfaces. Therefore, atomic-resolution STM images of semiconductor surfaces can be loosely described as images of the interatomic bonds and "dangling bonds" at the surface. This localization of the electronic states leads to somewhat larger corrugations than on metals, making atomic-resolution images relatively easier to acquire on semiconductor surfaces. However, because the states may not be localized around the atom cores and may be distributed differently for filled versus empty states, determining the structure from STM images is more complicated.

The local nature of the electronic states on semiconductor surfaces is very apparent in constant-current images of GaAs(110) (see Gallium Arsenide). The Ga and As atoms at the surface, which would both normally be in the surface plane, move slightly in and out of the plane in a simple (1 × 1) relaxation. The electronic effect of this structural relaxation, as shown in Fig. 9, is that the empty electronic states within a few volts of E_{F} are strongly localized on the surface Ga atoms, whereas the filled states are strongly localized on the As. The net result is that a different half of the surface atoms are seen in each image (as first reported by Feenstra *et al.,* 1987). Note that the surface shown in Fig. 9 also includes a point defect, a single missing As atom, highlighting the unique capability of STM to observe directly such structures and their affect on neighboring atoms (Lengel *et al.,* 1994).

The first surface structure ever observed with atomic resolution with an STM was the (7 × 7)-reconstructed (111) surface of silicon (Binnig *et al.,* 1983). STM images like those displayed in Fig. 10 were the critical pieces of the puzzle allowing the structure of this surface to be determined (Takayanagi *et al.,* 1985). Because constant-current images of Si(111)-(7 × 7) are so eye-catching (and relatively easy to acquire in UHV), they are probably the most widely displayed of all STM images and therefore warrant an expla-

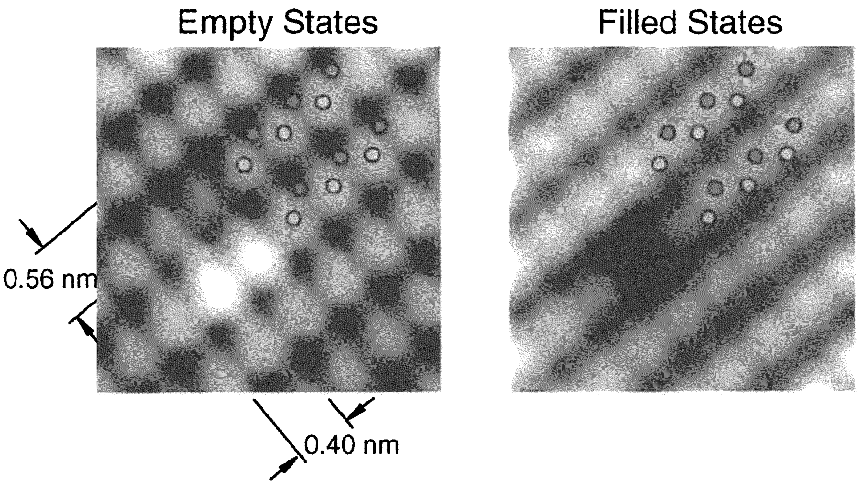

FIG. 9. Atomic-resolution empty- (2.0 V) and filled-state (1.8 V) constant-current images of the same area of a GaAs(110) surface in UHV. The locations of surface Ga atoms (light circles) and As atoms (dark circles) are indicated; the height change from atom-to-atom is about 0.02 nm in the images. The defect observed is a single surface As vacancy (after Lengel *et al.*, 1994).

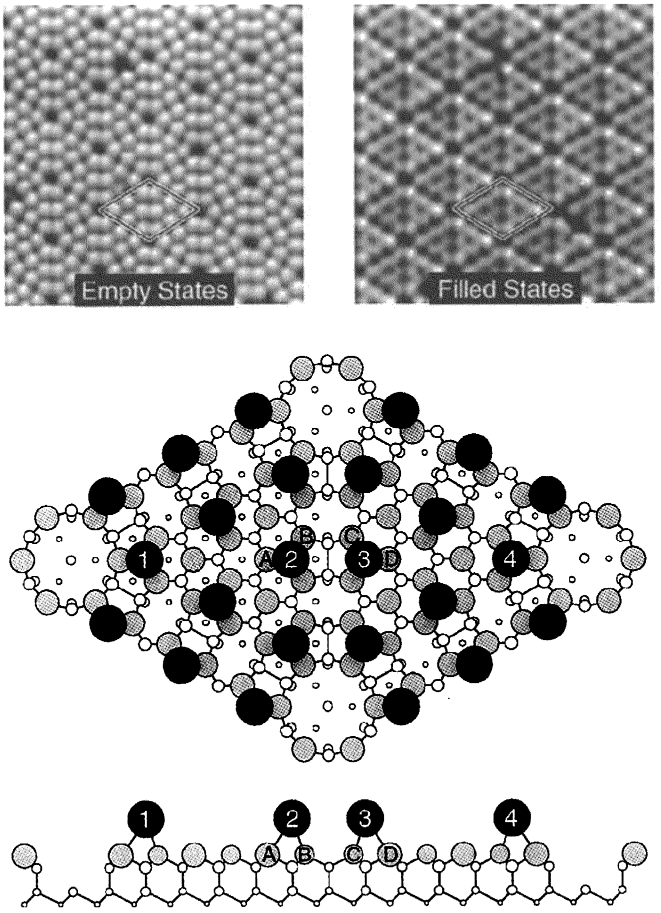

FIG. 10. The (7 × 7) reconstruction of Si(111) in UHV. Both empty- and filled-state constant-current images are shown (14 × 14 nm, ±2.0 V, 0.2-nm height range); one diamond-shaped (7 × 7) unit cell, 4.66 × 2.69 nm, is indicated (courtesy of A. A. Baski and the author). The structure of the unit cell is shown in top and side view (through the long diagonal) with the topmost atoms largest and darkest. Some of the atoms are labeled to aid correlation between the two views (after Takayanagi *et al.*, 1985).

nation. Although the structure observed in the images may appear relatively simple, the actual reconstruction is quite elaborate, involving atoms in the top three layers of the surface (Fig. 10). The topmost atoms, referred to as "adatoms," have one unfilled surface dangling bond each—the only features observed in the empty-state image. The atomic-scale features observed in the filled-state image are associated with the bonds between the adatoms and the second-layer atoms, and partially occupied dangling bonds on the second-layer atoms.

3.1.3 Oxides It is well established that the catalytic properties of oxide materials depend on the structure of their surfaces, including the density and configuration of steps, point defects, and extended crystal defects that intersect the surface. Although oxide materials are generally nonconductive, sometimes surface defects or some intrinsic material property makes them conductive enough for study with STM. For example, an STM image of the (100) surface of the mixed-valence compound $Mo_{18}O_{52}$ is shown in Fig. 11 (Rohrer *et al.*, 1993). The structure of this material can be described as a stack of MoO_3-like slabs shifted slightly from one to the next, with the shifts referred to as crystallographic shear planes. These shear planes are observed directly in the image where they intersect the surface and create a periodic array of steplike structures. Atomic-scale structures on the surface can also be resolved, including features associated with different MoO_x polyhedral groups.

3.2 Adsorption and Film Growth

Most surface-dependent technologies depend on the chemical interaction of a surface with gaseous or liquid reactants, and/or the properties of thin films deposited on a surface. Because surface chemistry is generally heterogeneous, occurring at different rates and by different mechanisms depending on the specific local arrangements of surface atoms, the direct atom-by-atom view afforded by STM makes it a particularly effective tool for studying surface reactions and thin-film growth. By observing a surface as a function of reaction parameters, such as time, temperature, and exposure to reactants, the rates and mechanisms for the individual surface sites and reaction pathways can be studied. Note that although STM can be used to study both atomic and molecular films, interpretation of atomic-resolution images is simpler when the types of chemical species on the surface (and their relative concentrations) are already known.

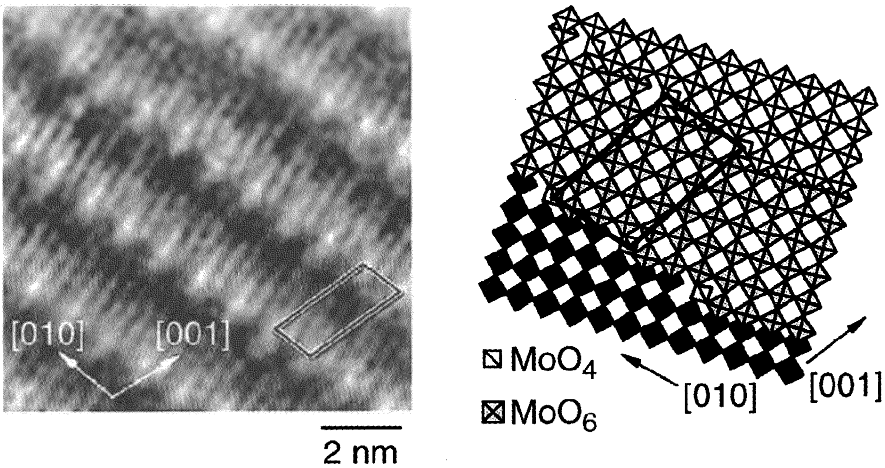

FIG. 11. Filled-state constant-current image of $Mo_{18}O_{52}(100)$ (1.6 V, 0.25-nm height range) and corresponding structural model. The planar repeat unit is indicated by the black box. The model is a polyhedral representation shaded by height (darker is lower). The sample was prepared in air and then imaged in UHV (Rohrer *et al.*, 1993, reprinted with permission of Elsevier Science—NL).

3.2.1 Atom-Resolved Surface Chemistry

The ability of STM to observe surface reactions in real time is demonstrated by the series of images displayed in Fig. 12 (Besenbacher *et al.,* 1994). These frames from an STM "movie" were recorded while a Ni(110) surface at room temperature in UHV was exposed to H_2S gas. Prior to exposure, the surface was covered with oxygen, forming a (2 × 1) reconstruction consisting of rows of alternating Ni and O atoms; only features associated with the Ni atoms are seen in the STM image [Fig. 12(a)]. The H_2S reacts with the O atoms to form H_2O gas, leaving S behind on the surface. As O atoms are removed at the start of the reaction, the neighboring Ni atoms accumulate in small S-covered islands, creating S-covered holes nearby [Figs. 12(b)–12(d)]. As the reaction proceeds, these small islands suddenly become unstable and are replaced by isolated atomic-width rows [Fig. 12(e)]. Eventually the surface becomes completely covered with such rows, forming an ordered S-induced (4 × 1) reconstruction, one that will not form at room temperature in the absence of O. The STM re-

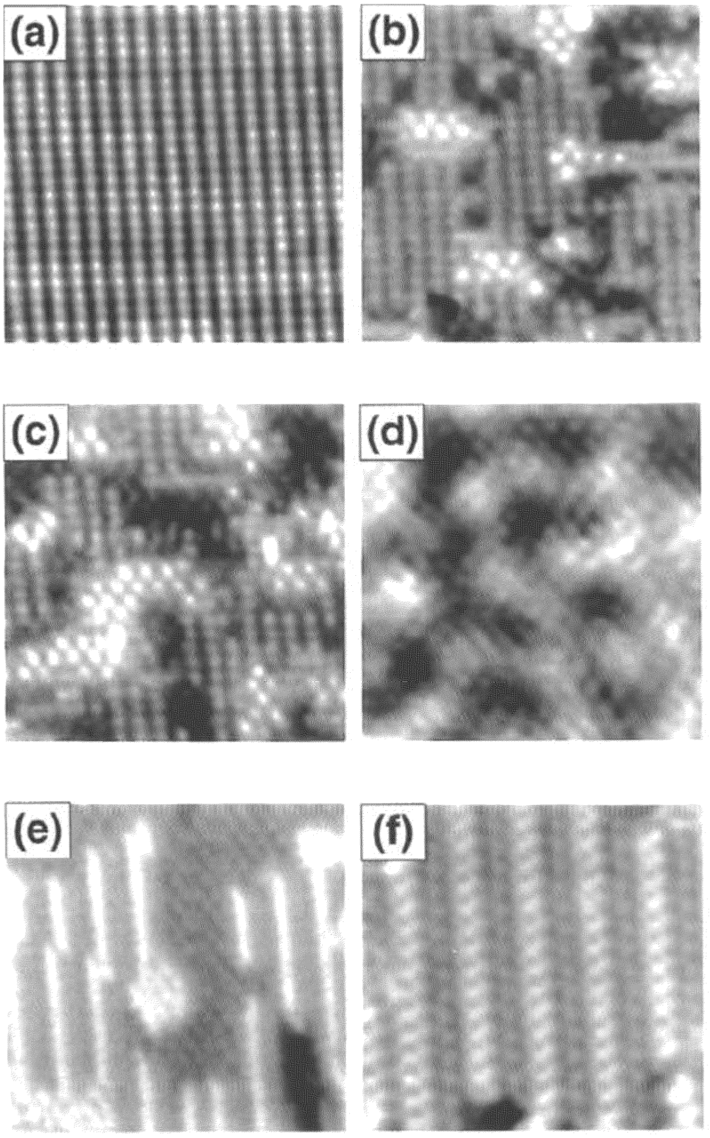

FIG. 12. Filled-state images recorded during the room-temperature reaction of H_2S with oxygen-covered Ni(110) in UHV; $H_2S_{(gas)} + O_{(surface)} \rightarrow H_2O_{(gas)} + S_{(surface)}$. (a) 0, (b) 4.0, (c) 10.6, (d) 26.6, (e) 33.3, and (f) 46.6 mbar·sec. All the images are about 9.5 × 9.5 nm, except (f), which is 7.5 × 7.5 nm (Besenbacher *et al.,* 1994, reprinted with permission of the American Vacuum Society).

sults revealed that the very small islands produced by the local removal of O create especially reactive sites for S adsorption, thereby facilitating the accumulation of S required to form the (4 × 1) structure.

The heterogeneous nature of surface reactions is highlighted by STM studies of the oxidation of Si(111)-(7 × 7), where the complex reconstruction (Fig. 10) creates a variety of inequivalent adsorption sites. As shown by the empty-state images in Fig. 13 (Martel *et al.*, 1996), when a clean Si(111) surface at room temperature is exposed to O_2 gas in UHV, the oxygen initially reacts almost exclusively with the atoms at the corners of the diamond-shaped unit cells (known as "corner hole adatoms"). The reaction results in two types of features at the corner holes, referred to as "bright" (B) and "dark" (D) sites. In addition, dark features are observed both individually and in pairs (DD). The STM images, when combined with the results of spectroscopy and other characterization techniques, revealed that the corner holes are preferred reaction sites by virtue of their higher density of states near E_F and that the reaction produces three types of species: an oxidized silicon adatom (B), an O_2 molecule bonded to an oxidized adatom (D), and an O_2 molecule bonded to a second-layer atom and interacting with two adjacent adatoms (DD). Identification of all three species is an important step toward attaining a complete understanding of the oxidation reaction.

3.2.2 Thin-Film Morphology Just as the reaction of a surface with a gas can be followed via STM, the morphology of deposited films can be monitored during growth. Film growth involves a host of processes, including adsorption, diffusion across the surface, nucleation of islands, and sticking to island or terrace edges; studies of how the morphology depends on growth conditions elucidate these reactions. For example, in Fig. 14 the temperature-dependent film morphology observed with a variable-temperature STM is shown for a tenth of a layer of Ag deposited on Pt(111) in UHV (Röder *et al.*, 1993). At 80 K the atoms migrate short distances upon deposition, forming small Y-shaped clusters (with an average of 19 atoms in each). As the temperature is increased, faster diffusion allows more of the deposited atoms to reach a Pt surface step where they become trapped, producing the morphology observed at 130 K. At this temperature migration of atoms along island edges is still hindered, however, giving the islands a jagged shape. At 170 K the temperature-dependent mobility is large enough to allow almost all nucleation to occur at step edges, which become noticeably smoother. Finally, at 300 K the film grows as round-edged is-

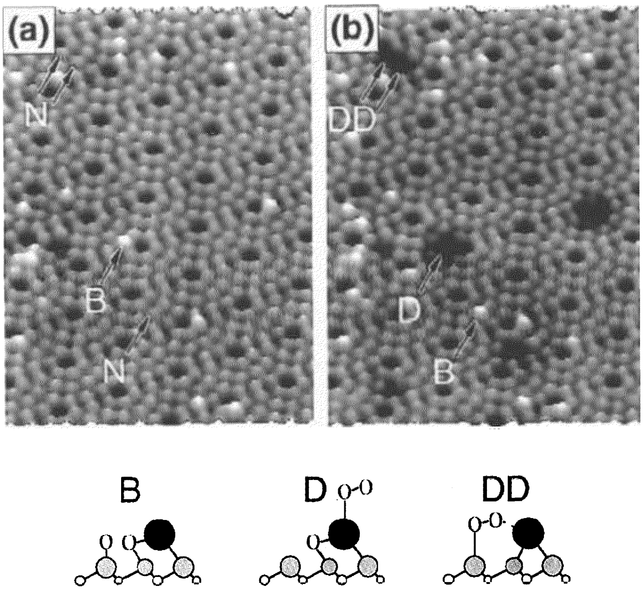

FIG. 13. Empty-state images (14 × 21.5 nm) of a Si(111)-(7 × 7) surface during the early stages of oxidation in UHV: after exposure to (a) 0.067 and (b) 0.133 mbar·sec of O_2. The arrows point to typical reaction sites: not reacted (N), bright (B), dark (D), and two adjacent dark sites (DD). Structural models for the reacted sites are also shown (after Martel *et al.*, 1996).

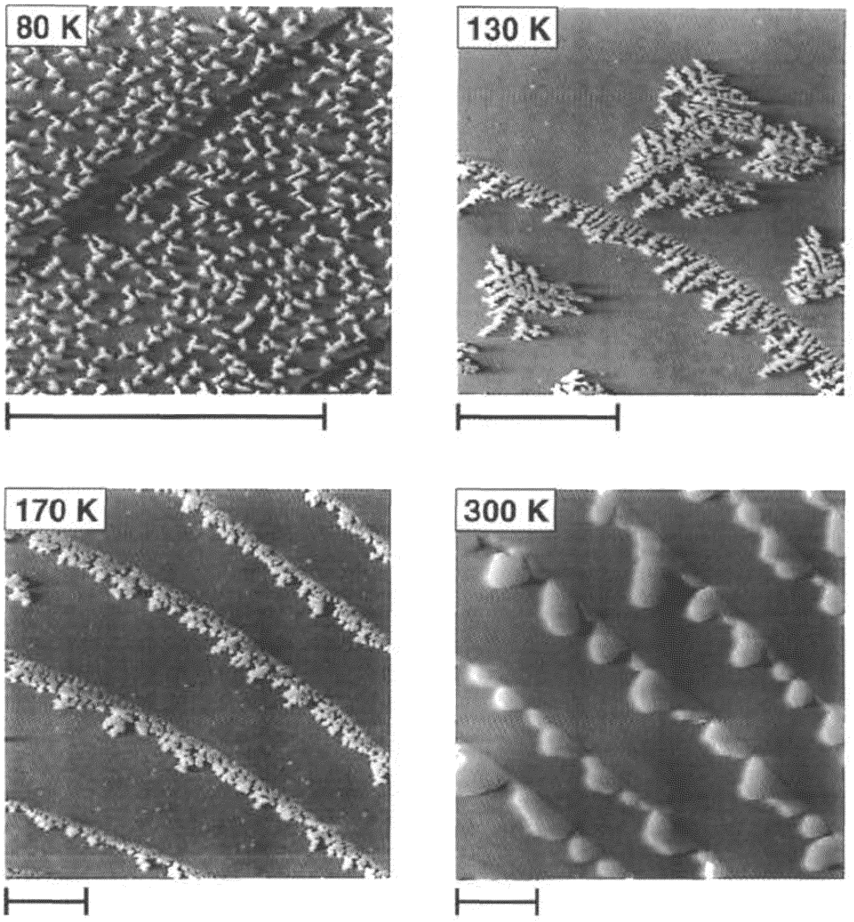

FIG. 14. Filled-state variable-temperature STM images of the growth of 0.1 layer of Ag on Pt(111) in UHV. The scale bars are all 50 nm, and the images are gray-scales of the *derivative* of the topography (which highlights edges) (Röder *et al.*, 1993, reprinted with permission of Elsevier Science—NL).

lands flowing out from the step edges, close to the equilibrium morphology. These STM results, combined with other images recorded as a function of film thickness, led to a detailed understanding of the atomic-scale kinetics and mechanisms of Ag growth on Pt(111), experimentally confirming general theories about nucleation and growth (Brune *et al.*, 1994).

In addition to furthering our general understanding of growth processes, STM can be used more simply as a direct method of characterizing the nanometer-scale morphology of multilayer films. This is an important application given that many thin-film–based technologies require films to be as nearly perfect as possible for optimum performance. The type of information attainable is illustrated in Fig. 15, where an STM image of a 4-μm-thick GaSb (gallium antimonide) film grown on a GaAs(001) substrate in UHV is displayed. The film was deposited by molecular-beam epitaxy (*q.v.*) for possible use in an electronic device (Thibado *et al.*, 1996). The surface consists of spiral-like mounds approximately 5 nm high, with each "step" on the surface a single atomic layer (about 0.3 nm). The striking morphology results from the crystal lattice of GaSb being 7% larger than that of GaAs, which causes crystal dislocations to form in the deposited films. These dislocations, which are detrimental to electronic devices, thread up through the film and emerge at the surface to produce a step. This type of characterization allows the effects of various growth con-

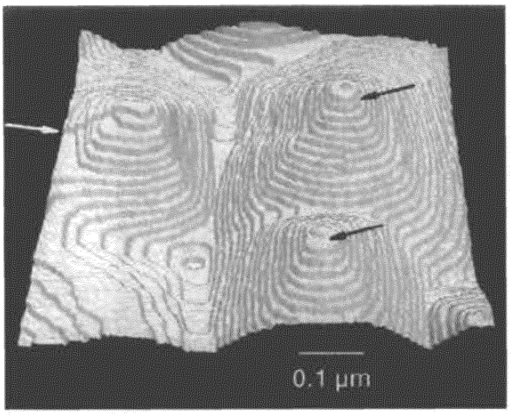

FIG. 15. The topography of a 4-μm-thick GaSb film grown on a GaAs(001) substrate by molecular-beam epitaxy in UHV. Each step in the constant-current image is about 0.3 nm high. The arrows point out where threading dislocations emerge at the surface and create steps. (Courtesy of P. M. Thibado, B. R. Bennett, B. V. Shanabrook, and the author.)

ditions to be directly observed, thereby simplifying the effort to improve the film quality.

3.3 Electrochemistry

Electrochemical technologies rely on processes that occur at the solid–liquid interface. Whereas in air and vacuum one can control the type and concentrations of reactants along with the substrate temperature; in an electrochemical environment an additional parameter can be varied: the electrochemical potential (see ELECTROCHEMISTRY). Control of this potential allows one to vary the thermodynamics of electrochemical reactions, including both reversible and irreversible processes such as oxidation, reduction, deposition, and changes in surface reconstruction; STM allows the atomic-scale nature of these processes to be directly observed *in situ*. One such process is exhibited in the series of STM images in Fig. 16, recorded during the oxidation of the naturally occurring semiconductor galena (PbS) in a solution of $NaClO_4$ (Higgins and Hamers, 1995). The reaction, $PbS \rightarrow Pb^{2+} + S + 2e^-$, is seen to occur almost exclusively via removal of atoms at step edges. Furthermore, impurities (I) are observed to cause etch pits (P) in the surface. Combined with STM images recorded during electrochemical reduction cycles, the complex surface electrochemistry of galena was elucidated on the atomic scale for the first time.

3.4 Surface Electronic Structure

Although some effects related to surface electronic structure were indirectly revealed in the applications discussed above, atomic-scale electronic properties can be more directly studied using the various imaging and spectroscopy modes described in Sec. 2. In this way the local electronic properties (including work function, conductivity, and density of states) of clean surfaces, point defects, isolated adsorbed atoms and molecules, nanometer-scale islands, and thin films can be observed. (See also ELECTRONIC STRUCTURE OF SURFACES.)

3.4.1 Conductance versus Structure One of the simplest and most direct applications of tunneling spectroscopy is the examination of the local conductance as revealed by the shape of *I-V* spectra recorded over different surface structures. A nice illustration of this application is the study of the conductance of different Cs structures observed on GaAs(110) in UHV (Whitman *et al.*, 1991). As Cs is deposited on the surface at room temperature, it first forms one-dimensional (1D) atomic chains, then a 2D array of planar clusters, and finally a disordered film of 3D clusters (images not shown). As displayed in Fig. 17, *I-V* spectra recorded on clean GaAs(110) have the expected tunnel-diode shape, with zero current (and zero differential conductance) within the semiconductor bulk band gap. Spectra recorded over the Cs structures, however, show that the local "band gap" decreases as the structures change from 1D to 3D. The 3D structures appear metallic, i.e., there is nonzero conductance at E_F (zero sample voltage). These spectroscopy results furthered our understanding of the properties of nanometer-scale metal-on-semiconductor structures.

3.4.2 Chemically Selective Imaging Local variations in the density of states that are associated with particular surface atoms or molecules can sometimes be exploited to provide atom-selective topographic contrast.

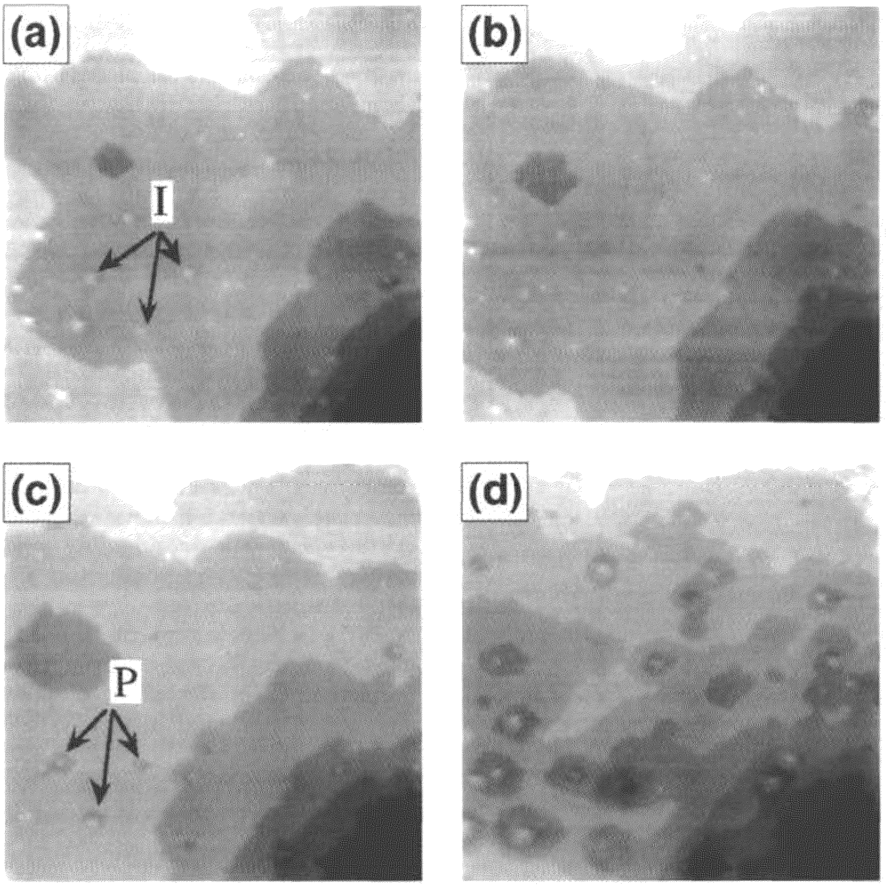

FIG. 16. A selection of constant-current images of PbS(001) acquired during electrochemical oxidation in $NaClO_4$. All images are 200 × 200 nm, with each successive gray level corresponding to a change in height of 0.3 nm. The images were recorded with a tunneling current of 0.17 nA, tip voltage of 0.94 V, and a sample voltage that changed from (a) 0.20 to 0.24, (b) 0.29 to 0.33, (c) 0.38 to 0.43, and (d) 0.38 to 0.34 V during image acquisition. Some surface impurities (I) and the resulting etch pits (P) are labeled (Higgins and Hamers, 1995, reprinted with permission of Elsevier Science—NL).

An extreme example is the (110) surface of GaAs discussed in Sec. 3.1.2 (Fig. 9), where only Ga (As) features are seen in empty (filled) states. A subtler, more elegant example is found in a study of Cr films deposited on Fe(001) (Davies *et al.*, 1996), the results of which are presented in Fig. 18. Metal "sandwiches" composed of alternating thin layers of ferromagnetic (e.g., Fe) and nonmagnetic (e.g., Cr) layers have novel magnetic properties of potential use in magnetic recording technologies. The properties are not always as expected, however, and this STM study revealed a possible cause. Rather than forming a simple overlayer, Cr alloys with the substrate, with Cr atoms randomly embedding into the surface. The Fe atoms displaced by this process intermix with additional Cr atoms to form isolated single-layer islands. Tunneling spectra recorded over the embedded Cr atoms differ from Fe spectra, with an extra peak in the conductance about 0.3 V below E_F; this extra conductance causes a small bump (~0.01 nm high) over every Cr atom in the constant-current topography recorded at negative sample bias. That spectral feature and the associated atom-selective images led directly to an understanding of the Cr/Fe interface structure.

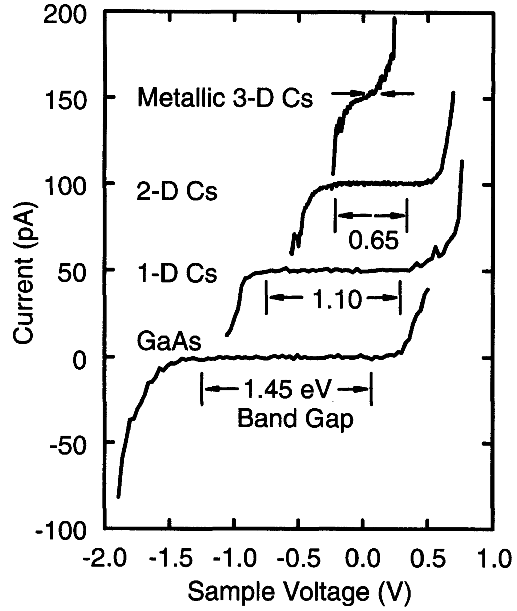

FIG. 17. Current-versus-voltage tunneling spectra recorded on clean and Cs-covered GaAs(110) in UHV. With increasing Cs coverage long 1D chains, then 2D planar clusters, and finally a disordered 3D overlayer are observed (images not shown). The apparent "band gap" observed for each structure is indicated (Whitman *et al.*, 1991).

3.5 Bulk Electronic Structure

The different chemical environment at a surface with respect to that in the bulk usually results in distinct structural and electronic properties. Exceptions to this situation occur in materials with strong covalent bonding in 2D layers and weak noncovalent bonds holding the layers together: When the weak bonds are broken to expose a surface, the surface atoms do not differ greatly from those at the interlayer interfaces, and STM can therefore be used to observe bulk properties. This is very fortuitous because the 2D nature of such materials often gives them scientifically fascinating properties. For instance, one important class of these layered materials, transition-metal dichalcogenides, exhibit both superconducting and charge-density-wave (CDW) phases (see COLLECTIVE PHENOMENA IN SOLIDS).

For superconducting phases, STM can be used to directly study spatial variations in the energy gap, the effect of point defects and impurities, and magnetic-field effects. Because charge-density waves are periodic variations of the charge density near E_F (caused by electron–phonon coupling), they

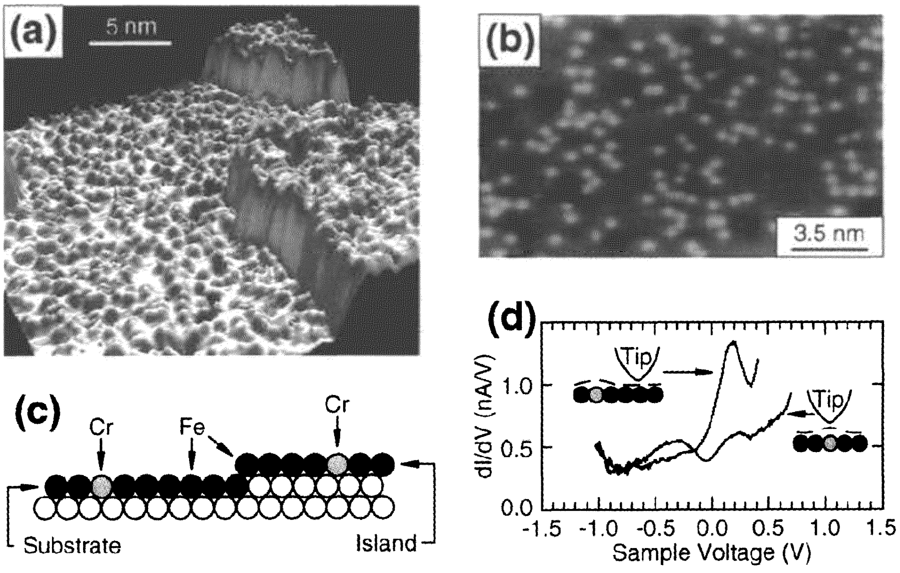

FIG. 18. (a) Filled-state image (1.1 V) of 0.4 layer of Cr deposited on Fe(100) at 290 C in UHV. Each step is 0.14 nm high, and the small bumps in the topography are approximately 0.01 nm high. (b) A higher-resolution view of the Cr-induced bumps. (c) A structural model for the Cr film. The Cr atoms are embedded in the substrate, with displaced Fe atoms forming single-layer islands. (d) Conductance spectra recorded over a Cr or Fe surface atom (after Davies *et al.*, 1996).

can also be directly imaged with STM, as shown in Fig. 19 for TaS_2 (Liu *et al.*, 1996). On pure TaS_2 crystals the CDW is observed as a large hexagonal corrugation (0.2–0.4 nm) with a period distinctly different from the smaller corrugation (<0.1 nm) associated with the underlying crystal lattice. Bulk impurity atoms (Nb) introduce dislocations and other forms of local disorder into the CDW but do not affect the long-range orientational order. More detailed image analysis as a function of impurity concentration showed that the CDW lattice evolves from a crystalline state to an amorphous liquidlike state with increasing concentration—a significant discovery about the nature of this material.

3.6 Atom Manipulation and Nanolithography

The nanometer-scale nature of the tip–surface interaction combined with the unprecedented degree of position control available with the STM make it a revolutionary instrument for surface modification. There are four fundamental processes inherent to surface modification: removal of surface atoms, deposition of new material, lateral translation of adsorbed atoms, and induction of surface chemical reactions. All of these processes can be actuated on the atomic scale via the three basic control parameters in an STM: the tip properties, including the material composition and shape of the tip; the tunneling current; and the electric field in the tunnel junction (a function of the gap and the bias voltage). Depending on the degree of control and the intensity of the tip–surface interactions, modifications can be made from the micron to the atomic scale. Although practical applications of such "nanolithography" have yet to be realized, the STM is a new tool with which to explore the properties of nanometer-scale structures and the limits of physical lithography.

3.6.1 Electron-Stimulated Patterning

The needlelike shape of etched STM tips provides extremely effective focusing of the emitted electrons, in both tunneling and field-emission modes. Even though the total current is very small, this focusing makes the current density quite large: 10^{11}–10^{12} times higher than a conventional electron beam with the same current. Combined with the large electric field in the tunnel junction ($\sim 10^7$–10^8 V/cm), that current density makes the STM an efficient tool for breaking surface chemical bonds. Figure 20 demonstrates this phenomenon as applied to lithography on Si(001) (Lyding *et al.*, 1994), the substrate on which the semiconductor industry is based. The surface was first covered with atomic hydrogen, which reacts with the surface Si atoms, passivating their dangling bonds and making the surface unreactive to most other gases. Raising the sample bias voltage above ~4 V gives the electrons enough energy to break the H–Si bond and

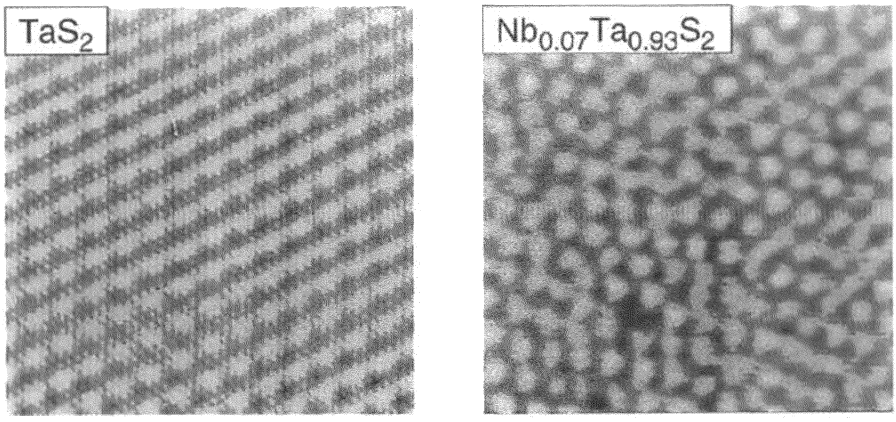

FIG. 19. Constant-current filled-state images (15 × 15 nm, 50 mV, 1 nA) of $Nb_xTa_{1-x}S_2$ crystals recorded in a nitrogen-filled glove box at 380 ($x = 0$) and 315 K ($x = 0.07$). The large-period corrugation is due to a charge-density wave that is incommensurate with the crystal lattice (the smaller, shorter-period corrugation) (Liu *et al.*, 1996, reprinted with permission of the American Vacuum Society).

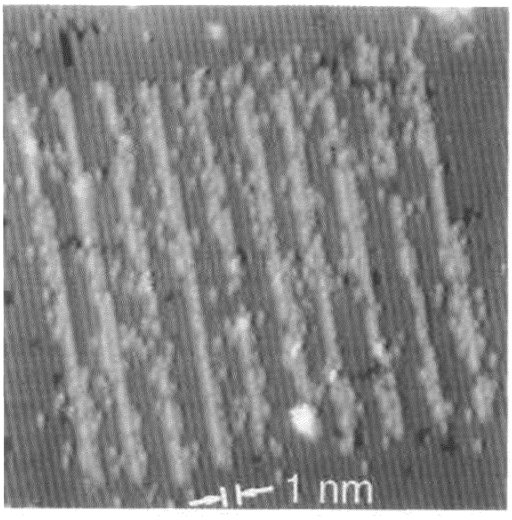

FIG. 20. Filled-state image (37.5 × 37.5 nm) of nanometer-scale lithography of Si(001) − (2 × 1) with an STM in UHV. The surface was first covered with atomic hydrogen, forming (2 × 1)-reconstructed rows of H–Si = Si–H (the dimmer rows visible in the image). The H was then locally removed (bright rows) by scanning the desired areas with a high sample bias and current (4.5 V, 2.0 nA) (Lyding *et al.*, 1994, reprinted with permission of the American Institute of Physics). The lines are ≤1 nm wide on a 3-nm pitch.

thereby locally remove H from the surface. (The H-free areas appear brighter in filled-state images because of a larger density of states on the bare Si atoms.) This removal can be confined to lines only one or two Si atoms wide. When the patterned passivated surface is exposed to a reactant such as O_2, only the H-free areas react (not shown).

3.6.2 Positioning Single Atoms Perhaps the most dramatic use of an STM has been to perform the ultimate surface modification: the controlled manipulation of individual surface atoms (Eigler and Schweizer, 1990). Using an STM operating in UHV at a cryogenic temperature (4 K), IBM scientists have precisely positioned adsorbed atoms by carefully placing the tip in contact with them one at a time and gently sliding them across the surface. (The atoms are otherwise immobile at this low temperature.) A sequence of images recorded during the assembly of a ring of 48 Fe atoms on Cu(111) is displayed in Fig. 21. When the first image was recorded, one-quarter of the ring had already been assembled and the additional Fe atoms required positioned nearby. A striking consequence of the construction of this "quantum corral" is the formation of electron standing waves within the ring, observable with the STM through its sensitivity to charge-density modulations (Crommie *et al.*, 1993). These waves are produced when surface electrons are reflected by the Fe atoms within the ring and interfere with themselves—a direct observation of the wave nature of electrons as described by quantum mechanics.

4. RELATED TECHNIQUES

The basic elements of the STM, a sharp probe held in close proximity to a sample surface with precise position control, have evolved into a remarkable number of related scanning-probe microscopies capable of measuring an astonishing array of structural, electrical, and chemical properties, often with nanometer-scale spatial resolution (see, for example, Wiesendanger, 1994). Four of the most common of these techniques, all commercially available, are briefly described here.

Atomic-force microscopy (AFM), the simplest extension of STM, allows the study of nonconductive samples by using the force of the tip–surface interaction as the control parameter. In an AFM the tip usually consists of a micron-sized silicon nitride pyramid on the end of a silicon nitride microcantilever a few hundred microns long, fabricated using semiconductor lithography and etching techniques. By maintaining a constant tip–surface force (as measured by the deflection of the cantilever and typically in the range of 10^{-8}–10^{-9} N), the tip can be rastered across a surface to observe topography routinely with ~1 nm resolution. Moreover, a variety of imaging and spectroscopy modes have been developed to measure force-related phenomena, including friction, adhesion, and indentation. In addition to being a powerful research tool, AFM is widely used for routine high-resolution microscopy on account of its ability to image almost any type of material under ambient conditions.

In magnetic-force microscopy (MFM) the spatial variation of the magnetic force interaction between a magnetic tip and sample is

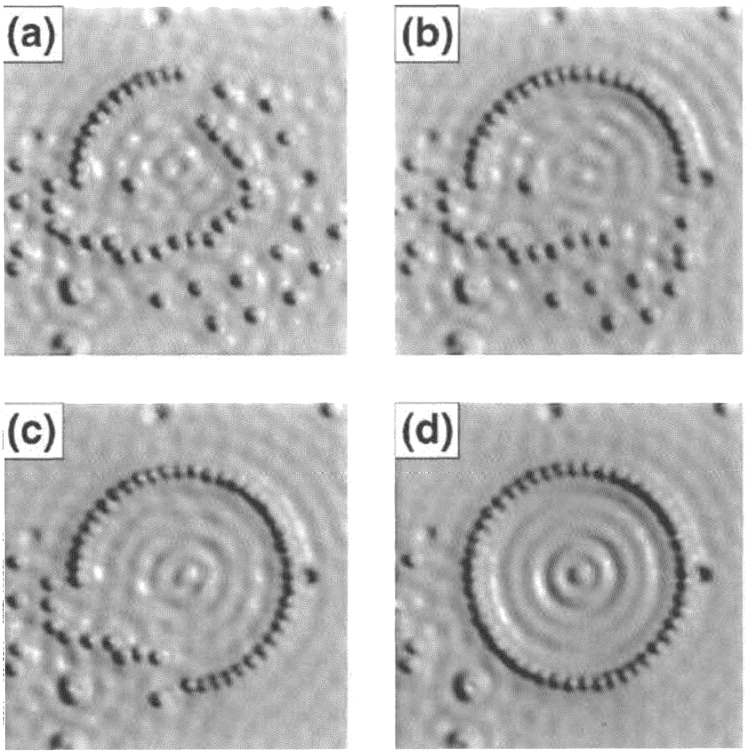

FIG. 21. (a)–(d) Sequential constant-current empty-state images (0.01 V, 1.0 nA) recorded during the assembly of a circle of Fe atoms on Cu(111) at 4 K in UHV. The average diameter of the circle is 14.25 nm. The structure was assembled one atom at a time using the STM tip. The concentric rings seen in the image are variations in the local density of states (not atomic heights) associated with standing waves created by surface electrons trapped inside the circle of Fe atoms (a "quantum corral"). (Courtesy of M. F. Crommie, C. P. Lutz, and D. M. Eigler.)

measured, usually by mounting a magnetic tip on the end of an AFM-style cantilever. In this way, images of the stray magnetic field (or its gradient) above a surface can be acquired with a spatial resolution of 20–100 nm. Because MFM works under ambient conditions and on samples with nonmagnetic coatings or surface contamination, it is particularly useful for characterizing magnetic recording media.

Near-field scanning optical microscopy (NSOM) or scanning near-field optical microscopy (SNOM) replaces the STM tip with an optical fiber drawn to a point of ~100 nm diameter. The fiber is coated with metal everywhere but on the end, thereby defining a nanometer-scale aperture that is held ~10 nm from the sample. The evanescent light either reflected from or transmitted through a sample is then collected with high spatial resolution. Whereas the resolution of conventional optics is limited by diffraction, collecting the light in this "near-field" regime enables a spatial resolution much less than the wavelength of the light. In addition to providing high-resolution optical microscopy, spectroscopy can be performed on the collected light allowing the optical characterization of material defects, nanometer-scale structures, and molecular properties (via fluorescence detection).

Ballistic electron-emission microscopy (BEEM) was developed in order to characterize electron transport through buried interfaces, typically between a thin film and a substrate with different electrical properties (e.g., a metal film on a semiconductor). Whereas STM is a two-electrode technique (the tip and the sample), BEEM adds a third electrode, a current collector on the edge of the thin film. The tip injects electrons into the thin film, almost all of which reach the buried interface without energy loss. Current observed at the collector is therefore a measure of electron scattering at the interface; it can be recorded as a function of the tip position along with an STM image, providing a simultaneous image of the buried interface. Furthermore, by measuring the dependence of the collector current on the sample bias, spectroscopy of interface electronic transport can also be performed.

ACKNOWLEDGMENT

Preparation of this article and the author's work described herein were supported by the U.S. Office of Naval Research.

GLOSSARY

Adsorption: The process whereby an atom or molecule becomes bonded to a surface.

Dangling Bond: An unbonded molecular orbital on a surface atom of a covalent or ionic crystal.

Density of States: The number of electronic levels per unit energy in a material.

Fermi Level (E_F): The energy of the highest occupied electronic state in a material at 0 K.

Miller Indices: Notation (*hkl*) used to describe a surface of a crystal. For a cubic crystal, the indices define a vector perpendicular to the surface, $h\hat{\mathbf{x}} + k\hat{\mathbf{y}} + l\hat{\mathbf{z}}$, where $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are the crystal axes.

Piezoelectric Ceramic: A material that expands or contracts when an electric field is applied across it. (See PIEZOELECTRIC DEVICES.)

Reconstruction: A rearrangement of the atoms on a crystal surface that changes their symmetry with respect to bulk positions. When the rearrangement is periodic it is denoted by the size of the unit cell, $(m \times n)$, with respect to the natural crystal structure, (1×1).

Surface Science: The study of the structural, electronic, and chemical properties of surfaces, interfaces, and thin films. (See AUGER SPECTROSCOPY; CATALYSIS; CHARACTERIZATION AND ANALYSIS OF MATERIALS; CVD (CHEMICAL VAPOR DEPOSITION); ELECTRON DIFFRACTION; ELECTRON MICROSCOPY; ELECTRONIC STRUCTURE OF SURFACES; FIELD ION MICROSCOPY; MOLECULAR BEAM EPITAXY; and PHOTOEMISSION AND PHOTOELECTRON SPECTRA.)

Surface Step: The place on a crystal surface where the height changes by one or more discrete layers of atoms.

Tunneling: The quantum-mechanical process whereby an electron is transported across a classically forbidden energy barrier.

Ultrahigh Vacuum (UHV): A pressure $<1 \times 10^{-7}$ Pa, low enough to prevent contamination of a surface by residual gases for many minutes.

Unit Cell: The periodic structural unit on a crystal surface.

Work Function: The energy required to move an electron at the Fermi level from the surface into the vacuum. (See ELECTRONIC STRUCTURE OF SURFACES.)

Works Cited

Altman, E. I., Colton, R. J. (1992), *Surf. Sci.* **279,** 49–67.

Besenbacher, F., Laegsgaard, E., Nielsen, L. Pleth, Ruan, L., Stensgaard, I. (1994), *J. Vac. Sci. Technol. B* **12,** 1758–1763.

Binnig, G., Smith, D. P. E. (1986), *Rev. Sci. Instrum.* **57,** 1688–1689.

Binnig, G., Rohrer, H., Gerber, Ch., Weibel, E. (1982a), *Appl. Phys. Lett.* **40,** 178–180.

Binnig, G., Rohrer, H., Gerber, Ch., Weibel, E. (1982b), *Physica* **109–110B + C,** 2075–2077.

Binnig, G., Rohrer, H., Gerber, Ch., Weibel, E. (1983), *Phys. Rev. Lett.* **50,** 120–123.

Brune, H., Röder, H., Boragno, C., Kern, K. (1994), *Phys. Rev. Lett.* **73,** 1955–1958.

Crommie, M. F., Lutz, C. P., Eigler, D. M. (1993), *Science* **262,** 218–220.

Davies, A., Stroscio, J. A., Pierce, D. T., Celotta, R. J. (1996), *Phys. Rev. Lett.* **76,** 4175–4178.

Eigler, D. M., Schweizer, E. K. (1990), *Nature* **344,** 524–526.

Feenstra, R. M., Stroscio, J. A., Tersoff, J., Fein, A. P. (1987), *Phys. Rev. Lett.* **58,** 1192–1195.

Higgins, S. R., Hamers, R. J. (1995), *Surf. Sci.* **324,** 263–281.

Lengel, G., Wilkins, R., Brown, G., Weimer, M., Gryko, J., Allen, R. E. (1994), *Phys. Rev. Lett.* **72,** 836–839.

Liu, J., Huang, J.-L., Lieber, C. M. (1996), *J. Vac. Sci. Technol. B* **14,** 1064–1069.

Lyding, J. W, Shen, T.-C., Hubacek, J. S., Tucker, J. R., Abeln, G. C. (1994), *Appl. Phys. Lett.* **64,** 2010–2012.

Martel, R., Avouris, Ph., Lyo, I.-W. (1996), *Science* **272,** 385–388.

Röder, H., Brune, H., Bucher, J., Kern, K. (1993), *Surf. Sci.* **298,** 121–126.

Rohrer, G. S., Lu, W., Smith, R. L., Hutchinson, A. (1993), *Surf. Sci.* **292,** 261–266.

Stroscio, J. A., Pierce, D. T., Dragoset, R. A., First, P. N. (1991), *J. Vac. Sci. Technol. A* **10,** 1981–1985.

Takayanagi, K., Tanishiro, Y., Takahashi, M., Takahashi, S. (1985), *J. Vac. Sci. Technol. A* **3,** 1502–1510.

Thibado, P. M., Bennett, B. R., Twigg, M. E., Shanabrook, B. V., Whitman, L. J. (1996), *J. Vac. Sci. Technol. A* **14,** 885–889.

Whitman, L. J., Stroscio, J. A., Dragoset, R. A., Celotta, R. J. (1991), *Phys. Rev. Lett.* **66,** 1338–1341.

Wiesendanger, R. (1994), *Scanning Probe Microscopy and Spectroscopy,* Cambridge, U.K.: Cambridge Univ. Press.

Wöll, Ch., Chiang, S., Wilson, R. J., Lippel, P. H. (1989), *Phys. Rev. B* **39,** 7988–7991.

Young, R., Ward, J., Scire, F. (1972), *Rev. Sci. Instrum.* **43,** 999–1011.

Further Reading

Binnig, G., Rohrer, H. (1987), *Rev. Mod. Phys.* **59,** 615–625.

Bonnell, D. A. (Ed.) (1993), *Scanning Tunneling Microscopy and Spectroscopy: Theory, Techniques, and Applications,* New York: VCH.

DiNardo, N. J. (1994), *Nanoscale Characterization of Surfaces and Interfaces,* New York: VCH.

Duke, C. (Ed.) (1994), "Surface Science: The First Thirty Years," *Surf. Sci.* **299/300,** 1–1054.

Güntherodt, H.-J., Wiesendanger, R. (1994), *Scanning Tunneling Microscopy I: General Principles and Applications to Clean and Adsorbate-Covered Surfaces,* 2nd ed., Springer Series in Surface Sciences Vol. 20, Berlin: Springer-Verlag.

Kubby, J. A., Boland, J. J. (1996), "Scanning Tunneling Microscopy of Semiconductor Surfaces," *Surf. Sci. Rep.* **26,** 61–204.

Kuk, Y., Silverman, P. J., (1989), *Rev. Sci. Instrum.* **60,** 165–180.

Magonov, S. N. (1996), *Surface Analysis With STM and AFM: Experimental and Theoretical Aspects of Image Analysis,* Weinheim: VCH.

Stroscio, J. A., Kaiser, W. J. (Eds.) (1993), *Scanning Tunneling Microscopy,* Boston: Academic.

Wiesendanger, R., Güntherodt, H.-J. (1995), *Scanning Tunneling Microscopy II: Further Applications and Related Scanning Techniques,* 2nd ed., Springer Series in Surface Sciences Vol. 28, Berlin: Springer-Verlag.

Wiesendanger, R., Güntherodt, H.-J. (1996), *Scanning Tunneling Microscopy III: Theory of STM and Related Scanning Probe Methods,* 2nd ed., Springer Series in Surface Sciences Vol. 29, Berlin: Springer-Verlag.

Zangwill, A. (1988), *Physics at Surfaces,* Cambridge, U.K.: Cambridge Univ. Press.

TURBINES, GAS

FREDRIC EHRICH, *GE Aircraft Engines, Lynn, Massachusetts, U.S.A.*

	Introduction	385
1.	**Physical and Technical Principles**	385
1.1	The Brayton Cycle and the Gas Generator	385
1.2	Energy Conversion and System Efficiency	386
1.3	Cycle Elaboration	387
1.3.1	Regeneration	387
1.3.2	Intercooling	387
1.3.3	Reheat	387
2.	**Design, Function, and Operation**	387
2.1	Key Components	387
2.1.1	Compressors	388
2.1.2	Combustors	389
2.1.3	Turbines	389
2.2	Subsystems	390
3.	**Materials and Manufacture**	390
3.1	Light-Weight Materials	390
3.2	High-Temperature Materials	390
3.3	High-Temperature Cooled Turbine Blades	391
4.	**Applications**	391
4.1	Stationary Power Plants	391
4.2	Vehicular Applications	392
4.3	Marine Applications	392
4.4	Aircraft Applications	392
5.	**Economic and Environmental Aspects**	392
5.1	System and Fuel Economy	392
5.2	Exhaust Emissions	394
5.3	Noise Emissions	394
	Glossary	395
	Works Cited	397
	Further Reading	397

INTRODUCTION

The gas turbine or gas turbine engine is a unique member of the family of heat engines that work on the principle of sequential compression, heat addition, and expansion of a working fluid. The gas turbine engine derives its uniqueness from its use of single-phase gas as working fluid rather than a two-phase liquid/vapor system such as characterizes the steam turbine; its use of steady flow compression and expansion in a turbo-compressor driven by a turbo-expander; and continuous heat addition at essentially constant pressure. The use of steady flow compression, heat addition, and expansion are the special characteristics that distinguish the gas turbine engine from those cycles (such as the Otto, Sterling, and Diesel) that are based on periodic batch processes that accomplish compression, combustion, and expansion in a closed volume such as the piston chamber of a reciprocating engine.

A typical modern gas turbine engine, used for marine propulsion and industrial power applications but, as is frequently the case, actually derived from the power section of an aircraft engine, is shown in Fig. 1.

1. PHYSICAL AND TECHNICAL PRINCIPLES

1.1 The Brayton Cycle and the Gas Generator

The gas turbine engine cycle involves a single-phase gaseous working fluid, most often air and the products of combustion. As with most heat engines, the gas turbine engine functions on the basis of compression of a relatively cool working fluid, followed by heat addition to the compressed working fluid and followed by expansion or work extraction from the hot, high-pressure working fluid. Since more work can be extracted from the hot high-pressure working fluid than is required to compress the cooler fluid, there is a surplus of mechanical energy left

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

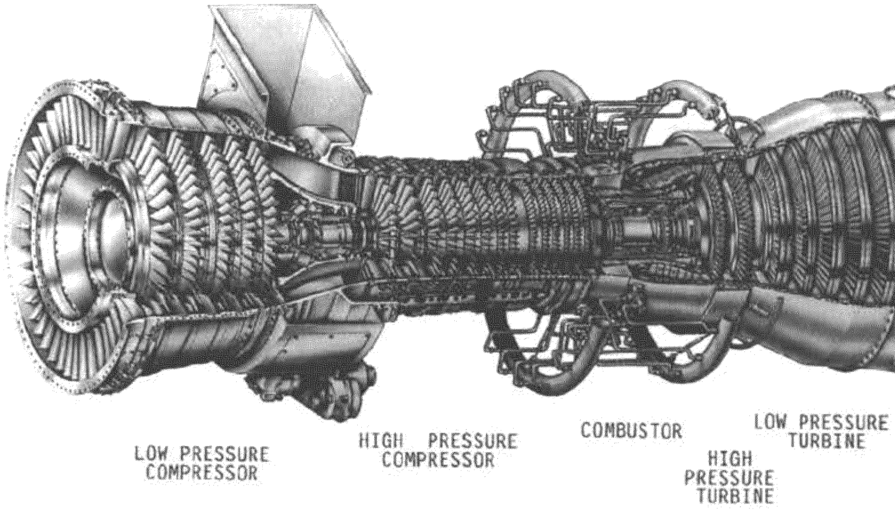

FIG. 1. A typical modern gas turbine engine, the GE LM6000 derived from an aircraft engine for marine and industrial application (courtesy of the General Electric Co.).

as a net output after subtracting the energy required to power the compression process.

The compression and expansion (or work extraction) are accomplished by, respectively, turbo-compression and turbo-expansion components, which impart energy to or extract energy from the steady stream of working fluid. The heating of the working fluid is accomplished in a steady flow process at essentially constant pressure either by heat exchange with an external heat source or, more usually, by continuous internal combustion of a gaseous or liquid fuel in the air that is the usual working fluid of the thermodynamic cycle. The constant-pressure combustion is one of the unique features of the gas turbine engine that characterize the Brayton cycle. When designed as an internal combustion engine, the system is, of necessity, an open cycle. That is, the existing working fluid cannot be cooled and returned to the inlet since its oxygen has been converted to products of combustion, and it cannot sustain combustion a second time. The vitiated air is discharged from the engine, and the working fluid is constantly replenished at the compressor inlet with fresh air.

The expansion process is most often carried out in two distinct steps, manifested by two distinct turbine components. In the first, or high-pressure, turbine, the working fluid is expanded through only a portion of the available pressure ratio, enough to supply the energy to drive the compressor. This combination of compressor, combustor, and high-pressure turbine is referred to as the gas turbine engine core or gasifier or gas generator. Its output, sometimes referred to as gas horsepower, is a continuous stream of high-pressure, high-temperature working fluid that contains the net useful energy output of the engine. In some instances such as with the jet engine, the stream of high-pressure, high-temperature gas is the appropriate net product of the gas turbine engine. But, more often, the core is followed by a second expansion component referred to as the low pressure or power turbine, operating on an independent rotor at an independent rotational speed, which extracts the available mechanical energy from the core discharge stream and delivers it to the load that is being driven. In this arrangement, the level of power can be modulated by varying fuel flow and letting the core rotational speed vary, without affecting the rotational speed of the driven load.

1.2 Energy Conversion and System Efficiency

The heat released by burning a typical hydrocarbon fuel in air is typically 18 650 BTU/lb_m (or 43 370 kJ/kg) of fuel. If this process were 100% efficient, it would produce power for every unit of fuel flow of 7.45 hp/(lb$_m$/h) [or 12.0 kW/(kg/h)]. In actual fact, the fundamental thermodynamic limitation termed the Carnot efficiency (η_{Carnot}), relating the peak gas temperature achieved in the cycle ($T_{\max}$) and the temperature of the inlet gas ($T_{\min}$), restricts the efficiency to a value of

$$\eta_{\text{Carnot}} = 1 - (T_{\min}/T_{\max}). \tag{1}$$

The cycle efficiency is further limited by the fact that the Brayton cycle involves heat addition to the working fluid at constant pressure rather than at the idealized constant temperature of the Carnot cycle. The peak pressure achieved in the Brayton cycle also affects the efficiency of energy generation. This then implies that the upper bound of the energy production per unit fuel flow is generally no more than 60% efficient, implying that the lower bound of specific fuel consumption (SFC) for an engine is 0.224 (lb$_m$/h)/hp [or 0.139 (kg/h)/kW]. In actual practice, the specific fuel consumption will be even higher than this lower bound because of less than ideal performance in the individual components of the prime mover, including leakage of high-pressure working fluid (air or products of combustion); leakage of heat from the high-temperature working fluid to the lower-temperature engine bay; pressure losses in the passages through which the working fluid moves from component to component; diversion of a portion of the air to cool high-temperature parts; and inefficiencies in the compression, combustion, and expansion components of the engine.

High cycle peak temperatures and pressure ratios are of considerable importance in achieving low SFC. Modern gas turbine engine power-producer sections, particularly where they are derived from advanced aircraft engines, may have cycle pressure ratios as high as 40:1 and peak cycle temperatures on the order of 2800 °F (or 1573 °C).

1.3 Cycle Elaboration

As shown in Fig. 2 in schematic form and in thermodynamic representation of variation of temperature vs entropy, variations of the basic Brayton cycle [Fig. 2(a)] such as regeneration [Fig. 2(b)], intercooling [Fig. 2(c)], and reheat [Fig. 2(d)] are used alone or in combination to improve the gas turbine engine's efficiency of energy conversion and/or to increase the amount of power available from a given size machine with a limited through-flow of working fluid (Wilson, 1984).

1.3.1 Regeneration Depending on the configuration of the heat exchanger used, this is also referred to as recuperation. As shown in Fig. 2(b), heat may be extracted from the low-pressure working fluid at the engine's discharge and used to preheat the high-pressure working fluid discharged by the compressor before it enters the combustor. The heat that is transferred makes possible a direct saving of fuel that would have otherwise been expended in the combustor of an unregenerated engine.

1.3.2 Intercooling The amount of energy required to compress a gas is proportional to the temperature of the gas. A significant saving in the energy consumed by the compression process can be achieved by splitting the compressor into two or more stages, as shown in Fig. 2(c), and cooling the working fluid between each pair of successive stages of compression. Intercooling generally results in an increase in power production for a given size engine but requires a heat sink for the heat rejected in the intercooler.

1.3.3 Reheat This is also referred to as interburning. The amount of energy that can be extracted from the expansion of a pressurized working fluid is directly proportional to the temperature of the fluid. Therefore, as shown in Fig. 2(d), additional work can be obtained from an engine of a given size by splitting the turbine into two or more stages and inserting a burner or burners between each successive pair of stages. The principle is very limited in its application because heat insertion at low pressure is less efficient than heat addition at high pressure, so that reheat generally degrades the fuel economy of the engine. The reheat combustor is also a challenging component design because of its operation at higher inlet temperature and lower pressure than the main combustor.

2. DESIGN, FUNCTION, AND OPERATION

2.1 Key Components

The basic components of the gas turbine engine are the compressor(s), the combustor, and the turbine(s).

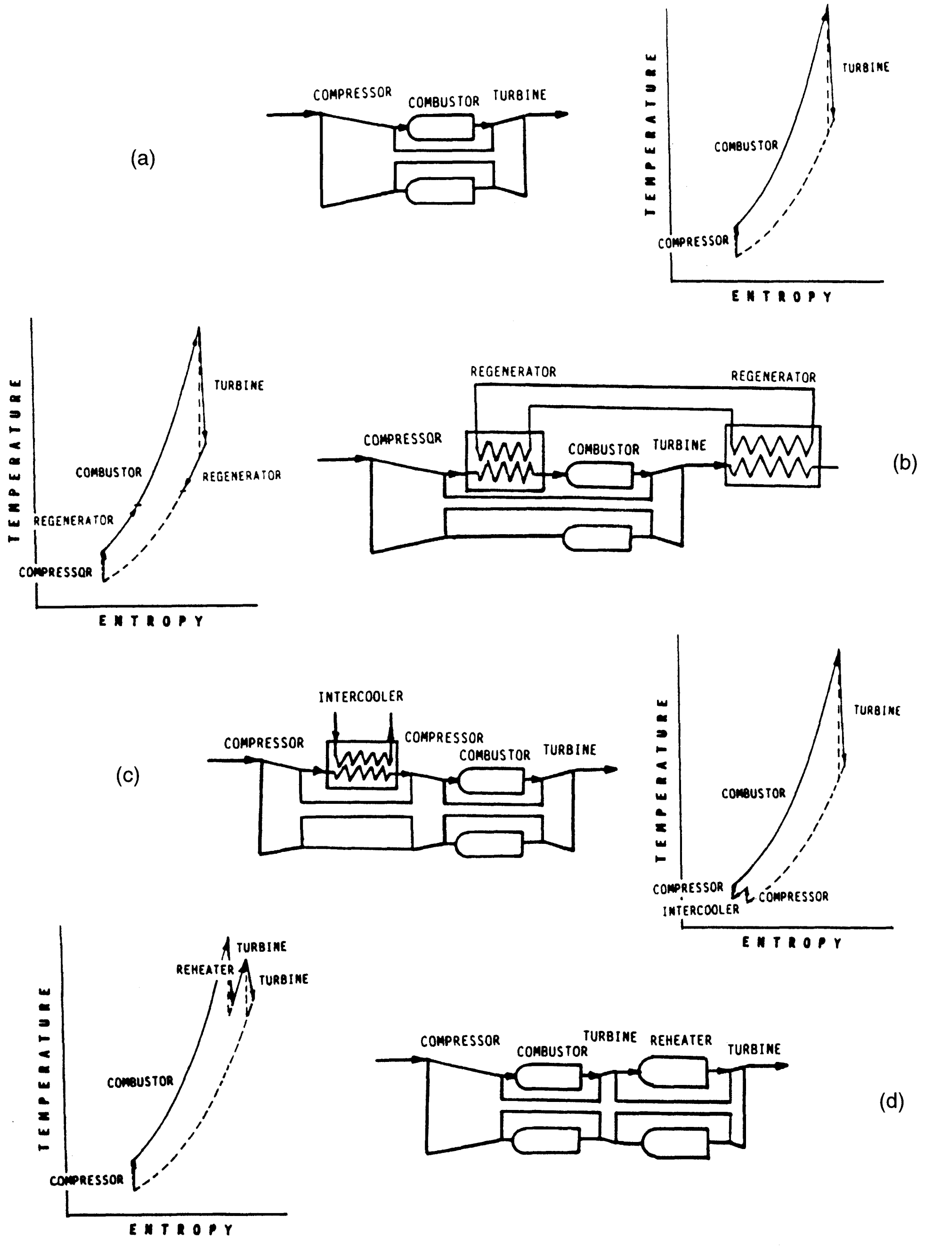
(a)
COMPRESSOR
COMBUSTOR
TURBINE
TEMPERATURE
ENTROPY
(b)
REGENERATOR
REGENERATOR
COMPRESSOR
COMBUSTOR
TURBINE
TEMPERATURE
ENTROPY
(c)
INTERCOOLER
COMPRESSOR
COMPRESSOR
COMBUSTOR
TURBINE
TEMPERATURE
ENTROPY
(d)
COMPRESSOR
COMBUSTOR
TURBINE
REHEATER
TURBINE
TEMPERATURE
ENTROPY

FIG. 2. Refinement of the basic gas turbine engine cycle: (a) the simple cycle, (b) regeneration, (c) intercooling, (d) reheat.

2.1.1 Compressors The efficient turbocompressor, which delivers a steady stream of high-pressure working fluid to the combustor, is fundamental to the gas turbine engine. The compression of the working fluid flowing in an annular passage is accomplished by means of an annular set of airfoils or blades mounted on the rim of a high-speed rotating disk, which accelerates the stream by imparting to it some of the tangential velocity of the rotor. The rotor is followed by an annular set of stationary airfoils or vanes that decelerate or diffuse the stream providing a net increase in the pressure.

Since the pressure rise in a single consecutive pair of rotating and stationary sets of airfoils (i.e., a stage) is limited by the diffusion (or stall) characteristics of the airfoils, the high pressures required by efficient gas turbine engine cycles are generally achieved by multiple-stage machines.

Although the compressors of early gas turbine engines were radial- or centrifugal-flow components, axial-flow compressors have been developed to achieve higher efficiency levels, lighter weight, and, for aircraft usage, lower frontal areas. They are therefore the prevailing approach in all gas turbine engines except for the smallest sizes where the axial-flow airfoils are not rugged enough for engine duty. In these small machines, we find single- and two-stage centrifugal compressors, as well as hybrid axi-centrifugal components—one or more axial stages followed by a centrifugal stage.

Multiple-stage compressors with pressure ratios above 5:1 or 10:1 involve extremely wide variation of flow conditions to the individual rows of airfoils in the front and the rear of the compressor as the speed and pressure ratio vary through the operating range. Such mismatch is aggravated by other operational circumstances such as nonuniform pressure distributions delivered by the inlet duct and/or rapid changes in the spool speed. Ultimately, the large angles of attack seen by individual airfoils will cause them to stall and result in unstable operation of the engine called surge. Such surge may cause flame-out or other malfunctions that result in loss of power and physical damage to the engine. Various approaches have evolved to afford high-pressure compressors the required amounts of stall margin. As shown in Fig. 1, a very high-pressure compressor may be split into two sections—the low- and high-pressure sections—each operating on a separate spool at its own speed and driven by its own turbine. An alternative or supplementary approach to the design of high-pressure compressors (called variable geometry) is to provide pivots for each stator vane in certain stages of the compressor along with actuators that will vary the pitch angle of the vanes as a function of the spool speed and compressor inlet conditions to avoid the large angles of attack and the stall and surge that might ensue.

2.1.2 Combustors The combustor of the modern gas turbine engine, sometimes referred to as the burner, is usually an annular chamber, in the middle of the core between the compressor exit and the turbine entry, and surrounding the center shaft(s) by which the turbine(s) aft of the combustor drives the compressor(s) forward of the combustor, as shown in Fig. 1. Fuel is introduced at the upstream end of the combustor through injectors that may exploit one of several mechanisms to condition the fuel for the combustion process. With a high-pressure fuel or atomizing system, the fuel is introduced through very small orifices in the nozzle, which generate a very fine spray. With an intermediate fuel-pressure system, sometimes referred to as the air-blast injection system, the droplets may not be fine enough to permit ignition-efficient combustion, and so various types of shear layers are introduced into the airstream in proximity to the injector to break up the droplets further. With a low-pressure fuel system or a vaporizing system, the injectors may also be located in a hot region of the combustor so that the fuel is preheated and completely vaporized before entering the combustor.

The combustion process is confined in a liner or shell, cooled by unburned air, which serves not only to contain the flame but also to provide passages for air jets, lateral to the main direction of the flow, which promote and stabilize the flame, and dilute its high-temperature zones to a more uniform level of temperature that can be tolerated by the downstream turbine. One or more electric spark igniters are placed strategically in the chamber to light the flame in the course of starting the engine.

2.1.3 Turbines These are also referred to as expanders. An annular row of stationary airfoils referred to as nozzles is used to accelerate and impart tangential motion to the high-temperature, high-pressure stream of working fluid exiting from the annular discharge passage of the combustor. That high-velocity stream impinges on a row of airfoils, referred to as blades or buckets, arranged in an annular space at the perimeter of a high-speed rotating disk, which acts in the manner of a windmill to extract energy from the stream. Since the amount of energy that can be extracted from a single pair of stationary and rotating blade rows (or turbine stage) is generally limited by such factors as maximum transonic gas velocities, a multiple stage turbine is often required to complete the energy extraction process, as illustrated in Fig. 1. Where compression is accomplished in more than one compressor component, a separate turbine component is required to drive each compressor in addition to what is usually a separate turbine to extract and deliver the useful power output.

2.2 Subsystems

The gas turbine engine control accepts a variety of signals from sensors situated around the engine that provide data such as measurements of key pressures and temperatures within the engine, rotative speed(s) of the engine's spool(s), actual position of the engine's variable geometry, torque in the engine's output shaft, rate of fuel flow to the engine's combustion system, and a signal from the operator or pilot or system controller indicative of the demand for power. On the basis of these data, the control then energizes actuators that manipulate fuel valves and other variable geometry of the engine to provide stable power at the demanded level and, when requested, provide smooth and fast changes in power level. The control must also provide for starting and shut-down of the engine, must protect the engine from surge, overspeed, over-temperature, and over-torque, and, in the event of any failure, must provide for residual power or for a safe shutdown. Early generations of gas turbine engines used hydro-mechanical controls, but modern systems are generally based on digital electronic devices.

Another important subsystem of gas turbine engines is an accessory gearbox, generally driven by the engine's core rotor, and feeding power to the engine's electric generator, and the engine's fuel, oil, and hydraulic pumps. A starting system (a hydraulic, pneumatic, or electric motor) is also mounted on the gearbox to drive the spool up to a speed where it is pumping enough air to permit combustion and is able to generate enough energy to be self-sustaining. The fuel system includes the fuel pumps, regulating valves, flow dividers, and fuel nozzles. A lubrication system is included for the rotors' bearings, including oil pumps, oil nozzles, scavenge pumps, oil filters, and an oil tank with an air/oil separator. The ignition system includes a high-voltage unit to energize the igniter. In aircraft, an anti-icing system provides hot air and/or electrical heating to parts in the engine inlet that might otherwise get clogged with ice during flight through super-cooled moisture in the atmosphere. Electrical and/or hydraulic and/or pneumatic actuators drive the engine's variable geometry.

3. MATERIALS AND MANUFACTURE

The gas turbine engine has been the focal point for much of the innovation and progress in advanced light-weight and high-temperature structural material development in recent decades, and in the new manufacturing processes that are required to produce the designs that fully exploit their potential.

3.1 Light-Weight Materials

Because of the high premium put on light weight, the aircraft gas turbine engine has been one of the primary application for the development of high-strength structural alloys of titanium. Although it is used extensively in high-speed rotating structures of aircraft engines, there has been some retrenchment from earlier, more ambitious applications because of problems encountered in titanium fires. Carbon fiber/organic resin matrix composites have been used for structural components and more recently for high-speed turbofan blading, which must withstand the extraordinarily high-impact loading of bird strikes and foreign object ingestion. For somewhat higher-temperature applications, future developments are aimed at exploiting the potential of metal matrix composites, and of intermetallics such as titanium aluminide, and ceramics in monolithic and composite form.

3.2 High-Temperature Materials

The considerable thermodynamic benefit of high turbine inlet temperature has placed enormous demands on virtually all the mechanical, aerodynamic, heat transfer, metallurgical, and manufacturing process arts. The blades of the turbine rotate at velocities at transonic or supersonic speed in a centrifugal acceleration field thousands of times that of gravity, and the entire turbine must operate in a chemically active transonic gas stream that contains abrasive particles and whose temperature is well above the melting temperature of the material of which the blades are made. Where the working fluid of a typical Otto or Diesel engine will reach peak cycle temperature for only an instant in the compression/combustion/expansion cycle so that the cylinder wall stabilizes at a temperature well below peak cycle temperature, the turbine of the gas turbine engine must stabilize at peak cycle temperature continuously when the engine is operated at peak power. The blades and other parts of the turbine must also endure thousands of cycles of large changes in temperature and stress as the engine is started and shut down and as power is varied during the course of its normal operation.

From the earliest application of high-temperature steel alloys, gas turbine engine metallurgy has progressed through development of cobalt- and nickel-based alloys strengthened by precipitation hardening. Many modern superalloys cannot be cast into an ingot with uniform properties and are produced by consolidating alloy powders by hot isostatic pressing. Mechanical alloying and dispersion strengthening are used in some cases. Future development is aimed at the use of intermetallics and ceramic matrix composites, as well as monolithic ceramics for stationary components.

At high temperatures, the grain boundary is the weakest link in a conventional superalloy system. The effect of grain boundaries is minimized in turbine blades by growing elongated grains, aligned in the direction of the principal stress, using a casting process known as directional solidification. In recent times, turbine blades have been directionally solidified as a single grain, thus eliminating grain boundaries altogether.

3.3 High-Temperature Cooled Turbine Blades

Not withstanding the very high useful operating temperature capability of currently available turbine blade alloys, gas turbine engines are routinely run at gas temperatures several hundred degrees Celsius above the melting point of these materials. This extraordinary feat is accomplished by cooling the blades with small quantities of high-pressure air that bypasses the combustor. The cooling schemes involve convective, impingement, film, and transpiration heat-transfer mechanisms, which require hundreds of small holes to be drilled into the blades and structural components of the high-temperature turbine. Such advanced drilling processes as electrochemical machining (ECM), electrical discharge machining (EDM), laser drilling, and abrasive water-jet machining have been perfected in this context. The cooling schemes also require minute convoluted passages to be built into the high-temperature structural parts. Where those parts are cast, those passages are usually produced by introduction of precisely molded and positioned ceramic cores into the casting molds. These delicate cores have been developed to withstand the abusive pouring temperatures of the high-temperature alloys for the long periods of time taken to grow the single crystals or directional crystals that constitute the finished cast part. The cores are removed from the casting by leaching in an autoclave.

4. APPLICATIONS

Gas turbines have found extensive application in stationary power plants, and in marine, vehicular, and aircraft applications.

4.1 Stationary Power Plants

Although the steam turbine is by far the most common power generation device in stationary power plants, gas turbine engines have found a variety of applications in special niches. Since their peak cycle temperature is much greater than the temperatures commonly achievable in steam turbine cycles, gas turbine engines are used as the topping cycle in combined (gas turbine/steam turbine) power plants where the gas turbine engine's exhaust gases are used to fire the boilers of the lower-temperature steam turbine system. Such plants are among the most efficient in operation. The gas turbine engine is also the system of choice in remote and relatively inaccessible locations such as ocean oil-drilling platforms and pipeline pumping stations, where relatively small power requirements do not justify the introduction of electrical power grids. Gas turbines are also used where emergency standby power, independent of the power grid, is needed on very short notice. The economic advantage of small gas turbine engine applications is magnified in cogeneration applications where the waste heat from the gas turbine engine exhaust can be exploited for space heating or industrial processes. The gas turbine engine is also applied in unique circumstances where its relatively short procurement cycle and more modest investment cost may overcome its inability to use lower-cost coal.

4.2 Vehicular Applications

Although they showed great promise in the 1970s and 1980s, gas turbine engines have never developed any major application in automobiles. In order to compete in thermodynamic efficiency with the common Otto-cycle reciprocating engine (particularly at important idle condition where the gas turbine engine is at a disadvantage), they require very high firing temperatures and the use of efficient, compact heat exchangers for regenerative thermodynamic cycles. Both of these requirements involve costly design features that make the final product cost noncompetitive, although the application of structural ceramics encourages many to think that this disadvantage may yet be overcome. Gas turbine propulsion has found application in railroad locomotives, in some large, long-distance trucks and buses, and in military vehicles where their light weight, compactness, and durability are greatly valued.

4.3 Marine Applications

Gas turbines are the modern powerplant of choice in small- and medium-sized naval ships because of their light weight, compact size, quick start-up characteristics, and competitive fuel economy.

4.4 Aircraft Applications

In all but the very smallest and lowest-speed general aviation aircraft, virtually all modern aircraft engines are based on the gas turbine engine core as prime mover. From helicopter turboshaft engines to supersonic transport powerplants, the gas turbine engine is harnessed to provide for aircraft propulsion. Propulsive thrust in aircraft is generated by a propulsor—one or more stages of turbomachinery that accelerate portion of the oncoming airstream. Generalizing Newton's second law of motion to deal with a continuous stream, the reaction force or thrust (F) on the device that accelerates the stream is found to be proportional to the product of the rate of mass flow (M) of the stream though the propulsor and the change in velocity of the stream (ΔV) imparted by the propulsor:

$$F = M(V_j - V_o) = MV_o(\Delta V/V_o), \tag{2}$$

where V_o is the inlet air velocity relative to the propulsor and is taken to be the flight velocity and V_j is the exhaust or jet velocity relative to the propulsor. An important consequence of this relationship is that a propulsor designed for very low-speed aircraft such as the rotor of a helicopter must accommodate large amounts of airflow compared to the airflow through the gas turbine engine that powers the propulsor. On the other hand, at very high supersonic flight speeds, the ideal propulsor requires approximately the same flow quantity as the gas turbine powerplant so that the gas turbine engine gas generator itself can be used as the propulsor (i.e., the jet engine) to accelerate the propulsion airstream without resort to a power turbine driving a separate propulsor. In the intermediate range of flight speeds between the low-speed helicopter and the supersonic transport, we find a continuous spectrum of propulsion systems, as shown in Fig. 3, which utilize both the jet from the gas turbine engine power producer and a propulsor (typified by a propeller and a high bypass fan) driven by energy extracted by a power turbine from the exit stream of the gas turbine power producer to create propulsive thrust.

5. ECONOMIC AND ENVIRONMENTAL ASPECTS

5.1 System and Fuel Economy

In fuel-burning stationary power plants, the gas turbine engine is at an economic disadvantage since it cannot be easily adapted

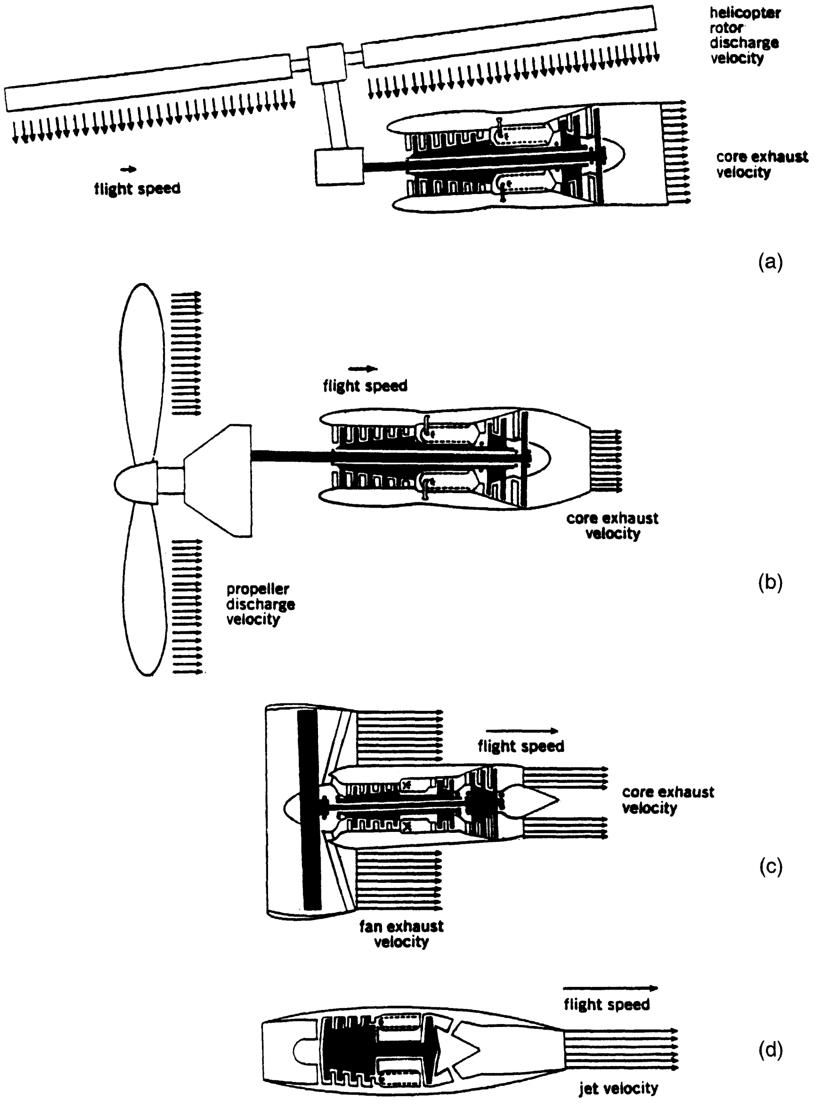
helicopter rotor discharge velocity
flight speed
core exhaust velocity
(a)
flight speed
core exhaust velocity
propeller discharge velocity
(b)
flight speed
core exhaust velocity
fan exhaust velocity
(c)
flight speed
jet velocity
(d)

FIG. 3. The spectrum of aircraft gas turbine engine propulsors (Ehrich, 1982): (a) turboshaft helicopter rotor drive, (b) turboprop, (c) high bypass turbofan, (d) turbojet.

to burn low-cost coal. But it has found important special applications where its compactness, light weight, simple site preparation requirements, speed of start-up, lower capital cost, lack of need for heat rejection, and ready availability of waste heat application in cogeneration applications make up for its fuel cost disadvantage.

In ship applications, its high-power fuel efficiency, light weight, and compactness make the gas turbine engine the power plant of choice for high-speed applications such as naval craft. The same attributes would make the gas turbine engine a principal choice for land vehicles but for two competitive disabilities: the gas turbine engine's poor fuel consumption at low part power and its relative high cost. Application of the regenerative cycle has made possible its successful application to military tanks, but its cost disadvantage is yet to be overcome in realization of its potential for mass application to land vehicles.

As previously noted, the gas turbine engine is virtually unchallenged as the power plant of choice for all aircraft from helicopters to supersonic transports, except for the very smallest helicopters and light aircraft requiring no more than several hundred horsepower. This universal advantage is due not only to the gas turbine engine's excellent fuel economy and versatility in achieving nearly ideal propulsive efficiency in any flight regime, but also is due, in no small part, to its light weight, which implies reduced lift and hence reduced drag for a given aircraft, and to its low frontal area that translates directly into low drag.

5.2 Exhaust Emissions

Since the working fluid in the turbine section of the gas turbine engine is composed of the products of combustion from the combustion chamber, the engine's exhaust emissions (Bahr, 1980; Barrett, 1991) are typical of those of any heating or power plant operated with fossil fuels. One class of emissions of key concern consists of incompletely burned hydrocarbon fuel—smoke (soot), unburned hydrocarbons (HC), and carbon monoxide (CO). Major reduction of all these emissions has resulted from deeper understanding of the combustion mechanisms that generate the emissions and development of design approaches that inhibit their formation. The attainment of low smoke levels requires engine combustor designs with both leaner primary-combustion–zone fuel-air mixtures and more effective fuel-air mixing within the primary zone, compared to those of then-existing combustors. HC and CO are primarily generated at idle and other low engine-power conditions. The combustor design features needed to reduce HC and CO levels involve provisions for improving fuel atomization quality at low-power conditions, enhancing primary-zone fuel-air mixing quality at these operating conditions and minimizing combustion quenching due to fuel impingement on the combustor liners.

A second class of nonvisible emissions of concern is oxides of nitrogen (NO_x), which are also by-products of combustion. When gases containing oxygen and nitrogen are heated to elevated temperatures, generally above 1900 K, some oxidation of the nitrogen occurs, mainly to nitric oxide (NO), together with small amounts of nitrogen dioxide (NO_2). Early measures at NO_x reduction, applicable primarily to stationary powerplants, involved water injection into the combustor to cool the local elevated temperature zones. Modern combustor design practice involves obtaining both the required lean and uniform primary zone mixtures at high-power conditions and of providing adequately rich primary zone mixtures at idle and other low-power conditions.

While not considered to be a pollutant emission in the same sense, the carbon dioxide (CO_2) emissions of gas turbine engines are also of possible concern in context of the greenhouse effect on world climate. Abatement of CO_2 can only be achieved by reduction of the specific fuel consumption of the engine or via the use of alternative fuels with lower carbon content such as natural gas. Some sulfur oxides (SO_x) emissions are also generated by gas turbine engines. However, the quantities are very small because the sulfur contents of typical turbine engine fuels are extremely low.

Regulation of emissions is usually specified on the basis of the application, with separately specified and administered regulations applicable to stationary power plants and aircraft.

5.3 Noise Emissions

Two basic types of noise signature can be emitted from gas turbine engines. First, broadband or "white" noise, characterized by a roaring or hissing sound, is generated when random turbulence is produced, either by flow over engine blade, vane, and duct wall surfaces or in the mixing layers of the engine airflow streams. The frequency spectrum is typically a broad, smooth "hump" having a characteristic peak and associated peak frequency. The second type of noise produced is discrete tone noise, characterized by a distinct whistle or whine, and is typically a sharp spike of noise occurring at a specific frequency, and often having harmonics or overtones.

In general, the entry compression turbomachinery (i.e., fan of an aircraft engine or the low-pressure compressor) and the discharge turbine (i.e., the low pressure or power turbine) produce both broadband or "white" noise and strong, discrete frequency tones, usually at integer multiples of blade-passing frequency of the rotor blades. The combustor and the jet exhaust mixing layers produce predominantly broadband or "white" noise. Also, stages of entry compression turbomachinery that operate at supersonic tip speeds will often emit multiple-pure-tone or "buzz-saw" noise, which is characterized by discrete tones occurring at integer multiples of fan shaft rotational frequency.

In stationary, vehicular, and marine applications, it is possible to install sound absorbing enclosures for the engine and sound attenuating inlet and exhaust systems. A major challenge remains in aircraft applications (Gliebe, 1992; Smith, 1989) where weight and volume are at a premium, where the propulsor (typically the fan) is considerably larger than the basic power generating core, and where the velocity of the exhaust jet must be maintained at relatively high levels, of the order of the aircraft's maximum flight speed.

Early generations of aircraft gas turbine engines were simple jet engines or low-bypass turbofans with extremely high exhaust jet velocities where broadband noise predominated. The principal means of suppression was to provide the exhaust stream with mixers that induced intense mixing between the jet stream and the external air stream within the confines of an enclosed duct. The current trend in the development of aircraft engine propulsion systems is to employ cycles with higher and higher bypass ratios, in order to achieve high propulsive efficiency. As bypass ratio increases, the noise from the broadband sources (i.e., the exhaust jet and the combustor) decreases rapidly, while the noise from the tone sources (i.e., the fan, low-pressure compressor, and low-pressure turbine) increases, until the dominant sources are the tone sources. Methods for reducing the tone noises caused by unsteady interactions include the following:

1. reducing the magnitude of the unsteady flow field felt by each turbomachinery blade—i.e., minimize upstream blade wake defects, minimize downstream obstacle pressure fields, reduce or eliminate inflow distortions and separations, etc.;
2. selecting the blade number to minimize the acoustic efficiency of the resulting unsteady interaction, including using cutoff, whereby the acoustic pressure waves decay exponentially inside the duct; and
3. employing duct liners, generally based on arrays of Helmholtz resonators embedded in the duct walls, to suppress the generated acoustic pressure waves inside the duct before they escape from the inlet or exhaust.

GLOSSARY

Angle of Attack: The angle between the gas or fluid flow and the mean line of the airfoil's leading edge.

Blade: A compressor or turbine airfoil.

Bucket: A rotating turbine airfoil.

Burner: A combustion system.

Bypass Ratio: The ratio of that portion of the mass flow discharged by the propulsor that bypasses the power producer to that portion that passes through the power producer.

Cogeneration: Power production by an integrated and self-contained power plant that makes practical use of the exhaust heat.

Combustor: A combustion system.

Composite: A material that has been strengthened and/or stiffened by incorporation of fibers of a dissimilar material.

Compressor: In gas turbine engine usage, the component composed of one or more pairs of stationary and rotating airfoils that impart(s) pressure to a continuous flow of the working fluid.

Compressor, Axi-centrifugal: A compressor with one or more axial flow stages and one centrifugal stage.

Compressor, Axial Flow: A compressor where the principal direction of flow is parallel to the centerline of rotation of the compressor.

Compressor, Centrifugal: A compressor where the principal direction of flow is in the radial direction or normal to the centerline of rotation of the compressor.

Compressor, Radial Flow: See **Compressor, Centrifugal.**

Cooling, Convective: Cooling effected by transfer of heat to a flow of coolant internal to the structure being cooled.

Cooling, Film: Cooling effected by the laying down of a barrier film of coolant over the external surface of the structure to be cooled.

Cooling, Impingement: A form of cooling where the coolant is sprayed in jets onto the surface of the structure to be cooled.

Cooling, Transpiration: Cooling by passing the coolant through a porous member of the structure to be cooled.

Core, Casting: An insert embedded in a casting mold to form a hollow chamber in the completed casting.

Core, Engine: That part of the engine composed of the basic compression component(s), the combustor, and the basic expansion component(s) that drive the compressor(s) and that, as an integrated unit, deliver mechanical energy in the form of a steady flow of high-pressure, high-temperature gas.

Cycle, Brayton: The basic gas turbine engine cycle, with continuous flow and with constant-pressure heat addition.

Cycle, Open: A thermodynamic cycle where the working fluid is *not* returned to the engine inlet.

Cycle, Thermodynamic: The sequence of variation in the thermodynamic properties of the working fluid of an engine that characterizes the sequence of processes of compression, heat addition, heat extraction, and expansion in the course of the engine's operation.

Cycle, Topping: A thermodynamic cycle with a very high exhaust temperature that is used to provide the heat input for a lower-temperature cycle.

ECM: Electrochemical machining.

EDM: Electrical-discharge machining.

Efficiency, Carnot: Efficiency of a reference thermodynamic cycle indicative of the maximum achievable fuel economy with ideal components.

Efficiency, Propulsive: The ratio of the energy expended in propelling a vehicle to the energy delivered to the propulsor.

Emission(s): Most often refers to the undesirable chemical constituents of the gas turbine engine exhaust but is also used to refer to the noise emitted by the engine.

Engine, Gas Turbine: See Introduction.

Entropy: A thermodynamic property whose variation in compression and expansion processes is indicative of the deviation of the process from the ideal.

Expander: See **Turbine.**

Fan: A turbo-compressor used in aircraft engines as a propulsor.

Flame-out: An uncommanded loss of combustion.

Gasifier: See **Core, Engine.**

Generator, Gas: See **Core, Engine.**

Injector, Fuel: See **Nozzle/Injector, Fuel.**

Interburning: See **Reheat.**

Intercooler: A heat exchanger that extracts heat from the working fluid in the midst of the compression process.

Intercooling: Extraction of heat from the working fluid in the midst of the compression process.

Jet Engine: The simplest form of aircraft engine, composed of a gas turbine engine core and a jet nozzle. Sometimes used generically to indicate any type of aircraft gas turbine engine.

Liner, Combustor: The shell in which the combustion process is contained.

NO_x: Oxides of nitrogen.

Noise, Broadband: Noise having a frequency spectrum that is typically a broad, smooth "hump" having a characteristic peak and associated peak frequency.

Noise, Buzz-Saw: See **Noise, Multiple-Pure-Tone.**

Noise, Multiple-Pure-Tone: Noise that is characterized by discrete tones occurring at integer multiples of fan shaft rotational frequency.

Noise, Tone: Noise, characterized by a distinct whistle or whine, which is typically a sharp spike of noise occurring at a specific frequency, and often has harmonics or overtones.

Noise, White: See **Noise, Broadband.**

Nozzle, Exhaust: See **Nozzle, Jet.**

Nozzle, Jet: A convergent or convergent/divergent duct that is used to accelerate the exhaust flow from a gas generator or fan in process of generating propulsive thrust.

Nozzle, Turbine: A single stationary turbine airfoil or an entire row of stationary turbine airfoils.

Nozzle/Injector, Fuel: The device used to precondition and inject a liquid or gaseous fuel into a combustion chamber.

Nozzle/Injector, Fuel, Air Blast: An injector that introduces a strong shear layer and turbulence in the surrounding airstream to assist in conditioning the fuel spray for combustion.

Nozzle/Injector, Fuel, Atomizing: An injector that depends primarily on impelling the fuel at very high pressure through a very small orifice to produce a very fine spray.

Nozzle/Injector, Fuel, Vaporizing: An injector that depends on preheating the fuel to condition it for combustion.

Pitch Angle: Angular orientation of a turbomachinery airfoil or propeller blade, as measured from the plane of rotation.

Propulsor: The propeller or helicopter rotor or fan that develops the propulsive thrust from the mechanical energy delivered by the gas turbine engine.

Recuperation: A form of regeneration.

Recuperator: A form of regenerator.

Regeneration: A variation of the gas turbine cycle wherein heat is extracted from the exhaust stream and added to the working fluid that enters the combustor, in order to reduce the fuel required by the cycle.

Regenerator: The heat exchanger that accomplishes regeneration.

Reheat: Addition of heat to the working fluid in the midst of the expansion process.

Resonator, Helmholtz: An aerodynamic oscillator, multiple units of which are used in arrays as noise suppressors in gas turbine inlet and exhaust ducts.

SFC: Specific fuel consumption, the ratio of rate of fuel burned to the thrust or power produced by the engine.

SO_x: Oxides of sulfur.

Spool: A rotor of the gas turbine engine that includes the rotating disks and airfoils of a turbine and the compressor it drives or a rotor composed of a turbine and a drive shaft to an external load.

Stage: An integrated pair of rotating and stationary airfoils constituting a subsystem of a compressor, or a pair of stationary and rotating airfoils constituting a subsystem of a turbine.

Stall: The breakdown of the airflow pattern over a row of compressor airfoils with consequent loss of diffusion and compression capability.

Stall Margin: The increment of pressure ratio that a compression system can obtain at a given operating rotational speed without encountering system surge.

Surge: Unstable operation of a gas turbine engine when operated about a limiting pressure ratio at a given rotational speed.

Thrust: Reaction force supplied to the mount points of an aircraft gas turbine and/or propulsor to propel the aircraft.

Turbine: The component composed of one or more pairs of stationary and rotating airfoils that extracts energy from a continuous flow of the working fluid.

Turbine, Gas: See Introduction.

Turbine, High Pressure: The first turbine in a system with more than one turbines.

Turbine, Low Pressure: The last turbine in a system with more than one turbines.

Turbine, Power: A turbine specifically dedicated to driving an external load.

Turbo-compressor: See **Compressor.**

Turbo-expander: See **Turbine.**

Turbofan: An integrated aircraft engine that includes a gas turbine engine and a fan.

Turbojet: See **Jet Engine.**

Turbomachinery: A gas turbine engine's compressor and turbine.

Turboprop: An integrated aircraft engine that includes a gas turbine engine and a propeller.

Turboshaft: A gas turbine engine that delivers mechanical power through a rotating shaft to an external mechanical load.

Vane: A stationary airfoil of a compressor.

Variable Geometry: Elements of the gas turbine engine's configuration such as the pitch angle of the stator vanes or the throat area of the jet nozzle that can be varied during operation.

Works Cited

Bahr, D. W. (1980), "Gas Turbine Engine Emission Abatement," in: *Proceedings of Project Squid Workshop on Gas Turbines Combustor Design Problems,* Hemisphere Publishing Co.

Barrett, M. (1991), *Aircraft Pollution—Environmental Impacts and Future Solutions,* WWF Research Paper.

Ehrich, F. F. (1982), "Turbine Propulsion," in: *Encyclopedia of Science and Technology,* 5th ed., New York: McGraw-Hill.

Gliebe, P. R. (1992), *Ultra-High Bypass Engine Acoustic Technology Overview,* DGLR/AIAA Paper No. 92-02-037, Aachen, Germany.

Smith, M. J. T. (1989), *Aircraft Noise,* Cambridge, U.K.: Cambridge Univ. Press.

Wilson, D. G. (1984), *The Design of High-Efficiency Turbomachinery and Gas Turbines,* Cambridge, MA: MIT Press.

Further Reading

Cohen, H., Rogers, G. F. C. (1987), *Gas Turbine Theory,* London: Longmans Green & Co.

Dixon, S. L. (1966), *Fluid Mechanics, Thermodynamics of Turbomachinery,* Oxford, U.K.: Permagon Press.

Harman, R. T. C. (1981), *Gas Turbine Engineering,* New York: Wiley.

Meetham, G. W. (1981), *The Development of Gas Turbine Materials,* New York: Wiley.

Whittle, F. (1981), *Gas Turbine Aero-Thermodynamics,* Oxford, U.K.: Pergamon Press.

Wilson, D. G. (1984), *The Design of High-Efficiency Turbomachinery and Gas Turbines,* Cambridge, MA: MIT Press.

TURBULENT FLOWS

Peter S. Bernard and James M. Wallace, *Department of Mechanical Engineering, University of Maryland, College Park, Maryland, U.S.A.*

Introduction 399
1. **Preliminaries** 401
1.1 Notational and Mathematical Considerations 401
1.2 Experimental Methods 402
1.3 Vorticity 404
1.3.1 Vorticity Equation 404
1.3.2 Wall Production of Vorticity 405
2. **General Turbulence Characteristics** 405
2.1 Turbulent Transport 405
2.2 Wall-Bounded Turbulent Flows 408
2.2.1 Flow Regions 408
2.2.2 Vortical Structures 411
2.3 Turbulent Kinetic Energy 412
2.4 Boundary-Free Turbulent Shear Flows 414
2.4.1 Similarity and Self-Preservation 415
2.4.2 Turbulence Spreading 416
2.4.3 Some Characteristics of Mixing Layers 416
3. **Theory of Idealized Turbulent Flows** 420
3.1 Isotropic and Homogeneous Turbulence 420
3.2 Homogeneous Shear Flow 421
4. **Turbulence Modeling and Theory** 422
4.1 Closure Schemes 422
4.2 Statistical Closures 423
4.3 Thermodynamics of Vortex Systems 424
4.4 Direct Numerical Simulations 424
4.5 Large-Eddy Simulations 425
5. **Summary** 426
Glossary 426
Works Cited 428
Further Reading 430

INTRODUCTION

Fluids in motion are most often in a complex disordered state that is referred to as turbulent. This behavior results from instabilities in laminar flow when the fluid velocity is increased beyond a critical value. Examples of turbulent fluid motion abound in nature and in technological applications. Sometimes the turbulence is evident, as in a buoyant turbulent plume of smoke billowing from an industrial plant on a cold clear night or in the complex eddies downstream of a bridge abutment. Sometimes it is invisible but obviously present, such as in the discomfort of buffeting on an airplane trip, and on other occasions the turbulence even goes unnoticed, as in the cylinder of an internal combustion engine or in the atmosphere where it has much to do with the daily weather patterns. So too, the white-hot burning gases on the surface of the sun are in turbulent motion, as are the gases between the stars. The Great Red Spot in the atmosphere of Jupiter has been conclusively shown to come from an instability in a turbulent mixing layer.

The need to understand and predict turbulence is a concern in many branches of engineering. Turbulent boundary layers forming on the wings and fuselage of aircraft greatly affect lift and drag characteristics. Pressure fluctuations in turbulent flow may lead to unwanted noise that has to be controlled. One of the most important characteristics of turbulence is its ability to transport mass, momentum, and heat efficiently. This is put to practical benefit by mechanical engineers in heat exchangers and by chemical engineers in mixing vessels. It is a problem to be coped with by environmental engineers concerned with air and water pollution. Even bioengineers have to take turbulence into account; the flow of air and blood in the human body is sometimes turbulent, especially under diseased conditions.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Although observations and sketches of turbulent flows were made by Leonardo da Vinci, the beginning of modern turbulence research required the formulation of the differential equations of fluid motion by Navier and Stokes in the nineteenth century. Osborne Reynolds (1884) was the first to study systematically the transition from laminar to turbulent flow in a pipe. He intuited that a flow can remain laminar only so long as viscous momentum diffusion is large enough to eliminate small disturbances in the flow. The latter continuously appear through natural causes, such as slight irregularities in the shapes of boundaries or in the initial state of the fluid, or even perhaps through microscopic flow perturbations (Tsuge, 1974).

A measure of the relative magnitude of inertial and pressure forces to viscous forces is given by the nondimensional Reynolds number, *Re*. When it is larger than a critical value, viscous forces are insufficient to damp out small disturbances, and they quickly amplify to measurable differences in the evolving velocity and pressure fields, appearing as the random motion associated with turbulence. In engineering practice, very large values of *Re* are common; e.g., $Re \approx 10^9$ in the flow past surface ships. This makes clear how far removed from the laminar state many practical flows are.

Reynolds (1895) later analyzed turbulent motion by decomposing the variables in the Navier–Stokes equations into mean and fluctuating parts. He initiated the practice of averaging the Navier–Stokes equations to derive a set of relations now referred to as the Reynolds-averaged Navier–Stokes equations or RANS. These may be made the basis for directly computing the mean velocity of a flow field without first computing actual realizations of the flow. A fundamental limitation of this approach is that the averaging process introduces additional unknowns into the equations in the form of quadratic products of the velocity fluctuations. These terms account for the turbulent momentum flux, and a separate method must be devised to predict them. Turbulence modeling is concerned with different strategies to approximate turbulent transport rates so that predictions of the mean velocity field can be made.

The problem of how to predict turbulent flow accurately in general circumstances remains for the most part unsolved. Nonetheless, many fundamental aspects of turbulent-flow behavior are reasonably well understood, helping to pinpoint the source of difficulty faced by prediction schemes. A primary problem is that turbulence occurs over a wide range of scales varying from the smallest, where mechanical energy is dissipated into internal energy, to scales as large as that of the bodies containing the turbulence or responsible for its generation. Further complicating the picture is the fact that despite being chaotic, turbulent fluid flow tends to be organized into a hierarchy of vortical structures at different scales that mutually interact with great complexity.

Much of the research in the first half of the twentieth century concentrated on analyzing idealized turbulent flows free of complications caused by anisotropies due to mean shear. Pioneering works by G. I. Taylor (1935), von Kármán and Howarth (1938), and many others introduced statistical analyses, which took advantage of special symmetries to gain an understanding of physical processes such as turbulence decay. Kolmogorov (1941) postulated that small-scale turbulent motions are isotropic at high Reynolds numbers and share common properties independent of how the turbulence is generated. There is a growing body of evidence confirming this hypothesis and delineating the conditions under which it is true (e.g., Saddoughi and Veeravali, 1994; Antonia and Kim, 1994).

Although it has long been known that in some sense turbulent flows have "structure," research in the last several decades has been marked by the discovery that the dynamics of shear flows, including the Reynolds stresses and other important properties, are closely associated with the presence of coherent and apparently self-replicating vortical structures (Wallace, 1990). The first investigations of coherency were almost exclusively experimental, using conditional statistical analysis made feasible by the advent of dedicated laboratory digital data-acquisition systems and computers. In recent years good numerical solutions of both wall-bounded and boundary-free turbulent flows have given an unprecedented view of the dynamics of vortical structures at relatively low Reynolds numbers. In such direct numerical simulations (DNS) all flow variables of interest are determined throughout the full flow domain, permitting novel explorations of the three-dimensional nature of the coherent structures and their relationship to turbulent transport (Bernard *et al.*, 1993).

DNS also represents a means for predicting turbulent flows of engineering interest in which no modeling is involved. Mean flow properties are obtained by averaging over the computed flow realizations. Because such schemes are computationally intensive and require enormous amounts of computer memory, they are severely limited in their range of application. The first successful simulations were performed for relatively simple turbulent flows, such as a channel at low Reynolds numbers (Kim *et al.*, 1987). Flows of somewhat greater complexity are now being studied, such as in a square duct (Huser *et al.*, 1994) or with riblets (Goldstein *et al.*, 1995).

Recent years have seen the rise of methods that are a compromise between direct simulations and RANS computations. Referred to as large eddy simulations (LES), these seek to decompose the velocity field into a filtered part, which is computed as the solution of the filtered Navier–Stokes equation, and a subgrid part that models the small-scale motions. Since the smaller-scale subgrid motions are more isotropic and thus more likely to have a universal character than larger ones, there is some hope that a generally applicable subgrid model can be developed. At the present time, many issues need to be resolved before it will be clear how best this goal can be achieved.

1. PRELIMINARIES

1.1 Notational and Mathematical Considerations

The components of the three-dimensional velocity field corresponding to the (x_1,x_2,x_3) coordinate directions will generally be denoted by U_i, $i = 1, 2, 3$, though on occasion the alternative notation (U,V,W) will be used to indicate velocities in the (x,y,z) directions, respectively. The streamwise direction will be denoted by x while y is the inhomogeneous and z the homogeneous cross-stream direction, respectively. In view of the importance of mean quantities in studying turbulent-flow problems, it is useful to make the Reynolds decomposition of the flow fields into mean and fluctuating parts. For example, the velocity and pressure fields may be written in the form

$$U_i = \bar{U}_i + u_i, \tag{1}$$

$$P = \bar{P} + p, \tag{2}$$

where the overbar denotes an appropriate average, and clearly $\bar{u}_i = 0$ and $\bar{p} = 0$. Averages may be considered to be over ensembles of realizations of the flow field as in

$$\bar{U}(\mathbf{x},t) = \frac{1}{N}\sum_{j=1}^{N} U^j(\mathbf{x},t), \tag{3}$$

where $U^j(\mathbf{x},t)$ is the jth realization out of N. If the mean flow is approximately steady, a time average

$$\bar{U}(\mathbf{x},t) = \frac{1}{T}\int_{t-T/2}^{t+T/2} U(\mathbf{x},s)ds \tag{4}$$

can also be defined, where it is understood that the time interval T is long compared to the time scales of the turbulent field. That the two definitions (3) and (4) give the same mean is likely, but it is not known to be true rigorously.

Products of fluctuating quantities give turbulence correlations. For example, the turbulent kinetic energy is a one-point correlation defined by

$$k = \overline{u_i u_i}/2, \tag{5}$$

where here and elsewhere repeated indices are to be summed for $i = 1, 2, 3$. Relative to the local mean velocity, k gives a useful measure of the local turbulence intensity since, whenever it is zero, the flow must be laminar. Consideration of the relative magnitude of the individual terms in (5), i.e., the variances $\overline{u_1^2}$, $\overline{u_2^2}$, and $\overline{u_3^2}$, gives an idea of the degree to which the local turbulence levels are anisotropic, i.e., have a directional dependence. If the turbulence is homogeneous, then all of the mean flow statistics, including $\bar{U}_i$, are independent of spatial position.

Multiple-point correlations may also be

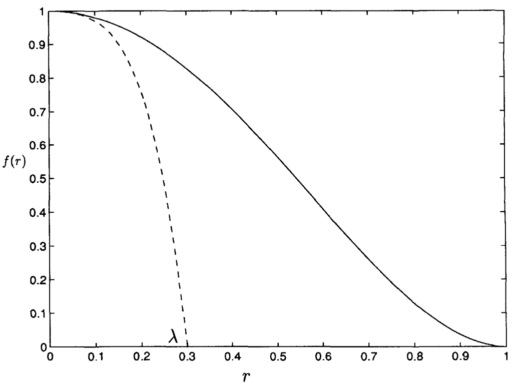

FIG. 1. The two-point longitudinal velocity correlation coefficient $f(r)$ (solid line). The intercept with the r axis of a parabolic approximation to $f(r)$ (dashed line) gives the Taylor microscale λ. The integral scale Λ is defined as the area under $f(r)$.

defined, as in the two-point double and triple correlations given by

$$\mathcal{R}_{ij}(\mathbf{x},\mathbf{r}) = \overline{u_i(\mathbf{x})u_j(\mathbf{x} + \mathbf{r})} \tag{6}$$

and

$$\mathcal{S}_{ijk}(\mathbf{x},\mathbf{r}) = \overline{u_i(\mathbf{x})u_j(\mathbf{x})u_k(\mathbf{x} + \mathbf{r})}, \tag{7}$$

respectively. For simplicity, the dependence on time t is not indicated explicitly in these relations. The particular cases $f(r) \equiv \mathcal{R}_{11}(r\mathbf{e}_1)/\mathcal{R}_{11}(0)$ and $g(r) \equiv \mathcal{R}_{22}(r\mathbf{e}_1)/\mathcal{R}_{22}(0)$ are referred to as the two-point longitudinal and transverse correlation coefficients, respectively. Here, $\mathbf{r} = r\mathbf{e}_i$, $\mathbf{e}_i$ is a unit vector in the x_i direction, $f(0) = g(0) = 1$, and $f(r)$, $g(r) \to 0$ as $r \to \infty$. Correlations such as (6) and (7) open up the possibility of examining structural characteristics of the flow field; i.e., they may reveal the presence of spatial organization to the velocity field that is not apparent from single-point statistics. $\mathcal{R}_{ij}$ and $\mathcal{S}_{ijk}$ also have benefit for characterizing phenomena at different scales. For example, $\mathcal{R}_{ij}$ provides a convenient basis for defining turbulence length scales at opposite extremes. Specifically, the Taylor microscale λ is a measure of small-scale dissipative phenomena, while the integral scale Λ measures the largest distance over which correlation persists in a turbulent flow field as well as the dimensions of the largest structures in the flow. The microscale is defined as the distance from the origin at which a parabolic approximation to $g(r)$ is zero, as illustrated in Fig. 1. For homogeneous flow conditions it is not hard to show that

$$\overline{\left(\frac{\partial u}{\partial y}\right)^2} = \frac{2\overline{u^2}}{\lambda^2}. \tag{8}$$

In contrast, Λ is defined as the integral under the $g(r)$ curve; i.e.,

$$\Lambda \equiv \int_0^\infty g(r)dr. \tag{9}$$

A fruitful way to explore the interactions between phenomena at different scales is through consideration of the energy spectrum tensor, $\mathcal{E}_{mn}(\mathbf{k})$, defined via Fourier transformation of $\mathcal{R}_{mn}(\mathbf{r})$:

$$\mathcal{E}_{mn}(\mathbf{k}) = (2\pi)^{-3} \int_{R^3} d^3 r e^{-i\mathbf{k}\cdot\mathbf{r}} \mathcal{R}_{mn}(\mathbf{r}). \tag{10}$$

Here, $\mathbf{k}$ is the wave-number vector, and possible dependencies of $\mathcal{E}_{mn}$ and $\mathcal{R}_{mn}$ on $\mathbf{x}$ or t are not written explicitly. The inverse Fourier transform implies that

$$\mathcal{R}_{mn}(\mathbf{r}) = \int_{R^3} d^3 k e^{i\mathbf{r}\cdot\mathbf{k}} \mathcal{E}_{mn}(\mathbf{k}). \tag{11}$$

Since $k = \frac{1}{2}\mathcal{R}_{ii}(0)$, (11) can be used to establish that

$$k = \int_0^\infty E(\kappa)d\kappa, \tag{12}$$

where the energy density $E(\kappa)$ is defined in terms of an average of $\mathcal{E}_{mm}$ over shells of radius $\kappa = |\mathbf{k}|$ in wave-number space. Equation (12) shows how the turbulent kinetic energy at each point in a flow can be decomposed among the different length scales as represented by κ.

Evolution equations for k, $\mathcal{R}_{ij}$, $\mathcal{E}_{ij}$, $E(\kappa)$, and other statistical quantities may be derived from the Navier–Stokes equation. Analysis and solutions of these form the basis for many studies of turbulence, as is made clear below.

1.2 Experimental Methods

Because of the impossibility of analytical solutions of the Navier–Stokes equations for turbulent flows and the restriction of DNS solutions to quite low Reynolds numbers for relatively simple geometries, experimental investigations of turbulence are still very much needed. There is a vast body of literature describing experimental results from a wide range of turbulent flows. There is only space here to mention some of the most important findings from these experiments.

Thermal anemometry, first developed in the 1920s, is still the experimental technique most frequently used in turbulence research (Blackwelder, 1981; Fingerson and Freymuth, 1983). The physical idea on which it is based is simple. A small-diameter (typically 1–5 μm) metal wire (usually tungsten, platinum, or platinum alloys) is heated to a temperature higher than the ambient temperature of the flow by an electrical current. The wire is cooled when heat is transferred to the flow, at a rate that depends nonlinearly on the flow velocity. Modern thermal anemometry usually incorporates the sensor wire as one leg of a Wheatstone bridge. The differential bridge voltage imbalance, caused by a differential change in the sensor resistance, is fed back through a high-gain amplifier. This brings the sensor resistance back to its operating condition by forcing more or less current through it. The voltage drop across the sensor, the resistance of which is held nearly constant in this mode of operation, is thus the physical variable measured. It is a nonlinear function of the velocity component cooling the sensor. Thermal anemometry has the advantage of providing long, continuous data records with good frequency response. Its major disadvantages are that it is intrusive, frequent calibrations are required, and additional sensors must be used to obtain more than one component of the velocity vector. Furthermore, at least nine sensors are needed to obtain all the components of the vorticity vector (Wallace and Foss, 1994). Measuring the velocity and/or the vorticity vectors at more that one location simultaneously in the flow requires even more sensors, together with the electronics to operate them. It is generally impractical to obtain such measurements.

In the last three decades various optical methods for measuring turbulence properties have been developed. The earliest of these is laser-Doppler velocimetry (LDV), a good introduction to which is given by Adrian (1983). This method utilizes the Doppler principle to the effect that relative to a particular observer, coherent light from a moving source exhibits a frequency shift depending on its velocity. Thus, the local fluid velocity can be determined from the frequency of the laser light beam scattered from tiny particles, faithfully following the fluid motion. LDV has very good frequency response and is nonintrusive except for the presence of the particles in the flow. The choice of size and density of the particles is important. Particles that are too large do not properly move with the local instantaneous flow field, while particles that are too small follow the Brownian motion of the fluid, responding to the fluid's temperature and not the continuum velocity field. Furthermore, too dense particles will not be neutrally buoyant. These constraints on the size and density of particles are particularly limiting in air flows. With more complex optics, more than one velocity component can be unambiguously sensed. Like thermal anemometry, however, use at more than one location in the flow requires a great increase in complexity and expense. Another drawback is that LDV data are obtained at irregular intervals depending on the random passage of particles through the measurement volume.

Recently, optical methods have been developed that rely on image processing of visualized particles to provide information about the velocity vector field projected onto full spatial planes of turbulent flows. Many of these methods, including particle tracking, particle imaging, digital particle imaging, and speckle velocimetry have been discussed and evaluated by Adrian (1991). Such methods also are conceptually simple. Most straightforwardly, in particle tracking velocimetry (PTV), where the particle number density is small, individual particles can be identified and their individual displacements tracked over short time intervals by a computer. With this information, the velocity components in the visualized planes can be calculated. With higher particle number densities, the displacement of local groups of particles can be tracked over two successive images using spatial correlations. This is commonly called particle imaging velocimetry (PIV). With even higher number densities, speckle patterns are formed from the phase differences of the overlapping images of individual particles. These local speckle patterns move with the turbulent motion and can be tracked in what is known as laser speckle velocimetry (LSV). Such particle motion methods are now being extended to three dimensions with holographic and stereoscopic imaging (Scherer and Bernal, 1988; Prasad and Adrian, 1993; Barnhart *et al.*, 1994; Meng and Hussain, 1995).

Even more recently, methods that invert the transport equation for a passive conserved scalar,

$$\frac{\partial C}{\partial t} + U_i \frac{\partial C}{\partial x_j} = \frac{1}{ReSc} \frac{\partial^2 C}{\partial x_j \partial x_j}, \tag{13}$$

where *Sc* is the Schmidt number, have been developed (Dahm *et al.*, 1992; Su and Dahm, 1996; Pearlstein and Carpenter, 1995). These methods require extremely high spatial and temporal resolution and imaging accuracy because the spatial and temporal concentration derivatives in Eq. (13) must be obtained from images of the rapidly evolving turbulent scalar. Furthermore, the velocity field computed from (13) is not unique. Additional constraints must be imposed on the solution in order to be assured of finding the physical velocity field.

Generally speaking, the same temporal and spatial resolution requirements in thermal anemometry also apply to the optical measurement methods. With the light sources, optics, film, and video technology available at present, it has been particularly difficult for imaging methods to measure the temporal evolution of turbulent flows.

1.3 Vorticity

The presence of prodigious amounts of vorticity is among the defining properties of turbulent flow. Regions of flow with significant vorticity are very often turbulent, whereas regions without vorticity are never turbulent. Understanding the dynamics of the vorticity field is indispensable to understanding turbulent flows. The vorticity vector, defined as the curl of the velocity vector, is expressed in tensorial component form as

$$\Omega_i = \epsilon_{ijk} \frac{\partial U_k}{\partial x_j}, \tag{14}$$

where ϵ_{ijk} is the alternating tensor. It may be shown using a vector identity that Ω_i is solenoidal, i.e., $\nabla \cdot \boldsymbol{\Omega} = 0$.

The vorticity corresponds to twice the instantaneous rate of rotation of a material fluid particle. Alternatively, given a surface enclosed by a small circuit in the fluid, the vorticity component normal to the surface at an arbitrary point may be interpreted as the circulation per unit area of the surface, where circulation is defined as the line integral of the tangential component of the velocity vector around the closed curve. This interpretation of vorticity comes from the fact that by the Stokes and divergence theorems, circulation is the integral of the normal vorticity component over an arbitrary surface bounded by a given circuit. By constructing the collection of vortex lines passing through a circuit, a vortex tube may be created. Its circulation is referred to as its "strength." By the divergence theorem, tube strength is constant for any closed circuit enclosing the vortex tube. The local vorticity increases inversely proportional to the cross-sectional area of the tube—as a reflection of conservation of angular momentum. On a solid surface the normal component of vorticity is zero, so that vortex lines must lie in the plane of the wall and are perpendicular to the streamlines there. Away from a bounding surface, vortex lines and streamlines are only orthogonal for two-dimensional and some axisymmetric flows and not for turbulent flows, which are inherently three dimensional. It also may be shown that vortex lines cannot end in the fluid. Instead, they must either form closed loops, extend to infinity, or intersect the wall at locations where the vorticity is zero.

1.3.1 Vorticity Equation Taking the curl of the Navier–Stokes equation gives the vorticity equation

$$\frac{\partial \Omega_i}{\partial t} + U_j \frac{\partial \Omega_i}{\partial x_j} = \Omega_j \frac{\partial U_i}{\partial x_j} + \nu \frac{\partial^2 \Omega_i}{\partial x_j \partial x_j}, \tag{15}$$

where ν is the kinematic viscosity. The left-hand side expresses the rate of change of the vorticity of fluid particles along Lagrangian paths. The first term on the right-hand side accounts for changes of Ω_i caused by stretching or reorientation of vortex lines, and the last one for effects of viscous diffusion. In inviscid flow, (15) can be used to prove Kelvin's theorem to the effect that the circulation of a material loop of fluid particles is conserved in time, as well as Helmholtz's theorem that vortex lines are also material lines in the absence of viscosity.

Finally, the conspicuous absence of pressure in the vorticity equation should be noted. Pressure and gravity forces act through the center of mass of particles and thus apply no moments to them. Only shear forces can change the vorticity of a fluid particle in incompressible flow.

1.3.2 Wall Production of Vorticity New vorticity entering an incompressible flow must originate at solid surfaces. These may be the bounding walls of confined flows or the nozzles, bluff bodies, or splitter plates that are the upstream source of free jets, wakes, or mixing-layer flows, respectively. If the Navier–Stokes equation is evaluated *at* the bounding surface (defined for simplicity here as the x-z plane), it reduces to (Lighthill, 1963)

$$\frac{\partial p}{\partial x} = -\mu \frac{\partial \Omega_z}{\partial y} \tag{16}$$

and

$$\frac{\partial p}{\partial z} = -\mu \frac{\partial \Omega_x}{\partial y}, \tag{17}$$

where μ is the dynamic viscosity. The right-hand sides of these equations represent the fluxes of the Ω_x and Ω_z components of vorticity normal to the wall. The wall-normal flux of Ω_y can be obtained from the solenoidal property of the vorticity vector and is given by

$$-\mu \frac{\partial \Omega_y}{\partial y} = \mu\left(\frac{\partial \Omega_x}{\partial x} + \frac{\partial \Omega_z}{\partial z}\right). \tag{18}$$

Thus, although the time rate of change of the vorticity of a particle within the flow is not affected by the pressure variations in the flow, the flux of new vorticity into the flow at the bounding surface depends on the pressure gradient instantaneously occurring over the surface.

2. GENERAL TURBULENCE CHARACTERISTICS

2.1 Turbulent Transport

The eddying motion associated with turbulence is considerably more effective in transporting momentum, vorticity, internal energy, scalar contaminants, and other flow properties than is molecular diffusion. Heightened transport rates near boundaries are of particular practical importance, as they lead to increased drag and augmented heat transfer. From the standpoint of turbulent flow prediction techniques based on solving directly for the mean velocity field, the ability to estimate turbulent transport rates is of primary importance. This may be seen from the RANS equations:

$$\rho\left(\frac{\partial \bar{U}_i}{\partial t} + \bar{U}_j \frac{\partial \bar{U}_i}{\partial x_j}\right) = -\frac{\partial \bar{p}}{\partial x_i} + \frac{\partial}{\partial x_j}\left(\mu \frac{\partial \bar{U}_i}{\partial x_j} - \rho\overline{u_i u_j}\right) \tag{19}$$

(where ρ is density), which could be solved for $\bar{U}_i$ and the pressure $\bar{p}$ if only a theory for predicting the additional Reynolds stress term, $R_{ij} \equiv -\rho\overline{u_i u_j}$, were available. Classical efforts at predicting it have used its physical interpretation as the turbulent flux of the ith component of momentum in the jth direction to assume an analogy between molecular and turbulent transport mechanisms. The analogy is imperfect, however, which has led to considerable inaccuracies in mean flow prediction when this route is followed.

To see why the molecular analogy is poor, it is instructive to consider in some detail the physical properties of the Reynolds stresses. This may be most easily viewed in the context of channel flow where (19) simplifies to

$$0 = -\frac{\partial \bar{p}}{\partial x} + \frac{d}{dy}\left(\mu \frac{d\bar{U}}{dy} - \rho\overline{uv}\right). \tag{20}$$

Here, only one component of the Reynolds stress representing (minus) the flux of streamwise momentum in the y direction (perpendicular to the channel walls) acts in addition to the mean viscous shear stress. Together they comprise the total stress, $\tau \equiv \mu d\bar{U}/dy - \rho\overline{uv}$. In laminar flow, $\overline{uv} = 0$ and the solution to (20) is the well-known Poiseulle flow. In the turbulent case, (20) is not sufficient to determine $\bar{U}$ unless $\overline{uv}$ can be ascertained independently.

In the kinetic theory of nondense gases (Chapman and Cowling, 1952), the relationship $\sigma_{12} = \mu dU/dy$ for the viscous shear stress is derived through consideration of molecular momentum transport. In this case $\sigma_{12} = -\rho\langle c_1 c_2\rangle$, where $c_i = C_i - U_i$ is the molecular fluctuation velocity, C_i is the velocity of a molecule, $U_i = \langle C_i\rangle$ is the hydrodynamic fluid velocity, and $\langle\ldots\rangle$ denotes averaging over a small sensing volume containing large numbers of molecules. To a very high degree of approximation, it may be shown that

$$-\rho\langle c_1 c_2\rangle = \mu\frac{d\langle C_1\rangle}{dy} = \mu\frac{dU_1}{dy}, \tag{21}$$

where the viscosity $\mu \sim \rho\lambda_m\langle C_i C_i\rangle^{1/2}$ and λ_m is the mean free path. The derivation of (21) involves formally evaluating the effect of random molecular migrations of momentum in the presence of an inhomogeneity in the y direction of the average rectilinear particle speed. Since turbulent transport is ostensibly a similar sort of phenomenon, namely, caused by random movements of momentum within a mean field—albeit at a very much larger scale—it has been thought that a generalization of the argument leading to (21) would be applicable in this case as well.

In fact, it can be shown that three essential properties of molecular transport used in deriving (21) are not satisfied in the turbulent case. In particular, (21) incorporates a well-defined mixing time as the time between molecular collisions; it assumes to good approximation that molecules carry their momentum undisturbed over the distance λ_m before exchanging it with the surrounding molecules; finally, λ_m is exceedingly small compared to the distance of greater-than-linear variation of the bulk fluid velocity. In contrast, for the turbulent case, a useful definition of the mixing time is not self-evident; nor can it be expected that momentum is conserved, even approximately, on fluid particle paths; and the distance over which mixing occurs is more than likely to be greater than that over which the mean field varies linearly.

To get more insight into the nature of turbulent transport, it is helpful to examine how the u and v correlation underlying the Reynolds stress gets established as a fluid particle moves in the mean velocity field of a turbulent flow. In particular, consider the identity

$$\overline{u_a v_a} = \overline{u_b v_a} + \overline{v_a(\bar{U}_b - \bar{U}_a)} + \overline{v_a(U_a - U_b)}, \tag{22}$$

where a is a given, unchanging point in the flow, and b denotes the position of a particular fluid particle in a particular realization at time $t - \tau$ ($\tau > 0$), which arrives at a at time t, as illustrated in Fig. 2. The correlation $\overline{u_b v_a}$ appearing on the right-hand side of (22) is between velocity components at different times and will be zero for sufficiently large τ. The turbulent mixing time may be defined as the minimum time interval at which u_b and v_a become decorrelated; i.e., $\overline{u_b v_a} = 0$. Thus τ may be viewed conceptually as the time necessary for the correlation between u_a and v_a to be created by the physical processes represented by the last two terms on the right-hand side. The first describes displacement transport due to fluid particles carrying the average momentum of their starting point, namely, $\bar{U}_b$, to point a over the mixing time; the second accounts for the effect on transport of changes in momentum, i.e., acceleration effects, along particle paths.

The term $\overline{v_a(\bar{U}_b - \bar{U}_a)}$ is a concise mathematical statement of the physical transport mechanism that historically has been modeled through the molecular analogy using a

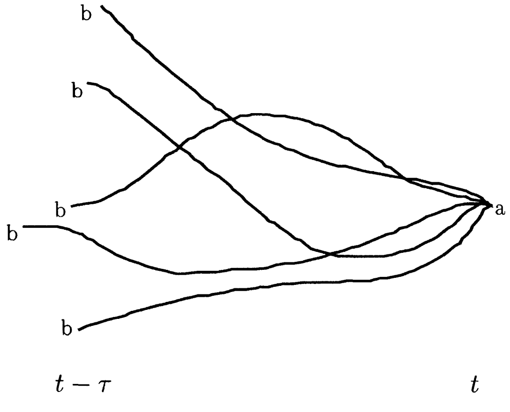

FIG. 2. Backward fluid particle paths over time interval τ between randomly located points b and fixed point a.

gradient formula. As it appears here, however, it is not limited by the gradient assumption. The latter may be deduced as an approximation if local linearity of $\bar{U}$ can be assumed. For example, in the case of a unidirectional mean flow $\bar{U}(y)$ if the Taylor series expansion

$$\bar{U}_b = \bar{U}_a - \int_{t-\tau}^{t} v(s)ds \, \frac{d\bar{U}}{dy} \tag{23}$$

is a good approximation to $\bar{U}_b$ over time τ, then it follows that

$$\overline{uv} = -T\overline{v^2}\,\frac{d\bar{U}}{dy}, \tag{24}$$

where $T \equiv \int_{t-\tau}^{t} \overline{v(0)v(s)}/\overline{v^2}$ is a Lagrangian integral time scale. In these formulas, $v(s)$ represents the Lagrangian velocity at time $s < t$ of the fluid particle in each realization that is at point a at time t, so that $v(t) = v_a$. The expression $\int_{t-\tau}^{t} v(s)ds$ in (23) is the vertical distance traveled by a fluid particle between points b and a. Clearly, (24) shows that the turbulent momentum-transport coefficient, usually called the eddy viscosity, should be given by $T\overline{v^2}$.

The properties of (22) have been examined using ensembles of fluid particle paths computed in a DNS of turbulent channel flow (Bernard and Handler, 1990; Handler *et al.*, 1992). Figure 3 shows the variation of the individual terms in (22) in the wall region. It may be noticed that the mixing term $\overline{u_b v_a} \approx 0$, indicating that the fluid particle paths were tracked at least as long as the mixing time. While displacement transport of the classical type, i.e., given by $\overline{v_a(\bar{U}_b - \bar{U}_a)}$, was found to be the dominant contributor to $\overline{uv}$, acceleration effects were not negligible. Near the wall, the latter acts to reduce Reynolds stress through deceleration of fluid particles approaching the surface, while further away from the wall it acts as an additional source of Reynolds stress. Both terms in (22) depend on physical factors occurring over a significant nonlocal region of the flow, implying that (24) is not well satisfied in this case.

A more complete understanding of the physical origins of turbulent transport requires unraveling the causes of the fluid motions leading to Reynolds stress. Considerable evidence for the dominant role of self-replicating quasistreamwise vortices in the dynamics of the wall-bounded flows now exists—as is discussed in Sec. 2.2.2. A definite connection between these and Reynolds stress has been established (Bernard *et al*, 1993) using particle paths as an intermediary. The streamwise vortices are directly responsible for the production of Reynolds stress by instigating transfer of high-speed fluid toward the wall or ejecting low-speed fluid away from the wall. Acceleration transport is linked to motions induced by those vortices in which fluid particles are accelerated as they travel over vortex cores or decelerated while traveling under them. Similar processes appear to occur in boundary-free flows, as is discussed in Sec. 2.4.3.

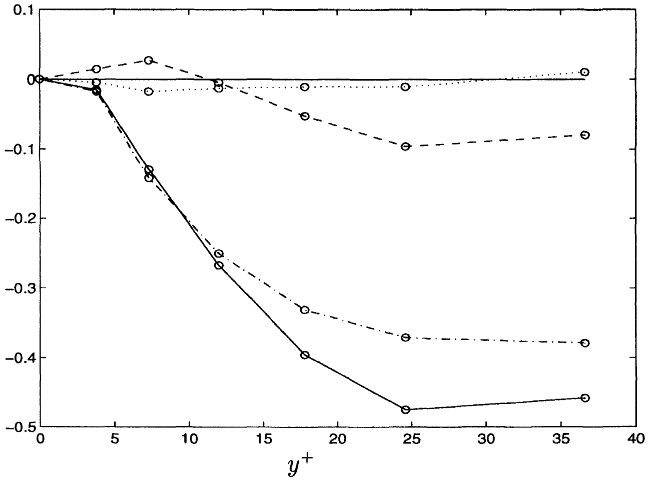

FIG. 3. Near-wall behavior of terms in Reynolds stress decomposition (22): $\overline{u_b v_a}$, · · ·; $\overline{v_a(\bar{U}_b - \bar{U}_a)}$, —·—; $\overline{v_a(U_a - U_b)}$, – – –; $\overline{uv}$, ———. Results from channel-flow DNS of Bernard *et al.* (1993) at $R_\tau = 250$.

The more the physics surrounding turbulent transport is understood, the more it is clear that its successful modeling requires taking into account nonlocal features of the flow. Chief among these is the presence of flow structures, including coherent vortices. Resolution of the self-replicating structures may be a necessary and sufficient condition for a successful prediction scheme, since the chief transport will have been taken into account; i.e., it seems likely that small-scale motions do not contribute much to the net turbulent transport. If this is the case, then LES of one sort or another may have some hope of being successful in predicting complex turbulent flows.

Not long after the development of the RANS approach, it was noted by Taylor (1915) that the closure problem could be equally well framed in terms of vorticity transport instead of momentum transport. This involves rewriting the RANS equations in such a way as to replace the Reynolds stress with the vorticity transport correlation tensor $\overline{u_i\omega_j}$. Taylor reasoned that this would be beneficial if vorticity transport was better described by a gradient law than momentum transport. The fact that vorticity is conserved on particle paths, in two-dimensional inviscid flow, suggested that this may be the case. In practice, Taylor (1932) noticed that gradient vorticity-transport laws cannot well account for vorticity transport near boundaries. In fact, it has recently been shown (Bernard, 1990) that vorticity transport is counter-gradient near boundaries and that this can be well explained by a modified form of the Lagrangian approach pursued by Taylor. In this, the effects of three-dimensional vortex stretching and reorientation are included in the transport law. This Lagrangian approach has also been put to practical use (Gorski and Bernard, 1995) in computing boundary-layer flows.

The transport issue is also faced in problems involving heat and mass transfer in turbulent flow. Here, Reynolds decomposition and averaging of the scalar equation (13) introduces the turbulent scalar flux correlation, $\overline{uc_i}$, where c denotes either temperature or concentration fluctuations of scalar contaminants. The deficiencies of gradient scalar-transport models are well known—particularly near sources. The various manifestations of turbulent transport require a nonlocal treatment to capture their complete physical origin. This is reflected in several recent approaches (Egolf, 1994; Dekker *et al.*, 1995; Bernard and Rovelstad, 1994; Bernard, 1995).

2.2 Wall-Bounded Turbulent Flows

Through numerous experiments and some recent computational work, a considerable amount has been learned about the distribution of velocity and Reynolds stress in flows near solid boundaries. This section describes some of the principal features of flows in boundary layers, channels, ducts, and pipes. Wall-bounded flows will be considered here only in their "fully developed" state, which means that their statistical properties do not change in the mean flow direction. That is to say all $\partial/\partial x$ gradients of average properties are zero. This condition is only an approximation for boundary layers that continuously develop downstream; however, in the wall layer, the mean flow is nearly parallel to the wall.

Experimental results characterizing wall-bounded flows are of two types. Earlier measurements, most of which have been confirmed or improved upon for low Reynolds numbers by DNS, were almost exclusively statistical analyses of long time series of velocity components and wall pressure fluctuations. As mentioned above, in the late 1950s and early 1960s, experimental evidence began to appear indicating that the properties of these flows were not simply a result of completely random turbulent fluctuations. Rather, it was observed that these flows had certain organized or "coherent" features that strongly influenced their statistical behavior. Much of the experimental research over the last 30 years has been an attempt to detect and characterize these coherent properties in the hope that this will lead to more physically realistic transport models.

2.2.1 Flow Regions The traditional division of wall-bounded turbulent flows is into four regions, illustrated in the sketch in Fig. 4 by Reynolds (1974). The thicknesses of the two layers nearest the wall (I and II), together denoted as the viscous layer, are exaggerated for clarity.

1. *Viscous Sublayer.* Here the mean velocity

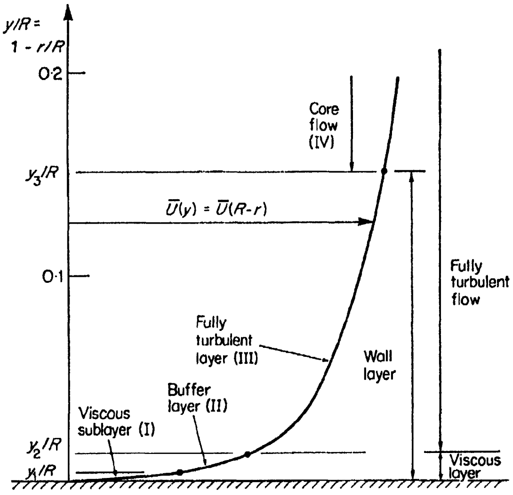

FIG. 4. Turbulent pipe (of radius R) flow velocity profile illustrating the (I) viscous sublayer, (II) buffer layer, (III) fully turbulent layer, and (IV) core-flow region. The thicknesses of layers (I) and (II) are exaggerated for clarity. From *Turbulent Flows in Engineering* by A. J. Reynolds, Copyright © 1974, John Wiley & Sons, Ltd. Reprinted by permission of John Wiley & Sons, Ltd.

varies linearly, i.e., $d\bar{U}/dy$ is constant, and the Reynolds shear stress, $-\rho\overline{uv}$, is very small as a result of the suppression of the wall-normal fluctuation, v, by the presence of the wall. The "intensity" of streamwise turbulence, $\sqrt{\overline{u^2}}/\bar{U}$, is largest in the viscous sublayer with a nearly constant value of 0.4 throughout (Alfredsson *et al.*, 1988). Thus the name "laminar sublayer," as this flow region was earlier called, is a misnomer. Velocity and length scales can be obtained from ρ, μ, and from the shear stress at the wall, $\tau_w \equiv \mu\partial\bar{U}/\partial y_{\text{wall}}$, in the form

$$u_\tau \equiv (\tau_w/\rho)^{1/2} \tag{25}$$

and

$$y_\tau \equiv \mu/(\rho\tau_w)^{1/2} = \nu/u_\tau. \tag{26}$$

For obvious reasons, u_τ is often called the "friction" velocity. If we scale the distance normal to the wall with y_τ, the viscous sublayer is usually defined as $y^+ \equiv y/y_\tau < 5$. At the outer edge of this region $d\bar{U}/dy$ begins to depart from linearity because of the increasing influence of the Reynolds stresses. In the viscous sublayer where the independent variables affecting the mean velocity variation are τ_w, μ, and y, dimensional analysis yields the linear relation

$$U^+ = y^+, \tag{27}$$

where $U^+ \equiv \bar{U}/u_\tau$. As noted by Reynolds (1974), for the pipe-flow data of Laufer (1953) at $Re_d = U_c d/\nu = 5 \times 10^5$, where U_c is the pipe centerline velocity and d its diameter, $-\rho\overline{uv}$ is approximately 10% of the wall shear stress at the edge of the viscous sublayer. For this flow the sublayer occupies only about 0.05% of the pipe radius, but the mean velocity at $y^+ = 5$ is about 18% of U_c. This clearly illustrates how large the mean velocity gradient is in this region.

2. *Buffer Layer.* The region $5 \leq y^+ \leq 30$ between the viscous sublayer and the logarithmic layer is referred to as the buffer region. Viscosity plays an asymptotically diminishing—but not negligible—role as the upper part of this region blends into region III. Thus, region II is included with region I as part of the viscous layer in Fig. 4. At $y^+ = 30$, for the $Re_d = 5 \times 10^5$ pipe-flow data of Laufer (1953), $-\rho\overline{uv}$ is almost 90% of the mean wall shear stress τ_w, and $\bar{U}$ is approximately 47% of U_c. However, the upper edge of the buffer layer is still only about 0.3% of the pipe radius. Thus, at the upper limit of the buffer layer, the mean velocity reaches almost $\frac{1}{2}$ of its maximum centerline value, and the total shear stress is dominated by the Reynolds shear stress.
3. *Logarithmic Layer.* In this layer, where the Reynolds shear stress is much greater than the viscous stress, the independent variables affecting the mean velocity distribution are τ_w, ρ, and y. Dimensional analysis yields a logarithmic mean velocity distribution, often called the "law of the wall" or "log law,"

$$U^+ = C_1 \ln y^+ + C_2, \tag{28}$$

where the constant $C_1 = 1/\kappa \approx 2.5$ (the inverse von Kármán constant) and $C_2 \approx 5.0$ (Coles and Hirst, 1968). This layer extends from $y^+ \approx 30$ to the location (depending on Reynolds number) where the mean velocity distribution no longer exhibits a logarithmic profile. Taken together, the three regions closest to the wall (I, II, and III in Fig. 4) are often called the *wall layer.* The wall layer has similar properties for all types of bounded flows, including boundary layers, channels, and pipes. The total shear stress τ, i.e., the sum of the viscous and Reynolds shear stress, varies slowly in this layer, as is seen in Fig. 6, and so it often also has been called the constant-stress layer. Recently, Panton (1996) used a matched asymptotic expansion analysis of wall flows for the inner (layers I, II, and III) and outer (layers III and IV) regions, earlier divided by Coles (1956) in this manner, where, respectively, viscous stresses and Reynolds shear stresses dominate. In this view the logarithmic region is an overlap region for which Panton has derived a Reynolds-number–dependent law for the Reynolds shear stress, scaled by the friction velocity:

$$-\frac{\overline{uv}}{u_\tau^2} = 1 - \frac{1}{\kappa y^+} - \frac{y^+}{Re_\tau}. \qquad (29)$$

Here $Re_\tau \equiv u_\tau \delta / \nu$ and δ is the boundary-layer thickness, channel half-width, or pipe radius.

4. *Core Region.* This is the outermost region of bounded flows. The physical processes occurring in this region are different for channels, ducts, and pipes compared to boundary layers. In the last it has the character of a boundary-free turbulent flow and is sometimes called the wake region, because the deficit velocity distribution, $[\bar{U}(y) - U_\infty]/u_\tau = f(y/\delta)$ (where U_∞ is the velocity in the irrotational free-stream region outside the boundary layer), is much like that of the wake of a bluff body. In the logarithmic (overlap) region (III) this velocity deficit distribution, as $y/\delta \rightarrow 0$, is given by

$$\frac{\bar{U} - U_\infty}{u_\tau} = C_1 \ln\frac{y}{\delta} + C_3. \qquad (30)$$

Here in the core or wake region IV, the "law of the wake" (Coles, 1956) applies and can be written as

$$W\left(\frac{y}{\delta}\right) = \frac{\bar{U} - U_\infty}{u_\tau} - \left(C_1 \ln\frac{y}{\delta} + C_3\right), \qquad (31)$$

with $C_3 \approx -2.7$ for a zero–pressure-gradient boundary layer (Panton, 1996). Values of C_3 can be chosen for which the function $W(y/\delta)$ fits the core-region velocity distribution of channel and pipe flows as well, in spite of the different physical processes occurring there for these flows as compared to boundary layers. The core region of a pipe is influenced by structures radially emanating from around the entire pipe circumference (Sabot and Comte-Bellot, 1976). In a high–aspect-ratio duct (wide in one cross-stream direction and narrow in the other, i.e., a channel), the core flow seems to be made up of large-scale motions emanating from opposite walls and "locking" together in the duct core region (Dean and Bradshaw, 1976).

When plotted in semilogarithmic form in Fig. 5, the mean velocity distributions in these four layers are clearly different. The Reynolds number dependence of the upper extent of the logarithmic layer and the concomitant beginning of the core layer also is quite evident.

Experimental measurements of the distribution of the Reynolds shear stress in the

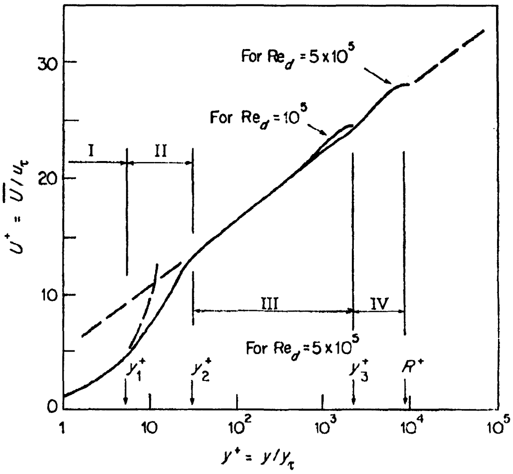

FIG. 5. Turbulent pipe-flow mean velocity distribution, normalized by wall-layer velocity and length scales, for data of Laufer (1953), which illustrates the (I) viscous sublayer, (II) buffer layer, (III) fully turbulent layer, and (IV) core-flow region. From *Turbulent Flows in Engineering* by A. J. Reynolds, Copyright © 1974, John Wiley & Sons, Ltd. Reprinted by permission of John Wiley & Sons, Ltd.

wall layer of a turbulent boundary layer at $R_\theta = 2420$, where θ is the momentum thickness, are shown in Fig. 6. These are spatially and temporally highly resolved LDV measurements of Karlsson and Johansson (1988). Compared to them in Fig. 6 are the DNS results of Spalart (1988) at $R_\theta = 300$, 600, and 1410. Also shown in the figure is the distribution of the total stress, i.e., $\mu d\bar{U}/dy - \rho\overline{uv}$. The observations made above about the varying relative proportions of the viscous and turbulent shear stresses in the different parts of the wall layer is clearly evident. Also visible is the Reynolds number dependence of the Reynolds shear-stress distribution accounted for by Eq. (29).

The Reynolds shear stress, $-\rho\overline{uv}$, is a covariance that is related to the joint probability density function (JPDF), $P(u,v)$, by

$$-\rho\overline{uv} = -\rho \iint_{-\infty}^{\infty} uvP(u,v)\,du\,dv. \qquad (32)$$

This integral represents the sum of contributions to $-\rho\overline{uv}$ from individual (u,v) pairs, weighted by how frequently they occur. The contours of $P(u,v)$ as well as of the covariance integrand, $uvP(u,v)$, shown in Fig. 7 from Ong (1992) for $y^+ = 35$, which agree well with those of Wallace and Brodkey (1977), demonstrate that "ejections" or quadrant-2 (Q2: $u < 0$, $v > 0$) events and "sweeps" or quadrant-4 (Q4: $u > 0$, $v < 0$) events are the dominant contributors to the Reynolds shear stress just above the buffer layer. In fact, nearer the wall, for $y^+ \leq 15$, Wallace *et al.* (1972) showed that Q2 (ejection) events dominate, while further from the wall Q4 (sweep) events take over as the major contributors.

2.2.2 Vortical Structures The same type of JPDF analysis considered in the previous section can be applied to the vorticity covariance given by

$$\overline{\Omega_x\Omega_y} = \iint_{-\infty}^{\infty} \Omega_x\Omega_y P(\Omega_x,\Omega_y)\,d\Omega_x\,d\Omega_y. \qquad (33)$$

Figure 8 shows $P(\Omega_x,\Omega_y)$ and the covariance integrand $\Omega_x\Omega_y P(\Omega_x,\Omega_y)$ measured by Ong (1992) at $y^+ = 35$. Clearly, the vorticity field has a preferred orientation of like-signed Ω_x and Ω_y inclined downstream at angles to the wall.

Robinson (1991) interrogated the $R_\theta = 670$ DNS database of Spalart (1988) to search for and characterize vortical structures important to the turbulence stress dynamics. Assuming that the pressure would be lower in the cores of vortices than in the surrounding fluid, Robinson (1991) detected vortical structures of the type shown in Plate 1, where two "loop"-like vortices are seen with "legs" stretched back at shallow angles with respect to the wall. Also shown are regions of high Q2 and Q4 Reynolds shear stress, ($u'v'2$ and $u'v'4$ in the plate). It is quite apparent that the turbulent stress has a strong spatial association with the vortices. The Q2 motions are primarily located on the inner side of the vortex legs and on the upstream side of the vortex "head," i.e., in regions of upwelling motion of low-speed fluid. On the other hand, the Q4 motions occur primarily on the outer side of the vortex legs where the fluid is swept down toward the surface.

In view of their importance to Reynolds stress production, a number of studies (Brooke and Hanratty, 1993; Bernard *et al.*,

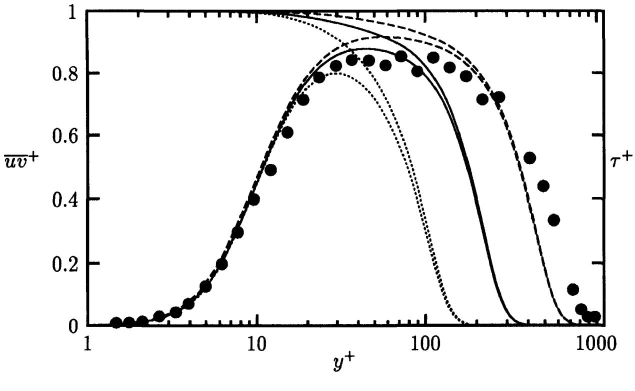

FIG. 6. Distribution of the Reynolds shear stress, $\overline{uv}$, and total stress, τ^+, in turbulent boundary layers. Courtesy of Spalart's (1988) DNS: (dotted lines) $R_\theta = 300$, (solid lines) $R_\theta = 670$, (dashed lines) $R_\theta = 1410$, and $\overline{uv}^+$ data from experiment of Karlsson and Johanson (1988): (●) $R_\theta = 2420$.

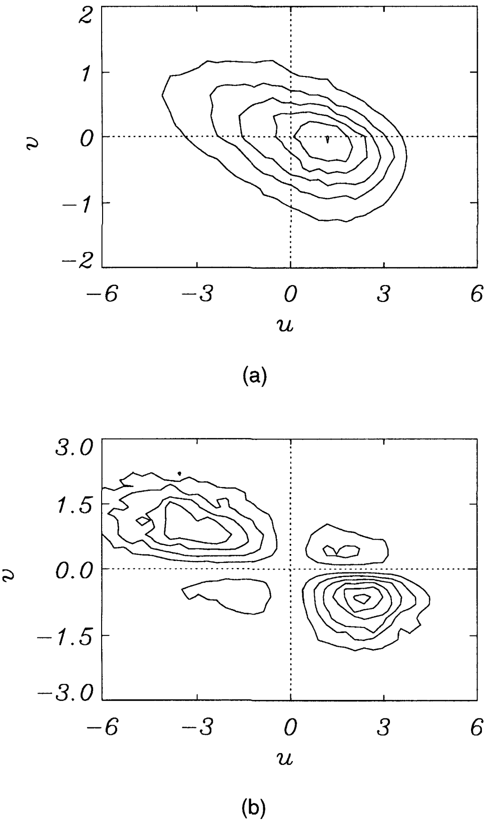

(a)

(b)

FIG. 7. (a) Joint probability density distribution $P(u,v)$, and (b) covariance integrand $uvP(u,v)$ at $y^+ = 35$ in a turbulent boundary layer at $R_\theta = 1070$. Contour values increase toward the interior in intervals of (a) 1.4×10^{-3} and (b) 8.6×10^{-4}. Courtesy of Ong (1992).

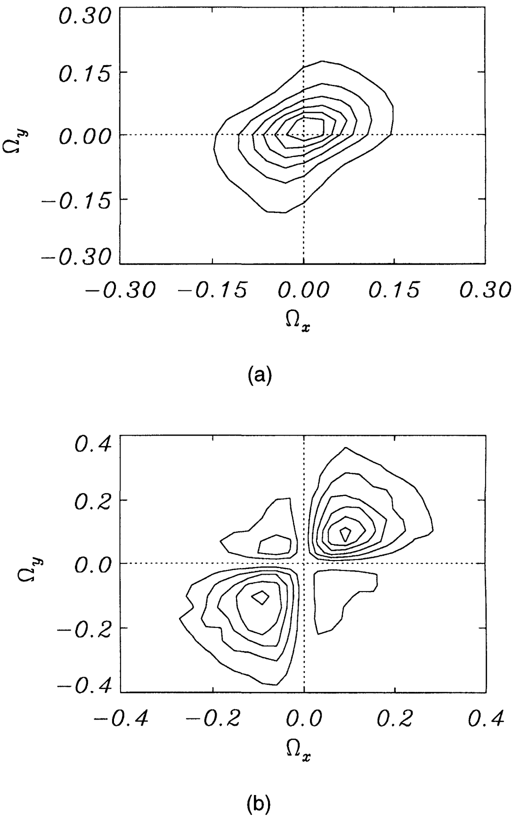

(a)

(b)

FIG. 8. (a) Joint probability density distribution $P(\Omega_x,\Omega_y)$, and (b) covariance integrand $\Omega_x\Omega_y P(\Omega_x,\Omega_y)$ at $y^+ = 35$ in a turbulent boundary layer at $R_\theta = 1070$. Contour values increase toward the interior in intervals of (a) 4.3×10^{-3} and (b) 1.1×10^{-5}. Courtesy of Ong (1992).

1993; Miyake *et al.,* 1995) have attempted to discover the mechanisms by which quasi-streamwise vortices are able to reproduce in the near-wall turbulent-flow environment. Though not agreeing in every detail, the different approaches have made many common observations. New vortices appear as a result of the influence of pre-existing structures of sufficient size, nearness to the boundary, and intensity of rotation. They are often nested, with the uplifted downstream parts of new vortices convecting over mature vortices lying close to the wall. Parent vortices induce new structures as a result of reorientation of spanwise vorticity into the wall-normal direction followed by its shearing into the streamwise direction. Streamwise vorticity produced at the wall surface to satisfy the no-slip condition in the lateral direction may also develop into new vortices on the upwash side of parent vortices. Individual vortices interact with other vortices at different locations along their length. This limits the time scale over which any individual structure can be expected to stay as a coherent entity. Through mutual induction, there appears to be a tendency for vortices to convect away from the wall and populate the outer layers of the wall region.

2.3 Turbulent Kinetic Energy

In investigating the physics of turbulent flow, it proves to be helpful to consider the exact governing equation for second-order moments. In particular, the Reynolds stress equation has the form

$$\frac{\partial R_{ij}}{\partial t} + \bar{U}_k \frac{\partial R_{ij}}{\partial x_k} = -R_{ik}\frac{\partial \bar{U}_j}{\partial x_k} - R_{jk}\frac{\partial \bar{U}_i}{\partial x_k} + \Pi_{ij} - \epsilon_{ij} - \frac{\partial \beta_{ijk}}{\partial x_k} + \nu\nabla^2 R_{ij}, \quad (34)$$

where

$$\Pi_{ij} \equiv \overline{p\left(\frac{\partial U_i}{\partial x_j} + \frac{\partial U_j}{\partial x_i}\right)}, \quad (35)$$

$$\epsilon_{ij} = 2\nu \overline{\frac{\partial U_i}{\partial x_k}\frac{\partial U_j}{\partial x_k}}, \quad (36)$$

$$\beta_{ijk} \equiv \overline{u_i u_j u_k} + \overline{p u_i}\delta_{jk} + \overline{p u_j}\delta_{ik}, \quad (37)$$

and for the purposes of the present discussion we take $R_{ij} = \overline{u_i u_j}$. The first two terms on the right-hand side of (34) represent production of Reynolds stress and the last its viscous diffusion. Equation (35) is the pressure–strain-rate correlation, (36) is the dissipation-rate correlation, and (37) forms the turbulent diffusion term. The properties of the terms of (34) have been studied by Mansour *et al.* (1988), using a channel-flow DNS.

Equation (34) yields relations for the individual normal stresses $\overline{u_i^2}$ ($i = 1, 2, 3$) or, upon contraction, an equation for the turbulent kinetic energy, k. Since $R_{ij}(\mathbf{x},t) = \mathcal{R}_{ij}(\mathbf{x},0,t)$, (34) is a special case of an even more general relation that can be derived for $\mathcal{R}_{ij}(\mathbf{x},\mathbf{r},t)$. The latter is quite complex and is not ordinarily used as an avenue toward the understanding of general flows. However, for simplified flows such as homogeneous isotropic turbulence, it has been extensively used, as is discussed in Sec. 3.1.

An understanding of the k field has intrinsic worth in the study of many turbulent phenomena such as flame propagation and sound generation. As a measure of local velocity scale, it is relevant to transport phenomena and so has been an essential aspect of many prediction schemes. Its governing dynamical equation may be put in the form

$$\rho\left(\frac{\partial k}{\partial t} + \bar{U}_j\frac{\partial k}{\partial x_j}\right) = -\rho\overline{u_i u_j}\frac{\partial \bar{U}_i}{\partial x_j} - \epsilon + \frac{\partial}{\partial x_j}\left(\nu\frac{\partial k}{\partial x_j} - \tfrac{1}{2}\overline{u_j u_i^2} - \overline{p u_j}\right), \quad (38)$$

where the first term on the right-hand side is generally positive and can be thought of as the work of the Reynolds stress in deforming the mean field to produce k. ϵ is the "isotropic" dissipation rate, i.e., the rate at which k would be dissipated in isotropic turbulence, and appears after substituting for the general dissipation rate using an identity. The last term accounts for diffusive effects, both viscous and turbulent, and causes no net increase or decrease in turbulent kinetic energy.

Physical experiments and simulations have provided some very detailed information about the k budget. Johansson and Karlsson (1989) compared the distribution in the viscous layer of the terms in Eq. (38), measured by LDV in the boundary-layer experiment at $R_\theta = 2420$ of Karlsson and Johansson (1988), with the $R_\theta = 1410$ DNS of Spalart (1988). As shown in Fig. 9, the dominant terms are production and dissipation, which are in balance outside of the viscous sublayer and buffer layer. In the lower part of the buffer layer, the production term has

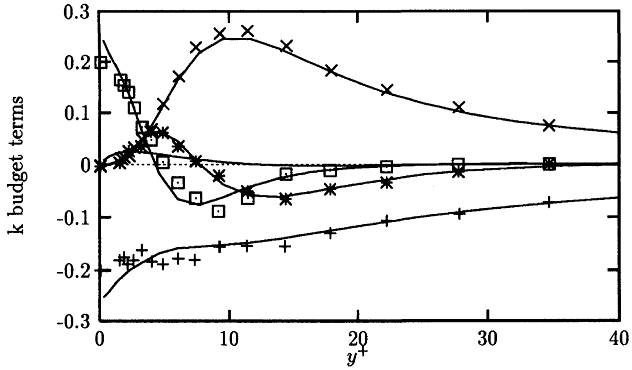

FIG. 9. Budget of turbulent kinetic energy equation terms for turbulent boundary layers. DNS results (solid lines) at $R_\theta = 1410$ courtesy of Spalart (1988) and LDV data at $R_\theta = 2420$ courtesy of Johansson and Karlsson (1989): (×) production, (+) dissipation, (*) turbulent diffusion, (□) viscous diffusion.

a maximum at about $y^+ = 10$, where it is balanced by the addition to the dissipation rate of negative viscous diffusion. The maximum dissipation rate occurs at the wall, where it is balanced by viscous diffusion, which is positive in the sublayer. Turbulent diffusion also changes signs just above the sublayer.

Further insight into the physics of the k field may be had via analysis of its spectrum, $E(\kappa,t)$. The latter, in the special case of homogeneous isotropic turbulence, satisfies

$$\frac{\partial E(\kappa,t)}{\partial t} = T(\kappa,t) - 2\nu\kappa^2 E(\kappa,t), \tag{39}$$

which may be derived from Fourier transformation of the $\mathcal{R}_{ij}(\mathbf{r},t)$ equation. Here, the last term in (39) gives the rate of dissipation of the turbulent kinetic energy at wave number κ. Summing this over wave-number amplitudes gives the total dissipation rate,

$$\epsilon(t) = 2\nu \int_0^\infty \kappa^2 E(\kappa,t) d\kappa. \tag{40}$$

Contrasting this with (12), it should be evident that wave numbers where kinetic energy is a maximum, i.e., where $E(\kappa,t)$ peaks, will be different—in fact lower—than those where dissipation is at a maximum, i.e., where $\kappa^2 E(\kappa,t)$ peaks. For large Reynolds numbers there may be a significant separation in wave numbers—equivalently length scales—between the peaks in energy and dissipation rate.

Evidently, turbulent flow requires a mechanism by which energy, which is introduced into the flow at the relatively large length scales associated with the stirring mechanism [and subsequently congregates in the region of large $E(\kappa,t)$], travels to higher wave numbers to be dissipated. The term $T(\kappa,t)$ in (39), which comes from a Fourier transform of $\mathcal{S}_{ijk}(\mathbf{r},t)$, accounts for the transfer process. It is driven physically by the stretching and folding of vortex filaments.

Among the first ideas about energy transfer between scales was that it occurs through a disequilibrium cascade in which energy primarily travels from large to small scales, with only modest amounts traveling in the reverse direction. Kolmogorov postulated that for sufficiently large Reynolds number, when dissipation and energy-containing scales are widely separated, there may be a universal structure to the small scales of the turbulence independent of the means by which energy is generated at large scales. Moreover, under these conditions, there may exist a range of scales, called the inertial range, which are at the same time not small enough for dissipation to be significant, nor so large as to be connected to the turbulence-containing eddies. Under these conditions, Kolmogorov assumed that $E(\kappa,t)$ would depend just on κ and ϵ, the latter representing the energy/second passing through this wave-number range. A dimensional argument gives $E(\kappa,t) \sim \epsilon^{2/3}\kappa^{-5/3}$. Numerous experimental evaluations of $E(\kappa)$ have confirmed the existence of this inertial range. Despite the validity of this $-\frac{5}{3}$ spectral law, the one-way cascade view of energy transfer is now widely recognized (Lesieur, 1990) to be incompatible with current understanding of the physics of the decay process. Dissipation is spatially intermittent, and the energy cascade can be viewed more accurately as a two-way street with nearly equal amounts of energy traveling in both directions, with the net transfer to smaller scales. Chorin (1994) has developed a statistical mechanical framework with which to explain the Kolmogorov spectrum in terms of such an equilibrium process (see Sec. 4.3).

The upper wave-number limit of the inertial range is that where viscosity begins to be important in causing dissipation. A length scale characterizing this, called the Kolmogorov scale, can be formulated from ν and ϵ as $\eta \sim (\nu^3/\epsilon)^{1/4}$. This has proven to be a useful parameter in establishing a minimum-resolution requirement in physical measurement techniques and in numerical simulation of turbulent flow, as well as in the scaling of turbulence properties to demonstrate their universal characteristics at high wave numbers.

2.4 Boundary-Free Turbulent Shear Flows

In contrast to the bounded flows considered previously, some of the principal properties of boundary-free turbulent shear flows are now described. The discussion is limited to steady, thin plane flows that are two dimensional in the mean, are developing as they proceed downstream from their source, and have a central plane at $y = 0$, and for which $\bar{U}(x,y) \gg \bar{V}(x,y)$ and $\partial/\partial y \gg \partial/\partial x$. Some flows in this category are a high–aspect-ratio rectangular jet, the wake of a two-dimensional bluff body—e.g., a cylinder—and a two-dimensional mixing (shear) layer. Sketches of these three boundary-free turbulent shear flows are given in Fig. 10. Axisymmetric developing free flows without swirl, such as round jets and the wake of a body of revolution, also can be treated in a similar manner to the plane flows considered here. The local length scale l in Fig. 10 is defined as the distance from the centerline at $y = 0$ to the cross-stream location where the velocity changes from its centerline value by $\frac{1}{2}U_c$. Here U_c is the mean velocity at the centerline for the jet, it is the velocity difference ($\Delta U \equiv U_h - U_e$) between the high- and low-speed sides of the mixing layer, and it is the velocity deficit at the centerline for the wake.

For such flows, the RANS equations may be approximated by

$$\bar{U}\frac{\partial \bar{U}}{\partial x} + \bar{V}\frac{\partial \bar{U}}{\partial y} \approx \frac{\partial \overline{uv}}{\partial y}. \tag{41}$$

For jets and wakes, this may be used to give the momentum flux per unit flow breadth as

$$M = \rho \int_{-\infty}^{\infty} [\bar{U}(\bar{U} - U_\infty) + \bar{v}^2 - \bar{u}^2]dy, \tag{42}$$

where U_∞ is the free-stream velocity in the case of the wake and zero for the jet with no coflow. The contribution to the momentum flux by the turbulent variances in (42) is usually small compared to that of the mean field and often can be neglected. For the plane jet, M is invariant with downstream distance, and (42) can be approximated by

$$M \approx \rho \int_{-\infty}^{\infty} \bar{U}^2 dy. \tag{43}$$

For the small-deficit plane wake, M in (42) should be interpreted as the mean momentum deficit flux per unit wake breadth and is also invariant. By contrast, for a plane mixing layer the momentum flux changes with distance downstream at the rate $V_\infty \Delta U$.

2.4.1 Similarity and Self-Preservation

Reynolds number similarity is the idea that, for sufficiently large Reynolds numbers, the structure of the turbulence no longer depends on viscosity and thus is Reynolds number independent. Self-preservation is a related idea. In this case local length and velocity scales can be found, which cause the scaled mean and other average statistical flow properties to have the same spatial distribution across the flow, independent of location in the mean flow direction. The self-preserving character of developing turbulent flows is obviously of great utility in giving experimental observations universal applicability. However, self-preservation is rigorously permitted by the equations of motion only in some restricted instances. Assuming

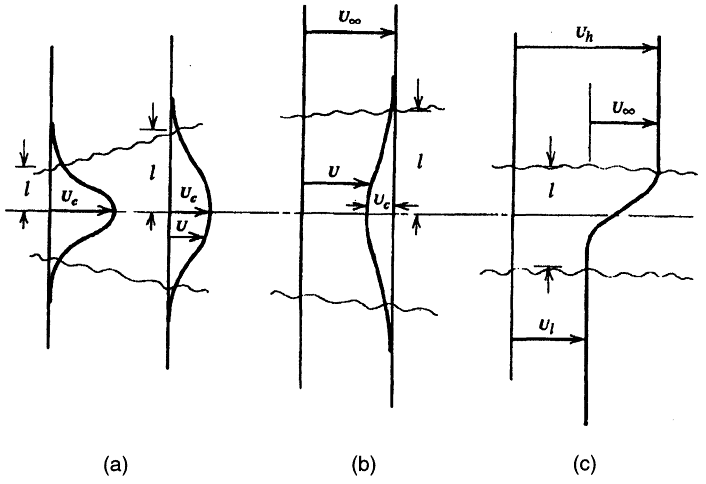

FIG. 10. Mean velocity profiles of a turbulent plane (a) jet, (b) wake, and (c) mixing layer. Note the length scale, *l*, and velocity scale, U_c. Adapted from *Incompressible Flow* by R. L. Panton Copyright © 1984. Reprinted by permission of John Wiley & Sons, Inc.

that l and U_c (ΔU for mixing layers) in Fig. 10 are the local scales that will yield self-preserving properties for plane jets, wakes, and mixing layers, then the self-preserving functional relations (Panton, 1984) for the mean velocities have the form, for a jet,

$$\frac{\bar{U}(x,y)}{\bar{U}_c(x)} = f\left(\frac{y}{l(x)}\right); \tag{44}$$

for a wake,

$$\frac{U_\infty - \bar{U}(x,y)}{\bar{U}_c(x)} = f\left(\frac{y}{l(x)}\right); \tag{45}$$

and for a mixing layer,

$$\frac{\bar{U}(x,y) - U_l}{\Delta\bar{U}(x)} = f\left(\frac{y}{l(x)}\right). \tag{46}$$

For the plane jet, substituting (44) into (43) results in

$$M = \rho U_c^2 l \int_{-\infty}^{\infty} \left[f\left(\frac{y}{l}\right)\right]^2 d\left(\frac{y}{l}\right). \tag{47}$$

The integral is a constant for self-preserving flow, as is the momentum flux M. Therefore $\rho U_c^2 l$ must also be invariant. Furthermore, assuming Reynolds number similarity, the dimensional relation

$$U_c(\rho x/M) = \text{const.} \tag{48}$$

may be derived. Taken together, (47) and (48) imply that $U_c \propto x^{-1/2}$ and $l \propto x$ for the plane jet. These relationships, based on the Reynolds-number–similarity and self-preservation assumptions, are demonstrated experimentally for the mean field of plane jets after about 5 jet widths downstream. This is where the two initial shear layers, emanating from the boundary layers on each side of the jet nozzle, have grown together to form the jet.

A similar analysis for the plane wake, with the same assumptions, gives $U_c \propto x^{-1/2}$ and $l \propto x^{1/2}$. In fact, far enough downstream, where the wake deficit is sufficiently small, plane wakes also exhibit these self-preserving characteristics. This is achieved first by the mean velocity profile, about 80 body diameters downstream in a cylinder wake, and further downstream for the turbulence intensities. Higher-order statistics require even greater distances to reach a self-preserving state.

The mean flow of mixing layers becomes self-preserving at about $Re_x = 4 \times 10^5$, at which point $U_c \sim x$ and $l \sim x$. The spreading rate of the mixing layer depends on the ratio of the freestream velocities, U_h/U_l.

2.4.2 Turbulence Spreading As the outer edges of a free turbulent shear layer is approached, many of its properties go to zero. For example, the turbulent kinetic energy, the Reynolds shear stress, and, most importantly, the vorticity all become negligible. Although the mean properties vary smoothly into the outer flow, the instantaneous turbulent/nonturbulent interface is rather sharp, as seen below in Fig. 13 for a turbulent mixing layer, and continuously changes position. Thus, the turbulence at any particular point is increasingly more intermittent, the further from the center axis it is. The outer edge region, sometimes called the "viscous superlayer," has been extensively studied by Corrsin and Kistler (1954) for a turbulent boundary layer whose core region has wakelike characteristics. The viscous superlayer is a highly sheared layer with thickness of the order of the Kolmogorov scale. A possible mechanism determining the rate at which the irrotational flow outside the turbulent layer is "entrained," i.e., made turbulent, is the engulfing of the irrotational fluid by layers of rotational fluid and subsequent viscous diffusion of vorticity, which is amplified by the straining action of these layers.

2.4.3 Some Characteristics of Mixing Layers The self-preserving length scale commonly used for mixing layers is the momentum thickness

$$\theta = \int_{-\infty}^{\infty} \frac{\bar{U} - U_l}{U_h - U_l}\left[1 - \frac{\bar{U} - U_l}{U_h - U_l}\right]dy, \tag{49}$$

which is proportional to the length scale l in Fig. 10. The self-preserving character of mixing layers, formed from two boundary layers with different free-stream velocities, U_h and U_l, meeting at the exit plane of a splitter plate, is shown in Fig. 11. The mean velocity distributions of an experimental mixing layer of Bell and Mehta (1992) at three streamwise locations and for R_θ up to 6000 are com-

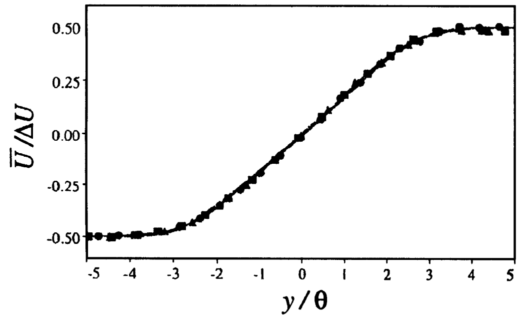

FIG. 11. Comparison of the mean velocity profiles of turbulent mixing layers, in their self-preserving regions, from the temporally developing DNS of Rogers and Moser (1994) at τ = 105.2, 116.7, 128.6, 141.0, and 150.8 and from the experiment of Bell and Mehta (1992): (■) x = 108.1 cm, (●) x = 128.4 cm, and (▲) x = 189.4 cm. The curves for the various values of τ are indistinguishable at the scale of the figure. Adapted from "Direct Simulation of a Self-Similar Turbulent Mixing Layer," by M. M. Rogers and R. D. Moser in *Physics of Fluids* **6**, Copyright © 1994, American Institute of Physics. Reprinted by permission of the American Institute of Physics.

pared with the temporally growing DNS mixing layer of Rogers and Moser (1994) at five different times in the self-preserving period of the simulation, with R_θ ranging from 800 to 2420. The Reynolds shear stress distribution shown in Fig. 12 for these two comparison flows also displays self-preserving characteristics. Brown and Roshko (1974) observed that coherent spanwise vortices with diameters of the order of the mixing layer width are a striking, persistent feature of mixing layers and appear to play a dominant role in their dynamics. They, and later Konrad (1976), also observed that, superimposed on the large spanwise roller vortices and in the "braid" region in between them, streamwise streaks occur, which Bernal and Roshko (1986) later demonstrated were due to vortices with axes oriented primarily in the streamwise direction. An instantaneous schlieren photograph of both the spanwise rollers and the streamwise streaks (Bernal and Roshko, 1986) is shown in Fig. 13. The spanwise rollers and the interspersed braid regions are also seen in Fig. 14, where contours of constant spanwise vorticity Ω_z are plotted from the temporally growing mixing-layer DNS of Rogers and Moser (1994). Views of this DNS flow in the cross-stream plane for cuts through a roller (indicated by the line marked B in Fig. 14) and through the braid (marked A) are shown in Fig. 15 for times (a) before self-preservation is reached and (b) during self-preservation. In these cross-stream views, contours of streamwise vorticity Ω_x and contours of a passive scalar for which Sc = 1 (in the high-speed free stream) are shown. Clearly, the flow in the self-preservation region is more disorganized than at the earlier time. Rogers and Moser (1994) note that they only observed distinct streamwise vortices and the "pairing" of spanwise rollers at the earlier times before the self-preserving state is achieved. Others have shown earlier (e.g., Hussain, 1986), as seen in Fig. 16, that Reynolds stress and kinetic energy production are concentrated primarily in the braid regions.

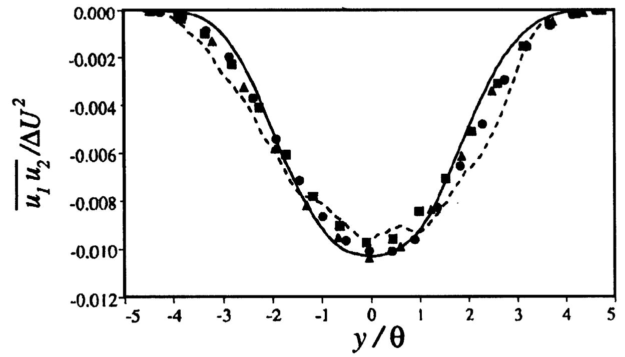

FIG. 12. Comparison of the Reynolds shear-stress profiles of turbulent mixing layers from the temporally developing DNS of Rogers and Moser (1994): (solid line) time-averaged DNS results, (dashed line) τ = 187.5, and (symbols) the experiment of Bell and Mehta (1992): (■) x = 108.1 cm, (●) x = 128.4 cm, and (▲) x = 189.4 cm downstream of origin. Adapted from "Direct Simulation of a Self-Similar Turbulent Mixing Layer," by M. M. Rogers and R. D. Moser in *Phys. Fluids* **6**, Copyright © 1994, American Institute of Physics. Reprinted by permission of the American Institute of Physics.

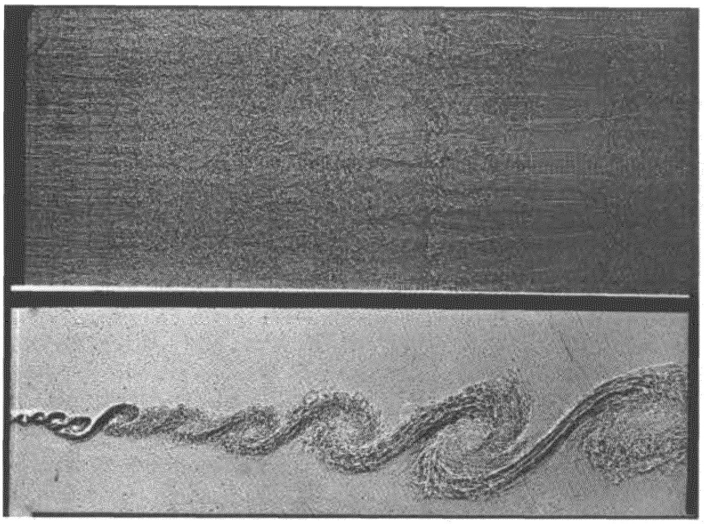

FIG. 13. Large-scale spanwise vortices and streamwise streaks in a turbulent mixing layer, (a) plan view, (b) spanwise view. Photo courtesy of A. Roshko. Also Fig. 2(b) of "On the Development of Three-Dimensional Small Scales in Turbulent Mixing," by L. P. Bernal, R. E. Briedenthal, G. L. Brown, J. H. Konrad, and A. Roshko in *Turbulent Shear Flows 2,* Copyright © 1980, Springer-Verlag GmbH & Co. KG. Reprinted by permission of Springer-Verlag GmbH & Co. KG.

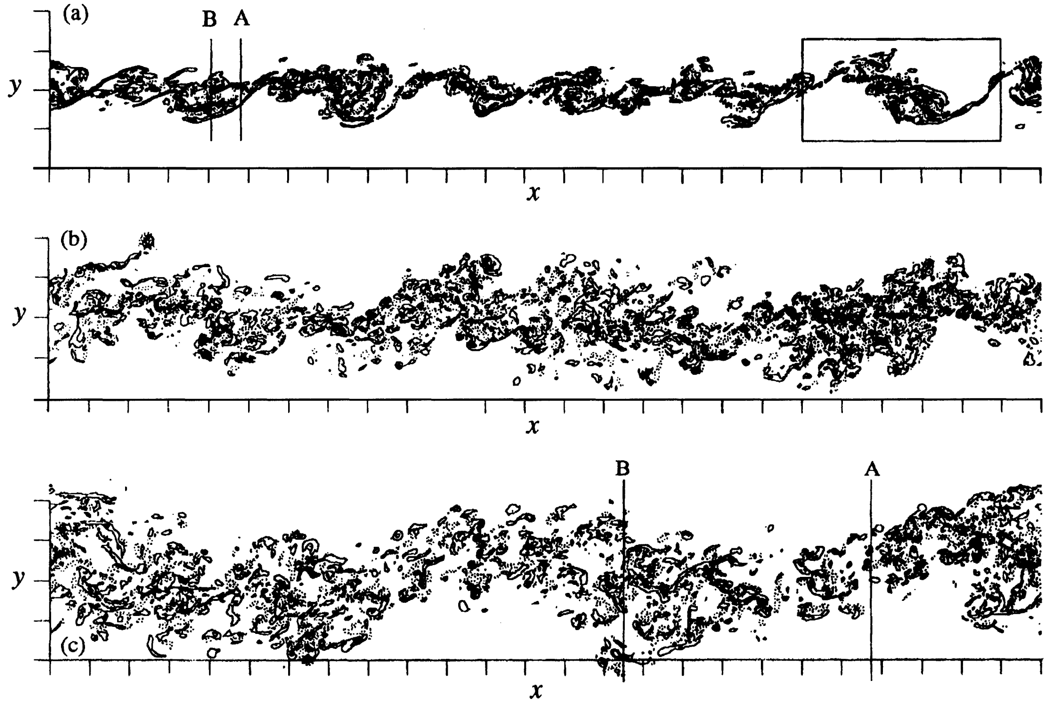

FIG. 14. Contours of spanwise vorticity in *x-y* planes at (a) $\tau = 78.5$, (b) $\tau = 150.0$, and (c) $\tau = 187.5$. The vertical lines marked A and B are the locations of the *y-z* planes shown in Figs. 15(a,b). Adapted from "Direct Simulation of a Self-Similar Turbulent Mixing Layer," by M. M. Rogers and R. D. Moser in *Phys. Fluids* **6,** Copyright © 1994, American Institute of Physics. Reprinted by permission of the American Institute of Physics.

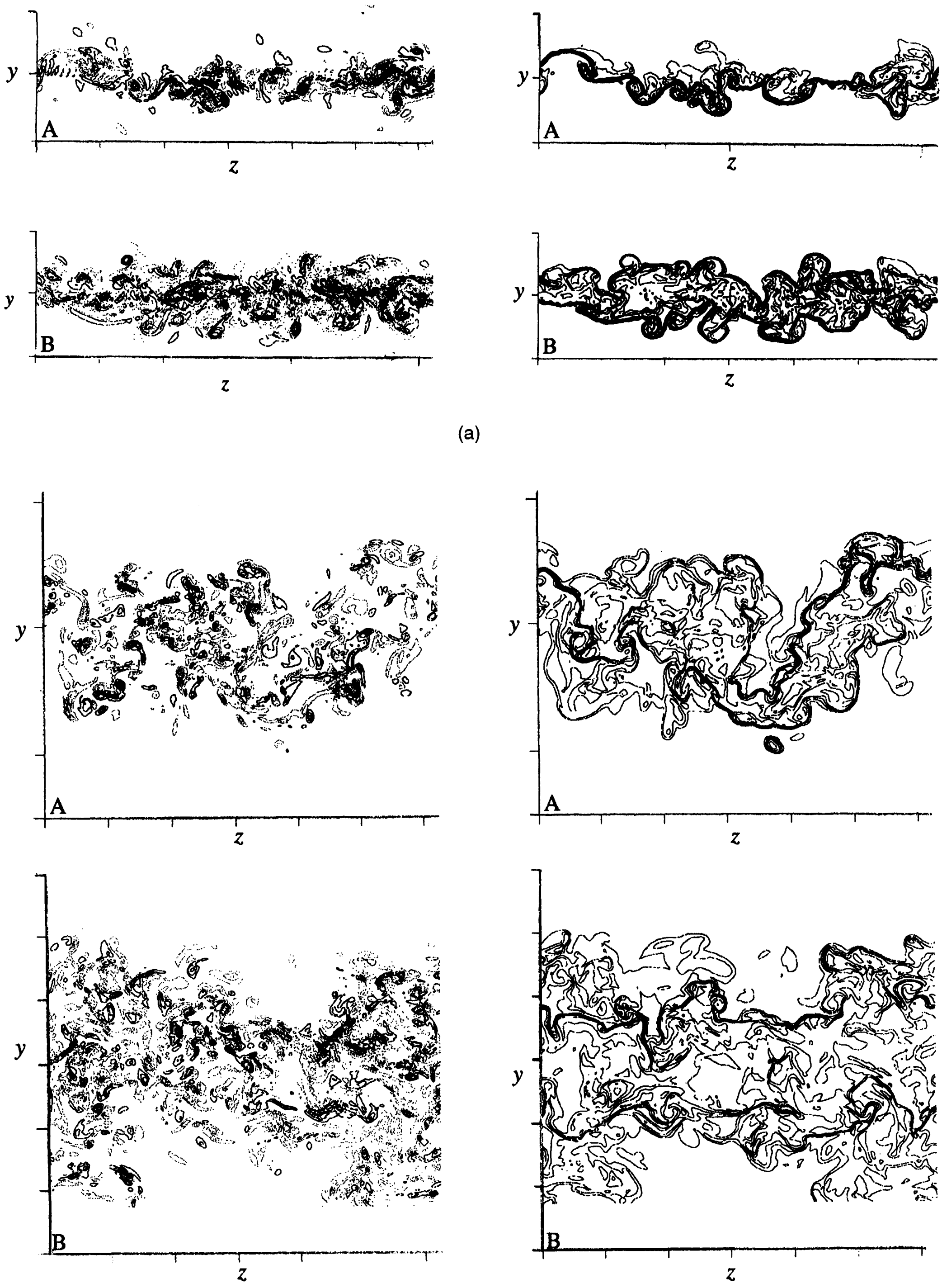

(a)

(b)

FIG. 15. Contours of streamwise vorticity (on left) and passive scalar (on right) in cross-stream *x*-*y* planes in the braid region (marked A in Fig. 14) and a large-scale roller vortex (marked B in Fig. 14) (a) before the self-preservation region at $\tau = 78.5$ and (b) within the self-preservation region at $\tau = 187.5$. Adapted from "Direct Simulation of a Self-Similar Turbulent Mixing Layer," by M. M. Rogers and R. D. Moser in *Phys. Fluids* **6,** Copyright © 1994, American Institute of Physics. Reprinted by permission of the American Institute of Physics.

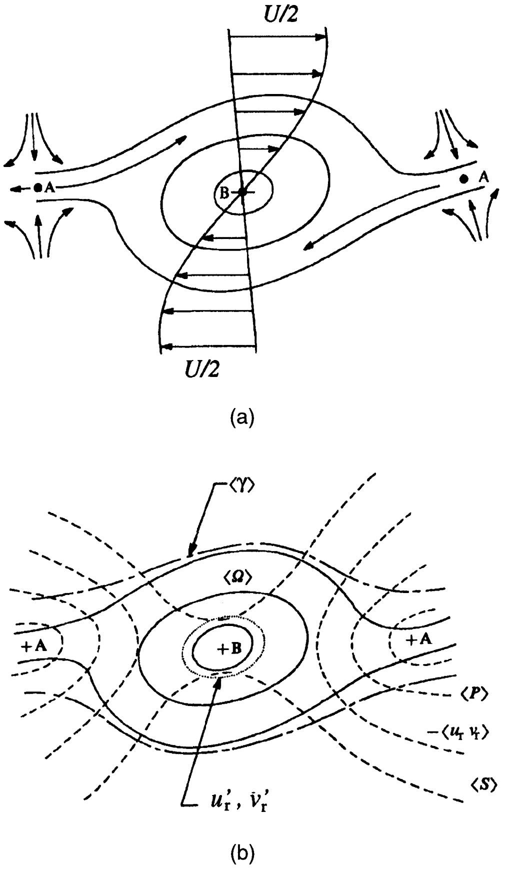

FIG. 16. Conceptual sketch of a coherent large-scale vortex and braid regions in a turbulent mixing layer. (a) Coherent vorticity contours showing saddle **A** and vortex center **B**. (b) Approximate contours of other coherent structure properties: intermittency $\langle\gamma\rangle$; turbulence intensities $\langle u_r'\rangle$, $\langle v_r'\rangle$; Reynolds stress $-\langle u_r v_r\rangle$; strain rate $\langle S\rangle$; production $\langle P\rangle$. Subscript *r* indicates incoherent field; other symbols are for coherent field. Adapted from Fig. 1 of "Coherent Structures and Turbulence" by A. K. M. F. Hussain in *J. Fluid Mech.* **173,** Copyright © 1986, Cambridge Univ. Press. Reprinted by permission of Cambridge University Press.

3. THEORY OF IDEALIZED TURBULENT FLOWS

3.1 Isotropic and Homogeneous Turbulence

A fundamental aspect of turbulent flow is the way in which energy decays from an arbitrary initial state in the absence of new turbulence produced by the presence of a mean shear. By assuming the turbulence to be homogeneous and isotropic, the decay problem is reduced to its simplest form. In this case, (38) becomes

$$dk/dt = -\epsilon, \tag{50}$$

which makes clear the essential role of the dissipation rate ϵ in determining the history of the kinetic energy field during the decay process. Further analysis of (50) depends on having information available about ϵ. This, in fact, can be usefully obtained from the exact ϵ equation, which is formally derivable from the Navier–Stokes equation:

$$\frac{d\epsilon}{dt} = \frac{7}{3\sqrt{15}} S_k R_t^{1/2} \frac{\epsilon^2}{k} - \frac{7}{15} G \frac{\epsilon^2}{k}. \tag{51}$$

The first term on the right-hand side of (51) represents the effect of vortex stretching in augmenting ϵ, while the second accounts for its destruction. Here, $R_t \equiv k^2/\nu\epsilon$ is the turbulence Reynolds number, $S_k \equiv -\overline{(\partial u/\partial x)^3}/[\overline{(\partial u/\partial x)^2}]^{3/2}$ is the skewness factor of $\partial u/\partial x$, and $G \equiv \overline{u^2}\ \overline{(\partial^2 u/\partial x^2)^2}/[\overline{(\partial u/\partial x)^2}]^2$ is the "palinstrophy" coefficient (see Lesieur, 1990, for a discussion of this terminology). A measure of the local intensity of the turbulent vorticity field is the turbulent enstrophy $\zeta \equiv \overline{\omega_i^2}$. In view of the isotropic identity $\zeta = \epsilon/\nu$, it is clear that there is a close connection between dissipation regions and those of high vorticity fluctuations. In fact, after division by ν, (51) may be written as the enstrophy equation.

Generally, S_k and G are time dependent, and so the system of equations formed by (50) and (51) is not closed. Further progress in the analysis of the isotropic decay problem can be had if additional information deriving from the two-point statistics is brought to bear on the situation. In particular, the isotropy condition introduces great simplification in the forms of $\mathcal{R}_{ij}$ and $\mathcal{S}_{ijk}$ so that they each, in fact, depend on just a single scalar function. In the case of $\mathcal{R}_{ij}$ it may be shown that

$$\mathcal{R}_{ij}(\mathbf{r},t) = \overline{u^2}(t)\left[\left(f(r,t) + \frac{r}{2}\frac{\partial f(r,t)}{\partial r}\right)\delta_{ij} - \frac{r}{2}\frac{\partial f(r,t)}{\partial r}\frac{r_i r_j}{r^2}\right], \tag{52}$$

where the longitudinal correlation coefficient

$f(r,t)$ was defined earlier. A similar result for $\mathcal{S}_{ijk}(\mathbf{r},t)$ shows that it depends on just the triple longitudinal correlation coefficient $K(r,t) \equiv -\mathcal{S}_{111}(r\mathbf{e}_1,t)/(\overline{u^2})^{3/2}$. If it is assumed that the turbulent flow maintains a structural self-similarity during the decay (von Kármán and Howarth, 1938), i.e., one in which $f(r,t)$ and $K(r,t)$ satisfy

$$f(r,t) = f(r/l(t)) \tag{53}$$

and

$$K(r,t) = K(r/l(t)), \tag{54}$$

where $l(t)$ is an appropriate length scale, then it may be shown that S_k and G are constants and $l(t) \sim \lambda(t)$. In this case (50) and (51) form a closed system so that they may be solved for the decay of $k(t)$ and $\epsilon(t)$ from an arbitrary initial state. These solutions can be expected to be germane to the general decay problem—where exact self-preservation is not present—since solutions to (50) and (51) are stable to perturbations of S_k and G; i.e., the solutions will not be qualitatively different if S_k and G show some time dependence.

Speziale and Bernard (1992) have extensively analyzed the self-similar decay solutions including interpreting their relationship to the general decay problem and to physical experiments. It was found that they admit two fixed-point solutions. One has $R_t = 0$, is approached at large times whenever $7G/15 < 2$, and is associated with the final period of decay. This corresponds to weak turbulence where $R_t \to 0$, $k \sim t^{-5/2}$, and f has the Gaussian form

$$f(r/\lambda) = \exp(-r^2/2\lambda^2), \tag{55}$$

as was first predicted and verified experimentally by Batchelor and Townsend (1948). It may be shown (Sedov, 1944; Bernard, 1985) via analysis of the separability properties of the exact dynamical equation for $f(r/\lambda)$ and $K(r/\lambda)$ (known as the von Kármán–Howarth equation), that (55) is compatible with a vanishingly small, but nonetheless finite, influence of vortex stretching.

A second fixed point of the decay is $R_t = [(7G/15 - 2)/(7S_k/3\sqrt{15})]^2$. It is approached rapidly for large initial values of R_t whenever $7G/15 > 2$ and is the unique high–Reynolds-number fully self-preserving decay law. For this, $k \sim t^{-1}$ is reached as an asymptotic state after the stretching and dissipation terms in (51) come into balance, typically after several eddy turnover times. Until this happens, decay with $k \sim t^{-\alpha}$, $1 \le \alpha \le 1.4$, is predicted. The smaller the initial R_t is, the longer it takes until t^{-1} decay is reached. This may explain the propensity of experiments to show $t^{-\alpha}$ decay with $\alpha \approx 1.25$ on average. The few experiments that have been performed at high R_t do indicate a t^{-1} decay.

A single self-similar solution cannot be maintained throughout the decay process. After the initial transient, in which R_t approaches a constant and the t^{-1} decay law is reached, a further change must eventually occur in which R_t decays to zero and the t^{-1} decay shifts to the final-period $t^{-5/2}$ solution. Whatever the initial form of f is, it must change over to (55) in due course.

3.2 Homogeneous Shear Flow

The ways in which the decay process described in the previous section is modified by the presence of mean shear—which acts to produce turbulence—is of considerable theoretical and practical interest. The essence of this phenomenon may be most easily examined in the case of an initially homogeneous isotropic turbulence subjected to a uniform mean shearing, say $S \equiv d\bar{U}/dy > 0$. In this case, (50) is modified by the addition of a single production term,

$$\frac{dk}{dt} = -\overline{uv}\frac{d\bar{U}}{dy} - \epsilon, \tag{56}$$

which is positive since $\overline{uv} < 0$.

A number of experimental and theoretical studies have attempted to predict the short- and long-term behavior of k and ϵ subjected to homogeneous shear. Generally, these have found exponential growth at approximately the same rate for k and ϵ over the dimensionless times that have been examined ($St < 30$). It is not certain what occurs at larger times, but two principal scenarios have been put forward. Townsend (1956) hypothesized that, given enough time, a production-equals-dissipation equilibrium would result in which terms on the right-hand sides of (56) and (51) achieve balance and k and ϵ attain constant values. An alternative viewpoint is that the exponential growth in k and ϵ, observed for relatively short times, continues unabated for long times as well.

The latter point of view is consistent with standard modeling assumptions that have been applied to the ϵ equation. In this, a scaling argument is used to conclude that the vortex-stretching term in (51) must be subsumed by the dissipation term. The resulting equations give exponential growth in k and ϵ. Note that, after substitution for R_t in (51) corresponding to the two equilibrium states discussed in the previous section, it is found that either $d\epsilon/dt = -(7G/15)\epsilon^2/k$ or $d\epsilon/dt = -2\epsilon^2/k$. In either case there is domination of vortex stretching by dissipation. Before equilibrium is reached, however, vortex stretching does make a contribution.

Unlike the decay case, in homogeneous shear flow, R_t increases with time, significantly affecting the relative behavior of the terms in the ϵ equation. The production-equals-dissipation equilibrium results if it is assumed merely that there is some residual effect of vortex stretching at high Reynolds numbers (Bernard and Speziale, 1992). The slight enhancement of dissipation that results is just enough to raise ϵ to the point that it slows and stops the growth in k and then consequently itself also. Numerical calculations show that this occurs at times that are generally beyond those seen thus far in physical experiments or computations.

Some indication that the production-equals-dissipation scenario is the more physical of the two may be found by considering the behavior of the enstrophy equation in isotropic decay in the limit of zero viscosity (Lesieur, 1990). For very small ν, i.e., high R_t, the vortex-stretching term in (51) will be unchecked by dissipation and the enstrophy will increase dramatically in consequence of nearly inviscid vortex stretching. Only later, after R_t drops precipitously because of the rise in ζ (i.e., ϵ) and the fall of k, does vortex stretching lose its dominance, i.e., when the t^{-1} equilibrium is reached. Turbulence models that limit the right-hand side of (51) to being purely dissipative miss the initial rise in vorticity; moreover, in the $\nu \rightarrow 0$ limit they erroneously predict constant enstrophy as the long-time solution, a physical impossibility. Keeping a net vortex stretching is equivalent to predicting enstrophy blowup in the $\nu \rightarrow 0$ limit, a result that is generally compatible with numerical solutions to the inviscid Euler equations (Pumir and Siggia, 1992). Whether or not this happens in finite or infinite time depends on the behavior of S_k as $\nu \rightarrow 0$; such information is not yet available.

4. TURBULENCE MODELING AND THEORY

4.1 Closure Schemes

Turbulence modeling is the name given to the various strategies that have been developed for obtaining closure to the average equations of turbulent flow. The simplest approaches model R_{ij} appearing in the RANS equation (19)—most often using a molecular analogy—or else model the vorticity transport correlation in the rotational form of the RANS equation. In second-order closures, the unknown terms in the R_{ij} equation (34) are modeled. Some alternative approaches, based mainly on the two-pont correlation equation, are considered in the next section.

Constituitive models for R_{ij} most commonly incorporate gradient transport in the general form

$$\overline{u_i u_j} = \tfrac{2}{3}k\delta_{ij} - \nu_t\left(\frac{\partial \overline{U}_i}{\partial x_j} + \frac{\partial \overline{U}_j}{\partial x_i}\right), \tag{57}$$

where $\nu_t > 0$ is termed the "eddy viscosity." Equation (57) is specially designed to be consistent with a contraction of indices, since $\partial \overline{U}_i/\partial x_i = 0$ for incompressible flows. It can also be thought of as the lowest-order term in a functional expansion of the Reynolds stress in terms of mean velocity differences (Speziale, 1991).

For unidirectional flows, such as in a channel, (57) takes the form

$$\overline{uv} = -\nu_t \frac{d\overline{U}}{dy}. \tag{58}$$

There is a wide variety of approaches to the selection of ν_t. For example, in the mixing-length model (Prandtl, 1925), the molecular transport analogy is invoked to justify the expression $\nu_t = l_m^2|d\overline{U}/dy|$, where l_m is a mixing length. In the k–ϵ closure (Launder and Spalding, 1974), dimensional reasoning gives $\nu_t = C_\mu k^2/\epsilon$, where C_μ is a constant, and additional model equations must be solved for k and ϵ independently of (19). For flow situations where (58) is invalid—e.g., because $\nu_t < 0$ or infinite—none of the various expressions for ν_t has meaning. The general experience with eddy viscosity models is that among the class of flows where (57) cannot be excluded *a priori,* no one model is suitable for every case. Among the important flows where (57) is inadequate are those with streamline curvature, body forces, and nonsteady mean velocity. Its implied isotropy of the normal stresses prevents (57) from reproducing secondary flows, such as can appear in noncircular ducts. Some of these fundamental limitations can be alleviated by forming nonlinear generalizations of (57) that include quadratic terms in the functional expansion (Speziale, 1987). While in some instances they improve upon linear models, they still fail to capture important curvature and relaxation effects.

At the outset, modeling of the R_{ij} equation (34) has the advantage that the convective and viscous diffusion terms are computed exactly. The modeling effort thus needs to focus just on the higher-order correlations Π_{ij}, ϵ_{ij}, and β_{ijk}. While the general outlines of a physical interpretation of these terms can be given, they are largely opaque to physical explanation in terms of classifiable events in turbulent flow. Modeling efforts have focussed instead on using such dynamical restrictions as frame invariance under the extended Galilean group, material frame invariance in the limit of two-dimensional turbulence, realizability (the condition that velocity variances such as $\overline{u_i^2}$ are predicted to be always positive), and consistency with rapid distortion theory to guide the development of functional forms (Speziale, 1991). The resulting closures are better able to include anisotropy and relaxation effects than lower-order closures, and, unlike the latter, they are able to distinguish between the presence of plane shear, plane strains, and rotating plane shear.

4.2 Statistical Closures

The modeling techniques discussed in the previous section are limited in the range of physical ideas about turbulence that they are able to incorporate in forming closure. Consequently, a number of other methodologies have been devised that provide an opportunity to encompass alternative physical ideas. Several of such theories depend on making hypotheses about the statistical properties of the velocity field as viewed from Fourier space. Others utilize probabilistic expansions of the velocity field. Such methods are often too mathematically complex to be useful for predicting general flows, though they can offer insights into the analysis of relatively simple flows, such as those discussed in Sec. 3. In recent years, renormalization-group methods have been brought to bear on the closure problem as well. Many of the details of these methodologies may be conveniently viewed in the monographs by Leslie (1973), Lesieur (1990), and McComb (1990).

At a given point in homogeneous, isotropic turbulence, experimental measurements show that the velocity components tend to be Gaussian random variables. The multiple-point statistics, however, are not Gaussian; i.e., the velocity field is not a Gaussian random field. If it is assumed that the departure from Gaussianity is not too severe, then there are a number of closely related techniques that can be used to generate approximate closures. These proceed in spectral space by essentially closing the hierarchy of Fourier transformed equations for $\mathcal{R}_{ij}$, $\mathcal{S}_{ijk}$, and so on.

An early idea in this direction was to close the system at the level of third-order moments—which have fourth-order moments as unknowns—by relating the fourth-order correlations to products of second-order moments using the properties of Gaussian random fields. Since the velocity field is not Gaussian, this "quasinormal" hypothesis cannot be exactly true, but it may have some partial validity. It was subsequently found that the solution to the closed system generates negative energy, on account of unwarranted growth in the third-order moments. Further hypotheses are needed to force their saturation, and thereby ensure realizability. A successful approach of this sort is to use "eddy damping" in coordination with a "Markovianization" assumption. The resulting closure is referred to as EDQNM. This has the virtue of being applicable to high–Reynolds-number flows and is much less expensive to use than DNS. Its mathematical complexity primarily limits its use to unbounded flows.

In the direct interaction approximation (DIA), and its close relatives (Kraichnan, 1958, 1959; Leslie, 1973), the velocity field is subjected to a probabilistic expansion about the Gaussian state. Fourier transformations of the Navier–Stokes equation for turbulence in a three-dimensional box driven by a homogeneous isotropic stirring force leads to a hierarchy of coupled equations for the Fourier coefficients of the velocity field. An equation is also generated for the response tensor associated with an infinitesimal perturbation force. Various statistical approximations are made to develop a fully realizable closed system of equations for the ensemble average of the response tensor and quadratic products of the velocity modes. As originally constituted, DIA does not produce the inertial-range spectrum, but later variants, in particular the Lagrangian history DIA, do so.

The RNG method, as developed by Yakhot and Orszag (1986), starts with a similar correspondence principle as used in DIA to the effect that the Navier–Stokes equation with the correct forcing can be made to reproduce an inertial-range spectrum. The forcing function is chosen to be analogous to that in equilibrium thermal systems. The velocity field is written as the sum of long- and short-wavelength modes. A dynamical equation for the former is derived from the space-time Fourier transform of the forced Navier–Stokes equation, by repeatedly using it to eliminate the high–wave-number part of the velocity field. The equation is then averaged over high–wave-number strips, leaving an equation for the long-wavelength modes. A zero–wave-number limit is taken, the variables are rescaled, and the procedure is repeated. In the asymptotic limit, a turbulence closure or LES scheme can be extracted.

4.3 Thermodynamics of Vortex Systems

Beginning with the important work of Onsager (1949) in two dimensions, there has been much interest in explaining some of the important properties of turbulence, such as the Kolmogorov inertial-range spectrum, in terms of the statistical mechanics of vortex systems. In recent years (e.g., Montgomery and Joyce, 1974; Miller, 1991; Robert, 1990) understanding of the 2D case has achieved a high degree of development, and attention has turned increasingly toward examining the situation in three dimensions. One of the principal goals is to describe the statistical and fractal properties of stretching, folding, and interacting vortical structures whose combined effect is to create the $-\frac{5}{3}$ inertial-range law. Chorin (1994) has provided a three-dimensional thermodynamics of vortex systems that fits within the framework of modern statistical mechanics. He demonstrates that vortex systems are distinguished by having negative temperatures and that turbulence evolves toward an equilibrium at the infinite-temperature limit. Using percolation theory and polymer statistics to develop a statistical description of vortex stretching, he shows how the $-\frac{5}{3}$ exponent appears as a critical exponent. This represents a distinct shift from earlier viewpoints, e.g., summarized by McComb (1990), in which turbulence in the Kolmogorov range is believed to be a strongly coupled nonequilibrium statistical system.

A second primary objective of a statistical mechanical formulation of turbulence, whether through vortex systems or otherwise, is to set up machinery for a renormalization procedure. As cited previously in regard to RNG, this is a methodology for systematically reducing the complexity of a physical system by absorbing the effect of small scales into computable larger-scale motions. The end result of renormalization is providing a formalism for large-eddy simulation techniques as discussed in Sec. 4.5. For vortex systems, renormalization can be applied by removing small hairpin vortices from the calculation (Chorin, 1993) or through rescaling of vortex rings (Chorin and Hald, 1995). Such methods can form the basis for efficient implementations of vortex methods for simulating turbulent flow (Bernard, 1996).

4.4 Direct Numerical Simulations

Numerical simulations of turbulent fluid flow are possible using the standard techniques of computational fluid dynamics (CFD), so long as adequate grid resolution is used to capture motions at the scale of the inertial range. The first successful simulations were of pressure-driven turbulent flow between two plane walls, i.e., channel flow. These used spectral methods, which fit in well with the use of periodic boundary conditions in the streamwise and spanwise directions. For successful simulation of turbulence, the computational domain must be at least as large as the distance over which significant correlation occurs in the velocity field. This presumably allows for a fair representation of the dynamical behavior of the principal vortical structures in the flow. Acceptable domain size for a channel-flow simulation is $L_x \approx 4\pi h$, $L_z \approx 2\pi h$, where h is the channel height. A minimal mesh that satisfies these requirements for Reynolds numbers just in the turbulent regime, i.e., $\leq$10 000, is on the order of 64^3 or 250 000 points.

The necessary grid density for turbulent-flow simulation goes up rapidly with Reynolds number, as may be seen from the following considerations. Assume that the smallest mesh spacing should be the Kolmogorov scale, $\eta \sim (\nu^3/\epsilon)^{1/4}$, and that the mesh must cover a region of the size of the scale of energy-containing eddies. The latter scale may be taken as $l_e = k^{3/2}/\epsilon = k^{1/2}T$, where the eddy-turnover time $T \sim k/\epsilon \sim k/(dk/dt)$ measures the time over which some appreciable decay might take place. The ratio $l_e/\eta = R_t^{3/4}$ (where $R_t = k^2/\nu\epsilon$ is the turbulence Reynolds number) is proportional to the number of grid points that are needed in each coordinate direction. For three dimensions, the required mesh must therefore have $R_t^{9/4}$ points. Furthermore, the minimal number of time steps of the calculation is $\sim R_t^{3/4}$, which is determined as the ratio of T to the maximum permissible time step in a numerical algorithm, $\sim \eta/k^{1/2}$. Thus, on the order of R_t^3 operations are necessary in performing a DNS, which is well beyond the capability of the current generation of computers once $R_t \sim 10^5$. The outlook is not good that high–Reynolds-number flows can be treated by such schemes any time in the foreseeable future.

4.5 Large-Eddy Simulations

Direct numerical simulation is not practical for most turbulent engineering flows by reason of the wide range of scales in need of resolution. Clearly, it is necessary to devise a means for limiting the range of computable scales—but not one so drastic as to put the entire range of physics into the transport correlations, such as occurs in Reynolds averaging. The resolution of these conflicting approaches is the large-eddy simulation (LES) technique that attempts to compute the energy-containing scales while modeling the rest. A variety of schemes for carrying out LES have been devised, though many issues remain to be resolved before it can be considered a generally applicable flow-prediction tool for engineering design.

The primary point of departure in formulating the LES approach has traditionally been through a filtering operation applied to the Navier–Stokes equations. A filtered velocity is defined via

$$\langle U\rangle(\mathbf{x},t) = \int_{R^3} G(\mathbf{x}|\mathbf{x}')U(\mathbf{x}',t)d\mathbf{x}', \tag{59}$$

where $G(\mathbf{x}|\mathbf{x}')$ is any one of a variety of filter functions. These can be functions of small spatial support or global functions that effectively filter high–wave-number modes from the velocity field. After applying the filtering operation to the Navier–Stokes equations, there results an equation for the resolvable fields $\langle U\rangle$ that is formally equivalent to the RANS equation, with the exception that the Reynolds stress is replaced by the subgrid stress tensor $\tau_{ij} = \langle U_iU_j\rangle - \langle U_i\rangle\langle U_j\rangle$. The latter ostensibly represents the effect of small-scale motions on the large. Unlike the Reynolds stress tensor, however, it lacks a clear-cut physical interpretation in terms of momentum flux. Subgrid models have tended to be reminiscent of (57) in that they assume

$$\tau_{ij} = \tfrac{2}{3}\tau_{kk}\delta_{ij} - \nu_t\left(\frac{\partial\langle U_i\rangle}{\partial x_j} + \frac{\partial\langle U_j\rangle}{\partial x_i}\right). \tag{60}$$

The most widely used subgrid model is that due to Smagorinsky (1963), in which

$$\nu_t = (C_S\Delta)^2|\langle S\rangle|, \tag{61}$$

where C_S is the Smagorinsky constant, Δ is the grid spacing of the numerical mesh, $|\langle S\rangle| \equiv (2\langle S\rangle_{ij}\langle S\rangle_{ij})^{1/2}$, and $\langle S\rangle_{ij} = (\partial\langle U_i\rangle/\partial x_j + \partial\langle U_j\rangle/\partial x_i)/2$. A motivation for (61) is that it is the outcome of assuming local equilibrium between molecular dissipation of small-scale turbulent energy and its production out of the resolved field. Such an equilibrium has more of a global character than a local one, however, since there are spatially intermittent regions of "backscatter" in which energy flows from small to large scales.

In an effort to correct for the purely dissipative form of the Smagorinsky model, there has been great interest in devising models that can adjust to local conditions in the flow, thereby producing backscatter when and if necessary. Such models are referred to as dynamic models. The approach of Germano *et al.* (1991) uses an extra coarse-grained filtering to sample the energy content of the smallest resolved scales as a means of adapting the coefficient C_S to local conditions.

Large-eddy simulations have been and continue to be applied to a range of flows at Reynolds numbers and complexity somewhat greater than are obtainable with DNS. A number of issues remain to be resolved before grid-based LES can be extended to engineering flows. One major problem is finding subgrid representations of the near-wall physics that permit a reduced grid coverage adjacent to boundaries. Finally, it should be noted that radically different formulations of LES are also possible, as discussed in Sec. 4.3. It remains to find out which of these methods is the best route to take for predicting turbulent flows.

5. SUMMARY

Driven by the wide range of technological problems requiring reliable prediction of turbulent forces, energies, and transport rates, the study of turbulence continues to be a priority among fluid dynamicists. A long history of experimental study has produced much valuable insight into many aspects of bounded and unbounded turbulent flows. In recent years this has been complemented by numerical schemes yielding detailed realizations of turbulent flow at modest Reynolds numbers in relatively uncomplicated geometries. Turbulence modeling can offer some benefit in predicting special classes of flows, though to obtain more accurate predictions, or treat flows in general complex configurations, approaches that are more closely aligned with the flow physics are required. Large-eddy simulations, in a variety of forms, attempt to balance the advantages of numerical simulations with the economies of small-scale modeling. Such methods are well poised to capitalize on recent advances in understanding the statistical equilibrium of vortical structures in the inertial range scales, and may offer the best long-range prospect for solving general turbulent-flow problems.

GLOSSARY

Anemometry: The measurement of velocity in fluids.

Aspect Ratio: The ratio of breadth to height when used to describe the geometry of a wind tunnel or channel in experimental fluid dynamics.

Backscatter: The flow of energy from small to large turbulent length scales or, equivalently, from high to low wave numbers.

Boundary Layer: A thin layer of fluid adjacent to boundaries where the mean velocity and/or temperature change rapidly with distance normal to the surface.

Braid Region: The region in boundary-free turbulent flows between two consecutive large spanwise-oriented vortices. In a frame of reference moving with the mean flow, a stagnation point occurs in this region, indicating a location of high compressive and extensional strain.

Buffer Layer: A thin layer near a solid surface in bounded flow, between the layer adjacent to the surface where the mean velocity varies linearly and the logarithmic layer.

Circulation: The line integral of the tangential component of the velocity vector around a closed circuit in a fluid. By the Stokes and divergence theorems, the circulation is also the integral of the normal vorticity component over an arbitrary surface whose boundary is the circuit. The circulation is the measure of the strength of a vortex tube.

Closure: A solvable system of equations formed from the averaged fluid-dynamics transport equations.

Computational Fluid Dynamics (CFD): Numerical methods for solving the time- and space-dependent equations describing fluid motion.

Core Region: The region above the logarithmic layer in bounded turbulent flows.

Correlation: The averaged product of two or more fluctuating flow variables at different times and/or spatial locations. When this product is normalized with respect to the standard deviations of the variables, it is called the correlation coefficient. In part it is a measure of the phase coincidence of the time series of the variables.

Diffusion: The movement of transportable properties such as heat (internal energy), momentum, species concentration, and vorticity through a fluid. Molecular diffusion occurs down gradients of the property. Turbulent diffusion is generally much greater and can be counter-gradient.

Direct Numerical Simulations (DNS): Numerical solutions of the Navier–Stokes equations for turbulent flow in which all the essential scales are resolved by a sufficiently fine computational grid.

Drag: The resistance to motion of a solid body in relative motion to a fluid caused by unbalanced viscous stresses and pressures acting on the body surface.

Eddy Turnover Time: A time scale over which some significant energy loss might occur in a turbulent flow.

Eddy Viscosity: Turbulent momentum-transport rate formulated in analogy to the molecular viscosity.

Energy Density: The kinetic energy contained in spherical shells of radius κ and thickness $d\kappa$, where $\kappa = |\mathbf{k}|$ is the magnitude of the wave-number vector.

Enstrophy: The average of the sum of the squares of the vorticity vector components.

Entrainment: The inclusion of surrounding nonturbulent fluid into a region of turbulent fluid.

Flux: The transport of a fluid or flow property across a surface either in the fluid or at its boundary.

Friction Velocity: A velocity scale defined in terms of the mean shear stress at the solid surface of a boundary layer. It is commonly used for scaling flow phenomena occurring near a solid surface.

Homogeneous Turbulence: Turbulent flow whose statistical properties are independent of spatial position.

Instability: The capacity of a fluid flow to become disordered in response to a small disturbance.

Integral Scales: A measure of the largest length or time scales in a turbulent flow. They are determined by integrating autocorrelation functions or two-point space or time correlation functions, respectively.

Irrotational Flow: Regions of a flow in which the curl of the velocity vector field is zero, i.e., where the vorticity is zero.

Isotropic Turbulence: Turbulent flow whose statistical properties have no preferred direction.

Kolmogorov Scale: A length scale of turbulence, named for the Russian mathematician Kolmogorov, characterizing dissipation of turbulent kinetic energy into heat.

Laminar Flow: Fluid flow lacking the disordered motions characteristic of turbulence. It may be thought of as the orderly motion of layers (laminae) of fluid moving relative to one another.

Large-Eddy Simulations (LES): A numerical method in which turbulent flow fields are computed by solving the filtered Navier–Stokes equations of fluid motion on a coarse grid. The large, energy-containing motions are computed while the more universal turbulent motions, of a size smaller than the grid resolution, are modeled.

Lift: The net fluid force on a solid body acting in the direction opposite to that of the force of gravity.

Logarithmic Layer: The flow layer in a bounded turbulent flow in which the mean velocity varies logarithmically with distance normal to the bounding surface.

Mixing Time: The time interval over which velocity fluctuations along fluid particle paths become decorrelated.

Momentum Thickness: A characteristic length scale of fluid motions measuring the momentum deficit in wake and boundary-layer flows.

Passive Scalar: A scalar property of a fluid having no influence on the velocity field.

RANS: The Reynolds averaged Navier–Stokes equations.

Reynolds Number: A nondimensional number, named for the British engineer and scientist Osborne Reynolds, defined as the ratio of inertial to viscous forces acting in a flow. Its magnitude characterizes the transition from laminar to turbulent flow and different regimes within fully turbulent flow.

Reynolds Stress: The correlation between products of pairs of fluctuating velocity components. It may be interpreted physically as an average momentum flux.

Schmidt Number: A nondimensional number formed by the ratio of the molecular viscosity to the mass concentration diffusivity. It measures the relative rates of momentum or vorticity diffusion to mass diffusion.

Solenoidal Vector Field: A divergence-free vector field.

Taylor Microscale: A turbulence-length scale, formulated by the British meteorologist and fluid dynamicist G. I. Taylor, that characterizes small-scale dissipative phenomena. It is defined in terms of a parabolic fit to two-point spatial correlation functions at their origin.

Turbulence Decay: The tendency for turbulent kinetic energy—in the absence of a production mechanism—to diminish in time through its dissipation into heat.

Turbulence Modeling: Procedures for deriving closure to the turbulent transport equations by approximating higher-order unclosed correlations (see **Closure**).

Turbulence Structures: Turbulent motions whose spatial and temporal features have recurring aspects that can be detected and characterized. These are frequently also called "coherent structures." Vortices are the most common turbulence structures.

Turbulent Kinetic Energy: Half the sum of the average variances of the turbulent velocity fluctuations.

Turbulent Transport: The net movement of fluid and flow properties such as heat, momentum, species concentration, and vorticity through a fluid by means of turbulent fluctuations (see **Diffusion**).

Viscous Sublayer: The very thin layer next to solid surfaces in bounded turbulent flows in which the mean velocity varies linearly. The relative magnitude of the turbulent fluctuation amplitudes to the mean velocity is greatest in this region.

Vortex Lines: Lines in a flow that are everywhere tangent to the vorticity vector field.

Vortex Strength: The circulation of a circuit enclosing a vortex tube. It is invariant along the tube.

Vortex Tube: A bundle of contiguous vortex lines, tubular in shape, passing through a closed circuit in a flow.

Vorticity: Formally the curl of the velocity vector, it is the measure of the rotative motion of fluid particles. It is a defining property of turbulence in that flow regions without vorticity are not turbulent.

Works Cited

Adrian, R. J. (1983), in: R. J. Goldstein (Ed.), *Fluid Mechanics Measurements,* Washington: Hemisphere Pub. Corp., pp. 155–244.

Adrian, R. J. (1991), *Annu. Rev. Fluid Mech.* **23,** 261–304.

Alfredsson, P. H., Johansson, A. V., Haritonidis, J. H., Eckelmann, H. (1988), *Phys. Fluids* **31,** 1026–1033.

Antonia, R. A., Kim, J. (1994), *Phys. Fluids* **6,** 834–841.

Barnhart, D. H., Adrian, R. J., Papen, G. C. (1994), *Appl. Optics* **33,** 7159–7170.

Batchelor, G. K., Townsend, A. A. (1948), *Proc. R. Soc. London* **194A,** 527–543.

Bell, J. H., Mehta, R. D. (1992), *J. Fluid Mech.* **239,** 213–248.

Bernal, L. P., Roshko, A. (1986), *J. Fluid Mech.* **170,** 499–525.

Bernard, P. S. (1985), *Int. J. Eng. Sci.* **23,** 1037–1057.

Bernard, P. S. (1990), *Theor. Comp. Fluid Dyn.* **2,** 165–183.

Bernard, P. S. (1995), in: *Proceedings of the Tenth Symposium on Turbulent Shear Flows,* University Park, PA: Penn State University, pp. 23-1–23-6.

Bernard, P. S. (1996), *FED* **238** (3), 33–40.

Bernard, P. S., Handler, R. A. (1990), *J. Fluid Mech.* **220,** 99–124.

Bernard, P. S., Rovelstad, A. L. (1994), *Phys. Fluids* **6,** 3093–3108.

Bernard, P. S., Speziale, C. G. (1992), *ASME J. Fluids Eng.* **114,** 29–39.

Bernard, P. S., Thomas, J. M., Handler, R. A. (1993), *J. Fluid Mech.* **253,** 385–419.

Blackwelder, R. (1981), in: R. Emrich (Ed.), *Methods in Experimental Physics: Fluid Dynamics Part A,* New York: Academic, pp. 259–314.

Brooke, J. W., Hanratty, T. J. (1993), *Phys. Fluids A* **5,** 1011–1022.

Brown, G. L., Roshko, A. (1974), *J. Fluid Mech.* **64,** 775–816.

Chapman, S., Cowling, T. G. (1952), *The Mathematical Theory of Non-Uniform Gases,* 2nd ed., London: Cambridge Univ. Press.

Chorin, A. J. (1993), *J. Comp. Phys.* **107,** 1–9.

Chorin, A. J. (1994), *Vorticity and Turbulence,* New York: Springer-Verlag.

Chorin, A. J., Hald, O. (1995), *Phys. Rev. D* **51,** 11969–11972.

Coles, D. (1956), *J. Fluid Mech.* **1,** 191–226.

Coles, D., Hirst, E. A. (1968), *Proceedings of Computation of Turbulent Boundary Layers,* vol. II, Department of Mechanical Engineering, Stanford University.

Corrsin, S., Kistler, A. (1954), NACA Technical Note No. 3133.

Dahm, W. J. A. Su, L. K., Southerland, K. B. (1992), *Phys. Fluids A* **4,** 2191–2206.

Dean, R. B., Bradshaw, P. (1976), *J. Fluid Mech.* **78,** 641–676.

Dekker, H., de Leeuw, G., Maassen van den Brink, A. (1995), *Phys. Rev. E* **52,** 2549–2558.

Egolf, P. W. (1994), *Phys. Rev. E* **49,** 1260–1268.

Fingerson, L. M., Freymuth, P. (1983), in: R. J. Goldstein (Ed.), *Fluid Mechanics Measurements,* Washington: Hemisphere Pub. Corp., pp. 99–154.

Germano, M., Piomelli, U., Moin, P. Cabot, W. H. (1991), *Phys. Fluids A* **3,** 1760–1765.

Goldstein, D., Handler, R., Sirovich, L. (1995), *J. Fluid Mech.* **302,** 333–376.

Gorski, J. J., Bernard, P. S. (1995), *ASME J. Fluids Eng.* **117,** 410–416.

Handler, R. A., Bernard, P. S., Rovelstad, A. L., Swearingen, J. (1992), *Phys. Fluids* **4,** 1317–1319.

Huser, A., Biringen, S., Hatay, F. (1994), *Phys. Fluids* **6,** 3144–3152.

Hussain, A. K. M. F. (1986), *J. Fluid Mech.* **173,** 303–356.

Johansson, T. G., Karlsson, R. I. (1989), in: R. J. Adrian, T. Asanuma, D. F. G. Durão, F. Durst, J. H. Whitelaw (Eds.), *Applications of Laser Anemometry to Fluid Mechanics,* Berlin: Springer Verlag, pp. 3–22.

Karlsson, R. I., Johansson, T. G. (1988), in: R. J. Adrian *et al.* (Eds.), *Laser Anemometry in Fluid Mechanics III,* Lisbon: Ladoan—Instituto Superior Tecnico, pp. 273–279.

Kim, J., Moin, P., Moser, R. (1987), *J. Fluid Mech.* **177,** 133–166.

Kolmogorov, A. N. (1941), *Dokl. Akad. Nauk SSSR* **30,** 299–303. English translation (1991), *Proc. R. Soc. London A* **434,** 9–13.

Konrad, J. H. (1976), Ph.D. Dissertation, The California Institute of Technology.

Kraichnan, R. H. (1958), *Phys. Rev.* **109,** 1047–1067.

Kraichnan, R. H. (1959), *J. Fluid Mech.* **5,** 497–543.

Laufer, J. (1953), NACA Technical Note No. 2954.

Launder, B. E., Spalding, D. B. (1974), *Comput. Methods Appl. Mech. Eng.* **3,** 269–289.

Lesieur, M. (1990), *Turbulence in Fluids,* 2nd ed., Dordrecht: Kluwer Academic.

Leslie, D. C. (1973), *Developments in the Theory of Turbulence,* Oxford, U.K.: Clarendon Press.

Lighthill, M. J. (1963), in: L. Rosenhead (Ed.), *Laminar Boundary Layers,* Oxford, U.K.: Clarendon Press, Chap. 2.

Mansour, N. N., Kim, J., Moin, P. (1988), *J. Fluid Mech.* **194,** 15–44.

McComb, W. D. (1990), *The Physics of Fluid Turbulence,* Oxford, U.K.: Clarendon Press.

Meng, H., Hussain, F. (1995), *Appl. Opt.* **34,** 1827–1840.

Miller, J. (1991), *Phys. Rev. Lett.* **65,** 2137–2140.

Miyake, Y., Tsujimoto, K. (1996), *FED* **238** (3), 41–48.

Montgomery, D., Joyce, G. (1974), *Phys. Fluids* **17,** 1139–1145.

Ong, L. (1992), Ph.D. Dissertation, The Univ. of Maryland at College Park.

Onsager, L. (1949), *Nuovo Cimento,* Suppl. to **6,** 279–287.

Panton, R. L. (1984), *Incompressible Flow,* 1st ed., New York: Wiley.

Panton, R. L. (1996), *Incompressible Flow,* 2nd ed., New York: Wiley.

Pearlstein, A. J., Carpenter, B. N. (1995), *Phys. Fluids* **7,** 754–763.

Prandtl, L. (1925), *Z. Angew. Math. Mech.* **5,** 136–138.

Prasad, A. K., Adrian, R. J. (1993), *Exp. Fluids* **15,** 49–60.

Pumir, A., Siggia, E. D. (1992), *Phys. Fluids A* **4,** 1472–1491.

Reynolds, A. J. (1974), *Turbulent Flows in Engineering,* New York: Wiley.

Reynolds, O. (1884), *Philos. Trans. R. Soc.* **174,** 935–982.

Reynolds, O. (1895), *Philos. Trans. R. Soc.* **186A,** 123–164.

Robert, R. (1990), *J. Stat. Phys.* **65,** 531–554.

Robinson, S. K. (1991), Ph.D. Dissertation, Stanford Univ., and NASA Technical Memorandum No. 103859.

Rogers, M. M., Moser, R. D. (1994), *Phys. Fluids* **6,** 903–923.

Sabot, J., Comte-Bellot, G. (1976), *J. Fluid Mech.* **74,** 767–796.

Saddoughi, S. G., Veeravali, S. V. (1994), *J. Fluid Mech.* **268,** 333–372.

Scherer, J., Bernal, L. P. (1992), AIAA Aerospace Sciences Meeting and Exhibit, 30th, Reno, NV, Jan. 6–9, 1992, *AIAA Paper* No. 92-0009.

Sedov, L. I. (1944), *Dokl. Akad. Nauk. SSSR* **42,** 116–119.

Smagorinsky, J. (1963), *Monthly Weather Rev.* **91,** 99–165.

Spalart, P. R. (1988), *J. Fluid Mech.* **187,** 61–98.

Speziale, C. G. (1987), *J. Fluid Mech.* **178,** 459–475.

Speziale, C. G. (1991), *Annu. Rev. Fluid Mech.* **23,** 107–157.

Speziale, C. G., Bernard, P. S. (1992), *J. Fluid Mech.* **241,** 645–667.

Su, L. K., Dahm, W. J. A. (1996), *Phys. Fluids* **8,** 1869–1906.

Taylor, G. I. (1915), *Philos. Trans. R. Soc. London* **215,** 1–26.

Taylor, G. I. (1932), *Proc. R. Soc.* **135A,** 685–705.

Taylor, G. I. (1935), *Proc. R. Soc.* **151A,** 421–444.

Townsend, A. A. (1956), *The Structure of Turbulent Shear Flow,* New York: Cambridge Univ. Press.

Tsuge, S. (1974), *Phys. Fluids* **17,** 22–33.

von Kármán, T., Howarth, L. (1938), *Proc. R. Soc. London* **164A,** 192–215.

Wallace, J. M. (1990), *Appl. Mech. Rev.* **43** (5), S201–S202.

Wallace, J. M., Brodkey, R. S. (1977), *Phys. Fluids* **20,** 351–355.

Wallace, J. M., Eckelmann, H., Brodkey, R. S. (1972), *J. Fluid Mech.* **54,** 39–48.

Wallace, J. M., Foss, J. F. (1994), *Annu. Rev. Fluid Mech.* **27,** 469–514.

Yakhot, V., Orszag, S. A. (1986), *J. Sci. Comput.* **1,** 3–51.

Further Reading

Batchelor, G. K. (1960), *The Theory of Homogeneous Turbulence,* Cambridge, U.K.: Cambridge Univ. Press. A classical text summarizing some early ideas on homogeneous turbulent flow.

Berge, P., Pomeau, Y., Vidal, C. (1986), *Order Within Chaos: Towards a Deterministic Approach to Turbulence,* New York: Wiley. An account of applications of chaotic dynamical systems to turbulence.

Bradshaw, P. (1971), *An Introduction to Turbulence and Its Measurement,* Oxford, U.K.: Pergamon Press. Provides a brief entree into the phenomena of turbulence with an emphasis on measurements.

Frost, W., Moulden, T. H. (1977), *Handbook of Turbulence,* Volume 1: *Fundamentals and Applications,* New York: Plenum Press. A collection of articles dealing with aspects of turbulence and transition.

Galperin, B., Orszag, S. A. (1993), *Large Eddy Simulation of Complex Engineering and Geophysical Flows,* Cambridge, U.K.: Cambridge Univ. Press. A selection of papers on recent trends in LES.

Hinze, J. O. (1976), *Turbulence,* 2nd ed., New York: McGraw-Hill. An extensive account of basic turbulence theory.

Lesieur, M. (1994), *La Turbulence,* Grenoble: Presses Universitaires de Grenoble. An introduction to turbulence in French for nonspecialists, including many results from recent direct numerical simulations.

Rotta, J. C. (1972), *Turbulente Strömungen,* Stuttgart: B. G. Teubner. A basic treatment of turbulent flows in German.

Taylor, G. I. (1958), in: G. K. Batchelor (Ed.), *Scientific Papers,* Cambridge, U.K.: Cambridge Univ. Press. Contains the many important contributions of Taylor to the statistical theory of turbulence.

Tennekes, H., Lumley, J. L. (1972), *A First Course in Turbulence,* Cambridge, MA: MIT Press. Develops the turbulent transport equations, discusses the statistical characterization of turbulence, and uses scaling arguments to obtain approximations of turbulent flow properties.

Townsend, A. A. (1976), *The Structure of Turbulent Shear Flow,* 2nd ed., Cambridge, U.K.: Cambridge Univ. Press. The second edition of a classic book on the physics of turbulence.

Tritton, D. J. (1988), *Physical Fluid Dynamics,* 2nd ed., Oxford, U.K.: Clarendon Press. Chapters 19, 20, and 21 present a useful account of turbulent-flow physics.

ULSI

See MICROELECTRONICS; VLSI—INTEGRATED CIRCUITS

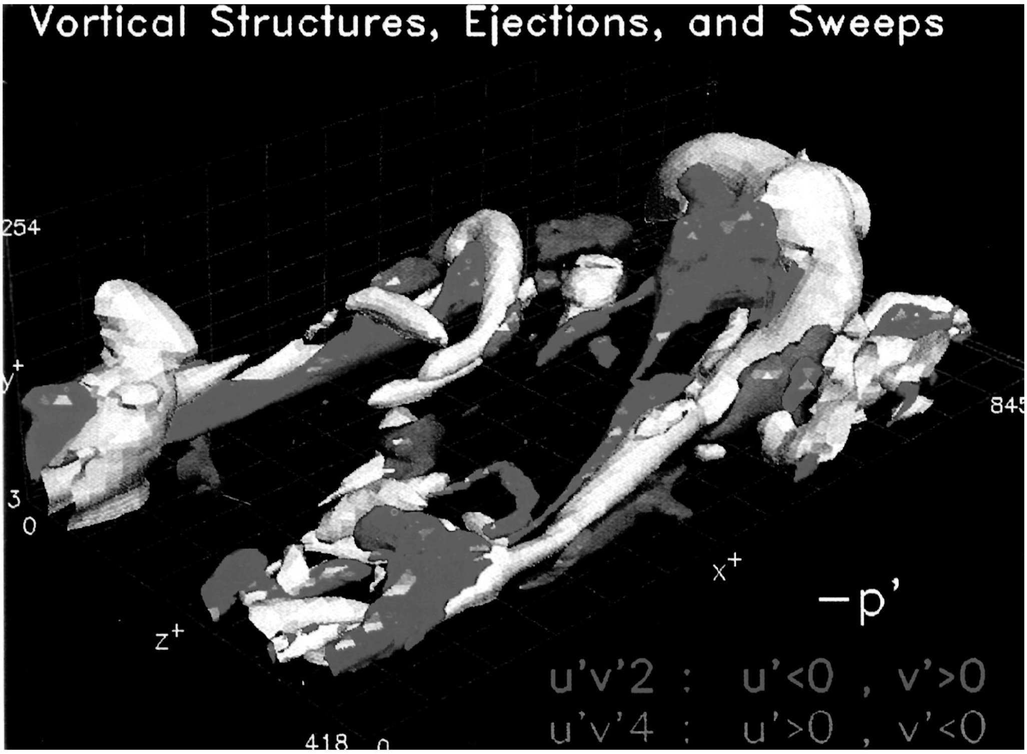
Vortical Structures, Ejections, and Sweeps
254
y+
3
0
845
x+
−p'
z+
418
0
u'v'2 : u'<0 , v'>0
u'v'4 : u'>0 , v'<0

Vol. 22 – PLATE 1. Surface contours regions of (white) low-pressure (threshold: $p^+ = -4.2$), (red) Q2 ($-u$, $+ v$ at threshold $uv_2^+ = -4.2$), and (blue) Q4 ($+u$, $-v$ at threshold $uv_4^+ = -4.2$) Reynolds shear stress, illustrating spatial relationship between Reynolds stress and wall-layer vortices. Courtesy of Robinson (1991).

22
Ultrafast Spectroscopy

Oliver Kühn and Stefan Lochbrunner

22.1 Introduction 771
22.2 Principles and Theoretical Concepts 772
22.2.1 Potential Energy Surfaces 772
22.2.2 Wave Packets 774
22.2.3 Theoretical Nonlinear Spectroscopy 776
22.2.3.1 Response Function Formalism 776
22.2.3.2 Nonperturbative Methods 778
22.3 Experimental Methods 778
22.3.1 Generation of Ultrashort Pulses 778
22.3.2 Probing with Linear Techniques 781
22.3.3 Probing Coherences 783
22.3.4 Pulse Shaping 784
22.4 Applications 786
22.4.1 Wave-Packet Dynamics in Small Molecules 786
22.4.1.1 Vibrational Wave Packets 786
22.4.1.2 Reactive Wave-Packet Dynamics 786
22.4.1.3 Influence of the Environment 787
22.4.2 Dynamics of Polyatomic Molecules 788
22.4.2.1 Vibrational Energy Redistribution and Cooling 788
22.4.2.2 Solvation 788
22.4.2.3 Electronic Dynamics: Internal Conversion, Conical Intersections, and Photochemistry 789
22.4.2.4 Coherent Wave-Packet Dynamics 790
22.4.2.5 Photoinduced Electron Transfer 792
22.4.3 Clusters and Liquids 793
22.4.3.1 Atomic Clusters 793
22.4.3.2 Water and Hydrogen-Bonded Networks 795
22.4.4 Biological Systems 795
22.4.4.1 Primary Steps of Photosynthesis 796
22.4.4.2 Rhodopsins 796
22.4.4.3 Peptide Dynamics 798

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

22.4.4.4 DNA 799
22.4.5 Solid State and Surfaces 801
22.4.5.1 Semiconductors 801
22.4.5.2 Polymers and Organic Materials 802
22.4.5.3 Surfaces 804
22.4.6 Laser Control 805
22.4.7 New Frontiers 807
22.4.7.1 Generation of High Harmonics and Attosecond Pulses 807
22.4.7.2 Application of High Harmonics and Attosecond Pulses 808
22.4.7.3 Carrier Envelope Phase and Frequency Comb 809
22.4.7.4 Time-Resolved X-Ray and Electron Diffraction 810
Acknowledgments 811
Glossary 811
References 811
Further Reading 815

22.1
Introduction

Ultrafast spectroscopy denotes the investigation of very fast processes with electromagnetic radiation at an extremely high time resolution. It covers the pico- and femtosecond timescale and has meanwhile reached the attosecond regime. The high impact of the field results from the fact that changes in all common forms of matter are related to the motion and rearrangement of nuclei and electrons. If we assume that an atom with a mass of 10 amu has a kinetic energy of 1 eV, then the atom needs about 2×10^{-14} seconds = 20 femtoseconds to transverse a distance of 1 Å. This shows that the femtosecond timescale is the natural scale of atomic motions and elementary processes in matter. This is also reflected by the time periods of vibrations in molecules and of phonons in condensed systems, which are in the range of 10–1000 femtoseconds.

The pump-probe technique is the most direct realization of ultrafast spectroscopy. An ultrashort light pulse, the pump pulse, excites the system under investigation and starts the process of interest (see Figure 22.2). A second ultrashort pulse, the probe pulse, arrives with a certain delay at the sample and makes a snapshot of the current stage of the process. To characterize the complete evolution of the process, the experiment has to be repeated with varying delay times. Using ultrashort light pulses instead of fast detectors and electronics results in an about 10^5 higher time resolution. To understand the observed dynamics, a description of an intrinsic nonstationary situation has to be given, usually on a quantum mechanical level, at least for a part of the system. The key elements that have to be considered are couplings between subsystems and superpositions of eigenstates and their evolution.

The development of this field started about 40 years ago when picosecond laser pulses became available. They were immediately applied to study photoinduced molecular processes, energy redistribution between vibrations, and vibrational coherences (Eisenthal, 1975; Laubereau and Kaiser, 1978; Zewail, 2000). The development of mode-locked dye lasers and in particular of the much more reliable Ti : sapphire lasers resulted in pulses with durations down to a few femtoseconds. As these systems were available, people started

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

to investigate basic photochemical processes on the subpicosecond timescale like dissociation of small molecules (Rosker, Dantus and Zewail, 1988; Rose, Rosker and Zewail, 1989; Zewail, 2000), electronic relaxation (Elsaesser and Kaiser, 1991), electron (Barbara, Walker and Smith, 1992) and proton transfer (Elsaesser and Kaiser, 1986) as well as isomerization reactions (Åberg *et al.*, 1994). These investigations promoted the understanding of molecular processes to a completely new level and were recognized by the award of the Nobel Prize in Chemistry to Ahmed Zewail in 1999 (Zewail, 2000). Owing to the development of new nonlinear conversion schemes (Riedle *et al.*, 2000) and the refinement of detection techniques, it is now possible to investigate almost every system with a time resolution down to 20 femtoseconds or even less and research dealing with such different topics like isolated molecules, solid-state physics, or biological systems make extensive use of ultrafast spectroscopy. Accordingly, almost every technique known from steady-state spectroscopy has meanwhile been transferred into the time domain. Broadband pulses provide, in addition, the unique possibility to construct very complex pulse shapes by manipulating the amplitude and phase of the frequency components in Fourier space. Using such appropriately designed electromagnetic fields for excitation allows to influence and control processes (Brixner and Gerber, 2003). By making attosecond pulses available, the time resolution was recently again dramatically improved (Corkum and Krausz, 2007). Those pulses are generated by recolliding electrons after an ultrashort photoionization event and the shortest pulse length reported so far is only 0.13 femtoseconds. Another current development is the implementation of femtosecond X-ray (Bargheer *et al.*, 2006) and electron diffraction techniques (Siwick *et al.*, 2003), which can measure the change of structures and might open the way to make a direct movie of ultrafast processes with subatomic spatial resolution.

Owing to the vastness of the field, it is not possible to give a comprehensive overview. We only describe a few selective examples to give an impression of the possibilities that the technique provides as well as the concepts that are important for theoretical modeling. Accordingly, we also restrict ourselves to weak field phenomena and have to exclude the fascinating area of high-intensity physics. The remainder of the article is organized as follows. Section 22.2 introduces the basic theoretical concepts used to describe ultrafast processes and experiments. In Section 22.3, some of the essential experimental tools and methods are sketched. In Section 22.4, different applications are described with a focus on molecular processes.

22.2
Principles and Theoretical Concepts

22.2.1
Potential Energy Surfaces

The quantum dynamics of a molecular system characterized by electronic coordinates $\mathbf{r}$ and nuclear coordinates $\mathbf{R}$ is determined by the time-dependent Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{r},\mathbf{R};t) = \hat{H}(t)\Psi(\mathbf{r},\mathbf{R};t) \quad (22.1)$$

where the Hamilton operator consists of a molecular part and an interaction term accounting for the coupling to the radiation field, $\mathbf{E}(t)$, here taken in dipole approximation ($\mathbf{d}$: dipole moment vector) :

$$\hat{H}(t) = \hat{H}_{\text{mol}} + \hat{V}_{\text{mol-field}}\,(t) \\ = \hat{T} + \hat{H}_{\text{el}} - \hat{\mathbf{d}} \cdot \mathbf{E}(t) \qquad (22.2)$$

Here, we introduced the nuclear kinetic energy, $\hat{T}$, and the electronic Hamiltonian, $\hat{H}_{\text{el}}$, which also contains the nuclear–nuclear repulsion. Using the Born–Oppenheimer separation of electronic and nuclear motions, the molecular wave function is written as

$$\Psi(\mathbf{r}, \mathbf{R}; t) = \sum_a \varphi_a(\mathbf{r}, \mathbf{R}) \psi_a(\mathbf{R}; t) \qquad (22.3)$$

where we introduced the adiabatic electronic states $\varphi_a(\mathbf{r}, \mathbf{R})$ and the nuclear wave packet on a given electronic state $\psi_a(\mathbf{R}; t)$. The latter is obtained from the set of coupled equations (May and Kühn, 2004)

$$i\hbar \frac{\partial}{\partial t} \psi_a(\mathbf{R}; t) \\ = \left(\hat{T} + \hat{V}_a(\mathbf{R}) - \hat{\mathbf{d}}_a(\mathbf{R}) \cdot \mathbf{E}(t) \right) \psi_a(\mathbf{R}; t) \\ + \sum_{b \neq a} \left(\hat{T}_{\text{ab}} - \hat{\mathbf{d}}_{\text{ab}}(\mathbf{R}) \cdot \mathbf{E}(t) \right) \\ \times \psi_b(\mathbf{R}; t) \qquad (22.4)$$

The first term describes the nuclear motion on the so-called potential energy hypersurface (PES) $V_a(\mathbf{R})$. This motion can be triggered, for example, by the coupling of an infrared (IR) laser field to the coordinate-dependent dipole operator of the considered electronic state. The laser field also triggers transitions between different electronic states via the transition dipole operator whose dependence on the nuclear coordinates is often neglected (Condon approximation), that is, $\hat{\mathbf{d}}_{\text{ab}}(\mathbf{R}) \approx \mathbf{d}_{\text{ab}}$. Transitions between adiabatic electronic states can also be caused by nonvanishing matrix elements of the nuclear kinetic energy. Such nonadiabatic transitions can become particularly important in situations where PESs of different electronic states come energetically close to each other. Nonadiabatic processes are at the heart of photochemistry. They cause internal conversion (IC) transforming electronic energy into vibrational energy, which comes along with a considerable shortening of the excited state lifetimes. Their description still provides a challenge to theory (Domcke, Yarkony and Köppel, 2004).

If the electronic states are well separated from each other, the transition probability follows the energy gap law, which results from the dependence of the Franck–Condon factors on the change of the vibrational quanta. On the other hand, nonadiabatic processes can be extremely efficient if a conical intersection is involved (Domcke and Stock, 1997; Domcke, Yarkony and Köppel, 2004). In this case, two electronic states are degenerate along a hyperline in the multidimensional space spanned by the nuclear coordinates. Of particular interest is the dynamics in the branching space, which is spanned by the coupling mode and the tuning mode. The nonadiabatic electronic coupling varies along the former and the energy gap between the corresponding so called diabatic states varies along the latter. This branching space is orthogonal to the hyperline of the intersection, while the shape of the adiabatic PESs in this space looks like a double cone (Figure 22.1). A conical intersection can be regarded as the extension of an avoided crossing into

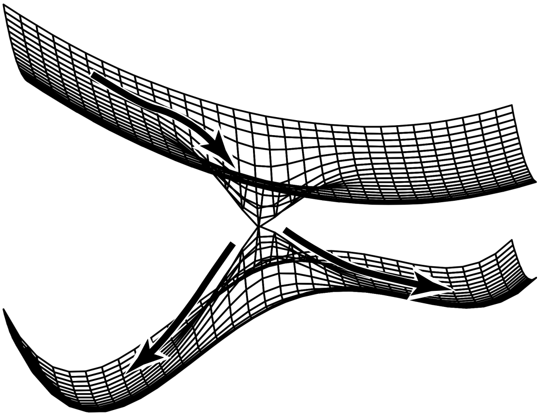

Fig. 22.1 Two PESs connected by a conical intersection. The relaxation pathway splits at the intersection and two minima corresponding to two different species are populated.

two dimensions. Owing to the double cone topology, a branching of the relaxation pathway at the intersection is quite likely. In this case, a part of the population usually reaches a minimum of the ground-state PES that differs from the original one and that is associated with a photoproduct. The rest of the population returns to the original minimum and completes the IC path. The speed of the process is given by the accessibility of the conical intersection, which depends on its energetic location and on whether the evolution of the excited-state population involves also coordinates that lead to the intersection. The quantum yield of a photoreaction depends on the dynamics at the intersection itself and how the wave packet propagates on the lower part of the double cone.

22.2.2
Wave Packets

As a consequence of the interaction with the laser field, the nuclear wave function will no longer describe a stationary molecular eigenstate on the considered PES, $\chi_{a,\xi}$ (**R**). Instead one has a coherent superposition (with coefficients $c_{a,\xi}(t)$) of such eigenstates called a *molecular wave packet*:

$$\psi_a(\mathbf{R};t) = \sum_{\xi} c_{a,\xi}\,(t)\;\chi_{a,\xi}\,(\mathbf{R}) \qquad (22.5)$$

Wave packets have been known since the early days of quantum mechanics. However, their observation in real time had to await the development of ultrafast spectroscopies during the last decades. In ultrafast spectroscopy, the molecule interacts with a sequence of N laser fields, that is,

$$\mathbf{E}(\mathbf{r};t) = \sum_{p=1}^{N} \mathbf{e}_p E_p(t-t_p) e^{-i(\omega_p t - \mathbf{k}_p\cdot\mathbf{r})} + \text{c.c.} \qquad (22.6)$$

Here, $\mathbf{e}_p$, ω_p, and $\mathbf{k}_p$ are polarization vector, carrier frequency, and wave vector, respectively, of the pth field. The pulse envelope, $E_p(t-t_p)$, is usually assumed to be of Gaussian shape centered at t_p.

In pump-probe spectroscopy, for instance, a first pump-pulse may excite a vibrational wave packet from the electronic ground to some excited

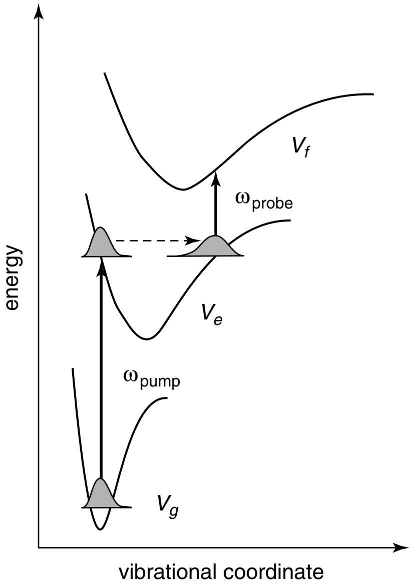

Fig. 22.2 Pump-probe spectroscopy of wave-packet dynamics along a vibrational coordinate. The pump-pulse with frequency ω_{pump} promotes the electronic ground-state (V_g) wave packet to the excited-state PES V_e, where it is probed by a time-delayed pulse with frequency ω_{probe} via a transition to a higher excited-state PES V_f.

state (see Figure 22.2). Owing to the different electronic structure, the nuclei are displaced from equilibrium and the wave packet no longer corresponds to a stationary state and propagates, that is, the bond is elongated. The ongoing dynamics can be interrogated by a time-delayed second pulse probing, for example, the transition to another electronic state from where fluorescence can be observed. The signal will reflect the oscillatory motion and dispersion of the wave packet in the intermediate state as well as concurring processes like photodissociation or nonadiabatic transitions to other electronic states.

Wave-packet dynamics as a modern research area has been initiated by Jortner and Berry's prediction of molecular quantum beats as a means for characterizing intramolecular radiationless transitions due to the breakdown of the Born–Oppenheimer approximation (Jortner and Berry, 1968). The development of numerical methods such as the split-operator approach and the utilization of fast-Fourier transforms for the calculation of the action of the kinetic energy operator onto the wave packet allowed for the efficient propagation of wave packets on general potentials in the early 1980s (Kosloff and Kosloff, 1982; Feit, Fleck and Steiger, 1982) marks a theoretical breakthrough, which enabled simulation of reactive processes at least for one-dimensional situations.

Straightforward extensions of these propagation methods face the limit of exponential scaling with the number of coupled degrees of freedom. This restriction is somewhat relaxed in the multiconfiguration time-dependent Hartree (MCTDH) method whose development has been a milestone toward the simulation of ultrafast multidimensional wave-packet dynamics (Meyer, Manthe and Cederbaum, 1990). Here, the wave packet for a D-dimensional system is expanded as follows:

$$\psi(\mathbf{R};t) = \sum_{j_1..j_D} A_{j_1..j_D}(t)\, \chi_{j_1}^{(1)}(R_1;t) \times \chi_{j_2}^{(2)}(R_2;t)\ \ldots\ \chi_{j_D}^{(D)}(R_D;t) \tag{22.7}$$

The use of a time-dependent basis set facilitates a flexible representation of the moving wave packet such that a considerably smaller number of basis functions is required as compared to the case of a fixed basis. Using this approach, the quantum dynamics of some tens of coordinates can be treated numerically exact; for special systems some thousand degrees of freedom are possible. Still, the propagation of the wave packet requires the knowledge of the complete PES including possible nonadiabatic couplings. Since its determination for complex systems is not feasible, methods that are based on classical trajectories with electronic structure theory-based forces determined on the fly have been developed. Quantum effects such as nonadiabatic transitions can be built into these simulations using the surface hopping method or semiclassical approaches (Stock and Thoss, 2005).

Despite the recent success of these all-atom simulations of complex systems at finite temperature the dynamics are more rigorously described by statistical methods such as the Liouville–von Neumann equation for the density operator $\hat{\rho}$. This facilitates, for instance, a reduced description of a relevant subsystem in interaction with a heat bath (May and Kühn, 2004). As a consequence of this, interaction energy and phase relaxation occurs, captured by the diagonal and off-diagonal elements of the reduced density matrix expressed in terms of the states of the relevant system {a}. The related damped wave-packet dynamics can be described in the simplest case of a Bloch model by

$$\frac{\partial \rho_{\mathrm{ab}}}{\partial t} = -i(\omega_{\mathrm{ab}} - i\gamma_{\mathrm{ab}})\rho_{\mathrm{ab}} - \delta_{\mathrm{ab}} \sum_c \Gamma_{\mathrm{ac}} \rho_{\mathrm{cc}} + i\mathbf{E}(t) \cdot \sum_c (\mathbf{d}_{\mathrm{ac}} \rho_{\mathrm{cb}} - \rho_{\mathrm{ac}} \mathbf{d}_{\mathrm{cb}}) \tag{22.8}$$

Here, ω_{ab} denotes the frequencies of transitions that are subject to dephasing with the rate γ_{ab}. The second term on the right-hand side describes the population redistribution with rates Γ_{ab} and the final term is due to the interaction with the laser field.

22.2.3
Theoretical Nonlinear Spectroscopy

22.2.3.1 Response Function Formalism

The different nonlinear spectroscopic techniques depend on the time delay between the pulses as well as on the phase-matching direction $\mathbf{k}_s$ for the detection of the macroscopic polarization field, which is defined as the expectation value of the dipole operator (n_{mol} : volume density of molecules in sample):

$$\mathbf{P}(\mathbf{r};t) = n_{\mathrm{mol}} \langle \hat{\mathbf{d}}(\mathbf{r};t) \rangle \tag{22.9}$$

The response function formalism is based on an expansion of the polarization in powers of the electric field strengths ("tr" denotes the trace operation)

$$\mathbf{P}(\mathbf{r};t) = \sum_n \mathbf{P}^{(n)}(\mathbf{r};t) = n_{\mathrm{mol}} \sum_n \mathrm{tr}\left[\mathbf{d}\hat{\rho}^{(n)}(\mathbf{r};t)\right] \tag{22.10}$$

Thus the nth order polarization is determined by a perturbation expansion of the density operator with respect to $\hat{V}_{\text{mol-field}}(t)$. This leads to

$$\mathbf{P}^{(n)}(\mathbf{r};t) = n_{\text{mol}} \int_0^\infty dt_n \ldots \int_0^\infty dt_1 \times S^{(n)}(t_n, \ldots, t_1)\mathbf{E}(\mathbf{r}; t - t_n) \times \ldots \times \mathbf{E}(\mathbf{r}; t - t_n - \ldots - t_1) \tag{22.11}$$

where the material properties are completely described by the response functions $S^{(n)}$, which are given by nth order correlation functions of the molecular dipole operator (Mukamel, 1995). For instance, the first-order response function that contains the information about linear spectroscopy is given by

$$S^{(1)}(t_1) = \frac{i}{\hbar}\theta(t_1)\,\text{tr}\,([\hat{\mathbf{d}}(t_1), \hat{\mathbf{d}}(0)]\hat{\rho}_{\text{eq}}) \tag{22.12}$$

where $[\hat{A}, \hat{B}]$ denotes the commutator and $\hat{\rho}_{\text{eq}}$ is the equilibrium density operator. For the case of an electronic two-level system coupled to intra- and intermolecular degrees of freedom Eq. (2.12) becomes (ω_{eg} is the transition frequency between states g and e, d_{eg} the transition dipole matrix element, and the expectation value is taken with respect to the ground state density)

$$S^{(1)}(t_1) = \frac{i}{\hbar}\theta(t_1)\left|d_{\text{eg}}\right|^2 \times \left(\exp\left[-i\langle\omega_{\text{eg}}\rangle t_1 - g(t_1)\right] - \text{c.c.}\right) \tag{22.13}$$

with the line shape function

$$g(t) = \frac{1}{\hbar^2}\int_0^t dt' \int_0^{t'} dt'' \langle\delta\omega_{\text{eg}}(t'')\delta\omega_{\text{eg}}(0)\rangle \tag{22.14}$$

Here, it has been assumed that the fluctuation of the transition frequency due to the interaction of the two-level system with its surroundings, $\omega_{\text{eg}}(t) = \langle\omega_{\text{eg}}\rangle + \delta\omega_{\text{eg}}(t)$, obey Gaussian statistics. In this formulation, the line shape function is the central quantity. It also enters the nonlinear response functions and is the starting point for simulations. Apart from models such as the multimode Brownian oscillator (Mukamel, 1995), the correlation functions can also be obtained from classical molecular dynamics simulations.

Comparing Eqs (22.6) and (22.11), we notice that the polarization can be written as

$$\mathbf{P}^{(n)}(\mathbf{r}, t) = \sum_L \mathbf{P}_L^{(n)}(t)e^{-i(\Omega_L t - \mathbf{K}_L\cdot\mathbf{r})} \tag{22.15}$$

Here, $L = \{l_1, l_2, \ldots, l_n\}$ comprises the numbers indicating possible combinations of fields for a given order n, $\Omega_L = \sum_p l_p\omega_p$, and $\mathbf{K}_L = \sum_p l_p\mathbf{k}_p$. In principle, a self-consistent solution of Schrödinger and Maxwell equations is required since the polarization enters as a source term into the latter. However, for optically thin samples, changes of incident fields in the medium can be neglected. Expanding the signal field into a Fourier series analogous to Eq. (22.15), one finds that the Fourier component along a certain direction $\mathbf{k}_s$ is proportional to the respective polarization term in Eq. (22.15) with wave vector selectivity introduced by the phase-matching condition, $\mathbf{k}_s \approx \mathbf{K}_L$. As an example, consider a third-order pump-probe experiment with $\mathbf{k}_1 = \mathbf{k}_2 = \mathbf{k}_{\text{pump}}$ and $\mathbf{k}_3 = \mathbf{k}_{\text{probe}}$. If the signal is detected in phase-matched probe direction, one has only terms with $L = \{\pm 1, \mp 1, 1\}$. The transient absorption signal can be calculated in slowly varying envelope approximation via

$$S(\omega_{\text{probe}}) = \frac{4\pi\,\omega_{\text{probe}}}{cn(\omega_{\text{probe}})} \times \frac{\operatorname{Im}\int dt\, E^*_{\text{probe}}(t) P_L^{(3)}(t)}{\int dt \left|E_{\text{probe}}(t)\right|^2} \tag{22.16}$$

where n is the index of refraction and we have skipped the vector character (Mukamel, 1995).

22.2.3.2 Nonperturbative Methods

In more complex situations that neither can be mapped onto the Brownian oscillator Hamiltonian nor can a diagonalization be performed, calculation of the multitime response functions becomes too tedious and a direct propagation approach is used (Domcke and Stock, 1997). This has also the advantage that it is, in principle, nonperturbative, that is, strong field effects can be considered. The idea can be illustrated for a pump-probe experiment with two incident pulses. To arbitrary order in the two fields, the polarization can be written as

$$\mathbf{P}(\mathbf{r}, t) = \sum_{n_1,n_2=-\infty}^{+\infty} \mathbf{P}_{n_1,n_2}(t) e^{i(n_1\mathbf{k}_{\text{pump}}+n_2\mathbf{k}_{\text{probe}})\cdot\mathbf{r}} \tag{22.17}$$

Defining the phases $\mathbf{k}_{\text{pump}}\mathbf{r} = \varphi_{\text{pump}}$ and $\mathbf{k}_{\text{probe}}\mathbf{r} = \varphi_{\text{probe}}$, the Schrödinger equation is solved using the time-dependent Hamiltonian where the wave vector dependence in the fields (22.6) is replaced by these phases. Calculating the polarization from Eq. (22.9), a certain component of the polarization can be extracted by inverting Eq. (22.17). For instance, $\mathbf{P}_{0,1}(t)$, contains the polarization component for the $\mathbf{k}_{\text{probe}}$ direction in any order with respect to the fields.

In combination with efficient propagation methods such as MCTDH, this direct approach is an important step toward first-principle quantum simulations of ultrafast spectroscopy (see, e.g., study of ultrafast electron transfer (Wang and Thoss, 2008)).

22.3
Experimental Methods

Figure 22.3 shows a typical layout of an ultrafast spectroscopic experiment. A mode locking oscillator generates ultrashort laser pulses that are subsequently amplified and transferred to other spectral regions by nonlinear optical processes. At some point of the setup, the laser beam is split into two arms to obtain a pump and a probe pulse, which are perfectly synchronized to each other. The pump pulse excites the sample while the interaction of the probe pulse is used to measure pump-induced coherences and population changes in the sample. Typically, a motorized delay line is incorporated in the path of the pump beam. By changing the pump delay, the time between the arrival of the pump and the probe pulses at the sample is systematically varied. The temporal evolution of a system after photo excitation is then reconstructed from the single point measurements taken at different delays. In the following, we shortly describe the components and techniques applied in such experiments.

22.3.1
Generation of Ultrashort Pulses

Ti : sapphire and fiber oscillators are commonly used to generate ultrashort

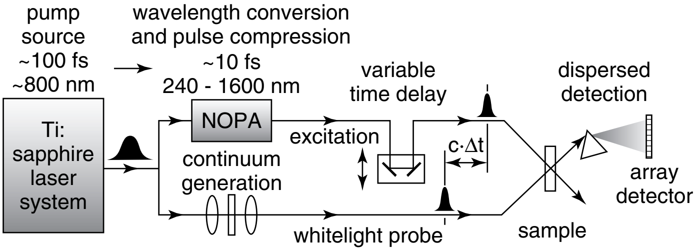

Fig. 22.3 Scheme of an ultrafast pump-probe experiment.

laser pulses at high repetition rates. Their output is directly used in applications that need only low energy pulses in a restricted wavelength range. Their layouts are shown in Figure 22.4. In these resonators, locking of longitudinal cavity modes and the formation of pulses occur since the pulses experience a higher gain or lower losses compared to less intense continuous-wave radiation. In the case of the Ti : sapphire oscillator, the intensity-dependent Kerr lens in the sapphire crystal causes for pulses a better spatial overlap with the tightly focused pump beam resulting in a higher gain (Spence, Kean and Sibbett, 1991). In the fiber oscillator, intensity-dependent birefringence of the fiber leads, in combination with appropriately aligned polarization tuning elements for pulses, to a higher transmission through the polarizer insulator (Tamura *et al.*, 1993). To generate short pulses, the total group delay dispersion of the resonator must be set to zero. In the Ti : sapphire oscillator, a prism sequence or chirped mirrors compensate for the dispersion of the other elements. The output pulses have a center wavelength of about 800 nm, energies of several nanojoules and can be as short as a few femtoseconds. In the fiber oscillator, dispersion compensation results from the combination of sections of a standard fiber that exhibits negative dispersion with an Erbium-doped fiber, which is the gain medium and has positive dispersion. The output is in the spectral region around 1550 nm and is usually frequency doubled if it serves as seed of an amplifier system.

To achieve pulse energies in the millijoules region, the oscillator output is amplified, typically with a regenerative amplifier operating at a repetition rate of 1 kHz (see Figure 22.5) (Backus *et al.*, 1998). Using a thin film polarizer (TFP), the seed pulse is coupled into the resonator of the amplifier. If a suitable voltage step is applied to the Pockels cell, the polarization of the pulse changes and it is captured in the resonator where it makes several round trips. A Ti : sapphire crystal pumped by a pulsed and frequency-doubled YLF or YAG laser serves as gain medium. When the inversion of the crystal is depleted, the pulse is released from the resonator by a second voltage change at the Pockels cell. The seed pulses are stretched to about hundred picoseconds by a grating stretcher before they enter the resonator. Otherwise, the high intensities of the amplified pulses would lead to nonlinear distortions of the pulse shape and to damage of the optical elements within the resonator. After the amplifier, a grating compressor removes the chirp introduced by the stretcher and the

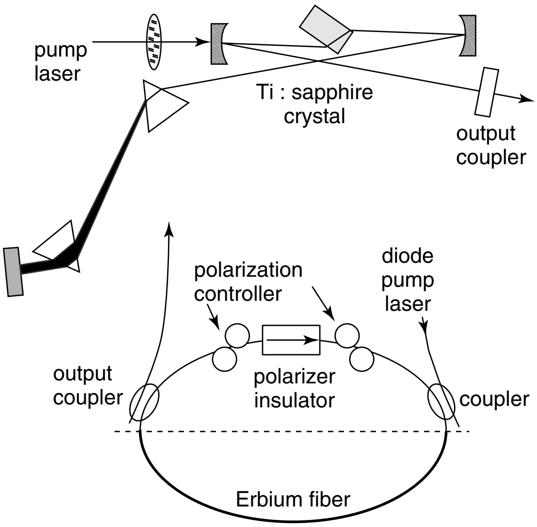
pump
laser
Ti : sapphire
crystal
output
coupler
polarization
controller
diode
pump
laser
output
coupler
polarizer
insulator
coupler
Erbium fiber

Fig. 22.4 Setup of a Ti : sapphire oscillator and a fiber oscillator.

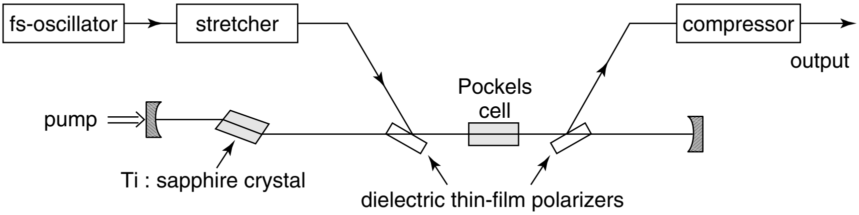
fs-oscillator
stretcher
compressor
output
Pockels
cell
pump
Ti : sapphire crystal
dielectric thin-film polarizers

Fig. 22.5 Scheme of a regenerative Ti : sapphire amplifier.

dispersion of the resonator resulting in pulse duration down to a few tens of femtoseconds.

Many different schemes have been successfully implemented to access other wavelength regions and to shorten the pulse duration further. Here, we restrict ourselves to an approach that is notably versatile and is being extensively applied in spectroscopic experiments. Its central element is the optical parametric amplification of a white light contiuum (Riedle *et al.*, 2000; Cerullo and De Silvestri, 2003), see Figure 22.6. Pulses from a Ti : sapphire amplifier system with a pulse length in the order of 100 femtoseconds are typically used as input. A small fraction in the order of 2 μJ is split off and tightly focused into a sapphire substrate where self-focusing and self-phase modulation lead to the generation of a white light continuum. The remainder of the input beam is frequency doubled and serves as pump light for the parametric amplifier. The nonlinear medium of the amplifier is a $\beta - \mathrm{B_aB_2O_4}$ (BBO) crystal cut for type I phase matching. The white light and the pump beam cross each other at a small angle of about 3° in the nonlinear crystal. Choosing the appropriate nonlinear geometry results in the remarkable situation where the phase-matching condition becomes, to first order, independent of the seed wavelength (Riedle *et al.*, 2000). Then a very broad amplification bandwidth is

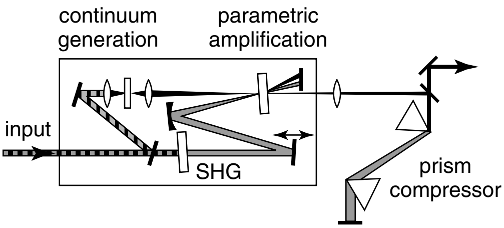

Fig. 22.6 Layout of a noncollinearly phase-matched optical parametric amplifier (NOPA) (Riedle *et al.*, 2000).

possible even with a thick crystal and a high gain for broadband pulses and pulse energies up to several 10 μJ can be achieved.

The dispersion of the amplified pulses is controlled with a prism compressor resulting in pulse durations down to 10 femtoseconds (see Figure 22.6). Using more sophisticated compression schemes, pulse durations as short as 5 femtoseconds have been demonstrated (Cerullo and De Silvestri, 2003). Since the white light has some chirp, the output of the noncollinearly phase-matched optical parametric amplifier (NOPA) can be easily tuned over the whole visible region and even far into the near infrared (NIR) by slightly changing the path length of the pump beam and thereby shifting its temporal overlap with the white light. The NOPA pulses can be efficiently frequency doubled extending the accessible wavelength range also far into the ultraviolet (UV).

22.3.2
Probing with Linear Techniques

Meanwhile, many different techniques are available to probe the pump pulse-induced changes of a system and their evolution with time. In this paragraph, techniques that are linear in the energy of the probe pulse are considered. The most straightforward one is perhaps the transient absorption spectroscopy (see Figure 22.3). A probe pulse transverses the sample in the same region as the pump and its energy loss due to absorption within the sample is measured. Very often, a white light continuum generated in a sapphire or a CaF_2-substrate is used as probe since the white light covers the complete spectral region from the NIR to the near UV and only very low pulse energies are necessary for probing (Kovalenko *et al.*, 1999). After the sample, the probe beam is dispersed by an imaging spectrograph and transmission changes over a broad spectral range are measured simultaneously with an array detector. An interesting feature of the technique is that absorption changes in the sample can be extracted in absolute numbers using Eq. (22.18).

$$\frac{Sig(\lambda)}{Sig_0(\lambda)} = \frac{T(\lambda)}{T_0(\lambda)} = e^{-n_{\mathrm{ex}}\cdot(\sigma(\lambda)-\sigma_0(\lambda))\cdot d} \tag{22.18}$$

$Sig(\lambda)$ is the dispersed probe signal measured with pump pulses hitting the sample and $Sig_0(\lambda)$ the signal when the pump beam is blocked. $T(\lambda)$ and $T_0(\lambda)$ stand for the corresponding sample transmissions. n_{ex} is the number density of excited particles, $\sigma(\lambda)$ is the time-dependent absorption cross section of the excited particle, $\sigma_0(\lambda)$ is its absorption cross section in the original state, and d is the sample thickness. The signal $S(\omega_{\mathrm{probe}})$ introduced in Section 2.3.1 is directly linked to these quantities by Eq. (22.19) and the experimental results can be compared to simulations by means of Eq. (22.16).

$$S(\omega_{\text{probe}} = 2\pi c/\lambda) = \frac{n_{\text{ex}}}{n_{\text{mol}}} \cdot (\sigma_0(\lambda) - \sigma(\lambda)) \tag{22.19}$$

Here, n_{mol} is the number density of the particles.

In Figure 22.7, the transient spectrum after excitation of a perylene bisimide dye is compared to the steady-state absorption and fluorescence spectrum (Schlosser and Lochbrunner, 2006). The depicted change of the optical density $\Delta OD = -\log(T/T_0)$ clearly shows the three contributions to the signal: (i) the bleach of the electronic ground state due to the excitation, (ii) stimulated emission from the electronically excited state to the ground state, and (iii) the excited state absorption caused by electronic transitions from the excited state to higher lying states. The bands contributing to the transient absorption in the visible and UV reflect the electronic state of a system.

Changes in vibrational structure and population show up as variations in the band shapes. However, those can be more specifically measured using an infrared pulse as probe pulse (Elsaesser and Kaiser, 1991). Another possibility is to measure the Raman signal associated with the probe pulse (Laubereau and Kaiser, 1978) or to perform stimulated Raman measurements by applying an additional picosecond pulse for pumping the Raman transition and by recording the stimulated emission at the Raman frequencies with the femtosecond probe pulse (Kukura, McCamant and Mathies, 2007).

Usually, several contributions of the transient absorption overlap spectrally. Then it is useful to look at the time-dependent emission. In most cases, this is done by fluorescence up-conversion (Yoshihara, Tominaga and Nagasawa, 1995; Schanz *et al.*, 2001).

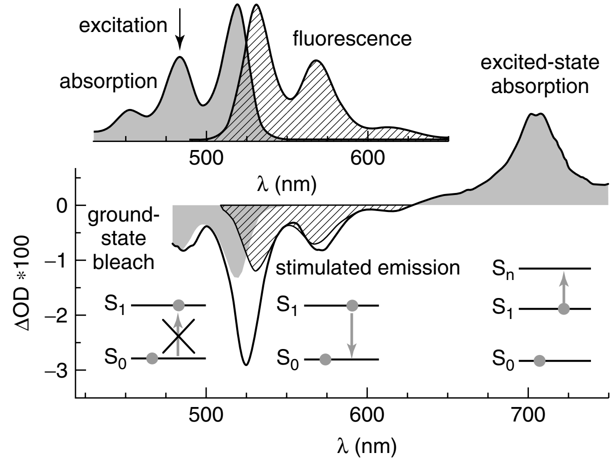

Fig. 22.7 Steady-state absorption and fluorescence spectrum of a perylene bisimide dye dissolved in chloroform and the transient absorption spectrum measured 5 picoseconds after optical excitation at 480 nm.

The spontaneous emission of the excited sample is imaged onto a nonlinear crystal. There, it is spatially overlaid with the probe pulsethat functions as a gate pulse in this case. The fraction of the emission that overlaps temporally with the gate pulse is subject to sum-frequency mixing and is transferred to shorter wavelengths. The strength and the spectrum of the sum frequency signal are measured with a spectrograph in dependence on the arrival time of the gate pulse at the crystal. The evolution of the emission spectrum is then directly reproduced by the time-dependent sum frequency signal.

For experiments in the gas phase or in a molecular beam, a very sensitive probe process is necessary due to the low sample densities. In this case, ionization techniques are used since ions and electrons can be detected with a very high efficiency (Stolow, Bragg and Neumark, 2004). One observable is the time-dependent mass spectrum, which reveals changes in the ionization probability and fragmentation pattern. Another one is the time-dependent photoelectron spectrum. It reflects the shape of the absorption bands associated with transitions from the populated electronic states to ionic states. Meanwhile powerful imaging and coincidence techniques were developed. They allow to assign each observed photoelectron to specific particles and to extract, besides the kinetic energy, the angular distribution, which is sensitive to the orbital from which the electron is emitted.

22.3.3
Probing Coherences

The probe schemes described above are not suitable to investigate electronic coherences. This is the domain of four-wave mixing (FWM) processes and interferometric techniques. In addition, nonlinear techniques of higher order have been applied. However, their discussion is beyond the scope of this article.

The most common realizations of ultrafast FWM spectroscopy are photon echo and two-dimensional spectroscopy. Three pulses are focused from somewhat different directions onto a common spot of the sample (see Figure 22.8) (Weiner, De Silvestri and Ippen, 1985). The first pulse induces a coherence between the ground and the excited state. The second pulse interacts with the coherence and creates a population in the excited-state the coherence or removes depending on the relative phase. A spatially modulated excited-state population results and constitutes a grating with the grating vector $\mathbf{k}_1 - \mathbf{k}_2$. The third pulse is diffracted by this grating and as signal a photon

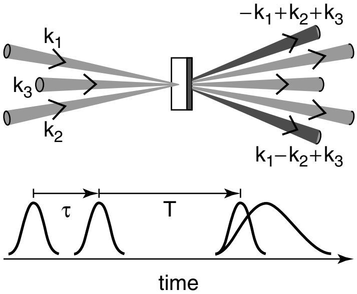

Fig. 22.8 Geometry and time diagram of a photon echo experiment.

echo with the wavevector $\mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3$ is generated. The time τ between the first and second pulse is often called *coherence time* since the signal strength depends on the coherence decay during this time interval. The time T between the second and the third pulse is named *population time* since during this interval the evolution of the populations is relevant. The role of the first and the second pulse can be exchanged and a second echo is emitted in the direction of the wavevector $-\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3$. Its dependence on the coherence time is mirrored compared to the former echo signal. In transient grating experiments, the first two pulses are simultaneously applied, that is, the coherence time is set to zero and not varied, and only the dependence on the population time is measured. Photon echo studies focus on the coherence time and deal predominantly with dephasing phenomena.

In photon echo experiments, usually the three applied pulses are identical. However, the arrangement can also be used for coherent Raman scattering. In the anti-Stokes configuration, the center frequency of the second pulse is shifted to lower frequencies with respect to the first pulse (Schmitt *et al.*, 1998). If the difference frequency fits to a vibrational mode, the vibrational resonance leads to a strongly increased diffracted signal at frequencies on the anti-Stokes side of the third pulse.

To extend the photon echo measurement to two-dimensional spectroscopy, a fourth laser pulse acting as local oscillator is interferometrically superimposed on the photon echo (Jonas, 2003). The resulting heterodyne signal is spectrally resolved and the coherence time τ is scanned keeping the population time T fixed. For each echo wavelength, the time-dependent signal is Fourier transformed. The result is a two-dimensional complex valued plot of the echo signal in dependence of the excitation and the signal frequencies. It shows which photon energies in the excitation contribute to the echo signal at a certain wavelength and if coherence transfer between excitation and echo generation occurs. The same approach using IR pulses is applied to investigate couplings between vibrational modes (Hamm, Helbing and Bredenbeck, 2008).

Wave-packet interference is another approach to look at the evolution of excitation-induced coherences. The sample is excited by two phase-locked collinear propagating pump pulses usually generated by means of a Michelson interferometer (Scherer *et al.*, 1991). The first pulse generates a wave packet that should at least partially return to the Franck–Condon region when the second pulse hits the sample. Then the degree of coherence between excited and ground state and the relative phase with respect to the field of the second pulse determine whether constructive, destructive, or incoherent superposition occurs. The corresponding signature is the appearance of interference fringes in scans of the delay between the excitation pulses.

22.3.4
Pulse Shaping

Ultrashort laser pulses constitute a broad band of phase-locked frequency components. If one has access to each of these components and one can tune its amplitude and phase, then it is possible to

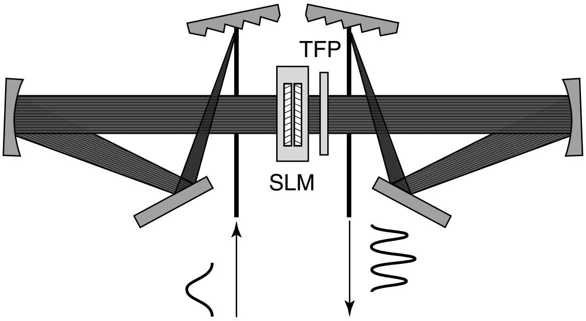

Fig. 22.9 Setup of a pulse shaper. SLM denotes a spatial light modulator containing two liquid crystal arrays and TFP a thin film polarizer.

generate completely arbitrary waveforms and pulse shapes. Only the carrier frequency would be similar and the shortest possible features are given by the original band width. Figure 22.9 shows a popular 4*f*-setup, which allows manipulating with high resolution amplitude and phase of the spectral components of a pulse. (Weiner, 2000). The incoming pulse is spectrally dispersed by a grating. A cylindrical mirror or lens placed in a distance to the grating equal to its focal length *f* collimates the beam and then focuses each spectral component on its own spot. The focal plane represents now a Fourier plane of the pulse. A modulator, typically a liquid crystal mask, consisting of two layers is placed in this plane. The birefringence of each mask pixel is separately controlled via applied voltages and thereby the phase shift of the transmitted light and, in combination with a TFP, also its amplitude. The spectral components are recombined via a second cylindrical lens or mirror and a second grating. The pulse spectrum is now composed of independent tunable components that can be as many as there are modulator pixels.

Pulse-shaping techniques opened the way to new experimental approaches that use optimized laser pulses to drive and control photoinduced processes. Usually they are implemented by means of a feedback loop. Using a shaperpulses with different shapes are applied to excite the system under investigation. The results of the excitations are analyzed and compared to each other. The pulse shapes giving the best results are selected, in different ways modified, and again applied. The procedure is repeated until no more significant improvement is observed. Meanwhile, many light-driven processes have been controlled in this way (Brixner and Gerber, 2003). This includes nonlinear frequency conversion, high harmonic generation, molecular processes, and imaging. In case several pathways interfere with each other, coherences in the system are relevant for the result. If the phase properties of the light pulse influence the relative phases between the pathways, the resulting interference is very sensitive on the pulse and coherent control of the process can be achieved (Rabitz *et al.*, 2000).

22.4
Applications

22.4.1
Wave-Packet Dynamics in Small Molecules

22.4.1.1 Vibrational Wave Packets

Small molecules in the gas phase provide good test cases for the concepts of nonstationary quantum mechanics. The time evolution of one-dimensional wave packets was, for example, studied in Na_2 (Baumert *et al.*, 1991). A 70-femtosecond-long laser pulse at 627 nm excites Na_2-molecules in a supersonic beam. A duplicate of the pulse is used to ionize the molecule via two-photon absorption at varying delay times (see Figure 22.10).

The $Na_2{}^+$-ion signal detected with a time of flight mass spectrometer shows a strongly oscillating time dependence, which is a clear signature of a wave-packet propagation in a bound potential. The spectrum of the excitation pulse overlaps with several vibrational levels of the $A^1\Sigma$ state and creates a coherent superposition of the vibronic states with the vibrational numbers $v' = 10–14$. The ionization probability is at the inner turning point of the $A^1\Sigma$ potential higher than at the one outer due to resonances with higher electronic states. Each recurrence of the wave packet at the inner turning point results in a transient rise of the $Na_2{}^+$ signal. The time delay between subsequent ionization maxima is 306 femtoseconds, which matches very well with the vibrational period that is given by the energetic separation of 109 cm^{-1} between the vibrational states. The long lifetime of the oscillations indicates that the wave packet broadens only slowly due to the quite moderate anharmonicity of the $A^1\Sigma$ potential at the involved vibrational levels.

22.4.1.2 Reactive Wave-Packet Dynamics

Couplings between electronic states and associated reaction channels strongly modify the wave-packet dynamics. The

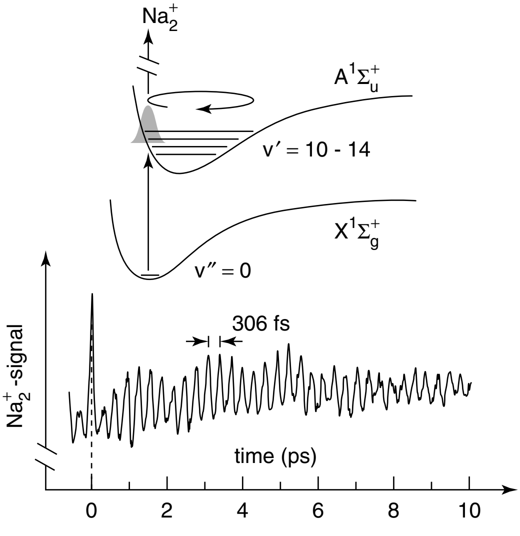

Fig. 22.10 Generation of a wave packet in Na_2 and oscillations in the ion signal due to the wave-packet propagation. (Adapted from Baumert, *et al.*, 1991.)

observed dynamics can then be used to analyze the reaction mechanisms. A one-dimensional example is the photodissociation of NaI (Zewail, 2000; Rose, Rosker and Zewail, 1989). A pump pulse at 307 nm launches in the first electronically excited state a wave packet, which then propagates toward the outer turning point (Figure 22.11). Thereby, it passes through an avoided crossing resulting from the diabatic covalent and the diabatic ionic PES. According to the Landau–Zener mechanism, the wave packet splits into two with one staying in the upper state while the other propagates along the lower covalent PES resulting in the formation of atomic iodide and sodium. The former is reflected by the binding ionic PES, returns to the Franck–Condon region, and is again reflected there. In this way, the wave-packet splitting is repeated several times until the excited state of NaI is depopulated.

With laser-induced fluorescence, the population of atomic sodium and electronically excited NaI can be followed in time (Zewail, 2000; Rose, Rosker and Zewail, 1989). At a probe wavelength of 589 nm, the resonance fluorescence of atomic sodium is dominant and one observes steplike increases of the sodium population every time the wave packet splits at the avoided crossing. In case the probe wavelength is off resonant the fluorescence of NaI dominates. The data shows recurrences of the wave packet which decay exponentially due to the losses at the avoided crossing.

If a laser pulse prepares a wave packet on a repulsive PES, the wave packet propagates in a ballistic fashion. The motion of its center of gravity follows almost purely classical dynamics and is given by the inertia of the involved nuclei and the slope of the PES. Accordingly, the dissociation products appear with some delay after the optical excitation. In addition, the wave packet exhibits broadening as a result of its wave nature. All these features have been demonstrated, for example, in the case of the photodissociation of ICN into I and CN (Rosker, Dantus and Zewail, 1988).

22.4.1.3 Influence of the Environment

In a dense environment, the wave-packet dynamics is strongly influenced by the interaction of the excited molecule with adjacent atoms or molecules. Several effects can occur (Gühr *et al.*, 2007). The forces exerted by the environment change the molecular PESs

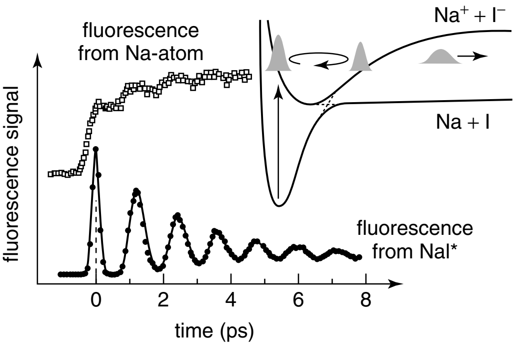

Fig. 22.11 Time-resolved observation of the photodissociation of NaI and the corresponding potential energy diagram (Zewail, 2000; Rose, Rosker and Zewail, 1989). The resonance fluorescence of the Na-atoms (squares) shows a stepwise buildup and the NaI-fluorescence a decay strongly modulated by the recurrences of the wave packet launched by the pump pulse in the electronically excited state of NaI. (Adapted from Zewail, 2000.)

and thereby the vibrational frequencies. The dynamics on originally purely repulsive PESs can even exhibit wave-packet recurrences due to reflections from adjacent particles. Moreover, energy is dissipated into the environment. Wave-packet oscillations are thereby efficiently damped and the population is collected at the potential minimum on a timescale in the order of 10 picoseconds. Fluctuations and inhomogeneities of the environment cause a dephasing of the coherent superposition generated by the optical excitation even if no energy exchange takes place. The decay of coherence-induced signal oscillations is therefore much faster than in the case of isolated molecules. On the other hand, it is possible to generate, in a matrix, coherently excited phonons with a short pulse excitation of an embedded chromophore.

22.4.2
Dynamics of Polyatomic Molecules

Larger molecules have usually several or many degrees of freedom that are involved in the dynamics, and the manifold of contributing states is very dense. This has two major consequences. First, the couplings between the intramolecular degrees of freedom have a strong impact. They cause energy redistribution between vibrational modes and often efficient radiationless electronic transitions. Secondly, the wave functions have a complex structure and observables are averages over population distributions as long as the temperature is not extremely low. Wave-packet signatures are therefore typically much less pronounced as in the case of small molecules. Quite often the dynamics is then described as a single or a series of transformations between states or configurations and the speed of the transformations is characterized by means of rates.

22.4.2.1 Vibrational Energy Redistribution and Cooling

Anharmonicities of the PES cause couplings between the vibrational modes of a molecule and thereby intramolecular vibrational energy redistribution (IVR). Energy transfer was measured for a number of molecules by IR pump and Raman probe spectroscopy (Laubereau and Kaiser, 1978; Hartl and Zinth, 2000). Transfer times in the range of 1 picosecond to almost 10 picoseconds were observed. This indicates that after several picoseconds, the energy is distributed over all vibrational modes and the molecule can be regarded as its own heat bath with an internal temperature.

Large molecules are often investigated in condensed phase. The interaction with the environment then causes an energy flow from the excited molecule to the environment and cooling of the molecule. This process was characterized by Raman measurements as well as via the shape of transient absorption spectra (Elsaesser and Kaiser, 1991; Schwarzer *et al.*, 2002). The intramolecular temperature decays exponentially to the temperature of the environment and the corresponding cooling times vary with the solvent and solute from slightly less than 10 picoseconds up to 40 picoseconds.

22.4.2.2 Solvation

The change of electronic states is usually accompanied by a change of the charge distribution in the molecule and of the static dipole moment. The environment reacts by adjusting its polarization so that the free enthalpy is again minimized. In solution, two contributions to the solvent polarization have to be considered. One is the electronic polarizability, which reacts instantaneous on changes of the charge distribution of the solute and is characterized by the index of refraction n. The second contribution results from the orientation and arrangement of the solvent molecules in the solvation shell around the solute. The energies associated with the two contributions $E_{\text{solv, elec}}$ and $E_{\text{solv, nuc}}$ are given by Eqs (22.20) and (22.21).

$$E_{\text{solv, elec}} = \frac{\mu^2}{4\pi\ \varepsilon_0\ a^3} \cdot \frac{n^2-1}{2n^2+1} \tag{22.20}$$

$$E_{\text{solv, nuc}} = \frac{\mu^2}{4\pi\ \varepsilon_0\ a^3} \cdot \left(\frac{\varepsilon-1}{2\varepsilon+1} - \frac{n^2-1}{2n^2+1}\right) \tag{22.21}$$

μ is the static dipole moment of the solute, a is its Onsager radius, ε is the static dielectric constant of the solvent, and ε_0 is the dielectric constant of vacuum. To adjust the second contribution, the nuclei of the solvent have to move. This includes vibrational motions, librations, and changes in the arrangement of the solvent molecules in the solvation shell. Accordingly, this response occurs on several timescales. Its evolution can be monitored by measuring the time-dependent shift of the emission spectrum after optical excitation of a chromophore that exhibits a reasonable strong change of the electric dipole moment like, for example, coumarin 153 (Horng *et al.*, 1995). The solvation dynamics is characterized by multiple exponential components with time constants in the range of a few hundred femtoseconds to several ten picoseconds. Smaller solvent molecules react faster and solvents, which form networks like the hydrogen-bonded networks of alcohols, exhibit slow response components associated with the rearrangement of the network. For certain solvents and a sufficient time resolution provided, the response can also exhibit oscillatory contributions caused by coherently excited motions of the solvent molecules (Pérez-Lustres *et al.*, 2005). It was also shown that the time dependence of the solvent response can be reconstructed from the frequency-dependent dielectric constant.

22.4.2.3 Electronic Dynamics: Internal Conversion, Conical Intersections, and Photochemistry

Most molecules exhibit excited-state lifetimes that are much shorter than the radiative lifetimes. Owing to the many nuclear degrees of freedom, nonadiabatic processes are likely to occur resulting in efficient IC and in photochemical products.

A key issue to understand molecular dynamics and photochemical processes is to locate conical intersections and to identify the involved coordinates. In an interplay of experimental (Fuß *et al.*, 1998) and theoretical work (Bernardi, Olivucci and Robb, 1996), several structural motifs were identified, which facilitate conical intersections. One important motif is, for example, a kink between a double and an adjacent single bond. It can lead to an energetic degeneracy between the electronic ground state and the first doubly excited singlet state. The molecular backbone is strongly twisted at this geometry and after the passage through the intersection, the twist motion can continue to a complete cis–trans isomerization. The rotation around the double bond is accompanied by a twist motion around the adjacent single bond. This can result in a Hula-twist isomerization where the two moieties on both sides of the double bond perform a bending motion instead of a rotation by 180° relative to each other. This allows for cis–trans isomerizations even in environments with low flexibility like proteins or organic matrices.

22.4.2.4 Coherent Wave-Packet Dynamics

Ultrashort excitation pulses generate, also in larger molecules, a coherent superposition of vibronic states. The wave packet launched by the pulse oscillates in those vibrational modes that contribute to the superposition. In time-resolved experiments with sufficient resolution, this results in oscillatory shifts of the bands contributing to the transient spectra and oscillations in kinetic traces measured at fixed probe wavelengths (see Figure 22.12). The frequencies of the oscillations are the frequencies of the contributing vibrations. Compared to small molecules, usually more

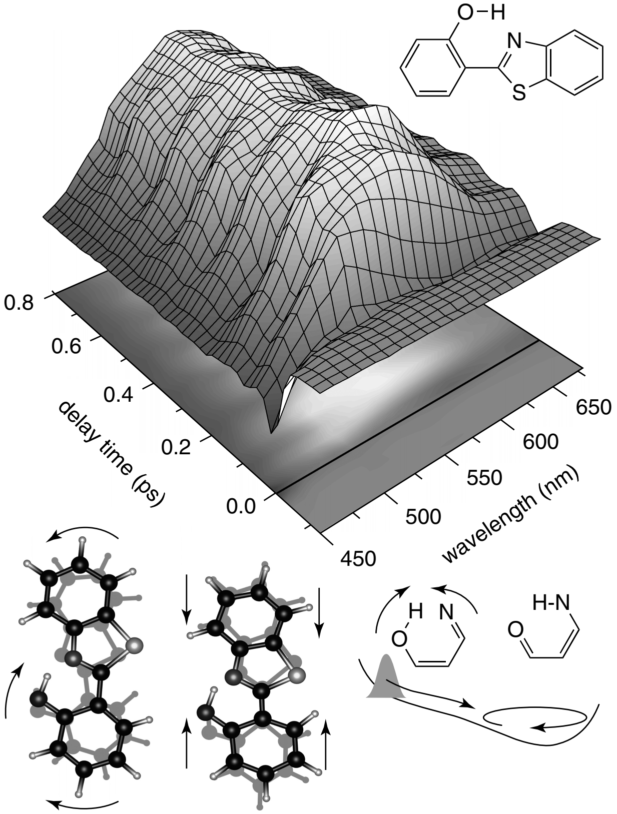

Fig. 22.12 Time-resolved transmission change after optical excitation of 2-(2′-hydroxyphenyl)benzothiazole. The excitation induces an intramolecular proton transfer, which proceeds as a ballistic wave-packet motion and excites coherently the depicted vibrational modes. (Lochbrunner, Wurzer and Riedle, 2003.)

vibrational modes and shorter damping times are observed due to the larger number of degrees of freedom.

A superposition of vibrational levels can be directly generated by the optical excitation. This is possible for Raman active modes and the superposition can then be reconstructed from the resonance Raman cross sections and the pulse spectrum (Takeuchi and Tahara, 2000). The optical excitation launches a wave packet not only in the electronically excited state but also very often in the electronic ground state via stimulated impulsive Raman scattering (SIRS). During the excitation process, population is pumped into the electronically excited state. The slope of the excited-state PES accelerates the corresponding wave packet along the Raman active modes. The ongoing interaction with the pulse can stimulate a part of the wave packet back into the electronic ground state where it oscillates in the Raman modes according to the gained momentum.

A coherent excitation of a vibrational mode can also result from the dynamics of a wave packet when the Franck–Condon region is already left and is not always caused by a direct optical excitation. For example, in azulene and in [2,2′-bipyridyl]-3,3′-diol, the coherent excitation of vibrational modes is observed that are optically inactive due to their symmetry (Wurzer *et al.*, 1999; Stock *et al.*, 2008). Accordingly, the analysis of oscillations reveals which coordinates contribute also beyond the Franck–Condon region to the path of the wave packet – information that can hardly be obtained by other methods.

The analysis of the coherent wave-packet motion is particularly valuable to reveal the mechanisms of photo reactions. Two examples are given here to demonstrate this approach. In the case of 2-(2′-hydroxyphenyl)benzothiazole, an excited-state intramolecular proton transfer takes place after optical excitation of the S_1-state. Femtosecond absorption experiments find that the proton transfer results in a coherent excitation of vibrational modes, which modulate the distance between the hydrogen donor and acceptor atom (Figure 22.12) (Lochbrunner, Wurzer and Riedle, 2003). This shows that prior to the transfer the donor–acceptor distance is shortened by a deformation of the molecular backbone until the hydrogen atom can be bound to the acceptor atom without the need for extra energy. Thereby the whole process can proceed along a barrierless and adiabatic pathway. The general notion is that if a process proceeds as a wave-packet motion, momentum is gained along coordinates that contribute strongly to the reaction path. After the process, the momentum results in oscillations in the associated normal modes of the product. This ringing of the product can be analyzed by time-resolved experiments and reveals which coordinates contribute to the reaction path.

In the case of the photodissociation of the metal hexacarbonyl compound $Cr(CO)_6$, oscillatory signal contributions with a low frequency of 95 cm^{-1} and with a higher frequency of about 400 cm^{-1} are observed in time-resolved, mass-selective dissociation experiments (Trushin *et al.*, 2008). The oscillatory component with the higher frequency is observed from the beginning of the dissociation process and results from a dominant optical excitation of the totally symmetric Cr-C stretch vibration. However, the low-frequency component is only observed in measurements that monitor the return of the wave packet at the electronic ground state. The frequency points to the coherent excitation of a totally symmetric C-Cr-C bending vibration. The results indicate that the electronic relaxation proceeds through a conical intersection between the S_1 and the S_0 state, which has a trigonal–bipyramidal geometry since after such an intersection the PES must have a slope along the C-Cr-C bending coordinate, resulting in an acceleration of the wave packet along this coordinate. This example demonstrates that the analysis of coherent wave-packet oscillations can give valuable information on the location of conical intersections.

22.4.2.5 Photoinduced Electron Transfer

The fundamental process of electron transfer can take place on timescales ranging from very slow for media in a glassy state to ultrafast for photoinduced reactions (Barbara, Meyer and Ratner, 1996). The charge transfer between a donor (D) and an acceptor (A) is strongly coupled to its environments. This necessitates a reorganization of intra- or intermolecular vibrational coordinates as well as of the polarization state of the solvent. Both effects can be described in a model where the potential or free energy is plotted along a reaction coordinate resulting, for example, in a double minimum potential, where D and A configurations are separated by a barrier. Possible transfer mechanisms encompass the limits of thermal activation and nuclear tunneling. Ultrafast photoinduced electron transfer occurs on timescales where the nonequilibrium effects with respect to the surroundings are important and the concepts underlying, for example, transition state theory break down. This might relate to the solvent which typically involves two timescales, that is, femtosecond inertial and picosecond diffusional motion. Measured intramolecular electron transfer times in $\mathrm{Ru}^{(\mathrm{II})}-\mathrm{Ru}^{(\mathrm{III})}$ mixed valence compounds in water yielded 80 femtoseconds, which is comparable to the inertial motion, but much faster than the diffusional one (Barbara, Meyer and Ratner, 1996). Furthermore, oscillatory components in the pump-probe signal have been demonstrated. They are a consequence of a coherent vibrational wave-packet motion, which is impulsively excited upon the Franck–Condon transition $\mathrm{D}-\mathrm{A} \rightarrow \mathrm{D}^{+}-\mathrm{A}^{-}$ (see also simulation in Wang and Thoss, 2008). Other examples are intermolecular D–A complexes in solution which have been studied, for example, by fluorescence up-conversion (Rubtsov and Yoshihara, 1999). Figure 22.13 shows data giving clear evidence for wave-packet motion, which has been assigned to an out-of-plane intramolecular vibration of the acceptor.

Owing to its technological importance, ultrafast heterogeneous electron transfer has attracted considerable attention. In particular, electron injection from a photoexcited adsorbed chromophore into a semiconductor is studied as a model for the primary process occurring in dye-sensitized electrochemical solar cells (Frischkorn *et al.*, 2007; Prezdho, Duncant and Prezhdo, 2008). In the case of perylene as dye and TiO_2 as semiconductor, injection times as short as 10 femtoseconds were observed. This fast dynamics results from the fact that the electron is transferred to a continuum of acceptor states in the conduction band of TiO_2. The insertion of bridge-anchor groups between the surface and the chromophore results in a slow down of the transfer rate, which is very sensitive to

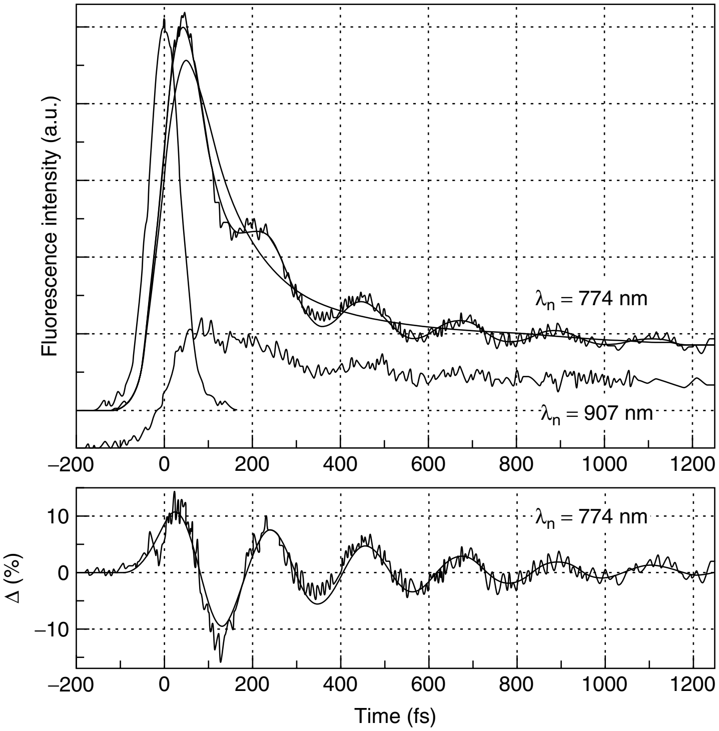

Fig. 22.13 Fluorescence decay of the hexamethylbenzene-tetracyanoethylene D–A complex in CCl_4 excited in the blue and red side of the time-integrated fluorescence spectrum. The oscillatory part ($\nu_{osc} = 155\ cm^{-1}$) is shown separately in the lower panel. The upper panel also contains the instrument response function (85 femtoseconds FWHM) as well as result of a biexponential plus oscillatory function fit. (Reprinted from (Rubtsov and Yoshihara, 1999), with permission from the ACS.)

the group type. If the group consists of conjugated bonds, it acts like a wire and only a moderate decrease of the rate is observed. If saturated bonds interrupt the transfer pathway, they induce a tunneling barrier and suppress the transfer strongly.

22.4.3
Clusters and Liquids

22.4.3.1 **Atomic Clusters**

Clusters are interesting objects since they allow to study how material properties change by going from the isolated atom or molecule to the condensed phase. They are generated in supersonic expansions of atomic or molecular gases. Typically, an inert gas is used as carrier to achieve low temperatures. Depending on the parameters of the expansion, clusters of various sizes between a few atoms or molecules up to many thousands of them can be generated (Hertel and Radloff, 2006). To measure the dynamics and evolution of electronically excited clusters, mass-selective ion and photoelectron spectroscopy are used in combination with the pump-probe technique (Stolow, Bragg and Neumark, 2004).

Atomic metal cluster anions are investigated to understand the transition from a single atom to a metal. They show very fast electronic relaxation processes even if the cluster contains only a few atoms (Dermota, Zhong and Castleman, 2004). The fast dynamics results from inelastic electron scattering, which is very similar as in the solid state although somewhat slower. The dynamics at stronger excitation intensities is governed by the finite size of the cluster. Multiphoton absorption results in multiple ionization and above a certain threshold in a Coulomb explosion of the cluster driven by the positive excess charges (Dermota, Zhong and Castleman, 2004). In the cluster a plasmon can be excited, that is, a collective oscillation of electrons. Its frequency depends on the density of quasi-free electrons and shifts to lower frequencies as the cluster expands during the Coulomb explosion (Fennel *et al.*, 2007). At some time, the plasmon frequency becomes resonant with the optical frequency of the exciting laser pulse if the pulse duration is sufficiently long. In this case, dramatically enhanced absorption occurs heating the cluster further up and resulting in extremely fast nonthermal electrons due to various acceleration processes.

In a supersonic expansion, molecules can form clusters that are typically stabilized by intermolecular hydrogen and van-der-Waals bonds. Optical excitation and photoionization of such clusters result often in complex fragmentation patterns and time dependences, and the interpretation relays on a careful analysis of the experimental findings. Ammonia clusters are a typical example (Hertel and Radloff, 2006). Nevertheless, it was possible to show for the ammonia dimer $(NH_3)_2$ that, after excitation to the $\tilde{A}$ state but prior to dissociation, a hydrogen transfer in the electronically excited state can occur leading to the $NH_4{-}NH_2$ complex. Interestingly, fast hydrogen transfer can also occur in mixed clusters containing one solute and a few solvent molecules like NH_3 or H_2O. For example, transfer in the electronically excited state from the solute to the solvent was observed on the subpicosecond timescale for indole ammonia and methanol water clusters. Such clusters give interesting insights into solvation processes since one can regard them as nanodroplets. The following two examples indicate that, in many cases, quite moderate cluster sizes result already in efficient solvation. The spectral and dynamical signatures of a solvated electron in water are almost fully developed when the cluster exceeds a size of about only 25 water molecules (Hertel and Radloff, 2006). In the case of the photodissociation of $I_2{}^-$ with NIR photons, it was shown that one solvation shell of CO_2 or OCS results in complete caging and recombination of the $I_2{}^-$ molecule (Sanov and Lineberger, 2004). Time-resolved experiments reveal that in the case of a full solvation shell the iodine atoms are directly reflected by the shell, whereas in the case of a partial solvation shell the weaker forces lead to a slowdown of the dynamics. If the excitation is performed with near-UV photons a higher lying dissociative state of $I_2{}^-$ with the spin-orbit excited $I^- + I^*(^2P_{1/2})$ asymptote is populated. Then, quenching of the spin-orbit excitation is necessary for recombination. This is only possible if the environment strongly disturbs the electronic structure of the $I_2{}^-$ anion. Therefore, the recombination yield and the dynamics depend strongly on the structure of the solvation shell and do not vary monotonically with the cluster size. However, almost complete caging is observed for about two solvation shells.

22.4.3.2 Water and Hydrogen-Bonded Networks

Disordered hydrogen-bonded networks are showing structural fluctuations on a broad range of timescales and have been targeted by ultrafast infrared spectroscopy (Nibbering *et al.*, 2007). Owing to its relevance, for example, for aqueous solvation, water has attracted most attention, although there are some efforts devoted, for example, to alcohols such as solvated methanol-OD oligomers (Asbury *et al.*, 2003). Due to technical limitations (high optical density of OH stretching fundamental transition in H_2O), most studies focused on isotopically diluted samples, that is, OH in HOD/D_2O or OD in HOD/H_2O. Upon excitation of the OH stretching vibration, one observes an ultrafast vibrational energy relaxation (~0.7 picoseconds), which has been assigned to be due to a Fermi resonance coupling with the HOD bending overtone vibration. Since the efficiency of this relaxation pathway depends on the proper resonance, it is not only influenced by the composition of the system but also on the fluctuation of the transition frequencies. These fluctuations are described by the correlation function of the energy gap, for example, for the OH stretch fundamental $\langle\delta\omega_{\mathrm{OH}}(t)\delta\omega_{\mathrm{OH}}(0)\rangle$, which determines the line shape function and therefore the nonlinear response (cf Eq. (22.14)). Three-pulse photon echo peak shift measurements on HOD/D_2O revealed a correlation function, which shows signatures of an underdamped motion (170 femtoseconds) related to the hydrogen-bond and a 1.2-picosecond decay assigned to be due to structural reorganization of the network (Fecko *et al.*, 2003). The subpicosecond dynamics of neat liquid water itself has been studied only recently (Cowan *et al.*, 2005). Here, one finds that correlations decay on a timescale of about 50 femtoseconds. This ultrafast loss of memory of the initially excited configuration was ascribed to the anharmonic coupling to librational motions in the hydrogen-bonded network.

The dynamics of an excess proton in water is another important issue that has been addressed by ultrafast IR spectroscopy (Woutersen and Bakker, 2006). Essentially there are two limiting structures the Eigen ($H_9O_4^+$) and the Zundel ($H_5O_2^+$) cations, which are related to proton migration. They can be distinguished by the absorption of the terminal OH stretching vibrations. Exciting the Eigen-like structure and probing for the Zundel-like one, it was found that the interconversion between these structures is faster than 100 femtoseconds.

22.4.4
Biological Systems

At first glance, stepping from molecules in solution to chromophores in biosystems seems to provide merely a different type of environment. However, biological systems are designed to fulfill a certain task, which is often made possible only by the concerted interplay of highly specific interactions. Although the biologically relevant response time to some external stimulus is in the range of microseconds or longer, the individual steps involving, for example, barrier crossings, can be substantially faster. These elementary steps and their importance for the structure–function relationship are in the focus of ultrafast spectroscopy.

22.4.4.1 **Primary Steps of Photosynthesis**

The conversion of sunlight into a transmembrane potential, which drives biochemical reactions in higher plants and certain bacteria builds on the ultrafast primary steps of excitation energy and electron transfer in pigment–protein complexes. Here, a variety of structures realize the same highly efficient cascade of energy funneling and charge separation. Purple bacteria such as *Rhodobacter (Rb.) sphaeroides* and *Rhodopseudomonas (Rps.) acidophila* have been the first organisms for which high-resolution X-ray structures became available. Here, light harvesting is performed by two types of ring-shaped antenna complexes. The peripheral antenna (LH2) consists of two pools of Bacteriochlorophyll a molecules, which are fixed by the protein scaffold. The closely packed B850 ring (16–18 pigments with 9 Å distance) gives rise to absorption around 850 nm, while a second absorption band around 800 nm originates from the B800 ring of 8–9 pigments being well separated (21 Å). These two pigment pools show an ultrafast dynamics, which can be described in terms of Frenkel exciton theory (Renger, May and Kühn, 2001). It has been extensively studied by various spectroscopic methods (Grondelle and Novoderezhkin, 2006; Sundström, 2008) leading to the following picture (time constants for 77 K): Excitation energy transfer among the B800 molecules takes place on timescales from 0.4 to 2 picoseconds, the transfer to the B850 pool covers the range from 0.4 to 1.1 picoseconds, while the intra-B850 transfer shows timescales of 150 femtoseconds and faster. This fast dynamics localizing the excitation energy in the red part of the spectrum is only possible due to the coupling between the electronic excitations to vibrations of the pigment and the protein, which causes energy dissipation (Renger, May and Kühn, 2001). The core antenna (LH1) receives the excitation energy from LH2 within about 5 picoseconds. Its structure is similar to that of the B850 ring, but it hosts the reaction center that traps the excitation energy within 30–40 picoseconds. The mechanism of the subsequent charge separation in the reaction center involves the initially excited special pair (P) of two strongly coupled bacteriochlorophyll molecules, a monomeric bacteriochlorophyll (B), a bacteriopheophytin (H), and a quinone (Q). Much effort had been focused on the role of B in the electron transfer chain, but by now there is clear evidence that at ambient temperatures the state P^+B^- forms within 3 picoseconds, suggesting a stepwise transfer mechanism. This intermediate decays within 0.9 picoseconds to give P^+H^- and after further 200 picoseconds P^+Q^- (Zinth and Wachtveitl, 2005) (see Figure 22.14).

Ultrafast two-dimensional spectroscopy in the visible range has been applied to the Fenna–Matthews–Olson light harvesting complex of sulfur bacteria to unravel the coupling patterns between its seven pigments, which are not arranged in some regular order (Brixner *et al.*, 2005).

22.4.4.2 **Rhodopsins**

Rhodopsins are photoreceptors where the absorption of a photon leads to a

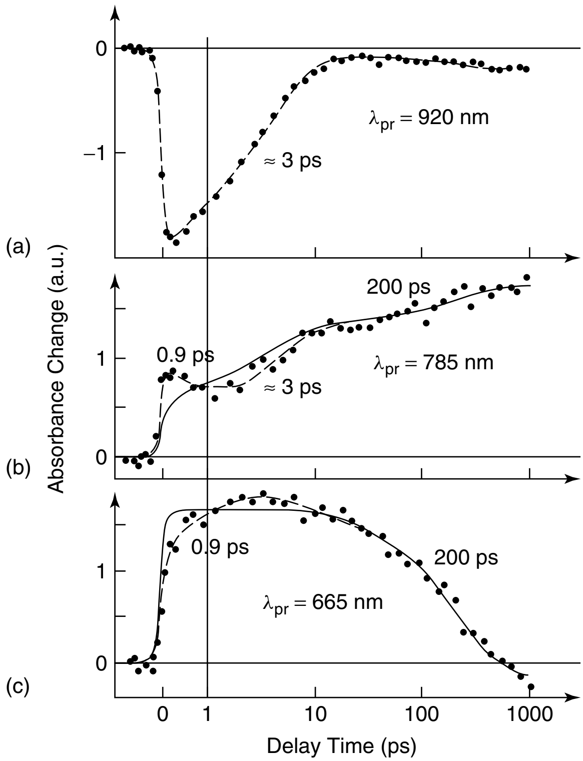

Fig. 22.14 Transient absorption of the reaction center of Rb. sphaeroides. At 920 nm, the signal is dominated by the decay of the primary donor (P*), at 785 and 665 nm the 0.9-picosecond component gives evidence for the P^+B^- to P^+H^- transfer. The 200-picosecond component corresponds to the electron transfer to the quinone. The solid and dashed lines represent fits with and without, respectively, the 0.9-picosecond component. Note that the axis is linear until 1 picosecond. (Reprinted from (Zinth and Wachtveitl, 2005), with permission from Wiley-VCH Verlag GmbH & Co. KgaA.)

conformational change, which as the result of a signaling cascade, eventually triggers some biological response. In rhodopsins, one has an 11-cis to all-trans conformational change in the retinal chromophore, which is key for the process of primary vision. In unicellular organisms, on the other hand, bacteriorhodopsin provides the means for ion transport across a membrane. The cis–trans isomerization of retinal has been one of the early targets of ultrafast spectroscopy (Sundström, 2008), see Figure 22.15. In fact, the formation of the primary photoproduct, bathorhodopsin, via an S_1-S_0 conical intersection occurs with high efficiency on a timescale of about 200 femtoseconds (Schoenlein *et al.*, 1991). The nature of the reaction coordinate has long been disputed. Here, valuable insight about the evolution on the excited S_1 state

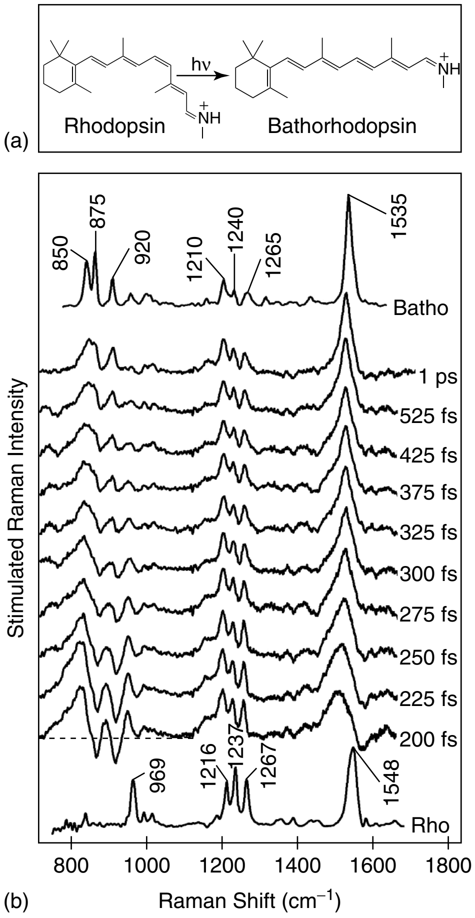

Fig. 22.15 (a) Photoinduced 11-cis to all-trans isomerization of the retinal chromophore of rhodopsin. (b) Characterization of this process after S_1-S_0 transition by means of time-resolved femtosecond-stimulated Raman spectra; lower and uppermost curve show resonance Raman spectra of ground-state rhodopsin and bathorhodopsin, respectively. (From (Kukura *et al.*, 2005), reprinted with permission from the AAAS.)

surface toward the conical intersection comes from on-the-fly adiabatic molecular dynamics simulations (Frutos *et al.*, 2007). It has been found that the retinal up to about 60 femtoseconds after excitation deforms to prepare for an asynchronous crankshaft type motion, which leads to the conical intersection after about 110 femtoseconds. The coherent evolution of the vibrations after the S_1-S_0 transition has been followed by femtosecond-stimulated Raman experiments as shown in Figure 22.15 (Kukura *et al.*, 2005). These data allow conclusions to be derived on the redistribution of the excess energy and the motions involved in forming the photoproduct. Dispersive features indicate changes in vibrational frequency along the reaction coordinate, which are particularly pronounced for the hydrogen out-of-plane motions observed around 800–900 cm^{-1}.

22.4.4.3 **Peptide Dynamics**

The function of peptides and proteins is largely determined by their three-dimensional structure. The latter is not static but fluctuates on rather different timescales, related to the intrinsic length scales. The fastest fluctuations in the picosecond and femtosecond range are not accessible to the well-established structure determination methods of nuclear magnetic resonance spectroscopy. However, they are likely to be of importance, for example, for energy transport or in the process of opening reaction channels leading to conformational changes. The most prominent signature in the IR spectrum of peptides is the amide I vibration (C=O stretch) of the peptide unit, which absorbs around 1620 cm^{-1}. Amide I modes on different peptide units are coupled leading to a delocalization of this vibration and therefore contain information on the secondary structure of the peptide and its fluctuations. This information, however, is congested in the linear IR spectrum. Two-dimensional (20) IR spectroscopy can be used to unravel this information by spreading the spectrum into a second dimension. For peptides, this was first demonstrated by Hochstrasser and coworker (Hamm, Lam and Hochstrasser, 1998). They used a dynamic hole-burning approach where a narrow band laser excites the sample and a broadband laser probes the spectral changes. If the dispersed signal is plotted as a function of the frequency of the pump laser, one obtains a 2D signal, which on the diagonal contains information on the amide I eigenstates, disentangling homogeneous (antidiagonal) and inhomogeneous (diagonal) contributions to the line width. Furthermore, in the off-diagonal part of the spectrum, cross peaks appear, which are related to the coupling of the considered transitions. The localized nature of the individual amide I vibrations and the sensitivity of the coupling strengths on the backbone structure allow for a simulation of these spectra in terms of vibrational Frenkel excitons (Mukamel and Abramavicius, 2004). The above-mentioned paper triggered host of investigations of the equilibrium conformational dynamics of various peptides in solution using different 2D IR spectroscopies (Woutersen and Hamm, 2002). More recently, the focus has been extended to membrane bound peptides (Mukherjee *et al.*, 2006).

One might ask whether this scheme can be applied to proteins containing hundreds of local amide I modes. It turns out that the 2D IR spectrum is rather sensitive to the structural motif as illustrated in Figure 22.16 (Ganim *et al.*, 2008). This gives direct access to the observation of large conformational changes, that is, protein folding and unfolding. In order to trigger these changes, the temperature-jump method has been used (Ganim *et al.*, 2008). An ultrafast alternative is the use of photoswitches built into the peptide structure. Photoexcitation of azobenzene, for instance, leads to a reversible cis–trans isomerization, which is capable of triggering a conformational change, for example, of small cyclic peptides (Hamm, Helbing and Bredenbeck, 2008).

22.4.4.4 **DNA**

The UV photochemistry and photophysics of nucleic acid bases has been studied extensively in gas and solution phase with the prospect that it contains the key for the understanding of DNA photoprotection and UV-damage (Crespo-Hernández *et al.*, 2004). Singlet excitation of DNA bases gives rise to a rapid radiationless deactivation,

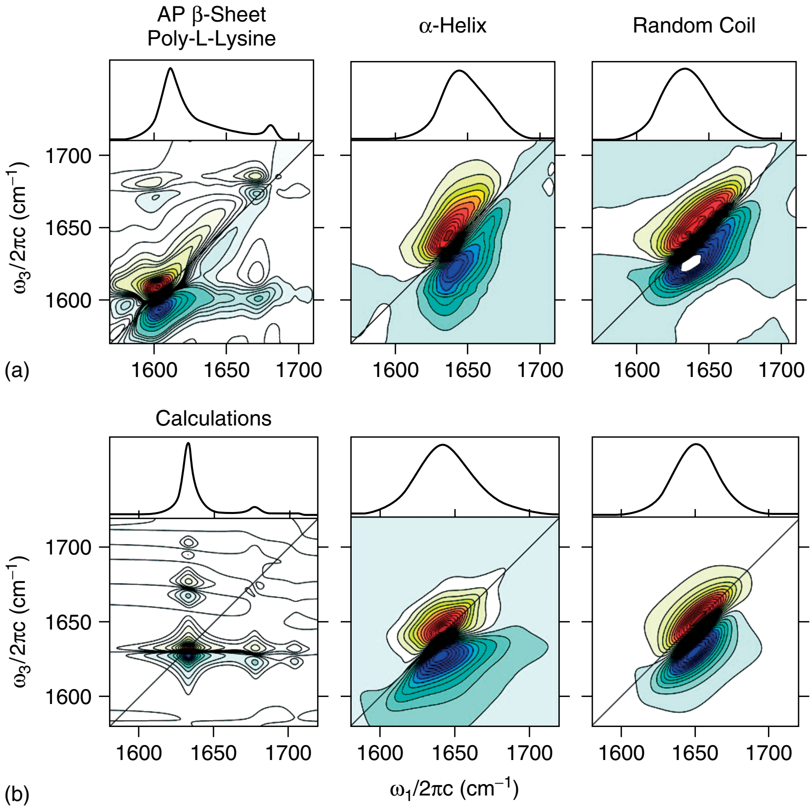

Fig. 22.16 Sensitivity of the 2D IR spectrum in the amide I region with respect to the secondary structure of poly(L-lysine) (panel (a) measured spectra, (b) model calculation). The upper/lower panels show the 2D IR spectra. The antiparallel (AP) β-sheet has two diagonal peaks as well as off-diagonal ones due to coupling pattern of amide I vibrations. These peaks merge into a single one for the α-helix and random coil, which can only be distinguished in the 2D IR spectrum. Notice that along ω_1 one has overlapping induced absorption and stimulated emission leading to an asymmetry with respect to the diagonal. (Reprinted from (Ganim *et al.*, 2008), with permission from the ACS.)

which necessitates the use of femtosecond methods for unraveling the coupled electron-vibrational dynamics. Quantum chemical calculations predict the relevant singlet manifold of the bases to be comprised of $\pi\pi^*$ and $n\pi^*$ excitations, where the former are carrying most of the oscillator strength. Time-resolved photoelectron spectroscopy of a supersonic jet expansion of adenine has been utilized to identify a 50-femtosecond $\pi\pi^*$ to $n\pi^*$ decay followed by a 750-femtosecond internal conversion into the electronic ground state for excitation well above the $\pi\pi^*$ origin (Ullrich *et al.*, 2004).

Upon base pairing new relaxation pathways become possible. For isolated base pairs, evidence for a deactivation along a reaction coordinate involving the hydrogen-bonded proton motion has been found. This leads to the stabilization of a charge transfer state, which converts into the S_0 ground state (Schultz *et al.*, 2004). Upon base stacking, the situation changes and the formation of an excimer has been observed. For double strands of adenine–thymine oligomers, the excimer formation is as fast as 400 femtoseconds and the excimer state decays within 50 picoseconds (Crespo-Hernández, Cohen and Kohler, 2005).

Base pairing does have a great impact on the vibrational dynamics in the electronic ground state. The formation of hydrogen bonds leads to substantial anharmonicities of the PES. For adenine–uracil pairs in deuterochloroform solution, pump-probe IR spectroscopy revealed a decrease of the lifetime of the hydrogen-bonded NH-stretching vibration by a factor of 3 as compared to monomeric uracil (Woutersen and Cristalli, 2004). DNA oligomers are investigated in the presence of water. For the IR spectrum, this gives rise to a broad absorption around 3000–3600 cm^{-1}, masking the fundamental transitions of the hydrogen-bonded NH and NH_2 stretching modes. Using two-color IR pump-probe spectroscopy, the symmetric NH_2 stretching band of adenine–thymine oligomers could be assigned by virtue of its anharmonic coupling to the NH_2 bending and carbonyl stretching modes in the fingerprint range (Heyne, Krishnan and Kühn, 2008). The anharmonic couplings related to base pair stacking and secondary structure of guanine–cytosine oligomers have been studied with the help of two-dimensional IR spectroscopy in the carbonyl stretching region (amide I modes) (Krummel and Zanni, 2006).

22.4.5
Solid State and Surfaces

22.4.5.1 Semiconductors

Ultrafast spectroscopy is applied to semiconductors and semiconductor nanostructures like single and multiple quantum wells to study many-body dynamics (Chemla and Shah, 2001; Axt and Kuhn, 2004). Pump probe and, in particular, FWM experiments are powerful methods to investigate coherences, correlations, and quantum kinetics in these systems. Deviations from the uncoupled quasi-particle picture are found to be quite strong and for many aspects the theory has to be extended beyond the mean-field level. The following examples illustrate these findings.

Early experiments on GaAs found already surprisingly efficient exciton–exciton scattering at mean exciton separations of more than 100 nm (Schultheis, Kuhl and Honold, 1986). Exciton scattering by free carriers was found to be even 10 times stronger. This demonstrates that in dense matter Coulomb forces lead even for formally neutral quasi-particles like excitons to strong interaction with other quasi-particles. In this respect, it is interesting to note that the formation of quasi-particles is not instantaneous. Femtosecond experiments on GaAs with terahertz-probe pulses found that Coulomb screening of free carriers takes several tens of femtoseconds (Huber *et al.*, 2001). This indicates that in the case of fast collisions screening is less efficient as for slow dynamics. The strong Coulomb interaction between the charge carriers leads, in general, to an ultrafast relaxation and equilibration

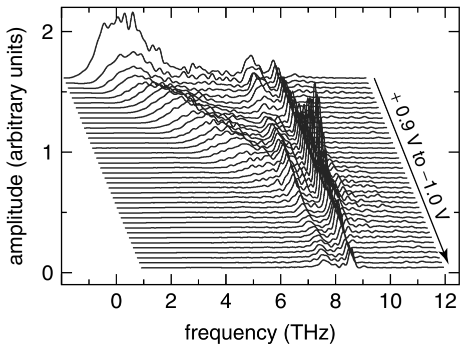

Fig. 22.17 Fourier transformations of time-resolved reflectivity measurements on a III-V semiconductor superlattice. The measurements were performed with different bias voltages as indicated at the right. (Reprinted from (Först *et al.*, 2007), with permission from Wiley-VCH Verlag GmbH & Co. KgaA.)

(Först *et al.*, 2007). However, for efficient recombination of electrons and holes, disorder and traps are necessary.

The analysis of quantum beats in time traces reveals information about carrier-phonon coupling. Time-resolved reflectivity experiments on III–V semiconductor superlattices showed, for example, that the coupling between optical phonons and Bloch oscillations can be efficiently tuned by an applied bias voltage (Först *et al.*, 2007) (Figure 22.17). The bias voltage changes the Bloch frequency linearly. If it is in resonance with an optical phonon, a strong increase of the phonon signal is observed. The largest signal is found for frequencies slightly below the frequency of the uncoupled phonon at 8.8 THz. This was interpreted as an indication that the dephasing rate of the Bloch oscillations increases by resonantly enhanced electron–phonon scattering. FWM experiments on GaAs quantum wells found clear evidence for two processes contributing to the dephasing of excitons (Carter, Chen and Cundiff, 2007): migration of the excitons and the interaction with acoustic phonons. The second contribution is identified on the basis of oscillatory features in time-dependent peak shifts of photon echo signals, which have a frequency in the expected range.

Ultrafast spectroscopy has also revealed strong evidence for biexcitons giving insight in Coulomb-induced correlations between electron hole pairs (Chemla and Shah, 2001; Axt and Kuhn, 2004). In several experiments, quantum beats with a frequency corresponding to the binding energy of biexcitons have been observed. Biexcitonic signatures might also be identified by means of their Stark shift, which is a red shift, if the excitation is just below the biexciton resonance contrary to mean-field contributions, which are blue-shifted (Axt and Kuhn, 2004).

22.4.5.2 Polymers and Organic Materials

Polymers and organic materials are very interesting for photonic applications like solar cells and organic light-emitting diodes (OLEDs). Ultrafast spectroscopy has contributed substantially to understand the exciton dynamics and the photoinduced charge generation in these materials (Lanzani *et al.*, 2005). The degree of disorder is much higher than in the case of semiconductors and the varying electronic structure of different molecular moieties has a strong impact on the dynamics (Scheblykin *et al.*, 2007). Electron hole pairs created by an optical excitation form Frenkel excitons, which are more or less localized on molecular chromophores and migrate by hopping processes through the material. Femtosecond absorption measurements show that after optical excitation, energy transfer on two timescales takes place. Within the first picoseconds, a red shift of the emission is observed indicating a relaxation toward the lowest state in the spatial vicinity of the excited chromophore. Thereafter, energy transfer occurs via thermally activated hopping. The transfer mechanism is of Förster type in both cases (Scheblykin *et al.*, 2007). Exciton migration becomes more efficient if the homogeneity of the site energies is improved. This was, for example, demonstrated for a polymer matrix highly doped with dye molecules, which interact only weakly with their environment but which can facilitate exciton migration by Förster energy transfer (Schlosser and Lochbrunner, 2006). Ultrashort pump pulses lead typically to exciton densities, which are high enough that bimolecular processes and exciton–exciton annihilation become important. If two excitons approach each other until they occupy two adjacent sites, they can interact and one of them is then usually annihilated (Scheblykin *et al.*, 2007). The associated time dependence of the exciton population can be analyzed to characterize the exciton mobility.

Photogeneration of charges in polymers is intensively investigated by ultrafast spectroscopy. The application of an external electric field increases the probability for the appearance of free charges (Kersting *et al.*, 1994). This allows to identify of the features and signatures of the generation process. The results show that primarily excitons are optically generated, which can then dissociate into free charges. The efficiency depends critically on the binding energy of the Frenkel excitons. In polymers with a significant binding energy charge pair generation occurs only if dissociation sites with an appropriate electronic structure are available, a sufficiently strong external field is applied, or the excitation brings a large amount of excess energy into the system. In the latter case, charge generation is restricted to a time window in the order of 100 femtoseconds limited by vibrational energy redistribution (Scheblykin *et al.*, 2007). For organic solar cells, the charge generation due to the interaction with electron acceptors like fullerenes is of particular importance. Experiments applying photoconductivity as probe process showed that the charge generation depends sensitively on the electron donating polymer (Müller *et al.*, 2005). In the case of methyl substituted poly(*p*-phenylene), coulombically bound polaron pairs are formed at the polymer fullerene interface. They show a high recombination probability resulting in a low photocurrent yield. In contrast, with poly(2-methoxy,5-(3,7-dimethyloctyloxy)-1,4-phenylenevinylene as donor free carriers are generated within 500 femtoseconds leading to a high yield.

J-aggregates of dye molecules show photoinduced dynamics, again determined by Frenkel excitons. Time-resolved absorption experiments found a clear signature for the transition from the one-exciton to a two-exciton state (van Burgel, Wiersma and Duppen, 1995). They also showed that efficient exciton–exciton annihilation can occur due to the mobility of the excitons. Calculations indicate that transient spectra should also exhibit features due to coherences between several exciton states (Heijs, Dijkstra and Knoester., 2007).

22.4.5.3 Surfaces

In the field of surface science, ultrafast spectroscopy plays an increasing role in order to understand the relevance of electronic surface states, the interaction with adsorbates, and the evolution of catalytic processes (Güdde, Berthold and Höfer, 2006). For most pump-probe techniques, the bulk contribution would lead to an overwhelming background signal. However, meanwhile several techniques have been developed, which are particularly sensitive to the surface. In the case of two-photon photoemission (2PPE), the kinetic energy of photoelectrons generated by the probe pulse is measured. Only electrons from the surface have a significant kinetic energy since photoelectrons from the bulk are subject to inelastic scattering before they reach the surface (Güdde, Berthold and Höfer, 2006). Other probe techniques rely on nonlinear second-order mixing processes like second harmonic generation (SHG) and sum frequency generation (SFG) (Bonn, Kleyn and Kroes, 2002). These processes are only possible if the medium exhibits no inversion symmetry. While for most materials this condition is not fulfilled in the bulk, it is an intrinsic property of surfaces and interfaces. Then the signal results only from the surface or the interface.

Image potential states are typical for metal surfaces. An electron is bound to the surface by attractive Coulomb forces resulting from the image charge in the metal. For well-ordered crystals, the electron can freely move parallel to the surface. However, the wavefunction extends into the bulk and couples strongly to the bulk states. This leads to ultrashort lifetimes in the range of 10–60 femtoseconds (Güdde, Berthold and Höfer, 2006) (Figure 22.18). If rare-gas adlayers are adsorbed on surfaces of metal crystals like copper or ruthenium, the lifetime increases strongly. The rare-gas layer can influence the lifetime in two ways. It pushes the wavefunction of the image potential state out off the bulk reducing the coupling to the conduction band (Figure 22.18). In addition, the state can be shifted in energy resulting in a larger energetic separation from the relevant bands of the bulk.

The chemistry on surfaces is characterized by a sequence of steps including adsorption of reactants, frequently dissociation, rearrangement, bonding, and desorption of products. Ultrafast spectroscopy provides new insight into the mechanisms of these processes (Bonn, Kleyn and Kroes, 2001). In addition, the laser pulse excitation of surfaces with adsorbates can also lead to new reaction pathways. Pump pulses can influence the dynamics in several ways. They can promote reactions if they are directly absorbed by the reactants. They can also be absorbed by the substrate and energy is subsequently transferred to the reactants by hot electrons or phonons. Electron-mediated processes are only possible within a short time window of about ~1 picosecond due to the strong electron–phonon coupling that cools the electrons efficiently whereas cooling of the phonon bath occurs on a timescale of ~50 picoseconds. The relevance of electron-mediated processes was demonstrated for the recombination of hydrogen atoms adsorbed on an Ru(100) surface (Frischkorn *et al.*, 2007). After the absorption of ultrashort laser pulses, hot electrons from the substrate cause a strong excitation in coordinates associated with the motion of

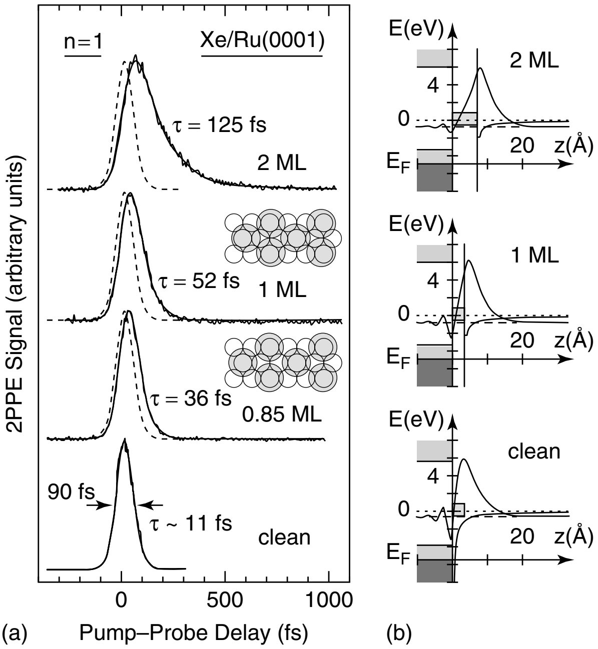

Fig. 22.18 (a) 2PPE lifetime (τ) measurements of the lowest image potential state of ruthenium(0001) for the clean surface (lowest panel and dashed curves) and with Xe coverages up to two monolayers (MLs). (b) Schematic energy diagrams and wavefunctions of the image potential state for no adlayer, one ML and two MLs of Xe. (Reprinted from (Güdde, Berthold and Höfer, 2006), with permission from the ACS.)

H-atoms parallel to the surface. This allows the formation of H–H complexes with some excess energy. Redistribution of the energy causes the excitation of vibrational motions perpendicular to the surface and finally desorption of H_2. Electron-mediated energy transfer was also found for the desorption of CO_2 from an Ru(0001) surface covered with CO and atomic oxygen (Bonn, Kleyn and Kroes, 2001).

22.4.6
Laser Control

Traditional chemistry controls reactions using a statistical approach, that is, changing equilibria via modifying external conditions. Ultrafast lasers open the possibility for controlling the wave-packet dynamics involved in elementary reaction steps. The field emerged in the 1980s triggered by the suggestion of two-photon pump-dump control of branching reactions (Tannor and Rice, 1985). Here, it is assumed that there are two exit channels on the electronic ground-state PES, $AB + C \leftarrow ABC \rightarrow A + BC$. Upon electronic excitation, the wave packet samples the available configurations on the excited-state potential and if dumped back onto the ground state, in the right moment, one can preferentially populate either of the two exit channels. A next breakthrough in terms of theoretical modeling has been the development of optimal control theory by Rabitz *et al.* (Peirce, Dahleh and Rabitz, 1988), where the problem of finding the proper field for populating a certain target is solved by means of functional optimization. Predicting laser fields, either by optimal control or by optimization of parameterized field forms and achieving control by tuning laser parameters in the laboratory, is known as open loop control. There are many examples for experimental realization of open loop control in atoms and diatomic molecules (Wollenhaupt, Engel and Baumert, 2005), but also in more complex systems such as the selective ground-state bondbreaking of diazomethane by chirped pulses (Windhorn *et al.*, 2003), the population of high vibrational states of the CO group in carboxyhemoglobin (Ventalon *et al.*, 2004), or the control of fragmentation products in the Coulomb explosion of silver clusters (Döppner *et al.*, 2005).

From the experimental point of view it would, however, be desirable to define the target and then let the molecule decide what kind of field is required. The appropriate approach is optimized feedback closed loop control and has been suggested by Rabitz *et al.* (Judson and Rabitz, 1992). Here, the laser pulse is iteratively optimized using a pulse shaper by incorporating the feedback from the experiment, that is, the reaction yield. For molecular reactions, the first successful demonstration has been by the group of Gerber in 1998 (see Figure 22.19) (Assion *et al.*, 1998).

Meanwhile, there are a number of applications (Brixner and Gerber, 2003) even to complex biological systems (Wohlleben *et al.*, 2005). From the theoretical point of view, this approach represents a serious challenge since even for simpler systems such as polyatomics no first-principle simulation is feasible. In certain cases, however, insight can be gained from appropriately chosen reduced model systems (Daniel *et al.*, 2003).

Using shaped laser pulses to control wave-packet dynamics is a rapidly developing field holding potential in different areas. Some examples include coherently

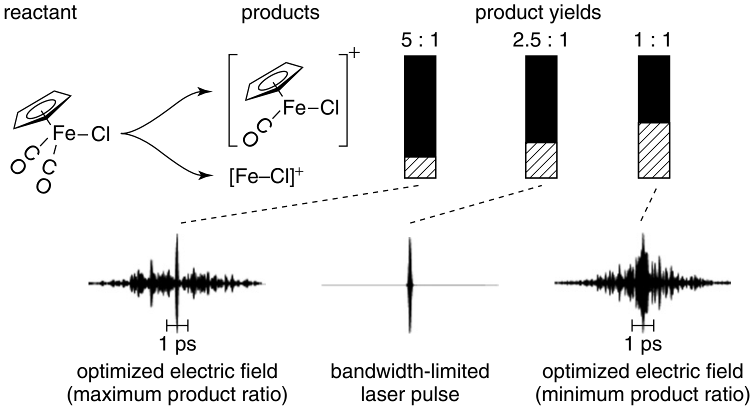

Fig. 22.19 Coherent feedback control of the $[CpFe(CO)Cl]^+/[FeCl]^+$ branching ratio. (Reprinted from (Brixner and Gerber, 2003), with permission from Wiley-VCH Verlag GmbH & Co. KgaA.)

controlled laser microscopy employing the coherent anti-Stokes Raman effect with a single pulse (Dudovich, Oron and Silberberg, 2002), the use of coherent control theory to perform logic operations with vibrational qubits in quantum information (de Vivie-Riedle and Troppmann, 2007), or even the manipulation of certain features in nonlinear optical spectra (Abramavicius and Mukamel, 2004).

22.4.7
New Frontiers

22.4.7.1 Generation of High Harmonics and Attosecond Pulses

If ultrashort high energy pulses interact with matter, coherent deep ultraviolet (XUV) and soft X-ray radiation is generated. In a typical experiment, 50–5-femtosecond short laser pulses with a center wavelength of 800 nm are focused into a gas jet of noble gas atoms or small molecules. After the jet XUV radiation, propagating collinear with the NIR-beam can be observed (Kapteyn *et al.*, 2007). A capillary filled with gas can be used instead of a jet. A schematic spectrum of the generated radiation is shown in Figure 22.20. It consists of many odd harmonics of the fundamental (Agostini and DiMauro, 2004; Pfeifer, Spielmann and Gerber, 2006). Under suitable conditions, harmonics with order numbers up to 100 can be observed and the shortest available wavelengths are only a few nanometers. In the low-frequency region, the intensity decreases with the order until the order number reaches a plateau region in which the intensity shows only a weak dependence on the order. At high frequencies when the photon energy is $\approx 3.2 \cdot U_P + I_P$, the plateau region ends and a “cutoff” region follows in which the intensity decays rapidly with the order number. I_P is the ionization potential of the medium and U_P the ponderomotive energy associated with the pulse field strength E_0 and the angular frequency ω of the fundamental via $U_P = e_0^2\, E_0^2/(4\, m_e\, \omega^2)$.

The appearance of this radiation is explained by the following three-step model (Corkum and Krausz, 2007). In the case of very intense laser pulses, the electric field strength, approaches for the valence electrons, the strength of the binding Coulomb potential. When the laser field adopts its maximal strength during an optical cycle, the tunneling probability becomes quite large and a significant fraction of the electrons is set free. Those electrons are subsequently accelerated by the laser field. Since the field reverses its direction in the next half cycle, the electrons are driven back to the atom or molecule. When they

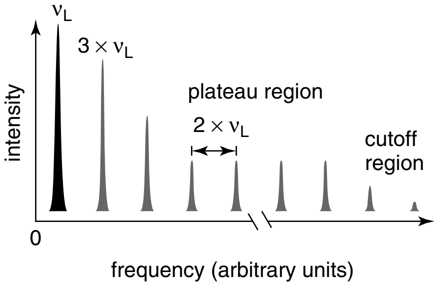

Fig. 22.20 Schematic spectrum of high harmonics generated by an ultrashort laser pulse with the center frequency ν_L. (Adapted from Agostini and DiMauro, 2004.)

return, they have a kinetic energy of about 50–1000 eV. Inelastic scattering and recombination of the electrons occur with high probability and transfer their kinetic energy into electromagnetic radiation. The recombination events themselves and the thereby emitted XUV pulses are extremely short, that is, in the order of 100 as. Since these events take place every half optical cycle, a train of attosecond pulses is generated and the spectrum consists of odd harmonics. In the maximum of the driving pulse, the recolliding electrons have the highest kinetic energy and lead to the contributions with the shortest wavelengths. In the case of a very short driving pulse, only a single attosecond pulse contributes to the highest frequencies. A single attosecond pulse can then be isolated from the train by selecting only the short wavelength contributions with an appropriate band pass filter. This approach led to isolated attosecond pulses with durations down to 130 as (Sansone *et al.*, 2006).

22.4.7.2 **Application of High Harmonics and Attosecond Pulses**

High harmonic generation is a promising technique for coherent radiation in the XUV and soft X-ray region. For example, in time-resolved photoelectron spectroscopy, such probe pulses provide access to high lying core electrons and their chemical shifts. Several techniques like quasi-phase matching and pulse shaping are applied to increase the high harmonic yield (Pfeifer, Spielmann and Gerber, 2006; Kapteyn *et al.*, 2007).

Attosecond time resolution is of particular interest for extremely fast electronic processes, like the decay of atomic core excitations. Such processes can be expected if no rearrangements or motions of nuclei are involved. Several experiments have been carried out using an attosecond pulse for excitation and the field of the associated NIR pulse for probing (Kling and Vrakking, 2008). The XUV attosecond pulse results in photoemission of valence as well as core-level electrons. The NIR field accelerates the photoelectrons and changes the photoelectron spectrum. This effect depends sensitively on the strength of the NIR field at the moment of the electron release and thereby on the phase of the NIR carrier wave at the time of the attosecond excitation. Accordingly, this "streaking" technique provides a time resolution comparable to the attosecond pulse length. Even more interesting, the ejection time of secondary photoelectrons, which result from relaxation processes like, Auger decays, can be precisely determined (Corkum and Krausz, 2007). In a similar way, tunnel ionization induced by the NIR field can be used as probe process with a time resolution much below the optical period.

Another approach uses the high harmonic spectrum to investigate the dynamics in small molecules during the generation process (Itatani *et al.*, 2004). After the tunnel ionization, the remaining ion is usually in a nonstationary configuration and its wavefunction evolves while the photoelectron first propagates away and then returns to the core. Since the whole process is coherent, the interference between the wavefunctions of the returning electron and the core has a strong impact on the emission of the XUV light (Kling and Vrakking, 2008). The characterization of the XUV spectrum, in dependence on the molecular orientation, can therefore provide a tomographic picture of the electronic wavefunction and its evolution.

22.4.7.3 Carrier Envelope Phase and Frequency Comb

In the case of very short laser pulses and nonlinear interactions, the carrier envelope phase (CEP), that is, the relative phase between the pulse envelope and the carrier wave becomes important. For a linearly polarized Gaussian pulse, a CEP of 0 indicates that the maximum of the envelope coincides with a field maximum of the carrier wave whereas a CEP of $\pm\pi/2$ indicates a zero crossing of the field at the maximum of the envelope (see Figure 22.21). The maximal field strength is therefore slightly higher in the former case than in the latter.

In a femtosecond oscillator, the phase delay and the group delay of a roundtrip are somewhat different and subsequent pulses exhibit different CEPs. In the ideal case of constant phase and group delay per roundtrip the shift of the CEP between the pulses is always the same. The spectrum of such a pulse train consists of equally spaced frequency components with a spacing given by the repetition rate of the oscillator (see Figure 22.21). Extrapolating this frequency comb toward lower frequencies results in a lowest component, which deviates from zero by the carrier envelope offset frequency (f_{CEO}) given by the CEP shift per pulse (Hänsch, 2006). f_{CEO} can be measured with a 1 f to 2 f interferometer (see Figure 22.21). This requires a pulse spectrum with reasonably strong components separated by an octave from each other. If the width of the pulse spectrum is insufficient, it can be broadened by self-phase modulation in a microstructured fiber. The low-frequency components are frequency-doubled with a nonlinear crystal and the SHG signal is overlaid with the high-frequency components on a fast photodetector. The interference between the frequency-doubled low-frequency components and the high-frequency components results in a beat signal with the carrier envelope offset frequency f_{CEO}. This radiofrequency as well as the pulse repetition rate f_{Rep} can be accurately measured and compared to high-precision standards like a cesium atomic clock. The frequency of a comb mode with the integer mode number m is strictly linked to these two frequencies by

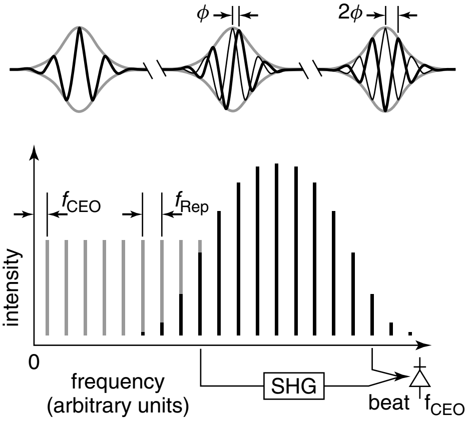

Fig. 22.21 A train of femtosecond pulses as it is generated by an oscillator. They differ only in their CEP ϕ. The lower panel shows the spectrum of the pulse train (black sticks). The gray sticks indicate an extrapolation of the observed frequencies toward zero. (Adapted from Hänsch, 2006.)

the relation $f_m = m \cdot f_{\text{Rep}} + f_{\text{CEO}}$. Accordingly, the frequency of each comb mode is known with an extremely high accuracy and can be used for high-precision spectroscopy. For this elegant approach, Theodor W. Hänsch was awarded the Nobel price in Physics in 2005 (Hänsch, 2006). The technique was, for example, applied to measure the 1S–2S transition frequency of the hydrogen atom with an accuracy of 1.4 parts in 10^{14} (Fischer *et al.*, 2004). Such experiments are highly relevant to test fundamental physical laws. The frequency comb allows also implementing atomic clocks based on optical transitions. They have the potential for a much higher accuracy than clocks operating at radiofrequencies and measurements with the same level of accuracy can be performed in much shorter times due to the high optical frequencies.

It is also possible to stabilize and control the f_{CEO} by this technique and thereby to generate pulses with deterministic CEP and electric field (Baltuška *et al.*, 2003). Adjusting the CEP to 0 or π results in the maximal field strength (see Figure 22.21), which is particularly useful in the case of nonlinear interactions of few cycle pulses. A prominent example is the generation of single attosecond pulses. In this case, the high-frequency components of the high harmonic spectrum are only generated during the central and strongest cycle of the electric field. Selecting these spectral components with a filter results in a single attosecond pulse (see above) (Sansone *et al.*, 2006).

22.4.7.4 **Time-Resolved X-Ray and Electron Diffraction**

One central vision in ultrafast spectroscopy is to follow, in real time, the structural and geometric changes during an ultrafast process. Ultimately, this should lead to a movie with femtosecond time resolution. Since structure and function are strongly related to each other, a much deeper understanding of the relevant mechanisms is expected for most processes.

To measure time-resolved structural changes, X-ray and electron diffraction are applied. X-ray diffraction itself is an extremely well established technique and the generation of short X-ray pulses has made large progress. Since the interaction of X-ray photons with matter is comparably weak and accordingly the diffraction efficiency low, a high X-ray flux is needed to observe a diffraction pattern. In several experiments, laser-driven plasmas have been used as X-ray sources (Bargheer *et al.*, 2006). An energetic laser pulse is focused onto a metal target like, for example, a copper tape. The high peak intensity in the order of 10^{17} W cm^{-2} results in a plasma with very energetic electrons. They penetrate into the metal and generate incoherent Bremsstrahlung and characteristic line emission, for example, K_α radiation. From this X-ray radiation, only the fraction that propagates toward the sample is used as probe beam. An optical pump pulse excites the sample and the X-ray diffraction pattern is recorded with a charge coupled device (CCD) camera in dependence on the delay time between the excitation and the X-ray pulse. Mostly, crystalline samples have been investigated so far since they provide the highest diffraction efficiencies. In this way, it was already possible to observe nonthermal melting, the dynamics of optical phonons, and the propagation of acoustic sound and shock waves. In the future, accelerator-based sources like the free electron laser will provide much higher X-ray fluxes (Pfeifer, Spielmann and Gerbr, 2006). This will make ultrafast X-ray diffraction applicable to a much wider class of samples and problems.

As an alternative to X-rays, electrons can be used as probe in ultrafast diffraction experiments (Zewail, 2006; Siwick *et al.*, 2003). If their kinetic energy is in the order of 30 keV, their de Broglie wavelength matches internuclear distances in the solid phase and results in diffraction patterns containing rich information. Since electrons interact very strongly with matter, their diffraction efficiency is several orders of magnitude higher than that of X-rays and quite low fluxes are needed to obtain a good signal-to-noise ratio. Owing to the Coulomb repulsion between the electrons, space charge effects are usually quite severe and electron pulses spread strongly in space and time when they consist of thousand or more electrons. This makes the experiments challenging and the available time resolution is about few hundred femtoseconds (Baum *et al.*, 2007). The electron pulses are generated by irradiating a photocathode with femtosecond laser pulses. The photoemitted electrons are accelerated by a high voltage to the appropriate kinetic energy. The sample is placed as near as possible to the exit of the acceleration section since short propagation distances minimize space charge effects. A femtosecond pump pulse initiates the dynamics in the sample and a CCD camera records the diffraction pattern behind it. In the case of dense matter the sample has to be extremely thin. Otherwise it cannot be penetrated by the electrons. If this is not possible, a grazing incidence geometry is used and the diffraction pattern is measured in reflection. Meanwhile, a number of pioneering experiments have been carried out and demonstrate that, for example, melting processes (Siwick *et al.*, 2003), phase transitions (Baum *et al.*, 2007), and structural changes during chemical reactions (Zewail, 2006) can be studied in detail.

Acknowledgments

This work has been financially supported by the Deutsche Forschungsgemeinschaft.

Glossary

Nonadiabatic Transition: The transition between two adiabatic electronic states triggered by the motion of the nuclei.

Pump-Probe Technique: A pump pulse initiates a process and a probe pulse measures the associated transient changes of observables.

Response Function: The function that describes the material response to the applied external fields.

Wave Packet: The superposition of quantum mechanical states that does not correspond to an eigenstate of the Hamiltonian.

References

Åberg, U., Åkesson, E., Alvarez, J.-L., Fedchenia, I., and Sundström, V. (1994) *Chem. Phys.*, **183**, 269–288.

Abramavicius, D. and Mukamel, S. (2004) *J. Chem. Phys.*, **120**, 8373–8378.

Agostini, P. and DiMauro, L.F. (2004) *Rep. Prog. Phys.*, **67**, 813–855.

Asbury, J.B., Steinel, T., Stromberg, C., Gaffney, K.J., Piletic, I.R., and Fayer, M.D. (2003) *J. Chem. Phys.*, **119**, 12981–12897.

Assion, A., Baumert, T., Bergt, M., Brixner, T., Kiefer, B., Seyfried, V., Strehle, M., and Gerber, G. (1998) *Science*, **282**, 919–922.

Axt, V.M. and Kuhn, T. (2004) *Rep. Prog. Phys.*, **67**, 433–512.

Backus, S., Durfee, C.D., Murnane, M.M., and Kapteyn, H.C. III (1998) *Rev. Sci. Instrum.*, **69**, 1207–1223.

Baltuska, A., Udem, Th., Uiberacker, M., Hentschel, M., Goulielmakis, E., Gohle, Ch., Holzwarth, R., Yakovlev, V.S., Scrinzi, A., Hänsch, T.W., and Krausz, F. (2003) *Nature*, **42**, 611–615.

Barbara, P.F., Walker, G.C., and Smith, T.P. (1992) *Science*, **256**, 975–981.

Barbara, P.F., Meyer, T.J., and Ratner, M.A. (1996) *J. Phys. Chem.*, **100**, 13148–13168.

Bargheer, M., Zhavorokov, N., Woerner, M., and Elsaesser, T. (2006) *ChemPhysChem*, **7**, 783–792.

Baum, P., Yang, D.-S., and Zewail, A.H. (2007) *Science*, **318**, 788–792.

Baumert, T., Grosser, M., Thalweiser, R., and Gerber, G. (1991) *Phys. Rev. Lett.*, **67**, 3753–3756.

Bernardi, F., Olivucci, M., and Robb, M.A. (1996) *Chem. Soc. Rev.*, **25**, 321–328.

Bonn, M., Kleyn, A.W., and Kroes, G.J. (2002) *Surf. Sci.*, **500**, 475–499.

Brixner, T. and Gerber, G. (2003) *ChemPhysChem*, **4**, 418–438.

Brixner, T., Stenger, J., Vaswani, H.M., Cho, M., Blankenship, R.E., and Fleming, G.R. (2005) *Nature*, **434**, 625–628.

van Burgel, M., Wiersma, D.A., and Duppen, K. (1995) *J. Chem. Phys.*, **102**, 20–33.

Carter, S.G., Chen, Z., and Cundiff, S.T. (2007) *Phys. Rev. B*, **76**, 121303–1–121303–4.

Cerullo, G. and De Silvestri, S. (2003) *Rev. Sci. Instrum.*, **74**, 1–18.

Chemla, D.S. and Shah, J. (2001) *Nature*, **411**, 549–557.

Corkum, P.B. and Krausz, F. (2007) *Nat. Phys.*, **3**, 381–387.

Cowan, M.L., Bruner, B.D., Huse, N., Dwyer, J.R., Chugh, B., Nibbering, E.T.J., Elasaesser, T., and Miller, R.J.D. (2005) *Nature*, **434**, 199–202.

Crespo-Hernández, C.E., Cohen, B., Hare, P.M., and Kohler, B. (2004) *Chem. Rev.*, **104**, 1977–2019.

Crespo-Hernández, C.E., Cohen, B., and Kohler, B. (2005) *Nature*, **436**, 1141–1144.

Daniel, C., Full, J., González, L., Lupulescu, C., Manz, J., Merlin, A., Vajda, S., and Wöste, L. (2003) *Science*, **299**, 536–639.

de Vivie-Riedle, R. and Troppmann, U. (2007) *Chem. Rev.*, **107**, 5082–5100.

Dermota, T.E., Zhong, Q., and Castleman, A.W. Jr. (2004) *Chem. Phys.*, **104**, 1861–1886.

Domcke, W. and Stock, G. (1997) *Adv. Chem. Phys.*, **100**, 1–169.

Domcke, W., Yarkony, D.R., and Köppel, H. (eds) (2004) *Conical Intersections*, World Scientific, New Jersey.

Döppner, T., Fennel, T., Diederich, T., Tiggesbäumker, J., and Meiwes-Broer, K.H. (2005) *Phys. Rev. Lett*, **94**, 013401.

Dudovich, N., Oron, D., and Silberberg, Y. (2002) *Nature*, **418**, 512–514.

Eisenthal, K.B. (1975) *Acc. Chem. Res.*, **8**, 118–124.

Elsaesser, T. and Kaiser, W. (1986) *Chem. Phys. Lett.*, **128**, 231–237.

Elsaesser, T. and Kaiser, W. (1991) *Annu. Rev. Phys. Chem.*, **42**, 83–107.

Fecko, C.J., Eaves, J.D., Loparo, J.J., Tokmakoff, A., and Geissler, P.L. (2003) *Science*, **301**, 1698–1702.

Feit, M.D., Fleck, J.A., and Steiger, A. (1982) *J. Comp. Phys.*, **47**, 412–433.

Fennel, Th., Döppner, T., Passig, J., Schaal, Ch., Tiggesbäumker, J., and Meiwes-Broer, K.-H. (2007) *Phys. Rev. Lett.*, **98**, 143401–1–143401–4.

Fischer, M., Kolachevsky, N., Zimmermann, M., Holzwarth, R., Udem, Th., Hänsch, T.W., Abgrall, M., Grünert, J., Maksimovic, I., Bize, S., Marion, H., Pereira Dos Santos, F., Lemonde, P., Santarelli, G., Laurent, P., Clairon, A., Salomon, C., Haas, M., Jentschura, U.D., and Keitel, C.H. (2004) *Phys. Rev. Lett*, **92**, 230802–1–230802–4.

Först, M., Nagel, M., Awad, M., Wächter, M., Dekorsy, T., and Kurz, H. (2007) *Phys. Stat. Sol. (b)*, **244**, 2971–2987.

Frischkorn, C., Wolf, M., Höfer, U., Güdde, J., Saalfrank, P., Nest, M., Klamroth, T., Willig, F., Ernstdorfer, R., Gundlach, L., May, V., Wang, L., Duncan, W.R., and Prezhdo, O.V. (2007) in *Analysis and Control of Ultrafast Photoinduced Reactions*, (eds O. Kühn and L. Wöste), Springer, Heidelberg, pp. 387–484.

Frutos, F.G., Andruniów, T., Santoro, F., Ferré, N., and Olivucci, M. (2007) *Proc. Natl. Acad. Sci.*, **104**, 7764–7769.

Fuß, W., Kompa, K.-L., Lochbrunner, S., Müller, A.M., Schikarski, T., Schmid, W.E., and Trushin, S.A. (1998) *Chem. Phys.*, **232**, 161–174.

Ganim, Z., Chung, H.S., Smith, A.W., Deflores, L.P., Jones, K.C., and Tokmakoff, A. (2008) *Acc. Chem. Res.*, **41**, 432–441.

Grondelle van, R. and Novoderezhkin, V.I. (2006) *Phys. Chem. Chem. Phys.*, **8**, 793–807.

Güdde, J., Berthold, W., and Höfer, U. (2006) *Chem. Phys.*, **106**, 4261–4280.

Gühr, M., Bargheer, M., Fushitani, M., Kiljunen, T., and Schwentner, N. (2007) *Phys. Chem. Chem. Phys.*, **9**, 779–801.

Hamm, P., Helbing, J., and Bredenbeck, J. (2008) *Annu. Rev. Phys. Chem.*, **59**, 291–317.

Hamm, P., Lim, M., and Hochstrasser, R.M. (1998) *J. Phys. Chem. B*, **102**, 6123–6138.

Hänsch, T.W. (2006) *ChemPhysChem*, **7**, 1170–1187.

Hartl, I. and Zinth, W. (2000) *J. Phys. Chem.*, **104**, 4218–4222.

Heijs, D.-J., Dijkstra, A.G., and Knoester, J. (2007) *Chem. Phys.*, **341**, 230–239.

Hertel, I.V. and Radloff, W. (2006) *Rep. Prog. Phys.*, **69**, 1897–2003.

Heyne, K., Krishnan, G.M., and Kühn, O. (2008) *J. Phys. Chem. B*, **112**, 7909–7915.

Horng, M.L., Gardecki, J.A., Papazyan, A., and Maroncelli, M. (1995) *J. Phys. Chem.*, **99**, 17311–17337.

Huber, R., Tauser, F., Brodschelm, A., Bichler, M., Abstreiter, G., and Leiternstorfer, A. (2001) *Nature*, **414**, 286–289.

Itatani, J., Levesque, J., Zeidler, D., Niikura, H., Pepin, H., Kieffer, J.C., Corkum, P.B., and Villeneuve, D.M. (2004) *Nature*, **432**, 867–871.

Jonas, D.M. (2003) *Annu. Rev. Phys. Chem.*, **54**, 425–463.

Jortner, J. and Berry, R.S. (1968) *J. Chem. Phys.*, **48**, 2757–2766.

Judson, R.S. and Rabitz, H. (1992) *Phys. Rev. Lett.*, **68**, 1500–1503.

Kapteyn, H., Cohen, O., Christov, I., and Murnane, M. (2007) *Science*, **317**, 775–778.

Kersting, R., Lemmer, U., Deussen, M., Bakker, H.J., Mahrt, R.F., Kurz, H., Arkhipov, V.I., Bässler, H., and Göbel, E.O. (1994) *Phys. Rev. Lett.*, **73**, 1440–1443.

Kling, M.F. and Vrakking, M.J.J. (2008) *Annu. Rev. Phys. Chem*, **59**, 463–492.

Kosloff, D. and Kosloff, R. (1982) *J. Comp. Phys.*, **52**, 35–53.

Kovalenko, S.A., Dobryakov, A.L., Ruthmann, J., and Ernsting, N.P. (1999) *Phys. Rev. A*, **59**, 2369–2384.

Krummel, A. and Zanni, M. (2006) *J. Phys. Chem. B*, **110**, 13991–14000.

Kukura, P., McCamant, D.W., and Mathies, R.A. (2007) *Annu. Rev. Phys. Chem.*, **58**, 461–488.

Kukura, P., McCamant, D.W., Yoon, S., Wandschneider, D.B., and Mathies, R.A. (2005) *Science*, **310**, 1006–1009.

Lanzani, G., Cerullo, G., Polli, D., Gambetta, A., Zavelani-Rossi, M., and Gadermaier, C. (2005) in *Physics of Organic Semiconductors* (ed. W. Brütting), Wiley-VCH, Weinheim.

Laubereau, A. and Kaiser, W. (1978) *Rev. Mod. Phys*, **50**, 607–665.

Lochbrunner, S., Wurzer, A.J., and Riedle, E. (2003) *J. Phys. Chem. A*, **107**, 10580–10590.

May, V. and Kühn, O. (2004) *Charge and Energy Transfer Dynamics in Molecular Systems*, Wiley-VCH Verlag GmbH, Weinheim.

Meyer, H.-D., Manthe, U., and Cederbaum, L. (1990) *Chem. Phys. Lett.*, **165**, 73–78.

Mukamel, S. and Abramavicius, D. (2004) *Chem. Rev.*, **104**, 2073–2098.

Mukherjee, P., Kass, I., Arkin, I., and Zanni, M.T. (2006) *Proc. Natl. Acad. Sci.*, **103**, 3528–3533.

Müller, J.G., Lupton, J.M., Feldmann, J., Lemmer, U., Scharber, M.C., Sariciftci, N.S., Brabec, C.J., and Scherf, U. (2005) *Phys. Rev. B*, **72**, 195208.

Nibbering, E.T.J., Dreyer, J., Kühn, O., Bredenbeck, J., Hamm, P., and Elsaesser, T. (2007) in *Analysis and Control of Ultrafast Photoinduced Reactions* (eds O. Kühn, L. Wöste), Springer, Heidelberg, pp. 619–687.

Peirce, A., Dahleh, M., and Rabitz, H. (1988) *Phys. Rev. A*, **37**, 4950–4494.

Perez Lustres, J.L., Kovalenko, S.A., Mosquera, M., Senyushkina, T., Flasche, W., and Ernsting, N.P. (2005) *Angew. Chem. Int. Ed.*, **44**, 2–6.

Pfeifer, T., Spielmann, C., and Gerber, G. (2006) *Rep. Prog. Phys.*, **69**, 443–505.

Prezhdo, O.V., Duncant, W.R., and Prezhdo, V.V. (2008) *Acc. Chem. Res.*, **41**, 339–348.

Rabitz, H., de Vivie-Riedle, R., Motzkus, M., and Kompa, K. (2000) *Science*, **288**, 824–828.

Renger, T., May, V., and Kühn, O. (2001) *Phys. Rep.*, **343**, 137–254.

Riedle, E., Beutter, M., Lochbrunner, S., Piel, J., Schenkl, S., Spörlein, S., and Zinth, W. (2000) *Appl. Phys. B*, **71**, 457–465.

Rose, T.S., Rosker, M.J., and Zewail, A.H. (1989) *J. Chem. Phys.*, **91**, 7415–7436.

Rosker, M.J., Dantus, M., and Zewail, A.H. (1988) *Science*, **241**, 1200–1202.

Rubtsov, I.V. and Yoshihara, K. (1999) *J. Phys. Chem. A*, **103**, 10202–10212.

Sanov, A. and Lineberger, W.C. (2004) *Phys. Chem. Chem. Phys.*, **6**, 2018–2032.

Sansone, G., Benedetti, E., Calegari, F., Vozzi, C., Avaldi, L., Flammini, R., Poletto, L., Villoresi, P., Altucci, C., Velotta, R., Stagira, S., De Silvestri, S., and Nisoli, M. (2006) *Science*, **314**, 443–446.

Schanz, R., Kovalenko, S.A., Kharlanov, V., and Ernsting, N.P. (2001) *Appl. Phys. Lett.*, **79**, 566–568.

Scheblykin, I.G., Yartsev, A., Pullerits, T., Gulbanis, V., and Sundström, V. (2007) *J. Phys. Chem. B*, **111**, 6303–6321.

Scherer, N.F., Carlson, R.J., Matro, A., Du, M., Ruggiero, A.J., Romero-Rochin, V., Cina, J.A., Fleming, G.R., and Rice, S.A. (1991) *J. Chem. Phys.*, **95**, 1487–1511.

Schlosser, M. and Lochbrunner, S. (2006) *J. Phys. Chem. B*, **110**, 6001–6009.

Schmitt, M., Knopp, G., Materny, A., and Kiefer, W. (1998) *J. Phys. Chem. A*, **102**, 4059–4065.

Schoenlein, R.W., Peteanu, L.A., Mathies, R.A., and Shank, C.V. (1991) *Science*, **254**, 412–415.

Schultheis, L., Kuhl, J., and Honold, A. (1986) *Phys. Rev. Lett.*, **57**, 1635–1638.

Schultz, T., Samoylova, E., Radloff, W., Hertel, I.V., Sobolewski, A.L., and Domcke, W. (2004) *Science*, **306**, 1765–1768.

Schwarzer, D., Hanisch, C., Kutne, P., and Troe, J. (2002) *J. Phys. Chem. A*, **106**, 8019–8028.

Siwick, B.J., Dwyer, J.R., Jordan, R.E., and Miller, R.J.D. (2003) *Science*, **302**, 1382–1385.

Spence, D.E., Kean, P.N., and Sibbett, W. (1991) *Opt. Lett.*, **16**, 42–44.

Stock, G. and Thoss, M. (2005) *Adv. Chem. Phys.*, **131**, 243–375.

Stock, K., Schriever, C., Lochbrunner, S., and Riedle, E. (2008) *Chem. Phys.*, **349**, 197–203.

Stolow, A., Bragg, A.E., and Neumark, D.M. (2004) *Chem. Rev.*, **104**, 1719–1757.

Sundström, V. (2008) *Annu. Rev. Phys. Chem.*, **59**, 53–57.

Takeuchi, S. and Tahara, T. (2000) *Chem. Phys. Lett.*, **326**, 430–438.

Tamura, K., Ippen, E.P., Haus, H.A., and Nelson, L.E. (1993) *Opt. Lett.*, **18**, 1080–1082.

Tannor, D.J. and Rice, S.A. (1985) *J. Chem. Phys.*, **83**, 5013–5018.

Trushin, S.A., Kosma, K., Fuß, W., and Schmid, W.E. (2008) *Chem. Phys.*, **347**, 309–323.

Ullrich, S., Schultz, T., Zgierski, M.Z., and Stolow, A. (2004) *J. Am. Chem. Soc.*, **126**, 2262–2263.

Ventalon, C., Fraser, J.M., Vos, M.H., Alexandrou, A., Martin, J.-L., and Joffre, M. (2004) *Proc. Natl. Acad. Sci.*, **101**, 13216–13220.

Wang, H. and Thoss, M. (2008) *Chem. Phys.*, **347**, 139–151.

Weiner, A.M. (2000) *Rev. Sci. Instrum.*, **71**, 1929–1960.

Weiner, A.M., De Silvestri, S., and Ippen, E.P. (1985) *J. Opt. Soc. Am. B*, **2**, 654–661.

Windhorn, L., Yeston, J.S., Witte, T., Fuss, W., Motzkus, M., Proch, D., Kompa, K.-L., and Moore, C.B. (2003) *J. Chem. Phys.*, **119**, 641–645.

Wohlleben, W., Buckup, T., Herek, J.L., and Motzkus, M. (2005) *ChemPhysChem*, **6**, 850–857.

Wollenhaupt, M., Engel, V., and Baumert, T. (2005) *Annu. Rev. Phys. Chem.*, **56**, 25–56.

Woutersen, S. and Bakker, H.J. (2006) *Phys. Rev. Lett.*, **96**, 138305.

Woutersen, S. and Cristalli, G. (2004) *J. Chem. Phys.*, **121**, 5381–5386.

Woutersen, S. and Hamm, P. (2002) *J. Phys.: Cond. Matt.*, **14**, R1035–R1062.

Wurzer, A.J., Wilhelm, T., Piel, J., and Riedle, E. (1999) *Chem. Phys. Lett.*, **299**, 296–302.

Yoshihara, K., Tominaga, K., and Nagasawa, Y. (1995) *Bull. Chem. Soc. Jpn.*, **68**, 696–712.
Zewail, A.H. (2000) *J. Phys. Chem. A*, **104**, 5660–5694.
Zewail, A.H. (2006) *Annu. Rev. Phys. Chem*, **57**, 65–103.
Zinth, W. and Wachtveitl, J. (2005) *Chem. Phys. Chem*, **6**, 871–880.

Further Reading

Cho, M. (2008) *Chem. Rev.*, **108**, 1331–1418.
Diels, J.-C. and Rudolph, W. (2006) *Ultrashort Laser Pulse Phenomena: Fundamentals, Techniques, and Applications on a Femtosecond Time Scale*, Academic Press, Burlington.
Fayer, M.D. (ed.) (2001) *Ultrafast Infrared and Raman Spectroscopy*, Marcel Dekker, New York.
Jortner, J. and Bixon, M. (eds) (1999) *Adv. Chem. Phys.*, **106–107**.
Kühn, O. and Wöste, L. (eds) (2007) *Analysis and Control of Ultrafast Photoinduced Reactions*, Springer, Heidelberg.
Manz, J. (1997) in *Femtochemistry and Femtobiology* (ed. V. Sundström), World Scientific, Singapore, p. 80.
Meier, T., Thomas, P., and Koch, S.W. (2007) *Coherent Semiconductor Optics*, Springer, Heidelberg.
Meyer, H.-D., Gatti, F., and Worth, G.A. (eds) (2009) *Multidimensional Quantum Dynamics: MCTDH Theory and Applications*, Wiley-VCH Verlag GmbH, Weinheim.
Mukamel, S. (1995) *Principles of Nonlinear Optical Spectroscopy*, Oxford University Press, Oxford.

ULTRASONIC BIOPHYSICS

LEON A. FRIZZELL, *Bioacoustics Research Laboratory, Department of Electrical and Computer Engineering, University of Illinois, Urbana, Illinois, U.S.A.*

Introduction 475
1. **Basic Principles** 476
1.1 Principles of Pulse-Echo Ultrasound 478
1.2 Therapeutic and Other Applications 479
2. **Acoustical Characteristics of Biological Media** 481
2.1 Solutions and Suspensions 481
2.2 Tissues 482
3. **Measurement Techniques** 482
3.1 Loss Parameters 482
3.1.1 Attenuation 482
3.1.1.1 Insertion Techniques 482
3.1.1.2 Variable-Path Pulse Techniques 483
3.1.1.3 Substitution Method 484
3.1.1.4 Resonant-Cavity System 484
3.1.2 Absorption 485
3.2 Speed 486
3.2.1 Time of Flight 486
3.2.2 Sing-Around Method 487
3.2.3 Variable-Path Method 487
3.2.4 Substitution Method 487
3.2.5 Continuous-Wave Techniques 488
3.2.6 Pulse-Superposition Method 488
3.3 Nonlinearity Parameter 490
4. **Recent Advances in Medical Imaging and Therapy** 492
Works Cited 493
Further Reading 493

INTRODUCTION

The field of ultrasonic biophysics is very broad, encompassing a large range of frequencies and many different biological systems. The modern use of ultrasound started several decades ago in the form of sonar for military applications. Since then it has developed to the point where it is used very extensively in industry, bioengineering, and medicine. Its original applications in medicine were for therapy, where the energy absorbed from ultrasound waves was used to heat a variety of tissues and structures lying below the body surface. Ultrasound could penetrate and deposit heat within these tissues, whereas insufficient heat could be conducted from surface sources. Improved instrumentation and techniques available after World War II led to the development of ultrasound systems for medical imaging. Over the past 35 years ultrasound has developed into an indispensable clinical diagnostic tool. Currently, ultrasound is used to image most parts of the body, and more than half of all pregnant women in the United States are examined with ultrasound. This widespread utilization has resulted from ultrasound's proven clinical utility for imaging soft tissues compared to more expensive imaging techniques. The development of ultrasound, particularly for fetal examinations, has also been fostered by its safety record; no case of an adverse biological effect induced by diagnostic ultrasound has ever been reported in humans (AIUM, 1993).

Ultrasound has been used for many years to characterize biological materials. This results in large part from a desire to understand better the interaction of ultrasound with biological tissues so that clinical ultrasonic systems might be made more efficacious. However, ultrasound has also found application to the determination of rate constants in chemical reactions and in recent years has been used, via ultrasonic cavitation, to enhance desired chemical reactions that normally occur very slowly or not at all. The latter application has spawned the area of sonochemistry (Suslick, 1988), which has the potential for significantly increasing industrial production of many important chemical products, but which is beyond the scope of this article and will not be discussed further.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Section 1 discusses some of the basic principles underlying the use of ultrasound in medical imaging, therapy, and other applications. This is followed by discussions of the acoustical characteristics of biological media and the techniques for measuring them. The article ends with a brief discussion of recent developments in medical imaging and therapy.

1. BASIC PRINCIPLES

Unlike electromagnetic waves, acoustic waves require a medium for propagation. The acoustic wave phenomenon causes displacement of particles (consisting of many molecules), which results in pressure and density changes within the medium. For a traveling sinusoidal wave the variation in acoustic pressure (the difference between the total and ambient pressure) can be represented by the form

$$p = Pe^{-Ax}\cos(\omega t - kx) \qquad (1)$$

for a wave propagating in the positive x direction, where p is the acoustic pressure, P is its amplitude, $\omega = 2\pi f$ is the angular frequency where f is the frequency in hertz, $k = \omega/c$ is the propagation constant where c is the propagation speed, A is the attenuation coefficient, and t is the time. The same form applies to the excess density, particle displacement, particle velocity, and particle acceleration. The wave can experience significant attenuation, as represented by the exponential decay of amplitude with distance, during propagation in tissues. This attenuation is due in the most part to absorption—conversion of the energy in the wave to heat—and, to a much lesser degree, scattering of energy in different directions. This loss of energy with propagation is important to both diagnostic and therapeutic applications. The attenuation coefficient increases roughly linearly with frequency, but its value varies greatly among tissues (Goss *et al.*, 1978, 1980; Haney and O'Brien, 1986). It is low for most body fluids, much higher for solid tissues, and very high for bone and lung; see Secs. 2.1 and 2.2 for more detail.

The relation given by Eq. (1) is a linear relation, which applies only to low-amplitude waves. In many of the applications discussed here, the amplitude is sufficiently high that significant nonlinear propagation effects may occur. These effects include harmonic generation, where an initially sinusoidal wave will become distorted as it propagates. The distortion can be viewed as the transfer of energy from the initial, or fundamental, frequency of the wave to higher harmonics. Associated with this is an increase in attenuation of energy because the higher frequencies have a higher attenuation coefficient.

Ultrasound is typically used to image soft body tissues such as liver, but the sound beam often travels through fluids—for example, through amniotic fluid when imaging the fetus. Generally, bone and lung are not imaged with ultrasound, largely because of their very large attenuation. The attenuation processes include absorption, which is the conversion of acoustic energy to heat, and scattering, which is addressed later. The attenuation increases roughly linearly with frequency in the 2–10-MHz range typically used for medical imaging. This range represents a compromise between increased penetration at lower frequencies (due to decreased attenuation) and improved resolution associated with higher frequencies as discussed below. Thus, the lower frequencies are used when greater penetration is required such as for fetal imaging in the obese patient and higher frequencies for lesser penetration such as the examination of peripheral vascular flow.

When an acoustic wave impinges on an interface between two media of different specific acoustic impedance, a portion of the incident energy is reflected. For normal incidence on an infinite plane interface, the pressure reflection coefficient is given by (Kinsler *et al.*, 1982)

$$R = (z_2 - z_1)/(z_2 + z_1), \qquad (2)$$

where z_1 and z_2 are the specific acoustic impedances of the incident and transmitting media, respectively. For a plane wave the specific acoustic impedance is equal to the characteristic impedance, which is the product of the density and acoustic speed in the medium. The speed is dependent upon the density and the elastic properties of the medium. Thus at an interface between media exhibiting different densities or elastic properties, e.g., compressibility, some acoustic energy will be reflected. Although the reflection coefficient at an interface between muscle and bone is large (approximately 0.54), the reflection coefficient between two soft tissues such as liver and muscle is quite small (approximately 0.006). Reflection at oblique incidence obeys Snell's law (Kinsler *et al.*, 1982).

In addition to the specular reflection that occurs at an interface between two media of different specific acoustic impedance as described above (where any curvature along the interface is negligible over distances comparable to a wavelength), energy may also be scattered in all directions by inhomogeneities in the medium. An acoustic image is formed by using this scattered energy as well as specular reflections. The fraction of the incident energy reflected or scattered is very small.

Though it is convenient to consider plane waves of infinite lateral extent, as was done above, real sources generate finite beams of ultrasound. These sources may be unfocused, but for many of the applications discussed here, they are focused. Figure 1 shows the acoustic field from a typical focused source. The source consists of a piezoelectric transducer, which converts electrical to acoustic energy and vice versa. Most transducers for bioengineering and medical applications are made from ceramic materials such as a lead zirconate titanate (PZT) mixture; however, composite materials consisting of a combination of ceramic and other materials are sometimes used as sources, and polyvinylidene difluoride material is used for hydrophone transducers for measuring the ultrasound fields of sources. A circular-aperture focused source may consist of a circular disk with a plano-concave lens mounted in front to produce spherical focusing. Alternatively, the transducer itself may be a spherical segment that produces a focused field without a lens. However, increasingly these sources utilize electronic focusing methods. Such a phased-array source consists of many small individual elements, which can be excited with signals having a controlled delay with respect to one another such that the acoustic pressure fields from all elements constructively interfere at the desired focal region. Similarly, in a phased-array receiver, the signals from the different elements are combined after applying delays that result in reinforcement of the signals from a receiving focal region.

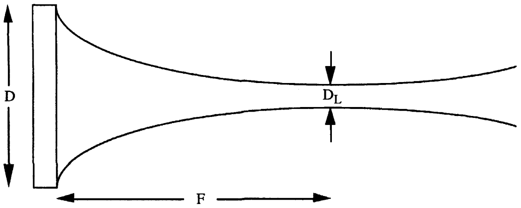

FIG. 1. Cross section of a typical focused circular ultrasound source of aperture diameter D and focal length F showing the focused beam of lateral beam width at the focus D_L.

The 3-dB lateral beam width D_L of the focused source is directly dependent upon the wavelength and focal length and inversely related to the aperture diameter (diameter of the transducer) D (Kino, 1987):

$$D_L = 1.02F\lambda/D. \tag{3}$$

Because $f\lambda = c$, the higher the frequency, the smaller is λ and D_L. For therapy systems, the energy flux per unit area (intensity) is increased when the source is strongly focused, i.e., D_L is small compared to D. For imaging systems, the smaller D_L, the better the lateral resolution near the focus, though the beam spread is greater with distance from the focus. Thus, for imaging, the strength of the focusing varies among transducers so that the user may choose very good resolution over a short region or somewhat poorer resolution that is maintained over a greater depth. With phased-array transducers, the focal region can be varied dynamically to optimize lateral resolution at all distances.

The discussion above is based on linear propagation. However, in medical imaging and therapy the amplitudes are usually so large that nonlinear propagation occurs. Nonlinear propagation is characterized by increasingly greater distortion of the wave toward the formation of a shock wave as it propagates. Thus, a wave that is initially a sinusoid will develop, in a low-loss medium, into a sawtooth wave over a propagation distance that varies inversely with initial amplitude. This distortion can also be represented in the frequency domain by the transfer of energy into higher-frequency harmonics as the wave propagates. Since energy is transferred to higher frequencies, the attenuation increases with the nonlinear propagation. In addition to the dependence on amplitude, the nonlinear distortion is dependent on properties of the medium. Specifically, the relation between acoustic pressure and density for liquids is nonlinear and varies with the medium. A parameter that characterizes the degree of nonlinearity in this relation is B/A, the ratio of the quadratic to linear coefficients in the Taylor series expansion of the relation between pressure and density; B/A is called the *nonlinearity parameter.* The larger B/A, the greater is the rate of wave distortion with propagation.

1.1 Principles of Pulse-Echo Ultrasound

Ultrasound imaging usually employs frequencies in the 2–10-MHz range, though some of the new intracavitary probes use higher frequencies. Images are formed by using a transducer within a probe to generate a short pulse (typically on the order of 1 μs or less in duration) of ultrasound that is propagated through the tissue. A portion of the energy in this pulse is reflected back toward the transducer from specular reflectors and from scatterers in the tissue. These acoustic echoes, with amplitudes much lower than the transmitted pulse, are converted by the transducer to electrical signals that are converted to a (rectified) video signal, amplified by a time-gain-controlled amplifier, and displayed. The time-gain-controlled amplifier increases the gain as a function of time so that later-arriving signals, coming from deeper within the body, are amplified more to compensate for the attenuation. The A-mode display is rarely used but simply involves display of the received echoes as amplitude versus time of arrival. The time of arrival is related by the wave speed to the tissue depth from which the echo returns, i.e., $d = ct/2$. This assumes that the speed does not vary with tissue depth, which is a good first-order approximation for soft tissues. Figure 2 provides a very simple representation of this process where the A-mode display associated with specular reflection from three different interfaces is illustrated. For clinical imaging the interfaces would not necessarily be perpendicular to the axis of the sound beam, and there would be a continuum of echoes, a continuous received signal, due to energy back-scattered from within the tissues. Since the ultrasound pulse is attenuated exponentially as it propagates, all ultrasonic imaging systems employ a logarithmic variation of amplifier gain with time to compensate the attenuation by the tissue. Thus, echoes from structures reflecting or back-scattering the same fraction of the incident signal will have the same amplitude after passing through the time-gain-controlled amplifier.

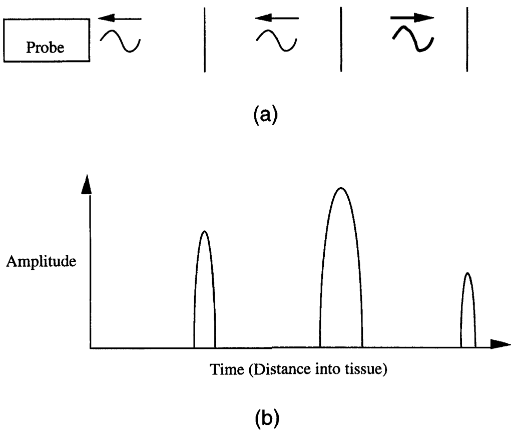

FIG. 2. Illustration of (a) the transmitted pulse (heavy wave) and echoes from reflecting structures and (b) the resulting A-mode display.

A B-mode display is typically used for ultrasound imaging. It involves display of the echoes at various brightness or gray levels corresponding to their amplitude. A two-dimensional B-mode display involves movement of the transducer (manually or automatically), movement of a mirror to change the direction of the field (automatically), or movement of the ultrasound beam directly (electrically) such that it scans a plane through the body. Figure 3 provides a simplified representation (again, echoes are shown as arising from interfaces only) of the formation of a B-mode image. The direction of the beam is monitored so that the received signals along each path are placed in their correct location on the display. Typically, the orientation information and echoes are processed by a digital scan converter for appropriate display of the two-dimensional image on a cathode-ray tube in the standard format used for television picture display. Most B-mode systems in use today create an image in 0.1 s or less, so that the image is displayed in real time for viewing of moving

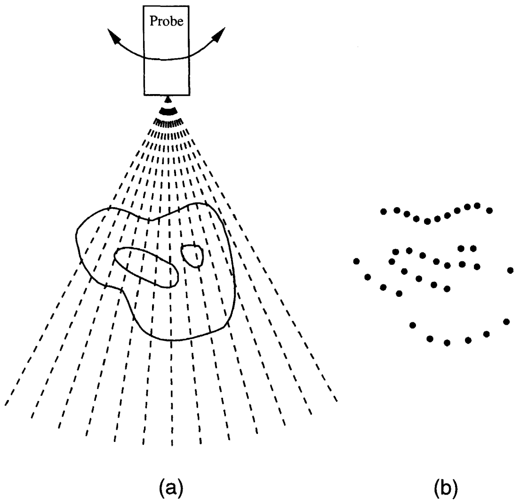

FIG. 3. Illustration of (a) the transmitted pulse paths for a rotating transducer probe and (b) the resulting two-dimensional *B*-mode display of echoes from the interfaces only.

structures, such as structures in the heart or the fetus moving within the womb. This is not possible with the typical magnetic-resonance or computed-tomography system.

Many systems now employ digital processing to enhance portions of the image. For example, it is possible to emphasize the large-amplitude, small-amplitude, or midrange signals. It is also possible to perform a more sophisticated analysis to enhance edges.

Many specialty probes have been designed for intracavitary examination. Examples include examination of the fetus with a vaginal probe, the prostate with a rectal probe, and blood vessel walls with intravascular probes. The intracavitary probe offers the advantage of decreasing the distance from the transducer to the tissue of interest and thus decreasing attenuation such that higher frequencies can be used for greater resolution. The lateral resolution of a focused probe is improved with frequency as discussed in the preceding section, but the axial (along the ultrasound beam) resolution also improves with frequency. The shorter the transmitted pulse, the better the axial resolution. Generation of shorter pulses requires sources with a larger bandwidth, which corresponds to a higher center frequency when the sources have bandwidths that are approximately the same fraction of the center frequency.

The use of ultrasound for motion detection and measurement has increased tremendously in recent years. These systems use either the Doppler principle or time-domain detection. In Doppler detection, if the ultrasound is reflected from a target moving at some speed v_t toward (away from) the source at an angle θ with respect to the beam axis, the frequency of the transmitted signal f is shifted up (down) by an amount f_D, the Doppler shift, given by

$$f_D = 2fv_t \cos\theta / c. \qquad (4)$$

In principle a measurement of f_D when f, c, and θ are known will yield the speed of the target v_t. However, it is often difficult to determine θ because the angle the transducer axis makes with a blood vessel, for example, is often unknown. Even when that angle is known, the flow is not necessarily along the direction of the vessel at every location and for all times. Time-domain detection of motion, by measuring the movement of specific echoes from one pulse to another, is an alternative to Doppler detection that is used in many ultrasound systems.

For many years duplex systems, which provide both a two-dimensional image and a Doppler signal, showing the change of target speed with time, from a particular selected target area, have been in wide use. More recently color-flow imaging has been employed, which provides a two-dimensional color (typically red or blue) image of flow toward and away from the transducer superimposed on the gray-scale image of stationary tissue structures. For these systems the speed, whether from Doppler or time-domain detection schemes, is indicated by color saturation, hue, or luminance. These systems have proven very valuable for detecting the existence of flow in a region, detecting obstructions to flow and the turbulence associated with this, detecting reduced flow, etc. Some other systems add to the color-flow image a display of speed versus time for a region that is defined by a user-movable box.

1.2 Therapeutic and Other Applications

Though therapeutic applications include the use of vibrating wires to remove plaque and augmentation of liposuction, this article is concerned with the use of ultrasound for therapeutic heating, hyperthermia, and surgery. These therapeutic applications rely almost exclusively on heating of tissues due to the absorption of the ultrasound. Therapeutic heating involves low temperature elevation in tissues and joints for a period of many minutes for physiotherapy. Typical time-averaged intensities are below approximately 3 W/cm^2. Hyperthermia produces low to moderate temperatures (in the range 43–50 °C) for approximately an hour to treat cancerous tumors, using time-averaged intensities below approximately 10 W/cm^2. In contrast, surgical treatment (sometimes called ablative hyperthermia) uses high-temperature treatment for only a few seconds to destroy tissue immediately. Time-averaged intensities below approximately 1000 W/cm^2 are used to treat the prostate, heart arrhythmia, and tumors in many regions of the body.

The beneficial effects of therapeutic heat have often been cited as increased extensibility of collagen tissues, decreased joint stiffness, pain relief, relief of muscle spasms, assistance in removing inflammatory infiltrates, edema, and exudates, and increased blood flow (Lehmann, 1990).

The rationale for hyperthermic treatment of tumor tissues is based on a belief that tumor tissues (at least some) are more susceptible than normal tissues to damage from elevated temperatures. This has been demonstrated for cells *in vitro* and clinically in patients for many tumor types. Hyperthermic treatment may involve heating the body (whole body hyperthermia), a region of the body that includes the tumor (regional hyperthermia), or the tumor alone (localized hyperthermia). Obviously, localized hyperthermia will have the least impact on normal tissues within the body. Hyperthermia is usually used in conjunction with radiation or chemotherapy. The combined modalities can be much more effective than either alone. For example, radiation has a much reduced effect on hypoxic cells, which are most affected by hyperthermia.

Several other modalities have been used for localized hyperthermia including externally applied microwaves, interstitial microwave antennas, and high-frequency currents. However, microwave frequencies useful for heating with external antennas have a longer wavelength than ultrasound and cannot be as well localized for treating the tumor and avoiding surrounding normal tissues. The use of interstitial microwave antennas involves implantation of many small antennas within the tumor to give a reasonably uniform temperature rise localized to the tumor. The use of high-frequency currents also requires invasive implantation of needle electrodes in arrays that would provide useful heating. On the other hand, bone and air spaces must be avoided with ultrasound, on account of very high attenuation and scattering, respectively. Additionally, "noninvasive" microwave or ultrasound treatment involves the use of invasive thermometry systems to monitor and control the temperature rise in the tumor and surrounding normal tissues.

Therapeutic heating and hyperthermic treatment of surface tumors can be accomplished using unfocused systems. However, since the absorbed power, $P_D = 2\alpha I$ where α is the absorption coefficient, is proportional to intensity I, the treatment depth is dependent upon the frequency because the attenuation coefficient increases with frequency and reduces the intensity much faster with depth. Thus, lower frequencies (on the order of 1 MHz or lower) are used for tumors extending deeper within the body and higher frequencies (on the order of 2–4 MHz) for shallow tumors or tumors with bone located closely behind.

Unfocused systems, which mimic plane-wave exposure, will have the greatest intensity and therefore heating at the body surface. Thus, for the treatment of deep tumors the ultrasound system must provide intensity gain to more than compensate for the attenuation along the path to the tumor so that the time-averaged intensities in the tumor are greater than those in the normal tissues between the body surface and the tumor. Simply stated, this means that the ultrasound entrance window (area) at the body surface must be substantially larger than the cross-sectional area of the tumor. Focusing systems are usually chosen to provide the required gain, but it must be remembered that the focus is scanned over the tumor and that gain in the time-averaged intensities is critical, not simply the static intensity gain of the focusing system.

Another reason that scanned focusing systems are generally used for hyperthermic treatment of deep tumors is so that the spatial power-deposition pattern within the tumor can be controlled to provide a uniform temperature throughout the tumor. Greater deposition at the tumor boundaries will lead to a more uniform temperature because of heat conduction into the tumor interior from the boundaries.

Surgical applications of ultrasound require systems with large intensity gain so that the tissue to be treated can be heated within seconds to a temperature that will destroy the tissue. Then the focal region can be relocated to treat another small volume until the tumor region is entirely destroyed.

Increasingly, systems for hyperthermic treatment of deep tumors and high-intensity focused surgical systems are changing from fixed-focus applicators that are mechanically scanned to electronic, phased-array, focusing systems. These allow rapid scanning without bulky and complex mechanical scanning systems. Additionally, the focal regions of these systems can be modified electronically. For example, the focal region can be made larger so that larger volumes can be affected with each placement of the focus.

In addition to these therapeutic applications of ultrasound, the attenuation and, to a lesser extent, the speed have been used for tissue characterization. This term generally refers to the quantitative measurement of tissue parameters as a means for diagnosis or treatment monitoring. It has been shown, for example, that the measurement of parameters associated with tissue attenuation, such as its magnitude and variation with frequency, can be quite useful to the differentiation between normal and diseased tissues. These applications have proven particularly fruitful in the eye and prostate (Feleppa *et al.*, 1996).

The main thrust of this discussion has been directed toward the medical applications of ultrasound. However, ultrasound biophysics also plays a role in determination of rate constants of chemical reactions through the measure of ultrasonic absorption and speed versus frequency, using horns to generate high intensities for breaking up of biological cells for study of their contents, and to generate cavitation for catalyzing reactions. The biophysical applications of ultrasound are many, and they cannot all be discussed in this article, but it is important to note that most of these applications rely in some way on the direct measurement of, or the knowledge of, the ultrasonic properties of the biological materials. Thus, the next sections of this article directly address the properties of biological materials and techniques for their measurement.

2. ACOUSTICAL CHARACTERISTICS OF BIOLOGICAL MEDIA

The acoustical characteristics of biological media to be discussed here include the loss parameters of attenuation and absorption, the speed of propagation, and the value of the nonlinearity parameter. Though attenuation and absorption are often used interchangeably, they are distinctly different. Absorption refers to the direct conversion of energy to heat as the acoustic wave propagates through the medium. Attenuation, on the other hand, includes absorption and all other processes by which energy is removed or redirected from the ultrasound beam, such as scattering and refraction.

Liquids and biological tissues exhibit a nonlinear relation between the acoustic pressure and the density that is important to the nonlinear propagation associated with high-amplitude waves in the medium. This nonlinear relation can be expressed as a Taylor series expansion with linear, quadratic, and higher-order terms. It has been found that the ratio of the coefficient of the quadratic term to that of the linear term, expressed as the nonlinearity parameter B/A, is very useful to describe nonlinear propagation in the medium.

2.1 Solutions and Suspensions

Solutions and suspensions of biological materials such as amino acids and proteins have been studied extensively in an effort to determine underlying mechanisms of absorption of ultrasound that might be applicable to whole tissues. They offer a simpler biological system for study and, since they are in liquid form, are amenable to measurement by more techniques. For these reasons and others many biological liquids have been measured.

The values for the attenuation and absorption coefficients are the same for solutions and nearly the same for most suspensions. Slight differences might be expected for suspensions of cells or other materials that exhibit an inhomogeneous distribution of density or compressibility that may cause scattering.

In general, all the propagation parameters (attenuation and absorption coefficients, speed, and nonlinearity parameter) increase with the concentration of the molecular species. Thus, it is not sufficient to specify a value for one of these parameters in biological liquids without specifying concentration. Specific values are listed for water, amniotic fluid, and blood in Table 1.

2.2 Tissues

Values for the attenuation coefficient, absorption coefficient, speed, and nonlinearity parameter are listed for water and selected tissues in Table 1. From Table 1 it is clear that bone and lung are very different from most soft tissues within the body. They both exhibit large loss and much higher and lower speeds, respectively. The nonlinearity parameter is generally larger in tissues than water and shows a tendency to increase with structure (Law *et al.*, 1985). On the other hand, the highest value measured for tissues to date is in fat, which also has the lowest speed.

3. MEASUREMENT TECHNIQUES

3.1 Loss Parameters

3.1.1 Attenuation Attenuation measurements may be subdivided into measurements on homogeneous and inhomogeneous media. In homogeneous materials, e.g., liquid solutions, the attenuation is equivalent to the absorption. For inhomogeneous media the attenuation includes absorption, scattering, refraction, etc., any process by which energy is removed from the wave field. While a number of attenuation-measuring techniques also yield absorption for homogeneous materials, there exist perhaps only one or two techniques that measure the absorption coefficient for inhomogeneous materials. The transient thermoelectric technique yields a measure of the absorption coefficient in tissues.

3.1.1.1 Insertion Techniques. This technique may be used with pulsed ultrasound and with phase-sensitive (first-order) or phase-insensitive (second-order) receivers. Continuous-wave (cw) signals may be used when reflection at the interfaces is negligible and when the receiver absorbs or diverts the sound beam so that a standing-wave field is not established (as with a radiation force target). The insertion technique involves the comparison of measurements of the received signal (pulsed or continuous source) with a sample of known thickness between source and receiver and with only the immersion medium between source and receiver; see Fig. 4. If the immersion fluid is of very low attenuation compared to the sample, then its attenuation can be neglected. Let A_0 or B_0 represent the amplitude of the received signal, first- or second-order quantity, respectively, with the sample removed, and let A_s or B_s represent the received signal, first- or second-order quantity, respectively, when a sample of thickness d is placed between the source and receiver. If the reflection at the interfaces is negligible or zero, the attenuation coefficient for the sample material is given by $(20/d)\log(A_0/A_s)$ or $(10/d)\log(B_0/B_s)$ dB/m and by $\ln(A_0/A_s)/d$ or $0.5\ln(B_0/B_s)/d$ Np/m, when d is in meters.

Table 1. Approximate ultrasonic propagation parameters for water and selected tissues at 3.5 MHz. Nonlinearity parameters from Law *et al.* (1985) and other data from Goss *et al.* (1978, 1980).

Tissue	Attenuation coefficient (m^{-1})	Speed (m/s)	Characteristic impedance (10^6 Pa·s/m)	Nonlinearity parameter
Water	0.2	1500	1.50	5.2
Amniotic fluid	0.3	1510	1.51	
Blood	7	1550	1.60	6.5
Fat	25	1470		11
Liver	35	1580	1.74	7.0–7.7
Muscle	50	1560	1.72	
Bone	800	3360	5.70	
Lung	1000	340	0.25	

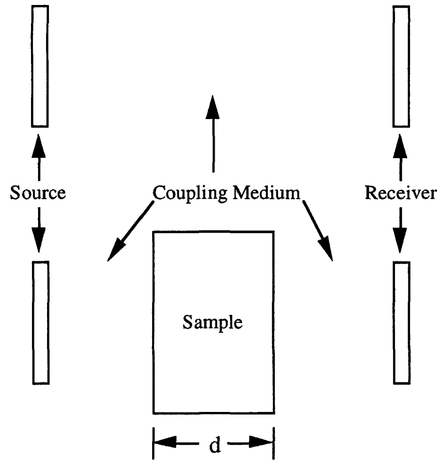

FIG. 4. Schematic representation of the insertion technique. In the upper part of the figure the source and receiver have only a coupling medium between them, whereas in the lower part of the figure a portion of the coupling medium is replaced with a sample of thickness *d*.

The effect of reflection at the interfaces can often be minimized by impedance matching. However, when solids are immersed in liquids the reflections can be very significant. To measure the attenuation coefficient within the sample for such cases, plot the measured attenuation (as opposed to the attenuation coefficient) versus the sample thickness d, as in Fig. 5. The slope gives the attenuation coefficient characteristic of the material, and the intercept ($d = 0$) gives the attenuation due to the interfaces.

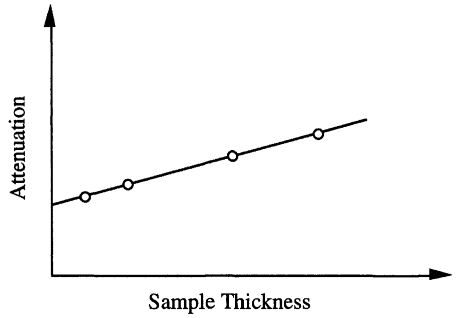

FIG. 5. Straight-line fit to attenuation versus sample thickness. The intercept is loss due to surface reflection, and the slope is the attenuation coefficient of the sample.

The insertion technique has also been used with broad-band transducers and short pulses. Such a system allows spectral analysis of the received signals, with and without the sample present, to determine the attenuation coefficient over a continuous range of frequencies limited by the bandwidth. A spectral-analysis approach has also been used for determination of the attenuation coefficient in fairly homogeneous tissues *in vivo* by measuring the decrease in signal with depth into the tissue (Kuc, 1980). This has been applied most successfully to tissues that can be examined rather directly without having to correct for effects of overlying tissues, such as tissues in the eye and the prostate (Lizzi *et al.*, 1983, 1986).

3.1.1.2 Variable-Path Pulse Techniques. Two measurement methods are described in this section. The first is a method applicable to solids and involves a sample preparation similar to that shown in Fig. 6. Transmitting and receiving transducers are mounted on opposite ends of the solid sample of length d, or a transceiving transducer is placed on one end and the other end left with an air interface to act as a reflector. When the received echos are plotted as a function of distance traveled x (determined from the speed and the length d of the rod), they look like the bold pulses (rectified and filtered rf pulses or video signals) of Fig. 7. The separation between echoes is $2d$. The attenuation coefficient A, neglecting loss at the transducer end, is determined by fitting an exponential curve of the form Be^{-Ax} to the signal levels shown in Fig. 7, where B is the amplitude at $x = 0$.

The same sort of analysis can be performed on a liquid system that has a reflector mounted at a distance d from the transducer. However, it is considerably more difficult to get good parallelism between the transducer and reflector. Therefore, in liquid systems a more common approach [based on

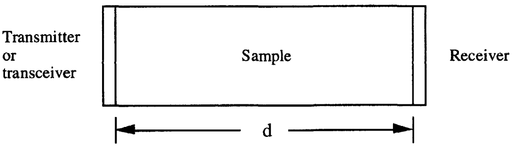

FIG. 6. Schematic diagram of solid sample of length *d* with one (transceiver) or two (transmitter and receiver) transducers mounted on the ends.

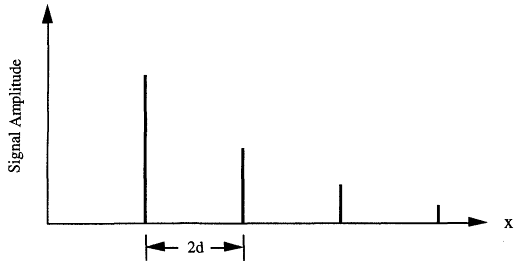

FIG. 7. Signal amplitude versus distance traveled. An exponential fit of the form Be^{-Ax} will yield the attenuation coefficient A, neglecting signal loss at the interfaces and beam diffraction effects.

a technique developed by Pellam and Galt (1946)] is to vary the distance between a source and receiver, smoothly or in discrete steps. These systems may be good over the frequency range from 3 to 100 MHz and higher; however, at the lower frequencies corrections must be made for diffraction effects in the beam.

3.1.1.3 Substitution Method. The effects of diffraction, at low frequencies, are eliminated by using the substitution technique; see Fig. 8. The fixed transducer separation minimizes the effects of diffraction. If the path in the reference medium is z and that in the test liquid is $d - z$, then the received signal A_z is

$$A_z = A_0 \exp(-\alpha_r z) \exp[-\alpha_t(d - z)], \qquad (5)$$

where α_r and α_t are the attenuation coefficients in the reference and test liquids, respectively, and A_0 is the signal with no attenuation. When the assembly is moved to a new position such that the path in the reference liquid is $z + \Delta z$, then the received signal $A_{z+\Delta z}$ is

$$A_{z+\Delta z} = A_0 \exp[-\alpha_r(z + \Delta z)] \times \exp[-\alpha_t(d - z - \Delta z)]. \qquad (6)$$

Taking the ratio of the two signals, we get

$$A_{z+\Delta z}/A_z = \exp[-(\alpha_r - \alpha_t)\Delta z]. \qquad (7)$$

Thus, the slope of the plot of the natural logarithm of the signal ratio versus Δz will give $\alpha_r - \alpha_t$. If α_r is either negligible or known, this gives α_t.

3.1.1.4 Resonant-Cavity System. The resonant-cavity system consists of two transducers separated by a fixed distance that, together with their housing, form a cylindrical cavity in which the test liquid is placed. One transducer serves as a transmitter, while the other is the receiver. When the transmitting transducer is excited with a sinusoidal voltage, standing waves result in the sample and resonances occur at particular frequencies (Eggers, 1968). At the resonance frequencies f_n, the receiving transducer registers pronounced voltage peaks, whose frequency separation depends on the sound speed in the sample, and the half-power (3 dB) bandwidth Δf of which is related to the attenuation per wavelength, $\alpha\lambda$. In fact, the attenuation coefficient of the sample may be calculated from the quality factor Q of the liquid-filled cavity, defined as the resonance frequency f_n divided by the 3 dB bandwidth Δf_n of the resonance. This quality factor is a function of the attenuation per wavelength $\alpha\lambda$ of the ultrasonic energy, and is expressed as

$$Q = (f_n/\Delta f_n) = \pi/\alpha\lambda. \qquad (8)$$

The measured Q, however, includes the effect of mechanical clamping of the transducers and losses associated with diffraction, wall effects, imperfect reflections at the transducer surface, etc., besides the desired attenuation due to the liquid. Assuming all of these energy losses are additive (Eggers and Funck, 1973), the measured quality factor Q is given by

$$1/Q_{\text{meas}} = 1/Q_{\text{liquid}} + 1/Q_{\text{extra}}, \qquad (9)$$

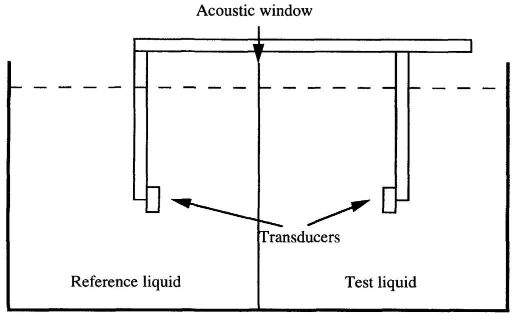

FIG. 8. Schematic diagram of the substitution apparatus.

where Q_{liquid} is the quality factor due to sound absorption in the liquid and Q_{extra} includes all the other cell losses. The absorption can be obtained by means of a reference measurement in the same cell at the same frequencies with a suitable reference liquid, having equal or very similar sound velocity to insure the same sound-field pattern for both measurements. The excess absorption per wavelength in the sample liquid is then obtained from

$$(\alpha\lambda)_{\text{excess}} = \pi(\Delta f_s - \Delta f_r)/f_n, \tag{10}$$

where Δf_s and Δf_r are the corresponding 3-dB bandwidths of the nth resonant peak in the sample and reference liquid, respectively.

In general, the transducer separation distance, the character of the ultrasonic field, the cavity shape, and the mechanical and electrical loading of the transducer, in addition to the acoustical properties of the sample liquid, all play significant roles in determining the performance of an ultrasonic resonator. Residual resonator losses such as those due to diffraction, transducer, and wall losses must be minimized so that the quality factor Q is predominantly determined by the sample-liquid absorption. Improvements of Q values for liquid-filled resonators by application of a slight static internal pressure to the cavity have been reported (Eggers and Funck, 1975). The reasons for this improvement seem to be related to the focusing of the sound beam due to the pressure-effected concavity of the transducers, reducing diffraction and side-wall effects, and to the elimination of effects of gaseous inclusions within the cavity.

With the internal pressure these systems have been used to frequencies as low as 0.2 MHz and without the pressure up to 30 MHz and more. These systems present the advantage of requiring only a small volume of sample.

3.1.2 Absorption As stated previously, for homogeneous media with no scatterers present, the absorption and attenuation coefficients are the same, and methods used to measure attenuation can be used to measure absorption. If this is not the case, as with most tissues, the transient thermoelectric technique has been used to measure the absorption coefficient.

The transient thermoelectric technique was developed in 1954 (Fry and Fry, 1954a,b) for measurement of the absorption coefficient in liquids and tissues. This method involves first imbedding a small thermocouple in a tissue sample, then irradiating the combination with an ultrasonic pulse of known intensity. The absorption coefficient α of the imbedding medium is then given as

$$\alpha = \frac{\rho CK}{2I_S}\left(\frac{dT}{dt}\right)_0, \tag{11}$$

where $(dT/dt)_0$ is the initial time rate of change of temperature (°C/s), I_S is the sound intensity at the site of the thermocouple junction (W/cm^2), and ρCK is the product of the surrounding medium's density (g/cm^3) and heat capacity at constant pressure (Cal/g-°C) and the mechanical equivalent of heat ($K = 4.186$ J/Cal) in units of J/cm$^3\cdot$°C. However, the temperature measured by use of the thermocouple junction is a true measure of the ultrasonic absorption of the surrounding medium only if the ultrasonic beam width is broad enough to minimize the effects of heat conducted away from the junction by the medium, and the size of the thermocouple is reduced to a point where the heat generated by the motion of the viscous medium relative to the thermocouple wire when subjected to ultrasound may be neglected.

The particle motion associated with the sound field causes a relative motion between the wire and the surrounding medium. Because the thermocouples are of small diameter (from a fraction of, to several, thousandths of an inch in diameter), the temperature that results from this relative motion in a viscous medium tends toward an equilibrium value in a short time (fraction of a second). However, the effect is present in a typical measurement of $(dT/dt)_0$; see Fig. 9. The heating from this source is dependent (Fry and Fry, 1954a,b) upon

1. viscosity of the medium,
2. wire size,
3. frequency,
4. the square of the particle velocity, and
5. other physical parameters.

For low-viscosity liquids and tissues, the ef-

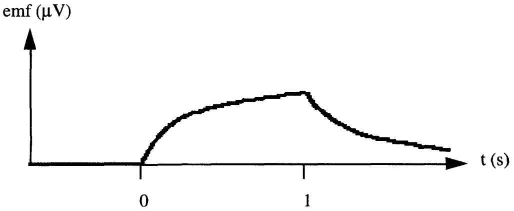

FIG. 9. Typical emf (proportional to temperature) from a thermocouple in an absorbing fluid when exposed to a 1-s ultrasound pulse.

fect of this heat source on the measurement of absorption, when $(dT/dt)_0$ is measured 0.5 s after the sound is initiated, is minimized when the wire diameter is 0.0005 in. (13 μm) or less. For additional information on this effect consult Goss *et al.* (1977).

Since real acoustic beams are not of infinite beam width, heat will diffuse from the position of peak intensity, where the thermocouple junction is located, to the surrounding lower-intensity (cooler) regions, affecting the temperature rise. The effect of this phenomenon on the measured absorption is included by adding a term to Eq. (11) to give

$$\alpha = (\rho CK/2I_S)(dT/dt)_0 + (\kappa/2I_S)\Delta^2 T, \tag{12}$$

where κ is the thermal conductivity (W/cm-°C). If the rate of heat diffusion, given by $\kappa\Delta^2 T$, is significant, then the α computed from Eq. (11) will be erroneously low. It has been shown (Goss *et al.*, 1977) that the half-power ultrasonic beam width should be at least 3–4 mm to minimize the errors from using Eq. (11) when determining α for most biological tissues.

The intensity at the site of the thermocouple junction I_S should be measured or calculated from the measured intensity with no sample present I_0 from the relation

$$I_S = I_0 e^{-2Ax}, \tag{13}$$

where A is the pressure attenuation coefficient and x is the path length in the attenuating material, i.e., the thermocouple depth. If the attenuation consists primarily of absorption, then an iterative procedure can be used to calculate α by starting with I_0 instead of I_S. The procedure involves a calculation of α from Eq. (11) with $I_S = I_0$, and the calculation of I_S from Eq. (13) with A equal to this initially calculated α. Then a new calculation of α is performed based on the computed I_S. A new I_S can then be calculated from the new α. This procedure is repeated until the value for α converges (it will converge provided that the product αx is not too large).

3.2 Speed

There exist a variety of speed-measuring techniques that are applicable to liquids and/or solids over specific frequency ranges. The techniques can be separated into pulsed and cw systems. Included in pulse techniques are the time-of-flight, sing-around, variable-path substitution, and pulse superposition methods. The cw systems employ interferometric techniques that involve either a fixed path length and a variable frequency or a fixed frequency and a variable path length. The pulse techniques are applicable to the measurement of both phase and group speeds. Broad-band pulses (short pulses) will yield group speed whereas narrow-band pulses (long pulses containing approximately 10 or 20 cycles or more) will yield the phase speed for the frequency of the rf signal constituting the pulse. Of course, the cw techniques provide phase speed. First the techniques utilizing pulses are discussed.

3.2.1 Time of Flight The straightforward measurement of the time for a pulse to travel through a known length of a specimen is perhaps the simplest means of measuring ultrasonic speed. Figure 6 illustrates the arrangement of source and receiver, or transceiver and reflector, at opposite ends of a solid specimen or a fixed path of liquid as discussed previously. The time of flight or transit time T from source to receiver or the round-trip time $2T$ may be measured by several approaches. The speed c is then given by

$$c = d/T, \tag{14}$$

where d is the length of the path.

One obvious but low-accuracy approach to determining the transit time is a simple measurement of the delay between transmitted and received pulses by measuring their separation on the time scale of an oscilloscope. Oscilloscopes are now generally available with special features that facilitate easy measurement of this time interval via digital readout after adjusting the delayed sweep to superimpose the two echoes. The accuracy of these scopes may be 1% or less, adequate for many simple measurements but rather low compared to some of the very accurate measurement techniques available.

Another means of measuring transit time is to count the number of cycles of a high-frequency timing signal that occur between triggering of the counter by the transmitted pulse and the ending of the count period by the received pulse. The errors in measurement of T for this case will depend upon the frequency of the timing signal and the timing errors of the start and stop control signals. The error arising from the count will be plus or minus one period ΔT of the wave to be counted. Thus for a timing signal of 200 MHz, ΔT will be 5 ns, resulting in a 0.3% error in the velocity of a solid of thickness 1 cm for which the true velocity is 6000 m/s.

Still another method of measuring transit time, the pulse-echo-overlap system, is discussed by Papadakis (1976). In this technique a transceiver system and several echoes from the end of the rod or the reflector are utilized. The echo signals are applied to the vertical axis of an oscilloscope, and a cw signal of variable frequency is applied to the horizontal axis. The principle of measurement is to adjust the frequency of the horizontal-axis signal so that successive echoes overlap. The transit time is then simply the period of this signal, which can be determined by an accurate measurement of frequency.

3.2.2 Sing-Around Method A block diagram of a sing-around system for velocity measurement is illustrated in Fig. 10. The principle of operation in this system is that the received pulse triggers another pulse so that a repetitive trigger signal occurs at a rate equal to the reciprocal of the propagation time. The frequency and thus the period between trigger pulses can be very accurately measured and the system is easily automated. However, any timing delays associated with the electronics will show up as errors in the determination of transit time. These electrical delays may be minimized by appropriate signal processing. This system is easily good to 1% (this is dependent upon path length) for absolute measurements but is several orders of magnitude more accurate when used as a comparison technique.

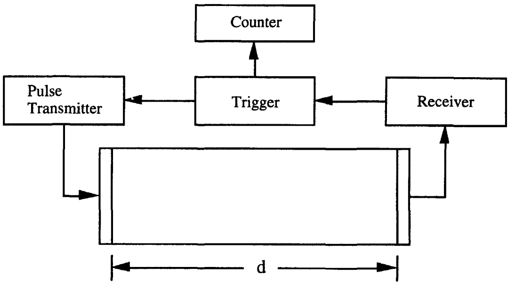

FIG. 10. Block diagram of the sing-around system for velocity measurement.

3.2.3 Variable-Path Method Another pulse method involves two transducers (a source and a receiver) whose separation is changed, e.g., increased at a constant rate by a motor and drive screw. Since the separation between transducers is continuously varied, these systems are restricted to measurements in liquids. Speed measurements are accomplished by algebraically adding a continuous-wave reference signal to the output of the receiver. The addition produces an interference pattern that exhibits a peak or minimum (a null if both signals are of equal amplitude) when the two signals are in or out of phase, respectively. The maxima occur every half wavelength, so that the velocity is easily determined after measuring the distance between maxima. These systems are used for measurements at higher frequencies where diffraction effects are smaller, e.g., over the frequency range from a few megahertz to 100 MHz.

3.2.4 Substitution Method The substitution or velocity-difference method is similar to the moving-transducer method in that the pulses are transmitted from a source to a receiver and the received signal is electronically mixed with a continuous-wave reference. However, in this method the two transducers are maintained at a fixed distance from each other, the changing phase of the received signal arising from substitution of the test liquid for a reference liquid over a portion of the signal path (Carstensen, 1954); see Fig. 8. One transducer is placed in the test liquid and the other in the reference liquid, which is separated from the test liquid by an acoustically transparent window. The substitution of one liquid for another is effected by moving the two-transducer assembly via micrometer or motor control.

The fixed transducer separation provides to this technique a way of virtually eliminating the effects of diffraction on the measurements of speed and attenuation. Thus it finds application in the low-frequency range, 0.3–15 MHz, where the errors incurred with the moving-transducer technique are unacceptably large. The accuracy of this method is good to 1 part in 10^4, provided adequate temperature control (± 0.1 °C) is maintained and the speed in the reference medium is known to that accuracy. This technique has proven useful to the investigation of dispersion in biological liquids (may be less than 0.1%). Typically, water, which is dispersionless in this frequency range, is used as the reference liquid.

To examine the technique in more detail, consider z to be the initial path length in the reference medium (taken as water in this example) and d to be the transducer separation. The number of acoustic wavelengths separating the transducers n is given by

$$n = z/\lambda_R + (d - z)/\lambda_X, \tag{15}$$

where λ_R is the wavelength in the reference medium and λ_X is the wavelength in the test liquid. If the transducer assembly is moved a distance Δz to go from one minimum in the received signal to another, the new relation for the number of wavelengths becomes, for $\lambda_X > \lambda_R$,

$$n + 1 = (z + \Delta z)/\lambda_R + (d - z - \Delta z)/\lambda_X. \tag{16}$$

This may be rearranged to give

$$c_X - c_R = c_R^2[1/(f\Delta z - c_R)], \tag{17}$$

where f is the frequency and c_R and c_X are the speeds in the reference and test liquid, respectively.

3.2.5 Continuous-Wave Techniques Continuous-wave techniques measure the variation in transducer parameters such as input impedance, voltage, or current in a standing-wave field as the transducer position or frequency is changed. Let us examine the specific acoustic impedance at the surface of a transducer located a distance L from a reflector; see Fig. 11. The acoustic pressure p_I of the wave incident on the reflector, the pressure p_R of the reflected wave, and their respective particle velocities u_I and u_R are given as follows:

$$\begin{aligned} p_I &= P_I e^{j(\omega t - kx)}, \\ p_R &= rP_I e^{j(\omega t + kx)}, \\ u_I &= (P_I/Z)e^{j(\omega t - kx)}, \\ u_R &= -(rP_I/Z)e^{j(\omega t + kx)}, \end{aligned} \tag{18}$$

where r is the complex reflection coefficient, Z is the characteristic impedance of the material, and the reflector is located at $x = 0$.

The specific acoustic impedance at a transducer located at position L will vary from infinity at a perfectly reflecting surface through zero at one-quarter wavelength from the reflector, back to infinity at one-half wavelength from the reflector, etc. Maxima in the impedance are spaced one-half wavelength apart, as are minima. Thus, the spacing for impedance maxima, or minima, can be measured to yield the speed.

3.2.6 Pulse-Superposition Method In some respects the pulse-superposition method is similar to the pulse-echo overlap technique. By superimposing the input signals over the echoes from the end of a rod of the material, the period T_0 between successive rf pulses can be measured and is given by

$$T_0 = hD - hv/2pf + n/f, \tag{19}$$

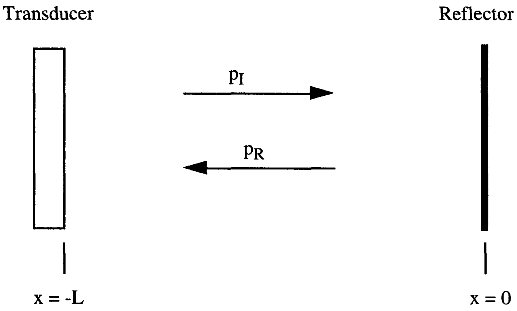

FIG. 11. A transducer and a reflector producing a standing-wave field.

where D is the round-trip delay time in the material (the parameter of interest), h is the number of round-trips through the material, v is the phase angle associated with reflection at the transducer end, n is the number of wavelengths by which pulses are shifted with respect to one another, and f is the frequency of the ultrasound pulse. The value of h is known as a part of the measurement technique, and values for v and n can be determined as described by McSkimin (1964). Thus D and the speed in a rod of known length can be determined. Consult McSkimin (1964) for more details of the technique.

A pulse-superposition technique can also be used to determine the specific acoustic impedance of shear waves in liquids. The complex reflection coefficient is determined for shear waves at the interface between a solid buffer rod of known impedance and the liquid. This is done by measuring the change in the period between successive pulses in the buffer rod (and amplitude of the pulses) when the rod termination is changed from air (perfect reflection) to the liquid to be measured.

The schematic of one system that can be used to accomplish the measurement is illustrated in Fig. 12. The repetition rate of a rf pulse generator is accurately controlled and measured. The pulse-train signal is split and each signal passed through a gate. The signal passed by gate B is attenuated before the two signals are recombined and applied to a shear transducer bonded to the buffer rod. The multiple echoes are received and displayed.

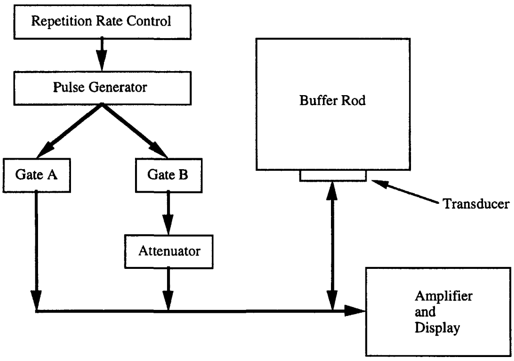

FIG. 12. Block diagram of a simple pulse-superposition system.

Figure 13 demonstrates the application of the technique. With an air interface at the end of the rod, the two gates [Fig. 13(b)] can be adjusted to let two rf pulses from the pulse train [Fig. 13(a), sinusoidal rf signal in pulses not shown], separated in time so that their echos do not overlap [Fig. 13(c)], pass to the transducer. The first echo of the B pulse is attenuated to equal the magnitude of a chosen echo of the A pulse. In Fig. 13(c) the video signals are illustrated wherein the third echo of the A pulse has been matched in amplitude by the first echo of the B pulse, $h = 2$ for this case. Then the B gate is adjusted to let an earlier pulse through so that

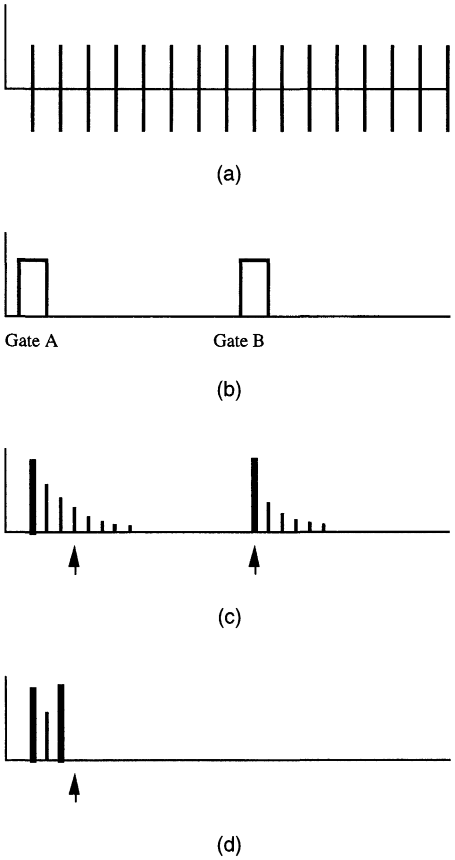

FIG. 13. Illustration of signals when using the superposition technique. (a) Selected rf signals are passed by (b) gates A and B. (c) In this example, the first echo from the gate-B series is matched in amplitude to the third echo from the gate-A series. (d) Gate B is moved and the pulse repetition frequency is adjusted to cancel echo 3 and higher from the gate-A series.

the two echoes mentioned are superimposed. The pulse repetition rate is finely adjusted to cancel the third and succeeding echos of the A pulse, Fig. 13(d). At this point the period T_0 between pulses A_3 and B_1 is given by Eq. (19) where $h = 2$. Then the liquid sample is placed on the end of the buffer rod and the same procedure followed to yield a new period:

$$T = hD - h(v - \epsilon)/2\pi f + n/f, \qquad (20)$$

where ϵ is the phase angle associated with reflection from the interface with the liquid. If T_0 and T are expressed in terms of the repetition rate frequencies $f_R = 1/T$ and $f_{R0} = 1/T_0$ and Eqs. (19) and (20) are combined, the following expression for ϵ is obtained:

$$\epsilon = 2\pi f(f_{R0} - f_R)/hf_Rf_{R0}. \qquad (21)$$

The magnitude of the reflection coefficient is determined by measuring the change in magnitude of A_4 (the change in attenuation, ΔM) that occurs upon application of the sample to the buffer rod. This is related to the magnitude of the reflection coefficient as

$$\Delta M = hB = -20h \log P, \qquad (22)$$

where ΔM is the change in the magnitude of the echo of interest in decibels, B is the change in the magnitude in decibels per impingement (echo), and P is the magnitude of the reflection coefficient (assuming total reflection from the buffer rod–air interface). This technique has proven useful in the 2–40-MHz frequency range for a number of liquids of interest (Dyro, 1972) and in tissues (Frizzell *et al.*, 1976).

3.3 Nonlinearity Parameter

The following is a brief summary of some of the techniques that have been used to measure the nonlinearity parameter B/A for biological materials. The thermodynamic method requires the accurate measure of the change in speed that occurs, either under adiabatic conditions or separately as a function of temperature and pressure; see Figs. 14 and 15. In the former approach, the speed is measured before and immediately after a rapid change in pressure, before significant heat flow can occur (change in speed measured at approximately 1 s in Fig. 15). The change in speed is determined from a measure of the change in phase of the received signal as illustrated in Fig. 14. Measurements of this type have been recently reported by Sehgal and Greenleaf (1984). From a knowledge of the change in speed with change in pressure at constant entropy, ρ_0, and c_0, B/A can be determined using the following relation:

$$\frac{B}{A} = 2\rho_0 c_0 \left(\frac{\partial c}{\partial p}\right)_{S,\rho=\rho_0}. \qquad (23)$$

It can also be shown that the expression above for B/A is equivalent to

$$\frac{B}{A} = 2\rho_0 c_0 \left(\frac{\partial c}{\partial p}\right)_{T,\rho_0} + \frac{2\beta_V T c_0}{C_P}\left(\frac{\partial c}{\partial T}\right)_{p,\rho_0}, \qquad (24)$$

where $\beta_V = V^{-1}(\partial V/\partial T)_P$ is the volume coef-

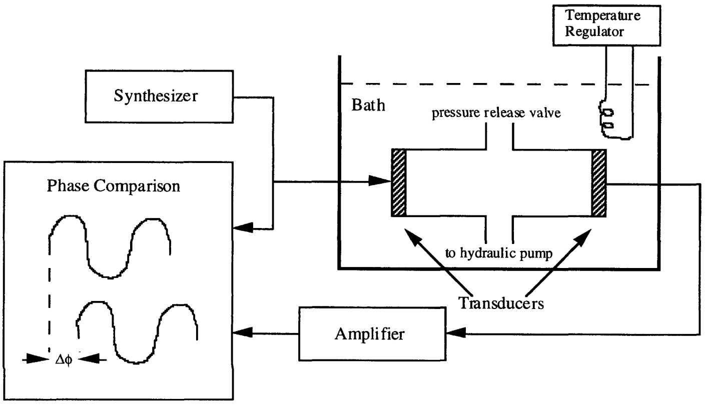

FIG. 14. Block diagram of system used to measure *B/A* by making speed measurements during a pressure change at constant entropy.

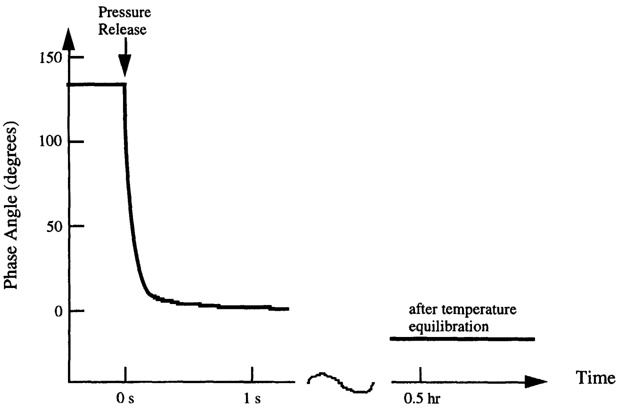

FIG. 15. Change in phase of received waveform (wave speed) with time after rapid pressure change. Final value after temperature equilibration.

ficient of thermal expansion, V is the volume, P is the pressure, C_p is the specific heat at constant pressure, and T is the temperature. Thus measurements of the speed as a function of temperature at constant pressure and as a function of pressure at constant temperature (measurement of change after approximately 0.5 h in Fig. 15) will give the two partial derivatives above. These and the values for β_v and C_p will allow computation of B/A. An example of a system that can be used to make such measurements was described by Law *et al.* (1985). This technique is also potentially very accurate but requires much longer measurement time for appropriate temperature equilibration.

Another class of measurement techniques has been referred to as finite-amplitude methods. These techniques employ the measurement of a harmonic component or components of the wave versus distance from the source. For example, for an ideal plane wave in a lossless material the Fubini solution to the nonlinear wave equation, describing the development of harmonics close to the source, gives the following rather simple form:

$$p_n(\sigma) = [2p_1(0)/n\sigma]J_n(n\sigma), \tag{25}$$

where $p_1(0)$ is the amplitude of the fundamental at the source, J_n is the nth-order Bessel function, and $\sigma = x/L$ where x is the distance from the source and L is the discontinuity distance (distance at which a shock wave starts to develop) (Beyer, 1974). Examining $n = 2$, i.e., the second-harmonic component, yields

$$\frac{B}{A} = \frac{2\rho_0 c_0^3}{\pi f}\frac{p_2(x)}{xp_1(0)^2} - 2, \tag{26}$$

where $p_2(x)$ is the value of the second harmonic as a function of distance from the source, x. This can be experimentally measured by using a receiver that is sensitive to the second harmonic and by varying the distance between source and receiver; see Fig. 16.

This is a valid determination of B/A if we take the intercept for $p_2(x)/xp_1(0)^2$. Thus

$$\frac{B}{A} = \frac{2\rho_0 c_0^3}{\pi f}\left[\frac{p_2(x)}{xp_1(0)^2}\right]_{xp_1(0)\to 0} - 2. \tag{27}$$

In lossless materials this bracketed function would be a horizontal line, but in lossy media it will intercept at an angle. However, it will be a straight line on a semilog plot (Fig. 17) such that the intercept is still valid (and easily obtained) for the determination of B/A. More details of this measurement approach can be found elsewhere (Law *et al.*, 1985). This technique is less accurate than the thermodynamic approach, but it does allow *in vivo* measurements that are not possible using thermodynamic systems, which require pressurization.

There are other acoustical methods used for measuring B/A that utilize "pump" and "probe" waves (Cain, 1986). The high-amplitude "pump" wave provides the pressure change for the low-amplitude "probe" wave. Such approaches present the possibility of imaging B/A; however, the details are quite complicated and one cannot directly measure B/A.

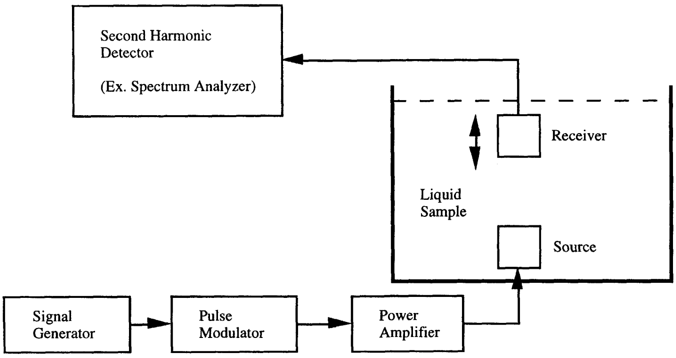

FIG. 16. Block diagram of system for measuring p_2 versus x.

4. RECENT ADVANCES IN MEDICAL IMAGING AND THERAPY

As mentioned previously, phased arrays have seen increasing use in both medical imaging and therapy. Linear arrays for medical imaging have become the norm for producing two-dimensional clinical images. However, ideally a full two-dimensional array would be used to generate a focus that could be scanned in three dimensions. The problem is that to eliminate grating lobes by keeping the interelement spacing on the order of one-half the wavelength leads to a very large number of elements that makes fabrication of fully filled transducers and the associated electronics too complex and costly. Thus, linear arrays, which allow focusing and movement of the focus in one plane, have been used almost exclusively. The thickness of this two-dimensional "slice" has been typically controlled by weak focusing in the other dimension using a cylindrical lens or a curved source.

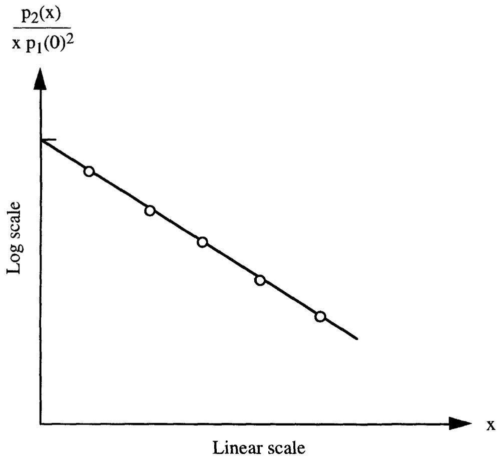

FIG. 17. Plot of parameter to determine *B*/*A*.

Recently the slice thickness has been decreased by the use of 1.5-D arrays. These are similar to the linear arrays except that they also have a few elements, to date typically no more than 5 or 7, in the other direction. Though these are not true two-dimensional arrays, they do allow the use of phasing to modify the slice thickness from what it would be with the standard mechanical focusing alone. These arrays are used increasingly in imaging systems.

Another advancement is three-dimensional imaging, the acquisition of signals from a three-dimensional volume of tissue (Tong *et al.*, 1996). This is accomplished by the systematic acquisition of two-dimensional images. The data from these can then be used, via specialized algorithms, to view two-dimensional images in various planes of choice and to determine volumes of organs, chambers, etc. much more accurately (Gilja *et al.*, 1995). Currently, three-dimensional image data are not viewed in real time, but only after some off-line processing time. However, the processing time is steadily decreasing, and these systems will be very useful in future ultrasound medical imaging.

Therapeutic ultrasound developments also center very much around the use of phased arrays for electronic control of treatment (Chapelon *et al.*, 1993; Fan and Hynynen, 1994; Wan *et al.*, 1995; Kluiwstra *et al.*, 1995; Hutchinson *et al.*, 1995; Goss *et al.*, 1996). Different groups have taken different approaches to the development of systems that employ phased-array technology but with a limited number of elements. Phased-array systems have been proposed for treatment of the heart (Kluiwstra *et al.*, 1995), the prostate (Hutchinson *et al.*, 1995), and other regions of the body (Wan *et al.*, 1995; Goss *et al.*, 1996).

Works Cited

AIUM (1993), *Bioeffects and Safety of Diagnostic Ultrasound,* Laurel, MD: American Institute of Ultrasound in Medicine, p. 86.

Cain, C. A. (1986), *J. Acoust. Soc. Am.* **80,** 28–32.

Carstensen, E. L. (1954), *J. Acoust. Soc. Am.* **26,** 858–861.

Chapelon, J. Y., Faure, P., Plantier, M., Cathignol, D., Souchon, R., Gorry, F., Gelet, A. (1993), in: M. Levy, B. R. McAvoy (Eds.), *Proceedings of the 1993 IEEE Ultrasonics Symposium,* Piscataway, NJ: IEEE, pp. 1211–1214.

Dyro, J. F. (1972), Ph.D. Thesis, University of Pennsylvania, Philadelphia.

Eggers, F. (1968), *Acustica* **19,** 323–328.

Eggers, F., Funck, Th. (1973), *Rev. Sci. Instrum.* **44,** 969–977.

Eggers, F., Funck, Th. (1975), *J. Acoust. Soc. Am.* **57,** 331–333.

Fan, X., Hynynen, K. (1994), *Med. Phys.* **22,** 297–306.

Feleppa, E. J., Kalisz, A., Sokil-melgar, J. B., Lizzi, F. L., Liu, T., Rosando, L., Shao, M. C., Fair, W. R., Wang, Y., Cookson, M. S., Reuter, V. E., Heston, W. D. W. (1996), *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **UFFC-43,** 609–619.

Frizzell, L. A., Carstensen, E. L., Dyro, J. F. (1976), *J. Acoust. Soc. Am.* **60,** 1409–1411.

Fry, W. J., Fry, R. B. (1954a), *J. Acoust. Soc. Am.* **26,** 294–310.

Fry, W. J., Fry, R. B. (1954b), *J. Acoust. Soc. Am.* **26,** 311–317.

Gilja, O. H., Smievoll, A. I., Thune, N., Matre, K., Hausken, T., Odegaard, S., Berstad, A. (1995), *Ultrasound Med. Biol.* **21,** 25–32.

Goss, S. A., Cobb, J. W., Frizzell, L. A. (1977), in: J. deKlerk, B. R. McAvoy (Eds.), *1977 Ultrasonics Symposium Proceedings,* Piscataway, NJ: IEEE, pp. 206–211.

Goss, S. A., Johnston, R. L., Dunn, F. (1978), *J. Acoust. Soc. Am.* **64,** 423–457.

Goss, S. A., Johnston, R. L., Dunn, F. (1980), *J. Acoust. Soc. Am.* **68,** 93–108.

Goss, S. A., Frizzell, L. A., Kouzmanoff, J. T., Barich, J. M., Yang, J. M. (1996), *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **UFFC-43,** 1111–1121.

Haney, M. J., O'Brien, Jr., W. D. (1986), in: J. F. Greenleaf (Ed.), *Tissue Characterization with Ultrasound,* Vol. 1, Boca Raton, FL: CRC Press, Chap. 2.

Hutchinson, E. B., Buchanan, M. T., Hynynen, K. (1995), in: M. Levy, S. C. Schneider, B. R. McAvoy (Eds.), *Proceedings of the 1995 IEEE Ultrasonics Symposium,* Piscataway, NJ: IEEE, pp. 1601–1604.

Kino, G. S. (1987), *Acoustics: Devices, Imaging, and Analog Signal Processing,* Englewood Cliffs, NJ: Prentice-Hall.

Kinsler, L. E., Frey, A. R., Coppens, A. B., Sanders, J. V. (1982), *Fundamentals of Acoustics,* 3rd ed., New York: Wiley.

Kluiwstra, J.-U., Zhang, Y., VanBaren, P., Strickberger, S. A., Ebbini, E. S., Cain, C. A. (1995), in: M. Levy, S. C. Schneider, B. R. McAvoy (Eds.), *Proceedings of the 1995 IEEE Ultrasonics Symposium,* Piscataway, NJ: IEEE, pp. 1605–1608.

Kuc, R. (1985), *Proc. IEEE* **73,** 1159–1163.

Law, W. K., Frizzell L. A., Dunn, F. (1985), *Ultrasound Med. Biol.* **11,** 307–318.

Lehmann, J. F., de Lateur, B. J. (1990), in: J. F. Lehmann (Ed.), *Therapeutic Heat and Cold,* 4th ed., Baltimore: Williams and Wilkins, Chap. 9.

Lizzi, F. L., Greenebaum, M., Feleppa, E. J., Elbaum, M., Coleman, D. J. (1983), *J. Acoust. Soc. Am.* **73,** 1366–1373.

Lizzi, F. L., Ostromogilsky, M., Feleppa, E. J., Rorke, M. C., Yaremko, M. M. (1986), *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **UFFC-33,** 319–329.

McSkimin, H. J. (1964), in: W. P. Mason (Ed.), *Physical Acoustics,* Vol. I, New York: Academic, Chap. 4.

Papadakis, E. P. (1976), in: W. P. Mason and R. N. Thurston (Eds.), *Physical Acoustics,* Vol. XI, New York: Academic, Chap. 5.

Pellam, J. R., Galt, J. K. (1946), *J. Chem. Phys.* **14,** 608–614.

Sehgal, C. M., Greenleaf, J. F. (1984), *J. Acoust. Soc. Am.* **76,** 1023–1029.

Suslick, K. (Ed.) (1988), *Ultrasound: Chemical, Physical, and Biological Effects,* New York: VCH Publishers.

Tong, S., Downey, D. B., Cardinal, H. N., Fenster, A. (1996), *Ultrasound Med. Biol.* **22,** 735–746.

Wan, H., VanBaren, P., Ebbini, E., Cain, C. (1995), in: M. Levy, S. C. Schneider, B. R. McAvoy (Eds.), *Proceedings of the 1995 IEEE Ultrasonics Symposium,* Piscataway, NJ: IEEE, pp. 1623–1628.

Further Reading

Beyer, R. T. (1974), *Nonlinear Acoustics,* Washington, DC: Naval Ship Systems Command, Department of the Navy.

Breazeale, M. A., Cantrell, J. H., Jr., Heyman, J. S. (1981), in: P. D. Edmonds (Ed.), *Methods of Experimental Physics: Ultrasonics,* Vol. 19, New York: Academic, Chap. 2.

Christensen, D. A. (1988), *Ultrasonic Bioinstrumentation,* New York: IOP Publishing.

Fullerton, G. D., Zagzebski, J. A. (Eds.) (1980), *Medical Physics of CT and Ultrasound,* New York: American Institute of Physics.

Hill, C. R. (1986), *Physical Principles of Medical Ultrasonics,* New York: Wiley.

Kremkau, F. W. (1993), *Diagnostic Sonography: Principles and Instruments,* 4th ed., Philadelphia: W. B. Saunders Co.

Tsui, B., Smith, M., Shung, K. K. (1992), *Principles of Medical Imaging,* New York: Academic.

Webb, S. (Ed.) (1988), *The Physics of Medical Imaging,* New York: IOP Publishing.

Wells, P. N. T. (1977), *Biomedical Ultrasonics,* New York: Academic.

Wells, P. N. T., Ziskin, M. C. (Eds.) (1980), *New Techniques and Instrumentation in Ultrasonography,* New York: Churchill Livingstone.

7
Ultrasonics

John H. Cantrell

7.1 Introduction 197
7.2 Ultrasonic Waves in Boundless Media 198
7.2.1 Solids 198
7.2.1.1 Stress, Strain, and Elastic Constants 198
7.2.1.2 Modal Wave Propagation 201
7.2.2 Fluids 202
7.3 Ultrasonic Waves in Bounded Media 203
7.3.1 Reflection and Refraction 203
7.3.2 Resonance Phenomena 205
7.3.3 Surface and Interface Waves 206
7.3.4 Guided Waves 208
7.4 Ultrasonic Attenuation 209
7.4.1 Scattering 210
7.4.2 Diffraction 214
7.4.3 Relaxation Processes 214
7.4.3.1 Relaxation in Solids 215
7.4.3.2 Relaxation in Fluids 217
7.4.3.3 Viscoelasticity 218
7.5 Wave Dispersion 219
7.6 Ultrasonic Generation and Detection 220
7.6.1 Piezoelectric Transducers 220
7.6.2 Noncontacting Methods 221
7.6.2.1 Capacitive Transducers 221
7.6.2.2 Electromagnetic Acoustic Transducers (EMATs) 222
7.6.2.3 Laser Generation and Detection 222
7.6.2.4 Modulated Particle Beam Generation 224
7.7 Measurement Techniques 224
7.7.1 Pulse Techniques 225
7.7.1.1 Basic Pulse-Echo Method 225
7.7.1.2 Pulse Superposition Method 225
7.7.1.3 Echo-Overlap Method 226

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

7.7.1.4 Pulse Interferometer Methods 226
7.7.2 Continuous Wave Techniques 227
7.7.3 Optical Techniques 228
7.7.3.1 Diffraction of Light by Sound 228
7.7.3.2 Brillouin Scattering 229
7.7.4 Sources of Error 230
7.8 Applications of Ultrasound 231
7.8.1 Medical Applications 231
7.8.2 Industrial Applications 232
7.8.3 Acoustic Microscopy 234
7.8.3.1 Scanning Acoustic and Laser Acoustic Microscopes 234
7.8.3.2 Scanning Electron Acoustic and Photoacoustic Microscopes 236
7.8.3.3 Acoustic-Atomic Force Microscopes 237
7.8.4 Underwater Acoustics 239
Glossary 239
References 240
Further Reading 241

7.1
Introduction

Generally, sound (sonics) refers to vibratory mechanical disturbances at all frequencies in all substances. Ultrasound (ultrasonics) refers to such disturbances in the range of frequencies above that of audible sound (approximately 20 kHz) up to pretersonic frequencies (10 GHz and higher). Recently, the term *hypersonics* has also come into use to describe sound at frequencies above 1 GHz. However, the term is somewhat ambiguous, because in aeroacoustics hypersonics refers to the study of phenomena associated with objects traveling faster than five times the velocity of sound in air.

The study and application of ultrasonics as a serious scientific discipline began about 1910 with the development of piezoelectric transducers. Early applications of ultrasound centered on the development of a primitive form of sonar, communications devices, and light modulators used in early experiments with television. Following improvement in high-frequency electronics during World War II, the pace of ultrasonics research and applications accelerated. The first acoustic delay lines were applied to radar systems to produce a moving target indicator and used as short term memories in early computers. Today, ultrasonic methods and techniques are widely used in such far-ranging fields of application as low temperature and solid state physics, signal processing and devices, materials processing and testing, nondestructive evaluation, medical diagnostics and imaging, Doppler sonar, microstructural materials characterization, acoustic microscopy, chemical kinetics, viscoelasticity, lubrication, fluid flow, ultrasonic motors, materials forming, and welding among others.

This article is divided into three broad categories: (i) physical and technical principles of ultrasound; (ii) ultrasonic generation, detection, and measurements; and (iii) applications. The first category is covered in four main sections (Sections 7.2–7.5), which establish the physical basis and phenomenological constructs of ultrasound. The development emphasizes the intrinsic connection to reversible and irreversible thermodynamics. The second category is covered in two main sections (Sections 7.6 and 7.7), which provide an overview of the most commonly used methods and equipment for the generation and detection of ultrasonic signals and techniques for measuring ultrasonic parameters. The article concludes in Section 7.8 with a survey of representative applications of ultrasound including some

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

recent applications in the final stage of development.

7.2
Ultrasonic Waves in Boundless Media

Many applications of ultrasound involve wave propagation that is minimally influenced by geometrical constraints or boundaries. Such conditions occur when the time duration of an ultrasonic wave, that is, an ultrasonic pulse or toneburst (gated continuous wave (cw)), is short compared to the propagation time through the medium. For sufficiently long tonebursts, the wave propagation may be modeled to a good approximation as a cw propagating in a boundless medium. The relative simplicity of such models emphasizes the basic physical properties of the propagation medium. We begin by considering the propagation of ultrasonic waves in thermodynamically reversible solids and fluids to explore the most basic physical features of these media that influence wave propagation.

7.2.1
Solids

We may consider a solid to be a material in which the thermodynamic state of equilibrium depends on the fixed mean configuration of the particles comprising the material and either the temperature T or the entropy S of the system. We denote the present configuration of the material particles by the set of particle vectors $\{\boldsymbol{x}\}$. The relevant state functions describing the system of particles are the Helmholtz free energy per unit mass $F(\boldsymbol{x}, \boldsymbol{T})$ and the internal energy per unit mass $U(\boldsymbol{x}, S)$.

7.2.1.1 **Stress, Strain, and Elastic Constants**

We assume that the present configuration of material particles of the solid is obtained by means of a small (infinitesimal) elastic deformation from some initial configuration of particles denoted by the set of particle vectors $\{\boldsymbol{a}\}$. We describe the elastic deformation by the set of transformation coefficients $\alpha_{ij} = \partial x_i/\partial a_j$ such that $x_i = \alpha_{ij} a_j$. The indices i and j take the values 1, 2, 3 representing three mutually orthogonal spatial reference axes. We define the resulting particle displacements u_i by $u_i = x_i - a_i$ and write $\alpha_{ij} = \delta_{ij} + u_{ij}$, where the displacement gradients $u_{ij} = \partial u_i/\partial a_j$, and δ_{ij} is the Kronecker delta ($\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ if $i \neq j$). The deformation of the solid results in a change in the mass density of the material described by the relation (Wallace, 1970)

$$\frac{\rho_0}{\rho} = J = \det[\alpha_{ij}] \tag{7.1}$$

where $\rho_0 = \rho_0(\boldsymbol{a})$ is the mass density in the initial state and $\rho = \rho(\boldsymbol{x})$ is the mass density in the final state.

Let $\Delta\boldsymbol{a}$ be the vector between any two material particles in the initial state and $\Delta\boldsymbol{x}$ be the vector between the same two particles in the final state. If the deformation is constant over the region defined by $\Delta\boldsymbol{x}$, then the relation (Murnaghan, 1951) (repeated suffix denotes summation)

$$|\Delta\boldsymbol{x}|^2 - |\Delta\boldsymbol{a}|^2 = 2\eta_{ij}\Delta a_i \Delta a_j \tag{7.2}$$

holds where

$$\begin{aligned}\eta_{ij} &= \frac{1}{2}(\alpha_{ki}\alpha_{kj} - \delta_{ij}) \\ &= \frac{1}{2}(u_{ij} + u_{ji} + u_{ki}u_{kj})\end{aligned} \tag{7.3}$$

The symmetric parameters η_{ij} are called the *Lagrangian strains*. For sufficiently small strains, the terms $u_{ki}u_{kj}$ can be neglected in Equation (7.3) and the Lagrangian strains reduce to the stain measure $\varepsilon_{ij} = (1/2)(u_{ij} + u_{ji})$ conventionally defined in linear elasticity. The advantage of the Lagrangian strains is that the distances between all pairs of material particles can be completely specified by the initial configuration of particles $\{\boldsymbol{a}\}$ and the Lagrangian strains η_{ij} defined over any region for which the η_{ij} are constant. Because the Helmholtz free energy and the internal energy depend only on the relative positions of the particles, we may write that $F(\boldsymbol{x}, S) = F(\boldsymbol{a}, \eta_{ij}, S)$ and $U(\boldsymbol{x}, S) = U(\boldsymbol{a}, \eta_{ij}, S)$. For small deformations, we may expand $U(\boldsymbol{a}, \eta_{ij}, S)$ about $\boldsymbol{a}$ (i.e., about $\eta_{ij} = 0$) to first order in η_{ij} as (Wallace, 1970)

$$\rho_0 U(a, \eta_{ij}, S) = \rho_0 U(a, 0, S) + C^S_{ij}\eta_{ij} + \frac{1}{2}C^S_{ijkl}\eta_{ij}\eta_{kl} + \cdots \quad (7.4)$$

where

$$C^S_{ij} = \rho_0\left(\frac{\partial U}{\partial \eta_{ij}}\right)_s, \quad C^S_{ijkl} = \rho_0\left(\frac{\partial^2 U}{\partial \eta_{ij}\partial \eta_{ij}}\right)_s \quad (7.5)$$

A similar expansion may be obtained for the Helmholtz free energy with coefficients C^T_{ij} and C^T_{ijkl} defined in terms of appropriate strain derivatives of the Helmholtz free energy per unit mass.

The coefficients C^S_{ij} are the isentropic initial stresses in the solid $(\sigma_0)_{ij}$ (sometimes called the *thermodynamic tensions*) and the coefficients C^S_{ijkl} are the isentropic elastic constants (often called the *second-order Brugger elastic constants*) of the solid. The C^T_{ij} and C^T_{ijkl} are the corresponding isothermal coefficients. For sufficiently small strains such that the terms $u_{ki}u_{kj}$ can be neglected in Equation (7.3), we may take the derivative of Equation (7.4) with respect to the strains to obtain the isentropic equation of state

$$\sigma_{ij} = (\sigma_0)_{ij} + C^S_{ijkl}\varepsilon_{kl} + \cdots \quad (7.6)$$

where $\sigma_{ij} = \rho_o(\partial U/\partial \varepsilon_{ij})$ are the stress components in the final state, $(\sigma_0)_{ij}$ are the stress components in the initial state, and ε_{ij} are the conventional strains. A similar isothermal expression holds using the Helmholtz free energy. Equation (7.6), known as the *generalized Hooke's law*, clearly shows that the second-order elastic constants serve to linearly couple the stress and the strain in perfectly elastic bodies.

Both the isentropic and the isothermal C_{ijkl} are fourth rank tensors that generally have 81 components. However, the stress symmetry $(\sigma_{ij} = \sigma_{ji})$ requires $C_{ijkl} = C_{jikl}$. The strain symmetry $(\varepsilon_{kl} = \varepsilon_{lk})$ requires $C_{ijkl} = C_{jilk}$. These two conditions reduce the number of independent second-order elastic constants to 36. The quadratic dependence of the energy on the strain, shown in Equation (7.4), means that the energy must remain unchanged not only under an interchange of strain suffix pairs i, j and k, l, but also under an interchange of i, k and j, l. Thus, the relation $C_{ijkl} = C_{jikl} = C_{ijlk} = C_{klij}$ holds for the elastic constants. This relationship reduces the number of independent constants to 21.

The crystal symmetry of a material may also reduce the number of independent elastic constants, because the constants are derivatives of the thermodynamic state functions with respect to crystallographically equivalent strains. A detailed derivation of the number of independent elastic constants may be found in Nye (1960). The number of independent constants for the different crystal systems is summarized in Table 7.1. The point groups in Table 7.1

Tab. 7.1 Number of independent elastic constants for different crystal systems.

Crystal system	Laue group	Point groups	Elastic constants
Triclinic	N	1, $\bar{1}$	21
Monoclinic	M	m, 2, $\frac{2}{m}$	13
Orthorhombic	O	2mm, 222, $\frac{2}{m}\frac{2}{m}\frac{2}{m}$	9
Tetragonal	TII	4, $\bar{4}$, $\frac{4}{m}$	7
Tetragonal	TI	4mm, $\bar{4}$2m, 422, $\frac{4}{m}\frac{2}{m}\frac{2}{m}$	6
Rhombohedral	RII	3, $\bar{3}$	7
Rhombohedral	RI	3m, 32, 3 $\frac{2}{m}$	6
Hexagonal	HII	$\bar{6}$, 6, $\frac{6}{m}$	5
Hexagonal	HI	$\bar{6}$2m, 6mm, 622, $\frac{6}{m}\frac{2}{m}\frac{2}{m}$	5
Cubic	CII	23, $\frac{2}{m}\bar{3}$	3
Cubic	CI	$\bar{4}$ 3m, 432, $\frac{4}{m}\bar{3}\frac{2}{m}$	3
Isotropic	I		2

reflect all symmetry operations possible for a given crystal system. A symmetry operation for a boundless crystal is an operation that interchanges the positions of the various atoms of the crystal into equivalent positions such that the crystal appears exactly the same as before the operation. For example, crystals belonging to the monoclinic system have one of three possible point groups given as m, 2, or $2/m$ in the full international notation. The symbol m denotes an operation corresponding to the reflection of the crystal atoms across a mirror plane, 2 denotes an operation corresponding to 180° rotations (twofold rotations) around a symmetry axis, and $2/m$ denotes an operation corresponding to reflection across a mirror plane normal to a twofold rotation axis. The point groups are given in the full international notation in Table 7.1 with the full set of possible point groups listed for each crystal system.

The full tensor notation C_{ijkl} for the elastic constants is often replaced by the abbreviated notation (sometimes called the *matrix* or *Voigt notation*) $C_{\alpha\beta}$ according to the prescription:

ij or $kl =$	11	22	33	23, 32	13, 31	12, 21
α or $\beta =$	1	2	3	4	5	6

It is instructive to write the set of elastic constants in Voigt notation for triclinic, cubic, and isotropic symmetry classes in the half matrix forms (the full matrices are symmetrical about the diagonal):

Triclinic : 21 constants

$$\begin{array}{cccccc}
c_{11} & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} \\
 & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} \\
 & & c_{33} & c_{34} & c_{35} & c_{36} \\
 & & & c_{44} & c_{45} & c_{46} \\
 & & & & c_{55} & c_{56} \\
 & & & & & c_{66}
\end{array}$$

Isotropic : 2 constants

$$\begin{array}{cccccc}
c_{11} & c_{12} & c_{12} & 0 & 0 & 0 \\
 & c_{11} & c_{12} & 0 & 0 & 0 \\
 & & c_{11} & 0 & 0 & 0 \\
 & & & c_{44} & 0 & 0 \\
 & & & & c_{44} & 0 \\
 & & & & & c_{44}
\end{array}$$

Cubic (all classes) : 3 constants

$$
\begin{array}{cccccc}
c_{11} & c_{12} & c_{12} & 0 & 0 & 0 \\
 & c_{11} & c_{12} & 0 & 0 & 0 \\
 & & c_{11} & 0 & 0 & 0 \\
 & & & \frac{1}{2}(c_{11}-c_{12}) & 0 & 0 \\
 & & & & \frac{1}{2}(c_{11}-c_{12}) & 0 \\
 & & & & & \frac{1}{2}(c_{11}-c_{12})
\end{array}
$$

For the remaining symmetry classes, see Nye (1960).

For isotropic symmetry Hooke's law, Equation (7.6) may be written as (Pollard, 1977) $\sigma_{ij} = (C_{11} - C_{12})\varepsilon_{ij} + C_{12}\varphi\delta_{ij} = 2G\varepsilon_{ij} + \lambda\varphi\delta_{ij}$, where the dilatation $\varphi = \varepsilon_{11} + \varepsilon_{22} + \varepsilon_{33} = \operatorname{div} \boldsymbol{u}$ ($\boldsymbol{u}$ is the displacement vector). The Brugger elastic constants in the last equality are replaced by the Lamé constants • and G, where $\lambda = C_{12}$ and $G = (1/2)(C_{11} - C_{12})$. The constant G is also called the *shear modulus* or *rigidity modulus*.

Another important modulus is the *bulk modulus B* defined as the ratio of the change in hydrostatic pressure, Δp, to the fractional change in the volume of the material, $\Delta V/V_0 = \Delta J$, produced when pressure is applied to an isotropic material (including fluids). It is related to the Brugger constants and the Lamé constants as $B = (1/3)(C_{11} + 2C_{12}) = \lambda + (2/3)G$.

A related constant is the compressibility $K = 1/B$.

7.2.1.2 Modal Wave Propagation

For a solid of arbitrary crystalline symmetry, we may substitute Equation (7.6) into Newton's Law, $\rho_0(\partial^2 u_i/\partial t^2) = (\partial\sigma_{ij}/\partial a_j)$, where t is time and body forces are neglected. We obtain the isentropic wave equation

$$\rho_0 \frac{\partial^2 u_i}{\partial t^2} = C^S_{ijkl} \frac{\partial^2 u_k}{\partial a_j \partial a_l} \tag{7.7}$$

It is important to note that for the case of small elastic deformations, the transformation coefficients $\alpha_{ij} \approx \delta_{ij}$ and the wave equation, Equation (7.7), may be written with the variables a_i (often called the *material or Lagrangian coordinates*) replaced by x_i (often called the *spatial or Eulerian coordinates*) (Thurston, 1964). Thus, in the "linear acoustic approximation," the coordinate variables in the wave equation may be either the set a_i or x_i. In this article, both sets are used depending on notational convenience.

An equation similar to Equation (7.7) holds for isothermal wave propagation. Whether isentropic or isothermal wave propagation occurs depends on the thermodynamic properties of the material (Pierce, 1989). Although isentropic (strictly adiabatic) conditions prevail in most cases, we shall drop the isentropic and isothermal designations except where it is appropriate to distinguish a specific condition.

We assume the propagation of plane harmonic waves of the form $u_i = w_i \exp[i(k_j a_j - \omega t)]$, where w_i are the Cartesian components of the normalized wave amplitude $\boldsymbol{w}(|\boldsymbol{w}| = 1)$, k_j are the components of the wave vector $\boldsymbol{k}$, and ω is the angular frequency of the wave. Substituting this form into Equation (7.7), defining $\boldsymbol{n} = \boldsymbol{k}/|\boldsymbol{k}|$ to be the unit vector along the wave propagation direction, and defining the wave phase velocity $c = \omega/|\boldsymbol{k}|$, we obtain the Christoffel equation

$$(\rho_0 c^2 \delta_{ik} - C_{ijkl} n_j n_l) w_k = 0 \qquad (7.8)$$

For a given direction of propagation $\boldsymbol{n}$, Equation (7.8) is a set of three homogeneous equations in three unknowns w_k, where $k = 1, 2, 3$. Nonzero solutions to Equation (7.8) exist if the determinant of the set of terms in parentheses (the secular determinant) is zero. Evaluation of the secular determinant leads to a cubic equation in c^2. The three roots of this equation correspond in general to three different propagation velocities. Substitution of each of these velocities back into Equation (7.8) yields a solution set of w_k, called the *wave polarization vector*, corresponding to that velocity. The polarization vectors are mutually orthogonal. If the polarization vector is parallel or nearly parallel to the propagation vector, the wave is called a *longitudinal or quasi-longitudinal wave.* If the polarization vector is perpendicular or nearly perpendicular to the propagation vector, the wave is called a *transverse (shear) or quasi-transverse (quasi-shear) wave.* In general, the one longitudinal wave and the two shear waves have different velocities.

From Equation (7.8), we find that the velocity of propagation of a wave of polarization $\boldsymbol{w}$ along the direction $\boldsymbol{n}$ is determined as $c = (C_{ijkl} n_j n_l w_i w_k / \rho_0)^{1/2}$. It is apparent from this expression that the crystalline symmetry influences the sound velocity through the elastic constants. For example, for plane wave propagation along the [100] crystalline axis of cubic solids, the longitudinal velocity $c_1 = (C_{11}/\rho_0)^{1/2}$ and the shear wave velocities $c_2 = c_3 = (C_{44}/\rho_0)^{1/2}$; along the [110] direction $c_1 = [(C_{11} + C_{12} + 2C_{44})/2\rho_0]^{1/2}$, $c_2 = (C_{44}/\rho_0)^{1/2}$, and $c_3 = [(C_{11} - C_{12})/2\rho_0]^{1/2}$; and along [111] $c_1 = [(C_{11} + 2C_{12} + 4C_{44})/3\rho_0]^{1/2}$ and $c_2 = c_3 = [(C_{11} - C_{12} + C_{44})/3\rho_0]^{1/2}$.

Along any wave propagation direction in materials having isotropic symmetry, the longitudinal velocity $c_1 = (C_{11}/\rho_0)^{1/2} = [(\lambda + 2G)/\rho_0]^{1/2} = [(B + (4/3)G)/\rho_0]^{1/2}$ and the shear wave velocities $c_2 = c_3 = [(C_{11} - C_{12})/2\rho_0]^{1/2} = (G/\rho_0)^{1/2}$. This result can also be obtained by substituting the isotropic Hooke's law, $\sigma_{ij} = 2G\varepsilon_{ij} + \lambda\varphi\delta_{ij}$, into Newton's law to obtain (after some manipulation) (Pollard, 1977)

$$\begin{aligned}\rho_0 \frac{\partial^2 \boldsymbol{u}}{\partial t^2} &= (\lambda + 2G)\nabla(\nabla \cdot \boldsymbol{u}) \\ &\quad - G\nabla \times (\nabla \times \boldsymbol{u}) \qquad (7.9)\end{aligned}$$

where ∇ is the vector operator $\nabla = (\partial/\partial a_1)\boldsymbol{e}_1 + (\partial/\partial a_2)\boldsymbol{e}_2 + (\partial/\partial a_3)\boldsymbol{e}_3$ and $\boldsymbol{e}_i$ are the unit vectors along the Cartesian axes. For the case of pure dilatational (compressional or longitudinal) waves, $\nabla \bullet \boldsymbol{u} \neq 0, \nabla \times \boldsymbol{u} = 0$ and Equation (7.9) reduces to $(\partial^2 \boldsymbol{u}/\partial t^2) = c_l \nabla^2 \boldsymbol{u}$, where $c_1 = [(\lambda + 2G)/\rho_0]^{1/2}$ is the longitudinal wave velocity. For the case of pure rotational (shear or transverse) waves, $\nabla \bullet \boldsymbol{u} = 0$, $\nabla \times \boldsymbol{u} \neq 0$ and Equation (7.9) reduces to $(\partial^2 \boldsymbol{u}/\partial t^2) = c_t \nabla^2 \boldsymbol{u}$, where $c_2 = c_3 = (G/\rho_0)^{1/2}$ is the shear wave velocity.

7.2.2
Fluids

We define a fluid as a material that will yield under the action of a shear stress, no matter how small it is. A fluid is ordinarily understood to be isotropic. We define a perfect fluid to be an idealized material in which there never exists a shear stress. Hence, the shear modulus G must be zero. The wave equation, Equation (7.7) or Equation (7.9), also holds for fluids and the results obtained for isotropic solids can be directly used for fluids with appropriate consideration of the isotropic symmetry and the vanishing of the shear modulus. Thus, the transverse wave velocity $c_2 = c_3 = 0$ and the longitudinal (also called *compressional*) wave velocity $c_1 = [B/\rho_0]^{1/2}$, where B is the bulk modulus.

It is instructive to consider an alternative form of the wave equation often used for fluid media. For a frictionless or nonviscous perfect fluid, the stress σ_{ij} is the thermodynamic pressure p obtained as $\sigma_{ij} = -p\ \delta_{ij}$. For small variations in the mass density ρ about the initial mass density ρ_0, the equation of state, Equation (7.6), may be approximated using the isotropic symmetry properties of the elastic constants and Equation (7.1) as $\rho = \rho_0 + (\rho_0/B_S)(p - p_0) + \ldots$. Substituting this equation into the equation of continuity $(\partial\rho/\partial t) = \partial(\rho v_i)/\partial a_i \approx \rho_0(\partial v_i/\partial a_i)$, where $v_i = \partial u_i/\partial t$ is the particle velocity, we obtain $(\partial p/\partial t) = -B_S(\partial v_i/\partial a_i)$. Taking the derivative of this expression with respect to time and using Newton's Law $\rho_0(\partial^2 u_i/\partial t^2) = \rho_0(\partial v_i/\partial t) = (\partial\sigma_{ij}/\partial a_j) = -(\partial p/\partial a_i)$, we obtain the pressure wave equation

$$\frac{\partial^2 p}{\partial t^2} = \frac{B_S}{\rho_0}\frac{\partial^2 p}{\partial a_i \partial a_i} = \frac{B_S}{\rho_0}\nabla^2 p \tag{7.10}$$

where ∇^2 is the Laplacian operator. It is apparent from Equation (7.10) that the compressional wave velocity $c_1 = (B_S/\rho_0)^{1/2}$.

The isentropic equation of state used in obtaining Equation (7.10) may be recast in the more traditional Hooke's Law form $p = p(\rho, S) = p_0 + (\partial p/\partial\rho)_S(\rho - \rho_o) + \ldots = p_0 + (B_S/\rho_0)(\rho - \rho_o) + \ldots$. This form reveals that the isentropic wave velocity for fluids, $c = (B_S/\rho_0)^{1/2}$, can be generally obtained from the equation of state as $c = [(\partial p/\partial\rho)\rho_0]^{1/2}$. For the special case of an ideal gas, the adiabatic condition is expressed as $p = p_0(\rho/\rho_0)^\gamma$, where γ is the ratio of isobaric to isochoric specific heats. Hence, the adiabatic sound velocity is obtained as $c = [\gamma(p_0/\rho_0)]^{1/2}$. The ideal gas equation of state involving temperature T may be written as $p = p(\rho, T) = \rho RT$, where $R = R_0/M$, R_0 is the universal gas constant, and M is the average molecular weight of the molecules comprising the gas. Thus, we find the temperature dependence of the adiabatic sound velocity to be $c = (\gamma RT)^{1/2}$.

7.3
Ultrasonic Waves in Bounded Media

The presence of surfaces and interfaces bounding two media having different acoustic properties gives rise to a variety of acoustic phenomena having important applications. We consider here conditions leading to wave reflection, refraction, acoustic mode conversion, inhomogeneous and evanescent waves, and guided waves.

7.3.1
Reflection and Refraction

We consider a plane harmonic pressure wave of unit amplitude of the form $p_{inc} = \exp[i(\boldsymbol{k}\bullet\boldsymbol{x} - \omega t)]$ incident on a plane boundary between two liquid or gas media as illustrated in Figure 7.1. We orient and center the coordinate axes such that the planar interface occurs at $z = 0$. We assume that the plane of the incident wave is coincident with the xz-plane such that $k_x = k\sin\theta$, $k_y = 0$, $k_z = k\cos\theta$, and $k = \omega/c$, where c is the sound velocity in the incident wave propagation medium (upper medium in figure) and θ is the angle of incidence formed by the normal to the wave front and the z axis. The density of

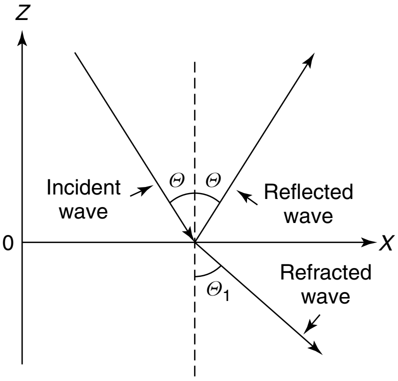

Fig. 7.1 Reflection and refraction of an incident sound wave.

the upper medium is ρ and the density and the sound velocity of the lower medium are ρ_1 and c_1, respectively.

In general, we expect that the incident wave of unit amplitude will be partially reflected from the interface with some amplitude R called the *stress (pressure) reflection coefficient*. The expressions for the incident and reflected waves are written respectively as $p_{inc} = \exp[ik(x\sin\theta - z\cos\theta)]\exp(-i\omega t)$ and $p_{refl} = R\exp[ik(x\sin\theta + z\cos\theta)]\exp(-i\omega t)$. Note that we assume specular (geometrical) reflection of the reflected wave such that the angle of incidence is equal to the angle of reflection. The total pressure field in the upper medium is then

$$p = p_{\mathrm{inc}} + p_{\mathrm{refl}} = (e^{-ikz\cos\theta} + Re^{ikz\cos\theta}) \times e^{i(kx\sin\theta - \omega t)} \tag{7.11}$$

The remaining energy of the incident wave will be transmitted into the lower medium with an amplitude W called the *stress (pressure) amplitude transmission coefficient*. The pressure field in the lower medium consists only of the transmitted wave, which we write as

$$p_1 = We^{ik_1(x\sin\theta_1 - z\cos\theta_1) - i\omega t} \tag{7.12}$$

where $k_1 = \omega/c_1$ is the wavenumber in the lower medium. The angle of refraction θ_1 does not generally equal the incident angle θ, indicating that the wave in the lower medium (the refracted wave) deviates from the incident wave direction.

The quantities R, W, and θ_1 are determined from the conditions that the pressure and the normal component of the particle velocity $v_z = \partial u_z/\partial t$ must be continuous across the boundary. In acoustics, often analogies between acoustic and electromagnetic quantities are used. The acoustic pressure can be identified with the electromotive force or voltage and the particle velocity with the electrical current (Pierce, 1989). Thus, in analogy with the concept of electrical impedance (ratio of voltage to current), we define the acoustic impedance Z normal to the surface as

$$Z = -\frac{p}{v_z} \tag{7.13}$$

The boundary conditions can now be expressed equivalently as the continuity of pressure and acoustic impedance at $z = 0$.

The continuity of pressure requires that $p = p_1$ at $z = 0$. From Equations (7.11) and (7.12), we find that $1 + R = W\exp[i(k_1\sin\theta_1 - k\sin\theta)x]$. Because the left side of this equation is independent of x, so must be the right side. We thus obtain

Snell's law

$$\frac{\sin\theta}{\sin\theta_1} = \frac{k_1}{k} = \frac{c}{c_1} = n \qquad (7.14)$$

where n is the index of refraction. Thus, we have $1 + R = W$.

The continuity of particle velocity may be replaced by the continuity of acoustic impedance. In the lower medium, the impedance $Z_1 = -i\omega\rho_1 p_1(\partial p_1/\partial z)^{-1} = \rho_1 c_1/\cos\theta_1$; in the upper medium $Z = (\rho c/\cos\theta)\,[\exp(-ikz\cos\theta) + R\exp(ikz\cos\theta)]/[\exp(-ikz\cos\theta) - R\exp(ikz\cos\theta)]$. Because $Z = Z_1$ at $z = 0$, we obtain from the above equations that

$$R = \frac{Z_1 - Z_{\text{inc}}}{Z_1 + Z_{\text{inc}}},\ Z_{\text{inc}} = \frac{\rho c}{\cos\theta} \qquad (7.15)$$

For fluid–solid and solid–solid interfaces, the number of reflected and refracted waves is dramatically increased, because both longitudinal and shear wave motion must be considered. Correspondingly, the reflection and transmission coefficients must include shear wave impedance terms as well as longitudinal impedance terms. For further information, see Brekhovskikh (1980).

7.3.2
Resonance Phenomena

Consider the one-dimensional case of a sample of length $L/2$ having flat and parallel opposing faces. We assume complete reflection of the sound waves normally incident on the surfaces. We assume that the face $x = 0$ is caused to vibrate with a time harmonic disturbance having particle velocity $v = \partial u/\partial t = \cos(\omega t)$. The disturbance propagates from the surface as a plane wave along the direction x described by the complex expression $v' = \exp[i(\omega t - kx) - \alpha x]$. The factor $\exp(-\alpha x)$ represents the decay of the waveform as the result of attenuation in the medium (see below); the parameter α is the attenuation coefficient. We assume that the wave is totally reflected at the opposing surface $x = L/2$ and the reflected wave after propagating back to the surface $x = 0$ is superimposed on the original driving vibration. The wave continues to reflect between the opposing surfaces giving rise to a superposition of waves of the form (Bolef and Miller, 1971)

$$\begin{aligned} v' &= \exp(i\omega t)(1 + \exp[-(ikL + \alpha L)] \\ &\quad + \exp[-2(ikL + \alpha L)] + \ldots) \\ &= \exp(i\omega t)/(1 + \exp[-(ikL + \alpha L)] \end{aligned} \qquad (7.16)$$

Taking the real part of the above expression, we obtain

$$v = \operatorname{Re} v' = v_1\cos\omega t + v_2\sin\omega t \qquad (7.17)$$

where

$$v_1 = \frac{e^{\alpha L} - \cos kL}{2(\cosh\alpha L - \cos kL)},$$
$$v_2 = \frac{\sin kL}{2(\cosh\alpha L - \cos kL)} \qquad (7.18)$$

The term v_1 is in phase with the driving vibration and the term v_2 is 90° out of phase (in quadrature). The amplitude of v is

$$\begin{aligned} |v| &= (v_1^2 + v_2^2)^{1/2} \\ &= \frac{e^{\alpha L/2}}{[2(\cosh\alpha L - \cos kL)]^{1/2}} \end{aligned} \qquad (7.19)$$

Plots of these terms as a function of $kL (= \omega L/c)$ are given in Figure 7.2. The curves show a resonance behavior at a value of kL corresponding to the condition that the sample length $L/2$ is equal to an integral number of half wavelengths (wavelength $\lambda = 2\pi c/\omega$). Such standing wave resonances occur at all frequencies

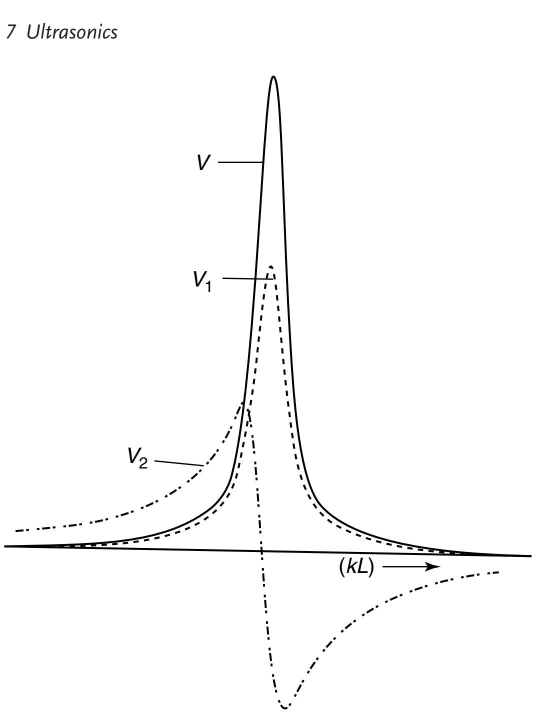

Fig. 7.2 Plots of ν (solid curve), ν_1 (dashed curve), and ν_2 (dashed-dotted curve) as functions of kL about a resonance value.

ω_m such that the resonance condition $L/2 = m(\lambda_{m/2})$ is satisfied, where m is an integer (harmonic number) and $\lambda_m = 2\pi c/\omega_m$. Using $k = \omega/c$, we obtain from the resonance condition that the mth resonance frequency is $\omega_m = 2\pi mc/L$ and that $\omega_m - \omega_{m+1} = 2\pi c/L$. The latter expression can be used to determine the ultrasonic velocity from the measurement of two successive resonance frequencies. The attenuation can be determined from the frequency width $\Delta\omega$ of the resonance at half power as $\alpha = \Delta\omega/(2c)$ (Bolef and Miller, 1971).

7.3.3
Surface and Interface Waves

We consider the situation at a fluid–fluid interface where the index of refraction $n < 1 (c_1 > c)$ and the angle of incidence of a plane wave exceed the critical angle $\theta_0 = \arcsin n$. For $\theta > \theta_0$, we rewrite Equation (7.15) in the complex form $R = [(\rho_1/\rho) \cos\theta - i(\sin^2\theta - n^2)^{1/2}]/[(\rho_1/\rho)\cos\theta + \mathrm{i}(\sin^2\theta - n^2)^{1/2}]$. Because $|R| = 1$, there is total reflection of the wave. Writing $R = \exp(i\phi)$, we find that the phase ϕ of the reflected wave varies, $\phi = -2\ \arctan[(\rho(\sin^2\theta - n^2)^{1/2}/\rho_1\cos\theta)]$. For the case of complete internal reflection, the field in the lower medium can be written as (Brekhovskikh, 1980)

$$z < 0,\ p_1 = (1 + R)\mathrm{e}^{\delta z}\mathrm{e}^{\mathrm{i}kx\sin\theta - \mathrm{i}\omega t},$$
$$\delta = k(\sin^2\theta - n^2)^{1/2} \qquad (7.20)$$

The wave represented by Equation (7.20) is an example of an inhomogeneous wave. Such waves propagate along a given direction with an amplitude that decays exponentially in a direction perpendicular to the direction of propagation. The wave of Equation (7.20) propagates along the boundary (a surface wave) with exponential decay perpendicular to the boundary.

Inhomogeneous waves can propagate on a boundary in the absence of an incident wave. Consider the case of an isotropic elastic halfspace that forms a planar interface with a vacuum on the $z = 0$ plane (we assume same axes orientation and centering as in Figure 7.1). A monochromatic wave propagating along the solid–vacuum surface (x direction) must satisfy the wave equation in Equation (7.7). In analogy with Equation (7.20), we assume that the form of the surface wave is given by $u_i = \exp(\delta z)\exp[i(kx - \omega t)]$. From Equation (7.14) (Snell's law) and the relations $k_1 = \omega/c_1, k_{1x} = k_1 \sin\theta_1$, we may recast the δ given in Equation (7.20) in the form $\delta = [k_{1x}^2 - (\omega/c_1)^2]^{1/2}$. We assume for the present case that δ can be written in the analogous form $\delta = [k^2 - (\omega/c)^2]^{1/2}$ where the subscripts have been dropped.

Let us resolve the displacement $\boldsymbol{u}$ (or u_i) into a longitudinal displacement $\boldsymbol{u}_l$ (or u_{li}) and a transverse displacement $\boldsymbol{u}_t$ (or u_{ti}) as $\boldsymbol{u} = \boldsymbol{u}_l + \boldsymbol{u}_t$(or $u_i = u_{l,I} + u_{t,i}$). The transverse component must satisfy the condition $\boldsymbol{\nabla}\bullet\boldsymbol{u}_t = (\partial u_{t,1}/\partial x) + (\partial u_{t,3}/\partial z) = 0$. We assume $u_{t,I} = \exp(\delta_t z)\exp[i(kx - \omega t)]$, where $\delta_t = [k^2 - (\omega/c_t)^2]^{1/2}$ and c_t is the shear wave velocity. We find from these relations that $u_{t,1} = \delta_t A_1 \exp(\delta_t z)\exp[i(kx - \omega t)]$ and $u_{t,3} = -ikA_t \exp(\delta_t z)\exp[i(kx - \omega t)]$, where A_t is a constant (Pollard, 1977).

The longitudinal part of the displacement satisfies the condition $\boldsymbol{\nabla} \times \boldsymbol{u}_l = (\partial u_{l1}/\partial z) - (\partial u_{l,3}/\partial x) = 0$. Performing an analysis similar to that for the transverse components above, we find that $u_{l,1} = kA_l \exp(\delta_l z)\exp[i(kx - \omega t)]$ and $u_{l,3} = -i\delta_l A_l \exp(\delta_l z)\exp[i(kx - \omega t)]$, where A_l is a constant. The boundary conditions to be satisfied for a free surface are $\sigma_{ij} n_j = 0$, where n_j are the components of the unit vector $\boldsymbol{n}$ normal to the surface (parallel to z axis). Application of these boundary conditions provides the relationship $2ik\delta_l A_l + (k^2 + \delta_t^2)A_t = 0$. We obtain from these expressions that the component of the displacement parallel to the surface $u_1 = u_{l,1} + u_{t,1}$ and perpendicular to the surface $u_3 = u_{l,3} + u_{t,3}$ are (in real form) (Briggs, 1992)

$$u_1 = A_2 k\left(e^{-\delta_l z} - \frac{2\delta_l\delta_t}{\delta_t^2 + k^2}e^{-\delta_t z}\right)\sin(kx - \omega t)$$

$$u_3 = A_2 \delta_l\left(e^{-\delta_l z} - \frac{2k^2}{\delta_t^2 + k^2}e^{-\delta_t z}\right)\cos(kx - \omega t) \qquad (7.21)$$

and that the wave velocity $c_R = \omega/k$ is given by

$$16G^2\left(1 - \frac{c_R^2}{c_l^2}\right)\left(1 - \frac{c_R^2}{c_t^2}\right) = \left[\frac{4G}{3} - \left(\lambda + \frac{4G}{3}\right)\frac{c_R^2}{c_l^2}\right]^2\left[2 - \frac{c_R^2}{c_l^2}\right]^2 \qquad (7.22)$$

According to Equation (7.21), both the longitudinal and transverse components decay exponentially in the direction z perpendicular to the propagation direction x. The wave represented by Equation (7.21) is thus a type of inhomogeneous wave known as the *Rayleigh wave*. An approximate solution to Equation (7.22) yields the expression (Scruby, Jones, and Anatoniazzi, 1987) $c_R \approx c_t(1.14418 - 0.25771\mu_P + 0.12661\mu_P^2)^{-1}$ for the Rayleigh wave velocity, where $\mu_P = [1 - 2(c_t/c_l)^2]/\{2[1 - (c_t/c_l)^2]\}$ is Poisson's ratio for the material (see Section 7.3.4).

Figure 7.3 shows the particle displacement field given by Equation (7.21) for a Rayleigh surface wave. The motion is generally elliptical with the major axis perpendicular to the surface. The motion reverses the sense of rotation below a depth of approximately 0.2λ.

If the solid surface on which the Rayleigh wave propagates is in contact with a fluid medium, the velocity of the Rayleigh wave

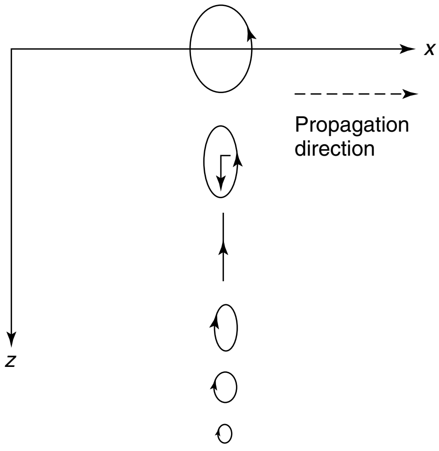

Fig. 7.3 Particle displacement field for a Rayleigh surface wave. The arrows indicate sense of particle motion. From Sabine and Cole, 1971.

will be slightly altered as a result of the reaction of the fluid medium. If the velocity of the Rayleigh wave is greater than the velocity of sound in the fluid medium, then the wave will be partially radiated into the fluid medium. The Rayleigh wave in this case belongs to a class of surface acoustic waves (SAWs) called "*leaky*" waves.

Many other types of surface waves can propagate in materials that depend on the properties of the materials forming the interface. The interface between two elastic media gives rise to an interface wave called the *Stonely wave*. The field distribution in each medium is that of two partial waves with particle displacements parallel and normal to the interface, the motion on either side of the interface being of the Rayleigh type. The Stonely wave velocity lies between that of the Rayleigh waves in each medium and the shear wave in the denser medium.

Other interface waves include the Love wave and the lateral wave. The Love wave involves solid–solid interfaces and is characterized by a transverse component parallel to the surface with particle displacement perpendicular to the direction of wave propagation. The lateral wave may occur at various interface combinations (solid–solid, fluid–solid, fluid–fluid) and is generally associated with the reflection of spherical waves. See Brekhovskikh (1980) for more information.

7.3.4
Guided Waves

Guided waves are characterized by energy flow primarily along the direction of some guiding configuration. Strictly, SAWs are guided waves, but we concern ourselves here with propagation in configurations such as cylinders and plates. These structures are capable of supporting an infinite number of waveguide modes. For simple configurations, the modes may be analyzed as composed of traveling waves along the waveguide axis and standing waves in the transverse direction (the transverse resonance principle). A general expression for the propagation of a sound wave along the direction x in a medium bounded in the (y, z) plane can written as an expansion in terms of the complete set of eigenfunctions $f_n(y, z)$ arising from the solution of the three-dimensional wave equation subject to transverse boundary conditions (i.e., boundary conditioned referred to the y and z dimensions) (Morse and Ingard, 1968). We thus write

$$v(x, y, z, t) = \sum_n b_n f_n(y, z) \mathrm{e}^{-\alpha_n x} e^{\mathrm{i}(\omega t - k_n x)} \tag{7.23}$$

where the constant $k_n = [(\omega/c)^2 - K_n^2]^{1/2}$ is the propagation constant for the nth mode and K_n is the eigenvalue corresponding to the eigenfunction $f_n(y, z)$, b_n is an expansion coefficient, and a_n is the attenuation coefficient for the nth mode.

If a solution corresponding to the eigenvalue $K_n = 0$ exists, it is called the *fundamental mode* and corresponds to one-dimensional wave propagation along the x direction. Higher order modes corresponding to $K_n^2 > 0$ reflect from the sides of the waveguide in a zigzag fashion but have a net propagation in the x direction. For a given frequency, eigensolutions exist only if $(\omega/c)^2 > K_n^2$. A given mode thus has a frequency cutoff below which the mode does not propagate. Such nonpropagating disturbances that are periodic only in time and decay exponentially in all spatial coordinates are called *evanescent waves.*

For a solid isotropic cylinder, the simplest modal solution may be obtained by considering the solid to be subjected to a stress σ_{11} along the cylindrical axis x and stress free conditions on the sides of the cylinder. The application of Hooke's law for an isotropic solid to this case yields the equations: $\sigma_{11} = 2G\varepsilon_{11} + \lambda\varphi; 0 = 2G\varepsilon_{22} + \lambda\varphi; 0 = 2G\varepsilon_{33} + \lambda\varphi$, where the dilatation $\varphi = \varepsilon_{11} + \varepsilon_{22} + \varepsilon_{33}$. From these equations, we define and obtain Young's modulus as $E = \sigma_{11}/\varepsilon_{11} = G(3\lambda + 2G)/(\lambda + G)$ and Poisson's ratio $\mu_p = -\varepsilon_{22}/\varepsilon_{11} = \lambda/[2(\lambda + G)]$. Substituting the defining relation for Young's modulus into Newton's law gives the wave equation $\rho_0(\partial^2 u_1/\partial t^2) = E(\partial^2 u_1/\partial a_1^2)$. We identify from this wave equation that $c_L = (E/\rho_0)^{1/2}$ is the velocity of propagation of bounded longitudinal waves in an isolated solid cylinder. This velocity is in general less than that of a plane longitudinal wave propagating in an unbounded solid. Rayleigh obtained a correction to c_L by accounting for "lateral inertia" and obtained the longitudinal phase velocity $c_L' = c_L[1 - \mu_p^2\pi^2(r_0/\lambda)^2]$, where r_0 is the radius of the cylinder. More general solutions require more powerful approaches such as that leading to Equation (7.23) or methods employing Green's functions. For more information, see Auld (1990).

For a free isotropic plate, horizontally polarized shear (SH) wave modes (characterized by having a polarization direction parallel to the surface) and Lamb wave modes (characterized by a coupling of vertically polarized shear waves and longitudinal waves) may propagate. In the limit $kL \to \infty$, where L is the thickness of the plate, the fundamental symmetrical and antisymmetrical Lamb wave modes degenerate into Rayleigh waves bound to the surfaces. An isotropic plate on an isotropic elastic substrate gives rise to generalized Lamb waves which yield tightly bound Stonely waves at the plate-substrate interface in the limit $kL \to \infty$. For further information on guided waves, see Brekhovskikh (1980) and Auld (1990).

7.4
Ultrasonic Attenuation

Attenuation refers to the total loss of amplitude (or energy) of a sound wave along its propagation path resulting from all mechanisms responsible for such losses, including but not limited to absorption, scattering, and diffraction. Absorption refers to the loss of amplitude (or energy) of an acoustic wave that results in an increase in temperature (however slight) in the propagation medium and is often associated with relaxation phenomena. Scattering results in a redirection of wave energy from the original propagation direction. Diffraction refers to scattering that occurs from the boundaries or edges of objects. Generally, attenuation produces an exponential decrease in the wave amplitude. For plane wave propagation along the x direction, we may write

$$u(x, t) = u_0 e^{-\alpha x} e^{i(kx-\omega t)} \tag{7.24}$$

where α is the attenuation coefficient. It is generally assumed that the contributions from all sources to the attenuation is additive such that $\alpha = \alpha_{\mathrm{abs}} + \alpha_{\mathrm{sc}} + \alpha_{\mathrm{diff}} \ldots$ where the subscripts refer to the absorption, scattering, and diffraction contributions, respectively.

7.4.1
Scattering

Generally, scattering refers to the redirection of wave energy from the original propagation direction or equivalently to the deviation of a ray or pencil of ultrasound from a straight path. Scattering may be classified into geometric contributions and diffraction-based contributions (Überall, 1997). Geometric contributions include specular reflections, transmission, refraction, and internal reflections of ultrasonic waves. Diffraction-based contributions refer to phenomena whereby an inhomogeneity in the path of a sound wave produces secondary waves spreading out from the inhomogeneity in various directions. Generally, an inhomogeneity refers to any object or region in the wave propagation path that exhibits a difference in density or elastic properties from that of the incident wave propagation medium. The object could be, for example, a defect in a crystal, a region of atmospheric turbulence, a fish in the ocean, or a tumor in the brain. Geometric scattering was considered in Section 7.3.1. We consider here general diffraction-based scattering. Diffraction is treated in Section 7.4.2.

If the region of inhomogeneity (having characteristic dimensions of order d) is small compared to the wavelength λ of the incident wave (i.e., $kd \ll 1$), the scattering is called *Rayleigh scattering*. The general features of Rayleigh scattering may be obtained from the consideration of constant frequency plane waves propagating along the direction $\boldsymbol{e}_k = \boldsymbol{k}/|\boldsymbol{k}|$ ($\boldsymbol{k}$ is the wave number vector) that impinge on a rigid immovable object centered at the origin of the coordinate system. The total acoustic pressure is $p = B\ \exp(i\boldsymbol{k}\bullet\boldsymbol{x}) + p_{\mathrm{sc}}(\boldsymbol{x})$, where B is the amplitude of the incident wave and $p_{\mathrm{sc}}(\boldsymbol{x})$ is the complex amplitude of the scattered wave. A rigid scattering surface imposes the requirement that $\nabla p\bullet\boldsymbol{n} = 0$ at the surface S of the scatterer, where $\boldsymbol{n}$ is the unit normal pointing outward from the surface (Pierce, 1989). Substituting the total acoustic pressure equation into the boundary condition gives $\nabla p_{\mathrm{sc}}\bullet\boldsymbol{n} = -iB[\exp(i\boldsymbol{k}\bullet\boldsymbol{x})](\boldsymbol{k}\bullet\boldsymbol{n})$. Thus, the scattered wave field may be viewed as that of a radiating body of the same size and shape as that of the scatterer with an amplitude that is the negative of that associated with the incident wave field.

For a small wavenumber, we may expand the exponential in the above equation in a power series to first order in k and integrate the results over the surface of the scatterer. The procedure yields the results (Pierce, 1989) $p_{\mathrm{sc}} = p_{\mathrm{sc}}$, mono $+p_{\mathrm{sc}}$, dipole. The term $p_{\mathrm{sc,mono}} = -(k^2 BV/4\pi)[\exp(ikr)/r]$, where r is the radial coordinate and V is the scatterer volume of order d^3, is due to monopole radiation. The term $p_{\mathrm{sc,dipole}} = (k^2 B/4\pi)\{\boldsymbol{e}_r \bullet \boldsymbol{M} \bullet \boldsymbol{e}_k[(1+(i/kr)]\}[\exp(ikr)/r]$, is due to dipole radiation. The factor $\boldsymbol{M}$ is a tensor, which scales as d^3, that is dependent on the scatterer size and shape. The sum of the monopole and the dipole terms produces a far-field scattered wave amplitude that scales as $(k^2 d^3/r)$. Thus, both the monopole and the dipole radiations are dependent on the square of the wavenumber (or frequency). Because the intensity of the scattered sound wave is proportional to the square of the pressure, the intensity of the scattered wave field is then dependent on the fourth power of the sound frequency. This strong dependence on the frequency in Rayleigh scattering has an analog in optics and leads to an explanation of why the sky is blue: higher frequency blue light scatters more than lower frequency red light.

In the limit $kd \gg 1$ (high frequencies), geometric scattering dominates and the scattering is more appropriately calculated in terms of impedance ratios described in Section 7.3. The transition region between the low- and the high-frequency limits is not easily assessed in terms of simple expressions. For wavelengths of the same magnitude as the dimensions of the scatterer, the scattered waves form complex patterns that must be calculated numerically. Figure 7.4 shows the angular distribution of scattered energy for waves incident from the left onto a sphere of radius d for $kd \ll 1$ (Rayleigh scattering), $kd = 6$, and $kd \to \infty$.

Relative motion between the scattering object (viewed as a source of sound) and the sound receiver causes a shift in the frequency of sound received relative to that of the source. This change in frequency is called the *Doppler shift* and can be understood by reference to Figure 7.5. We consider a sound source fixed in reference frame 2 moving with a velocity $\boldsymbol{V}_{21}$ relative to a sound receiver fixed in reference frame 1. A point vector $\boldsymbol{x}_1$ referred to frame 1 is related to a point vector $\boldsymbol{x}_2$ referred to frame 2 by the Galilean transformation $\boldsymbol{x}_1 = \mathbf{x}_2 + (t - t_0)\boldsymbol{V}_{21}$, where at $t = t_0, \boldsymbol{x}_1 = \boldsymbol{x}_2$. We describe the phases of the same sound disturbance in each of the two reference frames by $\phi_1(\boldsymbol{x}_1, t)$ and $\phi_2(\boldsymbol{x}_2, t)$. The wave vector $\boldsymbol{k}$ in either frame is defined by $\boldsymbol{k} = -\boldsymbol{\nabla}\phi$ and the angular frequency $\omega = \partial\phi/\partial t$. For the same sound disturbance, we must have $\phi_1(\boldsymbol{x}_1, t) = \phi_1(\boldsymbol{x}_2 + (t - t_0)\boldsymbol{V}_{21}, t) = \phi_2(\boldsymbol{x}_2, t)$. Differentiation of this expression with respect to the spatial and time coordinates, respectively, yields the relations $\boldsymbol{k}_1 = \boldsymbol{k}_2$ and $\omega_2 = \omega_1 - \boldsymbol{k}_1 \bullet \boldsymbol{V}_{21}$.

We first consider the case where reference frame 1 (with fixed receiver) is stationary relative to the sound propagation medium and frame 2 (with fixed source) is moving through the propagation medium. Assuming that the source generates sound at frequency ω_0, we may write $\omega_2 = \omega_0$. Because the receiver is stationary in the propagation medium, we may also write that $\boldsymbol{k}_1 = (\omega/c)\boldsymbol{e}_R$, where c is the sound velocity and $\boldsymbol{e}_R$ is the unit vector determined at the instant of sound emission along the direction from the sound source to the receiver. We thus find $\omega_2 = \omega_0 = \omega_1 - \boldsymbol{k}_1 \bullet \boldsymbol{V}_{21} = \omega - (\omega/c)\boldsymbol{e}_R \bullet \boldsymbol{V}_{21}$. Solving this equation for ω, we obtain $\omega = \omega_0(1 - c^{-1}\boldsymbol{e}_R \bullet \boldsymbol{V}_{21})^{-1}$.

We now consider the case where reference frame 1 (with fixed receiver) is moving in the propagation medium and frame 2 (fixed sound source) is stationary in the propagation medium. Then we may write $\omega_2 = \omega_0$ and $\boldsymbol{k}_1 = \boldsymbol{k}_2 = (\omega_0/c)\boldsymbol{e}_R$. Thus $\omega_2 = \omega_0 = \omega - (\omega_0/c)\boldsymbol{e}_R \bullet \boldsymbol{V}_{21}$. Solving for ω, we obtain $\omega = \omega_0(1 + c^{-1}\boldsymbol{e}_R \bullet \boldsymbol{V}_{21})$. In each case, we find that the frequency

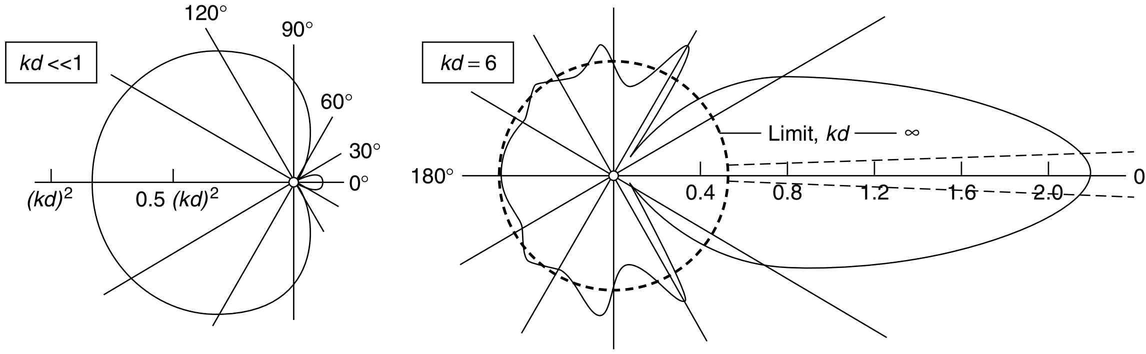

Fig. 7.4 Angular distribution of scattered wave energy for waves incident from the left onto a sphere of radius d for $kd \ll 1, kd = 6$, and $kd \to \infty$ (dotted curves). From Pierce, 1989.

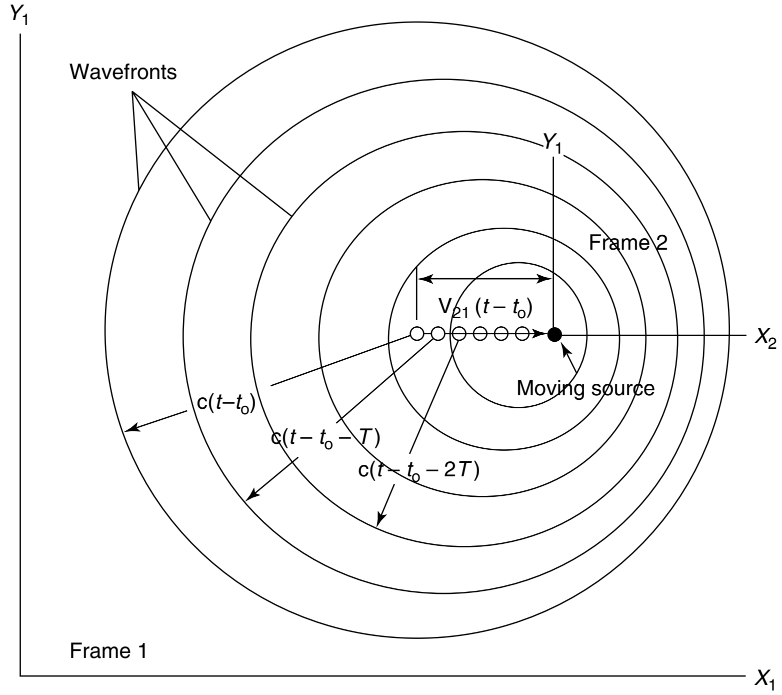

Fig. 7.5 Schematic of Doppler shifted crests of waveforms emitted with period T by source fixed in reference frame 2 moving at velocity V_{21} with respect to the propagation medium fixed in reference frame 1. Adapted from Pierce, 1989.

increases if the source and receiver are moving toward each other and decreases if the source and receiver are separating. The magnitude of frequency shift is determined by the component of the relative velocity along $\boldsymbol{e}_R$.

In many applications, a transmitting transducer fixed in a propagation medium (reference frame 1) is used to transmit sound to a scatterer fixed in a second propagation medium (reference frame 2) moving relative to reference frame 1 with velocity $\boldsymbol{V}_{21}$. A receiving transducer also fixed in frame 1 is then used to receive the scattered sound waves. The boundary between propagation media 1 and 2 is assumed to occupy well-defined instantaneous positions with respect to frame 1. With respect to the scatterer fixed in frame 2, the sound waves incident and scattered at the scatterer have the same frequency ω_2. However, with respect to frame 1, the transmitter frequency is $\omega_{\mathrm{tr}} = \omega_2 + \boldsymbol{V}_{21}\bullet\boldsymbol{k}_{1\mathrm{tr}}$ and the receiver frequency is $\omega_{\mathrm{rec}} = \omega_2 + \boldsymbol{V}_{21}\bullet\boldsymbol{k}_{1\mathrm{rec}}$, where $\boldsymbol{k}_{1tr}$ and $\boldsymbol{k}_{1\mathrm{rec}}$ are the transmitted and received wave vectors referred to frame 1. Eliminating ω_2 from these two equations yields the frequency shift between the transmitted and received signals as $(\omega_{\mathrm{rec}} - \omega_{\mathrm{tr}}) = \boldsymbol{V}_{21})\bullet\boldsymbol{k}_{1rec} - \boldsymbol{k}_{1\mathrm{tr}})$. Typical applications of Doppler frequency shift measurements include the assessment of flow velocities and fluid volume flowing per unit time in blood vessels ducts and tubes, the remote sensing of tropospheric wind velocities, and the measurement of subsurface ocean current velocities.

7.4.2
Diffraction

Generally, diffraction refers to scattering that occurs at the "edges" of objects. It also includes the influence of boundaries on radiation sources such as piston transducers. Mathematically, diffraction phenomena can be described in terms of monopole and dipole radiation patterns as is done with diffraction-based scattering phenomena generally (see Section 7.4.1). One of the most important applications of diffraction theory in ultrasonics is to account for beam spreading from circular transducers, the most common shape for ultrasonic sources.

Because the transducer is of finite size, the acoustic wave generated in the sample spreads out into a diffraction field that leads to an attenuation of the wave along the original propagation direction and to errors in velocity and source amplitude measurements if not corrected. The effects of diffraction are usually treated for the case of circular, axially concentric transmitting and receiving transducers of the same radius r. The transmitting transducer is treated as a finite piston source in a semi-infinite medium. The acoustic field is found at each point in the propagation medium and an integration is performed over the area in the field presented by the receiving transducer. In one treatment (Benson and Kiyohara, 1974), the effects of diffraction can be represented for waves of wavelength λ propagating along the z direction in an isotropic medium by the expression $u(z, t) = 2(I_x^2 + I_y^2)^{1/2}\cos[kc_p t - kz - \phi]$, where $\phi = \tan(I_y/I_x)$ and the quantities Ix and Iy are integral functions of the dimensionless parameter $S = z\lambda/r^2$.

These equations show that the diffraction involves an intensity variation in the diffraction pattern and a distortion of the otherwise planar wave front. Phase variations in the surface plane of the receiving transducer give rise to interference effects and a loss in signal (attenuation) when integrated over a phase-sensitive transducer surface. The factor ϕ in the above equation produces a variation in the phase of the propagating wave that results in an error in the velocity measurements, if not properly corrected.

A useful aspect of diffraction occurs for the focusing of sound fields from curved transducers. The field of a focused sound beam has diffraction characteristics similar to that of a plane piston transducer. Along the axis of the curved transducer, the intensity of the sound field goes through a series of maxima and minima in accordance with the interference patterns produced by diffraction. The maximum intensity occurs at an interference maximum corresponding to the focal zone of the transducer.

7.4.3
Relaxation Processes

The propagation of sound waves based on Hooke's Law is such that the stress (or pressure) at any point in the body is a linear function only of the instantaneous strain (or mass density) at that point. Upon removal of the stress, the strain fully and instantly recovers. The reversibility of the elastic deformation means that no dissipation or loss of energy in the wave occurs and that the wave propagates indefinitely. However, wave propagation in real materials always exhibits an amplitude loss resulting from the conversion of acoustic energy into other forms. Such conversions may involve, for example, thermal motion, the vibration or rotation of atoms in a molecule, structural rearrangements (including chemical reactions and electrolytic processes), or the motion of lattice constituents such as dislocations, point defects, and so on. The energy dissipation destroys the isentropic character of the sound propagation (although it may still be adiabatic) and requires the formalism of irreversible thermodynamics for a proper description.

7.4.3.1 **Relaxation in Solids**

We consider a process described by the reaction (Beyer and Letcher, 1969) $n_1 + n_2 + \ldots + n_k \leftrightarrow n_{k+1} + n_{k+2} + \ldots + n_m$, where $n_i (i \le k)$ is the number of moles of the ith reactant and $n_i (i > k)$ is the number of moles of the ith product involved in the reaction. The reaction process is driven by the acoustic wave and may describe, for example, the excitation of higher vibrational or rotational states of a gas or the jumping of interstitial atoms in a solid lattice to a new equilibrium site. We account for such reactions by modifying the internal and Helmholtz free energies per unit mass to include a dimensionless quantity ξ describing the extent to which the reaction has taken place. This variable, called the *degree of reaction*, ranges from a value of 0 when the state of the system is all reactant to a value of 1 when the state of the system is all product. In analogy to Equation (7.4), we may expand the modified internal energy per unit mass $U(\boldsymbol{a}, \eta_{ij}, S, \xi)$ to first order in all variables as

$$\begin{aligned} U(a, \eta_{ij}, S, \xi) &= U(a, 0, S_0, \xi_0) + \frac{1}{\rho_0} C^S_{ij} \eta_{ij} + \frac{1}{2\rho_0} C^S_{ijkl} \eta_{ij} \eta_{kl} \\ &\quad + T(S - S_0) + A(\xi - \xi_0) + \cdots \end{aligned} \tag{7.25}$$

The constant $A = (\partial U / \partial \xi)_{\eta, S}$ is called the *affinity* for the reaction and is related to n_i as $A = \Sigma (\mu_{CP})_i n_i$, where $(\mu_{CP})_i$ is the chemical potential for the reaction leading to a change in n_i.

It can be established from thermodynamic considerations (Beyer and Letcher, 1969) that (i) if $Ad\xi > 0$, where $d\xi$ is a small variation in ξ, then the process naturally occurs and will take place; (ii) if $Ad\xi = 0$, the process is reversible; and (iii) if $Ad\xi < 0$, the process cannot occur by itself. Thus, the value $A = 0$ corresponds to equilibrium at some value ξ_0 of the degree of reaction. Acoustically, equilibrium corresponds to zero frequency. Displacements from equilibrium result in natural system forces reacting to restore equilibrium such that the displacement rate $(d\xi / dt) = bA$, where b is a constant. A constant ξ occurs when the acoustic frequency is so large that the system cannot follow the amplitude changes in the sound wave.

It is clear that irreversible thermodynamics requires the independent thermodynamic variable ξ in addition to the usual reversible thermodynamic variables. As in reversible thermodynamics, the irreversible thermodynamic "variables" σ_{ij}, η_{ij}, T, S, A, and ξ are not completely independent, but are related, for example, through Equation (7.25). Any of the irreversible thermodynamic "variables" may be expressed in terms of any three other irreversible thermodynamic "variables." Thus, the affinity A, although introduced as a constant in Equation (7.25) (as is the temperature T), may also be regarded as a variable in a different thermodynamic state equation. In particular, the affinity may be regarded as a function of σ_{ij}, S, and ξ, which can be expanded in a triple Taylor series about the equilibrium point $A[(\sigma_0)_{ij}, S_0, \xi_0] = 0$. Performing the expansion to first order, using standard relations among partial derivatives, and recombining terms, we obtain the relation (Bauer, 1965)

$$A(\sigma_{ij}, S, \xi) = \left(\frac{\partial A}{\partial \xi}\right)_{\sigma,\mathrm{S}} [\xi - \xi'(\sigma_{ij}, S)] \tag{7.26}$$

$$\xi'(\sigma_{ij}, S) = \xi_0 + \left(\frac{\partial \xi}{\partial \sigma_{ij}}\right)_{\mathrm{A,S}} [\sigma_{ij} - (\sigma_0)_{ij}] + \left(\frac{\partial \xi}{\partial S}\right)_{\mathrm{A},\sigma} (S - S_0) \tag{7.27}$$

where the subscripted zero denotes the initial value of the parameter. Substituting Equation (7.26) into the rate equation $(d\xi/dt) = bA$, holding the stress and entropy constant, and integrating, we obtain

$$\xi = \xi'(\sigma_{ij}, S) + const \cdot \exp(-t/\ \tau_{\sigma\mathrm{S}})$$

$$\tau_{\sigma\mathrm{S}} = -(1/\ b)(\partial\xi\ /\ \partial A)_{\sigma\mathrm{S}} \tag{7.28}$$

The parameter $\tau_{\sigma S}$ is the relaxation time for processes in which the stress and entropy remain constant. A corresponding relaxation time $\tau_{\eta S}$ may defined for processes in which the strain and entropy remain constant. The two time constants are related as $(\tau_{\sigma S}/\tau_{\eta S}) = (M_\infty/M_0) = (c_\infty/c_0)^2$, where M_∞ is the appropriate modulus (combination of elastic constants) and c_∞ is the sound speed at infinitely large frequency; M_0 is the modulus and c_0 is the sound speed in the limit of zero frequency.

Combining the above triple Taylor series expansion of $A(\sigma_{ij}, S, \xi)$ with a similar expansion of the strain $\eta_{ij}(\sigma_{ij}, S, \xi)$ and performing several transformations of partial derivatives, we obtain the equation of state for a given propagation mode in the form

$$\tau_{\sigma\mathrm{S}}\frac{\partial \varepsilon}{\partial t} + \varepsilon = \frac{\tau_{\sigma\mathrm{S}}}{M_\infty}\frac{\partial \sigma}{\partial t} + \frac{1}{M_0}(\sigma - \sigma_0) \tag{7.29}$$

where ε and σ denote (quasi)longitudinal or (quasi)shear strains and stresses, respectively, as appropriate. Combining Equation (7.29) with Newton's Law in the form $(\partial^2\sigma/\partial a^2) = \rho_0(\partial/\partial t)(\partial\varepsilon/\partial t) = \rho_0$ $(\partial^2\varepsilon/\partial t^2)$ and using the relation $(\tau_{\sigma S}/\tau_{\eta S}) = (M_\infty/M_0)$, we obtain the wave equation

$$\frac{\partial^2\psi}{\partial t^2} + \tau_{\sigma\mathrm{S}}\frac{\partial^3\psi}{\partial t^3} = \frac{M_0}{\rho_0}\left(\frac{\partial^2\psi}{\partial a^2} + \tau_{\eta\mathrm{S}}\frac{\partial^3\psi}{\partial a^2\partial t}\right) \tag{7.30}$$

where ψ denotes either σ or ε.

We assume harmonic waves of the form $\psi = \psi_0\exp[i(k'a - \omega \mathrm{t})$, where $k' = k + i\alpha = (\omega/c) + i\alpha$ is the attenuation coefficient, and $c = (\omega/k)$ is the phase velocity of the sound wave. Substituting this form into Equation (7.30) and defining $\tau = (\tau_{\sigma S}\tau_{\eta S})^{1/2}$, we obtain that the phase velocity is given by

$$c = \frac{\omega}{k} = \left[\frac{M_\infty}{\rho_0}\left(1 - \frac{\Delta M\ /\ M_\infty}{1 + \omega^2\tau^2}\right)\right]^{1/2} \tag{7.31}$$

where the relaxation strength ε_r is defined by

$$\varepsilon_\mathrm{r} = \frac{\Delta M}{M_\infty} = \frac{M_\infty - M_0}{M_\infty} = \frac{c_\infty^2 - c_0^2}{c_\infty^2} \tag{7.32}$$

and the wave attenuation (absorption) coefficient α is given by

$$\alpha = \frac{\omega^2(\tau_{\eta\mathrm{S}} - \tau_{\sigma\mathrm{S}})}{2c(1 + \omega^2\tau^2)} \approx \frac{1}{2}\varepsilon_r\frac{\omega^2\tau}{1 + \omega^2\tau^2} \times \left(\frac{\rho_0}{M_\infty}\right)^{1/2} \tag{7.33}$$

Relaxation in solids can result from a number of different mechanisms. The diversity of mechanisms include atomic movement by diffusion, stress-induced ordering whereby point defects producing a localized strain will move from one state of equilibrium to another under the action of a sound wave, thermoelastic loss involving heat flow between hot and cold regions of the sound field, Bordoni peaks in metals involving the "jumping" of dislocation segments between equilibrium positions, and Akhieser loss in which time-varying strains continually change the equilibrium conditions for the thermal phonon distribution in the solid (Berry and Nowick, 1966).

7.4.3.2 **Relaxation in Fluids**

For a fluid medium, we again assume that the stress σ_{ij} is the thermodynamic pressure p obtained as $\sigma_{ij} = -p\delta_{ij}$. Assuming small variations in the mass density ρ about the initial mass density ρ_0 in Equation (7.1) and using the isotropic symmetry properties of the elastic constants, we may recast Equation (7.29) in the form

$$\tau_{\mathrm{pS}}\frac{\partial\rho}{\partial t} + \rho - \rho_0 = \frac{\tau_{\mathrm{pS}}}{c_\infty^2}\frac{\partial p}{\partial t} + \frac{1}{c_\infty^2}(p - p_0) \tag{7.34}$$

where we have used the relation $(t_{\mathrm{pS}}/\tau_{\rho S}) = (M_\infty/M_0) = (B_\infty/B_0) = (c_\infty/c_0)^2$, where B is the bulk modulus. Combining Equation (7.34) with Newton's law and assuming harmonic wave propagation as before, we obtain that the absorption coefficient at unit wavelength $\alpha\lambda$, where λ is the wavelength given as

$$\alpha\lambda = \left(\frac{c^2}{c_0^2}\right)\varepsilon_{\mathrm{r}}\frac{\pi\omega\tau_{\mathrm{pS}}}{1+\omega^2\tau_{\mathrm{pS}}^2} \tag{7.35}$$

where ε_r is the relaxation strength given by Equation (7.32). The phase velocity c is given by

$$\frac{c^2}{c_0^2} = \frac{1+\omega^2\tau_{\mathrm{pS}}^2}{1+\left(c_0^2\big/c_\infty^2\right)\omega^2\tau_{\mathrm{pS}}^2} \tag{7.36}$$

If we define $\tau = (\tau_{pS}\tau_{\rho S})^{1/2}$, then

$$\frac{c^2}{c_0^2} = 1 + \left(\frac{\varepsilon_{\mathrm{r}}}{1-\varepsilon_{\mathrm{r}}}\right)\left(\frac{\omega^2\tau^2}{1+\omega^2\tau^2}\right) \tag{7.37}$$

which can be recast in the form of Equation (7.31) using the definition of ε_r, Equation (7.32).

In the case of gases, the relaxation phenomena are often recast in terms of heat capacities. The ideal gas is assumed to consist of identical molecules all of which are capable of being excited into different energy states each characterized by specific energy differences per mole. The relaxation strength for ideal gases is often written as $\varepsilon_r = (C_p^0/C_p^\infty)[(C_\rho^\infty/C_\rho^0) - (C_p^\infty/C_p)]$, where the $C_p^j (j = 0, \infty)$ are the appropriately indicated heat capacities. An understanding of the physical phenomena involved for a given gas is obtained from an examination of the physical models used for the relaxation processes. It is generally assumed in gases that equilibrium involves collisions among pairs of molecules, so that the reaction rate (reciprocal of relaxation time) is directly proportional to the product of the probability that an excited molecule will lose its energy in a given collision and the average collision frequency. In an ideal gas, the collision frequency is a linearly proportional gas density or pressure, so the relaxation time $\tau \sim 1/p_0$. Thus, $\omega\tau \sim \omega/p_0$ and the absorption coefficient α and the phase velocity c are functions of the ratio of the sound frequency to gas pressure. Experimentally, this means that data can be taken equivalently as a function of sound frequency or gas pressure.

In liquids, the modeling is complicated by the fact that, unlike gases, the molecular collisions are not generally binary but involve both nearest neighbor and next nearest neighbor molecules. The collision rate in liquids is also usually much higher than in gases. This raises the relaxation frequency in most cases to ranges above that is experimentally attainable. Further, the number of vibrational states in liquids is considerable, complicating the relaxation processes. Nonetheless, the application of relaxation concepts has proven useful in connection with theories of the liquid state, and vibrational relaxations have been discovered in numerous liquids. The nature of the relaxation processes involved has been accurately established for liquids with rotational isomers. Generally, the attenuation in viscous fluids and in some biological tissues is proportional to the frequency or to some power of the frequency that is less than one. For more details, see Beyer and Letcher (1969) and Bauer (1965).

7.4.3.3 **Viscoelasticity**

Historically, viscoelasticity was introduced before the constructs of irreversible thermodynamics were established. The classical Maxwell (1867), Kelvin (1875) and Voigt (1892) models of viscoelasticity and the Zener (1948) anelastic solid model were based on various mathematical combinations of elastic and flow elements of the material. The elastic behavior is represented by springs and the viscous behavior is represented by a dashpot in these models. Although each of the classical models can be obtained from the more general relaxation model based on irreversible thermodynamics, it is instructive to consider each model independently.

Maxwell Material The Maxwell model (1867) was introduced to explain the behavior of fluids whereby the fluid responds to an applied shear stress with both elasticity and flow. Maxwell combined Hooke's law, $\sigma_{12} = G\varepsilon_{12}$, where σ_{12} is the shear stress, ε_{12} is the shear strain, and G is the shear modulus, with Newton's law for a viscous fluid, $\sigma_{12} = \mu(d\varepsilon_{12}/dt)$, where μ is the shear viscosity coefficient. He obtained the equation

$$\frac{d\varepsilon_{12}}{dt} = \frac{1}{G}\frac{d\sigma_{12}}{dt} + \frac{1}{\mu}\sigma_{12} \tag{7.38}$$

Using the relation $(\tau_{\sigma S}/\tau_{\eta S}) = (M_\infty/M_0) = (G_\infty/G_0)$ in Equation (7.29) and comparing the resulting expression with Equation (7.38) reveals that $G = G_\infty$ and that $\mu = \tau_{\eta S}G_\infty$. Assuming sinusoidal variation in σ_{12} and ε_{12} (i.e., sinusoidal wave propagation) and recalling the definition of the shear modulus (Hooke's law), we find from Equation (7.38) that

$$G = G_\infty \frac{\omega^2\tau_{\eta S}^2}{1+\omega^2\tau_{\eta S}^2} + iG_\infty \frac{\omega\tau_{\eta S}}{1+\omega^2\tau_{\eta S}^2} \tag{7.39}$$

Because the shear modulus G is complex, the sound velocity is also complex and gives rise to an attenuation of the sound wave along the propagation path. At very high frequencies, $G \to G_\infty$ and the fluid exhibits a shear rigidity that allows shear wave propagation. At very low frequencies, $G \to 0$ and the fluid loses its solidlike properties. Maxwell's model has been successfully applied to the characterization of the shear properties of such substances as pitch, simple liquids, and polymers.

Kelvin-Voigt Material Kelvin (1875) and Voigt (1892) independently suggested that

Hooke's law be modified to include a contribution from the strain rate as

$$\sigma_{ij} = C^S_{ijkl}\varepsilon_{kl} + \mu_{ijkl}\frac{\partial \varepsilon_{kl}}{\partial t} \qquad (7.40)$$

where μ_{ijkl} is the viscosity tensor. Substituting Equation (7.40) into Newton's law and assuming isotropic symmetry, we obtain the following equation (using Lamé elastic constants) for longitudinal wave propagation:

$$\rho_0\frac{\partial^2 u_1}{\partial t^2} = (\lambda + 2G)\frac{\partial^2 u_1}{\partial a_1^2} + (\chi + 2\mu)\frac{\partial^3 u_1}{\partial a_1^2 \partial t} \qquad (7.41)$$

where χ is the compressional viscosity coefficient.

We assume a plane harmonic wave disturbance of the form $u_1 = w_1 \exp[i(k'a - \omega t)]$, where the wave number $k' = k + i\alpha = (\omega/c_1) + i\alpha$, α is the attenuation coefficient, and c_1 is the longitudinal velocity. Substituting this form into Equation (7.41) and solving for c_1 and α gives $c_1 = [(\lambda + 2G)/\rho_0]^{1/2}$ as before and

$$\alpha \approx \frac{\omega^2(\chi + 2\mu)}{2\rho_0 c_1^2} \qquad (7.42)$$

A similar procedure for shear waves in isotropic materials yields the values $c_2 = (G/r_0)^{1/2}$ for the shear wave velocity and

$$\alpha \approx \frac{\omega^2 \mu}{2\rho_0 c_1^2} \qquad (7.43)$$

for the shear wave attenuation coefficient.

7.5
Wave Dispersion

The dependence of sound velocity on wavelength, such as that arising from modal propagation in waveguides, for example, is known as *wave dispersion*. A wave disturbance of finite duration such as an ultrasonic pulse or toneburst (a gated cw) may be viewed as composed of an integrated sum of monochromatic waves where the integration is performed over a continuous range of infinitesimally differing frequencies of the waves. In a linear, dispersive, propagation medium, the acoustic waveform $u(x,t)$ at position x and at time t in the medium is given by the expression (Whitham, 1974)

$$u(x,t) = F(x,t)e^{i\Omega(x,t)} + c.c.$$
$$\Omega(x,t) = kx - \omega t \qquad (7.44)$$

where *c.c.* denotes the complex conjugate of the preceding term providing

$$\frac{d\omega}{dk} - \frac{x}{t} = 0$$
$$\frac{d^2\,\omega}{dk^2} \neq 0 \qquad (7.45)$$

where ω is the angular frequency. The functional dependence of angular frequency on wave number k defines the dispersion relation $\omega = \omega(k)$. The condition $(d^2\omega/dk^2) \neq 0$ is the defining condition for a dispersive propagation medium and $(d\omega/dk) = c_g$ is defined to be the group velocity. For a given choice of x and t, Equation (7.45) may be solved for k if the dispersion function $\omega(k)$ is known. Thus, different parts of the waveform, corresponding to different given pairs (x, t), propagate with different group velocities $c_g(k)$ corresponding to specific wave numbers determined from Equation (7.45). The group velocity corresponds to the velocity of energy propagation of the wave (Whitham, 1974; Pierce, 1989).

Both the complex amplitude $F(x, t)$ and the phase function $\Omega(x,t)$ in Equation (7.44) are functionally dependent on x and t. Equation (7.44) has the form of an elementary waveform, but $F(x, t)$

is actually a position- and time-varying envelope of the waveform with the position- and time-varying phase $\Omega(x, t)$ defining nonuniformly spaced maxima and minima of the wave. Equation (7.44) thus defines a disturbance, which in general changes shape during its propagation through a dispersive medium.

The propagation of a given constant value of phase Ω_o defined by $\Omega(x, t) = \Omega_o$ moves with the phase velocity $c_p(x, t) = dx/dt = \omega/k$. The phase velocity $c_p(x, t)$ is generally different for different values of Ω_o and for different values of x and t. An observer following a particular phase (e.g., crest of a wave) moves with the local phase velocity having a locally varying frequency and wave number. Thus, the phase velocity of a wave of finite length propagating in a dispersive medium can only be defined locally (i.e., for given x and t) and for a given phase Ω_o. Hence, if a given pulse measurement technique (see Section 7.7) follows a given phase Ω_o of the waveform at different monotonically increasing or decreasing values of the space-time coordinates (x_i, t_i), then the local phase velocity at (x_α, t_α) for phase Ω_o can be obtained by plotting x_i as a function of ti and measuring the tangent to the curve at (x_α, t_α). For dispersive media, the curve is generally nonlinear, but for nondispersive media the curve is linear and coincides with the curve for the group velocity, $c_g = x/t$. But in general, the group velocity is related to the phase velocity as $c_g = c_p - \lambda(dc_p/d\lambda)$, where λ is the phase velocity wavelength (Pollard, 1977).

7.6
Ultrasonic Generation and Detection

The generation and detection of ultrasonic waves can be achieved using a variety of methods. One of the earliest and simplest methods for generating ultrasound was the Galton whistle, which produced sound waves in air by blowing high-pressure air against a sharp metal edge. Some methods require direct physical contact between the ultrasonic source and the propagation medium and are usually, but not always, associated with piezoelectric generation and detection of sound. Other methods, not requiring direct contact, are found to be useful in some applications.

7.6.1
Piezoelectric Transducers

When a mechanical stress σ_{ij} is applied to a piezoelectric crystal, an electric polarization P_i is produced according to the relation $P_i = d_{ijk}\sigma_{jk}$, where d_{ijk} are the piezoelectric coefficients. Similarly, when an electric field E_i is applied to a piezoelectric crystal, a strain ε_{jk} is produced in the crystal according to the relation $\varepsilon_{jk} = d_{ijk}E_i$. Thus, piezoelectric crystals can both generate and detect sound waves.

Although there are in general 27 piezoelectric coefficients, the crystal symmetry considerably reduces the number. Quartz, for example, has only five independent coefficients. If the quartz is cut such that a compressive stress σ_{11} occurs along the x (a_1) axis, then $P_1 = d_{111}\sigma_{11}$, $P_2 = P_3 = 0$. The electric polarization is also along the x axis. Conversely, if an electric filed is applied along the x axis, the resulting strains include a component $\varepsilon_{11} = d_{111}E_1$. Thus, the same piezoelectric crystal can be used to generate and detect compressional sound waves.

Generally, piezoelectric crystals can be cut in specific ways to generate and detect compressional waves or shear waves. The fundamental frequency of vibration f_1 for thickness longitudinal modes is

$f_1 = (1/2L_x)(c_{11}/\rho_0)^{1/2}$, where L_x is the thickness of the crystal along the x direction and c_{11} is the elastic coefficient. For shear modes, the fundamental vibration is determined from $f_1 = (1/2L_y)(c_{44}/\rho_0)^{1/2}$, where L_y is the thickness of the crystal along the y direction. Harmonics of the fundamental frequency in each case may be generated, but only for odd multiples of the fundamental. This occurs because the strain changes sign for each half wavelength and the polarization can be detected only when opposite charges appear at the opposing crystal surfaces.

Other single crystal materials used for piezoelectric transducers include ethylene diamine tartrate (EDT), sodium potassium tartrate (Rochelle salt), dipotassium tartrate (DKT), and ammonium dihydrogen phosphate (ADP).

Polycrystalline ceramic materials such as barium titanate, lead zirconate titanate (PZT), lead metaniobate, and sodium metaniobate have also been popular materials for ultrasonic transducers. In contrast to inherently piezoelectric materials such as quartz, these ferroelectric ceramics are initially isotropic. Because an essential feature of piezoelectricity is the absence of a center of crystallographic symmetry, the ceramics must be electrically polarized above the Curie temperature to induce the anisotropy in the materials responsible for the piezoelectric properties. A particular advantage of these materials is that the transducer can be made in a variety of shapes and focusing forms. Further, the resonances of the polycrystalline ceramic transducers (low Q value) are generally much less sharp than those of single crystal materials (high Q value). If sound absorbing material is added to the back of the ceramic material, the resonance will be flattened even more. This occurs because waves propagating in the backward direction will be eliminated from the superposition of resonance-forming waveforms (see Section 7.3.2). The result is a transducer having a broadband frequency capability for both generating and receiving sound.

High-frequency transducers made from thin film zinc oxide and cadmium sulfide deposited on a base material are capable of generating waves with frequencies in the pretersonic range of frequencies. Piezoelectric polymer film transducers such as polyvinylidene fluoride (PVDF) have a broad frequency bandwidth that is advantageous for use with short pulses. For more information on piezoelectricity, see Berlincourt, Curran, and Jaffe (1964) and O'Donnell, Busse, and Miller (1981).

7.6.2
Noncontacting Methods

In many applications, it is desirable to generate and detect sound by methods that are not in direct physical contact with the material to be measured. These applications may include, for example, materials for which the temperature is too high, materials that are too corrosive, or materials that are too inaccessible to use contacting transducers. A variety of noncontacting generation and detection techniques have been developed for such situations.

7.6.2.1 Capacitive Transducers

For solids, the capacitive transducer (also called the electrostatic acoustic transducer or ESAT) is a simple parallel plate arrangement in which a flat sample surface functions as the ground plate. The other plate is a flat electrode recessed (typically of the order $7-10\,\mu\text{m}$) from the sample surface. A dc bias V_b (typically 50–150 V) is applied through a large resistor to the electrode. An ultrasonic wave incident at the solid surface varies the gap spacing (hence capacitance) of the parallel plate arrangement. The measured output voltage V_{out} from this arrangement is related to the displacement amplitude u of the incident wave amplitude u as $V_{\text{out}} = V_{\text{b}}(2us^{-1})$, where s is the gap spacing between the plates in the absence of the sound wave. The capacitive transducer for solids has a flat frequency response from dc to well into the gigahertz range of frequencies.

The capacitive transducer may also be used as a sound generator by applying an ultrasonic drive voltage across the electrodes, which produces an electric field E at the sample surface. The resulting electrical force per unit area σ at this surface is $\sigma = (1/2)\varepsilon_0 E^2$, where ε_0 is the dielectric permittivity of air. For a sinusoidal drive voltage $E = E_0 \sin \omega t$, the driving stress $\sigma = (1/4)\varepsilon_0 E_0^2(1 - \cos 2\omega t)$. Thus, the driving stress and resulting ultrasonic wave have a frequency that is twice that of the driving voltage. Typically, in a capacitive driving transducer, a material with a high dielectric permittivity such as mica is sandwiched between the electrodes. The material allows higher drive fields to be obtained without electrical breakdown.

A capacitive transducer for immersion in liquids is generally constructed by stretching a conductive membrane over an electrically isolated, dc biased, cylindrical electrode again recessed approximately $10\,\mu$m from the membrane inner surface. Unlike the capacitive transducer for solids, the liquid immersible capacitive transducer does not have a flat frequency response. In the ultrasonic range of frequencies, the inertia of the membrane and the coupling to the fluid produces a frequency roll-off that varies as $[1 + (\sigma_A \omega/Z)]^{-1/2}$, where σ_A is the area density of the membrane and Z is the acoustic impedance of the immersion fluid (Cantrell and Yost, 1989).

7.6.2.2 **Electromagnetic Acoustic Transducers (EMATs)**

A useful noncontacting transducer for generating SAWs in electrically conductive materials is the electromagnetic acoustic transducer (EMAT). The transducers are based on the body forces produced by the presence of both static and time-varying magnetic fields. Figure 7.6 shows one configuration for generating surface waves. An alternating current is passed through a flat meander coil held near the surface of the conductor and produces an alternating magnetic field in the conductor. If a static magnetic field is also produced in the conductor having the field lines parallel to the surface, Lorentz forces on the electrons within the conductive material will couple to the material lattice. The lattice coupling will serve as a source for the generation of surface waves propagating along the static field lines. The elementary sources represented by each meander coil wire segment add coherently, if the periodic distance between the parallel wire segments equals one half of the acoustic wavelength at the frequency of operation. For more information on EMATs, see Thompson (1990) and Frost, Sethares, and Szabo (1977).

7.6.2.3 **Laser Generation and Detection**

The generation of sound waves by incident optical laser radiation occurs by a number of mechanisms including radiation pressure, material vaporization, thermal expansion, Brillouin scattering, and electrostriction (due to electric polarization-induced movement of atoms into regions of different optical intensity). In most cases, the thermal

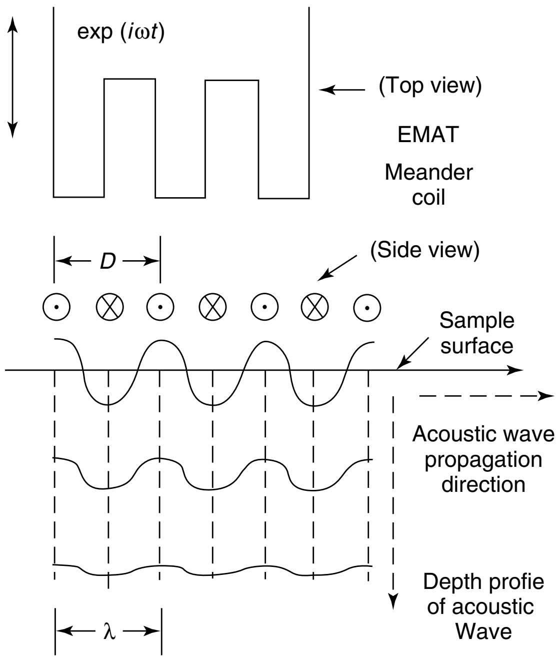
exp (iωt)
(Top view)
EMAT
Meander coil
D
(Side view)
Sample surface
Acoustic wave propagation direction
Depth profie of acoustic Wave
λ

Fig. 7.6 EMAT meander coil configuration for generating surface acoustic waves. Adapted from Thompson, 1977.

expansion mechanism dominates. If the incident laser energy is a pulse having energy E_0 that passes through a length L of a medium having an absorption coefficient α, then it can be qualitatively shown (Hutchins, 1988) that the resulting acoustic pressure amplitude p in the material is $p = K\beta cE_0\alpha/C_p$, where β is the volume expansion coefficient, C_p is the isobaric specific heat, c is the sound velocity, and K is a constant that depends on the configuration geometry and laser pulse shape. The details of the laser-induced acoustic pressure profile depends on whether the medium is solid or fluid, the shape of the acoustic source formed by the interaction of the laser light with the material, and the physical constraints placed on the propagation medium.

Laser detection of sound is commonly accomplished using an interferometric technique or a variation thereof. A variation useful for noncontact probing of pulsed SAWs is illustrated in Figure 7.7. Light from a laser is divided into two parallel and slightly separated collimated beams of equal intensity by a fixed beamsplitter. The beams are partially transmitted by a second beamsplitter and focused to points separated by a fixed distance on the surface of the specimen. Upon reflection, the beams are partially reflected by the second beamsplitter and superimposed to form a straight line interference pattern which is filtered with a spatial filter. This interferometer is sensitive to differential changes in the optical pathlengths of the two beams resulting from the passage of an SAW. The sound field causes a relative motion of the output fringe pattern with respect to the fixed spatial filter. For more information on optical detection of ultrasound, see Wagner (1990).

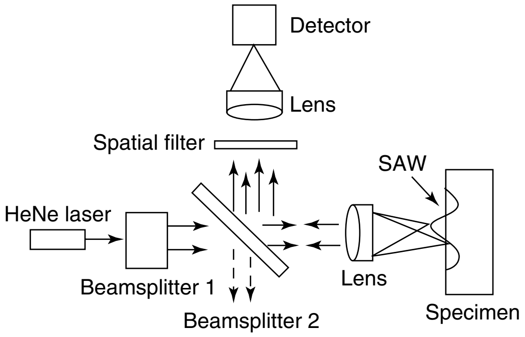

Fig. 7.7 Differential optical interferometric surface acoustic wave measuring system. From Claus and Cantrell, 1981.

7.6.2.4 Modulated Particle Beam Generation

Atomic particles incident on a material surface lose their energy in scattering or collision processes with target atoms as they penetrate into the material. Although the ions accelerated through an electric field have been used as particle beam sources, more typically accelerated electrons are used. Electron beams formed by accelerating electrons through a potential V (in kilovolts) will lose most of their energy due to atomic scattering in a roughly spherically shaped region (the interaction volume) of radius $d_s = V^{1.43}/(10\rho_0)$, d_s in microns, (Reimer, 1979). The electron scattering results in thermal generation as well as in momentum transfer. A modulated or chopped electron beam source generates both a thermal wave and acoustic waves simultaneously. A number of bulk and SAW modes are generated including modes resulting from thermal-to-acoustic wave mode conversion and from the driving source of electron in the interaction volume itself. Their amplitudes depend on the intensities of the driving source and the thermal waves and on their gradients at the material boundaries (Qian and Cantrell, 1989).

7.7
Measurement Techniques

Measurements of sound velocity and attenuation can provide considerable information about the physical properties of solids, liquids, and gases. Measurement of the variation of the sound velocity as a function of some intrinsic material variable, such as temperature or pressure, provides additional information about basic (often nonlinear) material properties. It is important to distinguish between group and phase velocities in the measurements of velocity. This distinction was addressed in Section 7.5 in the context of acoustic propagation in a dispersive medium. Similarly, it is important to distinguish between absorption and other attenuation mechanisms in acoustic measurements as addressed in Section 7.4.

Numerous techniques have been reported in the literature for making velocity and attenuation measurements. The most common techniques are based on optical, pulse, and cw methods, or on hybrids of these methods. The choice of technique is dictated in part by the availability of equipment and ease of setup for the particular application, the specific material and ultrasonic properties of the material to be measured, and the accuracy and precision desired in the measurement. Here, we briefly consider only representative techniques. For more information on measurement methods, see the review articles by McSkimin (1964), Papadakis (1968, 1976), Breazeale, Cantrell, and Heyman (1981), and Cantrell and Yost (1997).

7.7.1
Pulse Techniques

Pulse techniques utilize either broadband acoustic pulses or gated continuous acoustic waves (tonebursts) generated by an ultrasonic source.

7.7.1.1 Basic Pulse-Echo Method

An equipment arrangement for the basic pulse-echo system that allows for both velocity and attenuation measurements is shown in Figure 7.8. A pulsed ultrasonic signal is transmitted into the sample by means of a transducer attached to the sample surface. The ultrasonic pulse travels through the sample and reflects between the sample boundaries, eventually decaying away because of sample attenuation. Each time the ultrasonic pulse strikes the end of the sample coupled to the transducer, an electrical signal is generated, which is amplified and displayed on an oscilloscope. The received signals are commonly rectified and filtered (i.e., detected) before being displayed on the oscilloscope. The velocity of ultrasonic wave propagation is determined by measuring the transit time between the corresponding reference points of the reflected pulses and the corresponding pulse propagation distance in the sample. The ultrasonic attenuation can also be read directly from the calibrated attenuator by adjusting the attenuator to maintain a constant receiver signal level for each displayed echo. Accuracies of a few parts in 10^3 for absolute velocity measurements and five parts in 10^2 for attenuation measurements are typically achieved with the basic pulse-echo method. Numerous modifications and variations of the basic pulse-echo system have been developed that result in measurement accuracies of a few parts in 10^4 for absolute velocity measurements, one part in 10^7 for relative velocity measurements, and two parts in 10^3 for attenuation measurements. The reader is referred to McSkimin (1964), Papadakis, (1968, 1976), and Breazeale, Cantrell, and Heyman (1981) for details.

7.7.1.2 Pulse Superposition Method

The pulse superposition method uses a series of tonebursts introduced into the sample by a radiofrequency (rf) pulse generator. The repetition rate of these pulses is controlled by the frequency of a cw low-frequency oscillator. The repetition rate is adjusted to correspond to some multiples of an acoustic round-trip transit time in the sample corresponding to an "in-phase" superposition of pulses such that maximum amplitude is achieved. The measured time delay between the superimposed pulses is used to obtain the sound velocity. Accuracies of two parts in

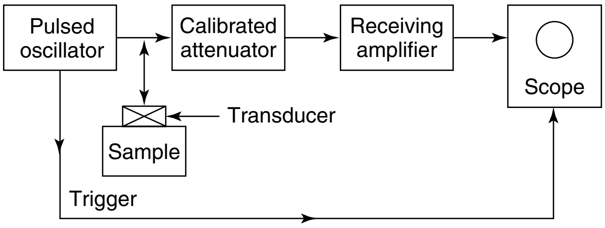

Fig. 7.8 Basic ultrasonic pulse-echo measurement system.

10^4 for absolute velocity measurements and one part in 10^7 for relative velocity measurements are possible with the pulse superposition method (see McSkimin, 1964, for more details).

7.7.1.3 Echo-Overlap Method

In the pulse-overlap method, a series of rf pulses from an rf-pulsed oscillator is transmitted into the sample. The trigger signal is derived by frequency dividing the output of a cw low-frequency oscillator (typically 100–1000 kHz) by a factor of 10^3. The trigger is used to control the pulse repetition rate and to trigger a double time delay (strobe) generator that actuates the z axis intensity gate on the oscilloscope. The delays are adjusted to intensify any chosen pair of displayed echoes. The cw oscillator frequency is adjusted to correspond to the reciprocal of the travel time between the echoes so that the echoes can be made to overlap cycle-for-cycle. The sound velocity is calculated from the measured travel between echoes after correcting for phase shifts due to reflections. Accuracies of a few parts in 10^4 for absolute velocity measurements are typical with the pulse superposition method (see Papadakis, 1976, for more details).

7.7.1.4 Pulse Interferometer Methods

In the pulse interferometer methods, an ultrasonic toneburst is transmitted into the sample by gating a cw reference oscillator. The received pulse (an echo in a single transducer arrangement) is combined with the cw reference signal in a phase-sensitive detector. If the relative phase between the received acoustic signal and the cw reference signal is then changed, the output of the phase-sensitive detector varies in response. The change in relative phase may result, for example, from a variation in the acoustic path or from a variation in the sound velocity as a function of temperature.

Two basic means are reported to measure the change in relative phase. One of the means employs a phase shift network that monitors the amount of phase added to the reference signal necessary to maintain a constant output from the phase-sensitive detector. The frequency of the cw reference signal is held fixed during this process. The second means uses a change in the frequency of the cw reference signal to maintain a constant detector output. In the pulse phase-locked loop system, a voltage controlled oscillator network replaces the phase shift network.

In both the constant frequency and variable frequency techniques, the total phase seen by the phase detector includes acoustic contributions related to the acoustic wavelength relative to sample length as well as collective nonacoustic contributions. The frequency-dependent nonacoustic contributions include those from electrical signal propagation paths and phase perturbations associated with various electrical components. In measurements of phase velocity in fluids, for example, the change in phase corresponding to a given change in path length may be measured. For the constant frequency technique, both the wavelength and the nonacoustic phase term are constants. The phase velocity is obtained from the measurement of the slope of the linear curve giving the change in phase as a function of the change in path length. For the variable frequency technique, the wavelength and the nonacoustic phase term are not fixed and the velocity can be calculated only after correcting for the variations in the nonacoustic phase term. The pulse interferometer methods have a sensitivity of parts in 10^8 for relative velocity measurements. The accuracy of absolute velocity measurements is comparable to that of the

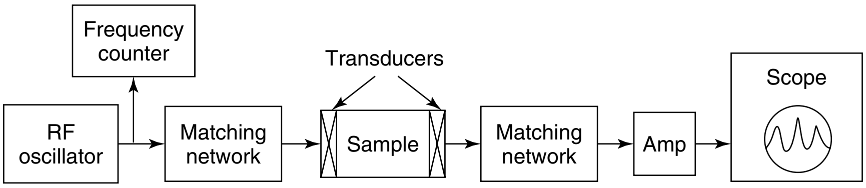

Fig. 7.9 Continuous wave ultrasonic interferometer.

pulse superposition and the echo-overlap methods. For more information on pulse interferometer methods, see Yost, Cantrell, and Kushnick (1992).

7.7.2
Continuous Wave Techniques

The cw techniques employ uninterrupted harmonic waves introduced into bounded material and are among the most expedient methods for measuring sound velocity and attenuation of materials with appropriate geometry. An advantage of cw methods is the increase in sensitivity to variations in velocity and attenuation resulting from the resonance conditions established by the material geometry and the physical properties. The effective gain in sensitivity can be measured in terms of "sensitivity enhancement factors."

Consider the one-dimensional resonances discussed in Section 7.3.2. For any observable A, the variation in amplitude $\Delta|A|$ may be expressed as $\Delta|A| = (\partial|A|/\partial\alpha)\Delta\alpha + (\partial|A|/\partial k)\Delta k$ for small variations in the attenuation α and wavenumber k. The partial derivatives are the sensitivity enhancement factors for attenuation and wavenumber, respectively. If the observable is the particle velocity $|v|$, the sensitivity enhancement factors are found from Equation (7.18) to be

$$\frac{\partial|v|}{\partial\alpha} = \frac{L(e^{-\alpha L/2} - e^{\alpha L/2})\cos kL}{8^{1/2}(\cosh\alpha L - \cos kL)^{3/2}} \tag{7.46}$$

and

$$\frac{\partial|v|}{\partial k} = \frac{Le^{\alpha L/2}\sin kL}{8^{1/2}(\cosh\alpha L - \cos kL)^{3/2}} \tag{7.47}$$

These factors together with similar factors for v_1 and v_2 can be used to separate the effects of dispersion and attenuation on the measurement parameters.

A simple transmission cw measurement system called the *continuous wave* ultrasonic interferometer is shown in Figure 7.9. In this technique, an oscillator is swept through a series of resonance peaks as observed on an oscilloscope screen. The frequencies at which the peak amplitudes occur are used to determine the ultrasonic velocity, and the frequency width $\Delta\omega$ at half power is used to calculate the attenuation in the manner described in Section 7.3.2. The addition of a mixer and 0 and 90° phase shifters to the system permits the assessment of the in-phase (v_1) and quadrature (v_2) resonance signals. Several modifications to the basic cw transmission spectrometer have been reported. The reader is referred to Breazeale, Cantrell, and Heyman (1981) for details.

For materials having a fixed length and sufficiently small dispersion, the phase velocity *cp* can be determined from the frequency difference ($f_n - f_{n-1}$) between two adjacent resonances in the manner discussed in Section 7.2.2. The attenuation α can be calculated from the determination of the frequency width at half power, $\Delta\omega$ as $\alpha = \Delta\omega/2c_p$. These results are based on the assumption that the sample responds as a simple resonator driven by noncontacting transducers. If contacting transducers are used, the sample and the transducers form a composite resonating system whose resonant frequencies are predicted by the roots of a transcendental equation. Approximate solutions to the transcendental equation have resulted in correction factors for the above formulas (Chern, Cantrell, and Heyman, 1981). However, errors in the velocity determination can exceed 10% under certain conditions.

For samples requiring analysis in a three-dimensional geometry, the method of resonant ultrasonic spectroscopy (Demarest, 1971; Ohno, 1976; Migliori *et al.*, 1993; Anderson, 1992; Visscher *et al.*, 1991) can be used. In this method, the modal frequencies for a sample of given three-dimensional geometry are compared to predictive analytical models of the sample modal frequencies. Analysis of the amplitude–frequency spectrum allows an assessment of the elastic constants as well as the internal friction (attenuation) of the sample.

7.7.3
Optical Techniques

Optical techniques can be used for sound velocity measurements in transparent media. Generally, the methods are based on the optical determination of acoustic wavelength either from diffraction or from scattering of light by the sound waves. The wavelength together with the acoustic frequency allows the determination of the phase velocity. For techniques based on diffraction, it is essential to know the type of diffraction experienced by the light.

7.7.3.1 Diffraction of Light by Sound

Consider an ultrasonic transducer generating compressional waves in a transparent propagation medium. If the propagating sound wave encounters light traversing the same medium at some angle incident to the sound wave, the light is diffracted. Assuming that the Fraunhofer conditions hold for the light diffraction (i.e., parallel light rays are brought to a focus), two distinct physical processes can produce the diffraction effects. One involves the formation of a corrugation in the phase fronts of the light due to the acoustically induced spatial variation of the index of refraction. This phenomenon is called *Raman–Nath diffraction*. The second process involves the reflection of light from the (evenly spaced) crests of the sound wave that occurs under conditions similar to that of X-ray diffraction. This phenomenon is called *Bragg diffraction*.

A dimensionless parameter has been introduced to determine which type of diffraction predominates for given experimental conditions. The parameter Q is defined by $Q = k^2 L/\mu_0 k^*$, where k is the ultrasonic propagation constant, L is the width of the ultrasonic beam, μ_0 is the index of refraction, and k^* is the propagation constant of the light in vacuo (Klein, Cook, and Mayer, 1965). If $Q > 9$, Bragg diffraction is predominant. If $Q < 1$, Raman–Nath diffraction occurs. If $1 < Q < 9$, the diffraction is mixed. Both types of diffraction can be used to determine the sound velocity of the material.

For Raman–Nath diffraction with the light incident at an angle $(\pi/2 - \phi)$ to the sound beam direction, the light is diffracted into diffraction orders given by the expression $\sin(\theta_n + \phi) - \sin(\phi) = n\lambda^*/\lambda$, where θ is the angle of diffraction with respect to the incident beam, λ^* is the wavelength of light in the medium, λ is the ultrasonic wavelength, and n is an integer denoting the diffraction order. The ultrasonic wavelength can be determined from the above equation.

For Bragg diffraction, the angle of diffraction θ_n is set to be equal to twice the negative angle of incidence $(-\phi)$ such that $n\ \lambda^* = 2\lambda \sin\phi_B$, where ϕ_B is the Bragg angle. Bragg diffraction is used to measure wave velocities in the frequency range from approximately 100 MHz to several gigahertz. The accuracy of velocity measurements using diffraction of light by sound methods is estimated to be of the order of 0.01–0.1% for homogeneous samples.

7.7.3.2 **Brillouin Scattering**

Lattice vibrations are caused by thermal and other excitations of the crystal lattice. These vibrations form vibrational modes (normal modes) within the lattice and play a major role in determining the thermal properties (e.g., heat capacity) of the solid. The normal modes form two major branches. The acoustic branch includes normal mode frequencies, which vanish linearly as the propagation vector $\boldsymbol{k}$ approaches zero (long wavelength limit). The optical branch includes normal mode frequencies that do not vanish as the propagation vector approaches zero.

The quantum mechanical treatment of the lattice vibrations predicts that the normal modes gain and lose energy in quantized units just as in the case of the harmonic oscillator. Because of the similarity of the lattice vibrations to the electromagnetic radiation field in a cavity, a quantum of lattice vibrational energy called the *phonon* is defined, primarily for convenience, in analogy to the electromagnetic quantum of energy called the *photon*. As with photons, phonons are given the property that each phonon carries an energy $\hbar\omega$, where $\hbar$ is the Planck's constant divided by 2π, and under certain conditions behaves as if it carries a crystal momentum of value $\hbar k$.

Light interacting with the normal modes in the material can gain or lose energy and momentum. Brillouin scattering occurs when the modes of the acoustic branch interact with light. Consider the interaction of a photon with a material whose index of refraction is μ_o. Application of the laws of conservation of energy and momentum, respectively, to the scattering process gives $\hbar\boldsymbol{\omega}^{*'} = \hbar\boldsymbol{\omega}^* \pm \hbar\omega$ and $\hbar\mu_o\boldsymbol{k}^{*'} = \hbar\mu_o\boldsymbol{k}^* + \hbar\boldsymbol{k} + \hbar\boldsymbol{K}$, where $\omega^{*\prime}$ and $\boldsymbol{k}^{*\prime}$ are the frequency and propagation vector, respectively, of the incident photon, ω^* and $\boldsymbol{k}^*$ are the frequency and propagation vector of the scattered photon, ω and k are the frequency and the crystal momentum assigned to the phonon, and $\boldsymbol{K}$ is the reciprocal lattice vector (assumed to be zero in this case). Because the photon frequency shifts are small for this process, the propagation vectors are very nearly of the same length. The vector addition expressed by the above equations gives an isosceles triangle whose apex angle is θ, the complement of the scattering angle. The equations lead to the result (Ashcroft and Mermin, 1976)

$$\omega = \pm 2\,\omega^* \left(\frac{c}{C_1}\right) \sin\left(\frac{\theta}{2}\right) \tag{7.48}$$

where C_1 is the speed of the photon in the medium, c is the speed of the phonon, and θ is the scattering angle. The frequency ω is equal to the difference in frequencies between the incident and scattered photons.

The technique produces a scattered photon spectrum of three lines, one of which is unshifted. The two frequency-shifted lines are called the *Stokes line* (a phonon is created) and the *anti-Stokes line* (a phonon is annihilated). These lines can be separated and measured with optical devices such as the Fabry–Perot interferometer to obtain ω. If desirable, the Bragg condition can be used for constructive reinforcement by adjusting θ. Equation (7.45) can be solved for the velocity c in terms of the other quantities measured.

Brillouin scattering is useful in the investigation of the sound velocity of solids and liquids especially near a phase transition where attenuation can become very large. The technique is also advantageous in that it is one of the few techniques capable of assessing sound velocity at frequencies of several gigahertz. The technique has been used for sound velocity measurements for both longitudinal and mixed (longitudinal-transverse) modes.

7.7.4
Sources of Error

A number of sources of error arise in the measurements of velocity and attenuation. Errors arising from time or length measurements can be evaluated directly, but among the less easily evaluated errors are those arising from velocity dispersion, material inhomogeneity, phase cancellation effects, diffraction, and the effects of thickness of the transducer and transducer bond. Many aspects of the sources of error are best addressed in the context of the measurement technique used, but certain characteristics of the error sources have more generic implications.

For example, as indicated in Section 7.5, any technique employing a change in frequency ω to measure the sound velocity c_p must account for dispersion, because the velocity is a function of the frequency. Measurements using gated cws (tonebursts) of finite length contain a spectrum of frequencies around the frequency of the generating continuous waveform. Such measurements can be made to approximate narrowband cw conditions, hence minimizing the effects of dispersion, if the pulse lengths are sufficiently long. The appropriate pulse length L is dependent on the intrinsic error M of the pulse measurement method and is given by the expression $L \geq \pi(d^2\omega/dk^2)/Mc_p$, where k is the wave number (Yost, Cantrell, and Kushnick, 1992).

A variation in the phase of the wave over the area presented by the surface of a phase-sensitive transducer can give rise to phase cancellation of different portions of the received signal that sets limits on the measurement accuracy of both velocity and attenuation. Sources of such phase variation include diffraction, media inhomogeneity, temperature gradients, wedging, and nonparallelism of the sample surfaces. Nonparallelism of opposite surfaces of the sample by an angle θ is representative of the effects of phase cancellation. For example, in measurements employing acoustic pulses of wave number k reflecting from the ends of the sample, the amplitude of the pulse received after n double reflections is diminished by the factor $[J_1(2kan\theta)/kan\theta]$ due to nonparallelism Truell, Elbaum, and Chick, 1969). The use of phase-insensitive devices such as the "acoustoelectric transducer" can minimize phase cancellation errors (Heyman, 1978), although such devices are generally less sensitive than the commonly used phase-sensitive piezoelectric elements.

A major source of error in measurements of velocity results from the effects of the transducer and the bond that couples the transducer to the sample. Both the transducer and bond introduce additional phase into the measurement system that can lead to errors of the order of parts in 10^2 for the measurement of velocity. Methods for correcting such errors are well established and have been addressed as appropriate to the specific measurement techniques discussed (for details see the review articles by McSkimin (1964), Papadakis (1968, 1976), Breazeale, Cantrell, and Heyman (1981), and Cantrell and Yost (1997). An alternative to such corrections is the use of noncontacting transducers such as capacitive transducers (Cantrell and Breazeale, 1977), electromagnetic transducers (Meridith, Watts-Tobin, and Dobbs, 1969), and optical techniques.

7.8
Applications of Ultrasound

Ultrasonic methods have long been used to characterize various properties of materials ranging from the fundamental to the most practical applications. Ultrasonic measurements of fundamental material properties, for example, have provided some of the first experimental evidence of the double well interatomic potential in vitreous silica, have been used to quantify the superconducting state of materials, and have provided well-established methods of testing lattice dynamical theories of the solid state. More practical uses of ultrasound span a wide range of disciplines including medical and industrial applications, nondestructive evaluation and testing of materials, acoustic microscopy, and underwater acoustics.

7.8.1
Medical Applications

Many medical applications of ultrasound center on the scanning or imaging of organs and tissues. The simplest scanning technique is the basic pulse-echo technique, often called the *A-scan method*. In the A-scan, an ultrasonic pulse is transmitted into the body. Changes in the acoustic impedance along the propagation path at tissue interfaces result in the partial reflection and transmission of sound from the interfaces. The reflected pulses (echoes) are received by the transducer and the echoes are displayed on the x trace of an oscilloscope and recorded. The amplitude of the echo is proportional to the magnitude of the reflection coefficient at the interface involved. A measurement of the time delays, calibrated in terms of distance, of ultrasonic echoes reflected from various anatomical landmarks provides the basis for the measurements. The A-scan is useful, for example, in the assessment of the anterior and posterior chambers of the eye. A variation of the A-scan is the M-scan in which a time record of variations in the x trace echo pattern is serially recorded along the y trace of the oscilloscope. The M-scan has proven useful in heart monitoring.

A B-scan image is formed by linking the time-base direction in the oscilloscope with the direction of propagation of the ultrasonic beam. The image is built up by using the amplitude of the received echoes to modulate the intensity of the oscilloscope beam as the ultrasonic beam is scanned.

In the ultrasonic C-scan, the x and y positions of the oscilloscope beam are made to track the x and y positions of the transducer as the transducer is made to areally scan a portion of the body. The z direction is the direction of ultrasonic propagation. A gated time delay is used to set the depth (z direction) at which ultrasonic echo information is received. The brightness of a given image point is dependent upon the amplitude of the received ultrasonic signal. In the C-scan mode, the ultrasonic images most closely resemble X-ray images.

Various modifications of the basic A-, B-, and C-scan imaging techniques have been developed including transducer array methodologies with electronic beam steering and acoustic holographic techniques. Among the many applications of ultrasonic imaging in its various forms are fetal monitoring, heart and cardiac structure imaging, assessment of the depth and severity of skin burns, and more recently high-resolution ultrasonic breast examination.

Ultrasound is also used in numerous medical applications other than imaging. For example, the Doppler shift associated with sound scattering from moving targets is used to measure both blood flow velocity and flow volume in blood vessels. Lithotripsy, a nonsurgical procedure, uses high-intensity sound focused into a small volume and directed onto hard objects such as kidney stones in the body. The acoustic impedance mismatch between the kidney stone and the surrounding soft tissue produces a sufficiently large pressure on the stones to break them into smaller pieces which can be excreted. The absorption of ultrasound directed into a muscle is used for heat therapy and message. The ultrasonic syringe utilizes the ability of ultrasound to temporarily disrupt surface tissue organization to permit the injection of medicine through the skin into the body. More recent applications of ultrasound (currently undergoing clinical evaluation) include the nonsurgical assessment of decubitus ulcers (bedsores) using spectral analysis profiles of tissue reflected echoes (amplitude versus frequency curve) and the noninvasive monitoring of intracranial dynamics and pressure (ICP) using pulse phase-locked loop ultrasonic interferometry.

7.8.2
Industrial Applications

Among the most common uses of ultrasound in industry is that in the nondestructive evaluation and testing of materials. Basic ultrasonic pulse-echo techniques are widely used for the detection and locations of cracks and flaws during routine inspection of machine and structural components and fixtures. The propagation of cracks is also passively monitored using the acoustic signals emitted during microfracture (acoustic emission monitoring). Ultrasonic imaging techniques are widely used for the assessment of material integrity, particularly for the assessment of delaminations and debonds in advanced composite materials. Ultrasonic velocity and attenuation measurements are often employed to obtain mechanical property data, such as the elastic constants or parameters associated with relaxation processes (internal friction) of the material. Such data can also be used for the assessment of the purity or the processing conditions of the material. The data is also useful in the assessment of basic microstructure-property relationships for the development and manufacturing of materials with improved or modified properties.

Residual stresses formed during materials processing or alterations in the stresses during the fatigue of materials are often of concern in critical components subjected to high cycle stresses. Because the magnitude of the sound velocity is sensitive to the stresses in the material, ultrasonic techniques are often used for the evaluation of such stresses. The acoustoelastic birefringence method relies on a stress-induced anisotropy in the material, which splits an incident shear wave into two components having different measured sound velocities. The difference in the two sound velocities is proportional to the stress in the material. Another method of assessing the presence of stresses depends on the measurement of stress-induced changes in the longitudinal velocity using either ultrasonic interferometry or pulse time-of-flight measurement techniques. The latter technique is also used for the precise tensioning of threaded fasteners in critical applications such as the disconnect studs linking the external fuel tank to the space shuttle orbiter.

Among the most important uses of SAWs is that in electronic SAW signal processing devices. Generally, the SAW device consists of an interdigital transducer formed by depositing an array of interlacing, alternately, charged electrodes on the surface of a piezoelectric substrate. The interdigital transducer generates an SAW. The SAW propagates to an output interdigital transducer. The intervening path between the input and output transducer is designed to produce a desired change in the input signal that is captured at the output. The intervening path may provide, for example, a propagation time delay, a waveform distortion, or alterations in the spectral frequency components of the waveform. Among the many SAW devices are delay lines, mixers, filters, correlators, resonators, coders, oscillators, and convolvers. Because the sound velocity is dependent on temperature and pressure, measurements of the SAW velocity on calibrated standards are also used to provide a sensitive monitor for both temperature and pressure. SAWs have been used to sense the position of a pointer (stylus or human finger) on a computer-controlled graphical input screen. SAW filters are widely used in cellular telephones. For more information on SAW devices, see Kino (1987).

Optical interactions with SAWs are used to make surface wave optical modulators, deflectors, and filters. For example, SAWs generated onto a thin optical-guiding film causes periodic variations in the optical index of refraction of the material. These periodic variations serve as Bragg diffraction gratings for incident optical beams coupled into the film. The incident optical waves are deflected in accordance with the Bragg reflection conditions as they encounter the SAWs. Optical output couplers can then be placed to receive only optical signals having a desired angle of deflection. Ultrasonic-induced frequency shifts of several megahertz in the diffracted light of optical modulators are used for single sideband mixing or demodulation of laser Doppler devices.

Applications of ultrasound other than materials characterization or SAW devices are also widely used in industry. For example, acoustic Doppler shift measurements are often used for monitoring the flow velocity of gases and molten solids. Doppler techniques are also used for measuring the velocity of objects moving in a fluid and ultrasonic Doppler sensors are commonly used in entry alarm systems. High-frequency acoustic vibrations in crystals are used in devices requiring precise frequency or time control. Basic pulse time-of-flight techniques are used for measuring the distance between an object and the sound source (range-finding). Ultrasonic rangefinders are found on many autofocusing cameras. The pressure in resonating high-intensity sound fields can be used to support and manipulate objects in air without mechanical contact with the object (acoustic levitation). Such manipulation is useful for handling contact sensitive or corrosive materials. SAWs can be used to manipulate and position small particles on a material surface. High-intensity sound is used in ultrasonic cutting and drilling machines, ultrasonic cleaners, and ultrasonic humidifiers (atomizers). High-intensity ultrasound is also used for soldering, welding (both metals and plastics), stress relieving previously welded joints, and metal forming and drawing.

Ultrasonic motors have become an important development for self-focusing cameras and robotic devices. Although there are many variations, the ultrasonic motor basically consists of a vibration-driven stator frictionally coupled to a rotor. A piezoelectric element is used to excite a flexural wave in the stator. The stator is placed in direct contact with the rotor. The transverse undulations of the flexural wave produces points on the stator surface that are in direct contact with the rotor and other points that are not in contact. As the wave propagates, a given point on the stator surface moves in an elliptically shaped, retrograde pattern similar to that of the Rayleigh wave. The motion of the stator–rotor contact point corresponding to the maximum wave amplitude is opposite to that of the wave propagation direction. The frictional forces between the stator and the rotor cause the rotor to move in that direction (Inaba *et al.*, 1987).

Thermoacoustic heat pumps have been developed that utilize a heat source and heat sink configured to an acoustic resonator having a stacked heat exchanger. The action of acoustically driven thermodynamics processes in the resonator gases has allowed the development of reliable, inexpensive engines, and refrigerators with a significant Carnot efficiency (∼40%) (Swift, 1995).

7.8.3
Acoustic Microscopy

The microstructural variation in the acoustic properties of materials (including biological material) has been the key to the development of a new class of microscopes that utilize acoustic waves in some fashion for obtaining microscopic images. The images so obtained are quite different from those obtained in conventional light or electron microscopes. Different types of acoustic microscopes use different energy sources such as electron beams or piezoelectric transducers for generating the acoustic images and are descriptively named, for example, the scanning acoustic microscope, the scanning electron acoustic microscope, the thermoacoustic microscope, and the scanning laser acoustic microscope (SLAM). Although the image contrast and spatial resolution attainable in such microscopes are in large measure dictated by the details of energy source-material interactions, the dependency on the local variations in the material elastic or thermoelastic properties, including residual stresses, is a necessary consideration in such interactions.

7.8.3.1 Scanning Acoustic and Laser Acoustic Microscopes

In scanning acoustic microscopy (SAM), plane acoustic waves are generated by a piezoelectric transducer attached to a cylindrical rod (typically made of sapphire) as shown in Figure 7.10. The opposite end of the rod is made into a concave spherical surface to form an acoustic lens. Coupling fluid (usually water) is placed between the rod surface and the specimen surface. Wave refraction at the acoustic lens-coupling fluid interface provides a focusing action for the plane waves incident from the transducer. The focused waves impinge on the specimen surface. In general, the waves incident on the specimen surface are specularly (geometrically) reflected back to the transducer. At certain critical angles of incidence (if the material properties permit), leaky surface

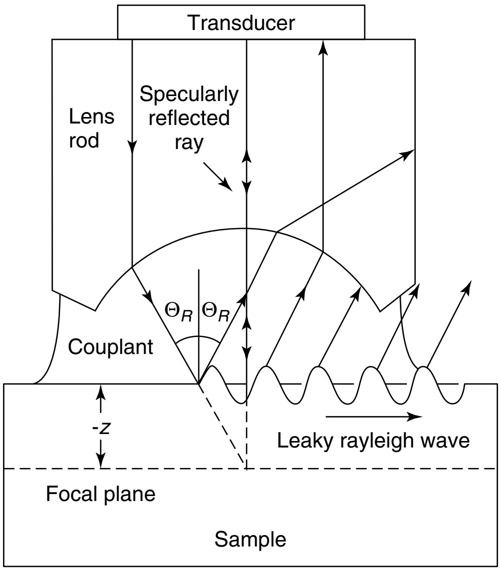

Fig. 7.10 Scanning acoustic microscope acoustic lens configuration for generating signals via the interference between "leaky" Rayleigh waves generated at the Rayleigh critical angle θ_R and specularly reflected waves (illustrated by the central specularly reflected wave). Adapted from Briggs, 1992.

waves (usually leaky Rayleigh waves) can also be generated. Under proper defocus conditions (lens-specimen distance), the leaky surface waves find their way back to the transducer. The total acoustic signal reaching the transducer in this mode of operation is composed of contributions from the geometrically reflected waves and the leaky surface waves. The interference of these two waves gives rise to the SAM output signal. If the acoustic lens–specimen distance is allowed to vary, the periodic variation in the signal amplitude as a function of this can be used to obtain the surface wave velocity of the material.

If the focused acoustic beam is rastered over the specimen surface, variations in the acoustic impedance of the specimen or in the surface wave velocity or geometrical irregularities in the specimen surface can change the magnitude and phase of the output signal. The variations in output signal at each point on the surface are recorded on a visual display to generate an SAM image. The images can be analyzed to obtain information related to the variations in the mechanical properties of the specimen over micron scale regions. Commercially available scanning acoustic microscopes use acoustic waves with frequencies as high as 2 GHz. For more information on SAM, see Briggs (1992).

The SLAM uses unfocused plane acoustic waves (usually in the range 100–500 MHz) to insonify the specimen. Internal elastic microstructures or other acoustic inhomogeneities in the specimen cause the propagating sound wave to locally scatter, distort, or attenuate. The sound field then impinges on and deforms a flat, polished area of the specimen surface producing a surface pattern reflective of the specimen inhomogeneity. A laser beam reflected from this surface is rastered to trace out an optical phase replica of the sound field at the surface. The laser signals are electronically processed and displayed on a television monitor to produce a visual image of the surface pattern. For more information on the SLAM, see Kessler and Yuhas (1979).

7.8.3.2 **Scanning Electron Acoustic and Photoacoustic Microscopes**

In scanning electron acoustic microscopy (SEAM), the signal generation source is derived by chopping or modulating the electron beam (most often in the range 0.1–1.0 MHz) in a scanning electron microscope (SEM) by means of electrostatic deflector plates inserted into the electron beam column of the SEM. The deflector plates are driven by a periodic signal (typically a square wave) of frequency f derived from a function generator. The chopped electron beam is focused on the specimen surface. The electrons lose their energies by means of atomic collision processes in a scattering interaction volume as discussed in Section 7.6.2.4, resulting in thermal and acoustic waves of the same frequency being generated simultaneously in the interaction volume. The acoustic waves are detected by a transducer (typically a piezoelectric element) attached to the specimen. The signals from the transducer are commonly processed by means of a lock-in amplifier that derives its reference signal from the modulation signal of the function generator. The signal from the lock-in amplifier is differentially amplified and used to drive the z axis of a cathode ray tube (CRT) display. As the chopped electron beam rasters the specimen surface, an electron-acoustic image is generated on the CRT display.

The acoustic SEAM signal output $V(f)$ may be approximated for isotropic materials as $V(f) = \alpha_T(3\lambda + 2G)R(f)/(\rho_0 C_H \kappa)^{1/2}$, where α_T is the thermal expansion coefficient, C_H is the specific heat, κ is the thermal conductivity, and ρ_0 is the mass density of the material (Qian and Cantrell, 1989). $R(f)$ is a function of the chopping frequency that collectively includes those factors not directly involving thermoelastic parameters. We see from this expression that the electro-acoustic signal is primarily governed by the thermoelastic properties of the material. The resulting SEAM images thus reflect variations in the thermoelastic parameters of the material and allow variations in material properties affecting the thermoelastic parameters to be imaged. Because the acoustic signal is derived from the electron interaction volume, both surface and subsurface features including residual stresses can be imaged with the SEAM, as illustrated in Figure 7.11. For further information on SEAM, see Cargill (1988).

The photoacoustic microscope uses intensity-modulated light rather than an electron beam to generate the acoustic signals. The acoustic signals are generally detected by a piezoelectric transducer bonded to the sample or by various optical methods. The original photoacoustic arrangement consists of a specimen enclosed in a gas-filled cell. Periodic heat flow and thermal expansion of the specimen leads to a corresponding pressure variation of the gas in the cell, which is detected by a microphone. The acoustic signals generated in the photoacoustic microscope are primarily governed by the thermoacoustic properties of the material in a manner similar to that of the SEAM. Measurements of the photoacoustic signal amplitude as a function of the light modulation frequency leads to a materials characterization modality known as *photoacoustic spectroscopy*. Among the uses of photoacoustic spectroscopy are the measurements of optical absorption and spectroscopic depth profile analyses of opaque materials. For more information on the phenomenological aspects of photoacoustic microscopy, see McDonald and Wetsel (1988) and for information on photoacoustic spectroscopy, see Sawada and Kitamori (1988).

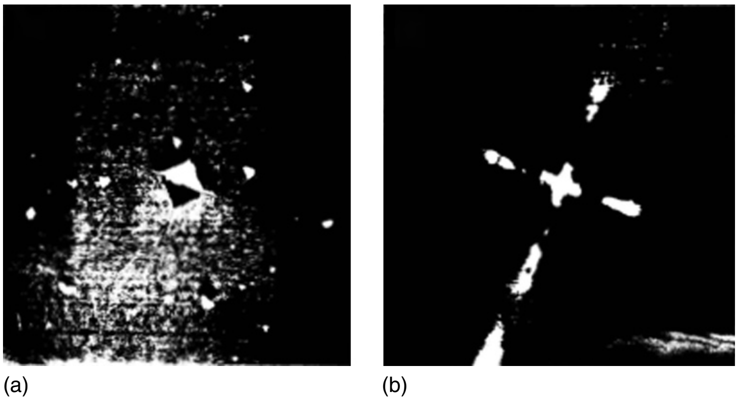

Fig. 7.11 Vickers indentation in reinforced alumina after a load of 15 kg has been applied and removed: (a) scanning electron microscope (SEM) secondary electron image showing surface cracks and deformation. (b) Scanning electron acoustic microscope (SEAM) image showing additionally the presence of residual stresses and subsurface cracks. From Cantrell *et al.*, 1990.

7.8.3.3 **Acoustic-Atomic Force Microscopes**

Acoustic-atomic force microscopy (A-AFM) is a dynamic implementation of atomic force microscopy (AFM) (Bennig, Quate, and Gerber, 1986) that combines the attributes of AFM with the bulk propagation properties of ultrasound to produce images of near surface and subsurface features within the material at the nanoscale. A-AFM includes such modalities as atomic force acoustic microscopy (AFAM) (Rabe *et al.*, 2002), force modulation microscopy (FMM) (Maivald *et al.*, 1991), ultrasonic force microscopy (UFM) (Kolosov and Yamanaka, 1993), heterodyne force microscopy (HFM) (Cuberes *et al.*, 2000), and resonant difference-frequency atomic force ultrasonic microscopy (RDF-AFUM) (Cantrell, Cantrell, and Lillehei, 2007). All modalities require that the AFM cantilever tip engage the sample surface, as indicated in Figure 7.12. Upon engagement, the cantilever tip encounters an interaction force that in general varies nonlinearly with the tip-surface separation distance. In hard tip-surface contact, the force-separation curve is linear to a good approximation. In soft tip-surface contact, the force-separation curve is extremely nonlinear at certain tip-separation distances.

Two basic equipment arrangements are used for A-AFUM depending on the modality. The simplest arrangement is that used for AFAM and FMM and is shown in Figure 7.12, where the indicated switches are in the open positions. AFAM and FMM utilize ultrasonic waves transmitted into the material by a transducer attached to the bottom of the sample. After propagating through the bulk of the material, the wave impinges on the sample top surface where it excites the engaged AFM cantilever. For AFAM and FMM, the cantilever tip is set to engage the sample surface in hard contact corresponding to the roughly linear interaction region of the force-separation curve. The response of the excited cantilever is monitored by a laser beam reflected from the cantilever

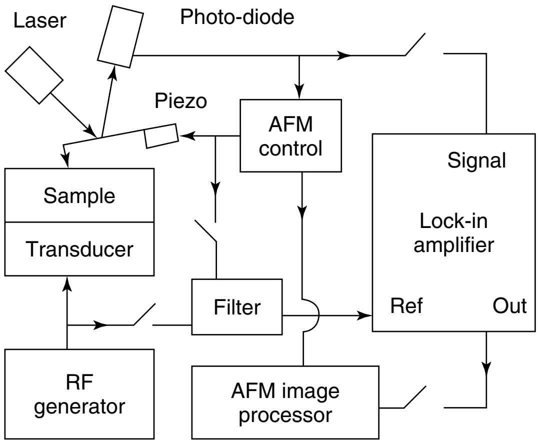

Fig. 7.12 Acoustic-atomic force microscope equipment configuration: switches open for AFAM, FMM, and UFM: switches closed for HFM and RDF-AFUM. Adapted from Cantrell, Cantrell, and Lillehei, 2007.

surface and detected by a photodiode. The signal from the photodiode is processed by the AFM image processor. The image processor and AFM controller are used to generate a raster image of subsurface, nanoscale, material features resulting from changes in the A-AFM signal produced by the features. The image may be an amplitude or phase-generated image depending on whether it is the change in the amplitude or the phase of the ultrasonic wave propagating through the material that is measured. The basic equipment arrangement used for UFM is the same as that for AFAM and FMM, except that the cantilever tip for UFM is set to engage the sample surface in soft contact, corresponding to the highly nonlinear interaction region of the force-separation curve. The UFM output response is a static or "dc" signal resulting from the interaction nonlinearity.

The equipment arrangement used for HFM and RDF-AFUM is shown in Figure 7.12, where the indicated switches are in the closed positions. In a manner similar to UFM, HFM and RDF-AFUM employ ultrasonic waves launched from the bottom of the sample, while the AFM cantilever tip engages the sample top surface in soft (nonlinear regime) contact. However, in contrast to UFM, the cantilever is also driven in the HFM and RDF-AFUM modalities. In RDF-AFUM, the cantilever drive frequency is set to differ from the ultrasonic frequency by one of the contact resonance frequencies of the cantilever. In HFM, the difference frequency is generally set to a value well below the lowest contact resonance of the cantilever. As shown in Figure 7.12, the cantilever drive and transducer drive signals are split and fed to a mixer. The mixer output signal, consisting of sum- and difference-frequency signals, is sent to the reference input of a lock-in amplifier that, because of its limited bandpass, filters out the sum frequency. The mixing of the oscillating cantilever and the ultrasonic wave in the nonlinear region defined by the cantilever tip-sample surface interaction force generates difference-frequency oscillations in the cantilever. The AFM photodiode signal, derived from the cantilever response at all frequencies, is sent to the signal input of the lock-in amplifier where all frequencies except the difference-frequency are filtered out. The lock-in amplifier measures both the amplitude and phase of the input difference-frequency signal. Variations in the amplitude and phase of the ultrasonic bulk wave due to the presence of subsurface nano/microstructures as well as variations in near surface material parameters affect the amplitude and phase of the difference-frequency signal. These variations are used to create spatial mappings generated by near surface and subsurface structures.

For a detailed mathematical analysis of signal generation, image contrast, and the relation of image generation to the physical properties of materials for each of the above-cited A-AFM modalities, see Cantrell and Cantrell (2008).

7.8.4
Underwater Acoustics

Because of the relative opacity of water to electromagnetic signals in the visible, infrared, radio, and radar ranges of frequencies, sound waves are used as the means of propagating information in underwater environments. The development of underwater sound systems has been among the most long-standing applications of ultrasound. Among the many developments are ultrasonic systems for both passive and active sound navigation ranging (sonar), depth sounding, mapping and profiling of the ocean floor, underwater geophysical prospecting, and fish detection.

Glossary

Acoustic Impedance: Ratio of the pressure to the particle velocity of the acoustic wave field.

Affinity: A measure of the ability of a reaction to occur and defined as the sum of the products of the chemical potential for a reaction and the number of moles of reactant–product involved.

Attenuation: The loss of acoustic amplitude (or energy) of a sound wave along its propagation path resulting from all mechanisms responsible for such losses, including but not limited to absorption, scattering, and diffraction.

Absorption: The loss of amplitude (or energy) of an acoustic wave that results in an increase in temperature (however slight) in the propagation medium.

Diffraction: Wave scattering occurring at the ‘‘edges’’ or boundaries of objects, including the influence of boundaries on radiation sources.

Dispersion: The variation of phase velocity with wavelength or wavenumber.

Elasticity: A material property characterized by a thermodynamically reversible relationship between stress and strain.

Evanescent Wave: A nonpropagating wavelike disturbance periodic only in time and decaying exponentially in all spatial directions.

Group Velocity: The velocity associated with the propagation of wave energy and defined as the derivative of the angular frequency with respect to the wavenumber.

Inhomogeneous Wave: A sound wave having an amplitude that decays exponentially in a direction perpendicular to the direction of propagation.

Phase Velocity: The sound velocity of a monochromatic wave defined as the ratio of the angular frequency to the wavenumber.

Refraction: The deviation in the propagation direction of a sound wave as the wave crosses the boundary at nonnormal incidence between two propagation media having different sound velocities.

Relaxation: A reaction process whereby the energy state of the material is caused to change from one state of equilibrium to another state of equilibrium by the driving action of an ultrasonic wave.

Relaxation Time: A time constant characterizing the time exponential approach to equilibrium following a sudden deviation from equilibrium of a relaxing medium.

Scattering: That phenomenon whereby an inhomogeneity in the path of a sound wave produces secondary waves spreading out from the inhomogeneity in various directions.

Viscoelastic Material: A material possessing both elastic and viscous properties.

Viscosity: A material property characterized by the dependence of the stress on the strain rate.

References

Anderson, O. (**1992**) *J. Acoust. Soc. Am*, **91**, 2245–2253.

Ashcroft, N.W. and Mermin, N.D. (**1976**) *Solid State Physics*, Saunders, Philadelphia.

Auld, B.A. (**1990**) *Acoustic Fields and Waves in Solids*, Kreiger Publishing Company, Malabar.

Bauer, H.J. (**1965**) in *Physical Acoustics*, vol. **IIA** (ed. W.P. Mason) Academic, New York, pp. 47–131.

Bennig, G., Quate, C.F., and Gerber, Ch. (**1986**) *Phys. Rev. Lett.*, **56**, 930–933.

Benson, G.C. and Kiyohara, O. (**1974**) *J. Acoust. Soc. Am.*, **55**, 184.

Berlincourt, D.A., Curran, D.R., and Jaffe, H. (**1964**) in *Physical Acoustics*, vol. **IA** (ed. W.P. Mason), Academic, New York, pp. 170–267.

Berry, B.S. and Nowick, A.S. (**1966**) in *Physical Acoustics*, vol. **IIIA** (ed. W.P. Mason), Academic, New York, pp. 1–42.

Beyer, R.T. and Letcher, S.V. (**1969**) *Physical Ultrasonics*, Academic, New York.

Bolef, D.I. and Miller, J.G. (**1971**) in *Physical Acoustics*, vol. **VIII** (eds. W.P. Mason and R.N. Thurston), Academic, New York, pp. 96–201.

Breazeale, M.A., Cantrell, J.H., and Heyman, J.S. (**1981**) in *Methods of Experimental Physics*, vol. **19** (ed. P.D. Edmonds), Academic, New York.

Brekhovskikh, L.M. (**1980**) *Waves in Layered Media*, Academic, New York.

Briggs, A. (**1992**) *Acoustic Microscopy*, Clarendon, Oxford.

Cantrell, J.H. and Breazeale, M.A. (**1977**) *J. Acoust. Soc. Am.*, **61**, 403–406.

Cantrell, J.H. and Cantrell, S.A. (**2008**) *Phys. Rev. B*, **77**, 165409.

Cantrell, S.A., Cantrell, J.H., and Lillehei, P.T. (**2007**) *J. Appl. Phys.*, **101**, 114324.

Cantrell, J.H., Qian, M., Ravichandran, M.V., and Knowles, K.M. (**1990**) *Appl. Phys. Lett.*, **57**, 1870–1872.

Cantrell, J.H. and Yost, W.T. (**1989**) *Rev. Sci. Instrum.*, **60**, 487–488.

Cantrell, J.H. and Yost, W.T. (**1997**) in *Encyclopedia of Acoustics*, vol. **2** (ed. M.J. Crocker), John Wiley & Sons, Inc., New York, pp. 629–637.

Cargill, G.S. (**1988**) in *Physical Acoustics*, vol. **XVIII** (eds. W.P. Mason and R.N. Thurston), Academic, New York, pp. 125–165.

Chern, E.J., Cantrell, J.H., and Heyman, J.S. (**1981**) *J. Appl. Phys.*, **52**, 3200–3204.

Claus, R.O. and Cantrell, J.H. (**1981**) *Acoust. Lett.*, **5**, 1–4.

Cuberes, M.T., Alexander, H.E., Briggs, G.A.D., and Kolosov, O.V. (**2000**) *J. Phys. D Appl. Phys.*, **33**, 2347–2355.

Demarest, H.H. Jr (**1971**) *J. Acoust. Soc. Am.*, **49**, 768.

Frost, H.M., Sethares, J.C., and Szabo, T.L. (**1977**) *J. Appl. Phys.*, **48**, 52.

Heyman, J.S. (**1978**) *J. Acoust. Soc. Am*, **54**, 243–249.

Hutchins, D.A. (**1988**) in *Physical Acoustics*, vol. **XVIII** (eds. W.P. Mason and R.N. Thurston), Academic, New York, pp. 21–118.

Inaba, R., Tokushima, A., Kawasaki, O., Ise, Y., and Yoneno, H. (**1987**) Proceedings of the IEEE Ultrasonics Symposium, Denver, Colorado, pp. 747–756.

Kelvin, L. (**1875**) *Encyclopedia Britannica*, 7th edition, p. 796.

Kessler, L.W. and Yuhas, D.E. (**1979**) *Proc. IEEE*, **67**, 526–536.

Kino, G.S. (**1987**) *Acoustic Waves*, Prentice-Hall, Englewood Cliffs.

Klein, W.R., Cook, B.D., and Mayer, W.S. (**1965**) *Acoustica*, **15**, 67–74.

Kolosov, O. and Yamanaka, K. (**1993**) *Jpn. J. Appl. Phys*, **32**, L1095–L1098.

Maivald, P., Butt, H.J., Gould, S.A., Prater, C.B., Drake, B., Gurley, J.A., Elings, V.B., and Hansma, P.K. (**1991**) *Nanotechnology*, **2**, 103–106.

Maxwell, J.C. (**1867**) *Philos. Trans. R. Soc. Lond.*, **157**, 49.

McDonald, F.A. and Wetsel, G.G.Jr (**1988**) in *Physical Acoustics*, vol. **XVIII** (eds. W.P. Mason and R.N. Thurston), Academic, New York, pp. 167–277.

McSkimin, H.J. (**1964**) in *Physical Acoustics*, vol. **IA** (ed. W.P. Mason), Academic, New York, pp 271–417.

Meridith, D.J., Watts-Tobin, R.J., and Dobbs, E.R. (**1969**) *J. Acoust. Soc. Am.*, **45**, 1393–1401.

Migliori, A., Sarrao, J.L., Visscher, W.M., Bell, T.M., Lei, M., and Fisk, Z. (**1993**) *Phys. B*, **183**, 1.

Morse, P.M. and Ingard, K.U. (**1968**) *Theoretical Acoustics*, McGraw-Hill, New York.

Murnaghan, F.D. (**1951**) *Finite Deformation of an Elastic Solid*, John Wiley & Sons, Inc., New York.

Nye, J.F. (**1960**) *Physical Properties of Crystals*, Oxford University Press, Oxford.

O'Donnell, M., Busse, L.J., and Miller, J.G. (**1981**) in *Methods of Experimental Physics*, vol. **19** (ed. P.D. Edmonds), Academics, New York, pp. 29–65.

Ohno, I. (**1976**) *J. Phys. Earth*, **24**, 355.

Papadakis, E.P. (**1968**) in *Physical Acoustics*, vol. **IVB** (ed. W.P. Mason), Academic, New York, pp. 269–328.

Papadakis, E.P. (**1976**) in *Physical Acoustics*, vol. **XII** (eds. W.P. Mason and R.N. Thurston), Academic, New York, pp. 277–374.

Pierce, A.D. (**1989**) *Acoustics*, Acoustical Society of America, Woodbury.

Pollard, H.F. (**1977**) *Sound Waves in Solids*, Pion, London.

Qian, M. and Cantrell, J.H. (**1989**) *Mater. Sci. Eng.*, **A122**, 57–64.

Rabe, U., Amelio, S., Kopychinska, M., Hirsekorn, S., Kempf, M., Goken, M., and Arnold, W. (**2002**) *Surf. Interface Anal.*, **33**, 65–70.

Reimer, L. (**1979**) *Scan. Electron Microsc.*, **2**, 111–124.

Sabine, H. and Cole, P.H. (**1971**) *Ultrasonics*, **9**, 103.

Sawada, T. and Kitamori, T. (**1988**) in *Physical Acoustics*, vol. **XVIII** (eds. W.P. Mason and R.N. Thurston), Academic, New York, pp. 347–401.

Scruby, C.B., Jones, K.R., and Anatoniazzi, L. (**1987**) *J. Nondestruct. Eval.*, **5**, 145–156.

Swift, G.W. (**1995**) *Phys. Today*, 22–28.

Thompson, R.B. (**1977**) Proceedings of the IEEE Ultrasonics Symposium, Phoenix, Arizona, p. 74.

Thompson, R.B. (**1990**) in *Physical Acoustics*, vol. **XIX** (eds. R.N. Thurston and A.D. Pierce), Academic, New York, pp. 157–200.

Thurston, R.N. (**1964**) in *Physical Acoustics*, vol. **IA** (ed. W.P. Mason), Academic, New York, pp. 2–109.

Truell, R., Elbaum, C., and Chick, B.B. (**1969**) *Ultrasonic Methods in Solid State Physics*, Academic, New York.

Überall, H. (**1997**) in *Encyclopedia of Acoustics*, vol. **1** (ed. M.J. Crocker), John Wiley & Sons, Inc., New York, pp. 55–67.

Voigt, W. (**1892**) *Ann. Phys.*, **47**, 671.

Wallace, D.C. (**1970**) in *Solid State Physics* (ed. H. Ehrenreich, F. Seitz, and D. Turnbull), Academic, New York, pp. 301–404.

Whitham, G.B. (**1974**) *Linear and Nonlinear Waves*, Wiley-Interscience, New York.

Yost, W.T., Cantrell, J.H., and Kushnick, P.W. (**1992**) *J. Acoust. Soc. Am.*, **91**, 1456–1468.

Visscher, W.M., Migliori, A., Bell, T.M., and Reinert, R.A. (**1991**) *J. Acoust. Soc. Am.*, **90**, 2154–2162.

Wagner, J.W. (**1990**) in *Physical Acoustics*, vol. **XIX** (eds. R.N. Thurston and A.D. Pierce), Academic, New York, pp. 201–266.

Zener, C. (**1948**) *Elasticity and Anelasticity of Metals*, University of Chicago Press, Chicago.

Further Reading

Bhatia, A.B. (**1967**) *Ultrasonic Absorption*, Clarendon, Oxford.

Crocker, M.J. (ed.) (**1997**) *Encyclopedia of Acoustics*, vols. **1–4**, John Wiley & Sons, Inc., New York.

Edmonds, P.D. (ed.) (**1981**) *Methods of Experimental Physics*, vol. **19**, Academic, New York.

Landau, L.D. and Lifshitz, E.M. (**1970**) *Theory of Elasticity*, Pergamon, Oxford.

Mason, W.P. and Thurston, R.N. (eds) (**1964–1988**) (eds. R.N. Thurston, A.D.

Pierce), (**1992–1995**), *Physical Acoustics*, Academic, New York. Each volume contains relevant information.

McDicken, N. (**1991**) *Diagnostic Ultrasonics*, Churchill Livingstone, New York.

Pierce, A.D. (**1989**) *Acoustics*, Acoustical Society of America, Woodbury.

Ziskin, M.C. and Lewin, P.A. (eds) (**1992**) *Ultrasonic Exposimetry* CRC Press, Boca Raton.

ULTRASOUND AND OTHER SOUND, PHYSICAL EFFECTS OF

E. Carr Everbach, *Engineering Department, Swarthmore College, Swarthmore, Pennsylvania, U.S.A.*

	Introduction	527
1.	**Heating Effects**	528
2.	**Thermoacoustics**	529
3.	**Effects on Electromagnetic Systems**	529
3.1	Nuclear Effects	529
3.2	Electron Effects	530
3.3	Optical Effects	530
4.	**Oscillations and Resonance**	530
4.1	Structural Resonance	530
4.2	Bubble Resonance	531
4.2.1	Acoustic Cavitation	531
5.	**Shock Waves**	534
5.1	Sonic Booms	534
5.2	Lithotripsy	535
6.	**Sonochemistry**	535
	Acknowledgment	536
	Glossary	536
	Works Cited	537
	Further Reading	538

INTRODUCTION

Sound waves, whether audible or not, are mechanical disturbances that propagate through a material and can cause physical changes to occur. In fluids with negligible viscosity, sound waves are longitudinal waves: the molecules move back and forth in the direction of propagation of the wave. In viscous liquids and solids, transverse waves can exist in which the molecules of the material move perpendicular to the direction of propagation. Other types of waves, such as shear waves or Rayleigh waves, can exist at interfaces between materials. Both energy and momentum are transferred by sound waves, although for audible frequencies and moderate sound levels in air, energy densities are typically less than a millionth of a joule per cubic meter. For very intense sounds, however, or for interactions between acoustic waves and highly resonant mechanical systems, sound energies can become concentrated enough to produce physical effects such as acoustic levitation, tissue heating, or even light production. The mechanical changes in the medium due to the sound waves can affect electrical, magnetic, or optical properties of the material, resulting in important secondary effects such as acoustic modulation of the optical index of refraction.

As sound waves propagate, they modify the spatial and temporal properties of the medium, producing compressional and rarefactional regions. Associated with these regions are changes in mechanical stress, strain, and motion that are described by such parameters as pressure, density, tension, particle velocity, and temperature. The kinetic energy of a sound wave is the energy of motion gained by the particles of the medium set in motion by the wave. The potential energy is due to the elasticity of the medium, the tendency of compressed or expanded material to return to its equilibrium state. The energy of the sound wave E is the sum of the kinetic and potential energies and is given by

$$E = \tfrac{1}{2}\rho_0 V_0(u^2 + p^2/\rho_0^2 c^2), \qquad (1)$$

where ρ_0 is the equilibrium density of the medium, u is the particle velocity, p is the instantaneous pressure above the equilibrium pressure, c is the sound speed, and V_0 is the volume occupied by the sound wave. The energy density ϵ of the wave is the ratio of E to V_0 and is an important measure of the strength of the wave. Other such measures are the *acoustic intensity I*, expressing the amount of energy per unit time (power) that flows through a given cross-sectional

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

area of a medium, and its logarithmic counterpart, the *intensity level* IL, expressed in decibels:

$$I = \frac{1}{T}\int_{t=0}^{T} pu\, dt, \tag{2}$$

$$\mathrm{IL} = 10 \log_{10}(I/I_{\mathrm{ref}}), \tag{3}$$

where T is the acoustic period, t is time, and I_{ref} is a reference intensity of 10^{-12} W/m^2 for sounds in air (Kinsler *et al.*, 1982). Physical effects begin when sound intensities exceed about 100 W/m^2 or 140 dB, although the nature of the effects varies depending upon the system considered.

Two important physical effects of intense sounds on fluids are acoustic radiation force and acoustic streaming. Acoustic radiation force, like that from light, results when a wave is reflected from an object with less momentum than it had previously (Henry, 1957), or when the momentum of the wave changes direction. The time rate of change of this momentum difference is the force exerted on the object in the direction of the momentum change. It follows that an object that is acoustically transparent and nondissipative will experience no acoustic radiation force, while one that reflects all of the incident energy 180° will experience the maximum acoustic radiation force. If the object perfectly absorbs the sound wave, it will experience half the maximum acoustic radiation force, and if the incident beam is deflected, the force on the object will be in the direction opposite to the surface normal. Streaming, on the other hand, is the bulk translational motion of an acoustically attenuating fluid due to its proximity to a sound source. The fluid is set in motion because the sound wave deposits more momentum into the fluid near the source than it does farther away. Thus streaming is the result of radiation force acting on a viscous liquid that is the propagation medium. Both streaming and radiation force are proportional to the intensity of the sound, while streaming also depends upon the acoustic absorption of the fluid and its shear viscosity (Nyborg, 1965). For more information, see the section on acoustic streaming in ACOUSTICS, NONLINEAR, and a description of acoustic radiation force in BIOLOGICAL EFFECTS OF SOUND AND ULTRASOUND.

The present article outlines the various consequences of the interaction of intense sound waves with the materials though which they propagate. We begin by considering the conversion of sound to thermal energy, both directly through absorption in the medium and indirectly through the interaction of thermal boundary layers with a standing wave field. We then consider nonthermal effects, beginning with electrical and resonance properties of the propagation medium. Finally, we move to specific phenomena related to these physical effects: acoustic cavitation and shock waves, along with some of their consequences. Treatment of SONOLUMINESCENCE and the BIOLOGICAL EFFECTS OF ULTRASOUND is left to the corresponding articles elsewhere. See ENGINEERING ACOUSTICS for a discussion of surface acoustic waves.

1. HEATING EFFECTS

As a mechanical disturbance, a sound wave ultimately dissipates its energy via friction into heat, a process called acoustic absorption. Depending upon the properties of the propagation medium, an acoustic wave may lose energy from different frequency components at different rates, causing distortion of the wave form as well as its lowered amplitude. For many homogeneous media, the intensity of the sound is reduced exponentially as the wave progresses in time or space:

$$I = I_0 e^{-\alpha z}, \tag{4}$$

where I_0 is the intensity at the source, z is the distance from the source, and α is the absorption coefficient of the medium. Energy in a plane acoustic wave will be converted into heat at the rate of $2\alpha I$ per unit volume (NCRP, 1983), and if there is no mechanism for conduction (or diffusion) of that heat out of the medium, the temperature Θ will rise linearly according to

$$\Theta(t) = \frac{2\alpha I}{\rho_0 C_p} t + \Theta_0, \tag{5}$$

where C_p is the specific heat at constant pressure and Θ_0 is the temperature of the medium at time $t = 0$. If convection or conduction is present, however, the temperature will reach a saturation level that balances the deposition of heat energy with the losses. Once sound energy is converted into heat, the physical effects that the heat has on the medium are indistinguishable from heating by any other means.

2. THERMOACOUSTICS

Under certain conditions, heat input may be made to cause a sound wave to be generated, and the result is a thermoacoustic engine or "prime mover" (Swift, 1988). Likewise, sound waves can be made to produce a temperature gradient in the thermal boundary layer of a surface, such as a stack of plates inserted in an intense standing wave field. The basic mechanism for this ability of sound energy to generate a temperature gradient is shown in Fig. 1. A small parcel of gas near a plate is forced back and forth by the sound wave, and it is compressed and expanded as it moves parallel to the plate surface. During the compression, the temperature of the gas parcel increases adiabatically, and the gas then transfers some of its heat at constant pressure to the plate, which warms slightly. During the rarefaction, the parcel expands and moves leftward, cooling adiabatically to a lower temperature. Heat from the leftmost portion of the plate enters the gas parcel at constant pressure and the cycle repeats, with a net amount of heat "lifted" from the plate at the left position and deposited at the right. Many microscopic interactions of this kind contribute to a "bucket brigade" of gas parcels that can move substantial amounts of heat from one end of the plate (or a stack of plates) to the other. For applications of this principle to specific devices such as acoustic refrigerators, please refer to the article THERMOACOUSTIC ENGINES AND REFRIGERATORS.

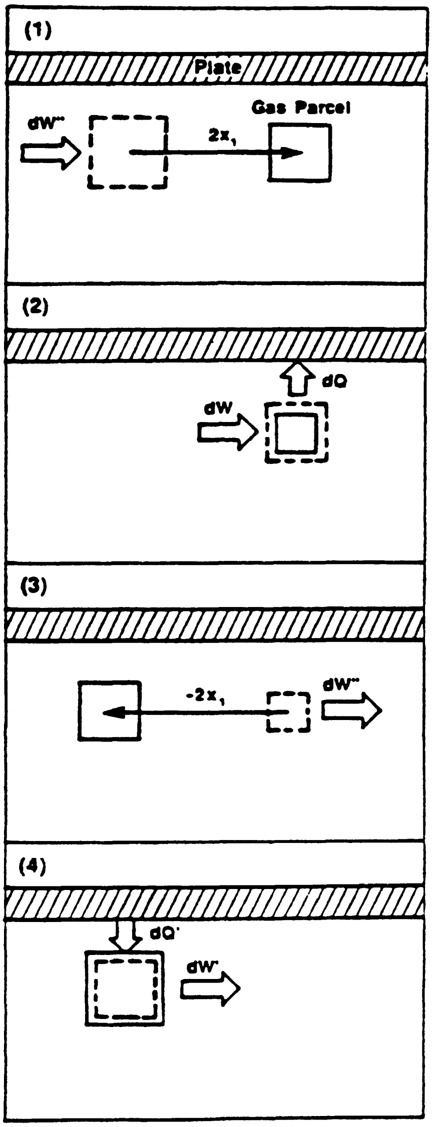

FIG. 1. Simplified mechanism of a thermoacoustic heat-pump cycle. A parcel of gas inside the resonator tube moves horizontally from left to right near a surface, compressing adiabatically in step (1), transferring heat dQ to the surface isobarically in step (2), expanding adiabatically in step (3), and picking up heat dQ′ from the surface isobarically in step (4). (Reprinted, with permission, from Swift, 1988. Copyright 1988 Acoustical Society of America.)

3. EFFECTS ON ELECTROMAGNETIC SYSTEMS

3.1 Nuclear Effects

Ultrasonic energy can induce transitions in the energy levels of nuclear spins subjected to electric or magnetic fields. This phenomenon, called nuclear acoustic resonance (NAR), is analogous to nuclear magnetic resonance (NMR) but relies not upon an externally generated magnetic field but rather upon ultrasound to cause internal interactions among the magnetic or electrical multipole moments of a nucleus. While NMR cannot be used to characterize the properties of metal alloys because of the poor penetration of radio-frequency electromagnetic fields in metals, NAR can couple phonons readily into bulk metals since ultrasonic absorption is low. The lattice vibrations caused by the acoustic wave couple energy into the nuclear spins by two primary mechanisms (Bolef and Sundfors, 1992): the dynamic nuclear electric quadrupole interaction and the dynamic nuclear dipolar (or Alpher-Rubin) interaction. In both interactions, the acoustic wave generates time-dependent gradients of electric field due to modulation of the distances between atoms. The changes in nuclear spin are usually detected with a superconducting quantum interference device (SQUID), which increases the sensitivity of magnetic measurements considerably (Holland and Sundfors, 1985).

3.2 Electron Effects

Acoustic paramagnetic resonance (APR) is similar to NAR except that ultrasonic phonons are used to excite transitions in electron spins rather than nuclear spins. The acoustoelectric effect describes the flow of electric charges in a conducting film that has been deposited upon a piezoelectric substrate through which a surface acoustic wave travels (Levy, 1992). The surface acoustic wave will lose energy through the dissipation caused by the motion of the charges, and this energy loss can be detected to yield information about the electrical conductivity of the surface layer. A similar phenomenon known as the magnetoelastic interaction occurs when surface acoustic waves traveling through thin nickel films over a quartz substrate are attenuated by changes in applied magnetic field (Levy, 1992). The attenuation is due to interactions between the phonons and electrons or magnetic moments in the surface film.

3.3 Optical Effects

Ultrasound interacts with materials to produce optical effects by modifying a material's index of refraction, n. The change in n is due entirely to instantaneous changes in density ρ that accompany the compressional and rarefactional pressure variations of the sound wave. Liquid water, for instance, has an index of refraction whose changes with pressure are given to within 5% at 20 °C by the following expression (Straudenraus and Eisenmenger, 1993):

$$\Delta n = 1.4 \times 10^{-4} \Delta p, \tag{6}$$

where the pressure variation Δp is in units of megapascals. Thus a ten-atmosphere variation in acoustic pressure Δp would produce a change Δn of about 1.4×10^{-4}, and for $p = 1000$ atm, $\Delta n \approx 1.4 \times 10^{-2}$ (Yadav *et al.*, 1973). Likewise, the changes of refraction index with pressure for silica glass are given by

$$\Delta n \approx 5 \times 10^{-6} \Delta p, \tag{7}$$

where Δp is again in units of megapascals (Barker and Hollenbach, 1970). The difference in changes of n with pressure between silica glass and water provides a method of using optical fibers to measure large-amplitude ultrasonic waves (Wurster *et al.*, 1994). Interferometric techniques have also been used to measure the changes of optical index of refraction caused by sound waves (Barker, 1972).

4. OSCILLATIONS AND RESONANCE

4.1 Structural Resonance

An acoustic wave that encounters a structure exerts a time-dependent force upon elements of that structure. This force is equal to the instantaneous acoustic pressure times the area upon which that pressure acts and is usually negligible for moderate sound intensities. Resonances of the structure, however, can allow significant oscillation amplitudes to occur after many acoustic cycles have each deposited energy in the structure over a long time. For instance, a structure with mass m, stiffness k, and damping b will have a well-defined resonance frequency at $f_R = \omega_d / 2\pi$, where

$$\omega_d = \sqrt{\omega_R^2 - \frac{b^2}{4m^2}}, \tag{8}$$

and

$$\omega_R = \sqrt{\frac{k}{m}}. \tag{9}$$

4.2 Bubble Resonance

Bubbles of a gas such as air or water vapor are more compressible than the surrounding fluid and hence may be considered to have an effective spring constant resulting from the increases in gas pressure that accompany any decrease in volume. If the gas in the bubble is considered an ideal gas, the adiabatic ideal gas relationship can be used (Everbach, 1989) to derive the linear, small-amplitude stiffness of the bubble as $12\pi a\gamma P_0$, where γ is the ratio of specific heats for the gas (approximately 1.4 for air), P is the instantaneous absolute pressure, P_0 is the mean absolute pressure, and a is the bubble radius. The effective mass of a bubble is due less to the gas inside it than to the inertia of the liquid surrounding the bubble when the bubble oscillates radially. Assuming the liquid is incompressible in comparison to the bubble, the effective mass m can be shown (Leighton, 1994) to be $4\pi a^2\rho_L$, where ρ_L is the density of the liquid. Thus a low-amplitude resonance frequency f_R exists for the bubble given by

$$f_R = \frac{1}{2\pi a}\sqrt{\frac{3\gamma P_0}{\rho_L}}, \tag{10}$$

where the R subscript signifies resonance. This relation was first derived by Minnaert (1933) to explain the "pitch" of bubbles in a babbling brook. Damping of the bubble oscillations occurs via three mechanisms:

1. acoustic waves produced by the bubble radiate energy away,
2. thermal conduction between the gas in the bubble and the surrounding liquid removes energy, and
3. viscous friction at the bubble wall dissipates energy.

Figure 2 shows graphically the variation of the different damping constants b as functions of frequency for resonant air bubbles in water (for algebraic expressions, see Eller, 1970).

When infinitesimally low-amplitude acoustic waves traveling in a liquid encounter a bubble, the bubble oscillates radially at the angular frequency ω of the incident wave with a radial amplitude given by (Leighton, 1994, Eq. 4.23, p. 295)

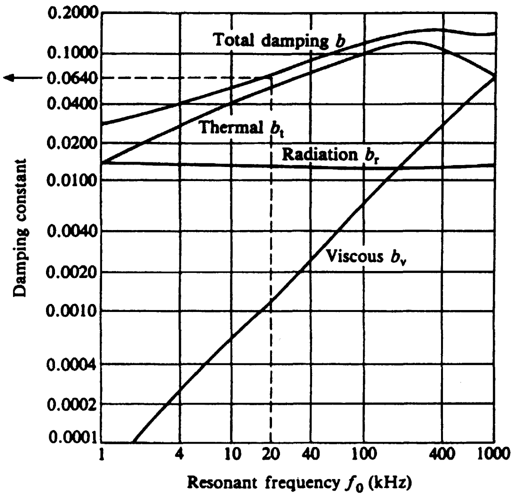

FIG. 2. Plot of damping constants b versus bubble resonance frequency for bubbles at resonance: b_v is damping due to the viscosity of the liquid surrounding the bubble, b_r is damping due to radiation losses, b_t is damping due to thermal losses via heat conduction, and b is the total damping constant (from Young, 1989; by permission).

$$\frac{P_A}{a\rho_L\sqrt{(\omega_R^2 - \omega^2)^2 + (2\beta_{\text{tot}}\omega)^2}}, \tag{11}$$

where P_A is the acoustic pressure amplitude, $\omega_R = 2\pi f_R$ is the resonance angular frequency of the bubble, and β_{tot} is the damping factor considering all dissipation mechanisms: $\beta_{\text{tot}} \approx b/2m$ for lightly damped bubbles. For finite-amplitude incident waves, oscillations of the bubble can have an effect on nearby objects, such as blood cells or ocean micro-organisms, because of the hydrodynamic flow patterns developed in the fluid around the bubble. In particular, shear forces sufficient to disrupt cell membranes have been observed to cause, for example, lysis of red blood cells (Williams *et al.*, 1986; Miller and Williams, 1983).

4.2.1 Acoustic Cavitation When a bubble is driven at or near resonance by an incident acoustic wave, it may oscillate in a continuous fashion (stable cavitation) or grow and collapse violently (transient or inertial cavitation), depending upon the bubble characteristics and the insonification parameters. Bubbles act as very nonlinear oscillators, and so even weakly driven bubbles may undergo changes in radius with time that demonstrate deterministic chaos (Cramer and Lauterborn, 1981; Parlitz *et al.*, 1990). As mentioned in the previous section, stable cavitation can produce microstreaming and large local shear forces in the fluid near the oscillating bubble walls. Transient cavitation, however, can result in additional physical effects including sonoluminescence, shock-wave production, liquid jet formation, and free-radical production. These effects are the result of a concentration of energy of up to 12 orders of magnitude from the relatively weak spatial ultrasonic field into the small volume occupied by the radially collapsing bubble.

For a bubble to grow and collapse violently, it must be inertially controlled: the bubble must grow during one or several rarefactional periods of the incident sound wave to two or three times its equilibrium radius. Once this rapid expansion occurs, the pressure is very low inside the expanded bubble, and during the next acoustic compression half-cycle, the walls of the bubble are accelerated radially inward. The stiffness of the gas inside the bubble is then far less important than the inertia of the moving liquid. The bubble walls thus converge, launching secondary shock waves and compressing whatever gas exists in the bubble to temperatures above 5000 K (Suslick, 1990). Such a spherically symmetric "hot spot" collapse is one mechanism suggested to account for the interesting chemistry and physics that can occur due to transient cavitation. Electrohydrodynamic mechanisms have also been proposed (Lepoint *et al.*, 1997). Please see the article on SONOLUMINESCENCE for more information.

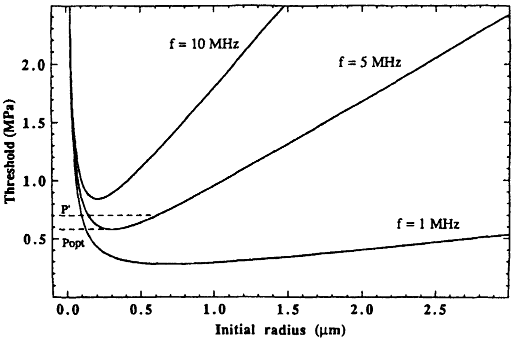

FIG. 3. Plot of the cavitation threshold (peak negative pressure) in water as a function of initial air-bubble radius for three frequencies of ultrasound: 1, 5, and 10 MHz, from an analytical model by Apfel and Holland (1991, by permission). As the threshold increases, a broader range of bubbles grow within a single acoustic cycle and collapse adiabatically to a temperature of 5000 K.

Bubbles are driven most efficiently at the resonance frequencies approximated by the formula in Sec. 4.2. In a bubble population containing a range of radii, therefore, those that are near resonance size collapse first and are called "prompt" cavitation nuclei. Figure 3 shows a plot of the peak negative acoustic pressure (called the cavitation threshold) that is sufficient to cause inertial growth and collapse air bubbles in water within a single acoustic cycle, as a function of initial bubble radius for three insonification frequencies. As the pressure amplitude increases at a particular insonification frequency (e.g., pressure *P'* for 5 MHz), a broader size range of nuclei act as "prompt" nuclei and cavitate within a single acoustic cycle. Those bubbles smaller than resonance size may grow via the process of rectified diffusion (Atchley and Crum, 1988) if the pressure amplitude is sufficiently high, as represented schematically in Fig. 4. Since the surface area of the bubble is slightly larger during a rarefactional half-cycle than during a compressional half-cycle, more gas diffuses into the bubble from the surrounding liquid during rarefaction than diffuses out during compression. Also, the layer of liquid surrounding the bubble tends to increase in thickness when the bubble contracts, decreasing the gas concentration gradient in the liquid closest to the bubble wall. The opposite effect occurs during the bubble expansion, increasing the local gas concentration gradient. Since the diffusion rate of gas is proportional to the concentration gradient, the result is a slow pumping of gas into the bubble over many acoustic cycles. Bubbles initially smaller than resonance size will tend to grow to resonance size and collapse, while bubbles larger than resonance size may grow large enough to rise via buoyancy out of the sound field. Under some conditions (Church, 1988), bubbles may become trapped in a stable cyclical process indefinitely, oscillating

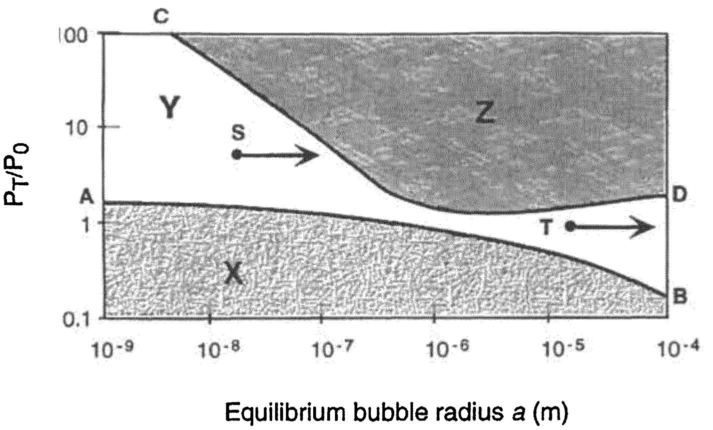

FIG. 4. Schematic representation of the domains of bubble growth via rectified diffusion for air-saturated water at 20 kHz (after Young, 1989). P_T is the peak negative pressure of the incident wave, P_0 is the ambient hydrostatic pressure (assumed to be 100 kPa), and *a* is the equilibrium bubble radius. In the region X delineated by curve A-B, acoustic forces cannot overcome surface tension and bubbles gradually dissolve. In the region Z delineated by curve C-D, bubbles collapse inertially. In region Y, bubbles grow via rectified diffusion until a boundary is crossed (e.g., *S* will grow and eventually collapse as it crosses C-D) or until buoyancy removes the bubble from the sound field (e.g., point *T*).

slowly around some mean size and never collapsing. Upon collapse, however, a bubble is likely to shatter into small fragments that can grow again by rectified diffusion, as shown schematically in Fig. 5.

During inertial cavitation, if a bubble is within one or two equilibrium radii of a solid surface, an asymmetrical collapse can occur. Because of the restriction of liquid flow on the side of the bubble closest to the solid surface, a liquid jet can form on the opposite side that penetrates the bubble and strikes the surface perpendicularly (Fig. 6).

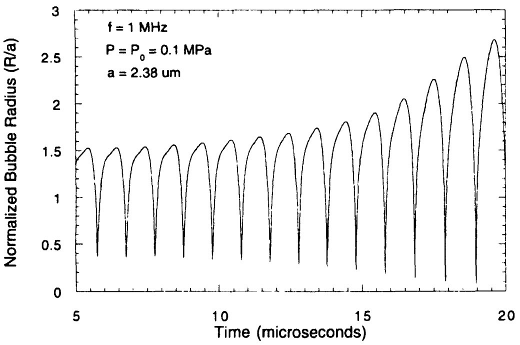

FIG. 5. Numerical calculations for a single CO_2 bubble in water exposed to continuous-wave ultrasound, showing growth via rectified diffusion and subsequent inertial collapse. Instantaneous bubble radius *R* divided by equilibrium radius $a = 2.38\ \mu m$ is plotted versus elapsed time for 1-MHz ultrasound at 0.1 MPa acoustic pressure amplitude (after Church, 1989).

Liquid jets have been observed at velocities of up to 200 m/s (Lauterborn and Bolle, 1975) and are thought responsible for pitting of foils (Coleman *et al.*, 1987), erosion of gallstones in lithotripsy (Vakil and Everbach, 1993), and ultrasonic cleaning of jewelry.

If the collapse of a bubble is symmetrical, the high temperatures, pressures, and shear forces that result can cause chemical and physical changes such as the fusing of metal particles (Doktycz and Suslick, 1990) and the formation of highly reactive free radicals (Riesz and Kondo, 1992), among other effects (see Sec. 6). The production of light (see SONOLUMINESCENCE) may also occur, although the precise mechanism for sonoluminescence is currently a matter of active research (Lepoint *et al.*, 1997).

In 1993, the U.S. Food and Drug Administration, in cooperation with the American Institute of Ultrasound in Medicine and the National Electronics Manufacturer's Association, took account of the potential for adverse bioeffects produced by inertial cavitation with a requirement for new output display on diagnostic ultrasound equipment (FDA, 1993). Diagnostic ultrasound devices must provide an on-screen descriptor of the likelihood of transient cavitation called the mechanical index (MI). The MI is defined as the derated peak rarefactional pressure in megapascals at the point of maximum pulse intensity integral, divided by the square root of the ultrasonic center frequency in megahertz (Meltzer, 1996). This descriptor was

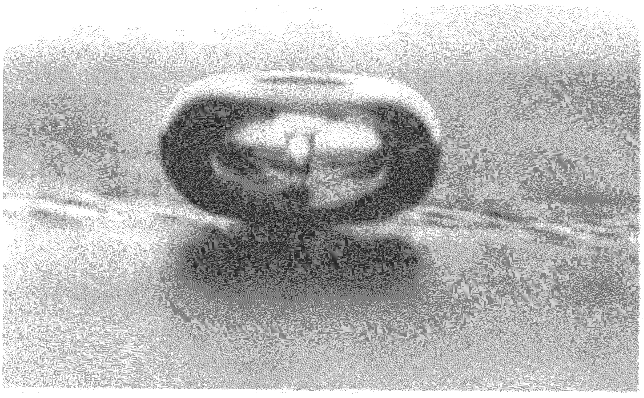

FIG. 6. As a result of the restriction of flow on the side of a collapsing bubble closest to a wall, a liquid jet can form that strikes the surface perpendicularly, causing pitting (from Coleman *et al.*, 1987; by permission).

adopted after reviewing the basic and theoretical research in cavitation inception available at the time (Apfel and Holland, 1991), which assumes that a population of cavitation nuclei of resonance size is always available (see BIOLOGICAL EFFECTS OF SOUND AND ULTRASOUND).

5. SHOCK WAVES

A shock wave in a gas or liquid is defined as a rapid discontinuity in pressure that travels as a wave with a speed faster than the small-signal sound speed of the medium. Shock waves can occur as a result of an abrupt source of pressure (e.g., an explosion), passage of an object traveling faster than the local sound speed (i.e., supersonic motion), or nonlinear wave steepening (see ACOUSTICS, NONLINEAR). Shock waves have physical effects on structures and materials due primarily to the energy they deposit and the forces they exert.

A derivation of the basic equations of shock theory, called the Rankine-Hugoniot shock relations (Lighthill, 1956), yields the following result for weak shocks:

$$v_{\text{shock}} \approx c_0 + \beta\left(\frac{u_1 + u_0}{2}\right), \tag{12}$$

where v_{shock} is the shock propagation speed, c_0 is the small-signal sound speed, $\beta = (\gamma + 1)/2$, γ is the ratio of specific heats of the medium, and u_1 and u_0 are the particle velocities behind the shock front and ahead of it, respectively. The above equation, valid to first order in the "shock strength" u_1/c_0, shows that a weak shock travels with the mean of the wavelet speeds just ahead of it and just behind it. For shocks in air, the spatial thickness of the discontinuity goes inversely as the shock strength, as Table 1 shows (Blackstock, 1982). Many applications of shock waves arise, but the most significant are those related to supersonic flight (sonic booms) and destruction of kidney stones (lithotripsy).

5.1 Sonic Booms

When an object moves faster than the local small-signal sound speed, sounds produced by the object (engine noise, wind turbulence noise) cannot propagate ahead of the object, and a conical shock is formed called a Mach cone. For supersonic aircraft,

Table 1. Shock thickness and SPL as a function of shock strength.

Shock strength u_1/c_0	Shock thickness (mm)	Equivalent SPL (dB)[a]
0.1	0.00524	174
0.01	0.0524	154
0.001	0.524	134
0.0001	5.24	114
0.00001	52.4	94

[a] Sound pressure level of a sinusoidal wave whose dimensionless amplitude is the same as the shock strength.

separate shocks occurring at cross-sectional discontinuities along the length of the aircraft are formed that merge into an "N"-shaped pressure discontinuity at large distances. Dissipation effects such as N_2 and O_2 molecular relaxation are strong at the vertical portions of the N wave, and so the rise time of the shock felt on the ground is strongly influenced by the altitude of the flight. Other important effects include relative humidity, wind turbulence aloft and near the ground, and pre-existing temperature gradients that serve to dephase the shock-wave frequency components.

By changing the length and shape of aircraft and by flying them as high as possible before supersonic flight occurs, the damage caused by sonic booms to windows and other structures on the ground can be minimized. So-called "shaped booms" have pressure signatures that deviate from N waves as a result of more streamlined aircraft designs (Maglieri, 1992). The trend to faster hypersonic flight, however, creates conditions that might reverse these advances. Such factors include a narrower Mach cone that produces more nearly vertical incident-ray angles and the potential for increased ground or water coupling (including shock waves launched into the ocean that are little affected by water depth).

5.2 Lithotripsy

Extracorporeal shock-wave lithotripsy (ESWL) is a medical procedure in which focused underwater shock waves impinge upon a kidney or gall stone, resulting in the pulverization of the stone into fragments that are subsequently passed by the patient. First introduced in the early 1980s (Chaussy *et al.*, 1980), the procedure has been used successfully on hundreds of thousands of patients worldwide. ESWL is considered much less traumatic and less expensive than abdominal surgery since it can be performed on an outpatient basis.

The precise mechanism by which the converging shock wave breaks the stone is unknown. However, stone breakage is thought to be governed either by the internal reflection of the compressional portion of the acoustic pulse (which becomes a tension wave) from the posterior of the stone (Ferrucci, 1989) or by the action of microscopic gas bubbles collapsing onto the stone surface in response to the rarefactional portion of the acoustic pulse (Coleman *et al.*, 1987). Both mechanisms may operate simultaneously or in combination, but each is dependent upon the temporal and spatial characteristics of the lithotripter pulse, as well as the material properties of the concretion (Vakil *et al.*, 1991).

Lithotripters of various designs produce acoustic shock waves using different techniques. Often an initial shock wave is created by an underwater explosion driven by a high-voltage spark discharge (Coleman and Saunders, 1989). The spark-gap electrodes are located at one focus of a hemi-ellipsoidal metal bowl so that the spherically expanding shock wave reflects from the bowl walls and converges at the opposite focus (Fig. 7). This second focus is placed (via ultrasound imaging or fluoroscopy) so as to coincide with the patient's kidney stone. Acoustic pulses can also be created via an electromagnetic device that is analogous to a loudspeaker, or via a piezoelectric crystal (or crystal array), but in these cases the waves produced are nearly sinusoidal and not shock waves. By focusing them with an acoustic lens or by placing the crystals on a spherical backing surface, the converging waves steepen and form shock waves *en route* to the stone. The "shape" of the lithotripter pulse varies from one lithotripter to another but shares the common characteristics of a rapid rise time (from ambient pressure to a maximum of nearly 1000 atmospheres in several nanoseconds) followed by a slower return to ambient pressure (in about 1 μs) and a longer negative tail (up to 200 atmospheres of tension lasting up to 10 μs).

6. SONOCHEMISTRY

Prior to 1940, investigators had established that certain chemical reactions could be accelerated in the presence of high-frequency sound waves. These reactions included the hydrolysis of dimethyl sulfate in basic aqueous solutions, the acceleration of the iodine clock reaction, the decomposition of certain pure organic liquids, the degradation of macromolecules in solution, and the production of free radicals (Riesz and Kondo, 1992). These and other effects are

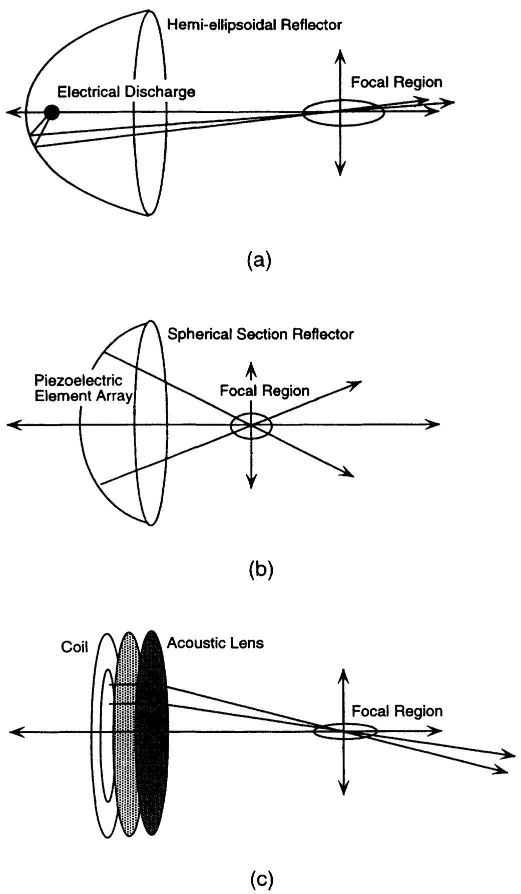

FIG. 7. Lithotripter designs: (a) underwater spark discharge is focused via hemi-ellipsoidal bowl, (b) piezoelectric mosaic mounted on spherical section causes convergence at center, and (c) electrodynamic coil or electrostrictive element launches pulse that is focused via a concave lens (after Coleman and Saunders, 1989).

thought by some researchers (Suslick, 1990; Suslick and Doktycz, 1991) to be due to the momentarily high temperatures (5000 K) and pressures (50 MPa) inside microbubbles undergoing inertial cavitation.

Reactions that take place inside the bubble include typical gas-phase pyrolysis reactions such as sonolysis of aqueous solutions. Unlike combustion pyrolysis, sonolysis can generate free oxygen inside the bubble by thermal dissociation of water, and so no external oxygen source is necessary for certain reactions. For example, in the sonolysis of aqueous solutions containing acetylene and argon, complex hydrocarbons such as benzene and naphthalene can be formed (Hart *et al.*, 1990). Reactions can also take place at the bubble wall in a region extending approximately 200 nm into the surrounding liquid (Suslick *et al.*, 1986), where temperatures reach approximately 2000 K. Here the concentration of OH radicals is the highest, however, and so the thermal decomposition of solutes may be affected by the scavenging of these radicals, depending upon the reaction. Finally, reactions can take place in the liquid away from the bubble wall as free radicals that have not been scavenged in the interfacial region react with solutes there (Riesz *et al.*, 1985). It is also the case that metal powder grains suspended in the solution may have their surface oxide layers removed by bubble jet formation or high-velocity interparticle collisions, and so increase their reactivity or catalytic function (Suslick, 1989).

ACKNOWLEDGMENT

This work was supported in part by a Presidential Faculty Fellowship of the National Science Foundation, MSS 92537777.

GLOSSARY

Cavitation Nuclei: Small pre-existing gas bubbles often stabilized in the cracks of dirt particles or by surfactants.

Derating: The procedure by which an acoustic field measurement in water is corrected to allow for the estimated attenuation caused by ultrasound propagation through tissue.

Deterministic Chaos: The family of characteristics of nonlinear systems that includes the following:

1. the system is governed by deterministic laws (i.e., not random), yet successive states appear random,
2. infinitesimal changes in initial conditions can produce wildly different final states, and
3. the dimensionality of a phase-space plot is of noninteger order.

Free Radicals: Unbound but highly chemically reactive ionic compounds such as OH^-.

Hypersonic: Faster than about Mach 4 (four times the local sound speed).

Mach Cone: The conical shock front, usually originating as a bow wave, from a supersonic object or aircraft.

Nuclear Spins: The rapid rotation of charged atomic nuclei that induces an effective magnetic field.

Particle Velocity: The speed and direction of fluid elements considered sufficiently small so that their properties may be considered uniform within them.

Pyrolysis: The breaking down, or the creation by rearrangement of bonds, of chemical compounds via heating.

Rarefaction: A region of negative gauge pressure and low density, in which the molecules of a medium are momentarily farther apart from one another than they are at equilibrium.

Resonance Size: The radius of a bubble at which an incident sound wave will cause maximum radial displacement oscillations.

Sonolysis: The breaking down, or the creation by rearrangement of bonds, of chemical compounds via action of ultrasound through the mechanism of inertial cavitation.

Wavelet: A single-frequency component of a wave, having infinitesimal amplitude.

Works Cited

Apfel, R. E., Holland, C. K. (1991), "Gauging the Likelihood of Cavitation from Short-Pulse, Low-Duty Cycle Diagnostic Ultrasound," *Ultrasound Med. Biol.* **17** (2), 179–185.

Atchley, A. A., Crum, L. A. (1988), "Acoustic Cavitation and Bubble Dynamics," in: K. S. Suslick (Ed.), *Ultrasound, Its Chemical, Physical, and Biological Effects,* New York: VCH Publishers, Inc., pp. 1–64.

Barker, L. M. (1972), "Laser Interferometry in Shock-Wave Research," *Exper. Mech.* **12** (40), 209–215.

Barker, L. M., Hollenbach, R. E. (1970), "Shock Wave Studies of PMMA, Fused Silica, and Sapphire," *Appl. Phys.* **41,** 4208–4226.

Blackstock, D. T. (1982), Notes on Nonlinear Acoustics ME/EE 384N.4, University of Texas at Austin, Austin, TX. Issued as part of an Acoustical Society of America Short Course on Nonlinear Acoustics, May 1993.

Bolef, D. I., Sundfors, R. K. (1992), "Continuous-Wave Ultrasonics with Applications to Acoustic Magnetic Resonance: A Review," in: Y. Fu, R. K. Sundfors, P. Auntharothok (Eds.), *Perspectives in Physical Acoustics,* Singapore: World Scientific Publ.,

Chaussy, C., Brendel, W., Schmiedt, E. (1980), "Extracorporeally Induced Destruction of Kidney Stones by Shock Waves," *Lancet* **20,** 1265–1268.

Church, C. C. (1988), "Prediction of Rectified Diffusion during Nonlinear Bubble Pulsations at Biomedical Frequencies," *J. Acoust. Soc. Am.* **83,** 2210–2217.

Church, C. C. (1989), "A Theoretical Study of Cavitation Generated by an Extracorporeal Shock Wave Lithotripter," *J. Acoust. Soc. Am.* **86,** 215–227.

Coleman, A. J., Saunders, J. E. (1989), "A Survey of the Acoustic Output of Commercial Extracorporeal Shock Wave Lithotripters," *Ultrasound Med. Biol.* **15** (3), 213–227.

Coleman, A. J., Saunders, J. E., Crum, L. A., Dyson, M. (1987), "Acoustic Cavitation Generated by an Extracorporeal Shock-Wave Lithotripter," *Ultrasound Med. Biol.* **13** (2), 69–76.

Cramer, E., Lauterborn, W. (1981), "Subharmonic Route to Chaos Observed in Acoustics," *Phys. Rev. Lett.* **47,** 1445–1448.

Doktycz, S. J., Suslick, K. S. (1990), "Interparticle Collisions Driven by Ultrasound," *Science* **247,** 1067–1070.

Eller, A. I. (1970), "Damping Constants of Pulsating Bubbles," *J. Acoust. Soc. Am.* **47,** 1469–1470.

Everbach, E. C. (1989), "Tissue Composition Determination via Measurement of the Acoustic Nonlinearity Parameter," Ph.D. Dissertation, New Haven, CT, Yale University, p. 49.

Ferrucci, J. T. (1989), "Biliary Lithotripsy—A Review," *J. Lithotripsy Stone Dis.* **1** (1), 3–15.

Food and Drug Administration (1993), *Diagnostic Ultrasound Guidance for 1993,* Revised 510(k), Rockville, MD: Center for Devices and Radiologic Health.

Hart, E. J., Fischer, Ch.-H., Henglein, A. (1990), "Pyrolysis of Acetylene in Sonolytic Cavitation Bubbles in Aqueous Solution," *J. Phys. Chem.* **94,** 284–290.

Henry, G. E. (1957), "Radiation Pressure," *Sci. Am.* **196** (6), 99–108.

Holland, M. R., Sundfors, R. K. (1985), *J. Phys. (Paris) Colloque. C10* **46,** 779.

Kinsler, L. E., Frey, A. R., Coppens, A. B., Sanders, J. V. (1982), *Fundamentals of Acoustics,* New York: Wiley.

Lauterborn, W., Bolle, H. (1975), "Experimental Investigations of Cavitation-Bubble Collapse in the Neighbourhood of a Solid Boundary," *J. Fluid Mech.* **72,** 391–399.

Leighton, T. G. (1994), *The Acoustic Bubble,* London: Academic, p. 137.

Lepoint, T., De Pauw, D., Lepoint-Mullie, F. (1997), "Sonoluminescence: An Alternative 'Electrohydrodynamic' Hypothesis," *J. Acoust. Soc. Am.* **101,** 130–152.

Levy, M. (1992), "Electron-Phonon, Acousto-Electric and Magneto-Elastic SAW Interactions with Superconducting Films, 2-D Electron Gas or Magnetic Films," in: Y. Fu, R. K. Sundfors, P. Auntharothok (Eds.), *Perspectives in Physical Acoustics,* Singapore: World Scientific Publ.

Lighthill, M. J. (1956), "Viscosity Effects in Sound Waves of Finite Amplitude," in: G. K. Batchelor (Ed.), *Surveys in Mechanics,* London: Cambridge Univ. Press, pp. 250–351.

Maglieri, D. J. (1992), "Overview of Current Knowledge and Activity in Sonic Boom Research," *J. Acoust. Soc. Am.* **92,** 2328(A).

Meltzer, R. S. (1996), "Food and Drug Administration Ultrasound Device Regulation: The Output Display Standard, the Mechanical Index, and Ultrasound Safety," *J. Am. Soc. Echocardiol.* **9,** 216–220.

Miller, D. L., Williams, A. R. (1983), "Further Investigations of ATP Release from Human Erythrocytes Exposed to Ultrasonically Activated Gas-Filled Pores," *Ultrasound Med. Biol.* **9,** 297–307.

Minnaert, M. (1933), "On Musical Air-Bubbles and Sounds of Running Water," *Philos. Mag.* **16,** 235–248.

NCRP (1983), "Biological Effects of Ultrasound: Mechanisms and Clinical Implications," Report No. 74, National Council on Radiation Protection and Measurements, Bethesda, MD.

Nyborg, W. L. (1965), "Acoustic Streaming," in: W. P. Mason (Ed.), *Physical Acoustics. Principles and Methods.* Vol. 2, pt. B, New York: Academic, p. 265.

Parlitz, U., Englisch, V., Scheffczyk, C., Lauterborn, W. (1990), "Bifurcation structure of bubble oscillators," *J. Acoust. Soc. Am.* **88,** 1061–1077.

Riesz, P., Kondo, T. (1992), "Free Radical Formation Induced by Ultrasound and Its Biological Implications," *Free Rad. Biol. Med.* **13,** 247–270.

Riesz, P., Berdahl, D., Christman, C. L. (1985), "Free Radical Generation by Ultrasound in Aqueous and Non-aqueous Solutions," *Environ. Health Perspect.* **64,** 233–252.

Straudenraus, J., Eisenmenger, W. (1993), "Fibre-Optic Probe Hydrophone for Ultrasonic and Shock-Wave Measurements in Water," *Ultrasonics* **31,** 267–273.

Suslick, K. E. (1989), "The Chemical Effects of Ultrasound," *Sci. Am.* **260** (2), 80–86.

Suslick, K. S. (1990), "Sonochemistry," *Science* **247,** 1439–1445.

Suslick, K. S., Doktycz, S. J. (1991), "Effects of Ultrasound on Surfaces and Solids," in T. J. Mason (Ed.), *Advances in Sonochemistry,* Vols. I and II, Greenwich, CT: JAI Press, p. 199.

Suslick, K. S., Hammerton, D. A., Cline, R. E., Jr. (1986), "The Sonochemical Hot Spot," *J. Am. Chem. Soc.* **108,** 5641–5642.

Swift, G. W. (1988), "Thermoacoustic Engines," *J. Acoust. Soc. Am.* **84** (4), 1145–1180.

Vakil, N., Everbach, E. C. (1993), "Transient Acoustic Cavitation in Gallstone Fragmentation: A Study of Gallstones Fragmented *in vivo,*" *Ultrasound Med. Biol.* **19,** 331–342.

Vakil, N., Gracewski, S. M., Everbach, E. C. (1991), "Relationship of Model Stone Properties to Fragmentation Mechanisms during Lithotripsy," *J. Lithotripsy Stone Dis.* **3,** 304–310.

Williams, A. R., Miller, D. L., Gross, D. R. (1986), "Haemolysis *in vivo* by Therapeutic Intensities of Ultrasound," *Ultrasound Med. Biol.* **12** (6), 501–509.

Wurster, C., Staudenraus, J., Eisenmenger, W. (1994), "The Fiber Optic Probe Hydrophone," 1994 in: M. Levy, B. R. McAvoy, (Eds.), *Proceedings of the 1994 IEEE Ultrasonics Symposium,* Vol. 2, Piscataway, NJ: IEEE.

Yadav, H. S., Murty, D. S., Veerma, S. N., Sinha, K. H. C., Gupta, B. M., Chand, D. (1973), "Measurement of Refractive Index of Water under High Dynamic Pressures," *J. Appl. Phys.* **44,** 2197–2200.

Further Reading

Fu, Y., Sundfors, R. K., Auntharothok, P. (Eds.) (1992), *Perspectives in Physical Acoustics,* Singapore: World Scientific Publ. (ISBN 981-02-0908-8).

Kinsler, L. E., Frey, A. R., Coppens, A. B., Sanders, J. V. (1982), *Fundamentals of Acoustics,* New York: Wiley.

Leighton, T. G. (1994), *The Acoustic Bubble,* London: Academic (ISBN 0-12-441920-8).

Suslick, K. S. (Ed.) (1988), *Ultrasound, Its Chemical, Physical, and Biological Effects,* New York: VCH Publishers, Inc.

Young, F. R. (1989), *Cavitation,* London: McGraw-Hill Book Company (ISBN 0-07-707094-1).

ULTRAVIOLET SPECTROSCOPY

See Supplement Volumes

UNDERWATER ACOUSTICS

See ACOUSTICS, UNDERWATER

UNDERWATER AND ATMOSPHERIC ACOUSTIC SIGNAL PROCESSING

DAVID HAVELOCK, *National Research Council of Canada, Ottawa, Ontario, Canada*

WILLIAM A. KUPERMAN, *Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California, U.S.A.*

ARTHUR BAGGEROER, *Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.*

Introduction 539
1. The Acoustic Environment .. 540
2. Signal Acquisition 543
2.1 Sensors 543
2.2 Signals 543
2.3 Spatial Distribution 544
2.4 Noise and Distortion Characteristics 544
3. Time-Series Processing 545
3.1 Complex-Signal Reconstruction 545
3.1.1 *I-Q* Demodulation of Narrow-Band Signals 545
3.1.2 Heterodyne Demodulation of Narrow-Band Signals 546
3.1.3 Analytic Reconstruction of Complex Signals 546
3.2 Narrow-Band Signal Filtering 547
3.2.1 FIR and IIR Filters 547
3.2.2 Processing Long Records 548
3.2.3 Efficient Baseband Signal Extraction 548
3.3 Zoom Filtering 548
3.4 Power Spectral Density Estimation 549
4. Array Processing 551
4.1 Linear Plane-Wave Beam Forming and Spatiotemporal Sampling 551
4.1.1 Frequency-Domain Processing 551
4.1.2 Time-Domain Processing 553
4.1.3 Some Properties of Array Processors 553
4.2 Adaptive Processors 554
4.2.1 Minimum-Variance Distortionless Processor 554
4.2.2 Eigenvectors and Beam Formers 555
4.3 Matched-Field Processing 556
5 Active Systems 559
5.1 Active Sonar Signal Processing 559
5.2 Communications and Seabed Mapping 561
5.3 Travel-Time Tomography 562
Glossary 562
Works Cited 564
Further Reading 564

INTRODUCTION

Underwater acoustics and atmospheric acoustics have many similarities. Both of these acoustic environments are often characterized by multipath or waveguide-type propagation (see ACOUSTICS, UNDERWATER; ATMOSPHERIC ACOUSTICS); the frequencies of interest are in the same regime, from a few hertz to tens of kilohertz; the media have many similar and analogous properties, such as stratification, currents or winds, internal waves, etc.; high-frequency attenuation results in band-limited propagation. Therefore, it is convenient to address many of the signal-processing issues for underwater and atmospheric acoustics in the same place.

In this article we discuss both temporal and spatial signal processing. We begin with a brief description of the atmospheric and underwater acoustic environment. Data acquisition and temporal and spatial processing are discussed with emphasis on ultimately taking advantage of digital signal processing. Both passive and active acoustic systems are treated. This article is a brief overview of a very large field. The references, therefore, provide much more detail.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1. THE ACOUSTIC ENVIRONMENT

Both the ocean and the atmosphere are nonhomogeneous fluid media in which the same general principles for acoustic sensing and signal processing apply, although differing boundary conditions, media parameters, and applications lead to variations in approaches and techniques. Descriptions of the media can be found in Clay and Medwin (1977) for the ocean and in ATMOSPHERIC ACOUSTICS and Panofsky and Dutton (1984) for the atmosphere. Characteristics that affect signal processing for the media are summarized here. For underwater acoustic propagation prediction models, see Jensen *et al.* (1994), and for a set of benchmark tests for propagation models in the atmosphere (excluding turbulence), see Attenborough *et al.* (1995).

Except in close proximity to the source, most sounds in the atmosphere and the ocean cause only small disturbances in the medium and propagate according to the linear wave equation for inhomogeneous media,

$$\rho\nabla\cdot\left(\frac{1}{\rho}\nabla p\right) - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = 0,$$

where p is the acoustic pressure, c is the sound speed, and ρ is the medium density. Both ρ and c may vary with position and time. For signals in this linear regime, particularly where moving sources and targets are concerned, sound propagation is treated as a time-varying, linear system. Higher-pressure sound waves in the nonlinear regime also occur, such as from explosions or sonic booms. They cause significant disturbances in the media, such as cavitation underwater, and require nonlinear models (see ACOUSTICS, NONLINEAR); however, nonlinear disturbances generally decay rapidly into the linear regime as they propagate.

Sound pressure levels are expressed in decibels relative to 1 μPa underwater and 20 μPa in the atmosphere. These different references result in sound pressure levels underwater being specified numerically as 26 dB above the equivalent level in the atmosphere. Sound pressure levels reported in the atmosphere are often weighted with frequency to account for human hearing characteristics. In the atmospheric acoustics literature, unweighted measurements are often referred to as being linearly weighted.

The acoustic energy flux, or acoustic intensity, for a plane wave is $\mathbf{I} = \mathbf{n}p^2/\rho c$, where $\mathbf{n}$ is a unit vector in the direction of propagation, p is the acoustic pressure, ρ is density (approximately 1.2 kg/m^3 for the atmosphere and 1000 kg/m^3 for the ocean), and c is the speed of sound (approximately 340 m/s for the atmosphere and 1500 m/s for the ocean). Accordingly, a plane wave with a given pressure amplitude will have an intensity in the atmosphere that is about 3500 times greater than that in the ocean.

A high sound pressure level within the linear regime in the atmosphere, such as a rocket launch measured at a few hundred meters, may exceed 140 dB re 20 μPa, linearly weighted. Background levels are often in the range of 40–60 dB re 20 μPa, linearly weighted. Background levels tend to increase at lower frequencies, as can be seen in Fig. 1, which shows an example of a background noise spectrum near the ground with a sound pressure level (linearly weighted) of about 60 dB re 20 μPa. Microphones outdoors must be protected by windscreens to reduce the noise generated at the sensor by wind. Rainfall generates high noise levels, making acoustic sensing difficult.

In the ocean, sound pressure levels are usually not weighted with frequency. A high

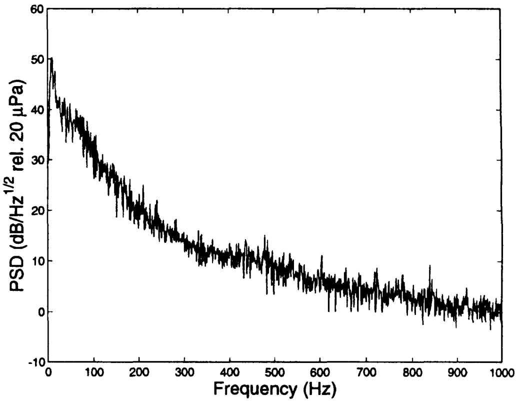

FIG. 1. An example power spectral density (PSD) for background noise in the atmosphere. The recordings were made near a hard ground in a rural environment. The PSD increases significantly at lower frequencies, although it declines below 10 Hz in the figure because of attenuation by the recording system.

sound pressure within the linear regime in the ocean, such as might be generated by sonar, is greater than 220 dB re 1 μPa. Background noise levels are of the order of 60–90 dB re 1 μPa, depending on frequency. As in the atmosphere, background noise level increases at lower frequencies.

Sound absorption loss is due primarily to molecular relaxation mechanisms (Piercy *et al.*, 1986), which involve the rate of exchange of energy between molecular states. In the atmosphere, there are two main relaxation mechanisms, which are catalyzed by water vapor and involve mechanical vibrational modes of O_2 and N_2 molecules. In the ocean, there are two chemical relaxation mechanisms, involving magnesium sulfate and boric acid. These absorption mechanisms exhibit a "resonance" sort of behavior, which results in humps in the spectral absorption curves. The locations of these humps are determined by the relaxation frequencies. While these frequencies change significantly for the atmosphere, because they are highly dependent upon the water-vapor content, they are relatively constant for the ocean. Figure 2 shows example sound absorption curves in the ocean and the atmosphere. The lower absorption of the ocean makes it possible to sense acoustic signals at much greater ranges than in the atmosphere. For long-range propagation in the atmosphere and in the ocean, frequencies of interest are constrained below about 1000 Hz on account of absorption. Although natural acoustic sources extend well into the infrasonic region, the greatest interest lies in the frequency ranges 10–1000 Hz because of the abundance of man-made and biological sources in this frequency range.

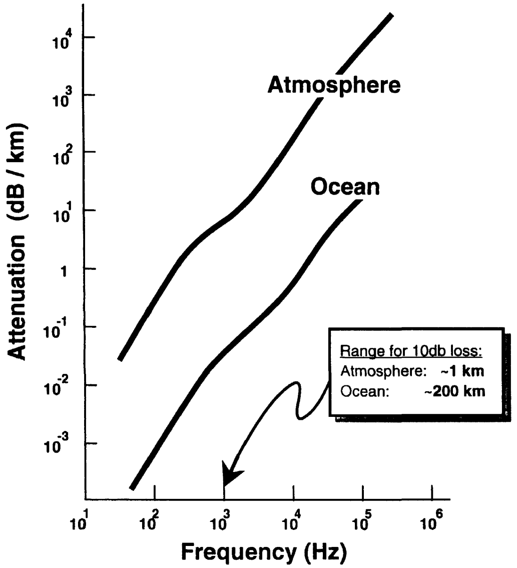

FIG. 2. Example absorption curves for sound propagating in the ocean and the atmosphere.

The speed of sound is nominally 1500 m/s in the ocean and 340 m/s in air; hence wavelengths are longer underwater, and larger arrays are needed for spatial processing.

The density, pressure, temperature, wind or currents, and salinity in the case of the ocean change with depth or height resulting in sound-speed variations that are expressed by *sound-speed profiles.* Typical sound-speed profiles for the ocean with their geographical disposition and a schematic of resulting ocean paths are shown in Figs. 1–3 of ACOUSTICS, UNDERWATER. An example of sound-speed profiles and propagation paths for the atmosphere are shown in Pierce (1989). Of particular interest is the existence of features such as a *surface duct,* the *deep-sea duct, acoustic shadow regions,* and periodic *convergence zones* in both the underwater and atmospheric sound propagation.

A surface duct is formed when sound is refracted toward the reflecting surface, thereby constraining propagation within the duct. Similarly, the deep-sea duct (or upper atmosphere duct) refracts sound back into the duct from its upper and lower extremities. Ducts may be cylindrically (azimuthally) symmetrical, as is typically the case underwater, resulting in cylindrical spreading, or they may be asymmetrical, as in the case of wind-induced profiles in the atmosphere. Sound pressure generally varies with range R as R^{-2} (spherical spreading), but when propagating through a duct, it varies as R^{-1} (cylindrical spreading).

For acoustic line-of-sight propagation there is at least one acoustic ray path between the source and receiver. For a receiver located within an acoustic shadow region, there are no ray paths between the source and receiver. Acoustic shadow regions can be generated by refraction patterns, often at a surface interface, or by masking due to large surface topographical features. At the edge of an acoustic shadow region, sound energy is present as a result of diffraction from the shadow boundary. Full wave propagation models are used to predict the sound field in these regions since the simpler ray-path models do not admit a ray path between the source and a receiver within an acoustic shadow region.

Deeper within the shadow region, sound energy is predominantly due to scattering from the turbulence in the surrounding ensonified regions. Sound pressure levels in acoustic shadow regions in the atmosphere near the ground can exhibit weak range dependence.

Convergence zones are regions of enhanced sound pressure levels at range intervals along a duct. They are caused by the coincidence, or focusing, of ray paths that are radiated (or launched) from the source at different elevation angles. Sound pressure levels in these convergence zones are well above that expected for cylindrical spreading and are determined by the shape of the sound-speed profile and the homogeneity of the medium.

Sound-speed profiles change considerably because of diurnal, seasonal, and weather system fluctuations, particularly near the ground in the atmosphere and near the surface underwater. An example of sound-speed profile fluctuations near the ground, and the corresponding short-range sound propagation, can be seen in L'Esperance *et al.* (1995).

Range dependence in the sound-speed profile complicates the structure (and prediction) of a propagating sound field. Atmospheric propagation and shallow-water propagation both exhibit significant range dependence in the sound-speed profiles. Winds in the atmosphere have a greater effect on propagation than do currents in the ocean.

Turbulence introduces random fluctuations in the index of refraction and fluctuations in the local velocity of the medium. These random perturbations generally moderate sound-level extremes that would be present in a nonturbulent, deterministic environment; sound levels in null regions, where interfering propagation paths would normally result in very low levels, are raised somewhat, and sound levels in regions of reinforcement, where constructive interference or convergence of ray paths would normally increase levels, are decreased somewhat.

Turbulence also distorts the propagating wave fronts and can cause mixing between propagation modes. In the presence of turbulence the sound field becomes a stochastic function of time and space with finite temporal and spatial coherence, limiting the achievable array processing gains. The effects of turbulence are more significant in the atmosphere than they are in the ocean. In the atmosphere, at frequencies of a few hundred hertz, temporal coherence is on the order of seconds and spatial coherence may be less than a meter (Daigle *et al.*, 1983; Havelock *et al.*, 1995). Coherence is greater in the ocean.

The fluctuations in the phase and magnitude of a pure tone propagating through turbulence increase with range, but the magnitude fluctuations reach a saturation level of approximately 6 dB at some distance, beyond which the magnitude fluctuations do not increase further. A 1000-Hz signal propagating in the atmosphere may reach saturation within a kilometer, whereas saturation will not be reached for hundreds of kilometers in the open ocean. Phase fluctuations, rather than magnitude fluctuations, are usually the main cause of spatial and temporal decorrelation of the sound field.

Acoustic sensing in the atmosphere is usually done near the ground, where the variable ground impedance, topology of the terrain, strong refraction, and surface structures such as buildings or forests play an important role. The frequent occurrence of acoustic shadow regions near the ground further complicates signal detection.

Deep-ocean measurements have a more stable environment than the atmosphere, but measurement near the ocean surface must contend with varying surface conditions, while measurements near the ocean floor or in shallow water must contend with a complex seafloor impedance and seafloor topography. The ocean seafloor typically consists of a sediment layer overlying a hard basement. Both types of materials are elastic media, which support shear waves as well as compressional waves. The shear-wave speed in sediments is typically of the order of 100 m/s (very slow). The existence of this wave acts as an additional degree of freedom to transport sound out of the water column; hence, it appears as a loss mechanism. On the other hand, the rock basement often has a shear-wave speed greater than the speed of sound in water and therefore acts as a good reflector (Jensen *et al.,* 1994). These layered structures combine to have complicated frequency-dependent properties. Accurate knowledge of the seafloor is difficult and expensive to ascertain but is important for many signal-processing applications.

2. SIGNAL ACQUISITION

2.1 Sensors

Signal acquisition begins at an acoustic transducer inserted into a sound field. Microphones (Wong and Embleton, 1995), used in the atmosphere, and hydrophones, used underwater, are capable of measuring wide ranges of signal frequencies and amplitudes; nevertheless, with acoustic signals of interest ranging over nine orders of magnitude in frequency and amplitude, the appropriate sensor type must be selected for the desired signals. Microphones can range in size from millimeters, for ultrasonic or insertion-type microphones, to several meters and beyond, for infrasonic microphones. In general, greater sensitivity and lower-frequency response is achieved with larger sensors.

In most cases, acoustic sensors are small relative to the signal wavelength and are considered to sample the sound field at a point. Acoustic sensors can measure the instantaneous acoustic amplitude, rather than just the envelope or intensity as is often the case with optical sensors. This capability allows multiple independent sensors to be used to construct an array that samples the sound field over a spatially distributed region and integrates the results coherently to obtain gains in sensitivity and directionality. The processing of signals gathered by such arrays occupies a large proportion of the subject of underwater and atmospheric acoustic signal processing.

2.2 Signals

A wide variety of signal sources are of interest in both underwater and atmospheric applications. The signals are broadly categorized as impulsive or continuous wave. Continuous-wave (cw) signals have fairly constant power levels over the signal monitoring interval and are generally easier to generate and detect than are impulsive signals. The term continuous wave is usually interpreted to mean pure sinusoidal tones, which have been studied extensively for detection, localization, and classification problems. Manmade equipment readily produces continuous-wave signals because of the periodicity of reciprocating mechanical parts.

Impulsive signals are characterized by short periods of high power interspersed with long periods of little or no signal power. The classical impulse signal is the Dirac delta function, which is not physically realizable. There are practical difficulties in generating and receiving pulses, due to the high instantaneous power required to obtain adequate signal-to-noise ratio. Active sonar systems utilize the high temporal resolution of impulsive signals to perform range estimation. Narrow-band short pulses, consisting of a tone that is amplitude modulated by short-duration pulses, are used for this purpose. Short pulse duration corresponds to wide frequency bandwidth, making pulses useful for examining system response. The time-varying environments of the ocean and the atmosphere are often probed with impulsive signals to determine the response at many frequencies simultaneously, before the environment changes.

Swept tones, which are generated by progressively increasing or decreasing the frequency of a pure-tone generator, are sometimes used to measure system response. These signals can provide high power in narrow frequency bands, but they have the disadvantage that the component tones do not propagate simultaneously. Periodic sequences, which are composed of combinations of harmonics of the repetition frequency, are better suited for this purpose. A repeated pulse is one example of such signals, and a repeated tone burst is another. Tone bursts are regularly spaced intervals of pure sinusoidal signal interspersed with silence.

An important parameter of test signals is the peak factor (or crest factor), which is a measure of the variations in the (instantaneous) power. It is defined as the ratio of the maximum signal amplitude to the rms signal amplitude. Techniques have been developed to adjust the phase and magnitude of harmonics so as to minimize the signal peak factor. (See Burgess, 1992 and the references therein.)

Maximum-length (ML) sequence signals can be efficiently generated by a feedback shift register (Kramer, 1965) and are well known for their low peak factor and constant power spectra. An ML sequence is a finite sequence of numbers, generated according to the principles of the theory of cyclic groups, with an autocorrelation function that approximates a delta function. Repetition of such a sequence results in a spectrum composed of the harmonics of the repetition frequency, and the envelope of the harmonics is reasonably flat. Binary ML sequences, which are widely used, are obtained from the high-order bit of the feedback shift register.

2.3 Spatial Distribution

Coherent detection with an array of sensors refers to adjusting the phase (and possibly the magnitude) of the signal received at each sensor so that they synchronously receive a specific signal. The spatial distribution of the signal sound field must be modeled for coherent detection. In many applications, when the source is a long distance from the receiver, the signal is modeled as a plane wave arriving at the receiver from a specific direction. Other models, based on ray propagation or modes, may be used in more complicated environments.

The sound field close to a source deviates significantly from a plane or spherical wave on account of nonpropagating modes arising from the specific geometry of the source. The *far field* refers to ranges sufficiently great that the sound field resembles the desired model (spherical spreading or plane-wave propagation, for example). Since the near field is complicated, acoustic measurement of the sound power output of a source is often made in the far field and the reported source levels are referred back to a standard range, such as 1 m, on the basis of a simple propagation model.

Refraction by the medium, reflection from boundaries, and diffraction from discontinuities can cause *caustics* and *nulls* in the sound field. Caustics are regions of intensification due to constructive interference of propagating wave fronts, and nulls are regions of rarefaction due to destructive interference of the wave fronts. An array of sensors spanning a significant portion of a null or caustic region will observe a sound field that may be significantly distorted from a plane wave. This problem is more acute in the atmosphere than underwater since wavelengths are shorter and the environment is generally more complicated. Modeling the received sound field under such conditions can be difficult.

Various models and computational implementations have been developed for sound propagation underwater (Jensen *et al.*, 1994) and in the atmosphere (Attenborough *et al.*, 1995; Gilbert and Di, 1993), for deterministic environments. Methods for including stochastic properties of the medium are only now beginning to emerge.

2.4 Noise and Distortion Characteristics

For signal-processing applications, noise is often modeled as a temporally uncorrelated, Gaussian process $n(t)$; that is, the random variable $n(t)$ has a Gaussian distribution for each t and the correlation $R_n(\tau) = E\{n(t)n(t + \tau)\}$ is zero for $\tau \neq 0$, where $E\{\cdot\}$ indicates the expected value in the ensemble sense. Such noise is referred to as *white Gaussian noise* because its power spectral density, given by the Fourier transform of $R_n(\tau)$, is flat in much the same way as white light is considered to contain all colors equally. When noise is temporally correlated, so that $R_n(\tau) \neq 0$ for some $\tau \neq 0$ and its spectrum is not flat, it is referred to as *colored noise.* The typical noise spectrum, both underwater and in the atmosphere, increases at lower frequencies, suggesting that the noise should be considered colored; however, in practical terms, when the signal bandwidth is narrow, the approximation to white noise within the signal bandwidth is often adequate.

If all statistics of the noise are independent of time (although they may depend upon time differences such as τ in the autocorrelation mentioned above), the noise is said to be *stationary,* or *strict-sense stationary.* The noise is *wide-sense stationary* if the mean and temporal correlations (but not necessarily other statistics) of the noise are independent of time. Noise in the ocean and in the atmosphere is not stationary, and some form of dynamic noise modeling may be required, depending upon the signal duration. The problem may be further complicated by the presence of impulsive noise sources (for example, from biological sources such as insects or marine mammals).

A noise field that has the same statistical distribution at all points in space is said to be *homogeneous*. When statistics of the noise are independent of the orientation of the frame of reference, the noise is said to be *isotropic*. In practical situations, the noise can often be assumed to be locally homogeneous but not necessarily isotropic. Knowledge of the noise anisotropy can be utilized to improve discrimination between signals and noise when using arrays of sensors.

Noise is usually considered to be additive, in that the received signal is the sum of a desired signal and a noise signal. However, there are various forms of signal distortion that are not additive. One example is *clipping*, which is the limiting of the system response to a maximum level for input signal levels that exceed the maximum receiver response. Another example of signal distortion is phase or amplitude modulation such as may be caused by sensor motion or variations in sensitivity. Signal perturbations due to turbulence are also a source of signal distortion.

Noise that is tonal in nature may be considered to be a jamming, or interfering signal. It is often treated in the spectral domain, where it can be localized within a small frequency band. Correspondingly, jamming that is impulsive may be treated best in the time domain. Repetitive impulsive noises may be treated with a combination of spectral and temporal processing. Spatial processing (beam forming) can be very effective for eliminating jamming. Reverberation is a signal-generated noise, or distortion, that is periodic in the spectral domain. It can be generated by the scattering of a probe signal from inhomogeneities in the medium and is a major cause of interference.

Noise sources underwater and in the atmosphere are discussed in ACOUSTICS, UNDERWATER and ATMOSPHERIC ACOUSTICS, respectively.

3. TIME-SERIES PROCESSING

The broad subject of digital signal processing for time-series data is treated in textbooks such as Marple (1987) Oppenheim and Schafer (1989), and Rabiner and Schafer (1978). Central to this subject is the discrete Fourier transform, implemented with a fast Fourier-transform (FFT) algorithm. Many of the tools used for signal processing of time series from underwater and atmospheric acoustics are generic, in that they are also used in other application areas. Prominent among the techniques that are more specific to our application are those that deal with constructing a complex-signal representation, with extracting narrow-band signals, and with reducing the large volumes of recorded time-series data that contain considerable amounts of noise power. Representative techniques for accomplishing these tasks are discussed in this section.

3.1 Complex-Signal Reconstruction

Analytical techniques of signal processing are often presented in the elegant framework of complex-signal theory. Acoustic sensors and digital data-acquisition equipment respond to the signal frequencies of interest to provide measured signals that are real valued, rather than complex. For direct implementation of the theoretical principles, it is common practice to reconstruct the complex acoustic signal from the measured real signal. While there is no perfect method for complex-signal reconstruction in practice, there are various techniques that are quite accurate if done correctly.

The simplicity and importance of the reconstructed complex signal is illustrated by an example in Fig. 3, which compares a real-valued pulse signal with its complex reconstruction. The location of the pulse peak is an important signal parameter that is more clearly determined in the complex signal than it is in the real signal.

3.1.1 *I-Q* Demodulation of Narrow-Band Signals The received signal $s_r(t)$ is real valued and often consists of a pure tone, called a carrier frequency, modulated by the signal of interest $x(t) = a(t)\exp[i\phi(t)]$:

$$\begin{aligned} s_r(t) &= \mathrm{Re}[x(t)\exp(2\pi i f_c t)] && (1)\\ &= a(t)\cos[2\pi f_c t + \phi(t)] && (2)\\ &= x_r(t)\cos(2\pi f_c t) - x_q(t)\sin(2\pi f_c t), && (3)\end{aligned}$$

where the bandwidth of $x(t)$ is assumed to be less than the carrier frequency. Equation (1)

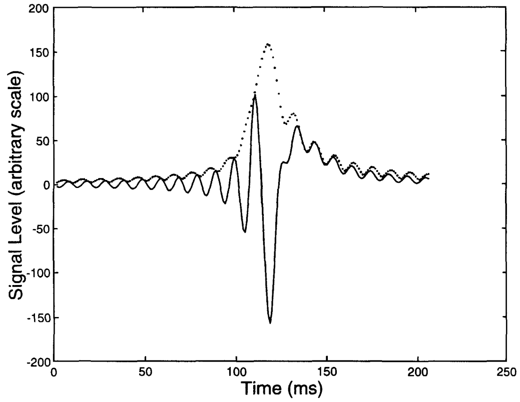

FIG. 3. The real-valued signal of a band-limited, distorted pulse (solid line) and the magnitude of the reconstructed complex signal (dotted line). The peak of the pulse is more easily located using the reconstructed signal.

is the *complex modulation* representation of the signal, Eq. (2) is the *envelope-phase* representation, and Eq. (3) is the *quadrature-carrier* representation. The frequency f_c is the carrier (center) frequency of the modulated signal $x(t)$. In the quadrature-carrier representation, the signal $x(t) = x_r(t) + ix_q(t)$ is the *complex envelope.* Interest in this signal characterization arises from a pure tone propagating through a perturbing ocean or atmosphere. In this situation, the perturbation $x(t)$, rather than the original carrier signal $\exp(2\pi i f_c t)$, is the main object of interest.

The real part of the complex envelope can be obtained by *I-Q* demodulation. In this method, a *quadrature filter* or *90° phase shifter* is first applied to the real received signal $s_r(t)$ to obtain the signal quadrature component $s_q(t)$, which is 90° behind the received signal, and then $x_r(t)$ is obtained using

$$x_r(t) = s_r(t)\cos(2\pi f_c t) + s_q(t)\sin(2\pi f_c t). \tag{4}$$

The accuracy of this method depends upon the accuracy in reconstructing $s_q(t)$, and on accurate knowledge of the value of f_c. The quadrature filter or a 90° phase shifter typically introduces a small error in either the magnitude or the phase, respectively (Bellanger, 1984, Chap. 9). Furthermore, the phase of the extracted signal $x(t)$ or $x_r(t)$ will depend upon the phase of the carrier signal at $t = 0$; we have assumed in the preceding that the carrier is of the form $\exp[2\pi i f_c(t + t_0)]$, where $t_0 = 0$. If the value of t_0 is shifted by $1/4f_c$, then the so-called in-phase component $s_r(t)$ and the quadrature component $s_q(t)$ are exchanged (with a sign reversal).

3.1.2 Heterodyne Demodulation of Narrow-Band Signals Equation (4) may be interpreted as the real part of the product of a complex version of the received signal, $s(t) = s_r(t) + is_q(t)$, and the complex exponential $\exp(-2\pi i f_c t)$. Multiplying by an exponential in this way is referred to as *heterodyning.* Heterodyning shifts the carrier frequency of the signal, and when the carrier frequency is shifted to dc (0 Hz), the process is called *homeodyning.* The bandwidth ω_x of $x(t)$ is assumed to be less than the center frequency f_c of the carrier frequency. The term *baseband* refers to the frequency band containing the signal of interest; thus, the baseband of $s(t)$ is $f_c \pm \omega_x/2$.

The mathematical description of the heterodyne process given above involves complex signals. Heterodyning can also be accomplished by multiplying real-valued signals by a real-valued sinusoid modulator, as is done in radio communications for example. This results in a spectrum that contains components at the sum and difference of the original signal frequencies. Suitable filtering can then be used to extract the desired component; extracting the difference component shifts the frequency down, and extracting the sum component shifts the frequency up.

3.1.3 Analytic Reconstruction of Complex Signals Complex-signal reconstruction can be treated as the analytic extension of a real function to the complex plane. This approach is not restricted to narrow-band signals and does not require knowledge of a carrier frequency. The real and imaginary components of an analytic complex signal are related by the Hilbert transform relationships, and these relationships also apply to discrete signals in the appropriate context (Oppenheim and Schafer, 1989). In the spectral domain, this has a particularly simple interpretation; a real-valued discrete signal $x_r(n)$ has a conjugate-symmetric discrete Fourier spectrum X_r [that is, $X_r(k) = X_r^*(N - k)$, where X^* is the complex conjugate of X and N is the length of the FFT], and the spectrum of the corresponding complex signal $x(n) = x_r(n) + ix_q(n)$ is obtained from just the lower half of X_r. More precisely, define the new spectrum $\tilde{X}$ by

$$\tilde{X}(k) = \begin{cases} X_r(k) & k = 0, N/2, \\ 2X_r(k) & 0 < k < N/2, \\ 0 & N/2 < k < N; \end{cases} \tag{5}$$

then the inverse Fourier transform of $\tilde{X}$, denoted $\tilde{x}$ (with the index suppressed), is complex and is equal to $x_r + ix_q$. The imaginary component $x_q = (\tilde{x} - x_r)/i$ is the *Hilbert transformation* of x_r. Discontinuity at the ends of the N-point FFT segment of x_r can result in large fluctuations in $x_q(n)$ near the ends of the segment. This problem can be overcome by using only the middle portion of each segment of the reconstructed signal. The segments must then be chosen so that they overlap to maintain continuity in the reconstructed signal.

Whereas Eq. (5) describes complex-signal reconstruction in the frequency domain, the same can be done in the time domain by convolving x_r with the impulse response obtained from the transfer function $\tilde{X}/X_r$.

3.2 Narrow-Band Signal Filtering

The processing of narrow-band signals is important in underwater and atmospheric acoustics because of the abundance of narrow-band signals that are generated by machinery and detectable at considerable range. Narrow-band digital filters are used to extract these narrow-band signals and reject noise outside of the signal baseband. Without this filtering, the noise power may be greater than the signal power and many sophisticated processing techniques would be ineffective. Computational efficiency is necessary to implement filtering for a multisensor (array) receiver that processes complicated signals in real time. Much of this efficiency is achieved by reducing the data volume (without losing signal information) as the data are processed.

3.2.1 FIR and IIR Filters Discrete time-series data are filtered using digital filters. The reader is directed to Oppenheim and Schafer (1989) and Rabiner and Gold (1975) for a more complete discussion of digital filters. Digital filters are often defined in terms of their response to an impulse. There are two principle types of digital filters: finite impulse-response (FIR) filters, for which there is zero response at the filter output beyond a finite time after the impulse, and infinite impulse-response (IIR) filters, for which the filter response remains nonzero (albeit diminishing) indefinitely.

Implementation of a digital filter is based on the constant-coefficient finite-difference equation

$$\sum_{k=0}^{N} a_k y(n-k) = \sum_{k=0}^{M} b_k x(n-k), \tag{6}$$

where $y(n)$ is the output sequence for an input sequence $x(n)$. The coefficients a_0, a_N, and b_M are always assumed to be nonzero. For an IIR filter, at least one of the coefficients a_k must be nonzero, besides a_0, so that $N > 0$. For a FIR filter, the coefficients a_k are all zero except a_0, so that $N = 0$. The order of the filter corresponding to Eq. (6) is the maximum of N and M.

The computational complexity of both the IIR and FIR filters increases with increasing order, and so it is important to achieve the desired filter characteristics with as few filter coefficients as possible. An IIR filter can often achieve, or approximate, a desired frequency response function using fewer coefficients than can a FIR filter; however, the FIR filter has several important advantages over the IIR filter. A FIR filter can be easily implemented in the frequency domain by multiplying the discrete spectrum of the signal with that of the impulse response of the FIR filter, on a point-by-point basis. The FFT algorithm makes this implementation very efficient. An IIR filter, on the other hand, must be implemented by a recursive algorithm in the time domain. Furthermore, FIR filters are easily designed to have a linear phase response, which is a desirable property for avoiding phase distortion.

An important class of filters consists of filters designed to reject frequencies outside of some frequency band. They are typically specified by the signal rejection in their *stopband,* the deviation from unity response in their *passband,* and the width of the *transition band* between the passband and stopband. Filters that have a very narrow passband and transition band, such as would be used to select a narrow-band signal as in Eq. (1), tend to be of high order.

3.2.2 Processing Long Records Processing of acoustic signals is often done over a long period of time relative to the rate at which data are being collected (the sampling rate). Efficient algorithms based on the FFT are inherently block oriented; that is, they process data in blocks of a specific size. Practical constraints on data storage and handling during computations places limits on the size of the data block that can be processed. Techniques to overcome these constraints are important in underwater and atmospheric acoustic signal processing.

One such technique is *overlap-save* (or its variation, *overlap-add*), which achieves the convolution of long sequences by breaking them into smaller, overlapping segments. The smaller segments are extended with a predetermined number of zeroes, their convolution is computed, and portions of the results are pieced together to recover the overall convolution of the long sequences.

This method is used to implement FIR filtering of long data sequences (Oppenheim and Schafer, 1989). A FIR filter with an impulse response $\{h_n\}$ of length $N + 1$ is convolved with a signal $\{x_n\}$ to obtain the filtered signal $\{y_n = \Sigma_{k=0}^{M} x_{n-k} h_k\}$. For a data record containing the samples x_k for $0 \leq k < N$, the convolution can be correctly calculated for y_n, where $M \leq n < N$, but not for $n < M$. The first M points can, however, be obtained from the previous record by arranging that the current record overlaps the previous record by M points.

3.2.3 Efficient Baseband Signal Extraction Reconstruction of a baseband signal using *I-Q* demodulation, as described in Sec. 3.1.1, can be applied to a discrete complex envelope signal $x(n)$ by filtering with a quadrature filter or a 90° phase shifter that is constructed as a FIR or an IIR filter and applying Eq. (4), with $t = nt_s$, where t_s is the sampling interval. Since the bandwidth of the envelope is much less than the carrier frequency f_c, and since the sampling rate t_s need only be twice f_c to prevent frequency aliasing, the reconstructed envelope time series x_r can be *decimated,* or subsampled by keeping only every kth sample. This greatly reduces the data volume in the reconstructed signal. The integer k is called the *decimation factor.* Decimation factors in excess of 100 are not uncommon.

As an alternative to *I-Q* demodulation, the baseband signal can be extracted by using a very narrow bandpass filter and homeodyning the filtered signal. If a FIR filter is used, with filtering implemented in the frequency domain, then the filtered signal is easily homeodyned by shifting the coefficients of its discrete Fourier transform. Computational efficiency and reduced data storage requirements are achieved by breaking the filter into several stages that are applied in series to reduce the width of the passband progressively. This approach requires lower-order filters because the passband of each stage does not need to be as narrow as that of the overall filter. Furthermore, by decimating the output of each stage of the filter, successive filter stages process less data and considerable computational savings are realized. This multistage approach to filtering is called *variable-rate filtering.*

The individual stages in a variable-rate filter can be designed to be very efficient. Modification or substitution of stages can provide an effective method for adjusting the overall filter characteristics. At each stage of the filter the signal may be examined, which provides an opportunity to analyze the signal at different frequency resolutions. The later stages of a variable-rate filter are sometimes referred to as *vernier modes.* Vernier modes may introduce a significant delay between the input signal and output response due to the block-processing nature of the FFT.

3.3 Zoom Filtering

Received signals, particularly those including multiple sources, can contain many narrow-band signals. The center frequencies of these signals may change as the source maneuvers or changes its operating conditions. Analysis of these signals involves isolating and possibly tracking each frequency with the use of very narrow passband filters. To make this task as computationally tractable and efficient as possible, various application-specific techniques are used. As an example of how the generic narrow-band signal extraction process can be made efficient, we discuss the *zoom filter* that is particularly well suited to processing complicated narrow-band signals. While alternative approaches exist, the zoom filter employs many of the basic principles in a compact algorithm.

The zoom filter procedure method utilizes a FIR filter that can be implemented particularly efficiently in the frequency domain. This is done by multiplying only a few bins of the signal spectrum by filter coefficients, corresponding to the filter response values in the transition bands. This is possible because the zoom FIR filter frequency response is designed to be zero at the spectral bins in the stopband and unity at the spectral bins in the passband.

Zoom filtering is an example of a variable-rate filtering technique. From an elemental building block, multiple parallel and sequential stages are combined to obtain the desired analysis bands and resolutions, as in Fig. 4. The building block of the zoom transform is implemented digitally by applying the following procedures in sequence.

1. The data $s_r(n)$ are segmented into records $s_r^m(n)$ of length N (usually a power of 2), with records overlapping by $N - M$ samples; $s_r^m(n) = s_r(mM + n)$ for $0 \le n < N$.
2. Each record is Fourier transformed, using a FFT.
3. The specialized zoom FIR narrow-band filter is applied, centered on a signal baseband of interest. This is done by zeroing the spectral bins that are in the filter stopband and multiplying the spectral bins in the transition bands by the optimally selected filter coefficients.
4. The filtered spectrum is shifted k_0 spectral bins so that an edge of the passband with flanking transition bands is near to zero hertz, thus heterodyning the signal baseband to a lower (but nonzero) frequency.
5. The signal is decimated by truncating the zero-valued, high-frequency portion of the shifted spectrum and applying an inverse FFT.
6. Nonoverlapping portions of the decimated time records are concatenated. A phase factor, which is constant over each record, must be applied to maintain phase continuity across data records. The phase factor $\exp(i\phi_m)$ for the mth data record is determined by the recursion $\phi_{m+1} = \phi_m - 2\pi M k_0/N$. For example, with 50% overlap the factor is -1 for a shift of an odd number of spectral bins and $+1$ otherwise.

The concatenated time series can be submitted to a second-stage zoom transform to obtain a narrower passband and finer spectral resolution. To extract more than one narrow-band signal, multiple narrow-band filters are applied to the same signal spectrum in parallel and their outputs processed separately (and possibly in parallel). The incremental computational cost for additional narrow-band signals or higher-resolution stages is small. With proper selection of the transition band coefficients, a multistage zoom filter can easily restrict aliasing artifacts and ripple in the stopband and the passband to less than -80 dB.

3.4 Power Spectral Density Estimation

The power spectral density (PSD) is an ensemble-averaged quantity describing the distribution of signal energy in the frequency domain. It is formally defined as the Fourier transform of the autocorrelation function $\rho(\tau) = \langle x(t)x^*(t + \tau)\rangle$ of a stochastic process $x(t)$ and is expressed in units of signal power per unit frequency interval.

In practice the power spectral density is

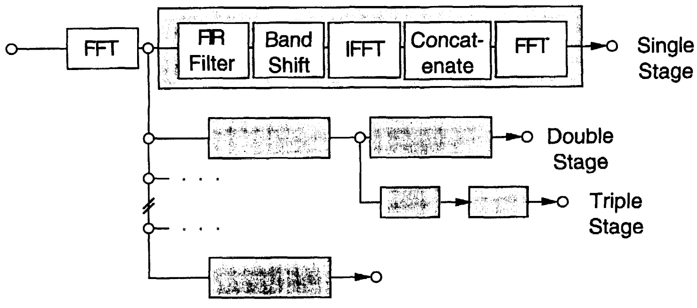

FIG. 4. Components of the zoom filter. The five basic components are FIR filter, heterodyne (band shift), inverse FFT, concatenate consecutive data segments, and FFT. These operations are grouped into a shaded box. Replicas of this box, operating with different spectral-resolution, bandwidth, and center-frequency parameters, are applied in series and parallel to achieve multiresolution processing in a computationally efficient and flexible manner.

estimated from the power spectra of windowed discrete data $y(n) = w(n)x(n\Delta t)$, as

$$\hat{P}(f/f_s) = \frac{2}{W_0}\left|\sum_{n=0}^{N-1} y(n)e^{-2\pi i n f/f_s}\right|^2, \tag{7}$$

where $\Delta t = 1/f_s$ is the sampling interval, the signal frequency f lies in the range $\pm f_s/2$, the window function $\{w(n)\}$ is zero outside the interval $0 \le n < N$, and $W_0 = \Sigma_{n=0}^{N-1} w(n)$. When the window function $w(n)$ is rectangular, with $w(n) = 1$ for $0 \le n < N$, the PSD estimate is called a *periodogram*. For a nonrectangular window function, $\hat{P}$ is called a *modified periodogram*.

The PSD estimate in Eq. (7) is typically calculated using a FFT and evaluates $\hat{P}$ at the normalized frequencies $f_k/f_s = k/N$. The numerical value of $\hat{P}(k/N)$ is the signal power in the kth bin (modulo N) relative to that of a unit discrete impulse. The PSD can be expressed in units of volts^2/hertz, as a function of conventional frequency, using

$$P_{\mathrm{Hz}}(f) = (\Delta V)^2 \Delta t \hat{P}(f/f_s), \tag{8}$$

where ΔV is the voltage quantization step of the digital data.

The variance, or uncertainty, in the power spectral density estimate $\hat{P}$ is independent of the duration $N\Delta t$ of the time record. In fact, the variance is approximately equal to the average value of the estimate. To obtain a better estimate of the PSD, that is, one with lower variance, it is necessary either to average adjacent frequency bins under the assumption that the PSD is approximately constant across these frequencies or to average multiple periodograms. The variance in the PSD estimate varies in proportion to $1/M$, where M is the number of adjacent bins or sequential periodograms being averaged. For individual PSD estimates to be accurate, the autocorrelation function of the signal must be small for lag times beyond $N\Delta t$.

A convenient technique for the display of time-varying signals is to plot the PSD as a function of frequency and time. This particular graphical data format has been widely used for the analysis of underwater acoustic data and is referred to as a *lofargram*, in reference to low frequency analysis and recording. The example lofargram in Fig. 5 shows time-varying frequency components from maneuvering small aircraft. Lofargrams for the analysis of underwater sources do not usually display such large variations in line frequency. Traditional lofargrams were recorded by a temperature-modulated hot needle that produced variable-density traces on long rolls of specially treated paper.

The visibility of lines in a lofargram is enhanced by techniques such as normalizing the spectrum with respect to the background noise. The lofargram has proven to be a valuable detector for narrow-band signals because the human visual system is effective at

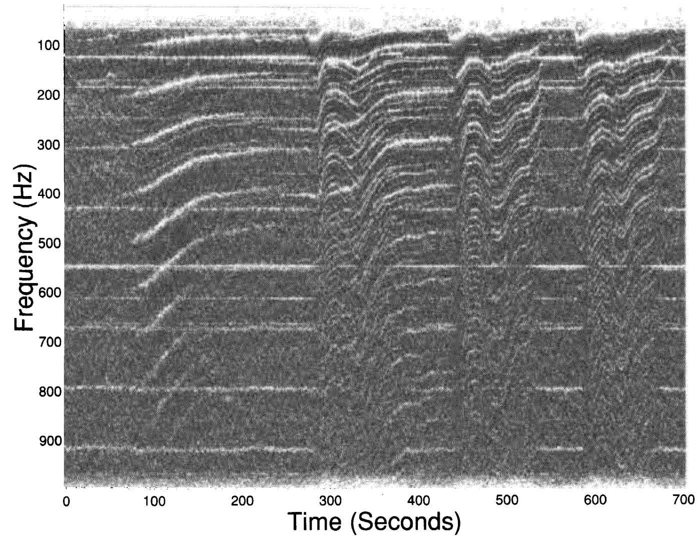

FIG. 5. An example lofargram of a maneuvering aircraft. White and black indicate high and low power spectral density, respectively. The density scale is arbitrary. The straight white lines correspond to stable tones. The wavy lines correspond to signals from aircraft, which exhibit frequency variations due to Doppler-shift and source variations. The signals are normalized with respect to the mean noise spectrum.

recognizing even faint narrow-band signals on long lofargram records. Furthermore, slight variations in line frequency and relationships between the line-frequency fluctuations of different lines are readily detected visually.

High-resolution spectral-estimation techniques are typically applied more for the purpose of frequency estimation and the resolution of closely spaced frequency components than they are for accurate PSD estimation. An overview of high-resolution spectral-estimation methods can be found in Marple (1987).

4. ARRAY PROCESSING

Spatial sampling of a sound field is usually done by an array of transducers, although the synthetic-aperture array, where a sensor is moved through space to obtain sequential measurements in both the time and space domains, is an important exception. Spatial sampling is analogous to temporal sampling with the sampling interval replaced by the sensor spacing vectors. The Nyquist criterion requires that the sensor spacing be at least twice the spatial wave number of the measured sound field. Aliasing and signal reconstruction for spatially sampled data follow the same principles as for temporal data, although there are some important differences such as the ambiguity that can arise in the reconstructed wave-number components due to symmetries in the array geometry.

Practical constraints associated with using multiple sensors make it necessary to minimize the number of sensors used; consequently, the conceptually simple uniformly spaced array distributed in three dimensions is seldom implemented in practice.

4.1 Linear Plane-Wave Beam Forming and Spatiotemporal Sampling

The simplest example of array processing is plane-wave beam forming, or delay and sum beam forming in the time domain. This processing is implemented by phase shading in the frequency domain or by sample delay in the time domain and is used to search for the bearing angle θ associated with a plane-wave signal as shown in Fig. 6 for a linear array.

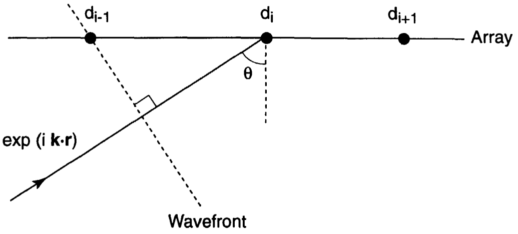

FIG. 6. Geometry for plane-wave beamforming.

4.1.1 Frequency-Domain Processing A plane wave with bearing angle θ and wave number $\mathbf{k}(\theta)$, observed at a position $\mathbf{r}$, can be represented as

$$s(\theta) = e^{i\mathbf{k}(\theta)\cdot\mathbf{r}}, \tag{9}$$

where we have suppressed the time dependence $\exp(-i\omega t)$, and $\omega = ck = c|\mathbf{k}|$ is the angular frequency of the signal. For an array of sensors at positions $\mathbf{d}_j$ the field is summed coherently (in phase) if the inputs r_j from the receiving elements (hydrophones or microphones) are first multiplied by the complex conjugate of the plane-wave phase factors (or *plane-wave replica*),

$$w_j = |w_j| e^{i\mathbf{k}\cdot\mathbf{d}_j} = e^{id(k\sin\theta_s)}, \tag{10}$$

where θ_s is a scanning angle (or *steering angle*) and $|w_j|$ are weighting factors (also called *tapers* or *array shading*) used to control sidelobes (Harris, 1978). The beam-forming process $B(\theta_s) = \Sigma w_j r_j$ will have a maximum when the scanning angle equals the incident angle of the signal.

Often it is the power (mean square of the magnitude) of the output of the beam former that is of interest:

$$\begin{aligned} |B(\theta_s)|^2 &= \left\langle \left| \sum_{i=1}^{m} w_i^*(\theta_s)(s_i(\theta) + n_i) \right|^2 \right\rangle \\ &= \sum_{i,j=1}^{m} w_i^*(\theta_s)(K_{s_{ij}} + K_{n_{ij}}) w_j(\theta_s), \end{aligned} \tag{11}$$

where the inputs are $r_i = s_i + n_i$, with s_i and n_i being the signal and noise at the ith receiving element, and where $K_{s_{ij}} + K_{n_{ij}}$ are elements of a *cross-spectral density matrix* (CSDM). The angle brackets indicate the expected (ensemble-averaged) value. The cross-spectral density matrix for the single-tone plane wave considered here corresponds to the signal covariance matrix of the array of receiving elements, but for more general signals it is obtained from the data using Fourier transforms and time averages, as discussed at the end of this section. To obtain the right-hand side of Eq. (11), the signal and noise fields are assumed to be uncorrelated.

We can write the above expression in matrix notation with boldface lower-case letters denoting vectors (in column form) and boldface upper-case letters denoting matrices. Define a *steering vector* $\mathbf{w} = \{w_i\}$ and a cross-spectral density matrix $\mathbf{K} = \{K_{ij}\}$ with $\mathbf{K} = \mathbf{K}_s + \mathbf{K}_n = \{K_{s_{ij}} + K_{n_{ij}}\}$ since the signal and noise are assumed to be independent.

In practice, to be precise, we cannot know the ensemble-average statistic **K**. Instead, we measure the *sample covariance*

$$\hat{\mathbf{K}} = \overline{\mathbf{q}^\dagger \mathbf{q}}$$

from time-averaged data **q**, where the dagger denotes the complex transpose operation and the overbar denotes time averaging, and assume ergodicity so that $\hat{\mathbf{K}}$ can be used to approximate the CSDM **K**. It is useful to note that the cross-spectral density matrix is Hermitian since it is constructed as the outer product of data vectors.

Equation (11) can be approximated as

$$|B(\theta_s)|^2 = \mathbf{w}^\dagger(\theta_s)\hat{\mathbf{K}}(\theta_{\text{true}})\mathbf{w}(\theta_s) \equiv \mathbf{w}^\dagger\hat{\mathbf{K}}\mathbf{w}, \qquad (12)$$

where the sample covariance of the field is composed of uncorrelated signal and noise covariances,

$$\hat{\mathbf{K}} = \hat{\mathbf{K}}_s + \hat{\mathbf{K}}_n. \qquad (13)$$

The data across the array, as represented in the matrix $\hat{\mathbf{K}}$, contain the information that the source is in the direction θ_{true}. Sometimes $\mathbf{w}(\theta_s)$ is referred to as a replica, and the above beam-forming process is viewed as matching the received data across the array with a replica. The type of beam former represented by Eq. (12) is called a *linear beam former,* or *Bartlett beam former,* as opposed to a nonlinear beam former, which utilizes data-dependent weights. We also mention that the beam-former process can also be considered as a wave-number–scanning process rather than an angle-scanning process.

For the sample covariance estimation we assume that we have an array with N sensors located at $\mathbf{d}_i$, $1 \le i \le N$, and a narrow-band model as illustrated in Fig. 7. While we assume in much of our discussion that the sample covariances are known, generally these covariances must be estimated from the data. For this estimation the signal is segmented into "snapshots," using a sampling window $w(t)$ that is unity for $0 \le t \le T_w$ and zero elsewhere, and then it is Fourier transformed to obtain the spectral estimates

$$R_i^l(f) = \int_{Tl}^{Tl+T_w} r_i(t)w(t - T_l)e^{-2\pi i f t}dt, \qquad (14)$$

where $0 \le i < M$ indexes the sensor and $0 \le l < L$ indexes the snapshot. We assume that window lengths and resolution bandwidths have been chosen so that bias errors are not an issue.

In most beam-forming algorithms these spectral estimates are used to obtain the sample covariance matrix

$$\hat{\mathbf{K}}(f) = L^{-1}\mathbf{R}(f)\mathbf{R}(f)^\dagger, \qquad (15)$$

where the matrix product sums across snapshots. We assume that there are L_{eff} "degrees of freedom" in the sample covariance matrix. (One can also average across frequency if relative phase constraints across the array are observed.) We assume that the data vectors have zero mean and that the expected value of the sample covariance is the ensemble covariance $\mathbf{K}(f)$. If we further assume that the statistics of the field are Gaussian, then the data vectors are Gaussian and the sample covariance matrix has a complex

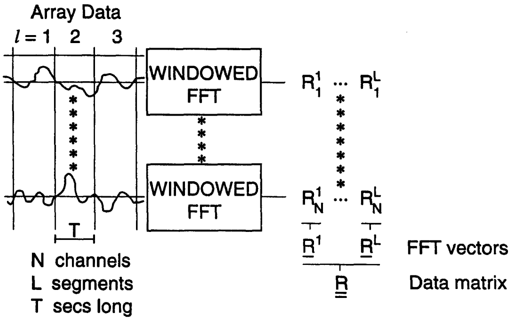

FIG. 7. Array, narrow-band model, sample matrix estimation, and data preconditioning for beam forming.

Wishart distribution of order N with L_{eff} degrees of freedom (Anderson, 1984).

4.1.2 Time-Domain Processing Time delay is the time-domain analogy to (linear) phase shading in the frequency domain. This can be derived formally by taking the Fourier transform of the beam-forming process. Using the property that the Fourier transform of a product of functions in one space is a convolution in the conjugate space, we can immediately write down from Eq. (11) that

$$B(t,\theta) = \sum_i \int w_i(t',\theta) r_i(t - t')dt', \tag{16}$$

where $r_i(t)$ is the signal plus noise time-domain data at the ith array element. The time-domain replica $w_i(t,\theta)$ is the Fourier transform of Eq. (10) [with the time dependence $\exp(-i\omega t)$ included],

$$w_i(t,\theta) = g_i\delta(t - (d_i/c)\sin\theta), \tag{17}$$

where δ is the Dirac delta function, and we have included the tapers g_i so that we can write the beam-former output as

$$B(t,\theta) = \sum_i g_i r_i\left(t - \frac{d_i}{c}\sin\theta\right). \tag{18}$$

This process is referred to as *delay and sum beam forming;* the delay is simply the time interval associated with the phase shift in the frequency-domain array processing. Here we note that we must now impose another requirement on the temporal sampling of the data in order to achieve the delay quantization consistent with the inherent angular resolution of the spatial array. The temporal sampling requirements for time-domain beam forming may be more than ten times the Nyquist requirements. Interpolation filters may be used to obtain samples between those obtained by the data acquisition system.

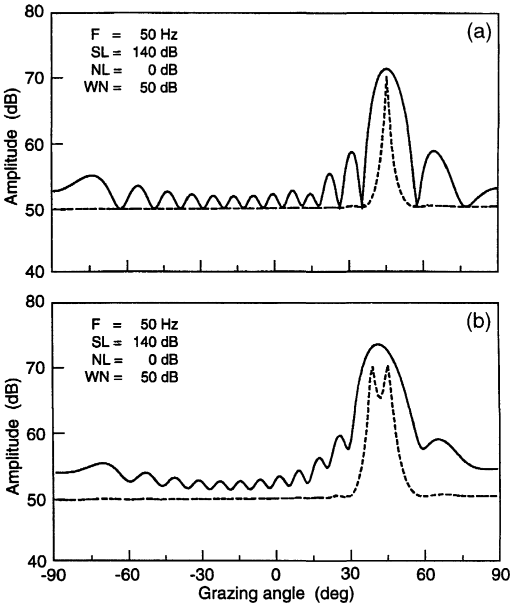

FIG. 8. Beam-former outputs. (a) Single source at a bearing of 45°. (b) Two sources with 6.3° angular separation. Solid line: linear processor (Bartlett). Dashed line: minimum-variance distortionless processor (MV).

4.1.3 Some Properties of Array Processors Figure 8 shows the output results of some plane-wave beam formers for the cases of one and two incident signals. To be noted are the sidelobes of a nontapered linear processor and the high-resolution performance of the adaptive processors (discussed in the next section). Some of the general attributes of an array processor are as follows:

- *The main response axis (MRA).* One often normalizes the beam-former output to have 0 dB, or unity, gain in the direction of maximum response, which under ideal conditions occurs when the array is steered in the direction of the signal.
- *Beamwidth.* An array with finite extent, or aperture, must have a finite beamwidth centered about the MRA that is termed the *main lobe.* A typical scale for the main-lobe beamwidth (in radians) of a line array is the reciprocal of the length of the array expressed in wavelengths. For linear arrays there is an extensive literature on weightings, often termed *tapers,* that produce desired main-lobe characteristics (see Harris, 1978, for example). For two-dimensional arrays, or those with an irregular sensor distribution, the situation is much more complicated, and adaptive processing (see below) plays a larger role in many contemporary applications.
- *Sidelobes.* Sidelobes are angular or wavenumber regions where the array has a

relatively strong response. Sometimes they can be comparable to the MRA, but in a well-designed array, they are usually -20 dB or lower; i.e., the response of the array is less than 0.1 to a signal in the direction of a sidelobe. Sidelobe performance is especially important in jamming situations where very strong signals entering through the sidelobes can overwhelm the desired signal. Again, for linear arrays there is a well-developed theory for sidelobe control (Harris, 1978); however, for more complicated arrays adaptive methods may be appropriate. It is also important to note that sidelobes are a function of frequency, and this can be exploited in many applications.

- *Wave-number processing.* Rather than scanning through incident angles, one can scan through wave numbers, $k \sin\theta_s \equiv \kappa_s$. Scanning through all possible values of κ_s results in nonphysical angles when $|\kappa_s| > k$, which correspond to waves not propagating at the acoustic-medium speed. Such waves may be associated, for example, with array vibrations. The beams associated with these wave numbers are sometimes referred to as *virtual beams.* An important aspect of these beams is that their sidelobes can be in the physical angle region, thereby interfering with acoustic propagating signals.
- *Array gain.* The array gain is defined as the ratio of the signal-to-noise ratio (SNR) of the array output to the SNR of a single-sensor output for a given ambient noise field. Array gain is expressed in decibels and, if the noise field is isotropic, the array gain is also termed the *directivity index.*

4.2 Adaptive Processors

The high-resolution methods to suppress sidelobes and perform main-lobe nulling are usually referred to as *adaptive methods* since the signal-processing procedure constructs weight vectors depending on the received data itself. We briefly describe one of these procedures: the minimum-variance distortionless processor (MV), sometimes also called the maximum-likelihood method (MLM) directional spectrum estimation procedure. The powerful tool of eigenvector analysis of beam formers is also presented. There are many other types of adaptive beam-forming algorithms (Johnson and Dudgeon, 1993). These adaptive methods are characteristically nonlinear functions of the received fields. As a matter of fact, these beam formers often have a signal-to-noise ratio threshold that must be exceeded for their desirable properties to appear. For this reason, nonlinear beam forming is often used interchangeably with adaptive beam forming. The processing described below is done in the frequency domain.

4.2.1 Minimum-Variance Distortionless Processor We seek a weight vector $\mathbf{w}_{\mathrm{MV}}$ applied to the matrix $\hat{\mathbf{K}}$ such that its effect will be to minimize the output of the beam former in Eq. (12) except in the "look" direction where we want the signal to pass through undistorted. Beam forming with such a weight vector is referred to as the *minimum-variance distortionless processor* (MV), which essentially optimizes the match to a signal in the "look" direction and therefore tries to reject signals and correlated noise from other directions. The weight vector is therefore chosen to minimize the functional

$$F = \mathbf{w}_{\mathrm{MV}}^{\dagger}\hat{\mathbf{K}}\mathbf{w}_{\mathrm{MV}} + \alpha[\mathbf{w}_{\mathrm{MV}}^{\dagger}\mathbf{w}(\theta) - 1]. \tag{19}$$

The first term is the mean square output of the array, and the second term incorporates the constraint of unity gain by means of the Lagrangian multiplier α. Note that for this complex-variable problem, there are two independent variables, which we can take as $\mathbf{w}_{\mathrm{MV}}$ and $\mathbf{w}_{\mathrm{MV}}^{*}$. Following the method of Lagrange multipliers, we obtain the MV weight vector,

$$\mathbf{w}_{\mathrm{MV}}(\theta) = \frac{\hat{\mathbf{K}}^{-1}\mathbf{w}(\theta)}{\mathbf{w}^{\dagger}(\theta)\hat{\mathbf{K}}^{-1}\mathbf{w}(\theta)}. \tag{20}$$

This new weight vector depends on the received data as represented by the estimate of the cross-spectral density matrix that is measured for a particular ambient noise field; hence, the method is "adaptive." Substituting back into the quadratic form of Eq. (19) (which is constructed so that the second term on the right-hand side will vanish) gives the output of our MV processor,

$$B_{\mathrm{MV}}(\theta_s) = [\mathbf{w}^{\dagger}(\theta_s)\hat{\mathbf{K}}^{-1}(\theta_{\mathrm{true}})\mathbf{w}(\theta_s)]^{-1}. \tag{21}$$

The MV beam former should have the same peak value at θ_{true} as the linear beam former in Eq. (12) but with the sidelobes suppressed and a narrower main beam, meaning that it is a higher-resolution beam former.

4.2.2 Eigenvectors and Beam Formers Eigenvector analysis provides a powerful tool to study and implement beam-forming algorithms. In principle, the estimate of the cross-spectral density matrix $\hat{\mathbf{K}}$ across a receiving array is made by taking an average over an appropriate time interval such that signals from sources are statistically independent of each other (uncorrelated) and of the background noise. For the set of source positions that the array can resolve, we can further interpret this to mean that the source position vectors are linearly independent, and, furthermore, that the noise subspace is orthogonal to the signal subspace. This linear independence of the source vectors, though, is not a necessary condition for implementing beam formers by utilizing the computational advantages of eigenvector methods.

We start by representing $\hat{\mathbf{K}}$ in terms of its M eigenvectors $\mathbf{v}_i$ and corresponding eigenvalues λ_i,

$$\hat{\mathbf{K}} = \sum_{i=1}^{M} \lambda_i \mathbf{v}_i \mathbf{v}_i^{\dagger}, \tag{22}$$

which satisfy the standard eigenvalue equation,

$$\hat{\mathbf{K}}\mathbf{v}_i = \lambda_i \mathbf{v}_i,\ i = 1, \ldots, M. \tag{23}$$

If there are p independent signal sources of strength σ_i with signal vectors $\mathbf{s}_i$, then $\hat{\mathbf{K}}_s$ of Eq. (13) can be expressed in terms of these signal vectors such that the sample covariance approximation to the total CSDM can be written as

$$\hat{\mathbf{K}} = \sum_{i=\mathrm{i}}^{p} \sigma_i \mathbf{s}_\mathbf{i} \mathbf{s}_\mathbf{i}^{\dagger} + \hat{\mathbf{K}}_n. \tag{24}$$

The simplest case is when the noise is spatially white (with strength σ_n at each sensor) for which the sample covariance of the noise is proportional to the identity matrix; then Eq. (24) becomes

$$\hat{\mathbf{K}} = \sum_{i=\mathrm{i}}^{p} \sigma_i \mathbf{s}_\mathbf{i} \mathbf{s}_\mathbf{i}^{\dagger} + \sigma_n \mathbf{I}. \tag{25}$$

It then follows that there are p eigenvectors that correspond to the signal vectors so that the remaining $M - p$ eigenvectors are orthogonal to the signal vector space. Note that the signal eigenvectors have eigenvalues $\sigma_i + \sigma_n$ that include the noise. We therefore have that

$$\hat{\mathbf{K}} = \sum_{i=1}^{M} (\sigma_i + \sigma_n) \mathbf{v}_i \mathbf{v}_i^{\dagger}, \tag{26}$$

and

$$\hat{\mathbf{K}}^{-1} = \sum_{i=1}^{M} (\sigma_i + \sigma_n)^{-1} \mathbf{v}_i \mathbf{v}_i^{\dagger}, \tag{27}$$

such that $\sigma_i = 0$ for $i > p$.

It is straightforward to simply use the eigenvalue decomposition of the CSDM to implement the linear and MV processors; however, there are some interesting additional variants of interest. For the linear processor, we might limit the eigenvector representation to the signal space. Alternatively, the eigenvector version of the MV processor suggests a direct application for the situation when eigenvectors can represent specific source vectors. In this case, we can eliminate signals by truncating the matrix representation so as not to include the signal subspace, and use as the CSDM

$$\hat{\mathbf{K}}_{\perp\text{sig}}^{-1} = \sum_{i=M-p}^{M} \lambda_i^{-1} \mathbf{v}_i \mathbf{v}_i^{\dagger}. \tag{28}$$

The replica scanning process produces a null for the source locations so that the MV prescription of the inverse of the process is infinite at source locations and finite otherwise. This provides an extremely high-resolution estimate of the source location. The *MUSIC* method (Schmidt, 1986) uses eigenvector decomposition where the eigenvalues of the (noise) subspace orthogonal to the signal subspace are set to unity by a "prewhitening" transformation of the CSDM. This kind of processor can estimate source direction but does not give a good estimate of source strength.

Finally we note that when the noise is

correlated so that the second term of Eq. (24) is not diagonal, then there is no one-to-one correspondence between eigenvectors and source location vectors. Nevertheless, one can expect that under certain conditions such as highly directional noise conditions or the existence of *a priori* knowledge about the noise field, eigenvector methods may prove beneficial. Examples of adaptive processing are shown in Fig. 8.

4.3 Matched-Field Processing

Matched-field processing (MFP) (Baggeroer *et al.*, 1993; Tolstoy, 1993) as first applied to ocean acoustics is the three-dimensional generalization of the conventional lower-dimensional plane-wave beam former. The one or two dimensions in the latter case correspond to bearing, and the matching is done to plane-wave replicas for these bearings. Localization in this case refers to determining a bearing direction. The generalized matched-field beam former matches the measured field at the array with replicas of the expected field for a point source. At each possible source location a replica of the expected field is derived from a propagation model, based on a specific ocean environmental model, as discussed in ACOUSTICS, UNDERWATER.

Localization in range, depth, and azimuth depends upon the degree of uniqueness of the spatial structure of the field, which in turn depends upon the array geometry and complexity of the ocean environment. The acoustic field generally has complex interference patterns that depend upon the source location and the ocean environment, thus allowing the field pattern to be *matched* to a source location and ocean environment. In terms of rays we can say that the refractive properties of the waveguide generate a pattern of arrival angles that can also be matched.

The process is shown schematically in Figs. 9 and 10. The process consists of systematically placing a test source at each point of a search grid, computing the acoustic field (replicas) at all the elements of the array, and then correlating this modeled field with the data from the real source whose location is unknown. When the test source is collocated with the true source, the correlation will be a maximum. The two main factors that limit performance of MFP are noise (both natural and man-made) and the ability to model the ocean acoustic environment accurately.

Using a specific ocean environment and an arbitrary (test) source location **a**, we construct a replica field **w** from one of the numerical solutions to the acoustic wave equation (see ACOUSTICS, UNDERWATER) to obtain a replica at each element of the array. Rather than finding just the direction of the source from the array, we can search for its actual location by matching the received data $\hat{\mathbf{K}}(\mathbf{a}_{\text{true}})$ from the true source location $\mathbf{a}_{\text{true}}$ against solutions of the wave equation based on the test source location **a**.

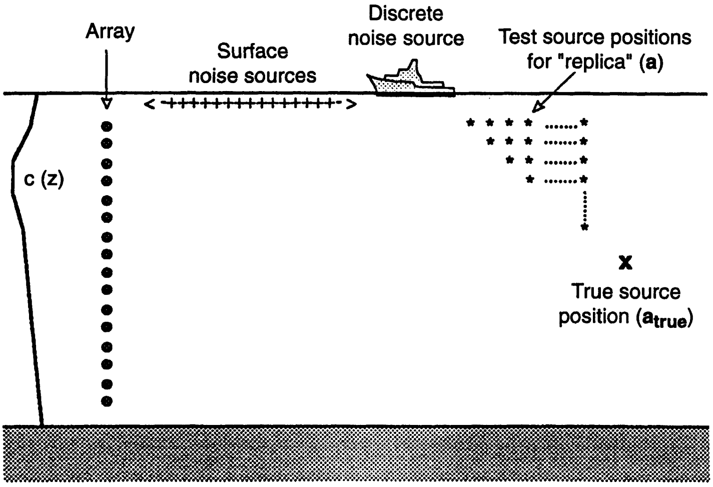
Array
Surface noise sources
Discrete noise source
Test source positions for "replica" (a)
c (z)
X
True source position ($\mathbf{a}_{\text{true}}$)

FIG. 9. Schematic of matched-field processing. The ocean is treated as a waveguide with a sound-speed profile *c*(*z*). Sound propagates from a source to an array of sensors. Ships and surface activity generate noise at the array. The source is located by systematically choosing test source positions and comparing the modeled data from the test positions with the actual signal on the array. When the test source position is at the true source position, correlation between modeled and real data is a maximum.

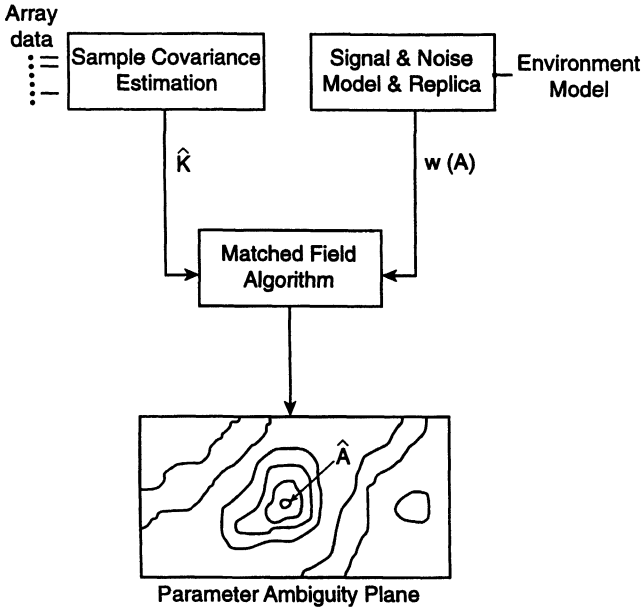
Array data
Sample Covariance Estimation
Signal & Noise Model & Replica
Environment Model
K̂
w (A)
Matched Field Algorithm
Â
Parameter Ambiguity Plane

FIG. 10. Block diagram of MFP processor.

The output of this matched-field processor, denoted by $S(\mathbf{a})$ to indicate the generalization beyond plane-wave beam forming, at each point in space $\mathbf{a} \equiv (x,y,z)$ is, in analogy to Eq. (12),

$$S(\mathbf{a}) = \mathbf{w}^{\dagger}(\mathbf{a})\hat{\mathbf{K}}(\mathbf{a}_{\text{true}})\mathbf{w}(\mathbf{a}). \tag{29}$$

where the peak of the output of the beam former, $S(\mathbf{a})$, is at $\mathbf{a}_{\text{true}}$. The beam-former output $S(\mathbf{a})$ is also referred to as the *Bartlett ambiguity function* (or surface) of the matched-field processor because it also contains ambiguous peaks, some of which are analogous to the sidelobes of a conventional plane-wave beam former. However, there is the added feature that the acoustic field in a waveguide has cyclic properties, an obvious example being adjacent convergence zones (see ACOUSTICS, UNDERWATER) that have similar spatial structures. Hence the localization of a source in one convergence zone may be confused with another, ambiguous convergence zone. Suppression of sidelobes, including those that originate from the waveguide nature of propagation, can often be accomplished by using a nonlinear beam former such as the maximum-likelihood method (MV) beam former:

$$S_{\text{MV}}(\mathbf{a}) = [\mathbf{w}^{\dagger}(\mathbf{a})\hat{\mathbf{K}}_{-1}(\mathbf{a}_{\text{true}})\mathbf{w}(\mathbf{a})]^{-1}. \tag{30}$$

A simulated vertical receiver array, as illustrated in Fig. 11, is used to demonstrate

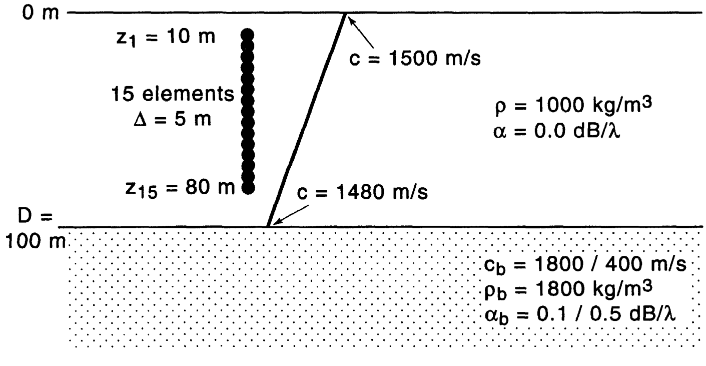
0 m
z1 = 10 m
15 elements
Δ = 5 m
z15 = 80 m
c = 1500 m/s
ρ = 1000 kg/m3
α = 0.0 dB/λ
c = 1480 m/s
D = 100 m
cb = 1800 / 400 m/s
ρb = 1800 kg/m3
αb = 0.1 / 0.5 dB/λ

FIG. 11. The environment for the matched-field processing results shown in Fig. 12.

an example of the Bartlett and MV types of MFP processors for an ocean acoustic waveguide with a high signal-to-noise ratio; the results are shown in Fig. 12. Eigenvector versions of these processors can be implemented as described above.

Mismatch between the replica fields and the actual acoustic field in MFP, usually due to incomplete knowledge of the environment, is an important problem to which much attention has been devoted. The technique of *matched-field tomography* (MFT) searches for the environmental parameters controlling the propagation (for example, the index of refraction, which may be a spatially dependent coefficient of the wave equation) rather than source location. As with the case of plane-wave beam forming, there are an assortment of processors that can be used for MFP/MFT (Baggeroer *et al.*, 1993).

There have been two approaches to deal with environmental mismatch in MFP. The first has been to construct processors that are tolerant to environmental uncertainty (Schmidt *et al.*, 1990; Richardson and Nolte, 1991; Krolik, 1992). These processors trade resolution for lack of precise knowledge of the environment. The second approach is *focalization* (Collins, 1991), which uses nonlinear optimization methods to search simultaneously for the environment and the source position. All of these methods have limita-

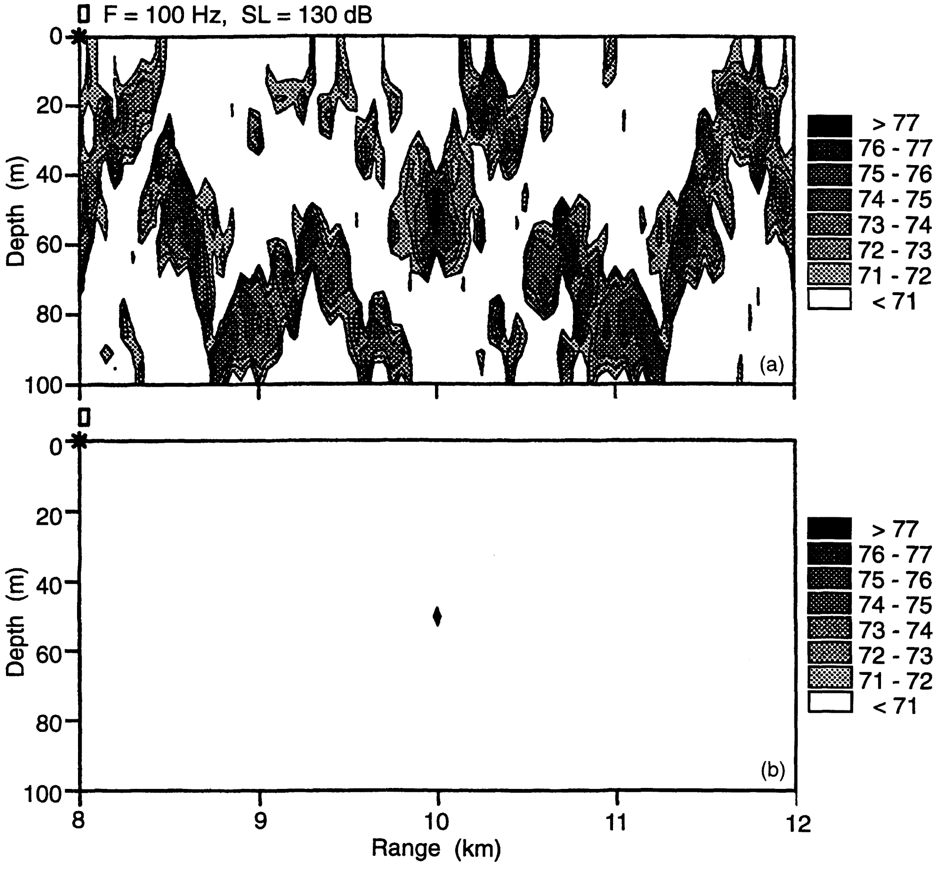

FIG. 12. Matched-field processing results. (a) Bartlett MFP ambiguity surface showing significant ambiguities (sidelobes) about the true location, which is at 10 km range and a depth of 50 m. (b) The MV processor with exact knowledge of the environment suppresses ambiguities.

tions, but under various conditions they have been successfully applied to experimental data.

5. ACTIVE SYSTEMS

Active systems (Baggeroer, 1978; Burdic, 1984) detect targets from reflections of a signal transmitted by the system and estimate their range (reflection travel time), speed (reflection Doppler), and bearing. They also can be used to infer environmental properties such as bottom bathymetry, currents and winds, particulate concentrations, etc. In an active system, a signal pulse with a known wave form is transmitted at known time and frequency. This contrasts to a passive system where the wave form is generally unknown and the transmission is asynchronous. This has two important signal-processing consequences. First, one can use coherent processing, in particular matched filtering, over the frequency band of the signal. Second, knowledge of the transmission time and carrier frequency enables one to estimate the travel times and Doppler shifts of the reflections, and hence the ranges and speeds of the targets producing the reflections.

The spatial-processing methods already discussed are applicable to the active problem, so that in this section we emphasize the temporal aspects of active signal processing. Here, two items that are of particular interest are target motion and signal design.

Wave forms are central to active sonar, where we are concerned with the complex envelope function [denoted here as $\tilde{s}(t)$] of a narrow-band real signal $s_r(t)$ with center frequency f_c and bandwidth $B < 0.5f_c$. The complex wave form and the envelope function can be reconstructed from the real signal by methods such as *I-Q* demodulation (see Sec. 3.1). Working with the complex envelope function provides considerable simplification in the mathematical notation for the signal processing, but it also has important practical consequences. Straightforward sampling of the original narrow-band signal requires a sampling rate of $2(f_c + B/2)$, where B is the bandwidth, whereas the complex envelope requires a sampling rate of only $2B$. Since the carrier frequency is usually much greater than the bandwidth, $f_c \gg B$, the reduction in data rate, and corresponding computational savings, can be considerable.

5.1 Active Sonar Signal Processing

Figure 13 schematically illustrates the basic elements of an active sonar: the (wave form) transmitter, the channel through which the signal, echo, and interference propagate, and the receiver. The receiver consists of some sort of matched filter, here a quadrature demodulated correlator, a square-law device, and possibly a threshold device for the detector and range, Doppler,

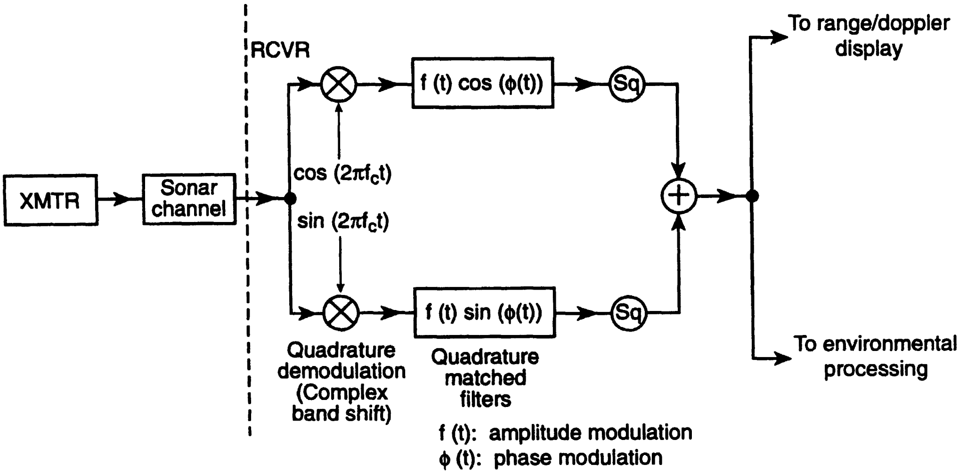

FIG. 13. Front-end processing structure for a quadrature-matched filtering of a sonar signal.

and bearing scanners for the estimator. (Some systems operate at the carrier or some intermediate frequency.)

The matched filter maximizes the ratio of the peak output signal power to the variance of the noise and is implemented by correlating the received signal with the transmitted signal. Implementing the narrow-band correlation process digitally with the complex signal, as suggested in Fig. 13, results in a lower sampling requirement than working with the original observed real signal. A simple description of the received signal, $r(t)$, is that it is an attenuated, delayed, and Doppler-shifted version of the transmitted signal, $s_t(t)$:

$$r(t) \approx \mathrm{Re}[\alpha e^{i\theta}\tilde{s}_t(t - T)e^{2\pi i f_c t}e^{2\pi i f_d t} + n(t)], \tag{31}$$

where α is the attenuation due to transmission loss and target cross section, θ is a random phase from the range uncertainty compared to a wavelength, T is the range delay time, and f_d is the Doppler shift caused by the target motion. The correlation process can be performed in either the frequency or time domain and will have an output related to the following process:

$$C(\mathbf{a}) = \left|\int \tilde{r}^*(t)\tilde{s}(t;\mathbf{a})dt\right|^2 \tag{32}$$

$$= \left|\int \tilde{R}^*(f)\tilde{S}(f,\mathbf{a})df\right|^2, \tag{33}$$

where $\tilde{s}(t;\mathbf{a})$ is a replica of the envelope of the transmitted signal as determined by a parameter set **a** that includes the propagation-reflection process. The parameter set may consist of the range delay and Doppler rate; a particular set of these parameters is called a *range–Doppler resolution* cell.

For the detection problem, the correlation receiver is used to generate a sufficient statistic that is the basis for a threshold comparison in making a decision whether a target is present. The parameters **a** determine the range and Doppler-shift value in the particular resolution cell of concern. If the range and Doppler shift are not known, this threshold must be generated over the entire plane of potential range–Doppler cells. The range–Doppler estimation problem has both global and local aspects. In the global aspect one simply wants to identify which range–Doppler resolution cell the target occupies. This yields an estimate with an accuracy of roughly the reciprocal of the bandwidth ($1/B$) for the range and the reciprocal of the duration ($1/T$) for the Doppler shift. At high SNR (signal-to-noise ratio) the accuracy can be further refined by finding the matched filter, or correlation, peak over a denser grid in the range–Doppler plane. This improves the accuracy by $1/\sqrt{\mathrm{SNR}}$ (where the SNR is expressed here in linear units, not in decibels).

For a matched filter operating in a background of white noise and detecting a point target in a given range–Doppler resolution cell, the detection SNR depends on the ratio of average energy to noise and not on the shape of the signal. The wave form becomes a factor when there is a reverberant environment and when one is concerned with estimating which range–Doppler resolution cell the target is in. The potential for range and Doppler resolution can be ascertained from the *ambiguity function* Θ of the transmitted signal,

$$\Theta(\tilde{T},T_t,\hat{f}_d,f_{d_t}) \propto \left|\int \tilde{s}_t^*(t - T_t)\tilde{s}_t(t - \hat{T}) \exp[2\pi i(f_{d_t} - \hat{f}_d)t]dt\right|^2 \tag{34}$$

where $\tilde{s}_t$ is the (complex reconstruction of the) transmitted signal, T_t and f_{d_t} are the true target range (time) and Doppler shift, and $\hat{T}$, $\hat{f}_d$ are the scanning estimates of range and Doppler shift. The ambiguity function is related to the correlation process of Eq. (32) for a transmitted signal scanned as a function of range and Doppler shift and can be used to design desirable wave forms for particular situations (VanTrees, 1971).

Figure 14 shows sketches of ambiguity functions of some typical wave forms. The range resolution is determined by the reciprocal of the bandwidth and the Doppler resolution by the reciprocal of the duration. The coded or pseudorandom sequence can attain good resolution of both by appearing as long-duration noise with a wide bandwidth.

Many environments and targets introduce travel-time and Doppler-shift spreading. The duration of the reflection of a short transmitted pulse is extended by travel-time spreading and the bandwidth of the reflection of a transmitted narrow-band tone is widened by

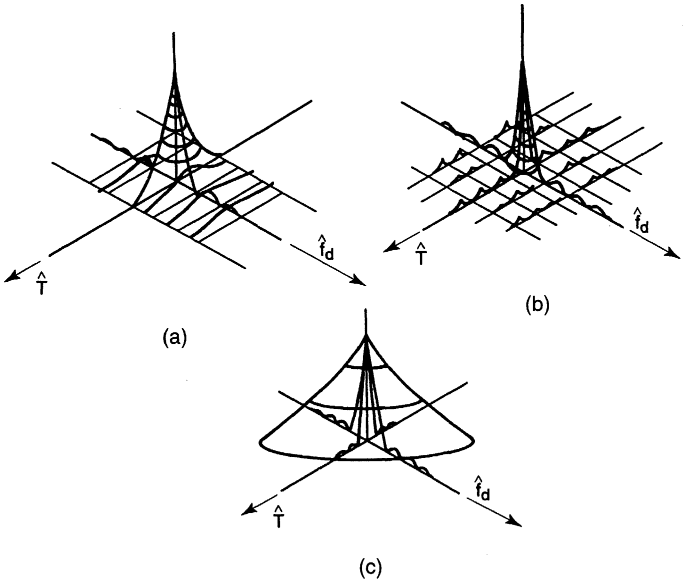

FIG. 14. Ambiguity function for several sonar signals: (a) rectangular pulse; (b) coded pulses; (c) chirped FM pulse.

Doppler-shift spreading. The *scattering function* (Westerfield *et al.*, 1960) describes how the energy of a transmitted signal is statistically redistributed in the reverberant ocean environment by multipath and Doppler-shift spread. This formulation leads to the result that the distribution of energy in the range–Doppler plane is given by the convolution of the ambiguity function of the transmitted signal and the scattering function of the reflection complex.

One of the important applications of scattering functions is for describing reverberation or clutter. This is signal-generated noise introduced by reflections and scattering not originating from the desired target. Increasing the signal energy in a reverberation-limited environment does not improve system performance, since it increases both the target and reverberation levels; consequently, good signal design is the only way to improve performance. For this, one can demonstrate that one wants to minimize the overlap of the result of convolution of the ambiguity function and the scattering function modeling the reverberation at the target range–Doppler location.

5.2 Communications and Seabed Mapping

Underwater acoustic telemetry is another type of active sonar system. The objective is to transmit a message, or a sequence of symbols, from a source to a user. Generally, one wants to do this at a maximum rate with a minimum number of errors. In acoustic telemetry this performance is limited by several characteristics of the underwater environment. First, the channel bandwidth is constrained by strong attenuation at high frequencies; second, the channel has both travel-time spreading due to the many different propagation paths and Doppler-shift spreading due to surface and internal waves and source-receiver motion. The power levels are often low because of restricted energy storage in typical applications, such as unmanned underwater vehicles and ocean-bottom sensors. These factors have made acoustic telemetry design a very challenging signal-processing problem.

In a communication or telemetry system, the input is a symbol sequence represented by a binary bit stream. These symbols may be encoded for cryptographic purposes, but, in any case, they are further encoded as a sequence of transmitted wave forms. The latter encoding algorithms are the subject of an extensive literature in communications theory (Proakis, 1989, 1990). Generally, there are two approaches: *block encoding,* where a fixed number of symbols are encoded together as a block, and *convolutional encoding,* where individual symbols are encoded using a linear, finite-field system. In both approaches additional (redundant) components are included in the encoded wave form for reliability, and the encoded signals, obtained from an alphabet of wave forms, are transmitted to the receiver.

The receiver design must accomplish several tasks. In the simplest, it must determine the sequence of blocks, for block encoding, or perform deconvolution to determine the system input, for the convolutional encoder. In practice, this is complicated by including

1. probing signals used to calibrate for the multipath propagation,
2. synchronization of the timing between source and receiver,
3. Doppler reference time for Doppler compensation, and
4. adaptive equalization for combatting multipath and intersymbol interference.

Communications in the undersea channel is especially challenging and depends a great deal on the acoustic environment. High-rate telemetry in shallow water where the range is much greater than the water depth is probably the most difficult; however, long-range telemetry at modest rates in deep water is certainly difficult as well. Presently, it is an area where a lot of signal-processing work is ongoing.

Maps and charts of the seabed are generated by plotting estimates of depths obtained by a sequence of reflections from the seabed (deMoustier, 1988). A single sounding does not convey an appreciable amount of information; a grid of soundings to determine the topographic relief is usually required. Systems can be as simple as plotting the matched-filter output for a sequence of transmissions as a ship steams a track, to very sophisticated multibeam systems that produce contour charts of bathymetry along a track. A variety of processing methods are used; some simply compensate for the finite beamwidth of the sounding system, while others compile the grid of points, interpolate the data, and contour the relief.

5.3 Travel-Time Tomography

Tomography generally refers to applying some form of inverse theory to observations in order to infer properties of the propagation medium (Munk, 1995). The received field from a source emitting a pulse is spread in time as a result of multipath structure in which different paths have different arrival times (or group speeds) that are determined by the sound-speed distribution in the ocean. Hence the arrival times sample the acoustic medium. (In contrast, medical tomography utilizes the different attenuation of the paths rather than arrival time for the inversion process.) Since the sound speed of the ocean is a function of temperature and other oceanographic parameters, the arrival structure can ultimately be related to a map of these oceanographic parameters. The same principle can be applied to the temperature and wind-velocity parameters in the atmosphere (Wilson and Thomson, 1994). In its most primitive form, the inversion can be described by an algorithm discretizing the ocean into cells, computing travel times from candidate acoustic paths through these cells, and solving a set of equations that equate these travel times to the measured travel times. Each cell is characterized by an unknown sound speed. Typically, some base-line oceanographic information is known so that one searches for departures from this base-line information. Tomographic experiments have been performed at ranges greater than 1000 km. Pulse-compression methods using sequences of signals are often employed in ocean tomographic experiments to enhance bandwidth and travel-time accuracy and received signal strength.

GLOSSARY

Array Gain: A measure of the effectiveness of a beam forming array. The response to a signal arriving on the main response axis relative to the response to isotropic noise.

Beam Forming: Combining signals from individual sensors in an array to obtain significantly greater response to signals that arrive from a specific direction. Beam forming takes into account the geometry of the array and the shape of the arriving wave front. By adjusting processing parameters the directional response of the array can be adjusted or steered without requiring that array elements be moved.

Beamwidth: The angular extent of the main lobe in the beam response pattern of a beamforming array, usually measured between points at which the response is 3 dB less than that of the main response axis.

Beam Response Pattern: The output of a beam-forming array as a function of the direction of arrival of the signal. The pattern generally exhibits a distinct maximum in the direction of the main response axis and a series of local maxima and minima that form a pattern of lobes. The lobe on the MRA is called the main lobe; all others are called sidelobes. It is usually desirable to achieve a narrow main lobe and small sidelobes.

Complex Signal: A signal containing both real and imaginary components. Generally, the real component is obtained from measurements, and the imaginary component is reconstructed from the real component.

Far Field: The sound field at distances from the source that are much greater than the signal wavelength and the source diameter.

Finite Impulse-Response (FIR) Filter: A linear time-invariant system whose response to an impulse has limited duration.

Frequency-Domain Beam Forming: Beam forming by applying a (complex coefficient) linear transformation to each frequency component output from the elements of an array.

Infinite Impulse-Response (IIR) Filter: A linear time-invariant system whose response to an impulse has infinite extent as a result of a feedback loop in the system.

***I-Q* Demodulation:** One of several methods for reconstructing a complex signal from its real component. This method is applicable to modulated pure tones.

Main Response Axis (MRA): The direction at which an array has its maximum response. The response is usually normalized to unity or 0 dB.

Matched-Field Processing (MFP): Simultaneous determination of source location and environmental parameters based on a propagation model and acoustic signals received at an array. Matched-field tomography is the application of this technique to estimating environmental parameters, as when the source location is known.

Matched Filter: A filter designed to maximize the signal-to-noise ratio of a receiver. For white noise, the filter impulse response is a time-reversed model of the signal.

Maximum-Length Sequence (or ML Sequence): A periodic signal with an autocorrelation function that is zero except at delays of zero or multiples of the signal period. These signals are generated by specialized pseudorandom-number generators that cycle through the entire output range in each period of the ML sequence.

MUSIC: An adaptive array-processing method for finding directional sources based upon a singular-value decomposition (SVD) of the sample spectral covariance matrix.

Minimum-Variance Distortionless Processor (MV): An adaptive processing method where the output variance in response to a noise field is minimized while the gain in the steered signal direction is unity producing no distortion of signal incident from the steered direction.

Power Spectral Density (PSD): The distribution of signal power in frequency. For practical measurements with finite record lengths the estimated signal power at each frequency is normalized to a unit spectral interval, such as 1 Hz. The power spectral density of white noise is independent of record length, but the peak in the PSD for a pure tone will increase in proportion to the record length.

Sample Spectral Covariance Matrix: An estimate of the spectral covariance matrix of a signal field on an array of sensors. It is a function of frequency.

Sidelobes: Angular regions away from the MRA where an array has relatively strong responses, e.g., -20 dB (or 0.1) compared to the maximum response of the MRA.

Singular-Value Decomposition (SVD): Representation of a linear transformation in terms of its eigenvectors. This technique is used to decompose the output of an array into mutually orthogonal signal and noise spaces. Beam formers employing SVD techniques are called eigenvector beam formers.

Sound Pressure Level (SPL): Mean acoustic pressure expressed in decibels relative to 20 μPa in the atmosphere and 1 μPa in the ocean.

Time-Domain Beam Forming: Beam forming by applying a time delay to the output of each array element and summing the result. Various gains may be applied to the individual elements prior to summing, these being referred to as tapers.

Tomography: A volumetric imaging technique using multiple planimetric views. In underwater and atmospheric acoustic signal processing, an array of sensors and an array of sources are used to determine the velocity of sound within a specific volume of the medium to investigate inhomogeneities.

White Gaussian Noise: Random noise that is temporally uncorrelated and has a Gaussian sample distribution.

Zoom Filtering: A variable-rate filtering technique used for high-resolution narrowband spectral analysis.

Works Cited

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis,* 2nd ed., New York: Wiley.

Attenborough, K., Teherzadeh, S., Bass, H. E., Di, X., Raspet, R., Becker, G. R., Güdensen, A., Chrestman, A., Daigle, G. A., L'Espérance, A., Gabillet, Y., Gilbert, K. E., Li, Y. L., White, M. J., Naz, P., Noble, J. M., van Hoof, H. A. J. M. (1995), *J. Acoust. Soc. Am.* **97,** 173.

Baggeroer, A. B. (1978), in: A. V. Oppenheim (Ed.), *Applications of Digital Signal Processing,* Englewood Cliffs, NJ: Prentice Hall.

Baggeroer, A. B., Kuperman, W. A., Mikhalevsky, P. N. (1993), *IEEE J. Ocean Eng.* **OE-18,** 401–424.

Bellanger, M. (1984), *Digital Processing of Signals,* Toronto: Wiley.

Burdic, W. S. (1984), *Underwater Acoustic Signal Analysis,* Englewood Cliffs, NJ: Prentice Hall.

Burgess, J. C. (1992), *J. Acoust. Soc. Am.* **91,** 1525–1530.

Clay, C. S., Medwin, H. (1997), *Acoustical Oceanography,* New York: Wiley-Interscience.

Collins, M. D. (1991), *J. Acoust. Soc. Am.* **90,** 1410–1422.

Daigle, G. A., Piercy, J. E., Embleton, T. F. W. (1983), *J. Acoust. Soc. Am.* **74,** 1505–1513.

deMoustier, C. (1988), *Int. Hydrographic Rev.* **65,** 25.

Gilbert, K. E., Di, X. (1993), *J. Acoust. Soc. Am.* **94,** 2343–2352.

Harris, F. J. (1978), *Proc. IEEE* **66** (1), 51–83.

Havelock, D. I., Di, X., Daigle, G. A., Stinson, M. R. (1995), *J. Acoust. Soc. Am.* **98,** 2289–2302.

Jensen, F. B., Kuperman, W. A., Porter, M. B., Schmidt, H. (1994), *Computational Ocean Acoustics,* New York: AIP Press.

Johnson, D. H., Dudgeon, D. E. (1993), *Array Signal Processing: Concepts and Techniques,* Englewood Cliffs, NJ: PTR Prentice Hall.

Kramer, C. (1965), *Electron. Eng.* **37,** 465–467.

Krolik, J. L. (1992), *J. Acoust. Soc. Am.* **92,** 1408–1419.

L'Esperance, A., Gabillet, Y., Daigle, G. A. (1995), *J. Acoust. Soc. Am.* **98,** 570.

Marple, S. L. (1987), *Digital Spectral Analysis with Applications,* Englewood Cliffs, NJ: Prentice Hall.

Munk, W. (1995), *Ocean Acoustic Tomography,* Cambridge, U.K.: Cambridge Univ. Press.

Oppenheim, A. V., Schafer, R. W. (1989), *Discrete-Time Signal Processing,* Englewood Cliffs, NJ: Prentice Hall.

Panofsky, H. A., Dutton, J. D. (1984), *Atmospheric Turbulence, Models and Methods for Engineering Applications,* Toronto: Wiley.

Pierce, A. D. (1989), *Acoustics: An Introduction to Its Physical Principles and Applications,* New York: Acoustical Society of America.

Piercy, J. E., Embleton, T. F. W., Sutherland, L. C. (1986), in: A. L. Sáenz, R. W. B. Stephens (Eds), *Noise Pollution,* New York: Wiley.

Proakis, J. G. (1989), *Digital Communications,* New York: McGraw-Hill.

Proakis, J. G. (1990), *IEEE J. Ocean Eng.* **OE-16,** 21–31.

Rabiner, L. R., Gold, B. (1975), *Theory and Application of Digital Signal Processing,* Englewood Cliffs, NJ: Prentice Hall.

Rabiner, L. R., Schafer, R. W. (1978), *Digital Processing of Speech Signals,* Englewood Cliffs, NJ: Prentice Hall.

Richardson, A. M., Nolte, L. W. (1991), *J. Acoust. Soc. Am.* **89,** 2280–2284.

Schmidt, R. O. (1986), *IEEE Trans. Antennas Propag.* **34,** 276–280.

Schmidt, H., Baggeroer, A. B., Kuperman, W. A., Scheer, E. K. (1990), *J. Acoust. Soc. Am.* **88,** 1851–1862.

Tolstoy, A. (1993), *Matched Field Processing for Underwater Acoustics,* Singapore: World Scientific.

VanTrees, H. L. (1971), *Detection Estimation and Modulation Theory,* New York: Wiley.

Westerfield, E. C., Prager, R. H., Stewart, J. L. (1960), *IRE Trans. Inf. Theory* **6,** 342–348.

Wilson, D. K., Thomson, D. W. (1994), *J. Atmos. Oceanic Technol.* **11** (3), 751–769.

Wong, G. S. K., Embleton, T. F. W. (1995), *AIP Handbook of Condenser Microphones: Theory Calibration and Measurement,* Woodbury: AIP Press.

Further Reading

Brown, E. H., Freeman, F. H., Jr. (1978), *Rev. Geophys. Space Phys.* **16** (1), 47–110.

Jensen, F. B., Kuperman, W. A., Porter, M. B., Schmidt, H. (1994), *Computational Ocean Acoustics,* New York: AIP Press.

Johnson, D. H., Dudgeon, D. E. (1993), *Array Signal Processing: Concepts and Techniques,* Englewood Cliffs, NJ: PTR Prentice Hall.

Knight, W. C., Pridham, R. G., Kay, S. M. (1981), *IEEE Proc.* **69,** 1451–1507.

Kuperman, W. A. (1991), "Acoustics, Underwater," in: G. L. Trigg (Ed.) *Encyclopedia of Applied Physics,* Vol. 1, New York: VCH Publishers, pp. 293–327.

Marple, S. L. (1987), *Digital Spectral Analysis with Applications,* Englewood Cliffs, NJ: Prentice Hall.

Munk, W. (1995), *Ocean Acoustic Tomography,* Cambridge, U.K.: Cambridge Univ. Press.

Oppenheim, A. V., Schafer, R. W. (1989), *Discrete-Time Signal Processing,* Englewood Cliffs, NJ: Prentice Hall.

Panofsky, H. A., Dutton, J. D. (1984), *Atmospheric Turbulence, Models and Methods for Engineering Applications,* Toronto: Wiley.

Papoulis, A. (1965), *Probability, Random Variables, and Stochastic Processes,* Toronto: McGraw-Hill.

Rabiner, L. R., Schafer, R. W. (1978), *Digital Processing of Speech Signals,* Englewood Cliffs, NJ: Prentice Hall.

Tolstoy, A. (1993), *Matched Field Processing for Underwater Acoustics,* Singapore: World Scientific.

UNIFIED FIELD THEORIES

Hitoshi Murayama, *Department of Physics, University of California, Berkeley, California, U.S.A., and Theoretical Physics Group, Lawrence Berkeley National Laboratory, University of California, Berkeley, California, U.S.A.*

	Introduction	1
1.	**Prototype Unification**	1
2.	**Four Forces**	2
2.1	Gravity	2
2.2	Electromagnetism	2
2.3	Strong Force	3
2.4	Weak Force	3
2.5	Similarities and Differences among Four Forces	4
2.6	Hints for Unification	4
3.	**Grand Unified Theories**	5
3.1	Original GUTs	5
3.2	Supersymmetric GUTs	6
4.	**Kaluza–Klein Theories**	8
5.	**Superstring Theory**	9
6.	**Summary**	10
	Glossary	11
	Works Cited	11
	Further Reading	13

INTRODUCTION

The idea of unification is to look for a single theory from which all observed phenomena can be ultimately derived. Even though the idea appears unrealistically optimistic at first sight, it has a good grounding in the context of elementary-particle physics. In fact, history has shown us examples that many varied phenomena could be understood in terms of a single and simple theory. This article intends to explain historic examples, attempts at finding a unified description of different phenomena, and the most recent activities.

1. PROTOTYPE UNIFICATION

Maxwell's theory of electromagnetism has been regarded as one of the best examples of unification. Another good example is the unified theory of electromagnetism and weak interactions. In both cases, there were early experimental hints of a need for a unified description, theoretical progress toward such a direction, prediction of new phenomena based on unified theories, and experimental confirmation of the unified theories.

By the end of the 19th century, it was clear that Faraday's notion of the field is the most useful description of both electric and magnetic forces. This was probably the only theoretical similarity between these two apparently distinct forces. On the other hand, there were experimental indications that the two forces were related. An electric current creates a magnetic field around it. A change in the magnetic flux induces an electric potential.

Maxwell pushed it to the extreme. Since a change in the magnetic flux can induce an electric field, they might be two manifestations of a single force. If this is true, a change in the electric field must also induce a magnetic field. Even though there were no experimental indications of it by then, Maxwell postulated the existence of such a new phenomenon, which allowed him to formulate the theory of electric and magnetic fields on an equal footing. In particular, such a unified description predicted the existence of an electromagnetic wave, which consists of electric and magnetic fields without any sources, propagating in the vacuum by inducing each other. Finally, the existence of electromagnetic waves was confirmed and Maxwell's theory was established.

Furthermore, Maxwell's theory of electromagnetism lead to a further revolution in physics. Maxwell's equations happen to have a symmetry that was first recognized by Lorentz and Poincaré. The symmetry, known as Lorentz invariance, was later emphasized as a true symmetry of space-time by Einstein in

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

his special theory of relativity, which required us to change our basic notion of absolute time.

The Glashow–Weinberg–Salam theory of electromagnetic and "weak" forces is another example that provided a unified description of two different forces by a single theory that further predicted a new force, the neutral-current weak interaction, and a new particle, the Higgs boson. We will discuss the electroweak theory further in the next section, after briefly describing the four known forces in nature.

2. FOUR FORCES

There are at least four known basic forces in nature: gravity, electromagnetism, "weak," and "strong" forces. Here, weak and strong refer to the proper names of two kinds of forces, which reflect their strengths as well.

The four forces manifest themselves in quite distinctive ways. Below are their brief descriptions. See Table 1.

2.1 Gravity

The gravitational force was first formulated in Newton's theory of gravity. It has an infinite range with a potential $V(r) \propto 1/r$ and acts on any bodies with a mass. Its remarkable feature is that the motions induced by gravity do not depend on the composition or mass of the body. Any body follows the same trajectory. This is called the equivalence principle. Because of this property, Einstein proposed his theory of general relativity, which describes gravity as a curvature of four-dimensional space-time. According to this theory, moving bodies follow geodesics: they appear to fall because the space-time is curved. This is a natural explanation of why gravity exhibits the equivalence principle. The quantum theory of gravity is still not available; in such a quantum theory, gravity is supposed to be mediated by a spin-2 carrier, the graviton. There is a single coupling constant in the theory, Newton's constant $G_N = (M_{\text{Planck}})^{-2} = (1.2 \times 10^{19}\ \text{GeV})^{-2}$, which has dimensions of inverse mass squared in the natural units $\hbar = c = 1$.

2.2 Electromagnetism

The electromagnetic force also has an infinite range with a Coulomb potential $V(r) \propto 1/r$. It acts on any matter with electric charges. Like gravity, the form of the interaction is completely determined only by the charge of an object, irrespective of its composition. This universality is now understood on the basis of the gauge principle, which allows us to change the phase of quantum-mechanical wave functions or quantum fields locally. The coupling constant of the theory is the fine-structure constant, $\alpha = e^2/4\pi\epsilon_0\hbar c \simeq 1/137$, which is a dimensionless number. With only this one parameter, electromagnetism and especially its quantum version (quantum electrodynamics, or QED) are extraordinarily successful and are in agreement with experimental data at the precision of 10^{-11} (Kinoshita, 1995). As a result of quantum effects, the fine-structure constant actu-

Table 1. Comparison of four known forces in Nature. The existence of the graviton is not confirmed yet. See the text for the meanings of "Universality" and "Gauge invariance."

	Gravity	Electromagnetism	Strong	Weak
Range	∞	∞	10^{-13} cm	10^{-16} cm
Force carrier (spin)	Graviton (?) (2)	Photon (1)	Gluon (1)	W, Z bosons (1)
Gauge invariance	General coordinate tr.	U(1)	SU(3)	SU(2)
Universality	Equivalence principle	Electric charge	SU(3) rep.	SU(2) rep.
Source	Anything with energy	Charged particles	Hadrons	Left-handed quarks, leptons
Theory	General relativity (classical)	Maxwell's theory (classical/ quantum)	QCD (quantum)	Electroweak theory (quantum)
Phase	Coulomb(?)	Coulomb	Confinement	Higgs
Coupling constant	Newton's constant	Fine-structure constant	Strong-coupling constant	Weak-coupling constant

ally varies slowly with energy scales, growing for larger energies.

2.3 Strong Force

The strong force is an interaction among the hadrons, such as baryons (proton, neutron, hyperons, etc.) and mesons (π, K, ρ, etc.), and is responsible for various phenomena, which include nuclear α decay, fission, fusion, the nuclear binding force, and hadron resonances. It has a finite range $\simeq 10^{-13}$ cm and appeared completely different from the above two forces. Now the hadrons are understood as composites made up of quarks: baryons as bound states of three quarks, and mesons of a quark and an antiquark. The quarks come in three kinds, called "colors," and one can rotate three kinds of colors arbitrarily [SU(3) gauge invariance] analogous to the phase rotation in QED. The color quantum number is the source of the gluon field, just like the electric charges for the electric field. The theory is characterized by a dimensionless coupling constant, the strong-coupling constant α_s. The gluon mediates a long-range force, but with a much stronger variation with energy scales than the fine-structure constant. Experimentally, $\alpha_s(91 \text{ GeV}) = 0.118 \pm 0.003$, which is much stronger than the fine-structure constant. The strong-coupling constant is smaller at higher energies (asymptotic freedom), grows logarithmically at lower energies, and becomes practically infinite around 300 MeV. Because of this behavior of the coupling constant, the potential among colored particles is approximately $V(r) \propto r$ at large distances $r > 10^{-13}$ cm, and the quarks are "confined" in bound states; i.e., it requires an infinite amount of energy to separate a quark out from a bound state.

The observed strong interactions among hadrons are understood as residual force among the bound states similar to the van der Waals force (residual Coulomb force) among the atoms (electron–nucleus composites). Such a strong residual force exhibited highly nonuniversal (composition and mass dependent) and nonperturbative phenomena, which hindered the formulation of theory until the 1970s. The experimental discovery that there is an approximate scaling in the deep inelastic electron–nucleon scattering experiments at SLAC in the late 1960s by the SLAC-MIT group (Bloom *et al.*, 1969; Breidenbach *et al.*, 1969), and a theoretical discovery that a non-Abelian gauge theory can exhibit asymptotic freedom by Gross and Wilczek (1973) and Politzer (1973), enabled people to formulate quantum chromodynamics (QCD) based on SU(3) gauge invariance. Because QCD becomes perturbative at high energies, many quantitative tests of QCD were and are being performed at high-energy experiments successfully.

2.4 Weak Force

The weak force manifested itself in nuclear β decay first and was later observed in various meson decays, muon decays, and neutrino interactions. It has the smallest range among the forces observed, $\simeq 10^{-16}$ cm. Fermi formulated the first theory of weak force, which showed a remarkable universality. The β-decay lifetimes of many different nuclei and mesons could be described by a single Fermi constant $G_F = 1.1664 \times 10^{-5}$ GeV^{-2}. In particular, the strengths of the weak force acting on the muon and "up" quark are measured accurately and agree at the level of a permille. Glashow (1961) speculated that the weak force could also be a gauge force based on $SU(2)_L \times U(1)_Y$ group because of this universality. An important point here is that the β decay changes the electric charge of the parent nuclei ("charged-current weak interaction"). If the weak force is mediated by a gauge boson (named the W boson) similar to the photon or gluon, it must carry an electric charge. The only way a gauge boson (W) can interact with another gauge boson (photon) is when they belong to the same non-Abelian gauge group [SU(2) in this case]. Therefore, a gauge theory of the weak interaction needs also to incorporate the electromagnetic force on equal footing: a need for a unified description. To be consistent with the observed pattern of electric charges, Glashow needed another group, $U(1)_Y$, and predicted a new kind of weak force: the neutral-current weak interaction mediated by the Z boson. Such a new type of weak force was discovered later by neutrino experiments, first by the Gargamelle Collaboration (Hasert *et al.*, 1973), and is now tested at permille accuracies at LEP (an e^+e^- storage ring at CERN, Geneva) and SLC (a linear e^+e^- collider at Stanford Linear Accelerator Center) (for a review, see Particle Data Group, 1996, p. 85). One feature of the weak force that turns out to be crucial in building unified models is that it distinguishes the left-handed from right-handed helicity states of particles. Thus the weak force is said to be *chiral,* as opposed to the other three forces that do not distinguish the helicities of particles of particles: *vector-like.*

A puzzle in Glashow's theory is why the weak force mediated by the *W* or *Z* boson is short ranged, while the theory also contains the photon. The explanation to this question is the Higgs mechanism. In the Meissner effect in a superconductor, a condensation of charges (Cooper pairs) results in a finite range of a gauge force, namely, the finite penetration length of the magnetic field into a superconductor. In a similar fashion, a condensation of the yet-to-be-found Higgs boson causes a finite range of the weak force. This theory was formulated by Weinberg (1967) and Salam (1968), based on a spontaneous breakdown of the SU(2) × U(1) gauge group to the electromagnetic U(1) gauge group by the Higgs mechanism at the energy scale of 250 GeV. It was later summarized as the standard model of particle physics together with QCD, based on the gauge group SU(3) × SU(2) × U(1).

2.5 Similarities and Differences among Four Forces

To arrive at a unified description of these forces, we need to identify the similarities and differences among them. One notable fact is that all four forces are based on gauge principles: general coordinate transformation (gravity), Abelian gauge invariance (electromagnetism), or non-Abelian gauge invariance (strong and electroweak forces). It is this fact that led to the idea that all four forces may have a unified description. The ranges of forces are quite diverse, but they are understood as a consequence of different phases that each gauge theory has: Coulomb (gravity and electromagnetism), confinement (strong), and Higgs (electroweak) phases. There are, however, two major differences between gravity and the other three.

1. Gravity is related to the space-time geometry described by a spin-2 field, the space-time metric. On the other hand, the other gauge fields are related to internal spaces, such as phase [U(1)], weak isospin [SU(2)], or color [SU(3)] degrees of freedom, and are described by vector fields, which carry spin 1.
2. The strength of gravity is described by the Newton constant with mass dimension -2 while the other three are described by dimensionless constants, such as the fine-structure constant in the natural units $\hbar = c = 1$.

These two points lead to a profound difference in the quantum field theory. Theories with constants with negative mass dimensions are not renormalizable; *i.e.,* there are irremovable infinities in the theory; while the ones with dimensionless coupling constants are. This fact makes the unification of electromagnetic, strong, and weak forces relatively easy from the field-theoretical point of view while the incorporation of gravity is rather difficult.

It must be noted that there are other kinds of forces in the standard model of particle physics. There is a self-interaction of the Higgs boson that generates the mass of the Higgs boson by its own condensate, and many Yukawa interactions of the quarks and leptons with the Higgs boson that generate their masses. They are all characterized by dimensionless coupling constants. They are, however, not based on a gauge principle by any means, and such interactions can be introduced with arbitrary strengths as opposed to the universal character of the gauge forces. They are often not called "forces" because we have not yet detected the effect of such interactions experimentally. However, they are present in the standard model and need to be understood as well if there is to be a unified description of *all* forces. In particular, the Yukawa interactions have a rather baroque pattern of strengths in order to reproduce the wide range of quark and lepton masses, i.e., 0.511 MeV/c^2 for the electron to 175 GeV/c^2 for the top quark discovered in 1995 (Abe *et al.*, 1995; Abachi *et al.*, 1995). A truly unified theory must also explain the pattern of Yukawa interactions.

2.6 Hints for Unification

Finally, we list two puzzles in the standard model that suggest a need for a possible unified description. In the standard model, the quarks and leptons appear completely independent from each other, and the different forces [namely, SU(3), SU(2) and U(1) forces] are *a priori* not related to each other. However, then, it is not clear in the standard model why the electric charges of both quarks and leptons come in integer multiples of $\frac{1}{3}$. Such a quantization is essential to guarantee the electric neutrality of matter, which balances the charge of a proton made up of two up and one down quarks, with the charge $\frac{2}{3} \times 2 + (-\frac{1}{3}) = 1$, and the charge of an electron -1. This neutrality is tested at 10^{-21}. Second, as a result of quantum effects, the gauge invariance could be lost (called anomaly), which potentially spoils the unitarity or renormalizability of the theory. In the standard model, however, this does not occur because of a nontrivial cancellation. For instance, it is necessary that the sum of all electric charges vanish in the standard-model particle content, which is true only after adding the contribution of the quarks *and* leptons, *and* with the correct multiplicities of color [3 for SU(3)], *and* the correct multiplicities for left-handed particles [2 for SU(2)]. Other necessary conditions are also satisfied thanks to contributions of both quarks and leptons with the correct multiplicities among the various groups. This observation suggests that different gauge groups are related at a deep level, and so are quarks and leptons.

3. GRAND UNIFIED THEORIES

If one leaves out gravity, the unification of the other three forces is not difficult because one can maintain the renormalizability of the quantum theory (for a detailed treatment, see Ross, 1984).

3.1 Original GUTs

The earliest model toward the unification of the electromagnetic, weak, and strong forces is probably the one by Pati and Salam (1973), where a gauge group SU(4) × SU(2) × SU(2) was employed. Later, Georgi and Glashow (1974) proposed a model based on a simple group SU(5), which became the reference group in almost all subsequent efforts. It was a remarkable observation that all gauge forces and quite complicated quantum numbers of quarks and leptons in the standard model fit nicely into a single gauge group SU(5) (see Table 2). Subsequently, Georgi *et al.* (1974) showed that the strong force becomes weaker at higher energies while the electromagnetic force becomes stronger, so that they have the same strength at a certain high energy: around 10^{14} GeV in the simplest example. This observation leads to the following picture of a possible unification of forces. The gauge forces (except gravity) are ultimately described by a simple group such as SU(5) [or SO(10) or E_6, as described below] with a single coupling constant. It breaks spontaneously to the SU(3) × SU(2) × U(1) gauge group at an extremely high energy scale (unification scale) such as 10^{14} GeV, and the gauge coupling constants of different groups run differently between the unification scale and the weak scale 250 GeV. A simple calculation shows that if one follows the running from the unification scale down to the weak scale, the coupling constant of a larger gauge group becomes larger than that of a smaller gauge group. Therefore, this picture, called grand unified theory or GUT, explains why the strong force is strong and the electromagnetic force is weak simply on the basis of the size of the groups, 3 > 2 > 1. Furthermore, quantization of the electric charge is ex-

Table 2. The quantum numbers of the quarks and leptons under the standard-model gauge groups and the way they can be embedded into representations under the grand unified gauge groups. In the first row, the first number is the number of color degrees of freedom under QCD, the second whether the particle couples to the electroweak group as doublets (**2**) or not (**1**), and the last the charge under the U(1) hypercharge group. In the latter two rows, the numbers indicate the dimensionality of SU(5) or SO(10) representations. The subscripts *L*, *R* refer to left- or right-handed helicity states.

	Q_L	u_R	e_R	d_R	L_L	ν_R
SU(3) × SU(2) × U(1)	(**3**, **2**, 1/6)	(**3**, **1**, 2/3)	(**1**, **1**, −1)	(**3**, **1**, −1/3)	(**1**, **2**, −1/2)	(**1**, **1**, 0)
SU(5)		**10**			**5***	**1**
SO(10)				**16**		

plained as a simple consequence of the group theory.

An interesting feature of SU(5) unification is that the quarks and leptons belong to the same multiplets under SU(5), so that they can transform into each other by the exchange of SU(5) gauge bosons, in the same way that an up quark and a down quark can transform into each other through the charged-current weak interaction mediated by the exchange of the *W* boson. Therefore, the SU(5) unification predicts that a quark can turn into a lepton, which allows a proton to decay (*e.g.*, $p \rightarrow e^+\pi^0$ in the simplest model), a new phenomenon that can be looked for by experiments. Such a violation of baryon number may explain the reason why the universe contains matter (protons, neutrons, and electrons) but no antimatter (antiprotons, anti-neutrons, and positrons) as a result of baryon-number nonconserving interactions that occurred when the universe was as hot as the unification energy scale (baryogenesis) (Yoshimura, 1978; Ignatiev *et al.*, 1978). Finally, it was also pointed out that the masses of the bottom quark and tau lepton, which differ roughly by a factor of 3 in laboratories, also become the same size when extrapolated up to a similar high energy scale. This observation led to a hope that the pattern of quark and lepton masses could be understood also in the context of unified theories.

Unfortunately, the original SU(5) unified model is now excluded for two reasons. First, the nonobservation of proton decay, especially by the IMB and Kamiokande experiments, has put lower limits on the proton lifetime that exceed the prediction of the model. Second, the LEP and SLC experiments measured the size of the three gauge coupling constants for three gauge groups SU(3) × SU(2) × U(1) accurately, which did not agree with the predicted pattern in the model. There were other problems pointed out concerning the original SU(5) unified model. The assumption that there are two very disparate energy scales (hierarchy), the unification scale and the weak scale, appears unnatural, as is discussed more below. The cancellation of the anomaly is still not obvious because one needs to add contributions of two multiplets (**5*** and **10**) under SU(5). The theory predicts the existence of magnetic monopoles of a mass higher than the unification scale, which have never been detected. The unification of quark and lepton masses works for the third generation (bottom and tau) but not so well for the first and second generations.

The question of anomaly cancellation led people to consider yet larger groups whose irreducible multiplet contains the whole generation that belonged to two irreducible multiplets under SU(5). One such candidate is SO(10), which carries certain unique features (see Table 2). It is the smallest group that allows chiral fermions while being automatically anomaly free. It predicts another particle, the right-handed neutrino, for each generation. If the right-handed neutrinos are allowed to acquire masses at a high scale, such as the unification scale, they become irrelevant to typical experiments. However, the diagonalization of the neutrino mass matrix leads to small but finite masses for the left-handed neutrinos, given approximately by $m_\nu \sim m_u^2/M$, where M is the mass of the right-handed neutrinos and m_u is the mass of the up-type quark in the same generation. Therefore, a large M results in a small m_ν well below 1 eV/c^2 (Yanagida, 1979; Gell-Mann *et al.*, 1979), which may explain the observed deficit in the solar neutrino flux, a part of the missing mass in the universe (hot dark matter), or various other observations in neutrino experiments (see Fukugita and Yanagida, 1994, for a review on neutrino physics). People discussed other anomaly-free gauge groups along the same line, such as E_6, or $[SU(3)]^3$.

3.2 Supersymmetric GUTs

The existence of a huge hierarchy between the unification scale and the weak scale posed a serious theoretical challenge. In particular, it was found that the correction to the Higgs boson mass squared from higher orders in perturbation theory is quadratically divergent, and either tends to bring the weak scale up to the unification scale or restores the electroweak group SU(2) × U(1) (no Higgs mechanism). A mechanism to stabilize the hierarchy in energy scales seemed necessary (see Murayama, 1995, for a review on this discussion). Supersymmetry (see Wess and Bagger, 1983, for a textbook) was postulated as such a mechanism. It is well known that the electron acquires a linearly divergent Coulomb self-energy in classical electrodynamics. If one extrapolates the electrodynamics to a much shorter distance scale, the Coulomb self-energy grows linearly with the distance scale and so does the electron mass. However, the existence of the positron and its appearance as a part of quantum fluctuation of the vacuum cancels the linearly divergent Coulomb self-energy, and the result is only logarithmically divergent for small distances. A symmetry, called chiral symmetry, which interchanges an electron with a positron, is responsible for the cancellation of wild behavior of the self-energy. After such a cancellation, quantum electrodynamics can be applied to much shorter distance scales, or, equivalently, much higher energy scales, without ruining the smallness of the electron mass. In a similar fashion, supersymmetry introduces a *superpartner* to each particle in the standard model, whose spins differ by $\frac{1}{2}$ and have opposite statistics (Bose vs Fermi). And their contribution in the quantum fluctuation cancels the quadratically divergent contribution to the Higgs boson mass squared and makes the standard model applicable up to the unification scale without spoiling the smallness of the weak scale.

Therefore, grand unified theories are often discussed with a supersymmetric extension (Dimopoulos and Georgi, 1981; Sakai, 1981). The precise determination of the three gauge coupling constants at LEP and SLC experiments running from 1989 revealed that they seem to unify remarkably well at an energy scale of 2 × 10^{16} GeV if we assume the minimal supersymmetric extension of the standard model, while they do not unify at all with a nonsupersymmetric standard model (see Fig. 1). This observation led to a renewed enthusiasm for supersymmetric grand unified theories. Superpartners of the standard-model particles are currently under intensive searches at various collider experiments, such as the upgraded LEP and Tevatron, and are expected to be found at latest at the LHC experiments that should begin data taking in 2005 (Baer *et al.,* 1996). The proton lifetime is expected to be within the reach of the Super-Kamiokande experiment that started in 1996, in the simple models (Murayama, 1997). Other experimental signatures of supersymmetric grand unified theories have been pointed out, such as $\mu \to e\gamma$ (Barbieri and Hall, 1994). They are under active experimental efforts as well. If the superparticles are found, there are numerous other tests possible on grand unified theories. The mass spectrum of superparticles has a definite pattern if the forces are grand unified. By measuring the masses of superparticles, one can either verify or exclude grand unified theories, which is possible at a percent level at future collider experiments (Tsukamoto *et al.,* 1995; Hinchliffe *et al.,* 1997).

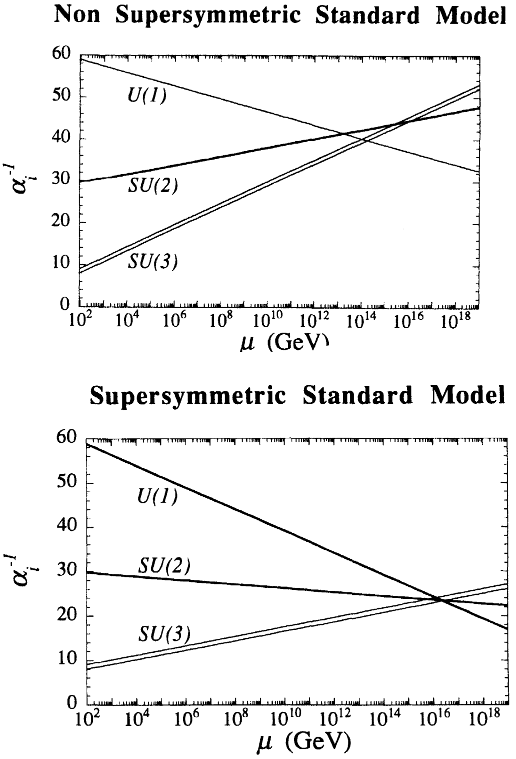

FIG. 1. The running of three gauge coupling constants in the standard model, assuming the minimal nonsupersymmetric standard model or its supersymmetric extension.

The supersymmetric grand unified theories are not without problems, however. Most importantly, supersymmetry *stabilizes* the hierarchy between the unification scale and the weak scale but does not explain why such a hierarchy exists to begin with. One common scheme assumed in literature is that supersymmetry is broken by a strong dynamics of a gauge theory, analogously to the chiral symmetry breaking due to the strong QCD dynamics, which is the origin of the masses of proton and neutron. Such a scheme naturally generates spontaneous supersymmetry breaking at an energy scale $\simeq M_{\text{Planck}}e^{2\pi/\alpha b_0}$ much lower than the unification scale or Planck scale, if the gauge coupling constant α is perturbative at the Planck scale, and if the gauge theory is asymptotically free implying $b_0 < 0$. Furthermore, the effect of broken supersymmetry is mediated to the standard-model particles only by gravitational interactions (Chamseddine *et al.*, 1982; Hall *et al.*, 1983). Therefore, incorporation of gravity into a unified theory appeared necessary.

Supersymmetry actually suggests a possibility to relate the gauge forces to gravity, whose force carriers have spin 1 and 2, respectively. Since supersymmetry relates particles with different spins, they may be related as well. Furthermore, supersymmetry can be promoted to a gauge invariance, *local* supersymmetry, which necessarily incorporates gravity (supergravity). Finally, supersymmetry makes the theory behave more reasonably at high energies, so that gravity, which has a severe problem of divergences at high energies and hence is nonrenormalizable, may be better behaved at the quantum level once it is made supersymmetric.

Supergravity itself does not make the quantum theory of gravity fully consistent. It still is nonrenormalizable. In the early 1980s there was a speculation that supergravity with the maximal supersymmetry (de Wit and Freedman, 1977; Cremmer and Julia, 1978, 1979; de Wit, 1979) (called $N = 8$ because it possesses eight times more supersymmetries than the smallest supergravity theory) may be finite and hence consistent. It also had a large internal symmetry that allowed 28 gauge bosons (de Wit and Nicolai, 1981, 1982; Hull, 1984), and, people hoped, it may contain the standard-model gauge groups. Unfortunately, neither of the expectations turned out to be true, and this direction of research has been basically abandoned. Possible counter terms were found at the seven-loop level (Howe and Lindstrom, 1981), which indicated that the theory has ultraviolet divergences. It was difficult to embed the standard-model gauge groups into the internal symmetry of the theory. And people could not embed chiral matter content into the theory. Supergravity is now regarded as an effective description of a true quantum theory of gravity valid only up to the energy slightly below the Planck scale.

Recently, a mechanism to mediate the effect of supersymmetry breaking to the standard model particles by usual gauge forces has attracted interest (Dine and Nelson, 1993; Dine *et al.*, 1995, 1996). In such a case, supergravity does not appear crucial for the purpose of generating supersymmetry-breaking effects that are phenomenologically necessary. The mechanism has a phenomenologically appealing feature that the quantum contribution of superparticles to rare phenomena are naturally suppressed below the experimental limits. There is ongoing active research to understand the phenomenological consequences of this mechanism (Dimopoulos *et al.*, 1996; Babu *et al.*, 1996). It appears, however, relatively contrived at the moment, and a substantial effort is being devoted to simplify it. Discussion of the phenomenology of grand unified models with this mechanism has just started (Carone and Murayama, 1996).

4. KALUZA–KLEIN THEORIES

Soon after Einstein's general theory of relativity, people tried to formulate electromagnetism on an equal footing. One possible direction is to obtain the electromagnetic gauge invariance [U(1)] as a part of the general coordinate transformation, which requires the extension of our four-dimensional space-time to higher dimensions. Such theories are called Kaluza–Klein theories according to the authors of early works (Kaluza, 1921; Klein, 1926). In general, theories of this class start with a higher-dimensional space-time with a certain theory of gravity and try to obtain other gauge forces in the standard model upon "compactifying" extra dimensions.

The simplest example by Kaluza and Klein starts with a five-dimensional space-time. One "compactifies" the extra spatial direction to a small circle, and the remaining large four dimensions are identified with our ordinary space-time. The metric tensor of the five-dimensional space-time $G_{\mu\nu}$, $\mu, \nu = 0, 1, 2, 3, 4$, now decomposes into the four-dimensional metric $g_{\mu\nu}$ ($\mu, \nu = 0, 1, 2, 3$), a vector $A_\mu = g_{4\mu}$, and a scalar $S = g_{44}$. The original five-dimensional general coordinate transformation acts as an ordinary gauge transformation on the vector A_μ. In general, an isometry of the compactified space generates an effective gauge invariance in the remaining four-dimensional space-time. Much later, it was found that such a space-time could be a solution to the (modified) Einstein's equation that could be regarded as a "spontaneous" compactification of a high-dimensional space-time.

This simple and beautiful idea suffers from all the difficulties that the Einstein's general relativity has, and more. First of all, quantum theory is not well-defined because the quantum theory of gravity is yet to be understood. Second, electromagnetism, which can be quantized and studied as a renormalizable quantum field theory, becomes nonrenormalizable in higher dimensions. Third, the five-dimensional theory cannot generate gauge fields for strong and weak forces (see, for a review and a collection of articles, Appelquist *et al.*, 1985).

Witten proposed a "realistic" version of the Kaluza–Klein theory (Witten, 1981). In order to incorporate all known forces, strong, weak, and electromagnetic, the space-time dimension has to be equal to or larger than 11. On the other hand, supersymmetry, as explained in the previous section, is probably an essential ingredient in unified theories. A space-time higher than 11 dimensions generates too many supersymmetry generators in four dimensions after compactifications, which results in theories with unwanted higher-spin fields, such as spin $\frac{5}{2}$ or 3, whose existence would make the theory even worse behaved at high energies. Therefore, 11 is the only dimensionality that may be realistic. In fact, solutions to Einstein's equation were found that allow compactified seven-dimensional space with SU(3) × SU(2) × U(1) isometry (Castellani *et al.*, 1984). This observation aroused a great enthusiasm in 11-dimensional supergravity theories (Cremmer *et al.*, 1978). However, the uniqueness of the dimensionality posed a serious difficulty in pursuing the direction further. Witten later (1985) showed that such a theory can never produce chiral fermions in four dimensions. Therefore, at least some extra ingredients (i.e., gauge fields) are needed before compactification to incorporate fully all known forces on equal footing. The apparent absence of the cosmological constant in our four-dimensional space-time was also very difficult to obtain in this framework. Finally, it did not solve the problem concerning the nonrenormalizability of the theory.

The Kaluza–Klein theories, at least with the original framework, do not seem to offer a successful unified description of all four forces. It is noteworthy, however, that 11-dimensional supergravity theories have renewed interest in the context of string duality, which is discussed later.

5. SUPERSTRING THEORY

The (partial) success of supersymmetric grand unified theories led to optimism on the basic concept of unification. There is no compelling and aesthetically appealing concrete model at hand, but the concept itself appears promising. On the other hand, the incorporation of gravity has been a serious challenge even with supersymmetry because of bad behavior of the quantum gravitational effects at high energies.

The string theory (see Green *et al.*, 1987, for a textbook) replaces particles by small strings with a typical size of the Planck length, $\sim 10^{-33}$ cm. The ordinary particles, with masses much less than the Planck mass, are identified with the lowest excitation modes of the string. Historically, the string model of particles was developed by Nambu and Goto to explain properties of hadrons that were not understood by then (see Scherk, 1975, for a review). Later, however, it became clear that the string theory contains massless spin-1 and spin-2 excitations that can potentially be identified with gauge bosons and the graviton. In particular, the equations of motion for spin-2 field derived from the string theory agreed with Einstein's theory of general relativity in low-energy approximation. Therefore, the focus of string theory changed to a consistent framework of quantum gravity (Scherk and Schwarz, 1974).

Green and Schwarz (1984, 1985) found that certain classes of string theory are free from anomalies and are ultraviolet finite. This observation aroused the hope that the string theory provides a consistent framework of quantum gravity. They were classified into the following categories: type I, type

IIA, type IIB, and heterotic. The heterotic string theory (Gross *et al.*, 1985) allowed only two choices for the gauge group, SO(32) or $E_8 \times E_8$, and the latter can incorporate the standard model or grand unified groups such as SU(5), SO(10), or E_6. The other E_8 can act as the sector that breaks supersymmetry well below the Planck scale. The theories could be formulated only in ten-dimensional space-time. It was further shown, however, that the extra six dimensions can be compactified as a solution to the equations of motions, on so-called Calabi–Yau manifolds or orbifolds, such that the remaining four-dimensional space-time has minimal supersymmetry and realistic gauge groups and chiral matter content (Candelas *et al.*, 1985). Furthermore, a string theory contains only a single coupling constant. If the theory is supposed to describe particle physics, it is supposed to explain *all* coupling constants by a single parameter. Therefore, if string theory is correct, it must be the "Theory of Everything."

There are a number of problems in string theory, however, at the moment. One of the major problems is that the string theory allows too many solutions to its equations of motion, which give different matter content, different gauge groups, different coupling constants, and even different space-time dimensions. It was hoped that nonperturbative effects would eliminate most of the solutions and could pick the unique solution that describes our universe. It requires a nonperturbative formulation of string theory, which is still not available. String theory is formulated only in the context of perturbation series at this moment. Phenomenological issues, such as whether the string theory is consistent with the observed pattern of gauge coupling constants (for a recent compilation of discussions, see Dienes *et al.*, 1996), are discussed mostly within perturbation theory.

The invention of string duality (Hull and Townsend, 1995; Witten, 1995) is changing the situation drastically. The behavior of string theory in the nonperturbative regime is now understood much better by using the duality transformation, which maps a strongly coupled regime of one string theory to a weakly coupled regime of another string theory. This is a generalization of electric-magnetic duality in electromagnetism or the duality in the two-dimensional Ising model, which interchanges small and large coupling constants. A much more nontrivial situation was revealed in supersymmetric gauge theories (for a review, see Intriligator and Seiberg, 1996). Similar techniques were employed in string theory, which led to discoveries of string duality. In particular, the heterotic string theory with $E_8 \times E_8$ gauge group, which appears to be the most promising one phenomenologically, is now believed to be dual to the 11-dimensional supergravity with a rather peculiar topology of the 11th direction (called M-theory) (Horăva and Witten, 1996). The theory contains no parameters at the Lagrangian level. The eleventh dimension is compactified on a finite interval $x^{11} \in [0,\rho]$, and the interval ρ plays the role of the coupling constant of the theory. All the gauge and matter fields live only inside the ten-dimensional hyperplanes at the boundaries $x^{11} = 0$ and $x^{11} = \rho$, while the supergravity multiplet can propagate both in ten-dimensional and 11th directions. Since there are gauge fields already present before the compactification to four dimensions, the theory is free from the problem of chiral fermions discussed before in the context of Kaluza–Klein theories. There are indications that the M-theory description is the most appropriate one for quantum gravity and the gauge forces among the elementary particles. Yet many problems are left unsolved, especially on the uniqueness of the solution.

6. SUMMARY

There have been many attempts to have a unified description of all observed forces in nature. Different attempts have different levels of success. Among them, supersymmetric grand unified theories appear to be the most promising phenomenological models that offer us a unified description of electromagnetic, strong, and weak forces. Many experimental efforts are being devoted to detect either superparticles or rare phenomena predicted by grand unified theories, and some of them are within the reach of near-future experiments. The incorporation of gravity, however, requires a formulation of quantum gravity, which is impossible within the standard quantum field theory. The most promising candidate formulation of quantum gravity is the string theory, and its nonperturbative dynamics has just begun to be understood. Concrete phenomenological predictions are yet to be worked out.

ACKNOWLEDGMENTS

The author thanks Christopher Carone for careful reading of and comments on the manuscript. This work was supported in part by the Director, Office of Energy Research, Office of High Energy and Nuclear Physics, Division of High Energy Physics of the U.S. Department of Energy under Contract DE-AC03-76SF00098, in part by the National Science Foundation under Grant PHY-90-21139, and also by the Alfred P. Sloan Foundation.

GLOSSARY

Anomaly: Suppose a theory defined at the level of classical Lagrangian has a symmetry. It sometimes happens that the symmetry is broken once the theory is quantized. This phenomenon in quantum field theory is called anomaly. For instance, the scale invariance of the QCD Lagrangian is broken quantum mechanically and the coupling constant runs as a function of energy; i.e., the theory is no longer scale invariant. This anomaly is needed to explain the asymptotic freedom of strong interaction and does not lead to any inconsistencies. On the other hand, the gauge invariance of a theory may also be broken quantum mechanically in chiral theories, which makes the theory inconsistent; it cannot maintain renormalizability and unitarity simultaneously. Therefore, one has to check that all gauge anomalies are canceled in chiral theories. This requirement puts a stringent constraint on the matter content of the theory.

Chiral: The word means handedness in Greek. A theory is called chiral if the right-handed and left-handed helicity states of a single particle interact differently with gauge fields at relativistic energies. A chiral theory necessarily breaks parity. The Glashow–Weinberg–Salam theory of electroweak forces is chiral.

Compactification: Suppose the true dimensionality of space-time is not four but is larger, such as ten in superstring theory. One assumes that extra dimensions ($10 - 4 = 6$ in this case) have very small spatial volume so that they are not visible in the current experiments. In this case, the extra dimensions are said to be compactified. The compactification can occur as a solution to the equations of motion of Einstein's equation with matter generalized to higher-dimensional space-time.

Supersymmetry: A hypothesized symmetry that interchanges bosons and fermions, whose spins differ by $\frac{1}{2}$. If supersymmetry exists, every particle in the standard model of particle physics, such as quarks, leptons, gauge bosons, and Higgs bosons, must have its superpartner.

Unified Theories: Theories of elementary particles and forces among them that attempt a unified description of many forces as different manifestations of a single force.

Works Cited

Abachi, S., *et al.* (D0 Collaboration) (1995), *Phys. Rev. Lett.* **74,** 2632.

Abe, F., *et al.* (CDF Collaboration) (1995), *Phys. Rev. Lett.* **74,** 2626.

Appelquist, T., Chodos, A., Freund, P. G. O. (Eds.) (1985), *Modern Kaluza–Klein Theories,* Frontiers in Physics Vol. 65, Reading, MA: Addison-Wesley.

Babu, K. S., Kolda, C., Wilczek, F. (1996), "Experimental Consequences of a Minimal Messenger Model for Supersymmetry Breaking," *Phys. Rev. Lett.* **77,** 3070–3073.

Baer, H., *et al.* (1997), in: T. L. Barklow, S. Dawson, H. E. Haber, J. L. Siegrist (Eds.), *Electroweak Symmetry Breaking and New Physics at the TeV Scale,* Advanced Series on Directions in High Energy Physics Vol. 16, River Edge, NJ: World Scientific.

Barbieri, R., Hall, L. J. (1994), *Phys. Lett.* **338B,** 212.

Bloom, E. D., *et al.* (1969), *Phys. Rev. Lett.* **23,** 930.

Breidenbach, M., *et al.* (1969), *Phys. Rev. Lett.* **23,** 935.

Candelas, P., Horowitz, G. T., Strominger, A., Witten, E. (1985), "Vacuum Configurations for Superstrings," *Nucl. Phys.* **B258,** 46–74.

Carone, C. D., Murayama, H. (1996), *Phys. Rev. D* **53,** 1658.

Castellani, L., D'Auria, R., Fre, P. (1984), "SU(3) × SU(2) × U(1) from $D = 11$ Supergravity," *Nucl. Phys.* **B239,** 610.

Chamseddine, A. H., Arnowitt, R., Nath, P. (1982), "Locally Supersymmetric Grand Unification," *Phys. Rev. Lett.* **49,** 970.

Cremmer, E., Julia, B. (1978), *Phys. Lett.* **80B,** 48.

Cremmer, E., Julia, B. (1979), *Nucl. Phys. B* **159,** 141.

Cremmer, E., Julia, B., Scherk, J. (1978), "Supergravity Theory in Eleven-Dimensions," *Phys. Lett.* **76B,** 409.

de Wit, B. (1979), *Nucl. Phys. B* **158,** 189.

de Wit, B., Freedman, D. Z. (1977), *Nucl. Phys. B* **130,** 105.

de Wit, B., Nicolai, H. (1981), *Phys. Lett.* **108B,** 285.

de Wit, B., Nicolai, H. (1982), *Nucl. Phys. B* **208,** 323.

Dienes, K. R., Faraggi, A. E., March-Russell, J. (1996), "String Unification, Higher Level Gauge Symmetries, and Exotic Hypercharge Normalizations," *Nucl. Phys.* **B467,** 44–99.

Dimopoulos, S., Georgi, H. (1981), *Nucl. Phys. B* **193,** 150.

Dimopoulos, S., Dine, M., Raby, S., Thomas S. (1996), "Experimental Signatures of Low-Energy Gauge Mediated Supersymmetry, *Phys. Rev. Lett.* **76,** 3494–3497.

Dine, M., Nelson, A. E. (1993), *Phys. Rev. D* **48,** 1277.

Dine, M., Nelson, A. E., Shirman, Y. (1995), "Low-Energy Dynamical Symmetry Breaking Simplified," *Phys. Rev. D* **51,** 1362–1370.

Dine, M., Nelson, A. E., Nir, Y., Shirman, Y. (1996), "New Tools for Low-Energy Dynamical Supersymmetry Breaking," *Phys. Rev. D* **53,** 2658–2669.

Fukugita, M., Yanagida, T. (1994), in: M. Fukugita, A. Suzuki (Eds.), *Physics and Astrophysics of Neutrinos,* Tokyo: Springer-Verlag, p. 1.

Gell-Mann, M., Ramond, P., Slansky, R. (1979), "Complex Spinors and Unified Theories," in: P. van Nieuwenhuizen, D. Z. Friedman (Eds.), *Supergravity,* Amsterdam: North-Holland.

Georgi, H., Glashow, S. L. (1974), *Phys. Rev. Lett.* **32,** 438.

Georgi, H., Quinn, H. R., Weinberg, S. (1974), *Phys. Rev. Lett.* **33,** 451.

Glashow, S. L. (1961), *Nucl. Phys.* **22,** 579.

Green, M. B., Schwarz, J. H. (1984), *Phys. Lett.* **149B,** 117.

Green, M. B., Schwarz, J. H. (1985), *Phys. Lett.* **151B,** 21; *Nucl. Phys. B* **255,** 93.

Green, M. B., Schwarz, J. H., Witten, E. (1987), *Superstring Theory,* Vol. 1, *Introduction,* Vol. 2, *Loop Amplitudes, Anomalies and Phenomenology,* Cambridge Monographs on Mathematical Physics, Cambridge, U.K.: Cambridge Univ. Press.

Gross, D. J., Wilczek, F. (1973), *Phys. Rev. Lett.* **30,** 1343.

Gross, D. J., Harvey, J. A., Martinec, E., Rohm, R. (1985), "The Heterotic String," *Phys. Rev. Lett.* **54,** 502–505.

Hall, L., Lykken, J., Weinberg, S. (1983), "Supergravity as the Messenger of Supersymmetry Breaking," *Phys. Rev. D* **27,** 2359–2378.

Hasert F. J., *et al.* (1973), *Phys. Lett.* **46B,** 138.

Hinchliffe, I., Paige, F. E., Shapiro, M. D., Soderqvist, J., Yao, W. (1997), "Precision SUSY Measurements at LHC," *Phys. Rev. D* **55,** 5520–5540.

Hořava, P., Witten, E. (1996), *Nucl. Phys. B* **460,** 506; **475,** 94.

Howe, P., Lindstrom, U. (1981), *Nucl. Phys. B* **181,** 487.

Hull, C. M. (1984), *Phys. Rev.* **D30,** 760; *Phys. Lett.* **142B,** 39.

Hull, C. M., Townsend, P. K. (1995), *Nucl. Phys. B* **438,** 109.

Ignatiev, A. Yu., *et al.* (1978), *Phys. Lett.* **76B,** 436.

Intriligator, K., Seiberg, N. (1996), *Nucl. Phys. Proc. Suppl.* **45BC,** 1.

Kaluza, Th. (1921), *Sitzungsber. Preuss. Akad. Wiss. Berlin, Math. Phys. Kl,* 966.

Kinoshita, T. (1995), in: H. Newman, T. Ypsilantis (Eds.), *History of Original Ideas and Basic Discoveries in Particle Physics,* New York: Plenum Press.

Klein, O. (1926), *Z. Phys.* **37,** 895.

Murayama, H. (1995), in: S. Yamada, T. Ishii (Eds.), *22nd INS International Symposium on Physics with High Energy Colliders,* Singapore: World Scientific.

Murayama, H. (1997), in: Z. Ajduk, A. K. Wroblewski (Eds.), *28th International Conference on High-energy Physics (ICHEP 96),* Singapore: World Scientific.

Particle Data Group (1996), *Phys. Rev. D* **54,** 1.

Pati, J. C., Salam, A. (1973), *Phys. Rev. D* **8,** 1240.

Politzer, H. D. (1973), *Phys. Rev. Lett.* **30,** 1346.

Ross, G. G. (1984), *Grand Unified Theories,* Menlo Park, CA: Benjamin/Cummings.

Sakai, N. (1981), *Z. Phys.* **C11,** 153.

Salam, A. (1968), in: N. Svartholm (Ed.), *Elementary Particle Theory, Proceedings of the Nobel Symposium,* Sweden, Stockholm.

Scherk, J. (1975), *Rev. Mod. Phys.* **47,** 123.

Scherk, J., Schwarz, J. H. (1974), *Nucl. Phys. B* **81,** 118.

Tsukamoto, T., Fujii, K., Murayama, H., Yamaguchi, M., Okada, Y. (1995), "Precision Study of Supersymmetry at Future Linear e^+e^- Colliders," *Phys. Rev. D* **51,** 3153–3171.

Weinberg, S. (1967), *Phys. Rev. Lett.* **19,** 1264.

Wess, J., Bagger, J. (1983), *Supersymmetry and Supergravity,* Princeton: Princeton Univ. Press.

Witten, E. (1981), *Nucl. Phys. B* **186,** 412.

Witten, E. (1983), in: R. Jackiw, N. N. Khuri, S. Weinberg, E. Witten (Eds.), *Shelter Island II: Proceedings,* Cambridge, MA: MIT Press.

Witten, E. (1995), IASSNS-HEP-95-63, hep-th/9507121, contributed to *STRINGS 95: Future Perspectives in String Theory,* Los Angeles, CA, 13–18 March 1995.

Yanagida, T. (1979), in: O. Sawada, A. Sugamoto (Eds.), *Proceedings of the Workshop on the Unified Theory and Baryon Number in the Universe,* KEK report, pp. 79–18.

Yoshimura, M. (1978), *Phys. Rev. Lett.* **41,** 281; Erratum (1979) *Phys. Rev. Lett.* **42,** 746.

Further Reading

Appelquist, T., Chodos, A., Freund, P. G. O. (Eds.) (1985), *Modern Kaluza–Klein Theories,* Frontiers in Physics Vol. 65, Reading, MA: Addison-Wesley.

Close, F. (1983), *The Cosmic Onion: Quarks and the Nature of the Universe,* London: Heinemann.

Georgi, H. (1982), *Lie Algebras in Particle Physics: from Isospin to Unified Theories,* Menlo Park, CA: Benjamin/Cummings.

Green, M. B., Schwarz, J. H., Witten, E. (1987), *Superstring Theory,* Vol. 1, *Introduction,* Vol. 2, *Loop Amplitudes, Anomalies and Phenomenology,* Cambridge Monographs on Mathematical Physics, Cambridge, U.K.: Cambridge Univ. Press.

Mohapatra, R. N. (1992), *Unification and Supersymmetry: the Frontiers of Quark-Lepton Physics,* Berlin: Springer-Verlag.

Nilles, H. P. (1984), *Phys. Rep.* **110,** 1.

Ross, G. G. (1984), *Grand Unified Theories,* Menlo Park, CA: Benjamin/Cummings.

2
Units

Bernd R. L. Siebert and Jörn Stenger

2.1 Introduction 9
2.2 The International System of Units (SI) 10
2.2.1 SI Base Units 10
2.2.1.1 The Meter 11
2.2.1.2 The Kilogram 12
2.2.1.3 The Second 12
2.2.1.4 The Ampere 12
2.2.1.5 The Kelvin 13
2.2.1.6 The Mole 13
2.2.1.7 The Candela 13
2.2.2 SI Derived Units 14
2.2.3 SI Prefixes 14
2.2.4 SI on the BIPM-homepage 14
2.3 Units Outside the SI 16
2.3.1 Non-SI Units Accepted for Use with the SI 16
2.3.2 Other Non-SI Units Not Recommended for Use 17
2.4 Rules for Expressing Quantities 17
2.5 International Metrology 23
2.5.1 Convention du Mètre 23
2.5.2 Mutual Recognition Arrangement 23
2.6 Further Development of the SI 26
Acknowledgment 27
Appendix: Conversion Factors for Some Non-SI Units 27
Glossary 27
References 31
Further Reading 31

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

2.1
Introduction

The use of units is omnipresent in our daily life. Everywhere in trade, manufacturing, and science, measurements are quantified or compared with each other by the use of units.

Already James Clerk Maxwell (1873) stated that every expression of a quantity consists of two factors or components. In today's notation, this is expressed by the relation

$$Q = \{Q\} \times [Q] \tag{2.1}$$

where $\{Q\}$ denotes the numerical value and $[Q]$ is the unit of a physical quantity. The unit is a reference magnitude.

In other words, any quantitative measurement can be imagined as a comparison of a property of an object such as its mass, length, or temperature, with the same property of an object that represents the unit.

To ensure the correctness of this reference, it must be compared with a "higher order" reference. This is called *calibration*. Primary standards are the highest order references. They cannot be calibrated. Comparisons of different primary standards for the same unit, in many cases operated in different countries, are used to achieve highest confidence in the comparability of measurements.

When a measurement is referenced to a primary standard, directly or via an unbroken chain of references, it is called *traceable*. The trustworthiness of a measurement needs to be quantified by stating the uncertainty associated with measured value.

The choice of a unit is, in principle, quite arbitrary. Historically, units have been represented by artifacts. The magnitude of the quantity had been chosen for the sake of convenience or, in case of length, had been set by the length of cells or feet of pharaohs, kings or in later times by the average obtained from a group of people. This led to a variety of units for one and the same quantity that made comparison of measurement results across borders difficult and acted as a barrier for trade and later also for science.

The Convention du Mètre, signed by 17 states in Paris in 1875 was the first international agreement on metrology with the aim to use the same units worldwide. This furthers international trade and cooperation in science and is the basis for international standardization. Presently, there is a multitude of quantities that needs to be measured in the context of trade, manufacturing, the protection of life and environment, and science. Fortunately, virtually all of these quantities can be expressed as a product of powers of few basic quantities, forming a "system of base quantities." The units used to express

Handbook of Metrology. Edited by Michael Gläser and Manfred Kochsiek
Copyright © 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40666-1

numerical values of base quantities are then termed *base units*.

The choice of the base quantities is to a certain degree arbitrary. The present worldwide convention is the International System of Quantities (ISQ). It is based on the seven base quantities of length, mass, time, electric current, thermodynamic temperature, amount of substance, and luminous intensity (BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML, 2008). The symbols for the dimensions of the base quantities are **L** for length, **M** for mass, **T** for time, **I** for electric current, **Θ** for thermodynamic temperature, **N** for amount of substance, and **J** for luminous intensity. Thus, the dimension of a quantity Q is denoted by

$$\mathrm{Dim}Q = \mathbf{L}^{\alpha}\mathbf{M}^{\beta}\mathbf{T}^{\gamma}\mathbf{I}^{\delta}\boldsymbol{\Theta}^{\varepsilon}\mathbf{N}^{\zeta}\mathbf{J}^{\eta} \qquad (2.2)$$

where the exponents, named *dimensional exponents*, are positive, negative, or zero.

Requirements that must be met in order to make the system acceptable for all users are as follows:

1. Consistency with the laws of physics: a measurement of pressure, for example, through a simultaneous measurement of force and area must not lead to different results when obtained by comparison with a realization of the unit of pressure.
2. Flexibility for modifications due to new needs, while being reliable and stable.
3. Practicability in all intended uses and simplicity.

The International System of Units (SI) represents such a system (BIPM, 2006). It is described in the following sections.

2.2
The International System of Units (SI)

In 1948, the Ninth Conférence Générale des Poids et Mesures (GCPM) adopted a draft for a system of units, which was originally based on six and meanwhile on seven base units. All other units are related with these base units just through multiplication and division. In 1960, the GCPM named the new system as *Système International d'Unités*, abbreviated as SI, and it decided that this abbreviation was to be used in all languages. All member states of the Meter Convention have adopted the metric system and have taken steps to eliminate most uses of traditional measurements.

The Bureau International des Poids et Mesures (BIPM) was entrusted to ensure worldwide uniformity of measurements and their traceability to the SI and publish a Brochure about the SI. The latest edition of the Brochure was published in 2006.

2.2.1
SI Base Units

The base units of the International System are listed in Table 2.1, which relates the base quantity to the unit name and unit symbol for each of the seven base units.

Sections 2.2.1.1–2.2.1.7 present the presently adopted definitions in italics followed by additional clarifications provided in the SI-brochure. Only brief remarks on history and realization are given as details can be found on the BIPM homepage (see Section 2.4).

For practical and economic reasons, the units are in many cases realized by reference standards optimized to the practical needs. In order to harmonize these realizations of units, the Comité international des poids et mesures (CIPM) drew up recommendations, which are generally referred to as the *mise en pratique* for the definition. The first *mise en pratique* was given in 1983 at the time of the adoption of the present definition of the

Tab. 2.1 SI base units.

Base quantity		SI base unit	
Name	Symbol	Name	Symbol
length	$l, x, r, \ldots$	metre	m
mass	m	kilogram	kg
time, duration	t	second	s
electric current	I, i	ampere	A
thermodynamic temperature	T	kelvin	K
amount of substance	n	mole	mol
luminous intensity	I_v	candela	cd

meter. In 1992, the CIPM acting on the advice of its Consultative Committee for the Definition of the Metre (CCDM), approved a revised *mise en pratique* for the definition of the meter (Quinn, 1994). Thereafter, it was decided to publish *mises en pratiques* on the homepage of the BIPM. The page is regularly updated if new significant progress has been achieved. Presently, *mises en pratiques* are available for the definition of the meter and the definition of the kelvin; see *http://www.bipm.org/en/publications.*

2.2.1.1 **The Meter**

The meter, unit of length, was redefined in 1983 by the 17th Conférence Générale des Poids et Pesures (CGPM) that specified the current definition, as follows: *The metre is the length of the path travelled by light in vacuum during a time interval of* 1/299 792 458 *of a second*.

It follows that the speed of light in vacuum is exactly 299 792 458 metres per second, $c_0 = 299\,792\,458$ m s.[1]

The first definition of the meter by the first CGPM in 1889 was based on the international prototype of platinum–iridium. This prototype is still kept at the BIPM under the conditions specified in 1889. The 11th CGPM replaced this first definition in 1960 and selected the wavelength of a well-defined transition in krypton 86 as a basis for defining the meter as a multiple of this wavelength. This was an important step toward defining units via fundamental constants.

The definition of the meter allows to use either the relation $l = c_0 t$, where l is the length of the path traveled by a plane electromagnetic wave in vacuum and t is the measured time of flight, or the relation $\lambda = c_0/f$, where λ is the wavelength and f the frequency of electromagnetic radiation. Both relations reflect that the unit of length depends on the unit of time.

The first of these relations is well suited for large distances, whereas the second relation is practical for small distances. In practice, laser radiation is used for realization of the unit of length.

For example, the distance between the earth and the moon is measured by directing a short pulse of a high-power laser beam at the moon where it is reflected from a mirror. The distance can be calculated from the time which the laser pulse needs to travel from the earth to the moon and back to the earth. The Global Positioning System (GPS), too, is based on time-of-flight measurements of electromagnetic radiation. The relative uncertainties associated with measured values of length using $l = c_0 t$ are determined by the uncertainty of the associated time measurement.

---

[1] This and the following definitions are cited from the 8th edition of the SI Brochure, BIPM 2006. With regard to the spelling of English-language words beside the quotations, we refer to the *United States Government Printing office Style Manual*, Which follows *Webster's Third New International Dictionary*. Form example, "meter" "liter," and "deka" are used instead of "metre," "litre," and "deca"

The use of the relation $\lambda = c_0/f$ requires the establishment of frequency standards representing reciprocal time realizations.

2.2.1.2 **The Kilogram**

The kilogram, unit of mass, is the only unit that is still explicitly based on an artifact, the international prototype of the kilogram that is made of platinum–iridium and kept at the BIPM since it was sanctioned by the first CGPM in 1889. The third CGPM confirmed this in 1903 by the definition: *The kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram.*

It follows that the mass of the international prototype of the kilogram is always 1 kilogram exactly, $m(\mathcal{K}) = 1\,\text{kg}$.

The masses of 1 kg secondary standards of the same alloy as the international prototype are compared in air with the mass of the international prototype by means of balances with a relative uncertainty approaching 1 part in 10^9.

2.2.1.3 **The Second**

The second, unit of time, was redefined in 1967/1968 by the 13th CGPM that specified the current definition, as follows: *The second is the duration of* 9 192 631 770 *periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium* 133 *atom.*

It follows that the hyperfine splitting in the ground state of the cesium 133 atom is exactly 9 192 631 770 hertz, $\nu(\text{hfs Cs}) = 9\ 192\ 631\ 770\,\text{Hz}$.

As affirmed by the CIPM at its 1997 meeting, this definition refers to a cesium atom at rest at a temperature of 0 K.

This note clarifies that the definition of the SI second is based on a cesium atom unperturbed by black body radiation, that is, in an environment whose thermodynamic temperature is 0 K. The frequencies of all primary frequency standards should therefore be corrected for the shift due to ambient radiation, as stated at the meeting of the Consultative Committee for Time and Frequency (CCTF) in 1999.

Presently, the best primary standard produced the SI second with a relative standard uncertainty of some parts in 10^{16}.

2.2.1.4 **The Ampere**

The ampere, unit of electric current, was proposed by the CIPM in 1946 and, after replacing the expression "MKS unit of force" by "newton," adopted by the ninth CGPM in 1948: *The ampere is that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross-section, and placed* 1 metre *apart in vacuum, would produce between these conductors a force equal to* 2×10^{-7} newton per metre *of length.*

It follows that the magnetic constant, μ_0, also known as the *permeability of free space*, is exactly $4\pi \times 10^{-7}$ henries per metre, $\mu_0 = 4\pi \times 10^{-7}\,\text{H/m}$.

Note that this definition implies the value of the electric constant, ε_0 (permittivity of vacuum), is also fixed, since $\varepsilon_0 = (\mu_0 c_0^2)^{-1}$.

2.2.1.5 **The Kelvin**

The kelvin, unit of thermodynamic temperature, was adopted in its present form by the 13th CGPM in 1968:

The kelvin, unit of thermodynamic temperature, is the fraction 1/273.16 *of the thermodynamic temperature of the triple point of water.*

It follows that the thermodynamic temperature of the triple point of water is exactly 273.16 kelvins, $T_{\text{tpw}} = 273.16$ K.

As affirmed by the CIPM at its 2005 meeting, this definition refers to water having the isotopic composition defined exactly by the following amount of substance ratios: 0.00015576 mol of ^{2}H per mol of ^{1}H, 0.000 379 9 mol of ^{17}O per mol of ^{16}O, and 0.002 005 2 mol of ^{18}O per mol of ^{16}O.

Already the 10th CGPM selected the triple point of water in 1954 as the fundamental fixed point and assigned to it the temperature of 273.16 K, so defining the unit.

Because of the manner in which temperature scales used to be defined, it remains common practice to express a thermodynamic temperature, symbol T, in terms of its difference from the reference temperature $T_0 = 273.15$ K, the ice point. This difference is called the *Celsius temperature*, symbol t, which is defined by the quantity equation: $t = T - T_0$.

The unit of Celsius temperature is the degree Celsius, symbol °C, which is by definition equal in magnitude to the kelvin. A difference or interval of temperature may be expressed in kelvin or degree Celsius, the numerical value of the temperature difference being the same.

The kelvin and the degree Celsius are also units of the International Temperature Scale of 1990 (ITS-90).

The *mise en pratique* for the definition of the kelvin (last update in 2006) is available on *http://www.bipm.org/en/publications/*.

2.2.1.6 **The Mole**

The mole, the unit for the amount of substance was adopted in its present form by the 14th CGPM in 1971:

1. *The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in* 0.012 kg *of carbon* 12; *its symbol is "mol."*
2. *When the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, other particles, or specified groups of such particles.*

It follows that the molar mass of carbon 12 is exactly 12 grams per mole, $M(^{12}\text{C}) = 12$ g/mol.

This definition refers to unbound atoms of ^{12}C at rest and in their ground state.

The definition of the mole also determines the value of the universal constant that relates the number of entities to amount of substance for any sample. This constant is called the *Avogadro constant*, symbol N_A or L. If $N(X)$ denotes the number of entities X in a specified sample, and if $n(X)$ denotes the amount of substance of entities X in the same sample, the relation is $n(X) = N(X)/N_A$.

Note that since $N(X)$ is dimensionless, and $n(X)$ has the SI unit mole, the Avogadro constant has the coherent SI unit reciprocal mole.

2.2.1.7 **The Candela**

The candela unit of luminous intensity was adopted in its present form by the 16th CGPM in 1979.

The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} hertz *and that has a radiant intensity in that direction of* 1/683 Watt per steradian.

It follows that the spectral luminous efficacy for monochromatic radiation of frequency of 540×10^{12} hertz is exactly 683 lumens per Watt, $K = 683\ \text{lm/W} = 683\ cdsr/W$.

The units of luminous intensity based on flame or incandescent filament standards in use in various countries before 1948 were replaced initially by the "new candle" based on the luminance of a Planck radiator (a black body) at the temperature of freezing platinum. This modification was then ratified in 1948 by the ninth CGPM, which adopted a new international name for this unit, the *candela*. In 1979, the present definition was given because of the difficulties in realizing a Planck radiator at high temperatures, and the new possibilities offered by radiometry, that is, the measurement of optical radiation power.

2.2.2
SI Derived Units

The unit of any quantity can be expressed by the relation

$$[Q] = \mathrm{m}^{\alpha}\ \mathrm{kg}^{\beta}\ \mathrm{s}^{\gamma}\ \mathrm{A}^{\delta}\ \mathrm{K}^{\varepsilon}\ \ \mathrm{mol}^{\zeta}\ \mathrm{cd}^{\eta}.$$

Units obtained by this relation are coherent derived SI units. The base and coherent derived units of the SI form a coherent set, designated the set of *coherent SI units*.

For convenience, certain coherent derived units have been given special names and symbols. There are 22 such units (see Table 2.2).

These units and symbols can be used in forming other coherent derived units. As an example, the unit of dynamic viscosity is expressed in terms of SI base units $\mathrm{m}^{-1}\,\mathrm{kg}\,\mathrm{s}^{-1}$, it can also be expressed in terms of coherent derived SI units pascal and the base unit second, it has then the unit symbol Pa. This convention allows to express the physical meaning of a given quantity.

2.2.3
SI Prefixes

The CGPM adopted, in 1960, a series of prefix names and prefix symbols to form the names and symbols of the decimal multiples and submultiples of SI units, ranging from 10^{12} to 10^{-12}. This range was increased in several steps. The presently approved prefix names and prefix symbols range from 10^{24} to 10^{-24}, and they are listed in Table 2.3. The prefixes are used to keep the numbers giving the values of a quantity in a practical order of magnitude. Names and symbols for decimal multiples and submultiples of the unit are generally formed by attaching prefix names to the unit name. However, one important exception is the base unit of mass, the kilogram, whose name and symbol, for historical reasons, already includes a prefix. For the unit of mass, decimal multiples and submultiples are formed by attaching a prefix to the gram, *g*.

The following points are noted:

- Units with a prefix are not coherent.
- SI prefixes refer strictly to powers of 10, and cannot be used for powers of 2. For example, 1 kilobit should not be used to represent 1024 bits (2^{10} bits), which instead is 1 kibibit (BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML, 2008). When SI prefixes are used, attention should be paid to the rules listed in Section 2.4.

2.2.4
SI on the BIPM-homepage

The BIPM issues the SI-brochure, which can be downloaded from its homepage *http://www.bipm.org*. The official version is the French text. The site allows both, online browsing and downloading of the entire brochure.

Tab. 2.2 Coherent derived units in the SI with special names and symbols.

		SI coherent derived unit		
Derived quantity	**Name**	**Symbol**	**Expressed in SI units**	**Expressed in SI base units**
plane angle	radian[a]	rad	1[a]	m/m
solid angle	steradian[a]	sr[a]	1[a]	m^2/m^2
frequency	hertz[b]	Hz	–	s^{-1}
force	newton	N	–	$m\ kg\ s^{-2}$
pressure, stress	pascal	Pa	N/m^2	$m^{-1}\ kg\ s^{-2}$
energy, work, amount of heat	joule	J	N m	$m^2\ kg\ s^{-2}$
power, radiant flux	watt	W	J/s	$m^2\ kg\ s^{-3}$
electric charge, amount of electricity	coulomb	C	–	s A
electric potential difference, electromotive force	volt	V	W/A	$m^2\ kg\ s^{-3}\ A^{-1}$
capacitance	farad	F	C/V	$m^{-2}\ kg^{-1}\ s^4\ A^2$
electric resistance	ohm	Ω	V/A	$m^2\ kg\ s^{-3}\ A^{-2}$
electric conductance	siemens	S	A/V	$m^{-2}\ kg^{-1}\ s^3\ A^2$
magnetic flux	weber	Wb	V s	$m^2\ kg\ s^{-2}\ A^{-1}$
magnetic flux density	tesla	T	Wb/m^2	$kg\ s^{-2}\ A^{-1}$
inductance	henry	H	Wb/A	$m^2\ kg\ s^{-2}\ A^{-2}$
Celsius temperature	degree Celsius	°C		K
luminous flux	lumen	lm	cd sr[a]	cd
illuminance	lux	lx	lm/m^2	$m^{-2}\ cd$
activity referred to a radionuclide(f)	becquerel[b]	Bq		s^{-1}
absorbed dose[d], specific energy (imparted), kerma	gray	Gy	J/kg	$m^2\ s^{-2}$
dose equivalent[d], ambient dose equivalent, directional dose equivalent, personal dose equivalent	sievert	Sv	J/kg	$m^2\ s^{-2}$
catalytic activity	katal	kat	–	$s^{-1}\ mol$

[a] The *radian* and *steradian* are special names for the number one that may be used to convey information about the quantity concerned. In practice, the symbols rad and sr are used where appropriate, but the symbol for the derived unit one is generally omitted in specifying the values of dimensionless quantities. However, in photometry, the name steradian and the symbol sr are usually retained in expressions for units.

[b] The *hertz* is used only for periodic phenomena, and the *becquerel* is used only for stochastic processes in activity referred to a radionuclide.

[c] Activity of a radionuclide is sometimes incorrectly called *radioactivity*.

[d] For details on quantities and units for ionizing radiation, see ICRU 1998a, 1998b.

Tab. 2.3 SI prefixes.

Factor	Prefix	Symbol	Factor	Prefix	Symbol
10^1	deca	da	10^{-1}	deci	d
10^2	hecto	h	10^{-2}	centi	c
10^3	kilo	k	10^{-3}	milli	m
10^6	mega	M	10^{-6}	micro	μ
10^9	giga	G	10^{-9}	nano	n
10^{12}	tera	T	10^{-12}	pico	p
10^{15}	peta	P	10^{-15}	femto	f
10^{18}	exa	E	10^{-18}	atto	a
10^{21}	zetta	Z	10^{-21}	zepto	z
10^{24}	yotta	Y	10^{-24}	yocto	y

Note that Appendix 2 of the brochure, titled "Practical realizations of the definitions of some important units" is published in electronic form only, as this allows updating in the course of advances in experimental realizations of units. This decision underlines the strict distinction between the definition and the realization of a unit.

Furthermore, this webpage features links to a "Brief history of the SI," a concise visualization of "Linking SI base units to fundamental and atomic constants" and to "Unit conversions" (see also Appendix).

2.3
Units Outside the SI

The consequent use of SI units and the few non-SI units that are accepted for use with the SI promotes communication in pure and applied science.

However, in nearly all countries traditional units are still in use, for example, units for time, such as minute, hour, and day, and there are also some fields of science that use non-SI units. The exclusive use of SI units cannot always be enforced; rather it is a process that is accelerated by evidencing its advantages. Along this line, the policy of the Consultative Committee for Units is to advocate the use of the SI instead of deprecating units that are still widely used. The eighth edition of the brochure promotes this process by providing listings of units accepted for use with the SI and tables for converting some units that are still widely used but not accepted to SI units and by referring to complete conversion table on the BIPM homepage.

2.3.1
Non-SI Units Accepted for Use with the SI

It is distinguished between
- non-SI units accepted for use with the SI (Table 2.4(a) and notes below);
- other non-SI units (Table 2.4(b));
- non-SI units whose values in SI units must be obtained experimentally (Table 2.5); and
- non-SI units associated with the CGS and the CGS-Gaussian system of units (Table 2.6).

Notes on Table 2.4(a):
- ISO 31 recommends that the *degree* (angle) be divided decimally rather than using the minute and the second. For navigation and surveying, however, the minute has the advantage that 1 minute of latitude on the surface of the Earth

Tab. 2.4a Non-SI units accepted for use with the International System of Units.

Quantity	Name of unit	Symbol for unit	Value in SI units
time	minute	min	$1\,\text{min} = 60\,\text{s}$
	hour	h	$1\,\text{h} = 60\,\text{min} = 3600\,\text{s}$
	day	d	$1\,\text{d} = 24\,\text{h} = 86\,400\,\text{s}$
plane angle	degree	$^\circ$	$1^\circ = (\pi/180)\,\text{rad}$
	minute	$'$	$1' = (1/60)^\circ = (\pi/10\,800)\,\text{rad}$
	second	$''$	$1'' = (1/60)' = (\pi/648\,000)\,\text{rad}$
area	hectare	ha	$1\,\text{ha} = 1\,\text{hm}^2 = 10^4\,\text{m}^2$
volume	liter	L, l	$1\,\text{L} = 1\,\text{l} = 1\,\text{dm}^3 = 10^3\,\text{cm}^3 = 10^{-3}\,\text{m}^3$
mass	tonne[a]	t	$1\,\text{t} = 10^3\,\text{kg}$

[a] In English-speaking countries, the unit tonne is usually called *metric ton*.

corresponds (approximately) to one nautical mile.

- The *gon* (or grad, where grad is an alternative name for the gon) is an alternative unit of plane angle to the degree, defined as $(\pi/200)$ rad. Thus there are 100 gons in a right angle. The potential value of the gon in navigation is that because the distance from the pole to the equator of the Earth is approximately 10 000 km, 1 km on the surface of the Earth subtends an angle of one centigon at the center of the Earth. However, the gon is rarely used.
- For applications in *astronomy*, small angles are measured in *arcseconds* (i.e., seconds of plane angle), denoted as milliarcseconds, microarcseconds, and picoarcseconds, denoted mas, μas, and pas, respectively, where arcsecond is an alternative name for second of plane angle.
- The unit *hectare*, and its symbol ha, were adopted by the CIPM in 1879. The hectare is used to express land area.
- The *liter*, and the symbol lowercase l, were adopted by the CIPM in 1879. The alternative symbol, capital L, was adopted by the 16th CGPM (1979, Resolutions 6; CR, 101) in order to avoid the risk of confusion between the letter l (el) and the numeral 1 (one).

2.3.2
Other Non-SI Units Not Recommended for Use

There are many more non-SI units. Some are of historical interest and others are still used for specific applications such as the barrel of oil, or used in particular countries, for example, inch, gallon, and pound.

Although the advantages of a consequent use of SI units appear obvious, it is expected that some of these non-SI units will be used in future. The CIPM has therefore decided to compile a list of the conversion factors to the SI for such units and to make this available on the BIPM home page.

For the convenience of the reader, some more often needed conversion factors are provided in the Appendix.

2.4
Rules for Expressing Quantities

The compliance with rules and conventions enhances the readability of technical and scientific papers.

Tab. 2.4b Other non-SI units.

Quantity	Name of unit	Symbol for unit	Value in SI units
pressure[a]	bar	bar	1 bar = 0.1 MPa = 100 kPa = 10^5 Pa
	millimeter of mercury	mmHg	1 mmHg = 133.322 Pa
length	ångström[b]	Å	1 Å = 0.1 nm = 100 pm = 10^{-10} m
distance	nautical mile[c]	M	1 M = 1852 m
area	barn[d]	b	1 b = 100 fm^2 = 100 × $(10^{-15}\ \text{m})^2$ = 10^{-28} m^2
speed	knot[c]	kn	1 kn = (1852/3600) m/s
logarithmic ratio quantities	neper[e,g]	Np	[see footnote[h] regarding the numerical value of the neper, the bel, and the decibel]
	bel[f,g]	B	
	decibel[f,g]	dB	

[a] Since 1982 1 bar has been used as the standard pressure for tabulating all thermodynamic data. It replaced the *standard atmosphere*, equal to 1.013 25 bar, or 101 325 Pa. The *millimeter of mercury* is a legal unit for the measurement of blood pressure in some countries.

[b] The *ångström* is widely used by x-ray crystallographers and structural chemists because all chemical bonds lie in the range 1–3 ångströms. However, it has no official sanction from the CIPM or the CGPM.

[c] The *nautical mile* is employed for marine and aerial navigation to express distance. As yet there is no internationally agreed symbol, but the symbols M, NM, Nm, and nmi are all used. One nautical mile on the surface of the Earth subtends approximately 1 minute of angle at the center of the Earth. The *knot* is defined as one nautical mile per hour. There is no internationally agreed symbol, but the symbol kn is commonly used.

[d] The *barn* is a unit of area employed to express cross sections in nuclear physics.

[e] The statement $L_A = n$ Np (where n is a number) is interpreted to mean that $\ln(A_2/A_1) = n$. Thus when $L_A = 1$ Np, $A_2/A_1 = \mathrm{e}$. The symbol A is used here to denote the amplitude of a sinusoidal signal, and L_A is then called the *neperian logarithmic amplitude ratio*, or the *neperian amplitude level difference*.

[f] The statement $L_X = m$ dB $= (m/10)$ B (where m is a number) is interpreted to mean that $\lg(X/X_0) = m/10$. Thus when $L_X = 1$ B, $X/X_0 = 10$, and when $L_X = 1$ dB, $X/X_0 = 10^{1/10}$. If X denotes a mean square signal or power-like quantity, L_X is called a *power level* referred to X_0.

[g] In using these units it is important that the nature of the quantity be specified, and that any reference value used be specified. These units are not SI units, but they have been accepted by the CIPM for use with the SI.

[h] The numerical values of the neper, bel, and decibel (and hence the relation of the bel and the decibel to the neper) are rarely required. They depend on the way in which the logarithmic quantities are defined.

Tab. 2.5 Non-SI units whose values in SI units must be obtained experimentally.

Quantity	Name of unit	Symbol for unit	Value in SI units[a]
Units accepted for use with the SI			
energy	electronvolt[b]	eV	1 eV = 1.602 176 53 (14) × 10^{-19} J
mass	dalton[c]	Da	1 Da = 1.660 538 86 (28) × 10^{-27} kg
	unified atomic mass unit	u	1 u = 1.660 538 86 (28) × 10^{-27} kg
length	astronomical unit[d]	ua	1 ua = 1.495 978 706 91 (6) × 10^{11} m
Natural units (n.u.)			
speed	n.u. of speed (speed of light in vacuum)	c_0	299 792 458 m s^{-1}(exact)
action	n.u. of action (reduced Planck constant)	$\hbar$	1.054 571 68 (18) × 10^{-34} J s
mass	n.u. of mass (electron mass)	m_e	9.109 382 6 (16) × 10^{-31} kg
time	n.u. of time	$\hbar/(m_e c_0^2)$	1.288 088 667 7 (86) × 10^{-21} s
Atomic units (a.u.)			
charge	a.u. of charge, (elementary charge)	e	1.602 176 53 (14) × 10^{-19} C
mass	a.u. of mass, (electron mass)	m_e	9.109 382 6 (16) × 10^{-31} kg
action	a.u. of action, (reduced Planck constant)	$\hbar$	1.054 571 68 (18) × 10^{-34} J s
length	a.u. of length, bohr (Bohr radius)	a_0	0.529 177 210 8 (18) × 10^{-10} m
energy	a.u. of energy, hartree (Hartree energy)	$E_\hbar$	4.359 744 17 (75) × 10^{-18} J
time	a.u. of time	$\hbar/E_h$	2.418 884 326 505 (16) × 10^{-17} s

[a] The values in SI units of all units in this table, except the astronomical unit, are taken (Mohr, P.J. and Taylor, B.N. (2005) *Rev. Mod. Phys.*, 77, 1–107). The standard uncertainty in the last two digits is given in parenthesis (see Section 2.4, below).

[b] The *electronvolt* is the kinetic energy acquired by an electron in passing through a potential difference of 1 V in vacuum. The electronvolt is often combined with the SI prefixes.

[c] The *dalton* and the *unified atomic mass unit* are alternative names (and symbols) for the same unit, equal to 1/12 times the mass of a free carbon 12 atom, at rest and in its ground state. The *dalton* is often used with SI prefixes to express the masses of large molecules or small mass differences of atoms or molecules.

[d] The astronomical unit is approximately equal to the mean Earth–Sun distance. It is the radius of an unperturbed circular Newtonian orbit about the Sun of a particle having infinitesimal mass, moving with a mean motion of 0.017 202 098 95 radians per day (known as the *Gaussian constant*). The value given for the astronomical unit is quoted from the International Earth Rotation and Reference Systems Service (IERS) Conventions 2003 (McCarthy, D.D. and Petit, G. (eds) (2004) *IERS Technical Note* 32, Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt am Main, p. 12). The value of the astronomical unit in meters comes from the Jet Propulsion Laboratory (JPL) ephemerides DE403 (Standish, E.M. (1995) Report of the IAU WGAS Sub-Group on Numerical Standards, in *Highlights of Astronomy* (ed. Appenzeller), Kluwer Academic Publishers, Dordrecht, pp. 180–184).

The rules for quantity calculus are independent from the language used; however, the usage of names for quantities, dimensions, and units in a text does depend on the language used. For the rules discussed below, numbers in brackets refer to the examples given in Table 2.7.

Names for quantities and units are ordinary nouns. Therefore, in English, they are declined and start with a lowercase letter, even units named after person [1]. The plural is formed regularly, for example, "lengths" and "henries" are the plurals of "length" and "henry." However for lux, hertz, and siemens the plural is identical with the singular. The names of multiples or submultiples of a unit are written without a hyphen or a space [2]. Derived units that are products of units are written with a hyphen or a space as "multiplication sign" [3], those representing ratios are written using the word "per" to indicate the division [4]. Modifiers such as "squared" or "cubed" are placed after the unit name. However, for area and volume the words "square" and "cubic" may be used and are then placed before the corresponding unit [5].

Symbols for quantities, dimensions, and units are mathematical entities. The normal rules of algebra apply and one must neither use the plural nor mix symbols and names within one expression [6].

Symbols for quantities are generally single letters of the Latin or Greek alphabets, printed in an *italic* font. Symbols for quantities are *recommendations* [7].

Symbols for units are printed in roman (upright) type regardless of the type used in the surrounding text. Symbols for units are mandatory [8]. Unless they are named after a person, unit symbols begin with lowercase letter [9]. However, for the accepted non-SI unit liter the symbol L or l is allowed; L should be used to avoid that l is interpreted as the Arabic numerical for one [10].

The *symbols* for derived units reflect a product of base units with integer exponents, that are negative, zero, or positive, see Section 2.2. Base units with negative exponents are divisors, those with zero as exponents are omitted and those with positive exponents are multipliers. Multiplication must be indicated a half space () or a half-high centered dot (·), but never by a cross (×) [11] and division is indicated by a horizontal line (—), a solidus (/), or by the negative exponent [12].

SI *prefixes* and the unit again form a unit; therefore, they are also printed in roman (upright) type without a multiplication sign between prefix symbol and unit symbol and their names are also written as one word [13]. They can be raised to a positive or negative power [14]. Compound prefixes, that is, prefixes formed by juxtaposition of two or more SI prefixes are not permitted [15] as prefixes are not stand-alone abbreviations for $10^{\pm n}$, n being an integer. Names of decimal multiples or submultiples of the unit of mass are formed by attaching prefixes to the name "gram" and the symbols are formed by attaching prefixes to the unit g [16]. SI prefixes should only be used in stating results and not in calculations in order to preserve coherence.

When multiplying or dividing the values of quantities, brackets should be used to exclude any ambiguities [17]. The terms *ppm*, *ppb*, or *ppt* should not be used to express relative values in the SI. In most English-speaking countries, 10^6, 10^9, 10^{12}, and 10^{15} are named *million*, *billion trillion*, and *quadrillion*, respectively. In French, German, and other languages, however, they are named *million*, *milliard*, *billion*, and *billiard*. On the other hand, the internationally recognized symbol % (percent) may be used with the SI to

Tab. 2.6 Non-SI units associated with the CGS and the CGS-Gaussian system of units.

Quantity	Name of unit	Symbol for unit	Value in SI units
energy	erg[a]	erg	$1\,\text{erg} = 1\,\text{g cm}^2\,\text{s}^{-2} = 10^{-7}\,\text{J}$
force	dyne[a]	dyn	$1\,\text{dyn} = 1\,\text{g cm s}^{-2} = 10^{-5}\,\text{N}$
dynamic viscosity	poise[a]	P	$1\,\text{P} = 1\,\text{g cm s}^{-1} = 0.1\,\text{Pa s}$
kinematic viscosity	stokes	St	$1\,\text{St} = 1\,\text{cm}^2\,\text{s}^{-1} = 10^{-4}\,\text{m}^2\,\text{s}^{-1}$
luminance	stilb[a]	sb	$1\,\text{sb} = 1\,\text{cd cm}^{-2} = 10^4\,\text{cd m}^{-2}$
illuminance	phot	ph	$1\,\text{ph} = 1\,\text{cd sr cm}^{-2} = 10^4\,\text{lx}$
acceleration	gal[b]	Gal	$1\,\text{Gal} = 1\,\text{cm s}^{-2} = 10^{-2}\,\text{m s}^{-2}$
magnetic flux	maxwell[c]	Mx	$1\,\text{Mx} = 10^{-4}\,\text{cm}^2\,\text{T} = 10^{-8}\,\text{Wb}$
magnetic flux density	gauss[c]	G	$1\,\text{G} = 1\,\text{cm}^{-2}\,\text{Wb} = 10^{-4}\,\text{T}$
magnetic field	œrsted[c]	Oe	$1\,\text{Oe} \mathrel{\hat{=}} (10^3/4\pi)\,\text{A m}^{-1} \cong 7.957\,747 \times 10\,\text{A m}^{-1}$

[a] This unit and its symbol were included in Resolution 7 of the 9th CGPM (1948; CR, 70).

[b] The gal is a special unit of acceleration employed in geodesy and geophysics to express acceleration due to gravity.

[c] These units are part of the so-called electromagnetic three-dimensional CGS system based on unrationalized quantity equations, and must be compared with care to the corresponding unit of the International System, which is based on rationalized equations involving four dimensions and four quantities for electromagnetic theory. The magnetic flux, Φ, and the magnetic flux density, B, are defined by similar equations in the CGS system and the SI, so that the corresponding units can be related as in the table. However, the unrationalized magnetic field, $H(\text{unrationalized}) = 4\pi \times H(\text{rationalized})$. The equivalence symbol $\hat{=}$ is used to indicate that when $H(\text{unrationalized}) = 1$ Oe, $H(\text{rationalized}) = (10^3/4\pi)\,\text{A m}^{-1}$.

represent the number 0.01, and a half space separates number and % [18].

Multiplication or division of several quantities again leads to a quantity; however, addition or subtraction is only physically meaningful for quantities with the same dimension and equivalent units. Equations involving physical quantities must have the same dimensions on both sides, and the dimensions must be the correct ones for the quantity calculated.

For a complete statement of the result of a measurement, it is required

- to indicate the measurand;
- to provide the numerical value and the unit; and
- to state a measure for the trustworthiness of the result.

The measurand should be defined in sufficient detail. For instance, the quantity ''temperature at the triple point of water'' is not well defined, unless the isotopic composition of water is given (cf. Section 2.2.1.5).

Even if in a given field, one and only one unit is used, it must be given explicitly and a prefix should be used, such that the leading digits range from 0.001 to 1000. If a result has many digits, it enhances the legibility to group them in blocks of three, separated by a half space [19]. However, within a column of a table, the same unit should be used. In stating the values of quantities of dimension one, powers of 10 are used.

The most informative measure for the trustworthiness of a numerical value is the expanded uncertainty.

The Guide to the expression of uncertainty in measurement (GUM) ISO

Tab. 2.7 Examples for demonstrating rules.

1	Several kelvins, *not* Kelvins or Kelvin
2	milligram, *neither* milli-gram *nor* milli gram
3	newton meter or newton-meter *not* newtonmeter
4	meter per second, *not* divided by second
5	meter per second squared, *not* square second, *but* kilogram per meter cubed *or* cubic meter
6	Potential difference is measured in volts, *not* in V; its value is 1000 V *not* 1000 volts
7	There are *several* symbols in use for length quantities, for example, l, x or r
8	*Only* the symbol m is allowed for meter
9	Ω is the symbol for the unit ohm, *not* ω lx is the symbol for the unit lux, *not* Lx
10	1 L but *neither* 1 l *nor* 1 l
11	Multiplication: *either* N m *or* N·m, but *neither* Nm *nor* N × m
12	Division: $\frac{\mathrm{m}}{\mathrm{s}}$, m/s, $\mathrm{m\,s^{-1}}$ or $\mathrm{m{\cdot}s^{-1}}$
13	The symbol for millimeter is mm, *neither* the symbol for milli meter is m m *nor* the symbol for milli-meter is m·m the symbol for the unit millisecond is ms, the symbol for the unit of speed is $\mathrm{m\,s^{-1}}$
14	$1\,\mathrm{cm^3} = (10^{-2}\,\mathrm{m})^3 = 10^{-6}\,\mathrm{m^3}$ or $1\mathrm{ns}^{-1} = (10^{-9}\,\mathrm{s})^{-1} = 10^9\,\mathrm{s}^{-1}$
15	10^5 Pa = 0.1 MPa or 100 kPa *but not* 10^5 Pa = 1 hkPa "(hectokilo)"
16	10^{-6} kg = 1 mg *not* 10^{-6} kg = 1 µkg
17	$(0.5\,\mathrm{m\,s^{-2}})(2.0\,\mathrm{s}) = (0.5\,\mathrm{m\,s^{-2}} \times (2.0\,\mathrm{s}) = 1\,\mathrm{m}^{-1})$ in some countries × is used for expressing areas $a = 5\,\mathrm{m} \times 5\,\mathrm{m} = 25\,\mathrm{m}^2$ or volumes $(a/b)/c \neq a(b/c)$, unless $c = 1$; $a/b/c$ is undefined
18	$w\ (x) = 0.032\%$, and $w(x) = 0.032\%$
19	$l = 1000$ mm is appropriate, $l = 1\,000000$ µm is *less* appropriate, $l = 987.654\ 32$ mm *or* $l = 0.987\ 654\ 32$ m, but *neither* $l = 0.000\ 987\ 654\ 32$ km *nor* $l = 987\ 654.32$ µm are easy to read

(1993) denotes the standard uncertainty associated with the best estimate x of the value of a quantity X by $u(x)$. The GUM provides rules for the evaluation of the uncertainty and for the determination of the so-called coverage factor. The product of the coverage factor k and the uncertainty is called *expanded uncertainty* for which the symbol U is used. For example, the following statement would be complete:

The measured value of the nominal 50-mm *gauge block is* (49.999 926 ± 0.000 073) mm. *The reported expanded uncertainty of measurement is stated as the standard uncertainty of measurement multiplied by the coverage factor* $k = 2$, *which corresponds to a coverage probability of approximately* 95% *for a normal distribution.*

In some literature, the uncertainty is often given by adding two digits in parenthesis. For the result given above, it is 49.999 926 (73) mm. Here, too, the used coverage factor is to be stated.

Finally, there are no rules on choosing a unit for a given nonbase quantity. However, if possible, the unit should reflect the physics of the addressed quantity, for example, for torque, one would use newton and not joule.

2.5
International Metrology

2.5.1
Convention du Mètre

International metrology aims at global comparability of measurements. The first international agreement on metrology is the Convention du Mètre, signed by 17 states in Paris in 1875. It is still effective and currently has (2008) 51 signatory states and 27 associated states and economies. This treaty gives authority to the

- CGPM;
- the CIPM; and the
- BIPM.

to act in matters of world metrology, particularly concerning the demand for measurement standards of ever-increasing accuracy, range, and diversity, and the need to demonstrate equivalence between national measurement standards.

Table 2.8 visualizes the structure of international metrology under the authority of the Convention du Mètre.

The CGPM is the highest authority. Representatives of all member states and observers from the associate states meet every four years, last in 2007. The CGPM is responsible for

- discussing and instigating the arrangements required to ensure the propagation and improvement of the SI;
- confirming the results of new fundamental metrological determinations and the various scientific resolutions of international scope, mainly prepared by the CIPM; and
- adopting the important decisions concerning the organization and development of the BIPM.

Principal task of the CIPM is to promote worldwide uniformity in units of measurement and this is done by direct action or by submitting draft resolutions to the CGPM. The CIPM meets annually at the BIPM. It has 18 members from different member states of the Meter Convention and the director of the BIPM is a member *ex officio*.

The BIPM was established by the Convention du Mètre. It operates under the exclusive supervision of the CIPM and has the task to ensure worldwide uniformity of measurements and their traceability to the SI. The BIPM carries out measurement-related research, takes part in and organizes international comparisons of national measurement standards, and carries out calibrations for National Metrology Institutes (NMIs).

In view of the extension of the work entrusted to the BIPM, the CIPM has set up Comités Consultatifs (CC) since 1927. These CCs are responsible for coordinating the international work carried out in their respective fields and they advise on matters of the system of units. At present, 10 CCs exist (see Table 2.8).

The BIPM also forms Joint Committees with other international organizations, for particular tasks of common interest (see Table 2.8).

2.5.2
Mutual Recognition Arrangement

The Comité International des Poids et Mesures Mutual Recognition Arrangement (CIPM-MRA) (CIPM, 1999), is an agreement of NMIs with the following objectives:

- to establish the degree of equivalence of national measurement standards;

Tab. 2.8 International Metrology.

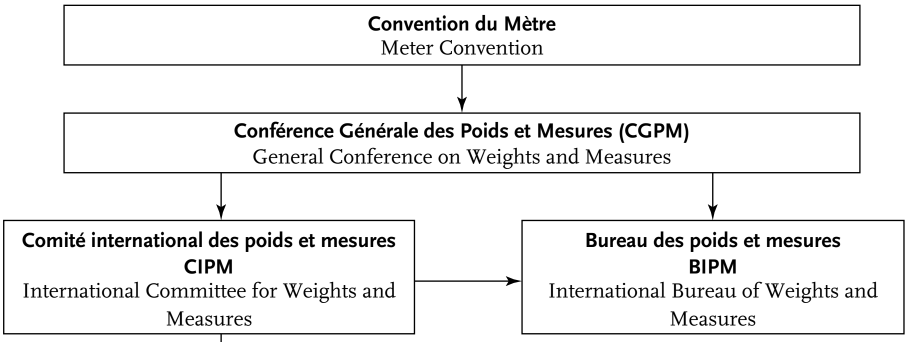
Convention du Mètre
Meter Convention
Conférence Générale des Poids et Mesures (CGPM)
General Conference on Weights and Measures
Comité international des poids et mesures
CIPM
International Committee for Weights and Measures
Bureau des poids et mesures
BIPM
International Bureau of Weights and Measures

Consultative committees			
CCEM	Electricity and Magnetism, 1927 CCE (d'électricité), renamed in 1997.	**CCPR**	Photometry and Radiometry , 1933 CCP (de photométrie), renamed in 1971.
CCT	Thermometry, 1937.	**CCL**	Length, 1952 CCDM (definition of the metre), renamed in 1997.
CCTF	Time and Frequency, 1956 CCDS (definition of the second), renamed in 1997.	**CCRI**	Ionizing Radiation, 1958 as CEMRI (étalons de mesure des rayonnements ionisants), renamed in 1997.
CCU	Units, set up in 1964.	**CCM**	Mass and Related Quantities, 1980.
CCQM	Amount of Substance, 1993, in 2001 the qualifier "metrology in chemistry" was added.	**CCAUV**	Acoustics, Ultrasound and Vibration, 1998.

Joint Committees of the BIPM and other international organizations

JCDCMAS:	Joint committee for the coordination of technical assistance to developing countries in metrology, accreditation and standardization.
JCGM:	Joint committee for guides in metrology. It has two working groups (WG). WG 1 promotes the use of the guide to the expression of uncertainty in measurement (GUM), prepares supplemental guides for its broad application. WG 2 revises and promotes the use of the International vocabulary of metrology (VIM), the acronym stands for it French version "Vocabulaire international de métrologie"
JCRB:	Joint committee of the regional metrology organizations and the BIPM. Its most important tasks is the coordination of the activities among the RMOs in establishing confidence for the recognition of calibration and measurement certificates, according to the terms of the CIPM MRA. It reports annually to the CIPM and to the signatories of the CIPM MRA.
JCTLM:	Joint committee for traceability in laboratory medicine. The goal of the JCTLM is to provide a worldwide platform to promote and give guidance on internationally recognized and accepted equivalence of measurements in laboratory medicine and traceability to appropriate measurement standards.

For more details see http://www.bipm.org/en/committees/

- to provide for the mutual recognition of calibration and measurement certificates; and
- thereby to provide governments and other parties with the metrological basis for wider agreements negotiated for international trade, commerce, and regulatory affairs.

The NMIs, one per country, are signatories of the CIPM-MRA. However, in many countries also, other institutes hold national standards. These institutes can be designated by the signatory NMI and participate in the CIPM-MRA. These institutes are referred to as designated institutes (DIs).

The CIPM-MRA was introduced in the year 1999 at the 21st General Conference on Weights and Measures (CGPM). Presently (2008), it has been signed by the representatives of 74 institutes – from 45 member states, 27 associates of the CGPM, and 2 international organizations – and covers a further 121 DIs.

The objectives of the CIPM-MRA are reached by the following key elements:

- peer review of the declared Calibration and Measurement Capabilities (CMCs);
- international comparisons of measurement standards (key comparisons or supplementary comparisons); and
- peer review of the quality systems.

International comparisons take many forms. Some are carried out by the BIPM; others are organized by a Consultative Committee and may or may not include the BIPM; many more are organized and carried out, usually as part of the activities of a regional metrology organization, by the NMI and other institutes. Most BIPM comparisons are among the key comparisons chosen by the Consultative Committees to establish the *degree of equivalence* between national measurement standards.

The *degree of equivalence* of a measurement standard is a quantified expression of the consistency of the value of a measurement standard with the *key comparison reference value*. This is expressed quantitatively by the deviation from the key comparison reference value and the uncertainty associated with that value. The degree of equivalence between two measurement standards is expressed as the difference between their respective deviations from the key comparison reference value and the uncertainty associated with the value of this difference.

A *key comparison* is a set of comparisons selected by a Consultative Committee to test the principal techniques and methods for measuring quantities they are concerned with. Since only a limited number of participants is technically feasible in a given key comparison, key comparisons are also carried out by Regional Metrology Organisations (RMOs):

- APMP: Asia Pacific Metrology Programme;
- COOMET: Euro–Asian Cooperation of National Metrological Institutions;
- EURAMET: European Association of National Metrology Institutes;
- SADCMet: Southern African Development Community Cooperation in Measurement Traceability;
- SIM: Sistema Interamericano de Metrologia.

The results of key comparisons are accessible from a database maintained by the BIPM. Only key comparisons carried out by a Consultative Committee or the BIPM lead to a key comparison reference value. For a key comparison carried out by a regional metrology organization the link to the key comparison reference value is obtained by reference to the results from those institutes that have also taken part in the CIPM key comparison. For procedures to evaluate key comparisons see, for instance, Cox [2002a, 2002b].

2.6
Further Development of the SI

The SI is, and must be, a stable and reliable system because it affects almost all aspects of human life. However, the specific definitions of the units eventually need to be revised, since the levels of uncertainties, which can be achieved with the primary standards built to realize the units for practical use, depend on the specific definition of the units.

In fact, a quantum-based definition of the volt and ohm through the Josephson and the Quantum-Hall effect enables the realization of these units with smaller uncertainties than those possible through the classical SI definition. This led to a parallel use of these quantum-based units since 1990. In addition, an unsatisfactory relative drift of national kilogram prototypes versus each other and versus the international prototype kept at BIPM, was observed.

These and other arguments motivated the CIPM at its 94th meeting in 2005 to ask the metrology community to take preparative steps toward a redefinition of the kilogram, ampere, kelvin, and mole so that these four base units are linked to fixed values of fundamental constants.

It was advised that a possible redefinition should not introduce a discontinuity in the value of the unit. Furthermore, a significant benefit for one or more units must be achieved. Indeed, even an increased uncertainty in one specific unit may be tolerated, depending on the practical demands, if it is obviously outweighed by a related benefit of reduced uncertainty of another unit or of the system as a whole.

In the following, options that are currently being discussed but have not been decided yet are described. Some papers concerned with that discussion are cited below (see Further Reading) and briefly summarized there. The selection mirrors the main ideas and propositions, but not claimed to be exhaustive.

The possible redefinitions are such that one or more of the following fundamental constants are fixed: the elementary charge e, Avogadro's number N_A, Boltzmann's constant k_B, and Planck's constant h.

For instance, the definition of the ampere could refer to the number of electrons per second carrying the elementary charge. Approaches to realize the ampere according to such a definition are the single-electron transport (SET) using a chain of tunnel contacts, quantum dots, or nanostructured semiconductor channels.

The definition of the mole could refer to the amount of substance that contains N_A elementary entities.

The kelvin could refer to a change of temperature such that the thermodynamic temperature and the thermal energy E expressed in Joule are linked via $E = k_{\mathrm{B}}T$.

For the definition of the kilogram, essentially two options are discussed: the first is the so-called Watt-balance experiment, in which the weight of a test body (which then represents a reference mass) is balanced by an electromagnetic force, thus effectively relating the mass to a measurement of a current and a voltage. If the voltage is measured via Josephson effect with the Josephson constant $K_{\mathrm{J}} = 2e/h$, the current is measured via Quantum-Hall effect with the von Klitzing constant $R_{\mathrm{K}} = h/e$, and applying

Ohm's law, the kilogram would be linked to Planck's constant h. As an alternative, the kilogram could refer to the atomic mass unit m_e or to Avogadro's constant N_A, as realized in the so-called Avogadro experiment.

Any definition of a system of units is based on the definition of a certain set of values of constants or material parameters. In the present SI, exact values are defined for the speed of light in vacuum c_0, the mass of the international prototype of the kilogram $m(K$, the triple-point temperature of water T_{tpw}, and the magnetic constant μ_0. All other constants, if not automatically exact by definition such as the electric constant ε_0 via $\varepsilon_0 = 1/(c_0^2\mu_0)$, are quantities that need to be measured with a particular uncertainty given by the experiment.

The possible redefinitions as outlined before that are based on defining exact values of e, N_A, k_B, h (and c_0), would require to measure the values of $m(\mathcal{K})$, $M(^{12}C)$, T_{tpw}, and μ_0 and nonvanishing uncertainties would inevitably be associated with those values. As one of the most prominent benefits, however, the quantum-based definitions of volt and ohm through Josephson constant K_J and the von Klitzing constant R_K would become part of SI. In addition, the definition of the kelvin would become independent of possible drifts in water triple point cells and the kilogram would become independent of drifts of the mass of the international kilogram prototype.

Some fundamental constants such as e, N_A, k_B, and h are linked by fundamental laws of nature. Thus, one is not free to choose their values. Any redefinition of units that defines an exact value of one or more of these fundamental constants requires the consistency of the current experiments on them. This is not yet satisfactorily the case for the Watt-balance and Avogadro experiments. Values for h can be deduced from both experiments, which differ significantly. This disagreement needs to be solved before the CGPM may adopt redefinitions as outlined earlier.

Acknowledgment

We wish to thank Professor A. J. Wallard, Directeur du Bureau International des Poids et Mesures. With his kind permission, passages and tables from the brochure "Le Systeme International d'Unites (SI)," 8th edition 2006, have been used. We are also grateful to the BIPM for their generous access to publications on its home page.

Appendix: Conversion Factors for Some Non-SI Units

This appendix provides a list of some more frequently used non-SI units (Table 2.9). For a complete list, the reader is referred to *http://www.bipm.org/en/si/si_constants.html* and to Appendix B in the National Institute of Standards and Technology (NIST) Special Publication 811, (Thompson and Taylor, 2008). Bold printing indicates *exactly defined* conversion factors.

Glossary

The glossary lists scientific terms of relevance to units. Main sources are the Guide to the Expression of Uncertainty in Measurement (GUM) (ISO, 1993), the Mutual Recognition Arrangement (MRA) (CIPM, 1999) and the Vocabulaire International de Métrologie (VIM) (JCGM, 2008).

Base Quantity: Quantity in a conventionally chosen subset of a given system of

Tab. 2.9

Unit name (symbol, when defined)	SI equivalent	Notes
Length		
1 inch (in)	**0.0254 m**	12 in = 1 ft
1 foot (ft)	**0.3048 m**	36 in = 3 ft = 1 yd
1 yard (yd)	**0.9144 m**	1760 yd = 1 mi
1 mile	**1.609 344 km**	
1 parsec	$3.085\,678 \times 10^{16}$ m	Used in astronomy
1 light year (l.y.)	$9.460\,73 \times 10^{15}$ m	Based on 1 d = 86 400 s, and 1 Julian century = 36 525 d
Volume		
1 barrel (bbl)	0.158 987 3 m^3	For trading crude oil = 42 US gallons
1 bushel (bu)	$3.523\,907 \times 10^{-2}$ m^3	2150.42 in^3, for trading grain
1 imperial gallon (gal)	**4.546 09 L**	Used in Canada and formally in the United Kingdom
1 gallon (US) (gal)	3.785 412 L	
1 quart (US liquid) (liq qt)	9.463 529 dL	4 liq qt = 1 gal, for example, beer and juice
1 pint (liquid) (US) (liq pt)	4.73 1765 dL	8 liq pt = 1 gal, for example, beer and juice
1 cup (US)	2.365 882 dL	2 cup = 1 liq pt, used in cookbooks
1 tablespoon (tbsp)	1.478 676 cL	16 tbsp = 1 cup, used in cookbooks
1 teaspoon (tsp)	4.928 922 mL	3 tsp = 1 tbsp, used in cookbooks
Velocity or speed		
1 knot (nautical mile per hour)	0.514 444 4 m/s	Navigation and aviation
1 mile per hour (mi/h)	**1.609 344 km/h**^{1}	
1 revolution per minute (rpm) (r min)	0.104 719 8 rad/s	
Mass		
1 ounce (avoirdupois) (oz)	$2.834\,952 \times 10^{-2}$ kg	16 oz = 1 lb
1 pound (avoirdupois) (lb)	**0.453 592 37 kg**	Usually, only 0.453 592 4 is used
Force		
1 kilogram-force (kgf)	**9.806 65 N**	for g_n= **9.806 65 m/s^2** (standard value)
1 poundal	1.383 550	1 poundal =1 lb ft s^{-2}
1 pound-force (lbf)	4.448 222 N	Rounded values for poundal and pound-force

Tab. 2.9 (*continued*)

Unit name (symbol, when defined)	SI equivalent	Notes
Pressure or stress		
1 atmosphere, standard (atm)	**$1.013\,25 \times 10^5$ Pa**	
1 torr (Torr)	$1.333\,224 \times 10^2$ Pa	760 torr = 1 atm
1 inch of mercury, conventional (in Hg)	$3.386\,389 \times 10^3$ Pa	
Energy (includes work)		
1 British thermal unit$_{\mathrm{IT}}$ (Btu$_{\mathrm{IT}}$)	$1.055\,056 \times 10^3$ J	
1 calorie$_{\mathrm{IT}}$ (cal$_{\mathrm{IT}}$)	**4.1868 J**	
Temperature		
Degree Fahrenheit (temperature) (°F) to kelvin (K):		$T(\mathrm{K}) = (t/^{\circ}\mathrm{F} + \mathbf{459.67})/\mathbf{1.8}$
Degree Fahrenheit (temperature) (°F) to degree Celsius (°C):		$t(^{\circ}\mathrm{C}) = (t/^{\circ}\mathrm{F} - \mathbf{32})/\mathbf{1.8}$
Power		
1 horsepower (550 ft-lbf s^{-1}) (hp)	$7.456\,999 \times 10^2$ W	
Radiology		
Curie (Ci)	**3.7×10^{10}Bq**	roughly the activity of 1 g of 226Ra
Röntgen (R)	**2.58×10^{-4}C kg^{-1}**	measure of exposure to X or γ radiation
Rad (rad)	**10^{-2} Gy**	1 rad = 10^2 ergs/g (CGS-unit)
Rem (rem)	**10^{-2} Sv**	absorbed dose × radiation quality
Motor vehicle efficiency		
Is expressed in the US by miles per gallon (mi gal^{-1}) and in metric countries by liter per 100 km. Upon introducing x mi gal^{-1} and y L/(10^2 km), one obtains $y = (2.352\,145 \times 10^2)/x$ or $x = (4.251\,437 \times 10^{-3})/y$.		

quantities, where no subset quantity can be expressed in terms of the others. The following amy be noted:

- The subset mentioned in the definition is termed the *Set of Base Quantities.*
- Base quantities are referred to as *being mutually independent* since a base quantity cannot be expressed as a product of powers of the other base quantities.
- *Number of entities'* can be regarded as a base quantity in any system of quantities.

Base Unit: Measurement unit that is adopted by convention for a base quantity.

Coherent Derived Unit: Unit that is defined as product of the number 1 and powers of the base units with at least two nonvanishing exponents. For example, kg/m is a coherent derived unit for the quantity mass density.

Degree of Equivalence of a Measurement Standard: The degree to which the value of a measurement standard is consistent with the key comparison reference value. This is expressed quantitatively by the deviation from the key comparison reference value and the uncertainty of this deviation. The degree of equivalence between two measurement standards is expressed as the difference between their respective deviations from the key comparison reference value and the uncertainty of this difference.

Derived Unit: Unit that is defined as product of a number and powers of the base units with at least two nonvanishing exponents. For example, g/cm is a derived unit for the quantity mass density, but it is not coherent.

Dimension: See quantity dimension.

International System of Units (SI): System of units, based on the International System of Quantities, their names and symbols, including a series of prefixes and their names and symbols, together with rules for their use, adopted by the General Conference on Weights and Measures (CGPM), see Sections 2.2 and 2.4 and Table 2.8.

Key Comparison: One of the set of comparisons selected by a Consultative Committee to test the principal techniques and methods in the field (note that key comparisons may include comparisons of representations of multiples and submultiples of SI base and derived units and comparisons of artifacts).

Key Comparison Database: The database maintained by the BIPM, which contains Appendices A, B, C, and D of the Mutual Recognition Arrangement (CIPM, 1999).

Key Comparison Reference Value: The reference value accompanied by its uncertainty resulting from a CIPM key comparison (Source: BIPM).

Measurement Unit: Real scalar quantity, defined and adopted by convention, with which any other quantity of the same *kind* can be compared to express the ratio of the two quantities as a number. Also called *unit of measurement or unit*.

Multiple of a Unit: Unit that is formed from a given unit according to a scaling convention and is larger than that unit; see also prefix.

Numerical Value of a Quantity: Ratio of that quantity to the unit used.

Prefix: SI-prefixes (see Table 2.3) are used to form a multiple or submultiples of a unit.

Quantity: Property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference.

Quantity Dimension: Expression of the dependence of a quantity on the base quantities of a system of quantities as a product of powers of factors corresponding to the base quantities, omitting any numerical factor.

Set of Base Quantities: Mutually independent quantities such as, for instance, the International System of Quantities (ISQ).

Submultiples of a Unit: Unit that is formed from a given unit according to a scaling convention and is smaller than that unit; see also prefix.

Symbol of a Unit: Conventional sign designating a unit.

System of Quantities: Set of quantities together with a set of noncontradictory equations relating those quantities.

Uncertainty (of Measurement): Parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand (ISO, 1993).

Unit: See measurement unit.

Value of a Quantity: Number and reference together expressing magnitude of a quantity; see also numerical value of a quantity.

References

BIPM (**2006**) *Le système International d' Unités* (SI), 8th édn, Bureau International des Poids et Mesures, Sèvres.

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML (**2008**) International Vocabulary of Metrology – Basic and General Concepts and Associated Terms(VIM), *http://www.bipm.org/en/publications/guides/vim.html*.

CGPM (**1948**) Resolution 7 of the 9th CGPM.

CIPM (**1999**) Mutual Recognition of National Measurement Standards and of Calibration and Measurement Certificates Issued by National Metrology Institutes, International Committee of Weights and Measures, Issued by the CIPM on 14 October 1999, and revised in October 2003: *http://www.bipm.org/utils/en/pdf/mra_2003.pdf*.

Cox, M.G. (**2002a**) Evaluation of key comparison data. An introduction. *Metrologia*, **39**, 587–588.

Cox, M.G. (**2002b**) Evaluation of key comparison data. *Metrologia*, **39**, 589–595.

ICRU (**1998a**) Conversion Coefficients for use in Radiological Protection Against External Radiation, Report No. 57.

ICRU (**1998b**) Fundamental Quantities and Units for Ionizing Radiation, Report No. 60.

ISO (**1993**) Guide to the Expression of Uncertainty in Measurement, International Standards Organisation, Geneva.

JCGM (**2008**) Evaluation of Measurement Data – Supplement 1 to the "Guide to the Expression of Uncertainty in Measurement" – Propagation of Distributions Using a Monte Carlo Method, *http://www.bipm.org/en/publications/guides/gum.html*.

Mohr, P.J. and Taylor, B.N. (**2005**) CODATA recommended values of the fundamental physical constants. *Rev. Mod. Phys.*, **77**, 1–107.

Quinn, T.J. (**1994**) Mise en Pratique of the Definition of the Metre (1992). *Metrologia*, **30**, 523–541.

Maxwell, J.C. (**1873**) *Treatise on Electricity and Magnetism*, Oxford University Press, London.

Thompson, A. and Taylor, B.N. (**2008**) Guide for the Use of the International System of Units (SI), NIST Special Publication 811 2008 Edition, *http://physics.nist.gov/cuu/pdf/sp811.pdf*.

Further Reading

History and Quantity Calculus

de Boer J. (**1995**) On the history of quantity calculus and the International System. *Metrologia*, **31**, 405–429.

The author discusses the foundation of quantity calculus and its algebraic structure as well as the interrelation between quantity systems and unit systems. Furthermore he analyzes the evolvement of the International System.

Selected Papers on the Envisioned Redefinition of Some SI Base Units

Kose, V., Siebert, B.R.L., and Wöger, W. (**2003**) General principles for the definition of the base units in the SI. *Metrologia*, **40**, 146–153.

The authors formulate general principles that should guide the redefinition of the SI base units and derive requirement from the demands and the need of economy, science and society. To meet this requirements, they propose to principally separate the definition and the realisation of the SI units.

Mills, I.M., Mohr, P.J., Quinn, T.J., Taylor, B.N., and Williams, E.R. (**2006**) Redefinition of the kilogram, ampere, kelvin and mole: a proposed approach to implementing CIPM recommendation 1 (CI-2005). *Metrologia*, **43**, 227–246.

The authors propose that the four base units kilogram, ampere, kelvin and mole should be given new definitions linking them to exactly defined values of the Planck constant h, elementary charge e, Boltzmann constant k and Avogadro constant N_A, respectively. They present the background and discuss the merits of these proposed changes, and also present possible wordings for the four new definitions.

Becker, P., De Bièvre, P., Fujii, K., Glaeser, M., Inglis, B., Luebbig, H., and Mana, G. (**2007**) Considerations on future redefinitions of the kilogram, the mole and of other units. *Metrologia*, **44**, 1–14.

The authors discuss alternatives for redefining the kilogram and the mole. They find good reasons for fixing a number as the ratio between the kilogram and the mass of an atomic particle, and they propose to convert the Avogadro constant to the 'Avogadro number', and to link the mole to this number.

Milton, M.J.T., Williams, J.M., and Bennett, S.J. (**2007**) Modernizing the SI: toward an improved, accessible and enduring system. *Metrologia*, **44**, 356–364.

The authors conclude that, while the proposed changes to the definitions of the mole and kelvin appear to have some advantages, the case for changes in the definitions of the ampere and the kilogram is less well made at the present stage of development of the key experiments involved.

UNIVERSALITY

VLADIMIR PRIVMAN, *Department of Physics, Clarkson University, Potsdam, New York, U.S.A.*

Introduction 31

1. Critical Phenomena at Phase Transitions 32

2. Scaling and Universality in Critical Phenomena 34

3. Renormalization, Two-Scale-Factor Universality, and Finite-Size Scaling 37

4. Crossover Scaling and Related Topics 38

5. Numerical and Analytical Methods and Selected Results 40

6. Polymers, Percolation, Chemical Reactions, Miscellaneous Results 42

Glossary 44

Works Cited 45

Further Reading 46

INTRODUCTION

Universality is a rather general term. Its uses in the sciences have been broad and varying. Strictly speaking, any scientific law implies at least a certain measure of universality. Indeed, in writing scientific relations, equations, and definitions, our aim is to provide description of the world in terms of the laws of nature. But any such description usually involves extrapolation; the laws of physics, in particular, are usually based on theoretical deductions suggested by, and checked through, analysis of experimental data. However, no matter how large the experimental data set is, the use of the equations of physics always assumes that they apply universally to all similar data. Thus, the widest use of universality in physics is in the context of extrapolation.

More "constructive" uses of the concept of universality are encountered in those applications where our knowledge of the underlying physical system is incomplete because of the size, cost, or complexity of the setup required to obtain experimental data sufficient to serve as a foundation for theoretical modeling, or because of difficulty with direct theoretical analysis of a complex system. Perhaps the most familiar approach is dimensional analysis (Schmidt and Housen, 1995) in which the number of independent variables is reduced by forming dimensionless combinations. A more definitive idea is that of *scaling*. In its broad sense, scaling consists of grouping physical variables according to some criteria in an attempt to reduce the number of independent combinations beyond that suggested by the simplest dimensional analysis. The scaling criteria usually refer to spatial or temporal scales (that is the source of the term "scaling") in the system. For instance, variables reflecting system behavior on macroscopic scales might be separated from variables corresponding to variation on the microscopic scales, in forming scaling combinations. Further discussion can be found in the article PHASE TRANSITIONS: RENORMALIZATION AND SCALING.

Universality then augments dimensional analysis and/or scaling in that the same relations are applied to different systems. Presumably, simpler systems, or numerical models appropriate for computer simulations, are used to determine the forms of the functional relations among the appropriate variable combinations. The results are then extended to more complex systems.

An elaborate use of universality and scaling ideas has developed during the past two to three decades in strongly fluctuating systems, near critical points and more generally at phase transitions. These effects are termed *critical phenomena* (*q.v.;* Domb and Green, 1976). Based on the renormalization-group ideas, formulation of scaling and universality for these systems has been developed, relat-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ing diverse systems in the regime of critical behavior. One of the most striking examples comprises the critical points of the liquid–gas phase transitions and of phase separation in certain binary mixtures, and the Curie points of uniaxial ferromagnets and of lattice models of ±1 "spins" (Ising models). All four systems fall in the same *universality class* in the sense that they are described by the same scaling relations. The Ising model, however, is much simpler than the other, experimental systems, and it has been used extensively in numerical simulations.

The rest of this article is devoted to universality in critical phenomena. Sections 1 through 4 introduce various aspects of the theoretical description of universal properties at phase transitions. Section 5 surveys the main calculation methods as well as selected results for the most familiar models. Additional topics are described briefly in Sec. 6.

1. CRITICAL PHENOMENA AT PHASE TRANSITIONS

Phase transitions are usually associated with singular or discontinuous behavior of thermodynamic quantities. As an illustration, in Fig. 1 the magnetization M of a ferromagnet is shown as a function of the temperature T. If there is no applied field, $H = 0$, then at low temperatures the system is spontaneously ordered. The magnetization is then nonzero, and in fact several directions of the magnetization can coexist. Some systems are *uniaxial* with "up" and "down" magnetization directions (termed Ising ferromagnets); others have vector magnetization pointing in an arbitrary direction in a plane (*XY* ferromagnets) or in three-dimensional space (Heisenberg ferromagnets), etc. In Fig. 1, two spin orientations are shown, thus assuming the simplest, uniaxial, Ising case. For temperatures above a certain transition point T_c, spontaneous magnetization disappears.

At the transition point, there may be a singularity, which is power-law in the leading order, $M \propto (T_c - T)^{\beta}$, as T approaches T_c from below. In that case we identify the critical or second-order phase transition, and the associated singular properties are termed critical phenomena. Figure 1 also illustrates a first-order phase transition for which the

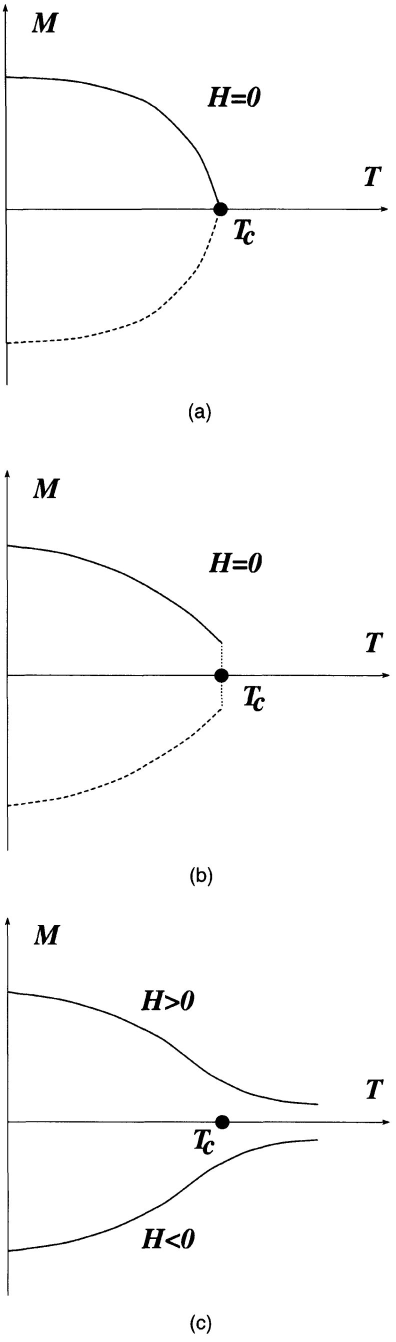

(a)

(b)

(c)

FIG. 1. Variation of the magnetization with temperature for a uniaxial (Ising) ferromagnet. In zero applied field, two directions of the magnetization are possible at low temperatures; see solid and dashed curves in (a) and (b). The magnetization vanishes singularly, (a), or discontinuously, (b), at the transition point, for second-order and first-order phase transitions, respectively. For nonzero applied field, the magnetization is nonsingular and it is aligned with the field; see (c).

variation of the magnetization across the transition point is discontinuous. The magnetization is termed the *order parameter* of the transition. Of course, the type and location of phase transitions for various substances, as well as the symmetry and the identification of the appropriate order-parameter variable, depend on the nature of the interactions of the microscopic particles (such as spins, atoms, and molecules) that are the constituents of the macroscopic thermodynamic system. In fact, singular or discontinuous behavior in equilibrium is normally associated with the *thermodynamic limit* of an infinite number of such particles, i.e., with a macroscopically large sample.

Other quantities are also singular at second-order phase transitions, and in fact they can actually diverge according to some power laws; expressions for quantities such as the susceptibility or the specific heat are introduced later. In a larger parameter space, the point of a second-order phase transition is usually a special point, termed *critical point*. For instance, Fig. 2 illustrates variation of the magnetization for an Ising ferromagnet near its critical point as a function of the applied field. We note that for each temperature $T < T_c$ the applied field actually drives the system through a first-order phase transition, i.e., discontinuous magnetization. The behavior is analytic for $T > T_c$. However, only for $T = T_c$ and at $H = 0$ do the singularities of the power-law type occur. Thus the ferromagnetic critical point is a special point at $(T,H) = (T_c,0)$.

Figure 3 shows phase diagrams of (a) ferromagnetic and (b) gas–liquid–solid phase coexistences. Large fluctuations develop only in a limited part of the parameter space. Specifically, for a uniaxial ferromagnet in applied field H and at temperature T, the line of coexistence of positive and negative magnetization M extends from $T = 0$ to the critical or Curie point at $T = T_c$. In the close vicinity of the critical point, i.e., for small field H and small dimensionless temperature deviation

$$t = (T - T_c)/T_c, \tag{1}$$

large regions of correlated fluctuations develop, and various thermodynamic quantities behave in a singular fashion. We note that the line of coexistence of positive and negative spontaneous magnetizations, at $H = 0$ and $T < T_c$, also corresponds to large-scale heterophase fluctuations. This first-order phase transition also has certain universal scaling properties, which are mentioned in Sec. 5. The main emphasis here is on the critical-point behavior.

The gas–liquid–solid phase diagram, Fig. 3(b), is more complicated. However, second-order–transition type singular behavior is only observed near the critical (or cloud) point at which the liquid–gas coexistence (evaporation) line ends. The other end of that line is the triple point, at which the melting and sublimation lines of first-order phase transitions meet with the evaporation line. In the vicinity of the critical point, one can define a temperature-like deviation variable "along" the coexistence line, analogous to t of the ferromagnet, as well as a fieldlike deviation variable in the direction "perpen-

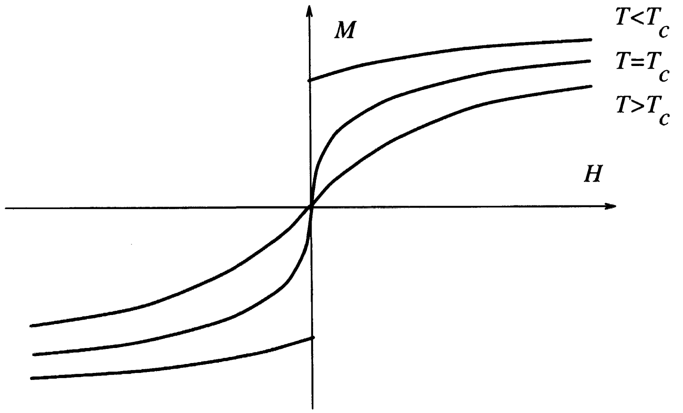

FIG. 2. Variation of the magnetization with applied field for a uniaxial (Ising) ferromagnet with a second-order phase transition point at $(T,H) = (T_c,0)$, at different temperatures.

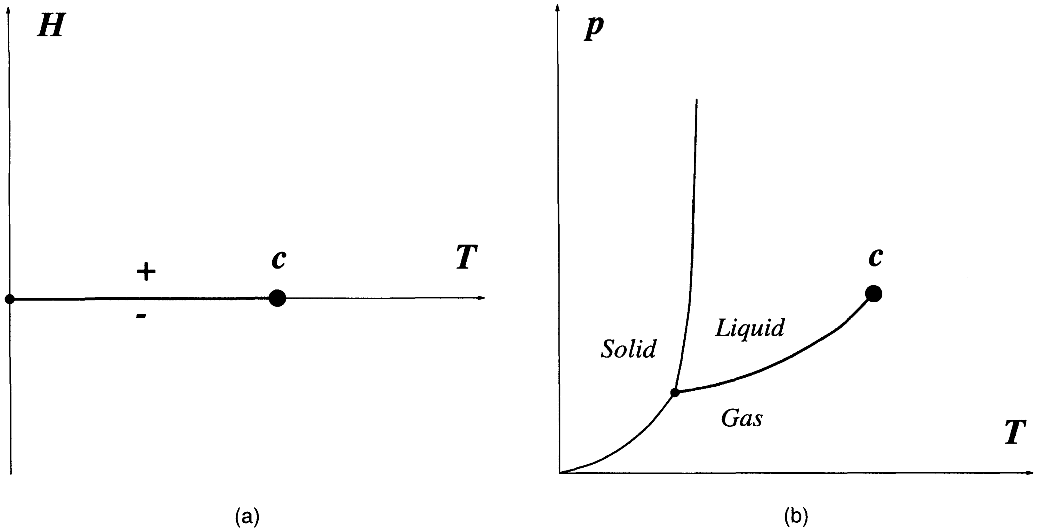

FIG. 3. (a) The ferromagnetic phase diagram showing the phase coexistence line for temperatures $T < T_c$, with the critical point (marked ***c***) at T_c; H denotes the magnetic field. (b) The gas–liquid–solid phase diagram with the liquid–gas phase coexistence line ending at the critical point (marked ***c***) at (T_c,p_c), where p denotes the pressure.

dicular" to the coexistence line. These are usually linear combinations of the temperature and chemical-potential deviations from their critical-point values (Privman *et al.*, 1991).

Let us also consider the nearest-neighbor Ising model defined by placing spins

$$s_i = \pm 1 \tag{2}$$

at sites i of a regular D-dimensional lattice. The energy of the spin configuration is

$$E\{s_i\} = -J \sum_{\text{pairs } m,n} s_m s_n - H \sum_k s_k. \tag{3}$$

Here the first sum is over all nearest-neighbor pairs. The coupling constant J is positive for the ferromagnetic behavior, while the applied field H can be positive or negative. For definiteness, we will use the dimensionless "magnetic field" variable of the Ising model,

$$h = H/k_B T, \tag{4}$$

where k_B is the Boltzmann constant. We note that this model has the phase diagram of Fig. 3(a). Similar dimensionless "applied field" variables analogous to H/k_BT can be defined for other systems.

2. SCALING AND UNIVERSALITY IN CRITICAL PHENOMENA

Near a critical point, singular behavior of thermodynamic quantities can be represented in the scaling form. In the limit $t \to 0$ and $h \to 0$, we have for the singular part of the free-energy density measured per k_BT

$$f_{\text{sing}}(t,h) \approx c_f |t|^{2-\alpha} Q_{\pm}(c_h h |t|^{-\Delta}), \tag{5}$$

where the scaling function $Q_{\pm}$ is different for $t > 0$ (+) and $t < 0$ (−). Both the scaling function $Q_{\pm}$ and the power-law exponents α and Δ are universal. The former is universal only provided we allow for the system-dependent constants c_f and c_h, frequently termed the *metric factors*. The origin of this universality is discussed in the next section.

Consider now various thermodynamic functions that can be derived from the free-energy density by differentiation. To be specific, we note that for the Ising model energy, Eq. (3), the free energy is given by

$$f(T,H) = -V^{-1} \ln \Lambda, \tag{6}$$

where V is the system volume (taken to infinity in the thermodynamic limit). The partition function Λ is defined as

$$\Lambda = \prod_i \left[\sum_{s_i=-1}^{1} \exp\left(-\frac{E}{k_B T}\right)\right]; \tag{7}$$

i.e., the Boltzmann factor $\exp(-E/k_B T)$ is summed over all system configurations. Then the thermodynamic average $\langle \mathbb{O} \rangle$ of any function $\mathbb{O}$ of the microscopic variables s_i can be calculated as

$$\langle \mathbb{O} \rangle = \Lambda^{-1} \prod_i \left[\sum_{s_i=-1}^{1} \mathbb{O} \exp\left(-\frac{E}{k_B T}\right)\right]. \tag{8}$$

Thus the magnetization,

$$M = \langle s_i \rangle, \tag{9}$$

is obtained as

$$M = \frac{\partial f}{\partial h}, \tag{10}$$

and similarly the specific heat, susceptibility, etc., are obtained as various derivatives of the free energy. We note that the choice of the free energy to "generate" the thermodynamic quantities depends on the physical system and on the independent variables selected; see Privman *et al.* (1991). For instance, for magnetic systems one could also use (T,M) as independent variables. The scaling expressions can then be written in terms of powers of t and M (Privman *et al.*, 1991).

Differentiation of Eq. (5) yields the singular behavior of the magnetization,

$$M \approx c_f c_h |t|^{\beta} W_{\pm}(c_h h |t|^{-\Delta}); \tag{11}$$

here

$$W_{\pm} = Q'_{\pm} \tag{12}$$

and

$$\beta = 2 - \alpha - \Delta. \tag{13}$$

Note that the symmetry $h \leftrightarrow -h$ implies that $W_{\pm}(w)$ are odd functions of their argument w. The meaning of the scaling relations is as follows. We consider behavior in the critical region, in the limit $t \to 0$, $h \to 0$, which corresponds to the vicinity of the critical point. While both t and h must be small for the scaling relations to apply (we mention corrections to scaling briefly in Sec. 5), the scaling combination $h/|t|^{\Delta}$ can be actually of order 1. Of course, the scaling function must correspond to the generally expected behavior of the appropriate quantity, here magnetization, outside the critical region. Specifically, we expect

$$W_{\pm}(w \gtrless 0) \gtrless 0, \tag{14}$$

$$W_{+}(0) = 0, \tag{15}$$

and

$$\lim_{w \to \pm 0} W_{-}(w) \gtrless 0, \tag{16}$$

where the latter property corresponds to spontaneous magnetization for $T < T_c$.

It is clear from the above discussion that exponents of the singular behavior, termed "critical exponents," and amplitudes entering scaling relations for various quantities derivable from the free-energy density are related. Indeed, they are expressed in terms of the exponents α and Δ and the metric factors c_f and c_h as well as the values of the universal functions $Q_{\pm}$ and their derivatives. For instance, for the magnetization, susceptibility χ (given by $\partial M/\partial h$), and specific heat C (proportional to $\partial^2 f/\partial t^2$), the leading-order zero-field critical behavior in the limit $t \to 0^-$, i.e., for temperatures below the critical point, is traditionally written as

$$M \approx B(-t)^{\beta}, \tag{17}$$

$$C \approx (A'/\alpha)(-t)^{-\alpha}, \tag{18}$$

$$\chi \approx \Lambda'(-t)^{-\gamma}. \tag{19}$$

Similarly, for $t \to 0^+$ we use the notation

$$C \approx (A/\alpha)t^{-\alpha}, \tag{20}$$

$$\chi \approx \Lambda t^{-\gamma}, \tag{21}$$

while $M \equiv 0$. Relations among exponents (termed *scaling relations*) then follow, such as

$$\Delta = \beta + \gamma, \tag{22}$$

$$\alpha + 2\beta + \gamma = 2. \tag{23}$$

Furthermore, universal amplitude combinations can be identified—for instance,

$$A/A', \tag{24}$$

$$\Gamma/\Gamma', \tag{25}$$

and

$$R_C = A\Gamma/B^2. \tag{26}$$

As emphasized, scaling relations are universally applicable to large classes of systems (see the next section for further explanation, as well as CRITICAL PHENOMENA and PHASE TRANSITIONS: RENORMALIZATION AND SCALING). Specifically, the scaling laws just defined apply not only to uniaxial ferromagnets but also to the liquid–gas and binary-mixture critical points. For instance, the $h = 0$ relations (17)–(21), etc., correspond to the liquid–gas critical isochore ($t > 0$) and phase boundary ($t < 0$). Other critical-point power-law singular behavior applies, for instance, along the critical isotherm. As an example, we consider $M(0,h)$; see Privman *et al.* (1991) for other quantities. The relation between M and h at $t = 0$ is written as

$$h \approx D_c M|M|^{\delta-1}. \tag{27}$$

In order to derive such a relation, we assume that the scaling functions $W_{\pm}(w)$ in Eq. (11) both have the same leading large-w asymptotic behavior,

$$W_{\pm}(w \text{ large}) \simeq \mathcal{W}_{\infty}\, \text{sign}(w)|w|^{\beta/\Delta}, \tag{28}$$

where the power law was selected to cancel out the $|t|$ dependence in the limit $t \to 0$. From this form, the exponent relation

$$\delta = \Delta/\beta \tag{29}$$

follows. Furthermore, since $\mathcal{W}_{\infty}$ is a universal coefficient, the nonuniversal factors in D_c follow as

$$D_c \propto c_f^{-\delta} c_h^{-\delta-1}. \tag{30}$$

The appropriate universal amplitude combination can then be identified as

$$R_\chi = AD_c B^{\delta-1}. \tag{31}$$

Behavior of other quantities along the critical isotherm, and the appropriate universal amplitude combinations, were reviewed by Privman *et al.* (1991).

Finally, let us mention the scaling of the two-point order-parameter correlation function. For ferromagnets, this is the spin–spin correlation function, while for liquid–gas systems, etc., it must be defined in terms of the local (microscopic) densities. The scaling of the correlation function $\mathcal{G}(\mathbf{r},t,h)$ is usually expressed in terms of its spatial Fourier transform $\hat{\mathcal{G}}(\mathbf{q},t,h)$ as follows:

$$\hat{\mathcal{G}} \approx c_{\mathcal{G}}|\mathbf{q}|^{-(2-\eta)}U(c_h h|t|^{-\Delta}, c_\xi|\mathbf{q}||t|^{-\nu}). \tag{32}$$

Note that new exponents were introduced as well as new metric factors. However, since the susceptibility can be obtained as the spatial integral of the correlation function (i.e., the $\mathbf{q} = 0$ value of its Fourier transform), two of the exponents and two of the amplitudes entering relation (32) are not independent (we already used the same Δ and c_h to simplify the notation). Specifically, we have the relation

$$2 - \eta = \gamma/\nu; \tag{33}$$

and, up to universal factors, $c_{\mathcal{G}}$ can be replaced by $c_f c_h^2 c_\xi^{-(\eta-2)}$.

The critical-point $\mathbf{q}$ dependence is frequently written as

$$\hat{\mathcal{G}}(\mathbf{q}, t = 0, h = 0) \approx D_\infty/|\mathbf{q}|^{2-\eta}. \tag{34}$$

Much attention has been paid to the quantity that scales the "inner" $\mathbf{q}$ dependence in relation (32), termed the *correlation length*, ξ. Actually, various correlation lengths can be identified, defined via the asymptotic fall-off of the correlation function with distance or via various spatial moments of its $\mathbf{r}$ dependence. The zero-field critical behavior is of the form

$$\xi = \xi_0 t^{-\nu} \quad (t > 0), \tag{35}$$

and

$$\xi = \xi_0'(-t)^{-\nu} \quad (t < 0), \tag{36}$$

with the obvious generalization to a scaling form for $h \neq 0$. Actually, for models with continuous symmetry of the order parameter, such as the O(N)-symmetric spin models with $N \geq 2$, the longitudinal and transverse correlation lengths must be identified separately for $t < 0$ (see Privman *et al.*, 1991).

Since $D_\infty \propto c_{\mathcal{G}}$ up to universal factors, while ξ_0 and ξ_0' are both proportional to c_ξ, these coefficients can be used to form universal amplitude combinations—for instance,

$$\xi_0/\xi_0' \tag{37}$$

and

$$Q_3 = D_\infty \xi_0^{2-\eta}/\Gamma. \tag{38}$$

However, a more direct relation between the "correlation" and "thermodynamic" amplitudes requires concepts of hyperscaling and hyperuniversality, discussed in the next section.

3. RENORMALIZATION, TWO-SCALE-FACTOR UNIVERSALITY, AND FINITE-SIZE SCALING

Universality in critical phenomena is described by the renormalization-group theory (Domb and Green, 1976; Phase Transitions: Renormalization and Scaling). Review of the renormalization-group theory is outside the scope of this article; see Critical Phenomena and Phase Transitions: Renormalization and Scaling. The renormalization-group transformations relate a system with correlations on the scale ξ to another system with correlations on the scale ξ/b, where $b > 1$.

This is accomplished by mapping the parameter space of "coupling constants" in a general-model Hamiltonian onto itself, while "integrating away" fluctuations on the shortest microscopic length scales. As a result, a modified, "coarse-grained" Hamiltonian is obtained with all the length scales, including the correlation length, scaled by b^{-1}, so that the new "microscopic" unit of resolution of fluctuations is of size b^D in terms of the original-model smallest resolution dimensions.

For $b \to \infty$, the critical point "flows" into a fixed point with $\xi = \infty$. Various systems are represented as special cases of the same general, all-encompassing model, with different initial parameters. The critical points of all models with the same universal properties, i.e., in the same universality class, are in the domain of attraction of the same fixed point of the renormalization-group transformation. The universality of critical exponents is obtained as the property of the renormalization-group flows linearized near the fixed point. The universality of scaling functions and various amplitude combinations is more subtle. It relates to the flow near the unstable manifold, i.e., that portion of the parameter space that is transformed away from the fixed point by the renormalization-group transformations.

The parameters of a given system are related to the parameters of the general model transformed by the renormalization group, in a complicated fashion. However, near criticality, a linear approximation is appropriate. For ordinary continuous phase transitions, such as the uniaxial ferromagnet transition that we use as a prototype, there are two directions in the parameter space that correspond to deviations from criticality. These are called the *relevant* renormalization-group fields: For a ferromagnet, they are proportional to t and h near the critical point. This implies that there are two independent metric factors for both thermodynamic and correlation properties. This result is termed two-scale-factor universality or *hyperuniversality*. It turns out that the metric-factor relation corresponds to the following: The combination $c_f c_\xi^D$ is universal, where D is the dimensionality of space. This is related to the observation that the free-energy density transforms like (length)$^{-D}$, while ξ transforms as length. A more widely used form is the universal (or rather hyperuniversal) amplitude combination

$$R_\xi^+ = A^{1/D}\xi_0. \tag{39}$$

The appropriate universal exponent relations, termed hyperscaling relations, also follow—for instance,

$$2 - \alpha = D\nu. \tag{40}$$

A physically interesting interpretation of the hyperscaling/hyperuniversality relations (39)–(40) is to consider the free energy of the large-scale fluctuations that presumably contribute the singular part of the total free energy. The free energy of such fluctuations in a volume ξ^D, when expressed as a dimensionless number (by measuring it per k_BT), is $\xi^D f_{\text{sing}}$. The metric factors entering this combination are $c_\xi^D c_f$, and so the hyperuniversality corresponds to the statement that the free energy of "singular" fluctuations in the volume defined by the correlation length is a universal multiple of k_BT. The underlying physical mechanism for this universality is the assertion that the correlation length is the only "macroscopic" length scale remaining after the microscopic fluctuations were excluded or "renormalized away" in the sense of being lumped into the metric factors.

It is important to note that these observations do not apply within the mean-field or effective-field theory of critical behavior (Stanley, 1971). The reason is that certain dependences on the microscopic parameters enter the mean-field singular behavior nontrivially. This relates to the fact that mean-field theories single out the uniform order-parameter fluctuations rather than all the large-scale fluctuations; see Stanley (1971). Thus the mean-field "law of corresponding states" (see EQUATIONS OF STATE) is a dimensional-analysis type rather than a scaling result.

For a finite-size system of linear dimension L, with its interaction parameters at their critical-point values ($t = 0, h = 0$), the total free energy (per k_BT) in the actual system volume $\propto L^D$ is universal (Privman, 1990). One can show that the appropriate scaling law takes the form

$$f_{\text{sing}} \approx L^{-D} Y(a_t t L^{1/\nu}, a_h h L^{\Delta/\nu}). \tag{41}$$

This entails a universal finite-size amplitude,

$$Y(0,0) = \lim_{L\to\infty} f_{\text{sing}}(t = 0, h = 0) L^D. \tag{42}$$

Similarly, the ratio of two macroscopic lengths,

$$X(0,0) = \lim_{L\to\infty} \xi(t = 0, h = 0)/L, \tag{43}$$

is universal, where the scaling relation for the correlation length takes the form

$$\xi \approx LX(a_t t L^{1/\nu}, a_h h L^{\Delta/\omega}). \tag{44}$$

Several comments are in order. First, we note that finite-size scaling relations (41)–(44) apply in the limit $t \to 0$, $h \to 0$, $L \to \infty$, with the appropriate scaling combinations kept of order 1. The last of these limits assumes fixed system shape, and in fact the universal scaling functions depend on the shape. Second, the "singular" subscript in the free energy refers to the bulk limit of $L \to \infty$ with fixed (small) t and h. For a fully finite system, the thermodynamic functions are of course not singular. Finite-size scaling relations also apply to systems such as "slabs" infinite in two dimensions and finite in one, or "strips" in $D = 2$ infinite in one dimension and finite in another; see Privman (1990) for details. The "bulk" limit $L \to \infty$ corresponds to the large-argument behavior of $Y(\lambda,\zeta)$ as a function of its arguments λ and ζ,

$$Y(\lambda \text{ large}, \zeta \text{ large}) \simeq |\lambda|^{D\nu} Q_\pm(\zeta |\lambda|^{-\Delta}), \tag{45}$$

where ± corresponds to the limits $\lambda \to \pm\infty$. To identify fully the bulk critical behavior (5), the hyperscaling relation (40) is needed, and we also obtain metric-factor relations

$$c_f = a_t^{D\nu} \tag{46}$$

and

$$c_h = a_h a_t^{-\Delta} \tag{47}$$

(we assume that all the metric factors are positive).

Our final comment on the finite-size scaling relations (41)–(44) is that they apply only to systems with periodic, helical, or other boundary conditions such that each microscopic system part is essentially in the "bulk." For more realistic boundary conditions, there will be surface, edge, and corner contributions to the free-energy density and other quantities. Their interplay with the leading scaling expressions such as (41) may be nontrivial. For instance, for short-range boundary-interaction models, the corner term in the free energy leads to an added universal term with a logarithmic size dependence; see Privman (1990) for details.

4. CROSSOVER SCALING AND RELATED TOPICS

Scaling forms appropriate for multicritical points, at which lines of phase transitions meet, are termed *crossover scaling*. Typically, additional relevant renormalization-group fields are needed to describe behavior near multicritical points such as bicritical, tricritical, tetracritical, etc. (Fisher, 1974; Aharony, 1983; Knobler and Scott, 1984).

As an illustration, we consider crossover scaling at tricritical points. A tricritical point can be found in the Ising model (3) when additional interactions are present that compete with the ferromagnetic nearest-neighbor interactions. For actual examples of systems with tricritical points, see Lawrie and Sarbach (1984), and Knobler and Scott (1984). The generic phase diagram is illustrated in Fig. 4. Here G denotes the added-interaction coupling constant. The lines of second-order and first-order transitions meet at the tricritical point (Fig. 4).

In addition to the temperature-like relevant scaling field t measuring deviations from the tricritical point at (T_t,G_t) "perpendicular" to the phase-transition lines, there is also the second relevant field, g, "along" the lines of phase transition. These variables are linear combinations of the deviations $(T - T_t)/T_t$ and $(G - G_t)/G_t$. With an applied field, not considered here, the phase-transition structure near the tricritical point is even richer (Lawrie and Sarbach, 1984).

The crossover scaling description for the magnetization corresponds to replacing (17) with the scaling form

$$M \approx B_t(-t)^{\beta_t}\mathcal{X}(c_g g(-t)^{-\phi}), \tag{48}$$

where ϕ is termed the crossover exponent. If $T(G)$ denotes the temperature of the transition line, then at fixed G, $M \equiv 0$ for $T > T(G)$. For other quantities, two scaling functions (labeled ± as before) are needed. The phase-transition behavior near (T_t,G_t) is "hidden" in the universal scaling function $\mathcal{X}$.

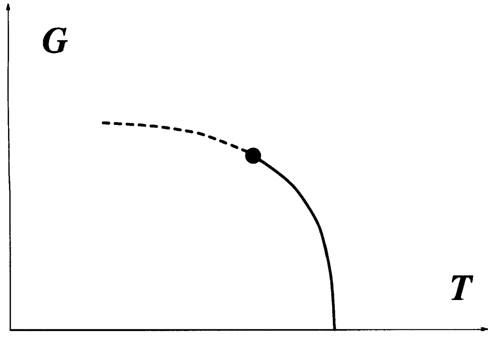

FIG. 4. Schematic illustration of the tricritical point (marked by heavy dot) and the lines of second-order (solid line) and first-order (dashed line) phase transitions, in the plane of temperature and added coupling constant, G; see text.

It is best to explain this by example. Let us consider the case of the mean-field theory. With the appropriate choice of the metric factors B_t and c_g, the mean-field expression for the scaling function is

$$\mathcal{X}(\omega) = \sqrt{\omega + \sqrt{\omega^2 + 3}}, \tag{49}$$

where we use the convention that $g > 0$ [$\omega = g(-t)^{-1/2} > 0$] in the direction of the first-order transition line, i.e., for $G > G_t$ in the notation of Fig. 4. Note that

$$\beta_t = \tfrac{1}{4} \text{ and } \phi = \tfrac{1}{2} \tag{50}$$

in the mean-field theory.

The form of the relation (49) for $g < 0$ corresponds to the vanishing magnetization at the phase-transition line in the limit $t \to 0^-$, i.e., to the continuous (second-order) phase transition. For small fixed $g < 0$, the exponent β equals $\frac{1}{2}$ since the scaling function behaves as

$$\mathcal{X} \propto (-\omega)^{-1/2} \text{ as } \omega \to -\infty. \tag{51}$$

The magnetization, therefore, vanishes as

$$M \propto (-t)^{1/4}[(-t)^{-1/2}]^{-1/2} = (-t)^{1/2}. \tag{52}$$

However, for small fixed $g > 0$, i.e., in the limit $\omega \to +\infty$, the magnetization remains finite, of order $\sqrt{g}$, corresponding to a discontinuous jump to zero at the phase transition line, as expected for a first-order transition.

Universal amplitude combinations at multicritical points were reviewed by Privman *et al.* (1991). Scaling for dynamics (Hohenberg and Halperin, 1977; Privman *et al.*, 1991), as well as for certain surface-interaction couplings in systems that are not infinite in all dimensions (Diehl, 1986), also involves new scaling combinations. Typically, there is a new exponent and a new nonuniversal amplitude (metric factor) for each added scaling variable. For instance, for dynamics, the large-time slow-relaxation modes near the critical point have their time scales (termed the relaxation times) scaled by

$$\xi^z \propto |t|^{-z\nu}, \tag{53}$$

where z is the new, dynamic critical exponent.

5. NUMERICAL AND ANALYTICAL METHODS AND SELECTED RESULTS

In this section we survey different techniques used to calculate universal exponents and amplitude combinations. Our aim is to enumerate the methods and provide literature citation; we avoid discussion of any details. However, we do comment on their relative advantages and limitations. Selected results are then given.

We first comment on two general principles that yield information mostly in two dimensions (2D), namely, conformal invariance (also termed conformal field theory) and duality. Conformal invariance (Cardy, 1987; Christe and Henkel, 1993) provides exact values of the 2D critical exponents and universal finite-size amplitudes, for a wide variety of universality classes. Certain conclusions regarding bulk amplitudes are also possible. The property of duality or self-duality relating quantities for $T > T_c$ to those for $T < T_c$ holds for many lattice models in two dimensions—e.g., Potts models (Wu, 1982) or percolation (Deutscher *et al.*, 1983). For example, we have the universal ratio

$$A/A' = 1, \tag{54}$$

due to duality, for the square-lattice q-state Potts models (Baxter, 1982; Wu, 1982), which have second-order phase transition for $0 \leq q \leq 4$. By universality, this result applies to all models in the appropriate universality classes. Note that $q = 2$ corresponds to the Ising model, while the limit $q \to 1$ can be interpreted to yield results for percolation.

Some models are exactly solvable. They yield important sources of universal values for approximate description of more realistic systems and for theoretical developments. These include the mean-field approach (infinite-range interactions; effectively, infinite dimensionality; see Stanley, 1971), and the spherical models (Joyce, 1972) solvable in general dimension. Ising and certain other lattice models have been solved in two dimensions (McCoy and Wu, 1973; Baxter, 1982). There are also several exact results for fractal and hierarchical lattices; see FRACTAL GEOMETRY.

Series expansions (Domb and Green, 1974; Guttmann, 1989; Baker, 1990) have traditionally yielded accurate information on the critical behavior. Expansions about the disordered state—the high-temperature series—are available to large orders for many lattice models. Low-temperature series (expansions about ordered states) are frequently more difficult to derive to high orders. Elaborate analysis methods have been devised for series expansions to evaluate behavior at "ordinary" critical-point singularities, including effects of corrections to scaling, etc. However, in mapping out phase diagrams that contain multicritical points, only limited results are available and the field is wide open to introduction of new, reliable analysis techniques.

Other numerical methods are based on finite-size scaling and related extrapolation techniques. They include the Monte-Carlo (Binder and Heermann, 1988; Privman, 1990; MONTE-CARLO METHODS) and transfer-matrix (Nightingale, 1982; Privman, 1990) approaches. The Monte-Carlo computer simulation method can be applied to nearly any equilibrium model, and, with stochastic evolution rules, also to dynamics. It is more versatile than other approaches, and it has become the method of choice with the advent of powerful modern computers. On account of computational difficulties, the transfer-matrix approach, in which the system is built up as a finite–cross-section "cylinder" from narrow slices, has been used extensively only in two dimensions.

Techniques based on the renormalization-group approach include ϵ expansions (Domb and Green, 1976), i.e., expansions in the deviation of the dimensionality of space from its "upper critical" value above which mean-field behavior applies. Expansions about the spherical-model limit, which corresponds to an infinite number of spin components, also can be cast in the renormalization-group form (Ma, 1973). Field-theoretical renormalization-group calculations directly in three dimensions (Domb and Green, 1976; Privman *et al.*, 1991), and certain other approximation schemes, also fall in this group. All renormalization-group methods yield qualitative results for the phase diagram, such as type of multicriticality, universality class, etc. Only for a few main universality classes (such as the three-dimensional Ising) have the renormalization-group calculations been carried out to high enough orders and analyzed in sufficient detail to yield quantitatively accurate results for universal exponents and amplitudes. However, as already mentioned, the renormalization-group methods are invaluable in deriving qualitatively the "global" properties, including various scaling relations, limiting properties and general behavior of universal scaling functions, etc.

In Table 1 we list selected results for universal exponents of the Ising universality class in two and three dimensions, as well as certain mean-field results. We list theoretical values of several universal quantities; for further results see Domb and Green (1976), Baker (1990), Privman (1990), and Privman *et al.* (1991). We note that all the nonexact results presented (mostly in three dimensions) are typical ranges obtained by combining results of several theoretical methods discussed above. Two universal quantities shown at the end of the table are defined in Sec. 6. Some mean-field entries are left blank because of unavailability of an estimate, inapplicability of hyperscaling above the upper critical dimension, which is $D = 4$ for the Ising universality class, or inapplicability of the amplitude definition. Specifically, $\alpha = 0$ in mean-field theory corresponds to a discontinuous jump in the specific heat, so that the ordinary definition of A/A' does not apply. Similarly, the $\alpha = 0$ value for $D = 2$ actually corresponds to a symmetric logarithmic divergence, represented nominally by the $A/A' = 1$ entry, although the amplitudes must be appropriately redefined. The finite-size results are for cubic (square in 2D) shapes with periodic boundary conditions; see Privman (1990) and Privman *et al.* (1991) for other results for different system shapes and boundary conditions.

Table 1. Selected universal exponents, amplitudes, and amplitude ratios of the Ising universality class in two dimensions (2D), three dimensions (3D), and the mean-field regime ($D > 4$).

	2D	3D	$D > 4$
α	0(log)	0.11 ± 0.01	0
β	$\frac{1}{8}$	0.325 ± 0.006	$\frac{1}{2}$
γ	$\frac{7}{4}$	1.241 ± 0.004	1
ν	1	0.630 ± 0.002	$\frac{1}{2}$
A/A'	1	0.53 ± 0.02	
Γ/Γ'	37.69 ...	4.9 ± 0.2	2
R_C	0.318 5699 ...	0.057–0.066	
R_χ	6.78 ± 0.01	1.7 ± 0.1	1
ξ_0/ξ'_0	2	1.91–1.98	$\sqrt{2}$
Q_3	0.291 ...	0.90–0.92	1
R_ξ^+	$1/\sqrt{2\pi} = 0.3989$ ...	0.27 ± 0.01	
$Y(0,0)$	-0.6399...	-0.64 ± 0.04	
g_*	14.66 ± 0.06	28.85 ± 0.02	
g_∞	1.83 ± 0.01	1.3–1.4	0.8–1.0

While Table 1 illustrates variation with dimensionality, Table 2 illustrates variation with the universality class. Thus, selected universal values are given for O(N)-symmetric order-parameter models in $D = 3$. The Ising model has uniaxial symmetry (scalar order parameter), $N = 1$. The $N = 2$, plane-vector–symmetry, *XY*-model results are also given, as are the results for the $N = 3$, Heisenberg spin models. These values are also compared with expressions from the exact solution of the spherical model, which, up to certain N-dependent factors, corresponds to the limit $N \to \infty$. Exponent values for $N = 0$, which formally map onto a polymer problem (to be discussed in Sec. 6), are also included in Table 2. For further theoretical results and tabulations for these and many other models in different universality classes, notably, Potts models, percolation, and various

Table 2. Selected universal exponents and amplitude ratios of the three-dimensional Ising ($N = 1$), *XY* ($N = 2$), Heisenberg ($N = 3$), and spherical-model ($N = \infty$) universality classes. All the values for $N < \infty$ are approximate. The results for $N = \infty$ are exact. Exponent values shown for $N = 0$ correspond to the polymer-conformation universality class.

	$N = 1$	$N = 2$	$N = 3$	$N \to \infty$	$N = 0$
ν	0.63	0.67	0.71	1	0.59
β	0.33	0.35	0.36	$\frac{1}{2}$	0.30
γ	1.24	1.32	1.39	2	1.16
A/A'	0.53	1.05	1.58	1.467 ...	
R_ξ^+	0.27	0.36	0.43	$\propto \pi N^{1/3}$	
R_C	0.06	0.35	0.17	$\propto N$	
R_χ	1.7	1.4	1.3	1	
Q_3	0.91		0.92	1	

multicritical points, see Fisher (1974), Domb and Green (1976), Wu (1982), Baker (1990), Privman (1990), Privman *et al.* (1991), as well as CRITICAL PHENOMENA and PHASE TRANSITIONS: RENORMALIZATION AND SCALING.

We note that most of the two- and three-dimensional results summarized in Table 1 have been tested experimentally, some exponents quite accurately and other values, especially amplitude ratios, semiquantitatively. The $N = 2$ and $N = 3$ values (Table 2) were tested accurately for some exponents and semiquantitatively for most other exponent and amplitude values. The experimental status of tests of universality in critical phenomena is generally quite satisfactory, especially for the most familiar universality classes such as Ising models, percolation, etc.; see Privman *et al.* (1991) for a comprehensive review.

6. POLYMERS, PERCOLATION, CHEMICAL REACTIONS, MISCELLANEOUS RESULTS

In this section we address several additional topics of interest in universality of various exponents, amplitudes, and amplitude ratios. First, let us define the "renormalized coupling constant" amplitude combination g_*, encountered in Table 1. We consider the leading nonlinear susceptibility,

$$\chi^{(\mathrm{nl})} = \frac{\partial^3 M}{\partial h^3}, \tag{55}$$

with the $h = 0$, $t > 0$ amplitude defined by

$$\chi^{(\mathrm{nl})} \approx \Gamma^{(\mathrm{nl})} t^{-2\beta-3\gamma}. \tag{56}$$

We then introduce the amplitude ratio

$$g_* = -\Gamma^{(\mathrm{nl})} / \xi_0^D \Gamma^2. \tag{57}$$

This quantity has received attention in the literature because its universality is a hyperuniversal property relatively easy to probe by various numerical methods. Furthermore, there exists a finite-size counterpart of this quantity, also encountered in Table 1, defined for a finite-size system of linear dimension L, at $h = 0$, $t = 0$, via

$$g_\infty = \lim_{L\to\infty} [-\chi^{(\mathrm{nl})} / L^D \chi^2]. \tag{58}$$

For further discussion and results for these quantities, see Privman (1990) and Privman *et al.* (1991).

Phase separation in polymer mixtures belongs to the Ising universality class. The length of the chains, i.e., the polymerization index $\mathcal{N}$, is a nonuniversal "microscopic" parameter. The $\mathcal{N}$ dependence must therefore cancel out in universal quantities. The $\mathcal{N}$ dependence of various amplitudes has been considered in the literature. There has been certain controversy over the precise form of this dependence and over experimental verification of the theoretical predictions; see Privman *et al.* (1991) for details.

Scaling for first-order phase transitions is different from that for critical points (second-order transitions) in several aspects. The only large-scale fluctuations are those associated with coexistence of phases (see PHASE EQUILIBRIA; PHASE SEPARATION). These lead to no power-law singularities at the Ising (scalar order parameter) transitions (although there is a weak "essential" singularity associated with droplet excitations). However, for continuous symmetries, such as the O(N)-symmetric spin models with $N \geq 2$, there are the soft-mode spin-wave "Goldstone boson" singularities present at the phase transition point. Finite-size scaling introduces a new length (system size). For finite-size systems, scaling forms involving the system size have been identified, and universal scaling functions were calculated exactly; see Privman (1990).

"Geometrical" critical phenomena include fractals, polymer conformations, percolation models, etc.; see FRACTAL GEOMETRY. In some instances there exist direct connections between these geometrical-criticality models and spin-system phase transitions. Specifically, let us consider polymer conformations in the fixed-fugacity ensemble, i.e., statistical averages over the configurations of $\mathcal{N}$-step chains (random walks with or without self-avoidance constraints) with statistical weight $y^{\mathcal{N}}$. Here y is the fugacity per step. The random-walk statistics can be mapped onto the $N \to 0$ limit of the O(N)-symmetric spin model (de Gennes, 1979), critical-point amplitudes and exponents defined, and their universality studied. We list $N = 0$ exponent estimates in Table 2.

Similarly, for the percolation model, defined by averaging over connected clusters of sites or bonds with the appropriately defined weights (Deutscher *et al.*, 1983), universal exponents and amplitudes can be defined in the "phase-transition" nomenclature; see Deutscher *et al.* (1983) or Privman *et al.* (1991) for tabulations of results. Furthermore, the percolation model can be mapped onto the q-state Potts model in the limit $q \to 1$. Another geometrical model where the phase-transition type universal quantities were extensively studied is the lattice-animal statistics, which corresponds to averaging over connected clusters by the site- or bond-content fugacity only (while in the percolation model there is a statistical weight "cost" of the reduced connectivity at the cluster boundary).

All three models—polymers (walks), percolation, lattice animals—have been studied in detail for their universal-exponent values. However, universal amplitude combinations have attracted less interest. Detailed results are available only for the percolation model; see Deutscher *et al.* (1983) and Privman *et al.* (1991). Instead, attention has been devoted to the universality of various cluster shape and size ratios. The shape and size of $\mathcal{N}$-element clusters is characterized by D-dimensional tensors such as the moment-of-inertia tensor or the squared-radius-of-gyration tensor. Let us consider the latter object. For an $\mathcal{N}$-site (or -bond) cluster, the (i,j) element of this tensor is defined by

$$\frac{1}{2\mathcal{N}^2}\sum_{k=1}^{\mathcal{N}}\sum_{m=1}^{\mathcal{N}}(r_k^{(i)} - r_m^{(i)})(r_k^{(j)} - r_m^{(j)}), \qquad (59)$$

where $r_n^{(p)}$ is the pth component of the coordinate vector $\mathbf{r}_n$ of the nth cluster site (bond). Universal shape measures can be defined by forming ratios of, and statistically averaging, various size measures derived from the radius-of-gyration tensor. The averaging in this case is over cluster shapes only, with fixed number of elements $\mathcal{N}$.

Consider, for instance, the most familiar of the cluster-size measures, the squared radius of gyration $R_{\mathcal{N}}^2$, which is obtained by averaging the trace of the tensor (59) over all the fixed-$\mathcal{N}$ cluster geometries, with the appropriate statistical weights. One can then identify the amplitude of the large-$\mathcal{N}$ asymptotic form,

$$R_{\mathcal{N}}^2 \approx \rho_G \mathcal{N}^{2\nu}, \qquad (60)$$

where ν is the familiar universal correlation-length exponent. The amplitude ρ_G is not universal. However, it can be used to form universal ratios involving other size-measure amplitudes. For instance, for self-avoiding random walks, one can define a similar amplitude of the squared averaged end-to-end distance, ρ_E. The ratio

$$\rho_E/\rho_G \qquad (61)$$

is universal; its values are 1, $\simeq$7.1, and $\simeq$6.4 in one, two, and three dimensions, respectively. Tabulation of several other universal amplitude ratios can be found in Privman *et al.* (1991).

As mentioned already in Sec. 3, scaling for dynamics in systems that reach equilibrium is reminiscent of the crossover scaling in that a new scaling combination involving the dynamic critical exponent enters (Hohenberg and Halperin, 1977; Privman *et al.*, 1991). There are, however, systems with irreversible dynamics, such as chemical reactions, certain deposition-with-relaxation models, and cluster growth and nucleation models, in which power laws with universal exponents and, in some instances, universal amplitudes, have been found (Privman, 1994). Specifically, we consider reaction–diffusion systems in low dimensions (Privman, 1997), which have been investigated extensively recently with emphasis on fluctuation-dominated effects leading to the breakdown of the standard chemical rate equations that correspond to the "mean-field" approximation of the reaction kinetics (see CATALYSIS; CHEMICAL REACTIONS).

As an illustration, we consider the simplest reactions of two-particle coagulation,

$$\mathcal{A} + \mathcal{A} \to \mathcal{A}, \qquad (62)$$

and annihilation,

$$\mathcal{A} + \mathcal{A} \to \text{inert}. \qquad (63)$$

The mean-field rate equations predict that the particle density decreases proportionally to $1/\mathcal{T}$, where $\mathcal{T}$ denotes the time. This large-time result gives the mean-field exponent 1 of the time dependence. It turns out that in dimensions below the upper critical dimension for these reactions, $D = 2$, an effect of fluctuations is to change the power law. For instance, in $D = 1$, the density is proportional to $u/\mathcal{T}^{1/2}$, where by the appropriate definition of dimensionless quantities, the amplitude u can be made universal (for each type of model). For heterospecies reactions such as

$$\mathcal{A} + \mathcal{B} \rightarrow \text{inert}, \tag{64}$$

with equal initial concentrations of both diffusing species, the density decreases according to a power law with the universal exponent $\frac{1}{4}$ in $D = 1$. The upper critical dimension for the heterospecies reactions turns out to be $D = 4$.

Finally, let us consider briefly corrections to scaling. Renormalization-group theory predicts that typically the leading power-law singularity will be accompanied by weaker, subleading singularities, and also by analytic corrections to scaling. Specifically, the ratio of the full quantity $\Phi(t)$ to its leading singular behavior, $\Psi t^{-\psi}$, has the following form for small t:

$$\Phi(t)/[\Psi t^{-\psi}] \approx 1 + \tau t^{\theta} + \sigma t + \dots . \tag{65}$$

The amplitudes of the first subleading correction-to-scaling terms in different physical quantities can be used to form certain universal ratios. The latter have been calculated theoretically and some measured experimentally (Privman *et al.*, 1991). The power-law exponent $\theta > 0$ is universal, and, unlike the leading exponent ψ or the amplitudes τ and σ, it is the same for various thermodynamic and correlation quantities as long as their t dependence is considered, where for simplicity we took $h = 0$ and $t > 0$ in writing (65). Amplitudes of the leading analytic corrections to scaling, σ, satisfy certain universal sum rules; details of all these developments are not discussed here; see Privman *et al.* (1991).

In summary, we have reviewed the concept of universality in modern physics and its applications to behavior of systems that in their "critical" range of parameters are dominated by large-scale spatial or temporal fluctuations.

GLOSSARY

Corrections to Scaling: Singular behavior subleading to the main singularities at critical points.

Correlation Length: Linear size of the region of correlated fluctuations in a thermodynamic system.

Critical Exponents: Power-law exponents of the singular behavior at critical points, as well as exponents of powers entering scaling combinations that in turn serve as arguments of scaling functions.

Critical Point: A point in the space of thermodynamic parameters at which the thermodynamic system develops anomalous fluctuations.

Critical Region: Vicinity of the critical point. See CRITICAL PHENOMENA.

Crossover Scaling: Scaling relations involving scaling variables other than the "basic" ones: temperature, applied field, system size. This term is sometimes reserved for those scaling relations in which the scaling function itself contains singularities corresponding to a crossover to phase transitions along lines ending at the point near which the scaling relation applies.

Heisenberg Model: Model of systems with the three-component, continuous-symmetry order parameter, such as the isotropic-symmetry magnets (see MAGNETIC ORDERING IN SOLIDS).

Hyperscaling: Relations between thermodynamic and correlation critical exponents that involve the dimensionality of space and are associated with the property of hyperuniversality (or two–scale-factor universality).

Hyperuniversality: Relation between free-energy and correlation amplitudes implying that the thermodynamic and correlation quantities have the same metric factors entering their scaling relations. See the related concept of **Two–Scale-Factor Universality.**

Ising Model: Lattice model of ±1 (or "up-down") spin variables.

Lattice-Animal Model: Statistics of connected clusters.

Metric Factors: Nonuniversal parameters relating the physical "deviation from criticality" variables to the scaling fields of the renormalization-group transformation. Metric factors are introduced in scaling relations in order to make the scaling functions universal.

Multicritical Points: "Critical" phase-transition points at which several lines of first-order and second-order transitions meet.

Percolation: Model of connectivity (conductivity) in a disordered medium, defined by considering connected (conducting) clusters. For lattice models, these clusters may be defined, for instance, as groups of connected bonds or sites.

Potts Model: Lattice model of scalar, q-value spins with interaction that favors spin "alignment" in the sense of adjacent spins having the same value out of the q possible. For $q = 2$ the Ising model is obtained.

Renormalization Group: Theory relating systems with fluctuations on different length scales; see PHASE TRANSITIONS: RENORMALIZATION AND SCALING.

Scaling: Reduction of the number of independent variables, achieved typically by separating "macroscopic" spatial and temporal scales from "microscopic" scales.

Tricritical Point: Point at which lines of first-order and second-order transitions meet.

Two-Scale-Factor Universality: Conclusion that a single metric factor is needed for each scaling variable, the same in thermodynamic and correlation quantities. See the related concept of **Hyperuniversality.**

Universality: Application of the same laws and relations, notably dimensional analysis and scaling, to different physical systems.

Universality Class: Different physical systems with similar singular behavior of thermodynamic and correlation quantities at their critical points.

***XY* Model:** Model of systems with a two-component, continuous-symmetry order parameter, including the plane-symmetry magnets (see MAGNETIC ORDERING IN SOLIDS), superfluids (SUPERFLUIDITY: LIQUID-HELIUM SYSTEMS), and superconductivity (SUPERCONDUCTIVITY, LOW-TEMPERATURE).

Works Cited

Aharony, A. (1983), in: F. J. W. Hahne (Ed.), *Critical Phenomena,* Lecture Notes in Physics, Vol. 186, Berlin: Springer-Verlag, pp. 209–258.

Baker, G. A., Jr. (1990), *Quantitative Theory of Critical Phenomena,* Boston: Academic.

Baxter, R. J. (1982), *Exactly Solved Models in Statistical Mechanics,* London: Academic.

Binder, K., Heermann, D. W. (1988), *Monte Carlo Simulation in Statistical Physics; An Introduction,* Springer Series in Solid State Sciences, Vol. 80, Berlin: Springer-Verlag.

Cardy, J. L. (1987), in: C. Domb, J. L. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena,* Vol. 11, London: Academic, Chap. 2, pp. 55–126.

Christe, P., Henkel, M. (1993), *Introduction to Conformal Invariance and Its Applications to Critical Phenomena,* Heidelberg: Springer-Verlag.

de Gennes, P. G. (1979), *Scaling Concepts in Polymer Physics,* Ithaca: Cornell.

Deutscher, G., Zallen, R., Adler, J. (1983), *Percolation Structures and Processes,* Annals of the Israel Physical Society, Vol. 5, Bristol: Adam Hilger.

Diehl, H. W. (1986), in C. Domb, J. L. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena,* Vol. 10, London: Academic, Chap. 2, pp. 75–267.

Domb, C., Green, M. S. (Eds.) (1974), *Phase Transitions and Critical Phenomena,* Vol. 3, London: Academic.

Domb, C., Green, M. S. (Eds.) (1976), *Phase Transitions and Critical Phenomena,* Vol. 6, London: Academic.

Fisher, M. E. (1974), *Rev. Mod. Phys.* **46,** 597–616.

Guttmann, A. J. (1989), in: C. Domb, J. L. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena,* Vol. 13, London: Academic, Chap. 1, pp. 3–234.

Hohenberg, P. C., Halperin, B. I. (1977), *Rev. Mod. Phys.* **49,** 435–479.

Joyce, G. S. (1972), in: C. Domb, M. S. Green (Eds.), *Phase Transitions and Critical Phenomena,* Vol. 2, London: Academic, Chap. 10, pp. 375–442.

Knobler, C. M., Scott, R. L. (1984), in: C. Domb, J. L. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena,* Vol. 9, London: Academic, Chap. 2, pp. 163–231.

Lawrie, I. D., Sarbach, S. (1984), in C. Domb, J. L. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena,* Vol. 9, London: Academic, Chap. 1, pp. 1–161.

Ma, S. K. (1973), *Rev. Mod. Phys.* **45,** 589–614.

McCoy, B. M., Wu, T. T. (1973), *The Two-Dimensional Ising Model,* Cambridge, MA: Harvard.

Nightingale, M. P. (1982), *J. Appl. Phys.* **53,** 7927–7932.

Privman, V. (Ed.) (1990), *Finite Size Scaling and Numerical Simulation of Statistical Systems,* Singapore: World Scientific.

Privman, V. (1994), *Trends Stat. Phys.* **1,** 89–95.

Privman, V. (Ed.) (1997), *Nonequilibrium Statistical Mechanics in One Dimension,* Cambridge, U.K.: Cambridge Univ. Press.

Privman, V., Hohenberg, P. C., Aharony, A. (1991), in: C. Domb, J. L. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena,* Vol. 14, London: Academic, Chap. 1, pp. 1–134.

Schmidt, R., Housen, K. (1995), *The Industrial Physicist* **1,** 21–24.

Stanley, H. E. (1971), *Introduction to Phase Transitions and Critical Phenomena,* New York: Oxford Univ. Press.

Wu, F. Y. (1982), *Rev. Mod. Phys.* **54,** 235–268.

Further Reading

Dimensional Analysis in Physics

Schmidt, R., Housen, K. (1995), *The Industrial Physicist,* **1,** 21–24.

Dimensional Analysis in Biology and Other Fields

Thompson, D'Arcy Wentworth (1942), *On Growth and Form,* Cambridge, U.K.: Cambridge Univ. Press.

Textbooks on Phase Transitions

Baker, G. A., Jr. (1990), *Quantitative Theory of Critical Phenomena,* Boston: Academic.

Ma, S.-K. (1976), *Modern Theory of Critical Phenomena,* London: Benjamin.

Stanley, H. E. (1971), *Introduction to Phase Transitions and Critical Phenomena,* New York: Oxford Univ. Press.

Yeomans, J. M. (1992), *Statistical Mechanics of Phase Transitions,* New York: Oxford Univ. Press.

Book Series Containing Comprehensive Technical Reviews of Various Developments in the Theory of Critical Phenomena

Domb, C., Green, M. S. (Eds.) (1974–1976), *Phase Transitions and Critical Phenomena,* Vols. 1–6, London: Academic.

Domb, C., Lebowitz, J. L. (Eds.) (1983-present), *Phase Transitions and Critical Phenomena,* Vols. 7 on, London: Academic.

Technical Books Covering Selected Topics and Techniques

Amit, D. J. (1978), *Field Theory, The Renormalization Group and Critical Phenomena,* New York: McGraw-Hill.

Baxter, R. J. (1982), *Exactly Solved Models in Statistical Mechanics,* London: Academic.

Binder, K., Heermann, D. W. (1988), *Monte Carlo Simulation in Statistical Physics; An Introduction,* Springer Series in Solid State Sciences, Vol. 80, Berlin: Springer-Verlag.

Christe, P., Henkel, M. (1993), *Introduction to Conformal Invariance and Its Applications to Critical Phenomena,* Heidelberg: Springer-Verlag.

de Gennes, P. G. (1979), *Scaling Concepts in Polymer Physics,* Ithaca: Cornell.

Deutscher, G., Zallen, R., Adler, J. (1983), *Percolation Structures and Processes,* Annals of the Israel Physical Society, Vol. 5, Bristol: Adam Hilger.

McCoy, B. M., Wu, T. T. (1973), *The Two-Dimensional Ising Model,* Cambridge, MA: Harvard.

Privman, V. (Ed.) (1990), *Finite Size Scaling and Numerical Simulation of Statistical Systems,* Singapore: World Scientific.

Privman, V. (Ed.) (1997), *Nonequilibrium Statistical Mechanics in One Dimension,* Cambridge, U.K.: Cambridge Univ. Press.

UNIVERSE, STRUCTURE OF THE

HANS J. HAUBOLD, *UN Outer Space Office, Vienna International Centre, Vienna, Austria*

A. M. MATHAI, *Department of Mathematics and Statistics, McGill University, Montréal, Québec, Canada*

Introduction 47
1. **Structure of the Ancient Universe** 48
1.1 Prehistoric Astronomy 48
1.2 Egypt 48
1.3 Mesopotamia 49
1.4 India and Persia 49
1.5 The Puranas and the Brahmapaksa 49
2. **Macrocosmos: From Anaximander to Einstein** 50
2.1 Greek and Roman World 50
2.2 Aristotle's Universe 50
2.3 Copernicus's Universe 52
2.4 Newton's Universe 52
2.5 Einstein's Universe 53
2.6 Big Bang 54
2.7 Beyond the Big Bang 54
3. **Microcosmos: From Leucippus to Yukawa** 55
3.1 Atomism 55
3.2 Atom 56
3.3 Electron 57
3.4 Atomic Nucleus and Photon 58
3.5 Neutron 59
3.6 Fundamental Particles 60
4. **Weak, Electromagnetic, and Strong Interactions: Structure of the Microcosmos** 60
4.1 Fundamental Interactions 60
4.1.1 Gravitational Interaction 60
4.1.2 Electromagnetic Interaction 61
4.1.3 Weak Interaction 61
4.1.4 Strong Interaction 62
4.2 Classification of Fundamental Particles 64
4.3 How Fundamental Particles Interact 64
4.3.1 Graviton 64
4.3.2 Photon 65
4.3.3 Intermediate Bosons 65
4.3.4 Gluons 65
5. **Gravitational Interaction: Structure of the Macrocosmos** 65
5.1 Structural Elements of the Universe 65
5.1.1 The Planck Scales 65
5.1.2 The Universe 66
5.1.3 Galaxies 66
5.1.4 Stars 67
5.1.5 White Dwarfs and Neutron Stars 68
5.1.6 Black Holes 68
5.2 Evolution of the Universe 69
5.3 Large-Scale Distribution of Matter 70
5.4 Morphology of Galaxies 72
5.5 Quasars 73
5.6 Stellar Evolution 74
6. **Mathematics and Physics and the Structure of the Universe** 76
6.1 General Relativity and Quantum Field Theory 76
6.2 Macroscopic Dimension of Space 76
6.3 Microscopic Dimension of Space 78
Glossary 78
Works Cited 80
Further Reading 81

INTRODUCTION

Every culture, from the ancient tale spinners of the Indus Valley to the modern technocrats of the Silicon Valley, has held its own unique view of the cosmos. Astronomy, the oldest science mainly based on observation, cannot be separated from physical law and mathematics. Big advances in space science and technology in this century have al-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

lowed astronomers to look both deeper into space and farther back in time, thereby discovering a close connection between particle physics and cosmology. The study of the creation, evolution, and structure of the universe has become a legitimate subject for astronomers, physicists, and mathematicians.

This article presents an outline of the current knowledge about the structure—more precisely the structural elements—of the universe from the point of view of the standard "big-bang" model (dubbed macrocosmos) and the standard model of fundamental particles (dubbed microcosmos). In this regard an important point has to be made that the article does not consider the time-dependent phenomena that ought to characterize the evolution of the universe but is focused on the structural elements of the universe as discovered by observation at the present stage of the evolution of the universe. One can ask the question, How is it possible to understand the universe on the basis of its structural elements? The ancient atomic philosophers may have answered: This objection is similar to that of someone saying, How can the entire poem of Homer, the *Iliad*, be composed of only 24 letters, those 24 letters of the Greek alphabet? Hidden in that thought is that the richness of the poem is not compromised by the fact that a method exists to trace it back ultimately to a few, very elementary and simple elements, such as the formal contents of the Greek language—to the 24 characters that are being referred to as the Greek alphabet. However, the degree of the current understanding of the structural elements of the universe (compared with the 24 letters of the Greek alphabet) reveals the fact that the scientific establishment is still far from understanding the beauty of the structure and evolution of the universe as a whole (compared with the *Iliad*).

The article is divided into six parts. The first three parts recount the development of macrophysical and microphysical knowledge from prehistory to the present. Part four is devoted to the elements of the standard model of fundamental interactions, namely, the leptons, quarks, and gauge bosons. The fifth part discusses the structural elements of the universe, from the large-scale structure down to the stars. Finally, the sixth part emphasizes the importance of mathematics for the physics and astronomy of the structure of the universe.

1. STRUCTURE OF THE ANCIENT UNIVERSE

1.1 Prehistoric Astronomy

From half a millennium before 4000 B.C., long barrows—elongated burial mounds of earth—were arranged so as to be aligned on the rising and setting of bright stars. In the middle of the third millennium, this technique was transferred to circular grave architecture, the so-called round barrows, and in this form it lasted for well over a thousand years. Why this particular phase in prehistorical astronomical activity is so important is that it testifies to the marriage of astronomy with geometry. There is clear evidence that the Sun and the Moon were observed. Stonehenge, shown in Plate 1, is the most famous prehistoric megalith (standing-stone monument) in Europe and lies north of Salisbury, England. Stonehenge had an exceptionally long history of use as a religious center and is also believed to have functioned as an astronomical observatory.

1.2 Egypt

Some concern with astronomy had already been shown in a cosmology associated with the rulers Seti I (r.c. 1318–c. 1304 B.C.) and Ramses IV (r.c. 1166–c. 1160 B.C.). The Egyptians had by then long been adept at measuring time and designing calendars, using simple astronomical techniques. They too aligned their buildings on the heavens. Some early Egyptian sources speak of a cult relating the Sun god Re (represented in art with a man's body and a falcon's head surmounted by a solar disk) and an earlier creator god Atum. The cult of Re-Atum was well established by the time of the first great pyramids, that is, about 2800 B.C. At first this was centered mainly on a temple to the north of the old Egyptian capital of Memphis. The place was known by the Greeks as Heliopolis, "City of the Sun," but by the Egyptians as On (located northeast of Cairo, Egypt). By historical times, the priests of Heliopolis had laid down a cosmogony that held Re-Atum to have generated himself out of Nun, the primordial ocean. His offspring were the gods of air and moisture, and only after them, and as their offspring, were Geb, the Earth god, and Nut, the sky goddess, created. The deities of Heliopolis (the Great Ennead) were made up with Osiris, Seth, Isis, and Nephthys, the offspring of Geb and Nut.

1.3 Mesopotamia

In the dynasty of Hammurabi (r.c. 1792–c. 1750 B.C.) in Babylonia, not all gods can be identified with stars. The three highest—Anu, Enlil, and Ea—corresponded to the heavens, Earth, and water.

1.4 India and Persia

The oldest of the Vedic writings in Hinduism, the Rigveda (dating from between 1500 and 500 B.C.), gives more than one account of the creation of the world (Fig. 1). The main version is that the world was made by the gods, as a building of wood, with heaven and Earth somehow supported by posts. Later it is suggested that the world was created from the body of a primeval giant. This last idea gave rise to the principle, found in later Vedic literature, that the world is inhabited by a world soul. Various other cosmogonies followed, with the creation of the ocean sometimes being given precedence, and place being made for the creation of the Sun and the Moon. There is a certain circularity in it all, however, since heaven and Earth are generally regarded as the parents of the gods in general, and water was sometimes introduced into the parentage. The Vedic literature gives no clear indication that mathematical techniques for describing the motions of heavenly bodies were discussed in India before the 5th century B.C.

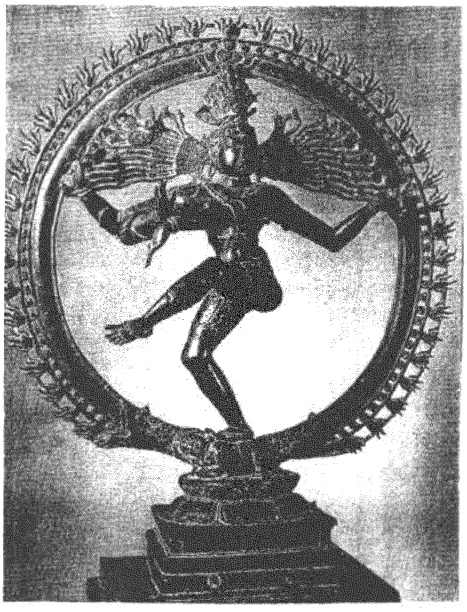

FIG. 1. Shiva is one of the two principal Hindu gods (the other being Vishnu), his most celebrated appearance being the one as Nataraja, the King of Dancers. Indian artists have represented Shiva's cosmic dance in magnificent bronze sculptures of dancing figures with four arms whose superbly balanced and yet dynamic gestures express the rhythm and unity of life. The upper right hand of the god holds a drum to symbolize the primeval sound of creation, the upper left bears a tongue of flame, the element of destruction. The balance of the two hands represents the dynamic balance of creation and destruction in the universe, accentuated further by the Dancer's calm and detached face in the center of the two hands, in which the polarity of creation and destruction is dissolved and transcended. The second right hand is raised in the sign of "do not fear," symbolizing maintenance, protection, and peace, while the remaining left hand points down to the uplifted foot, which symbolizes release from the spell of maya. The god is pictured as dancing on the body of a demon, the symbol of human ignorance that has to be conquered before liberation can be attained.

1.5 The Puranas and the Brahmapaksa

Although influenced by texts going back to Vedic times, and by Iranian sources too, the Puranas are Sanskrit writings about primordial times with cosmological sections dating from the 4th to 16th centuries A.D. of the Christian era. The Puranas deal with five topics: the creation of the Universe; the destruction and re-creation of the Universe, including the history of humankind; the genealogy of the gods; the reigns of the Manur; and the history of the lunar and solar dynasties. The Earth is now represented as a flat circular disc, with a world-mountain, Meru, at its center. The Meru is anchored to the yasti, which symbolizes the axis of the universe. The mountain is surrounded by alternating rings of sea and land, so that there are seven continents and seven seas. Wheels are conceived to carry the celestial bodies, these turned around the star at the north pole by Brahma, using cords made of mind. This cosmology was taken over by Jainism, a monastic religion that, like Buddhism, denies the authority of the Veda, but from the 5th and 6th centuries onward, it was undermined by the influx of a new form of Greek cosmology, with pre-Ptolemaic roots. In short, Aristotelianism reached India (see North, 1995).

2. MACROCOSMOS: FROM ANAXIMANDER TO EINSTEIN

2.1 Greek and Roman World

Anaximander of Miletus (c. 610–545 B.C.) is said to have made a map of the inhabited world, and to have invented a cosmology that could explain the physical state of the Earth and its inhabitants. The infinite universe was said to be the source of an infinity of worlds, of which ours was but one, that separated off and gathered their parts together by their rotary motion. Masses of fire and air were supportedly sent outward and became the stars. The Earth was some sort of floating circular disc, and the Sun and the Moon were ring-shaped bodies, surrounded by air. The Sun acted on water to produce animate beings, and people were descended from fish.

Anaximenes (c. 545 B.C.) elaborates on Anaximander's ideas and argues that air is the primeval infinite substance, from which bodies are produced by condensation and rarefaction; he produces logical arguments based on everyday experience. Again he introduces rotary motions as the key to understanding how the heavenly bodies may be formed out of air and water.

Pythagoras of Samos (c. 560–c. 480 B.C.) took the cosmic ideas of Anaximander and Anaximenes one stage further, saying that the universe was produced by the heaven inhaling the infinite so as to form groups of numbers. The Pythagoreans proposed a geometrical model of the universe, involving a central fire around which the celestial bodies move in circles. The central fire was not the Sun, although the Earth was certainly of the character of a planet to it. To account for lunar eclipses, the Pythagoreans postulated a counter Earth.

The discovery that the Earth is a sphere was traditionally assigned to Parmenides of Elea (c. 515–c. 450 B.C.), who was also said to have discovered that the Moon is illuminated by the Sun. A generation later, Empedocles (c. 484–c. 424 B.C.) and Anaxagoras (c. 500 B.C.) seem to have given a correct qualitative account of the reason for solar eclipses, namely, the obscuration of the Sun's face by the intervening Moon.

The discovery of the sphericity of the Earth, and of the advantages of describing the heavens as spherical, captured the imagination of the Greeks of the time of Plato (c. 428–c. 347 B.C.) and Aristotle (384–322 B.C.), in the 4th century, and of one man in particular: Eudoxus of Cnidus (c. 400–c. 355 B.C.), who produced a very remarkable planetary theory based entirely on spherical motions.

2.2 Aristotle's Universe

At the time of the Greek philosopher and naturalist Aristotle (384–322 B.C.), the Earth and the universe were seen as constructed out of five basic elements: earth, water, air, fire, and ether. The natural place of the motionless Earth was at the center of that universe. The stars in the heavens were made up of an indestructible substance called ether (aether) and were considered as eternal and unchanging. Aristotle's cosmology was the first "steady-state" universe. The other basic elements—water, air, and fire—were earthly elements. The celestial bodies, including the Sun, the planets, and the stars, were considered to be attached to rigid, crystalline spheres, which were supposed to revolve in perfect circles about the Earth. The three innermost spheres, closest to the Earth, contained water, air, and fire, respectively. The Egyptian–Greek astronomer Ptolemy (c. 100–170 A.D.) left the Earth to be located at the center of the universe but ascribed to the Sun and the planets a new place within the cosmological views of his time. The Sun and the planets would revolve in small circles whose centers revolve in large circles about the Earth ("epicycles"). The element of perfection and beauty ascribed to the divine heavens remained the circle out of which the orbits of heavenly bodies were composed. Basically, Ptolemaic ideas were devised to accommodate astronomical observations of planetary motions.

His "Almagest" described the motions of the heavens on this geocentric basis, employing wheels within wheels (epicycles) to obtain a plausible match with observations. Ptolemy's geocentric system dominated western thought on astronomy (including cosmology) until the time of Copernicus, fourteen centuries later (Fig. 2).

Eratosthenes (c. 270–c. 190 B.C.) was the first to measure accurately the radius of the Earth, in 196 B.C., by determining the minimum angle between the Sun's direction and the vertical at Alexandria on the day of the summer solstice. He knew that a zero angle occurs approximately when the Sun was at its highest point at the city of Syene (now

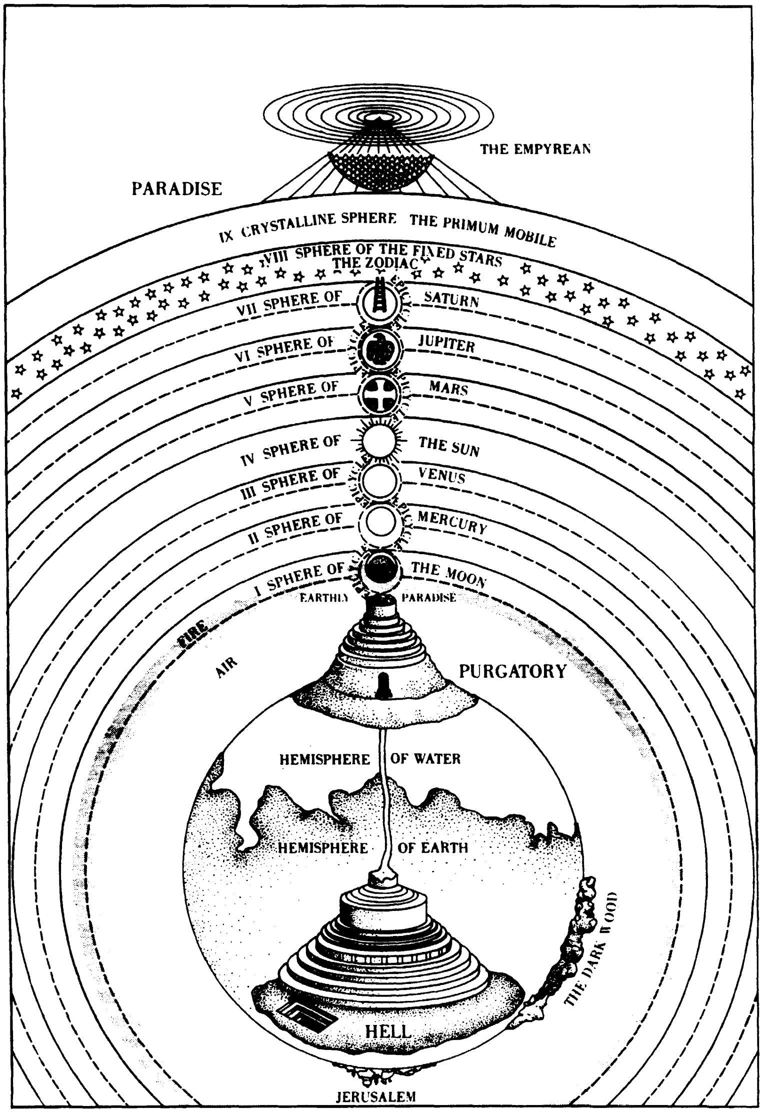
THE EMPYREAN
PARADISE
IX CRYSTALLINE SPHERE THE PRIMUM MOBILE
VIII SPHERE OF THE FIXED STARS
THE ZODIAC
VII SPHERE OF SATURN
VI SPHERE OF JUPITER
V SPHERE OF MARS
IV SPHERE OF THE SUN
III SPHERE OF VENUS
II SPHERE OF MERCURY
I SPHERE OF THE MOON
EARTHLY PARADISE
FIRE
AIR
PURGATORY
HEMISPHERE OF WATER
HEMISPHERE OF EARTH
THE DARK WOOD
HELL
JERUSALEM

FIG. 2. Dante Alighieri's (1265–1321) scheme of the universe in illustration from "Paradise" in "The Divine Comedy" extends Aristotelian cosmology in a modern way. Dante traverses the material world from the icy core of Earth, the abode of Lucifer, to the Mount of Purgatory. He continues through the *nine heavenly spheres*, each sphere larger and more rapidly turning than the last, until he reaches the Primum Mobile, the ninth and largest sphere and the boundary of space. His goal was to see the Empyrean, the abode of God.

Aswan), and he knew the base of the triangle, i.e., the distance from Alexandria to Syene.

2.3 Copernicus's Universe

Copernicus (1473–1543), a Polish astronomer, placed the Sun at the center of the solar system with the Earth orbiting around the Sun, thus proposing a heliocentric cosmology (*De revolutionibus orbium coelestium*, 1543). This entirely new basis of cosmological consideration did not fit the observations much better than the Aristotelian–Ptolemaic system but was justified by the "divine" principle of simplicity in comparison to the rather complicated construction using epicycles as employed in Ptolemy's cosmology. However, like Aristotle and Ptolemy, Copernicus retained the conventional idea that the planets moved in perfectly circular orbits and continued to believe that the stars were fixed and unchanging. Kepler (1571–1630), a German astronomer and physicist, adopted the Copernican system but introduced the concept of planets with elliptical orbits. He still believed that the Sun was the center of the universe. The cosmologies of Aristotle and Ptolemy had nevertheless been abandoned.

2.4 Newtons' Universe

The Italian philosopher Bruno (1548–1600) laid the groundwork for Newtonian cosmology by emphasizing that the universe is infinite and stars are scattered outward through infinite space (see, however, Fig. 3).

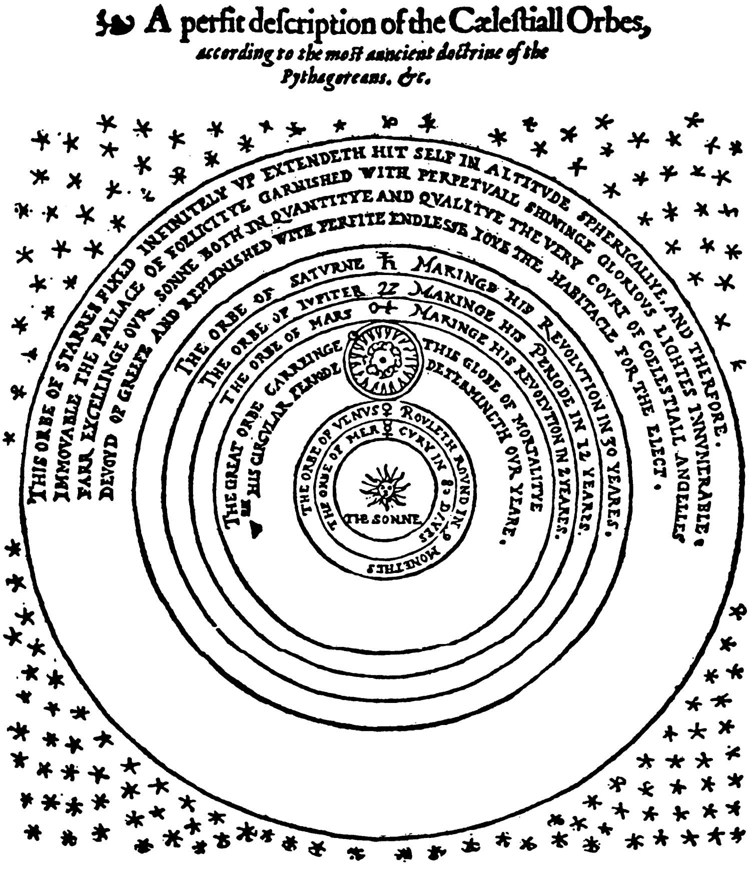
A perfit deſcription of the Cæleſtiall Orbes, according to the moſt auncient doctrine of the Pythagoreans, &c.
This orbe of starres fixed infinitely up extendeth hit self in altitude sphericallye, and therfore immovable the pallace of foelicitye garnished with perpetuall shininge glorious lightes innumerable, farr excellinge our sonne both in quantitye and qualitye the very court of coelestiall angelles. Devoyd of greefe and replenished with perfite endlesse ioye the habitacle for the elect.
The orbe of Saturne ♄ makinge his revolution in 30 yeares.
The orbe of Jupiter ♃ makinge his periode in 12 yeares.
The orbe of Mars ♂
The great orbe carryinge this globe of mortalitye
The orbe of Venus ♀ rouleth round in 9 monethes.
The orbe of Mercury ☿
The Sonne

FIG. 3. The first diagram to illustrate the proposal that the universe is infinite. From the edition by Thomas Digges of his father's *A Prognostication everlastinge* . . . , published in 1576 in London, eight years before its publication by Giordano Bruno to whom the idea is often credited. (By permission of The Royal Society.)

Bruno even went so far as to say that stars are Suns, perhaps with orbiting planets and life on them (*De l'infinito, universo e mondi,* 1584). This far-reaching statement signaled a transfer of attention away from the planets in the solar system to the stars in the Milky Way. The laws describing the behavior of planets are the same laws that describe the behavior of the stars. That the physical laws are of universal nature, and can be applied on Earth as well as in the heavens, was discussed on a philosophical basis by the French philosopher and mathematician Descartes (1596–1650). He compared the universe with a giant clock obeying mechanical laws, a concept that later on had a major influence on Newton's (1643–1727) thinking. It was his belief that the universe was infinite and that the primary qualities of the universe were mathematical in nature.

A major step in observational techniques was achieved through the invention of the reflecting telescope between, say, 1545 and 1559 by the Briton Leonhard Digges (c. 1520–1559). But it was not until 1609 that the Italian astronomer and physicist Galileo (1564–1642) realized observationally that the Milky Way is actually a collection of individual stars. Galileo also observed mountains on the Moon and discovered four satellites around Jupiter. He had taken the first step toward deducing the structure of the Milky Way. The discovery of heavenly bodies that evidently did not circle the Earth and his support for the Copernican heliocentric cosmology were described in his work *Dialogo sopra i due massimi sistemi del mondo, Tolemaico e Copernicano* (1632).

Cassini (1625–1712), an Italian–French astronomer, used the telescope to make accurate measurements of the "dimensions of the universe," particularly determining the distance of the Earth from the planet Mars. A rigorous mathematical foundation of Descartes's notion of the universe as a giant mechanical clock was provided by Newton's theory of gravity and his laws of motion (*Philosophiae naturalis principia mathematica,* 1687). From these he explained Galileo's results on falling bodies, Kepler's three laws of planetary motion, and the motion of the Moon, Earth, and tides. Newton clearly realized that gravity is the dominant force for understanding the structure of the universe; however, he argued that the universe must be static in a famous letter he sent to the theologian Richard Bentley (1662–1742) in 1692. Mainly for religious reasons, at this time, constancy and stability were associated with the perfection of God and change with friction and decay. Philosophical and religious ideas served as the scaffolding upon which scientific systems of thought developed. It was Bentley who derived for the first time, on the basis of Newton's gravitational theory, what is still considered to be one of the fundamental constants of nature, the gravitational constant.

In Kant's (1724–1804) cosmology, the gravitational attraction of stars for each other was exactly balanced by the orbital motions of the stars and he argued that forces can act at a distance without the necessity for a transmitting medium. In 1755 Kant proposed the nebular hypothesis for the formation of the solar system. In 1788 Laplace (1749–1827) attempted a mathematical proof of the stability of the solar system (*Système du monde,* 1796). Stability and order of the universe were considered as eternal principles in the heavens and on Earth. The cosmology of Copernicus, as refined by Kepler, is now believed to be essentially correct.

2.5 Einstein's Universe

In 1915 Einstein (1879–1955) put forth his general relativity theory (a new theory of gravity); he applied this theory to cosmology in 1917. The theory describes gravity as a distortion of the geometry of space and time. Unlike Newton's theory of gravity, general relativity was consistent with special relativity, which Einstein had introduced in 1905. Cosmology, based on general relativity, broadened the problem into one of finding a model of the space-time structure of the universe. Einstein's original solutions of his gravitational field equations left the universe in a stable state of static equilibrium, and he provided physical conditions required to maintain such static (time-independent) equilibrium (the "cosmological constant"). Only in 1922 did Friedmann (1888–1925) succeed in finding solutions of Einstein's field equations that evolved in time, describing an expanding (or contracting, if one cares to reverse the sense of time) universe.

Persuasive observational evidence that the universe is indeed expanding and changing in time was found by Hubble (1889–1953) in 1929 while employing the technique known as Doppler shift for measuring the redshift of colors in the spectrum of nebulae. "Modern cosmology" may be said to have begun with Einstein (1917) and Friedmann (1922, 1924), based on observations of cosmological relevance made by Slipher (1875–1969) in 1914 and Hubble in 1929. Slipher discovered the redshift of nebulae, later on identified by Hubble to be entire galaxies similar to the Milky Way; however, until 1929 their cosmological significance remained obscure. In 1929 Hubble had counted a great number of galaxies (to determine their distribution throughout the observable universe) and plotted the galaxies' redshifts against magnitudes for the brightest E-type cluster galaxies (Hubble diagram). Hubble found evidence that the outward speed of a galaxy is directly proportional to its distance away from the observer (Hubble law). This observational fact was exactly what would be expected if the universe is expanding, as discussed in an article of Lemaître (1894–1966) in 1927. Hubble's diagram reveals a linear increase of the magnitude of galaxies with increasing redshift. In 1956, Hoyle and Sandage developed the q_0 criterion that could be used to distinguish one cosmological model from others. Lemaître's model and Hoyle's steady-state model were ruled out by predicting $q_0 = -1$, whereas Hubble's linearity gave $q_0 = +1$. Cosmology focused on the search for two numbers: H_0, the rate of expansion of the universe at the position of the Milky Way; and q_0, the deceleration parameter, characterizing the change of the rate of universal expansion. The value of q_0 is believed to lie in the range between 0 and 0.5; the value of H_0 is thought to be uncertain by a factor of about 2 ($H_0 = 100h$ km s^{-1} Mpc^{-1}, where $0.5 \leq h \leq 1$). Theoretical ideas on which big-bang cosmology is based were contained in the publications of Friedmann and Lemaître by 1930 (for an in-depth review, see North, 1965).

2.6 Big Bang

Astrophysical arguments were introduced into the cosmological models with Gamow's (1946) prediction that helium (and possibly heavier elements) was generated at an early stage of the evolution of the universe. At this time it was believed that the relative abundance of cosmic nuclei represents truly cosmic abundances. On the basis of Gamow's arguments, the cosmological criterion of the origin of chemical elements in the primeval fireball was substantiated by Alpher and Herman (1949, 1950) and Alpher *et al.* (1953). However, after Burbidge *et al.* (1957) it became evident that the bulk of the chemical elements beyond helium were not synthesized at the early stages of the expansion of the universe but in the stars. Only the synthesis of helium (Hoyle and Tayler, 1964), which is not produced in sufficient quantities in stars, and deuterium (Peebles, 1966), which is destroyed during galactic evolution, continued to need Gamow's primordial nucleosynthesis to arrive at reasonable abundances as observed in the universe. Later on Wagoner *et al.* (1967) were able to show that in addition to D, ^{3}He, and ^{4}He, the only other cosmologically significant element was ^{7}Li. In 1965 the microwave background radiation of 3 K was discovered by Penzias and Wilson (1965), as predicted by Alpher and Herman in 1949 on the basis of Gamow's considerations of a hot and dense origin of the universe.

Today the big-bang cosmological model, the temperature of the microwave background radiation, the abundances of D, ^{3}He, ^{4}He, and ^{7}Li, and the astrophysically observed average density of galactic material are starting points for developments in cosmology.

2.7 Beyond the Big Bang

In the big-bang cosmology, the universe has been expanding throughout its history. Mathematical calculations would suggest that the temperature and the density were infinite at the instant of the big bang. The universe is supposed to originate in the big bang, but the mathematical and physical structure of the model does not permit matter to originate. Additionally, the big-bang model leaves unanswered several important questions regarding

1. the number of protons and neutrons in the universe, relative to the number of photons;

2. the large-scale homogeneity of the observed universe;
3. the actual density of the universe, which is close to the critical density (that average density of the universe dividing big-bang models that will keep on expanding forever from those that will ultimately recontract); and
4. the origin of density perturbations from which small-scale (stars and systems of them) and large-scale (galaxies and systems of them) inhomogeneities have been developed (Kolb and Turner, 1990).

A new approach to explaining some of the questions left over by big-bang cosmology began with Guth's (1981) paper inventing an "inflationary" period in the evolution of the universe. The phase transition associated with the breakup of the unified force in the Grand Unification epoch could leave the universe in a state of false vacuum, in which the vacuum has a very high energy density. This vacuum energy density acts like a cosmic repulsion, and the universe embarks on an exponential expansion that "inflates" the universe by a factor of 10^{30} in a brief instant of time. This inflationary period ends when the vacuum energy density transforms into matter and radiation, and the expansion of the universe continues. Besides offering explanation of the four questions above, it makes one concrete prediction: The present universal expansion should exhibit flatness to the extent of $\Omega_0 = 1 \pm \epsilon$ with $\epsilon < 10^{-6}$, if inflation was sufficient to extend to a distance $\geq cH_0^{-1}$ today. Observations of the luminous matter content of the universe reveal only $\Omega_0 < 0.002$, and the quantity of dark or nonbaryonic material necessary to explain the flat rotation curves of spiral galaxies and the viral equilibrium of large groups and clusters of galaxies requires at most $\Omega_0 \approx 0.2$. The only way that $\Omega_0 = 1$ from inflationary cosmological models can be reconciled with observation is by the existence of nonbaryonic weakly interacting particles. The unknown nature of dark matter has led to considerable activities in fundamental particle physics to speculate about new basic constituents of matter. Additionally, the inflationary universe picture has stimulated far-reaching speculation about the ultimate origin of the universe.

Cosmology based on general relativity has thus moved close to fundamental particle physics based on quantum field theory. A new challenge emerged for physical theory and observation subsumed under the term "quantum gravity." Interaction between mathematics and physics, as exemplified by the role of Riemannian geometry in general relativity and functional analysis in quantum mechanics, became lively again. Cosmology is currently based on the two fundamental theories in twentieth-century physics, general relativity and quantum field theory. General relativity describes the gravitational force on an astronomical scale, while quantum field theory describes the weak, electromagnetic, and strong interaction of fundamental particles. A formal quantization of general relativity leads to unphysical infinities; encompassing gravity in quantum field theories leads to a new connection of theoretical physics to modern mathematics, called string theory. However, there is still no quantum cosmology. Particularly, an inflationary era may have taken place in the early universe, but there is no proof that it did so (Manin, 1981; Schmid, 1992).

3. MICROCOSMOS: FROM LEUCIPPUS TO YUKAWA

3.1 Atomism

With regard to the structure of the universe, the belief persists that certain objects are fundamental and that others are derived, in the sense that the latter are composed of the former. In one version of this distinction, the fundamental objects are particles, or points of matter. Although the view of which objects qualify as fundamental particles has changed many times, the notion that the universe is ultimately made of such material points, moving through space, has endured in some form ever since the theory of atomism was first proposed by the Greeks Leucippus (5th century B.C.) and Democritus (c. 460–c. 370 B.C.) in the 5th century B.C. (Whittaker, 1951, 1953/1989).

An atom is the smallest unit of matter that is recognizable as a chemical element. Atoms of different elements may also combine into systems called *molecules*, which are the smallest units of chemical compounds (Fig. 4). In all these processes, atoms

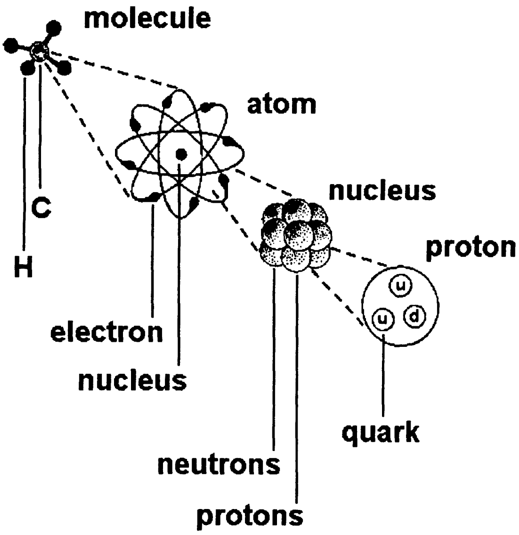

FIG. 4. The continuum of size and organizational level appears throughout the range of structural elements, from the fundamental particles to the highest-level systems of matter. Particles such as the quarks are known to be bound by strong forces. Protons and neutrons are bound within the atomic nucleus. The outer shell of atoms is bound to the nucleus by electromagnetic forces. A molecule may be thought of either as a structure built of atoms bound together by chemical forces or as a structure in which two or more nuclei are maintained in some definite geometrical configuration by attractive forces from a surrounding cloud of negative electrons. The evolution of the universe created a hierarchy of structural elements.

may be considered as the ancient Greeks imagined them to be: the ultimate building blocks of matter. When stronger forces are applied to atoms, however, the atoms may break up into smaller parts. Thus atoms are actually composites and not units and have a complex inner structure of their own.

The first recorded speculations that matter consisted of atoms are found in the works of Leucippus and Democritus. The essence of their views is that all phenomena are to be understood in terms of the motions, through empty space, of a large number of tiny and indivisible bodies. The name "atom" comes from the Greek word "atomos," for "indivisible." According to Democritus, these bodies differ from one another in shape and size, and the observed variety of substances derives from these differences in the atoms composing them.

Greek atomic theory was not an attempt to account for specific details of physical phenomena. It was instead a philosophical response to the question of how change can occur in nature. Little effort was made to make atomic theory quantitative—that is, to develop it as a physical theory for the study of matter. Greek atomism, however, did introduce the valuable concept that the nature of everyday things was to be understood in terms of an invisible substructure of objects with unfamiliar properties. Democritus stated this especially clearly in one of the few sayings of his that has been preserved: "color exists by convention, sweet by convention, bitter by convention, in reality nothing exists but atoms and the void."

Although adopted and extended by such later ancient thinkers as Epicurus (341–270 B.C.) and Lucretius (c. 95–55 B.C.), Greek atomic theory had strong competition from other views of the nature of matter. One such view was the four-element theory of Empedocles. These alternative views, championed by Aristotle among others, were also motivated more by a desire to answer philosophical questions than by a wish to explain scientific phenomena.

3.2 Atom

The atomic theory languished until the 18th and early 19th centuries, when physicists and chemists revived it to explain the properties of gases and some of the facts of chemistry. In these theories the fundamental particles, the atoms, remained indivisible points. The discovery in the late 19th and the 20th centuries that atoms were composite, rather than indivisible, set the stage for modern discoveries about fundamental particles (Pais, 1986; Whittaker, 1951, 1953/1989).

When interest in science revived in Europe in the 16th and 17th centuries, enough was known about Greek atomism to form the basis for further thought. Among those who revived the atomic theory were Pierre Gassendi (1592–1655), Robert Boyle (1627–1691), and especially Isaac Newton. The latter part of Newton's book *Optics* is a series of detailed speculations on the atomic nature of matter and light, indicating how some of matter's properties are to be understood in terms of atoms.

In the 19th century, two independent

lines of reasoning strengthened the belief in the atomic theory. Both approaches also began to reveal some quantitative properties of atoms. One approach, pioneered by John Dalton (1766–1844), involved chemical phenomena. The other, involving the behavior of gases, was carried out by physicists such as Rudolf Clausius (1822–1888) and James Clerk Maxwell (1831–1879).

Dalton's main step forward was his introduction of atomic weights. Dalton studied the elements then known and analyzed the data of their reactions with one another. He discovered the law of multiple proportions, which states that when several distinct reactions take place among the same elements, the quantities that enter the reactions are always in the proportions of simple integers—that is, 1 to 1, 2 to 1, 2 to 3, and so on. From this came the concept that such reacting quantities contain equal numbers of atoms and are therefore proportional to the masses of individual atoms. Dalton gave the lightest known element, hydrogen, an atomic weight of 1, and developed comparative atomic weights for the other known elements accordingly.

The study of gases in terms of atomic theory was begun by Daniel Bernoulli (1700–1782) in the 18th century. Bernoulli showed that the pressure exerted by a gas came about as the result of collisions of the atoms of the gas with the walls of its container. In 1811, Amedeo Avogadro (1776–1856) suggested that equal volumes of different gases, under the same conditions of pressure and temperature, contain equal numbers of atoms. Avogadro himself never estimated the magnitude of this value, although it is now known as the Avogadro number.

3.3 Electron

The history of particle physics has gone through four stages. In the first stage, Joseph J. Thomson (1856–1940) discovered (1897), by studying electricity passing through gases, that all atoms contain certain particles, called electrons, that carry a negative electric charge. Because atoms are electrically neutral, there must be a balancing of positive charges somewhere in the atom. Ernest Rutherford (1871–1937) proposed (1911), on the basis of a series of experiments by Hans Geiger (1882–1945) and Ernest Marsden (1889–1970), that these positive charges are concentrated in a very small volume, called the atomic nucleus, at the center of the atom.

By the end of the 19th century almost all scientists had become convinced of the truth of the atomic theory. By that time, however, evidence was just beginning to accumulate that atoms are not in fact the indivisible particles suggested by their name. One source of such evidence came from studies using gas discharge tubes. In such tubes, a gas at low pressure is subjected to intense electrical forces. Under these conditions various colored glows are observed to traverse the tube. One blue glow at one end of the tube, around the electrode known as the cathode, was observed for a wide variety of gases. The glow was shown by Joseph Thomson in 1897 to involve a stream of negatively charged particles with a charge-to-mass ratio indicating the existence of a particle with a very small mass on the atomic scale. These particles were called electrons, and they were soon recognized to be a constituent of all atoms. That is, atoms are not indivisible but contain parts.

In the late 19th and the early 20th century it was also found that some kinds of atoms are not stable. Instead they transform spontaneously into other kinds of atoms. For example, uranium atoms slowly change into lighter thorium atoms, which themselves change into still lighter atoms, eventually ending up as stable atoms of lead. These transformations, first observed by Antoine Henri Becquerel (1852–1908), came to be known as radioactivity, because the atomic changes were accompanied by the emission of several types of radiation.

Atoms are ordinarily electrically neutral. Therefore the negative charge of the electrons in an atom must be balanced by a corresponding positive charge. Because the electrons have so little mass, the positive constituents of an atom must also carry most of the atom's mass. The obvious question arose as to how these varied parts are arranged within an atom. The question was answered in 1911 through the work of Ernest Rutherford and his collaborators. In their experiments they passed alpha particles—a type of radiation emitted in some radioactive decays—through thin gold foils. They observed that in some instances the alpha particles emerged in the opposite direction from their initial path. This suggested a collision with a heavy object within the atoms of the gold. Because electrons are not massive enough to produce such large deflections, the positive charges must be involved. Analyzing the data, Rutherford showed that the positive charge in an atom must be concentrated in a very small volume with a radius less than 10^{-12} cm, or 1/10 000 the size of the whole atom. This part of the atom was called the nucleus.

3.4 Atomic Nucleus and Photon

Rutherford proposed an atomic model in which the atom was held together by electrical attraction between the nucleus and the electrons. In this model the electrons traveled in relatively distant orbits around the nucleus. The model eventually proved successful in explaining most of the phenomena of chemistry and everyday physics. Subsequent studies of the atom divided into investigations of the electronic parts of the atom, which came to be known as atomic physics, and investigations of the nucleus itself, which came to be known as nuclear physics. This division was natural, because of the immense difference in size between the nucleus and the electron orbits and the much greater energy needed to produce nuclear as compared to electronic changes.

The Rutherford model of the atom, however, had to face two immediate problems. One was to account for the fact that different atoms of the same element behaved in physically and chemically similar ways. According to the Rutherford model, electrons could move in any of the infinite number of orbits allowed by Newtonian physics. If that were so, different atoms of the same element could behave quite differently. This is actually a problem for any atomic model based on Newtonian physics, and it had already been recognized by Maxwell in 1870. The other problem was that, according to the principles of electromagnetism, electrons should continuously emit radiation as they orbit in an atom. This would cause the electrons to lose energy and to spiral into the nucleus.

An important step toward solving these problems was taken by Niels Bohr (1885–1962) in 1913. According to Bohr, the electrons in atoms cannot exist in arbitrary orbits. Instead they are found only in certain "states." The states in which they can exist are those in which the angular momentum of their orbits is an integer multiple of $h/2\pi$, where h is a quantity known as Planck's constant. This constant had been introduced by Max Planck (1858–1947) in his theory describing blackbody radiation.

According to the Bohr model of the atom, there is a so-called ground state for any atom. This ground state has the lowest energy allowed to the atom, and it is the same for all atoms containing the same number of electrons. An atom normally exists in this ground state, which determines the observed properties of a given element. Furthermore, according to Bohr, no radiation is emitted by an atom in its ground state. This is because energy must be conserved in the radiation process, and no available state of lower energy exists for the atom to balance any energy lost through radiation.

An atom can be removed from its ground state only when enough energy is given to it, by radiation or collisions, to raise an electron to an "excited" state. When the atom is excited, it will usually emit electromagnetic radiation rapidly and return to the ground state. The radiation is emitted in the form of individual packets, or quanta, of light, called photons. Each photon has an energy equal to the difference between the energy of the excited state and the ground state of the atom. According to a formula developed by Planck and Einstein, this energy corresponds to a specific wavelength of the emitted light. Using his assumption about the allowed angular momenta for electrons, Bohr was able to calculate the precise wavelengths in the spectrum of the simplest atom, hydrogen.

In 1869, Dimitri I. Mendeléev (1834–1907) stated the rule that chemical elements arranged according to the value of their atomic weights exhibit a clear periodicity of properties. Eventually, Bohr was able to extend his atomic theory to describe, qualitatively, the chemical properties of all the elements. Each electron in an atom is assigned a set of four so-called quantum numbers. These numbers correspond to the properties of energy, total orbital angular momentum, projection of orbital angular momentum, and projection of spin angular momentum. It is also assumed—as had first been suggested by Wolfgang Pauli (1900–1958) in 1924—that no two electrons in an atom can have the same values for all four quantum numbers. This came to be known as Pauli's exclusion principle. This principle influences the way in which the chemical properties of an element depend on its atomic number, that is, the number of electrons in each atom of the element. A maximum number of electrons can occur for each energy level, and no more than that. For example, the lowest energy level of an atom—the one in which the electrons have zero orbital angular momentum—can contain up to two electrons. The one electron in a hydrogen atom exists at this energy level, as do the two electrons in a helium atom. For the next heavier atom, lithium, one of its three electrons must exist in a higher energy state, and as a result, this electron can more easily be lost to another atom. Those electrons with approximately the same energy are said to form a "shell."

Although Bohr's model gives a qualitatively accurate description of atoms, it does not give quantitatively accurate results for atoms more complex than hydrogen. In order to describe such atoms, it is necessary to use quantum mechanics (see QUANTUM THEORY). This theory of atomic and subatomic phenomena was created by Erwin Schrödinger (1887–1961), Werner Heisenberg (1901–1976), Paul Dirac (1902–1984), and Pascual Jordan (1902–1980) in the 1920s. In quantum mechanics, the electron orbits are replaced by probability distributions that only indicate in which regions of space each electron is most likely to be found. An equation discovered by Schrödinger allows this distribution to be calculated for each atom. From the distribution, properties of the atom such as energy and angular momentum can be determined.

In the second stage, particle physics accommodated, through an analysis of isotopes of elements, that all atomic nuclei could be thought of as composed of two types of particles: the proton, which carries both mass and electric charge, and the neutron, which has about the same mass as a proton but is electrically neutral. This model was confirmed through the discovery (1932) of free neutrons by James Chadwick (1891–1974).

3.5 Neutron

Physicists by the late 1920s were convinced that they sufficiently understood the electronic structure of atoms. Attention therefore turned to the nucleus. It was already known that nuclei sometimes change into one another through radioactive decay. Rutherford had also shown, in 1919, that this could be accomplished artificially by bombarding nitrogen nuclei with high-energy alpha particles. In the process the nitrogen nucleus is converted into an oxygen nucleus, and a hydrogen nucleus, or proton, is ejected. It had further been discovered by Joseph J. Thomson, Francis W. Aston (1877–1945), and others that for a given element the nucleus sometimes occurs in several different forms that differ in mass. These chemically similar but physically distinct atoms were called isotopes. All of this provided evidence that atomic nuclei also had some kind of internal structure that could be explored through experiments and calculations.

Differences in the integer values of the electric charge and of the mass of many nuclei soon indicated that protons were not the only kind of particle to be found there. That is, the electric charge of a nucleus is always exactly an integer multiple of the charge of a proton, so that knowledge of this electric charge always indicates how many protons a nucleus contains. The mass of a nucleus is also approximately—but not exactly—an integer multiple of the mass of a proton. For many atoms, however, these two integer values are not the same. For example, a helium nucleus has twice the charge but four times the mass of a proton. Clearly, nuclei contain something other than protons.

This problem was solved in 1932 with the discovery by James Chadwick of the neutron. This particle has no electric charge and is slightly more massive than a proton. Thus most nuclei are composed of both protons and neutrons, which collectively are known as nucleons. A helium nucleus contains two protons and two neutrons, which correctly give the total charge and mass of the nucleus. The isotopes of any given element contain equal numbers of protons but different numbers of neutrons. For example, an isotope of hydrogen, called deuterium, contains one proton and one neutron, and a heavier isotope, called tritium, contains one proton and two neutrons.

The problem then arose as to how atomic particles could be held together in such a small region as the nucleus. The force holding them had to be different from others then known to physicists. It was stronger than the electric forces that can break electrons away from nuclei. On the other hand, the nuclear forces between different nuclei that are far apart are very weak, much weaker than electric forces at such distances.

3.6 Fundamental Particles

The third stage of particle physics came with the recognition that protons, neutrons, and electrons—the constituents of ordinary matter—were but three of a vast number of similar particles, which differed only in a few properties, such as their mass, and in their stability against spontaneous decay. Experiments with particle accelerators indicated that these many subatomic particles could be readily produced from protons and neutrons, provided that enough energy was available to produce the additional mass of the new particles predicted by the rules of Einstein's relativity theory. These discoveries in the 1940s and 1950s indicated that the proton and neutron were not really fundamental particles and that they would have to be understood as part of a much larger family of similar objects.

By 1932 nuclei were known to be composed of protons and neutrons. It was then necessary to explain how nuclei were held together, and in 1935, the Japanese physicist Hideki Yukawa (1907–1981) predicted a smaller fundamental particle that was the carrier of a theorized strong interaction, one of the four fundamental interactions, or forces. This particle, called a π-meson or pion, was discovered in 1947. Since then a host of particles smaller than protons and neutrons have been discovered in the nucleus, all falling within two classes: fermions, which obey the Pauli exclusion principle, and bosons, which carry the fundamental force. Modern nuclear physics centers on fundamental interactions between fermions and bosons. Protons and neutrons are composed of particles representing all four forces.

In the fourth stage, modern particle physics provided a successful explanation for the large number of particles. There are six different leptons: electron (e), muon (μ), tauon (τ), electron neutrino (ν_e), muon neutrino (ν_μ), and tau neutrino (ν_τ). There are also six quarks denoted up (u), down (d), charm (c), strange (s), top (t), and bottom (b). For each of these particles there exists an antiparticle. Many of the models for particle interactions pair the leptons and quarks into families: (e-ν_e), (μ-ν_μ), (τ-ν_τ), and (u-d), (c-s), (t-b). Experiments suggest that it is unlikely that there are more than these families. The interactions between these fundamental particles are mediated by gauge bosons (photon, intermediate bosons $W^\pm$ and Z^0, gluons).

4. WEAK, ELECTROMAGNETIC, AND STRONG INTERACTIONS: STRUCTURE OF THE MICROCOSMOS

4.1 Fundamental Interactions

At sufficiently low energies, there are four types of fundamental interactions whose existence is well established. Most studied are two of them: the gravitational and the electromagnetic interactions. The foundations of the classical (nonquantum) theory of the two interaction types were laid a long time ago by Newton and Maxwell.

In particle physics, the standard model encompasses all the particles known today and three of the fundamental interactions. The basic building blocks are two sets of matter particles, the quarks and the leptons (Table 1). These particles interact with each other through the exchange of gauge bosons (see Sec. 4.3, Table 2). The three fundamental interactions of the standard model are the electromagnetic, the strong, and the weak interactions, respectively. Gravity remains outside the standard model, but this does not invalidate the model as gravitational effects on particles are far smaller than the effects of the other interactions. The standard model has two components. One is the theory of strong interaction, called quantum chromodynamics. The other component is the theory that gives a unified description of electromagnetic and weak interactions, called the electroweak theory. The physical concepts used in the standard model are generalizations of concepts familiar in quantum electrodynamics (Kolb and Turner, 1990; Kaku, 1993).

4.1.1 Gravitational Interaction Gravitational interaction that governs the motion of celestial bodies is characterized by New-

Table 1. Fundamental fermions divided into two groups: leptons and quarks (Barnett *et al.*, 1996).

Leptons (spin = $\frac{1}{2}$)

Electric charge	Flavor	Mass (GeV/c^2)	Flavor	Mass (GeV/c^2)	Flavor	Mass (GeV/c^2)
0	ν_e, electron neutrino	$<1 \times 10^{-8}$	ν_μ, muon neutrino	1.7×10^{-4}	ν_τ, tau neutrino	$<2.4 \times 10^{-3}$
-1	e, electron	5.1×10^{-4}	μ, muon	0.106	τ, tau	1.777

Quarks (spin = $\frac{1}{2}$)

Electric charge	Flavor	Approx. mass (GeV/c^2)	Flavor	Approx. mass (GeV/c^2)	Flavor	Approx. mass (GeV/c^2)
$\frac{2}{3}$	u, up	$(2–8) \times 10^{-3}$	c, charm	1–1.6	t, top	180
$-\frac{1}{3}$	d, down	$(5–15) \times 10^{-3}$	s, strange	0.1–0.3	b, bottom	4.1–4.5

ton's constant $G = 6.7 \times 10^{-8}$ g^{-1} cm^3 s^{-2}. An excellent approximation that describes the gravitational interaction of two point masses m, a distance r apart, is Newton's formula,

$$F = Gm^2/r^2. \tag{1}$$

All fundamental particles are affected by gravity. The relativistic generalization of Newton's theory of gravity is Einstein's theory of general relativity.

4.1.2 Electromagnetic Interaction

Electromagnetic interaction determines the motion of charged bodies and acts only on charged particles. In the general case, their law of motion is described by Maxwell's equations. In the quasistatic approximation, an analog to Newton's law, the Coulomb approximation,

$$F = e^2/r^2, \tag{2}$$

proves to work well. Here, e denotes the charge of each point mass. The quantum field theory of electromagnetism, quantum electrodynamics, is the best theory available for describing effects of the electromagnetic force. One of this theory's important features is its gauge symmetry, which means that when independent changes to local field values are made at different points in space, the equations of quantum electrodynamics are not changed. This symmetry is ensured only if the quantum description of a charged particle contains an electromagnetic field with its gauge boson; i.e., the gauge symmetry demands the existence of the electromagnetic force and the photon. The symmetry is also linked to the ability to renormalize quantum electrodynamics so that it yields sensible, finite results.

The magnitudes of Gm^2 and e^2 depend on the choice of the system of units. To facilitate comparison in the framework of quantum field theory, one combines these quantities with universal physical constants, the Planck constant $\hbar$ and the velocity of light c, to obtain dimensionless constants. Thus, the nondimensional gravitational constant (see end of Sec. 4, Table 3),

$$\alpha_g = Gm^2/\hbar c, \tag{3}$$

and the nondimensional electromagnetic fine structure constant (see end of Sec. 4, Table 3),

$$\alpha_e = e^2/\hbar c, \tag{4}$$

are obtained, $e \approx 10^{-19}$ C being the electron (proton) charge. There is a difference in the definition of the two constants, α_e being in a way more universal than α_g. The definition of the number α_e contains fundamental constants only, whereas the constant α_g involves a mass m that is, generally speaking, arbitrary. To eliminate this arbitrariness, it is common to fix the value of m by setting it equal to the proton mass m_p. This choice is quite natural, for the proton is one of the two stable particles constituting the structure of the universe; the other one is the electron, with mass m_e. The choice between m_p and m_e is a matter of convention ($m_p \approx 1837 m_e$).

4.1.3 Weak Interaction The weak interaction governs the decay of particles into lighter ones and acts upon all quarks and leptons, including those with no electric charge. Historically, the first decay discovered was the decay of a neutron within an atomic nucleus (the β decay), according to the reaction $n \rightarrow p + e^- + \bar{\nu}$ (n, p, e^-, and $\bar{\nu}$ denoting a neutron, a proton, an electron, and an antineutrino, respectively). Subsequently, the discovery of new particles was intensified by progress in the development of accelerators. It turned out that all newly discovered particles have a common property: Heavy particles decay into lighter ones. Numerous investigations led to the conclusion that many decays are controlled by a unique interaction, referred to as the weak interaction, which is characterized by the Fermi coupling constant $g_F = 10^{-49}$ erg cm^3. The corresponding dimensionless coupling constant for the weak interaction is (see end of Sec. 4, Table 3)

$$\alpha_w = g_F m^2 c/\hbar^3. \tag{5}$$

The processes of collisions of neutrinos with matter are determined by the weak interaction as well.

There are many attempts to develop unified descriptions of all four interactions. In quantum theory every particle is associated with a particular field. How such a field transforms under the Lorentz transformation depends on the spin of the particle described by the respective field. A spin-0 particle can be described by a scalar field (Higgs field), a spin-$\frac{1}{2}$ particle by a spinor field, and a spin-1 particle by a vector field. The Lagrangian describing the field will carry information about the mass of the particle and its interactions. It is possible to construct a model that describes electromagnetic and weak interactions using a Lagrangian that possesses invariance under two transformation groups, U(1) and U(2).

The Higgs fields have a Lagrangian with a potential $V(\phi)$ that has nontrivial minima. If such a system comes into contact with matter fields that are in thermal equilibrium at some temperature T, then the effective potential energy will acquire a temperature dependence. That is, $V(\phi)$ will become $V(\phi,T)$. Such a temperature dependence can lead to several nontrivial effects—like phase transitions—in the early universe (Kolb and Turner, 1988). Even though the transformation group underlying the theory can be determined from some general principles, the detailed transformation properties of the fields representing specific particles cannot be derived from any fundamental considerations. These details are fixed using the known laboratory properties of these particles. For example, consider the fields describing the leptons. Given a spinor field ψ, one can construct its "right-handed" and "left-handed" components by the decomposition

$$\psi_L = \tfrac{1}{2}(1 - \gamma_5)\psi, \quad \psi_R = \tfrac{1}{2}(1 + \gamma_5)\psi, \tag{6}$$

where γ_5 is the 4×4 matrix

$$\gamma_5 = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}. \tag{7}$$

In the standard electroweak theory, the right-handed components behave as singlets (that is, they do not change) while the left-handed components transform as a doublet [that is, under an SU(2) transformation these fields are changed into linear components of themselves]. It is a consequence of this feature that in the simplest electroweak theory there is no necessity for the right-handed neutrino ν_R and that the left-handed neutrino ν_L is massless. Since the transformation properties of the fields are put into the theory by hand, it is possible to generalize these models in many ways. In particular, it is possible—though not necessary—to have massive neutrinos in the theory. This arbitrariness is of great importance for the existence of dark matter in the universe and for the solar neutrino problem.

4.1.4 Strong Interaction The strong interaction was identified with the nuclear interaction, which acts only on quarks and is ultimately responsible for binding protons and neutrons within the nucleus. The attempts to develop a consistent theory of nuclear interaction took a long time. A breakthrough was achieved with the progress of the dynamical theory of quark systems that led to the advent of quantum chromodynamics. In that scheme, the nuclear interaction was identified with the interaction in many-quark systems. It is instructive to trace briefly the evolution of the quark interpretation of nuclear interaction. To do so, we briefly outline the quark model proposed by Murray Gell-Mann (b. 1929) and George Zweig (b. 1937) in 1964. According to this model, each proton and neutron consists of three pointlike particles that are referred to as quarks and possess a charge that is a fraction of the electron charge e: $\pm\frac{1}{3}e$ or $\pm\frac{2}{3}e$. This theoretical conclusion was seemingly in contradiction to the experimental evidence that all the observable particles have integer electric charges. Nevertheless, numerous experimental confirmations of the quark hypothesis (such as systematics of the elementary particles, the magnitude of the magnetic moments, the ratios of the interaction cross sections, etc.) suggested that it deserves serious consideration. But then a profound question arose: How can the existence of quarks be reconciled with their nonobservability in experiments? At present, this problem is referred to as that of quark confinement. A postulate is invoked that has rather a character of magic: Quarks do exist, but in bound states. Even though no final solution of the confinement problem is available, one bases some expectations on the construction of a mathematical model that claims to provide a theory of the interaction between the quarks. It is this interaction that is identified with the strong interaction governing nuclear interaction. In 1954, Chen N. Yang (b. 1922) and Robert L. Mills (b. 1927) proposed a theory that is basically different from electrodynamics but accounts for the interaction caused by the transfer of zero-mass particles. The only such particle known at that time was the photon. The photon is the gauge particle in electrodynamics. That is why the Yang–Mills theory was considered just a mathematical exercise. The picture changed when a need emerged for a theory describing the dynamics of quarks. It seemed natural to consider the massless particles introduced by Yang and Mills to be responsible for the quark interaction. These particles were named gluons; by analogy with quantum electrodynamics, the quantum field theory of electromagnetism, one of the variants of the Yang–Mills theory is referred to as quantum chromodynamics. While gluons are analogs of photons, quarks are analogs of electrons. They carry not only color charges but also ordinary electric charge. For example, a proton consists of three quarks ($p = uud$), and a neutron consists of three quarks ($n = udd$), held together by continuing exchange of gluons. In the early 1970s the Yang–Mills equations were subjected to more scrutiny. As a result, the constant α_s was found to exhibit quite remarkable behavior, as distinct from quantum electrodynamics. This constant determines the quark–quark interaction that is currently believed to be the true strong interaction. It should be remembered that from the viewpoint of contemporary field theory the interaction is mediated by gauge bosons, i.e., quanta of the corresponding field. Energy momentum and hence—according to the special theory of relativity—mass is transferred along with a quantum. Elaborate calculations have demonstrated that the strong-interaction coupling constant α_s essentially depends on the energy-momentum and the mass m transferred. In a way, one had encountered a mass dependence of the constants α before (e.g., α_g and α_w), but quantum chromodynamics introduces a basic difference. In this theory, the dependence $\alpha_s(m)$ is deduced from quantum field theory, and not postulated, as was done earlier for the constants α_g and α_w on the basis of dimensional considerations. In addition, the variation of the constant α_s with the mass m has a specific feature: α_s decreases with increasing m. It should be remarked here that the terminology repeatedly used above might appear contradictory. On the one hand, one speaks of the constants α; on the other hand, one keeps stressing their dependence on m. In fact, the constants α are only constant at a fixed m; they vary with changing m. Accordingly, they are referred to as "running" constants. The final expression for the dependence of α_s on m reads, in the asymptotic approximation when $m \gg m_p$ (see end of Sec. 4, Table 3),

$$\alpha_s \sim \frac{a}{\ln(m/m_p)}\,. \tag{8}$$

The quantity a depends on N_q, the number of the sorts of quarks. In a standard theory ($N_q = 6$), $a \sim 1$. It follows from this formula that $\alpha_s \to 0$ as $m \to \infty$. This is the phenomenon of asymptotic freedom. A similar dependence also follows from a more exact expression. Unfortunately, the latter also has been obtained by methods whose validity breaks down for $m < m_p$. A "true" expression for α_s at small m is missing, because α_s is large, thus rendering standard computation techniques inapplicable. One can only state that for a small characteristic mass m, corresponding to the proton (or neutron) size, $r_N \approx 10^{-13}$ cm, the coupling constant is large. Furthermore, a rapid increase of the constant α_s with r approaching r_N inhibits progress in solving another problem, namely, that of nuclear forces. Today, quantum chromodynamics is considered as a theory that describes the interactions among quarks and gluons, out of which atomic nuclei are made.

4.2 Classification of Fundamental Particles

Fundamental particles are classified with respect to various parameters. A particle classification appears to be provided by the value of the spin, s. The behavior of particles depends on whether their spin is characterized by an integer (0,1,2,. . .) or a half-integer ($\frac{1}{2}$,$\frac{3}{2}$,$\frac{5}{2}$,. . .). Particles with a half-integer spin are referred to as fermions, those with an integer spin as bosons. In the framework of quantum mechanics, the difference in the behavior of fermions and bosons is expressed by the kind of symmetry of the wave functions describing these particles. A system consisting of fermions obeys Pauli's exclusion principle, as distinct from a boson system on which no such exclusion principle is imposed. Pauli's principle reads as follows: No two fermions may be in exactly the same quantum state.

An excellent illustration of Pauli's principle is the atomic level structure underlying the periodic system of the chemical elements. It is known that the first period of this system is composed of two elements, hydrogen and helium. For the first period, the principal quantum number equals unity. The atomic states associated with the first period are therefore determined only by the value of the spin projection of orbital electrons. There are two such values. Thus, only two elements can occur in the first period. For the second period, the principal quantum number equals two, giving rise to eight possible different states and thus to eight elements, etc. The Pauli principle is one of the foundations of the very structure of the periodic system. If this principle did not work, all the atomic electrons would populate the ground energy level (i.e., the hydrogen level), and, consequently, the periodicity of the system as well as the valency of chemical bonding would vanish. It is the Pauli principle that prevents atomic electrons from occupying the energetically most favorable ground state.

Another basis for the classification of the fundamental particles is their interaction. All particles participating in the strong interaction are referred to as hadrons (from the Greek word "hadros," meaning "strong"). All fermions that do not participate in the strong interaction are called leptons. A special place in this classification is reserved for the bosons, particles that mediate the interactions. The hadrons, in turn, are subdivided into the baryons and the mesons. The baryons are fermions; the lightest baryon is the proton. The hadrons with integer spin are referred to as mesons; the lightest meson is the pion ($m_\pi \approx 140$ MeV).

4.3 How Fundamental Particles Interact

Particles interact by exchange of gauge bosons; the exchange in the process of interaction involves not only energy, momentum, and mass but also the internal quantum numbers: spin, isospin, charge, and color. The properties of the exchange particles in the context of quantum field theory determine the interaction to a great extent. All exchange particles are bosons. The properties of the exchange particles are summarized in Table 2.

4.3.1 Graviton The graviton is a massless particle with zero charge and a spin $s = \pm 2$ and has not been detected because of its extremely weak interaction. Although most physicists have no doubts about the existence of gravitons, some caution is advisable, since the quantum theory of gravitation itself

Table 2. Properties of the force carriers (Barnett *et al.*, 1996).

	Bosons (spin = 0,1,2, . . .)	
	Electric charge	Mass (GeV/c^2)
Unified elektroweak (spin = 1)		
γ, photon	0	0
W^-	−1	80
W^+	+1	80
Z^0	0	91
Stronger or color (spin = 1)		
g, gluon	0	0

is far from being complete. Because of the weakness of the gravitational field, there is no hope for rapid progress in detecting and investigating gravitons. Because the graviton is a massless particle, the gravitational interaction is long range.

4.3.2 Photon Photons have spins of $s = \pm 1$ and a rest mass of zero, and are their own antiparticles. The electromagnetic interaction has a long range because the photon is massless.

4.3.3 Intermediate Bosons The weak interaction is mediated by three intermediate gauge bosons $W^{\pm}$ and Z^0, which have masses of ~80 and ~91 GeV, respectively. Since the range of an interaction is inversely related to the mass of the gauge boson, the weak interaction has an extremely short range.

4.3.4 Gluons The gluons, like the quarks, are not observable in the free state. However, in the late 1970s, considerable progress in the indirect verification of gluons was achieved by investigating the annihilation of high-energy positrons and electrons with hadron generation. It turned out that three hadronic jets occur in such processes; two jets are attributed to quarks, and the third one, to gluons. The experimental data on the three-jet processes accompanying positron–electron annihilation are in good agreement with the predictions of quantum chromodynamics, which indirectly confirms the existence of gluons, one of the basic elements of that theory. The strong interaction is mediated by a set of gauge bosons, containing at least eight gluons.

The data presented in Table 3 list the properties of the four interactions. A new entry here is the value of the interaction radius. For the gravitational, the weak, and the electromagnetic interaction, the magnitude of the radius r is determined from the uncertainty relation, $\hbar/m_B c$, m_B denoting the mass of an exchange particle. In the case of the strong interaction, the interaction radius r_N may be regarded either as an empirical constant or as the distance at which the value of the coupling constant α_s becomes unity.

5. GRAVITATIONAL INTERACTION: STRUCTURE OF THE MACROCOSMOS

5.1 Structural Elements of the Universe

The structure of the macrocosmos is manifested on many different scales, ranging from the universe on the largest scale, down to galaxies, stars, and planets. Only objects of the microcosmos, such as quarks and leptons, may be devoid of further substructure. (However, a recent experiment is suggesting that quarks and gluons may be composed of more fundamental particles; see Wilczek, 1996.) Most scales are determined to an order of magnitude by a few physical constants. In particular, the mass scale and length scale (in units of the proton mass m_p and the Bohr radius a_0) of structures down to the planets can be expressed in terms of the electromagnetic fine-structure constant α_e, Eq. (4), the gravitational fine-structure constant α_g, Eq. (3), and the electron-to-proton mass ratio ($m_p \approx 1837 m_e$).

The following considerations are based on order-of-magnitude arguments, factors of order unity (like π) being neglected. These results have been published in the astrophysical literature on several occasions and are collected here for easy reference (a detailed discussion is contained in Carr and Rees, 1979; Barrow and Tipler, 1988).

5.1.1 The Planck Scales The only quantities of dimensions mass and length that can be constructed from G, $\hbar$, and c are the Planck scales:

Table 3. Properties of the four fundamental interactions.

	Coupling constant		
Interaction	Analytic expression	Numerical value at $m = m_p$	Interaction radius (cm)
Gravitational	$\alpha_g = Gm^2/\hbar c$	$\sim 10^{-38}$	∞
Weak	$\alpha_w g_F m^2 c/\hbar^3$	10^{-5}	10^{-17}
Electromagnetic	$\alpha_e = e^2/\hbar c$	1/137	∞
Strong	$\alpha_s = a/[\ln(m/m_p)]$, $m \gg m_p$	≈ 1	10^{-13}

$$M_{\mathrm{Pl}} \sim (G/\hbar c)^{-1/2} \sim 10^{-5}\ \mathrm{g},$$
$$R_{\mathrm{Pl}} \sim (G/\hbar c^3)^{1/2} \sim 10^{-33}\ \mathrm{cm}. \tag{9}$$

Using the gravitational fine structure constant, Eq. (3), these scales can be expressed as

$$M_{\mathrm{Pl}} \sim \alpha_G^{-1/2} m_p, \quad R_{\mathrm{Pl}} \sim \alpha_G^{1/2} r_p. \tag{10}$$

Thus M_{Pl} is much larger than m_p but R_{Pl} is much smaller than r_p (r_p being the size of a proton, which can be taken to be the Compton wavelength associated with its rest mass, $r_p \sim \hbar/m_p c \sim 10^{-13}$ cm; accordingly, the corresponding time scale is $t_p \sim r_p/c \sim 10^{-23}$ s). The Planck length is the scale on which quantum gravitational fluctuations in the metric become of the order of unity, so that the concept of space breaks down at such small scales. M_{Pl} can be interpreted as the mass of a black hole of radius R_{Pl}. Space may be thought of as being filled with virtual black holes of this size. Such "instantons" may play an important role in quantum gravity theory.

5.1.2 The Universe In the simplest Friedmann cosmological model, the age of the universe t_0, is of the order of H_0^{-1} where H_0 is the Hubble parameter (this relation fails only if the universe is closed and near its maximum expansion). Since $H_0 \sim 50$ km s^{-1} Mpc^{-1}, this implies $t_0 \sim 10^{10}$ yr, a conclusion that is supported by several independent arguments. The associated horizon size (the distance traveled by light since the birth of the universe) satisfies the approximate relationship

$$ct_0 \sim \alpha_g^{-1}(\hbar/m_e c) \sim (\alpha_e/\alpha_g)a_0, \tag{11}$$

where a_0 denotes the radius of the lowest-energy electron orbit of a hydrogen atom, $a_0 \sim \hbar^2/m_e e^2 \sim 10^{-8}$ cm ~ 1 Bohr. The ratio of the size of the observable universe to the size of an atom is comparable with the ratio of the electrical (or nuclear) and gravitational forces between elementary particles. There is no explanation for this well-known coincidence within conventional physics, but Dirac (1937, 1938) has conjectured that α_g is always given by

$$\alpha_g \sim \hbar/m_e c^2 t \sim (t/t_e)^{-1}. \tag{12}$$

Assuming that $\hbar$, c, and m_e are constant in time, this requires that G decreases as t^{-1}, and so Dirac invokes Eq. (12) as the basis for a new cosmology (Barrow, 1996). Such a variation of G is inconsistent with current observations. The total mass associated with the observable universe (the mass within the horizon volume) is $\sim \rho_0 c^3 t_0^3$, where ρ_0 is the present matter density, given by the Friedmann equation,

$$\rho_0 = \frac{3H_0^2}{8\pi G} + \frac{Kc^2}{16\pi G}. \tag{13}$$

Here K is the scalar curvature of the universe. Provided that the K term is smaller than the others, one deduces that the mass of the universe is

$$M_u \sim c^3 t_0^3 G^{-1} H_0^2 \sim \frac{c^3 t_0}{G} \sim \alpha_g^{-2}\left(\frac{m_p}{m_e}\right) m_p. \tag{14}$$

The fact that the number of protons in the universe is of the order of α_g^{-2} is thus explained provided one can justify neglecting the K term in Eq. (13). It has been argued that K must always be zero by appealing to Mach's principle but, apart from this, there may be reasons for expecting that the K term is small. If K is negative, galaxies could not have condensed out from the general expansion unless $-K$ were less than $G\rho/c^2$ at their formation epoch. Otherwise, their gravitational binding energy would not have been large enough to halt their expansion. The term $-K$ may exceed $G\rho/c^2$ at the present epoch, but not by a large factor. If K is positive, it must be $<G\rho_0/c^2$, otherwise the whole universe would have recollapsed before $t \simeq t_{\mathrm{MS}}$ where t_{MS} is the lifetime of a typical main-sequence star—say, the Sun. Relationships (11) and (14) also mean that the universe has an optical depth of the order of unity to electron scattering.

5.1.3 Galaxies It is not certain how galaxies form, and so any estimate of their scale is very model dependent (Rees and Ostriker, 1977; Silk, 1977; Sciama, 1953). One can assume that galaxies originate from overdense regions in the gaseous primordial material and that they have a mass M and radius R_B when they become bound. After binding, motions in the protogalaxy randomize and equilibrate in the gravitational field of the galaxy at a radius $\sim R_B/2$. Thereafter, they will deflate on a cooling time scale, with a virial temperature

$$T \sim GMm_p/kR. \tag{15}$$

Provided kT exceeds one Rydberg, the dominant cooling mechanism is bremsstrahlung and the associated cooling time scale is

$$t_{\text{cool}} \simeq \frac{m_e^2c^3}{\alpha_e e^4 n}\left(\frac{kT}{m_ec^2}\right)^{1/2}. \tag{16}$$

The free-fall time scale is

$$t_{ff} \sim (GM/R^3)^{-1/2}, \tag{17}$$

and this exceeds t_{cool} when R falls below a mass-independent value

$$R_g \sim \alpha_e^4\alpha_g^{-1}(m_p/m_e)^{1/2}a_0, \tag{18}$$

which, from a more precise calculation, is 75 kpc. Until a massive cloud gets within this radius, it will contract quasistatically and cannot fragment into stars. This argument applies only if the mass is so high that $kT_{\text{virial}} > \alpha_e^2m_ec^2$ at the "magic radius" R_g; that is,

$$M \geq M_g = \alpha_g^{-2}\alpha_e^5(m_p/m_e)^{1/2} \simeq 10^{12}M_\odot. \tag{19}$$

Gas clouds of mass $<M_g$ cool more efficiently as a result of recombination and can never be pressure supported. Thus, M_g is a characteristic maximum galactic mass. Primordial clouds of mass $<M_g$ are inhibited from fragmentation and may remain as hot pressure-supported clouds. This type of argument can be elaborated and made more realistic (White and Rees, 1978; Rees and Ostriker, 1977; Silk, 1977; Sciama, 1953); but one still obtains a mass $\sim M_g$ above which any fluctuations are likely to remain amorphous and gaseous, and which may thus relate to the mass scale of galaxies. The quantities M_g and R_g may thus characterize the mass and radius of a galaxy. These estimates are the least certain. The properties of galaxies may be a consequence of irregularities imprinted in the universe by processes at early epochs.

5.1.4 Stars The virial theorem implies that the gravitational binding energy of a star must be of the order of its internal energy. Its internal energy comprises the kinetic energy per particle (radiation pressure being assumed negligible for the moment) and the degeneracy energy per particle. The degeneracy energy will be associated primarily with the Fermi momentum of the free electrons, $p \sim \hbar/d$, where d is their average separation. Provided the electrons are nonrelativistic, the degeneracy energy is $p^2/2m_e$, and so the virial theorem implies (Dyson, 1972; Weisskopf, 1975; Haubold and Mathai, 1994)

$$kT + \frac{\hbar^2}{2m_ed^2} \sim \frac{GMm_p}{R} \sim \left(\frac{N}{N_0}\right)^{2/3}\frac{\hbar c}{d}. \tag{20}$$

Here N is the number of protons in the star, $N_0 \equiv \alpha_g^{-3/2}$, and $R \sim N^{1/3}d$ is its radius. As a cloud collapses under gravity, Eq. (20) implies that, for large d, T increases as d^{-1}. For small d, however, T will reach a maximum,

$$kT_{\max} \sim (N/N_0)^{4/3}m_ec^2, \tag{21}$$

and then decrease, reaching zero when d is

$$d_{\min} \sim (N/N_0)^{-2/3}r_e, \tag{22}$$

where r_e is the size of an electron, $r_e \sim \hbar/m_ec \sim 10^{-10}$ cm; accordingly, the electron time scale is defined as $t_e = r_e/c \sim 10^{-20}$ s. A collapsing cloud becomes a star only if $T_{\max}$ is high enough for nuclear reactions to occur, that is $kT_{\max} > qm_ec^2$ where q depends on the strong and electromagnetic interaction constants and is $\sim 10^{-2}$. From Eq. (21), one therefore needs $N > 0.1N_0$. Once a star has ignited, further collapse will be postponed until it has burned all its nuclear fuel. An upper limit to the mass of a star derives from the requirement that it should not be radiation-pressure dominated. Such a star would be unstable to pulsations that would probably result in its disruption. Using the virial theorem [that is, Eq. (20) with the degeneracy term assumed negligible] to relate a star's temperature T to its mass $M \sim Nm_p$ and radius R, the ratio of radiation pressure to matter pressure can be shown to be

$$\frac{P_{\text{rad}}}{P_{\text{mat}}} \sim \frac{aT^4R^3}{NkT} \sim \left(\frac{N}{N_0}\right)^2, \tag{23}$$

so that the upper limit to the mass of a star is also $\sim N_0 m_p$. A more careful calculation shows that there is an extra numerical factor of the order of 10, and so one expects all main-sequence stars to lie in the range $0.1 < N/N_0 < 10$ observed. Only assemblages of 10^{56}–10^{58} particles can turn into stable main-sequence stars with hydrogen-burning cores. Less massive bodies held together by their own gravity can be supported by electron "exclusion principle" forces at lower temperatures (they would not get hot enough to undergo nuclear fusion unless squeezed by an external pressure); heavier bodies are fragile and unstable on account of radiation-pressure effects. The central temperature T adjusts itself so that the nuclear energy-generation rate balances the luminosity, the radiant energy content divided by the photon leakage time,

$$L \simeq acT^4R^4/\kappa M, \tag{24}$$

where κ is the opacity. The appropriate κ decreases as M increases (electron scattering being the dominant opacity for upper main-sequence stars), but the energy generation increases so steeply with T that M/R depends only weakly on M.

5.1.5 White Dwarfs and Neutron Stars When a star has burned all its nuclear fuel, it will continue to collapse according to Eq. (20), and, provided it is not too large, it will end up as a cold electron-degeneracy–supported "white dwarf" with the radius $R \propto M^{-1/3}$ indicated by Eq. (22). However, when M gets so large that $kT_{\max} \sim m_ec^2$, the electrons will end up relativistic, with the degeneracy energy being pc rather than $p^2/2m_e$. Thus the degeneracy term in Eq. (20) acts as d^{-1} instead of d^{-2}, and consequently there is no $T = 0$ equilibrium state. From Eq. (21) this happens if M exceeds the mass

$$M_c \sim \alpha_g^{3/2} m_p \sim M_\odot, \tag{25}$$

which characterizes stars in general. A more precise expression for this critical value of M (Chandrasekhar mass) is $5.6\mu^2M_\odot$, where μ is the number of free electrons per nucleon (Chandrasekhar, 1935). For a star that has burned all the way up to iron, $\mu \approx 1/2$, and the limiting Chandrasekhar mass, taking into account the onset of inverse β decay when electrons get relativistic, is $\sim 1.25M_\odot$. Stars bigger than M_c will collapse beyond the white-dwarf density but may manage to shed some of their mass in a supernova explosion. The remnant core will comprise mainly neutrons, the electrons having been squeezed onto the protons through the reaction $p + e^- \rightarrow n + \nu$, and this core, if small enough, may be supported by neutron-degeneracy pressure. The limiting mass for a neutron star is more difficult to calculate than that for a white dwarf because of strong-interaction effects and because, from Eq. (21), with $m_e \rightarrow m_p$, the particles that dominate the neutron star's mass are relativistic. The maximum mass is still of the order of $M_\odot$, however, and a neutron star bigger than this must collapse to a black hole. The maximum mass lies close to the intercept of a black hole line, given by Eq. (26), and the nuclear density line $\rho \sim m_p/r_p^3$. The intricacies of the line that bridges white-dwarf and neutron-star regimes reflect the effects of gradual neutronization and strong interactions. Stars on this bridge would be unstable and so are not of immediate physical interest.

The above order-of magnitude arguments show why the effects of radiation pressure and relativistic degeneracy both become important for masses $>\alpha_g^{-3/2}m_p$. Note also that general relativity is unimportant for white dwarfs because the binding energy per unit mass is only $\sim(m_e/m_p)$ of c^2 at the Chandrasekhar limit.

5.1.6 Black Holes The radius of a spherically symmetrical black hole of mass M is

$$R = 2GM/c^2 \sim \alpha_g(M/m_p)r_p. \tag{26}$$

This is the radius of the event horizon, the region from within which nothing can escape, at least classically. Black holes larger than $M_\odot$ may form from the collapse of stars or dense star clusters. Smaller holes require much greater compression for their formation than could arise in the present epoch, but they might have been produced in the first instants after the big bang when the required compression could have occurred naturally. Such "primordial" black holes could have any mass down to the Planck mass. In fact, Hawking (1974) has shown that small black holes are not black at all; because of quantum effects *they emit particles* like a blackbody of temperature given by

$$k\theta \sim \frac{\hbar c^3}{GM} \sim \alpha_g^{-1}\left(\frac{M}{m_p}\right)^{-1} m_p c^2. \tag{27}$$

This means that a hole of mass M will evaporate completely in a time

$$t_{\text{evap}} \sim \alpha_g^2 (M/m_p)^3 t_p [N(\theta)]^{-1}. \tag{28}$$

$N(\theta)$ is the number of species contributing to the thermal radiation: For $k\theta < m_e c^2$ these include only photons, neutrinos, and gravitons; but at higher temperatures other species may contribute. The evaporation terminates in a violent explosion. For a solar-mass hole, this quantum radiance is negligible: θ is only 10^{-7} K and $t_{\text{evap}} \sim 10^{64}$ yr. But for small holes it is very important. A Planck-mass hole has a temperature of 10^{32} K and only survives for a time $\sim R_{\text{Pl}}/c \sim 10^{-43}$ s. Those holes that are terminating their evaporation in the present epoch are particularly interesting. As the age of the universe is $t_0 \sim \alpha_g^{-1} t_e$, the mass of such holes would be

$$M_h \sim \alpha_g^{-2/3}(t_0/r_p)^{1/3} m_p \sim \alpha_g^{-1} m_p \sim 10^{15}\ \text{g}, \tag{29}$$

and, from Eq. (26), their radius would be r_p. The corresponding temperature is ~10 MeV: low enough to eliminate any uncertainty in $N(\theta)$ due to species of exotic heavy particles.

5.2 Evolution of the Universe

The origin of the universe is governed by laws of physics that are still unknown at the time of writing the *Encyclopedia of Applied Physics* (see Plate 2).

At $t = 10^{-43}$ s, $T = 10^{32}$ K: The strong, weak, and electromagnetic forces may appear as unified into one indistinguishable force. This period is often referred to as the Grand Unification epoch. During this epoch, there may have been a very rapid, accelerating expansion of the universe called "inflation." The inflation made the universe very large and flat, but also produced ripples in the space-time it was expanding.

At $t = 10^{-34}$ s, $T = 10^{27}$ K: The strong force becomes distinct from the weak and electromagnetic forces. The universe is a plasma of quarks, electrons, and other particles. Inflation ends and the expanding universe coasts, gradually slowing its expansion under the pull of gravity.

At $t = 10^{-10}$ s, $T = 10^{15}$ K: The electromagnetic and weak forces separate (see Fig. 5). An excess of one part in a billion of matter over antimatter has developed. Quarks are able to merge to form protons and neutrons. Particles have acquired substance.

At $t = 1$ s, $T = 10^{10}$ K: Neutrinos decouple and the electrons and positrons annihilate, leaving residual electrons but predominantly the cosmic background radiation as the main active constituent of the universe.

At $t = 3$ min, $T = 10^9$ K: Protons and neutrons are able to bind together to form nuclei since their binding energy is now greater than the cosmic background-radiation energy. A rapid synthesis of light nuclei occurs—first deuterium (D), then heavier elements, primarily helium (^{3}He, ^{4}He) but up to lithium nuclei ^{7}Li (Tytler *et al.*, 1996; Gloeckler and Geiss, 1996). About 75% of the nuclei are hydrogen and 25% are helium; only a tiny amount are other elements. The heavier elements are later formed by nuclear burning stars.

At $t = 3 \times 10^5$ yr, $T = 3 \times 10^3$ K: Matter and the cosmic background radiation decouple as electrons bind with nuclei to produce neutral atoms. The universe becomes transparent to the cosmic background radiation, making it possible for the Cosmic Background Explorer satellite (COBE) to map this epoch of last scattering (see Plate 3).

A $t = 10^9$ yr, $T = 18$ K: Clusters of matter have formed from the primordial ripples to form quasars, primordial stars, and pro-

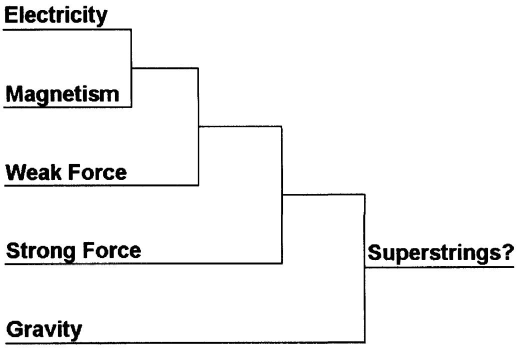

FIG. 5. Cosmology has been actively investigating the consequences of a new extension of the theory of matter within the evolution of the universe, in which the electricity, magnetism, weak force, strong force, and gravity are all unified at sufficiently high temperatures.

togalaxies. In the interior of stars, the burning of the primordial hydrogen and helium nuclei synthesizes heavier nuclei such as carbon, nitrogen, oxygen, and iron. These are dispersed by stellar winds and supernova explosions, making new stars, planets, and life possible.

At $t = 15 \times 10^9$ yr, $T = 3$ K: The present epoch is reached (see Plate 4). Five billion years earlier, the solar system condensed from the remnants of earlier stars. Chemical processes have linked atoms together to form molecules and then solids and liquids. Man has emerged from the dust of stars to contemplate the universe around him.

5.3 Large-Scale Distribution of Matter

The nearest large galaxy to the Milky Way is the "Andromeda galaxy," which is about 670 kpc away. Its mass is $\sim 3 \times 10^{11} M_\odot$ (or perhaps as much as $1.5 \times 10^{12} M_\odot$ if it has a massive halo), and it has a size of ~50 kpc. Studies show that it is a spiral galaxy. The Andromeda galaxy and the Milky Way lie nearly in each other's planes, but their spins are opposite. It may be noted that galaxies are packed in the universe in a manner very different from the way the stars are distributed inside a galaxy: The distance from the Milky Way to the nearest large galaxy is only 20 galactic diameters, while the distance from the Sun to the nearest star is 30 million times the diameter of such individual stars. There is some evidence to suggest that the Andromeda galaxy and the Milky Way are gravitationally bound to each other. They have a relative velocity—toward each other—of about 300 km s^{-1}. These two are only the two largest members of a group of more than 30 galaxies all of which together constitute what is known as the "Local Group." The entire Local Group is irregular in shape and can be contained within a spherical volume of ~2 Mpc in radius. This kind of clustering of galaxies into groups is typical in the distribution of the galaxies in the universe. A study within a size of about 20 Mpc from the Milky Way shows that only (10–20)% of the galaxies do not belong to any group; they are isolated galaxies, called "field" galaxies. Groups may typically consist of up to 100 galaxies; a system with more than 100 galaxies is conventionally called a "cluster." The sizes of groups range from a few hundred kpc to 2 Mpc. Clusters have a size of typically a few Mpc. Just like galaxies, one may approximate clusters and groups as gravitationally bound systems of effectively point particles. The large gravitational potential energy is counterbalanced by the large kinetic energy of random motion in the systems. The line-of-sight velocity dispersion in groups is typically 200 km s^{-1}, while that in clusters can be nearly 1000 km s^{-1}.

There are several similarities between clusters of galaxies and stars in an elliptical galaxy. For example, the radial distribution of galaxies in a cluster can be adequately fitted by the $R^{1/4}$ law with an effective radius $R_e \approx (1\text{–}2)h^{-1}$ Mpc ($0.5 \leq h \leq 1$ related to the Hubble constant). About 10% of all galaxies are members of large clusters. In addition to galaxies, clusters also contain very hot intracluster gas at temperatures 10^7–10^8 K. The two large clusters nearest to the Milky Way are the Virgo cluster and the Coma cluster. The Virgo cluster, located ~17 Mpc away, has a diameter of about 3 Mpc and contains more than 2500 galaxies. This is a prominent irregular cluster and does not exhibit a central condensation or a discernible shape. Virgo has an intracluster x-ray–emitting gas with a temperature of $\sim 10^8$ K, which has at least ten times the visible mass of the cluster. The Coma cluster, on the other hand, is almost spherically symmetric with a marked central condensation. It has an overall size of ~3 Mpc while its central core is ~600 kpc in size. The core is populated with elliptical and spheroidal galaxies with a density nearly 30 times larger than the Local Group. These values are typical for large clusters. Coma is located at a distance of ~80 Mpc and contains more than 1000 galaxies. The distribution of galaxies around the Local Group has been studied extensively. It turns out that most of the galaxies nearby lie predominantly in a plane—called the super-galactic plane—that is approximately perpendicular to the plane of the Milky Way galaxy. The dense set of galaxies in this plane is called the Local Supercluster (radius ~37 Mpc), and the Virgo cluster is nearly at the center of this highly flattened disclike system. The term "supercluster" is just used to denote structures bigger than clusters. Broadly speaking, the Local Supercluster consists of three components: About 20% of the brightest galaxies, forming the core, is the Virgo cluster itself; another 40% of galaxies lie in a flat disc with two extended, disjoint groups of galaxies; the remaining 40% is confined to a small number of groups scattered around. Nearly 80% of all matter in the Local Supercluster lies in a plane. The Local Supercluster is clearly expanding. Studies of distant galaxies show that there are many superclusters in our universe separated by large voids. The distribution of matter seems to be reasonably uniform when observed at scales bigger than about $100h^{-1}$ Mpc. For comparison, note that the size of the observed universe is about $3000h^{-1}$ Mpc. Thus, one may treat the matter distribution in the universe to be homogeneous while dealing with the phenomena at scales larger than 100 Mpc. The standard cosmological model is based on this assumption of "large-scale" homogeneity (see Fig. 6 and Table 4; Geller and Huchra, 1989).

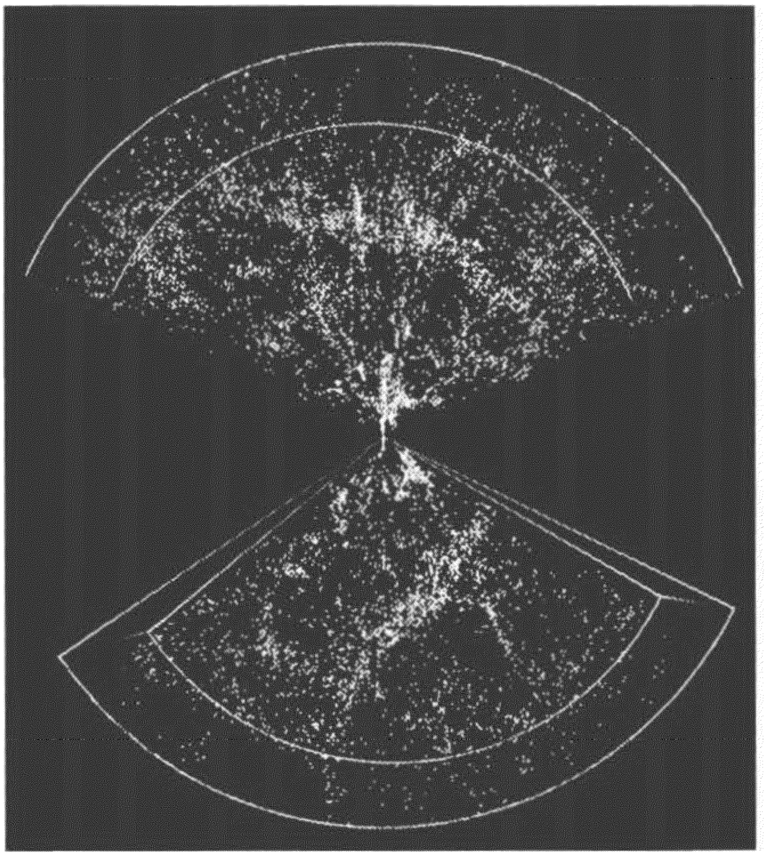

FIG. 6. A map of the nearby universe toward the north and south poles of the Milky Way. Each of the 9325 points in the image represents a galaxy similar to the Milky Way. The arcs that form the boundaries of the two wedgelike portions of the map are all at a distance of about 400 million light years from the Sun (Earth). The dark regions to the east and west are obscured by the plane of the Milky Way. The map shows that galaxies are arranged in patterns on an enormous scale. The Great Wall, a sheet containing thousands of galaxies, stretches nearly horizontally across the entire northern portion of the survey. A similar Southern Wall runs diagonally across the southern region. These walls delineate enormous dark voids where there are few if any galaxies. The voids are often 150 × 10^6 light years in diameter. The patterns in the north and south are similar. These large patterns are a tough challenge for attempts to model the development of structure in the universe. The curved boundaries are lines of constant declination (Galactic latitude): in the north they are at 8.5° and 44.5° (the wedge subtends an angle of 36° in the narrow direction). It runs from 8 to 17 h in right ascension (longitude) or of order 120°. In the south the declination runs from 0° to −40°, and the right ascension runs from 20.8 to 4 h. By permission of Margaret J. Geller, John P. Huchra, Luis A. N. da Costa, and Emilio E. Falco, Smithsonian Astrophysical Observatory, © 1994.

Table 4. Structural elements of the macrocosmos.

Object	Mass (g)	Radius (cm)
Jupiter	2×10^{30}	6×10^{9}
Sun	2×10^{33}	7×10^{10}
Red giant	$(2\text{–}6) \times 10^{34}$	10^{14}
White dwarf	2×10^{33}	10^{8}
Neutron star	3×10^{33}	10^{6}
Open cluster	5×10^{35}	3×10^{19}
Globular cluster	1.2×10^{39}	1.5×10^{20}
Spiral	$2 \times (10^{44}\text{–}10^{45})$	$(6\text{–}15) \times 10^{22}$
Elliptical	$2 \times (10^{43}\text{–}10^{45})$	$(1.5\text{–}3) \times 10^{23}$
Group	4×10^{46}	3×10^{24}
Cluster	2×10^{48}	1.2×10^{25}

5.4 Morphology of Galaxies

Galaxies range widely in their sizes, shapes, and masses; nevertheless, one may talk of a typical galaxy as something made out of $\sim 10^{11}$ stars. Taking the mass of a star to be that of the Sun, the luminous mass in a galaxy is $\sim 10^{11} M_\odot \approx 2 \times 10^{44}$ g.

This mass is distributed in a region with a size of ~20 kpc. Even though most galaxies have a mass of $\sim(10^{10}\text{–}10^{12})M_\odot$ and a size of ~10–30 kpc, there are several exceptions at both ends of this spectrum. For example, "dwarf galaxies" have masses in the range $\sim (10^{5}\text{–}10^{7})M_\odot$ and radii of ~1–3 kpc. There are also some "giant galaxies" with masses as high as $10^{13} M_\odot$. Galaxies exhibit a wide variety in their shapes as well and are usually classified according to their morphology (Hubble's classification of galaxies, for example). One may divide them into "ellipticals" and "discs." Ellipticals are smooth, featureless, distributions of stars, ranging in mass from $10^{8} M_\odot$ to $10^{13} M_\odot$. The proportion of elliptical galaxies in a region depends sensitively on the environment. They contribute ~10% of all galaxies in low-density regions of the universe but ~40% in dense clusters of galaxies. The second major type of galaxy is the "spiral" (or "disc") to which the Milky Way belongs. Spirals have a prominent disc, made of population I stars (see Sec. 5.6) and contain a significant amount of gas and dust. The name originates from the distinct spiral arms that exist in many of these galaxies. In low-density regions of the universe, ~80% of the galaxies are spirals while only ~10% of galaxies in dense clusters are spirals. This is complementary to the behavior of ellipticals. The stars in a disc galaxy are supported against gravity by their systematic rotation velocity. The rotational speed of stars $v(R)$ at radius R has the remarkable property that it remains constant for large R in almost all spirals; the constant value is typically 200–300 km s^{-1}. Most discs also contain a spheroidal component of population II stars. The luminosity of the spheroidal component relative to the disc correlates well with several properties of these galaxies. This fact has been used for classifying the disc galaxies into finer divisions. The stars in a galaxy are not distributed in a completely uniform manner. A typical galaxy contains several smaller stellar systems, each containing $\sim 10^{2}\text{–}10^{6}$ stars. These systems, usually called star clusters, can be broadly divided into two types called "open clusters" and "globular clusters." Open clusters consist of $\sim 10^{2}\text{–}10^{3}$ population I stars bound within a radius of ~1–10 pc. Most of the stars in these clusters are quite young. In contrast, globular clusters are population II systems with $\sim 10^{4}\text{–}10^{6}$ stars. The Milky Way contains ~200 globular clusters, which are distributed in a spherically symmetric manner about the center of the galaxy. Unlike open clusters, the stars in globular clusters are quite old. The mass density of stars in the core of a globular cluster, $\sim 10^{4} M_\odot\ \mathrm{pc}^{-3}$, is much higher than that of a typical galaxy, $\sim 0.05 M_\odot\ \mathrm{pc}^{-3}$. The core radius of the globular clusters is ~1.5 pc while the "tidal radius," which is the radius at which the density drops nearly to zero, is ~50 pc. In addition to the stars, the Milky Way also contains gas and dust that contribute ~(5–10)% of its mass. This interstellar medium may be roughly divided into a very dense, cold, molecular component (with about 10^{4} particles per cm^{-3} and a temperature of ~100 K), made of interstellar clouds, a second component that is atomic but neutral (with $n \approx 1\ \mathrm{cm}^{-3}$ and $T \approx 10^{3}$ K), and a third component that is ionized and very hot (with $n \approx 10^{-3}\ \mathrm{cm}^{-3}$ and $T \approx 10^{6}$ K). Though the interstellar medium is principally made of hydrogen, it also contains numerous other chemical species and an appreciable quantity of tiny solid particles ("dust"). The spiral arms are concentrations of stars and interstellar gas and are characterized by the presence of ionized hydrogen. This is also the region in which young stars are being formed. For a more detailed study of galaxies, one can divide ellipticals and spirals into subsets and also add two more classes of galaxies, called "lenticulars" and "irregulars." The ellipticals are subdivided as (E1, . . . ,En, . . .) where $n = 10(a - b)/a$ with a and b denoting the major and minor axes of the ellipticals. The lenticulars (also called SO) are the galaxies "in between" ellipticals and spirals. They have a prominent disc, which contains no gas, dust, bright young stars, or spiral arms. Though they are smooth and featureless like ellipticals, their surface brightness follows the exponential law of the spirals. They are rare in low-density regions (less than 10%) but constitute nearly half of the galaxies in the high-density regions. The spirals are subdivided into Sa, Sb, Sc, and Sd with the relative luminosity of the spheroidal component decreasing along the sequence. The amount of gas increases and the spiral arm becomes more loosely spiraled along the sequence from Sa to Sd. The Milky Way is between the types Sb and Sc. There also exists another class of spirals called "barred spirals," which exhibit a barlike structure in the center. They are classified as SBa, SBb, etc. Finally, irregulars are the galaxies that do not fall in the abovementioned morphological classification. These are low-luminosity, gas-rich systems with massive young stars and large HII regions (i.e., regions containing ionized hydrogen). More than one-third of the galaxies in the Milky Way's neighborhood are irregulars. They are intrinsically more difficult to detect at larger distances because of their low luminosity. As seen earlier, stars evolve and their luminosity and hence their color change. Since galaxies are made of stars, galaxies will also exhibit color evolution. Besides, the gas content and elemental abundances of the galaxies will change as the stars are formed and end as planetary nebulae or supernovae.

5.5 Quasars

Most galaxies that are observed have a fairly low redshift ($z < 0$–5) and an extended appearance on a photographic plate. There exists another important class of objects, called "quasars," that exhibit large redshifts (up to $z \approx 5$) and appear as point sources in the photographic plate (Shaver *et al.*, 1996). Estimating the distance from the redshift and using the observed luminosity, it is found that quasars must have a luminosity of about $L_q \approx 10^{46}$–10^{47} erg s^{-1}. It is possible to estimate the size of the region emitting the radiation from the time scale in which the radiation pattern is changing. It turns out that the energy from the quasar is emitted from a very compact region. One can easily show that nuclear fusion cannot be a viable energy source for quasars. It is generally believed that quasars are fueled by the accretion of matter into a supermassive black hole ($M \approx 10^8 M_\odot$) in the center of the host galaxy. The friction in the accretion disc causes the matter to lose angular momentum and hence spiral into the black hole; the friction also heats up the disc. Several physical processes transform this heat energy into radiation of different wavelengths. Part of this energy can also come out in the form of long, powerful, jets. The innermost regions of the quasar emit x and γ rays. Outer shells emit UV, optical, and radio continuum radiation in the order of increasing radius. Very bright quasars have an apparent magnitude of $m_B \approx 14$ (which corresponds to a flux of 10^{-25} erg s^{-1} cm^{-2} Hz^{-1}) while the faintest ones have $m_B \approx 23$. The absolute magnitudes of the quasars are typically $-30 < M_B < -23$; in contrast, galaxies fall in the band $-23 < M_B < -16$. The relation between m and M for any given quasar depends on the estimated distance to the quasar, which in turn depends on the cosmological model and the quasar's redshift. The luminosity function of the quasars, as a function of their redshift, has been a subject of extensive study. These investigations show that

1. the space density of bright galaxies (about 10^{-2} Mpc^{-3}) is much higher than that of quasars (about 10^{-5} Mpc^{-3}) with $z < 2$, and
2. bright quasars were more common in the past ($z \approx 2$) than today ($z \approx 0$).

Quasars serve as an important probe of the high-redshift universe. Quasars are believed to be one extreme example of a wide class of objects called "active galaxies." This term denotes a galaxy that seems to have a very energetic central source of energy. This source is most likely to be a black hole powered by accretion. One kind of active galaxy that has been studied extensively is radio galaxies. The most interesting feature about these radio galaxies is that the radio emission does not arise from the galaxy itself but from two jets of matter extending from the galaxy in opposite directions. It is generally believed that this emission is caused by the synchrotron radiation of relativistic electrons moving in the jets. The moving electrons generate two elongated clouds containing magnetic fields that, in turn, trap the electrons and lead to the synchrotron radiation.

5.6 Stellar Evolution

The time evolution of a star, which can be depicted as a path in the Hertzsprung–Russell diagram, is quite complicated because of the many physical processes that need to be taken into account. Detailed calculations, based on the numerical integration of the relevant differential equations, have provided a fairly comprehensive picture of stellar evolution. One of the primary sources of stellar energy is a series of nuclear reactions converting four protons into a helium nucleus. Since the simultaneous collision of four particles is extremely improbable, this process of converting hydrogen into helium proceeds through two different sequences of intermediate reactions, one called proton–proton chain and the other called carbon–nitrogen–oxygen cycle. In the *p–p* chain, helium is formed through deuterium and ^{3}He in the intermediate steps; this reaction is the dominant mechanism for hydrogen–helium conversion at temperatures below $\sim 2 \times 10^7$ K. In the CNO cycle, hydrogen is converted into helium through a sequence of steps involving ^{12}C as a catalyst; i.e., the amount of ^{12}C remains constant at the end of the cycle of reactions. Since the Coulomb barrier for carbon nuclei is quite high, the CNO cycle is dominant only at higher temperature. The evolution of a star like the Sun during the phase of hydrogen–helium conversion, called the "main sequence" phase, is fairly stable. The stability is essentially due to the following regulatory mechanism: Suppose the temperature decreases slightly causing the nuclear reaction rate to decrease. This will make gravity slightly more dominant, causing a contraction. Once the star contracts, the temperature will again increase, thereby increasing the rate of nuclear reactions and the pressure support. This will restore the balance. After the burning of core hydrogen ends, the core will undergo a contraction, increasing its temperature; if the star now heats up beyond the helium ignition temperature, then the burning of helium will start and stabilize the star. In principle, such a process can continue with the building up of heavier and heavier elements. But to synthesize elements heavier than He is not straightforward because He has a very high binding energy per nucleon among the light elements. Stars achieve synthesis of post-helium elements through a process known as "triple-alpha reaction" that proceeds as $^4\mathrm{He}(^4\mathrm{He},^8\mathrm{Be})\gamma$; $^8\mathrm{Be}(^4\mathrm{He},^{12}\mathrm{C})\gamma$. Once ^{12}C has been synthesized, production of heavier elements like ^{16}O, ^{20}Ne, ^{24}Mg, etc., can occur through various channels, provided temperatures are high enough, and the star can evolve through successive stages of nuclear burning. The ashes of one stage can become the fuel for the next stage as long as each ignition temperature is reached. Such stars will evolve into structures that contain concentric shells of elements. For example, a $15M_\odot$ star during its final phase can have layers of iron, silicon, oxygen, neon, carbon, helium, and hydrogen all burning at their inner edges.

The details of the above process, which occurs after the exhaustion of most of the fuel in the core, depend sensitively on the core mass. Consider, for example, a star with $M > M_\odot$. Its evolution proceeds in the following manner: Once the hydrogen is exhausted in the core, the core, containing predominantly helium, undergoes gravitational contraction. This increases the temperature of the material just beyond the core and causes renewed burning of hydrogen in a shell-like region. Soon, the core contracts rapidly, increasing the energy production and the pressure in the shell, thereby causing the outer envelope to expand. Such an expansion leads to the cooling of the surface of the star. About this time, convection becomes the dominant mechanism for energy transport in the envelope and the luminosity of the star increases due to convective mixing. This is usually called the "red-giant" phase. During the core contraction, the matter gets compressed to very high densities ($\sim 10^5$ g cm^{-3}) so that it behaves like a degenerate gas and not as a perfect gas. Once the core temperature is high enough to initiate the triple-alpha reaction, helium burning occurs at the core. Since degenerate gas has a high thermal conductivity, this process occurs very rapidly, called "helium flash." If the core had been dominated by gas pressure, such an explosive ignition would have increased the pressure and led to an expansion; but since the degeneracy pressure is reasonably independent of temperature, this does not happen. Instead the evolution proceeds as a runaway process: The increase in temperature causes an increase in the triple-alpha reaction rate, causing further increase in temperature, etc. Finally, when the temperature becomes $\sim 3.5 \times 10^8$ K, the electrons become nondegenerate; the core then expands and cools. The star has now reached a stage with helium burning in the core and hydrogen burning in a shell around the core. Soon the core is mostly converted into carbon and the reaction again stops. The process described above occurs once again, this time with a carbon-rich, degenerate core, and a helium-burning shell. This situation, however, turns out to be unstable because the triple-alpha reaction is highly sensitive to temperature. This reaction can overrespond to any fluctuations in pressure or temperature, thereby causing pulsations of the star with increasing amplitude. Such violent pulsation can eject the cool, outer layers of the star, leaving behind a hot core. The ejected envelope becomes what is known as a "planetary nebula." The above discussion assumes that the core could contract sufficiently to reach the ignition temperature for carbon burning. In low-mass stars, degeneracy pressure stops the star from reaching this phase and it ends up as a "white dwarf" supported by the degeneracy pressure of electrons against gravity. A more complicated sequence of nuclear burning is possible in stars with $M \gg M_{\odot}$. After the exhaustion of carbon burning in the core, one can have successive phases with neon, oxygen, and silicon burning in the core with successive shells of lighter elements around it. This process can proceed until ^{56}Fe is produced in the core.

The binding energy per nucleon is maximum for the ^{56}Fe nucleus; hence it will not be energetically feasible for heavier elements to be synthesized by nuclear fusion. The core now collapses violently reaching very high temperatures, about 10^{10} K. The ^{56}Fe photodisintegrates into alpha particles, and then even the alpha particles disintegrate at such high temperatures to become protons. The collapse of the core squeezes together protons and electrons to form neutrons, and the material reaches near-nuclear densities forming a "neutron star." There exist several physical processes that can transfer the gravitational energy from core collapse to the envelope, thereby leading to the forceful ejection of the outer envelope causing a "supernova explosion." A remnant with smaller mass is left behind. Numerical studies show that stars with $M > 8M_{\odot}$ burn hydrogen, helium, and carbon and evolve rather smoothly. During the final phase, such a star explodes as a supernova leaving behind a remnant that could be a white dwarf, neutron star, or black hole. It is generally believed that stars with masses in the intermediate range, $2M_{\odot} < M < 8M_{\odot}$, do not burn hydrogen and helium in degenerate cores but evolve through carbon burning in degenerate matter (for $M > 4M_{\odot}$) ending again in a supernova explosion. Stars with lower mass do not explode but end up as planetary nebulae. Low-mass stars with $M < 2M_{\odot}$ ignite helium in degenerate cores at the tip of the red-giant branch and then evolve in a complicated manner. The synthesis of elements as described above proceeds smoothly up to ^{56}Fe. Heavier elements are formed by nuclei absorbing free neutrons, produced in earlier reactions by two different processes called the "r-process" (rapid neutron-capture process) and the "s-process" (slow neutron-capture process). During the supernova explosion, a significant part of the heavy elements synthesized in the star will be ejected out into the interstellar space. A second generation of stars can form from these gaseous remnants. The initial composition of material in this second generation will contain a higher proportion of heavier elements (called "metals") compared to the first-generation stars. Both these types of stars are observed in the universe; for historical reasons, stars in the second generation are called population I stars while those in the first generation are called population II stars.

The foregoing discussion shows how stars could synthesize heavier elements, even if they originally start as gaseous spheres of hydrogen. The study of the spectra of stars allows us to determine the relative proportion of various elements present in the stars. Such studies show that population II stars are made of about 75% hydrogen and 25% helium; even population I stars consist of almost similar proportions of hydrogen and helium with a small percentage of heavier elements (~2%). It is possible to show, using stellar evolution calculations, that it is difficult for the stars to have synthesized elements in such a proportion if they originally had only hydrogen (called primordial or population III stars). Hence, such a universal composition leads us to conjecture that the primordial gas from which population II stars have formed must have been a mixture of hydrogen and helium in the ratio 3:1 by weight. Heavier elements synthesized by these population II stars would have been dispersed in the interstellar medium by supernova explosions. The population I stars are supposed to have been formed from this medium, containing a trace of heavier elements. The helium present in the primordial gas should have been synthesized at a still earlier epoch, the big bang, and cannot be accounted for by stellar evolution. Cosmological models provide an explanation for the presence of this primordial helium.

6. MATHEMATICS AND PHYSICS AND THE STRUCTURE OF THE UNIVERSE

6.1 General Relativity and Quantum Field Theory

A mathematician represents the motion of the planets of the solar system by a flow line of an incompressible fluid in a 54-dimensional phase space, whose volume is given by the Liouville measure; while a famous physicist is said to have made the statement that the whole purpose of physics is to find a number, with decimal points, etc. (!), otherwise you have not done anything (Manin, 1981). Despite such extreme points of view, it can be safely said that the relations between mathematics and physics (and astronomy) have been very productive in past epochs of the evolution of these disciplines (Barrow and Tipler, 1988).

Mathematical structures entered the development of physics, and problems emanating from physics influenced developments in mathematics. Examples are the role of differential geometry in general relativity, the dynamical theory of space and time, and the influence of quantum mechanics in the development of functional analysis and built on the understanding of Hilbert spaces. A prospective similar development occurred only recently when non-Abelian gauge theories emerged as the quantum field theories for describing fundamental-particle interactions. Yang–Mills theory found its mathematical formulation in the theory of principal fiber bundles. The understanding of anomalies in gauge theories involved the theory of families of elliptic operators and representation theory of infinite-dimensional Lie algebras and their cohomologies. Based on these developments, there are two fundamental theories in modern physics: general relativity and quantum field theory. General relativity describes gravitational forces on an astronomical scale. Quantum field theory describes the interaction of fundamental particles, electromagnetism, weak, and strong forces. The formal quantization of general relativity is leading to infinities, and the current understanding is that the mathematical machinery is missing to accomplish the unification of general relativity and quantum field theory. This is particularly due to the nonlinear character of general relativity, which is incompatible with the linearity needed for the superposition principle in quantum theory. More recently, attention has turned to the exploration of the mathematical structure of non-Abelian gauge theories and to the more ambitious attempts to construct unified theories of all the fundamental interactions of matter together with gravity. A success of such attempts may reveal new insights for structural elements of the microcosmos and macrocosmos (Schmid, 1992).

6.2 Macroscopic Dimension of Space

Of all the fundamental constants, the most familiar one is the dimension of physical space, $N = 3$. Variation of such a fundamental characteristic as the dimension N may lead to unpredictable changes of physical laws. It was Paul Ehrenfest (1880–1933) who, in 1917, was trying to answer the question of why physical space is three-dimensional. From physics, we are familiar with the analogy between Coulomb's law and Newton's law [Eqs. (1) and (2)]. In both cases, the force is $F \propto r^{-2}$. In physics these laws are treated separately. This lack of coherency obscures a profound relationship of the electromagnetic and the gravitational forces with the properties of space—in particular, with its dimension. Two properties are common to the gravitational and the electromagnetic interactions: Both are weak and long range. In modern language, this means that the mass of the gauge boson is zero in both cases, implying that the interaction radius is infinite and that the interaction constants are small, $\alpha_g, \alpha_e \ll 1$ [Eqs. (3) and (4)]. In the language of physics these properties mean that the lines of force, originating at the point where their source is located, run to infinity, not intersecting with each other, provided that no other source is present. The fact that the lines of force extend to infinity reflects the long-range character of the gravitational and electromagnetic forces; the absence of intercepts signifies that there is no reciprocal action between the lines of force, i.e., that the interactions under consideration are weak. The combination of the two properties, the weakness and the long-range action, is not characteristic for the other interactions. The force F exerted by one particle on another particle a distance r apart is proportional to the density n_i of the lines of force. Accordingly,

$$F \propto n_i = f/4\pi r^2. \tag{30}$$

The proportionality constant f in (30) is by definition equal to the product of the charges of the two particles (Coulomb's law) or of their masses (Newton's law). The denominator in (30) gives the surface area S of a sphere of radius r. For the three-dimensional space, this quantity equals $S_3 = 4\pi r^2 = a_3 r^{3-1}$. These considerations can be repeated for the more general case of an N-dimensional space. The surface area of a sphere in such a space is $S_N = a_N r^{N-1}$. Therefore, the force F_N acting in an N-dimensional space is

$$F_N = b_N/r^{N-1}. \tag{31}$$

Accordingly, the potential energy has the form

$$U_N = -b_N/(N-2)r^{N-2}, \tag{32}$$

where $N \neq 2$; for $N = 2$, the dependence is logarithmic. It should be stressed that these expressions apply only for integer N, and only for long-range forces in the quasistatic approximation, i.e., for motion in a central force field. The existence of stable orbits in a central force field in an N-dimensional space is determined by (31) and consequently by the dimension N. From mechanics, it is known that the existence of stable orbits depends on the form of the r dependence of the effective potential

$$U_{Ne} = U_N + M^2/2mr^2, \tag{33}$$

where $M^2/2mr^2$ is the centrifugal energy, M the angular momentum, and m the mass of the body moving at a distance r. A stable state is possible if the dependence $U_{Ne}(r)$ has a minimum at a value of r different from zero or infinity. Henceforth only the attractive forces are considered, for which $U_N < 0$. The results of an analysis of the function $U_{Ne}(r)$ with regard to the existence of an extremum are the following:

For $N > 4$, the dependence $U_{Ne}(r)$ has a maximum at $r \neq 0$ and a minimum at $r = 0$, which corresponds to merging of the two particles.

For $N = 4$, the dependence $U_{Ne}(r)$ is given by a monotonically decreasing function exhibiting no extrema.

For $N = 2$ and $N = 3$, the function $U_{Ne}(r)$ has a minimum at $r \neq 0$ and $r \neq \infty$.

For $N = 1$, the function $U_{Ne}(r)$ is monotonically increasing.

The existence of a minimum in the dependence $U_{Ne}(r)$ is a necessary condition for the stability of motion. For this reason, the existence and the properties of closed orbits are determined by the dimension of space, N. For $N > 4$, there is no minimum at $r \neq 0, \infty$; consequently, there are no stable closed orbits. Any motion caused by long-range forces would be of one of the following two types: Either it is infinite (the body escapes to infinity) or, otherwise, the moving body falls on a massive central body. At $N = 2$ or $N = 3$, all types of motion are possible: Infinite motion, fall onto a central body, and, notably, motion in stable, closed orbits. For $N = 1$, only finite motion is possible; a body cannot escape to infinity. In the one-dimensional case, no orbital motion occurs, and the centrifugal potential is zero ($M = 0$). The effective potential, Eq. (32), is then $U_1 = b_1 r$, and the force $F_1 =$ const. This corresponds to an infinitely deep potential well. To remove the body from this well, an infinitely large force has to be applied; this means it would be impossible for the body to escape to infinity. Hence, the degree of stability grows as the dimension N decreases. For $N > 4$, there are no analogues to planetary systems. Similar considerations in the framework of quantum mechanics have demonstrated that for $N > 4$, stable atomic systems do not exist, either. It appears that the absence of analogs of planets and atoms for $N > 4$ is a clue to the understanding of the significance of the space dimension $N = 3$, with regard to the existence of structural elements in the universe.

6.3 Microscopic Dimension of Space

In classical physics, particles have definite locations and follow exact trajectories in space-time. In quantum mechanics, wave packets propagate through space-time, their positions and velocities uncertain according to Heisenberg's uncertainty principle. In string theory, point particles are replaced by tiny loops with the result that the concept of space-time becomes "fuzzy" at scales comparable to the square root of a new fundamental constant, $\alpha' \approx (10^{-32}\ \mathrm{cm})^2$, introduced as string tension in string theory. Employing both string tension ($\alpha' \neq 0$) and quantum effects ($\hbar \neq 0$) leads to results that may change the conventional notion of space-time (Donaldson, 1996). The situation of introducing α' in string theory is similar to passing from classical to quantum mechanics by the introduction of Planck's constant $\hbar$. In string theory the one-dimensional trajectory of a particle in space-time is replaced by a two-dimensional orbit of a string. According to Heisenberg's uncertainty principle, at a momentum p one can probe a distance $x \approx \hbar/p$. However, the introduction of α' acts as if Heisenberg's uncertainty principle would have two terms, $\Delta x \geq \hbar/\Delta p + \alpha'(\Delta p/\hbar)$, where the second term reflects the fuzziness due to string theory. Then, the constant α' would be the absolute minimum uncertainty in length in any physical experiment. The consequences of this type of quantum field theory for fundamental particles and the structure of the universe are still to be discovered.

GLOSSARY

Antiparticle: A particle of opposite charge but otherwise identical to its partner. Most of the observable universe consists of particles and matter, as opposed to antiparticles and antimatter.

Asteroids: Small planetlike bodies of the solar system.

Atom: The basic building block of matter. Each atom consists of a nucleus with positive electric charge and a surrounding cloud of electrons with negative charge.

Baryon: Type of hadron, including the proton, the neutron, and the unstable hyperons (and their antiparticles).

Binary System: A system of two objects orbiting around a common center. The objects may be stars or black holes or a star and a black hole.

Black Hole: A region in which matter has collapsed to such an extent that light can no longer escape from it.

Boson: A class of particles with integer units of the basic unit of spin $h/2\pi$.

Cluster of Galaxies: An aggregate of galaxies. Clusters may range in richness from loose groups, such as the Local Group, with 10 to 100 members, to great clusters of over 1000 galaxies.

Cluster of Stars (Globular Star Cluster): An aggregate of stars. Globular star clusters contain the oldest stars in the galaxy and high-velocity stars.

Comet: A diffuse body of gas and solid particles that orbits the Sun in a highly eccentric trajectory.

Cosmic Microwave Background Radiation: Diffuse isotropic radiation whose spectrum is that of a blackbody at 3 K and consequently is most intense in the microwave region of the spectrum.

Cosmic String: A hypothetical one-dimensional, stringlike object that is made from a curvature of space.

Cosmogony: The study of the origin of celestial systems, ranging from the solar system to stars, galaxies, and clusters of galaxies.

Cosmology: The study of the large-scale structure and evolution of the universe.

Dark Matter: Matter whose presence is inferred from dynamical measurements but which has no optical counterpart.

Electron: A particle of matter with negative electric charge. All chemical properties of atoms and molecules are determined by the electrical interactions of electrons with each other and with the atomic nuclei.

Fermion: A particle with half-odd-integral units of the basic unit of spin, $h/2\pi$.

Galactic Nucleus: Innermost region of a galaxy exhibiting a concentration of stars and gas.

Galaxy: A large gravitationally bound cluster of stars that all orbit around a common center. Galaxies are basic building blocks of the universe.

Graviton: The particle that, according to wave-particle duality, is associated with gravitational waves.

Group of Galaxies: Gravitationally bound system of a few galaxies.

Group of Stars: Gravitationally bound system of a few stars (multiple stars).

HI Cloud: Cloud of cool, neutral hydrogen.

HII Region: Cloud of hot, ionized hydrogen, usually heated by a massive young hot star.

Hadrons: Particles (protons, neutrons, and mesons) that take part in strong nuclear reactions.

Halo: The diffuse, nearly spherical cloud of old stars and globular clusters that surrounds a spiral galaxy.

Hertzsprung–Russell Diagram: Plot of stellar luminosity (or absolute magnitude) against effective temperature (or color), in which the evolution of stars of different masses may be followed.

Hubble Classification of Galaxies: Organizes galaxies according to shape. They range from amorphous, relatively uniform elliptical systems to highly flattened spiral disks with prominent nuclei. It is not a classification based on evolution but one of different rates of star formation.

Hypergalaxy: A system consisting of a dominant spiral galaxy surrounded by a cloud of dwarf satellite galaxies, often ellipticals. The Milky Way and the Andromeda galaxy are hypergalaxies.

Intergalactic Gas: Matter that is present in the region between galaxies. It has been detected in considerable amounts in great clusters of galaxies, where the intergalactic gas is so hot that it emits copious amounts of x rays.

Interstellar Grains: Small needle-shaped particles in the interstellar gas with dimensions from 10^{-6} to 10^{-5} cm. They are primarily composed of silicates and strongly absorb, scatter, and polarize visible light in the far-infrared region of the spectrum.

Irregular Galaxy: A galaxy without spiral structure or smooth, spheroidal shape, often filamentary or very clumpy, and generally of low mass $[(10^7 \text{ to } 10^{10})M_\odot]$.

Leptons: Particles (neutrinos, electrons, and muons) that do not take part in strong interactions.

Local Group: Small group of 30 or so galaxies of which the Milky Way and the Andromeda galaxy are the two dominant members.

Meson: A class of strongly interacting particles with zero baryon number, among them the pi meson.

Meteorite: A solid portion of a meteoroid that has reached the Earth's surface.

Molecular Cloud: An interstellar cloud consisting predominantly of molecular hydrogen, with trace amounts of other molecules such as carbon monoxide and ammonia.

Molecule: An entity made of several atoms that share their electron clouds with each other.

Neutrino: A particle that resembles the photon, except that it interacts weakly with matter. Neutrinos come in at least three varieties, known as electron type, muon type, and tauon type.

Neutron: The uncharged particle found along with protons in atomic nuclei.

Neutron Star: Cold, degenerate, compact star in which nuclear fuels have been exhausted and pressure support against gravity is provided by the degeneracy pressure of neutrons.

Nucleons: Nuclear particles, i.e., neutrons and protons.

Observable Universe: The extent of the universe that we can see with the aid of large telescopes. Its ultimate boundary is determined by the horizon size.

Photon: A discrete unit of electromagnetic energy; the particle that, according to wave-particle duality, is associated with electromagnetic energy.

Population I: Young stars with relatively high abundances of metals, found in the disk of a galaxy, especially the spiral arms, and in dense regions of interstellar gas. The Sun is a population I star.

Population II: Old stars with relatively low abundances of metals, found in the nuclear bulge of a galaxy or in globular clusters.

Positron: The antiparticle of an electron, having positive charge but being otherwise similar.

Proton: The positively charged particle found along with neutrons in atomic nuclei.

Pulsar: A magnetized, spinning neutron star that emits a beam of radiation (radio waves and sometimes also light and x rays).

Quarks: Particles of which all hadrons are supposed to be composed.

Quasar: Luminous and compact quasistellar radio source related to violent events in the nuclei of a galaxy, believed to be powered by a massive black hole.

Red Giant: Phase in star's evolution after completion of hydrogen burning when outer layers become very extended.

Spiral Galaxy: A galaxy with a prominent nuclear bulge and luminous spiral arms of gas, dust, and young stars that wind out from the nucleus.

Star: The basic building blocks of the macroscopic universe.

Supercluster: A cluster of clusters of galaxies.

Universe: Our universe.

White Dwarf: Cool, degenerate, compact star, in which nuclear fuels are exhausted and pressure support against gravity is provided by the degeneracy pressure of electrons.

Works Cited

Alpher, R. A., Herman, R. C. (1949), *Phys. Rev.* **75,** 1089–1095.

Alpher, R. A., Herman, R. C. (1950), *Rev. Mod. Phys.* **22,** 153–219.

Alpher, R. A., Follin Jr., J. W., Herman, R. C. (1953), *Phys. Rev.* **92,** 1347–1361.

Barnett, R. M. *et al.* (Particle Data Group) (1996), "Review of Particle Physics," *Phys. Rev.* **D54,** 1–720.

Barrow, J. D. (1996), *Mon. Not. R. Astron. Soc.* **282,** 1397–1406.

Barrow, J. D., Tipler, F. J. (1988), *The Anthropic Cosmological Principle,* New York: Oxford Univ. Press.

Burbidge, E. M., Burbidge, G. R., Fowler, W. A., Hoyle, F. (1957), *Rev. Mod. Phys.* **29,** 547–560.

Carr, B. J., Rees, M. J. (1979), *Nature* **278,** 605–612.

Chandrasekhar, S. (1935), *Mon. Not. R. Astron. Soc.* **95,** 207–225.

Dirac, P. A. M. (1937), *Nature* **139,** 323.

Dirac, P. A. M. (1938), *Proc. Roy. Soc.* **A165,** 199–208.

Donaldson, S. K. (1996), *Bull. Am. Math. Soc.* **33,** 45–70.

Dyson, F. J. (1972), in: A. Salam, E. P. Wigner (Eds.), *Aspects of Quantum Theory,* Cambridge, U.K.: Cambridge Univ. Press, 213–236.

Einstein, A. (1917), *Sitzungsber. Preuss. Akad. Wiss. Berlin* 142–152. [English translation in Bernstein, J., Feinberg, G. (1986), *Cosmological Constants,* New York: Columbia Univ. Press.]

Friedmann, A. (1922), *Z. Phys.* **10,** 377–386. [English translation in Bernstein, J., Feinberg, G. (1986), *Cosmological Constants,* New York: Columbia Univ. Press.]

Friedmann, A. (1924), *Z. Phys.* **21,** 326–332. [English translation in Bernstein, J., Feinberg, G. (1986), *Cosmological Constants,* New York: Columbia Univ. Press.]

Gamow, G. (1946), *Phys. Rev.* **70,** 572–573.

Geller, M. J., Huchra, J. P. (1989), *Science* **246,** 897–903.

Gloeckler, G., Geiss, J. (1996), *Nature* **381,** 210–212.

Guth, A. H. (1981), *Phys. Rev. D* **23,** 347–356.

Haubold, H. J., Mathai, A. M. (1994), in: H. J. Haubold, L. I. Onuora (Eds.), *Basic Space Science,* New York: American Institute of Physics, 102–116.

Hawking, S. W. (1974), *Nature* **248,** 30–31.

Hoyle, F., Tayler, R. J. (1964), *Nature* **203,** 1108–1110.

Kaku, M. (1993), *Quantum Field Theory,* New York: Oxford Univ. Press.

Kolb, E. W., Turner, M. S. (1988), *The Early Universe: Reprints,* Redwood City, CA: Addison-Wesley Publishing Company.

Lemaître, G. (1927), *Ann. Soc. Sci. Bruxelles* **A47,** 49. [English translation in Bernstein, J., Feinberg, G. (1986), *Cosmological Constants,* New York: Columbia Univ. Press.]

Manin, Yu. I. (1981), *Mathematics and Physics,* Boston: Birkhäuser.

North, J. D. (1965), *The Measure of the Universe: A History of Modern Cosmology,* Oxford: Clarendon Press.

North, J. D. (1995), *The Norton History of Astronomy and Cosmology,* New York: W. W. Norton and Company.

Pais, A. (1986), *Inward Bound: Of Matter and*

Forces in the Physical World, New York: Oxford Univ. Press.

Peebles, P. J. E. (1966), *Astrophys. J.* **146,** 542–552.

Penzias, A. A., Wilson, R. W. (1965), *Astrophys. J.* **142,** 419–421.

Rees, M. J., Ostriker, J. P. (1977), *Mon. Not. R. Astron. Soc.* **179,** 541–559.

Schmid, R. (1992), *SIAM Rev.* **34,** 406–425.

Sciama, D. W. (1953), *Mon. Not. R. Astron. Soc.* **113,** 34–42.

Shaver, P. A., Wall, J. V., Kellerman, K. I., Jackson, C. A., Hawkins, M. R. S. (1996), *Nature* **384,** 439–441.

Silk, J. I. (1977), *Astrophys. J.* **211,** 638–648.

Smoot, G., Davidson, K. (1993), *Wrinkles in Time: The Imprint of Creation,* London: Abacus.

Tytler, D., Fan, X.-M., Burles, S. (1996), *Nature* **381,** 207–209.

Wagoner, R. V., Fowler, W. A., Hoyle, F. (1967), *Astrophys. J.* **148,** 3–49.

Weisskopf, V. F. (1975), *Science* **187,** 605–612.

White, S. D., Rees, M. J. (1978), *Mon. Not. R. Astron. Soc.* **183,** 341.

Whittaker, E. (1951, 1953/1989), *A History of the Theories of Aether and Electricity,* New York: Dover Publications.

Wilczek, F. (1996), *Nature* **380,** 19–20.

Further Reading

Beer, A. (1898/1961), *A Short History of Astronomy: From the Earliest Times through the Nineteenth Century,* New York: Dover Publications.

Bernstein, J., Feinberg, G. (Eds.) (1986), *Cosmological Constants: Papers in Modern Cosmology,* New York: Columbia Univ. Press.

Celnikier, L. M. (1989), *Basics of Cosmic Structures,* Gif-sur-Yvette, France: Editions Frontières.

Chandrasekhar, S. (1990), *Truth and Beauty,* Chicago: The University of Chicago Press.

Clayton, D. D. (1983), *Principles of Stellar Evolution and Nucleosynthesis,* Chicago: The University of Chicago Press.

Collins II, G. W. (1989), *The Fundamentals of Stellar Astrophysics,* New York: W. H. Freeman and Company.

Combes, F., Boisse, P., Mazure, A., Blanchard, A. (1995), *Galaxies and Cosmology,* Berlin: Springer.

Danby, J. M. A., Kouzes, R., Whitney, C. (1995), *Astrophysics Simulations,* The Consortium for Upper-Level Physics Software, New York: Wiley.

Kolb, E. W., Turner, M. S. (1990), *The Early Universe,* Redwood City, CA: Addison-Wesley Publishing Company.

Lang, K. R. (1980), *Astrophysical Formulae: A Compendium for the Physicist and Astrophysicist,* New York: Springer-Verlag.

Lawrie, I. D. (1994), *A Unified Grand Tour of Theoretical Physics,* Bristol: Institute of Physics Publishing.

Misner, C. W., Thorne, K. S., Wheeler, J. A. (1973), *Gravitation,* New York: W. H. Freeman and Company.

Padmanabhan, T. (1993), *Structure Formation in the Universe,* Cambridge, U.K.: Cambridge Univ. Press.

Pagels, H. R. (1982), *The Cosmic Code,* New York: Bantam Books.

Pagels, H. R. (1985), *Perfect Symmetry: The Search of the Beginning of Time,* New York: Bantam Books.

Pagels, H. R. (1988), *The Dreams of Reason,* New York: Bantam Books.

Peebles, P. J. E. (1980), *The Large-Scale Structure of the Universe,* Princeton, NJ: Princeton Univ. Press.

Peebles, P. J. E. (1993), *Principles of Physical Cosmology,* Princeton, NJ: Princeton Univ. Press.

Rowan-Robinson, M. (1996), *Cosmology,* Oxford: Clarendon Press.

Rozental, I. L. (1988), *Big Bang—Big Bounce: How Particles and Fields Drive Cosmic Evolution,* Berlin: Springer.

Stevens, C. F. (1995), *The Six Core Theories of Modern Physics,* Cambridge, MA: The MIT Press.

Thorne, K. S. (1994), *Black Holes and Time Warps,* New York: W. W. Norton and Company.

Trainor, L. E. H., Wise, M. B. (1981), *From Physical Concept to Mathematical Structure: An Introduction to Theoretical Physics,* Toronto: University of Toronto Press.

Weinberg, S. (1972), *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity,* New York: Wiley.

Vol. 23 – PLATE 1. It must be related to geometry, and it may have been a window to the stars: Stonehenge in Wiltshire, England, still the focus of hot debates.

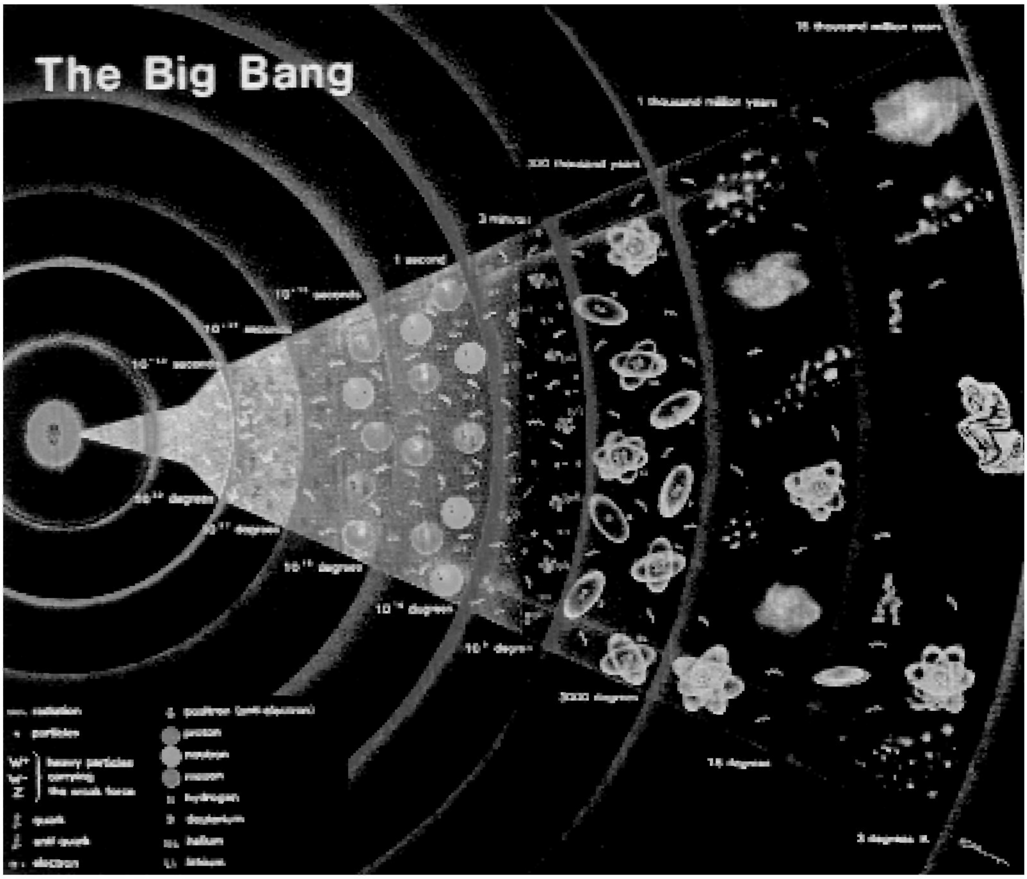
The Big Bang
15 thousand million years
1 thousand million years
300 thousand years
3 minutes
1 second
3000 degrees
3 degrees K
radiation
particles
heavy particles carrying the weak force
quark
anti-quark
electron
positron (anti-electron)
proton
neutron
meson
hydrogen
deuterium
helium
lithium

Vol. 23 – PLATE 2. The big bang: from the origin of the universe to its present epoch (Smoot and Davidson, 1993).

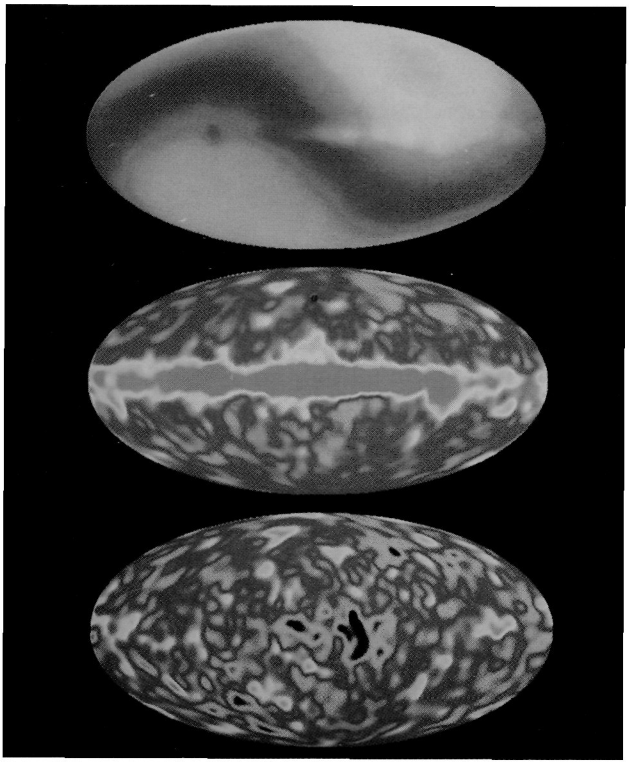

Vol. 23 – PLATE 3. Three full-sky maps made by the COBE satellite DMR instrument (Smoot and Davidson, 1993). Top: The dipole anisotropy caused by the Earth's motion relative to the cosmic background radiation (hotter in the direction we are going, cooler in the direction we are leaving). Center: The dipole-removed sky showing the emission from the plane of the galaxy – the horizontal red strip and the large-scale ripples in space-time. Bottom: A map of the wrinkles in time.

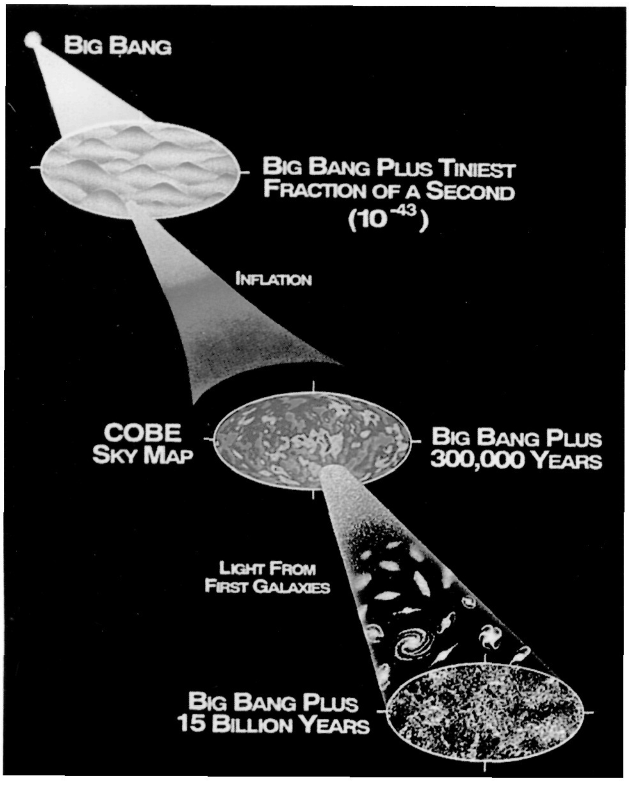
Big Bang
Big Bang Plus Tiniest Fraction of a Second (10^-43)
Inflation
COBE Sky Map
Big Bang Plus 300,000 Years
Light From First Galaxies
Big Bang Plus 15 Billion Years

Vol. 23 – PLATE 4. The big bang, with inflation producing space and the ripples in space-time mapped by COBE, and eventually evolving to become stars and galaxies and clusters of them (Smoot and Davidson, 1993).

VACUUM ELECTRONICS

T. Koryu Ishii, *Department of Electrical and Computer Engineering, Marquette University, Milwaukee, Wisconsin, U.S.A.*

Introduction 83
1. **Vacuum Diodes** 83
1.1 Detector Diodes 83
1.2 Power Rectifiers 85
1.3 Switching Tubes 85
1.4 Photodiodes 86
2. **Vacuum Triodes** 86
2.1 Audio and rf Triodes 86
2.2 Microwave Triodes 87
2.3 Field-Emission Triodes 88
3. **Vacuum Tetrodes** 89
3.1 Audio and rf Vacuum Tetrodes 89
3.2 Microwave Tetrodes 89
4. **Vacuum Pentodes** 89
4.1 Audio and rf Vacuum Pentodes 89
4.2 Microwave Pentodes 90
5. **Klystrons** 90
5.1 Multicavity Klystrons 90
5.2 Reflex Klystrons 92
6. **Magnetrons** 93
6.1 Heating Magnetrons 93
6.2 Radar Magnetrons 94
7. **Gyrotrons and Fast-Wave Tubes** 96
7.1 Gyrotrons 96
7.2 Novel Microwave Tubes 96
8. **Traveling-Wave Tubes (TWTs)** 97
8.1 O-Type TWT 97
8.2 M-Type TWT 98
9. **Display Tubes** 98
9.1 Cathode-Ray Oscilloscopes 98
9.2 Monitor Tubes 99
10. **Imaging Tubes** 100
10.1 Vidicon Tubes 100
10.2 Thermicon Tubes 101
11. **Scientific Vacuum Electronics** 101
11.1 Electron Microscopes 101
11.2 Particle Accelerators 103
Glossary 103
Works Cited 103
Further Reading 103

INTRODUCTION

Vacuum electronics is the field of electronics associated with a vacuum medium within a gas-tight envelope. The vacuum medium considered here is an evacuated region in which electron motion is essentially unaffected by the presence of residual gas or vapor. In this way vacuum electronics is different from gaseous electronics.

Vacuum electronics has a wide range of uses and capabilities. It is used for the generation, amplification, and detection of electromagnetic waves in many applications including, among other things, radio and video communications, navigation, and remote control and sensing. It can also be used for collecting and displaying video images. This versatility is why vacuum electronics is quite commonly employed in both industrial and consumer electronic products.

Two advantages of vacuum electronics over other types of devices are

1. high power-handling capacity and
2. the ability to operate in harsh environmental conditions, such as extreme temperatures, where most solid-state electronics cease to function.

The topics to be discussed in rest of this article include various kinds of electron tubes and specific vacuum electron devices for scientific investigation.

1. VACUUM DIODES

1.1 Detector Diodes

A vacuum tube consisting of a cathode and an anode is called a vacuum diode. An example of the cross-sectional view of a

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

vacuum diode with a coaxial cylindrical structure is shown in Fig. 1. The anode is often called the plate. When the plate is positive with respect to the cathode, electrons emitted from the cathode may reach the plate, and electric conduction occurs. When the plate is negative with respect to the cathode, electrons emitted from the cathode are pushed back toward the cathode, and few electrons, if any, reach the plate. Therefore, under such bias there is no electric conduction between the plate and the cathode. This property of the vacuum diode can be used for converting ac to dc. The process of converting ac to dc is called rectification. Similarly, the conversion of rf signals to dc is known as detection. Vacuum detector diodes are used in harsh environments in which solid-state diodes cannot function properly, or for cases where different functional characteristics of detection than solid-state diodes are required.

Vacuum diodes can be either planar or cylindrical. In a planar vacuum diode, both the anode and cathode are flat, or planar, and are parallel to each other. In a cylindrical vacuum diode, both the anode and the cathode are cylindrical in shape and coaxial to each other as shown in Fig. 1. Usually the center cylinder is the electron-emitting cathode, and the outer cylinder is the anode, as shown in the figure.

When the anode is positive with respect to the cathode under constant cathode temperature, the anode current increases with the anode voltage, up to a certain point. If the anode voltage is increased beyond this point, the anode current will increase no further. This condition is called saturation. The anode current under saturation is called the saturation current, or the temperature-limited current. The saturation current is limited by the finite rate of electron emission from the cathode at a given cathode temperature. The saturation current I_s is given by

$$I_s = 1.2 \times 10^6 A T^2 \epsilon^{-e\phi/kT}, \tag{1}$$

where A is the total emitting area of cathode in square meters, T is the cathode temperature in kelvins, $e\phi$ is the work function of the heated cathode in joules, and k is Boltzmann's constant, 1.38×10^{-23} J/K. If the anode voltage is low and the vacuum diode is not saturated, the anode current is called the space-charge current, or the space-charge limited current. The current is then primarily determined by the speed of the emitted electrons, which is controlled by the anode voltage. The space-charge current observed at the anode is given by the Child–Langmuir–Schottky equation (IEEE, 1972),

$$I_p = GV_p^{3/2}, \tag{2}$$

where I_p is the space-charge current, G is a proportionality constant called the perveance, and V_p is the anode voltage. The perveance is determined by the electrode geometry and the material of both the anode and the cathode. When a vacuum diode is used for rf detection, the operating rf signals are generally small, and so the vacuum diode will not enter the saturation condition. The vacuum diode is then operating under space-charge current conditions. The vacuum detector diode operates based on the principle that the diode conducts only for the positive half-cycles of the rf signals to be detected. This principle fails at high frequencies. When the electron transit time from the cathode to the anode, τ, is small in comparison to the period of the operating rf signals, T, the principle works well. If the electron transit time is comparable to or significantly large in comparison to the period of the rf signals, the diode does not work well in detection.

The minimum electron transit time, $\tau_{\min}$, is

$$\tau_{\min} = d/u_{\max} = d\sqrt{m/2ev}, \tag{3}$$

where d is the distance between the cathode

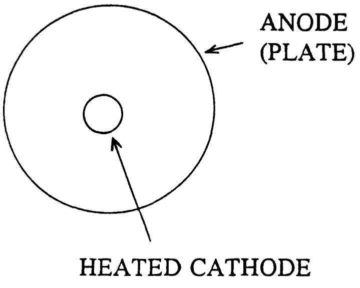

FIG. 1. Cross-sectional view of a vacuum diode with coaxial cylindrical structure.

and the anode, $u_{\max}$ is the maximum speed of the electrons (m/s) driven by the rf voltage v that is to be detected, m is the mass of an electron (kg), and e is the electric charge of an electron (C). For the satisfactory operation of a vacuum detector diode,

$$\tau_{\min} \ll T. \tag{4}$$

As a "rule of thumb,"

$$\tau_{\min} \leq 0.1T. \tag{5}$$

This means that the maximum operating frequency of a given vacuum detector diode is

$$f_{\max} \leqq \frac{1}{10\tau_{\min}} = \frac{\sqrt{2ev}}{10d\sqrt{m}}. \tag{6}$$

If the vacuum diode is constructed in such a way that the cathode and the anode are plane parallel at a distance d (m) apart, the electron current density due to the anode voltage V_a is (Nakamura, 1969)

$$J = 2.33 \times 10^{-6} V_a^{3/2}/d^2. \tag{7}$$

If the structure is cathode-centered coaxial, then the electron current density is (Nakamura, 1969)

$$J = 2.33 \times 10^{-6} V_a^{3/2}/\beta_1^2 r_a^2, \tag{8}$$

where r_a is the radius of the anode cylinder. The symbol β_1 represents the Langmuir constant. It ranges from 0 to $\sqrt{1.1}$ for an r_a/r_c ratio of 1 to 100, with $\beta_1 = \sqrt{1.1}$ at $r_a/r_c = 3$.

1.2 Power Rectifiers

For high-current or high-power rectifiers, vacuum diodes use a small amount of argon gas or mercury vapor. In the positive half-cycle of rectification, electrons from the heated cathode strike atoms of gas or vapor and ionize them. Thus, multiple-impact ionization takes place. The resultant electrons are driven toward the anode while the positive ions that are generated move toward the cathode. This multiplication of charge carriers makes for a larger forward current than that of a purely vacuum diode of the same size and structure. In the reverse cycle, nearly all charge-carrier movement is suppressed. Therefore, there is no reverse current in the reverse cycle.

There are high-current, low-voltage power rectifiers and low-current, high-voltage power rectifiers. In high-current, low-voltage power rectifiers, the cathode is made of tungsten or thoriated tungsten, and the anode is made of carbonated lead. The gas or vapor pressure inside the rectifier is relatively high, as much as 20 to 30 mm Hg. This high gas or vapor pressure aids in high-temperature (2400–2800 K) operation of the cathode (Nakamura, 1969). The high cathode temperature results in a large amount of electron emission, a large yield of multiple-impact ionization, a large forward current, and a small internal voltage drop across the rectifier tube. However, the large concentration of charge carriers per unit volume makes the inverse breakdown voltage relatively low during the reverse cycle, approximately 0.3 kV (Nakamura, 1969).

For low-current, high-voltage applications, mercury rectifiers are used. This type of vacuum power-rectifier diode contains a few tiny balls of mercury in it. The cathode is either a direct-heating tungsten cathode or an indirect-heating oxide cathode ($BaSrO_2$ coated on Ni), which has better electron emission efficiency. The anode is made of either carbonated metal or lead. The vapor pressure is kept low in order to obtain a high reverse breakdown voltage of 5 to 20 kV (Nakamura, 1969).

1.3 Switching Tubes

Vacuum diodes are sometimes used for turning on and off high-power microwave waveguides. This type of vacuum diode contains a small amount of gas, and there is no cathode or anode. Instead, there are two sharp, pointed metallic electrodes as illustrated in Fig. 2. The vacuum diode is mounted across the waveguide that is to be turned on and off. When a surge of high microwave power is present at the spark gap, the electrical spark short circuits the waveguide. Thus, high microwave power cannot propagate past the vacuum diode. When the microwave power is small enough, there is no spark at the spark gap, so that small microwave power can propagate beyond the diode. This type of vacuum diode is known as

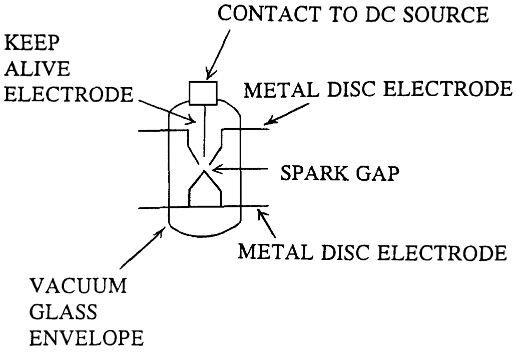

FIG. 2. Schematic diagram of a T-R tube (transmit–receive tube).

the transmit–receive tube, or the T-R tube (Reich *et al.*, 1953).

In order to increase sparking sensitivity to microwaves, the electron supply is kept alive by a dc discharge into the inside of one of the truncated cone-shaped electrodes as illustrated in Fig. 2. T-R tubes are used for protecting a radar receiver from the high microwave power of its own transmitter. When a T-R tube sparks, the waveguide is short circuited. When the T-R tube stops sparking, the waveguide is connected for transmission.

When a quarter-wavelength waveguide is attached to a T-R tube, the input impedance of the waveguide loaded with the T-R tube becomes infinity, effectively an open circuit. When the T-R tube is not sparking, the waveguide input impedance is zero, effectively a short circuit. This combination of a quarter-wavelength waveguide and a T-R tube is called the anti-transmit–receive tube, or the ATR tube.

1.4 Photodiodes

Photodiodes are used for converting light signals into electrical signals. They consist of an anode and a photocathode. The photocathode emits electrons when illuminated by light. The emitted electrons are attracted to the anode, forming the anode current. The number of photoelectrons emitted is proportional to the intensity of the incident light. The magnitude of the anode current, which is called the photocurrent, is proportional to the number of photoelectrons that reach the anode per unit time.

The volt-ampere characteristic of a generic photodiode under various levels of light intensity is shown in Fig. 3. When the anode voltage is low, there is a space-charge effect. The photoelectron emission is somewhat suppressed by the space charge of photoelectrons that have already been emitted. Therefore, for the lower anode voltage range, the anode current is proportional to the anode voltage.

Before the saturation, the anode current signal is calculated by

$$\Delta i_A = \frac{\partial i_A}{\partial I}\Delta I\left(1 + \frac{\partial i_A}{\partial v_A}Z\right)^{-1} \tag{9}$$

where i_A is the anode current (A), I is the light intensity (W), Z is the anode impedance (Ω), and v_A is the anode voltage.

2. VACUUM TRIODES

2.1 Audio and rf Triodes

If a metallic grid is inserted between the hot cathode and the anode of a vacuum diode as shown in Fig. 4, a vacuum triode is formed. A negative voltage is given to the metallic grid and the grid voltage controls the flow of electrons from the cathode to the anode by the repulsion between the negatively charged grid and electrons. The amount of the repulsive force is controlled by the amount of the negative voltage on the grid. Therefore the metallic grid is called the control grid. Vacuum triodes described here are used for the generation, amplification, and control of audio and rf signals. In most audio and rf vacuum triodes in practical use,

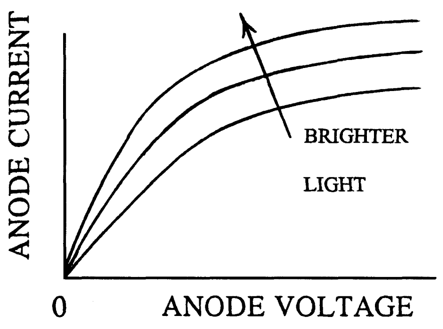

FIG. 3. An anode volt-ampere curve of a generic photodiode under various levels of light intensity.

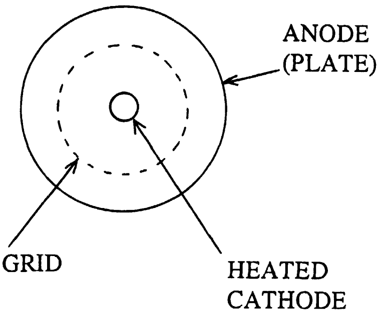

FIG. 4. A cross-sectional view of a vacuum triode with a coaxial structure.

for a given geometry of electrode structure and operating voltages, the electron transit time between electrodes is negligibly small in comparison to the operating period of the signal frequency. Interelectrode capacitance and lead inductance are also negligibly small in comparison to the lumped-parameter circuit components in the rest of the circuit. For a given geometry of a vacuum triode structure, if the operating frequency is increased, then the electron transit time among the electrodes and the effects of interelectrode capacitances and lead inductances become significant. At this point in operating frequency, proper function of the triode ceases.

Under normal operation in audio and rf, the performance of a vacuum triode tube is represented by the transconductance g_m, the internal resistance r_i, and the amplification factor μ. The transconductance g_m is defined as

$$g_m = \left.\frac{\partial I_A}{\partial V_g}\right|_{V_A=\text{constant}}, \tag{10}$$

where I_A is the anode current, V_g is the grid voltage, and V_A is the anode voltage. The internal resistance r_i is defined as

$$r_i = \left.\frac{\partial V_A}{\partial I_A}\right|_{V_g=\text{constant}}, \tag{11}$$

The amplification factor μ is defined as

$$\mu = \left.\frac{\partial V_A}{\partial V_G}\right|_{I_A=\text{constant}}. \tag{12}$$

When a vacuum triode is operated under space-charge limited condition before saturation, the anode current I_A, the grid voltage V_g, and the anode voltage V_A are related by

$$I_A = G(V_g + V_A/\mu)^{3/2}, \tag{13}$$

where G is the space-charge limited current coefficient (perveance) determined by the tube geometry and the space-charge density; then (Nakamura, 1969)

$$g_m = \tfrac{3}{2}G(V_g + V_A/\mu)^{1/2} = \tfrac{3}{2}G^{2/3}I_A^{1/2} \tag{14}$$

$$r_i = \tfrac{2}{3}\mu G^{-2/3}I_A^{-1/3}, \tag{15}$$

$$\mu = g_m r_i. \tag{16}$$

When the grid is negatively biased, there is a value of the grid voltage at which the anode current becomes zero. The grid voltage at that point is called the grid cutoff voltage and is given by

$$V_g = -V_A/\mu. \tag{17}$$

2.2 Microwave Triodes

For microwave-frequency operation of vacuum triodes, the lead inductance must be reduced, and the electron transit time between the cathode and anode must be much smaller than the period of the microwave operating frequency. To meet these requirements, the electrodes are of a planar shape and connected to the external circuit as illustrated in Fig. 5. Such a vacuum tube with sealed disc electrodes is called a lighthouse tube because its shape resembles that of a lighthouse.

The electron transit time between the anode and cathode is approximately given by

$$\tau \approx d(2eV_a/m)^{-1/2}, \tag{18}$$

where V_a is the potential difference, in volts, between the anode and the cathode; e is the charge of an electron, 1.6021×10^{-19} C; m is the mass of an electron, 9.1091×10^{-31} kg; and d is the distance between the anode and cathode in meters. It is desirable that τ be significantly smaller than the operating microwave period. As a general rule,

$$\tau < T/10, \tag{19}$$

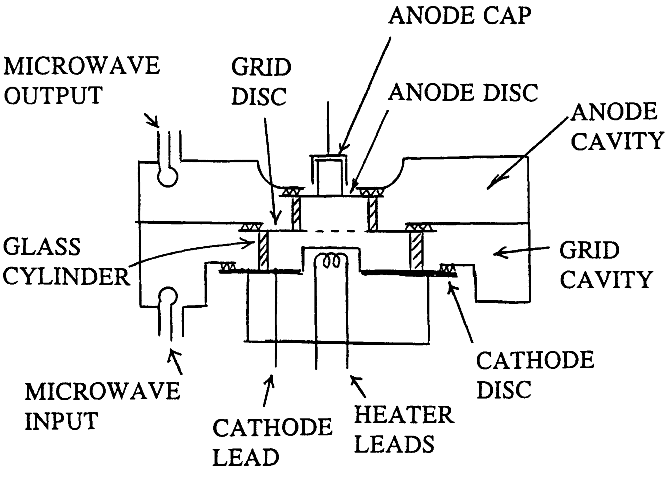

FIG. 5. A disc sealed lighthouse tube and its cavity resonator circuits.

where T is the period of the operating microwave frequency. Equations (10) through (17) also apply to microwave tubes.

2.3 Field-Emission Triodes

Field-emission triodes are fabricated using vacuum microelectronic techniques. The triodes are like transistors, with thousands of monolithic units fabricated simultaneously on a wafer.

Differing from transistors, the active region of the field-emission triode is a vacuum, and the materials used are not semiconductors. The device is made up of layers of metal and insulators as illustrated in Fig. 6 (Brodie, 1989). Since the materials are metals and insulators, the interelectrode distance can be reduced to submicrons instead of microns as in the case of semiconductor devices. In addition, electron mobility in a vacuum is much higher than electron mobility in semiconductors. For these two reasons the field-emission vacuum triode is well suited for high-speed and high-frequency applications.

The field-emission triode operates in the following manner. A bias voltage is applied to the metallic cathode and anode, setting up an electric field in the gap between them. The field is very strong because the distance from the cathode to the anode is small, approximately 0.5 μm. The electric-field intensity is even stronger at the pointed tip of the cathode, on account of its shape. This high field strength at the cathode tip causes field emission of electrons, which travel to the metal anode, forming the anode current. The electron flow is controlled by the metal gate. The metal gate has the same function as a grid in a conventional vacuum triode.

The insulator layers can be TeOS, SiO_2, or intrinsic Si. The field-emission metal can be molybdenum or some form of coated molybdenum. The metal for the base, gate, and anode may be W or TiW. The operating anode voltage can be as high as 20 V, though operation may begin with less than 5 V of anode voltage. The electron transit time from the cathode to the anode, ~0.5 μm, is on the order of picoseconds.

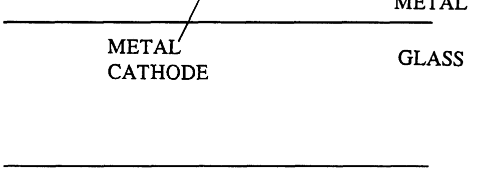

FIG. 6. A cross-sectional view of a monolithically developed field-emission vacuum triode.

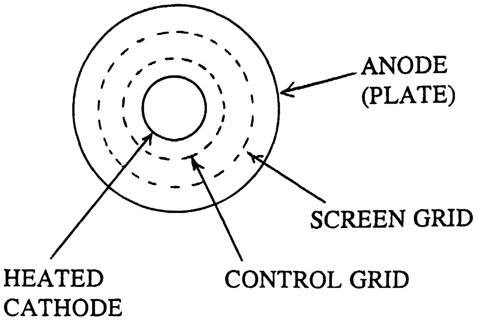

FIG. 7. A cross-sectional view of a vacuum tetrode with a coaxial cylindrical structure.

3. VACUUM TETRODES

3.1 Audio and rf Vacuum Tetrodes

If one more grid is added between the control grid and the anode of a vacuum triode, as illustrated in Fig. 7, the additional grid is called the screen grid, and the vacuum tube is called the vacuum tetrode. The purpose of the screen grid is to reduce the interelectrode capacitance between the grid and the anode. The screen grid is positively biased with a dc voltage slightly lower than the anode voltage, and the audio or rf signal on the screen grid is grounded through a bypass capacitor. The presence of the screen grid makes the internal resistance of the tube, r_i, larger.

As seen from Eq. (16), the larger r_i makes for a proportionately larger amplification factor μ. The reduced interelectrode capacitance is helpful for reducing interelectrode feedback so that the tetrode can be operated as a stable voltage amplifier at higher frequency. The volt-ampere characteristic is expressed by Eq. (13) with controlled G due to the screen grid. A specific volt-ampere curve of a specific vacuum tube can be obtained from the manufacturer of the tube.

3.2 Microwave Tetrodes

A disc-sealed–type microwave lighthouse-shaped vacuum tetrode is illustrated in Fig. 8. For microwave amplification, the input microwave cavity resonator is attached across the control grid disc and the cathode cylinder. The output microwave cavity is built around the anode block and the screen grid disc in a manner similar to the one shown in Fig. 5. A microwave tetrode is usually a medium-power microwave amplifier. Cooling fins are often attached to the anode block for forced-air cooling. The interelectrode distance is made as small as possible to reduce the electron transit time. The electron transit time should be significantly small in comparison with the period of operating frequency. The biasing scheme for the anode, screen grid, control grid, and cathode is identical to that of audio and rf tetrodes.

4. VACUUM PENTODES

4.1 Audio and rf Vacuum Pentodes

A cross-sectional view of a vacuum pentode with a coaxial cylindrical structure is illustrated in Fig. 9. A coarse metallic grid has been added between the anode and the screen grid as shown. This additional grid is called the suppressor grid. The suppressor grid is usually biased to the same dc electrical potential as the cathode to suppress sec-

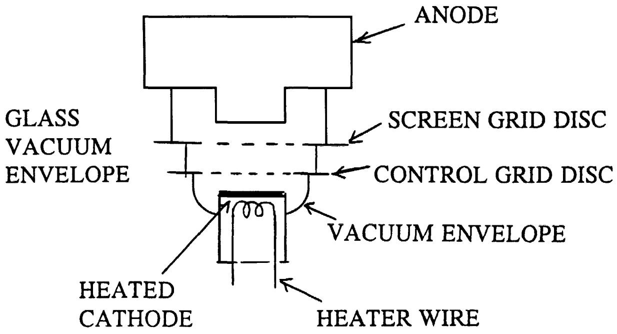

FIG. 8. A schematic diagram of a microwave tetrode.

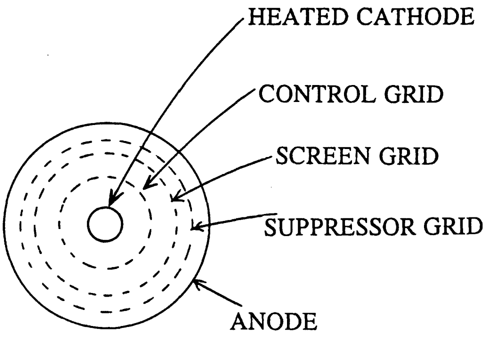
HEATED CATHODE
CONTROL GRID
SCREEN GRID
SUPPRESSOR GRID
ANODE

FIG. 9. A cross-sectional view of an audio and rf vacuum pentode having a coaxial cylindrical structure.

ondary electron emission from the anode. The volt-ampere curve of a pentode is represented by Eq. (13) with a modified value of *G* due to the presence of both the screen grid and the suppressor grid. Specific volt-ampere curve of a specific vacuum tube can be obtained from the manufacturer of the tube.

4.2 Microwave Pentodes

Disc-sealed lighthouse-type microwave pentodes are used in microwave power amplification, and their configuration is similar to the microwave vacuum tetrode shown in Fig. 8, except that a suppressor grid disc is added in front of the anode block. The biasing requirements of a microwave pentode are identical to those of audio and rf vacuum pentodes. As other vacuum tubes, the interelectrode distance is made as small as possible to minimize the electron transit-time effect.

5. KLYSTRONS

5.1 Multicavity Klystrons

A schematic diagram of a multicavity klystron is shown in Fig. 10. A multicavity klystron is a microwave power amplifier. Electrons are emitted from the electron gun, which shoots out an electrostatically focused electron beam. This electron beam is further focused to stay on course by a longitudinally applied magnetic flux density *B*. At the end opposite the electron gun, the electron beam is terminated by a collector. On the way from the electron gun to the collector, the electron beam goes through a series of microwave cavity resonators. Each cavity resonator is doughnut shaped, and a pair of grids are attached in the middle as illustrated in Fig. 10. These cavities resonate at

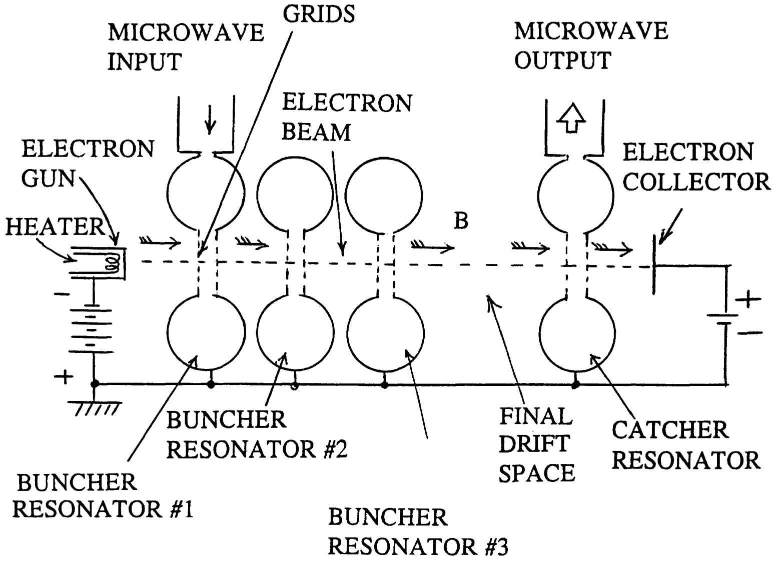
MICROWAVE INPUT
GRIDS
ELECTRON BEAM
MICROWAVE OUTPUT
ELECTRON GUN
HEATER
ELECTRON COLLECTOR
B
+
−
+
−
BUNCHER RESONATOR #2
BUNCHER RESONATOR #1
BUNCHER RESONATOR #3
FINAL DRIFT SPACE
CATCHER RESONATOR

FIG. 10. A schematic diagram of a multicavity klystron.

the operating frequency of microwave signals to be amplified.

The microwave input signals are fed to the first cavity resonator through a waveguide as shown in Fig. 10. This first cavity resonator is called the first buncher resonator. Due to the input excitation, a microwave input voltage

$$v = v_1 \sin\omega t \tag{20}$$

develops across the grid of the first buncher resonator. In this equation, v_1 is the amplitude of the microwave voltage across the grid of the resonator and $\omega = 2\pi f$ is the angular frequency of the signal. The electron beam is accelerated by the dc voltage V_a and the microwave voltage v across the grid of the first buncher resonator. The speed of electrons just before passing through the first grid is

$$u_0 = \sqrt{2eV_a/m}, \tag{21}$$

where e is the charge of an electron and m is the mass of an electron. But after passing through the first pair of grids,

$$u \approx u_0[1 + (Kv_1/2V_a)\sin\omega t], \tag{22}$$

where K is the beam coupling coefficient. This coupling coefficient is the ratio of microwave-frequency current in the cavity resonator to the microwave-frequency electron current in the beam. It is calculated to be

$$K = \frac{\sin(\omega l/2u)}{\omega l/2u}, \tag{23}$$

where l is the distance between the two grids in the first buncher resonator. As seen from Eq. (21), u_0 is a dc velocity, and u is a microwave-modulated velocity. This process is called velocity modulation, and the ratio $Kv_1/2V_a$ is called the depth of velocity modulation.

After the velocity modulation in the first buncher cavity, the electrons enter a drift space where they experience no acceleration or deceleration. The cavity resonators are electrically connected to each other and are therefore at the same dc voltage. Acceleration and deceleration of the electrons occur only between the grids of cavity resonators. The process of velocity modulation is repeated at the grids of the second and third buncher. These cavities are idler cavities. They are not directly connected to the input circuit. The cavities pick up tuned frequency voltage across the grids out of the velocity-modulated and weakly bunched electron beam. The induced voltage across the grids gives the velocity modulation back to the electron beam. By this time the depth of velocity modulation becomes deep. After passing through the third buncher resonator, the electrons enter the final drift space. With deeper velocity modulation, the electron bunching in the beam becomes tighter. The velocity modulation produces faster electrons and slower electrons. In the drift space accelerated electrons catch up with slower electrons that left the last buncher resonator grids earlier. Thus, the velocity modulation of the electron beam is transformed into density modulation.

The electric current associated with the velocity-modulated electron beam in the drift space is given by (Ishii, 1990)

$$I \approx I_0\left(1 + \frac{MKv_1}{2V_a}\frac{l_d}{u_0}\,\omega\cos\omega t\right), \tag{24}$$

where V_a is the anode voltage, I_0 is the dc beam current, l_d is the drifted distance in the last drift space, and M is the multiplication factor due to the number of buncher cavity stages. This is not primarily or directly related to the number of bunches created in the drift space. Both l_d and u_0 are adjusted so that the catcher-cavity resonator grids are located where the tightest bunch is. The density-modulated electron beam causes a large microwave voltage across the grids of the catcher cavity by electrostatic induction. In Eq. (24), the quantity $Kv_i l_d \omega/2V_a u_0$ is called the bunching parameter.

The induced microwave voltage across the grids of the catcher cavity resonator is

$$v_c = (MK^2 v_1 I_0/2V_a)R_c, \tag{25}$$

where R_c is the resonance resistance of catcher cavity resonator circuit at its resonant frequency. Therefore, the voltage gain of this klystron amplifier is

$$A_n = \frac{v_c}{v_1} = \frac{MK^2 I_0 R_c}{2V_a}. \tag{26}$$

A phasor voltage gain for large signals is (Ishii, 1990)

$$A_v = 2\frac{2MKZ_cI_0J_1(X_b)}{v_1}\epsilon^{-j(\theta-\pi/2)}, \qquad (27)$$

where Z_c is the catcher cavity impedance and X_b is the bunching parameter. $J_1(X_b)$ is the Bessel function of the first kind of order one with argument X_b. The quantity $\theta - \pi/2$ is the phase delay between the catcher gap voltage and the input voltage, including the electron transit delays between the multiple buncher stages and the final drift spaces and grid gaps.

The power gain for small-signal operation at the resonant frequency is

$$A_p = \frac{\omega^2MK^3R_cI^2l_d}{4G_gV_a^2u_0^2}. \qquad (28)$$

Multicavity klystrons are used for microwave power amplification, radar transmitters, UHF television transmitters, particle-accelerator transmitters, and industrial microwave-heater transmitters. The volt-ampere characteristic curve of a klystron is represented by Eq. (24) with both u_0 and I_0 functions of v_a as seen from Eq. (21). Specific volt-ampere curves of a specific klystron can be obtained from the specific manufacturer of the tube.

5.2 Reflex Klystrons

A reflex klystron is a low-power microwave-frequency oscillator vacuum tube. A schematic diagram of a reflex klystron is shown in Fig. 11. It operates on the basis of the principle of electron-beam velocity modulation in a manner similar to the klystron explained in Sec. 5.1. The electron beam is emitted from the electron gun, accelerated by the anode cavity, and then velocity modulated at the grid gap by a microwave signal voltage originating from noise already present. The noise includes thermal noise, shot noise, flicker noise, and partition noise. The microwave-frequency voltage is amplified by the high Q of the cavity resonator.

Since the repeller is negatively biased, the electron beam is repelled back to the grids of the cavity resonator. The velocity-modulated electrons are bunched on their return trip to the resonator gap. For the bunched returning electrons, the cavity resonator acts as the catcher cavity in multicavity klystrons. If the timing is correctly adjusted, the bunched returning electrons induce a large microwave voltage in phase with the existing microwave grid signal. Thus, positive feedback is accomplished and microwave oscillation begins.

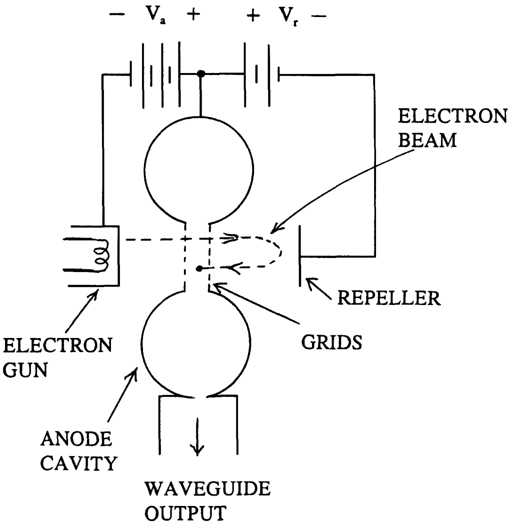

FIG. 11. A schematic diagram of a reflex klystron.

For positive feedback to occur, the number of electron transit cycles in the repeller space must be

$$N = n + \tfrac{3}{4}, \qquad (29)$$

where n is zero or an integer (Ishii, 1989). To produce a proper value of N for the positive feedback, the anode voltage V_a and the repeller voltage V_r must be adjusted to satisfy (Ishii, 1989)

$$N = 6.67 \times 10^{-3}(2l + d_g)f\frac{\sqrt{V_a}}{V_a - V_r}, \qquad (30)$$

where l is the distance from the second grid to the repeller, d_g is the gap distance between the two grids, and f is the operating microwave frequency.

Reflex klystrons are durable in harsh environment conditions, such as extreme high or low temperature, in which most solid-state devices are not convenient to operate. Reflex klystrons also have different tuning characteristics than solid-state oscillators. The inherent oscillator noise of reflex klystron oscillators is generally comparable to or smaller than solid-state oscillators. So, as a low-power microwave signal source or a local oscillator application, reflex klystrons are employed to complement solid-state oscillators wherever their superiority in performance is needed. The volt-ampere curves of a reflex klystron are represented by Eq. (24) with $M = 1$ and additional conditions that both the anode voltage and the repeller voltage satisfy conditions expressed by Eqs. (29) and (30). A specific volt-ampere curve of a specific reflex klystron can be obtained from the manufacturer of the tube.

6. MAGNETRONS

6.1 Heating Magnetrons

Magnetrons are high-power microwave oscillators. Heating magnetrons, in particular, are magnetrons made for microwave heating. In domestic, commercial, or industrial microwave heating, the heating is done by essentially unmodulated continuous-wave microwaves. Therefore, most heating magnetrons are designed to operate in continuous-wave operation. Since the application is domestic and commercial, cost effectiveness is one of the most important factors to be considered. Most heating magnetrons are designed with this in mind.

A heating magnetron consists of a heated cylindrical cathode and a resonator-loaded anode coaxially arranged as shown in Fig. 12. A strong magnetic flux is applied axially by either a permanent magnet or an electromagnet, depending on the design. As a result of the radially directed electric force and axially applied magnetic field, electrons emitted from the cathode circulate around the cathode without reaching the anode. The circulating electrons are velocity modulated by the cavity resonators around the anode, which act like buncher cavity resonators in a multicavity klystron. The velocity-modulated electrons form bunches called electron poles. A number of tightly bunched electron poles are formed as the electrons circle around the cathode. These electron poles induce high microwave voltage in the cavities as in the case of the catcher cavity in the klystron. In the case of a magnetron, all cavity resonators act as both bunchers and catchers. So if a pick-up antenna is coupled to one of the cavity resonators, the high-power microwaves induced in the cavity resonator are coupled out of the magnetron for external use.

Magnetron oscillation occurs when the anode voltage V_a and the axial magnetic flux density B_c (in gauss) are related to the anode radius r_a (cm) and the cathode radius r_c by

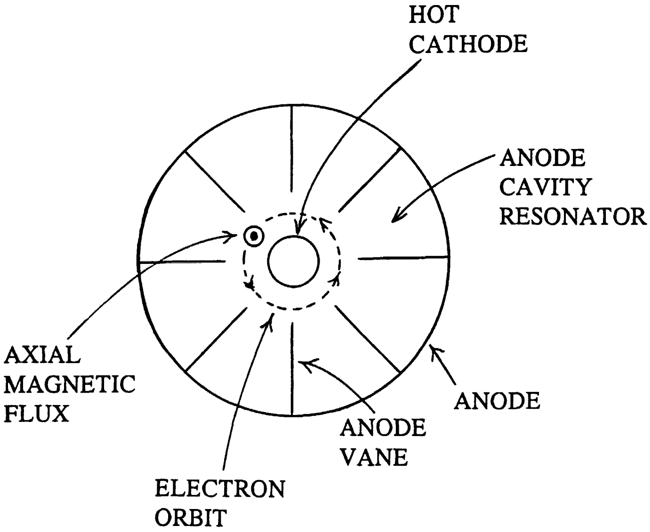

FIG. 12. A schematic cross-sectional view of a heating magnetron.

$$\frac{B_c^2}{V_a} = 45.5\left(\frac{r_a}{r_a^2 - r_c^2}\right)^2. \tag{31}$$

The microwave power generated by the magnetron is (Liao, 1988)

$$P = \tfrac{1}{2}Nv^2\omega C/Q \quad \text{(watt)}, \tag{32}$$

where N is the number of resonator cavities, v is the microwave voltage across the vane tips, ω is the operating angular frequency, C is the capacitance across the vane tips, and Q is the load Q (quality factor) of the resonator. The efficiency of the magnetron is

$$\eta = P/V_aI_a, \tag{33}$$

where V_a is the dc anode voltage and I_a is the dc anode current of the magnetron.

6.2 Radar Magnetrons

Radar magnetrons are designed for pulsed operation. To form an interrogation pulse for a long-range radar, a magnetron must produce a large amount of power for a short time. To allow a sufficient radar echo receiving time, the magnetron must wait for a while for the radar reflections to return before the next transmission of short-pulse modulated microwaves. A radar magnetron is on for a microsecond or so and off for a millisecond or so. The ratio of "on" time and the pulse repetition period is called the duty cycle. When the radar magnetron is on, it emits tens of kilowatts of microwaves, but the averaged power for the entire period is not so great. If the average power over the pulse repetition period T is P and the peak power in the transmission period is P_p for τ seconds, then

$$P = P_p\tau/T. \tag{34}$$

This is the reason that even high–pulse-power rated magnetrons are not excessively bulky. A typical radar magnetron configuration is shown in Fig. 13. Most radar magnetrons employ permanent magnets for magnetic biasing. The principle of radar magnetron oscillation is identical to the principle of operation for a heating magnetron. Most radar magnetrons are designed and operated in the π-mode. This means that phase-angle difference of the microwave voltage from one anode-block pole to the next is π radians. The anode-block pole is the part of the anode block that is protruding toward the cathode. To ensure π-mode operation, the two alternating sets of anode-block poles are strapped together by a pair of conducting rings called the strapping rings. A strapping ring equalizes microwave electric potential.

The performance characteristics of a radar magnetron are represented by a performance chart and a Rieke diagram. The performance chart is a family of volt-ampere

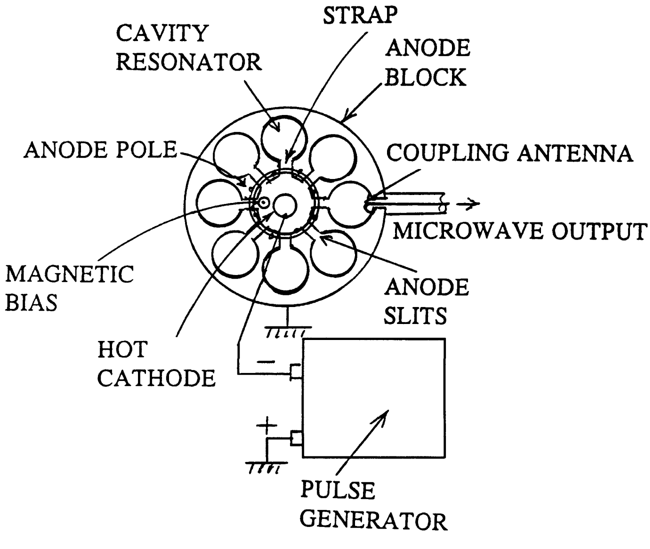

FIG. 13. A schematic cross-sectional view of a radar magnetron.

curves for various levels of output power, magnetic flux density, and efficiency, as illustrated in Fig. 14(a). The chart shows the value of the anode voltage, expected anode current, the required magnetic flux density, the expected microwave output power, and the expected operational efficiency.

A Rieke diagram is a family of plots of constant-frequency curves and constant-output power contours of a microwave oscillator tube on a Smith circuit admittance diagram as shown in Fig. 14(b). The Smith diagram is a diagram of microwave voltage reflection coefficient for various load admittances at various operating frequencies on a complex polar coordinate plot with the origin at the center, unmarked. The diagram was invented by P. H. Smith in 1939 (Smith, 1939, 1944). A point on the Rieke diagram in Fig. 14(b) shows the operating frequency, the output power, corresponding output admittance, voltage reflection coefficient, voltage standing-wave ratio, and the pulling figure of the magnetron (Ishii, 1989). The area marked as sink indicates the value of the load admittance with which the magnetron will not oscillate. The pulling figure is an amount of frequency deviation from the center frequency at the maximum power output where the voltage standing-wave ratio becomes 1.5. It is the indication of frequency stability against the variation of the load admittance. A small pulling figure indicates good frequency stability.

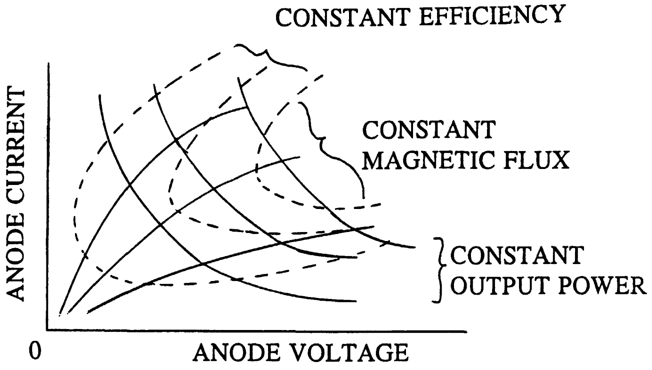

(a)

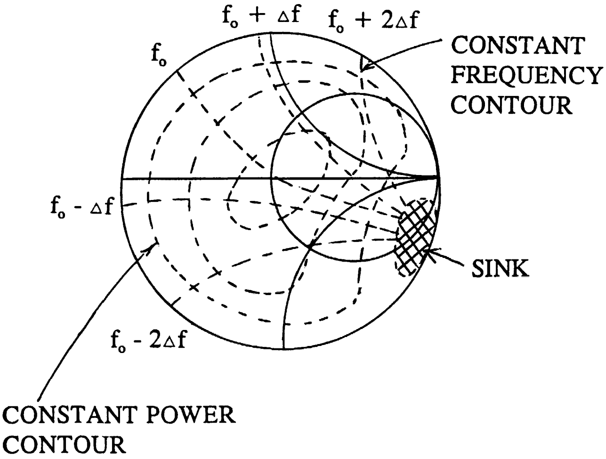

(b)

FIG. 14. (a) Performance chart and (b) Rieke diagram of a radar magnetron.

7. GYROTRONS AND FAST-WAVE TUBES

7.1 Gyrotrons

By properly adjusting the magnitude and direction of a dc magnetic field and a dc electric field, it is possible to impart gyrating motion to thermally emitted electrons. This is accomplished by adjusting both the axial acceleration velocity of electrons and axial and radial magnetic flux density together with the cathode geometry and the structure. Electrons are launched from a side-emitting cathode into a long cylinder-shaped anode that constitutes a cylindrical waveguide. The electron beam is formed by electrons in Larmor orbits. The frequency of electron gyrations in the Larmor orbits is Doppler-shifted relativistic cyclotron frequency (Granatstein, 1987; Coleman, 1982). Electrons in Larmor orbits interact with microwave electric fields. If the motion of electron gyration and microwave fields are synchronized so that the electrons are decelerated and lose kinetic energy, that lost energy is transferred into the interacting microwave electric energy fields, and the microwave fields grow stronger. The microwave electric fields are initially induced by the gyrating motion of the electrons and are subsequently strengthened by positive interaction between the gyrating electrons and the induced microwave field. The speed of electron gyration is fast, and the mass of the electron is weakly relativistic (Granatstein, 1987; Coleman, 1983). Vacuum electron devices based on gyrating electron beams are called gyrotrons. If the gyrating electron beam is launched into only one cylindrical waveguide, it is called the gyromonotron (Coleman, 1982).

Gyrotrons generally generate high frequencies, on the order of 100 GHz, and high power, on the order of 100 kW (Coleman, 1982).

In the gyromonotron, if a small microwave signal is fed at the input opening of the waveguide, the signal is amplified if the voltage and emission current are properly adjusted. Such a gyrotron amplifier is called the gyrotraveling-wave-tube. When the anode waveguide of a gyromonotron is split in the middle, the gyrotron is called the gyroklystron. The first waveguide resonator is called the buncher resonator. In this buncher waveguide resonator, the electron beam interacts with the microwave fields and becomes velocity modulated. Electrons are bunched together when they drift out from the buncher waveguide resonator and enter the second waveguide resonator, which is called the catcher. In the catcher waveguide, the bunched electrons amplify microwave electric fields by transferring their kinetic energy to the microwave electric-field strength due to relativistic variation of the mass of electrons.

For more detailed information, readers are encouraged to refer to Granatstein (1987).

7.2 Novel Microwave Tubes

Gyrotrons are considered to be fast-wave tubes. In gyrotrons, electrons are interacting with microwaves of natural speed in a natural, smooth waveguide. This microwave with natural speed is called the fast wave in this particular case. A number of novel microwave tubes are fast-wave tubes. The fast-wave tubes operate on the principle of interaction between the kinetic energy of bunched electrons and a microwave field traveling at its natural speed on a natural waveguide without any slow-wave structure. The features of the fast-wave tubes are, as seen in gyrotrons, high-frequency capability on the order of 100 GHz and, at the same time, high power capability on the order of a megawatt. Besides gyrotrons, Ubitrons and Peniotrons are also considered to be different versions of fast-wave tubes. The fast-wave microwave vacuum electron devices are also called free-electron masers or free-electron lasers depending on the operating wavelength. In Ubitrons, a spatially periodic transverse motion of electrons is created in an electron beam by periodically arranged dc magnets along its path. The periodicity and the phase of the transverse motion of electrons in the electron beam are adjusted to match with the periodicity of microwaves interacting in a multimode cavity resonator in which the electron beam is launched so that the transverse motion of electrons is decelerated by the microwave fields, thereby the kinetic energy is transferred to the energy of microwave fields. Thus the microwave is amplified (Ishii, 1989; Liao, 1988).

In Peniotrons a single helical electron

beam is launched along the center axis of a double-ridge dominant-mode rectangular waveguide. The transverse and tangential component of electron motion in the helical electron beam interacts with the high transverse microwave electric field created by the waveguide ridges. The diameter of the helix formed by the electron beam is adjusted to be equal to the center-to-center distance of the double ridges of the waveguide. The periods and phase of the helical motion of the electrons are synchronized and phased to the interacting microwave fields so that the transverse and tangential component of the electron motion is decelerated by the microwave electric fields. Thus the kinetic energy of the electron motion is transferred to the interacting microwave electric field, and the microwave is amplified (Ishii, 1989; Liao, 1988).

8. TRAVELING-WAVE TUBES (TWTs)

8.1 O-Type TWT

Traveling-wave tubes are vacuum tubes in which the speed of microwave propagation is nearly equal to the speed of electrons in an electron beam. By positive interaction of microwave electric fields and the electron beam, microwaves in TWTs are amplified. In an O-type TWT, an electron beam is focused by a longitudinal magnetic field that is parallel to the direction of electron beam as shown in Fig. 15. The trajectory of an individual electron is a helix.

The electron beam is accelerated by an anode that consists of a long helical conducting transmission line. The microwaves to be amplified are coupled to one end of this helical transmission line. The axial speed of microwave transmission is slowed down to

$$v = c \sin\psi, \tag{35}$$

where ψ is the pitch angle of the helix and c is the tangential speed of the microwaves propagating along the conductor of the helix, which is assumed to be equal to the speed of light in vacuum, 300 Mm/s. In a traveling-wave tube, the acceleration voltage V_a and the helix pitch angle are adjusted so that the electron speed

$$u \approx v. \tag{36}$$

Under this condition the electron beam interacts with the input microwaves. The electron beam is velocity modulated near the input. The acceleration and the deceleration relationships last a long period of time because $u \approx v$. The velocity-modulated electrons bunch as they travel inside the helical transmission line. As they approach the end of the helical line, the tightly bunched electrons lose a part of their kinetic energy through deceleration by microwave electric field, and the lost kinetic energy of the electron beam is transferred to the microwave energy. Thus, the microwaves are amplified. After the energy transfer and induction of the output signal, the electron beam is ter-

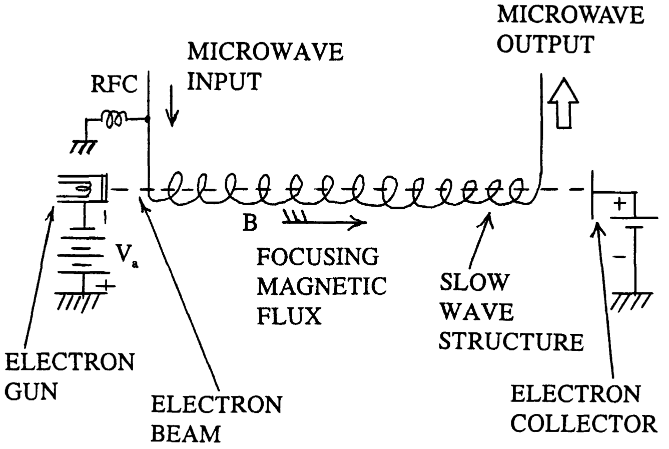

FIG. 15. A schematic diagram of an O-type TWT.

minated by the electron collector as shown in Fig. 15.

The output microwave voltage of a traveling-wave tube is

$$v_0 = v_i \exp\left(\frac{\sqrt{3}}{2} C \frac{\omega l}{u}\right), \tag{37}$$

where l is the length of the interaction region, ω is the operating angular frequency, u is the speed of electrons, v_i is the microwave input voltage, and C is the gain parameter given by

$$C = (Z_0 I_a / 4V_a)^{1/3}, \tag{38}$$

where Z_0 is the characteristic impedance, I_a is the beam current, and V_a is the anode voltage of the TWT.

The voltage gain of the TWT is

$$G \equiv \frac{v_0}{v_i} = \exp\left(\frac{\sqrt{3}C\omega l}{2u}\right) \tag{39}$$

or

$$G[\text{Neper}] = \sqrt{3}C\omega l/2u. \tag{40}$$

Because of the condition that $u \approx v$, the frequency bandwidth of the TWT is generally wide. It is not unusual for a TWT to have a bandwidth more than an octave wide.

8.2 M-Type TWT

The principle of microwave–electron-beam interaction and microwave amplification in an M-type TWT is similar to the principle in the O-type TWT. The difference is in the way that the electron beam is formed and focused. An example of a schematic diagram of an M-type TWT is shown in Fig. 16. Electrons are emitted from a side-mounted electron gun. The electron beam is focused by the vertical dc electric fields between the anode and the sole and the horizontal dc magnetic field, B, which is applied perpendicular to the plane of the paper. Under this condition, the individual electron orbit is a cycloid. As a whole, the electron beam is shaped like a ribbon.

The electron beam then undergoes velocity modulation, bunching, and electrodynamic induction, producing amplified microwaves at the output of the slow-wave transmission structure as in the case of the O-type TWT.

9. DISPLAY TUBES

9.1 Cathode-Ray Oscilloscopes

Electrical phenomena are often graphically displayed by means of a cathode-ray oscilloscope. A schematic diagram of a generic cathode-ray oscilloscope is shown in Fig. 17. An electron beam is launched into a vacuum into the space between two pairs of parallel-plate electrodes. The plate electrodes are for

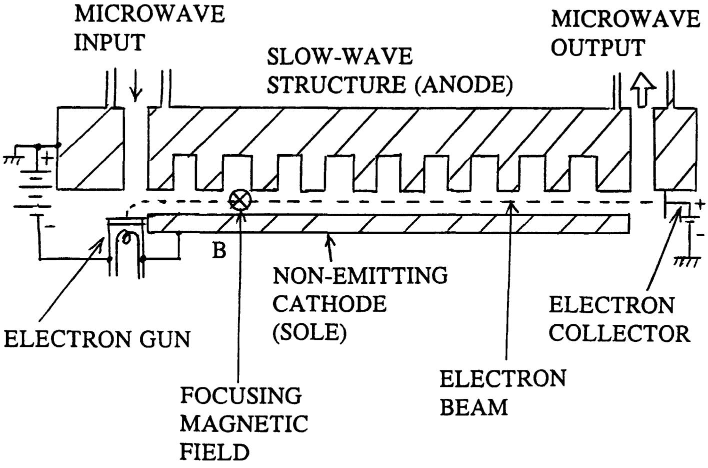

FIG. 16. A schematic diagram of an M-type TWT.

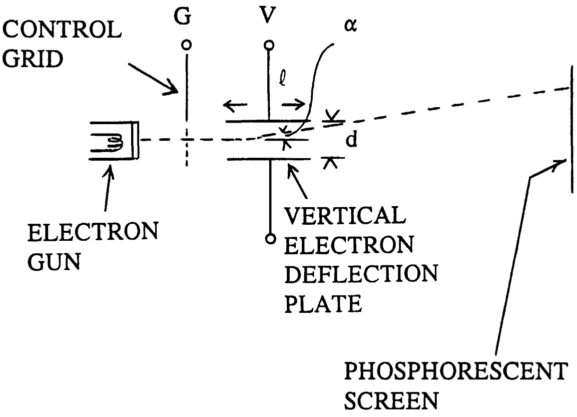
CONTROL GRID
G
V
α
l
d
ELECTRON GUN
VERTICAL ELECTRON DEFLECTION PLATE
PHOSPHORESCENT SCREEN

FIG. 17. A schematic diagram of a cathode-ray oscilloscope.

electron-beam deflection. Therefore, these plates are called the electron deflection terminals. If a voltage is applied across the pair of horizontally mounted deflection plates, the electron beam is vertically deflected by the Coulomb force. If a voltage is applied across the pair of vertically mounted deflection plates, the electron beam is horizontally deflected. Where the electron beam hits the phosphorescent screen, it produces visible light. Thus the landing spot of the electron beam on the screen is visible. The landing spot is controlled by the voltages applied across the deflecting plates. Therefore, electrical quantities applied across the deflecting plates are visible on the phosphorescent screen. Normally, a sawtooth wave is applied to the horizontal deflecting plates to produce a time axis.

The deflection angle α of the electron beam is given by (Martin, 1993)

$$\alpha = \tan^{-1}\left(\frac{1}{2}\frac{V_D}{V_0}\frac{l}{d}\right), \tag{41}$$

where V_D is the deflection voltage, V_0 is the anode voltage or the acceleration voltage of the electron beam, l is the length of the deflection plate, and d is the plate separation.

The control grid in Fig. 18 is negatively biased. The amount of negative voltage controls the brightness of the spot on the screen.

9.2 Monitor Tubes

In many video monitor tubes for computers and television picture display, the electron deflection is done magnetically by two

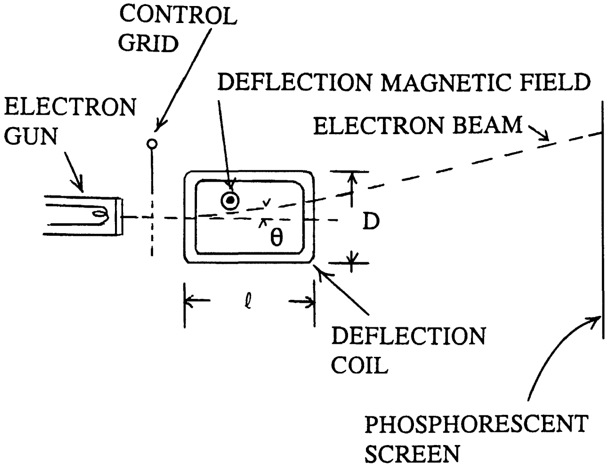
CONTROL GRID
DEFLECTION MAGNETIC FIELD
ELECTRON GUN
ELECTRON BEAM
v
θ
D
l
DEFLECTION COIL
PHOSPHORESCENT SCREEN

FIG. 18. A schematic diagram of a video monitor tube.

pairs of magnetic coils: one pair for the horizontal deflection and the other pair for the vertical deflection as shown in Fig. 18. These deflection coils control the landing location of the electron beam on the screen, and the video signals applied on the control grid controls the brightness of the spot on the screen. Usually the deflection coils are installed outside of the vacuum envelope.

The deflection angle of the beam, θ, is given by (Martin, 1993)

$$\theta = \sin^{-1}\frac{Nil}{2.68D\sqrt{V_0}}, \qquad (42)$$

where N is the number of turns in the deflection coil, l is the length of the magnetic interaction region, D is the height of the interaction region, V_0 is the anode acceleration voltage of the beam, and i is the coil excitation current.

For a color monitor, usually a multicolor screen is used. Electrons are directed to land at a specific location on the screen to produce the desired color by controlling the deflection coil current and the control grid voltage.

10. IMAGING TUBES

10.1 Vidicon Tubes

Vidicon tubes are vacuum tubes that convert optical images on the front face of the tube into electrical video signals. The electron beam is focused and scanned by the use of the longitudinal focusing magnetic coils and deflecting magnetic coils as in the case of video monitor tubes. However, the target that the electron beam strikes is not a phosphorescent screen, but a rather complicated optical image–electrical video conversion screen.

A schematic diagram of the video conversion screen of a vidicon tube is shown in Fig. 19. Using optics, an image is formed on the target plate. In front of it, there is a transparent semiconductor film a few micrometers thick, called the signal plate, which is electrically connected to the metal ring encircling the front of the tube. The signal plate is positively biased with respect to the electron gun that emits the electron beam. The target plate is coated with a thin film of a photoconductive coating. This photoconductive coating is almost an insulator in the dark. Therefore, in the dark the signal plate and scanning electron beam are virtually insulated from each other and there is very little conduction from the dc supply to the electron beam. When illuminated, the photoconductive coating becomes a conductor, connecting the signal plate and electron beam so that electric current flows from the dc supply to the electron beam. The amount of the current is proportional to the intensity of the light.

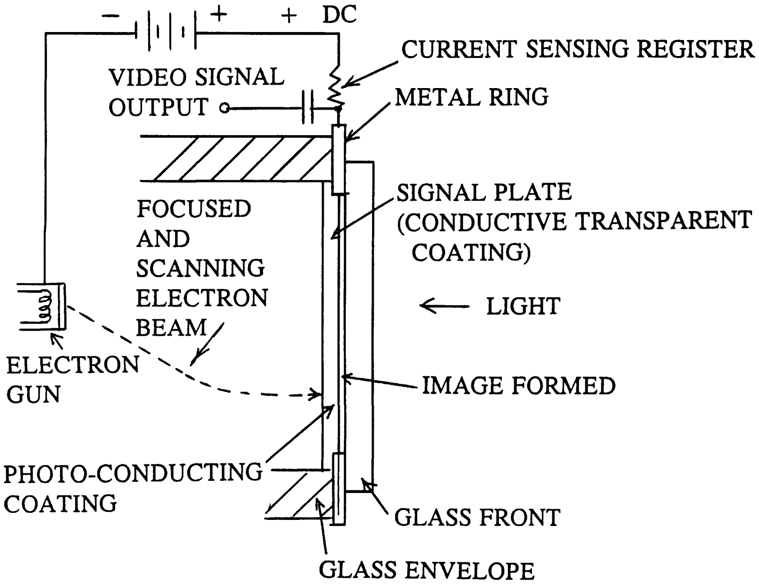

FIG. 19. A schematic diagram of the imaging area of a vidicon tube.

As the electron beam scans across the bright and dark spots of the image on the target plate, a video current proportional to incident light intensity at the beam landing spot flows from the dc supply through the current-sensing resistor, signal plate, and electron beam into the electron gun. Thus, a video voltage signal is developed across the current-sensing resistor shown in Fig. 19. From there the video voltage signal is capacitively coupled out and processed for remote television imaging systems.

A family of imaging tubes such as the infrared vidicon, silicon-target vidicon, plumbicon, and orthicon operate using basic principles similar to those of the vidicon tube (Liao, 1988).

10.2 Thermicon Tubes

The thermicon tube is a vacuum tube that converts infrared images into electrical video signals by temperature-dependent photoelectron emission. This is useful for seeing through darkness and fog. An essential part of the thermicon is schematically shown in Fig. 20. There is a plate that is electrically conductive, as well as thermally and optically transparent. One side of the plate is coated with a thin layer of material that absorbs infrared light but is transparent to visible light. An infrared image is formed by optics on this infrared absorbing layer. The infrared image heats up the plate according to its light and dark spots, creating a specific heat distribution pattern on the plate depending on the infrared image. The heat distribution is transmitted to the other side of the plate where photoelectron emitting layer is formed by coating with thermally sensitive photoemitting material. The photoemission efficiency depends on the temperature of the material.

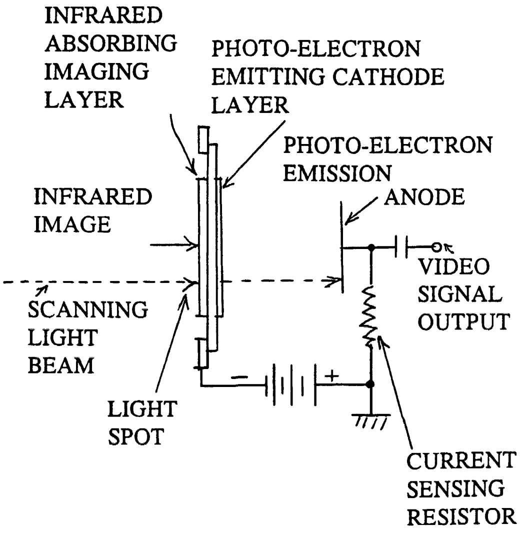

FIG. 20. A schematic diagram of the essential part of a thermicon.

If a bright infrared spot on the image plate is illuminated by a scanning spot of visible light, photoelectrons will be emitted. More infrared illumination will cause more photoelectron emission when spotlighted. The emitted electrons are collected by the anode, and the anode current is sensed by the current-sensing resistor. Across the current-sensing resistor, the video signal voltage develops as the light spot is scanned on the infrared image plate. This infrared image plate is called the retina of the thermicon (Liao, 1988).

11. SCIENTIFIC VACUUM ELECTRONICS

11.1 Electron Microscopes

Details of electron microscopes are covered in ELECTRON MICROSCOPY. Here a summary of the basic principles of electron microscopy is presented.

The main objective of electron microscopy is to obtain higher resolution and higher magnification than optical microscopy. The higher resolution is obtainable if the operating wavelength is shorter. The wavelength of visible light ranges from 400 to 700 nm. The wavelength of waves associated with the electron beam is (Harman, 1953)

$$\lambda = 1.23 \times 10^{-9}/\sqrt{V}, \qquad (43)$$

where V is the acceleration voltage of the electron. If V is 1000 V, the wavelength is only 0.04 nm. In practice, due to limitations in electron-focusing systems, the resolution obtainable by electron microscopy is on the order of 1 nm (Harman, 1953). In visible-light microscopy, the resolution is on the order of a micron.

It should be noted that the waves associated with an electron beam are electron waves, not electromagnetic waves like visible light (Harman, 1953). The wavelength of an electron wave associated with an electron beam is

$$\lambda = h/mv, \tag{44}$$

where h is Planck's constant, 6.62×10^{-34} J-s, m is the electron mass, and v is the electron velocity (Sproull, 1958);

$$mv = m_0 v/\sqrt{1 - v^2/c^2}, \tag{45}$$

where m_0 is the rest mass of an electron and c is the speed of light in a vacuum; mv is the momentum of an electron in the beam.

The moving electrons can be focused by the Lorentz magnetic force, which is

$$\mathbf{F} = e\mathbf{B} \times \mathbf{v} \quad \text{(Newton)}, \tag{46}$$

where $\mathbf{B}$ is the magnetic flux density vector, $\mathbf{v}$ is the velocity vector of the moving electrons, and e is the electric charge of an electron.

Therefore, if a magnetic focusing system is arranged as shown in Fig. 21, it is possible to form a microscope using electron waves. The electron gun emits an electron beam in a vacuum environment. The electron beam is then focused to form a paraxial beam by the longitudinally parallel magnetic flux created by the focusing magnet. The paraxial electron beam hits the specimen and, according to the nature of the specimen, experiences absorption and scattering of the electron waves. Therefore, the electron-wave beam transmitted through the specimen is influenced by the image of the specimen. The electron-wave beam transmitted through the specimen is focused again and diverged by properly shaped magnetic flux of the first electron-beam focusing and diverging magnet. This magnet is called the magnetic objective lens. Thus the first enlarged image is formed. The middle part of the first image is further focused and diverged by the second beam focusing and diverging magnet, which is called the magnetic projector lens. The electron-wave beam through this last magnet is projected onto an imaging screen that is either a photographic plate or a fluorescent screen. Thus a great magnification and resolution are obtainable using an electron microscope.

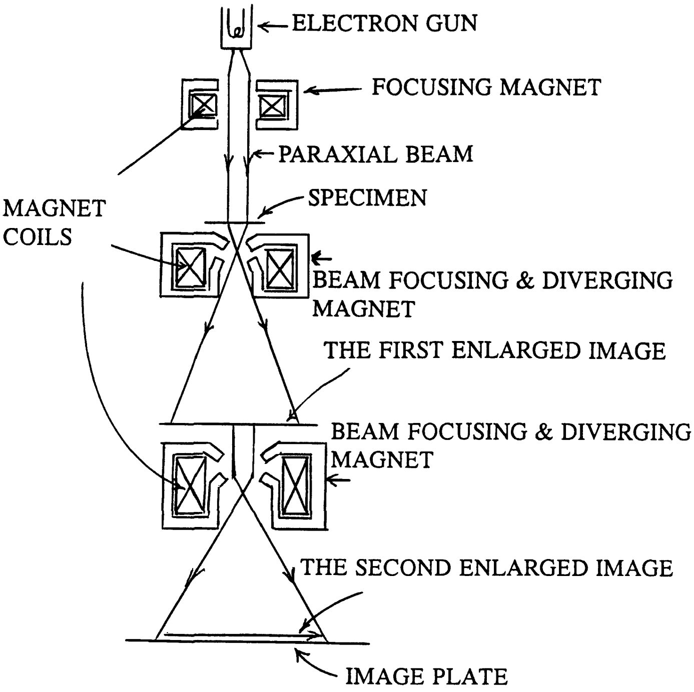

FIG. 21. A functional diagram of an electron microscope.

11.2 Particle Accelerators

Particle accelerators are used for the study of elementary particles and structure of atoms and molecules. There are a number of different types of particle accelerators including cyclotrons, synchrotrons, and linear accelerators. Details of these devices are given in CYCLOTRONS; SYNCHOTRONS; and ACCELERATORS, LINEAR. It should be noted here that particle accelerators are applications of vacuum electronics.

GLOSSARY

Convection Current: An electric current due to the motion of electrically charged particles.

Electrodynamic Induction: Induction of a time-varying electric current in a circuit by the motion of electrically charged particles in a space that is coupled to the circuit.

Electron Bunching: The formation of localized groups of electrons.

Electron Gun: A vacuum electron device consisting of a heated cathode, accelerating anodes, and focusing electrodes used as an electron-beam source.

Electron Multiplier: A thermionic device to multiply the number of electrons by the use of cascading electron multiplication by collision with secondary-electron producing electrodes and production and acceleration of secondary electrons.

Electron Transit Cycles: Number of microwave cycles elapsed while an electron is traveling a given distance.

Grid Cutoff Voltage: The value of negative grid voltage to make zero anode current in a vacuum tube.

Kinetic Potential: Electric potential generated by the motion of charged particles.

Perveance: Ratio of the anode current to the $\frac{3}{2}$ power of the anode or equivalent anode voltage of a vacuum tube.

Transconductance: The ratio of change of the anode current to the change of control grid voltage in a vacuum tube.

Works Cited

Brodie, Ivor (1989), "Physical Condition in Vacuum Microelectronics Devices," *IEEE Trans. Electron Dev.* **36,** 2641–2644.

Coleman, James T. (1982), *Microwave Devices,* Reston, VA: Reston Publishing Company.

Granatstein, Victor L. (1987), *High Power Microwave Sources,* Dedham, MA: Artech House.

Harman, Willis W. (1953), *Fundamentals of Electronic Motion,* New York: McGraw-Hill.

IEEE (1972), *IEEE Standard Dictionary of Electrical and Electronics Terms,* New York: Wiley.

Ishii, T. Koryu (1989), *Microwave Engineering,* San Diego, CA: Harcourt Brace Jovanovich.

Ishii, T. Koryu (1990), *Practical Microwave Electron Devices,* San Diego, CA: Academic Press.

Liao, Samuel Y. (1988), *Microwave Electron-Tube Devices,* Englewood Cliffs, NJ: Prentice-Hall.

Martin, Andre (1993), in: Richard C. Dorf (Ed.), *Electrical Engineering Handbook,* Boca Raton, FL: CRC Press.

Nakamura, Shintaro (1969), *Kaisetsu Denshi Soshi Kogaku,* Tokyo: Seibundo.

Reich, Herbert J., Ordung, Philip F., Krauss, Herbert L., Skalnik, John G. (1953), *Microwave Theory and Techniques,* Princeton, NJ: D. Van Nostrand.

Smith, P. H. (1939), "Transmission Line Calculator," *Electronics* **12,** 29–31.

Smith, P. H. (1944), "Transmission Line Calculator," *Electronics* **17,** 130–133, 318–325.

Sproull, Robert L. (1958), *Modern Physics,* New York: Wiley.

Further Reading

Brodie, I. (1988), "Very High Speed Integrated Microelectronic Tubes," U.S. Patent 4721885.

Brodie, I., Murray, J. J. (1982), *The Physics of Microfabrication,* New York: Plenum.

Dohler, G., Gallapher, Scafuri, F., Moats, R. (1978), "The Peniotron: A Fast Wave Device for Efficient High Power Millimeter Wave Generation," in: *IEEE Electron Devices Meeting: Digest,* New York: IEEE Electron Devices Society.

Dombrowski, G. E. (1988), "Numerical Solution of Magnetron Oscillator and Crossed-Field Amplifiers," *IEEE Trans. Electron Dev.* **ED-35,** 2060–2067.

IEEE (1996), *IEDM Technical Digest,* San Francisco, CA: IEEE Electron Devices Society.

Gilmour, A. S. (1986), *Microwave Tubes,* Dedham, MA: Artech House.

Phillips, R. M. (1960), "The Ubitron, a High-Power Traveling Wave Tube Based on a Periodic Beam Interaction in Unloaded Waveguide," *IRE Trans. Electron Dev.* **ED-7,** 231–241.

Pierce, J. R. (1950), *Traveling Wave Tubes,* Princeton, NJ: Van Nostrand.

Van DeRoer, Theo G. (1994), *Microwave Electronic Devices* London: Chapman and Hall.

VACUUM TECHNIQUES AND INSTRUMENTATION

JOHN F. O'HANLON, *Department of Electrical and Computer Engineering, University of Arizona, Tucson, Arizona, U.S.A.*

Introduction 105
1. **Pumping Techniques** 105
1.1 Displacement Pumps 105
1.1.1 Rotary Oil-Sealed 106
1.1.2 Lobe 106
1.1.3 Turbomolecular 107
1.1.4 Dry 107
1.1.5 Diffusion 108
1.1.6 Pump Fluids 109
1.2 Capture Pumps 109
1.2.1 Cryogenic and Sorption 109
1.2.2 Getter and Ion 110
2. **Gas Sources** 110
2.1 Volume 110
2.2 Surface 111
3. **Gas-Flow Fundamentals** 111
3.1 Molecular 111
3.2 Viscous Laminar 112
3.3 Choke 112
4. **System Performance** 112
4.1 Backstreaming 112
4.2 Outgassing and Permeation 113
4.3 Dynamic Effects 113
5. **Vacuum Instrumentation** 114
5.1 Total-Pressure Gauges 114
5.1.1 Thermocouple and Pirani Gauges 114
5.1.2 Capacitance Manometer 115
5.1.3 Ionization Gauge 115
5.2 Residual-Gas Analyzers 115
5.2.1 Quadrupole 116
5.2.2 Magnetic Sector 116
5.3 Thermal Mass-Flow Meter 117
5.4 Leak Detection 117
5.5 Calibration Metrology 117
6. **Safety** 118
Glossary 118
Works Cited 118
Further Reading 119

INTRODUCTION

Producing a low-pressure working environment requires understanding of pumping techniques, materials of construction, and sources of gas and the manner in which they flow, as well as practical issues such as chamber cleaning and leak detecting. This article briefly reviews these issues.

1. PUMPING TECHNIQUES

Removal of gas from an enclosed volume is accomplished by either displacement or capture pumping. Displacement pumps are compressors, and capture pumps are condensers or adsorbers. Displacement pumps transport gas to their exits, whereas capture pumps store gas.

1.1 Displacement Pumps

Displacement pumps admit and compress gas and exhaust it to another pump or to the atmosphere. The parameters that describe the performance of a displacement pump are pumping speed and compression ratio, and both are functions of inlet pressure. The compression ratio is the ratio of pump outlet pressure to pump inlet pressure. The (volumetric) speed is the ease with which gas can flow across a surface—the volumetric gas flow divided by the pressure at the plane of the measurement. The compression ratio of some pumps is so high that one rarely quotes numerical values. Oil-sealed vane, piston, lobe, turbomolecular, and diffusion pumps are displacement pumps used for research, development, and manufacturing applications. Dry pumps operate without using oil as a sealing medium.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

1.1.1 Rotary Oil-Sealed Rotary vane and piston pumps are oil-sealed mechanical pumps. The vane pump has a cylindrical rotor with, usually, two vanes located within, but offset from the axis of, a hollow stator cylinder. The vacuum seal is made between the vane tips and stator walls, and between the rotor cylinder and stator. See Fig. 1. Gas is admitted into the region between the rotor and stator, compressed, and forced past the exhaust valve to the atmosphere. A single-stage pump can exhaust to pressures of order 1 Pa, whereas a double-stage pump can exhaust to pressures of 0.01 Pa. In practical applications one rarely uses either pump below 10 Pa, or backstreaming to the chamber will result. Backstreaming is the flow of oil or other undesirable vapor in the direction counter to the desired flow.

Piston pumps are rugged, but their asymmetric design generates vibration, which can be controlled with a balancing design. The principle, sketched in Fig. 2, involves rotation of a cylindrical piston and hinge bar such that gas is allowed to enter the cavity between the cylinder and stator where it is compressed and ejected. The piston is mechanically constrained from touching the stator. The sealing surfaces are located at the hinge bar and the point of closest approach of the piston and stator wall. Two complete revolutions of the piston are required for each pumping cycle. They have the advantage of simplicity—no vanes to stick—and they must be physically mounted to the floor.

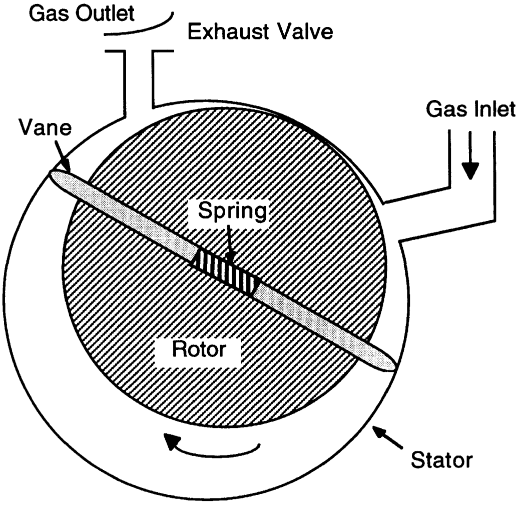

FIG. 1. Rotary oil-sealed mechanical pump; the axis of rotation of the cylindrical rotor is offset from the axis of rotation of the cylindrical stator housing.

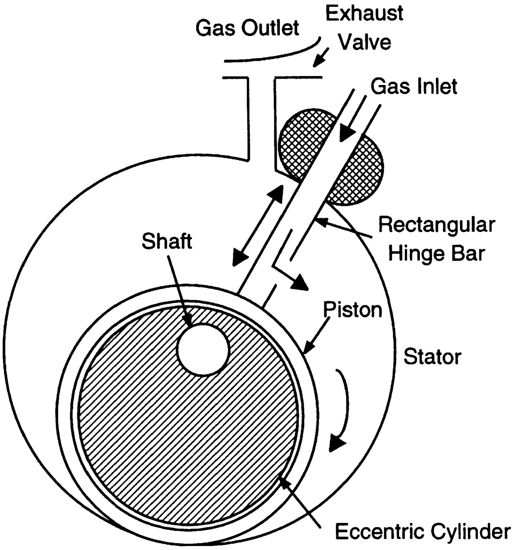

FIG. 2. Rotary piston pump; the axis of rotation of the shaft is concentric with the stator housing. The hinge bar is a rectangular cylinder that slides within the stator; it forms the path by which gas enters the pump. The hinge bar turns or hinges about the shaft such that a small gap remains between the rotor and stator.

1.1.2 Lobe Lobe pumps or blowers can move large volumes of gas efficiently. The two "figure 8" lobes, illustrated in Fig. 3, rotate in opposition and never touch each other or the stator housing. No fluid is used to seal the 0.9-mm gap between the two surfaces. Useful compression is obtained by spinning the rotors at speeds of order 3 500 rpm. The compression ratio is a maximum at about 100 Pa. For lobe blowers designed to work in vacuum, the work of compression will generate enough heat above 1 000 Pa to cause rotor expansion and pump damage, if the staging ratio is >1. The staging ratio is the ratio of lobe blower displacement to mechanical pump displacement. Thus, power is not applied to the shaft until this pressure is reached. Lobe blowers must exhaust into a mechanical pump. A typical staging ratio is 5:1 or 10:1. Lobe blower–mechanical pump sets are useful for achieving low pressures

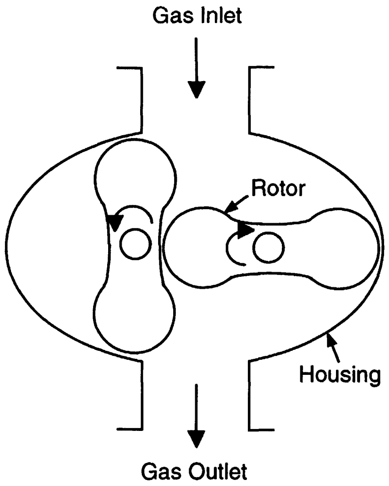

FIG. 3. Lobe pump, or Roots blower; the lobes rotate in opposite directions. They do not touch each other or the wall.

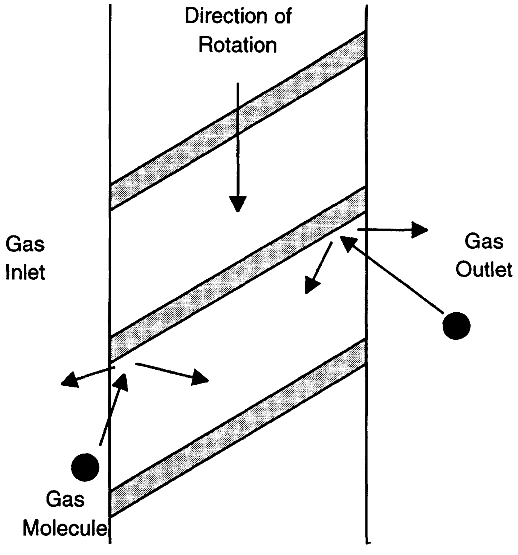

FIG. 4. End view of a moving blade in a turbomolecular pump illustrating how the moving blade preferentially adds momentum directed to the right. Molecules entering the pump are preferentially transmitted, whereas those molecules that have left the pump are prevented from backstreaming.

and high speeds. For a lower practical pressure limit of a mechanical pump—say, 10 Pa—to reduce oil backstreaming from the mechanical pump, the minimum inlet pressure of the lobe blower will be about 2 to 5 Pa. Speeds above 200 to 300 m^3/h are obtainable with either pump but are more economically obtained with a series combination of lobe and mechanical pump than with a large mechanical pump.

1.1.3 Turbomolecular Turbopumps use metal blades as "molecular bats." The blades compress gas by transferring momentum from moving blades to gas molecules, as sketched in Fig. 4. The blade tip velocity is comparable to the thermal velocity of a nitrogen molecule. Heavy gases are compressed well, but the maximum compression ratio, with no flow, for hydrogen is only 1 500 to 15 000 in modern pumps with about 20 blade pairs. See Kruger and Shapiro (1961). If the exhaust pressure of a turbopump is too high, say 20 Pa, the blade speed will decrease on account of viscous drag. In a similar manner, increasing the inlet pressure above 0.4 Pa will reduce the blade speed. Turbopumps are used where oil contamination must be minimized or where high gas flows must be pumped without interruption. They are manufactured with coated aluminum blades for pumping corrosive gases and with magnetically levitated blades for dry pumping applications. They can reach pressures of 10^{-8} Pa. When power is removed from a turbo pump, it must be properly vented to prevent vapors from the foreline side of the pump from reaching the inlet. This is best accomplished by closing the valve between the chamber and the turbo, followed by venting with dry nitrogen through the turbo pump inlet, after the blades have slowed to approximately 50% of their natural rotational velocity (Nesseldreher, 1976).

1.1.4 Dry In addition to the lobe pump, several other vacuum pumps have been designed that do not use organic fluids as a sealing medium. Carbon-vane, PTFE-sealed piston, claw, and scroll pumps have been used in a variety of applications. PTFE-sealed pumps and carbon-vane pumps are capable of roughing from atmosphere, but they generate contamination and do not operate well when pumping on the dust-laden gas streams characteristic of certain industrial processes. Wide-range turbopumps incorporating a molecular drag stage after a turbo stage can exhaust at 0.2 to 0.3 Pa into a diaphragm pump and provide "oil-free"

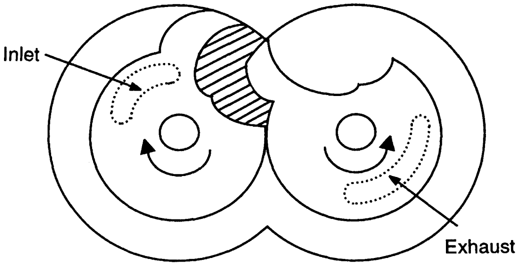

FIG. 5. Lobe pump using claw-shaped rotor; gas enters in the left-hand entrance arc and exits through the right-hand arc.

pumping. The claw, scroll, and lobe pumps are three pumps used where oil-free pumping of particle-laden gases is required. The claw mechanism, shown in Fig. 5, is a compressor, which produces a high compression ratio at pressures higher than in a lobe blower. One commercial design utilizes a lobe (inlet) blower in series with several claw pumps to produce a high gas flow, a high compression ratio, and a low ultimate pressure. The scroll pump consists of fixed and rotating scrolls that mesh to form crescent-shaped volumes. Gas is transported from the entrance slot to the exit slot, where it is discharged to atmosphere. Scroll pumps have usable speeds in the pressure range 1 000 Pa to atmosphere. Scroll pumps have only recently been considered for commercial use.

1.1.5 Diffusion Like the turbo pump, the diffusion pump is a momentum-transfer pump in which large organic molecules are the molecular "bats." See Fig. 6. Pumps are constructed with three or four jet stages in series. The fluid is heated in a boiler, and the vapors, traveling at supersonic speeds, exit the jets and travel to the walls where they are cooled and condense. The critical inlet pressure of a diffusion pump depends on its diameter and is typically 0.1 to 0.01 Pa for pumps smaller than 300 mm in diameter. Above this pressure the upper jet will collapse. The critical forepressure is of order 40–50 Pa, or about half the vapor pressure of the fluid at its operating temperature. If the exhaust is raised above this pressure, the back pressure will prevent the vapor jets

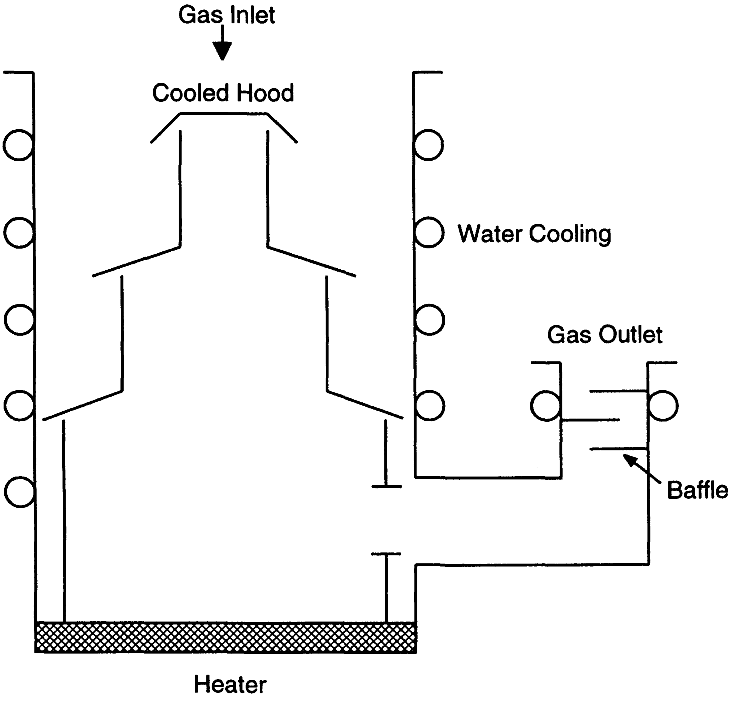

FIG. 6. Oil diffusion pump; oil vapors streaming from the jet assemblies travel to the wall where they condense and return to the boiler.

from reaching the walls and cause a cloud of pump fluid vapor to form at the top of the pump, where it can be transported to the chamber. Thus, the diffusion pump must be exhausted into a mechanical pump. The ultimate pressure of an oil diffusion pump will be determined by the fluid type, chamber materials, and cleaning techniques. Pressures as low as 10^{-10} Pa have been obtained with a liquid-nitrogen–cooled trap.

1.1.6 Pump Fluids A mechanical pump fluid must seal, lubricate, and cool. Turbomolecular and lobe-blower fluids cool and lubricate bearings, whereas diffusion-pump fluids must boil with low heat, be stable at high temperatures, and have extremely low vapor pressures. Petroleum (mineral) oils find application in all pumps. The best mineral oils are those containing mainly paraffin with lesser fractions of naphthalene and ring compounds. Fluids of different viscosity are blended to produce vane, piston, and lobe-blower fluids. Turbomolecular and diffusion-pump oils are distilled once to remove low–molecular-weight fractions, which would increase the fluid's vapor pressure. Pentaphenyl silicone and polyphenyl ether fluids are high-quality diffusion-pump fluids, with extremely low vapor pressures.

The fluids described above are not suitable for use in applications involving corrosive or oxidizing gases. For those applications a perfluoropoly ether (PFPE) fluid is preferred. PFPE fluids are rather expensive, but reasonably inert to a large range of chemicals. They are bases in the Lewis sense, and so the corresponding acids, such as boron trichloride (used for etching aluminum films) or aluminum trichloride (an etch by-product), will cause slow decomposition at the temperatures found in mechanical or turbo pumps. One would not pump on these gases with a PFPE-filled diffusion pump, as the high temperature would cause decomposition in minutes. Older fluids, e.g., phosphate esters, present disposal problems and are no longer used; they have been supplanted by PFPE. Reclamation has been reviewed by Whitman (1987).

1.2 Capture Pumps

Capture pumps condense, physisorb, and chemisorb gases and vapors, which are stored on surfaces or diffuse into deposited films. Their pumping speed is not constant with time but depends on the species being pumped, as well as the species and quantity of gas previously pumped. Capture pumps saturate periodically and require regeneration. Regeneration provides a fresh surface, either by desorbing the entrapped gas or by covering it with a newly deposited film. Cryogenic, sorption, titanium sublimation, nonevaporable getter, and ion pumps are important capture pumps.

1.2.1 Cryogenic and Sorption. The cryogenic pump captures gas molecules by condensation and sorption. A cryopump consists of a source of low temperatures and an array of surfaces on which molecules can become entrapped. The simplest source is a liquid cryogen such as nitrogen or helium, and indeed the former has long been used as a cold trap or Meissner trap to enhance the pumping of water vapor. Commercial cryopumps achieve low temperatures by means of a two-stage helium gas refrigerator (Gifford, 1966). The pumping arrays are connected to the two stages, which cool the arrays to ranges of 15 to 25 K and 70 to 100 K, respectively. The warm stage pumps water and some CO_2 by condensation, just as does a liquid-nitrogen trap. The cold stage has two surfaces: an outer (condensation) surface that pumps all other gases except H_2, He, and Ne, and an inner charcoal-coated (sorption) surface that pumps these three gases.

As the charcoal surface becomes saturated with sorbate, it ceases to pump. The cryocondensing surfaces will continue to pump. However, their "ice" surface temperatures increase, and ultimately the "ice" on the cold stage can make a thermal short to the warm stage, or the water ice on the warm stage can make a thermal short to the pump housing. Cryopumps are regenerated by admitting warm nitrogen gas directly onto the cold stage while the released gases are purged. Regeneration is needed rather infrequently on high-vacuum systems but frequently on sputter deposition systems, or other systems that pump large gas loads. The cryopump is widely used because it is organic free and not easily damaged if the system is accidentally mistreated.

Sorption pumps are containers filled with a zeolite, which is cooled until it equilibrates with its surrounding liquid-nitrogen bath. Zeolites possess an extremely large interior surface onto which gases can sorb to a depth of one monolayer. A quantity of zeolite the size of a softball contains one acre of surface one monolayer deep. Sorption pumps do not pump any noble gases. Noble gases can be partially pumped by using a series of pump modules. Each module can be opened, then closed quickly, while the gas remains in the viscous flow regime. During that brief time, all gases enter in viscous flow, without mass discrimination. A large fraction of the nonadsorbed gases are thus isolated from the system. A series of three pump modules can pump below the natural atmospheric partial pressure of neon. Sorption pumps are used as roughing pumps on systems in which the high vacuum is produced with titanium sublimation pumps (TSPs) and ion pumps to produce an all oil-free pumped environment.

1.2.2 Getter and Ion Getters are used alone, as in the TSP and nonevaporable getter (NEG) pump, or form part of the mechanism of the ion pump. The TSP is a getter pump that sorbs gases in freshly deposited titanium films. Water dissociates into oxygen and hydrogen. Oxygen and nitrogen are chemisorbed. Hydrogen diffuses inward to form a hydride, whereas methane is only weakly bonded—its adsorption is extremely small. Inert gases are not pumped. Periodically, fresh titanium films must be redeposited. This deposition frequency may be only once or twice daily at 10^{-10} Pa; however, at 10^{-3} Pa the filament must continually sublime. These pumps find application in research environments where, in combination with ion pumps, they provide a low-cost, oil-free method for achieving ultrahigh vacuum.

The NEG pumps find application in maintaining pressure in small isolated systems, as well as in distributed pumping for multikilometer-long accelerator systems. The NEG pumps are fabricated from Zr-Al or Zr-V-Fe alloys. Activation requires heating, which allows the oxide formed on the surface to diffuse inward, leaving a fresh surface to pump reactive gases. All gases except inert and hydrogen are pumped as compounds. Hydrogen is pumped as a hydride and can be expelled by heating, if regeneration is necessary. The NEG pumps begin to operate at 1 Pa. They have also found application as ultrapure gas purifiers.

The basic ion-pump cell consists of a cylindrical anode centered between two cathode disks. A dc voltage of order 5 000 V is applied between the anode and cathodes. A magnetic field of order 0.1 T is applied along the axial direction of the anode. An array of these cells forms a basic module. The module is located within the vacuum wall, which is surrounded by magnets located outside of the vacuum. This field creates an electromagnetic mirror, which traps electrons in circular orbits within the cell and increases their probability of colliding with and ionizing gas molecules. A number of modules can be connected to increase the total pumping speed. Active gases are pumped by burial in the titanium cathode or by reaction with titanium sputtered on the anode. Noble gases are only buried and have low pumping rates. Organic gases are first dissociated and then adsorbed. Ion pumps were the first to pump to ultrahigh vacuum without producing an organic contaminant. They are used in laboratory systems, high-power electron tubes, and small isolated systems and to supplement NEG pumps in large accelerator facilities.

2. GAS SOURCES

Gases and vapors to be removed from a vacuum vessel originate from within the volume of the chamber and from its internal surfaces. Volume sources are quite easily removed, whereas surface sources present the most serious obstacles in achieving low pressures.

2.1 Volume

The differential equation describing the time dependence of the pressure in a vacuum vessel during pumping of the gas within the chamber volume is given by

$$\frac{d(VP)}{dt} = -SP, \tag{1}$$

where VP is the volumetric gas quantity and S is the pumping speed. Volume V is measured in cubic meters, pressure P in pascals, and speed S in cubic meters per second. Alternatively, this equation may be recast in units commonly used in North America (pressure in units of Torr and speed in units of liters/second) or in units used in Europe (pressure in units of millibars, and speed in units of cubic meters per hour). VdP/dt is the rate at which gas flows from the chamber (the volumetric flow), and SP is the rate at which gas flows to the pump. These terms both have units of power—Pa·m^3/s or J/s. This "power" is the power dissipated in transporting the volumetric gas quantity. Volumetric pumping speed, flow, and gas quantity are not conservable. The volumetric gas quantity PV, according to the ideal gas law, equates to NkT; thus the number of molecules N in the volumetric quantity is dependent on the absolute temperature T. k is Boltzmann's constant. If the pumping speed is constant with pressure, the solution to (1) is

$$P = P_0 \exp(-St/V), \tag{2}$$

where $V/S = \tau$, the system time constant, is typically a few seconds, implying that a system could pump to a useful high vacuum pressure, say 10^{-5} Pa, from atmosphere in, say, 100 s. In practice it takes hours, because there is always gas released from real surfaces.

2.2 Surface

At low gas pressures the atoms and molecules originate from a number of surface sources. Gas adsorbed on the walls can be desorbed by thermal or other stimulated mechanisms. Gas dissolved within a solid can diffuse to the surface and desorb. Gas can permeate through some materials by the three-step processes of adsorption, diffusion, and desorption. Pumps are also sources of gas, as are chambers that contain trapped gas pockets or virtual leaks, such as those found at the end of a blind tapped screw thread, or closely spaced surfaces.

A quick mental exercise reveals the pressure at which volume gas sources become less important than surface gas sources. Envision a chamber containing no gas atoms in the volume and containing one atomic layer of adsorbed gas. The ideal gas law shows that the pressure associated with the instant release of this surface layer is of order 4 Pa for a 1-L chamber. Pressures below this range cannot be obtained without attention to reducing the surface gas load.

Surface gases release slowly. An approximate time dependence of the surface outgassing term Q_s is $Q_m t^{-n}$, where n ranges from 0.5 to 1.5 and m denotes the time at which the coefficient was measured. Permeation Q_p is constant with time. A quasistatic solution to (1) with these terms added is

$$P = P_0 e^{St/V} + (Q_m/S)t^{-n} + Q_P/S. \tag{3}$$

Outgassing and permeation data have been extensively reviewed in a number of textbooks and articles. Representative outgassing data are given in Dylla *et al.* (1993), Elsey (1975), and Nuovolone (1977). Peacock (1980) has reviewed the performance and selection of elastomer seals. Dylla *et al.* (1993) have described procedures for reducing the gas load in systems during pumping. Oxygen and water-vapor permeation data for new and established elastomers have been published by Ma *et al.* (1995).

3. GAS-FLOW FUNDAMENTALS

3.1 Molecular

Molecular flow occurs at low pressures, where Knudsen's number is >1. Knudsen's number is the ratio of mean free path to chamber dimension. The mean free path is the average distance that a molecule travels before colliding with another molecule. In molecular flow, molecules travel in straight lines and, because the mean free path is longer than a chamber dimension, collide with walls rather than other molecules. Molecules arriving at a wall lose their directed momentum and leave the wall in a diffuse (cosine) distribution, where the most likely direction of departure is normal to the surface. Molecular conductance, the gas flow through an object divided by the pressure difference across the object, is given by $C = A(v/4)a$ and is constant for pressures low enough for the mean free path to be longer than the dimension of the object, and is proportional to the cross-sectional area of the object. Here $v = (8kT/\pi m)^{1/2}$ is the thermal velocity, A is the cross-sectional area, and a is the transmission probability. In 1932 the transmission probability was first calculated for round tubes by Clausing (1971). Clausing's results, as well as the (Monte Carlo) transmission probabilities for many complex components, are found in most vacuum texts. For very long, round tubes the conductance is proportional to the third power of the diameter, because the transmission probability for long tubes is inversely proportional to l/d, the ratio of pipe length to pipe diameter.

Components in series have a net conductance less than any one component. Series conductances add as reciprocals, to within an accuracy of 20%. Oatley (1957) derived the following equation, which is valid to a few percent:

$$1/C_{\text{net}} = 1/C_1 + 1/C_2 \ldots + 1/C_m - (m - 1). \tag{4}$$

This relationship is valid for components with equal cross-sectional areas. Unequal areas are best dealt with by a method developed by Haefer (1980). The low series conductance of objects such as pipes, valves, and traps is a significant contributing factor to the low pumping speeds and slow pressure decrease observed in the molecular regime.

3.2 Viscous Laminar

Laminar flow is characterized by stream flow and occurs in pipes in which Knudsen's number is <0.01 and Reynolds number is <2 200—typically, in the pressure region 10^4 Pa to 0.1 Pa. For extremely long pipes, where the Hagen–Poiseuille equation is valid, the conductivity is proportional to the fourth power of the pipe diameter times the average pressure. However, this equation is not normally applicable to real pumping lines, as they are rarely long enough to meet the conditions of validity. Santeler (1986) has developed an excellent model where the viscous flow in a short tube is considered to be Poiseuille flow in series with a choked aperture.

3.3 Choke

Choke flow is the limiting turbulent flow through an orifice, determined by the speed of sound. Turbulent flow is chaotic flow characterized by a Reynolds number of >2 200 and a Knudsen's number <0.001. The magnitude of the choke flow through an orifice is proportional to the upstream pressure and the orifice area provided that the downstream pressure, for air, is $\leq 0.52\ P_{\text{upstream}}$. For other gases, this value will differ. The important concept is that the gas flow is limited by the speed of sound. This allows the construction of simple fixed-flow orifices, which are used in calibration fixtures as well as for improved safety on flammable gas cylinders. Choke flow is observed in the initial or "rough" pumping of a vacuum system. Often a choke is used to limit the flow until a pressure of approximately 0.1 atmosphere is reached, so that the chamber will not cool rapidly and cause condensation, as described in Sec. 4.3.

4. SYSTEM PERFORMANCE

A number of issues affect the performance of a vacuum system. The ultimate pressure of a pump is $P = Q/S$. The pumping speed can be doubled or tripled by use of a larger pump, but it is not physically possible to increase it much beyond that value; this is a costly and inefficient way to reduce pressure. Note that the conductances added in series with a diffusion pump (valve, trap, and pipe) typically reduce the speed at the chamber entrance to a value equal to $\frac{1}{3}$ of the basic diffusion pump speed. Gas sources contributing to the total load Q can be reduced by several orders of magnitude; therefore, anything that can be done to reduce the total load will contribute greatly to reducing the total pressure. Backstreaming, outgassing, permeation, surface treatments, conditioning procedures, exposures to ambient gases, and condensation of vapors are all issues that affect the total gas load. Doubling the speed of the pump would only decrease the ultimate pressure by a factor of 2.

4.1 Backstreaming

Backstreaming in displacement pumps results from a small but finite transmission probability in the reverse direction for exhausted gases and working fluids. Working-fluid vapors enter the chamber either from the inlet of the diffusion pump or from the inlet of the mechanical pump. Diffusion-pump vapors can be reduced to extremely small levels by use of a liquid-nitrogen cold trap and a low–vapor-pressure fluid and by not exceeding the critical inlet pressure. In systems not using a liquid-nitrogen trap, back-migration of diffusion-pump fluid is most easily reduced by using either pentaphenyl ether or pentaphenyl silicone, which have the lowest known vapor pressures of any commercially available fluid. These two fluids have vapor pressures, determined by extrapolation, of order 10^{-8} Pa at 25 °C. Contamination from the mechanical pump can be reduced by remaining in the viscous flow range where counterflow diffusion is dominant: >15 to 20 Pa in a 5- to 10-cm-diameter pipe. Residual-gas spectra from diffusion-pumped chambers invariably show that the contamination originates in the mechanical pump, so that correct mechanical pumping procedures are the single most important method of reducing organic vapor contamination.

In capture pumps backstreaming results from the release of entrapped gases. Backstreaming is often equated with organic fluid vapors; however, more generally it is any gas released by a pump that migrates to the chamber. In a cryopump it may be helium, hydrogen, or neon released from a warming surface. In an ion pump it may be release of buried argon. In a TSP or NEG pump it may be the retrograde solubility-limited release of hydrogen.

4.2 Outgassing and Permeation

Performance of representative systems can be estimated from three typical outgassing rates [units of Pa-m^3/(s-cm^2) or W]: unbaked stainless steel, aluminum, or Pyrex, 10×10^{-7}; baked stainless steel or aluminum, 2×10^{-11} (Ishimaru, 1993); and Viton elastomer, 1×10^{-3}. If one considers the three example systems described here, the relative roles of these materials become clear. Each example system has an internal area of 1 m^2, a pumping speed of 1 m^3/s, and, if using an elastomer O-ring, an added area of 100 cm^2. The calculated base pressures are 2×10^{-11} Pa (baked), 1×10^{-6} Pa (unbaked, no O-ring), and 1.1×10^{-5} Pa (unbaked with O-ring). To this we must add permeation. The permeation rate of Perspex is 3×10^{-12} m^2/s; thus, a 15-cm-diameter × 1-cm-thick Perspex window will raise the base pressure of this example system to 10^{-3} Pa. These simple examples illustrate the essence of vacuum technology: Ultrahigh-vacuum systems need to use all-metal gaskets, be cleaned extensively, and be thoroughly baked. The performance of unbaked vacuum systems is limited by the water permeation of elastomeric O-rings and water desorption from walls; dirty walls can contribute more outgassing than O-rings. Polymeric windows can result in significant water permeation.

A properly designed vacuum system should have short, large-diameter tubulations connecting the pumps and chamber, in order to reduce their conductance-limiting effect. Good design will also minimize locations where gas can become entrapped. Even with the best designs, which use load locks to prevent atmospheric water vapor from reaching the process chamber, the outgassing from all surfaces not baked will limit the purity of the process gas. Process gases, e.g., argon for sputtering, are available with impurity levels of order 1 part per billion (i.e., 10^{-9}). Even when gases of this purity are used, the outgassing load from the cleanest load-locked chamber will result in a net impurity in the parts-per-thousand range. Geyari (1976) discussed effects of corrosion on the properties of stainless steel.

4.3 Dynamic Effects

Adiabatic expansion causes gas cooling on sudden pumping. When gas expands, it does work on its surroundings and becomes colder. The ratio of initial to final temperature is related to the ratio of initial to final pressure in an adiabatic system. Real systems will act adiabatic initially but will soon become isothermal. That is, the chamber temperature will drop to a minimum of about -30 to -80 °C in about 0.5 to 1 s, corresponding to a pressure of about 80 000 Pa, and then recover to ambient in 20 to 30 s. During this brief temperature excursion, water and other condensable vapors become supersaturated and condense heterogeneously on ions or submicron particles or condense homogeneously. These (≈ 20 μm) droplets will fall and carry the entrapped particles to surfaces below. This effect, commonly observed when pumping from atmosphere, can also be seen when the sudden pressure drop is initiated from a low pressure, say 10 Pa; however, the amount of condensation is reduced proportionally. Several articles describe this effect in detail and provide solutions: Zhao *et al.* (1990), Wu *et al.* (1990), and O'Hanlon and Shieh (1990).

Pumping time will be increased if the chamber has had time to equilibrate with the ambient after opening. During the exposure to atmosphere, water will diffuse into material surfaces and elastomer gaskets. The amount of surface adsorption is related to the nature of the oxide layer on the chamber interior, and the amount of diffusion into elastomer gaskets is material dependent. Dayton (1995) and Ma *et al.* (1995) have developed models for surface and elastomer readsorption. Li and Dylla (1993) have shown that the amount of water adsorbed on stainless steel is proportional to the atmospheric exposure time and relative humidity. These studies quantitatively verify historically well-known practices. Most importantly, one should vent a system to atmosphere with nitrogen or other very dry gas. It is also important to minimize the atmospheric exposure time and to maintain an outward purge flow of dry nitrogen gas when transferring samples. In some systems, e.g., with large doors, the nitrogen flow is of no value. For those systems, one can maintain the walls at 40 to 50 °C to reduce readsorption.

5. VACUUM INSTRUMENTATION

Pressure and flow are two basic quantities measured routinely in vacuum. Both quantities need to be measured in ranges spanning as many as 16 decades, with the result that no one instrument can measure either quantity over this range. Additional quantities are derived from these measurements. Four total-pressure gauges, two partial-pressure gauges, and one flow meter are described here.

5.1 Total-Pressure Gauges

Total-pressure gauges either are direct—that is, they measure force on a solid or liquid surface—or they are indirect, and infer the pressure from a gas property, such as heat flow or ionization. A thorough discussion of total-pressure measurement has been given by Berman (1985).

5.1.1 Thermocouple and Pirani Gauges

Thermocouple and Pirani gauges are thermal-conductivity gauges, and they measure the rate of heat flow from a heated wire to a surrounding surface. At high pressure, the rate of heat flow due only to thermal conductivity is constant, and it decreases toward zero at low pressures. These gauges are useful in a range typically covering three orders of magnitude, from 0.1 to 100 Pa. In the thermocouple gauge, a wire is heated from an external source and the temperature is measured by means of a small thermocouple. See Fig. 7. It is also possible to measure the changing temperature of the wire by observing its change in resistance. This principal is used in the Pirani gauge, which has a shorter time response than the thermocouple gauge. Figure 8 sketches the heated wire used in the Pirani tube, which forms one arm of a Wheatsone bridge.

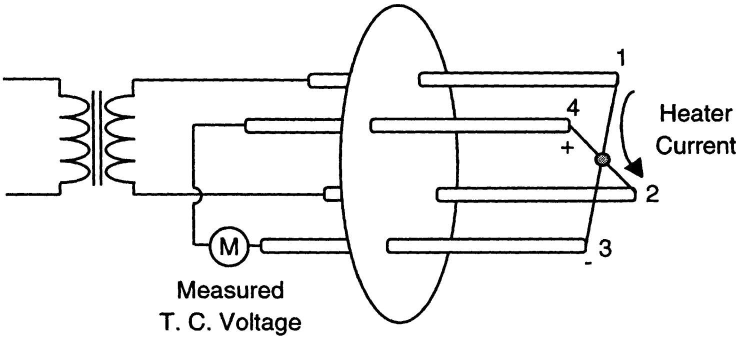

FIG. 7. Thermocouple gauge. A small ac voltage is applied between terminals 1 and 2, and the temperature is sensed across terminals 3 and 4. The two wires, 1–3 and 2–4, form a thermocouple at the weld junction.

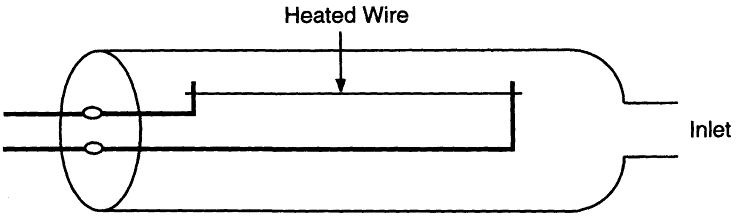

FIG. 8. Pirani gauge. The heated fine tungsten wire is strung between two support posts. An identical tube is exhausted and used as a balance arm in the bridge.

A version that can read to one atmosphere takes advantage of the small but measurable gas convection current flowing in the tube at high pressures. Thermal-conductivity gauges are not precise gauges but are extremely useful in monitoring the pressure of chambers and pumping lines in the region where pumps are first valved onto the chamber, and for monitoring the fore pressures of high-vacuum displacement pumps and the inlet pressures of oil-sealed pumps. Thermal-conductivity gauges are sensitive to gas species. All manufacturers can provide calibration curves.

5.1.2 Capacitance Manometer Total force is measured in a capacitance manometer by sensing the deflection of a thin diaphragm. A nearby, rigid counter-electrode forms the other half of a variable capacitor. See Fig. 9. This instrument can measure pressure over four orders of magnitude and is fabricated with full-scale deflections of 10^5, 10^4, 10^3, and 10^2 Pa. It can be used as a tertiary standard when the measuring head is temperature controlled. Capacitance manometers are typically constructed with the sense element on the vacuum side. These gauges are independent of gas species and are also useful for measuring corrosive and radioactive gases.

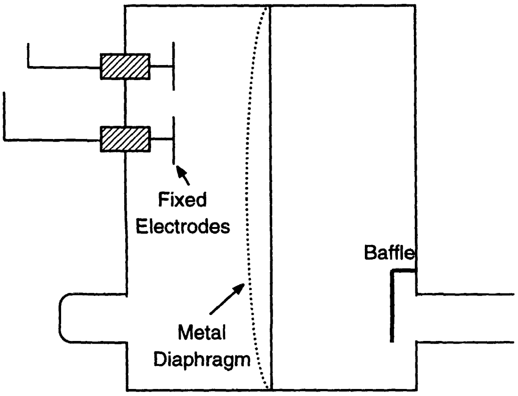

FIG. 9. Capacitance manometer. The metal diaphragm rests against the fixed electrodes at atmospheric pressure, whereas it is not deflected at its lowest pressure of operation.

5.1.3 Ionization Gauge The Bayard–Alpert hot-filament ionization gauge is the primary and most common high-vacuum gauge used in modern vacuum systems. It was developed over 47 years ago by Alpert and Bayard (1950). In the ionization gauge, sketched in Fig. 10, electrons are emitted from a hot filament, typically W or thoriated iridium, and collected on a cylindrical grid biased at a potential of 150 V. Molecules within this region have a probability of being ionized positively, after which they are collected by a fine wire cathode biased at 30 V negative with respect to the filament. External to the gauge tube, a sensitive amplifier is used to measure the collected ion current. The standard gauge tube is available in a glass tubulation or on a metal flange, and it will read pressures in the range 0.1 to 10^{-8} Pa. The lower pressure limit is due to electron-stimulated x-ray emission from the grid, which produces photoelectrons at the cathode. The ion gauge is sensitive to gas species, as ionization cross sections are species dependent. It is common practice to calibrate an ionization gauge for nitrogen (air) and to divide the meter reading by a relative sensitivity—that is, the sensitivity of the meter to the unknown gas divided by the sensitivity to nitrogen. Relative sensitivity data are available from the gauge manufacturer.

5.2 Residual-Gas Analyzers

Residual-gas analyzers (RGAs) are extremely useful diagnostic tools. They are commercially available with either magnetic

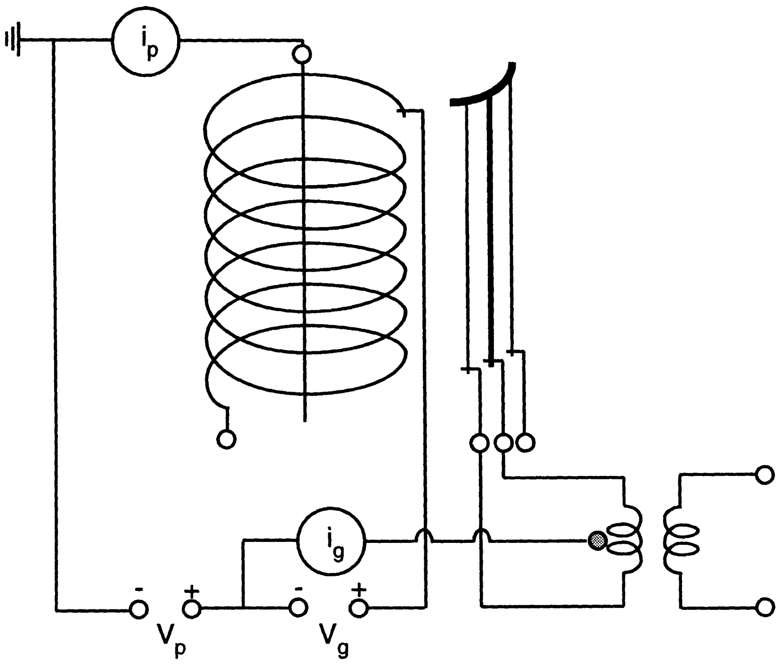

FIG. 10. Bayard–Alpert ionization gauge, showing the biasing potentials. Most gauges have two filaments, which can be electrically interchanged without removal from the system.

sector or quadrupole mass filters and are important in analyzing system problems, as well as determining the parameters required for a deposition, etch, or analytical process. A degree of skill is required to interpret the resulting mass spectra, as each individual molecular or atomic gas will have a cracking pattern determined by its isotopic abundance, degree of multiple ionization, and molecular fragmentation in the ion beam. One discussion of RGA analytical interpretation has been given by O'Hanlon (1989).

All RGAs have three subcomponents: an ionizer, a mass filter, and a detector. In all instruments the ionizer is a hot filament similar to that in a hot-filament ion gauge, with the exceptions that the ionization voltage will be typically 70 V; some ion sources are designed with shielding to minimize the ionization of desorbed gases from nearby walls. The detector will be a Faraday cup, often followed by an electron multiplier. Partial pressures of 10^{-7} and 10^{-12} Pa, respectively, are measurable with these detectors. The two mass filters differ considerably.

5.2.1 Quadrupole In one type, positive ions are axially injected in the region between four parallel rods, which form a radio-frequency mass filter. Only ions with the proper mass-to-charge ratio (m/e) exit the filter and are detected. The mathematical description of the filtering process has been reviewed by Dawson (1976). This mass filter is typically constructed with a resolution of one atomic mass unit, and with m/e ranges of 1 to several hundred. The mass-to-charge scans are linear and, thus, easy to read.

5.2.2 Magnetic Sector In the second type, positive ions leaving the ionizer are first accelerated to a constant voltage, and then directed to a magnetic sector where they travel in circular orbits following elementary electromagnetics. The orbit radii are given by $r = (1/B)(2mV/ze)^{1/2}$, where B is the magnetic field strength, V is the acceleration potential, and z is the number of elemental charges on the ion. Light ions are deflected more strongly than heavy ions, and the mass range is analyzed by scanning either the acceleration voltage or magnetic field while the filtered current is detected following an exit slit. Magnetic-sector instruments are less user friendly, in the sense that they need two magnets to analyze a wide m/e range and that the resulting plots are not linear with mass, unless the voltage or magnetic field are scanned in a nonlinear time-varying manner. However, they are less expensive than quadrupoles and are well suited for process control requiring the range below $e/m = 50$. In this range all the common atmospheric background gases—N_2, O_2, H_2O, CO, and CO_2—are easily observed.

5.3 Thermal Mass-Flow Meter

Thermal mass-flow meters are the most commonly used flow metering technique in modern systems over the range 0.01 to 10^{-5} Pa·L/s. They measure mass flow by heating the gas with known power and measuring its temperature rise with a thermocouple. Mass flow, the flow of gas through a pipe or orifice in conservable units, such as moles/second, is proportional to the heat input and inversely proportional to the heat capacity and temperature rise. The instrument is constructed around a small (0.25-mm-diameter) tube, which is heated at its center and thermally clamped at its ends. The symmetrically mounted thermocouples sense a temperature difference when gas is flowing, and this difference is converted to mass flow. Although the instrument is sensitive to mass flow, it is usually scaled in volumetric units (for nitrogen), which can be converted to mass flow. If the gas is other than nitrogen, the scale reading must be corrected by the ratio of the product heat capacity × density of nitrogen to that same quantity of the second gas.

5.4 Leak Detection

The primary instrument for detecting leaks in chambers whose pressures are $<10^{-2}$ Pa is the helium mass-spectrometer leak detector—a magnetic-sector instrument tuned to He, and containing its own high-vacuum, backing, and roughing pumps. Leaks are found by measuring the response to helium sprayed on the system exterior. He leak detectors respond to a helium flux as low as 10^{-9} Pa·L/s. Sensitivity and response time are important parameters. Sensitivity is determined by the fraction of the helium pumped through the detector, whereas the He response time is $\tau = V/S$.

Maximum sensitivity is obtained in a displacement-pumped system by connecting the leak detector to the foreline, valving off the forepump, and using the leak detector's pumps to back the system. In a capture pump this technique is not possible, so that maximum sensitivity is obtained, at the expense of response time, by connecting the leak detector to the chamber and reducing the main pump speed with a throttle valve. O-rings should be exposed only briefly to helium, 1 to 2 s, when leak checking, to minimize its permeation. In practice one leak checks from the top of the system and, if O-rings are used, periodically stops to allow permeated helium to outdiffuse from the elastomer. Permeation times are of order 15 to 20 min for most common elastomers.

Above 0.01 Pa alternative methods are used. Gross leaks can be observed by listening to the change in pitch of the mechanical pump when a leak is sprayed with He. Thermocouple and ion gauges are also sensitive to alcohol or He entering a leak. Cryopumps must be regenerated after helium leak checking. Helium leak detectors are available with "sniffer" probes. These instruments sample atmospheric air flowing through a capillary tube or porous orifice. They can detect helium exiting a leak in a pressurized vessel. Unless used in a helium-free environment, their sensitivity is limited by the background concentration of helium in the atmosphere.

5.5 Calibration Metrology

Rarely does one calibrate a pump, gauge, or flow meter. Speeds of mechanical and high-vacuum pumps are measured by manufacturers using specialized test domes. See, for example, O'Hanlon (1989). There are two simple methods, which are most helpful in checking system performance. If a mass flow meter is available, the net speed may be obtained by admitting a small volumetric flow Q and recording the change in pressure; $S = Q/(\Delta P)$ is the pumping speed. Without a flow meter, the speed may be obtained by measuring the temporal pressure increase after closing the main pumping valve. After pumping to the base pressure, a gas flow is admitted, raising the pressure well above background. The speed is given by $S = (V/P)dP/dt$. These two methods are extremely useful in checking performance quickly.

6. SAFETY

Safe operation of a vacuum system is the responsibility of the user. It is expected that the system installation and operational procedures will have been reviewed and approved by appropriate risk management personnel. Any system or component manufacturer will be able to answer questions regarding safe operation or installation. In this brief discussion, we note a few vacuum safety issues. Additional issues have been discussed by O'Hanlon and Fraser (1988).

The system controller should incorporate basic safety measures including disabling all high-voltage power supplies unless the chamber is closed and under vacuum; removal of power from all water-cooled pumps and furnaces on loss of cooling water; removal of power from diffusion and turbo pumps on loss of pumping or lubricating fluid; and closure of main chamber valves on loss of vacuum. Toxic, corrosive, or flammable gases need the approval of risk management prior to use. Other safety issues are pump or gauge specific. U.S. diffusion-pump makers prefer tetraphenyl silicone fluids, presumably for safety reasons. Cryopumps must have an operating safety gas release on the pump body, and in addition one should never mount a hot-filament gauge on a cryopump body, lest it ignite desorbed oxidizers and flammables. Convection-gauge readings are gas-species dependent, and it is possible to overpressurize a system when venting with certain gases—for example, Freons or argon. Substitutes for specialty pumping fluids, especially in turbo pumps, should never be used as they could damage the pump. Pump-oil vapors should always be vented into a proper exhaust duct. Liquid cryogens can absorb in clothing and cause serious burns. Lastly, never perform a helium leak check or maintenance on any system that has high-voltage power supplies operational. Insulators that are designed to withstand the voltage across an air gap will arc when surrounded by helium, because its breakdown voltage is only 1/10 that of air.

GLOSSARY

Backstreaming: The flow of a vapor or gas in the direction counter to the desired flow—for example, oil vapors from a mechanical pump to the chamber, hydrogen from an ion pump, process gas from a load lock to a process chamber.

Choke Flow: The limiting turbulent flow through an orifice, determined by the speed of sound.

Compression Ratio: The ratio of pump outlet pressure to inlet pressure.

Conductance (Volumetric): The ease with which gas can flow through an object; the volumetric flow divided by the pressure difference.

Flow (Mass): The flow of gas through a pipe or orifice given in conservable units such as moles/s, atoms/s, or kg/s. See also **Flow, Volumetric.**

Flow (Volumetric): The flow of gas through a pipe or orifice given in nonconservable units such as $\mathrm{Pa \cdot m^3/s}$, or $\mathrm{Torr \cdot L/s}$.

Knudsen's Number: The ratio of mean free path to chamber dimension.

Mean Free Path: The average distance a molecule travels before colliding with another, given by $\lambda = (2^{1/2} \pi d_0^2 n)^{-1}$.

Molecular Flow: Flow in a region in which the Reynolds number is <2 000 and Knudsen's number is >1.0. Characterized by molecular collisions with the walls of the system or pipe, i.e., wall–gas collisions dominate gas–gas collisions.

Pressure: The force on a wall divided by the wall area. $1\ \mathrm{N/m^2} = 1\ \mathrm{Pa}$, and $10^5\ \mathrm{Pa} = 760\ \mathrm{Torr} = 1\ 000\ \mathrm{mbar}$. Direct gauges measure force on a liquid or solid wall (capacitance manometer); indirect gauges measure force through some other property of the gas (thermal-conductivity gauge, ion gauge).

Speed (Volumetric): The ease with which gas can flow across a surface; the volumetric gas flow divided by the pressure at the plane of measurement.

Staging Ratio: The ratio of the displacement ratio of a turbopump or a lobe blower to that of its mechanical fore pump.

Turbulent Flow: Chaotic flow when the Reynolds number >2 200 and Knudsen's number <0.01.

Viscous Laminar Flow: Flow in a region of Reynolds number <2 000 and Knudsen's number <0.01. Characterized by laminar parallel streamlines gently flowing around obstructions.

Works Cited

Alpert, D. A., Bayard, R. T. (1950), in: W. Nottingham (Ed.), *Report on 10th Annual Conference on Physical Electronics*, Cambridge, MA: MIT

Press, pp. 88–93. See also Bayard, R. T., Alpert, D. A. (1950), *Rev. Sci. Instrum.* **21,** 571–572.

Berman, A. (1985), *Total Pressure Measurements in Vacuum Technology,* New York: Academic.

Clausing, P. (1932), *Ann. Phys.* **12** (5), 961–989. English translation (1971), *J. Vac. Sci. Technol.* **8,** 636–646.

Dawson, P. H. (Ed.) (1976), *Quadrupole Mass Spectrometry and Its Applications,* New York: Elsevier; reprint (1993), New York: American Institute of Physics.

Dayton, B. B. (1995), *J. Vac. Sci. Technol. A* **13,** 451–461.

Dylla, H. F., Manos, D. M., LaMarche, P. M. (1993), *J. Vac. Sci. Technol. A* **11,** 2623–2636.

Elsey, R. J. (1975), *Vacuum* **25,** 347–361.

Geyari, C. (1976), *Vacuum* **26,** 287–297.

Gifford, W. E. (1966), in: K. D. Timmerhaus (Ed.), *Adv. Cryog. Eng.* **11,** 152–159.

Haefer, R. (1980), *Vacuum* **30,** 217–223.

Ishimaru, H. (1993), *J. Vac. Sci. Technol. A,* **7,** 2439–2442.

Kruger, C. H., Shapiro, A. H. (1961), in: L. Talbot (Ed.), *Proceedings of the 2nd International Symposium on Rarefied Gas Dynamics, Berkeley, Calif, 1960,* New York: Academic, 117–140.

Li, M., Dylla, H. F. (1993), *J. Vac. Sci. Technol. A* **11,** 1702–1707.

Ma, C., Shero, E., Verma, N., Gilbert, S., Shadman, F. (1995), *J. IES* **38** (2), 43–46.

Nesseldreher, W. (1976), *Vacuum* **26** (7), 281–286.

Nouvolone, R. (1977), *J. Vac. Sci. Technol.* **14,** 1210–1212.

Oatley, C. W. (1957), *Brit. J. Appl. Phys.* **8,** 15–18.

O'Hanlon, J. F. (1989), *A User's Guide to Vacuum Technology,* 2nd ed., New York: Wiley.

O'Hanlon, J. F., Fraser, D. B. (1988), *J. Vac. Sci. Technol. A* **6,** 1226–1254.

O'Hanlon, J. F., Shieh, J. J. (1990), *J. Vac. Sci. Technol. A* **9,** 2802–2807.

Peacock, R. N. (1980), *J. Vac. Sci. Technol.* **17,** 330–336.

Santeler, D. J. (1986), *J. Vac. Sci. Technol. A* **4,** 338–343.

Whitman, C. B. (1987), *J. Vac. Sci. Technol. A* **5,** 255–261.

Wu, J. J., Cooper, D. W., Miller, R. J. (1990), *J. Vac. Sci. Technol. A* **8,** 1961–1968.

Zhao, J., Liu, B. Y. H., Kuehn, T. H. (1990), *Solid State Technol.* **33,** 85–89.

Further Reading

Dushman, S. (1980), in: J. M. Lafferty (Ed.), *Scientific Foundations of Vacuum Technique,* 2nd ed., New York: Wiley.

Espe, W. (1966), *Materials of High Vacuum Technology,* Vol. 2, New York: Pergamon Press.

Hablanian, M. H. (1990), *High Vacuum Technology, A Practical Guide,* New York: Marcel Dekker.

Kohl, W. H. (1967), *Handbook of Materials and Techniques for Vacuum Devices,* New York: Reinhold; reprint (1993), New York: American Institute of Physics.

O'Hanlon, J. F. (1989), *A User's Guide to Vacuum Technology,* 2nd ed., New York: Wiley.

Power, B. D. (1966), *High Vacuum Pumping Equipment,* New York: Reinhold.

Roth, J. P. (1982), *Vacuum Technology,* Amsterdam: North-Holland.

Saksaganski, G. L. (1988), *Molecular Flow in Complex Vacuum Systems,* New York: Gordon and Breach.

Santeler, D. J., *et al.* (1966), *Vacuum Technology and Space Simulation,* NASA SP-105, Washington, DC: National Aeronautics and Space Administration; reprint (1993), New York: American Institute of Physics.

Weston, G. F. (1985), *Ultrahigh Vacuum Practice,* London: Butterworths.

White, F. A. (1968), *Mass Spectroscopy in Science and Technology,* New York: Wiley.

VARIATIONAL METHODS

G. W. F. Drake, *Department of Physics, University of Windsor, Windsor, Ontario, Canada*

Introduction 121
1. **Techniques of the Calculus of Variations** 122
1.1 Variations with Constraints 124
1.2 Generalizations 124
2. **Applications to Classical Mechanics** 125
2.1 Introductory Concepts 125
2.2 Hamilton's Principle 126
2.2.1 Conservative Systems and First Integrals 127
2.3 The Hamilton–Jacobi Equation 127
2.3.1 The Principle of Least Action 128
2.3.2 Hamilton's Equations of Motion 129
2.3.3 Canonical Transformations 130
2.3.4 Interpretation of the Hamilton–Jacobi Equation 131
2.4 Relativistic Generalization 132
2.4.1 Inclusion of Electromagnetic Fields 132
3. **Applications to Quantum Mechanics** 132
3.1 Variational Derivation of the Schrödinger Equation 132
3.2 The Rayleigh–Schrödinger Variational Principle 133
3.3 The Rayleigh–Ritz Variational Method 133
3.3.1 Algebraic Solution for Linear Variational Parameters 134
3.3.2 Extension to Excited States 135
3.3.3 Variational Lower Bound 135
3.3.4 Illustrative Results for Helium 136
3.3.5 Extensions to More Complex Systems 138
3.4 Variation–Perturbation Methods 139
3.4.1 Variational Bounds 139
3.4.2 Spectral Representations and Pseudostates 140
3.4.3 Example: The Polarizability of Hydrogen 141
3.4.4 Time-Dependent Perturbations 141
4. **The General Sturm–Liouville Problem** 142
4.1 The Oscillation Theorem 142
4.2 Example: The Coulomb Problem 142
5. **Applications to Electrodynamics** 143
6. **Feynman's Path Integral** 144
6.1 Relation to Classical Dynamics 146
Glossary 146
Works Cited 148
Further Reading 149

INTRODUCTION

Variational principles derive from a certain aesthetic and metaphysical ideal of simplicity in the search for the principles underlying physical phenomena. The origins date back to the earliest Greek philosophers Thales (c. 600 B.C.) and Pythagoras (c. 550 B.C.). Aristotle (384–322 B.C.) clearly makes use of a variational principle to justify circular orbits for the planets when he says in *de Caelo II*

> Now of lines which return upon themselves, the line which bounds the circle is the shortest, and that movement is the swiftest which follows the shortest line.

This marks the first use of a "minimum" postulate, and the conclusion held sway until the time of Kepler (1571–1630). Hero of Alexandria (c. 125 B.C.) made the first rigorous use of a variational principle when he proved that when the angle of incidence equals the angle of reflection, the path taken by a ray of light from the object to the observer is shorter than any other possible path with fixed end points (see *Catoptrics* by Hero in

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Cohen and Drabkin, 1958). This later became Fermat's principle of least time in geometrical optics.

The belief that nature is in some sense "simple" and can be explained by some economically small number of postulates pervades the works of Galileo (1564–1642), Newton (1642–1727), and Leibniz (1646–1716). Although some of the early conclusions turned out to be scientifically unfounded, the philosophical basis for variational principles has great antiquity. They continue to guide the development of new physical theories at the most fundamental level and to provide powerful methods of practical computation. Perhaps most importantly, they bring out the structural analogies between superficially different phenomena and allow techniques developed in one field to be readily applied in another. At a fundamental level, practically all physical phenomena can be expressed in terms of variational principles that have a striking similarity.

The *calculus of variations* provides the basic mathematical tool for formulating and analyzing variational principles. The purposes of this article are first to give an overview of the calculus of variations, and then to discuss its application to a variety of physical phenomena. There is a vast literature on both aspects, and only a few of the most important points can be covered in the space available. Only a few principal references are given, with further general references in the reading list at the end. The main emphasis is on applications to classical mechanics and bound-state problems in quantum mechanics. Except for a brief discussion of the Feynman path integral, applications to scattering problems are not covered. An informal and instructive introduction to variational methods can be found in Hildebrand and Tromba (1985).

1. TECHNIQUES OF THE CALCULUS OF VARIATIONS

In its simplest form, the calculus of variations addresses the problem of finding the function $y(x)$ for which the integral

$$J = \int_{x_1}^{x_2} f(x,y,y_x)dx \tag{1}$$

is an extremum. The integrand $f(x,y,y_x)$ is some prescribed function of the indicated variables, where $y_x = dy/dx$, and x_1, x_2 are fixed end points. J is termed a *functional* of $y(x)$. As originally formulated by Euler, the problem is solved by considering infinitesimal variations $\delta y(x)$ about a particular path $y(x)$ connecting x_1 and x_2 (see Fig. 1) and demanding that the variation δJ induced in J vanish. For example, one might choose the variations to be $\delta y(x) = \epsilon\eta(x)$, where $\eta(x)$ is an arbitrary function such that $\eta(x_1) = \eta(x_2) = 0$ (to make the variation vanish at the end points), and ϵ is a small parameter controlling the size of the variation. Then $\delta y_x(x) = \epsilon d\eta(x)/dx$, and from a Taylor series expansion of $f(x,y,y_x)$ about $\epsilon = 0$ in Eq. (1), the induced variation in J is

$$\delta J = \epsilon \int_{x_1}^{x_2} \left[\frac{\partial f}{\partial y}\eta(x) + \frac{\partial f}{\partial y_x}\frac{d\eta(x)}{dx}\right]dx \tag{2}$$

up to terms of first order in ϵ. An integration of the second term by parts yields

$$\delta J = \epsilon \frac{\partial f}{\partial y_x}\eta(x)\Big|_{x_1}^{x_2} + \epsilon \int_{x_1}^{x_2} \left[\frac{\partial f}{\partial y} - \frac{d}{dx}\frac{\partial f}{\partial y_x}\right]\eta(x)dx. \tag{3}$$

The first term vanishes by the assumption that $\eta(x) = 0$ at the end points. Since $\eta(x)$ is otherwise an arbitrary function, the condition $\delta J = 0$ can be fulfilled only if the integrand of the second term vanishes identically for $x_1 < x < x_2$; i.e.,

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\frac{\partial f}{\partial y_x} = 0. \tag{4}$$

This is the basic Euler–Lagrange equation. For purposes of compactness, the functional

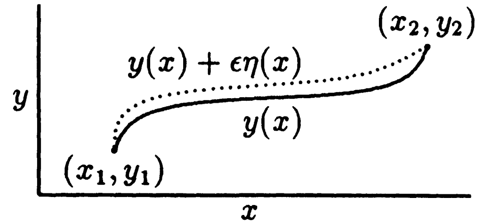

FIG. 1. Illustration of the actual path $y(x)$ and the varied path connecting fixed end points.

dependence of f on its various arguments is usually suppressed, as is done here.

Any solution to Eq. (4) satisfies the variational condition $\delta J = 0$. However, J itself could be a minimum, a maximum, or a point of inflection. One often knows from the geometrical nature of the problem being solved which case applies. Otherwise, it is necessary to extend the Taylor series expansion of $f(x,y,y_x)$ to terms of order ϵ^2 and determine the sign of the second-order variation $\delta^{(2)}J$ (see, e.g., Courant and Hilbert, 1966). If $\delta^{(2)}J < 0$, then J is a maximum; if $\delta^{(2)}J > 0$, then J is a minimum; if $\delta^{(2)}J = 0$, then J lies at a point of inflection.

If the end points are not fixed, then $\eta(x)$ does not vanish there. In this case, Eq. (4) still applies, subject to the condition that $\partial f/\partial y_x = 0$ at the end points (see Jeffreys and Jeffreys, 1972).

An important special case of Eq. (4) occurs if $f(x,y,y_x)$ does not depend explicitly on x because then the integrating factor is simply y_x. After multiplying Eq. (4) through by y_x and using

$$\frac{df}{dx} = \frac{\partial f}{\partial y} y_x + \frac{\partial f}{\partial y_x}\frac{dy_x}{dx} \tag{5}$$

(since $\partial f/\partial x = 0$ by assumption), the Euler–Lagrange equation becomes

$$\frac{d}{dx}\left(y_x \frac{\partial f}{\partial y_x} - f\right) = 0, \tag{6}$$

and so

$$y_x \frac{\partial f}{\partial y_x} - f = \text{const.} \tag{7}$$

A classic example is provided by the brachistochrone (shortest time) problem first propounded by John Bernoulli in 1696. It was solved by both him and his brother James, as well as by Newton and Leibnitz. Consider a bead sliding without friction on a wire of arbitrary length connecting two fixed points (x_1,y_1) and (x_2,y_2) in a vertical plane, as shown in Fig. 2. The problem is to find the shape that minimizes the travel time as the bead slides from rest under the force of gravity; i.e., to find the function $y = y(x)$ such that the integral

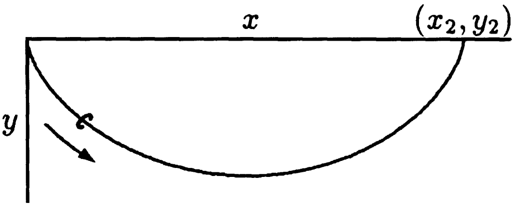

FIG. 2. The brachistochrone problem of a bead sliding without friction on a wire of arbitrary length in a vertical plane.

$$\tau = \int_{x_1}^{x_2} \frac{ds}{v} \tag{8}$$

for the travel time is a minimum. Here $ds = [(dy/dx)^2 + 1]^{1/2}dx$ is the element of arc length and v is the velocity. By conservation of energy, the velocity after falling a distance y (measured downward) is $v = \sqrt{2gy}$ (independent of x), where g is the acceleration due to gravity. The integral to be minimized is then

$$\tau = \int_{x_1}^{x_2} \frac{(y_x^2 + 1)^{1/2}}{(2gy)^{1/2}}\, dx. \tag{9}$$

With $f(x,y,y_x)$ defined by the above integrand, the Euler–Lagrange equation (7) then gives

$$\frac{y_x^2}{(y_x^2 + 1)^{1/2} y^{1/2}} - \frac{(y_x^2 + 1)^{1/2}}{y^{1/2}} = c^{1/2}, \tag{10}$$

and hence

$$y_x^{-1} \equiv \frac{dx}{dy} = \left(\frac{cy}{1 - cy}\right)^{1/2}, \tag{11}$$

where c is a constant of integration that determines the distance scale. Assuming that the bead starts from the origin, integration of this equation yields

$$cx = \sin^{-1}(cy)^{1/2} - (cy - c^2y^2)^{1/2}. \tag{12}$$

This is the equation of a cycloid symmetric about the minimum at $cy = 1$. The scale factor c is determined by the condition that the curve pass through the second terminus (x_2,y_2). If $y_2 = 0$ (as in Fig. 2), then $c = \pi/x_2$.

1.1 Variations with Constraints

There are many classes of problems where the functional J must be made an extremum subject to a subsidiary condition of the form

$$\int_{x_1}^{x_2} f_1(x,y,y_x)dx = \text{const.} \tag{13}$$

The example of a hanging chain of fixed length is discussed in the following paragraph. Such problems can be handled by applying the variational procedure to the functional

$$g(x,y,y_x) = f(x,y,y_x) + \lambda_1 f_1(x,y,x_y), \tag{14}$$

where λ_1 is called a *Lagrange undetermined multiplier.* The resulting Euler–Lagrange equation is

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\frac{\partial f}{\partial y_x} = -\lambda_1\left(\frac{\partial f_1}{\partial y} - \frac{d}{dx}\frac{\partial f_1}{\partial y_x}\right), \tag{15}$$

or, if f and f_1 are independent of x, the first integral is [cf. Eq. (7)]

$$y_x\frac{\partial f}{\partial y_x} - f + \lambda_1\left(y_x\frac{\partial f_1}{\partial y_x} - f_1\right) = \text{const.} \tag{16}$$

The idea is to solve Eq. (15) or Eq. (16) for a fixed but arbitrary value of λ_1. The equation of constraint provides the additional condition to determine λ_1 at the end of the problem, together with the two constants of integration.

As an example, consider the problem of finding the shape of a uniform hanging chain with both ends fixed. The shape is such that the potential energy due to gravity is a minimum, and so the quantity to be minimized is

$$J = \int_{x_1}^{x_2} \mu g y(y_x^2 + 1)^{1/2}dx, \tag{17}$$

subject to the constraint

$$\int_0^L ds \equiv \int_{x_1}^{x_2} (y_x^2 + 1)^{1/2}dx = L, \tag{18}$$

where μ is the mass per unit length and L the length. Thus $f = \mu g y(y_x^2 + 1)^{1/2}$ and $f_1 = (y_x^2 + 1)^{1/2}$. After dividing by μg and defining $\lambda = \lambda_1/\mu g$, Eq. (16) gives

$$(y + \lambda)\left[\frac{y_x^2}{(y_x^2 + 1)^{1/2}} - (y_x^2 + 1)^{1/2}\right] = -\frac{1}{c}. \tag{19}$$

As for the brachistochrone problem, this equation can be solved for $y_x^{-1} \equiv dx/dy$ and the result integrated to obtain

$$cy = -c\lambda + \cosh[c(x - a)], \tag{20}$$

where a is the second constant of integration. The three parameters a, c, and λ are determined by the three conditions that the curve pass through the points (x_1,y_1) and (x_2,y_2) at the ends of the chain, together with the equation of constraint. For example, if the two points are $(0,0)$ and $(x_2,0)$, then $a = x_2/2$, $c\lambda = \cosh(cx_2/2)$, and the equation of constraint becomes

$$\begin{aligned} L &= \int_0^{x_2} \{\sinh^2[c(x - \tfrac{1}{2}x_2)] + 1\}^{1/2}dx \\ &= \frac{2}{c}\sinh(cx_2/2). \end{aligned} \tag{21}$$

Solving this transcendental equation determines the remaining constant c. The quantity $T_0 = \mu g/c$ is the force of tension in the chain at the lowest point, where $y_x = 0$.

1.2 Generalizations

In applications to mechanics (see Sec. 2.1), the time t plays the role of the independent variable x marking the evolution of the system, but there are typically several dependent generalized coordinates q_i and $\dot{q}_i \equiv dq_i/dt$, $i = 1, \ldots, N$, in place of y and y_x, respectively. With this change in notation, f becomes a function of all the q_i's, $\dot{q}_i$'s, and t, and the generalization of Eq. (4) is

$$\frac{\partial f}{\partial q_i} - \frac{d}{dt}\frac{\partial f}{\partial \dot{q}_i} = 0, \quad i = 1, \ldots, N. \tag{22}$$

There are thus N coupled Euler–Lagrange equations, one for each degree of freedom of the system.

If in addition there are several independent variables $t_1, t_2, \ldots, t_r$, then Eq. (22) is further generalized to read

$$\frac{\partial f}{\partial q_i} - \sum_{j=1}^{r} \frac{\partial}{\partial t_j} \frac{\partial f}{\partial(\partial q_i/\partial t_j)} = 0, \quad i = 1, \ldots, N. \tag{23}$$

There can also be several equations of constraint of the form of Eq. (13) with integrands $f_k(t_j,q_i,\dot{q}_i)$, $k = 1, \ldots, m$. In this case, m Lagrange undetermined multipliers λ_k are introduced, and the function f in Eq. (22) or (23) is replaced by

$$g(t_j,q_i,\dot{q}_i) = f(t_j,q_i,\dot{q}_i) + \sum_{k=0}^{m} \lambda_k f_k(t_j,q_i,\dot{q}_i). \tag{24}$$

Constraints that can be expressed in integrated form, such as Eq. (13), are said to be *holonomic* (wholly named or specified). However, problems often arise in mechanics involving *nonholonomic* constraints that can only be expressed in differential form, such as a relation between velocities. An example is the problem of a vertical disk of radius R rolling without slipping on a plane. Four coordinates are required—the (x,y) Cartesian coordinates of the point of contact between the disk and the plane, a spinning angle of rotation θ about a vertical axis, and a rolling angle ϕ about an axis perpendicular to the disk. If the plane of the disk is initially perpendicular to the x axis (i.e., $\theta = 0$), then the constraint of "not slipping" corresponds to the differential relations

$$dx = R \sin\theta \, d\phi, \tag{25}$$

$$dy = -R \cos\theta \, d\phi. \tag{26}$$

These equations cannot be integrated without knowing in advance θ and ϕ as a function of t. However, the method of Lagrange undetermined multipliers can still be applied. If the general form of the differential constraints is written as

$$\sum_{i=1}^{N} a_{k,i} dq_i + a_{k,t} dt = 0, \quad k = 1, \ldots, s, \tag{27}$$

where the coefficients $a_{k,i}$ are, in general, functions of the q_i's and $\dot{q}_i$'s, then the generalization of Eq. (22) for nonholonomic systems is

$$\frac{\partial f}{\partial q_i} - \frac{d}{dt}\frac{\partial f}{\partial \dot{q}_i} + \sum_{k=1}^{s} \lambda_k a_{k,i} = 0, \quad i = 1, \ldots, N. \tag{28}$$

This equation can still be used even if the constraints are in fact holonomic. The term $\Sigma_{k=1}^{s} \lambda_k a_{k,i}$ can be identified with the generalized forces of constraint. In general, the λ_k are now functions of the q_i's and $\dot{q}_i$'s. The term $a_{k,t} dt$ in Eq. (27) does not contribute to Eq. (28) because the variations δq_i from the actual path are considered to occur at a particular instant of time (see Sec. 2.1).

As a simple illustration of the versatility of the method, consider the problem of finding the function $u(x,y,z)$ such that the square of its gradient is a minimum in a given volume of space. The problem is then to minimize

$$J = \iiint f dx dy dz, \tag{29}$$

with

$$f = (\nabla u)^2 = \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial u}{\partial z}\right)^2. \tag{30}$$

This can be regarded as an application of Eq. (23) with the three independent variables $t_1 = x$, $t_2 = y$, and $t_3 = z$, and a single degree of freedom ($N = 1$) with $q_1 = u$. Equation (23) then immediately gives

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0, \tag{31}$$

which is Laplace's equation. The term $\partial f/\partial q_1 \equiv \partial f/\partial u$ does not contribute because f does not depend explicitly on u, only on the partial derivatives $\partial u/\partial x$, $\partial u/\partial y$, and $\partial u/\partial z$. This problem clearly illustrates the way in which functional derivatives are to be interpreted and should be carefully studied.

2. APPLICATIONS TO CLASSICAL MECHANICS

The elements and techniques of classical mechanics are covered under MECHANICS, CLASSICAL. This section draws together the connections with variational principles.

2.1 Introductory Concepts

Consider a classical system of n interacting particles having masses m_s located at positions $\mathbf{r}_s = (x_{s,1}, x_{s,2}, x_{s,3})$, and acted on by forces $\mathbf{F}_s = (F_{s,1},F_{s,2},F_{s,3})$ due to the other particles and any external forces, including forces of constraint. The evolution of the system is obtained by solving Newton's equations of motion

$$m_s\ddot{x}_{s,j} = F_{s,j}, \quad s = 1, \ldots, n; j = 1, 2, 3. \tag{32}$$

These equations are completely general. However, it is often more convenient to replace the $3n$ Cartesian coordinates $x_{s,j}$ by *generalized coordinates* q_i defined through a system of connection equations of the form

$$x_{s,j} = x_{s,j}(q_1,q_2, \ldots ,q_{3n}). \tag{33}$$

The use of generalized coordinates is particularly effective in problems involving constraints. If the generalized coordinates are chosen such that their variations δq_i do no *virtual work* against the forces of constraint (i.e., δq_i is perpendicular to the *instantaneous* forces of constraint), then the number of independent q_i needed is reduced from $3n$ to $3n - m$, where m is the number of constraints.

To make these ideas concrete, consider the example of a bead sliding on a vertical wire hoop of radius R that is itself constrained to rotate about the z axis with angular velocity ω. In terms of the polar angles θ and $\phi = \omega t$, the connection equations are

$$\begin{aligned} x &= \sin\theta \cos\omega t, \\ y &= \sin\theta \sin\omega t, \\ z &= \cos\theta. \end{aligned} \tag{34}$$

The constraint of sliding on the hoop has reduced the number of independent coordinates from three to a single azimuthal angle θ. A variation $\delta\theta$ generates a displacement of the bead consistent with the instantaneous orientation of the hoop, but not consistent with the actual time evolution of the system, which includes the rotation of the hoop. Such variations are said to do no virtual work, and are called *virtual displacements.* From the connection equations (34), the kinetic energy of the bead is

$$\begin{aligned} T &= \tfrac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) \\ &= \tfrac{1}{2}m[(R\dot{\theta})^2 + (\omega R \sin\theta)^2]. \end{aligned} \tag{35}$$

If the system is conservative with a potential-energy function $V(x,y,z)$, then V can similarly be expressed in terms of θ.

2.2 Hamilton's Principle

In the absence of constraints, a direct transformation of Newton's equations of motion (32) from the $3n$ Cartesian coordinates $x_{s,j}$ to the $3n$ generalized coordinates q_i yields Lagrange's equations of motion

$$\frac{d}{dt}\frac{\partial T}{\partial \dot{q}_i} - \frac{\partial T}{\partial q_i} = Q_i, \tag{36}$$

where the Q_i are the generalized forces defined by

$$Q_i = \sum_{s=1}^{n} \mathbf{F}_s \cdot \frac{\partial \mathbf{r}_s}{\partial q_i}. \tag{37}$$

A comparison with Eqs. (3) and (22) shows immediately that with the identification $f = -T$, Lagrange's equations correspond to the variational condition

$$\int_{t_1}^{t_2} \left(\delta T + \sum_{i=1}^{3n} Q_i \delta q_i\right) dt = 0. \tag{38}$$

This is the most general form of Hamilton's principle in classical dynamics. The advantage gained is that $3n$ equations of motion have been consolidated into a single scalar variational condition that is invariant under coordinate transformation. In the absence of constraints, all the q_i can be varied independently so that each coefficient of δq_i must vanish separately, and Lagrange's equations are recovered. If m constraints are present, one need keep only $3n - m$ of the q_i whose virtual displacements δq_i are consistent with the instantaneous constraints, as in the rotating-hoop example of Sec. 2.1 where only a single parameter θ is required. One can still consider variations in the remaining q_i, even though they would violate the constraints. The only difference is that the corresponding Q_i are reinterpreted as the generalized forces required to maintain the constraints. They can be calculated by the method of Lagrange undetermined multipliers, as described in Sec. 1.2 [see especially Eq. (28)]. In this way, Lagrange's equations apply to the entire set, whether or not constraints are present. The invariance of Lagrange's equations and Hamilton's principle under coordinate transformation guarantees that if they are correct in Cartesian coordinates (as can easily be checked), they are correct in any other system of generalized coordinates.

2.2.1 Conservative Systems and First Integrals If the system is conservative, then the Q_i are derivable from a potential function $V(q_i)$ according to $Q_i = -\partial V/\partial q_i$, and Lagrange's equations reduce to

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} = 0, \tag{39}$$

where $L = T - V$ is the Lagrangian. It is then immediately evident from Eq. (4) that Hamilton's principle becomes

$$\delta \int_{t_1}^{t_2} L\, dt = 0. \tag{40}$$

This is the most useful form of Hamilton's principle for theoretical discussion.

If T and V do not involve time explicitly, then the first integral corresponding to Eq. (6) is

$$\sum_i \dot{q}_i \frac{\partial T}{\partial \dot{q}_i} - T + V = \text{const.} \tag{41}$$

If, further, T is a homogeneous quadratic function of the $\dot{q}_i$, then the above becomes

$$T + V = \text{const.}, \tag{42}$$

where the constant can now be identified with the conserved energy E of the system.

For the example of a bead on a rotating hoop discussed in Sec. 2.1, the first integral gives

$$\begin{aligned} &\tfrac{1}{2}m[(R\dot{\theta})^2 - (\omega R\sin\theta)^2] + V(\theta) \\ &\quad = T + V - m(\omega R\sin\theta)^2 = \text{const.} \end{aligned} \tag{43}$$

The constant of the motion here differs from $E = T + V$ because the time-dependent forces of constraint do work on the system. Overall conservation of energy is recovered only when the work required to keep the hoop rotating at a constant rate is included. If N_z is the required torque, then, with the use of Eq. (43), the rate at which it does work is

$$N_z\omega = \frac{d}{dt}(T + V) = \frac{d}{dt}m(\omega R\sin\theta)^2 = \omega\frac{dl_z}{dt}, \tag{44}$$

where $l_z = m\omega(R\sin\theta)^2$ is the angular momentum of the bead about the z axis. Thus $N_z = dl_z/dt$ as expected.

2.3 The Hamilton–Jacobi Equation

The variations considered thus far are taken between fixed end points t_1 and t_2 and, therefore, necessarily do not correspond to a possible dynamical evolution of the system. The actual evolution between fixed end points is uniquely defined, at least with respect to local variations, and so the varied path is unphysical. (But it is not necessarily so for nonlocal variations. For example, two points on a Kepler orbit are connected by two paths, depending on which way around the particle goes.)

The Hamilton–Jacobi equation comes from consideration of a different kind of variation Δq_i along a possible dynamical path between points allowed to vary in both space and time. In this case, it is necessary to keep the integrated term in Eq. (3) (or its generalizations). Suppose that t_1 and t_2 are replaced by $t_1 + \delta t_1$ and $t_2 + \delta t_2$. There is then a corresponding variation δt in the arrival time at each point along the path, so that $\Delta q_i(t) \equiv \delta q_i(t + \delta t)$ is the variation in path evaluated at the modified arrival time.

For definiteness, suppose that the paths are parameterized according to

$$q_i(\epsilon,t) = q_i(0,t) + \epsilon\eta_i(t), \tag{45}$$

where $q_i(0,t)$ is the actual path and $\eta_i(t)$ is an arbitrary differentiable function not assumed to vanish at the end points. Then $\delta q_i(t) = \epsilon\eta_i(t)$ and

$$\begin{aligned} \delta q_i(t + \delta t) &= q_i(\epsilon, t + \delta t) - q_i(0,t) \\ &\simeq q_i(0, t + \delta t) - q_i(0,t) + \epsilon\eta_i(t) \\ &= \dot{q}_i\delta t + \delta q_i(t) \end{aligned} \tag{46}$$

up to terms of first order in ϵ and δt. Application of the Δ variation to the integral in Hamilton's principle then yields

$$
\begin{aligned}
\Delta \int_{t_1}^{t_2} L\,dt
&= \int_{t_1+\delta t_1}^{t_2+\delta t_2} L\,dt - \int_{t_1}^{t_2} L\,dt + \int_{t_1}^{t_2} \delta L\,dt \\
&= L\delta t\Big|_{t_1+\delta t_1}^{t_2+\delta t_2} + \sum_i \frac{\partial L}{\partial \dot{q}_i}\,\delta q_i\Big|_{t_1}^{t_2} \\
&\quad + \int_{t_1}^{t_2} \sum_i \left(\frac{\partial L}{\partial q_i} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i}\right)\delta q_i dt. \qquad (47)
\end{aligned}
$$

The last term vanishes by the assumption that the varied path is a possible dynamical path, and so Lagrange's equations of motion are satisfied. Interest therefore centers on the integrated terms. The first term is simply $L(\delta t_2 - \delta t_1)$. Using Eq. (46) to replace $\delta q_i(t)$ by $\delta q_i(t + \delta t)$ in the second term, these terms become

$$
\Delta \int_{t_1}^{t_2} L\,dt = \left[\left(L - \sum_i \dot{q}_i \frac{\partial L}{\partial \dot{q}_i}\right)\delta t + \sum_i \frac{\partial L}{\partial \dot{q}_i}\,\delta q_i\right]_{t_1+\delta t_1}^{t_2+\delta t_2}. \qquad (48)
$$

With the definitions

$$
p_i = \frac{\partial L}{\partial \dot{q}_i}, \qquad (49)
$$

$$
H = \sum_i \dot{q}_i \frac{\partial L}{\partial \dot{q}_i} - L, \qquad (50)
$$

$$
S = \int_{t_1}^{t_2} L\,dt, \qquad (51)
$$

where the integral for S is taken along any dynamical path, Eq. (48) can be written in the form

$$
\Delta S = \left[-H\delta t + \sum_i p_i \delta q_i\right]_{t_1+\delta t_1}^{t_2+\delta t_2}. \qquad (52)
$$

H is called the *Hamiltonian* for the system, p_i is the *canonical momentum*, and S is called *Hamilton's principal function*. For an ordinary conservative system, $H = T + V$ is the total energy.

Equation (52) may be interpreted as follows. Since by assumption the system evolves along a possible dynamical path from an initial configuration $(q_i)_1$ at time t_1 to a final configuration $(q_i)_2$ at time t_2, there will in general be only a single set of initial velocities $(\dot{q}_i)_1$ that satisfy these requirements (at least with respect to small variations). The same applies to the initial momenta $(p_i)_1$ since they are connected to the $(\dot{q}_i)_1$ through Eq. (49), and hence the $(\dot{q}_i)_1$ can be eliminated from the problem. The p_i's as well as the q_i's are therefore uniquely determined from the initial conditions at time t_1. Since Eq. (52) remains true as t_2 is varied, it follows that

$$
\frac{\partial S}{\partial q_i} = p_i, \qquad (53)
$$

$$
\frac{\partial S}{\partial t} = -H(q_i,p_i,t) = -H\left(q_i,\frac{\partial S}{\partial q_i},t\right). \qquad (54)
$$

Equation (54) is called the *Hamilton–Jacobi* equation. It is a first-order partial differential equation in $N + 1$ variables and does not involve S explicitly. As it stands, a complete solution involving $N + 1$ constants of integration is determined from the initial conditions

$$
\left(\frac{\partial S}{\partial t}\right)_{t_1} = (H)_{t_1}, \quad \left(\frac{\partial S}{\partial q_i}\right)_{t_1} = -(p_i)_{t_1}. \qquad (55)
$$

The real significance of the Hamilton–Jacobi equation comes not from its practical utility in solving mechanical problems, but for the insight it gives into the structure of mechanics, and for applications involving the use of perturbation theory. The significance is further elaborated after a brief discussion of the principle of least action, Hamilton's equations of motion, and the theory of canonical transformations.

2.3.1 The Principle of Least Action

Assume that the system is conservative and holonomic so that H is a constant. The Maupertuis principle of least action follows by considering a restricted class of variations Δ_t that are identical to the Δ variations of Sec. 2.3, except that $\Delta_t q_i = \delta q_i(t + \delta t)$ is assumed to vanish at the end points; i.e., the system arrives at the same end point, but at the varied time $t + \delta t$, and H has the same constant value on the varied path. The varied path could be the same as the actual path except that it is traversed at a different rate. Under these conditions, Eq. (52) reduces immediately to

$$
\Delta_t S = -H(\delta t_2 - \delta t)_1. \qquad (56)
$$

However, with the use of Eq. (50) to replace L by H, a direct evaluation of $\Delta_t S$ from Eq. (51) yields

$$\Delta_t S = \Delta_t \int_{t_2}^{t_2} \sum_i p_i \dot{q}_i dt - H(\delta t_2 - \delta t_1), \qquad (57)$$

from which it follows that

$$\Delta_t \int_{t_1}^{t_2} \sum_i p_i \dot{q}_i dt = 0. \qquad (58)$$

The quantity

$$W = \int_{t_1}^{t_2} \sum_i p_i \dot{q}_i dt \qquad (59)$$

defines the classical action.

This is one of many possible ways of expressing the principle of least action. A purely geometrical form in which t is eliminated can be obtained as follows. If V is velocity-independent and T is a homogeneous quadratic function of the $\dot{q}_i$, then

$$\sum_i p_i \dot{q}_i = 2T = \sum_{j,k} M_{j,k}(q) \dot{q}_j \dot{q}_k, \qquad (60)$$

where the $M_{j,k}(q)$ are the coefficients appearing in the kinetic-energy expression. The $M_{j,k}(q)$ can be regarded as the elements of a metric tensor in a curvilinear coordinate space such that the element of path length is

$$(d\rho)^2 = \sum_{j,k} M_{j,k}(q) dq_j dq_k. \qquad (61)$$

The principle of least action can then be written in the form

$$\Delta_t \int_{t_1}^{t_2} T dt = 0 = \Delta \int_{\rho_1}^{\rho_2} \sqrt{T}\, d\rho, \qquad (62)$$

or equivalently,

$$\Delta \int_{\rho_1}^{\rho_2} \sqrt{H - V}\, d\rho = 0. \qquad (63)$$

For a single particle moving in a potential V, $d\rho$ is simply the element of arc length ds along the trajectory. In this form, the principle of least action is formally identical to Fermat's principle of geometrical optics. One need merely identify $\sqrt{H - V}$ with a variable index of refraction $n(s) = c_{\text{vac}}/c(s)$, which is inversely proportional to the velocity of light in the medium. The path taken by a beam of light is then such that the travel time given by

$$\tau = \frac{1}{c_{\text{vac}}} \int_{s_1}^{s_2} n(s) ds \qquad (64)$$

is a minimum (or, more strictly, an extremum). This justifies the solution to the brachistochrone problem in Sec. 1, and it demonstrates the formal equivalence between geometrical optics and the dynamics of conservative systems.

In a recent article, Gray *et al.* (1996) show that the Maupertuis principle of least action can be generalized to a form in which W is stationary with respect to varied paths on which H is held constant only *in the mean* and that there exists a reciprocal principle in which the mean value $\bar{H}$ is made stationary with respect to varied paths of constant W. They also show that the reciprocal principle can be derived directly from the classical limit of the Schrödinger variational principle of quantum mechanics (see Sec. 3.2). A similar reciprocity theorem for Hamilton's principle provides a set of four variational principles analogous to the four equilibrium principles of thermodynamics.

2.3.2 Hamilton's Equations of Motion

The basic approach in Lagrangian mechanics is to regard the N generalized coordinates $q_i(t)$ as the independent variables whose time dependence is determined by the N second-order Lagrangian equations of motion expressed by Eq. (39). The velocities $\dot{q}_i(t)$ enter only as derived quantities whose initial values, together with the initial $q_i(t)$, determine the required $2N$ constants of integration. The Hamiltonian approach differs in that the $\dot{q}_i(t)$ are eliminated in favor of the canonical momenta $p_i(t)$ defined by Eq. (49). The $p_i(t)$ are then elevated to an equal footing with the $q_i(t)$ so that the set $\{q_i, p_i | i = 1, \ldots, N\}$ forms a set of $2N$ independent variables satisfying a set of $2N$ coupled first-order differential equations called Hamilton's equations of motion. In what follows, we adopt the convention that a summation over repeated subscripts is implied, and q or p without a subscript stands for the entire set.

Hamilton's equations of motion can be

derived from Hamilton's principle if Eqs. (49) and (50) are first used to rewrite Eq. (51) in the form

$$S = \int [p_i\dot{q}_i - H(q,p,t)]dt. \tag{65}$$

With *fixed* end points t_1 and t_2, independent variations δp_i and δq_i then induce the variation

$$\begin{aligned}\delta S &= \int_{t_1}^{t_2} \left(p_i\delta\dot{q}_i + \dot{q}_i\delta p_i) + \frac{\partial H}{\partial q_i}\delta q_i + \frac{\partial H}{\partial p_i}\delta p_i\right)dt \\ &= p_i\delta q_i\Big|_{t_1}^{t_2} + \int_{t_1}^{t_2}\left(-\dot{p}_i\delta q_i + \dot{q}_i\delta p_i\right. \\ &\quad \left. + \frac{\partial H}{\partial q_i}\delta q_i + \frac{\partial H}{\partial p_i}\delta p_i\right)dt. \end{aligned} \tag{66}$$

The integrated term vanishes by assumption. Equating to zero the coefficients of δp_i and δq_i then yields the set of equations

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q_i}, \tag{67}$$

which are Hamilton's equations of motion. Thus, both Lagrangian and Hamiltonian mechanics are implied by Hamilton's principle, provided that the variations δp_i and δq_i can be regarded as independent. The latter assumption is important because it underlies the entire approach of Hamiltonian mechanics. It is justified by the fact that the Hamiltonian form can be obtained by a direct transformation of variables from the Lagrangian form.

2.3.3 Canonical Transformations

Transformations of the type (33) are called *point transformations* because they involve only the coordinates q_i. An advantage of Hamiltonian mechanics is that, with the q's and p's regarded as independent variables, more general types of transformation can be constructed that map the set $\{q,p\}$ to a new set $\{Q,P\}$ defined by connection equations of the form

$$Q_i = Q_i(q,p,t), \quad P_i = P_i(q,p,t). \tag{68}$$

Such a *contact transformation* is said to be *canonical* if the form of Hamilton's equations is left invariant; i.e., there exists a transformed Hamiltonian function $K(Q,P,t)$ such that

$$\dot{Q}_i = \frac{\partial K}{\partial P_i}, \quad \dot{P}_i = -\frac{\partial K}{\partial Q_i} \tag{69}$$

are the correct equations of motion for the transformed variables. This will be true if both sets of coordinates and momenta satisfy their respective variational principles

$$\delta \int [p_i\dot{q}_i - H(q,p,t)]dt = 0, \tag{70}$$

$$\delta \int [P_i\dot{Q}_i - K(Q,P,t)]dt = 0. \tag{71}$$

The integrands need not be equal, but they can differ by at most the total time derivative of a function F, called the *generating function* for the transformation. We can thus write

$$p_i\dot{q}_i - H(q,p,t) = \lambda\left[P_i\dot{Q}_i - K(Q,P,t) + \frac{dF}{dt}\right]. \tag{72}$$

Values of the constant $\lambda \neq 1$ just correspond to a scale transformation, and so without loss of generality, we can take $\lambda = 1$. With the choice $F = F_1(q,Q,t)$, the set $\{q,Q\}$ is regarded as the $2N$ separately independent variables, and the quantity dF/dt in Eq. (72) can be replaced by

$$\frac{dF_1}{dt} = \frac{\partial F_1}{\partial q_i}\dot{q}_i + \frac{\partial F_1}{\partial Q_i}\dot{Q}_i + \frac{\partial F_1}{\partial t}. \tag{73}$$

Equating to zero the coefficients of $\dot{q}_i$ and $\dot{Q}_i$ then yields

$$p_i = \frac{\partial F_1}{\partial q_i}, \quad P_i = -\frac{\partial F_1}{\partial Q_i}, \tag{74}$$

$$K = H + \frac{\partial F_1}{\partial t}. \tag{75}$$

For example, the choice $F_1 = q_iQ_i$ interchanges the roles of the q's and p's (except for a sign change), with the result

$$p_i = Q_i, \quad P_i = -q_i, \quad K = H. \tag{76}$$

This demonstrates explicitly that the q's and

p's stand on an equal footing as independent variables in Hamiltonian mechanics—their roles can be interchanged by a canonical transformation.

Other choices for the $2N$ independent variables can be achieved by application of one or more Legendre transformations (as is done in thermodynamics to change independent variables). Of particular relevance to the Hamilton–Jacobi equation is the choice $\{q,P\}$. One need merely replace $F_1(q,Q)$ by $F_2(q,P) = F_1 + Q_iP_i$. Then Eq. (72) becomes

$$p_i\dot{q}_i - H(q,p,t) = -Q_i\dot{P}_i - K(Q,P,t) + \frac{dF_2}{dt}, \tag{77}$$

with

$$\frac{dF_2}{dt} = \frac{\partial F_2}{\partial q_i}\dot{q}_i + \frac{\partial F_2}{\partial P_i}\dot{P}_i + \frac{\partial F_2}{\partial t}, \tag{78}$$

and so

$$P_i = \frac{\partial F_2}{\partial q_i}, \quad Q_i = \frac{\partial F_2}{\partial P_i}, \tag{79}$$

$$K = H + \frac{\partial F_2}{\partial t}. \tag{80}$$

For example, the choice $F_2 = q_iP_i$ generates the identity transformation with $Q_i = q_i$ and $P_i = p_i$.

2.3.4 Interpretation of the Hamilton–Jacobi Equation A comparison of Eq. (80) with Eq. (54) shows that the Hamilton–Jacobi equation can be regarded as a canonical transformation with $F_2 = S(q,P,t)$ such that the transformed Hamiltonian is $K = 0$. The transformed equations of motion are then

$$\frac{\partial K}{\partial P_i} = \dot{Q}_i = 0, \quad \frac{\partial K}{\partial Q_i} = -\dot{P}_i = 0. \tag{81}$$

Thus the transformed momenta $P_i = \alpha_i$ and the transformed coordinates $Q_i = \beta_i$ are all constants of the motion. The nature of the solution is now clear. Writing $S = S(q,\alpha,t)$, then the set of equations

$$p_i = \frac{\partial S(q,\alpha,t)}{\partial q_i}, \quad \beta_i = \frac{\partial S(q,\alpha,t)}{\partial \alpha_i} \tag{82}$$

evaluated at $t = t_1$ provides a set of $2N$ equations to determine $2N$ unknowns α_i and β_i in terms of the initial $(q_i)_1$ and $(p_i)_1$. For conservative systems, the remaining $(N + 1)$th constant of integration is the energy E of the system. For this case, the time variable is separable, and S can be written in the form

$$S(q,\alpha,t) = W(q,\alpha) - Et, \tag{83}$$

where $W(q,\alpha)$ is called *Hamilton's characteristic function.* The crucial point is that *any* complete integral to the Hamilton–Jacobi equation generates a possible dynamical motion of the system since it is the generating function for a canonical transformation. This result is called Jacobi's theorem.

As a consequence of Jacobi's theorem, any complete integral S contains within it all possible trajectories of the system as a function of the initial conditions, rather than one particular trajectory. In fact, surfaces of constant S move through configuration space like wave fronts of constant phase such that the particle trajectories follow the orthogonal set of curves. To see this, consider the example of a particle moving in a potential V, expressed in Cartesian coordinates. Equations (53) for $\{q_1,q_2,q_3\} = \{x,y,z\}$ can then be written as the single equation

$$\nabla S = \nabla W = \mathbf{p}, \tag{84}$$

which shows that the momentum $\mathbf{p} = m\mathbf{v}$ is everywhere perpendicular to surfaces of constant W, and the particle velocity is $v = |\nabla W|/m$. The Hamilton–Jacobi equation for W in this case is

$$(1/2m)(\nabla W)^2 + V = E, \tag{85}$$

so that

$$|\nabla W| = \sqrt{2m(E - V)}. \tag{86}$$

For the case of a freely falling particle, surfaces of constant W are just horizontal planes with the particle trajectories in the perpendicular direction. As time goes on, surfaces of constant S sweep through surfaces of constant W with a phase velocity given by

$$u = \frac{ds}{dt} = \frac{E}{|\nabla W|}, \tag{87}$$

where ds is a displacement in the direction normal to a surface of constant W. The above follows from the facts that the stationary phase condition $dS = 0$ corresponds to $dW = E\,dt$, and $dW = |\nabla W| ds$ is the change in W due to a displacement in the direction normal to the surface. Thus the phase velocity decreases as the particle velocity increases, just as is the case for the wave and particle pictures of light.

The above considerations in fact provide a wave picture of classical dynamics in the "geometrical optics" limit where the wavelength is infinitesimally small compared with the dimensions of the apparatus. The wave nature can then be ignored, and the trajectories determined by the principle of least action (or Fermat's principle in the case of geometrical optics).

2.4 Relativistic Generalization

For the case of a single particle acted on by forces derivable from a potential V, Hamilton's principle can be simply modified to incorporate the effects of special relativity. One need simply define

$$L = -mc^2\gamma - V, \tag{88}$$

where $\gamma = \sqrt{1 - v^2/c^2}$, and v is the velocity $|\dot{\mathbf{r}}|$ in a particularly chosen Lorentz frame. With this choice of L, Hamilton's principle and the Euler–Lagrange equations give the correct equations of motion

$$\frac{d}{dt}\left(\frac{m\dot{x}_i}{\gamma}\right) = -\frac{\partial V}{\partial x_i}. \tag{89}$$

The canonical momenta p_i and Hamiltonian H are given by

$$p_i = \frac{\partial L}{\partial \dot{x}_i} = \frac{m\dot{x}_i}{\gamma}, \tag{90}$$

$$H = p_i\dot{x}_i - L = T + V = E, \tag{91}$$

where $T = mc^2/\gamma$ is a relativistic generalization of the kinetic energy, including the rest-mass energy mc^2. After substituting for $\dot{x}_i$, H assumes the form

$$H = \sqrt{c^2p^2 + m^2c^4} + V. \tag{92}$$

2.4.1 Inclusion of Electromagnetic Fields In general, an electromagnetic field is derivable from a scalar potential $\phi(\mathbf{r},t)$ and a vector potential $\mathbf{A}(\mathbf{r},t)$ according to

$$\mathbf{E} = -\nabla\phi - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t}, \tag{93}$$

$$\mathbf{B} = \nabla \times \mathbf{A}, \tag{94}$$

where $\mathbf{E}$ and $\mathbf{B}$ are the electric and magnetic fields. The equation of motion for a particle of charge q is then

$$\frac{d}{dt}\left(\frac{m\dot{x}_i}{\gamma}\right) = qE_i + \frac{q}{c}(\mathbf{v} \times \mathbf{B})_i, \tag{95}$$

which now contains a velocity-dependent force term. This equation follows from the Lagrangian

$$L = -mc^2\gamma - q\phi + \frac{q}{c}\mathbf{A}\cdot\mathbf{v}, \tag{96}$$

or its nonrelativistic counterpart with the term $-mc^2\gamma$ replaced by T. The canonical momenta are then

$$p_i = \frac{\partial L}{\partial \dot{x}_i} = \frac{m\dot{x}_i}{\gamma} + \frac{q}{c}A_i. \tag{97}$$

A direct calculation shows that the Hamiltonian becomes [cf. Eq. (92)]

$$H = \sqrt{c^2[\mathbf{p} - (q/c)\mathbf{A}]^2 + m^2c^4} + V. \tag{98}$$

The same substitution $\mathbf{p} \to \mathbf{p} - (q/c)\mathbf{A}$ applies also in the nonrelativistic case. This simple prescription, together with $V = q\phi$, allows electromagnetic fields to be easily incorporated into the Lagrangian and Hamiltonian formulations of mechanics.

3. APPLICATIONS TO QUANTUM MECHANICS

3.1 Variational Derivation of the Schrödinger Equation

The considerations of Sec. 2.3.4 suggest that the Hamilton–Jacobi equation of classical dynamics expresses the short-wavelength limit of an underlying wave equation, with surfaces of constant S identified as surfaces of constant phase. In fact, Eq. (85) already bears a superficial resemblance to the time-independent Schrödinger equation, but it does not yet have the form of a wave equation. Following Schrödinger, a suitable wave equation can be obtained by first making the substitution

$$W = iC \ln \Psi \Rightarrow \Psi = e^{iW/C} \tag{99}$$

into the Hamilton–Jacobi equation

$$(1/2m)(\nabla W^* \cdot \nabla W) + V = E, \tag{100}$$

generalized for complex W, to obtain

$$(C^2/2m)(\nabla \Psi^* \cdot \nabla \Psi) + (V - E)\Psi^* \Psi = 0. \tag{101}$$

The left-hand side can be integrated over all space, provided that $\int \Psi^* \Psi d^3r$ remains finite. Application of the Euler–Lagrange equations to make the integral stationary with respect to arbitrary independent variations of Ψ and Ψ^* then yields the Schrödinger equation

$$-(C^2/2m)(\nabla^2 \Psi) + (V - E)\Psi = 0, \tag{102}$$

together with a similar equation for Ψ^*. The derivation is a simple extension of the one used to obtain Laplace's equation (31) in Sec. 1.2. The solution to Eq. (102) then determines the wave function $\Psi(\mathbf{r})$ for the system, subject to the constraint that $\int \Psi^* \Psi d^3r$ remain finite for bound systems; i.e., that Ψ is *normalizable.* Comparison with experiment shows that one should set $C = \hbar = h/2\pi$, where h is Planck's constant. Equation (102) can then be written in the form

$$H(q,p)\Psi = E\Psi, \tag{103}$$

where $H(q,p)$ is the Hamiltonian with the quantum-mechanical replacement $\mathbf{p} \rightarrow (\hbar/i)\nabla$. The constraint on $\int \Psi^* \Psi d^3r$ makes this an eigenvalue problem that determines the possible energies E of the system.

If there are n interacting particles, then the term $\nabla^2 \Psi$ is to be replaced by $\Sigma_i \nabla_i^2 \Psi$, and V includes all the interaction potentials. Also, the various integrals over $d\mathbf{r}$ are replaced by multiple integrals over $d^3r_1 d^3r_2 \cdots d^3r_n$.

3.2 The Rayleigh–Schrödinger Variational Principle

Consider a bound system, or one that is contained in a finite box. Under these conditions, Eq. (101) can be integrated over all space and the term $\nabla \Psi^* \cdot \nabla \Psi$ integrated by parts to obtain

$$\int \Psi^*(H - E)\Psi d^3r = 0. \tag{104}$$

The integrated term does not contribute under the assumed conditions because $\Psi(\mathbf{r}) \rightarrow 0$ sufficiently rapidly as $|\mathbf{r}| \rightarrow \infty$. The variational derivation of Sec. 3.1 guarantees that this integral is stationary with respect to arbitrary variations $\delta\Psi$ if Ψ satisfies the Schrödinger equation. However, the same variational condition can now be reinterpreted as the problem of making the integral $\int \Psi^* H \Psi d^3r$ stationary, subject to the constraint that

$$\int \Psi^* \Psi d^3r = \text{const.}, \tag{105}$$

with E playing the role of a Lagrange undetermined multiplier. In this guise, one can say that E obtained from the Rayleigh quotient

$$E = \frac{\int \Psi^* H \Psi d^3r}{\int \Psi^* \Psi d^3r} \tag{106}$$

is stationary. In fact, as discussed in the following section, E is a minimum under many circumstances.

3.3 The Rayleigh–Ritz Variational Method

The Schrödinger equation is a partial-differential equation that can be solved exactly only for certain special cases such as the Coulomb potential or the harmonic-oscillator potential. For arbitrary potentials, or for problems containing more than two bodies, the quantum-mechanical problem is no easier to solve than the corresponding classical one. In these cases, the Rayleigh–Schrödinger variational principle provides one of the most powerful methods for obtaining approximate eigenvalues E and wave functions Ψ.

Suppose one guesses by some means an approximate trial wave function $\Psi_{\rm tr}$ that conforms with the constraint of normalizability and approximates one of the exact solutions to

$$H\Psi_i = E_i\Psi_i, \quad i = 1, 2, \ldots. \tag{107}$$

The index i labels the spectrum of exact solutions. In general, the eigenvalue spectrum will have both discrete and continuous parts. In the latter case, summations over i include integrations over the continuous part. The crucial point is that even though the Ψ_i are not known, they form a complete basis set of functions in terms of which the trial function $\Psi_{\rm tr}$ can be expanded. In analogy with Fourier series, one can therefore write

$$\Psi_{\rm tr} = \sum_{i=1}^{\infty} c_i\Psi_i, \tag{108}$$

where the c_i are the expansion coefficients. Let the eigenvalue spectrum be ordered so that $E_1 < E_2 < E_3 < \cdots$, and assume that all the Ψ_i and $\Psi_{\rm tr}$ are normalized to unity; i.e., using Dirac bra-ket notation for integrals,

$$\begin{aligned}&\langle\Psi_{\rm tr}|H|\Psi_{\rm tr}\rangle \equiv \int \Psi^* H\Psi d^3r = E_{\rm tr},\\ &\langle\Psi_i|\Psi_j\rangle = \delta_{i,j},\\ &\langle\Psi_i|H|\Psi_j\rangle = E_i\delta_{i,j}.\end{aligned} \tag{109}$$

Substituting Eq. (108) into (106), and using Eqs. (109), one then obtains

$$E_{\rm tr} = |c_1|^2E_1 + |c_2|^2E_2 + |c_3|^2E_3 + \cdots \tag{110}$$

for the corresponding trial energy. Since, by assumption, $\langle\Psi_{\rm tr}|\Psi_{\rm tr}\rangle = 1$, it follows that

$$\sum_{i=1}^{\infty} |c_i|^2 = 1, \tag{111}$$

and so Eq. (110) can be rewritten in the form

$$\begin{aligned}E_{\rm tr} &= E_1 + |c_2|^2(E_2 - E_1)\\ &\quad + |c_3|^2(E_3 - E_1) + \cdots\\ &\geq E_1.\end{aligned} \tag{112}$$

Thus $E_{\rm tr}$ is an *upper bound* on the lowest eigenvalue E_1 for any normalizable $\Psi_{\rm tr}$.

The basic idea of variational calculations then is to write $\Psi_{\rm tr}$ in some arbitrarily chosen mathematical form with variational parameters (subject to normalizability and boundary conditions at the origin and infinity), and then adjust the parameters to obtain the minimum value of $E_{\rm tr}$. A lower $E_{\rm tr}$ is guaranteed to be closer to E_1. The power of the method stems both from this and the fact that, by the Rayleigh–Schrödinger variational principle, the error term linear in $\delta\Psi = \Psi_{\rm tr} - \Psi_1$ vanishes.

3.3.1 Algebraic Solution for Linear Variational Parameters

Suppose that $\Psi_{\rm tr}$ depends in some arbitrarily chosen way on a set of N variational parameters $a_1, a_2, \ldots, a_N$. [For example, in a one-dimensional case, one might choose $\Psi(r) = r^{a_1}e^{-a_2r}$ with a_1 and a_2 regarded as nonlinear variational parameters.] Then the variational condition corresponds to the system of equations

$$\frac{\partial E_{\rm tr}}{\partial a_p} = 0, \quad p = 1, \ldots, N. \tag{113}$$

In general, this is a set of transcendental algebraic equations that cannot be solved exactly.

However, the minimization problem for the case of *linear* variational coefficients can be solved exactly by matrix diagonalization. For example, let $\{\chi_p|p = 1, \ldots, N\}$ be a finite basis set of N arbitrarily chosen functions (subject to the boundary conditions and normalizability) that need have nothing to do with the exact Ψ_i, and write $\Psi_{\rm tr}$ in the form

$$\Psi_{\rm tr} = \sum_{p=1}^{N} a_p\chi_p. \tag{114}$$

Now the variational parameters a_p enter linearly, and the set of variational conditions (113) becomes exactly equivalent to the N-dimensional generalized eigenvalue problem

$$\mathbf{Ha} = \lambda\mathbf{Oa}, \tag{115}$$

where $\mathbf{a}$ is a column vector of coefficients a_p, and $\mathbf{H}$ and $\mathbf{O}$ have matrix elements $H_{pq} = \langle\chi_p|H|\chi_q\rangle$ and $O_{pq} = \langle\chi_p|\chi_q\rangle$. There are N eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$, of which the lowest is an upper bound to E_1.

Equation (115) is equivalent to the original Schrödinger equation (103) only if the basis set $\{\chi_p\}$ is complete, and in general this requires taking the limit $N \to \infty$. The significance of Eq. (115) is that it provides a computationally useful means of obtaining approximate solutions, even if the complete basis set of functions $\{\chi_p\}$ is truncated at some finite number N, and the lowest eigenvalue provides an upper bound that systematically decreases toward the exact E_1 as N is increased. As discussed in the following section, the bounds apply not just to the ground state but also to the lower-lying excited states.

3.3.2 Extension to Excited States By the Hylleraas–Undheim–MacDonald (HUM) theorem (see Hylleraas and Undheim, 1930; MacDonald, 1933), the remaining eigenvalues $\lambda_2, \lambda_3, \ldots$ are also upper bounds to the exact energies $E_2, E_3, \ldots$, provided that the spectrum is bounded from below. The HUM theorem is a consequence of the matrix eigenvalue interleaving theorem, which states that as the dimensions of **H** and **O** are progressively increased by adding an extra row and column, the N old eigenvalues λ_p fall between the $N + 1$ new ones. Consequently, as illustrated in Fig. 3, all eigenvalues numbered from the bottom up must move inexorably downward as N is increased. Since the exact spectrum of bound states is obtained in the limit $N \to \infty$, no λ_p can cross the corresponding exact E_p on its way down. Thus $\lambda_p \geq E_p$ for every finite N.

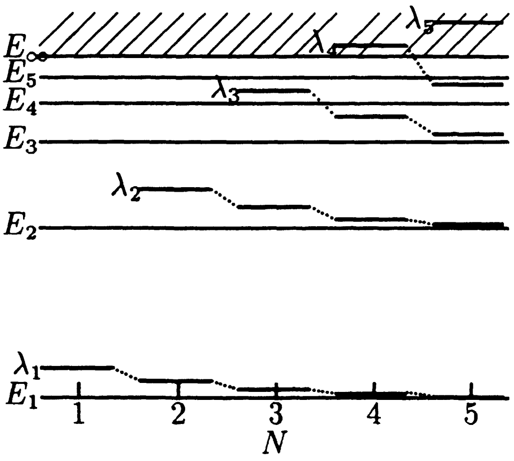

FIG. 3. Diagram illustrating the HUM theorem. The λ_p, $p = 1, \ldots, N$, are the variational eigenvalues for an N-dimensional basis set, and the E_i are the exact eigenvalues of H. The highest λ_p lie in the continuous spectrum of H.

If the exact Ψ_i can be formed from a linear combination of the χ_p included in the finite basis set, then the result of the variational calculation is the exact Ψ_i and E_i. Otherwise, one obtains the best variational approximation provided by the particular χ_p chosen. If the χ_p basis set becomes asymptotically complete as $p \to \infty$, then convergence to the correct answer is assured.

The HUM theorem no longer applies directly to the relativistic Dirac equation or similar problems because the spectrum is not bounded from below. However, finite basis-set methods can still be applied, provided that sufficient care is taken in their construction (see Drake and Goldman, 1988; Grant, 1996).

3.3.3 Variational Lower Bound If the Rayleigh–Ritz method is applied to the integral

$$\int \Psi^*(H - E)(H - E_2^<)d^3r, \tag{116}$$

then the quantity

$$E^< = \frac{\langle \Psi_{\rm tr}|H(H - E_2^<)|\Psi_{\rm tr}\rangle}{\langle \Psi_{\rm tr}|(H - E_2^<)|\Psi_{\rm tr}\rangle} \tag{117}$$

is made stationary. If the quantity $E_2^<$ is chosen to be a lower bound on the energy E_2 of the first excited state, then by an argument similar to that leading to Eq. (112), $E^<$ is a lower bound on the ground-state energy E_1, called the Temple lower bound (Temple, 1928) and denoted by $E_1^<$. In fact, if $\Psi_{\rm tr}$ is expanded as in Eq. (108), then after some algebra, $E^<$ from Eq. (117) becomes

$$E_1^< = E_1 + \frac{\sum_{i=2}^{\infty} |c_i|^2(E_i - E_1)(E_i - E_2^<)}{E_1 - E_2^< + D}, \tag{118}$$

where

$$D = \sum_{i=2}^{\infty} |c_i|^2(E_i - E_1) = E_1^> - E_1, \tag{119}$$

and $E_1^>$ is the variational upper bound on E_1. The numerator of the fraction in Eq. (118) is positive and the denominator is negative, provided that $E_1^> < E_2^<$, thus proving the bound. However, generally speaking, $E_1^<$ is much less accurate than $E_1^>$.

3.3.4 Illustrative Results for Helium

Application of the variational method to helium by Hylleraas (1928, 1929) played an important role in the early history of quantum mechanics because it provided the first test of the Schrödinger equation in a system more complicated than hydrogen. With its two electrons orbiting the nucleus, helium is the simplest atomic system that cannot be adequately described by the older Bohr–Sommerfeld quantum theory.

The Hamiltonian for helium (in the limit of infinite nuclear mass) is

$$H = -\sum_{i=1}^{2}\left(\frac{\hbar^2}{2m_e}\nabla_i^2 + \frac{e^2}{r_i}\right) + \frac{e^2}{r_{12}}, \tag{120}$$

where $r_{12} = |\mathbf{r}_1 - \mathbf{r}_2|$ is the electron–electron separation and Ze is the nuclear charge. The e^2/r_{12} term represents the Coulomb repulsion between the two electrons. Without this term, the Schrödinger equation would be separable, and the exact solution (including permutational symmetry) would be of the form

$$\Psi(\mathbf{r}_1,\mathbf{r}_2) = \psi_1(\mathbf{r}_1)\psi_2(\mathbf{r}_2) \pm \psi_1(\mathbf{r}_2)\psi_2(\mathbf{r}_1), \tag{121}$$

where $\psi_1(\mathbf{r})$ and $\psi_2(\mathbf{r})$ are exactly known hydrogenic wave functions, depending on the state in question. The Hartree–Fock approximation corresponds to the best variational representation that can be written in the form of a separable product with $\psi_1(\mathbf{r})$ and $\psi_2(\mathbf{r})$ regarded as arbitrary functions of r. However, even this is in error for the ground-state energy of $-2.903\,724$ a.u. [the atomic unit (a.u.) of energy is $e^2/a_0 = 27.211\,396$ eV, where a_0 is the Bohr radius] by 0.0247 a.u., or 0.673 eV. This difference, called the *correlation energy*, is much larger than typical chemical energies.

To obtain a better representation, Hylleraas suggested constructing a trial solution of the form

$$\Psi_{\text{tr}} = \sum_{i,j,k} a_{ijk} r_1^i r_2^j r_{12}^k \exp(-\alpha r_1 - \beta r_2), \tag{122}$$

which is of the form of a hydrogenic product $\psi_1(\mathbf{r}_1)\psi_2(\mathbf{r}_2)$, except that it contains explicit powers of r_{12} and is therefore not separable. The a_{ijk} are the linear variational parameters, and α and β are additional nonlinear parameters that can be separately varied to optimize the energy. Detailed formulas for the necessary matrix elements are given by Drake (1996). This basis set is provably complete in the limit of large $i_{\max}$, $j_{\max}$, and $k_{\max}$ (Klahn and Bingel, 1977). Typically, all combinations of powers are included in the basis set such that, for electrons with angular momenta l_1 and l_2,

$$i + j + k - l_1 - l_2 \le \Omega, \tag{123}$$

where Ω is an integer that is progressively increased until adequate convergence is obtained. Without further truncation, the number of terms obtained in this way is

$$N = \tfrac{1}{6}(\Omega + 1)(\Omega + 2)(\Omega + 3). \tag{124}$$

The effect of including powers of r_{12} is dramatic and immediate. The Hartree–Fock approximation corresponds to the limit of large $i_{\max}$ and $j_{\max}$ with $k_{\max} = 0$. As shown in Table 1, an increase of $k_{\max}$ to 1 reduces the error in E to only 0.006 20 eV, thereby accounting for 99% of the correlation energy. The results in Table 1 also demonstrate that the odd powers of r_{12} are much more effective than the even powers. This can be understood from the fact that r_{12}^2 can be written in the form

$$r_{12}^2 = r_1^2 + r_2^2 - 2r_1r_2\cos\theta_{12}, \tag{125}$$

where θ_{12} is the angle between the vectors $\mathbf{r}_1$ and $\mathbf{r}_2$. Thus r_{12}^2 is only linear in $\cos\theta_{12}$, while an expansion of $r_{12} = (r_{12}^2)^{1/2}$ contains all powers of $\cos\theta_{12}$.

Table 1. Energies for the ground state of helium obtained with various powers of r_{12} in the basis set.

r_{12} Terms	Energy (a.u.)	Error (eV)
No r_{12}	−2.879 029	0.672
r_{12}^2	−2.900 503	0.087 6
r_{12}^2, r_{12}^4	−2.902 752	0.026 4
r_{12}	−2.903 496	0.006 20
r_{12}, r_{12}^3	−2.903 700	0.000 65
All r_{12}	−2.903 724	0.000 00

Calculations of this type have been performed by many authors (see Drake, 1993a, for a review), and they have reached a high degree of sophistication. Problems typical of all variational calculations are a dramatic and progressive loss of accuracy for the more highly excited states, and numerical linear dependence in the basis set as it is enlarged. These problems can be avoided by doubling the basis set so that it contains a second set of terms with the same powers but different scale factors α and β. A complete optimization with respect to the α's and β's then leads to a natural partition of the basis set into two sectors with one representing the asymptotic form of the wave function and the other representing complex inner correlation effects. The results obtained by this method are essentially exact for all practical purposes. The nonrelativistic energies are known to better than one part in 10^{16} for the entire singly excited spectrum. A sample of results for the low-lying states is given in Table 2. The indicated convergence was obtained by progressively increasing Ω up to 17, corresponding to about 1700 terms in the doubled basis set. Table 3 shows an example of the convergence for the ground state. [For the case of S states, the basis-set sizes are smaller than indicated by Eq. (124) because terms with $i > j$ can be omitted by symmetry.] The ratios of successive differences in the last column provide a convenient method to monitor the convergence of the eigenvalue. They show that the differences themselves decrease in a fairly smooth and uniform fashion with increasing Ω. These high-precision results provide a benchmark for comparison with other less accurate methods of calculation such as Hartree–Fock and configuration interaction. Results for many other states are given by Drake (1993b, 1994).

Table 2. Nonrelativistic energies for several states of helium in the limit of infinite nuclear mass.

State	Energy (a.u.)
$1s^2\ ^1S$	−2.903 724 377 034 119 60(2)
$1s2s\ ^1S$	−2.145 974 046 054 419(2)
$1s2s\ ^3S$	−2.175 229 378 236 791 307(6)
$1s2p\ ^1P$	−2.123 843 086 498 101 35(5)
$1s2p\ ^3P$	−2.133 164 190 779 283 17(3)
$1s3s\ ^1S$	−2.061 271 989 740 911(5)
$1s3s\ ^3S$	−2.068 689 067 472 457 192(1)
$1s3p\ ^1P$	−2.055 146 362 091 943 33(7)
$1s3p\ ^3P$	−2.058 081 084 274 275 3(2)
$1s3d\ ^1D$	−2.055 620 732 852 246 51(8)
$1s3d\ ^3D$	−2.055 636 309 453 261 34(4)

Table 3. Convergence study for the ground-state energy of helium (in atomic units). The numbers in the last column give the ratios of successive differences.

Ω	$N_{\text{tot}}(\Omega)$	$E(\Omega)$	$R(\Omega)^a$
4	44	−2.903 724 131 001 531 810	
5	67	−2.903 724 351 566 477 006	
6	98	−2.903 724 373 891 109 909	9.88
7	135	−2.903 724 376 548 959 510	8.40
8	182	−2.903 724 376 960 412 587	6.46
9	236	−2.903 724 377 018 168 462	7.12
10	302	−2.903 724 377 030 786 217	4.58
11	376	−2.903 724 377 033 426 037	4.78
12	464	−2.903 724 377 033 966 492	4.88
13	561	−2.903 724 377 034 076 500	4.91
14	674	−2.903 724 377 034 107 875	3.51
15	797	−2.903 724 377 034 116 019	3.85
16	938	−2.903 724 377 034 118 518	3.26
17	1090	−2.903 724 377 034 119 239	3.47
18	1262	−2.903 724 377 034 119 479	3.01
Extrapolation		−2.903 724 377 034 119 597(15)	

[a] $R(\Omega) = [E(\Omega - 1) - E(\Omega - 2)]/[E(\Omega) - E(\Omega - 1)]$.

A comparison of the results in Table 2 with experiment is meaningful only after corrections for finite nuclear mass, special relativity, and quantum-electrodynamic (QED) effects (such as electron self-energy and vacuum polarization) are taken into account. A detailed discussion of these corrections can be found in Drake (1993b, 1994). When they are included, the calculated transition frequencies agree to within the estimated accuracy of the QED shift. If the measurements are expressed in terms of ionization energies for the various states, then their accuracies range from ±30 MHz ($\pm 5 \times 10^{-9}$ a.u.) for the ground state to ±0.1 MHz ($\pm 1.5 \times 10^{-11}$ a.u.) for the higher-lying P and D states. Since the nonrelativistic energies in Table 2 are much more accurate than this, the comparison with experiment is primarily a test of higher-order contributions to the QED shift (two-electron Lamb shift), which is the dominant source of uncertainty in the calculations.

As one example, the calculated ionization energy of the $1s2s\ ^1S$ state is (Drake *et al.*, 1993)

960 332 039.4 ± 1 MHz.

Of this total, −2808.5 ± 1 MHz comes from the calculated QED shift. For comparison, the two experimental values are

960 332 041.52 ± 0.21 MHz,
960 332 040.87 ± 0.15 MHz.

The first is obtained from an extrapolation of the $1s2s\ ^1S$–$1snp\ ^1P$ transition frequencies to the series limit (Sansonetti and Gillaspy, 1992), and the second from the $1s2s\ ^1S$–$1snd\ ^1D$ two-photon transition frequencies (Lichten *et al.*, 1991). Although the measurements do not quite agree with each other, taken together they determine the QED shift of the $1s2s\ ^1S$ state to an accuracy of about 100 parts per million and verify the calculated value to better than 0.1%. For the ground state, the calculated QED shift in the ionization energy has the much larger value $-(41\,233 \pm 35)$ MHz. This has recently been verified to an accuracy of ±45 MHz (±0.1%) from the total $1s^2\ ^1S$–$1s2p\ ^1P$ transition frequency (Eikema *et al.*, 1996).

In summary, the results in Table 2 provide a firm foundation of nonrelativistic energies upon which higher-order corrections can be built and compared with experiment. Further improvements in the QED part of the theory remain an important challenge for the future.

3.3.5 Extensions to More Complex Systems Fully correlated variational calculations of the type described in the previous section are difficult to extend to systems more complex than helium because of the rapid increase in the number of terms required. For an atom containing K electrons, there are K single-particle radial distances r_s and $K(K-1)/2$ interparticle distances r_{st} for a total of $P = K(K+1)/2$ radial coordinates. If all combinations of powers of the r_s and r_{st} are included in the basis set such that the sum of powers is $\leq \Omega$ [cf. Eq. (123)], then the generalization of Eq. (124) for the number of terms is

$$N = \frac{(\Omega+1)(\Omega+2)\cdots(\Omega+P)}{P!}. \tag{126}$$

Since the time required to calculate a single eigenvector increases in proportion to N^3, the overall complexity of the calculation increases roughly in proportion to

$$\left[\frac{6(\Omega+P)!}{P!(\Omega+3)!}\right]^3 \tag{127}$$

relative to helium with the same Ω.

As an example, from Table 3, an accuracy of 10^{-10} a.u. for the ground state of helium requires $\Omega = 8$. A similar accuracy for lithium with $K = 3$ and $P = 6$ therefore requires about 6000 times the computer resources, and for beryllium with $K = 4$ and $P = 10$, the factor from expression (127) becomes 1.4×10^{13}.

Because of this rapid increase of complexity with the number of electrons, fully correlated calculations of spectroscopic accuracy have only been extended as far as lithium (see Yan and Drake, 1995; Yan *et al.*, 1996; and earlier references therein). The pattern of convergence for the ground state is similar to that shown in Table 3. The results up to $\Omega = 8$ yield the extrapolated nonrelativistic eigenvalue

$$E(1s^2 2s^2\ S) = -7.478\,060\,323\,10(31) \text{ a.u.} \tag{128}$$

The uncertainty of $\pm 3 \times 10^{-10}$ is about what one would expect from Table 3 for $\Omega = 8$.

For systems more complex than lithium, one must resort to other methods of calculation that can be extended to arbitrarily complex systems, but typically having much lower accuracy ($\pm 10^{-6}$ a.u. or more). These methods include multiconfiguration Hartree–Fock (MCHF), configuration-interaction (CI), many-body perturbation-theory, finite-element, diffusion Monte Carlo (DMC), and variational Monte Carlo (VMC) techniques. The MCHF and CI methods are similar in concept to the fully correlated variational method described in Sec. 3.3.4, except that the members of the basis set χ_p are constructed from antisymmetrized products of one-electron orbitals corresponding to definite electronic configurations. The effect is analogous to including only the *even* powers of r_{12} as shown in Table 1, and so convergence with increasing angular momentum of the individual electrons is slow. Recently, Goldman (1994) has devised a modified CI method involving extrapolation procedures to overcome this problem, at least for simple systems. For recent work on finite-element and many-body perturbation-theory methods, see Ackermann (1995) and Plante *et al.* (1994), respectively.

The DMC and VMC Monte Carlo methods attempt to reduce the complexity problem for more complex systems by the use of random-sampling techniques. The DMC method takes advantage of the fact that the time-dependent Schrödinger equation is formally identical to the diffusion equation in imaginary time, and for large imaginary time, an arbitrary starting solution quickly decays to the ground state (see, e.g., Moskowitz *et al.*, 1982; Barnett *et al.*, 1995; and earlier references therein). A random sampling of initial configurations is then propagated forward in time to construct the wave function. The VMC method is more closely related to the standard variational techniques discussed in Sec. 3.3.4. The idea is to define a trial wave function Ψ_{tr} in terms of variational parameters, as in Sec. 3.3.4, and then to optimize them over a statistical distribution of sample points $\mathbf{r}_i$ by minimizing an expression for the variance such as

$$\frac{\sum_i (H\Psi_i - E_{\mathrm{ref}}\Psi_i)^2/w_i}{\sum_i \Psi_i^2/w_i}, \tag{129}$$

or the energy variance given by

$$\frac{\sum_i (H\Psi_i - E_{\mathrm{ref}}\Psi_i)^2\Psi_i^2/w_i^2}{\left[\sum_i \Psi_i^2/w_i\right]^2}. \tag{130}$$

Here, $\Psi_i = \Psi_{\mathrm{tr}}(\mathbf{r}_i)$ is the trial wave function evaluated at some particular set of values for the electronic coordinates collectively denoted by $\mathbf{r}_i$, and the weight function $w_i = w(\mathbf{r}_i)$ is the probability of choosing $\mathbf{r}_i$ if the sampling is nonuniform. The optimum strategy is to bias the sampling according to the value of a guiding function $g(\mathbf{r}_i)$ that resembles the actual Ψ^2 as closely as possible and to choose the reference energy E_{ref} as close as possible to the desired eigenvalue. Although the method could be applied to a direct optimization of $\langle H\rangle$, the advantage gained by optimizing the variance is that the sample space required for a given accuracy is much smaller. Several sample problems and illustrative examples are discussed by Alexander *et al.* (1991).

3.4 Variation–Perturbation Methods

For many problems, it is advantageous to split the Hamiltonian into two parts according to

$$H = H^{(0)} + gV, \tag{131}$$

where the eigenvalue problem for $H^{(0)}$ can be solved exactly (or to high precision), and V is a perturbation whose strength is controlled by the parameter g. If the wave functions and energies are similarly expanded,

$$\Psi = \Psi^{(0)} + g\Psi^{(1)} + g^2\Psi^{(2)} + \cdots, \tag{132}$$

$$E = E^{(0)} + gE^{(1)} + g^2E^{(2)} + \cdots, \tag{133}$$

and substituted into the Rayleigh–Ritz quotient (106), then the terms linear in g give

$$E^{(1)} = \frac{1}{\langle\Psi^{(0)}|\Psi^{(0)}\rangle}\Big[\langle\Psi^{(0)}|V|\Psi^{(0)}\rangle + 2\langle\Psi^{(0)}|H^{(0)} - E^{(0)}|\Psi^{(1)}\rangle\Big]. \tag{134}$$

This is stationary with respect to variations $\delta\Psi^{(0)}$ if $\Psi^{(1)}$ satisfies the first-order perturbation equation

$$(H^{(0)} - E^{(0)})|\Psi^{(1)}\rangle + (V - E^{(1)})|\Psi^{(0)}\rangle = 0. \tag{135}$$

Since by assumption

$$H^{(0)}|\Psi^{(0)}\rangle = E^{(0)}|\Psi^{(0)}\rangle, \tag{136}$$

it follows from Eq. (134) that

$$E^{(1)} = \langle\Psi^{(0)}|V|\Psi^{(0)}\rangle/\langle\Psi^{(0)}|\Psi^{(0)}\rangle. \tag{137}$$

The entire series of perturbation equations to all orders can be similarly generated from the Rayleigh–Ritz variational principle. Computational methods based on these results were first developed by Slater and Kirkwood (1931), and by Dalgarno and Lewis (1955, 1956) (see also Dalgarno and Stewart, 1956; Sternheimer, 1951, 1954, 1957; Schwartz, 1959). They have since been employed by numerous other authors for a wide variety of problems.

3.4.1 Variational Bounds A particular advantage of the variational derivation of the perturbation equations is its use in establishing bounds (see, e.g., Glover and Weinhold, 1976). For example, consider the second-order energy

$$E^{(2)} = \frac{1}{\langle \Psi^{(0)}|\Psi^{(0)}\rangle}\Big[2\langle \Psi^{(0)}|V|\Psi^{(1)}\rangle + 2\langle \Psi^{(0)}|H^{(0)} - E^{(0)}|\Psi^{(2)}\rangle + \langle \Psi^{(1)}|H^{(0)} - E^{(0)}|\Psi^{(1)}\rangle\Big], \qquad (138)$$

from terms quadratic in g. $E^{(2)}$ is stable with respect to variations $\delta\Psi^{(0)}$ if $\Psi^{(2)}$ satisfies the second-order perturbation equation

$$(H^{(0)} - E^{(0)})|\Psi^{(2)}\rangle + (V - E^{(1)})|\Psi^{(1)}\rangle = E^{(2)}|\Psi^{(0)}\rangle, \qquad (139)$$

from which it follows that

$$E^{(2)} = \frac{\langle \Psi^{(0)}|V - E^{(1)}|\Psi^{(1)}\rangle}{\langle \Psi^{(0)}|\Psi^{(0)}\rangle}. \qquad (140)$$

However, $\Psi^{(1)}$ is typically not known exactly and must be approximated in some way. From Eq. (138), $E^{(2)}$ is stable with respect to variations $\delta\Psi^{(1)}$ if $\Psi^{(1)}$ satisfies the first-order equation (135). Since the total $E = E^{(0)} + gE^{(1)} + g^2E^{(2)} + \cdots$ is an upper bound for sufficiently small g, and $E^{(0)}$ and $E^{(1)}$ are known exactly (by assumption), it follows that the value of $E^{(2)}$ calculated from Eq. (138) with some approximate $\Psi_{\mathrm{tr}}^{(1)}$ must be an upper bound.

In particular, if $\Psi_{\mathrm{tr}}^{(1)}$ is expanded in a finite basis set of functions χ_p, as in Eq. (114), then the variational condition $\partial E^{(2)}/\partial b_p = 0$ for the expansion coefficients b_p yields the set of N linear algebraic equations [cf. Eq. (115)]

$$(\mathbf{H}^{(0)} - E^{(0)}\mathbf{O})\mathbf{b} + \mathbf{V}^E = 0, \qquad (141)$$

where $\mathbf{V}^E$ is a column vector with elements $V_p^E = \langle \chi_p|V - E^{(1)}|\Psi^{(0)}\rangle$. [This works correctly if $\langle \chi_p|\Psi^{(0)}\rangle = 0$ by symmetry. However, if $\Psi^{(0)}$ can be expressed as a linear combination of the χ_p, then these equations are singular. In that case, it is sufficient just to delete one of the equations to obtain a nonsingular set.] It then follows that

$$\langle \Psi^{(1)}|(H^{(0)} - E^{(0)})|\Psi^{(1)}\rangle = -\langle \Psi^{(1)}|(V - E^{(1)})|\Psi^{(0)}\rangle \qquad (142)$$

within the finite basis set, and Eq. (138) for $E^{(2)}$ reduces to Eq. (140).

3.4.2 Spectral Representations and Pseudostates Although computationally less efficient, it is instructive and sometimes useful to transform the χ_p basis set to a new basis $\tilde{\chi}_p$ that diagonalizes $\mathbf{H}$. The matrix elements in the transformed basis are then

$$\langle \tilde{\chi}_p|H|\tilde{\chi}_q\rangle = \tilde{E}_p\delta_{p,q}, \qquad (143)$$

$$\langle \tilde{\chi}_p|\tilde{\chi}_q\rangle = \delta_{p,q}. \qquad (144)$$

The eigenvectors $\tilde{\chi}_p$ with eigenvalues $\tilde{E}_p$ are called *pseudostates* that form a discrete variational representation of the actual spectrum of the system (including the continuous spectrum of $H^{(0)}$; see Fig. 3). Although the diagonalization step is computationally slow, the advantage gained is that the system of equations (141) is brought into diagonal form with the immediate solutions

$$\tilde{b}_p = \frac{\langle \tilde{\chi}_p|V - E^{(1)}|\Psi^{(0)}\rangle}{E^{(0)} - \tilde{E}_p}, \quad p = 1, 2, \ldots, N. \qquad (145)$$

A term with $\tilde{E}_p = E^{(0)}$, if present, is simply omitted. Without loss of generality, one can then assume that $\langle \tilde{\chi}_p|\Psi^{(0)}\rangle = 0$ for the remaining $\tilde{\chi}_p$. Equation (140) then becomes

$$E^{(2)} = \sum_{p=1}^{N}{}' \frac{|\langle \tilde{\chi}_p|V - E^{(1)}|\Psi^{(0)}\rangle|^2}{E^{(0)} - \tilde{E}_p}, \qquad (146)$$

where the prime denotes that terms with $\tilde{E}_p = E^{(0)}$ are to be omitted. This expression for $E^{(2)}$ is formally identical to the standard second-order perturbation expression, except that the summation over the actual spectrum of $H^{(0)}$ (including an integration over the continuum) is replaced by a summation over the discrete variational pseudostates. The results are completely equivalent to those obtained by solving Eqs. (141) directly. A similar formal identity can be extended to all orders of perturbation theory.

From a more general point of view, the above results correspond to a discrete variational representation for the Green's function of complex variable z, defined by

$$(H^{(0)} - z)G^{(0)}(\mathbf{r},\mathbf{r}') = \delta(\mathbf{r} - \mathbf{r}'). \qquad (147)$$

Then

$$G^{(0)}(\mathbf{r},\mathbf{r}') \approx \sum_{p=1}^{N} \frac{|\tilde{\chi}(\mathbf{r})\rangle\langle\tilde{\chi}(\mathbf{r}')|}{\tilde{E}_p - z}. \tag{148}$$

The great advantage is that often a small number of appropriately chosen pseudostates can replace the infinite summation over bound states plus an integration over the continuous spectrum of $H^{(0)}$.

The pseudostate method has important applications in scattering theory as an alternative way to represent the scattering continuum. For example, the convergent close-coupling method of Bray and Stelbovics (1992) has yielded essentially exact solutions to the electron–hydrogen scattering problem.

3.4.3 Example: The Polarizability of Hydrogen Consider the problem of a hydrogen atom in its 1*s* ground state subjected to a static electric field of strength F pointing in the z direction. The total Hamiltonian in polar coordinates is then

$$H = H^{(0)} + eFr\cos\theta. \tag{149}$$

With $V = eFr\cos\theta$ as the perturbation, the first-order equation (135) can be solved analytically with the result

$$\Psi^{(1)} = -(1/\sqrt{3})(2r + r^2)e^{-r}Y_1^0(\hat{\mathbf{r}}). \tag{150}$$

Since V is of odd parity, $E^{(1)} = 0$, and, from Eq. (140), $E^{(2)} = -\frac{9}{4}a_0^3$, where $E^{(2)}$ is the coefficient of F^2 in the energy expansion. By definition, the dipole polarizability is

$$\alpha_{\mathrm{d}} \equiv -2E^{(2)} = \tfrac{9}{2}a_0^3. \tag{151}$$

If $E^{(2)}$ is written in the form of Eq. (146), summed over the actual spectrum of hydrogen, then nearly half of α_{d} comes from virtual transitions to the continuum.

Suppose now that a variational solution is constructed of the form

$$\Psi_{\mathrm{tr}}^{(1)} = -(1/\sqrt{3})(b_1 r + b_2 r^2)e^{-\lambda r}Y_1^0(\hat{\mathbf{r}}), \tag{152}$$

where b_1 and b_2 are linear variational parameters, and λ is an additional nonlinear variational parameter. This provides a two-dimensional basis set, with the exact solution being recovered for the case $\lambda = 1$. For $\lambda \neq 1$, the basis set provides the best variational representation of $\Psi^{(1)}$. After solving Eq. (141) for b_1 and b_2, the expression for α_{d} as a function of λ becomes

$$\alpha_{\mathrm{d}}(\lambda) = 6\lambda^5\left(\frac{2}{\lambda + 1}\right)^{12} \times \frac{9\lambda^4 - 12\lambda^3 + 14\lambda^2 - 10\lambda + 5}{5\lambda^4 - 10\lambda^3 + 18\lambda^2 - 10\lambda + 5}. \tag{153}$$

A graph of this function near its peak is displayed in Fig. 4. Since $E^{(2)}$ is an upper bound, α_{d} is a lower bound for any λ. It is clear that the exact value of $4.5a_0^3$ is recovered at $\lambda = 1$, where there is an absolute maximum. But what is also significant is the broad region of stability for $0.6 \leq \lambda \leq 1.1$, with a second local maximum near $\lambda = 0.657$. For any λ in this region, α_{d} is in error by no more than 0.6%, even with a basis set containing only two terms. In other words, the entire spectrum of hydrogen is being well represented by just two pseudostates, neither of which corresponds to a physical state of hydrogen. In fact, the accuracy is considerably worse at $\lambda = \frac{1}{2}$, corresponding to the physical 2*p* state. As higher powers of r are added to the basis set, the region of stability rapidly becomes broader and flatter.

3.4.4 Time-Dependent Perturbations Similar methods can be used to treat the case where the perturbation V is time depen-

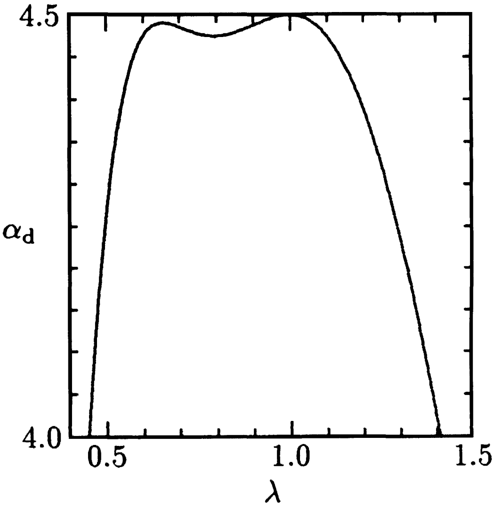

FIG. 4. Variational polarizability α_d of hydrogen, in units of a_0^3. The exact value at $\lambda = 1$ is $\alpha_d = 4.5\ a_0^3$.

dent. For example, with the replacement $E^{(0)} \rightarrow i\hbar\partial/\partial t$, the variational condition

$$\delta\left[\left\langle \Psi^{(1)}(t)\middle| H^{(0)} - i\hbar\frac{\partial}{\partial t}\middle|\Psi^{(1)}(t)\right\rangle + 2\langle\Psi^{(1)}(t)|V(t)|\Psi^{(0)}(t)\rangle\right] = 0 \qquad (154)$$

with respect to $\Psi^{(1)}(t)$ leads to the first-order time-dependent perturbation equation

$$\left(H^{(0)} - i\hbar\frac{\partial}{\partial t}\right)|\Psi^{(1)}(t)\rangle + V(t)|\Psi^{(0)}(t)\rangle = 0. \qquad (155)$$

This can be solved by Dirac's method of variation of constants. Many other techniques have been developed, but these will not be further pursued here. See, for example, Dalgarno (1966).

4. THE GENERAL STURM–LIOUVILLE PROBLEM

Many of the variational techniques discussed in Sec. 3 were developed long before the invention of quantum mechanics, in connection with boundary-value problems in classical physics such as vibrating membranes. Any linear second-order differential equation (of which the radial Schrödinger equation is just one example) can be written in the Sturm–Liouville form

$$\frac{d}{dx}\left[K(x)\frac{dy}{dx}\right] - G(x)y = 0, \qquad (156)$$

defined over some closed interval $[a,b]$, together with suitable boundary conditions. By application of the Euler–Lagrange equation with fixed end points, this equation follows from the variational condition

$$\delta\int_a^b\left[K(x)\left(\frac{dy}{dx}\right)^2 + G(x)y^2\right]dx, \qquad (157)$$

and so all the techniques discussed thus far can be applied. With the choice

$$G(x) = -\lambda g(x) + l(x), \qquad (158)$$

the Sturm–Liouville problem becomes an eigenvalue problem with λ adjusted to satisfy the boundary conditions.

4.1 The Oscillation Theorem

A great many theorems have been proven concerning the solutions to Sturm–Liouville problems (see, e.g., Ince, 1956). Of particular importance for physical applications is the oscillation theorem. Suppose that $K(x)$, $g(x)$, and $l(x)$ are all continuous, real, positive, monotonic decreasing functions of x in the interval $[a,b]$. It can then be proved that the two-point eigenvalue problem

$$\frac{d}{dx}\left[K(x)\frac{dy}{dx}\right] + [\lambda g(x) - l(x)]y = 0 \qquad (159)$$

has an infinite sequence of increasing eigenvalues $\lambda_1, \lambda_2, \ldots$, with corresponding eigenvectors $y_1(x), y_2(x), \ldots$ such that $y_m(x)$ has exactly $m - 1$ zeros in the open interval $[a,b]$. The eigenvalues are entirely discrete. If $g(x)$ changes sign in the interval, then the sequence of eigenvalues becomes doubly infinite with both an increasing $(\lambda_m^{(+)})$ and a decreasing $(\lambda_m^{(-)})$ set. In either case, the solutions are orthogonal with respect to the weight function $g(x)$.

The importance of this theorem (and its extensions) is that the $y_m(x)$ form the basis for a generalized Fourier series in terms of which an arbitrary function $f(x)$ can be expanded in the form

$$f(x) = \sum_{m=1}^{\infty} c_m y_m(x) \qquad (160)$$

with

$$c_m = \int_a^b f(x)g(x)dx. \qquad (161)$$

Since the eigenvalues are entirely discrete, there is no integration over a continuum in Eq. (160). Such a basis is called a *Sturmian* basis set. Most of the mathematical apparatus developed for Fourier analysis can be carried over directly. In fact, a Fourier series just corresponds to the choices $K(x) = 1$, $g(x) = 1$, $l(x) = 0$.

4.2 Example: The Coulomb Problem

Consider the radial Schrödinger equation

$$\left[-\frac{1}{2r^2}\frac{d}{dr}r^2\frac{d}{dr}+\frac{l(l+1)}{2r^2}-\frac{Z}{r}\right]R(r)=ER(r) \tag{162}$$

for an electron moving in the field of a nucleus with charge Z. The quantum-mechanical eigenvalue problem is solved on the interval $[0,\infty]$ by holding Z fixed and varying E such that the boundary conditions

$$\lim_{r\to 0} rR(r)=0,\quad \lim_{r\to\infty} R(r)=0, \tag{163}$$

are satisfied for the infinity of bound states with $E < 0$. There is in addition a continuum of scattering solutions with $E > 0$.

The Sturmian eigenvalue problem differs in that E is held fixed at some negative value $-\epsilon$ with $\epsilon > 0$, and Z is varied so as to satisfy the boundary conditions. Since the eigenvalues to the Coulomb problem are

$$E_n(Z) = -Z^2/2n^2, \tag{164}$$

it is clear that the Sturmian eigenvalue condition $E_n(Z) = -\epsilon$ can be satisfied infinitely many times by progressively increasing both n and Z. As Z increases, one eigenvalue after another from the original problem is pulled down through the value $-\epsilon$. The Sturmian eigenvalues are thus $Z_n = n\sqrt{\epsilon}$ and the corresponding eigenfunctions are

$$\begin{aligned} R_{nl}(r) = {} & \frac{1}{(2l+1)!}\left(\frac{(n+l)!}{(n-l-1)!2n}\right)^{1/2} \\ & \times (2\alpha)^{3/2}(2\alpha r)^l e^{-\alpha r} \\ & \times F(-n+l+1, 2l+2; 2\alpha r), \end{aligned} \tag{165}$$

where $\alpha = \sqrt{2\epsilon}$ and $F(a,b;z)$ is a confluent hypergeometric function. The $R_{nl}(r)$ form a complete set of finite Sturmian polynomials for $n \geq l+1$ that are orthogonal with respect to the potential $1/r$; i.e.

$$\int_0^\infty R_{n'l}(r)\frac{1}{r}R_{nl}(r)r^2dr = \epsilon\delta_{n',n}. \tag{166}$$

The Sturmian functions $R_{nl}(r)$ closely resemble the bound-state Coulomb wave functions. The main distinguishing feature is that α is a constant in the exponential factor instead of decreasing as $1/n$. The first N of them differ only by a transformation of the basis set from the functions used to construct a finite variational representation of $\Psi_{\mathrm{tr}}^{(1)}$ in Sec. 3.4.3. The theory of Sturmian functions therefore provides a rigorous foundation for the choice of basis functions in variational calculations, and their property of completeness ensures convergence to the correct answer as N increases.

5. APPLICATIONS TO ELECTRODYNAMICS

Consider an electromagnetic field propagating through a medium with a charge density $\rho(\mathbf{r})$ moving with velocity $\mathbf{v}(\mathbf{r})$. It follows from Maxwell's equations that the scalar and vector potentials introduced in Sec. 2.4.1 satisfy the equations

$$\nabla^2\mathbf{A} - \frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2} = -4\pi\rho\frac{\mathbf{v}}{\mathbf{c}}, \tag{167a}$$

$$\nabla^2\phi - \frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} = -4\pi\rho, \tag{167b}$$

provided that the Lorentz gauge condition

$$\nabla\cdot\mathbf{A} + \frac{1}{c}\frac{\partial\phi}{\partial t} = 0 \tag{168}$$

is imposed. Many other gauge choices can be made that leave the physical fields **E** and **B** invariant, but this one is simplest for the present discussion.

Unlike problems involving point particles, we are now dealing with fields that vary continuously in space. Equations (167a) and (167b) can be derived from a variational principle if a Lagrangian density $\mathcal{L}$ is first defined such that

$$L = \int \mathcal{L}dxdydz. \tag{169}$$

The action integral in Hamilton's principle then assumes the four-dimensional form

$$J = \int \mathcal{L}dxdydzdt. \tag{170}$$

The condition $\delta J = 0$ is obtained in a manner similar to that leading to Laplace's equation (31). The present case is an application of Eq. (23) with $f = \mathcal{L}$ and four independent variables $t_1 = x$, $t_2 = y$, $t_3 = z$, $t_4 = t$. There will be four equations corresponding to $q_1 = A_1$, $q_2 = A_2$, $q_3 = A_3$, and $q_4 = \phi$. The choice of $\mathcal{L}$ is severely restricted for fields *in vacuo* by the requirement that it be quadratic in the field components (since the field equations are linear), and relativistically invariant. The only quantity satisfying both requirements is a term proportional to $E^2 - B^2$. The inhomogeneous interaction terms on the right-hand sides of Eqs. (167a) are also included with the definition

$$\mathcal{L} = \frac{1}{8\pi}\left[\left(-\nabla\phi - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t}\right)^2 - (\nabla \times \mathbf{A})^2 - \left(\nabla\cdot\mathbf{A} + \frac{1}{c}\frac{\partial\phi}{\partial t}\right)^2\right] + \rho\left(\mathbf{A}\cdot\frac{\mathbf{v}}{c} - \phi\right), \tag{171}$$

where the first two squared terms correspond to E^2 and $-B^2$, respectively, and the third term generates the Lorentz gauge condition (168). The last term is the interaction term. A straightforward application of Eq. (23) with the terms identified as described following Eq. (170) then yields Eqs. (167a) and (167b).

The Lagrangian density $\mathcal{L}$ gives the equations of motion for the fields in the presence of a predefined matter distribution $\rho(\mathbf{r},t)$. For comparison, the Lagrangian L defined by Eq. (96) gives the converse equations of motion for particles in the presence of predefined fields $\mathbf{A}$ and ϕ. The remarkable point emerging from a comparison of $\mathcal{L}$ and L is that the matter-field interaction term $\rho(\mathbf{A}\cdot\mathbf{v}/c - \phi)$ in $\mathcal{L}$ is very similar in form to the second and third terms in L. In fact, the terms become identical with the choice $\rho(\mathbf{r}) = q\delta(\mathbf{r} - \mathbf{r}')$, corresponding to a point particle of charge q at position $\mathbf{r}'$. This suggests that the two Lagrangians can be combined into a single Lagrangian

$$L_{\text{tot}} = L_0 + \int \mathcal{L}_0 dxdydz + L_{\text{int}}, \tag{172}$$

where L_0 and $\mathcal{L}_0$ are the Lagrangians for free particles and free fields, respectively, and L_{int} is the remaining interaction term common to both L and $\int \mathcal{L}\, d\mathbf{r}$. Hamilton's principle and the Euler–Lagrange equations then give the equations of motion for the combined system of interacting particles and fields.

The above is of course not a proof that the resulting equations of motion provide an exact description of nature. The derivation is based on the supposition that something like Hamilton's principle remains valid for the combined system of interacting particles and fields, and it ignores the quantum nature of both matter and fields. However, Eq. (172) provides a basis for combining a quantized description of matter fields and electromagnetic fields into a single theory called *quantum electrodynamics*, whose predictions have been verified to an extremely high degree of precision (see, e.g., Kinoshita and Yennie, 1990). It can safely be described as the most successful theory ever invented. However, further discussion of this topic would take us beyond the scope of this article (see Further Reading and the article QUANTUM ELECTRODYNAMICS).

6. FEYNMAN'S PATH INTEGRAL

This article would not be complete without at least a passing reference to Feynman's path integral because of the way in which it provides an underlying coherent formalism unifying classical mechanics, quantum mechanics, and optics.

Consider for simplicity the x coordinate of a particle moving in a potential. The aim is to construct a path integral giving the quantum-mechanical transition amplitude for the particle to move from position x_0 at time t_0 to position x_f at time t_f. Let the state vector corresponding to a particle at position x be denoted by $|x\rangle$. In the coordinate representation

$$\langle x'|x\rangle = \delta(x - x'), \tag{173}$$

and by closure,

$$\int |x\rangle\langle x| = \mathbf{1}, \tag{174}$$

where $\mathbf{1}$ is the identity operator. In the momentum representation,

$$\langle p|x\rangle = (2\pi\hbar)^{-1/2}e^{-px/\hbar} = \langle x|p\rangle^*. \tag{175}$$

The remaining ingredient is the time-evolution operator $e^{-iHt/\hbar}$. Its inverse defines states $|x,t\rangle$ in the Heisenberg representation such that

$$|x,t\rangle = e^{iHt/\hbar}|x\rangle. \tag{176}$$

With these preliminaries, an initial expression for the desired transition amplitude is

$$\begin{aligned} K(x_f,t_f;x_0,t_0) &= \langle x_f,t_f|x_0,t_0\rangle \\ &= \langle x_f|\exp[-iH(t_f - t_0)/\hbar]|x_0\rangle. \end{aligned} \tag{177}$$

The key idea in constructing a path integral is to suppose that the system passes through a large number N of intermediate states $|x_k,t_k\rangle$ in making the above transition from $|x_0,t_0\rangle$ to $|x_f,t_f\rangle$ such that $t_{k+1} = t_k + \epsilon$, with $\epsilon = (t_f - t_0)/(N + 1)$. This can be formally achieved by making repeated use of the closure relation (174) to write

$$\begin{aligned} &K(x_f,t_f;x_0,t_0) \\ &\quad = \int dx_1\cdots dx_N\langle x_f,t_f|x_N,t_N\rangle \\ &\quad\quad \times \langle x_N,t_N|x_{N-1},t_{N-1}\rangle\cdots\langle x_1,t_1|x_0,t_0\rangle. \end{aligned} \tag{178}$$

Then for each intermediate step, the inner product is

$$\begin{aligned} &\langle x_k,t_k|x_{k-1},t_{k-1}\rangle \\ &\quad = \langle x_k|\exp[-i(t_k - t_{k-1})H/\hbar]|x_{k-1}\rangle \\ &\quad = \langle x_k|e^{-i\epsilon H/\hbar}|x_{k-1}\rangle. \end{aligned} \tag{179}$$

Using Eq. (175), this can be evaluated in the momentum representation to obtain (with symmetric or Weyl operator ordering)

$$\begin{aligned} &\langle x_k,t_k|x_{k-1},t_{k-1}\rangle \\ &\quad = \int \frac{dp_k}{2\pi\hbar}\exp\{i[p_k\Delta x_k - i\epsilon H(\bar{x}_k,p_k)]/\hbar\}, \end{aligned} \tag{180}$$

where $\Delta x_k = x_k - x_{k-1}$ and $\bar{x}_k = (x_k + x_{k-1})/2$. Substitution of this form into Eq. (178) then yields

$$K(x_f,t_f;x_0,t_0) = \int dx_1\cdots dx_N\frac{dp_1}{2\pi\hbar}\cdots\frac{dp_{N+1}}{2\pi\hbar}e^{iS_N}, \tag{181}$$

where

$$S_N = \frac{\epsilon}{\hbar}\sum_{k=1}^{N+1}\left[\frac{p_k\Delta x_k}{\epsilon} - H(\bar{x}_k,p_k)\right]. \tag{182}$$

Consider now the limit $N \to \infty$, $\epsilon \to 0$. Although each x_k separately ranges over all possible values due to the integrations in Eq. (181) (see Fig. 5), the quantity $\Delta x_k/\epsilon \equiv (x_k - x_{k-1})/\epsilon$ contributes a large and rapidly varying phase that averages to zero unless $x_k \approx x_{k-1}$. We can therefore identify $\Delta x_k/\epsilon$ with $\dot{x}$ in the limit $\epsilon \to 0$ and write

$$\lim_{\epsilon\to 0} S_N = \frac{1}{\hbar}\int_{t_0}^{t_f}[p\dot{x} - H(x,p)]dt = \frac{1}{\hbar}\int_{t_0}^{t_f} L\,dt. \tag{183}$$

This result provides a version of the Feynman path integral with the path integrated over all possibilities (including discontinuous ones) in phase space. However, the result can be carried a step further by performing the momentum integrations in Eq. (181). For example, if H has the form

$$H(x,p) = p^2/2m + V(x), \tag{184}$$

then a typical momentum integral has the Gaussian form

$$\begin{aligned} &\int_{-\infty}^{\infty}\frac{dp_k}{2\pi\hbar}\exp\left\{-\frac{i\epsilon}{\hbar}\left[\frac{p_k^2}{2m} - \frac{p_k\Delta x_k}{\epsilon}\right]\right\} \\ &\quad = \left(\frac{m}{2\pi\hbar i\epsilon}\right)^{1/2}\exp\left[\frac{im\epsilon}{2\hbar}\left(\frac{\Delta x_k}{\epsilon}\right)^2\right]. \end{aligned} \tag{185}$$

Using this in Eq. (181) gives

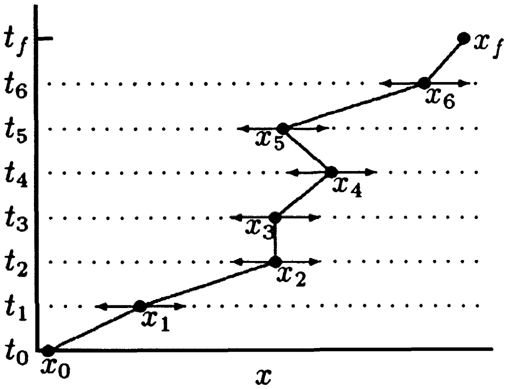

FIG. 5. A possible path for the Feynman integral with six intermediate states. Each of x_1 through x_6 varies independently over all possible values.

$$K(x_f,t_f;x_0,t_0) = \left(\frac{m}{2\pi\hbar i\epsilon}\right)^{(N+1)/2} \int dx_1 \cdots dx_N e^{iS_N}, \tag{186}$$

where now

$$S_N = \frac{\epsilon}{\hbar} \sum_{k=1}^{N+1} \left[\frac{m\Delta x_k}{2\epsilon} - V(\bar{x}_k)\right]. \tag{187}$$

Taking the limit $N \to \infty$, $\epsilon \to 0$ as in Eq. (183) leads to the final result

$$K(x_f,t_f;x_0,t_0) = A \int \mathcal{D}x \exp\left(\frac{i}{\hbar}\int_{t_0}^{t_f} L\,dt\right), \tag{188}$$

where A is a constant independent of the path, and $\mathcal{D}x$ is a shorthand notation for the infinitely nested integrals in Eq. (186). Since this structure is mathematically not well defined, any practical calculation must be done with a finite subdivision and the limit $\epsilon \to 0$ taken at the end.

An explicit calculation of K is possible only for certain special cases such as the free particle or harmonic oscillator. In the free-particle case ($V = 0$), K becomes

$$K(x_f,t_f;x_0,t_0) = \left(\frac{m}{2\pi i\hbar\Delta t}\right)^{1/2} \exp\left[-\frac{m(\Delta x)^2}{2i\hbar\Delta t}\right], \tag{189}$$

where $\Delta t = t_f - t_0$ and $\Delta x = x_f - x_0$. This is the known nonrelativistic kernel or Green's-function propagator for a free particle (see, e.g., Merzbacher, 1970). It controls the time evolution of the wave function according to

$$\Psi(x_f,t_f) = \int_{-\infty}^{\infty} K(x_f,t_f;x,t_0)\Psi(x,t_0)dx. \tag{190}$$

6.1 Relation to Classical Dynamics

As it stands, the Feynman path integral is not strictly a variational principle. However, in the limit $\hbar \to 0$, the exponential term

$$\frac{i}{\hbar}\int_{t_0}^{t_f} L\,dt \tag{191}$$

in general gives a large phase that varies rapidly with slight variations in the path, except very near the particular path where the integral is stationary; i.e., by Hamilton's principle, the classical trajectory. For any other path, the rapid phase variations cause the contributions to the integral to cancel. All possible paths contribute with equal weight, but the phase cancellation singles out the classical path. It similarly singles out the path of a light ray in geometrical optics by Fermat's principle. The connection between variational principles in classical and quantum mechanics is discussed from a different point of view by Gray *et al.* (1996), who show that the reciprocal form of the Maupertuis principle of least action (see Sec. 2.3.1) can be obtained from the classical limit of the Schrödinger variational principle without the use of the Feynman path integral.

On the other hand, the path integral embodies quantum mechanics, and, with the use of the Lagrangian density, electrodynamics. It therefore provides a powerful underlying formalism that unifies all of these diverse branches of physics. It provides a general starting point for the construction of quantum field theories and the description of new phenomena. These topics may be further pursued through some of the books in the reading list.

ACKNOWLEDGMENTS

The author is grateful to the Natural Sciences and Engineering Research Council of Canada for financial support, and to Mark Cassar and Alice Kolakowska for a careful reading of the manuscript.

GLOSSARY

Brachistochrone Problem: The problem of finding the curve giving the shortest travel time for a bead constrained to slide on a wire of arbitrary shape.

Calculus of Variations: A mathematical technique for finding the extrema (maxima, minima, or stationary points) with respect to variations in the function defining the path for an integration.

Canonical Momentum: The momentum p_i conjugate to a generalized coordinate q_i, defined by $p_i = \partial L/\partial \dot{q}_i$, where $\dot{q}_i = dq/dt$ and L is the Lagrangian.

Canonical Transformation: A transformation from one set of generalized coordinates and momenta to another such that the form of Hamilton's equations of motion is preserved.

Conservative System: A mechanical system for which the Hamiltonian is a constant of the motion.

Constraint: An equation that must be satisfied by the solution to a variational problem.

Correlation Energy: The difference between the exact and Hartree–Fock energies of a many-particle quantum-mechanical system.

Electrodynamics: The study of electromagnetic fields and their interactions with charged particles, treated as a single dynamical system.

Euler–Lagrange Equation: A second-order differential equation whose solutions determine the function that makes a path integral an extremum.

Fermat's Principle: A principle from geometrical optics stating that the path of a light beam through a medium of varying index of refraction is such that the travel time is a minimum.

Feynman Path Integral: An integral over all possible paths (both continuous and discontinuous ones) connecting the initial and final states of a system, with each path weighted by a phase factor of the form $\exp(iS/\hbar)$, where S is the path integral over the Lagrangian appearing in Hamilton's principle, and $\hbar$ is Planck's constant.

Functional: A function $F[y(x)]$ that depends on the functional form chosen for $y(x)$. For example, $y(x)$ might determine the path of an integration between fixed end points in a plane.

Generalized Coordinate: Any function of the Cartesian coordinates used to describe the evolution of a mechanical system, or its dynamical analogues involving continuous fields.

Generalized Force: The component of the work done by the actual force acting on a particle due to a change of a generalized coordinate.

Generalized Momentum: The momentum conjugate to a generalized coordinate—see **Canonical Momentum.**

Hamiltonian: A constant of the motion for time-independent systems, which for ordinary conservative systems becomes the sum $T + V$ of the kinetic energy T and the potential energy V.

Hamilton's Characteristic Function: The time-independent part of Hamilton's principal function.

Hamilton's Equations of Motion: A set of $2N$ first-order differential equations linking directly the N generalized coordinates and N generalized momenta through their time derivatives.

Hamilton's Principal Function: The function represented formally by the indefinite integral of the Lagrangian with respect to time. It corresponds to the general solution to the Hamilton–Jacobi equation.

Hamilton's Principle: A variational condition, involving a path integral over the Lagrangian with respect to time, that determines the equations of motion for the system under consideration.

Hartree–Fock Approximation: The best possible approximation to the wave function for a many-particle quantum-mechanical system that can be written in the form of a separable product of independent functions, with a separate function for each particle. The coupled differential equations to be solved follow from the Rayleigh–Schrödinger variational principle.

Holonomic Constraint: An equation of constraint involving the generalized coordinates that can be expressed in integrated form, as opposed to a relation among differential quantities.

Hylleraas–Undheim–MacDonald Theorem: A theorem establishing variational upper bounds for the energies of excited states of a quantum-mechanical system as well as for the ground state.

Lagrange Undetermined Multiplier: A mathematical technique for the introduction of constraints into variational problems.

Lagrange's Equations of Motion: A system of N second-order partial differential equations that determine the equations of motion for the generalized coordinates $q_i = q_i(t)$.

Lagrangian: For ordinary conservative systems, the Lagrangian L is the quantity $T - V$, expressed as a function of the N generalized coordinates q_i and their time derivatives $\dot{q}_i$.

Lagrangian Density: A generalization of the Lagrangian to continuous systems and fields in which the continuous system is first conceived as consisting of discrete elements.

Oscillation Theorem: A theorem applicable to two-point boundary-value problems of the Sturm–Liouville type. The theorem establishes that there is an infinite sequence of eigenvalues whose eigenfunctions have progressively more zeros between the boundary points, and hence progressively more oscillations.

Path Integral: The integral of a function $f(y(x))$ between given points in an xy plane (or its higher-dimensional generalizations) along a path specified by the function $y(x)$.

Perturbation Theory: A technique for the progressive approximation of more difficult problems involving the solution of differential equations (for example), starting from an exactly soluble simpler one. The difference between the two equations is called the perturbation term.

Principle of Least Action: A variational principle in classical dynamics, closely related to Hamilton's principle, which establishes a direct connection with Fermat's principle in geometrical optics.

Pseudostate: A member of a set of states obtained by diagonalization of the Hamiltonian matrix in a discrete variational basis set.

Quantum Electrodynamics: A quantized field theory describing the dynamical interactions of charged particles with electromagnetic fields.

Rayleigh–Ritz Variational Method: A method for the construction of an approximate wave function Ψ by expansion in a finite basis set of functions with expansion coefficients determined by the Rayleigh–Schrödinger variational principle.

Rayleigh–Schrödinger Variational Principle: A principle stating that the ratio $\langle\Psi|H|\Psi\rangle/\langle\Psi|\Psi\rangle$ is an upper bound to the lowest eigenvalue of H for any arbitrary (but normalizable) choice for the wave function Ψ.

Schrödinger Equation: A second-order partial differential wave equation that forms the basis of nonrelativistic quantum mechanics.

Sturmian Basis Set: The set of discrete eigenvalues and corresponding eigenfunctions obtained by solving a two-point eigenvalue problem of the Sturm–Liouville type.

Sturm–Liouville Problem: A class of second-order differential equations of the form $(d/dx)[K(x)dy/dx] - G(x)y = 0$, together with suitable boundary conditions.

Spectral Representation: A representation of the Green's function, or the terms in a perturbation series, in terms of explicit summations over the eigenvalue spectrum of the unperturbed problem.

Temple Bound: A method for constructing variational lower bounds to the energy based on the square of the Hamiltonian.

Variational Bound: An upper or lower bound on the energy or some other quantity obtained by means of a trial solution to the underlying differential equation, typically with parameters in the trial solution that can be adjusted to obtain the best possible solution.

Works Cited

Ackermann, J. (1995), *Phys. Rev. A* **52,** 1968–1975, and earlier references therein.

Alexander, S. A., Coldwell, R. L., Monkhorst, H. J., Morgan, J. D., III (1991), *J. Chem. Phys.* **95,** 6622–6633.

Barnett, R. N., Johnson, E. M., Lester, W. A., Jr. (1995), *Phys. Rev. A* **51,** 2049–2052.

Bray, I., Stelbovics, A. T. (1992), *Phys. Rev. A* **46,** 6995–7011, and earlier references therein.

Cohen, M. R., Drabkin, I. E. (1958), *A Source Book in Greek Science,* Cambridge, MA: Harvard Univ. Press.

Courant, R., Hilbert, D. (1966), *Methods of Mathematical Physics,* New York: Interscience, p. 190.

Dalgarno, A. (1966), in: C. H. Wilcox (Ed.), *Perturbation Theory and Its Applications in Quantum Mechanics,* New York: Wiley, pp. 145–183.

Dalgarno, A., Lewis, J. T. (1955), *Proc. R. Soc. (London) Ser. A* **233,** 70–74.

Dalgarno, A., Lewis, J. T. (1956), *Proc. Phys. Soc. (London) A* **69,** 628–630.

Dalgarno, A., Stewart, A. L. (1956), *Proc. R. Soc. (London) Ser. A* **238,** 269–275.

Drake, G. W. F. (1993a), in: F. S. Levin, D. A. Micha (Eds.), *Long-Range Casimir Forces: Theory and Recent Experiments on Atomic Systems,* New York: Plenum.

Drake, G. W. F. (1993b), *Adv. At. Mol. Opt. Phys.* **31,** 1–62.

Drake, G. W. F. (1994), *Adv. At. Mol. Opt. Phys.* **32,** 93–116.

Drake, G. W. F. (1996), in: G. W. F. Drake (Ed.) *Atomic, Molecular, and Optical Physics Handbook,* New York: American Institute of Physics, pp. 154–171.

Drake, G. W. F., Goldman, S. P. (1988), *Adv. At. Mol. Phys.* **25,** 393–416.

Drake, G. W. F., Khriplovich, I. B., Milstein, A. I., Yelkovski, A. S. (1993), *Phys. Rev. A* **48,** R15–R17.

Eikema, K. S. E., Ubachs, W., Vassen, W., Hogorvorst, W. (1996), *Phys. Rev. A* **55,** 1866–1880.

Glover, R. M., Weinhold, F. (1976), *J. Chem. Phys.* **65,** 4913–4926, and other references therein.

Goldman, S. P. (1997), *Phys. Rev. Lett.* **78,** 2325–2328.

Grant, I. P. (1996), in: G. W. F. Drake (Ed.), *Atomic, Molecular, and Optical Physics Handbook,* New York: American Institute of Physics, pp. 258–286.

Gray, C. G., Karl, G., Novikov, V. A. (1996), *Ann. Phys. (N.Y.)* **251,** 1–25.

Hildebrand, S., Tromba, A. (1985), *Mathematics and Optimal Form,* San Francisco: W. H. Freeman (Scientific American Library).

Hylleraas, E. A. (1928), *Z. Phys.* **48,** 469–494.

Hylleraas, E. A. (1929), *Z. Phys.* **54,** 347–366.

Hylleraas, E. A., Undheim, B. (1930), *Z. Phys.* **65,** 759–772.

Ince, E. L. (1956), *Ordinary Differential Equations,* New York: Dover.

Jeffreys, H., Jeffreys, B. S. (1972), *Methods of Mathematical Physics,* Cambridge, U.K.: Cambridge Univ. Press, p. 315.

Kinoshita, T., Yennie, D. R. (1990), in: T. Kinoshita (Ed.), *Quantum Electrodynamics,* Singapore: World Scientific, and other articles contained therein.

Klahn, B., Bingel, W. A. (1977), *Theoret. Chim. Acta* **44,** 27–43. See also Klahn, B., Bingel, W. A. (1978), *Int. J. Quantum Chem.* **11,** 943–957.

Lichten, W., Shiner, D., Zhou, Z.-X. (1991), *Phys. Rev. A* **43,** 1663–1665.

MacDonald, J. K. L. (1933), *Phys. Rev.* **43,** 830–833.

Merzbacher, E. (1970), *Quantum Mechanics,* New York: Wiley, p. 163.

Moskowitz, J. W., Schmidt, K. E., Lee, M. A., Kalos, M. H. (1982), *J. Chem. Phys.* **77,** 349–355.

Plante, D. R., Johnson, W. R., Sapirstein, J. (1994), *Phys. Rev. A* **49,** 3519–3530, and earlier references therein.

Sansonetti, C. J., Gillaspy, J. D. (1992), *Phys. Rev. A* **45,** R1–R3.

Schwartz, C. (1959), *Ann. Phys. (N.Y.)* **2,** 156–169, 170–177.

Slater, J. C., Kirkwood, J. G. (1931), *Phys. Rev.* **37,** 682–697.

Sternheimer, R. (1951), *Phys. Rev.* **84,** 244–253.

Sternheimer, R. (1954), *Phys. Rev.* **96,** 951–968.

Sternheimer, R. (1957), *Phys. Rev.* **105,** 158–169.

Temple, G. (1928), *Proc. R. Soc. (London) Ser. A* **119,** 276.

Yan, Z.-C., Drake, G. W. F. (1995), *Phys. Rev. A* **52,** 3711–3717, R4316–R4319.

Yan, Z.-C., McKenzie, D. K., Drake, G. W. F. (1996), *Phys. Rev. A* **54,** 1322–1327.

Further Reading

Techniques of the calculus of variations are covered in great detail by Courant and Hilbert (1966), and by Morse, P. M., and Feshbach, H. (1953), *Methods of Theoretical Physics,* New York: McGraw-Hill, along with most other books on the techniques of theoretical physics. The book by Yourgrau, W., and Mandelstam, S. (1968), *Variational Principles in Dynamics and Quantum Theory,* 3rd ed., Philadelphia: Saunders (Dover reprint 1979) provides an interesting and informative historical perspective.

Applications of variational principles to classical mechanics are covered in a very thorough, detailed, and readable manner by Goldstein, H. (1980), *Classical Mechanics,* 2nd ed., Reading, MA: Addison-Wesley. This book also contains a good pedagogical introduction to variational principles for continuous systems and fields. See also Lanczos, C. (1970), *The Variational Principles of Mechanics,* Toronto: University of Toronto Press (Dover reprint 1986).

The development of field theory from variational principles is covered in many recent books such as Ramond, P. (1981), *Field Theory, a Modern Primer,* Menlo Park, CA: Benjamin/Cummings; and Itzykson, C., and Zuber, J.-B. (1980), *Quantum Field Theory,* New York: McGraw-Hill. A good introduction to the Feynman path integral approach is given by Das, A. (1993), *Field Theory, a Path Integral Approach,* Singapore: World Scientific, and more detailed developments are contained in Rivers, R. J. (1987), *Path Integral Methods in Quantum Field Theory,* Cambridge, U.K.: Cambridge Univ. Press. The original development in Feynman, R. P., and Hibbs, A. R. (1965), *Quantum Mechanics and Path Integrals,* New York: McGraw-Hill, remains an authoritative source.

A wide variety of applications of variational principles in quantum-mechanical calculations can be found in numerous articles throughout Drake, G. W. F. (Ed.) (1996), *Atomic, Molecular, and Optical Physics Handbook,* New York: American Institute of Physics. In addition, the Kohn and Schwinger variational methods for scattering problems are covered in most books on scattering theory, such as Taylor, J. R. (1972), *Scattering Theory: The Quantum Theory on Nonrelativistic Collisions,* New York: Wiley. All of the above contain numerous other references to the literature and are intended only as a guide.

VERY LARGE-SCALE SIMULATION OF PHYSICAL SYSTEMS

Bruce M. Boghosian, *Center for Computational Science, Boston University, Boston, Massachusetts, U.S.A.*

Introduction 151
1. Computation and Its Relationship to Physics 152
1.1 The History of Computation 152
1.2 The Theory of Computation 154
1.3 Physical Computation 157
2. Languages for Scientific Computation 160
3. Data Structures and Algorithms 162
3.1 Algorithms and Computational Complexity 162
3.2 Parallel Computers 163
3.3 The Computer Representation of Real-Valued Data 166
4. Grid-Based Algorithms 167
4.1 Finite-Difference Methods 167
4.2 Banded-Matrix Equations 169
4.3 The Galerkin Approximation and Finite-Element Methods 172
5. Spectral Methods 173
5.1 Spectral Representation of Data 173
5.2 Operator Splitting 174
5.3 Spectral-Element Methods and Wavelets 176
6. Particulate Methods 179
6.1 Short-Range Interactions 180
6.2 Molecular Dynamics 181
6.3 Long-Range Interactions 182
6.4 Particle-in-Cell Methods 183
6.5 Lattice-Gas Hydrodynamics 184
7. Statistical Methods 186
7.1 Spin Systems 187
7.2 Metropolis Monte Carlo 188
7.3 Lattice Gauge Theory 190
8. The Future of Computational Physics 192
8.1 The State of the Art 192
8.2 Quantum Computation 192
9. Conclusion 195
Acknowledgments 195
Glossary 195
Works Cited 197
Further Reading 198

INTRODUCTION

The purpose of this article is to survey the history and the state of the art of the relationship between computation and physics. Traditionally, this topic has been divided into three major areas of study: *Physical computation* is the study of the use of physical systems, from simple mechanical devices to modern very large-scale integrated circuitry, to perform computation. *The theory of computation* is the mathematical specification of exactly what we mean by "computation" and of what computers can do and what they cannot do. *Computational physics* is then the study of how to use computers to simulate physical systems.

Recent developments in this field make it clear that all of these areas of study are closely related and that their relationship will only strengthen in the coming decades. The smallest electronic circuitry is already at the point where quantum-mechanical effects play a significant role. The consequences of the fact that certain physical systems are capable of computation have had a subtle but profound impact on physics, from the study of dynamical systems to the second law of thermodynamics. Finally, the recent realization that quantum computers may be able to do things that classical computers cannot (or at least do them much faster) reminds us that paradigms of computing are not unique mathematical abstractions but are rather dependent upon an underlying physical model. The observation that *computation is physical* is very analogous to—and every bit as profound as—the 19th century realization that

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Euclidean geometry is not a unique mathematical abstraction and the early 20th century addendum that the geometry of our physical world is in fact non-Euclidean.

This is a vast field, and so this article is necessarily in the nature of a very cursory survey. The object has been to give the reader many different and varied perspectives on how physicists use and think about computers, rather than a complete account of any one of them. Nevertheless, it is hoped that it will give the interested reader some appreciation of how and why people use physical systems to compute, of the nature of the computational process, of the implications of the theory of computation to physics, of the use of computation to study physics, and of the future of this entire enterprise.

1. COMPUTATION AND ITS RELATIONSHIP TO PHYSICS

1.1 The History of Computation

Since the first shepherds tied knots in strings to keep track of their flocks, mankind has employed physical systems to tabulate numerical data and perform computations. The reason then was the same as it is now: Physical devices can be used quite effectively to reduce the tedium of storing, manipulating, and retrieving large amounts of data. Other ancient calculational procedures may have involved the use of pebbles (indeed, the word "calculate" is derived from the Latin word for pebble), or marks on papyrus or clay to keep track of data. It is likely that early counters recognized that they could deal with data far more efficiently by employing representations for groups of objects, and then groups of groups of objects, etc. This observation gave rise to the *digital* representation of numbers (Schwartzman, 1996) and to computing machinery that stored and manipulated numeric data in this format. The *abacus* was developed in the Middle East at least 2500 years ago and uses wooden beads on a wire to add and subtract numbers in a digital representation.

Mechanical aids for multiplication and division are a much more recent development. In early 17th century Scotland, John Napier carved logarithmically spaced divisions on ivory sticks called *Napier's bones.* These could be used for multiplication and division and soon gave rise to the *slide rule,* which had both linearly and logarithmically spaced divisions and could be used to add, subtract, multiply, and divide. Slide rules remained in widespread use until the 1970s.

Meanwhile, the automation of tedious and difficult mechanical tasks in the Industrial Revolution gave rise to the idea that tedious and difficult calculations could likewise be performed by "computational engines." Thus, in Napoleonic France, Joseph Marie Jacquard developed a loom with an automatic card reader, so that information about desired patterns of colors on fabric could be easily encoded on punched cards. In Cambridge, England, in the mid-1800s, the mathematician Charles Babbage worked on the construction of a *difference engine,* and later an *analytical engine;* these would accept input from Jacquard-style punched cards and, by means of a complicated system of gears and cogs, compute and automatically print tables of numerical data. Though Babbage never finished these projects, he can be credited with designing the first general-purpose computer with data storage and printed output. Moreover, his collaboration with the mathematician Ada Augusta Lovelace gave rise to the first *computer programs,* including the *loop* construct, *subprograms,* and *conditional jumping.* By 1890, Herman Hollerith had convinced the United States Census Bureau to store information on punched cards, so that they could be processed by an electrical device.

The 20th century has seen spectacular advances—both quantitative and qualitative—in the field of computing. Between 1937 and 1944, Howard Aiken of Harvard University built a gigantic electrical computer called the Mark I, which could read and write punched cards and manipulate data by means of electrical switches and relays. It could add and subtract numbers in about 0.3 s, multiply them in less than 6 s, and divide them in less than 16 s. Meanwhile, at the University of Pennsylvania, Dr. John Mauchly and Dr. J. Presper Eckert developed the first electronic computer, called ENIAC, which used 18 800 vacuum tubes to perform computations. This represented a great advance in speed, as it could perform hundreds of multiplications and/or additions in 1 s; its memory, however, was limited to 20 ten-digit numbers. By 1949, the EDSAC (Electronic Delay Storage Automatic Calculator), built at Cambridge University in England, could perform both addition and multiplication in times on the order of a millisecond.

Thus was begun the exponential growth of computational speed and efficiency that, fueled by the microelectronics revolution of the late 20th century, continues to the present day. Currently, several figures of merit of modern computers seem to maintain an exponential rate of growth: At the time of this writing, modern processors can add and/or multiply hundreds of millions of *floating-point numbers* per second. (This figure of merit is sometimes called "FLOPS.") Clock rates are on the order of a nanosecond, and processor speed seems to double every 1.5 years or so. Figure 1 shows the amount of physical area required to store one *bit* of information as a function of time, beginning with relays as they were used in the Mark I and ending with modern integrated circuitry; the exponential rate of decrease of this quantity is apparent.

This stunning increase in computing capability has already had a profound impact on physics. First, and most clearly, it has elevated computation to a status on a par with that of theory and experimentation as a means of acquiring scientific knowledge. For example, though analytic methods were often able to discern the overall behavior of the *linear* equations of classical mathematical physics, such as the Maxwell equations, computation has made it possible to solve such equations in a much wider variety of geometries, with inhomogeneities, and irregular boundary and/or initial conditions. Moreover, certain *nonlinear* equations of mathematical physics, such as the Navier–Stokes equations of viscous fluid dynamics, have been known for more than a century but have only a very restricted set of known, nontrivial analytic solutions; the overall behavior of these equations is safely said to be unknown, as even the most elementary sta-

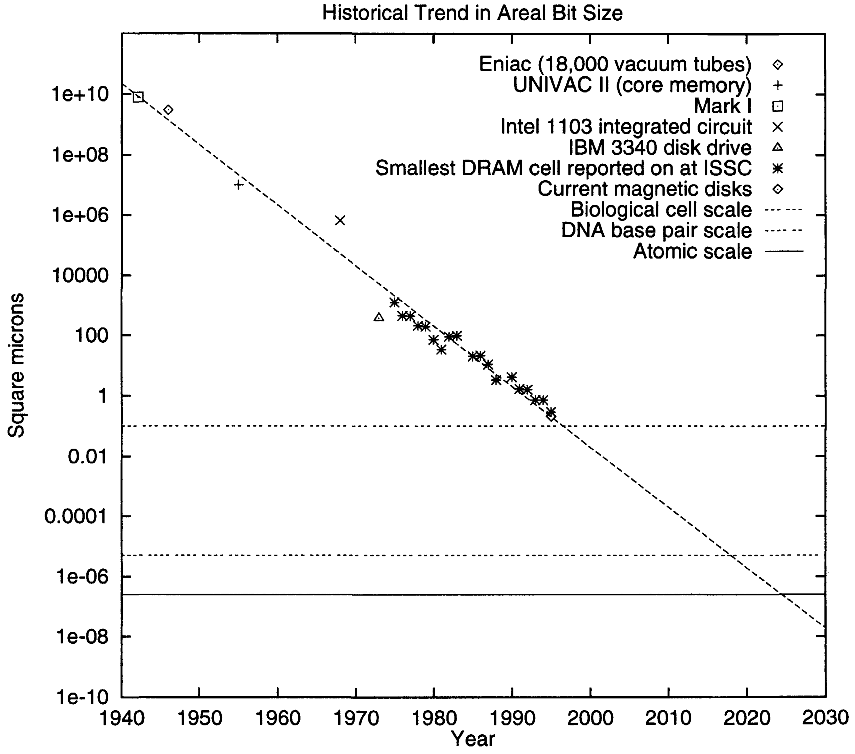

FIG. 1. Areal size required to store 1 bit of information, as a function of time (courtesy of Norman Margolus, MIT Laboratory for Computer Science).

tistical properties of high–Reynolds-number Navier–Stokes flow are analytically intractable. Computation has thus made it possible to examine the behavior of a much wider class of solutions to these equations and compare these to experiment, thereby enhancing our ability to test the laws of physics. Much of this article is devoted to outlining how computers are used for this purpose.

1.2 The Theory of Computation

Computers can simulate physical systems, but, at the same time, they are themselves physical systems. The computer simulation of physics can thus be viewed as the representation of one physical system by another. This observation has profound and surprising consequences. To begin to appreciate some of these, let us return to Fig. 1. In the physical world, no exponential rate of increase can go on forever, and the figure points to serious problems that will be faced by the computer industry in the coming decades. At the time of this writing, the physical area required to store 1 bit of information is seen to be on the order of the size of a biological cell. If the exponential rate of decrease were to continue unchecked, this size would be down to atomic scales before the year 2020. Such considerations have fueled new thinking about the relationship of computation to physics, and of possible physical limits to computation. The results of such thinking may ultimately be viewed as the second great contribution of the 20th century to the science of computation.

Though the underpinnings of the theory of computation date back to the 19th century, it was the mathematicians Alan Turing and John von Neumann who, in the 1930s and 1940s, developed the modern theory of computation that guided the first 50 years of electronic computer evolution. Turing provided a mathematical model of the computational process, known as a *Turing machine.* Conceptually, this can be imagined as a linear strip of paper tape, infinite in both directions and imprinted with symbols chosen from some finite set, passing through a machine with some finite number of internal states. At each time step, the machine can read a symbol from the paper tape and, on the basis of that symbol and the internal state of the machine at that moment, take one or more of four actions: (1) change the internal state, (2) print something on the tape, (3) advance the tape in one direction or the other, and (4) halt the machine. Note that the Turing machine can be thought of as a dynamical system with Markovian evolution. In spite of its conceptual simplicity, Turing showed that his "machine" was capable of performing all of the arithmetic operations of number theory, such as addition, subtraction, multiplication, division, and exponentiation, by initializing the paper tape with the proper (finite) "program." It is also possible to make the Turing machine perform more complex algorithmic procedures, such as looping and conditionally jumping.

More interestingly, it is even possible to make any Turing machine simulate any other Turing machine. That is, it is possible to write a program for a Turing machine *A* which accepts as input *both* the rules and input of another Turing machine *B*, and produces the same output that *B* would have produced. Because the Turing machine has this ability, we say that it is capable of *universal computation.* It therefore seemed eminently reasonable to regard the universal Turing machine as a complete mathematical paradigm of exactly what a computer is and does.

To be sure, the Turing machine is not the only possible computational paradigm that can be demonstrated capable of universal computation. One might imagine other schemes, with more tapes, more symbols, more internal state, etc. A particularly interesting computational paradigm, advanced by von Neumann just after World War II, is called a *cellular automaton.* This consists of an infinite Cartesian grid with some (usually small) number of states at each site, which are updated according to a rule that depends on the states of neighboring sites. As demonstrated later, this paradigm is also computationally universal, in that it is capable of simulating a Turing machine; thus, we say that it is *Turing equivalent.* Of course, this raises the question of whether some other digital computational paradigm might be found to be *super-Turing*—that is, capable of performing a computation that a Turing machine cannot perform. [Note that *analog computers,* such as analog recurrent neural nets, are often said to be super-Turing; for the moment, however, we limit our discussion to digital computing paradigms.] Turing's name, along with that of the American logician Alonzo Church, is attached to the hypothesis that the answer to this question is "no." The *Church–Turing hypothesis* states that any computer capable of universal computation may simulate any other computer capable of universal computation. That is, outside of quantitative considerations such as memory and speed, no universal computer is more powerful than another. There is no problem that one can do that another cannot also do, given sufficient time and memory. Thus, computer manufacturers may argue about whose product is faster and larger, but none can claim to perform any computation that their competitors cannot.

In addition to establishing the concept of universal computers and outlining what they could do, Turing addressed the even more interesting question of what they could *not* do. Strongly influenced by the contemporary work of Kurt Gödel who showed that there existed truths of number theory that could not be proven from its axioms, Turing recognized the close analogy between computation and the axiomatic systems of mathematics and demonstrated that there were some functions that could never be calculated by a universal computer. To see this, imagine that we would like to construct a Turing machine to accept a number x as input, compute a function $f(x)$, write it as output, and then come to a halt; such a procedure is called an *algorithm.* Since such programs must consist of a finite number of symbols chosen from a finite set, they can be enumerated; that is, they have the cardinality of the integers. It is therefore possible to create an (infinite) "alphabetized" list of *all* such algorithms. (We agree to discard from this list algorithms that are syntactically incorrect.) The question then reduces to whether or not every function that can be specified has an algorithm appearing on the list. Let us denote by $f_i(x)$ the result of applying the ith algorithm on the list to the numerical input value x; that is, f_i is the function corresponding to the ith algorithm on the list. From this, we can construct the function $g(x) \equiv f_x(x) + 1$. We can now prove that the function $g(x)$ cannot be on the list. If it were, there would be some j for which $f_j(x) = g(x)$ for all x; but this cannot be the case because when $x = j$ we would have $f_j(j) = f_j(j) + 1$, a contradiction. Thus, we have found a function that cannot be computed by any algorithm. This result is all the more odd because it *seems* straightforward to describe an algorithm for computing the function $g(x)$: Simply write a program that, given an input value x, causes the Turing machine to enumerate legal algorithms until it finds the xth such algorithm, apply this algorithm to the number x, output the result, and halt. Turing noted that the problem with this putative procedure is that it assumes that there is an algorithm for determining whether or not a given string of symbols is indeed a legitimate algorithm—that is, one that eventually produces output and halts. The conclusion is that no such "algorithm-verifying algorithm" can possibly exist.

The similarity of this line of reasoning to Gödel's theorem for axiomatic systems in mathematics should be evident. To see this, note that we can reduce any question about number theory to the computation of a mathematical function. For example, by knowing the value of

$$\min_{\substack{n>2,\\ i,j,k>0,}} (i^n + j^n - k^n)^2, \tag{1}$$

we can discern the truth or falsity of Fermat's last theorem (FLT). Why is it not always possible to specify an algorithm for the computation of such a function? After all, to pursue the foregoing example, it seems quite straightforward to write a computer program that will perform an open-ended search for a counterexample of Fermat's last theorem. One could simply loop over all possible i, j, k, and $n > 2$, evaluating expression (1), and stopping if it were ever found to be zero. Clearly, we can restrict the search to $i \geq j$ without loss of generality, so j can be taken from the finite set $\{1, \ldots, i\}$; likewise, it is easy to see that $k \in \{i + 1, \ldots, 2i\}$. This leaves infinite loops over two integer variables, i and n; these can be done by conducting a single open-ended loop over their sum $m \equiv i + n \geq 4$, restricting i to the finite set $\{1, \ldots, m - 3\}$, and taking $n = m - i$. (Of course, much better bounds are available, but these serve to illustrate the idea.) The problem with this grand scheme is that there is no *a priori* way of knowing if this program will ever halt (unless, of course, FLT is proven, as may be the case at the time of this writing). If it does not halt, then it is not a valid algorithm for computing the function in Eq. (1), since it does not provide a result in a finite number of steps. Thus, our result on uncomputable functions is related to the impossibility of solving what is often called the *halting problem*—the problem of making an *a priori* determination of whether or not a given computer program will ever halt.

The impossibility of solving the halting problem has profound implications for the simulation of dynamical systems on a computer and may indeed have profound implications for physics itself. To see this, let us return to von Neumann's idea of computing by means of *cellular automata.* One very heavily studied cellular automaton, developed by the mathematician John Horton Conway, is called the *Game of Life.* It is a dynamical system on a two-dimensional Cartesian grid, with only 1 bit of information (0 or 1) at every site, that is synchronously updated in discrete time steps according to rules that are easy to describe: Consider a site whose bit is b, and let the sum of the bits of that site's eight neighbors (at distances 1 and $\sqrt{2}$) be called S. The site then "updates" its bit to the value b', where

$$b' = \begin{cases} 1 & \text{if } (b,S) \in \{(1,2),(0,3),(1,3)\}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

It is to be emphasized that this rule is to be applied *simultaneously* to all of the sites on the lattice.

If this rule is applied to a lattice with randomized initial conditions, a wide variety of interesting emergent behavior is observed. Figure 2(a) shows configurations that are locally stable; that is, they are invariant under the action of the rule. Figure 2(b) shows configurations that oscillate in time, returning to their initial configurations after a definite period of time. In the language of dynamical systems theory, we might say that these kinds of patterns are *fixed points* and *limit cycles,* respectively. It is also possible to find patterns that are reminiscent of solitons or nonlinear waves; Fig. 2(c) shows examples of patterns that return to their original shape—but translated in space—after some finite period. The first of these translating patterns is often called a "glider." Figure 2(d) then shows a "glider gun"—that is, a pattern that oscillates (with period 24), emitting one glider at every oscillation.

Gliders are somewhat reminiscent of "bits" of information traveling along a wire in a computer. In this analogy, the glider gun plays the role of a "clock," emitting these bits at a set rate. It is also possible to show by construction that one can create the analog of *logic gates* for gliders—for example, a "NOT" gate, which emits gliders at the clock frequency unless another glider is inci-

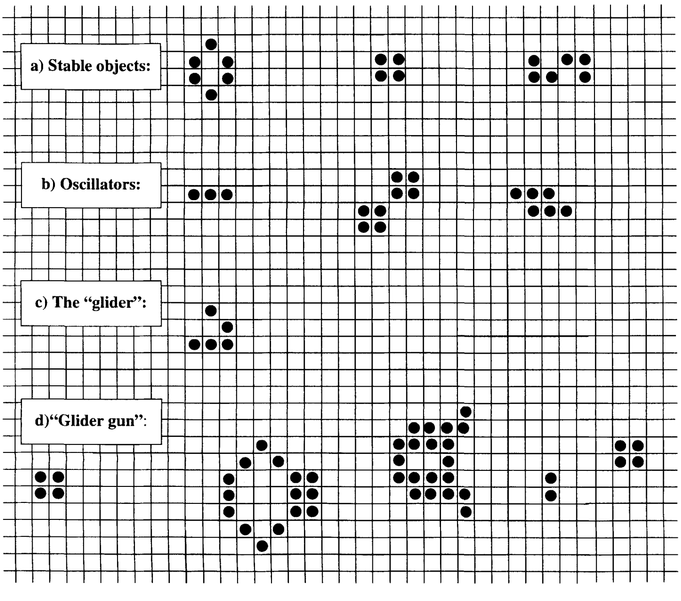

FIG. 2. Coherent structures of the Game of Life: (a) stable patterns, (b) oscillatory patterns, (c) translating patterns, and (d) the "glider gun."

dent, and an "AND" gate, which emits one glider if and only if gliders are present on both of two incoming channels. Since these logic gates are the basic building blocks of a digital computer, one could, in principle, use the patterns formed by the Game of Life to set up a universal computer on the lattice. This exercise may seem academic—indeed, almost frivolous—but it contains a very important lesson: It shows that the Game of Life is *Turing equivalent*—that is, its dynamics are capable of universal computation. It is highly significant that a simple dynamical system, defined by nearest-neighbor rules on a lattice, can function as a universal computer. Among other things, it means that it is *provably impossible* to know the asymptotic behavior of the Game of Life for every initial condition, since this would be tantamount to solving the halting problem. Note that there are many physical models, such as the heat equation, the asymptotic behavior of whose solutions can be determined by elementary methods of analysis. We have just demonstrated that the Game of Life is one dynamical system that *cannot* possibly admit such a simple, general asymptotic treatment.

Now there are many other dynamical systems in physics—such as the driven pendulum—that have thus far resisted a general asymptotic treatment. Could this be because they somehow set up structures that are capable of computation? A partial answer to this question is provided by the work of Moore (1990, 1991), who showed that bundles of trajectories in simple classical dynamical systems, as can be identified on a Poincaré surface of section ("puncture plot" in phase space), can shuffle about in successive time steps in a manner that is Turing equivalent. Moore demonstrated this by construction for a system consisting of an ideal classical particle in three dimensions, with a finite number of planar and parabolic reflectors. The asymptotic state of such a system is therefore unknowable for general initial conditions. Note that this *uncomputability* is a much stronger result than the demonstration of "chaos" for such a system. Certain dynamical systems, such as the logistic map with maximum multiplicative parameter [$x_{n+1} = 4x_n(1 - x_n)$, where $x_n \in [0,1]$], are certainly chaotic, in the sense that they have positive Liapunov exponents, but at the same time admit a simple asymptotic treatment. [The substitution $x_n = \sin(2\pi\theta_n)$ reduces this map to $\theta_{n+1} = (2\theta_n)_{\text{mod}1}$, which has the effect of chopping off the leading digit of the binary representation of the number at each time step. Thus, given the initial condition to infinite precision, one can immediately specify the result after iteration n.] Much less can be said about the time asymptotics of an *uncomputable* dynamical system.

1.3 Physical Computation

These mathematical limitations to computation are undeniably fascinating, but one might well wonder if they are simply academic exercises. After all, the proven existence of mathematical functions that cannot be described algorithmically certainly did not stop people from making good use of computers in the simulation of physical systems. At the same time, however, it should be kept in mind that real-world computers *are* themselves physical systems. Thus, as the power of real computers increased dramatically in the latter half of the 20th century, it was natural for people to begin to wonder if there were any *physical* limitation to their size and speed, in addition to the mathematical limitations studied by Turing. As we shall see, this issue remains an important topic of study to this day, and in addressing it we shall find that many of the issues raised by Turing reappear in a slightly different form.

Turing machines are mathematical idealizations of computers. How fast and how small may we make a computer in the real world? In the drive to make computers more powerful, one of the fundamental engineering constraints is due to the heat released by its components. As these components are packed ever more tightly together in three dimensions, the dissipation of this heat becomes an important problem. This leads to the question: Must computation always release heat? As late as the 1960s, it was widely believed that the answer to this question must be "yes." The reasoning was predicated on an ingenious argument due to Landauer (1961) that erasure of information in a computer—a process that goes on all the time in modern computers—must be accompanied by the dissipation of heat. This argument went as follows: Consider a microscopically reversible physical system that is capable of representing N macroscopically observable bits of information. Since each bit can take on one of two values, the system of N bits can be in any one of 2^N macroscopic states. Corresponding to each of these macroscopic states, let us say that there are Ω_0 microscopic states of the underlying physical structure, so that the full state-space volume can be written as $\Omega = 2^N\Omega_0$. Note that the volume associated with the *computational degrees of freedom* is 2^N, while that associated with the *noncomputational degrees of freedom* is Ω_0, and that these volumes are multiplicative since they correspond to independently specifiable degrees of freedom. Since the entropy of a system is Boltzmann's constant times the natural logarithm of its state-space volume, the entropy associated with our physical system is then $S = k_B \ln \Omega = k_B \ln(2^N\Omega_0) = k_B \ln \Omega_0 + Nk_B \ln 2$. Since we have assumed that the underlying physical system is reversible, this entropy must be constant. Now suppose that our physical system "erases" a bit of information at the macroscopic level; that is, suppose that it forces one of its bits to take on a specific value, regardless of its previous value. This constraint has the effect of contracting the state-space volume of the computational degrees of freedom by a factor of 2, since it can now be in one of only 2^{N-1} states; equivalently, we see that this has the effect of decreasing the entropy of the computational degrees of freedom by $\Delta S = k_B \ln 2$. Since the entropy of the system as a whole must be constant, we see that this lost computational entropy must be dumped into the noncomputational degrees of freedom. If the underlying physical structure corresponding to the noncomputational degrees of freedom is maintained at a temperature T, the increase in its entropy by ΔS corresponds to the appearance of an amount $T\Delta S = k_BT \ln 2$ of dissipated heat energy. Moreover, the full force of the second law of thermodynamics precludes the possibility of this heat energy ever being recovered as useful macroscopic information.

Landauer's argument indicates a close connection between the process of computation and thermodynamics. In particular, it indicates that any computer that erases bits of information must necessarily release heat. The trouble is that all of the familiar logic gates from which modern computers are built do this all the time. AND and OR gates, for example, are irreversible because their input cannot be deduced from their output. This is illustrated in Fig. 3(a). Since 1 bit of information is lost each time either of these gates is used, heat energy $T\Delta S = k_BT \ln 2$ must be dissipated. While this amount is many orders of magnitude smaller than the amount of heat released by the operation of a real computer gate, the result was considered interesting since it seemed to indicate that computation was fundamentally and necessarily irreversible.

In passing we note that this observation was used by Bennett (1987) and others to explain finally and convincingly the "Maxwell's demon" paradox that had plagued kinetic theory for a century. The issue has to do with the apparent ability of a "demon," poised at a small opening in a partition between chambers of high- and low-pressure gas, to violate the second law of thermodynamics by allowing molecules from the low-pressure side to enter the high-pressure side, but not vice versa. Such a demon effectively exploits the thermal energy of the molecules to create a pressure difference that could

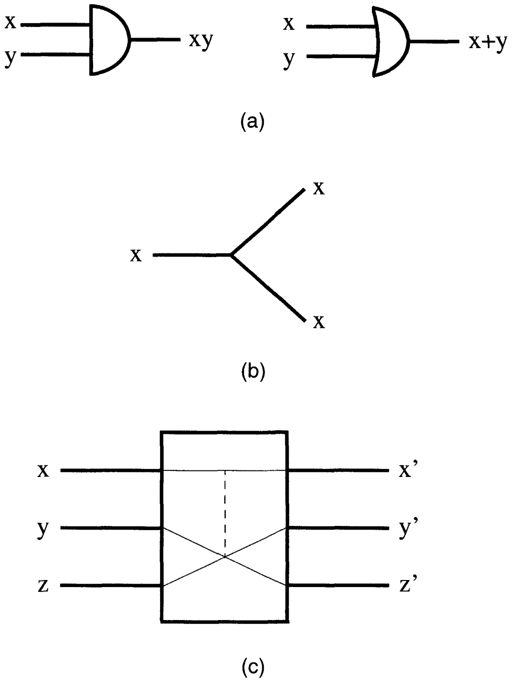

FIG. 3. Irreversible and reversible logic gates: (a) AND and OR gates (arithmetic shown is understood to be Boolean), (b) FANOUT gate, and (c) reversible, universal gate.

then be employed to do useful work—an egregious violation of the second law. A number of explanations of the demon have been put forth over the past century, but the issue was finally resolved only in the 1980s. Bennett observed that such a demon must perform what is essentially a computation in order to decide whether or not to allow a molecule through the partition. Indeed, Bennett was able to demonstrate this most clearly using a variant of the Maxwell demon argument due to Szilard. Requiring that the demon be able to perform computation is not, in and of itself, a fatal flaw, but the fact that the demon must *erase* his computational memory after each molecule is processed, in preparation for the next one, is the real explanation of why he is unable to violate the second law.

Returning to the lost information and concomitant heat generation inherent in computation, we note that one might try to retain the lost bits of information by considering two-input/two-output gates. Even then, however, one must restrict one's attention to those gates for which the input can be reconstructed from the output. It is not difficult to see that there are $2^8 = 256$ gates that accept 2 bits of input and yield 2 bits of output. Of these, $4! = 24$ are reversible, in that the 2-bit output states are a permutation of the 2-bit input states. Unfortunately, these gates are insufficient to construct a universal computer. It should be noted that, in addition to the usual gates necessary for this purpose (e.g., AND and NOT), it is necessary to worry about such things as the "FANOUT" function that is implicit every time one wire splits into two within a circuit, as in Fig. 3(b); though FANOUT is reversible given a legal output state, not all possible output states are legal, and the operation does seem to change the number of bits and hence the entropy of the computational degrees of freedom.

The observation that 2-bit reversible gates are insufficient to construct a universal computer, and the realization of the problematic nature of operations such as FANOUT long taken for granted by computer designers, tended to support the argument that universal computation is necessarily irreversible. In the 1970s, however, it was realized (Bennett, 1973; Fredkin and Toffoli, 1982) that this argument is specious. Figure 3(c) shows a three-input/three-output gate, known as the *Fredkin gate*, that is fully reversible and quite sufficient to construct a universal computer. It has three inputs (b_1, b_2, and b_3) and three outputs (b'_1, b'_2, and b'_3). If the first input is zero, then the other two outputs are exchanged; otherwise the outputs are the same as the inputs. The truth table for this gate is thus

b_1	b_2	b_3	b'_1	b'_2	b'_3
0	0	0	0	0	0
0	0	1	0	1	0
0	1	0	0	0	1
0	1	1	0	1	1
1	0	0	1	0	0
1	0	1	1	0	1
1	1	0	1	1	0
1	1	1	1	1	1

Note that if $b_3 = 0$, then $b'_2 = b_1 \times b_2$. (Here we identify 0 with *false* and 1 with *true*, as is customary.) This means that the gate captures all the functionality of the AND gate. If we fix $b_2 = 1$ and $b_3 = 0$, then $b'_3 = \sim b_1$ giving us the functionality of the NOT gate, and $b'_2 = b_1$ giving us FANOUT. These functions are sufficient to create a universal computer.

One can therefore imagine a universal computer that accepts input bits on the left, passes them through circuitry consisting entirely of Fredkin gates, and yields output on the right. There is, however, one remaining problem: In addition to the desired output, numerous "garbage" bits will have accumulated in the course of the calculation. For example, as we have seen, it is possible to arrange for the Fredkin gate to output the AND of two of its input bits, but it also produces other output bits when it does so. That other output, inessential for our calculation, cannot be simply thrown away because that would generate heat and violate reversibility. It seems possible that we have just postponed the problem of heat generation to the stage at which we read the computer's output and reset it for a new computation. As soon as we read the portion of the output that we want and unceremoniously erase the rest of it, heat will be generated in proportion to the number of garbage bits destroyed. It is therefore necessary to bound the number of garbage bits created.

Thanks to an ingenious argument of Bennett (Bennett, 1982), this is easy to do and necessitates only a factor of 2 or so more circuitry. Suppose that we set up a reversible circuit of Fredkin gates as above, with input on the left and both output and garbage bits on the right. Since the Fredkin gate is its own inverse, it is possible to recover the input from the output using a mirror image of the very same reversible circuit. Adding this on the right, we recover the original input bits that can then be channeled back to the left-hand side of the circuit and reused. At this point the circuit is fully reversible, and all of the garbage bits have been recycled. Recall, however, that the desired output resides somewhere in the middle of all this circuitry. By extracting it with FANOUT operations, we arrive at a perfectly reversible circuit that produces no garbage bits.

This discovery was quickly followed by that of a reversible cellular automaton that is capable of universal computation (Margolus, 1984), and by Fredkin's observation (Benioff, 1982) that a system of ideal billiard balls and reflectors could perform reversible logic. Figure 4(a) shows the paths taken by two billiard balls with perpendicular trajectories in the event that both balls are present and that either ball is absent. If the presence of a ball is regarded as a "bit" of information, examination of the final trajectories allows us to perform the AND operation on these bits. Given a steady supply of balls, emerging at the clock rate, the NOT operation is also easy to arrange: Simply collide the input balls with the stream balls in such a way that the former knock the latter (and themselves) out of the stream. The stream balls that do get through will then be the NOT of the incoming signal. Once again, AND and NOT are sufficient to attain universal computation.

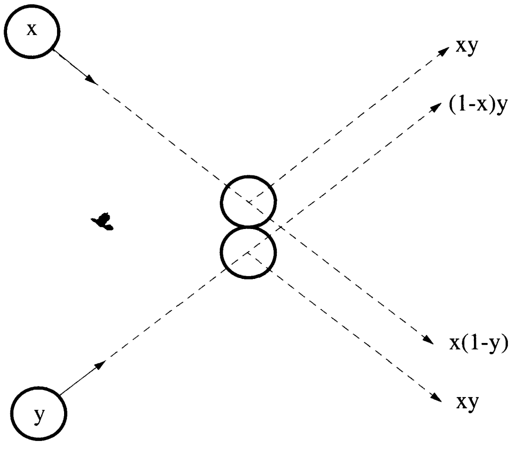

FIG. 4. Billiard-ball model of computation. The presence or absence of the two balls encodes the bits *x* and *y*, respectively. Note that the outgoing, post-collision trajectories yield the logical functions *xy*, *x*(1 − *y*), and (1 − *x*)*y* (arithmetic is understood to be Boolean).

One critique of billiard-ball computational paradigms is that they are inordinately error prone. The slightest deviation of any of the balls from their intended trajectories will tend to propagate and amplify in time according to the Liapunov exponent of the dynamics. It therefore seems that reversible computation could not be achieved experimentally. It turns out to be possible, however, to implement a simple and effective *error correction* scheme. Indeed, note that the NOT gate described above does precisely this. The output ball of the NOT gate is likely to be much more accurately aimed than the input, since the output balls have just emerged from their source. Thus, by passing a signal through two NOT gates, we can remove any errors that may have accumulated in their trajectories.

Thus, bits of information do not have to be destroyed in order to perform universal computation, and heat dissipation is not an unavoidable side effect of computation. Moreover, reversible computation can be achieved reliably, with error correction. The notion that computation can be reversible and robust is central to the idea of quantum computation, which is discussed later on in this article.

2. LANGUAGES FOR SCIENTIFIC COMPUTATION

We now descend abruptly from the lofty heights of ideal computational models to the ignominious realities of performing useful computation on contemporary computers. As noted above, the realization that computers could not compute everything did not stop people from trying to compute as many things as possible. In order for computers to become a practical tool for scientific computation, however, it was necessary that they become easy to program. At the very lowest level, computers accept instructions in *machine language,* which is really nothing more than a sequence of elemental instructions in the binary format expected by the processor. In fact, the very first computer programs input this binary data by setting hardware switches on the computer's front panel. This approach is obviously inappropriate for all but the smallest sets of instructions.

At the next level, *assembly language* uses mnemonic abbreviations for certain of these elemental operations and groups thereof. In the late 1950s, researchers at IBM developed the first *high-level language,* designed to simplify the programming of a computer for solving scientific problems. A high-level language typically employs a program called a *compiler* (or perhaps an *interpreter*) to translate instructions written according to a carefully defined language syntax into assembly language. This first high-level language was called *Fortran* (an abbreviation of "FORmula TRANslation"), and it quickly came to be the predominant high-level language for scientific computation—a status that it retains to this day.

Since its invention, Fortran has had three principal stages of evolution: In the early 1960s many divergent varieties of the language existed. To combat this tendency, computer users and vendors cooperated on the development and publication of a language standard. The result, called Fortran 66, was the first standardized computer language (in that programs written in this language could be compiled and executed on different computer platforms), and this stability contributed to its popularity among scientific programmers. Though powerful, this language was very undisciplined. It allowed for control to jump about from one part of the code to another in arbitrary ways, by means of a GOTO statement, leading some programmers to refer to it facetiously as "spaghetti code." Unless great care was taken in its use, it could be very difficult to read and understand an existing Fortran 66 code. For this reason, programs of more than 10 000 lines of instructions were very difficult to write, debug, maintain, and modify.

The desire to impose some structure on the language led to the next major leap in computer programming, called *structured programming,* in the late 1960s and early 1970s. The idea was that control in a computer code should have a definite and more or less unidirectional flow to it; the GOTO statement should be avoided. Frequently reused groups of instructions should be broken up into succinct and easily invoked *subroutines* or *function subprograms.* These subprograms should have the ability to declare and use *local variables,* and they should be capable of calling themselves in a recursive fashion. A number of new languages, including Pascal, Algol, PL1, and C, were invented to support this new paradigm. Eventually, however, most of these ideas found their way into the next major Fortran language standard, called *Fortran 77.* This was adopted by the International Standards Organization in 1980, and remains the dominant dialect of the language in use to this day. Using Fortran 77 or C, programmers routinely write and maintain computer codes with tens of thousands of lines. As larger and more complex software was developed, however, it soon became clear that codes larger than about 100 000 lines would be very difficult to develop and maintain.

At the time of this writing, the newest and most fashionable structuring paradigm for computer software is called *object-oriented programming.* It is defined by the following principal characteristics:

- **Encapsulation, persistency of state, and data hiding:** The notion of subprograms should be strengthened to allow the creation of "black box" software with local data that is completely protected from outside interference, and may be persistent in that it does not disappear when the subprogram completes execution. This allows for the development and maintainance of software libraries in complete isolation from the application programs that utilize them. This should be contrasted with more conventional software libraries, which often require the internal state to be passed back and forth to the calling routine.
- **Data abstraction:** In structured programming languages, users could specify if a given variable represented, e.g., a logical value, an integer, a real number, or a complex number. In object-oriented programming, the user should have the ability to create his/her own data structures and representations out of more fundamental

ones and to declare variables in them. For example, one might imagine creating a data structure for quaternions, representing them by four real numbers; a user could then declare variables of quaternionic type.

- **Polymorphism:** Since object-oriented computer languages allow for many new and varied data types, one might well worry about the management of the plethora of new operations necessary to manipulate them. Of course, mathematicians use the same symbol "+" to denote addition of integers, reals, complex numbers, and quaternions. Likewise, object-oriented languages allow for *operator overloading* so that the same operation may be applied to variables of different type. The compiler is able to determine the type of the arguments and arrange for the appropriate operations to be performed when the code is executed.
- **Inheritance and extensibility:** Object-oriented programs should be modular and should allow for data to acquire or inherit certain properties of other data. For example, if a program has data structures for cats, mammals, and animals, it ought to be possible for objects of type cat to inherit all properties of objects of type mammal, and for these to in turn inherit all properties of objects of type animal. This greatly facilitates hierarchical code development.

At the time of this writing, the most popular object-oriented computer languages are C++ (an extension of the C language) and Java. Nevertheless, Fortran has incorporated many of these ideas in its latest standard, called *Fortran 90*. Other languages developed for scientific programming, such as Ada, allow for some, but not all, object-oriented features. Most of the software examples used in the remainder of this article will be presented in Fortran 90, using an object-oriented style.

3. DATA STRUCTURES AND ALGORITHMS

3.1 Algorithms and Computational Complexity

The concept of a computer *algorithm* was described earlier, in reference to Turing machines. While modern algorithm development has become a science unto itself, the basic idea remains much the same: An algorithm is a sequence of basic instructions, or "recipe," for solving a given problem on a computer. Often the same problem will allow for several different algorithms, and it becomes important to be able to compare them to see which is better. One way that this comparison can be achieved is to compare the algorithms' *computational complexity*. This is simply a count of the number of elemental operations required to solve the problem; it can be more mathematically defined as the number of steps required for a standardized Turing machine to implement the algorithm.

Consider the procedure known as *Gaussian elimination* for the solution for x of

$$Ax = b, \tag{3}$$

where A is a dense $N \times N$ matrix, and x and b are N-vectors. Assuming that nothing is singular, we begin by dividing each element of the first row by the first element in that row (a total of N divisions, excluding that of the first matrix element since we know that will yield 1, but including that of the first element of b). We then multiply this row by the first element of each of the remaining rows, and subtract it from those rows, so that only the first row has a nonzero entry in the first column, and that entry is 1 (a total of $(N - 1)(N - 1)$ multiplications and subtractions for the nonzero matrix elements). To get to this point, we have performed a total of $2N^2 - 3N + 2$ operations (subtractions, multiplications, or divisions). We can now repeat this process on the lower right-hand $n \times n$ submatrices, where n ranges from $N - 1$ down to 1. At that point, we will have performed a total of

$$\sum_{n=1}^{N} (2n^2 - 3n + 2) = \frac{N}{6}(4N^2 - 3N + 5) \tag{4}$$

operations, and the problem will be solved. We say that the computational complexity of this algorithm is $O(N^3)$; that is, if the number of unknowns is increased by a factor of z, the number of operations required to solve the problem will increase by z^3 in the limit of large problem size.

Oftentimes, clever algorithmic design can drastically reduce the computational complexity required to solve a given problem. For example, the *optimal wedding problem* is well known in combinatorics: Given N men and N women who state quantitative preferences for each other, say as a number from 1 to 10, how can they be paired so as to maximize the sum of the preferences? At first glance, it would seem that all possible pairings of men and women must be considered in turn, leading to an algorithm with complexity $O(N!)$. In fact, an algorithm with $O(N^3)$ complexity is known for solving this problem (Khuller *et al.*, 1991).

Perhaps more pertinent to scientific computation is the example of the discrete Fourier transform of N data elements, f_j, where $j \in \{0, \ldots, N-1\}$. Since the definition of this transform,

$$\tilde{F}_k \equiv \sum_{j=0}^{N-1} e^{2\pi ijk/N} f_j, \tag{5}$$

involves a sum over N elements to obtain each of N results, it would appear that the transform requires a minimum of $O(N^2)$ elemental operations to compute. In 1965, however, Cooley and Tukey (1965) showed that it is possible to develop an algorithm, called the *fast Fourier transform*, with computational complexity of order $O(N \log N)$.

3.2 Parallel Computers

Since computers perform elemental operations at some rate fixed by an internal "clock," it follows that a single processor will perform a calculation in a time proportional to that calculation's computational complexity. A reduction of computational complexity therefore translates directly into faster execution time. Another way to decrease execution time is to apply more than one processor to a problem; this is called *parallel computing.*

Since the 1980s, parallel computers, containing anywhere from a few to thousands of processors, have been extensively applied to problems of computational science. There are several reasons for this trend. First, as has been noted, *very large-scale integration* (VLSI) of electronic components is approaching some fundamental physical limitations of size and speed (see Fig. 1). By the mid-1980s, the fastest processors had clock rates measured in nanoseconds. It was therefore natural that people would begin to investigate increasing computational power through the cooperative use of many such processors as an alternative to making them faster. Second, as processors and memory components became faster, and as memories became much larger, people began to worry that the classic computer architecture, in which a single processor reads a few words of data from memory and then writes a result back to memory, was becoming a bottleneck to speed. Since only those words of memory actually accessed by the processor were in use at any one moment, most of memory remains idle. Such an architecture was first advocated by von Neumann (not to be confused with his advocacy of cellular automata as a computational paradigm) and is therefore called a *von Neumann architecture.* This problem is therefore sometimes called the *von Neumann bottleneck.* Third, as processors became cheaper in the 1980s, so that processor and memory prices (per unit area of silicon) became comparable, it made sense to try to combat the von Neumann bottleneck by spreading more processing power out to the memory, so to speak.

In spite of these rather compelling arguments, parallel computers took some time to find applications to computational science because of an initial fear that they would be difficult to program. When it turned out that the regular structure of many scientific programs was quite amenable to parallelization and that most major parallel algorithmic paradigms could be easily captured in high-level languages, parallel computing exploded in popularity. At the time of this writing, virtually all major high-performance computer vendors offer one or more parallel computer platforms.

Early scenarios imagined *massively parallel* or *data parallel* computers, in which the number of processors available was comparable to the amount of data in a problem (Hillis and Boghosian, 1993). While computers with tens of thousands of processors have been built (Hillis, 1985), the current market seems to favor a coarser, more modest, "granularity" in which anywhere from a few to a few hundred processors, with sophistication comparable to that of modern workstations, can communicate as if they

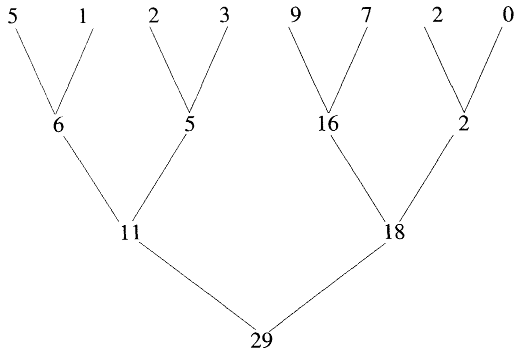

FIG. 5. Tree algorithm for parallelizing summation. Numbers are added together in a "tree" data structure, so that eight of them can be combined in only three parallelizable stages.

are connected to a very fast *local-area network*. Nevertheless, if we imagine spreading an algorithm with N data elements over M processors, it is interesting to try to determine the asymptotic behavior of the time required to execute the algorithm as a function of N and M; of course, as $M \rightarrow 1$ this result must approach the computational complexity of the algorithm. To demonstrate this idea, we can consider a few simple examples.

Suppose that we have an *array* of N numerical data elements, D_i, where $i = 1, \ldots, N$, and we wish to find its sum, $\Sigma_{i=1}^{N} D_i$. Clearly, this has computational complexity of order N, since $N - 1$ additions are required. We could, however, distribute N/M data elements in each of M processors. These processors would then require a time of order N/M to sum their elements, but they could all do this simultaneously. Once they each had their subtotal, the M data elements would need to be combined into one total, but this could be carried out in the "tree" structure shown in Fig. 5, in which all the operations at each level of the tree are performed in parallel. For a binary tree, the combination requires only $\log_2 M$ parallel steps to complete, and the time required to execute the algorithm scales like

$$T(M,N) \sim c_1 \frac{N}{M} + c_2 \log_2 M + c_3, \tag{6}$$

where c_1, c_2, and c_3 are architecture-dependent constants. Specifically, c_1 is the time required to add two numbers in the same processor, c_2 is the time required to communicate a number from one processor to another, and c_3 is whatever constant "overhead" or *latency* time is required to orchestrate the calculation. In Fig. 6 we plot the *speedup* due to parallelism, defined as

$$S(M,N) \equiv \frac{T(1,N)}{T(M,N)}, \tag{7}$$

as a function of the number of processors M, for $N = 100$ and $c_2/c_1 = c_3/c_1 = 5$. Note that this initially increases with M, due to the parallelism, but quickly reaches a point of diminishing returns, as the latency and interprocessor communication begin to dominate the performance. It is not difficult to see that a maximum speedup of

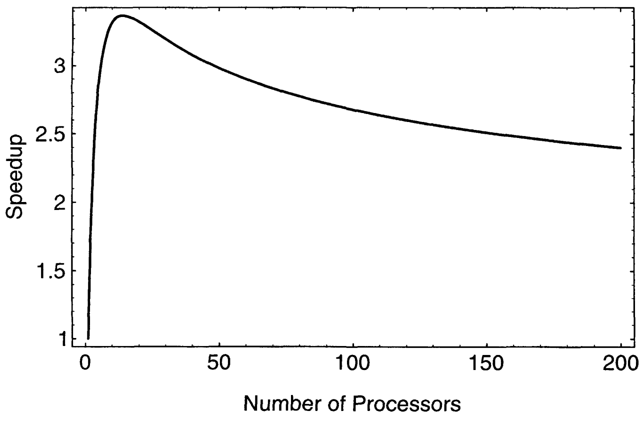

FIG. 6. Speedup versus number of processors, illustrating Amdahl's law, for $N = 100$ and $c_2/c_1 = c_3/c_1 = 5$.

$$S_{\max}(N) = \frac{c_1N + c_3}{c_2 \log_2(2c_1N/c_2) + c_3} \tag{8}$$

is obtained for M equal to

$$M_{\max} \equiv c_1N/c_2, \tag{9}$$

which is 20 for our example. For $M > M_{\max}$ the speedup actually begins to *decrease* with increasing M! This rather pessimistic assessment of the benefits of parallelism is a special case of what is known as *Amdahl's law*. This is the observation that, as the parallelism reduces the time required for the parallel portion of the computation toward zero, the latency (and interprocessor communication) will always end up dominating the speedup, which therefore must have an upper bound.

Though Amdahl's law seems compelling, in recent years researchers have very successfully applied parallel computers with thousands of processors to scientific problems. This seems like a contradiction and has been the source of much confusion, but the reason is simple: Amdahl's law considers the application of parallel computers to problems of a fixed size N. As problems become ever larger, however, note that both the maximum speedup, given by Eq. (8), and the number of processors for which it is obtained, given by Eq. (9), can become arbitrarily large. Thus, parallel computers are naturally most useful for the largest and most data-intensive problems of computational science. As we shall see, there is no shortage of such problems, and so parallel computers are here to stay.

Before leaving the topic of parallelism and the scaling of execution time with the number of processors, we note that different parallel data structures can result in quite different time scalings. To illustrate this, we consider the problem of multiplying a pair of dense $N \times N$ matrices, A and B, to obtain their product C. Since

$$C_{i,j} = \sum_{k=1}^{N} A_{i,k}B_{k,j} \tag{10}$$

are N^2 quantities, each of which requires N multiplications and additions to compute, it is clear that the problem has $O(N^3)$ computational complexity and thus requires a time proportional to N^3 to complete on a single processor. By contrast, on a parallel computer of N^2 processors, this problem can be done in $O(N)$ time. The algorithm used is the systolic or Cannon algorithm (Cannon, 1969), and it is naturally parallel. Both operands, call them A and B, are stored as parallel variables. We begin by cyclically shifting the ith row of A by $i - 1$ entries for all $i \in \{1, \ldots, N\}$, and the jth column of B by $j - 1$ entries for all $j \in \{1, \ldots, N\}$. This puts $A_{i,j+i-1}$ and $B_{j+i-1,j}$ in array element (i,j) (where all subscripts are taken modulo N). We then multiply these and accumulate into the product parallel variable $C_{i,j}$. These N^2 multiplications and accumulations can be done in fixed time, since we have assumed that our parallel computer has $O(N^2)$ processors. Continuing, we cyclically shift all the rows of A and all the columns of B by one, and then multiply and accumulate into C. We repeat this step a total of N times. After the kth such step, array element (i,j) contains $A_{i,k+j+i-1}$ and $B_{k+j+i-1,j}$, so that the N multiplications and accumulations into array element (i,j) are equivalent to a sum over the dummy index. Thus, after N steps, the result for $C_{i,j}$ is correct and in place. The algorithm takes $O(N)$ time because N steps are required. The Fortran 90 implementation of the Cannon algorithm is

```
SUBROUTINE CANNON(A, B, C)

 INTEGER :: i, n
 REAL, DIMENSION(:, :) :: A, B, C

 n = SIZE(A, 1)

 ! Initial skewing of operands:
 A = CSHIFT(A, DIM = 2, SHIFT =
   (/(i, i = 0, n-1)/))
 B = CSHIFT(B, DIM = 1, SHIFT =
   (/(i, i = 0, n-1)/))

 ! Multiplication of matrices:
 C = 0.0
 DO i = 1, n
  C = C + A*B
  A = CSHIFT(A, DIM = 2, SHIFT
  = 1)
  B = CSHIFT(B, DIM = 1, SHIFT
  = 1)
 END DO
 ! Restore operands:
  A = CSHIFT(A, DIM = 2, SHIFT =
    -(/(i, i = 0, n-1)/))
  B = CSHIFT(B, DIM = 1, SHIFT =
    -(/(i, i = 0, n-1)/))

END SUBROUTINE CANNON
```

Note that the `CSHIFT` command performs the cyclic shifts of either rows or columns; its first argument is the array to be shifted, and the other two keyword arguments are the dimension in which to shift it, and the amount of the shift. Also, we have added a step to restore the operands to their input values.

To take this one step further, suppose that we have a parallel computer with $O(N^3)$ processors. We can then allocate arrays `D`, `E`, and `F` of size $N \times N \times N$, and *spread* array `A` over the third dimension of array `D` so that $A_{i,k}$ is stored in $D_{i,k,j}$ for all j, and *spread* array `B` over the first dimension of array `E` so that $B_{k,j}$ is stored in $E_{i,k,j}$ for all i. We can then perform all N^3 multiplications of $D_{i,k,j} = A_{i,k}$ by $E_{i,k,j} = B_{k,j}$ in fixed time, since we have assumed a parallel computer with $O(N^3)$ processors. The answer for array `C` is then obtained by summing array $F_{i,k,j} = D_{i,k,j}E_{i,k,j}$ over its second (dummy) dimension. The Fortran 90 implementation of this algorithm is as follows:

```
SUBROUTINE FAST_MATMUL(A, B, C)

  INTEGER :: n
  REAL, DIMENSION(:, :) :: A, B, C
  REAL, DIMENSION(:, :, :),
    ALLOCATABLE :: D, E

  n = SIZE(A, 1)
  ALLOCATE(D(n, n, n), E(n, n, n))

  D = SPREAD(SOURCE = A, DIM = 3,
    NCOPIES = n)
  E = SPREAD(SOURCE = B, DIM = 1,
    NCOPIES = n)
  C = SUM(D*E, DIM = 2)
  DEALLOCATE(D, E)

END SUBROUTINE FAST_MATMUL
```

Note that there was no need actually to declare the array `F`. Note also that the most expensive operations were the *spreads* of array elements over a new array dimension, and the summation or *reduction* of array elements over an existing array dimension. Both of these operations, however, can be carried out in time of $O(\log_2 N)$, by the same arguments that we used to parallelize summation, illustrated in Fig. 5. Thus, on $O(N^3)$ processors, the multiplication of two $N \times N$ matrices takes $O(\log_2 N)$ time.

3.3 The Computer Representation of Real-Valued Data

Most physical models are formulated in terms of quantities that can take on a continuous range of values. Many physical models are described in terms of *fields* (either classical or quantum), for which the number of degrees of freedom itself is infinite. By contrast, digital computers can manipulate only a finite amount of finite-precision data. Thus, one of the principal challenges of computational physics is to reformulate physical laws in terms of finite and discrete mathematical models.

Real-valued quantities can be represented as finite-precision *floating-point numbers* in most all high-level languages available today. There is a standard for the storage of such numbers, adopted by the Institute of Electrical and Electronics Engineers (IEEE), whereby *single-precision* floating-point data is represented as

$$\sigma(1 + m)2^e, \qquad (11)$$

where the *mantissa* $m \in [0,1)$ is represented by 23 bits of data, the exponent $e \in \{-127, \ldots, 128\}$ is represented by 8 bits of data, and the sign of the number $\sigma \in \{-1, +1\}$ is represented by 1 bit, for a total of 32 bits. *Double-precision* floating-point numbers are also available in which the mantissa is represented by 52 bits, the exponent by 11 bits, and the sign by 1 bit, for a total of 64 bits. Thus, both single- and double-precision floating-point data take their values in a finite range, and the computer user must be ever vigilant not to *overflow* or *underflow* this range. There are many subtleties involved in this truncation; for some nice examples, see Kahan (1996).

Should we expect the truncation of real-valued data into finite-precision floating-point numbers to cause problems in physical simulations? Obviously, this finiteness circumvents the *computability* problem noticed by Turing and discussed earlier. Recall that Turing machines are idealized models of computers with infinite storage capacity. Real computers have finite storage capacity, and all finite strings of bits are computable, so that any number represented on a real computer is finite precision and computable. Thus, the issue of computability would seem to be an esoteric problem, wrapped up in infinities, that plagues idealized models of computing but not real computers. This sidestepping of the issue is satisfactory so long as the real world can be described in terms of computable numbers alone. Interestingly, Pour-El and Richards (1982) have proven that even simple linear physical models, such as the wave equation, when given computable initial conditions and evolved for a computable length of time, can give rise to uncomputable values of the dependent variable. Should this be a cause for concern? The real answer is that almost nobody worries about it. Nevertheless, we shall soon discuss some situations in which truncation of continuum data plays a very detrimental role in physical simulations, and so this issue is important at least to keep in mind.

4. GRID-BASED ALGORITHMS

4.1 Finite-Difference Methods

The most obvious way to discretize classical fields is simply to store their values on a fine spatial grid and to update those values in discrete time steps. Historically, this was the first approach adopted, and it is still in widespread use today. In this subsection, we consider this *finite-difference* method as applied to the heat equation in order to illustrate the various tradeoffs involved. We shall see that there are three main considerations when discretizing partial differential equations on grids: First, there is the *consistency* of the difference approximation—the question of whether or not it is really an approximation to the desired partial differential equation in some well-posed limit. Second, there is the *accuracy* of the difference approximation to the continuum equation—the question of how large an error may be expected for a given discretization. Third, there is the *stability* of the difference approximation—the question of whether or not its error remains bounded as the algorithm is applied.

Consider the heat equation,

$$\frac{\partial u(\mathbf{x},t)}{\partial t} = D\nabla^2 u(\mathbf{x},t) \tag{12}$$

in a d-dimensional box of side L, so that $\mathbf{x} \in [0,L]^d$, with periodic boundary conditions. We represent the spatial domain by a regular Cartesian grid with N points in each dimension, with points $\mathbf{x_j} = \mathbf{j}\Delta x$, where $\Delta x \equiv L/N$, indexed by $\mathbf{j} \in \{0, \ldots, N - 1\}^d$. Likewise, we consider the system at discrete times $t_k = k\Delta t$, where $k \in \{0,1,2,\ldots\}$ and Δt is the time step. We denote the dependent variable at the gridpoint indexed by j and time step k by $u_{\mathbf{j},k} \equiv u(\mathbf{x_j},t_k)$.

We can use the *centered* finite-difference approximation of the Laplacian operator,

$$\nabla^2 u(\mathbf{x_j},t_k) \approx \sum_{l=1}^{d} \frac{(u_{\mathbf{j}-\hat{e}_l,k} - 2u_{\mathbf{j},k} + u_{\mathbf{j}+\hat{e}_l,k})}{(\Delta x)^2}, \tag{13}$$

where $\hat{e}_l$ is the unit vector in dimension l. As the spatial grid spacing Δx goes to zero, this approximation is readily seen to become exact; hence the difference approximation is *consistent.* In fact, by Taylor expanding the right-hand side, we see that it is $\nabla^2 + O((\Delta x)^2)$, demonstrating both consistency and accuracy to $O((\Delta x)^2)$.

Likewise, a *two-point* finite-difference approximation of the time derivative is

$$\frac{\partial u(\mathbf{x_j},t_k)}{\partial t} \approx \frac{u_{\mathbf{j},k+1} - u_{\mathbf{j},k}}{\Delta t}. \tag{14}$$

This becomes exact as the time step Δt goes to zero. By Taylor expanding $u_{\mathbf{j},k+1}$ about $t = t_k$, we see that this approximation is accurate to $O(\Delta t)$.

The full finite-difference approximation to the heat equation is then

$$\frac{u_{\mathbf{j},k+1} - u_{\mathbf{j},k}}{\Delta t} = D \sum_{l=1}^{d} \frac{(u_{\mathbf{j}-\hat{e}_l,k} - 2u_{\mathbf{j},k} + u_{\mathbf{j}+\hat{e}_l,k})}{(\Delta x)^2}. \tag{15}$$

This is called an *explicit* finite-difference scheme, since it may be explicitly solved for the new dependent variable at each point on the grid $u_{\mathbf{j},k+1}$,

$$u_{\mathbf{j},k+1} = u_{\mathbf{j},k} + \alpha \sum_{l=1}^{d} (u_{\mathbf{j}-\hat{e}_l,k} - 2u_{\mathbf{j},k} + u_{\mathbf{j}+\hat{e}_l,k}), \tag{16}$$

where $\alpha \equiv D\Delta t/(\Delta x)^2$. By iterating this equation, beginning with some specified initial conditions on the grid, we can evolve the difference approximation forward in time to obtain an approximation to the solution of the heat equation. We have demonstrated that this approximation is consistent and accurate to $O((\Delta x)^2)$ and $O(\Delta t)$. We must next assess the *stability* properties of this algorithm. Since all arithmetic done on a computer is of finite precision, a certain amount of error will naturally be introduced into these calculations. If this error tends to grow exponentially in time, the algorithm will be unstable. We can determine this by applying a classical linear stability argument to the difference equation. Let

$$u_{\mathbf{j},k} = U_{\mathbf{j},k} + \xi_{\mathbf{j},k}, \tag{17}$$

where U is the exact solution to the difference equation and ξ is error. Since the difference equation is linear and homogeneous, it is clear that ξ will satisfy it,

$$\xi_{\mathbf{j},k+1} = \xi_{\mathbf{j},k} + \alpha \sum_{l=1}^{d} (\xi_{\mathbf{j}-\hat{e}_l,k} - 2\xi_{\mathbf{j},k} + \xi_{\mathbf{j}+\hat{e}_l,k}), \tag{18}$$

and that we can solve it by Fourier analysis. Letting

$$\tilde{\xi}_{\mathbf{m},k} = \sum_{\mathbf{j}} \xi_{\mathbf{j},k} \exp\left(-\frac{2\pi i \mathbf{j}\cdot\mathbf{m}}{N}\right), \tag{19}$$

so that

$$\xi_{\mathbf{j},k} = \frac{1}{N}\sum_{\mathbf{m}} \tilde{\xi}_{\mathbf{m},k} \exp\left(+\frac{2\pi i \mathbf{m}\cdot\mathbf{j}}{N}\right), \tag{20}$$

we substitute this into our difference equation to find the evolution equation of the **m**th mode,

$$\tilde{\xi}_{\mathbf{m},k+1} = \left[1 - 2d\alpha + 2\alpha \sum_{l=1}^{d} \cos\left(\frac{2\pi m_l}{N}\right)\right]\tilde{\xi}_{\mathbf{m},k}. \tag{21}$$

Clearly, this iteration will be stable for

$$-1 \leq 1 - 2d\alpha + 2\alpha \sum_{l=1}^{d} \cos\left(\frac{2\pi m_l}{N}\right) \leq +1. \tag{22}$$

The right inequality is easily seen to always be satisfied; the left requirement is most stringent for the $m_l = N/2$ mode, for which it demands that $2d\alpha \leq 1$, or

$$\Delta t \leq \frac{(\Delta x)^2}{2dD}. \tag{23}$$

This is called the *Courant condition* for the explicit finite-difference approximation to the heat equation. It states that the time step must scale as the square of the spatial grid size. This is a stringent requirement. It means that if one halves the grid spacing and doubles the number of grid points, presumably in an effort to increase accuracy, one must also decrease the time step by a factor of 4 and iterate for four times as many time steps.

The scaling of Δt with the square of Δx is characteristic of parabolic differential equations. As another physically relevant example, we consider the Navier–Stokes equations for a three-dimensional, viscous, incompressible fluid,

$$\nabla\cdot\mathbf{v}(\mathbf{x},t) = 0, \tag{24}$$

$$\frac{\partial\mathbf{v}(\mathbf{x},t)}{\partial t} + \mathbf{v}(\mathbf{x},t)\cdot\nabla\mathbf{v}(\mathbf{x},t) = -\nabla P(\mathbf{x},t) + \frac{1}{Re}\nabla^2\mathbf{v}(\mathbf{x},t), \tag{25}$$

where Re is the Reynolds number. Though a rigorous stability analysis for a nonlinear equation is generally much more difficult, since the equation is parabolic we might expect a Courant condition similar to that of the heat equation. Since the inverse of the Reynolds number plays the role of the diffusivity, we have

$$\Delta t \sim Re(\Delta x)^2, \tag{26}$$

so that the number of time steps needed to integrate one unit of physical time is

$$N_t \sim \frac{1}{\Delta t} \sim \frac{1}{Re(\Delta x)^2}, \tag{27}$$

and the number of grid points in one unit of physical spatial distance is

$$N_x \sim \frac{1}{\Delta x}. \tag{28}$$

The computational complexity of updating the entire three-dimensional lattice one unit of physical time is then

$$C \sim N_t N_x^3 \sim \frac{1}{Re(\Delta x)^5}. \tag{29}$$

We must then decide how large Δx needs to be according to accuracy considerations. Since the length scale of the smallest eddies in three-dimensional Navier–Stokes flow scale as $Re^{-3/4}$ (see Landau and Lifschitz, 1982 for details), to resolve these we must take $\Delta x \sim Re^{-3/4}$ so that

$$C \sim Re^{11/4}. \tag{30}$$

Thus, computational complexity scales very strongly with Reynolds number. At the time of this writing, the very highest Reynolds numbers that have been simulated are in the vicinity of 10^4 or 10^5. To increase this by an order of magnitude will require nearly three orders of magnitude more computing power. Since Reynolds numbers of interest to, for example, aircraft manufacturers are in the vicinity of 10^7, direct simulations of the Navier–Stokes equations are not an option anytime in the foreseeable future; instead, various approximations are used to avoid having to track the smallest eddies.

It is clear that there is much incentive to try to circumvent the Courant condition. One way of doing this is to use *implicit* finite-differencing schemes, as opposed to the explicit method just described. To pursue our foregoing example, suppose that we replace the finite difference in Eq. (16) by the following, in which the right-hand side is averaged over its values at t_k and t_{k+1}; one may think of this as an estimate of the rate of change at $t_{k+1/2}$. We get

$$\begin{aligned} u_{\mathbf{j},k+1} = u_{\mathbf{j},k} &+ \frac{\alpha}{2} \sum_{l=1}^{d} [(u_{\mathbf{j}-\hat{e}_l,k} - 2u_{\mathbf{j},k} + u_{\mathbf{j}+\hat{e}_l,k}) \\ &+ (u_{\mathbf{j}-\hat{e}_l,k+1} - 2u_{\mathbf{j},k+1} + u_{\mathbf{j}+\hat{e}_l,k+1})]. \end{aligned} \tag{31}$$

This is called the Crank–Nicholson differencing scheme for the heat equation. By arguments similar to those presented above, it can be shown to be *unconditionally stable.* The limitation on the time step Δt is therefore set by *accuracy,* rather than *stability,* considerations. The drawback to implicit schemes is that it is impossible to solve explicitly for the updated dependent variable at each site. Rather, one typically obtains a *banded-matrix* equation for the updated dependent variable. For example, specializing to one spatial dimension for simplicity, Eq. (31) may be written

$$\begin{aligned} -\frac{\alpha}{2} u_{j-1,k+1} &+ (1+\alpha) u_{j,k+1} - \frac{\alpha}{2} u_{j+1,k+1} \\ &= \frac{\alpha}{2} u_{j-1,k} + (1-\alpha) u_{j,k} + \frac{\alpha}{2} u_{j+1,k}, \end{aligned} \tag{32}$$

which makes it clear that $u_{j,k+1}$, for $j \in \{0, \ldots, N-1\}$, is the solution to a *tridiagonal-matrix* equation. (In two spatial dimensions this would be *pentadiagonal,* etc.) The gain of using a larger time step must thus be weighed against the expense of solving this type of equation. Methods for solution of banded-matrix problems are discussed in the next subsection.

4.2 Banded-Matrix Equations

The general tridiagonal system can be written

$$L_i X_{i-1} + D_i X_i + U_i X_{i+1} = R_i, \tag{33}$$

for $i \in \{1, \ldots, N\}$, with L_1 and U_N understood to be zero. Numerous methods for solving such systems of equations are known and are in widespread use. These tend to fall in one of three categories: *Direct solvers* provide an algorithm whereby the solution to the matrix equation is obtained in a finite number of steps, carried out in a prescribed sequence; *recursive solvers* break down the problem into two or more smaller problems, which are in turn broken down, in recursive fashion, until the resulting system is small enough to solve explicitly; *iterative solvers* provide an iteration for updating the unknown that converges to the correct answer. We consider each of these in turn.

A *direct tridiagonal solver* may be derived by rewriting Eq. (33) as follows:

$$L_i\left(X_{i-1} + \frac{D_i - P_i}{L_i} X_i\right) + P_i\left(X_i + \frac{U_i}{P_i} X_{i+1}\right) = R_i, \tag{34}$$

where the P_i are as yet unspecified. We define

$$Y_i \equiv X_i + (U_i/P_i)X_{i+1} \tag{35}$$

and determine the P_i by demanding that

$$U_{i-1}/P_{i-1} = (D_i - P_i)/L_i, \tag{36}$$

so that

$$L_iY_{i-1} + P_iY_i = R_i. \tag{37}$$

Thus, Eqs. (36) and (37) yield recursions for the P_i and Y_i,

$$P_i = D_i - L_iU_{i-1}/P_{i-1} \tag{38}$$

and

$$Y_i = (R_i - L_iY_{i-1})/P_i. \tag{39}$$

These may be recursed forward in i from the initial condition $P_1 = D_1$ (and remembering that $L_1 = 0$) to obtain all the P's and Y's. Once these are known, Eq. (35) yields a *backward* recursion in i for the desired unknowns,

$$X_i = Y_i - (U_i/P_i)X_{i+1}. \tag{40}$$

Since two recursive sweeps of N steps are required, note that this algorithm has computational complexity of $O(N)$. A Fortran 90 implementation of the above-described algorithm might be

```
SUBROUTINE TRIDIAGONAL_SOLVE(X, L,
 D, U, R)

 IMPLICIT NONE
 INTEGER :: j, n
 REAL :: p
 REAL, DIMENSION(:) :: X, L, D, U, R
 REAL, ALLOCATABLE, DIMENSION(:)
  :: Y

 n = SIZE(D)
 ALLOCATE(Y(n))

 p = D(1)
 IF(p == 0.0) THEN
  PRINT *, ™Error: Ill-conditioned
   matrix. Pivot: 1
  STOP
 END IF
 X(1) = R(1)/p

 DO j = 2, n
  Y(j) = U(j - 1)/p
  p = D(j) - L(j)*Y(j)
  IF(p == 0.0) THEN
   PRINT *, ™Error: Ill-conditioned
    matrix. Pivot: , j
   STOP
  END IF
  X(j) = (R(j) - L(j)*X(j - 1))/p
 END DO
 DO j = n - 1, 1, -1
  X(j) = X(j) - Y(j + 1)*X(j + 1)
 END DO

 DEALLOCATE(Y)

END SUBROUTINE TRIDIAGONAL_SOLVE
```

One problem with the above routine is that the recursions are inherently serial. If parallelism is desired, a different strategy must be adopted. One such strategy is that of a *recursive solver* in which the problem is successively broken down into smaller pieces. Consider, for example, the following three rows of the matrix relationship in Eq. (33):

$$\begin{aligned} L_{i-1}X_{i-2} + D_{i-1}X_{i-1} + U_{i-1}X_i \qquad\qquad\qquad\qquad &= R_{i-1}, \\ L_iX_{i-1} + \quad D_iX_i + \quad U_iX_{i+1} \qquad\qquad &= R_i, \\ L_{i+1}X_i + D_{i+1}X_{i+1} + U_{i+1}X_{i+2} &= R_{i+1}. \end{aligned} \tag{41}$$

Eliminating X_{i-1} using the first of these, and X_{i+1} using the last, the middle equation becomes

$$\left(-\frac{L_iL_{i-1}}{D_{i-1}}\right)X_{i-2} + \left(D_i - \frac{L_iU_{i-1}}{D_{i-1}} - \frac{U_iL_{i+1}}{D_{i+1}}\right)X_i + \left(-\frac{U_iU_{i+1}}{D_{i+1}}\right)X_{i+2} = \left(R_i - \frac{L_iR_{i-1}}{D_{i-1}} - \frac{U_iR_{i+1}}{D_{i+1}}\right). \tag{42}$$

Since this equation couples X_i only to X_{i-2} and X_{i+2}, this has had the effect of breaking the tridiagonal system of N equations into two tridiagonal systems of $N/2$ equations each. If N is a power of two, this process can be continued recursively, until a two-by-two system is obtained, for which the general solution is simple to write down. Since modern languages, such as Fortran 90, support recursion—subroutines that can call themselves—it is a simple matter to capture the above in code:

```
RECURSIVE SUBROUTINE
  NESTED_BISECTION(X, L, D, U, R)
 IMPLICIT NONE
 INTEGER :: n, nh
 REAL :: det
 REAL, DIMENSION(:) :: X, L, D, U, R
 REAL, DIMENSION(:), ALLOCATABLE
  :: LH, DH, UH, XH, RH

 n = SIZE(D)
 IF(n == 1) THEN
  X(1) = R(1)/D(1)
  RETURN
 ELSE IF(n == 2) THEN
  det = D(1)*D(2) - U(1)*L(2)
  X(1) = (D(2)*R(1) - U(1)*R(2))/
   det
  X(2) = (D(1)*R(2) - L(2)*R(1))/
   det
  RETURN
 ELSE IF(MOD(n, 2) == 0) THEN
  nh = n/2
  ALLOCATABLE(LH(nh), DH(nh),
   UH(nh), XH(nh), RH(nh))

  ! Solve for even variables:
  LH(2:nh) = -L(4:n:2)*L(3:n-1:2)/
   D(3:n-1:2)
  DH(1:nh-1) = D(2:n-2:2) -
   L(2:n-2:2)*U(1:n-3:2)/
   D(1:n-3:2) - & & U(2:n-2:2)*
   L(3:n-1:2)/D(3:n-1:2)
  DH(nh) = D(n) - L(n)*U(n-1)/
   D(n-1)
  UH(1:nh-1) = -U(2:n-2:2)*
   U(3:n-1:2)/D(3:n-1:2)
  RH(1:nh-1) = R(2:n-2:2) -
   L(2:n-2:2)*R(1:n-3:2)/
   D(1:n-3:2) - & & U(2:n-2:2)*
   R(3:n-1:2)/D(3:n-1:2)
  RH(nh) = R(n) - L(n)*R(n-1)/
   D(n-1)
  CALL NESTED_BISECTION(X(2::2), LH,
   DH, UH, RH)

  ! Solve for odd values:
  LH(2:nh) = -L(3:n-1:2)*
   L(2:n-1:2)/D(2:n-1:2)
  DH(1) = D(1) - U(1)*L(2)/D(2)
  DH(2:nh) = D(3:n-1:2) -
   L(3:n-1:2)*U(2:n-2:2)/
   D(2:n-2:2) - & & U(3:n-1:2)*
   L(4:n:2)/D(4:n:2)
  UH(1:nh-1) = -U(1:n-1:2)*
   U(2:n:2)/D(2:n:2)
  RH(1) = R(1) - U(1)*R(2)/D(2)
  RH(2:nh) = R(3:n-1:2) -
L(3:n-1:2)*
   R(2:n-2:2)/D(2:n-2:2) - &
   & U(3:n-1:2)*R(4:n:2)/D(4:n:2)
  CALL NESTED_BISECTION(X(1::2), LH,
   DH, UH, RH)

  DEALLOCATE(LH, DH, UH, XH, RH)

 ELSE
  PRINT *, ™Error: Problem size
   contains prime factors other than
   2.
  STOP

 ENDIF

END SUBROUTINE NESTED_BISECTION
```

Note that the above routine treats $N = 1$ and $N = 2$ as special cases. For even $N > 2$ it divides up the problem into two smaller ones as indicated above and calls itself for each of those two smaller problems. If an odd $N > 2$ is ever encountered, it means that the initial problem size was not a power of 2.

Note also that if N contains factors of 3 as well as 2 it is straightforward to work out a variant of the above algorithm that divides the system into three subsystems of size $N/3$, and to add that to the above recursive subroutine, with the test that `MOD(N, 3) == 0`. By working out this treatment for the first several prime numbers, a routine can be created that works for any N containing only those prime factors.

This routine is more amenable to parallel computation than the direct tridiagonal solver described above, since the arithmetic operations needed to compute the matrix coefficients of the smaller systems can be performed in parallel.

4.3 The Galerkin Approximation and Finite-Element Methods

Representing a function by its values at discrete grid points is, of course, only one way of approximating it. More generally, we might consider expanding a function in a series of (not necessarily orthogonal) basis functions, $\psi_\alpha(\mathbf{x})$. These may be a complete set if α is allowed to range over an infinite (discrete and/or continuous) set of values. We can then approximate the dependent variable $u(\mathbf{x},t)$ by restricting α to a finite subset $\mathcal{A}$ of its possible values. Thus, we write

$$u(\mathbf{x},t) \approx U(\mathbf{x},t) \equiv \sum_{\alpha\in\mathcal{A}} U_\alpha(t)\psi_\alpha(\mathbf{x}). \tag{43}$$

An error will generally be incurred if we substitute the approximate dependent variable $U(\mathbf{x},t)$ in any given dynamical equation for the field $u(\mathbf{x},t)$. Returning to our example of the heat equation, this error may be defined,

$$E(\mathbf{x},t) \equiv \frac{\partial U(\mathbf{x},t)}{\partial t} - D\nabla^2 U(\mathbf{x},t). \tag{44}$$

The *Galerkin approximation* fixes the coefficients $U_\alpha(t)$ by demanding that this error be orthogonal to all of the basis functions in the set α, and hence orthogonal to the entire subspace spanned by them. In order to have some notion of orthogonality, an inner product must be defined; this is most often taken to be an integral of the product of the functions,

$$\langle f|g\rangle = \int d^3x W(\mathbf{x}) f(\mathbf{x}) g(\mathbf{x}), \tag{45}$$

where $W(\mathbf{x})$ is a judiciously chosen weighting factor. Thus, we demand that

$$0 = \langle\psi_\alpha|E\rangle = \left\langle\psi_\alpha\left|\frac{\partial U}{\partial t} - D\nabla^2 U\right.\right\rangle \tag{46}$$

for all $\alpha \in \mathcal{A}$. Substituting Eq. (43), we quickly obtain

$$A_{\alpha\beta}\frac{dU_\beta(t)}{dt} = B_{\alpha\beta}U_\beta(t), \tag{47}$$

where there are understood summations over the repeated index $\beta \in \mathcal{A}$, and where we have defined the matrix A with elements

$$A_{\alpha\beta} \equiv \langle\psi_\alpha|\psi_\beta\rangle \tag{48}$$

and B with elements

$$B_{\alpha\beta} \equiv \langle\psi_\alpha|\nabla^2\psi_\beta\rangle. \tag{49}$$

As long as A is nonsingular, this can be solved for the $dU_\beta(t)/dt$, which can then be integrated by any one of a number of different numerical integration schemes. Fed back into Eq. (43), this gives the finite-element approximation to the solution $u(\mathbf{x},t)$.

As a concrete example, let us consider one spatial dimension and suppose that the functions $\psi_\alpha(\mathbf{x})$ are the linear basis splines,

$$\psi_\alpha(x) \equiv \begin{cases} (x - x_{\alpha-1})/(x_\alpha - x_{\alpha-1}), & x_{\alpha-1} \le x < x_\alpha, \\ (x_{\alpha+1} - x)/(x_{\alpha+1} - x_\alpha), & x_\alpha \le x < x_{\alpha+1}, \\ 0 & \text{otherwise,} \end{cases} \tag{50}$$

illustrated in Fig. 7. For the case of equally spaced grid points, $x_\alpha = \alpha\Delta x$ for $\alpha \in \{0, \ldots, N-1\}$, and if the weight function $W(x)$ is taken to be unity, we find that A and B are the tridiagonal matrices with elements

$$A_{\alpha\beta} = \begin{cases} 2\Delta x/3, & \alpha = \beta, \\ \Delta x/6, & |\alpha - \beta| = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{51}$$

and

$$B_{\alpha\beta} = \begin{cases} 2/\Delta x, & \alpha = \beta, \\ -1/\Delta x, & |\alpha - \beta| = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{52}$$

Thus, Eq. (47) can be solved using one of the previously discussed tridiagonal algorithms, and the system can be integrated forward in time. Note that this system is quite similar to that of Eq. (15), except that the matrix $A_{\alpha\beta}$

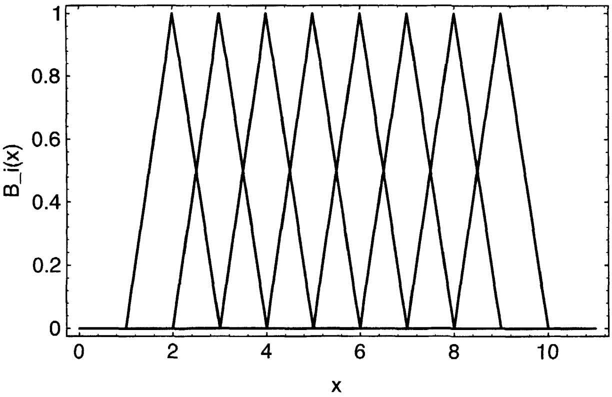

FIG. 7. Linear basis splines for finite-element method in one dimension. Function $B_j(x)$ increases linearly from 0 at $x = x_{j-1}$ to 1 at $x = x_j$, and then decreases back to zero at $x = x_{j+1}$. Functions $B_2(x)$ through $B_9(x)$ are shown for equally spaced grid points, $x_j = j$. These span the space of piecewise linear functions.

has the effect of smoothing the time differencing scheme a bit.

Because the dependent variable is represented as a sum of the linear splines illustrated in Fig. 7, it is clear that the function subspace that we are considering is that of all piecewise linear functions. It is also clear that we would still get a tridiagonal system if the grid points were spaced irregularly. Thus, finite-element methods can be easily applied to situations in which one needs more grid refinement in one part of a spatial domain than another. Moreover, the method could be used with a smoother set of basis splines than the linear splines used above. Cubic basis splines, assuring continuity of first and second derivatives at the gridpoints, are commonly used; in this situation, the resulting matrix equation will still be banded, though it will be pentadiagonal, rather than tridiagonal.

The method assumes great importance in more than one spatial dimension and for problems with irregular boundary conditions, such as are often encountered in engineering applications. In two dimensions, for example, it is possible to break up a domain into triangular or other polygonal regions and define spline functions on them with the appropriate continuity properties. Tensor products of one-dimensional basis splines can be used to create multidimensional basis splines, and the Galerkin approximation can still be employed to derive a finite-element method. In such situations, the matrix A is generally no longer neatly banded, though it is still likely to be *sparse* for problems involving purely local differential operators; a number of efficient schemes exist for storing and inverting such sparse matrices. Finally, we note that the method is also sometimes used for the solution of integrodifferential equations for which the matrix A is dense.

5. SPECTRAL METHODS

5.1 Spectral Representation of Data

In the applications of the Galerkin method that we examined in the previous section, we chose basis functions that were *local*—i.e., had compact support. This restriction is unnecessary. One might consider, for example, expanding a function on a periodic domain in a discrete Fourier series and representing it on a computer by the coefficients of this series. For problems in other geometries, or with other linear operators, other basis functions may be more appropriate. While Fourier representations diagonalize the ubiquitous del operator (and hence the gradient, divergence, and Laplacian) in periodic geometry, a Fourier–Bessel representation may be more appropriate for cylindrical geometry, spherical harmonics may be more appropriate for spherical geometry, and so on. Of these, only the Fourier transform with sine/cosine basis functions has a *fast* version (the so-called fast Fourier transform, or FFT), and so it is naturally the most useful in this regard.

It is easy to see that the method diagonalizes any linear, homogeneous equation, reducing it to a set of independent ordinary differential equations to be integrated. The challenge is therefore to deal with inhomogeneities and/or nonlinearities, which will couple the modes. The example of the heat equation, which served us well in the last section, is inadequate to illustrate this issue, and so we add a nonlinear term $|\nabla u(\mathbf{x},t)|^2$ on the left, and some inhomogeneous noise $\eta(\mathbf{x},t)$ on the right; the result is the *Kardar–Parisi–Zhang* (KPZ) equation,

$$\frac{\partial u(\mathbf{x},t)}{\partial t} + \tfrac{1}{2}|\nabla u(\mathbf{x},t)|^2 = D\nabla^2 u(\mathbf{x},t) + \eta(\mathbf{x},t), \tag{53}$$

which has received wide attention as a model of interfacial growth (Kardar *et al.*, 1986). If we expand $u(\mathbf{x},t)$ in a set of *orthogonal* basis functions,

$$u(\mathbf{x},t) = \sum_{\alpha} u_\alpha(t)|\psi_\alpha\rangle \tag{54}$$

for $\alpha \in \{1, \ldots, N\}$, we find that

$$\begin{aligned}\frac{du_\alpha(t)}{dt} &+ \tfrac{1}{2}\sum_{\beta<\gamma} u_\beta(t)u_\gamma(t)\langle\phi_\alpha|\nabla\phi_\beta\cdot\nabla\phi_\gamma\rangle \\ &= D\sum_{\beta} u_\beta(t)\langle\phi_\alpha|\nabla^2|\phi_\beta\rangle + \eta_\alpha(t),\end{aligned} \tag{55}$$

where $\eta_\alpha(t) \equiv \langle\phi_\alpha|\eta\rangle$.

To make this concrete, let us take periodic boundary conditions in a cube of side L,

$$|\phi_{\mathbf{j}}\rangle = \exp\left(\frac{2\pi i\mathbf{j}\cdot\mathbf{x}}{L}\right), \tag{56}$$

so that the basis functions are indexed by the wave number $\mathbf{j}$. The Laplacian operator is then diagonalized, and one sum drops out of the nonlinear term to yield

$$\begin{aligned}\frac{du_{\mathbf{j}}(t)}{dt} &- \frac{2\pi^2}{L^2}\sum_{\mathbf{k}} \mathbf{k}\cdot(\mathbf{j}-\mathbf{k})u_{\mathbf{k}}(t)u_{\mathbf{j}-\mathbf{k}}(t) \\ &= -\frac{4\pi^2 D}{L^2}|\mathbf{j}|^2 u_{\mathbf{j}}(t) + \eta_{\mathbf{j}}(t).\end{aligned} \tag{57}$$

Note that the quadratic nonlinearity in the original equation has given rise to a corresponding quadratic nonlinearity in the equation for the mode amplitudes. Suppose that we truncate our spectral representation of $u(\mathbf{x},t)$ at some finite number N of modes in each direction. For a d-dimensional spatial domain, there will thus be a total of N^d modes. It follows that the upper limit of each of the d summations on the left-hand side of Eq. (57) should be N. This yields a closed set of equations for the mode amplitudes. As long as N is chosen to be large enough that the smallest characteristic features of the flow can be represented, this is a valid numerical approximation.

On the other hand, note that the nonlinear term requires on the order of N^{2d} steps to compute, since there are N^d equations, each with d nested N-term sums. Since a fast Fourier transform of N^d data requires only on the order of $N^d \log(N^d) \sim dN^d \log(N)$ steps, it actually becomes less expensive to compute the gradient ∇u in Fourier space, square it in real space, and then transform the result back to Fourier space. This strategy is called a *pseudospectral algorithm.* Symbolically, this procedure may be written

$$\begin{aligned}\frac{du_{\mathbf{j}}(t)}{dt} &- \frac{2\pi^2}{L^2}\mathcal{F}[|\mathcal{F}^{-1}(\mathbf{k}u_{\mathbf{k}})|^2] \\ &= -\frac{4\pi^2 D}{L^2}|\mathbf{j}|^2 u_{\mathbf{j}}(t) + \eta_{\mathbf{j}}(t),\end{aligned} \tag{58}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier transform and its inverse, respectively.

5.2 Operator Splitting

Multiplication of the dependent variable by a spatially inhomogeneous term can also couple the modes. A classic example of this is provided by the time-dependent Schrödinger equation,

$$i\frac{\partial\Psi(\mathbf{x},t)}{\partial t} = -\tfrac{1}{2}\nabla^2\Psi(\mathbf{x},t) + V(\mathbf{x})\Psi(\mathbf{x},t). \tag{59}$$

We could proceed as above to solve this equation using spectral methods, but there are other considerations: Specifically, we know that Schrödinger evolution should be *unitary*; that is, it should conserve total probability $|\Psi(\mathbf{x},t)|^2$. It would be nice to develop a numerical method that respected this constraint.

Toward this end, we write the Schrödinger equation in the following format:

$$i\frac{\partial\Psi}{\partial t} = \hat{H}\Psi \equiv \hat{T}\Psi + \hat{V}\Psi, \tag{60}$$

where $\hat{T} \equiv -\nabla^2/2$ and $\hat{V} = V(\mathbf{x},t)$ are the kinetic and potential energy operators, respectively. This equation has the formal solution

$$\mathbf{\Psi}(t + \Delta t) = \exp(-i\Delta t\hat{H})\Psi(t) = \exp[-i\Delta t(\hat{T} + \hat{V})]\mathbf{\Psi}(t). \qquad (61)$$

Now we cannot write $\exp[-i\Delta t(\hat{T} + \hat{V})] = \exp[-i\Delta t\hat{T}]\exp[-i\Delta t\hat{V}]$ since the operators $\hat{T}$ and $\hat{V}$ generally do not commute. Motivated by the Baker–Campbell–Hausdorff theorem (see, e.g., Belinfante and Kelman, 1972), however, we can write

$$\exp[-i\Delta t(\hat{T} + \hat{V})] = \exp[-i\Delta t\hat{T}]\exp[-i\Delta t\hat{V}] + O(\Delta t^2), \qquad (62)$$

or

$$\exp[-i\Delta t(\hat{T} + \hat{V})] = \exp\left[-\frac{i}{2}\Delta t\hat{V}\right] \times \exp[-i\Delta t\hat{T}]\exp\left[-\frac{i}{2}\Delta t\hat{V}\right] + O(\Delta t^3), \qquad (63)$$

etc. If we integrate for $O(\Delta t^{-1})$ time steps, the first (second) of these will give rise to an accumulated error of order $O(\Delta t)$ $[O(\Delta t^2)]$. Choosing the first for simplicity, we have

$$\mathbf{\Psi}(t + \Delta t) \approx \exp(-i\Delta t\hat{T})\exp(-i\Delta t\hat{V})\mathbf{\Psi}(t). \qquad (64)$$

Now the first of these operators, $\exp(-i\Delta t\hat{T})$, is diagonalized in Fourier space since

$$\exp(-i\Delta t\hat{T}) = \exp\left(\frac{i\Delta t}{2}\nabla^2\right) = \mathcal{F}^{-1}\exp\left(-\frac{i\Delta t}{2}k^2\right)\mathcal{F}, \qquad (65)$$

where k is the wave vector; the second, $\exp(-i\Delta t\hat{V})$, is diagonalized in real space. We can thus write

$$\mathbf{\Psi}(t + \Delta t) \approx \mathcal{F}^{-1}\left\{\exp\left(-\frac{i\Delta t}{2}k^2\right)\mathcal{F}[\exp(-i\Delta t\hat{V})\mathbf{\Psi}(t)]\right\}. \qquad (66)$$

Given the wave function at time t, the complete algorithm for one time step is therefore to perform a local (in real space) multiplication by $\exp(-i\Delta t\hat{V})$, Fourier transform the result, perform a local (in Fourier space) multiplication by $\exp(-i\Delta tk^2/2)$, and invert the Fourier transform. Like the KPZ equation (53), note that the portion of the dynamics that does not mix modes is treated in Fourier space, while the portion that does mix modes is treated in real space; that is, it is pseudospectral. Also note that this algorithm has the virtue of being *exactly* unitary—or at least to within roundoff error—even for finite Δt.

Note that the presence of the global Fourier transform means that the algorithm is effectively implicit, insofar as the new wave function at any one point depends on the old one everywhere on the grid. In fact, the situation is actually worse than it was for the implicit finite-difference schemes discussed earlier, since those could be handled by a tridiagonal solver in $O(N)$ operations, where N is the number of grid points. The FFT, by contrast, will require $O(N \log N)$ operations. Thus, it would be nice to have a fully *explicit,* operator-splitting methodology for integrating Schrodinger's equation, maintaining exact unitarity. Such a method was discovered only fairly recently (DeRaedt 1987). To understand it, consider a one-dimensional, periodic, finite-difference representation of the Hamiltonian operator that was used above:

$$\hat{H} = -\tfrac{1}{2}\nabla^2 + V(\mathbf{x}) = -\frac{1}{2}\begin{pmatrix} 2 & -1 & 0 & 0 & 0 & \cdots & -1 \\ -1 & 2 & -1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & 0 & 0 & 0 & \cdots & 2 \end{pmatrix} + \begin{pmatrix} V_1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & V_2 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & V_3 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & V_4 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & V_5 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & V_N \end{pmatrix} = \hat{H}_{\text{even}} + \hat{H}_{\text{odd}}, \tag{67}$$

where we have defined $V_j \equiv V(x_j)$, and the matrix operators

$$\hat{H}_{\text{even}} \equiv \frac{1}{2}\begin{pmatrix} V_1 - 1 & & & & & \cdots & 1 \\ & V_2 - 1 & 1 & & & \cdots & \\ & 1 & V_3 - 1 & & & \cdots & \\ & & & V_4 - 1 & 1 & \cdots & \\ & & & 1 & V_5 - 1 & \cdots & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & & & & & \cdots & V_N - 1 \end{pmatrix} \tag{68}$$

and

$$\hat{H}_{\text{odd}} \equiv \frac{1}{2}\begin{pmatrix} V_1 - 1 & 1 & & & & \cdots & \\ 1 & V_2 - 1 & & & & \cdots & \\ & & V_3 - 1 & 1 & & \cdots & \\ & & 1 & V_4 - 1 & & \cdots & \\ & & & & V_5 - 1 & \cdots & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ & & & & & \cdots & V_N - 1 \end{pmatrix}. \tag{69}$$

We have purposely suppressed the zero matrix elements of these operators to emphasize the fact that both $\hat{H}_{\text{even}}$ and $\hat{H}_{\text{odd}}$ are direct sums of 2 × 2 matrices of the form

$$\hat{h}_j \equiv \begin{pmatrix} V_j - 1 & 1 \\ 1 & V_{j+1} - 1 \end{pmatrix}, \tag{70}$$

for j even and odd, respectively.

It is now possible to use operator splitting on the division of the Hamiltonian into $\hat{H}_{\text{even}}$ and $\hat{H}_{\text{odd}}$, rather than its division into $\hat{T}$ and $\hat{V}$. Because $\hat{H}_{\text{even}}$ and $\hat{H}_{\text{odd}}$ are direct sums of 2 × 2 blocks, however, their exponentiation will be likewise; that is, in order to exponentiate this Hamiltonian, all one needs to do is exponentiate a 2 × 2 Hermitian matrix. This is an elementary exercise, and the evolution will consist of the application of a sequence of 2 × 2 block unitary matrices to the wave function. Since these combine only neighboring values of the wave function on the grid, the evolution will be explicit as well as exactly unitary. This method also has the advantage of being straightforwardly parallelized (Richardson, 1991). Finally we note that, more recently, other fully explicit, exactly unitary methods have been developed for the integration of Schrödinger's equation (Boghosian and Taylor, 1998), but they are based on a quantum lattice-gas approach rather than on operator splitting.

5.3 Spectral-Element Methods and Wavelets

The orthogonality of the spectral basis functions in the previous examples has precluded the need to invert a tridiagonal matrix, as was necessary for the Galerkin method in Eq. (47). Moreover, sine and cosine basis functions have the virtue of diagonalizing differential operators, and possessing a "fast" transform—the FFT. At the same time, sine and cosine basis functions have certain disadvantages. All of our foregoing examples employed very simple boundary conditions. Indeed, sine and cosine basis functions are easiest for periodic boundary conditions. Moreover, since sine and cosine functions lack the compact support of basis splines, they are awkward for representing localized features.

A relatively recent numerical technique called the *spectral-element* method seeks to combine the best features of spectral methods and finite elements. Typically, it uses polyhedral elements; since these have a simple geometry, it is straightforward to adopt a spectral representation within each of them. The boundary conditions between elements give rise to a sparse-matrix problem, just as they did for the Galerkin method. By making the elements larger and using more modes to compensate for any loss of resolution, however, the size of this sparse-matrix problem can be reduced; this must be weighed against the benefit of using smaller, more localized elements to represent compact features.

The tradeoff, as stated earlier, would seem to be one between the *orthogonality* of the basis functions, useful for simplifying the solution procedure, and their *compact support,* useful for representing localized features. For example, basis splines are clearly compact but nonorthogonal, whereas sines and cosines are orthogonal but noncompact. From this point of view, one of the more exciting discoveries of the past ten years has been the fact that it is possible to construct sets of basis functions, called *wavelet bases,* that have both orthogonality and compact support. Like basis splines, these consist of different translations of a single canonical function, but they also include different scales as well. That is, a general wavelet expansion is of the form

$$f(x) = \sum_{j,k} b_{jk} W(2^j x - k), \tag{71}$$

where $W(x)$ can be chosen to satisfy a mixture of essential and optional requirements.

To get a general idea of how a wavelet basis is constructed, suppose that we are using the basis of normalized box functions,

$$\phi_j^{(l)}(x) \equiv \begin{cases} 1, & j \le x < j + 1, \\ 0 & \text{otherwise,} \end{cases} \tag{72}$$

where $j \in \{0, 1, 2, \ldots, 2^l - 1\}$, to represent piecewise constant functions on the interval $0 \le x < 2^l$. We can obviate the need for the subscript j by noting that

$$\phi_j^{(l)}(x) = \phi^{(l)}(x - j), \tag{73}$$

where $\phi^{(l)}(x) \equiv \phi_0^{(l)}(x)$; that is, the basis functions for different values of j are just translations of each other.

Next we consider the coarser basis obtained by scaling this one by a factor of 2. We define the scaled function

$$\phi^{(l-1)}(x) = \tfrac{1}{2}\phi^{(l)}\left(\frac{x}{2}\right) \tag{74}$$

and construct a basis with its translations,

$$\phi_j^{(l-1)}(x) = \phi^{(l-1)}(x - 2j). \tag{75}$$

The overall factor of $\frac{1}{2}$ on the right-hand side of Eq. (74) has been included to maintain normalization. Note that this coarser basis has only half as many basis functions as the one we started with. The finer basis is therefore capable of representing everything that the coarser basis can represent and more; specifically, it can represent a factor of 2 finer detail. To make this more explicit, note that each member of the coarse grid can be written as the sum of two members of the finer grid,

$$\phi_j^{(l-1)}(x) = \tfrac{1}{2}[\phi_{2j}^{(l)}(x) + \phi_{2j+1}^{(l)}(x)]. \tag{76}$$

It follows that the difference in resolution between the finer and coarser bases—that is, the space of functions that is spanned by the finer basis but is orthogonal to the coarser basis—is spanned by the set of functions

$$W_j^{(l-1)}(x) = \tfrac{1}{2}[\phi_{2j}^{(l)}(x) - \phi_{2j+1}^{(l)}(x)]. \tag{77}$$

These functions are called the *wavelets* corresponding to the basis of Eq. (72).

In more abstract terms, let us suppose

that the space of functions spanned by the $\phi_j^{(l)}(x)$ is called $\Lambda^{(l)}$, and that spanned by the $W_j^{(l)}(x)$ is called $\Omega^{(l)}$. The argument in the previous paragraph has demonstrated that

$$\Lambda^{(l)} = \Lambda^{(l-1)} \oplus \Omega^{(l-1)}. \tag{78}$$

That is, we have decomposed the original basis into the direct sum of its coarsening and a space of wavelets. We can now repeat this process with the space $\Lambda^{(l-1)}$, etc. The final result will be the decomposition of the original basis into a set of wavelet functions and their translations at each scale (plus possibly a constant function at the very coarsest scale). That is,

$$\Lambda^{(l)} = \Lambda^{(0)} \oplus \Omega^{(1)} \oplus \cdots \oplus \Omega^{(l-1)}. \tag{79}$$

It is clear from the method of construction that this decomposition results in an orthonormal basis.

Note that not all bases are capable of generating wavelets in the manner explained above. A fundamental requirement, satisfied by the hat function used above, is that each basis function at a given level of the hierarchy can be expressed as a linear combination of basis functions at the next finest level. That is, the basis functions ϕ must be such that

$$\phi(x) = \sum_k c_k \phi(2x - k), \tag{80}$$

for some constants c_k. It is not difficult to see that the condition that ϕ is normalized results in the constraint

$$\sum_k c_k = 2. \tag{81}$$

Not all functions will obey this *dilation equation.* Fortunately, as noted by Strang (1989), an amazing variety of familiar families of basis functions do obey it. For example, if $c_j = 2\delta_{j,0}$, then ϕ is the Dirac delta function. Our above example showed that the box function obeys the dilation equation with $c_0 = c_1 = 1$. Likewise, the piecewise linear functions of Eq. (50) on a uniform grid as illustrated in Fig. 7 also solve the dilation equation with $c_0 = c_2 = \frac{1}{2}$ and $c_1 = 1$.

The ability of these functions to represent polynomials in x can also be expressed as conditions on the c coefficients (Strang, 1989). Specifically, if the function

$$P(\xi) \equiv \tfrac{1}{2} \sum_k c_k e^{ik\xi} \tag{82}$$

has a pth-order pole at $\xi = \pi$, then the basis will be able to represent exactly all polynomials in x spanned by the set of monomials $\{1, x, x^2, x^3, \ldots, x^{p-1}\}$. Thus, the hat function used in the example mentioned above had $P(\xi) = \exp(ik\xi/2) \cos(k\xi/2)$, which has a pole of order $p = 1$, so it is capable of representing only piecewise constant functions.

Given a basis that obeys the dilation equation for some coefficients c, then the generalization of the hat-function argument given above shows that the corresponding wavelet basis is given by

$$W(x) = \sum_k (-1)^k c_{1-k} \phi(2x - k). \tag{83}$$

All of this formalism makes it straightforward to construct bases satisfying the dilation equation with the ability to represent polynomials of specified order and to construct the corresponding wavelet basis. Moreover, it can be shown that such wavelet bases, as a result of their hierarchical construction, possess a "fast" transform, like the FFT, so that they are of $O(N \log N)$ computational complexity for N data elements.

It is interesting to note that this procedure does not always result in smooth, well-behaved basis functions. In fact, Daubechies (1988) showed that the choice $c_0 = (1 + \sqrt{3})/4$, $c_1 = (3 + \sqrt{3})/4$, $c_2 = (3 - \sqrt{3})/4$, and $c_3 = (1 - \sqrt{3})/4$ results in a wavelet basis that is orthogonal, is compact on $[-1,2]$, and is capable of exactly representing any piecewise linear function. This function is illustrated in Fig. 8 (though it is differentiable almost nowhere, so that any graphical representation of it is rather dubious).

It is clear that wavelet bases that are able to represent higher powers of x exactly are a straightforward generalization of the above. On the other hand, wavelets do not diagonalize differential operators, as does a Fourier representation. Though a detailed discussion of the theory and usage of wavelets is outside the scope of this article, there are by now a very large number of references on this subject. At the time of this writing, this

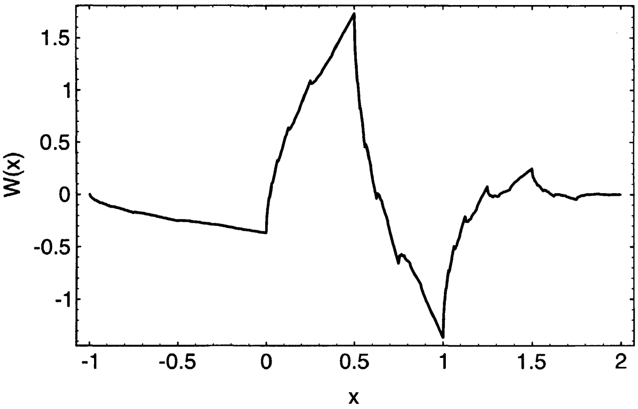

FIG. 8. Daubechies's wavelet function. In spite of this graphical representation, it should be noted that this function is not smooth and is differentiable almost nowhere.

remains a very active area of applied mathematical research.

6. PARTICULATE METHODS

From the above descriptions of finite-difference, finite-element, and spectral methods, the reader may have developed the impression that such methods are adequate for the simulation of any partial differential equation of interest in mathematical physics. In fact, nothing could be further from the truth. Certain partial differential equations are capable of generating structures on arbitrarily small length scales of their dependent variable(s). When we examined finite-difference approximations to the Navier–Stokes equations, for example, we discovered that the need for a grid fine enough to resolve the very smallest vortices was an important and severe restriction; in spite of this, we can feel at least somewhat fortunate that there is some nonzero lower bound on the structures developed by the Navier–Stokes equations at any finite Reynolds number. Imagine how much more difficult it is to simulate the Euler equations of inviscid fluid flow or the Poisson–Vlasov equations of plasma physics, which have no viscous terms to damp out the development of structures below a certain length scale. The structures developed by the evolution of such *Hamiltonian systems* can become arbitrarily filamentary, so that their representation on any fixed spatial grid, or using any fixed set of basis functions, is invalidated in finite time.

In spite of this seeming intractibility, it is possible to make numerical progress on such problems. One approach is to try to develop a spectral method with N modes that is a Hamiltonian system *for all N.* In this way, even the finite-dimensional truncated system is Hamiltonian, and has a symplectic phase-space structure—that is, the time rate of change of any phase-space function F is given by $\{F,H\}$, where H is a Hamiltonian and $\{\cdot,\cdot\}$ is a Poisson bracket that is bilinear and antisymmetric and obeys the Jacobi identity. For example, this was accomplished for Euler's ideal-fluid equations by Miller *et al.* (1992). While such methods are still not capable of representing arbitrarily filamentary modes, they at least respect the various constraints imposed by Hamiltonian evolution—conservation of phase-space volume and the integral invariants of Poincaré—to numerical accuracy (Channell and Scovel, 1990); this is analogous to the way that the operator-splitting method described above respected unitarity.

Another general method for simulating systems of this sort has been to abandon the spatial (or spectral) discreteness that has characterized all of the methods that we have considered thus far, and represent the field by the density of local particles, each of which is free to take its location in the continuum. That is, instead of an *Eulerian representation* of a dependent variable $\Psi(\mathbf{x}_j,t)$ at each point of a spatial grid, we might adopt a *Lagrangian representation* in which we keep track of the spatial coordinates, and perhaps weights, of some finite number of particles at each time step. This method does not require any kind of grid at all, though we note in passing that the finiteness of floating-point precision on any real computer is tantamount to the existence of a very fine grid, in that particle positions are restricted to a discrete set of values. Also, there exist hybrid algorithms, such as the *particle-in-cell* method, in which some quantities are represented on a grid and others are represented by discrete particles.

6.1 Short-Range Interactions

The first and most intuitive application of particulate methods is their use for tracking the positions and velocities of discrete particles, such as molecules, interacting according to a known classical force law. Such *molecular dynamics* (MD) algorithms are distinguished by their application to either short-range or long-range force laws (or both together). An example of the former is the Lennard-Jones potential,

$$V_{\mathrm{LJ}}(r) = V_0[(r_0/r)^{12} - (r_0/r)^6]. \tag{84}$$

This interaction potential drops off so sharply for large r that effects of particles outside of a certain *interaction sphere* can be simply ignored. Alternatively, one may employ a version of this potential for which this cutoff is explicitly enforced by taking

$$V(r) = \begin{cases} V_{\mathrm{LJ}}(r) - V_{\mathrm{LJ}}(r_1) & \text{for } r < r_1, \\ 0 & \text{otherwise,} \end{cases} \tag{85}$$

where the radius of the interaction sphere is $r_1 > r_0$. By contrast, the Coulomb potential

$$V(r) = e^2/r, \tag{86}$$

where e is the electric charge, can almost never be truncated because although the force decays like r^{-2}, the number of particles between distance r and $r + dr$ generally increases like r^2 so that effects of distant particles cannot be ignored.

In a purely Lagrangian representation of a system of particles with short-range interactions, only the coordinates of each particle need to be represented. The data structure might therefore consist of coordinates $\mathbf{x}_j$ for $j \in \{1, \ldots, N\}$, where N is the total number of particles. The force on a particle is computed by taking the vector sum of the forces due to all of the other particles in its interaction sphere. The challenge here is to find a way to compute this in a time that is linear in N, rather than quadratic. This ought to be possible, since for a substance at a given average density and temperature, there are only a fixed characteristic number of particles in an interaction sphere. If we denote this number by ν, then we need to compute on the order of νN forces. While this is linear in N, it is not obvious how to determine which other particles are in the interaction sphere of a given particle in a time that is linear in N. For example, everybody's first try at the problem,

```
DO i = 1, N
 DO j = 1, N
  IF(DISTANCE(i, j) < r1) FORCE(i) =
   FORCE(i) + GET_FORCE(i, j)
 END DO
END DO
```

(where the functionality of the variously called routines is obvious from their names), will decidedly *not* work. Even though the `IF` test is satisfied only on the order of ν times for each i, just the testing to see if the particles are farther than a distance r_1 apart will take on the order of N^2 operations in the above code. Since the very largest MD simulations run today use hundreds of millions of particles, no portion of the algorithm may be of order N^2 complexity.

Fortunately, there are numerous well-researched parallel algorithms for rapidly determining and updating the list of neighbors of a given particle (Allen and Tildesley, 1987). Interestingly, most algorithms for determining neighbors employ an Eulerian grid of some sort. If the grid is regular, each particle may easily determine which grid cell it is in, and the particles may be sorted according to cell number. "Bookkeeping" problems arise due to particles that are near cell boundaries, but they are surmountable. Moreover, once a list of neighbors has been determined for each particle, this *neighbor table* may be updated incrementally, rather than recomputed from scratch at each time step. Using all of these various tricks (and more) on large parallel computers, MD simulations with between 10^8 and 10^9 particles and with a characteristic performance of 10^{-8} s per particle update (the standard figure of merit for MD simulations) have been achieved (Zhou *et al.*, 1997).

6.2 Molecular Dynamics

The most straightforward type of MD simulation tracks the motion of a fixed number N of particles, constrained to move inside a fixed volume V. Periodic boundary conditions are often employed to emulate a large homogeneous sample of matter. Newton's equations for these particles,

$$\dot{\mathbf{r}}_i = \mathbf{p}_i/m_i, \tag{87}$$

$$\dot{\mathbf{p}}_i = \sum_{j\neq i}^{N} \mathbf{F}(i \leftarrow j), \tag{88}$$

where $\mathbf{F}(i \leftarrow j)$ is the force on particle i due to particle j, derivable from a time-independent potential $V(\mathbf{r})$, will conserve the total energy

$$E = \sum_{i}^{N} \frac{p_i^2}{2m_i} + \tfrac{1}{2}\sum_{i\neq j} V(|\mathbf{r}_i - \mathbf{r}_j|). \tag{89}$$

The state of the system thus samples the microcanonical ensemble. The average *kinetic* energy, and hence the temperature, will fluctuate. Typically, some desired property, such as a virial coefficient, a dielectric constant, or the shear viscosity, is time-averaged over the dynamics. For example, if the desired quantity is denoted by χ, the quantity

$$\langle\chi\rangle = \lim_{T\to\infty} \frac{1}{T}\int_0^T dt\chi(t) \tag{90}$$

is accumulated during the run. Its relationship to the true equilibrium value of this quantity thus rests on the assumption of ergodicity.

The first MD simulations were investigations of hard-disk fluids, carried out by Alder and Wainwright (1957) in the late 1950s. In a later article (Alder and Wainwright, 1967), these authors used such simulations to establish the existence of *long-time tails* in particles' velocity autocorrelation functions; that is, the quantity $\langle\mathbf{v}(0)\mathbf{v}(t)\rangle$, averaged over a large ensemble of particles, was found not to decay exponentially in time, as had been the common assumption, but rather to decay according to a power law. This unexpected result, of substantial physical import, was a great triumph of the MD method.

The number of particles that could be simulated at that time was in the tens or hundreds; as noted above, it is now in the hundreds of millions. Though this is undeniably a drastic improvement, it needs to be put into perspective: It is still only 10^{-15} moles of material. In terms of physical time, one time step of an MD simulation is on the order of 10^{-15} s. The very longest runs can carry out 10^7 time steps, for a total of about 10 ns of physical time. This is long enough for the energy to reach equipartition, and to compute many equilibrium properties of interest, transport coefficients, etc. Certain phenomena, however, such as protein folding and the dynamics of molecular clumps in glassy materials, occur on much longer time scales, and MD has been inadequate for their investigation. This is particularly true of *soft condensed-matter* applications, such as the development of micelles from dissolved surfactant, the response of liquid crystals to external fields, and the Brownian fluctuations of colloidal particles. Such *complex fluids* develop *mesoscale* structures—with characteristic lengths and times that are much larger than a mean free path and a mean free time, respectively—that dominate their macroscopic behavior.

Much work has been done over the last 20 years to accelerate the thermalization of molecular dynamics simulations. Two principal methodologies have been adopted for this purpose. One is to alter the particles' positions and/or momenta occasionally in stochastic fashion; the challenge here is to do this in a way that allows the particles to sample the phase space more rapidly than strict adherence to the Newtonian trajectories would allow, but to avoid doing such extensive damage to those trajectories that dynamic properties, such as time correlation functions and transport coefficients, are rendered meaningless. A number of recent methods adopt this strategy; these include, but are not limited to, lattice-gas automata (Frisch *et al.*, 1986, 1987), dissipative particle dynamics (Hoogerbrugge and Koelman, 1992; Espanol, 1995), the Langevin method (see, e.g., Toral, 1995), the lattice Boltzmann equation (Chen, 1998), direct-simulation Monte Carlo (Bird, 1994; Garcia, 1994), hybrid Monte Carlo (Heermann, *et al.*, 1990), and smoothed-particle hydrodynamics (Monaghan, 1992). The first of these is discussed further in Sec. 6.5.

Another method of accelerating the calculation of certain quantities of interest is to use an ensemble other than the microcanonical one (constant E, V, and N) described above. For example, Nosé (1986) and Hoover (1985) have shown how to modify particulate equations of motion so that the particles are no longer constrained to the energy surface, but rather provide a sampling of the full canonical ensemble at specified temperature.

6.3 Long-Range Interactions

Long-range interactions are ubiquitous in physics. Charged particles in a plasma may interact according to the Coulomb potential. Interactions between stars in a galaxy, and indeed between galaxies in the expanding universe, may be modeled by the Newtonian gravitational potential. Somewhat less well known is the fact that particulate methods with long-range interactions may also be used to model incompressible fluid flow. For example, defining the *vorticity* of a fluid as the curl of its hydrodynamic velocity

$$\boldsymbol{\omega} \equiv \nabla \times \mathbf{v}, \qquad (91)$$

and noting that

$$0 = \nabla \cdot \mathbf{v} \qquad (92)$$

for an incompressible fluid, we recognize the similarity of these equations to those of magnetostatics: Velocity plays the role of the magnetic field **B** (whose divergence vanishes), and vorticity plays the role of the current density **J** (equal to the curl of the velocity). It follows that vortex filaments give rise to velocity fields according to the same Biot–Savart law that describes how current filaments give rise to magnetic fields. We can thus represent the fluid motion by vortex filaments in three dimensions (or point vortices in two dimensions), rather than the usual scheme of representing the velocity directly. Furthermore, if we take the curl of both sides of the Navier–Stokes equation for momentum conservation, Eq. (25), we can derive the equation of motion of these vortex filaments (or points). It turns out that each moves in the velocity field of all the other vortices present and also has a random velocity component to model the viscous diffusion of momentum. To move the vortices at each step therefore requires summing the Biot–Savart law for the velocity at each one due to all of the others. Thus, the *vortex method* reduces the problem of incompressible fluid flow to that of long-range "particle" interactions (Gustafson and Sethian, 1991).

The most obvious way to sum up all of the forces is then something like

```
DO i = 1, N
 DO j = 1, N
  FORCE(i) = FORCE(i) +
   GET_FORCE(i, j)
 END DO
END DO
```

Clearly, this has order N^2 complexity, but this seems insurmountable since the forces are long range, and every particle really does interact with every other. In one of the more counterintuitive algorithmic results to be recently discovered in this field, it has been shown (Rokhlin and Greengard, 1987) that—for the special case of the Coulomb force—the required summation can be accomplished to machine precision with order-N computational complexity. This *fast multipole algorithm* makes use of the fact that forces due to distant particles can be approximated by a multipole expansion. It applies this observation in a recursive fashion: We first subdivide the spatial domain into eight parts using three mutually perpendicular planes in three dimensions (four parts using two perpendicular lines in two dimensions) and continue to subdivide the parts in the same way to obtain an *oct-tree* division of the spatial domain in three dimensions (*quad-tree* in two). The two-dimensional case is illustrated in Fig. 9. From its coordinates, it is straightforward for each particle to determine which "leaf" it occupies on such a tree, and the number of branchings is taken such that there is on the average of only a few particles per leaf. The algorithm begins by computing the multipole moments of each leaf. It then takes advantage of the fact that, for the Coulomb potential, multipole moments of "child" and "sibling" domains can be easily combined to get multipole mo-

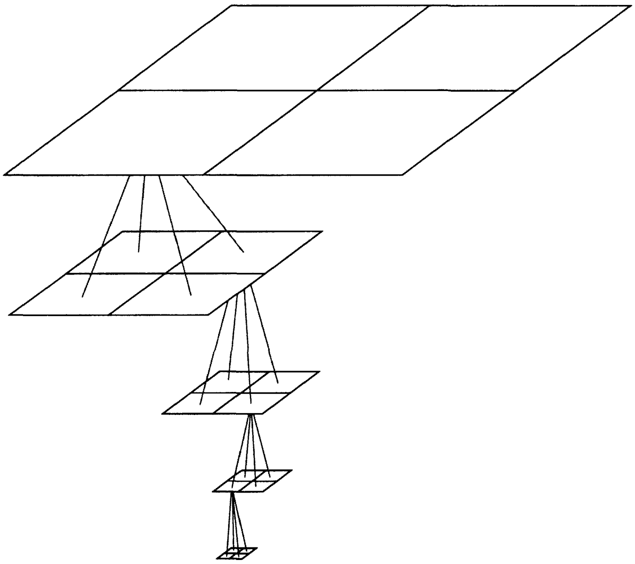

FIG. 9. Quad-tree division of spatial domain.

ments of "parent" domains. Thus, it works its way to the root of the tree, computing the multipole moments of each division at each level. Next, it recursively works its way back to the leaves, using the multipole moments of distant "parent" domains to obtain their contributions to the potential in every "child" domain. When it arrives at the leaves, it has the potential (or the force) at each leaf due to all of the other sites on the finest grid. Finally, it adds the contribution due to the other particles in the same cell. Since there are 4^j domains at level j and $\log_4 N$ levels in the tree, the number of operations performed is on the order of

$$\sum_{j=0}^{\log_4 N} 4^j = \frac{4^{\log_4 N} - 1}{4 - 1} \sim N \tag{93}$$

for a fixed desired accuracy. In practice, this accuracy can be made greater than machine precision, resulting in an algorithm that is, for all intents and purposes, exact.

6.4 Particle-in-Cell Methods

Even with the fast multipole algorithm, the number of Coulombic particles that may be simulated on a computer is far less than the corresponding number for short-range force laws. There is thus great interest in approximate methods with reduced computational complexity. An important example of such a method is the *particle-in-cell* algorithm. This is a hybrid Eulerian/Lagrangian technique, in which one keeps track of particle coordinates, but uses a grid to compute the fields by which these particles move.

As an example, let us consider a particle-in-cell simulation of a pure electron plasma confined by magnetic and electrostatic fields in a Penning trap (Chu, 1993). The plasma is confined in cylindrical geometry, with the magnetic field in the z direction. Each particle undergoes a very rapid gyration in the xy plane due to the magnetic field. Superposed on this is a bounce motion in the z direction due to the confining electrostatic potential. If we average over the cyclotron and bounce periods, we can consider the evolution of the density of gyration/bounce centers as these drift perpendicularly to the magnetic field,

$$\frac{\partial n(\mathbf{x},t)}{\partial t} + \mathbf{v}_{\mathbf{E}\times\mathbf{B}} \cdot \nabla n(\mathbf{x},t) = 0. \tag{94}$$

Here we have defined the $\mathbf{E} \times \mathbf{B}$ drift velocity,

$$\mathbf{v}_{\mathbf{E}\times\mathbf{B}} \equiv c\mathbf{E} \times \mathbf{B}/B^2 = (c/B)\hat{\mathbf{z}} \times \nabla\Phi(\mathbf{x},t), \tag{95}$$

which is a well-known systematic drift of charged particles confined in crossed electric and magnetic fields. In this case, the electric field is due to the plasma itself,

$$\nabla^2\Phi(\mathbf{x},t) = -4\pi(-e)n(\mathbf{x},t). \tag{96}$$

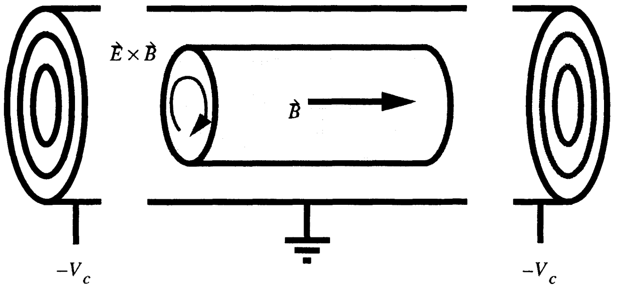

FIG. 10. Penning trap for pure-electron plasma. The plasma is confined radially by an axial magnetic field **B**, and axially by an applied electrostatic potential v_c. It drifts azimuthally with the drift velocity **E** × **B**.

The boundary condition is that the potential must be zero at the grounded outside wall. This is illustrated in Fig. 10.

In a particle-in-cell simulation, we keep track of the position coordinates of each particle. Since we have averaged over the motion in the z direction, our problem is really two dimensional, and so we keep two coordinates (e.g., r and θ) for each particle. At each time step we "smear" these particles onto a grid to get a charge density. This is done by the linear interpolation scheme illustrated in Fig. 11. Note that a polar-coordinate grid is used. The charge is divided among the four grid points at the corners of its cell, and we then divide by the area of that cell to get the local charge density.

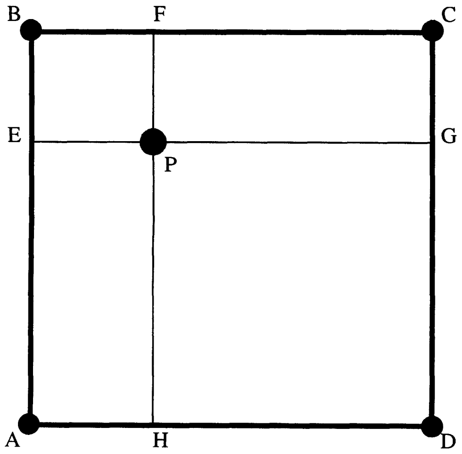

FIG. 11. Linear interpolation of particles to grid. The particle at position P contributes a fraction $a(PFCG)/a(ABCD)$ of its charge to grid point A, $a(PDGH)/a(ABCD)$ to grid point B, $a(PHAE)/a(ABCD)$ to grid point C, and $a(PEBF)/a(ABCD)$ to grid point D; here $a(ABCD)$ denotes the area of the quadrilateral $ABCD$.

Once the charge density is known on the grid, we can solve the Poisson equation, Eq. (96), to get the potential on the grid. This may be done by Fourier transformation in the θ direction (since it is periodic), and by solving a tridiagonal system of linear equations in the r direction. By finite differencing this potential, we can get the electric-field components on the grid. Using the same weighting factors that were used to spread the charge from the particles to the grid points, we can now interpolate to get the field at each particle's position. Each particle can then compute its local **E** × **B** drift velocity and take a step along the characteristics of Eq. (94),

$$\frac{d\mathbf{x}}{dt} = \mathbf{v}_{\mathbf{E}\times\mathbf{B}}. \tag{97}$$

The resulting simulation technique is capable of following the evolution of the plasma as it breaks down into wispy filaments that would be very difficult to capture in a purely Eulerian or spectral representation. Plate 1 illustrates a particle-in-cell simulation with 131 072 particles moving on a spatial grid of dimension 256 × 256.

A wide body of literature is available on this simulation technique. A particularly nice review of plasma-physics applications of the particle-in-cell method can be found in the book by Birdsall and Langdon (1985).

6.5 Lattice-Gas Hydrodynamics

In this subsection, we review a procedure for simulating the *nonequilibrium* behavior of viscous fluids. This technique, called *lattice-gas hydrodynamics,* deserves special mention because it introduces an entirely new paradigm for using the computer to simulate a physical system. Ironically, it is a return to von Neumann's *cellular automata* computational paradigm.

Hydrodynamic lattice gases take the same "minimalist" approach to fluids that the Ising model takes to ferromagnets. The object is not a precise model of the dynamics at the finest scales, but rather to invent a fictitious microdynamics whose *coarse-grained* behavior—in the thermodynamic limit—lies in the same universality class as the phenomenon under study.

At first glance, this approach is not unreasonable. Many different fluids found in nature obey the Navier–Stokes equations of viscous fluid dynamics; for example, oil and air both obey these equations in the proper regime, in spite of the fact that their molecular dynamics are radically different. Clearly there is something universal about the Navier–Stokes equations in the hydrodynamic limit. Indeed, kinetic theory tells us some of the reason for this: The *Chapman–Enskog* expansion in the Knudsen number, by which one derives the form of the hydrodynamic equations from the Boltzmann kinetic equation, yields the Navier–Stokes equations for a wide variety of fluids whose particles interact with short-range forces, as long as those interactions conserve mass and momentum.

A lattice gas is then an attempt to define the simplest system of interacting particles on a lattice, whose collisions conserve mass and momentum, and thus whose coarse-grained behavior can be expected to obey the Navier–Stokes equations. We work on a triangular lattice (Frisch *et al.*, 1986) in two dimensions; since this has a coordination number of 6, we take a dependent variable of 6 bits per site, reflecting the presence or absence of a particle moving along each of the six lattice vectors leaving the site. At each tick of the clock, the particles stream along their respective lattice vectors and collide with the other incoming particles at their new site. The collisions are required to conserve mass and momentum. The only nontrivial collisions for this two-dimensional model are enumerated in Fig. 12.

By taking block space and/or time averages of the velocities of these particles, one can derive a local hydrodynamic velocity. Kinetic theoretical arguments can then be used to demonstrate that this velocity field satisfies the Navier–Stokes equations. Note that this approach is fundamentally different from a finite-difference approximation of the Navier–Stokes equations. The intent here is not to start with the continuum equations and discretize them, but rather to start with a fundamentally discrete system of particles, and make them interact in such a way that the Navier–Stokes equations are *emergent,* just as they are for natural fluids. That is, lattice-gas automata (LGA), like molecular dynamics, are a particulate *simulation* of the physical system under study, rather than the *numerical solution* of hydrodynamic partial differential equations. Because of this approach, LGA share (at least) two important properties of real fluids: First, they have absolutely *exact* conservation laws—down to the last bit in the simulation. Second, approximate methods for computing their transport coefficients (i.e., predicting their viscosity using the Boltzmann molecular

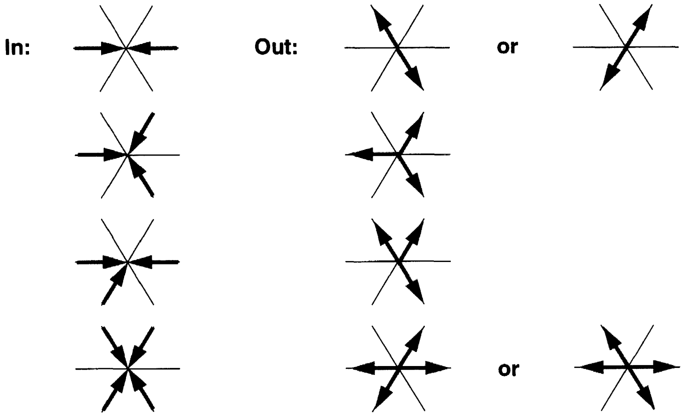

FIG. 12. Nontrivial collisions of a hydrodynamic lattice gas in two dimensions. The collisions shown conserve mass and momentum. In *all* other situations, the particles pass through each other without interacting.

chaos approximation) are often inaccurate. Consequently, practitioners of this method often have to measure such bulk properties in controlled flow situations, sometimes called *numerical experiments,* in order to calibrate runs for more interesting geometries or Reynolds numbers. This is, of course, exactly analogous to what is done by real experimentalists.

The triangular lattice is required for *isotropy* in two dimensions. Indeed, earlier versions of hydrodynamic lattice gases were tried on Cartesian grids in the 1970s (Hardy *et al.,* 1973, 1976) and were demonstrably anisotropic. The method became feasible for computational fluid dynamics only when it was shown by Frisch *et al.* (1986) that going from a Cartesian to a triangular grid did not just improve anisotropy—it eliminated it altogether! The reason for this is tied up with the symmetry properties of the lattice, and it is beyond the scope of this article. Lattices with requisite symmetry exist in three dimensions (Frisch *et al.,* 1987).

It is possible to introduce the Monte Carlo sampling of collision outcomes into this method, in order to hasten the approach to a local thermodynamic equilibrium for systems of interacting particles. By introducing multiple particle species and different types of interactions and by Boltzmann-weighting collision outcomes, immiscible fluid flow (Rothman and Zaleski, 1994), coexisting liquid and gas phases (Appert *et al.,* 1991), and amphiphilic fluids (Boghosian *et al.,* 1996) have all been modeled using this technique.

From a philosophical point of view, it is fascinating that the lattice-gas approach has effectively reduced hydrodynamics to logic. It simulates fluid motion by purely logical operations on bits—the kind of thing that a computer can do exactly, without approximation. Moreover, unlike floating-point representations of data, all of the bits present play essentially the same role in the simulation; this quality is sometimes referred to as "bit democracy," and vaguely indicates that the computer is being used in a more natural and efficient way than it is in a floating-point simulation. Indeed, the success of the lattice-gas paradigm has spurred more general studies of discrete particle mechanics on lattices. In particular, Baez and Gilliam (1994) have used ideas from algebraic geometry to generalize the methods of Lagrangian and symplectic mechanics to systems that are discrete in time, space, and dependent variable—including cellular automata, but also including more general discrete systems. They show how to construct action principles and derive an analog of the Euler–Lagrange equations for such systems. Their methods hold the promise of an entirely new way to discretize physical problems for computer implementation, and it is likely that this will be an active area of research in the coming years.

7. STATISTICAL METHODS

The study of particulate methods naturally leads us to study the statistical methods of computational physics. It is well known that stochastic methods are often useful for the estimation of multidimensional integrals. For example, students in introductory statistics courses are taught to estimate the value of π by placing points (x,y) randomly in the square $-1 \le x < +1$, $-1 \le y < +1$, and determining what fraction of the time they lie within the unit circle $x^2 + y^2 < 1$. As the number of points used goes to infinity, this fraction will approach $\pi/4$. A Fortran 90 program to implement this procedure might be

```
REAL FUNCTION ESTIMATE_PI(n)

 IMPLICIT NONE
 INTEGER :: i, n, count
 REAL :: x, y

 count = 0

 DO i = 1, n
  CALL RANDOM_NUMBER(x)
  x = 2.0*(x - 0.5)
  CALL RANDOM_NUMBER(y)
  y = 2.0*(y - 0.5)
  IF (x**2 + y**2 < 1) count = count
   + 1
 END DO

 estimate_pi = 4.0*FLOAT(count)/
  FLOAT(n)

END FUNCTION ESTIMATE_PI
```

This procedure is estimating the multidimensional integral

$$\int_{-1}^{+1} dx \int_{-\sqrt{1-x^2}}^{+\sqrt{1-x^2}} dy. \tag{98}$$

Suppose that we had checked to see if the points were in the circle of radius $\frac{1}{2}$, rather than the unit circle. Then the fraction of time that points were in the circle would have been $\pi/16$, and we could have multiplied by 16 to get π. This algorithm seems equivalent, but it is actually much worse. This is because the number of points that actually fall into the circle of radius $\frac{1}{2}$ is much smaller (by a factor of 4) than the number that fall into the unit circle, thereby worsening the statistics. In fact, it is an elementary exercise to determine that the fractional error of the result is $[(1 - p)/pn]^{1/2}$, where n is the number of samples and $p \equiv \pi R^2/4$ is the probability of the particle actually falling into the circle of radius $R \leq 1$. There is thus a strong incentive to increase p as much as possible, so the unit circle is preferable to that of radius $\frac{1}{2}$. Indeed, if we could find a way to make $p = 1$, the fractional error would vanish altogether. We could, for example, try to use the inscribed circle of a pentagon, hexagon, etc., rather than a square. Unfortunately, it is much more difficult to sample points efficiently in these polygons than in a square, so that there is a point of diminishing returns. This is, however, a special case of a more general rule: Monte Carlo methods work best when the samples can be taken from the problem's natural distribution function. We shall see this issue reappear in the following subsection.

Note that the standard deviation of the result for n points sampled goes as $n^{-1/2}$, so that something on the order of 10^8 points would be required to obtain π to four decimal places. While this may not be a particularly good way of obtaining the digits of π, it is often the only method available to estimate multidimensional integrals of the sort encountered in statistical physics. Recall that the equilibrium properties of a statistical mechanical system can be computed from the *partition function,*

$$Z = \sum_{s} e^{-\beta H(s)}, \tag{99}$$

where β is the inverse temperature, the sum extends over all possible states s of the system, and $H(s)$ is the energy of the system in state s. The challenge in computing Z is that the sum over states is typically huge. In fact, for systems with a continuum of states, this sum may be replaced by an integral, but in either case it is likely to be multidimensional so that the set of possible states s is enormous. The example of the Ising model, considered in the following subsection, illustrates this point.

7.1 Spin Systems

The *Ising model* is the prototype of a class of statistical physics models known as *spin systems.* It was developed by Lenz (1920) in 1920 as a model of ferromagnetism, and posed as a problem to his student, Ising, who solved the one-dimensional version of the model in 1925 (Ising, 1925). The model considers n "spins" described by $s_i \in \{-1, +1\}$ for $i = 1, \ldots, n$, on a lattice of D dimensions. In its simplest incarnation, the system's Hamiltonian has the form

$$H(s) = \frac{J}{2} \sum_{(i,j)} s_i s_j - B \sum_{i} s_i, \tag{100}$$

where the first sum is taken over all *neighboring pairs* of spins on the lattice. If b is the *coordination number,* the number of neighbors of any point on the lattice, then the first sum has bn terms; since this doubly counts the energy, a factor of $\frac{1}{2}$ is included with the coupling constant J in front of this sum. The second sum models an external field of strength B. The problem is to find the partition function of the system,

$$Z = \sum_{s_1=-1}^{+1} \sum_{s_2=-1}^{+1} \cdots \sum_{s_n=-1}^{+1} e^{-\beta H(s)}, \tag{101}$$

as a function of βJ and βB. While Ising's $D = 1$ solution is elementary, Onsager's $D = 2$ solution for $B = 0$ required almost another 20 years (Onsager, 1944) and is decidedly more complicated. The $D = 2$ solution for $B \neq 0$ is a relatively recent development, first published by Zamalodchikov (1989). The problem for $D = 3$ is outstanding, even for $B = 0$, and our knowledge of the behavior of Z in three dimensions rests almost entirely on numerical simulation.

At the root of the problem is the fact that the sum in Eq. (101) contains 2^n terms, and we would like to have the result in the *thermodynamic limit* as $n \to \infty$. While we may be able to perform this sum numerically for 20 spins (2^{20} terms in the summation) without much problem, it is clear that the problem is completely intractable if the spins number in the hundreds. Moreover, important equilibrium features, such as phase transitions, are sharply defined only in the thermodynamic limit. To see this, we compute the internal energy of the $D = 2$ Ising model by numerically summing Eq. (101) (and its derivative with respect to β) for 2×2, 3×3, and 4×4 lattices, and we plot these alongside Onsager's exact result in Fig. 13.

In analogy to our computation of π, we might suppose that we could *sample* the value of the internal energy by the following procedure:

1. Initialize the totals T_1 and T_2 to zero.
2. Perform the following loop N times:
   a. Randomly choose a state s.
   b. Compute the energy $H(s)$.
   c. Increment T_1 by $e^{-\beta H(s)}$.
   d. Increment T_2 by $H(s)e^{-\beta H(s)}$.
3. Estimate the energy by the ratio T_2/T_1.

This seems straightforward to implement, even for very large lattices, but it has a fatal problem: For all but the highest of temperatures, the vast majority of the total energy is accounted for by a miniscule fraction of the set of all possible states. Thus, if we sample this set *uniformly* in step **2**a of the above algorithm, we will almost never encounter the most important states, and our variance will be unacceptably high.

This is reminiscent of the problem we had when we made the radius of the circle smaller in the algorithm for calculating π in Sec. 7. Here, as there, the solution is to try to sample the states from the problem's natural distribution function. That is, we would ideally like to sample the states s with probability

$$f(s) = e^{-\beta H(s)}/Z. \tag{102}$$

Then we could simply add up the energies sampled, and divide by the total number of samples to obtain the average internal energy. It seems that we cannot do this directly, however, since the probability depends on the partition function Z, and that is, after all, one of the things that we are trying to use the Monte Carlo method to compute in the first place.

7.2 Metropolis Monte Carlo

The apparent impasse described in the last subsection was broken by an ingenious algorithm due to Metropolis *et al.* [1953]. Since then, a number of variants of this algorithm have been developed, but we describe the original one here.

The algorithm allows one to begin with a randomly selected state, but describes a Markov process for selecting all subsequent states. That is, if the system is currently in state s, the probability that this Markov process will induce a transition to a new state s'

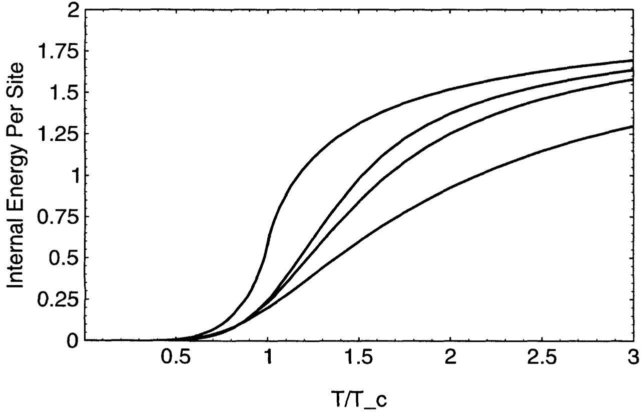

FIG. 13. Internal energy of two-dimensional Ising model. The curves are for 2×2, 3×3, and 4×4 lattices and an infinite lattice, in increasing order. The slope discontinuity at the critical temperature $T/T_c = 1$ is obtained only in the thermodynamic limit.

is called $K(s' \leftarrow s)$ and is specified by the algorithm in such a way that the time asymptotics of the Markov chain will sample the system with some desired probability function $f(s)$. It accomplishes this by demanding that the Markovian dynamics obey the condition of *detailed balance,* which ensures that it will converge to a thermodynamic equilibrium with distribution f. This condition may be written

$$f(s)K(s' \leftarrow s) = f(s')K(s \leftarrow s'). \tag{103}$$

The Metropolis selection process occurs in two steps: In the first, the new *trial state* s' is selected with some probability $T(s' \leftarrow s)$. In the second, the trial state is accepted with probability $A(s' \leftarrow s)$; otherwise, it is rejected and the system returns to the original state s. Thus, we may write

$$K(s' \leftarrow s) = T(s' \leftarrow s)A(s' \leftarrow s), \tag{104}$$

so that the condition of detailed balance becomes

$$f(s)T(s' \leftarrow s)A(s' \leftarrow s) = f(s')T(s \leftarrow s')A(s \leftarrow s'), \tag{105}$$

which may in turn be written

$$\frac{A(s' \leftarrow s)}{A(s \leftarrow s')} = \frac{f(s')T(s \leftarrow s')}{f(s)T(s' \leftarrow s)} \equiv Q(s' \leftarrow s). \tag{106}$$

Note that the identity

$$Q(s' \leftarrow s) = \frac{1}{Q(s \leftarrow s')} \tag{107}$$

is manifest from the definition of Q. We would like to be able to specify the desired distribution f; this is usually a Gibbsian equilibrium distribution for statistical physics applications. We would also like to specify the probability T of choosing a trial state in some convenient fashion. Thus, we would like to regard Q as an independent variable and demand that the acceptance probabilities satisfy the detailed-balance condition. To this end, we demand a functional relationship between the acceptance probability A and the independent variable Q; that is, we demand

$$A(s' \leftarrow s) = F(Q(s' \leftarrow s)) \tag{108}$$

for some function F. It immediately follows that

$$\frac{F(Q(s' \leftarrow s))}{F(1/Q(s' \leftarrow s))} = Q(s' \leftarrow s), \tag{109}$$

or, since this should hold for any Q, we look for functions F that obey

$$F(z) = zF(1/z) \tag{110}$$

for all $z \geq 0$. Fortunately, there is no shortage of functions that obey this condition. The one most often used in simulations of the Metropolis algorithm is

$$F(z) = \min(1,z), \tag{111}$$

though

$$F(z) = z/(1 + z) \tag{112}$$

also works, and there are many other possibilities as well.

We can now give a full statement of the algorithm for the sampling of a Gibbsian distribution, such as are encountered in statistical physics. The distribution is of the form of Eq. (102) for a system with Hamiltonian H and partition function Z. Since we do not know the latter, we cannot sample directly from this distribution. We can, however, sample a trial state from $T(s' \leftarrow s)$ and form the quantity Q described above, since this involves only ratios of the distribution f. In the simplest algorithms of this sort, the selection of a trial state is subject only to the restriction that

$$T(s' \leftarrow s) = T(s \leftarrow s'), \tag{113}$$

in which case

$$Q(s' \leftarrow s) = f(s')/f(s) = e^{-\beta[H(s')-H(s)]}. \tag{114}$$

The acceptance probability is then

$$A(s' \leftarrow s) = F(Q(s' \leftarrow s)) = \min(1,e^{-\beta[H(s')-H(s)]}). \tag{115}$$

The Markov process is then simple to describe: If one is in state s and proposes to move to a new state s', one first computes the change in energy $\Delta H(s' \leftarrow s) = H(s') - H(s)$ associated with this move. If it is negative—so that the move is to a state with less energy—one accepts the move unconditionally. If, on the other hand, $\Delta H(s' \leftarrow s)$ is positive—so that the move is to a state with greater energy—one accepts this move only with a probability $\exp[-\beta \Delta H(s' \leftarrow s)]$. Alternatively stated, one generates a uniformly distributed random number $r \in [0,1]$ and accepts the move to state s' only if $r < \exp[-\beta \Delta H(s' \leftarrow s)]$ [since this will always be true if $\Delta H(s' \leftarrow s) < 0$]. Note that for high temperatures this criterion is likely always to be accepted, so the sampling of states is really uniformly distributed; for low temperatures, on the other hand, the algorithm actively prevents one from visiting states with energies that are too high. In practice, since the randomly selected starting state is not likely to be particularly probable, one must "thermalize" this algorithm for some number of Markov steps before beginning to measure expectation values.

Using the Metropolis algorithm one can, for example, determine the equilibrium properties of a gas of classical molecules with arbitrary interaction potential. Recall that an analytic treatment of this problem results in an Ursell–Mayer cluster expansion (see, for example, Reichl, 1980). A serial algorithm for this problem is easy to describe: Begin with a random configuration of molecules, and try to move them at random, one at a time. For each proposed move, calculate the change of energy that would result and accept or reject that move according to the Metropolis criterion. After a while, the molecules will reach thermal equilibrium, and one can begin to take samples of any thermodynamic quantity of interest.

In recent years, a wide variety of improvements on this basic algorithm have been proposed. The most basic of these is the *biased sampling* of trial configurations, favoring those that are most consistent with the natural distribution function of the system under study—to the extent that it is known. This generally gives up the equality of $T(s' \leftarrow s)$ and $T(s \leftarrow s')$, but this can be done if their ratio is included in the Q factor that is used to determine acceptance, as indicated in Eq. (106).

The logical conclusion of the biased sampling procedure would be to bias the generation of new configurations toward the natural distribution function so successfully that every trial configuration is accepted with unit probability. Remarkably, this is sometimes possible. Swendsen and Wang (1987) showed how it could be done with the Ising and Potts models by sampling only those new configurations that differ from the original by the flipping of *clusters* of like spins. In this way they were able to reduce dramatically the problem of *critical slowing down* that plagues these methods for systems near a critical point. At criticality the correlation length tends to diverge, and so particles can be correlated with other particles that are arbitrarily far away. In this situation, the random flipping of individual spins is a "bull in a china shop" approach: Far more often than not, it badly damages the sensitive network of natural correlations and results in the trial configuration being rejected. Since the particles can be correlated on all scales, the solution is to flip them in clusters at corresponding scales, and this is precisely what Swendsen and Wang did. The sampling of the clusters can be accomplished by laying down "bonds" between each site with a probability that depends on the energy of the bond and the temperature; sites connected by an unbroken network of bonds are then said to be in the same cluster. For the Ising model, for example, the probability of such a bond is zero if the spins are not aligned, and $1 - e^{-2\beta J}$ if they are. Once the clusters are determined, they are flipped with probability $\frac{1}{2}$ (in the unmagnetized version of the model). All of this constitutes the sampling of a new configuration, which is then unconditionally accepted.

Subsequent work by Kandel and Domany (1989) generalizes the Swendsen–Wang procedure by introducing a hierarchical cluster algorithm capable of dealing with a wide variety of Hamiltonians. More recently, Machta *et al.* (1995) has developed an *invaded cluster* algorithm that is capable of sampling a critical system without prior knowledge of the critical temperature, and with no apparent critical slowing down.

7.3 Lattice Gauge Theory

There is a close similarity between statistical physics and quantum physics, since the Feynman path-integral formulation of the latter subject tells us that the complex amplitude for a particle at point $\mathbf{x}_1$ at time t_1 to propagate to point $\mathbf{x}_2$ at time t_2 is

$$Z[(\mathbf{x}_2,t_2) \leftarrow (\mathbf{x}_1,t_1)] = \int d[\mathbf{x}(t)] \exp\left[i \int_{t_1}^{t2} dtL(\mathbf{x},\dot{\mathbf{x}})\right], \tag{116}$$

where the outer *path integral* is over all paths connecting the two points in space-time, and L is the Lagrangian. Since a classical field $\psi(\mathbf{x},t)$ is just a function of the space-time coordinates, the analog of the above for a field is just

$$\mathscr{Z}[\psi_2(\mathbf{x},t) \leftarrow \psi_1(\mathbf{x},t)] = \int d[\psi(\mathbf{x},t)] \exp\left[i \int dt \int d^3x \mathscr{L}(\psi,\nabla\psi,\partial_t\psi)\right], \tag{117}$$

where $\mathscr{L}(\psi,\nabla\psi,\partial_t\psi)$ is the Lagrangian density of the field, so that $L = \int d^3\mathscr{L}$. One can then take expectation values of any (time-ordered) quantity by weighting it with the complex measure

$$\frac{1}{\mathscr{Z}[\psi_2(\mathbf{x},t) \leftarrow \psi_1(\mathbf{x},t)]} \exp\left[i \int dt \int d^3x \mathscr{L}(\psi,\nabla\psi,\partial_t\psi)\right], \tag{118}$$

in exact analogy to Eq. (102). In fact, by analytically continuing the time coordinate to the complex plane, $t = ix_4$, we can write the above in the form

$$\frac{1}{\mathscr{Z}} \exp\left[-\int d^4x \mathscr{L}(\psi,\boldsymbol{\partial}\psi)\right], \tag{119}$$

where $\mathbf{x}$ is now understood to be a space-time coordinate, and $\boldsymbol{\partial}$ is the four-gradient. The appearance of the manifestly covariant space-time volume element d^4x makes this formulation natural for relativistic field theories.

By discretizing the Lagrangian density onto a lattice, we can extract vacuum expectation values of any quantity of interest using the Metropolis procedure. Just as we did when introducing the finite-element method, we restrict the independent variable to a lattice $\mathbf{x_j} = \mathbf{j}\Delta x$, where Δx is the lattice spacing. The four-gradient is then approximated by

$$\partial_\alpha\psi(\mathbf{x_j}) \approx \frac{1}{\Delta x}[\psi(\mathbf{x}_{\mathbf{j}+\hat{e}_\alpha}) - \psi(\mathbf{x_j})]. \tag{120}$$

Gauge field theories, such as *quantum electrodynamics* (QED) or *quantum chromodynamics* (QCD), are associated with an underlying Lie group G; these are U(1) for QED and SU(3) for QCD. A set of group generators, denoted by λ^a, obey closed commutation relations,

$$[\lambda^a,\lambda^b] = if^{abc}\lambda^c, \tag{121}$$

where the f^{abc} are the structure constants of G (for this choice of generators). For example, if G is $SU(2)$, we could take the λ's to be (half) the Pauli matrices, in which case the structure constants are just the components of the Levi–Civita tensor $f^{abc} = \epsilon^{abc}$. An arbitrary element of the Lie algebra of G may then be expanded in terms of these generators; thus, the *gauge potentials,* which take their values in the Lie algebra, may be written

$$\mathbf{A}(\mathbf{x}) = \sum_a \mathbf{A}^a(\mathbf{x})\lambda^a. \tag{122}$$

Let us discretize the gauge field by storing $\mathbf{A}(\mathbf{x})$ at each lattice site. Exponentiation of the elements of the Lie algebra gives elements of G, so that

$$U(\mathbf{j},\alpha) \equiv \exp\left[\frac{ig\Delta x}{2} \sum_a \lambda_a \hat{\mathbf{e}}_\alpha \cdot \mathbf{A}^a(\mathbf{x_j})\right] \tag{123}$$

is an element of G that can be *associated with the link* connecting lattice site $\mathbf{x_j}$ with its neighboring site $\mathbf{x_j} + \hat{\mathbf{e}}_\alpha$. Here g is the coupling constant of the theory. The Lie group elements associated with the links have the virtue that their product around any closed path on the space-time lattice is gauge invariant. In terms of these quantities, Wilson (1971) showed that the discretized action ($S \equiv \int d^4\mathscr{L}$) of quantum chromodynamics could be written

$$S = -\frac{1}{g^2} \sum_{\text{plaquettes}} \text{Trace}(U_p), \tag{124}$$

where U_p is the product of the four elements of G around the edges of a closed unit square, or *plaquette,*

$$U_p = U(\mathbf{j},\alpha)U(\mathbf{j} + \hat{\mathbf{e}}_\alpha,\beta)U(\mathbf{j} + \hat{\mathbf{e}}_\alpha + \hat{\mathbf{e}}_\beta, -\alpha)U \times (\mathbf{j} + \hat{\mathbf{e}}_\beta, -\beta), \tag{125}$$

and where the sum is taken over all such plaquettes on the space-time lattice. Note that the U matrices are associated with *directed* links, so that

$$U(\mathbf{j} + \hat{\mathbf{e}}_\beta, -\beta) = U^\dagger(\mathbf{j},\beta). \tag{126}$$

One then computes expectation values of operators O depending on the $\mathbf{U}$ matrices (any gauge-invariant operator can be thus expressed) by

$$\langle O(\mathbf{U})\rangle = \frac{1}{Z}\int d[\mathbf{U}]e^{-S}. \tag{127}$$

This is exactly the kind of quantity that we were computing in Sec. 7.2, in the context of statistical physics, and can thus be estimated using the Metropolis procedure. Changing the gauge potential at any lattice point affects only those plaquettes that actually pass through the point, so that the change in the lattice action is a straightforward local calculation.

Because the coupling constant g of QCD is large, it is not useful as an expansion parameter, as it is in QED. Thus, many of the known consequences of QCD have been determined by numerical simulations of the sort described above. In particular, the form of the interquark potential has been estimated in this way, as have baryon masses, and the method has been used to investigate the elusive "glueball." A large literature exists on algorithms for lattice gauge theory, and on their supercomputer implementation; see Dasgupta *et al.* (1996) for details on the Fortran 90 implementation.

8. THE FUTURE OF COMPUTATIONAL PHYSICS

8.1 The State of the Art

In the previous sections, we have traced the evolution of the relationship between computation and physics. By implementing many of the methods described on the largest supercomputers available today, extremely large computations have been performed. Some of the largest computations thus performed have been compiled into a list of "World Scientific Records" by Stauffer (1996), and we mention some highlights here:

- At the time of this writing, the largest $D = 2$ Ising simulations on record employ a $1\,000\,000 \times 1\,000\,000$ grid (10^{12} sites) (Linke *et al.*, 1995), and the largest with $D = 3$ Ising simulations employ a $3072 \times 3072 \times 3072$ grid (2.9×10^{10} sites) (Munkel *et al.*, 1993). The fastest such simulations can perform Metropolis updates on 11 billion sites per second (Ito, 1996).
- *Exact* solutions of the $D = 3$ Ising model, using computer-assisted symbolic manipulation methods for carrying out the algebra, have been accomplished for a $5 \times 5 \times 6$ lattice (150 spins) (Guttmann and Enting, 1994). (Note that a brute-force approach to this problem would involve the summation of 2^{150} terms—an impossible task.)
- As mentioned earlier, molecular-dynamics simulations with between 10^8 and 10^9 particles and with a characteristic performance of 10^{-8} s per particle update have been achieved (Zhou *et al.*, 1997) on large parallel supercomputers.
- Quantum chromodynamics simulations have achieved 7 billion updates per second (Sexton *et al.*, 1995).
- Floating-point fluid-dynamics calculations in three dimensions have been performed on $512 \times 512 \times 512$ grids (Oran and Boris, 1992).

8.2 Quantum Computation

As the above-mentioned doubling of computer performance every 1.5 years continues unabated at the time of this writing, one can expect the above performance records to continue to improve substantially. When it became clear that this could not continue forever (see Fig. 1), computer manufacturers exploited parallelism so that the trend would continue. Most of the effort in the supercomputer industry today centers on making parallel clusters of computers on fast networks easy to use; one method of doing this is to develop efficient and elegant parallel programming paradigms, embodied in high-level languages, such as Fortran 90.

In spite of this success at warding off the problem for the time being, Fig. 1 still gives one the feeling that supercomputer manufacturers and users are heading for a "brick wall" in the not-too-distant future, and raises the question of whether or not any powerful new strategies for physical computation will come to the rescue, as did parallelism in the 1980s. In a development that has many a computer scientist scrambling to (re)learn quantum physics, one of the most interesting, exciting, and potentially explosive new computational paradigms currently under consideration centers on the computational properties of quantum-mechanical systems, and is called *quantum computation.*

Recall that all of classical computation theory, as prototyped by the Turing machine model, is carried out by logical operations on sequences of bits—quantities that take their values in the set {0,1}. By contrast, the fundamental element of information in a quantum computer is the *qubit* (pronounced "Q-Bit") or *quantum bit.* Like its classical counterpart, the measurement of a single qubit can yield one of two results, $|0\rangle$ or $|1\rangle$. Unlike its classical counterpart, however, a qubit can be in a complex superposition of these two states,

$$c_0|0\rangle + c_1|1\rangle, \tag{128}$$

where the coefficients obey the normalization condition $|c_0|^2 + |c_1|^2 = 1$. To borrow from the parlance of quantum mechanics, when the qubit is measured, it "collapses" to a classical bit of value 0 with probability $|c_0|^2$, and of value 1 with probability $|c_1|^2$. As long as we do not measure the qubit, however, its complex amplitude may combine and interfere with those of other qubits.

Just as a classical computer manipulates classical bits and combines them according to various logical operations, called *gates,* a quantum computer manipulates qubits and combines them according to unitary operations, called *quantum gates.* We note that even a single qubit with a unary quantum gate is capable of behavior that is totally unfamiliar from the classical theory of computation. For example, suppose that the unary quantum gate is the application of the unitary operator

$$\frac{1}{\sqrt{2}}\begin{pmatrix} +1 & +1 \\ -1 & +1 \end{pmatrix} \tag{129}$$

in the representation of Eq. (128). If we begin with a qubit that is assuredly (perhaps because of a measurement) in state $|0\rangle$, one pass through this quantum gate will yield

$$\frac{1}{\sqrt{2}}|0\rangle - \frac{1}{\sqrt{2}}|1\rangle. \tag{130}$$

If we were to perform a measurement at this point, the result would be 0 or 1, each with probability $\frac{1}{2}$. Thus, a random-bit generator is trivially realized on a quantum computer, unlike the classical case where we had to take a great deal of care to deterministically generate sufficiently uncorrelated pseudorandom numbers.

The fact that a measurement of the qubit yields a random bit does *not* mean that the quantum gate has destroyed the information that was in the originally prepared qubit. Indeed, quantum gates can never destroy information precisely because they are unitary; in this respect they are analogous to the *reversible gates* of classical computation, described earlier in this article. Of course, the measurement process is nonunitary and would indeed destroy the information, but suppose that instead of measuring the outgoing bit, we were to pass it back through the very same quantum gate one more time. The result would be $-|1\rangle$; that is, the result, if measured, would yield 1 with unit probability. Thus, two passes through the above unary quantum gate yield the classical "NOT operation," which is perfectly reversible. This has led one author (Hayes, 1995) to refer to the above quantum gate as "the square root of NOT."

Inspired by oddities of this sort, physicists and computer scientists began to examine what computations might be like on such computers back in the 1980s (Feynman, 1982, 1986; Deutsch, 1985, 1989). A central, if somewhat ill-posed, hope was that quantum computers would somehow be capable of calculations completely inaccessible to classical computers. At first, it is rather difficult to see what is meant by this. Consider, however, the above-described computation of an actual *random* (as opposed to pseudorandom) bit; this is clearly something outside the purview of a classical computer. An even more striking example is provided by DiVincenzo's (Barenco *et al.*, 1995) observation that 2-qubit quantum gates are sufficient to build a universal quantum computer; recall that 3-bit gates were necessary to do this with classical computation.

The net result of these observations is that quantum computation is a new paradigm of computation—completely distinct from the Turing-machine paradigm. Until the investigation of quantum computation, it was implicitly believed that the axiomatization of computation by the Turing machine paradigm was a unique mathematical construction. While Turing's ideas are often described with the cartoon of a "head" sliding along a "paper tape," this was considered to be nothing more than an intuitive picture to aid the understanding of an abstract mathematical concept. From a philosophical point of view, the observation that this is *not* true, and that the power of a computational paradigm does depend at least in part on its underlying physical realization, is just as profound as the 19th century observation that Euclidean geometry is not a unique mathematical construction, and the early 20th century discovery that the most natural geometry of the real world is in fact non-Euclidean.

While these issues are undeniably interesting, they seemed to be of exclusively academic interest until Shor proved that the integer factoring problem, long suspected to be of exponential complexity on a classical computer, could be solved in polynomial time on a quantum computer (Shor, 1994). Though the details of Shor's demonstration lie outside the scope of this article, one might intuitively imagine that a problem can be subdivided into exponentially many subproblems, and these may be superposed and "parallel processed" because of the linearity of quantum evolution. Shor showed that by selectively "interfering" the results in a certain way, it is possible to extract a desired final result in a time that is exponentially faster than the corresponding classical computation.

It is impossible to overstate the impact that such a result might have. As noted earlier, simulations of N quantum particles are necessarily of a classical complexity that is exponential in N. In analogy with Shor's result, it has been shown that such simulations are possible in polynomial time on a quantum computer, and an explicit quantum computational algorithm has even been developed for this purpose (Lloyd, 1996; Boghosian and Taylor, 1997). Indeed, the use of one quantum system to simulate another has a certain intuitive appeal, as it forcefully underscores the point that *all computation is physical,* and bodes for a future in which computation is regarded as the use of one physical system to emulate another.

Nowhere, however, is the result of Shor's observation more acutely felt than in the field of encryption. Though encryption algorithms are not commonly thought of as part of computational physics, and are therefore outside the scope of this article, suffice it to say that most all business and intelligence applications of encryption are predicated on the exponential difficulty of factoring large numbers. It is mostly for this reason that there has been an explosion of work in this area in the few years since Shor's article appeared.

Given that a quantum computer is demonstrably able to do these things, one may wonder why one is not being built today. In fact, a number of groups are actively trying to do exactly this, but it is difficult. The difficulty lies in the great care that must be taken to preserve the unitary evolution of the qubits involved. If the environment of the computer performs any "measurement" whatsoever on the qubits before they have completed their unitary evolution, they will collapse in a nonunitary fashion, and results that depend on their interference will no longer follow. In other words, the qubits must be kept coherent for the duration of the computation. This consideration sets upper bounds on the length of a computation that may be performed on a quantum computer, and hence the number of cycles in any algorithm that may be carried out. This bound is analogous to that of sensitivity to errors in trajectories such as we encountered in our earlier discussion of classical, reversible "billiard-ball" computers; since that problem had a solution, we might expect that this one does as well. Indeed, within the past year a number of interesting results have emerged on *quantum error correction* (Shor, 1996), and it has been shown that the reconstruction of a *damaged* qubit is possible in principle. It is hoped that this will provide some relief from the coherence requirements. In the meantime, researchers are trying to use a number of different physical systems to realize quantum computation (see, e.g., Turchette *et al.*, 1995; Mattle *et al.*, 1996; Cory *et al.*, 1997; Wineland *et al.*, 1997; and references therein).

9. CONCLUSION

We have surveyed the history and speculated about the future of the relationship between physics and computation. Though it is impossible to achieve completeness in any general survey of this sort, our intent has been to focus on examples that illustrate the principal issues involved, with emphasis on recent developments. We have surveyed physical computation, computational complexity, algorithms, languages, applications of computation to physics, and applications of physics to computation. Like any other important branch of physics, we have found that computational physics is interesting on many different levels; like any of the great revolutions of physics, we see that some of its most significant challenges and opportunities lie in questioning foundations and tenets long imagined to be inviolable.

ACKNOWLEDGMENTS

The author would like to thank Peter V. Coveney, Norman H. Margolus, Claudio Rebbi, and Francis Alexander for their careful critical reading of the draft version of this article and for the valuable comments they provided.

GLOSSARY

Abacus: Mechanical computing aid developed in Middle East at least 2500 years ago in which numerical data are represented by beads sliding on wires.

Analytical Engine: The first programmable general-purpose mechanical digital computer, designed by Charles Babbage in England in the mid-19th century.

Banded Matrices: Sparse matrices whose elements vanish except on certain diagonals, and which arise naturally from finite-difference problems.

Bit: Variable taking its value from a set of two elements, usually denoted {0,1} or `{false,true}`. The smallest unit of digital information.

Byte: Eight bits.

Cellular Automaton: Dynamical system that is discrete in space, in time, and in its dependent variable.

Church–Turing Hypothesis: Hypothesis, named after Alan Turing and Alonzo Church, that any universal computer can simulate and be simulated by a Turing machine.

Courant Condition: Condition for numerical stability when integrating partial differential equations, usually imposed as an upper bound on the time step.

Digital: The representation of numbers as $\Sigma_i d_i B^i$, where $d_i \in \{0, \ldots, B-1\}$. Here B is called the *base* of the representation.

Dissipative Particle Dynamics (DPD): A particulate method for accelerating molecular-dynamics simulations, using idealized collisions that conserve momentum exactly, but letting the energy fluctuate within the canonical ensemble at some fixed temperature. DPD is similar to the Nosé–Hoover method, but conserves momentum.

Finite-Difference: Method of integrating partial differential equations in which classical fields are discretized on a grid and partial derivatives are represented by differences of neighboring values.

Finite-Element: Method of integrating partial differential equations in which classical fields are represented as sums of localized spline functions.

Floating-Point Numbers: The computer representation of a real number by some number of bits of mantissa and exponent and by a sign bit. The format for this representation has been standardized by the IEEE.

Galerkin Approximation: Approximation at the heart of many finite-element methods that demands that the error incurred must lie outside of the space of basis functions used to represent the dependent variable.

Game of Life: Two-dimensional cellular automaton with 1 bit per site and simple nearest-neighbor rules, invented by John Horton Conway in the early 1970s, that is provably capable of universal computation.

Gödel's Incompleteness Theorem: Theorem stating that any axiomatic system capable of representing truths of number theory is necessarily *incomplete,* in that it contains truths that cannot be proven from its axioms.

Halting Problem: "Computer science version" of the Gödel incompleteness theorem, stating that there is no *a priori* way of knowing whether or not a given computer program is a legitimate algorithm, in that it is guaranteed to produce results and halt.

Ising Model: Prototypical statistical spin system defined on a grid at each site of which there is 1 bit of data, corresponding to a spin pointing either up or down. The system Hamiltonian is defined to be quadratic in the products of neighboring spin states.

Lattice Boltzmann Method: Similar to **Lattice-Gas Automata,** but keeps track of discrete-velocity, single-particle distribution function only, and lacks kinetic fluctuations. The distribution is represented by a floating-point number for each of the velocities present in the model.

Lattice-Gas Automata: Particulate hydrodynamic method in which particles are represented by individual bits moving about on a lattice, and suffering collisions that maintain exact mass and momentum conservation. Can be used as an alternative to molecular dynamics for the study of universal properties that depend on the presence—but not the details—of an underlying kinetic model.

Lattice Gauge Theory: Quantum gauge field theory on a lattice. Provides a Monte Carlo algorithm for sampling the expectation values of such theories.

Mark I: Electronic computer developed by Howard Aiken of Harvard University during World War II.

Metropolis Method: Markov chain method for the sampling of a large statistical ensemble without prior knowledge of its partition function.

Molecular Dynamics: The simulation of the dynamics of molecules to discern the bulk properties of materials.

Monte Carlo: Computational method that uses random numbers to estimate statistically a desired quantity.

Napier's Bones: Seventeenth-century computational aid that used logarithmically spaced divisions to multiply. Predecessor of the slide rule.

Nosé–Hoover Method: A particulate method for accelerating molecular-dynamics simulations that lets particles explore the canonical, rather than microcanonical, ensemble.

Operator Splitting: Method of integrating partial differential equations by splitting the time-development operator into several pieces that can be treated separately.

Parallel Computation: The coordinated use of many processors to share the work of, and thereby accelerate, a calculation.

Particle-in-Cell Method: Hybrid between grid and particle method, whereby individual particles move about according to self-consistent fields that are computed on a spatial grid.

Qubits: Elemental data of quantum computation that take their values in the Hilbert space spanned by the various classical states that the system can have. A single qubit is thus a complex superposition of $|\mathtt{true}\rangle$ and $|\mathtt{false}\rangle$.

Quantum Computation: Computational paradigm that is an alternative to the Turing machine model and that involves the unitary manipulation of *qubits.*

Slide Rule: Computational aid that uses logarithmically spaced divisions to multiply and divide. In widespread use until the 1970s.

Sparse Matrices: An $N \times N$ matrix with $O(N)$ nonvanishing elements.

Spin Systems: Statistical models, usually consisting of a lattice of "spins," each interacting with their neighbors according to some specified Hamiltonian. Closely related to cellular automata.

Symplectic Structure: Structure of the equations of classical mechanics that ensures the preservation of the invariant integrals of Poincaré, and an action principle based on a Lagrangian function. Invertible symplectic structures also give rise to a Poisson-bracket structure of the equations of motion. In recent years, numerical integration techniques have been developed to preserve the symplectic structure of the equations of motion, with the philosophy that this is probably more important than conserving other quantities such as the energy.

Theory of Computation: The theory of what computers can and cannot do, from both mathematical and physical considerations.

Turing Machine: Mathematical axiomatization of the computational process, developed by Alan Turing in the 1930s, and picturesquely described by a "machine" sliding along a "paper tape" imprinted with various symbols, and taking actions based on the values of those symbols.

Universal Computation: The capability to simulate any Turing machine.

Works Cited

Alder, B. J., Wainwright, T. E. (1957), *J. Chem. Phys.* **27,** 1208.

Alder, B. J., Wainwright, T. E. (1967), *Phys. Rev. Lett.* **18,** 988.

Allen, M. P., Tildesley, D. J. (1987), *Computer Simulation of Liquids,* Oxford: Clarendon Press.

Appert, C., Rothman, D. H., Zaleski, S. (1991), *Physica D* **47,** 85–96.

Baez, J. C., Gilliam, J. W. (1994), *Lett. Math. Phys.* **31,** 205–212.

Barenco, A., Bennett, C. H., Cleve, R., DiVincenzo, D., Margolus, N. H., Shor, P., Sleator, T., Smolin, J., Weinfurter, H. (1995), *Phys. Rev. A* **52,** 3457–3467.

Belinfante, J. G., Kolman, B. (1972), *A Survey of Lie Groups and Lie Algebras with Applications and Computational Methods,* Philadelphia: Society for Industrial and Applied Mathematics.

Benioff, P. (1982), *Int. J. Theor. Phys.* **21,** 177; *J. Stat. Phys.* **29,** 515; *Phys. Rev. Lett.* **48,** 1581.

Bennett, C. (1973), *IBM J. Res. Develop.* **6,** 525–532.

Bennett, C. (1982), *Int. J. Theor. Phys.* **21,** 905–940.

Bennett, C. (1987), *Sci. Am.* **257** (5), 108–116.

Bird, G. A. (1994), *Molecular Gas Dynamics and the Direct Simulation of Gas Flows,* Oxford: Clarendon.

Birdsall, C. K., Langdon, A. B. (1985), *Plasma Physics via Computer Simulation,* New York: McGraw-Hill.

Boghosian, B. M., Taylor, W. (1998), *Phys. Rev. E* **57,** 54–66.

Boghosian, B. M., Coveney, P., Emerton, A. (1996), *Proc. Roy. Soc. A* **452,** 1221–1250.

Cannon, L. E. (1969), *A Cellular Computer to Implement the Kalman Filter Algorithm,* Ph.D. Thesis, Montana State University.

Channell, P. J., Scovel, C. (1990), *Nonlinearity* **3,** 231–259.

Chen, S. (1998), *Ann. Rev. Fluid Mech.* **30** (to be published).

Chen, H., Doolen, G. D., Lee, Y. C., Rose, H. (1989), *Phys. Rev. A* **40,** 2850–2853.

Chu, R. Y.-Y. (1993), *Theoretical Studies of Pure Electron Plasmas in Asymmetric Traps,* Ph.D. Thesis, Department of Physics, Massachusetts Institute of Technology.

Cooley, J. W., Tukey, J. W. (1965), *Mathematics of Computation* **19,** 297.

Cory, D., Fahmy, A., Havel, T. (1997), *Proc. Nat. Acad. Sci USA* **94,** 1634–1639.

Dasgupta, I., Levi, A. R., Lubicz, V., Rebbi, C. (1996), *Comp. Phys. Comm.* **98,** 365–397.

Daubechies, I. (1988), *Comm. Pure Appl. Math.* **41,** 909.

DeRaedt, H. (1987), *Comput. Phys. Rep.* **7,** 1.

Deutsch, D. (1985), *Proc. Roy. Soc. London* **400A,** 96–117.

Deutsch, D. (1989), *Proc. Roy. Soc. London* **425A,** 73–90.

Espanol, P. (1995), *Phys. Rev. E* **52,** 1734–1742.

Feynman, R. (1982), *Int. J. Theor. Phys.* **21,** 467–488.

Feynman, R. (1986), *Found. Phys.* **16,** 507–531.

Fredkin, E., Toffoli, T. (1982), *Int. J. Theor. Phys.* **21,** 219–253.

Frisch, U., Hasslacher, B., Pomeau, Y. (1986), *Phys. Rev. Lett.* **56,** 1505.

Frisch, U., d'Humières, D., Hasslacher, B., Lallemand, P., Pomeau, Y., Rivet, J.-P. (1987), *Complex Systems* **1,** 75.

Garcia, A. L. (1994), *Numerical Methods for Physics,* Englewood Cliffs, NJ: Prentice Hall.

Gustafson, K. E., Sethian, J. A. (Eds.) (1991), *Vortex Methods and Vortex Motion,* Philadelphia: Society for Industrial and Applied Mathematics.

Guttmann, A. J., Enting, I. G. (1994), *J. Phys. A* **27,** 8007–8010.

Hardy, J., de Pazzis, O., Pomeau, Y. (1973), *J. Math. Phys.* **14,** 1746.

Hardy, J., de Pazzis, O., Pomeau, Y. (1976), *Phys. Rev. A* **13,** 1949.

Hayes, B. (1995), *Am. Sci.* **83,** 304–308.

Heermann, D. W., Nielaba, P., Rovere, M. (1990), *Comp. Phys. Comm.* **60,** 311–318.

Hillis, D. W. (1985), *The Connection Machine,* Cambridge, MA: MIT Press.

Hillis, D. W., Boghosian, B. M. (1993), *Science* **261,** 856–863.

Hoogerbrugge, P. J., Koelman, J. M. V. A. (1992), *Europhys. Lett.* **19,** 155–160.

Hoover, W. G. (1985), *Phys. Rev. A* **31,** 1695–1697.

Ising, E. (1925), *Z. Phys.* **31,** 253.

Ito, N. (1996), *Int. J. Mod. Phys. C* **7,** 99–106.

Kahan, W. (1996), *Beastly Numbers,* and *A Test for Correctly Rounded SQRT,* Technical Reports, Department of Electrical Engineering and Computer Science, University of California, Berkeley.

Kandel, D., Domany, E. (1989), *Phys. Rev. B* **43,** 8539–8548.

Kardar, M., Parisi, G., Zhang, Y. -C. (1986), *Phys. Rev. Lett.* **56,** 889.

Khuller, S., Mitchell, S. G., Vazirani, V. V. (1991), in: L. Leach Albert, B. Monien, M. Rodriguez Artalejo (Eds.), *Proceedings of the 18th International Colloquium on Automata, Languages and Programming,* Lecture Notes in Computer Science No. 510, Berlin: Springer-Verlag, pp. 728–738.

Landau, L. D., Lifschitz, E. M. (1982), *Fluid Mechanics,* Pergamon Press.

Landauer, R. (1961), *IBM J. Res. Devel.* **5,** 183–191.

Lenz, W. (1920), *Phys. Z.* **21,** 613.

Linke, A., Heermann, D. W., Altevogt, P., Siegert, M. (1995), *Physica* **A222,** 205–209.

Lloyd, S. (1996), *Science* **273,** 1073–1078.

Machta, J., Choi, Y. S., Lucke, A., Schweizer, T., Chayes, L. V. (1995), *Phys. Rev. Lett.* **75,** 2792–2795.

Margolus, N. (1984), *Physica* **10D,** 81–95.

Mattle, K., Weinfurter, H., Kwiat, P. G., Zeilinger, A. (1996), *Phys. Rev. Lett.* **76,** 4656–4659.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E. (1953), *J. Chem. Phys.* **21,** 1087.

Miller, J., Weichman, P. B., Cross, M. C. (1992), *Phys. Rev. A* **45,** 2328–2359, Appendix A.

Monaghan, J. J. (1992), *Annu. Rev. Astron. Astrophys.* **30,** 543–574.

Moore, C. (1990), *Phys. Rev. Lett.* **64,** 2354–2357.

Moore, C. (1991), *Nonlinearity* **4,** 199–230.

Munkel, C. *et al.* (1993), *Physica* **A193,** 540–552.

Nosé, S. (1986), *Mol. Phys.* **57,** 187–191.

Onsager, L. (1944), *Phys. Rev.* **65,** 117.

Oran, E. S., Boris, J. P. (1992), *Comput. Phys.* **7,** 523.

Pour-El, M. B., Richards, I. (1982), *Int. J. Theor. Phys.* **21,** 553.

Reichl, L. E. (1980), *A Modern Course in Statistical Physics,* University of Texas Press, Chap. 11.

Richardson, J. L. (1991), *Comput. Phys. Comm.* **63,** 84–94.

Rokhlin, V., Greengard, L. (1987), *J. Comp. Phys.* **73,** 325–348.

Rothman, D. H., Zaleski, S. (1994), *Rev. Mod. Phys.* **66,** 1417–1479.

Schwartzman, S. (1996), *Quantum* **6,** 4.

Sexton, J., Vaccarino, A., Weingarten, D. (1995), *Phys. Rev. Lett.* **75,** 4563–4566.

Shor, P. W. (1994), in: S. Goldwasser (Ed.), *Proc. 35th Annual Symposium on the Foundations of Computer Science,* Los Alamitos, CA: IEEE Computer Society Press, p. 124.

Shor, P. W. (1996), *Fault-Tolerant Quantum Computation,* AT&T preprint, quant-ph/9605011, and references therein.

Stauffer, D. (1996), *Int. J. Mod. Phys. C* **7,** 759–760.

Strang, G. (1989), *SIAM Rev.* **31,** 614–627.

Swendsen, R. H., Wang, J.-S. (1987), *Phys. Rev. Lett.* **58,** 86–88.

Toral, R. (1995), in: P. L. Garrido, J. Marro (Eds.), *Third Granada Lectures in Computational Physics: Proceedings of the III Granada Seminar on Computational Physics,* held at Granada, Spain, 5–10 September 1994, New York: Springer-Verlag.

Turchette, Q. A., Hood, C. J., Lange, W., Mabuchi, H., Kimble, H. J. (1995), *Phys. Rev. Lett.* **75,** 4710–4713.

Wilson, K. (1971), *Phys. Rev. B* **4,** 3174; *Phys. Rev. B* **4,** 3184.

Wineland, D. J., Monroe, C., Meekhof, D. M., King, B. E., Leibfried, D., Itano, W. M., Bergquist, J. C., Berkeland, D., Bollinger, J. J., Miller, J. (1997), *Proc. Roy. Soc. A* (to be published).

Zamalodchikov, A. B. (1989), *Int. J. Mod. Phys.* **A4,** 4235.

Zhou, S. J., Beazley, D. M., Lomdahl, P. S., Holian, B. L. (1997), *Phys. Rev. Lett.* **78,** 479–482; also see review of this article in Marder, M. (1997), *Nature* **386,** 219–220.

Further Reading

Allen, M. P., Tildesley, D. J. (1987), *Computer Simulation of Liquids,* Oxford: Clarendon Press.

Bennett, C. H. (1995), "Quantum information and computation," *Phys. Today* **48** (10), 24–30.

Birdsall, C. K., Langdon, A. B. (1985), *Plasma Physics via Computer Simulation,* New York: McGraw-Hill.

Coveney, P. V., Highfield, R. (1995), *Frontiers of Complexity,* New York: Fawcett Columbine.

Frenkel, D., Smit, B. (1996), *Understanding Molecular Simulation, From Algorithms to Applications,* New York: Academic.

Kenway, R. D., Pawley, G. S. (1987), in: *Computational Physics, Proceedings of the Thirty-Second Scottish Universities' Summer School in Physics,* Edinburgh: The Scottish Universities' Summer School in Physics.

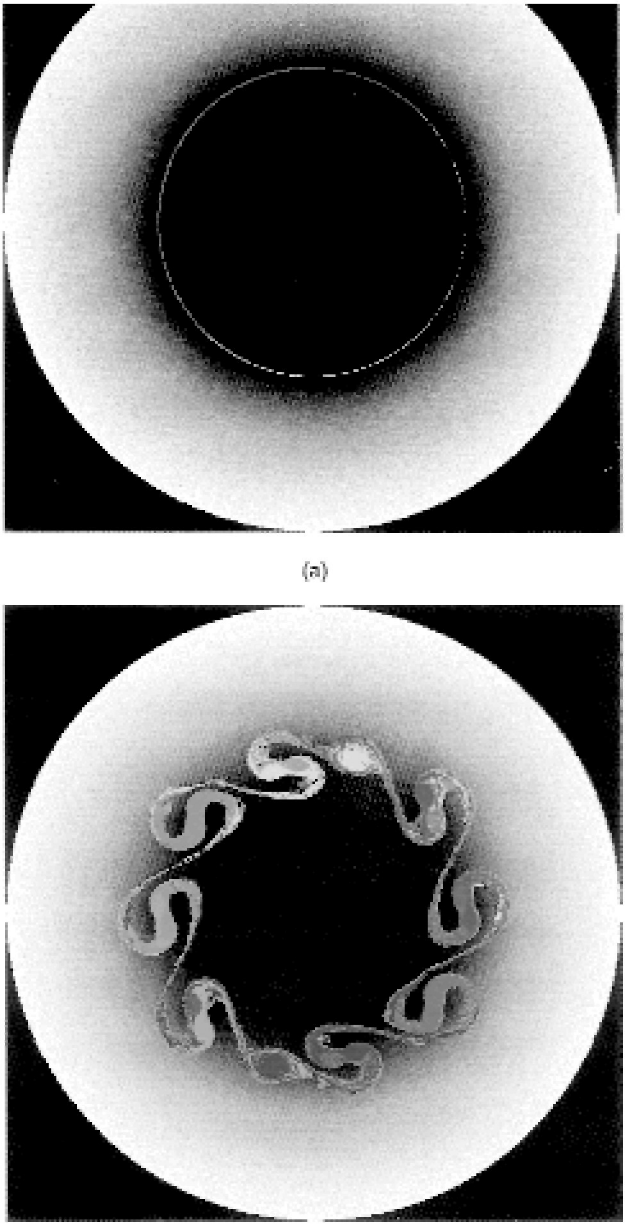

(a)

(b)

Vol. 23 – PLATE 1. Evolution of pure electron plasma in a Penning trap by the particle-in-cell method. The plasma particles are represented by colored dots and initialized on a ring of constant radius; their colors are determined according to their initial positions on the ring, and they retain their colors as they evolve. The gray scale in the background represents the electrostatic potential. (Graphics by Eric Brisson, Scientific Computation and Visualization Group, Boston University).

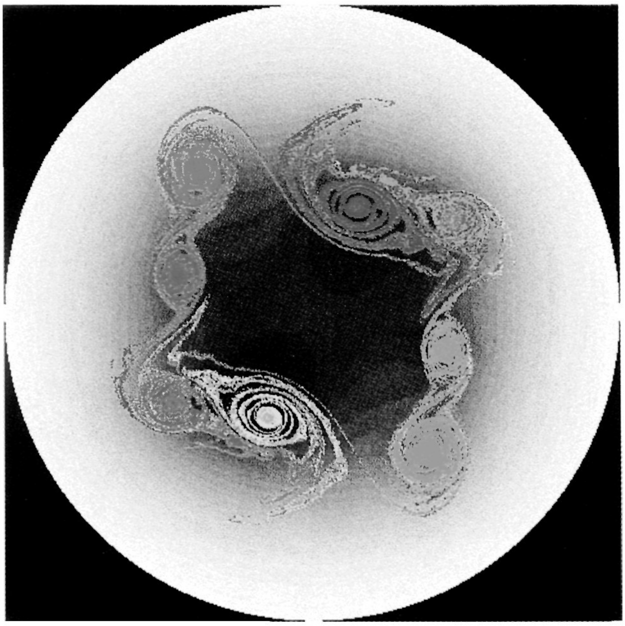

(c)

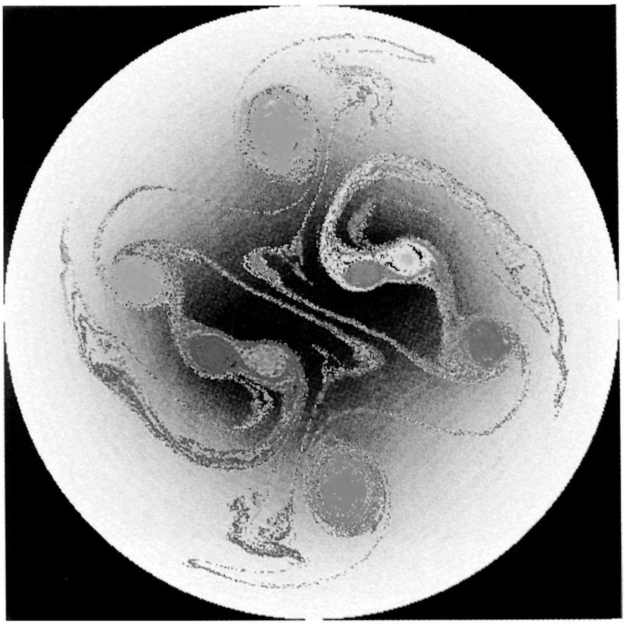

(d)

Vol. 23 – PLATE 1. Continued.

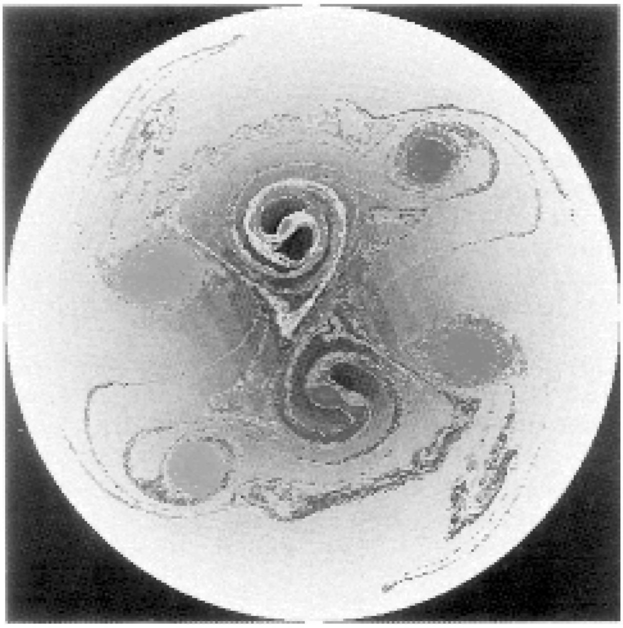

(e)

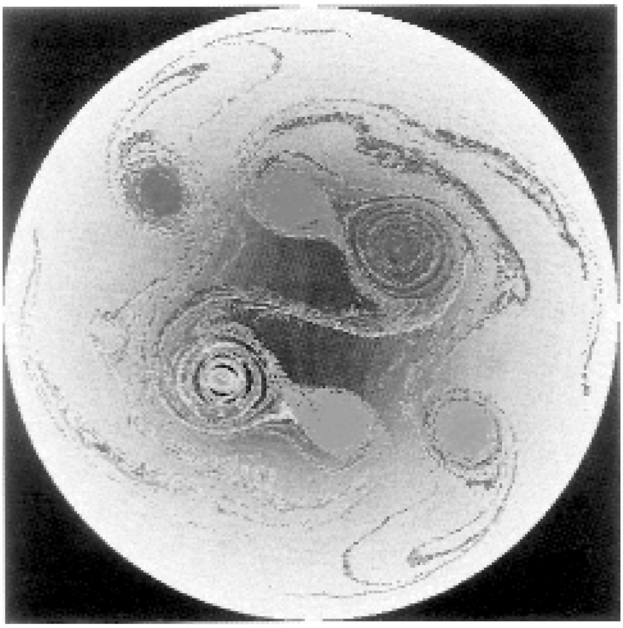

(f)

Vol. 23 – PLATE 1. Continued.

VESICLES AND BIOMEMBRANES

Reinhard Lipowsky, *Max-Planck-Institut für Kolloid- und Grenzflächenforschung, Teltow-Seehof, Germany*

Introduction 200
1. **From Cells to Vesicles** 201
1.1 Plasma Membrane and Internal Membranes 201
1.2 Evolution of Biomembranes 201
1.3 Universal Construction Principle 202
1.4 Fluidity of Biomembranes 202
2. **Molecular Structure of Lipid Bilayers** 202
2.1 Lipid Molecules in Water 202
2.2 Self-Assembly of Lipids 202
2.3 Lateral Diffusion and "Flip-flops" 203
2.4 Transport across Bilayers 203
3. **Elastic Properties of Fluid Membranes** 203
3.1 Stretching Deformations 203
3.2 Bending Deformations 204
4. **Related but Distinct Systems** 204
4.1 Soap Films 204
4.2 Surfactant Layers 205
4.3 Solidlike or Polymerized Membranes 205
5. **The Morphology of Vesicles** 205
5.1 Shape of Homogeneous Membranes 205
5.1.1 Curvature and Bending Energy 205
5.1.2 Asymmetric Membranes 207
5.1.3 Constraints on Area and Volume 207
5.1.4 Shape Transformations and Limit Shapes 207
5.1.5 Vesicles with Handles 207
5.2 Shape of Inhomogeneous Membranes 208
5.2.1 Composition and Shape 208
5.2.2 Domain-Induced Budding and Fission 208
5.2.3 Bud Size 209
5.2.4 Domain-Induced Budding of Biomembranes 209
5.3 Adhesion of Vesicles 209
5.3.1 Contact Potential 209
5.3.2 Adhesion Threshold 209
6. **Shape Fluctuations of Membranes** 210
6.1 Bending Modes or Undulations 211
6.1.1 Roughness Arising from Bending Modes 211
6.1.2 Measurement of the Bending Rigidity 211
6.2 Thermal Fluctuations on Molecular Scales 211
6.3 Persistence Length on Large Scales 211
7. **Generic Interactions of Two Membranes** 211
7.1 Direct Interactions between Rigid Membranes 211
7.2 Renormalization by Bending Undulations 212
7.2.1 Repulsive Interactions 212
7.2.2 Attractive Interactions and Continuous Unbinding 212
7.2.3 Potential Barriers and Discontinuous Unbinding 213
7.3 Tension-Induced Adhesion 213
7.4 Repulsive Interactions at Small Separations 213
8. **Bunches and Stacks of Membranes** 213
8.1 Free Membranes in Solution 214
8.2 Membranes at a Rigid Surface 215
8.3 Limit of Lyotropic Liquid Crystal 215
8.3.1 Power-Law Peaks of the Scattering Intensity 215
8.3.2 Melting of Lamellar States 216
9. **Polymer-Decorated Membranes** 216
9.1 Anchored Polymers 216
9.2 Dilute and Semidilute Regimes 216
9.3 Polymer-Induced Curvature 217

digital Encyclopedia of Applied Physics

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

9.4 Polymer-Induced Adhesion 217

10. Applied Electromagnetic Fields 218

10.1 Alternating Electric Fields 218

10.2 Electroporation 218

10.3 Electrofusion 219

10.4 Optical Tweezers 219

11. Outlook on Applications 219

Glossary 220

Works Cited 221

Further Reading 222

INTRODUCTION

The membranes and vesicles considered here are ultrathin and highly flexible sheets composed of lipids and other amphiphilic molecules (which have both a water-soluble, hydrophilic part and a water-insoluble, hydrophobic part). In biological systems, these membranes represent complex interfaces that partition space into different compartments and, thus, are responsible for the amazing architecture of these systems. One example of this architecture is shown in Fig. 1.

Biomembranes have been studied for a long time in biology, pharmacology, and medicine. On the micron scale, these membranes exhibit a unique combination of properties:

1. They form closed surfaces without edges;
2. they are highly flexible and, thus, can easily adapt their shape to external perturbations; and
3. in spite of this flexibility, they are rather robust and keep their structural integrity even for strong deformations.

This *combination of stability and flexibility* is a consequence of their molecular struc-

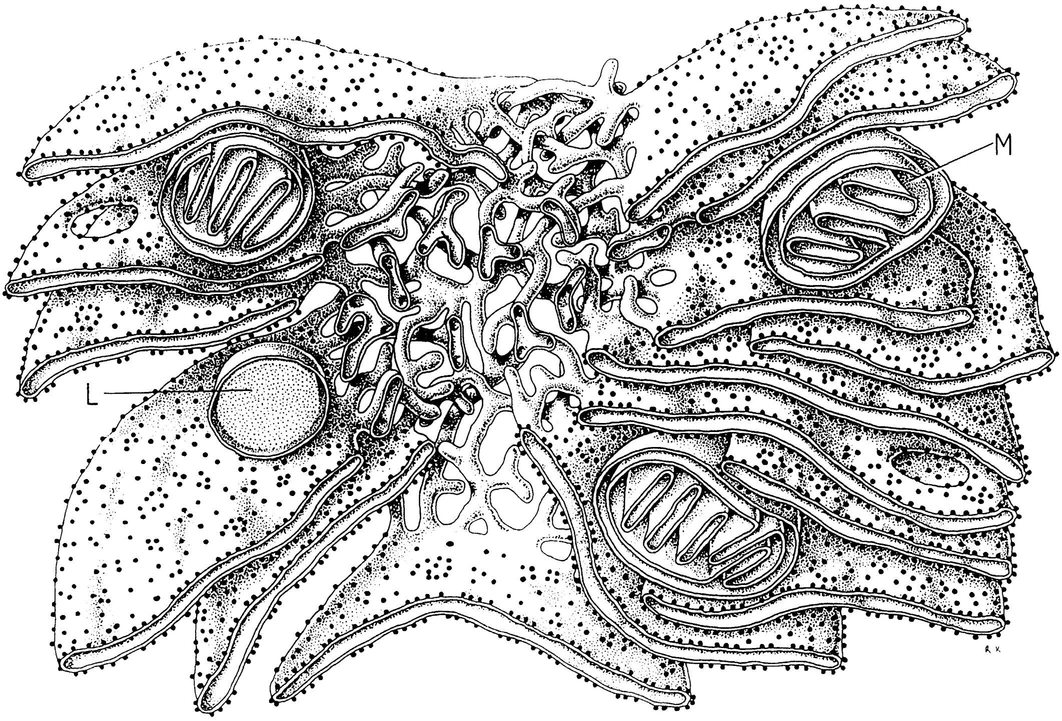

FIG. 1. Some internal membranes of a liver cell. The labyrinth of membrane sheets and tubes defines the endoplasmic reticulum; within the lamellar region, one sees the membranes of three mitochondria (M) and of one lysosome (L). (Krstic, 1976.)

ture. When viewed on the nanometer scale, each biomembrane consists of a specific mixture of many different amphiphilic molecules that reflect its diverse biological functions. However, in spite of this chemical complexity, all biomembranes are organized according to the same universal construction principle: their basic building block is provided by a *bilayer of lipid molecules.* The latter molecules are essentially *insoluble* in the aqueous solution, which ensures membrane stability. In addition, these lipid bilayers are maintained in a *fluid* state, which is the main mechanism underlying their enormous flexibility.

This article focuses on fluid bilayers that contain only one or a few lipid components. Even the simplest model membranes of this kind already exhibit several levels of *self-organization* that are reminiscent of biomembranes: The lipid molecules arrange themselves into bilayers; the bilayers spontaneously form closed bags or vesicles without edges; the vesicles adhere to one another and form various multilayer structures.

In addition, lipid bilayers in their fluid state can easily adapt to external forces by reorganization of their supramolecular structure. Indeed, one intriguing aspect of these membranes is their ability to undergo morphological changes: Depending on temperature and osmotic conditions, free membranes and vesicles exhibit a large variety of different shapes and shape transformations; similar shape changes can be induced by the formation of intramembrane domains; interacting membranes undergo transitions between bound and unbound states. A new level of self-organization is obtained when such bilayers are "decorated" with anchored polymers. These polymers form "mushrooms," "pancakes," or "brushes," which tend to curve the bilayers and to change their interactions.

Just like biomembranes, lipid bilayers are flexible but stable structures, which makes it possible to isolate them and to manipulate them in various ways: One can suck them into micropipettes, attach them to other surfaces, and grasp them with optical tweezers generated by focused laser beams. These bilayers can even tolerate local perturbations that lead to the formation of small holes: The bilayers restore their structural integrity since the holes close again spontaneously as long as the membranes do not experience a large lateral tension.

1. FROM CELLS TO VESICLES

1.1 Plasma Membrane and Internal Membranes

All living matter is built up from cells. Each cell is enclosed by its outer plasma membrane that controls the interaction between the cell and its environment. This applies both to the relatively small cells of bacteria or prokaryotes, which have no cell nucleus, and to the much larger cells of eukaryotes, which have such a nucleus. The latter class of organisms includes all animals and plants as well as single-celled microorganisms such as amoeba or yeast. In addition to the outer plasma membrane, all eukaryotic cells contain internal membranes that represent the boundaries of the internal organelles such as the nucleus, mitochondria, chloroplasts, etc. The membranes shown in Fig. 1, for example, represent the internal membranes that bound the endoplasmic reticulum and some mitochondria of a liver cell (Darnell *et al.,* 1990; Alberts *et al.,* 1994).

The total membrane area of an eukaryotic cell is relatively large. The membranes of a single liver cell, for example, have a total surface area of about 1.1×10^5 μm^2 while its volume is about 5×10^3 μm^3. About 98% of this large area belongs to the internal membranes and only 2% to the outer plasma membrane of the cell.

1.2 Evolution of Biomembranes

Membranes are quite old and thus have evolved over a long period of time. Indeed, membranes defined the boundaries of the first cells on earth and thus played a crucial role in the origin of life. The oldest microfossils that are viewed today as remnants of cells have an age of about 3.5×10^9 years. As a result of this long evolution, we now have an enormous variety of different cells and organelles, which live in very different environments. The interactions between the cells or organelles and their environments are mediated by their membranes. The basic function of these membranes is to act as highly selective barriers for the exchange of molecules between the different spatial regions and compartments. In this way, they sustain concentration gradients between their two sides, which are used, for example, in order to produce energetic molecules or to propagate localized excitations along the membranes. Likewise, these membranes act as transducers for chemical, optical, or mechanical signals and as supporting surfaces for anchored polymers and polymer networks.

1.3 Universal Construction Principle

In order to fulfill its specific biological functions, each biomembrane is composed of a specific mixture of hundreds of different molecules. However, in spite of this complex chemical composition, all biomembranes exhibit the same universal construction principle: a *bilayer of lipid molecules,* which provides a two-dimensional solvent for the hydrophobic anchors of membrane proteins. Therefore, the simplest model systems for biomembranes are lipid bilayers without any proteins. When dissolved in water, these bilayers form closed bags or vesicles that resemble the closed compartments as formed by biomembranes.

1.4 Fluidity of Biomembranes

Under physiological conditions, biomembranes are in a fluid state; i.e., the membrane molecules can diffuse rapidly along these membranes. On the molecular scale, this fluidity is necessary for the proper functioning of membrane proteins. On the supramolecular scale, it ensures that the biomembrane is highly flexible and can undergo shape changes such as the formation of small spherelike buds. Therefore, both prokaryotic and eukaryotic cells adjust the lipid composition of their membranes in such a way that they remain in a fluid state irrespective of the ambient temperature and of other external conditions. Prokaryotic cells achieve this by increasing the number of unsaturated double bonds within the hydrocarbon chains of the lipid molecules. These unsaturated bonds act as defects within the bilayer that prevent the freezing of these membranes. Eukaryotic cells, on the other hand, increase the concentration of cholesterol within their membranes in order to maintain the fluidity.

2. MOLECULAR STRUCTURE OF LIPID BILAYERS

2.1 Lipid Molecules in Water

Lipids are amphiphilic molecules with a hydrophilic head group and usually two lipophilic (= hydrophobic) hydrocarbon chains. The head group of phospholipids and glycolipids contains a phosphate group and some sugar groups, respectively.

Those lipid molecules that are contained in biomembranes are essentially insoluble in water. More precisely, single lipid molecules can be dissolved in water only up to a critical monomer concentration X_*. This concentration is very small and decreases with increasing length of the hydrocarbon chains. For phospholipids with two identical chains containing $2N_c$ carbon atoms, one has $X_* \sim \exp(-1.7N_c)$ at room temperature, as can be measured for relatively small values of $N_c \lesssim 10$ (Cevc and Marsh, 1987). For the lecithin DPPC (dipalmitoyl phosphatidyl choline) with $N_c = 16$, this leads to the estimate $X_* \sim 10^{-12}$, i.e., less than one monomer per 10 μm^3 or per 10^{-8} mL of water.

2.2 Self-Assembly of Lipids

As soon as the lipid concentration in the aqueous solution exceeds the critical monomer concentration X_*, lipid molecules assemble into supramolecular structures. These structures are built up from *lipid bilayers* in which the polar head groups of the lipid molecules form the two lipid/water interfaces whereas the hydrophobic chains are buried within the bilayer and have essentially no contact with the aqueous solution. This process is an example of the so-called hydrophobic effect, which has a large entropic contribution arising from the configurational entropy of the hydrogen bond networks within the water (Tanford, 1991). Because of this effect, the lipid bilayers also arrange themselves in such a way that they have no edges and thus form closed vesicles or liposomes.

In practice, there are several preparation methods in order to produce lipid vesicles starting with a lipid/water solution (see, e.g., Lasic, 1993). Small vesicles with a relatively narrow size distribution can be prepared from larger bilayer structures by sonification with ultrasound or by millipore extrusion. These vesicles are usually too small to be visible in the optical microscope. Relatively large vesicles with a linear size of the order of 10 μm are obtained if ordered stacks of membranes are swollen in a controlled way. In general, one then obtains a mixture of both unilamellar vesicles, consisting of a single bilayer, and multilamellar liposomes, consisting of several, closely packed bilayers.

In principle, the bilayers exchange lipid molecules with the aqueous solution surrounding them. However, since the monomer concentration X_* is so small, the exchange of molecules between the bilayers and the solution is extremely slow. As long as one considers phenomena that are fast compared to this rather slow exchange process, the chemical equilibrium that would arise from this exchange of molecules is blocked, and the number of lipid molecules within each bilayer is practically constant.

2.3 Lateral Diffusion and "Flip-flops"

Lipid bilayers are essentially two-dimensional systems: Their thickness is 4–5 nm, whereas their lateral extension often exceeds 10 μm. A two-dimensional system can exhibit distinct thermodynamic phases. Lipid monolayers at the water–air interface, for example, exhibit a large number of different phases, which have been studied by optical microscopy and x-ray and neutron scattering. Likewise, lipid bilayers always exhibit a fluid phase at high temperatures and one or several gel or solidlike phases at low temperatures. This article focuses on the *fluid states* of bilayers since these are the relevant states of biomembranes.

Within a fluid phase, the molecules can freely diffuse along the bilayer. The corresponding diffusion coefficients are typically of the order of 10^{-7}–10^{-8} cm^2/s. This implies that each lipid molecule covers of the order of 1 μm in 1 s.

While lateral diffusion within the fluid bilayer is rather rapid, "flip-flops," i.e., the transverse diffusion between the two monolayers, are much slower. In bilayers consisting of a single phospholipid, it takes usually several hours and more to exchange half of the phospholipid molecules between the two monolayers. In multicomponent bilayers, on the other hand, smaller molecules such as cholesterol can flip-flop more easily, and the corresponding time scale may be of the order of minutes.

2.4 Transport across Bilayers

Even though lipid bilayers are very thin, they provide selective barriers for the diffusive transport of molecules. Water and small uncharged molecules such as CO_2 or N_2 can permeate the bilayers: They first dissolve into their hydrophobic interior, cross it by simple diffusion, and finally dissolve into the aqueous solution on the other side of the membranes. The water permeability can be directly measured if one studies mixtures of H_2O with isotopically labeled water such as DHO or THO. On the other hand, lipid bilayers are essentially impermeable to ions and to larger uncharged molecules such as glucose. The transport of these latter molecules through biomembranes is facilitated by special membrane proteins such as ion channels and carriers. Some carriers represent ion pumps that provide an active, i.e., energy-consuming, transport mechanism against the concentration gradients (Darnell *et al.*, 1990; Alberts *et al.*, 1994).

3. ELASTIC PROPERTIES OF FLUID MEMBRANES

Fluid membranes have rather special elastic properties: Any shear deformation relaxes by flow within the membrane; i.e., the (zero-frequency) shear modulus of such a membrane vanishes. Thus, a fluid membrane exhibits only two types of elastic deformations: stretching and bending (Canham, 1970; Helfrich, 1973; Evans, 1974).

3.1 Stretching Deformations

The stretching of lipid bilayers is limited to rather small deformations, since they start to rupture as soon as their area is changed by about 1%. If a membrane segment with area $\mathcal{A}_0$ experiences the lateral tension Σ, one has the area increase $\Delta\mathcal{A}$ with $\Sigma = K_{\mathcal{A}}\Delta\mathcal{A}/\mathcal{A}_0$. The area compressibility modulus $K_{\mathcal{A}}$ is of the order of $K_{\mathcal{A}} \simeq 200$ mJ/m^2, as has been measured in micropipette aspiration experiments; see Fig. 2. By this method, one can also determine the tension of rupture, $\Sigma_{\max}$, which is of the order of a few mJ/m^2 (Evans and Needham, 1987).

3.2 Bending Deformations

Consider a planar membrane segment and let us focus on a cut perpendicular to such a membrane. If this membrane is bent, the surface of the cut is tilted; i.e., it is subject to a torque. This torque is proportional to the curvature of the membrane and to the length of the cut; the proportionality constant represents the *bending rigidity* κ, which has the dimension of an energy. For phospholipid bilayers, the bending rigidity is of the order of $\kappa \simeq 10^{-19}$ J $\simeq 20T$ as deduced from the thermally excited flickering of vesicles, see Sec. 6.1 (here and hereafter, the temperature T has energy units; i.e., T is a short-hand notation for Boltzmann constant $k_B \times$ temperature in kelvins).

As mentioned, the admixture of cholesterol to lipid bilayers increases their fluidity. At the same time, it also increases the bending rigidity of these membranes. For phospholipid bilayers, for example, the bending rigidity was measured to increase by a factor 3–4 if the bilayer contained 30% cholesterol. These combined effects of cholesterol are rather remarkable and must be the result of a long optimization process during evolution.

4. RELATED BUT DISTINCT SYSTEMS

4.1 Soap Films

The behavior of fluid membranes as discussed below is often counterintuitive. This is understandable since there is no macroscopic analog for such systems. The only fluid surfaces in our macroscopic world are soap films and soap bubbles. A soap film consists of a thin layer of water that is bounded by two monolayers of surfactant molecules. The latter molecules are also amphiphilic but they are soluble in water, at least to a certain extent: Their critical monomer concentration X_* typically varies between 10^5 and 10^7 molecules per 1 μm^3 or per 10^{-9} mL of water.

Each surfactant monolayer represents an air–water interface that is characterized by a relatively large interfacial tension. The latter tension represents the free energy per interfacial area and ensures that soap bubbles at-

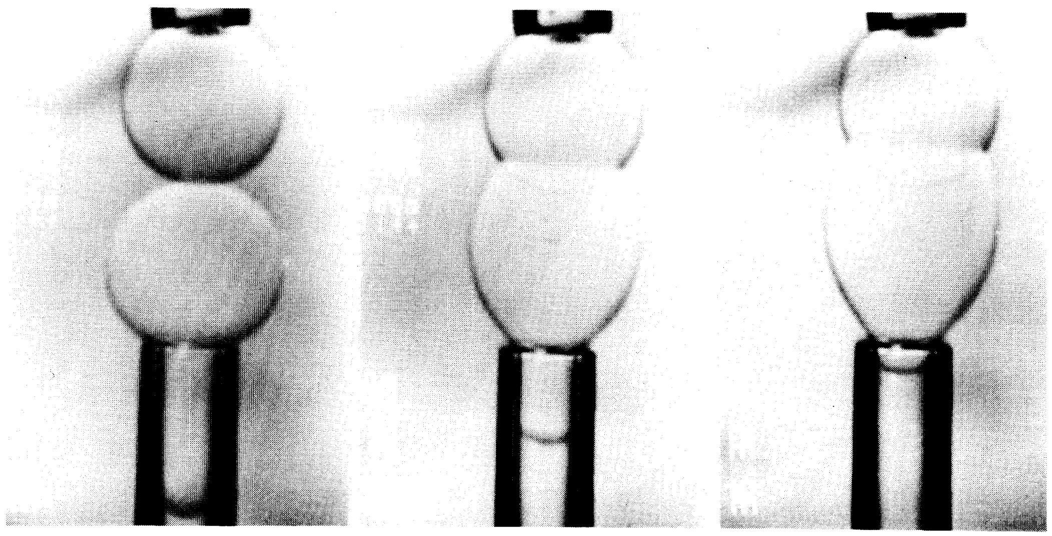

FIG. 2. Lipid vesicles may be sucked into small glass pipettes and can then be manipulated mechanically. The vesicle radii are about 10 μm. The top vesicle is almost spherical since it is exposed to a relatively large suction pressure. (Courtesy of E. Evans.)

tain a spherical shape. In addition, this interfacial tension is also responsible for the intrinsic instability of soap films: If one punches small holes into them, the tension acts to enlarge these holes and, thus, to rupture the soap films.

4.2 Surfactant Layers

Fluid monolayers of surfactant also form in multicomponent systems containing water and oil. Such mixtures exhibit many different phases in which these monolayers separate oil from water domains. Likewise, binary mixtures of water and surfactant may lead to liquid phases in which the surfactant molecules form bilayers separating two water domains. Since the critical monomer concentration X_* of the surfactant molecules is relatively large, these mixtures relax relatively fast toward states of thermal and chemical equilibrium, which are characterized by constant temperature and constant chemical potentials of the different components. These thermodynamic phases have been studied for a long time in physical chemistry and chemical engineering (for recent reviews, see Kahlweit and Lipowsky, 1996).

Surfactant layers are even more flexible than lipid bilayers, but they are also less stable and more sensitive to external perturbations. For example, it is usually not possible to transfer a single surfactant layer into a surfactant-free solution since this layer will dissolve rather rapidly. Likewise, it is hardly possible to manipulate surfactant layers by micropipettes or optical tweezers without destroying their structure. Therefore, in contrast to lipid bilayers, the behavior of *single* surfactant layers is not accessible to experimental studies.

4.3 Solidlike or Polymerized Membranes

Another class of membranes that has been recently studied in the theoretical physics community but will not be discussed in what follows consists of solidlike or polymerized membranes (for reviews, see Nelson *et al.*, 1989; Lipowsky, 1991). These latter membranes, which are characterized by a fixed connectivity of their building blocks, are more familiar since there are many examples for such membranes in our macroscopic world (such as an ordinary piece of paper, a thin film of rubber, or a fishnet). Solidlike membranes have a nonzero shear modulus, which leads to a coupling between bending and stretching deformations. Indeed, all bending deformations that change the Gaussian curvature of a solidlike membrane are necessarily coupled to strong stretching deformations of this membrane. Therefore, if the membrane (such as a piece of paper) is essentially unstretchable, it is impossible to deform it smoothly, i.e., without creating lots of folds or defects, from a planar into a spherical state. In contrast, such a shape transformation can be obtained easily for a fluid membrane; see Fig. 3.

5. THE MORPHOLOGY OF VESICLES

As mentioned, one can prepare relatively large vesicles with a linear size of the order of 10 μm that are bounded by a single bilayer. These vesicles can be directly observed in the optical microscope. One then finds that these vesicles exhibit a large variety of different shapes and various shape transformations as illustrated in Fig. 3.

A vesicle that is osmotically swollen experiences a lateral tension and, thus, attains a spherical shape just like a fluid droplet or a soap bubble. However, since water can permeate the membrane, the vesicle can adapt its volume and relax the tension. In this way, the membranes can attain an essentially tensionless state, the shape of which is governed by bending energies; compare Sec. 3 above. These shapes can be determined from systematic theories based on the relation between bending and curvature as described in the following section.

5.1 Shape of Homogeneous Membranes

First, consider a homogeneous fluid membrane with uniform elastic properties. As explained, the only elastic deformations that are relevant for fluid membranes are bending deformations governed by curvature.

5.1.1 Curvature and Bending Energy In general, each point of the membrane surface is characterized by its mean curvature

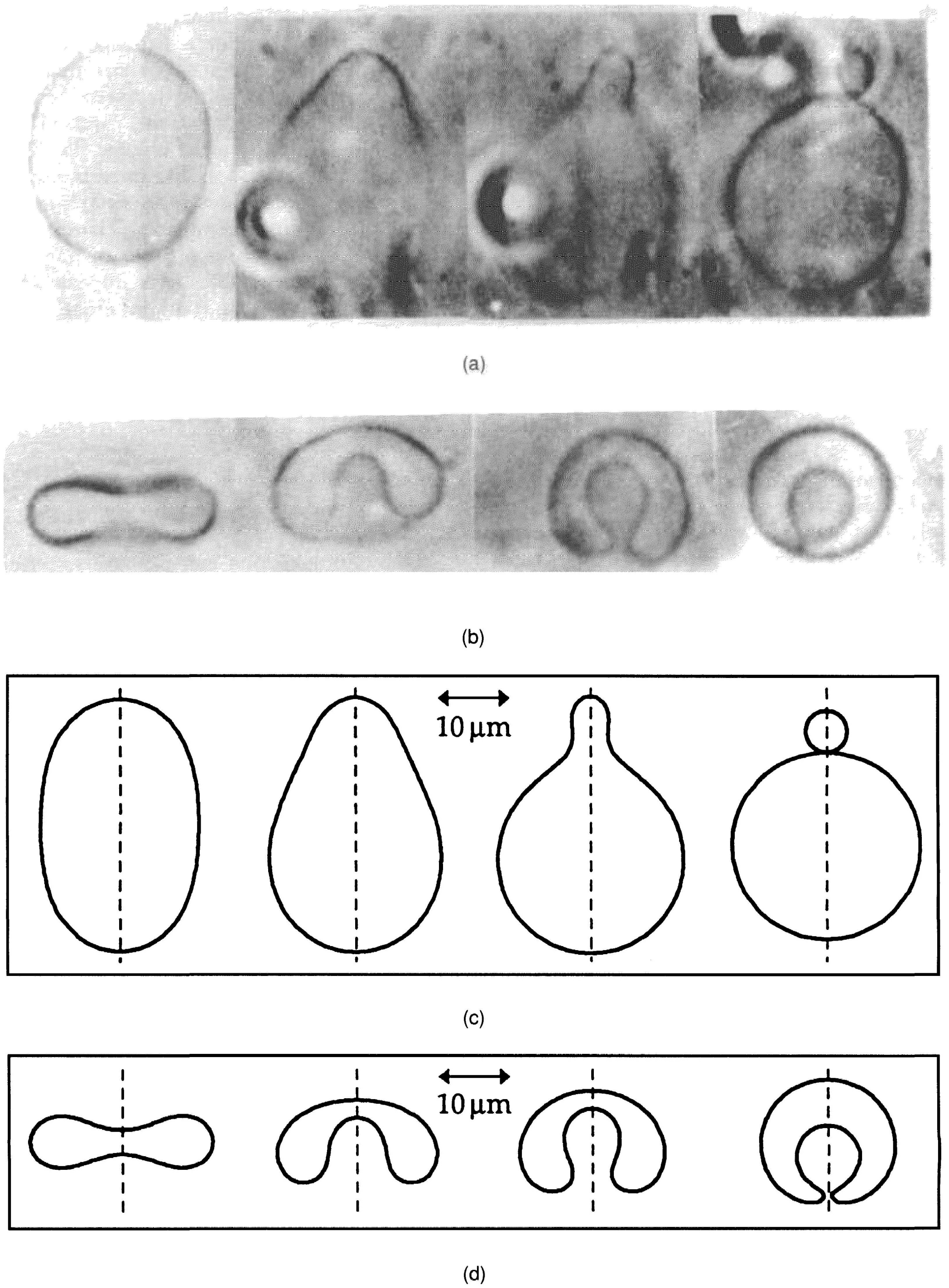

FIG. 3. Shape transformation of a single vesicle as observed via phase-contrast microscopy and calculated from curvature models. The transformations were induced by an increase in temperature that leads to an increase of the surface area of the vesicle. (a),(c) Budding, i.e., expulsion of a small vesicle from a larger one via the transformation from a prolate to a pear; (b),(d) inverse budding (or “endocystosis”) via the transformation from a discocyte to a stomatocyte. The shapes are axisymmetric with respect to the broken line. In both cases, the bud is connected to the large vesicle by a small neck.

$$M \equiv (C_1 + C_2)/2 \qquad (1)$$

and by its Gaussian curvature C_1C_2, where C_1 and C_2 are the two principal curvatures, which are equal to the inverse curvature radii.

If the two sides of the membrane are identical, the total bending energy has the form

$$\mathcal{H} = \oint d\mathcal{A}\,\tfrac{1}{2}\kappa(C_1 + C_2)^2 + \oint d\mathcal{A}\,\kappa_G C_1 C_2 \qquad (2)$$

up to second order in the principal curvatures C_1 and C_2. The surface integrals extend over the whole membrane surface, and $d\mathcal{A}$ is the intrinsic area element. The two parameters κ and κ_G have the dimensions of energy and represent the bending rigidity and the modulus of the Gaussian curvature, respectively. For closed membranes without edges, the integral over the Gaussian curvature does *not* depend on the shape of the surface but only on its topology, as follows from the Gauss–Bonnet theorem.

The mean curvature term $\sim\oint d\mathcal{A}(C_1 + C_2)^2$ is a dimensionless and thus scale-invariant quantity. In fact, it is even invariant under arbitrary conformal transformations of three-dimensional space (Willmore, 1982).

5.1.2 Asymmetric Membranes In general, the two monolayers of the bilayer membrane may differ in their chemical composition, and the two sides of the bilayer may face different surroundings. This asymmetry can be described by a spontaneous mean curvature M_{sp} (Helfrich, 1973). Furthermore, the difference in density between the two monolayers adapts locally to the curvature: If the bilayer is bent, one of the monolayers is compressed and the other one is stretched (Evans, 1974).

The coupling between the density difference and the mean curvature leads to a constraint on the total mean curvature

$$\mathcal{M} \equiv \oint d\mathcal{A} M = \oint d\mathcal{A}\tfrac{1}{2}(C_1 + C_2), \qquad (3)$$

which is proportional, for the *closed* bilayer, to the area difference of the two monolayers. This area difference can be changed by membrane proteins, which actively produce flip-flops of lipids between the two monolayers (Farge and Devaux, 1992). On the other hand, if flip-flops play no role, the total mean curvature would like to attain the preferred value, which corresponds to an unconstrained packing of the lipid molecules for which they all occupy the same (optimal) area (Miao *et al.*, 1994).

5.1.3 Constraints on Area and Volume In practice, vesicles exhibit a great variety of nonspherical shapes. This polymorphism arises, to a large extent, from two global constraints that are present for real vesicles:

1. The area of the bilayer membrane is constant (at constant temperature) since the exchange of molecules between the membrane and the solution can be neglected; and
2. the volume $\mathcal{V}$ of the vesicle does not adjust freely but is determined by the osmotic pressure P_{os} arising from those solutes that cannot permeate the bilayer; compare Sec. 2.4.

If X_{out} is the solute concentration outside of the vesicle and $\mathcal{N}_{in}$ is the number of solute molecules within the vesicle, one has $P_{os}\mathcal{V} \approx TX_{out}\mathcal{V} - T\mathcal{N}_{in}$ for small solute concentrations.

5.1.4 Shape Transformations and Limit Shapes As one changes a control parameter such as the osmotic pressure or the temperature, the shape of minimal bending energy usually evolves in a smooth way. However, for certain values of the control parameter, the shape undergoes a transformation that can be continuous or discontinuous. The shape transformations shown in Fig. 3 represent continuous transformations between a discocyte and a stomatocyte shape [Figs. 3(b) and 3(d)] and between a prolate and a pear shape at which the up–down symmetry of the shape is broken [Figs. 3(a) and 3(c)].

In addition, the shape may evolve into limit shapes in which different segments of the membrane surface start to come into contact. One type of limit shape consists of two segments connected by an infinitesimal neck that costs no bending energy; see the last shapes in Fig. 3.

5.1.5 Vesicles with Handles Recently, toroidal vesicles with one or several handles have also been studied both theoretically and experimentally. For vesicles with two or more handles, the shape of minimal bending energy is degenerate even if one includes the various constraints on the vesicle shape (Jülicher *et al.*, 1993). In this degenerate region, theory predicts a diffusion process in shape space along a conformal mode with constant area, volume, and total mean curvature as shown in Fig. 4. Such a process has been recently observed by phase-contrast microscopy (Michalet and Bensimon, 1995).

5.2 Shape of Inhomogeneous Membranes

In general, a bilayer is composed of different types of molecules. In such a multicomponent system, the composition can become inhomogeneous, which affects the elastic properties and thus the shape of the membrane. Several cooperative phenomena that are driven by this coupling between composition and shape have to be distinguished. One rather general effect is provided by the budding of intramembrane domains.

5.2.1 Composition and Shape Small vesicles that are composed of two different lipids often exhibit a strong asymmetry in the composition of the two monolayers. This compositional asymmetry reduces the mismatch or "frustration" between the spontaneous curvatures of the two monolayers. Likewise, conelike molecules within the membrane tend to diffuse toward membrane regions with an appropriate curvature. Thus, curvature may induce phase segregation.

The phase diagram of a multicomponent bilayer usually exhibits one-phase regimes and two-phase coexistence regimes depending on the temperature and the membrane composition. When such a bilayer is prepared in the two-phase regime, it will undergo phase separation, which leads to the formation of lateral domains or patches within the membrane. These intramembrane domains often form ordered patterns as originally observed by freeze fracture and electron microscopy (Sackmann, 1990). The theoretical work on these domain patterns is reviewed in Lipowsky (1995b).

5.2.2 Domain-Induced Budding and Fission Now, let us focus on a single domain within the membrane. In general, this domain may have a spontaneous curvature that differs from the spontaneous curvature of the surrounding matrix. In addition, the edge of the intramembrane domain has an energy that is proportional to its length. Therefore, the domain has a tendency to attain a circular shape in order to minimize its edge energy.

However, a *flat* circular domain does not represent the state of lowest edge energy since the length of the edge can be further reduced if the domain forms a bud: The domain edge now forms the neck of the bud, and this neck narrows down during the budding process; see Fig. 5. Because of the edge energy, the domain must bud as soon as its linear size exceeds a certain critical size even if it has no spontaneous curvature. Simple theoretical estimates also show that the bud is likely to break off from the membrane matrix. There is some evidence for such budding and fission processes from experiments on vesicles that contain mixtures of phospholipids and cholesterol.

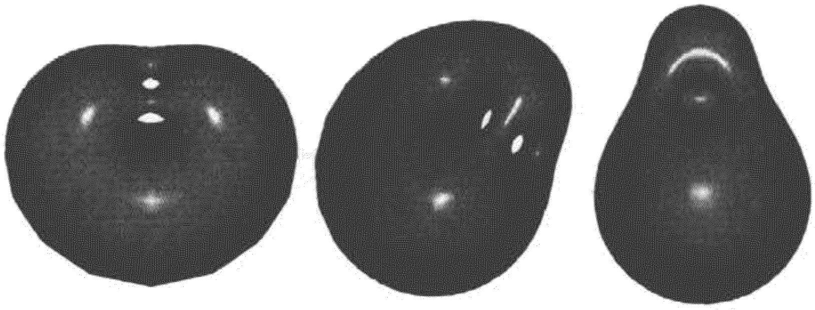

FIG. 4. Conformal diffusion of a vesicle with two handles: all three shapes have the same bending energy and the same area, volume, and total mean curvature.

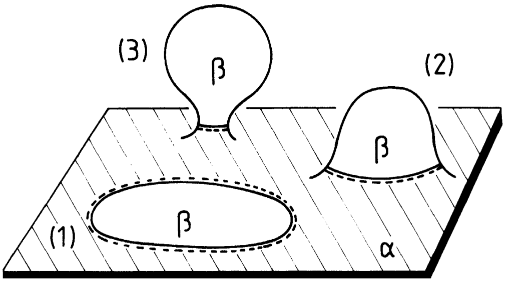

FIG. 5. Budding of the membrane domain β embedded in the membrane matrix α. The domain edge is indicated by the full-broken line. The length of this edge decreases during the budding process from (1) to (3).

5.2.3 Bud Size A vesicle that consists of two domains α and β could be obtained, for example, as a result of complete phase separation within the membrane. For such a vesicle, the shape of minimal bending energy exhibits a bud for a large range of parameters. This budded state attains a limit shape in which the bud β consists of a closed sphere that is connected to the mother vesicle α by an infinitesimal neck. The domain boundary with line tension σ is contained in this neck. The mean curvature M^β of this bud and the mean curvature M^α of the mother vesicle adjacent to this neck satisfy the general condition

$$\kappa^\alpha M^\alpha + \kappa^\beta M^\beta = \sigma/2 + \kappa^\alpha M^\alpha_{\text{sp}} + \kappa^\beta M^\beta_{\text{sp}}, \qquad (4)$$

where the bending rigidities κ^α and κ^β and the spontaneous curvatures M^α_{sp} and M^β_{sp} of the α and β domains may be different (Jülicher and Lipowsky, 1996).

The size of the bud is given by $1/M^\beta$. If the spontaneous curvatures are negligible, the bud size becomes $1/M^\beta \approx 2\kappa^\beta/\sigma$ in the limit of a relatively small β domain. In this limit, the size of the bud is determined by the elastic properties of the domain alone, and domain-induced budding then represents a *local* mechanism.

5.2.4 Domain-Induced Budding of Biomembranes In biological cells, budding of intramembrane domains represents the first step in the production of vesicles for the intracellular transport between different cell compartments. These domains can grow by diffusion-limited aggregation of molecules within the membrane or by the adsorption of molecules from the surrounding medium.

The budding of these domains could be governed by their spontaneous curvature. The aggregated molecules may have membrane-spanning anchors that induce such a curvature in the bilayer. Likewise, the adsorption of coat proteins onto one side of the bilayer leads to an asymmetric membrane. In any case, the reduction of the edge energy of these domains during the budding process will always act to facilitate this process.

5.3 Adhesion of Vesicles

A vesicle that is attracted toward a surface gains adhesion energy but increases its bending energy. For large vesicles, the adhesion energy must dominate since it is proportional to the contact area whereas the bending energy is scale-invariant and thus independent of the vesicle size.

5.3.1 Contact Potential A vesicle near a wall or substrate experiences various molecular forces; see Sec. 7. The typical range of these forces is usually small compared to the size of the vesicle. In order to study the overall shape of the bound vesicle, the microscopic interaction can be replaced by a contact potential. This potential is described by a single parameter W, which is equal to the adhesion energy per unit area. The value of the potential strength W is determined by the competition between direct molecular forces and fluctuation-induced forces; see Sec. 7.

5.3.2 Adhesion Threshold Since the bound vesicle is curved more strongly than the free vesicle, the contact potential $|W|$ has to exceed a certain threshold

$$|W_a| = w_a \kappa / \mathcal{A} \qquad (5)$$

before the vesicle starts to adhere to the wall. The dimensionless coefficient w_a depends on the reduced volume $\mathcal{V}/\mathcal{A}^{3/2}$ and is of order 1 (Seifert and Lipowsky, 1995).

For an ensemble of vesicles with different sizes, the relation (5) implies that large vesicles with surface area $\mathcal{A} > w_a\kappa/|W|$ stick to the wall, whereas vesicles with smaller surface area do not. This difference in the size distribution of bound and free vesicles is accessible to reflectivity measurements.

6. SHAPE FLUCTUATIONS OF MEMBRANES

The membranes of vesicles undergo thermally excited shape fluctuations, which can be directly observed in the light microscope. For adhering vesicles, one can use reflection interference microscopy as shown in Fig. 6 (Rädler *et al.*, 1995). In this way, one can probe fluctuations with wavelengths between about 0.4 μm and the vesicle size. There are, however, many more length scales involved in these fluctuations, which have wavelengths down to molecular dimensions.

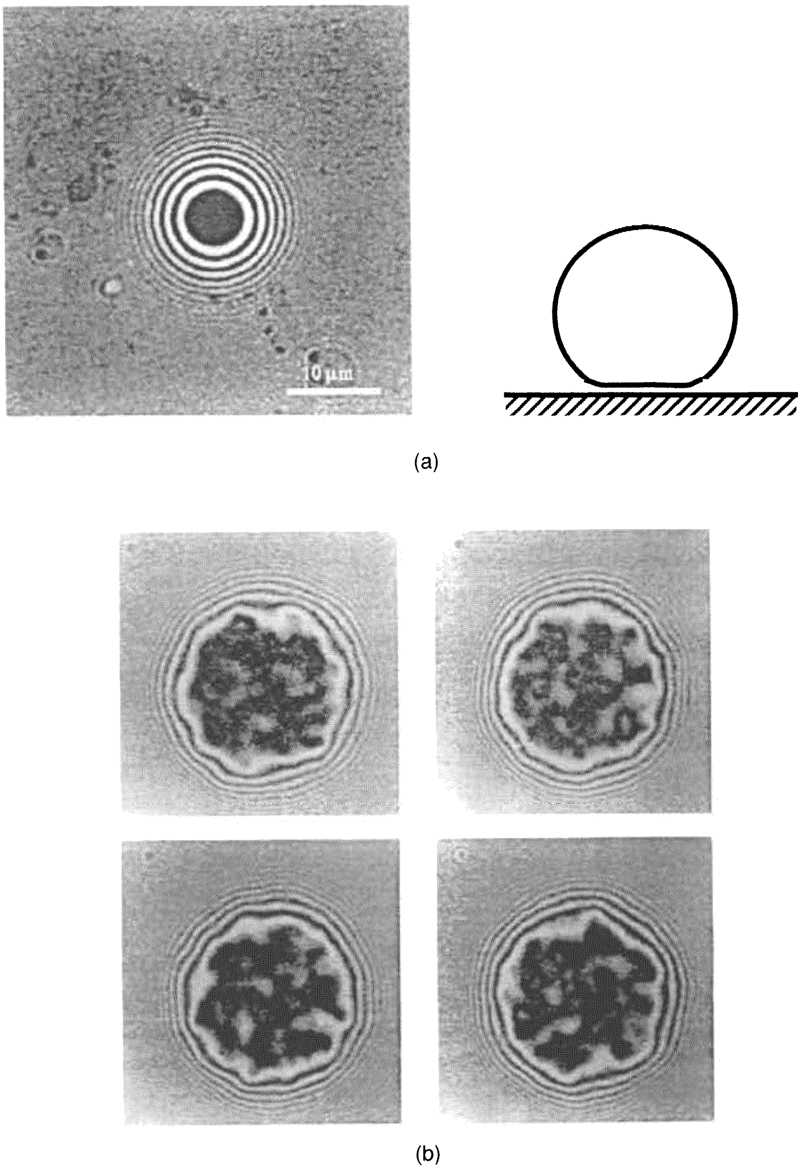

FIG. 6. Membrane of a bound vesicle as observed by reflection interference contrast microscopy. (a) Relatively large lateral tension, which suppresses all shape fluctuations, and (b) pronounced shape fluctuations for relatively small tension. (Courtesy of J. Rädler and E. Sackmann.)

6.1 Bending Modes or Undulations

On scales that are large compared to the membrane thickness, the typical shape fluctuations of fluid membranes are bending modes or undulations in which the surface area of the membrane remains unchanged.

6.1.1 Roughness Arising from Bending Modes An almost planar membrane segment that has bending rigidity κ and experiences the lateral tension Σ undergoes bending undulations that can be expanded in Fourier modes. The amplitude of these excitations depends on the temperature T. For a segment of linear size $L_{\|}$, these undulations lead to the membrane roughness $L_{\perp}$ with

$$\begin{aligned} L_{\perp}^2 &\approx (T/2\pi\Sigma)\ln(L_{\|}/\xi_{\Sigma}) \qquad \text{for } L_{\|} \gg \xi_{\Sigma}, \\ &\approx (T/16\kappa)L_{\|}^2 \qquad \text{for } L_{\|} \ll \xi_{\Sigma}, \end{aligned} \tag{6}$$

provided the crossover length $\xi_{\Sigma} \equiv (4\kappa/\Sigma)^{1/2}$ is large compared to the molecular scales (Brochard and Lennon, 1975). Thus, one has a rigidity-dominated regime for $L_{\|} \ll \xi_{\Sigma}$ and a tension-dominated regime for $L_{\|} \gg \xi_{\Sigma}$.

6.1.2 Measurement of the Bending Rigidity Since the roughness $L_{\perp} \sim (T/\kappa)^{1/2}L_{\|}$ within the rigidity-dominated regime, measurements of this roughness can be used to determine the bending rigidity κ. In practice, this is done by optical microscopy of nearly spherical vesicles for which the undulations are expanded in spherical harmonics. In this way, one finds the typical values $\kappa \simeq (10 - 20)T$ for phospholipid bilayers (for a recent list, see Seifert and Lipowsky, 1995). It is, however, difficult to obtain high-precision values for κ. This is understandable since the bending rigidity of lipid bilayers is affected by small changes in their molecular structure arising, e.g., from conelike defects or from the molecular roughness of the lipid–water interfaces.

6.2 Thermal Fluctuations on Molecular Scales

The concept of a bending mode is no longer meaningful as soon as its wavelength becomes of the order of the membrane thickness. On these small scales, the molecular structure of the lipid–water interface should be taken into account. This interface is roughened by thermal fluctuations, as has been observed in computer simulations (Pastor, 1994) and has been deduced from scattering experiments (Koñig *et al.*, 1992; Wiener and White, 1992; McIntosh and Simon, 1993). These thermal fluctuations correspond to relative displacements or deformations of the lipid head groups: Since they will, in general, change the surface area of the lipid–water interface, they are governed by an effective tension Σ_{eff}. The roughness arising from these fluctuations is set by the length scale $(T/2\pi\Sigma_{\text{eff}})^{1/2}$, which is expected to be of the order of a few ångströms.

6.3 Persistence Length on Large Scales

On sufficiently large scales, a fluid surface that does not experience any lateral tension starts to crumple, i.e., to behave as a random surface without any average orientation. This happens as soon as its size exceeds the so-called persistence length (de Gennes and Taupin, 1982; Gompper and Kroll, 1995). For phospholipid bilayers with bending rigidities $\kappa \gtrsim 10T$, the persistence length is, however, very large compared to the largest accessible sizes of these bilayers. Therefore, under normal circumstances, lipid bilayers (and biomembranes) do not behave as random surfaces with no average orientation.

7. GENERIC INTERACTIONS OF TWO MEMBRANES

The behavior of interacting membranes is governed by the interplay of direct interactions arising from the forces between the molecules and shape fluctuations such as bending undulations that act to renormalize these interactions.

7.1 Direct Interactions between Rigid Membranes

The direct interaction $V(l)$ between two rigid membranes at separation l can be measured by the surface force apparatus consisting of two mica surfaces onto which the membranes are immobilized (Marra and Israelachvili, 1985). The simplest example is provided by lipid bilayers that

1. are electrically neutral and
2. interact across a water layer that contains no macromolecules or colloids.

In this case, the interaction potential $V(l)$ is composed of a repulsive hydration and an attractive van der Waals interaction and has the schematic form shown in Fig. 7(a).

Lipid bilayers may become charged by adsorption of ions from the solution or by dissociation of their head groups. They then exhibit electric double layers, which usually lead to *repulsive* interactions between the surfaces as predicted by the classical Poisson–Boltzmann theory. The combination of van der Waals and electrostatic interactions often leads to a potential barrier as shown in Fig. 7(b).

7.2 Renormalization by Bending Undulations

Two interacting membranes that are in thermal equilibrium with the surrounding liquid undergo bending undulations. Since the configurational entropy of the undulations increases with the membrane separation, these fluctuations lead to a repulsive force that drives the membranes apart and, thus, reduces the attractive part of their interaction. It will be tacitly assumed here that the interaction potential $V(l)$ is effectively short-ranged and decays faster than $\sim 1/l^2$ for large l. Van der Waals forces eventually decay as $\sim 1/l^4$ and thus belong to this strong-fluctuation regime.

7.2.1 Repulsive Interactions If the interaction $V(l)$ between the membranes is effectively repulsive, the membranes can be kept together by an osmotic pressure P arising, e.g., from macromolecules in the surrounding solution that cannot permeate the membranes. In this case, the membranes are bound for finite $P > 0$ but unbind in the limit of zero P. The mean separation $\langle l \rangle$ behaves as $\langle l \rangle \sim 1/P^{\psi}$ for small P with the unbinding exponent $\psi = \frac{1}{3}$. As the membranes unbind, the probability $\mathcal{P}_2$ for pair contacts or two-membrane collisions decays to zero as $\mathcal{P}_2 \sim 1/\langle l \rangle^{\zeta_2}$ with the contact exponent $\zeta_2 = 3$. This is completely analogous to the behavior of one-dimensional strings, which are governed by a finite line tension.

The behavior of the mean separation $\langle l \rangle$ can be obtained from the entropic or fluctuation-induced interaction $V_{\text{fl}}(l) \sim 1/l^2$, as deduced by Helfrich (1978) using a heuristic scaling picture even though this picture implies the incorrect behavior $\mathcal{P}_2 \sim 1/\langle l \rangle^2$ for the contact probability.

7.2.2 Attractive Interactions and Continuous Unbinding For a direct interaction as shown in Fig. 7(a) that is dominated by attractive van der Waals forces, the membranes undergo a *continuous* unbinding transition from a bound state at low temperatures to an unbound state at high temperatures, as was first predicted theoretically (Lipowsky and Leibler, 1986). If the interaction potential $V(l)$ is parametrized by the effective potential depth $|U|$ and the effective potential range l_v, the unbinding temperature is given by

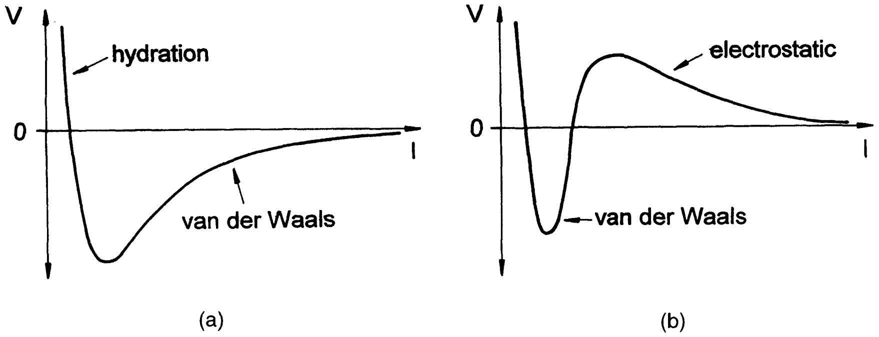

(a) (b)

FIG. 7. Direct interaction V between two planar membranes as a function of the membrane separation l: (a) hydration and van der Waals interaction between electrically neutral membranes; (b) hydration, van der Waals, and electrostatic interaction between electrically charged membranes leading to a potential barrier.

$$T_u \simeq \sqrt{\kappa |U|} l_v^2. \qquad (7)$$

As one approaches $T = T_u$ from below, the mean separation $\langle l \rangle$ of the membranes diverges as $\langle l \rangle \sim 1/|T - T_u|$, and the strength $|W|$ of the renormalized potential vanishes as $|W| \sim -|T - T_u|^2$. The same critical behavior is found for two interacting strings in two dimensions where the unbinding process represents a wetting transition.

The adhesion of large vesicles is also diminished by thermally excited fluctuations; compare Fig. 6. From the relation (5) and the behavior $|W| \sim -|T - T_u|^2$ of the adhesive strength, one obtains the unbinding temperature $T_u^{\text{ves}} \approx T_u - c\kappa l_v/\sqrt{\mathcal{A}}$ for large vesicles with surface area $\mathcal{A} \gg \kappa/|U|$.

7.2.3 Potential Barriers and Discontinuous Unbinding The direct interaction $V(l)$ may exhibit a potential barrier as shown in Fig. 7(b). In this case, the unbinding transition can be continuous or discontinuous depending on the relative strength of the potential barrier. If the potential barrier is relatively weak, the undulations "tunnel" through the barrier and the unbinding transition is continuous. This is again analogous to one-dimensional strings in two dimensions, which tunnel, however, through *any* potential barrier that decays faster than $1/l^2$ for large l. In contrast, two-dimensional membranes cannot tunnel through sufficiently strong barriers. In the latter situation, they are trapped by the barrier, and the unbinding transition is discontinuous.

7.3 Tension-Induced Adhesion

A lateral tension Σ acts to suppress the bending undulations and thus to decrease the fluctuation-induced repulsion. In fact, this interaction now becomes short-ranged and decays exponentially as $V_{\text{fl}}(l) \sim \exp(-l/l_\Sigma)$ for separations $l \gg l_\Sigma \equiv (T/2\pi\Sigma)^{1/2}$. Thus, for sufficiently large separations, this fluctuation-induced repulsion cannot compete with the attraction arising from van der Waals forces, and the membranes form a bound state in the presence of lateral tension. In the limit of vanishing tension, the membranes unbind provided the temperature exceeds the unbinding temperature T_u.

It has been argued by Helfrich (1989) that some lipid bilayers show a more complex behavior in the low-tension regime and that these layers develop a superstructure that acts as an additional reservoir for membrane area. This issue remains to be clarified.

7.4 Repulsive Interactions at Small Separations

It has been well established by many experiments that lipid bilayers in aqueous solution experience a strong repulsion at small separations of the order of 1 nm (Rand and Parsegian, 1989). The corresponding force per unit area or disjoining pressure, P, is observed to decay exponentially as $P \sim \exp(-l/l_{\text{eff}})$, where the effective decay length l_{eff} is of the order of a few ångströms; see Fig. 8. It was originally thought that this short-ranged repulsion represents a hydration effect and reflects the perturbed water structure in front of the polar head group (Marcelja and Radic, 1976). On the other hand, a similar repulsive force can also arise from the molecular roughness of the lipid–water interfaces (Israelachvili and Wennerström, 1990). In general, both mechanisms should contribute, and their relative importance will depend on the temperature and on the lipid–solvent composition (Lipowsky and Grotehans, 1994).

8. BUNCHES AND STACKS OF MEMBRANES

Lipid bilayers in solution often form bunches in which several membranes are, on average, parallel to each other. Two different geometries must be distinguished:

1. free bunches in solution, as shown in Fig. 9(a), and
2. bunches that adhere to a rigid surface or wall; see Fig. 9(b).

The latter geometry is obtained, e.g., by spreading a concentrated lipid solution on a glass slide.

The structure of such a bunch can be characterized by its density profile, which depends on the mean separations of the membranes within the bunch. If the shape fluctuations of these membranes are strong, they drive the membranes apart and lead to loosely bound or highly swollen states (Lipowsky, 1995a).

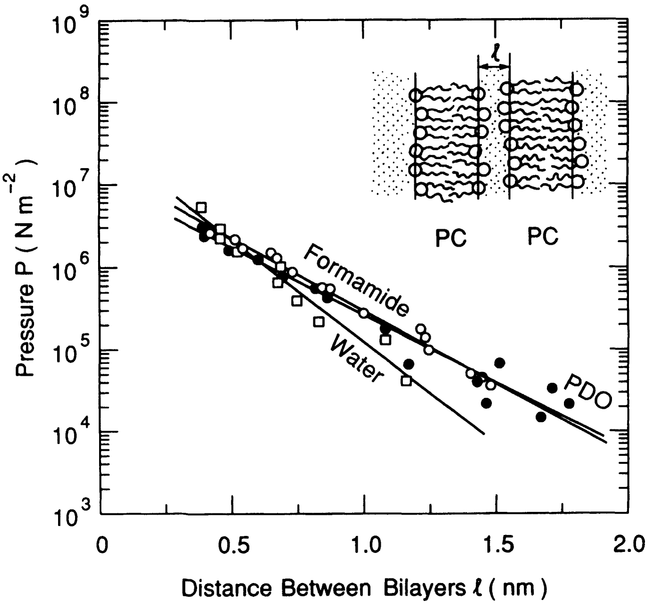

FIG. 8. Disjoining pressure P of phosphatidyl choline (PC) bilayers as a function of the mean separation $\langle l \rangle$ of lecithin or PC bilayers for the three solvents water, formamide, and 1,3-propanediol (PDO) (McIntosh *et al.*, 1989).

8.1 Free Membranes in Solution

For a free bunch of N identical membranes as shown in Fig. 9(a), all membranes have the same bending rigidity κ, and each adjacent pair of membranes interacts with the direct interaction $V(l)$. If the bunch is kept together by short-ranged attractive interactions arising, e.g., from van der Waals forces, it undergoes a *unique* unbinding transition as observed

1. in experiments for up to 20 sugarlipid membranes (Mutz and Helfrich, 1989) (see Fig. 10), and
2. in Monte Carlo simulations for up to four membranes (Netz and Lipowsky, 1993).

The unbinding temperature T_u^f of the free bunch is *independent* of N and thus identical to the unbinding temperature of two membranes.

Analytical work on bundles of strings (Hiergeist *et al.*, 1994) leads to the prediction that the mean separation $\langle l \rangle_n$ of the adjacent

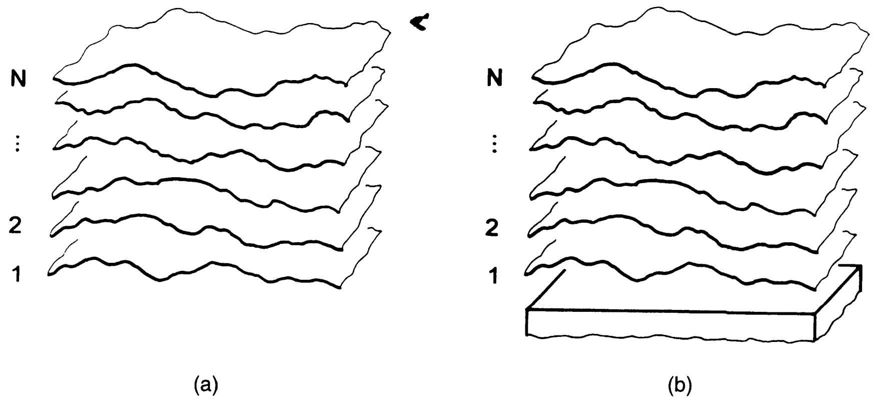

FIG. 9. Bunches of N fluctuating membranes: (a) Free bunch in solution; (b) bunch adhering to a rigid surface or wall.

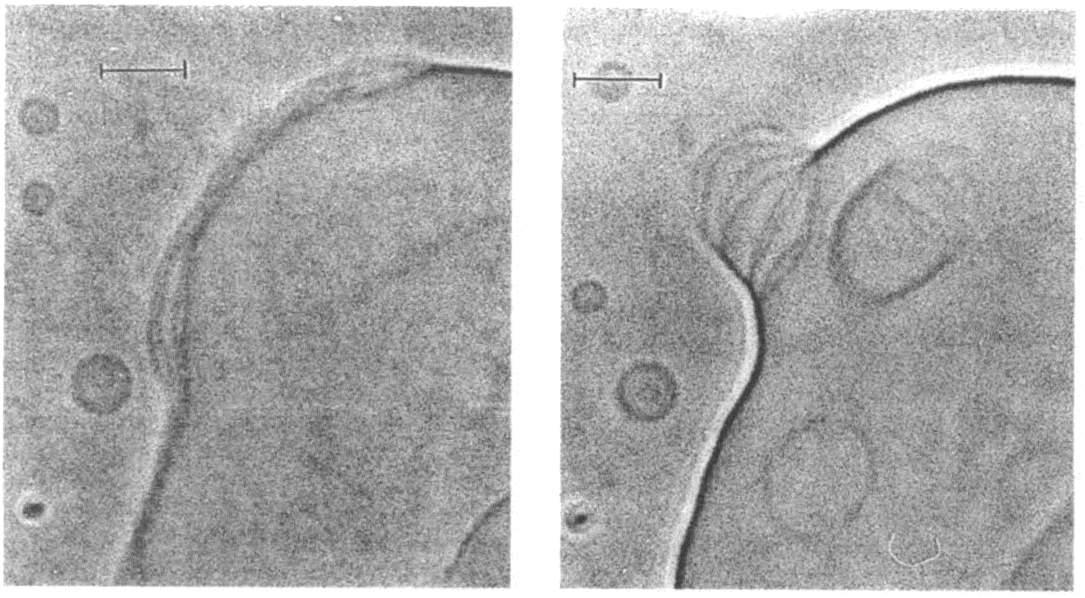

FIG. 10. Unbinding or adhesion transition as experimentally observed for a bunch of eight lipid bilayers in aqueous solution. (Left) For $T \simeq 22.4$ °C, the membranes undulate very strongly and then appear as thick fuzzy lines. (Right) For $T \simeq 22.1$ °C, the membranes form a bound state that corresponds to the sharp dark line. The water between the membranes has been squeezed into the large water pocket. The bars represent 10 μm. (Courtesy of W. Helfrich.)

membranes labeled by n and $n + 1$ is given by

$$\langle l\rangle_n \sim 1/n(N - n)|T - T_u^f|^{\psi} \qquad \text{with } \psi = 1. \tag{8}$$

Thus, all length scales should diverge with the universal unbinding exponent $\psi = 1$ but should exhibit an amplitude that depends strongly on N and on the position n within the bunch. This implies that the temperature interval in which the behavior is dominated by fluctuations shrinks as $\sim 1/N$ for large N. Therefore, for large N the continuous unbinding transition will appear to be discontinuous.

8.2 Membranes at a Rigid Surface

In the case of N membranes attracted toward a rigid surface or wall, as shown in Fig. 9(b), the unbinding behavior depends on the relative strength of the surface–membrane and the membrane–membrane attraction. There are essentially two possibilities:

1. If the attraction toward the substrate is sufficiently weak, one has a sequence of two transitions: the whole N-bunch unbinds from the substrate at the unbinding temperature $T = T_u^b < T_u^f$ and then undergoes the transition toward completely unbound membranes at $T = T_u^f$.
2. If the attraction toward the rigid surface is relatively strong, one has a sequence of N unbinding transitions and thus a sequence of unbinding temperatures T_u^n with $T_u^f < T_u^n \leq T_u^1$. At each of these transitions, a single membrane peels off from the bunch. Such peeling processes should always occur during the formation of vesicles from oriented samples. Thus, in this case, the presence of the rigid surface is still felt by the outermost membrane even in the presence of many intervening membranes.

8.3 Limit of Lyotropic Liquid Crystal

In the limit of large N, the membrane stack becomes a lyotropic liquid crystal, which can be studied by x-ray and neutron scattering.

8.3.1 Power-Law Peaks of the Scattering Intensity The renormalization of the direct interaction $V(l)$ by bending modes with small wavelengths leads to the effective potential $V_{\text{eff}}(l)$, which determines the mean separation $\langle l \rangle$ via $\partial V_{\text{eff}}(l)/\partial l = 0$. Expanding this effective potential up to second order in the displacements, one obtains a harmonic model, which leads to the prediction that the scattering intensity $S(q)$ exhibits the power-law behavior

$$S(q) \sim (q_z - q_m)^{-(2-X_m)} \quad \text{with } X_m \equiv Tq_m^2/8\pi\sqrt{\kappa v_{\text{eff}}} \qquad (9)$$

in reciprocal space along the q_z direction close to $q_z = q_m = 2\pi m/\langle l \rangle$ with $m = 1, 2, \ldots$ and $v_{\text{eff}} \equiv \partial^2 V_{\text{eff}}(l)/\partial l^2$ at $l = \langle l \rangle$. These power-law peaks are known as Landau–Peierls singularities and have been studied by x-ray scattering on lamellar phases of oil–water–surfactant mixtures (Meunier *et al.*, 1987).

8.3.2 Melting of Lamellar States Lamellar phases of oil–water–surfactant mixtures often melt into isotropic phases. If the melting is initiated by bending undulations, one may assume that the lamellar phase becomes unstable as soon as the roughness $\xi_\perp$ of the membrane separation satisfies $\xi_\perp > c_* \langle l \rangle$ (de Gennes and Taupin, 1982) (note that the roughness of a single "tracer" membrane within the liquid crystal increases logarithmically with the size of the membrane). This relation corresponds to the so-called Lindemann criterion for the melting of solids. Within the harmonic model, this implies that the lamellar phase becomes unstable as soon as the bending rigidity becomes too small and satisfies $\kappa < T^2/(2\pi c_*)^2 v_{\text{eff}} \langle l \rangle^4$.

9. POLYMER-DECORATED MEMBRANES

In addition to the lipid bilayer, biomembranes contain a complex mixture of macromolecules. These molecules, which are anchored in the membranes, are usually connected to a network of relatively stiff, rodlike filaments that belong to the so-called cytoskeleton within the cell. Likewise, the outer plasma membrane is often covered by a "coat" of polysaccharides, i.e., of branched polymers, which are also anchored within the membranes. These systems can be modeled, to a certain extent, by the decoration of lipid bilayers with polymers.

9.1 Anchored Polymers

Any polymer with one or several hydrophobic segments that fit into the hydrophobic interior of the lipid bilayer can be anchored to this bilayer. The simplest geometry is obtained for a polymer with a single anchor attached to one of its ends. Artificial systems of this kind are water-soluble polymers covalently bound to a single lipid molecule (Lasic, 1994) or block copolymers with a short hydrophobic block. Another possibility is to use polymers with hydrophobic side groups (Simon *et al.*, 1995). In this case, the polymers can be attached by several anchors. Such a system is shown in Fig. 11. In contrast to polymers grafted to solid surfaces, these anchored polymers can diffuse laterally along the fluid membrane.

If the nonanchored segments of the polymer are effectively repelled from or attracted to the membrane surface, the anchored polymer is in a desorbed or in an adsorbed state, respectively. In the desorbed state, a polymer that is attached to the membrane by a single anchor has the form of a "mushroom," the size of which is comparable to the dissolved polymer. Weak adsorption leads to squashed mushrooms; strong adsorption, to polymer "pancakes" that are tightly bound onto the membrane surface.

9.2 Dilute and Semidilute Regimes

First, consider effectively repulsive interactions between the anchored polymer and the membrane surface. If the concentration of anchors within the membrane is sufficiently small, the anchored polymers form well-separated mushrooms. This dilute regime applies up to the overlap concentration at which the membrane surface becomes completely covered by anchored polymers. If one increases the coverage beyond this overlap value, the polymers start to squeeze and to stretch each other and one enters the brush regime. This brush regime extends up to a maximal coverage at which the loss of entropy arising from the stretched state of the polymers exceeds their anchoring energy.

A similar distinction applies to anchored

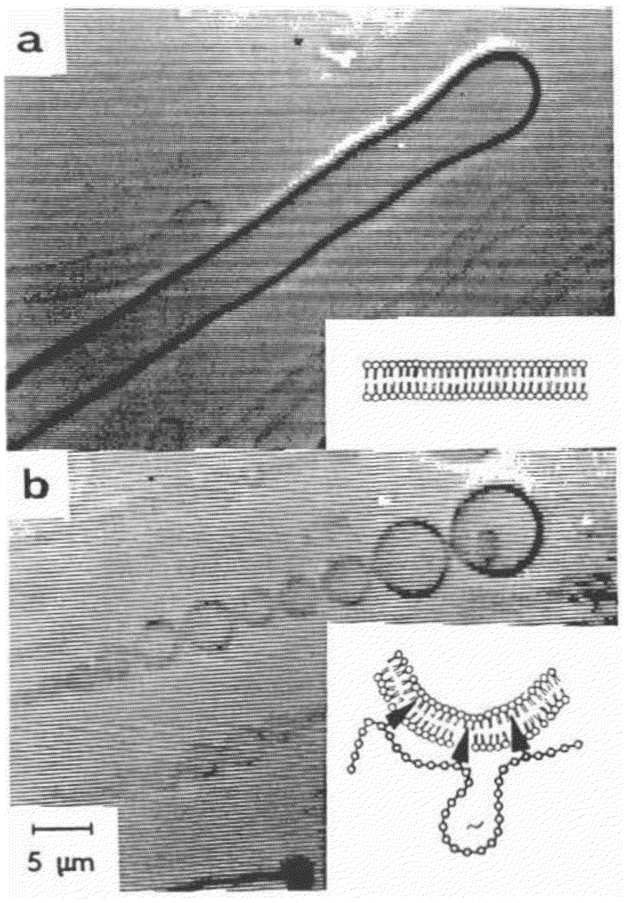

FIG. 11. Shape transformation of (a) a tubular vesicle into (b) a string of spheres as induced by the addition of polymers with hydrophobic side groups. This transformation is presumably governed by the interaction of these side groups with the lipid bilayer. (Courtesy of H. Ringsdorf.)

polymers that are adsorbed onto the membrane. In the dilute regime, one has well-separated pancakes. Beyond the overlap concentration, these pancakes start to interpenetrate each other and to form a semidilute adsorption layer.

9.3 Polymer-Induced Curvature

A bound polymer or mushroom suffers a loss of entropy compared to the free state. This entropy loss depends on the shape of the membrane; see Fig. 12. Therefore, the mushroom exerts entropic forces on the membrane that bend the membrane away from the polymer (Lipowsky, 1995b).

In the unbound state, a linear flexible polymer consisting of N monomers of linear size a forms a coil of linear dimension $R_{po} \simeq aN^{\nu}$ with $\nu \simeq 3/5$ for good solvents. In the bound mushroom state, the polymer induces the "spontaneous" mean curvature

$$M_{\mathrm{sp}} \sim T/\kappa R_{po} \sim T/\kappa a N^{\nu} \qquad (10)$$

of the decorated membrane. In a similar way, curvature is induced by a polymer brush that is attached to the membrane in a nonsymmetric way. If the polymer is adsorbed onto the membrane, on the other hand, it induces the opposite sign of the curvature: The membrane now bends toward the membrane in order to maximize the number of contacts.

9.4 Polymer-Induced Adhesion

The adhesion of cell membranes is mediated by rodlike polymers that have a linear extension of 10–30 nm. In the absence of these sticky rods, the cell membranes experience an effectively repulsive interaction arising from the electric charges on their surfaces. The adhesive strength mediated by

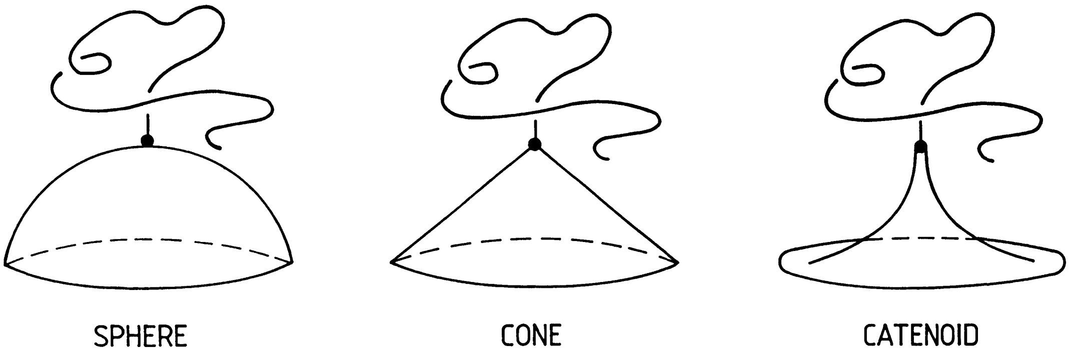

FIG. 12. A polymer attached to the membrane by a single anchor forms a mushroom that bends the membrane by entropic forces. For the three shapes of the membrane segment shown here, the cone leads to the highest entropy for the anchored polymer.

the sticky rods depends on the concentration X of their anchors within the membrane.

Since this strength must exceed a certain threshold value in order to overcome the repulsion arising from the membrane undulations, the anchor concentration must also exceed a critical value X_c. If the rod anchors attract one another, the bilayers may undergo phase separation into polymer-rich domains with $X > X_c$ and polymer-depleted domains with $X < X_c$. In the latter case, the membranes will only adhere via the polymer-rich domains.

10. APPLIED ELECTROMAGNETIC FIELDS

For the phenomena discussed in the previous sections, the molecular structure of the lipid bilayer is not perturbed in an essential way. In contrast, the application of pulsed electric fields or of optical tweezers built up from a laser beam strongly perturbs the bilayer structure, which leads to new cooperative effects.

10.1 Alternating Electric Fields

The lipid membrane has a dielectric constant (or permittivity) $\epsilon_M \simeq 2$, which is small compared to the dielectric constant $\epsilon_W \simeq 80$ of water. Likewise, the electrical conductivity γ_W of water is large compared to the conductivity γ_M of the lipids. This implies that the electric field is *amplified* within the membrane. This is the basic mechanism that is responsible for the electroporation and electrofusion of lipid bilayers and biomembranes (Neumann *et al.*, 1989; Chassy *et al.*, 1992).

If the alternating electric field $\mathbf{E}_W \exp[i\omega t]$ is applied in the water and the two lipid–water interfaces are treated as dividing surfaces with no intrinsic width, the transmembrane potential $\Delta\Phi_M$ across a planar membrane is given by

$$\begin{aligned} \Delta\Phi_M &= E_M a_\perp \\ \text{with } E_M &= \left[\frac{(\omega\epsilon_W)^2 + \gamma_W^2}{(\omega\epsilon_M)^2 + \gamma_M^2}\right]^{1/2} E_W, \end{aligned} \tag{11}$$

where $a_\perp$ represents the thickness of the membrane (this relation follows from Maxwell's equations and the continuity equation for the charge carriers). For water without salt, the square root prefactor is about 5×10^3.

10.2 Electroporation

If the lipid membrane is exposed to a transmembrane potential $\Delta\Phi_M$ that exceeds $\simeq 0.2$ V for an extended period of time, the membrane ruptures in an irreversible way. On the other hand, if one applies a potential $\Delta\Phi_M$ of the order of 1 V for a relatively short time period Δt within the range 10^{-7} s $\leq \Delta t \leq 10^{-4}$ s, one obtains a reversible electrical breakdown of the lipid bilayer: The permeability of the bilayer is strongly increased right after the pulse has been applied but then relaxes back to its original value. This behavior is observed both for artificial lipid bilayers and for biomembranes. In both cases, the high-permeability state arises from the formation of pores within the lipid membrane. These pores have been directly observed by freeze-fracture electron microscopy.

10.3 Electrofusion

Two membranes that are in close contact can be fused by applying a short electrical pulse. It is widely believed that this process of electrofusion is induced by the simultaneous electroporation of both membranes within the contact area. It has also been found, however, that fusion can even occur when the membranes are treated with electrical pulses before being brought into close contact. Electrofusion is a universal fusion mechanism that has been applied to many different types of biomembranes. An example in shown in Fig. 13.

The initial fusion step presumably consists in the juxtaposition of two pores in the two adjacent membranes. The next step should be the formation of a neck or "wormhole" that bridges the gap between the two membranes. Several intermediate structures that could be involved in this topological transformation have been proposed.

10.4 Optical Tweezers

Optical tweezers, arising from a laser beam that is focused through a microscope objective, provide localized light traps and can be used to manipulate various objects in the micron range. Recently, such tweezers have been directly applied to lipid bilayers and were found to induce novel shape transformations. When applied to cylindrical vesicles consisting of a single bilayer, the tweezers induce a pearling instability, which spreads from the trap and eventually produces a chain of pearls that again resembles the Plateau–Rayleigh instability of capillary tubes. When applied to bunches of several bilayers, the light trap acts to unbind the membranes locally and to induce a characteristic shape of the loosely bound structure (Bar-Ziv *et al.*, 1995).

It is presently believed that the focused laser beam destroys the bilayer structure locally and acts to pull lipid molecules into the light trap. This process should generate a lateral tension for the rest of the membrane, which is not exposed to the beam. A theoretical model has been developed that is based on this interpretation and has been used to calculate the selected wavelength of the pearling instability (Goldstein et al., 1996).

11. OUTLOOK ON APPLICATIONS

The tendency of lipids to form closed compartments such as vesicles is used in pharmacology, food industry, and cosmetics in order to enclose and transport drugs, enzymes, vitamins, moisturizers, and other in-

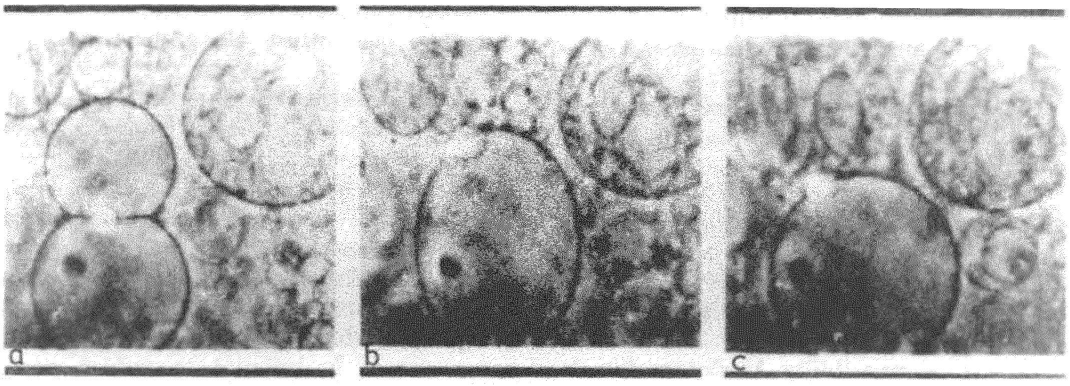

FIG. 13. Electrofusion of two lipid vesicles as observed by phase contrast microscopy: (a) The two vesicles, with diameters of about 45 and 37 μm, are brought into close contact by an alternating electric field; (b) fused vesicle about 1 s after the application of a short electrical pulse; (c) the fused vesicle, which has a diameter of about 51 μm, becomes roughly spherical when the alternating field is switched off. (Courtesy of U. Zimmermann.)

gredients. The delivery of drugs via vesicles can be improved by the decoration with polymers that are attached to the lipid head groups and that prevent the uptake of these vesicles by lymphocytes in the blood. Furthermore, vesicles composed of lipid mixtures might be used in order to transport drugs directly through the skin and, thus, to replace the use of hypodermic needles.

Other interesting areas are attempts to attach lipid layers to solid substrates in order to build up new types of biosensors, or to cover artificial bones by such layers in order to improve their biocompatibility.

The mechanisms of electroporation and electrofusion have been applied to many areas of biotechnology. For example, electroporation is now widely used in order to transfer genes into all kinds of cells.

In summary, flexible membranes already play an important role in various soft-matter technologies. It seems rather likely that an improved understanding of the physics of these systems will lead to new technological applications: The future is in the flexible!

GLOSSARY

Aggregation Colloids: Supramolecular structures that arise from the self-assembly of single molecules or monomers in solution.

Amphiphilic Molecules: Molecules that have a lyophilic part, which dissolves in the solvent, and a lyophobic part, which is insoluble in the solvent. One important example is molecules dissolved in water that have a polar hydrophilic part and a nonpolar hydrophobic part.

Anchored Polymer: Polymer chain attached to the bilayer membrane by one or several hydrophobic segments that fit into the hydrophobic interior of the bilayer. If the nonanchored segments of the polymer are effectively repelled or attracted to the membrane surface, the anchored polymer attains the form of a "mushroom" or "pancake," respectively.

Area Compressibility Modulus: Elastic modulus of membranes that governs their resistance against stretching.

Bending Rigidity: Elastic modulus of membranes that governs their resistance against bending.

Bilayer: Membrane consisting of two monolayers that are arranged in such a way that the polar head groups of the lipid molecules form the two lipid–water interfaces, whereas the hydrophobic chains are buried within the bilayer and have essentially no contact with the aqueous solution. The thickness of these bilayers is 4–5 nm.

Budding: Shape transformation in which the membrane produces a small bud that is essentially spherical and connected to the original membrane via a small neck.

Conformal Diffusion: For vesicles with two or more handles, the shape of the vesicle can undergo a diffusion process along degenerate states that are related by conformal transformations.

Critical Monomer Concentration: This concentration represents

1. the upper limit for the concentration of amphiphilic monomers in solution and
2. the critical concentration at which the amphiphilic molecules start to aggregate into supramolecular structures.

It is also called *critical micelle concentration* since many surfactant molecules assemble into micelles.

Curvature: Any point on a two-dimensional surface that is embedded in three-dimensional space can be characterized by two principal curvatures, C_1 and C_2, which are equal to the inverse curvature radii. The mean curvature is given by $M \equiv (C_1 + C_2)/2$.

Decorated Membrane: Lipid bilayer with anchored polymers. For fluid bilayers, these polymers can diffuse along the membrane surface.

Direct Interaction: Interaction between two rigid membranes arising from the forces between the molecules. This interaction can be measured by the surface-force apparatus in which the membranes are immobilized onto two mica surfaces.

Electrofusion: Two adjacent membranes that are separated by a small water layer can be fused by the application of an electrical pulse, which is believed to induce pores through both membranes.

Electroporation: Pores that go through a bilayer can be induced by the application of a short electrical pulse with a voltage of the order of 1 V.

Flip-flops: The slow exchange of molecules between the two monolayers of a bilayer membrane.

Fluid Membranes: Membranes for which shear deformations relax by flow within the membrane. The molecules within such membranes undergo rapid lateral diffusion.

Hydrophobic Effect: The tendency of amphiphilic molecules in water to form supramolecular structures in which the hydrophilic groups of these molecules form the outside interface toward the aqueous solution, whereas the hydrophobic groups are buried inside and have essentially no contact with the water; the corresponding free-energy change has a large entropic contribution arising from the configurational entropy of the hydrogen bond networks in the water.

Internal Membranes: Biomembranes within the cell that cover many intracellular organelles such as the cell nucleus, the endoplasmic reticulum, mitochondria, etc.

Lipids: Amphiphilic molecules for which the hydrophobic part consists of two hydrocarbon chains and the hydrophilic part consists of a polar head group.

Lyotropic Liquid Crystal: Liquid crystals that can be swollen by the addition of solvent. In the present context, these liquid crystals consist of stacks of bilayer membranes.

Monolayer: Monomolecular layer of amphiphilic molecules that can be found, e.g., at the interface between two immiscible liquids such as water and oil.

Optical Tweezers: Light trap with a linear size of about 0.5 μm, which is generated by a focused laser beam and which can be used to grasp and to move a single lipid membrane or a whole bunch of membranes.

Plasma Membrane: Biomembrane that represents the boundary between a biological cell and its environment; it consists of a lipid bilayer with many anchored proteins.

Polymer-Induced Curvature: Polymers that are anchored onto the membrane exert entropic forces that induce a "spontaneous" curvature of the decorated membrane.

Shape Fluctuations: Membranes flicker and thus undergo thermally excited undulations that are directly visible in the optical microscope.

Shape Transformations: The shape of vesicles can be transformed in a controlled way by changing the temperature or the osmotic conditions. Such shape transformations can occur in a continuous or discontinuous fashion.

Surfactants: Amphiphilic molecules for which the hydrophobic part usually consists of one hydrocarbon chain.

Tension of Rupture: Lateral tension, of the order of a few mJ/m^2, at which the lipid membrane ruptures.

Unbinding Transition: Phase transition of interacting membranes between a bound adhesive state at low temperatures and an unbound nonadhesive state at high temperatures.

Vesicle: Closed bilayer without edges, which can attain a large variety of different shapes. An onionlike structure consisting of several closed bilayers is usually called a liposome.

Works Cited

Alberts, B., Bray, D., Lewis, J., Raff, J. M., Roberts, K., Watson, J. D. (1994), *Molecular Biology of the Cell,* 3rd ed., New York: Garland.

Bar-Ziv, R., Menes, R., Moses, E., Safran, S.-A. (1995), *Phys. Rev. Lett.* **75,** 3356–3359.

Brochard, F., Lennon, J. F. (1975), *J. Phys. (Paris)* **36,** 1035–1047.

Canham, P. B. (1970), *J. Theoret. Biol.* **26,** 61–81.

Cevc, C., Marsh, D. (1987), *Phospholipid Bilayers: Physical Principles and Models,* New York: Wiley.

Chassy, B. M., Saunders, J. A., Sowers, A. E. (Eds.) (1992), *Guide to Electroporation and Electrofusion,* San Diego: Academic.

Darnell, J., Lodish, H., Baltimore, D. (1990), *Molecular Cell Biology. Scientific,* New York: W. H. Freeman.

de Gennes, P.-G., Taupin, C. (1982), *J. Phys. Chem.* **86,** 2294–2304.

Evans, E. A. (1974), *Biophys. J.* **14,** 923–931.

Evans, E., Needham, D. (1987), *J. Phys. Chem.* **91,** 4219–4228.

Farge, E., Devaux, P. F. (1992), *Biophys. J.* **61,** 347–357.

Goldstein, R. E., Nelson, P., Powers, T., Seifert, U. (1996), *J. Phys. II (France)* **6,** 767–796.

Gompper, G., Kroll, D. M. (1995), *Phys. Rev. E* **51,** 514–525.

Helfrich, W. (1973), *Z. Naturforsch.* **28c,** 693–703.

Helfrich, W. (1978), *Z. Naturforsch.* **33a,** 305–315.

Helfrich, W. (1989), *Liquid Cryst.* **5,** 1647–1658.

Hiergeist, C., Lässig, M., Lipowsky, R. (1994), *Europhys. Lett.* **28,** 103–108.

Israelachvili, J. N., Wennerström, H. (1990), *Langmuir* **6,** 873–876.

Jülicher, F., Lipowsky, R. (1996), *Phys. Rev. E* **53,** 2670–2683.

Jülicher, F., Seifert, U., Lipowsky, R. (1993), *Phys. Rev. Lett.* **71,** 452–455.

Kahlweit, M., Lipowsky, R. (Eds.) (1996), *Microemulsions—Experiment and Theory,* Berichte der Bunsengesellschaft für Physikalische Chemie No. 100, 181–393.

König, S., Pfeiffer, W., Bayerl, T., Richter, D., Sackmann, E. (1992), *J. Phys. II (France)* **2,** 1589–1615.

Krstic, R. V. (1976), *Ultrastruktur der Säugetierzelle,* Berlin: Springer.

Lasic, D. D. (1993), *Liposomes: From Physics to Applications,* Amsterdam: Elsevier.

Lasic, D. D. (1994), *Angew. Chem. Int. Ed. Engl.* **33,** 1685–1698.

Lipowsky, R. (1991), *Nature* **349,** 475–481.

Lipowsky, R. (1995a), *Z. Phys. B* **97,** 193–203.

Lipowsky, R. (1995b), *Curr. Opin. Struct. Biol.* **5,** 531–540.

Lipowsky, R., Grotehans, S. (1994), *Biophys. Chem.* **49,** 27–37.

Lipowsky, R., Leibler, S. (1986), *Phys. Rev. Lett.* **56,** 2541–2544.

Marcelja, S., Radic, N. (1976), *Chem. Phys. Lett.* **42,** 129–130.

Marra, J., Israelachvili, J. N. (1985), *Biochemistry* **24,** 4608–4618.

McIntosh, T. J., Simon, S. A. (1993), *Biochemistry* **32,** 8374–8384.

McIntosh, T. J., Magid, A. D., Simon, S. A. (1989), *Biochemistry* **28,** 7904–7912.

Meunier, J., Langevin, D., Boccara, N. (Eds.) (1987), *Physics of Amphiphilic Layers,* Springer Proceedings in Physics vol. 21, Berlin: Springer.

Miao, L., Seifert, U., Wortis, M., Döbereiner, H.-G. (1994), *Phys. Rev. E* **49,** 5389–5407.

Michalet, X., Bensimon, D. (1995), *Science* **269,** 666–668.

Mutz, M., Helfrich, W. (1989), *Phys. Rev. Lett.* **62,** 2881–2884.

Nelson, D., Piran, T., Weinberg, S. (Eds.) (1989), *Statistical Mechanics of Membranes and Surfaces,* Singapore: World Scientific.

Netz, R. R., Lipowsky, R. (1993), *Phys. Rev. Lett.* **71,** 3596–3599.

Neumann, E., Sowers, A. E., Jordan, C. A. (Eds.) (1989), *Electroporation and Electrofusion in Cell Biology,* New York: Plenum Press.

Pastor, R. W. (1994), *Curr. Opin. Struct. Biol.* **4,** 486–492.

Rädler, J., Feder, T. J., Strey, H. H., Sackmann, E. (1995), *Phys. Rev. E* **51,** 4526–4536.

Rand, R. P., Parsegian, V. A. (1989), *Biochim. Biophys. Acta* **988,** 351–376.

Sackmann, E. (1990), *Can. J. Phys.* **68,** 999–1012.

Seifert, U., Lipowsky, R. (1995), in: R. Lipowsky, E. Sackmann (Eds.), *Handbook of Biological Physics,* Vol. 1, *The Structure and Dynamics of Membranes,* Amsterdam: Elsevier.

Simon, J., Kühner, M., Ringsdorf, H., Sackmann, E. (1995), *Chem. Phys. Lipids* **76,** 241–258.

Tanford, C. (1991), *The Hydrophobic Effect: Formation of Micelles and Biological Membranes,* Malabar, FL: Krieger.

Wiener, C. M., White, S. H. (1992), *Biophys. J.* **61,** 434–447.

Willmore, T. J. (1982), *Total Curvature in Riemannian Geometry,* Chichester: Ellis Horwood.

Further Reading

Bloom, M., Evans, E., Mouritsen, O. G. (1991), *Quart. Rev. Biophys.* **24,** 293–397.

Cevc, C., Marsh, D. (1987), *Phospholipid Bilayers: Physical Principles and Models,* New York: Wiley.

Chassy, B. M., Saunders, J. A., Sowers, A. E. (Eds.) (1992), *Guide to Electroporation and Electrofusion,* San Diego: Academic.

Gennis, R. (1989), *Biomembranes: Molecular Structure and Function,* New York: Springer-Verlag.

Lasic, D. D. (1993), *Liposomes: From Physics to Applications,* Amsterdam: Elsevier.

Lipowsky, R., Sackmann, E. (Eds.) (1995), *Handbook of Biological Physics,* Vol. 1, *Structure and Dynamics of Membranes,* Amsterdam: North-Holland.

Seifert, U. (1997), "Configurations of Fluid Vesicles and Membranes," *Adv. Phys.* **46,** 13–137.

Shinitzky, M. (Ed.) (1993), *Biomembranes: Physical Aspects,* New York: VCH Publishers.

VIBRATIONAL STATES IN DISORDERED SOLIDS

R. O. POHL, *Department of Physics, Cornell University, Ithaca, New York, U.S.A.*

	Introduction	223
1.	**Experimental Methods**	223
2.	**Ordered Crystals**	225
3.	**Amorphous Solids**	227
3.1	Experiments	227
3.2	Theoretical Models	229
3.2.1	Low-Temperature Properties; Tunneling Model	229
3.2.2	High-Temperature Properties; Einstein's Model Revisited	232
4.	**Highly Disordered Crystals**	233
5.	**Closing Remarks**	236
	Glossary	237
	Works Cited	238
	Further Reading	238

INTRODUCTION

The atomic vibrations of solids play a fundamental role in many of their properties. These vibrations determine thermal energy contents and heat transport, and they also influence many other properties of solids, like optical absorption or electrical conduction. In their pioneering work Debye (1912) and Born and von Kármán (1912) have shown that in crystals the vibrational states can be well described as collective excitations, in which atoms oscillate with certain phase relations relative to their neighboring atoms. The motions are described as waves with wavelength λ and frequency of vibration ν. These waves travel through the crystal with velocity v and can pick up or give up energy like a harmonic oscillator. In the quantum picture, the quanta of these vibrations are called phonons.

The vibrational states of highly perfect crystals have been studied by now in detail, and the concept of elastic waves has been well confirmed. However, as more highly disordered solids, like amorphous solids and mixed crystals, have been investigated more thoroughly, it has been recognized that their vibrational states are more complex. In this article, some of the experimental observations of these vibrational states are reviewed, and the models proposed for their description are described.

1. EXPERIMENTAL METHODS

The measurements to be discussed below are thermal conductivity, specific heat, and propagation of ultrasound. These techniques are shown schematically in Fig. 1. Thermal conductivity Λ is defined as the proportionality factor between heat current density and temperature gradient; for the simple geometry shown in Fig. 1(a),

$$\Lambda = (d/A)P/\Delta T, \tag{1}$$

where P, the constant thermal power generated in an electric heater, enters the sample of cross section A at the top and leaves it at the bottom. The temperature difference ΔT is measured with two thermometers that are spaced along the sample by the distance d. To ensure that P only flows through the sample, it is mounted in an evacuated space. Since at elevated temperatures, typically above ~100 K, a non-negligible fraction of the heat also travels through vacuum as infrared radiation, a different technique is preferable at these temperatures (Cahill, 1990), as sketched in Fig. 1(b): A thin metal strip evaporated onto the sample is heated by an alternating current. It creates a temperature wave that diffuses into the sample. The amplitude of this temperature wave varies inversely as the thermal conductivity of the sample. Since the electrical resistance of the metal strip varies with its temperature, it

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

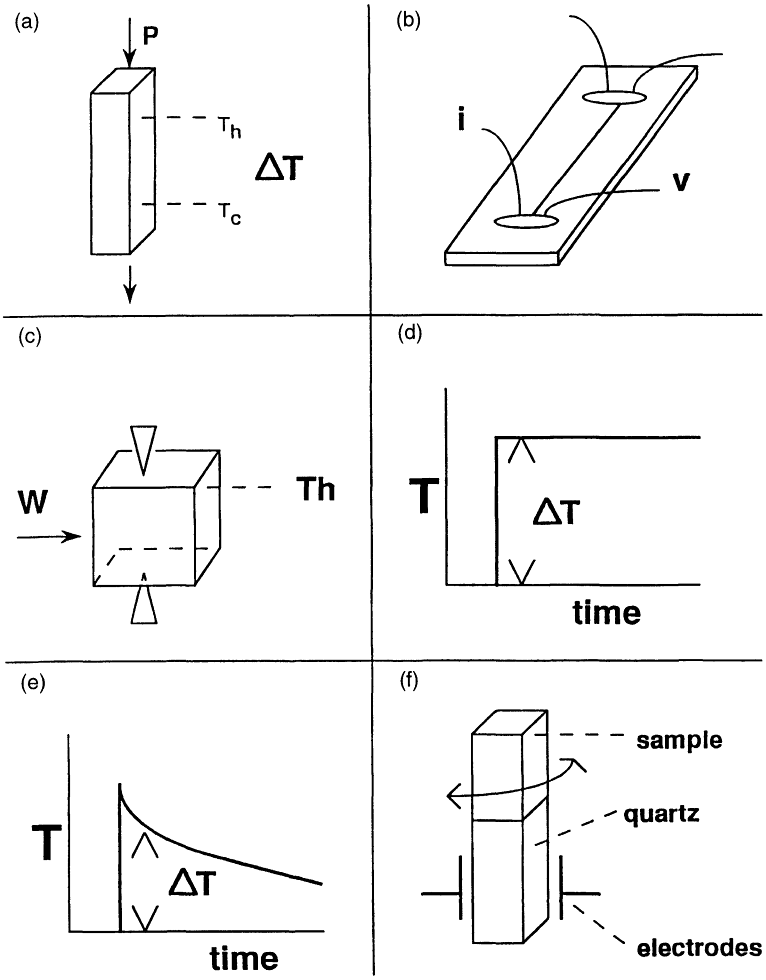

FIG. 1. Sketches of the experimental methods used. (a) Thermal conductivity. A steady heat flow sets up a temperature gradient along the sample. (b) Thermal conductivity measured with an alternating electric current flowing through a thin metal film deposited on a flat sample. The film is also used as thermometer. (Cahill, 1990.) (c) Specific heat. The sample (drawn as a cube) is supported in an evacuated chamber between pins, to ensure thermal isolation. When heat W is added (Joule heating of a heater attached to the sample), the temperature rise is monitored with a thermometer (Th). (d): After the heat W has been added in (c), the temperature T rises rapidly (ΔT) in a crystalline solid. In an amorphous solid, (e) thermal equilibrium is reached only gradually: "time-dependent specific heat," see text. (f): Torsional oscillator driven by an electric quadrupole field applied perpendicularly to the long axis of the quartz rod that acts as transducer.

can also be used to measure the latter. From this temperature amplitude, the thermal conductivity of the sample can be obtained. It can be shown that because of the geometry, the radiative heat losses are negligible.

For specific-heat measurements, the sample (mass m) is mounted such that it is essentially thermally isolated from its surroundings, Fig. 1(c). A very small electric heater is used to add the energy W to the sample, raising its temperature by ΔT; see Fig. 1(d). The specific heat C_p is then calculated through

$$C_p = (1/m)W/\Delta T. \tag{2}$$

(The subscript "p" indicates that the measurement is carried out at constant pressure, and not at constant volume; below 300 K, the difference is usually negligible.) If thermal equilibration is reached slowly, the temperature measured by the thermometer will at first rise steeply, and then will decrease with time, as the heat is slowly absorbed by the weakly coupled excitations; see Fig. 1(e). From such a measurement, the thermal relaxation of these excitations can be determined. If this time-dependent $\Delta T(t)$ is inserted into Eq. (2), one obtains a quantity $C_p(t)$, which is often called a time-dependent specific heat. This terminology is somewhat misleading, because the thermodynamic definition of C_p requires ΔT to be constant, i.e., the sample to be in thermal equilibrium. Nonetheless, in the following we will use this generally accepted term.

For the ultrasonic measurement, a variety of techniques can be used to measure either longitudinal or transverse sound attenuation. In Fig. 1(f), a setup is sketched in which the sample is attached to a quartz transducer that is excited by a time-varying electric field applied perpendicular to its axis to perform a torsional (twisting) motion around its axis (Cahill and VanCleve, 1989). Transducer and sample have been cut to the proper length so that a standing torsional wave is excited, with half a wavelength in the transducer and half a wavelength in the sample. For a typical driving frequency of 50 kHz, the quartz transducer must be ~3 cm long; the length of the sample depends on its transverse speed of sound. With this arrangement ultrasonic losses in the epoxy that is used to bond the sample to the transducer are minimized, since the glue line is in a plane of maximum torsional amplitude, and hence of minimum strain.

The quantity measured is the attenuation of the amplitude A of this compound oscillator:

$$A = A_0 \exp(-Q_{\mathrm{co}}^{-1}\pi\nu t), \qquad (3)$$

where A_0 is the amplitude of the wave at the time $t = 0$, ν is its frequency, and Q_{co}^{-1} is called the internal friction of the compound oscillator. The internal friction Q^{-1} of the sample itself is then calculated from

$$Q^{-1} = \frac{I_s + 1.06 I_t}{I_s} Q_{\mathrm{co}}^{-1}, \qquad (4)$$

where I_s and I_t are the rotational inertias of sample and transducer, respectively.

2. ORDERED CRYSTALS

In order to illustrate the properties of a solid in which the elastic excitations are waves, Figs. 2 and 3 show specific heat and thermal conductivity of several highly perfect crystals containing a minimum amount of chemical and physical imperfections. Over most of the temperature range, the specific heat C_p varies as the third power of the temperature T. The dashed lines are the predictions of the Debye theory based on the assumption that the entire thermal energy is contained in plane waves, all moving with

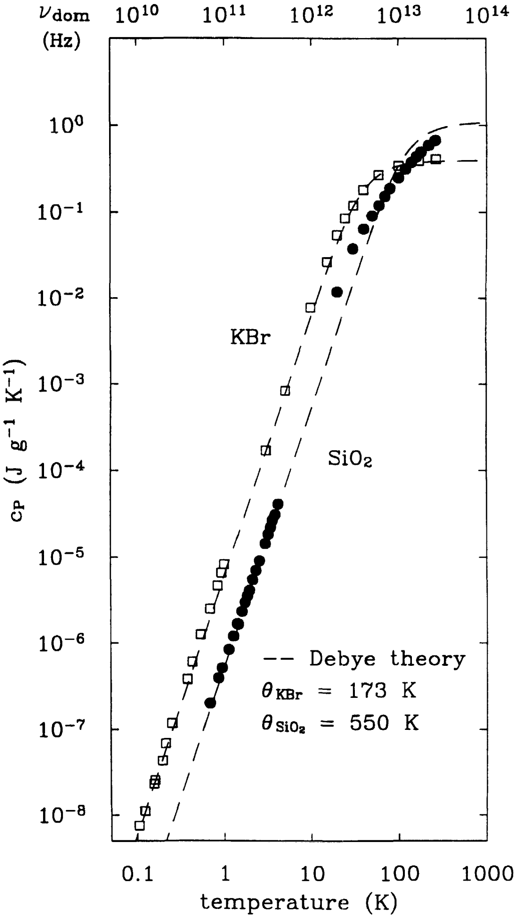

FIG. 2. Specific heat of two perfect single crystals. Measurements compared with the Debye theory (dashed lines). As the temperature increases, increasingly larger frequencies of vibration are thermally excited, shown as ν_{dom} on the upper horizontal scale (see text). Thus, by varying the temperature, one can vary the frequency of the waves studied.

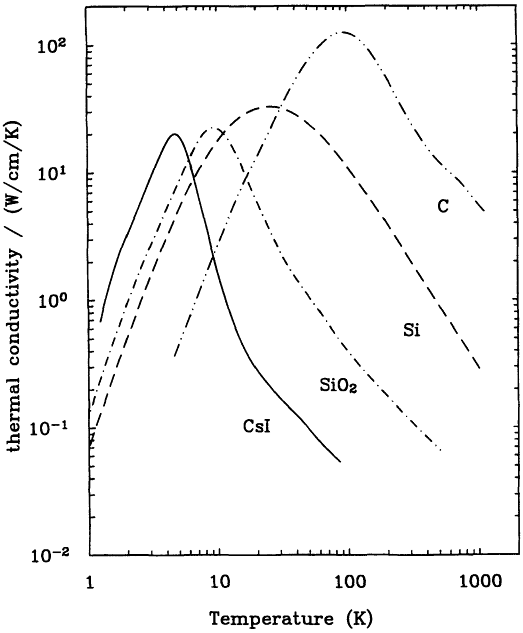

FIG. 3. Thermal conductivity of four perfect single crystals. C is diamond. By changing the measuring temperature, one can learn about the scattering of different phonon frequencies (see caption of Fig. 2).

the speed of sound, which is an average of the transverse and the longitudinal speeds of sound.

The Debye theory for elastic waves parallels that of Planck's theory for electromagnetic waves, in particular at low temperatures, where the energy contents of both vary as T^4, or the specific heat varies as T^3 (the only qualitative difference is that elastic waves can also be longitudinal; a quantitative difference is that the elastic waves have velocities about 10^5 times smaller than the electromagnetic ones). As the temperature increases, waves of increasingly larger frequency are excited. In Planck's theory this shift is called Wien's displacement law. It connects the frequency ν_{Wien} that, in a given frequency interval, carries most of the energy, with the temperature T. ν_{Wien} varies linearly with T:

$$\nu_{\mathrm{Wien}} = 5(k_B/h)T \approx (10^{11}\ \mathrm{Hz/K})T = \nu_{\mathrm{dom}}, \tag{5}$$

where $k_B = 1.38 \times 10^{-23}$ W s K^{-1} is Boltzmann's constant, and $h = 6.62 \times 10^{-34}$ W s^2 is Planck's constant. For elastic waves, this frequency is called the dominant frequency ν_{dom}. It is shown on the top axis in Fig. 2.

An important difference between Planck's and Debye's theories occurs only at high temperatures, at which the latter predicts a deviation of C_p from the T^3 behavior. In Fig. 2, this deviation occurs at ~10 K. The simple explanation is that at high temperatures the specific heat of solids must approach the classical Dulong–Petit value, which is derived from the fact that each atom in a solid can be viewed as a harmonic oscillator vibrating in the three spatial directions with the total energy

$$w = 3k_BT, \tag{6}$$

leading to a specific heat

$$C_p = 3Nk_B\rho = 3R\rho, \tag{7}$$

where $N = 6.02 \times 10^{23}$ mole^{-1} is Avogadro's number, ρ the mass density, and $R = 8.31$ W s mole^{-1} K^{-1} the universal gas constant. The high-temperature asymptotic value of the Debye theory is the Dulong–Petit value. Another way to understand the deviation of C_p from the T^3 behavior is that the total number density of normal modes (the traveling waves) cannot be larger than $3n$, where n is the number density of the atoms in the solid. This means that in the Debye theory an upper limit ν_D exists for the vibrational frequencies. It is called the Debye cutoff frequency ν_D and is given by

$$\nu_D = (6/8\pi)^{1/3}n^{1/3}v, \tag{8}$$

where v is the average speed of sound and $n^{1/3}$ the inverse of the interatomic spacing. One also talks about the Debye temperature, defined as $\Theta_D = \nu_D h/k_B$.

The elastic waves in solids carry energy, and this leads to conduction of heat (in metals, the electrons will also carry energy, but we restrict the discussion to electric insulators—also called dielectrics). The thermal conductivity of four highly perfect crystals is shown in Fig. 3. It differs significantly from that of the crystal's electromagnetic counterpart, vacuum. In empty space, the electromagnetic waves travel uninhibited, which leads to an infinite thermal conductivity. In the perfect crystal, scattering processes apparently exist at all temperatures, i.e., for waves of all frequencies (see the dominant frequency scale on the top axis in Fig. 2). At temperatures above the conductivity maximum, the anharmonic forces between the atoms lead to an increasingly larger scattering probability for the elastic waves (one talks about phonon–phonon scattering; Peierls, 1929). Below the maximum, this process vanishes in importance as a new process dominates: that of phonon scattering at the sample boundaries (Casimir, 1938). Using an analogy to the heat transport in a gas, we can express the thermal conductivity of the dielectric crystals as

$$\Lambda = \tfrac{1}{3}C_p\rho v l = \tfrac{1}{3}Cvl \qquad (9)$$

where $C = C_p\rho$ is the specific heat of the lattice vibrations or phonons per volume, v their average velocity, and l their mean free path between collisions (most easily visualized in the phonon picture). At low temperatures, when l is determined by collisions with the sample boundaries, $l \sim d$, where d is the sample diameter [see Fig. 1(a)]. Hence one expects $\Lambda \propto C \propto T^3$, as is indeed observed experimentally; see Fig. 3.

3. AMORPHOUS SOLIDS

3.1 Experiments

Having seen how the elastic waves in highly perfect dielectric crystals determine specific heat and thermal conductivity (we address the internal friction later), we now consider amorphous solids, interchangeably also called glasses. These solids can be viewed as having been quenched from the melt, although this is not the only way to produce them. Structurally, they are defined as solids lacking long-range order. Figure 4 shows the profound effect this disorder has on the thermal conductivity of SiO_2 (Zeller and Pohl, 1971). The thermal conductivity decreases monotonically with decreasing temperature, and even at the lowest temperature the Casimir boundary limit has not yet been reached. Most remarkable is the fact that all amorphous solids studied to date have very similar thermal conductivities, regardless of their chemical composition, their bonding, or their atomic arrangements. This is demonstrated in Fig. 5 on seven different amorphous solids with widely different bonding, some disordered on the atomic scale, some with some degree of order, like the polymeric solids. All amorphous solids measured lie in the range spanned by the seven solids shown. They all share the same temperature dependence: The thermal conductivity Λ varies at the lowest temperatures, up to a few kelvins, according to a power law:

$$\Lambda \propto T^\delta, \qquad (10)$$

with δ close to 2.0. The thermal conductivity levels off at intermediate temperatures, sometimes with some kind of dip or plateau, and near room temperature it becomes nearly independent of temperature.

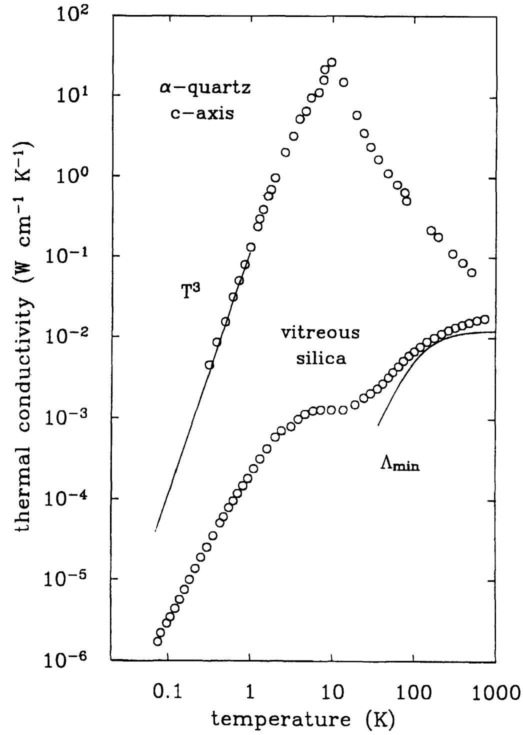

FIG. 4. Thermal conductivity of SiO_2 in the crystalline phase (α quartz) and in the amorphous phase (vitreous silica, a-SiO_2). The solid line indicates the Casimir prediction (boundary scattering). The curve called Λ_{min} is based on the model first proposed for crystals by Einstein (1911). It fits the amorphous data very well above ~50 K, considering that it contains no adjustable parameters.

The profound effect of the amorphous

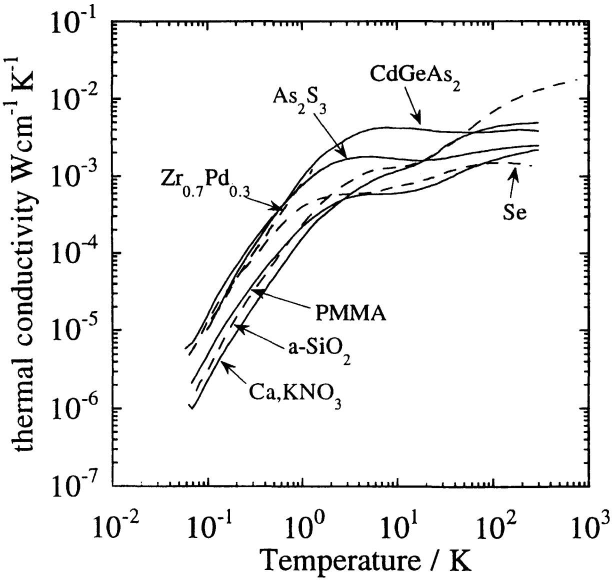

FIG. 5. Universal thermal conductivity of amorphous solids. At temperatures below 1 K the thermal conductivity of all amorphous solids becomes proportional to the square of the temperature. In addition, the magnitude of the thermal conductivity of all amorphous solids lies within a factor of 10 of that of *a*-SiO$_2$ at all temperatures. Zr$_{0.7}$Pd$_{0.3}$ is superconducting in the temperature range shown here and therefore behaves like a dielectric solid in this experiment.

structure on the lattice vibrations is emphasized in Fig. 6. Using Eq. (9), the phonon mean free path has been plotted for three amorphous solids, and for crystalline SiO$_2$ (α quartz). Beginning with quartz, the phonon mean free path l is found to be constant at low temperatures, and close to the sample diameter, as expected for boundary scattering. Above 10 K, l drops precipitously, as a result of phonon–phonon scattering. However, even at 300 K, l ~ 5 nm, which is much larger than the wavelength of the waves that carry most of the heat in quartz at this temperature (see the dashed line in Fig. 6). In the amorphous solids, l is much smaller at all temperatures. Above ~50 K, l becomes comparable to the dominant phonon wavelength. This is a particularly important observation, because a wave that does not even travel the distance of its wavelength before being scattered cannot be considered a wave any more. This is the first indication that the concept of elastic waves needs to be revised in disordered solids.

The specific heat of the amorphous solids shows equally strange, and also universal, behavior, as shown for three solids in Fig. 7. A comparison with the Debye prediction (the dashed lines) based on their known speeds of sound demonstrates the poor fits, which seem worst at the lowest temperatures. A fairly good fit to the experimental specific heat can be obtained with a polynomial,

$$C_p = AT + BT^3, \qquad (11)$$

where A is of very similar magnitude for all amorphous solids (to within about one order of magnitude), and B is close to, though never smaller than, and sometimes more than a factor of 2 larger than, the Debye value based on the speed of sound.

Ultrasonic attenuation in amorphous solids is shown in Fig. 8, measured through internal friction with the technique described in Fig. 1(f). All amorphous solids have a nearly temperature-independent internal friction in the temperature range 1–10 K, followed by a decrease at even lower temperature (Topp and Cahill, 1996). As in the case of the thermal conductivity, the magnitude of the temperature-independent internal friction Q_0^{-1} lies in a narrow range: $10^{-4} \lesssim Q_0^{-1} \lesssim 10^{-3}$.

The internal friction measured on a single crystal of silicon is very much smaller (see Fig. 8), and is determined by losses in the mounting and by accidental impurities (the latter causing the narrow peaks). The internal friction of a perfect crystal is expected to be orders of magnitude smaller yet.

The experiments described in this section have demonstrated strikingly different behavior from that of crystals, both at low (less than a few kelvins) and at high temperatures (~300 K), corresponding to low and high frequencies of the excitations, respectively.

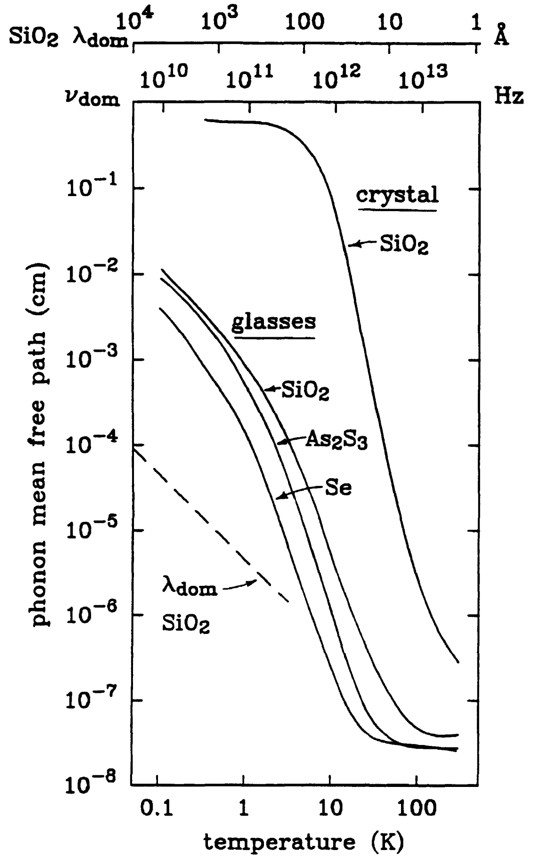

FIG. 6. Phonon mean free path l calculated with the help of Eq. (9). The specific heat used for the calculation is the Debye specific heat based on speeds of low-frequency sound. At the top is shown a scale of the dominant phonon frequency ν_{dom} (see text) and dominant phonon wavelength (for a-SiO_2, using the Debye speed of sound $v = 4.1 \times 10^5$ cm s^{-1}). The dashed line marked "λ_{dom}, SiO_2" shows how much shorter this quantity is below ~10 K than the mean free paths measured in the glasses (Zeller and Pohl, 1971).

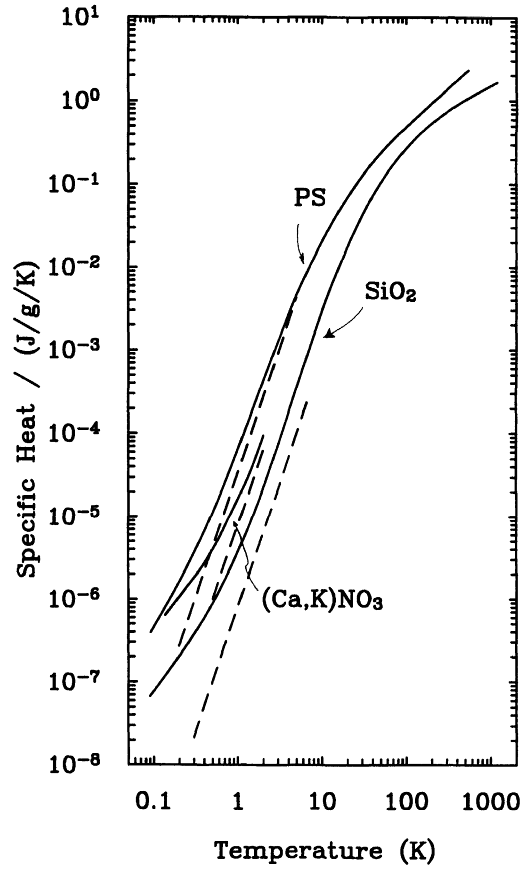

FIG. 7. Specific heat C_p of three amorphous dielectric solids: polystyrene (PS), a-SiO_2, and $Ca_{0.40}K_{0.60}NO_3$ [(Ca,K)NO_3]. Below 1 K, they all approach a linear temperature dependence, the magnitude of which is very similar, independent of chemical composition (and impurities) or sample preparation. The dashed lines are Debye-model calculations based on the measured speeds of sound: top, PS; bottom, SiO_2.

In the next section we review theoretical models proposed to explain these phenomena.

3.2 Theoretical Models

3.2.1 Low-Temperature Properties; Tunneling Model This model is based on the observation that some impurity atoms or molecules in otherwise perfect crystals are known to have more than one equilibrium position or orientation in their lattice sites, between which they move not only by thermally activated jumps over the potential barriers separating the wells but also by a quantum-mechanical process described as tunneling through the barriers (Narayanamurti and Pohl, 1970). As an example we mention KCl containing LiCl in small concentration ($\ll 1\%$); the small Li+ ion has eight equilibrium positions in its K^+ vacancy, which are displaced from the cavity center in the direction toward its eight Cl^- neighbors. The energies required for tunneling to occur between these positions are much smaller than those required for the

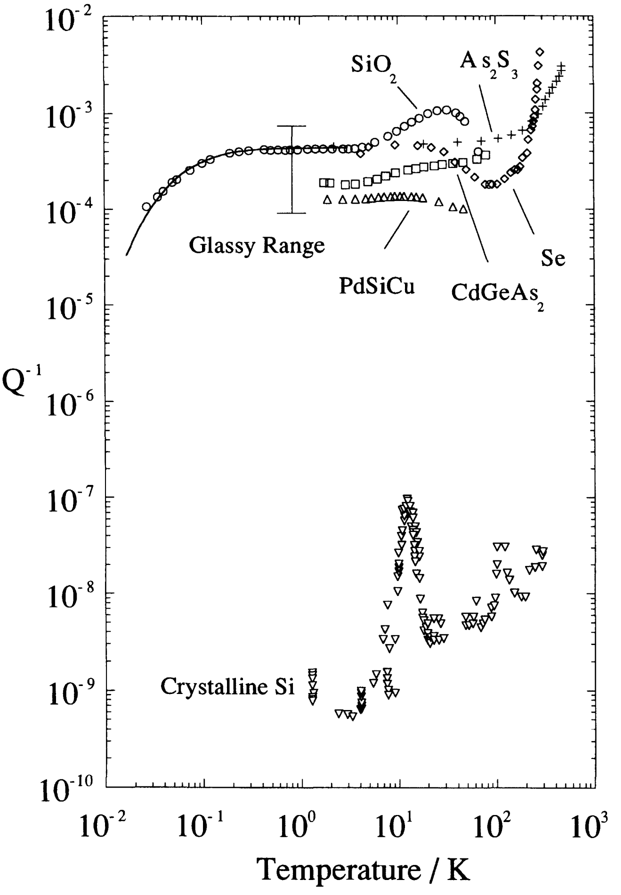

FIG. 8. Internal friction of amorphous and crystalline solids. At temperatures below ~10 K, the internal friction of amorphous solids becomes temperature independent with a magnitude between 10^{-4} and 10^{-3}, marked with the bar labeled "glassy range." This can be contrasted with the behavior found in crystalline solids where the internal friction can be some five orders of magnitude smaller and is then determined by mounting losses and accidental impurities. Data for amorphous solids are taken from Topp and Cahill (1996) while data for crystalline Si are from McGuigan *et al.* (1978). Solid line through the *a*-SiO$_2$ data is a fit based on the tunneling model. The dropoff of Q^{-1} in amorphous solids at the lowest temperatures is frequency dependent, also in agreement with the model. The large Q^{-1} is also observed in normally conducting amorphous metals (e.g., in PdSiCu).

thermally activated jumps between different equilibrium configurations; in KCl:Li, these so-called tunnel splitting energies are of the order of (1 K)k_B, but much smaller splittings have been observed in other doped solids. The thermal population of the excited tunneling states leads to specific-heat anomalies (so-called Schottky anomalies), which peak close to the temperature at which the dominant phonon energy $h\nu_{\mathrm{dom}}$ is equal to the tunnel splitting energy. An example of this is shown later (Fig. 15). Resonant scattering by these states has been observed through a depression of the thermal conductivity in the same temperature range.

In order to explain the low-temperature specific-heat anomaly varying linearly with temperature in all amorphous solids, the assumption was made (Anderson *et al.*, 1972; Phillips, 1972) that atomic or molecular tunneling is common to this state of matter. Because of its random structure, however, neighboring potential minima are separated by barrier heights of different magnitude [Figs. 9(a) and 9(b)]. Furthermore, asymmetries were postulated for the neighboring potential minima [Figs. 9(c) and 9(d)]. Under reasonable assumptions about the distribution of barrier heights and asymmetries, a constant density of states of the tunneling states could be derived with which both the linear specific heat anomaly and the thermal conductivity varying as T^2 below 1 K could be obtained, in agreement with the experiment. This model also predicted the existence of a temperature-independent plateau in the internal friction below a few kelvins, followed by a dropoff at the lowest temperatures, as subsequently indeed discovered in elastic measurements, shown in Fig. 8. The solid curve is the theoretical fit to the experimental data for *a*-SiO$_2$.

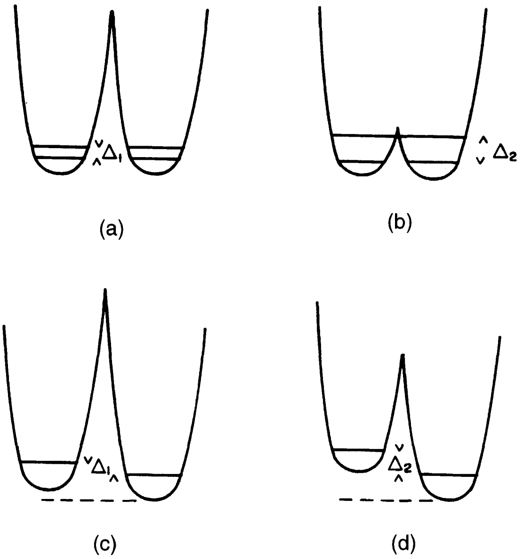

FIG. 9. Tunneling model (Anderson *et al.*, 1972; Phillips, 1972): The splitting of the tunneling states, Δ_i, depends on the barrier height [(a), (b)] and also on the asymmetry of neighboring potential minima [(c), (d)]. At any given temperature, states with one particular splitting Δ_i will be thermally excited.

This model made the additional prediction that the thermal relaxation in an amorphous solid should occur only gradually, a phenomenon that would lead to a time-dependent specific heat as described in Sec. 1. Its origin can be visualized with the help of Fig. 9: At a certain temperature T the states with a certain energy splitting, say Δ_2, will dominate the specific heat, where $k_BT \approx \Delta_2$. Nearly symmetric potential wells (Fig. 9b), leading to states for which Δ_2 is determined mostly by the tunneling probability, will relax much faster than states for which Δ_2 is determined largely by asymmetry (Fig. 9d), and for which the barrier height is high, hence the tunneling probability is small. The specific-heat measurements on a-SiO$_2$ shown in Fig. 10 largely confirmed this prediction. Large thermal-relaxation effects occur on measuring-time scales less than ~1 s, but even on the time scales of hours the thermal relaxation is still measurable.

Figure 10 also illustrates that on time scales exceeding those on which specific-heat measurements are usually carried out, 1–10^2 s, the increase of the specific heat with measuring time becomes rather small [note in Fig. 10 the increase expected between 10^4 s and 10^9 y (!), based on the tunneling model]. Thus, measurements of the kind shown previously (see Fig. 7) may to a good approximation be considered as having been performed in thermal equilibrium, although strictly speaking, the low-temperature data in Fig. 7 will grow forever—as illustrated in Fig. 10.

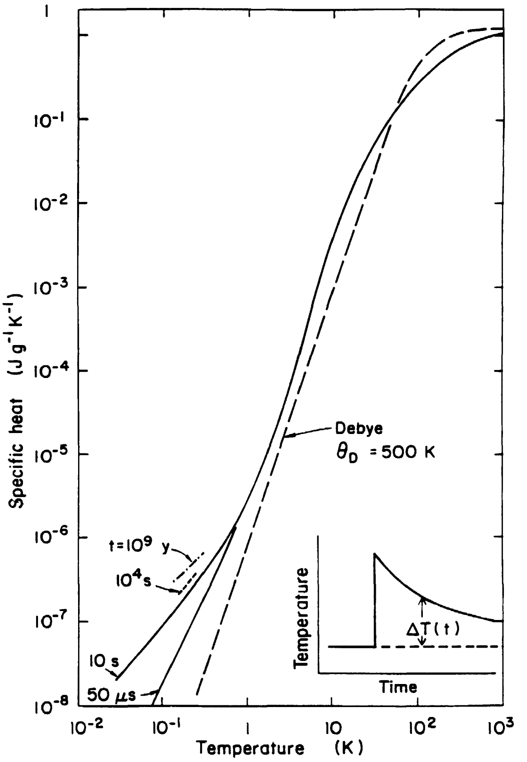

FIG. 10. Specific heat C_p of vitreous silica compared with the Debye theory based on elastic measurements (reviewed in Pohl, 1987). The inset shows an effect observed only below 1 K [see Fig. 1(e)]. After a heat pulse has been added to the sample, which is assumed to be perfectly thermally isolated from its surroundings, the temperature rise ΔT decays logarithmically with time, as an increasing number of the slowly relaxing low-energy excitations take up energy from the rest of the sample. As a consequence, the specific heat appears to increase with time. After 50 μs, the heat is distributed only over the most rapidly relaxing excitations, and thus C_p appears small. After 10 s (the usual time scale of specific heat measurements), C_p has become essentially independent of time, as shown with measurements after 10^4 s and an extrapolation to a time comparable to the age of the solar system.

One other point needs to be mentioned:

The tunneling model postulates the existence of the tunneling states in addition to the collective elastic excitations, the Debye waves. The existence of the latter in amorphous solids has indeed been proven for waves of frequencies as large as 10^{11} s^{-1}. The tunneling model considers the heat to be carried by the elastic waves; the latter are scattered by the tunneling states, which are spatially localized.

The tunneling model predicts that for sufficiently short measuring times, most of the tunneling states will not reach thermal equilibrium; the tunneling states will be "decoupled," and the specific heat should approach its Debye limit (the dashed curve in Fig. 10). This, however, has not been observed. According to the existing measurements, the short-time specific heat always exceeds the Debye prediction. If this excess were caused by strongly coupled defects other than the ones considered in the tunneling model, they should show up in thermal conductivity through their phonon scattering, which, however, does not seem to be the case. This is not understood.

One final point: The tunneling model makes no assumption as to why the tunneling density of states is nearly the same, within an order of magnitude, for all amorphous solids. This point, which appears to be an experimental fact, continues to be the major puzzle in our understanding of the tunneling model.

3.2.2 High-Temperature Properties; Einstein's Model Revisited It has been shown earlier (Fig. 6) that the assumption of elastic waves carrying the heat in amorphous solids at temperatures above, say, 50 K, corresponding to frequencies exceeding several THz, leads to unphysically short mean free paths. Thus, another model for the heat transport is needed in that temperature or frequency range. It has been shown experimentally that a model first considered for crystals by Einstein (1907), and also found by him to be inappropriate for these solids (Einstein, 1911), works very well for amorphous solids, yet for so-far unknown reasons. This is discussed next.

Einstein considered the lattice vibrations of solids to be individual atomic vibrations. Each atom vibrates with a characteristic frequency (the Einstein frequency) about its equilibrium position, with no well-defined phase relation to the motion of its neighbors, in contrast to wavelike motions. By applying the quantum concept to these mechanical oscillators, he successfully explained, at least qualitatively, the decrease of the specific heat from its high-temperature Dulong–Petit value, as the temperature decreased. However, when he applied the random-phase model of oscillations to the thermal conductivity of crystalline dielectrics, the model failed completely: Einstein showed that harmonic oscillators oscillating with equal frequencies but random phases should exchange their energies with the neighboring atoms on the time scale of one-half of a period of vibration. Setting this period equal to that of the Einstein oscillator, he calculated with no free parameters a thermal conductivity, as shown in Fig. 4, where it is labeled as $\Lambda_{\min}$ (this name is explained later). The computed curve disagrees completely both in magnitude and temperature dependence with the experimental thermal conductivity of crystalline SiO_2 (α quartz). This disagreement paved the way for the model of elastic waves proposed in 1912. What Einstein did not notice, however, was the good agreement of his theory with the thermal conductivity of amorphous SiO_2 (vitreous silica), also shown in Fig. 4 (Cahill and Pohl, 1989a). This close agreement suggested that above $\sim$50 K not elastic waves but randomly vibrating atoms were the proper description of the thermal vibrations in this amorphous solid. Note that this model closely resembles the one proposed to describe the heat transport in liquids viewed as a densely packed gas; each atom moves with its thermal velocity through the cavity it occupies, passing on its energy upon colliding with one of its neighbors (Bridgman, 1931). As a test for the model, which had worked so well for one amorphous solid, the Einstein picture was applied to a wide range of amorphous solids; note that in Fig. 5, the high-temperature thermal conductivity of amorphous solids ranges over about one order of magnitude. Using the measured elastic constants, an approximate average Einstein frequency was determined for each solid, and from that and the known average atomic mass and diameter, a thermal conductivity was calculated with no free parameters (Cahill *et al.*, 1992). Figure 11 shows the comparison of the cal-

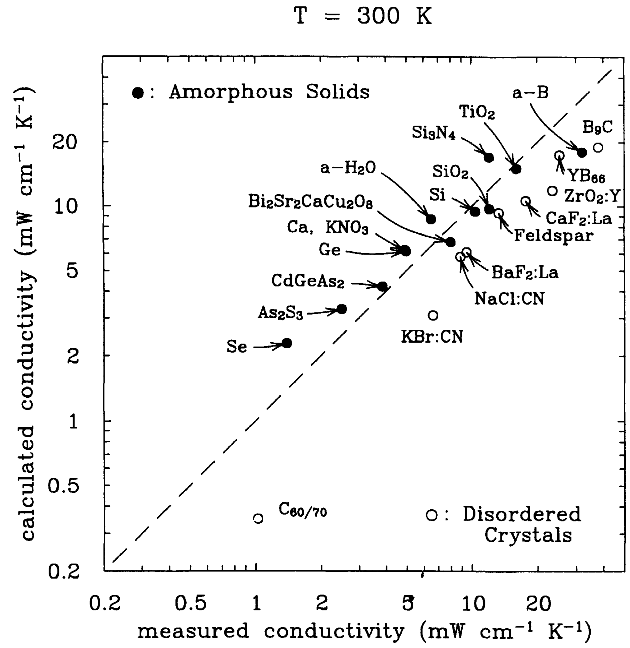

FIG. 11. Composite plot of the calculated minimum thermal conductivity at 300 K vs the measured values; all points lie close to the dashed line, which is drawn to indicate perfect agreement between experiment and Einstein's model (Cahill *et al.*, 1992). Amorphous solids are shown as filled circles and disordered crystals as open circles. The disordered crystals are single crystals, with the exception of $C_{60/70}$, which was compacted from microcrystalline powder (Olson *et al.*, 1993). While for perfect and also for lightly disordered crystals data to the right of the dashed line are common, no disordered (fully dense) solids have yet been found with thermal conductivities significantly smaller than the Einstein prediction (i.e., with data points to the left of the dashed line). This justifies calling this prediction the minimum thermal conductivity; see Fig. 4.

culated thermal conductivity of 11 amorphous solids with that measured at 300 K (ignore the open circles for the moment). The dashed line is drawn to indicate what would be perfect agreement between theory and experiment. Given the simplicity of the model, the agreement observed is quite satisfactory. It suggests that this model is indeed appropriate, although an understanding of the time for energy exchange is still lacking.

4. HIGHLY DISORDERED CRYSTALS

Intermediate between perfect crystals and amorphous solids are the disordered crystals. In general, their lattice vibrations depend critically on the type of disorder, although in most cases a combination of propagating elastic waves and of somehow localized excitations provides an adequate description. In this section, it is shown that the lattice vibrations observed in amorphous solids can also occur in certain highly disordered crystals. Thus, the universality stressed above for amorphous solids also extends to disordered crystals. Following common usage, we call these vibrations "glasslike." Since the disorder in crystals can be controlled, in contrast to that of amorphous solids, one may hope that a study of the glasslike lattice vibrations will provide a better handle for their understanding.

The evolution of glasslike lattice vibrations with increasing disorder is illustrated on the mixed crystal $Ba_{1-x}La_xF_{2+x}$ for increasing x. Its crystal structure is the cubic fluorite structure, Fig. 12, like that of BaF_2, as confirmed by x-ray diffraction. The third F ion introduced through the addition of the LaF_3, labeled F_i, occupies a vacant interstitial site in the center of a cube formed by eight fluorine ions. For $x = 8 \times 10^{-3}$, the thermal conductivity is nearly as expected for a perfect crystal; see Fig. 13 (Cahill *et al.*, 1992). As x increases, the thermal conductivity decreases, and approaches the low-temperature T^2 temperature dependence with the magnitude characteristic for amorphous solids. At high temperatures, it approaches a temperature-independent value with a magnitude close to that predicted by the Einstein model (see the curve labeled $\Lambda_{\min}$). Note that as x approaches 1.0 (pure LaF_3), the conductivity rises again and shows the temperature dependence characteristic for crystals. Measurements of the low-temperature specific heat and internal friction have also confirmed the glasslike lattice vibrations inferred

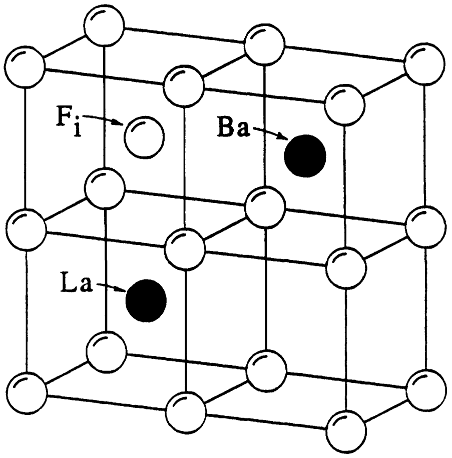

FIG. 12. Structure of an isolated defect in $Ba_{1-x}La_xF_{2+x}$. The La^{+++} ion substitutes on a Ba^{++} site. The charge-balancing fluorine interstitial, marked F_i, is shown in a nearest-neighbor position.

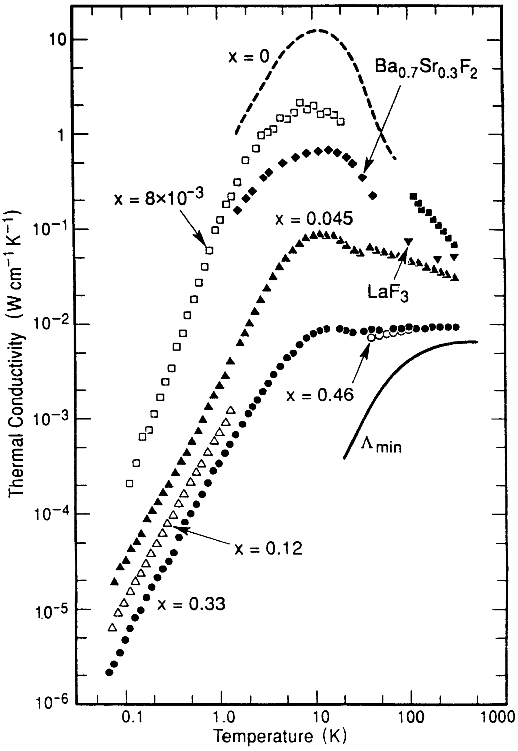

FIG. 13. Thermal conductivity of the mixed crystal $Ba_{1-x}La_xF_{2+x}$. Thermal conductivity for $x = 0$ below 30 K is shown as a dashed line. Solid line labeled Λ_{min} is the thermal conductivity calculated using Einstein's model as described in the text. Note the three data points for pure LaF_3. (Cahill *et al.*, 1992). The mixed crystal $Ba_{0.7}Sr_{0.3}F_2$ does not have glasslike lattice vibrations.

from thermal conductivity (Cahill and Pohl, 1989b).

Figure 13 also contains an example of a disordered crystal without glasslike lattice vibrations: The thermal conductivity of $Ba_{0.7}Sr_{0.3}F_2$, in which crystal the cations are distributed randomly over the cation sites, is that of a crystal with point defects.

Another example of glasslike lattice vibrations is shown in Fig. 14. $Or_1Ab_{33}An_{66}$ is a feldspar (labradorite), whose chemical composition can be described as a mixture of orthoclase (Or, which is short for $KAl\text{-}Si_3O_8$), albite (Ab, $NaAlSi_3O_8$), and anorthite (An, $CaAl_2Si_2O_8$). Its crystal structure can be approximately described as a three-dimensionally linked network of AlO_4 and SiO_4 tetrahedrons, with charge balancing alkali- and alkaline-earth ions occupying interstitial sites. Feldspars are of great geophysical interest, since they constitute 60% of the earth's crust. Thus, their very small, glasslike thermal conductivity, strikingly illustrated through the comparison with two amorphous solids, plays an important role in determining the heat flow out of the earth's interior.

The observation of glasslike lattice vibrations in the mixed BaF_2:LaF_3 crystal and in the feldspar suggests that large concentrations of interstitial ions may be important in causing these vibrations. However, glasslike lattice vibrations are also observed in disordered crystals without interstitials: The open circles in Fig. 11 show the comparison of the measured high-temperature thermal conductivity of a variety of disordered crystals with that predicted by the Einstein model. In addition to feldspar and the BaF_2:LaF_3 mixed crystals, glasslike lattice vibrations are, for example, established in yttria-stabilized zirconia, which contains vacancies in the oxygen sublattice, or in the mixed crystals of alkali halide–alkali cyanides, in which the oblong CN^- ions occupy halogen sites in the cubic lattice. All of the disordered crystals in Fig. 11 have random strains in common. However, such random strains alone are not sufficient to produce glasslike lattice vibrations either, since mixed crystals of $(BaF_2)_{1-x}(SrF_2)_x$, which also have random

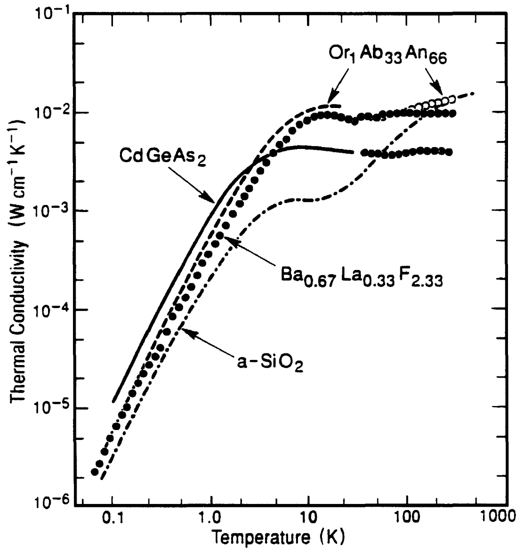

FIG. 14. Thermal conductivity of two disordered crystals, the feldspar labradorite (Or$_1$Ab$_{33}$An$_{66}$) and Ba$_{1-x}$La$_x$F$_{2+x}$, $x = 0.33$, compared with that of two amorphous solids, *a*-SiO$_2$ and CdGeAs$_2$ (Cahill *et al.*, 1992).

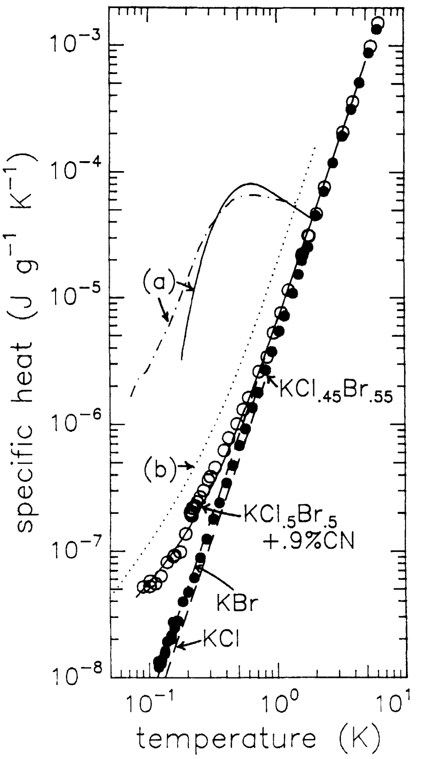

FIG. 15. The Schottky anomaly in KCl + 0.06 mol% CN [curve (a), dash-dotted line] is close to the expected contribution due to 0.06 mol% CN$^-$ (solid curve), indicating a narrow energy spectrum for the CN$^-$ tunneling states. In the similarly doped but also heavily strained crystal, (KCl)$_{0.5}$(KBr)$_{0.5}$ + 0.9 mol% CN, the specific heat is glasslike (open circles). The Schottky anomaly is replaced by a linear contribution, indicating vastly different energy spectra for the CN$^-$ ions, with a corresponding smaller contribution to the specific heat at any given temperature. The solid curve through the open circles is a fit to the data: $c_1T + (c_3 + C_D)T^3$ with $C_1 = 5 \times 10^{-7}$ J g^{-1} K^{-2}, $c_3 = 1.2 \times 10^{-6}$ J g^{-1} K^{-4}, and $C_D = 4.8 \times 10^{-6}$ J g^{-1} K^{-4} from the (KCl)$_{0.45}$ (KBr)$_{0.55}$ data. Also shown are measurements for a mixed crystal with glasslike excitations, (KBr)$_{59}$(KCN)$_{41}$ (curve b), crystalline T^3 behavior for (KCl)$_{0.45}$(KBr)$_{0.55}$ (solid circles), and Debye predictions for KBr and KCl. Both of the latter are well confirmed by experiment (see Fig. 2 for KBr). (Watson, 1995).

strains, do not have glasslike lattice vibrations, as shown in Fig. 13. The Debye model is also found to give the proper description for the low-temperature specific heat of (KBr)$_{1-x}$(KCl)$_x$ shown in Fig. 15.

Next, we show that if a crystal contains tunneling defects, like alkali halide crystals with small concentrations of CN$^-$ ions, which have tunneling states of well-defined energy splittings, then random strains can lead to a spreading of those tunneling states very much as in amorphous solids; see Fig. 15 (Watson, 1995). In a KBr or a KCl crystal containing a small concentration of KCN, the CN$^-$ tunneling states lead to a specific-heat anomaly of the Schottky type, which peaks at 0.6 K. However, if a small concentration of CN$^-$ is dissolved in the highly strained, mixed KBr:KCl crystal, a linear specific-heat anomaly arises, very similar to that found in amorphous solids or certain highly disordered solids; see for comparison the specific heat of (KBr)$_{0.59}$(KCN)$_{0.41}$, which has glasslike lattice vibrations (DeYoreo *et al.*, 1986). The comparison of the magnitude of this linear specific heat with that of the Schottky anomaly expected for the same CN$^-$ concentration [about 15 times larger than curve (a) in Fig. 15] shows that in this strained crystal most of the CN$^-$ ions have been frozen out so that they do not contribute to the specific heat at low temperatures. A very small fraction, however, has retained low-energy tunneling splittings as known in amorphous solids. This finding suggests that random strains can affect the reorientation of molecular entities with rotational degrees of freedom. The question whether this mechanism can explain the origin of the tunneling states in amorphous solids has to be left unanswered.

We return once more to Fig. 13. It illustrates how with increasing disorder the high-temperature thermal conductivity decreases and approaches the value predicted on the basis of Einstein's model. The same behavior has been found in the other disordered crystals shown in Fig. 11: The measured thermal conductivity decreases with increasing disorder without, however, ever dropping below the value predicted by Einsteins model. Although we have no explanation for this, the observations are by now sufficiently numerous to conclude that the value predicted by this model does represent a lower limit. This is our justification for calling it $\Lambda_{\min}$, the minimum thermal conductivity. Furthermore, $\Lambda_{\min}$ seems to describe closely the thermal conductivity of the amorphous phase above ~50 K (see, for example, Fig. 4). As a practical conclusion it follows that there is little hope of ever finding a disordered crystal with a thermal conductivity smaller than that of its amorphous phase.

5. CLOSING REMARKS

The concept that the elastic vibrations of solids can be described by waves, as proposed by Born and von Kármán (1912), and by Debye (1912), is inadequate for all amorphous solids, and also for a large—and growing—number of highly disordered crystals. These solids require for their description, in addition to waves, localized excitations or tunneling states at low energies. At high energies, the proper description of the vibrations is that of individual atoms vibrating with random phases, passing their energy along to their neighbors within half a period of vibration, in a random walk. This difference between crystals and highly disordered solids is illustrated schematically in Fig. 16. In this doubly logarithmic plot of the density of states of the excitations, the heavy solid line, which varies proportionally to the second power of the frequency ν, represents the elastic waves as described in Debye's model. At high frequencies, the line ends at a maximum frequency, called the Debye cutoff frequency [Eq. (8)]. For disordered solids, the density of states in the low-energy region is dominated by the tunneling states with a nearly constant value of their density of states, and a magnitude varying only within approximately one order of magnitude among chemically different solids. In the frequency regime above a few terahertz (marked by the vertical line in Fig. 16), the concept of propagating waves loses its significance and has to be replaced by short-lived vibrations of individual atoms.

Although a phenomenological description

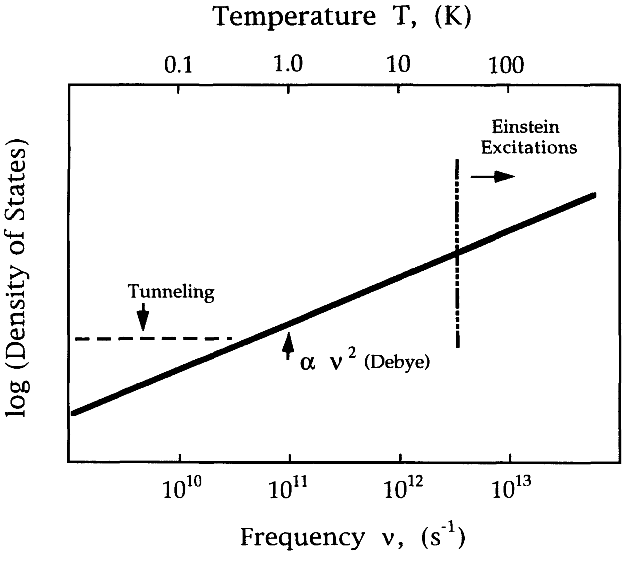

FIG. 16. Schematic summary of the present knowledge of the lattice vibration of disordered solids. The heavy solid line is based on the Debye model: All vibrations are waves, with a density of states varying as the square of the frequency ν, up to a cutoff frequency [given by Eq. (8)]. In disordered solids, the density of states is increasingly dominated by the localized tunneling states as the frequency decreases. Their density of states is nearly independent of frequency (dashed line). Above a certain frequency, traveling waves cease to exist in disordered solids and are replaced by localized excitations on the atomic scale. The excitations pass on their energy from atom to atom in a random walk. In the figure, they are labeled "Einstein Excitations." The upper horizontal scale marks the temperature at which excitations of certain frequency are thermally excited.

of the lattice vibrations of disordered solids has been created, many questions remain unanswered:

1. What is the physical nature of the tunneling entities? Specifically, can we apply the concepts developed from the study of disordered crystals to amorphous solids?
2. Why are the phenomena connected with the tunneling states so similar in all these solids, and specifically, why is their density of states so similar in all these disordered solids?
3. Which mechanism limits the high-temperature thermal conductivity of all dielectric solids to that of their amorphous phases?
4. Is there a connection between the tunneling states, which determine the properties at low temperatures, and the high-energy atomic vibrational states, which determine the heat flow at high temperatures?

It is remarkable indeed that the important and widely studied subject of the lattice vibrations of solids still contains so many unsolved mysteries.

GLOSSARY

Amorphous Solids: Solids in which the atoms are disordered topologically, i.e., their atoms lack long-range order. They can be viewed as frozen liquids.

Disordered Solids: Solids in which the atomic arrangement differs from that of perfect crystals. In this article, disordered solids are either amorphous (glassy), in which the atoms lack long-range order, or disordered crystalline, in which different atomic species are distributed randomly over well-defined crystalline lattice sites.

Elastic Waves: Sound waves, which can be either transverse or longitudinal. In a wave the motion of atoms is correlated; i.e., the atoms vibrate with fixed phase relations relative to their neighbors.

Internal Friction: A quantity describing the dissipation of vibrational energy. Defined in Eq. (3) through the decay of the vibrational amplitude. The internal friction Q^{-1} is related to the attenuation of an elastic (sound) wave by $Q^{-1} = (v/\omega)\alpha$, where v is the speed of the wave, ω is its frequency ν multiplied by 2π, and α is the absorption constant of the wave, which describes the attenuation of the elastic power carried in the wave.

Phonon: Quantum of elastic energy, analogous to the quantum of electromagnetic energy, which is called a photon. One may view a phonon as a wave packet moving through the solid with the speed of sound.

Quartz: Crystalline form of silica, SiO_2. The most common form is α quartz. The amorphous form is called vitreous silica, or briefly a-SiO_2.

Rotational Degrees of Freedom: Molecules in a gas can not only translate, but they can also rotate. The translational motion is frozen in most solids. Rotation can, however, occur. In that case, one talks about rotational degrees of freedom in the solid.

Specific Heat: The proportionality factor between the thermal energy (heat) added to a sample, its mass, and the resultant temperature rise, as defined in Eq. (2).

Stabilized Zirconia: ZrO_2 has a cubic crystal structure only near its melting point. As it cools to room temperature, its structure changes, and the crystal shatters. Through the addition of other oxides, like CaO or Y_2O_3, in small concentrations (~10 mol%), the cubic phase is stabilized to room temperature. Stabilized zirconia is used as an artificial gem stone of great hardness, resembling diamond in its optical properties.

Thermal Conductivity: The proportionality factor between the heat current density and the temperature gradient, as defined in Eq. (1).

Transducer: Any device used to change energy from one form to the other. In this article, it converts electric to elastic energy.

Tunneling: A quantum-mechanical process that demonstrates the wave nature of material particles (matter wave, wave-particle duality). Classically, a particle cannot cross into a region in which its potential energy exceeds the total energy of the particle. For example, a ball can never penetrate a solid wall. Quantum mechanically, however, the matter wave does have a nonzero probability to enter such an energetically forbidden region, just like an evanescent wave in the case of light waves.

Vibrational States: The vibrational motion of atoms is described by quantum-mechanical vibrational states.

Works Cited

Anderson, P. W., Halperin, B. I., Varma, C. M. (1972), *Philos. Mag.* **25,** 1–9.

Born, M., von Karman, Th. (1912), *Phys. Z.* **13,** 297–309.

Bridgman, P. W. (1931), *The Physics of High Pressure,* New York: MacMillan, pp. 316–320.

Cahill, D. G. (1990), *Rev. Sci. Instr.* **61,** 802–808.

Cahill, D. G., Pohl, R. O. (1989a), *Solid State Commun.* **70,** 927–930.

Cahill, D. G., Pohl, R. O. (1989b), *Phys. Rev. B* **39,** 10477–10480.

Cahill, D. G., VanCleve, J. E. (1989), *Rev. Sci. Instr.* **60,** 2706–2710.

Cahill, D. G., Watson, S. K., Pohl, R. O. (1992), *Phys. Rev. B* **46,** 6131–6140.

Casimir, H. B. G. (1938), *Physica* **5,** 495–500.

Debye, P. (1912), *Ann. Phys. (Leipzig)* **39,** 789–839.

DeYoreo, J. J., Knaak, W., Meissner, M., Pohl, R. O. (1986), *Phys. Rev. B* **34,** 8828–8842.

Einstein, A. (1907), *Ann. Phys. (Leipzig)* **22,** 180–190.

Einstein, A. (1911), *Ann. Phys. (Leipzig)* **35,** 679–694.

McGuigan, D. F., Lam, C. C., Gram, R. Q., Hoffman, A. W., Douglass, D. H., Gutche, H. W. (1978), *J. Low Temp. Phys.* **30,** 621–629.

Narayanamurti, V., Pohl, R. O. (1970), *Rev. Mod. Phys.* **42,** 201–236.

Olson, J. R., Topp, K. A., Pohl, R. O. (1993), *Science* **259,** 1145–1148.

Peierls, R. (1929), *Ann. Phys. (Leipzig)* **3,** 1055–1101.

Phillips, W. A. (1972), *J. Low Temp. Phys.* **7,** 351–360.

Topp, K. A., Cahill, D. G. (1996), *Z. Phys. B, Cond. Matter* **101,** 235–245.

Watson, S. K. (1995), *Phys. Rev. Lett.* **75,** 1965–1968.

Zeller, R. C., Pohl, R. O. (1971), *Phys. Rev. B* **4,** 2029–2041.

Further Reading

Ashcroft, N. W., Mermin, N. D. (1976), *Solid State Physics,* New York: Holt, Rinehart, Winston.

Braginsky, V. B., Mitrofanov, V. P., Panov, V. I. (1985), *Systems with Small Dissipation,* Chicago: University of Chicago Press.

Cahill, D. G., Pohl, R. O. (1988), "Lattice Vibrations and Heat Transport in Crystals and Glasses," *Annu. Rev. Phys. Chem.* **39,** 93–121.

Elliott, S. R. (1984), *Physics of Amorphous Materials,* London: Longman.

Hunklinger, S., Raychaudhuri, A. K. (1986), "Thermal and Elastic Anomalies in Glasses at Low Temperatures," *Prog. Low Temp. Phys.* **9,** 265–344.

Phillips, W. A. (Ed.) (1981), *Amorphous Solids: Low Temperature Properties,* New York: Springer.

Phillips, W. A. (1987), "Two-Level States in Glasses," *Rep. Prog. Phys.* **50,** 1657–1708.

Pohl, R. O. (1987), "Lattice Vibrations of Solids," *Am. J. Phys.* **55,** 240–246.

Zallen, R. (1983), *The Physics of Amorphous Solids,* New York: Wiley.

VIDEO DISPLAY TERMINALS

See DISPLAY DEVICES

VISCOSITY AND DIFFUSION

H. E. KING, JR., P. N. SEGRÈ, AND M. APPEL, *Corporate Research Science Laboratories, Exxon Research and Engineering Company, Annandale, New Jersey, U.S.A.*

Introduction 239
1. **Theory of Viscosity and Diffusion** 240
1.1 Continuum Hydrodynamics ... 240
1.1.1 Viscosity 240
1.1.1.1 Simple Motivation for Continuum Transport 240
1.1.1.2 Navier–Stokes Equation 241
1.1.1.3 Stokes Equation 243
1.1.2 Diffusion 243
1.1.2.1 Continuum Transport: Fick's Laws of Diffusion 243
1.1.2.2 Self-Diffusion 244
1.1.2.3 Diffusion Equation 244
1.1.2.4 Diffusion as a Random Walk 245
1.1.3 Stokes–Einstein Equation: Viscosity and Diffusion in Liquids 245
1.2 Atomistic Theories and Computer Simulations 246
1.2.1 Simple Kinetic-Theory Approximation 246
1.2.2 Boltzmann Development 248
1.2.3 Lennard-Jones Interactions 248
1.2.4 Hard-Sphere Model—Enskog Development 249
1.2.5 Hard-Sphere Model—Computer Simulations at High Density 249
1.2.6 Rough Hard-Sphere Model 250
1.2.7 Mode-Coupling Theory 251
1.2.8 Free-Volume Theory 252
1.3 Scaling Theories for Simple Liquids 253
2. **Measurement Techniques—Viscosity** 254
2.1 Rotational Viscometers 254
2.2 Capillary Viscometers 255
2.3 Falling/Rolling-Sphere Viscometer 255
2.4 Vibrational Viscometers 255
3. **Measurement Techniques—Diffusion** 256
3.1 Measurement of Mutual Diffusivity 256
3.2 Measurement of Self-Diffusivity 256
3.2.1 Molecular Tracer Methods 256
3.2.2 NMR 256
3.2.3 Scattering Techniques 257
3.2.3.1 Neutron Scattering 257
3.2.3.2 Light Scattering 257
4. **Materials/Systems** 258
4.1 Hard-Sphere Colloidal Suspensions 258
4.1.1 Diffusion 259
4.1.2 Viscosity 259
4.1.3 Stokes–Einstein Relation 260
4.1.4 Viscosity and Structural Relaxation 261
4.2 Polymers 261
4.2.1 Dilute Solutions 262
4.2.2 Melts 263
4.2.2.1 Viscosity of Melts 263
4.2.2.2 Diffusion of Melts 264
Glossary 266
Works Cited 268
Further Reading 270

INTRODUCTION

The subjects of viscosity and diffusion are parts (thermal conductivity is also included) of the general phenomena of transport in gases and liquids. As we discuss, viscosity is concerned with momentum transport and diffusion with mass transport. The length scale over which this transport is defined is macroscopic; hence, these phenomena are fundamentally long-wavelength properties. Thus their definitions and defining equations are part of the continuum description of matter.

When the structural relaxation time is short compared to time scales on which transport occurs, as is the case for the fluids we consider, this falls into the field of hydro-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

dynamics. As we will see, the fundamental definitions of viscosity and diffusion are imbedded in the fundamental hydrodynamic equations. If the transport rate approaches and exceeds the structural relaxation time, then one enters the viscoelastic regime. We only briefly consider this. It is a subject in itself, and we only need address it here in defining the shear rates at which slowly relaxing molecules, such as polymers, can be considered as acting as simple fluids.

Although the phenomena under consideration are continuum properties, we wish to understand their origins in terms of the intermolecular interactions that govern the motions of the individual particles of the fluid. This is a problem in statistical mechanics, one for which solutions to the gas-phase transport equations have been well developed. However, these analytical solutions to the equations of motion involve only binary collisions. Ultimately, as we approach higher densities, such as those of ordinary liquids, this approximation loses its validity, and the problem remains poorly understood. We touch upon the current thoughts on how this fundamental problem may be solved in the future through the use of mode-coupling equations.

In lieu of fundamental theories, the most common approach for describing the ordinary liquid state is scaling theories. These theories incorporate an approximate description of the effects of temperature, composition, or pressure on the change in viscosity or diffusion. These are in common use, and as we shall see, they range from close to fundamental descriptions of the underlying physics to strictly phenomenological equations used to systematize data.

One of the most important relationships between viscosity and diffusion for ordinary liquids is the Stokes–Einstein equation. This equation combines hydrodynamic drag on a spherical particle with the statistical mechanics of Brownian motion to yield a simple inverse relationship between viscosity and diffusion [see Eq. (26)]. Experimental tests of this relationship indicate that it is accurate over length scales ranging from atomic to macroscopic, and it finds wide application in interrelating particle diffusivity, particle size, and solvent viscosity.

1. THEORY OF VISCOSITY AND DIFFUSION

1.1 Continuum Hydrodynamics

1.1.1 Viscosity

1.1.1.1 Simple Motivation for Continuum Transport. We first consider viscosity. Imagine two plates, parallel to one another and separated by a distance Δz and perpendicular to the z axis of a coordinate system (Fig. 1). The region between the plates is filled with a fluid (gas or liquid). If we suddenly accelerate the top plate in the y direction to some velocity V, the intervening fluid between the plates will respond by also moving. In a short time, we find that a constant force F will maintain a constant velocity for the plate of area A. The resulting steady-state velocity gradient in the fluid is linear, going from zero at the stationary plate up to V at the moving one. We can write an equation for the required force F as follows:

$$F/A = \eta \Delta V/\Delta z. \tag{1}$$

This is Newton's law of viscosity. It defines the viscosity η as a proportionality constant between the force per unit area, F/A, and the velocity gradient, $\Delta V/\Delta z$. The force per unit area is a shear stress, and the velocity gradient is a shear rate, $\dot{\gamma}$. Several qualities of viscosity are apparent from this equation. First, we find that the dimensions of viscosity are F/A times time; in SI units this is given in Pa · s, whereas in cgs units the poise is used, which is dyne·s/cm^2. The second important feature is that in a plot of shear rate versus shear stress, Eq. (1) says that a straight line will result, the slope of which is the viscosity. For all fluids we consider here, this is true, and these are designated Newtonian fluids. Non-Newtonian behavior, seen in polymer melts for example, often results in curvature on such plots. See Bird *et al.* (1960) for a further discussion.

One can immediately see that the simplified geometry we have used to define viscosity can be generalized. In particular, if we write the left-hand side of Eq. (1) as a component of the shear stress tensor, with the second index indicating the direction of the force and the first designating the plane in

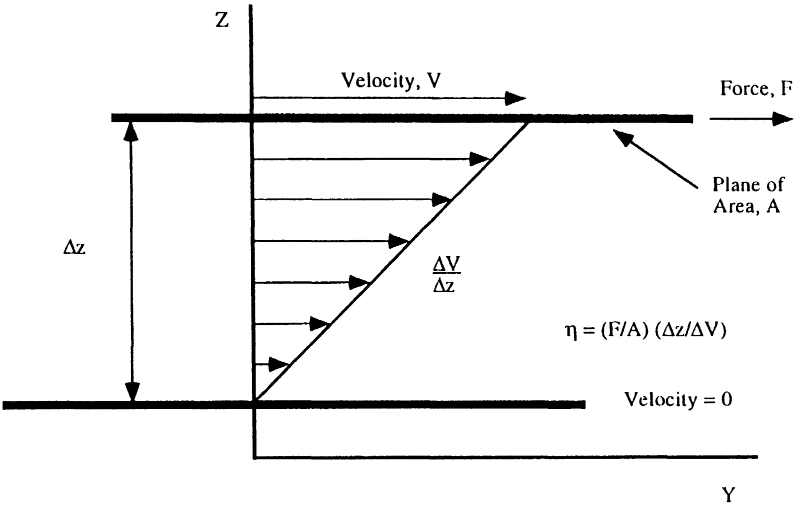

FIG. 1. Definition of the viscosity in the hydrodynamic limit. Movement of upper plane at constant velocity V leads to velocity gradient $\Delta V/\Delta z$ in the intervening fluid. The viscosity is the proportionality constant between the force per unit area and the velocity gradient produced. This defines the viscosity for a Newtonian fluid, one for which changes in the magnitude of the shear rate, equal to $\Delta V/\Delta z$, do not change the viscosity η.

which the response is measured, we can generalize the above equation:

$$\sigma_{zy} = -\eta \frac{dV_y}{dz}. \tag{2}$$

A further generalization is to write this equation in terms of the momentum flux. This will aid our understanding of the molecular origins of viscosity as well as connect us to the upcoming description of the general hydrodynamic approach to the role of viscosity in mass and momentum balances in a fluid. We write

$$\sigma_{zy} = -\mu \frac{d(\rho V_y)}{dz}, \tag{3}$$

where we have defined the kinematic viscosity μ as the ratio of viscosity to the density. Now we see that the gradient is written with respect to the momentum density ρV_y.

1.1.1.2 Navier-Stokes Equation. To understand the role of viscosity in the broader context of hydrodynamics, we need to understand the origin of the Navier–Stokes equation. This equation results from a consideration of mass and momentum balances. Imagine a cubic fluid element, and consider the mass flow into and out of its six faces. For each face the rate of mass inflow is the velocity at that point times the density. Similarly, the outflow is the velocity at the exit face times the density. For example, the net flow through the face perpendicular to the V_x direction is $\Delta y \Delta z[\rho V_x(x) - \rho V_x(x + \Delta x)]$. The total is then the sum over all three sets of faces. The resulting equation, written for the reference frame that moves with the fluid, is

$$\frac{d\rho}{dt} = -\rho(\nabla \cdot \mathbf{V}). \tag{4}$$

A similar procedure is used to determine the equation of motion. For our fictive cube, we consider the rate of momentum accumulation, which equals the rate of momentum inflow minus the outflow plus any forces acting on the system (such as gravity or pressure). Momentum flows in and out via two means: bulk fluid flow and stress transfer. The fluid-flow component depends upon the velocity at which the momentum density ρV enters and leaves. However, this situation is a bit more complex than that for mass. Each of the three components of ρV can be transferred by one of the three velocity components. Therefore, flow across all three pairs of faces contributes to the x-momentum flow, and there are similar contributions for y and z, a total of three equations with three components each. A similar situation results from consideration of the stress-transfer mechanism. This component is transferred via the stress, which we introduced earlier in Eq. (3), except that here we must additionally consider the on-diagonal stress-tensor terms. As shown above, each component of the stress tensor is itself the momentum flux. Hence our tabulation of the inflow and outflow relates to the spatial variation of this tensor at each of the faces of the imaginary cube. Thus for each component, x, y, and z, of momentum, the resulting equations contain three terms each for bulk-fluid-flow and stress-transfer momenta. We write out one of these for the x direction:

$$\begin{aligned}\frac{\partial}{\partial t}\rho V_x = &-\left(\frac{\partial}{\partial x}\rho V_x V_x + \frac{\partial}{\partial y}\rho V_y V_x + \frac{\partial}{\partial z}\rho V_z V_x\right)\\ &-\left(\frac{\partial}{\partial x}\sigma_{xx} + \frac{\partial}{\partial y}\sigma_{yx} + \frac{\partial}{\partial z}\sigma_{zx}\right)\\ &-\frac{\partial P}{\partial x} + \rho g_x. \qquad (5)\end{aligned}$$

There are two additional terms included in Eq. (5) to account for pressure forces and gravitational forces. This is one of three equations of motion. Clearly, the complete set are better written in vector notion. The result is

$$\frac{\partial}{\partial t}\rho \mathbf{V} = -(\nabla\cdot\rho\mathbf{V}\mathbf{V}) - (\nabla\cdot\sigma) - \nabla P + \rho\mathbf{g}. \qquad (6)$$

To solve this equation we need additional information, namely, the relationship between the stress-tensor components and velocity. These are of particular interest here since this is the basis for viscous effects in fluid flow. [Absent these terms one would have inviscid flow that is quite unrealistic in several ways; for example, a sphere moving in such a flow suffers no drag force (d' Alembert's paradox; Landau and Lifshitz, 1959).] We do not derive the needed relationships here (see Hirschfelder *et al.*, 1954) but give the results for two of the six independent stress-tensor components:

$$\begin{aligned}\sigma_{xx} &= -2\eta\frac{\partial V_x}{\partial x} + (\tfrac{2}{3}\eta - \kappa)(\nabla\cdot\mathbf{V});\\ \sigma_{xy} = \sigma_{yx} &= -\eta\left(\frac{\partial V_x}{\partial y} + \frac{\partial V_y}{\partial x}\right). \qquad (7)\end{aligned}$$

As can be seen, the on-diagonal terms now contain a new variable. This is the bulk viscosity κ, which describes the result of energy dissipation associated with volume (pressure) change. This term can be important when the rates of expansion and contraction are large or periodic such as in shock waves or ultrasound transmission (see Batchelor, 1967, pp. 155–156). However, it is common that the rate of shear dominates and that the bulk viscosity can be neglected. We do not consider it further. In our subsequent discussions, and in those likely to be encountered in other publications, the term viscosity is taken to mean the shear viscosity η.

Using Eq. (6), and mass balance, Eq. (4), a full description of fluid motion can be made. This includes the spatial and time derivatives of velocity, density, and pressure. Isothermal conditions are assumed, but an extension to include temperature variations is also possible (see Bird *et al.*, 1960, p. 310). Taking the case that the fluid is incompressible (a common and accurate assumption in nearly all instances, i.e., that the pressure generated during flow does not compress the fluid) and Newtonian viscosity, we obtain from Eq. (6) the celebrated Navier–Stokes equation,

$$\rho\frac{d\mathbf{V}}{dt} = \rho\frac{\partial\mathbf{V}}{\partial t} + \rho\mathbf{V}\cdot\nabla\mathbf{V} = \eta(\nabla^2\mathbf{V}) - \nabla P + \rho\mathbf{g}. \qquad (8)$$

This is the starting point for most hydrodynamic solutions to fluid flow. Indeed, a solution of this equation for steady-state flow, $dV/dt = 0$ in the absence of gravitational effects, will yield the linear velocity profile and a simple relation between viscosity and the stress and velocity gradients we discussed for Fig. 1.

A great difficulty in solving Eq. (8) comes about when the nonlinear term $\mathbf{V}\cdot\nabla V$ is nonnegligible. In most cases of interest for measuring viscosities, we can neglect the ΔP and $\rho\mathbf{g}$ terms. The condition that the nonlinear term is smaller than the remaining damping term is $\rho V^2/l < \eta V/l^2$ or $\rho Vl/\eta \ll 1$, where l is some characteristic length over which the velocity V varies. This ratio is defined as the Reynolds number

$$Re = \rho Vl/\eta = Vl/\mu, \qquad (9)$$

which plays a crucial role in hydrodynamics. When $Re \ll 1$, the Navier–Stokes Eq. (8) reduces to the simple result that is linear in the velocity, the equation on creeping flow,

$$\frac{\partial \mathbf{V}}{\partial t} = -\mu(\nabla^2 \mathbf{V}). \tag{10}$$

A direct result of the linearity of this equation is that for any given geometry and Re, it can be shown that solutions to the Navier–Stokes equation are identical; thus the solutions are appropriate to a wide range of conditions. For most situations, solutions to the Navier–Stokes equation are made in the low-Re limit. A feature of these low-Re solutions is that the flow lines (contours of constant velocity) are parallel, i.e., laminar flow. Such solutions, of course, become poorer as Re increases, and in general this would exclude values of $Re \geq 1$. However, the transition to turbulent flow is geometry specific, and the laminar flow regime can extend to higher Re numbers, perhaps as high as 10, before the onset of turbulence. Up to that point, viscosity continues to play the dominant role in the drag. As discussed later, for accurate results in most experimental viscosity measurements, turbulence should be absent.

1.1.1.3 Stokes Equation. An important solution to the Navier–Stokes equation is that for motion of a sphere in a fluid. From this we obtain the Stokes equation. It is a low-Re solution, typically $Re = 2RV\rho/\eta < 0.1$, made for steady-state flow, e.g., the terminal velocity of a falling sphere. The full derivation of this equation is outside the scope of the present work (Landau and Lifshitz, 1959; Bird *et al.*, 1960). However, we can understand the origin of the equation by noting that one obtains two forces on the sphere in considering its motion governed by Eq. (10): the force acting normal to the surface due to a pressure field set up by flow, F_n, and a force acting tangential to the surface due to stress transmitted tangentially, F_t. For the case when the fluid velocity at the sphere boundary equals that of the sphere, a nonslip boundary condition, the two net forces, integrated over the surface of the sphere, are given by

$$F_n = \tfrac{4}{3}\pi R^3 \rho g + 2\pi\eta RV,$$

$$F_t = 4\pi\eta RV;$$

thus the *total force* is

$$F = F_n + F_t = \tfrac{4}{3}\pi R^3 \rho g + 6\pi\eta RV, \tag{11}$$

where R is the sphere radius and V is its steady-state velocity. As one can see, the second term is velocity dependent and is termed the Stokes drag. The familiar form of the Stokes equation giving the terminal velocity of a falling sphere is obtained by setting the total force equal to the gravitational force, $\tfrac{4}{3}\pi R^3 \rho_S g$, on a sphere of density ρ_S.

One important feature of the Stokes-equation solution is that, to leading order, the decay of the flow field goes as $1/r$. Hence the disturbance is long ranged, and the equation is strictly applicable only to isolated spheres. A practical consequence is that the use of this equation in determining the viscosity through falling- and rolling-sphere viscometers requires an empirical correction for termination of the velocity field by various walls of a container.

1.1.2 Diffusion In the previous sections both the motion of fluids (Sec. 1.1.1.2) and the motion of particles in a fluid (Sec. 1.1.1.3) under the influence of external forces (e.g., shear or gravitational forces) have been described. However, molecular migration can be observed even without the existence of external forces. This is the subject of Brownian motion, i.e., the statistical movement of atoms or molecules. Brownian motion has been extensively studied, beginning with observations of Robert Brown in 1853 of the random incessant motions of dust pollen in solution. In the following sections we describe how Brownian motion and diffusion are related.

1.1.2.1 Continuum Transport: Fick's Laws of Diffusion. The term *mutual diffusion* (or transport diffusion) describes the process of molecular transport within systems caused by the statistical movement of atoms or molecules. This transport process becomes macroscopically visible during the process of equalizing concentration differences (e.g., of dye molecules in a solution). The reduction of a concentration gradient implies a particle flux that can be described with Fick's first law,

$$\mathbf{j}_i = -D_i \nabla c_i. \tag{12}$$

The proportionality constant between the particle flux density $\mathbf{j}_i$ and the concentration gradient ∇c_i is the diffusion coefficient D_i of the considered component i of a mixture. Although the units for $\mathbf{j}_i$ will depend upon the units of concentration (Bird *et al.*, 1960), the units for the diffusion coefficient are m^2/s in SI or cm^2/s in cgs. In general, the diffusion coefficient D_i is a function of the concentration of the component i.

The time derivative of the concentration of a certain component i in a volume element has to be equal to the total flux of this component through the surface of the considered volume:

$$\frac{\partial c_i}{\partial t} = -\nabla \cdot \mathbf{j}_i. \tag{13}$$

Using this mass balance constraint along with Eq. (12), we obtain Fick's second law,

$$\frac{\partial c_i}{\partial t} = \nabla \cdot (D_i \nabla c_i), \tag{14}$$

which describes the time derivative of the concentration for a system as it approaches its equilibrium state.

1.1.2.2 Self-Diffusion. Even in the absence of a concentration gradient, statistical fluctuations of particle positions continue. This process, called *self-diffusion,* is caused by the thermal energy of the particles. Therefore, the self-diffusion coefficient D_S arises from the same fluctuations that drive mutual diffusion, but there is no net mass transport in this case.

Because there is no macroscopic particle flux, to monitor the molecular displacement in such systems a certain portion of the molecules must be labeled in one way or another, e.g., by introducing observable isotopes into the system under study. With respect to the experimental setup, the proportionality constant between the gradient of the concentration of labeled particles and their flux density is called the coefficient of *tracer diffusivity.*

The coefficients of tracer and self-diffusion coincide if two assumptions are fulfilled: First, the molecular mobility of the molecules must not be altered by the process of labeling. Second, mutual interactions between the tracer molecules must be small enough to be negligible; i.e., the mobility of any labeled molecules must be unaffected by the existence of other (e.g., labeled) molecules. Experimentally, that requirement is fulfilled in sufficiently high dilution of the labeled species under consideration (Kärger *et al.*, 1988).

1.1.2.3 Diffusion Equation. In homogenous systems, Brownian motion is determined by the coefficient of self-diffusion only. Since in such systems the self-diffusivity is independent of the spatial position of the molecules, Eq. (14) simplifies to

$$\frac{\partial c^*(\mathbf{r},t)}{\partial t} = D_S \nabla^2 c^*. \tag{15}$$

In this equation c^* is the concentration of labeled molecules (see Sec. 1.1.2.2) and, therefore, the probability of finding a labeled particle on a certain position. Equation (15) is commonly referred to as the *diffusion equation* since that differential equation provides the probability of finding a molecule at time t at position $\mathbf{r}$ if it initially was at $\mathbf{r}_0$. With the initial condition that the concentration of the labeled molecules follows Dirac's delta function, $c^*(\mathbf{r},0) = \delta(\mathbf{r} - \mathbf{r}_0)$, the solution of the diffusion equation, Eq. (15), is found to be a Gaussian distribution of the following form:

$$\begin{aligned} c^*(\mathbf{r},\mathbf{r}_0,\mathbf{t})d\mathbf{r} &= \frac{1}{(4\pi Dt)^{3/2}} \exp\left\{ -\frac{(\mathbf{r} - \mathbf{r}_0)^2}{4Dt} \right\} \\ &= P(\mathbf{r},t). \end{aligned} \tag{16}$$

Equation (16) is commonly referred to as the *propagator P*(**r**,*t*), or *fundamental solution of the diffusion equation.* It allows one to calculate the mean square value of molecular displacement during the observation time t:

$$\langle \mathbf{r}^2(t) \rangle = \int_{-\infty}^{+\infty} P(\mathbf{r},t)\mathbf{r}^2 d\mathbf{r} = 6Dt. \tag{17}$$

Equation (17) is referred to as Einstein's equation of diffusion and provides a straightforward manner for defining molecular diffusivity.

More generalized, the exact form of Eq. (17) depends on the number of spatial dimensions k of molecular migration:

$$\langle x^2(t) + y^2(t) + z^2(t)\rangle = 2kDt. \tag{18}$$

Equation (18) coincides with Eq. (17) in isotropic systems where $\langle \mathbf{r}^2(t)\rangle = \langle x^2(t) + y^2(t) + z^2(t)\rangle$ and the dimension of diffusion, k, is equal to 3.

1.1.2.4 Diffusion as a Random Walk. On a microscopic level, the propagation of atoms or molecules can be understood as a stochastic jump process. The molecular migration becomes "random" if both the length and the direction of the elementary jumps do not correlate; i.e., this diffusion process has to be a Markovian process. In the simplest approach for deriving the diffusion equation, however, one considers only jump processes that are uncorrelated regarding their direction.

Imagine the one-dimensional diffusion of a particle that is situated at a certain position of an imaginary z axis and performs stochastic jumps of a *constant* length l toward the positive or negative direction of the z axis. The mean square displacement of the considered particle, $\langle z^2\rangle$, can be expressed as a function of the number of jumps, n, and their length l:

$$\begin{aligned}\langle z^2(n)\rangle &= \left\langle\left(\sum_{i=1}^{n} l_i\right)^2\right\rangle \\ &= \left\langle\left(\sum_i l_i^2\right)\right\rangle + \left\langle\left(\sum_{i\neq j} l_i l_j\right)\right\rangle,\end{aligned} \tag{19}$$

where l_i is the length of the ith jump in positive or negative z direction. Assuming a Markovian diffusion process where the mean displacements of all particles cancel each other, $\langle l_i\rangle = \langle l_j\rangle = 0$, and a constant jump length, $l_i \equiv l$, Eq. (19) simplifies to

$$\langle z^2(n)\rangle = nl^2 + \sum_{i\neq j} \langle l_i\rangle\langle l_j\rangle = nl^2. \tag{20}$$

The number of elementary jump processes, n, is the ratio of the total diffusion time t and the mean dwell time of the jumps, τ. Thus, it follows from Eq. (20) that

$$\langle z^2(t)\rangle = l^2 t/\tau, \tag{21}$$

which shows that the mean square displacement is directly proportional to the diffusion time if the displacements in subsequent intervals are uncorrelated.

The jump length l can be expressed using Einstein's equation, Eq. (18), for one-dimensional diffusion ($k = 1$) of the jumping particle:

$$D = l^2/2\tau. \tag{22}$$

Therefore, for Eq. (21) it follows that

$$\langle z^2(t)\rangle = 2Dt. \tag{23}$$

Equation (23) represents a definition of the diffusion coefficient equivalent to Eq. (17) or (18). Similarly to Eq. (18), Eq. (23) can be extended to three-dimensional diffusion.

For a given time interval, the mean square displacement is thus found to be directly proportional to the total observation (or diffusion) time. Molecular propagation that follows this relation is generally called *normal diffusion.* Since any confinements interfering with the diffusion process lead to a correlation of subsequent elementary jumps, the direct proportionality between mean square displacement and diffusion time breaks down; this is called *anomalous diffusion.* Depending on the nature of the confinement, quite distinct patterns of deviation from normal diffusion may be observed.

1.1.3 Stokes–Einstein Equation: Viscosity and Diffusion in Liquids The single most important relation between viscosity and diffusion in liquids is the Stokes–Einstein equation. Although its origin is from hydrodynamics, hence a continuum picture, it has been found to be an accurate equation down to the molecular length scale, and it finds wide application in relating molecular diffusion to fluid viscosity. We describe some of these throughout the text. Here we give the derivation.

Brownian motion of a single spherical body moving in an overall quiescent fluid is the target of our thinking. In the Stokes equation derivation (Sec. 1.1.1.3), we saw that such a body would experience a damping force equal to the viscous drag, $F/V = 6\pi\eta R$. The driving force for Brownian motion is thermal fluctuations. The equation of motion for a sphere of mass m can be written

$$m \frac{dV}{dt} = -F + F_R(t) = -6\pi\eta R V + F_R(t), \tag{24}$$

where $F_R(t) \propto k_B T$ is the random white-noise term that represents the incessant bombardment of the solute molecules that drive Brownian motion. The solution to Eq. (24) for the velocity correlation function is

$$\langle V(0)V(t)\rangle = (k_B T/m) \exp(-Ft/m). \tag{25}$$

The amplitude of the velocity fluctuations is proportional to the thermal energy, $k_B T$, and the decay is given by the frictional drag, m/F. For a typical 1-μm particle in water, $m/F \approx 0.1$ μs. An important consequence of Eq. (25) is that the velocity is correlated only for a short period; i.e., over long times the particle will appear to be moving randomly. This random walk is diffusion, with the magnitude of the diffusion coefficient being given by

$$D \equiv \int_0^\infty \langle V(0)V(t)\rangle dt = \frac{k_B T}{6\pi\eta R}. \tag{26}$$

This is the Stokes–Einstein equation. Despite its hydrodynamic origins, it is extensively utilized for both molecular and complex liquids in examining the relationship between single-particle (tracer) diffusion and solution viscosity. A description of particle positions in the random-walk, Brownian, regime can be found from Eqs. (25) and (26) and is given by

$$\langle \Delta r^2(t)\rangle = 6Dt, \tag{27}$$

providing a similar result for the time dependence of the mean square displacement to that derived in Secs. 1.1.2.3 and 1.1.2.4.

Experimental tests of the Stokes–Einstein equation have revealed the need to consider both the nonslip (as formulated) and slip boundary conditions for the fluid–particle interface. The slip boundary condition alters Eq. (26) by replacing the denominator with $4\pi\eta R$. We discuss this in Secs. 1.2.4 and 4.1.3.2.

Rotational motions, in addition to translational ones, also result from the incessant thermal motions imparted to individual molecules by the surrounding liquid. The randomness of the collisions results in the applicability of a diffusion equation to the description of the time-dependent rotational velocity, $\omega = d\Theta/dt$, given as

$$I \frac{d\omega}{dt} = -F_\omega \omega + F_\omega(t), \tag{28}$$

where I is the moment of inertia, F_ω the viscous drag force of a rotating sphere, and $F_\omega(t)$ the random white-noise term representing the incessant bombardment of the solute molecules that drive rotational diffusion. A solution to this equation is

$$D_\Theta \equiv \int_0^\infty \langle \omega(0)\omega(t)\rangle dt = \frac{k_B T}{8\pi\eta R^3}, \tag{29}$$

yielding an expression for the time-dependent angular displacements as

$$\langle \Delta\Theta^2(t)\rangle = 6D_\Theta t, \tag{30}$$

in exact analogy with the translational diffusion description above.

The specific study of rotational diffusion, as opposed to translational diffusion as referred to earlier, has been the subject of numerous works (Berne and Pecora, 1976). Both Alms *et al.* (1973) and Bauer *et al.* (1974) used dynamic light scattering (depolarized) to elucidate the nonsymmetric rotational relaxation mechanisms of benzene molecules in solution. Hu and Zwanzig (1974) determined the effective translation–rotation coupling constants by computer simulation for a wide range of elliptically shaped objects. Sillescu *et al.* studied both rotational and translational diffusion in highly viscous, glass-forming systems by forced Rayleigh scattering (FRS) and NMR-based diffusion techniques (Chang̈ *et al.*, 1994).

1.2 Atomistic Theories and Computer Simulations

1.2.1 Simple Kinetic-Theory Approximation As might be imagined, atomistic descriptions of the dynamics that lead to the energy dissipation mechanisms responsible for viscosity and diffusion are quite complex. Not only must the models describe the energy losses, but they must also be consistent with requirements for local equilibrium such as mass and energy balance. In the general case, such models derive from solutions to the Liouville equation (see Hansen and McDonald, 1986), which provides a full description of an N-particle system in the classical mechanics limit. However, a true solution to this equation is as yet unrealized; one instead finds solutions based upon consideration of subsets of the N-particle system. For example, the pair distribution function arises from this approach. The difficulty in solving this equation for viscosity and diffusion increases as one goes to higher and higher density. At low density, description of the dynamics through binary collisions is sufficient, but as the density increases, three-body interactions become increasingly important. It is in this regime that the current challenge in understanding the atomistic origins of viscosity and diffusion lie. Considerable theoretical and experimental efforts are now devoted to this. One theory, mode-coupling theory (Götze and Sjögren, 1992; Cummins *et al.*, 1993), which utilizes the memory function to describe the results of three-body interactions, is currently quite promising.

In the description that follows, we start with a simple description of a theory for viscosity applicable to low particle density and work our way forward considering increasingly complex treatments that will eventually include dense gases. This approach will give insight into the physical processes that lead to viscosity and diffusion without having to treat the involved mathematics necessary for more rigorous treatments. The reader is referred to several books (Hirschfelder *et al.*, 1954; Bird *et al.*, 1960; Hansen and McDonald, 1986) for more complete discussions.

Given a gas composed of hard spheres of diameter σ and mass m, at a concentration of n molecules per unit volume such that the average distance between molecules is considerably larger than σ, we obtain from kinetic gas theory the following expressions for the equilibrium properties. The average velocity is

$$\bar{u} = (8k_{\mathrm{B}}T/\pi m)^{1/2}. \tag{31}$$

The average distance traveled by a molecule between collisions is the mean free path λ:

$$\lambda = (\sqrt{2}\pi\sigma^2 n)^{-1}. \tag{32}$$

We assume that locally equilibrium is maintained during transport, making these equations an accurate description of local dynamics. Also, we consider only the local velocity u relative to the overall velocity $\mathbf{u}$. Of the ensemble of molecules, $\frac{1}{3}$ can be considered as traveling along each of three orthogonal directions, and half of each of these fractions are going in the positive direction, half in the negative one.

Now, consider the geometry depicted in Fig. 2, where we define planes N and P, separated from the reference plane O by one mean free path. Our interest is in the amount of some property, W (mass or momentum), that crosses plane O per unit area and unit time. Only $\frac{1}{6}$ of the molecules on plane N move in the $+z$ direction. The upward flux is $\frac{1}{6}\bar{u}W_N$; similarly, the downward flux is $\frac{1}{6}\bar{u}W_P$. The net flux Ψ_W of W is therefore

$$\begin{aligned}\Psi_W &= \tfrac{1}{6}\bar{u}(W_N - W_P) = -\tfrac{1}{3}\bar{u}\lambda\left(\frac{dW}{dz}\right)\\ &= -\xi\frac{\sqrt{k_{\mathrm{B}}Tm\pi}}{\pi m\sigma^2 n}\left(\frac{dW}{dz}\right),\end{aligned} \tag{33}$$

where $\xi = 2/3\pi$ and the subscripts N and P designate the values of W on the respective planes. This leads directly to equations for both the diffusion and viscosity of particles having mass m and concentration n:

$$\begin{aligned}\Psi_{(nm)} &= j_z = -\tfrac{1}{3}\bar{u}\lambda\left(\frac{d(nm)}{dz}\right) = -D\left(\frac{d(nm)}{dz}\right);\\ \Psi_{(nmv_y)} &= p_{yz} = -\tfrac{1}{3}\bar{u}\lambda nm\left(\frac{dv_y}{dz}\right) = -\eta\left(\frac{dv_y}{dz}\right).\end{aligned} \tag{34}$$

The expressions in the simple kinetic-theory (SKT) approximation for self-diffusion and viscosity are

$$\begin{aligned}D_{\mathrm{SKT}} &= \tfrac{1}{3}\bar{u}\lambda = \xi\frac{\sqrt{\pi mk_{\mathrm{B}}T}}{\pi\sigma^2}\frac{1}{\rho},\\ \eta_{\mathrm{SKT}} &= -\tfrac{1}{3}\bar{u}\lambda nm = \xi\frac{\sqrt{\pi mk_{\mathrm{B}}T}}{\pi\sigma^2} = \rho D.\end{aligned} \tag{35}$$

where $\rho = nm = pm/k_{\mathrm{B}}T$ is the density of the ideal gas and $\xi = 2/3\pi$ as in Eq. (33). This exceedingly simple model gives the correct functional form for both these transport properties.

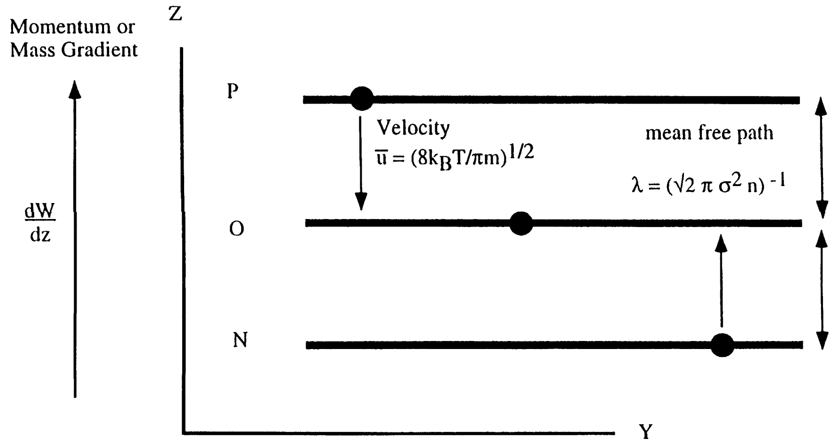
Momentum or
Mass Gradient
$\frac{dW}{dz}$
Z
P
O
N
Y
Velocity
$\bar{u} = (8k_BT/\pi m)^{1/2}$
mean free path
$\lambda = (\sqrt{2}\,\pi\,\sigma^2\,n)^{-1}$

FIG. 2. Illustration of the simple kinetic-theory approach to calculating the transport rate of the property W. For W equal to density, this yields the diffusion rate; for W equal to momentum, this yields the viscosity. A superimposed gradient in W along the vertical z axis in a dilute gas is transported by the movement of individual atoms or molecules. A molecule arriving at plane O from plane N brings with it the value of W from one mean free path away. A similar process transports the W values from plane P to O. On average, one-third of the molecules travel in each of three orthogonal directions, x, y, and z: one-sixth move along $+z$ and one-sixth along $-z$. Each molecule moves with an average thermal velocity $\bar{u}$. The transport equation is obtained by summing the $+z$ and $-z$ fluxes, as described in the text, where increasingly more complex approximations are also explained.

1.2.2 Boltzmann Development A more rigorous consideration of the nonequilibrium dynamics for hard spheres changes the values of ξ slightly. The expression in the Boltzmann (B) approximation are

$$D_{\mathrm{B}} = \tfrac{3}{8}\frac{\sqrt{\pi m k_{\mathrm{B}} T}}{\pi\sigma^2}\frac{1}{\rho},$$
$$\eta_{\mathrm{B}} = \tfrac{5}{16}\frac{\sqrt{\pi m k_{\mathrm{B}} T}}{\pi\sigma^2} = \tfrac{5}{6}\rho D, \qquad (36)$$

with $\xi_D = \frac{3}{8}$ and $\xi_\eta = \frac{5}{16}$. These equations are moderately accurate expressions for gaseous materials. As can be seen, they correctly predict that the viscosity is pressure independent and that the viscosity increases with increasing temperature. The square-root temperature dependence is not quite accurate, and consideration of more realistic intermolecular potentials helps correct this.

1.2.3 Lennard-Jones Interactions Extending the range of interaction for the binary collisions considered in the previous model now introduces an integral collision term into the equations. Typically used are potentials that are spherically symmetric, and the Lennard-Jones (6-12) potential has proven a useful way to model the interactions. The results for the derivation of the rigorous kinetic theory (see Hirschfelder *et al.*, 1954, p. 527) for the self-diffusion and the viscosity in the Lennard-Jones (LJ) interaction approximation are as follows:

$$D_{\mathrm{LJ}} = \tfrac{3}{8}\left(\frac{\sqrt{\pi m k_{\mathrm{B}} T}}{\pi\sigma^2\Omega^{(1,1)*}}\right)\left(\frac{1}{\rho}\right),$$
$$\eta_{\mathrm{LJ}} = \tfrac{5}{16}\left(\frac{\sqrt{\pi m k_{\mathrm{B}} T}}{\pi\sigma^2\Omega^{(2,2)*}}\right), \qquad (37)$$

where the terms $\Omega^{(n,n)*}$ are normalized collision integrals. These depend upon the interaction potential and are normalized to the values for a hard-sphere (HS) interaction. Hence, for hard spheres they equal unity. To utilize actual interatomic potentials, one must calculate the values for $\Omega^{(n,n)*}$. These are available in tabulated form in Hirschfelder *et al.*, (1954), Appendix *I*, for a Lennard-Jones potential. Although we give here the equations for single-component fluids, this same approach can be extended to mixtures and more complex interaction potentials. This is discussed in Hirschfelder *et al.* (1954), pp. 485–610.

1.2.4 Hard-Sphere Model—Enskog Development An extended version of Boltzmann's theory in dense gases and liquids was developed by Enskog and Chapman (Chapman and Cowling, 1939) more than 80 years ago. The Enskog theory still accounts only for binary collisions and assumes many features in common with dilute gases. Molecular chaos, as found in dilute gases, is assumed, which amounts to neglecting the correlations between successive particle collisions. Spheres collide with one another at random with velocities consistent with the Maxwell–Boltzmann distribution at that temperature. In particular, multiparticle correlations such as those seen in vortex or ringlike flows in computer simulations are neglected. Two fundamental assumptions in the Enskog theory that distinguish it from the dilute-gas theories can be stated as follows:

1. The molecules attain a finite size, which means that molecules close to the reference plane O in Fig. 2 will collide with those in the plane. This collisional transfer of momentum and energy must be accounted for.
2. Because of the higher density, the collision frequency differs from that of point particles. Both kinetic and potential energy contributions to the flux vector (mass or momentum) are included. The potential energy term arises from the collisional transfer, which is absent for low-density gases.

The Enskog theory (E) scales the Boltzmann results up to higher densities as

$$D_E = D_B/g(\sigma);$$
$$\eta_E = \eta_B[1/g(\sigma) + 3.2\phi + 12.18g(\sigma)\phi^2], \quad (38)$$

where D_B and η_B are given by Eq. (36) and ϕ is the volume fraction of particles. The quantity $g(\sigma)$ is the magnitude of the two-particle distribution function evaluated at contact and is a measure of particle collisions that are neglected in the Boltzmann approach.

The contact radial distribution function $g(\sigma)$ is related to the second virial coefficient in the expansion for the compressibility factor $Z(\phi)$ for the HS system,

$$Z = p/nk_BT = 1 + 4\phi g(\sigma). \quad (39)$$

Hoover and Ree (1964) as well as Carnahan and Starling (1969) calculated $Z(\phi)$, resulting in the following virial expansion equation:

$$Z = 1 + 4\phi + 10\phi^2 + 18\phi^3 + 28\phi^4 + \ldots; \quad (40)$$

$$Z = \frac{1 + \phi + \phi^2 - \phi^3}{1 - \phi^3}. \quad (41)$$

Substituting Eq. (41) into Eq. (39), we obtain an expression for the quantity of interest in the Enskog theory, the pair distribution function,

$$g(\sigma) = (1 - 0.5\phi)/(1 - \phi)^3. \quad (42)$$

In the dilute-gas limit, $\phi \to 0$; $g(\sigma) \to 1$, and the Boltzmann results are recovered. At higher densities, $\phi = 0.1, 0.2, 0.3$, and 0.4, we find $g(\sigma) = 1.3, 1.9, 2.3$, and 2.9, respectively.

An interesting result of the Enskog theory is that over a wide range of moderate densities the products of the diffusion coefficient and viscosity are nearly constant. This can be recast in terms of the Stokes–Einstein model, discussed in Sec. 1.1.3. We rewrite Eq. (26) as $D = k_BT/C\pi\eta R$, where C is 6 for nonslip and 4 for slip boundary conditions. Application of the Enskog model yields C = 3.55, 3.16, 2.88, 3.41, 3.50, and 3.03 at the volume fractions ϕ = 0.15, 0.25, 0.37, 0.41, 0.46, and 0.494, respectively. Hence this test shows that the Stokes–Einstein theory is not strictly obeyed by the model calculations but that solutions are closer to the *slip* boundary condition.

In practice, the Enskog theory is limited in its range of applicability to that of dense gases, not liquids. In the next section we consider corrections to the Enskog model into the fluid regime.

1.2.5 Hard-Sphere Model—Computer Simulations at High Density A critical examination of the Enskog transport theory of hard spheres was performed by Alder, Gass, and Wainwright (Alder *et al.*, 1970; Alder and Wainwright, 1970), who performed computer simulations over a wide range of densities. Several important effects were observed. First, the Enskog assumption of uncorrelated binary collision, molecular chaos, was violated at moderate densities. The velocity autocorrelation function, which should decay exponentially in time according to Enskog, was found to possess a long-time, algebraic persistence. This long-time tail results from multiparticle vortex-type flows. Second, vortex flows lead to an enhancement in diffusion, by about 30% at moderate densities. Third, at the highest densities, backscattering effects cause a marked reduction in D, as well as increase in η, upon approach to the freezing transition (Hoover and Ree, 1964) at $\phi = 0.494$. Finally, tests of the Stokes–Einstein equation can be made using values from these simulations. For $\phi = 0.15$, 0.25, 0.37, 0.41, 0.46, and 0.494, the values of C, as defined in Section 1.2.4, are $C = 4.1$, 4.3, 4.1, 4.3, 4.2, and 3.8, respectively, nearly independent of ϕ and consistent with the *slip* boundary condition.

The seminal work of Alder, Gass, and Wainwright led to considerable activity in the 1970s researching the transport behavior of HS fluids. In 1974 Dymond (1974a, 1974b) fitted the simulation results of Alder *et al.* (Alder and Wainwright, 1970) to an expression linear in the particle volume. In 1981, Woodcock and Angel (1981a, 1981b) performed HS simulations and extracted diffusivities at concentrations beyond those originally studied by Alder *et al.* (Alder *et al.*, 1970; Alder and Wainwright, 1970). One of their main findings was that the linear volume dependence of D seen at moderate densities crosses over to a free-volume (see Sec. 1.2.8) type dependence at the highest densities:

$$D = A \exp[-B/(V - 1.163V_0)]. \qquad (43)$$

It is interesting to note that the packing fraction at which the transport is predicted to cease, $\phi = 0.61$, is close to the value of random close packing (RCP) $\phi_{\mathrm{RCP}} = 0.638$, at which all the particles are touching and all particle motions necessarily stop. Woodcock and Angel also showed that experimental data on methylcyclohexane as well as model Leonard–Jones molecules are in excellent agreement with the predictions.

Accurate expressions for the transport properties of HS fluids provide a useful reference system against which measurements on real atomic and molecular liquids can be tested and modeled. However, it must be kept in mind that the theory only directly applies to *spherical* molecules that interact as hard spheres. The effects of asymmetric shapes, as seen in the next section, is to couple translational and rotational motions, resulting in reduced diffusion coefficients as compared to the HS model. Additionally, molecules are not explicitly hard spheres, typically displaying a steep short-range potential with an effective radius that can decrease slightly with increasing collision velocities, i.e., temperature. Several tests of the HS equations have been done on systems where it should best apply, noble gases such as argon and krypton (Dymond, 1974a). In these cases a linear dependence of the diffusion on molar volume was seen, with a close numerical accord of the results with the HS model. In the slightly more complex cases of nearly symmetric simple molecules such as methane and cyclohexane, the important linear dependence of D and $1/\eta$ on volume is seen, but the numerical values are not in close agreement. This situation is commonly found in many atomic and molecular liquids (Parkhurst and Jonas, 1975; Dymond and Brawn, 1979). Improvements in the HS model, to include the effects of the translation–rotation coupling, are shown in the next section to allow the extension to a wide variety of liquids.

1.2.6 Rough Hard-Sphere Model In 1975 Chandler proposed a simple extension to the hard-sphere model that includes the coupling of molecular translational and rotational degrees of freedom (Chandler, 1975). His calculations showed that the coupling provides an additional mechanism of relaxation of the velocity correlation function of a particle in a fluid. As a result, velocity correlations decay to zero more rapidly than if the coupling did not exist, leading to smaller values of D. The model is termed the *rough hard-sphere model,* described by D_{RHS} and η_{RHS}, and in this context the hard-sphere results are referred to as the *smooth hard-sphere model,* i.e., D_{SHS} and η_{SHS}.

The basic equation for the diffusion coefficients in this model is

$$D_{\mathrm{RHS}} = AD_{\mathrm{SHS}}; \qquad (44)$$

and similarly for the viscosity,

$$\eta_{\mathrm{RHS}} = B\eta_{\mathrm{SHS}}. \qquad (45)$$

The RHS model predicts that the effects of nonsphericity is such to produce $0 < A < 1$, and $B > 1$. In Chandler's derivation, the RHS model was developed as a rigorous result using the assumption of uncorrelated binary collisions, equivalent to the Enskog approximation, which has been shown above to be inaccurate at high densities. However, he suggested that the translation–rotation coupling parameters of the individual molecules, A and B, may be somewhat insensitive to the emergence of higher-order collisional correlations and suggested that the forms of Eqs. (44) and (45) be maintained with the use of the high-density computer simulation data results for D_{SHS}.

The RHS model is extensively used in the modeling of experimental diffusion and viscosity data onto the "hard-sphere" model. For instance, Chandler showed that even in a fairly spherical molecule such as CCl_4, the diffusion and viscosity data could be mapped onto the HS simulation results [Eqs. (44) and (45)] using $A = 0.54$ and $B = 1.74$. Parkhurst and Jonas (Parkhurst and Jonas, 1975) studied tetramethylisilane and benzene and found $B = 1.39$ and 1.34, respectively. Jonas *et al.* (1979) studied in detail the rather nonspherical molecule methylcyclohexane and contrasted its behavior against the more spherical cyclohexane and found $A \sim 0.3$ and $A \sim 0.75$, respectively. In addition they measured the temperature dependence of A; and while for cyclohexane the dependence was weak, for methylcyclohexane A decreased by a factor of 2 from $T = 300$ °C down to $T = 200$ °C. The marked increase of the translation–rotation coupling at low temperatures, or high densities, was attributed to the strong local caging of the nonspherical molecules.

1.2.7 Mode-Coupling Theory The discovery by Alder and Wainwright of long-time velocity correlations in dense fluids, i.e., the breakdown of the Enskog theory, has motivated numerous new theoretical approaches to describe the many-body relaxations that occur in normal liquids. Mode-coupling theory is the leading approach to deal with these phenomena in a rigorous mathematical framework. This theory finds widest application in describing behavior associated with the transition to the glassy state, but it is applicable over a wide temperature range.

We showed previously (Secs. 1.1.2.3 and 1.1.2.4) that thermally driven uncorrelated motion leads to self-diffusion. Fourier transform of Brownian diffusion, Eq. (16), leads to density correlation functions that in this case decay to zero in time via an exponential function, characterized by a single relaxation time. In dense fluids this simple picture changes in two important ways. First, the decay to zero becomes nonexponential, and second, there appear two distinct characteristic relaxation times. The initial fast rate of change of the density–density correlation is termed beta relaxation, while the long slow decay is called the alpha relaxation (see Cummins *et al.*, 1994, Fig. 1). Microscopically, the beta relaxation may be thought of as particle movement within a cage created by the surrounding molecules, whereas the alpha relaxation describes diffusional motion out of the cage. Mode-coupling theory provides a rigorous mathematical model to treat the entire relaxation spectrum. The intractable problem of treating multibody interactions directly is circumvented by recasting the problem as one of interactions between different density fluctuation modes.

In mode-coupling theory the density fluctuations are treated through a damped-oscillator equation, where a memory function describes the time correlations and controls the damping of the density fluctuations. Idealized mode-coupling theory (I-MCT) expresses this memory function in terms of linear and quadratic coupling coefficients among density fluctuation modes. For simple interactions, such as Lennard-Jones, numerical solutions for the coupling coefficients are obtained. These are, of course, temperature and density dependent, and this dependence leads to the changes in the relaxation spectra observed experimentally. For more complex interactions, asymptotic solutions lead to power laws whose exponents are obtained through fits to experimental data (Cummins *et al.*, 1993). I-MCT fails at low temperature because it predicts that at some temperature $T_C > T_g$ the alpha relaxation time will go to infinity and cause an arrest of structural relaxation. Extended mode-coupling theory (E-MCT) includes thermally activated hopping processes, which prevent this singularity, and adds another term to the memory function. For E-MCT, the alpha relaxation time continuously increases as temperature decreases.

Both viscosity and diffusion are associated with the alpha relaxation. As explained above, simple random-walk motions do not describe the behavior of very dense fluids but a nonexponential relaxation is found instead. It is common to fit this relaxation spectrum by a stretched exponential form, $\exp[(-t/\tau)^{\beta}]$, where β lies in the range $0 < \beta < 1$. These values can be associated with the fragility discussed in Sec. 1.3 (Böhmer *et al.*, 1993). Of greater interest for the present discussion is the relaxation time, τ_{α}. We can associate this time with the viscosity through use of the Maxwell relation, $\eta = \tau_{\alpha}/G_{\infty}$, where G_{∞} is the high-frequency shear modulus of the liquid. The assumptions underlying this simplified connection between viscosity and relaxation are discussed in Ferry (1980) and Cummins *et al.* (1997). Taking this approach and using the predictions of I-MCT, $\tau_{\alpha} \propto |T - T_C|^{-\gamma}$, we see that viscosity is predicted to vary as a power of the temperature difference with respect to a critical temperature. It is found that this form is an adequate representation of data at high temperatures (or low pressures; Li *et al.*, 1995), where values of γ may be obtained from fitting the viscosity data or from analysis of relaxation spectra (Cummins *et al.*, 1994). However, this covers only a few decades of viscosity change, and below T_C this fails. The T_C values have no apparent physical meaning, and this was one of the main objections to early formulations of MCT. The revised E-MCT prediction for viscosity variation with temperature has not been fully formulated, but an approximate form, including hopping transport, was recently shown (Cummins *et al.*, 1997) to fit accurately the experimental alpha relaxation spectrum over the entire experimental range. Developments in mode-coupling theory and its application to experiment are currently an active area of research, and other modifications will surely be discovered.

1.2.8 Free-Volume Theory In the absence of a rigorous theory for the intermolecular dynamics to treat real liquids, one can utilize empirical relations based upon various approximations to infer the temperature, pressure, and compositional dependence of the viscosity and diffusion. Several of these types of equations have been derived; among them is the Adam–Gibbs theory, based upon entropy theory (Adam and Gibbs, 1965), and the free-volume theory, based upon molecular diffusion via cooperative movement of intermolecular free space (Cohen and Turnbull, 1959). This latter theory is rather successful in describing the effects of pressure on the viscosity of many liquids. Its assumptions can be summarized as follows:

1. It is possible to associate a local volume of molecular scale with each molecule.
2. Molecular transport occurs only when voids having a volume greater than some critical value V_{∞} approximately equal to the molecular volume form by the redistribution of free volume.
3. No local free energy is required for free-volume redistribution.

In a dense liquid, each molecule is encaged by the surrounding molecules. Now it is asked what volume is required to allow a molecule to move a distance on the order of a molecular diameter. At a minimum this volume must equal the molecular volume. Under the assumption that movement to this new position costs no energy and that the distribution of free space surrounding each molecule is governed by entropy requirements, an equation for the diffusion as a function of temperature and relative density can be obtained. If it is further assumed that one can apply the Stokes–Einstein relationship, then this equation can also be used to describe the viscosity. The forms of the equations are as follows:

$$D \propto \exp[-\gamma V_{\infty}/(V - V_{\infty})];$$

$$\eta \propto \exp[\gamma V_{\infty}/(V - V_{\infty})]. \tag{46}$$

Equations (46) can accurately describe data over a limited temperature range (Cohen and Grest, 1979). Recently, these equations have been shown to have greater utility in treating isothermal, variable-pressure data. In this case a normalized form is used:

$$\ln(\eta) = \ln(\eta_{p=0}) - \frac{BV_{\infty}}{V_{p=0} - V_{\infty}} + \frac{BV_{\infty}}{V - V_{\infty}}, \tag{47}$$

where $\eta_{p=0}$ and $V_{p=0}$ are the ambient-pressure viscosity and volume; V_{∞} and B are fit parameters, V_∞ being the volume at infinite viscosity and B a parameter that describes the strength of the viscosity divergence as $V \rightarrow V_\infty$. This equation has been found to describe the viscosity rise correctly over a fifteen–order-of-magnitude range, from ordinary liquid to the glassy state (Cook *et al.*, 1993). Moreover, the molecular volume V_∞ obtained from fitting the equation to the viscosity data is in good agreement with other estimates of the molecular volume. This volume differs considerably from that obtained when the theory is applied to variable-temperature data (Herbst *et al.*, 1993). The variable B attains values that range from approximately 1 to 10 depending upon the fluid. In the original theory, written for hard-sphere molecules, it was expected that $\gamma \propto B \approx 1$. Indeed, for simple molecules such as isopropyl benzene (Li *et al.*, 1995) and *o*-terphenyl, values close to unity are found. For more complex molecular shapes and interactions, such as hydrogen-bonded molecules, larger values of B are found. Recently it was shown that this variation could be related to the "fragility" of the molecular liquid (Cook *et al.*, 1994), a concept originated to describe the degree of cooperativeness of motion within the liquid. Further theoretical development is needed to introduce this formally into the free-volume theory; however, even in its present form it can provide a quantitative means of describing the origin of pressure-induced viscosity changes. Its use to describe temperature effects is more problematic. The extended free-volume theory (Cohen and Grest, 1979) can provide an accurate description of the temperature-induced viscosity rise over a wide range, but this formulation is clearly inferior to the original theory when used for variable-pressure data. Reconciliation of this problem is an area for further theoretical development.

Nearly all liquids for which data are available exhibit characteristics of free-volume scaling in their pressure behavior. For example, one unusual feature in the pressure data is an inflection point in a semilogarithmic plot of the viscosity as function of pressure. Utilizing free-volume scaling, it is found (Herbst *et al.*, 1993) that this feature is a consequence of nonlinear compressibility coupled with density-dependent viscosity. The reader is referred to the original articles for a more complete discussion.

1.3 Scaling Theories for Simple Liquids

There is often a practical need to describe the viscosity changes that are caused by changes in the temperature, composition, or pressure of the liquid. As described in the previous section, current theories are inadequate to predict viscosity from first principles, and scaling theories are widely used. The reader is referred to an excellent review of these empirical equations (Monnery *et al.*, 1995). A second set of scaling relationships has arisen from a desire to ascribe a molecular or thermodynamic meaning to the viscosity. Among the widely used forms of these equations, the Adam–Gibbs formulation (Adam and Gibbs, 1965), the WLF (Williams–Landel–Ferry) equation (Ferry, 1980), and the VTF (Vogel–Tammann–Fulcher) equation (Böhmer *et al.*, 1993) are found. The WLF equation finds wide application to polymers (see Sec. 4.2.2.1). The VTF equation has received considerable attention for many fluids. This equation is as follows:

$$\ln(\eta) = A + DT_0/(T - T_0), \tag{48}$$

where A, D, and T_0 are fit parameters. The physics underlying these parameters has been widely discussed, and the reader is referred to a recent review (Böhmer *et al.*, 1993). In brief: T_0 is correlated with the *Kauzmann* temperature, a fictive temperature at which the entropy of the liquid becomes equal to that of the crystal. The parameter D is related to the fragility. This concept arose from the realization that the rate of viscosity change with temperature, the size of the heat capacity change at the glass transition, and the deviation from simple single-exponential molecular relaxations are all related (Angel, 1991). The implication is that intermolecular interactions lead to structures in the liquid that give rise to a cooperativeness of motion that is absent for hard-sphere-like fluids. Fluids range from strong to fragile as D varies from large to small. This is also mentioned briefly in Sec. 1.2.8. The parameter A is proportional to the viscosity in the high-temperature limit ($T \rightarrow \infty$). In many instances, the VTF equation can describe the viscosity change over many decades, and the resulting fits enable the definition of the fragility and the calculation of the glass transition temperature, defined in this case as the point at which $\eta = 10^{12}$ Pa s.

For some liquids the VTF equation does not fit a wide temperature range. In such instances, the low-temperature data, near the glass transition temperature, are typically used in obtaining D and T_0.

2. MEASUREMENT TECHNIQUES—VISCOSITY

As indicated in Eq. (1), the measurement of viscosity must at some level involve the imposition of a shear rate upon a fluid and a measurement of the resulting shear stress or vice versa. However, an accurate experimental realization of the conditions depicted in Fig. 1 requires rather specialized conditions of flow. As previously discussed, nonturbulent flow is required (low *Re*). In addition, there will be a minimum time, t, to set up the conditions of steady-state flow, and this leads to a requirement for a minimum strain, $\gamma = t\dot{\gamma}$. This, along with the influence of the plate edges, makes the plate-and-plate geometry depicted in Fig. 1 impractical for general use as a viscometer. In practice, viscometers having infinite strain are commonly used; examples are rotating geometries such as the Couette or cone-and-plate that attain homogeneous stress and strain fields. Nonhomogeneous fields are also useful, but their use requires that one solve the Navier–Stokes equation using certain assumptions, a typical one being Newtonian fluid behavior. The resulting equation will relate an observable, such as the amount of material that flows through a capillary with time, to the viscosity of the fluid. This explicit dependence on a flow model means that changes in the observable may not fully reflect the changes in fluid behavior. For example, change to a non-Newtonian fluid leads to an alteration of the entire velocity profile in the capillary and requires an altogether new set of equations to interpret the flow. A further complication in building any viscometer is that viscosity values can span many decades, from a few micropascal-seconds for a gas to a gigapascal-second for glassy materials. Consequently, a single viscometer capable of making measurements over the entire range is not found. Typically, each device is capable of covering a few decades of viscosity only. A change in dimensions of a given geometry or use of an entirely new device is then required to reach a new viscosity range.

In the following, we briefly review some commonly utilized geometries. We emphasize the determination of the shear viscosity of Newtonian fluids. For more detailed reviews of the techniques as well as the more generalized fields of rheology and finite-strain deformation, the reader is referred to Ferry (1980) and Macosko (1994). In most instances the compositional or temperature dependence of a fluid is measured, and these topics are covered in these two references. The reader is referred to a recent review (King, 1997) for a description of high-pressure viscometry.

2.1 Rotational Viscometers

In this section, two commonly used rotary viscometers, the Couette and cone-and-plate geometries, are described. Both devices produce homogeneous stress and strain fields. The first of these consists of two concentric cylinders with a small gap separating them. Rotation of one cylinder induces a torque in the other through the fluid filling the gap. If the ratio of the cylinder radii is near unity, the resulting velocity gradient is almost that of parallel plates. Hence, the torque on the stationary cylinder is proportional to the absolute shear viscosity, and the shear rate is given by $\dot{\gamma} = \Omega\langle R\rangle/\Delta R$, where Ω is the angular rotation speed, $\langle R\rangle$ is the average radius, and ΔR is the gap separation. The Couette geometry is especially good for study of shear rate effects on moderately viscous fluids and suspensions.

The cone-and-plate geometry consists of a stationary plate with a cone-shaped bobbin oriented perpendicular to the surface. The tip of the cone nearly touches the plate, and its cone angle, β, is small, about 6°. As the cone rotates, it induces a torque in the plate through the wedge-shaped annulus of fluid that is held in place through capillary action. The shallow cone angle gives a constant velocity gradient throughout the fluid, and the observed torque is proportional to the viscosity at a shear rate given by $\dot{\gamma} = \Omega/\beta$. The cone-and-plate geometry is advantageous for studying moderately high-viscosity fluids, particularly fluids that are viscoelastic.

Various experimental difficulties arise with the use of both geometries. For example, shear heating, cylinder-end effects for Couette flow, and fluid-edge effects for cone-and-plate geometry can complicate the experimental analysis, as discussed in detail in Macosko (1994).

2.2 Capillary Viscometers

Gravity-driven fluid flow through a narrow capillary was the basis for the earliest viscometers. The stress and strain fields are nonhomogeneous, but their variation is simple to describe. The shear rate is maximum near the wall and goes to zero at the center of the tube. For a Newtonian fluid, the resulting parabolic velocity profile leads to the familiar Hagen–Poiseuille law:

$$\mu = \frac{\eta}{\rho} = \frac{\pi g h R^4 t}{8VL}, \tag{49}$$

where μ = kinematic viscosity, η = absolute viscosity, ρ = fluid density, g = gravitational acceleration, R = radius of capillary, h = vertical distance between ends of capillary, V = volume of fluid flowing in time t, and L = length of capillary. Equation (49) yields the kinematic viscosity, requiring knowledge of the density for conversion to absolute viscosity. A more general form of the equation replaces $\rho g h$ with pressure and gives the absolute viscosity. Such pressure-driven devices are commonly used for high-viscosity fluids such as polymer melts. The shear rate can be quite high in capillary viscometers and is given by $\dot{\gamma} = 4V/\pi t R^3$. Although these viscometers might seem unsophisticated, in fact they provide some of the highest-accuracy data and are applicable for fluids ranging from gases to highly viscous polymer melts.

An important application is their use in defining the effect of small amounts of polymer upon the viscosity of a solution. For such measurements a capillary geometry described as an Ubbelohde tube viscometer is utilized (Young and Lovell, 1991). The use of such viscosity data in describing polymer conformation and molecular weight is described in Sec. 4.2.1. In polymer melts, a commonly cited quantity, proportional to viscosity, is the melt index. The capillary device used for its determination, however, does not produce a uniform velocity field. Thus the viscosity cannot be directly calculated from the melt index. True capillary geometry and various rotational or oscillatory viscometers can give viscosity values for such highly viscous melts.

2.3 Falling/Rolling-Sphere Viscometer

In the falling/rolling-sphere viscometers, the velocity of a falling or rolling sphere within the viscous liquid is monitored, and the viscosity of the liquid is extracted through the use of the Stokes equation, Eq. (11). As in capillary devices, the stress is inhomogeneous. However, the mathematical description is more complex because of spatial variations of the velocity field due to finite-sample-volume effects. This makes the shear rate in such a device very poorly defined; fortunately, the velocity of the moving object is typically slow, and with appropriate corrections these devices give accurate low-shear-rate viscosity values. Typically, the deviations from an ideal, infinite system are accounted for by an empirical correction factor that characterizes a particular geometry. Very accurate data can be obtained; accuracies better than 1% are commonly reported. Like the capillary devices, these viscometers can cover a wide range of viscosities, ranging from those of gases up to polymer melts near the glassy state. Because of their simple design and sealed geometry, they also find wide application for use at elevated pressures, as reviewed in King (1997).

2.4 Vibrational Viscometers

Damping of an oscillator immersed in a fluid can also be a sensitive measure of the fluid viscosity. Certain conditions must be met: the vibration frequency f must be well below the molecular relaxation rates, and the viscosity must be sufficiently low that the oscillation is strongly damped, i.e., the viscous penetration depth, $\sqrt{\eta/2\pi f \rho}$ where ρ is the fluid density, is small compared to the sample dimensions (see King, 1997, for a discussion of the constraints). One such device, used in the measurement of the temperature and pressure dependence of viscosity of gases and supercritical fluids, relies upon the damping of a vibrating wire (van Der Gulik *et al.*, 1988).

3. MEASUREMENT TECHNIQUES—DIFFUSION

3.1 Measurement of Mutual Diffusivity

Concentration gradients lead to particle fluxes that can be observed macroscopically. According to Fick's first law (see Sec. 1.1.2), a determination of the mutual diffusion coefficient of a certain pair of species of molecules can be based either on a measurement of both the flux density and the gradients of the concentration or on a measurement of the particle distribution at different times, supposing that the rearrangement within the sample is governed by diffusion phenomena (Crank, 1975).

Molecular concentrations can be determined by chemical as well as by various physical methods, since numerous properties such as weight, refractive index, birefringence, radioactivity, spectral absorbance, and emittance depend on the composition of the system. Therefore, a wide variety of experimental techniques for determining the mutual diffusion coefficients of systems have been developed (Jost and Hauffe, 1972; Kärger *et al.*, 1988).

3.2 Measurement of Self-Diffusivity

The random movements of individual molecules do not lead to a net transport of mass. Therefore, the displacement of the particles have to be monitored directly to determine their diffusivity. However, that requires either the labeling of a small minority of molecules (tracer methods) or the imposition of a field, varying on a molecular length scale, through which the ensemble of molecules move (NMR and scattering techniques).

3.2.1 Molecular Tracer Methods The traditional procedure of "tracer" experiments is to introduce different isotopes into the system under study. However, two major assumptions have to be fulfilled: First, the molecular mobility of the molecules must not be altered by the process of labeling. Second, in order to measure the coefficient of the self-diffusivity and not the coefficient of the mutual diffusivity (see Sec. 1.1.2), the concentration of the labeled molecules must be small enough to ensure that the mobility of any labeled molecules is unaffected by the existence of other labeled molecules.

There are several experimental techniques for monitoring the rearrangement of labeled molecules within an unlabeled matrix, like infrared spectroscopy (Klein, 1979; Klein and Briscoe, 1979) or forced Rayleigh scattering (FRS) (Hervet *et al.*, 1978; Antonietti *et al.*, 1984). Alternative ways of following the labeled species have been developed—e.g., by determining the recombination rate of free radicals from an analysis of electron-spin-resonance (ESR) spectra (Lebedev, 1967) or by measuring the variation of the local density of radioactive isotopes ("liquid scintillation counting") (Mills *et al.*, 1984). Generally, "tracer" techniques have the big advantage of being able to measure extremely slow molecular diffusivities (as small as 10^{-20} m^2/s). The often complicated process of molecular labeling sets limits in the application of these techniques.

3.2.2 NMR Mean square displacements of diffusing molecules can be directly monitored by the nuclear magnetic resonance (NMR) field-gradient technique. The self-diffusion coefficient can be calculated knowing the mean square displacement of the particles within a given time; see Eq. (17). It is required, however, that the displacements of the molecules be much larger than the mean lengths of the elementary jumps of diffusion. Up to now field-gradient NMR (FG NMR) is the only method that allows a direct measurement of molecular displacements over such large distances.

The principle of measuring molecular diffusivity using FG NMR is that the spatial position of each spin (and, hence, each molecule) is marked by its specific Larmor frequency within a well-defined, spatially inhomogeneous magnetic field. Diffusion of the marked molecules causes changes in the Larmor frequencies of the spins, which results in a measurable attenuation of the NMR signal. Recently, it has been shown (Fleischer and Fujara, 1994) that the FG NMR experiment can be interpreted in terms of a generalized scattering experiment, which implies the possibility of defining a "generalized scattering vector." The NMR field-gradient technique can cover length scales from about 100 Å (using the strong field gradients of the stray field of a cryomagnet) to a few microns. Since the accessible diffusion times vary from milliseconds up to seconds, a broad range of diffusion coefficients can be determined. This makes it possible to apply FG NMR to gases, liquids, and polymers. An excellent review of the principles and applications of the FG NMR can be found in Kärger *et al.* (1988).

3.2.3 Scattering Techniques Recent advances in laser and neutron spectroscopy have enabled scattering techniques to be widely applied in the investigations of diffusion processes. In these experiments information about the diffusion phenomena follows from an analysis of the Doppler-frequency shift of a monochromatic beam due to fluctuations in the observed systems that are caused by the motion of the molecular scatterers.

All scattering techniques measure the intensity of the scattered beam as a function of the energy difference $\hbar\Delta\omega$ and the momentum difference $\hbar\Delta\mathbf{q}$ between the incident and scattered beams. The scattering vector $\mathbf{q}$ is the difference between the wave vectors of the scattered and incident beam, $\mathbf{k}$ and $\mathbf{k}_0$, respectively, and is a function of the wavelength of the incident beam λ and the scattering angle θ:

$$|\mathbf{q}| = (4\pi/\lambda)\sin(\theta/2). \qquad (50)$$

3.2.3.1 Neutron Scattering. Among the scattering methods, neutron scattering is distinguished by the fact that the wavelength and the energy of the incident beam are respectively on the order of the atomic distance and thermal energy. The quantity measured in neutron scattering is the differential scattering cross section, $\partial^2\sigma/\partial\Omega\partial E$, i.e., the probability with which an incident neutron of energy E_0 is scattered into an angle element $d\Omega$. The motion of the individual nuclei can be studied by an analysis of the incoherent scattering process. Experimentally, one determines the so-called structure factor $S(\mathbf{q},\omega)$ from the intensity of a beam scattered into a certain angle element. For the case of incoherent scattering it can be shown (Springer, 1972) that the incoherent structure factor $S_{\text{inc}}(\mathbf{q},\omega)$ is the Fourier transform of the self-correlation function, $G_S(\mathbf{r},t)$:

$$S_{\text{inc}}(\mathbf{q},\omega) = \iint G_S(\mathbf{r},t)\exp[i(\mathbf{q}\cdot\mathbf{r} - \omega t)]d^3\mathbf{r}dt. \qquad (51)$$

Therefore, $S_{\text{inc}}(\mathbf{q},\omega)$ provides the information about the ensemble-averaged probability density that a particle has been shifted from position $\mathbf{r}_0$ to the position $\mathbf{r}_0 + \mathbf{r}$ within the time interval t.

The time scale over which molecular migration can be followed is determined by the energy resolution of the scattering equipment. In conventional neutron-scattering spectrometers (see SPECTROMETERS, NEUTRON), an energy resolution of $\Delta E_{\min} = \Delta(\hbar\omega) \approx 10^{-6}$ eV may be achieved, which corresponds to maximum observation times of about 10^{-8} s. Observation times longer by almost two orders of magnitude have been achieved by the introduction of neutron spin-echo spectroscopy. The largest scanned space scale, $r_{\max} \approx 2\pi/|\mathbf{q}_{\min}|$, is on the order of 30 to 50 nm and is determined by the minimum value of momentum transfer, or scattering vector $\mathbf{q}$, that is experimentally accessible (Mezei, 1983).

3.2.3.2 Light Scattering. Traditional light-scattering techniques using thermal sources (pre-1960) were confined to the investigation of nondiffusive phenomena because of the relatively low power and high bandwidths of conventional light sources and filters. Advances in optical and electronic detection devices together with the availability of laser light sources have eliminated these restrictions and have led to the development of "photon correlation" and "light beating spectroscopy" as powerful means for light scattering investigations.

The principle of diffusion measurements by light scattering is that fluctuations of the concentration on a molecular level lead to a line broadening of the scattered radiation with respect to the incident one. Since any fluctuations of the local density relax back to equilibrium through diffusion, studies of the line broadening can yield values of the relevant diffusion coefficients (Berne and Pecora, 1976).

Light scattering allows observation times from microseconds to several seconds, with a space scale, $2\pi/|\mathbf{q}_{\min}|$, typically on the order of a few microns. The probing lengths are generally considerably longer than the size of individual atoms or molecules. Self-diffusion can be studied by working in the limit of very high dilution (see Sec. 1.1.2.2) (Tyrrell and Harris, 1984).

Viscosities of liquids can also be determined using light scattering techniques by observing the diffusive motion of large (micron-sized) tracer particles placed in the liquid of interest. Viscosities can then be determined through the Stokes–Einstein equation given in Eq. (26). Recent detailed and accurate experiments have compared this technique to the more traditional mechanical rheometric methods and have demonstrated accuracies of order 1%.

A summary of the typical working regimes of classical techniques for measuring molecular self-diffusion is illustrated in Fig. 3.

4. MATERIALS/SYSTEMS

In the preceding sections we have dealt with concepts applicable to simple gases and liquids, the most common materials for which one would like to understand viscosity and diffusion. In the following sections we deal with systems that illustrate new concepts. In the first instance, colloidal suspensions, we encounter the effects of two-phase flow, but we also see that after rescaling, these systems can illustrate many behaviors characteristic of hard-sphere systems. They become model systems for molecular liquids when treated in this way. The second section deals with the behavior of polymers. The viscosity and diffusion of these materials is vital to their understanding, and because of their large molecular size, the mechanisms underlying their viscosities and diffusions differ substantially from those of the molecular systems covered in the main sections of this text.

4.1 Hard-Sphere Colloidal Suspensions

In this section we give a brief introduction to the subject of diffusion and viscosity of suspensions of micron-sized colloidal particles that interact like hard spheres. Conceptually such suspensions are very similar to hard-sphere atomic liquids, with the essential difference being the presence of an interparticle, or background, solvent. At low particle densities, the presence of a solvent leads to profound differences in the particle dynamics between the two systems—hard spheres move ballistically, colloidal particles execute Brownian motion. At high densities, the dominant effect in both systems that determines particle motions, and hence sample viscosities, is the direct particle–particle interactions due to the close crowding, or caging, of the spheres. Consequently, as outlined below, at high densities, many of the concepts derived originally in atomic liquids for the density dependence of the viscosity, as well as the use of the Stokes–Einstein equa-

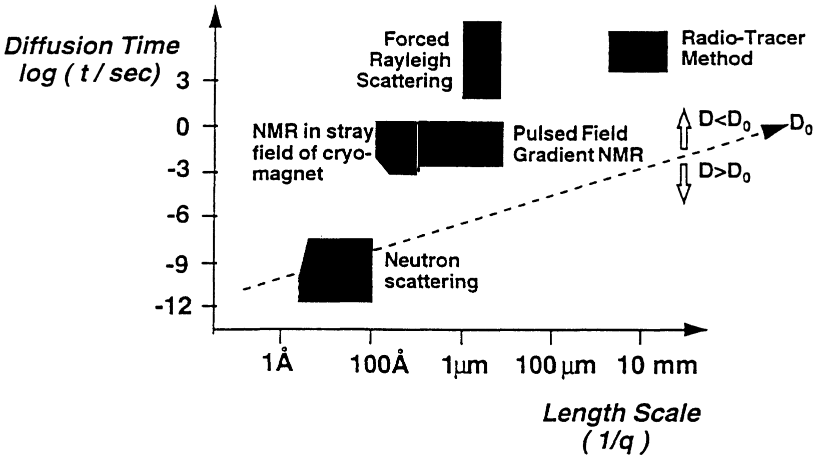

FIG. 3. Typical working regimes of classical techniques for measuring molecular self-diffusion. The dashed line has a slope of 1 and represents the trajectory of a particle diffusing according to normal self-diffusion, Eq. (26), with a certain diffusivity, D_0. A higher diffusion coefficient would displace this line downward; lower diffusivity, upward.

tion, are useful also in describing colloidal suspensions. One of the distinct advantages of studying colloidal systems is that the large particle sizes, of order 1 μm as compared to 1 nm in molecular liquids, lead to very slow dynamical processes, which are easily accessible via such commonly used techniques as dynamic light scattering (DLS) (Pusey, 1991).

4.1.1 Diffusion A single particle executes a random walk according to Eq. (27) with the Stokes–Einstein diffusion coefficient, D_0, given by Eq. (26). In semidilute suspensions individual spheres still execute Brownian displacements, albeit at a reduced diffusion coefficient, $D < D_0$, resulting from the influence of both the solvent-mediated hydrodynamic interactions and direct interactions between colliding spheres (Pusey, 1991). For concentrated suspensions it has been shown that particle motions proceed via a two-stage process. That is, there is an initial rate of diffusion for small displacements "inside" the cage of neighboring spheres, and a considerably slower diffusive process at longer times as the particles travel through the surrounding matrix (Pusey, 1991; Segrè and Pusey, 1996). As an example, Fig. 4 displays values of the (inverse) normalized short- and long-time self-diffusion coefficients, D_S^{Self} and D_L^{Self}, respectively, of suspensions of poly-(methyl methacrylate) (PMMA) spheres in *cis*-decalin (Segrè *et al.*, 1995a). The large disparity in the values at high concentrations, $\phi > 0.4$, reflects the tightening of the neighbor cages, which also shows up as an increase in the correlated short-range structure. This phenomenon is also found in simple liquids; see Sec. 1.2.7.

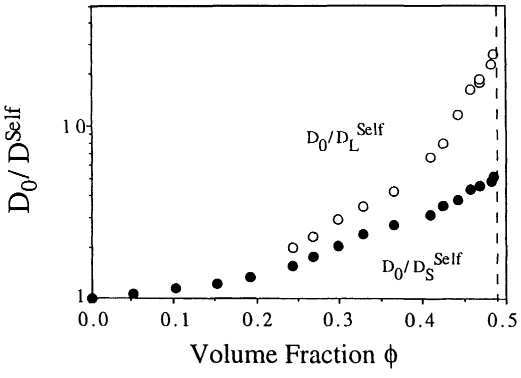

FIG. 4. Inverse normalized short- and long-time diffusion coefficients D_0/D_S^{Self} (closed circles) and D_0/D_L^{Self} (open circles) versus volume fraction ϕ. The line at $\phi = 0.494$ represents the disorder/order (freezing) transition. The considerable slowing down of D_L^{Self} represents the tightening of the neighbor cages as the volume fraction increases. Data from Segrè (1995).

Connections between the diffusion behavior of dense suspensions and molecular fluids have been the subject of numerous articles (Cohen, 1993) in recent years. Simulations of hard-sphere fluids have also exhibited two-stage diffusional processes similar to that seen in suspensions (Lust *et al.*, 1993). It is interesting to note that in suspensions, the onset of a long-time slow relaxation process occurs at a similar volume fraction ($\phi \sim 0.4$) to that at which the Enskog theory in simple liquids fails. This suggests that similar complex multiparticle correlations arise in both systems at similar densities. Further detailed comparisons, however, would benefit from more detailed relaxation spectra in simple liquids at moderate densities. At very high densities, $\phi > 0.5$, near to the glass transition, mode-coupling theories (see Sec. 1.2.7) have been developed for hard-sphere (atomic) fluids (Götze, 1991). Somewhat remarkably, these same theories also accurately account for the dynamics of very concentrated hard-sphere colloidal suspensions (van Megen and Underwood, 1993, 1994), the implication being that the effect of particle caging dominates the influence of the background solvent.

4.1.2 Viscosity When colloidal particles are placed in a solvent of viscosity η_0, the solution viscosity is found to increase. In the limit of low concentrations, Einstein found that the viscosity increase depends upon the volume fraction ϕ of the particles. He derived the linear term in ϕ, and later Batchelor (Batchelor, 1976) the ϕ^2 term, yielding

$$\eta = \eta_0(1 + \tfrac{5}{2}\phi - 6.55\phi^2). \qquad (52)$$

This expression, especially the linear term, finds wide application in expressing the viscosity of suspensions. The complete form accurately describes experimental data up to about $\phi = 0.10$. At higher densities Eq. (52) fails, and more complex theories are needed.

The viscosity in dense suspensions displays non-Newtonian behavior, with both a shear-rate and a frequency dependence. The key to understanding the behavior of the viscosity lies in examining the underlying suspension microstructure and its relaxation properties under shear (de Kruif *et al.*, 1985). The application of shear distorts the equilibrium microstructure, while Brownian motion tends to restore it. The dimensionless Peclet number, defined as the ratio of the time for a single particle to diffuse through its own radius to that needed to convect through its own radius under shear, is given by $Pe \equiv R^2\dot{\gamma}/D$. For values of the shear rate $\dot{\gamma}$ such that $Pe \ll 1$, the microstructure is only moderately perturbed, Brownian motion dominates the particle motions, and one measures the low-frequency shear viscosity η. For $Pe \gg 1$, the microstructure becomes severely distorted, and the high-frequency shear viscosity, η_∞, is found. Dense suspensions exhibit shear thinning, i.e., $\eta > \eta_\infty$, with the difference being attributed to the specific effects of Brownian motion. In practice, for submicron-sized particles, both regimes are experimentally observable. It is also worth noting that computer simulations of atomic liquids have also shown the effects of shear thinning. However, because of the small size of molecular liquids, $R \sim 0.1 - 1$ nm, experimentally unattainable shear rates of $\dot{\gamma} \geq 10^{12}$ s^{-1} would be required to see it (Evans *et al.*, 1981). For the sake of brevity we restrict the discussion in this section to the low-shear viscosity only.

In hard-sphere colloidal suspensions, the viscosity increase with ϕ has been measured under low–shear-rate conditions (Russel and Sperry, 1994; Segrè *et al.*, 1995b). At the equilibrium freezing point for monodisperse hard spheres, $\phi = 0.494$, a viscosity enhancement of about 50, relative to the background solvent, is found. At higher packing fractions the glassy state can eventually be approached. An interesting result has been obtained by de Kruif *et al.* (1985) in their studies, namely, that the results can be fitted by the relation $\eta/\eta_0 = 1/(1 - \phi_m)^2$, where $\phi_m = 0.64$, the random close-packing value at which all the particles are in direct contact and diffusion ceases. More recent studies (Phan *et al.*, 1996) have used the same fitting form and extracted $\phi_m = 0.58$, the value at which suspensions become glasslike. Overall, there is no canonical theory for this regime. Indeed, this is an area of great interest at present both for the intrinsic interest in colloidal suspensions and for the possibility that these systems can act as analogs for simple liquids.

4.1.3 Stokes–Einstein Relation Several studies have tested the validity of the Stokes–Einstein relation, Eq. (26), in concentrated colloidal suspensions (van Blaaderen *et al.*, 1992; Segrè *et al.*, 1995b). Segrè used the values of η and D_L^{Self} measured on hard-sphere-like suspensions and used the relation $D = k_BT/C\pi\eta R$ to extract the constant C, with the results shown in Fig. 5. In the dilute limit, i.e., for a single sphere, the exact relation for nonslip boundary conditions, Eq. (26), shows that $C = 6$. As shown in the figure, at the highest concentrations the measured values decrease to $C \sim 4$, which parallels what is generally also observed in dense hard-sphere/molecular fluids (see Sec. 1.2.4). These results indicate that there is no simple Stokes–Einstein equation (stick or slip) that holds over all volume fractions. This is perhaps not surprising given the assumption of an isolated sphere in the derivation of the Stokes–Einstein equation. A more complete theoretical understanding of these results, as well of viscosity of these systems in general, remains to be given.

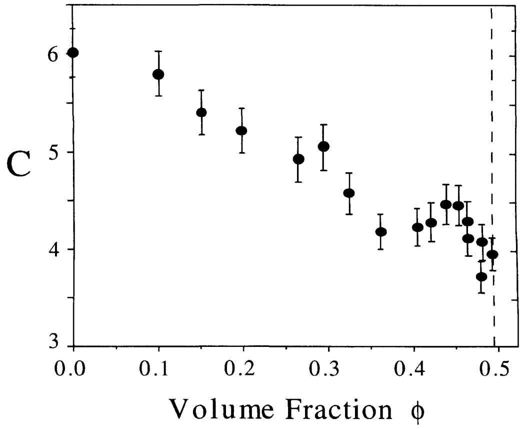

FIG. 5. Values of the constant C in the Stokes–Einstein equation [Eq. (26)] derived from measurements of the viscosity η, the long-time self-diffusion coefficient D_L^{Self}, and the particle radius R in suspensions of poly-(methyl methacrylate) (PMMA) spheres. Note the evolution from stick condition ($C = 6$) at low volume fractions to slip condition ($C = 4$) as volume fraction increases. Data from Segrè (1995).

4.1.4 Viscosity and Structural Relaxation One of the principal advantages of micron-sized colloidal suspensions as opposed to atomic liquids is that it is readily possible, using dynamic light scattering (DLS), to study accurately the dynamics of cage relaxation. Dense liquids, colloidal or molecular, do not display long-range order but possess strong short-range particle correlations on the scale of 1–3 diameters. This is exhibited in Fourier space as a strong peak in the static structure factor $S(q)$ at $q \approx q_m \approx \pi/R$. Under shear this ordering gets distorted, and its diffusional relaxation back to equilibrium plays a key role in several theories of viscosity (Geszti, 1983; de Schepper *et al.*, 1989; Götze, 1991; Brady, 1993). To study fluctuations in suspensions on this length scale is straightforward with DLS. However, it is only achievable for molecular liquids using neutron-scattering techniques, of which there are only limited results to date (Cohen, 1993). Consequently, most diffusion results in liquids concern either long–length-scale collective diffusion or self-(tracer) diffusion of isotopically labeled species. A recent DLS study in suspensions has examined directly the dynamics of structural relaxation—and characterized it by (long-time) diffusion coefficients $D_L(q_m)$. The results are shown in Fig. 6. First, the inset shows that, as a result of the cage effect, diffusion at the peak differs from single-particle self-diffusion, the former being consistently smaller. Second, for all values of ϕ up to 0.50, there is a close coincidence of the normalized viscosity and structural relation rates, consistent with the simple relation

$$\eta/\eta_0 = D_0/D_L(q_m). \tag{53}$$

Note that Eq. (53) can also be rewritten in the Stokes–Einstein-type format as $D_L(q_m) = k_BT/6\pi\eta R$. Since the long-time diffusion process is the α-relaxation, the proportionality in Eq. (53) is similar to the Maxwell relation between viscosity and τ_α as discussed in Sec. 1.2.7.

4.2 Polymers

The enormous variety of polymers and their uncounted applications have led polymer scientists to define many different types of viscosity and diffusion, and this can lead to confusion for those new in the field. The clearest division of behavior is between that of solutions of polymer molecules dispersed in a solvent (polymer solutions) and that of solvent-free systems (polymer melts). Moreover, there are five different regimes in the molar mass–concentration diagram of polymers. Figure 7 illustrates such a phase diagram for the example of a polystyrene–benzene system. At low volume fractions of polystyrene, Φ_{Polymer}, the polymer coils do not overlap one another. The threshold for overlapping coils is given by the so-called overlap-concentration c^*, which decreases with increasing molecular weight of the polymer. Systems with a polymer concentration greater than c^* are called semidilute so-

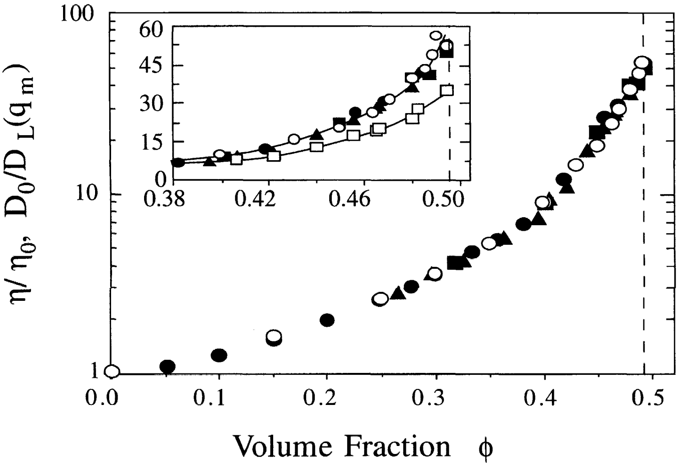

FIG. 6. Normalized low-shear viscosity η/η_0 (open symbols) and long-time peak diffusion $D_0/D_L(q_m)$ [see Eq. (53)] versus volume fraction ϕ for suspensions of poly-(methyl methacrylate) (PMMA) spheres. Inset: The same data on a linear scale at high volume fractions with in addition the open squares being measurements of the long-time self-diffusion coefficient D_0/D_L^{Self}. Note in the inset that diffusion at the peak differs from single-particle self-diffusion because of the cage effect. Note also the excellent scaling of the long-time peak diffusivity and viscosity over the entire range of volume fraction. Data from Segrè (1995).

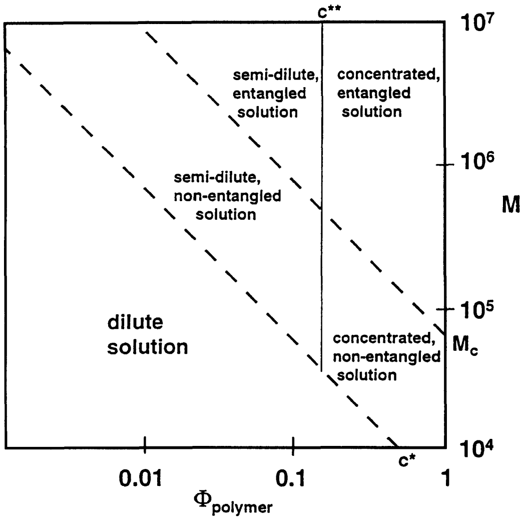
c**
semi-dilute, entangled solution
concentrated, entangled solution
semi-dilute, non-entangled solution
dilute solution
concentrated, non-entangled solution
c*
10^7
10^6
M
10^5
M_c
10^4
0.01
0.1
1
$\Phi_{polymer}$

FIG. 7. Different concentration regimes in polymer solutions illustrated through the example of a polystyrene/benzene system.

lutions. The semidilute regime turns into the concentrated regime with the onset of polymer–polymer interactions. The corresponding concentration is abbreviated c^{**}. Both the semidilute and the concentrated regimes have to be divided into entangled and nonentangled solutions because of the onset of topological constraints for long polymer chains, as described in Sec. 4.2.2.

In the following sections we focus on dilute solutions and polymer melts only. These two regimes represent the extreme cases of the dynamic behavior of polymers, on the one hand, and are understood best, on the other hand. The semidilute and concentrated solution behavior is too complex to be treated here but can be found elsewhere (Elias, 1971).

In the description of the diffusional behavior of polymer systems, we focus on self-diffusion only. The results of self-diffusion can easily be transferred to rotational diffusion as introduced in Sec. 1.1.3 for the case of dilute, simple molecular systems. In multicomponent systems one will encounter mutual or transport diffusivities. However, relating the mutual-diffusion coefficients to the underlying self-diffusion coefficients of the species in the solutions is nontrivial and beyond the scope of this encyclopedia.

4.2.1 Dilute Solutions

Low-shear-rate viscosity in the dilute-solution regime is closely related to the polymer conformation. Since this relation is well understood theoretically, it is a useful characterization tool; for example, the molecular weight of linear polymers can easily be obtained. Also, it is a powerful tool to measure the change in polymer conformation due to solvent effects. The hydrodynamics is similar to the two-phase flow we encountered in Sec. 4.1.2, and the Einstein equation, Eq. (52), gives the essential relationship for these dilute polymer solutions. The fluid consists of dispersed polymer coils with their concentration related to the mass of polymer per unit volume of fluid, c. Furthermore, each polymer coil is characterized by a mass, the molecular weight M_W, and a size, the hydrodynamic volume V_H. We saw earlier in the Einstein equation that the viscosity enhancement due to addition of spherical particles is proportional to the volume fraction, $\frac{5}{2}\phi$. We note here that $\phi = (c/M_W)N_AV_H$, and obtain an equation for the viscosity of a dilute solution. However, by convention polymer science chooses to define a new viscosity, the specific viscosity, η_{sp}, which equals $\eta/\eta_0 - 1$ of the Einstein equation; thus, $\eta_{sp} = \frac{5}{2}(c/M_W)N_AV_H$. Recall that the Einstein equation is applicable to noninteracting particles at low concentrations. This suggests that for real systems extrapolation to zero concentration would yield the most consistent results. In polymer science this extrapolated value of the specific viscosity is named the intrinsic viscosity, $[\eta]$, defined as follows:

$$[\eta] = \lim_{c\to 0}(\eta_{sp}/c);$$
$$\eta_{sp} = [\eta]c + k_H[\eta]^2c^2. \qquad (54)$$

The second equation is simply a Taylor's expansion and is known as the Huggins equation. It provides a tool to extract an intrinsic-viscosity value from experimental data and yields a parameter, k_H, proportional to the strength of polymer–polymer interaction in a given solvent.

The intrinsic viscosity is a key property of a polymer since it is the main parameter relating solvent viscosity to polymer conformation. For example, an important scaling relationship between average molecular weight and intrinsic viscosity is the Mark–Houwink equation (Young and Lovell, 1991),

$$[\eta] = K(\bar{M}_\nu)^a. \tag{55}$$

The quantities K and a are determined experimentally and are often tabulated for various polymer-solution combinations, which provides a simple means of determining the viscosity-averaged molecular weight $\bar{M}_\nu$ of a sample. The underlying molecular origin of the Mark–Houwink equation is seen in the Flory equation (Young and Lovell, 1991)

$$[\eta] = \Phi R^3/M_W, \tag{56}$$

where Φ is a constant and M_W is the weight-average molecular weight. This form of the equation can be used to probe the effective size of the polymer, R^3, under changing solvent conditions. For example, recent high-pressure viscosity measurements have shown that for certain water-soluble polymers the intrinsic viscosity changes significantly, indicative of a collapse of the polymer chain (Cook *et al.*, 1992b). This is in contrast to hydrocarbon solvents where no change is found (Cook *et al.*, 1992a). These effects arise from the unique role that hydrogen bonding plays in the solvation of nonionic water-soluble polymers (Sun and King, 1996).

Diffusion of individual polymer coils in dilute solutions is often measured via dynamic light scattering. The diffusion coefficient of the polymer coils is connected to their effective radius and the viscosity of the solvent through use of the Stokes–Einstein equation. The effective size of the polymer coil is typically reported as V_H, the hydrodynamic volume, as discussed above. Expansion and contraction of this volume can be related to the solvent–polymer interaction strength (Sun and King, 1996).

The diffusivity of the polymer coils in the dilute regime can be directly observed via field-gradient NMR if the NMR signal from the solvent molecules can be suppressed. This can be achieved, e.g., by use of deuterated solvents, which are invisible compared to the protonated polymers during ^{1}H NMR experiments.

4.2.2 Melts A flexible polymer chain is capable of assuming an enormous number of possible configurations by rotations of its chemical bonds. Therefore, a macroscopic deformation of a polymer causes complex structural rearrangements that extend from single chemical bonds to the displacement of whole polymer chains. The corresponding time scales span from a few picoseconds up to several days. This feature, specific for macromolecules, causes their mechanical properties to be very distinct from other materials: As a function of shear rate, the polymer can behave either as a liquid or as a rubber, depending upon the kinematics of the deformation, the temperature, and the molecular weight of the polymer. This behavior is called viscoelasticity. Although the phenomenon of viscoelasticity is too complex to be completely explained in this section (see, e.g., Ferry, 1980), two of its features have to be discussed because of their strong impact on the viscosity and the diffusion of polymer melts.

For polymer melts a continuous range of behavior ranging from viscous to elastic is encountered as the frequency of investigation is scanned from low to high. This effect becomes especially important for the understanding of the viscosity of polymer melts. The second remarkable feature of the dynamics of long polymer chains arises from the fact that these one-dimensional objects cannot cross each other. Therefore, the molecular rearrangement becomes restricted because of the onset of topological constraints, known as entanglements, between the polymer chains. Entanglements prevent the chains from displacing laterally beyond a certain distance; i.e., each polymer chain can be assumed to be confined within a tube that only allows displacements of the chains along their own backbone. The diameter of this tube represents an additional length scale that determines the transport in melts.

4.2.2.1 Viscosity of Melts. Because of the viscoelastic behavior, the complete description of the viscosity of polymer melts is well beyond our discussion. However, we can focus on the viscous component alone and consider the behavior in the low-shear limit. We find then that polymer melts exhibit two characteristic behaviors. First, since we are only considering here polymers that do not form crystals, as the polymer melt is cooled, the viscosity continuously rises and the glassy state is formed. Transitions into the glassy state for polymer melts are quite common; a good example of such a material is polystyrene. This increase in viscosity with decreasing temperature can be scaled through use of the Williams–Landel–Ferry (WLF) equation:

$$\log\left(\frac{\eta}{\eta_{T_g}}\right) = \frac{-C_1(T - T_g)}{C_2 + (T - T_g)}, \tag{57}$$

where C_1 and C_2 are constants and η_{T_g} is the viscosity at the glass transition. Values for the constants are tabulated; for example, see Sperling (1992).

The second important characteristic of polymer melt viscosity is its dependence on molecular weight. Here, a remarkable effect is that two different scaling relations are found: for $M < M_C$,

$$\eta = K_L M^{1.0}, \tag{58}$$

and for $M > M_C$,

$$\eta = K_H M^{3.4}.$$

For chains shorter than a critical molecular weight, M_C, the viscosity in the limit of zero shear is directly proportional to the molar mass after corrections are made to achieve a constant free-volume state, i.e., after adjusting for chain end effects. In contrast to this, it is experimentally found (Berry and Fox, 1968) that chains with a molecular weight above M_C scale with $M^{3.4}$.

The threshold molecular weight M_C is a material-dependent property. Empirically, it was found that M_C is roughly equal to $2M_e$, where the entanglement molecular weight M_e is the molar mass of that fraction of a chain between two entanglements. The entanglement length is an important concept in describing the full complexity of polymer viscoelasticity (Ferry, 1980; Kremer and Grest, 1995). The proportionality factors, K_L and K_H, are empirical constants valid for the low (L) and high (H) molecular-weight limits (Sperling, 1992).

4.2.2.2 Diffusion of Melts. It is obvious that topological constraints strongly influence the molecular mobility of polymer chains. Similar to the behavior of the viscosity discussed in the previous section, entanglements clearly distinguish the translational dynamics of unentangled chains from that of entangled ones. In this section we begin with a brief description of the Rouse model (Ferry, 1980) for unentangled chains, followed by a discussion of the effects of topological constraints on the diffusion of entangled polymers.

The Rouse model provides a reasonable description of the translational dynamics of unentangled polymer chains in the melt. In his attempt to understand the dynamics of chains in solution, Rouse argued that a chain could be considered as a series of beads connected by Hookean springs, where each spring constant is $3k_BT/b^2$; b is the separation of the beads. The position of each bead chain is determined by a set of coordinates $\{R_1, R_2, \ldots, R_N\}$, and each bead experiences an averaged frictional drag, the monomeric friction coefficient ξ. Therefore, the central assumption of this model is that the dynamics of a chain is governed by localized interactions along the chain. That assumption causes the Rouse model to fail at describing the translational dynamics in solutions, mainly because it neglects the effects of intramolecular hydrodynamics and excluded-volume interactions. However, it is well suited for understanding the dynamics of unentangled chains in the melt where long-range interactions are assumed to be screened.

According to the Rouse assumption, the motion of each bead can be described by a Langevin equation [see Sec. 1.1.3, Eq. (24)] with the frictional drag equal to the effective spring force $(3k_BT/b^2)(2R_n - R_{n+1} - R_{n-1})$. Consequently, the motion of the beads is similar to the Brownian motion of coupled oscillators, and the whole system can be described in terms of normal-mode coordinates where each mode behaves independently of the other. Solving the Langevin equation yields a spectrum of N relaxation times where each relaxation time τ_i corresponds to one normal mode. The major consequence of this approach is that the diffusion coefficient D_S is inversely proportional to the number of monomers, N:

$$D_S = k_BT/\xi N, \tag{59}$$

if the diffusion time t is long compared to the relaxation time of the longest relaxation mode,

$$\tau_1 = \frac{b^2}{3\pi^2 k_B T}\frac{1}{\xi N^2}. \qquad (60)$$

Entangled polymer melts show a totally different dynamic behavior than this and have challenged polymer science for some time. The breakthrough for understanding the dynamics of entangled polymer melts was made by de Gennes about 25 years ago when he introduced his reptation model (de Gennes, 1971). In this model, a polymer chain is imagined to undergo one-dimensional translational motion in the presence of a fixed network of obstacles. The motion is performed by a flow of a series of kinks along the contour of the chains, which can be visualized by the motion of a snake. After this qualitative understanding, Doi and Edwards (1978) theoretically described the snakelike motion of an entangled chain within a fixed tube to account for the effects of entanglements. This tube model allowed the qualitative prediction of most of the behavior of entangled melts. However, it should be mentioned that some of the experimental findings are not in agreement with the predictions of the Doi–Edwards theory but support more recent models and refinements of the Doi–Edwards theory. We briefly discuss those deviations at the end of this section.

The theoretical description of the reptation movement of an entangled chain by Doi and Edwards assumes that the chain performs a Rouse diffusion within a temporally fixed tube. The motion of the chain occurs by the curvilinear propagation along an average trajectory that Edwards called the primitive path. Edwards defined the primitive path as the shortest distance connecting the ends of the chains that has the same topology as the chain itself (Doi and Edwards, 1986); i.e., each primitive path can be imagined as a short cut along the random walk of the chain coil. The tube is created by the topological constraints set up by neighboring chains. According to Straube, the length of the primitive path, a, can be interpreted as a measure of the tube diameter d (Straube and Heinrich, 1984):

$$a = \sqrt{C_\infty N_E}\,b \approx 1.7d, \qquad (61)$$

where C_∞ is the characteristic ratio describing the stiffness of the polymer chain and N_E, the entanglement chain length, is the number of monomers between two entanglements.

The major result of the Doi–Edwards theory for describing the movement of an entangled chain is that the diffusivity of the center of mass of the polymer coil depends more strongly on the molecular weight than for unentangled chains:

$$D_S = k_B T/3\xi\, N_E/N^2. \qquad (62)$$

Comparing the equations for the diffusivity of unentangled and entangled chains [Eqs. (59) and (62), respectively] proves that the diffusivity of the center of mass of a reptating chain is equal to the diffusivity of an unentangled chain reduced by a factor of 3 (accounting for the crossover from three-dimensional Rouse diffusion to one-dimensional diffusion along the tube) and reduced by the ratio of N/N_E accounting for the influence of the entanglement density. Equation (62) is valid for diffusion times t that are longer than the so-called reptation time, t_{Rep}, i.e., longer than the time a reptating chain needs to propagate over a length equal to the contour length of the tube:

$$t_{\mathrm{Rep}} = \frac{a^2\xi}{2k_B T}\frac{N^3}{N_E^2}. \qquad (63)$$

The Doi–Edwards model is capable of describing most of the observed behavior of entangled polymers, especially the rubbery mechanical response. However, a variety of experimental observations are not qualitatively consistent with the simple reptation/tube picture. These include, e.g., the nonexponential decay of the stress relaxation (Ngai *et al.*, 1986), the 3.4 ± 0.2 exponent of Eq. (58) (3.0 is predicted by the Doi–Edwards theory), and the time dependence of the segmental diffusion of the chains within the tube that is stronger than the -0.5 prediction of the Doi–Edwards theory (Appel *et al.*, 1994). These experimental results support more recent theories assuming that the entangled chains have a considerably larger freedom for lateral motion than the strict reptation picture allows (Ngai *et al.*, 1986; Schweizer, 1989; Herman, 1992). The use of molecular dynamics and Monte Carlo computer simulations to address such questions is restricted because of their limitations to relatively small densities (Kremer and Grest, 1990).

If the lengths of the polymer chains, N, are comparable with the entanglement chain length N_E, the model of a temporally stable tube is not valid any more. Because of the diffusion of the surrounding chains, fluctuations of the tube occur that result in an increase of the effective tube diameter. Graessley described this process as lateral jumps of a primitive path step that occurs if an entanglement is released because of the diffusion of the neighboring chains and called this mechanism "constraint release" (Graessley, 1982). The effective diffusivity in this region of the chain length is found to be sum of the diffusivities according to the reptation model, $D_{S,\mathrm{Rep}}$, and the "constraint release" mechanism, $D_{S,\mathrm{c.r.}}$:

$$D_S = D_{S,\mathrm{Rep}} + D_{S,\mathrm{s.r.}} = \frac{k_B T}{\xi}\left(\frac{N_E}{3N^2} + \frac{\alpha N_E^3}{N^4}\right). \tag{64}$$

The factor α in Eq. (64) is an empirical value that represents the number of neighboring chains forming a constraint; its value is usually found between 3 and 5.

Figure 8 illustrates the theoretical dependence of the self-diffusivity of polymer melts as a function of their molecular weight. For molecular weights below the entanglement molecular weight M_e, the self-diffusion coefficient scales with M^{-1}. At very high molecular weight, the topological constraints described by the reptation model and the Doi–Edwards theory result in a scaling of the self-diffusivity with M^{-2}. Near the critical molecular weight, M_c of Eq. (58), additional fluctuations of the tube as explained by Graessley's constraint-release model result in an even stronger dependence of the self-diffusivity on the molecular weight. If the propagating chains are forced to remain within a tube, two degrees of freedom are lost for the process of migration. According to the Doi–Edwards theory, this loss results in a sudden reduction of the self-diffusion coefficient of the Rouse model at the entanglement molecular weight to $\frac{1}{3}$ if the dependence according to the reptation model is extrapolated to the entanglement molecular weight.

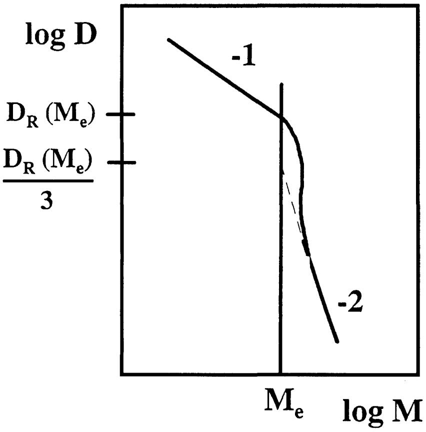

FIG. 8. Theoretical dependence of the self-diffusivity of polymer melts as a function of their molecular weight. The entanglement molecular weight M_e is indicated. Note that the crossover from three-dimensional Rouse diffusion below M_e to one-dimensional diffusion within the tubes (constraint release and reptation) above M_e is connected with a reduction of the molecular mobility to $\frac{1}{3}$ as a result of the reduced degree of freedom of the confined chains.

The dependence of the self-diffusivity on the molecular weight has been extensively studied. However, because of the broad range of molecular mobility to be covered and the drastic reduction of the diffusivity if the molecular weight exceeds its critical value M_e, only a few techniques are suited to perform such experiments. Figure 9 shows the result of pulsed field-gradient NMR measurements on different molecular weight fractions of poly(dimethyl)silioxane at a temperature of 60 °C (Appel and Fleischer, 1993). Note that the range of diffusivity covers about five orders of magnitude. The crossover from Rouse diffusion below M_e to constraint release and reptation above M_e is clearly visible. An interesting finding of those experiments is that reptation diffusion within stable tubes is only performed if the molecular weight exceeds the entanglement molecular weight, M_e, by about a factor of 8 to 10; i.e., tube fluctuations are found over a rather broad range of molecular weight.

GLOSSARY

Absolute Viscosity: Equivalent to **Shear Viscosity.** Compare with **Kinematic Viscosity.**

Anomalous Diffusion: Molecular migration that does not follow the direct propor-

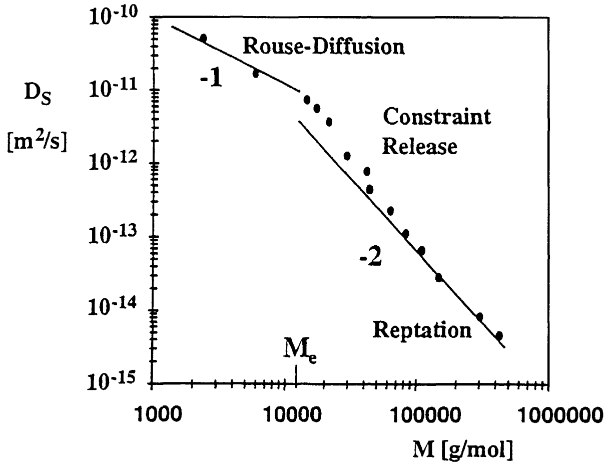

FIG. 9. Self-diffusion coefficients in the melt for different molecular weight fractions of poly(dimethyl)siloxane at $T = 60$ °C as measured by pulsed field-gradient NMR. The three different diffusion regimes (Rouse diffusion, constraint release, and reptation) can be clearly distinguished. Note that the crossover into the M^{-2} regime covers a large range of molecular weight; i.e., temporarily fixed tubes are only established if the molecular weight exceeds the entanglement molecular weight M_e by a factor of 8 to 10. Data from Appel *et al.* (1993).

tionality between mean square displacement and diffusion time because of correlations between the directions of successive elementary jumps.

α-Relaxation: Long, slow relaxation time observed in dense liquids. Its value is related to viscosity. Identified with diffusional motion of a particle out of the surrounding cage of near neighbors.

Boltzmann Theory: Relates the viscosity and diffusion for a dilute gas to the molecular mass and radius. A square-root temperature dependence for both viscosity and diffusion is predicted for a hard-sphere fluid. Extension to a Lennard-Jones fluid requires integration of the interparticle forces and leads to more realistic temperature dependence.

Bulk Viscosity: Coefficient describing energy dissipation due to expansion and contraction of a fluid. Except for specific conditions (see text), its contribution is negligible compared to shear viscosity.

Constraint Release: Theory for describing the diffusional behavior of entangled polymers near the entanglement molecular weight.

Critical Molecular Weight: Threshold molecular weight of polymers above which the viscosity strongly increases because of entanglements of the polymer chains.

Doi–Edwards Theory: Theory describing dynamic behavior of entangled polymers by assuming that a polymer chain is confined within a temporally fixed tube that is set up by neighboring chains.

Einstein Equation: Equation describing the viscosity increase of a solid/liquid mixture as a function of the volume fraction of solid.

Enskog Theory: Extension of the Boltzmann theory to higher density liquids but retaining assumption of uncorrelated binary collisions thus limiting its applicable density range.

Entanglement Length: Length of the fraction of a polymer chain between two entanglements.

Fick's Laws: Laws relating the particle flux to concentration gradient (first law) and describing the time derivative of the concentration for systems that are apart from their equilibrium state (second law).

Free-Volume Theory: Theory relating the diffusion and viscosity to the free volume available for molecules to move. Finds application to variable pressure data over wide ranges of both viscosity and density.

Hard-Sphere Model: A model for gases and liquid systems that assumes the molecules are identical hard spheres.

Intrinsic Viscosity: Limiting viscosity increment due to addition of polymer to a dilute solution.

Kinematic Viscosity: Ratio of absolute viscosity to density.

Markovian Process: A statistical process where the instantaneous movements are independent, for both length and direction, of those having occurred at earlier times. Describes particle motions, such as a random walk or Brownian diffusion.

Mode-Coupling Theory: Recently developed theory to describe the viscosity and diffusion in liquid and glassy systems at densities beyond those accessible to the Boltzmann and Enskog treatments.

Mutual Diffusion: Diffusion process caused by the gradient of the concentration of a certain species. Same as transport diffusion.

Navier–Stokes Equation: Fundamental hydrodynamic equation relating the bulk and shear viscosity to the pressure and velocity gradients in the fluid. It leads to several important relations: Stokes–Einstein equation, Einstein equation, Stokes equation, etc.

Reptation Model: Theory describing the migration of entangled polymer chains. Assumes that the lateral freedom of chain movement is restricted within a tube. Since longitudinal freedom of movement is unrestricted, polymer chain propagates in a snakelike movement along its own backbone.

Reynolds Number: Dimensionless ratio that describes the relative importance of the nonlinear inertial term to the shear, viscous-damping term in the Navier–Stokes equation. $Re \ll 1$ for laminar, or creeping, flow situations, $Re \gg 1$ for turbulent motions.

Rotational Diffusion: Description of the random motions of the orientation of molecules in solution.

Rough Hard-Sphere Model: Simple extension of the (spherical) hard-sphere model to account for the coupling of rotational and translational diffusion in asymmetric molecular systems.

Rouse Model: Theory describing diffusional behavior of unentangled polymer chains. Approximates polymer chain to a bead–spring model and reduces dynamic behavior of the bead–spring system to a superposition of relaxation modes having certain relaxation times.

Self-Diffusion: The random motion of individual species—atoms, molecules, colloids, polymers—caused by their thermal energy.

Shear Viscosity: Coefficient relating the shear stress to shear rate of a fluid. Contributions due to bulk viscosity are typically negligible; hence shear viscosity is usually equated with viscosity itself.

Specific Viscosity: Normalized solution viscosity utilized in polymer science. Defined as the ratio of the solution viscosity to the solvent viscosity minus unity.

Stokes Equation: Description of the drag on a sphere as it moves through a fluid. Assumes low-velocity flow, i.e., small Reynolds number.

Stokes–Einstein Equation: Relation between self-diffusion of a sphere in a fluid and the viscosity.

Structural Relaxation: Relaxation of the correlated structure between neighboring particles; found in both molecular and colloidal liquids. Strong influence on viscosity of systems.

Transport Diffusion: See **Mutual Diffusion.**

Viscoelasticity: Term describing the shear-rate–dependent viscous and elastic response of a material. Typical feature of entangled polymers and colloidal systems.

VTF Equation (Vogel–Tamman–Fulcher): Empirical equation relating the temperature-induced viscosity change to three material-dependent coefficients. Two of these coefficients find wide application in interpreting glassy state behavior.

WLF Equation (Williams–Landel–Ferry): Empirical equation relating the temperature-induced viscosity change to two material-dependent coefficients. This equation finds wide application in polymer science, where it plays an important role in describing viscoelastic behavior.

Works Cited

Adam, G., Gibbs, J. H. (1965), *J. Chem. Phys.* **43,** 139–146.

Alder, B. J., Gass, D. H., Wainwright, T. E. (1970), *J. Chem. Phys.* **53,** 3813–3826.

Alder, B. J., Wainwright, T. E. (1970), *Phys. Rev. A* **1,** 18–21.

Alms, G. R., Bauer, D. R., Brauman, J. I., Pecora, R. (1973), *J. Chem. Phys.* **59,** 5310–5320.

Angell, C. A. (1991), *J. Non-Cryst. Solids* **131–133,** 13–31.

Antonietti, M., Coutandin, J., Grütter, R., Sillescu, H. (1984), *Macromolecules* **17,** 798–802.

Appel, M., Fleischer, G. (1993), *Macromolecules* **26,** 5520–5525.

Appel, M., Fleischer, G., Kärger, J., Chang, I., Fujara, F. (1994), *Macromolecules* **27,** 4274–4277.

Batchelor, G. K. (1967), *An Introduction to Fluid Dynamics,* London: Cambridge Univ. Press.

Batchelor, A. J. (1976), *J. Fluid Mech.* **74,** 1–29.

Bauer, D. R., Bauman, J. I., Pecora, R. (1974), *J. Am. Chem. Soc.* **96,** 6840–6843.

Berne, B. J., Pecora, R. (1976), *Dynamic Light Scattering*, New York: Wiley.

Berry, G. C., Fox, T. G. (1968), *Adv. Polymer. Sci.* **5,** 261–357.

Bird, R. B., Stewart, W. E., Lightfoot, E. N. (1960), *Transport Phenomena*, New York: Wiley.

Böhmer, R., Nagi, K. L., Angel, C. A., Plazek, D. J. (1993), *J. Chem. Phys.* **99,** 4201–4209.

Brady, J. F. (1993), *J. Chem. Phys.* **99,** 567–581.

Carnahan, N. F., Starling, K. E. (1969), *J. Chem. Phys.* **51,** 635–636.

Chandler, D. (1975), *J. Chem. Phys.* **62,** 1358–1363.

Chang, I., Geil, B., Heuberger, G., Mangel, T., Sillescu, H. (1994), *J. Non-Cryst. Solids* **172–174,** 248–255.

Chapman, S., Cowling, T. G. (1939), *The Mathematical Theory of Non-Uniform Gases*, Cambridge, U.K.: Cambridge Univ. Press.

Cohen, E. G. D. (1993), *Physica* **A194,** 229–257.

Cohen, M. H., Grest, G. S. (1979), *Phys. Rev. B* **20,** 1077–1098.

Cohen, M. H., Turnbull, D. (1959), *J. Chem. Phys.* **31,** 1164–1169.

Cook, R. L., Herbst, C. A., King, H. E., Jr. (1993), *J. Phys. Chem.* **97,** 2355–2361.

Cook, R. L., King, H. E., Jr., Herbst, C. A., Herschbach, D. R. (1994), *J. Chem. Phys.* **100,** 5178–5189.

Cook, R. L., King, H. E., Jr., Peiffer, D. G. (1992a), *Macromolecules* **25,** 2928–2934.

Cook, R. L., King, H. E., Jr., Peiffer, D. G. (1992b), *Phys. Rev. Lett.* **69,** 3072–3075.

Crank, J. (1975), *The Mathematics of Diffusion*, New York: Oxford Univ. Press.

Cummins, H. Z., Du, W. M., Fuchs, M., Götze, W., Hildebrand, S., Latz, A., Li, G., Tao, N. J. (1993), *Phys. Rev. E* **47,** 4223–4239.

Cummins, H. Z., Li, G., Du, W. M., Hernandez, J. (1994), *Physica* **A204,** 169–201.

Cummins, H. Z., Li, G., Hwang, Y. H., Shen, G. Q., Du, W. M., Hernandez, J., Tao, N. J. (1997), *Z. Phys.* **B103,** 501–519.

de Gennes, P. G. (1971), *J. Chem. Phys.* **55,** 572–579.

de Kruif, C. G., Van Iersel, E. M., Vrij, A., Russel, W. B. (1985), *J. Chem. Phys.* **83,** 4717–4725.

de Schepper, I. M., Van Loef, J. J., Haffmans, A. F. E. M. (1989), *J. Stat. Phys.* **57,** 631–645.

Doi, M., Edwards, S. F. (1986), *The Theory of Polymer Dynamics*, Oxford, U.K.: Oxford Univ. Press.

Doi, M., Edwards, S. I. (1978), *J. Chem. Soc. Faraday Trans. II* **74,** 1789–1801.

Dymond, J. H. (1974a), *J. Chem. Phys.* **60,** 969–973.

Dymond, J. H. (1974b), *Physica* **B75,** 100–114.

Dymond, J. H., Brawn, T. A. (Eds.) (1979), *Symposium on Transport Properties of Fluids and Fluid Mixtures: Their Measurement, Estimation, and Correlation*, National Engineering Laboratory, East Kilbride, Glasgow.

Elias, H.-G. (1971), *Makromoleküle*, Basel/Heidelberg: Hüthig & Wepf.

Evans, D. F., Tominaga, T., Davis, H. T. (1981), *J. Chem. Phys.* **74,** 1298–1305.

Ferry, J. D. (1980), *Viscoelastic Properties of Polymers*, 3rd ed., New York: Wiley.

Fleischer, G., Fujara, F. (1994), in: P. Diehl, E. Fluck, H. Günther, R. Kosfeld, J. Seelig (Eds.), *NMR Basic Principles and Progress*, Vol. 30, Berlin: Springer Verlag.

Geszti, T. (1983), *J. Phys. C: Solid State* **16,** 5805–5814.

Götze, W. (1991), in: J. P. Hansen, D. Levesque, J. Zinn-Justin (Eds.), *Liquids, Freezing and the Glass Transition Part 1*, Amsterdam: Elsevier, Chap. 5.

Götze, W., Sjögren, L. (1992), *Rep. Prog. Phys.* **55,** 241–376.

Graessley, W. W. (1982), *Adv. Polym. Sci.* **47,** 67–117.

Hansen, J.-P., McDonald, I. R. (1986), *Theory of Simple Liquids*, New York: Academic.

Herbst, C. A., Cook, R. L., King, H. E., Jr. (1993), *Nature* **361,** 518–520.

Herman, M. H. (1992), *Macromolecules* **25,** 4925–4930.

Hervet, H., Urbach, W., Rondolez, E. (1978), *J. Chem. Phys.* **68,** 2725–2729.

Hirschfelder, J. O., Curtiss, C. F., Bird, R. B. (1954), *Molecular Theory of Gases and Liquids*, New York: Wiley.

Hoover, W. G., Ree, F. H. (1964), *J. Chem. Phys.* **40,** 939–950.

Hu, C. M., Zwanzig, R. (1974), *J. Chem. Phys.* **60,** 4354–4357.

Jonas, J., Hasha, D., Huang, S. J. (1979), *J. Chem. Phys.* **71,** 3996–4000.

Jost, W., Hauffe, K. (1972), *Diffusion*, Darmstadt: Steinkopff.

Kärger, J., Pfeifer, H., Heink, W. (1988), *Adv. Magn. Res.* **12,** 1–89.

King, H. E., Jr. (1997), in: W. B. Holzapfel, N. S. Isaacs (Eds.), *High-Pressure Techniques in Chemistry and Physics: A Practical Approach*, Oxford, U.K.: Oxford Univ. Press, Chap. 3.

Klein, J. (1979), *J. Contemp. Phys.* **20,** 611–629.

Klein, J., Briscoe, B. J. (1979), *Proc. Roy. Soc. London A.* **365,** 53–73.

Kremer, K., Grest, G. S. (1990), *J. Chem. Phys.* **92,** 5057–5086.

Kremer, K., Grest, G. S. (1995), in: K. Binder (Ed.), *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*, New York: Oxford Univ. Press, pp. 194–271.

Landau, L. D., Lifshitz, E. M. (1959), *Fluid Mechanics*, Vol. 6, New York: Pergamon Press.

Lebedev, Y. S. (1967), *Kinet. Katal.* **8,** 213–218.

Li, G., King, H. E., Jr., Oliver, W. F., Herbst, C. A., Cummins, H. Z. (1995), *Phys. Rev. Lett.* **74,** 2280–2283.

Lust, L., Valls, O. T., Dasgupta, C. (1993), *Phys. Rev. E* **48,** 1787–1798.

Macosko, C. W. (1994), *Rheology: Principles, Measurements, and Applications,* New York: VCH Publishers, Inc.

Mezei, F. (1983), *Physica* **120B,** 51–60.

Mills, P. J., Green, P. F., Palmstrm, J., Mayer, J. W., Kramer, P. J. (1984), *Appl. Phys. Lett.* **45,** 957–959.

Monnery, W. D., Svrcek, W. Y., Mehirotra, A. K. (1995), *Can. J. Chem. Eng.* **73,** 3–40.

Ngai, U. C., Rendell, R. W., Rajagopal, A. U., Teifler, S. (1986), *Ann. NY Acad. Sci.* **484,** 150–184.

Parkhurst, H. J., Jr., Jonas, J. (1975), *J. Chem. Phys.* **63,** 2698–2704.

Phan, S.-E., Russel, W. B., Cheng, Z., Zhu, J., Chaikin, P. M., Dunsmuir, J. H., Ottewill, R. H. (1996), *Phys. Rev. E* **54,** 6633–6645.

Pusey, P. N. (1991), in: J. P. Hansen, D. Levesque, J. Zinn-Justin (Eds.), *Liquids, Freezing and the Glass Transition Part 2,* Amsterdam: Elsevier, Chap. 10.

Russel, W. B., Sperry, P. R. (1994), *Prog. Org. Coat.* **23,** 305–324.

Schweizer, U. S. (1989), *J. Chem. Phys.* **91,** 5802–5839.

Segrè, P. N., Behrend, O. P., Pusey, P. N. (1995a), *Phys. Rev. E* **52,** 5070–5083.

Segrè, P. N., Meeker, S. P., Pusey, P. N., Poon, W. C. K. (1995b), *Phys. Rev. Lett.* **75,** 958–961.

Segrè, P. N., Pusey, P. N. (1996), *Phys. Rev. Lett.* **77,** 771–774.

Sperling, L. H. (1992), *Introduction to Physical Polymer Science,* 2nd ed., New York: Wiley.

Springer, T. (1972), *Quasielastic neutron scattering for the investigation of diffusive motions in solids and liquids,* Springer Tracts in Modern Physics No. 64, Berlin: Springer, pp. 1–100.

Straube, G., Heinrich, G. (1984), *Acta Polym.* **35,** 115–119.

Sun, T., King, H. E., Jr. (1996), *Phys. Rev. E* **54,** 2696–2703.

Tyrrell, H. J. V., Harris, K. R. (1984), *Diffusion and Heat Flow and Liquids,* London: Butterworths.

van Blaaderen, J., Peetermans, J., Maret, G., Dhont, J. K. G. (1992), *J. Chem. Phys.* **96,** 4591–4603.

van Der Gulik, P. S., Mostert, R., Van Den Berg, H. R. (1988), *Physica* **A151,** 153–166.

van Megen, W., Underwood, S. M. (1993), *Phys. Rev. E* **47,** 248–261.

van Megen, W., Underwood, S. M. (1994), *Phys. Rev. E* **49,** 4206–4220.

Woodcock, L. V., Angel, C. A. (1981a), *Adv. Chem. Phys.* **48,** 397–453.

Woodcock, L. V., Angel, C. A. (1981b), *Phys. Rev. Lett.* **47,** 1129–1132.

Young, R. J., Lovell, P. A. (1991), *Introduction to Polymers,* 2nd ed., London: Chapman and Hall.

Further Reading

Batchelor, G. K. (1967), *An Introduction to Fluid Dynamics,* London: Cambridge Univ. Press.

Berne, B. J., Pecora, R. (1976), *Dynamic Light Scattering,* New York: Wiley.

Bird, R. B., Stewart, W. E., Lightfoot, E. N. (1960), *Transport Phenomena,* New York: Wiley.

Crank, J. (1975), *The Mathematics of Diffusion,* New York: Oxford Univ. Press.

Doi, M., Edwards, S. F. (1986), *The Theory of Polymer Dynamics,* Oxford, U.K.: Oxford Univ. Press.

Ferry, J. D. (1980), *Viscoelastic Properties of Polymers,* 3rd ed., New York: Wiley.

Hansen, J.-P., McDonald, I. R. (1986), *Theory of Simple Liquids,* New York: Academic.

Hildebrand, J. H. (1977), *Viscosity and Diffusion. A Predictive Treatment,* New York: Wiley.

Hirschfelder, J. O., Curtiss, C. F., Bird, R. B. (1954), *Molecular Theory of Gases and Liquids,* New York: Wiley.

Jost, W., Hauffe, K. (1972), *Diffusion,* Darmstadt: Steinkopff.

Kirkaldy, J. S., Young, D. S. (1985), *Diffusion in the Condensed State,* London: The Institute of Metals.

Macosko, C. W. (1994), *Rheology: Principles, Measurements, and Applications,* New York: VCH Publishers, Inc.

Viswanath, D. S., Natarajan, G. (1989), *Data Book on the Viscosity of Liquids,* New York: Hemisphere Publishing Corporation.

Young, R. J., Lovell, P. A. (1991), *Introduction to Polymers,* 2nd ed., London: Chapman and Hall.

10
Ultraviolet and Visible Light Spectrometers

John L. Hardwick

10.1 Introduction 321
10.2 Design Considerations 322
10.2.1 Luminosity or Étendue 322
10.2.2 Resolution 323
10.2.3 Dispersion 323
10.2.4 Efficiency 323
10.3 Elements of a Generic Spectrometer 323
10.3.1 Entrance Aperture 323
10.3.2 Collimating Optics 324
10.3.3 Wavelength Selecting Element 324
10.3.4 Focusing Optics 324
10.3.5 Exit Aperture 325
10.3.6 Detector 325
10.4 Prism Spectrometers 325
10.4.1 Dispersion 325
10.4.2 Resolution 326
10.4.3 Luminosity 326
10.4.4 Specialized Designs 326
10.4.5 Aberrations 328
10.4.5.1 Spherical Aberration 328
10.4.5.2 Coma 329
10.4.5.3 Astigmatism 329
10.4.5.4 Curvature of Field 330
10.4.5.5 Chromatic Aberration 330
10.5 Grating Spectrometers 330
10.5.1 Dispersion 330
10.5.2 Resolution 331
10.5.3 Luminosity 331
10.5.4 Design and Manufacture of Gratings 332
10.5.4.1 Transmission Grating 332
10.5.4.2 Reflection Grating 332

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

10.5.4.3 Echelles 333
10.5.4.4 Holographic Gratings 333
10.5.5 Plane Grating Spectrometers 334
10.5.5.1 Aberrations 334
10.5.5.2 Types of Plane Grating Spectrometers 336
10.5.6 Concave Grating Spectrometers 338
10.5.6.1 Optical Aberrations of Concave Mirrors 338
10.5.6.2 Rowland Circle Mounts 339
10.5.6.3 Non-Rowland Circle Mounts 341
10.6 Interferometric Spectrometers 342
10.6.1 Michelson 342
10.6.2 Fabry–Perot 343
10.7 Hybrid and Compound Instruments and Imaging Filters 345
10.7.1 Order-Sorter for a Grating Spectrometer 345
10.7.2 Harrison Echelle 345
10.7.3 Fabry–Perot/Grating Instrument 346
10.7.4 Double and Triple Monochromators 346
10.7.5 Imaging Filters 347
10.7.6 The Sisam Spectrometer 347
10.8 Detectors 347
10.8.1 Ideal Detector 347
10.8.2 Types of Detectors 348
Glossary 351
References 351
Further Reading 352

10.1
Introduction

The visible part of the electromagnetic spectrum extends from about 400 to 750 nm. At longer wavelengths, the retina loses sensitivity, whereas at shorter wavelengths, the lens of the eye absorbs most of the incident radiation. The ultraviolet spectrum extends from 400 to 180 nm. At wavelengths shorter than 180 nm atmospheric oxygen strongly absorbs light. Wavelengths between 180 and about 1 nm are classified as vacuum ultraviolet; wavelengths shorter than this are classified as X-rays.

A spectrometer is an instrument for discriminating among different wavelengths of light, producing a record of intensity as a function of wavelength or frequency from a particular source. Spectrometers are commonly used as tools in chemical analysis, atomic and molecular physics, astronomical measurements, and shaping and filtering optical radiation in both the time and the frequency domains.

Dispersion of light by a prism was reported by Johannes Marcus Marci in 1647. In 1666, Isaac Newton used a prism to disperse a beam of sunlight onto the wall of a darkened room; by using a lens to focus the light from the prism, he was able to disperse the visible spectrum over a length of about 25 cm. The first observations of spectral lines occurred 140 years later, when W. H. Wollaston and, independently, Joseph Fraunhofer restricted the entrance aperture of Newton's basic device with an entrance slit, allowing images of that slit to be formed as single separate colors. This innovation allowed the first observation of the ''Fraunhofer lines'' of the solar spectrum. Fraunhofer cataloged several hundred of these lines for use as in defining colors of the visible spectrum so that the refractive index of glasses at those colors could be more carefully defined. Fraunhofer also ruled the first diffraction grating and using it, he measured for the first time the wavelength of the yellow light from the sodium D lines. In addition, he was the first to observe stellar and planetary spectra using a prism placed in front of an astronomical telescope.

Forty years later, in 1859, Gustav Kirchhoff and Robert Bunsen developed the first practical spectroscope and showed that it could be used for chemical analysis. They demonstrated that atoms have characteristic absorption and emission spectra and recognized that the absorption and emission lines were identical in

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

wavelength for a given atom; the appearance of a spectral line in absorption or emission depended only on the temperature of the sample. These developments mark the origin of optical spectroscopy as a branch of both analytical chemistry and chemical physics.

In 1862, G. G. Stokes used a quartz prism as a dispersing element and a fluorescent screen as a detector to make the first observations of ultraviolet atomic emission spectra. Although Stokes was unable to make wavelength measurements, he was able to observe spectra out to 186 nm. Extension of spectroscopic measurements into the vacuum ultraviolet was made by Victor Schumann in 1893. Using a fluorite prism and an evacuated spectrograph, Schumann was able to extend ultraviolet measurements to 120 nm, where absorption by the fluorite prism limited its useful range.

The production of diffraction gratings was greatly advanced in the late nineteenth century by the work of Henry Rowland on the development of an accurate ruling engine. Rowland also invented the concave grating – a grating ruled on a spherical mirror blank, which greatly simplified the construction of spectrographs. Using these gratings, Rowland was able to make accurate measurements of wavelengths of the ultraviolet spectrum down to 210 nm. In 1906, Theodore Lyman used such a concave grating to extend measurements in the vacuum ultraviolet down to 50 nm.

Theoretical understanding of atomic and molecular spectra had to await the rise of quantum mechanics in the early twentieth century. Beginning with the work of Niels Bohr, it became clear that spectra in the visible and ultraviolet regions depended on the internal electronic structure of atoms and molecules, and that the absorption and emission of radiation in this region corresponded to transitions of electrons between allowed energy levels.

10.2
Design Considerations

There are two figures of merit for any spectrometer, and generally they are mutually exclusive. They are *resolution*, the ability to distinguish between wavelengths of light that are nearly the same, and *luminosity*, the ability to accept light from a source and transmit all of it to the detector. Because improvement of one of these properties is generally at the expense of the other, a real spectrometer is always a compromise between the two.

10.2.1
Luminosity or Étendue

The light gathering power of a spectrometer is a function of both the area of the entrance aperture and the solid angle with which the instrument can accept incident light. The product of these two quantities is referred to as the *étendue of the instrument*. Terms that are often used synonymously are "throughput" and "luminosity." The étendue is defined as $L = a\Omega$, here a is the area of the entrance aperture and Ω is the solid angle subtended by the aperture stop of the instrument at the entrance aperture. For an optical system without losses (e.g., from scattering or absorption), L remains a constant through the system.

10.2.2
Resolution

The resolution of an instrument is the ability to discriminate between two adjacent wavelengths (or frequencies) of light. Resolution is most easily understood in the case of a monochromator, which transmits a certain range of wavelengths and rejects all others. The minimum attainable bandwidth of the instrument is its resolution, $\Delta\lambda$. A closely related quantity is the resolving power, R, defined as

$$R = \frac{\lambda}{\Delta\lambda} \qquad (10.1)$$

In a dispersive instrument, the maximum possible resolution is most conveniently calculated as the product of the reciprocal angular dispersion $d\lambda/d\theta$ and the diffraction-limited angular resolution $\Delta\theta$ from the aperture stop imposed by the dispersing element; the angle θ is the deviation of the dispersed light. For a nondispersing spectrometer, the resolution in wavenumber units is the reciprocal of the difference in optical retardation of rays.

The resolution desired depends on the experiment to be performed. Ideally, the resolution of the spectrometer will be matched to the line width of the sample to be examined. For solids or liquids at moderate temperatures, absorption lines have a width of several nanometers or more, and the resolution requirements on the spectrometer are modest. For gases, the width of an absorption or emission line is usually limited by the Doppler width, which is typically 2×10^{-6} times the wavelength. A high-resolution spectrometer is therefore defined as one whose resolving power is Doppler limited, that is, $R \geq 5 \cdot 10^5$.

10.2.3
Dispersion

The angular dispersion of a grating or prism is the change of the displacement angle with wavelength, $d\theta/d\lambda$. The linear dispersion of a spectrometer is the change of the distance along the focal plane with wavelength $dl/d\lambda$; this will determine the physical separation of two spectral lines at the slit or detector. A more frequently used quantity is the reciprocal linear dispersion, $d\lambda/dl$, also called the "plate factor."

10.2.4
Efficiency

The product of the resolving power and the étendue (the "resolution-luminosity product") is sometimes referred to as the *efficiency of the spectrometer*. It is a property of the instrument in question and, sometimes, the wavelength of operation. For a grating spectrometer operating at a given wavelength, for example, the efficiency is fixed, but it is possible within limits to trade resolution for throughput by opening the entrance and exit slits.

10.3
Elements of a Generic Spectrometer

There are several elements which are common to most spectrometers, and these are illustrated in Figure 10.1.

10.3.1
Entrance Aperture

The *entrance aperture* is used to define the source of light to be analyzed. In a prism or grating spectrometer, it is ordinarily a

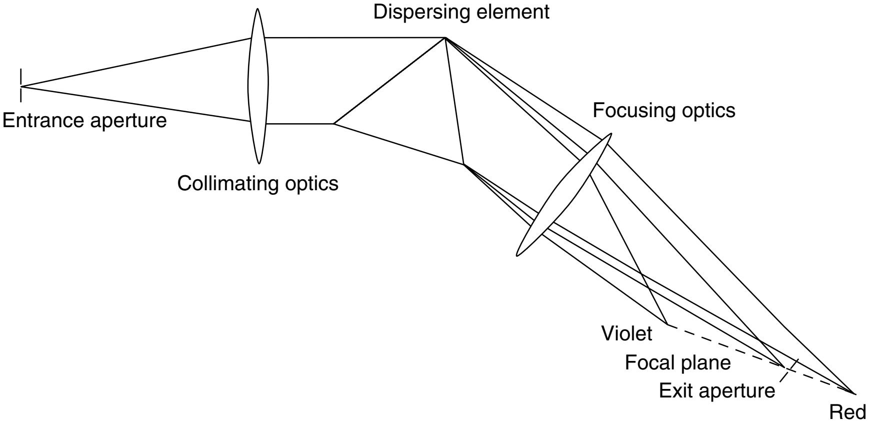

Fig. 10.1 A minimum deviation prism spectrometer, illustrating the elements common to most spectrometers.

slit, although a more complicated shape is sometimes used to enhance throughput. In an interferometric spectrometer, the entrance aperture is usually circular.

10.3.2
Collimating Optics

Collimating optics take the rays of light from the entrance aperture, which are typically diverging, and make the rays collinear. This is generally desirable so that all of the rays strike the dispersing element at the same angle; if they do not, often either the resolution is degraded, or the focus of the instrument changes with wavelength, or both. The collimating optics may be either refractive or reflective (i.e., either lenses or concave mirrors), although mirrors are generally used in modern instruments because of their lack of chromatic aberration.

10.3.3
Wavelength Selecting Element

The kind of wavelength selector in a spectrometer is usually classified as dispersive or nondispersive. Dispersive elements separate the wavelengths along some spatial dimension, while nondispersive elements pass some wavelengths and reject others without altering the optical path of the transmitted light. The most commonly used dispersing elements are the prism and the diffraction grating, although some interferometers, such as the Fabry–Perot, are also used in a dispersing mode. The most versatile nondispersing spectrometer is the Michelson interferometer, which is the basis of the modern Fourier transform spectrometer. Other examples of nondispersive elements are filters, which are used in simple colorimeters or in imaging spectrometers.

10.3.4
Focusing Optics

The focusing or condensing optics (in spectrographs often called the *camera* optics) take the collimated light passed by the wavelength selecting element and focus it onto an exit aperture (or several apertures) or a detector.

Together with the collimating optics, the focusing optics form an image of the entrance aperture at the exit aperture. The usual goal of spectrometer design is to make the collimating and focusing optics sufficiently aberration-free that the performance of the spectrometer will be limited only by the wavelength selecting element. As with the collimating optics, either lenses or mirrors may be used, with the preference being given to mirrors for most instruments.

10.3.5
Exit Aperture

The exit aperture must be matched to the image of the entrance aperture in such a way that it provides maximum discrimination of the wavelengths of interest while providing maximum allowable throughput of light. In a conventional grating or prism monochromator, for example, the exit aperture is a single exit slit in the shape of the image of the entrance slit.

10.3.6
Detector

The detector transforms the light into either a permanent image or an electronic signal. In addition to the requirement that it must be sensitive in the wavelength region of interest, the detector is chosen so as to minimize unwanted noise and provide a reproducible (ideally linear) response. Spectrometers are often distinguished from one another on the basis of the method of detection: an instrument which allows a spectrum to be observed visually is a *spectroscope*, one which makes a photographic record is called a *spectrograph*, and the term *spectrometer* refers to an instrument which records the spectrum photoelectrically. If photometric accuracy is of paramount importance, the device is called a *spectrophotometer*. In this chapter, the term *spectrometer* is used as an inclusive term covering all these types of instrument.

10.4
Prism Spectrometers

The earliest spectrometers were prism spectrometers, and the most commonly used grating spectrometer designs have their origins in prism spectrometers. Prism spectrometers, although they have generally poor optical performance in comparison with most modern designs, have the advantages of being inexpensive, rugged, easy to align, and technically undemanding.

10.4.1
Dispersion

Light incident on a prism with an angle α (measured normal to the face of the prism, as shown in Figure 10.2) will be refracted according to Snell's law, $\sin\alpha/\sin\alpha' = n_{\text{prism}}/n_{\text{air}}$, where n_{prism} and n_{air} are the refractive indices of air and the prism material at the wavelength of the incident light. On exiting the prism, the light will be refracted again by an angle β, for a total deviation θ of the beam of light. The deviation θ of the beam depends on the refractive index n of the prism, while the dispersion of the beam, $d\theta/d\lambda$, depends on the *change* of refractive index of the prism material with wavelength, $dn/d\lambda$. The angular dispersion of a prism is defined as the change of the angle of deviation θ with

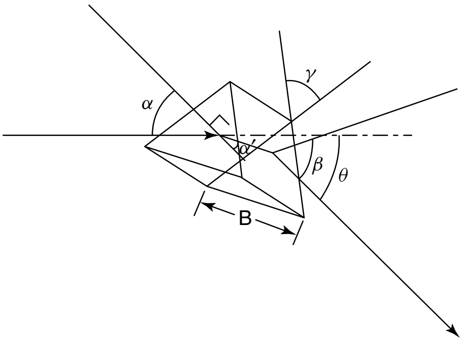

Fig. 10.2 Coordinates used in describing the dispersion of a prism.

wavelength and is given by the formula

$$\frac{d\theta}{d\lambda} = \frac{d\theta}{dn}\frac{dn}{d\lambda} \tag{10.2}$$

The first factor may be evaluated by differentiating Snell's law, while the second factor is a property of the material of the prism. If the angle of incidence α is nearly equal to the angle of departure and the refractive index of air is taken as 1, $d\theta/dn$ is given in terms of the apex angle γ of the prism as

$$\frac{d\theta}{dn} = \frac{2\sin(\gamma/2)}{\sqrt{1 - n^2\sin^2(\gamma/2)}} \tag{10.3}$$

The linear dispersion of the spectrometer $dl/d\lambda = fd\theta/d\lambda$ is the angular dispersion times the focal length of the condensing optics. Note, however, that the focal length of the condensing optics can often change considerably over the spectrum as a result of the chromatic aberration of the condensing lens.

10.4.2
Resolution

The angular resolution of a prism spectrometer, in the absence of optical aberrations of the lenses, will be determined only by the prism itself. It may be written in terms of the reciprocal angular dispersion as

$$\Delta\lambda = \frac{d\lambda}{d\theta}\Delta\theta \tag{10.4}$$

where $\Delta\theta$ is the diffraction limit imposed by the wavelength of light and the aperture of the prism itself. The resolving power for a prism at minimum deviation (Figure 10.2) is given in terms of the base B as (Sawyer, 1963)

$$R = B\left|\frac{dn}{d\lambda}\right| \tag{10.5}$$

10.4.3
Luminosity

The effective area of the aperture stop of a prism instrument is the product of the area of the prism face and the cosine of the angle of incidence. The solid angle of acceptance of the instrument will be this area divided by the square of the focal length of the collimating lens. The product of this acceptance angle and the area of the entrance slit will constitute the luminosity.

10.4.4
Specialized Designs

The design of a prism spectrometer depends on whether a large flat focal

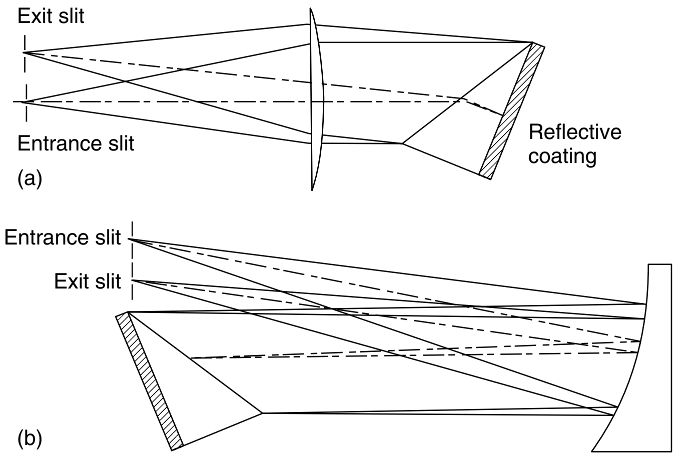

Fig. 10.3 Littrow mount prism spectrometers, (a) using lenses for imaging and (b) using an off-axis parabolic mirror for imaging. The entrance and exit slits would ordinarily be placed as close together as possible.

plane is most desirable (in the case of a spectrograph) or whether it is deemed necessary to scan the wavelength easily by rotating the prism (as in the case of a spectroscope or a spectrometer).

To minimize reflection losses, the total deviation of the light as it passes through the prism is kept to a minimum. This consideration is in conflict with the requirement that the base of the prism must be large to maximize both dispersion and resolution, and usually a 60° angle for the apex of the prism is chosen as a compromise between these two requirements. A typical minimum deviation spectrograph is illustrated in Figure 10.1. Light is collimated and focused by lenses in this design. If the lenses are of the same refractive index as the prism, then the focal surface is, to a first approximation, a plane that passes through the entrance slit (James and Sternberg, 1969).

A more compact design is the Littrow configuration, as shown in Figure 10.3(a). Here the prism is replaced by a prism having half the original apex angle whose rear surface is coated with a reflective coating. Light passes forward and then backward through the prism, and a single lens serves as both the collimating and focusing optic. The entrance and exit slits may be offset horizontally or, more frequently, vertically. This arrangement lends itself to scanning by simply rotating the prism, or an image of a spectrum may be formed as in the minimum deviation instrument. A variation which employs reflective optics is also used (Figure 10.3b). The mirror used to image the slit is chosen to be an off-axis paraboloid to minimize the astigmatism and spherical aberrations associated with a spherical mirror used off-axis. There are so many variations on the Littrow design that any spectrometer which disperses light back along the original path is often referred to as a *Littrow mount*.

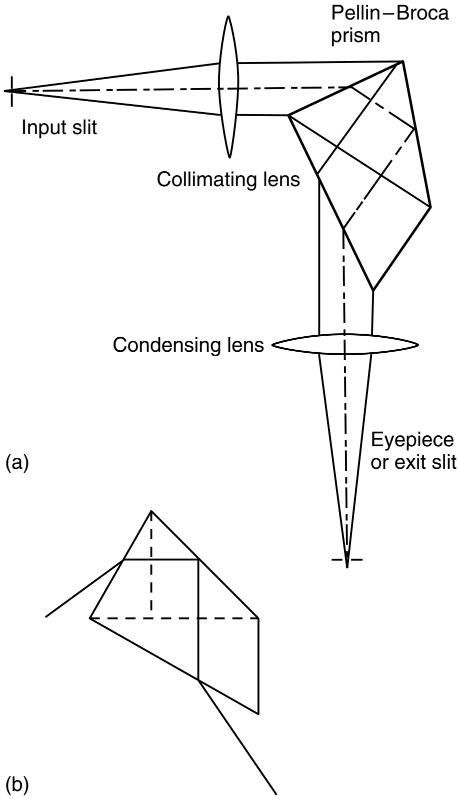

Fig. 10.4 (a) A Pellin–Broca spectrometer. Light is refracted at a constant 90° angle as the prism is rotated. (b) A Pellin–Broca prism, illustrating its conception as a composite of three separate prisms.

One of many *constant deviation* mounts of a prism is the Pellin–Broca spectrometer, illustrated in Figure 10.4(a). This design has the advantage that the prism may be rotated in such a way that the minimum deviation beam is always refracted at right angles to the incident beam. As indicated in Figure 10.4(b), the Pellin–Broca prism may be viewed as a pair of 30–60–90° prisms joined by a 45° total internal reflection prism.

10.4.5
Aberrations

The aberrations of the imaging optics that limit the resolution of a prism spectrometer include spherical aberration, coma, astigmatism, curvature of field, and, in some designs, chromatic aberration.

10.4.5.1 Spherical Aberration

In a spherical lens, the rays passing through the center of the lens do not focus in exactly the same point as the rays, which pass through the outer portion of the lens, as illustrated in Figure 10.5. The rays from the various parts of the lens converge to a minimum spot size, the *circle of least confusion*, which determines the resolution limit of a spherically figured lens. To a good approximation, a pair of plano-convex lenses used as collimating and condensing optics will

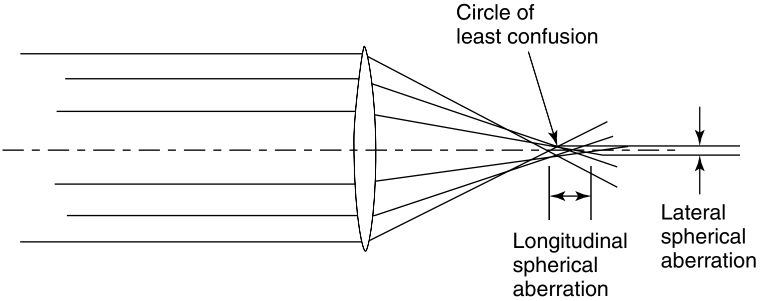
Circle of least confusion
Longitudinal spherical aberration
Lateral spherical aberration

Fig. 10.5 Spherical aberration in a lens.

provide the best form for reducing spherical aberration, with the plane side of the lens facing the image in each case.

10.4.5.2 **Coma**

Coma is an aberration, which affects only off-axis rays. As illustrated by Figure 10.6, coma is a result of rays passing through different annular rings of the lens being imaged, not into points, but into small circles of different size and located at different centers. The change in size with the changing focal point results in a *comatic flare* of the image resembling the shape of a comet tail, which is the origin of the term. Coma increases rapidly with the off-axis angle, severely restricting the size of the focal plane for wide-aperture spectrographs.

10.4.5.3 **Astigmatism**

Similar to coma, astigmatism is an off-axis aberration. If the rays striking the lens are off-axis in the vertical direction, the rays in the vertical plane will form a focus at a different point than those in the horizontal plane. The former will be focused into a line tangent to a circle about the principal optic axis (the tangential focus) and the latter into a line pointed toward the optic axis (the saggital focus). Astigmatism, which spreads a point source into a line parallel to the slit, will not necessarily

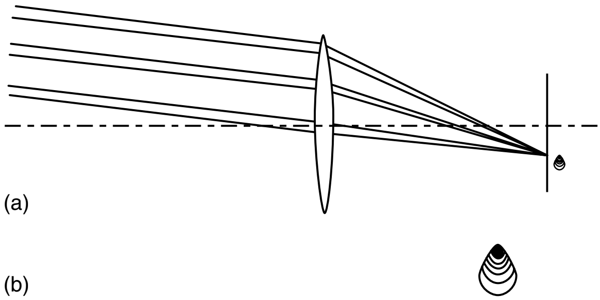

Fig. 10.6 Coma produced by a simple lens. (a) Off-axis marginal rays are imaged into different positions than the central ray. (b) Annular sections of the lens are imaged into circles whose size and position depend on the distance of the annuli from the center of the lens.

degrade the resolution, although it will reduce the brightness of the image; astigmatism, which spreads a point source perpendicular to the direction of the slit, will degrade both the throughput and the resolution.

10.4.5.4 **Curvature of Field**

The focal surface of an optical system is not usually planar, but instead is a curved surface. Whether this curvature is a problem depends in part on whether the curvature is fixed or changes according to the setting of the prism, and in part on whether the instrument is intended for use with an exit slit, an array detector, or a photographic plate. If the curvature changes with the prism angle, as it may with a minimum deviation mount, the prism cannot be rotated without defocusing the instrument, and the instrument is suitable only as a fixed prism spectrograph. If, on the other hand, the curvature and focus remain constant with the prism angle, the prism can be rotated to scan the spectrometer.

10.4.5.5 **Chromatic Aberration**

The high dispersion of a glass which makes it suitable for use as a prism means that lenses made from the same material will have different focal lengths at different wavelengths. If the spectrometer is to be used with photographic detection, as in the minimum deviation spectrograph of Figure 10.1, this is not a problem, because the plateholder can be adjusted once for optimum focus and fixed in that position. For a constant deviation mount such as the Pellin–Broca mount of Figure 10.4, both the collimating and the focusing lenses must be corrected for chromatic aberration if the spectrometer is to remain in focus as the prism is rotated.

10.5
Grating Spectrometers

Most modern ultraviolet and visible spectrometers are based on the diffraction grating, which offers improved resolution, luminosity, and dispersion compared with prism instruments of comparable size and expense. These may be classified as plane grating spectrometers, of which the Czerny–Turner spectrometer is an example, and concave grating spectrometers, which include the Eagle and Seya–Namioka mounts. The concave grating design is usually preferred whenever the number of optical surfaces must be minimized, as in the vacuum ultraviolet region, because the concave grating serves as both the dispersing element and the imaging optics. Plane grating instruments must rely on concave mirrors to image the spectrum; these mirrors lead to absorption losses at each optical surface, but the additional design freedom afforded by the imaging optics allow far greater control over aberration corrections than is possible for concave gratings. Accordingly, plane grating instruments are used whenever absorption losses at the mirrors are negligible and a high quality stigmatic image is desirable, as in the visible and near infrared.

10.5.1
Dispersion

Diffraction of light by a grating is determined by the grating equation:

$$m\lambda = a(\sin\alpha + \sin\beta) \tag{10.6}$$

where m, the order of diffraction, must be an integer; λ is the diffracted wavelength; α is the angle of incidence of the light with respect to the normal; and β is the angle of diffraction with respect to the normal. The reciprocal angular dispersion, $d\lambda/d\beta$, is given as $a \cos \beta/m$; the tuning of wavelength as the grating is turned is twice this number. The linear dispersion at the focus of the spectrometer is the angular dispersion times the focal length of the condensing optics.

10.5.2
Resolution

As with a prism, the resolution of a diffraction grating is the product of the dispersion and the diffraction-limited angular resolution imposed by the aperture of the grating. Unlike prism instruments, however, the dispersion is independent of the material of the grating, and so the resolution of a grating may be stated in several rather simple ways. Taking the ruled width of the grating as W and the angular resolution of light leaving the grating as $\Delta\beta = \lambda/(W \cos \beta)$, the resolution of the grating is found to be

$$\begin{aligned} \Delta\lambda &= \frac{d\lambda}{d\beta}\Delta\beta \\ &= \frac{a}{m}\cos\beta \cdot \frac{\lambda}{W\cos\beta} \qquad (10.7) \\ &= \frac{a}{m} \cdot \frac{\lambda}{W} \end{aligned}$$

The resolving power, in turn, is given by

$$\begin{aligned} R &= \frac{\lambda}{\Delta\lambda} \qquad (10.8) \\ &= m\frac{W}{a} \end{aligned}$$

which is just the order of diffraction times the number of grooves in the grating. Clearly, for a grating spectrometer to achieve its maximum resolving power, the entire width of the grating must be illuminated.

If $\alpha \approx \beta$, the resolution can be rewritten in terms of the wave number $\tilde{\nu}$ as

$$\begin{aligned} \Delta\tilde{\nu} &= \frac{\partial\tilde{\nu}}{\partial\lambda}\Delta\lambda \qquad (10.9) \\ &= \frac{1}{W \cdot 2(\sin\alpha)} \end{aligned}$$

which is just the reciprocal of the optical path difference of rays on opposite sides of the grating.

10.5.3
Luminosity

The solid angle of acceptance of the grating, Ω, is given to a good approximation by the equation

$$\Omega = \frac{A\cos(\alpha)}{f^2} \qquad (10.10)$$

where A is the area of the grating, α is the angle of incidence, and f is the focal length of the collimating optics. The luminosity of a grating spectrometer is thus equal to the product of Ω and the area a of the entrance slit, or

$$L = \frac{Aa\cos(\alpha)}{f^2} \qquad (10.11)$$

The "speed" of a spectrometer is often quoted in terms of its focal ratio, $F = f/d$, where d is the maximum diameter of the aperture stop dictated by the size of the grating. Thus, a spectrometer with a focal length of 1 m and a grating whose diagonal measures 12.5 cm will have a speed of $f/8$. This number obviously does not accurately measure the luminosity of the instrument as it is ordinarily used; it does, however, specify the minimum aperture of any external optics that are guaranteed to illuminate the entire surface of the grating. The luminosity of a spectrometer varies as the inverse square of the focal ratio.

10.5.4
Design and Manufacture of Gratings

Until recently, all gratings were ruled mechanically by inscribing parallel grooves on an optically polished master blank. The tolerances involved in the manufacture of such gratings generally require interferometric control of the ruling engine to compensate for mechanical errors in the lead screw; typically, modern ruling engines use a frequency stabilized laser for this purpose. Eliminating periodic errors in the rulings is extremely important, because periodic errors will result in diffraction at angles other than the design angle. The grating "ghosts" produced in this way can cause lines to appear at apparent wavelengths far removed from the true wavelength, creating problems ranging from nuisance to disaster depending on the experiment being performed.

More recently, gratings have been produced by holographically recording the groove pattern on a photosensitive surface. The main advantage of a holographic grating is the absence of the inevitable periodic errors that are found in mechanically ruled gratings. The principal advantage claimed for ruled gratings is a greater control over the shape of the groove and, consequently, a more sharply defined blaze angle and higher efficiency. These issues have been addressed in some detail theoretically (Loewen *et al.*, 1977). Gratings for use in the visible and ultraviolet are available with groove spacings from 300 grooves/mm to 4800 grooves/mm.

10.5.4.1 Transmission Grating

Conceptually, the simplest grating is the transmission grating. Transmission gratings may be fabricated either by imprinting a transparent substrate with closely spaced lines or by ruling a transparent substrate with shaped grooves. As schematically shown in Figure 10.7, the transmission grating is simply an extension of multiple slit interference to a very large number of slits. If two waves emerging from adjacent slits have a total path length which differs by an integral number of waves, constructive interference will occur and the intensity will be nonzero. For an angle of incidence α with respect to the grating normal and an angle of diffraction β, this condition requires $m\lambda = a(\sin\alpha + \sin\beta)$. This equation is always satisfied for $\alpha = -\beta$; therefore, transmission straight through the grating is allowed with $m = 0$ for all wavelengths. In addition, a wavelength may be diffracted at angles corresponding to nonzero integral values of m.

10.5.4.2 Reflection Grating

A more commonly used type of ruled grating is the *reflection grating*, which is typically ruled on a polished glass optical flat coated with a thin film of aluminum. Once again, the condition for constructive interference is that $m\lambda = a(\sin\alpha + \sin\beta)$. The reflection grating is preferred over the transmission grating for much the same reasons that mirrors

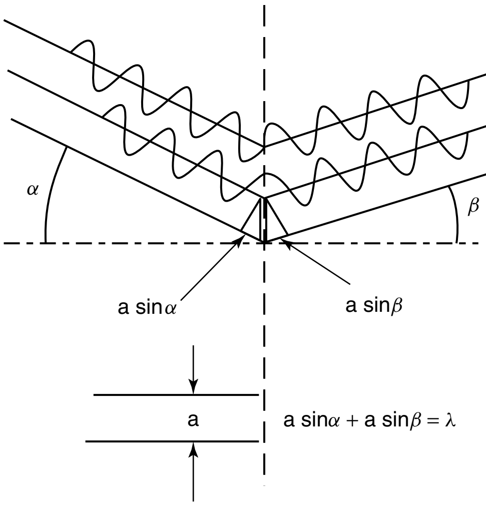

Fig. 10.7 Multiple-slit interference as observed in a segment of a transmission grating. If the waves are out of phase by an integral number of wavelengths, constructive interference will occur and intensity will be maximum.

are preferred over lenses as imaging optics: low absorption and scattering losses, high efficiency in the ultraviolet, and insensitivity to the homogeneity of the substrate.

By controlling the shape and angle of the diamond ruling tool, it is possible to control the shape of the grooves in such a way that light striking the grating normal to the face of the groves is diffracted with very high efficiency. When a grating is viewed at this angle, the grating appears to blaze with color, and this angle of highest efficiency is referred to as the *blaze angle* of the grating.

Ruled gratings are not ordinarily sold as originals, but are instead used as masters to produce replica gratings. This process allows a single ruling to be duplicated inexpensively, greatly reducing the cost of diffraction gratings in commercial spectrometers.

10.5.4.3 **Echelles**

If the grooves of a grating are ruled so that the sides are flat, the grating will have a sharply defined angle of maximum efficiency where the angle of diffraction coincides with the specular reflection from the groove facets. Such a grating is called an *echelle* or an *echelette*. It is ruled very coarsely so that the shape of the grooves can be controlled accurately, and therefore it is used in a high order of diffraction (typically $m = 100$). The free spectral range of such a grating is relatively small, so additional dispersing optics must be used to separate overlapping orders.

10.5.4.4 **Holographic Gratings**

Gratings may also be produced optically, without the use of a ruling engine. In this process, a pair of light beams is used to produce a pattern of interference fringes at a surface that has been coated with photoresist. Although the earliest such gratings predate the invention of the laser, the process did not become practical before the availability of high powered ion lasers (Hutley, 1982). Such gratings are generally referred to as *holographic gratings*.

Two intersecting coherent beams of light will produce interference fringes within their volume of intersection with a spacing D given by

$$D = \frac{\lambda}{2\sin(\theta/2)} \tag{10.12}$$

where θ is the angle of intersection. If a grating blank is placed at an angle ϕ to the bisector of θ, fringes will be produced with a spacing a of

$$a = \frac{\lambda}{2\sin\theta\sin\phi} \tag{10.13}$$

The blank, previously coated with a thin layer of photoresist, is subsequently developed to produce a corrugated surface with the spacing of the interference fringes. This surface is generally coated with a reflective metal layer and used as a reflection grating.

10.5.5
Plane Grating Spectrometers

10.5.5.1 Aberrations

Plane grating spectrometers almost invariably use spherical concave mirrors or, less frequently, paraboloids to image the entrance slit into a spectrum. Similar to lenses, the mirrors that form images in a typical grating spectrometer are subject to spherical aberration, coma, astigmatism, and curvature of plane; and, as is the case with prism instruments, the practical resolution of a grating spectrometer is frequently limited by optical aberrations. The only major aberration, which is completely removed using reflective optics, is chromatic aberration. As these aberrations usually define the limits of performance of an instrument, much of the design of grating spectrometers involves arranging the optics in such a way that aberrations induced by one optical element are corrected in part by others.

Spherical Aberration The rays reflecting from the center of a spherical mirror do not focus in exactly the same point as the rays, which are reflected from the outer portion of the mirror. As with a lens, the rays from the various parts of the lens converge to a minimum spot size, the *circle of least confusion*, which determines the resolution limit. For a spherical mirror, the diameter of this circle is $f/64F^3$, where F is the focal ratio and f is the focal length of the mirror. Thus, a 1-m focal length instrument with a 125 mm × 125 mm diffraction grating will have a spherical aberration amounting to 86 μm at the focal plane if uncorrected. If the same grating were placed in a 2-m focal length instrument, the uncorrected spherical aberration would be only 11 μm. For a parabolic reflector, the spherical aberration is zero: all rays parallel to the principal optical axis focus to a diffraction-limited point (about 4 μm for the above example of a 1-m focal length instrument with a 125 mm × 125 mm grating).

Coma Coma in a mirror, as in a lens, affects only off-axis rays. As with a lens, annular rings of a spherical mirror images off-axis rays into circles whose radius increases with the radius of the annular ring. The comatic flare is often more pronounced in reflective imaging systems than in systems which use lenses, because geometrical constraints require the entrance and exit slits to be placed off the principal optical axes of the imaging mirrors. For a spherical mirror, collimated light is imaged into a pattern whose height and width are given as (Harwit and Sloane, 1979)

$$H = \frac{3}{16}\frac{\theta f}{F^2}\left(1 - \frac{d}{2f}\right) \tag{10.14}$$
$$W = \frac{1}{8}\frac{\theta f}{F^2}\left(1 - \frac{d}{2f}\right)$$

where f is the focal length, F is the focal ratio, d is the distance from the mirror to the aperture stop (i.e., the grating), and θ is the angle the incident light makes with respect to the principal optical axis of the mirror. For the case of a 1-m focal length instrument with a 125-mm diffraction grating, the uncorrected coma at the ends of a 20-mm slit would increase the size of the image by about 10 μm. If the instrument is required to form an extended image of the spectrum, the coma would increase with the off-axis angle: at the edge of a 250-mm focal plane, the coma would increase to 100 μm.

The increase of comatic flare with the aperture of the imaging optics is the most serious practical problem preventing the design of a high resolution Ebert or Czerny–Turner grating spectrometer with a low focal ratio, because a wide collection aperture inevitably involves a large off-axis angle which limits the resolution due to unacceptably large coma.

Astigmatism Similar to a simple lens, a spherical mirror forms an astigmatic image if illuminated off-axis, as illustrated in Figure 10.8. The tangential and saggital foci, f_T and f_S, for a spherical mirror illuminated with collimated light at an

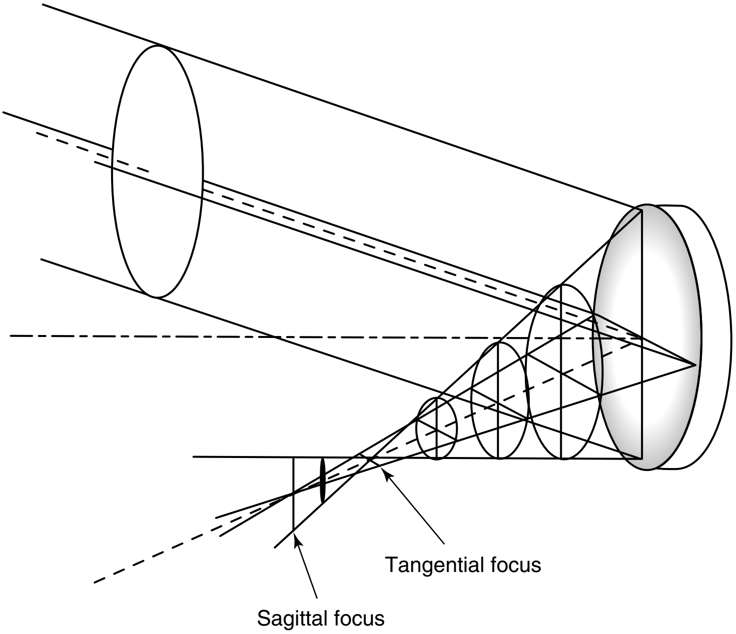

Fig. 10.8 Astigmatism of a spherical mirror. Off-axis rays of collimated light converge in different positions to form tangential and saggital foci.

angle θ to the principal axis are given as (Jenkins and White, 1976)

$$f_T = f\cos\theta \tag{10.15}$$

$$f_S = \frac{f}{\cos\theta}$$

While a paraboloidal mirror eliminates spherical aberration, it significantly exaggerates astigmatism, limiting its useful angular aperture.

Curvature of Field The focal surface of a spherical mirror is not in general a plane; this fact becomes important if an extended spectrum is to be imaged onto a photographic plate or an array detector. The exact curvature of the surface is determined by the curvature of the condensing optics, the distance of the condensing mirror from the grating, and the angle of the light with respect to the principal optical axis. It has been known since the work of Fastie (1952b) that, if the grating in an Ebert or Czerny–Turner spectrometer is placed at a distance of $2f/\sqrt{3}$ from the center of curvature of the condensing mirror, the focal surface is very nearly a plane. Reader (1969) has identified two other nearby positions which offer further corrections to the field curvature. The "corrected flat field" offers a somewhat flatter field than the Fastie position and insures that the focal plane is nearly perpendicular to the direction of propagation of light. Reader's "superflat field" provides the most nearly flat focal surface but inclines the plane slightly with respect to the direction of propagation. For the 1-m Czerny–Turner spectrometer, the offset of the grating from the Fastie flat-field position to the corrected flat-field position is 10 mm; the distance from the Fastie position to the superflat field position is 26 mm (Scheeline *et al.*, 1991).

10.5.5.2 **Types of Plane Grating Spectrometers**

Ebert-Fastie A particularly simple grating spectrometer was designed by Hermann Ebert in 1898. Dismissed and ignored by the spectroscopic community as flawed and impractical, Ebert's design was reinvented and improved in 1952 by William Fastie, who demonstrated its simplicity and optical superiority over any other instrument in use at that time (Fastie, 1952b, 1952a, 1991). The mount uses a single spherical mirror as both collimator and condenser, as illustrated in Figure 10.9(a). The grating is invariably mounted with its grooves aligned vertically, so Figure 10.9(a) is a view from above the spectrograph.

There are two implementations of the Ebert–Fastie mount: in the original design, the entrance and exit slit were placed side by side on either side of the grating (the "horizontal" Ebert mount, Figure 10.9b), while in some later versions, the slits were placed above and below the grating (the "vertical" Ebert mount, Figure 10.9c). The former has the advantage that the curvature or tilt of the image remains unchanged as the grating is rotated; the latter has the advantage of reducing stray light and thus simplifying the internal layout of the optics. In either case, both the entrance and exit slits must lie in the same plane to keep the instrument in focus over all angles of the grating; this point has been misunderstood in some commercial implementations of the Ebert mount.

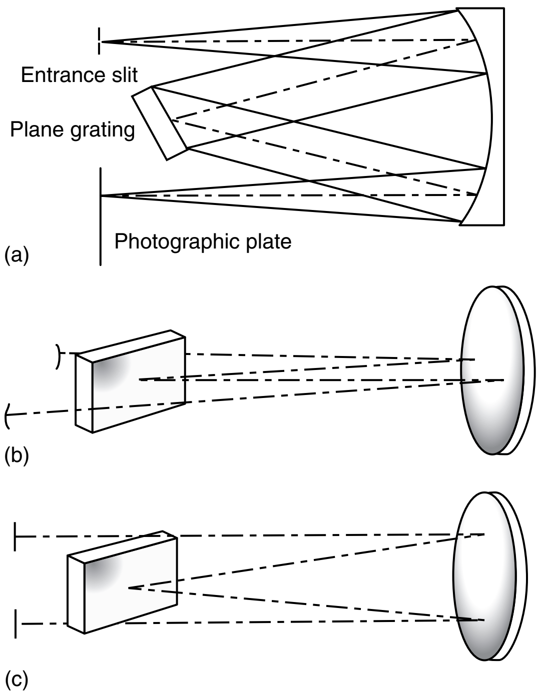

Fig. 10.9 The Ebert–Fastie grating mount, illustrating (a) a top view of an Ebert spectrograph, (b) a perspective view of a "horizontal" Ebert spectrometer, and (c) the "vertical" Ebert spectrometer.

Both designs have the effect of partially compensating for both spherical aberration and coma. In addition, Fastie has shown that the use of curved entrance and exit slits can reduce or eliminate astigmatism. The optimum shape for the slits of the horizontal Ebert spectrometer involves a pair of slits with a common center of curvature located on the line passing through the center of the grating and the center of curvature of the mirror.

Czerny–Turner If the single mirror of the Ebert–Fastie spectrometer is replaced by two independent mirrors, one for collimation and one for focusing, the Czerny–Turner spectrometer is obtained. The most commonly used design for this spectrometer is one in which two spherical mirrors are placed side by side as illustrated in Figure 10.10. This mount has all of the optical virtues of the Ebert mount (Czerny and Turner, 1930), but the extra degree of freedom afforded by the additional mirror allows superior correction of the image aberrations.

Shafer *et al.* (1964) and Reader (1969) have demonstrated that the performance of the Czerny–Turner spectrometer can be optimized by using mirrors of different focal lengths as collimator and focusing mirror. These authors have shown how the residual coma, astigmatism, and spherical aberration are affected by the choice of focal lengths. Reader has also provided equations for the focal surface of the focusing mirror in terms of the focal length and the distance of the mirror from the grating. Because of its high performance and flexibility, the Czerny–Turner mount in its many varieties forms the basis of most commercial spectrometers.

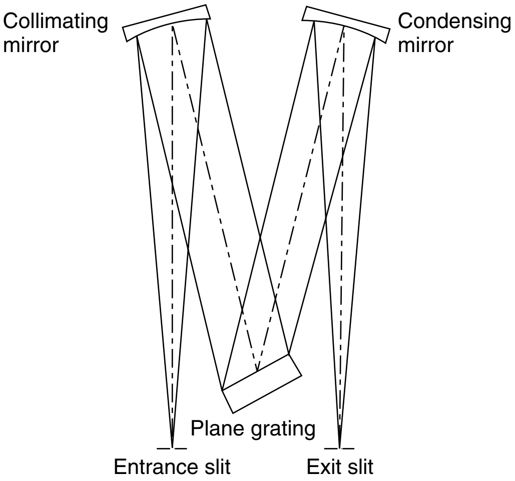

Fig. 10.10 The basic Czerny–Turner grating spectrometer.

Newtonian The most straightforward way to reduce the off-axis aberrations of coma and astigmatism is to bring the optical path as close to the principal optical axis as possible. This is the principle behind the Newtonian telescope, for instance, and it is possible to design a spectrometer based on this principle. A small plane diagonal mirror is used to bring the rays from the center of the entrance slit onto the principal optical axis of the collimating mirror, and a second plane diagonal mirror diverts the image of the spectrum from the condensing mirror to the focal plane. The problem of reducing optical aberrations then becomes the same as reducing optical aberrations in an astronomical telescope. Variations on this approach use the Schmidt or Cassegrain telescopes as collimator or condenser.

10.5.6
Concave Grating Spectrometers

If a grating is produced on a concave spherical mirror blank, the concave grating thus formed serves as both the dispersing element and as the image-forming optics. With rare exceptions, it is the only element in the optical train of a spectrometer in which it is used. There is one major advantage and one major disadvantage that follow from this choice: the advantage is that absorption losses at optical surfaces are considerably reduced, whereas the disadvantage is that the opportunity to correct optical aberrations is severely limited.

Concave grating spectrometers are of two general types: those which form images on the Rowland circle and those which do not. The former mounts include the Paschen–Runge mount, the Eagle mount, and some grazing incidence spectrometers. The latter include the Seya–Namioka and related mounts, the Wadsworth mount, and several modern designs which employ specially designed holographic gratings which compensate for certain aberrations.

10.5.6.1 Optical Aberrations of Concave Mirrors

The concave grating is ruled on a spherical mirror, and so inherits the spherical mirror's optical aberrations. Chief among these is astigmatism, because the grating is always used off the principal optical axis. In the vertical plane (i.e., the plane in which the light is dispersed), meridional rays will be focused according to the equation

$$\frac{1}{r_1}+\frac{1}{r_2}=\frac{2}{R} \tag{10.16}$$

where R is the radius of curvature of the spherical surface and r_1 and r_2 are the distances from the grating to the entrance and exit slits. The condition for horizontal focus, on the other hand, is

$$\frac{\cos\alpha}{R}-\frac{\cos^2\alpha}{r_1}+\frac{\cos\beta}{R}-\frac{\cos^2\beta}{r_2}=0 \tag{10.17}$$

where, as usual, α and β are the angles of incidence and diffraction at the grating.

For wide-aperture concave gratings, spherical aberration can also become a problem. In such a grating, rays diffracted from the extreme ends of the grooves do not come to a focus in exactly the same line as meridional rays. The effect of spherical aberration increases as the fourth power of the vertical acceptance angle, requiring a compromise between luminosity, which increases with the square of the aperture, and resolution, which decreases as the fourth power of the aperture.

10.5.6.2 Rowland Circle Mounts

If both the grating and the entrance slit are placed on a circle whose diameter is the radius of curvature of the concave grating, as indicated in Figure 10.11, then images of the slit will also be formed on this circle at diffraction angles which satisfy the grating equation, $m\lambda = a(\sin\alpha+\sin\beta)$, just as a plane grating. The horizontal focal equation for the concave grating is solved by setting

$$\frac{\cos\alpha}{R}-\frac{\cos^2\alpha}{r_1}=0 \tag{10.18}$$

and

$$\frac{\cos\beta}{R}-\frac{\cos^2\beta}{r_2}=0 \tag{10.19}$$

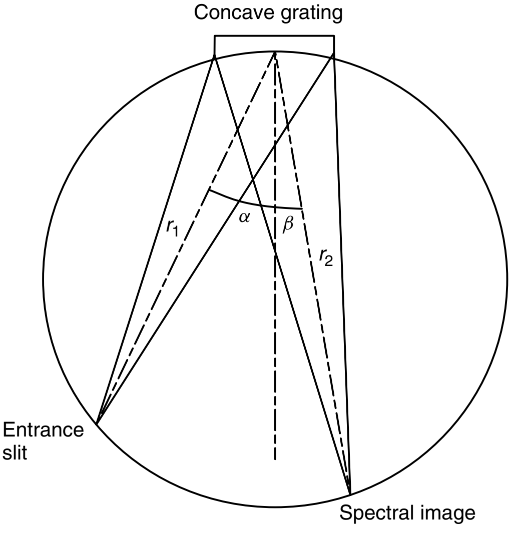

Fig. 10.11 The Rowland circle. Conjugate images of a concave grating are formed on a circle whose diameter equals the radius of curvature of the grating.

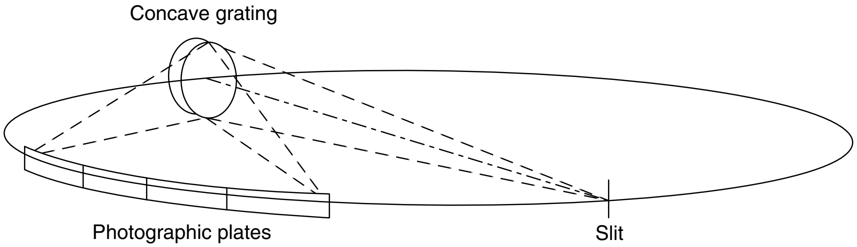

Fig. 10.12 The Paschen–Runge mount of a concave grating, illustrated for use as a spectrograph.

Paschen–Runge The most straightforward design using the Rowland circle is the Paschen–Runge spectrograph, illustrated in Figure 10.12. Here an entrance slit is placed at one point on the Rowland circle and photographic plates are located at points all around the circle wherever the spectrum is to be obtained. Using an instrument of this type, it is possible to cover the entire visible and ultraviolet spectrum at high resolution in a single exposure. The instrument occupies a very large area, and in addition is very difficult to isolate from building vibrations. As a result, it is used mostly in specialized circumstances requiring a high information bandwidth; an example would be to record the spectrum of a unique or precious sample such as a rare isotope.

Eagle Mount Among the many more compact mounts based on the Rowland circle, the Eagle mount is probably the most widely used (Eagle, 1910). In this mount, the entrance slit, grating and exit slit or plateholder are placed along a single optical axis, as shown in Figure 10.13. Variations of this mount place the plateholder in the same horizontal plane with the grating and entrance slit. In the Eagle mount, the plateholder is a segment of the Rowland circle, and it

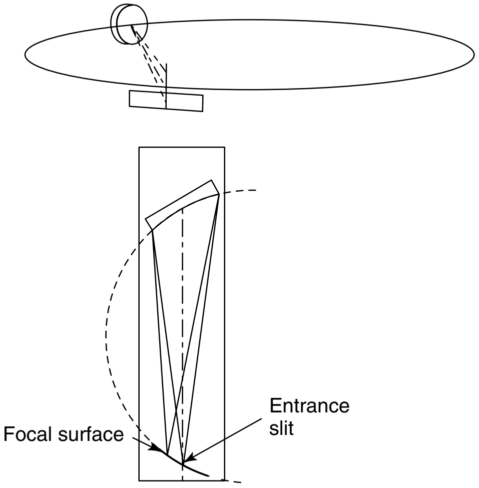

Fig. 10.13 The Eagle mount of a concave grating. The Rowland circle is indicated in both the perspective and the top view.

must be readjusted to fall on the focal surface of the grating each time the grating angle is changed. Moreover, the distance between the grating and the entrance slit must also be changed as the grating is rotated to keep the instrument in focus. Accordingly, the Eagle mount is not well suited for use as a scanning spectrometer. However, it functions very well as a high resolution spectrograph, and the compact design makes it suitable for a vacuum instrument.

Grazing Incidence Even the single surface of the concave grating produces unacceptable absorption losses at wavelengths shorter than 50 nm when used at normal incidence. At these wavelengths, use is made of the fact that short wavelength light striking a surface tangentially is reflected much more efficiently than light striking the same surface at normal incidence. This observation led to the *grazing incidence* mount of the concave grating, which is yet another mount making use of the Rowland circle. In this mount, the angle of incidence ranges from 60° to slightly less than 90°, and the angle of diffraction is only slightly less and on the opposite side of the grating normal.

At such high angles of incidence, the aberrations of the grating are quite pronounced. It has been shown that, in grazing incidence, the optimum width of a concave grating is much smaller than that for normal incidence; increasing the width above the optimum value degrades resolution without significantly increasing the central intensity of emission lines. Similarly, the useful length of a grating groove is limited by astigmatism.

10.5.6.3 Non-Rowland Circle Mounts

Seya–Namioka Other mounts of the concave grating are not based on the Rowland circle, but instead employ fixed entrance and exit slits while the grating is rotated about a point (Miyake, 1959). Such a mount is substantially simpler to use as a scanning monochromator than one based on the Rowland circle, and considerable effort has been made to find the best design to maintain focus of such an instrument as the grating is rotated.

Seya (1951) designed such a mount based on the requirement that the grating pivot about the center of its face. This design was refined by Namioka (1959), who used the condition that there be a minimum change of focus as the grating is rotated. The difference between the angle of incidence and the angle of diffraction is found to be 70°15′, and the distance of the slits from the surface of the grating is 0.8156 times the radius of curvature of the grating.

The Seya–Namioka solution of the focal equation for the concave grating involves taking

$$\frac{\cos\alpha}{R} - \frac{\cos^2\alpha}{r_1} + \frac{\cos\beta}{R} - \frac{\cos^2\beta}{r_2} = F(r_1, r_2, \alpha - \beta, \alpha + \beta) \qquad (10.20)$$

and setting F and its first three derivatives with respect to $(\alpha + \beta)$ equal to zero at some particular angle of incidence. While this condition does not guarantee a good focus over a wide range, the simplicity of the design makes it highly useful as a scanning vacuum monochromator.

Wadsworth A different kind of non-Rowland circle mount was devised by

Wadsworth to produce a stigmatic image using a concave grating. The grating is illuminated with collimated light from a concave mirror just as in a plane grating spectrometer, and the concave grating produces a stigmatic image at the focus. The image is brighter than the astigmatic image produced by the various Rowland circle mounts, provided the additional loss due to absorption at the first mirror surface can be tolerated. The focal surface is not flat, however, and both the focus and the curvature of the focal surface change as the grating is rotated, making the mount inconvenient as either a scanning spectrometer or a spectrograph. As a result, the Wadsworth mount is not widely used.

Compensated Holographic Yet another class of mounts has been designed around holographically recorded concave gratings which are specifically designed to compensate for one or more of the aberrations inherent in concave gratings. These aberration-corrected holographic gratings are not recorded with collimated light sources, and so the projection of the grooves onto a chord will not be equally spaced as is the case with a ruled grating or an ordinary concave grating. The greatest flexibility in correcting aberrations is achieved by allowing the focal surface of the grating to depart from that of the ruled concave grating.

10.6
Interferometric Spectrometers

10.6.1
Michelson

The Michelson interferometer, which has in the past four decades seen a rapid rise in popularity for use as a Fourier transform infrared spectrometer, has seen relatively little use as in the visible and ultraviolet regions. The reason for this is threefold: First, the requirement that the mirrors be kept in alignment to within a fraction of a wavelength during the full scan is a much more stringent condition in the ultraviolet than in the infrared. Second, the signal/noise ratio is not usually detector limited in the visible or ultraviolet, so that the multiplex advantage of a Fourier transform spectrometer is less important than in the infrared. Finally, the requirement for resolution of $0.001\ \mathrm{cm}^{-1}$ or better, which is important in the infrared and can most easily be achieved with a Fourier transform instrument, is less important in the visible or ultraviolet because of the much greater Doppler width.

Despite these obstacles, the Fourier transform spectrometer has seen some limited use in the shorter wavelength regions. There are two strategies for maintaining alignment during a scan. One of these involves passive stabilization of the fixed and moving mirrors using a self-aligning optical element. This may be either a corner cube or a cats-eye mirror, each of which is designed to reflect light back along the original path or parallel to it. The other strategy makes use of active stabilization: a monochromatic source such as a single frequency laser is passed through the interferometer, and the interference fringes are monitored to provide a feedback signal which corrects the tilt of one of the mirrors.

In addition to its use as a Fourier transform spectrometer, the Michelson interferometer has achieved considerable popularity in the measurement of the

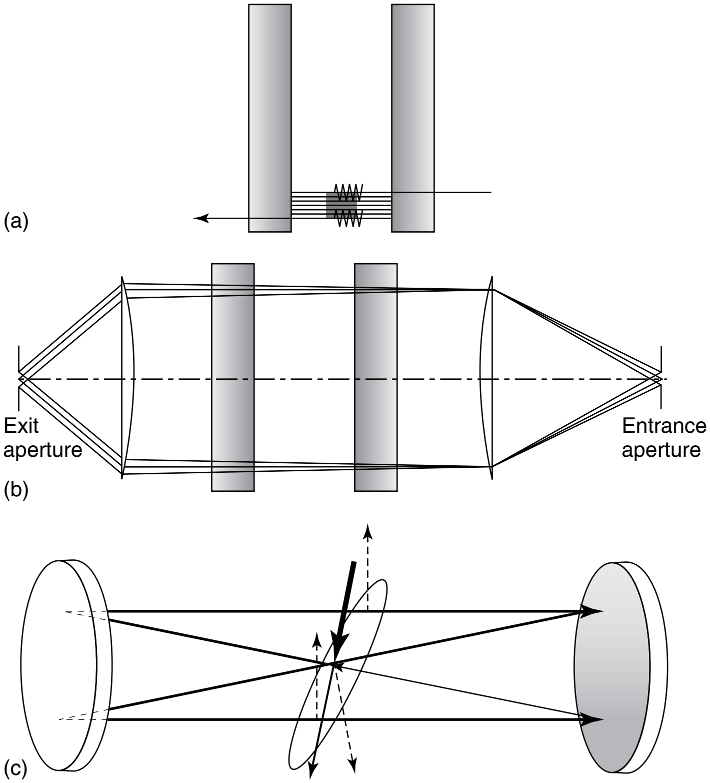

Fig. 10.14 The Fabry–Perot étalon. (a) The path of light indicating the resonance condition. (b) Illumination of the interferometer for use as a spectrometer. (c) The internally coupled confocal Fabry–Perot.

wavelength of single frequency lasers. In this application, the Fourier transform is unnecessary, as the wavenumber of an unknown laser can be obtained simply by counting the fringes during a sweep of the interferometer and comparing the number with the fringes from a laser of known frequency (Kowalski *et al.*, 1978; Fox *et al.*, 1999).

10.6.2
Fabry–Perot

The highest resolution spectrometer commonly used in the visible and ultraviolet is the Fabry–Perot interferometer or étalon (Fabry and Perot, 1899). Invented in 1899, the basic Fabry–Perot interferometer consists of a pair of partially reflective flat plates held parallel at a fixed distance d from each other (Figure 10.14a). This arrangement forms a resonant cavity which passes through light if the wavelength satisfies the condition $m\lambda = 2d\cos\theta$, where θ is the angle between the incident ray and the normal to the plate surface and m is an integer. At wavelengths which do not satisfy this condition, light is reflected. For a Fabry–Perot interferometer with perfectly flat plates and no absorption losses, the transmitted intensity is described in terms of the incident intensity I_0 and the reflectance r by the Airy formula:

$$I(\lambda) = \frac{I_0(1-r)^2}{1 - 2r\cos\left(\frac{4\pi d}{\lambda}\cos\theta\right) + r^2} \tag{10.21}$$

The Fabry–Perot interferometer can be used as the dispersing element of a spectrometer by using suitable collimating and condensing optics, as indicated in Figure 10.14(b). The entrance aperture is circular, and spectrometer produces at the output focal plane a set of concentric rings.

The resolving power R is given as

$$R = \frac{m\pi\sqrt{r}}{1-r} \tag{10.22}$$

while the free spectral range is found to be simply

$$\Delta\lambda = \frac{\lambda}{m} \tag{10.23}$$

The ratio of the free spectral range to the resolution, called the *finesse* of the Fabry–Perot interferometer, is given as

$$F_r = \frac{\pi\sqrt{r}}{1-r} \tag{10.24}$$

where the symbol F_r is used to emphasize that this is the part of the finesse due to the reflectance. The finesse is the most straightforward figure of merit for the Fabry–Perot interferometer, because it is the width of adjacent peaks of the instrument function divided by their separation. If absorption losses are considered, the peak intensity passed by the interferometer drops to

$$I_{\max} = I_0(1 - (A/(1-r))^2 \tag{10.25}$$

where A is the absorption coefficient of the reflectors.

If the interferometer plates are not perfectly flat, there will be both loss of intensity and a degradation of the resolution compared with the formulas above, and the instrument function will be a convolution of the Airy formula with these additional losses (Jacquinot, 1960). A contribution to the effective finesse can be determined from the surface irregularities if they are known: if the irregularities are characterized by a width Δd, then the finesse is limited to $F_d = \lambda/(2\Delta d)$. For example, if the surface imperfections are of the order $\Delta d = \lambda/200$, the flatness finesse F_d will be limited to 100. Increasing the reflectance to improve the reflectance finesse above this value will only introduce additional absorption and scattering losses.

In a similar way, a nonzero size for the entrance aperture of the collimating optics will result in a loss of resolution which can be cast into the form of a lower finesse. The solid angle of acceptance Ω will restrict the resolving power to a maximum value of $R_\Omega = 2\pi/\Omega$, so the aperture-limited finesse will be

$$F_\Omega = \frac{R_\Omega}{m} = \frac{2\pi}{\Omega m} \tag{10.26}$$

A variation on the plane-mirror Fabry–Perot is one which uses concave spherical mirrors placed so that their focal points coincide. This *confocal* Fabry–Perot has as its condition for constructive interference that $m\lambda = 4d$; consequently, it has a free spectral range (and resolving power) which is twice that of a plane Fabry–Perot of the same spacing. In addition, the confocal Fabry–Perot has a luminosity which exceeds that of the plane interferometer, since the interference condition is independent of the angle of incidence.

A Fabry–Perot interferometer is ordinarily used in an order $m = 10^5 – 10^6$, so that a modest finesse of 10–20 can easily produce a resolving power of a few million. However, the free spectral range is correspondingly limited so that this instrument is ordinarily used for uncluttered and isolated spectra, such as the nuclear hyperfine splitting of single atomic emission lines.

The combination of high resolution and luminosity make the Fabry–Perot interferometer the instrument of choice for monitoring and stabilizing high-resolution lasers such as dye lasers or diode lasers. One particularly useful configuration is an ‘‘internally coupled’’ confocal interferometer, Figure 10.14(c) (Reich *et al.*, 1986). The ‘‘internal coupling’’ refers to the means of coupling light into the cavity via a nearly transparent beam splitter placed near the common foci of the two spherical mirrors that make up the cavity. This method of coupling the light in and out of the cavity eliminates the need for specialized partially reflective dielectric coatings on the mirrors, greatly increasing the useful spectral range of the device: a single instrument can typically be operated without modification from the mid-infrared into the visible spectrum.

10.7
Hybrid and Compound Instruments and Imaging Filters

There are a large number of innovative designs for spectrometers which do not fit neatly into any of the categories above. Some of these are compound instruments with more than one dispersing element; others are hybrids which make use of diffraction gratings in what are essentially interferometers.

10.7.1
Order-Sorter for a Grating Spectrometer

The most common type of compound instrument involves a low-resolution spectrometer which serves as a predisperser or order-sorter for a high-resolution spectrometer with a limited free spectral range. As an example, consider a grating spectrometer employing a 300 groove/mm grating blazed at 65°. A 250-mm-wide grating of this type will produce a resolving power of 750 000 in the tenth order when used to observe a wavelength of 600 nm. It will also, however, diffract light at 545 and 666 nm in the ninth and eleventh orders, and these orders must be removed to record an absorption spectrum. This is usually accomplished by illuminating the slit of the grating spectrometer with a small prism monochromator centered at the wavelength of interest. The prism instrument can, at the modest resolution required to separate the adjacent orders, be designed to have high throughput so that the luminosity of the overall instrument is only slightly reduced.

10.7.2
Harrison Echelle

If the grating of the Czerny–Turner spectrometer is chosen to be an echelle, the spectrometer can be very efficient and deliver very high resolution. The disadvantage of such an arrangement is that the echelle must operate in a high order (typically $m = 50–100$ in the visible), and wavelengths from many overlapping orders are diffracted at the same angle. To separate the orders,

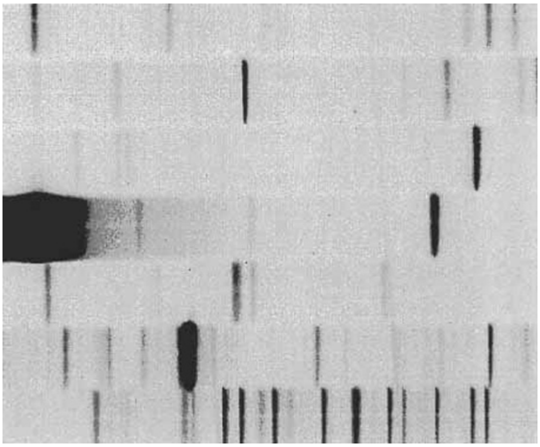

Fig. 10.15 A portion of an echelle spectrogram. Wavelength is dispersed horizontally within an order, while adjacent orders are separated vertically.

some other dispersing element must be used. One ingenious solution to this problem is to use a prism placed immediately in front of the grating as an order-sorting cross disperser. The prism-grating pair can be matched in such a way that the spectrum appears as short strips of adjacent orders as illustrated in Figure 10.15. A second solution which produces much the same result is to replace the condensing mirror of the echelle spectrometer with a concave grating whose grooves are perpendicular to those of the echelle.

10.7.3
Fabry–Perot/Grating Instrument

A second such tandem instrument is the Fabry–Perot interferometer in series with a grating spectrometer. In this case, it is the grating instrument which acts as a low-resolution order-sorter for the much higher resolution étalon. If the free spectral range of the étalon is matched to the resolution of the grating spectrometer, the maximum theoretical resolution is increased by a factor of the finesse of the Fabry–Perot. The instrument can be used either as a scanning spectrometer, by scanning the grating and the interferometer synchronously, or it can be used as an imaging spectrograph, in which case the image is a two-dimensional pattern of Fabry–Perot rings superimposed on the spectrometer slit.

10.7.4
Double and Triple Monochromators

In many instances, a tandem spectrometer combination is not intended to increase the resolution but rather to improve the signal/noise ratio. Double or triple monochromators are used to reduce scattered light from a prominent emission line and so improve the ability to observe a nearby faint feature. The exit slit of the first monochromator serves as the entrance slit of the second monochromator, which is scanned synchronously with the first. If the dispersion is additive, the entrance slit of the first and last monochromator determine the spectral resolution to a first approximation, while the central slit determines the rejection of scattered light. A typical use of such a spectrometer would be to observe low-frequency Raman or Brillouin scattering in the presence of a nearby exciting line. If the dispersion is subtractive, the double monochromator is essentially a sharp cutoff bandpass filter whose width is determined by the width of the central slit. The subtractive dispersion arrangement has the advantage in time-resolved experiments that all rays are transmitted through the instrument in the same time interval, because any optical retardation of a marginal ray through one monochromator is cancelled by the other.

10.7.5
Imaging Filters

Imaging spectrometers have long been of interest to astronomers, but they are also useful for any application requiring position-sensitive wavelength discrimination. One promising technology for imaging spectrometers is the acousto-optic tunable filter. These devices make use of Bragg diffraction by acoustic waves in an optoelastic medium. These acoustic waves are produced by a piezoelectric element in contact with the optoelastic element; and by tuning the acoustic excitation frequency, it is possible to tune the transmitted wavelength. Filters with a resolving power of a few hundred have been demonstrated that are tunable through much of the visible and into the near infrared (Georgiev *et al.*, 2002).

10.7.6
The Sisam Spectrometer

A hybrid instrument originally described by Connes (1959) attempted to capture the advantages of Michelson interferometer without the necessity of implementing the Fourier transform. The instrument consists of a Michelson interferometer in which the two mirrors have been replaced by diffraction gratings. The most significant technical problem with this arrangement is that the gratings are required to rotate in tandem; otherwise, the resolution of the instrument is degraded. This problem was effectively solved by Till *et al.* (1975), who constructed an optical system in which the two diffraction gratings were mounted back-to-back on a single rotating platform. Although the availability of fast and inexpensive computers has eliminated the main advantage of the Sisam spectrometer, it remains a useful design for highly specialized tasks; chief among these is the high resolution detection of a weak emission signal in the presence of a much stronger signal.

10.8
Detectors

10.8.1
Ideal Detector

The ideal detector would have unit quantum efficiency (that is, one photoelectric or photochemical event per photon), no dark current or readout noise, zero response time, and a flat response over a wide spectral range. In addition to these features, a detector which can increase the *information bandwidth* either by multiplexing a signal or by allowing parallel detection of several signals will improve the performance of a spectrometer enormously. Naturally, no single detector type embodies all of these features, and different compromises among the various desirable properties of detectors must be made depending on the nature of the experiment and the spectral region to be covered.

10.8.2
Types of Detectors

Photographic Plate The first detector used for making a permanent record of a spectrum was the photographic plate. Photographic plates can be obtained which cover the spectral region in the near infrared to about 1200 nm or through the vacuum ultraviolet. Photographic plates are especially suitable for high resolution measurements because of the ease of superimposing a wavelength calibration on a spectrum using an emission source which produces known wavelengths. In addition, the information bandwidth of a photographic plate is extremely high: if a single emission line is 10 μm wide (a typical width), a 250-mm-wide photographic plate has an advantage of 25 000 over a spectrometer using a single photoelectric detector with comparable quantum efficiency. If the spectrometer is of the Harrison echelle type, the advantage becomes even greater and can easily reach several million. This advantage is mitigated somewhat by a relatively low quantum efficiency (typically 0.01 or lower) and is severely compromised by the highly nonlinear response of photographic emulsions. The main difficulty with photographic detection is, however, the inconvenience of waiting several hours between performing an experiment and measuring the recorded spectrum. In addition, there is no way of adapting photographic recording to exploit modern time-dependent techniques. Despite these inconveniences, there are certain spectrograph designs such as the Paschen–Runge mount, for which the photographic plate has no cost-effective replacement.

Photomultipliers The photomultiplier tube (PMT) is the detector of choice for most spectroscopic applications. This detector consists of a vacuum tube containing a photoemissive surface (a *photocathode*) which ejects electrons when exposed to light. These electrons are accelerated by an electric field toward elements of a *dynode chain*, each of which will emit several electrons when struck by an energetic electron. The photocathode determines the spectral sensitivity and quantum efficiency of the photomultiplier, while the dynode chain serves to amplify the signal of a single photoelectron to a measurable signal. Photomultipliers can be obtained commercially with useful sensitivity from 200 to 900 nm. Quantum efficiency is typically 0.1–0.3 at peak sensitivity, and the internal amplification of the photomultiplier boosts the signal by a factor of about 10^6.

The photomultiplier produces a burst of current in response to each photon, and this feature may be used to count photons in favorable circumstances. The rise time of a pulse of electrons is a few nanoseconds, while the transit time is approximately 10 times that value; accordingly, the counting rate of photomultiplier pulses may range upward of several million per second before overlapping pulses begin to become a problem. The photon noise is simply the square root of the number of photons, so that the signal/noise ratio is about 10^3 for a 1-second sampling time under these circumstances. The dark count in selected photomultipliers may be as low as 10 counts per second, leading to photon noise limited behavior at count rates as low as a few hundred per second.

The impact of the availability of these detectors on spectrometer design cannot be overstated. In the infrared, even the best detectors have responses which are limited by the noise of the detector itself. In consequence, infrared spectrometers have always been designed to provide maximum efficiency of detecting a small signal in the presence of a large amount of noise. This has routinely been achieved by multiplexing the signal, leading to the popularity of Fourier transform or Hadamard transform spectrometers in the infrared. In the visible and ultraviolet regions, such techniques do not provide any great benefit because the detectors are of such high quality.

An additional advantage of electronic detectors such as photomultipliers is that they can be used for *phase sensitive detection*. If the source of light incident on a spectrometer is modulated (either in amplitude or in wavelength), the signal emerging from the spectrometer will also be modulated. A modern lock-in amplifier can then be used to amplify only that part of the signal from the detector that has both the same frequency as the source modulation and a constant phase with respect to that modulation. This scheme of detection can greatly reduce the bandwidth of the noise spectrum, improving the signal/noise ratio by several orders of magnitude.

Photodiode Several types of solid state detector have become available recently. Most of these depend on the ability of a photon to promote an electron in a semiconductor from one energy level to another where it is more mobile. The simplest of these is the photodiode, which is the solid state equivalent of an unamplified vacuum tube photocell. Each incoming photon with an energy in excess of the semiconductor band gap promotes one electron from the valence band to the conduction band, and this may be detected in the form of a change in potential or a change in current. The difficulty with this device is that the current carried by a single electron is well below the noise level of an ammeter, so the device is insensitive to a low light level. A typical silicon photodiode produces $1\ \mathrm{A\ W^{-1}}/$ of incident radiant power compared with $10^5\ \mathrm{A\ W^{-1}}$ for a good photomultiplier. At a radiant power of 10^{-12} W, the signal/noise of a photomultiplier is about 10^4 times higher than a photodiode.

The two advantages the photodiode have over the PMT are speed and infrared sensitivity. The photodiode can be configured so as to have a very large frequency bandwidth (cutting off at frequencies above 1 GHz). Silicon photodiodes have peak quantum efficiency in the red of 0.8 or better, retaining quantum efficiency as high as 0.1 out to 1000 nm. Indium gallium arsenide (InGaAs) photodiodes are sensitive to wavelengths as long as 1700 nm.

Avalanche Photodiode One improvement on the photodiode that makes it suitable for low light level detection is the avalanche photodiode (APD). This device has a built-in amplification similar in principle to the photomultiplier: the diode is run at a high reverse bias so that a photoelectron is accelerated by the applied field, ejecting a cascade of secondary electrons. This internal amplification boosts the current produced by a factor of about 100, improving the overall signal/noise ratio by a factor of 50. The amplification is high enough to allow APDs to be adapted for use in photon counting, and photon counting units with matched amplifier/discriminators are now available. With a quantum efficiency of 0.8 and a dark count as low as 25 counts/s, the APD is becoming competitive with the photomultiplier, especially at wavelengths longer than 800 nm.

Diode Array A second technique for improving the performance of the photodiode is to place a large number of photodiodes (128 up to 2048) in a linear array at the focal plane of a spectrometer. This tactic improves the information bandwidth of the instrument as a whole by a factor up to the number of parallel detectors. In spectrometers employing a diode array, sensitivity is further improved by the use of an image intensifier to amplify the image of the spectrum before it is imaged onto the diode array. The exit aperture is fixed at the size of a single photodiode, which is ordinarily 25 μm wide and up to 1 mm high.

The availability of inexpensive solid state array detectors such as the diode array and charged coupled device (CCD) has prompted a dramatic miniaturization of laboratory spectrometers. Instruments based on the Czerny–Turner mount are now commercially available which easily fit in the palm of the hand and weigh less than 300 g, including the detector and readout electronics. These offer coverage of the visible and UV spectrum at a resolving power of roughly 10^3.

CCD A different type of array detector is the CCD, which is commonly employed in video cameras. The CCD is similar to the photodiode in that the photon is used to promote an electron, producing an electron–hole pair. Instead of using the electron to produce a current, however, the CCD captures the electron on a small capacitor (a "gate"), allowing many electrons to accumulate before the charge packet is converted to an output voltage and digitized. Each pixel consists of such a gate, and a single CCD chip may contain several million pixels arranged as a rectangular array. The advantage of a CCD is that, by accumulating a charge over an exposure of several seconds or minutes, the integrated signal may be made to surpass the readout noise, even though the light level is low. For best performance, therefore, the dark current must be kept to a minimum, so that the dark current does not saturate the gate before the pixel is read; this is most often accomplished by cooling the CCD chip. Like the diode array, a CCD fixes the spectrometer exit aperture at the size of a single pixel. This size ranges from 25 μm × 25 μm down to 9 μm × 9 μm for CCDs, which are presently available commercially.

The simplest use of a CCD is to place it in the focal plane of the instrument in place of a photographic plate (Florek *et al.*, 1996). The CCD is becoming a sufficiently popular detector, however, that customized spectrometers are now being designed to take advantage of the properties of this detector (Scheeline *et al.*, 1991; Den Hartog and Holly, 1997; Furenlid and Cardona, 1988).

Conventional front-illuminated CCDs are silicon-based devices, and so are limited in their blue and ultraviolet sensitivity due to the absorption of short wavelength light by silicon before the photon can penetrate to the gate. This problem has been partly overcome by the use of thinned back-illuminated CCDs, with the result that CCDs are now available with useful sensitivity throughout the ultraviolet. As with photodiodes, CCDs based on indium gallium arsenide are now becoming available; these InGaAs detectors are sensitive from about 900 nm well into the near infrared.

Glossary

Aberration: The Gaussian approximation for a ray passing through a lens or reflected from a spherical mirror is that the angle of deviation θ is small enough for $\sin\theta$ to be replaced by θ. Any departure from this approximation is an *optical aberration*. For practical purposes, any departure from geometric optics that causes a point source to be imaged into a nonpoint image is an aberration.

Dark Current: Current from a photodetector, which is present even in the absence of illumination.

Information Bandwidth: The speed with which information channels may be recorded. In a spectrometer, the information bandwidth may be increased either by increasing the efficiency of recording a single wavelength or by recording several wavelengths simultaneously.

Monochromator: An instrument used to isolate a single wavelength of light.

Polychromator: An instrument capable of isolating several monochromatic rays simultaneously from an incident beam.

Spectrograph: An instrument which produces a graphical record (usually a photograph) of an optical spectrum.

Spectrometer: Any instrument that measures the intensity of light as a function of wavelength.

Spectrophotometer: A spectrometer specialized for accurately recording the intensity of light.

References

Connes, P. (1959) *Rev. d'Opt.*, **38**, 158.
Czerny, M. and Turner, A.F. (1930) *Z. Phys.*, **61**, 792.
Den Hartog, D.J. and Holly, D.J. (1997) *Rev. Sci. Instrum.*, **68**, 1036.
Eagle, A. (1910) *Astrophys. J.*, **31**, 120.
Fabry, C. and Perot, A. (1899) *Annal. Chim. Phys*, **16**, 115.
Fastie, W.G. (1952a) *J. Opt. Soc. Am.*, **42**, 641.
Fastie, W.G. (1952b) *J. Opt. Soc. Am.*, **42**, 647.
Fastie, W.G. (1991) *Phys. Today*, **44**, 37.
Florek, S., Beckerross, H., and Florek, T. (1996) *Fresenius Z. Anal. Chem.*, **355**, 269.
Fox, P.J., Scholten, R.E., Walkiewicz, M.R., and Drullinger, R.E. (1999) *Am. J. Phys.*, **67**, 624.
Furenlid, I. and Cardona, O.U.S.A. (1988) *Publ. Astronom. Soc. Pac.*, **100**, 1001.
Georgiev, G., Glenar, D.A., and Hillman, J.J. (2002) *Appl. Opt.*, **41**, 209.
Harwit, M. and Sloane, N.J.A. (1979) *Hadamard Transform Optics*, Academic Press, New York.
Hutley, M.C. (1982) *Diffraction Gratings*, Vol. 6, Academic Press, London, p. 330.
Jacquinot, P. (1960) *Rep. Prog. Phys.*, **23**, 268.
James, J.F. and Sternberg, R.F. (1969) *The Design of Optical Spectrometers*, Chapman and Hall, Ltd., London.
Jenkins, F.A. and White, H.E. (1976) *Fundamentals of Optics*, 4th edn, McGraw-Hill, New York, p. 111.
Kowalski, F.V., Teets, R.E., Demtröder, W., and Schawlow, A.L. (1978) *J. Opt. Soc. Am.*, **68**, 1611.
Loewen, E.G., Nevière, M., and Maystre, D. (1977) *Appl. Opt.*, **16**, 2711.
Miyake, K.P. (1959) *Sci. Light*, **8**, 39.
Namioka, T. (1959) *J. Opt. Soc. Am.*, **49**, 951.
Reader, J. (1969) *J. Opt. Soc. Am.*, **59**, 1189.
Reich, M., Schieder, R., Clar, H.J., and Winnewisser, G. (1986) *Appl. Opt.*, **25**, 130.
Sawyer, R.A. (1963) *Experimental Spectroscopy*, 3rd edn, Dover Publications, New York, p. 358.

Seya, M. (1951) *Sci. Light*, **2**, 8.

Scheeline, A., Bye, C.A., Miller, D.L., Rynders, S.W., and Owen, R.C. (1991) *Appl. Spectrosc.*, **45**, 334.

Shafer, A.B., Megill, L.R., and Droppelman, L. (1964) *J. Opt. Soc. Am.*, **54**, 879.

Till, S.M., Jones, W.J., and Shotton, K.C. (1975) *Proc. R. Soc. London, A*, **346**, 395.

Further Reading

Hutley, M.C. (1982) *Diffraction Gratings*, Academic Press, London.

James, J.F. (2007) *Spectrograph Design Fundamentals*, Cambridge University Press, Cambridge.

James, J.F. and Sternberg, R.F. (1969) *The Design of Optical Spectrometers*, Chapman and Hall, Ltd., London.

Jenkins, F.A. and White, H.E. (1976) *Fundamentals of Optics*, 4th edn, McGraw-Hill, New York.

Kitchin, C.R. (1995) *Optical Astronomical Spectroscopy*, Institute of Physics Publishing, Bristol and Philadelphia.

Loewen, E.G., Nevière, M., and Maystre, D. (1977), **16**, 2711–2721.

VLSI/ULSI—INTEGRATED CIRCUITS

STEVEN J. HILLENIUS, *Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey, U.S.A.*

Introduction 297
1. **Principles of Operation** 299
1.1 Metal–Oxide–Semiconductor (MOS) Field-Effect Transistors (FET) 299
1.2 Bipolar Transistors 300
2. **Structures and Devices** 301
2.1 Complementary MOS (CMOS) 301
2.2 Bipolar 302
3. **Manufacturing Techniques** 303
3.1 Implantation and Dopant Diffusion 303
3.2 Thermal Oxidation 306
3.3 Lithography 306
3.4 Deposition 308
3.4.1 Chemical Vapor Deposition (CVD) 308
3.4.2 Physical Vapor Deposition (PVD) 308
3.5 Etching 309
3.5.1 Wet Etching 309
3.5.2 Dry Etching 309
3.5.3 Chemical Mechanical Polishing 310
3.6 Constraints for VLSI Manufacturing 310
3.6.1 Process-Control Methodologies 310
3.6.2 Defect Control 311
3.7 Device Reliability 313
3.7.1 Hot-Carrier Aging 313
3.7.2 Dielectric Breakdown 314
3.7.3 Electromigration 315
4. **Applications** 315
4.1 Memory Devices 315
4.1.1 Dynamic Random-Access Memories (DRAM) Circuits 315
4.1.2 Static Random-Access Memories (SRAM) Circuits 316
4.1.3 Nonvolatile Memories 316
4.2 Microprocessors and Logic 317
5. **Economic Aspects and Industry Scaling Trends** 317
5.1 Device Scaling 317
5.2 Industry Scaling costs 318
Works Cited 319
Further Reading 319

INTRODUCTION

The term VLSI (very large-scale integration) refers to one stage of the continual process of putting more and more functions and individual components into a small area of single-crystal silicon. These devices and the techniques that are used to achieve this are the subjects of this article. There is not a hard rule on how the level of integration is described or on how many individual components are necessary to qualify an integrated circuit as VLSI. Circuits of various levels of complexity are referred to as either small, medium, large, very large, or ultra-large integration. This is according to a convention of using ULSI to describe devices with 2–64 million individual components (Broydo and Osburn, 1987). VLSI devices are described as having 0.06–2 million components in this convention, with each new category having a factor of 2^5 (32) more devices than the previous generation. Most of the state-of-the-art large circuits commercially available, such as memories and microprocessors, now have on the order of 10 million transistors, which would make them ULSI devices. The largest circuits that have been reported are the 1-gigabit memory circuits, which have over a billion (10^9) transistors (Horiuchi *et al.*, 1995; Sugibayashi *et al.*, 1995). The term VLSI is commonly used to describe these integrated circuits, and the content of this article is applicable to all of the levels of integration. The techniques of the most modern devices are emphasized in this article.

The integrated circuit was first proposed by Jack Kilby in 1957 (Kilby, 1976) as a way of connecting more than one device to form

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

a circuit. However, the widespread application of this technique and the economies of scaling of these devices did not become a reality until the late 1960s. Over the last 20 years, the scale of integration has doubled every 18 months. Two types of circuits, memories and microprocessors, have been most successful in taking advantage of the new levels of integration. This trend of putting more components and functionality on a single integrated circuit has been coupled with an enormous increase in the number of circuits produced per year. The result of all of the production improvements and the considerable economies of scale have resulted in the wide-scale application of integrated circuits in a growing number of consumer products. Integrated circuits now permeate the life of nearly everyone in modern society. The integrated circuit is used in electronic devices including telephones, calculators, computers, televisions, kitchen appliances, automobiles, video games, door knobs, and many others. The current rate of production of transistors is estimated to be 10^{17}/year, which equals about 4000 transistors per person per day for every man, woman, and child on Earth. The changes in the way of life we have witnessed in telecommunication, transportation, and information storage and retrieval have all been influenced by the widespread application of the VLSI technology.

In Fig. 1 is shown the relationship of the wafer, chip, and component sizes that illustrates the dimensions involved in ULSI manufacturing. This figure shows the scope of the topics that are discussed in this article. The device, circuit, and wafer dimensions are all the key indicators of the progress being made. Figure 2 shows the design rules and density trends for DRAMS (Komiya, 1993). This figure illustrates how the levels of integration have been growing over the last 10 years and the projection for this trend to continue into the future. The wafer sizes and the chip sizes are steadily increasing as the feature size gets smaller, as seen in Fig. 3. The result of all these trends in the complexity of the devices is that the ability to control yield-limiting defects becomes the single most important parameter to determine viability of a technology. The yield-limiting defects come in two varieties: process parameters that are out of the acceptable range and random defects due to particles or missing or extra features. The techniques to deal with these defects are quite different. These techniques, along with device structures and manufacturing techniques, are presented in this article. All of these compo-

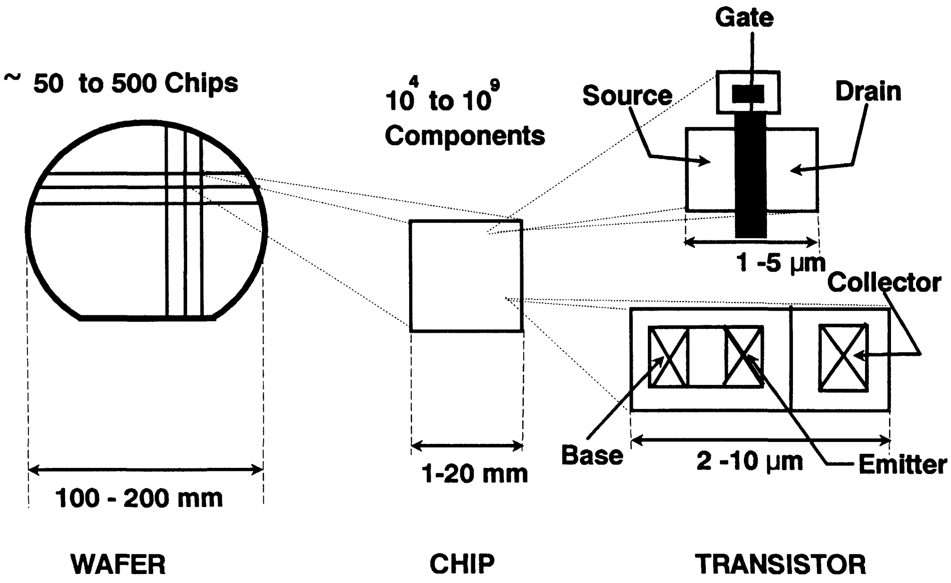

FIG. 1. Characteristic dimensions of the wafer, chip, and transistor for silicon ULSI devices.

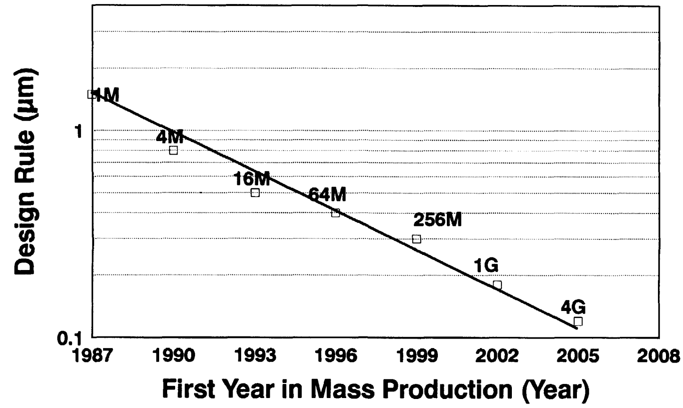

FIG. 2. Design rules for the generations of DRAMs from 1987 and projected to 2005 (from Komiya, 1993).

nents are critically important aspects of the ULSI technology.

1. PRINCIPLES OF OPERATION

1.1 Metal-Oxide-Semiconductor (MOS) Field-Effect Transistors (FET)

The most common type of active element used in VLSI devices is the metal-oxide-semiconductor (MOS) field-effect transistor (FET). This device is the most commonly used on account of the simplicity of operation, the low power consumption, and the favorable properties of the device as it is reduced in size. It is this scaling feature that has allowed the tremendous reduction in device size and increase in complexity. The basic operation of a MOS transistor is schematically indicated in Fig. 4. This figure indicates the four terminals of the device: source, gate, drain, and bulk. The current flows from the source to the drain in an n-channel MOS device depicted here, with the current flow regulated by the voltage on the gate. This is the basic operation and is described by

$$I_{ds} \propto \mu W/LD(V_{gs} - V_t)^2, \quad (1)$$

where μ is the electron mobility, W is the transistor width, L is the length of the transistor gate, V_{gs} is the voltage on the gate, D is the gate oxide thickness, and V_t is the threshold voltage of the transistor. The importance of the gate oxide thickness and the gate length can be seen in this expression. Decreasing either of these parameters results in an increase in the drive current. In the operating region where the gate voltage is below the threshold voltage, the current volt-

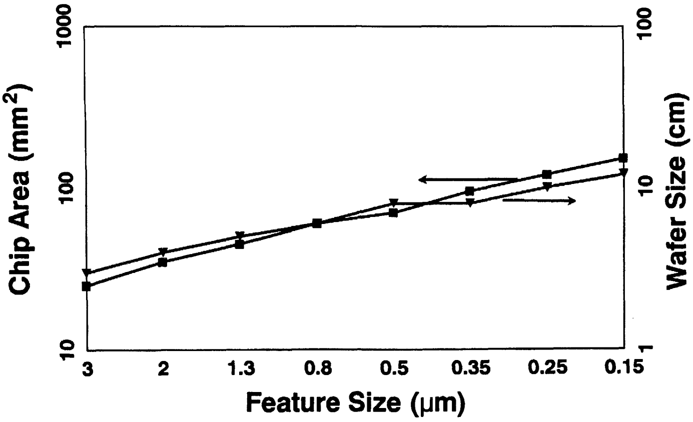

FIG. 3. Chip and wafer size trends for memory devices as a function of the minimum feature size (from Komiya, 1993).

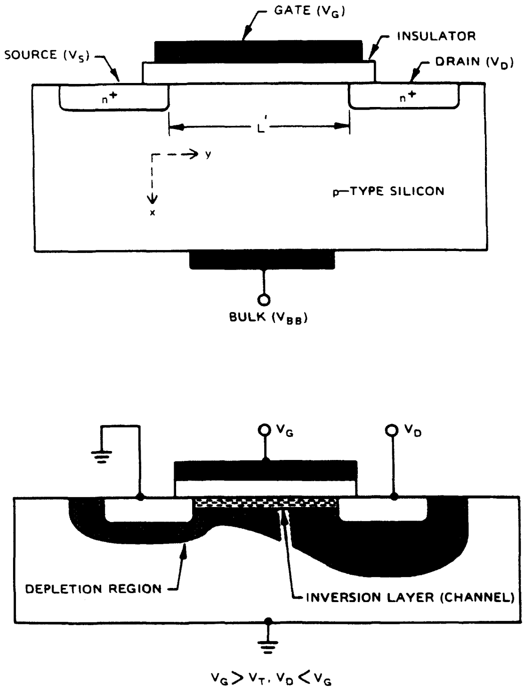

FIG. 4. Cross section of MOS transistor showing the components and the charge distribution in the linear region (from Beadle *et al.*, 1984).

age relationship is dominated by diffusion effects, and the drain current varies exponentially with the gate voltage. This is referred to as the subthreshold region of operation. The ratio of the current with the device on (I_{on}) to the current with the device off (I_{off}) is typically on the order of 10^6 to 10^8 for these devices. See TRANSISTORS, FIELD-EFFECT for a more detailed description of the operation of this device.

This basic transistor can be used in many different configurations, but the simplest building block of most digital circuits is the inverter. The inverter takes the input signal and produces an output signal that is the opposite state. The most common technology used for modern VLSI circuits is called complementary MOS (CMOS). CMOS technology uses NMOS transistors as described above and a PMOS or *p*-channel transistor, which operates just like the NMOS transistor, but the S/D region consists of p-type silicon and the channel region is *n* type. This gives the *p*-channel transistor a current that is exactly like the NMOS transistor except that the polarity is the opposite type. The basic inverter structure for the CMOS circuit is shown in Figs. 5(a) and 5(b). The most important characteristic of this circuit is that the steady-state current drawn by the inverter is only the "off" current of the individual transistors. This allows for a very low use of power in these circuits and is the primary reason why the CMOS type of circuit methodology is used in VLSI devices. Figures 5(c) and 5(d) show the voltage and current drawn by the inverter during the transition from one state to another.

1.2 Bipolar Transistors

Bipolar transistors are electronic devices with two *p-n* junctions in very close proximity that produce three distinct regions in the silicon described as the emitter (E), base (B), and collector (C) regions. A bipolar transistor can either be an *npn* or *pnp* according to the type of doping for the respective regions. A cross-sectional view for the basic *npn* and *pnp* bipolar transistor structure is shown in Fig. 6. The two distinct junctions formed are the emitter-base junction and the base-collector junction. Under equilibrium conditions, there is no net current flow across either of these junctions. When an external voltage is applied to the transistor, either of these junctions may be forward or reverse biased. When the E-B junction is forward biased in the *npn* transistor, a large number of electrons are injected into the *p*-type base region as a result of the forward bias applied to the *p-n* junction. If the distance between the emitter and collector is small enough, then most of the electrons will diffuse to the reverse-biased C-B junction. This effect is the essence of the bipolar transistor action, which is the large current flowing through a reverse-biased junction due to the close physical proximity of a forward-biased junction. The practical outcome of this transistor action is the ability to produce an amplifying effect on small currents or "current gain." This current gain is defined by the change in the emitter current (I_E), which consists of the electrons that reach the collector plus those that flow out of the transistor through the base (I_B), as the base current is changed. Two parameters that describe this transistor action are the common-base current gain,

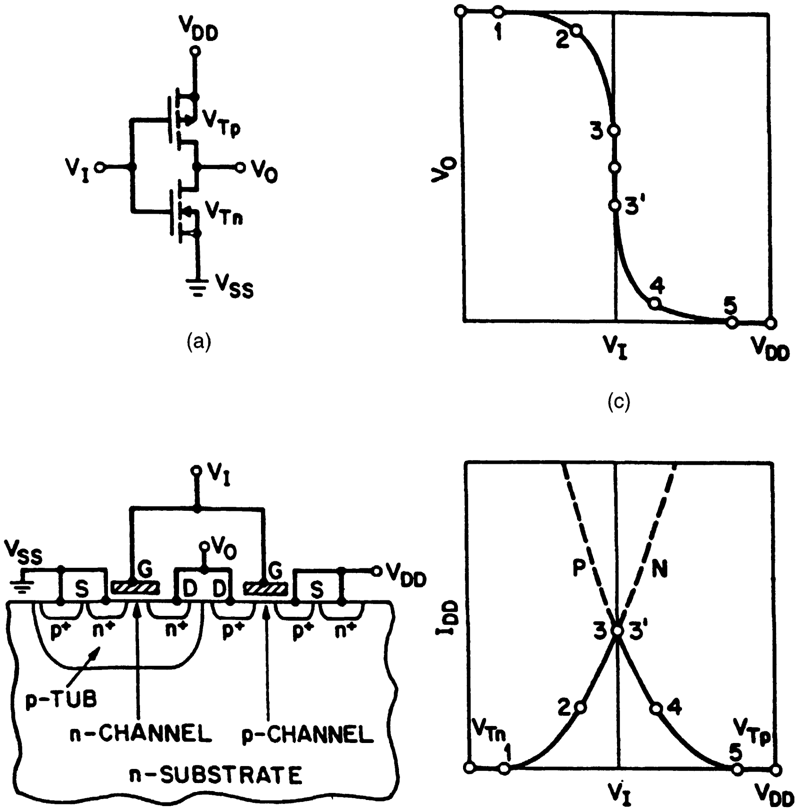

(a) (b) (c) (d)

FIG. 5. CMOS inverter showing the (a) circuit schematic, (b) device cross section, (c) output vs input voltage of inverter, and (d) current through inverters a function of input voltage (solid curve); *I–V* characteristics of *n*- and *p*-channel transistors (dashed curves). The numbers correspond to different points on the inverter transfer characteristic.

$$\alpha = \delta I_C/\delta I_E, \tag{2}$$

and the common-emitter gain,

$$\beta = \delta I_C/\delta I_B. \tag{3}$$

And with the total-current relation $I_E = I_C + I_B$, we get the relationship between the two gains as

$$\beta = \alpha/(1 - \alpha). \tag{4}$$

Typically, the value of α is very close to unity and the value of β is on the order of 100–200. A more detailed discussion of the physics of the transistor can be found in TRANSISTORS, BIPOLAR.

2. STRUCTURES AND DEVICES

2.1 Complementary MOS (CMOS)

The basic structure for building the transistors and interconnects consists of building successive layers of insulating, conducting, and semiconducting materials. The individual layers are patterned by the lithographic techniques and etching techniques described later in this article. The basic structure of the transistor is shown in Fig. 7. The basic device structure is formed by the combination of the basic steps shown in Fig. 8. The basic elements of any silicon device consist of various combinations of these steps. The CMOS device cross section is shown in Fig. 9(f), and the sequence of steps used to form the transistors is shown in Fig. 9. This shows

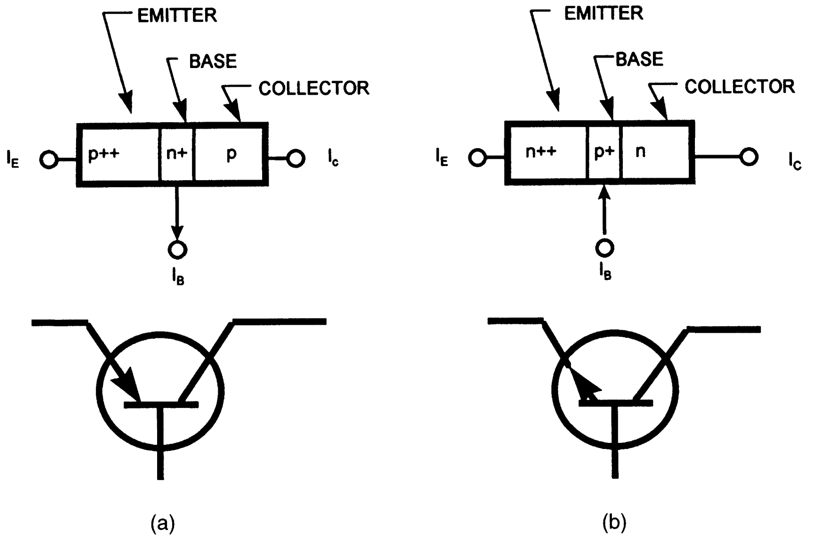

FIG. 6. Parts of the bipolar transistor and symbols used to denote (a) *pnp* and (b) *npn* transistors.

the two types of transistors (NMOS and PMOS) formed in a *p* tub and an *n* tub, respectively. The starting material for this process is an *n*-type silicon wafer with a lightly doped *n*-type epitaxial layer that is sometimes denoted with the Greek letter *ν*. The process consists of masking the implants using the patterned nitride layer shown in Fig. 9(a) as an ion implant (I^2) mask. This nitride layer is also used as an oxidation mask to form an oxide layer in the region of the *n* tub. This oxide layer is then used to mask the implant of the boron ions to form the *p* tub as is shown in Fig. 9(b). The oxide is then stripped in a wet etch, and the tub dopants are diffused into the silicon with a high-temperature anneal. The anneal is performed in a dilute oxygen ambient so that there is no loss of dopant [Fig. 9(c)]. The field oxide is then grown by a local oxidation process, which is then followed by a gate oxide formed and a polysilicon gate deposited and patterned [Fig. 9(d)]. The gate is then used as a mask for the source and drain implants of boron and arsenic. The boron is implanted in all the source and drain regions, but the arsenic is selectively implanted into the NMOS regions [Fig. 9(e)] to form the *n*+ source and drains for the NMOS and *p*+ source and drains for the PMOS transistors. The last step in Fig. 9(e) is to passivate the devices with an insulating layer of phosphorus-doped SiO_2 to insulate the devices electrically from the subsequent metal layers. Phosphorus-doped glass is used because the phosphorus prevents the diffusion of sodium and other mobile ions, which cause serious problems if they penetrate to the active regions of the transistors. These devices are then connected through the gate level or the metallization levels to form the inverters and other circuits as described in Sec. 1.

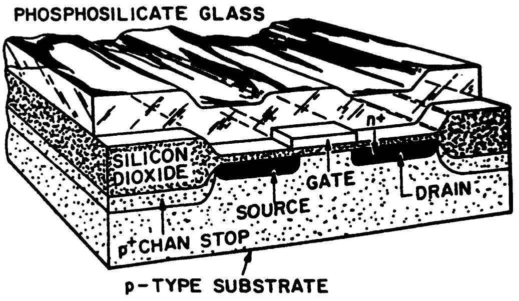

FIG. 7. Cross section of an NMOS transistor showing the structure and the individual parts (from Beadle *et al.*, 1984).

2.2 Bipolar

Bipolar devices are processed such that the structure is formed vertically into the silicon and with the critical dimensions for

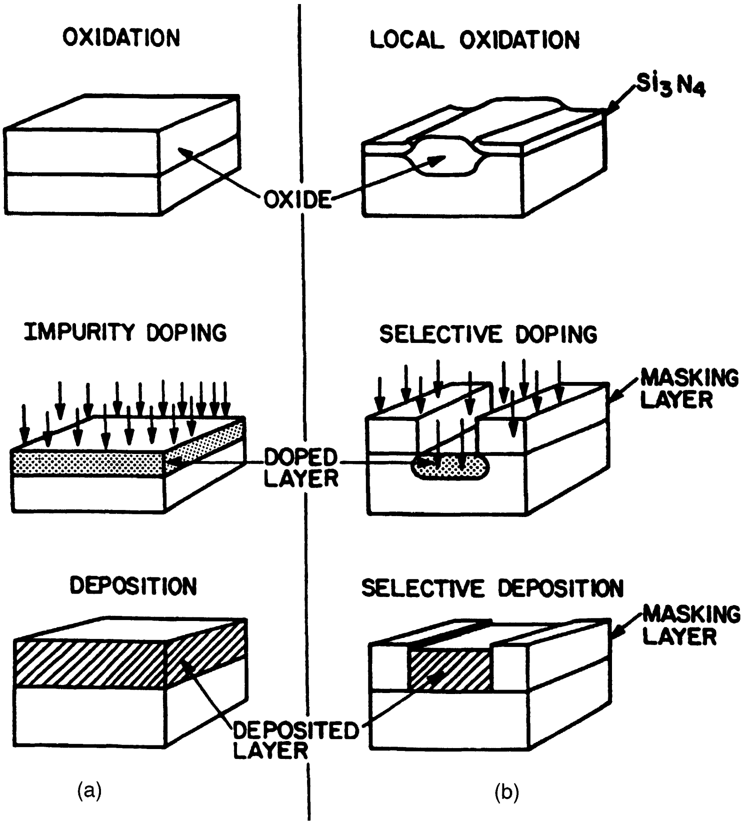

FIG. 8. Techniques in silicon processing to create either (a) uniform layers or (b) selective layers in silicon.

gain and performance in the vertical direction. The cross-sectional view of a bipolar structure is shown in Fig. 10(f), indicating the individual parts of the device. The formation sequence is similar to that of the CMOS process except that the base doping and thickness are critical parameters for the operation characteristics of the bipolar device. This requires the deposition of an epitaxial silicon layer to form a relatively lightly doped region of silicon in which the base of the transistor is formed. The doping levels of the bipolar transistor determine the gain of the device. The relationship between the doping level and the gain is determined by calculating the density of electrons and holes in each part of the bipolar devices. The gain is related to the doping in the emitter and the base of the device through the relationship

$$\beta \sim N_E/Q_B, \tag{5}$$

where N_E is the emitter doping, and

$$Q_B = \int_0^W N(x)dx, \tag{6}$$

where W is the width of the base. Q_B is called the Gummel number of the base. Equation (5) illustrates how the gain of the device is increased by maintaining a high emitter doping and a low base doping.

The most common structure for bipolar devices in VLSI/ULSI technology is the oxide-isolated device. The basic structure of this device is shown in Fig. 10(f). In the VLSI and ULSI applications, it is usually used in conjunction with a CMOS technology sometimes referred to as BiCMOS.

3. MANUFACTURING TECHNIQUES

3.1 Implantation and Dopant Diffusion

Diffusion of dopant through the silicon lattice is a process of random motion of particles in three dimensions. The dopant is dif-

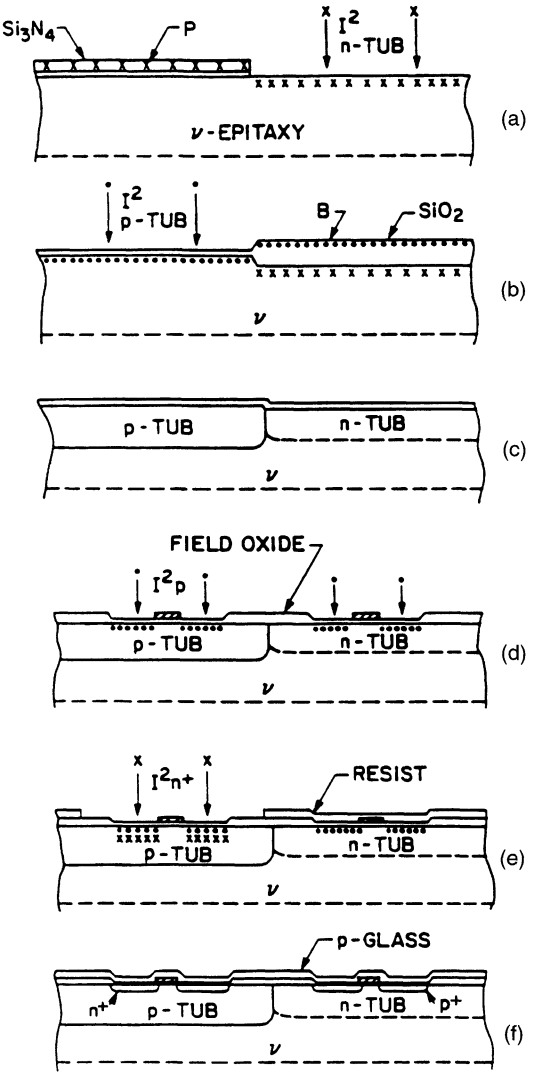

FIG. 9. Basic process sequence to make the CMOS device structures.

fused in the region of the device structure where it is necessary to produce the proper dopant profiles for device operation. The typical dopant species used in silicon VLSI are boron, for forming acceptor sites, and either arsenic, phosphorus, or antimony for forming donor sites. These dopants are used to form the *p*- and *n*-type regions of the semiconductor. The usefulness of each of the dopants is determined by a combination of the solid solubility, diffusion coefficients, and the compatibility of these dopants with the thermal cycles and materials used in the rest of the manufacturing process. Most of the impurity doping for modern VLSI devices is done through ion implantation of dopants, using ion accelerators that implant ions into silicon at energies somewhere between 10 and 300 keV. Table 1 shows the range of various ions illustrating the range that can be implanted. The implant energy would be chosen to place the dopant in the silicon such as to produce the correct distribution for the device characteristics required. The dopant will experience subsequent heat cycles that will diffuse the dopant into the silicon.

The dopant diffusion rates will vary considerably depending on the particular dopant and are the result of the random motion in three dimensions. The diffusion mechanism of these dopants is through an interaction between the dopants and point defects. Point defects can be vacancies in the crystal lattice, interstitials, or interstitial defects. Interstitials are atoms residing in between the lattice sites, and interstitial defects consist of two atoms that are in interstitial sites with a substitutional site between them.

The diffusion of these impurities can be described by Fick's laws regardless of what diffusion mechanism is involved. Fick's first law relates the number of particles passing through a unit area per unit time to the concentration gradient:

$$J_n = -D\frac{dn}{dx}. \tag{7}$$

Fick's second law of diffusion is based on the assumptions of conservation of matter and that diffusion is only a function of distance x and time t. If you also assume that the diffusivity at a given temperature is constant, then you get Fick's second law, which is

$$\frac{\partial N(x,t)}{\partial t} = D\frac{\partial^2 N(x,t)}{\partial x^2}. \tag{8}$$

The solutions of this equation depend on the boundary conditions, but one solution assuming a constant surface concentration is

$$C_s = C(0,t)\exp(-x^2/4Dt), \tag{9}$$

where C is the concentration of dopant and

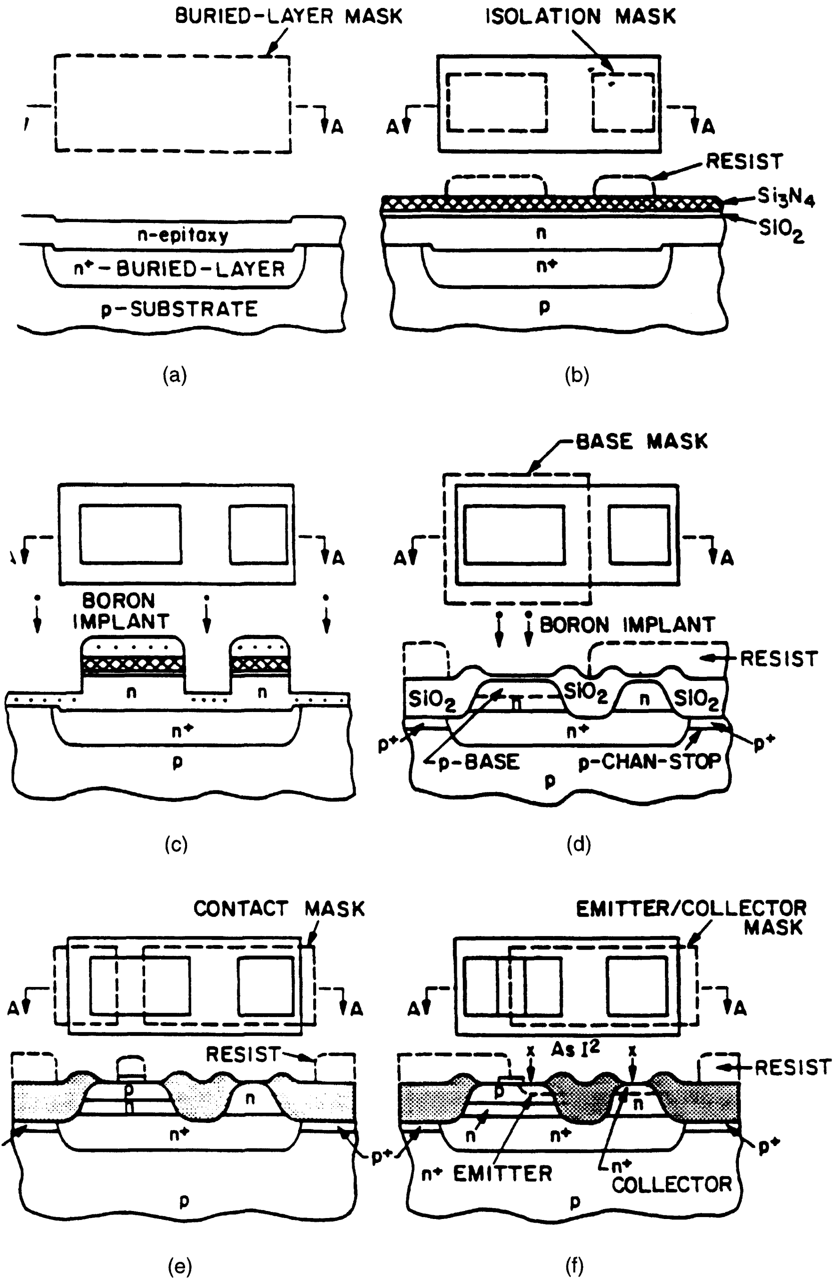
BURIED-LAYER MASK
A
A
n-epitaxy
n⁺-BURIED-LAYER
p-SUBSTRATE
(a)
ISOLATION MASK
A
A
RESIST
Si₃N₄
SiO₂
n
n⁺
p
(b)
A
A
BORON IMPLANT
n
n
n⁺
p
(c)
BASE MASK
A
A
BORON IMPLANT
RESIST
SiO₂
SiO₂
n
SiO₂
n⁺
p⁺
p⁺
p-BASE
p-CHAN-STOP
p
(d)
CONTACT MASK
A
A
RESIST
p
n
n
n⁺
p⁺
p
(e)
EMITTER/COLLECTOR MASK
A
A
As I²
RESIST
p
n
n
n⁺
p⁺
n⁺ EMITTER
n⁺ COLLECTOR
p
(f)

FIG. 10. Basic process sequence to make the bipolar device structures.

C_s is the concentration at the surface, which is held constant.

Another important aspect of diffusion is the temperature dependence of the diffusivity. The experimentally determined temperature dependence of the diffusivity is

$$D = D_0 \exp(-E_a/kT). \tag{10}$$

Equation (10) is the same as the equation describing heat conduction through a solid. The term E_a is the activation energy and k is the Boltzmann constant. D_0 is related to the

Table 1. Range of implants in silicon (in microns).

Energy (keV)	20	40	60	80	100	150	200	250	300
Implanted ion									
B	0.062	0.13	0.19	0.25	0.31	0.44	0.56	0.67	0.78
P	0.025	0.049	0.073	0.098	0.12	0.19	0.25	0.32	0.38
As	0.015	0.027	0.037	0.048	0.058	0.084	0.11	0.14	0.17
Sb	0.014	0.023	0.031	0.039	0.046	0.063	0.080	0.10	0.12

atomic jumping frequency or the lattice vibration frequency of the solid.

These simple equations describe the diffusion of impurities when the impurity concentration is low (smaller than the intrinsic carrier concentration n_i) and the conduction gradient is not large. If the concentrations are higher than n_i, then more complicated interactions with the point defects and different impurities causes the diffusion to become more difficult to model, and it is not simply represented by Fick's laws. This is called extrinsic diffusion. The extrinsic effects can also be caused by reactions at the surface of the silicon.

3.2 Thermal Oxidation

Thermal oxidation is the technique of forming a SiO_2 layer on a silicon surface by exposing the silicon to an oxidizing ambient at an elevated temperature. The typical method for forming this film is to insert the silicon wafers in a cylindrical quartz tube in a resistance-heated furnace that contains a source of either oxygen gas or water vapor. The temperature of oxidation is typically in the range of 800–1200 °C. The chemical reaction occurring on the surface of the silicon causes the silicon on the surface to be converted into SiO_2 with the ratio of the thickness of SiO_2 to silicon consumed equal to 2.3.

The properties of thermally grown SiO_2 are probably the single most important factor in the progress made in VLSI/ULSI technology. It is the interface of the Si/SiO_2 in the gate of the MOS transistor that determines the important electrical characteristics of the MOS device. The ability to grow defect-free films that can reliably withstand the high electric fields and resist damage from the other processing steps also determines the rate of development of the ULSI technologies.

3.3 Lithography

Lithography is the process through which a pattern is delineated in a layer of material in which the physical properties can be changed through exposure to photons, electrons, or ions. The most common application is to use a photosensitive material called a photoresist. This is a material that can be developed like a photosensitive emulsion to form regions in which the resist remains on the surface and other regions where the resist is removed. The area where the resist is removed exposes the layers below to chemical etching or doping. A schematic diagram of the basic projection-reduction optical lithography system is shown in Fig. 11. This consists of a light source with a mask and lens arrangement to focus an image onto the resist surface. This is the most common technique used for ULSI technologies today.

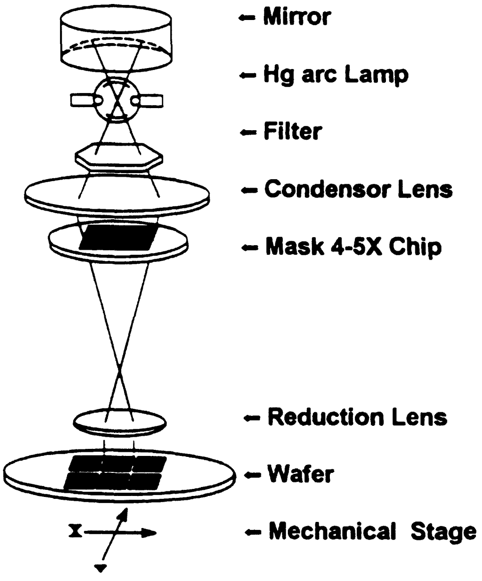

FIG. 11. Schematic of a reduction step-and-repeat projection system.

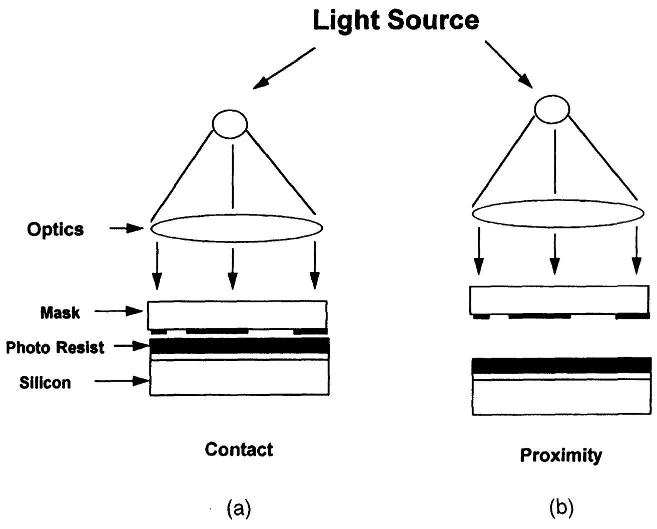

FIG. 12. Proximity and contact lithography.

The optical lithography techniques can be classified into three categories: contact, proximity, and projection. The contact and projection systems are schematically shown in Figs. 12(a) and 12(b). The disadvantages of contact and proximity systems are defect generation and the inability to make the fine features on the masks.

The most commonly used light source is ultraviolet (UV) light, formed in high-pressure mercury-vapor light arc sources that produce high-intensity emission in the wavelength range of 300–600 nm. The optical lithography generations are usually referred to by the particular lines of the emission spectrum that are used. Figure 13 shows the output spectrum of the high-pressure mercury arc source indicating the g, h, and i lines of this spectrum. The shorter the wavelength of the light, the finer the image size is. As the even shorter wavelengths become needed to pattern the very small features, the light sources will change to deep UV (DUV) light sources that use an excimer laser to produce the light at a wavelength of 248 nm. Beyond the use of 248 nm, the next optical wavelength that will be used will be 193 nm.

These optical lithography systems become increasingly difficult to manufacture at the shorter wavelengths and may become prohibitively difficult so that another lithography technique may be used. Possible alter-

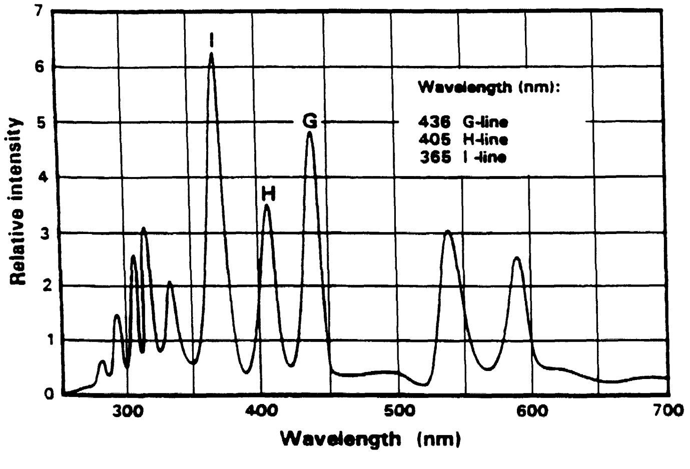

FIG. 13. Spectrum of high-pressure mercury-arc source indicating the high-intensity lines used for optical lithography.

natives such as using synchrotron radiation as a source of x rays are under development and may be useful in the future.

Another lithography technique that has its use in VLSI technology is electron-beam lithography. Electron-beam lithography uses accelerated electrons instead of photons to expose resist and to create the pattern. The use of electrons instead of photons allows the ultimate feature size to be reduced to below 0.1 μm. The disadvantages of *e*-beam lithography are the relatively more expensive equipment and the slow writing time for the technique. This means that this technique is used for prototyping and laboratory devices but is not commonly used for production of ULSI products. See MICROLITHOGRAPHY for more information on these techniques.

3.4 Deposition

The deposition of thin films of material to make very uniform and defect-free films of metals, insulators, and semiconductors is an essential part of the manufacture of silicon devices. The most important materials for manufacture of VLSI devices are polysilicon, silicon dioxide, silicon nitride, aluminum, various refractory metals such as titanium and tungsten, and metal compounds such as WSi_2 and $TiSi_2$. These materials are deposited by two basic techniques: chemical vapor deposition (CVD) and physical vapor deposition (PVD). There are also some applications that are combinations of CVD and PVD.

3.4.1 Chemical Vapor Deposition (CVD) Chemical vapor deposition (CVD) [see CVD (CHEMICAL VAPOR DEPOSITION)] is a process of depositing a film by causing a chemical reaction of a gas mixture or a decomposition of a gas. This is done at high temperatures on the wafer surface. The process can be performed at atmospheric pressure or at low pressure and is then denoted as APCVD or LPCVD, respectively. The general configuration of a CVD deposition system illustrating the various configurations is shown in Fig. 14 with the illustrations of the hot-wall reduced-pressure reactors and the continuous-deposition atmospheric reactors.

There are also several ways to enhance the deposition of films using light or plasma energy to enhance the deposition rate or film quality. These are referred to as photo-CVD and plasma-enhanced chemical vapor deposition (PECVD). Two configurations for the plasma-deposition reactors are shown in Fig. 15. The relative advantages of each deposition technique depend on the particular material properties of the thin film and the trade-offs that must be made between particle generation, step coverage, and the cost of deposition. More detailed explanations of the chemical processes can be found in CVD (CHEMICAL VAPOR DEPOSITION).

3.4.2 Physical Vapor Deposition (PVD) There are several important films used in VLSI technology that are applied using physical vapor deposition. Physical vapor deposition is done under a vacuum using a sputtering or an evaporation technique. The sputtering techniques (see SPUTTERING) used usually involve the acceleration of argon atoms through a potential gradient and the collision of these ions onto a target material at the cathode of the ion acceleration cham-

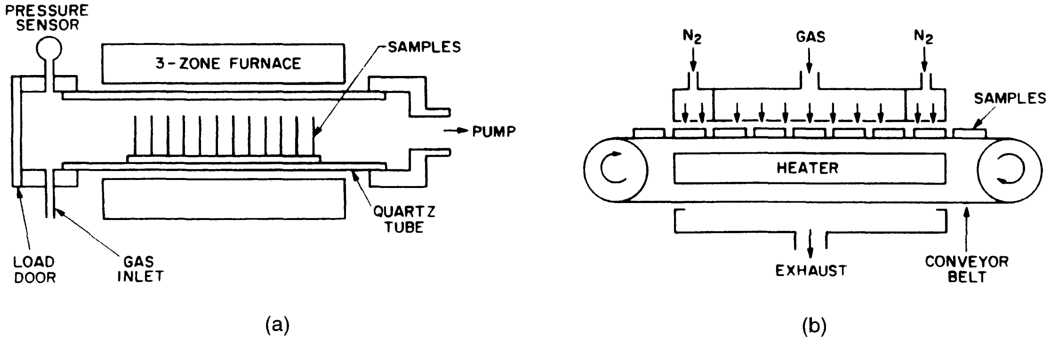

FIG. 14. Diagrams of two types of CVD reactors used to deposit dielectrics: (a) hot-wall, reduced-pressure reactor; (b) continuous atmospheric-pressure reactor. (From Sze, 1988.)

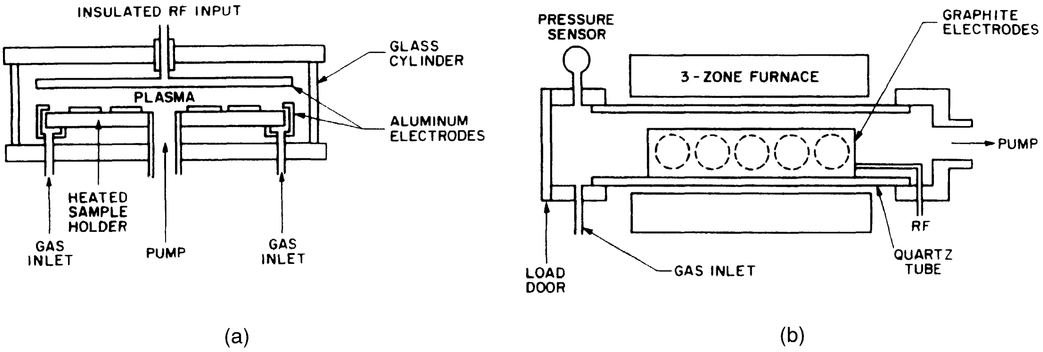

FIG. 15. Diagrams of plasma-deposition reactors: (a) parallel-plate, (b) hot-wall. (From Sze, 1988.)

ber. The ion causes the atoms of the target material to be dislodged, and the dislodged target atoms travel through the vacuum until they impinge upon the substrate that is receiving the film deposition. The ion energy must be sufficient to dislodge the atoms. A measure of the efficiency of this process is called the sputter yield. This number represents the number of atoms sputtered from the surface per incident ion. This number varies from about 0.5 to 1.5 depending on the material and the geometry of the targets.

Vacuum evaporation techniques are used when the source material can be easily heated above its melting point in a vacuum chamber allowing a direct evaporation of the material onto the surface receiving the thin-film deposition. Aluminum can be deposited this way. The rate of evaporation of the material depends on the temperature and the vapor pressure of the evaporated material. See THIN-FILM DEPOSITION.

3.5 Etching

Etching is the removal of material from the wafer surface either selectively, such as where photoresist is used as an etching mask, or globally, such as for cleaning a film or removing a layer of material.

3.5.1 Wet Etching Wet etching consists of the wafers being immersed in a liquid reagent to achieve a controlled chemical reaction that results in the particulate being suspended in the liquid or the material dissolved. The selective etching techniques needed to make modern high-density VLSI and ULSI circuits have become one of the critical enabling techniques for the progress of processing the features needed for these devices. Wet chemical etching is widely used for cleaning and for global removal of films, but the definition of the smallest features in modern devices is almost always done with a dry etching technique.

3.5.2 Dry Etching Figure 16 illustrates the difference between a pattern produced by wet etching, which is isotropic, and one produced by dry etching which is anisotropic. This is the reason the dry etching techniques allow the definition of the very smallest features. The wet-etched pattern allows for etching to occur under the masking layer, resulting in loss of control of the patterned feature size.

Dry etching techniques cover many different chemistries and equipment geometries, choice among which is determined by the selectivity, etch rates, uniformity, reproducibil-

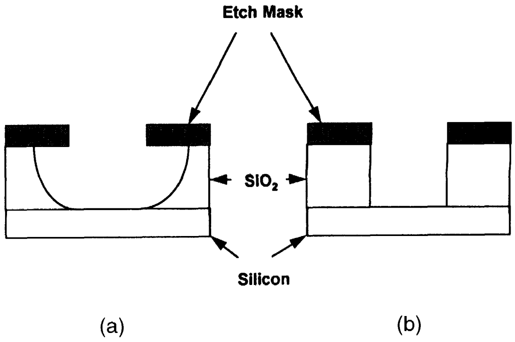

FIG. 16. Cross-sectional diagram of (a) the isotropic etch typical of a wet-etch process and (b) the anisotropic etch achieved by dry etching techniques.

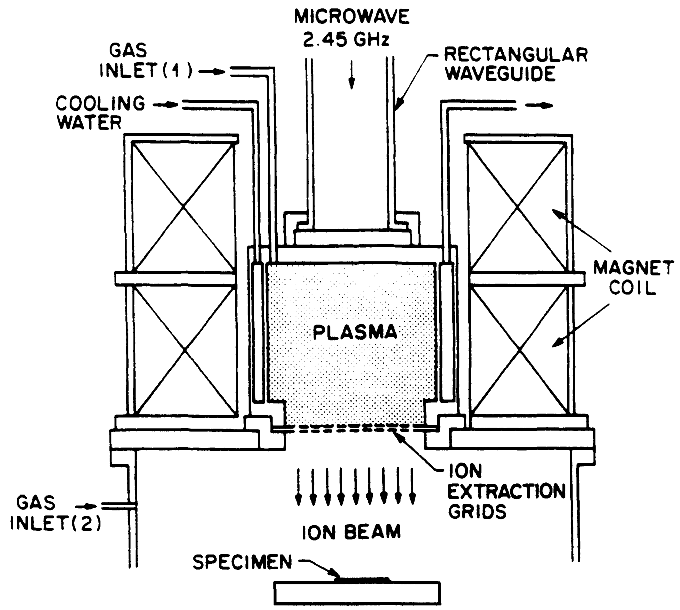

FIG. 17. Reactive ion-beam etcher using an electron cyclotron-resonance chamber as an ion source (from Sze, 1988).

ity, and cost. All of these factors are part of the decision made on what the final etch would be for a particular step in the process. Most commonly used dry-etching techniques use a low-pressure gas in the form of a plasma, leading to the term "plasma etching" or more commonly "reactive ion etching" (RIE). A schematic of the chamber of a RIE system is shown in Fig. 17. The particular gases used in the RIE of various materials are shown in Table 2 (Sze, 1988), which illustrates the variety of chemical reactions necessary to obtain the correct etch characteristics. More details of these etching techniques can be found in PLASMA ETCHING.

3.5.3 Chemical Mechanical Polishing

A technique that has recently come into use in the VLSI silicon industry is the application of chemical mechanical polishing (CMP) to achieve a flat surface across the entire wafer. This is an important quality of the surface in order to insure that the photolithography tools can print the finest features on the wafer surface. The surface of the silicon starts out flat, but topography develops as the dielectric insulating layers are deposited on top of the patterned metal layers. These nonuniformities in the surface due to the successive layers of metal and dielectric can be removed by polishing the surface with equipment that holds the wafers against a rotating polishing pad that is wetted by a suspension of SiO_2 in a high-*p*H liquid. The polishing process removes the topography since higher pressure is applied to the elevated portions of the wafer than to the recessed portions.

3.6 Constraints for VLSI Manufacturing

3.6.1 Process-Control Methodologies

The ability to control individual process parameters to the degree necessary to insure adequate yield is referred to as process control. The ability to quantify process control is very important for the complicated processes required for ULSI silicon manufacturing. It

Table 2. Source gases used for reactive plasma etching of silicides, silicon, and SiO_2.

Silicides	Si	SiO_2
$CFCl_3$	Cl_2	CHF_3
CF_2	F_2	$CF_4 + O_2$
CCl_4	HF	$CF_4 + H_2$
$BCl_3 + Cl_2$	$CFCl_2$	$SiCl_4$
$CF_4 + O_2$	CF_2Cl_2	
SF_6	CCl4	
NF_3	$BCl_3 + CL_2$	
	$CF_4 + O_2$	
	SF_6	
	NF_3	

is particularly important to quantify the contribution individual process steps make to the yield of the total process. One methodology will be described here that is in widespread use throughout the industry. It consists of measuring many parameters and characterizing the statistical distribution of these parameters. The specific characterization employs quantities called process-control indices and denoted by C_p and C_{pk}. The basic definition of C_p and C_{pk} is described by first defining a few terms:

USL, upper specification limit. This is the upper limit of a process-parameter value that will still yield working devices.
LSL, lower specification limit. This is the lower limit of a process-parameter value that will still yield working devices.
USL − LSL = design specification width.

It is assumed that the distribution of the values for the process is Gaussian with a mean value of X. The width of the distribution can be described with the value σ. The definition of C_p is

$$\begin{aligned} C_p &= \text{(design specification width)/(process width)} \\ &= (\text{USL} - \text{LSL})/6\sigma. \end{aligned} \tag{11}$$

This is a unitless quantity that would have a value of 1 when the total specification range is equal to 6σ. This gives no information about whether the distribution is centered or not, and so another parameter, k, is introduced. This is obtained by use of a quantity C_{pk}, which is defined as the smaller of

$$C_{pU} = (\text{USL} - X)/3\sigma \tag{12}$$

and

$$C_{pL} = (X - \text{LSL})/3\sigma; \tag{13}$$

then

$$C_{pk} = (1 - k)C_p \tag{14}$$

where

$$k = \frac{\text{process shift}}{\text{design specification width}}.$$

Figure 18 shows a schematic representation of these parameters for a typical Gaussian distribution of data points. In a typical ULSI process sequence there will be well over a hundred parameters that are measured and evaluated. If one of these parameters has a distribution that yields a $C_p = 1$ and it is centered, then there would be 1350 ppm that would be out of specification due to that process step alone. In practice it is found that in order to manufacture ULSI silicon devices economically, process control must be such that a $C_p = 2$ is achieved with a $C_{pk} = 1.5$ or greater. This requires very tight controls on the process steps.

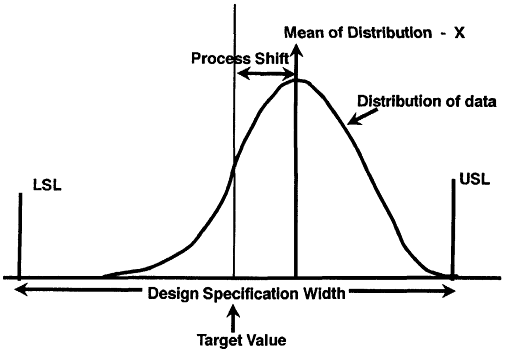

FIG. 18. Graphical representation of the components of the calculation of the process capability index.

A direction that will be common for the factory of the future is to have active control of the processes. This is used to compensate for the variation in the individual steps by adjusting subsequent steps. One example is shown in Fig. 19 in which the effect of active control is shown in one step of the process (Chattergee and Larrabee, 1993). This figure shows the result of having an active feedback for a particular step (polysilicon etching) that will adjust for differences in the material and etch rates. This allows the process-control problems of the polysilicon deposition step to be compensated for during the gate etching.

3.6.2 Defect Control Defect control is becoming the most important parameter in the choices made for the process sequences and the equipment selection. This means that much of the design and development resources go toward making the process as defect resistant as possible. Control of these defects has become one of the controlling

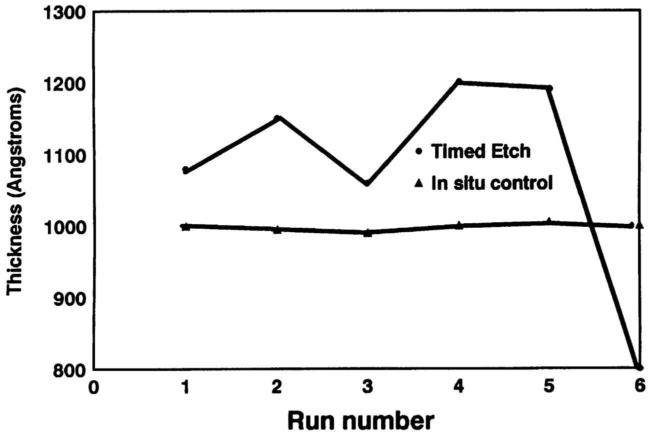

FIG. 19. Etch measurements for polysilicon etch for two different sets of runs without and with *in situ* control of the etch (from Chattergee and Larrabee, 1993).

parameters governing the rate of growth of the silicon device. Figure 20 shows the historical line of learning curves for a variety of memories (Stapper, 1990), which shows a curve of economic necessity that is needed to have a competitive manufacturing capacity. Over the last 10 years the necessity to control defects has required particle densities to be reduced by more than an order of magnitude. The size of the "killer defects" has also scaled with the minimum geometries, which compounds the problem, as shown in Fig. 21. Figure 21 also shows the detection limits that are currently limiting the ability to monitor defects during the processing. The detection limits are above the levels necessary to produce high yielding devices.

A table of defect mechanisms is shown in Table 3 to give an idea of how defects affect the yield of the process (Drum, 1991). This table shows the critical geometry present in a medium-sized VLSI circuit along with defect densities expressed in D_0 = defects/area and N_0 = defects/contact. The D_0 is set to 1/cm^2 or 1 ppm for contacts for this example. This is a technique that is used to model the projected yield of a device. This is done by determining the probability of the defect densities for each level of the process to cause a fatal failure. The yield loss of each individual mechanism may be quite small, but the product of these losses can result in yields close to zero.

This example shows that with one defect per cm^2 or ppm, the yield of a medium-sized

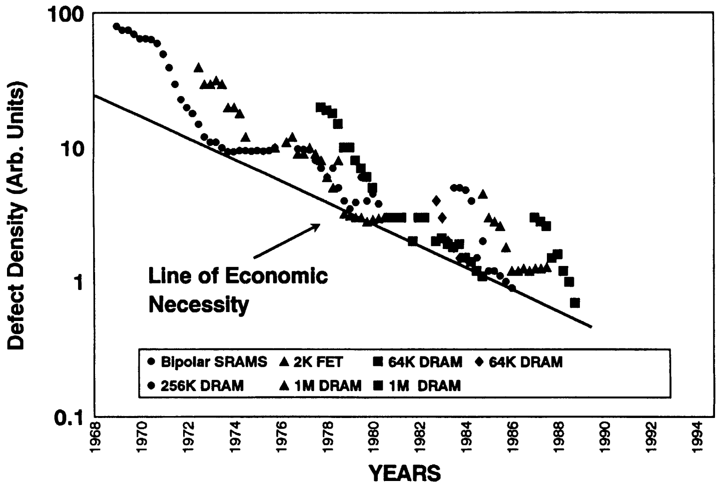

FIG. 20. Defects in memory production showing line of economic necessity for various generations of DRAMS (from Stapper, 1990).

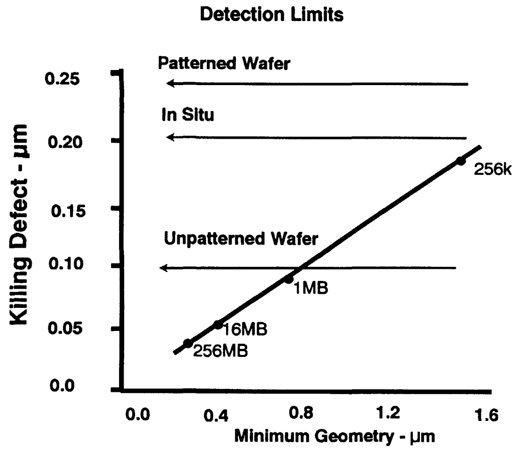

FIG. 21. Chart of the size of killing defects in the memory generations as a function of the minimum geometry. Detection limits are also shown for indicating current capabilities. (From Chattergee and Larrabee, 1993.)

circuit can be about 55%, which would not be acceptable for a modern manufacturing line.

The source of defects is changing as the geometries shrink. The dominant source of defects will be the equipment used to manufacture the devices and the processes chosen. The shift in sources of particles over the last 10 years is shown in Fig. 22 (Chattergee and Larrabee, 1993). The particles used to be coming from the people and the room. The problem is compounded by the fact that these smaller killer defects are more difficult and may be impossible to remove once they are attached to the silicon wafer. This requires the fundamental mechanisms of particle generation to be understood and controlled.

3.7 Reliability and Wearout

A discussion of VLSI devices cannot be complete without some discussion of reliability issues and wearout mechanisms. Although the materials used in making these devices are some of the most stable in the world, the fact that millions of devices are packed into a single area of silicon means that the one part per million failure mechanisms will be fatal for these devices. There are three main device wearout mechanisms that are a concern for silicon devices: hot-carrier charging, dielectric breakdown, and electromigration. These are all quite different in the physics involved but all result in a time-dependent failure mechanism.

3.7.1 Hot-Carrier Aging Hot-carrier aging is a problem for MOS devices where the gate oxide of the MOS device becomes charged by trapped charge at the interface of

Table 3. Yield loss for a medium-sized circuit showing the yield fraction loss for each type of defect. "Win" refers to the contact from the first level of metal (M1) to the silicon or the gate of the transistor. IL refers to interlevel shorts. Via is the connection between the first level of metal and the second level of metal.

Mechanism of random yield loss	Critical geometry	D_0 or N_0	Yield fraction
Metal-2 shorts	0.95 (m)	1.0 (/cm^2)	0.967
Metal-1 shorts	1.14	1.0	0.972
Silicide shorts	0.74	1.0	0.982
IL-M2-M1 shorts	0.0108 (cm^2)	1.0	0.989
IL-M1-Gt shorts	0.016	1.0	0.984
Nch-thinox-shorts	0.506 (m)	1.0	0.977
Pch-thinox-shorts	0.4337	1.0	0.981
Metal-2 opens	1.8207	1.0	0.938
Metal-1 opens	2.2645	1.0	0.883
Silicide opens	0.6875	1.0	0.963
Gate-oxide-shorts	0.0557 (cm^2)	1.0	0.946
N + junct'n leakage	0.0203	1.0	0.980
P + junct'n leakage	0.0128	1.0	0.987
Win-silicon-opens	0.0507 (million)	1.0 (ppm)	0.951
Win-gate-opens	0.0480	1.0	0.953
Via-opens	0.0412	1.0	0.960
Product			0.55

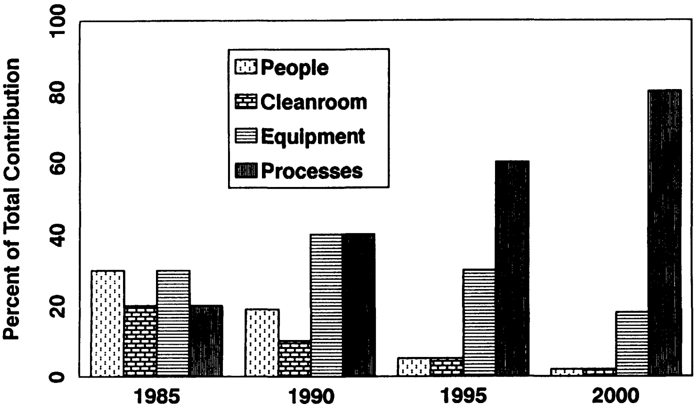

FIG. 22. Chart of the source of contamination over the last 10 years of manufacturing (from Chattergee and Larrabee, 1993).

the silicon and SiO_2. This problem arises as the device dimensions are reduced and the supply voltage is not reduced at the same rate. The lateral electric field generated in MOS devices increases, and if it becomes large enough, it can cause the hot-carrier effects. This has been a major concern in the design of VLSI MOSFETS. The problems associated with this effect are schematically shown in Fig. 23. The charge that is injected into the oxide in this situation can then cause a reduction in the V_t and the transconductance of the transistor. These changes can then cause the circuit to fail when these parameters fall below the values needed for proper operation.

One method that is commonly used to reduce the hot-carrier effects for the MOS devices is to introduce an extension of the drain region of the transistor, which has the effect of reducing the field in the drain region and therefore reducing the hot carriers. This extension is called a lightly doped drain and is schematically shown in Fig. 24.

3.7.2 Dielectric Breakdown Dielectric breakdown of thin oxides is a particularly important concern for the high-density DRAMS. The breakdowns that cause failure in these dielectrics are induced by electric fields. The breakdown can be an intrinsic property of the oxide film or can be due to small weak spots in the oxides caused by contamination, stacking faults, traps, asperities, etc. Any of these defects can cause a local minimum in the dielectric strength of the

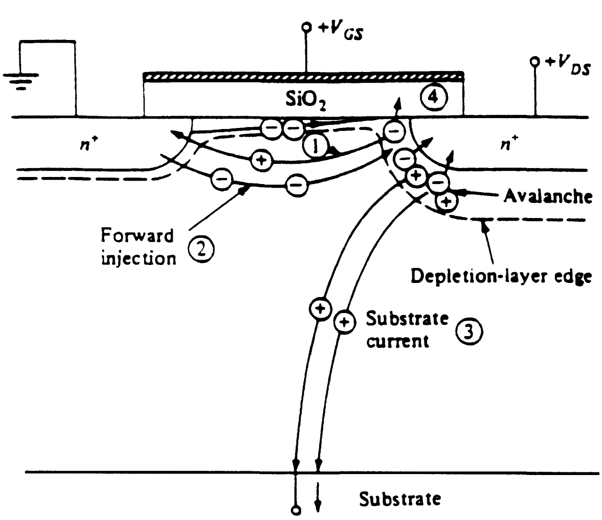

FIG. 23. Illustration of the generation of hot carriers in NMOS devices causing charging of the gate oxide.

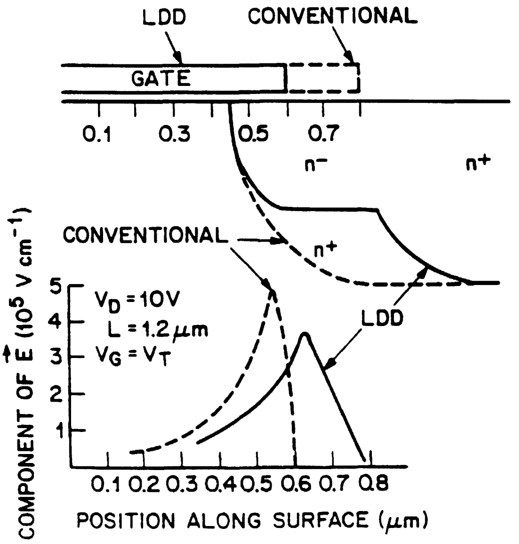

FIG. 24. Cross section of a device having an LDD structure and the electric field in the drain regions (solid lines). The dashed lines show the conventional device cross section and field. The *n*+ region for the conventional drain extends further under the gate than the *n*− region of the LDD structure. The *n*+ region of the LDD structure is formed with a mask (not shown).

oxide. These local minima will then fail in either a field-dependent or a time-dependent stress of the oxide.

3.7.3 Electromigration Electromigration is the phenomenon of movement of atoms caused by the collision of electrons with atoms. The material transport due to this movement of the atoms can cause voids to form in the metal lines used for the interconnect in modern silicon devices. This is particularly important in the modern VLSI devices, in which aluminum lines are used for the interconnect. The aluminum is gathered in the direction of the current flow, leading to a discontinuity in the current-carrying line. The transport mechanism can be much faster in thin-film aluminum lines than in the bulk metal because of the presence of grain boundaries, dislocations, and point defects. This rate of transport is therefore strongly dependent on the method of deposition, the alloying elements, the film thickness, and the temperature of the substrate during deposition.

Several techniques have been used to reduce the risk of failure of an aluminum line due to electromigration. One is simply to limit the current density of the aluminum lines as a design constraint so that it is always below the level to cause significant transport. This works but results in having metal lines that are too big. Another technique involves forming metal grains that have grain boundaries perpendicular to the current flow. This can reduce the electromigration effects but is quite difficult to implement reliably on a circuit. The most common technique is to form thin layers of a refractory material such as TiN on either the top, bottom, or both top and bottom of the aluminum lines. This insures that there is electrical continuity in the metal line even if there is a void formed in the aluminum.

4. APPLICATIONS

4.1 Memory Devices

The most widely used ULSI circuit application is the memory circuit. The memory circuit consists of many cells that store information in the form of individual binary bits. Memory chips come in a variety of forms but can be broadly grouped into three major categories: dynamic random-access memories (DRAM), static random-access memories (SRAM), and nonvolatile read-only memories (ROM). Nonvolatile memories come in programmable versions called PROMs, and electrically programmable versions called EPROMs, and electrically erasable versions called EEPROMs. Each of these types of circuits has specific uses, and in most computers and communication devices you will find all of these varieties used.

4.1.1 Dynamic Random-Access Memory (DRAM) Circuits A DRAM device is the most compact of all of the IC's and is the circuit that will be the first to take advantage of the advances in processing technology. The reason for this is that the basic building block of the DRAM circuit is the memory cell, which consists of a single transistor and a single capacitor. The basic structure is shown in Fig. 25. The cell is called a dynamic memory cell because the memory state is stored by charging the capacitor. The entire memory has to be periodically recharged to compensate for the loss of charge

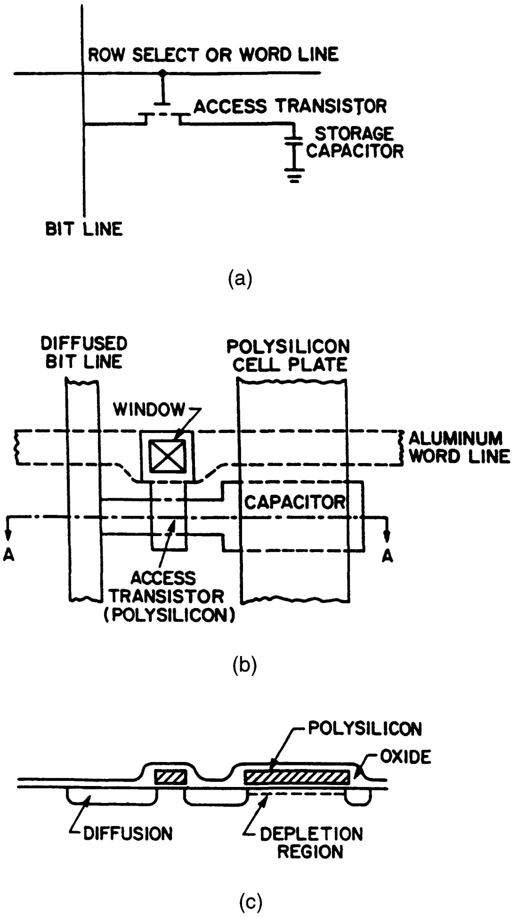
ROW SELECT OR WORD LINE
ACCESS TRANSISTOR
STORAGE CAPACITOR
BIT LINE
(a)
DIFFUSED BIT LINE
POLYSILICON CELL PLATE
WINDOW
ALUMINUM WORD LINE
CAPACITOR
A
A
ACCESS TRANSISTOR (POLYSILICON)
(b)
POLYSILICON
OXIDE
DIFFUSION
DEPLETION REGION
(c)

FIG. 25. Diagram of the device, top view, and cross section of the basic DRAM cell.

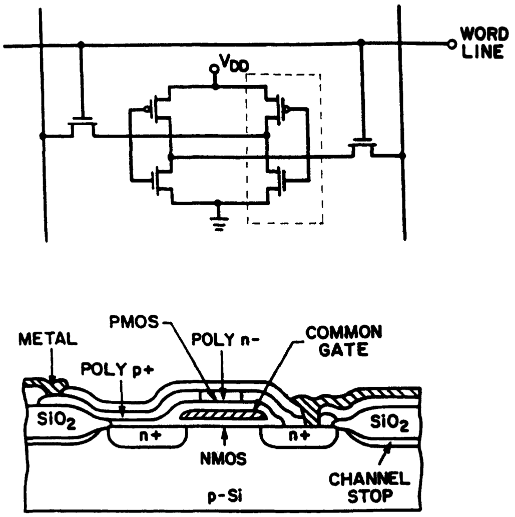
WORD LINE
V_{DD}
METAL
PMOS
POLY n-
COMMON GATE
POLY p+
SiO_2
n+
n+
SiO_2
NMOS
p-Si
CHANNEL STOP

FIG. 26. Diagram and cross section of the six-transistor SRAM cell showing the thin-film transistor. Box encloses the transistors shown in the cross section.

on the capacitor and maintain the data. The advantage the DRAM has is that the cell can be made extremely small and therefore supply the greatest density of RAM.

4.1.2 Static Random-Access Memory (SRAM) Circuits Static random-access memories consist of a more complicated cell structure consisting of six transistors. The cell is static in that it does not require a refresh cycle to maintain the information. The state of the cell is maintained as long as the power is supplied to the circuit. The cell consists of two inverters, each consisting of an NMOS and a PMOS transistor, and two access transistors. The schematic is shown in Fig. 26(a). The function of the cell is to drive one inverter high and the other low or vice versa to produce the high or low (0 or 1) memory state. The advantage of the SRAM is that reading and writing of the cells can be done faster than can be done with a DRAM. The disadvantage is that the cell takes up more space and therefore allows for less memory per unit area of silicon.

The most significant processing difference between the SRAM technology and other ULSI technologies is the need for the compact six-transistor cell, which for the most modern devices requires the use of thin-film transistors (TFTs). These TFTs allow the six-transistor cell to be folded to allow the cell area to be reduced by stacking transistors on top of one another. The cross section of the cell that results from this arrangement is shown in Fig. 26(b).

4.1.3 Nonvolatile Memories Nonvolatile memories are able to maintain the information stored even after the power supplies are disconnected. This is achieved by particular structures that allow information to be permanently stored by programming the information at the point of manufacture. These devices are called read-only memory (ROM). Other ROM devices can be programmed in the field and these are called programmable ROM or PROM. PROM devices usually require links to be fused for the connectivity, but once the links are fused, the device looks like a ROM. The last category of nonvolatile memory is the erasable PROM or EPROM. These devices have another transistor structure that allows charge to enter into the dielectric of a MOS transistor, which causes the transistor to be permanently on. This information is maintained until the device is exposed to UV or x-ray light. Another version of an EPROM is the electrically erasable EPROM or EEPROM. This device uses a two-transistor cell that allows the cell to be selected and selectively erased.

4.2 Microprocessors and Logic

Other than memories, microprocessors (*q.v.*) and other logic devices such as digital signal processors (DSP), application-specific logic (ASIC's), and other random logic devices represent the largest fraction of silicon VLSI devices. A microprocessor is defined as a computer central processing unit built on one MOS or bipolar chip with a complexity of at least 2000 equivalent gates. The block diagram of a simplified microprocessor architecture is shown in Fig. 27. The specific requirement for a microprocessor that differs from those of the other applications is the need for low-resistance, high-density interconnects. This need results in a large effort devoted to the material and processing issues involved with making multiple levels of metallization that can interconnect the logic and can function with the minimum of signal delay and power dissipation. The basic technology used is almost entirely CMOS, and the building blocks are individual elements such as inverters; NAND, NOR, and shift registers; adders; and dividers. These are assembled in a complicated architecture that requires millions of transistors. The microprocessor is noteworthy only because of the role it currently plays in driving the technology complexity. Figure 28 shows the complexity increase of the Intel microprocessors over the last 25 years.

5. ECONOMIC ASPECTS AND INDUSTRY SCALING TRENDS

5.1 Device Scaling

The historic trends shown in Figs. 2 and 28 are driven by the economies and the demand for including more functionality on a device. The microscopic impact of this trend drives the device design to deal with the fundamental physical limitations of the materials and device function. This trend at the microscopic level is referred to as device scaling. The first treatment of this was described (Denard, 1974), as the scaling rule where the scaling factor S is shown to affect the MOS device parameters such that

Lateral dimensions $X \rightarrow X/S$
Vertical dimensions $Z \rightarrow Z/S$
Supply voltage $V \rightarrow V/S$

This results in the device/circuit parameters:

Gate delay $\rightarrow 1/S$
Capacitance $\rightarrow 1/S$
Power dissipation/device $\rightarrow 1/S^2$
Power density $\rightarrow 1$

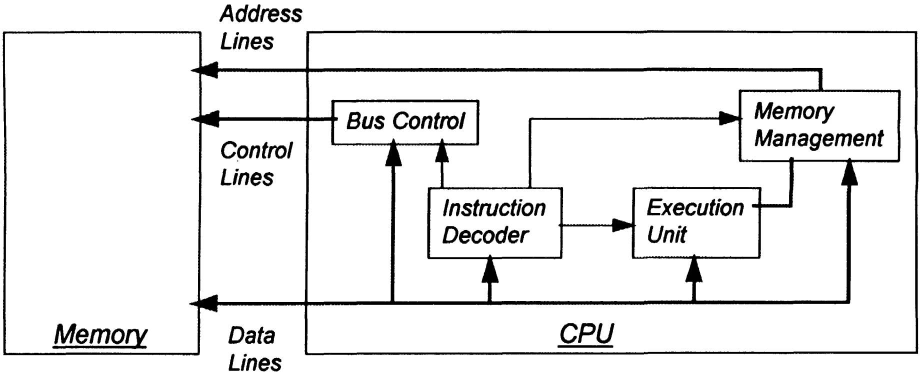

FIG. 27. Simplified block diagram showing the functions performed by a microprocessor.

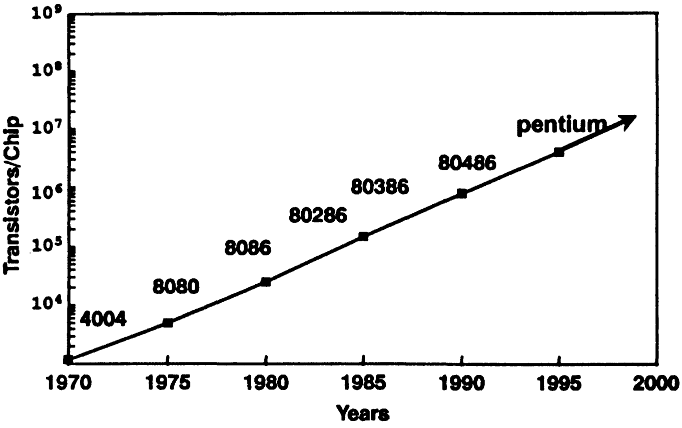

FIG. 28. Projection for transistor density on microprocessors from 1970 to 2000 (from Barrett, 1993).

Packing density → S^2
Speed-power product/device → $1/S^3$

The historic trend has not been toward this monotonic scaling of each part of the device but a scaling of lateral dimensions without a uniform scaling of the supply voltage. This is a result of the needs of the particular applications where the new devices needed to interface with devices from a previous generation and required the same voltages to be used between the chips. The power-supply voltages tended to be changed in steps as seen in Fig. 29. This caused the electric field in the devices (both across the gate oxide and in the drain region) to increase and made the reliability aspects of these high field effects dominate the trend in scaling the devices.

The trend in scaling is now becoming split into two major applications that have different sets of scaling goals. These applications are toward high-speed or toward low-power application. These two conflicting goals have been studied (Hu, 1994), and the studies show that the resulting scaling rules will be different. Projections for the impact that these different scaling rules have on device parameters such as $I_{d,\mathrm{sat}}$ and gate oxide thickness were made. These parameters will be the limiting cases for the ultimate scaling of devices and will determine the ultimate limits for either performance or power. Since these quantities can be individually optimized, the resulting devices will be different depending on the ultimate application.

5.2 Industry Scaling Costs

Although the functionality of the individual chips increases exponentially with successive generations of devices, the cost of

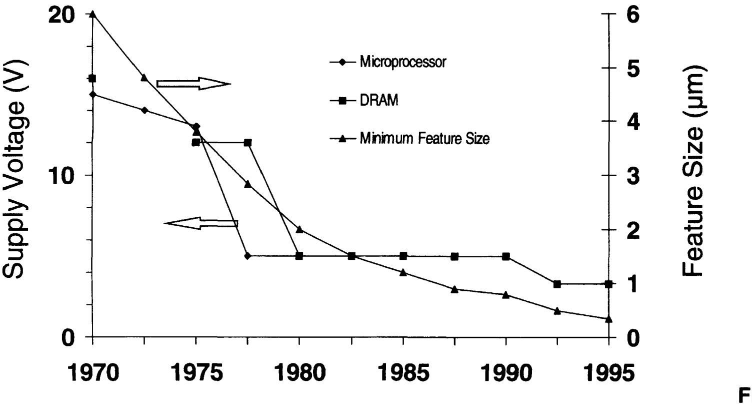

FIG. 29. Power-supply voltage reduction as a function of time.

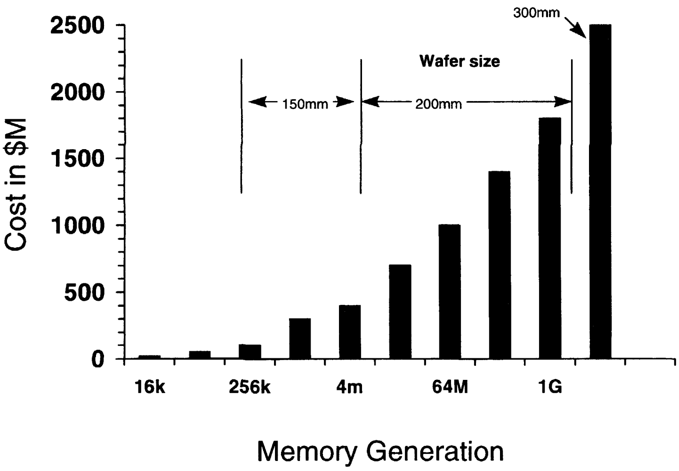

FIG. 30. Wafer fabrication costs as a function of the memory chip generations (from Chattergee and Larrabee, 1993).

manufacturing these devices also increases at an alarming rate. The reason for this is the fact that the individual pieces of equipment become more and more expensive as the technology becomes more complicated and the device dimensions become smaller. This requires the cost of a new manufacturing line to become greater and also requires the lines to be able to produce many more devices to pay for the huge capital investment. Figure 30 (Chattergee and Larrabee, 1993) shows the typical costs of a new line over the years. The cost has gone from about $\$1 \times 10^6$ in 1970 to about $\$1 \times 10^9$ in 1995. The huge costs of these fabrication facilities have driven many smaller companies out of the device fabrication business and create an environment where the device manufacturing for large-volume devices such as DRAM's are done in a few huge factories referred to as "Mega-fabs."

Works Cited

Barrett, C. R. (1993), in: *Symposium on VLSI Technology. Digest of Technical Papers,* Tokyo: Japan Society of Applied Physics, p. 7.

Beadle, W. E., Tsai, J. C. C., Plummer, R. D. (1984), *Quick Reference Manual for Silicon Integrated Circuit Technology,* New York: Wiley.

Broydo, S., Osburn, C. (1987), in: S. Broydo, C. Osburn (Eds.), *Proceedings of the First International Conference on Ultra Large Scale Science and Technology,* New Jersey: The Electrochemical Society.

Chattergee, P. K., Larrabee, G. B. (1993), *IEEE Trans. VLSI Syst.* **1,** 7.

Denard, R. H. (1974), *IEEE J. Solid State Circ.* 1974, 256.

Drum, C. M. (1991), *Proc. of the 1992 Int. Workshop on Defects and Fault Tolerance in VLSI Systems,* New York: Plenum Press, p. 60.

Horiuchi, M., *et al.* (1995), in: *ISSCC 95. Digest of Technical Papers,* New York: IEEE, p. 252.

Hu, C. (1994), *Semiconductor Int.* **17,** 105–114.

Kilby, J. S. (1976), *IEEE Trans. Electron Dev.* **ED-23,** 648.

Komiya, H. (1993), in: *ISSCC 93. Digest of Technical Papers,* New York: IEEE, p. 16.

Stapper, C. H. (1990), in: W. Maly (Ed.), *Defect and Fault Tolerance in VLSI Systems,* Vol. 2, New York: Plenum Press, p. 1.

Sugibayashi, T., *et al.* (1995), in: *ISSCC 95. Digest of Technical Papers,* New York: IEEE, p. 254.

Sze, S. (1988), *VLSI Technology,* 2nd ed., New York: McGraw-Hill.

Further Reading

Sze, S. (1988), *VLSI Technology,* 2nd ed., New York: McGraw-Hill.

Mead, C., Conway, L. (1980), *Introduction to VLSI Systems,* Reading, MA: Addison-Wesley.

VOLCANOLOGY

John C. Eichelberger, *Alaska Volcano Observatory, Geophysical Institute, University of Alaska Fairbanks, Fairbanks, Alaska, U.S.A.*

	Introduction	321
1.	**Properties and Processes of Molten Rock**	323
1.1	Composition of Volcanic Rocks	323
1.2	Melting Behavior	324
1.3	Volatiles in Melt	326
1.4	Magmatic Properties	326
1.4.1	Density	326
1.4.2	Viscosity	327
1.4.3	Bubble and Crystal Content	327
1.4.4	Mass and Thermal Diffusivity	328
1.5	Magmatic Domains	328
1.5.1	Source Region	328
1.5.2	Storage Region(s): Magma Chambers	328
1.5.3	Shallow System: Conduit and Vent	329
1.5.4	Eruption Plumes	329
1.5.5	Surface Deposits	330
1.6	Size of Eruptions	333
2.	**Formation and Ascent of Magma**	334
2.1	Tectonic Localization of Volcanoes	334
2.1.1	Mid-Ocean Ridges	334
2.1.2	Subduction Zones	335
2.1.3	Hot Spots	336
2.1.4	Other Occurrences	336
2.2	Ascent of Magma and Kinds of Eruptions	336
3.	**Societal Impact**	338
3.1	Geothermal Energy	338
3.2	Climate	338
3.3	Volcanic Hazards	340
4.	**Forecasting Eruptions**	340
4.1	Characteristics of Useful Predictions	341
4.2	Methods of Monitoring and Prediction	341
4.2.1	Eyewitness Reports	341
4.2.2	Seismology	341
4.2.3	Geodetic Measurements	342
4.2.4	Gas Measurements	342
4.2.5	Geology	343
4.2.6	Remote Sensing	343
4.2.7	Other Techniques	344
5.	**Case Histories**	344
5.1	Long Valley Caldera, California	344
5.2	Valley of Ten Thousand Smokes, Alaska	345
5.3	Crater Peak, Mount Spurr Volcano, Alaska, 1992	346
5.3.1	Eruption Chronology	347
5.3.2	27 June 1992 Eruption	347
5.3.3	18 August 1992 Eruption	348
5.3.4	16–17 September 1992 Eruption	348
5.3.5	Post-Eruption Seismic Swarms	348
5.3.6	Interpretation of Spurr Activity	349
6.	**The Future**	349
	Glossary	349
	Works Cited	351
	Further Reading	353

INTRODUCTION

Volcanology is the study of processes involved in ascent of magma to, and eruption onto, a planetary surface. Though waning somewhat from the early days of a planet-wide magma ocean, volcanism is still the dominant force in shaping Earth's surface: some 60% of Earth has been resurfaced by volcanism in the last 5% of the planet's history (Decker and Decker, 1989). As almost all of this new crust is covered by water, the abundance of volcanic activity on Earth is, if anything, underappreciated.

The processes that drive volcanism are fundamentally simple. Molten rock rises buoyantly from its melting source within the Earth because it is less dense than the col-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

umn of solid rock it traverses. Volatiles, which become concentrated in magma during its formation, exsolve during ascent and can give rise to fragmentation during final decompression. This explains both the reason for eruption and the cause of its often explosive character (Fig. 1).

Volcanism in operation is complex, however, and produces a tremendous array of phenomena and deposits. The reason is that magmas are themselves complex. Even the chemically simplest magma, rhyolite, cannot be usefully described as less than a six-component system. Most magmas are intimate and dynamically changing mixtures of gas bubbles and multiple crystal phases in silicate melt. Magmas are subjected to a wide range of pressure and temperature conditions during ascent and eruption, to which their response cannot be described in wholly equilibrium terms. Understanding the behavior of magma during ascent, which is at the heart of volcanology, requires the combined insights of laboratory experiments, fluid dynamical modeling, and field observations. A thorough understanding of volcanic systems requires synthesis of results from geophysics, of which seismic, gravity, and magnetic investigations are generally most helpful; geology; petrology; geochemistry; remote sensing; theoretical modeling; laboratory simulation; and direct sampling by drilling.

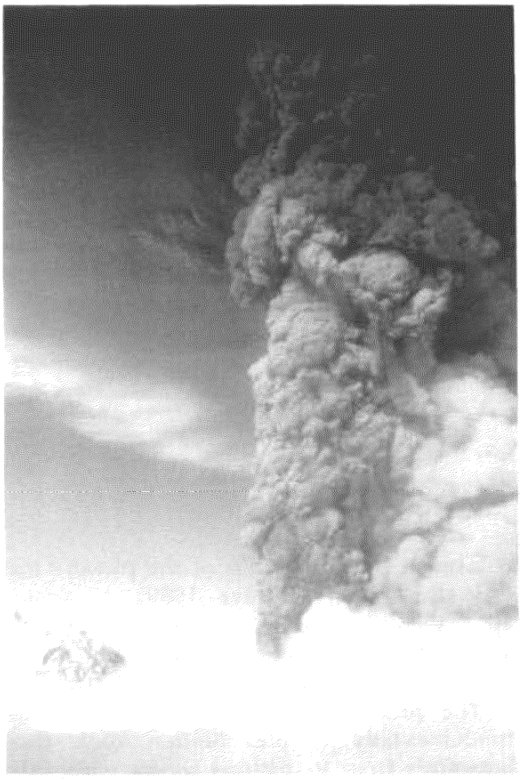

FIG. 1. Oblique aerial view of the 18 August 1992 eruption of Crater Peak, Mount Spurr, Alaska. The ash column is approximately 300 m in width at its base and, in its core, consists of a mixture of magma fragments and magmatic gas at about 1000 °C flowing at 100–300 m/s. Photograph by Robert McGimsey, Alaska Volcano Observatory.

Volcanism is important on a number of levels. From a basic scientific standpoint, it represents the transport of low-temperature–melting components to the planet's surface and must be understood in order to elucidate the development of Earth's crust, hydrosphere, and atmosphere. There are, in addition, several aspects of volcanism of immediate importance to society. Migration of fluids associated with volcanism results in economically important concentration of metals as ore deposits. Energy in the form of volcanic heat, or geothermal energy, is a viable and relatively nonpolluting source of power. Toxic volcanic emissions must be quantified in order to assess the seriousness of anthropogenic introduction of toxic materials into the environment. Volcanism may play a role in both short- and long-term perturbations in climate. Finally, volcanism has caused tens of thousands of deaths in the last few decades. The hazard to human life is increasing as populations living directly on active volcanoes increase and as the skies above even remote explosive volcanoes become crowded with large jet aircraft (Simkin, 1992; Casadevall, 1994). Our species is in fact fortunate that it has not experienced an eruption of the first order in its conscious memory. There was a time, however, when we were in an evolutionary "bottleneck" (e.g., Jones and Rouhani, 1986). How could a species destined to inhabit the Earth almost from pole to pole find itself so endangered? Perhaps it suffered a volcanic calamity in its volcano-rich African homeland. The impact need not have been direct. The relatively nonexplosive 14-km^3 Laki "fires" of Iceland in 1783 resulted in the deaths of over 50% of that island's livestock and 20% of its human population. We may be more shaped by seemingly esoteric natural forces like volcanism than we imagine. Volcanology can probably never mitigate the consequences of such a grand scale of disaster, but it has recently developed a predictive capability that is saving both lives and property during ordinary, "everyday" eruptive crises.

1. PROPERTIES AND PROCESSES OF MOLTEN ROCK

Magma can be defined as rock containing sufficient melt to be mobile. This is to identify it as the primary agent of mass transport in an igneous system and to distinguish it from static deep crust where some melt may be present but where large-scale flow is not occurring. Although melt is the defining phase of magma, it must be emphasized that magma refers to the entire system: entrained crystals and bubbles as well, not just the melt phase.

1.1 Composition of Volcanic Rocks

Because volcanic rocks are deposits observed to form from hot, molten material, they unquestionably form from magma. This material may or may not have changed in important ways since it was inside the Earth. The changes can be assessed by comparing characteristics of volcanic rocks that have experienced different rates of decompression and cooling. Pumice that emerges from the vent at velocities of hundreds of meters per second and is cooled in flight presumably represents magma that has seen the most rapid change to surface conditions of any samples obtainable: decompression in minutes to hours and cooling in seconds to minutes. Lava that emerges at centimeters per second and retains its heat in a slab tens to hundreds of meters thick as a tonguelike flow represents magma that has seen orders of magnitude slower decompression and cooling. Large intrusions that do not reach the surface but are later exposed by erosion may record the results of even slower rates of decompression and/or cooling. Inspection of these various deposits reveals that pumice gives a faithful (perhaps much inflated) view of magma as bubbles and crystals in melt that is now represented by glass (Fig. 2). The entrained crystals are called phenocrysts. Except for a chilled, glassy selvage, thick lava flows and shallow intrusions exhibit similar phenocrysts in a fine crystalline, rather than glassy, matrix. The normal inference is that phenocrysts represent crystals that grew from melt during pre-eruptive storage in a subsurface reservoir and that the fine matrix crystals grew after emplacement in the case of the relatively slowly cooled flow or intrusion interiors (McBirney, 1984).

A similar approach can be applied to bulk composition, in order to assess how closely bulk composition of volcanic rocks may reflect magmatic compositions. Among freshly erupted rocks, there appear to be no variations in major-element and nonvolatile-trace components that are correlative with decompression or cooling rates (e.g., Westrich *et al.*, 1988). Such concentrations apparently represent magmatic values.

The situation is quite different for volatile components. Water contents of fresh (unweathered) volcanic rocks are uniformly low. Exceptions are for magmas quenched to glass while still under high pressure such as in deep ocean or at the walls of conduits. Although low water contents of volcanic rocks could be interpreted as indicating that magmas are uniformly low in water, such a conclusion is at odds with known phase equilibria in silicate melts and with some direct evidence, as is discussed shortly. It is more likely, therefore, that water is lost almost instantaneously upon decompression and that magmas that reach the surface are anhydrous for that reason (Eichelberger, 1995). Chlorine concentration tends to be high in pumice, lower in glassy lava, and lower still in crystalline lava, indicating a relationship of its retention to decompression and cooling rates (Wolf and Eichelberger, 1996). Thus, from examining quickly cooled volcanic rocks, it is possible to determine directly the pre-eruption crystal and melt content and composition of the magma, though not the pre-eruption volatile content of the melt.

Volcanic-rock compositions are neither infinitely variable nor do they form discrete clusters. Rather, rocks associated in time and space—that is, those comprising a volcanic center—define chemical trends. Rocks of most volcanoes fall wholly within the basalt compositional field or within a portion of the "differentiated" basalt-andesite-dacite-rhyolite suite. Within this latter suite, as

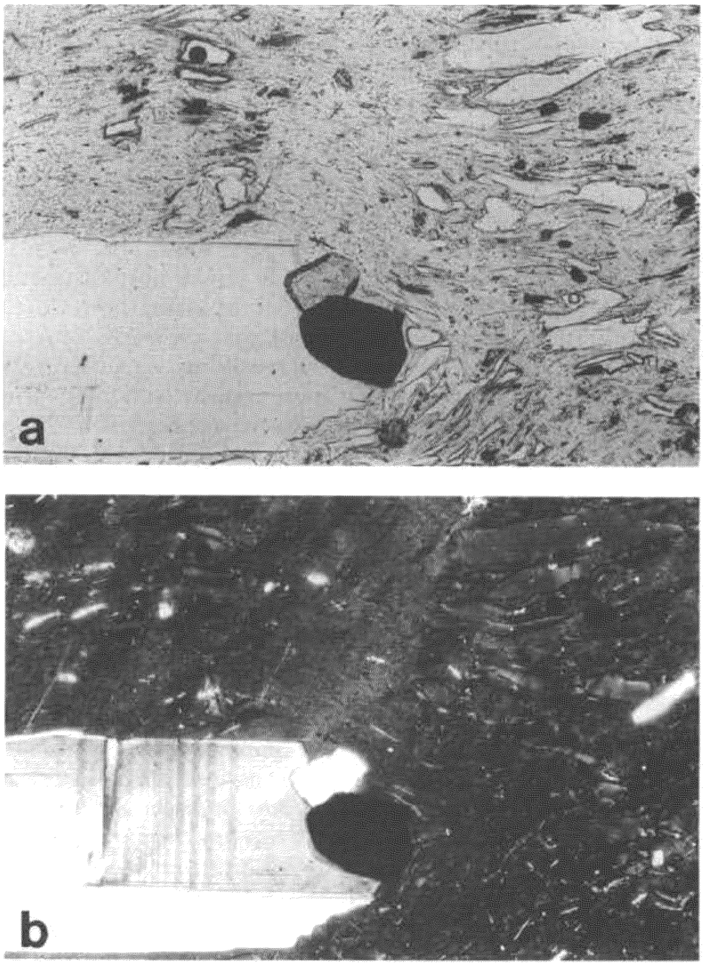

FIG. 2. Photomicrographs of Novarupta Dome rhyolite. The field width is 1.5 mm. The upper view is in normal transmitted light. In the lower view, polaroid filters above and below the same sample are crossed, so that only crystals that rotate polarized light are visible. From these photomicrographs of this quenched material, the three-phase (solid, liquid, and gas) character of magma is apparent. Gas bubbles (e.g., upper right) were inflated by exsolution of steam from silicate liquid during decompression. The matrix is glass (gray, upper view; black, lower view) that formed from melt upon rapid cooling during eruption. The large crystal at lower left is plagioclase feldspar that grew slowly from melt in the magma chamber prior to eruption. Smaller crystals are also of plagioclase and likely grew during ascent to the surface. Photograph by Daniel Wiesneth, Alaska Volcano Observatory.

SiO_2 content increases from about 48 to 77 wt%, the presumed direction of chemical evolution, CaO, MgO, and FeO decrease by an order of magnitude, and Na_2O and K_2O increase several fold. There is a corresponding variation in conditions and properties from hot and fluid to cool and viscous, which in turn have a profound effect on eruptive behavior. The chemical spectrum has classically been interpreted as representing a "liquid line of descent" from mantle-derived basalt by the process of fractional crystallization: growth and removal of crystals from melt. However, special water-fluxed mantle melting conditions, crustal melting, and crustal contamination may play a role in generation of the silicic members of the series (Hess, 1989). Peacefully flowing basalts, such as those of Kilauea Volcano, Hawaii, have long attracted the attention of volcanologists. In this century, disasters such as Mount Pelé in Martinique, the Valley of Ten Thousand Smokes in Alaska, and more recently Mount St. Helens in the United States, Mount Pinatubo in the Philippines, and Mount Unzen in Japan have drawn attention to the often-violently explosive silicic end of the chemical spectrum.

1.2 Melting Behavior

It is instructive to compare the melting of rock to the more familiar melting of ice. Water is a single-component system and consequently, at a given pressure, water and ice can coexist at only one temperature. Rocks are multicomponent chemical systems with more degrees of freedom. At a given pressure, melt and multiple crystalline phases coexist over a range of temperatures [Fig. 3(a)]. The lower temperature limit at which melt is present is the solidus, and the upper limit where crystals are present is the liquidus. Typically, there are hundreds of degrees Celsius between the liquidus and solidus, and most magmas exist in this "mush" state. A magma that is heated above its liquidus will merely dissolve its rock container until it has returned to solid-saturated conditions. Because the crystal and liquid fractions differ both in chemical composition and in physical properties, chemical fractionation can occur by phase segregation during storage or flow.

In contrast to ice, the melting of rocks involves approximately a 10% volume increase rather than a 10% volume decrease. This has two important consequences. First, silicate melts are buoyant within the Earth and tend to rise toward the surface. As mentioned before, this is the basic cause of volcanism and attendant planetary differentiation. Second, because increased pressure favors the solid (denser) phase, in the absence of effects of changing rock composition, temperature of melting increases with depth within the Earth.

Like water, silicate melt dissolves other chemical components and is therefore subject to freezing-point depression, where the solidification temperature is lowered with increased concentration in melt of a solute

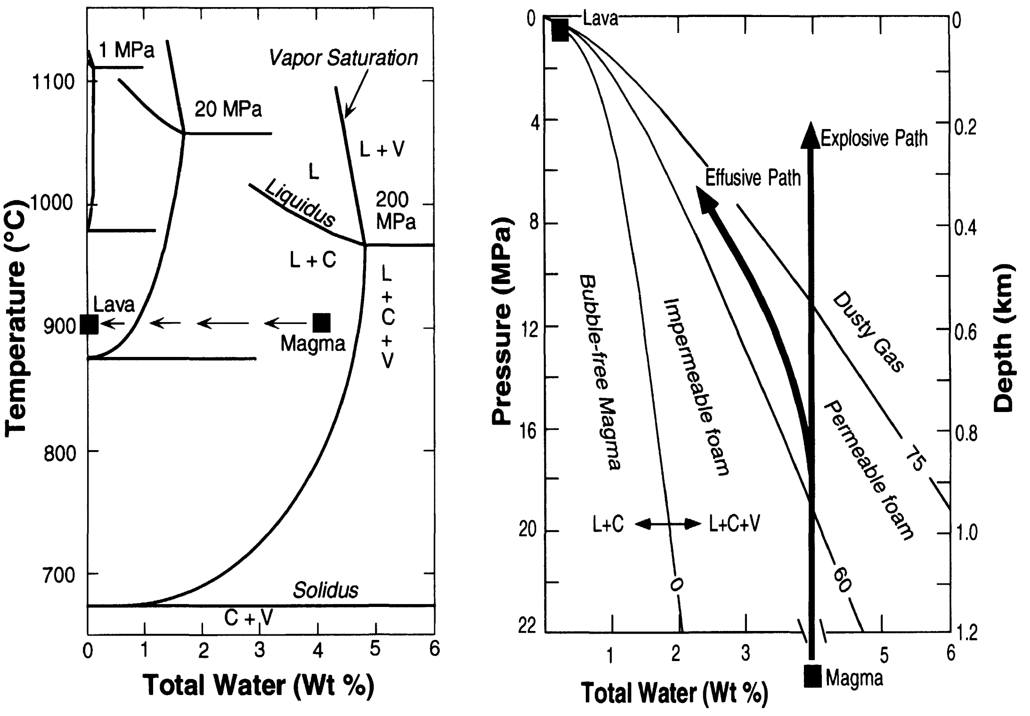

FIG. 3. Effects of water content and pressure on the crystal and bubble content of magma. (a) Phase fields (L = liquid, C = crystals, V = vapor) of silicic magma as a function of temperature and water content at pressures corresponding to depths of 8 km (200 MPa), 800 m (20 MPa), and 40 m (1 MPa) in the Earth. Pressure shifts the vapor saturation curve to higher water content and the solidus to lower temperature. At constant pressure, water lowers the liquidus. Note that loss of water during decompression causes magma to be below its solidus, that is, thermodynamically unstable at the surface. (b) Bubble content of silicic magma as a function of pressure and water content, assuming that the magma is crystal poor and that crystallization does not accompany water loss [violation of equilibrium assumptions of (a) but appropriate to most eruptions]. Curves are lines of constant porosity, as labeled. Solubility of water in melt decreases rapidly at shallow depth. Once exsolved, vapor expands inversely with pressure. 1 wt% water in melt produces approximately a 100-fold expansion of the system upon decompression to 1 atm. Removal of most of the vapor from the system is required for magma to emerge intact as dense lava (effusive path). After Eichelberger (1995).

that does not enter the solid phase. Water itself is highly soluble in silicate melt, and the addition of 1 wt% water to magma lowers the liquidus by tens of degrees Celsius. Because solubility of water in melt increases with pressure, the freezing-point depression effect increases with pressure, and both the solidus and the liquidus of the water-bearing system are shifted downward in temperature [Fig. 3(a)]. This may at first seem inconsistent with the expectation that pressure would favor denser crystals over melt, as was cited previously for dry melting. However, in the hydrous system, water is dissolved in melt when melt is present and free as vapor when melt is absent. The melt-bearing system therefore occupies less volume, and the stability of melt is enhanced by increased pressure as a consequence.

1.3 Volatiles in Melt

The presence of water and other volatiles in magmatic melt has another important effect: volatiles come out of solution and form a free vapor or gas phase during ascent. For most terrestrial magmas, water is the most abundant volatile component, and its solubility in melt decreases by two orders of magnitude from mid-crustal to surface pressures. The shape of the solubility curve, with the saturation value approximately proportional to the square root of pressure and hence of depth, is such that water is held in solution until shallow depth and then rapidly released with further decrease in pressure. 1 wt% dissolved water will produce roughly 100-fold expansion upon eruption. Given that water concentrations of a few to several weight percent water are thought to be common, it is easy to see that where a volcanic conduit constrains expansion to the vertical direction, very high velocities will be achieved. Consequently, exsolution and subsequent behavior of water vapor during ascent are the most important factors in determining the characteristics of an eruption (Sparks, 1978).

The fully expanded erupting material is no longer magma in the sense of having a continuous melt phase. Rather, it is a dusty gas, with a continuous vapor phase. The process by which melt-continuous magma becomes a vapor-continuous eruption plume is called fragmentation and results from a combination of coalescence and breakage of bubbles. Quenched, fragmented melt in the form of glass shards, as well as crystals from the magmatic suspension, constitute the small solid particles of the dusty gas and are known as ash. Larger, incompletely fragmented pieces of magmatic foam are pumice. About 75% porosity seems to be the normal limit of inflation without fragmentation, as represented by pumice [Fig. 3(b)].

At atmospheric pressure, the amount of water vapor required to inflate magma to pumice porosities is trivial, one or two hundredths of a weight percent. Given that this is three orders of magnitude less than the solubility of water in melt at the base of the crust, that water is likely present as hydrous minerals in magma source regions, and that water is strongly partitioned into the melt phase during magma formation (indeed, without the freezing-point depression effect of water, temperatures at depth may not reach the melting point), it might be expected that most magmas would contain significant water content and hence inflate and fragment to ash during ascent. A significant problem, therefore, is how to explain the intact emergence of initially hydrous magma as degassed lava in nonexplosive eruptions. The vapor pressure within entrained bubbles, which controls vapor–melt equilibrium, should be close to the pressure on the magma body, generally referred to as the lithostatic pressure (this may differ from fluid pressure in connected cracks outside the magma body, which will be less, often hydrostatic). If so, magma will not degas at depth to a vapor pressure lower than the lithostatic pressure it is under. But it also cannot emerge without fragmentation if vapor pressure is in excess of ambient pressure. It can therefore be inferred that some process of vapor–melt separation at shallow depth plays an important role in controlling whether eruption of ash or lava occurs.

1.4 Magmatic Properties

A number of important physical and chemical characteristics of magmas influence eruptive behavior.

1.4.1 Density Density of silicate melt is a strong function of composition and a weak function of temperature (Bottinga and Weill, 1970). Bulk density of magma is affected by these factors as well as the quantity of dense crystals and low-density bubbles entrained in the mixture. Although the coefficient of thermal expansion of silicate melt is quite small, it is large enough that even small bodies of viscous magma can be expected to convect (Bartlett, 1969). Because it is much smaller than the equivalent "coefficient of hydration expansion," the volume increase in response to introduction of water rather than heat into melt, heterogeneous distribution of dissolved water will tend to shut off convection (Shaw, 1974). Such magma will tend to stratify and stagnate with the wettest magma on top. Density-controlled stratification of magma in chambers has been suggested to explain chemically zoned eruption products and the most-explosive-first eruptive pattern of some volcanoes (e.g., Hildreth, 1981).

Most, but not all, magmas are less dense than the rocks they traverse during ascent. It may be that low-density continental crust acts as a density filter that discourages the rise of dense mafic, deep-seated magmas to the surface. If so, it is an imperfect one. Basalts do erupt through low-density continental crust, indicating that sufficient overpressure can be sustained within conduits to allow the magma to flow in artesian-well-like fashion.

1.4.2 Viscosity Viscosity of silicate melt is a strong function of temperature, water content, and silica content (Shaw, 1972). Silicate melt contains chains of silicate tetrahedrons. The longer these chains are, the more viscous the liquid. High silica content, low water content, and low temperature all favor polymerization. Thus, silicic magmas are more viscous than their hotter, more mafic associates unless these effects are offset by increased water content. Viscous magmas may be expected to rise more slowly and/or in larger bodies than fluid magmas. The most silicic magmas may sometimes rise as huge balloonlike bodies or diapirs many kilometers across, whereas mafic magmas often flow sheetlike through thin fissures centimeters to meters thick as dikes and sills (Fig. 4). High viscosity retards melt/crystal and melt/bubble segregation processes, reducing the amount of fractionation that can occur during a given storage or ascent time and requiring long times (thousands of years?) for evolution of silicic magmas.

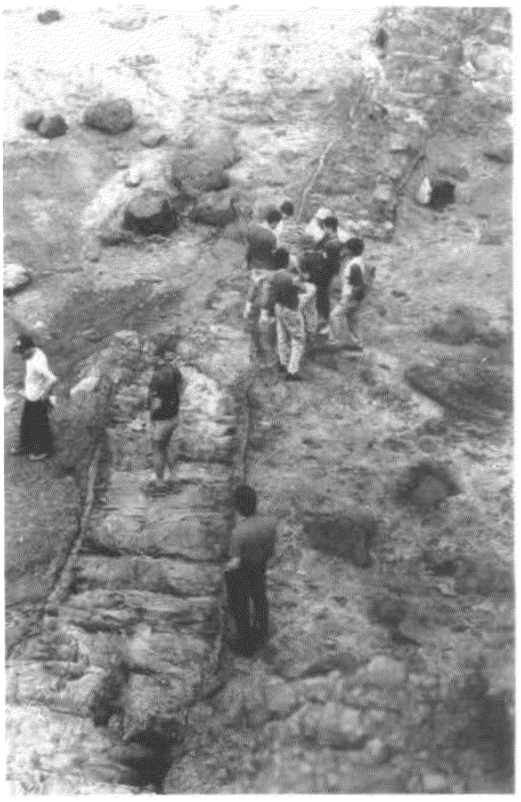

FIG. 4. An andesite dike cuts marine volcaniclastic rocks of similar composition. This exposure is on a wave-cut terrace on the coast of Hokkaido Island, Japan. Photograph by John Eichelberger, Alaska Volcano Observatory.

1.4.3 Bubble and Crystal Content Bubbles increase bulk viscosity of magma (Murase and McBirney, 1984) and represent a reservoir of gas available for instant expansion during decompression. Crystals increase bulk viscosity until they lock the melt in a rigid network of crystals (Marsh, 1981). There appears to be an eruption limit of about 50 vol% crystals in magma that represents onset of this immobility. Because most crystals contain little water compared to melt, crystallization concentrates water in the remaining melt. Where the solubility limit is exceeded, this causes exsolution of water vapor or "second boiling." Thus cooling, as well as decompression, causes vesiculation. This mechanism can produce very high overpressures in the crystallizing body, leading to dike injection, explosion, and/or intense fracturing (Burnham, 1979).

1.4.4 Mass and Thermal Diffusivity Water has the highest diffusivity of any major chemical component in magma, typically 10^{-7} cm^2/s (Shaw, 1974). Thermal diffusivity is about 4×10^{-3} cm^2/s (McBirney, 1984). The diffusivity of heat is therefore more than three orders of magnitude larger than the diffusivity of water. This means that magma bodies will always cool before significant long-distance diffusive chemical transport can occur. This results in the transformation, upon cooling, of melt to glass, and is the reason why chemical diffusion cannot operate over more than micrometer distances during an eruption.

1.5 Magmatic Domains

Although it is difficult to generalize over the entire spectrum of volcanic systems, it is possible to subdivide these systems into domains within which specific processes dominate (Fig. 5).

1.5.1 Source Region This is the domain of melting and melt coalescence. Depending upon whether melt is extracted immediately upon formation or lingers at its source until a critical melt fraction is reached, the chemical process may be described as fractional or equilibrium partial melting, respectively (McBirney, 1984). A melt-rich mass separates from a crystal-rich residue of the same temperature but of more refractory composition and rises diapirically (that is, balloonlike) through its plastic surroundings.

1.5.2 Storage Region(s): Magma Chambers Although magmas may rise directly from source to surface, they more often undergo substantial chemical modification by internal fractionation and external contamination en route. Such processes must be very slow, at least for silicic and therefore viscous magmas, and would seem to require protracted storage in a slow-cooling magma chamber. This ponding within the crust may occur when dense, unevolved magmas rise to their level of neutral buoyancy, or when diapiric ascent is stopped by the thermally controlled mid-crustal transition from ductile to brittle rock behavior. Natural exposures of deep levels beneath eroded volcanic systems

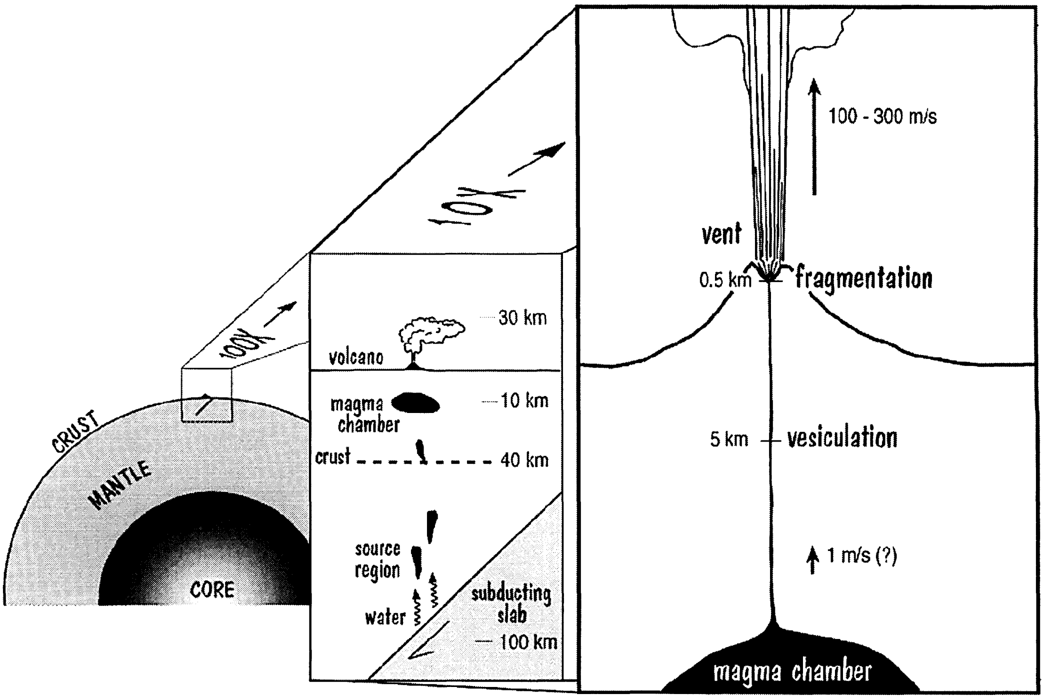

FIG. 5. Diagram illustrating the size relationships of parts of a subduction-zone volcano to the subduction zone and of the subduction zone to Earth. Figure by Deborah Coccia, Geophysical Institute.

provide direct geologic evidence of such features in the form of stocks or plutons—homogeneous or regularly zoned bodies of igneous rock that crop out over many square kilometers and lie in sharp contact with older host formations. Judging from eruption histories of volcanoes, storage times may range upward to hundreds of thousands of years. The chamber itself may last longer than magma within it; that is, it may repeatedly release magma to the surface and be replenished by magma from below. Little is known about the replenishment of such chambers, although petrologic evidence of mixing of primitive and reservoir magma just before eruption (Pallister *et al.*, 1992), together with supporting evidence from deep, magma-induced seismicity (White, 1996), suggests that replenishment may trigger eruptions.

Unlike the situation in the source region, crustal magma chambers will have a strong temperature contrast with the surrounding crust, and will therefore lose heat between replenishment events. This heat loss results in crystallization, which may in turn produce second boiling, overpressurization, and eruption (Tait *et al.*, 1989).

1.5.3 Shallow System: Conduit and Vent Upper-crustal portions of volcanic systems are frequently exposed by erosion and therefore are the portion about which there is the most direct information. This domain is characterized by extreme temperature contrast between magma and environment and by the possibility of direct interaction between magma and groundwater. Although a few volcanoes are open systems in constant eruption, most erupt episodically. They have repose periods of years to millennia interleaved with periods of activity, sometimes consisting of eruptive pulses lasting hours to days between comparable intervals of quiet. In most cases, then, the conduit is simply the preferred pathway for magma from source or holding chamber to surface, occupied by solid rock during repose. Although dikes would seem to be the energetically favored means of transport through cold, brittle crust and are indeed common, the form of shallow feeders ranges from dikes through elongate conduits to circular pipes (Carrigan *et al.*, 1992). Central, pipelike conduits of volcanic mountains are probably meters to tens of meters across.

Vents typically consist of a funnel-like flaring of the conduit toward the surface (Wilson *et al.*, 1980). This upward broadening can develop in three ways:

1. reaming of conduit walls as magma expands and accelerates during decompression,
2. slumping of shallow conduit-wall material into the open conduit during eruption, and
3. subsidence of the area surrounding the eruptive center due to withdrawal of magma from a chamber beneath it.

The first and second effects can be expected in small eruptions, whereas the third usually accompanies eruptions of a cubic kilometer of magma or more. Vents range in diameter from meters for fluid magma eruptions to kilometers for viscous, explosively erupted magmas. The largest structures associated with eruptions are tens of kilometers across, owe their form primarily to subsidence into a subjacent magma chamber, and are more properly termed calderas (Smith and Bailey, 1968). These may contain multiple, central post-collapse vents and/or a marginal circle of vents marking the ring fracture along which the magma chamber roof foundered. Doughnut-shaped "ring-dikes" of granite many kilometers in diameter, for example in the White Mountains of New Hampshire, may be the exposed intrusive roots of such systems.

In rising from a reservoir to the surface, magma rapidly encounters orders-of-magnitude changes in ambient temperature and pressure. Although many of the deep reservoir processes can be described in terms of phase equilibria, this is not the case in the conduit regime. Inspection of Fig. 3(a) shows that decompression will shift the solidus upward. Adiabatic rise of magma is close to isothermal (Rutherford and Hill, 1993), and so the solidus may well shift above temperatures prevailing in the magma during ascent. Thus, what erupts is actually a metastable, super-cooled liquid, and eruption requires that the rate of magma ascent outruns the kinetics of dewatering-induced crystallization.

1.5.4 Eruption Plumes Eruption plumes are an ephemeral portion of the volcanic system that penetrates the atmosphere during an explosive eruption (Fig. 1). The plume consists of magmatic gas, entrained crystals and quenched melt, condensates, vent-wall debris, and admixed atmosphere. A plume will ordinarily rise buoyantly to the tropopause, but convective thrust may carry it as high as 30 km. Plumes rapidly equilibrate thermally with the atmosphere, so that volcanic plumes that penetrate the stratosphere may actually be detected by satellite-borne infrared sensors as anomalously cold, rather than warm, clouds. After cessation of eruption and fallout of coarse material near the vent, the remaining ash cloud may remain as a coherent mass for days, drifting in air currents for tens of thousands of kilometers. Eruption plumes represent a significant contribution of particulate matter and gases to the atmosphere. If they reach flight levels, the resulting ash clouds pose a serious hazard to aviation (Fig. 6). If they are large, they may influence climate.

1.5.5 Surface Deposits Deposits formed when magma erupts can be naturally divided into those formed from magma that has fragmented upon eruption and those formed from magma that remained intact. The latter material is called lava. Depending upon viscosity, flux, total volume, and topography, lava flows may be meters to hundreds of meters thick and travel hundreds of meters or even tens of kilometers from their vents. Magma that does not flow outward appreciably forms a lava dome over the vent (Fig. 7). This often happens in the case of silica-rich magma. Heat is well conserved during extrusion of lava, so that flows and domes may remain hot for many months or years or even decades.

When magma reaches the near surface with external vapor pressure exceeding confining pressure, explosive fragmentation occurs. Some of the ensuing depositional processes are the same as those that form sedimentary rocks (Fisher and Schmincke, 1984). The extreme in energetic fragmental eruption produces a high-velocity eruptive column feeding a "mushroom cloud" of ash (Fig. 1). Fall-out from these cloud deposits comprises air-fall tephra. Such deposits are millimeters to tens of meters thick, blanket topography like snow, are size-sorted and crudely bedded, and may be traced to hun-

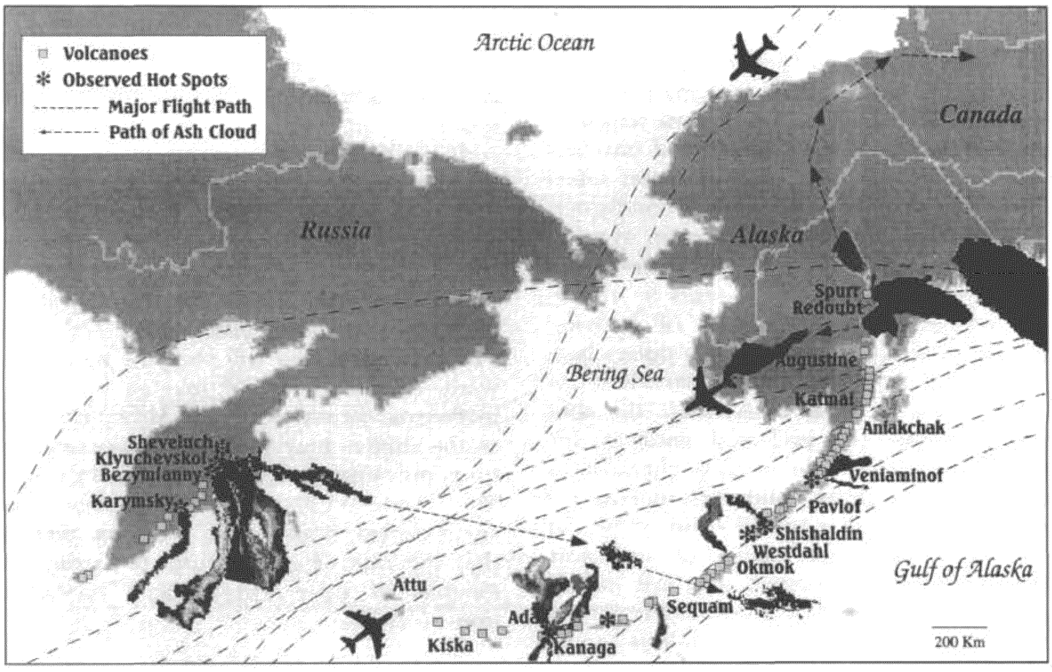
□ Volcanoes
✻ Observed Hot Spots
-------- Major Flight Path
←------ Path of Ash Cloud
Arctic Ocean
Canada
Russia
Alaska
Spurr
Redoubt
Augustine
Bering Sea
Katmai
Aniakchak
Veniaminof
Pavlof
Shishaldin
Westdahl
Okmok
Gulf of Alaska
Sequam
Shevelutch
Klyuchevskoi
Bezymianny
Karymsky
Attu
Kiska
Ad
Kanaga
200 Km

FIG. 6. Composite satellite image showing volcanic ash clouds that have invaded North Pacific air routes during 1990–1996. Images compiled by Ken Dean, Alaska Volcano Observatory.

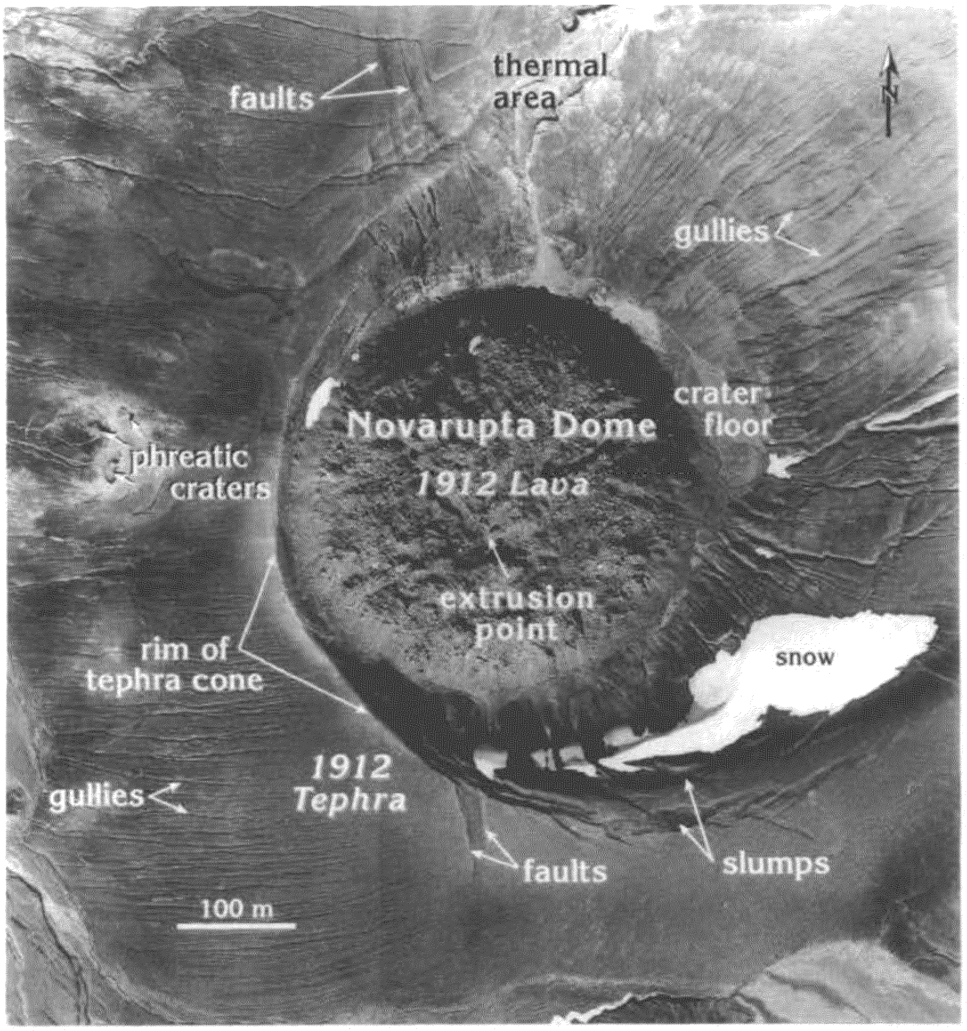
faults
thermal area
N
gullies
crater floor
Novarupta Dome
1912 Lava
phreatic craters
extrusion point
rim of tephra cone
snow
1912 Tephra
gullies
faults
slumps
100 m

FIG. 7. Vertical aerial view of Novarupta Dome, a 400-m-diam, 70-m-thick mass of rhyolite lava that concluded the great explosive eruption of 1912 at the head of the Valley of Ten Thousand Smokes, in what is now Katmai National Park, Alaska. The dome flowed from an east–west fissure buried beneath its center to fill the floor of the crater formed in the preceding explosive phase. Photograph by AeroMap Corp., Anchorage, Alaska.

dreds, or more rarely to thousands, of kilometers from the vent (Fig. 8). Except near the vent, most such deposits are substantially cooled when formed, because of the descent of constituent particles through cold atmosphere.

Still fragmental but, in a sense, intermediate between lava flows and tephra falls are pyroclastic flows. These behave as a fluid, like lava, but consist of fragmental material (though not size sorted because they are turbulently mixed during flow) like air-fall tephra. Pyroclastic flows form when magma contains sufficient gas to fragment upon eruption, but not enough to thrust it forcefully upward into the atmosphere. They may also be shed from the sides of energetic eruption columns, or issue from collapsing, incompletely degassed domes. The result is a relatively dense (compared to the atmosphere) mixture of magma fragments and magmatic gas that generally follows topographic lows as it flows downhill from the vent (Fig. 9). Heat may or may not be well retained, depending upon distance of travel, thickness of the flow, and amount of mixing

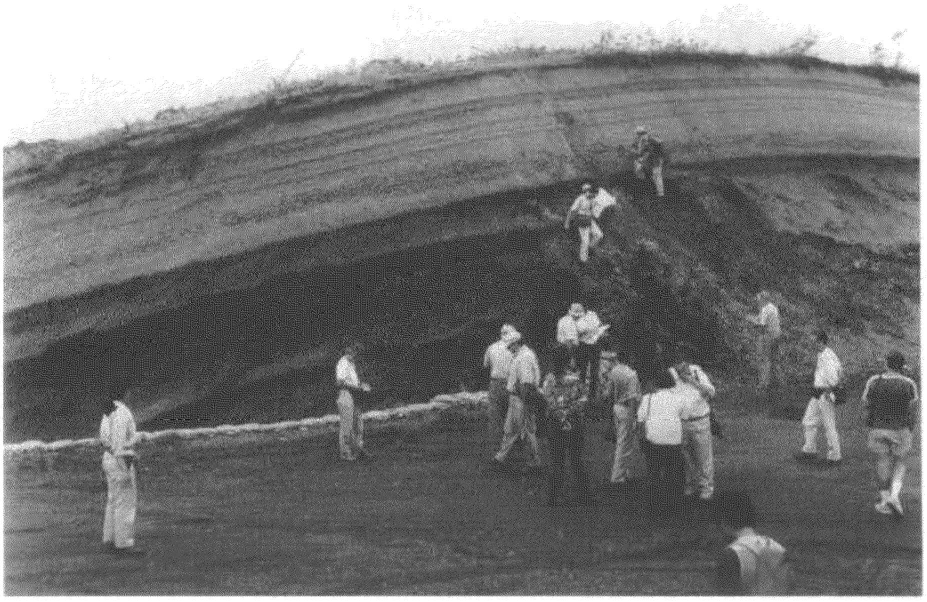

FIG. 8. Layers of air-fall tephra blanket the topography near Lake Toya Caldera, Hokkaido, Japan. Photograph by John Eichelberger, Alaska Volcano Observatory.

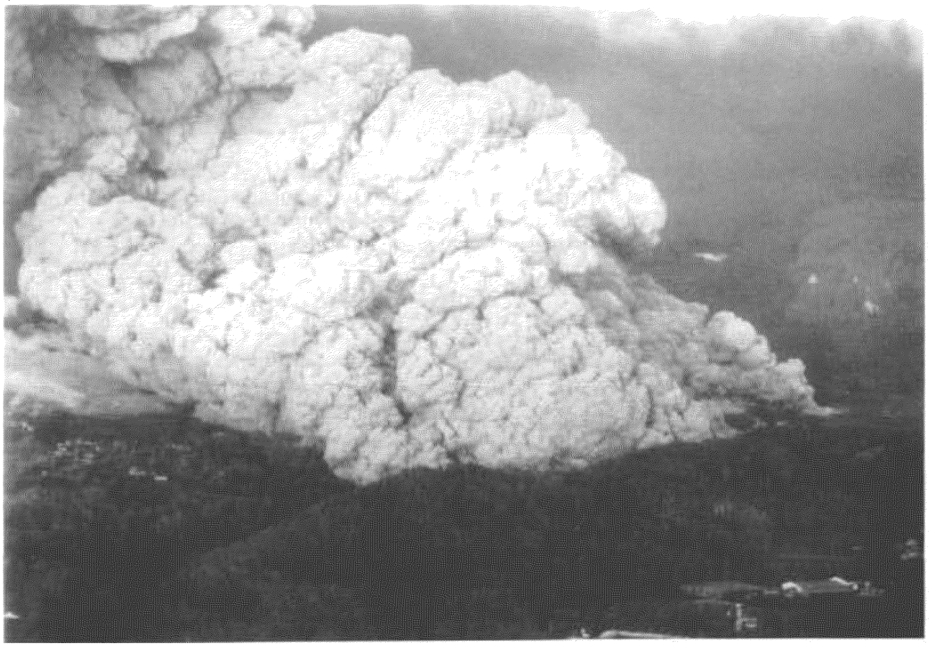

FIG. 9. A pyroclastic flow from Unzen Volcano sweeps down on the outskirts of Shimabara, Japan, during a June 1991 eruption. Photograph by Setsuya Nakada, University of Tokyo.

with air. Where heat is retained, the hot glass shards of the deposit will flatten and stick together under self-load in a process called welding. This compaction is revealed clearly by the flattening of included pumice clasts. The resultant cohesive rock is termed tuff (Fig. 10). Where welding has resulted in a complete loss of porosity, tuff is easily mistaken for lava.

An additional important class of deposits in volcanic terranes are mudflows or lahars. These result from the unstable accumulation of volcanic debris around vents. Mixing with surface water from torrential tropical rain as at Pinatubo in the Philippines or with eruption-triggered melt water from alpine glaciers as at Volcano de la Ruiz in Colombia creates a mobile slurry. While cooler (sometimes cold) than lava flows, lahars tend to be more fluid and far-reaching, spreading destruction down valleys as far as hundreds of kilometers from their source.

1.6 Size of Eruptions

Eruption is really only the end-member case for shallow intrusive events, in which magma sometimes stops short of the surface. Thus, the low extreme in volume erupted during magmatic events is zero, but the high extreme is more interesting (Smith, 1979). Typical subaerial eruptions of the Pacific "Ring of Fire," such as Mount Unzen, Japan (1990–1995), Mount Spurr, Alaska, U.S.A. (1992), and Mount St. Helens, Washington, U.S.A. (1980–1986), are 0.01–1 km^3 dense rock equivalent (DRE). The largest such event during this century, described later in this article, occurred on the Alaska Peninsula in 1912 and is 13 km^3 in magma volume. Also noteworthy are the 3-km^3 eruption of Mount Pinatubo, Philippines, in 1991, the 13-km^3 Krakatau, Indonesia, eruption of 1883, and the 14-km^3 Laki "fires" of Iceland of 1783 (Francis, 1993). The largest prehistoric silicic eruptions known, as inferred from mapping of voluminous tuff sheets, are 1000 km^3, such as Pleistocene Long Valley Caldera, California, and Yellowstone Caldera, Wyoming, eruptions, to 5000 km^3, such as the 28-Ma Fish Canyon Tuff in Colorado (Lipman, 1984). Vast Precambrian rhyolite terranes speak of colossal but ill-defined eruptions in the dim past.

Formation of a 1000-km^3 welded tuff sheet requires the sudden emptying of a

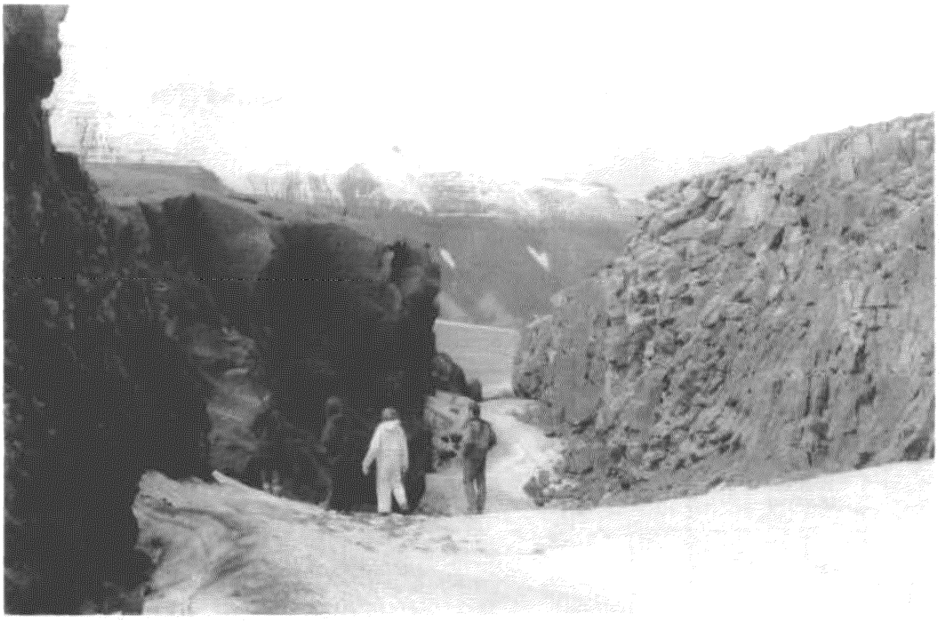

FIG. 10. Rapid erosion has partly exposed the interior of the tuff formed in 1912 that fills the Valley of Ten Thousand Smokes, Alaska. Pyroclastic flows emplaced the ash and pumice mixture at a temperature above the softening point of glass. Sintering of the glass and compaction or welding of the deposit under its own weight formed the hard, cohesive tuff. Photograph by Pavel Izbekov, Alaska Volcano Observatory.

crustal magma chamber with dimensions of order 10 × 10 × 10 km^3, although probably most are lenslike in form. That this actually happens is supported by syneruptive subsidence of the vent region to form a caldera of a volume comparable to that of the magma erupted. Thus, very large eruptions provide unequivocal proof of the existence of pluton-sized or even batholith-sized magma chambers (Lipman, 1984).

The largest cumulative volumes from multiple eruptions occur in the form of flood basalts, amounting to as much as a million cubic kilometers in less than a million years (Francis, 1993). Such volcanic provinces probably do not represent emptying of individual storage chambers, but rather rapid segregation of fluid melt from a huge mantle source volume.

2. FORMATION AND ASCENT OF MAGMA

2.1 Tectonic Localization of Volcanoes

A map of Earth's active volcanoes reveals that they are not uniformly distributed over the planet but generally occur in lines or arcs, with scattered outliers (Fig. 11). Some areas of the Earth have not experienced volcanism for a billion or more years. These areas correspond to the central regions of continents. By far, the largest class of volcanoes occur as a network of continuous ranges within the ocean basins, most prominently down the center of the Atlantic Ocean. The second most important class consists of a series of arcs that surround the Pacific Ocean and occur in a few other places as well. A third class occurs as chains, most notably the Hawaiian chain, that show a progression in age, so that one end is older and current activity is limited to the opposite end. Occurrence of these classes of volcanoes can be understood in terms of the plate-tectonics paradigm (Decker and Decker, 1989; PLATE TECTONICS).

2.1.1 Mid-Ocean Ridges Mid-ocean ridge volcanoes are almost entirely basaltic, with the exceptions of special locations such as Iceland. The total length of these ridges is 60 000 km or about 1.5 Earth circumferences. Few if any eruptions of these volcanoes perturb the ocean's surface, and so their importance is obscured. Being relatively volatile-poor magmas erupted under

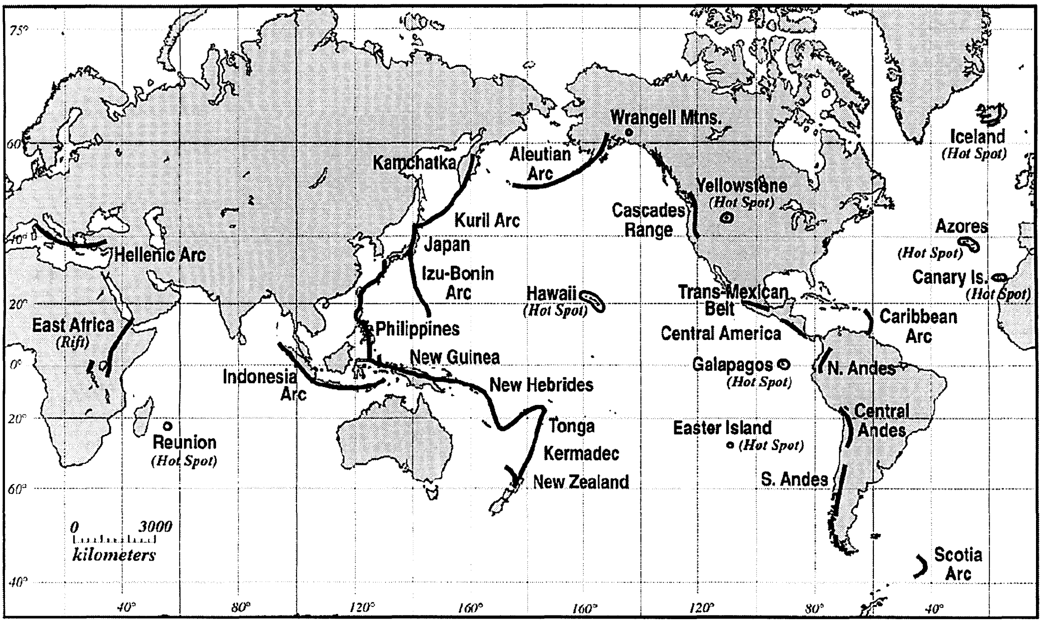
75°
60°
40°
20°
0°
20°
60°
40°
Wrangell Mtns.
Kamchatka
Aleutian Arc
Kuril Arc
Japan
Izu-Bonin Arc
Hellenic Arc
East Africa (Rift)
Philippines
New Guinea
Indonesia Arc
New Hebrides
Hawaii (Hot Spot)
Tonga
Kermadec
New Zealand
Reunion (Hot Spot)
Cascades Range
Yellowstone (Hot Spot)
Trans-Mexican Belt
Central America
Galapagos (Hot Spot)
Easter Island (Hot Spot)
N. Andes
Central Andes
S. Andes
Caribbean Arc
Iceland (Hot Spot)
Azores (Hot Spot)
Canary Is. (Hot Spot)
Scotia Arc
0 3000 kilometers
40°
80°
120°
160°
160°
120°
80°
40°

FIG. 11. Major active subaerial volcanic systems of the Earth. Modified from Smithsonian Institution, Global Volcanism Program.

high water pressure, ridge eruptions are effusive, and the lava flows produced are the primary constituent of the upper oceanic crust that floors the ocean basins. Associated shallow intrusive feeders and the crystallized source chamber comprise the crustal substrate. This dense, thin, and therefore topographically low crust covers about 60% of the planet (oceans lap onto the continental edges, and so cover a bit more, 70%, of the planet). Oceanic crust decreases in age from a maximum of about 200 000 000 years near some continental coasts to zero age at active mid-ocean ridges. This is known both from direct age dating and from a ridge-symmetric, mirror-image pattern of magnetic stripes on the ocean floor. Thus, the mid-ocean ridge volcanoes are responsible for generation of oceanic crust. Given that the crust averages about 8 km in thickness, this means an average production of magma of 20 km^3/year or 3×10^5 m^3/(km of ridge length)·year. Associated rates of crustal spreading range up to 10 cm/year. Such large-scale crustal motion is to be expected if the planet's skin rides on convection cells. Earth's upper mantle is apparently able to creep and thereby convect heat from the deeper mantle or core because it is near its melting point. The adiabatic lapse rate for mantle is considerably smaller than the rate of change of the mantle solidus temperature with pressure, and so a rising current of mantle material will undergo melting by decompression. The melt thus produced rises to the surface at crustal tears or rifts to produce mid-ocean ridge basaltic volcanism and generate a new crustal skin.

2.1.2 Subduction Zones If crust is being created from mantle at mid-ocean ridges, an equal area per unit time must be returned to the mantle, or subducted, elsewhere. The magnetic stripes on the ocean floor, which are isochrons of formation age caused by freezing-in of alternating polarities in Earth's magnetic field, are truncated at the deepest structures on the Earth's surface: ocean trenches. Intersecting the surface at these trenches are oblique planes of seismicity that extend to as much as 600 km depth and represent the deepest occurrence of brittle rock behavior in the Earth. The seismogenic zone coincides with the top of an inclined slab that exhibits anomalously fast transmission of seismic waves and hence is probably anomalously cold. From this, it is concluded that oceanic lithosphere (ocean crust and subjacent rigid mantle) is being thrust beneath another rigid plate with which it is converging. Where the collision is between a plate skinned with oceanic crust and one skinned with thicker and less dense continental crust, it is always the oceanic-crust-bearing plate that is subducted beneath continent-bearing crust. Inboard (continent-ward if collision is with a continent) of the trench and parallel to it, an arc of volcanoes is present on the surface. Often this line is quite narrow, as in the Aleutian Arc, and consistently overlies the ~100 km isobath on the subducting slab. Thus, subduction of the slab to some critical depth appears responsible for melt generation and volcanism. An exception is when the angle of descent of the slab is very shallow, as under the Alaska Range and parts of the Andes. Then, mountain building through crustal thickening, but not volcanism, occurs.

Subduction-zone volcanism is quite different from that of the mid-ocean ridges. It involves water-rich magmas, is consequently often explosive, and is on average more silicic in composition, although basalts are still present. Because the andesitic average composition of arc eruptives resembles the average composition of continental crust, it may be that arcs are the site of generation of new continental crust, just as mid-ocean ridges are the site of generation of oceanic crust (Marsh, 1979). Whether this is true depends on whether the silicic component of these arcs is newly derived from the mantle or merely recycled crust. Although the seismic and tectonic picture is quite clear, the petrologic picture is not. Subduction introduces a wealth of possible sources for melt generation: subducting mantle, subducting crust, subducting marine sediment, overlying mantle wedge, and overlying crust.

On the whole, subduction zones represent a cold rather than a hot region of the upper mantle. The planet's interior is being cooled here by introduction of cold crust. This is clear from thermal models and is shown by the anomalously deep earthquakes and high seismic velocities at depth (Marsh, 1979; Zhao *et al.*, 1995). These observations also show that thermal disequilibrium persists between slab and host mantle well past the depth of presumed melting. However, subduction does provide a route for introduction of water into the mantle. That this plays a role in melting is consistent with the water-rich character of subduction-zone magmas. Thus, it is logical to suggest that the profound freezing-point depression effect of subducted water is responsible for arc volcanism.

Direct evidence for rapid introduction of surface-derived volatiles into the mantle is provided by the presence of detectable ^{10}Be in these magmas (Morris *et al.*, 1990). Because ^{10}Be is generated only in the atmosphere and has a short half-life, its presence in subduction-zone magmas and apparent absence in other magmas is consistent with rapid (few million years) transport from surface to magma-source depth by subduction.

If water fluxing is responsible for melting, it likely causes melting in the overlying mantle wedge after release from the slab, rather than in the cold slab itself. Whether this primary melt is andesitic or is of basalt that later fractionates and/or causes additional melting in the crust to produce silicic associates is a matter of debate (Hess, 1989). Certainly, some role of crustal melting in andesite genesis is indicated by the correlation of the abundance of silicic volcanic eruptives with the continental character of the crustal substrate on which the volcanoes rest.

2.1.3 Hot Spots Trains of volcanoes that become regularly younger from one end to the other appear to represent the passage of a tectonic plate over a hot spot within the mantle. Hot spots are thought to reflect Earth-radial plumes arising from as deep as the core-mantle boundary (Sleep, 1992). The outstanding example is the Hawaii/Emperor Seamount chain, which ranges from zero age at the southeastern-most volcanoes of Loihi and Kilauea to 70-Ma Meiji Seamount, where the Emperor chain disappears into the Aleutian trench. The inferred rate of migration of the plate relative to the hot spot is 8 cm/year. Hot-spot volcanism is volatile poor and, in the ocean basins at least, predominantly basaltic. Like mid-ocean ridge volcanism, it most likely arises from decompression melting of ascending mantle. The relationship of such plumes to the large-scale mantle convection inferred from plate tectonics has not been satisfactorily resolved. Iceland appears to be a hot spot coincident with a mid-ocean ridge, and Yellowstone may be a plume impinging on continental crust, consequently causing crustal melting and voluminous silicic as well as basaltic magma production. Although this class of volcanoes is not as important as mid-ocean ridge and arc volcanoes in terms of total volumetric productivity, individual hot-spot volcanoes have the highest eruption rates of any volcanoes on Earth.

2.1.4 Other Occurrences Some volcanism does not fit neatly into the above categories but is usually associated with some sort of crustal deformation. Flood basalts may be the product of the initial arrival at the surface of the fattened head of a new mantle plume. Flood basalt provinces occur both on land in northwest coterminous United States, India, and Siberia, for example, and in the mid-Pacific as lava plateaus. Mafic volcanism behind volcanic arcs, known as back-arc volcanism, may reflect decompression melting in a small-scale mantle-wedge eddy, induced above the descending slab. Widespread, but infrequent and generally small-scale, volcanism in southwestern United States appears to reflect an exceptionally wide zone of crustal extension and shearing associated with the western transform boundary of the North American plate, perhaps inducing mantle diapirism and consequent decompression melting.

2.2 Ascent of Magma and Kinds of Eruptions

Although a few volcanoes, such as Kilauea Volcano in Hawaii, exhibit years-long periods of steady-state flow, most volcanoes are only intermittently active. The typical pattern for subduction-related volcanoes, Earth's dominant subaerial volcanic features, are repose periods of years to centuries interrupted by briefer episodes of multiple eruptive pulses. These magma systems are visualized as a balloon and straw, in which magma stored in a distinct chamber or diffuse mush zone is periodically tapped through a straw extending to the surface at the volcano's vent (Eichelberger, 1995). The straw or conduit is itself a remarkable feature, for it must be repeatedly reused to develop a volcanic mountain at the surface. Yet it is not an open pathway. In repose, it is filled with the solidified magma of the previous episode. Thus, rather than providing a weak route, the conduit may in some cases be stronger than its surroundings. Subsequent magma batches are guided by a common stress field and weak preheated pathway at depth, and perhaps the physical discontinuity represented by the cold conduit near the surface.

The episodic nature of volcanism may arise from the finite resistance of the system to magma flow, and the need to develop storage-region overpressure to overcome it. In detail, the cause of onset of activity is unknown, and may differ for the onset of an episode of activity versus a single eruptive event. One viable triggering mechanism is introduction of fresh primitive magma into the chamber. This suggestion, of course, merely moves the triggering problem deeper. Another possible trigger is second boiling. Crystallization of anhydrous phases, which must be occurring at the margins of a crustal magma reservoir, will necessarily cause exsolution of water vapor and, if chamber volume is held constant, high overpressure (Tait *et al.*, 1989). Triggering of eruptions by large earthquakes has been documented in a few cases.

One of the clearest demonstrations of an eruption trigger, though it does not explain the initial onset of the episode of rise of magma, is the 1980 Mount St. Helens eruption. Catastrophic failure of the volcanic edifice suddenly removed the confining pressure from a foamlike magmatic intrusion growing within the cone, causing it to explode (Eichelberger and Hayes, 1982).

Eruptions will stop when sufficient overpressure has been relieved in the source region or when enough lava has accumulated at the surface to balance the driving force of eruption (Jaupart and Tait, 1990; Stasiuk *et al.*, 1993). Both heat loss and water loss from the conduit, which act to cause crystallization of magma and increase viscosity of residual melt, will tend to stop flow. There may be a threshold flow rate required to offset these losses, below which the system locks up.

Regardless of why ascent to the surface begins, magma undergoes changes in response to the rapidly decreasing confining pressure and temperature of its environment. Release of water generally keeps pace with decompression, since nearly all eruption products are thoroughly dewatered. In contrast, crystallization lags such that dewatered melts may be erupted as supercooled liquids that congeal to glass, without any syneruptive crystallization at all.

Vesiculation of bubbles is an early effect of decompression, and one that can lead to violent explosive eruption. For a silicic, melt-rich magma containing 4 wt% water, bubble growth will begin at about 4 km depth (Sparks, 1978). By the time the confining pressure has declined to 100 bars, the vapor will occupy about 75 vol% of the system [Fig. 3(b)], a porosity typical of pumice. As mentioned previously, it is thought that further expansion then occurs by fragmentation to a dusty gas. This rock particulate/gas mixture is able to expand far more than the gas would by itself, because of the large amount of thermal energy contained in the entrained rock particles and their intimate contact with the expanding gas. If pressure and chemical equilibrium are assumed, with vapor pressure equal to the lithostatic load on the system, the 75% limit of bubble expansion would be reached, and fragmentation would occur when the magma ascends to about 500 m depth. There are, however, a number of untested assumptions in this scenario, and it may be that fragmentation actually occurs deeper or shallower. Laboratory experiments that simulate fragmentation by subjecting synthetic magma or analog fluids to sudden decompression may eventually solve this problem (Alidibirov and Dingwell, 1996).

If magmas often contain several weight percent water and closed-system decompression to one atmosphere pressure produces hundreds-fold expansion, then open-system behavior is required to explain effusive eruptions or lava flows. That is, a large amount of vapor must be separated from the magma prior to extrusion. The alternative is that effusive eruptions are the product of eruption of initially dry magma. In view of the petrologic similarity that coeruptive effusive and explosive eruptions products have in all aspects other than mode of eruption and the drastic effect that water has on phase equilibria, this seems unlikely. Rather, it appears that effusively erupted magmas must have begun with water contents similar to their explosive associates, and lost water such a short time before surface quenching that these losses are not reflected in the phase assemblage. The process must indeed be shallow, efficient, and rapid, because at plausible ascent rates of the order of 1 cm/s, magma with a few weight percent water would traverse the vapor-saturated pressure domain to the surface in hours or, at most, days.

That magmas can behave as open systems during the last few kilometers of ascent is suggested by permeability measurements and isotope-exchange experiments on pumice. It is also suggested by patterns of the variation of hydrogen isotopic composition in explosive to effusive eruption suites that are indicative of melt/vapor fractionation (Eichelberger, 1995). Permeability develops because the crowded bubbles in the magmatic foam begin to interconnect. Thus, there is a transition state between melt-continuous, closed-cell magmatic foam, and gas-continuous magmatic dusty gas in which both the melt and gas phases are continuous. To escape from the system, water need only diffuse micrometers out of melt into the nearest bubble, where it can flow as a vapor over meters to the margin or top of the conduit.

Attempts have been made at modeling this complex process. It seems clear that just as narrowness of conduit and conductivity of wall rock favor cooling, narrowness of conduit and permeability of wallrock favor degassing. Further, the more slowly magma rises, the more likely porous-flow degassing can keep pace with declining confining pressure, so that excess vapor pressure does not develop and fragmentation and explosive ejection do not occur. Other characteristics of lava as opposed to tephra, such as larger reaction rims surrounding hydrous phenocrysts in dewatered melt, higher matrix crystallinity, and lower chlorine content as compared with explosively erupted associates, are most easily explained as a consequence of slower rise of effusive magmas (Rutherford and Hill, 1993).

In accord with these observations, it may be that ascent velocity of magma is a fundamental controlling factor in eruptive behavior. At high ascent rates, neither degassing nor crystallization occurs, and explosive showers of fragmented glass are produced. At moderate ascent rates, degassing occurs but outpaces crystallization, so that glassy lava is extruded. At still slower ascent rates, both degassing and crystallization occur, with the result that partially crystalline domes appear. Ascent rate itself is in turn controlled by conduit size, magma viscosity, and chamber overpressure. Because of these relationships and contrary to common expectation, explosive volcanism is not limited to viscous silicic magmas, nor is effusive eruption to fluid mafic ones.

3. SOCIETAL IMPACT

3.1 Geothermal Energy

Associated with shallow magma chambers are subsurface reservoirs of natural steam or of superheated water (Smith and Shaw, 1975). If the energy content is sufficient, steam from these reservoirs can be used to run turbines and to generate electricity. If not, the steam or hot water can be used directly for space heating. Geothermal energy production offers greatly reduced CO_2 and SO_2 release compared to the burning of fossil fuels and is particularly attractive when fossil fuels are not locally available.

The United States, Philippines, Mexico, Italy, Japan, and New Zealand are the largest producers, in that order. Growth in use of this resource is limited by its rather restricted geographic distribution; the fact that in most cases it is not economically competitive with fossil fuels; and environmental concerns pertaining to noise, release of toxic gases, potential for damage to natural hydrothermal phenomena such as geysers, and aesthetic considerations. In regard to the latter, most active volcanic fields possess scenic and recreational values as well as geothermal potential and are often protected with park status. Indeed, objectives of productivity, preservation, research, and enjoyment are often at odds in volcanic areas (Eichelberger and Sattler, 1994). Nevertheless, geothermal energy is being vigorously pursued where it is a matter of national policy to lessen dependence on other energy sources, as in Japan.

3.2 Climate

The impact of volcanism on climate is one of the most interesting, important, and contentious topics in volcanology. It is interesting because it has likely shaped the past, and may shape the future, of humankind. It is important today because it is the only volcanic hazard that is global in extent and from which, therefore, there is literally no escape. It is contentious because cause and effect in climatic perturbations are very hard to assess, and once assessed they are laden with political implications concerning industrialization, have and have-not nations, and governmental intrusion into the private sector. It would be reassuring to know that all important sources of pollution and "global warming" are natural, beyond our control, and therefore without moral implications. This is not the case. However, neither is it the case that present and future climate anomalies can be attributed solely to humankind fouling its own nest.

Research into volcanic effects on climate has been driven by two factors. One is the anecdotal association of major eruptions with "strange" weather. Among the first to suggest a relationship between volcanism and climate was Benjamin Franklin who, while U.S. Ambassador to France, speculated that the faint sun and poor visibility in Europe in 1784 were caused by a voluminous eruption that was ongoing in Iceland (Francis, 1993). Icelanders were probably more convinced: they lost a quarter of their population and half of their farm animals to the ensuing famine and volcanic fluorine-induced toxicity. Denmark, which then ruled Iceland, considered evacuating the entire island. Suspicions of climatic influence have likewise been entertained following the 1813 eruption of Tambora in Indonesia, the 1883 eruption of Krakatau also in Indonesia, the 1912 eruption of the Valley of Ten Thousand Smokes in Alaska, and the 1991 eruption of Pinatubo in the Philippines. None of these, however, are within even an order of magnitude of the size of the largest eruptions that have occurred since *Homo sapiens* has been extant (Rampino and Self, 1993).

A second driving force in the study of volcanic effects on climate is the evaluation of the role of catastrophe in shaping life on Earth. The fossil record makes it clear that Earth has experienced several mass extinctions of species. The most obvious possible cause is global climate change, for which the source must be either within the planet or without. In particular, a great debate has flourished about the sudden end to the 60-million-year reign of dinosaurs. Present evidence, in the form of global dispersal of meteorite-abundant material at the stratigraphic level of the extinction and identification of a huge crater buried under the Yucatan Peninsula of appropriate age, favor the threat without. But the debate has shown large-scale volcanism to be of potentially serious impact to life worldwide (Rampino *et al.*, 1988).

Establishing a correlation between size of an eruption and perturbation of climate in a rigorous way has only recently become possible. It relies on satellite observations of atmosphere loading from large events and myriad surface and satellite-based observations of surface conditions. Some of the results of these investigations are counterintuitive.

Modest to large volcanic eruptions produce an ash veil that shrouds the Earth. Obviously, such a veil reflects solar energy, and so can cause cooling. However, it also acts as a blanket that rejects heat loss. Those who live at high latitudes can appreciate that there is little to be lost from diminished sunlight in winter, which is already minimal or nonexistent, but much to be gained from a blanket on a dark, several-month night. Thus, an early effect of great eruptions is warm winter temperatures at high latitudes. A longer-term effect is decreased solar input and consequent general cooling.

A number of parameters other than eruption size affect impact. Ash and gases play separate roles. The former affects planetary albedo, while the latter, especially SO_2, which forms H_2SO_4 aerosols, affect atmospheric absorption of incoming and reradiated solar energy. Low latitude of an eruption site facilitates global dispersal, whereas emissions from high-latitude volcanoes spread more slowly. Only those eruption plumes that penetrate the stratosphere, where high-velocity horizontal winds predominate, are likely to have widespread climatic effects. An exception is when emissions at low altitudes are sustained, as in Iceland's Laki "fires" of 1783 to 1784.

The largest eruptions since reliable, well-distributed weather records have been kept are several to ten cubic kilometers in magma volume. The only such "large" eruption that has occurred since satellite observations began is Mt. Pinatubo in the Philippines in June 1991. Various analyses place the long-term cooling effect of this size class of eruption at a few tenths of a degree Celsius and the immediate high-latitude winter warming effect at about a degree (Bluth *et al.*, 1992; Graf *et al.*, 1993). Although these changes seem slight, effects on length of growing season may bring significant consequences to humans.

But what of truly great eruptions of thousands of cubic kilometers of magma? The ash fall alone from such events would be catastrophic on a continent-wide scale, and the atmospheric effects would be comparable to the much-feared "nuclear winter." It is postulated that such eruptions could globally reduce daylight to about the brightness of a full moon, halting photosynthesis (Rampino and Self, 1993). Although the fervor for attributing mass extinctions to such events has subsided, their possible triggering of swings in glaciation remains intriguing. Moreover, severe impact to the populations of geographically restricted species, such as early hominids in East Africa, is possible. Such impacts might explain the "evolutionary bottlenecks" that our ancestors appear to have suffered. At least roughly coincident with the most recent bottleneck and the subsequent explosive expansion of *Homo sapiens* out of Africa was the 3000-km^3 eruption of Toba in Indonesia 74 000 years ago (Stringer and McKie, 1997). Perhaps even more basic to our existence, the very origin of the family *Homo*, the break with the family *Pan*, has been credited to the rise of the volcanic East African Rift, providing a barrier to migration and to climate (Coppens, 1994). *Pan* kept to the trees in the lush forests of the west, but *Homo* developed bipedal talents in its drying preserve to the east.

3.3 Volcanic Hazards

The potential for catastrophe was mild in the days of Pompeii compared to what it is today. Humankind is no more able to control volcanic forces than it was 1900 years ago, but orders of magnitude more people and property are at risk (Decker, 1986). Millions of people live on active volcanoes and millions more fly within range of potentially lethal ash clouds every year. Two rather small eruptions this century each killed almost 30 000 people, and civilization has not yet experienced a large eruption near a major population center. The possibility of small undetected eruptions from remote volcanoes is, for the first time, another cause for concern. An eruption-caused crash of a wide-bodied jet, while not as catastrophic as the destruction of a town or city, is a very serious event and a plausible one, given, for example, the invasion of North Pacific air routes by ash clouds on average four days of the year (Miller and Kirianov, 1995; Fig. 6). Ash clouds cannot be distinguished from weather clouds on aircraft radar and can cause engine failure due to abrasion and/or coating of engine parts.

Concern for volcanic hazards is now the largest driving force in the field of volcanology. Numerous governments and academic institutions have set up observatories for the dual purposes of monitoring volcanic events to mitigate hazards and of studying eruptions to understand magmatic processes. Emphasis has grown on investigating the shallow ascent and eruptive behavior of magma and on interpretation of surface deposits in order properly to assess hazards and identify precursory signals of eruption. As a result of this work, remarkable three-dimensional views of active magma systems have been produced.

4. FORECASTING ERUPTIONS

One of the most important achievements in volcanology in recent years is development of the capability to predict eruptions (Decker, 1986; McNutt, 1996). This capability is a useful but still a somewhat limited one. Although it is often possible to warm of the onset of an episode of activity, later eruptive pulses within an episode sometimes lack recognized precursors. It is seldom possible to determine when an eruptive episode has ended, information that could be as important to local residents as sensing its beginning. Finally, it is generally not possible to predict whether an imminent eruption will be explosive or effusive, or indeed whether the magma may stop just short of the surface as a shallow intrusion. The latter situation leads to troublesome "false-positive" results.

4.1 Characteristics of Useful Predictions

To be useful, a forecast must be accurate and specific. Inaccurate warnings may more than offset any benefits of correct ones. If vague, decisions to suspend or move affected activities cannot be implemented. For example, it does little good, other than for general readiness, to know that there is some possibility of an eruption in the next ten years. Once a warning of imminent eruption is issued, costly measures such as rerouting aircraft, suspending business and agriculture, and caring for evacuees must be undertaken. For minimizing economic losses, it is therefore essential that serious warnings be lifted as soon as the danger has passed.

4.2 Methods of Monitoring and Prediction

Although some observatories predate it, the idea of permanent stations from which to monitor active volcanoes came into its own following the obliteration of the town of St. Pierre by Mount Pelé on Martinique in 1902. The modern era of volcano monitoring began with the explosive eruptions of Mount Usu in Japan in 1977 and of Mount St. Helens in the United States in 1980. By this time, computer systems were able to process large amounts of data rapidly, and governments were willing to commit substantial resources to natural-hazard mitigation. The use of dense seismic networks on suspect volcanoes emerged as the primary tool of prediction, although other kinds of observations remain important.

4.2.1 Eyewitness Reports More than 1900 years after Pliny the Younger's account of the eruption of Mt. Vesuvius, eyewitness records are still the most useful information source for understanding eruptions. Such observations can be used as a predictive tool in the case of large eruptions where precursory ground shaking and cracking may be severe, or when eruptions are preceded by steam emission. Large explosive eruptions and their aftermath are best viewed from space, as lofted ash soon obscures proximal views. Thus, we have only obtained a complete picture of such events in the era of satellites.

Visual observations must be evaluated carefully. Even knowledgeable observers may mistake weather clouds for eruption clouds or old volcanic features for newly formed ones. Remotely placed and telemetered television cameras are very useful where it is impractical or dangerous to station human observers.

4.2.2 Seismology An obvious indicator that an earthquake may be of magmatic origin is that it occurs directly under a volcano (Fig. 12). Swarms of high-frequency earthquakes herald the rise of magma. They are caused by breakage of cold rock just outside the conduit or magma chamber as these structures deform with the movement of magma. Because these events involve brittle rock and therefore low temperature, a surface of seismicity enclosing an aseismic vol-

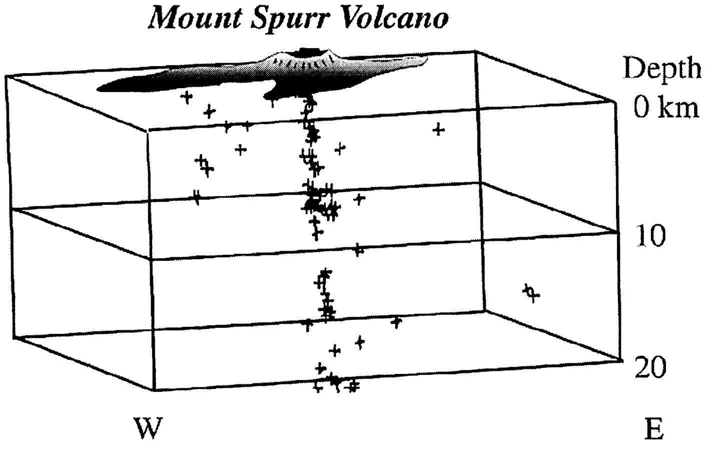

FIG. 12. Seismic events (crosses) beneath Spurr Volcano, Alaska, during its 1992 eruptive episode define the position of a magma conduit that leads from the base of the continental crust (40-km depth) to the surface (Power *et al.*, 1995). Figure by Gail Davidson, Alaska Volcano Observatory.

ume is evidence for a magma chamber. However, some high-frequency, tectonically induced events unrelated to magma movement occur beneath volcanoes. Volcanic earthquake swarms differ from tectonic swarms in that the largest events may occur anywhere in the swarm rather than at the start. In the tectonic case, much of the activity is "aftershock."

The most robust seismic indicators of impending eruption are low-frequency earthquakes and harmonic tremor (Power *et al.*, 1994; McNutt, 1996). Low-frequency events appear to arise from transient pressure pulses due to sudden release of magmatic or hydrothermal fluids (Chouet *et al.*, 1994). They accompany resonance of fluid flowing rapidly in pipes or cracks. During or sometimes just before eruption, these may punctuate or blend into a continuous signal with peak amplitude typically near 2 Hz, the constant humming that is volcanic tremor.

Precursor seismicity may occur over a period of days or even years prior to eruption. It may increase in intensity and decrease in depth. This is not necessarily the result of the slow fitful journey of a single magma body to the surface. Even at the beginning of eruptions following lengthy precursor activity, fresh-appearing magma giving no evidence of shallow storage may emerge. Rather, shallow pre-eruption seismic swarms may reflect repeated damming of intrusive pulses by the cold, water-soaked "lid" of the system (Eichelberger *et al.*, 1995). For reasons not yet understood, seismicity may also begin at shallow depth and progress to include deeper levels, until the entire conduit is seismically illuminated (Fig. 12).

If patterns of seismicity were regularly repeated, then eruption forecasting would be a simple matter and other tools would not be needed. However, this is not the case. Seismic crises occur at volcanoes without accompanying eruptions. Conversely, eruptions occur without recognized precursory seismicity. It is evident that magma can move without generating seismic events, perhaps when using preheated pathways. Sources unrelated to magma, such as hydrothermal systems, tectonic stresses, glaciers, and sea ice cause seismicity at volcanoes. Added to these complications are logistical considerations. Remote volcanoes are difficult to monitor seismically because of access, power, and data transmission requirements. Nevertheless, where seismic monitoring is feasible, it is relied upon as the first warning of unrest.

4.2.3 Geodetic Measurements Early surveying on Kilauea Volcano showed that volcanoes inflate prior to eruption and deflate during eruption (Decker, 1986). Inflation is thus a sign of impending activity, though, like seismicity, not an infallible one. Not all cycles of inflation and deflation are associated with eruption. Alternating pressurization and depressurization of a hydrothermal system are implicated in some cases (Fournier, 1989). However, coincidence of seismicity and inflation in time and space is strongly suggestive of magma migration. In addition, geodetic data constrain the volume of newly emplaced magma and can be inverted to estimate the location of the center of the inflating or deflating volume (e.g., Rundle and Hill, 1988). As a monitoring tool, geodetic measurement is undergoing a revolution because of changes in techniques. The original approach, surveying, was tedious and labor-intensive. The advent of public access to the Global Positioning System (GPS) made rapid measurement at multiple surface points relatively easy, and even real-time, seismic netlike, geodetic monitoring feasible. Still more recently and currently under development, interferometry using images from active synthetic-aperture radar (SAR) satellites has been applied to produce maps of centimeter-scale deformation covering entire volcanic fields (Lu *et al.*, 1997).

4.2.4 Gas Measurements Because gas is released from rising magma and can flow much more rapidly than the magma, it is natural to consider the surface emission of magmatic gases as a precursor to eruption. Unfortunately, there are many reasons why magmatic gas may not reach the surface and why, even if it does, it may not be identifiable. Magmatic gas undergoes a series of chemical reactions when passing from magma to a cooler environment (Symonds *et al.*, 1994), must follow a tortuous path to the surface, and will likely encounter a hydrothermal system during ascent. In many cases, the dominant magmatic volatile component is water that can only be distinguished from groundwater by its isotopic composition. Even then, the introduced magmatic component may be of such small volume relative to near-surface water that any shift in isotopic composition of springs or fumaroles due to rise of magma may not be detectable.

Carbon dioxide is relatively abundant in magmas and insoluble in water. Carbon dioxide emissions from magma, distinguishable in isotopic composition, have been detected with some success (Casadevall *et al.*, 1983). Unfortunately, carbon dioxide is also relatively insoluble in magma at crustal pressures and hence may be released while the magma is still at great depth. Its usefulness as a sign of imminent eruption is therefore limited. At this writing, release of carbon dioxide from a relatively deep intrusion beneath Mammoth Mountain, California, is killing trees and small animals over an area of many acres (Hill, 1996).

For using gas measurements to guide eruption forecasts, most successes to date have come from measuring SO_2 emissions. SO_2 is relatively soluble in magma at crustal pressures, so that close approach to the surface by magma is required for its release. Unfortunately, SO_2 is also highly soluble in water, so that it can be "scrubbed" from the magmatic gas stream by an intervening hydrothermal system (Doukas and Gerlach, 1995). High rates of SO_2 emission can therefore be taken as a sign that magma is approaching the surface, but low rates of emission do not necessarily mean it is not.

Measurement of SO_2 can be done remotely through spectroscopy because SO_2 is a strong absorber of ultraviolet light. Before or after eruptions, real-time estimates of SO_2 flux from a volcano can be obtained by combining wind-velocity data with ground or airborne spectroscopic observations. After an explosive event, an estimate of total erupted SO_2 in ash clouds can be obtained from satellites that were originally intended to measure concentrations of another UV-absorber, ozone. For example, from such data it was determined that the June 1991 eruptions of Mount Pinatubo in the Philippines released 2×10^7 tons of SO_2 (Bluth *et al.*, 1992), an amount that approaches the order of annual global anthropogenic production.

4.2.5 Geology As in the case of weather forecasting, persistence of a trend is a likely scenario. A volcano is likely to erupt in a way that it has been erupting during the most recent portion of its history (Decker, 1986). There are, of course, spectacular examples of failure of this guide, such as the 1912 eruption of the Valley of Ten Thousand Smokes in Alaska and the 1980 eruption of Mount St. Helens in Washington. In the current absence of better understanding, however, it is best to use seismology to sound the alarm and geology to suggest the likely course of events. Two parts of the geologic record must be considered. One is the volcanic edifice itself, which contains the entire history of the volcano, though often concealed by ice or young eruptive deposits. The other is the far-field tephra stratigraphy, which sometimes provides a more accessible record. It may also be a more datable record, if, as is common at lower elevations, the deposits are intercalated with or incorporate organic debris. To be useful, such stratigraphy must be positively tied to a particular volcano, a difficulty in volcanic fields where eruptive centers are closely spaced and of similar composition. This is often accomplished through use of mineralogical or trace-element clues. Even a good tephra sequence, say from sediments in a lake neighboring a volcano, will be incomplete, as it will only contain products of eruptions that had an explosive component, and then only major eruptions under conditions of certain wind directions. Small events, which nevertheless could significantly disrupt modern society, may leave little or no geologic record. In addition to establishing at least a lower limit on the frequency of activity, geologic observations can establish what areas around the volcano are at risk during an eruption from such hazards as pyroclastic flows, lahars (mudflows), debris avalanches, and lava flows. It is clearly wise to prepare maps of hazard zones in advance of eruptive activity, so that they can be used to advantage in the ensuing civil emergency.

4.2.6 Remote Sensing Because of the difficulty of access of many volcanoes, reliance on various satellite-based techniques may be required for monitoring them. It is weather satellites that are most often used for this purpose, as they provide frequent and thorough aerial coverage, though not with high resolution. With pixels of the order of 1×1 km^2, the typical volcano occupies only a small part of the image. Weather-satellite infrared data are useful for two purposes. One is detecting hot spots. Eruption products and erupting vents are hot, though they may be obscured by already cooled ash-cloud tops. Heating may significantly precede eruption, as magma rises within the cone. As with the rise of magmatic gas, however, it is not always a reliable indicator of impending activity. Magma flows much faster than rocks conduct heat, so that pre-eruption surface heating requires fortuitous convective transport of heat by water or gas.

Infrared data have also been found to be useful in distinguishing ash clouds from weather clouds after the cloud has left the vent and become translucent (Wen and Rose, 1994). The presence of ash in a cloud causes changes in the spectrum of infrared radiation that is emitted from the Earth's surface and passes through it. This is important because cloud plumes of purely meteorological origin are common in mountainous terrain. Hot-spot and ash-cloud detection are primary techniques of eruption confirmation, however, not of forecasting. The technology of detecting eruptive events by satellite has much in common with problems of detection of nuclear blasts, missile launches, wild fires, and anthropogenic pollution events.

4.2.7 Other Techniques A variety of other approaches have been applied to volcano monitoring with varying degrees of success. Ash eruptions are often accompanied by lightning, and so lightning-detection systems can be useful in confirming eruptive activity (Paskievitch *et al.*, 1995). Ground-based radar can see eruptive columns, but usually only the coarse, near-vent portion during actual eruption. Eruptions release acoustic energy as well as seismic energy, and this can be detected through the atmosphere by means of pressure transducers (Buckingham and Garcés, 1996). A particularly interesting and innovative suite of physical measurements was obtained in the recent eruption of Unzen Volcano, Japan (Yanagi *et al.*, 1992). This eruption was amenable to close observation because it was long-lasting and highly accessible, occurring in an almost urban setting.

5. CASE HISTORIES

Any theory of volcanism and any response to volcanic hazards must take into account the basic characteristics of volcanic activity. The following chronologies present examples of the events surrounding large (10^3 km^3), medium (10^1 km^3), and small (10^{-1} km^3) eruptions. These chronologies also show contrast in the kind and level of detail they contain. The first case is wholly reconstructed from the geologic record, though the current state of unrest of this volcano is cause for concern. The second is reconstructed from accounts of lay observers and from geological exploration beginning soon after the eruption. The last case is of an eruption of the modern era of instrumented volcanoes.

5.1 Long Valley Caldera, California

Among large continental volcanic systems, the Long Valley Caldera–Mono Craters volcanic complex in eastern California (Fig. 11) is one of the most thoroughly studied (Bailey *et al.*, 1976; Rundle and Hill, 1988). It is also presently the most active silicic system within the contiguous United States. Its most recent major eruption occurred only 550 years ago, and there is a possibility that a minor eruption occurred beneath Mono Lake in 1890. Beginning in October 1978, a sequence of moderate earthquakes occurred northwest of Bishop, California, spreading northward to Long Valley Caldera and southward into the Sierra Nevada mountains. Four of these events had magnitudes larger than 6.0. Recurrent swarms followed. Suspicion that this activity could be linked to the movement of magma arose when it was discovered that the central caldera region was undergoing rapid, blisterlike uplift and extension amounting to several tens of centimeters. A swarm beneath Mammoth Mountain on the southwest caldera rim was accompanied by the dangerously increased and continuing CO_2 emission cited earlier (Hill, 1996).

This volcano has experienced more seismicity, surface deformation, and gas emission than many volcanoes that have progressed from such signs of shallow magmatic intrusion to violent eruption. Even a small eruption would have serious social and economic impacts at the town of Mammoth Lakes, California, located within the caldera. A further reason for interest is that this volcano has produced gigantic eruptions in the past, including an eruption so large that ash beds from it are preserved in the mid-continent. The caldera may mark the site of a large modern magma body in the crust.

Long Valley Caldera lies astride huge, steeply inclined faults of the east front of the Sierra Nevada block. There can be little doubt that it owes its existence to the rise of magmas along this fault zone, though the broader issue of the volcano's relation to plate tectonic forces is problematic. The site became a locus of basaltic to andesitic volcanism 3–4 million years ago. Beginning about 2 million years ago, high-silica rhyolite flows and tephra began erupting from what is now the northeastern margin of the caldera. These products, which accumulated in substantial volume, appear to mark the development of a long-lived, chemically evolved magma chamber in the upper crust and set the stage for what was to follow. That silicic volcanism lagged behind basaltic and andesitic volcanism is a common pattern and suggests that the crust had to be thermally prepared for the establishment of such a reservoir by the rise of hotter, mantle-derived magmas. At 760 000 years ago, a catastrophic eruption of high-silica rhyolite began in plinian (so named because of Pliny the Younger's description of the early eruption column from Vesuvius in 79 A.D.) fashion as a vertically directed high-velocity flow from a central vent, causing widespread tephra fallout. Enough magma was withdrawn that the roof of the magma chamber soon foundered, and a ring fracture around the roof block ripped open (Hildreth and Mahood, 1986). The character of the eruption changed, and pyroclastic flows swept outward from the open ring vent toward the lowlands. In total, some 1000 km^3 of magma erupted, probably within a period of days or weeks. The roof block subsided about 2 km and, with the inward slumping and erosion that followed, left a caldera basin that is 20 $\times$ 30 km^2.

Was this eruption merely the skimming of a volatile-rich top, perhaps the upper 10% (Smith, 1979), of an even larger magma body that continues to evolve to this day, or did it completely exhaust the crustal supply of magma then existing? The answer is not clear. Persistent silicic volcanism associated with the caldera and some geophysical evidence suggests that such a large chamber still exists, but drilling to more than 2 km depth in the center of the caldera has failed to find the heat flow expected above a long-lived shallow heat source.

Following caldera formation, volcanism waned with smaller (though still substantial) rhyolite eruptions of increasing crystallinity. The system was apparently thermally rejuvenated 200 000–100 000 years ago when basaltic to andesitic magmas rose under the caldera and flooded its western floor. With the coming of late Pleistocene to Holocene time, crystal-poor rhyolites are once again being erupted, and the development of a large new crustal chamber whose heat has not yet reached the surface is a plausible, though by no means proven, scenario.

5.2 Valley of Ten Thousand Smokes, Alaska

By far the largest eruption of this century burst from the Mount Katmai region in the Aleutian Arc on 6 June 1912 (Hildreth, 1983). The site was among a cluster of imposing andesitic volcanic cones near where the Alaska Peninsula connects with the Alaska mainland. It may be that focusing of heat in the crust, as represented by this cluster, led to the silicic and voluminous character of the subsequent volcanic activity. But there was no prior hint of the cataclysm to come. A geologist who visited the site in 1898, 14 years before the great eruption, pronounced the volcanoes "inactive." The "persistence of a trend" guide for forecasting would have failed completely in this case, as the size, violence, and composition of the eruption were unprecedented for the district in which it occurred. So large was the eruption that, despite its extreme remoteness at the time, it became famous throughout the world and sparked some of the scientific debates that frame modern volcanology.

Seismicity in the weeks preceding the eruption was of such intensity that residents of the villages of Savinoski and Katmai Bay had evacuated by the time the eruption began. Consequently, there were no known casualties. The first and only reliable direct eyewitness report came from the ship Dora, sailing in the Shelikof Strait about 100 km to the east. At 1 p.m. on 6 June, passengers sighted an "immense" ash column ascending into rare blue sky above the Aleutian Range. This column likely reached 30 km altitude and, within a few hours, engulfed the Dora and then the town of Kodiak, 160 km from the eruption site. Most other eruption accounts are of the ash fall at Kodiak, which was cloaked in choking darkness for the next 60 h. Records from Kodiak and careful field work near the volcano make it possible to reconstruct the main aspects of the eruption (Hildreth, 1983, 1987; Fierstein and Hildreth, 1992).

The eruption site was not Mount Katmai, the prominent peak of the region that lost its summit in the event and that early workers held responsible, but a new vent that opened 10 km away at the base of neighboring Trident Volcano. The initial activity was the extremely powerful, vertically directed eruption seen by the Dora. This early stage was entirely of a crystal-poor, high-silica rhyolite, a magma type almost unknown in the Aleutian Arc and more commonly erupted from volcanoes of continental interiors. The ash column rose into the stratosphere and then spread out as a mushroom cloud that was stretched rapidly by strong northwest winds toward Kodiak, where fallout ultimately reached 30 cm in thickness. Within hours, dense mixtures of ash and gas began surging outward from the margin of the vent and sweeping down a broad glacial valley as pyroclastic flows. A combination of reaming and slumping of the vent formed a funnel-like structure of perhaps a kilometer in depth but at least a kilometer in surface radius. Growth of this hole undermined an adjacent, 700-m-high dacite dome, which partially collapsed into it. The pyroclastic flows lost energy as they flowed down valley and were in part dammed by a glacial moraine about 20 km away. The resulting tuff sheet attains a thickness of some 200 m in the upper valley (Fig. 10).

Sometime around midnight on 6 June, the eruption changed. Rhyolite was joined by crystal-rich dacite and andesite, apparently mixing together both in the conduit and after fragmentation within the eruption column. The source of this new influx of magma is clear, because at about the same time seismicity intensified and debris flows were shed from the towering Katmai cone. Mount Katmai began a fitful collapse, with most of the subsiding summit volume disappearing into the emptying subjacent magma chamber. The eruption became less energetic and soon ceased, but it resumed again at about noon on 7 June, this time with a plinian eruption of dacitic pumice and ash with subordinate andesite and virtually no rhyolite. Another similar pulse followed on the night of 8 June. This ended the explosive activity, but soon thereafter, a small rhyolite dome, Novarupta, emerged at the center of the vent crater (Fig. 7).

Altogether the eruption released about 30 km^3 of pumice and ash, equivalent to about 13 km^3 of bubble-free magma, in less than 60 h. Half this volume is represented by the extensive air-fall tephra deposits and the other half by the tuff flooring of what became known as the Valley of Ten Thousand Smokes. The volume of Mount Katmai lost in collapse, and now partially refilled by a spectacular caldera lake, is equivalent to about half the erupted magma volume. Through this catastrophe, Trident Volcano remained a close but innocent bystander (although it effused a half cubic kilometer of chemically distinct andesite and dacite lava rather placidly during the 1950s). This one huge eruption at Novarupta exhibited almost the entire common range of eruptive behavior and magma chemistry and produced a volume equivalent to that of a moderate-sized volcano in a single very short event. It demonstrated the close spatial, and therefore by implication the close chemical, relationship of contrasting magma types. It showed that even andesitic volcanoes that commonly produce small-volume eruptions sometimes have large-volume crustal magma chambers. Finally, by the 10-km separation of the sites of contemporaneous eruption and collapse, it showed the remarkable subsurface mobility of even silica-rich magmas. The system remains seismically active, and there is geophysical evidence for a large volume of magma remaining at shallow depth beneath the volcanic cluster (Ward, 1991).

5.3 Crater Peak, Mount Spurr Volcano, Alaska, 1992

Short explosive outbursts typify many volcanoes of the Pacific rim. While they leave little by way of geologic record, they can nevertheless be quite disruptive of human activities. One such eruption occurred recently at Mount Spurr, which lies 125 km west of Anchorage near the northeastern end of the Aleutian Volcanic Arc. The volcano is

100 km above the top of the subducting slab and 500 km inboard of the Aleutian Trench (Zhao *et al.,* 1995) on a basement of mainly intrusive rocks (Nye and Turner, 1990). Its large edifice, built of andesite lava and tephra, began to form more than 250 000 years ago. A few tens of thousands of years ago, the volcano collapsed in Mount St. Helens fashion, forming a 25-km-long debris avalanche and leaving the volcano with a horseshoe-shaped caldera rim. A large andesite dome complex, which now forms Mount Spurr's 3374-m summit, grew within the caldera about 5000 years ago. The presently active eruptive center, the 2309-m Crater Peak, is a more mafic satellitic cone that grew in the caldera's breach. This vent has produced about 40 recognizable tephra deposits in the last 5000 years (Riehle, 1985). Eruption products that are apparent mixtures of Crater Peak and summit-dome complex magmas indicate that the two systems are connected at depth despite the 4-km separation of their respective vents.

Prior to 1992, the only witnessed eruption of Mount Spurr had occurred on 9 July 1953. This eruption also issued from the Crater Peak vent and consisted of two explosive pulses, followed by steam and minor ash emission (Wilcox, 1959). In June, August, and September of 1992, short, violent eruptions each produced about 50×10^6 m^3 of tephra, each representing roughly 15×10^6 m^3 of magma. Ejecta ranged from proximal bombs to distal ash comprising narrow bands of fallout extending downwind (Neal *et al.,* 1995). Save for a tens-of-meters westward migration, vent geometry changed remarkably little during the episode.

5.3.1 Eruption Chronology Early indication of unrest at Mount Spurr was a conspicuous swarm of earthquakes during the last half of August 1991 (Eichelberger *et al.,* 1995). Unlike swarms that had occurred in 1982 and 1989, this one was directly under Crater Peak within a volume that had been aseismic for at least a decade, and it was followed by caldera-wide seismicity. By February of 1992, seismic activity had increased to such a level that discussions of Mount Spurr's unrest were included in weekly public updates from the Anchorage- and Fairbanks-based Alaska Volcano Observatory (AVO), which operates a seismic net on the mountain. By early June, AVO seismologists concluded that the seismicity was caused by intrusion of magma into conduits below Crater Peak and the summit-dome complex, and associated readjustments along caldera faults. AVO advised government agencies, communities, and the aviation industry to review plans for coping with an explosive eruption.

The number of earthquakes strong enough to be located peaked at 28 per day on 5 June and then declined to about three to six events per day. Volcanic tremor began on 6 June, and aerial observations on 8 June revealed upwelling in the crater lake, which had turned in color from green to gray. As many as 24 shallow tremor bursts per day, each burst lasting 1 to 10 min, were recorded at stations within 10 km of Crater Peak from 6 to 26 June (McNutt *et al.,* 1995). On 11 June, a field party measured a lake temperature of 50 °C and observed geysers erupting intermittently at the base of the crater wall.

5.3.2. 27 June 1992 Eruption Seismic behavior changed ominously on 24 June with occurrence of hours-long periods of tremor. AVO's weekly update on the morning of 26 June reported "well above normal" seismic activity but still cautioned that an eruption might not be imminent because low-frequency earthquakes, indicative of fluid flow at depth, had not been detected. Aerial observations of the crater revealed that the lake had almost completely drained or evaporated and that several large rocks had been ejected from the vent. That afternoon, a tremor that was continuous and stronger than earlier bursts began. AVO formally issued a warning that an eruption was likely and went on 24-h duty. Early the next morning, a swarm of high-frequency earthquakes, accompanied by the first low-frequency events, struck at 0- to 2-km depths beneath Crater Peak. Their rate soon increased to about one every 2 min. This swarm signaled the final surge of magma toward the surface. 4 h later at 7:04 a.m. on 27 June, tremor amplitude abruptly doubled and the eruption began. 12 min later, telemetry was lost from a seismic station 400 m from the vent and a commercial pilot reported that an eruption plume had risen 5000 m above the cloud cover. Tremor amplitude continued to increase and registered on seismic stations more than 100 km away. Using radar, the National Weather Service (NWS) measured a maximum plume height of 14 500 m. Wet debris flows swept southward off the snow-clad Crater Peak cone. At about 11:30 a.m., seismicity decreased abruptly and the eruption was over.

Clear weather the following day revealed a black, northward-broadening swath of tephra on white snow fields. The eruption cloud dusted Denali National Park, continued northward to the Beaufort Sea, then turned southeast into Canada and the contiguous United States, where it became indistinguishable from weather clouds about 2 July. Because of the initial northward direction of travel of the ash, there were few effects other than curtailment of trips by ski-equipped planes to the ash-covered glaciers near Mt. McKinley.

AVO declared the danger of eruption over on 8 July. This was based on greatly reduced seismicity and SO_2 emission, and on the precedent of a single day of activity in 1953. This conclusion proved to be premature.

5.3.3 18 August 1992 Eruption Recorded seismicity remained low through the first half of August, but this was in part because AVO was deafened by demise of the crater rim seismic station. On the afternoon of 18 August, the hum of weak tremor and low-frequency earthquakes reappeared. Minor ash emission prompted an alert from AVO, and shortly thereafter a violent eruption began at 4:42 p.m. ADT, when strong tremor was recorded by all Mount Spurr seismic stations. Within 16 min the eruption column reached a height of 11 000 m, ultimately reaching nearly 14 000 m. A satellite image 44 min after the onset of eruption shows the plume extending 80 km east from the volcano and covering an area of 2000 km^2. Three hours after onset of eruption, the leading edge of the plume was 300 km southeast of Mount Spurr, and its area had grown to 21 000 km^2.

Close observation from aircraft revealed a dark roiling cloud, periodically surrounded by lenticular shock waves, jetting from the full width of the Crater Peak vent (Fig. 1). Large blocks were being hurled nearly 1000 m above the vent. Small-volume pyroclastic flows descended the east and southeast flanks of Crater Peak. Other flows mixed with snow and ice were to become mudflows on the lower flanks of the cone. A late fusillade of blocks, some as large as 1 m, were thrown obliquely as far as 10 km southeastward. More than 170 lightning strikes were detected during the second half of the eruption. The eruption ended abruptly 3.5 h after it began. Ashfall from the drifting cloud forced the closing of Anchorage International Airport for 20 h, where about 3 mm of ash fell, and caused respiratory and visibility problems within the City of Anchorage.

5.3.4 16–17 September 1992 Eruption Following the August eruption, deep seismicity gradually increased, and by mid-September it had returned to levels comparable to mid-June. Teams visiting the crater on 7 and 16 September noted nothing unusual. However, at about 7:30 p.m. on 16 September, earthquakes and weak tremor were detected by the newly reinstalled crater rim seismic station. Tremor amplitude increased 3 h later, and AVO issued an emergency announcement of impending eruption. A few minutes later, a brief eruption began at 10:36 p.m. Weak tremor through the next hour foreshadowed the main phase of eruption, which began at 12:03 a.m. on 17 September. Intermittent bright incandescence was seen clearly from Anchorage, as a roman-candle-like flame playing hundreds of meters above the vent.

The 17 September eruption lasted 3.6 h. Pyroclastic flows again swept down the south-southeast and southeast flanks of Crater Peak. Tephra fallout on a flanking glacier generated a debris flow that temporarily dammed the Chakachatna River south of the volcano. The eruption closed with a strong swarm of about 50 high-frequency shocks between 5 and 10 km in depth, possibly caused by collapse of the conduit after magma withdrawal. The eruption cloud spread, but again remained as a coherent body that could be tracked for days in its drift across North America.

5.3.5 Post-Eruption Seismic Swarms Following the September eruption, the level of seismicity remained elevated, and several seismic crises caused concern. On 9 November, AVO issued an eruption warning an hour into an intense seismic swarm that lasted 3.5 h and contained 170 detected earthquakes. These earthquakes were localized at 1-km depth beneath the vent and had magnitudes as large as 1.7, but seismicity receded without eruption. Because the duration and energy of this swarm was similar to the three eruptions, the swarm may be viewed as a "failed" eruption, in which magma was rapidly emplaced at shallow depth but did not reach the surface. Soon after the beginning of 1993, Mount Spurr returned to seismic quiet.

5.3.6 Interpretation of Spurr Activity Available data for the Mount Spurr volcanic system suggest the rapid migration of relatively fluid, volatile-rich magma through a narrow conduit rooted at the base of the crust (Fig. 12). Because much of the seismicity defines a near-vertical line from the active vent to 40-km depth, there seems little doubt that the seismicity is associated with flow of magma. Each of the three eruptions was remarkably similar in style, duration, and volume. Quiet periods between events were also of similar duration. This suggests a conduit that gradually filled to the same volume prior to each event, emptied in each event, and was not physically altered during each event. Storage in a conduit extending to the base of the seismogenic zone (Power *et al.*, 1995) would require a conduit radius of 10 m. Chemical characteristics indicate that this was new magma, not a fractionated derivative of that erupted in 1953 (Nye *et al.*, 1995).

There was a mirror symmetry to the pattern of seismicity. Seismicity occurred without eruption at the beginning and end of the episode and aseismic eruption occurred in the middle. This pattern is consistent with thermal waxing and waning of the shallow system. When the system was hottest, less seismogenic brittle fracturing or interaction between magma and hydrothermal fluids occurred. Early and late in the activity, rising magma encountered a cold-wet "lid" and produced seismicity without eruption. This lid was also responsible for masking magmatic gas emission from the system except during eruption (Doukas and Gerlach, 1995).

6. THE FUTURE

Volcanology is concerned with phenomena spanning a tremendous range of time and size scales. Significant events occur over seconds in the case of blasts or over millions of years in the case of evolution of large volcanic fields. Explanations are sought for bubble and crystal growth on the scale of microns, and for volcanic chains on the scale of thousands of kilometers. Many of the recent advances in volcanology have resulted from technological advances in seismic monitoring, geodetic measurements, and geochemical analysis. Laboratory experiments simulating volcanic processes are becoming increasingly sophisticated and realistic. Spacecraft are giving us views and even chemical analyses of volcanism on other planets. Meanwhile, drilling technology, motivated by both geothermal rewards and scientific curiosity, is reaching closer and closer to the magmatic environment itself. Because of the quality and quantity of instrumentation that is now fielded during eruptions—a consequence of the serious hazards that volcanism poses—nearly every major eruption results in a significant gain in understanding of volcanism. Although it would be foolhardy to try to say where these developments will lead, improved eruption forecasts, use of volcano-related resources, and understanding of planetary evolution will almost certainly be forthcoming.

GLOSSARY

Andesite: Intermediate-composition member of the basalt-andesite-dacite-rhyolite family, with 53–63 wt% SiO_2.

Ash: Micron to tens of micron pieces of fragmented and devolatilized magma; individual particles may be of glass, crystals, or both.

Basalt: End member of the basalt-andesite-dacite-rhyolite family; usually defined as containing 45–53 wt% SiO_2.

Bomb: Blocks ejected from a volcano that follow a ballistic trajectory.

Caldera: Zone of kilometer-scale (larger than individual vents) subsidence related to eruption of a large volume of magma.

Conduit: High-aspect-ratio dike or pipe representing the path of magma to the surface.

Dacite: Intermediate-composition member of the basalt-andesite-dacite-rhyolite family, with 63–70 wt% SiO_2.

Dense Rock Equivalent (DRE): Erupted volume corrected by removal of porosity; approximately equal to magmatic volume.

Dike: Sheet of frozen, crack-filling magma that crosscuts the host rock.

Dome: Structure formed by viscous lava piling up at the vent.

Fractionation: Change in composition of a system due to separation of phases of different chemical composition; may occur during melting (fractional melting), crystallization (fractional crystallization), or decompression (degassing).

Fragmentation: Process of explosive disintegration of magma to fine particles called ash.

Gabbro: Intrusive equivalent of basalt.

Geothermal Energy: Extractable energy contained in hot fluids, rocks, or magma.

Hydrostatic Pressure: Ideal pressure condition controlled by loading from overlying water.

Island Arcs: Arcuate arrangement of oceanic, subduction-zone volcanoes—e.g., Aleutian Islands.

Lahar: A slurry of volcanic debris in water that flows at high velocities from volcanic centers, typically forming thick fragmental deposits on the flanks.

Lava: Degassed magma that has emerged without fragmentation at the surface.

Liquidus: Temperature at which a rock system contains an infinitesimal amount of crystalline material (i.e., mostly liquid).

Lithostatic Pressure: Ideal pressure condition within Earth due solely to loading from overlying rock.

Mafic: The magnesium- and iron-rich, silica-poor end portion of the basalt-andesite-dacite-rhyolite family; basalts and basaltic andesites.

Magma: Rock that contains a sufficient proportion of melt to flow.

Magma Chamber: Volume occupied by a stored, spatially continuous body of magma; a magma reservoir.

Mantle: Solid but near-melting, iron- and magnesium-silicate and -oxide-bearing portion of Earth between the crust and core, representing 80 vol.% of the planet.

Peridotite: Olivine-rich rock believed to represent upper mantle material.

Phenocrysts: Crystals in volcanic rocks that are significantly larger than crystals in the rock matrix; thought to represent pre-eruption crystallization.

Plate: Semirigid piece of crust and subjacent upper mantle, generally thousands of kilometers wide and a hundred kilometers thick, that moves as a unit.

Pluton: Kilometers-scale intrusion of magma; a frozen magma chamber.

Pumice: Millimeter- to meter-sized fragments of highly porous volcanic rock.

Pyroclastic Flow: Turbulent mixture of tephra in magmatic gas that moves outward from the vent as a distinct, high-velocity mass flow.

Quench: To cool and/or decompress so rapidly that phases present at equilibrium in the previous thermodynamic state are preserved.

Rhyolite: The most silica-rich volcanic rocks; usually defined as those exceeding 70 wt% SiO_2; the molten equivalent of granite; end member of the basalt-andesite-dacite-rhyolite family.

Second Boiling: Growth of gas bubbles in magma upon cooling because of crystallization of volatile-free minerals from volatile-bearing melt.

Silicic: The silica-rich, magnesium- and iron-poor end of the basalt-andesite-dacite-rhyolite family; rhyolite, dacite, and some andesites.

Sill: Sheet of frozen magma that intruded along horizontal bedding planes or formation contacts.

Solidus: Temperature at which a rock system contains an infinitesimal amount of liquid (i.e., mostly solid).

Subduction: Downward-inclined flow of oceanic crust and underlying mantle beneath an adjacent tectonic plate; the contact between plates is an inclined plane of high seismicity.

Tectonic: Referring to large-scale movements of the Earth's crust and upper mantle.

Tephra: Fragments ejected explosively from a volcano; ash, pumice, and/or bombs.

Tuff: Tephra that has welded to form a cohesive rock.

Vent: Site where magma reaches the surface.

Vesiculation: Growth of bubbles in magma.

Volcano: Land form built by emergence of magma, as lava or tephra or both, at the surface.

Welding: Process of compaction and cohesion of ash-rich tephra, when emplaced at temperature above the softening point of glass.

Works Cited

Alidibirov, M., Dingwell, D. B. (1996), "High Temperature Fragmentation of Magma by Rapid Decompression," *Nature* **380,** 146–149.

Bailey, R. A., Dalrymple, G. B., Lanphere, M. A. (1976), "Volcanism, Structure, and Geochronology of Long Valley Caldera, Mono County, California," *J. Geophys. Res.* **81,** 725–744.

Bartlett, R. W. (1969), "Magma Convection, Temperature Distribution, and Differentiation," *Am. J. Sci.* **267,** 2067–2082.

Bluth, G. J. S., Doiron, S. D., Schnetzler, C. C., Kruegre, A. J., Walter, L. S. (1992), "Global Tracking of the SO_2 Clouds from the June 1991 Mount Pinatubo Eruptions," *Geophys. Res. Lett.* **19,** 151–154.

Bottinga, Y., Weill, D. F. (1970), "Densities of Liquid Silicate Systems Calculated from Partial Molar Volumes of Oxide Components," *Am. J. Sci.* **269,** 169–182.

Buckingham, M. J., Garceś, M. A. (1996), "A Canonical Model of Volcano Acoustics," *J. Geophys. Res.* **101,** 8129–8151.

Burnham, C. W. (1979), "Magmas and Hydrothermal Fluids," in: H. L. Barnes, (Ed.), *Geochemistry of Hydrothermal Ore Deposits,* New York: Wiley, pp. 71–136.

Carrigan, C. R., Schubert, G., Eichelberger, J. C. (1992), "Thermal and Dynamical Regimes of Single- and Two-Phase Magmatic Flow in Dikes," *J. Geophys. Res.* **97,** 17377–17392.

Casadevall, T. J. (Ed.) (1994), *Volcanic Ash and Aviation Safety: Proceedings of the First International Symposium on Volcanic Ash and Aviation Safety,* U.S. Geological Survey Bulletin No. 2047, Washington, DC: U.S. GPO.

Casadevall, T. J., Rose, W. I., Gerlach, T. M., Greenland, L. P., Ewert, J., Wunderman, R., Symonds, R. (1983), "Gas Emissions and the Eruptions of Mount St. Helens through 1982," *Science* **221,** 1383–1385.

Chouet, B. A., Page, R. A., Stephens, C. D., Lahr, J. C., Power, J. A. (1994), "Precursory Swarms of Long-Period Events at Redoubt Volcano (1989–1990) and Their use as a Forecasting Tool," *J. Volcanol. Geotherm. Res.* **62,** 95–136.

Coppens, Y. (1994), "East Side Story: The Origin of Humankind," *Sci. Am.* **270** (5), 88–95.

Decker, R., Decker, B. (1989), *Volcanoes,* New York: W. H. Freeman.

Decker, R. W. (1986), "Forecasting Volcanic Eruptions," *Annu. Rev. Earth Planet. Sci.* **14,** 267–291.

Doukas, M. P., Gerlach, T. M. (1995), "Sulfur Dioxide Scrubbing during the 1992 Eruptions of Crater Peak, Mount Spurr Volcano, Alaska," in: Keith, T. E. C. (Ed.), *The 1992 Eruption of Crater Peak Vent, Mount Spurr Volcano, Alaska,* U.S. Geological Survey Bulletin No. 2139, Washington, DC: U.S. GPO, pp. 47–58.

Eichelberger, J. C. (1995), "Silicic Volcanism: Ascent of Viscous Magmas from Crustal Reservoirs," *Annu. Rev. Earth Planet. Sci.* **24,** 41–63.

Eichelberger, J. C., Hayes, D. B. (1982), "Magmatic Model for the Mt. St. Helens Blast of May 18, 1980," *J. Geophys. Res.* **87,** 7727–7738.

Eichelberger, J. C., Keith, T. E. C., Miller, T. P., Nye, C. J. (1995), "The 1992 Eruptions of the Crater Peak Vent, Mount Spurr Volcano, Alaska: Chronology and Summary," in: Keith, T. E. C. (Ed.), *The 1992 Eruption of Crater Peak Vent, Mount Spurr Volcano, Alaska,* U.S. Geological Survey Bulletin No. 2139, Washington, DC: U.S. GPO, pp. 1–18.

Eichelberger, J. C., Sattler, A. (1994), "Conflict of Values Necessitates Public Lands Research Policy," *Eos* **75,** 505–508.

Fierstein, J., Hildreth, W. (1992), "The Plinian Eruptions of 1912 at Novarupta, Katmai National Park, Alaska," *Bull. Volcanol.* **54,** 646–684.

Fisher, R. V., Schmincke, H.-U. (1984), *Pyroclastic Rocks,* Berlin: Springer-Verlag.

Fournier, R. O. (1989), "Geochemistry and Dynamics of the Yellowstone National Park Hydrothermal System," *Annu. Rev. Earth Planet. Sci.* **17,** 13–53.

Francis, P. (1993), *Volcanoes,* Oxford: Clarendon,

Graf, H.-F., Kirchner, I., Robock, A., Schult, I. (1993), "Pinatubo Eruption Winter Climate Effects: Model versus Observation," *Climate Dynamics* **9,** 81–93.

Hess, P. C. (1989), *Origins of Igneous Rocks,* Cambridge, MA: Harvard Univ. Press.

Hildreth, W. (1981), "Gradients in Silicic Magma Chambers: Implications for Lithospheric Magmatism," *J. Geophys. Res.* **86,** 10153–10192.

Hildreth, W. (1983), "The Compositionally Zoned Eruption of 1912 in the Valley of Ten Thousand Smokes, Katmai National Park, Alaska," *J. Volcanol. Geotherm. Res.* **18,** 1–56.

Hildreth, W., Mahood, G. A. (1986), "Ring Fracture Eruption of the Bishop Tuff," *Geol. Soc. Am. Bull.* **97,** 396–403.

Hildreth, W. (1987), "New Perspectives on the Eruption of 1912 in the Valley of Ten Thousand Smokes, Katmai National Park, Alaska," *Bull. Volcanol.* **49,** 680–693.

Hill, D. P. (1996), "Earthquakes and Carbon Dioxide beneath Mammoth Mountain, California," *Seism. Res. Lett.* **67,** 8–15.

Jaupart, C., Tait, S. (1990), "Dynamics of Eruptive Phenomena," in: J. Nicholls, J. K. Russell (Eds.), "Modern Methods of Igneous Petrology: Understanding Magmatic Processes," *Rev. Mineral.* **24,** 213–238.

Jones, J. S., Rouhani, S. (1986), "How Small Was the Bottleneck?" *Nature* **319,** 449–450.

Lipman, P. W. (1984), "The Roots of Ash Flow Calderas in Western North America: Window into the Tops of Granitic Batholiths," *J. Geophys. Res.* **89,** 8733–8750.

Lu, Z., Fatland, R., Wyss, M., Li., S., Eichelberger, J., Dean, K., Freymueller, J. (1997), "Deformation of New Trident Volcano Measured by ERS-1 SAR Interferometry, Katmai National Park, Alaska," *Geophys. Res. Lett.* **24** (6), 695–698.

Marsh, B. (1979), "Island-Arc Volcanism," *Am. Sci.* **67,** 161–172.

Marsh, B. (1981), "On the Crystallinity, Probability of Occurrence, and Rheology of Lava and Magma," *Contrib. Miner. Petrol.* **78,** 85–98.

McBirney, A. R. (1984), *Igneous Petrology,* San Francisco: Freeman, Cooper and Co.

McNutt, S. R. (1996), "Seismic Monitoring and Eruption Forecasting of Volcanoes: A Review of the State-of-the-Art and Case Historie," in: R. Scarpa, R. Tilling, (Eds.), *Monitoring and Mitigation of Volcano Hazards,* Berlin: Springer-Verlag, pp. 99–146.

McNutt, S. R., Tytgat, G., Power, J. A. (1995), "Preliminary analysis of volcanic tremor associated with the 1992 of Crater Peak, Mount Spurr, Alaska," in: Keith, T. E. C. (Ed.), *The 1992 Eruption of Crater Peak Vent, Mount Spurr Volcano, Alaska,* U.S. Geological Survey Bulletin No. 2139, Washington, DC: U.S. GPO, pp. 161–177.

Miller, T. P., Kirianov, V. Y. (1995), "Volcanic Ash Hazards along the North Pacific (NOPAC) Air Routes: Kurile Islands–Kamchatka–Alaska," *Proceedings of the '95 International Workshop on Volcanoes, University of Hokkaido, Sapporo, Japan,* 88–89.

Morris, J. D., Leeman, W. P., Tera, F. (1990), "The Subducted Component in Island-Arc Lavas: Constraints from Be Isotopes and B-Be Systematics," *Nature* **344,** 31–36.

Murase, T., McBirney, A. R. (1984), "Rheological Properties of Magma," *Annu. Rev. Earth Planet. Sci.* **12,** 337–357.

Neal, C. A., McGimsey, R. G., Gardner, C. A., Harbin, M. L., Nye, C. J. (1995), "Tephra-Fall Deposits from the 1992 Eruptions of Crater Peak, Mount Spurr, Alaska: A Preliminary Report on Distribution, Stratigraphy, and Composition," in: Keith, T. E. C. (Ed.), *The 1992 Eruption of Crater Peak Vent, Mount Spurr Volcano, Alaska,* U.S. Geological Survey Bulletin No. 2139, Washington, DC: U.S. GPO, pp. 65–80.

Nye, C. J., Harbin, M. L., Miller, T. P., Swanson, S. A., Neal, C. A. (1995), "Whole-Rock Major- and Trace-Element Chemistry of 1992 Ejecta from Crater Peak, Mount Spurr Volcano, Alaska," in: Keith, T. E. C. (Ed.), *The 1992 Eruption of Crater Peak Vent, Mount Spurr Volcano, Alaska,* U.S. Geological Survey Bulletin No. 2139, Washington, DC: U.S. GPO, pp. 119–128.

Nye, C. J., Turner, D. L. (1990), "Petrology, Geochemistry, and Age of the Spurr Volcanic Complex, Eastern Aleutian Arc," *Bull. Volcanol.* **52,** 205–226.

Pallister, J. S., Hoblitt, R. P., Agnes, A. G. (1992), "A Basalt Trigger for the 1991 Eruptions of Pinatubo Volcano?" *Nature* **356,** 426–428.

Paskievitch, J. P., Murray, T. L., Hoblitt, R. P., Neal, C. A. (1995), "Lightning Associated with the August 18, 1992 Eruption of Crater Peak Vent, Mount Spurr Volcano, Alaska," in: Keith, T. E. C. (Ed.), *The 1992 Eruption of Crater Peak Vent, Mount Spurr Volcano, Alaska,* U.S. Geological Survey Bulletin No. 2139, Washington, DC: U.S. GPO, pp. 179–182.

Power, J. A., Lahr, J. C., Page, R. A., Chouet, B. A., Stephens, C. D., *et al.* (1994), "Seismic Evolution of the 1989–90 Eruption Sequence of Redoubt Volcano, Alaska," *J. Volcanol. Geotherm. Res.* **62,** 69–94.

Power, J. A., Jolly, A. D., Page, R. A., McNutt, S. R. (1995), "Seismicity and Forecasting of the 1992 Eruptions of Crater Peak Vent, Mount Spurr Volcano, Alaska," in: Keith, T. E. C. (Ed.), *The 1992 Eruption of Crater Peak Vent, Mount Spurr Volcano, Alaska,* U.S. Geological Survey Bulletin No. 2139, Washington, DC: U.S. GPO, pp. 149–160.

Rampino, M. R., Self, S. (1993), "Climate-Volcanism Feedback and the Toba Eruption of ~74,000 Years Ago," *Quat. Res.* **40,** 269–280.

Rampino, M. R., Self, S., Stothers, R. B. (1988), "Volcanic Winters," *Annu. Rev. Earth Planet. Sci.* **16,** 73–99.

Riehle, J. R. (1985), "A Reconnaissance of the Major Holocene Tephra Deposits in the upper Cook Inlet Region, Alaska," *J. Volcanol. Geotherm. Res.* **26,** 37–74.

Rundle, J. B., Hill, D. P. (1988), "The Geophysics of a Restless Caldera—Long Valley, California," *Annu. Rev. Earth Planet. Sci.* **16,** 251–271.

Rutherford, M. J., Hill, P. M. (1993), "Magma Ascent Rates from Amphibole Breakdown: An Experimental Study Applied to the 1980–1986 Mount St. Helens Eruptions," *J. Geophys. Res.* **98,** 19667–19685.

Shaw, H. R. (1972), "Viscosities of Magmatic Silicate Liquid: An Empirical Method of Prediction," *Am. J. Sci.* **272,** 870–893.

Shaw, H. R. (1974), "Diffusion of H_2O in Granitic Liquids; Part I, Experimental Data"; "Part II, Mass Transfer in Magma Chambers," *Carnegie Inst. Wash. Publ. No. 634,* 139–170.

Simkin, T. (1992), "The Rate and Frequency of Volcanic Eruptions in Space and Time," *Annu. Rev. Earth Planet. Sci.* **21,** 427–452.

Sleep, N. H. (1992), "Hotspot Volcanism and Mantle Plumes," *Annu. Rev. Earth Planet. Sci.* **20,** 19–43.

Smith, R. L. (1979), "Ash-Flow Magmatism," in: C. E. Chapin, W. E. Elston (Eds.), *Ash-Flow Tuffs,* Geol. Soc. Am. Spec. Paper No. 180, Boulder, CO: Geological Society of America, pp. 5–27.

Smith, R. L., Bailey, R. A. (1968), "Resurgent Cauldrons," in: R. R. Coats *et al.* (Eds.), Studies in Volcanology, *Geol. Soc. Am. Mem.* **116,** 613–662.

Smith, R. L., Shaw, H. R. (1975), "Igneous-Related Geothermal Systems, in Assessment of Geothermal Resources of the United States," *U.S. Geol. Surv. Circ.* No. 726, 58–83.

Sparks, R. S. J. (1978), "The Dynamics of Bubble Formation and Growth in Magmas: A Review and Analysis," *J. Volcanol. Geotherm. Res.* **3,** 1–37.

Stasiuk, M. V., Jaupart, C., Sparks, R. S. J. (1993), "On the Variations of Flow Rate in Non-explosive Lava Eruptions," *Earth Planet. Sci. Lett.* **114,** 505–516.

Stringer, C., McKie, R. (1997), *African Exodus,* London: Random House.

Symonds, R. B., Rose, W. I., Bluth, G. J. S., Gerlach, T. M. (1994), "Volcanic Gas Studies: Methods, Results, Applications," in: M. R. Carroll, J. R. Holloway (Eds.), *Volatiles in Magmas: Reviews in Mineralogy No. 30,* Washington, DC: Mineralogy Society of America, pp. 1–66.

Tait, S., Jaupart, C., Vergniolle, S. (1989), "Pressure, Gas Content and Eruption Periodicity of a Shallow Crystallizing Magma Chamber," *Earth Planet. Sci. Lett.* **92,** 107–123.

Ward, P. L. (1991), "Seismic Evidence for Magma in the Vicinity of Mount Katmai, Alaska," *Geophys. Res. Lett.* **18,** 1537–1541.

Wen, S., Rose, W. I. (1994), "Retrieval of Sizes and Total Masses of Particles in Volcanic Clouds Using AVHRR Bands 4 and 5," *J. Geophys. Res.* **99-D3,** 5421–5431.

Westrich, H. R., Stockman, H. W., Eichelberger, J. C. (1988), "Degassing of Rhyolitic Magma during Ascent and Emplacement," *J. Geophys. Res.* **93,** 6503–6511.

White, R. A. (1996), "Precursory Deep Long-Period Earthquakes at Mount Pinatubo, Philippines: Spatio-Temporal Link to a Basalt Trigger," in: C. G. Newhall, R. Punongbayan (Eds.), *Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines,* Pinatubo Vol., Seattle: University of Washington Press, pp. 307–326.

Wilcox, R. (1959), "Some Effects of Recent Volcanic Ash Falls with Especial Reference to Alaska," *U.S. Geol. Surv. Bull.* **1028-N,** 409–476.

Wilson, L., Sparks, R. S. J., Walker, G. P. L. (1980), "Explosive Volcanic Eruptions, IV. The Control of Magma Properties and Conduit Geometry on Eruption Column Behavior," *Geophys. J. R. Astron. Soc.* **63,** 117–148.

Wolf, K., Eichelberger, J. (1996), "Syneruptive Mixing, Degassing, and Crystallization at Redoubt Volcano, Eruption of December 1989 to May 1990," *J. Volcanol. Geotherm. Res.* **75,** 19–37.

Yanagi, T., Okada, H., Ohta, K. (1992), *Unzen Volcano: The 1990–1992 Eruption,* Fukuoka, Japan: Nishinippon and Kyushu University Press, p. 137.

Zhao, D., Christensen, D., Pulpan, H. (1995), "Tomographic Imaging of the Alaska Subduction Zone," *J. Geophys. Res.* **100,** 6487–6504.

Further Reading

Decker, R., Decker, B. (1989), *Volcanoes,* New York: W. H. Freeman.

Fisher, R. V., Schminke, H.-U. (1984), *Pyroclastic Rocks,* Berlin: Springer-Verlag.

Francis, P. (1993), *Volcanoes: A Planetary Perspective,* Oxford: Clarendon.

Hess, P. (1989), *Origin of Igneous Rocks,* Cambridge, MA: Harvard Univ. Press.

Wood, C. A., Kienle, J. (1990), *Volcanoes of North America,* Cambridge, U.K.: Cambridge Univ. Press.

Wright, T. L., Pierson, T. C. (1992), "Living with Volcanoes," U.S. Geological Survey Circular No. 1073, 57 pp.

WASTE MANAGEMENT AND RECYCLING

JAMES N. JENSEN, MICHAEL E. RYAN, JOHN E. VAN BENSCHOTEN, AND A. SCOTT WEBER, *State University of New York at Buffalo, Buffalo, New York, U.S.A.*

Introduction 356

1. **Classifications and Origins of Waste** 356

1.1 Domestic/Household Wastes ... 357

1.1.1 Solid Wastes 357

1.1.2 Aqueous Wastes 357

1.1.3 Household Hazardous Waste 357

1.2 Industrial/Manufacturing Wastes 358

1.2.1 Solid Wastes 358

1.2.2 Aqueous Wastes 358

1.3 Other Special Wastes 358

1.3.1 Hazardous Wastes 358

1.3.2 Construction and Demolition Debris 359

1.3.3 Radioactive Wastes 359

1.3.4 Medical Wastes 359

1.3.5 Waste Tires 359

1.4 Intermedia Transport Considerations 360

2. **U.S. Public Policy Concerning Waste Management** 360

2.1 Federal Legislation 360

2.2 Regulatory Issues 361

3. **Waste Treatment** 361

3.1 Treatment and Disposal to a Receiving Water Body 361

3.1.1 Treatment Technologies 361

3.1.1.1 Physical Processes 361

3.1.1.2 Chemical Processes 362

3.1.1.3 Biological Processes 363

3.1.2 Management of Residuals 364

3.2 Treatment and Disposal to the Atmosphere 365

3.2.1 Thermal Oxidation (Incineration) and Pyrolysis ... 365

3.2.2 Air-Pollution Control Technologies 365

3.2.3 Management of Residuals 365

3.3 Treatment and Disposal on Land 365

3.3.1 Septic Systems 366

3.3.2 Composting 366

3.3.3 Subsurface Disposal 366

3.3.3.1 Secure Landfills/Burial 366

3.3.3.2 Deep-Well Disposal 366

4. **Other Waste-Management Strategies** 366

4.1 Introduction 366

4.2 Source Reduction 366

4.3 Recycling/Resource Recovery 367

4.3.1 Introduction 367

4.3.2 Collection and Sorting of Recyclable Materials 367

4.3.3 Paper and Paperboard Recycling 367

4.3.4 Glass Recycling 367

4.3.5 Aluminum Recycling 367

4.3.6 Recycling of Other Scrap Metal 368

4.3.7 Recycling of Other Materials 368

4.3.8 Energy Recovery from Waste Oils 368

4.4 Reuse 368

4.5 Substitution 368

4.6 Integrated Waste Management 368

4.6.1 Background 368

4.6.2 Management of Plastics 368

4.6.2.1 Disposal Options 368

4.6.2.2 Plastics Biodegradation 369

4.6.2.3 Plastics Recycling 369

4.6.2.4 Plastics Segregation and Sorting 369

4.6.2.5 Chemical Recycling 370

4.6.2.6 Practical Aspects of Plastics Reclamation 371

5. **Environmental Remediation and Restoration** 371

5.1 Emergency Spills and Releases 371

5.2 Inactive Municipal/Hazardous-Waste Disposal Sites 371

5.2.1 Risk Assessment 371

5.2.2 Remediation Strategies 371

5.2.2.1 Containment 371

digital Encyclopedia of Applied Physics

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

5.2.2.2 *In Situ* Treatment Technologies 372
5.2.2.3 *Ex Situ* Treatment Technologies 373
6. Future Trends in Waste Management and Recycling 373
6.1 Multi-Life-Cycle Products 373
6.2 Treatment and Recycling at Point of Generation 373
7. Summary 373
Glossary 374
Works Cited 374
Further Reading 374

INTRODUCTION

Waste refers to material deemed undesirable and thus discarded. Just as a rose in a corn field may be called a weed, many waste materials have positive characteristics. For example, sewage contains organic matter that may deplete dissolved oxygen levels when consumed by microorganisms in natural waters. However, sewage also contains nutrients of great value if used properly. In managing wastes, one attempts to maximize the use of the waste constituents, while minimizing adverse impacts on the environment and human health.

Waste materials encompass a tremendously large number of physical and chemical characteristics. Everything from photochemical smog to mine tailings to hospital refuse can be considered waste. The wide range of characteristics requires many management tools. To minimize environmental and human-health impacts, waste managers use an array of control technologies, usually coupled with economic and social incentives. The study and practice of waste management are made more interesting and challenging by the diversity of waste and control approaches.

The field of waste management is in a state of flux. Traditional views on waste characterization and waste-management strategies have been superseded by more holistic approaches. Regarding waste characterization, wastes traditionally have been classified by their discharge locations and labeled air, water, or solid waste pollution. Control technologies were developed for each pollutant type. Currently, intermedia transport considerations are being incorporated into waste-management policy. For example, air pollution at one location can result in water pollution from acid precipitation at another location. Regarding waste-management strategies, emphasis has shifted from pollution treatment to pollution prevention. Rather than constructing treatment works at the end of the pipe, the current approach is to avoid creating the waste. This concept of waste minimization has become a central tenet of modern waste management.

The field of waste management encompasses both a large number and a diverse group of disciplines. Contributors to the field must have strong backgrounds in the natural sciences, engineering, economics, and policy making. A thorough review of each aspect of waste management would require many volumes of text. The scope of this article has been designed to give the reader a broad overview of waste-management technology and practice. Toward this end, many different waste-management approaches are covered briefly here. It is assumed that the reader has a firm understanding of environmental science. Air-pollution technologies are reviewed concisely, with further information available in AIR POLLUTION: COMPONENTS, CAUSES, AND CURES. To give the reader the broadest introduction to waste management, examples and policy implications are generally limited to North America.

1. CLASSIFICATIONS AND ORIGINS OF WASTE

Wastes are characterized in three ways: by source (e.g., industrial, domestic, and institutional), by phase (e.g., solid, liquid, aqueous, and gaseous), or by other characteristics (e.g., hazardous, infectious, and biodegradable). Classification of wastes by source is important since discarded materials are managed according to laws and regulations related to their distinguishing characteristics or source of origin.

1.1 Domestic/Household Wastes

1.1.1 Solid Wastes Domestic solid waste is a part of municipal solid waste (MSW), which includes nonhazardous waste generated in households (residential solid waste), institutions (e.g., schools, jails, hospitals, nursing homes, and public buildings), and commercial and business establishments. The size of the solid-waste stream generated annually in the United States has increased at an average of approximately 2% per year for more than three decades. The United States currently (1995) generates approximately 2×10^8 tons of garbage per year that corresponds to a per capita generation of approximately 2 kg of garbage per day. On a global scale, the MSW generation rate for most European nations is about 0.6 to 1.2 kg/capita·year, while the rates in the United States, Canada, Korea, Australia, New Zealand, and Hungary are about 1.8–2.0 kg/capita·year.

The quantity and composition of the waste stream have a direct bearing on the selection of the most appropriate technologies for management and disposal. The average composition of MSW in the United States changes with season and demographics, but a typical composition is paper and paperboard, 37.6%; yard waste, 15.9%; plastics, 9.3%; metals, 8.3%; food waste, 6.7%; glass, 6.6%; wood, 6.6%; and other, 9.0% (Nebel and Wright, 1996). Significant changes in the composition of the municipal solid-waste stream have occurred because of improved food-processing techniques, the increased use of kitchen food-waste grinders, increased use of plastics for food and other packaging, and the growth in residential recycling and composting programs.

Yard waste refers to leaves, grass clippings, prunings, and other natural organic matter discarded from yards, gardens, and institutional establishments (e.g., golf courses, parks, and cemeteries), as well as material from commercial landscaping and grounds-maintenance activities. Reprocessing of yard waste into compost or mulch is becoming increasingly adopted for reducing the size or volume of the municipal solid-waste stream.

1.1.2 Aqueous Wastes Waste water is generated from a variety of sources including residential sanitary and storm sewers, as well as waste-water discharges from commercial, industrial, government, and other facilities. Aqueous pollutants also can be transported by means of uncontrolled runoff from nonpoint sources following rainfall and storm events. Approximately 3.4×10^{10} gallons of waste water per day are generated by households, commercial establishments, and industrial operations across the United States. This waste water is collected by sewers and transported to a sewage treatment plant commonly known as a publicly owned treatment works (POTW). After treatment in the POTW, the treated waste water typically is discharged to a river, lake, or other water body. There is increasing interest in the direct reuse of treated waste water in areas of scarce water resources.

Untreated or polluted waste water can be harmful to aquatic life and leave water unfit for human use. Potable (i.e., drinkable) water is a vital and indispensable element in an urban environment. In addition, high-quality water is essential to industrial processes and manufacturing operations, since these facilities often possess substantial needs for water as a raw material, cooling or heating medium, or cleaning agent, or for sanitary or other purposes.

1.1.3 Household Hazardous Waste Household hazardous wastes are wastes resulting from products purchased by the general public for household use that, because of their quantity, concentration, or physical, chemical, or infectious characteristics, may pose a substantial known or potential hazard to human health or the environment when improperly treated, disposed of, or otherwise managed (see Sec. 1.3.1). In essence, household hazardous waste refers to materials discarded from residences that meet the legal criteria defining a hazardous waste. Many communities have developed plans and programs to keep household hazardous waste out of the solid-waste stream and to provide options for the proper management of these wastes. The major constituents of household hazardous waste are (percentages in parentheses) oil-based paint (31.9%), solvents (15.4%), latex paint (12.1%), pesticides (9.9%), empty oil cans (8.8%), cleaners (8.3%), waste oil (5.0%), acids/bases (4.0%), petroleum products (1.7%), and other (2.9%) (Tchobanoglous *et al.*, 1993).

1.2 Industrial/Manufacturing Wastes

Different industrial and manufacturing sectors of the economy generate different types of wastes. Standard categories (called the Standard Industrial Classification Code or SIC Code) define manufacturing activity and the resulting waste streams.

1.2.1 Solid Wastes Although the management of MSW has captured public attention, the vast majority of solid waste in the United States (more than 98%) is of industrial and agricultural origin. Of the non-MSW solid waste, 75% is from mining (including oil and gas production), 13% from agriculture, 9.5% from industrial sources, and about 1% from municipal waste-water and drinking-water sludges (Miller, 1994). The characteristics of non-MSW solid waste vary significantly, with about 10–15% of industrial solid waste classified as hazardous.

1.2.2 Aqueous Wastes Aqueous industrial wastes are discharged most commonly to a POTW but also may be discharged to a receiving water body. The characteristics of industrial waste-water discharges can have a profound effect on water quality and on operation of the POTWs to which the industry discharges. Although discharges are permitted through the provisions of the Clean Water Act (see Sec. 2.1), some industrial facilities are frequently in noncompliance with the permitted discharge limits. In addition, industrial sources tend to generate a significant amount of the toxic metals and organic compounds that enter POTWs. Overall, POTWs are not designed to treat these substances effectively.

Pretreatment removes pollutants that may interfere with the treatment process, cause potential damage to the treatment facility, or pass through the treatment system directly into a receiving water body. It is imperative that appropriate pretreatment technologies be employed or developed that will meet site-specific or local water-quality standards and needs. This requires treatability studies and research and development efforts to provide guidance in the selection of the most cost-effective pretreatment alternative for the particular chemical constituents to be managed.

Although industrial waste waters have been treated for more than 50 years, a more desirable approach would be to eliminate or minimize the generation or presence of the contaminant rather than to adopt an end-of-pipe treatment approach. Pollution prevention initiatives would lessen the burden and impact on urban water supplies and treatment facilities. Pollution prevention is best accomplished by means of source reduction activities that effectively reduce or eliminate the generation of a waste or undesirable pollutant at the source (usually within the industrial process). Some typical measures that may be taken to accomplish this objective include modifications to the process (such as changes in design or operating procedures), feedstock substitutions or improvements in feedstock purity, product modification or reformulation, sound housekeeping and management practices, increases in process or machinery efficiency, and in-process recycling, among others. Source reduction refers to any action that results in a reduction in the amount of waste generated by a process that would otherwise require treatment, storage, or disposal.

1.3 Other Special Wastes

1.3.1 Hazardous Wastes The term "hazardous waste" has a very specific meaning as defined in the Resource Conservation and Recovery Act (RCRA) of 1976 (see Sec. 2.1). Under RCRA, a hazardous waste is defined as any waste or combination of wastes that, because of its quantity, concentration, or physical, chemical, or infectious characteristics, may

1. cause (or significantly contribute to) an increase in mortality or an increase in serious irreversible or incapacitating reversible illness, or
2. pose a substantial present (or potential) hazard to human health or the environment when improperly treated, stored, transported, or disposed of, or otherwise managed.

Four characteristics are used to determine if a waste is hazardous: ignitability, corrosivity, reactivity, and toxicity. Listed categories of hazardous wastes include nonspecific source wastes (generic wastes that might commonly be produced by manufacturing and industrial processes), specific source wastes

(wastes of specific industrial origin), and commercial chemical products (specific chemicals, excluding wastes from households, municipal resource recovery operations, agriculture, and mining overburden that is returned to the mine site).

1.3.2 Construction and Demolition Debris Construction and demolition (C&D) waste refers to solid waste, such as building materials and rubble, resulting from construction, remodeling, repair, and demolition operations on pavements, houses, commercial buildings, and other structures. Since the 1980s, C&D waste in the United States has generally been disposed of in separate landfills. C&D waste is primarily concrete (53.8%), wood (22.0%), and brick (21.1%).

1.3.3 Radioactive Wastes Radioactive wastes are generally divided into three distinct categories: high level, transuranic, and low level. High-level radioactive waste consists of spent nuclear fuel from military or commercial nuclear reactors or the wastes resulting from the reprocessing of the spent fuel. A large volume of high-level military waste is currently stored in underground tanks at three major government sites: Hanford, WA, Idaho Falls, ID, and Savannah River, GA. Transuranic wastes consist of isotopes of chemical elements above uranium in the periodic table with half-lives greater than 20 years and having a total activity greater than 100 nanocuries per gram of waste material. These wastes are typically by-products of fuel assembly, weapons fabrication, and reprocessing operations. Transuranic wastes generally have low radioactivity but contain several long-lived isotopes, thereby necessitating separate management of these wastes. Low-level wastes are defined as all radioactive wastes other than high-level and transuranic wastes. Low-level wastes originate from certain reactor operations (both utility- and Department of Energy–operated reactors), industry, and institutions (e.g., hospitals, universities, and research organizations). Mixed wastes refer to low-level wastes that contain both hazardous chemicals and radioactive substances.

1.3.4 Medical Wastes Medical wastes include all infectious waste, hazardous waste (including low-level radioactive waste), and other wastes generated from a variety of health-care institutions including hospitals, clinics, medical laboratories, and medical offices. The composition of this waste is extremely heterogeneous and includes general refuse (such as office paper, food waste, and noninfectious patient waste), infectious waste (such as pathological wastes, human blood and blood products, contaminated and anatomical wastes, etc.), hazardous wastes (such as waste pharmaceuticals, cytotoxic agents used in chemotherapy, mercury, etc.), and small quantities of radioactive wastes.

The major sources of medical waste are hospitals and other large health-care institutions. The quantity of this category of waste generated in the United States is difficult to estimate. In 1987, the U.S. Environmental Protection Agency (EPA) estimated the total generation of hospital wastes at 5900 tons/day, corresponding to a generation rate of 6 kg per bed per day. It is estimated that approximately 10–15% of all hospital wastes are infectious.

1.3.5 Waste Tires Waste tires constitute a ubiquitous component of the waste stream and are becoming an increasing environmental concern. An estimated (2 to 2.5) × 10^8 automobile and truck tires are disposed of annually in the United States. This estimate is consistent with the annual tire production statistics for domestic passenger vehicle and truck tires. Increasingly, landfills will not accept scrap tires because of either voluntary restrictions or legislative mandates. In addition to the annual generation of scrap tires, it is estimated that there may be as many as 2×10^9 waste tires stockpiled with more than 1000 known tire piles or dumps throughout the country. Tire stockpiles pose a variety of aesthetic and environmental problems. Exposed tires collect shallow pools of rainwater, creating warm moist breeding grounds for mosquitoes and rodents. Tire stockpiles also pose a significant fire hazard because of their energy content (approximately 35 000 kJ/kg). The shape of the tire enables it to carry a self-contained supply of oxygen, making it extremely difficult to extinguish a burning pile of tires. Burning tires generate considerable pollution in the form of pungent black smoke and uncontrolled oil runoff.

1.4 Intermedia Transport Considerations

Historically, the legislative framework has focused on regulatory enforcement targeted at a specific medium (e.g., air, water, or soil). This has at times encouraged waste disposal to a preferred medium. Current attention is being given to a broad perspective for waste disposal, instead of the historical fragmented approach to waste management and regulation. The problem of waste disposal encompasses a multifaceted array of challenging technical, social, economic, and political issues.

2. U.S. PUBLIC POLICY CONCERNING WASTE MANAGEMENT

2.1 Federal Legislation

Major federal environmental legislation related to waste management is summarized as follows:

Solid Waste Disposal Act (1965): First federal effort to regulate municipal solid waste.

Resource Recovery Act (1970): Greatly expanded federal authority over the management of municipal solid waste and encouraged waste reduction and resource recovery.

National Environmental Policy Act (1970): Required environmental impact statements for all federally permitted projects.

Clean Air Act (1970): Required EPA to establish national ambient air-quality standards for total suspended particulates, sulfur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), and hydrocarbons (with airborne lead, Pb, added in 1987); new source performance standards; and national emissions standards for hazardous air pollutants (asbestos, benzene, beryllium, mercury, and vinyl chloride).

Marine Protection, Research, and Sanctuaries Act (1972): Designed to prevent or limit ocean dumping of any wastes that might adversely affect human health or welfare or the marine environment by requiring a permit for the transportation of all wastes that are to be dumped at sea.

Federal Water Pollution Control Act (also called the Clean Water Act, 1972, amended in 1977 and 1981): Intended to restore and maintain the chemical, physical, and biological integrity of the nation's waters through the establishment of discharge permits (under the National Pollutant Discharge Elimination System or NPDES), water-quality standards, and use of the best available technology (BAT) to achieve the permitted limits.

Safe Drinking Water Act (1974): Set minimum national drinking-water standards (primary standards to protect human health and secondary standards to control aesthetics) and established mechanisms to protect ground water from potential pollution sources.

Hazardous Materials Transportation Act (1975): Provided a separate management structure for hazardous materials that may pose a threat to health, safety, property, or the environment when transported by air, water, rail, or highway.

Toxic Substances Control Act (TSCA, 1976): Regulated the introduction and use of new hazardous chemical materials, including a ban on the manufacture, processing, and distribution of products containing polychlorinated biphenyls (PCBs).

Resource Conservation and Recovery Act (RCRA, 1976): Enacted to protect the environment from contamination by solid-waste disposal through the listing of hazardous wastes and establishment of a manifest system for tracking the disposition of hazardous wastes from generators to treatment, storage, or disposal (TSD) facilities.

Hazardous and Solid Waste Amendments of 1984: Expanded RCRA by establishing new technological standards for land-disposal facilities, promulgating new requirements for underground storage tanks (USTs), placing restrictions on the future land disposal of some 400 chemicals and untreated hazardous wastes, and emphasizing "waste minimization" as the preferential strategy for enhancing environmental quality.

Comprehensive Environmental Response, Compensation and Liabilities Act (CERCLA, also called Superfund, 1980): Provided the framework for the liability, compensation, cleanup, and emergency response actions associated with contaminated sites resulting from accidental or nonpermitted releases or disposal (prioritized on the National Priority List, or NPL), and established a revolving fund from a tax on petroleum and 42 listed chemicals to ensure adequate financial coverage for the ultimate closure of present and future hazardous-waste–management operations.

Superfund Amendments and Reauthorization Act (SARA, 1986): Provided additional funding for the continued cleanup of contaminated waste sites from the following sources: taxes on petroleum and chemicals, a corporate environmental tax, general appropriations, Superfund penalties, and parties responsible for site cleanup (including historic owners or operators of the site, generators of the hazardous substances that have polluted the site, or transporters that have brought the hazardous wastes to the site).

Ocean Dumping Ban Act (1988): Extended the ban on ocean dumping to include sewage sludge and industrial waste.

Clean Air Act Amendments of 1990: Designed to reduce air emissions and improve ambient air quality through the establishment of nonattainment areas, with emphasis on alternative clean fuels for automobiles, hazardous air pollutants, SO_2-emissions reduction for acid-rain mitigation, and chlorofluorocarbons-emission restrictions (for protection of the atmospheric ozone layer). A phaseout of ozone-depleting substances was established by the Montréal Protocol, with an accelerated phaseout schedule agreed upon by 187 nations in December 1992 in Copenhagen.

2.2 Regulatory Issues

A broad combination of state and federal regulations has set quantitative standards for emissions and disposal of wastes in various forms. Most of these regulations are addressed to readily identified generators or point sources of pollution. The present political environment has shifted in the direction of deregulation and reform. Cost-benefit analyses, risk-based considerations, and economic incentives for voluntary action are being more widely adopted in formulating environmental policy. Some important regulatory issues include

1. the establishment of cleanup standards,
2. greater attention to nonpoint (i.e., diffuse) sources of environmental pollution, and
3. emphasis on pollution prevention.

The regulation of special wastes (such as tires, medical waste, batteries, etc.) must be formulated carefully to avoid improper disposal of these wastes if the regulations are perceived as excessively burdensome.

3. WASTE TREATMENT

3.1 Treatment and Disposal to a Receiving Water Body

3.1.1 Treatment Technologies Treatment technologies can be categorized in two ways: by treatment type and treatment objective. Treatment types include physical, chemical, and biological processes. The two primary treatment objectives are pollutant destruction and phase transfer (where, e.g., pollutants are removed from water and concentrated in a gas stream or on an adsorptive media). Typical processes for drinking-water, municipal waste-water, and industrial waste-water treatment are listed in Table 1.

3.1.1.1 Physical Processes. Physical processes generally are phase-transfer processes. (Incineration, a physical process re-

Table 1. Typical engineered treatment processes (less common processes in italics).

Process type	Type of water treated: Drinking water	Waste water	Industrial waste water
Physical	Pretreatment, sedimentation, filtration, *adsorption, disinfection, membranes*	Sedimentation, *filtration*	Sedimentation, filtration, adsorption, ion exchange, membranes
Chemical	Disinfection, *precipitation, oxidation*	Disinfection, *oxidation*	Oxidation, *disinfection*
Biological	Not used commonly in engineered systems	Suspended growth, attached growth	Suspended growth, attached growth

sulting in pollutant destruction, is discussed in Sec. 3.2.1). The major types of physical-treatment techniques include particle removal, sorption, transfer to the gas phase, disinfection, and membrane processes.

Particle-removal processes include pretreatment, sedimentation, and filtration. Pretreatment typically is a two-step process. In the first step, called coagulation, water containing particles is mixed rapidly with a coagulant (commonly alum or ferric chloride) to neutralize the negative charge on the particles. In the second step, called flocculation, coagulated water is mixed slowly to enhance particle growth. Larger particles will settle faster according to Stokes's law. Particle settling usually occurs by gravity sedimentation (although enhanced particle removal by centrifugation is employed in specialized applications). Sedimentation basins (also called settling tanks or clarifiers) have large surface areas and very low horizontal water velocities. In an analogous process called dissolved air flotation (DAF), small air bubbles are used to float pretreated particles to the surface of a tank where they are skimmed off. Filtration is often a polishing step. The most common filter type is a gravity-fed filter using sand or anthracite/sand layers. Pressure filters employing diatomaceous earth or wound fiber media can be used with small flows. During operation, the filter will clog, and the rate of filtration will decrease or particles will break through. When performance is unacceptable, filters are backwashed by passing clean water through the filter in the upflow (or reverse flow) mode.

Adsorption processes are used to remove adsorbable compounds from water (or air). The most common adsorbent is activated carbon, formed by heating natural materials (e.g., wood or coal products) to temperatures in excess of 900 °C. In a typical installation (called a fixed-bed absorber), granular activated carbon (GAC) is packed into a tower, and water is treated continuously. The bed life ends when the pollutant concentrations in the effluent exceed the regulatory limit. At this point, called breakthrough, the carbon must be removed and regenerated by heating it again to high temperatures. Another form of adsorption is ion exchange. In this process, ions are removed from water by passing the water through a fixed-bed adsorber packed with ion-exchange resin. Ions are exchanged for innocuous species (typically, cations are exchanged for H^+ or Na^+ and anions for OH^- or Cl^-). Ion exchange is common for the production of high-purity water and desalination.

Two types of gas-transfer processes are used to treat water and industrial wastes. Air-stripping reactors typically are designed with countercurrent flow of clean air and contaminated water (often sprayed from the top of the reactor). The interfacial area for mass transfer is increased by filling the reactor with small plastic shapes called packing. Distillation has limited use in environmental applications because of its high energy costs. Distillation is used primarily for desalination and purification of used solvents.

Disinfection refers to the inactivation (or death) of pathogens (i.e., harmful microorganisms) to levels consistent with maintaining public health. Some disinfection processes are purely physical and include ultraviolet (UV) irradiation, heat treatment, and gamma radiation. UV treatment is being used in North America with increasing frequency as an alternative to chlorine for the disinfection of municipal waste waters.

Membrane processes are gaining in popularity for the treatment of drinking water and some industrial waste streams. In a typical application, water diffuses across the membrane to produce a clean water stream (called the permeate), while pollutants are concentrated in the retentate. Most membrane processes are pressure driven, leading to high energy costs in some cases. Common membrane processes (in order of decreasing molecular-weight cutoff, MWCO) are microfiltration, ultrafiltration (UF), nanofiltration (NF), and reverse osmosis (RO).

3.1.1.2 Chemical Processes. Chemical processes are used for neutralization, precipitation, oxidation, and disinfection. Neutralization involves the addition of acids or bases to bring the *p*H of the water or waste to near neutral (*p*H 7). Chemical precipitation is used to remove metals from water. In industrial- and hazardous-waste treatment, metals commonly are precipitated as their hydroxide or sulfide solids by the addition of lime (CaO) and sulfide salts. In drinking-water treatment, high levels of calcium and magnesium (called the hardness cations) produce hard water, which prevents the lathering of soap and encourages scale [i.e., $CaCO_3(s)$] formation. Hardness is removed through the addition of lime and soda ash (Na_2CO_3), with calcium and magnesium precipitated as $CaCO_3(s)$ and $Mg(OH)_2(s)$, respectively. This process is called precipitative (or lime-soda ash) softening.

In chemical oxidation, strong chemical oxidants are added to oxidize many types of pollutants. Common chemical oxidants are air (or oxygen), chlorine, ozone, potassium permanganate, and hydrogen peroxide. In some cases, oxidants are used to oxidize inorganic pollutants. Examples include the oxidation of sulfide (for odor control in municipal waste-water treatment and some ground waters) and iron and manganese (which can cause discoloration of drinking water, pipes, and laundry). Chemical oxidants also are used to oxidize organics, including cyanide, dyes, and phenols in industrial waste streams and taste- and odor-causing compounds and naturally occurring organics in drinking-water treatment. The goal of chemical oxidation is to convert pollutants to less harmful forms, but complete oxidation to carbon dioxide is rare. Thus, the production of intermediates must be monitored closely. Chemical oxidation methods are being developed in which radical species (primarily the hydroxyl radical) are formed continuously. These methods are called advanced oxidation processes (or AOPs) and include UV/ozone, ozone/hydrogen peroxide (Peroxone process), UV/hydrogen peroxide, and other oxidant combinations.

Chemical oxidants are used for disinfection as well. The primary oxidants in drinking-water and waste-water treatment are chlorine and ozone. Chlorine reacts with ammonia in many municipal waste waters to form chloramines (primarily monochloramine, NH_2Cl), which are the main disinfectant. Chlorine reacts with naturally occurring organics during the disinfection of drinking water to produce many by-products, including trihalomethanes (THMs, primarily chloroform) and haloacetic acids. These by-products currently are regulated in drinking water. In waste-water treatment, residual chlorine discharged to receiving streams can result in fish kills. Thus, chlorination is being minimized in both drinking-water and waste-water treatment. Alternative disinfectants are being explored, including UV for waste water and ozone for drinking water.

3.1.1.3 Biological Processes. Microorganisms have been used to treat waste streams since the inception of engineered treatment systems. The most common microorganisms employed in engineered systems are bacteria. Bacteria can be classified by the energy and carbon sources they use. Heterotrophic bacteria use organic carbon as both the energy and carbon sources and are the most used for the removal of organic compounds from waste water.

Biological processes are characterized by the physical system and the availability of electron acceptors. The two main types of physical configurations are suspended-growth systems (where bacteria are free-floating in a reactor) and attached-growth systems (where organisms grow on a solid support). Biological processes are used primarily to oxidize pollutants, and thus an electron acceptor is required. The main categories of electron acceptor systems are

1. aerobic processes (where the electron acceptor is oxygen),
2. anoxic processes (where oxygen does not exist and the electron acceptors are other oxidized species such as nitrate and sulfate), and
3. anaerobic processes.

Biological processes can be used for the removal of specific pollutants or for the removal of general organic material in waste waters. Organic matter is of concern, because aerobic heterotrophs consume oxygen when the organic carbon is used as an energy source. Thus, where waste water is discharged without treatment, dissolved oxygen levels may decrease and fish kills may occur. Organic matter that results in the reduction of dissolved oxygen when consumed by microorganisms is called biochemical oxygen demand or BOD.

Nitrogen and phosphorus commonly limit the growth of algae. Thus, when nitrogen or phosphorus is supplied to a receiving water body from a waste-water discharge, algae blooms may occur. To prevent algal blooms, biological nutrient removal (BNR) processes are used to control the release of reduced nitrogen (ammonia), oxidized nitrogen (nitrite and nitrate), and phosphorus. Common biological processes are shown in Table 2.

Table 2. Common engineered biological processes and their uses.

Electron acceptor type	Physical configuration			
	Suspended growth		Attached growth	
	Process	Use	Processes	Use
Aerobic	Activated sludge	BOD removal	Biotowers	BOD removal
	Aerated lagoons	BOD removal	Rotating biological contactors	BOD removal
	Nitrification	Nitrification		
	Aerobic digestion	Stabilization		
Anoxic	Denitrification	Nitrogen removal	Denitrification	Nitrogen removal
Anaerobic	Anaerobic digestion	Stabilization	Anaerobic filter	BOD removal

The two most common biological processes for BOD removal are activated sludge and biotowers. The treatment train in the activated-sludge process consists of a primary clarifier, activated-sludge basin (also called an aeration basin), and final (or secondary) clarifier. Bacteria are allowed to grow in the activated-sludge basin to a set mass concentration. Oxygen is provided to the bacteria through air blowers or, less commonly, pure oxygen. Bacterial solids leaving the activated-sludge basin are recovered in the secondary clarifier and recycled. In this sense, the "sludge" (biological solids) is "activated" (i.e., previously exposed to the influent contaminants). Activated-sludge systems (a form of secondary treatment) are extremely common in the treatment of municipal waste water. Variations of the activated-sludge process also are used to treat industrial wastes. For example, sequencing batch reactors (batch versions of the activated-sludge process) are ideal for the biodegradation of some industrial and hazardous wastes.

The biotower (also called a trickling or roughing filter) is a vertical tank containing media (often rocks) onto which bacteria grow. Waste water passes through the packed bed and organics are consumed by the bacteria. Air is introduced, often by drafting, to provide an electron acceptor. Another attached-growth process in use is the rotating biological contractor (or RBC). In an RBC system, bacteria grow on large disks that are slowly rotated through the flowing waste water. Oxygen is provided by air as the disks rotate out of the waste water.

3.1.2 Management of Residuals

Waste streams (called residuals) are generated by nearly all of the processes discussed previously. In particular, sedimentation generates a settled sludge, filtration generates a backwash stream, and most biological processes generate a biological sludge from the clarifier(s). The quantity of sludge to be disposed of is significant. In the United States, rates of sludge production are 6.4×10^8 kg/year for municipal waste water and 1.8×10^9 kg/year for drinking-water treatment. The type of residuals management scheme depends on the nature of the solids.

Inorganic solids generally are thickened, dewatered, and disposed of. Thickening processes are used to increase the percentage of solids in the sludge to the 1–10% range. Common thickening processes are gravity, flotation, and centrifugation thickening. The purpose of dewatering is to increase the percentage of solids to the 20–40% range. Common dewatering devices are centrifuges, filter presses (where the water is squeezed out through a fabric), and vacuum filters (where the water is aspirated through fabric). The dewatered sludge may be disposed of by land spreading, lagooning, composting, and landfilling. Because of pressures on landfill space, recycling of sludges is becoming more economical. Ocean dumping of many sludges is no longer allowed.

Biological sludges require additional processing. The typical sludge-handling protocol for municipal waste-water sludges includes thickening, stabilization, dewatering, and final disposal. The goal of stabilization processes is to reduce pathogens and odors and decrease the potential for putrification. The most common stabilization process is digestion. In digestion, thickened sludge is placed in a closed tank for 10–30 days. Solids are converted to gases (methane and carbon dioxide) and volatile acids (short-chained organic acids). Digesters can be aerobic or anaerobic. The methane produced can be burned to recover heat or generate small amounts of electricity.

3.2 Treatment and Disposal to the Atmosphere

3.2.1 Thermal Oxidation (Incineration) and Pyrolysis An alternative disposal option for municipal solid waste is incineration. Incineration may be more properly regarded as a waste-processing technology for reducing the volume of waste requiring disposal. However, incinerators require a substantial capital investment and are unable to incinerate everything in the waste stream. Depending on the composition of the incinerated waste stream, approximately 10–30% by volume of the input waste will remain as ash. There are environmental concerns regarding air emissions and ash disposal. Consequently, the siting and permitting of new incinerators have received considerable public outcry and have become extremely controversial issues.

The most widely used incinerators for municipal solid waste are mass-burn facilities that are designed to incinerate the waste stream with virtually no preprocessing. Modular incinerators are typically smaller-scale facilities that are intended to incinerate unprocessed waste. Refuse-derived fuel (RDF) incinerators typically utilize a shredded-waste stream from which heavier, noncombustible items such as glass and metal have been removed. Most incinerators employ an energy-recovery system that captures the heat released during combustion and converts it to steam or electricity.

3.2.2 Air-Pollution Control Technologies This review focuses on technologies for the control of air pollutants from industrial sources. Three types of particulate control technologies are in use: force-field settlers, baghouse filters, and scrubbers. Force-field settlers use different forces to remove particles from air: gravitational settlers employ gravity, cyclones use centrifugal forces, and electrostatic precipitators (ESPs) use electrical forces. ESPs are the most efficient force-field settlers. In ESPs, particulates acquire a negative charge as they pass through a corona discharge and are removed on a positively charged electrode. Baghouse filters remove particulates by passing contaminated air through filter bags. A baghouse may contain hundreds of filter bags. Baghouse filters are among the most efficient technologies of particulate removal. Scrubbers transfer particulates to a water stream. Scrubber types (countercurrent spray, cross-flow spray, venturi, and cyclone spray) differ in the manner in which air and water are introduced to the chamber. Venturi scrubbers are the most efficient scrubber type.

Gaseous pollutants are removed by absorption, adsorption, or incineration. In absorption, the pollutant is dissolved in water. For example, SO_2 can be absorbed as sulfate (SO_4^{-2}) in a lime solution. Adsorption processes were described in Sec. 3.1.1.1. Adsorption on activated carbon is used to remove organics from air. Incineration can be used for the oxidation of certain air pollutants. For example, CO can be incinerated to form CO_2.

3.2.3 Management of Residuals Each air-pollution control technology creates a waste stream. For force-field settlers, baghouses, and adsorbers, a solid waste is generated. For scrubbers and absorbers, an aqueous waste is generated. Care must be taken to dispose of or reuse the waste streams to avoid resuspension of the air pollutants.

3.3 Treatment and Disposal on Land

In North America, early environmental legislation was focused on air- and water-pollution control that indirectly encouraged land disposal as the preferred waste disposal option. In addition, open land was plentiful and landfilling operations were simple and inexpensive. During the 1980s, concerns over the environmental impact of landfills began to emerge. As a result, more stringent regulations were enacted, land-disposal costs (called tipping fees) escalated dramatically, and the siting of new landfills became a controversial political issue. The opposition to landfills relates to the use of land that could be used for other purposes, the attraction of rodents and scavengers, odors, traffic, aesthetic and property value issues, and the potential for ground-water contamination. Communities began to experience limitations in existing landfill capacity, particularly in more densely populated areas.

3.3.1 Septic Systems Septic tanks are small physical systems for the treatment of waste water from individual homes or institutions. Septic tanks may have one or more chambers. In the two-chamber configuration, the first compartment is used for the settling, digestion, and storage of sludge, while the second compartment provides backup settling and storage capacity. In a typical system, the residence time for liquid is about 24 hours, and the sludge (called septage) is removed every two to three years. Septic tanks produce two waste streams: septage and a liquid effluent. The septage is pumped from the septic tank and discharged to a municipal waste-water treatment plant. Effluent flows to a leaching field for percolation into the ground. Septic tanks require sufficient percolation rates and are not recommended where drinking-water sources are nearby.

3.3.2 Composting Composting refers to the controlled biological decomposition of organic solid waste under aerobic conditions. Approximately 3×10^7 tons of yard waste (leaves, tree trimmings, grass) are disposed of in the United States annually. Yard waste accounts for approximately 20% of the municipal solid-waste stream. Composting transforms the organic waste materials into useful soil amendments such as humus or mulch.

Composting of municipal waste-water sludges is growing in popularity. Sludge composting is almost always performed under aerobic conditions. The compost temperature must reach 50–70 °C to kill pathogenic organisms. Co-composting of waste-water sludge and MSW can be accomplished, but a market for the final product has not yet developed in the United States.

3.3.3 Subsurface Disposal Many wastes are disposed of in landfills or through deep-well injection.

3.3.3.1 Secure Landfills/Burial. Historically, landfill disposal of municipal solid waste and industrial hazardous waste was economically attractive. The more restrictive regulations concerning the construction and operation of a secure landfill facility have resulted in increased costs being associated with this disposal option. A typical landfill consists of an excavation, frequently lined with clay or a synthetic geomembrane, and a water-collection and -treatment system. The solid waste is deposited in layers over a cell within the landfill. The waste layer is often compacted to reduce volume and covered daily with soil to minimize risk to human health and the environment. Drums or containers of industrial waste are disposed of in a similar fashion by placing them within a cell and covering them with soil. Once a cell is full, an impermeable cap is placed over the waste to prevent precipitation from percolating through the landfill.

3.3.3.2 Deep-Well Disposal. Historically, hazardous liquid wastes were commonly disposed of by pumping these materials through injection wells into well-confined geological formations (such as limestone or sandstone formations). Injection wells typically extend 2000 feet below the Earth's surface. The injected waste material remains in place unless disturbed. The primary concerns with injection-well disposal relate to the leaking potential (with subsequent drinking-water contamination) and the possibility of subsurface destabilization. To minimize these concerns, it is imperative that the geology of the site selected be compatible with the waste stream.

4. OTHER WASTE-MANAGEMENT STRATEGIES

4.1 Introduction

Recently, much activity has centered on the prevention of pollution before it is generated (rather than the treatment of pollution after it is generated). This approach, called pollution prevention or waste minimization, emphasizes the efficient use of materials and energy. (Pollution prevention with industrial aqueous wastes was discussed briefly in Sec. 1.2.2.) A hierarchy of waste-minimization strategies has been developed based on their effectiveness for reducing costs and environmental impacts. The four main approaches to pollution prevention are, in order, source reduction, recycling/recovery, reuse, and substitution. Each approach is discussed in more detail.

4.2 Source Reduction

The desired ideal with respect to waste management is to reduce or eliminate the generation of the waste material at the source. This process is called source reduction. Elimination of a waste obviates the need for the transportation, treatment, and disposal of the waste material. Source reduction is the most desirable pollution-prevention option, because it minimizes the amount (or toxicity) of material entering the waste stream.

4.3 Recycling/Resource Recovery

4.3.1 Introduction Recycling refers to the collection and reprocessing of resources into new products. The concepts of recycling and resource recovery are being championed as the preferred options for alleviating the problem of municipal solid-waste disposal. Apart from the avoided disposal costs, many communities and entrepreneurs believed that recycling technologies would be economically profitable. To date, this has often not been the case, and many recycling programs have been plagued by high implementation costs and a lack of stable and economically viable markets for recycled materials.

4.3.2 Collection and Sorting of Recyclable Materials A variety of collection strategies for recyclable materials have been implemented successfully, including drop-off centers, buyback and processing centers, commercial recycling programs (for office paper, corrugated boxes, construction and demolition debris, bulk waste, and additional materials), residential curbside collection, apartment-building collection, rural recycling programs, material recovery facilities, composting (including backyard composting, large-scale community composting in windows, and in-vessel composting and anaerobic digestion), and mixed-waste processing.

Municipal recycling facilities typically employ a variety of sortation techniques including manual, magnetic, and air classification in conjunction with crushers, grinders, and balers to produce clean, segregated recyclable materials in large quantities.

4.3.3 Paper and Paperboard Recycling Two common examples of recycling are paper (and paperboard) and glass. Business centers in large cities generate a large volume of high-grade office paper. Many enterprises also generate substantial quantities of bulky corrugated boxes. Many paper mills have been retrofitting their facilities for recycled paper feedstock and recycled paper, corrugated board, and paperboard and will most likely sustain a stable future demand.

4.3.4 Glass Recycling Only container glass used to make bottles and jars is being recycled in large quantities currently. Approximately 7×10^9 glass containers are being recovered annually in the United States. The special consideration in the recycling of glass is the need for color separation. Permanent dyes are used to make colored glass with the most common colors being clear (flint), brown (amber), and green (emerald).

4.3.5 Aluminum Recycling Of the various types of aluminum products, aluminum cans are the most feasible to recover for recycling. The packaging industry constitutes the largest market for raw aluminum, consuming 27% of total aluminum production. Aluminum cans are intended for short-term consumer usage and are discarded quickly (typically after a few days or weeks).

Many aluminum cans are returned to the point of purchase because of beverage container deposit laws in many states. In addition, aluminum cans are nearly always included in municipal and commercial recycling programs because of their relatively high market value and ease of identification. In 1989, recycling operations recovered 60.8% of all used aluminum beverage containers. The market for used aluminum beverage containers arises from the fact that the manufacture of new aluminum cans from used containers requires 95% less energy than production from virgin raw materials.

Household aluminum scrap can be collected as a source-separated material in conjunction with a curbside recycling program or included in a commingled recyclables program. Approximately 9×10^8 kg of aluminum are used annually to fabricate household products other than beverage containers. An additional 3.6×10^8 kg of aluminum are used annually to make wire and cable, including bare wire (4.5×10^7 kg), aluminum-conductor steel-reinforced and bare cable (1.4×10^8 kg), and insulated and covered wire and cable (1.8×10^8 kg). Some of this wire and cable also becomes household scrap.

Approximately 90% of scrap aluminum in municipal solid waste is in the form of wrought alloy products such as those mentioned above. However, most postconsumer aluminum scrap is used to produce casting alloys on account of their less-demanding product specifications. Thousands of aluminum scrap processors are currently operating in the United States and process a variety of grades of household scrap. Foreign end users also play an important role in the recycling market. Nearly 5 × 10^8 kg of aluminum scrap (not including used beverage containers) and remelt secondary ingot are shipped to foreign buyers annually.

4.3.6 Recycling of Other Scrap Metal Scrap metal typically refers to ferrous metal, including discarded automobiles and steel cans. Recycling of ferrous scrap dates back to antiquity. Producing steel from scrap metal is economically desirable from the perspective of energy conservation. Approximately four times as much energy is required to produce steel directly from iron ore as is required to produce steel from scrap. In addition, there are advantages in terms of reduced air pollution and reduction in landfill burden. Steel cans are extensively used as containers for food, beverages, paints, and other products. Steel cans may typically have a thin lining or coating of tin for corrosion protection.

"White goods" (i.e., large appliances such as refrigerators, freezers, washers, dryers, stoves, furnaces, and water heaters) also can be recycled. These items contain significant amounts of sheet steel and are typically replaced after approximately 15 years of use. An estimated 2.9 × 10^7 units were replaced in the United States in 1991. Because of their physical volume and their high content of recyclable material, it is highly advantageous to collect and recycle these items.

4.3.7 Recycling of Other Materials The generation and composting of yard wastes are discussed in Secs. 1.1.1 and 3.3.2, respectively. Plastics recycling is discussed in Sec. 4.6.2.3.

4.3.8 Energy Recovery from Waste Oils The heat value of waste oils can be recovered through the process of fuel blending. Waste oils typically must be treated to remove sediments, water, and bottom fractions. The treated waste oils can be mixed with coal or fuel oils to produce a waste-derived fuel.

4.4 Reuse

Reuse refers to the continued use of a resource in the same form. A glass bottle that is reprocessed into paving material is said to be recycled. However, a glass bottle that is refilled is said to be reused. Common reused materials include ferrous metals, aluminum (and other nonferrous metals), and glass bottles. Resource reuse often has the added benefit of energy savings because of the high energy requirements of mining and glass processing.

4.5 Substitution

Substitution refers to the replacement of a resource with a more renewable or less toxic substance. It is used as a form of pollution prevention in some specialized manufacturing processes.

4.6 Integrated Waste Management

4.6.1 Background Management of a particular resource often requires a combination of disposal, pollution prevention, and treatment approaches. These activities should be integrated to maximize resource utilization, while minimizing environmental harm. To illustrate the many management options, plastics disposal and recycling are discussed in more detail.

4.6.2 Management of Plastics At the present time, about 3 × 10^{10} kg of plastic resin are produced, sold, and consumed in the United States each year. Approximately 1.4 × 10^{10} kg of plastic are currently disposed of as municipal solid waste. This represents 7–8% by weight and 20% by volume of the roughly 2 × 10^8 tons of municipal solid waste generated annually in the country. The percentage of the solid waste stream consisting of plastics by weight in most European nations is virtually identical to that in the United States at 7.1–9.0%.

4.6.2.1 Disposal Options. Plastics are well suited to a variety of disposal options. For example, plastics are suitable for landfilling because they are so inert. They contribute to the stability of the landfill and the integrity of the cap and do not migrate or contaminate soil or water. In addition, plastics are attractive for incineration and energy recovery (quaternary recycling) because they possess a high calorific content since their constituents are generally hydrocarbons derived from petroleum or natural gas. They may be incinerated to recover some of their energy content to produce high-temperature steam and/or electrical power. They can also be used as a fuel supplement. The advantage of incineration is that no sorting, separation, or cleaning is required. However, because of the chemical composition of certain polymers (e.g., halogen-containing polymers such as polyvinyl chloride) and the pigments, stabilizers, and other additives used in plastics products, there are concerns regarding the possible emissions of heavy metals (lead, chromium, cadmium, zinc, mercury), acid gases (HCl, HF), and other pollutants such as SO_2, NO_x, dioxins, etc. It is therefore critical to operate the incinerator at the appropriate conditions and to employ proper emission-control systems. The ash generated (approximately 20% by volume) may contain heavy metals and require disposal in a secure landfill.

Plastics are also suitable for pyrolysis. Pyrolysis involves the liquefaction and/or gasification of the plastics waste stream at elevated temperatures (1000 °F) in the absence of oxygen to prevent the formation of oxygenated compounds. The process yields a variety of products, principally low–molecular-weight aliphatic hydrocarbons that are rich in hydrogen. The products generated depend on the specific process conditions such as temperature, residence time, and feedstock composition. Increasing temperature results in an increased gas fraction and a decrease in the solid and liquid oil fractions. The thermal cracking process associated with pyrolysis is endothermic and is carried out in one of three types of reactors, namely rotary kiln, fluidized bed, and tubular shaft. Like incineration, the advantage of pyrolysis is that no sorting, separation, cleaning, or feedstock preparation is required. The disadvantages are the substantial capital investment required and the fact that the generated hydrocarbons must compete economically with similar petroleum products manufactured from crude oil or natural gas.

4.6.2.2 Plastics Biodegradation. Consideration has been given to synthesizing polymeric structures that are inherently biodegradable or incorporating additives that would accelerate the biodegradation or photodegradation of the plastic. Unquestionably, this approach is most appropriate in certain niche markets (e.g., mulch film, compost bags, six-pack rings, baby diapers, medical disposables, etc.) but has serious limitations for long-life durable goods and may pose a barrier in terms of contamination of plastics waste streams targeted for recycling. The use of accelerant additives has been shown to be effective, but the ultimate fate of the actual polymeric matrix remains an issue. Biodegradable resins are generally more expensive than their nondegradable counterparts.

4.6.2.3 Plastics Recycling. Most plastics are inherently recyclable (thermoplastics) and the majority of plastics processors (including many thermoset processors) typically recycle regrind from unacceptable parts to be economically competitive. However, the recycling of postconsumer scrap plastics poses a more serious technological and economic challenge. The plastics materials must be collected and transported to a handling and sorting facility. This may involve a residential curbside collection program, a materials recovery facility (MRF), or some other entity. Some technologies have been developed to process a commingled plastics waste stream directly. The extrusion of plastics lumber for garden furniture, decks, parking barriers, park benches, etc., is an example of this approach. The variability in bulk density, moisture content, contaminant level, and processing characteristics associated with differences in the feedstock initially posed a technical challenge to the industry. However, this problem has been largely overcome by the adoption of appropriate process-control methodology. The technology has been successfully demonstrated and commercialized, but the product must compete economically with wood.

4.6.2.4 Plastics Segregation and Sorting. A wider range of technological and market options exist if the polymeric materials can be segregated by type. Because of the bottle deposit laws in many states and the fact that most plastic soft-drink containers consist of polyethylene terephthalate (PET), the recycling of PET beverage bottles has been highly successful. Table 3 provides current statistics on postconsumer recycling (PCR) of plastics.

The sorting of plastics by type can be achieved by various means. Plastics can be sorted manually. Many recyclable plastics containers are coded using the Society of the Plastics Industry, Inc. (SPI) labeling system (1 = PET, 2 = HDPE, 3 = PVC, 4 = LDPE, 5 = PP, 6 = PS, and 7 = other) to assist in identification of the plastic. This is particularly useful for source-separation programs involving the consumer. Trained personnel at a sorting facility are very efficient and adept at identifying the major commodity resins, but some remain difficult to distinguish (e.g., PET and PVC). Manual sorting at a centralized facility is expensive and has the potential for exposing workers to residual household and industrial chemicals from the collected containers.

4.6.2.5 Chemical Recycling. The depolymerization of a plastic back to its monomeric precursors can be achieved for certain polymers by chemical and/or thermal processes called chemical recycling. The monomers can then be repolymerized or transformed into other polymers. Condensation polymers, such as polyesters, polyamides, polyurethanes, and polyethers are produced by a reversible step-growth condensation polymerization. These materials are particularly amenable to chemical depolymerization via an alcoholysis-hydrolysis reaction mechanism. Alcoholysis includes methanolysis and glycolysis in which the polymerization reaction is reversed to yield monomers or oligomers. Goodyear has commercially demonstrated the recycling of PET by means of glycolysis. Complete depolymerization by methanolysis requires an excess of methanol and heat. For PET the resulting products are dimethylterephthalate (DMT) and ethylene glycol. The DMT can be purified by distillation and repolymerized. Hoechst Celanese, Eastman Chemical, and DuPont have commercially demonstrated the viability of this process. Complete depolymerization of PET can also be achieved by hydrolysis, which yields terephthalic acid and ethylene glycol. This process has been patented by Amoco.

The hydrolysis of polyurethanes results in the production of polyols, carbon dioxide, and diisocyanate. The reaction rate can be increased by increasing temperature and pressure. The glycolysis of polyurethanes is generally preferred because of economic considerations and technological convenience. Typically, ground polyurethane is fed into a continuously mixed reactor containing glycol at 200 °C. The reactor also contains an organometallic catalyst and is maintained under an inert nitrogen atmosphere.

The chemical depolymerization of polyamides generally takes place at high temperature in the presence of water and inorganic acids. Some polymers, such as polymethyl methacrylate, polystyrene, and

Table 3. Post-consumer plastics recycling (Powell, 1991). Abbreviations: PET = polyethylene terephthalate, HDPE = high density polyethylene, PVC = polyvinyl chloride, (L)LDPE = (linear) low-density polyethylene, PP = polypropylene, PS = polystyrene, N/A = not available.

Material	Product	Total sales (10^6 kg)	Postconsumer recycling (10^6 kg)	Postconsumer recycling (% of sales)
PET	Soft-drink Bottles	367	128.9	35.1
	Custom bottles	175	4.2	2.4
HDPE	Bottles	432	60.3	14.0
	Pigmented bottles	620	42	6.8
	Base cups	48	19.6	41.1
	Other packaging	500	3.0	0.6
PVC	Bottles	92	0.7	0.8
	Nonpackaging	N/A	2.7	N/A
LDPE and LLDPE	Bottles	28	0.05	0.2
	Other packaging	2099	21.3	1.0
PP	Custom bottles	61	0.6	1.0
	Other packaging	300	1.0	0.3
PS	All packaging	923	11.0	1.2

various acetal homopolymers, may be depolymerized to give high yields of their constituent monomers by heating in the absence of oxygen. The monomers may be isolated, purified, and synthesized into new polymers.

4.6.2.6 Practical Aspects of Plastics Reclamation. Myriads of potential applications exist for reclaimed plastics. Besides the direct reprocessing into finished products, reclaimed plastics can be employed in conjunction with virgin plastic materials. For example, multilayer extrudates and blow-molded containers can be fabricated with an interior layer consisting of recycled material.

The major barrier to more widespread recycling of plastics is essentially one of economics. With the relatively low cost of crude oil, the price of polymeric feedstocks and commodity virgin polymers is also relatively low. In a free market, recycled polymers must compete with virgin resin. The collection, transportation, separation, cleaning, and pelletizing results in a significant cost with very little or no margin compared with virgin material. Fluctuating market demand and limitations of supply have also posed difficulties. However, as the recycling infrastructure becomes established, markets will become more stabilized. In addition, technological developments will result in more efficient and economical operations.

There are currently several large-volume nuisance plastics waste streams besides rigid container packaging. These include flexible film, wire and cable coatings, medical disposables, and automotive shredder residue and represent substantial volumes of potentially recoverable materials. Further research efforts need to be directed at developing efficient and economical technologies for reprocessing these materials.

5. ENVIRONMENTAL REMEDIATION AND RESTORATION

5.1 Emergency Spills and Releases

Requirements for emergency planning were codified in the Superfund Amendments and Reauthorization Act (SARA) of 1986 (see Sec. 2.1). Specifically, emergency planning needs and procedures were included in the Emergency Planning and Community Right-to-Know Act (also called EPCRA or SARA Title III). Contingency planning was required in SARA Title II through emergency planning, emergency notification, community right-to-know reporting requirements, and toxic-chemical release reporting requirements. Certain Occupational Safety and Health (OSHA) regulations also cover emergency procedures.

5.2 Inactive Municipal/Hazardous-Waste Disposal Sites

5.2.1 Risk Assessment Risk assessment is the quantification of the probability of adverse effects from a pollutant. Risk assessment involves four steps: hazard identification, toxicity assessment, exposure assessment, and risk characterization. Hazard identification is the process of listing the potential harmful pollutants. Toxicity assessment involves the use of dose-response data to relate intake to adverse effects. In exposure assessment, the population at risk, exposure times, and exposure doses are quantified. The final step in risk assessment is the integration of the first three steps to characterize the risk fully. All data required to assess fully the risk from inactive municipal/hazardous-waste disposal sites rarely are available, and risk assessments must frequently embrace and quantify the uncertainty in the analysis.

5.2.2 Remediation Strategies

5.2.2.1 Containment. Containment refers to the restriction of movement of a contaminant in the environment by the construction of a low-permeability barrier. There are two general approaches to containment: passive control and active control. In passive control systems, ground-water migration into and/or out of the contaminated zone is reduced with vertical barriers, and surface-water infiltration is minimized with a cap. Vertical barriers typically extend to a confining layer of relatively impervious material (called an aquitard).

Vertical barriers are characterized by their placement and construction technique. Barriers can be circumferential (surrounding the contaminated zone and extraction wells), upgradient (blocking ground-water intrusion and allowing for extraction of the contaminated plume downgradient), or downgradient (preventing plume migration and allowing for extraction wells to remove contaminated ground water). In terms of construction, vertical barriers include slurry trench cutoff walls, grout curtains, and steel sheet pilings. Slurry trench cutoff walls can be made of soil-bentonite, cement-bentonite, or plastic concrete, while cement- or chemical-based grouts are used in grout curtains.

The primary active system for the containment of contaminated ground waters is pump-and-treat. In this approach, contaminated ground water is extracted from the contaminated zone. It is treated at the surface, usually by air stripping, activated carbon adsorption, chemical oxidation, or biological methods (see Sec. 3.1.1). The treated ground water is discharged or recharged to the subsurface. Extraction of contaminated ground water typically is accomplished by the use of recovery wells or drain-tile collection systems. In the drain-tile collection system (also called a french drain), contaminated ground water flows into a trench excavated around the site. The trench is filled with filter media and collection piping. The drain-tile collection approach is being used to collect ground waters in the vicinity of Love Canal in Niagara Falls, New York.

The primary advantage of pump-and-treat systems is that annual operation and maintenance costs are fairly low. However, the pump-and-treat approach has been criticized for the long times (typically years to decades) required for treatment. After an initial period of rapid decontamination, the rate of pollutant removal often slows dramatically during pump-and-treat.

An alternative to pump-and-treat is the use of permeable or reactive barriers. In this approach, barrier walls are constructed of materials that allow for treatment of the ground water as it passes through the wall. Examples include the use of activated carbon, ion-exchange resins, or zero-valent iron (which is capable of dechlorinating some chlorinated organics).

5.2.2.2* In Situ *Treatment Technologies. *In situ* treatment technologies treat contaminated soil in place without excavation. Treatment processes may be physical, chemical, or biological in nature. Physical *in situ* processes include soil flushing, soil vapor extraction, low-temperature desorption, electrokinetic treatment, and stabilization/solidification. Soil flushing is a modification of the pump-and-treat process whereby contaminant mobility is increased by the nature of the flushing solution. The flushing solution can be water, acids (for recovering metals and organic bases), bases (for phenol), complexation agents (for metals recovery), or surfactants (for organics). The flushing solution is introduced by injection wells, infiltration galleries, spray irrigation, or flooding. In soil vapor extraction (or SVE), air is passed through the soil to remove volatile organics. Air is introduced by drawing a vacuum on vapor extraction wells or perforated pipes located in the unsaturated (or vadose) zone. In a modification of SVE called air sparging, air is blown into the ground water, and volatilized contaminants pass into the unsaturated zone where they are collected by an SVE system. In low-temperature soil treatment, air is passed through soil that has been heated to 300–400 °F. Volatile organics are stripped into the air stream and subsequently treated. In electrokinetic (or electroosmotic) soil treatment, a direct current is applied to the soil to generate a voltage field. Pore water containing the contaminants moves in response to the field and can be captured and treated. Electrokinetic methods are used mainly for the removal of metals from soil.

Other physical processes that can be employed *in situ* or *ex situ* include stabilization and solidification. In stabilization (also called fixation), the mobility of contaminants is decreased significantly by the addition of binders. In solidification, the larger amounts of added materials are used to solidify the waste as well as bind contaminants. Stabilization and solidification are used for the treatment of sludges, soils, sediments, and oily wastes. Typical binding agents include lime, portland cement, pozzolanic materials (such as cement-kiln dust and fly ash), silicates, organically modified clays, and organic polymers. One form of solidification is vitrification. In vitrification, the soil is heated to 1600–2000 °C in a glassmaking furnace or *in situ* and cooled rapidly. The resulting amorphous mass can be crushed and used as a road aggregate.

Bioremediation refers to the treatment of contaminated sites with biological processes. The contaminated soil system replaces the biological reactors described in Sec. 3.1.1.

The primary concern of bioremediation systems is the proper introduction of electron acceptors (usually oxygen) or other nutrient. Oxygen and nutrients can be introduced into the ground water through injection wells or recharge trenches. Oxygen may be added in the form of air or hydrogen peroxide. Careful balance of extraction and reinjection wells is critical in maintaining adequate dissolved oxygen levels in the subsurface.

Modifications of the basic bioremediation scheme described earlier have been employed. For example, exogenous (non-native) bacteria may be added to the site. In addition, anaerobic biodegradation of some contaminants in the subsurface is possible. In some cases, addition of cometabolites can increase rates of biodegradation. Cometabolites induce enzyme systems that allow for the serendipitous degradation of the pollutant. For example, introduction of methane gas may enhance cometabolism of trichloroethylene (TCE).

5.2.2.3* Ex Situ *Treatment Technologies. *Ex situ* treatment technologies are used with excavated soils. Several of the *in situ* processes described in Sec. 5.2.2.2 (especially low-temperature desorption, electrokinetic treatment, stabilization/solidification, and bioremediation) also can be used with excavated soils. *Ex situ* bioremediation processes typically involve slurry treatment, land treatment, or composting. Other *ex situ* processes include disposal and soil washing. Excavation and disposal may be feasible if land disposal costs are low. Safety considerations during the transport of contaminated soils are critical. In soil washing, fluids (typically water, acids, chelating agents, surfactants, or solvents) are passed through the soil to solubilize the contaminants. For example, acids may be used to remove metals bound to the soil. Soil washing typically is performed on excavated soils or soil fractions.

6. FUTURE TRENDS IN WASTE MANAGEMENT AND RECYCLING

6.1 Multi-Life-Cycle Products

An important tool developed recently for environmental professionals is life-cycle analysis. In life-cycle analysis, the environmental impact of a product from "cradle to grave" is determined. This includes the impact of the procurement of the raw materials as well as the product design, manufacture, use, maintenance, repair, and final disposal. If used properly, life-cycle analysis can be a valuable tool in quantifying the overall impact of a product and can be used to determine real product costs.

As recovery, recycling, and reuse programs grow, it becomes necessary to think of products as having multiple lives. Product design may need to incorporate the future use of all material in the product. For example, car manufacturers are incorporating the final disposal and reuse of materials in the design of automobiles. This trend is likely to become more important in the future.

6.2 Treatment and Recycling at Point of Generation

Environmental technology is at a crossroads. Traditional "end-of-pipe" treatment technologies are being complemented with or supplanted by pollution-prevention concepts. However, treatment and pollution prevention are not mutually exclusive. Treatment of waste streams to allow for material recycling will have an increasingly important role in waste management.

On-site recycling or reuse is particularly attractive, because it avoids problems with transportation and storage of recycled materials. In addition, on-site recycling can eliminate concerns regarding inconsistencies in the markets for recycled materials. Especially where economies of scale are not important for treatment and recycling processes, on-site treatment and recycling may be very efficient.

7. SUMMARY

This review has provided a brief introduction to a large number of topics in waste management and recycling. Treatment processes will continue to play a major role in the management of solid wastes, municipal waste waters, hazardous wastes, and air pollutants. The complementary tools of pollution prevention (recovery, recycling, reuse, and substitution) will grow in importance as economic pressures increase. The need for efficient waste-management practices has never been more acute as we enter the twenty-first century.

GLOSSARY

Aerobic: Biochemical process or condition associated with the presence of oxygen.

Compost: Humuslike material resulting from the biological decomposition or breakdown of organic materials.

Effluent: A waste liquid from a manufacturing or treatment process that is discharged to the environment.

Fly Ash: Solids, including ash and other matter, that rise with the hot gases from combustion and are removed by emission-control devices.

Hazardous Waste: A waste material that, because of its quantity, concentration, or physical, chemical, or infectious characteristics, may pose a substantial potential hazard to human health or the environment when improperly treated, stored, transported, disposed of, or otherwise managed.

Postconsumer Material: Those products generated by a business or consumer that have served their intended end uses and that have been separated or diverted from solid waste for the purposes of collection, recycling, and disposition.

Recycling: The process of separating materials from the waste stream and using them for another purpose.

Refuse-Derived Fuel (RDF): A solid fuel obtained from municipal solid waste.

Resource Recovery: The extraction and utilization of materials and energy from the waste stream.

Reuse: The use of a product more than once in its same form and for the same purpose.

Sanitary Landfill: An engineered repository for the disposal of solid waste on land.

Secure Landfill: A landfill designed to prevent the entry of water and the escape of leachate through the use of impermeable barriers.

Yard Waste: Leaves, grass clippings, and other organic manner from yards and gardens.

Works Cited

Miller, G. T., Jr. (1994), *Living in the Environment,* 8th ed., Belmont, CA: Wadsworth Publ. Co.

Nebel, B. J., Wright, R. T. (1996), *Environmental Science,* 5th ed., Upper Saddle River, NJ: Prentice Hall.

Powell, J. (1991), *Resource Recycling* **10**(3), 72–77.

Tchobanoglous, G., Theisen, H., Vigil, S. (1993), *Integrated Solid Waste Management, Engineering Principles and Management Issues,* New York: McGraw-Hill.

Further Reading

Denison, R. A., Ruston, J. (Eds.) (1990), *Recycling & Incineration,* Washington, DC: Island Press.

Eckenfelder, W. W., Jr. (1989), *Industrial Water Pollution Control,* 2nd ed., New York: McGraw-Hill.

LaGrega, M. D., Buckingham, P. L., Evans, J. C. (1994), *Hazardous Waste Management,* New York: McGraw-Hill.

Lund, H. F. (Ed.) (1993), *The McGraw-Hill Recycling Handbook,* New York: McGraw-Hill.

Metcalf and Eddy, Inc. (1991), *Wastewater Engineering: Treatment, Disposal, Reuse,* New York: McGraw-Hill.

Murray, R. L. (1994), *Understanding Radioactive Waste,* 4th ed., Columbus, OH: Battelle Press.

WATER POLLUTION

D. J. Baumgartner, *Soil, Water and Environmental Science Department, University of Arizona, Tucson, Arizona, U.S.A.*

	Introduction	375
1.	**Water Sources and Their Susceptibility to Pollution**	376
1.1	Fresh Water	376
1.2	Marine Waters	376
2.	**Sources of Surface-Water Pollution**	379
2.1	Construction and Land Use	379
2.2	Domestic and Municipal Sewage	380
2.3	Intestinal Microorganisms	380
2.4	Organic Matter	380
2.5	Die-off of Indicator Organisms	381
2.6	Dilution of Effluents	382
2.7	Organic Matter and Dissolved Oxygen	384
2.8	Impact of Organics on Dissolved Oxygen of Receiving Waters	385
3.	**Toxic Substances**	386
4.	**Nonpoint (Diffuse) Sources of Pollution**	388
5.	**Ground-Water Pollution**	389
6.	**Regulatory Practice**	390
6.1	Risk Assessment	391
	Glossary	393
	Work Cited	393
	Further Reading	394

INTRODUCTION

Water serves many purposes in society, some of which have been recognized for eons. Other uses are the result of relatively recent industrial influences on our way of life. Some uses are compatible with all recognized uses of water in our society while others change the quality and quantity of the water, impairing other uses. Water pollution can be considered to result when contaminants are added or the physical quality, e.g., temperature, is changed to a degree sufficient to affect other uses adversely. Natural processes can in time ameliorate most water-pollution problems, but when limited time is a critical element in the intended reuse of a body of water, man-made adaptations of the natural processes can be employed to restore water quality.

Water quantity and quality are linked in determining when a given water use and subsequent reuses are compatible, simply because many water-pollution conditions are concentration dependent. That realization gave rise to an often maligned adage, "dilution is the solution to pollution." Although true to this day in the case of many contaminants, there is more than dilution alone that must be considered with regard to persistent organic compounds, for example, in preventing or resolving water-pollution problems. Through the last two centuries the demand for water has generated a hierarchy of uses, with human consumption at the top, considerable jockeying for intermediate positions, and relegation of waste disposal to the bottom of the list.

Legal doctrines on water use vary with geography and culture. In the eastern United States, riparian rights imply equal and unabridged right to water flowing along a property line, whereas in the western regions of the country, prior appropriation entitles a person first in time to claim all the water he needs, preventing others upstream from taking any at a later time. In the East, both quantity and quality issues were involved in adjudications, but in the West, quantity was (and is) the main issue. Emerging water quality concerns, even in the West, beginning in the 1950s and extending well into the 1990s have promoted new federal legislation and regulatory practices that will dominate municipal, industrial, and some agricultural uses of water into the 21st century, based on a wide range of quality issues. The major ef-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

fort in water-pollution regulation is administered by the U.S. Environmental Protection Agency (EPA), which in its 1994 inventory of water quality found that about 40% of the water bodies in the United States that were assessed by individual states were not suitable for such activities as fishing or swimming (Browner, 1994). While these uses of water are desirable amenities in more developed countries, it is clear that they are not the highest and best uses in the hierarchy of uses, and it is also clear that the effort to achieve the fishable/swimmable objective will impose significant additional requirements on other uses. The U.S. EPA employs strategies of technology-based controls and water-quality standards within an effluent-discharge permit system to achieve this objective.

In less-developed countries direct recreational uses may not be considered at all. Generally, drinking water and agriculture demand highest priority, transportation of goods (shipping) and industrial use a lower priority, and waste disposal the lowest priority. Local and regional interests can influence the priority rankings in any given society, however, and controversies may arise as water flows from one area to another within a watershed.

The problems and approaches to control of water pollution are similar in many industrialized nations, with control strategies frequently following the lead of the U.S. EPA. The major pollution problem in nonindustrialized nations is the lack of infrastructure for treatment of domestic sewage. This results in a serious impact on human health because of the lack of facilities for storage and treatment of drinking water.

1. WATER SOURCES AND THEIR SUSCEPTIBILITY TO POLLUTION

1.1 Fresh Water

The amount of fresh water in rivers and lakes, 125 000 km^3, is only 1/10 000 of the total water available on the planet (van der Leeden *et al.*, 1990). Another (10 to 20) × 10^6 km^3 are thought to exist in ground water. The total quantity, which is fairly constant decade to decade around the globe, is not nearly as important in considering water use and pollution as is its nonuniform distribution over terrestrial space scales of 10^4 to 10^5 km^2 and time scales of 10 to 1000 d. More than two-thirds of the water that falls on the earth as precipitation returns to the atmosphere as evaporation, and half of the remaining precipitation runs off in flood events and is not available for use. Most of the fresh surface water is in lakes, having a turnover time of about 100 a, while streams, with turnover times of 11 d on average, contain only 1% of the surface water. By comparison, the replenishment time of ground water is on the order of 1000 a.

The water stored in reservoirs and lakes, together with the water that flows perennially in streams, is subject to heavy stress due to use and reuse for water supplies, agriculture, industry, and recreation, frequently resulting in misuse. Most cities and industries in the United States discharge waste water to streams and rivers, rather than to lakes and reservoirs. Even though the waste waters are treated, large quantities of contaminants are carried by river systems. Within the last decade or so, most communities have installed treatment plants removing 80–90% of particulate and dissolved organic contaminants. The remaining 10–20% of these contaminants could still result in substantial pollution. For example, a community of a million people could add 9 to 18 metric tons of organics per day to the river running through their town, even with 80–90% sewage treatment efficiency. In addition, inorganic salts and metals are added, which are not as effectively removed in treatment. Similarly, ground water used for municipal and agricultural purposes is normally returned to streams. Whether downstream uses are impaired or not, i.e., whether pollution occurs, depends on the type and amount of contaminants added by incremental uses and on the physical, chemical, and biochemical processes that influence the fate of contaminants.

1.2 Marine Waters

Marine waters constitute most of the water on the planet. Even the tremendous volume of ocean water in which contaminants may be dispersed does not necessarily confer immunity from pollution for marine-water resources. Estuaries and near-shore coastal waters are most susceptible to pollution because of the load of contaminants brought to the coastal zone by rivers, local concentration of municipal and industrial activity, and the belief, sometimes correct, sometimes wrong, that less treatment is required for marine disposal of wastes.

Almost all of the water-carried wastes of a continent enter the coastal zone through an estuary. Estuaries take many physical forms, including fjords, broad deltas, and drowned river valleys, all of which have salinities modified from ocean concentrations as a result of evaporation and dilution by fresh-water inflows. Mixing of river flows and waste discharges with the ocean water in an estuary is influenced by the winds and tides, and by the density distribution in the water associated with both salinity and temperature differences. Depending on the degree of mixing and circulation of water masses within the estuary, contaminants may have a long residence time in the system, which can lead to pollution problems.

Although good circulation and short residence times tend to suppress the onset of pollution, there are still situations where pollution can result in well-mixed estuaries. In many of these cases the pollution is the result of seabed accumulation of organic or metal-laden particles. Two physical-chemical factors contribute to this phenomenon in addition to the amount and type of contaminants carried into the estuary by rivers and waste discharges. The first is that current speeds in the estuary are generally lower than in the rivers so that gravitational sedimentation is facilitated. Also, the mixing of oceanic salts with the essentially fresh-water character of rivers and waste water enhances precipitation and coagulation of dissolved and fine particulate matter, which adds to the sediment load. In some estuaries, especially in fjords where a strong density gradient restricts vertical mixing, particles that once settle into the bottom layer of the water column may be retained permanently. Sediment-quality criteria have been the subject of research within the EPA since the early 1980s. This research has focused mainly on toxicity to benthic invertebrates and on sediments as a food route for bioaccumulation of harmful organic chemicals. Science advisors to the EPA have suggested that additional research may lead to regulatory criteria within the next decade.

Offshore from the estuarine environment, the near-shore coastal zone has been the subject of pollution concern related primarily to the disposal of sediment removed by dredging of harbors, and to discharge of waste water from large municipalities and industries, e.g., the pulp and paper industry. The coastal zone extends from the low-tide line to the 200-m depth contour, coinciding with the geophysical demarcation of the continental shelf. The coastal zone can be as wide as 1400 km along some coasts and less than a kilometer along others. Dilution and off-shore transport of contaminants is generally more favorable in this physical environment than in rivers and estuaries. Consequently, there are advantages to disposal of waste-water effluents in the coastal zone over estuaries. A large outfall pipeline or tunnel, terminating in a multiple array of relatively small-diameter diffuser ports, is commonly used in the United States to transport the effluent several kilometers off shore. The practice is becoming more common for major municipalities around the world, although there are still coastal cities that discharge untreated waste water close to the beach through outfalls without multiple-port diffusers, thus failing to take advantage of the high degree of dilution that could be achieved.

Physical features of the coastal zone can interfere with effective waste disposal in the coastal zone. For example, continental headlands that protrude seaward from the beach can impede both circulation of water and exchange of near-shore water with open-ocean water. Outfalls are therefore best located farther off shore rather than inside the region of headland influence. Similarly, outfalls should not be located close to the entrance to estuaries or bays because tidal incursions can carry diluted wastes into the estuary, thereby diminishing one of the advantages offered by off-shore disposal (Baumgartner *et al.,* 1994). This problem also occurs sometimes with barge disposal of dredged materials off the mouth of estuaries, from the movement of deep off-shore coastal waters into an estuary. During parts of the tidal cycle, the deep, landward currents can be strong enough to carry the dumped particulates back into the estuary, where they again might settle in the channels that were just dredged.

Waste disposal has caused definite pollution problems in a few highly developed coastal areas in the United States, notably off of Los Angeles, New York, and Boston. In all cases the predominant manifestation has been the sediment accumulation on the seabed of particulate material discharged with partially treated waste water and digested sludge. Nutrients and persistent organic chemicals in and on the particulates have contributed to fish diseases and the disruption of indigenous biological communities in the region of the discharges. Billions of dollars have been expended to abate the pollution and prevent recurrence; however, many scientists and regulators disagree on the effectiveness of the technologies employed in these efforts. Furthermore, a report of the National Research Council (1990a) concludes that although over \$130 $\times 10^6$ is spent annually on coastal monitoring of water quality, most programs fail to provide the information necessary to assess the condition of the environment, anthropogenic effects on it, and the effectiveness of remediation and control measures. Two major difficulties with improving the state of affairs are the vast complexity of marine environmental processes and the limited knowledge possessed by the scientific community to describe and predict causal relationships.

Open-ocean waters beyond the coastal zone can become contaminated by atmospheric fallout, oil spills, and dumping of hazardous wastes and sewage sludge as practiced by some countries of the world. Barge dumping of treated municipal sludge into coastal and off-shore waters is prohibited by the United States, although discharge of treated waste-water effluent is allowed. Floatable and soluble materials tend to stay in the upper mixed layer of the ocean, the most active photosynthetic zone. The depth of this layer, which varies between 100 and 1000 m, changes with season and geographic location. Although mixing between the upper and deeper layers of the ocean is impeded by strong density gradients, particles formed in the upper mixed layer, or discharged there, may eventually settle so far that they can no longer be resuspended by surface-generated turbulence, and thus become incorporated with the natural detrital sediment load of the deep ocean waters.

Atmospheric fallout and washout of chemicals, primarily volatile organics, has resulted in the accumulation of chlorinated hydrocarbons in the fatty tissues of marine animals far from terrestrial industrial sources, highlighting a case where the large potential dilution capacity of the ocean is not necessarily an effective barrier to pollution. A large portion of petroleum hydrocarbons found in the ocean evidently derive from atmospheric transport of fugitive gases from terrestrial sources, although it is estimated that a third or more of the oil spilled at sea is evaporated and becomes part of the atmospheric burden. Major oil spills from vessels such as the *Exxon Valdez* represent only about 1% of the annual input of hydrocarbons to the world's oceans. Lead and mercury also have been found in sea-surface sampling off shore from municipal–industrial coastal developments. However, the direct connection to adverse impacts on marine resources from atmospheric inputs has not been documented.

Typical municipal and industrial waste-disposal operations have not represented a direct threat to global open-ocean water quality, contrary to the situation in many near-shore coastal waters. Indirectly, contaminants in low concentrations carried into the open ocean by currents flushing the coastal zone could conceivably contribute to diminished water quality. Unfortunately, through the decade of the 1990s this has been a matter of research interest only, and only cursory observation by regulatory authorities. To call it surveillance would overstate the magnitude of the attention dedicated to this effort.

When ocean disposal of certain materials is justified, technologies can be employed to avoid contamination of the upper mixed layer and facilitate transport to and long-term retention of the material in the deep waters of the open ocean. For example, containers have been proposed for disposal of such materials as hazardous chemicals or radioactive wastes. Although not an existing technology, proposals have been made to use pipelines to transport liquid carbon dioxide to the seabed, where it can be retained for a long time—conceivably long enough to help reduce the rate of carbon dioxide buildup in the atmosphere (Marchetti, 1977).

2. SOURCES OF SURFACE-WATER POLLUTION

Pollution in surface waters can result from many kinds and amounts of contamination. It may take just a small amount of contaminant to pollute a water body intended for a drinking-water supply. But the same water might not be considered polluted if the water were to be used, for example, for agriculture. Nor is pollution restricted to biological and chemical contaminants. Physical factors of the environment can also contribute to pollution. For example, heated water discharged from a power plant can change the temperature of an aquatic environment. It might not be a problem in a lake or a river during the winter, but it can certainly be a problem in the summertime. For example, young salmon in the Columbia River are especially vulnerable to a microbial infection when summer temperatures are raised only slightly above the normal midsummer temperatures. There are also many kinds of contaminants that can usually be accommodated by the natural environment without resulting pollution.

The major sources of surface-water contamination are construction, municipalities, agriculture, and industry. However, the water delivered to earth in the form of precipitation is not necessarily pure to begin with. Near the coast, it may contain particulate and dissolved sea salts, and further inland, it may contain organic compounds and acids scrubbed from contaminants added to the atmosphere both by natural processes and by anthropogenic activities. Gases from plant growth and decay and gases from geological activity are examples of naturally derived atmospheric contaminants that can be returned to earth via precipitation. The acid-rain problem of the New England states is a classic example of anthropogenically derived atmospheric contaminants that contribute to surface-water pollution.

2.1 Construction and Land Use

Because construction involves land clearing and grading, this activity is a major source of sediment contamination of streams. Agriculture and forestry are also operations traditionally responsible for pollution of streams with particulate matter. In addition, other land-use activities such as recreational off-road vehicle use can destabilize soil formations, thus increasing erosion. The runoff of fine-grained soil particles, especially in hilly terrain with high rainfall, can smother gravel beds that serve as a spawning habitat for fish. In addition, suspended particulates and colloidal material can harbor microorganisms, necessitating increased treatment for municipal water-supply facilities. Nor is the damage associated with construction limited to aquatic resources; erosion also decreases the topsoil resource needed for sustainable plant growth.

Erosion, both anthropogenic and natural, is a continuous process. All things considered, tens of thousands of hectares of riparian area are lost to erosion each year in the United States, and hundreds of millions of tons of topsoil are carried by rivers to the sea per year. Natural erosion of river banks in some areas masks the pollution caused by human activities. For example, such rivers as the Mississippi carry about 500 mg L^{-1} of suspended solids, so that even many tons of solids added over a section of the river by a construction event might not adversely affect the use of the water. In other areas, the amount of river-bank erosion can be exacerbated by poor land-use practices. Some streams are naturally very sediment free, so that the addition of as little as a ton of soil in a year can cause pollution of a major portion of the stream.

Beginning in the 1970s, U.S. pollution-control agencies began to restrict road construction and timber harvesting operations in forest areas to reduce soil erosion. Although these restrictions were primarily intended to reduce impacts on fish habitats, their serendipitous result was to enhance retention of topsoil. The 1980s saw a great deal of attention devoted to reducing soil runoff from individual home-site construction projects, as well as large home and shopping-center developments. Then, in the 1990s, the promotion of "no-till" planting in the agriculture industry helped to reduce wind- and water-carried loss of topsoil, and this in turn helped to reduce potential water-pollution problems in rivers and reservoirs. Measures designed to control soil loss to rivers may have reduced the concentrations of other contaminants, such as pesticides and metals, that can also pollute surface waters. However, the main sources of many of these contaminants are municipalities and industries as well as agriculture.

2.2 Domestic and Municipal Sewage

Domestic sewage includes human wastes and a variety of waste materials that arise from household products such as foods, soaps, paints, oil, and grease. These wastes contain microorganisms, metals, plastics, decomposable organic chemicals, and toxic chemicals. Municipal sewage comprises a matrix that includes everything in domestic sewage plus water-carried wastes from industries, schools, hospitals, and airports connected to the municipality's sewer system. In many cities, street washings and other wastes poured into streets contribute to the contaminant load in municipal sewage.

Pollution may result when these wastes are poorly treated or when even highly treated sewage is discharged into a watercourse that is physically or biologically incapable of accommodating the waste load. Human wastes make up a considerable portion of the contaminants in sewage. Many of the stream-water pollution problems that have occurred—and pollution-control technologies that have been developed—are related to human wastes. The components of human wastes that characteristically cause stream pollution problems worldwide are intestinal microorganisms and organic matter.

2.3 Intestinal Microorganisms

Many intestinal microorganisms can remain viable in surface waters for a long time and can cause illness, disease, or even death if they come in contact with humans. The primary method of defense for downstream water users against water-borne outbreaks of disease (such as typhoid and cholera) has traditionally been high-quality physical treatment (chemical precipitation and filtration) and disinfection of river water with chemicals such as chlorine prior to distribution to consumers. However, disease and illness can also be a problem for other uses of the stream resource, such as recreational bathing and harvesting of shellfish. For example, bacteria, viruses, and other microorganisms can cause upper respiratory and ear infections in bathers, while fairly large disease episodes have been attributed to the ingestion of improperly cooked shellfish harvested from polluted coastal waters. Also, some recreational users, misled by the clarity and coolness of streams in remote areas, become ill following ingestion of raw or poorly disinfected water. Visibly clear and ostensibly unpolluted mountain stream water can easily be polluted by wastes from one or two humans upstream, or even by the native warm-blooded animal population. For example, the gastrointestinal disorder known as giardiasis can be caused by the protozoan *Giardia lamblia* found in water and snow contaminated with animal feces. It is prudent—if not entirely necessary—to assume, without qualification, that all surface waters are contaminated with microorganisms of human health significance.

To some degree, the quality of surface water for recreation is protected by the treatment of domestic and municipal sewage at the point of discharge into the watercourse and by the natural purification processes in the water. High-quality sewage treatment processes can remove more than 99.99% of the microorganisms in sewage, although this level of treatment is not always provided. Even if it were provided, waste-water treatment alone is never sufficient to result in river water of suitable quality for direct human consumption. Major storm events can greatly increase the volume of water flowing through older sewage systems (combined sanitary and storm sewers), thus increasing the likelihood that untreated and poorly treated sewage will reach the watercourse. Such events frequently result in the closing of nearby bathing beaches because of the temporarily high concentrations of bacterial counts indicative of fecal pollution.

2.4 Organic Matter

Of the many organic chemicals in municipal sewage, two broad categories are important on the basis of the type of surface-water pollution problems they cause. One group, xenobiotics, contains the persistent synthetic chlorinated organic compounds that were produced by the petrochemical industry in ever-increasing amounts after World War II; these include insecticides and chlorinated hydrocarbons such as polychlorinated biphenyls (PCBs) and trichloroethylene (TCE). Such chemicals degrade very slowly in the environment and can cause long-lasting toxicity, either acutely or chronically as a result of bioaccumulation in tissues.

Other organics are grouped in a class labeled biodegradable organic matter, which includes natural organic compounds found in wastes, such as carbohydrates and proteins. Many industrial chemicals also fit in this category. The problem with these organics is the depression of the dissolved-oxygen content of stream waters caused by their microbial decomposition. The main feature of this pollutional effect in the stream is that the loss of dissolved oxygen is detrimental to aquatic organisms. A secondary effect is that if all the dissolved oxygen is consumed by aerobic microbial oxidation of the wastes, anaerobic decomposition becomes prominent, releasing noxious septic gases from the stream. Waste-water treatment methods, including trickling filters and activated sludge aeration, were developed in the first 50 years of the 20th century to deal with this problem.

Quantitative methods have been developed to assess the impact on surface waters and to calculate the degree of treatment needed before municipal wastes are discharged as effluents. Such methods are based on an understanding of the physical and biochemical processes controlling the decay of microbes and chemicals over time.

2.5 Die-off of Indicator Organisms

Early in the development of quantitative assessment methods, scientists realized that many different microorganisms that exist in human wastes could not be effectively cultured and counted. Consequently, they settled on the coliform group of organisms to serve as an indicator of fecal pollution because they could be cultured and counted easily. Known to exist in large numbers in the gut of all warm-blooded animals, the coliform group provides a good indication of fecal pollution; however, it was not very specific, and so other indicators such as fecal coliforms and streptococci may now also serve as indicators of the sanitary quality of water.

Tests have shown that 99.99% of the indicator bacteria can be removed by waste-water treatment. But the residual 0.01% remains a problem, raising concerns about the quality of water for recreational use, for example. The fact is that the number of bacteria in sewage is tremendous (500×10^6 to 2×10^9 per 100 mL), so that the percentage of removal is not as useful as is the actual remaining concentration in assessing quality. The microorganism concentrations allowed for various uses of water are relatively small. For drinking purposes, the concentration of fecal coliforms should, of course, be zero, but a concentration of less than 1 per 100 mL may be allowed; and for bathing, a concentration of 1000 per 100 mL is frequently accepted. If, for example, raw sewage has a population of 10^9 per 100 mL, a removal efficiency of 99.99% would still leave 100 000 per 100 mL. Simple models are used to determine if the concentration will decrease to acceptable levels through die-off and dilution before the sewage reaches a bathing beach.

Indicator-organism concentrations generally decrease with time of exposure in the aquatic environment, although under some conditions there appear to be brief periods of positive growth. The concentration as a function of distance downstream of the discharge point can be calculated by knowing the current speed and direction or by measuring the travel time from place to place with dyes or other tracers. Both physical and biological processes contribute to the die-off, and physical properties of the flow regime are used to compute the dilution factor. Depending on the location and proximity of uses, the degree of treatment, the point of discharge, and the method of discharge can be designed to optimize the rate of natural purification, further reducing downstream pollution problems.

The concentration C of bacterial indicators of fecal contamination has been observed to decrease with time t according to a first-order reaction, the same way in which radionuclides decay with time, where K is the die-off rate constant. Whereas radionuclide decay is characterized by the half-life of an isotope, a parameter frequently employed in water-pollution analyses obtained by solution of the first-order reaction equation is t_{90}, which is the time required for 90% die-off of the bacteria. Thus, given the value of K, we have

$$t_{90} = 2.3/K. \qquad (1)$$

The general solution used to find the concentration at any time C_t, after the initial concentration C_i is determined, is

$$C_t = C_i e^{-Kt}. \qquad (2)$$

To find the concentration of bacteria after effluent has traveled in the river for, say, 8 h, it is necessary to determine the value of K for the specific situation being studied. Results of many studies in lakes and streams have shown that K varies widely, depending on the temperature of the water, the amount of sunlight, and the depth at which the effluent plume travels with the current. Using an average value for fresh water of about $K = 0.038/\text{h}$, the initial concentration $C_i = 100\,000$ per 100 mL would be reduced to about 74 000 per 100 mL in $t = 8$ h through die-off alone. Dilution with ambient and tributary flows would additionally reduce the concentration.

In many cases, it is preferable to predict the concentration at a given distance downstream, rather than at time increments. Most U.S. rivers have a remarkably uniform, low-flow current of 1.5 to 2 km/h, which can be used to convert low-flow travel time to distance. In our example, this travel distance would be between 12 and 16 km. Travel time for other flow conditions vary from river to river, and so field measurements may be needed to relate bacterial concentrations to specific locations downstream.

In analyzing coliform die-off cases in the marine environment, an average value of 1.2/h can be used for K, nearly 30 times greater than that of fresh water. This rapid die-off rate is usually attributed to the salinity of the marine environment, although it may also be related to a greater concentration of predatory animals. In addition, the natural flocculation and sedimentation of particles that occur in estuaries could account for removal of bacteria from the water column. If the sewage data for the above example were instead considered in the marine environment, the concentration after 8 h of exposure would be reduced to 7 per 100 mL. There are, however, disposal practices in the marine environment that can reduce the die-off rate. For example, when an effluent is discharged at a great depth, the die-off can slow down considerably because sunlight cannot penetrate deeply enough for ultraviolet radiation to be an effective biocide.

Noncoliform pathogenic microorganisms may decay at rates different from those of coliform indicators. As methods of diagnosing water-borne illnesses become more sophisticated and specific in relation to identifying the causative organisms, it will be necessary to establish values that can be used to predict die-off rates for specific pathogens rather than the broad coliform indicator group.

In addition to the decrease of bacterial concentrations due to die-off in either fresh or marine surface waters, bacterial concentrations are decreased as the water is diluted with upstream ambient water at the point of effluent discharge and further diluted as it flows downstream. The effect of dilution may or may not be important in meeting water-quality criteria, depending on the initial mixing, the nature of the subsequent flow patterns, and the distance to water use areas.

2.6 Dilution of Effluents

Dilution can be—but is not necessarily—an effective way to prevent pollution of surface waters. Because many people apparently believe that dilution is always the only thing that needs to be done with wastes, while others believe that dilution only distributes and masks the ultimate adverse effect of wastes on the environment, it is necessary to use objective methods of analysis in evaluating aquatic waste-disposal practices. Through the analytical process of risk assessment, it is clear that certain principles underlie the utility of dilution in managing waste discharges. The first principle concerns the concentration dependence of the pollutant response mechanism. That is, if the concentration is reduced sufficiently, will the degree of effect be directly reduced? Further, once the contaminant concentration is reduced, can it subsequently become more concentrated? Reconcentration is a phenomenon often associated with sediments and persistent organic chemicals such as PCBs. Dilution of waste streams containing high concentrations of suspended solids may prevent significant pollution near the discharge site, but such sediments may eventually settle out of the water column and concentrate in depressions in the stream bed. There they can cause a variety of problems. With persistent organic chemicals, whose solubility in water is low and whose affinity for sorption to animal tissues is high, harmful levels can bioaccumulate even from highly diluted mixtures. Thus, dilution may or may not solve a potential concentration-related problem.

Aside from concentration–toxicity considerations, social and economic considerations enter into the decision to use dilution. In addition, its use is also dependent on the availability of a sufficient quantity of dilution water. In the arid southwestern United States, for example, many streams are ephemeral; that is, they contain water only after major rainfall events. In other streams in arid climates, the effluent from municipal treatment plants is the predominant flow for more than 50% of the year; that is, they are effluent-dependent streams. In such situations, dilution is somewhere between small and nil, except during storm runoff. For example, if the stream flow is 10 million liters per day (ML/d) and the effluent flow is 30 ML/d, the contaminants dissolved in the effluent will be reduced in concentration by just 25%, on the assumption that the concentration of each contaminant upstream is zero. (Generally, the effective dilution will be even less than that because there is almost always some measurable concentration of contaminants upstream.) The general equation used in determining the concentration after dilution is

$$c_f = \frac{c_e v_e + c_a v_a}{v_e + v_a}, \tag{3}$$

where c_f = cross-sectional–average final concentration in the stream, c_e = concentration in the effluent, v_e = volume flux of the effluent, c_a = concentration in the ambient dilution water upstream, and v_a = volume flux of the ambient dilution water.

In large rivers, the amount of dilution initially achieved depends on the method used to discharge the effluent into the river. To maximize the initial dilution, it is necessary to employ a diffuser. For example, if the river flow is 120 ML/d, and a diffuser is used to distribute the effluent across the whole width, a discharge of 30 ML/d could result in an 80% reduction in concentration of contaminants if the ambient concentration is zero. The final concentration would be one-fifth the effluent concentration. Of course, the ambient concentration is almost always greater than zero, so that Eq. (5) must be used to estimate the final concentration accurately.

In many streams construction of diffusers may be either inappropriate or prohibited. Consequently, the rate at which the concentration of a contaminant decreases through turbulent mixing (dispersion as distinguished from molecular diffusion) depends on natural processes that occur during stream flow. A rough guide to expected dilutions can be provided, based on assumptions of steady flow of the sewage and the river, using an empirical equation expressing the distance X to achieve complete mixing of the contaminant across the width and depth of the stream as a function of the average current speed U, the width W, and a measured transverse mixing coefficient ϵ (Fischer *et al.*, 1979):

$$X = 0.4UW^2/\epsilon. \tag{4}$$

But effluent flows are not steady, and there is a wide variation in widths and transfer coefficients, and so it is not possible to estimate a range of values for the dilution rate in cases like this. Without specific case-by-case data to generate reliable estimates, regulatory agencies often resort to "worst-case" scenarios to predict the impact of a discharge on downstream water uses. The rationale is that it is prudent to assume that without a physical structure, like a diffuser, to mix the effluent into the river, only the natural die-off processes can be considered in predicting downstream concentrations. This may result in overestimating downstream concentrations of many toxic pollutants that have very slow degradation rates, but, as will be shown subsequently, it is by far a necessary approach for predicting downstream bacterial concentrations.

Large bodies of water, particularly open coastal waters, offer much greater opportunity for effective initial dilution and long-term dispersion of waste streams. The dilution achieved with a multiple-port diffuser depends on the depth of water, along with other physical parameters. For most cases solutions can be estimated satisfactorily with normalized graphical results (Fischer *et al.*,

1979) or digital output from computer programs (Baumgartner *et al.*, 1994). The rate of dilution that can be achieved with a multiple-port ocean outfall is quite commonly well over 100 to 1 within a minute of discharge in waters 20 m deep or more, using a diffuser section less than a kilometer in length. An initial dilution of 100 to 1, commonly sought in design of outfalls, is sufficient to reduce toxicity of many chemical contaminants to an acceptable level.

In the effective placement of ocean outfalls, depth is not the only determining factor. If, for example, a location chosen for its great depth has poor circulation, the net result may be less effective dilution of wastes than that offered by placement in shallower but more open water. Such considerations are a major concern in the placement of outfalls in fjords, bays, and sometimes estuaries. When bacterial die-off is an important consideration, the *distance* from a designated use area is usually a more important factor than initial dilution *per se*. In this case, the time it takes microbes to travel a long distance increases the likelihood that they will be inactivated. Most modern coastal cities employ multiple-port ocean outfalls far off shore to protect beaches and near-shore recreational areas from the effects of bacterial contamination.

Following this initial dilution, waste streams undergo additional dilution, or dispersion, as they are transported by ambient currents and mixed with the surrounding water by turbulence. The process is analogous to the dispersion that takes place when smoke plumes dissipate in the atmosphere after the smoke has risen to an equilibrium level. In some aquatic systems, however, the effluent plume is not as easily observed.

One simplified equation that has been used successfully over the past 30 years for large bodies of water (Brooks, 1960) provides the maximum concentration at a distance X downcurrent from the point where horizontal transport begins:

$$c_{\max} = c_{pi} \operatorname{erf}\sqrt{\frac{Ub^2}{16\epsilon_0 X}}, \qquad (5)$$

where $c_{\max}$ = centerline (maximum) concentration at distance X, c_{pi} = plume concentration at the end of initial dilution, $\operatorname{erf}(x)$ = standard error function of (x), U = current speed in the X direction, b = width in the Y direction (orthogonal to X) at the end of initial dilution, ϵ_0 = constant horizontal (Y direction) eddy diffusivity, and X = travel distance [note that U/X can be replaced by $1/t$ (time)].

A typical problem is to estimate the highest concentration that would occur near a recreational shellfish area from an outfall 8 km away when the current speed (U) is 15 cm s^{-1}. The width of the plume after initial dilution, which would be obtained from local observations, would likely be about 1000 m. A commonly cited value for near-shore eddy diffusivity is $\epsilon_0 = 10^4$ cm^2 s^{-1}. The result is typically surprising: it shows that even with a travel distance of 8 km in a large, open body of water, the concentration is diluted to only 87%, that is, by a factor of about 1.2. Contrast this factor of 1.2 after traveling 8 km to an initial dilution factor of 100 achieved during a rise of 20 m.

The example shows that initial dilution is often more important in reducing harmful levels than is dilution due to dispersion associated with ambient currents. This is true for most contaminants, including heavy metals, ammonia, biodegradable organics, and toxic organics that are found in partially treated effluents. But in the case of indicator bacteria, the transport distance is more important than initial dilution because a long time is needed for die-off of bacteria to allowable levels. Assessments of the type shown in this example have demonstrated to the satisfaction of regulatory authorities that chlorination or other bacterial disinfection methods may not be required when treated municipal effluents are discharged to appropriate marine waters. Use of chlorine has been discouraged because of the fear of creating chlorinated organics in the effluent stream that could bioaccumulate in aquatic organisms to detrimental levels, or in human consumers.

2.7 Organic Matter and Dissolved Oxygen

Biodegradable organic compounds are decomposed by bacteria and other organisms that live in surface waters. While some organics are mineralized to carbon dioxide and oxides of nitrogen, others are synthesized into more microbial biomass, most of which is subsequently decomposed as well. All this decomposition consumes dissolved oxygen (DO), upon which many desirable species of fish, other aquatic organisms, and wildlife depend. Depressed dissolved-oxygen concentrations are inimical to these life forms. For example, some fish can survive at concentrations near 1 mg L^{-1}, but most are adversely affected at DO concentrations below 4 mg L^{-1}. Dissolved oxygen is a precious resource in surface waters because there is not very much to begin with. The maximum amount of oxygen a pure surface water can hold is a function of salinity and temperature; but, compared to the maxima of many other substances, it is remarkably low. Fresh water contains from 8 to 15 mg L^{-1} DO over the temperature range 30–0 °C, respectively, and sea water contains 6–11 mg L^{-1} over the same range.

Domestic sewage can contain about 300–400 mg L^{-1} of organic compounds, 60% of which is readily degradable by bacteria commonly found in nature. *Readily degradable* implies that most of the material will be decomposed within about a week in a suitable aquatic environment. The change in the concentration of the organic matter with time is conveniently described by the first-order decay equation used to describe bacterial die-off. However, the value of this K depends on the specific organic compounds in the sewage. For domestic sewage, an average value is about $K = 0.4/\mathrm{d}$, ranging from 0.1 to 0.7/d. As more industrial wastes are contributed to the sewer system, the rate constant may increase or decrease. The amount of organic material discharged to a surface water depends on the population served by the municipal sewer system.

2.8 Impact of Organics on Dissolved Oxygen of Receiving Waters

The biochemical oxygen demand (BOD) is the most commonly used parameter in the analysis of the effects of organics on oxygen resources in water. The BOD is the amount of oxygen used over the course of time, usually 5–20 d, as the organic matter is oxidized both microbially and chemically. A laboratory method is used to determine BOD from samples of treated sewage obtained over various time periods. The data obtained in the laboratory are used with the first-order BOD equation,

$$\mathrm{BOD}_t = \mathrm{BOD}_L(1 - e^{-Kt}), \tag{6}$$

to provide estimates of the reaction rate coefficient K and the ultimate value, called the limiting value of BOD, or BOD_L. (This K is not related to the K used in computing bacterial die-off.)

Typically, in more developed countries of the world, municipal sewage is treated to remove organics, and hence to reduce BOD, before discharge to surface waters. The amount of BOD remaining after treatment can range from 10% to 70% of the amount originally in the sewage. The impact on the DO resource of receiving waters depends on the amount of BOD added, the rate constant, the amount of DO in the water body before the sewage is added, and the rate at which additional oxygen is transferred from the atmosphere to the receiving water. Many rivers and streams have depressed initial DO concentrations, that is, a DO deficit, at the point where an effluent is discharged because of BOD added by cities upstream.

Most DO problem situations require that we determine the oxygen deficit resulting from the simultaneous effects of the oxygen demand of a waste and the competing restoration of oxygen from atmospheric reaeration. The combined result is termed the self-purification capacity of the water body. The deficit shown graphically as a function of time (or distance) is known as the oxygen sag curve because of its characteristic shape (Fig. 1). From this curve the lowest DO value likely to occur in the water course can be predicted, the location where minimum DO occurs, the amount of deficit caused by the discharge of the effluent, and the time required to return to the natural saturation level of DO. In this example, a BOD of 30 mg L^{-1} results from a continuous discharge of an industrial effluent mixed into a small stream saturated with DO at 10 mg L^{-1}, causing the DO to drop to 4.2 mg L^{-1} by the fifth day, i.e., causing a deficit of 5.8 mg L^{-1}. By this time the effect of reaeration surpasses the rate of oxygen demand and the deficit decreases with time as more and more DO is restored. By 40 d the DO level is nearly back to the saturation concentration. In many cases an effluent is only mixed with part of the stream flow and continuing downstream dilution with uncontaminated water would restore the DO more quickly.

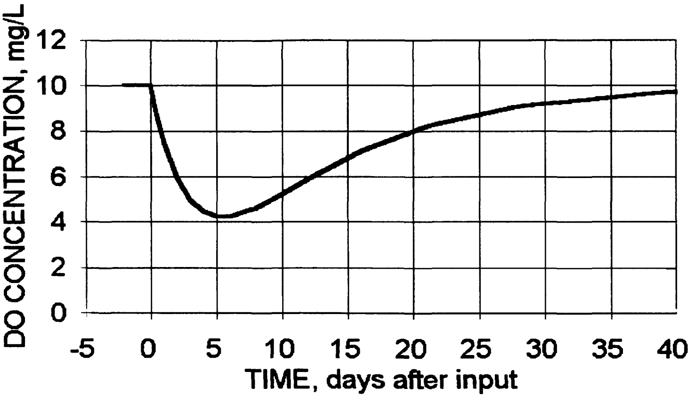

FIG. 1. Oxygen sag curve, for BOD of 30 mg/L, K = 0.1/d, and R = 0.3/d.

Stream-flow rates for near worst-case low-flow conditions are used to convert time to distance in order to identify stream locations where critical conditions might occur. Steady-state solutions of this type are useful for many regulatory purposes, although time-varying analyses can be performed with computer programs for more precise estimates.

Sometimes, water-quality regulators impose limits both on the amount of deficit per se and on the resulting DO value itself. For example, a regulation could require that a waste discharge must neither increase the deficit by more than 10% nor depress resulting DO concentration below 5 mg L^{-1}. Additional regulations may require more rapid return to saturated DO conditions to protect resources farther downstream. If the results of this evaluation are unfavorable, a lesser environmental impact can be achieved by use of one or more alternatives, including more treatment, moving the point of discharge, and reducing the amount of organic material entering the sewer system.

When water systems are heavily used, or highly valued water uses are threatened, additional factors require consideration in the DO analysis:

1. the diurnal demands and supplies of oxygen resulting from photosynthesis;
2. the oxygen demand of organic materials deposited in the sediment layer;
3. the oxygen demand of nitrogen compounds discharged in the effluent;
4. variations in the reaeration coefficient R with travel time due to flow conditions in the water body; and
5. the wide temporal range of sewage flows.

Comprehensive computer programs are available to describe the concentration of oxygen in large watersheds consisting of dozens of interconnecting streams and dozens of waste-water inputs; however, data on plume travel, K values, and R values still have to be obtained with time-consuming laboratory studies and physically demanding field studies. Dilution of wastes is one of the most important factors to consider in assessing impacts. The methods used to assess effects on bacterial concentrations are applicable to assessment of BOD effects on the dissolved-oxygen resource of a water body. If the dilution capacity of the water body is not exploited, only the decomposition of organics and reaeration can be considered in the restoration and maintenance of DO.

3. TOXIC SUBSTANCES

In addition to water-pollution problems caused by bacteria and BOD, referred to in some regulations as "conventional pollutants," chemicals added to a municipal sewer system can escape complete removal in standard waste-water treatment units and cause harm to organisms in the receiving water. In the United States the Toxic Substances Control Act gives EPA the authority to regulate certain types of chemicals that may present an unreasonable risk to health and the environment. Of the tens of thousands of chemicals that might be covered under this rubric, EPA has listed several dozen organic and inorganic compounds and a handful of pesticides, labeled as "priority pollutants" for regulation in waste waters discharged to water courses. Individual states may add others. The list includes benzene, dioxin, methyl chloride, phenol, and vinyl chloride. Many are from industrial discharges to the system, but the source of some is household and municipal use of pesticides and cleaners. Even though toxic chemicals occur in relatively small concentrations in waste waters compared to conventional pollutants, frequently in the range of micrograms to milligrams per liter, the annual mass emissions are impressive amounts because of the large volumes of treated effluents returned to the aquatic environment. A sense of the scale involved is provided by data from the seven large municipalities in southern California that discharge treated effluents into coastal waters of the Pacific Ocean (National Research Council, 1990b). In 1985 the amounts of DDT and PCB carried in 440×10^9 gallons of effluents totaled about one metric ton apiece for the year. Chromium, copper, and zinc emissions were approximately 100, 250, and 400 t, respectively, while BOD and suspended particulate matter emissions were each on the order of 2×10^5 t. Nearly 15×10^6 people are served by these treatment facilities, and although not all population centers have the same mix and substantial contribution of industrial wastes as does this region, the relative proportions could be applied to other areas to acquire a first approximation of the mass of contaminants involved in treated waste-water discharges. While the population and the waste-water volume have grown considerably in the area in the last two decades, the mass of contaminated material has decreased significantly for many industrial chemicals as a result of control at the source, and for domestic contributions as a result of greater use of secondary treatment processes. The greater removal of particulate material (in Los Angeles as well as in every city employing secondary treatment) has resulted in a materials handling problem from the resulting digested and dried municipal sludge. To encourage recycling, EPA has developed stringent criteria for acceptable use of this resource under the term "Biosolids." The use of this material is growing in agricultural areas, although public acceptance is still problematic.

Agriculture is also a source of pesticide contamination of surface waters. The major concern with pesticides is the effect they have on aquatic organisms and riparian wildlife, although human exposure is also considered in establishing maximum allowable concentrations in receiving waters. Allowable concentrations for aquatic species are generally determined from laboratory dose-response studies of organisms selected from several taxa, exposed for several time periods to pure chemicals diluted in a small container of clean water. The shortcomings of this methodology are abundant. In short, it does a poor job of representing the exposure conditions seen in the natural environment. Attempts are continuously being made to incorporate more realistic exposure conditions in laboratory assays. For several years research has been conducted in simulated stream segments (Zischke and Arthur, 1987), and observations in natural settings are being incorporated into the process of promulgating water-quality criteria (U.S. EPA, 1991).

The majority of published work on water-quality criteria is based on 96-h acute toxicity studies in which the concentrations are chosen to produce lethal effects on a portion of the test organisms. A great deal of the pioneering work of this type involved metal ions such as chromium, cadmium, silver, lead, nickel, mercury, and zinc. The form of the metal as introduced into the test water is a factor in the degree of toxicity expressed, as is the ion balance in the water, conveniently represented in a crude way by the carbonate hardness concentration. In the mid 1960s, possibly related to the publication of *Silent Spring* (Carson, 1962), a new wave of research was focused on pesticides, detergents, and other organic materials. The test results are used to calculate a concentration expected to result in lethality for 50% of the test organisms. The 96-h LC_{50}, as it is called, is analogous to medical studies where median effective doses (ED_{50}) and median lethal doses (LD_{50}) are developed. The allowable stream concentration is established as a fraction (e.g., 0.01) of the LC_{50} to account for uncertainties and to attempt to avoid even sublethal adverse impacts. A more direct approach currently being used is to conduct chronic exposure studies to assess one or more of a variety of impacts on growth, fecundity, behavior, disease, and tissue accumulation of contaminants. Attempts are also being made in the use of data for regulatory practice to take into account the sensitivity of aquatic organisms to the frequency of exposures and the duration of each exposure.

For many years criteria were available only for fresh-water systems, but these were applied to marine waters as well. Now criteria are being developed using marine species, and attempts are being made to develop criteria for specific watershed units such as the Laurentian Great Lakes and the ephemeral and effluent-dependent water courses of the U.S. arid West. Regulatory agencies have recognized that an overwhelming resource requirement would be needed to develop criteria for a variety of biogeochemical environments, and an increasing number of specific chemicals, to say nothing about the question of the synergistic and antagonistic effect of a mixture of chemicals in a waste stream. Currently (1996) the U.S. EPA is proposing to require the use of whole-effluent toxicity testing (WET or WETT) as a way to focus available resources on this approach that cuts directly to the question of local standards, realistic exposure scenarios, and composite wastes.

WET testing is a new element in the repertoire of water-pollution regulatory practice (discussed subsequently) that targets point sources as the highest-priority input class for control. That is not to say that responsible regulators view publicly owned sewage treatment plants as villains, necessarily, but rather that they present a convenient point for removing a large amount of contamination from a large volume of water about to enter a water course. Some large industries would also be viewed the same way. Just how effective conventional treatment plants are in removing toxics is somewhat uncertain. Almost all plants have been designed from theory and practice to remove settleable particulate matter and BOD. That bacterial indicators and toxics are removed is due in large part to their fortuitous association with particulate matter. With conventional primary and secondary treatment followed by a modest amount of initial dilution, most toxic standards can be met by conventional plant effluents.

When effluents are subject to a WET test standard, the options for response are more limited for a municipal treatment works than an industry. The municipal plant may have received a slug dose of an unknown chemical from an illegal or accidental one-time discharge to the sewer system. If it is a routine occurrence, technicians might be able to find the input source and prohibit its recurrence. If the causal agent is properly a constituent of the municipal sewage coming into the plant, the plant operators may be able to improve the efficiency of solids and BOD removal, or add supplemental units such as air strippers, carbon filters, or chemical precipitation, all of which would be expensive for treating the whole flow of the plant. Industrial dischargers have only their plant processes to contend with and may be able to sort out the causative agent and control it with process changes, substitution of chemicals, or pretreatment of individual small process streams. The major concern with employment of WET testing is that test results could force the removal of constituents that while toxic to organisms exposed in laboratory bioassays would not be a problem once discharged to a flowing stream environment where there would be a decay of the toxic agents in addition to dilution (Parkhurst and Mount, 1991). One example is the case of ammonia in municipal effluents, which could be lethal in sufficient concentration to some fish species used in the test but which are not present in the water course, as for example in many arid West washes. Furthermore, the ammonia is oxidized to nitrate that is of a substantial nutritional value to trees and other foliage along these washes (Baumgartner *et al.*, 1993).

4. NONPOINT (DIFFUSE) SOURCES OF POLLUTION

Agriculture is the largest persistent diffuse source of contaminants to surface waters. The primary contaminants are pesticides, fertilizers, salts, and particulate matter. The components of this industry include traditional grain and table-crop farming, dry-land farming dependent on irrigation, production of fiber (e.g., cotton), forestry, cattle grazing and pasture, animal feed lots, and aquaculture. Depending on size, feed lots and aquaculture operations may be subject to point-source discharge regulations and require individual permits. Cattle operations can be a significant source of fecal coliforms to receiving waters.

Overland runoff and channelized irrigation return flows contribute turbidity and nutrients to water courses, causing photosynthetic growths that diminish the aesthetic quality of the water. In some cases, especially during warm, low-flow times of the year, the growth of aquatic plants and animals depletes oxygen to dangerously low levels. In the arid sections of the nation, irrigation return flows are high in dissolved salts as a result of high evapotranspiration rates and the initially high solids concentration of ground water used for irrigation.

Control efforts directed at agricultural runoff include development of Best Management Practices (BMPs) and education to help practitioners conserve water and chemicals, retard and capture water tending to run off too quickly, and treat and reuse water contaminated with animal waste on site. BMPs provide economic incentives as well as environmental payoffs. In some states farmers can be fined if they are not able to demonstrate compliance with BMPs.

In addition to diffuse agricultural contaminant inputs, regulators have discovered that significant input occurs from runoff from urbanized areas. This discovery was enhanced by the discovery that water-quality deterioration persisted even after vigorous control of point-source discharges was achieved. The search for additional inputs led to the conclusion that there were large pollution loads distributed over many diffuse sources. These inputs can occur at a rather uniform rate in regions such as around Puget Sound in the state of Washington where winter rains provide a steady runoff, or more sporadically as in areas prone to thunderstorms of short duration but high intensity. Technological advances have been made in dealing with the latter, consisting of devices and reservoirs to divert and capture the first flush of stormwater systems.

Rural areas, irrespective of agriculture, are also a source of domestic-sewage pollution of water supplies. A large percentage of rural residences are not connected to sewerage systems, relying instead on individual septic tanks for waste disposal. Even when septic tanks are operated properly, and many are not, the effluent can contribute large quantities of nutrients, BOD, and intestinal microorganisms to surface waters and ground-water aquifers. In particular, the nitrogen content of the effluent, in the form of either ammonia or organic compounds, is oxidized by soil bacteria to nitrates, which can be a health hazard for infants consuming water containing greater than 10 mg/L nitrate nitrogen (U.S. DHEW, 1962). The natural organisms in the infant's gut reduce the nitrate to nitrite, and nitrite interferes with hemoglobin uptake of oxygen, causing methemoglobinemia, also known as blue-baby syndrome.

5. GROUND-WATER POLLUTION

Pollution of ground-water aquifers has become a matter for scientific and regulatory interest because of high-visibility pollution cases involving inadequate disposal of industrial materials, primarily chlorinated organic solvents used in cleaning electronic components such as circuit boards. Another category of organic compounds frequently identified as contaminants is "BETX" (benzene, ethylbenzene, toluene, and xylene) associated with the petroleum industry and leaking underground storage tanks (LUST). There is evidence that these chemicals may be of human health concern, primarily as cancer-causing agents. An additional concern is that chlorination of ground water in water-purification plants can create chlorinated organics that may be hazardous. Municipal landfills of household and industrial wastes can contribute some of the above chemicals and are frequently a major contributor of metals.

Previously, the major concern with ground-water quality was the incursion of sea water into the aquifer due to reduction of the water table by excessive withdrawal in coastal municipalities. This resulted in reduced quality in domestic and agricultural well water from the increase in dissolved solids but was not a direct health concern. In some areas the rate of intrusion was controlled by the injection of treated municipal sewage effluent into the aquifer at the intrusion front, which in itself was an action that caused health concerns due to the unknown fate of viruses and other contaminants in the effluent. In arid regions, effluents and surface waters (stream flows and storm-water flows) are being used to replenish aquifers, either by injection wells or infiltration basins, again raising questions about contamination (Kanarek and Michail, 1994).

A number of treatment techniques have been developed, many still in the exploratory stage, to deal with the problem of pollutants in ground water. In some cases it may be possible to remove the soil layer that contains the pollutants and either treat the soil or dispose of it in an acceptable location. In other cases the water in the soil layer can be extracted and treated, to be either returned to the soil or used directly for some aboveground purpose. For volatile organics, an acceptable treatment method is air-stripping, caution being required to assure compliance with air-quality standards. Depending on the soil type, organics may be adsorbed to the clays so that the concentration in the water is established by a partition coefficient. As the water is removed from the affected area by pumping, the concentration decreases because a contact time is needed for a new equilibrium to be established. There is a trade-off between pumping too rapidly, resulting in lower concentrations, and pumping too slowly, increasing the time required to remove the pollutant. Some organic chemicals are more readily transported with the water (e.g., polar compounds), while others (nonpolar organics) have high retardation coefficients with respect to the transport of water.

In situ methods have been used to treat pollutants in the soil–aquifer matrix using a variety of methods. Sparging (injection of air) is one way to effect volatilization of organic compounds and to add oxygen to aid in aerobic biodegradation of contaminants. Also, chemicals have been added to aid in the destruction of pollutants, a good example being nitrates that can be reduced to nitrogen gas if sufficient carbonaceous organics are added as an energy source for soil microbes. Ethanol, methanol, or other easily biodegradable organics have been used. One method that has lost favor is the injection of (presumably) special strains of bacteria purported to have specific ability to attack certain chemical forms. More recently the emphasis has been on supplying nutrients, such as nitrogen and phosphorus, and relying on naturally occurring microbial communities to grow and complete the job of biodegradation. Details of these and other methods are being published regularly, as, for example, by Zoller (1994).

Phytoremediation is a method being researched using plants to sequester metals from polluted soil or ground water. This is an extension of the methods using plants such as water hyacinths in constructed or natural wetlands as one step in the treatment of municipal sewage. One application in the treatment of ground water is to use the plants solely as a concentrator, intending to dispose of the plants along with the contaminants in an acceptable manner. A variation is to extract the concentrated contaminants for reuse (e.g., uranium) or to treat the plant tissues in some way before disposal. Ground water contaminated with nitrates can be used to irrigate certain crops that not only tolerate nitrates but grow luxuriantly exposed to high nitrates. Selenium, a contaminant of many western soils and ground waters, can present a hazard to animals if they are fed crops that accumulate selenium. However, if the crops are harvested and used as feed supplements in areas where cattle experience selenium deficiency, two favorable objectives are achieved: water reuse and improved crop quality. In arid climates, however, the quality of ground water may deteriorate over all, in spite of loss of certain contaminants to irrigated plants, as a result of high water loss to evapotranspiration through the plants. Overall, the evaporative loss of water will cause the concentration of total dissolved solids to increase to the level that interferes with subsequent reuse of the irrigation return flows.

6. REGULATORY PRACTICE

Water-pollution regulatory practice in the United States is founded on a system of discharge permits—known as the National Pollution Discharge Elimination System (NPDES). Holders of these permits, e.g., municipal sewage treatment authorities and industries, must comply with the restrictions and requirements of their particular permit, such as limits on concentrations and mass emission rates of specific constituents. They also have to meet water-quality standards established for the water body into which they discharge their effluents. Some permits, especially for coastal-water discharges, require elaborate environmental monitoring projects. The permit holder is required to conduct monitoring activities and report the results to demonstrate compliance with permit conditions specifically, and generally to demonstrate the preservation of environmental quality. On occasion, regulatory agencies conduct special studies to verify and revise monitoring programs. Monitoring data reflect the wide variations of conditions found in the natural environment, and dischargers and regulators are often challenged to rationalize monitoring results with predictions used in setting permit conditions.

Permit conditions of regulatory agencies usually allow exceptions to one or more of the water-quality criteria within a mixing zone adjacent to the point of discharge. A mixing zone might be established by purely arbitrary considerations or by use of data and simulations with mathematical models. Many mixing-zone determinations are made on the basis of the expected dilution rate that will be provided by efficient diffuser designs intended to optimize initial dilution. For large bodies of water, a common approach is to describe the width of the zone as the depth of water at the disposal site, and the length as the length of the diffuser. For large rivers, it is common to restrict a mixing zone so it does not extend completely across the river, thereby leaving a "safe passage" for aquatic species to avoid high concentrations of waste-water constituents. But sometimes the shape of a mixing zone is entirely arbitrary, say a rectangular zone downstream from a discharge pipe equal to one-fourth the width of the stream and extending downstream for 1 km. Frequently, there are two or three mixing zones for different groups of contaminants and degrees of toxicity.

Regulatory interest may be appropriately directed toward both discrete and average values of contaminants. For example, the state of California and the U.S. EPA specify maximum allowable instantaneous values for some parameters as well as several temporal average values (e.g., 30-day and 6-month arithmetic means). In some cases, these regulations are based on knowledge of the effects on aquatic organisms. In other cases, these values are specified to acquire statistics on the performance of the waste-water treatment plant.

Beyond the mixing zone, there may be regions where current streams of diluted effluent, each leaving the zone at a different time in different direction, would converge over a reef, a kelp forest, or a swimming area. In this case, the frequency and duration of exposure may be more important than the highest observed concentration in assessing the overall impact on these resources.

6.1 Risk Assessment

Except for certain special considerations, NPDES permits require a degree of treatment that is standardized throughout the country. That is, there is a "technology" standard that is applied irrespective of the impact of the effluent on the environment. Generally, the standard technology is identified as "secondary treatment," which involves a biological treatment method following plain sedimentation to remove 90% of BOD and 90% of the suspended solids. A temporary modification of the secondary treatment requirement has been granted to several coastal communities that have demonstrated that this level of treatment is not necessary to meet water-quality requirements and to protect the marine environment. In other cases where the receiving water is termed "water-quality limited," additional treatment requirements have been imposed.

Critics of the technology-standard approach have argued that a rational approach should be applied to all discharges, e.g., a water-quality–based approach. They argue that valuable resources are wasted if a technology is required for technology's sake unless it can be demonstrated that a problem would arise if the technology were not employed. They also point to places where in spite of the employed technology there is no improvement in environmental quality. The scientific arguments persist on both sides and frequently are resolved in court. The modification desired by coastal communities was incorporated into federal law in 1972 and was in effect for about a dozen years. It was the subject of federal legislation introduced in 1994 to grant a second opportunity for San Diego, California, to apply for an exemption from conventional biological secondary treatment, following improvements made in the treatment methods and extension of the outfall. San Diego's attempt to acquire a permit modification under the 1972 provision was not successful.

There is a possibility that a new paradigm can be developed to regulate water quality using the methods of risk assessment applied in a variety of health programs. The promise posed by risk-assessment techniques is not guaranteed to be entirely applicable in water-quality science, because there are social considerations, including economic and political implications, in addition to the physical and biological science questions that are addressed (Freudenburg, 1988). Nevertheless, in 1992 the U.S. EPA published a framework for ecological risk assessment as a basis for making environmental decisions, hoping to provide a means "... of targeting environmental protection resources at the problems and the geographic areas posing the greatest risks" (Norton *et al.*, 1992). Risk assessment is a scientific process that interfaces with a regulatory role of risk management (Fig. 2). On the science side are distinct tasks to characterize the resource (or site) at risk in terms of the biological communities and the physical parameters that represent the natural environment of the resident and transitory species. The site needs to be characterized also in terms of the values assigned by societal interests, such as recreation, and preservation of certain species. An important technical task is to characterize the nature of the stress on the site, such as the magnitude or intensity of the contaminant, the variability of its input function, and its temporal and spatial fate within the environment at risk. Results of site and source characterizations are integrated to produce a quantitative exposure assessment, i.e., a determination of what the concentration (or other measure of intensity) will be for the resource components at the site. Next, a quantitative description of the response expected is generated, based on laboratory dose-response studies, etc. The information and data resulting from exposure and response assessments are used to calculate a measure of risk, sometimes as simply as the quotient of response over exposure (Norton *et al.*, 1992), or in other cases a qualitative expression of risk.

The output of the risk assessment, according to at least one model (Baumgartner and Duedall, 1990), is then used by a regulatory program to make risk-management decisions on the content of permits, to monitor recovery and preservation of the resource and to pursue legal remedies as needed. From the beginning of the regulatory process and at various times during the permit period, there will be a need for evaluation of alternative strategies, providing a continuing interface for the scientific and regulatory roles. Many

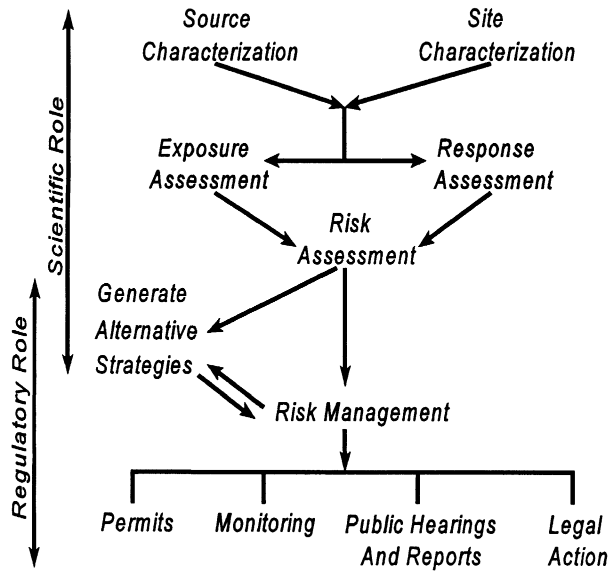

FIG. 2. Risk assessment process (after Baumgartner and Duedall, 1990).

of the steps in this process, and the scientific methodology for obtaining the necessary data, are quite similar to the process used by the U.S. EPA to develop water-quality standards and the water-quality–based approach for toxics control described earlier. The difference, if there is one, may be that under risk assessment the general application of technology standards would be abandoned as a way to control pollution.

GLOSSARY

Anthropogenic: Pertaining to human activity as distinguished from natural processes.

Aquifer: A soil stratum saturated with flowing water.

Biogeochemical: Pertaining to the differences and similarities in geographic regions as defined by biological, physical, and chemical characteristics of the environment.

Carbonate Hardness: The sum of calcium and magnesium ion concentrations in water expressed as their carbonate equivalents.

Combined Sanitary and Storm Sewer: A municipal sewer that transports water-carried household wastes and storm-water runoff from street inlets. Distinguished from separate sewers for each purpose in modern municipalities.

Diffuser: A device to spread a liquid flow over a wide expanse of space, usually on the end of a pipe.

Digested Sludge: Slurry of dense particulates obtained from sedimentation basins in a sewage treatment plant and retained for several weeks in a heated reactor to facilitate anaerobic microbial decomposition of organic matter.

Dilution: Reduction in concentration due to mixing a fluid with relatively high concentration into a fluid with a lower concentration.

Dispersion: A natural process of turbulent energy exchange resulting in mass transfer from regions of high concentration to low concentration.

Estuary: Body of water between the open ocean and a fresh-water stream in which ocean water is exchanged and diluted with stream water.

Fecundity: Success at producing offspring.

Fjord: A type of estuary, usually narrow and deep, characterized by a sill extending from the seabed to near the surface.

Henry's Law: The linear relationship between the equilibrium concentration of noble gases in adjoining gaseous and aqueous phases through a constant specific for each gas.

Initial Dilution: Dilution of a waste in the near-field flow region of an effluent discharge pipe due to turbulence caused by the effluent being pumped into the stream with high energy, generally through many small outlets.

Metric Ton (t): One thousand kilograms, thus 2205 pounds or 1.1 U.S. tons.

Mixing Zone: A portion of a water body where waste-water streams are dispersed and pollution-control standards are relaxed.

Riparian: Of, situated, or dwelling on the bank of a river or other body of water.

Spawning Habitat: Section of a water body used by fish to deposit and fertilize eggs.

Taxa: Categories in a hierarchical classification system of species.

Xenobiotic: Of or pertaining to an organic compound that is foreign in the evolutionary experience of organisms and possessing an adverse biotic effect on organisms, e.g., a synthetically produced pesticide.

Works Cited

Baumgartner, D. B., Duedall, I. W. (1990), in: D. B. Baumgartner, I. W. Duedall (Eds.), *Oceanic Processes in Marine Pollution,* Vol. 6, Malabar, FL: Robert E. Krieger, Chap. 1.

Baumgartner, D. J., Smith, E. L., Frye, R. J., Riley, J. J., Smith-Otero, K. A. (1993), *Rationale for a Program of Research to Develop Water Quality Criteria for Effluent-Dependent Ephemeral Streams and Riparian Habitats in the Arid West,* Report to Pima County Wastewater Management, Tucson, AZ: Environmental Research Laboratory.

Baumgartner, D. B., Frick, W. E., Roberts, P. J. W. (1994), *Dilution Models for Effluent Discharges,* Washington, D. C.: U.S. Environmental Protection Agency.

Brooks, N. H. (1960), in: E. A. Pearson (Ed.), *Proceedings of the First Conference on Waste Disposal in the Marine Environment,* New York: Pergamon, p. 246.

Browner, C. M. (1994), *EPA J.* **20,** 6–9.

Carson, R. (1962), *Silent Spring,* Boston: Houghton Mifflin.

Fischer, H. B., List, E. J., Koh, R. C. Y., Imberger, J., Brooks, N. H. (1979), *Mixing in Inland and Coastal Waters,* New York: Academic.

Freudenburg, W. R. (1988), *Science* **242,** 44–49.

Kanarek, A., Michail, M. (1994), in: U. Zoller (Ed.), *Groundwater Contamination and Control,* New York: Dekker, p. 457.

Marchetti, C. (1977), *Clim. Change* **1,** 59–68.

National Research Council (1990a), *Managing Troubled Waters: The Role of Marine Environmental Monitoring,* Washington, D.C.: National Academy Press.

National Research Council (1990b), *Monitoring Southern California Coastal Waters,* Washington, D.C.: National Academy Press.

Norton, S. B., Rodier, D. J., Gentile, J. H., van der Schalie, W. H., Wood, W. P., Slimak, M. W. (1992), *Environ. Toxicol. Chem.* **11,** 1663–1672.

Parkhurst, B. R., Mount, D. I. (1991), *Water Environ. Technol.* **3,** 45–47.

U.S. Department of Health, Education, and Welfare. Public Health Service (1962), *Drinking Water Standards,* Washington, DC: U.S. GPO.

U.S. Environmental Protection Agency. Office of Water (1991), *Technical Support Document for Water Quality-Based Toxics Control,* Washington, DC: EPA.

van der Leeden, F., Troise, F. L., Todd, D. K. (1990), *The Water Encyclopedia,* Chelsea, MI: Lewis Publishers.

Zischke, J. A., Arthur, J. W. (1987), *Arch. Environ. Contam. Toxicol.* **16,** 225–231.

Zoller, U. (Ed.) (1994), *Groundwater Contamination and Control,* New York: Dekker.

Further Reading

Bacow, L. S., Wheeler, M. (1984), *Environmental Dispute Resolution,* New York: Plenum Press.

Jolley, R. L. (1978), *Water Chlorination Environmental Impact and Health Effects,* Vol. 1, Ann Arbor, MI: Ann Arbor Science.

Jolley, R. L., Gorchev, H., Hamilton, D. H., Jr. (1978), *Water Chlorination Environmental Impact and Health Effects,* Vol. 2, Ann Arbor, MI: Ann Arbor Science.

Tchobanoglous, G., Maitski, F., Thompson, K., Chadwick, T. H. (1989), *Res. J. WPCF* **61,** 1625–1635.

U.S. Environmental Protection Agency. Office of Research and Development (1988), *Design Manual Constructed Wetlands and Aquatic Plant Systems for Municipal Wastewater Treatment,* Cincinnati, OH: Center for Environmental Research Information.

Vigil, K. M. (1996), *Clean Water: The Citizen's Complete Guide to Water Quality and Water Pollution Control,* Portland, OR: Columbia Cascade.

WAVES, ELECTROMAGNETIC

See ELECTROMAGNETIC RADIATION; ELECTROMAGNETIC WAVE PROPAGATION

WAVES, MECHANICAL AND ELASTIC

See MECHANICAL AND ELASTIC WAVES IN SOLIDS

4
The Standard Model of Elementary Particles

Jens Erler

4.1 Introduction 121
4.2 Principles and Theoretical Foundations of the Standard Model 121
4.2.1 Quantum Field Theories 121
4.2.2 Local Gauge Symmetries 123
4.2.3 Gauge Group of Electroweak Interaction 124
4.2.4 Gauge Bosons 125
4.2.5 Spontaneous Symmetry Breaking 125
4.2.6 Higgs Mechanism 127
4.2.7 Infinities 128
4.3 The Standard Theory 129
4.3.1 Lepton Replication and Muon Decay 129
4.3.2 Quarks and Quantum Chromodynamics 131
4.3.3 Quark Mixing 132
4.3.4 CP Violation 133
4.4 Experimental Tests 134
4.4.1 Z^0 Pole Physics 134
4.4.2 Low Energy Measurements 136
4.4.3 Global Analysis 137
4.5 Beyond the Minimal Model 139
4.5.1 Accidental Symmetries and Nonrenormalizable Corrections 139
4.5.2 Lepton Number Nonconservation and Neutrino Mass 140
4.5.3 Baryon Number Nonconservation and Proton Decay 141
Glossary 142
References 143
Further Reading 145

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

4.1
Introduction

All known phenomena in the physical world are aspects of only four basic forces: gravity, electromagnetism, the strong force, and the weak force. While these forces appear to have very different characters, all of them are based on quantum field theories (QFTs) exhibiting some form of *local gauge symmetry*. Although the historical development was different, with hindsight local gauge invariance is now seen not as a fundamental principle but as a consequence of the more basic principles of Lorentz invariance and cluster decomposition, as well as the axioms of quantum mechanics. Section 4.2 first reviews how the concepts of a QFT and of local gauge symmetry naturally arise before they are applied to the prototype electroweak model of electrons, e^-, positrons, e^+, and electron-neutrinos, ν_e. Section 4.3 discusses a series of extensions of this model to include quarks and the strong interaction, family replication and quark mixing, and CP violation. Experimental tests and constraints on the Standard Model (SM) are the subject of Section 4.4. Finally, Section 4.5 describes limitations of the minimal model and goes beyond Section 4.3 to incorporate neutrino mass and mixing. It also discusses accidental symmetries and their possible violation by nonrenormalizable corrections.

4.2
Principles and Theoretical Foundations of the Standard Model

Rather than reviewing how the SM emerged historically as the correct theory of weak and electromagnetic interactions, we first discuss the relevance of QFTs in general (Section 4.2.1) and gauge theories in particular (Section 4.2.2), before specializing in the electroweak gauge theory in Section 4.2.3. Key features like non-Abelian gauge symmetry, spontaneous symmetry breaking and the Higgs mechanism, renormalizability, and anomaly cancellation in chiral gauge theories are covered.

4.2.1
Quantum Field Theories

The mathematical framework to study relativistic particles and their interactions is provided by QFTs. A particle can be defined as an irreducible unitary representation (irrep) of the inhomogeneous Lorentz group (Poincaré group), which is characterized (classified) by its mass, m, and either its spin (the eigenvalue of the square of the angular momentum operator, $\vec{J}$) or

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

in the massless case its helicity,[1] h (the eigenvalue of the third component, J_3, of $\vec{J}$). It is convenient to introduce quantum mechanical operators, $a^\dagger$ and a, respectively creating and annihilating particles of a particular species with a specific relativistic four-momentum (energy-momentum) and spin (including its direction). This serves to satisfy (also in nonrelativistic theories) the general scientific requirement that distant experiments should yield independent results (locality), that is, quantum mechanical probabilities factorize (cluster decomposition). The most general Hamiltonian is then a sum of products of any number of $a^\dagger$ and a operators, which are recalled to be defined in momentum space.

However, a relativistic theory requires, in addition to particles transforming according to representations of the Poincaré group, also Lorentz-covariant scattering amplitudes (S-matrix). Invariance under Lorentz transformations (boosts) presents an obstacle, because (according to Noether's theorem) there must be a conserved vector related to them, but (unlike linear and angular momentum operators) boost generators fail to commute with the Hamiltonian. This complication introduces an additional condition (with no counterpart in nonrelativistic theories), which often takes the form of a causality requirement, that is, Hamiltonian densities at spacelike separation ought to commute. The important point is that causality conditions are expressed in configuration (position) space.

Therefore one integrates the creation and annihilation operators over all spatial momenta[2] $\vec{p}$ (Fourier transformation)

$$\phi_l(x) = \int \frac{d^3p}{(2\pi)^{3/2}} \frac{1}{\sqrt{2p^0}} \times (u_l\, a_{\vec{p}}\, e^{-ipx} + v_l\, a^\dagger_{\vec{p}}\, e^{ipx}) \qquad (1)$$

yielding functions of space-time with a set of discrete Lorentz and spinor indices (collectively called l) giving rise to finite-dimensional nonunitary (unless a Lorentz scalar) and in general reducible representations of the homogeneous Lorentz group. The Fourier transforms (1) are the quantum fields. The Hamiltonian is now a sum of products of quantum fields (and their Hermitian adjoints) and one can construct Lorentz scalars in much the same way as one constructs rotational scalars by the use of Clebsch–Gordon coefficients.

The causality condition on the Hamiltonian translates into a similar commutation relation (or anticommutation relation for fermions) for fields and this in turn imposes constraints on the coefficients, u_l and v_l. In the course of solving these constraints, one encounters the fundamental theorems of QFT:

Field equations: Every free field must satisfy the second-order differential wave equation, $(\partial^\mu \partial_\mu + m^2)\phi_l(x) = 0$ (Klein-Gordon equation), and many fields must obey other first-order differential equations or algebraic constraint equations, as well.

Antiparticles: Antiparticles (Dirac, 1928) exist for any particle species (not just for spin-1/2 fermions as in the Dirac formalism) with an additive conserved quantum number (e.g. electric charge, Q).

---

[1] In contrast to the massive case, the commutator algebra of symmetry generators (Lie algebra) does not restrict the possible (real) eigenvalues of J_3, but there is a topological restriction allowing only half-integer of integer values of h.

[2] $p^0 = \sqrt{m^2 + \vec{p}^2}$ is the particle energy; x_0 appearing in $px \equiv p^0x_0 - \vec{p}\vec{x}$ denotes time multiplied by the speed of light.

Spin-statistics connection: Fields describing particles with (half)-integer spin are (fermions) bosons.

CPT theorem: Any QFT is invariant under the combined operations of charge (particle–antiparticle) conjugation (C), parity (space) reflexion (P), and time reversal (T).

4.2.2
Local Gauge Symmetries

In some cases it is impossible to find any solution to the constraints on u_l and v_l. In particular, no four-vector field, A^μ, can be constructed from the annihilation and creation operators for a particle of helicity ± 1 and $m = 0$ (Weinberg, 1965). Of course, one might use some other Lorentz representation, such as an antisymmetric tensor field, $B^{\mu\nu}$, which does allow a solution. However, the interactions in such a theory have a faster falloff at large distances than the inverse-square law of the electrostatic coulomb force. Such interactions may well exist in nature, but they would be strongly suppressed relative to inverse-square law interactions. But while it may suffice to have a covariantly transforming four-vector field, A^μ, to construct a Lorentz-invariant theory, this is not actually necessary. Indeed, under a general homogeneous Lorentz transformation, Λ, one finds

$$U(\Lambda)A^\mu(x)U^{-1}(\Lambda) = \Lambda^{\nu\mu}A_\nu(x) + \partial^\mu\Omega(x) \tag{2}$$

That is, the A^μ in this case transforms covariantly up to an extra term, where $\Omega(x)$ is a linear combination of $a^\dagger$ and a operators. Therefore, if for arbitrary functions, $\omega(x)$, we require invariance under local (space-time-dependent) ‘‘gauge’’ transformations of the form

$$A^\mu(x) \longrightarrow A^\mu(x) - \partial^\mu\omega(x) \tag{3}$$

then the failure of A^μ to transform covariantly would be immaterial. Thus, local gauge invariance provides us with an additional option to construct long-range forces mediated by particles of $h = \pm 1$ and $m = 0$, and (if present) we should expect gauge-invariant theories to play a dominant role at large distances. This is precisely what is observed and the electromagnetic, strong, and weak forces, are all gauge theories invariant under the replacement (3).

This discussion generalizes straightforwardly to the case of gravity. No symmetric rank-two tensor field, $h^{\mu\nu}$, can be constructed from the annihilation and creation operators for a particle of $h = \pm 2$ and $m = 0$ (Weinberg, 1965). To construct an inverse-square law (Newton's law) one still needs to work with $h^{\mu\nu}$, where the failure to transform covariantly is cured by requiring invariance under general coordinate transformations as in the General Theory of Relativity (Einstein, 1916) generalizing the local gauge transformation (3).

The mathematics of all this generalizes further to massless particles of higher Helicities, $h > 2$. Now, Noether's theorem implies that conserved quantities must be associated with these transformations, which are electric charge (and generalizations thereof) for gauge theories and energy-momentum conservation in the case of gravity. There would be associated conserved quantities for the case $h > 2$, but the abundance of conservation laws would rule out the possibility of nontrivial scattering processes of any kind. Thus, the physics of this does not generalize beyond gravity. It is not yet known whether the only other possible case, $h = \pm 3/2$, is realized in nature.[3] If it is, the corresponding transformations generalizing the local gauge transformation (3) are the local "supersymmetry" transformations. Supersymmetry (Golfand and Likhtman, 1971; Wess and Zumino, 1974) is an active subfield of elementary particle physics and among the leading candidates of possible physics that may modify the SM at higher energies.

The historical development was, of course, rather different. The theory of electromagnetism (Maxwell, 1865) was not constructed from Lorentz and local gauge invariance but rather gave rise to these concepts (Lorentz, 1909).

4.2.3
Gauge Group of Electroweak Interaction

The gauge symmetry must and does (Fock, 1926) persist after the field A^μ is coupled to matter. In quantum electrodynamics (QED), one therefore extends the gauge transformation (3) to electrons and positrons to undergo the local $U(1)$ (phase) transformation (e is the electromagnetic coupling constant)

$$\Psi(x) \longrightarrow e^{\pm ieQ\omega(x)}\Psi(x) \tag{4}$$

In the minimal electroweak theory, one assumes that the electron combines with the electron-neutrino to form a doublet, extending the Abelian $U(1)$ gauge symmetry of QED to some (in general) non-Abelian symmetry group. In non-Abelian gauge theories, particles can be transformed into other kinds of particle, making them suitable candidates for describing the weak interaction that can transform, for example, electrons into neutrinos, or neutrons into protons. Positivity of quantum mechanical probabilities implies that the associated Lie algebra is the direct sum of commuting compact simple and $U(1)$ subalgebras. These have been classified[4] by Cartan (1894), leaving only a few possibilities.

To see what these are, a unique feature of the weak interaction (not shared by the other known forces) must now be mentioned, namely that electroweak gauge transformations act distinctly on left- and right-handed particles (chirality). This possibility arises if one postulates invariance under proper Lorentz transformations only (i.e., excluding parity reflections) with respect to which left- and right-handed particles form separate irreps. There is presently no piece of experimental evidence that would incontrovertibly require the existence of a right-handed neutrino and it is excluded from the minimal theory. And since right-handed particles are not independent from left-handed antiparticles, we may restrict our attention to the left-handed triplet of fields, $(\nu_e, e^-, e^+)_L$, with the understanding that the physics of right-handed particles and antiparticles is determined in terms of the left-handed ones.

Now the largest possible symmetry group that may act on this triplet is $U(3) = SU(3) \times U(1)$, but such a theory would imply lepton number violation at rates that are in conflict with observation. The largest possible subgroup of $U(3)$ consistent with lepton number is $U(2)_L \times U(1)_R = SU(2)_L \times U(1)_L \times U(1)_R$ with $SU(2)_L$ (isospin) acting nontrivially on the doublet of fields, (ν_e, e^-). The diagonal linear combination of $U(1)_L$ and $U(1)_R$ would give rise to a long-range force

---

[3] The cases $h = 0$ and $h = \pm 1/2$ do not present the kind of complications discussed here.

[4] For an extensive discussion with emphasis on the application to particle physics, see Slansky (1981).

(unless spontaneously broken) coupled to lepton number, but experiments provide no evidence in its favor and it is dropped. The gauge group for this leptonic model (Weinberg, 1967; Salam, 1968) is then

$$G = SU(2)_L \times U(1)_Y \tag{5}$$

with the remaining combination, $U(1)_Y$, coupling to "hypercharge," Y. The hypercharge assignment must reproduce the observed electric charges, so that (in an appropriate normalization)

$$Y \equiv Q - T_3 \tag{6}$$

where T_3 generates the $U(1)$ subgroup of $SU(2)_L$. T_3 must be traceless for each irrep, and one can take $T_3 = \mathrm{diag}(1/2, -1/2)$ ($T_3 = 0$) and consequently $Y = -1/2$ ($Y = +1$) when acting on the doublet (singlet).

4.2.4
Gauge Bosons

Local gauge invariance requires vector bosons to transform in the defining (adjoint) representation of the gauge group, that is, there must be precisely one gauge boson for each group generator. Thus, there is an isotriplet, $\vec{W}_\mu$, transforming under $SU(2)_L$, and an isosinglet, B_μ, for hypercharge. The linear combinations with definite mass and electric charge (shown as superscript) are

$$W^\pm_\mu = \frac{1}{\sqrt{2}}(W^1_\mu \mp iW^2_\mu) \tag{7}$$

$$Z^0_\mu = \cos\theta_W W^3_\mu - \sin\theta_W B_\mu \tag{8}$$

$$A_\mu = \sin\theta_W W^3_\mu + \cos\theta_W B_\mu \tag{9}$$

where the weak mixing angle, θ_W, is defined so that A_μ can be identified with the photon, that is, it couples to Q and describes the electromagnetic current. Similarly, the $W^\pm$ bosons produce charged current processes, such as observed in radioactive β decays, and the doublet–singlet structure of the SM reproduces the vector minus axial-vector ($V - A$) law of the weak interaction. On the other hand, at the time when the SM was under construction (Weinberg, 1967; Salam, 1968) no process indicating the presence of the Z^0 boson was known, and the associated neutral current is a genuine prediction of the model. Denoting the gauge coupling constants for $SU(2)_L$ and $U(1)_Y$, respectively, by g and g', one finds the relations

$$e = g\sin\theta_W = g'\cos\theta_W = \frac{gg'}{\sqrt{g^2 + g'^2}} \tag{10}$$

An important property of non-Abelian gauge theories (Yang and Mills, 1954) (not shared by Abelian ones) is the appearance of nonlinear terms in the Lagrangian density

$$\mathcal{L} = -\frac{1}{4}\left(\partial_\mu \vec{W}_\nu - \partial_\nu \vec{W}_\mu - g\vec{W}_\mu \times \vec{W}_\nu\right)^2 - \frac{1}{4}\left(\partial_\mu B_\nu - \partial_\nu B_\mu\right)^2 \tag{11}$$

This implies the existence and precise form and strength of triple and quartic gauge boson self-interaction vertices.

The gauge-invariant interactions of vector bosons with leptons (and other spin-1/2 and spin-0 matter particles) are implemented by replacing the ordinary derivatives acting on the matter fields by gauge-covariant derivatives

$$D_\mu = \partial_\mu + ig\vec{T}\vec{W}_\mu + ig'YB_\mu \tag{12}$$

4.2.5
Spontaneous Symmetry Breaking

If the short-range weak force is to be described by the $W^\pm$ and Z^0 bosons, these would have to be massive. Simply adding a mass term to the Lagrangian (11) would break local gauge invariance, which (as discussed) is unacceptable. This is a very severe problem, which touches upon the basic principles of QFTs as reviewed in Section 4.2.1. Indeed, a massless vector boson has only one degree of freedom (dof) which is characterized by its helicity, h, and which cannot be altered by a proper Lorentz transformation.[5] On the other hand, a massive vector boson is characterized by the three spin states of an ordinary spatial vector.

The solution to this problem involves the phenomenon usually called spontaneous symmetry breaking (SSB), although the symmetry of the Lagrangian remains fully intact. But the lowest energy (vacuum) state of the theory is degenerate, with various possible ground states related by the symmetry of the Lagrangian. Whichever is the actual vacuum state and whatever determines it, the physical consequences are indistinguishable because of the symmetry but the latter may be obscured or entirely hidden.

The SSB idea applied to *continuous, global* symmetries in QFTs (i.e., excluding local gauge symmetries) reveals the existence of a massless spin-0 field (Nambu–Goldstone boson) for each broken symmetry generator (Nambu, 1960; Goldstone, 1961). This can be proved (Goldstone, Salam and Weinberg, 1962) in general (Goldstone theorem).

For example, consider a doublet of complex scalar fields

$$\Phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1(x) + i\phi_2(x) \\ \phi_3(x) + i\phi_4(x) \end{pmatrix} \tag{13}$$

with $\phi_i = \phi_i^\dagger$ and with Lagrangian

$$\begin{aligned} \mathcal{L}_\Phi &= \partial_\mu \Phi^\dagger \partial^\mu \Phi - m_\Phi^2 \Phi^\dagger \Phi - \tfrac{1}{2}\lambda^2 (\Phi^\dagger \Phi)^2 \\ &= \tfrac{1}{2}\partial_\mu \phi_i \partial^\mu \phi_i - \tfrac{1}{2} m_\Phi^2 \phi_i \phi_i - \tfrac{1}{8}\lambda^2 (\phi_i \phi_i)^2 \end{aligned} \tag{14}$$

where here and in the following summation over i is understood. $\mathcal{L}_\Phi$ is manifestly invariant under continuous $SO(4) = SU(2) \times SU(2)$ rotations in the space spanned by the ϕ_i. The minimum energy state receives no contribution from the kinetic energy (the first term in $\mathcal{L}_\Phi$). For $m_\Phi^2 > 0$, the minimum of the potential, V (the negative of the last two terms in $\mathcal{L}_\Phi$), is reached for vacuum expectation values (VEVs) $\langle \phi_i \rangle = 0$, which define a unique and trivially $SO(4)$ symmetric vacuum. For $m_\Phi^2 < 0$, however, the ground state condition is

$$|\langle \Phi \rangle| = \sqrt{-\frac{m_\Phi^2}{\lambda^2}} \equiv \frac{v}{\sqrt{2}} \tag{15}$$

which has many solutions and none of them possesses $SO(4)$ symmetry.[6] We may choose, $\langle \phi_i \rangle = 0$ for $i = 1, 2, 4$ and $\langle \phi_3 \rangle = v$, or we define a new field variable, $\eta \equiv \phi_3 - v$ so that $\langle \eta \rangle = 0$, and write instead

$$\Phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} r\phi_1(x) + i\phi_2(x) \\ v + \eta(x) + i\phi_4(x) \end{pmatrix} \tag{16}$$

---

[5] To reach the second dof characterized by $-h$, one needs to invoke a parity transformation.

[6] One may object that there are nonzero matrix elements of the Hamiltonian, H, between different such individual vacuum solutions and that the linear superpositions that one must take to diagonalize H will then be symmetry eigenstates (including the lowest energy state). But for macroscopic systems these off-diagonal matrix elements are much smaller (and in infinite volume theories like QFTs they even vanish) than any imperceptibly small symmetry violating perturbation, which would diagonalize H in the individual basis. Moreover, vacuum states should also satisfy the principle of cluster decomposition (see Section 4.2.1) and one can show that general linear combinations do not.

The potential now reads

$$V = -\frac{\lambda^2 v^4}{8} + \frac{\lambda^2 v^2}{2}\eta^2 + \frac{\lambda^2 v}{2}\eta\phi_i\phi_i + \frac{\lambda^2}{8}(\phi_i\phi_i)^2 \tag{17}$$

showing a mass term for η ($m_\eta = \lambda v$), while the ϕ_i are massless in accord with the Goldstone theorem.

4.2.6
Higgs Mechanism

If the field (16) is assumed to transform nontrivially under $SU(2) \times U(1)$ gauge transformations, one can arrange for a particular transformation U to pass to a gauge (unitary gauge) such that

$$U\Phi(x) = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \tag{18}$$

that is, $\phi_1(x)$, $\phi_2(x)$, and $\phi_4(x)$ can be eliminated. If we apply this to the $SU(2)_L \times U(1)_Y$ electroweak gauge theory, we must replace the derivatives in Eq. (14) by the covariant derivatives (12), which introduces extra gauge boson dependent terms. The bilinear ones of these are

$$\Delta\mathcal{L} = -\frac{1}{8}v^2\left[g^2\left(W_\mu^1\right)^2 + g^2\left(W_\mu^2\right)^2 + \left(g' B_\mu - g W_\mu^3\right)^2\right] \tag{19}$$

showing that, as anticipated, the $W_\mu^\pm$ Eq. (7) and Z_μ^0 Eq. (8) bosons are indeed mass eigenstates, with mass, respectively, given by

$$M_W = \frac{1}{2}gv, \quad M_Z = \frac{1}{2}\sqrt{g^2 + g'^2}\,v \tag{20}$$

while the photon (9) remains exactly massless. This shows how the puzzling mismatch of d.o.f. counting mentioned at the beginning of Section 4.2.5 is rectified: the extra (longitudinal) dof carried by a massive gauge boson is provided by a scalar dof of what in the context of the breaking of the corresponding *global* symmetry would have been a Goldstone boson (would-be Goldstone boson). This is the Higgs mechanism (Englert and Brout, 1964; Higgs, 1964; Guralnik, Hagen and Kibble, 1964) and the dof represented by η is the yet to be discovered Higgs boson, H, with mass, $M_H = \lambda v$.

The covariant derivatives will also induce trilinear and quadrilinear terms producing couplings of two gauge bosons to respectively, one and two Higgs bosons. These are fixed in terms of the gauge couplings g and g'.

We have chosen an isodoublet of complex scalars, but the Higgs mechanism would also work with other nontrivial representations. But if in the course of SSB the electron is to acquire a mass term, as well, the only renormalizable possibility (see Section 4.2.7) is indeed a nonderivatively coupled $SU(2)_L$ doublet (which can combine with the lepton doublet to form a singlet) with hypercharge $Y = -1/2$ (to cancel the hypercharges of the electron and the positron). This allows to add to the theory, a Yukawa term (Yukawa, 1935)

$$\mathcal{L}_Y = -\sqrt{2}\lambda_e \overline{(\nu_e, e)}_L \Phi e_R + \text{H.c.} \tag{21}$$

where the overline notation is introduced to form a Lorentz-invariant scalar product for fermions and adding the Hermitian conjugate assures a real valued Lagrangian. If again the Higgs field, $\Phi(x)$, is replaced by its VEV (15) a mass term for the electron arises,

$$m_e = \lambda_e v \tag{22}$$

4.2.7
Infinities

Quantum mechanical amplitudes for physical processes can be computed in an expansion in powers of the coupling constants (perturbation theory) and are given by a sum of Feynman diagrams (Feynman, 1949) translating into concrete mathematical expressions. If one attempts to calculate higher order radiative corrections to these processes (containing closed particle loops) one frequently encounters divergent expressions. One can ''renormalize'' a theory if the infinities match a set of infinite counterterms that one may add to the Lagrangian density defining the theory. In this case, one can absorb all infinities into redefined coupling constants, masses, and fields, provided one works with the most general Lagrangian and the complete set of counterterms allowed by the assumed symmetries.

The number of counterterms may in principle be infinite. There is a stricter sense of renormalizability, which refers to field theories that need only a finite number of counterterms and which (for historical reasons) are the ones actually called *renormalizable*. The coefficients of the counterterms in this case are restricted to be either dimensionless[7] or their dimension is a positive power of mass (power counting renormalizability). In gauge theories one still needs to demonstrate that gauge invariance constrains the infinities in the same way as the interactions. This has been achieved first for QED (Dyson, 1949) and later for non-Abelian theories ('t Hooft and Veltman, 1972) including those with spontaneously broken gauge symmetries (Lee and Zinn-Justin, 1972) showing that the SM is renormalizable in the stricter sense.

There is another class of infinities (Adler, 1969; Bell and Jackiw, 1969) that can arise and for which there are no counterterms to absorb them. They appear in Feynman diagrams containing a closed spin-1/2 fermion triangle with gauge bosons attached (possibly including gravitons) of which at least one couples chirally – just like the $SU_L \times U(1)_Y$ gauge bosons in the SM. Such diagrams can be interpreted as corrections to the three-gauge-boson vertices. Since there are no corresponding divergences in the corrections to the four-gauge-boson vertices or to the vertices generated by the covariant derivatives, the subtle structure required by gauge invariance (e.g., in Eq. (11)) breaks down. This phenomenon of symmetry violation by quantum effects (anomaly) does not necessarily spoil the mathematical consistency of the theory (see Section 4.5.3 for an example), but *chiral gauge anomalies* of the type encountered here do so.

It is perhaps surprising that symmetries of the action can be violated by the interactions derived from this very same action. Fujikawa (1979) has given an interpretation of chiral anomalies, which reveals their fundamental nature. In the space-time approach (Feynman, 1948) to quantum mechanics, amplitudes are given as integrals over all possible path histories (path integral formulation) weighted by a complex phase factor given in terms of the action. The chiral anomaly is here seen as the Jacobian determinant (which can be written as an extra phase factor) arising from a change of integration variables of the form of a local gauge transformation. The nontrivial Jacobian by itself would not affect the perturbation theory because it is field-independent and would pose no problem. But it is infinite and needs regularization in the process of which (gauge) field-dependence is introduced. For this reason, in addition to a gauge-invariant action, one also has to require a gauge-invariant path integral measure (more strictly, only the combination of the two needs to be gauge-invariant).

---

[7] As is customary among elementary particle physicists, we are using physical units in which, by definition, the fundamental values of the speed of light, c, and of Planck's constant, $h/(2\pi)$, are exactly equal to one. The dimensions of mass, energy, linear momentum, and inverse length are then the same.

The model of leptons (Weinberg, 1967; Salam, 1968) described in this section is plagued by chiral gauge anomalies and is therefore – as it stands – inconsistent. However, one may add additional fermionic dofs to the model that may precisely cancel the anomalies. These dofs are provided by the quarks discussed in Section 4.3.2.

4.3
The Standard Theory

Section 4.3.1 extends the model to include the muon and tau sectors whose existence is not explained in the SM but which incidentally help to determine the basic coupling parameters, respectively, of the weak and strong interactions, with great precision. The quark sector and the modern theory of the strong interaction (quantum chromodynamics (QCD)) are added in Section 4.3.2, completing the particle content of the SM. Sections 4.3.3 and 4.3.4 are dedicated to quark mixing and CP violation, respectively.

4.3.1
Lepton Replication and Muon Decay

Besides Yukawa interactions as, for example, in $\mathcal{L}_Y$, the leptonic model has only four adjustable parameters, namely, the gauge couplings, g and g', as well as from the Higgs Lagrangian (14) the Higgs-self coupling, λ, and the only fundamental parameter with dimension of mass in this theory, m_Φ. The combination of the latter two in Eq. (15) can be determined experimentally with very good accuracy. This is because of the existence of the muon, which forms (with its antiparticle and the left-handed muon-neutrino, ν_μ) the same left-handed $SU(2)_L$ doublet and right-handed singlet structure with the same hypercharges as the electron, providing an exact duplication except for an approximately 200 times larger value for the corresponding Yukawa coupling (21), λ_μ as compared to λ_e. Thus, the μ^- is much heavier and can decay into a ν_μ, e^-, and $\overline{\nu}_e$ with an amplitude[8], which to good approximation is proportional to g^2/M_W^2. According to Eq. (20), this determines the Higgs VEV,

$$v = \frac{2M_W}{g} = \sqrt{\frac{1}{\sqrt{2}G_F}} = 246.22\ \text{GeV} \tag{23}$$

where $G_F = 1.166367 \pm 0.000005 \times 10^{-5}$ GeV^{-2} (4 parts per million precision), is the four-fermion coupling constant appearing in the original Fermi theory (Fermi, 1934) of the weak interaction.

It should be mentioned that the extraction of fundamental parameters to such high precision requires the calculation of higher order corrections. These are often difficult and lengthy computations but are feasible within renormalizable field theories. For example, in the case of the Fermi constant, one needs the muon lifetime formula to fourth-order (two-loop) precision (Marciano and Sirlin, 1988; van Ritbergen and Stuart, 2000).

---

[8] The Lagrangian (21) and the gauge interactions are such that the total number of e^- and ν_e minus e^+ and $\overline{\nu}_e$ entering and leaving any Feynman diagram must be the same (electron number conservation) and likewise for muons. This is why an antineutrino is emitted together with an electron.

The relations (10) allow to fix another combination of couplings since the QED fine structure constant

$$\alpha \equiv \frac{e^2}{4\pi} = \frac{1}{137.0359997 \pm 0.0000001} \tag{24}$$

is known to within 700 parts per trillion. It is extracted from the anomalous magnetic moment of the electron (see Section 4.4.2) and is based on an eighth-order (four-loop) calculation (Gabrielse *et al.*, 2006).

The gauge boson masses (20) are then (to first order) determined by the weak mixing angle alone,

$$\begin{aligned} M_W &= \frac{\sqrt{4\pi\alpha(M_Z)}}{2\sin\theta_W} v \approx \frac{38.59\ \text{GeV}}{\sin\theta_W} \\ &\approx 86 \pm 7\ \text{GeV} \qquad (25) \\ M_Z &= \frac{\sqrt{4\pi\alpha(M_Z)}}{2\sin\theta_W\cos\theta_W} v \approx \frac{77.18\ \text{GeV}}{\sin 2\theta_W} \\ &\approx 96 \pm 6\ \text{GeV} \qquad (26) \end{aligned}$$

These equations use the value, $\alpha^{-1}(M_Z) \approx 127.9$, appropriate for high-energy (weak-scale) observables instead of the low energy value in Eq. (24). This is because leading radiative corrections – those enhanced by large logarithms of the form $\ln M_Z/m_f$, where m_f is a fermion mass – have been taken into account. Effectively, these logarithms lead to the concept of the energy scale (μ) dependence[9] (running) of the coupling constants (Stueckelberg and Petermann, 1953; Gell-Mann and Low, 1954), described by the Callan–Symanzik β-function (Symanzik, 1970; Callan, 1970)

$$\mu^2 \frac{d}{d\mu^2}\alpha(\mu) = \beta(\mu) \tag{27}$$

Note, that α grows with μ (screening), and so $\beta > 0$. The last step in the relations (25) and (26) shows the predictions for the value $\sin^2\theta_W = 0.20 \pm 0.03$ obtained by the 1978 experiment (Prescott *et al.*, 1978) in electron–deuteron fixed-target scattering (see Section 4.4.2). The $W^\pm$ (Arnison *et al.*, 1983a, Banner *et al.*, 1983) and Z^0 (Arnison *et al.*, 1983b, Bagnaia *et al.*, 1983) bosons were discovered in 1983 in proton–antiproton collisions at the Super Proton Synchrotron (SPS) at the European Organization for Nuclear Research (CERN) near Geneva, Switzerland. The reconstructed masses, $M_W \sim 80\,\text{GeV}$ and $M_Z \sim 92\,\text{GeV}$, were in perfect agreement with the predictions (25) and (26).

The most general amplitude for muon decay (Michel, 1950) is described by 19 real parameters (Michel parameters), which reduce to 6 parameters if the outgoing neutrinos are not observed. The Michel parameters have been used to confirm the $V-A$ structure of the SM. The muon sector also played a crucial role in the first observation of a neutral current process. The 1973 discovery (Hasert *et al.*, 1973) at CERN was based on a single event in $\nu_\mu e^-$ elastic scattering.

The lepton sector is completed by yet another replication, composed of the $\tau^\pm$ and the ν_τ. The $\tau^\pm$ is about 17 times heavier than the $\mu^\pm$, and is the only known lepton heavy enough to decay into hadrons, providing a valuable laboratory for detailed and quantitative studies of the strong interaction (Davier, Hocker and Zhang, 2006).

---

[9] There are various definitions arising from different calculational (renormalization) schemes, which differ by higher order terms. Values quoted here correspond to what is known as the *modified minimal subtraction* ($\overline{\text{MS}}$) scheme ('t Hooft, 1973).

4.3.2
Quarks and Quantum Chromodynamics

It was stressed in Section 4.2.7 that the leptonic model by itself has a gauge anomaly, which can be cured by adding additional fermions (quarks). As can be seen from Eq. (6), a left-handed quark doublet (u, d) with hypercharge $Y = +1/6$ and two antiquark singlets $\overline{u}$ ($Y = -2/3$) and $\overline{d}$ ($Y = +1/3$) correspond to $Q_u = +2/3$ and $Q_d = -1/3$. These give a contribution to the pure (cubic) hypercharge anomaly of $Y^3 = N_C(2 - 64 + 8)/216 = -N_C/4$. If one chooses $N_C = 3$ (the number of "colors") this cancels the contribution, $Y^3 = -1/4 + 1 = 3/4$, from electron-type leptons. Doublets with other kinds of quarks (flavors) (c, s) and (t, b) along with their antiquarks are needed to cancel the anomaly contributions from the muon and tau sectors. Anomaly cancellation for other gauge boson combinations can also be checked.

Neither the color quantum number nor the free quarks have ever been observed directly. Qualitatively this can be understood if one assumes that there is another non-Abelian gauge interaction with gauge group $SU(3)_C$, which is acting on color triplets of quarks and color antitriplets of antiquarks[10], and if this new gauge interaction became stronger at large distances[11] so that it would be too costly in energy to isolate a quark (confinement hypothesis). For this, the β-function of QCD would have to be negative (antiscreening) in contrast to the case of QED (see Section 4.3.1) and any other renormalizable QFT. Non-Abelian gauge theories (if not coupled to too many matter fields) have indeed this property (Gross and Wilczek, 1973a; Politzer, 1973).

The reason why the $SU(3)_C$ gauge bosons (the gluons) have not been observed directly is not because they are very massive, but rather because they are massless and confined into colorless hadrons (Weinberg, 1973; Gross and Wilczek, 1973b). The known hadrons are quark–antiquark bound states (mesons) or bound states of either three quarks or three antiquarks (baryons and antibaryons). Other possibilities such as bound states of four quarks and an antiquark (pentaquarks) or purely gluonic states (glueballs) have been speculated about and in some cases their mass predicted, but they have not yet been uncontroversially observed. Both gluons and quarks can be indirectly observed in high-energy collisions as directionally clustered collections of hadrons (jets). For example, the gluon was discovered (Berger *et al.*, 1979; Bartel *et al.*, 1980) in 1979 at the Positron Electron Tandem (PETRA) Ring Accelerator (PETRA) at the German Electron Synchrotron (DESY) in Hamburg as planar three jet events, which could be interpreted as the radiation of a gluon by one of the quarks (gluon Bremsstrahlung). This completes the elementary particle spectrum of the SM, which is summarized in Table 4.1.

By precisely measuring the total production rate of jet events one can not only determine N_C (to which it is proportional to first order) but also

$$\alpha_s \equiv \frac{g_s^2}{4\pi} \qquad (28)$$

---

[10] Triplets and antitriplets of $SU(3)$ are complex conjugates of each other and form mathematically inequivalent irreps. The only other gauge group in Cartan's classification (Cartan, 1894) with a three-dimensional irrep (for $N_C = 3$) is $SU(2)$, but its irreps are (pseudo)real and therefore would give a hadron spectrum different from what is observed.

[11] Conversely, it ought to become weaker at short distances (asymptotic freedom) in agreement with experiments in deep inelastic electron–nucleon scattering (Bloom *et al.*, 1969).

Tab. 4.1 Standard model particles.

Multiplet			**Spin**	**SU(3)$_C$**	**SU(2)$_L$**	**U(1)$_Y$**	
Higgs			**0**	**1**	**2**	$\frac{-1}{2}$	
$(\nu_e, e^-)_L$	$(\nu_\mu, \mu^-)_L$	$(\nu_\tau, \tau^-)_L$	$\frac{1}{2}$	1	2	$\frac{-1}{2}$	
e_R^-	μ_R^-	τ_R^-	$\frac{1}{2}$	1	1	-1	Plus
$(u, d)_L$	$(c, s)_L$	$(t, b)_L$	$\frac{1}{2}$	3	2	$\frac{+1}{6}$	anti
u_R	c_R	t_R	$\frac{1}{2}$	3	1	$\frac{+2}{3}$	particles
d_R	s_R	b_R	$\frac{1}{2}$	3	1	$\frac{-1}{3}$	
Gluons							
			1	8	1	0	
$\vec{W}$			1	1	3	0	
B			1	1	1	0	

where g_s is the strong gauge coupling associated with $SU(3)_C$ and which enters at second order. Examples for such measurements include hadron production in e^+e^- annihilation at sufficiently high energy, the τ lepton lifetime, and various Z^0 decay observables (see Section 4.4.1).

QCD by itself individually conserves the number of quarks of each flavor (minus the number of each antiflavor). Moreover, QCD perturbation theory also conserves P and C invariance (just like QED), but (unlike in QED) these are generally expected to be broken non-perturbatively (see Section 4.3.4).

4.3.3
Quark Mixing

When one adds to the u (up) and d (down) quarks of the first family the second generation s (strange) and c (charm) quarks, then the Yukawa couplings (21) become 2×2 matrices. These complex matrices can be brought to diagonal form with real eigenvalues (the quark masses) by means of bi-unitary[12] transformations (changes of basis). The weak interaction, however, is not diagonal in this new (mass) basis whenever the left-handed unitary transformation, U_L^u, acting on up and charm quarks differs from U_L^d, acting on down and strange quarks. As a result, the linear combinations

$$\begin{aligned} d' &= \cos\theta_c d + \sin\theta_c s \\ s' &= -\sin\theta_c d + \cos\theta_c s \end{aligned} \tag{29}$$

of mass eigenstates d and s are the ones that actually form the $SU(2)_L$ doublets with u and c, respectively. Here, θ_c is the so-called Cabibbo angle (Cabibbo, 1963), and the right-handed transformation matrices, U_R^u and U_R^d, have been used to ensure that $u' = u$ and $c' = c$ remain unchanged. In the SM,

---

[12] In general, different matrices are necessary for the left- and right-handed quark fields. Each matrix must be unitary, $U^\dagger = U^{-1}$, so that the quark kinetic energy terms maintain their canonical forms.

U_R^u and U_R^d have no observable effect, which can be traced to the fact that there are only left-handed $SU(2)_L$ doublets. Thus, in the quark sector there is a numerically smaller Fermi constant, $G_F \to G_F \cos\theta_c$, more effective than in the lepton sector, with $\theta_c \approx 13^\circ$. Flavor changing charged current transitions between the first two families such as kaon (light mesons containing an s quark) β decays, $K^- \to \pi^0 e^- \overline{\nu}_e$, are possible but suppressed by $\sin^2\theta_c \approx 0.05$.

Flavor changing neutral current (FCNC) transitions, such as K^0–$\overline{K^0}$ oscillations or $K^0 \to \mu^+ + \mu^-$ decays, are also predicted. However, the contributions from two full generations of quarks tend to cancel each other leaving only a small residual effect, which can mostly be attributed to the differences in quark masses. Historically, this cancellation (Glashow-Iliopoulos-Maiani (GIM)mechanism) was the basis for the prediction (Glashow, Iliopoulos and Maiani, 1970) of the c quark and an estimate of its mass before the 1974 discovery of the first c–$\bar{c}$ bound state (the J/Ψ) almost simultaneously at the Alternating Gradient Synchrotron (AGS) at the Brookhaven National Laboratory (BNL) (Aubert *et al.*, 1974) and at the Stanford Positron Electron Accelerating Ring (SPEAR) at the Stanford Linear Accelerator Center (SLAC) (Augustin *et al.*, 1974).

4.3.4
CP Violation

Upon inclusion of the t (top) and b (bottom) quarks, 3×3 Yukawa (mass) matrices and the possibility of CP violation arise (Kobayashi and Maskawa, 1973). CP violation typically occurs in the presence of complex phases leading to different quantum mechanical interference effects between charge-conjugate amplitudes. Many phases in the Cabibbo–Kobayashi–Maskawa (CKM) matrix (Cabibbo, 1963; Kobayashi and Maskawa, 1973)

$$V_{\mathrm{CKM}} = U_L^u U_L^{d\dagger} \tag{30}$$

can be removed by redefinitions. Indeed, no diagonal phase transformation of the form $U_R = U_d = \mathrm{diag}(\alpha_1, \alpha_2, \alpha_3)$, affects the real masses (and corresponding quark-Higgs couplings) or anything else except for V_{CKM} itself. These can be used (with the exception of one redundant overall phase redefinition) to remove phases from V_{CKM} (in the three quark generation case a total of five phases can be eliminated). Being defined as the product (30) of two unitary matrices, V_{CKM} is itself unitary, and since it is unconstrained otherwise it has nine (real) parameters. Of these only four are then observable, which can be interpreted as three mixing angles (generalizing θ_c) and one observable (CP violating) phase, $e^{i\delta}$. The same counting reveals $(N-1)^2$ parameters for an N generation V_{CKM} of which $(N-1)(N-2)/2$ are complex phases. Therefore, as long as $e^{i\delta} \neq \pm 1$, the three generation SM predicts CP nonconservation in the electroweak sector. Combining the fact that CP violation occurs for three but not for two generations of quarks with the small observed values of V_{CKM} elements connecting the three generations with each other, one expects rather small CP violating effects even for δ not very close to 0 or π. CP violation has been observed in kaons and in mesons containing b-quarks. All results can be understood in terms of a common value of the CKM phase δ.

There is another source of CP (and P) violation from the QCD sector. A term may be added to the QCD Lagrangian, which in some sense extends the gauge coupling to complex values, with an imaginary part proportional to a real parameter θ (QCD θ angle). It can be written as a total derivative, which renders it harmless in QCD perturbation theory because it does not modify the Feynman rules[13]. Even nonperturbatively it may be absorbed by a chiral phase redefinition of the quark fields but at the expense of having to allow complex quark masses, modifying the phase counting above. The QCD θ angle would give, for example, a contribution to the CP violating electric dipole moment of the neutron, which has not yet been observed. This implies a stringent bound, $\theta \lesssim 10^{-10}$. It is an open question why θ is so small (strong CP problem).

4.4
Experimental Tests

After the discoveries of the weak neutral current (Hasert *et al.*, 1973), the $W^{\pm}$ (Arnison *et al.*, 1983a; Banner *et al.*, 1983) and Z^0 (Arnison *et al.*, 1983b; Bagnaia *et al.*, 1983) bosons, and the observation (Prescott *et al.*, 1978) of parity violation in deep inelastic electron–deuteron scattering (see Section 4.4.2), the SM was widely accepted as the correct theory of electroweak interactions, at least to some first approximation. High precision measurements were then needed to establish the SM as a renormalizable field theory and at the level of quantum corrections. Since the $SU(2)_L$ gauge coupling usually enters in the combination $g^2/4\pi^2 \approx 0.01$, one typically has to achieve relative accuracies of better than 1% in genuine electroweak observables.

4.4.1
Z^0 Pole Physics

Z^0 bosons can be produced copiously in e^+e^- annihilation experiments if the center of mass energy is tuned to M_Z (Z resonance). The high statistics achievable on the Z^0 pole was exploited by the Large Electron Positron (LEP) collider at CERN and the Stanford Linear Collider (SLC) at SLAC (Z factories). A joint document (ALEPH, DELPHI, L3, OPAL, SLD Collaborations, LEP Electroweak Working Group, SLD Electroweak and Heavy Flavour Groups, 2006) combines the results and accounts for common systematic and theoretical uncertainties in the interpretation of the individual experiments.

The LEP program recorded about 17 million Z^0 decays and included an energy scan around the Z^0 pole. The fit to the Z^0 resonance lineshape resulted in a 23 parts per million determination of $M_Z = 91.1876 \pm 0.0021$ GeV. Together with the fine structure constant (24) and the Fermi constant appearing in Eq. (23) this fixes three of the four fundamental electroweak input parameters mentioned in Section 4.3.1. Even though the Higgs boson has not been discovered yet, the missing input parameter, M_H, can be constrained by a global fit to all electroweak precision data (see Section 4.4.3). In addition to M_Z, the Z^0 lineshape scan also yielded 0.9 permille determinations of the total Z^0 decay width, $\Gamma_Z = 2.4952 \pm 0.0023$ GeV, and the resonance peak cross section for hadronic final states, $\sigma_{\text{had}} = 41.541 \pm 0.037$ nb. The cross sections for specific quark or lepton flavors, f, are then expressed relative to σ_{had} as ratios, R_f. The observables, Γ_Z, σ_{had}, and R_ℓ ($\ell = e, \mu, \tau$), serve as the theoretically cleanest determinations of α_s. If one subtracts the partial Z^0 decay widths into hadrons and charged leptons from Γ_Z

---

[13] The fact that a total derivative term in the Lagrangian can still have a nontrivial physical effect is best understood in the path integral formulation. One points to the existence of extended space-time-dependent (topological) field configurations (such as the so-called instantons), which one has to include (Belavin *et al.*, 1975).

one obtains the "invisible" Z^0 width, Γ_{inv}, which in the SM represents the decay width into neutrinos. Conversely, by computing the SM decay width into one standard neutrino, one can use the measured Γ_{inv} to extract the number of massless neutrinos, $N_\nu = 2.985 \pm 0.009$, which rules out further generations of fermions (unless their properties are qualitatively different from the three known ones). By measuring the cross-section asymmetry, $A_{FB}(f)$, for fermions f going into the forward hemisphere of the detectors relative to the backward direction (forward–backward asymmetry), the LEP Collaborations obtained a number of observables, which are proportional to lepton-Z vector couplings, v_ℓ. Since (to leading order) the v_ℓ are proportional to the combination $1 - 4\sin^2\theta_W$ and $\sin^2\theta_W \approx 0.23$ is numerically close to 1/4, these observables have an enhanced sensitivity to $\sin^2\theta_W$ and allow permille level determinations. This also applies to the final state τ polarization asymmetry, $A_{\text{pol}}(\tau)$, a measurement that included information from the angular distribution, as well.

The SLC operated with a 75% polarized e^- beam and percent level polarimetry permitting competitive measurements of $\sin^2\theta_W$ with a much smaller number of approximately 600 000 Z^0 bosons. This is because left-right (polarization) asymmetries (A_{LR}) are also proportional to v_ℓ, while being much larger than leptonic LEP asymmetries (yielding better data statistics). They are also cleaner (only counting of hadrons or leptons is required) than the quark asymmetries at LEP, which need to tag the flavors of heavy quarks and distinguish quarks from antiquarks. The combination of measurements of the weak mixing angle at Z^0 pole energies yields

$$\sin^2\theta_W(M_Z) = 0.23124 \pm 0.00017 \qquad (31)$$

The results obtained from the Z factories are in remarkable agreement with the SM predictions. In particular, the decay properties of the Z^0 boson and its couplings to leptons and quarks have been verified at the subpercent level tightly constraining any possible physics beyond the SM. On the other hand, direct contributions to the scattering amplitudes from outside the SM are suppressed relative to the resonating Z^0 amplitude and are best studied away from the Z^0 pole.

One possibility is to upgrade the colliders to higher energies as was done with LEP (LEP 2), which operated up to 209 GeV (ALEPH, DELPHI, L3, OPAL Collaborations, LEP Electroweak Working Group, 2006). Here one can study directly the electroweak energy scale, but high precision measurements are generally difficult owing to small data samples. Nevertheless, various key measurements were achieved, most notably the very precise measurement of $M_W = 80.376 \pm 0.033$ GeV. Similarly, the proton–antiproton collider Tevatron at the Fermi National Accelerator Laboratory (FNAL) (FNAL) near Chicago is producing collisions up to 2 TeV. It provided a comparable measurement of $M_W = 80.429 \pm 0.039$ GeV (Abazov *et al.*, 2004; Aaltonen *et al.*, 2007), and was also able to discover the very massive t quark (Abe *et al.*, 1995; Abachi *et al.*, 1995) and determine its mass, $m_t = 170.9 \pm 1.8$ GeV (Tevatron Electroweak Working Group, 2007). The Tevatron experiments are expected to increase their data samples by another order of magnitude before the final shutdown of the collider. But complementary measurements are also possible at much lower energies if sufficient precision can be achieved (see Section 4.4.2).

4.4.2
Low Energy Measurements

The 1978 deep inelastic scattering (DIS) experiment (Prescott *et al.*, 1978) of polarized electrons from deuterium at SLAC was crucial in establishing the SM. Rival models based on different gauge groups or with different $SU(2)_L \times U(1)_Y$ irrep assignments could not simultaneously explain contemporary ν-scattering results or were altogether inconsistent with the observed parity nonconservation in eD-DIS. At energy (momentum transfer) scales Q of about 1 GeV, the parity violating interference term between the photon and the Z^0 is suppressed by $Q^2/M_Z^2 \sim 10^{-4}$ relative to the parity conserving QED cross section. In other words, an asymmetry measurement of the order of 10^{-5} corresponds to a 10% determination of the Z^0 mediated amplitude.

The same idea has later been applied (Anthony *et al.*, 2005) to e^-e^- (Møller) scattering at even lower momentum transfer, $Q^2 = 0.026\,\text{GeV}^2$. Just like at the Z^0 pole the parity violating asymmetry, A_{PV}, is here further suppressed by v_ℓ yielding a very small $A_{PV} = (-1.31 \pm 0.17) \times 10^{-7}$. Owing to the enhanced sensitivity to the weak mixing angle, this yields $\sin^2\theta_W(Q) = 0.2397 \pm 0.0013$, the most precise value away from the Z resonance. The central value is higher than the one in Eq. (31) establishing that $\sin^2\theta_W$ is a running parameter (see Figure 4.1) just like the QED coupling α discussed in Section 4.3.1 and the QCD coupling α_s.

The interference between photon and Z^0 mediated amplitudes can also induce mixing between opposite parity states in atoms. Atomic parity violation (APV) has been seen in various heavy atoms where the effect is larger but it must be enhanced further. For example, it can be isolated as a small modulation of the level mixing induced by an external electric field (Stark mixing). The most precise measurement has been obtained by a group in Boulder, CO, in cesium (Wood *et al.*, 1997) and may be interpreted as a precise measurement of $\sin^2\theta_W$ (see Figure 4.1). However, the interpretation requires a solid understanding of the structure of many-electron atoms (Ginges and Flambaum, 2004). For cesium the uncertainties associated with atomic wave functions are relatively small and comparable to the experimental ones.

Since neutrinos are electrically neutral, they are directly sensitive to the weak interactions, but their cross sections are proportional to G_F^{-2} and are very small. DIS of neutrinos from nuclei have been measured to percent accuracies but here the interpretation of the results is hampered by theoretical uncertainties from the strong interaction. These can be reduced by considering ratios of neutral and charged current cross sections and by using nuclei with (approximately) equal numbers of protons and neutrons (isoscalars). If a high-intensity, high-energy antineutrino beam is also available, one can use the ratio of differences of ν and $\overline{\nu}$ cross sections (Paschos-Wolfenstein ratio), reducing the theoretical uncertainties even further (Paschos and Wolfenstein, 1973). This was possible by studying neutrinos produced at the Tevatron and resulted in the most precise determination (Zeller *et al.*, 2002) of $\sin^2\theta_W$ from ν-DIS dominating the world average shown in Figure 4.1.

By far the most precise observable testing the SM (as opposed to fixing one of the input parameters) is the anomalous magnetic moment[14] of the muon, a_μ. At

---

[14] The term anomalous is of historical origin and refers to quantum corrections (Schwinger, 1948) to the lowest order magnetic moment g of massive particles with spin.

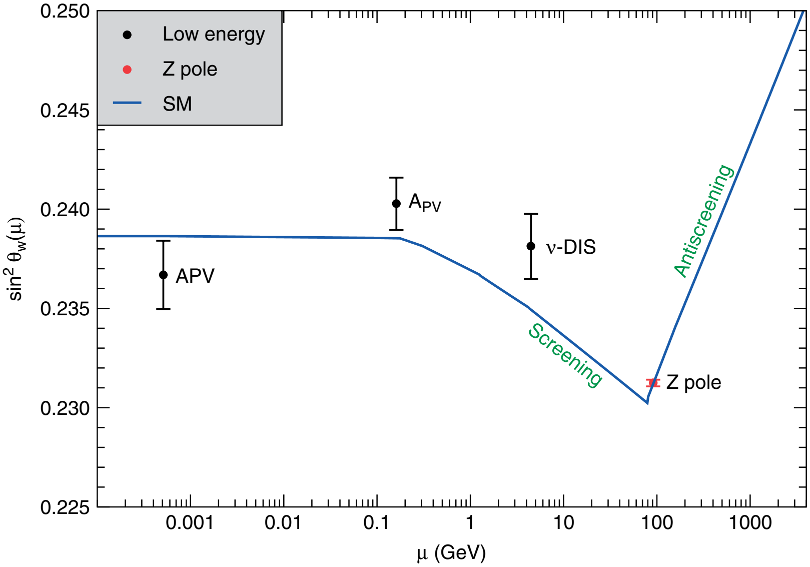

Fig. 4.1 The scale dependence of the weak mixing angle (Czarnecki and Marciano, 1998; Erler and Ramsey-Musolf, 2005) as predicted in the SM, compared to various low energy measurements and the Z^0 pole value (1.31). At $\mu = M_W$, the β-function of $\sin^2\theta_W$ changes sign because only with the inclusion of the $W^\pm$ bosons above M_W does the electroweak theory become a non-Abelian gauge theory with the characteristic antiscreening.

the level of the impressive experimental precision achieved at BNL (Bennett *et al.*, 2004),

$$a_\mu \equiv \frac{g_\mu - 2}{2} = (1165920.80 \pm 0.63) \times 10^{-9} \qquad (32)$$

receives nonnegligible corrections from the weak interaction and is easily affected by new physics contributions. The SM prediction (Miller, de Rafael and Roberts, 2007) of a_μ is not inconsistent with Eq. (32), but there is a tantalizing difference at the level of three standard deviations. Conclusions as to whether this deviation is from new physics at higher energy scales must await further theoretical work because the interpretation of a_μ is complicated by hadronic contributions entering at two-loop order (where it can be constrained by experimental data) and at three-loop order (which are much smaller but very difficult to estimate).

4.4.3
Global Analysis

To fully exploit the information contained in the various experimental results, it is best to perform a simultaneous analysis of all data (Amaldi *et al.*, 1987; Erler and Langacker, 1995; Yao *et al.*, 2006). Some of the measurements used for such a study are shown in Table 4.2.

The low energy data and other (less precise) observables[15] are not shown in the table but included. The column denoted

---

[15] For these and many other measurements and tests, including the determination of fermion masses, the parameters in V_{CKM}, and limits on SM forbidden processes, see the Review of

Tab. 4.2 Standard model precision tests.

Observable	**Experimental value**	**SM prediction**	**Pull**	**Deviation**
Γ_Z (GeV)	2.4952 ± 0.0023	2.4968 ± 0.0010	-0.7	-0.5
σ_{had} (nb)	41.541 ± 0.037	41.466 ± 0.009	2.0	2.0
R_ℓ	20.767 ± 0.025	20.758 ± 0.011	0.4	0.5
R_b	0.21629 ± 0.00066	0.21584 ± 0.00006	0.7	0.7
R_c	0.1721 ± 0.0030	0.17228 ± 0.00004	-0.1	-0.1
$A_{FB}(\ell)$	0.0171 ± 0.0010	0.01627 ± 0.00023	0.8	1.1
$A_{FB}(b)$	0.0992 ± 0.0016	0.1033 ± 0.0007	-2.5	-2.0
$A_{FB}(c)$	0.0707 ± 0.0035	0.0738 ± 0.0006	-0.9	-0.7
$A_{\text{pol}}(\tau)$	0.1465 ± 0.0033	0.1473 ± 0.0011	-0.2	0.1
A_{LR}	0.1514 ± 0.0022	0.1473 ± 0.0011	1.9	2.4
M_W (GeV)	80.398 ± 0.025	80.375 ± 0.015	1.4	1.7
m_t (GeV)	170.9 ± 1.8	171.1 ± 1.9	-0.1	-0.8

"Pull" shows the deviation (normalized to the total experimental plus theoretical uncertainty added in quadrature) of the experimental values from the SM best fit with M_H allowed as a free parameter, while the column "Deviation" refers to the SM best fit with $M_H = 117$ GeV fixed. The agreement is generally very good and there are no major discrepancies.

The global fit can be used to constrain the SM parameters including the Higgs boson mass,

$$M_H = 77^{+28}_{-22}\ \text{GeV} \tag{33}$$

the weak mixing angle, $\sin^2\theta_W(M_Z) = 0.23119 \pm 0.00014$, and the QCD coupling, $\alpha_s(M_Z) = 0.1217 \pm 0.0017$. The constraints by group of observable in the M_H–m_t plane are shown as contours in Figure 4.2. Good agreement is observed except that the low energy data (driven mostly by the NuTeV (Zeller *et al.*, 2002) result) tend to imply larger M_H. Also, most of the 90% CL ellipse is excluded by the Higgs boson searches at LEP 2, $M_H > 114.4$ GeV (Barate *et al.*, 2003), and indeed the one-standard-deviation range (33) is already ruled out. This may be due to a statistical fluctuation or could be an early indication of loop contributions from new particles beyond the SM affecting the precision observables. Including the results of LEP 2 searches (which include a slight excess of candidate Higgs boson events) in the analysis yields the 95% CL upper bound,

$$M_H \leq 167\ \text{GeV} \tag{34}$$

Thus, if the SM is correct, then the Higgs boson is not expected to be much heavier than the massive gauge bosons, and may be producible at the Tevatron and should certainly be detected at the Large Hadron Collider (LHC). The LHC is a soon-to-be-inaugurated accelerator placed in the LEP tunnel at CERN, which will mostly study proton–proton collisions at very high energies (around 14 TeV).

Particle Properties (Yao *et al.*, 2006). This reference also contains in-depth reviews on many SM-related topics and is updated biennially.

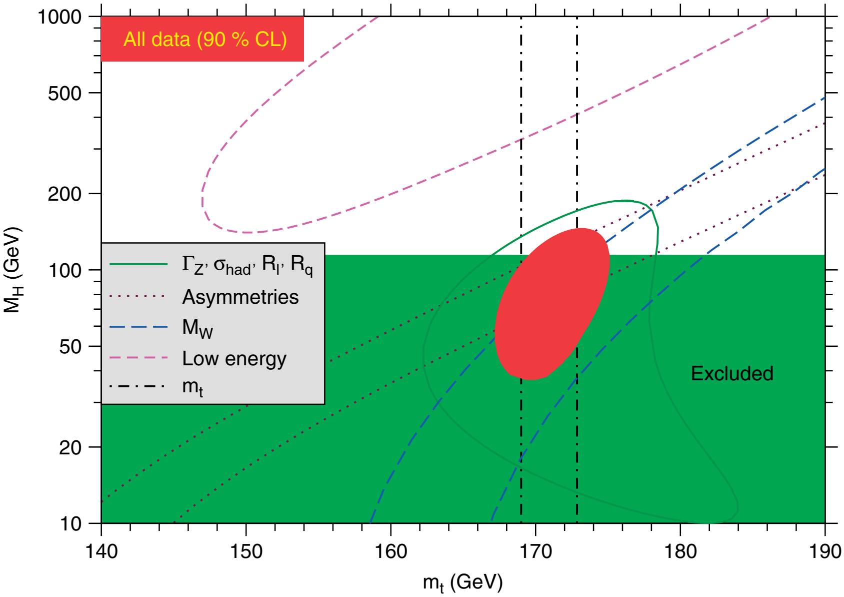

Fig. 4.2 One-standard-deviation (39.35%) uncertainties in M_H as a function of m_t for various inputs, and the 90% confidence level (CL) region allowed by all data. The 95% direct exclusion limit from LEP 2 (Barate *et al.*, 2003) is also indicated.

One can also perform a fit to the precision data alone, that is, excluding the direct m_t from the Tevatron. The result, $m_t = 175^{+10}_{-8}$ GeV, is in perfect agreement with the direct determination and provides an impressive confirmation of the SM at the level of radiative corrections.

4.5
Beyond the Minimal Model

Despite all its successes, the SM is widely believed to be correct only up to small nonrenormalizable (see Section 4.2.7) correction terms suppressed by some high energy scale Λ_{new}. This is because the mass parameter, m_Φ, in the Higgs potential generally receives large radiative corrections of the order of Λ_{new}, destabilizing the electroweak scale. Moreover, the gauge group, the irreps, and the values of the input parameters are chosen in an ad hoc way to conform with observation, but are not understood at any deeper level.

4.5.1
Accidental Symmetries and Nonrenormalizable Corrections

Whatever may be the underlying theory beyond the SM that cures these problems, at the electroweak scale one can organize it in an expansion in inverse powers of Λ_{new}, where each term (operator) contains only SM fields and where the unsuppressed (renormalizable) terms define the SM. The coefficients of the nonrenormalizable operators are to be adjusted to reflect the full theory (matching). In other words, extra heavy particles and fields are integrated out (their low energy effects are included as quantum mechanical averages) leaving their footprints in these coefficients. The SM is then considered as the low energy limit (effective field theory) of the more fundamental theory. Since at present one can only speculate as to what this theory might be, one may alternatively take the coefficients of the nonrenormalizable operators to parameterize arbitrary extensions of the SM constrained only by the local SM gauge symmetries.

In addition to these nonnegotiable local gauge symmetries, the renormalizable SM Lagrangian happens to have additional symmetries, which arise simply because any term that would violate them would also violate local gauge invariance. For example, electron number, muon number[16], and tau number are (at least perturbatively) conserved, and likewise the total quark number and thus baryon number. And as discussed in Section 4.3.4, the two generation SM conserves CP in the electroweak sector, leading to the suppression of CP violation in the full three generation model. Thus, one can understand such symmetries as *accidental* without the need to postulate them as additional physical principals. But by the same token, one would generally expect them to be broken by nonrenormalizable terms.

4.5.2
Lepton Number Nonconservation and Neutrino Mass

At order $\Lambda_{\text{new}}^{-1}$ baryon number is still accidentally conserved, but one can form Lorentz and locally gauge-invariant terms (Weinberg, 1979) out of any two lepton doublets and two Higgs doublets. Replacing each Higgs doublet by the VEV (1.15) yields a sum of terms, which can be interpreted as mass terms for neutrinos,

$$\mathcal{L}_M = -\sum_{ij} \frac{\lambda_{ij}}{\Lambda_{\text{new}}} \overline{\nu_i^c} \nu_j \, v^2 \tag{35}$$

But unlike the mass terms for charged leptons and quarks, which originate from the Yukawa terms (21) and connect left- and right-handed fields (Dirac mass terms), $\mathcal{L}_M$ connects the left- and right-handed components of conjugate fields (Majorana mass terms).

As a result, Majorana mass terms (Majorana, 1937) for neutrinos violate lepton number (the number of leptons minus antileptons) by two units and the same holds true for the corresponding trilinear and quadrilinear neutrino–Higgs interactions. This implies the important prediction of the possibility of neutrinoless double β decays ($0\nu\beta\beta$ decays), such as $K^- \to \pi^+ e^- e^-$ and more importantly of nuclei. $0\nu\beta\beta$ decays are searched for in experiments with large volume detectors but have not been observed so far.

Another consequence of $\mathcal{L}_M$ is the lepton number conserving but lepton flavor number violating process of neutrino oscillations (the change of one ν flavor into another over long flight paths), provided the matrix of parameters, λ_{ij}, has nonvanishing off-diagonal entries, in which case the mass and weak interaction eigenstates of neutrinos are not the same.[17] This mirrors the situation from the quark sector, and indeed a neutrino mixing matrix (Maki, Nakagawa and Sakata, 1962), V_{MNS}, analogous to V_{CKM} in Eq. (30), arises. The only difference is that left-handed fields are not independent of right-handed antifields so that the counting of observable phases is modified. The result is that there are $N_\nu - 1$ additional (CP violating) phases (Majorana phases) because, in contrast to the quark sector, phase redefinitions of right-handed fields do not enter here. While it appears to be extraordinarily difficult to design an experiment to measure the two Majorana phases (for $N_\nu = 3$), several of the other parameters in V_{MNS} have already been measured in experiments studying neutrinos originating from the sun (Cleveland *et al.*, 1998), the earth's atmosphere (Fukuda *et al.*, 1998), nuclear reactors (Eguchi *et al.*, 2003), and particle accelerators (Aliu *et al.*, 2005). What is observed in these experiments is the disappearance (rate decrease) of neutrinos of a particular flavor from a source with known flavor composition. Conversely, one may detect a neutrino flavor not initially present in the source (appearance experiment). On the other hand, the sum of all neutrino flavors appears to be unchanged (Ahmad *et al.*, 2001), as is expected if V_{MNS} is indeed the (only) origin of neutrino oscillations. It is conceivable that the extra phase in V_{MNS} that has its direct analog in V_{CKM} (Dirac phase) may also be accessible in future ν oscillation experiments. In any case, all experimental results are consistent with neutrino mass differences of the order of 10^{-1} eV (from atmospheric neutrinos) and 10^{-2} eV (from solar neutrinos) and generally large mixing angles. Assuming that the λ_{ij} are of $\mathcal{O}(1)$ or smaller,[18] one obtains estimates, $\Lambda_{\mathrm{new}} \lesssim \mathcal{O}(10^{15}\ \mathrm{GeV})$ and $\lesssim \mathcal{O}(10^{16}\ \mathrm{GeV})$, respectively. This is interestingly close to the Planck scale, $M_P \sim \mathcal{O}(10^{18}\ \mathrm{GeV})$, defined as the scale where quantum corrections to gravity are no longer negligible and where fundamental and far reaching modifications to current theories of elementary particles are expected.[19]

---

[16] These forbid processes such as $\mu \to e\gamma$.

[17] Of course, this also requires that the neutrino masses differ from each other. Hence, neutrino oscillation amplitudes are proportional to neutrino mass differences.

The term (35) was introduced without reference to any particular model of the physics at higher energy scales. The most important (because very simple) concrete realization is the so-called seesaw mechanism (Yanagida, 1979; Gell-Mann, Ramond and Slansky, 1979; Mohapatra and Senjanovic, 1980) in which right-handed neutrinos, ν_R, are added to the SM. They are singlets under the SM gauge group and one would generally expect them to have very large (Majorana) masses (there is no symmetry reason why they should be of the order of the electroweak scale). The role of Λ_{new} is played here by the right-handed neutrino mass scale, and the left-handed neutrino masses are generated by mixing terms analogous to Eq. (21). The term (35) would now be the result of integrating out the ν_R and one arrives at the SM as an effective field theory as discussed at the beginning of this section.

Alternatively, since the ν_R do not participate in the SM gauge interactions, one may also assume that for some reason they cannot have Majorana masses. In this case, one would be left with the Yukawa terms (21) and Dirac neutrino masses, and V_{MNS} would have the same form as V_{CKM}.

4.5.3
Baryon Number Nonconservation and Proton Decay

At order $\Lambda_{\mathrm{new}}^{-2}$ there are interaction terms (Weinberg, 1979; Wilczek and Zee,

---

[18] The λ_{ij} cannot be much larger than one, since their β-functions (see Section 4.3.1) would drive them back to order unity at the electroweak scale even if they started out much larger at Λ_{new}.

[19] Additional modifications may become necessary at additional scales between M_Z and M_P.

1979) that violate both baryon and lepton number. These involve one lepton and three quark fields and predict proton decay rates at $\mathcal{O}(\Lambda_{\text{new}}^{-4})$. The proton lifetime, τ_p, can be studied with large water detectors, but so far proton decay has not been observed and a lower (mode-independent) limit, $\tau_p > 2 \times 10^{29}$ years (Yao *et al.*, 2006), could be set. On dimensional grounds, τ_p is roughly of the order $\Lambda_{\text{new}}^4/m_p^5$ ($m_p = 938$ MeV is the proton mass), so that (assuming that all relevant coupling constants are of order unity) $\Lambda_{\text{new}} \gtrsim \mathcal{O}(10^{15}$ GeV), which is not inconsistent with the scales indicated by ν oscillations.

A realization of these model-independent observations is the *Grand Unified theories* (Langacker, 1981) in which quarks and leptons are conjectured to inhabit the same irreps of a larger gauge group, the simplest possibility being $SU(5)$. This implies proton decay (e.g. $p \to e^+\pi^0$), with the role of Λ_{new} being played by the masses of the additional gauge bosons predicted by the larger local gauge symmetry (SSB induced by extra Higgs fields would have to be at work).

Baryon number is violated in the SM even without the nonrenormalizable operators. As mentioned in Section 4.2.7, anomalies can violate the otherwise conserved quantities. In the SM, this happens both to baryon and lepton number (but not the difference), and can be understood as an instanton effect ('t Hooft, 1976). However, this is an extremely small (by any standards) quantum mechanical tunneling effect (from one vacuum to a gauge transformed vacuum) and is therefore only of academic interest.

Even though proton decay has not yet been observed, baryon number violation is a pressing problem: if baryon number is exactly conserved, then the baryon asymmetry of the universe (BAU), that is, why the universe contains more baryons than antibaryons (no large concentrations of antimatter are observed), would be difficult to understand. Moreover, baryon number violation is necessary but not sufficient to produce the BAU (Sakharov, 1967); a fairly strong source of CP violation beyond the CKM phase δ is necessary, as well. Thus, there is a large variety of reasons to strongly suspect that there is physics beyond the SM.

Glossary

Action: The action is defined to generate the classical equations of motions as its stationary "points" (Euler–Lagrange equations). It is particularly useful to construct theories that are invariant under some symmetry such as gauge invariance. A quantity called *action* also enters the weight factor in the path integral formalism, but is not always identical to its classical counterpart.

Anomaly: Symmetry violation by quantum effects. Gauge anomalies may appear in chiral theories, in which case they spoil their mathematical consistency. Anomalies of other symmetries are permissible and may lead to interesting physical effects.

Chiral Theory: A theory that fundamentally distinguishes left-handed and right-handed particles and fields (handedness).

Feynman Diagrams: Effective and intuitive way to visualize the perturbative contributions to quantum mechanical amplitudes. Each line in a Feynman diagram represents a particle. Each diagram translates into a concrete mathematical expression.

Gauge Symmetry: Gauge invariance of the first kind is invariance under global (space-time-independent) phase multiplications and their non-Abelian generalizations.

Gauge invariance of the second kind is the generalization to local (space-time-dependent) transformations.

Hamiltonian: Symmetry generator of time translations and thus one of the elements of the Poincaré algebra. It is an operator with energy eigenvalues. In QFTs it usually takes the form of a space integral over a Hamiltonian density, which is also often referred to simply as *Hamiltonian*.

Lagrangian: The action usually takes the form of a time integral over a function called the *Lagrangian*. The Hamiltonian (density) can be obtained in terms of the Lagrangian (density).

Loops: Closed line in a Feynman diagram. A loop represents virtual particles of arbitrary energy and momentum over which one has to sum (integrate). Feynman diagrams with loops are often divergent.

Path Integral: Space-time approach to quantum mechanics in which amplitudes are given as functional integrals over all possible path histories weighted by a complex phase factor given in terms of the action.

Poincaré Group: The fundamental symmetry group of space-time consisting of translations in space and time, rotations, and Lorentz boosts. The Poincaré group can be generated by *infinitesimal* symmetry transformations. The corresponding symmetry generators form the associated Poincaré algebra.

Regularization: Technique to organize the divergent parts of Feynman diagrams. Physical results should not depend on the regularization method.

Renormalizable Theory (Modern Sense): Theory in which all divergences can be absorbed by counterterms.

Renormalizable Theory (Strict Sense): Theory in which all divergences can be absorbed by a *finite* number of counterterms.

Renormalization: The process to absorb the divergences in counterterms. Only the finite sum of the divergent Feynman diagrams and infinite counterterms are physically meaningful.

References

Aaltonen, T. *et al.* (CDF Collaboration) (**2007**) *http://arxiv.org/abs/*0708.3642.

Abachi, S. *et al.* (D0 Collaboration) (**1995**) *Phys. Rev. Lett.*, **74** 2632–37.

Abazov, V.M. *et al.* (CDF and D0 Collaborations) (**2004**) *Phys. Rev. D*, **70** 092008.

Abe, F. *et al.* (CDF Collaboration) (**1995**) *Phys. Rev. Lett.*, **74** 2626–31.

Adler, S.L. (**1969**) *Phys. Rev.*, **177** 2426–38.

Ahmad, Q.R. *et al.* (SNO Collaboration) (**2001**) *Phys. Rev. Lett.*, **87** 071301.

ALEPH, DELPHI, L3, OPAL, SLD Collaborations, LEP Electroweak Working Group, SLD Electroweak and Heavy Flavour Groups (**2006**) *Phys. Rept.*, **427** 257–454.

ALEPH, DELPHI, L3, OPAL Collaborations, LEP Electroweak Working Group (**2006**) *http://arxiv.org/abs/hep-ex/0612034*.

Aliu, E. *et al.* (K2K Collaboration) (**2005**) *Phys. Rev. Lett.*, **94** 081802.

Amaldi, U. *et al.* (**1987**) *Phys. Rev. D*, **36** 1385–407.

Anthony, P.L. *et al.* (SLAC–E–158 Collaboration) (**2005**) *Phys. Rev. Lett.*, **95** 081601.

Arnison, G. *et al.* (UA1 Collaboration) (**1983a**) *Phys. Lett. B*, **122** 103–116.

Arnison, G. *et al.* (UA1 Collaboration) (**1983b**) *Phys. Lett. B*, **126** 398–410.

Aubert, J.J. *et al.* (BNL–E–598 Collaboration) (**1974**) *Phys. Rev. Lett.*, **33** 1404–6.

Augustin, J.E. *et al.* (SLAC–SP–017 Collaboration) (**1974**) *Phys. Rev. Lett.*, **33** 1406–8.

Bagnaia, P. *et al.* (UA2 Collaboration) (**1983**) *Phys. Lett. B*, **129** 130–40.

Banner, M. *et al.* (UA2 Collaboration) (**1983**) *Phys. Lett. B*, **122** 476–85.

Barate, R. *et al.* (ALEPH, DELPHI, L3, OPAL Collaborations, LEP Working Group for Higgs boson searches) (**2003**) *Phys. Lett. B*, **565** 61–75.

Bartel, W. *et al.* (JADE Collaboration) (**1980**) *Phys. Lett. B*, **91** 142–47.

Belavin, A.A., Polyakov, A.M., Shvarts, A.S., and Tyupkin, Yu.S. (**1975**) *Phys. Lett. B*, **59** 85–87.

Bell, J.S. and Jackiw, R. (**1969**) *Nuovo Cimento A*, **60** 47–61.

Bennett, G.W. *et al.* (Muon g-2 Collaboration) (**2004**) *Phys. Rev. Lett.*, **92** 161802.

Berger, C. *et al.* (PLUTO Collaboration) (**1979**) *Phys. Lett. B*, **86** 418–25.

Bloom, E.D. *et al.* (**1969**) *Phys. Rev. Lett.*, **23** 930–34.

Cabibbo, N. (**1963**) *Phys. Rev. Lett.*, **10** 531–32.

Callan, C.G. (**1970**) *Phys. Rev. D*, **2** 1541–47.

Cartan, E. (**1894**) *Sur la Structure des Groupes de Transformations Finis et Continue*, Paris.

Cleveland, B.T. *et al.* (**1998**) *Astrophys. J.*, **496** 505–26.

Czarnecki, A. and Marciano, W.J. (**1998**) *Int. J. Mod. Phys. A*, **13** 2235–44.

Davier, M., Hocker, A., and Zhang, Z. (**2006**) *Rev. Mod. Phys.*, **78** 1043–109.

Dirac, P.A.M. (**1928**) *Proc. R. Soc. London A*, **117** 610–24.

Dyson, F.J. (**1949**) *Phys. Rev.*, **75** 1736–55.

Eguchi, K. *et al.* (KamLAND Collaboration) (**2003**) *Phys. Rev. Lett.*, **90** 021802.

Einstein, A. (**1916**) *Ann. Phys.*, **49** 769–822.

Englert, F. and Brout, R. (**1964**) *Phys. Rev. Lett.*, **13** 321–22.

Erler, J. and Langacker, P. (**1995**) *Phys. Rev. D*, **52** 441–50.

Erler, J. and Ramsey-Musolf, M.J. (**2005**) *Phys. Rev. D*, **72** 073003.

Fermi, E. (**1934**) *Z. Phys.*, **88** 161–77.

Feynman, R.P. (**1948**) *Rev. Mod. Phys.*, **20** 367–87.

Feynman, R.P. (**1949**) *Phys. Rev.*, **76** 749–59.

Fock, V. (**1926**) *Z. Phys.*, **39** 226–32.

Fujikawa, K. (**1979**) *Phys. Rev. Lett.*, **42** 1195–98.

Fukuda, Y. *et al.* (Super-Kamiokande Collaboration) (**1998**) *Phys. Rev. Lett.*, **81** 1562–67.

Gabrielse, G. *et al.* (**2006**) *Phys. Rev. Lett.*, **97** 030802.

Gell-Mann, M. and Low, F.E. (**1954**) *Phys. Rev.*, **95** 1300–312.

Gell-Mann, M., Ramond, P., and Slansky, R. (**1979**) Complex spinors and unified theories, in *Supergravity* (eds P. Van Nieuwenhuizen and D.Z. Freedman), North-Holland, Amsterdam, pp. 315–21.

Ginges, J.S.M. and Flambaum, V.V. (**2004**) *Phys. Rept.*, **397** 63–154.

Glashow, S.L.J., Iliopoulos, J., and Maiani, L. (**1970**) *Phys. Rev. D*, **2** 1285–92.

Goldstone, J. (**1961**) *Nuovo Cim.*, **19** 154–64.

Goldstone, J., Salam, A., and Weinberg, S. (**1962**) *Phys. Rev.*, **127** 965–70.

Golfand, Yu.A. and Likhtman, E.P. (**1971**) *JETP Lett.*, **13** 323–26.

Gross, D.J. and Wilczek, F. (**1973a**) *Phys. Rev. Lett.*, **30** 1343–46.

Gross, D.J. and Wilczek, F. (**1973b**) *Phys. Rev. D*, **8** 3633–52.

Guralnik, G.S., Hagen, C.R., and Kibble, T.W.B. (**1964**) *Phys. Rev. Lett.*, **13** 585–87.

Hasert, F.J. *et al.* (Gargamelle Neutrino Collaboration) (**1973**) *Phys. Lett. B*, **46** 138–40.

Higgs, P.W. (**1964**) *Phys. Rev. Lett.*, **13** 508–509.

't Hooft, G. (**1973**) *Nucl. Phys. B*, **61** 455–68.

't Hooft, G. (**1976**) *Phys. Rev. Lett.*, **37** 8–11.

't Hooft, G. and Veltman, M.J.G. (**1972**) *Nucl. Phys. B*, **44** 189–213.

Kobayashi, M. and Maskawa, T. (**1973**) *Prog. Theor. Phys.*, **49** 652–57.

Langacker, P. (**1981**) *Phys. Rept.*, **72** 185–385.

Lee, B.W. and Zinn-Justin, J. (**1972**) *Phys. Rev.*, **D5** 3121–60.

Lorentz, H.A. (**1909**) *The Theory of Electrons and Its Applications to the Phenomena of Light and Radiant Heat*, B. G. Teubner, Leipzig.

Majorana, E. (**1937**) *Nuovo Cim.*, **14** 171–84.

Maki, Z., Nakagawa, M., and Sakata, S. (**1962**) *Prog. Theor. Phys.*, **28** 870–80.

Marciano, W.J. and Sirlin, A. (**1988**) *Phys. Rev. Lett.*, **61** 1815–18.

Maxwell, J.C. (**1865**) *Phil. Trans. R. Soc. London*, **155** 459–512.

Michel, L. (**1950**) *Proc. Phys. Soc. A*, **63** 514–31.

Miller, J.P., de Rafael, E., and Roberts, B.L. (**2007**) *Rept. Prog. Phys.*, **70** 795–881.

Mohapatra, R.N. and Senjanovic, G. (**1980**) *Phys. Rev. Lett.*, **44** 912–15.

Nambu, Y. (**1960**) *Phys. Rev. Lett.*, **4** 380–82.

Paschos, E.A. and Wolfenstein, L. (**1973**) *Phys. Rev. D*, **7** 91–95.

Politzer, H.D. (**1973**) *Phys. Rev. Lett.*, **30** 1346–49.

Prescott, C.Y. *et al.* (SLAC–E–122 Collaboration) (**1978**) *Phys. Lett. B*, **77** 347–52.

van Ritbergen, T. and Stuart, R.G. (**2000**) *Nucl. Phys. B*, **564** 343–90.

Sakharov, A.D. (**1967**) *JETP Lett.*, **5** 24–27.

Salam, A. (**1968**) Weak and Electromagnetic Interactions, in *Elementary Particle Theory: Relativistic Groups and Analyticity: Proceedings of the Eighth Nobel Symposium*

(ed N. Svartholm), Almqvist and Wiksells, Stockholm, pp. 367–77.
Schwinger, J.S. (**1948**) *Phys. Rev.*, **73** 416–17.
Slansky, R. (**1981**) *Phys. Rept.*, **79** 1–128.
Stueckelberg, E.C.G. and Petermann, A. (**1953**) *Helv. Phys. Acta*, **26** 499–520.
Symanzik, K. (**1970**) *Commun. Math. Phys.*, **18** 227–46.
Tevatron Electroweak Working Group (CDF and D0 Collaborations) (**2007**) *http://arxiv.org/abs/hep-ex/0703034*.
Weinberg, S. (**1965**) *Phys. Rev.*, **138** B988–B1002.
Weinberg, S. (**1967**) *Phys. Rev. Lett.*, **19** 1264–66.
Weinberg, S. (**1973**) *Phys. Rev. Lett.*, **31** 494–97.
Weinberg, S. (**1979**) *Phys. Rev. Lett.*, **43** 1566–70.
Wilczek, F. and Zee, A. (**1979**) *Phys. Rev. Lett.*, **43** 1571–73.
Wess, J. and Zumino, B. (**1974**) *Nucl. Phys. B*, **70** 39–50.
Wood, C.S. *et al.* (**1997**) *Science*, **275** 1759–63.
Yanagida, T. (**1979**) Horizontal gauge symmetry and masses of neutrinos, in *Proceedings of the Workshop on the Baryon Number of the Universe and Unified Theories* (eds O. Sawada and A. Sugamoto), KEK, Tsukuba, pp. 95–98.
Yang, C.N. and Mills, R.L. (**1954**) *Phys. Rev.*, **96** 191–95.
Yao, W.M. *et al.* (Particle Data Group) (**2006**) *J. Phys. G*, **33** 1–232.
Yukawa, H. (**1935**) *Proc. Phys. Math. Soc. Jpn.*, **17** 48–57.
Zeller, G.P. *et al.* (NuTeV Collaboration) (**2002**) *Phys. Rev. Lett.*, **88** 091802.

Further Reading

The non-historical presentation followed here is strongly inspired by the technical but very pedagogical treatise in

Weinberg, S. (**1995**) *The Quantum Theory of Fields, Vol. 1: Foundations*, Cambridge University Press, Cambridge.
Weinberg, S. (**1996**) *The Quantum Theory of Fields, Vol. 2: Modern Applications*, Cambridge University Press, Cambridge.

A selection of more traditional sources is

Abers, E.S. and Lee, B.W. (**1973**) *Phys. Rept.*, **9** 1–141.
Bohm, M., Denner, A., and Joos, H. (**1981**) *Gauge Theories of the Strong and Electroweak Interaction*, Teubner Verlag, Stuttgart.
Cheng, T.P. and Li, L.F. (**1984**) *Gauge Theory of Elementary Particle Physics*, Oxford University Press, Oxford.
Commins, E.D. and Bucksbaum, P.H. (**1983**) *Weak Interactions of Leptons and Quarks*, Cambridge University Press, Cambridge.
Georgi, H. (**1984**) *Weak Interactions and Modern Particle Theory*, Addison-Wesley Publishing Company, Redwood City.
Peskin, M.E. and Schroeder, D.V. (**1995**) *An Introduction to Quantum Field Theory*, Westview Press, Boulder.
Renton, P. (**1990**) *Electroweak Interactions: an Introduction to the Physics of Quarks and Leptons*, Cambridge University Press, Cambridge.

For precision tests of the SM, see

Bardin, D. and Passarino, G. (**1999**) *The Standard Model in the Making: Precision Study of the Electroweak Interactions*, Oxford University Press, Oxford.
Cvetic, M. and Langacker, P. (eds) (**1991**) *Testing the Standard Model: Proceedings of the 1990 Theoretical Advanced Study Institute in Elementary Particle Physics*, World Scientific Publishing Company Inc, Singapore.
Erler, J. and Ramsey-Musolf, M.J. (**2005**) *Prog. Part. Nucl. Phys.*, **54** 351–442.
Langacker, P. (ed) (**1995**) *Precision Tests of the Standard Electroweak Model*, World Scientific Publishing Company Inc, Singapore.

WEATHER FORECASTING

STEVEN R. SILBERBERG, *Meteorology Program, Department of Geography, Northern Illinois University, DeKalb, Illinois, U.S.A.*

Introduction 409
1. **Observations of Weather Systems** 411
1.1 Time and Space Scales of Weather Systems and Sampling Considerations 411
1.2 Observing Systems 412
1.2.1 *In situ* Observations 412
1.2.2 Remote-Sensing Observations 414
2. **Numerical Weather Prediction** 417
2.1 Historical Background 417
2.2 Data Assimilation 419
2.3 Forecast Models 421
2.4 Statistical Forecasting 428
2.4.1 Stochastic-Deterministic Forecasting 429
3. **Forecast Evaluation** 432
3.1 Model Performance and Predictability 432
Glossary 436
Works Cited 437
Further Reading 439

INTRODUCTION

Weather exerts a powerful influence over human affairs, intimately touching the lives of all living beings and affecting the surrounding physical environment. During pleasant conditions it is hardly noticed; with the development of adverse conditions, however, the weather becomes the foremost topic of discussion on account of its significant economic and social impacts. Although the weather cannot be controlled, timely and accurate communication of impending changes are important for decisions at the personal, business, and government levels. Such decisions include, for example, determining proper dress; travel decisions; scheduling of activities, resources, and equipment; operation of transportation and utility systems; deployment of government resources; and, in rare situations, evacuations and emergency response for the protection of life and property. The nature and quality of these decisions depends upon a reliable and accurate assessment of future weather conditions—the weather forecast.

Throughout history civilizations have endeavored to foretell future weather conditions. The first forecasting methods involved keen nonquantitative observation of local conditions and how these observations were related to future weather events (Godske *et al.*, 1957). The Greek civilization built a tower of winds containing sculptures that described in graphical form the types of winds and weather they typically brought to Athens. Local weather folklore and indigenous weather knowledge is also based on the local forecast method. For example, the Basotho, inhabitants of the mountain kingdom of Lesotho in southern Africa, have developed many local forecast tools for weather and climate (Pepin, 1996). These local forecast tools include observations of cloud development, coverage, height, and color; the behavior of animals and birds; human physical conditions; the presence of a halo around the moon; the shape of the moon in its new phase; wind direction; the length of the rainy season in summer; and the mildness of winter. Some of these forecast tools are supported by scientific understanding and are used today, while others have no scientific basis and are considered to be meteorological folklore.

Development of quantitative observations of weather elements at a fixed point allowed further advances in the use of the local method of forecasting. Quantitative observations of rainfall commenced in the 15th century in Korea, temperature and pressure were first measured in the 17th century in

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Europe, and the vertical temperature structure was first observed during the 18th century in Scotland. In the United States, Ben Franklin carefully observed that precipitation in an east-coast storm traveled northeastward from Philadelphia to Boston, synthesizing that midlatitude storms moved opposite to the direction of the surface wind on their cold side. In the 19th century, cloud species were scientifically classified; humidity and wind measurements were started; detailed vertical profiles of temperature, pressure, and humidity were made; and self-recording instruments were introduced. By applying the local method to these observations in the 19th century, scientists formulated early ideas of air masses, fronts, cyclones, and midlatitude weather. Further advances were made in the early 20th century with the establishment of a formal single-station forecast method that was employed during the Second World War. Radar and electrical observations made at single stations were also incorporated into the local forecast method during wartime. Adaptations of the local forecast method appear in many general and introductory textbooks in meteorology and indicate forecasts based on values and changes of barometric pressure, wind speed, and direction.

The local forecast method is currently applied as an important supplement to numerical prognoses of weather. A form of the local method known as model output statistics (Glahn and Lowrey, 1972) utilizes numerical forecasts of meteorological variables and multiple-regression techniques to predict temperature, precipitation, cloud cover, wind direction and speed, and other forecast variables at individual stations. With an appropriate developmental database, model output statistics partially accounts for model errors and local effects. This method is described further in Sec. 2.4 of this article. Forecasters also employ the local forecast method to incorporate their experience and understanding of local and regional influences on weather systems to tailor forecasts more precisely, and to predict and warn of small space- and time-scale phenomena that cannot be resolved by observing systems, numerical forecasts, and statistical techniques.

With the establishment of national meteorological services and meteorological observing stations in Europe and North America in the 19th century, and the communication of these observations to a central location, the synoptic method of weather forecasting emerged and began to develop (Kutzbach, 1979). By plotting and analyzing simultaneous meteorological observations over a region, meteorologists began to define and describe the eastward-moving migratory high- and low-pressure systems of midlatitudes, the relationship between the wind and pressure distribution, variations of the pressure field, squall lines, and their associated precipitation, cloud, and wind. Forecasts made by extrapolation from these analyses met with modest success that decreased westward across both Europe and North America because of a lack of upstream observations over the Atlantic and Pacific Oceans. It is interesting to note that a similar westward decrease of forecast skill still occurs today across North America and Europe because of a similar lack of upstream observations.

Synoptic analysis and forecasting methods advanced during the early part of the 20th century as further investigations revealed the air-current structure of midlatitude low-pressure systems or cyclones, the identification of fronts and air masses, and the conceptualized life cycle of the polar-front cyclone and its relationship to the larger-scale circulation (Bjerknes and Solberg, 1922). Better availability of upper-air meteorological measurements led to the use of vertical atmospheric soundings of temperature and water vapor to forecast clouds, precipitation, temperature, thunderstorms, fronts, air masses, and fog. Analyses of the upper air flow revealed the jet stream, the large-scale wave motions of the midlatitude westerlies, and their relationship to surface cyclones and anticyclones (Palmén and Newton, 1969). From the analyses of upper-air wave patterns, Rossby (1939) provided a simplified theoretical model that predicted the movement and energy propagation of the large-scale waves in the midlatitude westerlies, and Sutcliffe and Forsdyke (1950) illustrated application of a simplified model that explained the formation of midlatitude cyclones and anticyclones.

With new observations, theoretical tools, and conceptual models, and an archive of synoptic analyses both at the surface and in the upper air, the synoptic method of forecasting began to assert itself as the primary forecast method in the mid-20th century. Synoptic forecasting methods employed in weather services included weather types and analogs; various types of extrapolation using nonphysical and quantitative physical reasoning, statistical techniques; and tropospheric models incorporating dynamical, thermodynamical, and synoptic experience (Petterssen, 1940, 1957; Riehl *et al.*, 1952; Bergeron, 1959). These techniques were used to prepare prognostic synoptic maps of surface and upper-air systems as well as forecasts of individual weather elements.

Despite the development and incorporation of local and synoptic forecasting methods into national weather services, advances in weather-forecasting skill were limited. Meager progress was attributed to poorly trained forecasters, numerous forecasting techniques, inability to determine the forecast problem, and the lack of a focused effort to address the fundamental scientific problems of meteorology (Willett, 1951). It was recognized that future forecasting progress would result from scientific weather forecasting: a completely mathematical and physical forecast of the atmosphere based on scientific meteorological principles.

Scientific weather forecasting was described in a landmark article nearly a century ago by Vilhelm Bjerknes (1904). On the basis of an initial-value problem approach, Bjerknes indicated that observations of the initial state of the atmosphere combined with knowledge of the physical laws governing future changes of the initial state provided a scientific basis for weather prediction. Obstacles to the successful implementation of this plan in 1904 were inadequate observations in space and time, incomplete knowledge of the physical laws, and a lack of sufficient mathematical knowledge and computing resources. Implementation of the Bjerknes approach to scientific forecasting through numerical integration of the physical laws by L. F. Richardson (1922) was unsuccessful, and this poor result deterred further work on the problem. Significant progress has been made in overcoming these obstacles during the past century, with the most rapid progress occurring during the past 40 years through advances in observing systems, computing technology, and mathematical techniques that have culminated in operational numerical weather prediction (Thompson, 1983; Shuman, 1989; Wiin-Nielsen, 1991). With the Bjerknes/Richardson plan in operation around the world, weather-forecasting skill in midlatitudes has advanced in predictive range and accuracy (Bengtsson, 1991). Forecast errors still result from inadequate observations in space and time, incomplete knowledge of the physical laws and processes, and a lack of sufficient computing resources to resolve the large range of spatial and temporal scales of meteorological phenomena.

In this article the practice of weather forecasting by operational numerical weather prediction in national weather services in the Northern Hemisphere midlatitudes is described. This practice involves observation of the four-dimensional structure of the atmosphere, assimilation of the data into a numerical forecast model for initial conditions, forecasting with a numerical model, and statistical and ensemble forecasting. The various observing systems that provide initial data for numerical forecasts are described first. Next, the theoretical principles and their application to operational numerical weather prediction are discussed. Applications of numerical prediction such as statistical and ensemble forecasting are reviewed, and the article concludes with an assessment of forecast skill and a discussion of the limits of predictability of a nonlinear dynamic system.

1. OBSERVATIONS OF WEATHER SYSTEMS

1.1 Time and Space Scales of Weather Systems and Sampling Considerations

Meteorological observations are essential for the analysis of weather systems and specification of the initial state of the atmosphere for operational numerical weather prediction. Observations and analyses have revealed a large variety of weather systems whose spatial and temporal scales typically increase together and range over seven orders of magnitude (Table 1). Of critical importance is that the smallest systems have the shortest duration, are typically the most intense, and may often require emergency

Table 1. Approximate time and space scales of selected weather systems (adapted from Fujita, 1986).

System	Time scale	Space scale
Dust devil	10 s	10 m
Tornado	10 min	100 m
Microburst	10 min	1 km
Mesocyclone	1 h	10 km
Thunderstorm	1 h	10 km
Supercell thunderstorm	3 h	50 km
Low-level jet streak	12 h	500 km
Mesoscale convective system	24 h	500 km
Warm front	2 d	1 000 km
Jet streak	3 d	1 000 km
Hurricane	5 d	1 000 km
Cold front	5 d	1 000 km
Anticyclone	10 d	2 000 km
Midlatitude cyclone	10 d	2 000 km
Short wave	10 d	5 000 km
Long wave	>0.5 month	10 000 km

response to protect life and property. In addition, weather systems and processes of all scales interact nonlinearly with each other. This nonlinear interaction and the large range of weather-system space and time scales imposes constraints on observing-system design in order to measure, analyze, and predict meteorological phenomena.

Weather observations sample a continuous four-dimensional atmosphere at discrete temporal and spatial resolution. A physical constraint of this discretization, world geography and politics, and the historical development of national and international observing networks is that large portions of the world are poorly sampled. An unavoidable consequence of this discretization is that complete four-dimensional measurements of atmospheric structure essential for the conduct of the Bjerknes/Richardson method of scientific weather forecasting are difficult to realize. This results in incomplete specification of the initial state of the atmosphere and errors in specification of the smallest scales of motion. Analysis errors are propagated across scales by nonlinear interaction during numerical forecasts and degrade forecast accuracy in space and time.

1.2 Observing Systems

Sampling issues and technological advances have guided development of a global observing system for weather forecasting. This system combines *in situ* and remote-sensing observing-system platforms distributed on the Earth's surface, in the atmosphere, and in space to optimize the temporal and spatial sampling of meteorological observations. The local and synoptic forecast methods and numerical weather prediction rely on observations gathered by a variety of operational observing-system platforms for the analysis and specification of the initial state of the atmosphere. The various platforms are operated by national weather services with observations first transmitted to national weather-service centers and then to other forecasting centers by a global telecommunications system for worldwide use in weather analysis and numerical forecasting.

1.2.1 *In situ* Observations One cornerstone of the observing system is *in situ* measurements made by surface stations. Surface observations of the horizontal wind, temperature, pressure, water vapor, clouds, precipitation, snow cover, and visibility are made at nearly hourly intervals by national and international networks of manual and automated stations on land. These hourly observations are augmented more frequently in rapidly changing weather conditions. Over the oceans, merchant, military, research, and weather service ships as well as fixed and drifting buoys provide the basic meteorological observations noted above and contribute additional observations of sea-surface temperature and salinity, ocean waves, and currents. The land-surface observing network is relatively dense in the populated midlatitudes of the Northern Hemisphere, and observation density decreases over tropical land areas, the polar regions, and less affluent countries of the world (Fig. 1). Oceanic observations are sampled with less spatial resolution than land areas, with the highest density over the North Atlantic and Pacific Ocean trade routes, sparse sampling over the tropical oceans, and very sparse sampling over the Southern Hemisphere oceans. The surface-observing network provides the highest spatial- and temporal-resolution measurements of meteorological variables of any observing system deployed worldwide.

A second cornerstone of the observing system is *in situ* measurements of the upper atmosphere made by balloon-borne instru-

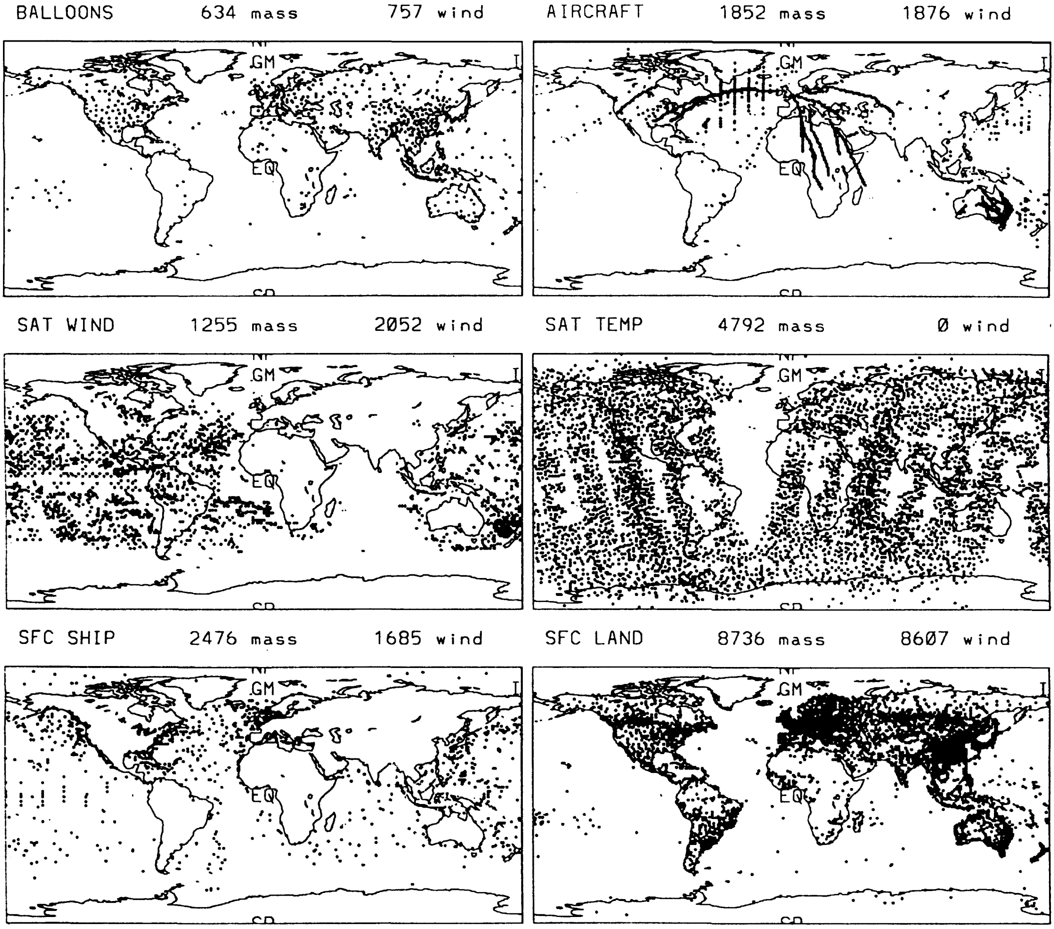
BALLOONS 634 mass 757 wind
AIRCRAFT 1852 mass 1876 wind
SAT WIND 1255 mass 2052 wind
SAT TEMP 4792 mass 0 wind
SFC SHIP 2476 mass 1685 wind
SFC LAND 8736 mass 8607 wind
GM
EQ
I

FIG. 1. Weather-observation data distribution for the globe for a 3-h window surrounding 0000 UTC 9 October 1996. "Balloons" represent pilot balloon and rawinsonde observations. "Sat Wind" indicates geostationary satellite-derived winds from infrared and water-vapor images. "Sat Temp" are vertical temperature profiles retrieved from polar-orbiting satellite radiance measurements. "Sfc Ship" represents observations from ships of opportunity, moored and drifting buoys, and other ocean observation platforms such as oil rigs. "Sfc Land" are automated and nonautomated observations from stations on land. Figure courtesy of E. Kalnay, U.S. National Centers for Environmental Prediction.

ment packages known as pilot balloons, rawinsondes, and dropwinsondes. Through visual tracking, pilot balloons measure wind speed and direction in clear air. Rawinsondes are tracked by radar to obtain horizontal wind observations. During their ascent they transmit pressure, temperature, and relative humidity measurements to ground stations at regular intervals, furnishing detailed vertical profiles of atmospheric structure. Dropwinsondes are similar to rawinsondes and are deployed by special observing aircraft in remote ocean regions where significant weather is occurring, such as hurricanes or typhoons and intense wintertime ocean storms. They transmit the same meteorological information as rawinsondes during their descent.

Upper-air balloon-borne observations are routinely made every 12 hours using a network of stations that is more sparse than the surface land-observing network (Fig. 1); dropwinsonde measurements are made infrequently at higher spatial resolution. With the exception of a few weather ships, ships of opportunity, and islands, the world oceans are poorly sampled by pilot balloons, rawinsondes, and dropwinsondes. Although high-resolution vertical profiles of the atmosphere are obtained by rawinsonde, twice-daily ascents and coarse spatial resolution limit these measurements to describing weather systems that last at least one day and are no smaller than approximately 1000 km in spatial scale.

Other *in situ* upper-air measurements are made by wide-body jetliners. Advanced wide-body jets make observations of temperature and horizontal wind speed and direction at nearly minute intervals along their routes. These observations provide detailed resolution of conditions during climbout and descent and along the flight path in the upper troposphere and lower stratosphere. Vertical profiles are concentrated at major world airports, and flight-level observations are clustered along major air routes. The high temporal and spatial resolution of the observations defines the structure of smaller-scale phenomena only to the extent that they lie along flight paths (Fig. 1).

1.2.2 Remote-Sensing Observations A third cornerstone of the world observing system is passive remotely sensed radiation measurements by satellite platforms in polar and geostationary orbits. Polar-orbiting satellites from the United States and the former Soviet Union orbit the Earth every 100 min in sun-synchronous orbits that are inclined 98.8° from the equator at an altitude of approximately 850 km. This nearly pole-to-pole orbit allows different satellites to pass over the equator at a defined local solar time. As the Earth rotates beneath the satellite, different swaths of approximately 2600 km are viewed with each orbit. Examples of these swaths can be seen in Fig. 1 (middle right) and Fig. 2 (upper left).

Geostationary satellites are located over the equator and orbit the Earth at an altitude of approximately 36 000 km. At this orbit altitude, the satellite completes one revolution of the Earth every 24 h and remains fixed over the equator at a specific longitude. Five geostationary satellites, from the United States, Japan, India, and Europe, are positioned at specific longitudes with the view from each satellite shown in Fig. 2. Meteosat, the European geostationary satellite, is located at 0° longitude; Insat, the Indian satellite, is located over 75° E longitude; GMS, the Japanese satellite, is located over 140° E longitude; and GOES-West and East, two U.S. satellites, are located at 135° W and 75° W longitude, respectively.

A combination of both polar-orbiting and geostationary satellites is necessary to view the entire globe, with each satellite furnishing a unique perspective of the Earth. For example, animation of successive geostationary satellite images shows cloud and circulation features in motion, but animation cannot be performed with polar-orbiting satellites that view a different Earth scene on each orbit. Geostationary satellites cannot measure Earth radiation at latitudes poleward of approximately 70° latitude on account of the spherical nature of the Earth; however, polar-orbiting satellites regularly view a large portion of the polar regions with every orbit.

Visible and infrared imagery from the first polar-orbiting and geostationary satellites in the 1960s furnished a global perspective of weather systems, revolutionized weather analysis and forecasting, and provided qualitative meteorological information over formerly data-sparse regions such as the oceans and Southern Hemisphere. Satellite imagery immediately became indispensable for analysis over data-sparse regions and for short-term forecasts of weather phenomena such as tropical storms, jet streams, upper-level disturbances, convective storms, clouds, and precipitation. Imagery was utilized to monitor the quality of numerical model initialization and forecasts, providing qualitative guidance where weather systems are poorly defined in model initial analyses and where model prognoses diverge from interpretations of satellite imagery. In particular, the analysis and monitoring of tropical cyclones by satellite imagery led to improved warnings, and to the development of pattern-recognition techniques to forecast intensity changes and storm motion and to direct more accurately aircraft missions supplying critical observations of tropical storm environment and structure. The global perspective of satellite imagery also cultivated conceptual advances in understanding extratropical-storm structure, land-based convective storms, and circulation links between the tropics and extratropics, and sowed the seeds of global interdependence that led to future international experiments (Kidder and Vonder Haar, 1995; Bader *et al.*, 1995; Scofield and Purdom, 1986).

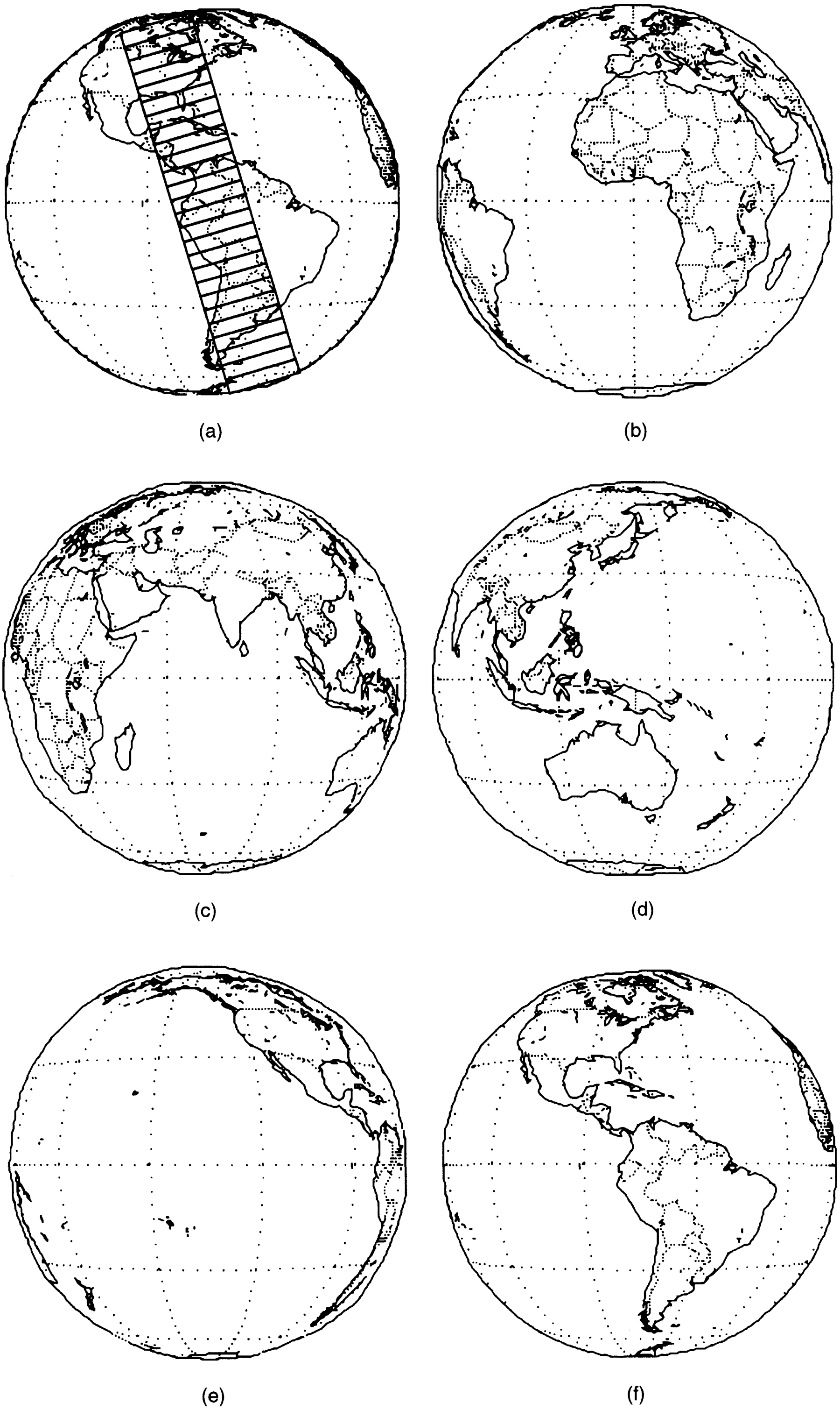
(a) (b) (c) (d) (e) (f)

FIG. 2. Schematic views of Earth scenes viewed from (a) a polar-orbiting satellite, (b) METEOSAT, (c) INSAT, (d) GMS, (e) GOES-West, and (f) GOES-East.

In addition to the qualitative and conceptual qualities of satellite imagery that revolutionized weather analysis and forecasting, a fundamental advantage of satellite platforms is their ability to observe and provide quantitative measurements of meteorological variables over poorly sampled portions of the Earth. Satellite remote sensing supplies tens of thousands of multispectral radiation measurements over data-sparse areas of the globe, which are converted to meteorological variables through retrieval techniques. Using remotely sensed radiances in the ultraviolet, infrared, and microwave portions of the electromagnetic spectrum on polar-orbiting and geostationary satellites, precipitation rates and vertical profiles of temperature, water vapor, and trace gases have been retrieved along with measurements of the Earth's radiation budget (Ferriday and Avery, 1994; Kidder and Vonder Haar, 1995). Additional variables derived from satellite remote sensing include the temperature of the land and sea surface, surface winds, sea level, sea ice, snow cover, soil moisture, precipitation rate, the pressure and temperature at cloud top, the amount of clouds covering the sky, and the amount of water vapor and liquid water in an atmospheric column. Severe-storm forecasters make use of further derived meteorological fields such as horizontal and vertical temperature gradients, indices that measure static stability, and horizontal wind fields derived from thermal gradients to forecast convective storms. Satellite-derived meteorological variables are routinely incorporated into data-assimilation systems to produce initial conditions for numerical forecast models (Fig. 1, middle left and right). Recent advances have allowed radiance measurements to be directly assimilated into numerical forecast models without the need for retrieval calculations.

Satellite retrieval techniques for determining meteorological variables have the highest quality in clear and partly cloudy atmospheres (Kidder and Vonder Haar, 1995). While cloud contamination degrades the accuracy of vertical profiles of atmospheric variables, the cloud field itself has been used to advantage for the generation of meteorological data. Successive images from geostationary satellites in the visible, infrared, and water-vapor absorption portions of the electromagnetic spectrum are used to determine cloud-motion wind vectors at lower and upper tropospheric levels. Height assignment of a wind vector is based on cloud-top temperature and its relationship to the background vertical temperature profile as well as other information that includes past history of the clouds tracked, experience of a manual operator, and correspondence with numerical model analyses. Cloud wind data can also be used to infer the temperature field through application of the geostrophic thermal wind relationship in midlatitudes. Winds estimated from successive water-vapor images complement cloud-tracked winds from infrared imagery in that water-vapor winds can be determined in regions of clear air where water-vapor imagery is able to discern circulation features. Cloud and water-vapor motion winds derived from geostationary satellites over data-sparse regions are routinely incorporated into numerical forecast models at forecasting centers worldwide (Fig. 1, middle left).

Future efforts at satellite remote sensing include active systems such as the Laser Atmospheric Wind Sounder (LAWS) and the Global Positioning System (GPS). Spaceborne LAWS systems measure backscattered radiation from aerosols and clouds to infer horizontal wind and cloud-top heights. LAWS wind measurements from a polar-orbiting platform will be able to resolve the global wind field with approximate horizontal resolution of 100 to 300 km and vertical resolution of 300 to 1000 m, and cloud-top heights to within 50 m at approximately 50-km horizontal resolution (Baker *et al.*, 1995). Observing-system simulation experiments have indicated that lidar wind measurements substantially improve analyses and forecasts over sparsely sampled areas of the globe such as the Southern Hemisphere and the tropical oceans. With the easing of cold war tensions, the military GPS, a network of orbiting satellites that allow precise measurement of location on Earth, has become available for nonmilitary applications. Research projects are applying GPS technology to determine horizontal winds through balloon tracking, vertically integrated water vapor, and vertical temperature and pressure profiles through measurement of the deviation of GPS signals passing through the atmosphere (Businger *et al.*, 1996; Ware *et al.*, 1996).

Active remote-sensing components of the global observing system are Doppler weather radar and wind profilers (Doviak and Zrnic, 1993; Strauch *et al.*, 1987). Doppler weather radars are fixed systems that measure backscattered microwave radiation to infer precipitation amount and intensity, and vertical profiles of the horizontal wind where appropriate microwave scattering bodies are present. Depending on radar operations and atmospheric conditions, horizontal wind and precipitation amount and intensity estimates are available every 5 min with a horizontal resolution of 250 m and vertical resolution of 300 m in the United States. Outside the United States Doppler weather radar is widely deployed in Europe and Asia but is not as common elsewhere.

Doppler wind profilers measure backscattered VHF and UHF radiation resulting from refractive-index gradients to infer vertical profiles of the horizontal wind in the troposphere and stratosphere. Nearly 50 wind profilers are installed across the United States with an approximate horizontal resolution of 300 km, a sampling frequency of 10 h^{-1}, and vertical resolution of 300 m. Outside the United States, nearly 40 wind profilers have been installed in South America, Europe, Asia, Australia, the Indonesian Archipelago, and tropical islands in the Pacific Ocean.

Other remote-sensing techniques, some of which are employed operationally, are microwave radiometers, sodar, lidar, and synthetic-aperture radar. Surface-based passive microwave radiometers measure thermodynamic profiles of temperature and water-vapor content throughout the lower troposphere with nearly continuous temporal resolution, but less vertical resolution than rawinsondes. Synthetic-aperture radar successfully measured sea-surface winds and rainfall with high spatial resolution for a brief period on a polar-orbiting satellite. Sodar or acoustic radar surface stations actively measure the structure, depth, and horizontal winds in the boundary layer of the atmosphere. Surface-based lidar or optical radar determines depth, temperature, water vapor, and gas constituents in the boundary layer by measuring backscattered energy from aerosols and air molecules. Infrared lidar is capable of measuring horizontal wind in the boundary layer (Golden *et al.*, 1986). A small number of passive microwave and active sodar and lidar systems are operational at various locations around the world.

The modern weather-forecasting observation system is a combination of *in situ* and both passive and active remote-sensing methods and systems. These systems are employed to measure atmospheric variables required for analysis of the current state of the atmosphere and climate. Advances in technology coupled with scientific analysis of the observation problem have played important roles in creating and developing the new observing systems. These advances continue to address the ongoing problem of incomplete observations of the four-dimensional state of the atmosphere.

2. NUMERICAL WEATHER PREDICTION

2.1 Historical Background

The scientific basis for modern numerical weather prediction, wherein future states of the atmosphere evolve from preceding ones according to known physical laws, originated with the initial-value problem approach described by Vilhelm Bjerknes (1904). In this approach, initial conditions of the atmosphere combined with knowledge of the three equations of motion on a rotating Earth, the first law of thermodynamics, the continuity equation (or conservation of mass) for air, the equation of state for air, and an equation for water vapor constitute a specification of the forecast problem where there are as many variables as equations. Bjerknes pointed out that exact analytical integration of this system of partial differential equations was impossible and indicated that graphical and numerical methods, particularly finite-difference methods, would have to be employed. L. F. Richardson (1922) began a "scheme of weather prediction," based upon finite-difference integration of the system of equations set forth by Bjerknes and performed the first detailed attempt at numerical weather prediction. Richardson pioneered formulation of finite-difference equations, solution techniques, and use of a three-dimensional grid upon which meteorological variables were tabulated and prognoses stored (Charney, 1950). Prediction of a 145-mbar pressure rise in 6 h for a point in western Europe suggested difficulties with the Richardson method, and these difficulties were later ascribed to imbalances between the mass and wind fields in the initial conditions (Charney, 1951; Lynch, 1992; CAeM Working Group, 1992). Although Richardson's noble attempt to solve numerically the predictive equations of atmospheric motion was not entirely successful, the advances he made paved the way for future progress in numerical prediction.

Observational, theoretical, and technical advances led to the birth of modern numerical weather prediction after the Second World War. A comprehensive *in situ* surface and observing system network was in place after the war that furnished a spatially dense set of initial conditions over continental regions of the Northern Hemisphere. Introduction of the barotropic vorticity equation by Rossby (1939) explained the relationship between the configuration and movement of the upper-level waves in the midlatitude westerlies (called Rossby waves) and the strength of the zonal wind. Downstream wave amplification commonly observed on upper-air analysis maps was also explained by the dispersive nature of Rossby waves, wherein energy propagation occurred by the faster group velocity rather than with the slower phase speed of ridge and trough movement (Rossby, 1938, 1940). Further theoretical advances included the analysis of baroclinic instability of waves in the midlatitude westerlies, the quasi-geostrophic approximation of atmospheric motion, the filtering of gravity waves from meteorological motions, and development of an equivalent barotropic model of the atmosphere (Charney, 1947, 1948, 1949; Charney *et al.*, 1950; Eliassen, 1949). Advances in mathematical theory by Courant *et al.* (1928) provided a theoretical basis for finite-difference solutions to approximate the continuous partial-differential-equation solutions of the meteorological equations. These observational and theoretical advances along with development of electronic computers created the opportunity to renew the work of Richardson in the mid 1940s.

Two numerical weather prediction programs were begun in the United States and Sweden with much collaboration between them. Recognizing the limited computing resources and the desire to avoid diagnosing complicated systems, the programs began with simplified models and gradually added increasing complexity (Charney, 1951; Wiin-Nielsen, 1991). The first successful forecasts were made with a one-level equivalent barotropic model (Charney *et al.*, 1950; Staff Members, 1952) that took 24 h of computation to complete a 24-h forecast. With slightly faster electronic computers, computation time decreased such that in 1954 the Swedish program performed the first barotropic forecasts completed in time for use by operational forecasters. Extension of numerical weather prediction models to include baroclinic effects (Charney and Phillips, 1953) and a successful forecast of a strong midlatitude storm along the east coast of North America convinced the United States to begin operational numerical weather prediction in 1955 (Shuman, 1989). Although the first operational forecasts began with more complicated baroclinic models, poor performance guided operational prediction back to the simpler equivalent barotropic model.

With numerical weather prediction established operationally, development and refinement of the analysis of observations, numerical methods, and models continued. Objective analysis, methods to interpolate irregularly spaced observations to a regular grid for numerical forecasts, were developed by Panofsky (1949) and developed further by Gilchrist and Cressman (1954), Bergthórsson and Döös (1955), and Cressman (1959). Advances in numerical methods and incorporation of a terrain-following vertical coordinate system led to improved forecasts and the development of more complicated models based on the so-called primitive equations discussed by Bjerknes in 1904 (Charney, 1955; Phillips, 1957). A hemispheric forecast model with six vertical levels based on the primitive equations was developed in the 1960s (Shuman and Hovermale, 1968), and its operational performance nearly equaled the skill of manually produced forecast maps. With this success, primitive-equation models established themselves as the primary tool for numerical weather prediction, and descendents of these models are still in use today (Shuman, 1989; Bengtsson, 1991).

During the past 40 years operational numerical weather prediction has evolved from a one-level equivalent barotropic model solved over continental domains on a relatively coarse grid mesh of approximately 400 km (the typical spacing of the upper air rawinsonde network that provides initial conditions for the forecasts) to global- and local-scale multilevel primitive-equation models solved on grid meshes as fine as 10 km. Throughout this advancement, the basic procedures of numerical weather prediction have remained unchanged. These procedures involve

1. observations of the four-dimensional state (x,y,z,t) of the atmosphere with the highest resolution possible,
2. assimilation of data from various observing platforms to produce an initial analysis of meteorological fields,
3. initialization to balance the analyzed fields dynamically,
4. forecasting future states by integrating the model forward in time, and
5. post-processing model output to compute quantities not forecast by the model and to produce forecast maps for interpretation and verification.

Procedure **1** was discussed in the first section of this article. The remaining portions of this section describe how procedures **2–5** are carried out at national forecast centers in the Northern Hemisphere midlatitudes.

2.2 Data Assimilation

Construction of three-dimensional analyses of meteorological variables for use as initial conditions in a primitive-equation forecast model is the purpose of data assimilation. The various observing-system platforms described in Sec. 1 produce nearly continuous measurements with differing spatial and temporal resolution and error characteristics, which must be blended together to estimate initial meteorological fields. The data assimilation process in practice at meteorological centers worldwide involves quality control of observations, objective analysis of irregularly spaced observations onto a three-dimensional lattice valid at a specific time, initialization of the analysis fields if required, and a short forecast from a data assimilation/forecast model to produce a background first-guess field for the next assimilation cycle (Dey, 1989; Daley, 1991; Ghil and Malanotte-Rizzoli, 1991; Benjamin *et al.,* 1991; Lorenc *et al.,* 1991). This process has evolved to provide the most accurate initial state of the atmosphere for numerical weather forecasting. The quality of a data-assimilation–produced initial analysis depends on observation accuracy, objective analysis frequency, and the accuracy of the forecast model supplying the first-guess field for the analysis. Details of data assimilation processes vary between forecasting centers with each center using different objective analysis methods and assimilation models, and either continuous or intermittent assimilation of data.

Data assimilation begins with comprehensive quality control of meteorological data to prevent incorrect observations from influencing the analysis (Hollingsworth *et al.,* 1986; Gandin, 1988; Collins and Gandin, 1990). Meteorological data continuously received at forecast centers is produced by a variety of observing platforms, each with different types of errors. These errors may be the result of human error in measurement, or to instrument errors due to failure, improper calibration or design, or the data itself may have been incorrectly coded and transmitted. Once an observation has been received at a forecast center, it is subjected to a number of quality-control inspections. Such checks search for gross errors such as incorrect coding, time, or location, and consistency of observations with neighboring observations, climatology, and a background first-guess field. After passing the gross-error checks, the observations are examined to see if they are dynamically consistent with themselves through use of the geostrophic relationship between the wind and height fields, the hydrostatic relationship between the temperature and height fields, and the geostrophic thermal wind relationship between the geostrophic wind and temperature fields. These quality-control inspections have advanced to the point where they are capable of discerning instrument errors, which can then be corrected (Hollingsworth *et al.,* 1986). Observations that pass scrutiny are then sent along to the objective analysis portion of the process.

Objective analysis has advanced remarkably from its initial beginnings with polynomial interpolation of irregularly spaced observations to a regular grid (Panofsky, 1949). Gilchrist and Cressman (1954) adapted polynomial interpolation to operational use and introduced the concept of a region of influence of observations by weighting observations on the basis of their distance from an analysis point. They also suggested the usefulness of having a prior estimate of the analysis field known as a first-guess field. Bergthórsson and Döös (1955) developed a technique that incorporated the suggestion of Gilchrist and Cressman and is the forerunner of current operational statistical interpolation methods. This technique uses a background first-guess field that is subtracted from the observations to produce observation increments. Weighted combinations of observation increments are then used to correct the background first-guess field within a particular distance from an analysis point. Cressman (1959) modified the Bergthórsson and Döös method to include successive corrections to the background first-guess field, with each iteration employing a reduced region of influence to define small-scale structure better. Global-scale objective analysis was developed in the late 1960s through an interpolation technique involving Hough function solutions to LaPlace's tidal equation.

When numerical weather prediction began in the early 1950s, most of the weather observations necessary to depict the three-dimensional structure of the atmosphere were from rawinsondes available twice daily at the synoptic times of 0000 and 1200 UTC. Technological developments in observational platforms such as satellites began to supply new observations over poorly sampled regions; however, these observations were made at asynoptic times and provided data of a different character than rawinsondes. Asynoptic data included portions of a meteorological observation such as vertical temperature profiles without wind information from satellites or just wind information from aircraft or cloud drift winds. Statistical interpolation techniques were developed and implemented (Gandin, 1963; Dey, 1989) to define mathematically the blending of various meteorological data streams into a three-dimensional analysis of the atmosphere. In this technique, estimates of observation and first-guess (model forecast) error are quantified as weighting values based on the ratio of the observation error to the first-guess error. Using these weighting values, the method blends observations of varying quality with the background first-guess field to produce an analysis whose mean square error is statistically minimized. The statistical analysis allows a defined number of local observations within a region to influence an analysis point. In addition, the statistical analyses are multivariate such that multiple types of local observations can influence the value of a particular analysis variable through various meteorological relationships.

Spectral statistical interpolation methods have been introduced into operational data assimilation systems in recent years to incorporate asynoptic data better and to determine more accurately analysis values over the globe (Derber *et al.,* 1991; Parrish and Derber, 1992). This method uses the entire global observation database to analyze model variables directly within the horizontal and vertical coordinates of the forecast model through statistical methods involving appropriate weights for various observing systems. By conducting the analysis in model coordinates, the number of information-losing interpolations between analysis and model coordinate systems is reduced, and, more importantly, dynamical constraints specific to the forecast model are incorporated into the interpolation. These attributes result in smoother corrections to the first-guess field and a better fit of the analysis to the observations, and they produce such a good balance between the wind and mass fields that initialization of the analysis fields to remove high-frequency oscillations prior to forecasting is no longer necessary. Use of this technique has allowed inclusion of satellite radiance measurements directly into the analysis, eliminating the need for retrieval calculations. Improvements to the analysis from this technique have reduced the root mean square error between the 6-h forecast and rawinsonde measurements to 16 m for geopotential height and 0.75 K for temperature, values that approach instrument error. Forecasts produced from spectral statistical interpolation generated analyses have shown improved performance.

Analyses produced by statistical interpolation methods often have imbalances between the mass and momentum fields that when integrated in a forecast model create high-frequency oscillations. These oscillations may have larger amplitude than meteorologically important motions and may not decay rapidly enough that at the next statistical interpolation step in the data assimilation cycle, good data may be rejected. The initialization step in data assimilation is designed to adjust the mass and momentum fields within consistent dynamical constraints to achieve balanced analyses and avoid rejection of good data in the assimilation system (Bengtsson, 1991). With the development of the spectral statistical interpolation method of objective analysis, initialization of the analyses is no longer required at some forecast centers.

Two variations of data assimilation systems have developed at various forecast centers: intermittent and continuous. Figure 3 illustrates an example of a 6-hourly intermittent data assimilation system. Analyses are performed every 6-h at 00, 06, 12, and 18 UTC using data accumulated during a 6-h window surrounding a particular analysis time. The 6-h window allows for delays in transmission of important observations from less accessible regions of the globe. A 6-h forecast from the previous analysis cycle provides first-guess fields for a statistical interpolation method that blends quality-controlled data with the first guess. Depending on the type of statistical interpolation method, analyses are either incorporated directly into a forecast model or initialized before forecasting begins. Forecasts from these analyses are made for operational requirements and to provide a first-guess field for the next analysis cycle. In continuous forward data assimilation, observations are assimilated directly into an assimilation/forecast model in near real time after they are received at a forecast center and subjected to quality control. The data assimilation model is run continuously; and at each time step in the forward integration, new observations are inserted and the analyses are updated and adjusted to the new observations over a period of hours. Through these methods, a nearly continuous analysis of the four-dimensional state of the atmosphere is maintained.

2.3 Forecast Models

Operational numerical weather prediction models solve the governing or primitive equations of atmospheric motion and pro-

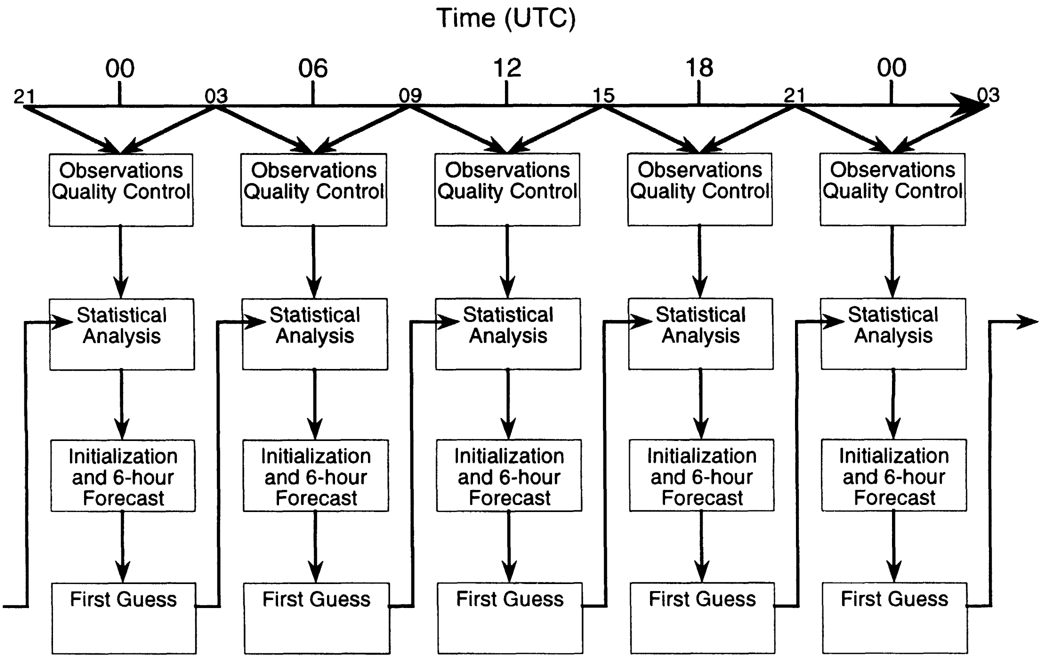

FIG. 3. Schematic illustrating a 6-hourly intermittent data assimilation system. Arrows indicate flow of time, data, and assimilation-system steps.

cesses over domains ranging from the globe to small regions on the order of a few hundred square kilometers (e.g., Haltiner and Williams, 1980; Pielke, 1984). The governing equations include three momentum components, the first law of thermodynamics, mass conservation, and water vapor. As discussed in the article METEOROLOGY AND CLIMATOLOGY, through application of the hydrostatic approximation to the vertical momentum equation, vertical motion becomes a diagnostic quantity, and, more importantly, sound waves and other nonmeteorological motions are filtered while meteorologically important Rossby waves and gravity waves are retained. The resulting equations are often solved using a vertical coordinate system called sigma coordinates (Phillips, 1957). The sigma coordinate is a normalized pressure vertical coordinate that allows the Earth's topography to be incorporated as part of the bottom boundary.

The primitive equations are the horizontal momentum,

$$\frac{D\mathbf{V}}{Dt} = -\nabla\phi - f\mathbf{k} \times \mathbf{V} - RT\nabla(\ln p_s) + \mathbf{F}; \tag{1}$$

hydrostatic,

$$\frac{\partial\phi}{\partial\sigma} = -\frac{RT}{\sigma}; \tag{2}$$

thermodynamic,

$$\frac{DT}{Dt} = \frac{RT\omega}{c_p p_s \sigma} + \frac{H}{c_p}; \tag{3}$$

mass conservation,

$$\frac{Dp_s}{Dt} = -p_s(\nabla\cdot\mathbf{V}) - p_s\frac{\partial\dot{\sigma}}{\partial\sigma}; \tag{4}$$

water vapor,

$$\frac{Dq}{Dt} = E - P; \tag{5}$$

and the definition of the sigma vertical coordinate;

$$\sigma = p/p_s, \tag{6}$$

where the operator D/Dt is the Lagrangian or material derivative following air parcel motions in (x,y,σ) space, ∇ is the two-dimensional gradient operator on a constant-σ surface, $\mathbf{V} = (u,v,0)$ is the horizontal velocity vector, ϕ is geopotential height, f is the Coriolis parameter, $\mathbf{k}$ is the unit vertical vector, R is the gas constant for dry air, T is temperature, p_s is air pressure at the Earth's surface, p is pressure, $\mathbf{F}$ represents frictional effects, c_p is the specific heat at constant pressure, ω is the Lagrangian derivative of pressure, H is heat addition and subtraction per unit mass, $\dot{\sigma}$ is the Lagrangian derivative of the vertical coordinate σ, q is specific humidity, E is evaporation, and P is precipitation. Sigma coordinates, by virtue of incorporating p_s, the air pressure at the Earth's surface, represent the Earth's surface topography and have a value of 1 at the Earth's surface and 0 at the top of the atmosphere. Sources and sinks of the prognostic variables are often referred to as model physics. Model physics includes momentum changes due to frictional effects, energy gain and loss resulting from the transfer of sensible heat at the surface and latent heat release during phase changes of water vapor, precipitation processes, and processes involving short-wave and long-wave radiation.

The lack of an analytical solution for the primitive equations requires numerical techniques to solve the equations at a finite number of locations arranged in a three-dimensional lattice over a particular domain. The three-dimensional atmospheric structure is subdivided as a sequence of horizontal slices arranged on top of each other at discrete intervals in the vertical. Almost all operational models employ the terrain-following vertical coordinate [Eq. (6)] that is based on the ratio of pressure at a level to the pressure at the Earth's surface. Sigma surfaces in forecast models are not regularly spaced; they are closely spaced near the ground and in the upper troposphere to resolve heat and momentum transfers between the atmosphere and the Earth's surface and to resolve the jet stream, the tropopause, and upper tropospheric frontal zones (Fig. 4). Hybrid sigma-pressure and sigma-theta (where theta is potential temperature) vertical coordinates have been developed that combine the terrain-following qualities of sigma coordinates in near-surface layers with the advantages of

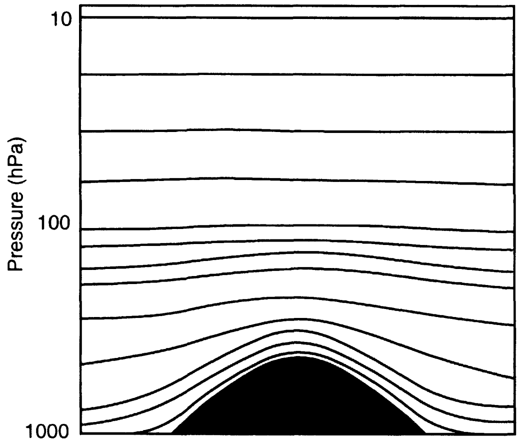

FIG. 4. Schematic showing the vertical distribution of sigma vertical coordinate levels in an operational numerical forecast model containing a mountain at the lower boundary.

pressure or isentropic coordinates in the free atmosphere (CAeM, 1992; Bleck and Benjamin, 1993). The eta vertical coordinate (Mesinger, 1984; Mesinger and Black, 1992), a more generalized version of the sigma vertical coordinate, has also been incorporated in operational weather-prediction models. Instead of defining mountains in sigma coordinates as smoothed terrain, in eta coordinates, mountains are depicted as a sequence of steps with vertical faces that resolve regions of steep topography more accurately. Operational models typically contain between 30 and 40 vertical levels and use finite differencing to approximate vertical derivatives. Terrain height commensurate with the horizontal and vertical resolution of the model is also specified.

On each vertical coordinate surface, meteorological fields are defined as either grid-point values or coefficients of spherical harmonic functions. Regional models use grid-point formulations and employ finite-difference or finite-element methods to approximate horizontal derivatives. Some global domain models also use finite differences but require careful treatment of pole-point singularities in spherical coordinates. Spectral methods have been adopted for use in global models and recently extended to regional models because of advantages in calculating the horizontal derivatives of nonlinear advective terms, the treatment of the pole in spherical geometry, and the ability to implement efficient time-integration methods that allow increased horizontal resolution for similar computation cost. Spectral models represent meteorological variables as amplitudes of spherical harmonic functions of varying wave numbers in the east-west and north-south directions around the globe. These functions are typically truncated at wave numbers 126 or 213 giving approximate equivalent grid-point resolution of under 100 km, rivaling the resolution of many limited-area models. In spectral models, the horizontal momentum equations are reformulated to predict the divergence of the horizontal velocity and vorticity (the vertical component of the curl of the momentum), quantities that are better represented in spherical harmonics. However, model physics is calculated in grid-point space, necessitating transformations between spectral and grid-point representations with each time step of the model.

Time integration of the primitive equations is accomplished by first reformulating the Lagrangian equations into Eulerian partial-derivative format where the time tendency of either momentum, temperature, surface pressure, or water vapor is produced by horizontal and vertical advection and by Lagrangian sources and sinks. The time tendency terms are approximated by time-integration methods such that successive forward jumps in time can be made that satisfy numerical stability considerations (Courant *et al.*, 1928). Advances in time-integration methods have allowed forward time steps to be increased while maintaining numerical stability, with a typical forward time step averaging approximately 1–5 min. A 24-h forecast requires between 300 and 1000 time steps involving numerical solution of the primitive equations over the three-dimensional model lattice. With approximately two million entries in a three-dimensional lattice, a 24-h forecast for five variables over 1000 time steps will require on the order of one trillion calculations. This extraordinary magnitude of calculations to forecast just one day ahead demonstrates the importance of computer limitations such as computational speed, memory, and storage capacity to the operational forecasting process.

The horizontal and vertical lattice spacing of forecast models determines the minimum scale of motion and weather-system size that can be resolved. Weather systems and physical processes smaller than the resolvable scale, such as individual thunderstorms, are not directly predicted by operational forecast models, and the effects of unpredicted sub-grid-scale motions on resolvable circulations are parametrized to maintain physical consistency with the full range of atmospheric motions.

Sub-grid-scale processes that are parametrized are model physics such as surface and boundary-layer processes, radiation and energy-transfer processes, and large-scale stratiform and convective scale precipitation. Surface processes include sources and sinks of momentum, energy, and water vapor that occur over various bottom boundaries such as land, mountains, water, and ice. The importance of momentum and energy sources and sinks makes appropriate specification of the atmosphere–surface interface an important component of forecast models. The surface boundary conditions used in sub-grid-scale calculations are determined from satellite, ship, buoy, and surface observations of water temperature, snow cover, and surface conditions. Sources and sinks of momentum occur through surface friction and the effects of mechanically and thermodynamically generated turbulent eddies near the Earth's surface. Sensible and latent heat transfer are sources and sinks of energy and water vapor with latent heat transfer being a vital component of the hydrological cycle. Soil, surface energy, and water-balance models have been incorporated within forecast models to calculate the redistribution of heat and momentum and changes in static stability more accurately. Boundary effects due to mountains include the contribution of gravity waves, internal atmospheric oscillations that are produced by flow over mountains in a stably stratified atmosphere. These oscillations transfer momentum vertically into the stratosphere, producing a sink of lower-atmospheric momentum. Radiation processes include the diurnal cycle of solar insolation and its attenuation by albedo and cloud effects, while long-wave radiation parametrizations consider the absorption and emission of infrared radiation by greenhouse gases and clouds.

Precipitation is an important source and sink of energy and water vapor and tends to occur on scales smaller than model resolution. Two types of precipitation are parametrized: large-scale stratiform and convective-scale or thunderstorm precipitation. Large-scale stratiform precipitation occurs in a stably stratified atmosphere and results from the large-scale lifting of air until it has cooled to saturation. Large-scale precipitation parametrization algorithms initiate falling snow or rain from a saturated layer that can evaporate as it falls, simulating the gradual descent of precipitation from altostratus clouds in warm fronts. The algorithm computes the amount of condensed water vapor, latent heat release, evaporative cooling, and precipitation reaching the ground. Convective precipitation parametrizations quantify the bulk effects of thunderstorms on the larger-scale mass, momentum, energy, and water vapor fields. These algorithms first determine if an unstable stratification with respect to a saturated atmosphere will lead to buoyant updrafts. If buoyant updrafts are diagnosed, then the method quantifies momentum transfer inside and outside convective clouds, the amount and location of latent energy release, and the amount of precipitation created while adjusting the thermal and water-vapor fields to stabilize the atmosphere.

A number of different convective precipitation methods have been utilized at operational forecast centers to quantify convective precipitation forecasts and the influence of convective scale motions on the atmosphere. Convective scale motions affect tropical and midlatitude prediction because much of the tropical atmosphere responds to differential solar heating through convective overturning. These overturning motions force ascending branches of global-scale circulations in the tropics. Variations of tropical convection change the location and intensity of these tropical circulations, which, in turn, affect the extratropical circulation and short-term climate.

As future states of the atmosphere are calculated, model output is post-processed to create graphics of forecast fields. Figures 5 and 6 show examples from global- and regional-scale models to illustrate the kinds of graphical products available to weather forecasters. A 120-h forecast produced by the global model of the U.S. National Centers for Environmental Prediction is shown in Fig. 5. The forecast shows fields of mean–sea-level

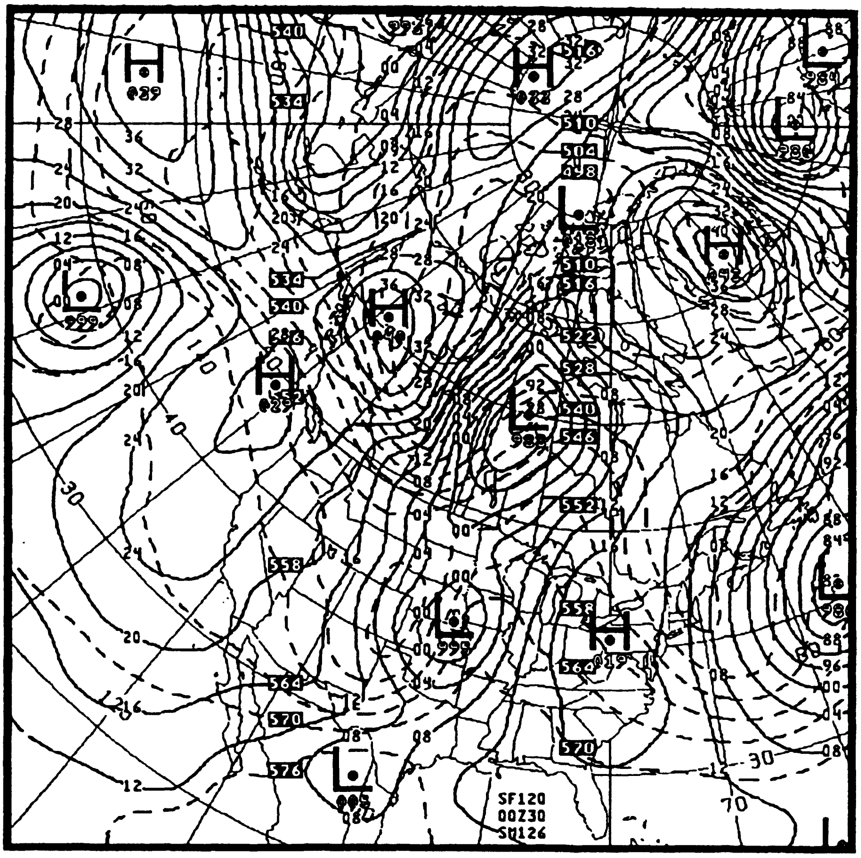

FIG. 5. An operational forecast from the U.S. National Centers for Environmental Prediction Medium Range Forecast Model, a global spectral model with triangular truncation of 126 waves and 28 vertical sigma levels. The forecast was made from initial conditions at 00 UTC 25 October 1996 and is valid 120 h later at 00 UTC 30 October 1996. The figure shows analyses of mean-sea-level pressure (solid contours every 4 mbar with the one's and ten's digit labeled, such that 24 ⇒ 1024 mbar) and the thickness between the 1000- and 500-mbar pressure levels (dashed contours every 6 dekameters and labeled with white numbers on a black rectangle such that 540 ⇒ 5400 m). The thickness field represents the mean temperature in the lower troposphere with higher values indicating warmer conditions. Figure courtesy of U.S. National Centers for Environmental Prediction.

pressure isobars as solid contours every 4 mbar, and the 1000- to 500-mbar thickness contoured every 60 m in dashed contours for a portion of the Northern Hemisphere. The mean sea-level pressure distribution depicts the forecast position and intensity of migratory and semipermanent low- and high-pressure systems while the 1000- to 500-mbar thickness field describes the lower mean tropospheric temperature distribution. Overlaying the two fields allows forecasters to determine the position, movement, and development of fronts, the important process of temperature advection for forecasting temperature variations, clouds, and precipitation, and the relationship of pressure features to the thermal field that allows forecasters to subjectively determine future development of pressure systems.

Forecasts from regional-scale models usually provide more meteorological detail, but for a shorter forecast projection. Figure 6 shows a 48-h forecast from a regional-scale model of the U.S. National Centers for Environmental Prediction. This four-panel forecast displays

1. 500-mbar geopotential height and absolute vorticity isopleths,
2. mean sea-level pressure and 1000- to 500-mbar thickness,

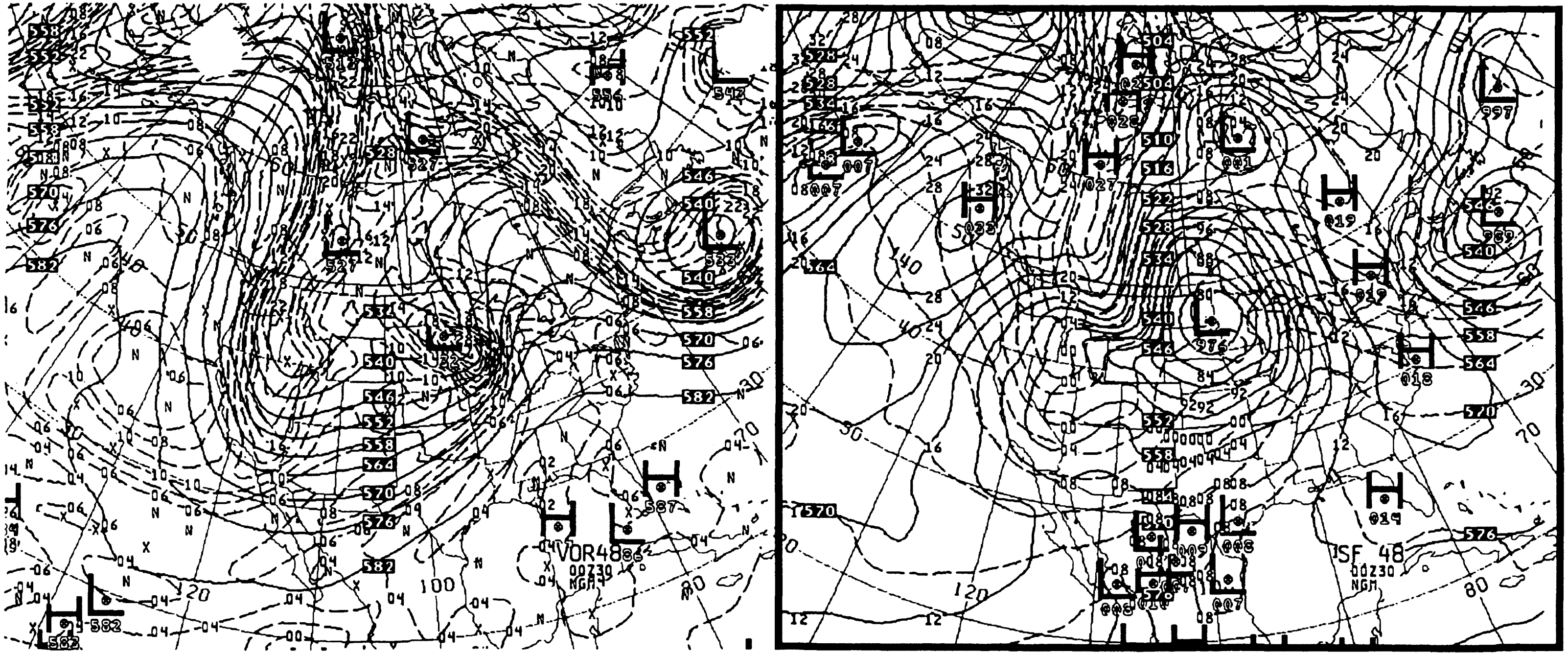
VOR48
00Z30
NGM
SF 48
00Z30
NGM

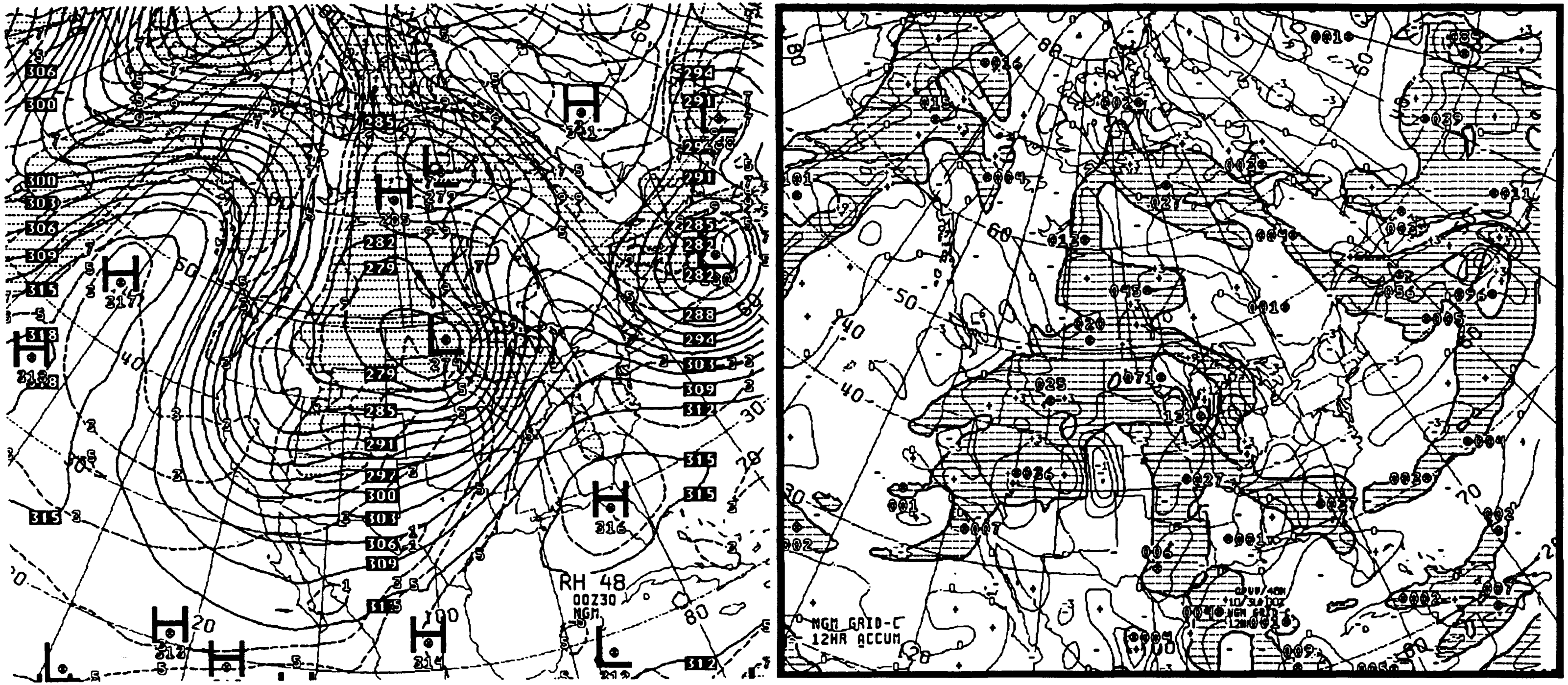

FIG. 6. 48-h operational forecast from initial conditions at 00 UTC 28 October 1996 by the U.S. National Centers for Environmental Prediction Regional Analysis and Forecast System, a hemispheric grid-point model with 80-km horizontal resolution over North America, 200 km over the remainder of the hemisphere, and 16 vertical sigma levels. Upper left panel shows 500-mbar geopotential heights (solid contours every 6 dekameters and labeled with white numbers on a black rectangle such that 540 ⇒ 5400 m) and absolute vorticity (dashed contours every $2 \times 10^{-5}\ s^{-1}$). Upper right panel shows mean-sea-level pressure (solid contours every 4 mbar with the one's and ten's digit labeled, such that 24 ⇒ 1024 mbar) and the thickness between the 1000 and 500 mbar pressure levels (dashed contours every 6 dekameters and labeled with white numbers on a black rectangle such that 540 ⇒ 5400 m). Lower left panel shows 700-mbar geopotential heights (solid contours every 3 dekameters and labeled with white numbers on a black rectangle such that 300 ⇒ 3000 m) and the average relative humidity from the surface to approximately 500 mbar (dashed contours every 20% with values over 70% shaded and the units digit suppressed). Lower right panel shows 700-mbar vertical velocity (thin solid contours every 3 μbar s^{-1}, with "+" denoting ascent and "−" descent) and 12-h accumulated precipitation from 12 UTC 29 October to 00 UTC 30 October [contoured every 12.7 mm (half inch); with shading between 0.254 mm (0.01 in.) and 12.7 mm (half inch); no shading between 12.7 mm (half inch) and 25.4 mm (1 in.); and shading from 25.4 to 38.10 mm (1.5 in.) and alternating every 12.7 mm thereafter]. Precipitation maxima are shown by crosses within a circle and are labeled in hundredths of inches such that 131 ⇒ 1.31 in. = 33.27 mm. Figure courtesy of U.S. National Centers for Environmental Prediction.

3. 700-mbar geopotential heights and 1000- to 500-mbar relative humidity, and
4. 700-mbar vertical velocity and 12-h accumulated precipitation ending at the forecast time.

The familiar mean-sea-level pressure and 1000- to 500-mbar thickness map is included while other forecast fields provide information on meteorological processes such as vorticity advection, a process that produces vertical motion and the development of low- and high-pressure systems. Relative-humidity fields indicate where saturation is forecast and gives an assessment of cloud fields associated with fronts and low-pressure systems. Vertical-velocity forecasts denote where ascent and descent are forecast; this information is important for the forecasting of precipitation and convective scale systems. The 12-h accumulated quantitative precipitation forecast describes the distribution and amounts of precipitation produced by both the stratiform and convective-scale precipitation algorithms. Many other model forecast fields are available from forecast centers worldwide, and they provide a vast amount of information that forecasters use to produce weather forecasts ranging from the next few hours to two weeks.

2.4 Statistical Forecasting

Numerical models provide predictions of model-equation variables such as wind, temperature, specific humidity, geopotential height, quantitative precipitation amount, and pressure. Weather forecasters, however, need to predict meteorological quantities that are not explicitly forecast by the models for public, businesses, government, and transportation interests. These quantities include, for example, visibility, obstructions to visibility, cloud amount, ceiling height, precipitation type and amount, probability of precipitation, thunderstorm probability, maximum and minimum temperatures, frontal windshift, surface wind, clear-air turbulence, snow amount, solar energy, and sunshine. In order to meet these needs, statistical techniques are applied to create forecasts of meteorological variables not forecast by numerical models (Klein *et al.*, 1959; Klein, *et al.*, 1982).

Two approaches commonly utilized in statistical forecasting are perfect prognosis and model-output statistics (Carter *et al.*, 1989; CAeM, 1992). Perfect prognosis defines statistical relationships between observed weather elements and predictor weather elements that are observed concurrently or at previous observation times. For example, a statistical relationship may be defined between observed visibility at time t and observed temperature and dew-point temperature at the same time (t) and/or at a previous time $t - dt$. Once these relationships have been defined through multiple linear-regression equations, the equations are applied with model forecast conditions taking the place of the concurrent or antecedent observations used to define the statistical relationship. Substitution of model prognosis conditions for observed conditions is equivalent to having a perfectly observed prognosis in the statistical relationship. Perfect prognosis relies on the accuracy of the substituted model prognoses and cannot correct for model biases and the diminishment of forecast accuracy with time. Model-output statistics (MOS) is a variation of the perfect prognosis method wherein statistical relationships are defined from predictor weather elements obtained from model forecasts as opposed to observed conditions. In application, MOS is applied to forecast model output, and these forecasts can partly account for model systematic errors and the degradation of forecast accuracy with time.

Application of perfect prognosis or MOS to forecasting individual weather elements requires a development sequence. A meteorological variable to be forecast is first defined on the basis of user needs and the ability of a statistical method to forecast that quantity. The meteorological variable may be a continuous variable such as temperature or a categorical value such as precipitation type that consists of liquid, freezing, or frozen categories. Depending on the choice of perfect prognosis or MOS, each meteorological variable requires an appropriate historical developmental dataset of observations or model forecasts, respectively, from which relationships with potential predictors of that variable can be defined. This training dataset must be large enough to create a sample size that ensures statistical stability of the defined relationships. Next, a particular statistical method such as multivariate linear regression, regression estimation of event probabilities, multiple discriminant analysis, or classification and regression trees is chosen. Equations are developed to quantify the statistical relationships between predictors and the meteorological variable of interest, and these relationships are tested on a training dataset. The equations are next tested on an independent dataset to establish how well these relationships perform in operational forecasting, to measure how well the method forecasts extreme events, and to discover strengths and weaknesses that will require further refinement.

Examples of MOS forecasts produced from global and regional forecast systems of the U.S. National Centers for Environmental Prediction for O'Hare Airport in Chicago, Illinois, are shown in Tables 2 and 3. Global model statistical guidance provides 7-d forecasts of maximum and minimum temperature, 12-h probability of precipitation, a conditional probability of snow, percentage of cloud cover, wind speed, and a 24-h probability of precipitation that can be compared with climatological values. Similar weather elements are forecast for O'Hare from the regional scale model with greater temporal resolution over a 48-h time period. Additional weather elements forecast from the regional scale model are dew-point temperature, cloud amount categories, wind direction, categorical quantitative precipitation amount, thunderstorm probability, categorical precipitation type, probability of freezing precipitation and snow, categorical snow amount and probability, and categorical ceiling, visibility, and obstructions to visibility.

An advantage of statistical forecasting methods is their ability to interpret model forecast variables in terms of operationally important weather elements. Both perfect prognosis and MOS techniques offer strengths and weaknesses. Perfect prognosis techniques suffer from decreases of model forecast skill at greater forecast projections. Perfect prognosis can be applied to different models and models undergoing continual development because the technique employs observational data and is independent of a forecast model. MOS is directly coupled to model forecasts with performance suffering if the model is changed by continual development; however, MOS offers the advantage of partly accounting for model systematic errors. Statistical forecast guidance developed from these two techniques at forecast centers worldwide has been shown to offer skillful forecasts of weather elements (Carter *et al.*, 1989; CAeM, 1992) and is an important element in the human–computer integration in weather forecasting.

2.4.1 Stochastic-Deterministic Forecasting The weather forecasting approach outlined by Bjerknes (1904), applied by Richardson (1922), adopted by world forecasting centers, and described in this article follows an initial-value or deterministic approach to the forecast problem. In this approach it is assumed that accurate specification of the initial state coupled with a sufficiently detailed and accurate forecast model will produce an optimal weather forecast. Over the last 15 years, numerical model resolution and skill have increased to the point where a

Table 2. Statistical or objective forecast produced from the medium-range forecast model of the U.S. National Centers for Environmental Prediction in Washington, D.C. Temperatures have been converted from Fahrenheit and rounded to the nearest whole degree Celsius. Legend: ORD = KORD, O'Hare International Airport, Chicago, IL, USA; MN/MX = minimum/maximum surface temperatures in °C; POP12 = 12-h probability of 0.25 mm or more of liquid equivalent precipitation in %; CPOS = conditional probability of snow in %; CLDS = mean opaque cloud cover in %; WIND = average wind speed in knots; POP24 = 24-h probability of 0.25 mm or more of liquid equivalent precipitation in %; CLIMO = climatological values of forecast variables for the 7-day period.

MRF-based objective guidance 10/05/96 0000 UTC

ORD	SAT 05	SUN 06		MON 07		TUE 08		WED 09		THU 10		FRI 11		SAT 12		CLIMO	
MN/MX	19	11	24	11	18	7	17	8	18	7	17	5	13	3	15	7	19
POP12	0	0	7	30	28	19	15	39	30	20	16	16	21	16	10	23	19
CPOS	0	0	0	0	2	1	4	2	2	2	3	5	14	18	3	4	5
CLDS	31	34	48	54	59	46	56	41	62	43	56	39	44	40	45	47	52
WIND	10	9	11	7	9	7	8	6	10	7	11	8	11	8	10	7	9
POP24		7		43		25		51		27		28		20		33	

Table 3. Statistical or objective forecasts produced from the nested-grid model (NGM) of the U.S. National Centers for Environmental Prediction in Washington, D.C. Temperatures have been converted from Fahrenheit and rounded to the nearest whole degree Celsius. Legend: ORD = KORD, O'Hare International Airport, Chicago, IL, USA; HOUR = time in UTC; MN/MX = minimum/maximum surface temperatures for the 12-h period in °C; TEMP and DEWPT = time-specific surface temperature and dew-point temperature in °C valid every 3 h; CLDS = categorical amount of opaque cloud cover with SC = scattered (1 to 5 tenths opaque clouds), BK = broken (6 to 9 tenths opaque clouds), and OV = overcast; WDIR = surface wind direction in degrees true; WSPD = surface wind speed in knots; POP06 = 6-h probability of 0.25 mm or more of liquid-equivalent precipitation in %; POP12 = 12-h probability of 0.25 mm or more of liquid-equivalent precipitation in %; QPF = categorical quantitative precipitation forecast; TSV06 = thunderstorm/conditional probability of a severe thunderstorm for the 6-h period in %; TSV12 = thunderstorm/conditional probability of a severe thunderstorm for 12 h in %; PTYPE = categorical precipitation type, R = rain; POZP = condition probability of freezing precipitation such as freezing rain or glaze, sleet or ice pellets, or snow grains in tens of %; POSN = conditional probability of snow in tens of %; SNOW = categorical forecast of snow amount for 6/12-h periods; CIG = categorical forecast of cloud ceiling; VIS = categorical forecast of visibility; OBVIS = obstructions to visibility, N = none, F = fog, H = haze.

```
ORD  EC   NGM MOS Guidance  10/05/96  1200 UTC
DAY /OCT   5 /OCT  6                 /OCT   7                 /
HOUR   18 21 00 03 06 09 12 15 18 21 00 03 06 09 12 15 18 21 00
MN/MX                    14          24          10          15
TEMP   19 21 18 17 16 16 16 18 22 23 19 17 14 12 11 13 14 14 12
DEWPT   6  7  9 10 11 12 12 13 13 13 12 12 11 10  8  8  7  6  6
CLDS   SC SC SC SC SC SC BK BK BK SC BK OV OV OV OV OV OV OV BK
WDIR   17 17 15 18 20 21 22 22 21 22 21 24 30 35 35 34 34 35 01
WSPD   10 10 09 12 11 12 13 16 15 15 12 11 13 14 11 11 12 09 12
POP06         0     0     2    12    29    35    26    17    12
POP12                     2          30          42          21
QPF         0/    0/   0/0   0/    0/0   0/    0/1   0/    0/0
TSV06      3/ 5  7/ 3  0/ 2  3/ 0 16/18 33/ 6 22/ 1 11/ 0  0/ 0
TSV12            6/ 7        3/ 2       32/22       24/ 0
PTYPE   R  R  R  R  R  R  R  R  R  R  R     R     R     R     R
POZP    0  0  0  2  0  2  4  0  0  0  0     0     0     0     0
POSN    2  0  1  0  0  0  0  0  1  0  0     0     0     5     1
SNOW        0/    0/   0/0   0/    0/0   0/    0/0   0/    0/0
CIG     7  7  7  7  7  7  6  7  7  7  7     5     4
VIS     5  5  5  5  5  5  5  5  5  5  5     5     5
OBVIS   N  N  N  N  N  N  F  H  H  N  N     F     N
```

6-h first-guess forecast can now indicate observational errors. However, accurate specification of the initial state is hindered by geographical and technical limitations of the global observing system and approximations within data assimilation systems and forecast models, as well as the inherent nonlinearity of the forecast equations. Thus the "best" weather forecast is only an estimate of the atmosphere's future state.

Lorenz (1963) pointed out that even with a perfect model, inability to specify initial conditions accurately results in increasingly divergent (with time) forecasts from slightly different initial conditions, producing a range of possible forecast outcomes. Recognition of the chaotic nature of apparently deterministic dynamical systems such as the atmosphere, and of initial analysis uncertainties that reduce forecast skill with increasing forecast projection, led to the adoption of a stochastic-deterministic viewpoint of atmospheric prediction (Lorenz, 1965, 1969; Tracton and Kalnay, 1994; Toth and Kalnay, 1993; Molteni *et al.*, 1996). Implementation of this viewpoint, called ensemble prediction, is based on multiple predictions from slightly different initial conditions with a deterministic forecasting system. If the ensemble is large enough and is generated by reasonably accurate forecast models from initial conditions that are representative of initial analysis uncertainty, the ensemble average will yield a better forecast than one high-resolution control forecast. The ensemble-average forecast is more accurate because it is assumed that varying initial analysis errors will tend to cancel in a large ensemble average. Ensemble forecasting provides a statistical measure of forecast uncertainty by quantifying the range of multiple forecasts and defining the probability of occurrence of a particular forecast through clustering and averaging of multiple forecasts. Knowledge of forecast spread and probability of occurrence of various scenarios allows forecasters to assign confidence to portions of their weather forecasts.

Ensemble forecasting requires knowledge of analysis errors. Unfortunately, analysis errors are not known, and different techniques are employed to estimate them at world forecast centers. One approach assumes that the most important component of the analysis error results from the most energetic and rapidly growing errors. These errors are introduced in the data assimilation cycle, which contains errors from observations and the model-derived first-guess forecast, and grow faster than amplifying weather systems to contaminate forecasts. Growing errors are estimated by first identifying analysis differences between arbitrarily perturbed and unperturbed forecasts and then using these differences to produce perturbations in the subsequent analysis cycle. By continuing this cycle in time and varying the initial arbitrary perturbation, this technique breeds estimates of the fastest-growing analysis errors. A second approach employs mathematical techniques to estimate the maximum-growing analysis error structures through consideration of model formulation and perturbation energy. Initial analysis perturbations are then dynamically defined from the maximum-growing error structures. In both approaches the estimated perturbations are applied to vary initial conditions for the ensemble members.

Ensemble forecasts are routinely produced at forecast centers in Europe and North America for both short-range (0–48 h, Brooks *et al.*, 1995) and long-range forecasts with a variety of limited-area and global-domain models. Ensembles of 10–50 members are currently produced at individual operational forecast centers; and as computer speed and capacity increase, model resolution along with the number of ensemble members will multiply. Forecast centers are sharing ensembles and finding that the larger sample size of shared ensembles performs better than an individual forecast-center ensemble based on a single model and data assimilation system. This result supports increased cooperation between world forecast centers because model formulation and data assimilation system diversity improves the definition of atmospheric analysis uncertainty and the ability to establish an envelope of future atmospheric states.

A challenge in ensemble forecasting is the communication of multidimensional information to forecasters. The time domain is shown through multiple maps and clusters of forecast fields while uncertainty is portrayed by probability maps of meteorological variables. An example of the time variation and probability envelope of 500-mbar geopotential height for a 7-d forecast produced by a 17-member ensemble of the United States National Centers for Environmental Prediction for the Northern Hemisphere is shown in Fig. 7. The figure is an example of limited contour analysis that portrays midlatitude and subtropical circulations associated with the 564- and 585-dekameter geopotential height contours of the 500-mbar pressure surface. Thin light contours at higher latitudes represent individual members of the 564-dekameter contour with thin darker contours at lower latitudes depicting members of the 585-dekameter contour. The ensemble indicates confidence in the forecast where the thin lines of a particular height contour are close together and less confidence where the thin lines show large divergence from each other. Separation and orientation of both contours with respect to each other indicates the geostrophic wind speed and direction. The figure also includes climatological values (thickest contours) and the high-resolution control forecast (medium thickness contours) of the two contours. This day-7 forecast is fairly confident of the midlatitude circulation over the North Pacific Ocean, North America, and the Atlantic Ocean. Divergent 564-dekameter contours over Europe suggest less confidence in the forecast, with possible solutions ranging from warm and dry to cool and stormy. The range of solutions for the subtropical circulation indicates confidence over northern Africa and southern Asia, and relatively high confidence regarding a ridge over southern North America and the flow over the western North Atlantic Ocean. Increasing uncertainty is noted in the trough/ridge structure over the central North Pacific Ocean, with much uncertainty concerning the subtropical cir-

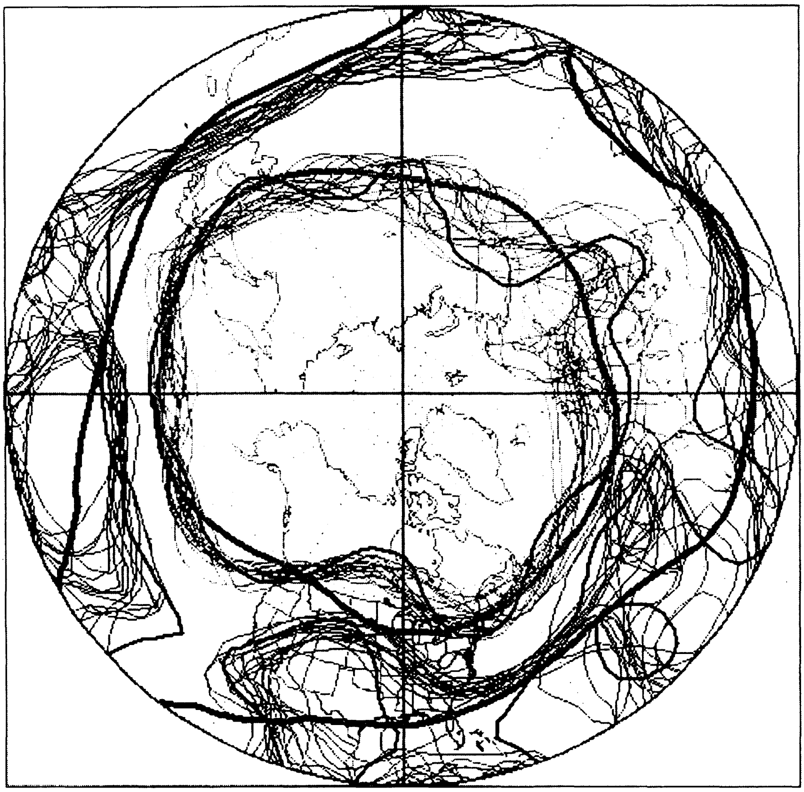

FIG. 7. Ensemble forecast of day-7 500-mbar geopotential heights over the Northern Hemisphere produced by the U.S. National Centers for Environmental Prediction Ensemble Prediction System and created by the United States NOAA Environmental Research Laboratory Climate Diagnostics Center (http://www.cdc.noaa.gov). The ensemble contains 17 members all valid at 00 UTC 14 October. Twelve members are from 00 UTC runs of the Medium Range Forecast Model (MRF) and 5 from 12 UTC runs of the aviation version of the MRF. The 564-dekameter height contours are shown as thin light lines at higher latitudes and the 585-dekameter height contours are shown as darker thin lines at lower latitudes. Climatological values of these contours are shown by the thickest line and the high-resolution control-run forecast is shown by the medium-thickness lines. The ensemble indicates confidence in the forecast where the thin lines of a particular height contour are close together and less confidence where the thin lines show large divergence from each other.

culation over the eastern North Atlantic Ocean. Graphical products such as Fig. 7 provide forecasters with meteorological information to assign confidence to a particular forecast.

3. FORECAST EVALUATION

3.1 Model Performance and Predictability

Forecasts made by operational centers are routinely evaluated to measure the long-term trend in forecast skill, impacts of changes to the forecast system, and the performance of a particular forecast. Evaluations consist of verification statistics of meteorological fields by a variety of scoring techniques. A result of these evaluations has demonstrated that the investment in numerical weather prediction systems at world forecast centers during the past few decades has produced significant improvement in weather forecasts.

An example of this prodigious improvement is shown in Fig. 8, a time-series graph of the skill of 36-h forecasts of 500-mbar geopotential height over North America. The skill score shown in Fig. 8 is a modified version of the S1 skill score of Teweles and Wobus (1954), which indicates improved skill with a higher score. The figure shows a dramatic increase in the ability to forecast

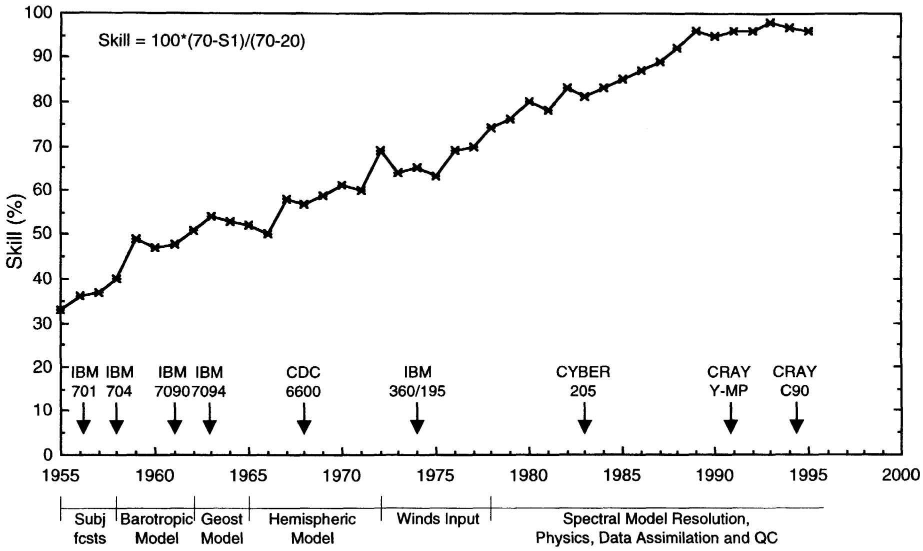

FIG. 8. Skill of 36-h forecasts of the 500-mbar geopotential height field over North America in forecasts made by the U.S. National Centers for Environmental Prediction. Noted on the graph are acquisition dates and types of computing systems and the type of operational-forecasting technique or numerical model. Figure courtesy of E. Kalnay, U.S. National Centers for Environmental Prediction.

the 500-mbar circulation from approximately 33% to a sustained 95% correct since 1989. The skill score may also be viewed from the perspective of the length of predictive skill. For example, as measured by the correlation of circulation anomalies between forecast and observed maps, current 5-d forecasts of 500-mbar circulation are as skillful as 36-h forecasts made in the 1960s (Peixoto and Oort, 1992). These measures also indicate that forecasts of the 500-mbar circulation show predictive skill to day 7, and forecasts of the stratospheric circulation (50 mbar) exhibit skill past day 10 (Bengtsson, 1991).

A result of this lengthening of the time of forecast skill is the positive impact and increased confidence society places in weather forecasts. A prominent example of this positive impact is the major winter snowstorm of 12–14 March 1993, which paralyzed the eastern seaboard of North America with heavy snowfall, blizzard conditions, coastal flooding, high winds, and tornadoes (Kocin *et al.*, 1995; Uccellini *et al.*, 1995; Caplan, 1995). Achievements of forecasts made by models from operational forecasting centers in Europe, Asia, and North America were

1. indications of a major winter storm in the correct time and place as early as 5 d beforehand,
2. 3-d warning of the strength and severe nature of the storm, and
3. 2-d warning of the areal extent and amount of snow.

Forecasters were able to warn the public well in advance of the storm such that this particular case heralds the longest lead time for warning of a severe storm in the history of the U.S. National Weather Service.

The increase in forecast skill has not been evenly distributed over the globe, which partially reflects the type and distribution of observations shown in Fig. 1. Bengtsson (1991) showed that the limit of predictability of circulation and temperature forecasts for the Southern Hemisphere lagged the Northern Hemisphere by approximately 2 d, which is suggestive of the reduced spatial sampling by the global observing system in the Southern Hemisphere. Errors are also larger over the eastern ocean basins and western continents of the Northern Hemisphere midlatitudes (Arpe and Klinker, 1986; CAeM, 1992) connoting the lack of upstream observations over the oceans and the ability of forecast models to simulate, accurately, cyclone development, maturation, and decay along the two primary midlatitude oceanic storm tracks.

Quantitative precipitation is a difficult problem in numerical weather prediction because precipitation occurs on time and space scales smaller than the circulations that produce precipitation (e.g., Ray, 1986). Progress in precipitation forecasting has been achieved through systematic increases in model horizontal and vertical resolution and through improvements in precipitation algorithms. The quality of areal precipitation forecasts produced by two regional scale models of the U.S. National Centers for Environmental Prediction, the 80-km/16-level Regional Analysis and Forecast System and the recently implemented 29-km/50-level Mesoeta model, are shown in Fig. 9. The quality is measured by the equitable threat score, which indicates the percentage of grid points that were correctly forecast to receive a particular threshold amount of precipitation. Figure 9 indicates that the accuracy of areal precipitation forecasts is relatively low and decreases with increasing precipitation amount. However, it is encouraging to note that the accuracy for heavy precipitation amounts has nearly doubled for the Mesoeta model. Since most heavy-precipitation events are associated with convective-type precipitation, the increase in accuracy indicates the importance of high model resolution and appropriate precipitation algorithms to capture precipitation processes better.

An example illustrating the issues discussed in the verification section is shown through the verification of the two forecasts presented in Figs. 5 and 6. Figure 10 shows the verifying analyses of mean sea-level pressure and 1000- to 500-mbar thickness for 00 UTC 30 October produced by the Regional Analysis and Forecast System of NCEP. Comparison of Fig. 10 with the 5-d forecast shown in Fig. 5 indicates that the spatial distribution of high- and low-pressure systems was predicted reasonably well over North America and the adjacent oceans. Intensity forecasts of individual systems are fairly accurate; however, the strength of the major

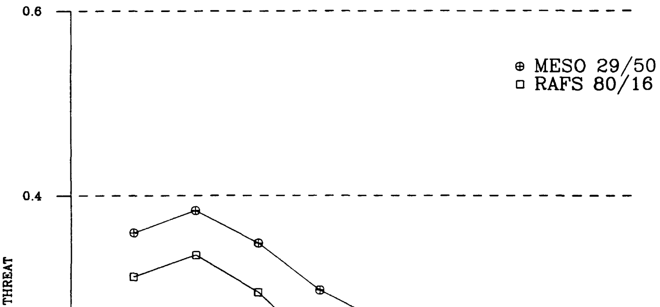

FIG. 9. Equitable threat scores for precipitation from two regional-scale forecast models from the U.S. National Centers for Environmental Prediction for the period 1 March 1995 to 29 February 1996. Square data points are for the Regional Analysis and Forecast System running at 80-km horizontal resolution and 16 vertical sigma levels (RAFS 80/16). Circled data points are for the mesoscale Eta model running at 29-km horizontal resolution and 50 vertical Eta levels (MESO 29/50). The equitable threat score is on the ordinate. The abscissa contains two lines of information. The top line of the abscissa shows threshold categories of precipitation (in inches) that are defined for computation of the equitable threat score. The second line of the abscissa indicates the number of model grid points that received this threshold amount of precipitation. An 80-km grid from the NCEP Eta model was used to compute the scores for both models. For example, 233 568 grid points were forecast to receive at least 0.01 in. of precipitation during the period 1 March 1995 to 29 February 1996. The equitable threat score is calculated separately for each threshold category with a higher score indicating increasing skill. Figure courtesy of E. Kalnay, U.S. National Centers for Environmental Prediction.

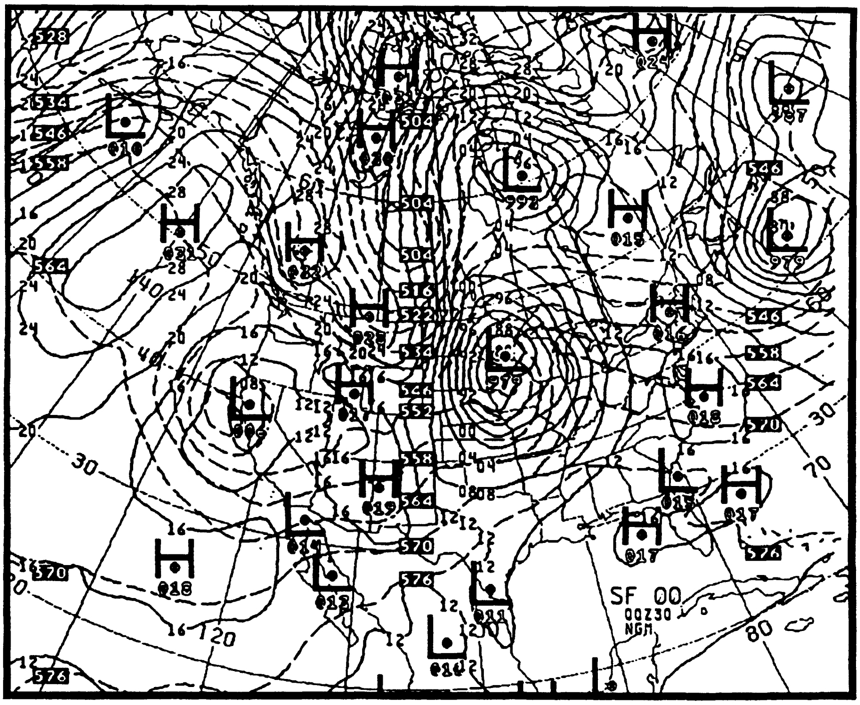

FIG. 10. Analysis from the U.S. National Centers for Environmental Prediction Regional Analysis and Forecast System for 00 UTC 30 October 1996. Shown are mean-sea-level pressure (solid contours every 4 mbar with the one's and ten's digit labeled, such that 24 ⇒ 1024 mbar) and the thickness between the 1000- and 500-mbar pressure levels (dashed contours every 6 dekameters and labeled with white numbers on a black rectangle such that 540 ⇒ 5400 m). Figure courtesy of U.S. National Centers for Environmental Prediction.

low-pressure system over the central United States is underforecast, and a low-pressure system over California is not forecast at all. These problems illustrate difficulties in forecasting downstream of a nearly data-void ocean because the synoptic features of the major storm in the central United States and over California were over the central Pacific Ocean 5 d earlier at initialization time for Fig. 5. Comparison of Fig. 10 with the upper right panel of Fig. 6 reveals that the 48-h forecast from the regional scale model performed well with regard to the intensity and position of the low- and high-pressure systems over the map domain. In particular, the difficulty in accurately predicting the low-pressure system over California exemplifies the problems faced by west-coast forecasters in Europe and North America with the intensity and timing of synoptic scale weather systems.

The advancements made during the past century have realized the vision of Vilhelm Bjerknes and L. F. Richardson to forecast the weather on the basis of the principles of physics. Weather forecasting has made scrupulous use of many branches of physics and mathematics to solve the problems of initial conditions, to create accurate forecast models, and to determine the range of forecast probability. These advances have increased the length of forecast skill to slightly greater than 1 week at present. Lorenz (1969) estimated the limit of deterministic predictability to lie between 2 and 3 weeks because of inherent limitations in the ability accurately to determine and initialize all scales of motion. Progress toward that goal has been substantial during the past decades, and with the development of ensemble prediction systems, it will be possible to assess a probabilistic range of weather forecasts for short-range (1 to 2 weeks) and short-term climate for future seasons (McPherson, 1994).

GLOSSARY

Absolute Vorticity: Rotation about a local vertical axis that includes wind motion relative to the Earth and the Coriolis parameter, which is the amount of rotation a local vertical axis makes with respect to the Earth's axis rotation.

Albedo: The amount of incident radiation that is reflected.

Altostratus Cloud: A middle-level cloud located between 3000 and 6000 m characterized by a layered structure and a dull gray color.

Analog: A weather pattern that bears similarities to another pattern.

Anticyclone: A high-pressure system that circulates clockwise in the Northern Hemisphere and counterclockwise in the Southern Hemisphere.

Baroclinic Instability: A theory that describes how infinitesimal perturbations in a flow characterized by temperature gradients and vertical changes in wind speed and direction can grow exponentially.

Barotropic Vorticity Equation: An equation that describes how changes in the rotation of the wind can change in an atmosphere characterized by no temperature gradients. The equation states that the absolute vorticity is conserved following the motion of an air parcel.

Boundary Layer: The layer of the atmosphere, typically 1–2 km thick, that is closest to the ground and conveys the effects of friction and of sensible and latent heat between the ground and the atmosphere.

Clear-Air Turbulence: Atmospheric flow in which the velocity changes direction and magnitude rapidly and erratically in air that is free from clouds.

Convective Storms and Precipitation: Storms and precipitation produced by buoyant updrafts and characterized by strong vertical motions in statically unstable air. Thunderstorms are a subset of convective storms.

Coriolis Parameter: A quantity that expresses the amount of rotation a local vertical axis makes with respect to the Earth's axis of rotation.

Cyclone: A low-pressure system that circulates counterclockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere. There are tropical cyclones and midlatitude cyclones and each is produced and maintained by different atmospheric processes.

Dew-Point Temperature: The temperature to which air must be cooled in order for saturation and the formation of dew to occur. The dew point is a temperature measure of the amount of water vapor in the air.

Discretization: The process whereby a continuous variable is subdivided into distinct measurements at particular points in space and time.

Divergence: A concept that describes the magnitude of air parcels separating from each other or coming together in a particular region or volume of air.

Dropwinsonde: An instrument package typically dropped from an airplane and carried down by parachute that measures temperature, relative humidity, and wind speed and direction, and telemeters these measurements to a receiving site.

Equivalent Barotropic Model: A forecast model based on the barotropic vorticity equation in which wind speed but not direction is allowed to vary with elevation.

Finite Differencing: A numerical method developed from Taylor series expansion that expresses continuous derivatives as the quotient of differences of the variables contained within the derivative.

Geostrophic Thermal Wind: A concept that relates changes of the geostrophic wind with height to the strength of the horizontal temperature gradient.

Geostrophic Wind: An approximation to the true wind where the wind speed is proportional to the strength of the horizontal pressure gradient and blows at a constant speed parallel to straight isopleths of pressure. The geostrophic wind is defined as the balance between the horizontal pressure gradient and the Coriolis forces.

Gravity Waves: A wave disturbance produced by the restoring force of gravity and exhibiting oscillatory motion about a level due to the displacement of an air parcel from a state of hydrostatic balance.

Hydrostatic Balance: A state where vertical accelerations of air parcels are absent because of the balance between gravity and the vertical pressure gradient force.

Initial-Value Problem: A class of problems in which the solution proceeds from initial values.

Numerical Weather Prediction: Weather prediction that uses numerical approximations to integrate the laws of physics forward in time.

Objective Analysis: Techniques that interpolate irregularly spaced meteorological observations to a regularly spaced lattice.

Polar-Front Cyclone: A low-pressure system in middle latitudes that forms and develops along the polar front, a thermal boundary in the midlatitudes that separates polar air from milder midlatitude air.

Quasi-geostrophic: An approximation to the governing equations of atmospheric flow in which the flow is assumed to be in nearly hydrostatic and geostrophic balance. Departures from hydrostatic and geostrophic balance allow for the development of vertical motion, divergence, cyclones, and anticyclones.

Rawinsonde: An instrument package carried aloft by balloon that measures temperature and relative humidity, telemeters these measurements to a ground station, and is tracked to measure wind speed and direction.

Ridge: A feature of a contour field in which an axis of local maximum exists.

Squall Lines: Convective storms organized in a line.

Static Stability: A measure of an air parcel's resistance to vertical displacements. Positive static stability means that an air parcel displaced from its original position will return to that position due to the restoring force of gravity. Static stability can change as a result of temperature and water vapor changes in the atmosphere.

Stratiform Precipitation: Precipitation that is produced from the widespread lifting of air under statically stable conditions.

Synoptic: Depiction, viewing, and analysis of many synchronous weather observations.

Temperature Advection: A process in which air parcels of different temperature are blown across a region.

Trough: A feature of a contour field in which an axis of local minimum exists.

Vorticity: Rotation of the wind about a local vertical axis.

Works Cited

Arpe, K., Klinker, E. (1986), *Quart. J. Roy. Meteorol. Soc.* **112,** 181–202.

Bader, M. J., Forbes, G. S., Grant, J. R., Lilley, R. B. E., Waters, A. J. (1995), *Images in Weather Forecasting,* Cambridge, U.K.: Cambridge Univ. Press.

Baker, W. E., Emmitt, G. D., Robertson, F., Atlas, R. M., Molinari, J. E., Bowdle, D. A., Paegle, J., Hardesty, R. M., Menzies, R. T., Krishnamurti, T. N., Brown, R. A., Post, M. J., Anderson, J. R., Lorenc, A. C., McElroy, J. (1995), *Bull. Am. Meteorol. Soc.* **76,** 869–888.

Bengtsson, L. (1991), *Quart. J. Roy. Meteorol. Soc.* **117,** 855–902.

Benjamin, S. J., Brewster, K. A., Brummer, R. L., Jewett, B. F., Schlatter, T. W., Smith, T. L., Stamus, P. L. (1991), *Mon. Weath. Rev.* **119,** 888–906.

Bergeron, T. (1959), in: B. Bolin (Ed.), *The Atmosphere and The Sea in Motion,* New York: Rockefeller Institute Press, 440–474.

Berghthórsson, P., Döös, B. (1955), *Tellus* **7,** 329–340.

Bjerknes, J., Solberg, H. (1922), *Geophys. Publisk.* **3,** 1–17.

Bjerknes, V. (1904), *Meteorol. Z.* **21,** 1–7 (English translation by Y. Mintz).

Bleck, R., Benjamin, S. G. (1993), *Mon. Weath. Rev.* **121,** 1770–1785.

Brooks, H. E., Tracton, M. S., Stensrud, D. J., DiMego, G., Toth, Z. (1995), *Bull. Am. Meteorol. Soc.* **76,** 1617–1624.

Businger, S., Chiswell, S. R., Bevis, M., Duan, J., Anthes, R. A., Rocken, C., Ware, R. H., Exner, M., VanHove, T., Solheim, F. S. (1996), *Bull. Am. Meteorol. Soc.* **77,** 5–18.

CAeM Working Group on Advanced Techniques Applied to Aeronautical Meteorology (1992), *Methods of Interpreting Numerical Weather Prediction Output for Aeronautical Meteorology,* Technical Note No. 195, Geneva: World Meteorological Organization.

Caplan, P. M. (1995), *Bull. Am. Meteorol. Soc.* **76,** 201–212.

Carter, G. M., Dallavalle, J. P., Glahn, H. R. (1989), *Weath. Forecasting* **4,** 401–412.

Charney, J. (1947), *J. Meteorol.* **4,** 135–162.

Charney, J. (1948), *Geophys. Publish.* **17,** No. 2.

Charney, J. (1949), *J. Meteorol.* **6,** 371–385.

Charney, J. (1950), *Bull. Am. Meteorol. Soc.* **31,** 231–236.

Charney, J. (1951), in: T. F. Malone (Ed.), *Compendium of Meteorology,* Boston: American Meteorological Society, 470–482.

Charney, J. (1955), *Tellus* **7,** 22–26.

Charney, J., Phillips, N. A. (1953), *J. Meteorol.* **10,** 71–99.

Charney, J., Fjørtoft, R., von Neumann, J. (1950), *Tellus* **2,** 237–254.

Collins, W., Gandin, L. (1990), *Mon. Weath. Rev.* **118,** 2752–2767.

Courant, R., Friedrichs, K., Lewy, H. (1928), *Math. Ann.* **100,** 32–74.

Cressman, G. P. (1959), *Mon. Weath. Rev.* **87,** 367–374.

Daley, R. (1991), *Atmospheric Data Analysis,* Cambridge, U.K.: Cambridge Univ. Press, 457 pp.

Derber, J. C., Parrish, D. F., Lord, S. J. (1991), *Weath. Forecasting* **6,** 538–547.

Dey, C. H. (1989), *Weath. Forecasting* **4,** 297–312.

Doviak, R. J., Zrnic, D. S. (1993), *Doppler Radar and Weather Observations,* San Diego: Academic.

Eliassen, A. (1949), *Geophys. Publish.* **17,** No. 3.

Ferriday, J. G., Avery, S. K. (1994), *J. Appl. Meteorol.* **33,** 1587–1596.

Fujita, T. T. (1986), in: P. S. Ray (Ed.), *Mesoscale Meteorology and Forecasting,* Boston: American Meteorological Society, 18–35.

Gandin, L. (1963), *Objective Analysis of Meteorological Fields,* Leningrad: Gidrometeorologicheskoe Izdatelstvo [translation (1965), Jerusalem: Israel Program for Scientific Translations].

Gandin, L. (1988), *Mon. Weath. Rev.* **116,** 1137–1156.

Ghil, M., Malanotte-Rizzoli, P. (1991), *Adv. Geophys.* **33,** 141–266.

Gilchrist, B., Cressman, G. P. (1954), *Tellus* **5,** 309–318.

Glahn, H. R., Lowrey, D. A. (1972), *J Appl. Meteorol.* **11,** 1203–1211.

Godske, C. L., Bergeron, T., Bjerknes, J., Bundgaard, R. C. (1957), *Dynamic Meteorology and Weather Forecasting,* Boston and Washington: American Meteorological Society: Carnegie Institution of Washington, 619–790.

Golden, J. H., Serafin, R., Lally, V., Facundo, J. (1986), in: Peter S. Ray (Ed.), *Mesoscale Meteorology and Forecasting,* Boston: American Meteorological Society, 50–69.

Haltiner, G. J., Williams, R. T. (1980), *Numerical Prediction and Dynamic Meteorology,* New York: Wiley.

Hollingsworth, A., Shaw, D. B., Lönnberg, P., Illari, L. (1986), *Mon. Weath. Rev.* **114,** 861–879.

Kidder, S. Q., Vonder Harr, T. H. (1995), *Satellite Meteorology,* San Diego: Academic, 466 pp.

Klein, W. H., Lewis, B. M., Enger, I. (1959), *J. Meteorol.* **16,** 672–682.

Klein, W. H., Lewis, B. M., Enger, I. (1982), *Bull. Am. Meteorol. Soc.* **63,** 170–177.

Kocin, P. J., Schumacher, P. N., Morales, R. F. Jr., Uccellini, L. W. (1995), *Bull. Am. Meteorol. Soc.* **76,** 165–182.

Kutzbach, G. (1979), *The Thermal Theory of Cyclones,* Boston and Washington: American Meteorological Society.

Lorenc, A. C., Bell, R. S., Macpherson, B. (1991), *Quart. J. Roy. Meteorol. Soc.* **117,** 59–89.

Lorenz, E. N. (1963), *J. Atmos. Sci.* **20,** 130–141.

Lorenz, E. N. (1965), *Tellus* **17,** 321–333.

Lorenz, E. N. (1969), *Tellus* **21,** 287–307.

Lynch, P. (1992), *Bull. Am. Meteorol. Soc.* **73,** 35–47.

McPherson, R. D. (1994), *Bull. Am. Meteorol. Soc.* **75,** 363–373.

Mesinger, F. (1984), *Riv. Meteorol. Aeronautica* **44,** 195–202.

Mesinger, F., Black, T. L. (1992), *Meteorol. Atmos. Phys.* **50,** 47–60.

Molteni, R., Buizza, R., Palmer, T. N., Petroliagis, T. (1996), *Quart. J. Roy. Meteorol. Soc.* **122,** 73–119.

Palmén, E., Newton, C. W. (1969), *Atmospheric Circulation Systems,* San Diego: Academic.

Panofsky, J. (1949), *J. Meteorol.* **6,** 386–392.

Parrish, D. F., Derber, J. C. (1992), *Mon. Weath. Rev.* **120,** 1747–1763.

Peixoto, J. P., Oort, A. H. (1992), *Physics of Climate,* New York: American Institute of Physics.

Pepin, N. (1996), *Weather* **51,** 242–248.

Petterssen, S. (1940), *Weather Analysis and Forecasting,* New York: McGraw-Hill.

Petterssen, S. (1957), *Meteorol. Monogr.* **3,** 114–151.

Phillips, N. A. (1957), *J. Meteorol.* **14,** 14–37.

Pielke, R. A. (1984), *Mesoscale Meteorological Modelling,* San Diego: Academic.

Ray, P. S. (1986), *Mesoscale Meteorology and Forecasting,* Boston: American Meteorological Society.

Richardson, L. F. (1922), *Weather Prediction by Numerical Process,* Cambridge, U.K.: Cambridge Univ. Press.

Riehl, H., *et al.* (1952), *Meteorol. Monogr.,* No. 5, Boston: American Meteorological Society.

Rossby, C.-G. (1938), *J. Marine Res.* **1,** 239–263.

Rossby, C.-G. (1939), *J. Marine Res.* **2,** 38–55.

Rossby, C.-G. (1940), *Quart. J. Roy. Meteorol. Soc.* **66** (Suppl.), 68–87.

Scofield, R. A., Purdom, J. F. W. (1986), in: P. S. Ray (Ed.), *Mesoscale Meteorology and Forecasting,* Boston: American Meteorological Society, 118–150.

Shuman, F. G. (1989), *Weath. Forecasting* **4,** 286–296.

Shuman, F. G., Hovermale, J. B. (1968), *J. Appl. Meteorol.* **7,** 525–547.

Staff Members, Inst. of Meteorology, University of Stockholm (1952), *Tellus* **4,** 21–30.

Strauch, R. G., Weber, B. L., Frisch, A. S., Little, C. G., Merritt, D. A., Moran, K. P., Welsh, D. C. (1987), *J. Atmos. Oceanic Technol.* **4,** 563–571.

Sutcliffe, R. C., Forsdyke, A. G. (1950), *Quart. J. Roy. Meteorol. Soc.* **76,** 189–217.

Teweles, S. Jr., Wobus, H. B. (1954), *Bull. Am. Meteorol. Soc.* **35,** 455–463.

Thompson, P. D. (1983), *Bull. Am. Meteorol. Soc.* **64,** 755–769.

Toth, Z., Kalnay, E. (1993), *Bull. Am. Meteorol. Soc.* **74,** 2317–2330.

Tracton, S., Kalnay, E. (1994), *Weath. Forecasting* **8,** 379–398.

Uccellini, L. W., Kocin, P. J., Schneider, R. S., Stokols, P. M., Dorr, R. A. (1995), *Bull. Am. Meteorol. Soc.* **76,** 183–199.

Ware, R., Exner, M., Feng, D., Gorbunov, M., Hardy, K., Herman, B., Kuo, Y., Meehan, T., Melbourne, W., Rocken, C., Schreiner, W., Sokolovskiy, S., Solheim, F., Zou, X., Anthes, R., Businger, S., Trenberth, K. (1996), *Bull. Am. Meteorol. Soc.* **77,** 19–40.

Wiin-Nielsen, A. (1991), *Tellus* **43AB,** 36–52.

Willett, H. C. (1951), in: T. F. Malone, *Compendium of Meteorology,* Boston: American Meteorological Society, 731–746.

Further Reading

CAeM Working Group on Advanced Techniques Applied to Aeronautical Meteorology (1992), *Methods of Interpreting Numerical Weather Prediction Output for Aeronautical Meteorology,* Geneva: World Meteorol. Org., Tech. Note No. 195.

Daley, R. (1991), *Atmospheric Data Analysis,* Cambridge, U.K.: Cambridge Univ. Press.

Haltiner, G. J., Williams, R. T. (1980), *Numerical Prediction and Dynamic Meteorology,* New York: Wiley.

Kocin, P. J., Uccellini, L. W. (Eds.) (1989), *Weath. Forecasting* **4,** 273–443.

Ray, P. S. (1986), *Mesoscale Meteorology and Forecasting,* Boston: American Meteorological Society.

WHISKERS AND DENDRITES

R. Trivedi, *Ames Laboratory, U.S. Department of Energy and Department of Materials Science and Engineering, Iowa State University, Ames, Iowa, U.S.A.*

A. Karma, *Department of Physics, Northeastern University, Boston, Massachusetts, U.S.A.*

Introduction 441
1. **Dendritic Growth** 442
1.1 Background 442
1.2 Self-Consistent Dendrite-Growth Model 443
1.3 Ivantsov Model 444
1.4 Marginal-Stability Criterion 445
1.5 Microscopic Solvability Theory 446
1.6 Dynamical Models 447
1.7 Comparison between Theory and Experiments 448
1.8 Side-Branch Instability 450
2. **Whisker Growth** 451
2.1 Diffusion Model 452
2.2 Steady-State Solution 453
2.3 Time-Dependent Model 453
2.4 Kinetic Model 455
2.5 Whisker-Growth Experiments 456
Glossary 457
Works Cited 457
Further Reading 459

INTRODUCTION

Whiskers and dendrites are fine filamentary crystals with a very high aspect ratio of length to the diameter, and they can be formed from a vapor or a liquid phase (Kurz and Fisher, 1992; Trivedi and Kurz, 1994; Wagner, 1970; Givargizov, 1978). The filamentary nature of dendrites is illustrated in Fig. 1, from the studies by Trivedi and Somboonsuk (1984), which shows an array of dendrites in a thin sample of directionally solidified (i.e., solidified under a positive temperature gradient in the liquid) transparent system of succinonitrile-acetone. The shape of the dendrite front and the dendrite tip radius do not vary significantly for different dendrites in an array. Thus, a definite criterion exists for dendrite tip-radius selection. On the other hand, the formation of silicon whiskers from a vapor is illustrated in Fig. 1(b), which shows that there is no unique selection process, and a range of diameters is formed with the thinner whiskers growing at a lower rate (Givargizov, 1978).

The importance of dendritic growth comes from the fact that most alloys freeze with a dendritic structure. This dendritic freezing is accompanied by solute redistribution, which can either result in microsegregation or lead to the formation of a new phase in the interdendritic region. Either of these effects will significantly alter the properties of the solidified material. From a practical viewpoint, the characteristics of dendritic structures strongly influence the properties of cast and welded materials. Significant theoretical and experimental studies have thus been carried out to correlate quantitatively the dendrite growth rate and dendrite tip radius with the liquid undercooling.

The importance of whiskers can be seen through the consideration of the cohesive strength of the material. The theoretical strength of metals, calculated by determining the stress required to move lattice planes along the slip planes, was estimated to be in the range $\mu/30$–$\mu/15$, where μ is the elastic modulus (Frenkel, 1926). The maximum resolved shear stress required for plastic flow in real metals, however, was found to be significantly lower, and was in the range $(10^{-3}$–$10^{-4})\mu$. The lower strength of real crystals was attributed to the presence of dislocations, which cause plastic deformation. Thus, if dislocations could be removed or minimized, the strength of the crystals could be increased significantly. Since dislocations near the surface will experience image forces that will remove them, a very fine crystal will approach its theoretical strength (Frenkel,

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

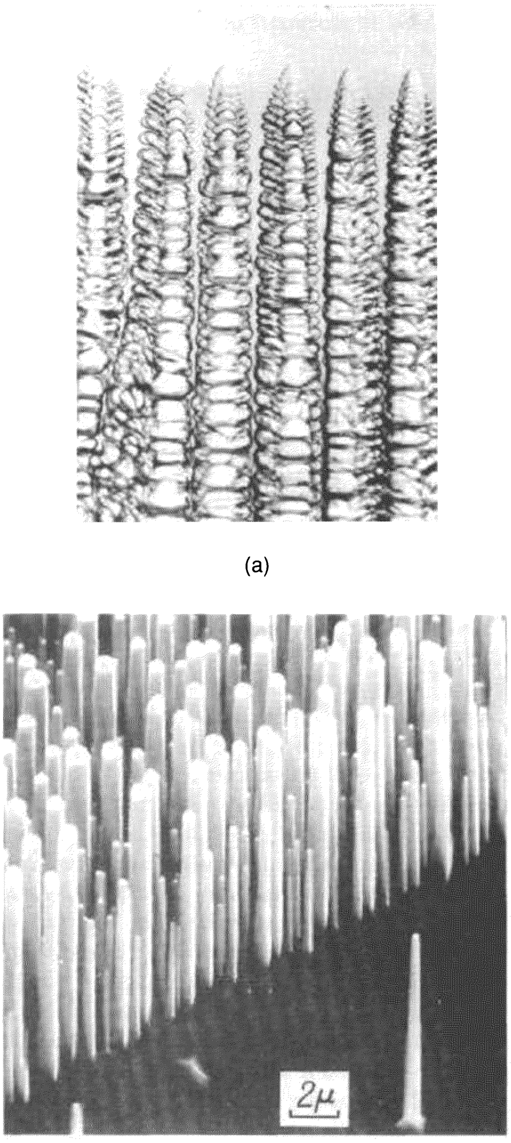

FIG. 1. (a) A dendritic array in directionally solidified succinonitrile–4.0 wt% acetone system (Trivedi and Somboonsuk, 1984); (b) silicon whiskers grown from a vapor (Givarzigov, 1978).

1926; Hirth and Lothe, 1968). Herring and Galt (1952) first showed that the strength of some of the whiskers approached the theoretical strength of the crystal. This high strength of whiskers generated much interest in their use as reinforcement materials in the design of composite materials. Specifically, whiskers of ceramic materials combine this high strength with the high modulus, low density, and high melting points of ceramics, so that ceramic whiskers can be effectively used in the design of ceramic:metal and ceramic:polymer composites. Whiskers, because of their small size, also show quantum size effects on optical, magnetic, and electrical properties so that whiskers of semiconductor materials have also been studied extensively.

The common technique for whisker production is growth from the vapor phase, and models have been developed that assume that the whisker formation is a vapor-to-solid transformation. Experimental studies by Wagner and Ellis (1964, 1965), however, have shown that whisker growth generally occurs through the formation of a liquid droplet at the advancing front of the whisker, giving rise to the vapor–liquid–solid (VLS) growth mechanism.

We shall first examine in detail the current status of the model of dendritic growth from an undercooled melt, and then examine models of whisker growth from the vapor phase.

1. DENDRITIC GROWTH

1.1 Background

The term dendrite comes from the Greek word dendron, which means tree, and like a tree the dendrite is a branched structure with primary, secondary, tertiary, and eventually higher-order branches. A microstructure of a dendrite grown from a pure undercooled succinonitrile (Huang and Glicksman, 1981) is shown in Fig. 2. Such a dendritic interface represents one of the most complicated patterns that evolve through the dynamical processes of growth. In addition, even the steady-state dendrite-growth model of an unbranched dendrite is complex since it requires proper physics to obtain a shape-preserving solution that gives a unique dendrite growth rate for a given undercooling.

Dendritic structures are ubiquitous structures in crystal growth. They can be formed by a variety of processes including growth from an undercooled melt, from a vapor, or by electrodeposition. In solidification, dendritic structures evolve under two distinctly

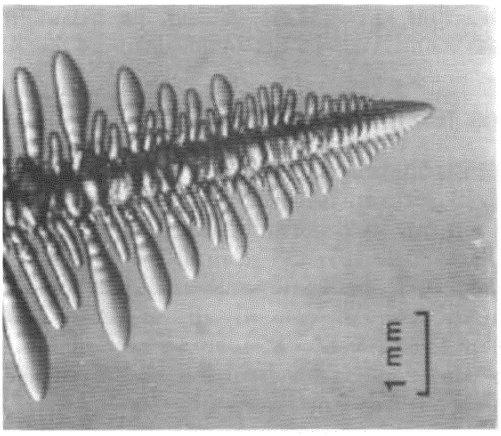

FIG. 2. A dendritic structure formed from an undercooled melt of pure succinonitrile (Huang and Glicksman, 1981).

different growth conditions, which differ in the way in which the latent heat of fusion is carried away from the interface. These are the following:

1. Growth from an undercooled melt, in which generally an equiaxed dendritic crystal forms, the latent heat of fusion being dissipated through the cooler liquid ahead of the interface. In this case the temperature gradient in the liquid at the interface is negative, whereas that in the solid is nearly zero.
2. Directional solidification or constrained growth in which a positive temperature gradient in the liquid is imposed so that the latent heat of fusion is dissipated through the solid.

Note that the dendrite growth direction and the heat flow direction are parallel for solidification in the undercooled melt condition, whereas they are antiparallel in the directional solidification process. In an equiaxed crystal, the thermal and solute fields ahead of any given dendrite can be significantly influenced by the presence of other dendrites in neighboring grains. In order to emphasize the basic physics of dendrite growth, we shall consider only the case of dendritic growth in a pure undercooled melt. In this case one obtains equiaxed dendrites, and the latent heat of fusion is dissipated through the cooler liquid ahead of the interface. The extension of this model to alloy dendrite growth is discussed by Langer (1986) and Trivedi and Kurz (1994).

During solidification of a pure melt, nonplanar structures form because they can effectively dissipate the latent heat of fusion required for growth. Consequently, dendrite growth models have been based on the solution of the thermal diffusion equation. The steady-state dendrite growth, however, is a free-boundary problem in which the shape of the interface is not known *a priori,* but must be determined in a self-consistent manner such that the solution of the diffusion equation establishes the shape that is preserved during the steady-state growth; i.e., the interface velocity V_n must satisfy the relationship $V_n = V\cos\theta$, where θ is the angle of the normal with respect to the growth axis and V is the velocity of the dendrite tip.

1.2 Self-Consistent Dendrite-Growth Model

The rigorous steady-state dendrite growth problem is governed by the following set of equations for the thermal fields in the liquid and solid:

$$\nabla^2 T_i + \left[\frac{V}{D}\right]\left[\frac{\partial T_i}{\partial z}\right] = 0, \qquad (1)$$

where $i = S$ or L, V is the axial dendrite growth rate, and D is the diffusion coefficient in the liquid. The boundary condition at the interface is governed by the Gibbs–Thomson and linear interface kinetic effects, which give

$$T_I = T_M - \left(\frac{T_M}{L}\right)\sum_{i=1}^{2}\left[\gamma(\mathbf{n}) + \frac{\partial^2\gamma(\mathbf{n})}{\partial\theta_i^2}\right]\frac{1}{R_i} - \frac{v_n}{\mu(\mathbf{n})}, \qquad (2)$$

where T_I is the interface temperature, R_i are the principal radii of curvature, θ_i are the angles between the normal $\mathbf{n}$ and the two local principal directions on the interface, and $\mu(\mathbf{n})$ is the linear interface kinetic coefficient. The other boundary condition requires that the temperature field approach the initial temperature of the undercooled liquid, T_∞, far from the interface.

The normal velocity of the interface is obtained from the condition of heat conservation at the interface:

$$v_n = D\left(\frac{c_p}{L}\right)\left[\left(\frac{\partial T}{\partial n}\right)_s - \left(\frac{\partial T}{\partial n}\right)_L\right] \tag{3}$$

in which c_p and L are the specific heat per unit volume and the latent heat per unit volume, respectively. For a crystal with cubic symmetry, the anisotropic interface energy $\gamma_n(n)$ is defined as

$$\gamma_n(n) = \gamma_0\,\bar{a}_s[1 + \epsilon_4'(n_x^4 + n_y^4 + n_z^4)] \tag{4}$$

in which ϵ_4' is the anisotropy parameter and γ_0 is the isotropic interface energy, i.e., interface energy for $\epsilon_4' = 0$. In practice, the anisotropy parameter ϵ_4' is not directly measured experimentally. The anisotropy parameter that is measured is the amplitude of the first Fourier cosine mode of $\gamma(\theta)$, where θ is the polar angle in a two-dimensional cross section of the equilibrium shape parallel to a (100) plane. If we write

$$\gamma(\theta) = \gamma_0 + \gamma_4 \cos 4\theta, \tag{5a}$$

then the ratio $\epsilon_4 = \gamma_4/\gamma_0$ can be extracted by examining the cross-sectional equilibrium shape

$$R(\theta) = R_0(1 + \epsilon_4 \cos 4\theta). \tag{5b}$$

It is then easy to work out that $\bar{a}_s = 1 - 3\epsilon_4$, and $\epsilon_4' = 3\epsilon_4/(1 - 3\epsilon_4)$, by expanding $\cos^4\theta$ and $\sin^4\theta$ in Eq. (4) in terms of $\cos 4\theta$.

1.3 Ivantsov Model

Papapetrou (1935) first showed qualitatively that an isothermal parabolic interface will grow at a constant rate and satisfy the shape-preserving constraint. A mathematical model of the steady-state dendrite growth model was subsequently developed by Ivantsov (1947), who obtained an exact solution of the steady-state diffusion equation under the boundary conditions that $T(z \to \infty) = T_\infty$, and the interface temperature T_I is constant, $T_I = T_M$, where T_M is the melting point. For a parabolic needle dendrite growth, he obtained a relationship between dimensionless undercooling Δ and the dimensionless Peclet number P as

$$\Delta = Pe^P E_1(P), \tag{6a}$$

where $\Delta = (T_M - T_\infty)(c_p/L)$ and $P = V\rho/2D$, in which ρ is the dendrite tip radius. $E_1(P)$ is the exponential integral function given by

$$E_1(P) = \int_P^\infty \frac{e^{-z}}{z}\,dz. \tag{6b}$$

For a given undercooling, the Ivantsov solution gives a continuum of solutions between V and ρ, and no specific velocity or radius is selected (Fig. 3). Experimental studies by Huang and Glicksman (1981) and Glicks-

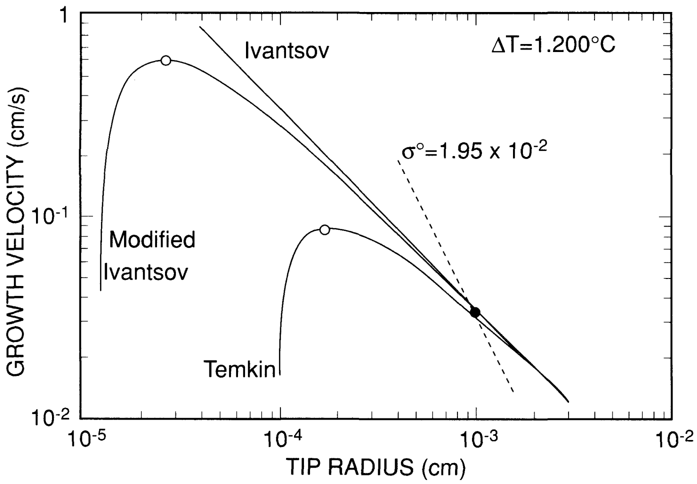

FIG. 3. The variation in dendrite velocity with the dendrite tip radius for a given undercooling in high-purity succinonitrile (Huang and Glicksman, 1981). The dashed line is the marginal-stability criterion for $\sigma^* = 0.0195$.

man *et al.* (1976) confirm that the shape of the dendrite is very close to a parabola, and the Ivantsov relationship, Eq. 6(a), is satisfied when experimental values of D, V, and ρ are used. Thus, the experimental results fall on the Ivantsov curve in Fig. 3, but the Ivantsov model does not predict this specific selection of V and ρ. From a theoretical viewpoint, if only thermal transport process were important for the formation of dendrites in an undercooled melt, then the dendrite tip would be unstable with respect to tip splitting, and no steady-state growth would occur. Thus, another consideration, such as interfacial free energy or interface kinetics that can stabilize the dendrite tip, is essential to define properly the dendrite growth problem. The quantitative characterization of this stabilizing effect is key to the unique solution of the steady-state dendrite growth problem.

Several approximate models were proposed to incorporate the effect of interface energy into the dendrite growth model. The modified Ivantsov model assumed that the dendritic interface is parabolic and isothermal, but the constant temperature corresponds to the tip temperature, which will be different from the melting point because of the curvature effect. This model still gave a continuum of solutions, as shown in Fig. 3, but the velocity–radius relationship exhibited a maximum since the velocity must reduce to zero at the critical radius for nucleation. Subsequently, Temkin (1960) extended the Ivantsov model by assuming the interface to be parabolic, and considered the Gibbs–Thomson effect and the kinetic effect to characterize the variation in temperature along the interface. The nonisothermality of the interface also required the solution of heat flow in the solid. Temkin solved Laplace's equation for the thermal field in the solid and in the liquid, and a solution for the steady-state diffusion equation for the parabolic interface was subsequently obtained by Trivedi (1970), who obtained the result

$$\Delta = [Pe^{P}E_1(P)][1 + (d_0/2P\rho)F_0 + (d_k/\rho)F_k], \tag{7}$$

where $d_0 = T_M c_p \gamma_0 / L^2$ is the capillary length and $d_k = c_p D / L\mu_0$ is the kinetic length in which μ_0 is the linear kinetic coefficient. The functions F are approximately given by $F_0 = (3.8410 + 4P) + (K_s/K_l)[2.06/(1 + P)]$ and $F_k = (1.4050 + 2P) + (K_s/K_l)0.5954/(1 + P)$. K_s and K_l are thermal conductivities in the solid and liquid, respectively. The three terms in the second factor show the undercoolings required for the diffusion process, for the interface energy, and for the interface kinetics, respectively. This model gives a lower growth rate due to the additional heat flux from the sides of dendrite to the tip, in the liquid and in the solid, due to the nonisothermal interface.

The consideration of interface energy in the Ivantsov model gave rise to a continuum of solutions with a maximum in the V–ρ relationship for a given undercooling, as shown in Fig. 3. The tip selection was assumed to correspond to the maximum growth rate. However, all these modifications are not self-consistent since they do not satisfy the shape-preserving condition.

A self-consistent model of dendritic growth was first proposed by Nash and Glicksman (1974), who developed an integral formulation using the Green's function of the diffusion equation in which the interface shape appeared as an unknown. They assumed isotropic interface energy, and solved the equation only approximately, which again gave a continuum of solutions. Thus, the selection of dendrite tip radius and velocity at a given undercooling still remained unexplained. A definitive experimental study by Glicksman *et al.* (1976) in a transparent system in which both V and ρ can be accurately measured showed that, although the Ivantsov equation is satisfied, the dendrite tip selection does not correspond to the maximum growth-rate value, but the selected radius is larger than that corresponding to the maximum growth rate.

1.4 Marginal-Stability Criterion

The problem of dendrite tip selection was addressed by Langer and Müller-Krumbhaar (1977), who carried out a linear stability analysis of the dendrite tip around a parabolic dendrite. They considered isotropic interface energy but neglected the interface kinetics effect. They assumed that a stable steady-state shape exists under isotropic interface energy that departs only slightly from a parabola. They showed that only a narrow band of dendrite tip radii will be stable from the continuum solution of Ivantsov, and the lower limit of stable radius corresponded to the maximum growth-rate condition. They proposed a marginal-stability criterion that assumed that the system selects only the largest stable radius value, as shown by a dashed line in Fig. 3. This marginal-stability criterion, when coupled with the Ivantsov solution, was found to agree remarkably well with the experimental results of Glicksman *et al.* (1976) in pure undercooled succinonitrile (SCN). The model showed that, in addition to the dimensionless parameters Δ and P in the Ivantsov solution, the inclusion of surface energy gives a new dimensionless parameter σ^*, defined as

$$\sigma^* = 2D_l d_0/\rho^2 V \tag{8}$$

in which d_0 is the capillarity length, $d_0 = T_M c_P \gamma_0/L^2$. The marginal-stability condition was found to occur at $\sigma^* = 0.025$. For very small undercoolings, the modification to the Ivantsov equation due to interface energy was found to be negligible. Equations (6a) and (8) thus give the unique values of V and ρ for a given undercooling.

1.5 Microscopic Solvability Theory

Although the marginal-stability criterion gives a good agreement with experimental results, it is based on the erroneous assumption that steady-state growth is possible for a continuum of velocities and tip radii. In the early 1980s, the introduction of simplified models of dendritic growth with local growth laws, such as the geometrical (Kessler *et al.*, 1985) and boundary-layer (Ben-Jacob *et al.*, 1984) models, made it possible to solve the steady-state growth equations without any uncontrolled approximations. These studies lead to two important conclusions:

1. Surface tension ($d_0 \neq 0$) acts as a singular perturbation that breaks the continuous spectrum of steady-state growth solutions. This singular role originates mathematically from the fact that capillary effects enter into the equations as the product of the surface tension length d_0 and the curvature κ that involves the highest-order spatial derivative of the interface shape [e.g. $\kappa \sim d^2z(x)/dx^2$ in two dimensions]. Physically, this singular role of capillarity is already reflected in the stability spectrum of a planar interface. Without capillarity, a sinusoidal perturbation of the interface is always unstable, no matter how small is its wavelength, whereas with capillarity it is only unstable if its wavelength is larger than the stability length $\lambda_s \sim 2\pi(d_0 l)^{1/2}$.
2. Anisotropy has a critical role. For an isotropic surface energy ($\epsilon_4 = 0$), there is *no* possible steady-state solution. In this case dendritic growth becomes replaced by the growth of seaweed-like disordered structures controlled by repeated tip-splitting events. In contrast, with a finite amount of anisotropy ($\epsilon_4 \neq 0$), even small, there is a discrete spectrum of possible steady-state solutions, only the one with the largest velocity being linearly stable against tip splitting.

Numerous studies showed subsequently that these findings extend to the complete diffusive free-boundary problem of dendritic growth described by Eqs. (1)–(5), so that these conclusions are now firmly established.

For growth in two dimensions, the interface is one-dimensional, and the dendrite growth problem can be solved accurately by the Green's function method (Pelcé and Pomeau, 1986; Caroli *et al.*, 1986; Kessler and Levine, 1986; Meiron, 1986; Ben-Amar and Moussallam, 1987). The model predicts that the parameter σ^*, defined in Eq. (8), is not a numerical constant but is a function of the interface energy anisotropy parameter ϵ_4. For vanishing anisotropy value the limiting value of σ^* is obtained as $\sigma^* = \sigma_0 \epsilon_4^{7/4}$, where σ_0 is of the order of unity. Note that the variation in σ^* departs from this result at very small values of ϵ_4. In the absence of anisotropy, i.e., $\epsilon_4 = 0$, the model shows that there are no steady-state solutions.

The basic prediction that anisotropy is necessary for dendrite growth provides a natural explanation for the observed preferred growth direction of dendrites. Previous models, which did not take into account the anisotropy, could only attribute the symmetry to the role played by anisotropy during the initial development of a spherical nucleus.

Rigorous solvability predictions in three dimensions, for arbitrary anisotropy and undercooling values, have been made only for an axisymmetric case in which the interface shape $z(r,\phi)$ is independent of the polar angle ϕ in the x-y plane perpendicular to the growth direction. This axisymmetric approximation reduces the three-dimensional steady-state growth problem to a tractable problem that is effectively two dimensional, and the interface shape $z(r)$ approaches the Ivantsov paraboloid of revolution far from the tip region (Barbieri and Langer, 1989). This paraboloid becomes the exact asymptotic shape in this case. The nonlinear boundary integral equation for this problem was developed by Kessler and Levine (1988) as

$$\Delta - \Delta_m = \int_0^{2\pi} \frac{d\phi'}{2\pi} \times \int_0^{+\infty} \frac{r'dr'}{l} \frac{\exp[z(r') - z(r) - d/l]}{d}, \tag{9}$$

where $l = 2D/V$ is the diffusion length,

$$d = \{r^2 + r'^2 - 2rr' \cos(\phi - \phi') + [z(r') - z(r)]^2\}^{1/2},$$

and

$$\Delta_m = -d_0\left(a_s + \frac{\partial^2 a_s}{\partial \bar{\theta}^2}\right)\left(\frac{z_{rr}}{(1 + z_r^2)^{3/2}}\right) - d_0\left(a_s + \cot\bar{\theta}\,\frac{\partial a_s}{\partial \bar{\theta}}\right)\left(\frac{z_r}{(1 + z_r^2)^{1/2}}\right)$$

represents the Gibbs–Thomson effect for the axisymmetric case in terms of the anisotropy function

$$a_s(\bar{\theta}) = \bar{a}_s[1 + \epsilon_4'(\cos^4\bar{\theta} + \tfrac{3}{4}\sin^4\bar{\theta})] \tag{10}$$

that is averaged in the azimuthal direction and only depends on the angle $\bar{\theta}$ between the normal to the interface and the z axis defined by $\tan\bar{\theta} = -z_r$. $\bar{a}_s$ is defined in Eq. (4).

The nonaxisymmetric three-dimensional model, however, has remained difficult to solve because the two-dimensional interface is harder both to parametrize and to find numerically. In addition, in contrast to the case of two dimensions, the Ivantsov paraboloid needle crystal solution no longer represents the correct asymptotic solution to the steady-state growth equations far behind the tip, which makes it difficult to impose a rigorous boundary condition on the form of the steady-state shape far behind the tip. A solution only in the limit of vanishing anisotropy, $\epsilon_4 \rightarrow 0$, and small undercooling, $\Delta \ll 1$, has been obtained by Ben-Amar and Brenner (1993) by analyzing the behavior of the solution in the tip region separately from the tail. This model yields the classic solvability prediction that at vanishingly small values of ϵ_4, the relationship $\sigma^*(0,\epsilon_4) = B\epsilon_4^{7/4}$ is obtained. This result, which shows the same $\frac{7}{4}$ exponent as in the two-dimensional model, is conceptually important in that it provides a convincing theoretical argument that the structure of the solvability theory does indeed extend from two to three dimensions. The model, however, is only quantitatively valid for very small values of ϵ_4. Thus, for the purpose of testing the solvability theory against experimental results, nonaxisymmetric models are required for arbitrary anisotropy and undercooling values, and such models have recently been developed by Karma and Rappel (1996a, 1996b, 1997) through phase-field simulations.

1.6 Dynamical Models

Several dynamical simulations of dendritic growth have been carried out that establish the critical role of crystalline anisotropy on the growth morphology. Accurate quantitative tests of solvability predictions have also been performed by Ihle and Müller-Krumbhaar (1994) using sharp interface and by Karma and Rappel (1996a, 1996b, 1997) using phase-field models. Kobayashi (1994) first carried out phase-field simulations of dendritic growth in three dimensions with a kinetic anisotropy. More recently, Schmidt (1996) has simulated three-dimensional dendrite morphologies using a finite-element code. Both these studies have yielded dendrite morphologies that resemble those seen in experiment, although they have not compared the results quantitatively with the microscopic solvability model. More detailed phase-field calculations have recently been carried out by Karma and Rappel (1997), and they have compared their results with those of the numerical solvability model and experiments.

The phase-field method is rooted in continuum models of phase transition. It is rapidly emerging as a method of choice for simulating interface pattern-formation phenomena in solidification and other systems. The appeal of this approach is to avoid tracking of macroscopically sharp phase boundaries by introducing a nonconserved order parameter or "phase field," ψ, to distinguish between different phases. The equation of motion for this field is derived variationally from a phenomenological free-energy functional F, and is given by

$$\tau(\mathbf{n})\frac{\partial\psi}{\partial t} = -\frac{\delta F}{\delta t} \qquad (11)$$

in which

$$F = \int d\mathbf{r}\left[\frac{W^2(\mathbf{n})}{2}|\nabla\psi|^2 + f(\psi,u)\right], \qquad (12)$$

$$f = -\frac{\psi^2}{2} + \frac{\psi^4}{4} + \lambda u\psi\left(1 - \frac{2\psi^2}{3} + \frac{\psi^4}{5}\right), \qquad (13)$$

and

$$u = (T - T_M)/(L/c_p). \qquad (14)$$

Equation (13) has minima at $\psi = -1$ and $\psi = +1$ that correspond to the liquid and solid phases, respectively. The anisotropic surface energy defined by Eq. (4) is recovered by choosing

$$W(\mathbf{n}) = W_0\gamma(\mathbf{n})/\gamma_0, \qquad (15)$$

where $\mathbf{n} = -\nabla\psi/|\nabla\psi|$ is the vector field that corresponds to the normal to the boundary and $\gamma(\mathbf{n})$ and γ_0 are defined by Eq. (4). This representation of anisotropy is a direct generalization in three dimensions of the way anisotropy has been previously included in two dimensions. The equation for the phase field is now coupled to that of the diffusion field:

$$\frac{\partial u}{\partial t} = D\nabla^2 u + \frac{1}{2}\frac{\partial\psi}{\partial t}, \qquad (16)$$

in such a way that the equations for the two fields reduce to those of the original free-boundary problem in the limit where the interface width W_0 is small compared to the principal radii of curvature of the boundary. This thin-interface limit can be taken by using the analysis of Karma and Rappel (1996b, 1997). In this approach, the necessary relation for choosing λ and the function $\tau(\mathbf{n})$ is taken such that $1/\mu(\mathbf{n})$ in Eq. (2) vanishes, thereby purposely eliminating the kinetic undercooling at the interface. The equations of the phase-field model, Eqs. (11) and (12), can now be simulated explicitly on large cubic lattices for a fixed undercooling and for different anisotropy values.

A typical three-dimensional dendrite morphology resulting from the simulated growth of a small spherical seed is illustrated in Fig. 4. The results of these runs are summarized in Figs. 5(a) and 5(b), which also shows the result of the numerical and analytical solvability models for an axisymmetric dendrite.

1.7 Comparison between Theory and Experiments

The dependence of P and σ^* on anisotropy predicted by the numerical solvability theory and the phase-field simulations are in good overall agreement (Fig. 5). The Peclet number P is found to agree quantitatively over the entire range of anisotropy, while σ^* values predicted by solvability are systematically lower than the phase-field values. The σ^* value predicted by solvability is about 70% smaller than the one observed in phase-field simulation for $\epsilon_4 = 0.047$, and only about 20% smaller for $\epsilon_4 = 0.0066$. This difference can be attributed to the nonaxisymmetric departure of phase-field dendrites from a shape of revolution that was found to

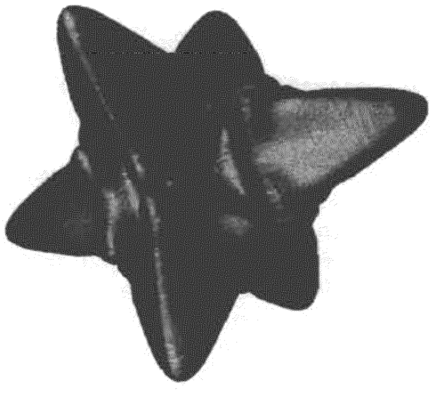

FIG. 4. A three-dimensional nonaxisymmetric dendritic structure predicted by the phase-field method (Karma and Rappel, 1997).

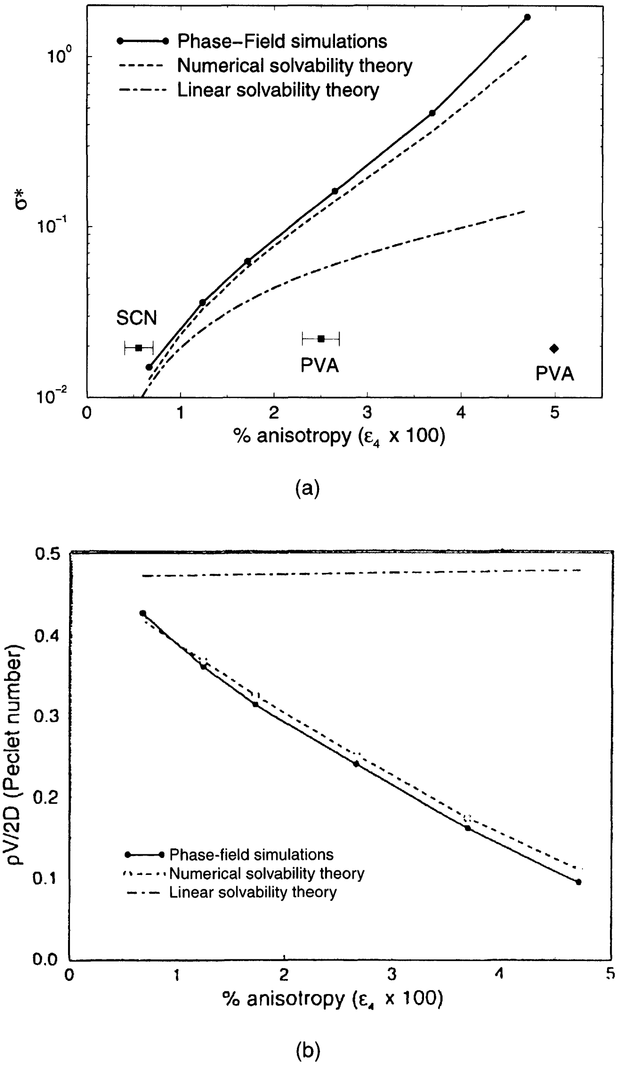

FIG. 5. The relationship between (a) σ^* and anisotropy, and (b) Peclet number and anisotropy, as predicted by the phase-field simulations, numerical solvability, and linear solvability methods. Experimental results for SCN and PVA are also shown for comparison. Two PVA results are plotted: the anisotropy measurements of Glicksman and Singh (1989) shown by a filled diamond, and of Muschol *et al.* (1992) shown by filled squares. (From Karma and Rappel, 1997.)

increase with anisotropy. For the nonaxisymmetric model, the actual tip radius, which is also the value measured experimentally, that is used to calculate σ^* is different from the tip-radius value that corresponds to the asymptotic Ivantsov shape. The results of Fig. 5(b), however, indicate that even at very low anisotropy, these two Peclet numbers still differ in such a way that the choice of definition is quantitatively relevant.

The results of experimental studies in succinonitrile and pivalic acid are also compared in Fig. 5(a). The present comparison is restricted to SCN and PVA, which are the only pure materials with an underlying cubic symmetry for which detailed measurements of both anisotropy and σ^* are available (Huang and Glicksman, 1981; Glicksman and Singh, 1989; Muschol *et al.*, 1992; Rubinstein and Glicksman, 1991). For succinonitrile, the phase-field model predicts $\sigma^* = 0.015$, which is slightly smaller than the experimental value obtained by Glicksman *et al.* (1976), as shown in Fig. 5(a).

The agreement between theory and experiment, however, remains very poor for PVA, both if we use the anisotropy measurement of Mushol *et al.* (1992), or the higher 5% anisotropy value measured previously by Glicksman and Singh (1989). Historically, the proximity of the σ^* values of SCN and PVA appeared as a strong point in favor of marginal stability theory. The departure of σ^* in PVA from the solvability model, however, is significant.

There seem to be at least three possible reasons for this discrepancy. The first is that kinetic effects at the interface, which were purposely excluded from the phase-field computations, could play a significant role at low undercooling in PVA, unlike in SCN where σ^* is measured to be only slightly dependent on undercooling. Kinetic effects have been studied recently by Wang and Sekerka (1996) by two-dimensional phase-field simulations at large Δ, and they found kinetic effects to lead to a decrease in σ^* with increasing Δ that is qualitatively similar to the one observed in experiment in PVA. However, detailed phase-field simulations that take into account the kinetic effects need to be developed to resolve an order-of-magnitude disagreement between theory and experiment.

The second possibility is that experiments, which have intrinsically a finite optical resolution, cannot measure the tip radius ρ sufficiently accurately, and can overestimate its value. The shapes of the dendrite tip region predicted by the axisymmetric numerical model, the nonaxisymmetric phase-field model, and the Ivantsov model are shown in Fig. 6. Thus, it becomes very difficult to measure ρ accurately at large anisotropy because the curvature varies very rapidly in the tip region. Special care needs to be taken to obtain reliable answers. It is possible that only an averaged value of the curvature in the tip region is measured experimentally. This aspect needs to be reexamined carefully for strongly anisotropic materials.

A third and perhaps least likely possibility is that the dendrite tip selects a different dynamical attractor than the underlying steady-state needle crystal at low undercooling. In this case the solvability theory would not be applicable. When the diffusion length becomes much larger than ρ, there is the possibility of a nonlinear feedback mechanism between the side branches and the tip, which could potentially modify the operating state of the tip, although one would not expect an order-of-magnitude shift in σ^*.

1.8 Side-Branch Instability

Everything described so far pertains to the determination of the operating state of the dendrite tip. We conclude this section with a brief description of sidebranching that leads to treelike dendrites. According to current theories, the formation of secondary branches behind the tip should be the result of the selective amplification of some form of noise along the sides of dendrites (Langer, 1989). Current quantitative estimations indicate that thermal noise should, alone, be large enough to account for the experimentally observed sidebranching (Brenner and Temkin, 1995). Although it is not possible experimentally to switch off thermal noise, it is possible to do so in dynamical simulations. Simulations to date do indeed confirm that dendrites do not form secondary branches without noise (Warren and Boettinger, 1995; Karma and Rappel, 1996c). Experimental measurements of sidebranching characteristics also seem consistent with a noise-induced sidebranching mechanism (Dougherty *et al.*, 1987).

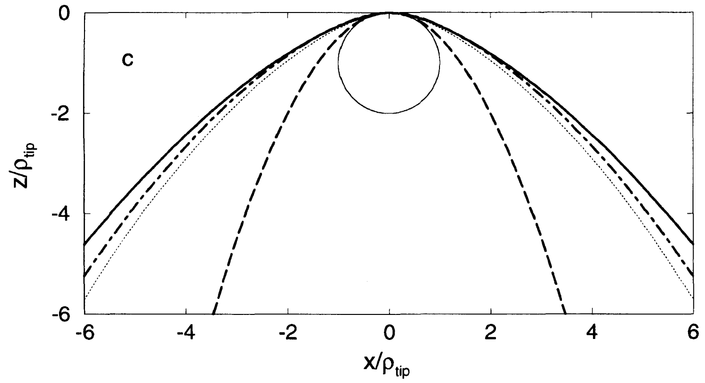

FIG. 6. The shape of the dendrite tip region predicted by the nonaxisymmetric phase-field model, the solid and dash-dotted lines corresponding respectively to $\phi = 0°$ and $\phi = 45°$. The result of axisymmetric numerical calculations (dotted line), Ivantsov parabola (dashed line), and a circle of unit radius that is normalized with tip radius obtained by the phase-field method are shown for comparison. (Karma and Rappel, 1997.)

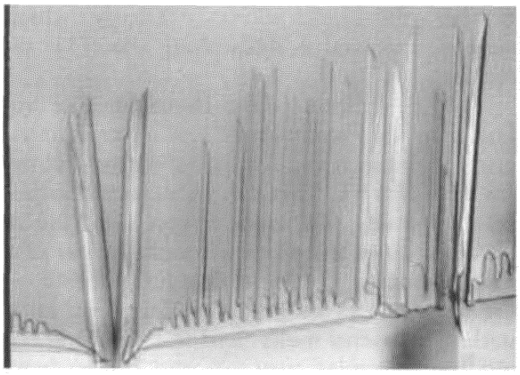

FIG. 7. Fine needle crystals formed during the directional solidification of as-received *tert*-butyl alcohol (Fabietti and Trivedi, 1997).

The fact that thermal noise can produce sidebranching may seem at first surprising. Thermal fluctuations are microscopically small (with an energy scale ~$k_B T_M$, where k_B is the Boltzmann constant) and, as a rule of thumb, do not affect the dynamics of macroscopic objects. Dendrites, however, provide a rather unique exception to this rule. The microscopically small fluctuations become important because their magnitude becomes amplified by several orders of magnitude by the same morphological instability that produces the primary branches in the first place. Some broader questions remain to be understood: how do secondary, tertiary, and higher-order compete dynamically so as to produce an anisotropic fractal structure at the scale of an entire dendritic grain? How can this structure be characterized? Progress on these more complex issues is currently being made (Brenner *et al.*, 1996) but much remains to be understood.

2. WHISKER GROWTH

When crystals are grown from a liquid in a very anisotropic material, very thin and elongated whiskers form. Figure 7 shows an array of such crystals in a directionally solidified transparent material, *tert*-butyl alcohol (TBA) (Fabietti and Trivedi, 1998). The crystal structure plays an important role in the formation of such needlelike crystals (Nabarro and Jackson, 1958). Besides the growth from a liquid, whiskers are commonly grown from the vapor phase. Several theoretical models of whisker growth have been developed. Initial theories by Sears (1953, 1955) were based on the formation of whiskers directly from a vapor phase, i.e., vapor-to-solid transformation, in which the preferential growth leading to a whisker shape was assumed to be due to the presence of an axial screw dislocation that would provide sites for atomic attachment during the crystal growth (Frank, 1949; Burton *et al.*, 1951), as shown in Fig. 8(a).

Several experimental studies have been carried out to test the dislocation model. Axial dislocations in whiskers have been observed only in a few materials that support

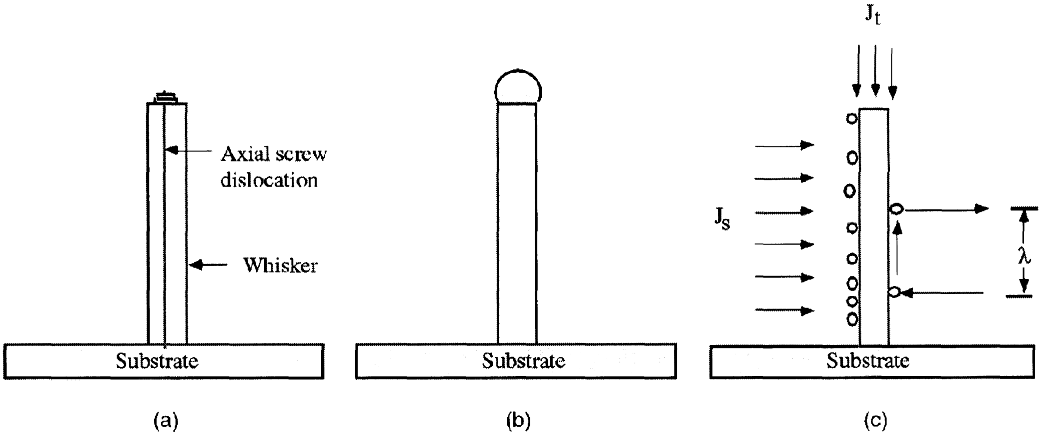

FIG. 8. Schematic illustrations of whisker growth by (a) screw dislocation mechanism, (b) liquid droplet mechanisms, and (c) the diffusion model showing the impingement flux and the desorption of adsorbed atoms.

the formation of whiskers. However, detailed experimental examination of whiskers in several systems found no dislocations in the whisker (Webb, 1958; Wagner *et al.*, 1964; Drum and Mitchell, 1964; Edwards and Huang, 1966; Minagava and Gejyo, 1969; Gabor and Blocher, 1969; Gotoh, 1972; Regis, 1969; Hasiguti *et al.*, 1970; Givargizov and Sheftal, 1972; Koparanova *et al.*, 1994; Hasimoto and Yamaguchi, 1996). In addition, even if screw dislocations were present in the substrate, the growth of steps leading to a spiral growth would be preferred over the formation of a whisker during the early stages of growth. Typically, whiskers have been found to form only under high supersaturation conditions, whereas the screw-dislocation mechanism is dominant only under very low supersaturations. Finally, experimental studies showed the need for impurities to form whiskers. When high-purity materials were used, no whisker formation occurred (Wagner, 1970). Thus another mechanism may also be operative for the whisker growth.

Wagner and Ellis (1964, 1965), in their experimental studies on Si whisker growth, observed that a droplet of liquid was present on the growing front of each whisker, as illustrated schematically in Fig. 8(b). They proposed an alternative mechanism in which the growth occurs through the transport of matter from vapor to liquid to solid, a process called the vapor–liquid–solid (VLS) growth. The presence of the liquid droplet has now been observed in several whisker-growth studies, and the droplet forms as a result of the presence of impurity, which alloys with the substrate or deposits, causing the alloy region to melt at the substrate temperature. Wagner and Ellis (1964, 1965) and Wagner (1967) proposed a model for the growth of silicon whiskers with gold as an impurity element. When the substrate is heated above the eutectic temperature, a gold-silicon liquid alloy is formed, and the liquid takes the shape of a droplet. A vapor in the form of a mixture of $SiCl_4$ and H_2 is introduced, which react at the liquid droplet surface to produce silicon. Silicon then transports to the solid–liquid interface where it gets deposited and causes the whisker to grow. The whiskers were found to grow along the ⟨111⟩ directions, and the unidirectional growth was a result of the anisotropy in the solid–liquid interface energy, which gave preferential growth along the ⟨111⟩ direction.

The recognition of the importance of impurity in the formation of a whisker enabled one to obtain a regular two-dimensional array of whiskers. This can be achieved either by depositing gold on a silicon substrate at predetermined locations through the use of a mask or by depositing a thin film of gold on the substrate. This gold film forms an alloy with silicon and melts, and the resulting liquid film becomes unstable and forms droplets in two dimensions with a characteristic wavelength.

2.1 Diffusion Model

A theoretical model of whisker growth was first proposed by Sears (1953, 1955) in which the diffusion of adsorbed atoms was considered to control the growth of whiskers. The axial velocity of the whisker due to the diffusive flux of adsorbed atoms at the advancing front can be calculated by considering the number of atoms impinging on the advancing surface less the number of atoms evaporating from the advancing surface, and assuming that the net impinged atoms get attached to the solid and increase the length L of the whisker. This gives

$$V_0 = \frac{dL}{dt} = \frac{(p - p_0)\Omega}{(2\pi mkT)^{1/2}}, \tag{17}$$

where Ω is the atomic volume in the solid, m is the mass of an atom, and p_0 is the equilibrium vapor pressure at temperature T. When this result was compared with the experimental data for mercury whisker growth, Sears observed that the experimental growth rate was more than three orders of magnitude larger. He thus postulated that the whisker growth occurs by the adsorption of atoms on the substrate and on the sides of the whiskers, and these adsorbed atoms move by surface diffusion toward the advancing front of the whisker, as shown in Fig. 8(c). This is known as the diffusion model for the whisker growth.

Besides the direct impingement of atoms on the advancing front, the flux J_t in Fig. 8(c), the whisker growth model considers the following ideas:

1. impingement of atoms on the lateral sides of the whisker, J_s, and temporary adsorption of the atoms,
2. diffusion of adsorbed atoms to the tip of the whisker, and
3. evaporation of adsorbed atoms that do not reach the tip before lifetime τ_s, or when they reach the mean free surface diffusion length, λ_s, with $\lambda_s = (D_s\tau_s)^{1/2}$, as shown in Fig. 8(c).

Based on these ideas, the surface diffusion equation was developed by Dittmer and Neumann (1960) as

$$\frac{\partial n}{\partial t} = J_s + D_s\left(\frac{\partial^2 n}{\partial z^2}\right) - \left(\frac{1}{\tau_s}\right)n, \tag{18}$$

where $n(z,t)$ is the concentration of atoms along the whisker length, z the axial growth direction, and J_s the impingement flux. Solutions of this diffusion equation have been developed by several investigators: Dittmer and Neumann (1958, 1960), Blakely and Jackson (1962), Ruth and Hirth (1964), and Simmons *et al.* (1964, 1967). The diffusion model is valid only when the adsorbed layer is only a fraction of a monolayer.

2.2 Steady-State Solution

Since diffusion occurs much faster than whisker growth, a steady-state condition can be assumed, i.e., $\partial n/\partial t = 0$, and a quasi-stationary surface diffusion profile can be calculated. This steady-state solution is obtained as

$$V = \frac{dL}{dt} = V_0 + \left[\frac{4V_0\lambda_s}{d}\right]\tanh\left(\frac{L}{\lambda_s}\right). \tag{19}$$

Equation (19) was developed in a fixed frame of reference. Ruth and Hirth (1964) have obtained the steady-state solution in a moving frame of reference attached to the whisker front, and under most whisker-growth conditions, the solution was found to approach that for the fixed frame of reference.

When whiskers are grown by chemical vapor deposition in the VLS technique, the first term in Eq. (19) can be dominant for larger-diameter whiskers, so that $V \approx V_0$. On the other hand, for a whisker grown from its own vapor, or when the deposition process is physical vapor deposition, the first term, i.e., V_0, is negligible. The general behavior of the result in this case can be visualized by examining the limiting cases. For short whiskers, or during the early stages of growth, i.e., for $L \ll \lambda_s$, the second term gives for the growth rate

$$V = 2V_0L/d. \tag{20}$$

Substituting $V = dL/dt$ and integrating gives

$$L = L_0e^{(2V_0/d)t}, \tag{21}$$

where L_0 is the whisker length at $t = 0$. Thus, an exponential growth law is predicted initially.

For long whiskers, when $L \gg \lambda_s$, the whisker growth rate reduces to

$$V = \frac{4V_0\lambda_s}{d} \tag{22}$$

which gives

$$L = \left[\frac{4V_0\lambda_s}{d}\right]t, \tag{23}$$

so that a linear growth law is predicted. For intermediate lengths, a smooth transition from an exponential to linear growth is observed. The whisker-growth model does not predict a unique velocity or whisker diameter at a given undercooling. Rather, a continuum of solutions is shown to exist that relates velocity to the diameter.

2.3 Time-Dependent Model

A rigorous non-steady-state solution of Eq. (18) was obtained by Simmons *et al.* (1964, 1967) to determine the conditions under which steady-state approximation is valid, and to present a complete solution of the time-dependent whisker-growth problem. They rewrote Eq. (18) in terms of a parameter $U = n - n_e$, in which n_e is the equilibrium concentration of adatoms on a stationary whisker of radius r that is in equilibrium with the vapor phase. For the origin of the coordinate system at the base of the whisker, the diffusion equation was obtained as

$$\frac{\partial U}{\partial t} = D_s \frac{\partial^2 U}{\partial z^2} - \frac{U}{\tau} + J, \tag{24}$$

where $J = J_s - J_0$. The boundary conditions at $z = 0$ and $z = \mathrm{L}$ are given by

$$\left[\frac{\partial U}{\partial z}\right]_{z=0} = 0, \tag{25}$$

$$U(L(t),t) = 0. \tag{26}$$

The flux balance at the whisker tip gives the growth rate as

$$\left[\frac{\partial L}{\partial t}\right]\alpha = -D_s\left[\frac{\partial U}{\partial z}\right]_{z=\mathrm{L}}, \tag{27}$$

where $\alpha = v/p$, in which p is the perimeter of the whisker and v is the number of atoms in 1 cm length of the whisker including the equilibrium number of adsorbed atoms on the surface. Since experimental studies show that the number of atoms along the radius is larger than 20, the value of α for a cylindrical whisker of effective diameter d is given by $\alpha = d/4\Omega$.

An exact analytical solution of the above problem was obtained by Simmons *et al.* (1967). Except for a short transient period, the solutions in terms of dimensionless length and dimensionless time were found to depend only on one dimensional parameter, σ, defined as

$$\sigma = (J/\alpha)\tau \approx 4V_0\tau/d. \tag{28}$$

The results of the model for two different values of σ are shown in Fig. 9, in which the predictions of the steady-state model of Dittmer and Neumann (1960) are also shown for comparison.

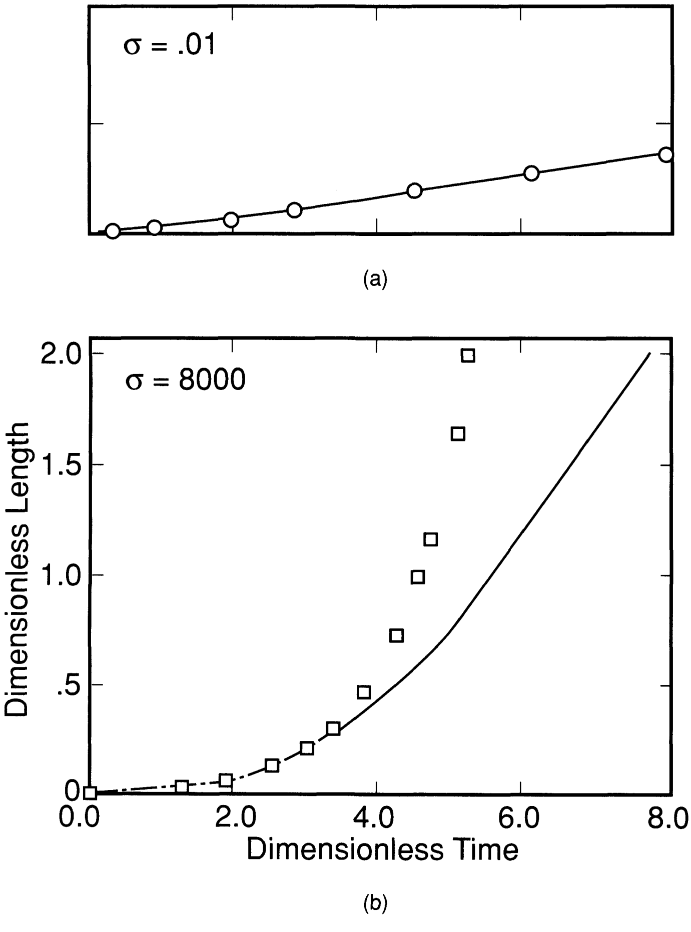

(a)

(b)

FIG. 9. The variation in the dimensionless length, $L\ (J/4D_s\alpha)^{1/2}$, with the dimensionless time, $t' = (J/\alpha)t$, after Simmons *et al.* (1967). The square points are for the steady-state model of Dittmer and Neumann (1958).

The limiting values of velocities for the time-dependent model were also obtained. For $L \ll \lambda_s$, the result can be written in terms of dimensional parameters as

$$V = 2V_0L/d. \tag{29}$$

This limiting case gives an identical result to that obtained by Dittmer and Neuman (1960); i.e., an exponential growth law is predicted for the initial whisker length. Thus, the results of the time-dependent and the steady-state models, shown in Fig. 9(b), match quite precisely for the initial time.

For long whiskers, when $L \gg \lambda_s$, the whisker growth rate was obtained as

$$V = \left(\frac{4V_0D_s}{d}\right)^{1/2}\left(\frac{\sigma}{1 + \sigma}\right)^{1/2}, \tag{30}$$

which gives a linear growth law for the whisker length. The slope of the line, however, depends upon the value of σ. For $\sigma \gg 1$, the above result gives

$$V = \left(\frac{4V_0D_s}{d}\right)^{1/2}, \tag{31}$$

whereas for $\sigma \ll 1$, it predicts

$$V = \left(\frac{4V_0\lambda_s}{d}\right)^{1/2}. \tag{32}$$

The result for $\sigma \ll 1$ is identical to the result of the steady-state model of Dittmer and Neumann, and an excellent agreement between the steady-state and time-dependent models is seen in Fig. 9(a) for $\sigma = 0.01$. For very large values of σ, the slopes of the lines for the two models in the linear regime are different. The steady-state model overestimates the whisker length in the linear region, as seen in Fig. 9(b).

2.4 Kinetic Model

The above surface-diffusion–controlled models are valid for whisker growth with a screw dislocation or for the VLS mechanism when the growth of whisker is by the physical vapor deposition. The screw dislocation or the liquid droplet serves as an active sink for the atoms to join at the advancing front. If both the screw dislocation and the liquid droplet are present, the whisker growth will be faster. For VLS growth, when the diameter is large, i.e., $d > 1$ μm, the capillarity effects are negligible, and the velocity does not vary significantly with diameter. In this case, the second term in Eq. (19) is negligible so that the growth is controlled by the direct impingement of atoms at the droplet surface, i.e., $V = V_0$.

When the whisker is grown by the chemical vapor deposition technique, the kinetics of the reaction can also be a rate-limiting factor. For example, for the silicon whisker growth in the chloride process, the following chemical reaction produces silicon:

$$SiCl_4 + 2H_2 \rightarrow Si + 4HCl. \tag{33}$$

The reaction is believed to occur at the interface. Since the vapor–liquid interface is rough, the reaction rate will be greater at the vapor–liquid interface than at the vapor–solid interface. Consequently, most silicon atoms are supplied by the liquid droplet, and the contribution of surface diffusion is negligible. The whisker growth, in this case, will be governed by the following processes:

1. chemical reaction at the vapor–liquid interface,
2. diffusion of silicon in the liquid phase, and
3. incorporation of atoms into the crystal at the liquid–solid interface.

The transport of atoms through the liquid would be quite rapid compared to the diffusion of adsorbed atoms on the surface of the whisker. This is due to the fact that the diffusion in the liquid is faster, and the diffusion distances in the droplet are very small. In case of compounds, it is also possible to have transport by surface diffusion (Wang and Fischman, 1994). Thus, if the diffusion in the liquid and the decomposition reaction are rapid, then the whisker growth will be controlled by the atomic attachment kinetics.

The growth rate of whiskers, controlled by the interface kinetics, is given by

$$V = A(\Delta\mu/kT)^n \tag{34}$$

in which A is the kinetic coefficient and n is an exponent whose value depends upon the attachment mechanism. $\Delta\mu$ is the effective difference in chemical potential between the Si atom in the nutrient phase (vapor or liquid) and in the solid, which, for very small whisker diameters, is given by the Gibbs–Thomson effect as

$$\Delta\mu = \Delta\mu_0 - 4\Omega\gamma_{vs}/d, \qquad (35)$$

where $\Delta\mu_0$ is the chemical potential difference at a plane interface, or for $d \to \infty$, and γ_{vs} is the vapor–solid interface energy. Substituting this value in the whisker-growth equation gives

$$V^{1/n} = A^{1/n}\frac{\Delta\mu_0}{kT} - A^{1/n}\frac{4\Omega\gamma_{vs}}{kT}\frac{1}{d}. \qquad (36)$$

This equation predicts a linear relationship between $V^{1/n}$ and $1/d$. A critical diameter, d^*, is thus obtained at which the whisker velocity goes to zero, and this is given by

$$d^* = 4\Omega\gamma_{vs}/\Delta\mu_0. \qquad (37)$$

Thus, if the value of d^* is measured experimentally, then the whisker velocity can be obtained as

$$V^{1/n} = A^{1/n}(\Delta\mu_0/kT)(1 - d^*/d). \qquad (38)$$

Experimental studies on silicon whisker growth show that the whisker growth-rate dependence on d appears to be close to $n = 2$, which is for the dislocation mechanism of crystal growth. However, no dislocations have been observed in the experiments on Si growth by the chemical vapor deposition method. Thus, a mechanism of two-dimensional nucleation is proposed by Mutaftschiev *et al.* (1965) and by Chernov and Lyubov (1968), which gives

$$V = B\exp\left(-\frac{\pi\Omega\gamma_e}{3kTh\Delta\mu}\right), \qquad (39)$$

where γ_e is the specific free energy of the edge of the nucleus, h is the height of the critical nucleus (see Sec. 2.5), and B is a constant.

2.5 Whisker-Growth Experiments

Several experimental studies have been carried out to examine the growth characteristics of whiskers (Dittmer and Neumann, 1960; Price, 1960; Parker *et al.*, 1963; Parker and Hardy, 1964; Budurov *et al.*, 1972). The main driving force for the whisker growth is the supersaturation, which controls the growth rate and diameter of whiskers. Controlled studies by physical vapor deposition have been carried out for different supersaturations to obtain quantitative information on whisker-growth velocity and diameter by using the field-emission microscope and by the direct observation of whisker growth. For small whiskers, examined in the field-ion microscope, an exponential growth law is observed (Gomer, 1958; Melmed and Gomer, 1961). For large whiskers a linear law is found by Dittmer and Neumann (1960). Thus, an exponential growth rate is observed initially, which becomes linear with time or length of the whisker, as predicted by the diffusion theory.

Whisker-growth experiments have also been carried out by using the VLS technique in which whiskers can be formed at a predetermined site and their growth kinetics can be quantitatively characterized. The basic characteristics of whiskers can be seen by examining the results of the Si whiskers grown from the vapor of different supersaturations by the VLS process. The supersaturation is controlled by the ratio of $SiCl_4/H_2$ in the vapor phase. A typical picture of the Si whiskers is shown in Fig. 10(a), which illustrates the relationship between the growth rate and the diameter of the whisker; i.e., thin whiskers grow more slowly than thick whiskers. This observation shows that capillarity effects, given by the Gibbs–Thomson effect, play a crucial role in the development of whiskers since a finer whisker consumes a larger fraction of driving force to form the surfaces. The quantitative measurements of the growth rate of whiskers are shown in Fig. 10(b), which shows a linear relationship between $V^{1/2}$ and $1/d$, where d is the diameter of the whisker. Different curves are for different supersaturations, with supersaturation increasing from line (1) to line (4). The extrapolation of the lines to $V = 0$ gives the critical diameter of a hemispherical nucleus for a given supersaturation, which is given

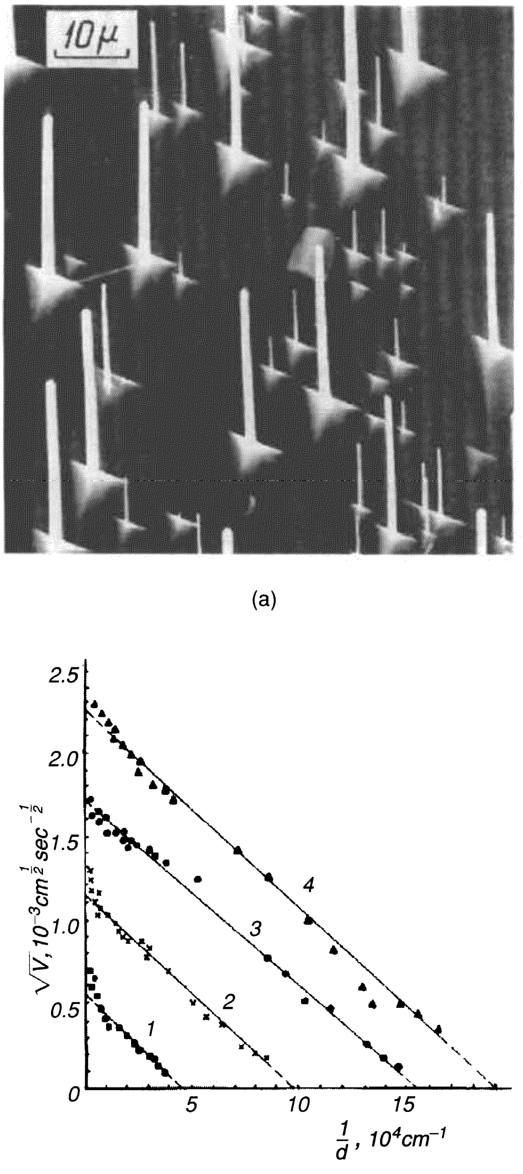

(a)

(b)

FIG. 10. (a) Growth of silicon whiskers from the vapor by the chloride process. (b) The variation in the whisker growth rate as a function of whisker diameter. (Givarzigov, 1978.)

by $4\gamma/\Delta G_V$, in which γ is the solid–vapor interface energy and ΔG_V is the Gibbs free-energy change per unit volume. These results are in agreement with Eq. (36) for $n = 2$, although they could also be fitted by the exponential growth law predicted for the two-dimensional nucleation process.

An unambiguous test of the whisker-growth theories has not been possible since many of the parameters are not known accurately. In fact, experimental results are generally compared to the theory to determine the growth law and then obtain the values of the parameters from the model.

ACKNOWLEDGMENTS

This work was carried out in part (RT) at the Ames Laboratory, which is operated for the U.S. Department of Energy by Iowa State University, under Contract No. W-7405-Eng-82, supported by the director of Energy Research, Office of Basis Energy Sciences. This work was supported (AK) by DOE Grant DE-FG02-92ER45471.

GLOSSARY

Anisotropy: Orientation dependence of interface energy and interface kinetic coefficient.

Anisotropy Parameter: Half of the percentage difference between the surface energy measured along the (100) direction and the (110) direction for a material with an underlying cubic symmetry.

Capillary Length: A characteristic length of the surface-energy effect.

Directional Solidification: Solidification under a positive temperature gradient in the liquid. The latent heat of fusion is dissipated through the solid.

Phase Field: A scalar field used to distinguish between the solid and liquid phases within a phenomenological mathematical description of first-order phase transformation.

Undercooling: A difference in temperature between the equilibrium freezing temperature and the bath temperature. The equilibrium temperature for a pure melt is given by the melting point, and that for an alloy is given by the liquidus temperature of the alloy.

Works Cited

Barbieri, A., Langer, J. S. (1989), *Phys. Rev. A* **39,** 5314–5325.

Ben-Amar, M., Brenner, E. (1993), *Phys. Rev. Lett.*, **71,** 589–592.

Ben-Amar, M., Moussallam, B. (1987), *Physica* **D25**, 155–164.

Ben-Jacob, E., Goldenfeld, N. D., Koteliar, B. G., Langer, J. S. (1984), *Phys. Rev. Lett.* **53**, 2110–2113.

Blakely, J. M., Jackson, K. A. (1962), *J. Chem. Phys.* **37**, 428–430.

Brenner, E., Temkin, D. (1995), *Phys. Rev. E* **51**, 351–359.

Brenner, E., Muller-Krumbhaar, H., Temkin, D. (1996), *Phys. Rev. E* **54**, 2714–2722.

Budurov, S., Stojcev, N., Avramov, I. (1972), *Kryst. Technik* **7**, 387–393.

Burton, W., Cabrera, N., Frank, F. C. (1951), *Phil. Trans. Roy. Soc. A* **243**, 299–358.

Caroli, B., Caroli, C., Roulet, B., Langer, J. S. (1986), *Phys. Rev. A* **33**, 442–452.

Chernov, A. A., Lyubov, B. Ya. (1968), in: N. N. Sheftat (Ed.), *Growth of Crystals*, Vol. 5A, New York: Consultants Bureau, pp. 7–23.

Dittmar, W., Neumann, K. (1958), in: R. H. Doremus *et al.* (Eds.), *Growth and Perfection of Crystals*, New York: Wiley, pp. 121–125.

Dittmar, W., Neumann, K. (1960), *Z. Electrochem.* **64**, 297–305.

Dougherty, A., Kaplan, P. D., Gollub, J. P. (1987), *Phys. Rev. Lett.* **58**, 1652–1655.

Drum, C. M., Mitchell, J. W. (1964), *Appl. Phys. Lett.* **4**, 164–165.

Edwards, P. L., Huang, S. (1966), *J. Am. Cer. Soc.* **49**, 122–125.

Fabietti, L. M., Trivedi, R. (1998), *J. Cryst. Growth* (in press).

Frank, F. C. (1949), *Disc. Faraday Soc.* **5**, 48–66.

Frenkel, J. (1926), *Z. Phys.* **37**, 572–609.

Gabor, T., Blocher, J. M. (1969), *J. Appl. Phys.* **40**, 2696–2704.

Givargizov, E. I. (1978), in: E. Kaldis (Ed.), Current Topics in Materials Science, Vol. 1, Amsterdam: North-Holland Publishing Company, pp. 79–145.

Givargizov, E. I., Sheftal, N. N. (1972), *Kryst. Technik* **7**, 37–41.

Glicksman, M. E., Schaefer R. J., Ayers, J. D. (1976), *Metall. Trans. A* **7**, 1747–1759.

Glicksman, M. E., Singh, N. B. (1989), *J. Cryst. Growth* **98**, 277–284.

Gomer, R. (1958), *J. Chem. Phys.* **28**, 457–464.

Gotoh, Y. (1972), *Japan J. Appl. Phys.* **11**, 1403–1412.

Hasiguti, R. R., Yagi, E., Nishike, U., Sakai, T. (1970), *J. Cryst. Growth* **7**, 117–119.

Hasimoto, S., Yamaguchi, A. (1996), *J. Mat. Sci. Eng.* **31**, 317–322.

Herring, C., Galt, J. K. (1952), *Phys. Rev.* **85**, 1060–1061.

Hirth, J. P., Lothe, J. (1968), *Theory of Dislocations*, New York: McGraw-Hill, p. 7.

Huang, H.-C., Glicksman, M. E. (1981), *Acta Metall.* **29**, 701–715.

Ihle, T., Müller-Krumbhaar, H. (1994), *Phys. Rev. E* **49**, 2972–2991.

Ivantsov, G. P. (1947), *Dokl. Akad. Nauk SSSR* **58**, 567–569.

Karma, A., Rappel, W.-J. (1996a), in: L.-Q. Chen (Ed.), *Mathematics of Microstructure Evolution*, Warrendale, PA: TMS and SIAM.

Karma, A., Rappel, W.-J. (1996b), *Phys. Rev. E* **53**, R3017–R3020.

Karma, A., Rappel, W. J. (1996c), *Phys. Rev. Lett.* **77**, 4050–4053.

Karma, A., Rappel, W.-J. (1997), *Phys. Rev. E* **5**, 1–30.

Kessler, D. A., Levine, H. (1986), *Phys. Rev. B* **33**, 7867–7870.

Kessler, D. A., Koplik, J., Levine, H. (1985), *Phys. Rev. A* **31**, 1712–1717.

Kessler, D. A., Levine, H. (1988), *Acta Metall.* **36**, 2693–2706.

Kobayashi, R. (1993), *Physica* **D63**, 410–423.

Kobayashi, R. (1994), *Exp. Math.* **3**, 60.

Koparanova, N., Zlatev, Z., Genchev, D., Popovich, G. (1994), *J. Mat. Sci. Eng.* **29**, 103–109.

Kurz, W., Fisher, D. J. (1992), *Fundamentals of Solidification*, Zurich: Trans Tech Publ.

Langer, J. S. (1980), *Rev. Mod. Phys.* **52**, 1–28.

Langer, J. S. (1986), *Phys. Rev. A* **33**, 435–441.

Langer, J. S. (1989), *Science* **243**, 1150–1153.

Langer, J. S., Müller-Krumbhaar, H. (1977), *J. Cryst. Growth* **42**, 11–14.

Meiron, D. I. (1986), *Phys. Rev. A* **33**, 2704–2715.

Melmed, A. J., Gomer, R. (1961), *J. Chem. Phys.* **34**, 1802–1812.

Minagava, S., Gejyo, T. (1969), *J. Am. Cer. Soc.* **52**, 282–284.

Muschol, M., Liu, D., Cummins, H. Z. (1992), *Phys. Rev. A* **46**, 1038–1050.

Mutaftschiev, B., Kern, R., Georges, G. (1965), *Phys. Lett.* **16**, 32–33.

Nabarro, F. R. N., Jackson, K. A. (1958), in: R. H. Doremus *et al.* (Eds.), *Growth and Perfection of Crystals*, New York: Wiley, pp. 13–101.

Nash, G. E., Glicksman, M. E. (1974), *Acta Metall.* **22**, 1283–1290.

Papapetrou, A. (1935), *Z. Kristallogr.* **92**, 89–130.

Parker, R. L., Anderson, R. L., Hardy, S. C. (1963), *Appl. Phys. Lett.* **3**, 93–95.

Parker, R. L., Hardy, S. C. (1964), in: E. Rutner *et al.* (Eds.), *Condensation and Evaporation of Solids*, New York: Gordon and Breach, p. 649–662.

Pelcé, P., Pomeau, Y. (1986), *Stud. Appl. Math.* **74**, 245–258.

Price, P. B. (1960), *Philos. Mag.* **5,** 473–484.

Regis, M. (1969), *Acta Metall.* **17,** 1281–1289.

Rubinstein, E. R., Glicksman, M. E. (1991), *J. Cryst. Growth* **112,** 84–96.

Ruth, V., Hirth, J. P. (1964), *J. Chem. Phys.* **41,** 3139–3149.

Schmidt, A. (1996), *J. Comput. Phys.* **125,** 293–312.

Sears, G. W. (1953), *Acta Metall.* **1,** 457–459.

Sears, G. W. (1955), *Acta Metall.* **3,** 361–366.

Simmons, J. A., Oser, H., Coriell, S. R. (1967), in: H. S. Peiser (Ed.), *Crystal Growth,* New York: Pergamon Press, pp. 255–264.

Simmons, J. A., Parker, R. L., Howard, R. E. (1964), *J. Appl. Phys.* **35,** 2271–2272.

Temkin, D. E. (1960), *Dokl. Acad. Nauk SSSR* **132,** 1307–1310.

Trivedi, R. (1970), *Acta Metall.* **18,** 287–296.

Trivedi, R., Somboonsuk, K., (1984), *J. Mat. Sci. Eng.* **65,** 65–74.

Wagner, R. S. (1967), *J. Appl. Phys.* **38,** 1554–1560.

Wagner, R. S., Ellis, W. C. (1964), *Appl. Phys. Lett.* **4,** 89–90.

Wagner, R. S., Ellis, W. C. (1965), *TMS-AIME* **233,** 1053–1064.

Wagner, R. S., Ellis, W. C., Jackson, K. A., Arnold, S. M. (1964), *J. Appl. Phys.* **35,** 2993–3000.

Wagner, R. S. (1970), in: A. P. Levitt (Ed.), *Whisker Technology,* New York: Wiley pp. 47–119.

Wang, H., Fischman, G. S. (1994), *J. Appl. Phys.* **76,** 1557–1562.

Wang, S.-L., Sekerka, R. F. (1996), *Phys. Rev. E* **53,** 3760–3776.

Warren, J. A., Boettinger, W. J. (1995), *Acta Metall. Mater.* **A43,** 689–703.

Webb, W. W. (1958), in: R. H. Doremus *et al.* (Eds.), *Growth and Perfection of Crystals,* New York: Wiley, p. 230–238.

Further Reading

Kessler, D. A., Koplik, J., Levine, H. (1988), *Adv. Phys.* **37,** 255–339.

Langer, J. S. (1987), in: J. Souletie, J. Vannimenus, R. Stora (Eds.), *Chance and Matter,* Lectures on the Theory of Pattern Formation, Les Houches, Session XLVI, Amsterdam: North-Holland, pp. 629–711.

Pelcé, P. (1988), *Dynamics of Curved Fronts,* Boston: Academic.

Trivedi, R., Kurz, W. (1994), *Int. Mat. Rev.* **39,** 49–74.

Givargizov, E. I. (1978), in: E. Kaldis (Ed.), *Current Topics in Materials Science,* Vol. 1, Amsterdam: North-Holland pp. 79–145.

WIND ENERGY

JERRY MCNERNEY, *Pleasanton, California, U.S.A.*
DAVE RICHARDSON, *Electric Power Research Institute, Palo Alto, California, U.S.A.*
ED MCCARTHY, *Martinez, California, U.S.A.*
JOHN HULLS, *Point Reyes, California, U.S.A.*

Introduction 461
1. **History** 461
2. **The Wind** 462
2.1 The Wind Resource 462
2.2 Statistical Variation and Turbulence 463
2.2.1 Wind-Speed Frequency Distribution 463
2.2.2 Turbulence 464
3. **Turbines** 464
3.1 Power in the Wind 464
3.2 Mechanical Aspects of Wind Turbines 466
3.3 Horizontal-Axis Wind-Turbine Trade-offs 466
3.4 Horizontal-Axis Wind-Turbine Input Loads 466
3.5 Horizontal-Axis Wind-Turbine Dynamics 468
3.6 Vertical-Axis Wind Turbines 469
4. **Airfoils** 470
4.1 Aspects of Two-Dimensional Flow about an Airfoil 470
4.1.1 Forces on the Airfoil 470
4.1.2 Viscous Effects and Boundary Layer 472
4.1.3 Definition of the Airfoil Shape 473
4.2 Determination of Airfoil Characteristics 473
5. **Electrical-Power Generation** 474
5.1 Generators 474
5.2 Power Electronic Converters 475
6. **Economics** 476
Glossary 477
Works Cited 477

INTRODUCTION

Wind-turbine technology has developed from a crude science in the 1970s to sophisticated economic devices utilizing the latest in power electronics, aerodynamics, and mechanical drive-train designs. Wind turbines now available in the market place offer renewable energy at 5 cents per kilowatt-hour economics (Smith and Ilyin, 1991). The design of wind turbines involves using low-speed aerodynamics, complex mechanical system design, and power conversion to provide low-cost and reliable power. This article covers a brief history of recent wind-energy technology, the wind resource, mechanical design principles, power conversion, and the economics of wind energy from a design and commercialization perspective.

1. HISTORY

Interest in renewable sources of energy, including the contemporary wind turbine, as commercial power sources erupted in the mid 1970s when concerns about the environmental effects of conventional energy sources gained momentum simultaneously with the OPEC oil embargoes. The U.S. federal and California state renewable-energy tax credits together with the Public Utilities Regulatory Policy Act of 1987 (PURPA) requiring public utilities to purchase power from independent power producers at the avoided cost created the necessary economic climate. Investors joined with innovative technologists to form several small enterprises dedicated to the development of wind-turbine designs capable of competing with

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

more conventional power sources in the utility market. Lessons learned during the early years of utility-grade wind-turbine development have led to understanding of the resource and technology sufficient for the evolution of successful economic wind turbines. Governmental assistance, in the form of basic research at national institutions and some forms of federal subsidies, also contributed to necessary understanding of the technology and eventual success of the industry. After the tax credits expired in 1986, a number of wind-turbine manufacturers and developers remained in business, and California dominated the world in wind-energy production.

Today, wind-power plants are going on line in different sites around the world. Given the advanced state of the wind-turbine technology, the available worldwide wind resource, and an expected increase in demand for energy sources not damaging to the environment, it is sensible to expect that wind energy will become a significant component of energy supplies in the near-term future. In this article, some of the basics on how a wind turbine operates as well as some of the issues in wind-turbine utility applications are discussed.

2. THE WIND

Wind is the movement of the air. This movement can be consistent or inconsistent; can occur all the time, only at night, or only during the day. There can be a seasonal variation with the strongest winds occurring during the summer or during the winter. Understanding the wind and integrating that understanding with technology that can harvest the electrical energy is the goal of wind-resource assessment.

2.1 The Wind Resource

The wind resource is a general term used to describe the wind energy at a location (EPRI, 1995). The characteristics of the wind resource include wind speed, wind direction, wind turbulence, wind shear, and air density. To describe a wind resource means to describe quantitatively the primary features of each of these parameters. Since wind power is directly related to the cube of the wind speed, it is obvious that wind speed is the most important parameter in resource assessment.

To identify promising sites for wind-energy development, areas with annual average wind speeds of 6.5 m/s at typical hub heights of wind turbines (25 to 40 m) should be considered. Annual wind speed is the first of several important indicators that can be used to characterize a wind resource. However, average wind speed alone is not a reliable indicator of the power in the wind. Sites having the same annual average wind speed may not have the same wind-power density as a result of differences in the wind-speed frequency distribution. The wind-speed frequency distribution gives the number of hours that the wind occurs at any given wind speed. Table 1 (Elliott, 1987) is a comparison of the average wind speeds and wind-power densities at 10 m for three sites—Culebra, Puerto Rico; Tiana Beach, New York; and San Gorgonio, California. The wind-speed frequency distributions are substantially different at the three locations, resulting in different wind-power densities.

Wind direction is another important parameter. Individual wind turbines, used for producing electrical energy, are usually sited in large parks or arrays. Understanding the wind direction is critical for siting the individual wind turbines on account of the possible wake effects between turbines and for proper placement of turbines on terrain features. An area with a single prevailing wind direction will have a different array design than an area with multidirectional winds, which requires more spacing between turbines.

Turbulence in the wind can influence operations and maintenance costs associated with a wind park and can influence the fatigue life of wind-turbine components. Turbulence is the result of rapid changes in the

Table 1. Comparison of annual average wind-power density at three sites.

Site	Wind speed (m/s)	Wind-power density (W/m^2)
Culebra, Puerto Rico	6.3	220
Tiana Beach, New York	6.3	285
San Gorgonio, California	6.3	365

wind speed or the wind direction. A commonly used indicator of wind-speed turbulence is the term *turbulence intensity* (TI), which is defined as the standard deviation of the wind speed divided by the average wind speed for the period.

Wind shear is the change in wind speed with increasing elevation. Depending upon the topography, flow structure, and surface roughness, the wind shear can be positive (wind speed increasing with increasing height), neutral (wind speed unchanged with increasing height), or negative (wind speed decreasing with increasing height). Wind-energy researchers have found that wind measurements at hub height are critical to the success of any project. Extrapolation techniques exist for estimating the wind speed at other levels in the atmosphere. The first is the standard wind-speed power law, and the second is the logarithmic power law. The only adjustment is a height adjustment from the published measurement level to a standardized height of 30 m. Several formulas exist for adjustment of wind speed from one height to a second height. The first formula is the logarithmic wind profile:

$$V(z) = (V_*/\kappa)\ln(Z/z_0), \tag{1}$$

where $V(z)$, in meters per second, is the wind speed at height Z, Z is the elevation above the ground, V_* is the friction velocity (m/s), κ is the von Karman constant (0.4), and z_0 is the roughness length. The second formula is the wind-speed power law:

$$V_2 = V_1(Z_2/Z_1)^{\beta}, \tag{2}$$

where V_2, in meters per second, is the calculated wind speed at height 2; V_1 is the observed wind speed at height 1; Z_1 is the elevation of the observed wind speed, in meters, above the ground; Z_2 is the elevation of the calculated wind speed, in meters, above the ground; and β is the wind-speed power-law exponent. The value $\beta = \frac{1}{7}$ is the most common value used in the wind industry.

Air density will also influence the production of wind energy. Air density is calculated from daily, monthly, seasonal, or annual temperature and pressure information for the site. Basic formulas for calculating air density are found in Rohatgi and Nelson (1994).

2.2 Statistical Variation and Turbulence

2.2.1 Wind-Speed Frequency Distribution The most critical information concerning the wind resource is the wind-speed frequency distribution. Several statistical relationships are available that are commonly used to develop these frequency distributions—the Weibull function and a special case of the Weibull function, the Rayleigh function.

The Weibull function is a two-parameter function that is used to estimate wind-speed frequency distribution. It is defined as follows:

$$f(V) = (k/c)(V/c)^{k-1}\exp(-V/c)^{k}, \tag{3}$$

where $f(V)$ is the Weibull probability density function for V in (m/s), and c, expressed in meters per second, is the Weibull scale factor. k is the Weibull shape factor, which describes the distribution of the wind speeds. The Weibull function is commonly used because it has been found to give a good fit to the observed wind-speed data at the surface (Wentink, 1976; Justus *et al.*, 1976) and in the upper air (Barnes *et al.*, 1975).

A special form of the Weibull probability density function, known as the Rayleigh distribution, has been found typically to represent the wind-speed frequency distributions at most sites. In the Rayleigh distribution, the shape factor (k) is assumed to have a value of 2.0. The Rayleigh distribution is expressed as

$$R(V) = (\pi V/2\mathcal{V}^2)\exp(-\pi V^2/4\mathcal{V}^2). \tag{4}$$

In this equation, $\mathcal{V}$ is the long-term average wind speed in meters per second. Corotis (1977) and Cliff (1977) found that the Rayleigh distribution provided a very good fit between the theoretical and actual wind-speed distributions.

If actual wind-speed data are available, a widely used procedure in wind-energy analysis is to assemble the wind-speed data into a histogram that shows the number of hours or observations of wind speed in discrete 1-m/s bins (Rohatgi and Nelson, 1994). A wind-speed frequency histogram is then obtained by dividing the number of observations in each bin by the total number of observations in the data set.

2.2.2 Turbulence Wind turbulence causes random load and output power fluctuations, which stress the entire wind-turbine structure—rotor, gearbox, generator, tower, and foundation. As wind-speed variations are highly irregular and cannot be described in a deterministic manner, these variations are best described by TI and other, more sophisticated methods, which include aspects of frequency, amplitude, and time. An excellent discussion on the aspects of wind turbulence is included in Rohatgi and Nelson (1994).

While TI is commonly used as a gross measure of turbulence, Barnard *et al.* (1995) and Wendell *et al.* (1993) have pointed out that there may be better indices to classify wind turbulence. On the basis of the turbulence characterization studies undertaken by Battelle, Wendell *et al.* (1993) proposed two other terms to characterize more adequately the turbulence characteristics of a site, specific to wind-turbine technologies. These two terms are turbulence engulfing gust (TEG), which refers to those eddies that affect the entire rotor disk at once, and turbulence fluctuating shear (TFS), which refers to those eddies that only affect a portion of the rotor area through shear flow. Recently, Barnard *et al.* (1995) proposed a model relating conventional wind-speed measurements and wind-turbine blade fatigue. Based on case studies obtained from the National Renewable Energy Laboratory, this model is shown to be a much better predictor between low-turbulence and high-turbulence cases than the TI index.

3. TURBINES

3.1 Power in the Wind

Before any mechanical aspects of wind turbines can be discussed, the driving force of the machines must be considered. For a given wind speed V, the power in the wind depends on the area presented through the equation

$$P_T = \tfrac{1}{2}\rho A V^3, \tag{5}$$

where the subscript T of P refers to the total power existing in the wind, ρ is the air density, and A is the presented area, or swept area as it is commonly referred to by wind engineers. There is a theoretical limit, called the Betz limit, of approximately 59% to the efficiency a wind turbine can extract power from the wind. The Betz limit can be derived by an actuator disc model as in Eggleston and Stoddard (1987). Any wind turbine will have an intrinsic efficiency, called the power coefficient C_P, which varies with wind speed. It is customary to give the power coefficient as a function of the ratio of linear blade-tip speed to wind speed or simply the tip-speed ratio, λ, the dimensionless quantity

$$\lambda = r\Omega/V, \tag{6}$$

where r is the radius of the rotor and Ω is the angular velocity of the wind turbine in radians per second. The numerator in this equation is the tip speed of the blade. The aerodynamic performance of a wind turbine is specified when the power coefficient is given as a function of the tip-speed ratio over the range of wind speeds in which the wind turbine operates. C_P is determined experimentally, or through a detailed aerodynamic calculation as described below. Figure 1 is a typical C_P curve. C_P has a theoretical maximum dictated by the Betz limit, but most wind turbines have a maximum C_P of 50% to 52%. This formulation allows the power extracted by a wind turbine to be given by the closed-form equation

$$P = \tfrac{1}{2}\rho A V^3 C_P. \tag{7}$$

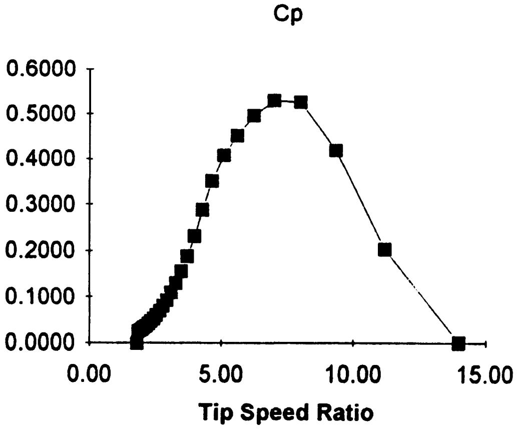

FIG. 1. Hypothetical curve of C_P.

The aerodynamics of the blades determines the power coefficient. If the lift and drag coefficients of the blades are known, then the power coefficient can be calculated using an iterative approach such as the one derived in Wilson *et al.* (1976), or in Eggleston and Stoddard (1987). Many computer programs are available that perform this calculation. Once the power coefficients are known, a power curve as a function of wind speed can be calculated. Figure 2 is a typical curve of power versus wind speed. There are three critical wind speeds for any wind turbine. The first is the cut-in wind speed, v_{cutin}, that at which the wind turbine can first produce power. The second critical point, v_{rated}, is the lowest wind speed at which the turbine reaches its rated power. The last critical point, v_{cutout}, is the high-wind cutout speed above which the wind turbine is shut down. These three points identify four distinct regions of the power curve. In the first region, left of v_{cutin}, the wind speed is too low to generate positive power. In the second region, between v_{cutin} and v_{rated}, power increases from zero to rated power. In the third region between v_{rated} and v_{cutout}, the power is constant at the wind turbine's rated level. This is idealized, and most wind turbines vary in output power somewhat after rated power. In the last region, above v_{cutout}, the wind turbine is shut down to prevent excessive winds from damaging the equipment.

The power of a wind turbine is regulated in one of two ways. The first method, called pitch regulation, depends on a mechanism to pivot the blades to higher angles of attack, making the blades operate less efficiently aerodynamically and shedding excess power in the wind. This approach requires that the blade pitch angle be controlled on a continuous basis when the wind speed is greater than v_{rated}, to track the changing wind speed. The second method used to regulate wind-turbine power, called stall regulation, depends on the blades being designed to go into aerodynamic stall above v_{rated}. In this passive approach, no blade-pitching mechanisms are required, although some stall-regulated wind turbines use some sort of aerodynamic spoiling mechanism to stop the turbine.

A useful quantity used in comparing the performance of wind turbines is the annual energy capture per square meter of swept area. This quantity depends on the wind resource. If the wind speed is partitioned into uniform intervals, with V_i being the center of interval i, then annual energy per square meter of swept area for a wind turbine is given by the equation

$$e = \tfrac{1}{2}\rho \sum_{i=1}^{V_{\max}} [F(V_i + 0.5) - F(V_i - 0.5)]V_i^3 C_{P,i}, \tag{8}$$

where $F(V_i)$ is the wind-speed distribution function for the site. When the wind-speed distribution is not known, the Rayleigh distribution is normally used. The standard air density at sea level is $\rho = 1.2$ kg/m^3. Using an annual average wind speed of 8 m/s, and Betz limit for C_P, the annual energy captured per square meter of swept area is 3038 kW · hr/yr · m^2. This represents the theoretical limit for energy capture for a wind turbine. The power coefficient can be as high as 0.53

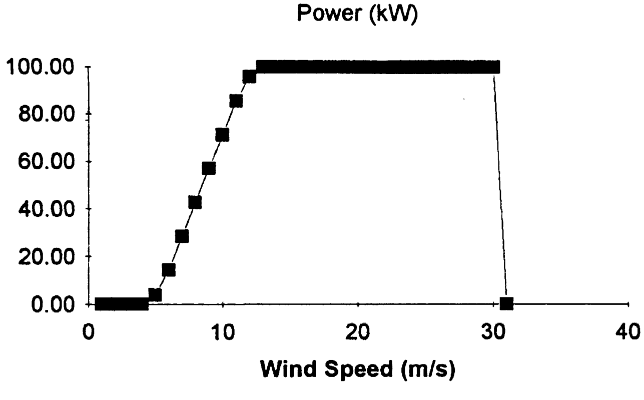

FIG. 2. Hypothetical power curve for 100-kW wind turbine.

for the optimum wind speed but is much less than that value for other operating wind speeds. The annual energy per square meter for the hypothetical 100-kW wind turbine of Figures 1 and 2 is 1700 kW·h/yr·m^2, which is a more realistic value. Variable-speed operation of a wind turbine is an advantage in terms of energy capture, because the rotor changes speed allowing the blade tips to follow the maximum power coefficient over a wide range of operating speeds, thereby capturing more energy. Constant-speed wind turbines are often designed so that the maximum power coefficient is aligned with the design wind-speed distribution to give the maximum energy capture.

3.2 Mechanical Aspects of Wind Turbines

Wind turbines can be broadly categorized into two types: horizontal-axis wind turbines (HAWTs) in which the rotating axis is horizontal, parallel to the wind direction, and supports the blades in a plane perpendicular to the ground, and vertical-axis wind turbines (VAWTs) in which the rotating axis is vertical and the wind turbine has the appearance of an egg beater. HAWTs are by far the more common type of wind turbine and are commercially available in any size ranging from a couple of hundred watts to more than 1 MW. VAWTs are commercially available through a small number of manufacturers but appear to have a number of advantages that may make them cost competitive in the near future. HAWTs and VAWTs are each discussed in separate sections below so that the fundamental properties of each system can be considered.

One property that all wind-turbine types have in common, referred to as the "square-cubed law," illustrates how the cost effectiveness of a wind-turbine architecture is governed by the size of the wind turbine. The square-cubed law states that the loads on the components of a wind turbine are proportional to the cube of the diameter of the rotor, while the energy capture is proportional to the square of the diameter of the rotor. The consequence of this law is that the cost of the wind turbine will increase faster than the benefit for large wind turbines, and for any design type, an optimum design size will exist at the point at which the increasing variable costs overtake the fixed costs of the system. Holley *et al.* (1987) reported this law for particular wind-turbine designs. The optimum square-cube rotor diameter varies widely with the design. Typical optimum values fall in the diameter range corresponding to between 300 and 600 kW for HAWTs. The square-cubed law optimum size is not as easy to define for VAWTs, because the height of a VAWT can vary independently from the diameter, but the optimum sizes of VAWTs on a cost-of-energy (COE) basis are thought to be somewhat larger than those of HAWTs.

3.3 Horizontal-Axis Wind-Turbine Trade-offs

Figure 3 is an illustration of a typical two-bladed HAWT, in which the basic components have been identified. HAWTs have the following design options and trade-offs:

1. 2 or 3 blades. Two blades cost less than three blades, but two-bladed systems require a more complicated teetered hub.
2. Free yaw or active yaw control. Free-yaw systems risk higher yaw rates, which induce large gyroscopic loads, but effective yaw-control systems are notoriously difficult to design.
3. Upwind or downwind orientation. Upwind wind turbines are quieter and capture more energy but usually require an active yaw-control system.
4. Power regulation by aerodynamic stall or by blade pitch. Stall offers mechanical simplicity but very difficult aerodynamics for limiting power. Blade pitch offers good control of the average aerodynamic power, but it requires complicated mechanical control mechanisms and has large load fluctuations.
5. Constant-speed or variable-speed operation. Variable-speed operation allows the rotor to operate at the maximum aerodynamic efficiency in below-rated winds, and offers potentially better power quality, but it requires costly and complicated electronics and has higher losses.

3.4 Horizontal-Axis Wind-Turbine Input Loads

Most loads entering a wind turbine originate from aerodynamic forces at the blades. Figure 4 shows a blade section. Figure 4(a)

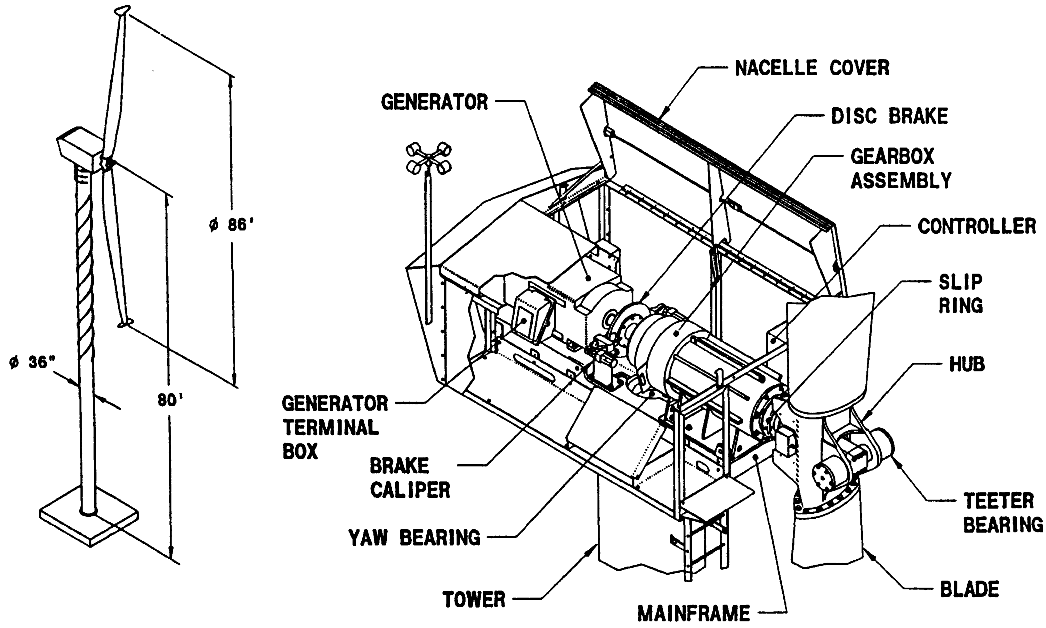
ϕ 86'
ϕ 36"
80'
GENERATOR
NACELLE COVER
DISC BRAKE
GEARBOX ASSEMBLY
CONTROLLER
SLIP RING
HUB
GENERATOR TERMINAL BOX
BRAKE CALIPER
YAW BEARING
TOWER
MAINFRAME
TEETER BEARING
BLADE

FIG. 3. Typical two-bladed horizontal-axis wind turbine.

is an illustration of the wind input on the blade section, and 4(b) is an illustration of the resulting loads. The forces on an airfoil are normally resolved into lift force F_L, and drag force F_D. F_L and F_D can be combined to form the resultant aerodynamic force, F_A. For the purposes of wind-turbine load analysis, it is more convenient to resolve F_A into the lead-lag force F_{LL}, in the direction of the blade chord, and the flatwise force F_{FW}, perpendicular to the blade chord. In addition to the aerodynamic forces, there is an aerodynamic moment M_A, which tends to pitch the blade; i.e., the moment vector points in the direction of the blade length. If the blade element is twisted, pitched, or coned, if the rotor is tilted, if the rotor is yawed out of the wind, or if blade vibration has translated or changed the orientation of the blade element, then appropriate coordinate systems

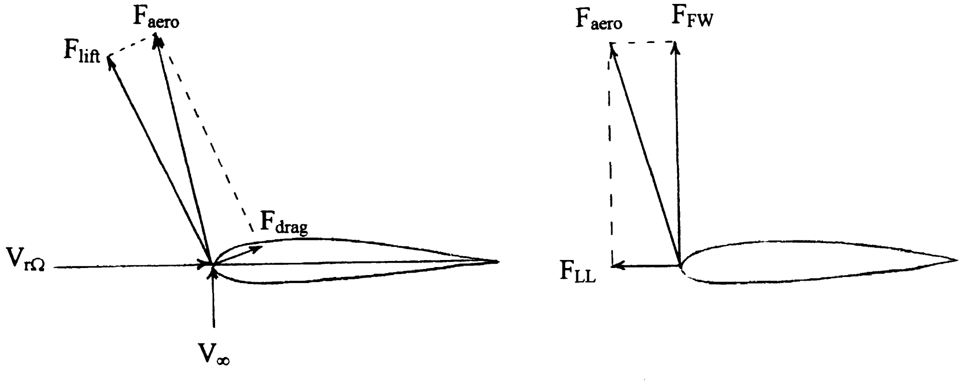
F_{aero}
F_{lift}
F_{drag}
$V_{r\Omega}$
V_∞
F_{aero}
F_{FW}
F_{LL}

(a) (b)

FIG. 4. Blade element with wind input and resulting aerodynamic loads.

need to be invoked. At any time increment, the loads F_{LL}, F_{FW}, and M_{A} can be calculated for each blade section and summed over the entire blade to find the forces and moments at the blade root–hub intersection. When the loads are combined at the hub from the two or three blades, the lead-lag forces combine to form the productive torque on the main shaft, and the flatwise forces combine to form the thrust load on the system. The aerodynamic moments affect the blade root loads, and the blade pitch link mechanism if one exists, but normally do not combine to form global loads. As mentioned above, any changes from the simple orientation of the blade elements result in off-axis loads, which combine to form the yaw forces, among others. Yaw forces cause the rotor to turn into or out of the prevailing wind direction. The yaw rate combines with the rotor angular velocity and the rotor inertia to form gyroscopic bending moments on the main shaft. These loads are transmitted to the bearings, transmission, frame, yaw bearing, tower, and foundations.

It is significant to note that for horizontal-axis wind turbines, the wind loading is highly stochastic. The inertial subrange of the wind, at which most energy is created, has a strong random component. The blades rotate through turbulent eddies, ground boundary-layer–induced wind shear, and tower shadow, which all contribute to the varying nature of the loads. Because of the highly varying nature of the loads, several HAWT components will be designed primarily by random loads. The design driver can be either fatigue considerations or extreme limits. Components will be designed by high-wind survival loads or by dynamic operating load cases. International standards now exist that specify which load cases must be satisfied to obtain certification of wind turbines. The International Electrotechnical Commission (IEC) (1994) establishes four wind-speed classes to be used in establishing loads, with annual average wind speeds of 10, 8.5, 7.5, and 6 m/s, respectively, and an average turbulence intensity of 0.17 for all four classes. Standard IEC 1400-1 should be consulted for international design standards.

3.5 Horizontal-Axis Wind-Turbine Dynamics

To obtain a complete picture of the loads on a wind turbine, the dynamics of the system must be addressed. Blade vibrations have a considerable effect on loads both at the blade-hub connection and on the blade itself. Drive-train dynamics affect how the loads at the hub are transmitted to the rest of the system, influencing the power generated as well as the loads that affect design of the transmission, frame, bolts, yaw bearing, tower, and foundation.

The blades can best be modeled dynamically either by a finite-element approach or by assigning modal coordinates in a generalized coordinate system. The drive train can be modeled as a two–degree-of-freedom lumped-mass system as shown in Fig. 5.

Using Fig. 5 as a guide, the following two equations of motion may be written down:

$$I_R\alpha_R + d(\Omega_R - \Omega_G) + k(\theta_R - \theta_G) = Q_R, \tag{9}$$

and

$$I_G\alpha_G - d(\Omega_R - \Omega_G) - k(\theta_R - \theta_G) = D(\Omega_G - \Omega_0) = Q_G. \tag{10}$$

Given the input torque Q_R, the rotor and generator inertias I_R and I_G, the torsional

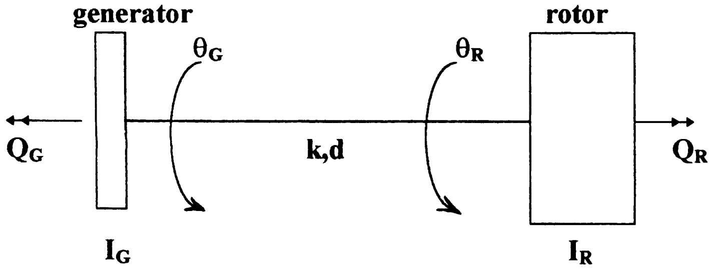

FIG. 5. Drive-train free-body diagram.

stiffness k, the structural damping d, the generator synchronous speed Ω_0, and the generator slip constant D, the two equations may be solved numerically to determine the rotor and generator speeds, Ω_R and Ω_G. This simple model will predict the dynamic motion of the drive train. Notice that in the equations, the absolute angles θ_R and θ_G are never needed. Since the model includes the transmission, transmission losses may be added. Also, care needs to be taken to insure that the inertias I_R and I_G are referred to the same speed. If a more detailed model is needed to predict, for example, the behavior of the yaw system, then the equations above need to be expanded to include additional variables and constants.

3.6 Vertical-Axis Wind Turbines

Vertical-axis wind turbines have historically played a less significant role in the commercial wind-energy industry. The reasons for this range from the lower aerodynamic efficiency of VAWTs compared to HAWTs, to the tendency of markets to accept the more familiar technology. Since the VAWT blade does not require twist, uniform blades are satisfactory. This allows the potential cost of the blades to be significantly lower per foot than for HAWTs, thus offsetting the requirement of significantly more blade length than for HAWTs.

The concept for the modern vertical-axis wind turbine was introduced by the French engineer Darrius, who observed that if the blades are formed into the shape that a spinning rope would assume if it were constrained on the ends, then the operating stresses in the blades would be mostly in tension, the load direction in which blades offer the most strength. In addition, constraining the blades at the top allows guy cables to be utilized to stiffen the tower top. This concept was put into practice in the 1970s and 1980s by industries and government laboratories in the United States and Canada. The early VAWTs showed potentially excellent longevity but were not cost competitive with many HAWTs. However, improved aerodynamics and reduced blade manufacturing costs now provide the technology to make VAWTs potentially as cost effective as HAWTs, particularly with larger sizes.

Vertical-axis wind turbines present the designer a different type of problem than HAWTs. The primary design loads for HAWTs are likely to result from the random fluctuations in the wind. For VAWTs, the primary consideration is the stiffness of the system. The fundamental frequency of the central tube decreases with increasing rotational speed and must be kept above zero. This requires the central tube to have a large inertia and the guy cables to provide considerable stiffness. Once the stiffness requirements are satisfied, then the system will most likely be able to withstand whatever operating and high-wind–survival loads are generated. The blades have a similar problem. Rotation causes centrifugal softening of the blades, which must maintain enough stiffness to resist divergence. Larger VAWTs are also subject to blade flutter problems, requiring the blades to have added stiffness.

VAWTs are subject to large periodic loads that are generated during operation. VAWT blades move through a wide range of angles of attack during each rotation, causing large fluctuations in aerodynamic loads with frequencies that are multiples of the rotational frequency. One tool that is useful for identifying potential resonance problems is a Campbell diagram, which displays how the excitation frequencies cross the structure's natural frequencies as a function of rotor speed. Figure 6 shows a Campbell diagram for the Sandia 34-m VAWT in Bushland, Texas (Ashwell, 1992). The abscissa of Fig. 6 is the operating speed of the wind turbine, and the ordinate is the frequency. The horizontal lines represent the different natural frequencies of the system, which vary with operating speed. The diagonal lines emanating from the origin are the 1 per revolution or 1P, 2P, 3P, 4P, and 5P frequencies at which structural members can be excited by aerodynamic forces as they rotate in and out of the wind. Crossings of any natural frequency with any per revolution excitation at a rotational speed will result in system resonances at that rotational speed and must be avoided. The abbreviations represent the different natural frequencies; for example, 1F, 2F, and 3F represent the first, second, and third blade flatwise motion of the blade flapping perpendicular to the direction of rotation. The tower in-plane and out-of-plane (of the blades) vibrations are denoted by TI and

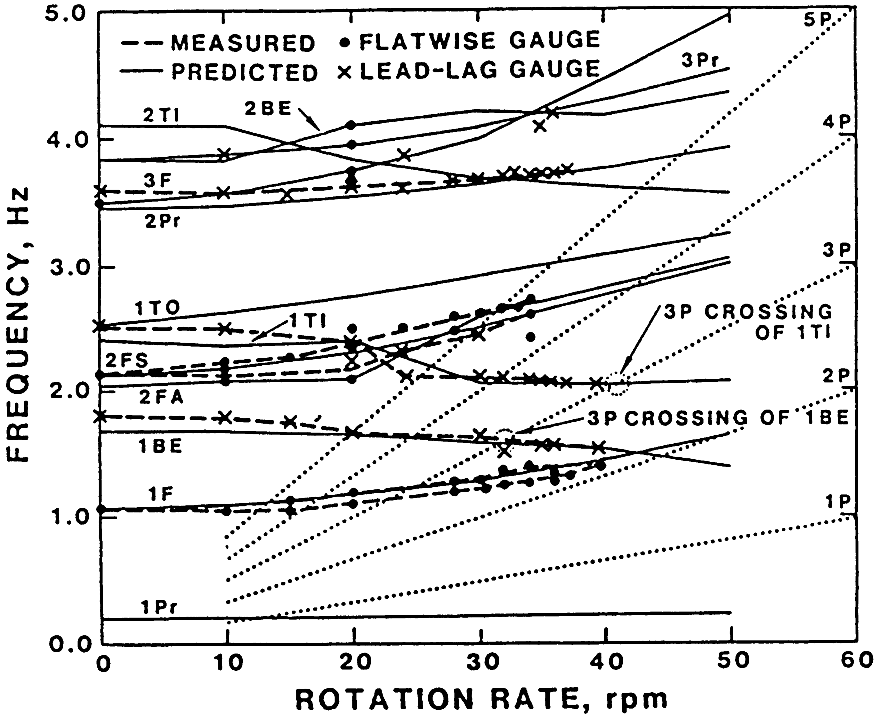

FIG. 6. Campbell diagram for the Sandia 34-m VAWT.

TO, respectively. This type of resonance is avoided in three-bladed vertical-axis wind turbines.

The rotor of a VAWT can also be modeled by a simple system, such as that given for the HAWT.

4. AIRFOILS

Development of suitable airfoils is a crucial element of modern wind-turbine design. The airfoil's function is to exert an influence on the free-stream flow so that lift is produced in such a manner as to enable the rotor system of the turbine to produce power. Successful wind-turbine airfoil design combines three main factors, which are

1. an understanding of the more important aspects of two-dimensional airfoil theory,
2. an analysis of the aerodynamic environment in which the airfoil will operated, and
3. a determination of the airfoil characteristics that are most desirable for the wind-turbine operational applications.

4.1 Aspects of Two-Dimensional Flow about an Airfoil

4.1.1. Forces on the Airfoil Basic airfoil data are set forth in the form of nondimensional force and moment coefficients as shown in Fig. 7. Of principal interest are the lift, drag, and moment coefficients, expressed for two-dimensional airfoils as c_l, c_d, and c_m, respectively. They are usually expressed in terms of angle of attack α, or with respect to each other, the most common being $c_l(\alpha)$, $c_d(\alpha)$, $c_m(\alpha, c_l)$. From c_l and c_d, the normal (c_n) and chordwise (c_c) force coefficients can be calculated. While the difference between c_l and c_n is usually small, the relationship between c_c and c_d can be very significant. When the c_c component is directed forward, a driving force results, thus enabling the rotor to produce power.

The pitching moment and aerodynamic center (designated by the subscript ac) are of significance in turbine design because of structural considerations. Following the generally accepted NACA (National Advisory Committee on Aeronautics) practice, the

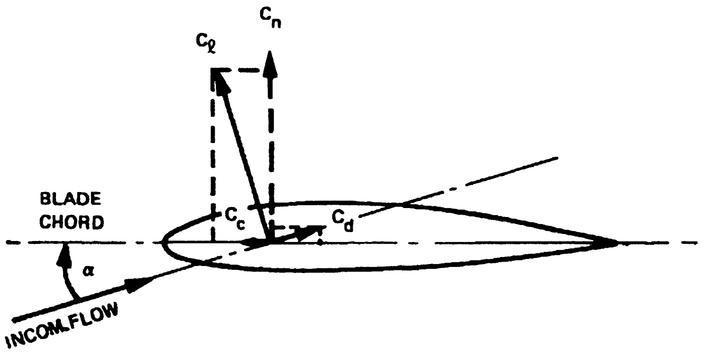

FIG. 7. Basic airfoil data.

pitching moment in reference to an axis passing through the aerodynamic center, designated by $c_{m,ac}$, remains constant with respect to c_l until stall is reached. The aerodynamic center is usually considered to be close to the $\frac{1}{4}$-chord point of the airfoil. The chordwise component of aerodynamic force does not contribute significantly to the moment about the ac, as the major contributor is the normal component, approximately equal to lift. Given this assumption, the location of the center of pressure (designated by the subscript cp) can be expressed as $x_{cp}/c = c_{m,ac}$, where c is the blade chord and x_{cp} is the chordwise distance from the point of application of the normal force to the aerodynamic center. The value of x_{cp} is positive when measured ahead of the aerodynamic center and negative when aft of the aerodynamic center. The value of $c_{m,ac}$ has a very significant effect on blade aeroelastic phenomena. When $c_{m,ac}$ is greater than zero, center-of-pressure movement resulting from increased angle of attack generates a force that resists torsional deflection of the blade. When $c_{m,ac}$ is less than zero, additional torsional loads can result. For this reason, early successful helicopters used symmetrical sections where $c_{m,ac} = 0$. Further developments in the understanding of aeroelastic phenomena have allowed designers to use cambered airfoils, which can produce higher lift coefficients in spite of $c_{m,ac}$ values of less than zero.

The primary purpose of the airfoil shape is to produce lift. The simplest way to visualize the lift generation is shown in Fig. 8 where an infinite mass of an ideal fluid flows with a velocity V around a cylinder of infinite length about which exists a circulation of strength Γ. By applying the Kutta–Jukowski law (Abbot and Von Doenhoff, 1959), an important element in the application of classic vortex theory, the relationship between the lift on the unit length of the cylinder and the strength Γ of the circulation about the cylinder can be expressed as $l = \rho U_\perp \Gamma$, where Γ is the strength of circulation

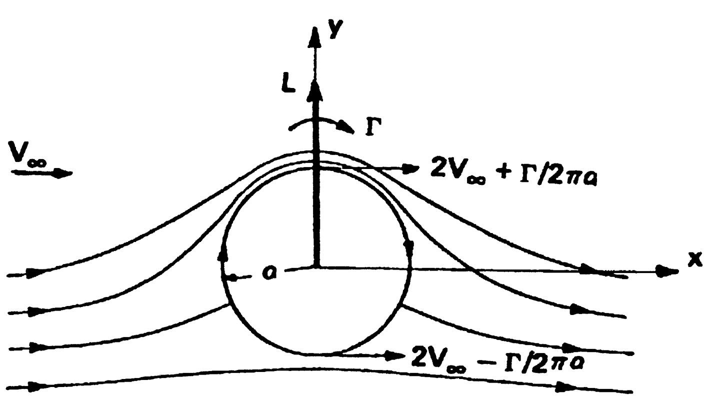

FIG. 8. Circulation around a cylinder.

around the airfoil, ρ is the air density, and $U_\perp$ is the air velocity component perpendicular to the vortex filament at that section.

The flow pattern can be described through the so-called complex potential $w = \theta + i\psi$, where θ is the velocity potential, ψ is the stream function, and i is the square root of negative 1. This equation can be represented as a function of the complex variable $z = x + iy$, $w = f(z)$. From this relationship, the velocity of the flow at a given point z can be obtained from the following differentiation: $dw/dz = u + iv$, where u is the velocity component in the x direction and v in the y direction.

The complex potential of uniform flow around a cylinder with circulation is

$$w(z) = V_{\text{freestream}}[z + (a^2/z)] + i(\Gamma/2\pi) \ln z,$$

where a is the cylinder radius.

By performing the previous differentiation, the classic flow patterns shown in the illustration can be obtained from this equation. The illustration shows that the streamlines at the top of the cylinder are closer together at the top of the cylinder than at the bottom, indicating that the velocity of local flow in the upper region is higher than the lower. From Bernoulli, the pressure on the top of the cylinder will be lower and the pressure on the bottom higher than the freestream. Thus, an upwards lift force is generated. Note that because of the symmetry of flow with respect to the y axis, no drag is present in this model. The stagnation points should also be noted, being those areas where the incoming flow has only a radial velocity component with respect to the surface of the cylinder.

From this model, much of the work of the development of the theory of finite-thickness airfoil sections began. In the early part of the century, Jukowski (Stepniewski and Keys, 1984), by using conformal transformations, developed a way of transforming a cylinder into an airfoil shape, while the flow about the cylinder with circulation was transformed into a corresponding flow about the airfoil. The conformal transformation as described by Durand (1943) consists of a method in which a geometrical field 1, characterized by an assemblage of points and lines, may be transformed into another field 2, point by point and line by line, in such a manner that the infinitely small element of area in field 1 shall transform, in field 2, into an element of similar geometrical form and proportions, while at the same time, the aggregate in field 2 may be quite different from that in field 1. While this approach produced fair approximations with practical tests, its lack of consideration of viscous effects did not permit the definition of a pressure gradient to produce desired boundary-layer conditions, which is the basis of much modern airfoil design.

4.1.2 Viscous Effects and Boundary Layer The simple visualization previously discussed considers air as an ideal incompressible, nonviscous fluid and ignores the tangential shearing forces resulting from viscosity. It should be noted that all aerodynamic forces acting on the airfoil are influenced by viscosity, the magnitude of these influences being dependent on Reynolds number. The dependence of aerodynamic forces on Reynolds number is sometimes called the "scale effect."

Of principal interest to the airfoil designer is the condition of the boundary layer. Here, viscosity is of great significance to the boundary-layer mechanics. The air flowing next to the surface has no velocity with respect to the surface, whereas at some distance from the airfoil, the velocity is that of the freestream. This layer of air is called the boundary layer. The flow of air over an airfoil is laminar near the leading edge, with the boundary layer being relatively thin. The flow then passes through a transition zone and becomes turbulent, but still attached to the airfoil. The boundary layer is much thicker in the turbulent zone.

Under certain conditions the boundary layer will separate entirely. The flow over the upper surface will experience an initial reduction in pressure with the resultant increase in velocity and then recover against an area of increasing, or adverse, pressure gradient. Against an adverse pressure gradient, both laminar and turbulent boundary layers will eventually separate. At the point of separation, the surface friction is zero. Ahead of this point, the flow in the boundary layer is downstream, and after the point of separation, it is upstream. It will be appreciated that the skin friction drag for laminar flow is significantly lower than for turbulent flow, while the turbulent boundary layer is more stable than the laminar flow because of the greater mixing in the turbulent boundary layer. Faster moving air particles from the outer edge of the boundary layer with more inertia mix within the boundary layer contributing to the total momentum of the boundary layer. This enhances the boundary layer's ability to move against an adverse pressure gradient. It must be noted that the condition of the boundary layer is significantly influenced by roughness and will strongly influence the selection of an airfoil for wind-turbine application. A pressure distribution curve for a typical nonlaminar airfoil is shown in Fig. 9.

Separation is commonly associated with stall, in which the boundary layer separates such that it influences the circulation about the airfoil, with a rapid decrease in lift with increasing angle of attack and a large increase in drag. The type of stall encountered is greatly influenced by the geometry of the leading edge and by the pressure distribution. While the airfoil designer has considerable leeway on the sharpness of stall, there is comparatively little understanding of deep stall conditions. Such conditions exist especially in wind turbines that use aerodynamic stall to regulate rotor power.

4.1.3 Definition of the Airfoil Shape

It is evident that the design of an airfoil can best be accomplished by establishing the desired boundary-layer condition with a pressure distribution that effectively defines the shape of the airfoil. While Jukowsi's approach gave fair correlation with practical tests, it was not until the 1930s that the path to modern airfoil design was started. Theodorsen (1932) developed a theory of wing sections of arbitrary shape. Jacobs (1939) subsequently inverted this method in order to determine an airfoil shape that would produce a desired pressure distribution. Thus the way was open to the modern methods of defining a boundary layer that would produce the desired airfoil characteristics and the shape of the airfoil itself.

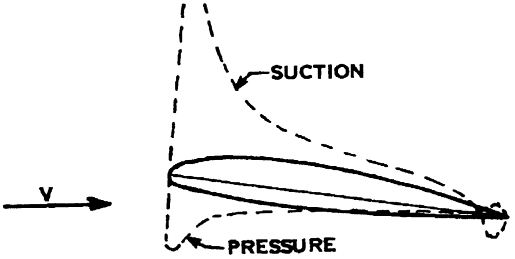

FIG. 9. Pressure-distribution curve for a typical nonlaminar airfoil.

One major environmental influence is the accumulation of dirt and insects on the leading edge of the airfoil. Sufficient debris can accumulate to degrade seriously the performance of the airfoil. Significant decreases in maximum lift coefficient and large increases in drag can result from dirt accumulation, and these factors must be taken into account in the airfoil specifications. Early wind turbines suffered catastrophic power losses well in excess of 50% because of environmental degradation of the blades.

4.2 Determination of Airfoil Characteristics

The airfoil design objectives and specifications will be selected to produce the most effective rotor system for a given wind resource, environmental conditions, and turbine generator capacity. The airfoil specification will be tailored to a particular type of rotor system and the location of the airfoil section on the blade. For instance, the requirements for the root-section airfoil of a horizontal-axis wind turbine are significantly different from the tip region because of the great difference in speed and angle of attack. A stall-regulated machine will have an airfoil where the boundary layer is designed to separate and the airfoil stall at a given windspeed, thus preventing the rotor from overpowering the turbine. In this case, the stall regulation is achieved by having the airflow separate at a designed angle of attack and lift coefficient. The maximum lift coefficient may also be suppressed to avoid overpowering the turbine in high winds, thus allowing a larger rotor and greater energy capture in lower winds. The airfoil designer will also look at the constraints on the airfoil. For instance, a designer may specify a minimum airfoil thickness for structural reasons. Another example is the requirement for vertical-axis wind turbines to produce lift on both the upwind and downwind passes of the blade, which further restricts the airfoil design options. The requirements may also specify a limit on the amount of laminar flow on the upper section, especially where manufacturing techniques might restrict the achievement of reduced drag by maximizing laminar flow. The airfoil design objectives can also specify defined stall characteristics, susceptibility to roughness, and numerous other less significant characteristics.

Once the wind turbine designer has specified the airfoil requirements, a modern design and analysis program such as the Eppler Airfoil Design and Analysis Program can be used (Eppler and Somers, 1980). This program consists of a conformal-mapping method for prescribing velocity-distribution characteristics and producing the physical airfoil shape required to achieve them, a panel method for analysis of existing airfoil shapes, and a boundary-layer method. From the design objectives, a drag polar can be produced from which the program can specify the necessary velocity distribution. Given the capabilities of such programs, the designer can directly determine the velocity distributions that will produce the desired boundary-layer conditions.

Thus, with correct definition of the airfoil requirements and skilled use of tools such as the Eppler program, the profile of the wind-turbine blade can be accurately defined. These results can then be used with a variety of methods to determine the actual rotor performance. Many of these methods are adapted from propeller and helicopter rotor design, such as momentum theory, vortex theory, and blade element theory. These methods give reasonably accurate predictions of wind-turbine performance. The main area of development still remains the achievement of a better understanding of the airfoil requirements of the wind turbine in service conditions.

5. ELECTRICAL-POWER GENERATION

The airfoils and mechanical drive train of the wind turbine harness the energy of the wind and make it available for use as mechanical power. This power can be used, as in earlier times, for turning mill stones or pumping water. However, the use of wind energy for these purposes is most efficient when the point of collection and the point of use are coincident. Converting the mechanical motion of the wind to electrical energy provides the means to collect the energy at the point of the greatest wind resource, transport it to points of use, and convert it back to mechanical or other forms for use in other activities. This energy conversion and transportation requires an electrical generator, possibly a power electronic converter, and an electrical network or "balance of plant."

5.1 Generators

Many generation technologies have been proposed and studied for this application, but the most commonly used generators are the induction and synchronous generators. Of these, the induction generator is by far the most widely used generator for grid-connected applications. Induction generators require reactive power for excitation and are generally not well suited for stand-alone application because of this need for reactive power.

Constant-speed wind turbines have traditionally used induction generators directly connected to the synchronous utility network. The reactive demand of the induction generators is then supplied by the network but can be compensated at the generators with a series of switched capacitor banks so that the power factor of the turbines will be 1.0 at full power output. The reactive demand of the substation and collection system is then corrected at the substation by additional banks of switched capacitors. Using this method, the wind-power plant can generate power at near-unity power factor at any power level. During periods of light winds, when the wind turbines are not operating, this capacitance is available to be switched to aid in voltage stability by the utility.

The mechanical power delivered by the turbine to the generator shaft is given by the product of the rotor shaft speed and the torque on the mechanical shaft:

$$P_m = \omega_m T_m. \tag{11}$$

The mechanical power is converted to electrical power in an induction generator according to the relationship

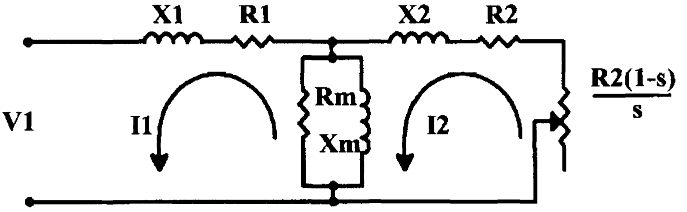

FIG. 10. Induction generator schematic.

$$P_m = 3I_2^2R_2(1 - s)/s, \tag{12}$$

where I_2 is the rotor current, R_2 is the rotor resistance, and s is the slip speed given by $s = (\omega_0 - \omega_m)/\omega_0$. ω_0 is the electrical synchronous speed, $3600/(p/2)$ rpm where p is the number of magnetic poles of the generator, always an even number, usually 2 or 4. Figure 10 is a schematic diagram of one phase of a three-phase induction generator connected to a utility, represented by V_1. Knowing the generator parameters and using Kirchhoff's laws to solve for the stator phase current I_1 will give the real and reactive power delivered to the utility as $S = V_1I_1^*$.

A constant-speed wind turbine using directly connected induction generators will generate no appreciable harmonics and have no synchronization problems. With power-factor correction capacitors connected at the wind turbines and at the wind-power plant substation, the power factor can be controlled to be near unity at any power level. A wind-power plant using constant-speed wind turbines with induction generators can be made to generate power with no power-quality issues at any power level.

Synchronous generators (Fig. 11) are capable of supplying both real and reactive power to the network but require synchronization to the network before direct connection, are more expensive than induction generators, and are less tolerant of transient overtorques that may occur from wind gusts. Small- to medium-sized wind turbines will commonly use induction generators and larger wind turbines will sometimes use stall-regulated air foils and synchronous generators.

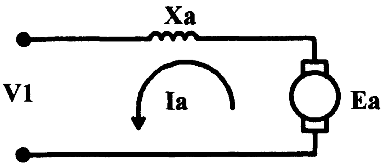

FIG. 11. Synchronous generator schematic.

The mechanical power is converted to electrical power in a synchronous generator according to the relationship

$$P_m = 3V_1|E_a| \sin\delta/X_a, \tag{13}$$

where V_1 is the terminal voltage, X_a is the synchronous reactance, E_a is the excitation voltage, and δ is the excitation phase angle.

A constant-speed wind turbine using directly connected synchronous generators will generate no appreciable harmonics and be able to control reactive power and the terminal voltage by varying the excitation phase angle δ. A wind-power plant using constant-speed wind turbines with synchronous generators can be made to generate power with no power-quality issues at any power level.

5.2 Power Electronic Converters

Variable-speed wind turbines use an electronic power converter to convert the variable frequency of the wind turbine to the synchronous frequency of the utility network. Power electronic converters typically used in wind-turbine application include both line-commutated and force-commutated topologies. Line-commutated converters using thyristor technology are less expensive but require reactive power and are used with synchronous generators only. These line-commutated converters are highly efficient, but because of their switching control, which allows for control of device turnon but requires that the line current go through a zero crossing for device turnoff, generate a high level of current harmonics.

State-of-the-art force-commutated converters using solid-state devices such as isolated-gate bipolar-transistor (IGBT) power-electronic devices can operate to control both the real and reactive power of the wind turbine. The converter can operate to control the amplitude of the line-current wave form by determining the power generation from the wind turbine, and the frequency and phase angle of the line-current wave form by monitoring the utility voltage. This method ensures that the wind turbine is always synchronized with the utility voltage regardless of the utility frequency, either 50 or 60 Hz. This method also controls the power factor of the wind turbine by controlling the current displacement angle.

Power electronic converters will typically generate some level of harmonics due to the switching, with higher switching frequencies generating cleaner wave forms with lower harmonics. The high switching frequencies that are possible using IGBT devices produce a current wave form of very good fidelity and low harmonic components. The step-up transformers and transmission lines that connect the wind turbines to the utility further reduce the already low harmonics so that at the point of common coupling, there is very little harmonic distortion due to the wind-turbine converters.

Recent release of recommended practices such as IEEE-519 have specified current-harmonic limits at the point of common coupling with the utility for power converters or other electrical equipment. This recommended practice calls for a maximum total harmonic distortion of the current and a maximum distortion due to any single harmonic at the point of common coupling to the high-voltage utility system. A wind-power plant using variable-speed wind turbines with power electronic converters can be made to generate power within these harmonic-power quality limits at any power level.

The capability to control reactive power can provide transmission efficiencies and voltage stability. This capability is available during high winds when generating real power and also during light winds when no real power is being generated. This feature is particularly valuable on weak grids to enhance voltage stability.

6. ECONOMICS

The method described below is the one recommended by the National Renewable Energy Laboratory (NREL) of Boulder, Colorado, for most applications (Cohen *et al.*, 1989.) Other methods, including yearly cash-flow analyses or the utility revenue requirements method (EPRI, 1986), may be more appropriate, depending on the developer and customer.

The primary economic figure of merit for wind turbines is the cost of energy (COE). This number should be provided in constant dollars and calculated for a standard size installation, a 50-MW (rated) wind farm with an expected life of 30 years. The COE depends on several economic factors to bring the long-term costs and revenues in relation to the initial capital cost. The NREL approach reduces the year-by-year analysis into a simple formula by introducing a fixed charge rate (FCR), which incorporates the effects of the present value discount rate and interest rates into a single number. Using this formulation, the COE is calculated by the equation

$$\mathrm{COE} = \frac{\mathrm{FCR} \times \mathrm{ICC}}{\mathrm{AEP}_{\mathrm{net}}} + \mathrm{LRC} + \mathrm{O\&M}, \qquad (14)$$

where COE = levelized cost of energy ($/kW·h), FCR = fixed charge rate (1/yr) = 0.109, ICC = initial capital cost ($), O&M = annual operating and maintenance cost ($/kW·h), LRC = levelized replacement cost ($/kW·h), typically 0.2 cents/kW·h, $\mathrm{AEP}_{\mathrm{net}}$ = net annual energy production (kW·h/yr). The $\mathrm{AEP}_{\mathrm{gross}}$ of a single turbine can be expanded into the output of a larger wind facility using the following formula:

$$\mathrm{AEP}_{\mathrm{net}} = \mathrm{AEP}_{\mathrm{gross}} \times (1 - \mathrm{EL}) \times \text{availability}, \qquad (15)$$

where EL is the product of individual energy losses calculated as below, and availability is the percentage of calendar time that the turbines are functional and ready to deliver power to the grid. EL is calculated as follows:

$$\begin{aligned}\mathrm{EL} = 1 &- (1 - L_{\mathrm{array}}) \times (1 - L_{\mathrm{soiling}}) \\ &\times (1 - L_{\mathrm{collect}}) \times (1 - L_{\mathrm{control}}) \\ &\times (1 - L_{\mathrm{icing}}),\end{aligned} \qquad (16)$$

where L_{array} are array losses (2%–8%), L_{soiling} are blade soiling losses (1%–3%), L_{collect} are collection system losses from turbine to grid

(1%–3%), L_{control} are control, grid, and miscellaneous losses (2%–4%), and L_{icing} are blade icing losses (2%–8%). The typical ranges of losses are indicated in the parentheses.

COE values typically fall in the range of 4–7 cents/kW·h. Wind energy requires a good wind resource to be economic and requires a utility customer or other readily available load to be practical. An average annual average wind speed of about 6 m/s is the minimum needed to obtain a COE of below 6–7 cents/kW·h. Some countries, particularly in Europe and North America, require utilities to buy power at the avoided cost of the utility, which is the dollar amount of fuel saved by operating the wind turbines. Many local situations may lead to power purchase contracts that are more favorable than the avoided cost. If a wind-power plant can displace the need to install a conventional power plant, capacity credit may be available, which can result in a large increase over the avoided cost of energy. Government incentives are also frequently available to developers and owners of renewable-energy generation, which can make otherwise marginal economic installations very attractive.

GLOSSARY

Apparent Power: The complex phasor sum of real and reactive power.

Balance of Plant: The electrical infrastructure to collect and transport the energy from the wind turbines to the point of use.

Betz Limit: Theoretical limit of wind turbine energy capture = 59% of wind-power density.

Campbell Diagram: A modal plot for VAWTs used to identify potential resonant conditions.

Pitch Regulation: A method of wind-turbine power regulation in which the blades are pivoted to larger angles of attack in high winds, making the blades operate less efficiently aerodynamically.

Power Factor: The cosine of the phase angle between the voltage and current wave forms of the same phase.

Reactive Power: Electrical energy stored in inductance or capacitance during one part of an alternating-current cycle and released during another part of the cycle without doing work.

Real Power: Electrical energy generated or consumed from doing work.

Square-Cubed Law: Law stating that the loads of a wind turbine increase as the cube of the diameter while the power increases as the square, thus implying a practical limit for the size of a cost-effective wind turbine.

Stall Regulation: A passive approach to wind-turbine power regulation in which the blades are designed to go into aerodynamic stall above the rated speed and no blade-pitching mechanisms are required.

Turbulence Intensity: The standard deviation of the wind speed divided by the average wind speed for the period.

Wind Power Density: Fluid power in the wind per unit of cross-sectional or swept area, measured in units of watts per square meter.

Wind Shear: The change in wind speed with increasing elevation.

Works Cited

Abbot, I., Von Doenhoff, A. (1959), *Theory of Winged Sections,* New York: Dover.

Ashwell, T. (1992), "Measured Data for the Sandia 34-Meter Vertical Axis Wind Turbine," Sandia National Laboratories Report No. SAND91-2228.

Barnard, J., Wendell, L. (1995), *A Single Method of Estimating Wind Turbine Blade Fatigue at Potential Wind Turbine Sites,* PNL-10630.

Barnes, J., Davenport, A. (1975), in: *Fourth Conference on Probability and Statistics in the Atmospheric Sciences,* Tallahassee: American Meteorological Society, pp. 1–7.

Cliff, W. (1977), *The Effect of Generalized Wind Characteristics on Annual Power Estimates from Wind Turbine Generators,* DOE Report No. PNL-2436.

Cohen, J., Schweizer, T., Hock, S., Cadogen, J. (1989), *Windpower '89 Proceedings,* San Francisco: NREL/TP-257-3628.

Corotis, R. (1977), *Stochastic Modeling of Site Wind Characteristics,* ERDA Report No. EY-76-S-06-2342.

Durand, W. (1943), *Aerodynamic Theory,* California Institute of Technology: Durand Reprinting Committee.

Eggleston, D., Stoddard, F. (1987), *Wind Turbine Engineering Design,* New York: Van Nostrand Reinhold.

Electric Power Research Institute (1986), *Technical Assessment Guide,* Report No. P-4463-SR, Palo Alto, CA: EPRI.

Electric Power Research Institute (1994), *Planning Your First Wind Power Project,* Report No. TR-104398, Palo Alto: EPRI.

Elliott, D. (1937), *Wind Energy Resource Atlas of the United States,* U.S. Department of Energy Report No. DOE/CH100094-4, Washington, DC: U.S. GPO.

Eppler, R., Somers, D. (1980), *A Computer Program for the Design and Analysis of Low Speed Airfoils,* NASATM-80210.

Holley, W., Aitkenhead, W., McNerney, G., Rogers, E. (1987), *Sixth ASME Wind Energy Symposium,* Dallas: SED-Vol. 3, 43–50.

International Electrotechnical Commission (1994), *IEC Standard 1400-1,* Geneva: IEC.

Jacobs, E. (1939), *Preliminary Report on Laminar-Flow Airfoils and New Methods Adopted for Airfoil and Boundary-Layer Investigations,* National Advisory Committee on Aeronautics Report No. WR L-345, Washington, DC: U.S. GPO.

Justis, C., Hargraves, W., and Yalcin, A. (1976), "Nationwide Assessment of Potential Output from Wind Power Generators," *J. Appl. Meteorol.* **15,** 673–678.

Rohatgi, J., Nelson, V. (1994), *Wind Characteristics: An Analysis for the Generation of Wind Power,* Canyon, TX: Alternative Energy Institute, West Texas A&M University.

Smith, D., Ilyin, M. (1991), *Tenth ASME Wind Energy Symposium,* Houston: SED-vol. 11, 29–32.

Stepniewski, W., Keys, C. (1984), *Rotary Wing Aerodynamics,* New York: Dover.

Theodorsen, T. (1932), *Theory of Wing Sections of Arbitrary Shape:* NACA Rep 411.

Wendell, L., Barnard, J., Morris, V. (1993), in: *Proceedings of the Windpower '93 Conference,* San Francisco: American Wind Energy Association.

Wentink, T. (1976), *Study of Alaskan Wind Power Potential and Its Possible Applications,* Geophysical Institute, University of Alaska, Report No. NSF/RANN/SE/AER 74-00239/FR-76/1.

Wilson, R., Lissaman, P., Walker, S. (1976), "Aerodynamic Performance of Wind Turbines," Oregon State University, Corvallis, Report No. AER74-04-014A03.

X-RAY AND NEUTRON REFLECTIVITY

Moshe Deutsch, *Physics Department, Bar-Ilan University, Ramat-Gan, Israel*

Benjamin M. Ocko, *Physics Department, Brookhaven National Laboratory, Upton, New York, U.S.A.*

Introduction 479
1. **Theory of Reflectivity** 480
1.1 General Features 480
1.2 Dynamical Theory of Reflectivity 481
1.3 Kinematical Theory of Reflectivity 482
1.4 Unique Features of Neutron Reflectivity 482
1.5 Examples 483
1.5.1 Rough Surfaces 483
1.5.2 A Thin Film on a Substrate 483
1.5.3 Crystal Surface 483
2. **Instrumentation** 484
3. **Applications** 485
3.1 Simple Liquids: Water 485
3.2 Thin Films: Water on Glass 486
3.3 Polymer Films: Combined X-Ray and Neutron Reflectivity 486
3.4 Polarized-Neutron Reflectivity: Ultrathin Fe Films 487
3.5 Crystalline Surfaces: Clean Au(001) in Vacuum 487
4. **Related Techniques** 488
Works Cited 489
Further Reading 490

INTRODUCTION

The intensity of x rays and neutrons reflected from an interface at near-grazing incidence depends sensitively on the refractive-index profile normal to the interface. As the refractive index depends on the density, the measured reflectivity vs incidence angle can be used to determine the interfacial density profile with sub-ångstrom resolution. The density sampled is, however, different for the two probes. As neutrons interact with the atomic nuclei, they probe the scattering amplitude density profile of the nuclei. This amounts effectively to probing the average mass density profile, where the contribution of each atomic species is weighted by its nucleus–neutron interaction strength, which varies irregularly throughout the periodic table. X rays, however, interact with the electronic charge and thus probe the total electronic density profile of the interface. The two probes provide, therefore, slightly different, though related, information on the interface structure. Since modern synchrotron x-ray sources are 5–6 orders of magnitude brighter than the best neutron sources, x-ray reflectivity measurements routinely achieve much higher resolutions and accuracies than their neutron counterparts. However, the magnetic interaction and the large variation of the neutron scattering length upon isotopic substitution renders neutron reflectivity a method of choice in special cases such as studies of surface magnetism and the adsorption of polymers at liquid and solid surfaces.

The use of x-ray reflectivity to measure the index of refraction of matter was originally suggested by Compton (1923). Parratt (1954) revived this method and used it in his seminal study of glass-deposited thin metal films. Marra *et al.* (1979) first applied the very small penetration depth for low angles of incidence to study the surface structure of crystals, opening up a new and very active field of research, which has provided a wealth of new knowledge about surface reconstruction and quasi-2D phase transitions. The x-ray reflectivity studies of the free surfaces of liquid metals (Lu and Rice, 1978) and of liquid crystals (Als-Nielsen *et al.*, 1982; Pershan and Als-Nielsen, 1984) laid the foundations for the fast-growing field of surface studies of liquids and soft condensed matter. These methods provided true *microscopic* information on important phenomena

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

like surface freezing and melting and the structure of the various phases of monomolecular Langmuir films. These techniques were later extended to solid–solid crystalline, liquid–solid, and even liquid–liquid interfaces.

The first detailed theoretical treatment of x-ray reflectivity was given by Picht (1929). Abeles (1950) and Parratt (1954) provided the first dynamical treatments, based on ideas drawn from classical optics of multilayers. The distorted-wave Born-approximation theory of Vineyard (1982) was an important step beyond the kinematical theory (Born approximation). Since then a large number of theoretical studies of both specular reflection and the accompanying diffuse scattering have been published, taking into account an increasing number of real effects like surface roughness, capillary surface waves in liquids, static and dynamic disorder in solid surfaces, and the stepped nature of crystalline surfaces.

Neutron reflectivity was first employed by Fermi and Zinn (1946) to measure the scattering length and phase shift of the neutron upon reflection. Hughes and Burgy (1951) pioneered the use of neutron reflectivity to study surface magnetism, a method further developed by Felcher (1981) and co-workers. More recently, Penfold and Thomas (1990) studied polymer adsorption at, and surfactant monolayers on, the free surface of mixtures of D_2O and H_2O by ingeniously using the isotope variation of the scattering length to eliminate the signal originating from the water. Surfaces of solid polymers and interfaces between polymeric bulks were also studied extensively by Russell (1990) and others.

The fast and wide dissemination of x-ray reflectivity and related surface methods since the 1980s is mainly due to two important factors. The first is the advent of powerful synchrotron radiation sources, dedicated to the production of x rays. They provide intense, highly collimated x-ray beams, which are prerequisites for these grazing-angle surface studies. The second is the explosive growth in computer power since the 1970s, which put powerful computer fitting, modeling, and simulation methods within easy reach of small groups and individual researchers. A similar growth rate was enjoyed by neutron reflectivity. This growth was largely due to the construction of high-power research reactors and spallation sources in the United States, United Kingdom, and France. With the development of brighter thermal-neutron sources, a faster increase is expected in the use of neutron reflectivity for surface studies.

1. THEORY OF REFLECTIVITY

1.1 General Features

Standard electromagnetic theory (Born and Wolf, 1975) predicts the splitting of a plane wave incident on an interface between two media of refractive indices $n_1 \neq n_2$ into reflected and refracted waves, as shown in Fig. 1(a). While $\alpha_i = \alpha_r$ regardless of n_1 and

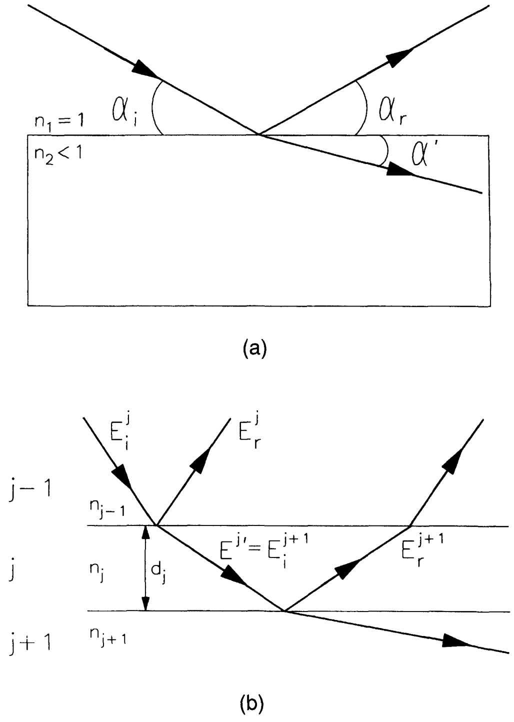

(a)

(b)

FIG. 1. (a) Reflection and refraction diagram at a free surface. Note that for x rays the refractive index of matter is less than unity. Hence, the refraction angle α' is smaller than the incidence or reflection ones, α_i and α_r. (b) The various waves reflected and refracted at the lower and upper surfaces of one layer, j, in a multilayer structure. n_j and d_j are the layer's refractive index and thickness, respectively.

n_2, the angle of refraction α' is determined by Snell's law:

$$n_1 \cos\alpha_i = n_2 \cos\alpha'. \tag{1}$$

For both x rays and neutrons, the complex refractive index n is given by

$$n = 1 - \delta + i\beta. \tag{2}$$

For x rays,

$$\delta = r_0\rho_e\lambda^2/2\pi, \quad \beta = \mu\lambda/4\pi, \tag{3}$$

where r_0 is the classical electron radius, λ is the wavelength, ρ_e is the electron density, and μ is the mass absorption coefficient. For neutrons, $r_0\rho_e$ is replaced by the neutron scattering-length density, $\Sigma_i N_i b_i$, where b_i is the empirically determined scattering length of nucleus i and N_i is its number density. The weak interactions render δ in Eq. (2) small, of order 10^{-5}–10^{-6} for most materials. The absorption term $\beta \ll \delta$ in most cases and, for simplicity, will be neglected from now on.

Free surfaces of solids and liquids are by far the most studied interfaces. Here $n_1 = 1$, and since for x rays $\delta > 0$, $n = n_2 < 1$. Snell's law yields $\alpha' < \alpha_i$, and thus, upon decreasing α_i, a finite angle $\alpha_c = \sqrt{2\delta}$ is reached where $\alpha' = 0$. For an angle of incidence smaller than this critical angle, total external reflection of the incident wave occurs; i.e., the reflectance, defined as the complex quantity $r = E_r/E_i$, has a modulus that is unity for $\alpha_i < \alpha_c$, and so is, of course, the reflectivity $R = |r|^2$.

For angles $\alpha_i > \alpha_c$, $|r|$ and R are no longer unity and for an infinitely sharp and flat surface are given by Fresnel's law:

$$r = \frac{q - q'}{q + q'} \quad \text{and} \quad R_F(q) = \left|\frac{q - q'}{q + q'}\right|^2, \tag{4}$$

where $q = (4\pi/\lambda)\sin\alpha_i$ is the momentum transfer perpendicular to the surface, $q' = \sqrt{q^2 - q_c^2}$, and the absorption is assumed to be negligible. The reflectivity R is a fast decreasing function, and for $q \gg q_c$ is very closely approximated by $R_F(q) \approx (q_c/2q)^4$. In the field of small-angle bulk scattering, this is long known as Porod's law. In practice, critical angles are a few tenths of a degree, and reflectivities at $\sim 20q_c$, just a few degrees, are of order $\sim 10^{-7}$. Thus, the well-collimated and very intense x-ray beams available at synchrotron sources are often the tools of choice for reflectivity measurements. Note that because of the low angles of incidence the reflectivity is practically polarization independent.

Real interfaces are neither abrupt nor ideally flat. On an atomic scale the transition between adjacent media is of a finite width. For example, thermally induced capillary waves in liquids, and structural defects or crystalline surface steps in solids, modify the density profile. The reflectivity off a real surface will therefore deviate from Fresnel's law. Two main approaches exist for calculating the reflectivity in these cases. The first is the dynamical approach originally proposed by Parratt (1954). The second is the kinematical approach, based on the first Born approximation. While the first of these is more rigorous, the second is much easier to use in practice and yields in most cases accurate and reliable density profiles. We now discuss each of these methods separately.

1.2 Dynamical Theory of Reflectivity

The density variation perpendicular to the interface can be approximated by a piecewise-constant function, thus replacing the actual interface by a series of N slabs having each a constant refractive index $n_j = 1 - \delta_j$ and a width d_j. By increasing the number and decreasing the width of the slabs, the approximation can be made arbitrarily close to the actual density profile. The first slab, $j = 1$, denotes the medium above the free surface (e.g., the vacuum), and the last, Nth, slab is the bulk where $d_N = \infty$. Consider now the jth slab, depicted in Fig. 1(b). The continuity requirements of the field at each interface yield for the reflectance r_j of slab j (Born and Wolf, 1975)

$$r_j = a_{j-1}^4(r_{j+1} + F_j)/(r_{j+1}F_j + 1), \tag{5}$$

where $a_j = \exp(-iq_jd_j/4)$, $F_j = (q'_{j-1} - q'_j)/(q'_{j-1} + q'_j)$, and $q'_j = \sqrt{q^2 - q_{c,j}^2}$, with $q_{c,j}$ denoting the critical angle for slab j. Equation (5) is solved iteratively, by noting that since $d_N = \infty$, the reflectance off the bottom of the Nth slab is zero, and the sole contribution of that slab is from the upper interface. This yields $r_N = a^4_{N-1}F_N$, which is then used to calculate r_{N-1} and so on until the free-surface reflectance r_2 is obtained, whence $R(q) = |r_2|^2$. This method properly accounts for the effects of refraction and is therefore exact. Absorption can also be conveniently included by using a complex momentum transfer $q'_j = \sqrt{q^2 - q^2_{c,j} + i\mu_j/2}$. An additional advantage of the dynamical theory is that it gives a natural description of the many important cases involving layered interfaces such as ordered diblock polymers, thin films on a solid substrate, or monomolecular Langmuir films on water. For more than a few layers, however, its use becomes very tedious analytically, although the iterative algorithm is very easy to program for computer calculations. By contrast, the approximate kinematical theory, given below, has the advantage of yielding a closed-form expression for the reflectivity that is sufficiently accurate for most cases.

1.3 Kinematical Theory of Reflectivity

Consider again the multislab approximation discussed above. This time, however, we assume small reflectivities and neglect refraction effects so that each slab reflects the incident wave independently. The amplitude ratio of the reflected to incident wave for a single slab j at depth z below the nominal surface is given by (Als-Nielsen and Kjær, 1989; Schlossman and Pershan, 1990)

$$E^j_r/E_i = 4\pi i r_0 \langle \rho_e(z) \rangle d_j/q, \tag{6}$$

where $\langle \cdots \rangle$ denotes averaging in the x-y plane, parallel to the surface. Making the slabs of equal infinitesimal thickness dz, and summing the contributions over z, multiplying each by the appropriate phase factor $\exp(iqz)$, one obtains the so-called "master equation":

$$R(q) = R_F(q) \left| \frac{1}{\rho_\infty} \int_{-\infty}^{\infty} \frac{d\langle \rho_e(z) \rangle}{dz} \exp(iqz) dz \right|^2, \tag{7}$$

where ρ_∞ is the bulk electron density. Thus, the Fresnel reflectivity is modified by the absolute value of the Fourier transform of the density gradient, which is, within a constant, the standard crystallographic structure factor of the interface. Strictly speaking, the master equation applies in the weak-reflectivity limit, i.e., where $R \ll 1$. For many cases, however, it is sufficiently accurate at all q. It has the advantage of being easily applicable to analytic models of the density profile and is therefore widely used. It should be noted, however, that since Eq. (7) employs the square of the structure factor, the phase information is lost just as in standard crystallography, and a unique inversion of the measured $R(q)$ to obtain $\rho_e(z)$ is not possible in the general case. However, through judicious use of additional information and general physical constraints, it is possible in almost all cases to obtain a well-defined and accurate description of the density profile.

1.4 Unique Features of Neutron Reflectivity

Neutrons have several unique properties that can be used to advantage in particular cases. For some materials (e.g., silicon) the scattering and absorption cross sections are extremely small; thus neutrons can travel through these materials with little attenuation. This feature can be used to advantage when studying thin organic films in contact with liquid interfaces. Also, for x rays the scattering-length density always has the same sign, whereas for neutrons δ can assume both negative and positive values for different atoms and isotopes. In particular, H and D have b values of opposite signs. Hence an appropriate mixture of D_2O and H_2O can be prepared (molar ratio of 0.088:1) such that its average scattering length is equal to that of air. The free surface of such a mixture will not reflect or refract an incident neutron beam, and the general concept is referred to as contrast matching. The cancellation of the bulk signal greatly improves the signal/noise ratio in experiments involving monolayers on water, surface adsorption of polymers in solution, and solute concentration near a free surface of water. Furthermore, by deuterating a specific part of an organic molecule or biomolecule, one can study by neutron reflection the behavior of that part only near the surface. Although anomalous dispersion of x rays near an absorption edge was also employed for similar measurements, neutrons are superior to x rays in this respect.

Another advantage of neutrons is their

magnetic interaction with matter. For magnetic materials the total scattering length is $b \pm C\mu$, where μ is the magnetic moment and C is a constant (Felcher, 1981). The sign depends on whether the neutron polarization is parallel (+) or antiparallel (−) to the magnetization. Thus, polarized-neutron reflectivity permits studies of magnetic structures near the surface. The measurement of the so-called flipping ratio, the ratio of the parallel to antiparallel reflectivities, is particularly advantageous since the interference from many nonmagnetic variables cancels out.

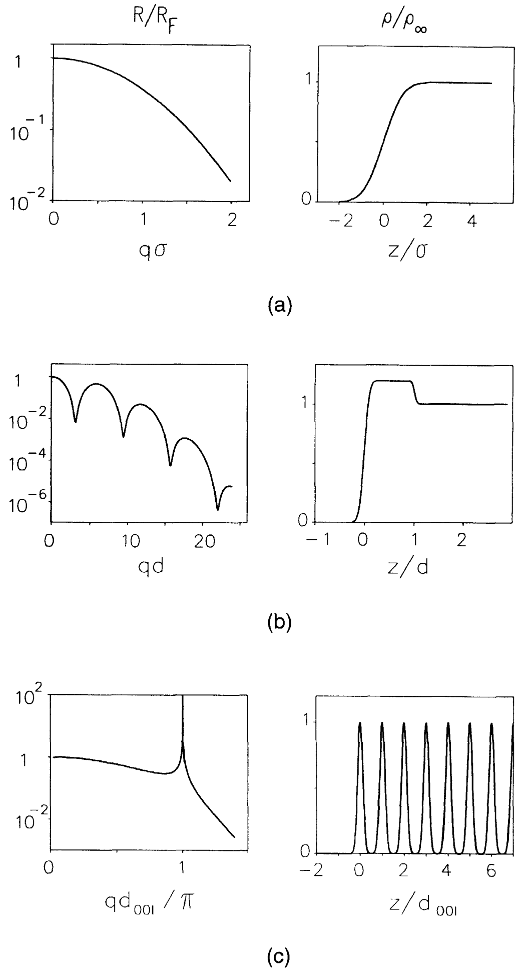

FIG. 2. Prototypical reflectivity curves R divided by the Fresnel-law R_F (left) and their corresponding surface-density profiles (right). The curves are discussed in Sec. 1.5.

1.5 Examples

Figure 2 shows a few typical examples of surface profiles and the resulting reflectivity curves. We now discuss each separately.

1.5.1 Rough Surfaces An important case, shown in Fig. 2(a), is the "rough" surface, where the density profile perpendicular to the surface has a finite width. This may be the free surface of a liquid, broadened by thermally excited capillary waves; an amorphous solid surface with a random height distribution; or a crystal surface with extra unit cells missing or piled up at random positions on the crystallographic plane constituting the nominal "surface." Averaged over the x-y plane, the perpendicular density profile of such a diffuse interface is well represented in many important cases by an error function. The master equation yields in this case

$$R(q)/R_F(q) = \exp(-q^2\sigma^2), \tag{8}$$

where σ is the width parameter of the height distribution. This Debye–Waller type term is used to represent diffuseness or roughness in profiles of more complicated structure. It is easily recognized by plotting R/R_F vs q^2 on a semilog scale, yielding a straight line, where the slope is $-\sigma^2$. An example of this is shown in the inset in Fig. 3.

1.5.2 A Thin Film on a Substrate The reflectivity from a thin film of thickness d and a uniform electron density ρ_e on a substrate with $\rho_e' \neq \rho_e$ and an equal roughness σ at the upper and lower interfaces is shown in Fig. 2(b). The master equation yields

$$R(q)/R_F(q) = [a^2 + b^2 + 2ab\cos(qd)] \times \exp(-q^2\sigma^2), \tag{9}$$

where $b = \rho_e/\rho_e'$ and $a = 1 - b$. The typical oscillatory reflectivity curve, often called Kiessig fringes, is the signature of such a film. The film thickness d can be determined immediately from the curve by $d \approx 2\pi/(q_i - q_{i+1})$, where q_i, q_{i+1} are the positions of two consecutive extrema.

1.5.3 Crystal Surface Assume a cubic crystal, having an abrupt surface parallel to a set of lattice planes $00l$ of spacing d_{00l}. Summation of the contributions from all

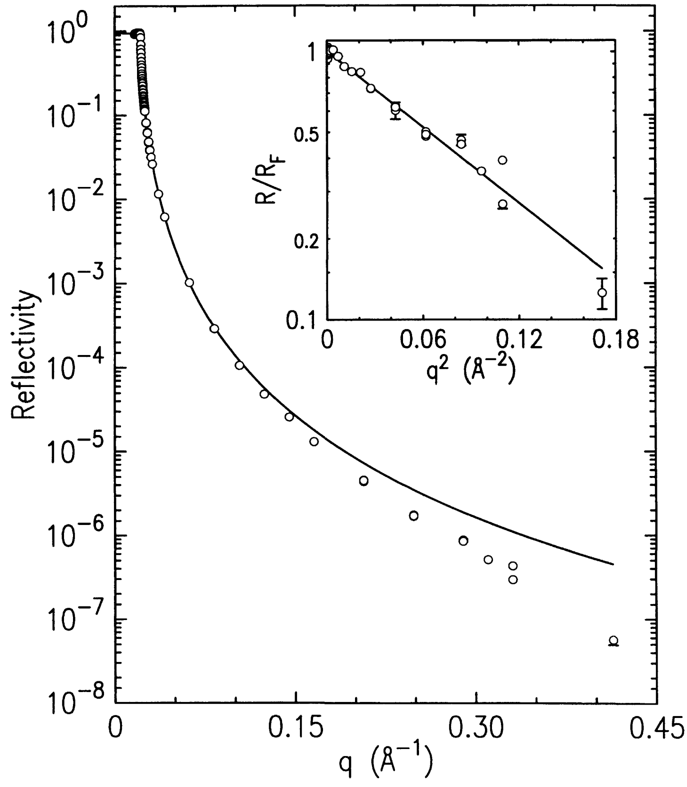

FIG. 3. X-ray reflectivity from the water–vapor interface (solid dots). The deviation between the data and the Fresnel law (solid line) is due to capillary-wave–dominated surface roughness. The inset shows that the Gaussian roughness model in Eq. (8) fits the measured reflectivity well with a roughness width parameter σ = 3.3 ± 0.1 Å. Adapted from Braslau *et al.* (1988). For further details see Sec. 3.1.

planes in the kinematical approximation yields

$$R(q) \sim [q \sin(qd_{00l}/2)]^{-2} \exp(-q^2\sigma_b^2), \qquad (10)$$

where σ_b is the bulk Debye–Waller factor. For small q, $\sin(qd_{00l}/2) \sim qd_{00l}/2$ and Eq. (10) assumes the right q^{-4} form of Fresnel's law. For $q = q_{00l} = 2\pi l/d_{00l}$, however, Bragg peaks are obtained, as shown in Fig. 2(c). The "wings" of these peaks fall off as $\sim|q - q_{00l}|^{-2}$. This shape is, of course, modified in real crystals by the molecular form factor, surface roughness, faceting, reconstruction, and adsorption of foreign molecules at the surface.

2. INSTRUMENTATION

The spatial resolution in reflectivity measurements is inversely proportional to the angular range measured. High resolution entails, therefore, measuring out to large angles of incidence and consequently very low reflectivities. The use of high-intensity x-ray beams available at synchrotron sources is thus very beneficial for many of these measurements. Present-day second-generation synchrotron sources and the associated beam-line optics are capable of producing monochromatic fluxes of the order of 10^{11} photons per second in a $\leq$1-mm^2 spot. This represents a flux increase of $\sim10^4$ over conventional rotating-anode laboratory-based sources.

For solid samples, a conventional two-circle diffractometer can be used to measure the reflectivity. The sample is oriented in this case perpendicular to the rotation plane of the circles (usually the horizontal) and the reflectivity is measured in a θ-2θ scan as in a single-crystal Bragg-reflection measurement. Because of the small angle of incidence, the illuminated area ("footprint") on the sample is magnified greatly as compared to the beam width. For sufficiently large samples a narrow slit is positioned close to the sample to keep the footprint smaller than the size of the sample at all angles of incidence. For samples that cannot be made larger than the footprint, an angle-dependent illuminated-area correction (e.g., $\sin\theta$) is required.

Reflectivity measurements are often accompanied by complementary surface-scattering measurements including grazing-incidence diffraction, truncation rod, and diffuse scattering. This requires exploring a large volume in reciprocal space, and thus four- or six-circle diffractometers are often employed for these measurements, sometimes in conjunction with ultrahigh-vacuum environmental chambers to keep the surface of the samples atomically clean (Feidenhans'l, 1989; Robinson and Tweet, 1992).

Liquid samples dictate a different geometry. By virtue of their horizontal surfaces, the plane of reflection is always vertical. The incident beam must be, therefore, bent down to impinge on the surface at the required angle of incidence. As the beams generated at synchrotron sources are highly collimated and horizontal, the beam is reflected downward by a tilted crystal. At laboratory-based rotating-anode sources, which produce a highly divergent beam, the incident beam is selected by a collimator looking back at the source at the required downgoing angle. In both cases, the sample is translated in the up–down direction as the angle of incidence is varied, to intercept the downgoing beam. Finally, the reflected beam is detected by a scintillator–phototube combination, an ion chamber, a diode, or a one- or two-dimensional detector.

Neutron-reflectivity measurements are carried out at both steady-state thermal-neutron reactors and pulsed spallation sources. At the former, the reflectometers function similarly to x-ray reflectometers since Bragg optics can be used for neutrons as well as x rays. Such reflectometers have been built in either the standard or liquid configurations. Reflectometers at spallation sources have been developed to take advantage of the relationship between the neutron wavelength and velocity (Felcher *et al.*, 1987). The sample and detector are fixed to provide the desired incident and exit angles α to satisfy the specular condition. The incident beam contains a spread of wavelengths with a known spectrum. The reflected beam is analyzed by a neutron time-of-flight technique to provide the reflected intensity for each wavelength. Through the relation $q = 4\pi/\lambda\sin\alpha$, this provides the required $R(q)$ curve. An important advantage of this energy-dispersive mode of measurement is that the reflectivity is acquired for a range of q values simultaneously without moving the sample or the incidence and exit slits and with a fixed area illuminated on the sample. Although similar energy-dispersive x-ray reflectivity measurements have been done, they are more complicated experimentally and seldom used in practice.

Several current reflectometers are described in detail by Russell (1990).

3. APPLICATIONS

3.1 Simple Liquids: Water

The liquid–vapor interface is intrinsically very sharp and flat. Its reflectivity should therefore closely follow the Fresnel-law $R_F(q)$. The water–vapor interface is perhaps the most common interface in nature and one of the first studied with x-ray reflectivity. The measured reflectivity (Braslau *et al.*, 1988) is reproduced in Fig. 3 along with the Fresnel law. Below $q_c = (4\pi/\lambda)\sin\alpha_c = 0.0217$ Å^{-1}, the reflectivity is nearly unity as expected. In the range $q_c \leq q \leq 0.4$ Å^{-1}, the reflectivity decreases by over seven orders of magnitude. While at low q values the measured curve is seen to follow closely Fresnel's law (solid line), as expected, above $q = 0.15$ Å^{-1} the data deviate increasingly from Fresnel's law. This deviation is exceptionally well described by the Gaussian function given in Eq. (8) as is apparent from the inset to Fig. 3. The best fit to the data yields an interfacial profile width of $\sigma = 3.3 \pm 0.1$ Å. This was shown to be predominantly due to thermally induced capillary waves in addition to a small intrinsic interfacial width. The dominant capillary contribution is predicted to have a width ~2.8 Å, in excellent agreement with the measurements. The capillary-wave model also predicts the dependence of the width upon the spectrometer resolution, surface tension, and temperature. All of these effects have been confirmed experimentally. Measurements of the diffuse scattering close to the specular direction also support the capillary-wave model. For liquid metals, the surface tensions of which are up to tenfold higher than that of water, recent measurements (Magnussen *et al.*, 1995; Regan *et al.*, 1995) reveal a surface roughness of ≤1 Å.

3.2 Thin Films: Water on Glass

For thin films on a substrate, the waves reflected from the upper and lower interfaces of the film interfere and yield a modulated reflectivity curve. The x-ray reflectivity curve of a thin water layer on a smooth borosilicate float-glass substrate is shown in Fig. 4 (Garoff *et al.*, 1989). Periodic oscillations in the curve are apparent. The monotonic decrease of their amplitude with q is due to the effects of roughness. The period of $\Delta q = 0.069\ \text{Å}^{-1}$ corresponds to a water-layer thickness of $d = 2\pi/\Delta q = 91$ Å. The solid line shows a model reflectivity curve where the density profile incorporates a water–vapor interfacial width of 4.2 Å, an average water-layer thickness of 91 Å, and a water–substrate interface width of 3 Å.

In addition to their scientific interest, thin films also have important industrial applications in the fields of coatings, electronics, biotechnology, and medicine. X-ray reflectivity has become a powerful probe of the structure of these films, and many films and multilayers have been studied including semiconductors, metals and alloys, polymers, Langmuir–Blodgett films, Langmuir films on water, liquid crystals, and liquid helium.

3.3 Polymer Films: Combined X-Ray and Neutron Reflectivity

The study of block copolymer films on solid substrates, using both x-ray and neutron reflectivity, demonstrates the complementary nature of these two techniques. In Fig. 5, x-ray and neutron reflectivity data (Mayes *et al.*, 1994) are shown for a diblock copolymer film, cast onto a flat polished silicon substrate and annealed for 24 hours. The polymer consists of a deuterated polystyrene block (d-PS) and a protonated polymethyl methacrylate block (PMMA). The PS, having a lower surface energy, aggregates upon annealing at the surface, while the PMMA resides at the silicon surface. As the electron densities of the two components are similar, the internal layered structure of the film does not give rise to optical-density variations for x rays. Thus only the fringes due to the total thickness of the film are observed in the x-ray data. A constant-density model fit to the x-ray data, shown in Fig. 5 as a solid line, shows very good agreement and yields a uniform film thickness of ~1120 Å, which corresponds to $3\frac{1}{2}$ bilayers. By contrast, the different neutron scattering lengths of hydrogen and deuterium, the first being positive and the other negative, result in very different optical densities for the internal layers for neutrons. This gives rise to additional interference beyond that reflecting the film's total thickness, as can be seen in the neutron data in Fig. 5. A model fit to the data, shown as the solid line, yields a thickness of ~320 Å for each of the PM and PMMA bilayers, with a diffuse transition layer of 50 Å between them and a surface roughness of 6 Å. As can be seen, the agreement between the model and data is excellent.

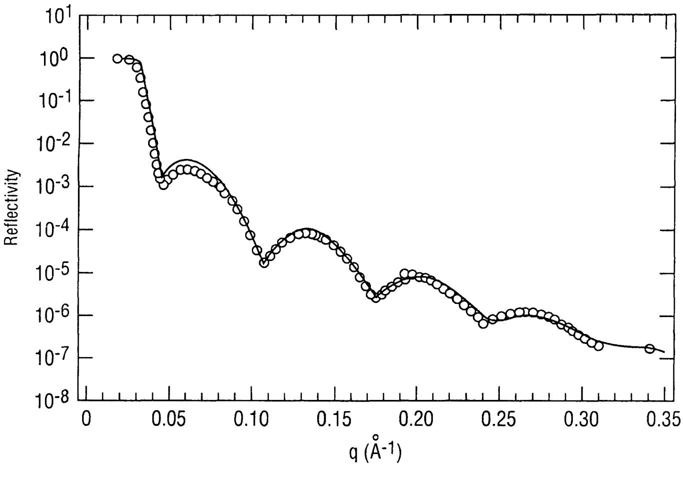

FIG. 4. X-ray reflectivity from a 91-Å-thick water layer on a glass substrate. The oscillations in the reflectivity curve originate from the interference between rays reflected from the water–vapor and the glass–water interfaces. Adapted from Garoff *et al.* (1989). For further details see Sec. 3.2.

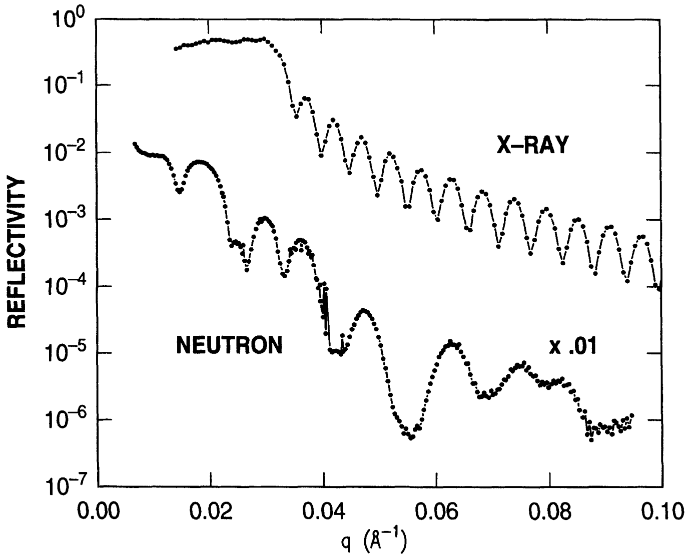

FIG. 5. X-ray and neutron reflectivities from an annealed deuterated-protonated di-block copolymer film cast onto a silicon surface. The irregular oscillations in the neutron reflectivity result from interference due to the alternating deuterated-protonated internal layered structure of the film, and the large variation of the neutron scattering length upon isotopic substitution. By contrast, since the charge density is equal for both types of layers, no internal structure is observed in the x-ray reflectivity, and only the regular fringes due to the total thickness of the film, ~1120 Å, are observed. Adapted from Mayes *et al.* (1994). Further details are given in Sec. 3.3.

A large number of similar studies on different polymers have been carried out, providing a wealth of information about the statics and kinetics of phase separation in polymer blends and block copolymers.

3.4 Polarized-Neutron Reflectivity: Ultrathin Fe Films

Theoretical studies indicate that magnetic moments of atoms at the surface of metals should be greater than those of bulk atoms. The same should hold for atoms in ultrathin metallic films. Polarized-neutron reflectivity has been successfully implemented to study this and related issues in a number of metallic films.

One such study (Huang *et al.*, 1993) addressed ultrathin Fe films, ranging in thickness from 4 to 16 Å. The reflectivity from a 16-Å film in a 200-Oe magnetic field at room temperature is shown in Fig. 6. As discussed earlier (Sec. 1.4), for neutron polarization parallel to the field, the usual nuclear scattering length of the Fe atom is enhanced by a magnetization term generated by an assembly of aligned atomic magnetic moments μ. For the other polarization, the scattering amplitude is *decreased* by the same amount. The difference between the parallel (filled circles) and antiparallel (open circles) polarizations is apparent in the figure. By fitting the reflectivities for both polarizations with a simple model, the effective iron layer thickness, 14 Å, and the iron scattering density, 7.0×10^{-6} Å^{-2}, have been obtained. These values indicate a (not unreasonable) 13% decrease in the density of Fe in the film as compared to the bulk. Contrary to theoretical predictions, the magnetic moment per atom is determined to be $(2 \pm 0.2)\mu_B$, the same as that of bulk Fe, regardless of thickness. However, temperature-dependent measurements show that the thinnest films exhibit a lower Curie temperature than that of the bulk and a magnetization axis tilted out of the film surface.

3.5 Crystalline Surfaces: Clean Au(001) in Vacuum

For crystalline surfaces, x-ray reflectivity provides information on the atomic arrangement of the topmost layers. As an example, the reflectivity from the Au(001) surface in ultrahigh vacuum is shown in Fig. 7 (Gibbs *et al.*, 1988). Note that the maximum q measured, ~5 Å^{-1}, is nearly an order of magnitude greater than what is typically achieved for simple liquids or thin films. For Au(001), the sample size limits the q range to values greater than q_c. The Bragg peak at $q = 3.08$ Å^{-1} corresponds to the bulk (002) reflection. The calculated reflectivity for an ideally ter-

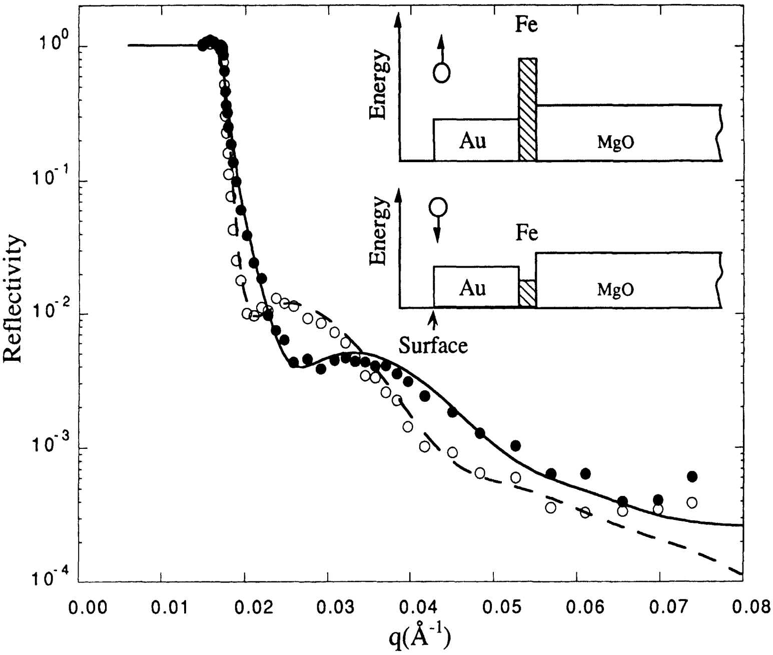

FIG. 6. Spin-dependent neutron reflectivity from a 16-Å-thick Fe film. A 200-Oe magnetic field aligns the spins parallel to the surface and transverse to the beam. Filled and open circles indicate data for neutron spin parallel (+) and antiparallel (−) to the applied field, respectively. Note the strong dependence of the measured curves on polarization. The inset describes the different potentials for neutron spin parallel (top) and antiparallel (bottom) to the magnetic field. Reproduced from Huang *et al.* (1993) with permission of The American Physical Society. See Sec. 3.4 for further details.

minated gold surface (dashed curve) deviates from the measured reflectivity, especially where the reflectivity is weak. This deviation results from the top gold layer, which does not have the same symmetry and gold density as the underlying gold layers. Rather, the top gold layer is "reconstructed" such that these top-layer atoms are arranged in a hexagonal pattern with a density 25% greater than the underlying layers and the topmost interlayer spacing is 20% larger than the bulk (001) layer spacing of 2.04 Å. The predicted reflectivity for this model (solid line) describes the essential features of the data. The prominent asymmetry around the (002) peak results from the 20% lattice expansion. Above 1170 K the reconstruction vanishes. A similar reconstruction was observed by x-ray reflectivity under electrochemical conditions (Ocko *et al.,* 1990). Studies have been carried out on other crystalline facets, different metals, and semiconductor surfaces, and at the electrode–electrolyte interface.

4. RELATED TECHNIQUES

There are several important techniques, related to reflectivity, whence additional information on the structure of the surface can be derived. For example, the surface "roughness" reduces the specularly reflected intensity by scattering intensity out of the specular direction. This diffuse scattering is dominated by, and can be analyzed to yield information on, the nature of the in-plane correlations of the "roughness," and their range and spatial dependence (Sinha *et al.,* 1988).

A few techniques take advantage of the refracted beam, which, for angles of incidence smaller than the critical angle, becomes evanescent. It travels along the surface and has an amplitude that decays exponentially with depth below the surface. As the penetration depth is of order ~50 Å, the diffraction of this wave by the surface layer provides information on the structure of the surface layer, free from any obscuring signal from the

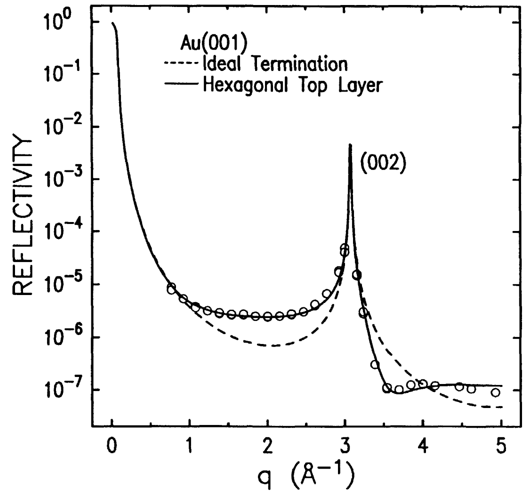

FIG. 7. X-ray reflectivity from the (001) surface of a single crystal of gold in ultrahigh vacuum. The Bragg peak at $q = 3.08$ Å^{-1} is the bulk (002) reflection. The measured reflectivity (open circles) deviates from the one calculated for an ideally terminated gold surface (dashed curve) but is well described by a surface model (see Sec. 3.5 for details) where the top gold layer is reconstructed in a hexagonal structure different from that of the bulk (solid curve). Adapted from Gibbs *et al.* (1988).

bulk. This effect is the basis for the grazing-incidence diffraction (GID) technique, a powerful tool for surface structure determination. Although it has been used to determine the liquid-surface structure factor of liquid metals and alloys, it is used most extensively for studying ordered surfaces and monolayers. Here the order in the surface translates into Bragg peaks, often unique to the surface and nonexistent for the bulk. However, perpendicular to the surface the order is restricted: a single molecular layer for Langmuir films or a crystalline half-space for the free surface of a crystal. Thus, the intensity distribution in the GID peaks is not pointlike as in a bulk Bragg peak; rather it is elongated perpendicular to the surface. These surface-normal scattering distributions, often called Bragg rods, are the two-dimensional equivalents of the Laue spots obtained in diffraction from bulk crystals, where the order is three dimensional. The intensity distribution along the rod can be analyzed to yield information mostly on the structure of the scattering medium perpendicular to the surface. Note that unlike specular reflectivity, which yields similar information but averaged over all of the surface, the Bragg rods originate in the ordered regions and hence yield information on these regions only. This is particularly useful for studies of partially ordered surfaces and monolayers.

The variation of the penetration depth of the refracted beam with the angle of incidence has also been used to investigate the concentration dependence of a specific atom on depth below the surface. This is done by using radiation of a fixed wavelength above the absorption edge of the atom and measuring the intensity variation with the *angle* of incidence of the fluorescence from that atom. A related technique, called SEXAFS for Surface Extended Absorption Fine Structure, uses the intensity variation with incident *wavelength* of the fluorescence at wavelengths close to, and above, an absorption edge. Since the angle of incidence is kept below the critical for all wavelengths, the fluorescence originates in the surface layer only, on account of the short penetration depth. This technique is useful for investigating the local structure near the given atom up to its second or third coordination sphere.

Works Cited

Abeles, F. (1950), *Ann. Phys. (Paris)* **5,** 596.

Als-Nielsen, J., Christensen, F., Pershan, P. S. (1982), *Phys. Rev. Lett.* **48,** 1107–1110.

Born, M., Wolf, E. (1975), *Principles of Optics,* Oxford: Pergamon.

Braslau, A., Pershan, P. S., Swislow, G., Ocko, B. M., Als-Nielsen, J. (1988), *Phys. Rev. A* **38,** 2457–2470.

Compton, A. H. (1923), *Phil. Mag.* **45,** 1121–1131.

Felcher, G. P. (1981), *Phys. Rev. B* **24,** 1595–1598.

Felcher, G. P., Hilleke, R. O., Crawford, R. K., Haumann, J., Kleb, R., Ostrowski, G. (1987), *Rev. Sci. Inst.* **58,** 609–619.

Fermi, E., Zinn, W. (1946), *Phys. Rev.* **70,** 103.

Garoff, S., Sirota, E. B., Sinha, S. K., Stanley, H. B. (1989), *J. Chem. Phys.* **90,** 7505–7514.

Gibbs, D., Ocko, B. M., Zehner, D. M., Mochrie, S. G. J. (1988), *Phys. Rev. B* **38,** 7303–7310.

Huang, Y. Y., Liu, C., Felcher, G. P. (1993), *Phys. Rev. B* **47,** 183–189.

Hughes, D. J., Burgy, M. T. (1951), *Phys. Rev.* **81,** 498–506.

Lu, B. C., Rice, S. A. (1978), *J. Chem. Phys.* **68,** 5558–5567.

Magnussen, O. M., Ocko, B. M., Regan, M. S., Penanen, K., Pershan, P. S., Deutsch, M. (1995), *Phys. Rev. Lett.* **74,** 4444–4447.

Marra, W. C., Eisenberger, P., Cho, A. Y. (1979), *J. Appl. Phys.* **50,** 6927–6933.

Mayes, A. M., Russell, T. P., Bassereau, P., Baker, S. M., Smith, G. S. (1994), *Macromolecules* **27,** 749–755.

Ocko, B. M., Wang, J., Davenport, A., Isaacs, H. (1990), *Phys. Rev. Lett.* **65,** 1466–1469.

Parratt, L. G. (1954), *Phys. Rev.* **95,** 359–369.

Penfold, J., Thomas, R. K. (1990), *J. Phys. Conden. Matter* **2,** 1369–1412.

Pershan, P. S., Als-Nielsen, J. (1984), *Phys. Rev. Lett.* **52,** 759–762.

Picht, J. (1929), *Ann. Phys.* **5,** 433–496.

Regan, M. J., Kawamoto, E. H., Lee S., Pershan, P. S., Maskil, N., Deutsch, M., Magnussen, O. M., Ocko, B. M., Berman, L. E. (1995), *Phys. Rev. Lett.* **75,** 2498–2501.

Russell, T. P. (1990), *Mater. Sci. Rep.* **5,** 171–271.

Sinha, S. K., Sirota, E. B., Garoff, S., Stanley, H. B. (1988), *Phys. Rev. B* **38,** 2297–2311.

Vineyard, G. (1982), *Phys. Rev. B* **26,** 4146–4159.

Further Reading

Review Articles

Als-Nielsen, J. (1991), in: G. S. Brown, D. E. Moncton (Eds.), *Handbook on Synchrotron Radiation,* Amsterdam: North-Holland, pp. 471–503.

Als-Nielsen, J., Kjær, K. (1989), in: T. Riste, D. Sherrington (Eds.), *Phase Transitions in Soft Condensed Matter,* New York: Plenum, pp. 113–138.

Feidenhans'l, R. (1989), *Surf. Sci. Rep.* **10,** 105–188.

Hamley, I. W., Pedersen, J. S. (1994), *J. Appl. Crystall.* **27,** 29–35, 36–49.

Penfold, J., Thomas, R. K. (1990), *J. Phys. Cond. Matter* **2,** 1369–1412.

Robinson, I. K., Tweet, D. J. (1992), *Rep. Prog. Phys.* **55,** 599–651.

Russell, T. P. (1990), *Mater. Sci. Rep.* **5,** 171–271.

Schlossman, M. L., Pershan, P. S. (1990), in: D. Langevin (Ed.), *Light Scattering by Liquid Surfaces and Complementary Techniques,* New York: Dekker, pp. 365–404.

Books

Bienfait, M., Gay, J. M. (Eds.) (1989), "Proceedings of the First International Conference on Surface X-ray and Neutron Scattering," *J. Phys. (Paris) Colloq.* **C7,** 1–304.

Dosch, H. (1992), *Critical Phenomena at Surfaces and Interfaces,* Berlin: Springer.

Felcher, G. P., Russell, T. P. (1991), "Methods of Analysis and Interpretation of Neutron Reflectivity Data," *Physica B* **173,** 1–210.

Zabel, H., Robinson, I. K. (Eds.) (1992), *Surface X-ray and Neutron Scattering,* Springer Proceedings in Physics, Vol. 61, Heidelberg: Springer.

X-RAY DIFFRACTION

Boris W. Batterman, *Applied & Engineering Physics and Cornell High Energy Synchrotron Source (CHESS), Cornell University, Ithaca, New York, U.S.A.*

Introduction: X Rays and Visible Light 491
1. **The Structure of the Atom** 491
2. **Periodic Structures** 492
3. **Kinematic versus Dynamical Diffraction** 494
3.1 Kinematic Diffraction 494
3.2 Dynamical Diffraction 495
4. **Relationship between X-Ray Crystallography and Microscopy** 495
5. **Crystallography and the Phase Problem** 496
6. **Structural Determination of Nonperiodic Systems** 498
6.1 Gases 498
6.2 Liquids 499
7. **Intensity versus Amplitude-Derived Structural Information** 500
8. **X-Ray Instrumentation** 501
8.1 Detectors 501
8.2 CCD Devices 502
8.3 Storage Phosphors 502
9. **X-Ray Diffraction Methods** 502
9.1 Powder Diffraction 502
9.2 Laue Methods 502
9.3 Single-Crystal Techniques 503
Works Cited 503
Further Reading 504

INTRODUCTION: X RAYS AND VISIBLE LIGHT

Light is scattered by matter because its electromagnetic field drives electrons into oscillatory motion, which causes them to radiate some of the incident flux. This fundamental phenomenon is valid for visible as well as x-ray light, but the competition with other types of interaction can depend strongly on the energy (frequency) of the light. For this treatment we will consider x-ray interactions as elastic events (no energy loss) in which many centers collectively and coherently contribute to the scattering process.

Visible light and x rays behave similarly in this type of geometric scattering, the important parameter being the relative value of the wavelength (λ) to some characteristic length (d) in the scatterer. For comparable values of λ/d, diffraction is basically independent of λ.

Diffraction with visible light is usually from one- or two-dimensional structures such as slits, edges, or arrays of slits (gratings), where λ/d basically defines the problem. On a more limited front, optical diffraction from a three-dimensional object can be usefully pursued. With x rays of wavelengths the order of 1 Å (10^{-8} cm), the atomic granularity of the basic structure determines the characteristic length of the diffraction.

X rays, because of their high penetrating power compared to visible light, diffract from three-dimensional structures. The conditions for strong diffraction from three-dimensional periodic structures are much more restrictive than from lower-dimensional systems, making strong interferences a much rarer event. This is discussed in more detail in Sec. 2.

1. THE STRUCTURE OF THE ATOM

The diffraction pattern observed from an object illuminated by a plane wave is the summation of the amplitudes of scattered waves from each component of the object taking into account the path-length–induced phase differences. The amplitude of scattering E from an arrangement of matter defined

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

by the vectors **r**, with scattering amplitude per unit of volume $f(\mathbf{r})$, is*

$$E = \int_{\text{object}} f(\mathbf{r}) e^{i(\mathbf{k}-\mathbf{k}_0)\cdot\mathbf{r}} dv, \tag{1}$$

where $\mathbf{k}_0$ and **k** are the incident and diffracted wave vectors with magnitude $|\mathbf{k}| = 2\pi/\lambda$. $i(\mathbf{k} - \mathbf{k}_0)\cdot\mathbf{r}$ is the phase angle of a scatterer at position **r** with respect to some reference scatterer at an arbitrary origin and is merely 2π times the optical path-length difference in units of λ. $\mathbf{k} - \mathbf{k}_0$ is called the scattering vector (Fig. 1).

X rays are scattered by electrons with an amplitude given by $f_e = e^2/mc^2$, the Thompson scattering length, so that (1) becomes

$$E = f_e \int \rho(\mathbf{r}) e^{i(\mathbf{k}-\mathbf{k}_0)\cdot\mathbf{r}} dv, \tag{2}$$

where $\rho(r) = \Psi\Psi^*$ is the density of electrons of wave function Ψ. (e is the charge of an electron, m its mass, and c the velocity of light; $e^2/mc^2 = 1.38 \times 10^{-13}$ cm.) If the object is an atom, then the quantity

$$f = \int \rho(\mathbf{r}) e^{i(\mathbf{k}-\mathbf{k}_0)\cdot\mathbf{r}} dv \tag{3}$$

is called the *atomic scattering factor,* or simply the *form factor.* [Note that if all the electrons are treated as point scatterers at the atom center, then $f = \int \rho(r) dv = Z$, the number of electrons in the atom.] If an atom (or a dilute gas of atoms) is illuminated by a plane wave of x rays, the scattered intensity I, as a function of the scattering vector $\mathbf{k} - \mathbf{k}_0$, will be proportional to the quantity $|f|^2$. From a complete intensity data set, Fourier inversion of $\sqrt{I}$ will give $\rho(r)$, the radial distribution of electron density in the atom.

Figure 2 shows the form factor for neon gas derived from the experimental scattering intensities. 2θ is the total scattering angle. The theory curves are from theoretical calculations of the electron wave functions for neon (Compton and Allison, 1935).

2. PERIODIC STRUCTURES

Structures that are periodic, where each structural element is precisely repeated on a grid determined by translational vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, have scattering patterns with sharp maxima in certain directions. For example, in one dimension, a set of scattering centers spaced $\mathbf{a}_1$ apart along a straight line will produce maxima when the following condition with integer h is met [see Fig. 3(a)]:

$$(\mathbf{k} - \mathbf{k}_0)\cdot\mathbf{a}_1 = 2\pi h. \tag{4}$$

This condition shows that maxima exist in directions **k** when the path difference between the incident $\mathbf{k}_0$ and scattered waves **k** from each element is a multiple h of the wavelength of the x rays. The diffraction pattern in this case is a series of cones of scat-

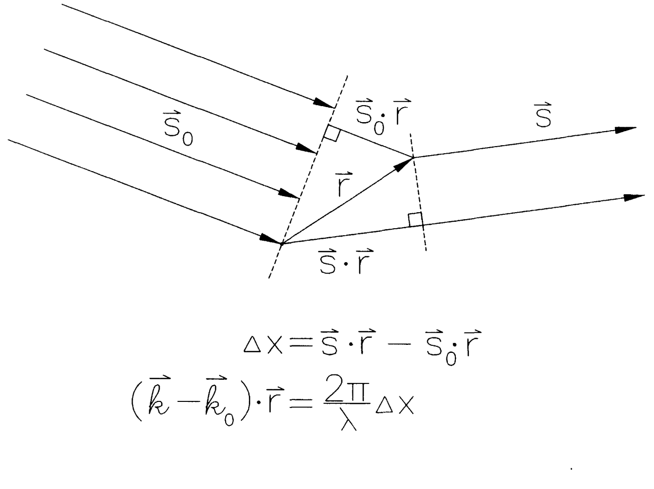
$\vec{S}_0$
$\vec{S}_0 \cdot \vec{r}$
$\vec{S}$
$\vec{r}$
$\vec{S} \cdot \vec{r}$
$\Delta x = \vec{S}\cdot\vec{r} - \vec{S}_0\cdot\vec{r}$
$(\vec{k} - \vec{k}_0)\cdot\vec{r} = \frac{2\pi}{\lambda}\Delta x$

FIG. 1. The scattering vectors showing the path-length phase of a scatterer at vector position **r**. **S** and $\mathbf{S}_0$ are unit vectors in the diffracted and incident directions, respectively. Δx is the path-length difference with respect to a scatterer at position 0.

---

*We use complex notation $e^{i\mathbf{k}\cdot\mathbf{r}} = \cos\mathbf{k}\cdot\mathbf{r} + i \times \sin\mathbf{k}\cdot\mathbf{r}$ to represent a plane wave traveling in the **k** direction. $\mathbf{k}\cdot\mathbf{r} = 2\pi x/\lambda$, where x is distance along the **k** direction.

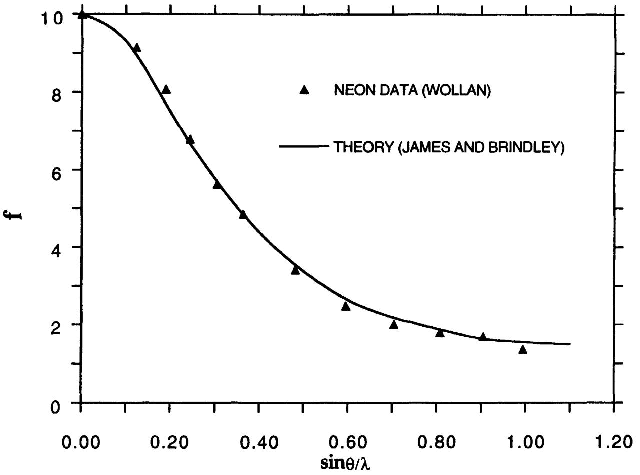

FIG. 2. The atomic x-ray scattering factor for neon as determined by Wollan (Compton and Allison, 1935). The total scattering angle is 2θ.

tering centered about $\mathbf{a}_1$ with the cone angle depending on the incident direction with respect to the line of scatterers.

For two-dimensional scattering, an additional similar condition must be satisfied for a periodicity defined by the translation vector $\mathbf{a}_2$, viz:

$$(\mathbf{k} - \mathbf{k}_0)\cdot\mathbf{a}_2 = 2\pi k \quad (k \text{ integer}). \tag{5}$$

Scattering from a two-dimensional array (such as a mesh screen) will take place when conditions (4) and (5) are simultaneously satisfied, and this is shown geometrically in Fig. 3(b) as the intersections of two cones centered around the respective axes $\mathbf{a}_1$ and $\mathbf{a}_2$.

If we now go to three-dimensional (3-D) periodic structures and consider the problem geometrically, we see that a third equation similar to (4) and (5) must be satisfied (involving the vector $\mathbf{a}_3$ and a third integer l), which requires the mutual intersection of three cones in space. Such a common intersection is clearly a rare occurrence. In two dimensions, two cones will, in general, intersect producing strong diffracted rays. With 3-D gratings, conditions must be carefully set to obtain a strong diffracted beam. Shooting a well-collimated monochromatic x-ray beam at a crystal will produce a diffracted beam with vanishingly small probability. However, as we show later, certain experimental conditions make 3-D diffraction a more likely event.

Equations (4) and (5) with the additional one for a third translational vector $\mathbf{a}_3$ are

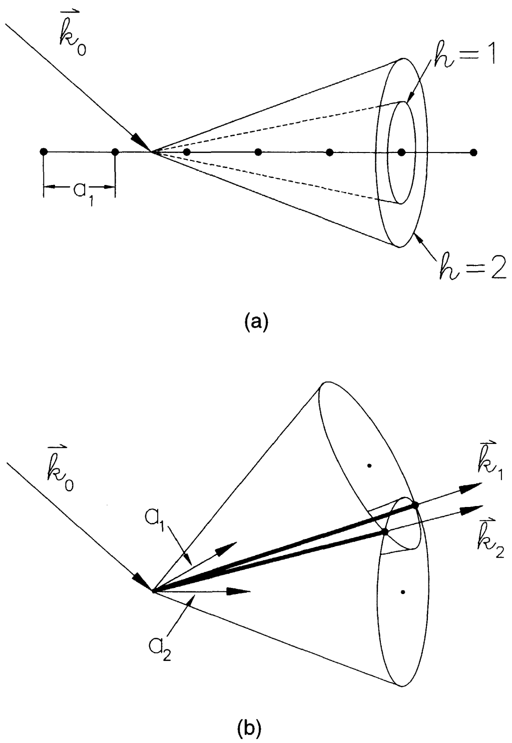

FIG. 3. Diffraction from one- and two-dimensional gratings. (a) The diffraction pattern from a one-dimensional array of scatterers consists of a series of concentric cones centered on the line of scatterers. The indices h represent different orders. (b) Diffraction from a 2-D grating will, in general, produce a pair of diffracted rays for each order pair (h,k). The diffracted rays are the intersection of the cones about $\mathbf{a}_1$ and $\mathbf{a}_2$.

known as the *Laue equations,* which we consolidate in terms of unit vectors $\mathbf{s}$ and $\mathbf{s}_0$ in the diffracted and incident beam directions respectively:

$$
\begin{aligned}
(\mathbf{s} - \mathbf{s}_0)\cdot\mathbf{a}_1 &= \lambda h,\\
(\mathbf{s} - \mathbf{s}_0)\cdot\mathbf{a}_2 &= \lambda k,\\
(\mathbf{s} - \mathbf{s}_0)\cdot\mathbf{a}_3 &= \lambda l.
\end{aligned} \tag{6}
$$

In a remarkable simplification, Bragg (Ewald, 1962) showed that these three equations are equivalent to mirror reflection from planes of atoms. For planes spaced d apart, strong reflections will occur when the incident beam of wavelength λ makes an angle θ with respect to the planes such that

$$\lambda = 2d \sin\theta. \tag{7}$$

This is known as *Bragg's law* and is the fundamental equation in x-ray scattering (see Fig. 4). When (7) is satisfied, the path-length difference between successive planes d equals one wavelength.

For crystalline materials, Bragg's law determines the condition for strong diffraction from an arbitrarily large number of parallel planes (Miller planes) defined by the integers h, k, l. We show in a later section that the intensities of Bragg reflections from the Miller plane (hkl) can be related to structure (i.e., the actual distribution of atoms).

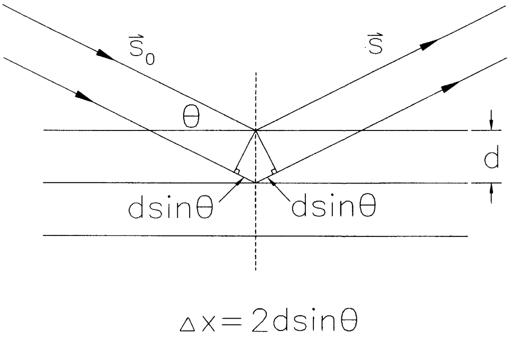

FIG. 4. A simple schematic showing how Bragg's law can be derived by requiring that the scattering path-length difference between two successive planes of scatterers be a multiple of a wavelength.

3. KINEMATIC VERSUS DYNAMICAL DIFFRACTION

3.1 Kinematic Diffraction

In the usual simplified treatment of the scattering of x rays, it is assumed that when an atom scatters an x-ray beam, the scattered wave has a negligible interaction with the atoms on its way out of the crystal. This approximation is known as the *kinematic theory of diffraction.* It predicts an experimentally observable Bragg intensity of scattering by a small crystal of volume Δv as

$$I_{hkl} = I_0(e^2/mc^2)^2 D|F_{hkl}|^2\Delta v, \tag{8}$$

where F_{hkl}, the structure factor for the Bragg (Miller) planes h, k, l, is related to the atomic scattering factors f_i of Eq. (3) with phase factors determined by the positions $(xyz)_i$ of the ith atom in the basic repeating unit cell of the crystal:

$$
\begin{aligned}
F_{hkl} &= \sum_{\text{atoms}} f_i[\cos 2\pi(hx_i + ky_i + lz_i)\\
&\quad + i \sin 2\pi(hx_i + ky_i + lz_i)]\\
&= |F_{hkl}|e^{i\Phi_{hkl}}.
\end{aligned} \tag{9}
$$

The structure factor F_{hkl} contains the positional coordinates of all the atoms and is in essence the crystal structure. The quantity D in (8) contains geometric and extrinsic parameters such as detector sizes, distances, and angular speeds of crystal rotation, which depend on the particular diffraction technique used.

If one could measure F_{hkl} for a large number of reflections, the coordinates of each atom could be computed. However, the quantity F_{hkl} has both amplitude $|F|$ and phase Φ as represented in Eq. (9). The amplitude is

$$
\begin{aligned}
F = \Bigg\{&\left[\sum_i f_i \cos 2\pi(hx_i + ky_i + lz_i)\right]^2\\
&+ \left[\sum_i f_i \sin 2\pi(hx_i + ky_i + lz_i)\right]^2\Bigg\}^{1/2}.
\end{aligned} \tag{10}
$$

The phase angle is

$$\Phi_{hkl} = \tan^{-1}\left[\frac{\sum_i f_i \sin 2\pi(hx_i + ky_i + lz_i)}{\sum_i f_i \cos 2\pi(hx_i + ky_i + lz_i)}\right]. \tag{11}$$

Experimentally one measures the intensity [Eq. (8)], which only gives the magnitude $|F_{hkl}|$ and not the phase information intrinsic in Eq. (9). It is for this reason that a careful measurement of large numbers of Miller reflections from a crystal will *not* provide sufficient information to extract the positional coordinates $(xyz)_i$ of each atom.

In x-ray crystallography, the art of deriving the atomic structure of a crystal, the crystallographer must use other means to arrive at the phases of the structure factor F. The need to derive, or guess at, the phases has spawned a variety of innovative procedures (too arcane to discuss in this treatment), some of which have led to Nobel prizes for their creators (Karle and Hauptman, 1953).

Another way of formulating the diffraction pattern from crystalline materials is to treat the crystal as a periodic, continuous distribution of matter $\rho(\mathbf{r})$. $\rho(\mathbf{r})$ is the electron density at point $\mathbf{r}$. Using very basic applications of Fourier series, it can be straightforwardly shown that $\rho(\mathbf{r})$ can be expressed as a 3-D Fourier series whose coefficients are F_{hkl}, the x-ray structure factors,

$$\rho(xyz) = \sum_h \sum_k \sum_l F_{hkl} e^{2\pi i(hx+ky+lz)}. \qquad (12)$$

With complete knowledge of F_{hkl}, the crystal structure $\rho(x,y,z)$ can be calculated. This somewhat different approach requires that both the magnitude and phase of the structure factor F_{hkl} be known and demonstrates again that intensity data alone, which gives only the amplitude $|F_{hkl}|$ and not the phase ϕ, is *not* sufficient to solve the structure.

In those cases (perfect crystals and strong reflection) where one cannot neglect multiple scattering, the quantitative description of intensities discussed earlier is no longer valid, and we enter the regime commonly described as dynamical diffraction.

3.2 Dynamical Diffraction

Multiple diffraction strongly affects the strength and phase of the propagating primary wave in the crystal. The intensity is no longer proportional to Δv, the volume of the crystal illuminated, and in the ideal case the intensity in a Bragg peak is proportional to F, not F^2.

It is not possible in this limited treatment of diffraction to delve deeply into the intricacies of dynamical diffraction, but a few points are worth mentioning. Bragg reflection curves from the face planes of a perfect crystal are predicted to have a particular top-hat shape with an angular range of total reflection and a finite width. This is called the Darwin curve and represents accurately the reflectivity curve from monochromator crystals used with synchrotron radiation sources.

In certain cases, standing waves are generated in the crystal that dramatically affect the diffracted intensities. In one case, called the Borrmann effect, the effective absorption of x rays from a crystalline material can be reduced by several orders of magnitude when these standing waves are created. See Batterman and Cole (1964).

4. RELATIONSHIP BETWEEN X-RAY CRYSTALLOGRAPHY AND MICROSCOPY

Why does the crystallographer have to go through the agony of measuring a million or so x-ray reflections while the microscopist (optical or electron) puts the sample in an instrument and literally looks at the magnified object? Why, in those cases, is the phase problem not a problem?

The main difficulty in doing x-ray microscopy is that there are no lenses that will refract and collect x rays to an image plane. X rays emanating from a point on the object cannot be focused to a corresponding point on an image plane.

When light interacts with matter, it scatters, creating a diffraction pattern. The eye, or lens in a microscope, collects the multibeamed diffraction pattern and simultaneously combines the beams together at an image point such that the relative amplitudes and phases of all scattered wavelets are coherently combined.

It is not necessary to identify the diffraction pattern as a part of the imaging process. Since all the diffracted beams occur and are collected at the same time from the initial incident plane wave, there is no phase problem because the diffraction pattern is not measured separately. The waves continue on to an objective lens, which combines all the diffracted rays keeping track of all the relative phases and amplitudes to illuminate the image plane.

In x-ray diffraction, because of the technical problems of simultaneously creating and recombining diffracted beams, each beam must be separately recorded, thereby losing the phase. The intensities of the diffracted beams are proportional to $F \cdot F^* = |F|e^{i\phi} \cdot |F|e^{-i\phi} = |F|^2$ and therefore do not give up phase information.

5. CRYSTALLOGRAPHY AND THE PHASE PROBLEM

The human rhinovirus-14 contains about 10^6 atoms. The actual virus unit is a soccer ball–shaped structure whose outer coat (the leather in the ball analogy) contains about 0.75×10^6 atoms. The outer coat is the sensitive part that attaches itself to the human cell. Knowing its precise topological structure is an essential step in the key to understanding how the virus infects the cell and, more to the point, provides certain essential information to design an antiviral drug.

The virus can be crystallized with about four virus units in the spatially repeating unit cell. The solving of the structure requires the coupling of knowledge of the biology of the system with sophisticated experimental techniques for crystal growth and a subtle intuitive understanding of the mathematical and computational techniques.

This *tour de force* in structure determination was performed by Michael Rossmann (Rossmann *et al.*, 1985) of Purdue University. The measurement involved determining the intensities of 2.5×10^6 Bragg reflections in order to establish the positions of about 10^6 atoms.

The experimental setup was quite straightforward by modern standards. A monochromatic beam of synchrotron-radiation x rays of energy about 8 keV (1.5 Å) was incident on a crystal of the virus, which was rotated through a fraction of a degree while exposing an x-ray film to the diffracted beams. Crystals of the rhinovirus are shown in the photograph of Fig. 5. A typical x-ray exposure is shown in Fig. 6. The relative intensities of the spots were determined from the films, and the $|F_{hkl}|^2$ were determined using Eq. (8).

As a help in determining the phases, a crystal of the virus was grown that was "doped" with gold atoms in such a way that these heavy atoms resided on specific sites in

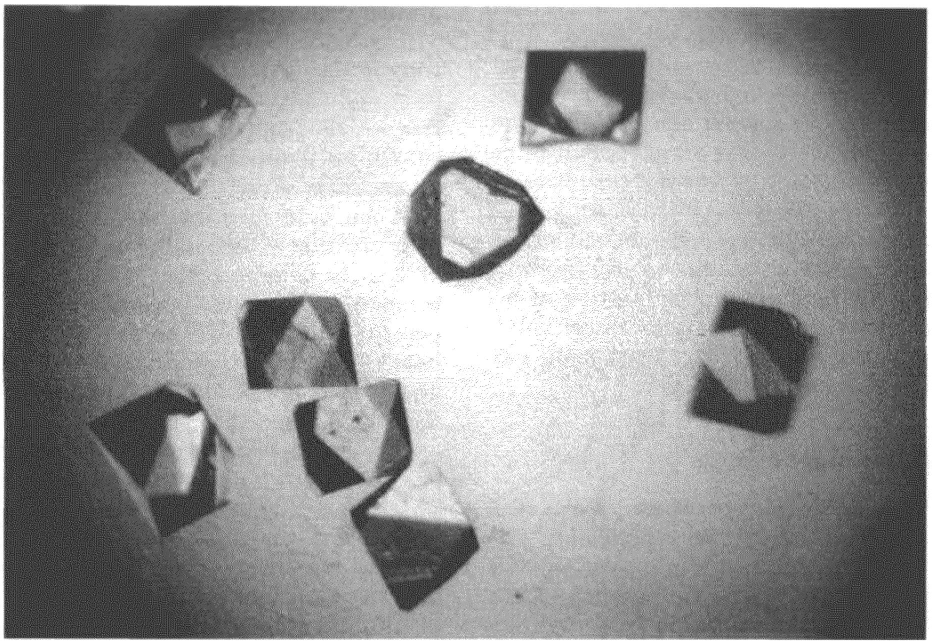

FIG. 5. Crystals of the rhinovirus-14 grown by Michael Rossmann. Typical crystal sizes are about 0.5 mm.

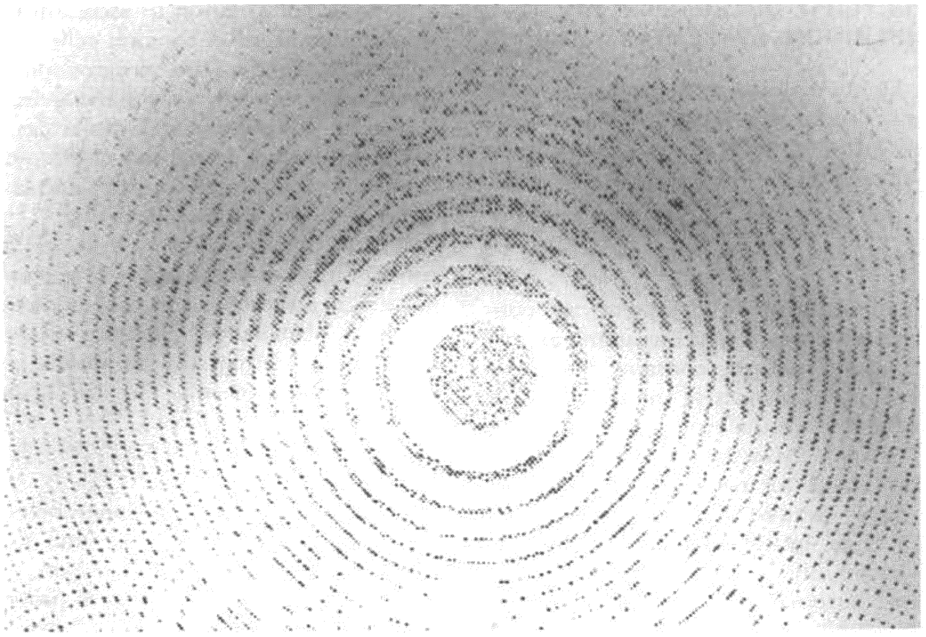

FIG. 6. Diffraction photographs of the HRV14 virus taken at the Cornell High Energy Synchrotron Source (CHESS). The crystal was oscillated through 0.3°, using x rays of 1.56-Å wavelength. Typical modern exposure times for a diffraction pattern of this quality would be considerably less than 1 min.

the structure. The gold atoms contain many electrons and therefore scatter more strongly than the typical atoms making up the virus and position themselves in the structure so as not to strongly perturb the native structure. Comparing the intensities of the peaks (see Fig. 6) between the native and the gold-doped "isomorphous replacement" structure gives important phase information. In this instance, the gold atoms act as a reference phase in a similar way that a reference beam in light holography allows the hologram to record phase information on a film. From the rough phases obtained in this comparison, and the F_{hkl} derived experimentally, an electron density map is computed using Eq. (12). This first map has poor resolution primarily because of noise (errors) in the data, which limit the resolution to about 6 Å.

This first try at the electron density uses roughly estimated phases because of errors in the data and experimental values of $|F_{hkl}|$. The resulting map has poor resolution and does not display the known symmetry of the structure very precisely. Here, the first trick in a boot-strapping iterative procedure is initiated. The computed electron density is averaged and made to comply with the known symmetry of the virus. Using this averaged density $\rho(r)$, the structure factors are calculated using Eq. (3), which is the same as F_{hkl} in Eq. (9) but for a continuous distribution of matter. New phases are calculated from F_{hkl} using Eq. (11). Now, these new phases and the *old* $|F_{hkl}|$, experimentally determined (the latter are the only things known for sure), are used in Eq. (12) to recalculate $\rho(r)$, which is the improved estimate of the structure. This process is repeated, each time providing new phases and incrementally adding higher-order terms in the Fourier series of Eq. (12), thus improving the resolution until a discernible high-resolution map is determined.

This simplified explanation is merely to give the reader a feeling for the complexities of getting around the phase problem through innovative experimental and computational techniques. In this manner, Rossman *et al.* (1985) obtained the first mammalian structure determined by x-ray crystallography.

6. STRUCTURAL DETERMINATION OF NONPERIODIC SYSTEMS

In a highly ordered and periodic system such as a crystal, the x-ray diffraction patterns consist of a sharp periodic distribution of x-ray beams as we have seen in Fig. 6.

6.1 Gases

In the most disordered system, we consider a gas of random noninteracting atoms. The diffraction pattern contains no sharp peaks, but one could derive the radial distribution of charge in the atom from the intensity data. In Fig. 2 we show the experimental scattering curves for neon gas with a theoretical curve derived from Thomas–Fermi wave functions. In this case, each atom scatters independently of all the others, and the diffraction pattern is simply the sum of the intensities from each atom illuminated by the beam. For a crystalline material, on the other hand, all atoms contribute to each diffraction spot in a highly complex manner, precisely determinable from their known positions in the crystal.

The order present in real materials continuously covers the range between these two extremes; and, not surprisingly, the complexity of the diffraction patterns ranges from trivial to intriguingly subtle. In the simplest cases, such as scattering from a gas, structural information such as the radial distribution of electrons in the atom can be obtained from intensity data alone. In the more complex structures we must have intensity and phase information to ascertain the positions of the atom in the unit cells.

Let us follow the progression toward more complex structures starting from scattering by gas atoms and molecules. If our gas sample were composed of diatomic molecules of O_2 instead of the simple single-atom rare gas discussed in Fig. 2, we can still treat the scattering of N molecules as N times that obtained from a single molecule, since the molecules are noninteracting and randomly distributed. Then the structure factor, Eq. (9), defines the amplitude scattered from a single molecule. For the O_2 molecule

$$F = \sum f_i e^{i(\mathbf{k}-\mathbf{k}_0)\cdot\mathbf{r}} = f_0(1 + e^{i(\mathbf{k}-\mathbf{k}_0)\cdot\mathbf{r}_{ab}}),$$

where $\mathbf{r}_{ab}$ is the vector distance between the two atoms of the molecule and f_0 is the atomic scattering factor for oxygen.

The intensity of scattering is proportional to $|F|^2$, which, averaged over all orientations of the O_2 dumbbells, can be shown to be

$$I \approx |F|^2 = 2f_0^2\left(1 + \frac{\sin kR}{kR}\right), \qquad (13)$$

where R is the distance between two oxygen atoms.

The intensity of scattering is simply the sum of the scattering from two random oxygen atoms [the first term in Eq. (13)] plus a sinusoidal modulation with a period determined by the separation R of the two atoms in the molecule. Figure 7 shows the experimental curve and theoretical prediction for the O_2 molecule (Compton and Allison,

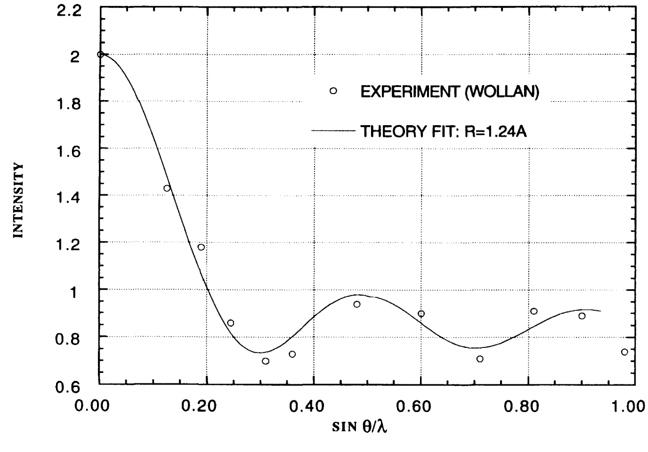

FIG. 7. X-ray scattering from diatomic oxygen. The solid curve is the fit of Eq. (13) with a value of R = 1.24 for the separation of the two atoms in the O_2 molecule.

1935). From a fit of the intensity vs scattering angle according to Eq. (13), the value of R can be readily obtained. The data in Fig. 7 give a value of $R = 1.24$ Å for the separation of the oxygen atoms in the O_2 molecule.

For more complex molecules, more sinusoidal periods will be present in the intensity function—in fact, one for each interatomic distance. The general expression for the intensity from an N-atom molecule will be

$$I = I_e \sum_{i=1}^{N} \sum_{j=1}^{N} f_i f_j \frac{\sin kr_{ij}}{kr_{ij}}, \tag{14}$$

where r_{ij} is the interatomic distance between atoms i and j and I_e is the Thompson scattering from a single electron. It can be readily seen that Eq. (14) reduces to the result for O_2 [Eq. (13)] when only two atoms are present.

Note that the structural information derived from the intensity of x-ray scattering gives interatomic distances r_{ij}, but not the spatial orientation of the vector distances between atoms. However, in many instances this limited structural information can be the key to deriving the complete structure.

6.2 Liquids

Liquids and amorphous materials have complex structures where atoms interact strongly but do not produce a periodic crystalline structure. Nevertheless, there can be strong local ordering between neighboring atoms that strongly affects the macroscopic properties of the liquid.

In a manner similar to the scattering from dilute gases, x rays can give important information about the near-neighbor environment of the individual constituents. An analysis expanding on the concepts discussed earlier for gases leads to an expression for the intensity of x rays scattered from a monatomic liquid as

$$I = I_e \int_{\text{specimen}} 4\pi r^2 [\rho_0 - \rho(r)] \frac{\sin kr}{kr}\, dr, \tag{15}$$

where $4\pi r^2 \rho(r) dr$ is the number of atoms in a spatial shell dr of radius r about a given atom at the origin. This is the so-called radial distribution function of the material.

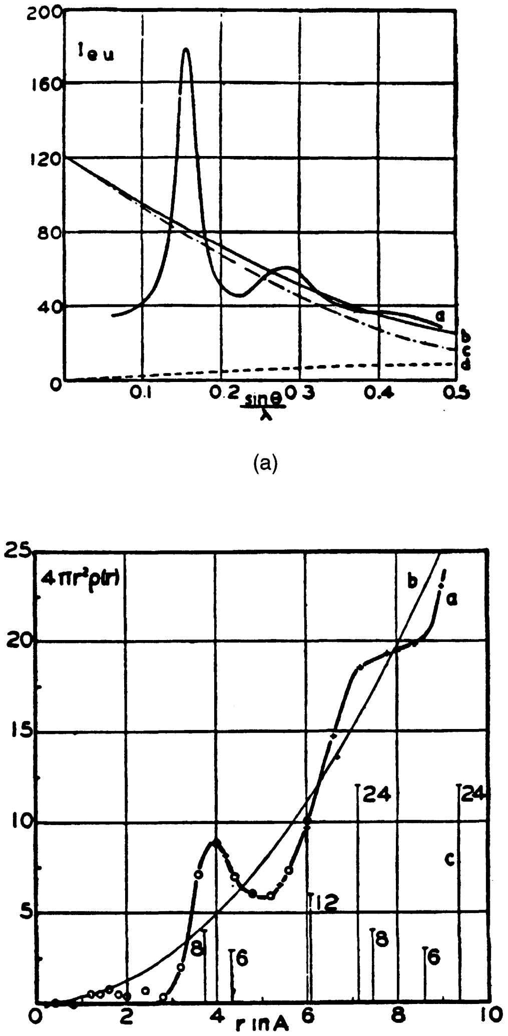

(a)

(b)

FIG. 8. (a) The experimental intensity curve for liquid sodium as a function of $\sin\theta/\lambda$ where 2θ is the total scattering angle and λ is the wavelength. (b) The radial distribution functions $4\pi r^2 \rho(r)$ derived from the data in (a). The curve b is the distribution in a homogeneous specimen. The bar graph c shows the distribution of neighbors in crystalline sodium (Tarasov and Warren, 1936).

We show in Fig. 8 the intensity curve (James, 1950) from liquid sodium together with the derived radial distribution function. Liquid sodium has a structure similar to that of a hard-sphere liquid, which can be modeled by the radial distribution function of a container randomly filled with ball bearings. For more complex liquids—for example, H_2O

(Morgan and Warren, 1938)—the effect of the molecular basis of the structure is apparent in the diffraction pattern of Fig. 9 compared to Fig. 8.

7. INTENSITY- VERSUS AMPLITUDE-DERIVED STRUCTURAL INFORMATION

Section 6 on scattering from nonperiodic structures shows that intensity information without phases can give important parameters of the structure of a material. One can generalize these results as follows: The intensity of scattering of x rays, without the phase information, gives only information on the interatomic distances present in a structure. This is valid irrespective of whether the structure is crystalline or amorphous. If one needs further information about how the interatomic distances are vectorially related to one another, then additional information about scattering phases is necessary.

Even for highly complicated structures such as the cold virus HR14 (see Sec. 5), one can, from intensity data alone, get the crystalline analog of the radial distribution function in liquid scattering. If the density of matter in a crystal is defined by $\rho(\mathbf{r})$, then this quantity gives the electron density at the point $\mathbf{r}$ in the crystal. If an atom is at that point, then the density is high. If the vector $\mathbf{r}$ lies between atoms, the density is low. If we now take the product of that density with the one a vector distance $\mathbf{X}$ away,

$$P_r(\mathbf{X}) = \rho(\mathbf{r})\rho(\mathbf{r} + \mathbf{X}),$$

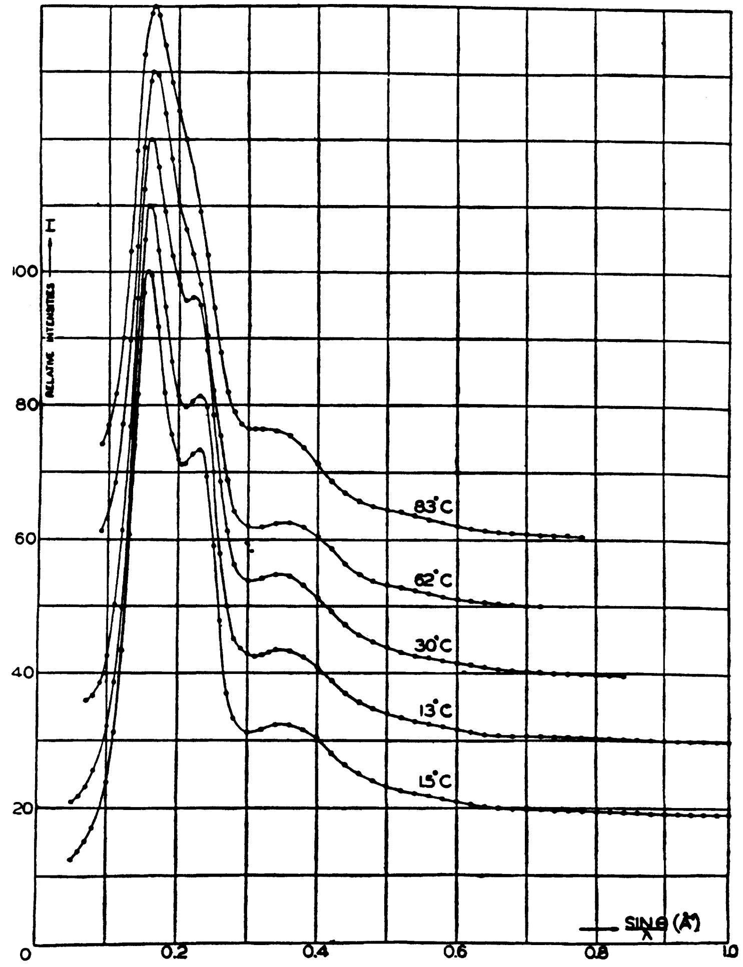

FIG. 9. The diffraction pattern of liquid H_2O at various temperatures (Morgan and Warren, 1938).

then $P_r(\mathbf{X})$ is the correlation function of the distribution of matter at a point $\mathbf{r}$ in the unit cell with a similar quantity for a point at vector distance $\mathbf{X}$ away. The average of this quantity over the repeat unit cell of the structure,

$$P(\mathbf{X}) = \iiint_{\text{cell volume}} \rho(r)\rho(\mathbf{r} + \mathbf{X})dv,$$

is a quantity called the autocorrelation function of the structure. Using methods similar to those leading to Eq. (12), we can show that the function $P(\mathbf{X})$ can be expressed as a Fourier series whose coefficients are not F_{hkl} as in Eq. (12) but F_{hkl}^2, which does not include the phase of the structure factor:

$$P(\mathbf{X}) = \sum_h \sum_k \sum_l F_{hkl}^2 e^{2\pi i(hX_x + kX_y + lX_z)}. \tag{16}$$

$P(\mathbf{X})$ is called the Patterson function of the crystal. F_{hkl}^2 is directly determined from the intensities of the x-ray diffraction Bragg peaks. The vector $\mathbf{X}$ is a vector in the real crystal. So, measuring a set of Bragg peak intensities, deriving F^2 from Eq. (8), and plugging them into (16) gives, without phase information, the self-correlation function of the structure. This is precisely what was done to get the radial distribution function of a liquid in Eq. (15) from the intensity of scattering. For a crystal, this function does not produce the structure, but it gives unambiguous information about relative atomic positions in the unit cell. This can be a significant clue to crack the structure.

In Fig. 10 we show a projection of the Patterson function of FeS_2 together with corresponding actual structure. Note that the Patterson function is more complicated than the actual crystal. This is because every vector distance between atoms is represented as a peak with respect to the origin of the Patterson function ($X = 0$). Thus the relative position of the atoms is lost.

Since the general rule that you get nothing for nothing applies, the Patterson function, easily obtainable, gives a very complicated semblance of the real crystal. However, in many cases, important clues about the structure can be quickly obtained.

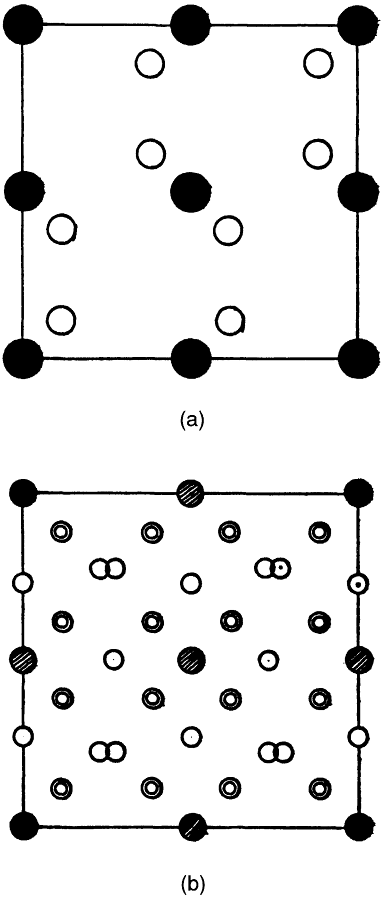

FIG. 10. (a) The atoms in a unit cell of FeS_2 as viewed from a cube face of the structure. (b) The Patterson function, derived solely from intensity data in the same projection.

8. X-RAY INSTRUMENTATION

8.1 Detectors

In recent years, x-ray diffraction techniques have undergone a major technological revolution due to remarkable advances in source and detector development. The development of synchrotron radiation sources has provided unprecedented new possibilities in the field of x-ray diffraction. It is analogous to the impact that the invention of the laser has had in the field of optics. Synchrotron radiation makes available fluxes up to six or more orders of magnitude greater than conventional x-ray tube sources with added ability to tune the photon energy continuously.

This source is also stroboscopically pulsed so that timing experiments of structural changes at the nanosecond level can be measured.

8.2 CCD Devices

Detectors have advanced well beyond Geiger and scintillation counters and film. (Film is still popular because of its relatively low cost, large active areas, and integrating properties.) Two major detector advances in the last decade have been the application of CCD detectors (similar to television cameras) and the development of storage phosphors. The CCD development involves the coupling of an x-ray image from a fluorescent screen to a CCD pixel array.

8.3 Storage Phosphors

Storage phosphors are a film replacement with some remarkable recording properties. The recording medium is a plastic sheet of sizes comparable to those used in standard medical applications of x-ray photography. An x-ray photon strikes the plastic that contains complex molecules of rare earth elements. These molecules are promoted to a metastable excited state with lifetimes of days. The plate is scanned by a narrow laser beam that stimulates the excited molecules to return their ground state emitting a distinctive pulse of visible light. These photons can be correlated to the laser beam position.

In many x-ray applications, storage phosphors have sensitivities that can be more than an order of magnitude greater than x-ray film and with dynamic ranges exceeding 10^4.

9. X-RAY DIFFRACTION METHODS

The diffraction techniques most often used in materials research and structure determination these days show the influence of the new synchrotron sources. Surprisingly, the more direct and simpler techniques survive.

9.1 Powder Diffraction

Most crystalline matter is available in polycrystalline form. Growing relatively large single crystals can be a difficult procedure, but in many cases the polycrystalline forms can be filed and ground to obtain large-volume samples of randomly oriented particles. When these are placed in a monochromatic x-ray beam, diffraction cones are generated and are explainable through a simple application of Bragg's law [Eq. (7)]. The cones intersect a film plane or detector slit and can be recorded with a single slit detector or a 2-D plate (CCD film or storage phosphor).

Because of the high collimation and monochromaticity of the source, each *hkl* reflection can be recorded and separated from its neighbors, allowing precise intensity measurements of each Bragg reflection. Highly sophisticated computer techniques can deconvolute the instrumental broadening. The final data can then be fitted to a model for the structure defined through the structure factor [Eq. (9)]. These new computer techniques allow one to solve moderately large unit cell structures without having to grow single crystals.

The powder method (also called Debye–Scherrer technique) is generally useful as an analytical tool for general-purpose material identification and is remarkable for its simplicity.

9.2 Laue Methods

The original discovery of the diffraction of x rays by von Laue in 1912 (Ewald, 1962) used the most primitive of all techniques: a collimated beam of polychromatic x rays strikes a crystal, and a film is placed behind the specimen to record the diffracted beams. For each Bragg plane (*hkl*) a particular wavelength is selected from the incident rays according to Bragg's law [Eq. (7)].

From the characteristic pattern one can identify the crystal orientation and make some comment about the crystal's symmetry. Until recently, the Laue method has not been an important tool for structure determination but is widely used as a general-purpose diagnostic. The development of synchrotron radiation has had a significant impact in increasing the technique's capabilities because of the intense broad-band fluxes it provides.

A conventional x-ray tube produces a few sharp characteristic lines and a weak polychromatic brehmsstrahlung background. The continuous beam from a synchrotron source is at least nine orders of magnitude higher.

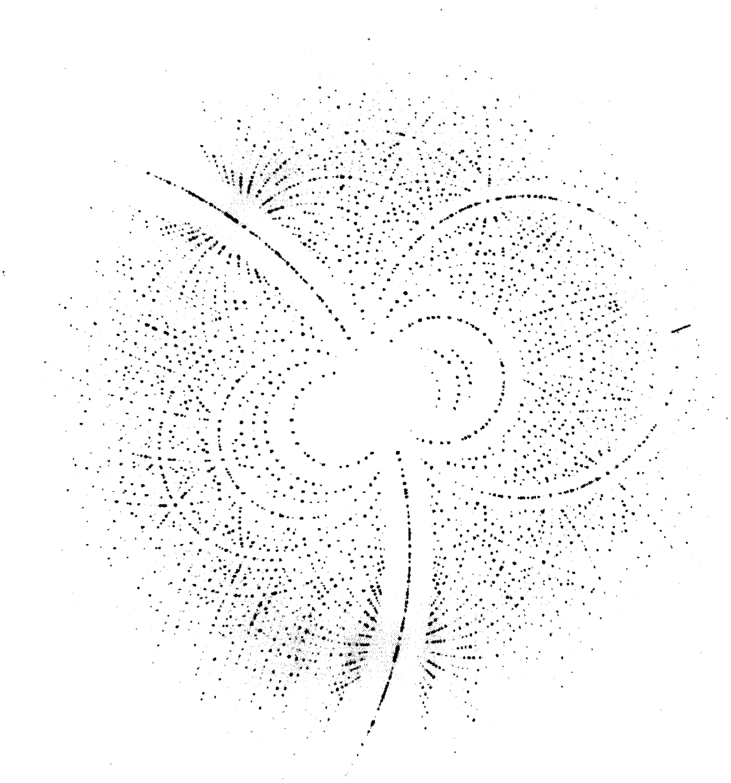

FIG. 11. The Laue pattern of an organic single crystal taken with synchrotron radiation.

Figure 11 shows a Laue pattern containing many thousands of spots. In appropriate cases, these exposures can be recorded in millisecond and even microsecond time scales. Attempts are under way to develop these techniques for studies of real time changes in structures promoted by external stimuli.

9.3 Single-Crystal Techniques

As we pointed out earlier, a highly parallel monochromatic x-ray beam incident on a crystal will be an unlikely candidate to produce a diffracted beam because of the geometrical restriction in satisfying Bragg's law. In the two techniques described above, these conditions are relaxed by allowing a wide range of wavelengths in the Laue case or a wide range of orientations with powder diffraction. For single crystals one relaxes the angular restrictions by rotating the crystal slowly so that different *hkl* planes can satisfy the Bragg diffraction condition. The pattern of the rhinovirus in Fig. 6 was taken in this manner. The crystal is exposed to a monochromatic synchrotron radiation beam as it oscillates over several degrees and a film is exposed. This is repeated over many crystal orientations to cover a large range of Miller planes *hkl.* For complex structures such as viruses and other large unit cell structures, the oscillation technique is the most simple and effective.

Works Cited

Batterman, B. W., Cole, H. (1964), "Dynamical Diffraction of X-Rays," *Rev. Mod. Phys.*, **36,** 681–717.

Compton, A. H., Allison, S. K. (1935), *X-Rays in Theory and Experiment,* New York: D. Van Nostrand Co., Chap. 3.

Ewald, P. P. (1962), *50 Years of X-Ray Diffraction,* Utrecht: N.V.A. Oosthoeks, Uitgeversmaatschappij.

James, R. W. (1950), *Optical Principles of the Diffraction of X-Rays,* London: G. Bell and Sons, Chap. 9.

Karle, J., Hauptman, H. (1953), *Acta. Cryst.* **6,** 131.

Morgan, J., Warren, B. E. (1938), *J. Chem. Phys.* **6,** 666.

Rossmann, M. G., Arnold, E., Erickson, J. W., Frankenberger, E. A., Griffith, J. P., Hecht, H.-J., Johnson, J. E., Kamer, G., Luo, M., Mosser, A. G., Rueckert, R. R., Sherry, B., Vriend, G. (1985), "Structure of a Human Common Cold Virus and Functional Relationship to Other Picornaviruses," *Nature* **317,** 145–153.

Tarasov, L. P., Warren, B. E. (1936), *J. Chem. Phys.* **4,** 236.

Further Reading

Batterman, B. W., Cole, H. (1964), "Dynamical Diffraction of X-rays by Perfect Crystals, *Rev. Mod. Phys.* **36,** 681.

Compton, A. H., Allison, S. K. (1935), *X-Rays in Theory and Experiment,* New York: D. Van Nostrand Co.

Cullity, B. D. (1956), *Elements of X-ray Diffraction,* Reading, MA: Addison-Wesley.

Ewald, P. P. (1962), *50 Years of X-Ray Diffraction,* Utrecht: N.V.A. Oosthoeks, Uitgeversmaatschappij.

James, R. W. (1950), *Optical Principles of the Diffraction of X-Rays,* London: G. Bell and Sons.

Schwartz, L. H., Cohen, J. B. (1977), *Diffraction from Materials,* New York: Academic.

Warren, B. E. (1969), *X-Ray Diffraction,* Reading, MA: Addison-Wesley.

X-RAY IMAGING, MEDICAL: TOMOGRAPHY AND RADIOGRAPHY

William R. Hendee, *Medical College of Wisconsin, Milwaukee, Wisconsin, U.S.A.*

Introduction 505

1. **Projection Radiography** 506

1.1 Nature and Production of X Rays 506

1.2 Interactions of X Rays 507

1.3 Image Production on Screens and Film 509

1.4 Fluoroscopy 509

1.5 Digital Imaging 510

1.6 Image Detail 512

2. **Image Interpretation** 513

3. **X-Ray Tomography** 514

3.1 Analog Tomography 514

3.2 Computed Tomography 514

4. **Image Networking** 516

5. **Applications** 518

5.1 Mammography 518

5.2 X-Ray Treatment Planning 520

6. **Conclusions** 521

Glossary 521

Works Cited 522

Further Reading 523

INTRODUCTION

Medicine is largely devoted to preventing disease or injury, diagnosing what is wrong when disease or injury may be present, and treating the patient to alleviate pain and, when possible, to cure the problem. Diagnosis is often a major challenge, as the patient's symptoms are frequently relatively unspecific and might be caused by a variety of conditions. Meeting this challenge usually requires conducting various tests, including laboratory analyses and imaging procedures. These tests are all products of technological advances in the practice of medicine that have occurred in the 20th century. A century ago, physicians were severely limited in the tools available to help them arrive at a correct diagnosis. Most diagnoses were made from a verbal description of the patient's symptoms, combined with the patient's history and supplemented by use of the physician's senses to detect unusual odors, visual appearances, abnormal sounds, physical sensations, and, at times, an odd taste. Although significant problems may have been present within the patient's body, they frequently provided few external clues, and the physician was guided toward a diagnosis largely by his (or, in a few cases, her) intuition and experience. Needless to say, the final diagnosis was uncertain or incorrect in more than a few cases.

A major advance in the physician's ability to detect and diagnose disease and injury occurred in 1895 with the discovery of x rays by the physicist Wilhelm Röntgen in Würzburg, Germany. Experimenting with a Crookes tube, Röntgen noted that crystals of barium platinocyanide on a bench across the laboratory gave off light when the tube was energized. He concluded that the fluorescence was caused by a new type of radiation, which he named x rays, that was emitted by the Crookes tube. He discovered that the radiation was reduced in intensity but not totally absorbed when an object was placed in its path and that an image of the object could be photographed by replacing the crystals with photographic film. If the object was an anatomical specimen such as the human hand, the image, termed a "projection radiograph," was a picture of the internal structure of the specimen. The first image, a radiograph of Frau Röntgen's hand, was produced a few days after x rays were discovered. When published a few months later, these results stimulated interest in many countries (Röntgen, 1895). In 1901, Röntgen was awarded the first Nobel prize in physics. By the turn of the century x rays were being widely employed for a variety of medical

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

conditions. X rays have been a very major contributor to the many advances in medicine that have occurred over the present century.

Today, x-ray imaging systems are used in hospitals, clinics, and physicians' offices worldwide for the detection and diagnosis of a wide spectrum of illnesses and injuries. In general, an x-ray imaging system consists of several components. An x-ray tube serves as the source of x rays that penetrate the patient and impinge on an image receptor such as a fluorescent screen or digital detector. The image is "captured" photographically or electronically and processed to accentuate anatomic features of interest in the patient. This image is interpreted by the physician to arrive at an opinion concerning the health status of the patient. The image may be stored physically or electronically for future referral.

1. PROJECTION RADIOGRAPHY

1.1 Nature and Production of X Rays

X rays are a form of electromagnetic radiation similar to infrared, visible, and ultraviolet light except higher in energy. They are widely used in medicine to produce images (radiographs) of virtually every region of the body. To achieve acceptable images, a source of x rays is required that

1. provides a sufficient number of x rays in a short interval of time,
2. allows the user to vary the x-ray energy,
3. produces x rays in a reproducible fashion, and
4. meets current standards of safety and operational efficiency.

The source almost universally employed for imaging is the x-ray tube. In this device x rays are released as high-speed electrons impinge on a target of a heavy metal such as tungsten.

The main components of a modern x-ray tube are shown in Fig. 1. As electrical current heats the tungsten filament in the cathode, electrons are boiled off (thermionic emission) and accelerated toward the target in the anode through a potential difference of from 20 000 to 150 000 V (20–150 kVp, where kVp denotes the peak kilovoltage across the tube). A focusing cup around the filament helps direct the electrons onto a small area (the focal spot) of an annular target that is embedded in a larger disk (the anode). X rays are released in all directions as the electrons interact in the target. The x rays are restricted by collimators outside the x-ray tube to confine the x-ray beam to the anatomic region of interest in the patient. Although intense bursts of x rays can be produced in this manner, most (>99%) of the electron energy is dissipated as heat in the target (Hendee and Ritenour, 1992). The anode rotates at several thousand revolutions per minute to spread the heat over a larger volume of target metal to keep the temperature of the metal from reaching its melting point.

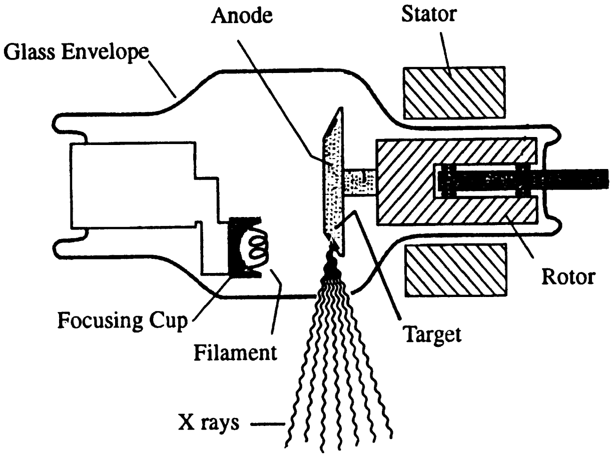

FIG. 1. Components of a modern diagnostic x-ray tube.

As the high-speed electrons penetrate the target metal, they interact with the electrons and nuclei of the target atoms through Coulombic (charged particle) interactions. During interactions with electrons, target atoms are ionized by ejection of inner-shell electrons, and x rays are released as the affected atoms regain stability by reordering of the electron shells. These x rays are termed "characteristic" x rays because their energy is characteristic of the type of metal employed as the x-ray target. For example, characteristic x rays of 58 and 67 keV are released from a tungsten target, where keV stands for kiloelectron volt, and the electron volt is a unit of energy equal to that gained by an electron accelerated through a potential difference of 1 V.

As electrons interact with nuclei of the target atoms, they are slowed down, and energy is released in the form of x rays termed bremsstrahlung (braking radiation). Bremsstrahlung provides a continuum of x-ray energies with a maximum value defined by the energy of the impinging electrons. For most x-ray beams used for medical imaging, bremsstrahlung is a more important contributor compared with characteristic x rays. A notable exception is the significant role played by characteristic x rays in beams employed for mammography (imaging of breast tissue), an application discussed later in this article.

To prevent the accelerated electrons from interacting with gas molecules on their way to the target, components of the x-ray tube are housed in a glass tube that is evacuated to very low pressure. The tube is contained within an oil-filled shielded housing that provides electrical insulation and prevents x rays from emerging in unwanted directions. During an x-ray exposure, the technique factors [tube voltage (kVp), tube current (mA), and exposure time (ms)] are monitored carefully to prevent overheating of the target and other components of the x-ray tube.

1.2 Interactions of X Rays

For a parallel, monoenergetic beam of x rays, the intensity I of x rays transmitted through a material of thickness z is

$$I = I_0 \exp\left[-\int \mu(z)dz\right],$$

where the line integral is taken over all tissues along the x-ray beam, I_0 is the intensity with no material present, and μ is the linear x-ray attenuation coefficient of the material, which may vary depending on the material's composition at each depth. If the material varies in composition laterally from one point to the next, the transmitted x rays also vary in intensity across the x-ray beam. By capturing the transmitted x-ray beam on a film or fluorescent screen, the pattern of variation in the material's composition is reflected in the image. As mentioned earlier, this image is termed a projection radiograph because it is formed through projection of a material into an image by transmitted x rays.

The linear attenuation coefficient μ is the sum of coefficients for various types of x-ray interactions that may occur in the material. For the energy range (20–150 keV) of x-ray energies used to produce medical images, two types of interactions are dominant. These interactions are the photoelectric effect, described by the photoelectric linear attenuation coefficient τ, and the Compton effect, described by the Compton linear attenuation coefficient σ. That is, $\mu = \tau + \sigma$. The dependence of these coefficients on x-ray energy is shown in Fig. 2 for interactions in human soft tissue. Additional information on x-ray interactions is available in INTERAC-

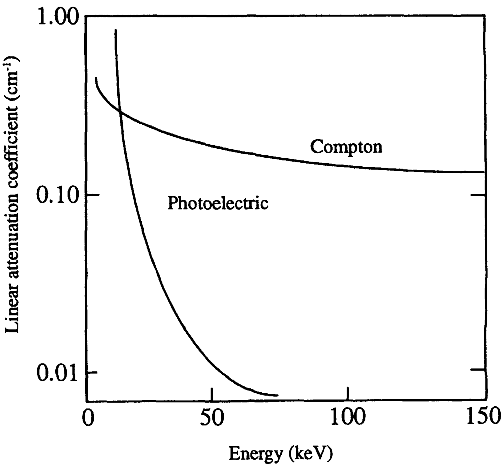

FIG. 2. The linear attenuation coefficient for x rays in soft tissue is the sum of two dominating contributions. Attenuation coefficients for these two contributions, Compton scattering and photoelectric absorption, are plotted as functions of the x-ray energy.

TION OF SOLIDS WITH PARTICLES AND RADIATION.

In a photoelectric interaction, the x ray interacts almost always with a tightly bound electron in an inner shell of an atom of the absorbing material. Because of the importance of tightly bound electrons, photoelectric interactions are strongly influenced by the atomic number Z, with the probability of photoelectric interaction per unit mass of material varying directly with Z^3 of the material. All of the energy of the x ray is transferred to the electron, which is ejected from the atom with a kinetic energy E_k of

$$E_k = E_{\text{x-ray}} - E_b,$$

where E_b is the binding energy of the ejected electron and $E_{\text{x-ray}}$ is the x-ray energy, which can also be expressed as $h\nu$, where h is Planck's constant (6.62×10^{-34} J-s) and ν is the frequency of the x ray. Since the x ray is only energy, it disappears during a photoelectric interaction and is essentially replaced by a high-speed electron and an ionized atom. Photoelectric interactions occur predominantly with low-energy x rays, and decrease in likelihood as $(h\nu)^{-3}$. A photoelectric interaction is depicted in Fig. 3(a).

A Compton interaction results in the scattering of an x ray in a new direction at reduced energy, rather than complete absorption of the x ray. Compton scattering produces most of the scattered radiation that degrades the quality of x-ray images. In this type of interaction, the x ray interacts with a loosely bound electron in an outer shell of an atom of the absorbing material. The electron is ejected with an energy equal to the difference in energy of the incoming and scattered x rays, with the assumption that the binding energy of the ejected electron is so small it can be neglected. Because the interaction occurs almost always with outer, loosely bound electrons, the Compton interaction is independent of the atomic number Z of the absorbing material. Compton interactions occur predominantly for the higher range of x-ray energies employed for medical imaging. The likelihood of these interactions decreases gradually with increasing x-ray energy. A Compton interaction is depicted in Fig. 3(b).

For both photoelectric and Compton interactions, the linear attenuation coefficient varies with the physical density (kg/m^3) of the absorbing material. Hence, the intensity of x rays transmitted through a material reflects the density of the material and, to varying degrees, the atomic number of the material, depending on the energy of the x rays and, therefore, the degree to which photoelectric interactions are responsible for reducing the intensity. By choosing the x-ray energy through selection of the voltage across the x-ray tube, an x-ray beam may yield a projected radiograph that accentuates or suppresses differences in atomic number among different constituents of an absorbing material. This characteristic is used to great advantage in medical imaging. For example, low-energy x rays may be used to amplify small differences in tissue composition that may be important for the detection of subtle abnormalities in soft tissue. Breast imaging is one example where low-energy x rays are used to amplify subtle soft-tissue differences. On the other hand, high-energy x rays may be needed to suppress relatively large differences in tissue composition so that more subtle features of tissue are not hidden by "shadows" cast by strongly absorbing struc-

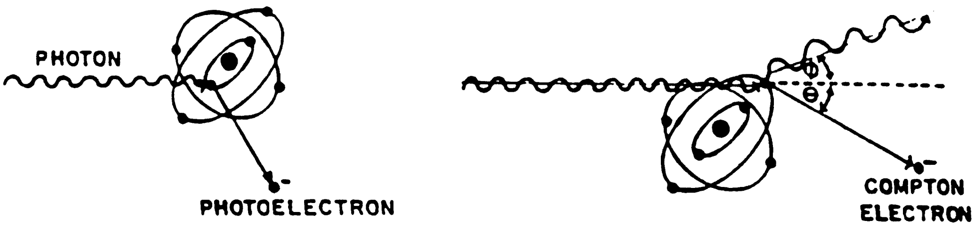

(a) (b)

FIG. 3. (a) Photoelectric interaction; (b) Compton scattering.

tures. Chest imaging is one example where high energy x rays are used to penetrate the ribs and other strongly absorbing structures to reveal more subtle features of the underlying lung parenchyma. For x rays at any energy, those that are not transmitted through the patient are absorbed or scattered in the patient. These processes result in energy deposition in the tissues of the patient, where the energy deposited per unit mass of tissue is known as the "absorbed dose."

1.3 Image Production on Screens and Film

Image receptors employed for x-ray imaging include film, usually used in combination with fluorescent screens; fluorescent screens contained within an image intensifier for fluoroscopic imaging of dynamic processes; and solid-state detectors such as semiconductor detectors and photostimulatable phosphorescent plates for generation of electronic signals. Photographic emulsions are sensitive to x rays, and an x-ray image of exceptionally high quality can be formed by direct exposure of photographic film to an x-ray beam. This approach to medical imaging is occasionally used when exquisite detail is required in the image, as might be the case when hairline fractures in the bones of extremities are suspected. However, film by itself is an inefficient absorber of x rays, and for most applications it would require long exposure times that would yield images with unacceptable amounts of motion blurring. In almost all cases, permanent x-ray images are produced by sandwiching the film between "intensifying screens" that absorb x rays with reasonable efficiency (up to 50%) and emit visible light that in turn exposes the photographic film.

The first intensifying screens were developed by Edison and contained calcium tungstate ($CaWO_4$) as the active light-emitting ingredient. These screens remained popular until the 1970s, when they were replaced by screens with phosphors containing a rare-earth element such as gadolinium, lanthanum, or yttrium complexed with oxysulfide or oxybromide crystals embedded in a plastic matrix. Compared with $CaWO_4$ screens, rare-earth screens are better absorbers of x rays, and some yield more light for each x ray absorbed. Screens with thicker emulsions of light-emitting material are said to be "faster" (i.e., they have a higher "speed") because they require less exposure to release enough light to expose the film. This advantage is not without cost, however, because thinner screens (so-called detail screens) yield images with improved spatial resolution (less geometrical blurring). The choice of intensifying screens is a compromise between increased geometric blurring with fast screens and increased motion blurring with detail screens.

Most x-ray images are produced by sandwiching an x-ray film between intensifying screens in an imaging cassette that is tightly sealed (sometimes even evacuated) to produce intimate contact between the film and screens, thereby reducing geometric blurring. Cassettes are available to hold film of different sizes from 4 × 4 inches to 14 × 17 inches used for chest radiography. Once the film has been exposed to x rays or light from the intensifying screens, it contains a "latent image" that can be made visible by chemical processing. In the visible image, darker areas represent regions penetrated by a larger number of x rays, whereas lighter areas depict regions where more x rays are absorbed. In many cases, the lighter areas reveal bony structures because the higher density and atomic number of bone results in greater absorption of the impinging x rays.

1.4 Fluoroscopy

Projection radiography yields exquisite images that represent a "snapshot in time" of the patient's anatomy. In many cases these images are sufficient to detect an abnormal condition and to arrive at a diagnosis of the cause of the abnormality. But this is not true in all cases. Sometimes a continuous image is required that yields information about how the region of anatomy is changing or functioning over the course of a few seconds or minutes. The technique that yields such a continuous image is referred to as "fluoroscopy" because it captures the image instantaneously on a fluorescent screen and displays it in real time to the viewer.

Early applications of fluoroscopy employed a fluorescent screen that was viewed directly by the physician. The image was so dim, however, that the observer's vision had to be "dark-adapted," and even then only gross features could be detected. The viewing conditions resembled those present in a darkened movie theater. Although this approach was highly unsatisfactory, it was the only method available for fluoroscopy until the 1950s when the image intensifier was developed.

An image intensifier is depicted in Fig. 4. X rays impinging on the CsI input screen produce light that dislodges electrons from the photocathode. These electrons are accelerated toward the anode and are focused to pass through the aperture in this structure and onto the output screen. Here they produce a minified image that is thousands of times brighter than the image on the input screen. This image may be captured optically and projected through a system of lenses and mirrors for direct viewing by the observer. More often it is projected into a television camera and converted into an electrical signal that can be transmitted to a remote television monitor for reconfiguration and viewing. Since the image is converted from light to an electrical signal during this process, it can be converted easily to a digital image for integration into a digital imaging network.

Fluoroscopic methods are an essential part of medical x-ray imaging. They are widely employed in many types of x-ray studies, including those of the gastrointestinal tract and angiographic studies of the central and peripheral circulatory system. They also are used for image-guided therapy (e.g., coronary angioplasty, interventional neuroradiology, and minimally invasive surgery), a therapeutic approach that is growing rapidly in application because of its reduced morbidity and, often, lower cost compared with alternative therapeutic approaches.

1.5 Digital Imaging

The combination of intensifying screens and photographic film has many advantages for capturing and recording x-ray images. It is simple, portable, and inexpensive, and it yields images with excellent anatomic detail. This approach, however, has some disadvantages as well. It is limited to a rather narrow range of acceptable exposures and offers little flexibility for image enhancement or data compression. Film images are bulky to store and easy to misplace, and they must be transported physically from one location to another. Replacing this analog method of image production with an approach that yields digital images would eliminate these disadvantages. Hence, considerable interest in digital approaches to projection radiography has surfaced in recent years. In all likelihood, digital methods will capture an increasing fraction of the projection radiography market over the next several years.

In the digital radiographic unit shown in Fig. 5, the x-ray source and receptor are computer controlled to yield digital images that are displayed in real time on a video screen. Digital images can be stored on magnetic media or optical disk, and a film-writ-

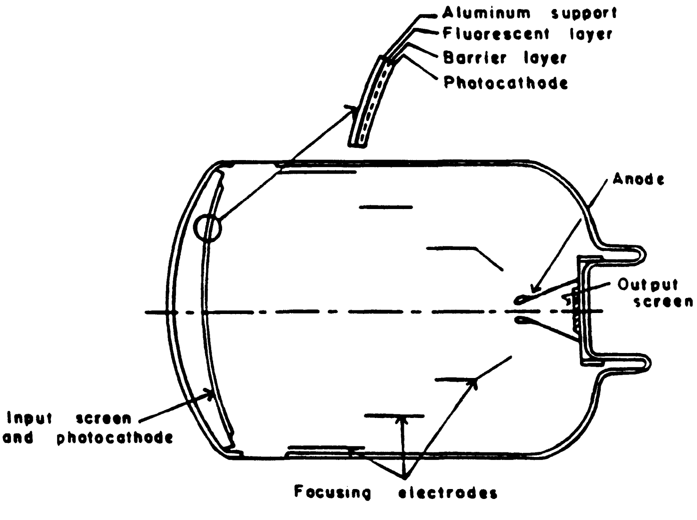

FIG. 4. X-ray image intensifier.

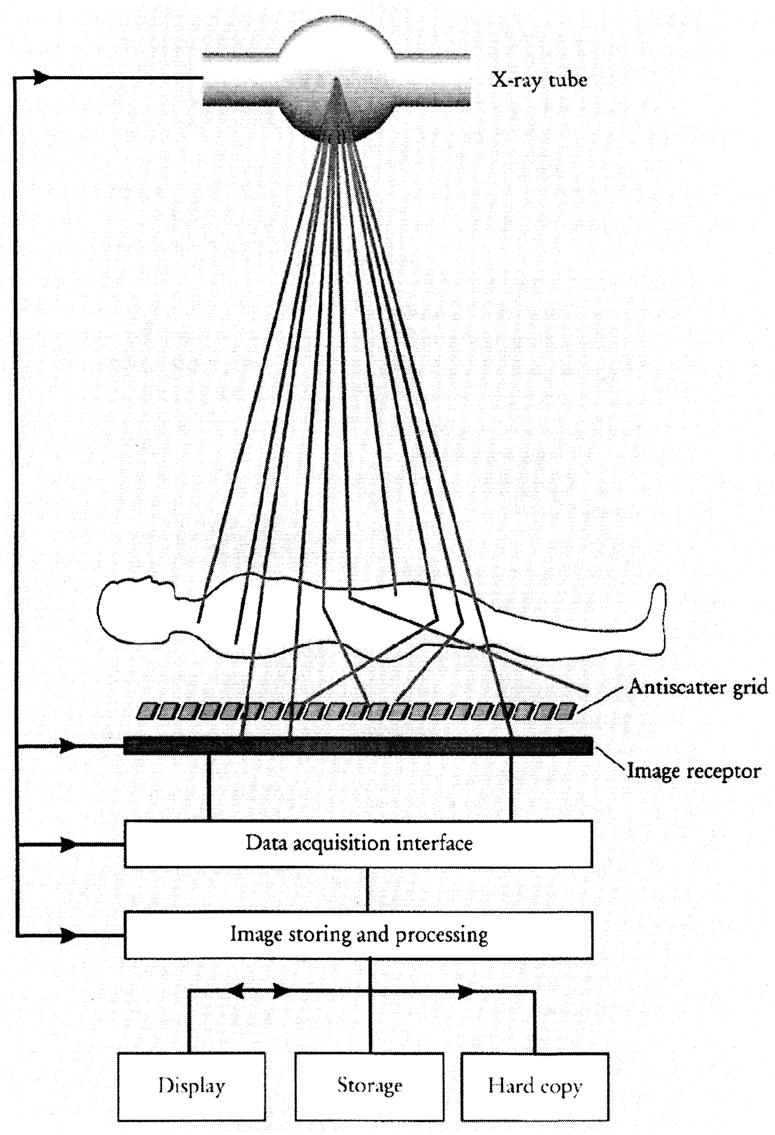

FIG. 5. Digital x-ray imaging system creates images that can be displayed in real time. X rays that traverse the patient unscattered are recorded by the digital image receptor, and the data are passed on for processing and display, storage and, when desired, production of hard-copy images. The antiscatter grid prevents scattered x rays from reaching the receptor. (Used with permission from Hendee, 1995.)

ing device can be used to produce permanent analog copies of the images if they are needed. Two types of image receptors currently are used in digital radiography. One type records the entire image simultaneously, while the other acquires a complete image by scanning the x-ray beam and receptor in synchrony across the patient, a technique known as scan-projection radiography. Various methods are employed for digital data acquisition, including the use of image intensifiers with frame-grabbing television cameras, ionographic chambers, solid-state detector arrays, and photostimulated-luminescence plates (Hendee, 1995a). Of these methods, all but photostimulated plates are impractical today for routine projection radiography. Commercially available stimulated-luminescence plates employ a phosphor that absorbs x rays and reradiates their energy as visible light when scanned with a laser beam. The emitted light is detected by a photomultiplier tube and digitized with an 8- or 16-bit analog-to-digital converter. The digitized image is displayed on a video monitor and can be stored for later recall.

Conventional projection radiographic films can be digitized with a scanning microdensitometer to yield digital radiographs. As with any other digital image, a digital radiograph can be transmitted electronically to distant sites for interpretation and consultation, a technique known as teleradiology. Digital radiographs can be integrated with other digital images into an electronic image network for correlation with other digital images such as those produced by computed tomographic techniques, as discussed later.

Digital x-ray images offer several advantages over screen and film radiographs. They can be computer-processed to highlight selected features (edge sharpening, contrast enhancement, and noise smoothing), and the data can be compressed for efficient transmission electronically to distant sites. They require less storage space, and copies can be easily presented for viewing without risk of misplacing the original data. Digital x-ray images sometimes require less radiation exposure compared with analog projection radiographs. Finally, computer subtraction of digital images taken at different times can often reveal changes that are not visible in the original images. Disadvantages of digital radiographs include the cost of the equipment required for their production compared with simple film-screen techniques and possible losses in image detail occasioned by the finite number of picture elements (pixels) used for image display.

1.6 Image Detail

The level of detail visible in an x-ray image is a compromise among four kinds of resolution: spatial, contrast, temporal, and statistical. Spatial resolution is related to the geometry of the image-forming process and the resolving capacity of the imaging device (often called the “image receptor”). Contrast resolution is affected by the differences in the absorption and scattering characteristics of the various tissues depicted in the image and by the impact of scattered x rays on the receptor. Temporal resolution refers to the blurring caused by the patient’s voluntary or involuntary movement (e.g., breathing, peristalsis, cardiac motion, or vascular pulsatile motion) during exposure to x rays. Statistical resolution is related to the number of x rays absorbed in the receptor to produce the image; the noise is inversely proportional to the square root of the number of x rays used to form the image. In most cases, improving any one of these resolution factors degrades one or more of the others. The compromise among them represents a balance so that no single factor strongly dominates the degradation of detail in the image.

X rays scattered from the patient during Compton interactions can be a major degrading influence on the visibility of detail in the image. Usually this influence is controlled by inserting a mechanical grid between the patient and the image receptor. The grid contains thin strips of lead or other high-*Z* material oriented parallel to the direction of impinging primary x rays. These x rays pass through the grid, while x rays that are scattered away from the direction of the primary radiation strike the grid strips and are absorbed.

The challenge of x-ray imaging is to capture subtle differences (i.e., modulation) in tissue composition by passing an x-ray beam through the tissue, and to transfer these captured differences onto an image receptor where they can be seen by the human eye. That is, the purpose of medical imaging is to transfer modulation in the tissue (object) into modulation in the image. The degree to which this modulation is transferred faithfully by an imaging system is described as its modulation transfer function (MTF). With most imaging systems, object modulation can be transferred to the image with high fidelity if the modulation is of relatively low frequency (i.e., the object is composed of relatively large structures that are well separated from adjacent structures). For smaller structures that are closer together, however, the integrity of modulation transfer may be less exact. Hence, the MTF is frequency dependent, as shown in Fig. 6 for two imaging systems. In this illustration, the fidelity of system 1 is superior to that of system 2, as evidenced by the higher values of MTF at higher frequencies.

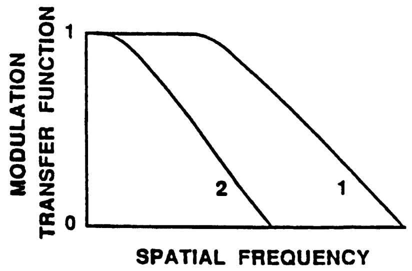

FIG. 6. Modulation transfer function (MTF) for two x-ray imaging systems, with system 1 providing higher spatial fidelity compared with system 2.

The MTF is a relatively simple (and somewhat simplistic) measure of the performance of an imaging system. To determine the MTF, one obtains an x-ray image of a very thin transparent slit in an otherwise opaque object. The image of the slit is scanned with a microdensitometer to determine the spread of the slit in the image. The resulting curve is described by a "line spread function" that can be characterized in frequency space through the process of Fourier transformation. The result is the MTF, expressed as a function of spatial frequency, as demonstrated in Fig. 6. The spatial frequency, expressed in units such as line pairs per millimeter or cycles per millimeter, is a measure of the frequency of variation per unit distance. It may be compared with the temporal frequency, expressed in units such as cycles per second or hertz, which is a measure of the frequency of variation per unit time.

2. IMAGE INTERPRETATION

Interpretation of medical images is a complex process that is as much art as science. It requires years of training, much of which is provided in a "master and apprentice" style since it has not been possible to model the interpretive process quantitatively. On the other hand, computer programs have been written to perform selected analyses of the patterns of information present in certain types of images such as chest radiographs and mammograms. By comparing these analyses to databases compiled from a large number of normal and abnormal images, a tentative diagnosis can be arrived at for presentation to the interpreter for confirmation. Such a decision-support system is referred to as "computer-aided diagnosis." Studies of this approach have shown that it increases the accuracy of diagnosis of the typical radiologist (Doi *et al.,* 1993).

Computer-aided diagnosis is a far step from the use of expert artificial-intelligence systems to provide a primary diagnosis from medical images without human intervention. Such systems would essentially have to mimic the thought processes of physicians interpreting images, including the incorporation of ancillary information such as the patient's history and data acquired by physical examination. These thought processes are largely intuitive, acquired through apprenticeship training and experience, and are almost impossible to articulate clearly. Replacement of the human interpreter of images with a computer algorithm for image interpretation would be a major challenge.

It is important to recognize, however, that an expert system for image interpretation does not have to be perfect. Instead, it must function only at a level at least comparable to the accuracy of the average radiologist. With this criterion in mind, exploration is occurring of expert systems for image interpretation in medicine. Some of these explorations employ "neural networks," algorithms that use feedback information from humans to build decision rules based on the successes and failures of past decisions. Medical imaging is still a long way from automated diagnosis, as contrasted with computer-aided diagnosis, which relies ultimately on human confirmation or rejection of a proposed diagnosis. Still, progress in expert systems in other imaging disciplines, including military surveillance and space exploration, has been spectacular over the past two decades, and the application of such systems to medical imaging remains promising.

The ability of a computer to transcribe voice dictation of radiology reports automatically would be a major advance in getting the reports quickly to physicians who are caring for ill and injured patients. This approach would also reduce the costs associated with manual transcription of dictated reports, especially since radiologists dictate over 130 million reports annually in the United States alone. Speech-processing computer-based systems are available commercially, and their application to radiology reporting is under exploration in a number of imaging centers. Voice patterns are as distinctive as fingerprints, and automatic dictation could enhance the security and reduce unauthorized access of patient information associated with medical imaging procedures.

3. X-RAY TOMOGRAPHY

In a conventional radiograph, all structures within a three-dimensional volume of tissue exposed to x rays are projected "in focus" onto the two-dimensional image receptor. As a result, an anatomic structure of interest is often obscured in the image by the shadow of objects that lie above or below it in the patient. Removing the presence of these obscuring shadows often helps reveal the structure of interest. Two approaches, analog tomography and computed tomography, are available for removing the image of obscuring structures. As computed tomography has grown over the past two decades, analog tomography has diminished in popularity.

3.1 Analog Tomography

Analog tomographic images are produced by keeping structures in focus in a selected image plane in the patient, while blurring the image of structures above and below the imaging plane. The blurring is accomplished by moving the x-ray tube and image receptor during exposure through an angle with a pivot (fulcrum) in the image plane. In actuality, the image plane has a finite thickness (and is more properly referred to as an image section) that can be reduced by increasing the angle of motion. The location of the section within the patient may be changed by positioning the patient appropriately with respect to the pivot of the tomographic angle. Although the shadows of overlying and underlying structures are blurred, they are not completely removed from the image, and they frequently appear as "ghostlike" artifacts that interfere with the clarity of structures of interest. In addition, analog tomography has the limitations that conventional image receptors are unable to resolve small differences (1–2%) in the intensity of incident radiation, and the wide-area x-ray beams employed produced considerable scattered radiation that interferes with the visualization of subtle detail in the image. Analog tomography is still used for certain applications such as a supplement to chest radiography. For other tomographic needs, however, it has been replaced by computed tomography because this approach reduces the limitations often encountered with conventional tomography.

3.2 Computed Tomography

The mathematical model used in computed tomography (CT) was described in 1917 in a remarkable work by the Austrian mathematician Radon (Radon, 1917). The model was first used in applications of image reconstruction to radio astronomy, electron microscopy, and optics. In 1972, the first commercially available x-ray transmission CT unit for studies of human anatomy was announced. In 1979, the developer of this unit, G. Hounsfield (1973), shared the Nobel Prize in Medicine with a physicist, A. Cormack (1963), who did early work on image reconstruction. The first commercial CT unit employed a narrow beam of x rays scanned across the patient in synchrony with a scintillation detector moving on the opposite side of the patient. The intensity I of x rays measured by the detector is given by

$$I = I_0 \exp\left(-\sum \mu_i x_i\right),$$

where μ_i represents the linear attenuation coefficient of each of several structures in the path of the narrow x-ray beam and x_i represents the thickness of each of the structures.

With a single measurement of x-ray transmission, the separate attenuation coefficients cannot be determined. With multiple transmission measurements obtained at different orientations, however, the separate coefficients can be distinguished to yield a two-dimensional distribution of linear attenuation coefficients across the imaging section defined by the thickness of the scanning x-ray beam. This distribution can be mapped as CT numbers determined from the formula

$$\text{CT number} = 1000(\mu - \mu_w)/\mu_w,$$

where μ is the coefficient at a specific location in the section and μ_w is the linear attenuation coefficient of water for the x-ray energy employed for CT scanning. From this expression it can be appreciated that the CT number for water is zero and the CT number for air, with a very small value of μ, is close to -1000. The CT number would be about $+1000$ for a structure such as compact bone or colon containing barium contrast medium, because the physical density would be almost twice that of water. The CT numbers can be presented as various shades of gray to provide a visual display of the distribution of tissues across the image section. This display is referred to as a CT image.

The x-ray tube and scintillation detector of early CT units scanned the patient along a linear path perpendicular to the axis of rotation of tube and detector. Measurements of x-ray transmission were obtained every 2.5 mm over the linear path of 40 cm length to yield a total of 160 transmission measurements. Then the angular positions of the x-ray tube and detector were advanced 1°, and another scan of 160 measurements was obtained. This process was continued over 180° to provide 28 800 transmission measurements. These data were then transmitted to a computer equipped with a mathematical package for reconstructing a cross-sectional image of the scanned section. Initially, reconstructed images were computed by a time-consuming iterative mathematical model. More recently, faster algorithms using a convolution (filtered back-projection) model for image reconstruction have been employed (Kak and Slaney, 1988; HOLOGRAPHY, ACOUSTICAL). These algorithms rely on a Fourier transformation of x-ray projection data into frequency space, permitting use of ramp and cutoff-frequency filters to improve image quality and enhance subtle features in the image (see FOURIER AND OTHER MATHEMATICAL TRANSFORMS).

The combination of translational and rotational motion employed in early CT units yielded satisfactory CT images, but several minutes were required for data acquisition, limiting the method to studies of structures such as the head and extremities where voluntary and involuntary motion of the patient is not a problem. There was great interest in applying CT to thoracic and abdominal studies, but the scan time had to be decreased from minutes to seconds if these areas were to yield satisfactory images. Major efforts to reduce scan time led to the development of successive generations of CT scanners as depicted in Fig. 7. The initial translate-rotate scanner (first generation) is shown in Fig. 7(a), whereas an array of detectors (second generation) that tracks the linear motion of the x-ray beam is shown in Fig. 7(b). A purely rotational design (third generation) for CT scanning is depicted in Fig. 7(c), and a fourth-generation design is shown in Fig. 7(d) in which an x-ray tube rotates around the patient within a stationary ring of x-ray detectors. With the rotational scanners, data for a single section can be acquired in just a few seconds. By combining these geometries with the patient gantry moving continuously along its long axis, cross-sectional images of many slices can be obtained in minimum time, a process known as "spiral" or "helical" scanning (Heath *et al.*, 1995). Finally, images of vertical (sagittal or coronal) slices through the body can be produced by compiling arrays of attenuation coefficients along axes parallel to the patient's long axis.

Although the evolution of CT scanning geometries shortened the time to acquire images, none of the geometries permitted data acquisition in a time (less than 0.1 s) short enough to capture images of the heart and other blood-perfused organs without significant degradation caused by motion. For CT images of these organs, a way is needed to produce x-ray projection data without mechanical motion in the scanner. A scanner of this sort is known as a fifth-generation scanner (Hendee, 1983). It employs an electron-beam gun that generates x-ray beams in different directions by scanning over a stationary concave metal target. The resulting scan times are as fast as 50–100 ms.

CT images are most commonly displayed as cross-sectional representations of patient anatomy. By examining continuous cross-sectional images, radiologists can intuitively integrate the imaging information into a three-dimensional model of anatomy. This integration can also be accomplished physically by recognizing that a data set of contiguous two-dimensional sections in computer memory is in actuality a three-dimensional database. By mathematically compiling data across planes that cut across this database in preferred directions, coronal and sagittal images, as well as images at oblique

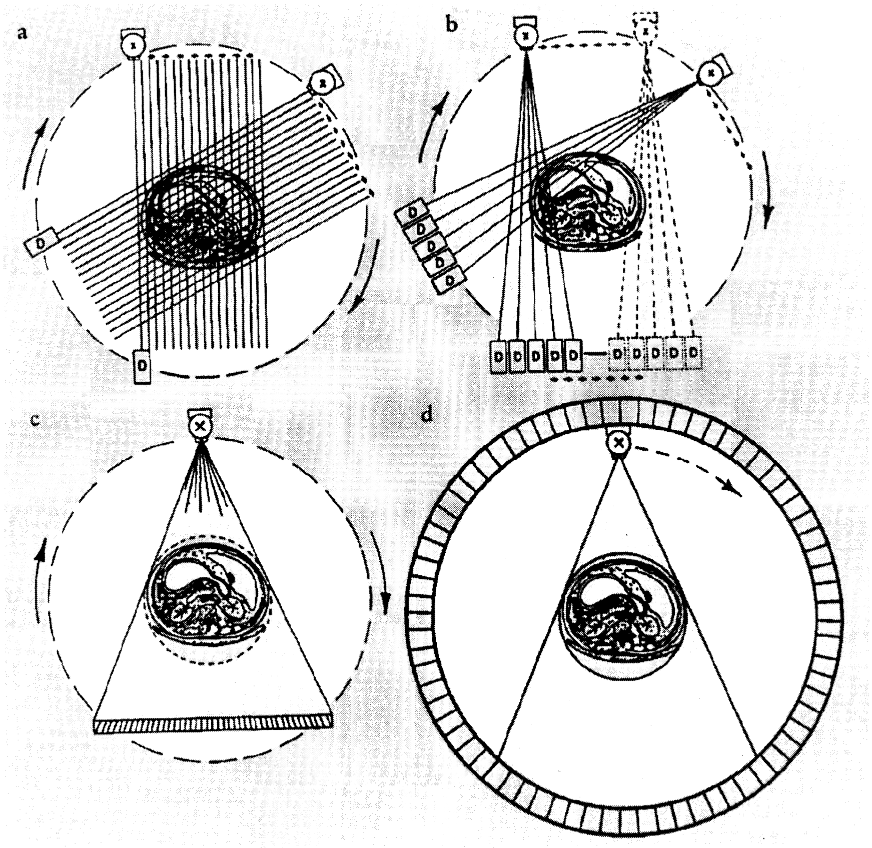

FIG. 7. Evolution of geometries of x-ray CT scanners. (a) First-generation scanner, in which the pencil x-ray beam is both translated and rotated to cover the region of the body being imaged; (b) second-generation scanner, with a diverging fan beam and detector array that are translated and rotated; (c) third-generation scanner, with a fan-beam source that rotates around the body, together with its detector bank; (d) fourth-generation scanner, with a rotating fan-beam source and a stationary ring of detectors.

angles, can be produced. Furthermore, the three-dimensional database can be analyzed to yield images that give the distinct impression of being three-dimensional. They also can be windowed to yield images of specific ranges of CT numbers corresponding to selected types of tissue (Fishman *et al.*, 1989). An example of a three-dimensional CT image is shown in Fig. 8.

4. IMAGE NETWORKING

Medical imaging includes several systems, including x-ray computed tomography and magnetic resonance imaging, that present imaging data in a digital format. Linking these systems electronically in a picture archiving and communications system (PACS) [referred to as an image management, archiving and communication system (IMACS) when connected into other information networks such as the radiology and hospital information systems] offers several advantages, including these: images can be stored electronically in a central location, images from different systems can be called up for display and comparison in a viewing center, and images can be transmitted to remote locations for viewing by specialists without loss of the central file of data. A PACS has the potential of expediting diagnoses, improving diagnostic accuracy, enhancing information to other physicians, and eliminating misplaced and lost films. With the availability of digital x-ray imaging methods, PACS can be expanded

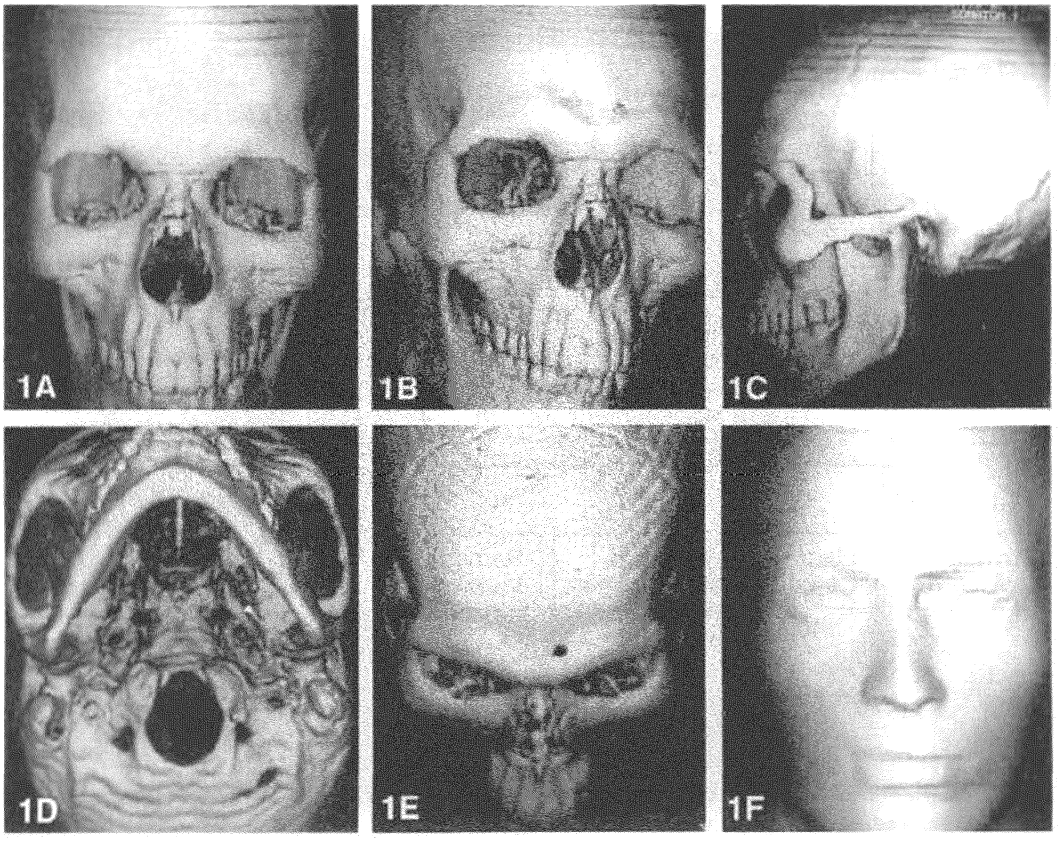

FIG. 8. 1.3-D standard views: (a) full frontal face, (b) right and left 45° oblique, (c) right and left lateral, (d) worm's view, and (e) bird's-eye view. (f) 3-D images can also be displayed in soft tissue format. (Used with permission from Fenlon and Breatnach, 1995.)

to encompass essentially all imaging systems employed in medicine. Many leaders in medical imaging believe that a fully integrated PACS represents the radiology department of the future. Others are convinced that the expense of converting fully to a PACS (several million dollars for a typical imaging facility) will prevent most departments from converting completely to PACS, at least in the near future. Instead, they believe their conversion will occur incrementally, with film remaining as the preferred detector for some applications (e.g., mammography and chest imaging) for some time to come.

In the common star network configuration depicted in Fig. 9 (Wolbarst, 1993), a computer and PACS management system can have access to images from any one of several peripheral digital imaging systems and can transmit these images to any other peripheral site. Some of the peripheral sites consist of stand-alone imaging systems complete with their own image and display workstation, image archive, and image-processing computer. Other peripheral sites are "dumb terminals" that can receive transmitted images for viewing but lack the capacity for image modification. The latter sites are typically placed in clinical locations, such as critical-care units, emergency rooms, and outpatient clinics, where images are important to decisions about the care of patients.

Within a hospital or clinic, imaging data are usually transmitted on-line by copper or optical-fiber cable. For communication over greater distances, modems are used with direct or acoustic coupling to transmit image data over standard telephone lines, typically at rates of 14.4 or 28.8 kilobits per second

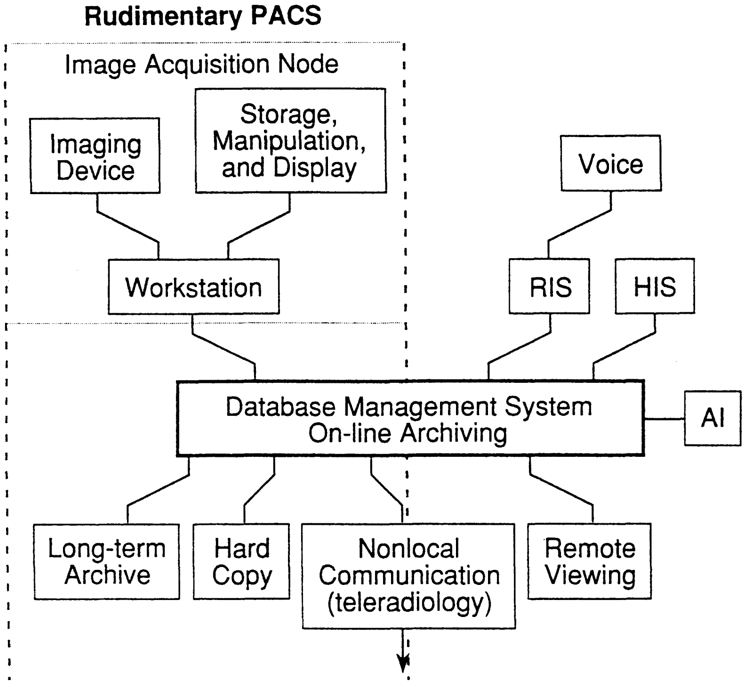
Rudimentary PACS
Image Acquisition Node
Imaging Device
Storage, Manipulation, and Display
Workstation
Voice
RIS
HIS
Database Management System On-line Archiving
AI
Long-term Archive
Hard Copy
Nonlocal Communication (teleradiology)
Remote Viewing

FIG. 9. Actual PACS designed according to a star configuration. This image-acquisition node represents one of several peripheral nodes, and the nodes on the upper right depict integration of the PACS with the radiology information system (RIS), hospital information system (HIS), voice transcription services, and computer programs for diagnostic decision making (AI). (Used with permission from Wolbarst, 1993.)

(kbps). Higher transmission rates up to 2 megabits per second (Mbps) can be achieved with special "T1" lines that can be leased from the local telephone company. Also, microwave carriers can be used either directly for line-of-sight transmission or by satellite transmission for longer distances. The replacement of standard "twisted pair" telephone lines with fiber-optic cabling in many locations in the United States greatly enhances the transmission of large databases such as those employed for medical imaging. The transmission of medical images over long distances is termed "teleradiology."

5. APPLICATIONS

The use of x rays is so ubiquitous in medicine that almost everyone has had an x-ray examination. Most examinations are for routine purposes such as to detect dental caries, confirm that a fractured bone has been set properly, provide a routine examination of the chest to rule out pneumonia, or perform a barium swallow or enema to study the integrity of the gastrointestinal tract. Many more applications of x rays to medicine exist, however. Even though these applications may be encountered less frequently, they are critically important to the patients who need them. Two specific applications are described here as examples of the many interesting challenges presented to physicists and engineers working in the discipline of medical physics.

5.1 Mammography

One of every eight women will develop breast cancer in her lifetime, and many will die as a result. The most effective measure against this disease is periodic screening of the breasts with x rays for detection of cancer before it has a chance to spread. Many voluntary and professional organizations recommend annual screening with x rays for women 50 years of age and older, and some recommend screening every 1 to 2 years for women between 40 and 50 years of age. X-ray imaging of the breasts is referred to as mammography.

A mammographic x-ray image must be able to reveal very subtle differences in the composition of soft tissues of the breast, as well as the presence of tiny (<100 μm) specks of calcium hydroxyapatite (calcifications) that suggest the presence of malignant cells. That is, mammographic images must bring out very small variations in contrast and must also yield exquisite detail within a background of relatively low noise. These images must depict the relatively thick tissues near the chest wall without losing detail in thinner regions of the breast. These demands place inordinate requirements on systems for production of satisfactory mammographic images. They also necessitate periodic evaluation of image quality and radiation dose by a physicist experienced in inspection and quality control of mammographic units (American College of Radiology, 1993).

Subject contrast is enhanced in mammography by the use of low-energy x rays that interact strongly by photoelectric processes. Reliance on these interactions yields the greatest difference in transmission among structures that are similar, although not identical, in chemical composition. The use of low-energy x rays also aids in the identification of calcifications. In most cases today, mammography is performed with special x-ray units that employ an x-ray tube operated at 20–35 kVp and a film and single-screen image receptor designed specifically for mammography. The x-ray tube has a molybdenum target that produces characteristic *K* x rays of 17.4 and 19.8 keV superimposed on a bremsstrahlung x-ray spectrum with a maximum energy equal numerically to the peak voltage expressed as kVp. The x-ray beam is filtered with a thin sheet of molybdenum that is relatively transparent to the characteristic x rays but absorbs higher-energy x rays with high efficiency because of the existence of the *K* absorption edge above 20 keV. A typical x-ray energy spectrum suitable for mammography is depicted by the solid line in Fig. 10. The dashed line in this illustration depicts the x-ray spectrum for a nondedicated x-ray unit with a tungsten target.

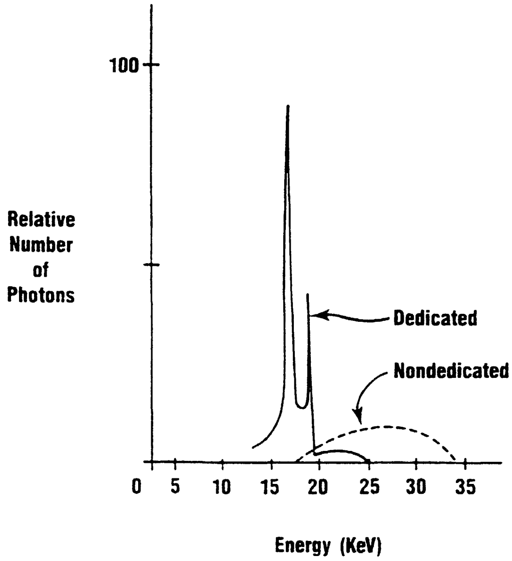

FIG. 10. X-ray spectra (normalized to equal area) for a molybdenum-target dedicated x-ray unit and a tungsten-target nondedicated x-ray unit for mammography. The dedicated unit employs a molybdenum (Mo) target, a 0.03-mm Mo filter, and a tube voltage of 25 kVp; the nondedicated unit uses a tungsten target, aluminum filtration, and a tube voltage of 35 kVp. (Used with permission from Hendee, 1995b.)

Screen–film mammography is an exquisitely sensitive and useful imaging method. Still, it may be possible to improve breast imaging by replacing screen–film receptors with digital detectors. Unlike film, digital systems acquire, display, and store images independently, so that each of these functions can be optimized without interfering with the others. Also unlike film, the response of digital detectors varies linearly over a wide range of exposures. In addition, digital images can be "windowed" to enhance the contrast of specific regions of tissue. Both of these attributes of digital detectors improve the contrast sensitivity of digital mammographic methods compared with analog screen–film methods. These advantages may well outweigh the disadvantage of a slightly lower spatial resolution with digital methods. Although digital detectors have high potential, the practicality, dependability, and low cost of film continue to make it the preferred detector for mammography in clinical settings.

Radiation can induce cancer as well as detect it. (Radiation can also be used to treat cancer, which is the medical specialty of radiation oncology.) In the past, concern has been expressed that screening large numbers of asymptomatic women for breast cancer could induce as many cancers as are cured through the process of early detection. This concern was primarily a result of the relatively high doses delivered to patients with earlier, less sensitive methods of image production. With today's ultrasensitive screen–film combinations, together with the use of low-energy, highly filtered x-ray beams, the induction of cancer is much less likely. The overwhelming balance of benefit over risk in screening programs for older women has largely dispelled concern over the possible induction of cancer in the screened women (Kimme-Smith, 1992). Nevertheless, efforts to reduce radiation dose further in mammography should be continued until each examination uses no more than the essential number of x rays to reveal underlying pathology.

5.2 X-Ray Treatment Planning

Cancer is characterized by an abnormally rapid proliferation of cells. The successful treatment of cancer with x rays requires that the cancer is localized and that that cancer and its microscopic extensions into normal tissue receive a dose of radiation sufficient to kill the cancer cells. At the same time the dose to nearby normal tissues must be kept low enough to avoid serious complications. The balance between these doses is depicted in Fig. 11 (Wolbarst, 1988). The probability $P_{\text{tumor}}(D)$ that the cancer will be controlled by killing tumor cells increases with D, the radiation dose. The probability $P_{\text{organ}}(D)$ that life-threatening complications from radiation damage can be avoided is, of course, a decreasing function of D. The probability $P_{\text{patient}}(D)$ that the patient will survive both the disease and the treatment is

$$P_{\text{patient}}(D) = P_{\text{tumor}}(D) \times P_{\text{organ}}(D).$$

The patient's survival probability as a function of radiation dose is the lower curve in Fig. 11. If the dose is low, the tumor is likely to recur. At high doses, the therapy itself is likely to cause unacceptable complications, perhaps threatening the patient's life. Between these extremes the physician seeks to maximize the patient's survival of both the cancer and the treatment. If the radiation dose that optimizes the balance between curative treatment and complications is so high that the patient experiences extreme discomfort, the physician may be forced to settle for a lower dose even though the patient's chance of surviving the cancer is reduced.

Radiation-dose optimization demands accurate delineation of the margins of the cancer. That in turn demands good medical imaging. Before CT, cancer localization and treatment planning depended principally upon projection radiography combined with clinical and surgical (if available) information about the patient, together with knowledge of the natural course of the disease. But projection radiography has many disadvantages. It is frequently hard to see the cancer clearly, and it is often difficult to transcribe

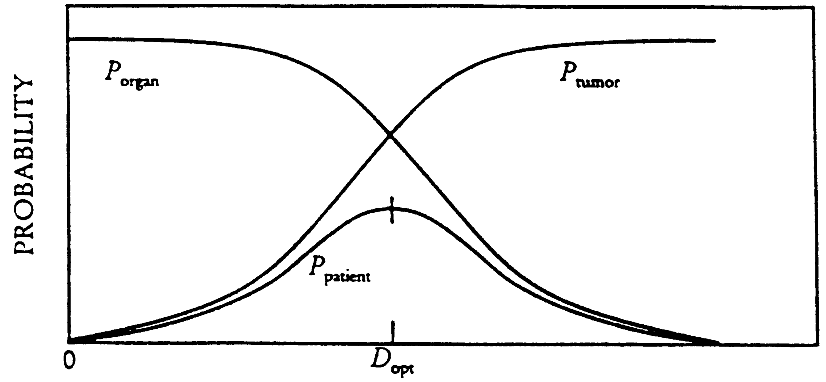

FIG. 11. Therapeutic outcome probabilities for cancer patients undergoing x-ray treatment, plotted schematically as functions of radiation dose D. The probability of success (P_{patient}) is the product of the probability that the tumor will be destroyed (P_{tumor}) and the probability (P_{organ}) that the treatment will not cause life-threatening damage to normal tissue.

the radiographic information onto a cross-sectional plane through the patient for designing the treatment plan describing the orientation and relative weighting of radiation beams used for treatment.

CT and, more recently, magnetic resonance imaging have largely overcome these problems. Many treatment centers have a CT unit dedicated to treatment planning and monitoring. The images provide not only a cross-sectional picture of the patient's anatomy, including the cancer and surrounding tissues, but also an accurate representation of the body contour and organs that are particularly sensitive to radiation. Because the data are digital, they can be entered immediately into the treatment-planning computer so that proposed treatment plans can be superimposed directly onto the CT images. Once an acceptable plan is accepted, it can be implemented on the x-ray treatment machine. This entire procedure must be carefully monitored to ensure that no misalignment occurs and that the treatment plan accurately reflects the radiation administered during treatment. Monitoring of this process is the responsibility of the medical physicist.

In most radiation therapy centers, commercial imaging units are used to generate CT information for treatment planning. In an effort to improve alignment of the CT information with the treatment geometry, a few medical physicists have built CT scanners on treatment-simulating gantries. Although these units yield spatial resolution inferior to that with commercial CT units, they duplicate the treatment geometry and provide images good enough to be useful for treatment planning.

Most x-ray treatments involve the use of multiple fixed radiation fields converging on the cancer from different directions. In this manner the dose is concentrated in the cancer while much lower doses are administered to surrounding normal tissues. In some cases a better dose distribution can be achieved by continually rotating the gantry of the treatment machine around the patient during treatment so that only the cancer is constantly in the path of the x-ray beam. One complication is that cancers are almost always asymmetrical. Therefore, the size of the x-ray beam must be expanded or contracted continuously during rotation. The dose distribution could be improved even more by simultaneously varying the dose rate during rotation. This approach, termed "conformal therapy," is being pursued in several centers in the United States and overseas (Mageras *et al.*, 1994). It requires detailed three-dimensional knowledge of the anatomy of the irradiated tissue, together with exquisite computerized control of the x-ray unit itself, as well as the x-ray gantry and patient couch position.

A significant advance in conformal therapy would be the convergence of CT and x-ray therapy in a single gantry, so that tomographic images could monitor alignment and dose distribution continually during treatment. This hybrid approach remains to be developed and is being pursued by a few groups of medical physicists (Convery and Rosenbloom, 1995; Mackie *et al.*, 1993). It presents magnificent technical challenges, but the potential of real benefits to patients as well (Hendee, 1995a).

6. CONCLUSIONS

The use of x rays for medical imaging is a prototypical example of "high-tech" medicine. It illustrates the contributions of medical physicists working with physicians to improve the health and welfare of patients. Much remains to be done, especially in refining the growing convergence of imaging with therapeutic techniques to cure disease and prevent the debilitating effects of injuries.

Physicists have played an instrumental role in the evolution of x-ray imaging. The American Association of Physicists in Medicine, a member society of the American Institute of Physics, includes almost 4000 members (AAPM, 1996). Graduate and postdoctoral training programs in medical physics increasingly attract bright young physicists. These individuals are among those destined to accomplish many of the advances that remain to be made.

GLOSSARY

Absorbed Dose: The energy absorbed per unit mass of an irradiated material.

Artificial Intelligence: The property of a machine capable of logical, systematic reasoning by which it can learn functions normally associated with human intelligence.

Characteristic X Rays: Electromagnetic radiation emitted as a result of rearrangements of the electrons in the inner shells of atoms; the spectrum consists of lines whose wavelengths depend only on the element concerned and the energy levels involved.

Compton Effect: Scattering of x rays by electrons of an atom when the energy of the x rays is large enough that the binding energy of the electrons can be neglected.

Computed Tomography: The process of producing an image showing body organs in cross section by first electronically detecting the variation in x-ray transmission through the body section along various paths at different angles, and then processing this information in a digital computer to reconstruct the x-ray absorption coefficients of the tissues at an array of points representing the cross section. Abbreviated CT. Also known as computerized axial tomography (CAT).

Digital Radiography: An x-ray imaging method in which x rays transmitted through the patient are recorded as a series of digital signals.

Dynamic Range: The amplitude range of signals that can be detected by an imaging system without distortion or disappearance into background noise.

Expert System: A computer system composed of algorithms that perform a specialized, usually difficult professional task at the level of (or sometimes beyond the level of) a human expert.

Fluoroscopy: Use of a fluorescent screen or image intensifier to convert an x-ray signal into an optical image to permit direct visual observation of x-ray images of objects interposed between an x-ray tube and the screen or intensifier.

Gantry: An extended mechanical structure holding elements of an imaging system.

Image Intensifier: An electronic device that receives an x-ray beam as an input signal and produces a bright and/or contrast-enhanced optical image as an output signal.

Intensifying Screen: A device used in x-ray imaging in which incident x rays cause the emission of light that in turn exposes a photographic film.

Linear Attenuation Coefficient: Fractional rate of attenuation of x rays in an absorbing medium.

Mammography: X-ray imaging of the breast.

Modulation Transfer Function (MTF): A measure of the ability of an imaging system to register fine detail present in the object.

Neural Network: An information-processing system that utilizes a large number of simple modules, to transmit, receive, and store digital information.

Photoelectric Effect: The liberation of an electron by electromagnetic radiation incident on a substance.

Pixel: The smallest area that can be displayed by an imaging system; a larger image is composed of many pixels.

Projection Radiography: The technique of producing a photographic image of an opaque specimen by transmitting a beam of x rays through it onto an adjacent photographic film; the image results from variations in thickness, density, and chemical composition of the specimen; used in medicine and industry.

Real-Time Display: A display that is able to follow changes in the object that occur at physiologic rates.

Subtraction: A process of image-contrast enhancement in which the brightness of corresponding parts of images of the same region made under different imaging conditions are subtracted from each other to remove distracting background structures and highlight structures of interest such as blood vessels.

Teleradiology: The transmission of medical images over long distances.

Tomography: The technique of making radiographs of plane sections of a body or an object; its purpose is to show detail in a predetermined plane of the body, while blurring the images of structures in other planes. Also known as laminography, planigraphy, and stratigraphy.

Works Cited

American Association of Physicists in Medicine (1996), One Physics Ellipse, College Park, MD 20740-3846 for information.

American College of Radiology (1993), *The Overview of Mammography Accreditation,* Reston, VA: American College of Radiology.

Convery, D. J., Rosenbloom, M. E. (1995), "Treatment Delivery Accuracy in Intensity-Modulated Conformal Radiotherapy," *Phys. Med. Biol.* **40,** 979–999.

Cormack, A. (1963), "Representation of a Function by Its Line Integrals, with Some Radiological Applications," *J. Appl. Phys.* **34,** 2722–2727.

Doi, K., Giger, M., Nishikawa, R., MacMahon, H., Schmidt, R. (1993), "Artificial Intelligence and Neural Networks in Radiology: Application to Computer-Aided Diagnostic Schemes," in: W. R. Hendee, J. Trueblood (Eds.), *Digital Imaging,* Madison, WI: Medical Physics Publishing, pp. 301–322.

Fenlon, H. M., Breatnach, M. B. (1995), "3-D CT Aids Planning of Craniofacial Surgery," *Diagnostic Imaging* **17,** (8), 47.

Fishman, E., Magid, D., Ney, D., Macrae, K., Kuhlman, J., Robertson, D. (1989), "Three-Dimensional Imaging: Advanced Medical Applications," in: J. Anderson (Ed.), *Innovations in Diagnostic Radiology,* New York: Springer-Verlag, pp. 47–68.

Heath, D., Soyer, P., Kuszyk, B., Bliss, D., Calhoun, P., Bluemke, D., Choti, M., Fishman, E. (1995), "Three-Dimensional Spiral CT during Arterial Portography: Comparison of Three Rendering Techniques," *Radiographics* **15,** 1001–1012.

Hendee, W. (1983), *The Physical Principles of Computed Tomography,* Boston: Little, Brown & Co.

Hendee, W. (1995a), "X Rays in Medicine," *Phys. Today* **48** (11), 51–56.

Hendee, W. R. (1995b), "History and Status of X-Ray Mammography," *Health Phys. J.* **69,** 636–648.

Hendee, W., Ritenour, R. (1992), *Medical Imaging Physics,* 3rd ed., St. Louis: Mosby-YearBook.

Hounsfield, G. (1973), "Computerized Transverse Axial Scanning (Tomography). I. Description of System," *Br. J. Radiol.* **46,** 1016–1022.

Kak, A., Slaney, M. (1988), *Principles of Computerized Tomographic Imaging,* New York: IEEE Press.

Kimme-Smith, C. (1992), "New and Future Developments in Screen-Film Mammography Equipment and Techniques," *Radiol. Clin. North Am.* **30,** 55–66.

Mackie, T. R., Holmes, T., Swerdloff, S., Reckwerdt, P., Deasy, J. O., Yang, J., Paliwal, B., Kinsella, T. (1993), "Tomotherapy: A New Concept for the Delivery of Dynamic Conformal Radiotherapy," *Med. Phys.* **20,** 1709–1719.

Mageras, G. S., Fuks, Z., O'Brien, J., Brewster, L. J., Burman, C., Chui, C. S., Leibel, S. A., Ling, C. C., Masterson, M. E., Mohan, R. (1994), "Initial Clinical Experience with Computer-Controlled Conformal Radiotherapy of the Prostate Using a 50-MeV Medical Microtron," *Int. J. Rad. Oncol. Biol. Phys.* **30,** 971–978.

Radon, J. (1917), "Über die Bestimmung von Funktionen durch ihre Integralwerte laengs gewisser Mannigfaltigkeiten (on the Determination of Functions from Their Integrals along Certain Manifolds)," *Ber. saechs. Akad. Wiss. Leipzig, Math.-Phys. Kl.* **69,** 262–277.

Röntgen, W. (1895), "Über eine neue Art von Strahlen (vorläufige Mitteilung)," *Sitzungs ber. physik.-med. Ges. Würzburg* **9,** 132.

Wolbarst, A. (1988), "Optimization of Radiotherapy Treatment Planning," *Encyclopedia of Medical Devices and Instrumentation,* Vol. 4, New York: Wiley-Interscience, p. 2488.

Wolbarst, A. (1993), *Physics of Radiology,* East Norwalk, CT: Appleton & Lange.

Further Reading

Grigg, E. R. N. (1965), *The Trail of the Invisible Light: From X-Strahlen to Radio(bio)logy,* Springfield IL: Charles C. Thomas.

Hasegawa, B. (1991), *Medical X-Ray Imaging,* 2nd ed., Madison, WI: Medical Physics Publishing.

Hendee, W. (1983), *The Physical Principles of Computed Tomography,* Boston: Little, Brown & Co.

Hendee, W., Ritenour, R. (1992), *Medical Imaging Physics,* 3rd ed., St. Louis: Mosby-YearBook.

Macovski, A. (1983), *Medical Imaging Systems,* Englewood Cliffs, NJ: Prentice-Hall.

Webb, S. (1988), *The Physics of Medical Imaging,* Philadelphia: Adam Hilger.

Webb, S. (1990), *From the Watching of Shadows: The Origins of Radiological Tomography,* Bristol, England: IOP Publishing Ltd.

Wolbarst, A. (1993), *Physics of Radiology,* East Norwalk, CT: Appleton & Lange.

X-RAY MICROSCOPY AND ANALYSIS

See Supplement Volumes

ATMOSPHERIC ICE

DENNIS LAMB, *Meteorology Department, The Pennsylvania State University, University Park, Pennsylvania, U.S.A.*

Introduction 3
1. **Observed Forms** 4
1.1 Observational Methods 4
1.1.1 *In Situ* Methods 4
1.1.2 Remote-Sensing Methods 5
1.2 Types of Atmospheric Ice 6
1.3 Distributions 9
2. **Origins of Atmospheric Ice** 11
2.1 Primary Nucleation 11
2.2 Secondary-Ice Formation 12
3. **Growth Processes** 13
3.1 Vapor Deposition 13
3.2 Aggregation 15
3.3 Accretion 16
4. **Effects of Atmospheric Ice** 18
4.1 Cloud Glaciation 18
4.2 Electrical and Chemical Phenomena 21
4.3 Interactions with Radiation 21
5. **Snow and Ice Storms** 22
5.1 Meteorological Settings 22
5.2 Formation of Snow and Ice 23
6. **Current Uncertainties and Research Areas** 24
Glossary 24
Works Cited 25
Further Reading 25

INTRODUCTION

Ice is a common constituent of the atmosphere and influences our world in diverse ways. Much of this ice, occurring as discrete particles of solid water within or emanating from clouds, resides entirely within the atmosphere over the course of its existence. Individual ice particles may form in the cold parts of clouds, where they grow for a time before encountering drier or warmer environments and evaporate or melt. Some ice particles, however, grow large enough to fall against the updrafts that spawn the clouds, appearing later as snow, sleet, or hail on the ground and then often hindering our routine activities as winter snow and ice storms and during summer thunderstorms. Surface snow tends to remain in place, because of its solid and immobile character, and accumulates when the local temperatures remain sufficiently low. In this way atmospheric ice contributes importantly to the maintenance of mountain glaciers and the formation of the seasonal snowpack that serves as a valuable resource for freshwater and recreation. Having had its origin in the atmosphere, such surface ice differs fundamentally from other types of ice on earth, such as sea or lake ice, but it similarly influences the radiation budget of the earth and therefore climate. Under some meteorological conditions, ice also arises on surfaces in the form of frost, rime ice, or glaze *via* processes closely related to those that form ice in clouds. The processes of atmospheric ice formation are thus important not only in their own right, but also through their impact on surface ice and human society.

Man's perception of atmospheric ice undoubtedly started with his encounter with snow. As nowadays, people of ancient times probably found snow to be both a nuisance and a wonder. The properties of snow cause it to hinder visibility, especially during high winds, and to pile up in most inconvenient ways. At the same time, one must marvel at the structure of snow, both in the macroscopic patterns that form on and around objects and in the intricate beauty of the individual crystals. It has been recognized for centuries that snow crystals exhibit a sixfold symmetry, but it was not until the end of the nineteenth century that the crystals were formally classified broadly as columnar (developed mainly along the principal axis) and plate-like or tabular (developed laterally). The endless variations in shape afforded by single ice crystals, as well as by the many polycrystalline forms found in

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

snow, hail, and sleet, make atmospheric ice one of the most mysterious solids in man's experience.

Virtually all natural ice as we know it on earth, whether appearing in the atmosphere or on the surface, is a crystalline solid in which the water molecules are arranged in an open lattice exhibiting hexagonal symmetry. This crystallographic structure, formally designated "ice Ih" to distinguish it from the several other possible solid phases or polymorphs of water, gives ordinary ice many of its unique properties. The external shape or morphology of single ice crystals grown by vapor deposition reflects this hexagonal symmetry rather directly. Other formation processes, such as those involving the freezing of liquid water, may yield ice particles of different morphologies, but invariably the molecular structure of all ordinary ice is fundamentally similar. The wide variety of ice forms found in the atmosphere results in large measure from the diversity of environmental conditions under which atmospheric ice originates and subsequently grows.

This article aims to provide an overview of the types of ice particles found in the atmosphere and to show how such forms arise from the set of physical processes we know to take place in the atmosphere. In the first section, some methods for observing and studying atmospheric ice are described, followed by a presentation of the diverse forms actually found. Then, in Secs. 2 and 3, extended discussions are given of the processes by which ice particles form initially and then grow into their various types. Some effects of ice on clouds and other atmospheric phenomena are considered in Sec. 4 before the main discussion ends with an overview of how nature produces snowstorms, in Sec. 5. Finally, Sec. 6 highlights some areas needing further study and affirms that atmospheric ice offers a fascinating and still incompletely understood topic for scientific investigation.

1. OBSERVED FORMS

1.1 Observational Methods

Atmospheric ice particles may be observed in a number of ways, depending on the type of information one wishes to obtain. When the properties of individual particles are desired, then one needs to use one of several *in situ* methods. On the other hand, the properties of a population of particles are best investigated with remote sensors. Here, some of the more commonly used techniques are mentioned briefly.

1.1.1 *In Situ* Methods *In situ* techniques are especially valuable when the sizes or shapes of the particles need to be known with some precision. Inspection of particle morphology, in particular, tells much about the origins and growth mechanisms of the ice particles, from which one can sometimes infer the environmental conditions during their growth. Often *in situ* methods start with the actual collection of ice particles near the ground or higher in the atmosphere from aircraft or balloons, although high-speed electro-optical devices now make it possible to characterize the particles nondestructively.

The direct collection of ice particles is usually done on a surface that has suitable natural characteristics or one that has been specially prepared. Probably the simplest way of sampling snow is to let the crystals fall onto a dark-colored, thermally insulating material, such as black felt. Direct visual observation or photography is then possible, once suitable precautions have been taken to avoid the melting or evaporation of the particles. Many of the classic studies of natural and artificial snow crystals, such as those by Nakaya (1954), Bentley and Humphreys (1962), and Magono and Lee (1966), were conducted with this straightforward method. Some of the earliest *in situ* observations of ice crystals in the mid- to upper troposphere were made by Weickmann (1945, 1949), who used microphotography aboard aircraft to identify the types of crystals that arise in different classes of clouds. Knowledge of the environmental conditions that give rise to specific crystal types allows one in turn to use the crystals themselves as aerological sondes and infer the state of the atmosphere at the time of their formation (Weickmann, 1957).

When immediate characterization is not possible, the sampled particles may be either preserved intact or replicated. Bulk snow samples and hailstones are typically collected in

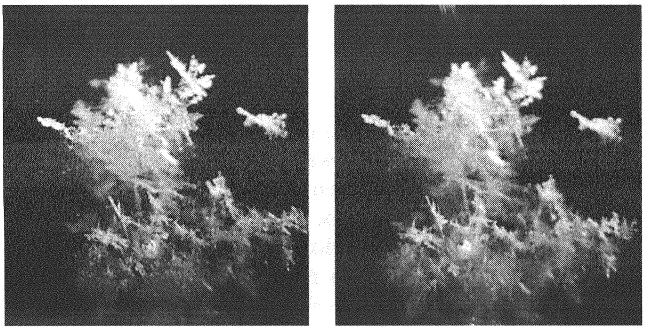

FIG. 1. Stereophotograph of a snow aggregate nearly 1 cm across.

clean containers on the ground and stored in freezers until appropriate physical or chemical analyses can be performed. Cryogenic preservation of snow samples has been used only occasionally, but astonishingly high-resolution photographs are possible when the crystals thus preserved are later viewed with a suitable electron microscope (*e.g.*, Wergin *et al.*, 1995). Replication involves physical contact between the sampled crystals and a prechilled liquid solution of an organic compound [*e.g.*, poly(vinyl formal) resin, known under the trade names "Formvar" and "Vinylec"] that then conforms to the shape of each crystal without melting it. After the solvent has evaporated, a thin solid film remains that retains the form of the original particle, allowing the replica to be viewed some time later at room temperature. Variations of this technique include briefly exposing precoated slides containing the ice samples to the solvent vapors, thereby preserving fine details of the crystals (see Hobbs, 1974, p. 652). The liquid replication technique has often been used aboard research aircraft that fly directly into the regions of clouds where the particles form and grow. Special balloon-borne replicators have also been developed to enable sampling in regions not readily accessible by aircraft. When relatively large ice particles, such as hailstones, are encountered, research aircraft may be equipped with special impactors in which imprints of the particles are made on strips of aluminum foil.

With modern electronic capabilities it is becoming increasingly common to characterize individual ice particles directly, without forcing them to strike a surface or otherwise destroying them. Commercial electro-optical instruments exist that yield discrete shadow images of the particles as they pass rapidly through a laser beam. Huge numbers of particles may be imaged within clouds from aircraft or during snowfalls on the surface. Although such methods have provided a wealth of information about the evolution of ice in clouds, shadow images lack depth information and resolve features only larger than about 50 μm. When greater detail is needed, one may again have to resort to photographic methods. As shown in Fig. 1, it is possible to capture the three-dimensional detail of delicate snow particles as they fall naturally through the air. Some holographic methods are also under development and show great promise for imaging ice particles in their natural state. New developments in the detection of small ice particles, those in the size range of 1–50 μm, are very much needed.

1.1.2 Remote-Sensing Methods The extraction of information about cloud properties from remote locations typically depends on the interaction of electromagnetic radiation with the suspended bits of matter that comprise the clouds. Remote sensing not only circumvents the need to travel to the cloud itself, but the type of information obtained usually represents a much more integrated measure of cloud properties than can be obtained *via in situ* methods. These advantages are, however, offset somewhat by the limited spatial resolution and the relative lack of detail about the individual particles that is obtainable. Moreover, detailed radiative-transfer models must often be employed in order to obtain the parameters of interest. Still, remote-sensing methods are being developed that yield an ever-expanding wealth of information about clouds and the ice they contain.

Remote sensors vary greatly in operation and how they can be used. The sensors may be deployed on the ground, aboard aircraft that may circle the clouds under study, or within satellites orbiting the earth. Typically, the sensors are classified by the wavelengths of radiation employed and whether they are "active" (energy-emitting and -receiving) or "passive" (energy-receiving only). Most commonly, remote sensors operate in the microwave, visible, and infrared regions of the electromagnetic spectrum. Modern science, of course, depends heavily on sophisticated electronic equipment, but we should not forget that much early and valuable information about clouds, even the ice they contain, was derived using the most natural of passive remote sensors operating in the visible range of wavelengths, the human eye. Astute observations by experienced personnel can still lead to important insights about the physical processes active in clouds.

Examples of the two classes of remote sensors are radar (active) and radiometers (passive). Radars have been the traditional means by which clouds are remotely interrogated, but recent innovations in technology have placed modern research radars far ahead of the weather-surveillance radars that emerged during the 1940s. Beyond simply measuring the received signal power (reflectivity), current radars can also measure velocity spectra and various polarization states of the returned signals. Vertically pointed radars operating at frequencies of 94 or 35 GHz have proven particularly useful for the study of cirrus clouds from the ground (Mace *et al.*, 1997). Radiometers gather path-integrated radiation emitted by cloud particles. Because the rotations of water molecules within the ice lattice are hindered, ice particles emit microwave radiation only very weakly. Nevertheless, microwave radiometers sensitive to liquid water are useful for identifying clouds conducive to later ice formation, especially when that water is supercooled. Infrared sensors aboard satellites are commonly used to identify cold cloud tops where ice is likely to exist. Field studies, such as that conducted in the mid-1980s to study the feasibility of augmenting the snowpack of the Sierra Nevada of California, are often designed around a coordinated suite of *in situ* and remote-sensing systems (see Reynolds and Dennis, 1986).

In the visible range, the most useful active device is the lidar (light detection and ranging). One or more wavelengths of radiation are transmitted (often vertically from the ground), while a telescope gathers the radiation scattered from the cloud particles. As with radars, some lidars can measure the Doppler-shifted frequencies, as well as the polarization state of the returned signal. Changes in the polarization of the returned radiation from that of the emitted beam are particularly sensitive to the physical phase of the particles, so that ice particles can often be uniquely identified in high clouds, even those in the stratosphere. Especially when used in combination, remote-sensing systems can yield much information about the occurrence of ice in the atmosphere and even some important microphysical properties, such as size, fallspeed, and crystal habit.

1.2 Types of Atmospheric Ice

Atmospheric ice is found in an extremely wide variety of forms, both as single crystals and as polycrystals. Single crystals, whether occurring individually or clumped together as "aggregates" (such as that shown in Fig. 1), exhibit some of the most spectacular forms in appearance and provide a unique record of their individual growth histories. Despite the broad range of possible shapes, single crystals tend to form regular geometric solids bounded by two hexagonal "basal" planes that lie perpendicular to the principal crystallographic (c) axis and six rectangular "prism" faces aligned parallel to the c axis (see Fig. 2). The "aspect ratio" of a crystal is defined quantitatively as the ratio of its dimension along the c axis to its maximum dimension between opposing prism-face edges, that is, along one of the lateral a axes. The c/a ratio thus identifies the primary habit of the crystal, which may be broadly classified as "plate", when $c/a < 1$, and "column", when $c/a > 1$. As we shall see later, the primary habit exhibited by any given crystal depends on the relative growth rates of the basal and prism faces.

The secondary habit of a single ice crystal

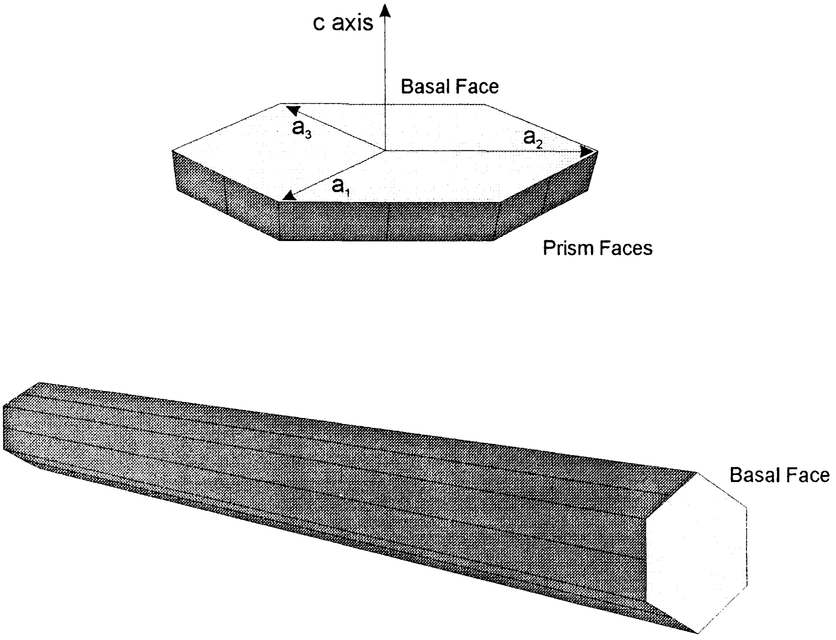

FIG. 2. Schematic diagram of the primary habits of vapor-grown ice crystals, showing perspective views of the hexagonal symmetry and the crystallographic axes. Top: plate. Bottom: column.

represents one of the myriad of possible variations in form that may be superimposed upon a given primary habit. The corners of a plate may grow preferentially, for instance, yielding a set of protruding branches of the type shown in Fig. 3(a). In this example, we clearly see the underlying hexagonal shape of the basal face as an interior feature, upon which the branches are superimposed. The brighter spots on the branches are the places where supercooled cloud droplets collided with the crystal and froze on impact once the crystal had grown large enough and gained an appreciable fall-speed. Despite the complexity of its overall morphology, this lightly rimed crystal is clearly classified as a plate because the dimensions across the branches (along any of the three *a* axes) greatly exceed its thickness (dimension along the *c* axis, here perpendicular to the plane of the photograph). Other subclasses of plates, many examples of which have been presented by Bentley and Humphreys (1962), include sector plates, broad-branch plates, stellar crystals, and dendrites.

Columnar crystals tend to exhibit a narrower range of secondary habits compared with those found on plates. As shown in Fig. 3(b), a "needle" crystal arises when the *c*-axis dimension greatly exceeds the *a*-axis dimension and fine points emerge from the corners of the basal faces. Other columnar crystals may exhibit hollowing of the basal faces, in the extreme yielding almost completely hollow forms called "sheaths". Sometimes, several columns originate from a common frozen droplet, yielding a combination form called a "bullet rosette". If such crystals break apart, each of the individual columns is called a "bullet" because of the pointed end that appears. Other columnar forms include "scrolls", especially common in some hoarfrost, and "capped columns", a combination form in which plates grow on the ends of the columns.

Polycrystalline ice particles contribute greatly to the variety of ice types found in snow and cloud ice. Figure 4 gives some idea of the diversity that can arise in individual snowfalls. In addition to a number of columnar crystals, we see frozen drops ("sleet") and heavily rimed particles called "graupel", both of which are subclasses in the category of "ice pellets". Careful inspection of the photograph reveals that broken branches from dendrites that subsequently rimed probably contributed to this particular snowfall as well. When the temperature of the ground is below the thermodynamic melting point (0 °C), slightly supercooled rain may form a heavy coating of "glaze ice" on all exposed surfaces, a phenomenon particularly prevalent in valleys that trap cold air in winter. At the other extreme of small drops that are strongly supercooled, surface structures may develop a thick layer of "rime ice", a situation common in mountainous regions when wintertime clouds intercept the elevated terrain. In extremely cold environments ice fogs may yield "droxtals", frozen droplets that exhibit crystallographic features on close inspection. In the summertime, some regions of the world experience damaging hail, precipitation composed of polycrystalline ice particles larger than 5 mm in diameter and exhibiting numerous layers alternating between clear and opaque ice when dissected. With the wide varieties of forms possible within any one category of ice, coupled with the huge number of combination forms that are conceivable, we begin to appreciate how complex the subject of atmospheric ice can be.

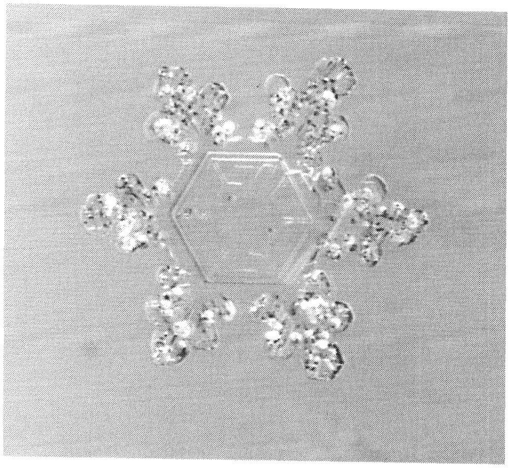

(a)

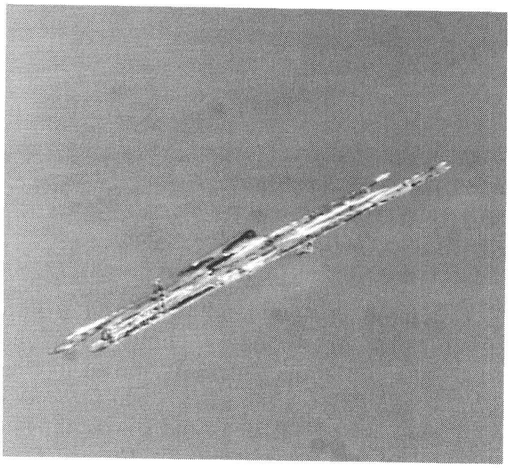

(b)

FIG. 3. Examples of single snow crystals about 1 mm across. (a) Plate with light riming on the sector-like extensions. (b) Needle.

1.3 Distributions

The diverse forms of atmospheric ice are found over broad geographic regions, depending on the local conditions of temperature and

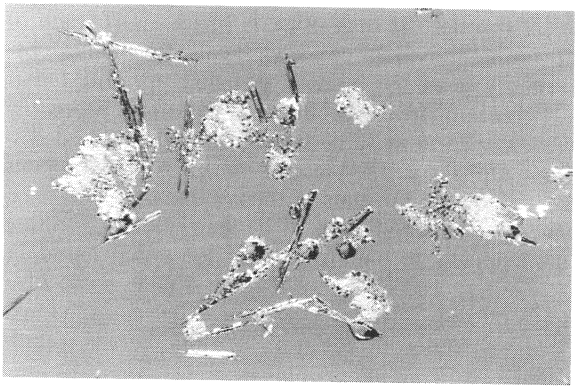

FIG. 4. Polycrystalline forms of ice comprising an aggregate.

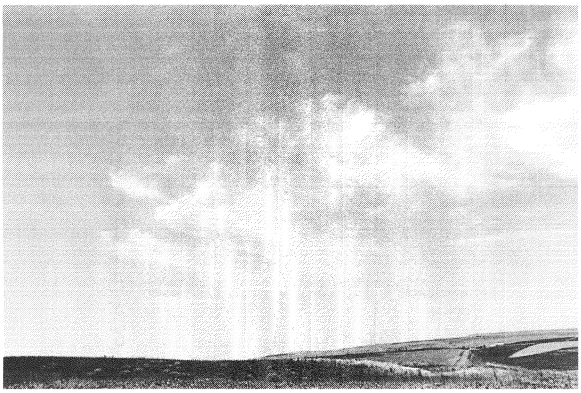

FIG. 5. Cirrus uncinus, a common form of ice cloud in the upper troposphere, showing fallstreaks.

humidity. Because the air temperature typically decreases with increasing height in the troposphere, the altitude of the terrain or of the cloud becomes a relevant factor in determining the distribution of atmospheric ice. Thus, whereas polar regions are generally conducive to the formation of ice even near sea level, high mountains in all latitude belts may allow for the formation and accumulation of ice and snow. Even in summer, most thunderstorms involve ice in their upper portions, as evidenced even from the ground by the fibrous appearance to the cloud tops, the occasional appearance of hail, and, very likely, the occurrence of lightning. Air exiting a storm near its top carries residual moisture that may later deposit on aerosol particles in the upper troposphere, yielding ice clouds, such as the cirrus uncinus shown in Fig. 5. Small ice particles also form, despite relatively dry conditions, in the polar regions of the lower stratosphere and upper mesosphere, where temperatures can fall below −100 °C. Ice must be viewed as an important component of the atmosphere during all seasons of the year, not just in winter, and in all climatic regimes, in the tropics as well as at high latitudes.

The occurrence of specific types of vapor-grown ice crystals has been shown, through repeated observation in the atmosphere and experimentation in the laboratory, to depend strongly on both the temperature and humidity of the air at the time of growth. To a good first approximation, the temperature seems to be the environmental variable that most controls the primary habit, as shown in Fig. 6. As ice crystals originate and grow at progressively lower temperatures, the primary habit is found to alternate between plate and column, with transition temperatures at approximately −3, −8, and −22 °C. The abundance of "excess" moisture, that is, the vapor density in excess of the equilibrium value, seems to determine the secondary habit: the larger the excess vapor density, the more extreme the secondary growth features. Thus, as summarized in Fig. 6, in the plate regimes we find simple plates at low excess vapor densities, but progressively longer extensions and complexities of the branches protruding from the corners as the excess vapor density increases. The simultaneous presence of liquid water in a cloud restricts the magnitude of excess vapor density available (shown by the dashed curve), and so the most intricate (dendritic) structures are found in the middle of the colder plate regime, where the liquid-water saturation effect is close to a maximum. Similarly, columns with the greatest aspect ratio (*i.e.*, needles) appear in the presence of liquid cloud droplets when the growth temperature is near −6 °C. Compact columns, crystals with near unity aspect ratio, form near the primary-habit transition temperatures.

Vapor-grown ice crystals exhibit a number of interesting features that depend on the conditions of growth. The rate with which water mass accumulates on a crystal in clouds near liquid-water saturation varies with the growth temperature in an unusual manner. As shown in Fig. 7, the mass growth is maximum at two temperatures, not just one, as might be ex-

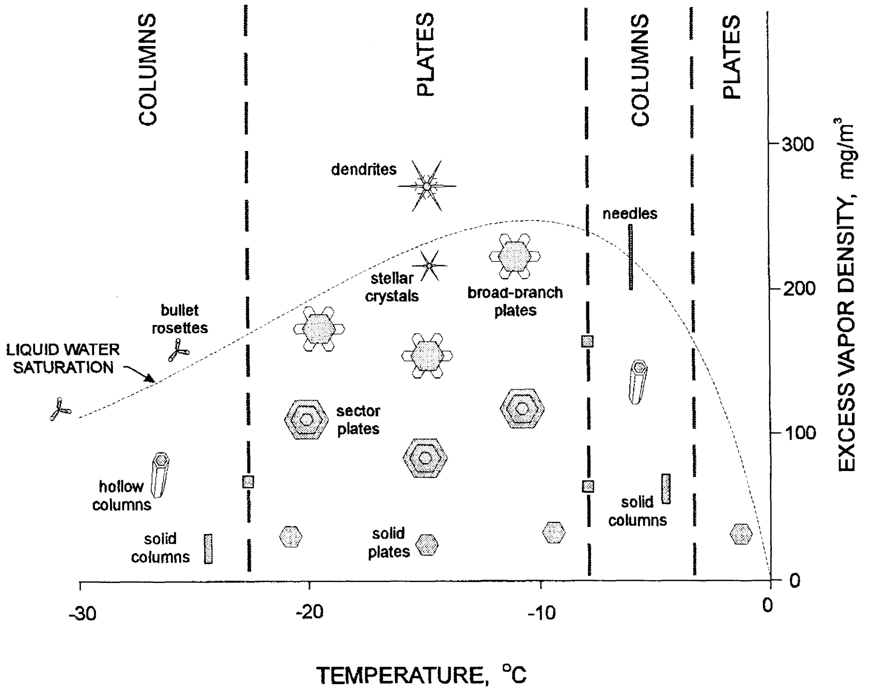
COLUMNS
PLATES
COLUMNS
PLATES
dendrites
needles
stellar crystals
broad-branch plates
bullet rosettes
LIQUID WATER SATURATION
sector plates
hollow columns
solid columns
solid plates
solid columns
-30
-20
-10
0
TEMPERATURE, °C
0
100
200
300
EXCESS VAPOR DENSITY, mg/m³

FIG. 6. The distribution of ice crystal habits with growth temperature and excess vapor density (the density of vapor above the equilibrium value at each temperature).

pected from the distribution of excess vapor density with temperature when liquid water is present (dashed curve in Fig. 6). It appears that the mass growth rate is linked closely to the habit of the crystal, the highest growth rates appearing when the axial ratios are most extreme (as with needles, an extreme form of columnar growth, and with dendrites, an extreme type of plate-like growth).

The axial ratios of vapor-grown crystals depend not only on the temperature and supersaturation, but also on the dimensions and habits of the crystals themselves. As a summary of many observations of crystals, Fig. 8 shows that the dimensions of vapor-grown ice crystals along the *a* and *c* axes follow power-law relationships only up to a certain point. Once the dimension along the rapidly growing axis exceeds about 100 μm, the growth along the other axis seems to be limited. Although one cannot readily ascertain whether the growth along the minor axis actually de-

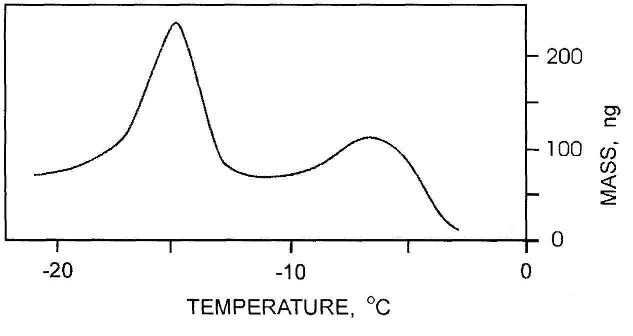

FIG. 7. Variation of crystal mass with temperature after a growth period of 150 s. Data from Ryan *et al.* (1976).

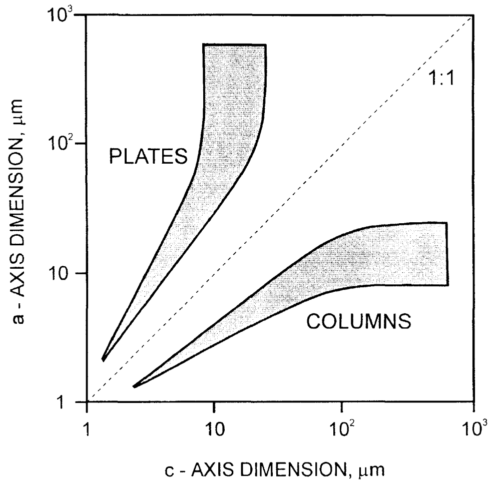

FIG. 8. Relationship between the dimensions along the *c* and *a* axes of crystals grown from the vapor in the plate and column regimes. Adapted from Chen and Lamb (1994), based on data from Ono (1970).

creases, the fact that this change in growth behavior occurs just when the crystal fallspeed becomes appreciable suggests that the growth along the major axis preferentially increases as a result of the enhanced ventilation of the most protruding parts of the crystal (Chen and Lamb, 1994).

2. ORIGINS OF ATMOSPHERIC ICE

The processes that give rise to ice particles in the atmosphere are as yet only partially understood. Typically, clouds develop initially as a collection of liquid water droplets that cool gradually as they are carried aloft by atmospheric motions. Once the local temperature decreases below 0 °C, the droplets begin to supercool and persist for a time in this metastable state. The first ice to appear in the supercooled cloud forms by one of several "primary" nucleation mechanisms. Later, additional ice-formation mechanisms may operate to generate a second generation of ice particles from the first, leading to "secondary" ice. Here, an overview of some of the more important ice-initiation processes is given.

2.1 Primary Nucleation

Ice in most clouds originates in the presence of supercooled liquid-water droplets at temperatures below the thermodynamic melting point of ice. Current nucleation theory suggests that pure water can supercool by as much as 45 °C, depending on the sizes of the droplets (Pruppacher and Klett, 1997, p. 341). The first ice to appear in some upper-level clouds seems to be consistent with a mechanism of homogeneous nucleation of liquid solution droplets, but at lower levels in the atmosphere the first ice appears in the −8 to −12 °C temperature range. Thus, a heterogeneous mechanism is needed to explain most primary ice formation.

Heterogeneous nucleation refers to the catalyzing influence that some foreign (nonaqueous) substances can have on the nucleation rate. Although a detailed understanding of the mechanism is still lacking, it appears that a solid substrate is needed that has lattice characteristics similar in some way to those of ice. For instance, crystalline silver iodide, often produced artificially to induce ice formation in supercooled clouds, is particularly effective because its lattice exhibits hexagonal symmetry and has cell dimensions differing only slightly from those of ice. Suitable ice-nucleating materials in nature are typically insoluble substances of either mineral or biogenic origins, such as clay particles or fragments of bacteria, respectively. Natural ice nuclei always comprise a minute subset of the atmospheric aerosol (often less than 1 in 10^6), but these few particles are responsible for much of the ice found in the atmosphere.

Once brought into an appropriate part of a supercooled cloud, the ice nuclei may interact with the cloud droplets in one of several ways. The nucleus is said to act in the "freezing" mode if it has already become immersed inside the cloud droplet sometime prior to the freezing event. Other nuclei, particularly small and mobile particles, seem to initiate freezing of the cloud drops on "contact" with the liquid surface. A third mode of action occurs when supersaturated water vapor deposits directly onto the particle and forms ice without the obvious presence of any liquid water. This "deposition" or "sublimation" mode of nucleation thus differs in concept from the freezing and contact modes.

The ability of any given particle to activate the solid phase is a strong function of the temperature and supersaturation with respect to ice. Whereas earlier views held that the temperature and supersaturation effects were independent, many now think that the main factor is supersaturation, temperature playing a secondary and largely implicit role. Measurements of ice nuclei acting in the freezing or deposition modes exhibit an exponential dependence on supersaturation in the temperature range between −5 and −20 °C, yielding a concentration of ice in the order of 10^3 ice particles m^{-3} at −10 °C in the presence of liquid cloud droplets (Meyers *et al.*, 1992). Because of the difference in the equilibrium vapor pressures of ice and liquid water, higher concentrations of ice can be expected in supercooled clouds at lower temperatures.

The first ice to form in supercooled clouds typically arises from the freezing of a small fraction of the cloud droplets. The freezing of these droplets takes place in two sequential stages. In the first stage, a dendritic mesh of ice spreads rapidly throughout the drop volume, releasing latent heat that warms the slush to 0 °C within about 1 ms. This initial stage is effectively adiabatic, because of the small time available for thermal exchange with the environment, and allows an approximate fraction $c_w \Delta T/\lambda_f \cong \Delta T/80$ of the liquid to be converted into ice, where ΔT is the supercooling, c_w is the heat capacity of liquid water, and λ_f is the latent heat of fusion. The remaining liquid in the slush freezes more slowly during the second stage, the process being limited by the rate with which the latent heat can be dissipated by conduction to the surrounding air. At the relatively high temperatures (between about −8 and −12 °C) at which first ice forms in many clouds, the probability of generating a stable embryo of ice within the liquid is low, so that these frozen droplets are usually single crystals.

These first frozen droplets, by virtue of the numerous, still-liquid droplets present, experience a high supersaturation and rapid deposition of vapor onto their surfaces. This growth takes place mainly on the prism faces because the nucleation occurs in the plate regime (−8 to −22 °C). As depicted in Fig. 9, the form of the first ice, commonly detected by *in situ* sampling, is typically a double plate, two plate-like crystals separated by the original frozen droplet. Ice that forms by droplet freezing at much lower temperatures (below about −20 °C) typically shows evidence of polycrystalline nucleation and preferential growth along the *c* axes of each crystal, yielding bullet rosettes.

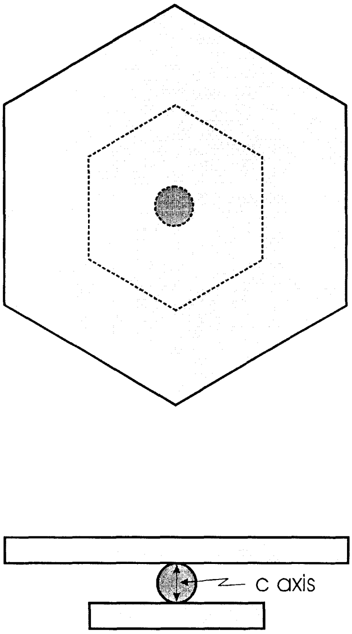

FIG. 9. Schematic diagram of a "double plate". Top: view along the *c* axis. Bottom: view normal to the *c* axis (arrow).

2.2 Secondary-Ice Formation

Secondary ice forms as a consequence of physical processes involving ice particles that already exist. The concept of second-generation ice is needed in part to account for high concentrations of ice particles found in some clouds. In some cases, the observed concentration of ice particles exceeds the concentration expected from primary nucleation events by a thousand times or more. Some snowstorms, too, seem to be prolific sources of needles, a crystal type known to grow only within a narrow temperature range around −6 °C, under conditions unfavorable to primary nucleation. Both laboratory experimentation and field observations confirm that new ice particles can be generated in supercooled clouds once some ice has already formed in the cloud (Young, 1993, p. 97).

A search for the likely mechanisms by which secondary ice forms has been under way for nearly three decades with modest success. Difficulties arise in most investigations in part because the conditions conducive to the generation of secondary ice appear to be narrow. That is, whereas clouds and primary ice form under a broad set of relatively predictable conditions, secondary ice tends to form only in narrow temperature zones and possibly only when a restricted set of cloud particles is present. Reproducing the required set of conditions in the laboratory or finding them in nature can be difficult. Nevertheless, some generalities can be made.

The mechanisms responsible for secondary-ice formation typically involve the interactions of cloud particles. Sometimes these interactions are rather indirect; other times they involve actual collisions between particles. One line of research has suggested that the so-called secondary ice is actually primary ice stimulated by local fluctuations in conditions. Thus, by way of example, an aerosol particle that would not normally be active as an ice nucleus under average cloud conditions (say, liquid-water saturation at $-10\,^\circ$C) might happen to pass through the strong gradients of temperature and moisture in the wake of a graupel particle or of a large drop in the process of freezing. Such a particle might become active in the transient high-supersaturation environment of the existing ice particle and cause a new ice particle to form (Fukuta and Lee, 1986). Direct collisions between two ice particles are relatively common and have been thought by some to yield daughter particles. Such a "fragmentation" mechanism is expected to operate when either or both of the particles is structurally delicate, as dendrites are naturally or as many crystals become during evaporation or melting. Of particular importance, the process of accretion (discussed in Sec. 3.3) involves many collisions of supercooled cloud droplets with graupel particles, some fraction of which yields small ice "splinters". Considerable evidence (see Pruppacher and Klett, 1997, p. 357, for history) has been amassed in both the laboratory and the field that this "rime-splintering" mechanism can operate in the columnar-growth regime (-3 to $-8\,^\circ$C) when the graupel particles are sufficiently large (≈ 0.5 mm in diameter). The consequences of secondary-ice formation are discussed later within the context of cloud glaciation.

3. GROWTH PROCESSES

Once formed, ice particles must grow by the accumulation of water mass to sizes that allow them to interact with other cloud particles or fall out of the cloud as precipitation. The mechanisms by which such growth takes place are discussed here. The various processes may all take place simultaneously, but for clarity each is considered separately.

3.1 Vapor Deposition

The first stage of growth usually involves the direct deposition of water vapor onto the exposed surfaces of an ice particle. As water molecules are removed from the vapor phase in the immediate vicinity of the growing ice crystal, gradients of vapor density are set up that drive the molecular diffusion of vapor from more "remote" regions (several particle diameters) about the particle. The process of vapor deposition is sometimes referred to as "diffusional" growth because the rate of growth is typically limited by this phenomenon of molecular diffusion through the surrounding air. However, unlike the diffusional growth of liquid droplets, which can be explained fully by consideration of radial gradients alone, the nonspherical shapes assumed by ice crystals implies that nonradial vapor gradients and the processes responsible for incorporation of the impinging molecules into the ice lattice, *i.e.*, the "surface kinetics", must also be taken into account.

The probability that a given vapor molecule is incorporated into the crystal lattice and contributes to the growth of the ice particle is related to the surface characteristics of the particular face it contacts. Laboratory measurements show that this efficiency factor, often called a "deposition coefficient", varies strongly with the crystallographic orientation of the face. Whereas the deposition coefficients

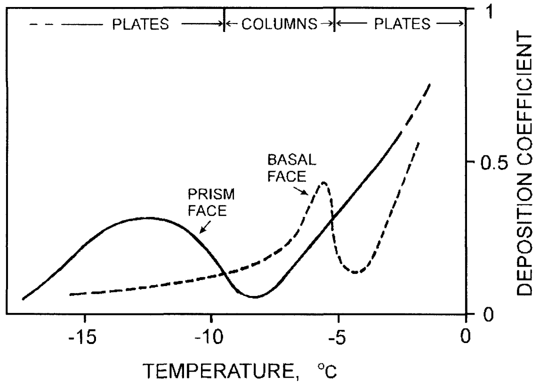

FIG. 10. Dependence on temperature of the deposition coefficients of the basal face and prism faces. Adapted from Lamb and Scott (1974).

of all high-index faces are close to unity, those of the low-index basal and prism faces vary strongly with the temperature and supersaturation at the time of growth. As shown in Fig. 10 for constant excess vapor density, these deposition coefficients each decrease with decreasing temperature from values near unity at the thermodynamic melting point, but not monotonically. The deposition coefficient of the basal face has a local maximum near −6 °C, whereas that of the prism face exhibits a broad maximum between −10 and −15 °C. The temperatures at which the curves cross are close to the observed primary-habit transitions and represent equal likelihood of growth on the basal and prism faces. At lower excess vapor densities the deposition coefficients become smaller, but the crossing points remain relatively invariant.

The reason for the complicated variation of the deposition coefficient (and therefore the primary habit) with temperature is probably related to surface structure and to systematic changes that may occur on the basal and prism faces of ice. Each of these low-index faces advances by the propagation of terraces or steps across the surface. The advancement of each step relies on a supply of adsorbed water molecules that migrate by surface diffusion to "kink sites" along the step, favorable locations for molecular incorporation into the lattice (see Fig. 11). The greater the surface migration distance, the greater the molecular supply rate and the larger the probability of incorporation. In this mechanistic picture the linear growth rates and, ultimately, the habits of ice crystals depend on a balance between the probabilities of molecular desorption (which determines the lifetime in the adsorbed state) and surface diffusion (which determines the speed of migration). The strong temperature dependence of the deposition coefficient is therefore likely to be related to the adsorption enthalpy and the activation energy for diffusion, both of which depend on the as yet unknown molecular structure of the crystal surface. A real understanding of crystal growth must await further research in this area.

As complicated as the molecular-scale processes responsible for the vapor growth of ice crystals are, they must also be put into the context of the growth environment experienced by the particle as a whole. The interactions that occur between the surface processes and the ambient vapor field introduce complexities that ultimately account for the diversity of crystal habits found in nature. The incorporation of the water molecules into the ice particle locally alters the vapor density and its gradients in the gas phase. Any changes in the local vapor density in turn affect the magnitude of the deposition coefficient (*i.e.*, the efficiency of growth).

When reasonable allowance is made for the interactions arising between the ice surface and the surrounding vapor field, we can see a likely mechanism for the formation of the many observed secondary habits, as exemplified in Fig. 12 for the case of a spherical ice particle growing in the plate regime. Initially, the high-index faces grow out rapidly, leaving only the low-index basal and prism faces. At low supersaturations (Fig. 12, top), the weak vapor gradients allow growth steps to propagate uniformly across each prism face and the form of a simple hexagonal plate to be maintained. At

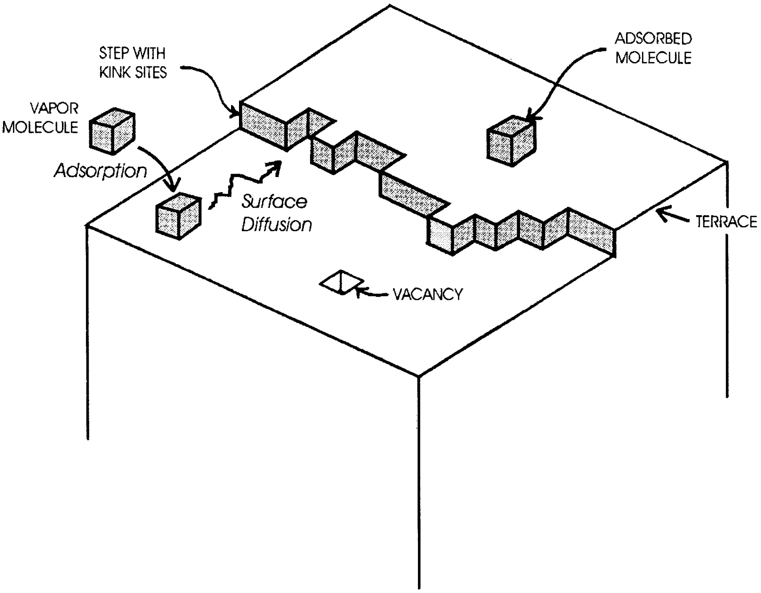

FIG. 11. Schematic representation of molecules incorporating into a low-index face.

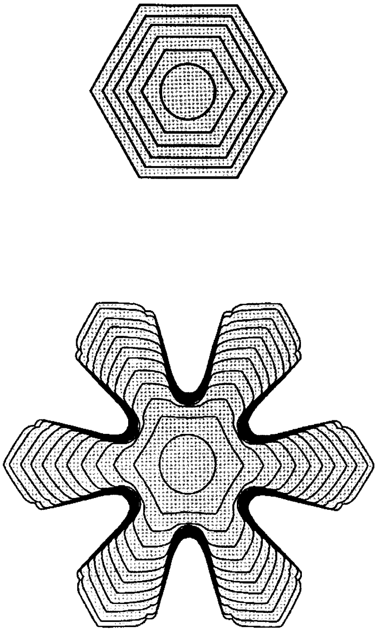

FIG. 12. Schematic illustration of the evolution of secondary habits in the plate regime from an initially spherical frozen droplet. Growth at low (top) and high (bottom) supersaturations. Adapted from Yokoyama (1993).

larger supersaturations (Fig. 12, bottom), the gradients of vapor density become appreciable, and so steps that propagate toward the center of each face, where the local supersaturation is relatively small, slow up and "bunch" together. The edges bounding neighboring prism faces experience the largest supersaturations, leading to rapid growth locally, the formation of protruding branches and crystals classified as "broad-branch plates". At still larger ambient supersaturations, the gradients and their effects on the growth become larger still, to the point where a "stellar" crystal forms. At high supersaturations, the shape of the crystal itself affects the distribution of excess vapor across its surfaces and hence the subsequent growth of the crystal. Protruding parts invariably command the greatest vapor supply and the highest probability of continued growth.

3.2 Aggregation

The growth process of aggregation differs fundamentally from that of vapor deposition. Crystals that have already achieved a certain size by the deposition of vapor may, by chance, collide with neighboring crystals and adhere, forming ice particles of larger mass and generally more complicated morphology. When continued for some time under appropriate conditions, this process can lead to "aggregates" containing dozens or even hundreds of separate crystals that each grew individually from the vapor phase. The large aggregate shown in Fig. 1 provides an example of the morphological complexity that can result. One must also appreciate that such ice particles are rather fragile and may fragment during some collisions.

Aggregation, because it involves collisions between particles, is one of several growth processes classified as "collection". As with all collection processes, aggregation tends to form fewer, but larger, particles at the expense of the large number of smaller particles in a cloud. The collection process is inherently discrete and stochastic in nature. In principle, when an appropriate account is taken of all of the pairwise interactions in a population of particles, a collection equation arises that can be used to evaluate the rate of change of the particle size distribution:

$$\frac{\partial n(m)}{\partial t} = \frac{1}{2}\int_0^m K(m_x, m - m_x)n(m_x) \times n(m - m_x)\,\mathrm{d}m_x - n(m)\int_0^\infty K(m, m_x)n(m_x)\,\mathrm{d}m_x. \quad (1)$$

Here, $n(m)$ is the particle mass distribution function, and the collection kernel is

$$K(m_\mathrm{L}, m_\mathrm{S}) = E(m_\mathrm{L}, m_\mathrm{S})\pi(r_\mathrm{L} + r_\mathrm{S})^2(v_\mathrm{L} - v_\mathrm{S}), \quad (2)$$

where the effective "radius" of the large ("collector") particle is r_L and the radius of the small ("collected") particle is r_S. The corresponding particle fallspeeds are given by v_L and v_S, and the collection efficiency is $E(m_\mathrm{L}, m_\mathrm{S})$. The factor of $1/2$ prevents the over-counting that would otherwise occur from the first integral. Note that the particle concentration (n) appears in the equation as a product with itself, and so the collection process is inherently nonlinear.

Whereas solutions to the collection equation have been used numerous times in the past to calculate the spectrum of liquid drops in clouds, complications associated with the nonspherical shapes of ice particles make this approach difficult and uncertain. Nonspherical solid particles often fall irregularly, especially at high particle Reynolds numbers. Vortices formed in the rear of some particles may cause them to slip sideways or even tumble, for instance, so that large uncertainties exist in both the area swept out and the fallspeeds used in Eq. (2). The collection efficiency, too, is complicated by the fact that it represents a conditional probability of two events, collision and coalescence, the sticking of the two particles together once contact has been made. Relative to droplets, the nonspherical nature of ice particles may actually augment the collision probability, but solid particles can often collide without adhering to each other. Such bounce-off events have positive implications for some cloud processes (*e.g.*, electrification), but they will not lead to growth of aggregates.

Empirical studies have most often yielded the more definitive information regarding the aggregation process. Measurements from both inside and below clouds show that aggregation is favored at relatively high temperatures, although the shapes of the crystals may also play a role. The largest aggregates, their diameters sometimes exceeding a centimeter, are found when the observation temperature is higher than about −5 °C. Conventional thinking holds that ice within a few degrees of the melting point forms a "quasiliquid" or at least a highly disordered layer of water on its surface. The mechanism of coalescence may therefore be associated with sintering. A secondary peak in particle size is often found between −10 and −15 °C and has usually been ascribed to the mechanical interlocking of the dendritic arms of separate crystals. In general, it appears that the main mechanistic limitation to aggregation is crystal coalescence.

3.3 Accretion

Accretion is the technical term given to the process by which ice particles or other structures collect supercooled drops, yielding a milky, granular deposit of rime, or alternatively a clear, smooth deposit of glaze, depending on the accretion rate and degree of supercooling. Ice mass accumulates as the drops collide with the object and freeze on contact. Accretion is an important process by which precipitation is formed in summertime convection, as well as in winter snow storms, but it can also lead to the buildup of ice on aircraft or structures attached to the Earth's surface, such as buildings, power lines and trees. Surface accretion is particularly prevalent in elevated terrain, as when supercooled clouds envelop mountain tops and yield thick coatings of rime on trees and rocks. Similarly, the sedimentation of supercooled drizzle and rain onto surface objects can quickly leave them with a thick coating of clear glaze ice.

As with aggregation, accretion is a collection process, but it otherwise differs in many ways from aggregation. Accretion involves the collection of liquid droplets that invariably freeze. This phase change has several consequences that cause rimed particles to behave uniquely in the atmosphere. Whereas aggregation leads to snowflakes with an open, low-density structure, accretion leads to compact, dense ice particles that tend to be structurally strong. Their compact forms mean that rimed particles offer relatively large mass and low aerodynamic resistance, so that their terminal fallspeeds are higher and even increase in rough proportion to the degree of riming.

Growth of ice particles by accretion proceeds by stages. Individual ice crystals first begin collecting supercooled cloud droplets after having achieved certain minimal sizes by vapor deposition. For plates, the threshold for collection is crossed once the semi-major dimension approaches 150 μm, whereas for columns the onset of riming occurs when the radius along the minor (*a*-axis) dimension exceeds about 25 μm. Figure 3(a) provides an example of a lightly rimed crystal. With continued accretion, crystals gradually become encased in a mass of frozen droplets and eventually lose their identity. At this point, the ice particles are termed "graupel" (sometimes "snow pellets"), as exemplified by the larger particles shown in Fig. 4. At some point, typically when the graupel diameter exceeds 5 mm, the particle fallspeed and collection rate increase to such an extent that the droplets do not have time to complete their freezing process before subsequent droplets are collected. The accreted water mass then flows across the surface and freezes relatively slowly, resulting in a layer of clear ice and the birth of a hailstone. The alternating layers of opaque and clear ice characteristic of hailstones arise as the individual hailstones successively encounter regions in a cloud of relatively low or high liquid water concentration and grow in "dry" or "wet" regimes, respectively.

These phenomenological aspects of accretion can be understood from a consideration of the energy budget associated with an individual ice particle that is simultaneously collecting supercooled droplets and interacting with the vapor field. Once well beyond the threshold size for riming, an ice particle (graupel or hail) of equivalent radius r_p accretes water mass continuously at a rate that can be expressed approximately as

$$\frac{\mathrm{d}m_\mathrm{p}}{\mathrm{d}t} = K\omega_\mathrm{L}, \tag{3}$$

where $K = \pi r_\mathrm{p}^2 E v_\mathrm{p}$ is the collection kernel, E is the collection efficiency, v_p is the particle fallspeed, and ω_L is the liquid water concentration (collective mass of cloud droplets per unit volume) encountered by the falling particle. The freezing rate implied by this accretion rate causes the particle to gain enthalpy at the rate $\lambda_\mathrm{f} K\omega_\mathrm{L}$, where λ_f is the latent heat of fusion. The particle warms, decreasing or even reversing the rate of vapor deposition. In a steady state, the energy gained by the particle through the freezing of accreted water must be balanced by the energy lost by sublimation and heat conduction to the surrounding air. Thus,

$$\lambda_\mathrm{f} K\omega_\mathrm{L} = 4\pi r_\mathrm{p} f_\mathrm{v} D_\mathrm{V}(\rho_s - \rho_\infty)\lambda_\mathrm{s} + 4\pi r_\mathrm{p} f_\mathrm{h} k_\mathrm{T}(T_\mathrm{s} - T_\infty), \tag{4}$$

where f_v and f_h are ventilation coefficients for vapor and heat transfer, respectively, while D_V is the diffusion coefficient of water vapor in air, λ_s is the latent heat of sublimation, k_T is the thermal conductivity of the air, ρ_s is the vapor density immediately over the particle surface at temperature T_s, and ρ_∞ is the vapor density remote from the particle, where the temperature is T_∞. The threshold value of ω_L for wet growth, the so-called Schumann–Ludlam limit, occurs when the particle temperature reaches the melting point T_0 and is

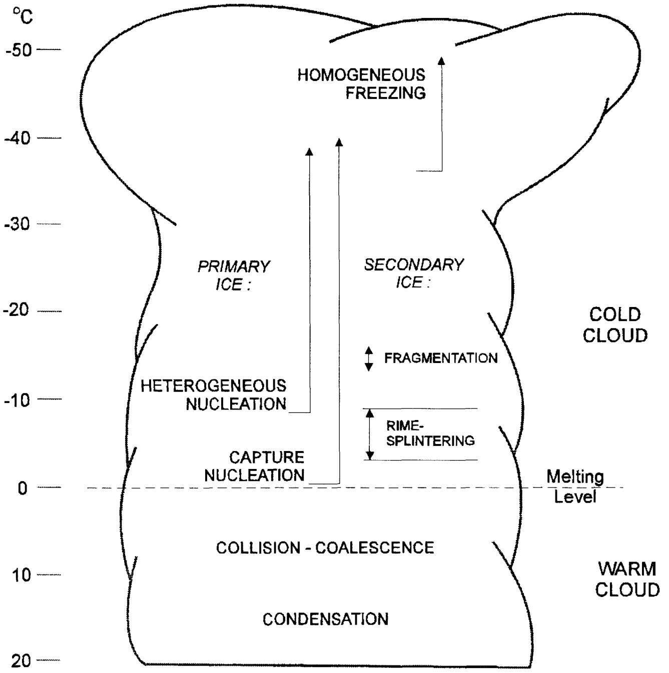

FIG. 13. An overview of the mechanisms that contribute to cloud glaciation.

given by

$$\omega_{\mathrm{L,wet}} \geq \frac{4}{E r_{\mathrm{p}} v_{\mathrm{p}} \lambda_{\mathrm{f}}} [f_{\mathrm{v}} D_{\mathrm{V}} (\rho_{\mathrm{s}} - \rho_{\infty}) \lambda_{\mathrm{s}} + f_{\mathrm{h}} k_{\mathrm{T}} (T_0 - T_{\infty})]. \qquad (5)$$

As a large ice particle, such as a hailstone, falls relative to the surrounding air and encounters liquid water concentrations alternately above and below this instantaneous threshold value (ω_{L}), it forms layers of clear and opaque ice, respectively. At very high liquid-water concentrations the rate of accretion may exceed the particle's ability to retain all of the water, in which case "shedding" of excess water occurs.

4. EFFECTS OF ATMOSPHERIC ICE

4.1 Cloud Glaciation

In an atmospheric context, glaciation refers to the formation of ice in a cloud, usually the complete transformation of an initially all-liquid (supercooled) cloud to an all-ice cloud. The mechanisms by which this phenomenon can occur are still speculative, for it must be appreciated that clouds represent water in a highly dispersed phase. Freezing events in some parts of a cloud do not automatically affect other parts of the cloud, yet it often appears as if an entire supercooled cloud, especially when convective, converts into ice within a matter of minutes. The processes responsible for rapid glaciation very likely involve secondary-ice generation and can account for the relatively high concentrations of ice found in some clouds, as discussed earlier in Sec. 2.2. Here, some mechanistic possibilities and the effects associated with glaciation are explored.

Ice can arise in a number of ways within supercooled clouds. The schematic diagram given in Fig. 13 provides an overview of the possibilities in convective clouds. Homogeneous freezing can generate many ice particles (frozen droplets) very rapidly, but not unless the temperature is below about −35 °C, meaning that only the upper parts of a cloud glaciate by this process. Heterogeneous nucleation is needed to explain ice lower in the cloud, but this mechanism is weak at temperatures higher

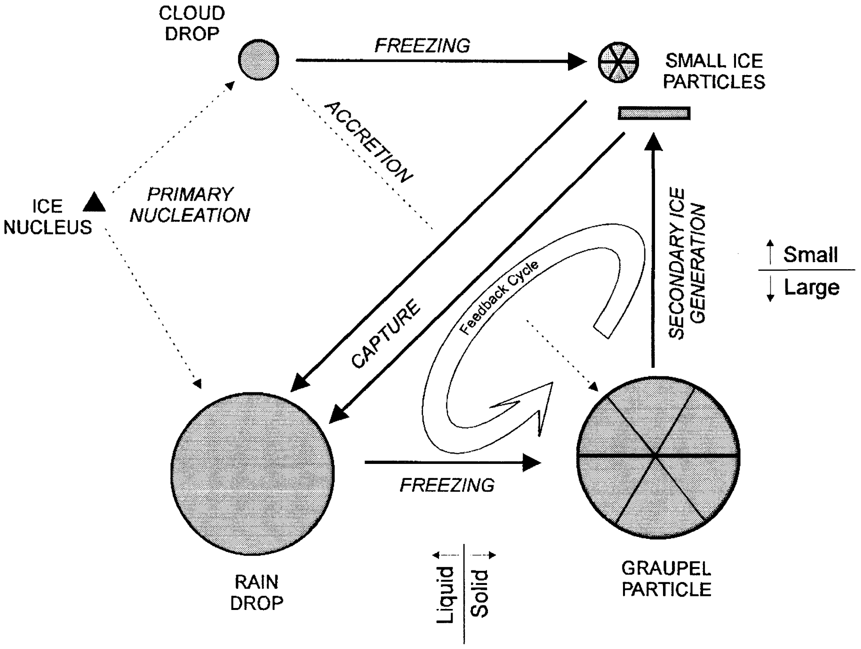

FIG. 14. Schematic diagram of possible interactions between various classes of cloud particles during the glaciation of a supercooled cloud. Adapted from Lamb *et al.* (1981).

than about −8 °C. "Capture nucleation", a freezing event initiated by the hydrodynamic collision (or capture) of a small ice particle with a relatively large supercooled drop, is effective for enhancing the mass of ice anywhere within a supercooled cloud, but it presupposes that such small ice particles exist. The formation of secondary ice *via* fragmentation or rime-splintering, each operating within a relatively narrow band of temperatures, generates new ice particles and so offers the main hope for explaining rapid glaciation and high ice concentrations.

Ice multiplication has been offered as a mechanism for the rapid transformation of some supercooled convective clouds to ice. As the name implies, the number concentration of existing ice is increased by some multiplier greater than unity. The secondary-ice formation mechanisms do this naturally, by definition, but the real power of multiplication emerges when the secondary ice can itself induce additional secondary ice. The stage is then set for the possibility of regenerative feedback and the exponential buildup of the ice population.

A simple conceptual model, based on the rime-splintering mechanism of secondary-ice generation in a way first developed by Chisnell and Latham (1976), helps illustrate the principle of ice multiplication. As suggested by the schematic diagram of Fig. 14, cloud particles may be classified broadly as small or large, liquid or solid. The "small" particles are the supercooled liquid cloud droplets or pristine (unrimed) ice crystals originating by either primary or secondary mechanisms. Such particles are typically some tens of microns in size, have small fallspeeds, and therefore follow the air motions. The "large" particles, often greater than 500 μm in diameter, are either supercooled rain drops, that likely grew by the collision–coalescence process within the "warm" cloud below the 0 °C level, or frozen raindrops that subsequently serve as riming centers (*i.e.*, graupel particles) that accrete the small cloud droplets. Given the difference in fallspeeds of the raindrops and the small ice particles, it is inevitable that some "capture" events will take place whenever these two classes of particles coexist in the same volume of cloud. Then, if the temperature is between

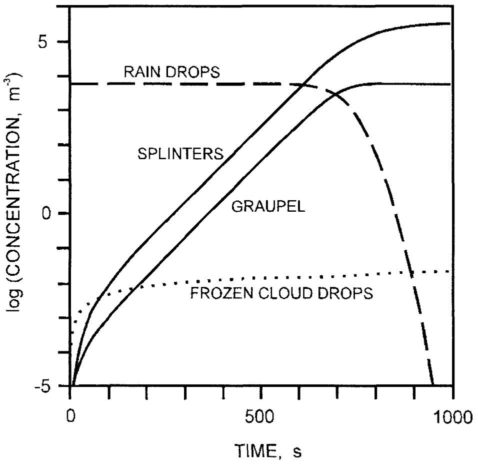

FIG. 15. Computed concentrations from the glaciation model employed by Lamb *et al.* (1981).

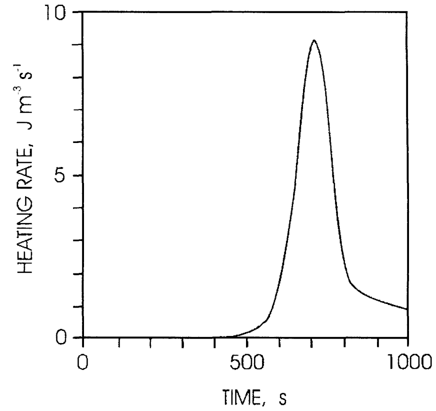

FIG. 16. Computed heating rates evolved during the buildup of the ice shown in Fig. 15. Most of the heating arises from the freezing of the large supercooled rain drops depicted in Fig. 14.

−3 and −8 °C, secondary ice ("splinters") would likely be released during the riming growth of the newly frozen raindrops. In this way, a cycle is set up (curved arrow in Fig. 14) whereby splinters are captured by the raindrops, which then freeze rapidly and throw off a new generation of splinters to perpetuate the process until all of the raindrops have been exhausted.

A straightforward analysis of the interactions depicted in Fig. 14 can be used that shows how the population of ice particles builds up with time when the dynamic structure of the cloud permits the "large" ice and liquid drops to remain for a time in the active secondary-ice production zone (between the −3 and −8 °C levels). The result of calculating the rates of interactions among the various cloud particles is shown in Fig. 15. The concentrations of both the splinters and the graupel are seen to increase exponentially until the process is cut off by the depletion of the raindrop population. Note that the initially all-liquid cloud was completely transformed into an ice cloud within less than about 12 min. Such an exercise demonstrates the complicated nature of cloud processes and the power of ice multiplication processes when opportunities for positive feedback exist.

Ice formation in supercooled clouds has a number of consequences. In some cases, the effects are immediate and strong; in other regards, they may be more persistent and perhaps subtle. When the glaciation proceeds rapidly, as when positive feedback processes are active, the sudden change in phase strongly affects the energy budget of the cloud: The latent heat released, particularly that due to the freezing of supercooled rain drops, leads to a warming of the cloudy air and an increase in the buoyancy of the cloud at mid levels. Results of the computations of the heating rate from the same glaciation model that was just considered are shown in Fig. 16. Because of the exponential nature of the glaciation (refer again to Fig. 15), significant heating appears only near the end of the process. A pulse of heating then results, the rate of which is comparable to that from the formation of the liquid cloud in the first place. Such large pulses of heating and buoyancy generation can occur naturally, or intentionally through cloud seeding, and lead to the "explosive" growth of convective clouds. More frequently, less dramatic effects occur, but the circulations of the cloud are stimulated nevertheless, often at an important stage in its development.

Precipitation of all types, both liquid and solid, is frequently initiated by the formation of ice in supercooled clouds. At one time, mainly during the middle decades of the twentieth century, it was thought that virtually all precipitation in the temperate zones resulted from the so-called cold-cloud or ice-crystal mechanism. Because of the inherent difference in vapor pressure of the solid and liquid phases of water, any ice particles that form in the presence of supercooled droplets grow rapidly by vapor deposition at the expense of the droplets. This phase change in effect destabilizes the colloidal nature of the cloud, permitting water mass to be transferred from the many small, nonsedimenting cloud droplets to a relatively few ice particles that can fall effectively against the updraft and later gather additional water by accretion. This mechanism of precipitation initiation appears to be important in many clouds, but we now recognize that the growth of precipitation by the all-liquid collision–coalescence mechanism is important for producing precipitation itself under some meteorological conditions, as well as for setting the microphysical stage for cloud glaciation.

4.2 Electrical and Chemical Phenomena

Cloud electrification is another likely consequence of ice formation, especially when it occurs in the large convective cells of summertime thunderstorms. Although the details by which the observed, vertically oriented dipole arises in convective clouds are still uncertain, it appears that collisions between graupel particles and relatively unrimed ice crystals play a fundamental role. The electrical nature of the surfaces of these two classes of particles is likely to differ, so that small amounts of electric charge would be transferred from one particle to the other when they collide and rebound without coalescing. Laboratory studies have shown that the sign of this microphysical charge transferred depends on the temperature, the liquid water concentration, and the degree of chemical impurities in the environment. For typical clouds with about $1\,\mathrm{g\,m^{-3}}$ of liquid water, the graupel particles gain negative charge and the ice crystals gain positive charge when the temperature is below about $-20\,^{\circ}\mathrm{C}$ (Saunders, 1993). Both the frequency of collisional interactions and the rate of macroscopic separation of the charges within the cloud are aided by the relatively large difference in fallspeeds of these two types of particles. The upper, glaciated reaches of thunderstorms thus become positively charged relative to the lower, mixed-phase portion of the cloud. With sufficiently vigorous convection, the potential difference between these charge centers, or between one of them and ground, can lead to electric breakdown of the air and the violent discharge known as lightning.

Interactions between ice and the various trace chemicals in the air can have potentially important consequences. Some electrification studies have implicated contaminants in the exchange of charge during collisions between ice crystals and graupel particles, but the mechanism is unclear. The supercooled cloud droplets accreted by the graupel particles invariably contain ions acquired at the time of their formation and subsequently during in-cloud scavenging events, and so it is logical to expect the chemical composition and therefore the electrical nature of graupel and vapor-grown ice to differ. However, past studies have indicated that the chemical effects are weak. Some researchers, by contrast, argue that the "contact potential" that may be responsible for the charge transfer on the particle scale depends only on the growth rates and temperatures of the colliding surfaces. Much research is needed on the nature of the ice surface and how it interacts with atmospheric trace chemicals before cloud electrification is understood.

Other interactions of trace chemicals with ice may have other, more direct impacts, however. Recent studies indicate that some trace gases, such as HNO_3 and SO_2, adsorb onto the ice surface over a broad range of time scales. In some cases, the ice surface seems to serve as a catalytic medium for reactions as well, leading to possible roles for heterogeneous reactions involving ice in the chemistry of the upper troposphere and lower stratosphere. This area of research is in its infancy and should be worth watching.

4.3 Interactions with Radiation

By interacting with electromagnetic radiation, atmospheric ice gives rise to numerous optical phenomena and affects the radiation budget of the Earth. The beautiful 22° halo frequently seen around the sun or moon, for instance, occurs as light is refracted through two of the prism faces of regular hexagonal ice crystals in thin ice clouds. Some optical phenomena, among them the familiar parhelia or "sun dogs", provide the observer with unique information about the type and orientation of the crystals contributing to the display. In this case, the crystals tend to be plates with their principal axes vertical. Such optical displays arise because the regular geometry of the crystals causes solar radiation to be scattered in well-defined directions.

More commonly, however, the ice-particle morphology is not so regular, and the incoming radiation is scattered more diffusely. Some fraction of the solar energy scattered from thin ice clouds returns to space, but most of it is forward-scattered toward the earth, so that the impact of cirrus clouds on the short-wave radiation budget is minimal. At infrared wavelengths, on the other hand, high-level ice clouds can have substantial influence. Ice readily absorbs upwelling terrestrial (infrared) radiation, but the re-emission of thermal radiation is weak at the low temperatures found in the upper troposphere. High ice clouds therefore tend to contribute to a general warming of the earth in the same sense that the radiatively active trace gases do.

5. SNOW AND ICE STORMS

Snow and ice storms represent one of the most direct intersections of atmospheric ice with humans. When ice particles in the atmosphere grow large enough to fall to the surface as snow, the impacts may be as benign as a beautiful blanket of white over the landscape or as devastating as a raging blizzard. Large snow storms may yield millions of tons of snow that clog highways, close airports, and generally create havoc on human activities. Floods may ensue once warmer weather arrives and cause another bout of devastation. On the positive side, however, snow that falls in mountainous terrain may persist well into the warm season and supply much-needed water to urban areas during drier times of the year. As much as half of the land mass of the northern hemisphere may become covered with snow each winter, generating a seasonal snowpack that reflects a significant fraction of the incoming solar radiation and thus affects the Earth's climate. The storms that give rise to such effects are described in a general way here, first from a meteorological perspective, then in terms of the microphysical processes discussed earlier.

5.1 Meteorological Settings

Snow and ice storms typically result from cyclones that form during wintertime in the mid latitudes. Driven by the sharp contrasts in temperature that develop across the polar front, these storms convert atmospheric potential energy into kinetic energy and serve to reduce the meridional temperature gradients of the earth. In a typical sequence of events, waves of relatively short wavelength develop in the airflow (jet stream) of the upper troposphere and cause the surface pressure to decrease beneath the poleward-flowing branch of the wave. The cyclonic (counterclockwise in the northern hemisphere) circulation that is established about the center of low pressure causes warm air to flow poleward and cold air to flow equatorward. At this point, the polar front is transformed into well-defined warm and cold fronts that radiate out from the center of the low-pressure system. Because of density differences in the cold and warm air masses, vertical components to the air motions are established. Adiabatic cooling of the air during upward motion leads to supersaturation development and cloud formation along both fronts. The latent heats of condensation and freezing that are released from the changes in phase augment the circulations and help maintain the energy of the cyclone.

Although mid-latitude cyclones exhibit a life cycle lasting several days and so cannot be treated as steady-state phenomena, a generic structure may be described that pertains to the system during the mature stage, the period during which the warm and cold fronts remain distinct. Figure 17 depicts the airflow pattern and conveys the three-dimensional structure of a typical storm during this stage. The relatively warm, moist air arising in the warm sector is lifted up and over the cooler air flowing beneath the warm front, by virtue of its relatively low density. This airstream, the so-called "warm conveyor belt", provides the cyclone

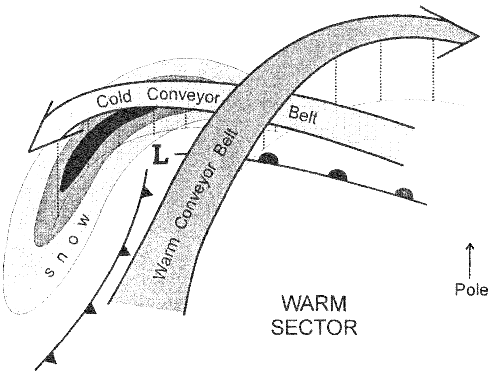

FIG. 17. The airflow pattern during the mature stage of a typical mid-latitude cyclone. Based in part on the study of Browning (1983).

with a steady supply of moisture for cloud and precipitation formation. Poleward of the warm front, the "cold conveyor belt" brings a steady stream of surface air beneath the warm conveyor belt before it, too, ascends over the still colder air behind the cold front. In some situations, such as those of the "Nor'easters" that frequent the east cost of North America, the air of the cold conveyor belt can be relatively very moist, so contributing greatly to the precipitation of the system.

5.2 Formation of Snow and Ice

Precipitation forms in the ascending branches of cyclonic systems by the same microphysical processes described in Sec. 3. As moist air cools by adiabatic ascent, such as in the warm and cold conveyor belts shown in Fig. 17, supersaturations develop and cause cloud to form. The temperatures vary nonuniformly in the vertical, first decreasing with altitude (beneath the warm front), then increasing in a narrow band through the front before again decreasing with altitude in the warm overrunning air. The warmest air is thus found in a layer aloft, about where the cloud base is located. Still, in a typical wintertime situation, this maximum temperature is near 0 °C, so that most of the cloud is supercooled. The coldest parts of the system are found near cloud tops. Especially along the warm conveyor belt, the air can become convectively unstable on account of the lifting of relatively warm, moist air beneath the colder, drier air of the upper troposphere. Convective towers form on top of the otherwise smooth, stable flow and lead to localized regions of very large supersaturations and high probabilities of ice nucleation.

As primary ice particles form, they grow rapidly by vapor deposition due to the large difference in the equilibrium vapor pressures of ice and liquid water. This "Bergeron process" leads to relatively large single crystals that sediment against the updraft and grow further by aggregation and accretion. The relative contributions of the various growth mechanisms to the overall mass gain of the snow depend on the relative abundances of liquid and ice particles, as well as on the geometry of the storm-scale flow patterns. In some cases, it appears that a "seeder–feeder" mechanism leads to the efficient production of snow when the vapor-grown ice crystals (seed particles) formed in the warm conveyor belt fall into the underlying cold conveyor belt, where they continue their growth by collection (Browning, 1983). The growth of the snow particles in the lower, feeder cloud is particularly effective when the air of the cold conveyor belt is forced to rise against mountainous terrain.

The distribution of snow and ice formed in a wintertime cyclone depends on a number of factors. As a rule, the maximum rate of snow formation occurs where the vertical flux of moisture is maximum. Thus, where moist air ascends rapidly, as where it crosses a frontal boundary or where it encounters a topographic barrier, the rate of condensate and precipitation formation will be large. The total accumulation on the ground, however, depends also on the wind speed and the movement of the storm as a whole. Relative to the moving system, the main region of snowfall typically appears in a crescent-shaped band behind the cold front and wrapping around the center of low pressure. If the temperatures are significantly below zero everywhere, then one can expect all precipitation to arrive at the ground in the form of snow; otherwise one may encounter rain in some parts, possibly freezing rain in other parts. The favored location for freezing rain is on the cool side of the warm front, where the lowest layer of air can be below the freezing point at the same time that the warmer air above the sloping frontal zone is above 0 °C. Such situations are not uncommon in the northeastern United States of America and are exacerbated by the evaporative cooling of precipitation falling below the cloud base. Only minor changes in the temperature profile above a given location are sufficient to alter the form of precipitation, but the consequences to humans can be profound.

6. CURRENT UNCERTAINTIES AND RESEARCH AREAS

As can be appreciated from the discussion of snowstorms, precipitation on the ground represents the culmination of a number of meteorological and microphysical processes operating on a hierarchy of spatial and temporal scales. The processes interact with one another, so that precipitating systems tend to be complicated. Predicting the arrival of a new weather system, or its consequences, is difficult in part because the complexities in the systems tend to make them sensitive to the initial conditions. Especially in the case of freezing rain, considerable precision is needed in forecasts because of the sudden change in the phase of water near 0 °C. Through ongoing research efforts, considerable progress is being made in the area of regional forecasting of wintertime precipitation. Similar arguments apply to the prediction of hail storms, but there the problem is further complicated by the small scales of thunderstorms and our incomplete understanding of the evolution of hail in a convective environment.

Major uncertainties still exist in understanding some of the microphysical processes of ice and snow formation. Whereas we can categorize processes and have general understandings of how they operate, details are lacking in such areas as the concentrations of ice nuclei and their modes of action, the causes of the peculiar variation of crystal habit with temperature, the mechanistic basis of secondary ice formation, and quantitative relationships of atmospheric ice concentration to rainfall from cumuliform clouds. Both the warm-cloud and cold-cloud mechanisms of precipitation formation occur simultaneously, but we cannot yet quantify their relative contributions to precipitation amount. Both field programs and computer modeling serve to enhance our understandings of cloud-scale phenomena, but often laboratory experimentation is needed to resolve the details of specific microphysical mechanisms. Of particular importance to the advancement of our understandings on the particle scale are experimental studies aimed at characterizing the molecular nature of the ice surface under realistic atmospheric conditions. The findings from such research would undoubtedly yield insights into the reason for the variation of deposition coefficients and habits with temperature, for instance, as well as provide a foundation upon which to explain the phenomenon of charge transfer during collisions between ice particles. Chemical interactions, too, require detailed research at the molecular level. It is clear that, despite much progress in recent years, diverse methods of research are still needed before we really understand atmospheric ice.

GLOSSARY

Bergeron Process: The microphysical mechanism by which water is transferred *via* the vapor phase from liquid drops to ice particles that then serve as embryos for precipitation. Named after Tor Bergeron (1891–1977), the Swedish meteorologist who recognized that the difference in phase leads to the colloidal instability of clouds.

Cold Cloud: The generally upper parts of a cloud containing ice particles.

Cold Front: The leading edge of an advancing cold air mass.

Dendrite: A platelike crystal grown from the vapor that exhibits a treelike branching structure.

Freezing Rain: Rain that falls as supercooled liquid drops and freezes upon contact with the ground, resulting in a coating of relatively clear ice (glaze).

Graupel: Snow pellets, soft ice particles up to 5 mm in diameter formed by the accretion of supercooled droplets.

Hail: Precipitation in the form of irregular lumps of ice greater than 5 mm in diameter that grew partially by the "wet" accretion of supercooled droplets.

Polar Front: The quasistationary boundary between cold, polar air and the warmer air toward the equator.

Sleet: Precipitation formed by the freezing of supercooled rain before contact with the ground.

Troposphere: The lowest 10–20 km of the atmosphere, characterized by decreasing temperature with increasing height and vertical air motions that contribute to the formation of clouds and weather.

Ventilation Coefficient: A factor to account for the motion of a particle relative to the air. In the case of crystal growth from the vapor, for example, the mass ventilation coefficient is the ratio of the actual vapor flux toward a sedimenting crystal to the flux toward a similar, but stationary, crystal.

Warm Cloud: The lower, warmer parts of a cloud containing only liquid drops.

Warm Front: The leading edge of an advancing warm air mass.

Works Cited

Bentley, W. A., Humphreys, W. J. (1962), *Snow Crystals*, New York: Dover.

Browning, K. A. (1983), *Quart. J. R. Meteor. Soc.* **109**, 225–242.

Chen, J. P., Lamb, D. (1994), *J. Atmos. Sci.* **51**, 1206–1221.

Chisnell, R. F., Latham, J. (1976), *Quart. J. R. Meteor. Soc.* **102**, 133–156.

Fukuta, N., Lee, H. J. (1986), *J. Atmos. Sci.* **43**, 1833–1843.

Hobbs, P. V. (1974), *Ice Physics*, Oxford: Clarendon Press.

Lamb, D., Hallett, J., Sax, R. I. (1981), *Quart. J. R. Meteor. Soc.* **107**, 935–954.

Lamb, D., Scott, W. D. (1974), *J. Atmos. Sci.* **31**, 570–580.

Mace, G. G., Ackerman, T. P., Clothiaux, E. E., Albrecht, B. A. (1997), *J. Geophys. Res.* **102**, 13581–13593.

Magono, C., Lee, C. W. (1966), *J. Fac. Sci. Hokkaido Univ.* Ser. VII **2**, 321–335.

Meyers, M. P., DeMott, P. J., Cotton, W. R. (1992), *J. Appl. Meteor.* **31**, 708–721.

Nakaya, U. (1954). *Snow Crystal, Natural and Artificial*, Cambridge: Harvard University Press.

Ono, A. (1970), *J. Atmos. Sci.* **27**, 649–658.

Pruppacher, H. R., Klett, J. D. (1997), *Microphysics of Clouds and Precipitation*, Dordrecht: Kluwer Academic Pub.

Reynolds, D. W., Dennis, A. S. (1986), *Bull. Am. Meteor. Soc.* **67**, 513–523.

Ryan, B. F., Wishart, E. R., Shaw, D. E. (1976), *J. Atmos. Sci.* **33**, 842–850.

Saunders, C. P. R. (1993), *J. Appl. Meteor.* **32**, 642–655.

Weickmann, H. K. (1945), *Beitr. Phys. freien Atmos.* **28**, 12–52.

Weickmann, H. K. (1949), *Ber. Deutsch. Wetterdienst* No. 6, 1–54.

Weickmann, H. K. (1957), *Meteor. Res. Rev.*, **3**, 226–255.

Wergin, W. P., Rango, A., Erbe, E. F. (1995), *Scanning* **17**, 41–49.

Yokoyama, E. (1993), *J. Cryst. Growth* **128**, 251–257.

Young, K. C. (1993), *Microphysical Processes in Clouds*, New York: Oxford University Press.

Further Reading

Braham, R. R. (1996), in: J. R. Fleming (Ed.), *Historical Essays on Meteorology, 1919–1995*, Boston: American Meteorological Society, pp. 181–223.

Greenler, R. (1980), *Rainbows, Halos, and Glories*, Cambridge: Cambridge University Press.

Hallett, J. (1984), *Am. Scientist* **72**, 582–589.

Hobbs, P. V. (1974), *Ice Physics*, Oxford: Clarendon Press, Chaps. 8 and 10.

Pruppacher, H. R., Klett, J. D. (1997), *Microphysics of Clouds and Precipitation*, Dordrecht: Kluwer Academic Publishers.

Rogers, R. R., Yau, M. K. (1989), *A Short Course in Cloud Physics*, Oxford: Pergamon Press.

Schaefer, V. J., Day, J. A. (1981), *A Field Guide to the Atmosphere*, Boston: Houghton Mifflin.

Weickmann, H. K. (1981), *PAGEOPH* **119**, 538–547.

BIOMIMETIC PROCESSING OF CERAMICS

Liisa T. Kuhn, *Laboratory for the Study of Skeletal Disorders and Rehabilitation, Children's Hospital, Harvard Medical School, Boston, Massachusetts, U.S.A.*

Arthur H. Heuer, *Department of Materials Science and Engineering, Case Western Reserve University, Cleveland, Ohio, U.S.A.*

Introduction 27
1. **Conventional Ceramic Processing** 28
1.1 Powder Processing 28
1.2 Thin-Film Deposition and Layered Materials 29
2. **Ceramic Processing in Nature** 29
2.1 Examples of Biomineralization 29
2.1.1 Bone Structure and Formation 29
2.1.2 Seashell Structure and Formation 32
2.1.3 Eggshell Structure and Formation 34
2.2 General Biological Processing Principles 36
3. **Biomimetic Ceramic Processing** 37
3.1 Compartmental Processing of Particles and Particulate Networks 37
3.1.1 Nanoparticles from Single Compartments 37
3.1.2 Nanoparticle Networks 38
3.2 Specific Mineralization (Application to Thin Films) 38
3.2.1 Self-Assembled Monolayers 38
3.2.2 Biomimetic Lithography 40
3.2.3 Langmuir–Blodgett Films 40
3.3 Net Shape Formation with Layered Architecture 41
3.4 Living Systems Used for Materials Synthesis 42
4. **Conclusions** 43
Acknowledgements 43
Works Cited 43
Further Reading 45

INTRODUCTION

Biomimetic ceramic processing is a relatively new field within materials engineering that seeks to imitate or mimic the procedures Nature has used to produce biological ceramics such as bones, eggshells, and seashells. The interest in biomimetic ceramic processing stems primarily from an admiration of the complex shapes and remarkable physical and chemical properties displayed by biological ceramics, given their generally mundane chemical makeup. For example, seashells, which contain 98–99% calcium carbonate, can be twice as hard and several thousand times more fracture-resistant than nonbiological calcium carbonate (Laraia and Heuer, 1990; Kuhn *et al.*, 1996a, Jackson *et al.*, 1988). The exceptional properties of biologically produced ceramics can be directly correlated to the complex microstructures (microarchitectures) created by organisms through a highly controlled and precise sequence of assembly. In the case of the seashell, for example, the high toughness is directly related to structure on several levels, a hierarchical structure. While shell microstructures vary between marine organisms (Bøggild, 1930), the toughest materials exhibit a laminated multi-layered microarchitecture containing structure on five length scales; each macroscopic layer is composed of microscopic layers that are in turn composed of even finer layers. The macroscopic toughness is increased because the layers act to deflect cracks, and thus dissipate the energy associated with incipient mechanical instability among many fine-scale noncatastrophic microcracks (Laraia and Heuer, 1990; Kuhn *et al.*, 1996a). A technological advancement potentially capable of improving the toughness of ceramics by several orders of magnitude is especially significant; it may further the acceptance of ceramics

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

by structural designers in a variety of advanced applications, a long-held goal.

Biological ceramics are laminated microcomposites, with each layer of biomineral separated from its neighbor by organic "matrix" constituents (primarily acidic proteins). These matrix proteins regulate and control the structure and composition of the biomineral phase during initial formation. It is this organic matrix control that forms the basis of biomimetic processing approaches. Other natural systems, in particular bacteria and unicellular organisms (Mann, 1996), can also provide inspiration for novel magnetic, electronic, and mechanical devices, and are briefly mentioned below.

Although many of the detailed steps involved in the formation of biological ceramics are not yet well understood, general processing paradigms have emerged that have significant implications for materials science. The three primary processing principles examined in this chapter are illustrated with examples involving the formation of bones, seashells, and eggshells. In turn, examples of how each principle has been used to develop a biomimetic ceramic processing approach will be discussed. From the description of property improvements available from a biomimetic approach, biomimetic processing of ceramics appears to hold much promise.

1. CONVENTIONAL CERAMIC PROCESSING

In order to appreciate the recent and potential advances of biomimetic ceramic processing, the conventional methods of producing ceramic components are first briefly reviewed.

1.1 Powder Processing

Ceramic substances of interest here (refractory oxides, nitrides, carbides, and similar materials) cannot easily be melted and cast into near net-shape components because of their high melting point; therefore, powder processing approaches are the rule. Furthermore, their hard and brittle nature makes machining of ceramic components to exacting tolerances difficult and expensive, and so net-shape processes are preferred.

The processing of ceramics generally begins with formation of a powder compact, and the sintering shrinkage that invariably accompanies the subsequent densification of powder compacts must be accounted for at this stage. Basic consolidation methods used to form complex shapes at ambient temperatures from ceramic powders include dry powder pressing, isostatic pressing, injection molding, extrusion, slip casting of slurries, wet plastic forming, gel casting, or cementatious bonding. With the exception of cementatious bonding, consolidation is followed by a densification step at elevated temperatures (pressureless sintering, hot pressing, or hot isostatic pressing), which "fuses" or "bonds" the individual powder grains and allows pore shrinkage. The resulting sintered body is a cohesive, dense (or nearly dense) monolithic structure.

Care must be employed to avoid introduction of deleterious defects during powder processing and subsequent finishing to final dimensions. Common ceramic processing concerns include control of residual porosity and other two-dimensional or three-dimensional defects that may weaken the part, as well as differential shrinkage, which can introduce residual stresses and crack-like defects. To obtain the high strength and reliability necessary for advanced or "high-tech" non-clay-based ceramics for structural or electronic applications, it is necessary to have stringent control of the size and shape of the starting powders. Organic processing additives—dispersants, deflocculants, binders, surfactants, *etc.*—that improve powder flow and packing, and allow for molding into "green" (unfired) parts with good handling characteristics, are commonly employed. A mixture of powder sizes packs most efficiently, provided the particles exhibit good flowability. Ceramic powders may be obtained from grinding and milling of natural raw materials, but control of particle size and shape is more usually achieved by a variety of chemical powder synthesis methods. Precipitation from aqueous and nonaqueous solutions, thermal decomposition of inorganic salts, hydrolysis of alkoxides and other organometallics, sol–gel techniques, pyrolysis of polymeric precursors, flame hydrolysis, hydrothermal techniques, and condensation from a vapor phase are commonly used to obtain micrometer- and submicrometer-sized powders.

Detailed descriptions of conventional powder synthesis can be found in standard references (Messing *et al.*, 1987; Segal, 1989). Some of these methods are relatively expensive and do not lend themselves to large-scale production.

Colloidal technologies are an important component of modern ceramic processing approaches. These usually involve the use of suspended particles in a slurry containing various additives. The aim is to control interparticle forces with adsorbed charged additives to achieve stable suspensions, which lend themselves to production of high green densities during subsequent consolidation. [Without additives, interparticle interaction from van der Waals forces can cause uncontrolled aggregation that effectively increases the particle size and leaves undesired voids in the finished product (Lange, 1983).]

1.2 Thin-film Deposition and Layered Materials

There is much interest in the formation of thin films of ceramic materials for electronic, magnetic and optical devices, as well as for thermal, corrosion and abrasion protection. Important thin-film materials include high-T_c oxide superconductors, oxide and nitride dielectric interlayers, nonmetallic magnetic films, and ferroelectric films for electro-optical and memory devices (Ohring, 1992). The conventional methods for formation of thin ceramic films on substrates include sol–gel deposition or other solution techniques and vapor-phase techniques including plasma spray deposition, molecular-beam epitaxy (*q.v.*), laser ablation, sputtering (*q.v.*), and chemical vapor deposition [see CVD (Chemical Vapor Deposition)] (Feldman and Mayer, 1986).

There are a number of limitations associated with conventional thin-film deposition methods. In order to control physical and chemical properties of thin films, it is necessary to control film density, crystal orientation, grain size, degree of crystallinity, and thickness. This is difficult with sol–gel and colloidal processes because they generally undergo significant densification during final sintering. Furthermore, the high processing temperatures prohibit the use of polymer substrates. Vapor-phase deposition methods usually require expensive fabrication systems and deposition rates are very slow. Crystal orientation is very important because many magnetic, ferroelectric, and superconducting properties are anisotropic, yet the choice of substrates for oriented growth is very limited. These limitations provide a strong motivation for the development of biomimetic thin-film processing.

2. CERAMIC PROCESSING IN NATURE

There are a myriad of fascinating biominerals found in Nature, with complex shapes and microstructures and equally fascinating and complex methods of fabrication (Omori and Watabe, 1980; Nancollas, 1982; Mann *et al.*, 1989; Lowenstam and Weiner, 1989). The structure and formation of the most widely studied biological ceramics—mammalian bones, eggshells, and seashells—are described in this section. We illustrate the complex microarchitectures and their development, which demonstrate an exquisite level of control of the spatial regulation of both the nucleation and growth of the biomineral. Given the relatively modest properties of the basic building materials, the resultant structures demonstrate the remarkable properties that are possible with cell-mediated, organic matrix-modulated assembly.

2.1 Examples of Biomineralization

2.1.1 Bone Structure and Formation Much study has been devoted to the structure of mammalian bone, primarily motivated by biomedical applications (Glimcher, 1959; Pechak *et al.*, 1986; Giraud-Guille, 1988; Landis *et al.*, 1993; Kim *et al.*, 1995; Robey and Boskey, 1996; Ziv *et al.*, 1996); bone development is thus a natural source of inspiration for biomimetic ceramic processing. Bone is a complex material with unique architectures at many different length scales that are customized for the several functions that bone must provide (Fig. 1). In addition to the obvious structural applications, bone acts as an ion reservoir for maintaining chemical homeostasis and as a repository for healing cells (Glimcher, 1959; Pechak *et al.*, 1986; Robey and Boskey, 1996).

Macroscopic features of note include dense,

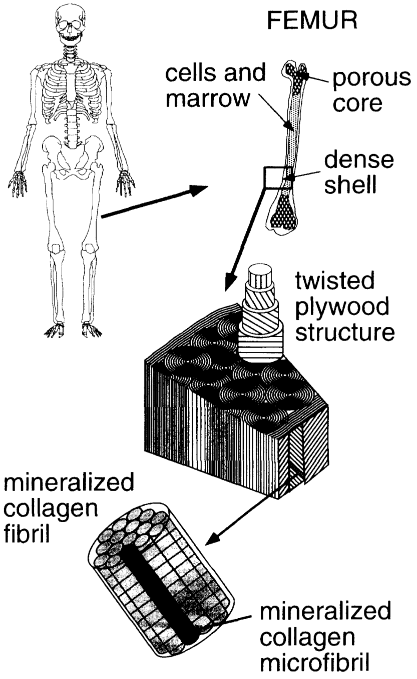
FEMUR
cells and marrow
porous core
dense shell
twisted plywood structure
mineralized collagen fibril
mineralized collagen microfibril

FIG. 1. Bone structure, part I. Bone is a complex material with unique architectures at many different length scales. The long bones such as the femur have a dense shell of cortical bone with a central cell-filled marrow cavity. Porous trabecular bone reinforces the ends of the bone. A twisted plywood structure can be observed at the microstructural level, with each layer comprised of lamellae of mineralized collagen fibrils, which themselves are comprised of mineralized microfibrils.

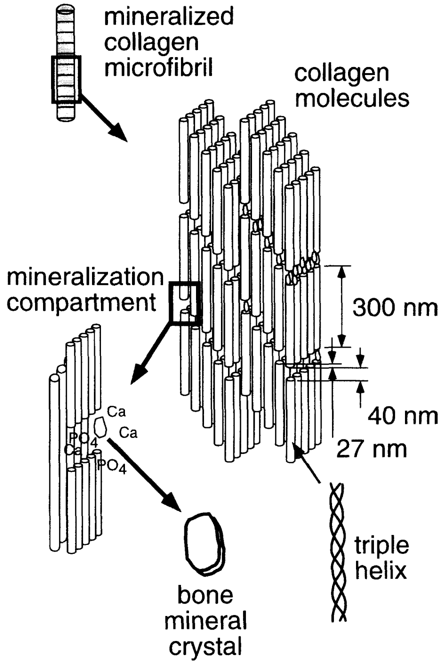
mineralized collagen microfibril
collagen molecules
mineralization compartment
300 nm
40 nm
27 nm
Ca
Ca
PO4
Ca
PO4
bone mineral crystal
triple helix

FIG. 2. Bone structure, part II. The hierarchical microarchitecture of bone extends to the nanometer level. The mineralized microfibrils are composed of many collagen molecules arranged in an overlapping staggered array. The mineral phase of bone nucleates in the gaps or hole zones between the ends of the collagen molecules, which also have a complex architecture: a triple helix of three strands of protein.

but hollow, shells of cortical bone enclosing cell-filled marrow cavities, and larger, broader ends of bone reinforced with porous trabecular bone to absorb mechanical energy and optimize weight transfer and friction across joints (Fig. 1). The microarchitecture is one of repeating layers of mineral-reinforced bundles of structural fibrillar proteins, offset in a twisted plywood structure (Giraud-Guille, 1988; Ziv *et al.*, 1996) (Fig. 2); the mineral is calcium phosphate in the form of carbonated apatite containing other ions such as HPO_4 and Na and Mg (Kim *et al.*, 1995), *not* hydroxyapatite as is frequently suggested in the literature. Each layer is made up of sheets containing collagen fibers arranged side by side. Each fiber is actually a bundle of layered microfibrils that contain layered collagen molecules in a triple helix assembly (Fig. 2); the mineral "reinforcement" occurs at the microfibrillar level. In addition to apatite (65 wt%) and collagen (30 wt%), there are a variety of other proteins, proteoglycans, and glycoproteins present in bone that are well described by Robey and Boskey (1996).

Materials selection charts that allow a rapid selection of the optimal material for a given application indicate that the best strength/modulus ratio can be obtained when the size of the reinforcement decreases (Ashby, 1993). Bone mineral, with its nanometer-sized apatite crystals, represents a near-ideal configuration that no synthetic system has yet attained.

In long bones like the femur, bone formation is accomplished by cellular activity around a structural precursor composed of cartilage (Fig. 3). During embryogenesis, cells first de-

(a) Cartilage model

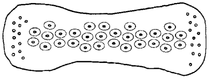

(b) Pre-bone cells form a layer around the model

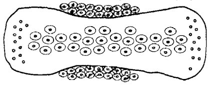

(c) Bone cells secrete mineralizing organic matrix

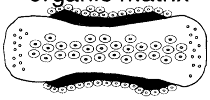

(d) Bony collar starves cells inside, degradation begins.

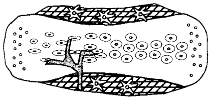

(e) Finished bone with cartilage on ends and marrow within.

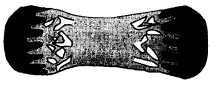

FIG. 3. Bone formation begins with the formation of a cartilage model that defines the macroscopic dimensions of the bone. Pre-bone cells form a layer completely around the model and then differentiate into bone-producing cells, osteoblasts. A bony collar is formed around the model when the matrix secreted by the cells mineralizes. An ingrowth of blood vessels brings in osteoclasts, which degrade the cartilage and open a cavity for marrow cells and bone trabeculae formation.

posit a collagen-based cartilage form that defines the dimensions of the bone, and then assemble in layers around it (Pechak *et al.*, 1986). The cells then secrete a different, bone-specific collagen that self-assembles into the fibrillar sheets described earlier that have compartments in which mineralization occurs. Prior to mineralization, the cells secrete matrix vesicles that contain high concentrations of calcium and phosphorus throughout the collagen network (Wuthier, 1997). There are acidic proteins [*e.g.*, bone sialaprotein (Kinne and Fisher, 1987)] within the compartments of collagen that are believed to concentrate calcium ions. Mineralization commences when enzymatic processes release phosphorus from the phosphoproteins in the surrounding milieu to effect locally high concentrations of phosphate ions. After a rigid, bony collar has formed around the cartilage preform, the cartilage is intentionally degraded by other cells that enter through the collar of bone and proceed to break down (decompose) the cartilage [Fig. 3(d)]. Vasculature, *i.e.*, the formation of blood-carrying vessels, develops and allows transport of the cartilage residues, emptying the central canal, which is then filled with marrow. The lamellar structure is not well defined upon initial formation, and this early structure is referred to as woven bone. It is only during "remodeling" that the lamellar structure develops; remodeling is a phenomenon whereby certain bone cells called osteoclasts resorb bone, and other cells called osteoblasts rebuild the bone.

Mineralization of bone involves the following coupled mechanisms:

1. Cells adjacent to the site to be mineralized secrete organic components that self-assemble into an organic collagen scaffold with uniformly sized compartments that subsequently mineralize.
2. A specific mineral of a defined shape in specific orientation is nucleated with assistance from acidic proteins bound to the structural proteins. Bone mineral crystals are unique in that they are nearly monodisperse, nanometer-sized plates of carbonated apatite of similar dimensions to the organic compartment. No other mineral phases are present and the crystallographic *c* axes of the apatite crystals are parallel with the collagen fibril axis.
3. Complex hollow and porous macroscopic shapes are generated by the repetition of self-assembly of structural proteins that arrange themselves in layered sheets with a twisted plywood structure around a predefined, space-holding preform. The cartilage preform is temporary, as it defines the macroscopic shape of the bone to be formed and is afterwards degraded. The layered architecture of bone hinders crack propagation and enhances toughness. Strict biomimesis of mineral–polymer composites based on the structure of bone would involve regularly spaced aligned filler particles

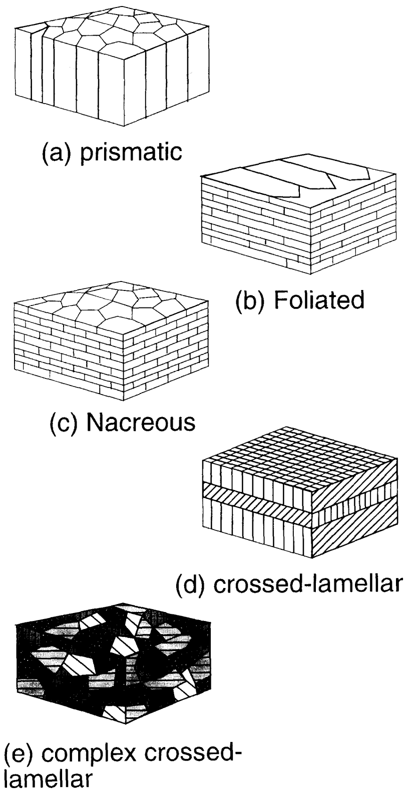

FIG. 4. Five common mollusk shell microstructures.

that are nanometer-sized plates present at a high volume fraction (more than 60% in bone).

2.1.2 Seashell Structure and Formation Most mollusk shells are calcium carbonate with a small amount of organic material (0.01–5 wt%), known as conchiolin. A variety of material architectures within mollusk shells have evolved, based on the environment of the marine organism. The currently accepted material architectures are still those defined by Bøggild (1930), several of which are illustrated in Fig. 4. The crossed-lamellar structure is the most specialized, and the most common within the phylum; it can be found in conch shells. The best known structure is mother of pearl, or nacre, the iridescent inner shell layer of oysters and abalone. These two microstructures are discussed in detail in this section.

The *Strombus Gigas* (pink conch) is a burrowing mollusk with an extremely hard and tough crossed-lamellar shell structure (Fig. 5). The microstructure within this shell consists of layers of aragonitic calcium carbonate at five length scales (macroscopic layers, first-, second- and third-order lamellae, and polysynthetic twinning at the 10 nm level), separated by a tenuous organic matrix present at less than 1 wt% (Taylor and Layman, 1972). The internal structure of the three macroscopic layers [each 2–5 mm thick and denoted i (inner) next to the organism, m (middle), and o (outer) in Fig. 5] are rotated by approximately 90° from one layer to the next. The first-order lamellae within layers are 25–40 μm thick and several mm long and are oriented such that adjacent lamellae are rotated by 70–90°. The second-order lamellae are 1–10 μm thick and 20–50 μm wide, and are aligned parallel within a first-order lamella; the third-order lamellae within them are also aligned. Each of the third-order lamellae are twinned single crystals of aragonite, 100–500 nm wide and less than 50 nm thick, the very-fine growth twins being about 10 nm wide (Laraia and Heuer, 1990). The aspartic-acid–rich organic matrix has been observed by TEM (transmission elecron microscopy) as a thin layer that envelops each of the third-order lamellae (Weiner, 1984).

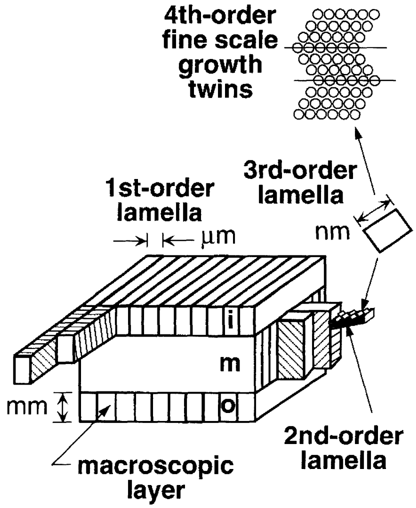

FIG. 5. The crossed-lamellar microstructure is composed of multiple layers of lamella at five length scales ranging from millimeter to nanometer thicknesses: i, m, and o denote the inner (closest to the organism), middle, and outer (exterior) layers, respectively.

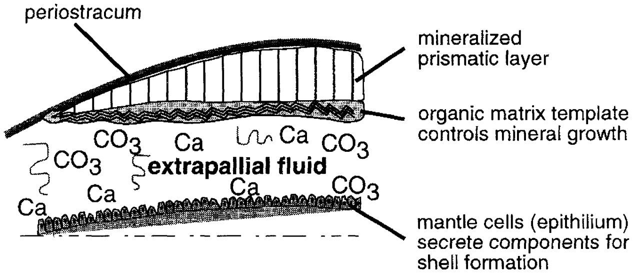
periostracum
mineralized prismatic layer
organic matrix template controls mineral growth
CO3
Ca
CO3
Ca
CO3
extrapallial fluid
Ca
Ca
CO3
Ca
mantle cells (epithilium) secrete components for shell formation

FIG. 6. Mollusk shell formation proceeds with the deposition of an organic scaffold, the periostracum, that acts as a substrate for the deposition of nucleating and growth controlling organic matrix molecules. The mantle cells secrete the extrapallial fluid, which contains all the components for biomineralization.

This highly specialized mollusk shell structure possesses a fracture resistance around 1000 times that of mineral aragonite, and a hardness substantially greater (factor of 2) than that of mineral aragonite (Taylor and Layman, 1972; Currey and Kohn, 1976; Laraia and Heuer, 1990; Kuhn *et al.*, 1996a). Fractographic observations of test specimens taken from mature shells reveals delocalized damage in the form of multiple microcracks, crack bridging, crack branching, and delamination cracking that greatly increases the total crack area and frictional dissipation of strain energy during fracture.

The shell of the abalone has a nacreous inner structure [Fig. 4(c)], which has 1–4 wt% organic material and is the strongest of the shell structures (Taylor and Layman, 1972; Currey and Taylor, 1974). The abalone shell also has an external layer of vertically oriented calcite crystals arranged in the prismatic structure [Fig. 4(a)]. In nacre, small platelets of aragonite are stacked in a brick-and-mortar fashion. The platelets in different layers are crystallographically related to each other, with the *c* axis normal to the shell surface and the *a* or *b* axis twisted relative to the layer (Sarikaya, 1994). The polypeptide chain directions in three different mollusk species are parallel to the direction of the *b* crystallographic axis of the aragonite crystals (Weiner *et al.*, 1983); this may be a general phenomenon.

Shell formation proceeds with secretion by the mantle cells of a nonmineralizing organic layer called the periostracum, that hardens and acts as a substrate for the subsequent deposition of proteins that form mineral templates (Gregoire *et al.*, 1955; Wilbur, 1973; Krampitz and Graser, 1988) (Fig. 6); the external crystal form and pattern of the biomineral conforms to the microtopography of the periostracum (Wilbur, 1973). In addition to the secretion of a group of matrix constituents with significant polysaccharide content, the ions necessary for mineral formation are also secreted into the space between the periostracum and the mantle tissue. This extrapallalial fluid has a severalfold supersaturation with respect to calcium carbonate, similar to the supersaturation of sea water. A mineral-confining envelope is defined by the self-assembly of freshly secreted matrix above the periostracum, after which nucleation and growth of mineral commences in the compartment. Crystal growth continues inwards towards the organism and normal to the shell until impingement upon the next matrix layer, and lateral growth continues until terminated by impingement with other mineral grains. The five-layered organic "mortar" of the nacreous structure is presumed to self-assemble above the growing crystals, and the beta-sheet acidic protein structure of the outer proteins, combined with the acidic proteins, is believed to concentrate calcium and induce oriented nucleation (Fig. 6) (Lowenstam and Weiner, 1989). Recent studies on nacre have shown that the mineral formation is in fact continuous from layer to layer and not separated completely by an organic matrix compartment (Morse *et al.*, 1993). In the crossed-lamellar structure, the organic mortar is only a thin layer of acidic proteins (Weiner, 1984).

The following are key features of biomineralization in marine organisms:

1. Cells adjacent to the site to be mineralized secrete organic matrix constituents that self-assemble against the organism into a structural organic scaffold. The texture of this structure defines the surface texture of the shell. There is subsequent secretion of additional matrix components that assemble

into uniformly sized mineralizing compartments. Mineralization occurs within the compartments, whose thickness thus limits the size of the crystals. The structure is approximately monodisperse—95% of the crystallites within seashells are within 30% of the average thickness of 0.5 μm.

2. Fixed polyanions assist nucleation by providing a negatively charged surface that interacts electrostatically with cations. The distribution of the polyanions on a structural organic scaffold is believed to provide a stereochemical template for adsorbing a layer of cations, such that a particular phase of mineral is nucleated with a specific crystal orientation.
3. Biominerals of specific shape are built by self-assembly of incremental units that assemble in concert in a predetermined order.

2.1.3 Eggshell Structure and Formation

The common avian eggshell represents another example of how careful materials processing can transform ordinary materials into exceptional articles—in this case the production of a seamless, elliptical, crack-resistant protective container of $CaCO_3$ for embryo development. Eggshell architectures are less complex than those of bone or seashell, which is probably related to the fact that they develop more rapidly than either bone or seashell (5 g of $CaCO_3$ are formed in 24 h). Despite this, the material serves several functions for the developing embryo—providing structural protection, acting as an ion reservoir (a source of Ca for development of the skeleton of the developing embryo), and providing continuous micropores for the ready transport of gases during respiration. In the schematic of the cross section of an eggshell (Fig. 7), large columnar grains of calcitic calcium carbonate can be seen to traverse the entire shell thickness. The calcite columns ($\approx 70\,\mu m$ in diameter and 200–300 μm thick) are oriented single crystals and show uniform birefringence and single-crystal electron-diffraction patterns (Dennis *et al.*, 1996). The size of the crystals is unusually large for crystals grown in a few tens of hours.

The columns of mineral, known as palisades, are attached to the collagen-based shell membranes that surround the yolk and white of the egg. There is both an outer and an inner shell collageneous membrane, with the outer membrane generally more porous and composed of larger-diameter collagen fibers than the inner membrane (Fig. 7). At the base of the columnar mineral structure is the so-called calcium reserve assembly (CRA), which is a cluster of organic vesicles that are calcium-rich and carbonate-poor (Dieckert *et al.*, 1989). The CRA is preferentially resorbed by the developing embryo as a calcium source during skeletal development (Dieckert *et al.*, 1989). The outermost layer of the eggshell, the cuticle, is $\approx 10\,\mu m$ thick and is a sugar-rich proteinaceous layer containing calcium phosphate needles (Dennis *et al.*, 1996) that cover the entire shell, sealing the biomineral structure against microbial invasion and excessive H_2O loss. Coloration is added to this layer in certain species.

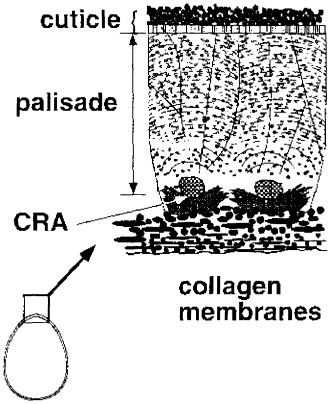

FIG. 7. Diagrammatic representation of the microstructure within the eggshell of a domestic chicken. Large columns of calcite form the palisade layer, which is anchored to the collagen membranes through the calcium reserve assembly (CRA) baseplates.

Although the avian eggshell is essentially a brittle material, quail eggs can undergo "graceful" failure and sustain loads after first cracking of as much as 70% of the load required to initiate fracture (Bond *et al.*, 1986). This behavior is related to the intermittent attachment between the compliant, fibrous membrane and the hard, brittle calcite columns. When growing microcracks reach the unattached regions, the crack is "blunted" as the load is spread over a much larger volume.

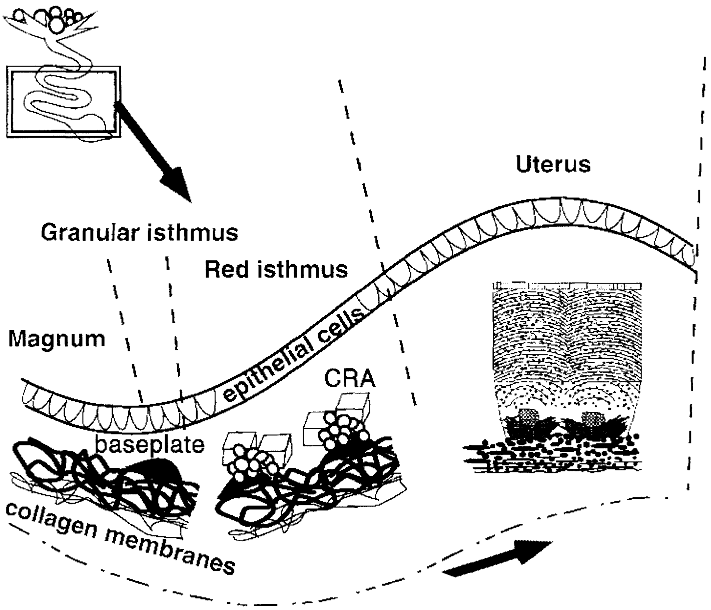

FIG. 8. Schematic partial cross sections of the avian eggshell at various stages of development in the oviduct. The bulk of eggshell crystal growth occurs in the uterus, where the egg remains for 20 h. Mineral nucleation occurs before this in the red isthmus on the organic structures known as baseplates, which are attached to the collagen membranes.

A schematic drawing of an avian oviduct, which serves as a kind of tubular bioreactor operating on a 24 h cycle, is shown in Fig. 8. During the passage of the egg through the oviduct, a number of extracellular matrix layers are sequentially processed onto the avian egg yolk by specialized secretory cells populating the subregions of the oviduct (Solomon, 1983). The first structural layers covering the egg, after albumen deposition in the magnum, are the inner and outer shell membranes. These membranes are secreted in the isthmus and self-assemble into random mats of collagenous fibers.

Mineralization of the eggshell is initiated in the final region of the isthmus (Creger *et al.*, 1976; Stemberger *et al.*, 1977). This process is first observed when the CRA particles, 1–10 μm in diameter, reach a maximum density of 2700 mm^{-2} on the outer surface of the shell membrane (Dieckert *et al.*, 1989). These particles are anchored in the membrane and serve as the bases of the large columnar calcite columns that develop to their full size while the egg rotates in the uterus.

In contrast to bone, the cells involved in eggshell formation are not intrinsic components of the shell; mineral deposition is effected by specialized oviduct cell populations in an "assembly-line" sequence as the egg is transported down the oviduct. The oviduct cells have more specialized control of distinct parts of eggshell fabrication—the cells in the isthmus portion of the oviduct secrete membrane components, and also the initial calcified structures upon which subsequent crystal growth proceeds. However, the bulk of the mineral phase is deposited in the uterus (the shell gland). Because the secretory cells do not become an intrinsic part of the eggshell, and because the secretory events occur in a relatively short time frame, there is only a limited portfolio of regulatory mechanisms available to the cells of the oviduct.

In terms of biological processing principles, the following key steps of biomineralization in eggshells have been identified:

1. Avian eggshell crystallization sequences are accomplished under cellular control. Cells

lining the hen's oviduct secrete organic matrix components that self-assemble around the shape-defining egg yolk to form a structural organic scaffold that will subsequently mineralize.
2. The spatial location and initiation of mineralization is defined by the clustering of certain interactive proteins that appear to bind to a specific region of the structural matrix and, at the same time, sequester large amounts of one of the mineral-phase constituents. The interactive proteins are predominantly anionic, rich in specific arrays of carboxyl phosphate groups, and may present β-sheet-like structures.
3. Successive layers of mineral/matrix composite are then deposited onto the membrane to form the eggshell proper.

2.2 General Biological Processing Principles

Mineralization processes used in biological systems exhibit a number of interesting features that have significant implications for material scientists and engineers seeking highly efficient and/or unique ceramic processing strategies. Biomineralization is of potential value because it is a form of environmentally sound ("green") colloidal ceramic processing based on biocompatible materials and mild conditions with no toxic by-products. In contrast to the high temperatures and/or high pressures associated with conventional shape-forming processes, the goal of a high-density, net-shape structure is accomplished with apparent ease by cellular deposition of an organic "matrix" that regulates where, when, and in what form mineralization will occur. The result of this molecular assembly is often a nanostructured composite with a complex microstructure, containing microarchitectures on a number of length scales. Dense crystalline films are synthesized, and sometimes patterned, without subsequent thermal treatment, on polymer substrates of complex geometry. Exceptional chemical and mechanical properties can derive from the fine composite structure, as described earlier.

Although many of the details of biomineralization are not yet well understood, general processing paradigms have emerged from analyses of biomineralizing systems. The three general biological processing principles governing the composition, architecture and methods of assembly of a variety of biological ceramics can be succinctly summarized (Heuer *et al.*, 1992).

1. *The compartment principle*. The initial stage in the fabrication of biominerals is the construction of an organized reaction environment or compartment. Biomineralization occurs within specific subunit compartments or microenvironments, which implies stimulation of crystal production at certain functional sites and inhibition or prevention of the process at all other sites.
2. *The specific mineralization principle*. A specific mineral is nucleated within the compartment with a defined crystal size (frequently in the nanometer range), shape, and orientation. Nucleation and growth of the biomineral phase are carefully and exquisitely orchestrated by a complex spectrum of organic matrix biopolymers. These matrix molecules appear to have been engineered through evolutionary processes to regulate a single, precise step in either the nucleation or the growth portion of the production of the mineral phase. Although specific details are not yet available for any organism, it is generally considered that the assembly of mineral nuclei is governed by electrostatic, structural, and stereochemical complementarity at the inorganic–organic interface (Mann, 1988).
3. *Net shape by self-assembly of incremental units principle*. Macroscopic shape forming is accomplished by self-assembly of many incremental units around a genetically defined scaffold or mandrel of structural proteins. Slow, repetitive processing results in a net-shape composite with layered microarchitecture that can impart exceptional mechanical properties. Rapid mineralization may be characterized by columnar growth patterns and one-time exposure to nucleating proteins.

An important feature of biomineral processing is that integration of the three general principles is the norm rather than the exception, as is illustrated in the examples of formation of bones and shells. Another powerful feature of biomineral processing is that, in some systems, remodeling and regeneration of the original mineral structure occurs, as needed, to optimize strength, accommodate organism growth, maintain mineral ion equilibrium, and effect repairs. While this extremely "intelligent" processing is currently too advanced for exact imitation, a few examples of novel synthetic routes inspired by biological mineralization are discussed next to suggest the potential of the biomimetic approach.

3. BIOMIMETIC CERAMIC PROCESSING

The next three sections offer representative examples of the three main areas of biomimetic ceramic processing: compartmentalized processing of particulates, thin-film formation, and net-shape formation of structures with layered architectures. Each area has developed around one of the three general biological processing principles. In a few cases, two of the three principles have been integrated with even greater advantage.

3.1 Compartmental Processing of Particles and Particulate Networks

Historically, the control of crystallization has been achieved empirically. The size, morphology, and polymorphic expression of a crystalline material can be controlled by changing the ambient temperature, supersaturation, solvent, and impurity level during crystallization. Biomimetic approaches seek to attain specific control over crystal habit, texture, and morphology by selective use of organic molecules.

3.1.1 Nanoparticles from Single Compartments Unicellular creatures (bacteria, algae, protozoa) use single, enclosed protein cages and lipid vesicles (compartments) to mineralize discrete nanometer-sized crystals and crystal clusters that do not assemble into complex, hierarchical structures such as bone and shell (Blakemore *et al.*, 1979). Recognizing the unique electrical, optical, and chemical quantum-sized effects possible in nanoscale particles (Ozin, 1992), as well as the high strength and superplastic deformation associated with nanostructured materials (Gleiter, 1992; Siegel, 1992), many different nanoparticles have been synthesized using isolated organic compartments (Table 1). Specific references can be found in two review articles (Mann, 1996; Kuhn *et al.*, 1996b).

Table 1. Biomimetic approaches to nanoscale particle synthesis.

Product	Systems	Materials
Clusters	Reverse micelles	CdS
	Microemulsions	Pt, Co, metal borides, Fe_3O_4, $CaCO_3$, $LaNiO_3$, $BaFe_{12}O_{19}$
Nanoparticles	Vesicles	Pt, Ag, CdS, ZnS, Ag_2O FeOOH, Fe_3O_4, $CoFe_2O_4$ Al_2O_3, Ca phosphates
	Ferritin	MnOOH, UO_3, FeS, Fe_3O_4
	Langmuir–Blodgett films	CdS
	Polystyrene resin	γ-Fe_2O_3

Microemulsions and reverse micelles are versatile systems because the reaction environment can have variable diameters (1–500 nm) and the surface functional group can be modified by straightforward chemistry. For example, phospholipid vesicles and reversed micelles formed by surfactant–water mixtures have been used to obtain unimodal, nonagglomerating, nanometer-sized metal oxide powders (Table 1) (Fig. 9) (Archibald and Mann, 1993), while the inverse microemulsion technique has been used to synthesize a fine $LaNiO_3$ powder (Gan *et al.*, 1994) (Fig. 10). Since each particle is surrounded by a bilayer membrane of several nanometers thickness, particle-to-particle interactions are negligible and reaction rates can be diffusion-controlled.

The utility of these techniques is limited, on account of the sensitivity of phospholipid or surfactant assemblies to changes in phase behavior and the dynamic behavior of reverse micelles. Protein cages with nanometer-sized internal cavities are more robust and have been successfully utilized to synthesize uniform and monodisperse Fe_3O_4 particles (Meldrum

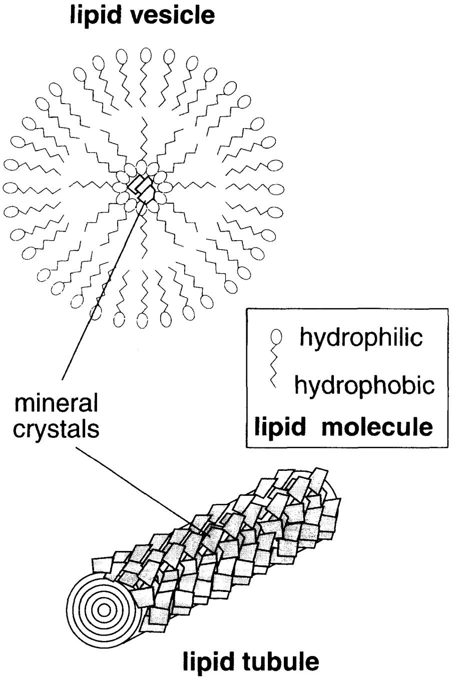

FIG. 9. Lipids provide versatile compartments for nucleation of nanometer-sized mineral particles.

et al., 1992); use of such protein cages appears to represent an integration of principles (1) and (2).

3.1.2 Nanoparticle Networks A combination of traditional sol–gel processes and polymer chemistry has been used to fabricate biomimetic films of particle-filled polymer composites. The inorganic reinforcing phase is formed when the alkoxide in a polymer–alkoxide–solvent mixture undergoes hydrolysis. Both SiO_2- and TiO_2-filled organic–inorganic hybrids have been formed with a wide range of polymers (Mark and Calvert, 1994). As might be expected, sol–gel composites, which are interpenetrating networks of organic and inorganic phases, are much stiffer and glass-like than the polymer-matrix materials.

The surge of activity in mesoporous silica and zeolites is based on the principle of co-organization of mineral and organic components in a synergistic process and can also be considered biomimetic. One-, two-, and three-dimensional liquid-crystal-like periodic arrays of inorganic–organic assemblies have been formed from surfactant molecules and inorganic ions (Firouzi *et al.*, 1995).

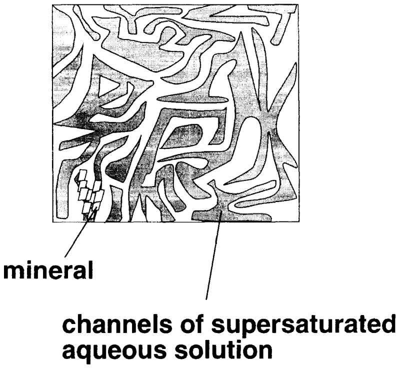

FIG. 10 Nanoparticle networks are formed in the supersaturated aqueous channels of a microemulsion.

3.2 Specific Mineralization (Application to Thin Films)

3.2.1 Self-assembled Monolayers The concept of a functionalized surface that will initiate specific crystallization [principle **2**] has been used to produce ceramic coatings through low-temperature processing routes. In some ways, these approaches are the most overtly biomimetic processing techniques yet studied.

Methods to create surfaces derivatized with specific functional groups that will initiate crystal nucleation and growth have embraced a variety of chemical approaches, including supramolecular templates and templates made from inorganic polymers, organic polymers, or biomolecules. The aqueous formation of dense polycrystalline films of FeOOH, Fe_2O_3, SnO_2, CdS, and $Al(OH)_3$ onto surfaces functionalized by self-assembled monolayers (SAMs) have been reported (Bunker *et al.*, 1994). The surfactant molecules within SAMs consist of three parts:

1. a silane coupling agent to the substrate;
2. the alkyl chain; and
3. a variety of functionalized end groups with variable chemistry, *i.e.*, $-COOH$, $-SO_3H$, $-H_2PO_4$, $-CH_3$, and $-NH_2$ (Fig. 11).

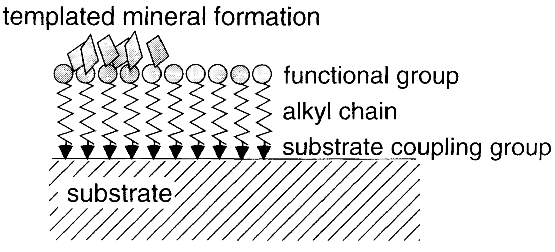

(a) Self Assembled Monolayer

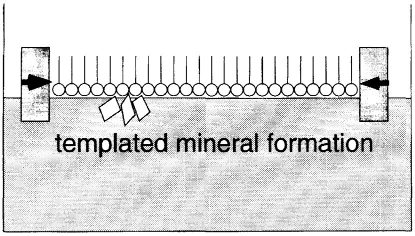

(b) Langmuir Blodgett film

FIG. 11. Templated mineral formation on functionalized monolayers: (a) self-assembled monolayers (SAMs) or (b) Langmuir–Blodgett films.

The ordered structure and durability of siloxane-anchored organic SAMs are related to the monolayer-to-substrate bonding and to the van der Waals interactions among the linear alkyl chains of the SAM. The chemical attachment of the SAM is based on the siloxane network formed by the spontaneous, self-assembling reaction (Ulman, 1991) of the trichlorosilyl ($-SiCl_3$) group of the alkylsilane with the hydrated surface of the oxide on the substrate. The strong interactions with the substrate and the short-range van der Waals forces between the chains allow for the formation of a highly ordered and close-packed monolayer film. This attachment yields monolayers highly resistant to adverse conditions such as prolonged exposure to aggressive solvents. Significant advantages for oxide film formation are thus accrued. First, a functional group on the SAM surface can be chemically modified without disrupting the monolayer. Thus, a directly deposited SAM, bearing a chemical functionality that is suited for incorporation into the alkyltricholorosilane and desired for SAM deposition, can be chemically altered after the SAM is deposited. In the context of coating technology, such functionality may be desirable for its ability to participate in and/or induce the formation of an inorganic oxide layer, though it could not stably coexist with the trichlorosilane used for attachment to the substrate in the initial SAM formation. Second, SAMs can survive temperatures approaching 100 °C while exposed to solutions ranging in *p*H from strongly acidic to mildly basic, thereby providing a workable range of conditions for the formation of various oxides from aqueous or organometallic solutions. This technology has been applied to the formation of high-quality TiO_2 films with possible microelectronic applications (Collins *et al.*, 1996).

These biomimetic processes have significant advantages over vapor deposition techniques for the formation of oxide thin films. Vapor techniques primarily involve line-of-sight deposition, whereas the solution process can be applied to complex shapes and a wide variety of surfaces. With vapor techniques, the coverages and properties of the deposited films depend on the design of the reaction chamber and (typically) empirical determination of the optimum deposition conditions. Most importantly, achievement of the desired properties with vapor deposition routes generally requires heating the substrate (often to several hundreds of degrees Celsius), either during deposition or subsequently, to convert the usually amorphous as-deposited material into a well-ordered crystalline film. Finally, with vapor deposition, when a specific crystallographic orientation of the film is required, the choice of substrate is usually restricted to a single crystalline material whose surface interatomic spacing closely matches that of the desired crystalline film, enabling epitaxial growth. However, the use of organic substrates may accommodate some lattice mismatch because of their high compliance.

3.2.2 Biomimetic Lithography Traditional lithography methods can generate patterns of electronic circuits on substrates with dimensions of the order of 0.3 μm or less, and straightforward photolithography was used to form patterned films of TiO_2. Thioacetate (SCOOH)-terminated hexadecyl trichlorosilane SAMs were deposited onto Si(100) wafers, providing modestly hydrophobic films, which do not encourage oxide thin-film formation. The thioacetate functionality can be photolytically oxidized through a mask to an acidic sulfonate (SO_3H) moiety, rendering the exposed portions of the film hydrophilic. Upon immersion into an aqueous medium, even one that is strongly acidic, the sulfonate groups will deprotonate, becoming negatively charged, while the nonphotolyzed thioacetate-terminated regions remain unchanged. Thin films of nanocrystalline TiO_2 (anatase polymorph) form via heterocoagulation of homogeneously nucleated colloidal particles on the SO_3H-terminated regions of the film (Collins *et al.*, 1996).

A complementary technique has been used for patterned deposition of dielectric oxides (Pb, La)TiO_3 and $LiNbO_3$ (Jeon *et al.*, 1995). In this approach, a hydrophobic SAM acts as a resist layer to which sol–gel deposited oxides do not adhere. This approach relies on the removal of poorly adsorbed oxides from the SAM-coated regions by mechanical abrasion.

Electron- and ion-beam lithographic techniques can also be used to pattern SAMs with functional organic groups that either inhibit or promote mineral deposition (Fig. 12). Patterned films of iron oxyhydroxide several micrometers in width were deposited on areas patterned with nucleation sites (Rieke *et al.*, 1994). A hydrophobic SAM was first deposited over the entire substrate (again a silicon wafer). In the second step selected regions were etched with an ion or electron beam to define the pattern. These regions were backfilled with a second SAM terminated with end groups functionalized to promote mineralization, which commenced after immersion of the treated substrate in an aqueous solution.

A number of other biomimetic approaches to surface patterning have recently been developed that utilize self-assembly of organic molecules, such as protein or lipid systems (Douglas *et al.*, 1986), or the microcontact printing technique (Kumar *et al.*, 1994), to create specific patterns on the substrate surface. These systems are potentially capable of creating patterns on a nanometer scale.

3.2.3 Langmuir–Blodgett Films Surfactants or other amphiphilic molecules compressed to form monolayers on the surface of an aqueous liquid are known as Langmuir–Blodgett films [Fig. 11(b)]. These monolayers can be removed intact from the surface by dipping a substrate into the film, and they can also be stacked to form multilayers. Both the phase and orientation of mineral crystals that nucleate on these films can be controlled in a biomimetic manner, by appropriate choice of the Langmuir–Blodgett monolayer substrate and the solution concentration. For example, stearate ($C_{17}COO$) monolayers have a distance between the psuedohexagonal net of long-chain carboxylate molecules that matches the distances between the coplanar Ca ions on the $\{1\bar{1}0\}$ plane of calcite, which facilitates the oriented nucleation of crystals with this habit (Heywood *et al.*, 1991). Similar geometric correspondences are present for the (100) face of $BaSO_4$ crystals nucleated under long-chain alkyl sulfate or phosphonate monolayers (Heywood *et al.*, 1992).

Particulate films at the air–water interface have been formed by dispersing surfactant-stabilized nanoparticles on an aqueous subphase contained in a Langmuir trough (Meldrum and Fendler, 1996). The particles are prepared prior to their incorporation into the films, thus enabling their dimensions, physical properties, and particle size distribution to be precisely controlled. Films of CdS, TiO_2, Fe_3O_4, colloidal and micellar Ag, Pt, Pd, and $BaTiO_3$ have been formed by this method (Meldrum and Fendler, 1996).

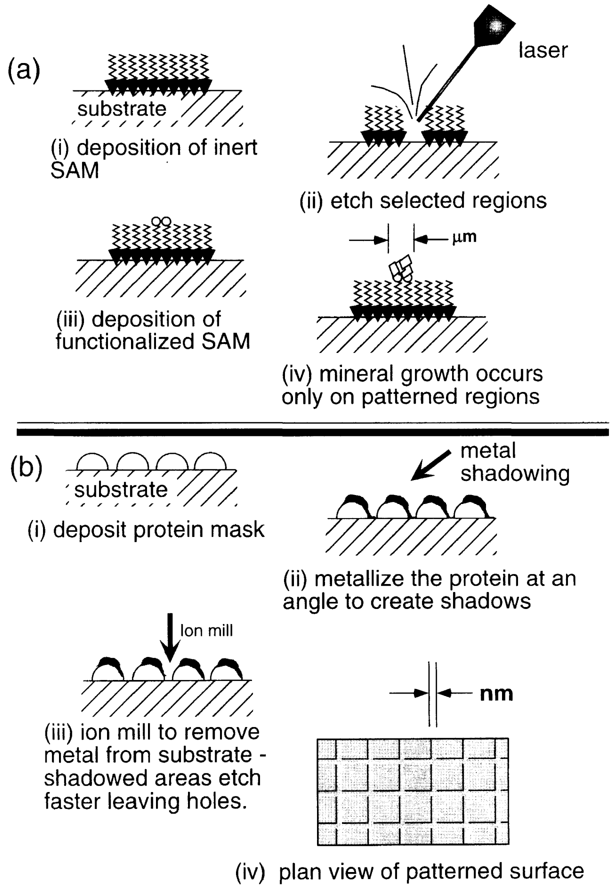
(a)
substrate
(i) deposition of inert SAM
laser
(ii) etch selected regions
(iii) deposition of functionalized SAM
µm
(iv) mineral growth occurs only on patterned regions
(b)
substrate
(i) deposit protein mask
metal shadowing
(ii) metallize the protein at an angle to create shadows
Ion mill
(iii) ion mill to remove metal from substrate - shadowed areas etch faster leaving holes.
nm
(iv) plan view of patterned surface

FIG. 12. Biomimetic lithography at two length scales: (a) using SAMs to create surface patterns with micrometer-sized features and (b) protein masks for nanometer-sized features.

3.3 Net Shape Formation with Layered Architecture

The layered architecture of biominerals such as bone and seashell has inspired new material designs for layered ceramics that exhibit increased fracture resistance over that of monolithic ceramics. Al-B_4C cermets were produced by stacking layers of B_4C tapes, which were then laminated using thermocompression, sintered and subsequently infiltrated with Al (Yasrebi *et al.*, 1990). These graded structures showed increasing strength and toughness with decreasing lamina width, with as much as a 40% improvement over that of monolithic materials. Graceful failure reminiscent of the conch shell fracture described earlier, due to a 400% increase in fracture toughness, has been observed in a synthetic composite material composed of SiC and graphite layers (Clegg, 1992). These are relatively crude imitations (100 times coarser) of the fine laminae in biological ceramics, and yet lead to a significantly improved fracture resistance.

While microstructural control is important, the most significant limitation of lamination techniques (as practiced to date) is in relation to the manufacture of complex shapes. Recent innovations have been made in fabricating components with well-defined complex geometries by methods that are reminiscent of the cellular deposition strategy of biological ceramics. These methods are generally referred to as solid free-form fabrication (SFF) and can in principle be used for rapid prototyping or for production of small quantities of engineering components. Complex three-dimensional parts are made through sequential deposition of layers of material that represent a two-dimensional slice of the three-dimensional object being formed. A specific cross section may be defined by cutting a layer of tape with a laser to a contour specified by a CAD (computer-aided design) representation, or by depositing a layer of binder onto a powder layer preferentially within the desired cross-sectional shape. The high cost of tooling, normally an obstacle to the implementation of design changes, or to material substitutions, can be overcome with these methods.

A number of approaches are now being pursued (Sachs *et al.*, 1990; Griffin *et al.*, 1994; Thissell and Marcus, 1994; Cawley *et al.*, 1996; 3D Systems in Valencia, CA, USA); we describe one here, the process known as CAM-LEM (for computer-aided manufacturing of laminated engineering materials). Fabrication begins with a combination of green ceramic tape and "fugitive" tape (Fig. 13) (Cawley *et al.*, 1996). As is common to all SFF technologies, each part originates from a computer description, which is analyzed and decomposed into boundary contours of thin slices. A laser is used to cut the two-dimensional shape into the green tape sheet stock. The resulting part-slice regions are extracted from the sheet stock and stacked to assemble a physical 3D realization of the original CAD description. Appropriately cut fugitive supports are placed, layer by layer, along with the cut tapes (Fig. 13). The green assembly is then laminated and is conventionally sintered. The use of laminated ceramic tapes is well suited to fabrication of high-density advanced structural ceramics, of functionally gradient materials, and of laminated composites. An impressive variety of Al_2O_3 components, many with complex internal channels, have been fabricated by this method (Cawley *et al.*, 1996).

There are very few synthetic processes that generate a shape or an architecture by microscopic remodeling of an existing ceramic structure. The primary limitation of synthetic additive or subtractive methods is the damage introduced by the chemical or mechanical removal techniques to the surface, which cannot be easily repaired by addition of new material. However, if the procedure could be done at the molecular level, it might be possible to effect new levels of control of microstructure.

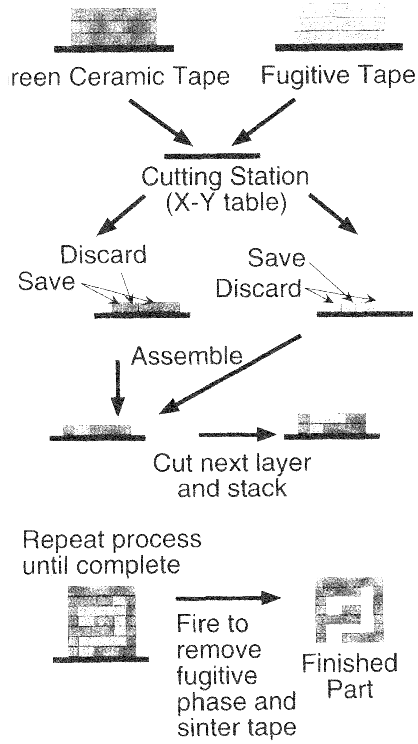

FIG. 13. A schematic representation of the principal steps involved in the CAM-LEM process for forming ceramics with a complex shape. To obtain complicated internal porosity, both green ceramic tape and a fugitive tape that supports overhanging elements are used. The tapes are separately cut, stacked, and then fired to sinter the laminae, remove the fugitive phase, and produce a finished product.

3.4 Living Systems Used for Materials Synthesis

More than 60 distinct minerals are known to be produced by living cells, many being produced by unicellular organisms (Lowenstam and Weiner, 1989). For example, magnetotactic bacteria such as *Aquaspirillum magnetotacticum*, discovered in 1979 (Blakemore *et al.*, 1979), produce magnetic minerals (Fe oxides and sulfides) that are believed to align anaerobic bacteria along the Earth's magnetic field, thereby directing the organism into the anaerobic zone. These bacteria might prove to be a source of particles that could be used directly in the production of Fe oxide magnetic coatings. The Fe oxide and CdS produced in such organisms are nonagglomerated 10–100-nm particles, a form that is difficult to prepare synthetically. Genetic manipulation of single-celled organisms is now widely practiced and presents an opportunity for the production of naturally occurring biopolymers in commercial amounts. The opportunity exists to use these organisms or their products in a biomimetic cell-mediated fabrication of synthetic ceramic/polymer composites.

4. CONCLUSIONS

Studies on biological ceramics have shown that many of the remarkable properties exhibited by natural materials arise from the complex microarchitectures that organisms can fabricate. Biomimetic approaches to ceramic processing seek to imitate the microarchitecture of biological materials, which are primarily layered or fibrillar. The production of biological ceramics involves an exquisite level of control of both the spatial regulation of the nucleation and growth of mineral and of the development of unique microarchitectures during formation. This is accomplished by organic "matrix" constituents secreted by cells. Conventional processing methods cannot replicate the fine-scale structure of biological ceramics, and thus several innovative methods have developed based on the three general biological processing principles that govern the composition, architecture, and methods of assembly of a variety of biological ceramics—the compartment principle, the specific mineralization principle, and the net shape by self-assembly of incremental units principle.

The research cited herein indicates that it is possible to extend the properties (most notably the fracture resistance) of synthetic materials by creating a fine-scale architecture within the material. While biomimetic material architectures improve toughness, it is usually at the cost of stiffness and strength. This may require a redesign of the macroscopic shape. Since Nature uses systems and not just materials in isolation, a concept focused on structures and not just materials is considered a further extension of biomimetics. The ultimate extension of biomimetics may involve the incorporation of genetically manipulated living cells modulating materials synthesis with tailored biomolecules and enzymatic reactions.

ACKNOWLEDGMENTS

The authors would like to thank the Department of Energy for financial support of the biomimetic processing work, the Electric Power Research Institute for supporting their ongoing research in this area, and the National Institutes of Health for financial support for L. Kuhn in the form of an NRSA training grant to the Department of Orthopaedic Surgery at Harvard Medical School administered by the Laboratory for the Study of Skeletal Disorders and Rehabilitation at Children's Hospital, Boston. The authors greatly appreciate the secretarial support provided by Carla Hines and Peggy Mayer.

WORKS CITED

Archibald, D. D., Mann, S. (1993), *Nature (London)* **364**, 430–433.

Ashby, M. F. (1993), *Acta Metall. Mater.* **41**, 1313–1335.

Blakemore, R. P., Maratea, D., Wolfe, R. S. (1979), *J. Bacteriol.* **140**, 720–729.

Bøggild, O. B. (1930), *K. danske Vidensk. Selsk. Skr.* **2**, 232–325.

Bond, G. M., Scott, V. D., Board, R. G. (1986), *Zool. Lond. A* **209**, 225–237.

Bunker, B. C., Rieke, P. C., Tarasevich, B. J., Campbell, A. A., Fryxell, G. E., Graff, G. L., Song, L., Liu, J. W., McVay, G. L. (1994), *Science* **264**, 48–55.

Cawley, J. D., Heuer, A. H., Newman, W. S., Mathewson, B. B. (1996), *Amr. Ceram. Soc. Bull.* **75**, 75–79.

Clegg, W. J. (1992), *Acta Metall. Mater.* **40**, 3085–3093.

Collins, R. J., Shin, H., DeGuire, M. R., Heuer, A. H., Sukenik, C. (1996), *Appl. Phys. Lett.* **69**, 860–862.

Creger, C. R., Phillips, H., Scott, J. T. (1976), *Poult. Sci.* **55**, 1717–1723.

Currey, J. D., Taylor, J. D. (1974), *J. Zool. Lond.* **173**, 395–406.

Currey, J. D., Kohn, A. J. (1976), *J. Mater. Sci.* **11**, 1615–1623.

Dennis, J. E., Xiao, S.-Q., Agarwal, M., Fink, D. J., Caplan, A. I. (1996), *J. Morphol.* **228**, 287–306.

Dieckert, J. W., Dieckert, M. C., Creger, C. R. (1989), *Poult. Sci.* **68**, 1569–1584.

Douglas, K., Clark, N. A., Rothschild, K. J. (1986), *Appl. Phys. Lett.* **48**, 676–679.

Firouzi, A., Kumar, D., Bull, L. M., Besier, T., Sieger, P., Huo, Q., Walker, S. A., Zasadzinski, J. A., Glinka, C., Nicol, J., Margolese, D., Stucky, G. D., Chmelka, B. F. (1995), *Science* **267**, 1138–1143.

Gan, L. M., Chan, H. S. O., Zhang, L. H., Chew, C. H., Loo, B. H. (1994), *Mater. Chem. Phys.* **37**, 263–268.

Giraud-Guille, M. M. (1988), *Calcif. Tissue Int.* **42**, 167–180.

Gleiter, H. (1992), *Nanostruct. Mater.* **1**, 1–20.

Glimcher, M. J. (1959), *Rev. Mod. Phys.* **31**, 359–393.

Gregoire, E. M., Duchateau, G. H., Florkin M. (1955), *Ann. Inst. Oceanogr. Monaco* **31**, 1–36.

Griffin, C., Daufenbach, J., McMillin, S. (1994), *Am. Ceram. Soc. Bull.* **73**, 109–113.

Heuer, A. H., Fink, D. J., Laraia, V. J., Arias, J. L., Blackwell, J., Calvert, P. D., Kendall, K., Messing, G. L., Rieke, P. C., Thompson, D. H., Wheeler, A. P., Veis, A., Caplan, A. I. (1992), *Science* **255**, 1098–105.

Heywood, B. R., Rajam, S., Mann, S. (1991), *J. Chem. Soc. Faraday Trans.* **87**, 735–743.

Heywood, R. R., Rajam, S., Mann, S. (1992), *J. Am. Chem. Soc.* **114**, 4682–4686.

Jackson, A. K., Vincent, J. F. V., Turner, R. M. (1988), *Proc. R. Soc. London B* **234**, 415–440.

Jeon, N. L., Clem, PG, Nuzzo, R. G., Payne, D. A. (1995), *J. Mater. Res.* **10**, 2996–2999.

Kim, H.-M., Rey, C., Glimcher, M. J. (1995), *J. Bone Min. Res.* **10** 1589–1610.

Kinne, R. W., Fisher, L. W. (1987), *J. Biol. Chem.* **262**, 10206–10211.

Krampitz, G., Graser, G. (1988), *Agnew. Chem., Int. Ed. Engl.* **27**, 1145–1159.

Kuhn, L. T., Kessler, H., Chateau, E., Spearing, S. M., Ballarini, R., Heuer, A. H. (1996a), *J. Mater. Sci.* **31**, 6583–6594.

Kuhn, L. T., Fink, D. J., Heuer A. H. (1996b), in: S. Mann (Ed.), *Biomimetic Materials Chemistry*, New York: VCH Publishers, Chap. 2.

Kumar, A., Bierbuyck, H. A., Whitesides, G. M. (1994), *Langmuir* **10**, 1498–1511.

Landis, W. J., Song, M. J., Leith, A., McEwen, L., McEwen, B. F. (1993), *J. Struct. Biol.* **110**, 39–54.

Lange, F. F. (1989), *J. Am. Ceram. Soc.* **72**, 735–741.

Laraia, V. J., Heuer, A. H. (1990), in: J. J. Bentzen, J. B. Bilde-Sørensen, N. Christiansen, A. Horsewell, B. Ralph (Eds.), *Structural Ceramics—Processing, Microstructure and Properties*, Denmark: Risø National Laboratory, pp. 79–96.

Lowenstam, H. A., Weiner, S. (Eds.) (1989), *On Biomineralization*, Oxford, UK: Oxford University Press.

Mann, S. (1988), *Nature (London)* **332**, 119–124.

Mann, S. J., Webb, J., William, R. J. P. (Eds.) (1989), *Biomineralization: Chemical and Biochemical Perspectives*, New York: VCH Publishers, p. 145.

Mann, S. (1996), in: S. Mann (Ed.), *Biomimetic Materials Chemistry*, New York: VCH Publishers, Chap. 1.

Mark, J. E., Calvert, P. D. (1994), *Mater. Sci. Eng. C* **1**, 159–173.

Meldrum, F. C., Heywood, B. R., Mann, S. (1992), *Science* **257**, 522–523.

Meldrum, F. C., Fendler, J. H. (1996), in: S. Mann (Ed.), *Biomimetic Materials Chemistry*, New York: VCH Publishers, pp. 175–220.

Messing, G. L., Mazdiyasni, K. S., McCauley, J. W., Haber, R. A. (Eds.) (1987), *Ceramic Powder Science*, Advances in Ceramics Vol. 21, Westerville, OH: American Ceramic Society.

Morse, D. E., Cariolou, M. A., Stucky, G. D., Sarembia, C. M., Hansma, P. K. (1993), in: C. Viney, S. T. Case, J. H. Waite, (Eds.) *Biomolecular Materials*, Materials Research Society Symposium Proceedings Vol. 292, Pittsburgh: Materials Research Society, pp. 59–67.

Nancollas, G. H. (Ed.) (1982), *Biological Mineralization and Demineralization*, Amsterdam: Springer-Verlag.

Ohring, M. (1992), *The Material Science of Thin Films*, Boston: Academic Press.

Omori, M., Watabe N. (Eds.) (1980), *The Mechanisms of Biomineralization in Animals and Plants*, Tokyo: Tokai University Press.

Ozin, G. A. (1992), *Adv. Mater.* **4**, 612–649.

Pechak, D. G., Kujawa, M. J., Caplan, A. I. (1986), *Bone* **7**, 441–472.

Rieke, P. C., Tarasevich, B. J., Wood, L. L., March, B. D., Song, L., Fryxell, E., Engelhard, M. H., Bear, D. R., John, C. M. (1994), in: M. A. Alper, H. Bayley, D. Kaplan, M. Navia (Eds.), *Biomolecular Materials by Design*, Materials Research Society Proceedings, Vol. 330, Pittsburgh: Materials Research Society, p. 139.

Robey, P. G., Boskey, A. L. (1996), in: R. Marcus, D. Feldman, J. Kelsey (Eds.), *Osteoporosis*, San Diego: Academic Press, Chap. 4.

Sachs, E., Cima, M., Williams, P., Brancazio, D. (1990), *Trans. NAMRI/SME*, 41–45.

Sarikaya, M. (1994), *Microsc. Res. Tech.* **27**, 360–375.

Segal, D. (1989), *Chemical Synthesis of Advanced Ceramic Materials*, Cambridge, UK: Cambridge University Press.

Siegel, R. W. (1992), in: F. E. Fujita (Ed.), *Physics of New Materials*, Heidelberg: Springer-Verlag.

Solomon, S. E. (1983), in: B. M. Freeman (Ed.), *Physiology and Biochemistry of the Domestic Fowl*, New York: Academic Press, Chap. 4.

Stemberger, B. H., Mueller, W. J., Leach, R. M., Jr. (1977), *Poult. Sci.* **56**, 537–543.

Taylor, J. D., Layman, M. (1972), *Palaeontology* **15**, 73–87.

Thissell, W. R., Marcus, H. L. (1994), in: H. L. Marcus, J. J. Beaman, J. W. Barlow, D. L. Bourell, R. H. Crawford, (Eds.), *Proceedings of the Solid Freeform Fabrication Symposium*, Austin, TX: University of Texas, pp. 311–320.

Ulman, A. (1991), *An Introduction to Ultrathin Organic Films—from Langmuir Blodgett to Self Assembly*, Boston: Academic Press.

Weiner, S., Talmon, Y., Traub, W. (1983), *Int. J. Biol. Macromol.* **5**, 325–328.

Weiner S. (1984), *Am. Zool.* **24**, 945–951.

Wilbur, K. M. (1973), in: *Symposium on Hard Tissue Growth, Repair and Mineralization*, Ciba Foundation Symposium 11, London: Elsevier, pp. 7–33.

Wuthier, R. E. (1977), *Calcif. Tissue Res.* **23**, 125–133.

Yasrebi, M., Kim, G. H., Gunnison, K. E., Milius, D. L., Sarikaya, M., Aksay, I. A. (1990), in: B. J. J. Zelinski, C. J. Brinker, D. E. Clark, D. R. Ulrich, (Eds.), *Better Ceramics through Chemistry IV*, MRS Symposium Proceedings Vol. 180, Pittsburgh: Materials Research Society, pp. 625–635.

Ziv, V., Sabanay, I., Arad, T., Traub, W., Weiner, S. (1996), *Microsc. Res. Tech.* **33**, 203–213.

FURTHER READING

Brighton, C. T., Friedlaender, G., Lane, J. M. (1994), *Bone Formation and Repair*, Rosemont, IL: American Academy of Orthopaedic Surgeons.

Bunker, B. C., Rieke, P. C., Tarasevich, B. J., Campbell, A. A., Fryxell, G. E., Graff, G. L., Song, L., Liu, J. W., McVay, G. L. (1994), *Science* **264**, 48–55.

Harasewych, M. G. (1991), *Shells: Jewels from the Sea*, Pennsylvania: Courage Books.

Lowenstam, H. A., Weiner, S. (Eds.) (1989), *On Biomineralization*, Oxford, UK: Oxford University Press.

Mann, S. J., Webb, J., William, R. J. P. (Eds.) (1989), *Biomineralization: Chemical and Biochemical Perspectives*, New York: VCH Publishers.

Mann, S. (Ed.) (1996), *Biomimetic Materials Chemistry*, New York: VCH Publishers.

Sarikaya, M., Aksay, I. A. (1995) *Biomimetics: Design and Processing of Materials*, Woodbury NY: AIP Press.

Vincent, J. (1990) *Structural Biomaterials*, New Jersey: Princeton University Press.

Wainwright, S. A., Biggs, W. D., Currey, J. D., Gosline, J. M. (Eds.) (1982), *Mechanical Design in Organisms*, New Jersey: Princeton University Press.

MICROMAGNETICS AND MICROMAGNETIC DEVICES

TOSHIAKI YACHI, *NTT Integrated Information & Energy Systems Laboratories, Musashino-shi, Tokyo, Japan*

MASATO MINO, *Research Planning Department, NTT Multimedia System Laboratory Group, Yokosuka-shi, Kanagawa Prefecture, Japan*

Introduction 47
1. **General Principles** 48
1.1 Concept of Micromagnetics 48
1.1.1 Theory of Micromagnetics 48
1.1.2 Advantages of Micromagnetic Devices 51
1.2 Materials of Micromagnetic Devices 52
1.3 Design of Micromagnetic Devices 52
2. **Types of Micromagnetic Device** 54
2.1 Planar Inductors and Planar Transformers 54
2.1.1 Meander Coil 54
2.1.2 Spiral Coil 55
2.2 Thin-Film Magnetic Heads 57
2.3 Microactuators and Microstators 57
3. **Applications of Micromagnetic Devices** 59
3.1 Switching Power Supplies 59
3.2 Magnetic Disk Storage 61
3.3 Micromachines 61
Acknowledgments 62
Glossary 62
Works Cited 63
Further Reading 63

INTRODUCTION

The past three decades have witnessed the rapid advance of solid-state electronics, from the replacement of hand-wired discrete elements to the integration of many circuit elements into a single semiconductor chip. Active devices such as semiconductors as well as resistive and capacitive elements have already been incorporated into modern integrated circuit (IC) technology. The only fundamental discrete circuit elements that have been left behind in the revolution are those involving magnetic materials.

In recent years, the field of micromagnetics and micromagnetic devices has attracted the attention of an increasing number of researchers. The objective of this field is to develop extremely small magnetic microdevices. Recent developments in microelectronic technology have revolutionized the fabrication techniques of magnetic devices and broadened the range of useful applications, such as in portable electronics and micromachines.

Micromagnetic devices use magnetic films, insulators, and planar coils. A planar inductor having a spiral coil sandwiched between Permalloy magnetic films was first fabricated by Saleh *et al.* (1970). Later, the importance of the IC inductor with magnetic films was pointed out by Soohoo (1979). In parallel with the development of computers, the bit-packing density of magnetic disk storage was steadily being increased following the proposal in 1977 of a thin-film head that combined a magnetic thin film and a planar coil. The thin-film head permits high-speed data transfer with good read/write performance, since it uses a magnetic film with a smaller magnetic circuit and better high-frequency characteristics than those of a bulk-type core. The use of lithography allows several heads to be produced on one substrate, resulting in the mass production of heads with high dimensional accuracy and stable characteristics.

In the latter half of the 1980s, active research on microinductors and microtransformers sought to resolve high-frequency application requirements for more compact switching power supplies for portable elec-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

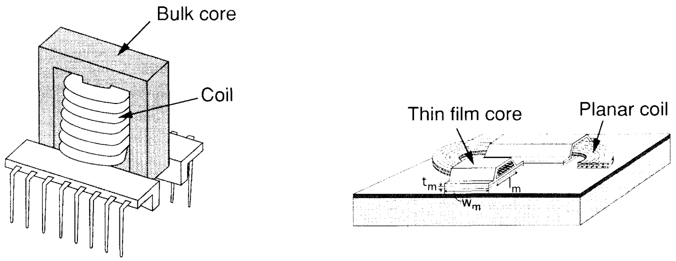

(a) Bulk

(b) Planar

FIG. 1. Schematic diagrams of magnetic devices: (a) bulk device and (b) planar device.

tronic devices that were proliferating then. Many reports were published on micromagnetic device materials, device structures, manufacturing processes, and characteristics, reflecting marked advances in the field of micromagnetics. A buck-converter switching power supply using a thin-film inductor was made by Mino *et al.* (1996). This power supply had an efficiency of 85%, which is sufficient for practical use. In the same year, a step-up switching power supply was made by Sato *et al.* (1996), using a thin-film inductor operating at 5 MHz.

In the 1990s, micromachine technology suddenly burgeoned, and ideas based on micromagnetic technology also came to the fore. Further research and development activities are expected in the promising field of micromachines in the future.

This article reviews the fundamental theory of micromagnetics and micromagnetic devices in terms of designs, fabrication techniques, and applications. No treatment of the basic physics of magnetic devices is attempted (see MAGNETIC DEVICES AND MAGNETIC MATERIALS).

1. GENERAL PRINCIPLES

1.1 Concept of Micromagnetics

Applied micromagnetics is a new concept aiming to develop more compact, more highly integrated magnetic devices in the same way that advanced microcircuit technologies led to the development of more compact, more highly integrated solid-state electronic devices. For example, the microinductor shown in Fig. 1(b) is a planar magnetic device formed using advanced magnetic thin-film technology. The microinductor differs from the three-dimensional, bulk-core type inductor shown in Fig. 1(a) in that it is nearly two-dimensional—its thickness being extremely small compared with its area.

1.1.1 Theory of Micromagnetics The field of micromagnetics developed along with high-frequency applications of magnetic devices. Magnetic properties and their relationship with high-frequency applications are particularly interesting subjects in this field. Other concerns include magnetic domain-wall displacement and magnetization rotation, which cause the reversal of magnetization in magnetic materials. Domain-wall displacement is the dominant factor in the low-frequency region, and the use of thin material is most effective for this. In the high-frequency region of several megahertz, high-speed displacement of the domain wall causes a large eddy-current loss, causing the characteristics in the direction of the wall movement to deteriorate. Thus, only the direction of magnetization rotation, in other words the hard axis of magnetic anisotropy, is utilized. Fortunately, comparatively large uniaxial anisotropy can be given to the thin-film material used for micromagnetic devices to decrease the loss, thanks to its magnetic anisotropy, magnetoelastic effect, and anisotropic shape. In the higher-frequency region, however, the eddy-current loss due to magnetization rotation will increase, causing deterioration of the magnetic properties. An effective countermeasure against this deterioration may be to increase the electrical resistance of the magnetic material. However, this cannot be the final solution because the loss due to magnetic resonance will increase in the higher-frequency region.

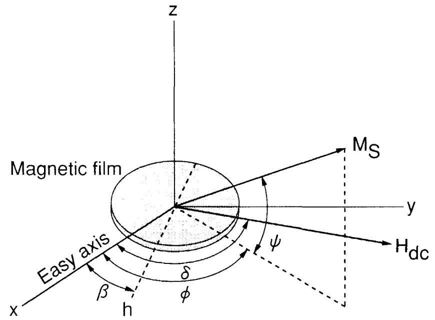

FIG. 2. Notation for Eqs. (1)–(8).

To gain a clear understanding of the frequency characteristics described above, let us consider permeability behaviors as previously reported (Shimada, 1993). The experiment on Ni–Zn ferrite known as Snoek's limit is famous for evaluating permeability behavior in the high-frequency region. This arises from the relationship between the above-mentioned natural resonance and magnetic anisotropy, and in a thin film, shape anisotropy is an additional major cause of magnetic anisotropy. The shape anisotropy works to hold magnetization within the thin-film surface, thus greatly increasing the resonant frequency.

The equation of motion of thin-film magnetization is expressed as below from the Landau–Lifshitz equation:

$$\frac{d^2\phi}{dt^2} + 4\pi\lambda\frac{d\phi}{dt} + \frac{\gamma^2}{\mu_0}\frac{\delta E}{\delta\phi} = 0, \tag{1}$$

where $\lambda = \alpha\gamma M_s/4\pi\mu_0$, α is the damping constant, γ is the gyromagnetic ratio, and μ_0 is the permeability of the vacuum.

Figure 2 shows the directions of external magnetic field H_{dc}, alternating current magnetic field h, and magnetization M_s with the x–y plane being the film surface. Energy E is expressed by the following equation when the easy axis of magnetic anisotropy is parallel to the x axis, as previously reported (Smith, 1958):

$$\begin{aligned} E = {} & K_u \sin^2\phi - M_s(H_{dc}\cos\delta + h\cos\beta) \\ & \times \cos\psi\cos\phi - M_s(H_{dc}\sin\delta + h\sin\beta) \\ & \times \cos\psi\sin\phi + (M_s{}^2/2\mu_0)\sin^2\psi, \end{aligned} \tag{2}$$

where K_u is the uniaxial anisotropy constant. The first three terms are obviously important in the static case, and the fourth term indicates the demagnetizing field energy produced by the thin-film shape; it acts to increase the resonance frequency.

The static energy E_0 is obtained from Eq. (2) with the magnetization directions stabilized by H_{dc} being ϕ_0 and ψ_0 when $h = 0$:

$$\begin{aligned} E_0 = {} & K_u \sin^2\phi - M_s H_{dc}\cos(\phi - \delta) \\ & + (M_s{}^2/2\mu_0)\sin^2\psi. \end{aligned} \tag{3}$$

The following equation is also obtained from Taylor's series of $\delta E/\delta\phi$ at (ϕ_0, ψ_0):

$$\frac{\delta E}{\delta\phi} = M_s h \sin(\phi_0 - \beta) + \left.\frac{\delta^2 E_0}{\delta\phi^2}\right|_{\phi_0,\psi_0} \Delta\phi \tag{4}$$

Equations (1), (3) and (4) give the equation of motion of thin-film magnetization:

$$\begin{aligned} & \frac{d^2(\Delta\phi)}{dt^2} + 4\pi\lambda\frac{d(\Delta\phi)}{dt} + \left.\frac{\gamma^2}{\mu_0}\frac{\delta^2 E}{\delta\phi^2}\right|_{\phi_0,\psi_0} \Delta\phi \\ & = -(\gamma^2/\mu_0) M_s h \sin(\phi_0 - \beta). \end{aligned} \tag{5}$$

The following equation is obtained from Eq. (5) with a small alternating current magnetic field $h = h_m \exp(j\omega t)$ and $\Delta\phi = \phi_m \exp(j\omega t)$:

$$\begin{aligned} \phi_m = {} & -(\gamma^2/\mu_0)[M_s h_m \sin(\phi_0 - \beta)] \\ & \times \frac{(\omega_0{}^2 - \omega^2) - j4\pi\lambda\omega}{(\omega_0{}^2 - \omega^2)^2 + (4\pi\lambda\omega)^2}, \end{aligned} \tag{6}$$

where ω_0 is the resonant frequency.

The permeability μ_i is obtained by dividing ΔB, the change in magnetic flux density in the direction of h, by h:

$$\begin{aligned} & \mu_i = \Delta B/\mu_0 h = \mu_i{}' + j\mu_i{}'', \\ & \mu_i{}' = \frac{\gamma^2}{\mu_0^2}\frac{M_s{}^2\sin^2(\phi_0 - \beta)(\omega_0{}^2 - \omega^2)}{(\omega_0{}^2 - \omega^2)^2 + (4\pi\lambda\omega)^2}, \\ & \mu_i{}'' = \frac{\gamma^2}{\mu_0{}^2}\frac{4\pi\lambda\omega M_s{}^2\sin^2(\phi_0 - \beta)}{(\omega_0{}^2 - \omega^2)^2 + (4\pi\lambda\omega)^2}. \end{aligned} \tag{7}$$

If h is parallel to the hard axis for $H_{dc} = 0$, then $\omega_0 = \gamma(H_k M_s/\mu_0)^{1/2}$ for $\beta = \pi/2$, and so

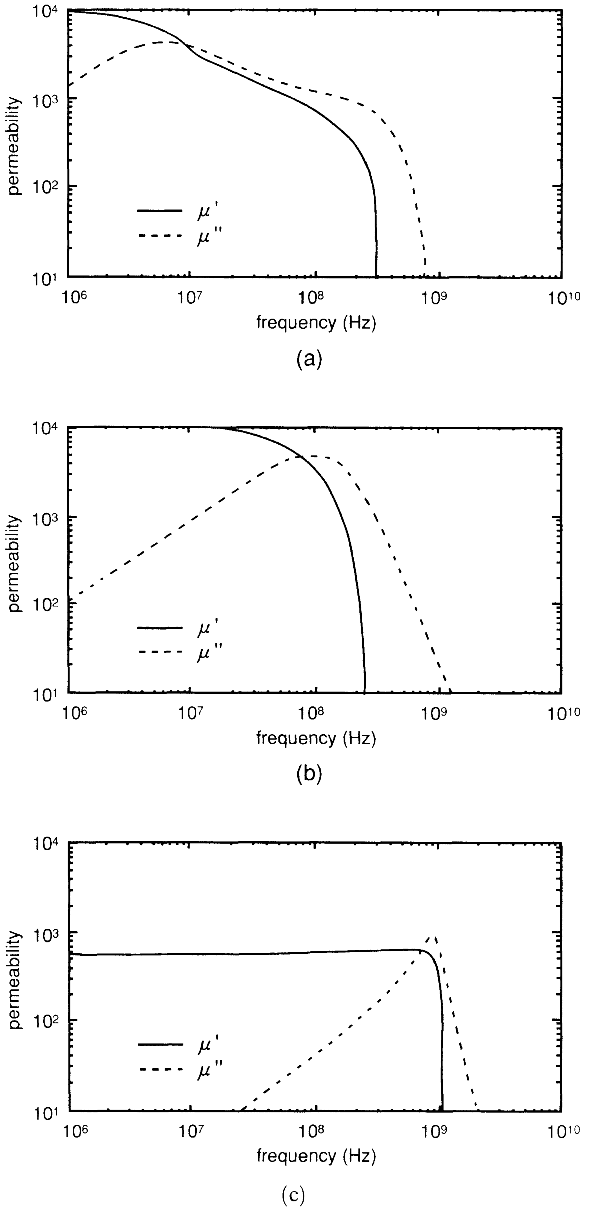

(a)

(b)

(c)

FIG. 3. Permeabilities of magnetic films: (a) $\rho = 120\,\mu\Omega\,\mathrm{cm}, H_k = 80\,\mathrm{A\,m^{-1}}$; (b) $\rho = 3000\,\mu\Omega\,\mathrm{cm}$, $H_k = 80\,\mathrm{A\,m^{-1}}$; (c) $\rho = 3000\,\mu\Omega\,\mathrm{cm}, H_k = 1280\,\mathrm{A\,m^{-1}}$.

the natural resonance frequency of the magnetic thin film depends not only on the anisotropy field H_k but also on saturation magnetization M_s.

For a metallic magnetic thin film with high conductivity, the eddy-current loss is added to the above characteristics. When the permeability is determined by magnetization rotation, the equation of the eddy-current loss is obtained from Maxwell's formula, and the permeability is expressed by

$$\mu_e = \mu_i \tanh(pd/2)/(pd/2),$$
$$p = (j\omega\mu_i/\rho)^{1/2}, \qquad (8)$$

where d is the film thickness and ρ is the film resistivity.

Figure 3 shows the results from Eq. (7), assuming the film thickness $d = 5\,\mu\mathrm{m}$, saturation magnetization $M_s = 10^4\,\mathrm{G}$, anisotropy field $H_k = 80$ or $1280\,\mathrm{A\,m^{-1}}$, and resistivity $\rho = 120$ or $3000\,\mu\Omega\,\mathrm{cm}$, as reported by Shimada

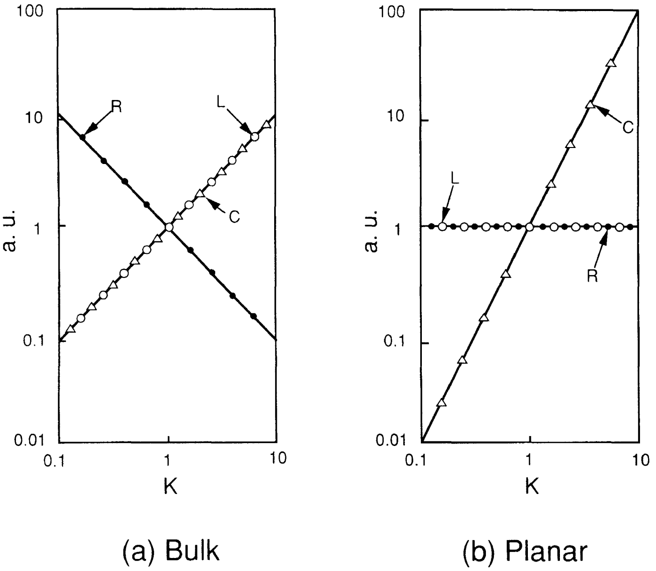

(a) Bulk

(b) Planar

FIG. 4. The inductances, capacitances, and resistances depend on the size reduction factor *K*: (a) bulk device and (b) planar device.

(1993). Figure 3(a) shows the case of a high-permeability magnetic thin film, such as amorphous metal. The decrease in μ' due to the eddy current and natural resonance is clearly seen. Figure 3(b) is the case with increased resistivity, where the loss in the high-frequency region is reduced. Figure 3(c) shows the case where the anisotropy field is also increased with the result that resonance loss is markedly reduced, although permeability is decreased.

As discussed above, it is necessary to control the uniaxial anisotropy precisely, while increasing the resistivity of the magnetic thin film.

1.1.2 Advantages of Micromagnetic Devices

Micromagnetic devices are formed two-dimensionally using magnetic thin film. This feature allows the device to be integrated with solid-state electronic devices such as large-scale integrated circuits (LSIs). This implementation is promising for a variety of applications of such micromagnetic devices. Let us now consider the features of the micromagnetic device, by comparing the characteristics of two- and three-dimensional magnetic devices, taking the inductor as an example.

Let us change the size of the inductor, keeping the number of coil windings constant. Multiply the length, width, and height (l, w, t) of the bulk inductor by *K*, and multiply the length and width (l, w) of the planar inductor by *K*, while keeping its height t' constant because it is a thin-film structure. Then, the inductance, resistance and capacitance of the bulk-type inductor are expressed by

$$L_b \propto wtK^2/lK \propto K$$

$$R_b \propto lK/wtK^2 \propto 1/K$$

$$C_b \propto wlK^2/tK \propto K,$$

while those of the planar inductor are expressed by

$$L_p \propto wt'K/lK = \text{constant}$$

$$R_p \propto lK/wt'K = \text{constant}$$

$$C_p \propto wlK^2/t' \propto K^2.$$

Figure 4 shows their behavior as reported by Shirae *et al.* (1992). As the size of the bulk-type inductor becomes larger, its inductance increases and its resistance decreases, improving its characteristics as a magnetic device. When the size of the planar inductor is reduced, its inductance and resistance remain constant, and its capacitance decreases as K^2. The fact that the inductance of the planar inductor remains constant at a certain value as its size is reduced indicates that a compact planar inductor can be made using a thin-film structure. The fact that the capacitance of the planar inductor decreases indicates that it is suitable for use at high frequencies. These characteristics or features are common to both the planar inductor and the micromagnetic device formed with a magnetic thin film.

1.2 Materials for Micromagnetic Devices

The minimum requirements for a micromagnetic device material are that its loss should be low in the high-frequency region and that it can be made small. Thus, a thin film is the ideal magnetic material.

An amorphous metal film (see AMORPHOUS MATERIALS) is a magnetic material that can be processed into a thin film. Co-based amorphous metal, produced through sputtering deposition, is widely used for micromagnetic devices. This type of material first attracted attention as a magnetic head material because its magnetostriction is nearly zero and its saturation magnetization is large. Highly practical, soft magnetic thin films are now obtained as a result of the success in independently controlling the characteristics of magnetostriction, saturation magnetization, and crystallization temperature through the production of ternary amorphous films such as CoZrTa, CoMnNb and CoFeNb.

Films having multilayer structures have been studied in order to improve their high-frequency characteristics. A soft magnetic film demonstrating high permeability in the high-frequency region is desired to have the magnetic domain structure illustrated in Fig. 5(a). However, the actual soft magnetic single layer takes a thermodynamically stable form as shown in Fig. 5(b). The 180° opposite domains and closing magnetic domains are produced in the specimen to close the magnetic path to decrease magnetostatic energy, thereby preventing a free magnetic pole from being produced at each end of the specimen. The closure magnetic domain decreases the permeability of the magnetic film at high frequencies. Magnetic domains are controlled by using ceramics and thin nonmagnetic films to obtain multilayered soft magnetic thin films as shown in Fig. 5(c). Magnetization is closed between magnetic layers to suppress the generation of free magnetic poles or closure magnetic domains, since magnetostatic coupling occurs at the edges of magnetic layers that are separated by thin nonmagnetic films or ceramics. Thus, the domain geometry will have only 180° domains, resulting in high permeability. For example, with CoBN film, which has a high coercive force, the single layer shows the property of having coercive force in the perpendicular direction, while the multilayer film having in-plane magnetic films shows markedly decreased coercive force, as seen in Fig. 6 (Shirae *et al.*, 1992).

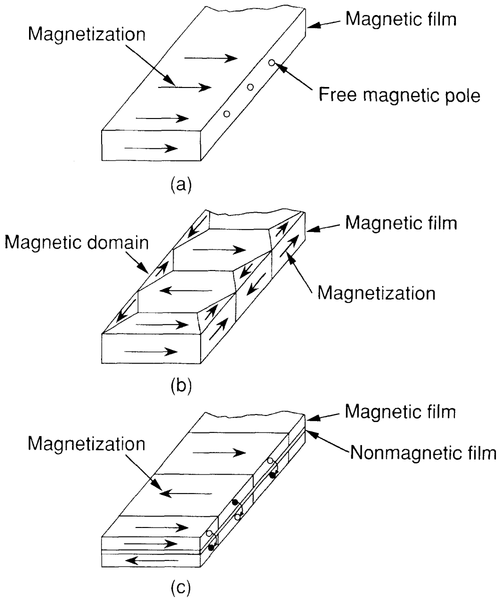

FIG. 5. Schematic diagrams of the magnetic domains in magnetic heads.

A multilayer film can improve the frequency characteristics. The eddy-current loss is proportional to the square of the frequency and the film thickness, and inversely proportional to the resistivity of the magnetic film. Thus, the thinner the film, the lower the loss, resulting in improved frequency characteristics.

1.3 Design of Micromagnetic Devices

Although micromagnetic devices employ many different types of structure, the most basic types are the meander and the spiral. A schematic diagram of the meander-type inductor is shown in Fig. 7. It is composed of a folded meandering conductor pattern and magnetic thin films sandwiching the pattern. The self-inductance of the meander-type in-

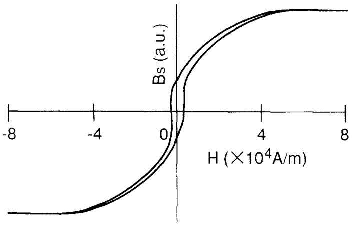

(a) Co·BN film

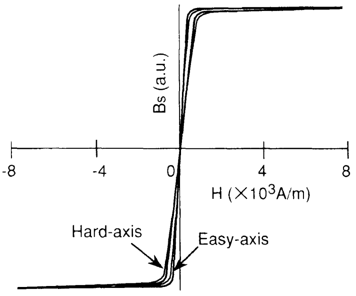

(b) Co·BN/BN multilayers (12nm/3nm) × 200

FIG. 6. *B–H* curves of CoBN single-film and multilayers.

ductor is generally expressed by the following equation, provided that there are $2N$ conductors of length b (Kawabe *et al.*, 1984):

$$L_l = \frac{\mu_0 ab}{2(2N-1)g/N + a^2/(2N-1)N\mu' t_m}, \tag{9}$$

where μ_0 is permeability of the vacuum, N is the number of turns of a pair of conductors, g is the gap between the upper and lower magnetic films, μ' is the relative permeability, and t_m is the magnetic film thickness. The following equation expresses the inductance when the inductance of the folded sections is taken into account:

$$L = L_l(1 + a/2Nb). \tag{10}$$

This reveals that the self-inductance decreases as the pattern width a is reduced while the number of turns is kept constant. Making the inductor more compact markedly decreases its inductance.

Let us now consider the spiral-type inductor. A model of an elliptic spiral-type inductor was shown in Fig. 1(b). In this form a spiral thin-film conductor pattern is sandwiched by the upper and lower magnetic thin films. The following equation expresses the inductance of the elliptical spiral-type inductor if its air core without a magnetic film is ignored (Soohoo, 1979):

$$L = \mu_0 \mu' t_m w_m N^2 / l_m, \tag{11}$$

where w_m is the magnetic film width and l_m is the magnetic path. This reveals that the self-inductance increases as l_m is reduced while the number of turns is kept constant. More specifically, the structure of the elliptical spiral-type inductor is effective for micromagnetic devices because its inductance increases as it is made more compact.

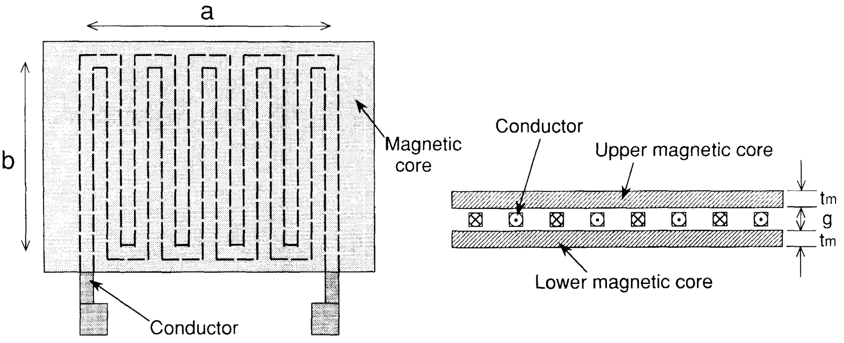

(a) (b)

FIG. 7. Meander-type inductor geometry, in which a planar conductor is sandwiched between magnetic thin film cores.

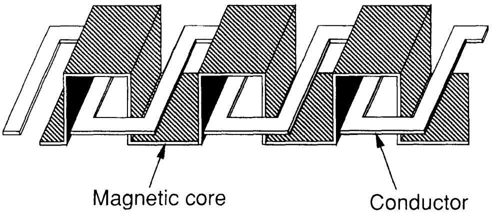

FIG. 8. Toroidal-meander–type inductor geometry, in which a magnetic core is wrapped around a planar conductor.

Structures other than the meander and spiral types include the toroidal-meander (Ahn *et al.*, 1994), cloth (Shirakawa *et al.*, 1990), and outer-conductor (Mino *et al.*, 1992) types. The toroidal-meander type inductor (Fig. 8) is formed by a meandering conductor pattern enfolded with two layers of magnetic film. This structure can reduce the leakage flux without increasing the conductor pattern resistance. The cloth-type inductor (Fig. 9) is structured so that magnetic thin-film strips are sandwiched by two layers of conductor pattern. This structure allows one to form an inductor with less leakage flux in a similar manner to the toroidal-meander type. Both the toroidal-meander and cloth-type inductors only have a low inductance of a few microhenries because they are based on the meander method.

The outer-conductor type transformer (Fig. 10) is structured so that the magnetic thin film is wrapped by the upper and lower conductor patterns via an isolation layer. The primary and secondary coils are formed in parallel and through-holes connect the upper and lower conductor patterns. This type keeps the leakage flux comparatively low, but has the disadvantage of increased conductor pattern resistance due to the contact resistance in the through-holes.

Micromagnetic devices, like LSIs, are usually produced by photolithography. Figure 11 shows the photolithography process. Thin films of magnetic, conductive, and isolation materials are deposited on a substrate through processes such as sputtering deposition and vacuum evaporation, then photoresist is applied to the top film layer. The photoresist surface is exposed to light through a mask pattern made of, for example, a glass plate, and is then developed to produce a photoresist pattern on the top thin film. Etching the thin film, using the photoresist pattern as a protective film, forms the desired thin film pattern on the substrate. Repeating the above photolithography process produces a micromagnetic device composed of thin films over the substrate.

Micromagnetic devices must be made flat because they often handle power and use films thicker than those of the LSI. Figure 12 illustrates an example of the planarization etch-back process. A planar isolation layer is deposited over the conductor pattern formed on the substrate by applying photoresist to the isolation layer through, for example, spin coating, to produce a planar surface. Then the surface is etched back evenly using an etching process with a low etching ratio.

2. TYPES OF MICROMAGNETIC DEVICE

2.1 Planar Inductors and Planar Transformers

2.1.1 Meander Coil Many researchers have studied and experimented on the mean-

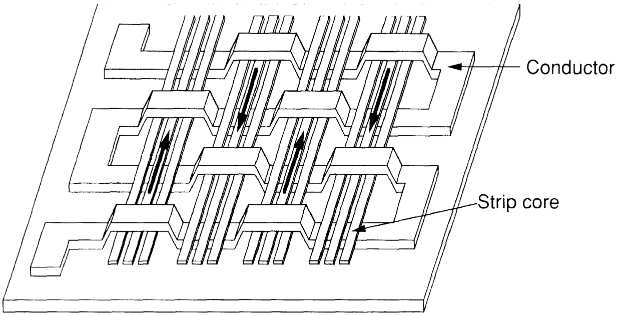

FIG. 9. Cloth-type inductor geometry with a magnetic strip core array.

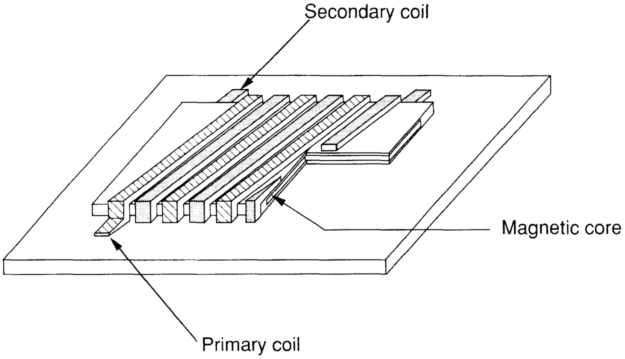

FIG. 10. Outer-conductor-type transformer geometry, in which a thin magnetic film core is wrapped around planar conductors.

der coil because it has a simple structure and is relatively easy to fabricate. Figure 13 shows the results of an experiment conducted by the present authors, as an example of the magnetic properties of the planar inductor using a meander coil. This coil used a copper pattern deposited by sputtering. Amorphous CoZrRe layers deposited by sputtering were used as the two magnetic layers to sandwich the coil. The coil had eight turns with a pattern width of 100 μm and a pattern separation of 50 μm. In addition, the half-turn length of the coil was 3.5 mm and the coil copper film thickness was 5 μm. The magnetic film was 5 μm thick.

The inductance of the meander-type inductor remained constant, up to approximately 10 MHz, and then decreased in the higher-frequency region. The equivalent resistance of the inductor also remained constant, up to approximately 10 MHz, and then increased in the higher-frequency region. The inductance on a chip a few millimeters square was as low as 20–30 nH. Inductance increased in proportion to the number of turns or the half-turn length of the coil. It is, however, extremely difficult with the presently available magnetic material to obtain an inductance of a few microhenries, which is appropriate for a switching power supply with an output as low as 1 W.

2.1.2 Spiral Coil Although its fabrication process is somewhat complex, a micromagnetic device using a spiral coil provides a relatively large value of inductance. The photograph in Fig. 14(a) shows an example of the spiral-type inductor, the top view of the planar inductor manufactured by the present authors.

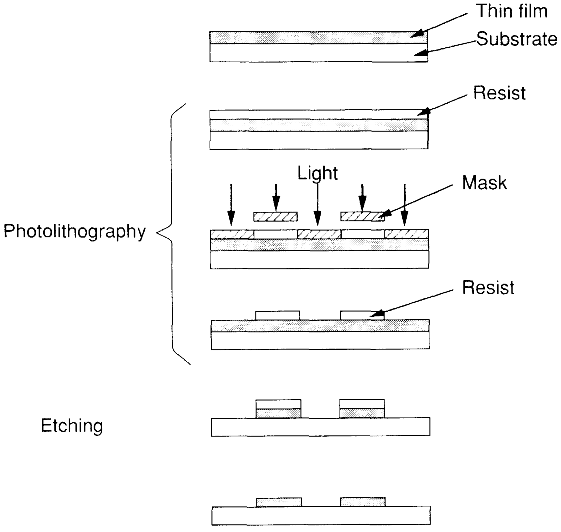

FIG. 11. Fabrication sequence of micromagnetic devices.

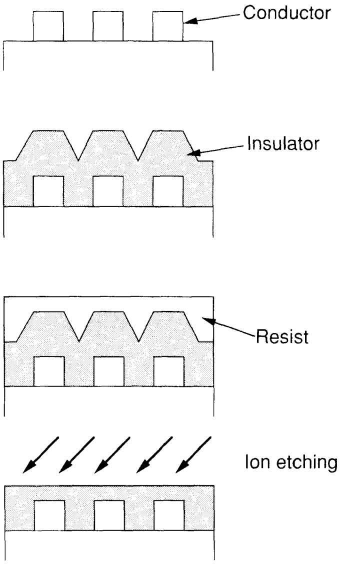

FIG. 12. Planarization process using the dry-etching technique.

Its coil, made of thin copper films, has six turns, and is 100 µm wide. Its magnetic film, made of CoZrTa amorphous film, is a total of 4 mm wide from both sides, and the magnetic path is 3 mm long. Compared with a circular spiral coil, it is easier to provide an elliptical spiral coil with uniaxial anisotropy. Figure 14(b) is a photograph showing the cross-section of the planar inductor. On a silicon wafer a thermal oxidization film was used for the substrate, on which the lower magnetic film was deposited, over which two layers of coil patterns were formed, which was then covered with the upper magnetic film. Polyimide was used for the intermediate insulation film. A multilayer magnetic film structure was used, using SiO_2 films for intermediate insulation films to decrease the eddy-current loss. The coil film was 15 µm thick and the coil separation was 20 µm. A smooth shape with little height variation was obtained, even though relatively thick multilayer films of about 120 µm were deposited on the substrate. A planar transformer was obtained by using the upper and lower coil patterns of the above planar inductor for its primary and secondary coils.

Figure 15 shows the magnetic properties of a spiral-coil type inductor as previously reported (Mino *et al.*, 1996). Similar to that of the meander-type inductor, the inductance of the spiral-coil type planar inductor remains constant, up to 10 MHz, and then decreases in the higher-frequency region. Its equivalent resistance, on the other hand, remains constant up to several hundred kilohertz, and then increases in the higher-frequency region. Thus inductance of a few microhenries, applicable to a converter with an output of approximately 1 W, is obtained with a chip only a few millimeters square. Inductance increases in proportion to the number of turns, but the equivalent resistance increases. The *Q* value has a its maximum of 4.5 at around 1 MHz.

To improve the high-frequency characteristics of the planar inductor, an anisotropy

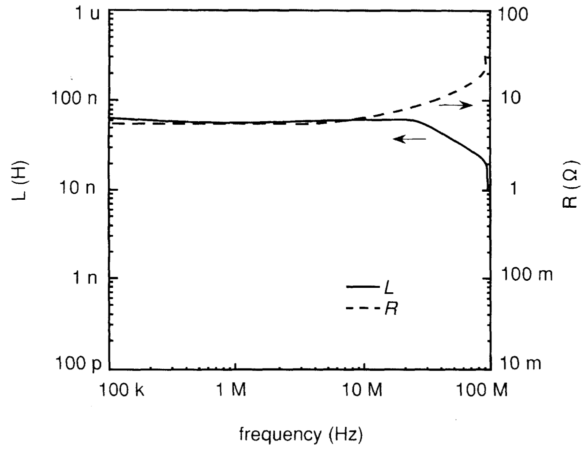

FIG. 13. Characteristics of a meander-type inductor: equivalent inductance and equivalent resistance versus frequency.

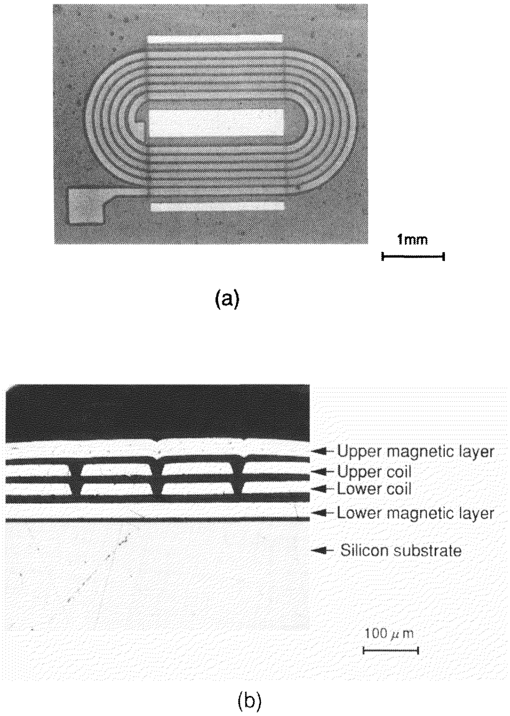

FIG. 14. Spiral-type inductor: (a) top view and (b) cross-sectional photograph.

field is generally induced into the magnetic thin-film deposition. This gives a hard-axis characteristic, providing inductance with a flat characteristic up to approximately 100 MHz. This, however, causes a problem in that the absolute inductance value is lowered by the decreased relative permeability. Because of this, the anisotropy field of the actual planar inductor is adjusted according to the conversion frequency of the switching power supply. Fox example, the desired properties are obtained with an amorphous CoZrTa magnetic thin film by annealing at around 200 °C in a magnetic field.

2.2 Thin-Film Magnetic Heads

This article reviews only thin-film magnetic heads because magnetic recording heads are described elsewhere (see MAGNETIC DEVICES). Figure 16 shows a schematic cross-sectional view of a thin-film head. The thin-film head is composed of spiral coils deposited in an insulator in the middle of the head, and upper and lower magnetic thin films surrounding the coils and the insulator. A read/write gap of 0.2 μm is provided in part of the magnetic film. The entire thin-film head is approximately 20 μm thick, and its gap is approximately 10 μm wide.

Basically, there are two ways of fabricating a thin-film head: the wet method using electroplating and the dry method using ion milling. Both methods are fundamentally the same as the thin-film inductor manufacturing method. The dry-etching method for manufacturing a thin-film head is outlined below. An SiO_2 film is deposited on an Al_2O_3–TiC substrate, and a lower magnetic thin-film pattern is deposited by sputtering. The coil patterns are produced by sputtering a copper film onto an isolation layer. An isolation layer is deposited over the coil patterns for planarization, and a through-hole is placed for contact between the upper and lower magnetic films. Then, an SiO_2 film is deposited in the gap section to form the gap, and the upper magnetic film is deposited. Finally, a protective film is deposited, terminals are opened, and cutting, grinding, and slider processing are carried out.

Today, recording (write) heads exclusively use the inductive-type thin-film head. This is because increases in the speed of magnetic disk drives need to take into account that reproduction (read) output is proportional to the relative speed between the head and the disk. High-sensitivity heads utilizing the magnetoresistance effect (MR) or the giant magnetoresistance effect (GMR) are used for reproduction (read) heads.

2.3 Microactuators and Microstators

The field of micromachines has been actively studied along with the advancement of technology for processing silicon semiconductor devices. Of the various reported ideas concerning micromachines, many use electrostatic force as their driving force, but some use electromagnetic force instead.

Figure 17 shows, as an example of a microactuator using electromagnetic force, a microvalve made of soft magnetic material whose thin film spring moves vertically with an external coil (Yanagisawa *et al.*, 1994). Figure 18 shows the microvalve manufacturing method using photolithography, and film de-

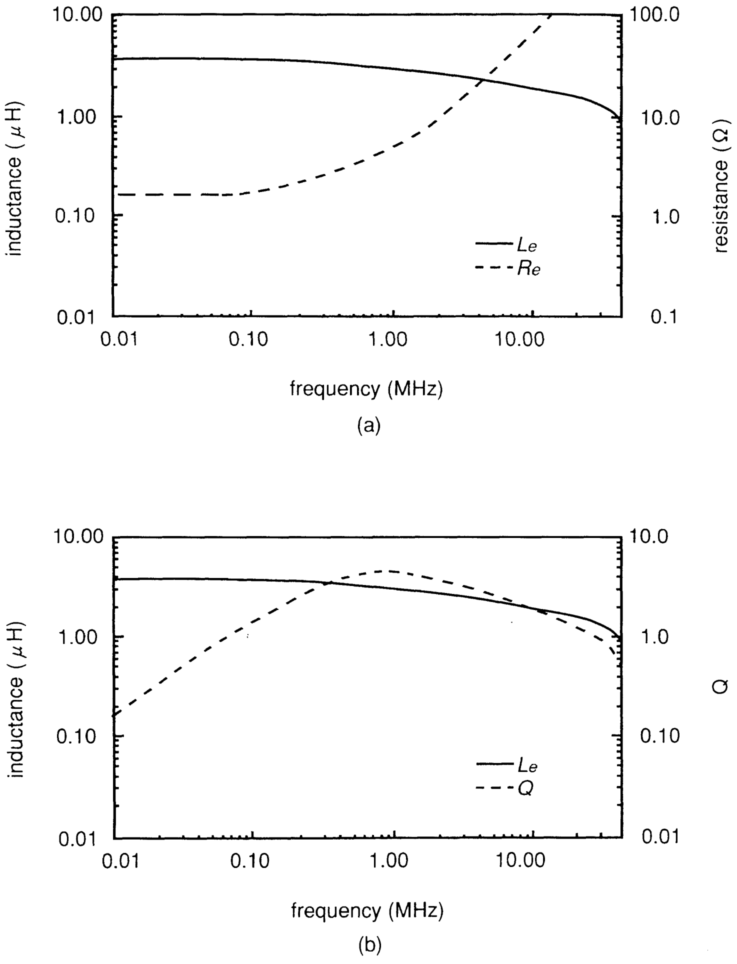

FIG. 15. Characteristics of a spiral-type inductor: (a) equivalent inductance and equivalent resistance versus frequency and (b) equivalent inductance and *Q* versus frequency.

position and patterning by ion beams. An SiO_2 film is deposited on the back of a silicon wafer, and a valve plate of Permalloy film is formed on the front of the silicon wafer. A sacrificial SiO_2 film is deposited on the valve plate, over which is produced a valve cap of Permalloy. Then, the SiO_2 film on the rear side of the substrate is windowed and the silicon is etched anisotropically. Finally, the sacrificial film and the SiO_2 film on the rear side of the substrate are removed, completing the manufacturing process.

Since the microactuator employs an angular-spiral spring structure, the stress is released perpendicular to the spring motion. This ensures precise positioning between the valve plate and the valve cap. The angular-spiral valve cap measures $300 \times 300\,\mu m^2$, the beam width is $20\,\mu m$, and the film thickness is $2\,\mu m$. The valve-plate hole in the center of the valve plate is $30\,\mu m$ in diameter.

Another example of a micromachine using electromagnetic force is the microstator. A larger torque can be obtained with electromagnetic force than with electrostatic force. It is more advantageous to use electromagnetic force for a device a few millimeters in size. Figure 19 shows a schematic diagram of a microstator employed for a microgenerator (Ota *et al.*, 1995). The cylindrical stator is 1 mm in

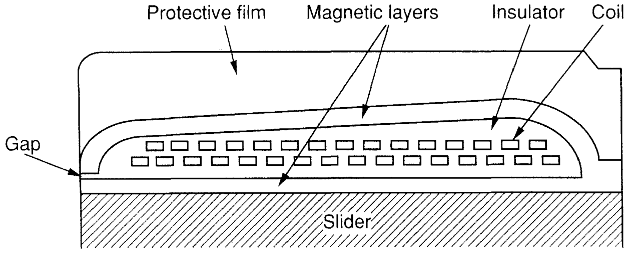
Protective film
Magnetic layers
Insulator
Coil
Gap
Slider

FIG. 16. Schematic cross-sectional view of thin-film magnetic head.

diameter and 0.5 mm high. The microstator is composed of a core die made of Permalloy and six coils manufactured by photolithography and plating that are inserted in the core die. Each coil consists of 20 turns of a copper pattern whose cross-section measures $7 \times 15\,\mu\mathrm{m}^2$. It is reported that a microgenerator 1.2 mm in diameter has been manufactured using the above microstator, and that a voltage of a few millivolts was obtained.

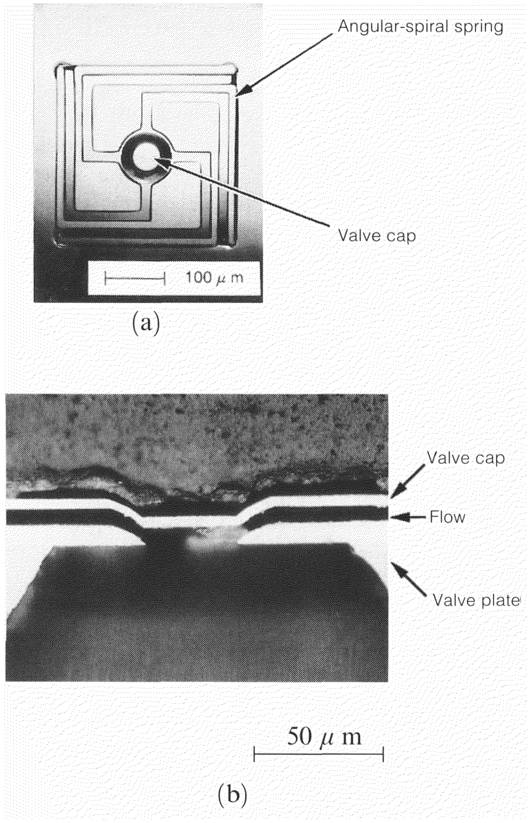
Angular-spiral spring
Valve cap
100 μm
(a)
Valve cap
Flow
Valve plate
50 μm
(b)

FIG. 17. Microactuator: (a) top view and (b) cross-sectional view.

3. APPLICATIONS OF MICROMAGNETIC DEVICES

Interest has been growing in the field of micromagnetics and micromagnetic device applications. Some of the more novel aspects of micromagnetic technology are presented below to illustrate further this diversity.

3.1 Switching Power Supplies

Explosive growth has been seen in the field of portable electronics such as cellular telephones, radio pagers and notebook computers. These devices need more compact and highly efficient switching converters. Micromagnetic components based on thin-film magnetic materials, constructed with microfabrication

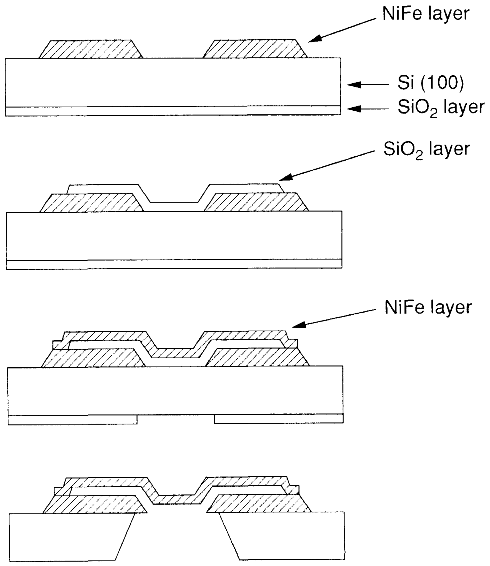
NiFe layer
Si (100)
SiO$_2$ layer
SiO$_2$ layer
NiFe layer

FIG. 18. Fabrication sequence of the microactuator.

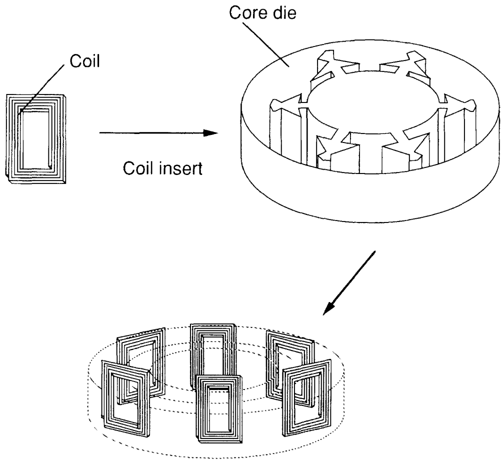

FIG. 19. Schematic diagram of the microstator.

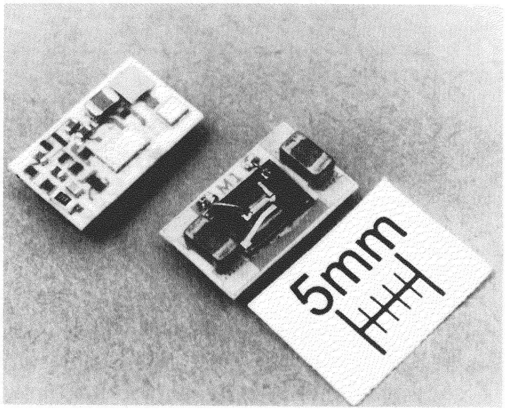

Fig. 20. Microconverter with the micromagnetic inductor.

techniques, show great promise for compact and efficient high-frequency power conversion circuits.

An example of a microswitching converter using a spiral type microinductor is shown in Fig. 20 as previously reported (Mino *et al.*, 1996). It is a one-watt class, compact buck-converter, whose circuit topology is shown in Fig. 21. The circuitry that is shaded and enclosed by the dashed lines is a microinductor. The control circuit, which regulates the output voltage using square-wave pulse-width modulation, is fabricated by BiCMOS 1-μm process technology. This technology keeps the power consumption at 10 mW; it can operate up to 2 MHz. To enable high-frequency operation, a low-capacitance MOSFET switch, a Schottky-barrier rectifier diode, and multilayer ceramic capacitors are used. The semiconductor devices are mounted on one side of a four-layer printed ceramic board with printed resistors, while the microinductor and input and output capacitors are mounted on the other side. These design conditions result in high-frequency operation, making it possible to achieve low power dissipation with an extremely small volume of 0.2 cm^3.

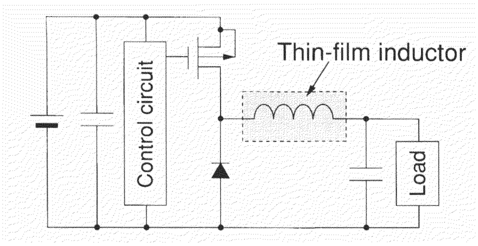

FIG. 21. Circuit diagram of a buck-converter.

The voltage wave forms for the microswitching converter are shown in Fig. 22, where V_{in} is the input voltage, V_{G} is the gate voltage of the MOSFET, V_{D} is the voltage across the diode, V_{out} is the output voltage, and V_{ac} is the output ripple voltage. Switching

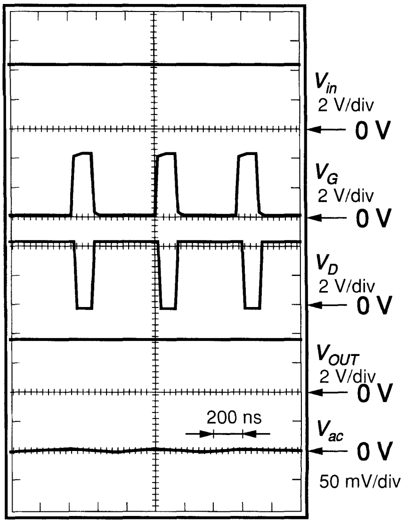

FIG. 22. Wave forms of the microconverter.

frequency is 1.6 MHz, V_{in} is 4.4 V, V_{out} is 3.3 V, output current I_{out} is 0.12 A, and output power is 0.4 W. The wave form of the output ripple voltage is very small. The maximum overall efficiency reached 85%.

Another microconverter using a spiral-type transformer has also been reported (Sullivan *et al.*, 1996). It achieved a power-handling density of 22.4 W cm^{-3} at 61% efficiency. The circuit topology was a simple half-bridge forward converter. These microconverters using micromagnetic devices are expected to meet the demands of both today's and tomorrow's portable-equipment market.

3.2 Magnetic Disk Storage

Because magnetic disk-drive technology is a key technology supporting today's computer age, efforts have been increasing the bit-packing density of drives, boosting recording density by a factor of more than 10^4 during the past 30 years. The thin-film head, the essential component of the magnetic disk drive, contributed greatly to the development of drives in the 1980s. Figure 23 shows an arm with a thin-film head. The thin-film head is mounted on a slider at the tip of the arm.

The combined magnetic head, which has a recording (write) head with a coil and MR or GMR heads for reproduction (read), was brought into use in the early 1990s as described in Sec. 2 (see Fig. 24). A recording density of 0.5 Gbit cm^{-2} was attained with a track width of 1.5 μm. This translates into a capacity of 5–50 Gbytes for a 3.5-in. hard disk drive.

The GMR head is considered to be more promising than the thin-film head as the magnetic head of the future. However, micromagnetic technology using the planar coil remains the key technology for the GMR head.

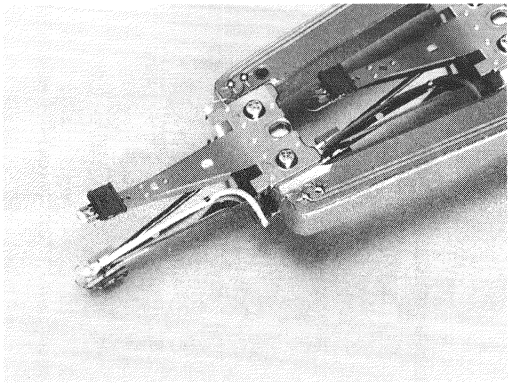

FIG. 23. Magnetic heads.

3.3 Micromachines

The study of micromachines has become suddenly active since the early 1990s, but micromachines have not yet been put to practical use. Although many reports on micromachines concern the use of electrostatic force, some reports involve micromagnetic technology.

Figure 25 shows the basic configuration of a microvalve utilizing an external coil, described in Sec. 2. The thin-film spring, made of soft magnetic material, is moved vertically by the external coil to control the flow of fluids. Figure 26 illustrates an example of fluid flow control by the microvalve shown in Fig. 25. Air was used as the motive gas. The valve was operated at driving frequencies of 0.1–100 Hz, and changes in the pressure were observed. A good flow response characteristic was obtained.

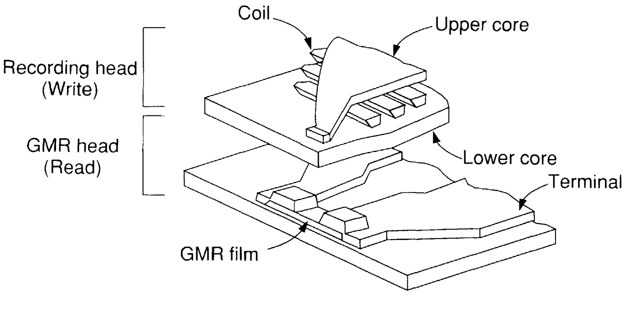

FIG. 24. Schematic diagram of giant magnetoresistive head.

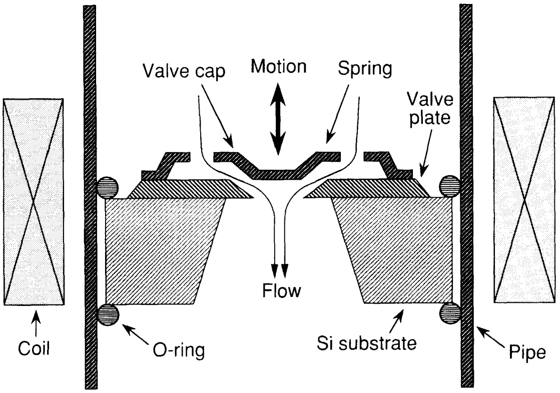

FIG. 25. Schematic cross-sectional view of the magnetic microvalve.

Further developments in the field of micromachines are expected as we enter the 21st century. The authors hope that great efforts in micromagnetic technology research and development will be made in the future.

ACKNOWLEDGMENTS

The authors wish to acknowledge the help afforded by Dr. H. Kuwano and Mr. K. Yanagisawa through frequent discussions.

GLOSSARY

Amorphous Metal: A glass-like metallic alloy with a random atomic arrangement. Excellent corrosion resistance, strength, and magnetic properties because of random atomic arrangement, which is different from an alloy consisting of fine crystals. Used for devices such as high-performance magnetic heads and low-loss transformers.

Eddy-Current Loss: Energy loss caused by an eddy current flowing through a metallic material.

Inductive Head: A head that uses the principle of electromagnetic induction to write to and read from a magnetic disk.

Magnetic Anisotropy: The phenomenon in which a thermodynamic quantity of a magnetic crystal varies with the direction of magnetization against the crystal axis. Often referred to as the phenomenon in which the internal energy varies with the spontaneous magnetization inside a ferromagnetic crystal.

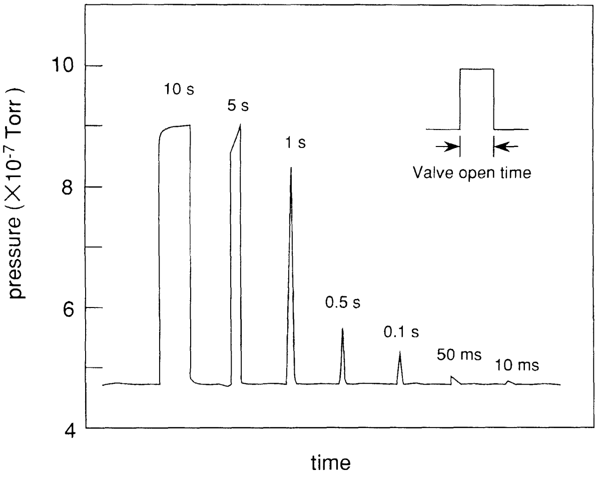

FIG. 26. Example of fluid flow characteristics using the magnetic microvalve. (1 Torr = 133.322 Pa.)

Magnetic Thin Film: Magnetic alloy deposited from 100 to 1000 nm thick over a substrate by vacuum evaporation or electrodeposition. If the deposition process is performed in a magnetic field, inductive magnetic anisotropy will be produced, and the magnetization will take place in the positive or negative direction along the easy axis.

MR (Magnetoresistive) Head: A read-only head utilizing the anisotropic magnetization effect.

Sputtering Deposition: A method of depositing a film over a substrate by sputtering atoms from a target material with ions of argon or other gas discharged in a vacuum.

Works Cited

Ahn, C. H., Allen, M. G. (1994), "A New Toroidal-Meander Type Integrated Inductor With A Multilevel Meander Magnetic Core," *IEEE Trans. Magn.* **30**, 73–79.

Kawabe, K., Koyama, H, Shirae, K. (1984), "Planar Inductor," *IEEE Trans. Magn.* **20**, 1804–1806.

Mino, M., Tsukamoto, K., Yanagisawa, K., Tago, A., Yachi, T. (1996), "A Compact Buck-Converter Using a Thin-Film Inductor," in: *Proceedings of the IEEE Applied Power Electronics Conference, San Jose*, Parsippany, NJ: IEEE, pp. 422–426.

Mino, M., Yachi, T., Tago, A., Yanagisawa, K., Sakakibara, K. (1992), "A New Planar Microtransformer for Use in Micro-Switching Converters," *IEEE Trans. Magn.*, **28**, 1969–1973.

Ota, H., Oda, T., Kobayashi, M. (1995), "Development of Coil Winding Process for Radial Gap Type Electromagnetic Micro-Rotating Machine," in: *Proceedings of Micro Electro Mechanical Systems*, pp. 197–202.

Saleh, N., Qureshi, A. H. (1970), "Permalloy Thin-Film Inductors," *Electron. Lett.* **6**, 850–852.

Shimada, Y. (1993), "Magnetically Soft Thin Films for Micro-Magnetic Devices," *J. Magn. Soc. Jpn.* **17**, 667–671.

Shirae, K., Arai, K., Shimada, Y. (1992), *Micromagnetic Devices*, Tokyo: Kogyo-Cyosa-Kai.

Shirakawa, K., Yamaguchi, K., Hirata, M., Yamaoka, T., Takeda, F., Murakami, K., Matsuki, H. (1990), "Thin Film Cloth-Structured Inductor for Magnetic Integrated Circuits," *IEEE Trans. Magn.*, **26**, 2262–2264.

Smith, D. O. (1958), "Magnetization Reversal and Thin Films; Static and Dynamic Behavior of Thin Permalloy Films," *J. Appl. Phys.* **29**, 264–273.

Soohoo, R. F. (1979), "Magnetic Thin Film Inductors for Integrated Circuit Applications," *IEEE Trans. Magn.*, **15**, 1803–1805.

Sullivan, C. R., Sanders, S. R. (1996), "Measured Performance of a High-Power-Density Microfabricated Transformer in a DC-DC Converter," in: *Records of IEEE Power Electronics Specialists Conference, Baveno*, Parsippany, NJ: IEEE, pp. 287–294.

Yanagisawa, K., Kuwano, H., Tago, A. (1994), "An Electromagnetically Driven Microvalve," in: *Proceedings of the 7th International Conference on Solid-State Sensors and Actuators, Koube*, Tokyo: IEEJ, pp. 102–105.

Further Reading

Brown, W. F., Jr. (1963), *Micromagnetics*, New York: Interscience Publishers.

Brown, W. F., Jr. (1966), *Magnetoelastic Interactions*, New York, Berlin, Heidelberg: Springer-Verlag.

Information on design and new applications and designs for micromagnetic devices can generally be found in conference proceedings for applications to specific technology areas. The International Magnetic Conference, the Power Electronics Specialists Conference, the IEEE Micro Electro Mechanical Systems Workshop, the IEEE Transactions on Magnetics, and the IEEE Transactions on Power Electronics are good sources of related designs, fabrications and applications for micromagnetic devices.

NITRIDES

D. P. Thompson, *Materials Division, Department of Mechanical, Materials and Manufacturing Engineering, University of Newcastle, Newcastle upon Tyne, United Kingdom*

Introduction 65
1. **Classification** 66
1.1 Definition 66
1.2 Bonding 66
1.3 Common Structure Types 66
2. **Preparation** 67
2.1 Preparation of Nitride Powders 67
2.2 Preparation of Nitride Thin Films 68
3. **Ionic Nitrides** 68
3.1 Lithium Nitride 68
3.2 Group II Metal Nitrides 69
4. **Covalent Nitrides** 69
4.1 Silicon Nitride 69
4.2 Aluminum Nitride 70
4.3 Other Group III Nitrides 70
4.4 Boron Nitride 71
5. **Transition Metal Nitrides** 71
5.1 Bonding and Structure 71
5.2 Classification 72
5.3 Titanium Nitride 72
6. **Complex Nitrides** 72
6.1 Mixed Metal Nitrides 73
6.2 Mixed Nonmetal Nitrides 74
7. **Applications** 74
7.1 Thin Films and Coatings 74
7.2 Bulk Nitrides 75
8. **Summary** 76
Glossary 77
Works Cited 78
Further Reading 79

INTRODUCTION

Nitrogen is one of the most familiar elements known to man. In its gaseous molecular form, N_2, it comprises some 80% of the Earth's atmosphere; moreover, it is essential to life, and provides an important area of study in disciplines such as agriculture, biology and food technology. Here, the nitrogen is normally present in the form of groups such as ammonium (NH_4), nitrate (NO_3), and amide (NH_2). The term *nitrides* is reserved for those compounds in which nitrogen is acting as a nonmetal in its own right, not as a group in combination with either oxygen or hydrogen, and in that capacity exercises its conventional valence of three (very occasionally five). Nitrogen does not occur naturally in the form of nitrides because most nitrides react very rapidly with water even at room temperature, with evolution of ammonia. For this reason, they must be prepared synthetically, under controlled conditions of water vapor, oxygen, and other oxygen-containing species, and need to be kept under similarly controlled conditions if subsequent hydrolysis is to be avoided. Many of the uses of nitrides rely on the presence of a thin passivating layer of surface oxide to prevent attack by atmospheric moisture.

The difficulties associated with the preparation and handling of nitrides are an important reason why, despite the abundance and familiarity of nitrogen, the detailed study of nitrides has developed relatively recently, in response to the general progress of science and technology in requiring materials with specific combinations of properties that only nitrides can provide. Table 1 summarizes the main properties of nitrides that have caused them to find application in the last 50 years.

Adamantine (i.e., diamond-like) nitrides (Parthé, 1972) are of interest because of their combination of high melting point, high hardness, and good thermal conductivity, and silicon nitride and aluminum nitride fall into this category, along with the recently developed carbon nitrides. Transition-metal nitrides (Toth, 1971) were originally far less important than their structurally related carbide analogs, but their occurrence as inclusions in steels (Obara and Tsunoyama, 1995), their use as hard coatings on metals (Sundgren and Hultman, 1995), and their novel magnetic proper-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Table 1. Useful properties of nitrides.

Property	Example (s)
High melting point	BN, CN_x, TiN, ZrN
High-temperature structural properties	Si_3N_4
Hardness	TiN, ZrN
Thermal conductivity	AlN
Ionic conductivity	Li_3N
Semiconductor properties	AlN, GaN, InN
Electronic insulating layers	Si_3N_4, AlN
Magnetic properties	$Ln_2(Fe,M)_{17}N_{3-x}$, M = transition metal

ties (Hadjipanayis *et al.*, 1992) have resulted in their receiving a considerable level of attention recently. The drive to find new materials for electronic applications has resulted in the preparation of nitrides in the form of thin films; aluminum nitride has emerged as a good electronic substrate material because of its combination of low electrical conductivity but high thermal conductivity. More recently, gallium and indium nitrides have received attention because their band gaps make them useful for light-emitting diode (LED) applications. Lithium nitride is the sole stable alkali-metal nitride, and finds limited application as an ionic conductor.

1. CLASSIFICATION

1.1 Definition

The definition of nitrides given above excludes compounds in which the nitrogen is present combined as a group with other elements (especially oxygen), but includes compounds where other nonmetals may be present. Examples of the latter include nitrofluorides (*e.g.*, Ba_2NF, Mg_2NF, Mg_3NF_3) and oxynitrides [*e.g.*, ALON aluminum oxynitride spinel, $\approx 3Al_2O_3 \cdot AlN$; sialon ceramics (Ekström and Nygren, 1992)]. Mixed metal nitrides have been developed and are finding applications; a special group of these comprises nitrided interstitial alloys (*e.g.*, $Sm_2Fe_{17}N_{3-x}, NdFe_{12}N_y$), where the small amount of nitrogen introduced exercises an important role in modifying magnetic properties (Nagata and Fujii 1991).

1.2 Bonding

A convenient way of classifying nitrides is in terms of their bonding. The Group I and Group II metals that form nitrides produce compounds that contain essentially ionic bonds, and display very rapid hydrolysis reactions. Elements in Groups III and IV form covalently bonded nitrides, which are often compared with diamond in the combination of properties they exhibit. Transition metals form nitrides that are interstitial in form and show intermediate properties between their parent metals and the covalent nitrides in Groups III and IV. Rare-earth nitrides show mixed behavior in between those of the transition metal nitrides and the ionic nitrides. More complex ternary and quaternary nitrides have been prepared that exhibit bonding made up of ionic, covalent, and metallic contributions, in proportion to the particular metallic elements they comprise.

1.3 Common Structure Types

Morgan (1977) made a detailed study of bonding in nitrides and concluded that it was surprising in view of the considerable electronegativity difference between oxygen and nitrogen to find that most nitride structures were the same as those observed in oxide crystal chemistry. Thus many transition-metal and rare earth nitrides have the rocksalt (NaCl or MgO) structure, TaON has a rutile (TiO_2) structure, UN_2 has a fluorite (ZrO_2) structure and silicon nitride has the phenacite (Be_2SiO_4) structure. This similarity in structure is very convenient for the classification of nitrides (Marchand *et al.*, 1991). Clearly, the generalisation that oxides are "more ionic" and nitrides are "more covalent" is not a classification that can be translated into structural implications.

Lang (1977) used cation coordination as a convenient parameter for categorizing the structures of nitrides. Thus in lithium nitride (Sec. 3.1), lithium atoms are 2- and 3-coordinated by nitrogen, whereas in boron nitride (graphite structure) the boron atoms are 3-planar coordinated by nitrogen atoms in the form of a triangle, with much weaker bonds joining them to nitrogen atoms in adjacent layers.

Metal atoms 4-coordinated by nitrogen represent perhaps the most common situation in nitrides. Nitrogen prefers to stack in a hexagonal close-packed manner compared with the cubic close-packed arrangements observed for the other elements in Group V. Thus 1:1 compounds involving Group III metals and nitrogen usually adopt the wurtzite structure, exemplified by AlN, GaN and InN. This arrangement forms the basis of more complex superlattices for mixed metal nitrides represented in group number terminology by the formulas II/IV/N_2 and I/IV_2/N_3 (Lang, 1977). Divalent metal nitrides of formula M_3N_2 commonly adopt the antibixbyite structure (Sec. 3.2), and metal nitrides of overall formula M_2N (*e.g.*, LiMgN, LiZnN, Li_3AlN_2) generally adopt the antifluorite structure or ordered variants of this.

Carbon nitrides occur either with graphite-like layers or with a silicon nitride arrangement. Silicon and germanium nitrides are isostructural, and both occur in two forms α and β, with arrangements related to the mineral phenacite (Wild *et al.*, 1972). The structures of sialon ceramics are extremely varied (Thompson, 1989) and in many cases can be considered as nitride analogs of the mineral silicates. This is discussed in more detail in Sec. 4.1.

Metal atoms 6-coordinated by nitrogen are most commonly observed in transition-metal nitrides. Structures are determined by the radius ratio ($r_N : r_M$), and if this is less than ≈ 0.59, simple structures with the metal atoms in hexagonal or cubic close-packed arrangements occur, whereas above this value, more complex structures result (Evans, 1964). Mixed metal nitrides with the nitrogen in 6-coordination often display the perovskite structure (Lang, 1977).

Specific details of the structures of the more important nitrides are given later under the relevant sections.

2. PREPARATION

Routes for preparing nitrides can be divided into two groups depending on whether the final product is a powder or a thin film or surface coating. Nitride powders are often subsequently converted into bulk solids by appropriate densification procedures. Techniques for this are dealt with in the sections covering the particular nitrides concerned. There is also some interest in the production of single crystals of nitrides, especially for electronic applications (Motojima and Iwanaga 1992). Gas-phase growth by CVD (chemical vapor deposition) (*q.v.*) is generally easier than growth from the melt.

2.1 Preparation of Nitride Powders

Direct nitridation of metals using molecular nitrogen is the most commonly used technique for the production of nitride powders, especially when the metal melts at high temperature; this has been used extensively for the production of the nitrides of lithium, silicon, and aluminum. Even though ammonia can be used for nitriding, and offers advantages of a higher nitrogen potential and hence lower nitriding temperatures, it has the disadvantage that only the surface of the metal powder grains is nitrided, and diffusion into the centers of the grains requires higher temperatures or long times. Nevertheless, many ionic nitrides can be successfully prepared in pure form by reacting appropriate metal salts with ammonia.

Carbothermal reduction of oxides in nitrogen (Van Dijen *et al.*, 1985) is generally the cheapest method of nitride production and, because of the large volumes of carbon monoxide generated, always produces a powder as the finished product. This process has been used for producing AlN, Si_3N_4, and BN powders, and is also a cheap process for producing more complex nitrides and oxynitrides starting from aluminosilicates or more chemically diverse naturally occurring starting materials. An excess of carbon is generally beneficial; this is burnt off later at a temperature low enough not to oxidize the nitride. Transition metal nitrides can also be made by the carbothermal route, but the final product is invariably deficient in nitrogen, and contains either carbon or oxygen depending on the metal:carbon ratio in the starting mix.

The purest nitride powders are produced by methods in which metal–nitrogen complexes are synthesized by low-temperature chemical routes, and are then subsequently decomposed by heating to higher temperatures to form the nitride (Büchner *et al.*, 1984). In the case of silicon, either silicon amide, $Si(NH_2)_4$, or silicon imide, $Si(NH)_2$, is formed by reaction between silicon tetrachloride and ammonia in ammoniacal solution, and this is then heated to form initially an amorphous powder at ≈900 °C, which then subsequently crystallizes into the nitride on heating to higher temperatures. It is also possible to prepare low-temperature Si-N–based polymers starting with carbosilazane derivatives, which on decomposition break down to give silicon nitride at high temperatures. In general these methods are expensive, especially when large volumes of powder are being produced; they also require careful attention to safety features because of the presence of chlorine and ammonia; nevertheless, they yield powders of high purity and controlled grain size. The production of pure, ultrafine (*i.e.*, nanometer range) particles is becoming increasingly important, and CVD-based routes allow particles to be produced of uniform size in the range 10–500 nm (Kato 1992). Alternative methods involving the use of either dc or radio-frequency coupled plasmas (Leflamme *et al.*, 1992) are increasingly being introduced.

2.2 Preparation of Nitride Thin Films

The last five years has seen an enormous increase in the number of methods that have been used to produce nitride thin films or coatings. In this article, the term coating is used to denote a thin film that has been deposited to provide improved corrosion resistance and/or mechanical protection, and where it is important to have a strong bond to the underlying substrate; the term thin film is used to denote layers of material deposited on a substrate where the properties of the deposited layer provide the principal interest. The latter term is more commonly used to denote thin films deposited for the purpose of imparting particular electronic or magnetic properties.

CVD techniques (Hoffmann, 1994) were some of the first to be used to produce nitride thin films; these have been further developed using novel precursors (Gordon, 1994), and techniques have been devised to speed up the rate of deposition (Lackey *et al.*, 1991). However, CVD procedures often require high temperatures, which, in addition to being expensive, may be beyond the capability of the substrate to withstand. For this reason they have given way to a range of PVD (physical vapor deposition) techniques, which allow much lower temperatures to be used. Sputtering (Howson, 1994) is a technique that is cheap, easy, and safe to use, but cannot easily control the chemistry and grain size of the final film. It is feasible to produce mixed nitrides by this technique (Belkind *et al.*, 1991), and concentration gradients can be introduced (Monaghan *et al.*, 1993) Plasma-based methods (Korhonen *et al.*, 1990; Jehn, 1994), pulsed laser deposition (Donley and Zabinski 1994), and arc ion plating (Khono *et al.*, 1993; Vetter 1995) are commonly employed to deposit hard-metal coatings. Ion-implantation routes (Bolse 1994; Takano and Isobe, 1991; Armour, 1994) provide an alternative way of nitriding a metal surface, and this can also be used to achieve varied compositions (Lieb *et al.*, 1994).

3. IONIC NITRIDES

The number of compounds in this category is small because of the poor stability of alkali-metal nitrides. Lithium forms an ionic nitride of formula Li_3N, but the other Group I metals do not form equivalents; instead they form azides (NaN_3, KN_3, *etc.*) of limited stability. In Group II, beryllium nitride is a fairly covalent solid, but is dealt with here for comparison with the other nitrides in Group II. Ionicity increases on moving down Group II, and all metals in this group form stable nitrides.

3.1 Lithium Nitride

Figure 1 shows the unusual crystal structure of lithium nitride (Zintl and Brauer 1935). In most compounds, lithium adopts fourfold coordination, but clearly this is impossible in Li_3N because the nitrogen would be 12-coordinated by Li. Instead, the two 3-coordinated lithium atoms occupy positions on triad axes in the basal plane of the hexagonal unit cell with nitrogen occupying the origin site, and with the third (2-coordinated) lithium atom linking adjacent layers of these planes in the *z* direction. As a result, each nitrogen atom is 8-

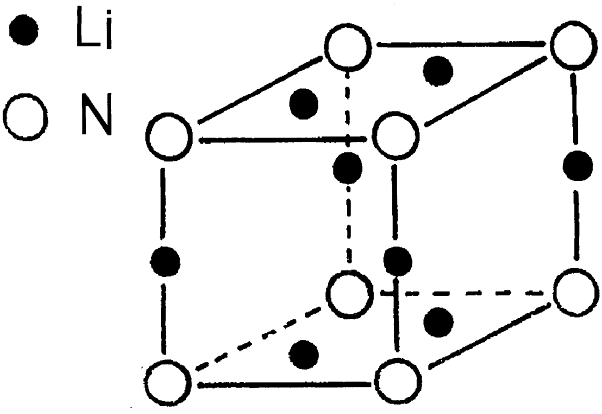

FIG. 1. Atomic arrangement in the hexagonal structure of Li_3N.

coordinated by lithium atoms. Small amounts of other metals (Cu, Co, Ni) can substitute in the 2-coordinated lithium site. The closeness of the 3-coordinated lithium atoms in the basal plane permits easy diffusion from one site to the next, and this forms the basis of the ionic conductivity observed for this compound. As with ionic nitrides, lithium nitride hydrolyses readily in the presence of water with emission of ammonia gas.

3.2 Group II Metal Nitrides

Group II nitrides have the general formula M_3N_2, and the basic structure is the cubic antibixbyite (anti-Mn_2O_3) type, which adopts the space group I*a*3 (Lang, 1977). This structure can be considered as comprising eight fluorite unit cells placed in a 2 × 2 × 2 arrangement along each crystallographic axis, but in each subcell, one quarter of the metal atoms have been removed in an ordered fashion, thus changing the metal:nonmetal ratio from 4:2 to 3:2. As a result, the cations are in a fairly regular 4-coordinated environment, wheras the nitrogen atoms are in a distorted 6-coordination.

Beryllium nitride occurs in two structural modifications, α and β. The α form has the antibixbyite structure discussed above, whereas the β form is hexagonal, with a polytypic structure (Thompson, 1989). This compound forms the basis for a series of polytypoid ternary and quaternary compounds in the Be–Si–O–N system (Roebuck and Thompson, 1977). Magnesium nitride, Mg_3N_2, occurs only in the antibixbyite form, and is characterized by a greenish-brown color.

Calcium nitride occurs in several uncharacterized forms in addition to the antibixbyite type (Lang, 1977). It is very surprising to find calcium in a 4-coordinated environment; moreover, the Ca–N distances are large, and this, combined with the unoccupied sites characteristic of the antibixbyite structure, combine to give the compound a low density. Its sensitivity to moisture means that it must be handled very carefully to avoid hydrolysis.

Strontium forms two nitrides of composition SrN and Sr_2N, the former having a sodium chloride structure and the latter a rhombohedral variant with nitrogen atoms omitted (Lang, 1977). Barium nitride is similar, and data for the various barium nitrides have been published, but further characterization is needed (Linke and Schroedter, 1975).

4. COVALENT NITRIDES

4.1 Silicon Nitride

Forty years ago, silicon nitride appeared as one of several new "special ceramics", which showed promise for development as a high-temperature engineering material because of a good combination of mechanical properties up to ≈1400 °C. Government research programs carried out in the U.K., the U.S.A., Europe and Japan confirmed this promise, but showed that the production of reliable components from this material was difficult because of its poor sinterability and the very high temperatures ($\approx 1750\,^\circ$C) needed. As a result, silicon nitride has subsequently found more applications at low ($< 1000\,^\circ$C) temperatures, where its good wear resistance combined with good thermal shock and chemical resistance are the principal advantages.

An important feature of silicon nitride chemistry is the extent to which simultaneous replacement of silicon and nitrogen by aluminum and oxygen can occur to produce a wide range of alternative structures. These materials (sialons) more readily sinter to high density at lower temperatures than silicon nitride itself because of the presence of low–melting-temperature metal-silicon-aluminum oxynitride liquid phases, which aid densification. The name β-sialon is given to the range of solid solutions of composition $Si_{6-z}Al_zO_zN_{8-z}$, where $0 < z < 4$, which form by direct replacement of silicon and nitrogen in β-Si_3N_4 by aluminum and oxygen; the related compound, α-sialon, has a composition $M_xSi_{12-m-n}Al_{m+n}O_nN_{16-n}$, and has a unit cell similar to that of α-silicon nitride but with $m+n$ Al atoms replacing Si, and valency balance being restored by inserting x metallic cations (M = lithium, magnesium, calcium, yttrium and most rare earths) into the large interstices in the structure. Ranges for the various compositional parameters in the formula are approximately $0 < x < 2$, $0 < n < 3$, and $0.5 < m < 3.5$.

In the early years of silicon nitride development, germanium nitrides were also prepared and it was found that there were also two forms, α and β, isostructural with the α and β forms of silicon nitride (Wild *et al.*, 1972). However, the germanium analogs are not as stable, decomposing into the elements below 1000 °C.

The importance of silicon nitride is such that a complete article of the Encyclopedia has been devoted to it; see SILICON NITRIDE. Further information on sialon and other oxynitride derivatives is given in Sec. 6.2.

4.2 Aluminum Nitride

Aluminum nitride is a refractory, covalently bonded nitride, stable up to temperatures of $\approx 2500\,^\circ$C. It is prepared in powder form either by the nitridation of aluminum powder, by carbothermal reduction of alumina, or by decomposition of preprepared organic derivatives containing aluminum and nitrogen. Aluminum nitride powder hydrolyses easily even at room temperature, but solid samples resist hydrolysis because of partial passivation. Bulk samples are prepared by sintering the powder either at high temperatures (1750–1950 °C) with additions of small amounts of metal oxides (most commonly MgO, CaO, Y_2O_3). The oxygen on the surface of the AlN grains reacts with the oxide additive to form a liquid at sintering temperatures, which aids densification, and then cools to form mainly a crystalline ternary oxide phase in the grain boundaries.

Even though aluminum nitride was originally developed for its refractoriness and mechanical properties, it does not find extensive application in this area. A more important area of application is as an electronic substrate, where its high thermal conductivity (up to $200\,\mathrm{W\,mK^{-1}}$) is the main property of interest. However, dense samples of AlN sintered using oxide additives exhibit much lower thermal conductivities (typically $50\,\mathrm{W\,mK^{-1}}$) because of residual oxygen present both in the aluminum nitride grains and also in the grain boundaries. To produce a high-conductivity material, it is necessary to remove oxygen (especially lattice oxygen within the grains), and this can be achieved either by a higher temperature post-treatment, or by use of reducing additives that extract the oxygen in gaseous form during the sintering stage (Kuramoto *et al.*, 1986).

4.3 Other Group III Nitrides

The preference of nitrogen for a hexagonal close-packed stacking sequence results in the wurtzite (AlN-type) structure being adopted by both gallium and indium nitrides, even though the cubic form can occur for GaN. As with AlN, small amounts of oxygen easily remain dissolved in the crystal lattice. The main interest in both compounds is in their electronic properties, because their band gaps are 3.5 and 1.9 eV, respectively. GaN is a good source material for high-brightness blue-light–

emitting diodes; InN is used for similar device applications and also for high-efficiency solar cells.

4.4 Boron Nitride

Boron nitride is a unique material, which in many ways can be regarded as a white, insulating form of graphite. Normally prepared by reaction between boric oxide and ammonia, it is produced as a white powder, with a graphite-like crystal structure, which is stable in reducing environments up to ≈3000 °C. In steam BN hydrolyses to boric oxide and oxidizes in air at higher temperatures, and since boric oxide melts at 420 °C, this can result in the formation of a viscous glass. Thus BN is useful as a lubricant, as an inert refractory filler and in other applications where its electrically insulating properties offer advantages over carbon.

As with carbon, BN can be prepared in cubic form with a diamond crystal structure and dense forms (with trade names such as Amborite, Borzan, or Sumiboron), exhibiting similar mechanical properties to diamond. Because of the different chemical reactivities of diamond and boron nitride, cubic boron nitride is often used as a cutting tool for materials which show extensive reaction with diamond.

5. TRANSITION METAL NITRIDES

Transition metals form interstitial nitrides that are generally similar in formula and in chemical behavior to the analogous carbides.

5.1 Bonding and Structure

The formula of interstitial nitrides bears no relationship to the valency of the cation, and most transition-metal nitrides have relatively simple formulae of the type MX, M_2X, M_3X, etc., with the metal usually present in the larger proportion. In the absence of balanced valences, there are obviously free electrons as in the pure metal, but the nitrogen changes the outer electronic band structure imparting a covalent influence, which as a consequence increases the resistivity compared with the pure metal. Usually, the arrangement of metal atoms in interstitial nitrides is either face-centered cubic or hexagonal close-packed, but other arrangements are known. A measure of the increased bonding in these compounds compared with the parent metals can be seen from the increased melting point (see Table 2). However, this increase is not constant for all compounds, and on moving across a period of transition metals, the stability of the nitrides decreases from left to right (*i.e.*, with increas-

Table 2. Physicochemical data for selected transition-metal nitrides. T_M = metal melting point; T_{MN} = metal nitride melting point (if known).

Metal	$r_N : r_M$	Nitride	Structure	T_M (°C)	T_{MN} (°C)
Th	0.39	ThN	NaCl	1750	2630
Zr	0.44	ZrN	NaCl	1852	2980
Sc	0.44	ScN	NaCl	1541	2650
U	0.45	UN	NaCl	1132	2800
Nb	0.48	NbN	NaCl	2468	2050
Ti	0.48	TiN	NaCl	1660	2930
W	0.50	W_2N	NaCl	3410	
Mo	0.50	Mo_2N	NaCl	2610	≈895
V	0.52	VN	NaCl	1890	2320
Mn	0.53	Mn_4N	Fe_4N[a]	1244	d.[b]
		Mn_2N	NiAs	1244	≈900
Cr	0.55	Cr_2N	NiAs	1857	1770
		CrN	NaCl	1857	1450
Fe	0.56	Fe_4N	Fe_4N[a]	1535	d.[b]
		Fe_2N	NiAs	1535	d.[b]

[a] Fe_4N has the nitrogen atom in the body-centered position of a face-centered cubic distribution of iron atoms.

[b] d = Decomposes.

ing atomic number). Thus, for example, scandium nitride can be regarded as a stoichiometric compound, in which both metal and nonmetal atoms exhibit their conventional valences, and the bonding is mixed ionic/covalent with very little metallic contribution. Therefore ScN has a very high melting point compared with metallic Sc. The next element, titanium, forms TiN, and the loss of balanced valences results in an increased metallic contribution, but nevertheless there is an appreciable difference in melting points between the metal and the nitride. At the other end of the first-row transition metals, both copper and zinc form extremely unstable nitrides that decompose at low temperatures. The elements in between show an increase in metallic behavior and a decrease in ionic/covalent influence with increasing atomic number. As a result of this the chromium–, manganese–, and iron–nitrogen systems show a larger number of relatively unstable nitrides, which are not easy to prepare. Chapter 5 in Goldschmidt (1967) should be consulted for a detailed discussion of these compounds. The work of Jack (1950) was particularly important in elucidating the compositions and structures of the iron nitrides. The same general trends are evidenced across the second (Y–Cd) and third (La–Hg) rows of transition metals. The rare-earth metals, in contrast, generally form nitrides of formula LnN; this might appear to be consistent with the general metal valence of 3 in this series, but is also true for the elements which show other valences under normal conditions (Ce, 4; Eu, 2). These compounds behave more like ionic nitrides than transition-metal nitrides, but their electrical conductivities still show some evidence of a metallic contribution.

It is rare to find transition-metal nitrides as pure nitrides. Both oxygen and carbon can readily displace nitrogen in the interstitial sites with very little effect on either the chemical or physical properties. Most "nitrides" are in fact "oxynitrides", but demonstrating this, and achieving an accurate value for the oxygen content, is often quite difficult.

5.2 Classification

The simple formulas adopted by transition-metal nitrides form the easiest basis for their classification. As mentioned in Sec. 1.3, if the radius ratio $r_N : r_M$ is less than a value of ≈ 0.59, the nitrogen will behave interstitially in an fcc/hcp framework, and Table 2 shows that this covers most of the transition-metal nitrides of interest. If the radius ratio is greater than 0.41, the nitrogen will occupy a 6-coordinated interstice, and this is the normal case. For 6-coordination, and an overall formula of the type MN, the structure adopted is that of a normal fcc metal, with the nitrogen occupying sites half way between metals. For M_3N structures, a hcp metal arrangement is common, with (in conventional ABC close-packing terminology) nitrogen atoms occupying A-type sites in between three neighbors above and below in B- and C- type layers, respectively. For M_2N structures a similar hcp metal-atom arrangement occurs, with a different proportion of the interstitial sites occupied. What is clear from Table 2 is how common the rock-salt (sodium chloride) structure is for MN-type transition metal nitrides. Detailed studies have been made of defect structures, electronic behavior and properties, especially for those nitrides used as hard coatings on metals (Johansson, 1995).

5.3 Titanium Nitride

This is singled out for special mention here as being the most common of the transition metal nitrides. This is because it offers greater stability compared with the oxide than most other nitrides, and the thin layer of oxide that does form as a result of hydrolysis gives total protection against further degradation. In the competition with other nitrides as a hard surface coating material, TiN is cheaper than most other competitors, can be readily produced as a good-quality, adherent coating by the methods discussed in Sec. 2.2, and remains chemically and thermally stable up to high temperatures (Rickerby and Matthews, 1991). Its characteristic gold color may also add to its attractiveness for certain applications.

6. COMPLEX NITRIDES

A wide range of ternary nitrides exist in which nitrogen is bonded to more than one cation. These compounds have structures that can often be related to the binary nitrides described in Secs. 3 and 4, and have similar chemical and physical properties. This section also deals with mixed nonmetal nitrides, where the nitrogen occurs in conjunction with (commonly) oxygen or fluorine. Both crystalline and vitreous oxynitrides are fields of considerable research interest.

6.1 Mixed Metal Nitrides

These can be divided into either more ionic or more covalent types, in the same way that the binary nitrides were subdivided earlier (Sec. 1.2).

By far the largest group of complex ionic nitrides comprises those based on lithium (Lang, 1977). There is a group of strongly ionic nitrides such as LiMgN, LiZnN, and Li_3AlN_2, which adopt anti-fluorite structures and exhibit very strong hydrolysis reactions. Related nitrides exist with the general formula $Li_{2n-3}M^{n+}N_{n-1}$ (M = Mg, Zn, Al, B, Ba, Si, Ti, V, Ta, Mn, Cr), which also adopt the anti-fluorite structure. Other mixed nitrides with this structure include $Li_{15}Cr_2N_9$ and Li_8SiN_4. A further group of compounds have the anti-La_2O_3 structure and the general formula Li_2MN_2, where M = Si, Zr, Ce. Although these compounds do exhibit ionic conductivity, they are not suitable for battery applications (Metselaar, 1994).

Barium is another element that forms a wide range of ternary nitrides in combination with other metals. Compounds of the type $BaMN_2$ (M = Ti, Zr, Ce) and Ba_2MN_3 (M = Nb, Ta) (Green and Hughbanks, 1993) form with tetragonal and monoclinic structures, respectively, and more complex variants (*e.g.*, $Ba_2Si_5N_8$, $Ba_3Ga_2N_4$, and $Ba_5Ti_2N_6$) have been reported. Interstitial nitrides based on barium are also common. Compounds of BaMN with M = Li, Ni, Co exist; CaNiN has a body-centered arrangement of Ca and Ni with –N–Ni–N– chains arranged in criss-cross fashion. A review of this and related structures is given by Röhr (1996). A series of barium sodium nitrides exist, in which nitrogen is octahedrally coordinated by Ba, with Ba_6N units present either as individual octahedra (as in $Na_{16}Ba_6N$) or linked together in chains (in $NaBa_3N$ and Na_5Ba_3N). More complex units form when calcium is also present. A series of quaternary nitrides of the type $Ba_2Li_3MN_4$ with M = Nb, Ta has been reported. In these compounds, chains of interconnected LiN_4 and NbN_4 tetrahedra are linked by barium atoms (Chen and Eick, 1984).

Ternary nitrides with the overall composition MT_3N are common, where M is a nontransition metal and T a transition metal (Metselaar, 1994). A similar formulation and structure is also exhibited by Ca-containing compounds of the type Ca_3MN, where M = P, As, Sb, Bi, Ge, Sn, and Pb. The latter compounds exhibit interesting electrical properties, with the P and As derivatives being insulating, the Bi and Sb derivatives being semiconducting, and the Pb, Sn and Ge derivatives being metallic.

Molybdenum and tungsten form a variety of ternary nitrides in combination with other metals. $LiMoN_2$ can be formed by ammonia treatment of lithium molybdate at 800 °C, and metal tungsten nitrides of formula MWN_2, with M = Mn, Fe, Co, Ni can be prepared in a similar way. η–Fe_3Mo_3N and Fe_3W_3N nitrides (and related variants with iron replaced by other first-row transition metals) are all cubic with a unit-cell dimension of ≈ 1.13 nm. Compositionally similar compounds of the type Ni_2W_3N are cubic with a cell dimension of 0.67 nm and have a stuffed β-manganese structure.

Iron-rich magnetic alloys with formulations of the type LnT_{12} and Ln_2T_{17} (Ln = rare earth; T = first-row transition metal) were developed in the 1980s because of their increased magnetic moments, and these phases readily accommodate interstitial atoms (boron, carbon, nitrogen). The most explored nitride in this series is $Sm_2Fe_{17}N_x$, where x is a maximum of three but more typically has a value in between 1 and 2. The crystal structure is that of Th_2Zn_{17} (Helmholdt and Buschow 1989), which includes three 6-coordinated interstitial sites per unit cell in which the nitrogen is accommodated. An important effect of the nitrogen in this series of compounds is to increase the Curie temperature (Nagata and Fujii, 1991).

Covalent mixed metal nitrides occur, based on the covalent binary nitride structures discussed in Sec. 4. Thus, a range of compounds exist of the type $M^{II}M^{IV}N_2$, where M^{II} = Be, Mg, Mn, Zn, and M^{IV} = Si, Ge, and also a smaller number of ternary and quaternary nitrides of stoichiometry 1:2:3, exhibited by the compounds $LiSi_2N_3$ and $MAlSiN_3$, where M = Mg, Mn, Ca. Both types of compounds exhibit superlattices of the basic wurtzite type, and the reader is directed to more detailed works (*e.g.*, Thompson, 1989) for a full description of these structures. Boron readily forms ternary nitrides with rare-earth metals. Thus neodymium forms $Nd_3B_2N_4$ and praseodymium $PrBN_2$ (Klesner and Rogl, 1992).

6.2 Mixed Nonmetal Nitrides

Nitrofluorides were mentioned briefly in Sec. 1.1. However, these have little significance apart from demonstrating that the combination of one nitrogen and one fluorine atom can generally behave in the same way as two oxygen atoms. The compound ZrNCl finds some application as a source of nitrogen. A much more important group of compounds in this category comprises the oxynitride glasses and crystalline solids, which have been researched extensively during the last 25 years.

Oxynitride glasses (Thompson, 1992) are derived from oxide glasses of similar cation ratio by the substitution of up to 25 at.% of oxygen by nitrogen. Compositions are often quoted in equivalent % of nitrogen, which refers to the percentage of nitrogen valences expressed relative to the total number of nonmetal valences. Metal atoms present almost invariably include silicon and aluminum, plus also one of the cations Y, Mg, Ca, Li, and Ln. Nitrogen solubility increases with the valence of the third metal cation, with a maximum of $\approx$25 at.% being observed for the low-Z rare earths. These glasses are of interest partly because they occur in the grain boundaries of the commercially important group of sialon ceramics and partly because in comparison with their oxide analogs, they have better refractoriness and higher hardness, viscosity, and refractive index. Amorphous P-O-N oxynitrides form another interesting emergent group of materials, which have been mainly explored in the Zr-P-O-N and Al-P-O-N systems (Fripiat *et al.*, 1997).

Crystalline oxynitrides are best known in combination with higher-valence metals. Silicon oxynitride, Si_2N_2O, is a similar material to silicon nitride, and in the Al-O-N system the spinel compound, $3Al_2O_3 \cdot AlN$ (commonly referred to as ALON), has importance as a more easily sintered, cubic form of alumina, which can be produced in transparent form, suitable for use as infrared windows. Three oxynitrides occur in the ZrO_2-ZrN system, of composition Zr_2N_2O, $Zr_7O_8N_4$, and $Zr_7O_{11}N_2$; the first is an antibixbyite derivative and the last two are defect fluorite structures. Mixed-cation oxynitrides are numerous, and a good summary is given by Thompson (1989).

Considerable interest has been expressed in the four Y-Si-O-N compounds nitrogen melilite ($Y_2Si_3O_3N_4$), nitrogen apatite [$Y_{10}(SiO_4)_6N_2$], nitrogen wollastonite [$(YSiO_2N)$], and nitrogen YAM ($Y_4Si_2O_7N_2$). These compounds often occur in silicon nitride ceramics densified with yttrium oxide, and amply demonstrate that the replacement of oxygen by nitrogen can take place readily with the retention of mineral silicate structures. These compounds all have analogs in equivalent rare-earth systems.

7. APPLICATIONS

Nitrides have found increasing application during the last 25 years either because of good mechanical performance (at both room and elevated temperatures) or, in other cases, because of their good electronic/optoelectronic performance. It is convenient to distinguish between applications that require a thin film or coating of the nitride material and those that require a bulk sample.

7.1 Thin Films and Coatings

The use of transition-metal nitrides as hard metal coatings on metallic substrates was discussed in Sec. 6 as an important application for TiN (see Fig. 2). However, TiN is not unique in this respect, and other transition-metal nitrides have been used. Indeed, the whole field of coatings has grown as a result of these developments, and in consequence, coated tungsten carbide tools have remained competitive despite the onset of new ceramic cutting tools.

In the electronic field, thin layers of silicon nitride are often used to provide insulating

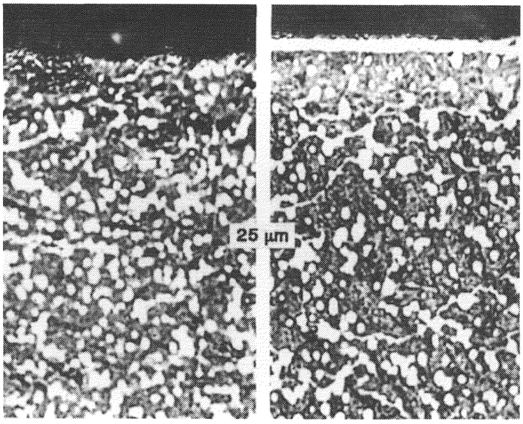

FIG. 2. Optical micrograph of a medium-carbon steel before and after coating with TiN (after Matthews *et al.*, 1993).

layers on silicon wafers (Fig. 3). In a similar way, thin layers of AlN provide a good thermal conductivity substrate material, which is particularly useful for circuits operating at high power levels that need to dissipate large amounts of heat (Fig. 4). The use of Group III metal nitrides (AlN, GaN, InN) as optoelectronic materials for devices because of their suitable band gaps has been discussed in Sec. 4.3. This area represents a significant growth area for these nitrides.

7.2 Bulk Nitrides

Silicon nitride is made into many shapes for components that find application in the aerospace and automotive industries, primarily as wear parts (Fig. 5). These include seals and dies (Fig. 6), the latter for both wire and tube drawing. For these purposes, alumina is added to the starting mix, and the resulting ceramic is strictly speaking a sialon product; this is preferable because the alumina not only helps densification, to produce a dense, fully sintered body, but sintering takes place at a lower temperature, with associated cost benefits. The original intention of using silicon nitride in high-temperature gas turbine engines (as envisaged in the 1960s) has been slow to take off because of the demanding nature of the application, and is unlikely to ever be a large-volume application. The use of refractory nitrides as crucible materials has been more successful; reaction-bonded silicon nitride is excellent for handling molten aluminum. For higher-melting metals, interaction between slags on the metal and the thin films of oxynitride glass in the grain boundaries of the silicon nitride causes degradation and limits lifetimes. Nitride-bonded carbide refractories, consisting of silicon carbide grains bonded by silicon nitride, are examples of high-performance materials finding application as steel-furnace refractories.

Boron nitride finds some application as a high-temperature crucible material for use in inert atmospheres. Aluminum nitride was originally developed for high-temperature structural applications, but its ease of hydrolysis combined with a worse portfolio of properties (hardness, strength, and toughness) compared with silicon nitride and silicon carbide has restricted its development.

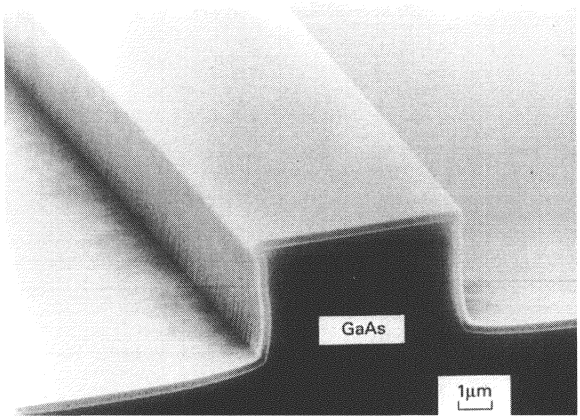

FIG. 3. 250-nm-thick layer of plasma-enhanced CVD silicon nitride deposited on an etched GaAs wafer (after Pearton and Ren, 1994).

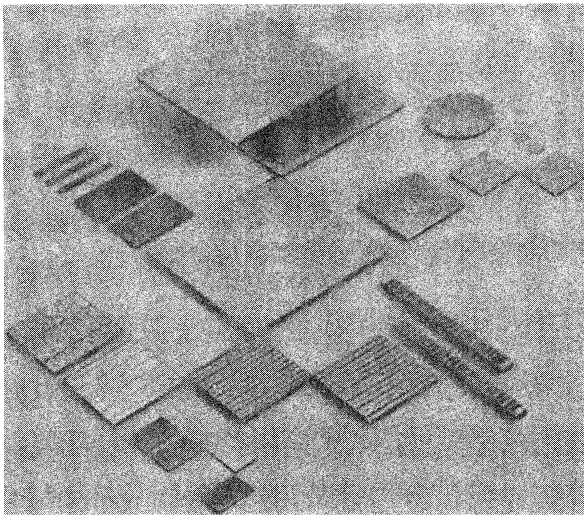

FIG. 4. A selection of partly metallized AlN substrates (after Werdecker and Aldinger, 1984).

8. SUMMARY

Nitrides are an emerging class of materials offering a very wide range of properties, which have been developed during the last 40 years to meet specific niche requirements in the spectrum of materials applications. New nitride materials are still emerging; well-established materials are being refined and new processing routes developed. This article has focused

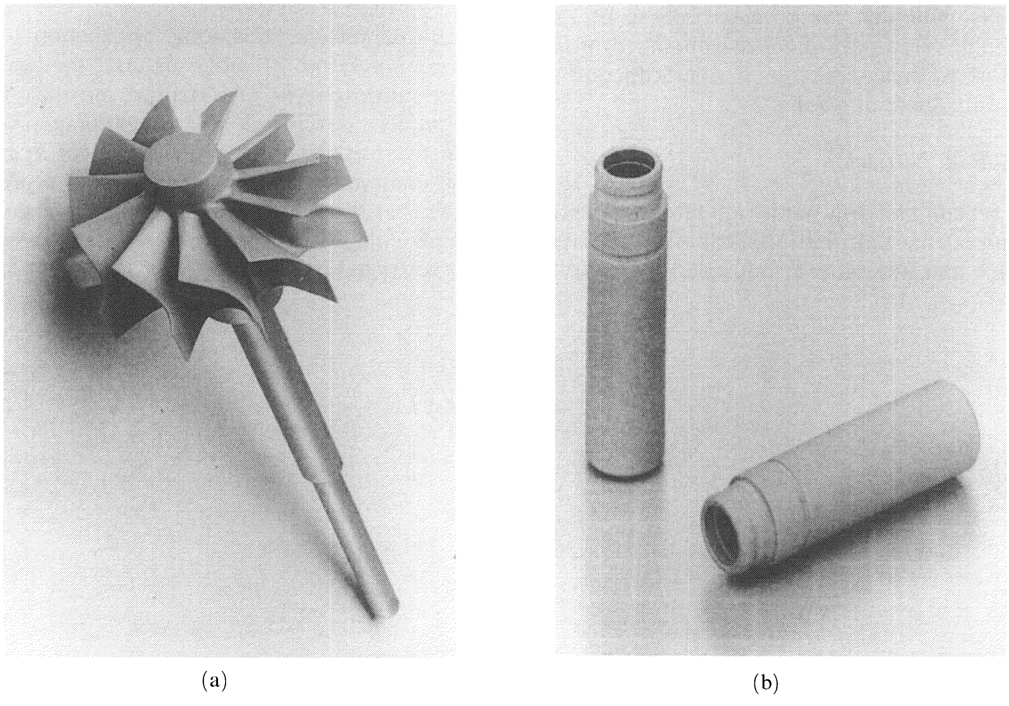

(a)

(b)

FIG. 5. Examples of silicon nitride automotive parts: (a) turbocharger rotor, (b) valve guides (after Heinrich and Kruner, 1996).

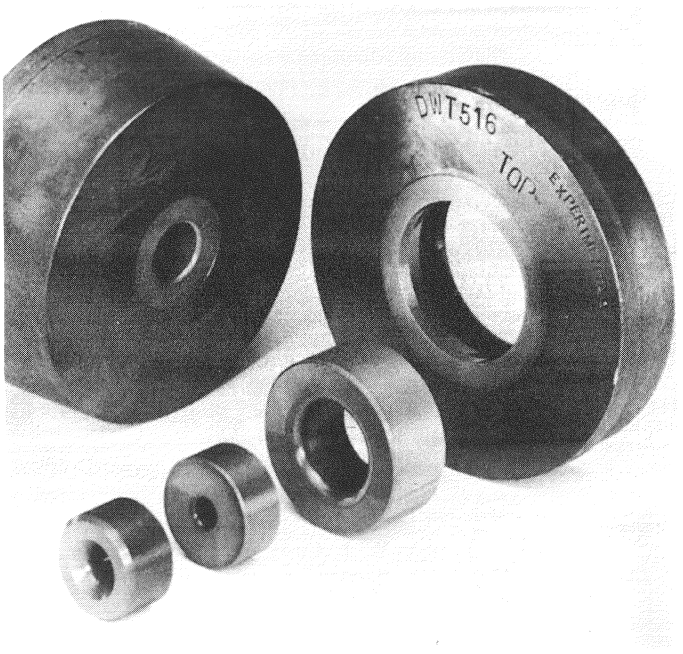

FIG. 6. A selection of sialon die inserts, shrunk-fit into steel surrounds (Lucas Cookson Syalon, 1985).

mainly on binary nitrides because these are currently the most important as far as applications are concerned. However, new ternary and quaternary compounds are emerging, which in time will also find application. The field of thin films, for both electronic and mechanical applications, has rapidly developed in the last 15 years and will undoubtedly expand further. An important emphasis of future activity will be the development of low-cost processes for nitride production; this is necessary if structural nitrides are to become more competitive and is expected also to have a similar impact on markets in the thin films and coatings area.

GLOSSARY

Adamantine: Literally "diamond-like", and applied to those covalent solids formed between low atomic number elements in Groups III, IV and V of the periodic table and displaying high hardness and high melting/decomposition temperature.

Carbothermal Reduction: The reduction of an oxide by heating with carbon, the reaction forming either the metal (if carried out in inert atmosphere) or the nitride (if carried out in nitrogen).

Chemical Vapor Deposition: The deposition of a thin film at high temperature using mixtures of gases that react at that temperature to form the desired product.

Coordination: The number of neighbors surrounding an atom in a crystal structure.

Equivalent %: A term used especially when describing compositions of sialon ceramics, when the proportions of individual elements are expressed relative to the number of valences of that species. Usually metals and nonmetals are dealt with separately, and therefore $YSiO_2N$ would be described as containing 43 equivalent % of yttrium and nitrogen, and 57 equivalent % of silicon and oxygen.

Interstitial: Used to describe compounds containing small atoms, in which the small atoms can geometrically fit into the spaces between the large atoms.

Physical Vapour Deposition: Deposition of thin layers of a substance using nonchemical methods (*e.g.*, evaporation or sputtering).

Sialons: Compounds containing the elements silicon, aluminum, oxygen, and nitrogen (plus others).

YAM: The yttrium aluminate of formula $Y_4Al_2O_9$.

Works Cited

Armour, D. G. (1994), *Nucl. Instr. Methods Phys. Res., Sect. B* **89**, 325–331.

Belkind, A., Gerristead, W., Orban, Z., Dow, D., Felts, J., Laird, R. (1991), *Surf. Coat. Technol.* **49**, 155–160.

Bolse, W. (1994), *Mater. Sci. Eng.* **R12**, 53–121.

Büchner, W., Schliebs, R., Winter, G., Büchel, K. H. (1984), *Industrielle Anorganische Chemie*, Weinheim: Verlag Chemie, Chap. 5.

Chen, A. X. Z., Eick, H. (1984), *J. Solid State Chem.* **113**, 362–366.

Donley, M. S., Zabinski Z. S., (1994), *Thin Films* **22**, 431–453.

Ekström, T., Nygren, M. (1992), *J. Am. Ceram. Soc.* **75**, 259–276.

Evans, R. C. (1964), *Crystal Chemistry*, Cambridge, UK: Cambridge Univ. Press, Chap. 13.

Fripiat, N., Conanec, R., Marchand, R., Laurent, Y., Grange, P. (1997), *J. Eur. Ceram. Soc.* **17**, 2011–2015.

Goldschmidt, H. J., (1967), *Interstitial Alloys*, London: Butterworths, Chap. 5.

Gordon, R. G., (1994), in: S. B. Desu (Ed.) *MRS Publication No.* **335**, Pittsburgh: Materials Research Society, pp. 9–19.

Green, M. T., Hughbanks, T. (1993), *Inorg. Chem.* **32**, 5611–5615.

Hadjipanayis, G. C., Tan, Z. X., Ging, W., Singleton, E. W. (1992), in: *Proceedings of the 7th International Symposium On Magnetic Anisotropy and Coercivity of Rare Earth Transition Metal Alloys*, Ames, IA: Rare Earth Information Center, pp. 403–424.

Heinrich, J. G., Kruner, H. (1996), in: D. S. Yan, X. R. Fu, S. X. Shi (Eds.), *Proceedings of the 5th International Symposium On Ceramic Materials & Components for Engines*. Singapore: World Scientific, pp. 76–98.

Helmholdt, R. B., Buschow, K. H. J. (1989), *J. Less. Comm. Metals* **155**, 15–20.

Hoffmann, D. M. (1994), *Polyhedron* **13**, 1169–1179.

Howson, R. P. (1994), *Pure Appl. Chem.* **66**, 1311–1318.

Jack, K. H. (1950), *Acta Crystallogr.* **13**, 392.

Jehn, H. A. (1994), in: H. A. Jehn, P. Kumar, M. Uz (Eds.), *Proceedings of the Klaus Schulze Symposium On Processing and Applications of High-Purity Refractory Metal Alloys*, Warrendale, PA: Minerals Metals and Materials Society, pp. 153–163.

Johansson, L. I. (1995), *Surf. Sci. Rep.* **21**, 177–250.

Kato, A. (1992), *Trans. Ind. Ceram. Soc.* **51**, 6–16.

Khono, M., Ikenaga, M., Ichimura, H. (1993), *Hyomen Gijutsu* **44**, 708–714.

Klesner, H., Rogl, P. (1992), *J. Am. Ceram. Soc.* **75**, 2825–2827.

Korhonen, A. S., Molarius, J. M., Harju, E., Forsen, O. (1990) in: T. S. Sudarshan, D. G. Bhat, M. Jeandin (Eds.), *Proceedings of the 4th International Conference on Surface Modification Technology*, Warrendale, PA: Minerals, Metals and Materials Society, pp. 311–326.

Kuramoto, N., Tanaguchi, H., Aso, I. (1986), *IEEE Trans. Compon. Hybrid. Manuf. Technol.* **9**, 386–390.

Lackey, W. J., Carte, W. B., Hanigofsky, J. A., Hill, D. N., Barefield, E. (1991), *PCT Int. Appl.*, No. WO91 07, 236.

Lang, J. (1977), in F. L. Riley (Ed.), *Proceedings of the NATO Advanced Study Institute on Nitrogen Ceramics*, Leyden: Noordhoff, pp. 89–108.

Leflamme, C. B., Soucy, G., Jurewicz, J. W., Boulos, M. I. (1992), *J. High Temp. Chem. Processes* **1**, 283–301.

Lieb, K.-P., Bolse, W., Uhrmacher, M. (1994), *Nucl. Instrum. Methods Phys. Res., Sect. B* **89**, 277–289.

Linke, K. H., Schroedter, K. (1975), *Z. Anorg. Allg. Chem.*, **413**, 165–170.

Lucas Cookson Syalon [now International Syalons (Newcastle) Ltd., Cookson, House, Willington Quay, Tyne & Wear NE28 6UQ, U.K.] (1983), publication No. 4242.

Marchand, R., Laurent, Y., Guyader, J., L'Haridon, P., Verdier, P. (1991), *J. Eur. Ceram. Soc.* **8**, 197–213.

Matthews, A., Leyland, A., Fancey, K. S., James, A. S., Robinson, P. A., Holiday, A., Dehbi-Alaoui, A., Stephenson, P. R. (1993), in: P. K. Datta, J. S. Gray (Eds.), *Surface Engineering Volume III: Process Technology and Surface Analysis*, Special Publication No. 128, Cambridge, UK: Royal Society of Chemistry, pp. 99–110.

Metselaar, R. (1994), *Pure Appl. Chem.*, **66**, 1815–1822.

Monaghan, D. P., Teor, D. G., Laing, K. C., Efeoglu, I., Arnell, R. D. (1993), *Surf. Coat. Technol.* **59**, 21–25.

Morgan, P. E. D. (1977), in: F. L. Riley (Ed.), *Proceedings of the NATO Advanced Study Institute on Nitrogen Ceramics*, Leyden: Noordhoff, pp. 23–40.

Motojima, S., Iwanaga, H. (1992), *J. Chem. Vap. Dep.* **1**, 87–129.

Nagata, H., Fujii, H. (1991), *Jp. J. Appl. Phys.* **30**, L367–370.

Obara, T., Tsunoyama, K. (1995), *Materia* **34**, 296–300.

Parthé, E. (1972), *The Crystal Chemistry of Tetrahedral Structures*, Paris: Gordon and Breach.

Pearton, S. J., Ren, F. (1994), *J. Mater. Sci. Electron.* **5**, 1–12.

Rickerby, D. S., Matthews, A. (1991) in: C. A. Brookes (Ed.), *Reviews on Powder Metallurgy and Physical Ceramics*, Vol. 4, London: Freund, pp. 155–295.

Roebuck, P. H. A., Thompson, D. P. (1977), in: F. P. Glasser, P. Potter (Eds.), *High Temperature Chemistry of Inorganic and Ceramic Materials*, Special Publication No. 30, London: The Chemical Society, pp. 222–228.

Röhr, C. (1996), *Angew. Chem.* **35**, 1199–1201.

Sundgren, J. E., Hultman, L. (1995), in: Y. Pauleau (Ed.), *NATO ASI Series E*, Boston: Kluwer, **290**, 453–474.

Takano, I., Isobe, S. (1991), *Shinku* **34**, 745–751.

Thompson, D. P. (1992) in: M. Cable, J. M. Parker (Eds.), *High-Performance Glasses*, Glasgow: Blackie and Sons, pp. 50–62.

Thompson, D. P. (1989) in: T. Y. Tien, D. A. Bonnell (Eds.), *Preparation and Properties of Silicon Nitride Based Ceramics*, Zurich: Trans Tech, pp. 21–42.

Toth, L. E. (1971), *Transition Metal Carbides and Nitrides*, New York: Academic Press.

Van Dijen, J. K., Metselaar, R., Siskens, C. A. M. (1985), *J. Am. Ceram. Soc.* **68** 16–19.

Vetter, J. (1995), *Surf. Coat. Technol.* **76–77**, 719–724.

Werdecker, W., Aldinger, F. (1984), *IEEE Trans. Compon. Hybrids. Manuf. Technol.* **7**, 399–404.

Wild, S., Grieveson, P., Jack, K. H. (1972), in: P. Popper (Ed). *Special Ceramics*, Vol. 5, Stoke-on-Trent, UK: British Ceramic Research Association, 385–395.

Zintl, E., Brauer, G. (1935), *Zeit. Electrochem.* **41**, 102.

Further Reading

Marchand, R. (1998), in: K. A. Gschneidner Jr., L. Eyring (Eds.), *Handbook on the Physics and Chemistry of Rare-earths*, Vol. 25, Amsterdam: Elsevier, Chap. 166.

PERCOLATION PROCESSES

MUHAMMAD SAHIMI, *University of Southern California, Los Angeles, California, U.S.A.*

Introduction 82
1. **What is Percolation?** 82
1.1 Percolation Processes 83
1.1.1 Bond Percolation 84
1.1.2 Site Percolation 84
1.2 Percolation Thresholds 85
1.3 Characteristics of Percolation Systems 85
1.4 Flow and Transport Properties of Percolation Systems 86
1.5 Scaling Laws of Percolation 87
1.6 Fractal Structures in Percolation Systems 89
1.7 Correlated Percolation 90
1.8 Continuum Percolation 92
1.9 Calculation of Flow and Transport Properties of Percolation Systems 94
1.9.1 Monte-Carlo Simulation 94
1.9.2 Effective-Medium Approximation 95
1.9.3 Position-Space Renormalization-Group Method 96
1.9.4 Renormalized Effective-Medium Approximation 98
2. **Application to Composite Materials** 99
2.1 Conductivity of Granular Media 99
2.2 Percolation and Transport Properties of Metal–Insulator Films 101
2.3 ac Conductivity and Dielectric Properties of Composite Media 102
2.4 Hall Conductivity of Composite Media 103
3. **Application to Porous-Media Problems** 104
3.1 Characterization of Porous Media 104
3.1.1 Mercury Porosimetry 105
3.1.2 Pore-Space Connectivity and Sorption Phenomena 106
3.2 Single-Phase Flow in Porous Media 108
3.3 Anomalous Diffusion in Porous Media 109
3.4 Immiscible Displacements in Porous Media 111
3.4.1 Spontaneous Imbibition 111
3.4.2 Quasistatic Imbibition 111
3.5 Percolation Models of Two-Phase Flow in Porous Media 112
3.5.1 Random Percolation Models 112
3.5.2 Invasion Percolation 113
3.5.3 Gradient Percolation 114
3.5.4 Crossover from Fractal to Compact Displacement 114
4. **Application to Hopping Conductivity of Amorphous Semiconductors** 115
4.1 Percolation Models of Hopping Conductivity 117
4.2 Effect of a Variable Density of States 118
4.3 Effect of Coulomb Interactions 119
4.4 Fractal Structures and Superlocalization 119
5. **Application to Granular Superconductors** 120
6. **Application to Elastic and Viscoelastic Properties of Heterogeneous Materials** 121
6.1 Rigidity Percolation 122
6.2 Bond-Bending Models 124
6.3 Elastic Properties of Composite Solids 126
6.4 Elastic and Viscoelastic Properties of Gel Polymers 127
6.5 Structural Properties of Branched Polymers and Gels 129
7. **Application to Glass Formation** 131
8. **Application to Vibrations and Density of States of Materials** 132
9. **Application to Localization: Quantum Percolation** 135
9.1 Numerical Studies of Quantum Percolation 136
9.2 Crossover Between Anderson Localization and Percolation 137
Glossary 137
Works Cited 138

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

INTRODUCTION

We all know that Nature is disordered. We live in a world of composites and mixtures, ranging from man-made polymers and plastics that are used for manufacturing a variety of products, to oil reservoirs and groundwater aquifers, to our own body, which is perhaps the most complex system in Nature. Only theoreticians can think of ordered, pure, and perfectly characterized and geometrically immaculate systems. Even the experimentalist who focusses on the purest of subtances can seldom escape the effects of defects, trace impurities, and finite boundaries. There are few concepts in science more elegant to contemplate than an infinite, perfectly periodic system, and few systems as remote from experimental reality. We are therefore obliged to come to terms with disordered media, variation in shape and constitution often so ill characterized that we must deem it to be random if we are to describe it. In particular, we have to live with the apparent randomness in the morphology of many important systems, which is their topology—the interconnectivity of their individual microscopic elements—and geometry—the shape and size of these individual elements. However, this is only half of the story. We believe, at least above the quantum mechanical level, in the doctrine of determinism, yet important continuum systems exist in which deterministic descriptions are beyond hope. A typical example is Brownian motion where, over certain length scales, we observe an apparent random process, or disordered dynamics.

Nature, then, is disordered both in her structure and in the processes that she supports. The two types of disorder often conspire. An example is fluid flow through a porous medium (*e.g.*, flow of oil in a reservoir) where the interplay between the disordered morphology of the pore space and the dynamics of fluid motion gives rise to a rich variety of phenomena. Despite the rather obvious randomness in Nature, for several decades the problem of how to model a disordered system and the phenomena that occur there was more or less "swept under the rug." Research in this field made progress by taking advantage of periodic structures. However, as one always has to confront the real world, it became apparent that a statistical physics of disordered systems must be devised to provide methods for deriving macroscopic properties of such systems from laws governing the microscopic world, and by taking into account the effect of both the geometry and interconnectivity of their microscopic elements. Although the role of geometry was already appreciated in the early years of this century, the effect of the connectiveness, or the topology of the system, was ignored for many decades, or was treated in an unrealistic manner, simply because it was thought to be too difficult to be taken into account.

The effort for devising such a statistical physics of disordered systems was begun in earnest in the late 1960s and early 1970s. Over the past 30 years, statistical physics of disordered media has been in a rapid stage of progress, and the reason for it is fourfold.

1. Standard theoretical methods for calculating the average properties of such systems have been established.
2. The vast increase in computational power has enabled us to study by computer simulations very complex systems.
3. The advent of novel experimental techniques has allowed more precise data to be accumulated.
4. Perhaps, however, the most important reason for this rapid progress is that the role of interconnectivity of the microscopic elements of disordered media and its effect on their macroscopic properties have been appreciated.

Point (4) has been possible through the development and application of percolation theory, the subject of this article.

1. WHAT IS PERCOLATION?

Consider a square network in which each bond represents a resistor. We select at random a fraction p of the bonds and assign a unit conductance to them, while the rest of the bonds are assigned zero conductance. Thus, the network is a composite of conducting and nonconducting bonds. We now apply a unit voltage gradient between two opposite edges of the network and ask, What is the minimum value of p, the fraction of the conducting bonds, in order for the electrical current to flow from one face of the network to the opposite face? From a technological viewpoint, this is an important question because, for example, its answer tells us what (volume) fraction of a disordered material made of conducting and nonconducting constituents, such as carbon black composites that are routinely used in many applications, has to be conducting in order for the composite as a whole to be conducting. Clearly, if too many bonds are nonconducting, no macroscopic current will flow in the network, whereas for sufficiently large p one has conduction between the two opposite faces.

Consider another example. Imagine that the bonds of the square network are pores of a porous medium, *e.g.*, those of an oil reservoir. In reality, no porous medium looks as ordered as the square network, but as an idealization this network serves a useful purpose. Now suppose that the pores are filled with oil, and that there are two wells in the system at two opposite corners of the network. We would like to push oil out of this porous medium by injecting water into the system at one corner (the injection well) and producing oil at the opposite corner (the production well). Oil and water do not mix with each other, and therefore we assume that each pore is filled with either oil or water. As the water is pushed into the reservoir, it tries to find the smallest pores that it can reach and expel the oil from it (why this is so is explained in Sec. 3). In reality, the process is more complex than this, but we ignore all the complications. The key question is: What fraction of the pores are filled with water when it reaches the production well for the first time (this is called the *breakthrough point*)? In other words, we would like to know how much oil is produced up to the breakthrough point. This is obviously an important question, scientifically, technologically and economically.

Questions such as these are answered by percolation theory, which tells us when a system is macroscopically connected so that a given phenomenon can occur in it. For example, it tells us when one can have a flow of electrical current from one side of a composite solid to the opposite side, or how much oil one can hope to extract from an oil reservoir. The point at which a sample-spanning cluster of the conducting elements, or of the pores of an oil reservoir that are filled with the injected water, is formed for the first time is called the *percolation threshold* of the system, and the behavior of the system close to this point is of prime interest and importance.

This article attempts to outline the most important concepts of percolation theory and discuss their applications. Over the past 25 years percolation theory has been applied to modelling a wide variety of phenomena in disordered media. It is impossible to discuss all such applications in one article. Thus, we discuss only those percolation models for which not only is there broad practical and scientific importance, but also there is quantitative (numerical) agreement between their predictions and the relevant experimental data. Stauffer and Aharony (1992), Bunde and Havlin (1996), and Hughes (1996) provide the necessary theoretical foundations of percolation theory. Hence, we only summarize the main concepts that will be used in the rest of this article.

1.1 Percolation Processes

Percolation processes were first developed by Flory (1941) and Stockmayer (1943) to describe polymerization in which small branching molecules react and form large macromolecules (see Sec. 6). However, they did not use the terminology of percolation theory, and also they developed their theory for a special type of lattice, called the Bethe lattice, an endlessly branching structure without any closed loops. Percolation processes, as we know them today, were introduced by Broadbent and Hammersley (1957) who were interested in the spread of a hypothetical fluid through a random medium. Generally speaking, such a process may involve some random elements, but the underlying randomness may be of two very different types. In one type, the randomness is dictated by the *fluid*, in that it is the fluid that decides which path to take in the system; this is the classical diffusion process. In the other type, the randomness in the path of the particle is imposed by the *medium*: this was the situation that was newly considered by Broadbent and Hammersley (1957). They called this phenomenon a *percolation process*; the spread of the fluid through the random

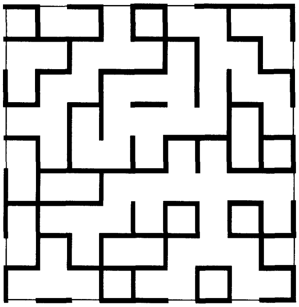

FIG. 1. Configuration of a square lattice in bond percolation. Thick lines are the occupied bonds, whose fraction is 0.57.

medium resembles the flow of coffee in a percolator. Since percolation theory deals with various properties of a disordered medium, we need to have a model of the medium in order to discuss its concepts. We represent the medium by a network or lattice. What the bonds and sites of the lattice represent depends on the particular application of percolation theory. In their original paper, Broadbent and Hammersley focussed on two problems, site and bond percolation, which we now discuss.

1.1.1 Bond Percolation In this problem the bonds of the network are either occupied or intact, randomly and independently with probability p, or they are vacant or removed with probability $1 - p$. For a large network, this assignment is equivalent to removing a fraction $1 - p$ of all the bonds at random. Figure 1 shows configuration of a bond percolation process on the square network. The occupied bonds may represent the conducting elements of a composite material, or the high-permeability regions of a pore space through which most of the fluid flow and transport take place, or the chemical bonds between the monomers during polymerization, and so on. The vacant bonds represent, *e.g.*, the low-permeability, or impermeable, regions of the pore space that contribute very little to flow or transport. Of course, in many heterogeneous media the disorder is not necessarily distributed randomly, but for now we ignore such complications.

Two sites are called *connected* if there exists at least one path between them consisting solely of occupied bonds. A set of connected sites bounded by vacant bonds is called a *cluster*. If the network is very large, and if p is of sufficiently small, the size of any connected cluster is likely to be small. However, if p is close to 1, the network should be entirely connected, apart from occasional small holes. At some well-defined value of p, there is a transition in the topological structure of the network; this value is called the *bond percolation threshold* p_{cb}. For $p < p_{\mathrm{cb}}$ there is no sample-spanning cluster of the occupied bonds, whereas for $p > p_{\mathrm{cb}}$ a cluster of the occupied bonds spans the system. In the context of the problem of composite solids, for example, for $p < p_{\mathrm{cb}}$ the system is not conducting, whereas for $p > p_{\mathrm{cb}}$ the overall conductivity of the composite is nonzero.

1.1.2 Site Percolation In this problem, sites of the network are occupied with probability p and vacant with probability $1 - p$.Two nearest-neighbor sites are connected if they are both occupied, and connected clusters on the network are again defined in the obvious way. For example, in the context of the polymerization problem, the occupied sites are the monomers and the connected sites are those that have reacted with each other and have formed a chemical (covalent) bond between them. Similar to bond percolation, there is also a site percolation threshold p_{cs} above which a sample-spanning cluster of occupied sites spans the network. Figure 2 shows site percolation clusters on a square network. Depending on a specific application of percolation theory to a problem of interest, many variants

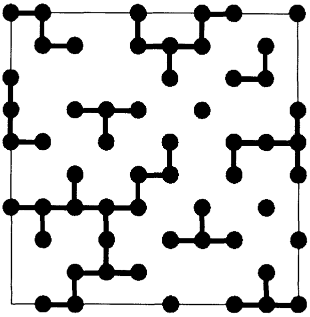

FIG. 2. Configuration of a square lattice in site percolation. Circles denote the occupied sites, whose fraction is 0.57.

of the classical percolation processes have been developed. The interested reader is referred to Stauffer *et al.* (1982) for a list of many variants of the classical percolation problem.

From a practical view point, generating a percolation network by randomly removing sites or bonds is not suitable for most applications, because in addition to the sample-spanning cluster, this method also generates isolated finite clusters. In most applications one uses only the sample-spanning cluster, or at least the phenomenon of interest starts with a single cluster, and therefore one must first delete all the isolated clusters from the system. In such cases, one can use an efficient method due to Leath (1976) and Alexandrowicz (1980) that generates only the sample-spanning (or the largest) cluster. In this method one starts with a single occupied site at the center of the network. One then identifies the nearest-neighbor sites of the occupied site and considers them occupied and adds them to the cluster if random numbers r, uniformly distributed in $(0, 1)$ and attributed to the sites, are less than the fixed value p. The perimeters (the nearest-neighbor empty sites) of these sites are found and the process of occupying the sites continues in the same way. If a selected perimeter site is not made occupied, then it remains unoccupied forever. The generalization of this method for generating clusters of occupied bonds is obvious. Figure 3 presents a sample-spanning cluster on the triangular lattice that was generated by this method.

1.2 Percolation Thresholds

Exact values of p_{cb} and p_{cs} are known only for certain lattices related to the Bethe lattice and for a few two-dimensional (2D) networks. For the Bethe lattice one has

$$p_{\text{cb}} = p_{\text{cs}} = 1/(Z-1), \tag{1}$$

where Z is the coordination number of the lattice, *i.e.*, the number of bonds connected to any one site. The current estimates of p_{cb} and p_{cs} for three common 2D lattices are compiled in Table 1, while the most accurate estimates of p_{cb} and p_{cs} for four common 3D lattices are compiled in Table 2. A very accurate approximation to p_{cb} is given by $p_{\text{cb}} \simeq d/Z(d-1)$.

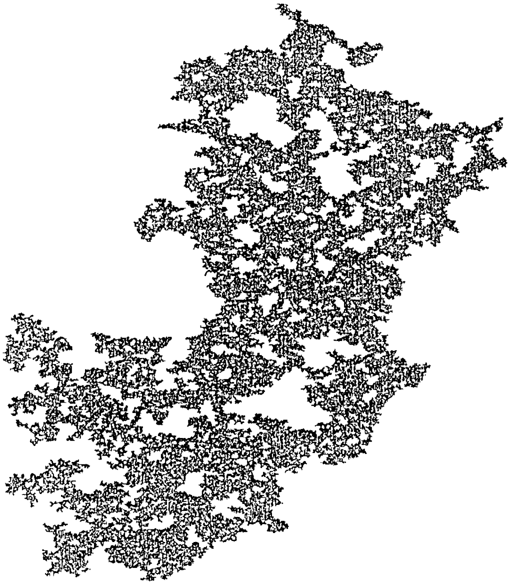

FIG. 3. The sample-spanning cluster in site percolation on a very large triangular lattice, generated by the Leath–Alexandrowicz method.

Thus, the product $B_{\text{c}} = Zp_{\text{cb}} \simeq d/(d-1)$, the average number of conducting bonds at p_{cb}, is an essentially invariant quantity that depends only on the dimensionality d of the system. This is demonstrated in Tables 1 and 2 where we also present B_{c}. The significance of B_{c} is discussed in Sec. 1.8.

1.3 Characteristics of Percolation Systems

In addition to p_{cb} and p_{cs}, the structure of percolation networks is characterized by several other quantities, the most important of which are as follows (for brevity, we use p_{c} to denote p_{cb} or p_{cs}).

The *percolation probability* $P(p)$ is the probability that, when the fraction of occupied bonds is p, a given site belongs to the sample-spanning cluster of occupied bonds.

The *accessible fraction* $P_{\text{A}}(p)$ is the fraction of the occupied bonds (or sites) belonging to the sample-spanning cluster.

The *backbone fraction* $P_{\text{B}}(p)$ is an important quantity in flow and transport in percolation systems. In such phenomena, the sample-spanning cluster, through which flow and

Table 1. Values of bond percolation threshold p_{cb}, site percolation threshold p_{cs}, and $B_c = Zp_{cb}$ for three common 2D networks.

Network	Z	p_{cb}	B_c	p_{cs}
Honeycomb	3	$1 - 2\sin(\pi/18) \simeq 0.6527$[a]	1.96	0.6962
Square	4	1/2[a]	2	0.5927
Triangular	6	$2\sin(\pi/18) \simeq 0.3473$[a]	2.08	1/2[a]

[a] Exact result.

transport occur, can be divided into two parts: the dead-end part that carries no flow or current, and the backbone, which is the multiply connected part of the cluster through which a fluid flows. The fraction of the occupied or conducting bonds in the backbone is $P_B(p)$, and clearly, $P_A(p) \geq P_B(p)$. Near the percolation threshold, the bonds in the backbone are also divided into two groups (Stanley, 1977): those in the *blobs* that are multiply connected and make the transport paths very tortuous, and the *red* bonds, which are those that, if cut, would split the backbone into two parts, and which thus control transport and flow through the backbone. Figure 4 shows a sample-spanning percolation cluster and its backbone on the square network.

The *correlation length* $\xi_p(p)$ is the typical radius of percolation clusters for $p < p_c$, and the length scale over which a percolation network is macroscopically homogeneous for $p > p_c$. For a very large network, ξ_p is the dominant length scale in the system. If the network is finite with a linear size L, then for $L > \xi_p$ the system is macroscopically homogeneous, while for $L < \xi_p$ it is inhomogeneous.

The average number of clusters of size s (per lattice site) $n_s(p)$ is an important quantity in some of the problems of interest here because it corresponds, for example, to the number of islands or blobs of a fluid that are formed during the displacement of that fluid by another fluid in a porous medium. The important task of calculating n_s is done efficiently by an algorithm due to Hoshen and Kopelman (1976). This algorithm is described in detail by Stauffer and Aharony (1992), where a computer program is also given for counting the clusters.

Table 2. Numerical estimates of bond percolation threshold p_{cb}, site percolation threshold p_{cs}, and $B_c = Zp_{cb}$ for four common 3D networks.

Network	Z	p_{cb}	B_c	p_{cs}
Diamond	4	0.3886	1.55	0.4299
Simple-cubic	6	0.2488	1.49	0.3116
BCC	8	0.1795	1.44	0.2464
FCC	12	0.119	1.43	0.199

1.4 Flow and Transport Properties of Percolation Systems

Because a major application of percolation theory has been modeling of flow and transport in disordered media, and in particular porous media and composite solids, we must also consider the effective flow and transport properties of percolation systems, namely, their conductivity, diffusivity and permeability. We first consider the conductivity of percolation networks. Consider a two-component network in which each bond has a conduc-

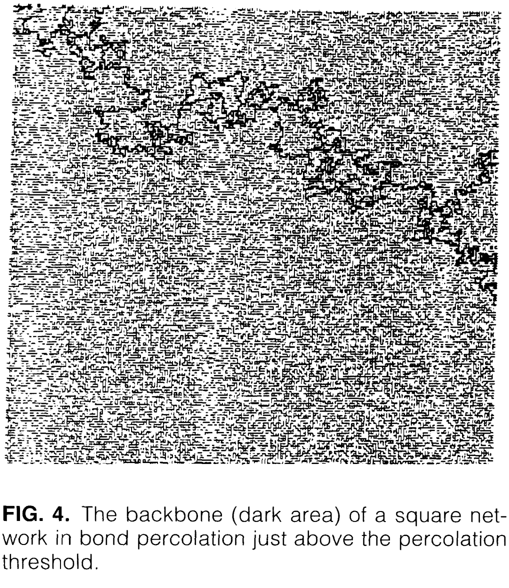

FIG. 4. The backbone (dark area) of a square network in bond percolation just above the percolation threshold.

tance g_1 with probability p or g_2 with probability $q = 1 - p$. It is straightforward to show that the effective conductivity G of the network is a homogeneous function and takes on the following form

$$G(p, g_1, g_2) = g_1 F(p, h), \tag{2}$$

where $h = g_2/g_1$, and F in a homogeneous function. By definition G is invariant under the interchange of g_1 and g_2, and therefore we must have

$$G(p, g_1, g_2) = G(q, g_2, g_1) \quad \text{and}$$

$$F(p, h) = hF(q, 1/h). \tag{3}$$

The limit in which $g_2 = 0$ and g_1 is finite corresponds to conductor-insulator mixtures. In this case, as $p \to p_c$, more and more bonds are insulating, and therefore the conduction paths become very tortuous and G decreases; at p_c one has $G(p_c) = 0$. The conductance g_1 can be distributed according to a given distribution, which is the case in most real disordered systems, such as porous media and composite solids.

The limit in which $g_1 = \infty$ and g_2 is finite represents superconductor–conductor mixtures. All quantum-mechanical aspects of real superconductors are ignored in this definition, and we are concerned only with the effect of the connectivity of the percolation system on this conductivity. (Application of percolation theory to granular superconductors is discussed in Sec. 5.) It is clear that the macroscopic conductivity G of this system is dominated by the superconducting bonds. If $p < p_c$, then a sample-spanning cluster of the superconducting bonds does not exist, and G is finite. As $p \to p_c^-$, G increases. At $p = p_c$ a sample-spanning cluster of the superconducting bonds is formed for the first time, and G diverges. The significance of superconducting percolation networks will be discussed in Sec. 2.3. Note that both cases correspond to $h = 0$, and therefore the point $h = 0$ at $p = p_c$ is particularly important.

In a similar way, the effective diffusivity D and hydrodynamic permeability K of the system are defined; namely, since fluid flow or diffusion occurs in the occupied bonds, the effective permeability and diffusivity are defined in a manner similar to the conductivity of conductor–insulator mixtures. Figure 5 shows the typical behavior of some of these properties for a simple-cubic network in site percolation.

1.5 Scaling Laws of Percolation

One of the most important characteristics of percolation processes is their universal properties. The behavior of many percolation

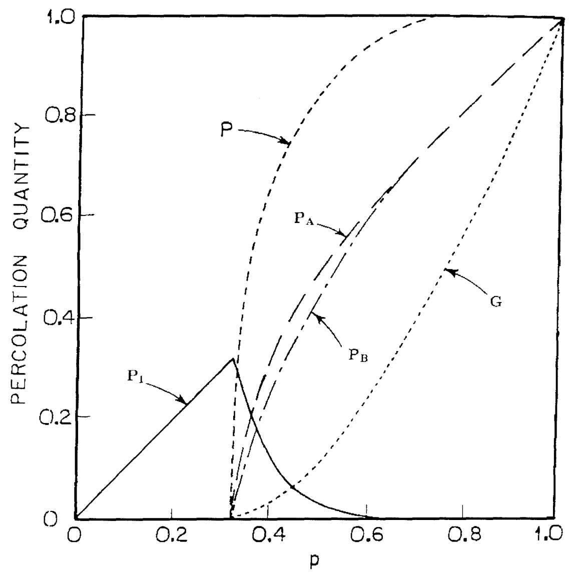

FIG. 5. Typical variations of the percolation quantities in site percolation on the simple cubic lattice. $P_I = P_A - P_B$ is the fraction of the isolated occupied sites.

quantities near p_c is insensitive to the network structure and to whether the percolation process is a site or a bond problem. The quantitative statement of this universality is that many percolation properties obey power laws near p_c, and that the critical exponents that characterize such power laws are universal and depend only on the Euclidean dimensionality d of the system. In general, near p_c we have the following power laws:

$$P(p) \sim (p - p_c)^{\beta}, \tag{4}$$

$$P_A(p) \sim (p - p_c)^{\beta}, \tag{5}$$

$$P_B(p) \sim (p - p_c)^{\beta_B}, \tag{6}$$

$$\xi_p(p) \sim |p - p_c|^{-\nu}, \tag{7}$$

$$G(p) \sim (p - p_c)^{\mu}, \tag{8}$$

$$G(p) \sim (p_c - p)^{-s}, \tag{9}$$

$$K(p) \sim (p - p_c)^{\chi}, \tag{10}$$

where Eqs. (8) and (9) refer to the conductor–insulator and conductor–superconductor cases, respectively. For length scales $L < \xi_p$, the resistance R between two end points of a box of linear size L scales with L as $R \sim L^{\tilde{\zeta}}$. It is not difficult to show that

$$\mu = (d - 2)\nu + \zeta, \tag{11}$$

where, $\zeta = \tilde{\zeta}\nu$. In 2D it has been shown that $\mu = s$.

Equations (8) and (9) can be unified by using the two-component resistor network described above. In the critical region where both $|p - p_c|$ and $h = g_2/g_1$ are small, G obeys the following scaling law (Efros and Shklovskii, 1976; Straley, 1976):

$$G \sim g_1 |p - p_c|^{\mu} \Phi_{\pm}(h|p - p_c|^{-\mu-s}). \tag{12}$$

where Φ_+ and Φ_- are two homogeneous functions corresponding, respectively, to the regions above and below p_c, and are, like μ and s, universal. For any fixed and nonzero h, G has a smooth dependence on $p - p_c$. This becomes clearer if we rewrite Eq. (12) as

$$G \sim g_1 h^{\mu/(\mu+s)} \Psi(|p - p_c| h^{-1/(\mu+s)}), \tag{13}$$

where

$$\Psi(x) = x^{\mu}\Phi_+(x^{-\mu-s}) = (-x)^{\mu}\Phi_-((-x)^{-\mu-s}).$$

The power law that characterizes the behavior of the effective diffusivity $D(p)$ near p_c is derived from that of $G(p)$. According to Einstein's relation, $G \sim \rho D$, where ρ is the density of the carriers. Although a diffusing species can move on all the clusters, above p_c only diffusion on the sample-spanning cluster contributes significantly to D, so that $\rho \sim P_A(p)$. Hence, $G(p) \sim P_A(p)D(p)$, and therefore

$$D(p) \sim (p - p_c)^{\mu-\beta}. \tag{14}$$

For large clusters near p_c, $n_s(p)$ obeys the following scaling law:

$$n_s \sim s^{-\tau} f((p - p_c)s^{\sigma}), \tag{15}$$

where τ and σ are two more universal critical exponents, and $f(x)$ is a scaling function such that $f(0)$ is not singular.

These topological exponents are not all independent. For example, one has $\tau = 2 + \beta\sigma$ and $\nu d = \beta + 1/\sigma$, and in fact knowledge of ν and another exponent is sufficient for determining most of the percolation exponents. However, no exact relationship is known between the flow and transport and the topological exponents. This is perhaps because the transport exponents describe dynamical properties of percolation systems, whereas the topological exponents describe the static properties, and in general there is no reason to expect a direct relationship between the two. The implied prefactors in the above power laws depend on the type of lattice and are not universal. If two phenomena are described by two different sets of critical exponents, the physical laws governing the two must be fundamentally different. Thus critical exponents can help one to distinguish between different classes of problems and the physical laws that govern them. Moreover, since the numerical values of the percolation properties are not universal and vary from one system to another system, but the scaling laws that they obey near p_c are universal and do not depend on the details of the system, the critical exponents are used for establishing the relevance of percolation to a given phenomenon in disordered media. In

Table 3. Values of the critical exponents and fractal dimensions of percolation systems. Rational numbers represent exact results.

Exponent	$d = 2$	$d = 3$
β	5/36	0.41
β_B	0.48	0.99
τ	187/91	2.18
σ	36/91	0.45
ν	4/3	0.88
D_F	91/48	2.53
D_B	1.64	1.87
D_{min}	1.13	1.37
D_R	3/4	1.13
μ	1.3	2.0
s	1.3	0.735

Table 3 values of the critical exponents in 2D and 3D are compiled.

1.6 Fractal Structures in Percolation Systems

As mentioned above, the correlation length ξ_p has the physical significance that for length scales $L > \xi_p$ the percolation system is macroscopically homogeneous. However, for $L < \xi_p$ the system is not homogeneous, and the macroscopic properties of the system depend on L. In this regime, the sample-spanning cluster is statistically self-similar at all length scales less than ξ_p. This means that if at such length scales we magnify or shrink the sample-spanning cluster, it looks the same. Under this condition the mass M of the sample-spanning cluster (its total number of occupied bonds or sites) scales with L as

$$M \propto L^{D_F}, \quad L < \xi_p, \tag{16}$$

where D_F is called the fractal dimension of the cluster. For $L > \xi_p$, $M \propto L^d$ and thus $D_F = d$, where d is the Euclidean dimension of the system. However, for fractal objects, $D_F < d$, implying that they do not fill the space in which they are embedded. The crossover between the two regimes takes place at $L \simeq \xi_p$. Thus, similar to Eqs. (12) and (13), we can combine the two regimes and write down a unified scaling equation for $M(L, \xi_p)$:

$$M(L, \xi_p) = L^{D_F} h(L/\xi_p), \tag{17}$$

where $h(x)$ is a scaling function. For $L \gg \xi_p$ ($x \gg 1$) we expect to have $M \sim L^d$, and hence $h(x) \sim x^{D_F - d}$, whereas for $x \ll 1$ we must have $h(x) \simeq$ constant. In the case of the sample-spanning percolation cluster D_F can be related to the other percolation exponents. The mass M of the sample-spanning cluster is proportional to $P_A(p)\xi_p^d$, since only a fraction $P_A(p)$ of all the occupied bonds or sites are in the cluster. Therefore, we must have $M \propto \xi_p^{d-\beta/\nu}$. For $L < \xi_p$ we must replace ξ_p with L, since now L is the dominant length scale of the system. Thus, $M \propto L^{d-\beta/\nu}$, which when compared with Eq. (16), yields

$$D_F = d - \beta/\nu, \tag{18}$$

so that $D_F(d = 2) = 91/48 \simeq 1.9$ and $D_F(d = 3) \simeq 2.53$. Similarly, for $L < \xi_p$, the backbone is also a fractal object, and its fractal dimension D_B is given by

$$D_B = d - \beta_B/\nu. \tag{19}$$

Herrmann and Stanley (1984) argued that the backbone at p_c is a *volatile* fractal, *i.e.*, its structure is not stable under a length scale change. Note that at $p = p_c$ the correlation length is divergent, so that at this point the sample-spanning cluster and its backbone are fractal objects at any length scale.

Near and at p_c, the number M_r of the red bonds in a box of linear size L, with $L < \xi_p$, scales with L as $M_r \propto L^{D_R}$, where D_R is the fractal dimension of the set of the red bonds. It has been shown (Coniglio, 1981) that

$$D_R = 1/\nu. \tag{20}$$

Another important concept is the minimum or chemical path between two points of a percolation cluster, which is the shortest path between the two points. For $L < \xi_p$ the length L_{min} of the path scales with L as

$$L_{min} \sim L^{D_{min}}, \tag{21}$$

so that D_{min} is the fractal dimension of the minimal path.

If $L < \xi_p$, *i.e.*, if the percolation system is in the fractal regime, all the percolation properties become scale dependent. Since $|p - p_c| \sim$

$\xi_p^{-1/\nu}$, one can rewrite Eqs. (4)–(10) and (14) in terms of ξ_p. For example, Eq. (4) is rewritten as $P(p) \sim \xi_p^{-\beta/\nu}$. For $L < \xi_p$ one replaces ξ_p by L, since now L is the dominant length scale in the system (for $L > \xi_p$ the correlation length is the dominant length scale), and therefore Eqs. (4)–(10) and (14) become

$$P(L) \sim L^{-\beta/\nu}, \tag{22}$$

$$P_{\mathrm{A}}(L) \sim L^{-\beta/\nu}, \tag{23}$$

$$P_{\mathrm{B}}(L) \sim L^{-\beta/\nu}, \tag{24}$$

$$G(L) \sim L^{-\mu/\nu}, \tag{25}$$

$$G(L) \sim L^{s/\nu}, \tag{26}$$

$$K(L) \sim L^{-\chi/\nu}, \tag{27}$$

$$D(L) \sim L^{-\theta}, \tag{28}$$

where $\theta = (\mu - \beta)/\nu$, and Eqs. (25) and (26) refer to the conductor–insulator and conductor–superconductor mixtures, respectively. Thus, *scale-dependent properties are a signature of a fractal structure.* Similar to the mass of the sample-spanning cluster for which we could write scaling Eq. (17), we can also write down scaling equations for length-scale dependence of all other percolation properties. For example, Eqs. (8) and (25) can be combined to yield

$$G(L, \xi_p) = \xi_p^{-\mu/\nu} w(L/\xi_p), \tag{29}$$

where $w(x)$ is a scaling function. Once it is established that a system is a fractal, many classical laws of physics have to be significantly modified. For example, Fick's law of diffusion with a constant diffusivity is no longer appropriate for describing diffusion processes in fractal systems; this is discussed in Sec. 3.3.

For random percolation, or for correlated percolation in which the correlations are short-ranged, all the topological critical exponents and the fractal dimensions are completely universal. The exponents μ, s, and χ are also universal if two conditions are satisfied. One is that there should not be any long-range correlations in the system. If the conductances of the bonds are distributed according to a distribution $\psi(g)$, then the second condition is (Kogut and Straley, 1979; Straley, 1982; Sahimi *et al.*, 1983a)

$$\psi_{-1} = \int_0^\infty \mathrm{d}x\, \psi(x)/x < \infty. \tag{30}$$

Moreover, for any percolation system that satisfies these two conditions, $\chi = \mu$. Percolation in some correlated or continuum systems violates one or both of these conditions. We now discuss percolation in such systems, which are also of practical importance.

1.7 Correlated Percolation

In the classical site- and bond-percolation problems discussed above, there are no correlations between various segments of the system. However, the nature of disorder in many important physical systems and phenomena is not completely random, and there usually exist some correlations. For example, in packing of solid particles, whose mechanical properties are described by the elastic percolation networks discussed in Secs. 6.1–6.3, there are some short-range correlations. In some other system, such as field-scale porous media, there are long-range correlations.

Thus, we divide all the correlated percolation models into two groups. In one group are those that contain short-range correlations. An example of this type of model is the correlated bond-percolation model due to Kirkpatrick (1973). In his model, which was intended for hopping conduction in disordered solids (see Sec. 4), to each site of a lattice is assigned a random number s_i, uniformly distributed between -1 and $+1$. Random numbers b_{ij} for the bonds are calculated by

$$b_{ij} = (|s_i| + |s_j| + |s_i - s_j|)/2, \tag{31}$$

and all the bonds with $b_{ij} > \varepsilon$ are removed, where ε is some selected value. Thus a bond remains conducting only if b_{ij} is sufficiently small, implying that $|s_i|$ and $|s_j|$ must both be small. This also means that if a bond is conducting, its neighbors are also likely to be conducting, and hence such bonds are clustered together. Another example of this type of model is the site-bond–correlated percolation problem developed by Coniglio *et al.* (1979). Their model was intended for gelation processes in which the effect of the solvent, which had been ignored in the original theories of Flory and Stockmayer, is taken into account. In their model each site of a lattice is occupied by either a monomer or a solvent molecule, and the probability that a bond is present depends on the relative interaction energy between its two end sites (monomer–monomer, monomer–solvent, solvent–solvent) and thus on the temperature of the system. A third example is a correlated-site percolation model developed by Stanley (1979; see also Stanley *et al.*, 1981) which was intended for explaining some unusual features of supercooled water. In this model, the bonds of a lattice are occupied randomly, as in random percolation. One then observes that each site can be considered to be a member of one of $Z+1$ different species, depending on whether $0, 1, 2, \ldots, Z$ bonds emanate from that site. Thus, a site is occupied only if, for example, all the Z bonds emanating from it are also occupied, and hence the presence or absence of a site depends on the bonds that are connected to it.

In the second class of correlated percolation models are those that contain long-range correlations, by which we mean those in which the extent of the correlations is comparable to the linear size of the system. A prime motivation for developing such models has been the observation that the permeability, porosity, and other properties of field-scale porous media, such as oil reservoirs and aquifers, are correlated with each other, with the extent of the correlations being comparable to the linear size of system. The first of such models was probably developed by Weinrib (1984). An example of such percolation models is that of Prakash *et al.* (1992), who considered a percolation model in which the correlation function $C(r)$ in a d-dimensional system, defined by

$$C(r) = \langle v(r')v(r+r')\rangle, \tag{32}$$

where $v(r)$ is a random variable obeying a distribution with long-range correlations and $\langle \cdot \rangle$ denotes an average over all values of r', given by

$$C(r) \sim r^{-(d-\lambda)}, \tag{33}$$

where $-2 \leq \lambda \leq 2$ is a parameter of the model, such that $0 \leq \lambda \leq 2$ represents positive correlations, while $-2 \leq \lambda \leq 0$ corresponds to negative correlations. Another model was developed by Sahimi and Mukhopadhyay (1996), in which the correlations between the conductances of the bonds of a networks were of the same type that one observes in the hydraulic conductances of field-scale porous media. They used their percolation model to derive several scaling laws for transport coefficients in flow through geological formations, with the results being in agreement with field observations.

Positive correlations usually reduce the percolation threshold of the system. For example, in Kirkpatrick's model, $p_{\mathrm{cb}} \simeq 0.1$ for the simple-cubic lattice, sharply lower than $p_{\mathrm{cb}} \simeq 0.2488$ for random bond percolation. However, Bug *et al.* (1985) showed by computer simulations that positive correlations may increase or decrease p_{c}, depending on the distance between two occupied sites that are considered, the dimensionality, and the temperature of the system. The effect of correlations in experimental percolation systems is discussed below.

As discussed above, the scaling properties of percolation systems with short-range correlations are the same as those of random percolation, if the length scale of interest is larger than the correlation length. For example, if the correlation function decays as r^{-d} or faster, where r is the distance between two points and d is the dimensionality of the system, then the critical properties of the systems are identical with those of random percolation. This is not totally unexpected because random percolation itself already has nonzero correlation function, and hence the introduction of any type of short-range correlation cannot change its scaling properties. However, with long-range correlations the scaling properties of such phenomena are nonuniversal. In the case of flow through porous media, this nonuniversality is consistent with field observations and data that indicate that there is not a unique scaling law (*i.e.*, unique critical exponent) for a given transport property, *i.e.*, the exponents change from field to field; see Sahimi and Mukhopadhyay (1996) for more details.

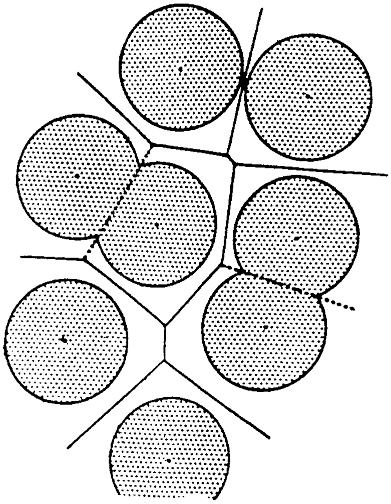

FIG. 6. Two-dimensional Swiss-cheese model.

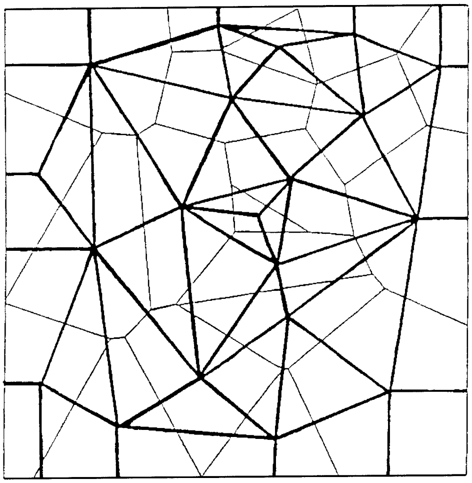

FIG. 7. Two-dimensional Voronoi tesselation of space (thin lines) and Voronoi network (thick lines).

1.8 Continuum Percolation

Although percolation in regular networks has been used extensively for investigating flow and transport in disordered media, percolation in continua is also of great interest, since most real disordered media can be represented as a continuum, at least up to certain length scales. There are at least three ways of realizing a percolating continuum. In the first method, one has a random distribution of inclusions, such as circles, spheres or ellipses, in an otherwise uniform system. Figure 6 shows the so-called Swiss-cheese model in which spherical (circular in 2D) inclusions are punched at random in an otherwise uniform system. If the spheres or circles are not overlapping, and if transport takes place through the channels between the inclusions (as in, for example, flow through a packed bed of solid particles), then the problem can be mapped exactly onto the same transport process on the edges of a Voronoi tessellation of space, a 2D example of which is shown in Fig. 7. Voronoi tessellation (see, for example, Winterfeld *et al.*, 1981) is made by distributing points randomly in the plane and drawing the perpendicular bisector between each pair of points. From the line segments formed by intersections of the perpendicular bisectors between a given point A and its neighboring points is constructed the convex polygon within which every point is closer to A than to any other point. The extension to 3D is obvious. If in the Voronoi tessellations one constructs a network by connecting the centers of the neighboring 3D polyhedra (polygons in 2D) to each other, one obtains a Voronoi network, which is used as a model of a random network (see Fig. 7). One can also construct an inverted Swiss-cheese model in which the roles of the two different media in the Swiss-cheese model are switched, *i.e.*, transport takes place through the inclusions (as in, for example, conduction in granular materials), in which case the problem can be mapped onto the Voronoi network. In both models it is assumed that all the inclusions have the same size. If this assumption is relaxed, then van der Marck (1996) has proposed a general method for mapping such continua onto equivalent network models.

In the second method, one generates a percolating continuum by dividing the space into regular or random polyhedra, a fraction of which is conducting while the rest of them are insulating. This is particularly useful for modelling transport in composite solids. Finally, in the third method, one distributes at random conducting sticks of a given aspect ratio, or plates of a given extent, and studies transport in such a system. This model has been proposed for representing fractured rock, and

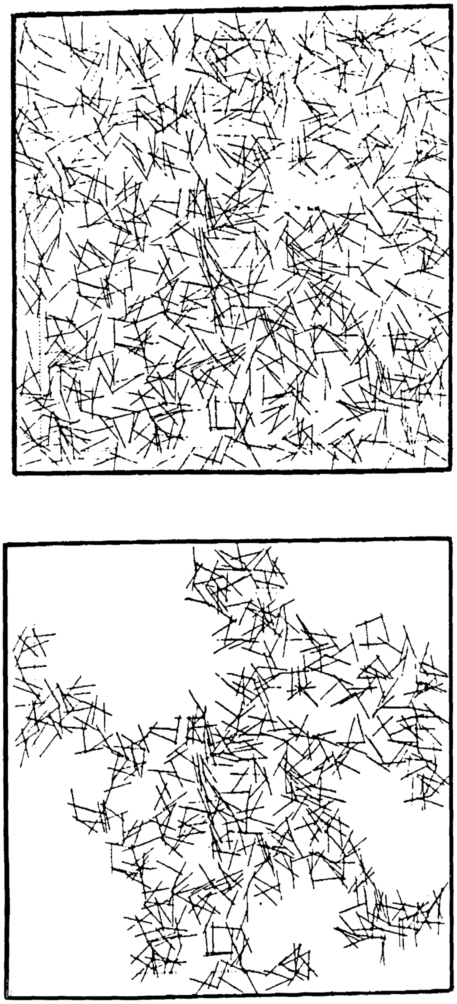

FIG. 8. A two-dimensional model of fibrous material (top) and its backbone (bottom).

also fibrous materials. An example is shown in Fig. 8.

One of the most important discoveries for continuum percolation (Scher and Zallen, 1970) is that the critical value of the occupied volume fraction ϕ_c, defined as

$$\phi_c = p_{cs} f_l, \tag{34}$$

where f_l is the filling factor of a lattice when each site of the lattice is occupied by a sphere in such a way that two nearest-neighbor impermeable spheres touch one another at one point, appears to be an almost invariant of the system, whose value is about 0.17 for 3D systems. Shante and Kirkpatrick (1971) generalized this idea to permeable spheres, and showed that the average number of bonds per sites $B_c = Zp_{cb}$ at p_c is related to ϕ_c by

$$\phi_c = 1 - \exp(-B_c/8), \tag{35}$$

and that the continuum B_c is the limiting value of $p_{cs}Z$ when Z becomes large. It is clear from Table 2 that in 3D, $B_c \simeq 1.5$, and hence $\phi_c \simeq 0.17$, in agreement with the discovery of Scher and Zallen.

It has been established that the topological exponents, defined by Eqs. (4)–(7) and (15), are the same for lattice and continuous systems. For example, Jerauld *et al.* (1984) showed that these exponents for random networks, such as the Voronoi networks, are the same as those for regular networks. Moreover, they established that as long as the average coordination number of a regular network and that of a topologically random network are about the same, most transport properties of the two systems are, for all practical purposes, identical, so long as the same distribution of the transport properties is used in both networks. For example, a 2D Voronoi network has an average coordination number of about six, and the triangular network's coordination number is exactly six.

However, if one uses the equivalent network of a given continuum, one can no longer assign the transport properties of the bonds from an arbitrary distribution, because there is a natural distribution of the conductances of the transport channels in the continuum that can be constructed on the basis of the particular shapes and sizes that the channels take on. Such natural distributions give rise to new scaling laws which cannot be predicted by the network models, unless the same natural distributions are also used in the networks. To see the difference between flow and transport in random percolation networks and continua, we consider a few examples that have been studied thoroughly. It was shown by Halperin *et al.* (1985) and Feng *et al.* (1987) that the scaling of the transport properties of the Swiss-cheese and the inverted Swiss-cheese models near their percolation thresholds is very different from those of percolation networks. In particular, if μ_{sc} and μ_{isc} denote the critical exponents of the conductivity of the Swiss-cheese and the inverted Swiss-cheese models, then in 2D $\mu_{\text{sc}} = \mu_{\text{isc}} = \mu$, while in 3D $\mu_{\text{isc}} = \mu$ and

$$\mu_{\text{sc}} \simeq \mu + 1/2. \tag{36}$$

Moreover, unlike the case of lattice percolation discussed above (for which $\chi = \mu$), the permeability exponent of continuum percolation is very different from the conductivity exponent μ. In 2D, $\chi_{\text{isc}} = \mu$ and

$$\chi_{\text{sc}} \simeq \mu + 3/2, \tag{37}$$

whereas in 3D

$$\chi_{\text{sc}} \simeq \mu + 5/2, \quad \chi_{\text{isc}} = \mu + 1/2, \tag{38}$$

so that the permeability is much more strongly affected by the continuous structure of the system. It was also shown by Straley (1982), Halperin *et al.* (1985), and Feng *et al.* (1987) that if the conductance distribution of a random continuum is of power-law form, $\psi(g) = (1-p)\delta(g) + p(1-\alpha)g^{-\alpha}$, where $0 \leq \alpha \leq 1$, then the conductivity exponent $\mu(\alpha)$ of this systems is not universal and is given by

$$\begin{aligned}\mu_1(\alpha) &= 1 + (d-2)\nu + \frac{\alpha}{1-\alpha} \\ &\leq \mu(\alpha) \leq \mu + \frac{1}{1-\alpha}.\end{aligned} \tag{39}$$

Such power-law conductance distributions are actually very common among disordered continua. For example, for the Swiss-cheese models in d dimensions, one has $\alpha = (2d-5)/(2d-3)$. In 2D, $\alpha = -1$, and the conductance distribution is not singular at $g = 0$ [inequality (30) is satisfied], and therefore $\mu_{\text{sc}} = \mu$. In 3D, $\alpha = 1/3$, and the conductance distribution is singular at $g = 0$, hence giving rise to Eq. (39). Moreover, Straley (1982) argued that

$$\mu(\alpha) = \min[\mu, \mu_1(\alpha)]. \tag{40}$$

These results have not only been confirmed by computer simulations, they have also received experimental verification (see, for example, Garfunkel and Wiessman, 1985).

1.9 Calculation of Flow and Transport Properties of Percolation Systems

Let us now discuss how flow and transport properties of percolation systems are estimated. Here we review briefly a few of the methods that have proven to be efficient and, at least under certain conditions, accurate.

1.9.1 Monte-Carlo Simulation This method is the most accurate technique for estimating flow and transport properties of percolation networks, although extensive numerical work may be involved. We first discuss calculation of the effective conductivity of a percolation network. According to Kirchhoff's law, for every site of the network the sum of all the currents reaching it is zero. Thus, if J_{ij} is the current in the bond ij, then for site i of the network we must have $\sum_j J_{ij} = 0$. Since according to Ohm's law, $J_{ij} = (V_i - V_j)g_{ij}$, where V_i is the voltage or potential at i, and g_{ij} is the conductance of the bond ij, we obtain

$$\sum_j (V_i - V_j)g_{ij} = 0. \tag{41}$$

Equation (41), when written for every interior site of the network, yields a set of simultaneous linear equations for the nodal voltages or potentials. As the boundary conditions, one typically imposes a fixed voltage drop between two opposite faces of the network and periodic boundary conditions in the other direction(s). Alternatively, one can use a constant-current boundary condition in which a fixed current is injected into the network at one or several sites. The set of equations is typically solved by an efficient iterative algorithm such as the conjugate-gradient method. Having solved the set, one can easily determine the overall conductivity G of the network. For example, if as the boundary condition a constant voltage drop has been imposed on the network, G is determined by calculating the total current that flows through the network, and dividing it by the applied voltage drop. One can use exactly the same method for calculating the effective thermal conductivity of a composite solid.

The number of iterations for the convergence of the set of equations to their true solution increases rapidly as $p \to p_c$. To avoid this difficulty, special methods have been devised that build up the network one bond at a time (Derrida *et al.*, 1984b; Duering and Bergman, 1990), or take advantage of laws of electrical networks (Gingold and Lobb, 1990). However, these methods are most efficient near p_c, but not far from it. For an efficient method both near and away from p_c see Mehrabi and Sahimi (1997).

The same type of Monte-Carlo simulation can also be used for calculating the effective permeability K or diffusivity D of a porous medium, modeled by a percolation network. Consider, for example, the problem of estimating K. Each occupied bond represents a pore throat through which a fluid flows, while each site represents a pore body at which the throats meet. A porous medium is typically represented by a pore size distribution $\Psi(r)$, so that we assign an effective radius r, selected from $\Psi(r)$, to the bonds. Assuming, for example, that the flow is laminar, the volumetric flow rate in the throat ij is given by, $Q_{ij} = \pi(P_i - P_j)r_{ij}^4/(a\eta l_{ij})$, where P_i is the pressure at i, η is the viscosity of the fluid, l_{ij} is the throat (bond) length, and a is a constant geometrical factor that depends on the shape of the throats. We must have conservation of mass at every site i of the network, *i.e.*, the net mass of the fluid reaching i must be zero, and therefore, $\sum_j \rho Q_{ij} = 0$. If the fluid density ρ is constant, and if the pressure drops in the sites are negligible (which is often true, as the pore bodies are typically much larger than the throats which are long and narrow), then we must have

$$\sum_j \pi(P_i - P_j)r_{ij}^4/(a\eta l_{ij}) = 0, \qquad (42)$$

which is completely analogous to Eq. (41). The rest of the analysis is obvious: writing down Eq. (42) for each site i of the network yields a set of simultaneous equations for the nodal pressures, from the solution of which the effective permeability K of the network is computed in a manner completely similar to G. The same type of model and method can also be used for estimating the effective diffusivity of the network.

1.9.2 Effective-Medium Approximation

One of the simplest, and in many cases accurate, analytical methods of estimating effective transport properties of disordered media is the effective-medium approximation (EMA). The original EMA was developed by Bruggeman (1935), and independently by Landauer (1952; for a review of history of the EMAs see Landauer, 1978) for a continuum. It was extended by Kirkpatrick (1973) to regular networks of random conductances, and has been shown to be accurate under a variety of conditions.

We first consider estimating the effective conductivity of a percolation network by an EMA. Consider a network of coordination number Z, and suppose that the conductance of the bonds is selected from a distribution $\psi(g)$. In an EMA approach, the random conductances of all the bonds are replaced by G to create a uniform network. The potential differences along the bonds of such a network are known exactly. One then replaces, in the uniform network, the effective conductance of one bond with its true value g, which causes fluctuations in the potential difference in that bond with respect to its value in the uniform network. One insists that the average of these fluctuations, when the averaging is taken with respect to $\psi(g)$, should be zero. This results in a simple equation for G given by (Kirkpatrick, 1973)

$$\int_0^\infty \frac{G - g}{g + G(Z/2 - 1)} \psi(g)\, \mathrm{d}g = 0. \qquad (43)$$

If we use a percolation distribution (which corresponds to bond percolation)

$$\psi(g) = (1 - p)\delta(g) + (1 - p)\delta(g - g_0), \qquad (44)$$

then it is easy to see that Eq. (43) predicts that G vanishes at $p_{cb} = 2/Z$. Thus, EMA predicts the exact value $p_{cb} = 1/2$ for the square lattice, and $p_{cb} = 1/3$ for the simple cubic lattice, which should be compared with the numerical estimate $p_{cb} \simeq 0.2488$. Moreover, EMA predicts that the critical exponent $\mu = 1$, which does not agree with the numerical estimates (see Table 3). For a d-dimensional disordered continuum with randomly distributed spherical or circular inclusions, Eq. (43) can still be used if $Z/2$ is replaced by d.

We can also use an EMA for estimating the permeability and diffusivity of a porous medium. Consider an effective-medium network in which each bond or pore throat has a conductance G. We fix the pressures at two opposite faces of the network so as to produce an average pressure gradient $\langle \nabla P\rangle$. The total volume Q of fluid crossing any plane perpendicular to the average pressure gradient is the sum of the individual fluid volumes in the bonds intersecting the plane. If we approximate the local pressure difference as the projection of the average pressure gradient along the bond length l, we find

$$Q = \sum \frac{G}{\eta} \langle \nabla P\rangle \cdot \boldsymbol{l}. \tag{45}$$

If we divide Q by the area S of the plane, we obtain an average velocity and hence an estimate of the permeability:

$$K = G\left\langle \frac{1}{S} \sum \boldsymbol{l} \cdot \boldsymbol{n} \right\rangle, \tag{46}$$

where $\mathbf{n}$ is a unit vector along the pressure gradient. However, if the porous medium is statistically homogeneous and isotropic, any unit vector can be used. As in the case of the conductivity, EMA predicts that the critical exponent of the permeability is $\chi = 1$.

The effective diffusivity D of the porous medium can also be estimated by an EMA. It is not difficult to show that

$$D = \frac{1}{3\pi} \frac{Gl}{\langle r^2\rangle} \varphi, \tag{47}$$

where φ is the porosity of the porous medium, and $\langle r^2\rangle = \int_0^\infty r^2 \Psi(r)\,\mathrm{d}r$, where $\Psi(r)$ is the pore size distribution of the porous medium. Note that in both Eqs. (46) and (47), G is the EMA estimate of the conductivity of the network. The accuracy of the EMA has been checked under a wide variety of conditions. In general, it is very accurate if the system is not close to its percolation threshold, regardless of the structure of $\psi(g)$ or $\Psi(r)$. It is also more accurate for 2D systems than 3D ones. We refer the reader to Koplik (1981) and Doyen (1988) for typical experimental and Monte-Carlo tests of the EMA's accuracy.

1.9.3 Position-Space Renormalization Group Method If the disorder is strong, or if the system is near its percolation threshold, then the EMA is no longer accurate. In such cases, a position-space renormalization-group (PSRG) method is more accurate. Consider, for example, a square or a cubic network in which each bond is occupied (conducting) with probability p. In a PSRG approach one partitions the network into $b \times b$ or $b \times b \times b$ cells (this corresponds to bond percolation), where b is the number of bonds in any direction, and calculates their percolation properties, which are hopefully representative of the properties of the original network. The shape of the cell can be selected arbitrarily, but it should be chosen in such a way that it preserves, as much as possible, the properties of the original network. Figure 9 presents examples of such RG $b = 2$ cells for the square and cubic networks, which have been shown to yield the most accurate results for these networks. The next step in a PSRG method is to replace each RG cell with one bond in each principal direction. If in the original network each bond is conducting with probability p, then the bonds that replace the cells would be conducting with probability $p' = R(p)$ (see Fig. 9), where $R(p)$ is called the *renormalization-group transformation*, and is the sum of the probabilities of all possible conducting configurations of the RG cell, which is obtained as follows. We solve for the percolation and transport properties in each cell by applying a fixed potential difference across the cell in a given direction. For exam-

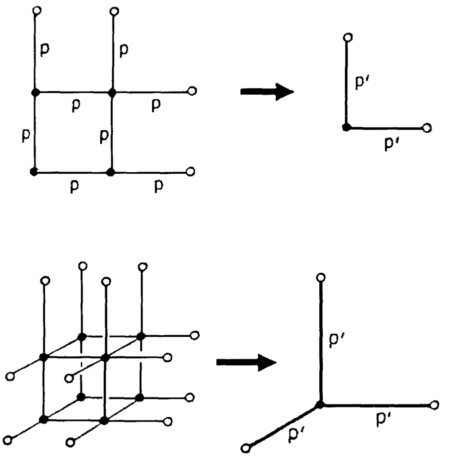

FIG. 9. Two- and three-dimensional RG cells with $b = 2$. $p' = R(p)$ is the RG transformation.

ple, as far as percolation and transport are concerned, the 2 × 2 cell of Fig. 9 is equivalent to the circuit shown there. Thus for this RG cell we only need to deal with five bonds, and for the 2 × 2 × 2 cell we have an equivalent 12-bond circuit (see Fig. 9). To obtain $R(p)$, we find all the conducting configurations of such circuits, with some bonds conducting and some insulating. Thus for the 2 × 2 cell we obtain

$$p' = R(p) = p^5 + 5p^4q + 8p^3q^2 + 2p^2q^3, \tag{48}$$

where $q = 1 - p$. It is easy to see how Eq. (48) is obtained: there is only one conducting cell configuration with all five bonds conducting (probability p^5), five conducting configurations with four bonds conducting and one bond insulating (probability $5p^4q$), and so on.

To obtain an estimate of p_c, we solve the following equation:

$$R(p_c) = a, \tag{49}$$

where $0 < a < 1$ is any arbitrary number. Often $a = p_c$ gives the most accurate results. For example, for the RG cells of Fig. 9 we obtain the exact result $p_c = 1/2$ for both $b = 2$ and $b = 3$ cells, if we use $a = p_c$. In fact, it can be shown (Bernasconi, 1978) that the RG transformation for such 2D cells for any b always predicts $p_c = 1/2$, if one uses $a = p_c$. For the 2 × 2 × 2 cell we obtain $p_c \simeq 0.208$, compared with the numerical estimate $p_{cb} \simeq 0.2488$. One then calculates p_c for several cell sizes b and uses finite-size scaling according to which

$$p_c(\infty) - p_c(b) \sim b^{-1/\nu}, \tag{50}$$

where $p_c(\infty)$ is the true percolation threshold of the network and ν is the critical exponent of the percolation correlation length ξ_p, to extrapolate the results to $b \to \infty$ and estimate p_c (Reynolds *et al.*, 1980). The critical exponent ν of the correlation length is estimated from $\xi'_p = \xi_p/b$, or

$$\nu = \ln b/\ln \lambda_p, \tag{51}$$

where $\lambda_p = dR(p)/dp$, evaluated at p_c. Thus for the 2 × 2 and 3 × 3 cells we obtain $\nu \simeq 1.43$ and 1.38, respectively. For the 2 × 2 × 2 cell we obtain $\nu \simeq 1.03$, compared with the numerical estimate, $\nu \simeq 0.88$.

In a PSRG approach for calculating the conductivity (or permeability) of a network (Stinchcombe and Watson, 1976), one starts with a conductance (or permeability) distribution $\psi_0(g)$ for the bonds of the RG cell and replaces it with a new distribution $\psi_1(g)$, the distribution for the conductance of a bond in the renormalized RG cell. The $\psi_1(g)$ value is calculated by determining the equivalent conductance of the RG cell. Thus one obtains a recursion relation relating $\psi_1(g)$ to $\psi_0(g)$:

$$\psi_1(g) = \int \psi_0(g_1)\, dg_1 \psi_0(g_2)\, dg_2 \cdots \psi_0(g_n) \times dg_n \delta(g - g'), \tag{52}$$

where $g_1, \ldots, g_n$ are the conductances of the n bonds of the cell and g' is the equivalent conductance of the RG cell. For example, for the five-bond cell of Fig. 9 one has

$$g' = \frac{g_1(g_2g_3 + g_2g_4 + g_3g_4) + g_5(g_1 + g_2)(g_3 + g_4)}{(g_1 + g_4)(g_2 + g_3) + g_5(g_1 + g_2 + g_3 + g_4)}, \tag{53}$$

and it is not difficult to show that, if we use distribution (44), then for the five-bond RG cell of Fig. 9, Eq. (52) yields

$$\begin{aligned}\psi_1(g) = {} & [1 - R(p)]\delta(g) + 2p^3q^2\delta(g - g_0/3) \\ & + 2p^2(1 + 2p)q^2\delta(g - g_0/2) \\ & + 4p^3q\delta(g - 3g_0/5) \\ & + p^4\delta(g - g_0).\end{aligned} \tag{54}$$

One iterates Eq. (52) again to obtain a new distribution $\psi_2(g)$ by substituting $\psi_1(g)$ into the right-hand side of Eq. (52). The iteration process continues until a distribution $\psi_\infty(g)$ is obtained, the shape of which does not change under further iterations. This is called the *fixed-point distribution*, and the conductance of the original network is simply an average of this distribution. However, it is difficult to iterate Eq. (52) analytically many times. The common practice is to replace $\psi_i(g)$ after iteration i by an *optimal* distribution $\psi_i^{\rm o}(g)$, which mimics $\psi_i(g)$ closely. The optimal $\psi_i^{\rm o}(g)$ is usually taken to have the form

$$\psi_i^{\rm o}(g) = [1 - R(p)]\delta(g) + R(p)\delta[g - g^{\rm o}(p)], \tag{55}$$

where $g^{\rm o}(p)$ is an optimal conductance. One of the most accurate approximations to $g^{\rm o}(p)$ was proposed by Bernasconi (1978). If after i iterations $\psi_i(g)$ is given by

$$\psi_i(g) = [1 - R(p)]\delta(g) + \sum_i a_i(p)\delta(g - g_i), \tag{56}$$

then $g^{\rm o}(p)$ is approximated by

$$g^{\rm o}(p) = \exp\left[\sum_i a_i(p) \ln g_i / R(p)\right]. \tag{57}$$

Once $g^{\rm o}(p)$ is calculated, Eq. (52) is iterated again, a new distribution $\psi_{i+1}(g)$ and the associated $\psi_{i+1}^{\rm o}(g)$ are determined, and so on. In practice, after a few iterations even an initially broad $\psi_0(g)$ converges quickly to a distribution whose shape does not change under further rescaling. When this happens, $\lambda_{\rm c}\psi_{n+1}(\lambda_{\rm c}g) \simeq \psi_n(g)$, where $\lambda_{\rm c}$ is a constant. The critical exponent μ of the conductivity is then given by

$$\mu = \nu \ln \lambda_{\rm c} / \ln b. \tag{58}$$

For the 2×2 RG cell of Fig. 9 we obtain $\mu \simeq 1.32$, and for the 3D cell of Fig. 9, $\mu \simeq 2.14$. These results should be compared with the numerical estimates given in Table 3. Moreover, the predicted conductivities in 2D are in good agreement with the simulation results, but are not as accurate in 3D (Bernasconi, 1978).

Renormalization methods are usually very accurate for 2D systems. However, they have a main drawback for 3D systems, in that the predicted conductivity or permeability for a percolation network with any type of $b = 2$ cell is not accurate, if the network is near $p_{\rm c}$. Moreover, even after the first iteration of (52) the renormalized conductance distribution $\psi_1(g)$ is very complex, making the computations very difficult.

1.9.4 Renormalized Effective-Medium Approximation Sahimi *et al.* (1983b, 1984) proposed a new method that combines EMA and PSRG, and is called the renormalized EMA (REMA). Their method takes advantage of two main strengths of EMA and PSRG, which are the facts that an EMA is accurate away from $p_{\rm c}$, and that when we rescale an RG cell, the renormalized cell is farther away from $p_{\rm c}$ than the original cell (because in the renormalized cell, $\xi_p' = \xi_p/b$). Thus one may use Eq. (43) with the renormalized conductance distribution $\psi_1(g)$, instead of $\psi_0(g)$. This method is drastically more accurate than both EMA and PSRG in both 2D and 3D. For example, with the 3D RG cell of Fig. 9, REMA predicts that $p_{\rm cb} \simeq 0.267$, only 7% larger than $p_{\rm cb} \simeq 0.2488$ for the cubic network. In general, if $R(p)$ is the RG transformation, REMA predicts that $p_{\rm cb}$ is the root of $R(p_{\rm cb}) = 2/Z$. Using REMA, Sahimi *et al.* (1983b, 1984) and Sahimi (1988) obtained very accurate predictions of transport properties of various 2D and 3D networks; an example is given in Fig. 10.

Having defined and discussed the basic concepts of percolation theory, and various methods for estimating the flow and transport properties of percolation systems, we now embark on discussing their applications.

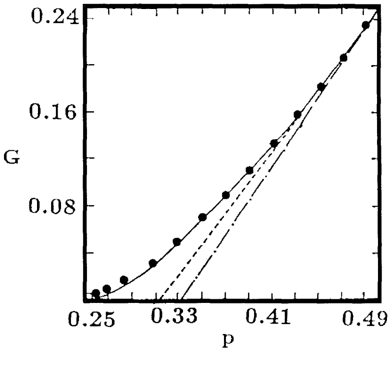

FIG. 10. Comparison of the predictions of the REMA for the conductivity G of a simple cubic network (solid curve) with those of the EMA (dash-dotted line), the cluster EMA (dashed line) and Monte Carlo results (circles).

2. APPLICATION TO COMPOSITE MATERIALS

Composite materials and media constitute a large and important class of naturally occuring or man-made disordered systems. Included in this class are granular materials, composite solids, metal–insulator films, porous media, polymers and aggregates, colloids, and many others. In this section we discuss the application of percolation theory to predicting the effective properties of such materials. Far from p_c, the effective properties of composite materials are predicted accurately by the EMA and PSRG methods discussed above; the interested reader should consult Lafait and Tanner (1989) and Cody *et al.* (1990) for discussion and references. Thus, we restrict our discussion to the scaling behavior of the composites near p_c where percolation effects are dominant and universal power laws should be observed. The existence of the universal scaling laws of percolation theory is the most stringent test of applicability of the theory. It is also often true that the critical region is quite broad, and therefore such universal scaling laws can be very useful for predicting the effective properties of a composite medium over a broad range of the volume fraction of its components.

2.1 Conductivity of Granular Media

One of the simplest 3D composite solids is a random packing of hard spheres, whose effective properties are relevant to those of a wide variety of systems of practical importance. For example, if all or some of the spheres are conducting, the packing is a model for the (electrical or thermal) conductivity of powders and other granular media. If in the pore space between the spheres a fluid flows, the packing is used for studying flow phenomena in unconsolidated porous media.

The first systematic and extensive study of percolation and conduction properties of packings of particles was carried out by Ottavi *et al.* (1978). They used molded plastic spheres a fraction of which were electroplated with a copper coating to make them conducting. Their data, shown in Fig. 11, indicate that the effective conductivity G vanishes at $p_c \simeq 0.29 \pm 0.02$. Since the filling factor f_1 (see Sec. 1.8) of such packings is about 0.6, one obtains

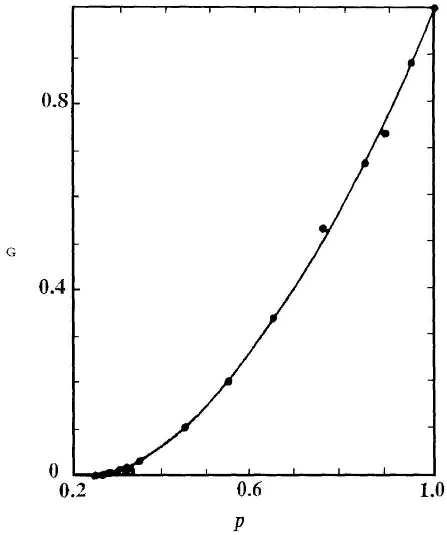

FIG. 11. Conductivity of a mixture of conducting and nonconducting spheres *versus* the fraction p of the conducting spheres (after Ottavi *et al.*, 1978).

$\phi_c = f_1 p_c \simeq 0.17$, in agreement with the prediction of Scher and Zallen (1970) for continuum percolation discussed above. In the critical region, their data followed Eq. (8) with $\mu \simeq 1.7 \pm 0.2$, not too far from $\mu \simeq 2.0$ for 3D percolation (see Table 3).

Similar studies were carried out by Deptuck *et al.* (1985), who investigated conduction in a sintered, submicrometer silver powder (and also its Young's modulus; see Sec. 6.3) with a volume fraction $\phi \sim 0.4$, commonly used for millikelvin and submillikelvin cryostats, and found that the conductivity vanishes with a critical exponent $\mu \simeq 2.15 \pm 0.25$, completely consistent with $\mu \simeq 2.0$ for 3D percolation (see Table 3); by Lee *et al.* (1986), who used silver-coated-glass–Teflon composites and found that $\phi_c \simeq 0.17$, in good agreement with the percolation threshold of 3D continuum percolation predicted by Scher and Zallen (1970), and $\mu \simeq 2.0 \pm 0.2$, in excellent agreement with that of 3D percolation; by Deprez *et al.* (1989), who measured the electrical conductivity of sintered nickel samples, and also found that Eq. (8) (with $\mu \simeq 2.0$) was applicable over the *entire range* of the volume fraction; and by Hsu and Berzins (1985), who studied percolation effects in a polymer composite, namely, perfluorinated ionomers, which have the general formula $[(CF_2)_n\ CF]_m$–O–R–SO_3X, where typically $n = 6$–13, R is a perfluoro alkylene group which may contain ether oxygen, and X is any monovalent cation. These composites possess exceptional transport, chemical, and mechanical properties, and have been used as membrane separators, acid catalysts, and polymer electrodes. Percolation effects are important in these composites because of a spontaneous phase separation occuring in the wet polymer, which is the conductive aqueous phase distributed randomly in the insulative fluorocarbon phase. Hsu and Berzins (1985) found that their data for the electrical conductivity of these polymer composites follow Eq. (8) with $\mu \simeq 2.0$ over the wide range $0.01 \leq p - p_c \leq 0.8$. For the most recent experimental work see Andrade *et al.* (1998), who studied the conductivity of vapor-grown carbon fiber (VGCF). They observed a percolation transition as the density of compressed VGCF was changed, with a conductivity exponent $\mu \simeq 1.91 \pm 0.06$, fully compatible with 3D percolation.

Another important class of composite solids whose electrical conductivity has been modeled by percolation theory is dispersed ionic conductors. Mixtures of solid ionic conductors that consist of fine particles and an insulating second phase exhibit a remarkable increase in their conductivity when compared with pure and homogeneous systems. For example, if we add small particles of an insulating phase, *e.g.*, Al_2O_3, to ionic conductors, such as LiI or AgI, the conductivity of the composite is enhanced by a few orders of magnitude. It is generally believed that this effect is caused by an enhanced conductivity along the interface between the conducting matrix and the dispersant particles. Bunde *et al.* (1985) and Roman *et al.* (1986) proposed a percolation model for predicting the conductivity of such composites. In their model one starts with a simple cubic lattice in which the elementary cubes are occupied randomly with probability p. The bonds that connect two neighboring sites are considered as resistors, and three types of resistors are recognized:

- **(1)** highly conducting bonds with interface conductance g_1;
- **(2)** normally conducting bonds with conductance g_2;
- **(3)** insulating bonds with zero conductance.

The bonds are highly conducting if they belong to one, two or three occupied cubes, and are insulating if they belong to four occupied cubes. The rest of the bonds are normal conductors. Thus, the fractions of the insulating, normally conducting and highly conducting bonds are, respectively, $p_3 = p^4$, $p_2 = (1-p)^4$, and $p_1 = 1 - p^4 - (1-p)^4$. There are also two percolation thresholds in the system. One of them, p_{c1}, defines the lowest concentration of the highly conducting bonds at which a sample-spanning cluster of such bonds is formed. The second one, p_{c2}, defines the percolation threshold where all the normal and insulating bonds get disrupted. In the simple cubic network, $p_{c1} \simeq 0.097$ and $p_{c2} = 1 - p_{c1} \simeq 0.903$. At both thresholds one has $p_3 \simeq 0.335$, somewhat larger than $p_{cb} \simeq 0.2488$ for the bond-percolation threshold of random two-component systems discussed above. Bunde *et al.* (1985) and Roman *et al.* (1986) showed that the electrical conductivity of this system exhibits a maximum as p is varied, in agreement with the experimental data, if p is interpreted as the mole fraction of the insulating phase. Roman (1990) extended this lattice model to continuous systems using random distribution of spheres; see also Lauer and Maier (1992).

Many composite materials are anisotropic, and the anisotropy influences their effective properties near p_c. However, the anisotropy should vanish at p_c, since near p_c the conducting paths are so tortuous that the current cannot distinguish between different directions. Indeed, Shklovskii (1978) proposed that the anisotropy ratio G_l/G_t obeys the following power law near p_c:

$$G_l/G_t - 1 \sim (p - p_c)^{\lambda}, \tag{59}$$

where G_l and G_t are the longitudinal and transverse conductivities and λ is a universal critical exponent; in 3D, $\lambda \simeq 0.5$ (Mukhopadhyay and Sahimi, 1994). If the system always remains conducting (as in a mixture of good and poor conductors), then the anisotropy should be minimum at p_c. The conductivity of anisotropic systems can also be predicted by an EMA (Bernasconi, 1974), and indeed EMA predicts that the conductivity anisotropy vanishes at p_c. These predictions of percolation theory have been confirmed by the experimental data of Smith and Lobb (1979), who measured the conductivities of 2D conductor–insulator networks generated photolithographically from laser speckle patterns; by those of Troadec *et al.* (1981), who measured thermal and electrical conductivity of conducting (WTe_2) and semiconduting (WSe_2) powders, characterized both by a geometrical anisotropy (grain shapes and sizes) and by anisotropy in a transport property; and by the results of Balberg *et al.* (1983), who measured the resistivity of a composite composed of elongated carbon black aggregates embedded in an insulating plastic, poly(vinyl chloride) (anisotropy of the system is due to elongation of the aggregates).

2.2 Percolation and Transport Properties of Metal–Insulator Films

Thin metal films are two-phase mixtures of metal and nonmetal components, which have interesting structural and transport properties. If the volume fraction of the metallic phase is large enough, it forms a sample-spanning cluster such that the electrical conductivity of the film is large and its temperature coefficient of resistance is positive. On the other hand, if the nonmetallic phase forms a sample-spanning cluster, the system is in the dielectric regime with a very small electrical conductivity. Transmission electron microscopy shows that thin metal films have a discontinuous structure in which the metallic phase constitutes only a fraction of the total volume of the system. Generally speaking, a metal film is considered as thin if its thickness l is less than the percolation correlation length ξ_p of the corresponding 3D film. Thick composite films also have interesting properties. Examples of such composites are Ni–SiO_2, Al–Si and Al–Ge. There are also strong correlations in the structure of thick metal films, so that a sort of correlated percolation may be applicable. Deutscher *et al.* (1978) proposed that thick metal films should be considered as random composites if the insulating phase is crystalline.

Percolation studies of thin (2D) metal films appear to have been begun by Liang *et al.* (1976), using SiO films deposited on a glass slide from a microscope. They found that the conductivity of the film near ϕ_c, the percolation threshold of the conducting phase, follows Eq. (8) with $\mu \simeq 1.15$, reasonably close to $\mu \simeq 1.3$ for 2D percolation. Important studies of morphological properties of thin metallic films were carried out by Voss *et al.* (1982) and Kapitulnik and Deutscher (1982). Voss *et al.* prepared Au films by electron-beam evaporation onto 30-nm-thick amorphous Si_3N_4 windows grown on an Si wafer frame. Figure 12 shows one of their samples at a fractional Au coverage $p \simeq 0.71$, just below $p_c \simeq 0.74$. At p_c the largest Au cluster was a fractal object with $D_F \simeq 1.9$, in excellent agreement with 2D percolation (see Table 3). Kapitulnik and Deutscher (1982) prepared samples by successive deposition of amorphous Ge as the substrate and of thin Pb films as the metal. Their study indicated that the sample-spanning cluster and its backbone exhibit fractal behavior, with $D_F \simeq 1.9$ and $D_B \simeq 1.65$, in excellent agreement with 2D percolation (see Table 3). Also analyzed was n_s, the number of metallic clusters of mass s below p_c, which was found to follow Eq. (15) with $\tau \simeq 2.1 \pm 0.2$, completely consistent with $\tau = 187/91 \simeq 2.054$ for 2D percolation. Similar results were obtained by Laibowitz *et al.* (1982), who used thin Al–Al_2O_3 films, by Gordon and Goldman (1988), who used Al thin films, and by Papandreou and Nédellec (1992), who used Pd films with a

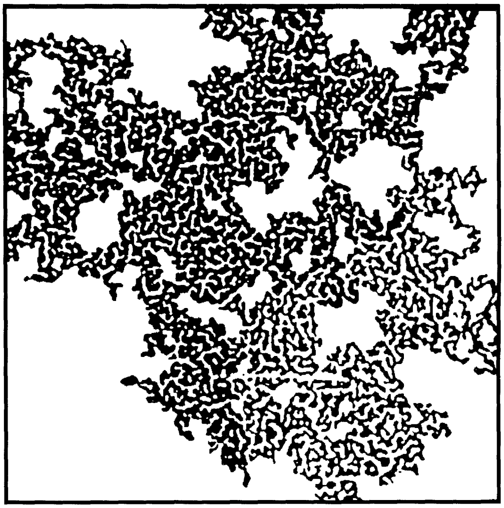

FIG. 12. Two-dimensional Au clusters at $p \simeq 0.71$ (after Voss *et al.*, 1982).

typical thickness of about 10 nm. The last authors also found a conductivity exponent $\mu \simeq 1.3$, in complete agreement with 2D percolation (see Table 3).

In a seminal paper, Abeles *et al.* (1975) studied the growth of the grains of the metals W or Mo in the insulators Al_2O_3 or SiO_2 using co-sputtering. Their measured conductivity exponent, $\mu \simeq 1.9 \pm 0.2$, is in very good agreement with $\mu \simeq 2.0$ for 3D percolation (see Table 3). Similar results were also reported by Kapitulnik and Deutscher (1983), who used Al–Ge films and found that $\mu \simeq 2.1 \pm 0.5$. More recent experimental data for In–Ge thin films were given by Tessler and Deutscher (1989), and for Al–Ge films by Kapitulnik *et al.* (1990). Finally, we should mention that optical properties of percolating metal films have also been measured and explained. The reader is referred to Yagil and Deutscher (1988), Gadenne and Gadenne (1989), Gadenne *et al.* (1989), and Robin and Souillard (1989) for details.

2.3 ac Conductivity and Dielectric Properties of Composite Media

An EMA can be used for predicting the ac conductivity of disordered media that are not near their percolation threshold (see, for example, Odagaki and Lax, 1981; Sahimi *et al.*, 1983a). A comprehensive review of this subject is provided by Clerc *et al.* (1990). Here we only summarize the main theoretical results, and discuss their experimental verification. Consider the two-component percolation network of resistors described in Sec. 1.4. In a percolation model for the ac conductivity and dielectric properties, one views g_1 and g_2 as complex conductances, with $g_1 = 1/R$ and $g_2 = \mathrm{i}C\omega$, where R is the resistance of the bond, C is the capacitance, ω is the frequency, and $\mathrm{i} = \sqrt{-1}$. The conductance ratio h is then given by $h = \mathrm{i}\omega/\omega_0$, where $\omega_0 = 1/RC$. In the static limit ($\omega = 0$) the capacitors become insulators, and the model reduces to the usual percolating conductor–insulator mixture. The key result is that near p_c one has

$$G(p,\omega) \sim \frac{1}{R}|p - p_c|^{\mu}\Phi_{\pm}\left(\frac{\mathrm{i}\omega}{\omega_0}|p - p_c|^{-\mu-s}\right), \tag{60}$$

which follows directly form Eq. (12). An immediate consequence of Eq. (60) is the existence of a time scale t_s that diverges as p is approached from *both sides*:

$$t_s \sim \omega_0^{-1}|p - p_c|^{-(\mu+s)}. \tag{61}$$

We now define the frequency-dependent complex dielectric constant $\varepsilon(p,\omega)$ of the system by $\varepsilon(p,\omega) = G(p,\omega)/i\omega$, which is a generalization of the usual static dielectric constant ε_0. For an insulating dielectric medium $G \simeq i\omega\varepsilon_0$ as $\omega \to 0$. Bergman and Imry (1977) derived the following relationships:

$$G(p_c,\omega) \sim \omega^x, \tag{62}$$

$$\varepsilon(p_c,\omega) \sim \omega^{-y}, \tag{63}$$

where the exponents x and y satisfy

$$x + y = 1. \tag{64}$$

Gefen *et al.* (1983) argued that the fractal nature of diffusion and percolation clusters at length scales up to ξ_p dominates the contributions to the ac properties, and hence proposed that

$$x = \mu/\nu(2 + \theta), \tag{65}$$

$$y = (2\nu - \beta)/\nu(2 + \theta), \tag{66}$$

where the exponents ν, β, and θ were defined in Secs. 1.5 and 1.6.

Another important consequence of Eq. (60) is that

$$\varepsilon_0 \sim A_{\pm}C|p - p_c|^{-s}, \tag{67}$$

that is, the static dielectric constant *diverges* as p_c is approached from *both sides*, and the critical exponent that characterizes this divergence is s, the critical exponent of a percolation network of superconductors–conductors defined by Eq. (9). This spectacular result was first derived by Efros and Shklovskii (1976). Although Eq. (67) is supposed to be valid in the limit $\omega = 0$, its validity extends to higher frequencies as long as $\omega \ll 1/t_s$, where t_s is given by Eq. (61). The amplitude ratio A_+/A_- is universal. If we now write $G(p,\omega) =$

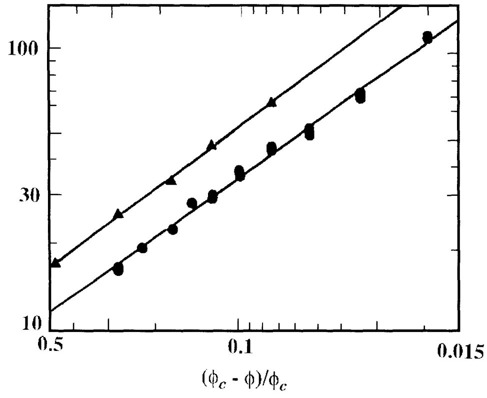

FIG. 13. Logarithmic plot of the dielectric constant of two series of Ag–KCl composites *versus* the volume fraction φ of Ag. The solid lines are the best fit of the data (after Grannan *et al.*, 1981).

$G' + iG'' = i\omega\varepsilon(p,\omega) = i\omega(\varepsilon' - i\varepsilon'')$, then a loss angle is defined by

$$\tan\delta = G'/G'' = \varepsilon''/\varepsilon', \tag{68}$$

and it is clear that $0 \le \delta \le \pi$. At p_c the loss angle δ_c is universal and is given by

$$\delta_c = \tfrac{1}{2}\pi s/(\mu + s). \tag{69}$$

Although Eq. (69) is supposed to be valid exactly at $p = p_c$, its validity extends to a broad frequency range if $|p - p_c|$ is small enough, *i.e.*, $1/t_s \ll \omega \ll \omega_0$. Moreover, since for 2D percolation we have $s = \mu$, we must have $\delta_c = \pi/4$, another remarkable result.

Equations (64)–(66) were confirmed by the experiments of Hundley and Zettl (1988), who measured the ac conductivity and dielectric constant of thin Au films in the frequency range 100 Hz–1 GHz. Their measurements indicated that at high frequencies $x \simeq 0.32$ and $y \simeq 0.8$, in reasonable agreement with the predictions of Eqs. (65) and (66), $x \simeq 0.34$ and $y \simeq 0.66$. A definitive verification of Eq. (67) was provided by Grannan *et al.* (1981), who used a composite consisting of small spherical Ag particles randomly distributed in a nonconducting KCl host. Figure 13 shows the scaling plot of their data as a function of the volume fraction of Ag, indicating that ε_0 diverges as ϕ_c is approached with an exponent $s \simeq 0.73 \pm 0.01$, in perfect agreement with the 3D percolation prediction $s \simeq 0.735$. Similar results were obtained by Nicklasson and Grangvist (1984).

2.4 Hall Conductivity of Composite Media

In the presence of a magnetic field H the conductivity of a composite is described by a tensor $\boldsymbol{G}$ which has nonzero off-diagonal terms, even if the system is isotropic. Some of the off-diagonal terms are symmetric and even in the magnetic field H, while others are antisymmetric and odd in H. Suppose that an electric field $\mathbf{E}$ has also been imposed on the system. Then the system has an electrical (Ohmic) conductivity G and a Hall conductivity $\boldsymbol{G}_H$. For an isotropic medium and small H, $\boldsymbol{G}_H$ is proportional to H and is defined by the Kirchhoff's law relating the current density $\boldsymbol{J}$ and the electric field $\boldsymbol{E}$,

$$\boldsymbol{J} = G\boldsymbol{E} + \boldsymbol{E} \times \boldsymbol{G}_H. \tag{70}$$

If Eq. (70) is inverted for small H, the Hall coefficient $R_H = G_H/HG^2$ appears along with the Ohmic resistivity $\rho = 1/G$: $\boldsymbol{E} = \rho\boldsymbol{J} + R_H(\boldsymbol{H} \times \boldsymbol{J})$. Cohen and Jortner (1973) developed an EMA for this problem. Note that, because of the presence of the magnetic field, the general circuit element of a bond of the network for calculating G_H and R_H is not a simple resistor but a conductance matrix. Straley (1980b) and Bergman *et al.* (1990) proposed network models for calculating G_H and R_H.

Study of the scaling behavior of G_H and R_H near p_c was initiated by Levinshtein *et al.* (1976), Shklovskii (1977), and Straley (1980a). However, a complete scaling theory was developed by Bergman and Stroud (1985). Consider a two-component network in which each bond has good Ohmic and Hall conductivities g_1 and g_{H1} with probability p, and poor Ohmic and Hall conductivities g_2 and g_{H2} with probability $1 - p$. Bergman and Stroud (1985) proposed that for $g_{H2}/g_{H1} \ll 1$ and for p close to p_c,

$$\frac{G_H - g_{H2}}{g_{H1} - g_{H2}} = |p - p_c|^{\tau} F\left(\frac{g_{H2}/g_{H1}}{|p - p_c|^{\mu+s}}\right), \tag{71}$$

where τ is a new critical exponent characterizing the scaling behavior of G_H near p_c when $p > p_c$, $g_2 = 0$, and $g_{H2} = 0$. Equation (71) has been written in complete analogy with Eq. (12). The scaling function $F(x)$ has the properties that $F(x) \simeq$ const. if $x \ll 1$ and $p > p_c$ (Regime I), $F(x) \sim x^2$ for $x \ll 1$ and $p < p_c$ (Regime II), and $F(x) \sim x^{\tau/(\mu+s)}$ for $x \gg 1$ and $p \simeq p_c$ (Regime III). However, a better way of understanding Eq. (71) is in terms of the Hall coefficient R_H. To each bond we assign a Hall coefficient $R_{H1} = g_{H1}/g_1^2$ with probability p, or $R_{H2} = g_{H2}/g_2^2$ with probability $1 - p$. Then one has (Bergman and Stroud, 1985)

$$R_H \sim a_1 R_{H1}|p - p_c|^{-\zeta} + b_1 R_{H2}(g_2/g_1)^2 \times |p - p_c|^{-2\mu}, \quad \text{Regime I;} \tag{72}$$

$$R_H \sim a_2 R_{H1}|p - p_c|^{-\zeta} + b_2 R_{H2}|p - p_c|^{2s}, \quad \text{Regime II;} \tag{73}$$

$$R_H \sim a_3 R_{H1}(g_2/g_1)^{-\zeta/(\mu+s)} + b_3 R_{H2}(g_2/g_1)^{2s/(\mu+s)}, \quad \text{Regime III;} \tag{74}$$

where a_i and b_i are constant, and ζ is a critical exponent that characterizes the divergence of R_H as p_c is approached from below and is related to μ and τ by

$$\zeta = 2\mu - \tau. \tag{75}$$

In 2D, $\zeta = 0$ exactly (Shklovskii, 1977), *i.e.*, the Hall coefficient remains finite at $p = p_c$, and in 3D, $\zeta \simeq 0.5$. The important point to remember is that in Regime I the ratio of the second to the first term is of order $(g_{H2}/g_{H1}) \times |p - p_c|^{-\tau}$, and therefore, depending on the parameters of the system, either term may dominate, so that experimental verification of these scaling laws is not straightforward.

Early experimental studies of Hall conductivity and coefficient in percolating composites were carried out by Levinshtein *et al.* (1976) and Sichel and Gittleman (1982). Experimental verification of Eqs. (72)–(74) was provided by Palevski *et al.* (1984), who showed that the Hall coefficient does indeed remain finite in thin (2D) metal films, and by Uri *et al.* (1987), who studied the problem in thick (3D) Al–Ge metal–insulator films. In Regime I, the poor conductor was Al–doped Ge, which was found to dominate the Hall effect. To make the first term of Eq. (72) dominant, Uri *et al.* dissolved the metallic aluminum in KOH, leaving the doped Ge as the good conductor, the poor-conductor role now being played by the vacuum. In such a situation, the first term of Eq. (72) is dominant, and therefore ζ and τ could be measured directly. The results were $\tau \simeq 3.8 \pm 0.2$ and $\zeta \simeq 0.38 \pm 0.05$, compatible with simulation results, $\tau \simeq 3.56 \pm 0.06$ and $\zeta \simeq 0.49 \pm 0.06$.

3. APPLICATION TO POROUS-MEDIA PROBLEMS

A major application of percolation theory has been characterization of the morphology of heterogeneous porous media, and the description of various flow phenomena therein. It is not possible to describe all the efforts that have been made in this direction, as the literature on the subject is too extensive (see, for example, Sahimi, 1993b, 1995). Thus, in what follows we only outline some of the major issues, and discuss applications of percolation theory for addressing them.

3.1 Characterization of Porous Media

An important quantity for characterization of a porous system is the capillary pressure P_c. If two immiscible fluids, *e.g.*, water and oil, are in contact in a bounded system, such as a cylindrical tube, then there is a discontinuity in the pressure field as one moves across the interface from one fluid phase to the other. That is, if we consider two points in the immediate vicinity of the interface, one in the water phase at pressure P_w and one in the oil phase at pressure P_o, then the pressure difference $P_c = P_o - P_w$, usually called the capillary pressure, is given by the Laplace equation

$$P_c = \sigma \cos\theta(1/r_1 + 1/r_2). \tag{76}$$

Here r_1 and r_2 are the principal radii of curvature of the interface, θ is the contact angle, and σ is the interfacial tension between the two fluids. For an interface in a cylindrical tube

$r_1 = r_2 = r$, where r is the tube's radius. In a cylindrical tube P_c depends uniquely on the amount of the fluid in the system. However, in a porous medium with many irregularly shaped pores different amounts of fluids, measured by their saturations (the volume fraction of the fluids), may yield the same P_c, and *vice versa*, since there are several interfaces in the pore space at different locations. For example, suppose that we initially fill our porous medium with water, and then expel it gradually by injecting oil into the pore space. For each value of the water saturation S_w we have a corresponding P_c. After continuing this for long enough, we reach a point where no more water is produced, although there is still some water left in the pore space. The saturation of water at this value of P_c is called the *irreducible* saturation S_{wi}, and the process of expelling water (more generally, a wetting fluid) by oil (more generally, a nonwetting fluid) is called *drainage*. Now, if we start with the porous system at the end of drainage (at the highest P_c) and expel the oil gradually by injecting water into the system, a process called *imbibition*, we obtain another curve of P_c *versus* S_w, very different from what we obtained during drainage, and thus there is hysteresis between the two curves. Figure 14 shows a typical P_c–saturation diagram during both drainage and imbibition, indicating that even when $P_c = 0$, there is still some oil left in the system. The saturation S_{or} of oil at this point is called the *residual* saturation. The all-important goal of enhanced oil-recovery methods is extracting this residual oil from the reservoir.

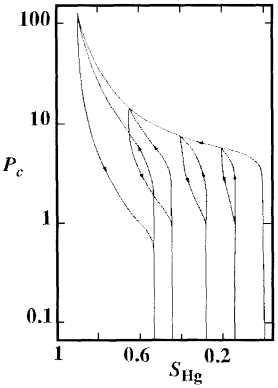

FIG. 14. Typical capillary pressure P_c *versus* mercury saturation S_{Hg} during mercury porosimetry.

Two important morphological characteristics of any porous medium are its pore size distribution (PSD) and pore connectivity. Careful examination of natural porous media reveals that what are usually referred to as pores are in fact either pore bodies, where most of the porosity resides, or pore throats, which are the narrow channels that connect the pore bodies and through which a fluid flows. One usually assigns effective radii to the pore bodies and throats that are nothing but the radii of spheres that have the same volumes as the pores and throats. In a network representation of the pore space, the pore bodies and throats are represented by the sites and bonds of the network, respectively. There are several experimental techniques for measuring the PSD of a porous medium, two of which, mercury porosimetry and adsorption–desorption experiments, require the concepts of percolation theory for proper interpretation of the data. Here we discuss mercury porosimetry. Adsorption techniques have also proven to be very valuable for measuring the connectivity of a pore space, and hence we discuss them in that context.

3.1.1 Mercury Porosimetry In this method, the porous medium is first evacuated and then immersed in mercury. The pressure is then increased (incrementally or continuously) and the volume of mercury injected into the porous medium is measured as a function of the applied (capillary) pressure. This corresponds to a drainage process, since mercury acts as a nonwetting fluid. Larger and larger pressures are needed to penetrate progressively smaller pores. In general if a nonwetting fluid (in this case mercury) is to displace a wetting fluid, it must overcome a capillary pressure P_c at the pore throat of radius r_t [see Eq. (76)],

$$P_c = 2\sigma / r_t. \tag{77}$$

Equation (77) means that the nonwetting phase penerates the largest throats (that require the smallest P_c) that it can reach. The pressure is then lowered back to atmospheric pressure to retract the mercury from the pores (imbibition), but with a characteristic shift or hysteresis between the injection and retraction curves. During imbibition, if the wetting phase is to displace the nonwetting phase in the pore body, the capillary pressure must be

$$P_c = 2\sigma / r_b, \tag{78}$$

implying that the wetting phase will occupy the smallest pore bodies. Some mercury stays in the medium even after the pressure is lowered back to atmospheric pressure.

While mercury porosimetry is a relatively straightforward experiment, the interpretation of the data is not simple. By measuring P_c and the mercury saturation, one tries to deduce the PSD of the pore space. The sequence of the pores that are penetrated by mercury during mercury porosimetry depends not only on the PSD, but also on the topology—the way the pores are connected to each other. Although at a given P_c a throat may be large enough for mercury penetration, a connection between the throat and the external surface of the pore space, *via* the already penetrated pore bodies and throats, has to be established, *i.e.*, that throat has to be accessible in the percolation sense described in Sec. 1.3. Therefore, mercury porosimetry represents a percolation phenomenon, and hence interpretation of porosimetry data and obtaining the PSD requires use of a network model of the pore space. However, it was only in 1977 that the connection between this phenomenon and percolation was recognized. Chatzis and Dullien (1977) and Larson and Morrow (1981) were among the first to recognize this connection. Many other workers have studied this problem, a list of whom is too long to be given here (for reviews see, for example, Sahimi, 1993b, 1995).

To give the reader some idea about percolation modelling of capillary pressure, we outline the model for mercury injection (drainage). Suppose that r_m is the minimum throat radius into which the mercury penetrates. Then, the fraction of the throats that can potentially be penetrated by mercury, *i.e.*, the allowed fraction, is given by

$$p(r_m) = \int_{r_m}^{\infty} \Psi(r)\, dr, \tag{79}$$

where $\Psi(r)$ is the PSD of the pore space. However, not all the allowed throats will be penetrated by mercury, but only a fraction $P_A(p(r_m))$ of them, where $P_A(p)$ is the percolation accessiblity function defined in Sec. 1.3. Therefore, the PSD $\Psi_i(r)$ of the throats that are actually penetrated by the injected mercury is

$$\Psi_i(r) = \Psi(r)/p(r_m), \quad r \geq r_m, \tag{80}$$

and clearly $\Psi_i(r) = 0$ for $r < r_m$. The saturation S_{Hg} of mercury during injection is given by

$$S_{Hg} = \frac{P_A(p(r_m))}{p(r_m)} \frac{\int_{r_m}^{\infty} \Psi(r) V(r)\, dr}{\int_0^{\infty} \Psi(r) V(r)\, dr}, \tag{81}$$

where $V(r)$ is the volume of a throat of radius r. Corresponding to any S_{Hg} there is a capillary pressure given by Eq. (77). Clearly, data on P_c and S_{Hg}, together with Eqs. (79)–(81), can be converted into a PSD for the pore space. This model has been used with great success for interpreting porosimetry data and obtaining the PSD of porous media; see the reviews by Sahimi (1993b, 1995) for detailed discussion.

3.1.2 Pore-space Connectivity and Sorption Phenomena One of the simplest concepts for characterizing the topology of a porous medium is the coordination number Z, loosely defined as the number of pore throats that meet at a pore body. For an irregular pore space one has to define an average coordination number $\langle Z \rangle$, with the average taken over a large enough sample. Although several experimental techniques have been developed for determining $\langle Z \rangle$, percolation theory has proven to be invaluable for interpreting adsorption–desorption data and estimating $\langle Z \rangle$ from them. Thus, let us first discuss typical adsorption–desorption experiments and why percolation theory is relevant to them. Liquid nitrogen is usually used in such experiments, and thus consider nitrogen adsorption in a single pore. As the pressure is increased an adsorbed film of nitrogen forms on the pore walls whose thickness increases with increasing pressure. At condensation pressure P_{co} the pore is filled with a (liquid-like) condensed phase, which results in a steep increase in the adsorption isotherm. The condensation pressure P_{co} is given by the Kelvin equation which, for a pore of radius r, is given by

$$\frac{P_{\mathrm{co}}}{P_0} = \exp\left[-\frac{2\sigma V_{\mathrm{L}}}{RT(r-l)}\right] \qquad (82)$$

in which P_0 is the saturation pressure, σ the liquid–vapor surface tension, R the gas constant, T the temperature, V_{L} the molar volume of the liquid, and l the thickness of the adsorbed layer present on the surface of the pore when condensation occurs. Equation (82) tells us that for any P_{co}/P_0 the adsorption process can be uniquely parameterized by an effective pore radius, which hereafter is referred to as r_{a}. Hence, adsorption or the reverse process of desorption corresponds, respectively, to an increase or decrease in r_{a}, with the change being caused by the disordered morphology of the pore space.

Now consider adsorption in a porous medium. During this process all the pores are equally accessible, vapor condenses in all pores of radius $r > r_{\mathrm{a}}$, and liquid nitrogen fills the pores. For $r < r_{\mathrm{a}}$ the pores fill rapidly and continuously with nitrogen. Thus, during adsorption connectivity of the pores plays no role. In contrast, desorption does depend on the pore-space connectivity. As the pressure is reduced, the desorption isotherm does not retrace that of adsorption but, similar to mercury porosimetry, forms a hysteresis loop before rejoining the adsorption isotherm. Two mechanisms contribute to this hysteresis, the first of which is thermodynamic in nature: a metastable phase may exist beyond the coexistence pressure during adsorption and/or desorption. This means that during adsorption a vapor phase may exist at pressures above P_{co}, or during desorption a liquid phase may exist below P_{co}. Thus, this is a pore-level mechanism and has nothing to do with pore connectivity, *i.e.*, percolation. The second mechanism is due to the connectivity of the pore space: a pore with an effective radius r is allowed to desorb (allowed to contain vapor) if $r > r_{\mathrm{a}}$ (so that at given pressure the liquid phase is metastable with respect to the vapor phase), *and* if it is connected to the bulk vapor in desorption. Expressing it in another way, a pore can desorb if it is accessible from the bulk vapor phase, and thus this is a percolation effect. Unless the PSD of the pore space is very narrow, the percolation effect is more important in the formation of the hysteresis loops than the thermodynamic mechanism. Typical adsorption–desorption isotherms are shown in Fig. 15.

Seaton (1991) (see also Liu *et al.*, 1992, for further elaboration) developed a percolation method for estimating $\langle Z \rangle$ on the basis of adsorption–desorption experiments. This method is based on finite-size scaling analysis, Eq. (23), according to which

$$P_{\mathrm{A}}(p) = L^{-\beta/\nu} f((p-p_c)L^{1/\nu}), \qquad (83)$$

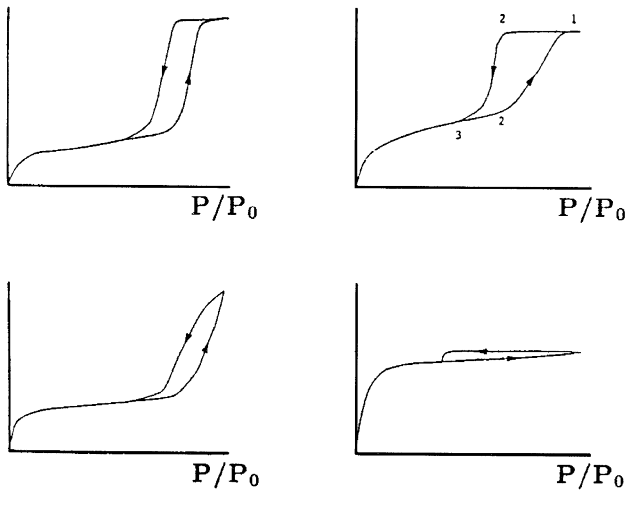

FIG. 15. Typical adsorption–desorption isotherms as a function of the pressure *P*. Segments 3–2–1 represent the adsorption isotherm, while 1–2–3 (decreasing *P*) indicates the desorption isotherm.

which can be rewritten as

$$\langle Z\rangle P_A(p) = L^{-\beta/\nu} f((\langle Z\rangle p - B_c)L^{1/\nu}), \qquad (84)$$

using $B_c = \langle Z\rangle p_{cb}$ (see Table 3). Thus, it remains to calculate $P_A(p)$ during the experiments. Consider, as an example, Fig. 15. The desorption curve has three segments indicated by 1, 2, and 3 in the figure. In the 1–2 interval, the isotherm is almost linear and occurs because of decompression of the liquid nitrogen in the pores. In the corresponding percolation network, the fraction p of the open pores (*i.e.*, those in which the nitrogen pressure is below P_{co}) increases, but $P_A(p)$ is still zero because a sample-spanning cluster of open pores has not yet formed. At point 2, the network reaches its p_c, a sample-spanning cluster of open pores is formed, and the metastable liquid nitrogen in the pores of the cluster vaporizes. If one decreases the pressure further, the number of pores containing metastable nitrogen and the number of pores whose nitrogen has vaporized both increase. At point 3, almost all the pores in which the nitrogen pressure is below their condensation pressure can also vaporize, and therefore $P_A(p) \simeq p$.

Thus a percolation method for estimating $\langle Z\rangle$ consists of two steps:

1. $P_A(p)$ is determined from the adsorption–desorption data; and
2. $\langle Z\rangle$ and L are determined by fitting Eq. (84) to $P_A(p)$.

Note that $P_A(p)/p$, the ratio of the number of pores in the sample-spanning cluster and the number of pores below their P_{co}, can also be written as N_p/N_b, where N_b is the number of moles of nitrogen that would desorb if all the pores containing nitrogen below its P_{co} had access to the vapor phase, and N_p is the number of moles of nitrogen that actually have desorbed at that pressure. Now, if N_A is the number of moles of nitrogen in the pores at a given pressure during the adsorption experiments, N_D is the number of moles of nitrogen in the pores at that pressure during the desorption experiment, and N_F is the number of moles of nitrogen that would have been present in the pores at that pressure during the desorption experiment if no nitrogen had vaporized from the pores that contain nitrogen below its P_{co}, then $N_p = N_F - N_D$, and $N_b = N_F - N_A$, and therefore

$$\frac{P_A(p)}{p} = \frac{N_F - N_D}{N_F - N_A}. \qquad (85)$$

The final step is to determine p, so that $P_A(p)$ can be calculated from Eq. (85). If $\Psi(r)$ is the PSD of the pore space, then for a given pressure one has

$$p = \int_r^\infty \Psi(x)\,\mathrm{d}x, \qquad (86)$$

where r is the pore radius in which nitrogen condenses at that pressure. Therefore, given $\Psi(r)$, p and hence $P_A(p)$ are determined. Using this method, the average connectivities of a variety of porous materials have been determined very accurately.

3.2 Single-phase Flow in Porous Media

Although the effective permeability K and conductivity G of porous media can be estimated by EMA and PSRG methods discussed above, these methods require very detailed information about the morphology of porous media. Katz and Thompson (1986, 1987) used percolation theory successfully for predicting K and G. Their method was based on a concept called critical-path analysis, first proposed by Ambegaokar *et al.* (AHL) in 1971. These authors argued that transport in a disordered medium with a broad distribution of heterogeneities is dominated by those conductances whose magnitudes are larger than some characteristic g_c, which is the smallest conductance such that the set of conductances $\{g|g > g_c\}$ forms a sample-spanning cluster called the critical path. Therefore, transport in a disordered medium with a broad conductance distribution reduces to a percolation problem with threshold conductance g_c. Shante (1977) and Kirkpatrick (1979) modified this idea by assigning g_c to all the local conductances with $g \geq g_c$, and setting all conductances with $g < g_c$ equal to zero (since the contribution of such bonds to the overall transport properties is negligible), and obtained a trial solution for G which has the same form as Eq. (8):

$$G = a g_c[p(g_c) - p_c]^\mu, \qquad (87)$$

where $p(g_c)$ is the probability that a given conductance $g \geq g_c$ and a is a constant. Equation (87) was then maximized with respect to g_c to obtain an estimate of g_c and thus G.

Katz and Thompson (1986, 1987) extended the ideas of AHL to estimate the permeability and electrical conductivity of porous media. In a porous medium the local hydraulic conductance is a function of a length scale l. Therefore there is a characteristic length scale l_c corresponding to g_c. The length that signals the percolation threshold in the flow problem also defines the threshold in the conduction problem. Thus we rewrite Eq. (87) as

$$G = \varphi g_c(l)[p(l) - p_c]^{\mu}, \tag{88}$$

where the porosity φ ensures a proper normalization of the fluid or the electric-charge density; $g_c(l)$ is equal to $c_f l^3$ for the flow problem and $c_e l$ for conduction. For appropriate choices of $p(l), G(l)$ achieves a maximum for some $l_{\max} \leq l_c$. In general $l^{\rm f}_{\max}$ for the flow problem is different from $l^{\rm c}_{\max}$ for the conduction problem, because the transport paths have different weights for the two problems. If $p(l)$ allows for a maximum in the conductance which occurs for $l_{\max} \leq l_c$, then for $l^{\rm f}_{\max} = l_c - \Delta l_{\rm f}$ and $l^{\rm c}_{\max} = l_c - \Delta l_c$, and $l^{\rm c}\mu p''(l_c)/p'(l_c) \ll 1$ (which is valid if the PSD is broad), we can write

$$\begin{aligned} l^{\rm f}_{\max} &= l_c\left[1 - \frac{\mu}{1 + \mu + l_c\mu p''(l_c)/p'(l_c)}\right] \\ &\simeq l_c\left(1 - \frac{\mu}{1+\mu}\right) \simeq \frac{1}{3}l_c, \end{aligned} \tag{89}$$

$$\begin{aligned} l^{\rm c}_{\max} &= l_c\left[1 - \frac{\mu}{3 + \mu + l_c\mu p''(l_c)/p'(l_c)}\right] \\ &\simeq l_c\left(1 - \frac{\mu}{3+\mu}\right) \simeq \frac{3}{5}l_c, \end{aligned} \tag{90}$$

where we used $\mu \simeq 2$ for 3D percolation. Thus, writing

$$G = a_1\varphi[p(l^{\rm c}_{\max}) - p_c]^{\mu} \tag{91}$$

and

$$K = a_2\varphi(l^{\rm f}_{\max})^2[p(l^{\rm f}_{\max}) - p_c]^{\mu}, \tag{92}$$

we obtain, to first order in Δl_c or in $\Delta l_{\rm f}$, $p(l^{\rm f,c}_{\max}) - p_c = -\Delta l_{\rm f,c}p'(l_c)$. To obtain a_1 and a_2, Katz and Thompson (1986) assumed that at the local level the conductivity of the porous medium is $g_{\rm f}$, the conductivity of the fluid (usually brine) that saturates the pore space, and that the local pore geometry is cylindrical, implying that $a_1 = g_{\rm f}$ and $a_2 = 1/32$. Therefore, one obtains

$$K = \frac{1}{226}\frac{G}{g_{\rm f}}l_c^2 \tag{93}$$

and (Katz and Thompson, 1987)

$$\frac{G}{g_{\rm f}} = \frac{l^{\rm c}_{\max}}{l_c}\varphi S(l^{c}_{\max}), \tag{94}$$

where $S(l^{\rm c}_{\max})$ is the volume fraction of connected pore space involving pore widths of size $l^{\rm c}_{\max}$ and larger. Note that it has been assumed that the permeability and conductivity exponents are equal, $\chi = \mu$, which is true if inequality (30) is satisfied.

Equations (93) and (94) involve no adjustable parameters. To obtain the characteristic length l_c, Katz and Thompson (1986, 1987) used mercury porosimetry discussed above, and proposed that $l_c = -(4\sigma\cos\theta)/P_c$, where σ is the surface tension between the mercury and the vacuum, θ the contact angle between the mercury and the pore surface, and P_c is the capillary pressure at the point when a sample-spanning cluster of mercury is formed in the pore space for the first time. Thus, a single set of porosimetry measurements yields l_c and hence K and G. Figure 16 compares the logarithm of the calculated K for a set of sandstones with the measured values. The dashed lines mark a factor of two; the agreement between the predictions and the data is very good. Equally impressive predictions were obtained for the electrical conductivity of the same porous media. This percolation method has also been extended for estimating the permeability of a porous medium saturated by a *non-Newtonian* fluid (Sahimi, 1993a), such as polymer solution.

3.3 Anomalous Diffusion in Porous Media

As mentioned in Section 1.6, diffusion in a porous medium at length scales $L < \xi_p$ is

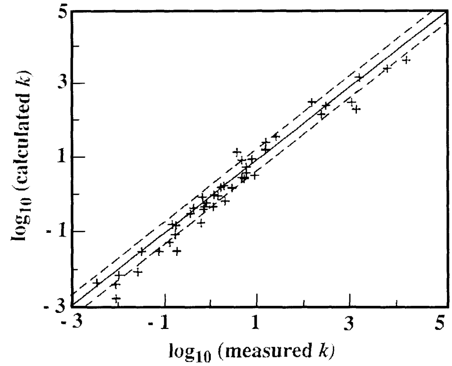

FIG. 16. Percolation predictions of the permeability k (in millidarcy) of various porous media *versus* the predicted values. Dashed lines denote a factor two of possible difference (after Katz and Thompson, 1986).

anomalous or fractal. A useful characteristic of diffusion is the mean square displacement $\langle R^2(t)\rangle$ of the diffusing particles, which is related to the effective diffusivity D by

$$\langle R^2(t)\rangle = 2dDt, \tag{95}$$

and therefore, depending on how $\langle R^2(t)\rangle$ grows with t, one can have three distinct diffusion regimes. Let us write

$$\langle R^2(t)\rangle \sim t^{2/D_W}, \tag{96}$$

where D_W is called the fractal dimension of the random walk or diffusion. In general, there are three possibilities.

1. $D_W = 2$, which means that $\langle R^2(t)\rangle$ grows linearly with t, and D is a constant. This happens when $R_d = \sqrt{\langle R^2(t)\rangle}$ is larger than the dominant length scale of the medium, which for percolation disorder is the correlation length ξ_p.
2. If $D_W > 2$, $\langle R^2(t)\rangle$ grows with t slower than linearly. In this case diffusion is called anomalous (Ben Avraham and Havlin, 1982; Gefen *et al.*, 1983) or fractal (Sahimi *et al.*, 1983a). In this regime $D \sim t^{2/D_W-1}$ vanishes as $t \to \infty$, *i.e.*, diffusion is very slow and inefficient. This happens when R_d is smaller than the dominant length scale of the medium, which for percolation disorder is the case if $R_d < \xi_p$. In general, diffusion in *all* fractal media is anomalous. Note that ξ_p is divergent at p_c, and therefore one always has fractal diffusion at p_c.
3. Finally, if $D_W < 2$, then $\langle R^2(t)\rangle$ grows with t faster than linearly. This is called *superdiffusion* (Sahimi, 1987).

Turbulent flow is an example of a transport process that gives rise to superdiffusion. We should point out that neither fractal diffusion nor superdiffusion is described by the classical diffusion equation, and the exact form of the governing equation is unknown.

In case **2** for $L < \xi_p$ (or equivalently at p_c), we can relate D_W to the critical exponents of percolation defined above. Suppose that diffusion occurs only in the sample-spanning cluster. Equations (14) and (20) tell us that near p_c, $D \sim \xi_p^{-\theta}$, where $\theta = (\mu - \beta)/\nu$. Because fractal diffusion occurs only if $R_d < \xi_p$, we can replace ξ_p with $R_d = \langle R^2(t)\rangle^{1/2}$ and write $D \sim \langle R^2(t)\rangle^{-\theta/2} \sim d\langle R^2(t)\rangle/dt$, which after integration yields $\langle R^2(t)\rangle \sim t^{2/(2+\theta)}$, implying that (Gefen *et al.*, 1983)

$$D_W = 2 + \frac{\mu - \beta}{\nu} = 2 + \theta. \tag{97}$$

Thus, we find that (see Table 3) $D_W(d = 2) \simeq 2.87$ and $D_W(d = 3) \simeq 3.8$. It can be shown (Sahimi *et al.*, 1983a) that EMA predicts that $D_W = 4$ for both $d = 2$ and $d = 3$. The crossover between fractal and Fickian diffusion occurs at a time t_{co} such that

$$t_{co} \sim \xi_p^{D_W} \sim (p - p_c)^{-2\nu+\beta-\mu}, \tag{98}$$

so that for $t < t_{co}$ diffusion is fractal, whereas for $t > t_{co}$ it is Fickian. Equation (98) tells that t_{co} diverges as $p \to p_c$, which is because as $p \to p_c$ the correlation length ξ_p diverges, and thus the percolation medium is a fractal object on a rapidly increasing length scale. If diffusion takes place on all the percolation clusters, then it can be shown (Gefen *et al.*, 1983) that

$$D_W = 2 + \frac{2\mu}{2\nu - \beta}. \tag{99}$$

Equations (96) and (97) were verified experimentally by Knackstedt *et al.* (1995), who obtained, with a model porous medium, $D_W \simeq 3.7 \pm 0.2$, consistent with Eq. (97).

3.4 Immiscible Displacements in Porous Media

We now turn our attention to percolation models of immiscible displacements and two-phase flows in porous media. A large number of factors affect this class of phenomena, among which are capillary, viscous, and gravity forces, the viscosities of the two fluids and the interfacial tension between them, chemical and physical properties of the surface of the pores, *i.e.*, whether there are surface active agents, or whether the surface is rough and fractal, the morphology of the pore space, and the wettability of the fluids. Thus, two-phase flow in a porous medium involves a set of complex phenomena, and to date no model has been developed that can take into account the effect of all the factors.

We already discussed, in the context of mercury porosimetry, a percolation model of drainage processes, *i.e.*, the displacement of a wetting fluid by a nonwetting one. In Section 3.5 we discuss such models in more detail. A more complex process is imbibition, by which a wetting fluid displaces a nonwetting fluid in a porous medium. The main difficulty is due to formation of thin films of the wetting fluids on the surface of the pores: even when the bulk of the pore is occupied by the nonwetting fluid, the presence of the wetting film on the pore surface makes the process very complex (Blunt and Scher, 1995; Hashemi *et al.*, 1998). Depending on the boundary conditions and a variety of other factors, one can have several types of imbibition processes, not all of which can be described by a percolation model. We describe here two of them that are described by percolation.

3.4.1 Spontaneous Imbibition The driving force for this process is capillary suction, because of which the smallest pore bodies that are next to the interface are always invaded. At any given time step, many pore-level interfaces advance in the porous medium. Usually, the displacement takes place at small but finite capillary numbers, the ratio of viscous and capillary forces defined by

$$Ca = \eta v/\sigma, \tag{100}$$

where η is the average viscosity and v is the average fluid velocity. However, Ca does not remain a constant, but varies over a range of values. When relatively large values of Ca are created, a transition zone develops in which there is a high saturation gradient. As the interface advances in the medium, two separate regions develop. One is in front of the transition zone where the saturation of the displaced fluid is high. The other moves behind the transition zone where the saturation of the displacing fluid is high. This region expands as the interface advances. The transition zone remains essentially the same throughout the displacement, except when the interface nears the end of the porous medium.

3.4.2 Quasistatic Imbibition A main difference between this process and spontaneous imbibition is that in this process at any given step only one pore is invaded by the displacing fluid. This can be achieved by adjusting the backpressure so that the *narrowest* pore throat is invaded, while the interface at other larger throats remains essentially motionless. Since even the largest pore throats are smaller than the pore body to which they are connected, once a pore body is invaded, so also are all the throats that are connected to it. As soon as the interface enters such throats, the smallest pore body that is connected to them is invaded, and so on. Thus at any given step of the displacement the smallest pore body that is accessible from the surface through a continuous path of the displacing fluid is invaded. This implies that a percolation model is applicable to this phenomenon.

A breakthough is achieved when the displacing fluid forms a sample-spanning cluster. Just before the breakthrough the displaced fluid is mostly connected. However, as the displacement proceeds small blobs of the displaced fluid are formed and get trapped. At the end of this process, one ends up with a large number of isolated blobs whose volume fraction in unconsolidated media varies between 0.14 and 0.2 and in consolidated porous media between 0.4 and 0.8. Experimental data (see, for example, Chatzis *et al.*, 1983) indicate that the size distribution of the blobs, when expressed in terms of the number of pore bodies they occupy (remember that most of the porosity and volume of a porous medium are provided by its pore bodies), follows Eq. (15), confirming the applicability of a percolation model.

3.5 Percolation Models of Two-phase Flow in Porous Media

In this section we discuss percolation models of two-phase flow and displacement in porous media, which, strictly speaking, are applicable only when *Ca* is small, although their range of applicability is often larger. At the outset we should point out that a fundamental assumption in all percolation models of two-phase flow is that the (bond or site) occupation probability p is proportional to the capillary pressure. Without such an assumption, it would be difficult to make a one-to-one correspondence between a percolation model and the two-phase flow problem.

3.5.1 Random Percolation Models The first random percolation model of two-phase flow in porous media was proposed by Larson *et al.* (1977, 1981a,b). To model drainage, the porous medium was represented by a cubic network. It was assumed that a bond is penetrated by the displacing fluid if the P_c at that point exceeds a critical value, implying that the bond's radius $r \geq r_m$, where r_m is defined by Eq. (79). All the bonds that are connected to the displacing fluid by a path of throats with $r \geq r_m$ were considered accessible, the accessibility being defined in the sense of percolation discussed in Sec. 1.3. It was also assumed that all the accessible throats with $r \geq r_m$ are filled by the nonwetting fluid (in reality this is not true, as the interface has to travel along a certain path before it reaches an accessible throat; see below). The displaced fluid was assumed to be compressible, so that it could not be trapped by the displacing fluid. Larson *et al.* (1981b) also proposed a percolation model of imbibition for calculating the residual nonwetting phase saturation S_{nwr} and its dependence on *Ca*. They modelled the creation of isolated blobs of the nonwetting phase by random site percolation. At p_{cs} they calculated the fraction $\hat{a}(s)$ of the active sites that are in clusters of length s in the direction of flow, and argued that this represents the desired blob size distribution. To calculate S_{nwr}, they assumed that once a blob is mobilized, it is permanently displaced (which is not always the case, because a blob can get trapped again, or join another blob and create a larger one).

Despite their shortcomings, the predictions of percolation models of Larson *et al.* were in qualitative agreement with experimental data. The fundamental assumption behind Larson *et al.*'s models is that pore-level events are controlled by capillary forces. Simple scaling arguments can be used for estimating the values of *Ca* for which this assumption is valid. It turns out that the ratio of viscous pressure drop and the capillary pressure is given by

$$P_{vis}/P_c \sim Ca/K_d, \tag{101}$$

where $K_d = K/l^2$ is a small dimensionless permeability (of order 10^{-3} or smaller) and l is the typical grain size. Thus, for capillary-controlled displacements, one must have $Ca \ll 1$ (typically, $Ca \sim 10^{-5}$–10^{-6}); experimental data do support this conclusion.

Heiba *et al.* (1982, 1983, 1984, 1992) further developed these ideas by distinguishing between pore throats that are allowed to a phase, and those that are actually occupied by the phase (*i.e.*, those that are accessible to it). Then, given the PSD of the pore space, they calculated the PSD of the allowed and occupied pores. For example, during drainage the PSD of the throats occupied by the nonwetting fluid is given by Eq. (80). Such PSDs are essential for calculating the permeability of each fluid phase and a related quantity, the *relative* permeability. It can be shown that if two immiscible fluids β and γ flow in a porous medium, then the governing equations for the average flow velocities and saturations of the fluids, for $Ca \ll 1$, are given by

$$\mathbf{v}_\beta = -\frac{\boldsymbol{K}_\beta}{\eta_\beta} \cdot (\boldsymbol{\nabla} \langle P_\beta \rangle^\beta - \rho_\beta g) + \boldsymbol{K}_{\beta\gamma} \cdot \mathbf{v}_\gamma, \tag{102}$$

$$\varphi \partial S_\beta / \partial t + \boldsymbol{\nabla} \cdot \mathbf{v}_\beta = 0, \tag{103}$$

with similar equations for fluid γ. Here $\boldsymbol{K}_\beta$, P_β, and ρ_β are respectively the permeability, pressure, and density of fluid β, and g is the acceleration due to gravity. The first term of (102) is the usual Darcy's law, while the second term is a coupling between the two phases, with $\boldsymbol{K}_{\beta\gamma} = \boldsymbol{K}_{\gamma\beta}$. Generally speaking, $\boldsymbol{K}_{\beta\gamma}$ is small unless $\eta_\beta \simeq \eta_\gamma$ (when a thin film of one phase coats the walls of a pore whose bulk volume is filled with the other phase). In practice one writes

$$\boldsymbol{K}_\beta = \boldsymbol{K} k_{\beta r}, \tag{104}$$

with a similar equation for the γ-phase, where $k_{\beta r}$ is called the relative permeability (RP) to the β phase, and $\boldsymbol{K}$ is the absolute permeability tensor of the porous medium. A major problem in multiphase flows in porous media is the prediction of the RPs. Unlike the absolute permeability $\boldsymbol{K}$, $k_{\beta r}$ depends on many parameters, including saturation and saturation histories of the fluids, pore-space morphology, the wetting characteristics of the fluids, the viscosity ratio, and the capillary number *Ca*. Experiments show that (see below) the RP to a phase typically becomes very small when its saturation is less than a critical value, which is distinctly larger than zero. This is, of course, the signature of a percolation problem.

If we use the PSDs of the pores and throats occupied by each fluid during drainage and imbibition (see Sec. 3.1.1), calculating the phase and relative permeabilities becomes a percolation problem. Therefore, EMA, PSRG, or REMA can be used for calculating the phase permeabilities and hence the RPs (Heiba *et al.*, 1982, 1983, 1984, 1992; Sahimi *et al.*, 1986). Figure 17 shows the results obtained with a cubic network, which are in excellent agreement with experimental data.

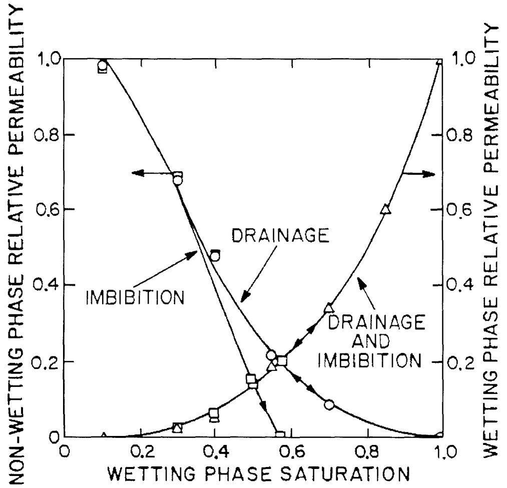

FIG. 17. Two-phase relative permeabilities as predicted by the random percolation model, using a simple cubic network model of porous media. The individual drainage and imbibition curves represent the relative permeabilities to the nonwetting phase during these processes, while the curve marked drainage and imbibition gives the relative permeability to the wetting phase during both processes (*i.e.*, there is no hysteresis in the wetting phase).

3.5.2 Invasion Percolation This model was first proposed by Lenormand and Bories (1980) and Chandler *et al.* (1982). In this model the network is initially filled with a fluid called the defender. To each site of the network is assigned a random number uniformly distributed in (0,1]. Then the displacing fluid (the invader) is injected into the medium and displaces the defender, at each time step, by choosing the site next to the interface that has the smallest random number. If we interpret the random numbers as the resistance that the sites offer to the invading fluid, then choosing the site with the smallest random number is equivalent to selecting a pore with the largest size, and hence this model simulates a drainage process. A slightly more tedious procedure can be used for working with bonds instead of sites. Two versions of the model have been developed. In one model the defender is incompressible, and therefore if its blobs are surrounded by the invader they become trapped (Chandler *et al.*, 1982; Wilkinson and Willemsen, 1983). In the second model, trapping is ignored, i.e., the defender is infinitely compressible (Wilkinson and Barsony, 1984).

There is a close connection between invasion percolation without trapping and random percolation (Wilkinson and Barsony, 1984). Define an *acceptance profile* $a_n(r)$ such that $a_n(r)\,\mathrm{d}r$ is the probability that random number r, selected at the nth step of the invasion, is in the interval $[r, r+\mathrm{d}r]$. Then as $n \to \infty$ one has

$$a_\infty(r) = 1/p_c, \quad r < p_c, \tag{105}$$

and $a_\infty = 0$ for $r > p_c$, where p_c is the percolation threshold. In practice a cumulative acceptance profile $b_n(r)$ is computed by

$$b_n(r) = (n + \langle n_b \rangle_n - 1)^{-1} \sum_{m=1}^{n} a_m(r), \tag{106}$$

where $\langle n_b \rangle_n$ is the average number of sites on the boundary immediately before the site is selected. In the limit $n \to \infty$ one has

$$b_n(r) = 1, \quad r < p_c, \tag{107}$$

and $b_n(r) = 0$ for $r > p_c$. Wilkinson and Barsony (1984) also defined two other quantities

$$B_1(n) = \int_0^{p_c} [1 - b_n(r)]\, \mathrm{d}r, \quad B_2(n) = \int_{p_c}^{1} b_n(r)\, \mathrm{d}r, \tag{108}$$

and postulated that $B_1(n) \sim B_2(n) \sim n^{-1/\Delta}$, where $\Delta = \nu D_F$ is called the gap exponent of percolation and D_F is the fractal dimension of the sample-spanning percolation cluster. Their Monte-Carlo simulations indicated that Δ is the same for invasion and random percolation.

For invasion percolation with trapping, Monte-Carlo simulations by Wilkinson and Willemsen (1983) indicated that the fractal dimension D_F of the sample-spanning cluster is $D_F(d = 2) \simeq 1.82$, somewhat smaller than $D_F(d = 2) = 91/48 \simeq 1.896$ for random percolation (see Table 3). This was confirmed by the experiments of Lenormand and Zarcone (1985) who displaced oil (the wetting fluid) by air (the nonwetting fluid) in a large and transparent 2D etched network and obtained $D_F \simeq 1.82$, consistent with 2D invasion percolation with trapping. Moreover, Shaw (1987) showed that if a porous medium, filled with water, is dried by hot air, the dried pores (*i.e.*, those filled with air) form a percolation cluster with the same D_F as that of invasion percolation. In 3D, recent results (Porto *et al.*, 1997) indicate that there may be significant differences between random percolation and invasion percolation with trapping, but a definitive picture has not yet emerged. A large number of workers have used an invasion percolation model or its variants to simulate two-phase flow in porous media; for comprehensive reviews see Sahimi (1993b, 1995).

3.5.3 Gradient Percolation So far, we have neglected the effect of gravity on an immiscible displacement. However, this effect cannot be neglected in a real porous medium, as the hydrostatic component of pressure adds to the applied pressure, and this creates a vertical gradient in the effective injection pressure. Because of this gradient, the fraction of the accessible pores decreases with the height of the system. A modification of the invasion percolation by Wilkinson (1984), by Sapoval *et al.* (1985) and by Gouyet *et al.* (1988) succeeded in taking into account such effects. The competition between gravity and capillary forces is usually quantified by the Bond number Bo defined as

$$Bo = gl^2 \Delta\rho/\sigma, \tag{109}$$

where $\Delta\rho$ is the density difference between the two fluids, g the acceleration due to gravity, and l the typical size of the grains or pores of the porous medium. Wilkinson (1984) showed that in an immiscible displacement under gravity, the correlation length ξ_g does *not* diverge (unlike random and invasion percolation which have a diverging correlation length ξ_p), but reaches a maximum value given by

$$\xi_g \sim Bo^{-\nu/(1+\nu)}, \tag{110}$$

where ν is the critical exponent of percolation correlation length, so that $\xi_g \sim Bo^{-0.47}$ and $\xi_g \sim Bo^{-4/7}$ in 3D and 2D, respectively. In 3D there is a transition region where both fluid phases can span the system (percolate), and the width w of this region is given by

$$w \sim Bo^{-1}. \tag{111}$$

Similar results were obtained by Sapoval *et al.* (1985) and Gouyet *et al.* (1988) in the context of *gradient percolation*, which is a model in which a gradient $\mathcal{G}$ for the occupation probability p is imposed on one direction of the network. They used arguments similar to Wilkinson's to show that

$$\xi_g \sim \mathcal{G}^{-\nu/(1+\nu)}, \tag{112}$$

which is completely similar to (110), in which Bo has been replaced with $\mathcal{G}$. The experiments of Hulin *et al.* (1988) and Birovljev *et al.* (1991) were completely consistent with these results. For example, Birovljev *et al.* (1991) used transparent 2D models consisting of a monolayer of 1-mm glass beads placed at random and sandwiched between two plates. The system was filled with a glycerol–water mixture, which was displaced by air invading the system at one end. They obtained $\xi_g \sim Bo^{-0.57}$, where the exponent 0.57 agrees with the prediction $\nu/(1 + \nu) = 4/7 \simeq 0.57$.

3.5.4 Crossover from Fractal to Compact Displacement Although we discussed the relative permeabilities for both imbibition and drainage in terms of a percolation model, there are certain qualitative differences between the two that need to be discussed. A clue to these differences is already evident in Fig. 17: the RP to the nonwetting phase during primary imbibition by a strongly wetting fluid vanishes only at zero residual nonwetting saturation, *i.e.*, the nonwetting phase is completely expelled from the medium and the wetting phase fills the system, indicating that imbibition is an essentially compact displacement. However, during drainage the RP to the wetting phase vanishes at a finite value of residual water saturation, *i.e.*, the nonwetting phase does not fill the porous medium, and a fractal percolation cluster is formed. This was nicely demonstrated by Lenormand and Zarcone (1984), who used a 2D etched network, injected mercury into the system (drainage) and then withdrew it (imbibition). The cluster formed during imbibition was compact and filled the network.

The crossover between the two types of displacement was clarified in a definitive study by Cieplak and Robbins (1988, 1990). In their study the porous medium was represented by a 2D array of disks with random radii, where the underlying lattice was either a triangular or a square network. The limit of low *Ca* was considered, and the displacement dynamics was modeled as a stepwise process where each unstable section of the interface moved to the next stable or nearly stable configuration. Trapping of the defending fluid was also allowed. If all the disks have the same radius, the resulting patterns are very regular and faceted which also preserve the symmetry of the lattice, consistent with experimental observations. However, when the radii of the disks are randomly distributed, then the behavior of the system depends on the contact angle θ [Eq. (76)]. To quantify the effect of θ, we define an interface width w. When θ is near 180°, *i.e.*, one has a drainage process, then the phenomenon is an invasion percolation and w is of the order of pore size. However, Cieplak and Robbins (1988, 1990) showed that as θ decreases, the cluster of the invading fluid becomes more compact, and w increases and finally diverges at a critical contact angle θ_c according to a power law

$$w \sim (\theta - \theta_c)^{-\nu_\theta}, \qquad (113)$$

where $\nu_\theta \simeq 2.3$ in 2D. The critical angle θ_c was found to depend on the porosity φ of the system; for example, $\theta_c \simeq 29°$ for $\varphi = 0.322$ and $\theta_c \simeq 69°$ for $\varphi = 0.73$. The exponent ν_θ was found to be universal. The compactness of the cluster for $\theta < \theta_c$ is consistent with the imbibition picture discussed above.

4. APPLICATION TO HOPPING CONDUCTIVITY OF AMORPHOUS SEMICONDUCTORS

Miller and Abrahams (1960) were probably the first to study hopping conductivity of amorphous semiconductors in the framework of a random resistor network. Over a decade after their work, it was realized that hopping conduction in semiconductors can be modeled successfully by percolation theory. This section summarizes the important elements of this successful application of percolation to a technologically important problem. Let us first discuss how the Miller–Abrahams network model was constructed, and then outline the percolation approach to this problem. The starting point in the Miller–Abrahams work was the Boltzmann equation

$$\frac{\partial P_i}{\partial t} = \sum_j [W_{ji}P_j(1 - P_i) - W_{ij}P_i(1 - P_j)], \qquad (114)$$

where P_i is the probability that site i of the network is occupied. It is assumed that the P_i's are uncorrelated. If repulsion can cause strong correlations, then the exclusion factor $1 - P_i$ is omitted. In the linear (Ohmic) regime, the current is proportional to the applied field, and we linearize Eq. (114) by writing $P_i = P_i^0 + \Delta P_i$ and $W_{ij} = W_{ij}^0 + \Delta W_{ij}$, where 0 denotes the equilibrium value and Δ an increment proportional to the applied field. Since $\Delta W_{ij} = -\Delta W_{ji}$, we obtain the linearized version of Eq. (114),

$$\frac{\partial \Delta P_i}{\partial t} + \sum_j A_{ij}\Delta P_i - \sum_j A_{ji}\Delta P_j = \sum_j B_{ji}\Delta W_{ji}, \qquad (115)$$

where $A_{ij} = W_{ij}^0(1 - P_j^0) + W_{ji}P_j^0$ and $B_{ij} =$

$P_i^0(1-P_j^0)+P_j^0(1-P_i^0)$. Equation (115) is a set of linear equations for the unknowns ΔP_i. The equilibrium values P_i^0 are given by the Fermi distribution

$$P_i^0 = [\exp(E_i/k_\mathrm{B}T)+1]^{-1}, \tag{116}$$

where E_i is the energy of a carrier on site i, as measured from the Fermi level, and k_B is the Boltzmann constant. The equilibrium values W_{ij}^0 are given by

$$W_{ij}^0 = u_{ij}/|\exp[(E_j-E_i)/k_\mathrm{B}T]-1|, \tag{117}$$

with

$$u_{ij}=u_{ji}=\exp(-2r_{ij}/a)/\tau_0. \tag{118}$$

Here $1/\tau_0$ is of the order of a phonon frequency, r_{ij} is the distance between i and j, and a is a Bohr radius. It is assumed that τ_0 depends only weakly on r_{ij} and the temperature T.

Suppose that $\boldsymbol{F}$ is the intensity of an applied field, and $\boldsymbol{r}_i$ is the radius vector of site i. The applied field changes the difference $\Delta_{ij}=E_i-E_j$ between the energies of sites i and j. We define a new variable V_i by

$$P_i = P_i^0+\Delta P_i \equiv |\exp[(E_i-eV_i)/k_\mathrm{B}T]+1|^{-1}, \tag{119}$$

where e is the charge of an electron. In the linear regime V_i is proportional to $\boldsymbol{F}$. Thus, one can transform the set of linear equations for P_i to one for V_i given by

$$\frac{P_i^0e^2(1-P_i^0)}{k_\mathrm{B}T}\frac{\partial V_i}{\partial t} = \sum_j[(V_i-\boldsymbol{F}\cdot\boldsymbol{r}_i)-(V_j-\boldsymbol{F}\cdot\boldsymbol{r}_j)]g_{ij}. \tag{120}$$

Equation (120) represents the Miller–Abrahams network which, at steady state, represents a resistor network if we think of $V_i-\boldsymbol{F}\cdot\boldsymbol{r}_i$ as the potential at site i. Then $z_{ij}=1/g_{ij}$ is the resistance (more generally, an impedance) between i and j, and Eq. (120) is simply Kirchhoff's law for site j. For the unsteady-state case, we define a capacitance $C=P_i^0e^2(1-P_i^0)/k_\mathrm{B}T$ with a potential V_i across it, and refer all the potentials to "ground" potential which is zero. Because $\boldsymbol{F}\cdot\boldsymbol{r}_i$ is the applied potential at i, it is represented as an output from a generator connected in series with C between the ground and site i. The impedance z_{ij} connects two junctions i and j, and the capacitor C_i is in series with a generator connected to the ground. Restricting our attention to the case when various site energies are of the order or larger than $k_\mathrm{B}T$, we obtain

$$z_{ij} = k_\mathrm{B}T\frac{\exp[(|E_i|+|E_j|+|E_i-E_j|)/2k_\mathrm{B}T]}{e^2u_{ij}}$$

$$= \frac{k_\mathrm{B}T}{e^2}\exp\left(\frac{E_{ij}}{k_\mathrm{B}T}+\frac{2r_{ij}}{a}\right)\tau_0, \tag{121}$$

$$C_i = \frac{e^2}{k_\mathrm{B}T}\exp\left(-\frac{E_i}{k_\mathrm{B}T}\right), \tag{122}$$

where either E_{ij} is the energy of the site farther from the Fermi energy, or $E_{ij}=(|E_i|+|E_j|+|E_i-E_j|)/2$. Equation (121) has an important implication: even if E_i is moderately distributed, the exponential dependence of z_{ij} and C_i on the energies makes them enormously broadly distributed quantities. This fact is used to reduce the computations of the effective properties of the network, since the broadness of the distribution of z_{ij} implies that there are many small conductances that can be removed from the network. The resulting network is called the *reduced* network, and the procedure for constructing it is discussed in detail by Pollak (1978).

In the Miller–Abraham model the hopping is between the nearest-neighbor sites. Mott (1968) pointed out that, if the activation energy of a nearest-neighbour site is large, a hop to a distant site whose energy is lower may be easier than one to a nearest-neighbour site. The range l_h of the hopper depends on the ease of activation to higher energies, and therefore l_h (and thus the resistance) depends on the temperature. This mechanism of hopping conduction, usually called *variable-range* hopping, is believed to be operating at low temperatures, whereas nearest-neighbour hopping occurs at high temperatures. Moreover, Mott showed that at low temperatures

$$G = G_0\exp[-(T_0/T)^\alpha]. \tag{123}$$

In general, α depends on the density of states near the Fermi level (see below). In Mott's theory, $\alpha = 1/(d+1)$ for a d-dimensional system, and G_0 and T_0 are some constants with

$$T_0 = \frac{\lambda a^3}{k_B N(E_F)}, \tag{124}$$

where $N(E_F)$ is the density of states at the Fermi level E_F, and λ is a dimensionless parameter. Equation (123) is particularly accurate for amorphous germanium in the range $60\,\mathrm{K} \leq T \leq 300\,\mathrm{K}$, with $T_0 \simeq 7 \times 10^7\,\mathrm{K}$. Similar results have also been found for amorphous silicon and carbon and vanadium oxide.

4.1 Percolation Models of Hopping Conductivity

Ambegaokar *et al.* (AHL) (1971), Shklovskii and Efros (1971), Brenig *et al.* (1971), and Pollak (1972) proposed that hopping conductivity of amorphous semiconductors can be modelled by percolation theory. As already mentioned in Sec. 3.2, AHL argued that an accurate estimate of G is the critical percolation conductance g_c, which is the largest value of the conductance such that the subset of the network with $g_{ij} > g_c$ still contains a conducting sample-spanning cluster. Thus, the network is divided into three parts:

1. a set of isolated clusters of high conductivity, each cluster consisting of a group of sites connected together by conductances $g_{ij} \gg g_c$;
2. a small number of resistors with g_{ij} of order g_c, which connect together a subset of high-conductance clusters to form the sample-spanning cluster, called the critical subnetwork, which is essentially the static limit of reduced network discussed above and dominates the overall conductance of the network;
3. the remaining resistors with $g_{ij} \ll g_c$.

Thus, the condition that $g_{ij} > g_c$, together with Eq. (121), is expressed as

$$\frac{r_{ij}}{r_m} + \frac{|E_i| + |E_j| + |E_i - E_j|}{2E_m} \leq 1, \tag{125}$$

where $r_m = a\ln(G_0/g_c)/2$ is the maximum distance between any two sites between which a hop can occur, $E_m = k_B T \ln(G_0/g_c)$ is the maximum energy that any initial or final state can have, and r_{ij} is the distance between nodes i and j.

To construct the critical subnetwork, AHL started with the smallest resistors and inserted them in an empty network one by one. As more resistors are inserted, clusters of connected resistors are formed until the critical value $z_c = 1/g_c$ is reached at which a sample-spanning cluster is formed. To calculate z_c, AHL used Eq. (121) and assumed that τ_0 is constant and that $N(E)$ is constant near the Fermi level, $N(E) = N(E_F)$. Thus, z_{ij} is a monotonic function of the random variable $\zeta = E_{ij}/k_B T + 2r_{ij}/a$, and z_c defines a corresponding critical value ζ_c. Around each site i such that $E_i < \zeta k_B T/2$, a sphere of radius

$$r_i = \frac{a}{2}\left(\frac{\zeta}{2} - \frac{E_i}{k_B T}\right), \tag{126}$$

is drawn. The radius of the sphere increases with ζ. When two spheres overlap, a bond is inserted between the two sites with overlapping spheres, which happens only if inequality (125) is satisfied. Percolation occurs at ζ_c, corresponding to a critical radius r_c. Equation (126) tells us that there is a maximum radius $r_m = a\zeta_c/4$ and a maximum energy $E_m = \zeta_c k_B T/2$, with $\zeta_c = 2\ln(G_0/g_c)$. Thus, if two sites are separated by a distance larger than r_m, or farther from energy level E_m, they will not contribute significantly to G. The volume of the sphere defined by Eq. (126) is $(\pi/6)a^3(\zeta_c/2 - E_i/k_B T)^3$, and the volume $\langle V\rangle$ averaged over all sites with a sphere with a non-zero radius (*i.e.*, those for which $E_i < \zeta k_B T/2$) is given by

$$\langle V\rangle = \frac{\pi}{384}a^3\zeta_c^3 = \frac{\pi}{48}r_m^3. \tag{127}$$

The volume fraction ϕ_c of the spheres at the percolation threshold is given by $\phi_c \simeq n\langle V\rangle$, where $n = N\zeta_c k_B T$ is the fraction of the sites with a sphere, *i.e.*, those with an energy in the interval $(-E_m, E_m)$. AHL estimated that $\phi_c \simeq 1/4$, somewhat larger than $\phi_c \simeq 0.15 - 0.17$ for 3D continuum percolation. One can also use the number of bonds per sites $B_c \simeq d/(d-1)$ (see Sec. 1.8). For 3D amorphous materials (or continuum percolation) $B_c =$

$4\pi n r_c^3/3$ and is related to $N(E)$ by

$$B_c = \frac{4\pi n}{3} \frac{\int_{-E_m}^{+E_m} N(E_i)\,dE_i \int_{-Em}^{+E_m} (r_m^3 + 3r_m^2 D_m + 3r_m D_m^2) N(E_j)\,dE_j}{\int_{-E_m}^{+E_m} N(E_i)\,dE_i}, \tag{128}$$

where D_m is the mean size of the sites. Combining $\phi_c = n\langle V\rangle \simeq 1/4$ with Eq. (127), or using Eq. (128) with the lattice or continuum value of B_c and the appropriate expression for r_m, one finally obtains Eqs. (123) and (124) with $\alpha = 1/4$ and $\lambda = 16$.

Equation (123) gives a lower bound to the true hopping conductivity of a network whose individual conductances vary over a broad range, because the critical subnetwork corresponds to replacing all $g_{ij} < g_c$ by 0 and all $g_{ij} \geq g_c$ by g_c in the original network. Equation (123) is exact only in the limit $T \to 0$. If $T > 0$, hops with conductance $g < g_c$ also contribute to the macroscopic conductivity, which means that the optimal cutoff should be somewhat larger than g_c. Moreover, the percolation approach of AHL cannot be used for 1D or quasi-1D conductors, since percolation disorder divides a linear chain into finite segments and the problem becomes meaningless. An example is a class of compounds which consist of weakly coupled parallel chains of strongly coupled atoms or molecules. The conductivity of such composites is highly anisotropic, and therefore they may be treated as essentially 1D conductors. Well-known examples of such systems are salts of the organic ion-radical tetracyanoquinodimethane (TCNQ) and the square planar complexes of transition metals such as platinum and iridium. Shante (1977) proposed a modification of the percolation model of AHL that took into account the effect of such complexities. His model was a bundle of chains in which hopping could occur along the chains and between them. However, interchain hopping was assumed to be much more difficult than intrachain hopping, but was still allowed. Shante's model also allowed the possibility of such intrachain hoppings in either 1D or 2D. At low temperatures Shante's model corresponded to 2D and 3D percolation, respectively, and consequently $T^{-1/3}$ and $T^{-1/4}$ behaviors were obtained (when ln G was plotted versus T). At high temperatures the percolation model is no longer applicable, and Shante obtained a T^{-1} behavior. The predictions were in excellent agreement with data. A completely different verification of Eq. (123) was provided by Ball *et al.* (1994), who studied the metal–insulator transition in excimer-laser–irradiated polyimide. The hopping conductivity of these composites was found to follow Eq. (123) with $\alpha = 1/4$. The electrical conductivity of the composite near its percolation threshold was found to follow Eq. (8) with $\mu \simeq 2 \pm 0.05$, completely consistent with 3D percolation (see Table 3). The pre-exponential factor G_0 has also been calculated by a percolation model. For example, Shklovskii and Efros (1975) proposed that $1/G_0 = r_c(2r_c/a)^{\nu} R_0$, where ν is the correlation length exponent of 3D percolation and R_0 is the resistance for $\zeta = \zeta_c$.

Finally, we should mention that hopping conductivity of disordered solids has also been studied by an EMA. For nearest-neighbor hopping the predictions of the EMA (see, for example, Odagaki and Lax, 1981; Sahimi *et al.*, 1983a) and the percolation model agree with each other, as well as with the experimental data. For variable-range hopping conductivity, Movaghar *et al.* (1980) showed that Mott's law, Eq. (123), can be predicted by a modified EMA. More recent works on this subject are those of Dyre (1994), Bleibaum *et al.* (1996) and Parris (1998).

4.2 Effect of a Variable Density of States

In the above percolation model, it was assumed that $N(E)$ is constant near the Fermi level. Although the basic $T^{-1/4}$ law has been observed in many semiconductors, criticism of this model was raised by many (see, *e.g.*, Szpilka and Viscor 1982), mainly because the predicted and measured G_0 differed by several orders of magnitude. This large difference was attributed to the assumption of a constant $N(E)$. Ortuno and Pollak (1983) proposed that if $N(E)$ is concave, then a modified percolation model can predict G_0 quantitatively. They used Eq. (128) with $N(E) = N(E_F)\exp(E/E_0)$, where E_0 is the exponential decay rate, and found good agreement between the predictions and the data for amorphous Si and Ge. Maloufi *et al.* (1988) used Ortuno and Pollak's theory to fit their conductivity data for amorphous Si_ySn_{1-y} with $y = 0.47$–1, and found excellent agreement over much of the temperature range. Moreover, if we assume that $N(E) \sim |E|^{\varpi}$, where ϖ is a positive constant, we find $\alpha = (\varpi + 1)/(\varpi + 4)$, which reduces to Eq. (123) when $\varpi = 0$. The limit $\varpi \to \infty$ is also interesting because it corresponds to a system that has a sudden onset of states away from the Fermi level, and we thus obtain a T^{-1} behavior for such systems. An experimental realization of a power-law density of states was provided by Redfield (1973), who carried out a careful study of hopping conductivity of heavily doped and strongly compensated GaAs; see also Sheng and Klafter (1983) and Mehbod *et al.* (1987) for related experiments.

4.3 Effect of Coulomb Interactions

Despite its success, the origin of a power-law density of states was not clear for some time until Efros and Shklovskii (1975) suggested that Coulomb interactions between localized electrons can create a soft gap (called Coulomb gap) in $N(E)$ near the Fermi level E_F. This means that $N(E)$ is not a constant in a narrow gap centered around E_F, whereas outside the gap $N(E)$ vanishes. If δ is the width of this gap, Efros and Shklovskii (1975) showed that $N(\delta) \sim \delta^{d-1}$. They also suggested that in 3D there is a crossover from a $T^{-1/4}$ behavior at relatively high temperatures, where the Coulomb gap is not effective, to $T^{-1/2}$ at lower temperatures. We point out that the Coulomb gap affects only those systems which have localized electronic states, such as doped crystalline semiconductors. Convincing experimental evidence for this crossover was provided by Zhang *et al.* (1990), who measured resistivity of five insulating compensated *n*-type CdSe samples, which are doped semiconductors. Their data clearly indicated a crossover from $T^{-1/4}$ behavior to $T^{-1/2}$ as the temperature was lowered by about three orders of magnitude from 15 to 0.04 K. The crossover temperature T_{co} was found to decrease as the concentration of the donor was increased. Aharony *et al.* (1992) (see also Meir, 1996) proposed that the resistivity $Z = 1/G$ of the samples obeys the scaling law

$$\ln(Z/Z_0) = Ah(T/T_{co}), \tag{129}$$

where A and T_{co} both depend on the sample properties, but the scaling function $h(x)$ is universal with $h(x) \sim x^{-1/4}$ for $x \gg 1$ and $h(x) \sim x^{-1/2}$ for $x \ll 1$. Moreover, using the percolation model, Aharony *et al.* proposed that

$$h(x) = \frac{x + [(x+1)^{1/2} - 1]}{x[1 + x^{1/2} - 1]^{1/2}}. \tag{130}$$

According to Eq. (129), if $\ln(Z/Z_0)$ for various samples is plotted as a function of T/T_{co}, all the data should collapse onto a single curve. Figure 18 shows such a collapse for the data of Zhang *et al.*, and the agreement is excellent.

4.4 Fractal Structures and Superlocalization

As discussed above, a $T^{-1/2}$ behavior for hopping conductivity can be explained in terms of the Coulomb interactions and gap. However, Coulomb interactions do not play any role in many semiconductors that do exhibit a $T^{-1/2}$ behavior. There has been some experimental evidence that one reason for the $T^{-1/2}$ behavior can be a fractal structure for the semiconductor. Lévy and Souillard (1987) suggested that in a fractal system impurity quantum states are *superlocalized*, *i.e.*, their wave functions decay with the distance r like $\exp(-r^{\gamma})$, with $\gamma > 1$ (in the classical Anderson

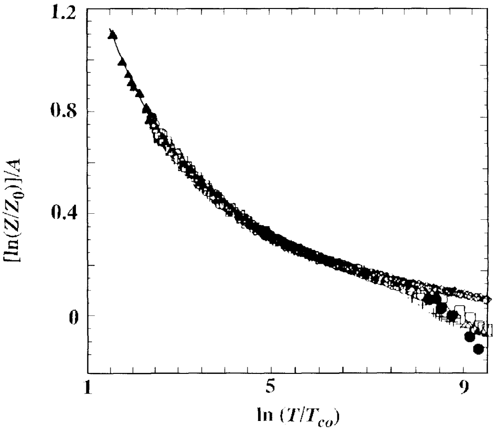

FIG. 18. Collapse of hopping conductivity $G = 1/Z$ of carbon black–polymer composites versus temperature T (in kelvins) (after Aharony *et al.*, 1992).

localization, $\gamma = 1$). Although no experimental evidence for superlocalization of electrons has been found yet, the problem has been studied theoretically by many workers. For example, Harris and Aharony (1987) and Aharony and Harris (1990) argued that one has to distinguish between the behavior of a *typical* geometry and that of a system that has been averaged over *all* the possible geometries. Moreover, they argued that an average over all the possible geometries yields $\gamma = 1$ (*i.e.*, there is no superlocalization of electrons). Early numerical studies of this problem were carried out by Lambert and Hughes (1991). The most definitive numerical study of this problem was carried out by Kantelhardt and Bunde (1997), who studied localization of electronic eigenfunctions and vibrational excitations (usually called *fractons*; see Sec. 8) on percolation clusters, in both the topological (r_t) and Euclidian (r_e) spaces, and found three different regimes. In the first regime, which corresponds to small r_t and r_e, electrons and fractons are superlocalized in the r_t space, but the localization is not exponential in the r_e space. In the intermediate regime, stretched-exponential *sublocalization* was found in both spaces with $\gamma \simeq 0.6$. In the third regime, corresponding to large r_t and r_e, the average behavior was found to depend strongly on the number of percolation configurations used in the averaging, but the asymptotic regime was a simple exponential decay with $\gamma = 1$, in agreement with the argument of Aharony and Harris.

However, none of these results supports the view that the existence of a fractal structure can by itself explain the observed $T^{-1/2}$ behavior of hopping conductivity of semiconductors in which Coulomb interactions are unimportant. van der Putten *et al.* (1992) measured the conductivity of carbon-black–polymer composites as a function of carbon-black concentration from a point in the vicinity of the percolation threshold up to $33p_c$, and in the temperature range from 4 to 300 K. Their data indicated that $\alpha \simeq 2/3$. To interpret this, they assumed that $N(E) \sim E^{\varpi}$, and proposed that

$$\alpha = \frac{\gamma(\varpi + 1)}{D_F + \gamma(\varpi + 1)}, \tag{131}$$

so that with $\alpha = 2/3$ and $D_F \simeq 2.53$ one obtains $\varpi \simeq 11/4$. Thus, their data can be best explained by a combination of a percolating fractal structure and a power-law $N(E)$. Moreover, if we take $\varpi = 1$, then Eq. (131) predicts that $\alpha \simeq 1/2$, indicating that a linear density of states and a percolating fractal structure may be responsible for the observed $T^{-1/2}$ behavior of hopping conductivity of semiconductors in which Coulomb interactions are unimportant. However, Aharony *et al.* (1993) argued that there is more than one way of interpreting the data of van der Putten *et al.* (1992), and therefore their interpretation is not unique. For the most recent experimental data for the conductivity of carbon-black–polymer composites see Mandal *et al.* (1997).

5. APPLICATION TO GRANULAR SUPERCONDUCTORS

Percolation theory can also explain some of the observed properties of granular superconductors, which are composed of superconducting grains separated by thin insulating grains that are called Josephson junctions. A granular superconductor is characterized by the size of the grains and the energy barrier between the grains, which determine the Josephson energy coupling E_J. Garland (1989) has reviwed many granular properties of superconducting materials.

Deutscher *et al.* (1980) proposed a percolation model for the onset of superconductivity by assuming that the grains are coupled if $E_J > k_B T$. Because the coupling energy depends on T, more and more grains become coupled as T is lowered. The coupling was obtained randomly with a T-dependent probability for which, when it was equal to the percolation threshold, a sample-spanning cluster of coupled superconducting grains was formed. If the temperature is still further lowered, the sample-spanning cluster of the superconducting grains grows in size. The distributions of the grain sizes and the junction resistances give rise to randomness in the coupling energy. The model was used to calculate the temperature dependence of the specific heat of the superconducting system; the agreement with the experimental data for granular Al–Al_2O_3 films was excellent.

Although there is Josephson tunnelling through the junctions, it does not generate any potential difference between the grains up to a critical current J_c. If T is decreased below the superconducting critical temperature T_c of the system, the number of superconducting links N_l (per unit cross-section area) increases, and percolation theory can be used for estimating N_l, J_c, and T_c. For example, consider the scaling behavior of J_c. Near p_c one has the power law

$$J_c \sim (p - p_c)^{\upsilon}, \tag{132}$$

where υ is a new critical exponent. Near p_c the granular superconductor can be viewed as a network of superconducting blobs connected by links (see Sec. 1.3). The links are quasi-linear chains of Josephson junctions. The current that flows through parallel links is inversely proportional to N_l, and the distance between the links is roughly proportional to the percolation correlation length ξ_p. Thus, for a d-dimensional system N_l is of the order $1/\xi_p^{d-1}$. Since $J_c \sim N_l$, we obtain

$$J_c \sim (p - p_c)^{\upsilon} \sim (p - p_c)^{\nu(d-1)}. \tag{133}$$

This prediction was confirmed by Monte-Carlo simulations in 2D percolation networks (Kirkpatrick 1979; Lobb and Frank 1979), and by careful experiments of Deutscher and Rappaport (1979) who prepared thin superconductor–semiconductor films of Pb–Ge and measured their critical currents. They obtained $\upsilon \simeq 1.3 \pm 0.1$, in excellent agreement with the prediction $\upsilon = \nu = 4/3$.

How does a magnetic field affect the behavior of a granular superconductor? There are two types of superconductors that behave quite differently in an external magnetic field.

1. Type-*I* superconductors remain superconducting and expel the magnetic field up to a critical field H_c. They then abruptly cross over to the normal state.
2. Type-*II* superconductors allow the magnetic field to penetrate into the system if H is greater than a lower critical field H_{c1}, but lose their superconductivity at an upper critical field H_{c2}.

A characteristic length scale of granular superconductors is their superconducting coherence length ξ_s, which is defined as the diffusion length over a characteristic time t_c for the relaxation of the order parameter at a given T. Therefore, one has $\xi_s \sim \sqrt{Dt_c}$, where D is a diffusion coefficient. Moreover, t_c is given by $t_c \sim (T_c - T)^{-1}$. If $\xi_s > \xi_p$, one is in the homogeneous regime in which $H_{c2} \sim T_c - T$ near T_c. The upper critical field of a type-II superconductor is proportional to ξ_s^{-2}. In the inhomogeneous regime, $\xi_s \ll \xi_p$, one has fractal diffusion (see Sec. 3.3) instead of Fickian diffusion. Since ξ_s is defined as a diffusion length scale, ξ_s^2 is proportional to $\langle R^2(t)\rangle$, the mean square displacement of a random walker given by Eqs. (95) and (96). Therefore, one obtains $H_{c2} \sim \xi_s^{-2} \sim t_c^{-2/(2+\theta)}$, which implies that

$$H_{c2} \sim (T_c - T)^{2/(2+\theta)}, \tag{134}$$

where $\theta = (\mu - \beta)/\nu$, defined in Sec. 1.5. Equation (134) was confirmed by Gerber and Deutscher (1989), who measured the upper critical field of thin semicontinuous lead films on germanium substrates and found that $H_{c2} \sim (T_c - T)^{\gamma}$ with $\gamma = 0.66 \pm 0.06$. Since in 2D $\theta \simeq 0.87$, the percolation prediction $\gamma = 2/(2 + \theta) \simeq 0.7$ is in very good agreement with the measured value.

6. APPLICATION TO ELASTIC AND VISCOELASTIC PROPERTIES OF HETEROGENEOUS MATERIALS

So far, we have discussed those applications of percolation theory in which only the connectivity of the clusters was important. We now consider more general percolation models—those in which one needs more than mere connectivity of percolation clusters in order to have macroscopic phenomena happening. An important example of such models is a class of elastic percolation networks that are examples of vector percolation. To begin with, let us consider a percolation network in which each bond represents an elastic element, *e.g.*, a spring, with an elastic constant e which takes on values from a probability density function $\psi(e)$. Since we focus mainly on the universal properties of such models, we consider the simplest case in which e is distributed according to a binary distribution,

$$\psi(e) = p\delta(e-a) + (1-p)\delta(e-b). \qquad (135)$$

If a is finite and $b = 0$, we obtain an elastic percolation network (EPN). Another case of interest is when $a = \infty$ and b is finite, *i.e.*, a fraction p of the springs are totally rigid and the rest are soft. We call this a superelastic percolation network (SEPN), which is the elastic analog of percolation models of conductor–superconductor mixtures discussed in Sec. 1.4. The percolation threshold p_{ce} of such models is not necessarily the same as p_{c}, the threshold of scalar and random percolation discussed above. As $p \rightarrow p_{\mathrm{ce}}^{+}$, all elastic moduli E of the system decrease and eventually vanish at p_{ce}. Near p_{ce} the elastic moduli of an EPN obey the power law

$$E \sim (p \sim p_{\mathrm{ce}})^{f}, \qquad (136)$$

whereas in an SEPN all the elastic moduli diverge as $p \rightarrow p_{\mathrm{ce}}^{-}$ according to

$$E \sim (p_{\mathrm{ce}} - p)^{-T}. \qquad (137)$$

The elastic percolation threshold p_{ce} and the critical exponents f and T may depend on the microscopic force laws between the sites of the network.

Let us briefly discuss how Monte-Carlo calculations with EPNs and SEPNs are carried out. To calculate elastic properties of these models we minimize their elastic energy $\mathscr{H}$ with respect to the displacement $\boldsymbol{u}_i$ of node i (*i.e.*, with respect to its components u_{ix}, u_{iy}, and u_{iz}), $\partial\mathscr{H}/\partial\boldsymbol{u}_i = 0$. Writing down this equation for every interior node of the network results in dN simultaneous equations for the nodal displacements $\boldsymbol{u}_i$ of a d-dimensional network of N internal nodes. The boundary conditions depend on the modulus that we would like to calculate. For example, to calculate the elastic constant C_{11} we stretch two opposite faces of the network by a given strain $\mathscr{S}$ (a tensile experiment), and impose periodic boundary conditions in the other directions. The resulting set of linear equations is solved typically by an iterative method such as the conjugate-gradient method. From the solution of the set of equations for nodal displacements we calculate $\mathscr{H}$, and hence the elastic modulus $C_{11} = 2\mathscr{H}/\mathscr{S}^2$. Alternatively, one can use a transfer-matrix method (Zabolitzky *et al.*, 1986), an efficient technique for calculating the elastic properties of percolation networks and their scaling properties.

6.1 Rigidity Percolation

The simplest model of vector percolation is the central-force (CF) model in which only stretching or central forces are present. The elastic energy of this model is given by

$$\mathscr{H} = \frac{\alpha}{2}\sum_{ij}[(\boldsymbol{u}_i - \boldsymbol{u}_j)\cdot\boldsymbol{R}_{ij}]^2 e_{ij} \qquad (138)$$

where α is the CF constant. Percolation on such networks is called *rigidity percolation.* Such networks are of both theoretical and practical interest. In many engineering problems, structures composed of bars or beams connected at nodes, which are called trusses, acquire their rigidity mainly from the tensile and compressive stiffness of the beams, and these are CF type of contributions. In contrast, those in which angular forces, *e.g.*, covalent bonds at the molecular level, are the most important are usually referred to as frames. For example, in the absence of friction between the particles of a granular packing the mechanical behavior of the packing is similar to that of rigidity percolation. It is not difficult to see that rigid systems in which angular forces dominate their behavior support an applied stress as long as they are simply connected. In contrast, CF systems require higher degrees of connectivity, and because of their importance to granular media, glasses, and gels they must be carefully studied. It was established in the mid 1980s (Feng and Sen, 1984) that the bond percolation threshold $p_{\mathrm{ce}}^{\mathrm{B}}$ of rigidity percolation is much larger than $p_{\mathrm{c}}^{\mathrm{B}}$, the corresponding threshold of scalar percolation. The reason for this is clear: on a CF percolation network there are many connected configurations of the network that, although sample-spanning, are not rigid, *i.e.*, their deformation does not cost any energy, and the elastic energy of the system remains unchanged. For example, for a d-dimensional cubic network one has

$p^B_{ce} = p^S_{ce} = 1$, where p^S_{ce} is the site percolation threshold of the system, which should be compared with those in Tables 1 and 2. Therefore, a meaningful study of rigidity percolation is restricted to certain lattices, *e.g.*, the triangular and bcc networks, which are regular lattices, or random lattices. Some typical values of p^B_{ce} and p^S_{ce} are

$$p^B_{ce} \simeq 0.641 \pm 0.001,$$

$$p^S_{ce} \simeq 0.713 \pm 0.002.$$

for the triangular network (Hansen and Roux, 1989; Arbabi and Sahimi, 1993), and

$$p^B_{ce} \simeq 0.737 \pm 0.002,$$

for the bcc network (Arbabi and Sahimi, 1993). These estimates should be compared with the corresponding p^B_c and p^S_c given in Tables 1 and 2. For rigidity percolation a correlation length ξ_e is defined that, similar to ξ_p for scalar percolation, diverges as p_{ce} is approached as

$$\xi_e \sim (p - p_{ce})^{-\nu_e}, \tag{139}$$

where in 2D $\nu_e \sim 1.1$.

Scaling properties of the elastic moduli of CF percolation networks have been studied extensively. For bond percolation in 2D, numerical simulations (Hansen and Roux, 1989; Sahimi and Arbabi, 1989; Arbabi and Sahimi, 1993) show that

$$f/\nu_e \simeq 2.95, \tag{140}$$

while for site percolation (Arbabi and Sahimi, 1993), $f/\nu_e \simeq 1.12 \pm 0.05$, very different from that of bond percolation. For bond percolation on the bcc network, Monte-Carlo simulations indicate (Arbabi and Sahimi, 1993) that $f/ve \simeq 2.1 \pm 0.2$, which is close to the critical exponent of conductivity of random percolation networks, $\mu \simeq 2.00 \pm 0.04$. Superelastic percolation networks with with CFs have also been studied by various techniques. For bond percolation on the triangular network the most accurate estimate of T (Arbabi and Sahimi, 1993) is $T/\nu_e \simeq 0.92 \pm 0.02$, while for bond percolation on a bcc network (Arbabi and Sahimi, 1993), $T/\nu_e \simeq 0.80 \pm 0.03$. We should point out that these estimates of f and T were obtained with regular lattices. It has been shown (Moukarzel and Duxbury, 1995; Jacob and Thorpe, 1995, 1996) that rigidity percolation on random lattices behaves quite differently than on the regular lattices. It is not yet clear whether f and T in random lattices will be different from the above estimates.

Bond percolation in CF percolation networks has also been studied by an EMA. For a d-dimensional network of coordination number Z, the effective force constant α^e is given (Feng *et al.*, 1985) by

$$\int_0^\infty \frac{\alpha^e - \alpha}{\alpha + \alpha^e(Z/2d - 1)} \psi(\alpha)\, d\alpha = 0, \tag{141}$$

which is very similar to Eq. (43) for scalar percolation, where $\psi(\alpha)$ is the distribution of the force constant α (note that αe_{ij} can be interpreted as the force constant of the bond i–j, so that a distribution for e_{ij} implies a distribution for α). From the effective force constant α^e, one can calculate the elastic moduli of the network. For example, for the triangular network the shear modulus is given by $4\alpha^e/\sqrt{3}$. It is not difficult to see that Eq. (141) predicts that

$$p^B_{ce} = 2d/Z. \tag{142}$$

According to (142), the average coordination number of a d-dimensional CF network at p_{ce} is given by $\langle Z \rangle = p^B_{ce} Z = 2d$, independent of the coordination number of the network itself. This implies that the coordination number of a d-dimensional CF network must be greater than $2d$ in order for it to have nonzero elastic moduli, thus explaining why the study of this model on the simple cubic lattices is meaningless. The same is true for the hexagonal and the diamond lattices. Equation (142) can also be derived by a constraint-counting argument (Feng *et al.*, 1985). When p is small, the CF network consists of disconnected pieces (floppy regions) and hence has many zero-frequency modes whose number is given by the number of degrees of freedom, $N_d = Nd$, minus the number of constraints, $N_c = ZNp/2$, where N

is the number of nodes in the lattice. Thus, the fraction of zero-frequency modes is given by $(Nd - ZNp/2)/Nd = 1 - Zp/2d$, which vanishes at $p = p_{\mathrm{ce}}^{\mathrm{B}} = 2d/Z$, identical with Eq. (142).

Equation (142) predicts that for the triangular network $p_{\mathrm{ce}}^{\mathrm{B}} = 2/3 \simeq 0.666$, which should be compared with the numerical estimate $p_{\mathrm{ce}}^{\mathrm{B}} \simeq 0.642$ given above. For the bcc lattice Eq. (142) predicts $p_{\mathrm{ce}}^{\mathrm{B}} = 3/4 = 0.75$, compared with the numerical estimate $p_{\mathrm{ce}}^{\mathrm{B}} \simeq 0.737$. Thus, unlike the EMA for percolation conductivity, Eq. (43), the performance of the EMA for rigidity percolation improves as the dimensionality of the system increases. Equation (141) also predicts that the critical exponent of the elastic moduli is $f = 1$, which does not agree with its numerical estimates. However, except for a very small region near $p_{\mathrm{ce}}^{\mathrm{B}}$, the predictions of Eq. (141) are in very good agreement with numerical simulation results.

6.2 Bond-Bending Models

One main shortcoming of CF percolation networks is that their percolation thresholds p_{ce} are much large than those of scalar percolation p_{c}. Thus, for $p_{\mathrm{c}} \le p \le p_{\mathrm{ce}}$, where p is the fraction of the intact springs, an EPN with only central forces is geometrically connected, but its elastic moduli are all zero, which is unphysical. Another shortcoming of CF percolation networks is that the significance of a straight-bond chain in transmitting elastic forces in them is ambiguous, because in a nonlinear model the straight bonds could buckle under compression but not under extension.

A more general vector percolation model with both central and bond-bending (or angle-changing) forces is free of such shortcomings. One of the main advantages of the bond-bending (BB) model is that its percolation threshold can be the same as that of scalar percolation, if the appropriate microscopic force laws between the sites of the network are used, and hence the model can be used for investigating mechanics of disordered solids such as polymers, ceramics, and powders. In general, the elastic energy of any bond-bending model is given (Kantor and Webman, 1984) by

$$\mathscr{H} = \frac{\alpha}{2}\sum_{\langle ij\rangle}[(\boldsymbol{u}_i - \boldsymbol{u}_j)\cdot \boldsymbol{R}_{ij}]^2 e_{ij} + \frac{\gamma}{2}\sum_{\langle jik\rangle}(\delta\theta_{jik})^2 e_{ij}e_{ik}, \tag{143}$$

where α and γ are the central and BB force constants, respectively. Here $\langle jik\rangle$ indicates that the sum is over all triplets in which the bonds j–i and i–k form an angle whose vertex is at i. The first term in Eq. (143) represents the usual contribution of the CFs, while the second term is due to BB forces. The precise form of $\delta\theta_{jik}$ depends on the microscopic details of the model. If bending of the bonds that make 180° with one another (*i.e.*, collinear bonds) is not allowed, then

$$\delta\theta_{jik} = (\boldsymbol{u}_i - \boldsymbol{u}_j)\cdot \boldsymbol{R}_{ik} + (\boldsymbol{u}_i - \boldsymbol{u}_k)\cdot \boldsymbol{R}_{ij}. \tag{144}$$

We refer to this particular version of the model as the Kirkwood–Keating (KK) model. Kirkwood (1939) used this model to study vibrational properties of rod-like molecules, while Keating (1966) studied the elastic properties of covalent crystals with essentially the same model (but not in the context of percolation). If bending of the collinear bonds is allowed, then

$$\delta\theta_{jik} = \begin{cases} (\boldsymbol{u}_{ij}\times\boldsymbol{R}_{ij} - \boldsymbol{u}_{ik}\times\boldsymbol{R}_{ik})\cdot(\boldsymbol{R}_{ij}\times\boldsymbol{R}_{ik})/|\boldsymbol{R}_{ij}\times\boldsymbol{R}_{ik}|, \\ \qquad \boldsymbol{R}_{ij} \text{ not parallel to } \boldsymbol{R}_{ik}, \\ |(\boldsymbol{u}_{ij}+\boldsymbol{u}_{ik})\times\boldsymbol{R}_{ij}|, \\ \qquad \boldsymbol{R}_{ij} \text{ parallel to } \boldsymbol{R}_{ik}, \end{cases} \tag{145}$$

where $\boldsymbol{u}_{ij} = \boldsymbol{u}_i - \boldsymbol{u}_j$. For all 2D systems, Eq. (145) simplifies to

$$\delta\theta_{jik} = (\boldsymbol{u}_i - \boldsymbol{u}_j)\times \boldsymbol{R}_{ij} - (\boldsymbol{u}_i - \boldsymbol{u}_k)\times \boldsymbol{R}_{ik}. \tag{146}$$

We refer to this as the BB model. For most materials to which the KK or the BB models are applicable, one has $\gamma/\alpha \le 0.3$.

Phillips and Thorpe (1985) used a mean-field theory to predict that the bond percolation threshold $p_{\mathrm{ckk}}^{\mathrm{B}}$ of a d-dimensional KK model with the coordination number Z is given by

$$p_{\text{cck}}^{\text{B}} \simeq \frac{1}{Z}\frac{d^2+d}{2d-1}. \qquad (147)$$

On the other hand, as mentioned in Sec. 1.2, for scalar percolation one has

$$p_{\text{c}}^{\text{B}} \simeq \frac{1}{Z}\frac{d}{d-1}. \qquad (148)$$

Thus, for 2D systems, $p_{\text{ckk}}^{\text{B}} = p_{\text{c}}^{\text{B}} \simeq 2/Z$, and simulations have confirmed that the percolation threshold of the KK model in 2D is the same as that of the scalar percolation. However for 3D systems Eq. (147) predicts that

$$p_{\text{ckk}}^{\text{B}} \simeq 2.4/Z, \qquad (149)$$

so that the average coordination number of the network at the percolation threshold should be about 2.4. On the other hand, Eq. (148) yields $p_{\text{c}}^{\text{B}} \simeq 1.5/Z$. Equation (149) was confirmed by numerical simulations (He and Thorpe, 1985). The BB model in d dimensions, on the other hand, has the same percolation threshold as p_{c} if each site of the network interacts with at least $d(d-1)/2$ of its nearest-neighbors. Therefore, by using suitable forms of $\mathscr{H}$ one obtains EPNs whose percolation thresholds are the same as p_{c}, a desirable property as far as many practical applications are concerned.

The critical exponents f and T have also been estimated for the BB models by a variety of techniques. In 2D the most accurate estimate of f was obtained by a transfer-matrix method (Zabolitzky *et al.*, 1986), while in 3D f was estimated (Arbabi and Sahimi, 1988) by large-scale Monte-Carlo simulations. The numerical results are listed in Table 4. Moreover, Sahimi (1986a), and later Roux (1986), proposed that

$$f = \mu + 2\nu, \qquad (150)$$

where μ is the critical exponent of the conductivity of percolation networks. The predictions of Eq. (150) are in excellent agreement with the numerical results given above, and thus it is likely to be exact. The origin of Eq. (150) is clear: elasticity introduces two extra factors of ξ_p, and therefore one must have, elastic moduli/conductivity $\sim \xi_p^2$, which results in Eq. (150).

The 2D and 3D superelastic BB models have been studied extensively. Such models were first proposed by Sahimi and Goddard (1985) for modeling the divergence of the viscosity η of a gelling solution near and below the gel point (see below). In addition to their relevance to modelling viscoelastic properties of gel polymers, SEPNs can be thought of as a model of random reinforcement of disordered materials. The most accurate estimates of the critical exponent T for BB models were also obtained (Arbabi and Sahimi, 1990; Sahimi and Arbabi, 1993) by large-scale Monte-Carlo simulations, and are also listed in Table 4. These estimates are somewhat smaller than the corresponding values for the exponent s of conductor–superconductor mixtures defined by Eq. (9). It has been proposed (Arbabi and Sahimi, 1990; Sahimi and Arbabi, 1993) that

$$T = \nu - \frac{\beta}{2}. \qquad (151)$$

The predictions of Eq. (151) agree with the numerical estimates of T.

The BB models have also been studied with an EMA (Schwartz *et al.*, 1985; Mall and Russel, 1987). The general form of the EMA for this case is similar to Eqs. (43) and (141). It

Table 4. Values of the critical exponents of scalar and vector percolation in d dimensions. Values of f and T for the central-force model refer to bond percolation, while those of μ and s are independent of the model.

d	μ/ν	s/ν	f/ν	T/ν	Model
2	0.9745 ± 0.0015	0.9745 ± 0.0015	2.97 ± 0.03	0.92 ± 0.03	Bond bending
	—	—	2.95 ± 0.25	0.92 ± 0.02	Central force
3	2.27 ± 0.01	0.835 ± 0.005	4.3 ± 0.1	0.74 ± 0.04	Bond bending
	—	—	2.1 ± 0.1	0.80 ± 0.03	Central force

predicts that $p_c^B = 2/Z$, the same as that of scalar percolation. The numerical predictions of the EMA for the BB models in 2D are also accurate, except very near p_c.

We should mention two other related models. One is the beam model (Roux and Guyon, 1985) in which each bond is a beam whose elastic behavior is governed by three material-dependent constants, $a_1 = l/E_{\mathrm{Y}}S$, $a_2 = l/E_{\mathrm{s}}S$, and $a_3 = l^3/E_{\mathrm{Y}}I$, where E_{Y} and E_{s} are the Young and shear moduli, S is the cross section of the beam, and I is the moment of inertia for flexion. Each site of the network is characterized by a displacement vector $\boldsymbol{u}_i$ and a rotational angle θ_i. If a site is rotated, the beams bend accordingly, and thus the local momenta are taken into account. Consider, for example, a square network. For a horizontal beam that connects sites i and j one has a longitudinal force F_{lj} acting at j, $F_{lj} = \alpha_1(u_{ix} - u_{jx})$, a shear force $F_{sj} = \alpha_2(u_{iy} - u_{jy}) + \alpha_2 l(\theta_i + \theta_j)/2$, and a flexural torque $\tau_j = \alpha_2 l(u_{iy} - u_{jy} + l\theta_j)/2 + \alpha_3 l^2(\theta_i - \theta_j)$, where $\alpha_1 = 1/a_1$, $\alpha_2 = 1/(a_2 + a_3/12)$, and $\alpha_3 = \alpha_2(a_2/a_3 + 1/3)$. In a similar way, the analogous equations for the vertical direction are written down. In mechanical equilibrium the sums of all forces and of all torques acting on a site must be zero, giving rise, in the continuum limit, to the classical Cosserat equations. The percolation threshold of this model is of course the same as that of scalar percolation. Numerical simulations also indicated that the scaling properties of this model are identical with those of the BB model.

The second model is the so-called granular or disk model (Schwartz *et al.*, 1984) which has been used for studying vibrational properties of granular rock, such as sandstone. The elastic energy of the system is given by

$$\begin{aligned}\mathscr{H} = \frac{1}{2}\sum_{\langle ij\rangle} e_{ij}\{&\alpha[(\boldsymbol{u}_i - \boldsymbol{u}_j)\cdot \boldsymbol{R}_{ij}]^2 \\ &+ \gamma\{\boldsymbol{R}_{ij} \times [\boldsymbol{R}_{ij} \times (\boldsymbol{u}_i - \boldsymbol{u}_j)] \\ &+ R(\theta_i + \theta_j)\boldsymbol{z} \times \boldsymbol{R}_{ij}\}^2 \\ &+ \delta(\theta_i - \theta_j)^2\}, \qquad (152)\end{aligned}$$

where R is the radius of the disks, δ is the elastic constant associated with the restoring torque that is necessary when two disks counterrotate, and $\boldsymbol{z}$ is a unit vector normal to the plane of the disks. The rest of the notation is the same as before. With such couplings, any relative displacements or rotations of two neighboring disks can be done at a cost to the energy $\mathscr{H}$ of the system. Therefore, the percolation threshold of this system is the same as that of scalar percolation. In terms of a lattice model, the granular model is essentially equivalent to the beam model described above. Numerical studies indicate that the scaling properties of the elastic moduli of this model are the same as those of the BB model. Therefore, the BB model is the simplest elastic percolation model that possesses two important and fundamental properties that most models of heterogeneous materials must have: a rotationally invariant elastic energy, and a percolation threshold that is identical with that of scalar percolation. We summarize in Table 4 the currently accepted estimates of the various critical exponents for both rigidity percolation and the BB model.

Let us now discuss applications of EPNs and SEPNs to modelling of various properties of composite solids and gel polymers (the application to glasses is discussed separately).

6.3 Elastic Properties of Composite Solids

Deptuck *et al.* (1985) measured the Young's modulus of a sintered, submicrometer powder that was described in Sec. 2.1. Their measurements indicated that the Young's modulus of the powder obeys scaling law (136) with $f \simeq 3.8$, in excellent agreement with the numerical estimate (see Table 4) of the BB model in 3D. Allen *et al.* (1988) carried out dynamical measurements of the shear modulus of a model 2D percolation lattice, by exciting small torsional linear oscillations on a cylindrical screen as links were cut to approach p_c. They obtained $f \simeq 3.6 \pm 0.2$, relatively close to $f \simeq 3.96$ in 2D. Measurements of the elastic moduli of 2D composites were also carried out by Benguigui (1986), Lobb and Forrster (1987), and Sofo *et al.* (1987). Benguigui's experiments were done with sheets of metal in which holes had been punched; he obtained $f \simeq 5.0 \pm 0.5$, consistent with the prediction of Halperin *et al.* (1985) and Feng *et al.* (1987) that the critical exponent f for such 2D percolating continua should be larger than the corresponding value for lattice systems by about 3/2 (hence, $f \simeq 3.96 + 1.5 = 5.5$). Measurements of Lobb and Forrester (1987) were also carried out with metal sheets into which holes had been punched. They obtained $f \simeq 4.95 \pm 1.1$, again consistent with Halperin *et al.*'s prediction. Sofo *et al.* (1987) used randomly holed metallized Mylar and obtained $f \simeq 5.3 \pm 0.7$, in even better agreement with the prediction of Halperin *et al.* (1985).

On the other hand, Benguigui and Ron (1993) constructed a mixture of a very soft material, in their case a gel, and a very rigid material, which was alumina and zirconia, and measured the Young's modulus of the mixture as the volume fraction of the rigid materials approached p_c, the percolation threshold of the rigid component. They found that the modulus diverges at p_c, and that near the percolation threshold the Young's modulus of the mixture obeys scaling law (140) with $T \simeq 0.66 \pm 0.04$, in excellent agreement with the value of superelasticity exponent in 3D (see Table 4). Thus, there remains no doubt that elastic and other mechanical properties of disordered solids can be described by EPNs and SEPNs. Elastic percolation networks have also been used for modeling wave propagation in porous media, for modelling fracture of rock at large length scales (Sahimi *et al.*, 1993), and for explaining the spatial distribution (Sahimi *et al.*, 1993) and dynamics (Sahimi and Arbabi, 1996) of earthquakes.

An important characteristic of any solid is its Poisson ratio $\mathscr{P}$. For a 2D isotropic solid $\mathscr{P}$ is given by

$$\mathscr{P} = 1 - 2E_s/E_t, \tag{153}$$

where E_s and E_t are the shear and tensile moduli, respectively. For a 3D isotropic solid the corresponding equation is

$$\mathscr{P} = \frac{3(E_b/E_s) - 2}{2 + 6(E_b/E_s)}, \tag{154}$$

where E_b is the bulk modulus of the solid. The Poisson ratio is a measure of the strain in the transverse direction that results from a stress applied longitudinally. In most materials the experimentally measured $\mathscr{P}$ is positive, which corresponds to materials that shrink transversely when stretched longitudinally. The definition of $\mathscr{P}$ for an anisotropic solid is arbitrary. Theoretically (Landau and Lifshitz, 1959) one can have highly compressible isotropic materials with Poisson ratios as low as -1, and more generally one has $-1 \leq \mathscr{P} \leq 1/2$. For anisotropic materials one has the general bound $|\mathscr{P}| \leq \sqrt{E_{yM}/E_{ym}}$, where E_{yM} and E_{ym} are the major and minor orthotropic Young's moduli.

The elastic percolation models discussed here can also possess negative values of the Poisson ratio. On the experimental side, there are some manufactured foams (Lakes, 1987) that are isotropic and their Poisson ratio is as low as -0.5. There are also many anisotropic materials, such as microporous poly(tetrafluoroethylene) (Evans and Caddack, 1989), microporous ultrahigh-molecular-weight polyethylene (Alderson and Evans, 1992) and iron pyrites, pyrolitic graphite, hexagonal metals, and fiber-reinforced composites, that have large negative Poisson ratios, as low as -1.2, regardless of how $\mathscr{P}$ is defined. Of course, 3D elastic percolation models on all regular lattices are also anisotropic. Thus, negative Poisson ratios, as demonstrated by elastic percolation models and by many real materials, should not be considered as unphysical and, in fact, they may point to the possibility of designing a new class of materials with interesting, unusual, and useful properties. For example, they would have enhanced flexural rigidity and plane strain-fracture toughness, but lower bulk modulus.

6.4 Elastic and Viscoelastic Properties of Gel Polymers

Another class of materials to which elastic and superelastic percolation models are applicable comprises gel polymers. To compare experimental results for the scaling properties of the elasticity of gels with the predictions of EPNs, we divide the gel polymers into two groups. In the first group are what we call physical gels, in the formation of which chemical reactions play no role. Two examples are gelation of silica particles in NaCl solutions and in pure water (Gauthier-Manuel *et al.*, 1988) and silica aerogels (Woignier *et al.*, 1988). The attachment of the particles in such gels is by relatively weak association. The BB forces are important in such gels since touching particles that form long chains, when deformed, roll on top of one another and this motion, together with the displacement of the centers of any three mutually touching particles, creates forces that are equivalent to the BB forces. Experimental measurements (Gauthier-Manuel *et al.*, 1988; Woignier *et al.*, 1988) of f for such gels confirmed this: f was found to be about 3.8, in excellent agreement with the critical exponent f of the 3D BB model (see Table 4).

In the second group are what we call chemical gels, which are the usual branched polymers are formed by a chemical reaction between the monomers. In most chemical gels the BB forces are usually not important and the only important force between the monomers seems to be the central (stretching) forces. There are numerous experimental measurements of f for such gels. Examples include measurement of f for hydrolyzed polyacrylamide (Allain and Salomé, 1987b, 1990), tetraethylorthosilicate reactions (Hodgson and Amis, 1990; Takahashi *et al.*, 1994), gelatin solutions (Djabourov *et al.*, 1988), polycondensation of polyoxypropylated trimethylolpropane with hexamethylenediisocyanate (Durand *et al.*, 1987), and several other measurements (Gauthier-Manuel and Guyon, 1980; Adam *et al.*, 1981; Tokita *et al.*, 1984). These measurements all yielded a value of f in the range 1.9–2.2, which does not agree with that of the BB model, but agrees nicely with the exponent of the 3D CF percolation (see Table 4).

A value of f in the range 1.9–2.2 can also be interpreted by another percolation model. Alexander (1984) argued that in some gels and rubbers that are under internal or external stress, there are terms in the elastic energy of the system that are similar to the Born model described by

$$\mathscr{H} = \frac{1+\mathscr{P}}{1-\mathscr{P}} \sum_{ij} [(\boldsymbol{u}_i - \boldsymbol{u}_j) \cdot \boldsymbol{R}_{ij}]^2 e_{ij} + \frac{1-3\mathscr{P}}{4(1-\mathscr{P})} \sum_{ij} (\boldsymbol{u}_i - \boldsymbol{u}_j)^2, \qquad (155)$$

where all the notations are as before. In 3D the critical exponent of this system is $f = \mu \simeq 2.0$, because near p_c the contribution of the second term on the right-hand side of Eq. (155), which is a purely scalar term, dominates that of the first term, which is due to the central forces. However, Eq. (155) has the unphysical feature that it is not rotationally invariant. Moreover, because rubbers and gels differ from the Born model in several important ways, such as the presence of nonlinear terms in their elastic energy and the possibility of negative as well as positive Born coefficients, it is not clear that the result $f \simeq \mu$ is applicable to such materials.

We should mention that there are some chemical gels in which the BB forces seem to be important, and thus their scaling behavior may be described by the BB model. Examples include the measurements of Adam *et al.* (1985) for polycondensation, $f \simeq 3.3 \pm 0.5$, and those of Martin *et al.* (1988) and Adolf *et al.* (1990) for gels made from 89 wt% of the diglycidyl ether of bisphenol A cured with 11 wt% of diethanolamine, which yielded $f \simeq 3.3 \pm 0.3$. These values are much closer to $f \simeq 3.75$ for the 3D BB model than to $f \simeq 2.1$ for the 3D CF model. We must also point out that in some polymers entropic effects are important (leading to the formation of long chains and entanglement), whereas in the elastic percolation models described here such effects have not been taken into account (the elastic energy of the EPNs considered here is enthalpic). It is not clear that if the entropic effects are taken into account (a difficult task), f would remain the same as the one that one finds with the BB models. Daoud and Coniglio (1981) and Martin *et al.* (1988) suggested that for such cases

$$f = \nu d. \qquad (156)$$

In 3D Eq. (156) predicts $f \simeq 2.64$, quite different from the exponents of the CF and BB models. Recent measurements of the shear modulus of an end-linking gel polymer by Takahashi *et al.* (1994) yielded $f \simeq 2.7$, in very good agreement with Eq. (156).

Numerous experimental measurements indicate that the viscosity η of a gelling solution below and near the gel point obeys the power law

$$\eta \sim (\phi_c - \phi)^{-k}, \qquad (157)$$

where ϕ is the gel fraction and ϕ_c its value at the gel point, which is a percolation threshold. To explain the experimental data on the scaling behavior of η near the gel point, it was proposed (Arbabi and Sahimi, 1990; Sahimi and Arbabi, 1993) that the scaling properties of η near the gel point are analogous to those of the shear modulus of a SEPN near p_c. This analogy was made simply because both quantities diverge at p_c (the gel point). However, even the analogy between η and the shear modulus of an SEPN is not nearly enough to explain the scaling of η near the gel point, because experimental data indicate that the value of k is either in the range 0.6–0.9 (see, for example, Adam *et al.*, 1979, 1985; Allain and Salomé, 1987a; Durand *et al.*, 1987), or in the range 1.3–1.5 (see, for example, Djabourov *et al.*, 1988; Martin *et al.*, 1988, Martin and Wilcoxon, 1988), whereas the shear modulus of a 3D SEPN is characterized by a *unique* value of T. The reason for having two distinct values of k is (Arbabi and Sahimi, 1990; Sahimi and Arbabi, 1993) that the dynamics of the two solutions that yield the two distinct values of k are totally different. In one case, the system is presumably close to the Zimm limit in which there is little or no polymer diffusion in the reaction bath, because there are strong hydrodynamic interactions between the monomers, and also between polymers of various sizes. Hence, a SEPN, which is a static system with fixed rigid clusters, may be suitable for simulating the Zimm regime. Indeed, the estimate $T \simeq 0.65$ for 3D SEPNs supports the idea that those gelling solutions whose value of k is in the range 0.6–0.9 are in fact close to the Zimm regime and can be described by a *static* SEPN.

On the other hand, the gelling solution can also be near the Rouse regime in which there are no hydrodynamic interactions between the polymers of various sizes, and therefore the polymers can diffuse essentially freely in the reaction bath. To simulate this regime a dynamic superelastic percolation model was proposed (Arbabi and Sahimi, 1990; Sahimi and Arbabi, 1993) in which each cluster of the totally rigid bonds represents a polymer. The soft bonds (those for which the elastic constant e_{ij} is finite) represent the liquid (solution) medium in which the rigid clusters move randomly, with equal probability, in any direction of the network. This simulates the diffusion of the polymers in the reaction bath. Two rigid clusters cannot overlap, but they can temporarily join and form a larger cluster, which can be broken up again at a later time. It was shown (Arbabi and Sahimi, 1990; Sahimi and Arbabi, 1993) that the shear modulus of this dynamic SEPN diverges with an exponent T' given by

$$T' = 2\nu - \beta, \tag{158}$$

which in 3D predicts that $T' \simeq 1.35$, supporting the idea that the gelling solutions whose values of k are in the range 1.3–1.5 are close to the Rouse limit, and can be described by a *dynamic* SEPN.

6.5 Structural Properties of Branched Polymers and Gels

After a polymer is formed by a chemical reaction, the experimentalist usually analyzes its structure by diluting it in a good solvent. Such branched polymers in a dilute solution of a good solvent may swell and have a radius larger than their extent at the end of the chemical reaction. Thus, it is important to consider both typical polymers and the swollen ones. Consider first a swollen branched polymer in a good solvent whose radius is larger than the correlation length. The structural properties of such branched polymers are described by *lattice animals*, which are percolation clusters below p_c whose radii are larger than the percolation correlation length ξ_p. However, the scaling properties of lattice animals are different from those of percolation clusters (Lubensky and Isaacson, 1978; Family and Coniglio, 1980). For example, their fractal dimension D_A is quite different from D_F, the fractal dimension of the sample-spanning percolation clusters at p_c. In fact, Parisi and Sourlas (1981) showed that in 3D, $D_A = 2$ exactly, quite different from $D_F \simeq 2.52$.

One way of checking this prediction is by small-angle scattering. The intensity $I(q)$ of scattering at an angle θ_s is related to the magnitude of the scattering vector, $q = 4\pi\sin(\theta_s/2)/\lambda$ (λ is the wavelength of radiation), by

$$I(q) \sim q^{-D_A}. \tag{159}$$

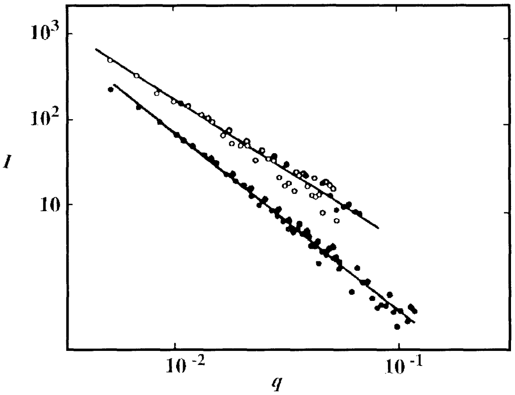

FIG. 19. SANS data for branched polymers. The upper curve is for a polydisperse polymer solution, while the lower curve is for a single polymer in a dilute good solvent (after Bouchaud *et al.*, 1986).

Bouchaud *et al.* (1986) carried out small-angle neutron-scattering (SANS) experiments on a monodisperse polyurethane sample and measured the scattered intensity as a function of q. Figure 19 presents their results, from which one obtains $D_A = 1.98 \pm 0.03$, in excellent agreement with the Parisi–Sourlas prediction, $D_A = 2$. In real applications, however, polymer solutions are almost always polydisperse and contain polymers of all sizes with radii smaller or larger than ξ_p. Thus, one needs to define *average* properties, where the averaging is taken over the polymer size distribution. For example, an average polymer radius $\langle R \rangle$ is used to define an *effective* fractal dimension for all the branched polymers in the solution through the relation $n \sim \langle R \rangle^{D_E}$, where n is the average number of the monomers. Daoud *et al.* (1984) showed that for a polydisperse polymer in a dilute solution of a good solvent

$$D_E = D_A(3 - \tau). \qquad (160)$$

Thus, Eq. (160) mixes the animal fractal dimension D_A with the percolation exponent τ defined by Eq. (15). Thus, one should have $D_E \simeq 1.64$ for 3D polymers. Bouchaud *et al.* (1986) synthesized a natural polydisperse polyurethane sample and carried out SANS experiments on a dilute solution of it. Figure 19 also shows their results, from which one obtains $D_E \simeq 1.6 \pm 0.05$, in good agreement with the theoretical prediction of 1.64. For related experiments see Adam *et al.* (1985), who carried out static light-scattering experiments in dilute polydisperse polyurethane solutions; Leibler and Schosseler (1985), who measured the average radius of polystyrene, crosslinked by irradiation by elastic light scattering; Patton *et al.* (1989), who performed both quasielastic and elastic light scattering experiments on branched polyesters; and Dubois and Cabane (1989), who used a silica gel, a physical (rather than chemical) gel, all of whom found evidence supporting the theoretical predictions.

Note that Eq. (160) can also be used for estimating the percolation exponent τ, since D_E and D_A can be measured directly from the small-angle experiments. Indeed, using the experimental values for D_E and D_A, one obtains $\tau \simeq 2.2$, in excellent agreement with the percolation prediction of 2.18 (see Table 3). In their experiments with irradiated polystyrene solution in cyclopentane, Leibler and Schosseler (1985) coupled gel permeation chromatography and light scattering and obtained the polymer size distribution, which provides a direct means of measuring τ, using Eq. (15). They obtained $\tau \simeq 2.3 \pm 0.1$, consistent with the percolation prediction. Lapp *et al.* (1989) further checked this result by doing similar experiments in a system made by chemical end-linking of polydimethylsiloxane, and Patton *et al.* (1989) did the same in a system in which polyester was made by bulk condensation polymerization. The results of both groups were consistent with the percolation prediction for τ. Note that since the fractal dimension D_F is related to τ by

$$D_F = d/(\tau - 1), \qquad (161)$$

these results also imply that $D_F \simeq 2.5$, in agreement with the percolation prediction, $D_F \simeq 2.52$. For other experiments confirming the applicability of percolation theory to describing the structure of branched polymers and gels see Adam *et al.* (1985), who performed static light-scattering measurements on a polyurethane sol; Candau *et al.* (1985), who carried out their experiments on polystyrene systems crosslinked with di(vinyl benzene); Schmidt and Burchard (1981), who carried out anionic copolymerization of di(vinyl benzene) with styrene and obtained branched polymers and gels and used light scattering to measure various properties of interest; and Trappe *et al.* (1992). More recent theoretical developments in this area include the work of Lucena *et al.* (1994) who introduced models of branched polymers that seem different from those of lattice animals discussed above. Bunde *et al.* (1995) studied this model extensively, and showed that the branched polymers that are generated by this model are similar to percolation clusters (*i.e.*, gels) rather than lattice animals. However, no experimental verification of this result is yet available.

7. APPLICATION TO GLASS FORMATION

In an attempt to explain the strong glass-forming propensity of some chalcogenide systems, Phillips (1979) proposed for the first time that covalent bonding in glasses can be optimized when the average coordination number $\langle Z\rangle$ of the 3D covalent network in mechanical equilibrium is about 2.4. This idea was significantly refined by Thorpe (1983), who established the connection between this phenomenon and a vector percolation problem. According to the Phillips–Thorpe (PT) ideas, in covalent networks for $\langle Z\rangle < 2.4$, the number N_f of interatomic (central and BB) forces per atom that act as constraints (thus $N_f = N_c$, where N_c is the number of constraints) is less than the network dimensionality $d = 3$, which is the same as the number of degrees of freedom N_d per atom. Under a shearing force such networks easily deform and possess a finite number of zero-frequency modes $f_m = N_d - N_c$. When $\langle Z\rangle > 2.4$, N_c is larger than N_d, and thus such networks are macroscopically rigid and possess nonzero elastic moduli. Hence, when $\langle Z\rangle = 2.4$, one has $N_c = N_d$ and therefore $f_m = 0$, which represents the vector percolation threshold of the system. Two crucial points should be noted here. First, the arguments leading to $\langle Z\rangle \simeq 2.4$ are all mean-field type of reasoning. However, computer simulations of He and Thorpe (1985) confirmed that this mean-field estimate is highly accurate. Secondly, $\langle Z\rangle \simeq 2.4$, which agrees with Eq. (149), represents the average coordination number of 3D networks at the percolation threshold of the KK model.

There have been many experimental tests of these ideas, the first of which was probably provided by Bresser *et al.* (1986). They used binary Ge_xSe_{1-x} glasses, which are perhaps the simplest of such materials. The coordination numbers of Ge and Se are 4 and 2, respectively, over the composition range $0 < x < 1/3$, and thus the average coordination number of the network is $\langle Z\rangle = 2(x+1)$. Therefore, with $\langle Z\rangle \simeq 2.4$, one has $x_c \simeq 0.2$. Bresser *et al.* (1986) used such binary glasses in the composition range $0 < x < 1/3$ using ^{129}I Mössbauer emission spectroscopy, and presented strong evidence that $x_c \simeq 0.23 \pm 0.02$. The Mössbauer-site intensity ratio I_B/I_A indicates dramatic threshold behavior. At $x = 0$ (*i.e.*, *g*-Se) only B sites are observed, and thus I_B/I_A is infinite. For $0 < x < 0.15$, I_B/I_A decreases with increasing x, reaching a minimum around $x \simeq 0.15$, but as $x_c \simeq 0.23$ is approached this ratio increases and reaches a maximum. Beyond x_c the ratio I_B/I_A decreases monotonically with x again. The molar volumes V_m of Ge_xSe_{1-x} also indicate a threshold behavior. For $0 < x < 1/3$ the molar volume decreases monotonically with x, reaching its minimum at $x_c \simeq 0.23 \pm 0.02$, beyond which it increases again monotonically as x does. Note that $x_c \simeq 0.23 \pm 0.02$ agrees nicely with the PT estimate $x_c = 0.2$, corresponding to an average coordination number of about 2.4.

Further experimental evidence of the applicability of the PT ideas was provided by Halfpap and Lindsay (1986), who measured the elastic moduli C_{11} and C_{44} of the Ge–As–Se glasses over a wide range of composition. The coordination number of As is 3, and therefore if the glass has the composition $Ge_xAs_ySe_{1-x-y}$, then its average coordination number is $\langle Z\rangle = 4x + 3y + 2(1-x-y) = 2x + y + 2$. Thus, varying x and y allows one to obtain glasses over a wide range of compositions and test the applicability of vector percolation to glass formation. Despite different local structures of the glasses that are obtained when x and y are varied, measurements of Halfpap and Lindsay (1986) indicated that at $\langle Z\rangle \simeq 2.4$ there is a dramatic increase in the elastic moduli C_{11} and C_{44}; an example is shown in Fig. 20. Further

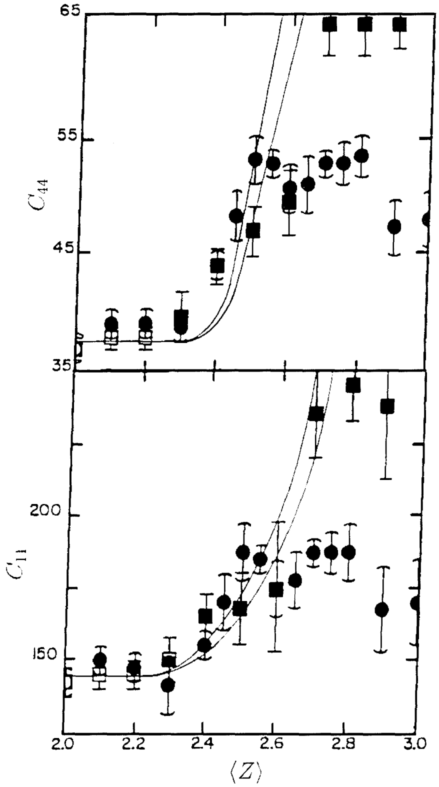

FIG. 20. Threshold behavior of the elastic moduli (in kbar; 1 bar = 10^5 Pa) of glasses versus the average coordination number $\langle Z\rangle$. Circles and squares correspond, respectively, to high- and low-As series of glasses. The theoretical curves on the left are for amorphous Ge, while those on the right are for crystalline Ge (after Halfpap and Lindsay, 1986).

evidence for the applicability of the PT ideas was provided by the experimental data of Mahadevan and Giridhar (1989), who also measured the elastic moduli C_{11} and C_{44} of various glasses formed with As, Ge, S, Sb, Se and Te, and found essentially the same behavior that had been found by Halfpap and Lindsay (1986). Finally, Feng *et al.* (1997) presented Raman-scattering data for $\mathrm{Ge}_x X_{1-x}$ glasses (X = S or Se) that indicated that there is a stiffness threshold at a mean coordination number $\langle Z\rangle \simeq 2.46 \pm 0.01$, in close agreement with the PT ideas. For related experimental works see Yun *et al.* (1989), Tatsumisago *et al.* (1990), and Böhmer and Angell (1992). Grest and Cohen (1983) give a review of older works that attempted to establish a connection between glass transition and percolation.

8. APPLICATION TO VIBRATIONS AND DENSITY OF STATES OF MATERIALS

An important property of materials is $N(\omega)$, their vibrational density of states, which is used for estimating their specific heat and thermal conductivity. $N(\omega)\,d\omega$ is the number of vibrational modes with frequencies between ω and $\omega + d\omega$. On the basis of the similarity between the Fourier transform of the equation of motion for a system of Hookeian springs,

$$-\omega^2 \tilde{\boldsymbol{u}}_i = \sum_{\langle ij\rangle} e_{ij}[(\tilde{\boldsymbol{u}}_j - \tilde{\boldsymbol{u}}_i)\cdot \boldsymbol{R}_{ij}], \qquad (162)$$

and the Laplace transform of the discretized diffusion equation,

$$\lambda \hat{P}_i(\lambda) = \sum_{\langle ij\rangle} W_{ij}(\hat{P}_j - \hat{P}_i) \qquad (163)$$

where $P_i(t)$ is the probability of finding a diffusing particle at site i at time t and W_{ij} is the transition rate, *i.e.*, the probability per unit time that the diffusing particle jumps from i to j, Alexander *et al.* (1978) suggested that

$$N(\omega) = -2\omega\, \mathrm{Im}\langle \hat{P}_0(-\omega^2)\rangle/\pi, \qquad (164)$$

where Im denotes the imaginary part of the complex number. It is then straightforward to show that

$$N(\omega) \sim \omega^{d-1}. \qquad (165)$$

Equation (165) is valid at low frequencies (*i.e.*, long wavelengths or large length scales over which the system is homogeneous) such that $\omega < \omega_{\mathrm{co}}$, where ω_{co} is the cutoff or crossover frequency. Vibrational states in homogeneous systems are usually called phonons, which are weakly localized.

If the network is fractal, *e.g.*, a percolation network at p_c or at length scales $L < \xi_p$, then Eq. (165) is no longer valid. The density of states of fractal networks was first discussed by Alexander and Orbach (1982) and Rammal and Toulouse (1983), who argued that in this case

$$N(\omega) \sim \omega^{D_S - 1}, \tag{166}$$

where D_S is the spectral or fracton dimension, already mentioned in Sec. 4.4. For percolation networks

$$D_S = 2\frac{D_F}{D_W} = 2\frac{\nu d - \beta}{2\nu + \mu - \beta}. \tag{167}$$

Equation (165) is valid as long as the scalar contributions to the elastic energy of the material dominate the vectorial contributions (as in, for example, the Born model). However, as we discussed above, the nature of the forces that are exerted on the particles of the network and their contributions to the elastic energy are most important. Hence, if we wish to take into account the vectorial nature of the vibrational density of states, then we are forced to introduce a new quantity D_{ES}, which is called the *elastic* or *flexural* spectral dimension and is defined by replacing in Eq. (167) the conductivity exponent μ by f, the critical exponent of the elastic moduli, so that (Webman and Grest, 1985)

$$D_{ES} = 2\frac{\nu d - \beta}{2\nu + f - \beta}. \tag{168}$$

Accordingly, we are also forced to have two different types of densities of states for vibrations of fractal networks. One is $N(\omega)$, the density of states when the scalar contributions to the elastic energy of the system dominate all other contributions, for which Eq. (166) is valid. The second type is what we call the *vectorial* or *elastic* density of states $N_e(\omega)$, which is the case when the vectorial contributions, such as those of the central or BB forces, dominate the elastic energy. In this case, in analogy with Eq. (166), the density of states is given by

$$N_e(\omega) \sim \omega^{D_{ES} - 1}. \tag{169}$$

Both Eqs. (166) and (169) are valid for high frequencies (*i.e.*, short wavelengths or length scales) such that $\omega > \omega_{co}$. Vibrational states on fractal structures, expressed by Eqs. (166) and (167), are called *fractons* (phonons on fractals!). Such vibrational states are localized as long as D_S or D_{ES} is less than 2 (Rammal and Toulouse, 1983). Using values of the critical exponents given in Tables 3 and 4, we see that while $D_S \simeq 4/3$ for all $d \geq 2$, D_{ES} varies continuously with d and is *less than 1*.

The crossover between Eqs. (165) and (166) is similar to the crossover between normal and anomalous diffusion that occurs at a time scale t_{co} given by Eq. (98), which corresponds, in the Laplace transform space, to a crossover value of $1/\lambda_{co}$, and since $\omega_{co}^2 \sim \lambda_{co}$, we obtain

$$\omega_{co} \sim (p - p_c)^{\nu + (\mu - \beta)/2} \sim (p - p_c)^{\nu D_W/2}$$

$$\sim \xi_p^{-D_W/2} \sim \xi_p^{-D_F/D_S}. \tag{170}$$

For an L^d system such that $L < \xi_p$, Eq. (170) is equivalent to

$$\omega_{co} \sim L^{-D_W/2} \sim L^{-D_F/D_S}, \tag{171}$$

so that as L increases (but $L/\xi_p \ll 1$), the cutoff or crossover frequency decreases. Alternatively, we can convert Eq. (171) to an equation for a cutoff length scale L_{co} at which a crossover between Eqs. (165) and (166) takes place:

$$L_{co} \sim \omega_{co}^{-D_S/D_F}. \tag{172}$$

These results can be written in a unified scaling form

$$N(\omega) \sim \omega^{D_S - 1} h(\omega/\omega_{co}), \tag{173}$$

where $h(x)$ is a scaling function with the properties that $h(x) \sim x^{d - D_S}$ for $x \ll 1$, and $h(x) \sim$ const. for $x \gg 1$. This also implies that for $L > \xi_p$ we have

$$N(\omega) \sim (p - p_c)^{\nu D_W (D_S - d)/2} \omega^{d-1} \sim \omega_{co}^{D_S - d} \omega^{d-1}. \tag{174}$$

We note here that $N(\omega)$ can also be calculated by an EMA (Sahimi, 1984; Derrida *et al.*, 1984a). The results are $N(\omega) \sim \omega_{co}^{-d/2} \omega^{d-1}$ for the phonon regime and $N(\omega) =$ const. for the fracton regime.

The difference between Eqs. (166) and (169) is significant. Recall that for both 2D and 3D percolation networks at $p_c, D_S \simeq 4/3$, and hence

$$N(\omega) \sim \omega^{1/3}, \tag{175}$$

implying that the medium is mechanically stable [since $N(\omega) \to 0$ as $\omega \to 0$]. Equation (175) is valid for materials in which the scalar contributions dominate their elastic energy. On the other hand, if we use the value of f for 3D BB models, $f \simeq 3.75$, we obtain $D_{ES} \simeq 0.87$, so that

$$N_e(\omega) \sim \omega^{-0.13}. \tag{176}$$

That is, as ω decreases (as the length scales increase) there are a large number of vibrating modes, implying that such systems cannot become too large, because if they do they lose their mechanical stability, and therefore have to restructure themselves into a more stable structure. This type of density of states should be observed in materials in which the contributions of the BB forces dominate their elastic energy.

Feng (1985) argued that in a percolation network in which both the central and BB forces are present, another length scale L_{bb}, in addition to ξ_p, is important, which, according to him, is given by

$$L_{bb} \sim \sqrt{\gamma/\alpha}, \tag{177}$$

where α and γ are, respectively, the central and BB force constants defined by Eq. (143). In effect, L_{bb} is the length scale below which the motion due to central forces is energetically more favorable and above which the BB forces dominate. If $L_{bb} > \xi_p$, then at low frequencies Eq. (165) and at higher frequencies Eq. (166) governs the density of states, with a crossover at ω_{co}. However, if $L_{bb} < \xi_p$, then at low frequencies one is in the phonon regime until a characteristic frequency ω_{bb} is reached, at which there is a crossover from Eq. (165) to Eq. (169). When the higher frequency ω_{co} is reached, there is a second crossover from Eq. (169) to (166). It is not difficult to guess that [see Eq. (171)]

$$\omega_{bb} \sim \xi_p^{-D_F/D_{ES}}. \tag{178}$$

For a review of computer simulation efforts for confirming these predictions see Nakayama *et al.* (1994). Here we discuss a few experimental verifications of the above theoretical predictions.

Most of the experimental studies involved aerogels, which are highly porous solid materials that have a very tenuous structure. Their porosity can be as high as 99%, and for this reason they often have very unusual and unique properties and a wide range of applications, ranging from catalyst supports to thermal insulators and radiators. Among all aerogels silica aerogels have received the widest attention. They appear to have the structure of percolation networks, the evidence for which comes from two different directions. One is the work of Woignier *et al.* (1988) who measured the Young's modulus of silica aerogels as a function of density and volume fraction; their work was already discussed in Sec. 6.4. The other evidence comes from measurements of the fractal dimension of the gel by SANS using Eq. (162). Vacher *et al.* (1988) presented extensive SANS data from which they obtained $D_F \simeq 2.45$, very close to $D_F \simeq 2.52$ for 3D sample-spanning percolation clusters (see Table 3). Therefore, at least one class of silica aerogels have a structure which is more or less similar to that of the percolation cluster at p_c.

Tsujimi *et al.* (1988) carried out very low-frequency depolarized Raman-scattering measurements on fractal silica aerogels. If $I(\omega)$ is the total scattering intensity at frequency ω, then it can be shown that

$$I(\omega)/N(\omega) \sim \omega^{-2(1-2\gamma/D_W)}, \tag{17}$$

where D_W is the fractal dimension of the random walk discussed in Sec. 3.3. Measurements by Tsujimi *et al.* yielded $\gamma > 1$, indicating apparent superlocalization of the vibrational modes (see Sec. 4.4). Another bit of experimental evidence for the above results is provided by verifying the scaling law for the density of states, Eq. (166) or (169). One way of doing this is by plotting the crossover or cutoff frequency ω_{co} as a function of q, the magnitude of the scattering vector. Since q is inversely proportional to the length scale L, Eq. (171) tells us that

$$\omega_{co} \sim q^{D_F/D_S}, \tag{180}$$

so that such a plot should yield $\omega_{\rm co} \sim q^{1.88}$, if silica aerogels are percolation fractals and $D_{\rm S} \simeq 4/3$. Indeed, Vacher *et al.* (1989) presented data obtained with Brillouin scattering of visible light, which indicated that $\omega_{\rm co} \sim q^{1.9}$, in very good agreement with the theoretical prediction. The density of states itself can be measured directly by various methods, and the experiments of Schaefer *et al.* (1990) yielded $D_{\rm S} \simeq 1.22 \pm 0.14$, again consistent with the theoretical expectation.

9. APPLICATION TO LOCALIZATION: QUANTUM PERCOLATION

Electronic transport in disordered media that was discussed in Secs. 2 and 4 is considered classical, because in those systems conduction ceases only when the volume fraction of the conducting component is below the percolation threshold. Anderson localization has motivated the development of a variant of percolation theory, namely, quantum percolation, in which disorder is brought about by a distribution of the matrix elements of a tight-binding representation of the Hamiltonian. As its name suggests, quantum percolation is the quantum counterpart of the classical percolation theory discussed so far, and has proven to be useful in the understanding of wave propagation through a disordered heterogeneous medium and certain modes of electronic transport in metal–insulator composites. For example, as discussed in Sec. 2, if the volume fraction of the nonmetallic component of a metal–insulator composite is large enough, the medium is in the dielectric regime, in which the conduction of charge carriers is *via* thermally activated hopping. Even when the metallic component is sample-spanning, one may not observe the expected metallic behavior if the temperature of the system is low enough.

Similar to rigidity percolation discussed in Sec. 6.1, in quantum percolation the formation of a sample-spanning cluster is not enough for ensuring the existence of a nonzero quantum transmittance. Moreover, as discussed in Sec. 1.3, in classical percolation the dead-end bonds or sites play no important role in flow and transport through the system, whereas in quantum transport (percolation) such bonds or sites do play an important role. When a quantum particle encounters an unoccupied bond or site, scattering takes place, giving rise to the important phenomenon of coherent backscattering (see, for example, Khemel'nitskii, 1984). For large enough scattering strength the classical diffusivity vanishes and electrons become trapped, even if a sample-spanning cluster exists for their transport. This immediately implies that the quantum percolation threshold $p_{\rm cq}$ is *larger than* $p_{\rm c}$, its classical counterpart, and numerical simulations as well theoretical works confirm this.

As is well known, Anderson (1958) proposed that, in addition to the extended states, the electronic states in heterogeneous media may be localized, if the strength of the heterogeneities is strong enough. Such heterogeneities can be morphological, and thus quantum percolation can help one understand the Anderson transition. The study of this localization usually starts with the following tight-binding Hamiltonian (or one of its variants):

$$\mathscr{H} = \sum_i \varepsilon_i a_i^\dagger a_i + \sum_i \sum_j V_{ij} a_i^\dagger a_j. \tag{181}$$

The diagonal site energy terms ε_i are distributed uniformly in the interval $[-W/2, W/2]$. If i and j are nearest neighbors, then $V_{ij} = V$; otherwise $V_{ij} = 0$. Edwards and Thouless (1972) extended the ideas of Anderson by suggesting that the conductance of a piece of a disordered medium depends on the sensitivity of its energy levels to the changes in the boundary conditions. Suppose that $\delta\mathscr{E}$ is the energy difference between the eigenstates, and $\Delta\mathscr{E}$ is the change in the energy levels as the result of changing the boundary conditions from periodic to antiperiodic. Then, it can be shown that $\Delta\mathscr{E}/\delta\mathscr{E}$ is equal to the dimensionless conductivity of the sample. These ideas were further refined by Abrahams *et al.* (1979), who related the energy levels of a system with volume $2L^d$ with those of a system with volume L^d, in order to quantify the perturbations caused by a change in the boundary conditions. They defined a quantity β given by

$$\beta = \frac{d \log G}{d \log L}, \tag{182}$$

and assumed that it is a function of the conductance G alone; $\beta(G)$ turns out to be the relevant scaling function for the Anderson transition. If $\beta(G) > 0$, then $G \to \infty$ as $L \to \infty$, implying that all the states are extended with a continuous spectrum. However, if $\beta(G) < 0$, then $G \to 0$ as $L \to \infty$, which means that the system is insensitive to boundary perturbations, and thus all the states all localized. This scaling theory also suggests that in a disordered medium with $d \leq 2$ all the states are localized, while for $d \geq 3$ they are all extended, depending on the strength of disorder, and therefore $d = 2$ is the lower critical dimension of the problem. This means, in the context of quantum percolation, that for $d \leq 2$ one has to have $p_{\mathrm{cq}} = 1$, whereas for $d \geq 3$ one must have $0 < p_{\mathrm{cq}} < 1$. Numerical studies of quantum percolation do confirm this (in $d = 1$ there is no difference between the quantum and classical percolation). Therefore, let us discuss these studies briefly.

9.1 Numerical Studies of Quantum Percolation

Unlike classical percolation for which one does not need any Hamiltonian, the study of quantum percolation requires a Hamiltonian. The simplest model is one in which the Hamiltonian is given by Eq. (181), and the site energies ε_i are distributed according to a binary distribution:

$$\psi(\varepsilon_i) = p\delta(\varepsilon_i - \varepsilon_{\mathrm{A}}) + (1-p)\delta(\varepsilon_i - \varepsilon_{\mathrm{B}}). \tag{183}$$

Consider the limit $\varepsilon_{\mathrm{B}} \to \infty$. In this case, an electron initially on a site A may propagate across the sample if it can find a continuous path of the A sites or bonds. This is the necessary condition for quantum percolation (Kirkpatrick and Eggarter, 1972). However, a quantum percolation model, as defined by Eqs. (181) and (183), differs from the Anderson model in several aspects. In the latter case disorder is introduced by assuming that the site energies are distributed randomly, whereas in the former case disorder is of binary type. Moreover, the density of states for the quantum percolation model consists of split bands (see, for example, Soukoulis and Grest, 1992) with a central spike and a dip around it where the states are localized. Thus, in this model the Anderson transition can occur only through the merging of at least two pairs of mobility edges, which is different from the usual Anderson model. There are also certain difficulties associated with the moments of the distribution of ε (Odagaki, 1980), which we do not discuss here. Despite such differences, quantum percolation has proven to be very useful for understanding Anderson localization.

Similar to the classical percolation, analytical studies of quantum percolation are very difficult, and thus many numerical methods have been developed for studying quantum percolation. Some of these methods are similar to those developed for the classical percolation. For example, renormalization-group methods, paralleling those for the classical percolation (see Sec. 1.9.3), have been developed for quantum percolation, beginning with the work of Odagaki and Chang (1984) and Root and Skinner (1986). The latter authors also defined a correlation length $\xi_q(p)$ for quantum percolation:

$$\xi_q(p) = \left\langle \sum_j P_{ij} r_{ij} \right\rangle, \tag{184}$$

where the average is taken over all the occupied sites i and all the configurations for a given p. Here P_{ij} is the time-averaged probability that an initial state $|i\rangle$ evolves into state $|j\rangle$; $\xi_q(p)$ diverges as p approaches p_{cq}:

$$\xi_q \sim (p - p_{\mathrm{cq}})^{-\nu_{\mathrm{q}}}, \tag{185}$$

where ν_q is the analog of ν for the classical percolation. Renormalization-group methods for quantum percolation involve the usual errors that are associated with such methods. They are more difficult than those for the classical percolation, but become more accurate as the size of the renormalization cell increases.

Another method of studying quantum percolation is by series expansion and Padé approximation (Shapir *et al.*, 1982; Meir *et al.*, 1986, 1989; Harris and Meir, 1987). Similar to the classical percolation, this method uses the fact that at p_{cq} certain quantities diverge. For example, in classical percolation the mean cluster size diverges as p_{c} is approached. For quantum percolation, one can study a variety of quantities. For example, Meir *et al.* (1989) studied

$$\mathscr{T}(\mathscr{E}) = \int \mathscr{T}(r, \mathscr{E})\mathrm{d}^d r, \tag{186}$$

where $\mathscr{T}(r, \mathscr{E})$ is the average transmittance of a wave with energy $\mathscr{E}$ between two points separated by a distance r. For a localized state one expects to have

$$\mathscr{T}(r, \mathscr{E}) \sim \xi_{\mathrm{q}}^{\gamma_{\mathrm{q}}+d} \sim (p - p_{\mathrm{cq}})^{-\nu_{\mathrm{q}}(\gamma_{\mathrm{q}}+d)}, \tag{187}$$

where γ_{q} is the associated exponent. While this method yielded reasonable results for $d \geq 3$, the results for $d = 2$ were not conclusive. Another method used to study quantum percolation is the slice recursion method, which has been used widely for localization problems, including quantum percolation (Taylor and MacKinnon, 1989). In this method one calculates the Lyapunov exponent, which is the inverse of the localization correlation length ξ_{l}, and by developing a suitable scaling function for it one finds the associated critical exponents. There are several other numerical methods for studying quantum percolation, which are discussed in detail by Mookerjee *et al.* (1995).

Summarizing, the most accurate numerical analysis of quantum percolation indicates that in two dimensions all the states are localized, in agreement with the scaling theory of Abrahams *et al.* (1979), and thus $p_{\mathrm{cq}} = 1$. In $d = 3$ the most recent work (Stadler *et al.*, 1996) indicates that $\nu_{\mathrm{q}} \simeq 1.6$, and on a simple cubic lattice, $p_{\mathrm{cq}} \simeq 0.44$, which, as expected, is larger than the classical percolation threshold $p_{\mathrm{c}} \simeq 0.311$ (see Table 2).

9.2 Crossover Between Anderson Localization and Percolation

The crossover between percolation and Anderson localization, which is relevant to the conductivity and the onset of superconductivity in thin metal–insulator films, is also an important issue. Deutscher *et al.* (1985) introduced two parameters, $z = \xi_{\mathrm{l}}/\xi_p$, where ξ_{l} is the localization correlation length which diverges as $(p - p_{\mathrm{c}})^{-\nu_l}$, and $w = R_0 e^2/L\hbar$, where R_0 is the resistivity of the sample at a minimum length scale L over which it can be defined, $\hbar$ is Planck's constant, and e is the electron's charge. Using the ideas of Abrahams *et al.* (1979) regarding the location of Anderson localization (see above), Deutscher *et al.* (1985) showed that $z = w^{\nu/(\mu-\nu)}(p - p_{\mathrm{c}})^{-\nu_{\mathrm{l}}}$, and proposed the following scenario for the crossover.

1. If $z \ll 1$, then classical percolation dominates and the resistance R of the sample is given by the usual scaling law [Eq. (8)], $R \sim (p - p_{\mathrm{c}})^{-\mu}$.
2. If $z \gg 1$, then localization effects are dominant and $R \sim (p - p_{\mathrm{c}})^{-\nu_{\mathrm{l}}}$.

Deutscher *et al.* (1985) showed that this scenario is completely consistent with the experimental data.

GLOSSARY

Accessible Fraction: The fraction of the microscopic elements of a disordered medium that can be reached from two opposite external surfaces of the medium.

Anomalous Diffusion: A diffusion process in which the mean square displacement of the diffusants varies slower than linearly with the process time.

Backbone: The current- or flow-carrying part of a disordered (percolation) medium.

Capillary Pressure: The difference between the pressures at two points, one on each side of an interface between two immiscible fluids.

Drainage: Displacement of a wetting fluid by a nonwetting fluid in a porous medium.

Effective-Medium Approximation: The idea that a disordered medium may be represented by an ordered medium whose properties mimic those of the disordered medium. These properties are calculated in a self-consistent manner.

Fractals: A word to describe objects with a noninteger dimensionality. Such objects are exactly or statistically scale invariant. This means that there is no length scale associated with the objects, and thus they appear self-similar under magnification.

Fractons: Vibrational modes on fractal objects. Such modes are localized on most fractal structures.

Hysteresis: If a system does not come back to its original state as a control parameter is varied in a cyclic manner, the system is said to show hysteresis. An example is capillary pressure in porous media discussed in this article.

Imbibition: Displacement of a nonwetting fluid by a wetting fluid in a porous medium.

Porosimetry: A method for obtaining the pore size distribution of a porous medium by injecting mercury into the pore space and retracting it, and measuring the pressures that are needed during this process.

Renormalization Group: The group formed by the set of operators used to coarsen a system up to a length scale.

Rigidity Percolation: A percolation model in which each bond of a lattice is a Hookeian spring. The mere connectivity of the springs and their formation of a sample-spanning cluster is not enough for percolation of stress or strain in the lattice.

Scaling Laws: The equations that various properties of a disordered medium follow near its critical point, such as the percolation threshold. Such equations are often in the form of power laws, characterized by critical exponents.

Superlocalization: A quantum state in a disordered medium in which all the wave functions decay with the distance r as $\exp(-r^{\gamma})$ with $\gamma > 1$. In the usual Anderson localization $\gamma = 1$.

Universality: The idea that certain quantities, such as the critical exponents, are independent of microscopic details of a system, and depend only on its dimensionality.

Vector Percolation: A percolation model in which each bond of a lattice is an elastic element, such as a spring or beam. Thus, vector percolation describes transport of vectorial properties in disordered media, such as transport of stress or strain.

Works Cited

Abeles, B., Pinch, H. L., Gittleman, J. I. (1975), *Phys. Rev. Lett.* **35**, 247–250.

Abrahams, E., Anderson, P. W., Licciardello, D. C., Ramakrishnan, T. V. (1979), *Phys. Rev. Lett.* **42**, 673–676.

Adam, M., Delsanti, M., Durand, D. (1985), *Macromolecules* **18**, 2285–2290.

Adam, M., Delsanti, M., Durand, D., Hild, G., Munch, J. P. (1981), *Pure Appl. Chem.* **53**, 1489–1495.

Adam, M., Delsanti, M., Okasha, R., Hild, G. (1979), *J. Phys. (Paris) Lett.* **40**, L539–L544.

Adolf, D., Martin, J. E., Wilcoxon, J. P. (1990), *Macromolecules* **23**, 527–533.

Aharony, A., Harris, A. B. (1990), *Physica* **A163**, 38–43.

Aharony, A., Harris, A. B., Entin-Wohlman, O. (1993), *Phys. Rev. Lett.* **70**, 4160–4163.

Aharony, A., Zhang, Y., Sarachik, M. P. (1992), *Phys. Rev. Lett.* **68**, 3900–3903.

Alderson, K. L., Evans, K. E. (1992), *Polymer* **33**, 4435–4439.

Alexander, S. (1984), *J. Phys. (Paris)* **45**, 1939–1945.

Alexander, S., Bernasconi, J., Orbach, R. (1978), *J. Phys. (Paris) Colloq.* **39 (C6)**, 706–711.

Alexander, S., Orbach, R. (1982), *J. Phys. (Paris) Lett.* **43**, L625–L632.

Alexandrowicz, Z. (1980), *Phys. Lett.* **A80**, 284–288.

Allain, C., Salomé, L. (1987a), *Macromolecules* **20**, 2957–2963.

Allain, C., Salomé, L. (1987b), *Polym. Commun.* **28**, 109–112.

Allain, C., Salomé, L. (1990), *Macromolecules* **23**, 981–987.

Allen, L. C., Golding, B., Haemmerle, W. H. (1988), *Phys. Rev. B* **37**, 913–916.

Ambegaokar, V., Halperin, B. I., Langer, J. S. (1971), *Phys. Rev. B* **4**, 2612–2619.

Anderson, P. W. (1958), *Phys. Rev.* **109**, 1492–1505.

Andrade, J. S., Auto, A. M., Kobayashi, Y., Shibusa, Y., Shirane, K. (1998), *Physica A* **248**, 227–234.

Arbabi, S., Sahimi, M. (1988), *Phys. Rev. B* **38**, 7173–76.

Arbabi, S., Sahimi, M. (1990), *Phys. Rev. Lett.* **65**, 725–728.

Arbabi, S., Sahimi, M. (1993), *Phys. Rev. B* **47**, 695–702.

Balberg, I., Binenbaum, N., and Bozowski, S. (1983), *Solid State Commun.* **47**, 989–992.

Ball, Z., Phillips, H. M., Callahan, D. L., Sauerbrey, R. (1994), *Phys. Rev. Lett.* **73**, 2099–2102.

Belitz, D., Kirkpatrick, T. R. (1994), *Rev. Mod. Phys.* **66**, 261–310.

Ben Avraham, D., Havlin, S. (1982), *J. Phys. A* **16**, L691–L695.

Benguigui, L. (1986), *Phys. Rev. B* **34**, 8176–8179.

Benguigui, L., Ron, P. (1993), *Phys. Rev. Lett.* **70**, 2423–2426.

Bergman, D. J., Dureing, E., Murat, M. (1990), *J. Stat. Phys.* **58**, 1–44.

Bergman, D. J., Imry, Y. (1977), *Phys. Rev. Lett.* **39**, 1222–1225.

Bergman, D. J., Stroud, D. (1985), *Phys. Rev. B* **32**, 6097–7001.

Bernasconi, J. (1974), *Phys. Rev. B* **9**, 4575–4582.

Bernasconi, J. (1978), *Phys. Rev. B* **18**, 2185–2193.

Birovljev, A., Furuberg, L., Feder, J., Jøssang, T., Maloy, K. J., Aharony, A. (1991), *Phys. Rev. Lett.* **67**, 584–587.

Bleibaum, O., Böttger, H., Bryskin, V. V. (1996), *Phys. Rev. B* **54**, 5444–5449.

Blunt, M. J., Scher, H. (1995), *Phys. Rev. E* **52**, 6387–6403.

Böhmer, R., Angell, C. A. (1992), *Phys. Rev. B* **45**, 10091–10097.

Böttger, H., Bryskin, V. V. (1985), *Hopping Conduction in Solids*, Berlin: VCH.

Bouchaud, E., Delsanti, M., Adam, M., Daoud, M., Durand, D. (1986), *J. Phys. (Paris) Lett.* **47**, L539–L546.

Brenig, W., Döhler, G., Wölfe, W. (1971), *Z. Phys.* **246**, 1–8.

Bresser, W., Boolchand, P., Suranyi, P. (1986), *Phys. Rev. Lett.* **56**, 2493–2396.

Broadbent, S. R., Hammersley, J. M. (1957), *Proc. Camb. Philos. Soc.* **53**, 629–639.

Bruggeman, D. A. G. (1935), *Ann. Phys. (Leipzig)* **24**, 636–645.

Bug, A. L. R., Safran, S. A., Grest, G. S., Webman, I. (1985), *Phys. Rev. Lett.* **55**, 1896–1899.

Bunde, A., Dieterich, W., Roman, E. (1985), *Phys. Rev. Lett.* **55**, 5–8.

Bunde, A., Havlin, S. (1996), *Fractals and Disordered Systems*, 2nd ed., Berlin: Springer.

Bunde, A., Havlin, S., Porto, M. (1995), *Phys. Rev. Lett.* **74**, 2714–2716.

Candau, S. J., Ankrim, M., Munch, J. P., Rempp, P., Hild, G., Osaka, R. (1985), in: B. Sedlacek (Ed.), *Physical Optics of Dynamic Phenomena and Processes in Macromolecular Systems*, Proceedings of the 27th Microsymposium on Macromolecules, Prague, July 16–19, 1984, Berlin: De Gruyter.

Chandler, R., Koplik, J., Lerman, K., Willemsen, J. F. (1982), *J. Fluid Mech.* **119**, 249–257.

Chatzis, I., Dullien, F. A. L. (1977), *J. Can. Pet. Tech.* **16**, 97–107.

Chatzis, I., Morrow, N. R., Lim, H. T. (1983), *Soc. Pet. Eng. J.* **23**, 311–321.

Cieplak, M., Robbins, M. O. (1988), *Phys. Rev. Lett.* **60**, 2042–2045.

Cieplak, M., Robbins, M. O. (1990), *Phys. Rev. B* **41**, 11508–11520.

Clerc, J. P., Giraud, G., Laugier, J. M., Luck, J. M. (1990), *Adv. Phys.* **39**, 191–232.

Cody, G. D., Geballe, T. H., Sheng, P., eds. (1990), *Physical Phenomena in Granular Materials*, MRS Proceedings Vol. 195, Pittsburgh: Materials Research Society.

Cohen, M. H., Jortner, J. (1973), *Phys. Rev. Lett.* **30**, 696–699.

Coniglio, A. (1981), *Phys. Rev. Lett.* **46**, 250–253.

Coniglio, A., Stanley, H. E., Klein, W. (1979), *Phys. Rev. Lett.* **42**, 518–521.

Daoud, M., Coniglio, A. (1981), *J. Phys. A* **14**, L301–L306.

Daoud, M., Family, F., Jannik, G. (1984), *J. Phys. (Paris) Lett.* **45**, L199–L206.

Deprez, N., McLachlan, D. S., Sigalas, I. (1989), *Physica* **A157**, 181–189.

Deptuck, D., Harrison, J. P., Zawadzki, P. (1985), *Phys. Rev. Lett.* **54**, 913–916.

Derrida, B., Orbach, R., Yu, K. W. (1984a), *Phys. Rev. B* **29**, 6645–6651.

Derrida, B., Zabolizky, J. G., Vannimenus, J., Stauffer, D. (1984b), *J. Stat. Phys.* **36**, 31–40.

Deutscher, G., Entin-Wohlman, O., Fishman, S., Shapira, Y. (1980), *Phys. Rev. B* **21**, 5041–5047.

Deutscher, G., Goldman, A. M., Micklitz, H. (1985), *Phys. Rev. B* **31**, 1679–1686.

Deutscher, G., Rappaport, M. L. (1979), *J. Phys. (Paris) Lett.* **40**, L219–L225.

Deutscher, G., Rappaport, M. L., Ovadyahu, Z. (1978), *Solid State Commun.* **28**, 593–596.

Djabourov, M., Lebland, J., Papon, P. (1988), *J. Phys. France* **49**, 333–340.

Doyen, P. M. (1988), *J. Geophys. Res.* **93**, 7729–7739.

Dubois, M., Cabane, B. (1989), *Macromolecules* **22**, 2526–2532.

Duering, E., Bergman, D. J. (1990), *J. Stat. Phys.* **60**, 363–373.

Durand, D., Delsanti, M., Adam, M., Luck, J. M. (1987), *Europhys. Lett.* **3**, 297–304.

Dyre, J. C. (1994), *Phys. Rev. B* **49**, 11709–11717.

Edwards, J. T., Thouless, D. J. (1972), *J. Phys. C* **5**, 807–816.

Efros, A. L., Shklovskii, B. I. (1975), *J. Phys. C* **8**, L49–L54.

Efros, A. L., Shklovskii, B. I. (1976), *Phys. Status Solidi b* **76**, 475–478.

Evans, K. E., Caddack, B. D. (1989), *J. Phys. D* **22**, 1883–1890.

Family, F., Coniglio, A. (1980), *J. Phys. A* **13**, L403–L407.

Feng, S. (1985), *Phys. Rev. B* **32**, 5793–5796.

Feng, S., Halperin, B. I., Sen, P. N. (1987), *Phys. Rev. B* **35**, 197–209.

Feng, S., Sen, P. N. (1984), *Phys. Rev. Lett.* **52**, 216–219.

Feng, S., Thorpe, M. F., Garboczi, E. (1985), *Phys. Rev. B* **31**, 276–281.

Feng, X., Bresser, W. J., Boolchand, P. (1997), *Phys. Rev. Lett.* **78**, 4422–4425.

Flory, P. J. (1941), *J. Am. Chem. Soc.* **63**, 3083–3092.

Gadenne, M., Gadenne, P. (1989), *Physica* **A157**, 344–350.

Gadenne, P., Yagil, Y., Deutscher, G. (1989), *Physica* **A157**, 279–286.

Garfunkel, G. A., Weissman, M. B. (1985), *Phys. Rev. Lett.* **55**, 296–299.

Garland, J. C. (1989), *Physica* **A157**, 111–120.

Gauthier-Manuel, B., Guyon, E. (1980), *J. Phys. (Paris) Lett.* **41**, L503–L508.

Gauthier-Manuel, B., Guyon, E., Roux, S., Gits, S., Lefaucheux, F. (1988), *J. Phys. (Paris)* **48**, 869–875.

Gefen, Y., Aharony, A., Alexander, S. (1983), *Phys. Rev. Lett.* **50**, 77–80.

Gerber, A., Deutscher, G. (1989), *Phys. Rev. Lett.* **63**, 1184–1187.

Gingold, D. B., Lobb, C. J. (1990), *Phys. Rev. B* **42**, 8220–8225.

Gordon, J. M., Goldman, A. M. (1988), *Phys. Rev. B* **38**, 12019–12024.

Gouyet, J. F., Rosso, M., Sapoval, B. (1988), *Phys. Rev. B* **37**, 1832–1837.

Grannan, D. M., Garland, J. C., Tanner, D. B. (1981), *Phys. Rev. Lett.* **46**, 375–378.

Grest, G. S., Cohen, M. H. (1983), in: G. Deutscher, R. Zallen, J. Adler (Eds.), *Percolation Structures and Processes*, Bristol: Adam Hilger, p. 187.

Halfpap, B. L., Lindsay, S. M. (1986), *Phys. Rev. Lett.* **57**, 847–850.

Halperin, B. I., Feng, S., Sen, P. N. (1985), *Phys. Rev. Lett.* **54**, 2391–2394.

Hansen, A., Roux, S. (1989), *Phys. Rev. B* **40**, 749–752.

Harris, A. B., Aharony, A. (1987), *Europhys. Lett.* **4**, 1355–1360.

Harris, A. B., Meir, Y. (1987), *Phys. Rev. B* **36**, 1840–1847.

Hashemi, M., Sahimi, M., Dabir, B, (1998), *Phys. Rev. Lett.* **80**, 3248–3251.

Havlin, S., Ben-Avraham, D. (1987), *Adv. Phys.* **36**, 695–798.

He, H., Thorpe, M. F. (1985), *Phys. Rev. Lett.* **54**, 2107–2110.

Heiba, A. A., Davis, H. T., Scriven, L. E. (1983), *Society of Petroleum Engineers (SPE)*, Paper 12172.

Heiba, A. A., Davis, H. T., Scriven, L. E. (1984), *Society of Petroleum Engineers (SPE)*, Paper 12690.

Heiba, A. A., Sahimi, M., Scriven, L. E., Davis, H. T. (1982), *Society of Petroleum Engineers (SPE)*, Paper 11015, New Orleans, Louisiana.

Heiba, A. A., Sahimi, M., Scriven, L. E., Davis, H. T. (1992), *SPE Reservoir Eng.* **7**, 123–131.

Herrmann, H. J., Stanley, H. E. (1984), *Phys. Rev. Lett.* **53**, 1121–1124.

Hodgson, D. F., Amis, E. J. (1990), *Macromolecules* **23**, 2512–2518.

Hoshen, J., Kopelman, R. (1976), *Phys. Rev. B* **24**, 3438–3446.

Hughes, B. D. (1996), *Random Walks and Random Environments*, Oxford: Oxford University Press.

Hulin, J. P., Clément, E., Baudet, C., Gouyet, J. F., Rosso, M. (1988), *Phys. Rev. Lett.* **61**, 333–336.

Hundley, M. F., Zettl, A. (1988), *Phys. Rev. B* **38**, 10290–10295.

Hsu, W. Y., Berzins, T. (1985), *J. Polym. Sci. Polym. Phys. Ed.* **23**, 933–942.

Jacob, D. J., Thorpe, M. F. (1995), *Phys. Rev. Lett.* **75**, 4051–4054.

Jacob, D. J., Thorpe, M. F. (1996), *Phys. Rev. E* **53**, 3682–3691.

Jerauld, G. R., Scriven, L. E., Davis, H. T. (1984), *J. Phys. C* **17**, 3429–3436.

Kantelhardt, J. W., Bunde, A. (1997), *Phys. Rev. E* **56**, 6693–6701.

Kantor, Y., Webman, I. (1984), *Phys. Rev. Lett.* **52**, 1891–1894.

Kapitulnik, A., Deutscher, G. (1982), *Phys. Rev. Lett.* **49**, 1444–1447.

Kapitulnik, A., Deutscher, G. (1983), *J. Phys. A* **16**, L255–L259.

Kapitulnik, A., Hsu, J. W. P., Hahn, M. R. (1990), in: G. D. Cody, T. H. Geballe, P. Sheng (Eds.), *Physical Phenomena in Granular Materials*, MRS Proceedings Vol. 195, Pittsburgh: Materials Research Society.

Katz, A. J., Thompson, A. H. (1986), *Phys. Rev. B* **34**, 8179–8182.

Katz, A. J., Thompson, A. H. (1987), *J. Geophys. Rev. B* **92**, 599–606.

Keating, P. N. (1996), *Phys. Rev.* **152**, 774–784.

Khemel'nitskii, D. E. (1984), *Physica* **B126**, 235–245.

Kirkpatrick, S. (1973), *Rev. Mod. Phys.* **45**, 574–588.

Kirkpatrick, S. (1979), in: R. Balian, R. Maynard, G. Toulouse (Eds.), *Ill-Condensed Matter*, Amsterdam: North-Holland.

Kirkpatrick, S., Eggarter, T. P. (1972), *Phys. Rev. B* **6**, 3598–3607.

Kirkwood, J. G. (1939), *J. Chem. Phys.* **7**, 506–513.

Knackstedt, M. A., Ninham, B. W., Monduzzi, M. (1995), *Phys. Rev. Lett.* **75**, 653–656.

Kogut, P. M., Straley, J. P. (1979), *J. Phys. C* **12**, 2151–2157.

Koplik, J. (1981), *J. Phys. C* **14**, 4821–4829.

Kramer, B., MacKinnon, A. (1993), *Rep. Prog. Phys.* **56**, 1496–1580.

Lafait, J., Tanner, D. B. (Eds.) (1989), special issue on Electrical Transport and Optical Properties of Inhomogeneous Media II, *Physica* **A157**, 1–676.

Laibowitz, R. B., Allessandrini, E. I., Deutscher, G. (1982), *Phys. Rev. B* **25**, 2965–2971.

Lakes, R. (1987), *Science* **235**, 1038–1041.

Lambert, C. J., Hughes, G. D. (1991), *Phys. Rev. Lett.* **66**, 1074–1077.

Landua, L. D., Lifshitz, E. M. (1959), *Theory of Elasticity*, Oxford: Pergamon.

Landauer, R. (1952), *J. Appl. Phys.* **23**, 779–783.

Landauer, R. (1978), in: J. C. Garland, D. B. Tanner (Eds.), *Electrical Transport and Optical Properties of Inhomogeneous Media*, AIP Conference Proceedings No. 40, New York, AIP, pp. 2–43.

Lapp, A., Leibler, L., Schosseler, F., Strazielle, C. (1989), *Macromolecules* **22**, 2871–2877.

Larson, R. G., Davis, H. T., Scriven, L. E. (1981a), *Chem. Eng. Sci.* **36**, 75–86.

Larson, R. G., Morrow, N. R. (1981), *Powder Technol.* **30**, 123–132.

Larson, R. G., Scriven, L. E., Davis, H. T. (1977), *Nature (London)* **268**, 409–412.

Larson, R. G., Scriven, L. E., Davis, H. T. (1981b), *Chem. Eng. Sci.* **36**, 57–70.

Lauer, U., Maier, J. (1992), *Ber. Bunsenges. Phys. Chem.* **96**, 111–120.

Leath, P. L. (1976), *Phys. Rev. B* **14**, 5046–5053.

Lee, P. A., Ramakrishnan, T. V. (1985), *Rev. Mod. Phys.* **57**, 287–324.

Lee, S.-I., Song, Y., Noh, T. W., Chen, X.-D., Gaines, J. R. (1986), *Phys. Rev. B* **34**, 6719–6724.

Leibler, L., Schosseler, F. (1985), *Phys. Rev. Lett.* **55**, 1110–1113.

Lenormand, R., Bories, S. (1980), *C. R. Acad. Sci. Paris* **B291**, 279–285.

Lenormand, R., Zarcone, C. (1984), in: F. Family, D. P. Landau (Eds.), *Kinetics of Aggregation and Gelation*, Amsterdam: Elsevier.

Lenormand, R., Zarcone, C. (1985), *Phys. Rev. Lett.* **54**, 2226–2229.

Levinshtein, M. E., Shur, M. S., Efros, A. L. (1975), Zh. Eksp. Teor. Fiz. **69**, 2203–2211 [Sov. Phys. JETP **42**, 1120–1128].

Lévy, Y.-E., Souillard, B. (1987), *Europhys. Lett.* **4**, 233–236.

Liang, N. T., Shan, Y., Wang, S.-Y. (1976), *Phys. Rev. Lett.* **37**, 526–529.

Liu, H., Zhang, L., Seaton, N. A. (1992), *Chem. Eng. Sci.* **47**, 4393–4350.

Lobb, C. J., Frank, D. J. (1979), in: D. U. Gubser, T. L. Francavilla, J. R. Laibowitz, S. A. Wolf (Eds.), *Inhomogeneous Superconductors*, AIP Conference Proceedings, No. 58, New York: AIP, pp. 308–311.

Lobb, C. J., Forrester, M. G. (1987), *Phys. Rev. B* **35**, 1899–1902.

Lubensky, T. C., Isaacson, J. (1978), *Phys. Rev. Lett.* **41**, 829–832.

Lucena, L. A., Araujo, J. M., Tavares, D. M., da Silva, L. R., Tsallis, C. (1994), *Phys. Rev. Lett.* **72**, 230–233.

Mahadevan, S., Giridhar, A. (1989), *J. Non-Cryst. Solids* **110**, 118–125.

Mall, S., Russel, W. B. (1987), *J. Rheol.* **31**, 651–661.

Maloufi, N., Audouard, A., Piecuch, M., Vergant, M., Marchal, G., Gerl, M. (1988), *Phys. Rev. B* **37**, 8867–8871.

Mandal, P., Neumann, A., Jansen, A. G., Wyder, P., Deltour, R. (1997), *Phys. Rev. B* **55**, 452–457.

Martin, J. E., Adolf, D., Wilcoxon, J. P. (1988), *Phys. Rev. Lett.* **61**, 2620–2623.

Martin, J. E., Wilcoxon, J. P. (1988), *Phys. Rev. Lett.* **61**, 373–376.

Mehbod, M., Wyder, P., Deltour, R., Pierre, C., Geuskens, G. (1987), *Phys. Rev. B* **36**, 7627–7632.

Mehrabi, A. R., Sahimi, M. (1997), *Phys. Rev. Lett.* **79**, 4385–4388.

Meir, Y. (1996), *Phys. Rev. Lett.* **77**, 5265–5267.

Meir, Y., Aharony, A., Harris, A. B. (1986), *Phys. Rev. Lett.* **56**, 976–979.

Meir, Y., Aharony, A., Harris, A. B. (1989), *Europhys. Lett.* **10**, 275–286.

Miller, A., Abrahams, E. (1960), *Phys. Rev.* **120**, 745–753.

Mookerjee, A., Dasgupta, I., Saha, T. (1995), *Int. J. Mod. Phys. B* **9**, 2989–3024.

Mott, N. F. (1968), *J. Non-Cryst. Solids* **8–10**, 1–8.

Moukarzel, C., Duxbury, P. M. (1995), *Phys. Rev. Lett.* **75**, 4055–4058.

Movaghar, B., Pohlmann, B., Sauer, G. W. (1980), *Phys. Status Solidi* **B97**, 533–540.

Mukhopadhyay, S., Sahimi, M. (1994), *J. Stat. Phys.* **74**, 1301–1308.

Nakayama, T., Yakubo, K., Orbach, R. L. (1994), *Rev. Mod. Phys.* **66**, 381–440.

Nicklasson, G. A., Grangvist, C. G. (1984), *J. Appl. Phys.* **55**, 3382–3387.

Odagaki, T. (1980), *Solid State Commun.* **33**, 861–864.

Odagaki, T., Chang, K. C. (1984), *Phys. Rev. B* **33**, 1612–1615.

Odagaki, T., Lax, M. (1981), *Phys. Rev. B* **24**, 5284–5290.

Ortuno, M., Pollak, M. (1983), *Philos. Mag.* **47**, L93–L96.

Ottavi, H., Clerc, J. P., Giraud, G., Roussenq, J., Guyon, E., Mitescu, C. D. (1978), *J. Phys. C* **11**, 1311–1317.

Palevski, A., Rappaport, M. L., Kapitulnik, A., Fried, A., Deutscher, G. (1984), *J. Phys. (Paris) Lett.* **45**, L367–L373.

Papandreou, N., Nédellec, P. (1992), *J. Phys. I France* **2**, 707–724.

Parisi, G., Sourlas, N. (1981), *Phys. Rev. Lett.* **46**, 891–894.

Parris, P. E. (1998), *J. Chem. Phys.* **108**, 218–226.

Patton, E., Wesson, J. A., Rubinstein, M., Wilson, J. E., Oppenheimer, L. E. (1989), *Macromolecules* **22**, 1946–1953.

Phillips, J. C. (1979), *J. Non-Cryst. Solids* **34**, 153–164.

Phillips, J. C., Thorpe, M. F. (1985), *Solid State Commun.* **53**, 699–702.

Pollak, M. (1972), *J. Non-Cryst. Solids* **11**, 1–10.

Pollak, M. (1978), in: L. R. Friedman, D. P. Tunstall (Eds.), *The Metal-Non-metal Transition in Disordered Systems*, Edinburgh: University of Edinburgh Press.

Porto, M., Havlin, S., Schwarzer, S., Bunde, A. (1997), *Phys. Rev. Lett.* **79**, 4060–4062.

Prakash, S., Havlin, S., Schwartz, M., and Stanley, H. E. (1992), *Phys. Rev. A* **46**, R1724–R1727.

Rammal, R., Toulouse, G. (1983), *J. Phys. (Paris) Lett.* **44**, L13–L19.

Redfield, D. (1973), *Phys. Rev. Lett.* **30**, 1319–1322.

Reynolds, P. J., Stanley, H. E., Klein, W. (1980), *Phys. Rev. B* **21**, 1223–1245.

Robin, T., Souillard, B. (1989), *Physica* **A157**, 285–290.

Roman, H. E. (1990), *J. Phys.: Condens. Matter* **2**, 3909–3917.

Roman, H. E., Bunde, A., Dieterich, W. (1986), *Phys. Rev. B* **34**, 3439–3445.

Root, L. J., Skinner, J. L. (1986), *Phys. Rev. B* **33**, 7738–7743.

Roux, S. (1986), *J. Phys. A* **19**, L351–L354.

Roux, S,. Guyon, E. (1985), *J. Phys. (Paris) Lett.* **46**, L999–L1003.

Sahimi, M. (1984), *J. Phys. C* **17**, 3957–3966.

Sahimi, M. (1986a), *J. Phys. C* **19**, L79–L83.

Sahimi, M. (1986b), *J. Phys. C* **19**, 1311–1316.

Sahimi, M. (1987), *J. Phys. A* **20**, L1293–L1298.

Sahimi, M. (1988), *Chem. Eng. Commun.* **64**, 179–195.

Sahimi, M. (1993a), *AIChE J.* **39**, 369–386.

Sahimi, M. (1993b), *Rev. Mod. Phys.* **65**, 1395–1534.

Sahimi, M. (1994), *Applications of Percolation Theory*, London: Taylor and Francis.

Sahimi, M. (1995), *Flow and Transport in Porous Media and Fractured Rock*, Weinheim, Germany: VCH.

Sahimi, M. (1998), *Phys. Rep.* **306**, 213–395.

Sahimi, M., Arbabi, S. (1989), *Phys. Rev. B* **40**, 4975–4980.

Sahimi, M., Arbabi, S. (1993), *Phys. Rev. B* **47**, 703–712.

Sahimi, M., Arbabi, S. (1996), *Phys. Rev. Lett.* **77**, 3689–3692.

Sahimi, M. Goddard, J. D. (1985), *Phys. Rev. B* **32**, 1869–1872.

Sahimi, M., Heiba, A. A., Davis, H. T., Scriven, L. E. (1986), *Chem. Eng. Sci.* **41**, 2123–2136.

Sahimi, M., Hughes, B. D., Scriven, L. E., Davis, H. T. (1983a), *J. Chem. Phys.* **78**, 6849–6862.

Sahimi, M., Hughes, B. D., Scriven, L. E., Davis, H. T. (1983b), *Phys. Rev. B* **28**, 307–311.

Sahimi, M., Mukhopadhyay, S. (1996), *Phys. Rev. E* **54**, 3870–3880.

Sahimi, M., Robertson, M. C., Sammis, C. G. (1993), *Phys. Rev. Lett.* **70**, 2186–2189.

Sahimi, M., Scriven, L. E., Davis, H. T. (1984), *J. Phys. C* **17**, 1941–1948.

Sapoval, B., Rosso, M., Gouyet, J. F. (1985), *J. Phys. (Paris) Lett.* **46**, L149–L157.

Schaefer, D. W., Brinker, C. J., Richter, D., Farago, B., Frick, B. (1990), *Phys. Rev. Lett.* **64**, 2316–2319.

Scher, H., Zallen, R. (1970), *J. Chem. Phys.* **53**, 3759–3761.

Schwartz, L. M., Feng, S., Thorpe, M. F., Sen, P. N. (1985), *Phys. Rev. B* **32**, 4607–4614.

Schwartz, L. M., Johnson, D. L., Feng, S. (1984), *Phys. Rev. Lett.* **52**, 831–834.

Schmidt, W., Burchard, W. (1981), *Macromolecules* **14**, 370–376.

Seaton, N. A. (1991), *Chem. Eng. Sci.* **46**, 1895–1906.

Shante, V. K. S. (1977), *Phys. Rev. B* **16**, 2597–2607.

Shante, V. K. S., Kirkpatrick, S. (1971), *Adv. Phys.* **20**, 325–336.

Shapir, Y., Aharony, A., Harris, A. B. (1982), *Phys. Rev. Lett.* **49**, 486–489.

Shaw, T. M. (1987), *Phys. Rev. Lett.* **59**, 1671–1674.

Sheng, P., Klafter, J. (1983), *Phys. Rev. B* **27**, 2583–2586.

Shklovskii, B. I. (1977), *Zh. Eksp. Teor. Fiz.* **72**, 288–291 [*Sov. Phys.-JETP* **45**, 152–155].

Shklovskii, B. I. (1978), *Phys. Status Solidi b* **85**, K111–K114.

Shklovskii, B. I., Efros, A. L. (1971), *Zh. Eksp. Teor. Fiz.* **60**, 867–872 [*Sov. Phys.-JETP* **33**, 468–473].

Shklovskii, B. I., Efros, A. L. (1975), *Pis'ma Zh. Eksp. Teor. Fiz.* **1**, 174–177 [*Sov. Phys.-JETP Lett.* **1**, 83–86].

Shklovskii, B. I., Efros, A. L. (1984), *Electronic Properties of Doped Semiconductors*, Berlin: Springer.

Sichel, E. K., Gittleman, J. I. (1982), *Solid State Commun.* **42**, 75–78.

Smith, L. N., Lobb, C. J. (1979), *Phys. Rev. B* **20**, 3653–3656.

Sofo, J., Lorenzana, J., Martinez, E. N. (1987), *Phys. Rev. B* **36**, 3960–3964.

Song, Y., Noh, T. W., Lee, S.-I., Gaines, J. R. (1986), *Phys. Rev. B* **33**, 904–909.

Soukoulis, C. M., Grest, G. S. (1992), *Phys. Rev. B* **45**, 7724–7730.

Stadler, A. W., Kusy, A., Sikora, R. (1996), *J. Phys.: Condens. Matter* **8**, 2981–2990.

Stanley, H. E. (1977), *J. Phys. A* **11**, L211–L217.

Stanley, H. E. (1979), *J. Phys. A* **12**, L329–L334.

Stanley, H. E., Teixeira, J., Geiger, A., Blumberg, R. L. (1981), *Physica A* **106**, 260–271.

Stauffer, D., Aharony, A. (1992), *Introduction to Percolation Theory*, 2nd ed., London: Taylor and Francis.

Stauffer, D., Coniglio, A., Adam, M. (1982), *Adv. Polymer Sci.* **44**, 105–132.

Stinchcombe, R. B., Watson, B. P. (1976), *J. Phys. C* **9**, 3221–3230.

Straley, J. P. (1976), *J. Phys. C* **9**, 783–789.

Straley, J. P. (1980a), *J. Phys. C* **13**, L773–L776.

Straley, J. P. (1980b), *J. Phys. C* **13**, 4335–4340.

Straley, J. P. (1982), *J. Phys. C* **15**, 2343–2349.

Stockmayer, W. H. (1943), *J. Chem. Phys.* **11**, 45–59.

Szpilka, A. M., Viščor, P. (1982), *Philos Mag.* **45**, 485–490.

Takahashi, M., Yokoyama, K., Masuda, T. (1994), *J. Chem. Phys.* **101**, 798–805.

Tatsumisago, M., Halfpap, B. L., Green, J. L., Lindsay, S. M., Angell, C. A. (1990), *Phys. Rev. Lett.* **64**, 1549–1552.

Taylor, J. P. G., MacKinnon, A. (1989), *J. Phys.: Condens. Matter*, **1**, 9963–9970.

Tessler, L. R., Deutscher, G. (1989), *Physica* **A157**, 154–160.

Thorpe, M. F. (1983), *J Non-Cryst. Solids* **57**, 355–363.

Tokita, M., Niki, R., Hikichi, K. (1984), *J. Phys. Soc. Jpn.* **53**, 480–482.

Trappe, V., Richtering, W., Burchard, W. (1992), *J. Phys. II France* **2**, 1453–1459.

Troadec, J. P., Bideau, D., Guyon, E. (1981), *J. Phys. C* **14**, 4807–4814.

Tsujimi, Y., Courtens, E., Pelous, J., Vacher, R. (1988), *Phys. Rev. Lett.* **60**, 2757–2760.

Uri, D., Palevski, A., Deutscher, G. (1987), *Phys. Rev. B* **36**, 790–795.

Vacher, R., Woignier, T., Pelous, J., Courtens, E. (1988), *Phys. Rev. B* **37**, 6500–6509.

Vacher, R., Woignier, T., Pelous, J., Coddens, G., Courtens, E. (1989), *Europhys. Lett.* **8**, 161–165.

van der Marck, S. C. (1996), *Phys. Rev. Lett.* **77**, 1785–1788.

van der Putten, D., Moonen, J. T., Brom, H. B., Brokken-Zijp, J. C. M., Michels, M. A. J. (1992), *Phys. Rev. Lett.* **69**, 494–497.

Voss, R. F., Laibowitz, R. B., Allessandrini, E. I. (1982), *Phys. Rev. Lett.* **49**, 1441–1444.

Webman, I., Grest, G. (1985), *Phys. Rev. B* **31**, 1689–1692.

Weinrib, A. (1984), *Phys. Rev. B* **29**, 387–395.

Wilkinson, D. (1984), *Phys. Rev. A* **30**, 520–531.

Wilkinson, D., Barsony, M. (1984), *J. Phys. A* **17**, L129–L134.

Wilkinson, D., Willemsen, J. F. (1983), *J. Phys. A* **16**, 3365–3372.

Winterfeld, P. H., Scriven, L. E., Davis, H. T. (1981), *J. Phys. C* **14**, 2361–2376.

Woignier, T., Phalippou, J., Sempere, R., Pelons, J. (1988), *J. Phys. France* **49**, 89–94.

Yagil, Y., Deutscher, G. (1988), *Appl. Phys. Lett.* **52**, 373–376.

Yun, S. S., Li, H., Cappelletti, R. L., Enzweiler, R. N., Boolchand, P. (1989), *Phys. Rev. B* **39**, 8702–8709.

Zabolitzky, J. G., Bergman, D. J., Stauffer, D. (1986), *J. Stat. Phys.* **44**, 211–219.

Zhang, Y., Dai, P., Levy, M., Sarachik, M. P. (1990), *Phys. Rev. Lett.* **64**, 2687–2690.

Further Reading

The Works Cited represent a very small fraction of the vast literature on percolation theory and its applications. Most papers that deal with percolation problems appear in *Physical Review B, Physical Review E, Physical Review Letters, Journal of Physics A* and *Journal of Statistical Physics*. For a simple but extensive discussion of theoretical foundations of percolation theory, see Stauffer and Aharony (1992). Bunde and Havlin (1996) and Hughes (1996) provide more theoretical discussion of percolation theory. For extensive discussions of various applications of percolation theory see Sahimi (1994). A comprehensive review of electrical and dielectrical properties of composite solids is given by Clerc *et al.* (1990). Porous-media problems are extenstively discussed in Sahimi (1993b, 1995). Anomalous diffusion in disordered media is discussed exhaustively by Havlin and Ben-Avraham (1987). Very detailed discussions of hopping conductivity are given by Shklovskii and Efros (1984), and by Böttger and Bryskin (1985). Applications of percolation theory to modeling of elastic and viscoelastic properties of composites, polymers, ceramics, and glasses, as well as nonlinear composites, are discussed extensively by Sahimi (1998). Localization and quantum percolation are reviewed by Lee and Ramakrishnan (1985), Kramer and MacKinnon (1993) and Belitz and Kirkpatrick (1994).

PERIODIC SURFACES IN PHYSICS, CHEMISTRY AND BIOLOGY

Robert Hołyst, *Institute of Physical Chemistry and College of Science, Polish Academy of Sciences, Warsaw, Poland*

Wojciech Góźdź, *Institute of Physical Chemistry and College of Science, Polish Academy of Sciences, Warsaw, Poland, and Max-Planck Institut für Kolloid und Grenzflächenforschung, Teltow-Seehof, Germany*

Introduction: Historical Survey .. 145
1. **Types of Periodic Surfaces** 146
1.1 Minimal and Constant-Mean-Curvature Surfaces 147
1.2 Zero-Potential and Equipotential Surfaces.......... 149
1.3 Nodal Surfaces................. 149
2. **Experimental Observation of Periodic Surfaces**............... 150
2.1 Diffusion on Surfaces........... 150
2.2 Amphiphilic Solutions.......... 150
2.3 Copolymers.................... 151
2.4 Chemical Structures............ 152
2.5 Periodic Surfaces in Living Matter......................... 152
3. **Functional Generation of Periodic Surfaces**............... 152
3.1 Landau–Ginzburg Model...... 153
3.2 Minimization Procedure in Real and Fourier Space 153
3.3 High-Genus Periodic Surfaces.. 154
3.4 Multiply Continuous Structures 154
4. **Applications** 155
4.1 Mesoporous Sieves and Catalysts....................... 156
4.2 Soft Contact Lenses 156
4.3 Crystallization of Membrane Proteins on Periodic Surfaces... 157
Glossary 157
Works Cited................... 158
Further Reading 160

INTRODUCTION: HISTORICAL SURVEY

A periodic surface is a surface that moves onto itself under a unit translation in one, two or three coordinate directions, similarly as in the periodic arrangement of atoms in regular crystals. The most interesting are triply periodic surfaces, which are periodic in all three directions forming structures that have various crystallographic symmetries. Here we mostly concentrate on smooth surfaces with cubic symmetry.

The paradigm structures for all periodic surfaces are minimal periodic surfaces. Such surfaces have zero mean curvature at every point of the surface, as was shown by Meusnier (Nitsche, 1989) and Laplace in the XVIII century. The simple patch of the minimal surface can be visualized in a simple experiment of Plateau (Belgian physicist, XIX century): use a soap solution and dip a metal frame (not necessarily planar) in it. The film that forms on the frame will adopt the shape such as to minimize the surface free energy, *i.e.*, as to minimize the area of the surface (hence the name minimal surface).

The first periodic (in one direction only) minimal surface discovered in 1776 (Hoffman, 1996) was a helicoid: the surface swept out by a horizontal line rotating at a constant rate as it moves at a constant speed up a vertical axis. The next example (periodic in two directions) was discovered in 1830 by Herman Scherk. The first triply periodic minimal surface was discovered by Herman Schwarz in 1865. The revival of interest in periodic surfaces was due to the observation (Luzzati and Spegt, 1967; Luzzati *et al.*, 1968a,b,c) that bilayers of lipids in water solutions form triply periodic surfaces under suitable thermodynamic conditions, and the discovery of new triply periodic minimal surfaces (among them the gyroid phase) by the mathematician Schoen (1970). If we draw a surface through the middle of a triply periodic lipid bilayer it follows from geometrical con-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

straints that it must be a triply periodic minimal surface (Charvolin and Sadoc, 1987).

The self-assembled structures formed by lipids or surfactants in water solutions are used nowadays as templates for the three-dimensional polymerization reaction leading to mesoporous sieves (Davis, 1993), contact lenses (Anderson and Ström, 1991), or bulk catalysts, and this adds a technological dimension to the study of periodic surfaces. The formation of periodic surfaces in etioplast in plants and the usage of periodic surfaces for the crystallization of high–molecular-weight membrane proteins (de Kruijff, 1997) shows that these surfaces are related to biological problems. After the discovery of various structures formed by carbon—*i.e.*, wires and fullerenes—it has been also shown theoretically that the carbon atoms can be arranged to form triply periodic surfaces (Terrones and Mackay, 1993; Hyde and O'Keeffe, 1996). In ionic crystals the points at which the electrostatic potential is constant form a periodic surface; its symmetry can be determined by the symmetry-adapted distribution of charges (von Schnering and Nesper, 1987). It has also been shown that there is a very strong connection between chemical structures and periodic surfaces (Klinowski *et al.*, 1996; Mackay and Klinowski, 1986).

1. TYPES OF PERIODIC SURFACES

Triply periodic surfaces can be nodal (Mackay, 1993, 1994, 1995; von Schnering and Nesper 1991), equipotential surfaces (von Schnering and Nesper, 1987; Mackay, 1988), and minimal surfaces. They are described below in terms of their geometrical properties and generating equations.

Surfaces are characterized locally by the mean (H) and Gaussian (K) curvature: K and H can be calculated according to $H = \frac{1}{2}(1/R_1 + 1/R_2)$, $K = 1/R_1R_2$, where R_1 and R_2 are principal radii of curvature (Fig. 1). From the point of view of physics, especially important are those of constant mean curvature, since they are widespread in nature. A soap film spanned on a wire frame is a surface with zero mean curvature, and a soap bubble is a surface of constant mean curvature. The curvature of the surface is related to thermodynamic quantities by the Poisson–Laplace equation $H = \Delta P/2\sigma$, where H is the mean curvature at the interface of two homogeneous media, ΔP is the pressure difference across the interface, and σ is the surface tension. The notion of curvature was used to develop the effective Hamiltonian (Canham, 1970; Helfrich, 1973)

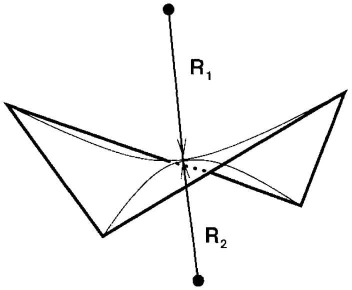

FIG. 1. Piece of surface with nonpositive Gaussian curvature. R_1 and R_2 are the principal radii. The Gaussian (K) and the mean (H) curvatures are expressed in terms of the principal radii as follows: $H = 1/2R_1 + 1/2R_2$, $K = 1/R_1R_2$. If $R_1 = -R_2$, at every point, the surface is called minimal. This implies that K is nonpositive at every point.

$$\mathscr{H} = \kappa \int dS(H - H_0)^2 + \bar{\kappa} \int dS\,K, \qquad (1)$$

where κ and $\bar{\kappa}$ are bending and Gaussian rigidities, H is the mean curvature, H_0 is the spontaneous curvature, and K is the Gaussian curvature. The integration is over the entire surface. The effective Hamiltonian has been successfully used in studies of lipid membranes (Lipowsky, 1993; Gompper and Schick, 1994).

Every triply periodic surface can be characterized by genus, an integer number that describes the number of holes or handles in the surface. The genus of a surface is related to its local properties by the Gauss–Bonnet theorem:

$$\int dS\,K = 2\pi\chi = 4\pi(1 - g), \qquad (2)$$

where K is the Gaussian curvature, χ is the Euler–Poincaré characteristic, and g is the genus. It is possible that different surfaces have the same genus. The genera for the best known triply periodic surfaces, reported for a translational unit cell, are 3 for the Schwarz (P), 5 for

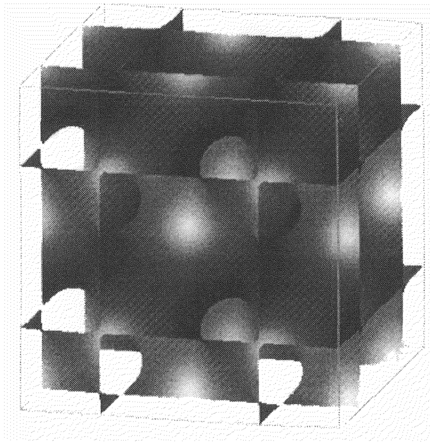

FIG. 2. Schwarz's minimal surface *D*.

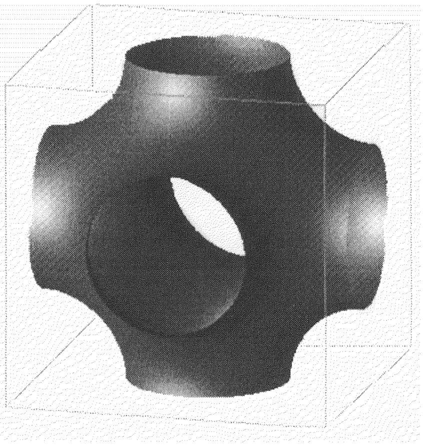

FIG. 4. Schwarz's minimal surface *P*.

the Schoen (G) and 9 for the Schwarz (D) surfaces (see Figs. 2–4). Please note that the value of the genera for P, D, and G is the same for a lattice fundamental region (different from the unit cell) (Anderson *et al.*, 1990). The latter fact should be expected on the basis of the local isometry of these surfaces.

Periodic surfaces can also be characterized by distribution of normal vectors. In the case of polymerized surfactant membranes, such a distribution is accessible experimentally *via* ^{2}H NMR technique (Anderson, 1990b; Góźdź and Hołyst, 1997). An example of the normal vector distribution for the P, D and G surfaces is given in Fig. 5.

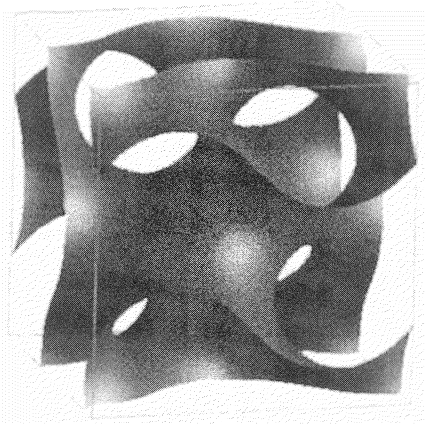

FIG. 3. Schoen's minimal surface *G*.

1.1 Minimal and Constant–Mean-Curvature Surfaces

Triply periodic minimal surfaces (TPMS) are surfaces of constant mean curvature equal to zero. The name "minimal" comes from the fact that the surface spanned on a frame has

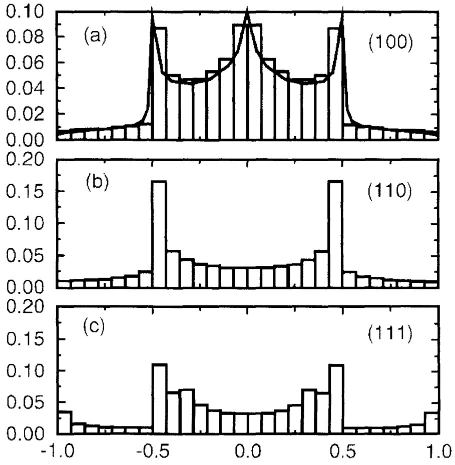

FIG. 5. The normalized NMR distribution function for *P*, *D*, and *G* surfaces (same for all) calculated numerically (Góźdź and Hołyst, 1997) and analytically (Anderson, 1990b) for the following directions of the magnetic field $\boldsymbol{B}$: (a) (100); the solid line shows the results of analytic calculations (Anderson, 1990b). (b) (110). (c) (111). $x = [3\cos^2(\theta_{nB}) - 1]/2$. Here θ_{nB} is the angle between the normal to the surface and the magnetic field. For a sphere this distribution (given for simplicity in θ_{nB}) is $f(\theta_{nB}) = 0.5\sin(\theta_{nB})$.

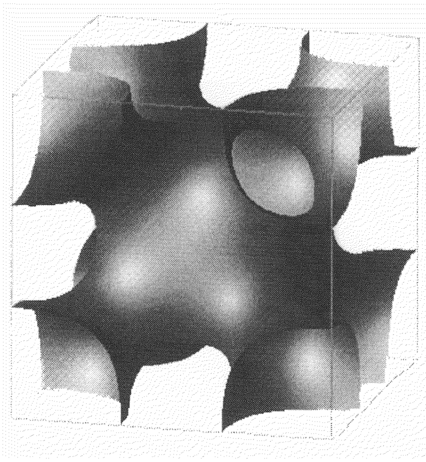

FIG. 6. Schoen's minimal surface *I-WP*.

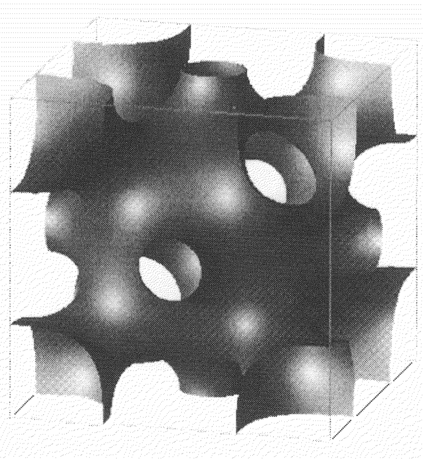

FIG. 7. Schoen's minimal surface *O-CT*0.

the lowest, "minimal" surface area if it is the surface of zero mean curvature. TPMS are infinite and are not bounded by a frame, but the mean curvature for these surfaces is zero at every point on the surface. That is why they are called "minimal". The condition of zero mean curvature implies that the Gaussian curvature is nonpositive at every point. The points where the Gaussian curvature is zero are called the flat points. These points are isolated.

Most of the TPMS divide the volume into two equal subvolumes (balanced surfaces; Fischer and Koch, 1986, 1996) (*P*, *D*, and *G* are the most prominent examples). There are also TPMS that divide volume into unequal subvolumes (unbalanced surfaces) (*I-WP*, *O-CT*0); see Figs. 6 and 7.

TPMS are described locally by the Enneper–Weierstrass equations (Schoen 1970; Terrones, 1990):

$$x = \tau \operatorname{Re} \int_{\omega_0}^{\omega_1} \mathrm{d}\omega \, e^{i\theta} R(\omega)(1-\omega^2),$$

$$y = \tau \operatorname{Re} \int_{\omega_0}^{\omega_1} \mathrm{d}\omega \, e^{i\theta} R(\omega) i(1+\omega^2), \tag{3}$$

$$z = \tau \operatorname{Re} \int_{\omega_0}^{\omega_1} \mathrm{d}\omega \, e^{i\theta} R(\omega) 2\omega,$$

where Re denotes the real part of the complex integral and integration is carried on an arbitrary path from ω_0 to ω_1 in the complex plane. The Enneper–Weierstrass equations guarantee that the surface is minimal but not necessarily embedded, *i.e.*, free of self-intersections. $R(\omega)$ is called the Weierstrass function and is characteristic for a given minimal surface. The parameter τ sets the length scale. The angle θ is called the Bonnet angle. The transformation that results from the change of θ is called the Bonnet transformation. Surfaces related by Bonnet transformation are called associate. The Bonnet transformation is isometric and conformal: all the lengths and angles are preserved. The surface is only bent without stretching. The surfaces related by Bonnet angle equal to 90° are called adjoint or conjugate surfaces, since they are described by adjoined (conjugate) complex functions. *D*, *P*, and *G* are associate surfaces with the Weierstrass function $R(\omega) = (1 - 14\omega^4 + \omega^8)^{-1/2}$. The Bonnet angles are respectively $\theta = 0°, 90°, 38.015°$ (Schoen, 1970). *P* and *D* are adjoint surfaces.

The properties of the TPMS follow uniquely from the Weierstrass function $R(\omega)$ (Terrones, 1990). $R(\omega)$ is known only for a few TPMS of simple topology. It is possible to construct $R(\omega)$ if there exists a surface patch from which the whole TPMS can be generated by reflection or rotation of the patch boundary, provided the surface belongs to the so-called regular class (Fogden, 1990, 1992; Fogden *et al.*, 1993; Fogden and Haeberlein, 1994;

Cvijović and Klinowski, 1992a,b,c). Not every surface has this property.

TPMS can be also constructed by the conjugate-surface method invented by Schoen (1970) (see also Karcher, 1989; Karcher and Polthier, 1996). The idea is to convert the free-boundary problem to the Plateau problem by using a Bonnet transformation and solve the simpler Plateau problem next. The method has been applied to prove the existence of many conjectured TPMS (Karcher, 1989) and to generate new examples by so called handle insertion (Schoen, 1970; Karcher and Polthier 1996). The conjugate-surface method has been applied to low-genus simple surfaces, but in principle can be used to generate high-genus surfaces.

The analysis of the crystallographic-group rotational symmetries led to the discovery of several new TPMS (Fisher and Koch, 1986, 1996; Koch and Fischer 1988). They deduced a number of group–subgroup pairs of space groups compatible with TPMS. Similar analysis has been done in the Fourier space (Sheng, 1994; Sheng and Elser 1994), leading to the discovery of a new minimal surface. The Fourier space is especially convenient in studies of symmetries of TPMS (Mermin, 1992).

TPMS belong to the broader class of periodic surfaces of nonzero constant mean curvature (CMC). The existence of periodic CMC surfaces was proven in 1970 (Lawson, 1970) for two doubly periodic surfaces of hexagonal and square symmetry. The first triply periodic CMC surfaces were calculated numerically for *P*, *D*, *I-WP* *F-RD*, *C*(*P*), and S'-S'' surfaces (Anderson *et al.*, 1990). The proof of existence was given by Karcher (1989) by the solution of the analogous Plateau problem in S^3. Recently numerical calculations have been done, using the Surface Evolver program (Brakke, 1992), for the *G* surface (Grosse–Brauckmann, 1997).

The Surface Evolver presents a convenient way of generating minimal and CMC surfaces. It minimizes the integral

$$\int (H - H_0)^2 \mathrm{d}S \tag{4}$$

for a discretized surface, *i.e.*, a surface given as a set of triangles.

1.2 Zero-potential and Equipotential Surfaces

Periodic equipotential surfaces are the level surfaces of the electric potential field generated by a periodic, discrete distribution of point charges in the space. The potential $\Omega(\boldsymbol{r})$ at a point $\boldsymbol{r} = (x, y, z)$ is defined according to

$$\Omega(\boldsymbol{r}) = \sum_{i,j,k}^{\infty} q(i,j,k)|\boldsymbol{r} - (i,j,k)|^{-1}, \tag{5}$$

where $q(i,j,k)$ is a point charge at the lattice site (i,j,k).

It turns out that many surfaces $\Omega(\boldsymbol{r}) = 0$ defined for charge distributions characteristic for ionic crystals of chemical compounds have very similar topology to TPMS (von Schnering and Nesper, 1991). The zero-potential surface for the CsCl ionic crystal charge distribution shares the same topology as the Schwarz *P* TPMS. However, analysis of the mean curvature shows that periodic equipotential surfaces are not TPMS (Barnes *et al.*, 1990).

1.3 Nodal Surfaces

Nodal surfaces are defined by $\Psi(\boldsymbol{r}) = 0$, for $\Psi(\boldsymbol{r})$ given by the following Fourier series (Mackay, 1993; Klinowski *et al.*, 1996):

$$\Psi(\boldsymbol{r}) = \sum_{\boldsymbol{k}} F(\boldsymbol{k}) \cos[2\pi \boldsymbol{k} \cdot \boldsymbol{r} - \alpha(\boldsymbol{k})], \tag{6}$$

where $\boldsymbol{k}$ describes the reciprocal lattice vectors for a given lattice, $\alpha(\boldsymbol{k})$ is a phase shift, and $F(\boldsymbol{k})$ is an amplitude associated with a given $\boldsymbol{k}$ vector.

Periodic surfaces can be approximated in terms of nodal surfaces. The quality of approximation depends on the number of terms in the Fourier series. The topology of a surface is in some simple cases already well reproduced by a series containing only the first terms in the expansion. Several such surfaces have been tabulated (von Schnering and Nesper, 1991). For example, the simplest approximations for *P*, *D*, and *G* surfaces based on Eq. (6) are respectively

$$\cos(X) + \cos(Y) + \cos(Z) = 0,$$

$$\cos(X)\cos(Y)\cos(Z) + \sin(X)\sin(Y)\sin(Z)$$
$$= 0,$$

$$\sin(X)\cos(Y) + \sin(Y)\cos(Z) + \cos(X)\sin(Z) = 0, \tag{7}$$

where $X = 2\pi x$, $Y = 2\pi y$, $Z = 2\pi z$.

Although nodal surfaces are neither minimal nor of constant mean curvature, they can be used as an *Ansatz* for such surfaces (Lambert *et al.*, 1996). The huge advantage of nodal surfaces is that they are described by simple mathematical equations. This fact makes them easily accessible to a broad range of scientists.

2. EXPERIMENTAL OBSERVATIONS OF PERIODIC SURFACES

Scattering techniques applied to a solution of water and lipids (Luzzati *et al.*, 1968a,b,c; Luzatti and Spegt, 1967) demonstrated the existence of a periodic bilayer of lipids dividing the water into two disjoint continuous subvolumes, forming a bicontinuous structure. Apart from scattering techniques, the single most effective probe of bicontinuous structures is the pulsed-gradient NMR self-diffusion technique, which can provide the self-diffusion rates of all the components in the cubic structure (Lindman *et al.*, 1989), providing a direct check on the continuity and extension of the region occupied by the components. If any of the components of the system is closed inside the micelle, then its effective diffusion constant (measured as a mean squared displacement divided by time) goes to zero, whereas for one in the continuous structure the effective diffusion coefficient is nonzero. Also TEM (transmission electron microscopy) micrographs often play a crucial role in the identification of bicontinuous structures (Thomas *et al.*, 1986).

2.1 Diffusion on Surfaces

For any closed surface the effective diffusion coefficient is zero. For a cylindrical surface the effective diffusion constant would be $\frac{1}{3}D_0$ (D_0 is the local diffusion coefficient), because of the reduction of the available spatial dimensions from 3 to 1. Analogously for the plane we find that the diffusion constant is $\frac{2}{3}D_0$. The coefficient in front of D_0 is known as an obstruction factor. It can be shown exactly that the diffusion coefficient for any minimal periodic surface of cubic symmetry is $\frac{2}{3}D_0$ (Anderson and Wennerström 1990), which also sets an upper limit for the diffusion constant on any surface. Actually for any surface of small mean curvature the value of the diffusion constant would be close to the value for the minimal surface. For the diffusion of a surfactant on a surface of constant mean curvature one can use the approximate formula given by Anderson and Wennerström (1990),

$$D = D_0[2/3 - \mu(v - v_0)^2], \tag{8}$$

where μ is a constant specific for a given symmetry and/or topology and v is the volume fraction occupied by one of the subvolumes, while v_0 is the volume fraction of this subvolume for the minimal surface of the same symmetry and topology as the surface studied. For balanced TPMS [*e.g.*, *G*, *D*, *P*, and *C*(*P*)], $v_0 = \frac{1}{2}$. For unbalanced TPMS (*e.g.*, *I-WP* and *F-RD*), v_0 differs from $\frac{1}{2}$ by about 6%. The constant $\mu = 1.8$ for the *P* family for monolayers and $\mu = 1.8/4 = 0.45$ for the bilayers. We note that additional study of the diffusion of water in the subvolume yields information about the hydration of water molecules at the polar periodic surface. Short-time diffusion of water should yield additional information about the surface area *S* of the periodic surface per unit volume, since the diffusion constant for short times *t* is (Mitra *et al.*, 1992)

$$\frac{D(t)}{D_0} = 1 - \frac{4}{9\sqrt{\pi}}\sqrt{D_0 t}S. \tag{9}$$

A detailed comparison of the theoretical calculation and experimental observations for various periodic surfaces is given by Anderson and Wennerström (1990). The typical diffusion constant D_0 for surfactant or phospolipids in bilayers is usually between 2×10^{-7} and $5 \times 10^{-8}\,\mathrm{cm^2\,s^{-1}}$; for comparison the diffusion of water is $2 \times 10^{-5}\,\mathrm{cm^2\,s^{-1}}$.

2.2 Amphiphilic Solutions

In an ordered mixture of water and surfactant or lipid, diffusion measurements provide direct information about the bicontinuity of the structure and therefore about the existence of a periodic surface. X-ray scattering provides information about the symmetry of the structure and therefore discerns between different periodic surfaces: the set of reflections for the *P* surface (*Im*3*m* symmetry) indexes to $\sqrt{2} : \sqrt{4} : \sqrt{6} : \sqrt{10} : \sqrt{12} : \sqrt{14} : \sqrt{16}$; that of the *D* surface (*Pn*3*m* symmetry), to $\sqrt{2} : \sqrt{3} : \sqrt{4} : \sqrt{6} : \sqrt{8} : \sqrt{9} : \sqrt{10}$; and that of the *G* surface (*Ia*3*d* symmetry), to $\sqrt{6} : \sqrt{8} : \sqrt{14} : \sqrt{16} : \sqrt{20} : \sqrt{22} : \sqrt{24}$ (Henry and Lonsdale, 1952). Finally measurements of the surface area inside the cubic cell can yield partial information about the topology of the structure. One can expect a small unit cell and a small surface area per side of the unit cell in structures of simple topology (Barois *et al.*, 1990) and large values for structures of complex topology. Typically the surface area per side of the cubic cell is between 2 and 4 for simple topology surfaces such as *P*, *D*, *G*, *I-WP*, *O-CT*0, etc., while it can be larger than 7 for complex-topology surfaces (Góźdź and Hołyst, 1996c).

In binary mixtures of water and surfactants or lipids the most commonly observed periodic structure is the gyroid phase, existing usually on the phase diagram between the hexagonal and lamellar mesophases. This structure has been observed in a very large number of systems (Luzzati *et al.*, 1968a,b,c; Arvidson *et al.*, 1985; Larsson, 1989; Clerck *et al.*, 1990, Turner *et al.*, 1992, Landh, 1994) and in computer simulations of the systems (Larson, 1996). The gyroid phase is found at rather high surfactant concentrations, usually much above 50% by weight. The similarities of various phase diagrams suggest that the transitions are driven mainly by the packing constraints, entropies and energies of mixing, and conformations of surfactant chain (Larson, 1996; SOLUBILITY AND MIXING IN FLUIDS).

Other cubic phases even at very low concentration of surfactants can be found in ternary mixtures (with water and oil) (Anderson and Wennerström, 1990; Barois *et al.*, 1990). For the system didodecyldimethylammonium bromide, water, and styrene, periodic surfaces are found over the huge range of water fractions from 11 to over 80%. Also in the system clear evidence for the *G* to *D* to *P* transition is found both in the mentioned experiments and in theory (Ciach, 1996).

The phase transitions between surfaces of different topology are accompanied by thermally excited topological fluctuations in the form of passages between the surfaces, as observed in computer simulations (Larson, 1992, 1996; Gompper and Kraus, 1994; Hołyst and Góźdź, 1997). Each passage changes the Euler characteristics by -2; thus this quantity is the direct measure of the number of passages in the system. The passages should be expected close to phase boundaries.

2.3 Copolymers

A diblock *AB* copolymer consists of two chemically distinct homopolymers *A* and *B* bound together by a covalent bound. The immiscibility (see SOLUBILITY AND MIXING IN FLUIDS) of *A* and *B* monomers drives the system towards ordered structures, such as to minimize the contacts between different monomers. The system separates into *A*-rich and *B*-rich domains, with the covalent bond resting at the interface between the domains. The typical structures formed are lamellar and hexagonal structures, where the domains are planar or cylindrical. Between the hexagonal and lamellar phases the gyroid phase (similarly as in lipid solutions) has been observed in x-ray experiments (Schultz *et al.*, 1994; Hajduk *et al.*, 1994; Föster *et al.*, 1994), where a periodic surface of topology and symmetry of the *G* surface is formed between *A*-rich and *B*-rich domains. It is worth mentioning that the first report on the observation of a periodic surface in block copolymers (starblocks) was done in 1986 (Thomas *et al.*, 1986), but the correct identification of the gyroid structure for the starblocks was done in 1995 (Hajduk *et al.*, 1995).

The gyroid phase for diblock copolymers exists in a narrow window of diblock copolymer [polyisopropene (PI)–polystyrene (PS)] composition v_{PI} (polyisoprene volume fraction) between 35 and 40% (Föster *et al.*, 1994). The gyroid phase exhibits pronounced metastability when cooled below the transition to lamellar or hexagonal phases. The theory also predicts the existence of the gyroid phase in the diblock copolymer system (Matsen and Bates, 1996).

Multiply continuous structures were observed in an *ABC* linear triblock copolymer system (Mogi *et al.*, 1992). In this structure two periodic surfaces of *D* symmetry and topology separate a volume into three disjointed subvolumes *A*-rich, *B*-rich, and *C*-rich, therefore the name ordered tricontinuous double-diamond structure.

2.4 Chemical Structures

The periodic zero-potential surfaces (POPS) occur most naturally in ionic crystals, where we have a periodic distribution of charges (von Schnering and Nesper, 1987). Although POPS are not minimal surfaces (Mackay and Klinowski, 1986; Barnes *et al.*, 1990), they may resemble them closely (Klinowski *et al.*, 1996).

The cesium chloride CsCl structure has the same topology and symmetry as the *P* Schwarz surface and is called the P^* surface to distinguish it from the *P* minimal surface. The aluminosilicates of the zeolite sodalite ($Na_4Al_3Si_{12}O_{26}Cl$) form a P^* surface, similarly as does the perovskite structure ($CaTiO_3$).

The D^* surface can be found in the zeolite structure of faujasite ($M^{1}_{57}Al_{57}Si_{135}O^{[53]}_{384}$) (Andersson *et al.*, 1984), similar to the cubic Laves phase $MgCu_2$.

Finally Y^{**}, the analog of the *G* surface, is found in compounds with the garnet structure $Ca_3[Al_2Si_3O_{12}]$, materials important for their applications as optoelectric color displays and magnetic bubble memory. Y^{**} is also found in the compound Ta_6Cl_{15}. A detailed description is given by von Schnering and Nesper (1987).

Mackay and Terrones (1991) predicted that a new type of ordered graphite foams, related to fullerenes, with topology similar to periodic surfaces may be found. These new structures are constructed by introducing seven- or eight-membered rings of carbon atoms into sheets of six-membered rings, thus giving saddle-shape surfaces. In fact such structures should be more stable than fullerene C_{60} (Terrones and Mackay, 1993) mainly because the 120° bond angles in ordinary graphite are almost preserved in the seven-membered and eight-membered carbon rings. In this way *P*, *D*, *G*, and *I-WP* surfaces were constructed of energies lower than the C_{60} form of carbon. The saddle and splay moduli (Helfrich, 1973) for these structures are respectively 190 kT and −65 kT (Hyde and O'Keeffe, 1996), making the graphite surfaces an order of magnitude stiffer than bilayers of an amphiphilic system (Chung and Caffrey, 1994).

2.5 Periodic Surfaces in Living Matter

Chroloplasts are organelles involved in photosynthesis, presenting a complex system of parallel bilayers. When a plant is grown for several days in the absence of light, the lamellar structure of chloroplast transforms into a periodic surface of the same topology and symmetry as the *P* Schwarz surface (also called a prolamellar body or etioplast), although it is not a minimal surface (Günning, 1965; Bouligand, 1990).

The cell membranes in a living organism—*e.g.*, endoplasmic reticulum (where cell proteins and lipids are synthesized)—are known to have a complex three-dimensional morphology, which, as shown by Landh (1995), is the same as for the triply periodic minimal surfaces. Three fundamental structures identified so far in cellular organelles are the gyroid *G* Schoen surface and the *D* and *P* Schwarz surfaces. The partition of space by the surface in the cell organelle enables the cell to control the concentration of various molecules and their transport across the bilayers. There are indications that multiply continuous structures exist in the mitochondria of the very large cone cells in tree shrews (Samorajski *et al.*, 1966).

The skeleton elements of sea stars, sea urchins, or sea cucumbers are monocrystals of a magnesium-rich calcite, each crystal being limited not by a planar surface but by a surface of the topology of the *P* surface (Nissen, 1969).

We note that it is possible using the mathematical modeling of periodic surfaces to represent the DNA helicoidal structure by a singly periodic surface (Jacob, 1997).

3. FUNCTIONAL GENERATION OF PERIODIC SURFACES

Periodic surfaces can be generated as minima of a free-energy functional. Here we present the functional used to generate periodic structures and discuss the methods of minimization (in real and Fourier space) and the resulting solutions of the minimization problem.

3.1 Landau–Ginzburg Model

In studies of microemulsion and ordered phases that appear in systems containing surfactants, the following Landau–Ginzburg functional was proposed (Teubner and Strey, 1987; Gompper and Schick 1990):

$$\mathscr{F}[\phi(\boldsymbol{r})] = \int \mathrm{d}^3 r(|\Delta\phi(\boldsymbol{r})|^2 + g(\phi(\boldsymbol{r}))|\nabla\phi(\boldsymbol{r})|^2 + f(\phi(\boldsymbol{r}))). \tag{10}$$

The functions $g(\phi(\boldsymbol{r}))$, $f(\phi(\boldsymbol{r}))$ are given by

$$g(\phi(\boldsymbol{r})) = g_2\phi(\boldsymbol{r})^2 + g_0, \tag{11}$$

$$f(\phi(\boldsymbol{r})) = [\phi(\boldsymbol{r})^2 - 1]^2[\phi(\boldsymbol{r})^2 + f_0]. \tag{12}$$

The values of the field $\phi(\boldsymbol{r})$ are proportional to the difference in oil and water concentrations and are negative for water and positive for oil regions or *vice versa*. The sign is just a matter of convention. The surface

$$\phi(\boldsymbol{r}) = 0 \tag{13}$$

describes the interface between oil and water. The function $f(\phi(\boldsymbol{r}))$ is the bulk free energy. The surfactant degrees of freedom are considered as being integrated out and the surfactant properties enter the functional (10) through the form of the functions $g(\phi(\boldsymbol{r}))$ and $f(\phi(\boldsymbol{r}))$. The functional (10) can be also used to model binary mixtures of water and surfactants. In such a case the negative values of $\phi(\boldsymbol{r})$ are interpreted as the interior part of a bicontinuous phase and the positive as the exterior part of the phase (or *vice versa*).

The function $f(\phi(\boldsymbol{r}))$ has by construction three minima, which guarantees three-phase coexistence, *i.e.*, oil, water, and a microemulsion. The minima for oil-rich and water-rich phases are of equal depth, which makes the system symmetric. Varying the parameter f_0 makes the middle phase (microemulsion) more or less stable with respect to the two bulk phases. The phase diagram for the functional (10) contains homogeneous oil and water phases, microemulsion, and a lamellar phase (Gompper and Zschoke, 1992).

3.2 Minimization Procedure in Real and Fourier Space

In real space, the functional (10) is discretized first on the cubic lattice to find its minima. Thus the functional $\mathscr{F}[\phi(\boldsymbol{r})]$ becomes a function $F(\{\phi_{i,j,k}\})$ of N^3 variables, where $L = (N-1)h$ is the linear dimension of the cubic lattice, h is the lattice spacing and $\{\phi_{i,j,k}\}$ stands for the set of all variables of the function. Each variable $\phi_{i,j,k}$ represents the value of the field $\phi(\boldsymbol{r})$ at the point $\boldsymbol{r} = (i,j,k)h$ and $i,j,k = 1,\ldots,N$.

The periodicity is incorporated into the functional (10) by the boundary conditions $\phi_{1,j,k} = \phi_{N,j,k}$, $\phi_{2,j,k} = \phi_{N+1,j,k}$, $\phi_{3,j,k} = \phi_{N+2,j,k}$ $\phi_{0,j,k} = \phi_{N-1,j,k}$, and $\phi_{-1,j,k} = \phi_{N-2,j,k}$ and similarly in the y and z directions. The points outside the unit cell, given by the periodic boundary conditions, enter the functional through the calculation of derivatives of points at the boundary and near the boundary of the lattice, *i.e.*, when at least one of the indices i,j,k is equal to $1, 2, N-1$, or N.

The symmetry of the structure is imposed on the field $\phi(\boldsymbol{r})$ by building the field inside a unit cell of a smaller polyhedron, replicating it by reflections, translations, and rotations (Góźdź and Hołyst, 1996c).

In Fourier space $\phi(\boldsymbol{r})$ for periodic structures is represented by the Fourier series

$$\phi(\boldsymbol{r}) = A_0 + \sum_i \sum_{\boldsymbol{k}^i} [A_i \cos(\boldsymbol{k}^i \cdot \boldsymbol{r}) + B_i \sin(\boldsymbol{k}^i \cdot \boldsymbol{r})] \tag{14}$$

where $\boldsymbol{k}^i$ are the reciprocal lattice vectors in the ith shell and A_i and B_i are the amplitudes of the ith shell. Here the shell means a set of reciprocal lattice vectors of the same length related to each other by the symmetry operations characteristic for a given lattice. The approximation by the Fourier series ensures that the function $\phi(\boldsymbol{r})$ is continuous at the unit-cell boundaries. Next, the functional (10) is minimized with respect to the amplitudes A_i and

B_i. As a result one gets the nodal surface given by the equation $\phi(\mathbf{r}) = 0$. The quality of approximation is better the larger the number of shells included.

From the computational point of view the Fourier-space approach requires fewer variables to be minimized for, but the speed of calculations is significantly decreased by the evaluation of a trigonometric function, which is computationally expensive.

3.3 High-Genus Periodic Surfaces

High-genus surfaces are easily generated by functional minimization in real space. It is enough to set a larger value for the unit cell than the value for a low-genus surface of the same symmetry and to start minimization from an arbitrary configuration. The minimization will lead to the high-genus surface with the mean curvature close to zero (Góźdź and Hołyst, 1996a,b,c). The high-genus surfaces are most easily generated for the values of parameters f_0 and g_0 in (10) taken near the boundary between lamellar and microemulsion phases. Other methods present great difficulties in generating high-genus periodic minimal surfaces.

Here a few examples of high-genus surfaces of different symmetry are presented The structure *SCN*1, Fig. 8, has the symmetry of the Schwarz *P* surface; the structure *CD*, Fig. 9, has the symmetry of double diamond *D* surface; and the structure *GX*5, Fig. 10, has the symmetry of the gyroid *G* surface. All these surfaces are balanced. The genus for *SCN*1 is 43; for *CD*, 73; and for *GX*5, 157.

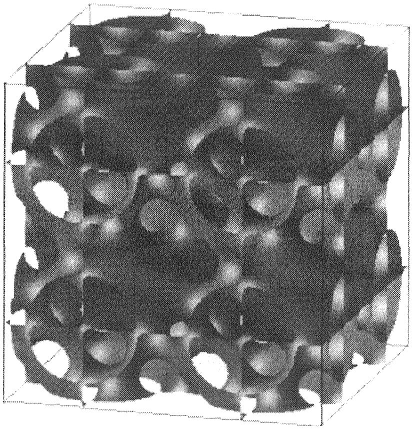

FIG. 9. High-genus surface *CD* of the same symmetry as the *D* surface.

3.4 Multiply Continuous Structures

Multiply continuous structures were discovered during functional generation of periodic surfaces (Góźdź and Hołyst, 1996a,c). The possibility of the existence of multiply continuous structures has so far never been discussed with respect to ternary mixtures of oil, water, and a surfactant, although it has

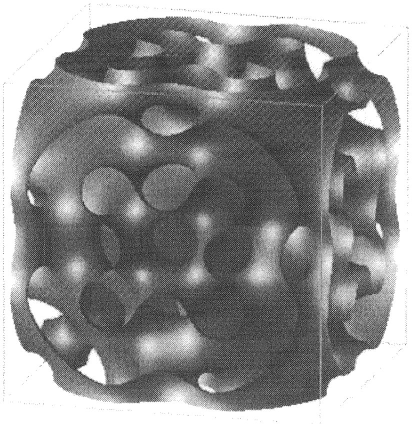

FIG. 8. High-genus surface *SCN*1 of the same symmetry as the *P* surface.

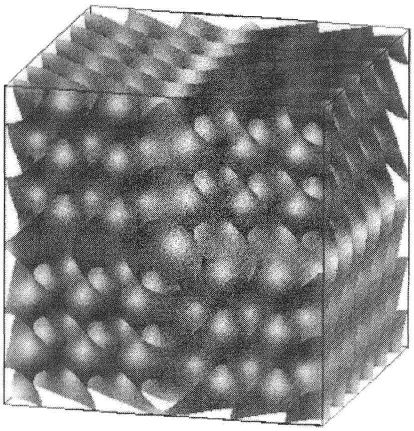

FIG. 10. High-genus surface *GX*5 of the same symmetry as the *G* surface.

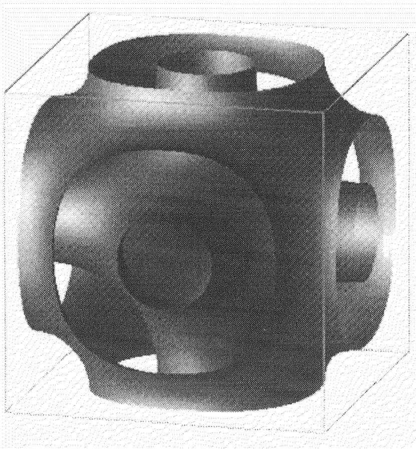

FIG. 11. Triply continuous structure of the same symmetry as the *P* surface.

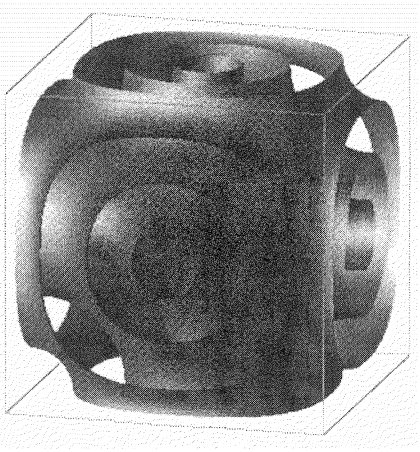

FIG. 12. Quadruply continuous structure of the same symmetry as the *P* surface.

been observed in the *ABC* diblock copolymers (Mogi *et al.*, 1992). In the bicontinuous structure a single surface separates the volume into two disjoint subvolumes. In multiply continuous phases there is more than one periodic surface, which disconnects the volume into three or more disjoint subvolumes. The multiply continuous structures are most easily generated for the values of parameters f_0 and g_0 in (10) taken near the boundary between lamellar and water (oil) phases and for a bigger unit cell length, $L = (N - 1)h$, than the length of structures of low genus. The experimental results do not rule out the existence of such structures (Gompper and Schick, 1994).

We present here multiply continuous structures of the simple cubic phase symmetry (*SCL*1, *SCL*2, Figs. 11 and 12) and of the gyroid phase symmetry (*GL*2, Fig. 13). The *SCL*1 structure is triply continuous, *SCL*2 is quadruply continuous, and *GL*2 is sextuply continuous. It is remarkable that the volume fraction of oil and water is 0.5 for all these structures. The genera for all surfaces in a given structure are the same.

The *SCL*1 surface is especially interesting. Although the outer and the inner surfaces look different, they have the same surface area. In fact they are built from the same piece of the surface.

The *SCL*2 structure is composed of three different embedded periodic surfaces. The middle surface is the Schwarz minimal surface *P*. Similarly, the middle surface in the *GL*2 structure is the Schoen minimal surface *G*.

The existence of a structure such as *GL*2 suggests that one can generate arbitrary *n*-uply continuous structures. In the functional method it is only necessary to set the cell length sufficiently large.

4. APPLICATIONS

Although the periodic structures formed by lipids, surfactants, and soaps in water sol-

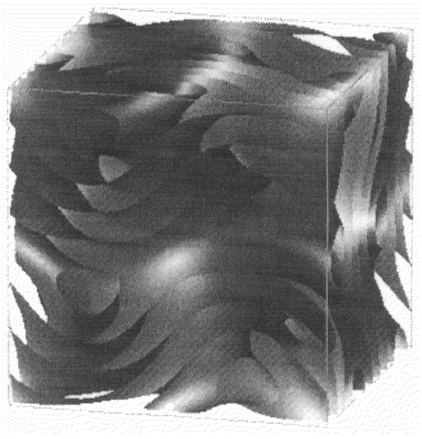

FIG. 13. Sextuply continuous structure of the same symmetry as the *G* surface.

utions have been known since 1967 (Luzatti and Spegt, 1967), the practical applications started in the 1990s. The discovery that an organic periodic surface can be used as a template for organic or inorganic polymerization reactions where the final product is an ordered mesoporous material of well defined pore sizes and shapes—*e.g.*, mesoporous silica (Kresge *et al.*, 1992) or hydrogels used as contact lenses (Anderson and Ström, 1991)—brought the subject of periodic surfaces into the realm of technology.

4.1 Mesoporous Sieves and Catalysts

The mesoporous silica structures are prepared in a water–surfactant system (Kresge *et al.*, 1992). After polymerization of the silica oligomers the surfactant can be burned and the inorganic sieve emerges. The sieves are remarkable since they have pore sizes ranging from 20 Å to hundreds of angstroms; the pores have well defined shape and size; the structure is ordered over a micron length scale. The process is complicated and still not fully understood (Cheng *et al.*, 1995). Monnier *et al.* (1993) describe the process as follows: there is a multidentate binding of anionic silicate oligomers to the cationic surfactant, the polymerization takes place in the interfacial region, and there is a charge-density matching between the surfactant and the silicate. The synthesis can be carried out under conditions in which silicate alone does not condense (at *p*H from 12 to 14 and silicate concentration between 0.5% and 5%) and/or at a small concentration of surfactant (Monnier *et al.*, 1993; Attard *et al.*, 1995). They point out that under certain conditions formation of the ordered structure results from the interplay between the silica and the surfactants. The formation of the ordered structure at high temperature is fast in comparison with the polymerization reaction; the polymerized silica mesostructure obtained in this two-stage process is more stable than the one obtained at low temperature. The polymerization of silicate at the interface is favored because the surfactant partially screens the charges and the concentration of silicate species is high in this region. There is an indication that the silicate and the polymerization process may take part in the ordering process, especially when the polymerization is fast enough. Apart from the lamellar and hexagonal microstructures, the bicontinuous gyroid structure is formed. In the system of CTA/SiO_2 for a high fraction of both components a single infinite silicate sheet separates the surfactant into two equal and disconnected subvolumes. The midplane of the silicate sheet is the minimal gyroid periodic surface.

The surfaces formed by surfactants can also be used as templates for inorganic oxide crystallization (Archibald and Mann, 1993), and presumably nature uses them in biomineralization processes.

The coating of mesoporous silicates with sulfur compunds can be achieved and the coated powder can be used as an effective agent for water purification (Feng *et al.*, 1997).

4.2 Soft Contact Lenses

The periodic surfaces formed by amphiphilic molecules can also be used as templates for the polymerization reaction, leading to the formation of material that can be used for the production of contact lenses (Anderson and Ström, 1991; Anderson, 1987, 1990a). For example, a water solution of hydrophilic monomers such as 2-hydroxyethylmethacrylate (HEMA) mixed with a surfactant in the right proportions self assembles into a triply periodic structure, in which monomers and surfactants form a periodic surface. Next, the polymerization is initiated by, *e.g.*, ultraviolet light, heat, an electron beam, electroinitiation, or redox initiation. After polymerization the surfactant is removed (*e.g.*, electrophoretically) and the mesoporous rigid structure is created. These structures have three properties important for contact lenses. First of all they have a very regular pore structure, unlike the conventional hydrogels where the pore structure is very irregular. The pore sizes can be made large enough to allow the passage of high–molecular-weight proteins through the lens into the eye. They can have a large water content, allowing for the rapid permeability of oxygen, which is an exponential function of the water concentration in hydrogel. Finally, despite large pores (which can be selected in a range 20–400 Å), the structure can have still a large shear modulus, unlike the simple hydrogel typically used as a contact-lens material. In a simple hydrogel at equilibrium swollen by a solvent, the shear modulus G_s is given by (Flory 1953)

$$G_s = -RT\frac{\ln(1-\phi_m)+\phi_m+\chi\phi_m^2}{v_1(\phi_m^{1/3}-\phi_m)}, \tag{15}$$

where ϕ_m is the volume fraction of monomers, v_1 is the molar volume of the solvent, χ is the Flory–Huggins interaction parameter between the solvent and the monomer molecules, T is the temperature, and R is the gas constant. A simple estimate for poly-*cis*-1,4-butadiene shows that when the solvent (water) content increases from 52 to 82% the shear modulus decreases from 1.35×10^7 to 2.56×10^5 dyn cm^{-2} (*i.e.*, fifty-fold). This can be compared with the mesoporous material made on the surfactant template for which the decrease would be only 50%, because in this case the strength of the material is only a linear function of the pore size.

In the simple hydrogel the large pore size usually means a strong decrease in the cross-linking density. On the other hand, in the mesoporous material prepared on a periodic surface the average distance between the monomers is independent of the pore size.

4.3 Crystallization of Membrane Proteins on Periodic Surfaces

The prediction of the biological function of a given protein is based on its known structure. This in turn can be deduced from the x-ray spectra, NMR spectra (Thompson *et al.*, 1992), or electron microscopy (Unwin, 1993). In the case of x-ray spectra one has to prepare a crystal of a given protein. The high-molecular-weight membrane proteins that are responsible for many biological functions—such as active transport through the membrane, signal transduction, charge separation, energy conversion, and catalysis of vital reactions—have anisotropic shape and hydrophobic surfaces. This is a serious obstacle in producing well-ordered three-dimensional crystals suitable for x-ray analysis. The natural environment of membrane proteins is the lipid bilayer that forms the membrane. In solutions the proteins lose their integrity, denaturate, or aggregate. However, if the protein removed from the bilayer is put back, it regains stability (Page *et al.*, 1988). The idea for obtaining crystals of membrane proteins is to use the cubic phases consisting of water and the periodic lipid bilayer as hosts for polymerization of membrane proteins (Landau and Rosenbusch, 1997). In this experiment monoolein or monopalmitolein forms a periodic surface that incorporates bacteriorhodopsin, a membrane protein extracted from the purple patches of *Halobacterium halobium*. Nucleation starts at the surface, and the crystal grows while the proteins diffuse along the bilayer to the nucleation site. As the crystal grows the lipid surface recedes, so that no lipidic components are incorporated into the growing crystal. The structure of the crystal depends on the structure of the bilayer. For monoolein it was a hexagonal crystal, while in the case of monopalmitolein it formed a rhombic crystal. The periodicity and continuity of the lipid surface is crucial; when the lipids form a micellar solution—that is, the lipid surface forms closed structures—the crystals could not be grown. The crystallization process takes about 14 d to obtain crystals of sizes $0.1 \times 0.1 \times 0.03\,\text{mm}^3$, and the current resolution of x-ray spectroscopy is 2 Å.

This method seems to be very promising for the routine crystallization of labile membrane proteins.

GLOSSARY

Amphiphilic Molecule: A molecule that consist of two parts, one hydrophobic and one hydrophilic (examples: soap molecules, detergents, lipids, and phospholipids).

Bicontinuous Structure: A structure consisiting of two disjoint continuous subvolumes, separated by a continuous (*i.e.*, not closed) surface.

Diblock Copolymer: Two linear polymers consisting of different repeated units joined by a covalent bond at the ends.

Euler Characteristic: $\chi = 2(1-g)$, where g is genus.

Gaussian Curvature: The inverse of the product of two principal radii of curvature. For a sphere of radius R it is equal to $1/R^2$. For an infinite cylinder it is 0.

Genus: Number of holes in a closed surface. The genus of a sphere is 0; for a torus it is 1; and for a pretzel it is 2.

Hydrophilic: Having strong affinity towards water ("water loving").

Hydrophobic: Opposite to hydrophilic ("water hating").

Mean Curvature: One half of the sum of the inverse of two principal radii of curvature. For a sphere of radius R it is $1/R$. For a cylinder of the same radius it is $1/2R$.

Mesoporous Material: A porous material with pore sizes intermediate between the microscopic scale (*i.e.*, angstroms) and the macroscopic scale (*i.e.*, thousands of angstroms).

Micellar Solution: A solution of, *e.g.*, water and surfactant where the surfactant molecules form a closed structure—spherelike, disklike, or rodlike. The bilayer surface is closed in this case in contrast to the periodic surfaces.

Periodic Surface: A surface that moves onto itself under a unit translation in one, two, or three coordinate directions.

Polymer: A molecule of high molecular weight consisting of repeated units (monomers or mers).

Works Cited

Anderson, D. M. (1987), *US Patent Appl.* No. 052, 713.

Anderson, D. M. (1990a), *World Patent Appl.* WO 90/07575.

Anderson, D. M. (1990b), *J. Phys. (Paris), Colloq.* **51** C7-1–C7-18.

Anderson, D. M., Ström, P. (1991) *Physica A* **178**, 151–167.

Anderson, D. M., Wennerström, H. (1990), *J. Phys. Chem.* **94**, 8683–8694.

Anderson, D. M., Davis, H. T., Scriven, L. E., Nitsche, J. C. C. (1990), *Adv. Chem. Phys.* **77**, 337–396.

Andersson, S., Hyde, S. T., von Schnering, H.-G. (1984), *Z. Kristallogr.* **168**, 1–17.

Archibald, D. D., Mann, S. (1993), *Nature (London)* **364**, 430–433.

Arvidson, G., Brentel, I., Khan, A., Lindblom, G., Fontell, K. (1985), *Eur. J. Biochem.* **152**, 753–759.

Attard, G. S., Glyde, J. C., Gölter, C. G. (1995), *Nature (London)* **378**, 366–368.

Barnes, I. S., Hyde, S. T., Ninham, B. W. (1990), *J. Phys. (Paris), Colloq.* **51**, C7-19–C7-24.

Barois, P., Eidam, D., Hyde, S. T., (1990), *J. Phys. (Paris), Colloq.* **51**, C7-25–C7-34.

Brakke, K. A. (1992), *Exp. Math.* **1**, 141–165.

Bouligand, Y. (1990), *J. Phys. (Paris), Colloq.* **51**, C7-35–C7-52.

Canham, P. (1970), *J. Theor. Bio.* **26**, 61–75.

Charvolin, J., Sadoc, J. F. (1987), *J. Phys. (Paris), Colloq.* **48**, 1559–1569.

Cheng, C.-F., Luan, Z., Klinowski J., (1995), *Langmuir* **11**, 2815–2819.

Chung, H., Caffrey, M. (1994), *Nature (London)* **368**, 224–226.

Ciach, A. (1996), *J. Chem. Phys.* **104**, 2376–2384.

Clerck, M., Levelut, A. M., Sadoc, J. F. (1990), *J. Phys. (Paris), Colloq.* **51**, C7-97–C7-104.

Cvijović, D., Klinowski, J. (1992a), *J. Phys. I* **2**, 137–147.

Cvijović, D., Klinowski, J. (1992b), *J. Phys. I* **2**, 2191–2205.

Cvijović, D., Klinowski, J. (1992c), *J. Phys. I* **2**, 2207–2220.

Davis, M. E. (1993), *Nature (London)* **364**, 391–393.

De Kruijff, B. (1997), *Nature (London)* **386**, 129–130.

Feng, X., Fryxell, G. E., Wang, L.-Q., Kim, A. Y., Liu, J., Kerner K. M., (1997), *Science* **278**, 923–925.

Fisher, W., Koch, E. (1986), *Krystallogr.* **179**, 31–52.

Fisher, W., Koch, E. (1996), *Philos. Trans. R. Soc. London. A* **354**, 2105–2142.

Flory, P. (1953), *Principles of Polymer Chemistry*, Ithaca, NY: Cornell Univ. Press.

Fogden A. (1990), *J. Phys. (Paris), Colloq.* **51**, C7-149–C7-156.

Fogden A. (1992), *J. Phys. I France* **2**, 233–239.

Fogden A., Haeberlein (1994), *J. Chem. Soc. Faraday Trans.* **90**, 263–270.

Fogden A., Haeberlein M., Lidin S. (1993), *J. Phys. I France* **3**, 2371–2385.

Föster, S., Khandpur, A. K., Zhao, J., Bates, F. S., Hamley, I. W., Ryan, A. J., Bras, W. (1994), *Macromolecules* **27**, 6922–6935.

Gompper, G., Kraus, M. (1994), *Phys. Rev. E* **47**, 4301–4312.

Gompper, G., Schick, M. (1990), *Phys. Rev. Lett.* **65**, 1116–1119.

Gompper, G., Schick, M. (1994), in: C. Domb and J. L. Lebowitz (Eds.), *Self-assembling amphiphilic systems*, Vol. 16 of *Phase transitions and critical phenomena*, London: Academic Press.

Gompper, G., Zschoke (1992), *Phys. Rev. A* **46**, 4836–4851.

Góźdź, W., Hołyst, R. (1996a), *Macromol. Theory Simul.* **5**, 321–332.

Góźdź, W., Hołyst, R. (1996b), *Phys. Rev. Lett.* **76**, 2726–2729.

Góźdź, W., Hołyst, R. (1996c), *Phys. Rev. E* **54**, 5012–5026.

Góźdź, W., Hołyst, R. (1997), *J. Chem. Phys.* **106**, 9305–9312.

Grosse-Brauckmann, K. (1997), *J. Coll. Int. Sci.* **187**, 418–428.

Günning, B. E. S. (1965), *Protoplasma* **60**, 111–130.

Hajduk, D. A., Harper, P. E., Gruner, S. M., Honeker, C. C., Kim, G., Thomas, E. L., Fetters, L. J. (1994), *Macromolecules* **27**, 4063–4075.

Hajduk, D. A., Harper, P. E., Gruner, S. M., Honeker, C. C., Thomas, E. L., Fetters, L. J. (1995), *Macromolecules* **28**, 2570–2573.

Helfrich, W. (1973), *Z. Naturforsch.* **28c**, 693–703.

Henry, N. F. M., Lonsdale K. (Eds.) (1952), *International Tables for X-ray Crystallography*, vol. I, Birmingham, U.K.: Kynoch.

Hoffman, D. (1996), *Nature (London)* **384**, 28–29.

Hołyst, R., Góźdź, W. (1997), *J. Chem. Phys.* **106**, 4773–4782.

Hyde, S. T., O'Keeffe, M. (1996), *Philos. Trans. R. Soc. London A* **354**, 1999–2008.

Jacob, M. (1997), *J. Phys II France* **7**, 1035–1044.

Karcher, H. (1989), *Manuscr. Math.* **64**, 291–357.

Karcher, H., Polthier, K. (1996), *Philos. Trans. R. Soc. London A* **354**, 2077–2104.

Klinowski, J., Mackay, A. L., Terrones, H. (1996), *Philos. Trans. R. Soc. London A* **354**, 1975–1987.

Koch, E., Fisher, W. (1988), *Z. Krystallogr.* **183**, 129–152.

Kresge, C. T., Leonowicz, M. E., Rothe, W. J., Vartuli, J. C., Beck, J. S. (1992), *Nature (London)* **359**, 710–712.

Lambert, C. A., Radzilowski, R. H., Thomas L. E. (1996), *Phil. Trans. R. Soc. London A* **354**, 2009–2023.

Landau, E. M., Rosenbusch, J. P. (1996), *Proc. Natl. Acad. Sci. USA* **93**, 14532–14535.

Landh, T. (1994), *J. Phys. Chem.* **98**, 8453–8462.

Landh, T. (1995), *FEBS Lett.* **369**, 13–17.

Larson, R. G. (1992), *J. Chem. Phys.* **96**, 7904–7914.

Larson, R. G. (1996), *J. Phys. II France* **6**, 1441–1463.

Larsson, K. (1989), *J. Phys. Chem.* **93**, 7304–7314.

Lawson, B. H. (1970), *Ann. Math.* **92**, 335–374.

Lindman, B., Shinoda, K., Olsson, U., Anderson, D., Karlström, D., Wennerström, H. (1989), *Colloids Surf.* **38**, 205–224.

Lipowsky, R. (1993), *Physica A* **194**, 114–127.

Luzzati, V., Spegt, P. A. (1967), *Nature (London)* **215**, 701–704.

Luzzati, V., Tardieu, A., Gulik-Krzywicki, T. (1968), *Nature (London)* **217**, 1028–1030.

Luzzati, V., Gulik-Krzywicki, T., Tardieu, A. (1968b), *Nature (London)* **218**, 103–106.

Luzzati, V., Tardieu, A., Gulik-Krzywicki, T., Rivas, E., Reiss-Husson, F. (1968c), *Nature (London)* **220**, 485–488.

Mackay, A. L. (1988), *Angew. Chem., Int. Ed. Engl.* **27**, 849–850.

Mackay, A. L. (1993), *Proc. R. Soc. London A* **442**, 47–59.

Mackay, A. L. (1994), *Chem. Phys. Lett.* **221**, 317–321.

Mackay, A. L. (1995), *Curr. Sci.* **69**, 151–161.

Mackay, A. L., Klinowski, J. (1986), *Comput. Math. Appl.* **12B**, 803–824.

Mackay, A. L., Terrones, H. (1991), *Nature (London)* **352**, 762.

Matsen, M. W., Bates, F. S. (1996), *Macromolecules* **29**, 1091–1098.

Mermin, N. D. (1992), *Rev. Mod. Phys.* **64**, 3–48.

Mitra, P. P., Sen, P. N., Schwartz, L. M., Le Doussal, P. (1992), *Phys. Rev. Lett.* **68**, 3555–3556.

Mogi, Y., Mori, K., Matsushita, Y., Noda, I. (1992), *Macromolecules* **25**, 5412–5415.

Monnier, A., Schüth, F., Huo, Q., Kumar, D., Margolese, D., Maxwell, R. S., Stucky, G. D., Krishnamurty, M., Petroff, P., Firouzi, A., Janicke, M., Chmelka, B. F. (1993), *Science* **261**, 1299–1302.

Nissen, H. U. (1969), *Science* **166**, 1150–1152.

Nitsche, J. C. C. (1989) *Lectures on Minimal Surfaces*, Vol. 1, Cambridge, U.K.: Cambridge University Press.

Page, M. G., Rosenbusch, J. P., Yamato, I. (1988), *J. Biol. Chem.* **263**, 15897–15905.

Samorajski, T., Ordy, J. M., Keefe, J. R. (1966), *J. Cell. Biol.* **28**, 489–504.

von Schnering, H. G., Nesper, R. (1987), *Angew. Chem.* **26**, 1059–1200.

von Schnering, H. G., Nesper, R. (1991), *Z. Phys. B* **83**, 407–412.

Schoen, A. H. (1970), *Infinite Periodic Minimal Surfaces Without Self-intersections*, NASA Technical Note No. D-5541, Washington, DC: U.S. GPO.

Schultz, M. F., Bates, F. S., Almdal, K., Mortensen, K. (1994), *Phys. Rev. Lett.* **73**, 86–89.

Sheng, Q. (1994), Ph. D. Thesis, Cornell University.

Sheng, Q., Elser, V. (1994), *Phys. Rev. B* **49**, 9977–9982.

Terrones, H. (1990), *J. Phys. (Paris), Colloq.* **51** C7-345–C7-363.

Terrones, H., Mackay, A. L. (1993), *Chem. Phys. Lett.* **207**, 45–50.

Teubner, M., Strey, R. (1987), *J. Chem. Phys.* **87**, 3195–3200.

Thomas, E. L., Alward, D. B., Kinning, D. J., Martin, D. C., Handlin, D. L., Fetters, L. J. (1986), *Macromolecules* **19**, 2197–2202.

Thompson, L. K., McDermott, A. E., Raap, J., van der Wielen, C. M., Lugtenberg, J., Hershfeld J., Griffin, R. G. (1992), *Biochemistry* **31**, 7931–7938.

Turner, D. C., Wang, Z-G., Gruner, S. M., Mannock, D. A., McElhaney, R. N. (1992), *J. Phys. II France* **2**, 2039–2063.

Unwin, N. (1993), *J. Mol. Biol.* **229**, 1101–1124.

Further Reading

Anderson, D. M., Davis, H. T., Scriven, L. E., Nitsche, J. C. C. (1990), "Periodic Surfaces of Prescribed Mean Curvature," in: I. Prigogine, S. A. Rice (Eds.), *Advances in Chemical Physics*, Vol. 77, New York: Wiley, pp. 337–396.

Coxeter, H. S. M. (1968), *Twelve Geometric Essays*, Carbondale, IL: Univ. of Southern Illinois Press.

E. Dubois-Violette, B. Pansu (Eds.) (1990), "International Workshop on Geometry and Surfaces" *J. Phys. (Paris), Colloque* **51** (C7).

Giusti, E. (1984), *Minimal Surfaces and Functions of Bounded Variations*, Boston–Bassel–Stuttgart: Birkhauser.

Gompper, G., Schick M. (1994), *Self Assembling Amphiphilic Systems*, 1st ed. Vol. 16 of *Phase Transitions and Critical Phenomena*, C. Domb and J. L. Lebowitz (Eds).

Hoffman, D., Concus, P. Finn, R. (Eds.) (1991), *Geometric Analysis and Computer Graphics*, MSRI-Book Series, Berlin: Springer Verlag.

Hopf, H. (1983), *Differential Geometry in the Large*, Springer Lectures Notes in Mathematics No. 1000, Berlin: Springer.

Isenberg, C. (1978), *The Science of Soap Films and Soap Bubbles*, New York: Dover Publications.

J. Klinowski, A. L. Mackay (Eds.) (1996), "Curved Surfaces in Chemical Structures," *Philos. Trans. R. Soc London A*, **354**.

Lawson, H. B. (1980), *Lectures on Minimal Submanifolds*, Vol. 1, Berkeley: Publish or Perish, Inc.

Mackay, A. L., Klinowski, J. (1986), "Towards a Grammar of Inorganic Structures" *Comput. Math. Appl.* **12B**, 803–824.

Nitsche, J. C. C. (1989), *Lectures on Minimal Surfaces*, Vol 1., Cambridge: Cambridge University Press.

von Schnering, H. G., Nesper, R. (1987), "How Nature Adopts Chemical Structures to Curved Surfaces" *Angew. Chem.* **26**, 1059–1200.

Schoen, A. H. (1970), *Infinite Periodic Minimal Surfaces Without Self-intesection*, NASA Technial Note No. D-5541.

Spivak, M., (1979), *A Comprehensive Introduction to Differential Geometry*, Vol. IV, Berkeley: Publish or Perish Inc.

QUANTUM COMPUTATION

Samuel L. Braunstein, *SEECS, University of Wales, Bangor, United Kingdom*

	Introduction	239
1.	**Computing at the Atomic Scale**	240
2.	**Reversible Computation**	240
3.	**Classical Universal Machines and Logic Gates**	241
3.1	FANOUT and ERASE	241
3.2	Computation without ERASE	242
4.	**Elementary Quantum Notation**	243
5.	**Logic Gates for Quantum Bits**	243
6.	**Logic Gates in the Laboratory**	245
7.	**Model Quantum Computer and Quantum Code**	246
8.	**Quantum Parallelism: Period of a Sequence**	247
9.	**The Complexity of Factoring**	249
10.	**Security and RSA**	249
11.	**Shor's Result: Factoring Numbers**	250
12.	**Quantum Error Correction**	251
13.	**Prospects**	252
	Glossary	253
	Appendix	254
	Works Cited	254

INTRODUCTION

A quantum computer is a device that can arbitrarily manipulate the quantum state of a part of itself. The field of quantum computation is largely a body of theoretical promises for some impressively fast algorithms that could be executed on quantum computers. However, since the first significant algorithm was proposed in 1994 (Shor, 1994) experimental progress has been rapid with several schemes yielding two– (Turchette *et al.*, 1995; Monroe *et al.*, 1995) and three–quantum-bit manipulations (Gershenfeld and Chuang, 1997). At the writing of this article it does not seem unreasonable to expect that small quantum computers capable of manipulating the quantum states of five or six two-level systems will be available within around five years. In addition, with the discovery of quantum error-correction schemes (Shor, 1995), such machines have the promise of providing long-term storage of quantum information and possibly allowing the ability to manipulate many more bits. It is still too early to tell whether the promises of rapid computation are achievable, nor is it yet well understood how broad a class of problems could be significantly speeded up by quantum computers.

Quantum computers were first discussed by Benioff (1980, 1981, 1982) in the context of simulating classical Turing machines (very elementary conventional computers) with quantum unitary evolution. Feynman (1982, 1986) considered the converse question of how well classical computers can simulate quantum systems. It was concluded that classical computers invariably suffer from an exponential slowdown in trying to simulate quantum systems, but that quantum systems could, in principle, simulate each other without this slowdown. It was Deutsch (1985a, 1985b), however, who first suggested that quantum superposition might allow quantum evolution to perform many classical computations in parallel.

To demonstrate where such capabilities may lie hidden, we review an elementary quantum mechanical experiment. The two-slit experiment is prototypic for observing one key feature of quantum mechanics: a source emits photons, electrons, or other particles that arrive at a pair of slits. These particles undergo unitary evolution and finally measurement. We see an interference pattern, with both slits open, which wholly vanishes if either slit is covered. In some sense, each particle passes through both slits in parallel. If such unitary evolution were to represent a calculation (or an operation within a calculation) then the quantum system would be performing computations in parallel. In some sense this quantum parallelism comes for free without our having to construct many copies of the "processing

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

unit." The output of this system would be given by the constructive interference among the parallel computations.

In this article we give a tutorial on how quantum mechanics can be used to improve computation. We concentrate on the only known algorithm that demonstrates an exponential speedup relative to the best known classical algorithms. Our challenge: solving a problem that is exponentially difficult for a conventional computer—that of factoring a large number. As a prelude, we review the standard tools of computation, universal gates and machines. These ideas are then applied first to classical, dissipationless computers and then to quantum computers. A schematic model of a quantum computer is described as well as some of the subtleties in its programming. The Shor algorithm (Shor, 1994; Ekert and Jozsa, 1996) for efficiently factoring numbers on a quantum computer is presented in two parts: the quantum procedure within the algorithm and the classical algorithm that calls the quantum procedure. The mathematical structure within the factoring problem is discussed, making it clear what contribution the quantum computer makes to the calculation. The complexity of the Shor algorithm is compared with that of factoring on conventional machines, and its relevance to public-key cryptography is noted. In addition, we discuss the experimental status of the field and also quantum error correction, which may in the long run help solve some of the most pressing difficulties. We conclude with an outlook as to the feasibility and prospects for quantum computation in the coming years.

1. COMPUTING AT THE ATOMIC SCALE

Quantum computers will perform computations at the atomic scale (DiVincenzo, 1995b). We might ask at this point how close conventional computations are to this scale already. Figure 1 shows a survey made by Keyes (1988): the number of dopant impurities in the bases of bipolar transistors used for digital logic against the year. This plot may be thought of as showing the number of electrons required to store a single bit of information.

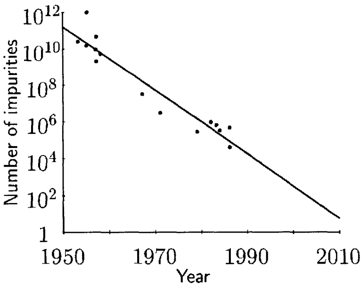

FIG. 1. Number of dopant impurities involved in logic in bipolar transistors versus year. (From Keyes, 1998; copyright 1988 by International Business Machines Corporation, reprinted with permission.)

An extrapolation of the plot suggests that we might be within reach of the atomic-scale computations within the next two decades.

Another way of viewing this plot is perhaps even more relevant for the development of quantum computation: conventional computers have been improving in speed and miniaturization at an exponential rate since their earliest days. Clearly there is a bound to our ability to miniaturize conventional electronics, and we will likely be touching that limit within the next 20 years. The question is raised: can we continue to expect to see an exponential improvement in performance 20 and more years from now? As we approach some of the physical limits to conventional computational construction we may begin to see a slowdown of this exponential rate. A detailed study of quantum computation may help us understand the fundamental physical limitations upon computation, conventional or otherwise.

2. REVERSIBLE COMPUTATION

What are the difficulties in trying to build a classical computing machine on such a small scale? One of the biggest problems with the program of miniaturizing conventional computers is the difficulty of dissipated heat. As early as 1961 Landauer studied the physical limitations placed on computation from dissipation (Landauer, 1961). Surprisingly, he was able to show that all but one operation required in computation could be performed in a reversible manner, thus dissipating no heat! The first condition for any deterministic device to be reversible is that its input and output be uniquely retrievable from each other. This is called logical reversibility. If, in addition to being logically reversible, a device can actually run backwards then it is called physically reversible and the second law of thermodynamics guarantees that it dissipates no heat.

The work on classical, reversible computation laid the foundation for the development of quantum mechanical computers. On a quantum computer, programs are executed by unitary evolution of an input that is given by the state of the system. Since all unitary operators U are invertible with $U^{-1} = U^{\dagger}$, we can always "uncompute" (reverse) a computation on a quantum computer.

3. CLASSICAL UNIVERSAL MACHINES AND LOGIC GATES

We now review the basic logic elements used in computation and explain how conventional computers may be used for any "reasonable" computation. A reasonable computation is one that may be written in terms of some (possibly large) Boolean expression, and any Boolean expression may be constructed out of a fixed set of logic gates. Such a set (*e.g.*, AND, OR, and NOT) is called universal. In fact we can get by with only two gates, such as AND and NOT or OR and NOT. Alternatively, we may replace some of these primitive gates by others, such as the exclusive-OR (called XOR or often controlled-NOT); then AND and XOR form a universal set. The truth tables for these gates are displayed in Table 1. Any machine that can build up arbitrary combinations of logic gates from a universal set is then a universal computer. Further, which universal set of gates is chosen makes little difference: a theorem by Muller (1956) states that the complexity of the simplest circuits needed to compute any reasonable Boolean function is affected by at most a constant multiplicative factor.

Which of the above gates is reversible? Since AND, OR, and XOR are many-to-one operations, information is lost and they are not, as they stand, logically reversible. Before we discuss how these logic gates may be made reversible we consider some nonstandard gates that we shall require.

Table 1. Truth table defining the operation of some simple logic gates. Each row shows two input values A and B and the corresponding output values for gates AND, OR, and XOR. The output for the NOT gate is shown only for input B.

A	B	AND	OR	XOR	NOT B
0	0	0	0	0	1
0	1	0	1	1	0
1	0	0	1	1	1
1	1	1	1	0	0

3.1 FANOUT and ERASE

Although the above gates are sufficient for the mathematics of logic, they are not sufficient to build a machine. A useful computer will also require the FANOUT and ERASE gates (Fig. 2).

First consider the FANOUT gate: is it reversible? Certainly no information has been destroyed, and so it is at least logically reversible. Landauer (1961) showed that it could also be physically reversible. Let us describe a simple model for FANOUT based on Bennett's (1988) scheme for a reversible measurement (Fig. 3). [We note, however, that the concepts involved in this scheme come from Toffoli (1980) and Fredkin and Toffoli (1982).] Here a dark ball is used to determine the presence or absence of a second (light) ball inside a trap. The trap consists of a set of mirrors and may be thought of as a one-bit memory register. If the trap is occupied then the dark ball is reflected and leaves along direction M (with the

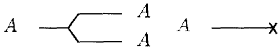

(a) FANOUT (b) ERASE

FIG. 2. Two nonstandard gates that are required to build a computer, in addition to a universal set of logic gates: (a) the FANOUT gate, which duplicates an input A; and (b) the ERASE gate, which deletes its input.

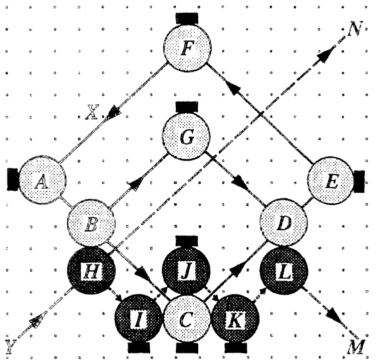

FIG. 3. A reversible measurement of the existence of a (light) ball in a trap of mirrors (dark rectangles). A (dark) ball enters the trap from *Y*. In the absence of a light ball in the trap the dark ball will follow the path *HN*. In presence of a light ball (timed to start at *X*) the dark ball will deflect the light one from its unhindered trajectory *ABCDEF* to *ABGDEF* and will follow the path *HIJKLM* itself. (From Bennett, 1998; copyright 1988 by International Business Machines Corporation, reprinted with permission.)

light ball continuing along its original trajectory); otherwise it passes unhindered towards *N*. Upon leaving the trap, the dark ball's direction is used to populate, or not, another trap.

Let us now consider the ERASE operation, required to "clean out" the computer's memory periodically. One type of erasure can be performed reversibly: if we have a backup copy of some information, we can erase further copies by uncomputing the FANOUT gate. The difficulty arises when we wish to erase our last copy, referred to here as the primitive ERASE.

Consider a single bit represented by a pair of equally probable classical states of some particle. To erase the information about the particle's state we must irreversibly compress phase-space by a factor of two. If we allowed this compressed phase space to expand, at temperature *T*, to its original size, we could obtain an amount of work equal to $k_B T \ln 2$ (where k_B is Boltzmann's constant). Landauer (1961) concluded, on the basis of simple models and more general arguments about the compression of phase space, that the erasure of a bit of information at temperature *T* requires the dissipation of at least $k_B T \ln 2$ heat (a result known as Landauer's principle).

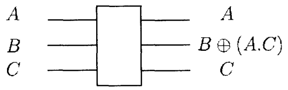

FIG. 4. Three-input, three-output universal reversible Toffoli gate. This gate is clearly reversible since a second application of it retrieves the original input.

3.2 Computation without ERASE

Fortunately, the primitive ERASE is not absolutely essential in computation. To see why, consider what is required to compute arbitrary functions using reversible logic (where the primitive ERASE is forbidden). Landauer showed how any function $f(a)$ could be made one-to-one by keeping a copy of the input:

$$f : a \rightarrow (a, f(a)). \tag{1}$$

Here the bold parentheses represent an ordered set of values, in this case, two. Extra "slots" will be added (or removed) as required in our discussion below.

How can this trick be used to perform reversible logic? One solution, known as the Toffoli gate (Toffoli, 1980; Fredkin and Toffoli 1982; Landauer 1961), is shown in Fig. 4. The output of this gate may be decomposed into various gates:

$$B \oplus (A \cdot C) = \begin{cases} A \cdot C, & \text{for } B = 0 \quad \text{(AND)}, \\ A \oplus B, & \text{for } C = 1 \quad \text{(XOR)}, \\ \bar{A}, & \text{for } B = C = 1 \quad \text{(NOT)}, \\ A, & \text{for } B \neq C = 1 \quad \text{(FANOUT)}, \end{cases} \tag{2}$$

where $A \cdot B$ represents an AND gate, $A \oplus B$ represents an XOR gate, and $\bar{A}$ represents a NOT gate. We see that this gate is universal, because it performs AND, XOR, NOT, or FANOUT depending on its inputs. A combination of many such gates could then be used for any computation and would still be reversible.

As noted by Landauer, this procedure leads to an immediate problem because of the absence of the primitive ERASE. The more gates we employ, the more "junk" bits we accumulate: at each gate we must save input bits in order to preserve reversibility. In other words a computer built out of reversible logic instead of conventional, irreversible logic gates would behave like

$$f : a \to (a, j(a), f(a)), \tag{3}$$

with many extra junk bits $j(a)$.

Bennett (1973 and 1989; Li *et al.*, 1997) solved this problem by showing that the junk bits could be reversibly erased at intermediate steps with minimal run-time and memory costs. The spirit of Bennett's solution may be understood in terms of the following procedure:

$$f : a \to (a, j(a), f(a)) \tag{4a}$$

$$\text{FANOUT} : (a, j(a), f(a)) \to (a, j(a), f(a), f(a)) \tag{4b}$$

$$f^{\dagger} : (a, j(a), f(a), f(a)) \to (a, f(a)), \tag{4c}$$

where $f^{\dagger}$ denotes uncomputing f, as opposed to computing f^{-1}. First, f is computed, producing both junk bits and the desired output. Then the FANOUT gate is applied to duplicate the output. Finally, we uncompute the original function f by running its computation backwards. This procedure removes the junk bits and the original output. The duplicate, however, remains!

This completes our discussion of the construction of classical, reversible computers. We have found that reversibility does not bar the logical design of computing machines. Before mapping these ideas to quantum systems, however, we introduce some elementary quantum mechanical notation.

4. ELEMENTARY QUANTUM NOTATION

A simple quantum system is the two-level spin-$\frac{1}{2}$ particle. Its basis states, spin down $|\downarrow\rangle$ and spin up $|\uparrow\rangle$, may be relabeled to represent binary zero and one, *i.e.*, $|0\rangle$ and $|1\rangle$, respectively. The state of a single such particle is described by the wave function $\psi = \alpha|0\rangle + \beta|1\rangle$, *i.e.*, a linear superposition among any of the possible "classical" states of the system. The squares of the complex coefficients $|\alpha|^2$ and $|\beta|^2$ represent the probabilities for finding the particle in the corresponding states. Generalizing this to a set of k spin-$\frac{1}{2}$ particles we find that there are now 2^k basis states (quantum mechanical vectors that span a Hilbert space) corresponding, say, to the 2^k possible bit strings of length k. Freely moving between decimal, binary, and spin labels then we might write, for example for $k = 5$, a state $|25\rangle = |11001\rangle = |\uparrow\uparrow\downarrow\downarrow\uparrow\rangle$.

The dimensionality of the Hilbert space grows exponentially with k. In some very real sense quantum computations make use of this enormous size latent in even the smallest systems.

5. LOGIC GATES FOR QUANTUM BITS

In this section we describe how arbitrary logic gates may be constructed for quantum bits. We start by considering various one-bit unitary operations and a single two-bit one—the XOR operation. Combinations of these are sufficient to construct a Toffoli gate for quantum bits or, indeed, *any* unitary operation on a finite number of bits.

Start with a single quantum bit. If we represent the states $|\downarrow\rangle$ and $|\uparrow\rangle$ (*i.e.*, $|0\rangle$ and $|1\rangle$) as the vectors $\binom{1}{0}$ and $\binom{0}{1}$, respectively, then the most general unitary transformation corresponds to a 2×2 matrix of the form

$$U_\theta \equiv \begin{pmatrix} e^{i(\delta+\sigma/2+\tau/2)}\cos(\theta/2) & e^{i(\delta+\sigma/2-\tau/2)}\sin(\theta/2) \\ -e^{i(\delta-\sigma/2+\tau/2)}\sin(\theta/2) & e^{i(\delta-\sigma/2-\tau/2)}\cos(\theta/2) \end{pmatrix}, \tag{5}$$

where we typically take $\delta = \sigma = \tau = 0$ (Barenco *et al.*, 1995a). Using this operator we can flip bits via

$$U_\pi|0\rangle = -|1\rangle, \quad \text{and} \quad U_\pi|1\rangle = |0\rangle. \tag{6}$$

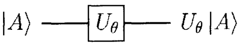

FIG. 5. Schematic of the quantum circuit diagram for a one-bit gate. The line represents a single quantum bit (such as a spin-$\frac{1}{2}$ particle). Initially, this bit has a state described by $|A\rangle$; after it has "passed" through this circuit it comes out in the state $U_\theta|A\rangle$.

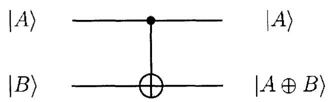

FIG. 6. Quantum circuit diagram for an XOR gate. The lower bit $|B\rangle$ is flipped whenever the upper bit $|A\rangle$ is set.

The extraneous sign represents a phase factor that does not affect the logical operation of the gates and may be removed if we wish, now or at a later stage. Such one-bit computations are illustrated schematically as a quantum circuit in Fig. 5 (Barenco *et al.*, 1995a; DiVincenzo 1995a).

Another important one-bit gate is $U_{-\pi/2}$, which maps a spin-down particle to an equal superposition of down and up:

$$U_{-\pi/2}|0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle). \tag{7}$$

Consider a string of k spin-$\frac{1}{2}$ particles initially spin-down. If we apply this gate independently to each particle we obtain a superposition of every possible bit string of length k:

$$|0\rangle \rightarrow \frac{1}{\sqrt{q}}\sum_{a=0}^{q-1}|a\rangle, \tag{8}$$

where $q = 2^k$. Our computer is now in a superposition of an exponentially large number of integers a from 0 to $2^k - 1$. Suppose further that we could construct a unitary operation that maps a pair of bit strings $|a;0\rangle$ into the pair $|a;f(a)\rangle$ for some function $f(a)$. Then such a unitary operator acting on the superposition of states

$$\frac{1}{\sqrt{q}}\sum_{a=0}^{q-1}|a;0\rangle \rightarrow \frac{1}{\sqrt{q}}\sum_{a=0}^{q-1}|a;f(a)\rangle, \tag{9}$$

would compute $f(a)$ in parallel an exponentially large number of times for the various inputs a. This follows from the linearity of quantum mechanics.

To see how such unitary operators may be constructed from a few elementary ones we must also consider the XOR gate (Barenco *et al.*, 1995a; DiVincenzo 1995a). Writing the two-particle basis states as the vectors

$$|00\rangle = \begin{pmatrix}1\\0\\0\\0\end{pmatrix}, \qquad |01\rangle = \begin{pmatrix}0\\1\\0\\0\end{pmatrix}, \tag{10a}$$

$$|10\rangle = \begin{pmatrix}0\\0\\1\\0\end{pmatrix}, \qquad |11\rangle = \begin{pmatrix}0\\0\\0\\1\end{pmatrix}, \tag{10b}$$

we may represent the XOR gate as a unitary operator

$$U_{\mathrm{XOR}} = \begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&0&1\\0&0&1&0\end{pmatrix}. \tag{11}$$

Here the first particle acts as a conditional gate to flip the state of the second particle. It is easy to check that the state of the second particle corresponds to the action of the XOR gate given in Table 1. The quantum circuit for an XOR gate is illustrated in Fig. 6. This circuit is equivalent to the elementary instruction

$$\texttt{if } (|A\rangle = \texttt{1})\ |B\rangle = \mathrm{NOT}\ |B\rangle \tag{12}$$

which may be thought of as example of quantum computer code (suggest by DiVincenzo at the 1995 Quantum Computation, Villa Gualino Workshop). The ket brackets $|\ \rangle$ are reminders that we are dealing with quantum rather than classical bits. The XOR gate allows us to move information around, as is

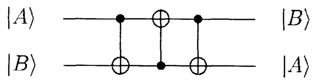

FIG. 7. Circuit for swapping a pair of bits.

illustrated in Fig. 7.

How do we construct the Toffoli gate? One major problem with this gate is that it requires three bits in and three out. Quantum mechanically, this seems to correspond to a scattering process involving three-particle collisions (Deutsch, 1989) calling for a (possibly) unreasonable control of the particles. Fortunately, the Toffoli gate may be constructed by two-particle scattering processes alone (DiVincenzo, 1995a; Barenco *et al.*, 1995b; Sleator and Weinfurter, 1995; Deutsch *et al.*, 1995; Lloyd, 1995). In particular, we show a construction here involving the XOR gate and some one-bit gates U_θ (Fig. 8) (Barenco *et al.*, 1995a). Not only is the XOR sufficient for all logic operations on a quantum computer, but it can be used to construct arbitrary unitary transformations on any finite set of bits. Numerous proposals for producing such gates have been considered (Ekert and Jozsa 1996; DiVincenzo 1995b), and we discuss some promising experimental results in the next section.

6. LOGIC GATES IN THE LABORATORY

In this section we briefly review two recent experiments that demonstrate conditional dynamics of a type that is promising for constructing quantum logic gates (Turchette *et al.*, 1995; Monroe *et al.*, 1995). These two results appeared back to back in the same issue of *Physical Review Letters*.

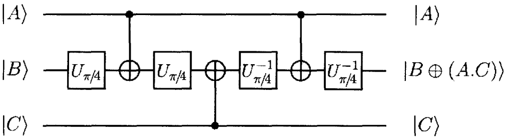

FIG. 8. Toffoli gate built from two-bit XOR gates plus some one-bit gates (Barenco *et al.*, 1995a). This circuit introduces some extra signs in the unitary matrix U_{XOR}, which may be removed at a later stage.

The first experiment, by Turchette *et al.* (1995), demonstrated that the phase of a weak coherent optical field could be controlled by the intensity of a second coherent field at a slightly different frequency. The chief result is that such a high nonlinear susceptibility was achieved that a large phase shift (up to 16°) was produced by a change in intensity corresponding to a single photon in the second field. The coupling between the optical fields was obtained using the hyperfine level structure of cesium. A stream of Cs atoms was dropped through an optical cavity, which effectively restrained the atomic decay modes to the cavity modes. This allowed the atoms to couple strongly to the optical fields passing through the cavity with minimal incoherent emission into free space. Since coherent instead of single-photon states were used in this experiment, however, there could be no direct demonstration of the coherence retained in the final state of the optical fields. In this scheme a qubit would need to be represented by single-photon states rather than weak coherent states, but that does not appear to be a great difficulty.

The second experiment, by Monroe *et al.* (1995), involved a direct demonstration of an XOR gate in a radio-frequency ion trap (also known as a Paul trap). Here the lowest vibrational excitation of a single $^9\mathrm{Be}^+$ ion in the trap and its hyperfine state represented the two qubits. A pair of off-resonant laser beams were used to drive stimulated Raman transitions between the basis states of these two qubits. The XOR gate was executed using three laser pulses, as had been suggested by Cirac and Zoller (1995). The coherence of the qubits in this system was reported to have survived for around to 10–20 XOR operations (Thompson, 1996).

Cirac and Zoller had suggested using a linear ion trap to hold a set of ions in a well localized manner by mutual electrostatic repulsion. Each ion would be "addressed" separately by its own laser. By tuning the lasers shining on individual ions to the appropriate levels, either single-qubit operations could be performed or the internal ionic state could be transferred to that of the lowest two vibrational modes of the trap. In this way two-qubit gates could be simulated between even noncontiguous ions via their interactions with the trap's vibrational modes. This ability significantly reduces the complexity of elementary operations over other proposals (Lloyd, 1993). The fact that this theoretical proposal was implemented within a few months, though in a slightly modified form, suggests that few-qubit processors could become a reality within a relatively short period of time. In Sec. 13 we discuss the short-term prospects for such machines.

In quantum computation we normally aim at having the qubits couple maximally to each other and minimally to the outside world. This lessens the actions of environment-induced decoherence. From this perspective the latter scheme, involving ion traps, appears more nearly ideal. However, there is another class of quantum logic processor, which aims at communicating quantum information between various, possibly distantly, separated subsystems. This might allow for the combining of smaller quantum computers into larger ones operating on more qubits through combination. It is likely also to be important for the area of quantum communication and potential technologies that few-bit quantum logic processors might enable. For such tasks the former scheme involving "flying" qubits appears a more likely direction. It is possible that a mature quantum computation technology would have components incorporating the positive aspects of both the above schemes.

7. MODEL QUANTUM COMPUTER AND QUANTUM CODE

In this section we describe a simple abstract model for a quantum computer based on a classical computer instructing a machine to manipulate a set of spins. This model has some intrinsic limitations that make designing algorithms in a high-level language somewhat tricky. We discuss some of the rules for writing such quantum computer code as a high-level language and give an example.

Consider the following model for the operation of a quantum computer. Several thousand spin-$\frac{1}{2}$ particles (or two-level systems) are initially in some well defined state, such as spin down. A classical machine takes single spins or pairs of spins and entangles them (performing an elementary one-bit operation U_θ or the two-bit XOR gate); see Figs. 9(a), 9(b), and 9(c). These stages are repeated on different pairs of spins according to the instructions of a conventional computer program. Since the spins are entangled, we must not look at the spins at intermediate stages. We must keep the quantum superposition intact. Furthermore, nothing else may interfere with the spins that could destroy their orientation or interrupt their unitary evolution. Once this well-defined cycle of manipulation is complete the orientations of the spins are measured [Fig. 9(d)]. This set of measured orientations is the output of the computation.

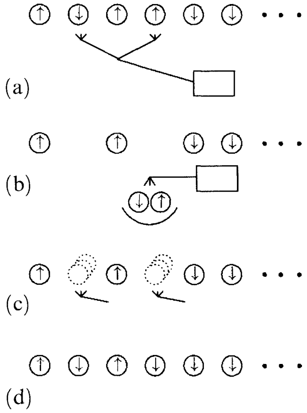

FIG. 9. Model quantum computer as pictured by Shor (presented at the 1994 Quantum Computation, Villa Gualino Workshop). Initially all particles are spin-down. In stage (a) a classical machine takes a single or pair of spins and in stage (b) it performs a selected one-bit or two-bit operation; in stage (c) the "entangled" particles are returned to their original locations. These three stages are repeated many times in accord with the instructions given by an ordinary classical computer. When this cycle is complete stage (d) consists of measuring the state of the particles (leaving them is some particular bit string); this bit string is the result of the computation.

Given this paradigm for a quantum computer, what might its high-level language (its computer code) look like? The most serious difficulty that must be dealt with is that the quantum information is manipulated by a conventional computer in a completely blind manner—without any access to the values of this quantum information. This means that the program cannot utilize "shortcuts" conditional on the value of a quantum variable (or register or bit). For example, loops must be iterated through exactly the same number of times independent of the values of the quantum variables. Similarly, conditional branches around large pieces of code must be broken down into repeated conditions for each step. In addition, each instruction performed upon the quantum bits must be logically reversible. Thus, ordinary assignments of a value to a variable, such as `|a⟩ = n`, are not legal and must instead be performed as increments on an initially zeroed variable, such as `|a⟩ = |a⟩ + n`.

An example of such code that could run on this machine might look like this (suggested by DiVincenzo at the *1995 Quantum Computation, Villa Gualino Workshop*):

```
   do 10 k = 1, worstdiv
   |a⟩ = |a⟩ - n
   if ( |a⟩ >=0 ) |q⟩ = |q⟩ + 1
10 continue
   do 20 k = 1, worstdiv
   if ( k > |q⟩ ) |a⟩ = |a⟩ + n
20 continue
```

This code fragment could be used to calculate the quotient and the remainder, placed in `|q⟩` and `|a⟩`, respectively, for the division of `|a⟩` by `n`; the constant `worstdiv` is the worst-case number of times the loop must be traversed. Here `|q⟩` is initially zero. Each instruction here is either a conventional computer instruction or one involving some quantum variables. The former are direct instructions for the external computer, while the latter must be interpreted as a sequence of manipulations to be performed upon the quantum bits. As it stands, this code is *not* reversible (neither is it very efficient); e.g., the label `10` gives no specification of which routes might be used to get to it. It can, however, be easily rewritten.

8. QUANTUM PARALLELISM: PERIOD OF A SEQUENCE

We now have sufficient ingredients to understand how a quantum computer can perform logical operations and compute just like an ordinary computer. In this section we describe an algorithm that makes use of the quantum parallelism that we have hinted at already: finding the period of a long sequence.

Consider the sequence

$$f(0), f(1), \ldots, f(q-1), \tag{13}$$

where $q \equiv 2^k$; we shall use quantum parallelism to find its period. We start with a set of initially spin-down particles, which we group into two sets (two quantum registers, or quantum variables):

$$|0;0\rangle = |\downarrow, \downarrow, \ldots; \downarrow, \downarrow, \ldots\rangle, \tag{14}$$

the first set having k bits, the next having sufficient for our needs. (In fact other registers are required, but by applying Bennett's solution to space management they may be suppressed in our discussion here.) On each bit of the first register we perform the $U_{-\pi/2}$ one-bit operation, yielding a superposition of every possible bit-string of length k in this register:

$$\rightarrow \frac{1}{\sqrt{q}} \sum_{a=0}^{q-1} |a;0\rangle. \tag{15}$$

The next stage is to break down the computation corresponding to the function $f(a)$ into a set of one-bit and two-bit unitary operations. The sequence of operations is designed to map the state $|a;0\rangle$ to the state $|a; f(a)\rangle$ for any input a. Now we see that the number of bits required for this second register must be at least sufficient to store the longest result $f(a)$ for any of these computations. When, however, this sequence of operations is applied to our exponentially large superposition, instead of the single input, we obtain

$$\rightarrow \frac{1}{\sqrt{q}} \sum_{a=0}^{q-1} |a; f(a)\rangle. \tag{16}$$

An exponentially large amount of computation has been performed essentially for free.

The final computational step, like the first, is again a purely quantum mechanical one.

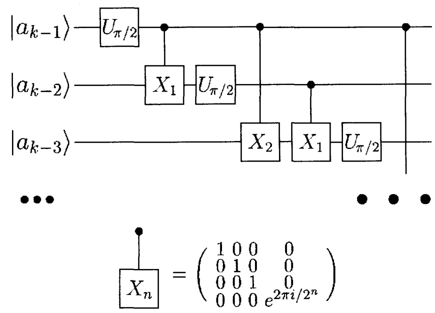
$|a_{k-1}\rangle$
$|a_{k-2}\rangle$
$|a_{k-3}\rangle$
$U_{\pi/2}$
X_1
X_2
X_n
$= \begin{pmatrix} 1\,0\,0 & 0 \\ 0\,1\,0 & 0 \\ 0\,0\,1 & 0 \\ 0\,0\,0 & e^{2\pi i/2^n} \end{pmatrix}$

FIG. 10. Circuit for the quantum Fourier transform of the variable $|a_{k-1} \ldots a_1 a_0\rangle$ using Coppersmith's (1994; Cleve 1994; DiVincenzo 1995b) fast Fourier-transform approach. The two-bit "X_n" gate may itself be decomposed into various one-bit and XOR gates (Barenco *et al.*, 1995a).

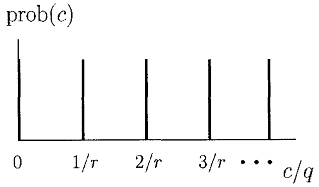

FIG. 11. Idealized plot of the probability of each result prob(c) *versus* c/q. Constructive interference produces narrow peaks at multiples of the inverse period of the sequence $1/r$. (The discrete approximation means that the peaks will actually have a nonzero width.)

Consider a discrete "quantum" Fourier transform on the first register

$$|a\rangle \rightarrow \frac{1}{\sqrt{q}} \sum_{c=0}^{q-1} e^{2\pi iac/q} |c\rangle. \tag{17}$$

It is easy to see that this is reversible via the inverse transform, and indeed it is readily verified to be unitary. Further, an efficient way to compute this transform with one-bit and two-bit gates has been described by Coppersmith (1994; Cleve 1994; DiVincenzo 1995b) (Fig. 10).

When this quantum Fourier transform is applied to our superposition, we obtain

$$\rightarrow \frac{1}{q} \sum_{a=0}^{q-1} \sum_{c=0}^{q-1} e^{2\pi iac/q} |c; f(a)\rangle. \tag{18}$$

The computation is now complete and we retrieve the output from the quantum computer by measuring the state of all spins in the first register (the first k bits). Indeed, once the Fourier transform has been performed the second register may even be discarded (Chuang *et al.*, 1996).

What will the output look like? Suppose $f(a)$ has period r so that $f(a+r) = f(a)$. The sum over a will yield constructive interference from the coefficients $e^{2\pi iac/q}$ only when c/q is a multiple of the reciprocal period $1/r$. (In fact, we must be careful that the discrete Fourier transform yields sufficient resolution to extract the multiple of the inverse period from c/q. This is always possible provided the number of bits k in the first quantum register satisfies $r^2 \leq q = 2^k$.) All other values of c/q will produce destructive interference to a greater or lesser extent. Thus, the probability distribution for finding the first register with various values is shown schematically by Fig. 11.

One complete run of the quantum computer yields a random value of c/q underneath one of the peaks in the probability of each result prob(c). That is, we obtain a random multiple of the inverse period. To extract the period itself we need only repeat this quantum computation roughly $\log\log(r/k)$ times in order to have a high probability for at least one of the multiples to be relatively prime to the period r—uniquely determining it (Shor 1994). Thus, this algorithm yields only a probabilistic result. Fortunately, we can make this probability as high as we like.

All the above work may appear a little anticlimactic. We have gone to a lot of trouble to design a quantum computer to find the period of a sequence. The point is, however, that the sequence is calculated in parallel and is exponentially long—even for a small value of, say, $k = 270$ bits in the first register, the quantum computer has calculated and stored more results than there are particles in the universe.

This algorithm for finding the period of an exponentially long sequence on a quantum computer lies at the heart of efficiently factoring numbers. We first proceed to review the computational difficulty of factoring for conventional computers. Then we discuss the implications this computational difficulty has had for the secure transmission of private information via public key cryptosystems. We then follow these discussions with Shor's (1994) new result for efficient factoring.

9. THE COMPLEXITY OF FACTORING

How can we quantify the difficulty of solving a problem with a conventional computer? Surely once the computer program is written and debugged we may simply let it run and wait for the answer. But this brings us to the crux of the difficulty. For a given problem how long must we expect to wait for the solution? When more carefully phrased this becomes the simplest measure of computational difficulty of an algorithm, yielding the "algorithmic complexity" of the problem.

To be more specific, without getting into technicalities, let us consider the problem of factoring a number N into its prime factors (*e.g.*, the number 51 688 may be decomposed as $2^3 \times 7 \times 13 \times 71$). A convenient way to quantify how quickly a particular algorithm may solve this, or any, problem is to ask how the number of steps to complete the algorithm scales with the size of the "input" the algorithm is fed. For the factoring problem, this input is just the number N we wish to factor; hence the length of the input is $\log N$. (The base of the logarithm is determined by our numbering system. Thus a base of 2 gives the length in binary; a base of 10 in decimal. For example, the number 51 688 requires 16 binary digits, but only five decimal digits, to specify it.) "Reasonable" algorithms are ones that scale as some small-degree polynomial in the input size (with a degree of perhaps 2 or 3). One famous example of a fast algorithm is the fast Fourier transform, which requires roughly $O(M \log_2 M)$ steps to perform the discrete Fourier transform of M points (so for a fixed precision the input scales as M); by contrast, a conceptually simpler algorithm equivalent to matrix multiplication would require $O(M^2)$ computational steps (Press *et al.*, 1988). This modest improvement from a quadratic to a roughly linear complexity has made many image-processing applications possible with even quite modest computers.

On conventional computers the best known factoring algorithm runs in $O(\exp[(64/9)^{1/3} \times (\ln N)^{1/3} (\ln \ln N)^{2/3}])$ steps (Odlyzko, 1995). This algorithm, therefore, scales exponentially with the input size $\log N$. For instance, in 1994 a 129-digit number (known as RSA129; Rivest *et al.*, 1978) was successfully factored using this algorithm on approximately 1600 workstations scattered around the world; the entire factorization took eight months (Atkins *et al.*, 1995). Using this to estimate the prefactor of the above exponential scaling, we find that it would take roughly 800 000 years to factor a 250-digit number with the same computer power; similarly, a 1000-digit number would require 10^{25} years (significantly longer than the age of the universe). This difficulty is, however, almost certainly exaggerated since it takes no account of the constant improvement in factoring algorithms and the constant speedup of computer hardware. In fact, both of these components have shown a more or less exponential improvement over the last few decades (Odlyzko, 1995), with each contributing roughly equally to the increased computational power. The difficulty of factoring large numbers is crucial for public-key cryptosystems, such as ones used by banks. There, such codes rely on the difficulty of factoring numbers with around 250 digits.

10. SECURITY AND RSA

Cryptography as a discipline aims at minimizing the affect of the dishonest. One such situation is the need for secure communication between two parties across an insecure channel. The sender encrypts a plain-text message with an encryption key yielding the cypher text. The message is sent across an insecure medium where, we must assume, an eavesdropper may have access. The receiver takes the cypher text and uses a decryption algorithm to restore the plaintext. If we assume that the eavesdropper has access to the decryption algorithm, then the receiver must have something in addition to ensure the security of the transmitted message; this is the decryption key. According to Shannon's information theory the cypher text must contain some information about the plain-text message unless the decryption key is at least as long as the message itself (Goldreich, 1995). Such a perfect cipher was invented in 1917, and is known as the Vernam cipher or one-time pad; it requires a key equal in size to the plain-text message. Thus, for perfect security we have the problem of distributing the key itself, which must be done over a secure channel such as by trusted courier. In many situations, such as banking transactions where the volume of information is very large, this is unreasonable.

An understanding of computational complexity allows us to "circumvent" the restriction of Shannon's theory. A pseudorandom-number generator can be used to generate a long almost random key from a much smaller secret key. If the lack of randomness cannot be discovered except through an unreasonable amount of computational effort by the eavesdropper then we have a secure encryption system. An excellent example of such a scheme is the U.S. Data Encryption Standard (DES), which uses a 56-bit key (Press *et al.*, 1988). One estimate of the security of DES suggests that for around one million dollars a special-purpose machine could be built to try all 2^{56} keys in a few hours, though security can be easily enhanced by multiple application and a larger effective key (Odlyzko, 1994). Clearly short keys reduce the burden of key distribution amongst single pairs of users, but for n users $n(n-1)/2$ keys would be required to allow any pair to communicate securely. This becomes unwieldy for commercial applications where millions of users may be involved.

Another approach, also based on computational complexity, is known as public-key encryption. The most popular scheme, and one used in may commercial applications, is RSA encryption (Rivest *et al.*, 1978). A person wishing to receive secret communications simply publishes a pair of numbers (N, e) that form the public key. Encryption involves converting the message to numerical data and dividing it into blocks of numbers m_j each smaller than N. Each block m_j of the message is then encrypted by its modular exponentiation

$$c_j \equiv m_j{}^e (\operatorname{mod} N), \tag{19}$$

where mod N represents modulo arithmetic (the expression is computed and only the remainder after division by N is retained). The encrypted blocks c_j are then transmitted to the receiver via a public channel. Thus, RSA (and any other public key scheme) efficiently solves the key distribution problem.

Decryption by the receiver requires knowing the inverse operation, *i.e.*, knowing the d such that

$$m_j \equiv c_j{}^d (\operatorname{mod} N), \tag{20}$$

reconstructs the original message m_j from the encrypted data c_j. The size of N makes the direct determination of d too difficult. Instead, d is constructed along with the public key pair (N, e) in an efficient manner. This construction involves choosing $N = pq$ as the product of a pair of comparably sized primes, with e relatively prime to both $p-1$ and $q-1$, and solving the much simpler problem

$$ed \equiv 1 (\operatorname{mod} p-1) \tag{21a}$$

$$ed \equiv 1 (\operatorname{mod} q-1). \tag{21b}$$

It is important to note that the best algorithms for finding d proceed by first factoring N; thus the security of RSA relies on the assumed difficulty of factoring (Schneier, 1994).

11. SHOR'S RESULT: FACTORING NUMBERS

Recently, an algorithm was developed by Shor (1994; Beckman *et al.*, 1996) of AT&T for factoring numbers on a quantum computer that runs in $O((\log N)^3)$ steps. This is cubic in the input size, so that factoring a 250-digit number with such an algorithm would require only a few billion steps. The implication is that public-key cryptosystems based on factoring may be breakable. In this section we give the classical portion of Shor's algorithm, which relates factoring to finding the period of an exponentially long sequence and hence makes the problem tractable for a quantum computer.

We wish to factor the number N. It will be sufficient to find even a single factor, since then we can reduce the problem to a simpler one. First, select a number x. Euclid's algorithm (see Appendix) could be used to compute efficiently the common factors between N and x, hence reducing our problem. We therefore assume that these numbers are co-prime. Next, consider the sequence formed by the function $f(a) = x^a (\bmod N)$. This sequence has the form

$$\begin{array}{l} 1, x, \ldots, x^{r-1}, \quad x^r, x^{r+1}, \ldots \\ \underbrace{1, x, \ldots, x^{r-1},}_{r\ \text{terms}} \quad \underbrace{1, x, \ldots,}_{r\ \text{terms}} \quad \underbrace{1, x, \ldots}_{r\ \text{terms}} \end{array} \tag{22}$$

Here the top row is just the sequence of powers $\{x^a\}$; the bottom row is the same sequence written in modulo arithmetic, namely $\{x^a (\bmod N)\}$. The number r is just the first nontrivial power where $x^r \equiv 1 (\bmod N)$. A close look at this sequence shows that it has a periodic structure with period r. Using standard algorithms this period would not be readily accessible for a long sequence. However, with the quantum computer algorithm described in Sec. 8 it could be calculated efficiently. This possibility opens up a novel way to find the factors of N, as we shall now describe.

Let us suppose that we have obtained the period r by the above quantum computer algorithm. [Note that since the period r is not known beforehand, we require $N^2 \leq q = 2^k$ for the Fourier transform step to yield sufficient resolution (Shor, 1994; Ekert and Jozsa, 1996)]. If now this period is even we may proceed with our factoring algorithm. If not, we must select another x and start again. A randomly chosen x will yield a suitably even period r 50% of the time, and so not too many trials will be needed (Shor, 1994; Ekert and Jozsa, 1996).

Having chosen an x so that the sequence $\{x^a (\bmod N)\}$ has an even period r, we rewrite the expression $x^r \equiv 1 (\bmod N)$ as the difference of two squares:

$$(x^{r/2})^2 - 1 \equiv 0 \ (\bmod N). \tag{23}$$

Expressing the left-hand-side as a product between a sum and difference we obtain

$$(x^{r/2} + 1)(x^{r/2} - 1) \equiv 0 \ (\bmod N). \tag{24}$$

This says that the product of the two terms on the left is a multiple of the number N we wish to factor. Thus, either one or the other of these terms must have a factor in common with N. The final step in the algorithm then is to calculate the greatest common divisor of these terms individually with N (see the Appendix for an efficient classical algorithm); any nontrivial common divisor will be a factor we have sought. This completes our search.

As an example, consider the number $N = 91$. Choosing $x = 3$ we find that the sequence $3^a (\bmod 91)$ has the form:

$$\begin{array}{l} a : 0, 1, 2, 3, 4, 5, 6, 7, \ldots \\ 3^a : 1, 3, 9, 27, 81, 243, 729, 2187, \ldots \\ 3^a (\bmod 91) : 1, 3, 9, 27, 81, 61, 1, 3, \ldots . \end{array} \tag{25}$$

A quantum computer could calculate the period in parallel; however, it is sufficient here to see by eye that this sequence has a period of $r = 6$ (since it is even we may proceed with the algorithm). Rearranging the expression $3^6 \equiv 1 (\bmod 91)$ as discussed above we conclude that $28 \times 26 \equiv 0 (\bmod 91)$. This implies that either $\gcd(28, 91)$ or $\gcd(26, 91)$ will be a nontrivial factor of 91 (where gcd is the greatest common divisor function). In fact, in this case, the two terms yield different factors, 7 and 13, respectively. This completes the prime factorization of 91 yielding $91 = 7 \times 13$.

12. QUANTUM ERROR CORRECTION

Building a quantum computer is a daunting task. Even within apparently small atomic-scale systems, quantum computation runs on the enormous size of Hilbert space. Quantum computation involves building a trajectory from a standard initial state to a complex final state. The main difficulty is keeping to this trajectory. To fail is to be lost in Hilbert space. The largest problem is hypersensitivity to perturbations, shifting the computational trajectory randomly from its path. Such perturbations come from an unintentional coupling to external noise (Unruh, 1995). In this section we briefly touch on quantum error correction and fault-tolerant computing, both introduced by Shor (1995, 1996), which promise to alleviate greatly the problem associated with un-

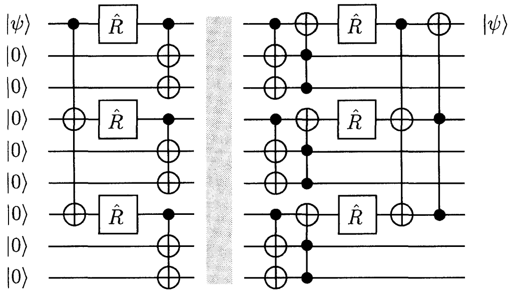

FIG. 12. Schematic of Shor's (1995) original error-correction scheme to protect one qubit in a nine-qubit code against arbitrary one-qubit decoherence. The unprotected qubit enters as qubit 1 from the left. Its error protection encoded state is just to the left of the shaded region. This region represents an arbitrary one-qubit error that may be introduced through coupling to the environment. To recover the original qubit one performs the remainder of the circuit to the right of the decoherence. (The text describes the new circuit elements.)

wanted perturbations.

In straight quantum error correction the state of a fragile quantum system is encoded into a quantum system having more degrees of freedom. By choosing the mapping to a suitable subspace of the larger system, a limited class of errors that occur on this larger space may be corrected. In Fig. 12 we see a circuit for Shor's original scheme. [Since then much work has been done in the last year (Steane, 1996, 1997a, 1997b; Calderbank and Shor, 1996; Calderbank *et al.*, 1997; Laflamme *et al.*, 1996; Knill and Laflamme, 1997; Gottesman, 1996, to cite just a few).] The unprotected single qubit $|\psi\rangle$ is processed by the left half of the circuit shown in Fig. 12 until just before the shaded region. The resulting combined nine-qubit state represents the now error-protected encoded state. Any single-qubit error (represented by the shaded region) on this encoded state may now be "undone" by sending the state through the decoding and correcting circuit shown to the right of the shaded region.

We have introduced three new circuit elements in Fig. 12, which we now explain. The first three-qubit gate involves a single control qubit (at the heavy dot) from qubit 1 to qubits 4 and 7. This gate is a shorthand notation for a pair of XOR gates (or controlled-NOT gates) between gates 1 and 4 and gates 1 and 7. The second new three-qubit gate involves two control qubits (see, for example, the last gate in the decoding circuit). This is a Toffoli gate with the condition to flip qubit 1 given by the logical AND of qubits 4 and 7 (recall that these logical operations occur separately for each branch of the wave function). (In this circuit a Toffoli gate without extraneous phases is required.) Finally, the single-qubit gate labeled $\hat{R}$ is defined by

$$\hat{R}|0\rangle = (|0\rangle + |1\rangle)/\sqrt{2}, \tag{26a}$$

$$\hat{R}|1\rangle = (|0\rangle - |1\rangle)/\sqrt{2}, \tag{26b}$$

which is a variation of the operation $\hat{U}_{-\pi/2}$ discussed earlier.

The above scheme and its variations allow for the long-term storage of quantum information. Unfortunately, they require that the encoding and decoding circuitry operate without error. Further, the above circuitry only slows the rate of decoherence, but does not eliminate its effect. By contrast, fault-tolerant quantum computation can get around these problems and, in principle, allows for unlimited quantum computation provided a threshold in accuracy for each elementary gate's operation in the presence of decoherence can be surpassed. Fault-tolerant computation involves a redesign of the error-correction and computation circuitry so as to allow for the imperfect operation of individual gates or for a decoherence event during computation. The future success of quantum computation will almost certainly rely on these techniques and upon our ability to determine theoretically, and experimentally surpass, the threshold (see Shor, 1996; DiVincenzo and Shor, 1996; Plenio *et al.*, 1997; Aharonov and Ben-Or, 1996; Steane, 1997d; Zalka, 1996; Gottesman, 1997).

13. PROSPECTS

We now describe the likely prospects for quantum computation. In this article we have discussed a single algorithm yielding an exponential speedup over conventional methods: effectively the calculation of the period of a long sequence. To date this is the only algorithm displaying such a speedup. This algorithm was applied to a traditional computer-science problem, factoring, only by recognizing a deeper structure within that problem. This requirement appears to be a general one: quantum parallelism will only yield an exponential speedup in problems whose structure avoids the need to try exponentially many solutions (Jozsa, 1991; Deutsch and Jozsa, 1992; Yao, 1993; Bennett *et al.*, 1997). Thus, a brute-force approach to some of the hardest computational questions, known as NP-complete problems, will not succeed with the aid of quantum parallelism. Any progress for such problems will require finding a deeper structure within them. Thus, it is possible that the direction the field of quantum computation will take will be along the line of simulating (or manipulating) small quantum systems (Feynman, 1982, 1986; DiVincenzo, 1995b; Lloyd, 1996). Notwithstanding the above comments, another class of algorithms for quantum computers gives more modest gains, though still sizable in comparison with the best classical algorithms. This class is based on Grover's (1996a, 1996b, 1997; Boyer *et al.*, 1996) database-searching algorithm which searches a "virtual" database for a specific entry. Unlike a classical database, a virtual database is a function that must be computed each time it is used. For searching a unique item from such a virtual database of N items, only $\pi\sqrt{N}/4$ iterations calling the virtual database are required for a close to 100% efficiency compared with only a 50% probability of success requiring $N/2$ iterations for a conventional computer. Certainly if quantum computers ever get to rival the complexity of conventional computers, the Grover algorithm will play a central role in their surpassing their rival's performance for the traditionally hard problems of computer science.

How difficult will it be to build a quantum computer? Currently, several implementations are being considered by theoreticians and experimentalists worldwide (Barenco *et al.*, 1995b; Sleator and Weinfurter, 1995; Turchette *et al.*, 1995; Monroe *et al.*, 1995; Cirac and Zoller, 1995; Pellizzari *et al.*, 1995; Lloyd, 1993; Gershenfeld and Chuang, 1997; Loss and DiVincenzo, 1997). The most promising scheme to date appears to involve linear ion traps (Monroe *et al.*, 1995; Cirac and Zoller, 1995; Pellizzari *et al.*, 1995). There are already several theoretical studies investigating the limitations to these systems, and for numerous reasons it seems that ion-trap quantum computers will be limited to computations involving no more than around 10–20 ions (Plenio and Knight, 1996a, 1996b; Hughes *et al.*, 1996; Steane, 1997c). It seems likely, then, that the first generation of quantum computers will not be performing traditional computations but will be used for the manipulation of small amounts of quantum information: a quantum information processor employed possibly for quantum cryptography, quantum teleportation, quantum storage, and quantum communication of quantum information. Indeed, these nontraditional tasks will probably lead to new types of technology, even with the relatively modest quantum computers we will be capable of building within a few years.

Over the next couple of decades we will approach computing at the atomic scale. Heat dissipation will become an ever increasing problem. The lessons from reversible classical computation and quantum computation may help us overcome this engineering hurdle and may even open doors to new faster algorithms.

GLOSSARY

Algorithmic Complexity: See **Computational Complexity**.

Computational Complexity: A measure of the complexity of a question as measured by the number of computational steps that must be performed to find the answer. The scaling relation of the number of steps required to the size of the input information described in the question is typically quoted as the associated complexity of that question (or class of questions).

Environment: Any uncontrolled external degrees of freedom.

Fault-Tolerant Computation: Computation that is robust even when a modest amount of noise is present in the logic gates performing operations and the wires connecting them.

Modular Exponentiation: The operation of taking the remainder after division by N of some number m to a power e, is called the modular exponential; it is written $m^e (\bmod N)$.

Public-Key Cryptography: A method of encoding secret information to a specific recipient who has made his/her encoding key publically known. The security of this scheme relies on the computational complexity of any attempt to invert the known encoding procedure.

Quantum Bit: See **Qubit**.

Quantum Error Correction: A scheme to "reverse" the affects of dissipation, decoherence, or dephasing due to inadvertent coupling to an environment.

Quantum Fourier Transform: A quantum mechanical analog of the discrete Fourier transform acting on a superposition of quantum states.

Qubit: Quantum bit of information; the amount of information that can be stored by a two-level quantum system.

Reversible Computer: A hypothetical computer that can actually "run" backwards, yielding its input from its output.

RSA: An accronym based on the names, Rivest, Shamir, and Adleman, of the inventors of the prototypically public-key cryptographic system.

Toffoli Gate: A specific three-bit gate that is universal for classical reversible computations.

Turing Machine: Prototypical classical universal computer using a read/write head and one or more paper tapes.

APPENDIX

Here we describe Euclid's algorithm for finding the greatest common divisor (gcd) between a pair of numbers $n_0 \geq n_1$ (Hardy and Wright 1979). The algorithm proceeds by calculating the sequence of divisions with remainder for these numbers:

$$\begin{aligned} n_0 &= d_1 \times n_1 + n_2 \\ n_1 &= d_2 \times n_2 + n_3 \\ &\vdots \\ n_{m-2} &= d_{m-1} \times n_{m-1} + n_m \\ n_{m-1} &= d_m \times n_m + 0, \end{aligned} \tag{A1}$$

where the d_m are the quotients and $n_{m-1} \geq n_m$ at each stage. The last nonzero remainder n_m yields the answer, *i.e.*, $\gcd(n_0, n_1) = n_m$. For example, the sequence

$$91 = 3 \times 28 + 7 \tag{A2a}$$

$$28 = 4 \times 7 + 0, \tag{A2b}$$

shows that $\gcd(28, 91) = 7$ in just two steps. The worst-case number of steps required to complete Euclid's algorithm is $O(\log \log n_1)$.

Works Cited

Aharonov, D., Ben-Or, M. (1996), LANL preprint No. quant-ph/9611025.

Atkins, D., Graff, M., Lenstra, A. K., Leyland, P. C. (1995), in: J. Pieprzyk, R. Safavi-Naini (Eds.), *Advances in Cryptology—ASIACRYPT '94*, Lecture Notes in Computer Science No. 917, Berlin: Springer Verlag, pp. 263–277.

Barenco, A., Bennett, C. H., Cleve, R., DiVincenzo, D. P., Margolus, N., Shor, P., Sleator, T., Smolin, J. A. (1995a), *Phys. Rev. A* **52**, 3457–3467.

Barenco, A., Deutsch, D., Ekert, A. (1995b), *Phys. Rev. Lett.* **74**, 4083–4086.

Beckman, D., Chari, A. N., Devabhaktuni, S., Preskill, J. (1996), *Phys. Rev. A* **54**, 1034–1063.

Benioff, P. (1980), *J. Stat. Phys.* **22**, 563–591.

Benioff, P. (1981), *J. Math. Phys.* **22**, 495–507.

Benioff, P. (1982), *Phys. Rev. Lett.* **48**, 1581–1585.

Bennett, C. H. (1973), *IBM J. Res. Develop.* **17**, 525–532.

Bennett, C. H. (1988), *IBM J. Res. Develop.* **32**, 16–23.

Bennett, C. H. (1989), *SIAM J. Comput.* **18**, 766–776.

Bennett, C. H., Bernstein, E., Brassard, G., Vazirani, U. V. (1997), *SIAM J. Comput.* **26**, 1510–1523.

Boyer, M., Brassard, G., Hoeyer, P., Tapp, A. (1996), Fortschr. Phys. **46**, 493–505.

Calderbank, A. R., Shor, P. W. (1996), *Phys. Rev. A* **54**, 1098–1105.

Calderbank, A. R., Rains, E. M., Shor, P. W., Sloane, N. J. A. (1997), *Phys. Rev. Lett.* **78**, 405–408.

Chuang, I. L., Laflamme, R., Shor, P. W., Zurek, W. H. (1996), *Science* **270**, 1633–1635.

Cirac, J. I., Zoller, P. (1995), *Phys. Rev. Lett.* **74**, 4091–4094.

Cleve, R. (1994), University of British Columbia preprint.

Coppersmith, D. (1994), IBM Research Report No. RC19642.

Deutsch, D. (1985a), *Int. J. Theor. Phys.* **24**, 1–41.

Deutsch, D. (1985b), *Proc. Roy. Soc. London, A* **400**, 97–117.

Deutsch, D. (1989), *Proc. Roy. Soc. London, A* **425**, 73–90.

Deutsch, D., Jozsa, R. (1992), *Proc. R. Soc. London, A* **439**, 553–558.

Deutsch, D., Barenco, A., Ekert, A. (1995), *Proc. R. Soc. London, A* **449**, 669–677.

DiVincenzo, D. P. (1995a), *Phys. Rev. A* **51**, 1015–1022.

DiVincenzo, D. P. (1995b), *Science* **270**, 255–261.

DiVincenzo, D. P., Shor, P. W. (1996), *Phys. Rev. Lett.* **77**, 3260–3263.

Ekert, A., Jozsa, R. (1996), *Rev. Mod. Phys.* **68**, 733–753.

Feynman, R. P. (1982), *Int. J. Theor. Phys.* **21**, 467–488.

Feynman, R. P. (1986), *Found. Phys.* **16**, 507–531.

Fredkin, E., Toffoli, T. (1982), *Int. J. Theor. Phys.* **21**, 219–253.

Gershenfeld, N., Chuang, I. (1997), *Science* **275**, 350–356.

Goldreich, O. (1995), *Foundations of Cryptography (Fragments of a book)*, http://theory.lcs.mit.edu/~oded.

Gottesman, D. (1996), *Phys. Rev. A* **54**, 1862–1868.

Gottesman, D. (1997), LANL preprint No. quant-ph/9702029.

Grover, L. K. (1996a), in: *Proceedings, 28th Annual ACM Symposium on the Theory of Computing* New York: ACM, pp. 212–219.

Grover, L. K. (1996b), LANL preprint No. quant-ph/9607024.

Grover, L. K. (1997), *Phys. Rev. Lett.* **79**, 325–328.

Hardy, G. H., Wright, E. M. (1979), *An Introduction to the Theory of Numbers*, Oxford: Clarendon Press.

Hughes, R. J., James, D. F. V., Knill, E. H., Laflamme, R., Petschek, A. G. (1996), *Phys. Rev. Lett.* **77**, 3240–3243.

Jozsa, R. (1991), *Proc. R. Soc. London, A* **435**, 563–574.

Keyes, R. W. (1988), *IBM J. Res. Develop.* **32**, 24–28.

Knill, E., Laflamme, R. (1997), *Phys. Rev. A* **55**, 900–911.

Laflamme, R., Miquel, C., Paz, J. P., Zurek, W. H. (1996), *Phys. Rev. Lett.* **77**, 198–201.

Landauer, R. (1961), *IBM J. Res. Develop.* **3**, 183–191.

Li, M., Tromp, J., Vitanyi, P. (1997), *Physica D* **120**, 168–176.

Lloyd, S. (1993), *Science* **261**, 1569–1571.

Lloyd, S. (1995), *Phys. Rev. Lett.* **75**, 346–349.

Lloyd, S. (1996), *Science* **273**, 1073–1078.

Loss, D., DiVincenzo, D. P. (1997), *Phys. Rev. A* **57**, 120–126.

Monroe, C., Meekhof, D. M., King, B. E., Itano, W. M., Wineland, D. J. (1995), *Phys. Rev. Lett.* **75**, 4714–4717.

Muller, D. E. (1956), *IRE Trans. Elec. Comput.* **5**, 15–19.

Odlyzko, A. M. (1994), *AT&T Tech. J.* **73**, 17–23.

Odlyzko, A. M. (1995), *Cryptobytes: The Technical Newsletter of RSA Laboratories*, Summer.

Pellizzari, T., Gardiner, S. A., Cirac, J. I., Zoller, P. (1995), *Phys. Rev. Lett.* **75**, 3788–3791.

Plenio, M. B., Knight, P. L. (1996a), *Phys. Rev. A* **53**, 2986–2990.

Plenio, M. B., Knight, P. L. (1996b), *Proc. R. Soc. Lond. A* **453**, 2017–2041.

Plenio, M. B., Vedral, V., Knight, P. L. (1997), *Phys. Rev. A* **55**, 4593–4596.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T. (1988), *Numerical Recipes: The Art of Scientific Computing*, Cambridge, U.K.: Cambridge Univ. Press, p. 228 and p. 390.

Rivest R., Shamir, A., Adleman, L. (1978), *Communications ACM* **21**, 120–126.

Schneier, B. (1994), *Applied Cryptography*, New York: Wiley.

Shor, P. W. (1994), in: S. Goldwasser (Ed.), *Proceedings of the 35th Annual Symposium on the Foundations of Computer Science*, Los Alamitos, CA: IEEE Computer Society Press, pp. 124–134.

Shor, P. W. (1995), *Phys. Rev. A* **52**, R2493–R2496.

Shor, P. W. (1996), in: *Proceedings of the 37th Annual Symposium on the Foundations of Computer Science*, Los Alamitos, CA: IEEE Computer Society Press, pp. 56–65.

Sleator, T., Weinfurter, H. (1995), *Phys. Rev. Lett.* **74**, 4087–4090.

Steane, A. M. (1996), *Phys. Rev. Lett.* **77**, 793–797.

Steane, A. M. (1997a), *Proc. Roy. Soc. London*, **452**, 2551–2576.

Steane, A. M. (1997b), *Phys. Rev. A* **54**, 4741–4751.

Steane, A. M. (1997c), *Appl. Phys. B* **B64**, 623–643.

Steane, A. M. (1997d), *Phys. Rev. Lett.* **78**, 2252–2255.

Thompson, T. (1996), *Byte*, **21**(4), 45–54.

Toffoli, T. (1980), in: J. W. de Bakker, J. van Leeuwen (Eds.), *Automata, Languages and Programming*, New York: Springer-Verlag, pp. 632–644.

Turchette, Q. A., Hood, C. J., Lange, W., Mabuchi, H., Kimble, H. J. (1995), *Phys. Rev. Lett.* **75**, 4710–4713.

Unruh, W. G. (1995), *Phys. Rev. A* **51**, 992–997.

Yao, A. C.-C. (1993), in: *Proceedings of the 34th Annual Symposium on Foundations of Computer Science*, New York: IEEE, pp. 352–361.

Zalka, C. (1996), LANL preprint, quant-ph/9612028.

SEMICONDUCTORS, ISOTOPICALLY ENGINEERED

E. E. Haller *Lawrence Berkeley National Laboratory and University of California at Berkeley, Berkeley, California, U.S.A.*

Introduction 257
1. **Materials Aspects** 258
2. **Phonons and Related Properties** 259
2.1 Lattice Dynamics in Bulk Crystals 259
2.1.1 Phonons and Thermal Conductivity 259
2.1.2 Lattice Constants and Bulk Modulus 265
2.1.3 Band Gaps 267
2.2 Isotope Superlattices 269
3. **Local Vibrational-Mode Spectroscopy** 269
4. **Diffusion** 271
5. **Neutron Transmutation Doping (NTD) and Electronic Transport Studies** 274
5.1 Bulk Crystals 274
5.2 Selectively Doped Isotopic Superlattices 276
5.3 Electronic Transport and Metal–Insulator Transition 277
6. **Future Developments and Conclusions** 279
Acknowledgments 280
Works Cited 281
Further Reading 283

INTRODUCTION

Most compound semiconductors contain at least one element that consists of more than one stable isotope. Silicon, the most commonly used semiconductor, for example, is composed of the three isotopes $^{28}_{14}$Si, $^{29}_{14}$Si and $^{30}_{14}$Si with abundances of 92.23, 4.67 and 3.10%, respectively. The important compound semiconductor GaAs consists of monoisotopic $^{75}_{33}$As and two Ga isotopes $^{69}_{31}$Ga and $^{71}_{31}$Ga, with abundances of approximately 60 and 40%, respectively. The isotopic abundances are a direct consequence of the nucleosynthesis that led to the formation of the elements making up the matter in our region of the universe.

Most physical properties of semiconductor crystals depend to some degree on their isotopic composition (Berezin, 1989). While these dependences may be rather weak for many properties, they nevertheless teach us a lot about the origin of some basic physical processes occurring in a crystal and enhance our theoretical understanding. When determined with very high precision, they may establish new, rugged standards of length and mass and new, highly accurate values of Avogadro's number.

Not all isotope effects are subtle, though. For example, the room-temperature thermal conductivity of isotopically controlled diamond has been shown to change by tens of percent depending on the isotopic composition. By far the largest effects due to isotopic composition can occur after the capture of a thermal neutron by the nucleus of a specific isotope. Either the new nucleus is stable and the element remains unchanged, or it decays, transmuting into a new element. This so-called neutron-transmutation doping (NTD) is used to introduce phosphorus donors into high-purity floating-zone (FZ) silicon (Schnöller, 1974). The phosphorus atoms are created in the beta decay of $^{31}_{14}$Si, an unstable Si isotope, formed when a $^{30}_{14}$Si isotope captures a thermal neutron. The extreme homogeneity and the excellent control of the phosphorus concentration that can be achieved with NTD led to the rapid development and acceptance of NTD as a doping technique in the 1970s for Si single crystals used in the fabrication of high-power, high-voltage rectifiers and thyristors. The NTD process has in recent years been extensively applied to ultrapure Ge used in the fabrication of very low-temperature thermistors operating in the few-kelvin to millikelvin

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

range (Haller *et al.*, 1994). These thermal sensors work in the hopping conduction regime and require extremely homogeneous doping. They are used as single sensors or in two-dimensional arrays in terrestrial and space-borne infrared, millimeter-wave, and radio telescopes (Kreysa *et al.*, 1996; Holland *et al.*, 1996).

The natural isotopic composition of some compound semiconductor crystals in which the various isotopes are randomly distributed can be used to determine the sublattice on which certain impurities reside. Carbon impurities in gallium arsenide show localized vibrational modes (LVM) with several frequencies indicating that they occupy the As sublattice and that the two stable Ga isotopes, $^{69}_{31}$Ga and $^{71}_{31}$Ga, form nearest neighbors in several distinct combinations. If carbon were residing on the Ga sublattice, it would be surrounded by monoisotopic As neighbors forming identical surroundings. Such a configuration would produce one sharp line in the LVM spectrum. A second example, which is discussed in further detail, involves oxygen impurities in both Si and Ge single crystals. Occupying a bond-centered interstitial position, oxygen has two nearest neighbors. The atomic masses of these neighbor atoms slightly affect the various oxygen vibrational-mode frequencies, again creating the opportunity of a precise determination of the oxygen position.

The natural isotopic composition of a crystalline solid leads to a subtle state of disorder, which affects several intrinsic physical properties. For example, the thermal conductivity, which has been determined as a function of temperature for many semiconductor crystals, is rather sensitive to isotopic composition. Reduction or elimination of the isotope disorder through isotopic enrichment was used by Geballe and Hull (1958) to demonstrate experimentally the increase in the low-temperature thermal conductivity in Ge that was predicted by Pomeranchuk (1942) as early as 1942. In recent years the interest in the thermal conductivity of isotopically enriched diamond (treated in this article as a wide–band-gap semiconductor) has led to some extreme speculations. Careful experimentation and a good theoretical understanding have produced record values of thermal conductivity, which are currently only limited by the size of the diamonds and by the available experimental techniques. It is surprising that the 1% $^{13}_{6}$C in natural diamond leads to a reduction of the room-temperature thermal conductivity of 30% (Banholzer and Anthony, 1992; Anthony *et al.*, 1990) as compared with an isotopically pure diamond. At lower temperatures these effects become very much larger, as is shown in Sec. 2. Less dramatic effects are observed for the phonon dispersion relationship, the lattice constant, and the Raman frequencies and intensities, to name a few properties.

One of the main reasons why we may experience a strong increase in the scientific and technical interest in the control of the isotopic composition of semiconductors is closely related to the recent geopolitical changes. Russia and the Ukraine have a well developed isotope-enrichment industry and possess sizable quantities of certain highly enriched isotopes. Many of them are elements forming semiconductors. The availability of highly enriched isotopes at affordable costs will be required to change the attitude of semiconductor researchers and technologists towards isotopically controlled semiconductors. The availability of many different isotopes in tens-of-grams quantities, together with advanced thin-film crystal growth techniques such as molecular-beam epitaxy (MBE), is opening completely new opportunities in semiconductor research. Isotope superlattices of $^{70}_{32}$Ge and $^{74}_{32}$Ge, grown recently with the MBE technique, exhibit fascinating phonon spectra (Spitzer *et al.*, 1994). $^{69}_{31}$GaAs/$^{71}_{31}$GaAs multilayer thin-film structures have been grown for the study of Ga self-diffusion (Wang *et al.*, 1996; Tan *et al.*, 1992) in GaAs. Using the NTD technique, isotope superlattices can be selectively doped after they have been grown. The doping level may be increased in finished devices by reimmersion of the crystal in a thermal-neutron bath.

It is the aim of this article to familiarize the reader with the recent developments in semiconductor isotope engineering in the hope that this will stimulate many creative ideas for studies and structures generating new scientific results and in due time new semiconductor devices.

1. MATERIALS ASPECTS

The periodic table of elements shows that approximately 80% of all the elements consist of more than one stable isotope. The remaining elements are monoisotopic. Focusing on those elements that are of importance for semiconductors either by being a constituent of the semiconductor lattice or by acting as a neutral, shallow- or deep-level impurity, we find that nature offers an interesting and useful mixture of mono- and polyisotopic elements. As a general rule we find that the number of different stable isotopes for a given element increases with atomic number. There are important exceptions including Al, P, As, and Au, all of which are monoisotopic. Table 1 shows a selection of elements that are of importance for this article.

The information contained in the first two columns, the isotopes and their natural abundances, illustrates the very large number of isotopic compositions that can be formed, at least in principle, when we consider binary compound or ternary and quaternary alloy semiconductors. Because excessively small abundances and/or difficulties in forming molecules suitable for separation will lead to very large isotope separation costs, it will not be feasible to obtain large enough quantities to grow bulk single crystals or thin films of all these isotopes. However, certain experiments with isotopically engineered compounds have already been performed. For example, the phonon dispersion of natural-cadmium–based II–VI compounds including CdS, CdSe, and CdTe cannot be deduced with inelastic neutron scattering because of the extremely large thermal-capture cross section of ${}^{113}_{48}\mathrm{Cd}$, which occurs with an abundance of 12.22%. [This large cross section is the reason why Cd metal rods are used to control (to absorb!) the neutron flux in a nuclear reactor.] In order to circumvent this problem, Rowe *et al.* (1974) have used ^{114}CdTe single crystals to study lattice dynamics. Measurement of the phonon-dispersion relationships with inelastic neutron scattering of CdS and CdSe await the growth of crystals with negligible ^{113}Cd content.

A key question regarding broad-scale semiconductor isotope engineering must address the availability and costs of a given isotope. From extensive discussions with many experts it appears to this author that every stable isotope can in principle be separated and enriched with one or more techniques. However, it seems to be very difficult to establish realistic cost estimates, since these depend in complicated ways on the required quantity of the desirable isotope, on the existing reserves, on electricity, and on manpower costs. For small quantities of highly enriched isotopes (>90%) we find that isotopes present at tens of percent in naturally occurring elements may cost several US dollars per milligram (several million dollars per kilogram). As soon as several kilograms are produced, the costs decrease by factors of between 10 and 100. From this we conclude that isotope engineering will become feasible once a substantial demand for certain isotopes arises.

2. PHONONS AND RELATED PROPERTIES

2.1 Lattice Dynamics in Bulk Crystals

2.1.1 Phonons and Thermal Conductivity The strong coupling between neighboring atoms in a crystal together with the symmetry of the crystal lattice leads to a number of extended vibrational modes, *i.e.*, phonons, which can propagate through the solid. The simplest dependence of the phonon frequency ω on mass M can be described within the virtual-crystal approximation (VCA) (Vegard, 1921; Nordheim, 1931) by a "spring and ball" model and is given by

$$\omega \propto M^{-1/2}. \quad (1)$$

Figure 1 shows the phonon dispersion relationship for single-crystal Ge (Nilsson and Nelin, 1971). The two optical-phonon branches become degenerate at $\mathbf{k} = 0$, the center of the Brillouin zone. The energies of the phonons of the acoustical branches go to zero at this point. At the zone boundary $\mathbf{k} = (2\pi/a)(1/2, 1/2, 1/2)$ we find four different phonons, TO, LO, LA, and TA in order of descending energy. Inelastic neutron scattering allows the measurement of the energies of phonons spanning the full Brillouin zone. Optical techniques allow much simpler and highly accurate measurements at $\mathbf{k} = 0$ and $\mathbf{k} = (2\pi/a)(1/2, 1/2, 1/2)$. Raman spectroscopy, a photon-scattering process involving vanishingly small changes in $\mathbf{k}$, is ideally suited for the study of optical phonons at $\mathbf{k} = 0$. Pho-

Table 1. Stable isotopes of importance to semiconductors and their neutron transmutation products.

Isotope	Abundance[a] (%)	Thermal-neutron capture cross section[b] (10^{-24} cm^2)	Neutron transmutation product[b]
${}^{10}_{5}\mathrm{B}$	19.8	$\begin{cases} 3838(\mathrm{n}, \alpha) \\ 0.52 \end{cases}$	${}^{7}_{3}\mathrm{Li} + {}^{4}_{2}\mathrm{He}$ ${}^{11}_{5}\mathrm{B}$
${}^{11}_{5}\mathrm{B}$	80.2	0.005	${}^{12}_{6}\mathrm{C}$
${}^{12}_{6}\mathrm{C}$	98.89	0.0034	${}^{13}_{6}\mathrm{C}$
${}^{13}_{6}\mathrm{C}$	1.11	9×10^{-4}	${}^{14}_{7}\mathrm{N}$
${}^{14}_{7}\mathrm{N}$	99.63	0.076	${}^{15}_{7}\mathrm{N}$
${}^{15}_{7}\mathrm{N}$	0.37	4×10^{-5}	${}^{16}_{8}\mathrm{O}$
${}^{16}_{8}\mathrm{O}$	99.76	1.78×10^{-4}	${}^{17}_{8}\mathrm{O}$
${}^{17}_{8}\mathrm{O}$	0.038	$0.235(\mathrm{n}, \alpha)$	${}^{14}_{6}\mathrm{C} + {}^{4}_{2}\mathrm{He} \rightarrow {}^{14}_{7}\mathrm{N}$
${}^{18}_{8}\mathrm{O}$	0.204	1.58×10^{-4}	${}^{19}_{9}\mathrm{F}$
${}^{27}_{13}\mathrm{Al}$	100	0.231	${}^{28}_{14}\mathrm{Si}$
${}^{28}_{14}\mathrm{Si}$	92.23	0.17	${}^{29}_{14}\mathrm{Si}$
${}^{29}_{14}\mathrm{Si}$	4.67	0.10	${}^{30}_{14}\mathrm{Si}$
${}^{30}_{14}\mathrm{Si}$	3.10	0.11	${}^{31}_{15}\mathrm{P}$
${}^{31}_{15}\mathrm{P}$	100	0.18	${}^{32}_{16}\mathrm{S}$
${}^{32}_{16}\mathrm{S}$	95.02	0.53	${}^{33}_{16}\mathrm{S}$
${}^{33}_{16}\mathrm{S}$	0.75	$0.095(\mathrm{n}, \alpha)$	${}^{30}_{14}\mathrm{Si} + {}^{4}_{2}\mathrm{He}$
${}^{34}_{16}\mathrm{S}$	4.21	0.24	${}^{35}_{17}\mathrm{Cl}$
${}^{36}_{16}\mathrm{S}$	0.017	0.15	${}^{37}_{17}\mathrm{Cl}$
${}^{64}_{30}\mathrm{Zn}$	48.6	0.78	${}^{65}_{29}\mathrm{Cu}$
${}^{66}_{30}\mathrm{Zn}$	29.7	1.0	${}^{67}_{30}\mathrm{Zn}$
${}^{67}_{30}\mathrm{Zn}$	4.1	7.0	${}^{68}_{30}\mathrm{Zn}$
${}^{68}_{30}\mathrm{Zn}$	18.8	0.88	${}^{69}_{31}\mathrm{Ga}$
${}^{70}_{30}\mathrm{Zn}$	0.62	0.098	${}^{71}_{31}\mathrm{Ga}$
${}^{69}_{31}\mathrm{Ga}$	60.1	1.7	${}^{70}_{30}\mathrm{Zn}$ (0.2%) ${}^{70}_{32}\mathrm{Ge}$ (99.8%)
${}^{71}_{31}\mathrm{Ga}$	39.9	4.7	${}^{72}_{32}\mathrm{Ge}$
${}^{70}_{32}\mathrm{Ge}$	20.5	3.25	${}^{71}_{31}\mathrm{Ga}$
${}^{72}_{32}\mathrm{Ge}$	27.4	1.0	${}^{73}_{32}\mathrm{Ge}$
${}^{73}_{32}\mathrm{Ge}$	7.8	15.0	${}^{74}_{32}\mathrm{Ge}$
${}^{74}_{32}\mathrm{Ge}$	36.5	0.36	${}^{75}_{33}\mathrm{As}$
${}^{76}_{32}\mathrm{Ge}$	7.8	0.16	${}^{76}_{34}\mathrm{Se}$
${}^{75}_{33}\mathrm{As}$	100	4.4	${}^{76}_{34}\mathrm{Se}$
${}^{74}_{34}\mathrm{Se}$	0.89	52	${}^{75}_{33}\mathrm{As}$
${}^{76}_{34}\mathrm{Se}$	9.36	85	${}^{77}_{34}\mathrm{Se}$
${}^{77}_{34}\mathrm{Se}$	7.63	42	${}^{78}_{34}\mathrm{Se}$
${}^{78}_{34}\mathrm{Se}$	23.78	0.7	${}^{79}_{35}\mathrm{Br}$ (very long half-life $\approx 6.5 \times 10^{4}$ y)
${}^{80}_{34}\mathrm{Se}$	49.61	0.68	${}^{83}_{36}\mathrm{Kr}$
${}^{106}_{48}\mathrm{Cd}$	1.25	1.0	${}^{107}_{47}\mathrm{Ag}$
${}^{108}_{48}\mathrm{Cd}$	0.89	1.2	${}^{109}_{47}\mathrm{Ag}$ (long half-life 453 d)

Table 1 (continued)

Isotope	Abundance[a] (%)	Thermal-neutron capture cross section[b] (10^{-24} cm^2)	Neutron transmutation product[b]
$^{110}_{48}$Cd	12.49	11	$^{111}_{48}$Cd
$^{111}_{48}$Cd	12.80	24	$^{112}_{48}$Cd
$^{112}_{48}$Cd	24.13	2	$^{113}_{48}$Cd
$^{113}_{48}$Cd	12.22	1.98×10^4	$^{114}_{48}$Cd
$^{114}_{48}$Cd	28.73	0.34	$^{115}_{49}$In (95%) $^{115}_{50}$Sn (5%)
$^{116}_{48}$Cd	7.49	0.075	$^{117}_{50}$Sn
$^{113}_{49}$In	4.3	10.8	$^{114}_{50}$Sn (94.8%) $^{114}_{48}$Cd (5.2%)
$^{115}_{49}$In	95.7	202	$^{116}_{50}$Sn
$^{121}_{51}$Sb	57.36	6.2	$^{122}_{52}$Te (97%) $^{122}_{50}$Sn (3%)
$^{123}_{51}$Sb	42.64	4.0	$^{124}_{52}$Te (20%) $^{124}_{50}$Sn (80%)
$^{120}_{52}$Te	0.096	2.3	$^{121}_{51}$Sb
$^{122}_{52}$Te	2.603	3	$^{123}_{52}$Te
$^{123}_{52}$Te	0.908	400	$^{124}_{52}$Te
$^{124}_{52}$Te	4.816	6.7	$^{125}_{52}$Te
$^{125}_{52}$Te	7.139	1.6	$^{126}_{52}$Te
$^{126}_{52}$Te	18.95	1.0	$^{127}_{53}$I
$^{128}_{52}$Te	31.69	0.22	$^{129}_{53}$I (decays into $^{129}_{54}$Xe with a half-life of 1.6×10^7 y)
$^{130}_{52}$Te	33.80	0.23	$^{131}_{54}$Xe
$^{196}_{80}$Hg	0.15	3120	$^{197}_{79}$Au
$^{198}_{80}$Hg	9.97	0.018	$^{199}_{80}$Hg
$^{199}_{80}$Hg	16.87	2000	$^{200}_{80}$Hg
$^{200}_{80}$Hg	23.10	<60	$^{201}_{80}$Hg
$^{201}_{80}$Hg	13.18	<60	$^{202}_{80}$Hg
$^{202}_{80}$Hg	29.86	5.0	$^{203}_{81}$Tl
$^{204}_{80}$Hg	6.87	0.4	$^{205}_{81}$Tl

[a] From De Laeter *et al.* (1991).
[b] From Lederer and Shirley (1978).

toluminescence (PL) studies based on the radiative decay of free excitons allow the accurate determination of zone-boundary phonons in Ge because of the particular band structure of this semiconductor (Fig. 2). The recombination of an electron at the conduction-band minimum [$\mathbf{k} = (2\pi/a)(1/2, 1/2, 1/2)$, L point] with a hole at the valence-band maximum ($\mathbf{k} = 0$, Γ point) requires the emission or absorption of a phonon to satisfy the conservation of wave vector momentum. Raman spectroscopy and PL have been used for the study of phonons in isotopically enriched and deliberately mixed germanium crystals.

Figure 3 shows superimposed Raman lines of natural and of isotopically enriched Ge single crystals recorded near liquid-nitrogen temperatures (90 K) (Fuchs *et al.*, 1992). The lines

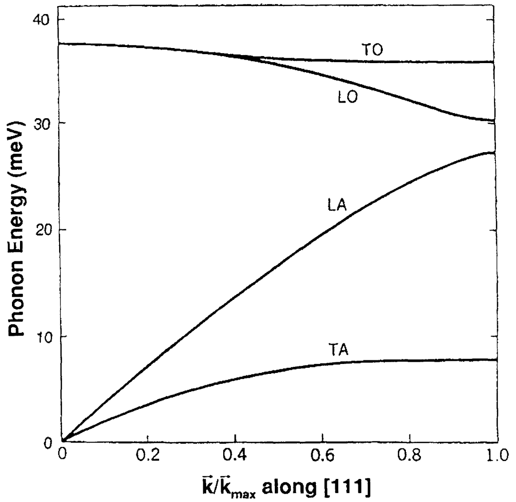

FIG. 1. Phonon dispersion relations in the [111] direction in germanium at 80 K. The two TA-phonon branches are horizontal at the zone-boundary position, ($\mathbf{k} = \pi/a$, L point). The LO and TO branches are degenerate at the center of the Brillouin zone ($\mathbf{k} = 0$, Γ point); this also is a consequence of the crystal symmetry of Ge. The results were obtained with neutron inelastic scattering. (After Nilsson and Nelin, 1971.)

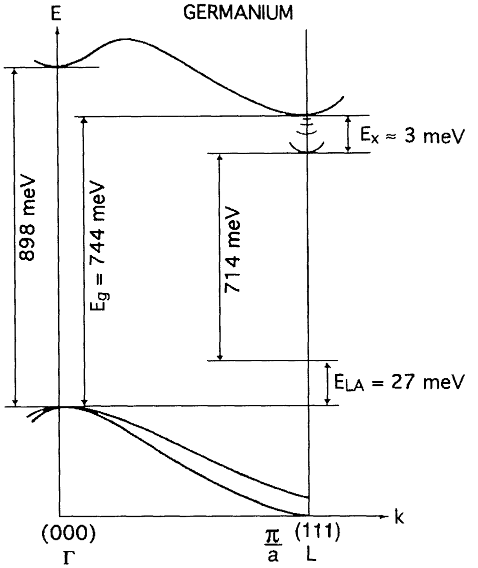

FIG. 2. Energy bands $E(\mathbf{k})$ for electrons and holes in Ge versus wave vector $\mathbf{k}$ in the $\langle 111 \rangle$ direction. The two lower lines are valence-band edges, the upper is the conduction-band edge. An electron–hole pair binds into an exciton with an energy $E_x = 3\,\text{meV}$ below the band edge; it decays by luminescence at 714 meV. The exciton decay is assisted by an LA phonon for the case shown. Other types of phonons (TA, LO, TO) can also assist exciton decay, leading to several replicas in the PL spectrum.

in these spectra are fully resolved instrumentally and their width is caused by homogeneous broadening. The centroid of the Raman line shifts following Eq. (1). What is surprising at first sight is the very weak dependence of the phonon linewidth of $\approx 1\,\text{cm}^{-1}$ on isotopic composition. The close to constant linewidth has been explained with phonon delocalization (Fuchs *et al.*, 1991a, 1991b; Etchegoin *et al.*, 1993; Cardona *et al.*, 1992, 1993), which can be discussed within the framework of Anderson localization (Anderson, 1955). If we express the relative mass fluctuations $\Delta M/\overline{M}$ in the form of a frequency bandwidth $(\Delta M/\overline{M})\omega_0 = 12\,\text{cm}^{-1}$, with ω_0 being the angular optical-phonon frequency at $\mathbf{k} = 0$, and compare it with the bandwidth of the optical phonons in the first Brillouin zone, $\Delta\omega_0 \approx 100\,\text{cm}^{-1}$, we immediately see that the mass-fluctuation bandwidth is much smaller than the intrinsic optical-phonon bandwidth. Under these circumstances phonons do not become localized on individual atoms but remain spread out over a large number of unit cells, leading to a close to perfect linear averaging of the various isotope masses.

The Raman linewidth has been shown to be due to the phonon lifetime for the anharmonic decay into two phonons. In good approximation one writes

$$\Delta\omega_0 \cdot \tau \simeq 1. \tag{2}$$

Time-resolved measurements have yielded lifetimes of 6 ps, in very good agreement with the observed $\Delta\omega_0 = 1\,\text{cm}^{-1}$ (equivalent to $3 \times 10^{10}\,\text{Hz}$) (Fuchs *et al.*, 1991a). This natural linewidth (or lifetime) is dominated by anharmonic interactions leading to the decay of an optical phonon at $\mathbf{k} = 0$ into two phonons. At the low temperature where these Raman experiments were conducted, the rate for anharmonic decay (τ^{-1}) is proportional to the amplitudes u of the two phonons created in the process:

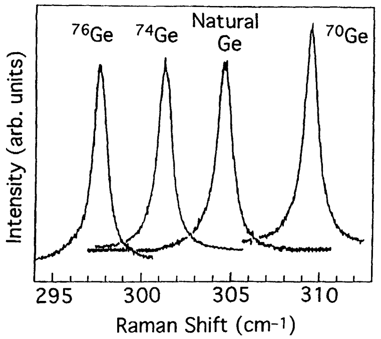

FIG. 3. Raman spectra of the optical zone-center phonons of several Ge isotopes and natural Ge at 90 K (courtesy of Fuchs *et al.*, 1992).

$$\tau^{-1} = \Delta\omega_0$$

$$\propto \langle u_1^2\rangle\langle u_2^2\rangle = \hbar^2/4M^2\omega_1\omega_2. \qquad (3)$$

Using Eq. (1), $\omega \propto M^{-1/2}$, we find

$$\Delta\omega_0 \propto M^{-1}. \qquad (4)$$

The widths of the Raman lines obtained from the four isotopically enriched Ge crystals indeed follow such a simple dependence (Fig. 4). The Raman linewidths for the crystals of natural composition and of maximum isotope disorder are slightly larger. Cardona *et al.* (1993) have shown that this extra broadening requires an anharmonic effect on the density of phonon states for three-dimensional crystals. They further showed that on going to two dimensions, isotope disorder broadening can be explained in the harmonic approximation.

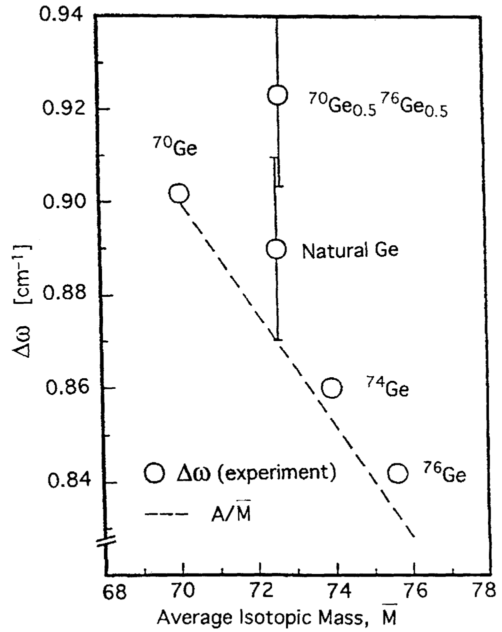

FIG. 4. Full width at half maximum of the Raman lines shown in Fig. 3 and of a ^{70}Ge$_{0.5}$/^{76}Ge$_{0.5}$ sample. The dashed line represents the M^{-1} dependence for isotopically pure samples. (Courtesy of Fuchs *et al.*, 1992.)

Photoluminescence studies of several isotopically enriched Ge crystals at low temperatures were reported by Davies *et al.* (1992, 1993, 1994) and by Etchegoin *et al.* (1992). Figure 5 shows representative luminescence spectra of highly enriched ^{70}Ge and ^{74}Ge crystals. The LA- and the TO-phonon assisted free-exciton (FE) lines are clearly shifted in energy. Excitons bound to deliberately introduced substitutional Cu acceptors generate both phonon-assisted (TA, LA, TO) and no-phonon (NP) lines. In addition one observes a sharp no-phonon line due to P donors. Whereas the phonon-assisted lines suffer energy shifts due to phonon and band-gap energy changes, the NP lines are displaced solely because of band-gap energy changes. This information allows the quantitative determination of isotope-related phonon and band-gap energy changes. After correction for the band-gap shifts, the phonon energies accurately follow the frequency dependence given in Eq. (1).

Stimulated in part by speculations regarding very large increases of the thermal conductivity in isotopically pure diamonds, a series of diamond crystal growth and characterization studies with extensive theoretical interpretation has been conducted. Using the well established multiple-scattering formalism of the coherent-potential approximation (CPA), Hass *et al.* (1991, 1992) were able to explain quantitatively the large effects of isotopic disorder on the frequencies of optical phonons at $\mathbf{k} = 0$ and for the Raman linewidths for a range of isotopic compositions. Figure 6 shows the Raman frequencies and linewidths as functions of the normalized ^{13}C concentration. Isotopic disorder causes the upwards bowing of the Raman frequency dependence as well as the large asymmetric increase in linewidth with

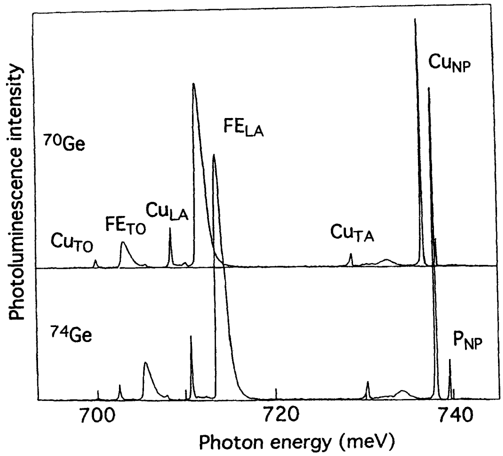

FIG. 5. Luminescence from the decay of free excitons and excitons bound to Cu acceptors and P donors in ^{70}Ge and ^{74}Ge at 4.2 K. The subscripts indicate the *L*-point phonons assisting in the exciton decay. The impurity-bound excitons can decay without emission of a phonon (NP). (Courtesy of Davies *et al.*, 1992.)

isotopic composition. Both dependences are well described with CPA theory, especially when the intrinsic linewidth of the isotopically purest samples is subtracted from all the data points.

It is interesting to note that the success in describing optical-phonon frequencies and Raman linewidths of isotopically mixed diamonds does not help in the understanding of the very strong dependence of the thermal conductivity (which depends on acoustic phonons) on small levels of isotopic disorder.

The thermal conductivity of dielectrics is given by

$$K = \tfrac{1}{3}CvL, \tag{5}$$

with C = specific heat, v = phonon velocity, and L = phonon mean free path (Ziman, 1963). Neither the velocity of low-frequency phonons nor the specific heat, which has been determined by Morelli *et al.* (1991), are predicted to change much at a 1% isotope impurity level. The large thermal-conductivity increases that are observed for diamonds with increasing isotopic purity must be due to changes in the *phonon scattering rates*. This suggestion goes again back to Pomeranchuk (1942).

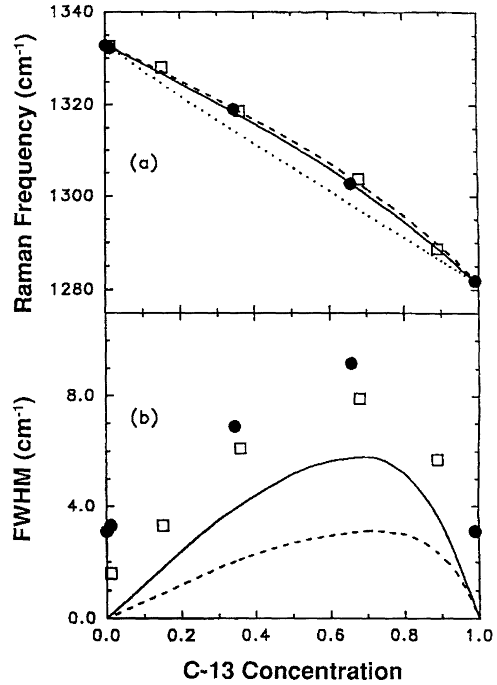

FIG. 6. Isotopic composition dependence of the diamond (a) Raman frequency and (b) linewidth (FWHM). CPA results for pure and adjusted "TPZ" models and VCA results in (a) shown as dashed, solid, and dotted lines, respectively. Open squares courtesy of Chrenko (1988); filled dots courtesy of Hass *et al.* (1992).

Wei *et al.* (1993) have measured a record thermal conductivity of 410 W cm^{-1} K^{-1} at 104 K for a 99.9% ^{12}C-enriched diamond (Fig. 7). They fitted their data with the Callaway model (Callaway, 1959), using four independent scattering processes. The functional form of the individual scattering rates were chosen as follows:

$$\tau^{-1}_{\text{normal}} = AvT^3\lambda^{-1}, \tag{6}$$

$$\tau^{-1}_{\text{Umklapp}} = BvT\lambda^{-1}\exp(-C/T), \tag{7}$$

$$\tau^{-1}_{\text{boundary}} = v/D, \tag{8}$$

$$\tau^{-1}_{\text{isotope}} = 4\pi^3 v\lambda^{-4}V_0 x(1-x)(12+x)^{-2}, \tag{9}$$

with v and λ corresponding to the average phonon velocity and wavelength, respectively;

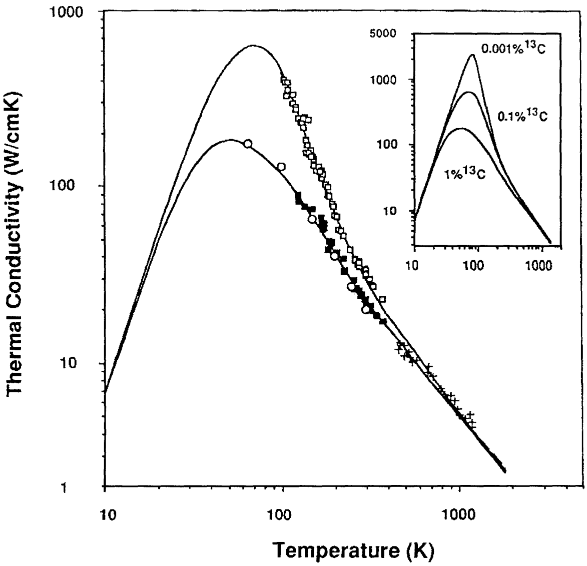

FIG. 7. Thermal conductivity of natural-abundance (1.1% ^{13}C) diamond (lower squares), isotopically enriched (0.1% ^{13}C) diamond (upper squares), together with the low-temperature data of Slack (1973) (circles) and the high-temperature data of Olson *et al.* (1993) (crosses). The solid curves are the result of fitting the Callaway theory (Callaway, 1959) to the data, using the same set of fitting parameters. Inset: The calculated thermal conductivity corresponding to 1%, 0.1%, and 0.001% ^{13}C concentrations according to the Callaway theory. (Adapted from Wei *et al.*, 1993.)

x is the isotope fraction of ^{13}C, V_0 is the atomic volume, and T is the absolute temperature. The constants A, B, C, and D were determined from the fits: $A = 7.2 \times 10^{-11}\,\text{cm}\,\text{K}^{-3}$, $B = 1.5 \times 10^{-12}\,\text{cm}\,\text{K}^{-1}$, $C = 670\,\text{K}$, and $D = 0.33\,\text{cm}$. Once the constants A through D have been determined it is possible to predict the dependence of thermal conductivity on temperature and isotopic composition. This has led to their prediction of a thermal conductivity as great as $2000\,\text{W}\,\text{cm}^{-1}\,\text{K}^{-1}$ near 80 K, exceeding Geballe's prediction (Geballe, 1990) that the thermal conductivity of pure diamond may surpass $1000\,\text{W}\,\text{cm}^{-1}\,\text{K}^{-1}$ around liquid nitrogen temperature (77 K). Similar results have very recently been reported by Olson *et al.* (1993). It should be pointed out that measurements of such extreme thermal conductivities for millimeter-size samples are extremely difficult. The samples act like close to perfect thermal short circuits.

Thermal conductivity studies have also been performed on very highly enriched, ultrapure ^{70}Ge crystals (Ozhogin *et al.*, 1996). In these 99.99% enriched single crystals the thermal conductivity exceeds the maximum values recorded by Geballe and Hull (1958) by a factor of $\gtrsim 3$. Further detailed measurements have been conducted, and an in-depth analysis has been reported recently (Asen-Palmer *et al.*, 1997).

Increased thermal conductivity in isotopically enriched semiconductors could become interesting for Si integrated circuits, for which the thermal budget is always pushed to the limit. At this point in time it is difficult to estimate if the cost increases related to increased isotopic purity can be justified by the gains in thermal conductivity.

2.1.2 Lattice Constants and Bulk Modulus

The lattice constants a_0 of large, chemically pure and crystallographically highly perfect crystals can be measured with very high precision. Because of this "naturally" occurring precision there exists an interest in using the lattice constant of a suitable crystal as a length standard (Becker, 1995). However, as a consequence of the zero-point motion the atomic volumes change with isotopic composition. The resulting effect on a_0 has been carefully studied for an enriched ^{74}Ge crystal (Buschert *et al.*, 1988) and for a series of diamonds with varying ^{12}C/^{13}C composition (Holloway *et al.*, 1991).

The theoretical treatment of the dependence of the lattice constant on isotopic composition is straightforward. Starting with the equation of state

$$P + \left[\frac{\partial F}{\partial V}\right]_T = 0, \tag{10}$$

we evaluate the Helmholtz free energy F in regard to volume (V), temperature (T) and atomic mass (M):

$$F = U_0(V) + \sum_i \left\{ \frac{1}{2}\hbar\omega_i + kT \ln\left[1 - \exp\left(\frac{-\hbar\omega_i}{kT}\right)\right]\right\}. \tag{11}$$

Here $U_0(V)$ is the binding energy of the a crystal with infinite-mass atoms, ω_i are the frequencies of the phonon modes, and k is Boltzmann's constant. The atomic mass is reflected by the inverse square-root dependence of ω on M.

For a small range of isotope masses ΔM, the equation of state yields

$$\frac{\Delta V}{V} = -\Delta M \frac{\partial^2 F/\partial V \partial M}{V \partial^2 F/\partial V^2} = -C\Delta M \frac{\partial^2 F}{\partial V \partial M}, \tag{12}$$

with C the isothermal compressibility. The above expression can now be further modified by inserting the expression for F [Eq. (11)], giving the following equation for the relative changes in the lattice constant a_0 with isotopic mass at low temperatures ($\exp[-\hbar\omega/kT] \ll 1$):

$$\frac{\Delta a}{a} = -\frac{C}{a^3}\frac{\Delta M}{M}\left(\gamma_0 \hbar\omega_0 + \frac{3}{4}\gamma_a k\theta\right), \tag{13}$$

where $\gamma_0 = 1.12$ and $\gamma_a = 0.40$ are the Grüneisen parameters for optical and acoustical phonon modes in Ge, $\theta = 374\,\mathrm{K}$ is the Debye temperature, and the optical phonon energy $\hbar\omega_0 = 37.3\,\mathrm{meV}$. Buschert *et al.* (1988) evaluated the above equation for a 95.8% enriched ^{74}Ge crystal [the same crystal that was measured by Geballe and Hull (1958)] in comparison with a natural Ge crystal. Theory, Eq. (13), predicts a 12-ppm and a 6-ppm reduction in a_0 for $T = 0\,\mathrm{K}$ and $T = 300\,\mathrm{K}$, respectively. Buschert *et al.* (1988) found experimentally reductions by 14.9 ppm and 6.3 ppm at $T = 77\,\mathrm{K}$ and at $T = 300\,\mathrm{K}$, respectively. Though this agreement between theory and experiment is considered adequate, better experimental results may be obtained today with more highly enriched crystals of at least four isotopes of Ge.

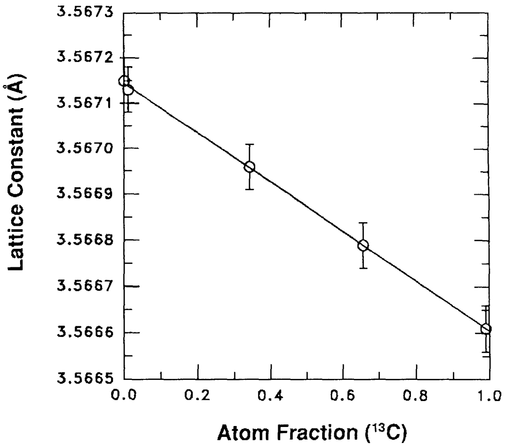

FIG. 8. Isotopic dependence of the lattice constant of diamond at 25 °C. The line is a least-squares fit to the data. (Courtesy of Holloway *et al.*, 1991.)

Similar lattice-parameter determinations have been conducted for several isotopically mixed diamonds. Figure 8 shows the experimental data obtained by Holloway *et al.* (1991). The data can be fitted by

$$a_0(T = 25\,^\circ\mathrm{C}, x) = [3.56714 - 0.00054x]\ \text{Å} \tag{14}$$

Theoretical estimates of the dependence of a_0 on isotopic composition by Collins *et al.* (1988, 1990) based on the London theory of the variation of crystal volume with isotopic substitution (Berezin and Ibrahim, 1988) are larger by a factor 1.4–3.2 than the experimental findings and have been attributed to the oversimplified assumptions made in the theory. It is unlikely that the crystalline quality of the diamonds is the cause for the discrepancy, since all the crystals exhibited x-ray rocking curves with full width at half maximum between 10 and 20 arc sec, indicating a very high degree of perfection.

The decrease in lattice constant of diamonds with increasing ^{13}C content has led to the speculation that ^{13}C diamonds may be harder than their natural counterparts (Banholzer and Anthony, 1992). In a recent study of the Brillouin and Raman scattering in natural and isotopically controlled diamond, Vogelgesang *et al.* (1996) explored the effects of zero-point motion and the anharmonicity of the lattice vibrations. On the basis of a valence-force model explicitly incorporating the isotopic composition, they predicted the isotopic mass dependence of the elastic moduli deduced from Brillouin spectra, of the zone center optical mode frequency, and of the lattice constant. On the basis of the anharmonicity parameter, deduced from a comparison of the theory with the experimentally reported dependence of the lattice constant discussed above, they predicted that the bulk modulus of ^{13}C diamond should exceed that for ^{12}C diamond by one part in 10^3, just below the sensitivity of their high-precision Brillouin-scattering experiment. The predicted increase in the bulk modulus of ^{13}C diamond, if confirmed, would imply that it is the hardest terrestrial material [Ramdas *et al.*, (1993)]. Such diamonds would also represent solids with the highest atomic density.

2.1.3 Band Gaps The width and the character (direct or indirect) of the lowest-energy band gap is of paramount importance for a large number of semiconductor properties and in turn for all semiconductor devices. Because of this great significance there exists a strong interest in all effects that influence the band structure.

The isotopic composition affects the band gaps through the electron–phonon coupling and through the change of volume with isotopic mass. Several groups have conducted low-temperature studies of the direct and indirect band gaps of natural and isotopically controlled Ge single crystals. Agekyan *et al.* (1989) used photoluminescence, infrared absorption, and Raman spectroscopy with a Ge crystal of natural composition and a crystal with 85% ^{76}Ge and 15% ^{74}Ge. They found an indirect–band-gap change of $\Delta E_G = 0.09$ meV and a direct–band-gap change of $\Delta E_G = 1.25$ meV with an error of ±0.05 meV. Etchegoin *et al.* (1992) and Davies *et al.* (1992, 1994) reported PL studies of natural and several highly enriched, high-quality single crystals of Ge. Measurement of the energies of impurity-bound excitons shown in Fig. 5 permits the direct determination of band-gap shifts with the crystal isotope mass because the radiative recombination does not require phonon participation. Figure 9 shows the no-phonon energies of excitons E_{NP} bound to P and Cu in several isotopically controlled crystals. As may be expected from the very large Bohr orbits of the excitons, their binding energy only depends on the average isotope mass and not on the isotopic disorder. The rate of band-gap energy change, which is equal to the exciton binding-energy change, with isotope mass as determined by Davies *et al.* (1992, 1994) is

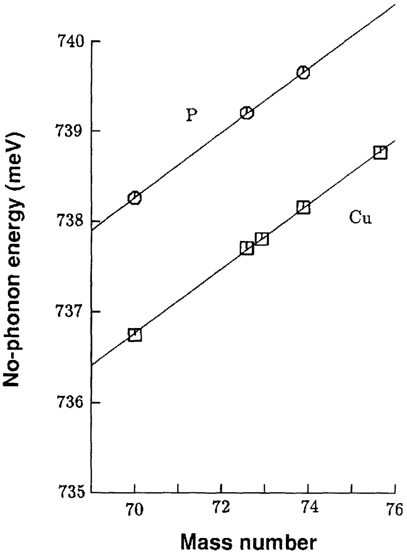

FIG. 9. Energies of the no-phonon lines of excitons bound to Cu acceptors (squares) and P donors (circles) (courtesy of Davies *et al.*, 1992; Davies, 1994).

$$\frac{dE_{IG}}{dA} = \frac{dE_{NP}}{dA} = 0.35 \pm 0.02\,\mathrm{meV\,u^{-1}}. \tag{15}$$

Etchegoin *et al.* (1992) obtained a very similar value.

The contribution to the band-gap shift originating in the volume change can be estimated using the results of Buschert *et al.* (1988) for lattice-constant change with isotope mass and the published dependence of E_{IG} with volume (Schmid *et al.*, 1990). One finds

$$\left.\frac{dE_{IG}}{dA}\right|_{vol} = 0.132\,\mathrm{meV\,u^{-1}}. \tag{16}$$

This is the smaller contribution to the experimentally determined energy-gap change with isotope mass. It is in reasonable agreement with the earlier estimates of Agekyan *et al.* (1989).

The main contribution to $\mathrm{d}E/\mathrm{d}A$ can be directly related to the change of the energy gap with temperature. This change is described by structure factors that contain electron–phonon interaction terms (Debye–Waller factors) and self-energy terms. For practical calculations these terms are expanded in a power series of the atomic displacements. The leading terms are proportional to the mean square displacements $\langle u^2\rangle$ of each atom. Describing the lattice atoms in terms of harmonic oscillators, we find

$$\langle u^2\rangle = \hbar\left(\tfrac{1}{2}+n\right)/M\omega. \tag{17}$$

With increasing temperature both n and $\langle u^2\rangle$ increase, leading to the observed reduction of the energy gap. At low temperature, $n = 0$ and we only deal with the zero-point oscillation. Combining the dependence of ω on M with the above equation we find

$$\langle u^2\rangle \propto 1/\sqrt{M}. \tag{18}$$

Zollner *et al.* (1992) have performed a numerical calculation of the electronic bands using an empirical pseudopotential method including the necessary lattice dynamics. They find for Ge $(\mathrm{d}E_{\mathrm{IG}}/\mathrm{d}A)_{\text{e-p}} = 0.41\,\mathrm{meV}$. The total calculated shift of the indirect band-gap energy with isotope mass adds up to $(\mathrm{d}E_{\mathrm{IG}}/\mathrm{d}A)_{\mathrm{total}} = 0.48\,\mathrm{meV}$. This result compares favorably with the experimental values stated above by Davies *et al.* (1992, 1994), and by Etchegoin *et al.* (1992) who reported $(\mathrm{d}E_{\mathrm{IG}}/\mathrm{d}A)_{\mathrm{total}} = 0.37 \pm 0.01\,\mathrm{meV\,u^{-1}}$.

Measurements of the direct band gap at the Γ point ($\mathbf{k} = 0$) in the Brillouin zone have also been performed. Though the direct band gap is technologically less important than the minimum indirect band gap, determining the dependence of this gap on isotope mass is of the same fundamental significance as the indirect-band-gap studies. Davies *et al.* (1993) used low-temperature optical-absorption measurements on very thin samples of Ge single crystals with natural composition and of three different, highly enriched isotopes. They found

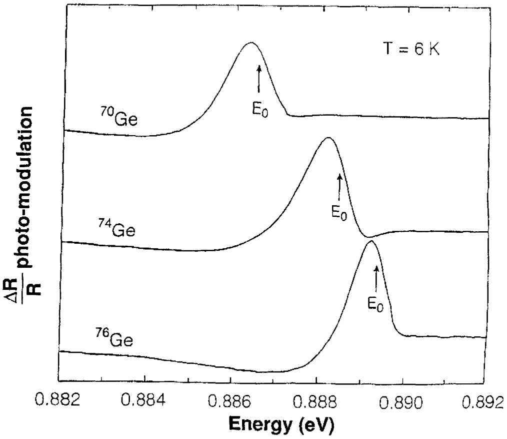

FIG. 10. Photomodulated reflectivity showing the direct band gap of single crystals of ^{70}Ge, ^{74}Ge, and ^{76}Ge at $T = 6\,\mathrm{K}$ (courtesy of Parks *et al.*, 1994).

$$\frac{\mathrm{d}E}{\mathrm{d}A} = 0.49 \pm 0.03\,\mathrm{meV\,u^{-1}} \tag{19}$$

for the temperature extrapolated to zero. Parks *et al.* (1994) have used piezo- and photomodulated reflectivity spectra of four monoisotopic and one natural Ge crystals. These techniques do not require the extreme sample thinning that is necessary for optical-absorption measurements, and the derivative nature of the spectra emphasizes the small changes. Figure 10 shows the photomodulated reflectivity of three monoisotopic Ge crystals at $T = 6\,\mathrm{K}$. The position of the direct energy gap M is indicated by an arrow marked E_0. The excellent signal-to-noise ratio and the superb spectral resolution allow a very accurate determination of the dependence of E_{DG} on isotopic mass. At very low temperatures an inverse square-root dependence accurately describes band-gap dependence:

$$E_{\mathrm{DG}} = E_{\mathrm{DG}}^{\infty} + C/\sqrt{M}. \tag{20}$$

A fit through five data points yields $E_{\mathrm{DG}}^{\infty} = 959\,\mathrm{meV}$ and $C = -606\,\mathrm{meV\,u^{1/2}}$. Written as a linear dependence for the small range of isotopic masses, Parks *et al.* (1994) find

$dE_{DG}/dA = 0.49\,\mathrm{meV\,u^{-1}}$, in perfect agreement with the results of Davies *et al.* (1993). Parks *et al.* (1994) also determined the isotope mass dependence of the sum of the direct gap and the split-off valence band (Δ_0) and found $d(E_{DG} + \Delta_0)/dA = 0.74\,\mathrm{meV\,u^{-1}}$.

The experimental results can be compared with Zollner *et al.*'s calculations (1992), which are found to be of the correct order of magnitude. The theoretical estimates for the contribution of the linear isotope shift of the minimum indirect gaps, which are caused by electron–phonon interaction, are too large by a factor of ≈1.7, and for the smallest direct gap they are too large by a factor ≈3.2.

We conclude this discussion of the bandgap shifts with isotope mass by observing that the effects are exceedingly small and most likely will not have any consequences for Ge nor any other semiconductors for technological applications. However, this exercise is an excellent demonstration of the advanced state of our quantitative theoretical understanding of the subtle effects of temperature, pressure, and isotope mass on the electronic band structure.

2.2 Isotope Superlattices

The discussion of phonons, their dispersion, and the dependence of their properties on the isotopic composition of the host crystal has so far been restricted to bulk crystals, crystals sufficiently large so as not to affect phonon properties. What do we expect when the dimensions of the crystal are significantly reduced? This is the focus of the discussion of the fascinating features shown by short-period isotope superlattices consisting of layers each made up of a few atomic planes of ^{A}Ge and ^{B}Ge. These are ideal structures for phonon studies because the electronic band structure is not significantly affected by the isotope superlattice, *i.e.*, the material looks bulklike for electrons throughout the crystal layers. In superlattices with very significant periodic band-structure discontinuities and large changes in host atom masses (*e.g.*, AlGaAs/GaAs), it is impossible to perform phonon studies using Raman spectroscopy.

The first problem to be addressed relates to the normal phonon modes of a short-period isotope superlattice. The vibrational modes and the phonon dispersion relation for a $^{70}\mathrm{Ge}_4/^{76}\mathrm{Ge}_4$ [001] superlattice at $\mathbf{k} = 0$ were calculated using the planar force-constant model (Fuchs *et al.*, 1993). These theoretical results were tested with Raman studies of a series of MBE-grown, short-period isotope superlattices of ^{70}Ge and ^{74}Ge (Spitzer *et al.*, 1994). The experimental and theoretical results are summarized in Fig. 11. The shortest periods studied consisted of two atomic layers each and produce a single Raman line corresponding to the zone-center optical phonon in a Ge crystal with isotope mass $A = 72$, the linear average of the two masses. The experts who grew this superlattice are quite certain that this result is not due to intermixing of more than one atomic layer during growth because they used very low growth temperatures ($T = 350\,°\mathrm{C}$) and low growth rates ($4\,\mathrm{Å\,min^{-1}}$). The $n = 4$ superlattice shows the Raman line predicted at $\approx 290\,\mathrm{cm^{-1}}$, though at a slightly lower intensity. In the $n = 12$ superlattice we observe at least three strong Raman lines. For $n = 32$ the Raman spectra begin to approach the structure we would expect to observe for two isotopically pure bulk crystals, one made up of ^{70}Ge, the other of ^{74}Ge.

These first Raman spectroscopy studies of isotope superlattices demonstrate Brillouin-zone folding in a very pure form. The relative intensities of the Raman lines reflect the densities of states at $\mathbf{k} = 0$. It is quite impressive how well the experimental data are reproduced by the planar bond-charge and bond-polarizability models. For layers thicker than approximately 30 atomic planes we observe close to perfect localization in the individual layers, *i.e.*, bulklike behavior.

3. LOCAL VIBRATIONAL-MODE SPECTROSCOPY

Local vibrational-mode (LVM) spectroscopy is a well established research tool for the study of impurities and impurity complexes in semiconductors (Barker and Sievers, 1975; Newman, 1973; Spitzer, 1971; Haller, 1995). A very large body of information on static and dynamic structural information has been ac-

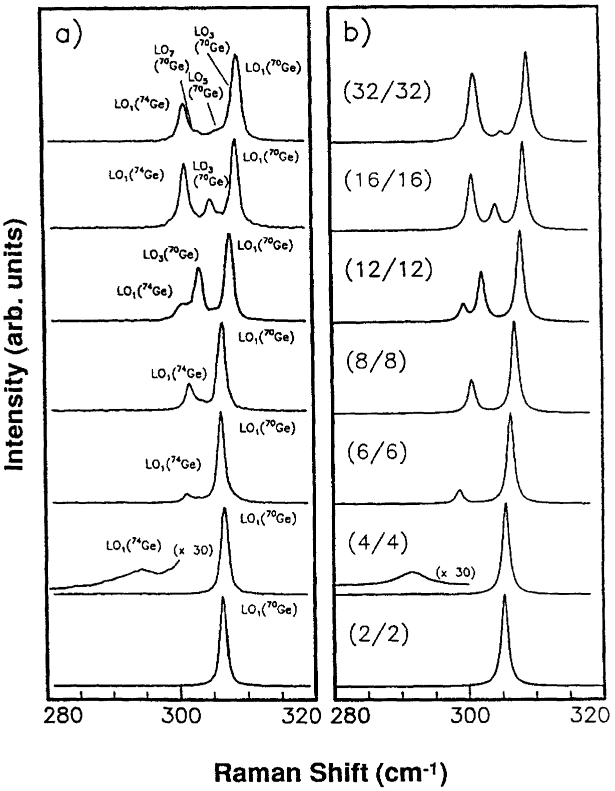

FIG. 11. (a) Measured at $T = 10\,\text{K}$ and (b) calculated Raman spectra for a series of isotope superlattices $^{70}\text{Ge}_n/^{74}\text{Ge}_n$ grown along [001] with $n = 2, 4, 6, 8, 12, 16$ and 32. A number of modes confined to the ^{70}Ge and ^{74}Ge layers are evident. (Courtesy of Spitzer *et al.*, 1994.)

cumulated using LVM spectroscopy. In general LVM studies are performed on impurities that have a mass lower than the average mass of the host-lattice atoms. Depending on the details of the lattice location of the impurity, particularly on symmetry and on how many bonds are formed with host-lattice atoms, one observes a number of normal vibrational modes with well defined frequencies above the vibrational spectrum of the host lattice. In order to obtain a crude understanding of an LVM spectrum, it usually suffices to model the impurity and the nearest bound neighbors as a quasimolecule imbedded in the host crystal. The number of normal modes of this molecule, their frequencies, and even their intensity ratios can be extracted with limited accuracy from such a simple approach. The model can be improved substantially by taking into account next-nearest neighbor atoms in the form of an additional mass correction, the "host interaction mass" and by extending the linear force dependence on impurity displacement with higher terms (anharmonicity).

Because of the simple dependence of vibrational frequencies on the inverse square root of the vibrating mass,

$$\nu_1/\nu_2 = \sqrt{M_2/M_1}, \tag{21}$$

isotope substitutions have already been used since the earliest studies to obtain unambiguous proof for the association of an LVM feature with a certain impurity. The largest isotope shift caused by such a substitution of stable isotopes is close to 40% and is obtained with hydrogen and deuterium. Extensive use of this particular isotope substitution has been made in recent years by a large number of investigators studying the very many properties exhibited by hydrogen in crystalline semiconductors (Haller, 1978; Stavola *et al.*, 1988; Watkins *et al.*, 1990; Denteneer *et al.*, 1989; McMurray *et al.*, 1987; Hoffman *et al.*, 1992; McCluskey *et al.*, 1995, 1996; Götz *et al.*, 1996).

Hydrogen-related studies are not the only ones that have profited from LVM spectroscopy and isotopic substitution. A large number of impurities in many semiconductors have been investigated. Whenever available, isotopes have been used to help identify LVM lines. Newman (1993) has written a comprehensive review of LVM studies of defects in III–V compounds.

What then can isotope engineering contribute to this already highly successful characterization technique? One group of studies that can profit concerns LVM spectroscopy of impurities and impurity complexes in semiconductors that are naturally composed of several stable isotopes with some of these present at significant atomic fractions. LVM spectra of impurities in such crystals are usually composed of many closely spaced or overlapping lines because the impurity and its host lattice neighbors form quasimolecules with a number of different reduced masses (Sciacca *et al.*, 1996). Isotope enrichment of the host material can lead to drastic simplification of the LVM spectra in such cases.

In order to give an illustration of the effect of isotope enrichment on spectral complexity we want to review briefly the LVM structure of O in natural and isotopically controlled Ge. To model the LVM of this bond-centered impurity one takes into account the two nearest neighbors, while the rest of the lattice appears as a small correction, called interaction mass. The ^{x}Ge–O–^{y}Ge molecule has three fundamental modes. The "wag" mode (ν_3) has the highest frequency and was observed in oxygen-rich Ge quite some time ago (Kaiser *et al.*, 1956). High-resolution studies show 11 lines, each one split into several components whose intensities depend on temperature. The 11 lines result from the different mass combinations of ^{x}Ge and ^{y}Ge. Combinations with identical linear mass averages (*e.g.*, ^{72}Ge/^{74}Ge and ^{73}Ge/^{73}Ge) lead to almost the same oxygen vibrational frequencies.

The complexity of such spectra can be reduced very significantly through the use of isotopically pure (or at least highly enriched) crystals. Figure 12 shows two spectra that were reported by Mayur *et al.* (1994). They used one natural and four highly enriched Ge crystals doped with oxygen. The reduction from 11 to one LVM line is significant because it allows a quantitative study of the linewidth and of the coupling between ν_3 and the other modes as a function of temperature. The splitting is a result of the nonlinear superposition of the ν_2 and the ν_3 modes. The low-frequency ν_2 modes can already be thermally populated near to the temperature of liquid helium.

The vibrational modes of carbon impurities on As sites in GaAs represent a similar physical system. C_{As} vibrations are modeled with an X_4Y molecule with two different masses for X. Experimentally one has resolved five very closely spaced lines (Theis and Bajaj, 1982; Theis *et al.*, 1983) of the nine lines predicted by theory (Leigh and Newman, 1982). Even though the experimental findings are in good agreement with theory, it would be instructive to see what happens when only one Ga isotope is present in an isotopically pure GaAs crystal. One extremely sharp line of a triply degenerate mode should be observed.

4. DIFFUSION

One of the most fundamental processes characterizing all matter is the random motion of its atomic constituents. In semiconductors much has been learned in recent years about the motion of host and impurity atoms as well as native defects such as vacancies and interstitials. A number of comprehensive review papers on this subject have been written (Gösele, 1989, 1994; Frank *et al.*, 1984; Tan *et al.*, 1991; Hu, 1992; Chen *et al.*, 1994; Shaw, 1988; Palfrey, 1989). When a concentration gradient $\mathrm{d}N/\mathrm{d}x$ is introduced, random motion leads to a net flux of matter J, which is proportional to the concentration gradient:

$$J = -D\frac{\mathrm{d}N}{\mathrm{d}x}. \tag{22}$$

The diffusion coefficient D, in many (though not in all) cases, can be described by a thermally activated constant

$$D = D_0 \exp(-E/kT). \tag{23}$$

Impurity diffusion in semiconductors plays a

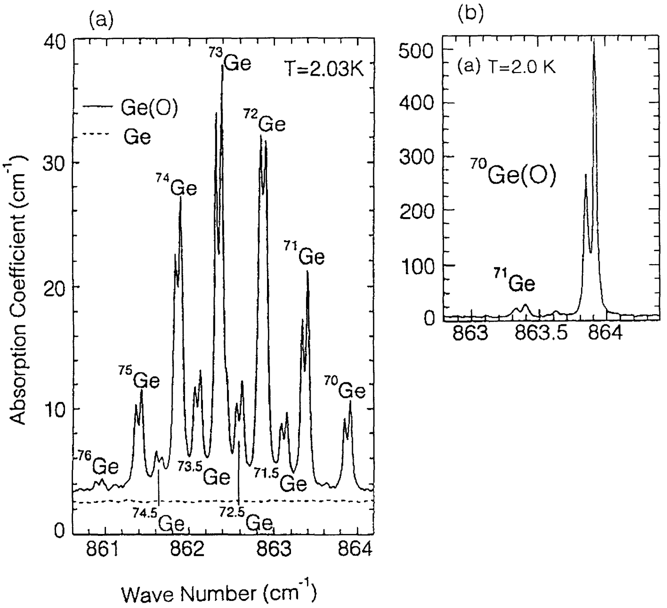

FIG. 12. (a) Spectrum of the Ge_2O quasimolecules in natural Ge at 2.05 K. Eleven lines split by $\nu_2 + \nu_3$ coupling excitations correspond to the distinct isotopic mass combinations of the two Ge atoms in the quasimolecule. An oxygen-free sample produced the dashed line. (b) The spectrum of an oxygen-doped, highly enriched sample of ^{70}Ge. The lines labeled ^{71}Ge are caused by traces of ^{72}Ge in ^{70}Ge. (Courtesy of Mayur *et al.*, 1994.)

key role in the fabrication of electronic devices. By way of example, diffusion can be utilized as a desirable process enabling the introduction of dopant impurities into areas defined by a mask on a semiconductor wafer. Diffusion can act also as a detrimental process, broadening narrow impurity implantation profiles or admitting rapidly diffusing, undesirable impurities. There exists a very extensive literature on diffusion studies for most semiconductors. The field is extremely active, especially for the semiconductors which have become important recently for high-temperature electronics (*e.g.*, SiC) and light-emitting devices working in the green and blue range of the visible spectrum (*e.g.*, ZnSe, GaN).

An impressive body of experimental and theoretical work on Si has accumulated over the past 30 years. A major advance was made in the late 1960s and early 1970s when electron-microscopy observations of A-type swirl defects in dislocation-free Si confirmed the theoretical prediction that Si interstitials are present in addition to vacancies at high temperatures. It took many years before it became generally accepted that both interstitials and vacancies were the native defects assisting self- and impurity diffusion in Si. A series of metal diffusion experiments, which started with Au and were later extended to Pt and most recently to Zn (Bracht *et al.*, 1995), eventually convinced the most critical experts of the coexistence of interstitials and vacancies in Si at high temperatures. Au is especially suitable to explore the role of Si interstitials because it diffuses via the so-called "kick-out" mechanism. Interstitials arriving at Au on substitutional sites (Au_s) exchange their place and form a highly mobile Au interstitial species (Au_i). The process reverses at a new location returning a Au_s. Au_s concentration profiles after diffusion from a surface source into a Si wafer exhibit a shape that is very characteristic for the kick-out process and allow the determination of the equilibrium concentrations of interstitials (C_I^{eq}) and vacancies (C_V^{eq}) as well as the individual equilibrium diffusion coefficients (D_I^{eq}, D_V^{eq}) at the diffusion temperature.

Despite progress in recent years, significant errors remain, and the large disparities between the values reported by numerous authors (Gösele, 1989; 1994; Frank *et al.*, 1984; Tan and Gösele, 1985; Willoughby, 1994; Zimmermann and Ryssel, 1992) illustrate the limited state of our knowledge of self-diffusion in Si. Diffusion in compound semiconductors and their alloys as well as impurity diffusion adds considerable further complexities to the problem leading to even greater uncertainties.

Isotope control is expected to be able to contribute with new experimental approaches to the study of diffusion and hopefully can help in improving our understanding of the many interacting factors. Tan *et al.* (1992) were the first to make use of a GaAs isotope superlattice to study Ga self-diffusion. Their basic idea was to grow a superlattice consisting of approximately 1000-Å thick, undoped layers of ^{69}GaAs and ^{71}GaAs. Only ^{69}Ga/^{71}Ga interdiffusion can be studied because As is monoisotopic. Because the isotopically pure layers were grown on a Si-doped substrate, the Ga self-diffusivity may have been affected and may not reflect the intrinsic Ga self-diffusivity in GaAs. Wang *et al.* (1996) have designed and executed a new experiment on Ga self-diffusion in GaAs using undoped, semi-insulating substrates with undoped 2000 Å isotopically enriched ^{69}GaAs and ^{71}GaAs layers. Instead of capping layers, which can introduce strain and impurities, individual samples were annealed in ampoules with very tight As ambient control. This allowed the determination of the Ga diffusion constant over six orders of magnitude. All the data points lie on one straight line in the Arrhenius plot shown in Fig. 13. A best fit leads to

$$D = (43 \pm 25)\exp[(-4.24 \pm 0.06\,\mathrm{eV})/k_{\mathrm{B}}T], \tag{24}$$

using for the pre-exponential factor the expression

$$\mathrm{D}_0 = ga^2\nu_0\exp(S^{\mathrm{SD}}/k_{\mathrm{B}}) \tag{25}$$

where g is a geometric factor, a is the lattice constant, ν_0 is the highest vibrational mode frequency, S^{SD} is the self-diffusion entropy, and k_{B} is Boltzmann's constant.

Assuming $g \simeq 1$, Wang *et al.* (1996) found a Ga self-diffusion entropy value $S^{\mathrm{SD}} = 8k_{\mathrm{B}}$. Recent molecular-dynamics calculations confirm this value as well as the activation energy to within a few percent (Bockstedte and Scheffler, 1996). This exceptionally good agreement may be a sign that diffusion experiments and theory begin to measure and describe the proper diffusion physics, at least for these well defined conditions.

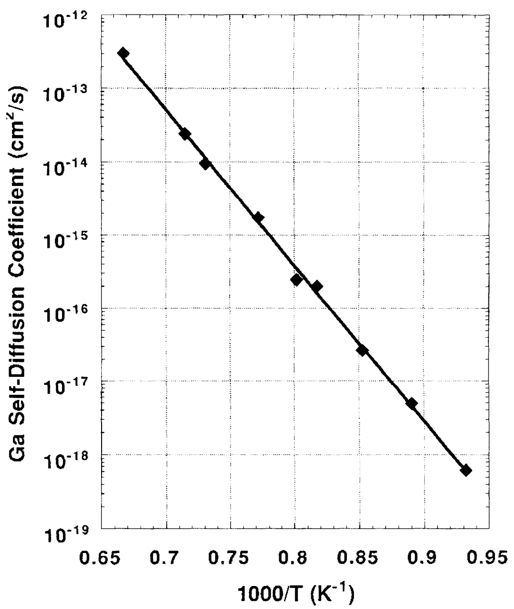

FIG. 13. Arrhenius plot of the Ga self-diffusion coefficient in GaAs determined with ^{69}Ga/^{71}Ga isotope interdiffusion (adapted from Wang *et al.*, 1996).

Very recently the same type of experiment has been performed for GaP (Wang *et al.*, 1977). The activation energy for Ga self-diffusion in this semiconductor is very similar to that in GaAs but the entropy is only $4k_{\mathrm{B}}$. These results await a detailed theoretical evaluation.

In order to test the new approach to self-diffusion studies based on isotopically controlled multilayer structures with a well understood system, the Ge self-diffusion coefficient has been determined with bilayers of ^{70}Ge and ^{74}Ge (each 1000 or 2000 Å thick) which were grown by MBE on a natural Ge substrate (Fuchs *et al.*, 1995). Disregarding the small differences in diffusivity caused by different isotope masses, (Campbell, 1975), we expect the Ge isotopes to diffuse symmetrically into each other following complementary error functions. Figure 14 shows SIMS results for the as-grown sample and for the sample diffused at 586 °C for 55.55 h. Because the isotopes are only enriched into the high 90 percentile range, we obtain SIMS data from some of the residual minor Ge isotopes. This redundancy in data is very useful in the deconvolution of the SIMS instrument function and in

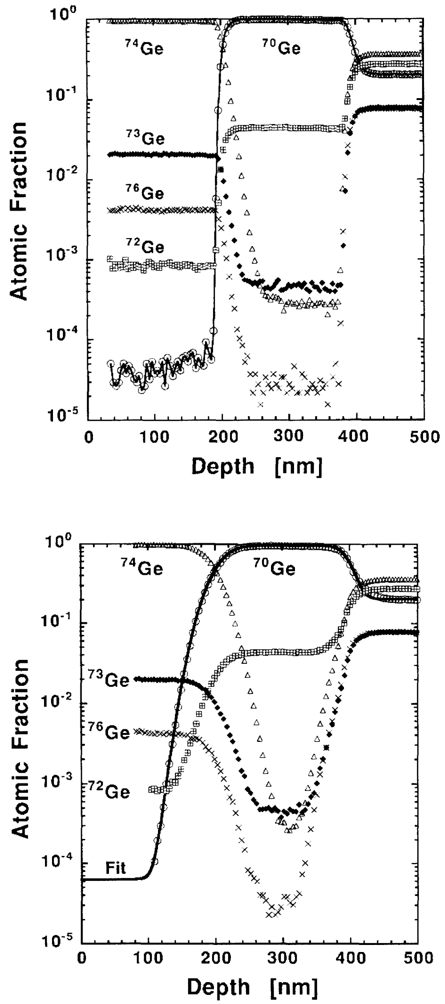

FIG. 14. Experimental concentration–depth profiles obtained with SIMS of the atomic fractions of ^{70}Ge, ^{72}Ge, ^{73}Ge, ^{74}Ge and ^{76}Ge of a ^{74}Ge/^{70}Ge bilayer (top) on a natural Ge substrate of the as-grown sample and (bottom) after annealing at 586°C for 55.55 h. The solid line is a fit to the ^{70}Ge profile. (Courtesy of Fuchs *et al.*, 1995.)

improving the accuracy of the data. The results for the Ge self-diffusivity are in excellent agreement with the earlier results (Vogel *et al.*, 1983; Werner *et al.*, 1983).

Summarizing the opportunities offered by internal isotope interfaces and multilayer structures for diffusion, we can make the following observations. First, the combination of isotope multilayers and SIMS makes studies with radioactive tracers obsolete. This happens at a convenient time when numerous regulations for laboratory work are making the use of any radioactive substance at least tedious but sometimes even impossible. Second, we can study diffusion at buried interfaces, away from surfaces, which always influence the intrinsic point-defect concentrations, the drivers for most diffusion processes. Third, the redundancy in data obtained from the major and the minor isotopes helps in obtaining improved data. Fourth, we foresee self- and impurity diffusion studies with doped isotope multilayers and SIMS. The case of superlattices selectively doped with NTD appears especially attractive (see Sec. 5). Heavily doped layers can be created between undoped ones with NTD, allowing the simultaneous observation of self- and impurity-diffusion. A Si superlattice consisting of alternating layers of undoped ^{28}Si and ^{30}Si represents such a case. Self-diffusion studies could be performed with the as-grown layers. NTD can be used to dope the ^{30}Si layers n-type with different P concentrations. Diffusion followed by SIMS analysis will generate one consistent and complete set of data, especially when the P concentration is chosen sufficiently high so as to exceed the SIMS sensitivity limit. Si self-diffusion studies using isotope heterostructures have been conducted recently by Bracht *et al.* (1998).

5. NEUTRON TRANSMUTATION DOPING (NTD) AND ELECTRONIC TRANSPORT STUDIES

5.1 Bulk Crystals

Compared with the relatively subtle effects generated by small isotope mass differences on a number of physical properties, some of which we have discussed, the consequences of thermal-neutron absorption by the various isotopes of a given element can be quite spectacular. Table 1 is a selective list of isotopes that are relevant to semiconductor science and technology. Every nucleus has a finite cross section for thermal neutron capture because neutrons carry no electrical charge and can reach the vicinity of a nucleus, where they experience the strong force that binds nuclei together.

Table 2. NTD Ge thermistor data (Fig. 15).

NTD No.	n Dose (cm^{-2})	N_{Ga} (cm^{-3})	N_{As} (cm^{-3})	N_{Se} (cm^{-3})	$N_{\mathrm{Ga}} - \sum N_{\mathrm{D}}$ (cm^{-3})
4	3.38×10^{17}	9.94×10^{15}	2.83×10^{15}	1.86×10^{14}	6.74×10^{15}
5	7.50×10^{17}	2.21×10^{16}	6.29×10^{15}	4.13×10^{14}	1.50×10^{16}
13	1.24×10^{18}	3.65×10^{16}	1.04×10^{16}	6.83×10^{14}	2.47×10^{16}
15	1.54×10^{18}	4.53×10^{16}	1.29×10^{16}	8.49×10^{14}	3.07×10^{16}
16	2.07×10^{18}	6.09×10^{16}	1.73×10^{16}	1.14×10^{15}	4.13×10^{16}
18	2.61×10^{18}	7.67×10^{16}	2.19×10^{16}	1.44×10^{15}	5.20×10^{16}
26	2.82×10^{18}	8.29×10^{16}	2.36×10^{16}	1.55×10^{15}	5.62×10^{16}
28	3.07×10^{18}	9.03×10^{16}	2.57×10^{16}	1.69×10^{15}	6.11×10^{16}
12	3.33×10^{18}	9.79×10^{16}	2.79×10^{16}	1.83×10^{15}	6.64×10^{16}
25	4.20×10^{18}	1.23×10^{17}	3.52×10^{16}	2.31×10^{15}	8.37×10^{16}

Neutron transmutation is especially intriguing for semiconductors because it can lead to doping (Lark-Horowitz, 1951; Larrabee, 1984; Fritzsche and Cuevas, 1960; Haller *et al.*, 1984). For example, P donors are formed in Si through the following NTD reaction:

$$^{30}_{14}\mathrm{Si} + \mathrm{n} \rightarrow ^{31}_{14}\mathrm{Si} \xrightarrow{T_{1/2}=2.62\,\mathrm{h}} ^{31}_{15}\mathrm{P} + \beta^{-} + \bar{\nu}_e \qquad (26)$$

The β^{-} is an electron of nuclear origin, and it shares the total energy of the reaction with an electronic antineutrino. These two particles are leptons and in nuclear reactions the sum of the leptons is conserved.

The P concentration is proportional to the concentration of $^{30}_{14}\mathrm{Si}$ (3.1% of the natural Si), the integrated thermal neutron flux n, and the thermal neutron capture cross section σ_{n}:

$$[\mathrm{P}] = \sigma_{\mathrm{n}}[^{30}_{14}\mathrm{Si}]\, n \qquad (27)$$

The major advantage of NTD is the extreme doping uniformity achieved. This is due to three factors. First, the $^{30}_{14}\mathrm{Si}$ isotopes are ideally randomly distributed in Si crystals; second, the thermal-neutron cross-sections are so small that several centimeters of Si will absorb an insignificant fraction of the thermal neutrons; and third, the thermal-neutron field is very much larger than any Si crystal produced today, which, in turn, guarantees a homogeneous flux. Disadvantages include radiation damage caused by residual fast neutrons and, in the specific case of natural Si, the small fraction of $^{30}_{14}\mathrm{Si}$. The latter disadvantage could be minimized with $^{30}_{14}\mathrm{Si}$ enrichment. A pure $^{30}_{14}\mathrm{Si}$ crystal would have an over 30 times faster transmutation rate than a natural Si crystal! The fast-neutron exposure and the concentration of deep levels would be reduced accordingly. Again cost-benefit issues will have to be resolved before any large-scale $^{30}_{14}\mathrm{Si}$ enrichment will become economically feasible.

The first NTD of Si used in the fabrication of high-voltage, high-power devices was reported in 1974 by Schnöller (1974). Since that time NTD has matured into a well controlled technology (Larrabee, 1984), and most high-power rectifiers and thyristors are fabricated from floating-zone Si doped with the NTD process.

Natural Ge contains five stable isotopes, three of which transmute into dopants after thermal-neutron capture (see Table 1). Fritzsche and Cuevas (1960) have used NTD of Ge for their electronic transport studies. In 1982 Haller *et al.* (1984) began doping ultrapure Ge with NTD to obtain crystals with a wide range of dopant concentrations for use in low-temperature sensors. The acceptor and donor concentrations are sufficiently high (Table 2) to lead to thermally assisted hopping conduction (Shklovskii and Efros, 1984) in the kelvin and millikelvin temperature range (Fig. 15). This material is now widely used for the fabrication of extremely sensitive low-temperature thermal detectors, also called thermistors (Haller, 1994; Haller *et al.* 1994). These thermistors are used as far-infrared photon detectors in a wide variety of laboratory instruments and in focal planes of terrestrial and balloon- and space-borne telescopes for far-infrared and millimeter- and radio-wave observations of sources in the universe. Currently there is a well focused worldwide effort under way to develop large thermistor arrays oper-

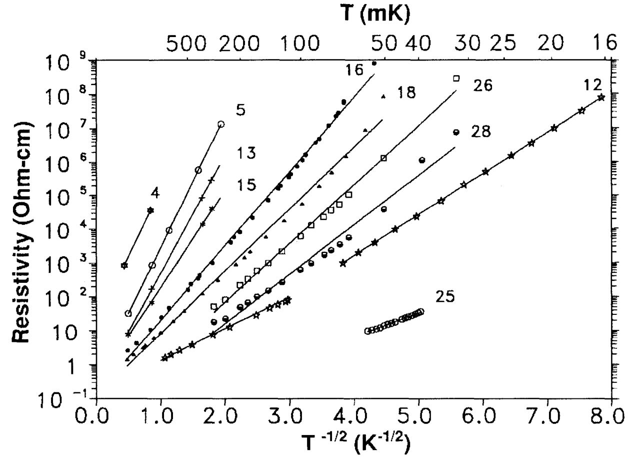

FIG. 15. Resistivity as a function of $T^{-1/2}$ for a number of NTD natural Ge samples. Hopping conduction is dominant and the simple dependence $\rho = \rho_0 \exp(T_0/T)^{-1/2}$ holds over nine orders of magnitude. Table 2 contains neutron fluences and dopant concentrations for the various samples. (Haller *et al.*, 1994.)

ating at subkelvin temperatures (Kreysa *et al.*, 1996; Holland *et al.*, 1996). This interest is a direct consequence of the spectacular successes already achieved with small NTD Ge bolometer arrays.

5.2 Selectively Doped Isotopic Superlattices

Homogeneous and heterogeneous superlattices and quantum wells can be formed with subatomic layer control by a number of epitaxial growth techniques. They have led to a variety of new devices and physical phenomena. The expression "band-gap engineering" has been introduced to describe the deliberate manipulation of the elemental composition of thin films and superlattices, which in turn creates changes in the band structure.

In view of the rich spectrum of the multilayer structures and the techniques to synthesize them, it is surprising that isotope superlattices consisting of layers with differing isotopic composition have only recently been considered. The idea of forming isotope superlattices by an epitaxial crystal growth technique and doping the various layers *selectively* with NTD was first proposed by Haller (1990).

Isotope superlattices offer a host of new possibilities and improved processing control. The main advantage of isotope superlattices is the possibility to separate the multilayer-layer growth and the doping into two completely independent processes. The following advantages arise from this separation:

1. The growth of multilayered structures of chemically pure isotopes automatically excludes impurity auto doping;
2. layer interdiffusion is limited to host isotope autodiffusion, the slowest diffusion process in all semiconductors; and
3. using NTD, one can selectively dope the various layers after the growth process.

The dopant concentration and species can be controlled by the proper selection of the isotope and its atomic concentration. The radiation damage caused by the fast neutrons accompanying the NTD process can be removed by thermal annealing at temperatures that are modest compared with the epitaxial growth temperatures.

The potential of superlattices selectively doped by NTD can be best illustrated with the formation of a Ge *nipi* structure. The necessary donors are created by neutron transmutation of ^{74}Ge into ^{75}As while the acceptors are obtained from ^{70}Ge by electron capture and ^{71}Ga formation. ^{72}Ge and ^{73}Ge can be used for the intrinsic layer or to reduce the atomic concentration of ^{70}Ge and ^{74}Ge. This may be necessary if one wants to adjust the relative doping concentrations in the *n* and *p* layers. Isotopic composition and doping of a hypothetical Ge *nipi* structure may look like this:

$$[^{74}\mathrm{Ge}_x : \mathrm{As}/^{72\text{ or }73}\mathrm{Ge}_{1-x}] : [^{72\text{ or }73}\mathrm{Ge}] :$$

$$[^{70}\mathrm{Ge}_y : \mathrm{Ga}/^{72\text{ or }73}\mathrm{Ge}_{1-y}] : [^{72\text{ or }73}\mathrm{Ge}] : \cdots .$$

The dopant concentrations after NTD are

$$[\mathrm{As}] = n_{\mathrm{thermal}}\sigma_{^{74}\mathrm{Ge}}x[^{74}\mathrm{Ge}]t, \tag{28}$$

$$[\mathrm{Ga}] = n_{\mathrm{thermal}}\sigma_{^{70}\mathrm{Ge}}y[^{70}\mathrm{Ge}]t. \tag{29}$$

Choosing the ratio of x and y to be the inverse of the ratio of the thermal neutron cross sections $\sigma_{^{74}\mathrm{Ge}}$ and $\sigma_{^{70}\mathrm{Ge}}$, respectively, leads to identical concentrations of As and Ga for any neutron exposure time t.

Isotope superlattices of the other semiconductors offer similar opportunities. Elemental heterogeneous superlattices (Si/Ge) as well as homogeneous and heterogeneous superlattices and quantum wells of III–V and II–VI semiconductors can be grown and doped with NTD. As Table 1 shows, most transmutation products form donors. This strong asymmetry in favor of n-type doping can be balanced to some degree by chemical doping during growth, with the known shortcomings.

A unique aspect of isotope multilayer structures selectively doped via NTD is the option of repeated doping of the same structure. This makes possible doping-dependent studies with the identical structure or device redoped with NTD as many times as necessary.

The idea of isotope multilayer structures predated our access to highly enriched isotopes of semiconductor elements. Now, where larger quantities of enriched isotopes have become available, we expect that such structures will soon be synthesized for doping purposes.

5.3 Electronic Transport and Metal–Insulator Transition

The electrical conductivity σ in a semiconductor is described by

$$\sigma = e(n\mu_n + p\mu_p) \tag{30}$$

with e = charge of the electron (-1.6×10^{-19} A s); n, p = concentration of electrons and holes, and μ_n, μ_p = mobilities of the electrons and holes, respectively. The motion of free charge carriers, *i.e.*, electrons in the conduction band and holes in the valence band, is affected by a number of scattering processes, each dominating over certain temperature ranges.

The total mobility can be related to the effective mass m^* of the free carrier and to an average relaxation time $\langle\tau\rangle$,

$$\mu = e\langle\tau\rangle/m^*. \tag{31}$$

The average relaxation time is obtained using the Maxwell–Boltzmann integration:

$$\langle\tau\rangle = \frac{4}{3\sqrt{\pi}}\int_0^\infty \frac{x^{3/2}\exp(-x)}{\tau_{\mathrm{ac}}^{-1} + \tau_{\mathrm{ion}}^{-1} + \tau_{\mathrm{neutral}}^{-1}}\,\mathrm{d}x, \tag{32}$$

with x being the ratio of the kinetic energy of the carrier and the thermal energy kT. The acoustic relaxation time τ_{ac}, or its inverse the scattering rate, are caused by phonons modulating the bandgap described by the deformation potential. This process dominates at high temperatures:

$$\tau_{\mathrm{ac}} \propto T^{-3/2}. \tag{33}$$

At lower temperatures Coulomb scattering from ionized impurities becomes important:

$$\tau_{\mathrm{ion}} \propto T^{3/2}. \tag{34}$$

In doped, low-compensation crystals the impurities de-ionize at very low temperature, and ionized-impurity scattering diminishes while a third process, neutral-impurity scattering, begins to dominate. This scattering process is rather weak and does not depend strongly on temperature. Several theoretical studies have been performed on neutral-impurity scattering (Erginsoy, 1950). They are all based on the scattering of an electron from a neutral hydrogen atom. The results of the various studies could not be tested satisfactorily up to now, because of a lack of suitable crystals.

The availability of highly enriched isotopes of Ge, which can be purified to residual dopant levels $<10^{12}\,\mathrm{cm}^{-3}$, has provided the first opportunity to measure neutral-impurity scattering over a wide temperature range (Itoh *et al.*, 1993). Itoh *et al.* (1994) reported on transport studies with a series of specially grown high-purity crystals of ^{70}Ge and ^{74}Ge. These crystals were doped with NTD followed by appropriate thermal annealing leading to well defined and ideally homogeneous Ga acceptor and As donor concentrations, respectively.

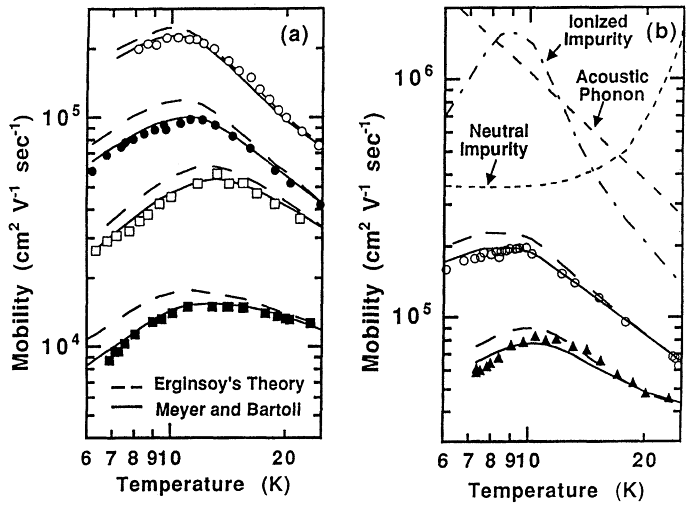

FIG. 16. Mobility measured for (a) four differently doped ^{74}Ge/As samples and (b) two ^{70}Ge/Ga samples. The theoretically determined mobilities are shown as broken lines (Erginsoy, 1950) and continuous lines (Meyer and Bartoli, 1981). The contributions of the different scattering mechanisms to the total mobility of the high-mobility sample ^{70}Ge/Ga are also shown in (b). (Courtesy of Itoh *et al.*, 1994.)

Doping levels between 10^{14} and $10^{16}\,\mathrm{cm}^{-3}$ with less than 1% compensation were chosen. Extensive Hall-effect and resistivity measurements from room temperature down to 4.2 K yielded very accurate free-carrier concentrations and mobilities as functions of temperature and doping level. Below 10 K the dopant impurities rapidly de-ionize and the neutral-impurity concentration rises to more than 100 times the residual ionized-impurity concentration. It is this large excess of homogeneously distributed neutral scattering centers that is necessary for an accurate study of τ_{neutral} and that can be achieved through the combination of the control of isotopic composition and NTD. Figure 16 shows data points for four *n*-type ^{74}Ge crystals [16(a)] and two *p*-type ^{70}Ge crystals [16(b)]. The continuous lines are not fits to the data but represent the theoretical prediction of the model proposed by Meyer and Bartoli (1981). No adjustable parameters had to be used to obtain this excellent agreement between theory and experiment. The older and less sophisticated theory of Erginsoy (1950) clearly leads to a less satisfactory agreement. The approach taken for these transport studies in Ge could easily be adapted to other semiconductors. Si and GaAs would be good candidates because high-quality crystals with low residual dopant concentrations can be grown. Both materials become *n*-type through the NTD process, which limits the studies to electron transport.

A long-standing problem in solid state physics is the metal–insulator transition (MIT). A very large body of experimental and theoretical studies exists but is far too extensive to be reviewed here. In principle, semiconductors are ideally suited for the study of this important transition of a solid from the insulating to the conducting state. All that has to be done is to increase the doping concentration in a semiconductor in a systematic fashion from values below to values above the MIT and to measure the transport properties. Close to the MIT the conductivity becomes very sensitive to the homogeneity of the dopant distribution. In melt-doped, diffused, and ion-implanted crystals the doping homogeneity seems to be the major cause for large discrepancies between the results reported by various groups. In the absence of consistent data, the theoretical contributions cannot be tested thoroughly.

The combination of isotope enrichment, ultrapurification, and NTD has led to *p*-type ^{70}Ge crystals that are homogeneously doped with a truly random dopant distribution and a donor compensation of the order 1% (Itoh *et al.*, 1996). A series of crystals with Ga acceptor concentrations ranging from 3×10^{16} to $2.6 \times 10^{17}\,\mathrm{cm}^{-3}$ have been studied in detail on the insulator and the metal sides of the transition. Figure 17 shows that the experimentally determined parameter T_0 as a function of the Ga concentration on the insulator side and the

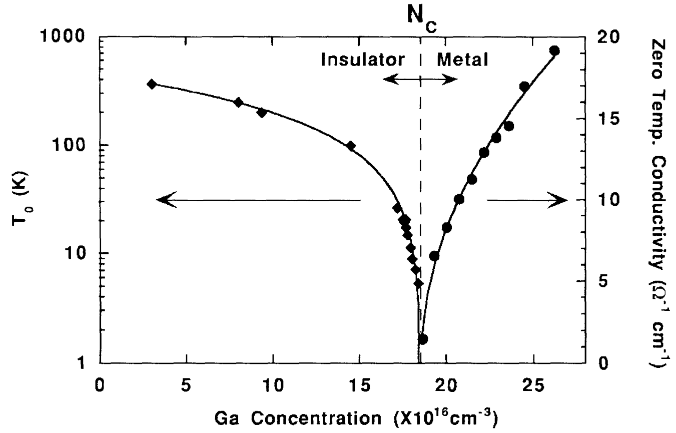

FIG. 17. Conductivity parameters for NTD ^{70}Ge below and above the metal–insulator transition. The left-hand side shows the experimentally determined values of T_0 [Eq. (35)] of 14 insulating samples as a function of Ga concentration (◆). The right-hand side shows the zero-temperature conductivity $\sigma(0)$ obtained from ten metallic samples as a function of Ga concentration (●). The critical concentrations obtained on the insulator and on the metallic side agree to within a few percent. (Courtesy of Itoh *et al.*, 1996.)

zero-temperature conductivity on the metal side converge precisely at the critical concentration of $1.856 \times 10^{17}\,\text{cm}^{-3}$. The resistivity ρ on the insulator side follows the expression developed by Shklovskii and Efros (1984):

$$\rho = \rho_0 \exp(T_0/T)^{\nu} \qquad (36)$$

with ρ_0 the prefactor, T_0 a constant depending on the dopant concentration, and ν the critical exponent.

Temperature-dependant resistivity measurements show that $\nu \cong 0.5$ for all doping concentrations below the MIT. The value of ν remains 0.5 for donor compensation up to $\gtrsim 40\%$. Most of the published MIT theories find values for ν different from 0.5. Stimulated by the MIT studies with isotopically engineered Ge, Phillips (1997) has formulated new theoretical treatments that produce the experimentally determined value of ν. Work is in progress to repeat the same type of experiment with ^{74}Ge, which turns *n*-type with As donors after NTD.

6. FUTURE DEVELOPMENTS AND CONCLUSIONS

Considering what has been achieved with isotopically controlled semiconductors, we are confident that a very large number of interesting experiments can be performed if there is enough commitment to creating the necessary materials. There can be little doubt that we will see experiments with isotopically controlled Si bulk crystals and perhaps ^{28}Si/^{30}Si thin films and superlattices. The studies performed with Ge should definitely be repeated with Si in order to verify and consolidate the understanding gained with Ge. Once they are grown, Si isotope superlattices can be doped *n*-type in the ^{30}Si layers.

Isotopically controlled III–V and II–VI compound semiconductors and their alloys are most promising in connection with NTD. Zn-containing II–VI semiconductors can be doped either with Cu or Ga or not at all, depending on the particular choice of the Zn isotope (Wheeler *et al.*, 1997). The Group VI partners S, Se, or Te offer an abundance of doping possibilities including donors and acceptors. Post-growth doping of II–VI semiconductors with the standard techniques (diffusion and ion implantation) has found rather limited success, and NTD in combination with isotope control should be explored as a possible solution to the doping problem. In exchange one will have to deal with nuclear-decay recoil and fast-neutron radiation damage.

NTD will play a role in diffusion studies. Self-diffusion in several semiconductors has been studied with isotope multilayer structures and superlattices. Such isotope superstructures can be doped over a broad range of impurity concentrations. With such an approach one can obtain large sets of data for self- and dopant diffusion with built-in redundancy and a minimal chance of errors introduced by hidden materials parameters (*e.g.* electrically inactive impurities, dislocations, precipitates). If undoped superlattices are grown, extremely sharp interfaces, unaffected by impurity-promoted interdiffusion, can be obtained.

A family of semiconductors that can profit from isotope control are the Cd-containing II–VI compounds. Rowe (*et al.*, 1974) led the way with CdTe crystals, in which ^{113}Cd, the isotope with one of the largest thermal-neutron capture cross sections, was essentially eliminated. The other members of this family, *i.e.*, CdS, CdSe, and alloys, can also be studied with neutron diffraction after ^{113}Cd has been removed.

Moving to atomic-level dimensions we can foresee exciting experiments with scanning tunneling microscopy (STM) (see TUNNELING MICROSCOPY AND SPECTROSCOPY). Recent reports indicate that donors have been observed with STM in GaAs (Zheng *et al.*, 1994). Let us imagine that a semiconductor is exposed to a certain neutron fluence and that the half-life of the transmutation process lies in the range of hours to days. The crystal surface can be observed with an STM after neutron exposure. If the concentration of the STM-observed impurities should increase exponentially with the same half-life as that of the transmuting isotope, we would have strong proof for STM observations of dopant impurities or other impurities, as well. With very high-speed STM it may become possible to observe relaxation around a defect or dopant formed by neutron transmutation or another nuclear reaction in real time!

Local vibrational-mode studies could become even more powerful than they already are, if isotopically controlled host crystals were available. The enormous simplification gained from isotopically pure host crystals has been clearly demonstrated with LVM studies of O in Ge. LVM studies of Ca and Ti in CdTe and Mg in ZnTe have shown spectrally resolved isotope effects of both the Te host-crystal isotopes and the impurity isotopes (Sciacca *et al.*, 1996). "Isotopically doubly controlled" experiments in which both the impurity and the host crystal are isotope controlled (pure as well as deliberately mixed isotopes) present ideal solid state systems for LVM studies.

Nuclear reactions with specific isotopes of semiconductor-forming elements offer a range of opportunities that have so far not been explored. There exists information on a very large number of nuclear reactions between energetic ions and practically all the stable isotopes of the periodic table. We foresee experiments focusing on radiation damage, on ion implantation, and on nuclear transmutation. Isotope control will simplify many of these studies.

Isotope control and engineering of semiconductors and other solids offers numerous possibilities for solid state studies and novel devices. The post-Cold War era has brought great opportunities for making highly enriched isotopes more readily available to the science and technology community. Inventiveness and creativity are required to take full advantage of the vast array of natural, stable isotopes, which so far have played only a minor role in a few special cases.

ACKNOWLEDGMENTS

First and foremost, I wish to thank Valerii Ozhogin of the Kurchatov Institute in Moscow, who had the courage to make the first macroscopic quantities of highly enriched Ge available for many exciting experiments. I am indebted to many colleagues at Berkeley and all over the world with whom we have been collaborating on various experiments with isotopically controlled semiconductors. Their names are presented in alphabetical order: Gerhard Abstreiter, Jeffrey W. Beeman, Fernando Briones, Manuel Cardona (and the members of his group), Gordon Davies, Karl Eberl, Jon Erickson, John Farmer, Heinz D. Fuchs, William L. Hansen, Kohei M. Itoh, Anant K. Ramdas (and the members of his group), Wladek Walukiewicz, and Lei Wang.

The author acknowledges the Max-Planck-Research Prize, which he shared with Manuel Cardona, offered by the Max-Planck Society, Germany, in support of our research on isotopically controlled semiconductors. The work at Berkeley was also supported in part by the Lawrence Berkeley National Laboratory Directed Research and Development Program and in part by the Office of Energy Research, Office of Basic Energy Sciences, Materials Sciences Division of the U.S. Department of Energy, under contract DE-AC03-76SF00098.

Works Cited

Agekyan, V. F., Asnin, V. M., Kryukov, A. M., Markov, I. I., Rud', N. A., Stepanov, V. I., Churilov, A. B. (1989), Fiz. Tverd. Tela **31**, 101–104 (*Sov. Phys. Solid State* **31**, 2082–2084).

Anderson, P. W. (1955), *Phys. Rev.* **109**, 1492–2084.

Anthony, T. R., Banholzer, N. F., Fleischer, J. F., Wei, L.-H., Kuo, P. K., Pryor, R. L., Thomas, R. L. (1990), *Phys. Rev. B* **42**, 1104–1111.

Asen-Palmer, M., Bartkowski, K., Gmelin, E., Cardona, M., Zhernov, A. P., Ozhogin, V., Inyushkin, A., Taldenkov, A., Itoh, K., Hansen W. L., Haller, E. E. (1997) *Phys. Rev. B* **56**, 9431–9447.

Banholzer, W. F., Anthony, T. R. (1992) *Thin Solid Films* **212**, 1–10.

Barker, A. S., Sievers, A. J. (1975), *Rev. Mod. Phys.* **47**, Suppl. 2.

Becker, P. (1995), personal communication.

Berezin, A. A., Ibrahim, A. M. (1988), *Mater. Chem. Phys.* **19**, 407–430.

Berezin, A. A. (1989), *J. Phys. Chem. Solids* **50**, 5–8.

Bockstedte, M., Scheffler, M. (1996), *Bull. Am. Phys. Soc.* **41** (1), 562.

Bracht, H., Stolwijk, N. A., Mehrer, H. (1995), *Phys. Rev B* **52**, 16542–16560.

Bracht, H., Haller, E. E., Clark-Phelps, R. (1998), *Phys. Rev. Lett.* **81**, 293–6.

Buschert, R. C., Merlin, A. E., Pace, S., Rodriguez, S., Grimsditch, M. H., (1988), *Phys. Rev. B* **38**, 5219–5221.

Callaway, J. (1959), *Phys. Rev.* **113**, 1046–1051.

Campbell, D. R. (1975), *Phys. Rev. B* **12**, 2318–2324.

Cardona, M., Etchegoin, P., Fuchs, H. D., Molinàs-Mata, P. (1993), *J. Phys. Cond. Matter* **5**, A61–72.

Cardona, M., Grein, C. H., Fuchs, H. D., Zollner, S. (1992), *J. Non-Cryst. Solids* **141**, 257–264.

Chen, B., Zhang, Q.-M., Bernholc, J. (1994), *Phys. Rev. B* **49**, 2985–2988.

Chrenko, R. M. (1988), *J. Appl. Phys.* **63**, 5873–5875.

Collins, A. T., Davies, G., Kanda, H., Woods, G. S. (1988), *J. Phys. C* **21**, 1363–1376.

Collins, A. T., Lawson, S. C., Davies G., Kanda, H. (1990), *Phys. Rev. Lett.* **65**, 891–894.

Davies, G., Lightowlers, E. C., Itoh, K., Hansen, W. L., Haller, E. E., Ozhogin, V. (1992), *Semicond. Sci. Technol.* **7**, 1271–1273.

Davies, G., Lightowlers, E. C., Hui, T. S., Ozhogin, V., Itoh, K., Hansen, W. L., Haller, E. E. (1993), *Semicond. Sci. Technol.* **8**, 2201–2204.

Davies, G. (1994), *Semicond. Sci. Technol.* **9**, 409.

De Laeter, J. R., Heumann, K. G., Rosman, K. J. R. (1991), *J. Phys. Chem. Ref. Data* **20**, 1327–1337.

Denteneer, P. J. H., Van de Walle, C. G., Pantelides, S. J. (1989), *Phys. Rev. Lett.* **62**, 1884–1887.

Erginsoy, C. (1950), *Phys. Rev.* **79**, 1013–1014.

Etchegoin, P., Weber, J., Cardona, M., Hansen, W. L., Itoh, K., Haller, E. E. (1992), *Solid State Commun.* **83**, 843–848.

Etchegoin, P., Fuchs, H. D., Weber, J., Cardona, M., Pintschovius, L., Pyka, N., Itoh, K., Haller, E. E. (1993), *Phys. Rev. B* **48**, 12661–12671.

Frank, W., Gösele, U., Mehrer, H., Seeger, A. (1984), in: G. E., Murch, A. S. Norwick (Eds.), *Diffusion in Crystalline Solids*, New York: Academic Press, pp. 63–142.

Fritzsche, H., Cuevas, M. (1960), *Phys. Rev.* **119**, 1238–1245.

Fuchs, H. D., Grein, C. H., Devlen, R. I., Kuhl, J., Cardona, M., (1991a), *Phys. Rev. B* **44**, 8633–8642.

Fuchs, H. D., Grein, C. H., Thompson, C., Cardona, M., Hansen, W. L., Haller, E. E., Itoh, K. (1991b), *Phys. Rev. B* **43**, 4835–4842.

Fuchs, H. D., Grein, C. H., Cardona, M., Hansen, W. L., Itoh, K., Haller, E. E. (1992), *Solid State Commun.* **82**, 225–228.

Fuchs, H. D., Molinàs-Mata, P., Cardona, M. (1993), *Superlatt. Microstruct.* **13**, 447–458.

Fuchs, H., Walukiewicz, W., Haller, E. E., Dondl, W., Schorer, R., Abstreiter, G., Rudnev, A. I., Tikhomirov, A. V., Ozhogin, V. I. (1995), *Phys. Rev. B* **51**, 16817–16821.

Geballe, T. H. (1990), *Science* **250**, 1194.

Geballe, T. H., Hull, G. (1958), *Phys. Rev.* **110**, 773–775.

Gösele, U. (1989), in: R. A. Levy (Ed.), *Microelectronic Materials and Processes*, Dordrecht: Kluwer Academic, pp. 583–634.

Gösele, U. (1994), in: D. Bloor, M. C. Flemings, S. Mahajan, R. W. Cahn (Eds.), *Encyclopedia of Advanced Materials*, Vol. **1**, New York: Pergamon, pp. 629–635.

Götz, W., Johnson, N. M., Bour, D. P., McCluskey, M. D., Haller, E. E. (1996), *Appl. Phys. Lett.* **69**, 3725–3727.

Haller, E. E. (1978), *Phys. Rev. Lett.* **40**, 584–586.

Haller, E. E. (1990), *Semicond. Sci. Technol.* **5**, 319–321.

Haller, E. E. (1994), *Infrared Phys. Technol.* **35**, 127–146.

Haller, E. E. (1995), in: S. Ashok, J. Chevallier, I. Akasaki, N. M. Johnson, B. L. Sopori (Eds.), *Defect and Impurity Engineered Semiconductors and Devices*, MRS Symposium Proceedings No. 378, Pittsburgh: Materials Research Society, pp. 547–565.

Haller, E. E., Itoh, K. M., Beeman, J. W., Hansen, W. L., Ozhogin, V. I. (1994a), in: D. L. Crawford, E. R. Craine (Eds.), *Instrumentation in Astronomy VIII*, SPIE Proceedings Vol. **2198**, Bellingham, WA: SPIE, pp. 630–637.

Haller, E. E., Palaio, N. P., Rodder, M., Hansen, W. L., Kreysa, E. (1984), in: R. D. Larrabee (Ed.), *Proceedings of the Fourth International Conference on Neutron Transmutation Doping of Semiconductor Materials*, National Bureau of Standards, Gaithersburg, MD, June 1–2, 1982, New York: Plenum Press, pp. 21–36.

Hass, K. C., Tamor, M. A., Anthony, T. R., Banholzer, W. F. (1991), *Phys. Rev. B* **44**, 12046–12049.

Hass, K. C., Tamor, M. A., Anthony, T. R., Banholzer, W. F. (1992), *Phys. Rev. B* **45**, 7171–7182.

Hofmann, G., Madok, J., Haegel, N. M., Roos, G., Johnson, N. M., Haller, E. E. (1992), *Appl. Phys. Lett.* **61**, 2914–2916.

Holland, W. S., Ade, P. A. R., Griffin, M. J., Hepburn, I. D., Vickers, D. G., Cunningham, C. R., Hastings, P. R., Gear, W. K., Duncan, W. D., Baillie, T. E. C., Haller, E. E., Beeman, J. W. (1996), *Intl. J. Infrared MM Waves* **17**, 669–692.

Holloway, H., Hass, K. C., Tamor, M. A., Anthony, T. R., Banholzer, W. F. (1991), *Phys. Rev. B* **44**, 7123–7126; erratum (1992), *Phys. Rev. B* **45**, 6353.

Hu, S. M. (1992), *J. Electrochem. Soc.* **139**, 2066–2075.

Itoh, K. M., Haller, E. E., Beeman, J. W., Hansen, W. L., Emes, J., Reichertz, L. A., Kreysa, E., Shutt, T., Cummings, A., Stockwell, W., Sadoulet, B., Muto, J., Farmer, J. W., Ozhogin, V. I. (1996), *Phys. Rev. Lett.* **77**, 4058–4061.

Itoh, K., Hansen, W. L., Haller, E. E., Farmer, J. W., Ozhogin, V. I., Rudnev, A., Tikhomirov, A. (1993), *J. Mater. Res.* **8**, 1341–1347.

Itoh, K. M., Walukiewicz, W., Fuchs, H. D., Beeman, J. W., Haller, E. E., Farmer, J. W., Ozhogin, V. I. (1994), *Phys. Rev. B* **50**, 16995–17000.

Kaiser, W., Keck, P. H., Lange, C. F. (1956), *Phys. Rev.* **101** 1264–1268.

Kreysa, E., Beeman, J., Haller, E. E. (1996), in: G. Pilbratt (Ed.), *Proceedings of the 30th ESLAB Symposium on "Submillimetre and Far Infrared Space Instrumentation,"* Noordwijk, Netherlands, Sept. 24–27, 1996, ESA Special Publication No. 388, Paris: European Space Agency, pp. 111–114.

Lark-Horowitz, K. (1951), in: H. K. Henish (Ed.), *Proceedings of the Conference on Semi-Conducting Materials*, London: Butterworth Sci. Publ., pp. 47–69.

Larrabee, R. D. (Ed.) (1984), *Neutron Transmutation Doping of Semiconductor Materials*, New York: Plenum Press.

Lederer, C. M., Shirley, V. S. (Eds.) (1978), *Table of Isotopes*. 7th ed., New York: John Wiley.

Leigh, R. S., Newman, R. C. (1982), *J. Phys. C* **15**, L 1045–1051.

Mayur, A. J., Sciacca, M. D., Udo, M. K., Ramdas, A. K., Itoh, K., Wolk, J., Haller, E. E. (1994), *Phys. Rev. B* **49**, 16293–16299.

McCluskey, M. D., Haller, E. E., Walker, J., Johnson, N. M. (1995), *Phys. Rev. B* **52**, 11859–11864.

McCluskey, M. D., Haller, E. E., Walukiewicz, W., Becla, P. (1996), *Phys. Rev. B* **53**, 16297–16301.

McMurray, Jr., R. E., Haegel, N. M., Kahn, J. M., Haller, E. E. (1987), *Solid State Commun.* **61**, 27–32.

Meyer, J. R., Bartoli, F. J. (1981), *Phys. Rev. B* **24**, 2089–2100.

Morelli, D. T., Smith, G. W., Heremans, J. P., Anthony, T. R., Banholzer, W. F. (1991), in: R. Messier, J. T. Glass, J. E. Butler, R. Roy (Eds.), *Proceedings of the 2nd International Conference on New Diamond Science and Technology*, Pittsburgh: MRS, pp. 869–873.

Newman, R. C. (1973), *Infrared Studies of Crystal Defects*, London: Taylor & Francis.

Newman, R. C. (1993), in: E. R. Webr (Ed.), *Semiconductors and Semimetals*, Vol. 38, New York: Academic Press, pp. 117–187.

Nilsson, G., Nelin, G. (1971), *Phys. Rev. B* **3**, 364–369.

Nordheim, L. (1931), *Ann. Phys.* **9**, 607–678.

Olson, J. R., Pohl, R. O., Vandersande, J. W., Zoltan, A., Anthony T. R., Banholzer, W. F (1993), *Phys. Rev. B* **47**, 14850–14856.

Ozhogin, V. I., Inyushkin, A., V., Taldenkov, A. N., Tikhomirov, A. V., Popov, G. E., Haller, E. E., Itoh, K. (1996), *Pis. Zh. Eksp. Teor. Fiz.* **63**, 463–467 (*JETP Lett.* **63**, 490–494).

Palfrey, H. D. (1989), *J. Crystal Growth* **94**, 778–782.

Parks, C., Ramdas, A. K., Rodriguez, S., Itoh, K. M., Haller, E. E. (1994), *Phys. Rev. B* **49**, 14244–50.

Phillips, J. C. (1997), personal communication.

Pomeranchuk, I. (1942), *J. Phys.* (Moscow) **6**, 237–250.

Ramdas, A. K., Rodriguez, S., Grimsditch, M., Anthony, T. R., Banholzer, W. F. (1993), *Phys. Rev. Lett.* **71**, 189–192.

Rowe, J. M., Nicklow, R. M., Price, D. L., Zanio, K., (1974), *Phys. Rev. B* **10**, 671–675.

Schmid U., Christensen, N. E., Cardona, M. (1990), *Solid State Commun.* **75**, 39–43.

Schnöller, M. S. (1974), *IEEE Trans. Electr. Dev.* **ED-21**, 313–314.

Sciacca, M. D., Mayur, A. J., Kim, H., Miotkowski, I., Ramdas A. K., Rodriguez, S. (1996), *Phys. Rev B*, **53**, 12878–12883.

Shaw, D. (1988), *J. Crystal Growth* **86**, 778–796.

Shklovskii, B. I., Efros, A. L. (1984), in: *Electronic Properties of Doped Semiconductors*, Solid State Series Vol. 45, Berlin: Springer-Verlag.

Slack, G. A. (1973), *J. Phys. Chem. Solids* **34**, 321–335.

Spitzer, W. G. (1971), in: O. Madelung, (Ed.) *Festkörperprobleme XI/Advances in Solid State Physics*, Vol. 11, Braunschweig: Pergamon-Vieweg, pp. 1–44.

Spitzer, J., Ruf, T., Cardona, M., Dondl, W., Schorer, R., Abstreiter, G., Haller, E. E., (1994), *Phys. Rev. Lett.* **72**, 1565–1568.

Stavola, M., Pearton, S. J., Lopata, J., Dautremont-Smith, W. C. (1988), *Phys. Rev. B* **37**, 8313–8318.

Tan, T. Y., Gösele, U. (1985) *Appl. Phys. A* **37**, 1–17.

Tan, T. Y., Gösele, U., Yu, S. (1991), *Crit. Rev. Solid State Mater. Sci.* **17**, 47–106.

Tan, T. Y., You, H. M., Yu, S., Gösele, U., Jäger, W., Boeringer, D. W., Zypman, F., Tsu, R., Lee, S.-T. (1992), *J. Appl. Phys.* **72**, 5206–5212.

Theis, W. M., Bajaj, K. K. (1982), *Appl. Phys. Lett.* **41**, 70–72.

Theis, W. M., Bajaj, K. K., Litton, C. W., Spitzer, W. G. (1983), *Physica B + C* **117/118**, 116–118.

Vegard, L. (1921), *Z. Phys.* **5**, 17–26.

Vogel, G., Hettich, G., Mehrer, H., (1983), *J. Phys. C* **16**, 6197–6204.

Vogelgesang, R., Ramdas, A. K., Rodriguez, S., Grimsditch, M., Anthony, T. R. (1996), *Phys. Rev. B* **54**, 3989–3999.

Wang, L., Hsu, L., Haller, E. E., Erickson, J. W., Fischer, A., Eberl, K., Cardona, M. (1996), *Phys. Rev. Lett.* **76**, 2342–2345.

Wang, L., Wolk, J. A., Haller, E. E., Erickson, J. W., Cardona, M., Ruf, T., Silveira, J. P., Briones, F. (1997), *Appl. Phys. Lett.* **70**, 1831–1833.

Watkins, G. D., Fowler, W. B., Stavola, M., DeLeo, G. G. (1990), *Phys. Rev. Lett.* **64**, 467–470.

Wei, L., Kuo, P. K., Thomas, R. L., Anthony, T. R., Banholzer, W. F. (1993), *Phys. Rev. Lett.* **70**, 3764–3767.

Werner, M., Mehrer, H., Siethoff, H. (1983), *J. Phys. C* **16**, 6185–6195.

Wheeler, E. D., Boone, J. L., Farmer, J. W., Chandrasekhar, H. R. (1997) *J. Appl. Phys.* **81**, 524–526.

Willoughby, A. F. W. (1994), in: S. Mahajan, (Ed.), *Handbook on Semiconductors*, Vol. **3**, Amsterdam: Elsevier, Chap. 16, pp. 1231–1249.

Zheng, J. F., Liu, X., Newman, N., Weber, E. R., Ogletree, D. F., Salmeron, M. (1994), *Phys. Rev. Lett.* **72**, 1490–1493.

Ziman, J. M. (1963), *Electrons and Phonons*, London: Oxford University Press, pp. 288–319.

Zimmermann, H., Ryssel, H. (1992), *Appl. Phys. A* **55**, 121–134.

Zollner, S., Cardona, M., Gopalan, S. (1992), *Phys. Rev. B* **45**, 3376–3385.

Further Reading

Haller, E. E. (1995), "Isotopically Engineered Semiconductors," *J. Appl. Phys.* **77**, 2857–2878.

Larrabee, R. D. (Ed.) (1984), *Neutron Transmutation Doping of Semiconductor Materials*, New York: Plenum.

Olander, D. R. (1978), "The Gas Centrifuge," *Scientific American*, **239** (2) 37–43.

Yu, P. Y., Cardona, M. (1996), *Fundamentals of Semiconductors*, New York: Springer.

SEMICONDUCTORS, NITRIDE

F. A. Ponce, *Department of Physics and Astronomy, Arizona State University, Tempe, Arizona, U.S.A.*

Introduction 285
1. **Crystal Chemistry** 286
1.1 General 286
1.2 Chemical Bonding 286
1.3 Crystal Structures 287
1.3.1 Boron Nitride 287
1.3.2 Aluminum Nitride 288
1.3.3 Gallium Nitride 288
1.3.4 Indium Nitride 289
1.4 Band Gap 289
1.5 Mechanical Properties 289
1.6 Thermal Stability and Etching 290
2. **Growth Techniques** 290
2.1 General 290
2.2 Bulk Growth 291
2.2.1 Aluminum Nitride 291
2.2.2 Gallium Nitride 291
2.2.3 Indium Nitride 291
2.3 Substrates for Heteroepitaxy 291
2.3.1 Silicon Carbide as a Substrate 292
2.3.2 Sapphire as a Substrate 292
2.4 Growth with Molecular Beams 292
2.5 Growth by Vapor-Phase Transport 292
2.6 Doping 293
2.6.1 *n*-Type 293
2.6.2 *p*-Type 293
3. **Materials Properties** 293
3.1 Lattice Defects 293
3.1.1 Point Defects 293
3.1.1.1 Native Defects 293
3.1.1.2 Impurities 293
3.1.2 Dislocations 294
3.1.3 Inversion Domains 294
3.1.4 Stacking Faults 294
3.2 GaN on Silicon Carbide 294
3.3 GaN on Sapphire 295
3.3.1 Microstructure 295
3.3.2 Electrical 296
3.3.3 Optical 297
4. **Applications** 297
4.1 General 297
4.2 Light-Emitting Diodes 297
4.2.1 Traffic Lights 298
4.2.2 Signals 299
4.2.3 Interior Automobile Lighting 299
4.2.4 Full-Color Large-Panel Displays 299
4.2.5 Interior Lighting 299
4.3 Diode Lasers 299
4.3.1 High-Density Data Storage 299
4.3.2 High-Resolution Printing 300
4.3.3 Projection Displays 300
4.3.4 Underwater Communications 300
4.4 Solid-State Electronics 301
4.4.1 High-Temperature Electronics 301
4.4.2 High-Power Electronics 301
4.4.3 Solar-Blind Electronics 301
4.5 Cold-Cathode Applications 301
4.6 Challenges and Future Developments 301
5. **Summary** 301
Glossary 301
Works Cited 302
Further Reading 303

INTRODUCTION

The Group III metals (B, Al, Ga, In) when combined with nitrogen form a special family of materials that are extremely hard, refractory, and resistant to chemical attack. Their physical characteristics include strong chemical bonds, large electronic band gaps, and low electron affinities. Because of these properties the Group III materials have found a niche is several important technological applications. Boron nitride is second only to diamond in terms of hardness and thermal conductivity, and it is used for coating surfaces of tools used for abrasive purposes and in other adverse environments. The alloys resulting from the combination of AlN, GaN, and InN can be grown with an electronic band gap appropriate for visible- and ultraviolet-light emitters and detectors, as well as with robust mechanical and thermal properties that make them useful for high-temperature and high-power electronic applications.

This article provides a concise review of the

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

development of the nitrides, their properties, and their applications. Extensive journal reviews and books have recently been produced, and some are listed at the end for the interested reader.

Brief history. The Group III nitrides are not really new materials, and their recent success results from much effort over the last three decades. A brief account of the development of the nitrides follows; more details and references on the evolution of the nitrides are found in Ponce and Bour (1997). AlN was first reported in 1907, placing it among the first compound semiconductors ever synthesized. The wurtzite lattice structure of GaN was determined in 1937. In the last four decades, much of the research in the field has been stimulated by the desire to achieve short-wavelength light emission such as for green and blue solid-state lamps. Growth of GaN crystalline films by hydride vapor-phase epitaxy was first reported by Maruska and Tietjen (1969). Pankove *et al.* (1971) produced the first blue GaN LEDs using metal–insulator–semiconductor structures. Also in the same year Manasevit grew GaN by metalorganic chemical vapor deposition for the first time, a technique that is widely used today; and Dingle reported stimulated light emission at 2 K from GaN needles. The search for high-quality materials encountered three important problems that appeared insurmountable for many years: the first thin films were rough and fractured, *n*-type conductivity was difficult to control, and *p*-type conductivity appeared virtually impossible to achieve. Akasaki and Amano conquered these obstacles: smooth morphologies were achieved by Amano *et al.* (1986) using low-temperature AlN nucleation layers. Once specular films had been achieved, it was shown that the use of silane (SiH_4) could produce effective and controllable *n*-type doping. It took a few more years to achieve *p*-type conductivity. Magnesium was expected to be a good acceptor (*p*-type dopant), but the first films with high concentrations of Mg failed to exhibit *p*-type characteristics. Amano *et al.* (1989) showed that low-energy electron-beam irradiation (LEEBI) converted the highly resistive Mg-doped GaN into conductive films with high hole concentrations ($10^{16}\,cm^{-3}$). Under these conditions they were able to produce the first *p–n* junction LED. Nakamura (1991) introduced low-temperature GaN nucleation layers, with further improvement of the electronic properties, and then demonstrated (Nakamura *et al.*, 1992) that thermal annealing caused activation of Mg in GaN in a similar manner as the LEEBI technique. At that point few could predict the rapid development and commercialization of GaN-based optoelectronic devices in the following years.

Current relevance. The nitride semiconductors have positioned themselves for high impact in important emerging technologies. Blue and green LEDs based on GaN are now commercially available and, in combination with red AlInGaP LEDs, they can provide solid-state illumination with substantially better power efficiencies than incandescent tungsten-filament light bulbs. Blue laser diodes operating continuously at room temperature with long lifetimes ($>10\,kh$) have been demonstrated and are now commercially available. Much excitement and activity currently exists for the further commercialization of these new materials, and substantial technological innovations are foreseen. One area is in the production of the three primary colors for large-panel displays, outdoor signs, and ultimately full solid-state indoor lighting. Another area results from the ability to focus shorter wavelengths into smaller spots without increment in size and cost of the optical system. This will allow high-density optical information storage and readout, full-color photographic laser printing, and high-resolution xerographic laser printing. The development of green and blue laser diodes will make possible the implementation of small and inexpensive full-color projection television.

1. CRYSTAL CHEMISTRY

1.1 General

This section deals with the chemical aspects of the Group III nitrides, their crystal structures, and the electronic and thermal properties. For in-depth data and references consult Madelung (1982). A detailed treatment of the properties of the Group III nitrides by Edgar (1994) is also recommended.

1.2 Chemical Bonding

Tetrahedral coordination is a characteristic of the atomic bonding configuration in many

Table 1. Electronic properties of the elements found in the nitride semiconductors. Values are for atoms in tetrahedrally coordinated environments. Values taken from Kittel (1976) and Philips (1973).

	Ground-state electron configuration	Electronegativity	Atomic radius (Å)
B	[He] $2s^2$ $2p^1$	2.0	0.88
Al	[Ne] $3s^2$ $3p^1$	1.18	1.26
Ga	[Ar] $3d^{10}$ $4s^2$ $4p^1$	1.13	1.26
In	[Kr] $4d^{10}$ $5s^2$ $5p^1$	0.99	1.44
N	[He] $2s^2$ $2p^3$	3.0	0.70

semiconductor materials such as silicon, germanium, the III–V and II–VI compounds, and SiC. Table 1 shows the ground-state configurations of the elements in the nitride semiconductors. The s and p orbitals in the outer electron shells combine to produce hybrid sp^3 orbitals where the probability of finding an electron in a p state is three times as much as finding it in an s state. The directions of greatest electronic charge density in sp^3 orbitals is such that bonding will occur with fourfold coordination, forming a regular tetrahedron about a particular lattice site. The atomic radii can be used to obtain a good approximation of the bond lengths (actual values for the binary compounds are given in Table 2). The chemical properties of the binary nitrides, such as the cohesive energy per bond, increase with the value of the electronegativity of the Group III element (see Tables 1 and 2). The melting characteristics also follow a similar trend (Table 3).

1.3 Crystal Structures

The tetrahedral coordination produces a characteristic atomic array that can be viewed as consisting of hexagonal closed-packed layers consisting of two sublayers; one sublayer is occupied by nitrogen, while the other contains the Group III elements. The cubic zincblende structure [space group $F\bar{4}3m$, T_d^2 (No. 216)] occurs when the hexagonal layers are stacked in a periodic ABCABC sequence along the ⟨111⟩ directions. The hexagonal wurtzite structure (space group $P6_3mc$, C_{6v}^4 [No. 186]) follows an ABABAB arrangement along the [0001] direction. The unit cells and the layer sequences of the two lattices are shown in Fig. 1. The lattice parameters and other properties of the wurtzite structure are given in Table 4. The ideal wurtzite structure consists of regular tetrahedra with a c/a ratio of 1.633. Other lattice structures are also possible; in particular, the rocksalt (NaCl) structure is typically observed in the nitrides under high-pressure conditions.

1.3.1 Boron Nitride Boron nitride exists in the high-density tetrahedrally coordinated forms (zincblende and wurtzite structures), as well as in a low-density hexagonal layered structure. These structures and their properties are often compared with those of carbon. The zincblende form has a similar structure and properties to those of diamond. The low-density, hexagonal form of BN has similar lattice properties to those of graphite. The differences between BN and C originate mostly in the nature of their chemical bonds, mixed co-

Table 2. Chemical properties of the nitride semiconductors. This table shows the experimental cohesive energy per bond (Harrison, 1980), bond lengths of the Group III nitrides (with wurzite structure), the standard molar enthalpy (heat) of formation $\Delta_f H^0$, the Gibbs free energy of formation $\Delta_f G^0$, the standard entropy ΔS^0, and the molar heat capacity C_p at 298.15 K.

	Cohesive energy per bond (eV)	Bond length (Å)	$\Delta_f H^0$ (kJ mol^{-1})	$\Delta_f G^0$ (kJ mol^{-1})	ΔS^0 (J mol^{-1} K^{-1})	C_p (J mol^{-1} K^{-1})
BN	3.34	1.57	−254.4	−228.4	14.8	19.7
AlN	2.88	1.89	−318.0	−287.0	20.2	30.1
GaN	2.24	1.94	−110.5	−118.6	29.7	40.8
InN	1.93	2.15	−138.2	−151.1	43.5	41.7

Table 3. Melting characteristics of the nitrides (wurtzite form). These values are estimated in various ways and have not been measured directly.

	Melting temp., T_M (K)	Vapor pressure at T_M (10^6 Pa)	Melting entropy (J mol^{-1} K^{-1})	Melting enthalpy (kJ mol^{-1})	Reference
BN	3000				
AlN	3487	3	69.53	242.4	Van Vechten, 1973
GaN	2791	200	67.01	187.0	Karpinski *et al.*, 1984
InN	2146	$>10^4$	59.02		Grzegory *et al.*, 1993

valent–ionic in BN versus strongly covalent in diamond. BN has a slightly higher energy band gap and lattice constant, but lower mechanical strength, thermal conductivity, and Debye temperature than diamond. The rocksalt form of BN is stable under ambient conditions. The zincblende form is the most stable under normal conditions, and the wurzite form is metastable under all conditions. Lattice constants and densities for various forms of BN are given in Table 5.

1.3.2 Aluminum Nitride AlN crystallizes in the wurtzite structure under normal conditions, with a c/a ratio reportedly ranging form 1.600 to 1.602; the deviation from the ideal value is attributed to ionicity. A metastable zincblende form has been reported with a lattice parameter of 4.38 Å. A phase transition to rocksalt has been observed at 14 GPa.

1.3.3 Gallium Nitride GaN usually crystallizes in the wurtzite form with a c/a ratio of

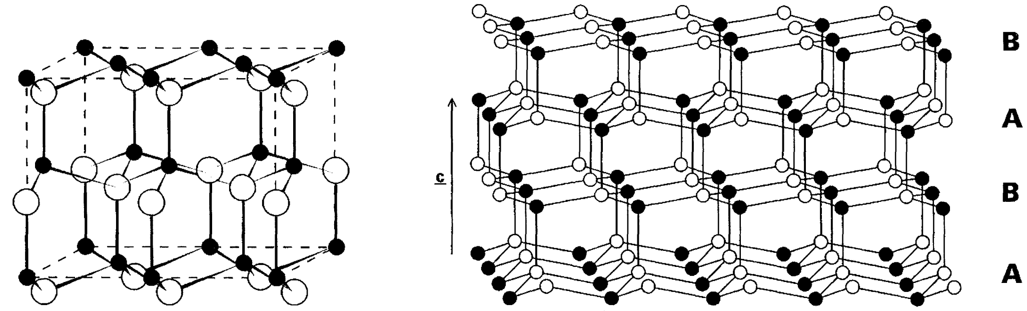

(a) Wurtzite (hexagonal GaN) **(b)**

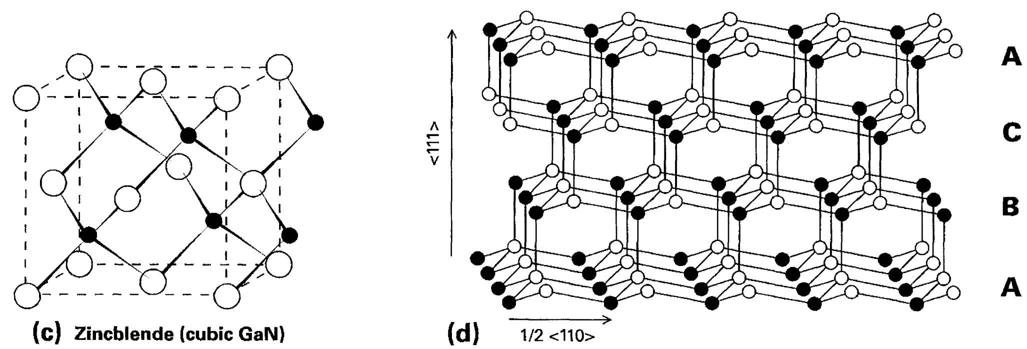

(c) Zincblende (cubic GaN) **(d)**

FIG. 1. Atomic arrangement in tetrahedrally coordinated structures. Unit cells for (a) hexagonal wurtzite and (c) cubic zincblende, and their respective crystalline plane layer sequences (b) ABABAB and (d) ABCABC.

Table 4. Lattice parameters, thermal expansion, and energy gap (D = direct, I = indirect) for the Group III nitrides and substrates. All hexagonal forms of SiC and the AlGaN system belong to the space group $P6_3mc$ (186). The coefficent of linear thermal expansion for the nitrides varies with temperature in a non-linear fashion (for more complete data see, *e.g.*, Madelung, 1996).

Crystalline properties		BN	AlN	GaN	InN	α-Al_2O_3	6H-SiC
Lattice parameter (wurtzite structure) (Å)	a	2.555	3.1114	3.186	3.5446	4.758	3.081
	c	4.2131	4.9792	5.178	5.7034	12.991	15.092
	c/a	1.652	1.600	1.625	1.609	2.730	1.633(×3)
Density (g cm^{-3})		3.49	3.26	6.10	6.99		
Thermal expansion coefficient ($10^{-6}C^{-1}$)	a	2.0	4.2	5.59	5.7	7.5	4.2
	c	2.6	5.3	3.17	3.7	8.5	4.68
Energy gap (at room temperature) for wurtzite (W) and zincblende (Z) forms (eV)	W	5.2 *I*	6.2 *D*	3.39 *D*	1.89 *D*		
	Z	6.4 *I*	5.1 *D*	3.2 *D*	2.2 *D*		
Lattice parameter (zincblende structure) (Å)	a	3.615	4.38	4.52	4.98		

1.6252. A zincblende form is also observed in thin-film form with a lattice parameter of 4.55 Å. Several high-pressure forms have been reported (see Edgar, 1994).

1.3.4 Indium Nitride Small InN crystallites have been produced in the wurtzite and zinblende structures. It has been noted that nitrogen vacancies are common and may be responsible for a large uncertainty in the lattice constant *c*. A transition to rocksalt at about 245 GPa has been experimentally and theoretically reported.

1.4 Band Gap

The nitride semiconductors have a direct band gap. This is the energy difference between the conduction band and the valence band, and it determines the frequency (color) of absorbed and emitted light. A *direct* band-gap transition involves only the emission of a photon, while an *indirect* band-gap transition requires the combination of a photon and a phonon (lattice vibration). As a result, direct band-gap materials are much better light detectors and emitters. Figure 2 illustrates the values of band gap and lattice constant for binary compound semiconductors commonly used in optoelectronic applications. The electronic bandgap in the Group III nitrides extends from 2.0 to 6.4 eV, including most of the visible and ultraviolet portions of the electromagnetic spectrum. The ternary alloy $In_xGa_{1-x}N$ ranges over most of the visible spectrum. This range of band gaps is a very useful property of the nitride semiconductors.

1.5 Mechanical Properties

The elastic constants and moduli are shown in Table 6. These values are generally significantly greater that those of other semiconductors.

Table 5. Lattice constants and densities of boron nitride (adapted from Edgar, 1994).

	Structure	a (Å)	c (Å)	Density (g cm^{-3})
h-BN	Hexagonal	2.5040	6.6612	2.28
r-BN	Rhombohedral	2.507	10.00	2.276
c-BN	Zincblende	3.615		3.48
w-BN	Wurtzite	2.556	4.175	3.47

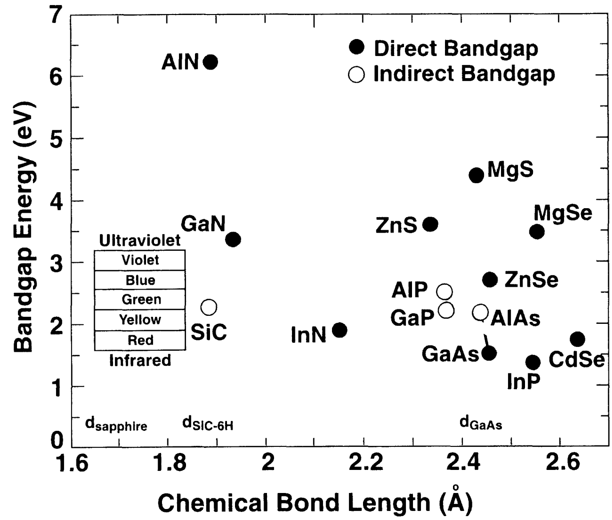

FIG. 2. Band gap *versus* chemical bond lengths of semiconductor compounds that emit in the visible range of the electromagnetic spectrum. The chemical bond length is used instead of the lattice parameter in order to allow comparison between interatomic separations in the zincblende (cubic) structure of the phosphides, arsenides, and II–VI compounds with the wurtzite (hexagonal) structure of the nitrides (Ponce and Bour, 1997).

1.6 Thermal Stability and Etching

The strength of the chemical bond makes these materials thermally stable and mechanically hard. Tables 2 and 7 show the thermal properties of the semiconductor nitrides. With increasing atomic number of the anions, the electronegativity decreases, their ionic nature increases, and the covalent bonding becomes weaker. It is also natural to expect less stability towards oxidation. Even though the heats of formation (such as in $2GaN + 2O_2 \rightarrow Ga_2O_3 + N_2O$) predict strong tendency toward oxidation, no such reactions are observed in practice. This may be due to kinetic barriers introduced by the lattice structure. These materials are also very difficult to etch. No wet etching method is known for GaN. Chemically assisted ion-beam etching (CAIBE) has been successfully used for device fabrication using Cl_2/BCl_3 reactive gas.

2. GROWTH TECHNIQUES

2.1 General

Nitride formation is commonly observed by heating the metal in ammonia at temperatures usually between 500 and 550 °C. The first synthesis of GaN was reported in 1932 by reaction of metallic Ga in an NH_3 stream following the reaction $2Ga + 2NH_3 \rightarrow 2GaN + 3H_2$. Using the same technique, AlN was syn-

Table 6. Mechanical properties of the Group III nitrides. The values for the wurtzite and zincblende structures are shown in the top and bottom (in parenthesis) positions, respectively. The elastic constants (c_{ij}), Poison's ratio (v), Young's Modulus (E) and shear modulus (μ) values are adapted from Edgar (1994).

Properties	c_{11} (GPa)	c_{12} (GPa)	c_{13} (GPa)	c_{33} (GPa)	c_{44} (GPa)	v	E (GPa)	B (GPa)	μ (GPa)
BN	876	119	80	916	346		790±90	410±80	330±40
	(712)	(80)			(334)		(794)	(290)	
AlN	345	125	120	395	118	0	308	160	
GaN	296±18	130±11	158±6	267±18	24±2	0.372	150	195	
	(264)	(153)			(68)	(0.366)		(173)	
InN								165	

Table 7. Thermal properties of the Group III nitrides. Thermal diffusivity is important for heat dissipation in LEDs and LD structures (Lide, 1998).

	Thermal diffusivity ($cm^2\ s^{-1}$)	Thermal conductivity ($W\,cm^{-1}\ K^{-1}$)
BN	0.4	0.8
AlN	0.13	0.3
GaN		1.3
InN		

thesized from Al and InN from $InF_6(NH_4)_3$. Single-crystal nitride semiconductors can be grown in bulk and thin-film forms. Growth of semiconductor-grade material requires the use of special techniques for the control of chemical purity and crystalline quality. An in-depth review of growth techniques and references is given by Strite and Morkoc (1992).

2.2 Bulk Growth

The growth of large single crystals of the nitride semiconductors has been inhibited by the high vapor pressures and extremely high temperatures at the melting point (see Table 3). These characteristics make virtually impossible the growth from stoichiometric melts by conventional methods such as the Bridgman and Czochralski techniques. Further information can be found in Chap. 9 of Pearton (1997).

2.2.1 Aluminum Nitride Bulk single crystals of AlN were first produced by Slack and McNelly (1976) using sublimation techniques in closed tungsten crucibles. Crystals with dimensions of a few millimeters were produced. The technique shows promise for growth of centimeter-size crystals.

2.2.2 Gallium Nitride GaN single crystals have been grown from gallium melt saturated with molecular nitrogen at ≈1600 °C under nitrogen vapor pressures of about 15 kbar. This approach has produced GaN platelets with dimensions of several millimeters and thickness of ≈0.1 mm, and with a typical high *n*-type conductivity due to high levels of impurities (oxygen being the most abundant).

2.2.3 Indium Nitride Attempts to grow bulk crystals of InN have resulted in small crystallites (5–50 μm). Pressures beyond 20 kbar are believed necessary.

2.3 Substrates for Heteroepitaxy

In the absence of bulk nitride single crystals, other materials have been extensively evaluated as substrates for thin-film growth. Two important considerations for the selection of substrates are the lattice characteristics and the thermal/chemical stability of the substrate surface. ZnO and SiC have similar lattice characteristics and thermal expansion rates to those of the nitrides. Both have tetrahedral atomic coordination and exhibit polarity along the *c* axis. Wurzitic ZnO crystals are available, but they are unstable in the highly corrosive environment and high temperatures needed for growth of the nitrides. The Si face of SiC is quite stable at typical growth temperatures. Sapphire (α-Al_2O_3) has hexagonal symmetry and highly chemically stable surface properties. Lattice-mismatch value are given in Table 8.

Table 8. Lattice mismatch (%) and resulting misfit dislocation separation (in Å) corresponding to complete misfit relaxation for basal plane interfaces.

Crystalline properties	Substrate	GaN	AlN	InN	α-Al_2O_3	6H-SiC
Lattice mismatch (%) with	Sapphire	14.8	12.5	25.4	—	11.5
	SiC	3.3	1.0	14.0	−11.5	—
	GaN	—	−2.4	10.6	−14.8	−3.3
Dislocation distance (Å) on	Sapphire	17.2	20.3	10.6	—	21.9
	SiC	80.9	276.7	20.4	21.9	—
	GaN	—	114.4	27.3	17.2	80.9

2.3.1 Silicon Carbide as a Substrate The lattice of SiC bears many similarities to the wurtzite lattice of the nitrides. It is tetrahedrally coordinated and exhibits a polar structure along the *c* axis. The arrangement of the basal planes in SiC occurs in a variety of ways, resulting in a lattice structures with cubic (C), hexagonal (H), and rhombohedral (R) symmetries. The lattice arrangement in SiC is usually described by its symmetry (C, H, or R), and by the number of planes in a period along the direction normal to the basal planes. Thus, the notation 2H refers to the wurtzite structure and 3C refers to the zincblende structure. The 6H form is the most common form that is available commercially. The lattice properties of SiC are listed in Table 4. Note that the lattice constant and the thermal expansion rates along the basal plane between AlN and SiC are very similar. High-quality epitaxy can be routinely achieved using SiC substrates; blue LEDs are commercially available, and blue laser diodes have been demonstrated.

2.3.2 Sapphire as a Substrate α-Al_2O_3 is currently the most commonly used substrate for the epitaxial growth of nitride semiconductor thin films. It has a hexagonal lattice structure with constants given in Table 4. The (0001) basal planes consist of hexagonal close-packed planes of oxygen, alternated with hexagonal arrays of aluminum planes. The Al planes have one third of the sites vacant. The 2/3 occupation of aluminum sites accounts for the Al/O stoichiometry in its chemical formula. The (0001) planes are favored for substrates, although the $\{\bar{2}110\}$ planes have also been recently used for the fabrication of state-of-the-art devices.

2.4 Growth with Molecular Beams

Molecular-beam epitaxy (*q.v.*) (MBE) is performed in ultrahigh-vacuum (UHV) conditions at relatively low temperatures. Gallium beams can be produced from gallium melt effusion cells, and from gaseous trimethylgallium (TMG) and GaCl. Growth under UHV conditions is difficult because of the requirement of large nitrogen equilibrium overpressures. The two most common sources of nitrogen, N_2 and NH_3, have strong bonds and are highly thermally stable. Growth at reasonable rates requires the breaking of the molecular structure to produce more reactive radicals. This can be accomplished at very high temperatures, where growth is difficult to handle. New nitrogen sources and methods to break the molecular structure into more reactive atomic or ionic species at low temperatures ($\approx$600 °C) are of high current interest. A technique that has achieved limited success is gas-source molecular-beam epitaxy (RMBE) using NH_3 or the more reactive hydrazine (N_2H_4) as nitrogen sources. Plasma excitation (such as by electron-cyclotron resonance, ECR) has been used frequently to increase the reactivity of the nitrogen source. However, ECR-plasma sources produce highly energetic particles that result in somewhat poor growth efficiencies and in the formation of subsurface defects due to atomic impact during growth.

2.5 Growth by Vapor-phase Transport

Since 1971, vapor-phase transport [see CVD (CHEMICAL VAPOR DEPOSITION)] has been the most commonly used technique for the growth of Group III nitrides. The hydride vapor-phase epitaxy (HVPE) technique uses NH_3 and GaCl (formed by HCl flowing over a Ga melt) as sources, with purified N_2 as the carrier gas. Growth is typically performed at about 1100 °C. HVPE is characterized by high growth rates and is suitable for the production of thick films (tenths of a millimeter in thickness). Metalorganic chemical vapor deposition (MOCVD) uses NH_3 and metalorganic Group III sources (such as the trimethyl and triethyl forms of Ga, Al and In). The reaction is of the form $(CH_3)_3Ga + NH_3 \rightarrow GaN + 3CH_4$. To date, the best optoelectronic properties have been achieved from materials grown by MOCVD. Low-temperature buffer layers (deposited at about 550 °C) are used to overcome three-dimensional nucleation in the early stages of growth. Epitaxial thin films of AlN, GaN, InN, and their alloys have been grown by this technique at temperatures in the range of 700–1100 °C.

2.6 Doping

As with other semiconductor systems, the electronic properties of the nitrides are very sensitive to doping impurities. Unintentional background impurities resulting from all the growth techniques have a typical *n*-type behavior. The lowest background concentrations currently achieved are around $1 \times 10^{16}\,\mathrm{cm}^{-3}$.

2.6.1 *n*-Type Silicon and germanium are the most commonly used *n*-type dopants. Silane (SiH_4) and germane (GeH_4) are commonly used as doping sources in MOCVD. The Si donor state is located at about 12–22 meV from the conduction band. Electron carrier concentrations up to $10^{19}\,\mathrm{cm}^{-3}$ are obtained by use of silicon as the dopant. Room-temperature electron mobilities as high as $900\,\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$ have been reported.

2.6.2 *p*-Type Zn and Mg are the most common *p*-type dopants for GaN, AlN, and InN. *Bis*-cyclopentadienyl magnesium is commonly used in MOCVD. As-grown GaN doped with Mg is usually highly resistive, but a thermal anneal at about 800 °C activates the *p*-type doping. It is believed that Mg–H complexes are responsible for this behavior. The Mg acceptor state is located at 165–182 meV above the valence band. Room-temperature hole concentrations in the mid-$10^{17}\,\mathrm{cm}^{-3}$ range (corresponding to magnesium concentrations of $\approx 10^{20}\,\mathrm{cm}^{-3}$) are commonly achieved using the thermal activation technique. Room-temperature hole mobility values as high as $20\,\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$ have been observed.

3. MATERIALS PROPERTIES

3.1 Lattice Defects

Lattice defects are closely associated with the properties of semiconductors. Point defects in the nitrides are responsible for the light-emission properties. Dislocations are abundant in current state-of-the-art epitaxy and may strongly affect the distribution of point defects and therefore the optical and electrical properties. The properties of lattice defects in the nitride semiconductors are subject of much current interest.

3.1.1 Point Defects Point defects can be classified as native defects (vacancies, self-interstititals and antisites) and impurities (unintentional impurities and dopants).

3.1.1.1 Native Defects. Theoretical formation energies of various native defects in GaN have been reported by Neugebauer and Van de Walle (1994). The nitrogen vacancy (V_N) is suspected to be abundant in these materials, and manifests itself as a shallow donor, about 0.8 eV above the conduction band edge. The Ga vacancy (V_Ga) is expected to be a shallow triple acceptor. The Ga interstitial (I_Ga) is energetically favorable. The N interstitial (I_N) is expected to form a split-interstitial configuration by joining a host N atom and sharing the same lattice site. Antisite defects can be largely interpreted in terms of the covalent radii of the cation and anion species. A nitrogen on a gallium site (N_Ga) defect should involve N–N bonds similar to those found in the N_2 molecule; and since the covalent radius of N is much smaller than Ga, such defects are under tensile stress with a formation energy of about 6 eV. Similarly, gallium on a nitrogen site ($\mathrm{Ga_N}$) is under compressive stress with a formation energy ranging from 7 to 11 eV, depending on the Fermi level.

3.1.1.2 Impurities. The most common growth-related impurities are hydrogen, oxygen, carbon, and silicon. High concentrations of hydrogen ($\approx 10^{21}\,\mathrm{cm}^{-3}$) are easily incorporated into GaN (during growth as well as during processing). Hydrogen is commonly present in most crystal-growth techniques. Hydrogen can easily form stable complexes resulting from interaction with dangling bonds, and it is expected to easily passivate shallow acceptors and sometimes shallow donors. Hydrogen passivation or compensation of Mg in GaN has been historically an important effect that was overcome by thermal annealing (see Sec. 2.6.2). Carbon is present in the organometallic precursors, and silicon and oxygen originate from the quartz parts in growth equipment. As substitutional impurities, C, Si, and O are single donors. Doping species are discussed in Sec. 2.6.

Table 9. Magnitude of the Burgers vectors for dislocations observed in the nitrides (in Å).

Burgers vectors	GaN	AlN	InN
$a=\frac{1}{3}\langle 11\bar{2}0\rangle$	3.186	3.111	3.545
$c=[0001]$	5.178	4.979	5.703
$c+a=\frac{1}{3}\langle 11\bar{2}3\rangle$	6.080	5.871	6.715

3.1.2 Dislocations Dislocations form following the wurtzite $\{0001\}\langle 11\bar{2}0\rangle$ and the zincblende $\{111\}\langle 110\rangle$ slip systems. Dislocations of the screw, edge and mixed types have been observed in nitrides with the wurtzite structure. Table 9 shows the Burgers vectors and expected relative energies of these dislocations. The large values of the Burgers vectors and the elastic constants (Table 6) imply large strain energies. It has been observed that sometimes screw dislocations with $\mathbf{b}=[0001]$ exhibit hollow cores (nanopipes); this behavior may be due in part to the large dislocation energies.

3.1.3 Inversion Domains Inversion domains are a result of the polar nature of the Group III nitrides and have been observed in GaN thin films.

3.1.4 Stacking Faults Faults in the stacking sequence of the lattice occur during growth by nucleation of a rotated layer, or subsequently by motion of a dislocation along a slip plane. Intrinsic stacking faults involve the absence of a plane in the lattice (ABCAB$\underline{\text{CB}}$CABC in zincblende, and ABAB$\underline{\text{AC}}$ACAC in wurtzite). Extrinsic stacking faults involve the presence of an extra plane (as in ABCABC$\underline{\text{B}}$ABCABC in zincblende and ABABAB$\underline{\text{C}}$ABABAB in wurtzite). In the nitride semiconductors, stacking faults are observed in the early stages of growth, and are commonly observed in low-temperature nucleation layers. Stacking faults are rarely observed in regions more than 1 μm away from the substrate.

3.2 GaN on Silicon Carbide

High-quality nitride films and devices have been reported based on GaN epitaxy on SiC using AlN and GaN buffer layers. As described in Sec. 2.3.1, SiC exhibits many similarities to the nitride semiconductors and serves well as a substrate. Because of the close lattice match and very similar thermal expansion characteristics with AlN, very high crystalline quality growth is possible. The $[0001]_{\mathrm{Si}}$ substrate surface, which has silicon on the top position of the basal plane, gives best results. Figure 3 shows the atomic arrangement at the interface between AlN and SiC. The projection of the 6H structure exhibits the characteristic sixfold zigzag in the stacking sequence of the basal planes, compared with the vertical stacking sequence of the wurtzite (2H) structure. The atomic bonding at the interface is coherent with both lattices, producing a bonding sequence C–Si—C–Si—N–Al—N–Al or C–Si—C–Al—N–Al, and charge balance is possible leading to stable interfacial structures and defect-free epitaxy (Ponce *et al.*, 1996). A current limitation is in the preparation of the substrate, where surface steps cause lateral offsets (due to the zigzag of the 6H structure) in the position of the atomic sites on the basal

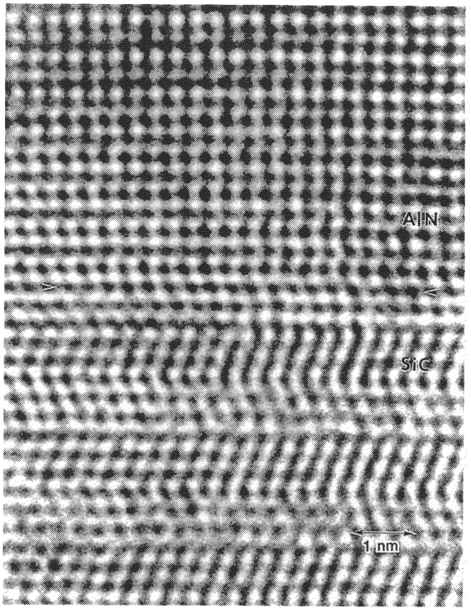

FIG. 3. Transmission electron microscopy lattice image of the AlN/SiC interface. Note the transition between the 6H-SiC and the 2H-AlN as shown by the arrangement of the lattice spots (sixfold zig-zag in the SiC, *versus* vertical columns in the AlN) (Ponce *et al.*, 1996).

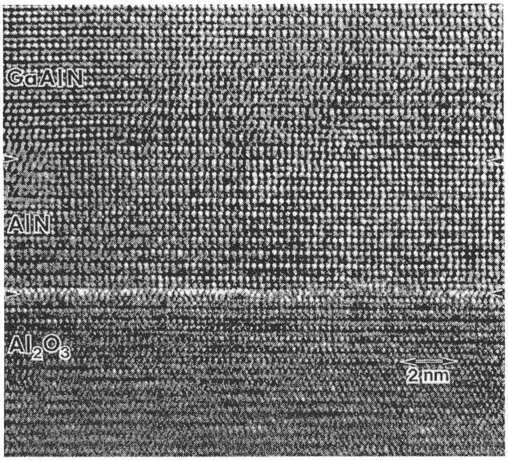

FIG. 4. AlN/sapphire interface (Ponce *et al.*, 1994).

planes, which are difficult to remedy with the 2H wurtzite structure, causing the appearance of complex faults in the epilayer.

3.3 Epilayers on Sapphire

Nitride semiconducting thin films grown on sapphire have been shown to have superior device quality despite the large differences in lattice structure and thermal-expansion characteristics that exist between GaN and α-Al_2O_3.

3.3.1 Microstructure The difference in lattice structures between sapphire and the nitrides is resolved at their interface. Figure 4 shows an image of the AlN/sapphire interface. The interface is planar and abrupt, consisting of a single atomic layer. The lattice mismatch of 12.5% (see Table 8) is accommodated by the atomic arrangement at the interface plane. The buffer layer plays a key role in determining the properties of the film. With an initial amorphous structure, it undergoes solid-phase crystallization as the temperature is increased. The resulting crystallites act as templates for growth, producing a columnar array of crystals with the *c* axis closely aligned along the growth direction. The distribution of the columnar orientation is on the order of 5 arcmin for the tilt of the *c* axis, and 8 arcmin for the twist (rotation of the *c* axis) component shown in Fig. 5. Dislocations exist to maintain the coherency in these very-low-angle tilt boundaries. Figure 6 shows a cross-section transmission electron microscope image of a high-efficiency light-emitting diode (Ponce and Bour, 1997). Dislocations are observed propagating in the growth direction with a density of $\approx 10^{10}\,\mathrm{cm}^{-2}$. Initially, it was surprising that materials with such high dislocation density could be adequate for high–light-emission efficient LEDs. Data from other systems, shown in Fig. 7, indicate that dislocation densities above $10^3\,\mathrm{cm}^{-2}$ have a significantly adverse

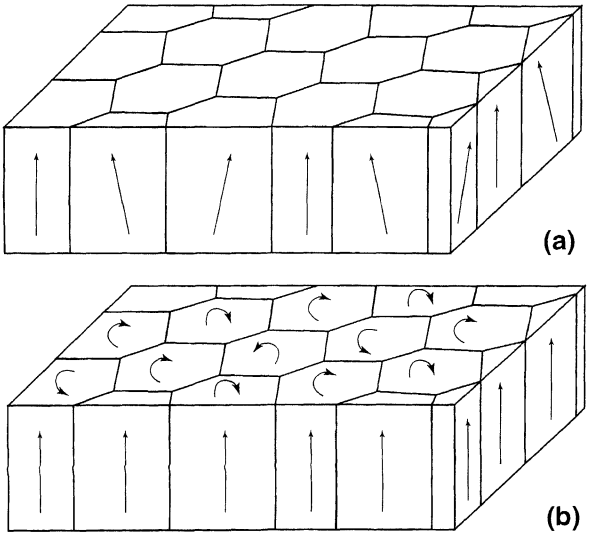

FIG. 5. Columnar structure model for the III–V nitrides. The epilayer consists of a columnar array with a distribution in orientations distinguished by two components: (a) tilt of the *c* axis (typically about 5 arcmin) and (b) rotation about the *c* axis (typically about 8 arcmin).

Table 10. Electrical and optical properties of the nitride semiconductors (Madelung, 1996).

	BN	AlN	GaN	InN
Energy gap (eV)	6.4 *I*	6.02 *D*	3.34 *D*	2.00 *D*
Electron effective mass, m_e	0.752		0.27	
Hole effective mass, m_h			0.8	
Heavy	0.375			
Light	0.150			
Electron mobility ($cm^2 V^{-1} s^{-1}$)			>900	≈250
Hole mobility ($cm^2 V^{-1} s^{-1}$)		≈14	≈10	
Refractive index, n	2.117		2.29	2.56
Dielectric constant $\varepsilon(0)$	7.1	9.14	10.4	
Dielectric constant $\varepsilon(\infty)$	4.5	4.84	5.8	9.3

spond to significant lower mobility values than the best achievable.

3.3.3 Optical Light emission and absorption depend on many factors. One is the band-edge emission of the material, determined by the composition as shown in Fig. 2. Other factors are the nature and concentration of the doping impurities that introduce states in the band gap. Silicon introduces shallow donor states and causes in GaN only a slight reduction in the light energy. Magnesium introduces deep acceptor states and causes in GaN a shift of the emission and absorption peaks towards longer wavelengths. Higher–band-gap AlGaN alloys are used for ultraviolet applications, while InGaN alloys are used for applications involving blue and green. The optical properties of these materials are reviewed in Chap. 5 of Pearton (1997).

4. APPLICATIONS

4.1 General

The nitride semiconductors have found a niche in several important technologies such as short-wavelength light emitters and detectors, high-power and high-temperature electronics, and cold-cathode electron sources. These applications are based on their unique properties that include large chemical-bond energies, large electronic band gap, and low electron affinities.

4.2 Light-Emitting Diodes

Much of the GaN research has been driven by the search for short-wavelength light-emitting diodes. An important milestone was reached by the introduction in 1994, by Nichia Chemical in Japan, of blue LEDs with emissivities above one candela. In order to understand the significant impact of such event it is important to view the progress of visible LEDs depicted in Fig. 8. All commercially available LEDs operating in the infrared to visible range (≈450–1500 nm) are fabricated using III–V compound semiconductors grown by chemical vapor-deposition techniques. The first LEDs were produced by General Electric in 1962, and consisted of $GaAs_{1-x}P_x$ epitaxial layers. The efficiency of LEDs has increased since then by about an order of magnitude per decade. Nitrogen was introduced as a dopant in the mid-1970s in order to enhance the emission efficiency of GaAsP, allowing the production of brighter red, yellow, and green LEDs. The use of heterojunctions and double heterostructures in the 1980s allowed greater confinement of free carriers and increased the efficiency of AlGaAs red LEDs beyond that of filtered incandescent lamps. The use of AlGaInP in the 1990s placed the efficiency of red and orange LEDs above the level of unfiltered incandescent lamps. Bright materials operating in the green to blue range were eagerly sought during this period. ZnSe-based materials were emphasized, but the ease of generation of lattice defects in these II–VI compounds impeded their commercialization. Severe difficulties were encountered in the initial efforts with the nitrides that were overcome in due course by

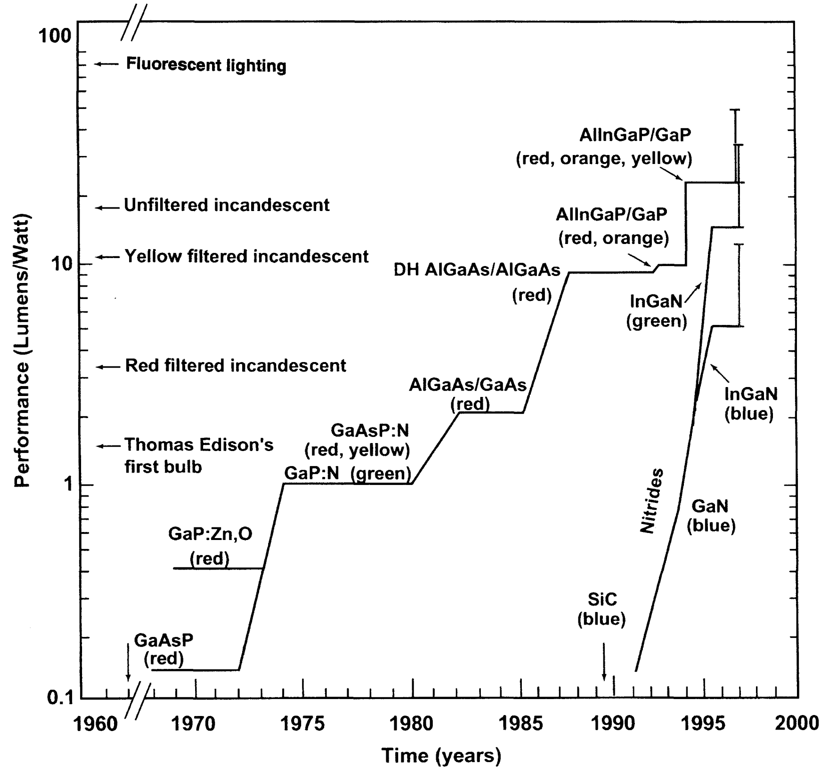

FIG. 8. Performance of light-emitting diodes through the years. The original branch was based on the arsenides and phophides, and started in 1962. Since then the light output (measured in lumens per watt) of devices has increased by about one order of magnitude per 10 years. In the 1990s, such performance surpassed that of the incandescent lamps used for interior lighting. The development of the nitrides advanced rapidly during the mid-1990s to reach similar levels of performance as the phosphides. (Ponce and Bour, 1997.)

key contributions described in Sec. 2.5. Great excitement resulted from the announcement by Nakamura (1994) of high-efficiency blue LEDs based on Group III nitride thin-film technology. Green LEDs were announced shortly after. With the introduction of quantum wells in the active region of the devices, Nakamura was able to gain another order of magnitude in power output. As a result, LEDs covering the full visible spectrum were commercially available by 1996, with efficiencies exceeding those from incandescent tungsten-filament lamps ($>10\,\mathrm{lm\,W^{-1}}$), long lifetimes (estimated at $>50\,000$ h for the nitride-based LEDs), and quick response times, and at competitive prices (Ponce and Bour, 1997). LEDs are currently being introduced into a number of applications; these are described as follows.

4.2.1 Traffic Lights Traffic lights utilize high amounts of energy since they must be turned on continuously. Tungsten-filament incandescent lamps lose a significant amount of power in the filtering process ($\approx 85\%$ for red, $\approx 30\%$ for green and amber). This makes monochromatic LEDs highly attractive. For example, a red (filtered) incandescent lamp using 170 W can be replaced with an LED array using only 19 W. In addition LEDs have longer lifetimes than incandescent lamps (several years versus several months, respectively), reducing replacement costs. Therefore, the use of LEDs causes significant economic savings

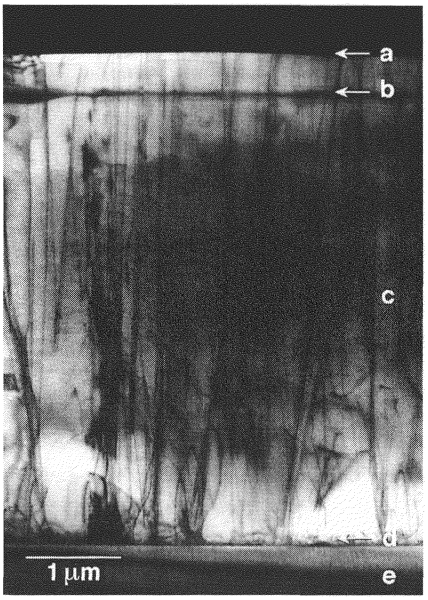

FIG. 6. High-efficiency blue LED viewed by cross-section transmission electron microscopy. A high dislocation density of $\approx 10^{10}\,\text{cm}^{-2}$ is observed. *a*, top surface of device; *b*, InGaN active region of device; *c*, GaN film; *d*, low-temperature buffer layer; and *e*, sapphire substrate. (Ponce and Bour, 1997.)

effect on device performance. The nitride semiconductors exhibit six orders of magnitude higher dislocation densities than other systems used for light emitters. Another interesting feature of GaN epitaxy on sapphire is the relatively low thermal stresses, which are inconsistent with the high growth temperature and the large difference in thermal expansion between the nitrides and sapphire. Dislocations should play a role in the relaxation of such thermal stresses and hence the relation between high dislocation densities and device performance.

3.3.2 Electrical Some electrical and optical properties of the nitrides are summarized in Table 10. As discussed in Sec. 2.6, high carrier densities are possible with use of Si and Mg as *n*-type and *p*-type dopants, respectively. Electron concentrations of $10^{19}\,\text{cm}^{-3}$ and hole concentrations in the mid 10^{17}-cm^{-3} range, at room temperature, are commonly achieved in laboratories today. Carrier mobility values are very sensitive to doping concentrations. Under low background impurity concentrations and low doping concentrations mobility values as high as $900\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$ and $20\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$ have been observed for *n*-type and *p*-type materials, respectively. Low-resistive device materials are achieved by optimizing the product of the carrier concentration and the mobility. Actual values used in devices typically corre-

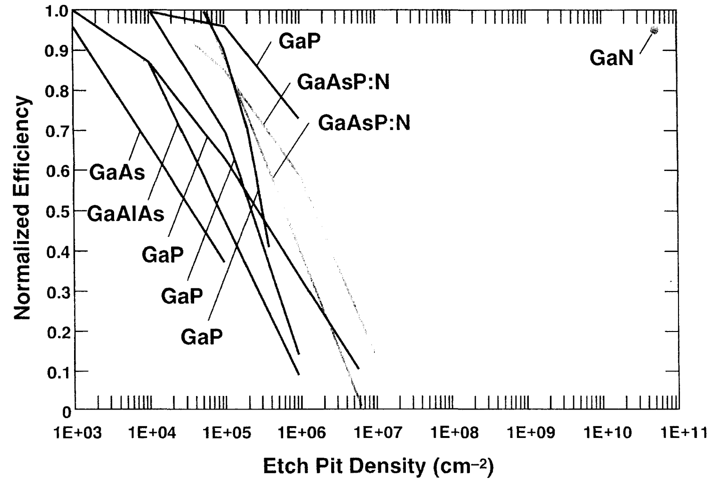

FIG. 7. Relationship between light-emitting efficiency and the dislocation density in commercial LEDs based on compound semiconductors (Lester *et al.*, 1995).

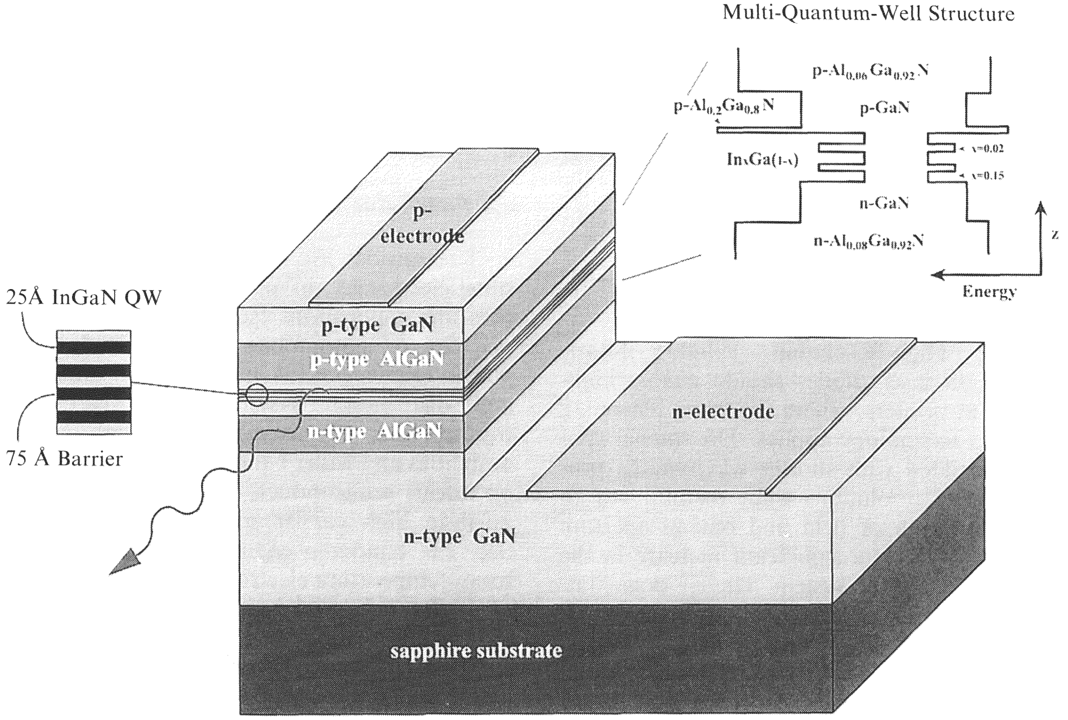
Multi-Quantum-Well Structure
p-Al$_{0.06}$Ga$_{0.92}$N
p-Al$_{0.2}$Ga$_{0.8}$N
p-GaN
In$_x$Ga$_{(1-x)}$
x=0.02
x=0.15
n-GaN
n-Al$_{0.08}$Ga$_{0.92}$N
z
Energy
25Å InGaN QW
75 Å Barrier
p-electrode
p-type GaN
p-type AlGaN
n-type AlGaN
n-electrode
n-type GaN
sapphire substrate

FIG. 9. Structure diagram of the InGaN multiple–quantum-well diode laser reported by Nakamura. The diagram in the inset shows the energy-band diagram of the device structure. (Ponce and Bour, 1997.)

and is bound to have a substantially desirable environmental impact.

4.2.2 Signals Bright outdoor signs, such as for advertising in streets and signaling in highways, require high light emissivity and longevity.

4.2.3 Interior Automobile Lighting Longevity makes LEDs highly desirable for interior automotive lighting and for monochromatic panel indicators.

4.2.4 Full-Color Large-panel Displays Large full-color, outdoor electroluminescent displays are currently available using AlInGaP LEDs for red and InGaP LEDs for green and blue.

4.2.5 Interior Lighting As manufacturing costs drop with increased mass production, considerations of energy savings and longetivity will apply to all types of lighting. This includes the home and the office, where permanent light fixtures could be installed to last the lifetime of the building.

4.3 Diode Lasers

Nakamura (1996) achieved pulsed blue laser operation based on the InGaN multiple-quantum-well structure shown in Fig. 9. Using a similar layered structure, Nakamura achieved continuous-wave operation for more than an extrapolated 10 000 h at room temperature (Ponce *et al.*, 1998). In early 1999 commerical blue-violet laser diodes operating at 400 nm were introduced by Nichia Chemical. Several important applications are envisioned for short-wavelength laser diodes based on nitride semiconductors; among them are the following.

4.3.1 High-Density Data Storage There exists a large market for data storage in the compact disk (CD) industry, which currently operates with 780-nm (near-infrared) diode lasers. The size of the bits of information in the CDs is associated with the optical diffraction limit of the optical lenses used during recording and reading of data. In this limit, the spot size is proportional to the wavelength of light. The use of violet–blue lasers operating at about 410 nm should produce a fourfold increment in the recording density. This is sufficient to produce a feature-length movie on a digital video disk (DVD) using current compression technology.

4.3.2 High-Resolution Printing Short-wavelength laser diodes should enable high-resolution printers, where the same reasoning as in CD technology applies. The smaller spot size associated with shorter wavelengths provides higher resolutions while maintaining an adequate depth of field and optical aperture size, *i.e.*, without a significant increase in the cost of the optical system. Higher power attainable from the nitride semiconductors should allow higher printing speeds. Higher on/off speeds (rise/fall times of ≈1 ns) should enable improved image quality. Such high-resolution raster-output systems would have application for printing directly on photographic film as well as for xerography (using photoconductive plates and charged toner). In the latter, the use of shorter wavelengths should minimize decomposition of commonly used organic photoconductor materials.

4.3.3 Projection Displays These include projection high-definition television where the current liquid-crystal modulation system is replaced by three diode lasers with $\lambda = 655$ (red), 530 (green), and 455 nm (blue). Current technology involves costly optical systems for collimating, splitting, polarizing, and recombining the primary color beams. The optical design is greatly simplified and the costs significantly reduced by the coherent, monochromatic, and polarized nature of the light from the diode lasers. In addition, a broader color gamut can be achieved compared with that of phosphor or of filtered light. Some big challenges lie ahead, however. The most important is perhaps achieving green laser action, which would require the use of InGaN (see following section on challenges).

4.3.4 Underwater Communications Nitride lasers operating in the blue–green range of the visible spectrum are desirable because water has a passband between 450 and 550 nm (blue–green region where transmission of light is maximized), and can be used for optical communications between submarine vehicles.

4.4 Solid-State Electronics

The nitrides are very attractive for solid-state electronics, in particular for high-temperature applications (such as inside nuclear reactors or automobile engines) and high-power electronics (as in power transmission and distribution systems, and microwave transmitters). This is due to the high carrier mobilities and wider band gap of devices based on nitride semiconductors. In addition to the intrinsic high carrier mobilities reported in Sec. 2.6, band-gap engineering has produced room-temperature electron mobilities in excess of $1700\,\mathrm{cm}^2\,\mathrm{V}^{-2}\,\mathrm{s}^{-1}$ using modulation-doped AlGaN/GaN structures. AlN has excellent electronic insulating properties and is used as a gate dielectric in field-effect transistors (FETs). Together with InGaN, high-transconductance FETs have been reported capable of handling large currents. For further information, see Chap. 16 in Pearton, 1997.

4.4.1 High-Temperature Electronics The wide band gap of the nitrides allows for semiconductor behavior at significantly higher temperatures. For instance, an intrinsic carrier concentration of $10^{15}\,\mathrm{cm}^{-3}$ is reached at ≈300 °C in silicon, at ≈500 °C in GaAs and at ≈1300 °C in GaN. Thus, cooling may not be necessary when the nitrides are used, bringing down the costs of fabrication of high-power devices (see e.g. Khan, 1995).

4.4.2 High-Power Electronics Higher breakdown fields are also a result of the wide band gap. For instance, breakdown fields for silicon and GaN are $(\approx 0.2 \text{ and } 3.0) \times 10^6\,\mathrm{V\,cm}^{-1}$, respectively. Applications for GaN power electronics include switching on power grids, automotive (specially the all-electric vehicle), military avionics, and industrial motor drives (see, *e.g.*, Wu, 1996).

4.4.3 Solar-Blind Electronics The ozone layers absorbs all the solar energy in the ultraviolet between 210 and 290 nm. Electronic devices operating in this range of frequencies enjoy the absence of optical ambient noise and are highly desirable for defense applications.

4.5 Cold Cathodes

Aluminum-containing nitrides, such as $Al_xGa_{1-x}N$, are known to have negative electron affinities. This characteristic makes them useful for electron emitters in a number of applications such as vacuum electronics, flat-panel displays, and electron sources for microscopy. The surface properties of the nitrides make them rugged to electron-induced decomposition, and provide the ability to operate in poor vacuum conditions, with longer operating lifetimes and with higher energy efficiency.

4.6 Challenges and Future Developments

It is surprising to see a high optoelectronic performance of the nitrides in spite of their high dislocation densities. Following conventional wisdom, it is believed that reducing the defect densities should produce increasingly superior materials. Currently blue–violet laser diodes are available. True blue laser diodes will require mastering the growth of $In_xGa_{1-x}N$ thin films with $x > 0.10$. Successful development of green laser diodes will require the growth of high indium concentrations of $In_xGa_{1-x}N$. If pure InN films become available ($\lambda \approx 605\,\mathrm{nm}$), a GaN/InN/GaN double heterostructure should provide orange–red diode lasers as well. On the shorter wavelength range, the development of ultraviolet diode lasers from $Al_xGa_{1-x}N$-based heterostructures should further increase optical data storage capacity by approximately another four times ($\lambda \approx 200\,\mathrm{nm}$ for AlN). The use of $B_xAl_{1-x}N$ should establish the maximum limit for data storage density within the nitride semiconductors.

5. SUMMARY

The nitride semiconductors have become important technological materials during the decade of the 1990s. This is based on their unique properties that include light emission in the visible and ultraviolet range, semiconductor properties at elevated temperatures, and low electron affinity. With the ability to produce high-quality semiconducting thin films, a number of applications are becoming commercially feasible. The advent of bright LEDs operating in the green and blue brings to an end the long search for bright solid-state sources covering the full range of the visible spectrum. The nitride-based LEDs have superior energy efficiency when compared with other means of light production, and their long device lifetimes make them economically desirable. Blue and ultraviolet laser diodes promise high-density storage, high-resolution printing, and compact, high-quality projection TV.

GLOSSARY

Band Gap: The minimum energy difference between the conduction band and the valence band. A direct band-gap transition involves only the emission of a photon, while an indirect band-gap transition requires the combination of a photon and a phonon (lattice vibration).

Bridgman Growth Technique: Method for crystal growth from the liquid phase, resembling the static freeze method. The ampoule containing the melt is moved within the furnace, while the freezing-point temperature remains fixed.

Chemical-Assisted Ion-beam Etching (CAIBE): Method to remove materials with use of highly reactive ion beams.

Cohesive Energy: Energy required to separate a solid into isolated atoms.

Cold Cathode: A surface that emits electrons purely by the effect of a voltage gradient, without thermal assistance. This is in contrast to thermionic emission, where the cathode is heated until electrons boil off the surface.

Czochralski Growth Technique: A common method to grow crystals from the melt. A seed is used to trigger growth at the top surface of the melt. With the use of appropriate thermal gradient and by pulling the seed away from the melt, a large boule can be produced. This technique is popular in the growth of silicon crystals for microelectronics.

Direct Band gap: See **Band Gap**.

Double Heterostructure: A sandwich of materials with two heterostructure interfaces. The use of double heteostructure improves carrier and optical confinement.

Electroluminescence: Light emission by direct conversion from electrical energy. An example is a *p–n* junction where electron–hole recombination generates light. See also **Light–Emitting Diode**.

Heterojunction: The junction of two dissimilar semiconductors that differ in their chemical composition and their doping levels.

High-Density Data Storage: Storage of information in the form of bits, using photographically defined structures and a laser probe.

High-Power Electronics: Electronic circuitry that operates at elevated current densities, such as found in electric utilities.

High-Temperature Electronics: Electronic circuitry designed for operation at high temperatures.

Incandescent Lamp: A light source that operates by heating of a filament by means of electronic current.

Indirect Band Gap: See **Band Gap**.

Light-Emitting Diode: A *p–n* junction that emits light when forward biased. The energy of light is typically close to the value of the band gap.

Metalorganic Chemical Vapor Deposition: A technique for growth of thin films by vapor transport, using organic compounds that react at the surface of the growth front.

Nitride Semiconductors: The semiconducting form of the Group III nitrides, including the binary compounds BN, AlN, GaN, and InN, as well as their alloys.

Nucleation Layer: Also referred as initiation layer and buffer layer. Thin film that is deposited at the very beginning of growth, at typically lower temperatures, in order to overcome surface tension and equilibrium nucleation kinetics. After the substrate is covered with the nucleation layer, temperature is raised for growth of the main film.

Sapphire: A crystalline variety of the mineral corundum (α-Al_2O_3), exhibiting a rhombohedral unit cell, with a three-fold axis along the diagonal. Sapphire is hard (9 on the Mohs scale), and highly stable at high temperatures and under corrosive environments. It is a substrate of choice for the growth of nitride semiconductors.

Wurtzite Lattice: Hexagonal crystalline form of zinc sulfide. A common lattice structure observed in most of the Group III nitrides, and in many II–VI compounds such as CdS.

Zincblende Lattice: Cubic form of zinc sulfide. A common lattice structure found in the binary III–V compounds such as GaAs and GaP and in binary II–VI compounds such as CdTe and ZnSe. It is also referred as the sphalerite structure.

Works Cited

Akasaki, I., Amano, H. (1994), *J. Electrochem Soc.* **141**, 2266.

Amano, H., Kito, M., Hiramatsu, K. Akasaki, I. (1989), *Jpn. J. Appl. Phys.* **28**, L2112–L2114.

Amano, H., Sawaki, N., Akasaki, I., Toyoda, Y. (1986), *Appl. Phys. Lett.* **48**, 353–355.

Edgar, J. H. (Ed.) (1994), *Properties of Group III Nitrides*, London: INSPEC.

Grzegory, I., Jun, J., Krukowski, S., Bockowski, M., Wroblewski, W., Porowski, S. (1993), *Physica B* **185**, 99–102.

Harrison, W. A. (1980), *Electronic Structure and the Properties of Solids*, San Francisco: Freeman, pp. 175–176.

Karpinski, J., Jun, J., S. Porowski, S. (1984), *J. Cryst. Growth* **66**, 1–10.

Khan, M. A., Shur, M. S., Kuznia, J. N., Chen, Q., Burm, J., Shaff, W. (1995), *Appl. Phys. Lett.* **66**, 1083–1085.

Kittel, C. (1976), *Introduction to Solid State Physics*, 5th ed., New York: Wiley, p. 100.

Lester, S. D., Ponce F. A., Craford, M. G., Steigerwald, D. A. (1995), *Appl. Phys. Lett.* **66**, 1249.

Lide, D. R. (Ed.) (1998), *CRC Handbook of Chemistry and Physics*, 79th ed., Boca Raton: CRC Press.

Madelung, O. (Ed.) (1996) *Semiconductors—Basic Data.* Berlin: Springer.

Nakamura, S. (1991), *Jpn. J. Appl. Phys.* **30**, L1705–L1707.

Nakamura, S. (1996), in: F. A. Ponce, T. D. Moustakis, I. Akasaki, B. Monemar (Eds.), *III–V Nitrides*, Materials Research Society Symposium Proceedings Vol. 449, Pittsburgh: Materials Research Society, pp. 1135–1142.

Nakamura, S., Mukai, T. Senoh, M. (1994), *Appl. Phys. Lett.* **64**, 1687–1689.

Nakamura, S., Mukai, T., Senoh, M. Iwasa, N. (1992), *Jpn. J. Appl. Phys.* **31**, L139–L142.

Neugebauer, J., Van de Walle C. G. (1994), *Phys. Rev. B* **50**, 8067.

Pankove, J. I., Miller, E. A. & Berkeyheiser, J. E. (1971), *RCA Rev.* **32**, 383–392.

Pearton, S. J. (Ed.) (1997), *GaN and Related Materials*, Amsterdam: Gordon and Breach.

Philips, J. C. (1973), *Bonds and Bands in Semiconductors*, New York: Academic Press, p. 54.

Ponce, F. A., Bour, D. P. (1997), *Nature (London)*, **386**, 351–359.

Ponce, F. A., Major, J. S., Plano, W. E., Welch, D. F. (1994), *Appl. Phys. Lett.* **65**, 2302.

Ponce, F. A., Van de Walle, C. G., Northrup, J. E. (1996), *Phys. Rev. B* **53**, 7473–7478.

Ponce, F. A., DenBaars, S. P., Meyer, B. K., Nakamura, S., Strite, S. (Eds.) (1998), *Nitride Semiconductors*, Materials Research Society Symposium Proceedings Vol. 482, Pittsburgh: Materials Research Society.

Slack, G. A., McNelly, T. F. (1976), J. *Cryst. Growth* **34**, 263–279.

Strite, S., Morkoc, H. (1992), *J. Vac. Sci. Technol. B* **10**, 1237–1266.

Van Vechten, J. A. (1973), *Phys. Rev. B* **7**, 1479–1507.

Wu, Y. F., Keller, B. P., Keller, S., Kapolnek, D., Kozodoy, P., Denbaars, S. P., Mishra, U.K. (1996), *Appl. Phys. Lett.* **69**, 1438–1440.

Further Reading

Gil, B. (Ed.) (1998), *Physics and Applications of Group III Nitride Semiconductor Compounds*, Oxford: Oxford University Press.

Mohammad, S. N., Salvador, A. A., Morkoc, H. (1995), *Proc. IEEE* **83**, 1306–1355.

Morkoc, H., Strite, S., Gao, G. B., Lin, M. E., Sverdlov, B., Burns, M. (1994), *J. Appl. Phys.* **76**, 1363.

Nakamura, S., Fasol, G. (1997), *The Blue Laser Diode*, Berlin: Springer.

Ponce, F. A., Dupuis, R. D., Nakamura, S., Edmond, J. A. (Eds.) (1996), *Gallium Nitride and Related Materials*, Materials Research Society Symposium Proceedings Vol. 395, Pittsburgh: Materials Research Society.

Ponce, F. A., Moustakas, T. D., Akasaki, I., Monemar, B. (Eds.) (1997), *III–V Nitrides*, Materials Research Society Symposium Proceedings Vol. 449, Pittsburgh: Materials Research Society.

SYMMETRIES AND CONSERVATION LAWS

GINO SEGRÈ, *Department of Physics, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.*

Introduction 305
1. **Symmetries of Rigid Bodies** 305
2. **Symmetries of Crystals** 306
3. **Symmetries and the Lorentz Group** 307
4. **Conservation Laws and Noether's Theorem** 309
5. **Discrete Symmetries: *T*, *C*, and *P*** 310
6. **Internal Symmetries** 311
7. **Broken Symmetries and Goldstone Bosons** 313
8. **Gauge Symmetries** 315
9. **The Higgs Mechanism** 317
10. **Experimental Limits on Symmetry Breaking** 319
Glossary 320
Works Cited 321
Further Reading 321

INTRODUCTION

The topic of symmetry is a vast one. Books have been written about each one of the topics in the table of contents, and so our treatment perforce covers only some of the highlights.

Simple notions of symmetry are present in many considerations, but the central role of symmetry was not realized until the 20th century. In fact Wigner repeatedly argued that progress in physics was largely based on the ability to separate a physical problem into analyses of the laws of nature independently of the initial conditions. The latter are arbitrary and complicated while the former are general. Symmetries aid in formulating the regularities in physical laws without reference to the specific dynamics. To quote Gross (1995) in his tribute to Wigner, "Wigner argued that invariance principles provide a structure and coherence to the laws of nature just as the laws of nature provide a structure and coherence to a set of events".

In this article we attempt to provide a birds-eye view of the application of symmetry considerations to problems in atomic and molecular physics, condensed matter physics, particle physics, and field theory. The article is weighted to the last because conservation laws appear naturally as a consequence of continuous symmetries of the Lagrangian and are seen most clearly in this context. The brief description of symmetry in other phenomena is intended as an attempt at completeness.

The central mathematical tool in a discussion of symmetry is group theory. This was also realized by Wigner (1959) soon after the development of quantum mechanics. Much of the advance of physics has been based on the employment of techniques it provides —point groups, Lie groups, or more exotic creations.

1. SYMMETRIES OF RIGID BODIES

The symmetry transformations of a finite system such as an atom or a molecule consist of the set of transformations that leave the body unchanged. All such transformations are combinations of rotations and reflections. Rotations are through definite angles about specified axes and reflections are through a plane.

The usefulness of such an analysis is considerable. Not only does the symmetry restrict the form of the allowed interactions, but it also provides a simple method for finding selection rules of transition-matrix elements.

One must specify the order in which one performs operations since, in general, two operations do not commute: that is, performing a given rotation and then a reflection generally gives a different result from that of performing these operations in the reverse order.

A body is said to have an axis of symmetry of *n*th order if it is left unaltered by a rotation

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

of $2\pi/n$ about a given axis. We denote this operation symbolically by R_n. Repeating this operation n times gives us back the identity transformation, which we denote by E, and so we write

$$R_n{}^n = E. \tag{1}$$

Symmetry transformations of finite bodies must leave at leave at least one point invariant. Symmetry groups having this property are known as *point groups*. Landau and Lifshitz (1981) give the description of these groups. They are classified as $C_n, S_{2n}, C_{nh}, C_{nv}, D_n, D_{nh}$, D_{nd}, the tetrahedron group, the octahedron group and the icosahedron group. The simplest example, C_n, has already been described. C_{nh} and C_{nv} are obtained by adding to the axes in C_n planes of symmetry that respectively are perpendicular to or pass through a symmetry axis. D_n has two symmetry axes, one of nth order and a second one perpendicular to the original axis. D_{nh} and D_{nd} are defined by adding planes of symmetry in addition to the axes of symmetry. S_{2n} is a group with $2n$ elements defined by symmetry transformations about axes defined by combinations of rotations and reflections. Two continuous groups are also possible symmetry groups, namely the full rotation group and the group of rotations about a fixed axis. It is clear that a spherically symmetric body's appearance is unchanged by an arbitrary rotation and a cylindrically symmetric one's by a rotation about the axis of symmetry.

The analysis of the matrix elements of operators between atomic or molecular states is enormously simplified through the classification of the states by their transformation properties under the rotation group—*e.g.*, for atoms the S, P, D, etc. states. To pick a more sophisticated example consider the ammonia molecule, NH_3, which one may picture in terms of the plane defined by the location of the three hydrogen atoms with the nitrogen atom free to oscillate through the plane. The group of transformations that leave the molecule in its equilibrium state invariant is C_{3v}. The analysis proceeds by studying the group properties of C_{3v} on the space of the coordinates of the four atoms. For an excellent pedagogical review of this case see Lax's book (Lax, 1974).

2. SYMMETRIES OF CRYSTALS

The symmetries of atoms and molecules are described by rotations and reflections. At least one point in the body must be left unchanged by the combination of transformations; *e.g.*, a single body cannot be transformed into itself by successive rotations about two nonintersecting axes.

For infinite crystals, there is another type of symmetry transformation, namely translations. This is a vast subject, of which we can do no more than present a few highlights. For good introductions to the subject see the text by Ashcroft and Mermin (1976).

As we already said, symmetry operations are combinations of rotations, reflections, and translations. Rotations about a given axis combined with a translation perpendicular to the axis do not give a new symmetry; they are simply equivalent to the original rotation, but now about an axis parallel to the first one. Similarly reflections through a plane combined with displacements perpendicular to the plane do not give new symmetries.

On the other hand, if we combine a rotation about an axis with a displacement along the same axis we do obtain a new type of symmetry. The axis is known as a screw axis. In an n-fold screw, the system repeats itself after a rotation by $2\pi/n$ and a displacement by some distance d.

Similarly reflection through a plane combined with a displacement parallel to the plane leads to a so-called glide reflection plane symmetry, in which the lattice has symmetry through the simultaneous operations of reflection and translation.

The structure of a crystalline lattice is described most easily making use of the concept of the Bravais lattice, which specifies the the periodic ordering of the units of the lattice. These units may be atoms, molecules or other entities. All we are interested in now is their location. The Bravais lattice is defined as the set of all position vectors $\boldsymbol{R}$ such that

$$\boldsymbol{R} = n_1\boldsymbol{a}_1 + n_2\boldsymbol{a}_2 + n_3\boldsymbol{a}_3 \tag{2}$$

where $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ are three vectors not all lying in the same plane and n_1, n_2, n_3 are allowed to range through all the values of the integers. The vectors $\boldsymbol{a}_i$ are called primitive vectors and are said to span the lattice. In general there are many inequivalent choices of primitive vectors for any given Bravais lattice.

The primitive unit cell is defined as that volume of space which, when translated by the allowed vectors of the Bravais lattice, fills all of space without leaving any voids. For a cubic lattice, *e.g.*, it is simply a cube. Once again there is no unique choice of the primitive cell.

The symmetries of a crystal are usually classified by the simpler set of symmetries of the underlying Bravais lattice. The set of operations that take the Bravais lattice into itself is known as the symmetry group or space group of the Bravais lattice. The space group can be shown to consist of

1. translations through Bravais lattice vectors,
2. operations, *i.e.*, rotations and reflections, that leave a particular point of the lattice fixed, and
3. combinations of the first two.

The subset of the symmetry group that leaves a particular point in the Bravais lattice fixed is known as the point group of the Bravais lattice. There are only seven distinct point groups and hence seven crystal systems. Relaxing the restriction to point groups, one finds 14 distinct symmetry groups and hence 14 independent Bravais lattices. The symmetry group of a crystal structure is still larger, depending on both the symmetry of the object and the symmetry of the underlying Bravais lattice. For crystals of arbitrary shape based on arbitrary Bravais lattices, there are 230 possible space groups. For a discussion of these we refer the reader to the text by Ashcroft and Mermin (1976).

A fundamental concept to which one is led in the analysis of periodic structures is that of the reciprocal lattice. In particular, it is the key tool for systematizing discussions of diffraction scattering. Consider a plane wave $e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$. In general the plane wave will not have the symmetries of the Bravais lattice; *i.e.*, translation through a vector $\boldsymbol{R}$ will not leave the plane wave unchanged; but there is a set of vectors $\boldsymbol{K}$ for which the wave is unchanged by translations through $\boldsymbol{R}$. They satisfy

$$e^{i\boldsymbol{K}\cdot\boldsymbol{R}} = 1. \tag{3}$$

The vectors $\boldsymbol{K}$ form what is called the reciprocal lattice. It is straightforward to show that the reciprocal lattice is itself a Bravais lattice. The primitive cell of the reciprocal lattice is known as the first Brillouin zone. The great usefulness of the reciprocal lattice lies in the fact that, because of the above equation, when rays or neutrons are scattered from a crystal, the scattering is coherent if the momentum transfer is $h\boldsymbol{K}$, with $\boldsymbol{K}$ a reciprocal lattice vector and h equal to Planck's constant. The pattern of bright spots due to the coherent scattering allows us to reconstruct the crystal structure.

An interesting development in crystal symmetry of the past decade is the discovery of *quasicrystals*, an intermediate between crystal structures and glassy structures. The former, as we have already seen, have

1. long-range translational order and
2. long-range order under rotations and reflections.

A glassy structure, by contrast, has none of the above. Quasicrystals have long-range translational and rotational order. However, the translational order is of a subtler kind: it is quasiperiodic rather than periodic. Quasicrystals also may have rotational symmetries that are not allowed for ordinary crystals. Their group-theoretical analysis leads to some interesting new properties (Steinhardt and Ostlund, 1987).

3. SYMMETRIES AND THE LORENTZ GROUP

We now turn to the main subject of this review, namely symmetries and conservation laws. In quantum mechanics an operator A that has no explicit time dependence is said to be conserved if its commutator with the Hamiltonian of the system, H, vanishes:

$$[H, A] = 0. \tag{4}$$

A state is represented by a normalized vector Ψ in Hilbert space, where Hilbert space is a complex vector space (Weinberg, 1995). Observables are represented by Hermitian operators A, linear mappings $\Psi \to A\Psi$ of the Hilbert space into itself. The probability of a measurement yielding the result that a state Ψ is in one of a set of states represented by mutually orthogonal vectors Ψ_n is given by

$$P = (\Psi, \Psi_n)^2. \tag{5}$$

A symmetry operation is one that does not change the results of possible experiments; *i.e.*, if Ψ and Ψ' are two states related to one another by a symmetry transformation, then

$$(\Psi, \Psi_n)^2 = (\Psi', \Psi'_n)^2. \tag{6}$$

What we are saying is that if an observer O makes a measurement and obtains a result, an observer O' related to O by a symmetry transformation will obtain the same result.

In Wigner's (1959) pioneering work in the 1930s, he showed that to satisfy the above equation Ψ had to be related to Ψ' by either a unitary or an antiunitary transformation U. Both take the form

$$U^\dagger = U^{-1} \tag{7}$$

Unitary transformations can be discrete or continuous. We treat discrete symmetry transformations in Sec. 5 and here concentrate on continuous ones. The transformation $U = 1$ is of course just the identity. Continuous transformations that differ infinitesimally from the identity are represented by

$$U = 1 + i\varepsilon t \tag{8}$$

where ε is an infinitesimal real parameter and t is an operator. The fact that U is unitary means that t is Hermitian and therefore a physical observable.

Symmetry transformations forming a group are of particular interest. A group has multiplication and an inverse defined for each element. A set of symmetry operations forming a group corresponds to a set of unitary transformations $U(T)$ acting on the vectors Ψ of the Hilbert space.

Of particular interest in physics are the connected Lie groups. They are continuous groups; any element of the group can be connected to the identity by a path that lies within the group space. This allows us to describe the group elements by infinitesimal transformations, because any finite transformation can be built up by an infinite sequence of infinitesimal transformations. We therefore write

$$U(T(\theta)) = 1 + i\theta_a t_a + \cdots \tag{9}$$

The t_a are the generators of infinitesimal transformations, repeated indices are assumed to be summed over, and we take $a = 1, 2, \ldots, n$. The condition of group multiplication requires the t_a to satisfy a set of consistency conditions, which can be expressed as commutation relations:

$$[t_a, t_b] = iC_{abc}t_c \tag{10}$$

with C_{abc} a set of numbers, the so-called structure constants of the group. This set of commutation relations form what is called a Lie algebra and characterize the group. If all the commutators vanish, or eqivalently all the structure constants are zero, we have what is called an Abelian Lie group. In this case the finite transformations can be trivially built up out of infinitesimal ones and we obtain

$$U(T(\theta)) = \exp(i\theta_a t_a). \tag{11}$$

In general, Lie groups are non-Abelian.

The most important symmetry operations are the rotations, the translations, and the Lorentz boosts. To these we append the discrete reflection transformations. These operations form the so called Poincaré group or inhomogeneous Lorentz group (the homogeneous Lorentz group is the latter minus translations). The coordinate transformation corresponding to such an operation is

$$x'_\nu = \Lambda^\mu_\nu x_\mu + a_\nu. \tag{12}$$

In the above a_ν is the four-vector corresponding to translations and Λ^μ_ν the tensor corresponding to rotations and boosts. The equivalence of inertial frames dictates that the latter must satisfy

$$\eta_{\mu\tau}\Lambda^\mu_\sigma\Lambda^\tau_\rho = \eta_{\sigma\rho} \tag{13}$$

where $\eta_{\mu\nu}$ is the diagonal matrix

$$\eta_{11} = \eta_{22} = \eta_{33} = 1, \quad \eta_{00} = -1. \tag{14}$$

For the identity transformation, we simply have Λ^ν_μ equal to the Kronecker delta. For infinitesimal transformations we have

$$\Lambda^\mu_\nu = \delta^\mu_\nu + \omega^\mu_\nu$$

and

$$U(\omega,\varepsilon) = 1 + \frac{1}{2}i\omega_{\mu\nu}J^{\mu\nu} - i\varepsilon_\mu P^\mu. \tag{16}$$

The four components of P_μ, the generator of infinitesimal translations, are respectively P_0, the Hamiltonian or energy of the system, and P_i, the three-vector momentum (the latin indices take on values $1,2,3$ as opposed to greek indices which take values $0,1,2,3$). $J_{\mu\nu}$ is an antisymmetric tensor with six independent components: J_{23}, J_{31}, J_{12} are the generators of infinitesimal rotations, labelled as $J_i = \varepsilon_{ijk}J_{jk}$. The J_{i0} are the generators of Lorentz boosts, K_i.

The commutators, which specify the algebra of the group, are

$$[J_i, J_j] = -[K_i, K_j] = i\varepsilon_{ijk}J_j, \tag{17}$$

$$[J_i, K_j] = i\varepsilon_{ijk}K_k, \tag{18}$$

$$[J_i, P_j] = i\varepsilon_{ijk}P_k, \tag{19}$$

$$[K_i, P_j] = iH\delta_{ij}, \tag{20}$$

$$[K_i, H] = iP_i, \tag{21}$$

$$[J_i, H] = [P_i, H] = [H, H] = 0. \tag{22}$$

The above imply that $\boldsymbol{J}$, $\boldsymbol{K}$, $\boldsymbol{P}$, whose Cartesian components are J_i, K_i, P_i, are all vectors under ordinary spatial rotations and that J_i, P_i, H are all conserved quantities, *i.e.*, constants in time. To rephrase these results in a more intuitive manner, note that if H is invariant under a continuous symmetry U, we have $UHU^{-1} = H$. This implies that the generator of infinitesimal transformations commutes with the Hamiltonian. The last of the array of equations above simply restates then that the Hamiltonian is invariant under rotations and space-time translations.

4. CONSERVATION LAWS AND NOETHER'S THEOREM

Classical mechanics can be formulated by the principle of least action. This states that for all paths that go from a fixed q_1 at time t_1 to a fixed q_2 at time t_2, the physical trajectory corresponds to a stationary value of the action

$$I = \int_{t_1}^{t_2} \mathrm{d}t\, L(q(t), \dot{q}(t)), \tag{23}$$

where $L(q,\dot{q})$ is the Lagrangian of the system. The stationarity of the action implies in turn Lagrange's equations

$$\frac{\partial L}{\partial q} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}} = 0. \tag{24}$$

For the simplest cases L is the difference between the kinetic and the potential energy of the system and is unchanged by the addition of a total time derivative.

An alternative formulation of mechanics makes use of the Hamiltonian, which is simply related to the Lagrangian by

$$H = p\dot{q} - L, \tag{25}$$

with p the momentum conjugate to q, defined as

$$p = \frac{\partial L}{\partial \dot{q}}. \tag{26}$$

Hamilton's equations can be shown to follow from the principle of least action (Itzykson and Zuber, 1980). For a system described by N independent coordinates and hence $2N$ independent variables, we extend the principle straightforwardly and obtain N independent Lagrange's equations.

Let us now consider continuous infinitesimal changes of coordinates $q_i \rightarrow q_i + \varepsilon\Delta q_i$ where ε is an infinitesimal parameter and Δq_i is a deformation of q_i. Such transformations are symmetry operations if they leave the equations of motion unchanged. This is ensured if L is invariant or, more generally, only changes by a total time derivative (in both cases the action is unchanged):

$$\begin{aligned}\Delta L &= \frac{\partial L}{\partial q_i}\Delta q_i + \frac{\partial L}{\partial \dot{q}_i}\Delta \dot{q}_i + \frac{\mathrm{d}C}{\mathrm{d}t}\\ &= \frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{\partial L}{\partial \dot{q}_i}\Delta q_i + C\right] + \Delta q_i\left[\frac{\partial L}{\partial q_i} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}\right].\end{aligned} \tag{27}$$

Repeated indices (in this case $i = 1, 2, \ldots, N$) are assumed to be summed over. The second term in the above equation vanishes because of

Lagrange's equations. We see that a continuous symmetry of L, which implies $\Delta L = 0$, leads to the existence of a conserved quantity, namely a quantity whose time derivative is zero,

$$Q = \frac{\partial L}{\partial q_i} \Delta q_i + C = p_i \Delta q_i + C. \qquad (28)$$

This is Noether's theorem.

For systems with an infinite number of degrees of freedom, we introduce the concept of a field $\phi(x)$ where x is a continuous variable denoting time and spatial locations. Again this is readily generalizable to N independent fields. We introduce a Lagrangian density $\mathscr{L}$ and the action is now

$$I = \int \mathrm{d}^4 x \mathscr{L}(x). \qquad (29)$$

The Lagrange equations for the fields take the form

$$\frac{\partial \mathscr{L}}{\partial \phi(x)} - \partial_\mu \frac{\partial \mathscr{L}}{\partial [\partial_\mu (\phi(x))]} = 0 \qquad (30)$$

The least-action principle once again leads to conservation laws. Consider, for instance, the effect of translation of the coordinates by $x \to x + a$. The change in the Lagrange density is

$$\mathscr{L}[x + a] = \mathscr{L}[\phi(x + a), \partial_\mu \phi(x + a)]. \qquad (31)$$

For an infinitesimal displacement, the variation of the action is given by

$$\delta I = \int \mathrm{d}^4 x \left[\partial_\nu \mathscr{L} - \partial_\mu \left[\frac{\partial \mathscr{L}}{\partial (\partial_\mu \phi)} \partial_\nu \phi(x) \right] \right] \delta a^\nu = 0 \qquad (32)$$

From the vanishing of δI for arbitrary δa^μ we deduce the existence of a tensor $T^{\mu\nu}$ satisfying the conservation law

$$\partial_\mu T^{\mu\nu} = 0 \qquad (33)$$

where the tensor, known as the energy-momentum tensor, is defined as

$$T^{\mu\nu} = \frac{\partial \mathscr{L}}{\partial (\partial_\mu \phi)} \partial_\nu \phi - \eta_{\mu\nu} \mathscr{L} \qquad (34)$$

The $\eta_{\mu\nu}\mathscr{L}$ term in the above equation is the analogue of the $\mathrm{d}C/\mathrm{d}t$ term defined earlier. From the form of the Lagrangian density, it usually follows that the energy-momentum tensor is explicitly symmetric in μ and ν. If this is not the case, one defines a modified energy-momentum tensor, which has the desired symmetry. Since $T^{\mu\nu}$ is conserved, we can also define four conserved charges

$$P^\nu = \int \mathrm{d}^3 x T^{0\nu}, \qquad (35)$$

which are time independent if the fields vanish sufficiently fast at spatial infinity so that

$$\int \mathrm{d}^3 x \partial_i T^{i\nu} = 0. \qquad (36)$$

As before latin indices $i = 1, 2, 3$ refer only to spatial components. This implies that

$$\partial_0 P^\nu = \int \mathrm{d}^3 x \partial_0 T^{0\nu} = \int \mathrm{d}^3 x \partial_\mu T^{\mu\nu} = 0. \qquad (37)$$

The conserved quantities P^ν we have just defined are of course the four components of a four-vector and are simply the energy and momentum of the system as described in the previous section. They are also the generators of infinitesimal time and space translations.

So far we have discussed space-time symmetries. As we shall see shortly, Noether's theorem is also useful for internal symmetries.

5. DISCRETE SYMMETRIES: *T*, *C* AND *P*

Discrete symmetries play an important role in field theory. Space reflection corresponds to the improper Lorentz transformations, which reverse the signs of coordinates leaving time unchanged—*i.e.*, under space reflection

$$\mathbf{x}' = -\mathbf{x}, \quad t' = t; \qquad (38)$$

and, under time reflection, $t' = -t$ while spatial coordinates are unchanged. Under a parity transformation scalars and axial vectors are invariant while vectors change sign. A pseudoscalar, which may be thought of as the invariant product of a vector and an axial vector, is of course an invariant under rotations, but changes sign under parity transformations.

Charge conjugation C relates particles with equal masses and spins, but opposite charges, *i.e.*, it relates particles to antiparticles. As an example, it changes electrons into positrons, protons into antiprotons, and π^+ into π^-. Some particles are their own antiparticles. The photon is one such example: charge-conjugation invariance of the theory implies that the photon changes under C into minus itself since electrons change into positrons; these couple to photons as electrons do, but with the opposite sign.

In a second-quantized formalism the Lagrangian is written in terms of fields, which are themselves linear combinations of creation and annihilation operators. For a charged scalar field $a(k)$ and $b(k)$ are the annihilation operators for quanta of opposite charges, so that, *e.g.*, $a(k)$ annihilates quanta of charge $+1$ while $b^\dagger(k)$ creates quanta of charge -1. In both cases the charge of the system is increased by the action of the operator. One of the types of particles associated with the quanta is called particle and the other antiparticle. As mentioned, some particles, such as the photon, may be their own antiparticles.

There exists a unitary operator $\mathscr{C}$ which takes the fields for particles into those for antiparticles. This is the field-theoretical version of charge conjugation. For instance, the photon field $A_\mu(x)$ transforms into minus itself, *i.e.*,

$$\mathscr{C} A_\mu(x) \mathscr{C}^\dagger = -A_\mu(x). \tag{39}$$

For the charged scalar field

$$\mathscr{C} a(k) \mathscr{C}^\dagger = b(k) \tag{40}$$

In a classic series of experiments in the 1950s, following the suggestion of Lee and Yang (1956), conservation of parity was found to be maximally violated. It was believed at the time that CP, the combined operation of parity and charge reflection, was a good symmetry, but this was disproved by the observation in 1964 of the long-lived kaon decaying into two pions (Christenson *et al.*, 1964).

It is still believed, however, that though T, C, and P are individually not conserved, invariance under the combined operation of TCP holds. The validity of this belief, derived from the existence of local relativistic quantum field theory (Streater and Wightman, 1968), ensures the equality of masses and lifetimes of particles and antiparticles. We discuss tests of these invariances in the last section of the article.

6. INTERNAL SYMMETRIES

We have so far considered symmetries associated with Lorentz transformations. There are also many internal symmetries of systems, generally associated with new quantum numbers. An example is isospin, the symmetry that relates protons to neutrons. By analogy with an ordinary spin-$\frac{1}{2}$ field for which the $S_3 = \pm\frac{1}{2}$ components are related by spatial rotations, we imagine the proton and neutron to be an isospin $I_3 = \pm\frac{1}{2}$ doublet.

The internal symmetries we consider in this section are global symmetries as distiguished from local symmetries; in the former the transformations are independent of space-time, *i.e.*, all protons in a nucleus are rotated by the same amount when we apply an isospin rotation to the state of a nucleus. For local or gauge symmetries, on the other hand, the transformation itself is a function of space-time, *i.e.*, protons in different locations would be rotated by different amounts.

There is another important difference between global and local symmetries: the former are generally broken; in fact many theorists believe there are no exact global symmetries. On the other hand local or gauge symmetries are believed to be conserved in the absence of spontaneous symmetry breaking (this occurs when the Lagrangian is symmetric, but the ground state, or vacuum, of the system is not symmetric). Local symmetries are discussed in Sec. 8.

Taking for granted that the symmetries (from now on in this section we refer to global symmetries) are broken, they may nevertheless be very useful to study, particularly if the breaking is small. Isospin symmetry is a good example; the strong interactions that bind nucleons in a nucleus are believed to be approximately symmetric under an isospin rotation. Electromagnetism obviously distinguishes between protons and neutrons since the former have electric charge and the latter do not. On the other hand electromagnetism usually leads to only small perturbations in nuclei and, after all, the proton and neutron masses only differ by one part in two thousand; isospin is a useful symmetry for nuclei.

In the early 1960s the concept of isospin was extended by Gell-Mann and Neeman (Cheng and Li, 1984) to unitary symmetry, or SU(3), with spectacular success. In SU(3) strangeness, an additional quantum number of hadrons that is conserved in the strong interactions was unified with isospin. The baryons, the pseudoscalar mesons, and vector mesons were found to lie in octet representations, and the baryon resonances in a decuplet representation; and a variety of predictions regarding masses, mixings and decays were found to be true with remarkable precision. Once again symmetry proved a powerful tool.

For the above reasons it is useful to begin by treating these symmetries as if they were exact and look at Noether's theorem (see Sec. 5). We also saw in Sec. 4 how continuous Lie groups were defined. Let us begin to apply some of these notions.

Consider a set of N real scalar fields $\phi_1 \cdots \phi_N$, which we will label as ϕ_i with i running from 1 to N. We assume that the fields belong to a representation of a group G; that is, the fields may be thought of belonging to an N-dimensional vector space acted on by the rotations in G. Calling the Lagrangian density $\mathscr{L}$ as before and assuming that it is a function of ϕ_i and its derivatives, we consider variations of the fields

$$\delta\phi_i(x) = i\varepsilon_a(t_a)_{ij}\phi_j(x) \tag{41}$$

where ε_a is taken to be an infinitesimal quantity assumed independent of space and time. As before, repeated indices are summed over.

If the Lagrangian density is invariant under the rotation in question we have, using Lagrange's equations,

$$\delta\mathscr{L} = \partial_\mu \frac{\delta\mathscr{L}}{\delta(\partial_\mu\phi_i)}\delta\phi_i + \frac{\delta\mathscr{L}}{\delta(\partial_\mu\phi_i)}\partial_\mu(\delta\phi_i), \tag{42}$$

which, on substituting the value for $\delta\phi_i$, implies

$$\delta\mathscr{L} = \varepsilon_a\partial_\mu\left[\frac{\delta\mathscr{L}}{\delta(\partial_\mu\phi_i)}i(t_a)_{ij}\phi_j\right]. \tag{43}$$

Since the Lagrangian is invariant and the above vanishes for any ε_a, we see that the quantity in the square brackets is a conserved current

$$J_{\mu,a} = -i\frac{\delta\mathscr{L}}{\delta(\partial^\mu\phi_i)}(t_a)_{ij}\phi_j. \tag{44}$$

The charges, defined by

$$Q_a = \int d^3x J_{0,a}(x), \tag{45}$$

can also be written, using the definition of conjugate momentum,

$$Q_a = -i\int d^3x \pi_i(t_a)_{ij}\phi_j. \tag{46}$$

Using the equal-time commutation relations between fields and their conjugate momenta, one can easily prove that the charges satisfy in turn the same commutation relations as the generators of the group G, *i.e.*,

$$[Q_a(t), Q_b(t)] = iC_{abc}Q_c(t). \tag{47}$$

The charges are all evaluated at the same time since the commutation relations of the fields and their conjugate momenta are specified at equal times. These relations are called, for obvious reasons, the algebra of charges.

Interestingly enough, since the commutation relations of fields and conjugate momenta hold in any case, these relations hold even if the symmetry is broken. This observation was the starting point for the major development of current algebra in the 1960s (Adler and Dashen, 1968).

This subject is a broad one in itself, but at its roots there is the important identification of the currents associated with the generators of the symmetry with the currents observed in the weak interactions. By this we mean the following: the form of the weak interactions, originally proposed by Fermi in 1934, with suitable modifications, was seen to be an adequate phenomenological description of the weak interactions. It states that the effective

Hamiltonian is

$$H_{weak} = G_F J_\mu J^\mu, \tag{48}$$

where $J_\mu(x)$ is a charged current of the $V - A$ form, namely with equal vector and axial vector parts (Commins and Bucksbaum, 1983).

The initial step was to identify the vector part of this current with the current of the isospin-raising generator; this has many important physical consequences. It not only dictates that the weak current must connect any two particles that lie in an isospin multiplet and differ by one unit of charge (*e.g.*, π^+, π^0), but also specifies the relative normalization of the matrix elements.

This hypothesis, the so called conserved vector-current hypothesis, was confirmed experimentally and extended to an interpretation of the strangeness-changing currents as the currents in the generators of the strangeness operators in the algebra of SU(3). It also became natural then to look for a group-theoretical or symmetry counterpart of the axial currents in the weak interactions.

The algebra of SU(2), with charges Q_a, was extended to an algebra of SU(2) × SU(2) by the inclusion of pseudoscalar charges Q_a^5 with commutation relations

$$[Q_a^5, Q_b^5] = i\varepsilon_{abc}, Q_c \tag{49}$$

$$[Q_a, Q_b^5] = i\varepsilon_{abc} Q_c^5, \tag{50}$$

$$[Q_a, Q_b] = i\varepsilon_{abc} Q_c. \tag{51}$$

This of course can then be generalized to SU(3) × SU(3). This is the historical path. The modern point of view relies heavily on these notions, but incorporates in a central way the notion of spontaneous symmetry breaking, which we turn to next.

7. BROKEN SYMMETRIES AND GOLDSTONE BOSONS

Symmetries of the Lagrangian may be broken explicitly by the introduction of noninvariant terms. More subtly, the Lagrangian may be invariant, but not the physical vacuum or ground state. We distinguish the physical vacuum from what we call the bare vacuum: the former is the true ground state of the system while the latter is the state with no particles or excitations. The physical or true vacuum and the bare vacuum may not coincide, as is the case in, *e.g.*, superconductivity, where the true vacuum contains a superposition of Cooper pairs.

If a continuous symmetry of the Lagrangian is not a symmetry of the physical vacuum, we say that the symmetry is spontaneously broken. To refer back to the case of superconductivity, the Lagrangian has a U(1) symmetry associated with gauge transformations of electromagnetism. The vacuum, on the other hand, does not possess this symmetry since it contains electron pairs.

In general the spontaneous breaking of a continuous symmetry leads to the the existence of massless scalar bosons, usually called Nambu–Goldstone bosons or simply Goldstone bosons (Weinberg, 1995). To see how such a phenomenon arises we can begin by observing that a global rotation of the vacuum (of course here and in what follows we mean the physical vacuum or ground state) costs no energy, or, in other words, vacua differing only by a global rotation are degenerate in energy. Vacua differing by a local rotation differ in energy by the shear energy associated with the deformation of the vacuum. The energy of the shear goes to zero as the wave number goes to zero. In a quantum field theory, there exist states, connected to the physical vacuum by these local rotations, that are degenerate in energy and are populated by quanta whose energy goes to zero as their momentum goes to zero. They are therefore zero-mass bosons, referred to usually as Goldstone bosons. An important caveat, which we discuss in more detail later, is that these massless quanta are not observable in all cases. The number of types or species of such bosons is related to the degree of reduction of the symmetry, *i.e.*, how much less symmetry the vacuum has than the Lagrangian. An explicit field-theoretical example may be helpful.

Consider a theory of scalar fields invariant under the three-dimensional rotation group O(3), now, however, an internal symmetry group. Let the scalars be in the vector, or three-dimensional, representation of the group. The scalars are described by fields $\phi_a(x)$ with

$a = 1, 2, 3$, and an effective potential

$$V(\phi) = -\mu^2\phi^2/2 + \lambda\phi^4/4 \tag{52}$$

with μ a mass and λ a dimensionless coupling constant. As shorthand we have used ϕ^4 equal to the square of $\phi^2 = \phi_a\phi_a$.

The fields ϕ cannot have the usual interpretation as quantum fields since the mass term is negative and hence unphysical. Corresponding to this, the minimum of the potential is not at $\phi^2 = 0$, but rather at $\phi^2 = \mu^2/\lambda = v^2$. Normally in quantum field theory one expands about the minimum of the potential in fields that are described by excitations about the minimum. In this case the minimum of the potential corresponds to a value of ϕ with modulus equal to v, and so the quantum excitations are about this point and not about $\phi = 0$. We say then that in the true vacuum ϕ takes on a nonzero value and the expansion in quantum excitations needs to be made about this value. Pick then a direction in O(3), say the third direction, and let

$$\tilde{\phi}_a = \phi_a - (0, 0, v). \tag{53}$$

The vacuum expectation value (vev) of the field $\tilde{\phi}_a$ is zero while the vev of ϕ_a is

$$\langle 0|\phi_a|0\rangle = (0, 0, v). \tag{54}$$

The fields $\tilde{\phi}_a$ have zero vev and hence may be described in terms of the usual creation and annihilation operators, *i.e.*, they have a standard quantum interpretation. Let us now examine the potential given above as a function of the physical fields. We find, after some algebra, that

$$V(\tilde{\phi}_a) = \mu^2\tilde{\phi}_3^2 + 2v\tilde{\phi}_3\tilde{\phi}^2 + \lambda\tilde{\phi}^4/4 \tag{55}$$

where we have dropped an overall constant and $\tilde{\phi}^2 = \tilde{\phi}_a\tilde{\phi}_a$.

Looking at the quantum fields, $\tilde{\phi}_a$, we see that the first two components are massless while the third has a positive mass. Let us recapitulate what we have found, suitably phrased so that we may generalize our conclusions. We started with a Lagrangian which was explicitly O(3) invariant, but the vacuum was not invariant, as we saw when the field ϕ_3 was shown to have a nonzero vev. This breaking selected a preferred direction in O(3) space which we took to be the third direction, but which was chosen arbitrarily. In terms of canonical or physical fields, we found one massive field, $\tilde{\phi}_3$, and two massless scalar fields, namely the other two components of $\tilde{\phi}_a$.

The symmetry of the Lagrangian, which was O(3), was reduced to rotations about the third axis, so that the symmetry associated with rotations about the 1 and 2 axes was lost. At the same time we saw the appearance of two massless scalar fields. We studied a particular example, but the situation is general. The massless scalars, also known as Goldstone bosons or Goldstone modes, are a general feature of spontaneously broken symmetries. They appear in particle physics, but they also appear in a variety of condensed matter physics problems, *e.g.*, as spin waves in antiferromagnets (Fradkin, 1991).

SU(2) × SU(2) and even SU(3) × SU(3) appear to be good symmetries of the strong interactions. Why is this so? In a simple model of a free proton and neutron, the breaking of vector SU(2) is proportional to the proton-neutron mass difference:

$$\begin{aligned}[Q_a, H] &= i\int d^3x\partial_0 J_a^0 \\ &= \int d^3x\partial_\mu J_a^\mu \sim (M_p - M_n);\end{aligned} \tag{56}$$

but the breaking of the axial SU(2) is proportional to the sum of the neutron and proton masses and hence is apparently not small (Cheng and Li, 1984); however, the symmetry may be spontaneously broken. This requires a triplet of massless pseudoscalars (pseudoscalars rather than scalars since the symmetry is an axial symmetry); they are the Goldstone bosons of the theory. In this case we identify them with the pions. A theory with nucleons and pions can have SU(2) axial symmetry if either nucleons or pions are massless.

One way to understand the smallness of the symmetry breaking is to begin a discussion of the quark model as proposed by Gell-Mann and Zweig (Cheng and Li, 1984). Just as the constituents of nuclei are taken to be protons and neutrons, a doublet or two-component representation of SU(2), so we take the fundamental constituents of neutrons, protons, and other elementary particles to be quarks, the triplet or three-dimensional representation of SU(3).

The three species of quarks are known as the up, the down and the strange quarks or the u, d and s quarks, often written as q_i with $i = 1, 2, 3$. Mesons are bound states of a quark and an antiquark, while baryons are bound states of three quarks. The strong interaction that binds these quarks is taken to be invariant under SU(3) × SU(3) (more about this in the next section) as is the kinetic energy term, so that the only breaking of the symmetry in the Lagrangian, or equivalently the Hamiltonian, is due to the quark masses:

$$\mathscr{L} = \mathscr{L}_{symm} + \mathscr{L}_{mass} = \mathscr{L}_{symm} + m_i \bar{q}_i q_i \tag{57}$$

The success of the quark model has been amply confirmed with the caveat that a free quark has never been seen (more about this in the next section). The values of the quark masses have been found to be

$$m_{\mathrm{u}} \sim 4\,\mathrm{MeV}, \quad m_{\mathrm{d}} \sim 7\,\mathrm{MeV}, \quad m_{\mathrm{s}} \sim 130\,\mathrm{MeV}; \tag{58}$$

so one sees why SU(2) × SU(2) is such a good symmetry [and even why SU(3) and SU(3) × SU(3) are relatively good symmetries]. In the limit of $m_{\mathrm{u}} = m_{\mathrm{d}} = 0$, SU(2) × SU(2) is an exact symmetry of the Lagrangian, broken spontaneously by the vacuum with the appearance of a massless triplet of Goldstone bosons, the pions. In reality the pions have small masses, proportional to m_{u}, m_{d}. It is the smallness of the pion mass which is the true measure of the symmetry breaking.

Of course another puzzle then rears its head, namely how does one build a 940 MeV object out of three objects with masses in the 5–10 MeV range? There is no simple answer to this: quark dynamics create so-called constituent quarks of approximately 300 MeV, while the current-algebra quarks, as described by $\mathscr{L}$, have much smaller values of mass. A major effort to unravel these questions has been embarked on by studies of quarks on lattices (Creutz, 1983).

8. GAUGE SYMMETRIES

So far we have considered rotations of fields by parameters θ_a that are independent of space-time. Can we generalize this notion? The problem is seen immediately by looking at the free Lagrangian of a spin-$\frac{1}{2}$ particle,

$$\mathscr{L} = \bar{\psi}(x)[i\gamma^\mu \partial_\mu - m]\psi(x), \tag{59}$$

which is invariant under the transformation $\psi(x) \to e^{i\theta(x)}\psi(x)$ only for constant θ. If the rotation is space-time dependent, the derivative in $\mathscr{L}$ will introduce extra terms. If we consider not a free theory, but one with minimal coupling to a spin-1 gauge boson, the Lagrangian is invariant under space-time dependent transformations. Minimal coupling means that

$$\partial_\mu \to \partial_\mu + igA_\mu(x) \tag{60}$$

where g is a dimensionless coupling constant. The Lagrangian is invariant under a rotation of ψ by $\theta(x)$ if at the same time we shift

$$A_\mu(x) \to A_\mu(x) - \theta(x)/g. \tag{61}$$

Having introduced the gauge field (photon), we must now include in $\mathscr{L}$ a term for the kinetic energy: this must also be invariant under the gauge transformation and takes the form

$$\mathscr{L}_{\mathscr{A}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{62}$$

with the so called field strength defined as

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{63}$$

A further requirement for local gauge invariance is that the photon be massless because a mass term in $\mathscr{L}$ for the photon would have to be of the form $M^2 A_\mu A^\mu$ and hence would not be invariant under the translation of $A_\mu(x)$ given in Eq. (61).

Invariance under local symmetry transformations requires the existence of minimally coupled massless gauge bosons.

The example we treated above corresponds to a U(1) gauge transformation, namely the fields were only rotated by a phase $\theta(x)$. Suppose now that we want the more general transformations, as illustrated in Eqs. (9) and (11), to be dependent on space-time, *i.e.*, $\theta_a \to \theta_a(x)$. The solution to this problem was given by Yang and Mills (1954), who showed how to maintain invariance of the Lagrangian when the gauge fields themselves form a multiplet of the group G.

The prescription for invariance under an infinitesimal transformation $\delta\theta_a = \varepsilon_a(x)$ whereby

$$\psi_i(x) \to [1 + it_a\varepsilon_a(x)]_{ij}\psi_j(x) \tag{64}$$

is that ordinary derivatives be replaced by covariant derivatives

$$\partial^\mu \to D^\mu = \partial^\mu + igt_a A^\mu_a \tag{65}$$

where g is the dimensionless coupling constant, and that the gauge fields transform as

$$A^\mu_a(x) \to A^\mu_a(x) - \partial^\mu\varepsilon_a(x)/g - C_{abc}\varepsilon_b(x)A^\mu_c(x). \tag{66}$$

For continuous groups it is sufficient to specify the transformations under infinitesimal transformations. From the above we see that that the number of gauge bosons must equal the number of generators of the group G, or equivalently the gauge bosons must belong to the adjoint representation of G. For instance, if G is the unitary group SU(N), there have to be $N^2 - 1$ gauge bosons, *i.e.*, three for SU(2) and eight for SU(3).

As is the case for electromagnetism, the gauge bosons must be massless. The field strength needs to be of the form

$$F^{\mu\nu}_a = \partial^\mu A^\nu_a - \partial^\nu A^\mu_a - gC_{abc}A^\mu_b A^\nu_c \tag{67}$$

in order for the Lagrangian to be invariant.

The prototype of a gauge theory is the theory of the strong interactions known as quantum chromodynamics. It assumes that each type of quark comes in three species known as colors and that there is a gauged SU(3) symmetry group that acts on them. There is a red, a white, and a blue up type quark and similarly three down quarks and three strange quarks. Since the gauge group is SU(3), there must be eight massless spin-1 particles, the gauge bosons. These are commonly called gluons.

The field strength has terms proportional to the square of the field A^μ_a, and therefore the so-called kinetic term, proportional to the field strength squared, has terms proportional to the third and fourth powers of the field A_μ. This means that $\mathscr{L}$ has nonlinear interactions of the gauge field. This is only true when we have a so called non-Abelian symmetry group, *i.e.*, one for which $C_{abc} \neq 0$. For electromagnetism such terms are not present.

The nonlinear interactions of the gluons cause the effective quark–gluon coupling constant (coupling constant as modified by radiative corrections) to have a very different behavior than the electromagnetic coupling constant. In fact the quark–gluon coupling becomes smaller as one increases the momentum transferred to the gluon; for electromagnetism the electron's coupling to the photon becomes smaller as one decreases the momentum transferred to the photon. These changes in the effective coupling constants are quantum effects due to the dielectric behavior of the vacuum. They are not present in the corresponding classical field theories.

Conversely, quark interactions at small momentum transfer are very strong, leading to quark confinement: only SU(3) color singlet states are observable. For instance, the quark–antiquark potential grows with separation while the electron–positron potential decreases. Since quark interactions become weaker at large momentum transfer, the quarks become almost free, a property known as asymptotic freedom. Both aspects have been confirmed experimentally (Aitchison and Hey, 1989).

As of now the evidence indicates that both the U(1) symmetry of electromagnetism and SU(3) of color are unbroken, or equivalently, unbroken local SU(3) × U(1) is a symmetry of nature.

Gravity is also an unbroken gauge symmetry of sorts. The unbroken symmetry is general covariance, the conserved quantity is the energy-momentum tensor, and the gauge boson is the graviton, a quantum fluctuation of the metric tensor (Weinberg, 1972). Intuitively it is not surprising that if the quantum field that couples to a conserved current has spin 1, the quantum field that couples to a conserved second rank tensor should have spin 2.

The notion that gravity and electromagnetism might be reunited was of course Einstein's dream. This concept is presently being pursued by theorists exploring string theory (Gross, 1995), but that is outside the scope of this article.

We have seen that three of the four fundamental forces (electromagnetic, strong, and gravity) seem to be described by gauge theories with massless gauge bosons as mediators. What about the fourth known force, the weak force that describes beta decay, neutrino interactions, *etc*. Is it like the other three? The answer is that it is indeed similar to the other three forces in that there is a gauge invariance associated with the weak interactions, but there is also an important difference in that, although the Lagrangian is invariant under the corresponding transformations, the physical vacuum is not. This implies, as we shall see in the next section, that the corresponding gauge bosons are massive. In turn, this leads to the conclusion that the force generated by the weak interactions is short range, as opposed to the electromagnetic or gravitational forces.

Equation (48) gives an effective Hamiltonian describing the weak interactions, but we now understand that this is only valid to second order in perturbation theory for an underlying theory with interaction

$$\mathscr{L}_{weak} = gJ^{a\mu}A_{a\mu} \tag{68}$$

where $a = 1, 2, 3$ is the label for a set of three currents associated with the three gauge bosons of the weak interactions. The g and A^{μ} in the above equation are not meant to be confused with those we wrote for the strong interactions.

The breakthrough in understanding the connection of such a model with gauge theories came about with the development of the so-called Glashow–Weinberg–Salam model (Weinberg, 1995). It was realized that the underlying theory was a gauge theory with SU(2) invariance and hence three massless spin-1 gauge bosons and three associated conserved currents.

The quarks and leptons, which constitute the fermionic components of the theory, are placed in doublets and singlets of SU(2), and, of course, the Lagrangian is invariant under SU(2) symmetry transformations. Having the currents conserved is only possible if all fermion masses are zero, since the divergences of the weak currents are, as we have seen, proportional to sums and differences of fermion masses. This is achieved by putting all left chiral fermions in doublets and right chiral fermions in singlets (mass terms connect left to right chiral fields) and, of course, SU(2) invariance forbids direct coupling of doublets to singlets.

The resulting model is elegant, but far from describing nature since neither the mediators of the weak interactions nor the fermions are massless. On the other hand, introducing mass terms in the Lagrangian would break the local symmetry and destroy the renormalizability and hence the finiteness of the theory.

The resolution of this problem using the technique of spontaneous symmetry breaking is explained in the next section. It succeeds in giving masses to the three gauge bosons of the weak interactions: two of them, called $W^{\pm}$, are charged and the remaining one, the Z^0, is neutral. In summary, we say that the so-called "standard model" of the strong, weak and electromagnetic interactions is described by a SU(3) × SU(2) × U(1) gauge theory, spontaneously broken to SU(3) × U(1). At present the "standard model" includes as fermionic building blocks three negatively charged leptons, e^-, μ^-, τ^-, each with its associated neutrino, three charge-$\frac{2}{3}$ quarks, the u, c, and t quarks, and three charge-$(-\frac{1}{3})$ quarks, the d, s, and b quarks. In addition each and every quark comes in three so-called colors, corresponding to the triplet representation of the unbroken color SU(3) group.

An active field of research is the study of whether there exists some much larger gauge group, *e.g.*, SU(5), which spontaneously breaks to the standard-model invariance group (Aitchison and Hey, 1989).

9. THE HIGGS MECHANISM

There is one very important exception to the rule that spontaneously broken symmetries lead to Goldstone modes. This is the so-called Higgs mechanism, anticipated and discussed in the context of condensed matter physics by Y. Nambu and by P. W. Anderson (Weinberg, 1995). If the theory includes minimally coupled massless gauge bosons, they necessarily interact with the scalar bosons because the latter's kinetic energy terms need to have ordinary derivatives replaced by covariant derivatives. For instance, for O(3), we find

$$\begin{aligned}\partial^\mu\phi_a\partial_\mu\phi_a \to D^\mu\phi_a D_\mu\phi_a &= (\partial_\mu\phi_a + ig\varepsilon_{abc}A_{\mu,b}\phi_c)\\ &\quad \times (\partial^\mu\phi_a + ig\varepsilon_{ade}A^\mu_d\phi_e).\end{aligned} \tag{69}$$

If we now introduce into the above equation the vacuum expectation value of ϕ_a, namely the O(3) vector $(0, 0, v)$, we find an effective mass term for the gauge bosons

$$g^2v^2\varepsilon_{ab3}\varepsilon_{ad3}A_{\mu,b}A^\mu_d. \tag{70}$$

From this, we see that the third component field A^μ_3 has no mass term since $\varepsilon_{33a} = 0$ while the other two fields have acquired a mass proportional to gv. This acquisition of mass is puzzling since a massless gauge boson like the photon has only two polarization degrees of freedom while a massive gauge boson can have a longitudinal polarization as well as the two transverse polarizations.

The resolution of the puzzle lies in the Higgs mechanism. We apparently acquired two new degrees of freedom corresponding to the longitudinal modes of the gauge bosons. If we study the problem somewhat more carefully, we find that two massless scalar bosons can be removed from the Lagrangian by a gauge transformation and hence disappear from the spectrum. The colloquial expression is that they have been eaten by the gauge bosons.

Recapitulating, we see that spontaneous symmetry breaking leads to massless scalar bosons: if the scalar bosons are coupled minimally to massless gauge bosons, some or all of the massless scalars may disappear from the spectrum, replaced by the longitudinal modes of the now massive gauge bosons. Gauge bosons coupled to currents associated with unbroken symmetries remain massless. In the O(3) case rotational symmetry about the third axis is unbroken and A^μ_3 remains massless.

For the theory of electroweak interactions our starting gauge theory is SU(2) × U(1) with four massless gauge bosons. In addition we have a complex SU(2) scalar doublet, which transforms nontrivially under U(1) as well, so that there are four independent scalar fields (read and imaginary parts of the doublet). The fermions are left-handed chiral SU(2) doublets and right-handed chiral SU(2) singlets.

To begin with, the gauge bosons are massless, the fermions are massless [left-right couplings are forbidden by SU(2) invariance], and the scalars have a negative squared mass. The symmetry is spontaneously broken from SU(2) × U(1) to simply a U(1) symmetry. This leads to one massive scalar and three massless scalars. The latter are eaten by the four gauge bosons, three of which become massive, and one, the photon, stays massless. The three massive gauge bosons are the $W^\pm$ and the Z^0, the mediators of the charged and neutral weak interactions.

The fermions also acquire a mass, since Lagrangian interaction terms of the form

$$f\bar{q}_{L,a}\phi_a q_R \tag{71}$$

are allowed, and, once ϕ_a acquires a vacuum expectation value equal to v, fermions get masses proportional to fv. The predictions of what has come to be known as the standard model are extraordinarily successful. The $W^\pm, Z^0$ have been discovered at exactly the predicted masses, decay modes agree, *etc.* The one remaining unconfirmed piece of the puzzle is the detection of the single massive scalar, commonly called the Higgs boson.

The Higgs phenomenon has its counterpart in condensed matter physics as well. Let us illustrate this again with the example of superconductivity. The gauge symmetry is U(1), with the photon as the gauge boson. The scalar field is the order parameter, the complex scalar electron-pair wave function, which also transforms nontrivially under U(1) since it has charge −2. The order parameter has a nonzero vev, signalling a phase transition, and the phase of the order parameter is a massless excitation, corresponding to a Goldstone mode. The latter disappears from the spectrum, being eaten by the photon. Inside a superconducting medium, the photon is essentially massive: this is known as the Meissner effect (Weinberg, 1995).

10. EXPERIMENTAL LIMITS ON SYMMETRY BREAKING

We have discussed at some length the notion of symmetry. Let us now consider, at least briefly, some of the experimental limits on conservation laws. Begin with discrete symmetries; *TCP* follows, as we stated, from general principles of relativistic quantum field theory. The best limits on *TCP* invariance come from the equality of the K^0 and the $\bar{K}^0$ masses. The experimental limits on the mass difference Δ are that

$$\Delta / m_{K^0} \leq 10^{-18} \tag{72}$$

at a 90% confidence level. Unless otherwise indicated the experimental results cited in this section are all obtained from the Particle Data Group's compilation (Particle Data Group, 1994). *TCP* also predicts the equality of lifetimes *etc.* for particles and antiparticles. Tests have been performed to compare lifetimes for electrons and positrons, $\mu^{\pm}$ mesons, and so forth. As an example, the lifetimes of the latter are given by

$$\tau_\mu^+ / \tau_\mu^- = 1.00002 \pm 0.00008. \tag{73}$$

Another example is the equality of the anomalous magnetic moments of the electron and the positron. If *TCP* invariance holds, *T* and *CP* symmetries are presumably equal, but again it is worth testing them separately. Experiments in atomic, molecular, nuclear, and particle physics are being conducted to test these symmetries. At present the most sensitive test of time-reversal invariance is probably the limit on the electric dipole moment (edm) of the neutron. The present limit for the neutron is that its edm $< 1.1 \times 10^{-25} e\,\mathrm{cm}$. The bound on the electron edm is comparable to that of the neutron and is a more stringent test of unification theories. It is worth mentioning that a neutron nonzero edm requires violation of both *T* and *P*. Nuclear physics searches for *T* nonconservation are also being pursued (Haxton *et al.*, 1994).

CP symmetry is a vast topic (Jarlskog, 1988) At present the best limits on *CP*-invariance violation come from the analyses of *K* decay. In the neutral kaon system the strength of *CP*-invariance violation relative to the standard weak interactions is characterized by a dimensionless small parameter $\varepsilon \simeq 10^{-3}$. It is crucial to determine if ε is so small because *CP*-invariance violations are very weak or because they involve mixing with heavier quarks, which are only minor consituents of the kaons.

Large experimental facilities in Japan (KEK) and the USA (SLAC) are presently under construction that will provide further tests of *CP* nonconservation and begin to address these questions. An example is the so called BABAR facility at the Stanford Linear Accelerator which will study the B^0-$\bar{B}^0$ system (B^0 is a bound state of a b quark and either a d or an s antiquark).

Baryon-number conservation is presumed to be violated at some level if a grand unified theory that incorporates all known gauge theories exists. The present best limit on baryon-number conservation comes from the stability of the proton, which is known to have a lifetime greater than 10^{31} years.

Baryon-number nonconservation is one of the three key ingredients to the generation of a baryon asymmetry in the early universe, as was pointed out by Andrei Sakharov in a prescient 1967 paper (Kolb and Turner, 1990). The other two are *CP*-invariance violation and a departure from thermal equilibrium. Baryon asymmetry is, of course, present since our known universe is preponderantly made of matter, not of antimatter, but a wholly successful explanation of the asymmetry has not yet been advanced.

Lepton number is tested in neutrinoless double beta decay, particularly in isotopically enriched germanium (Avignone, 1995). The limits state the ratio of neutrinoless decays to neutrino decays to be less than one in a thousand.

The separateness of muon number and electron number is tested by the nonobservation of $\mu \rightarrow e + \gamma$, which has a branching ratio of $< 5 \times 10^{-11}$ to the total decay rate of the muon. The dominant muon decay is into a state of an electron, a neutrino, and an antineutrino. Though it is not obvious, lepton numbers are automatically conserved if neutrinos are massless (technically it would be sufficient if all neutrinos were mass degenerate). There are numerous ongoing searches for evidence of nonzero neutrino mass. Lepton mixing and nonzero neutrino masses may first be observed in so-called neutrino oscillation experiments. In these one species of neutrino, say an electron neutrino, is produced but it then oscillates between that and a second species, *e.g.*, a muon neutrino, while travelling from the source to the observer.

Local gauge invariance requires massless gauge bosons. One obvious verification is the masslessness of the photon. Present limits are

$$m_\gamma < 3 \times 10^{-27}\text{eV}. \tag{74}$$

Trying to set limits on gluon (the presumed mediator of the strong interactions) masses, on the other hand, is very difficult, since gluons are not directly observable. In fact, observation of a free gluon or a free quark would represent a departure from the presently held view that these particles are only observable when combined with other like such particles to form color singlets. Incidentally, all recent searches for free quarks have yielded null results.

Conservation of electric charge is presumed to hold. If it were not, we could imagine an electron decaying into a neutrino and a photon. Present limits on the electron lifetime are $\tau_e > 2.7 \times 10^{23}$ years.

One prominent symmetry we have not mentioned in the text so far is supersymmetry, the unique allowed extension of space-time translation symmetry to a group that includes particle transformations. The symmetry requires paired multiplets, necessarily degenerate in mass, of fermions and bosons. Thus, if supersymmetry exists in nature, it must be spontaneously broken. Some current theories suggest that the partners of the known elementary particles, differing in spin by one half unit, should lie in the mass range of a few hundred GeV. It is therefore an interesting question for the next round of accelerator experiments whether this extended symmetry could be a reality (Wess and Bagger, 1983).

GLOSSARY

Baryon: A strongly interacting particle obeying Fermi statistics, such as the proton or the neutron.

Bravais Lattice: The elementary specification of the location of the units in an infinite crystalline lattice, making manifest the periodic ordering and the translational symmetry.

Charge Conjugation: An operation that transforms a particle into its antiparticle. A theory which incorporates particles and antiparticles in a symmetric manner may then be invariant under the operation of charge conjugation or said to have charge-conjugation symmetry. Similar statements can be made about time-reversal and parity symmetries.

Gauge Boson: A particle of spin 1 that couples to matter in such a way as to maintain gauge invariance, the invariance under continuous space-time dependent transformations. The photon is an example of a gauge boson.

Goldstone Boson: A massless particle of spin 0 that arises when a continuous symmetry is broken spontaneously.

Group: A set of elements and a rule for combining them, commonly called group multiplication. The product of any two group elements must itself be a group element, multiplication must be associative, a unique identity element must exist, and every element must have a unique inverse such that the product of an element and its inverse yields the identity element. Groups may be continuous, such as the full rotation group, or discrete, such as the group of rotations about the z axis by 90°, which has only four elements: rotations through 90, 180, 270 and 360°.

Lepton: The name given to fermionic elementary particles with no strong interactions, as for example the electron and its neutrino.

Quantum Field: The generalization of the concept of a particle satisfying a quantum mechanical equation of motion. Particles are the quantum excitations of the appropriate field, *e.g.*, the photon is an excitation of the electromagnetic field. Quantum field theory allows naturally antiparticle as well as particle creation and annihilation.

Quark: The presumed structureless, *i.e.*, pointlike, elementary fermionic constituents of strongly interacting particles such as the neutron and the proton. Quarks are believed to be bound together by the mediators of the strong force, conventionally called gluons.

Spontaneous Symmetry Breaking: A situation that arises when the Lagrangian and hence the equations of motion of a system are invariant under a continuous symmetry, but the vacuum is not. This comes about because the minimum of the potential and hence the stable point corresponds to a nonzero expectation value of a scalar field.

Standard Model: The term commonly used for the model of matter described as consisting of pointlike fermionic constituents (quarks and leptons) interacting by means of gauge bosons, which in turn are self interacting. The model features an invariance under a set of specified symmetry transformations, corresponding to the group SU(3) × SU(2) × U(1).

Works Cited

Adler, S., Dashen, R. (1968), *Current Algebras*, New York: Benjamin Press.

Aitchison, I. J., Hey, A. G. (1989), *Gauge Theories in Particle Physics*, Philadelphia: Adam Hilger.

Ashcroft, N., Mermin, D. (1976), *Solid State Physics*, Philadelphia: Saunders.

Avignone, F. (1995), in: *Particle and Nuclear Astrophysics and Cosmology in the Next Millennium*, E. Kolb, R. Peccei (Eds.), Singapore: World Scientific.

Cheng, T. P., Li, L. F. (1984), *Gauge Theory of Elementary Particle Physics*, Oxford, UK: Clarendon Press.

Christenson, J., Cronin, J., Fitch, V., Turlay, R. (1964), *Phys. Rev. Lett.* **13**, 138–142.

Commins, E., Bucksbaum, A. (1983), *Weak Interactions of Leptons and Quarks*, Cambridge: Cambridge University Press.

Greutz, M. (1983), *Quarks, Gluons and Lattices*, Cambridge: Cambridge University Press.

Fradkin, E. (1991), *Field Theories of Condensed Matter Systems*, New York: W. A. Benjamin.

Gross, D. J. (1995), "Symmetry in Physics: Wigner's Legacy," *Phys. Today* **48**(12), 47–55.

Haxton, W., Horing, A., Musolf, M. (1994), *Phys. Rev. D* **50**, 3422–3454.

Itzykson, C., Zuber, J. B. (1980), *Quantum Field Theory*, New York: McGraw-Hill.

Jarlskog, C. (1988), *CP Symmetry*, Singapore: World Scientific.

Kolb, E. W., Turner, M. S. (1990), *The Early Universe*, New York: Addison–Wesley.

Landau, L. D., Lifshitz, E. M. (1981), *Quantum Mechanics- Non Relativistic Theory*, London: Pergamon Press.

Lax, M. J. (1974), *Symmetry Principles in Solid State and Molecular Physics*, New York: Wiley.

Lee, T. D., Yang, C. N. (1956), *Phys. Rev.* **104**, 254–266.

Particle Data Group, M. Aguilar-Benitez *et al.*, (1994) *Phys. Rev. D* **50**, 1173–1827.

Steinhardt, P. J., Ostlund, S. (1987), *Quasi-Crystals*, New York: World Scientific.

Streater, R. F., Wightman, A. S. (1968), PCT, *Spin and Statistics and All That*, New York: Benjamin Press.

Weinberg, S. (1972), *Gravitation and Cosmology*, New York: John Wiley.

Weinberg, S. (1995), *The Quantum Theory of Fields*, Cambridge: Cambridge University Press.

Wess, J., Bagger, J. (1983), *Supersymmetry and Supergravity* Princeton: Princeton University Press.

Wigner, E. P. (1959), *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra*, New York: Academic Press.

Yang, C. N., Mills, R. L. (1954), *Phys. Rev.* **96**, 191–205.

Further Reading

Coleman, S. (1989), *Aspects of Symmetry*, Cambridge UK: Cambridge University Press. A review of modern topics in quantum field theory, with particular emphasis on the gauge theory and its geometric underpinnings.

Lee, T. D. (1981), *Particle Physics and Introduction to Field Theory*, Chur, Switerland: Harwood Academic. A review of field theory and how it is used in elementary particle physics.

Mermin, D. (1992), *Rev. Mod. Phys.* **64**, 3–122. A learned review of space groups in crystallography.

Peskin, M., Schroeder, R. (1995), *An Introduction to Quantum Field Theory*, Reading, MA: Addison-Wesley. A modern reference for field theory and elementary particle physics.

Weyl, H. (1931), *The Theory of Groups and Quantum Mechanics*, New York: Dover Publications. A classic introduction to group theory and its applications to quantum mechanics.

Wigner, E. P. (1967), *Symmetries and Reflections: Scientific Essays of Eugene P. Wigner*, Bloomington: Indiana University Press. An overview of the notions of symmetry.

X-RAY MICROSCOPY AND MICROANALYSIS

WERNER MEYER-ILSE, *Center for X-ray Optics, Lawrence Berkeley National Laboratory, Berkeley, California, U.S.A.*

Introduction 323
1. **Contrast Mechanism** 324
1.1 X-Ray Interaction with Matter 324
1.2 Hard and Soft X Rays 324
1.3 Dichromatic Microanalysis 325
1.4 Spectromicroscopy 326
1.4.1 Orientation- and Polarization-Dependent Contrast 326
1.5 Radiation Dose 326
2. **Optics for X-ray Microscopy** 327
2.1 Zone-Plate Lenses 327
3. **X-Ray Microscopes** 329
3.1 X-Ray Sources and Detectors 329
3.2 Focusing Transmission X-Ray Microscopes 329
3.2.1 Conventional X-Ray Microscopes 329
3.2.2 Scanning X-Ray Microscopes 331
3.3 Examples of Other Focusing Microscopes 332
3.3.1 X-Ray Fluorescence Microprobes 332
3.3.2 Photoemission X-Ray Microscopes 333
3.4 Nonfocusing X-Ray Microscopes 334
3.4.1 Contact X-Ray Microscopy 334
3.4.2 Holography and Diffraction 334
3.5 Microtomography 334
3.6 Cryogenic Fixation 334
3.7 Antibody Labeling 335
4. **Applications** 335
4.1 Biology 336
4.1.1 Malaria 336
4.1.2 Sperm 336
4.2 Environmental Sciences 336
4.3 Materials Sciences 337
4.3.1 Chemistry, Civil Engineering 337
4.3.2 Cement and Concrete 337
4.3.3 Polymers 338
4.4 Future Applications 338
Glossary 339
Works Cited 340
Further Reading 341

INTRODUCTION

Use of shorter-wavelength radiation, like electrons or x rays, provides significantly better spatial resolution than visible-light microscopy (see OPTICAL MICROSCOPY). Although important, a high spatial resolution alone is not sufficient to make a microscope a useful tool. Little can be learned about a sample that is too opaque or too transparent to provide useful contrast. Individual living cells are a few micrometers thick, transparent to visible light, and opaque to electrons. That is why phase-contrast methods are commonly used in visible-light microscopy and sectioning and staining is used in electron microscopy. For soft x rays, however, such cells provide a natural contrast.

The desire for x-ray microscopes arose soon after the discovery of x rays by Wilhelm Röntgen in 1895, but the lack of materials with strong refraction and weak absorption prevented an early development of x-ray lenses and thus of high-resolution x-ray microscopy.

The unique interactions of x-rays with matter were extensively used soon after the discovery of x rays. X-ray spectroscopy, using gratings, was greatly developed by the 1920s. It was realized that x rays in the water window (see Sec. 1.1) would provide a natural contrast mechanism for x-ray microscopy of hydrated biological specimens. The moderate absorption and phase shifts of x rays when passing through matter, which make x-ray microscopy desirable, make it also impractical to build x-ray microscopes based on refractive optical elements.

X-ray microscopes that avoid x-ray optics altogether and record a shadow of the sample were therefore developed first (Coslett and Nixon, 1960). The best spatial resolution of shadow microscopes is achieved by placing the sample in close proximity to an x-ray point source or a high-resolution detector. When the

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

sample is close to a point source of x-rays, we speak of projection x-ray microscopy. When the sample is placed near the detector, it is called microradiography or contact x-ray microscopy.

Since refractive optics are impractical for x rays, focusing x-ray microscopes rely on reflective or diffractive optical elements. Kirkpatrick and Baez built the first focusing x-ray microscope (Kirkpatrick and Baez, 1948) using an arrangement of two focusing mirrors that operate through reflection at grazing incidence (see X-RAY OPTICS). Zone-plate lenses for x-ray microscopes were first discussed by Baez (1952) and are today used in the instruments with the highest spatial resolution. Present x-ray microscopes exceed the resolution of visible-light microscopes by more than a factor of 5, with further improvements expected in the future. The most complete progress reports on x-ray microscopy are documented in the proceedings of a series of conferences that started in 1984 in Göttingen (Schmahl and Rudolph 1984) and have been held every third year since (Sayre *et al.*, 1988; Michette *et al.*, 1992, Aristov and Erko 1994; Thieme *et al.*, 1998).

1. CONTRAST MECHANISM

Contrast in x-ray microscopy is from the interaction of x-ray photons with the electrons in the sample molecules. The specific dependence of the x-ray interaction with the sample on its elemental composition makes x-ray spectroscopy a very useful tool. When combined with high spatial resolution we speak about spectromicroscopy.

1.1 X-Ray Interaction with Matter

The interaction of soft x rays with matter is described by the complex index of refraction

$$n = 1 - \delta - i\beta. \tag{1}$$

The parameters δ and β are called the refractive index decrement and absorption index, respectively. An electromagnetic wave of wavelength λ passing through a sample with a thickness z will be attenuated in accordance to Beer's law:

$$I/I_0 = \mathrm{e}^{-\beta(4\pi/\lambda)z} = \mathrm{e}^{-\mu\rho z}, \quad \mu = 4\pi\beta/\lambda\rho, \tag{2}$$

and experience a phase shift of

$$\varphi = (2\pi/\lambda)\delta z. \tag{3}$$

The absorption coefficient μ ($\mathrm{cm}^2\,\mathrm{g}^{-1}$) and density ρ ($\mathrm{g\,cm}^{-3}$) are often used instead of the absorption index β.

The refractive index depends in a characteristic way on the photon energy of the x rays and the electronic configuration of the sample. X-ray photons are essentially probing the electronic structure of the molecules in the sample. As an example, we show in Fig. 1 the $1/e$ attenuation lengths of water and a generic protein as a function of photon energy. The absorption increases almost abruptly where the photon energy exceeds the electron binding energies of the elements in the sample. For carbon and oxygen the K-shell binding energies are at 284 and 543 eV, giving rise to the so-called water window, where organic matter absorbs nearly an order of magnitude more than water.

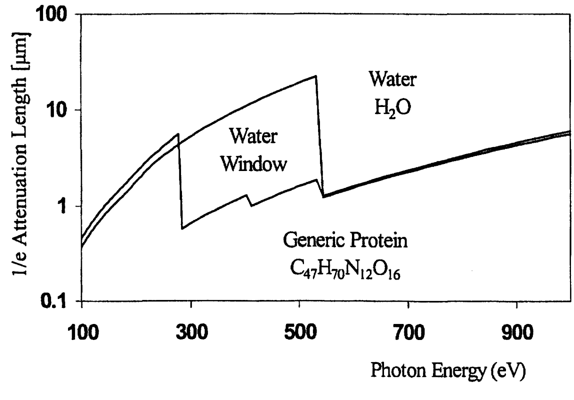

FIG. 1. X-ray $1/e$ attenuation lengths of a generic protein and water as a function of photon energy. The range between 284 and 543 eV is called the water window because water absorbs almost an order of magnitude less than organic material.

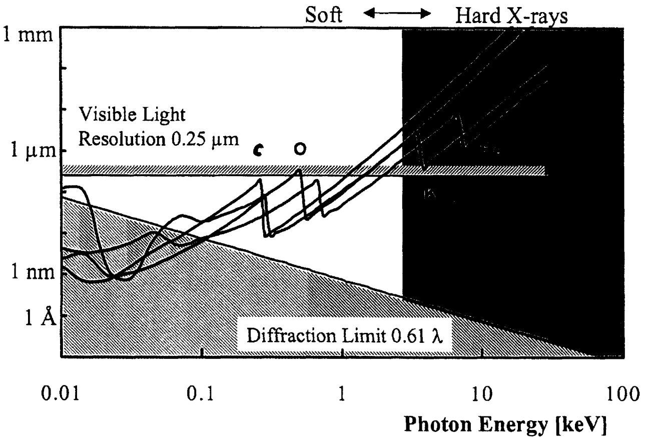

FIG. 2. Material thickness to produce 3% contrast from absorption in a transmission x-ray microscope as a function of photon energy for a variety of elements with a density of 1 g cm^{-3}.

The index of refraction is related to the atomic cross sections f_1 and f_2, which are well characterized and published for x-ray energies away from absorption edges (Henke *et al.* 1993) (see also X-RAY AND NEUTRON REFLECTIVITY):

$$\delta = \frac{r_0 \lambda^2}{2\pi} \frac{N_A}{A} \rho f_1, \qquad \beta = \frac{r_0 \lambda^2}{2\pi} \frac{N_A}{A} \rho f_2, \tag{4}$$

with r_0 the classical radius of an electron, A the atomic weight (g mol^{-1}), and N_A Avogadro's number (mol^{-1}).

Most high-resolution imaging in x-ray microscopy to date uses contrast primarily from photoelectric absorption in the water window, but phase contrast is used as well (Schmahl *et al.*, 1993). For x-ray energies just below the oxygen edge (*e.g.*, 517 eV, $\lambda = 2.4$ nm), the penetration depth of x rays is also ideally suited to image whole cells with a thickness of a few micrometers (see Fig. 1).

The photoelectric interactions of x rays near binding energies are more complex than shown in Fig. 1 and the Henke tables, because the binding state of the electrons affects the cross sections as well. This is used to map the chemical status of the sample.

1.2 Hard and Soft X Rays

The penetration depth of x rays in matter varies significantly throughout the photon energy range that is called x rays. Figure 2, for instance, shows the thickness of a variety of materials needed to produce a 3% absorption contrast. Also indicated are the resolution of visible-light microscopes (approximately $\frac{1}{4}$ µm) and the diffraction-limited resolution due to the finite wavelength of the x rays (0.61 λ). Soft x rays, that is for 100 to 2500 eV photon energy, are ideally suited to provide a spatial resolution exceeding that of visible-light microscopes. Harder x rays, with photon energies more than approximately 2500 eV, do not provide sufficient absorption contrast for high-resolution transmission x-ray microscopy. Phase contrast is also reduced when used with harder x rays, but at a lesser degree. The presence of K-shell absorption edges makes harder x rays also very useful for spectroscopic methods with spatial resolutions down to about 1 µm.

1.3 Dichromatic Microanalysis

The discontinuities of the x-ray absorption cross section are used to produce elemental or chemical distribution maps. Assume that we have two x-ray images I_1 and I_2 of the same sample taken at wavelengths λ_1 and λ_2. The two wavelengths are chosen so that the photon energies are just below and just above an absorption edge of the element to be studied. Further, assume that the cross section μ of this element changes between those wavelengths by $\Delta\mu$, and that the cross section of the other ele-

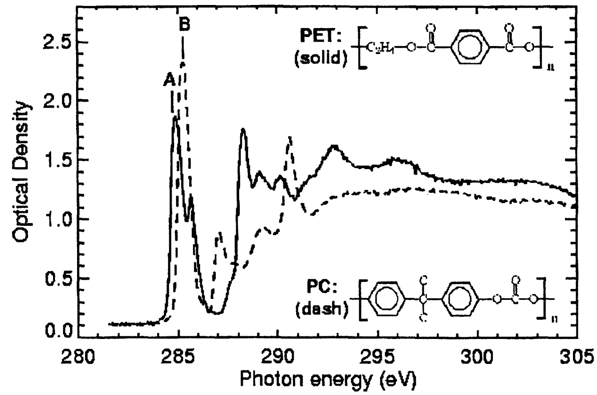

FIG. 3. Carbon NEXAFS spectra from (0.1 μm)^2 areas of a polymer blend. The solid line is from poly(ethylene terephthalate) (PET) and the dashed line from polycarbonate (PC). Images of the polymer blend are shown in Fig. 25. [Adapted from Ade *et al.* (1995), courtesy of H. Ade.]

ments in the sample are constant. Then we can map the areal density (g cm^{-2}) as

$$\rho z = \frac{1}{\Delta\mu} \ln\left(\frac{I_1}{I_2}\right). \qquad (5)$$

The sensitivity of this method is ultimately limited by photon noise. At this limit we can assume $I_1 \approx I_2 \approx I$ and the error of I being $\sqrt{I}$; then the uncertainty of the areal density measurements is

$$\Delta(\rho z) \approx \sqrt{2}/\Delta\mu\sqrt{I}. \qquad (6)$$

It should be noted here that the detectable quantities of the element in question relative to the bulk are on the order of one part in a hundred. If compared with other methods that detect relative quantities on the order of parts per million, this does not appear impressive at first sight. Because of the microscopic nature of the method, however, the absolute quantities that can be detected are quite small. For instance, an iron particle might have a $\Delta\mu = 10^4\,\mathrm{cm}^2\,\mathrm{g}^{-1}$. With a 40-nm spatial resolution and 1000 photons per resolution element detected, we could see 7.2×10^{-17} g of iron or 7.7×10^5 atoms concentrated within the (40-nm)^2 pixel.

1.4 Spectromicroscopy

Closer views of the x-ray spectrum, especially near the binding energy levels, reveal additional structures (see SPECTROMETERS). These structures are based on the energy for promoting electrons into a near-vacuum molecular energy level. Near-edge x-ray absorption fine-structure (NEXAFS) spectra therefore carry information on the chemical environment of the atoms (Ade *et al.*, 1992). Figure 3, for instance, shows the carbon NEXAFS spectra of two polymers: polycarbonate (PC) and poly(ethylene terephthalate) (PET).

The NEXAFS technique is also called XANES (x-ray absorption near-edge spectroscopy).

1.4.1 Orientation- and Polarization-Dependent Contrast The absorption contrast does not depend only on the chemical state of the electrons in the sample, but it is also dependent on the orientation and polarization of the incident radiation. This is especially noticeable with magnetic material. Magnetic circular dichroism provides significantly different absorption of magnetic domains for right and left circularly polarized radiation. The first demonstration of magnetic contrast with an x-ray microscope was done with the Göttingen x-ray microscope at BESSY (Schmahl *et al.* 1996). Ade and Hsiao (1993) have used soft x-ray linear dichroism to map the orientation of polymer molecules in Kevlar fibers.

1.5 Radiation Dose

The photoelectric absorption that provides contrast in soft x-ray microscopy also damages the sample (see BIOLOGICAL EFFECTS OF ELECTROMAGNETIC AND PARTICLE RADIATION), especially at the high doses the sample receives in an x-ray microscope.

The dose, measured in Gray (Gy), is the energy (J) absorbed per mass (kg). The energy of one photon depends on the wavelength and is (h = Planck constant, c = speed of light)

$$E_{\text{photon}} = hc/\lambda. \tag{7}$$

In practical units,

$$E = (1239.8\,\text{eV nm})/\lambda. \tag{8}$$

The dose D at the surface of a sample that is irradiated with n photons per area a (= pixel) is

$$D = (n/a)(hc/\lambda)\mu. \tag{9}$$

For instance, 1000 photons at 2.4 nm wavelength in a (40-nm)^2 size pixel will deposit a dose of 6.7×10^4 Gy ($\mu = 1.3 \times 10^4\,\text{cm}^2\,\text{g}^{-1}$). This is a typical dose for scanning x-ray microscopes. Conventional x-ray microscopes use about an order of magnitude higher dose (see Sec. 3.2.1).

When compared with electron microscopy, x-ray microscopy requires about an order of magnitude less dose to form an image of the same quality from samples that are a couple of micrometers thick (Sayre *et al.*, 1977). This, however, is not low enough to allow direct observation of ongoing living processes (see BIOLOGICAL EFFECTS OF ELECTROMAGNETIC AND PARTICLE RADIATION).

The dose tolerated by a sample in an x-ray microscope before structural changes become visible depends on the exposure time and the sample. Very short x-ray exposure times can be used to investigate initially living samples without limitations to the dose tolerated. At present, however, there are no single-pulse x-radiation sources available that routinely provide the brightness and photon flux needed for images of the quality produced by the best x-ray microscopes, although they are used in x-ray contact microscopy.

Exposure times with x-ray microscopes at synchrotron radiation light sources, which provide the highest-resolution images to date, are on the order of seconds. Sample changes due to the dose during exposure have to be considered with these instruments. The dose tolerated varies widely depending on the sample as well as its preparation. Unfixed chromatin in dried sperm, for instance, is resistant enough to allow imaging at high spatial resolution and good signal-to-noise ratio without visible structural damage. Living green algae can only be imaged, before structural damage degrades the image quality, with a reduced signal-to-noise ratio or spatial resolution.

Chemical fixation—*e.g.*, with glutaraldehyde—and cryogenic fixation are frequently used to reduce the effects of radiation dose to radiation sensitive materials.

2. OPTICS FOR X-RAY MICROSCOPY

X-ray optical elements are essential for focusing x-ray microscopes. Most high-resolution x-ray microscopes use zone-plate lenses, which are covered here. Holographically made zone plates were the first to provide a spatial resolution in x-ray microscopy exceeding that of visible-light microscopy (Schmahl and Rudolph, 1969). Even higher resolution is achieved with electron-beam lithography (Anderson and Kern, 1992). Please refer to X-RAY OPTICS for a more general discussion of x-ray optics other than zone plates.

2.1 Zone-plate Lenses

Zone plates are diffractive focusing elements made of concentric circular zones (see Fig. 4). The radii r_n of the zones $n = 1$ to N are chosen so that mth order diffraction will focus radiation of wavelength λ as shown in Fig. 5 and the equation

$$\sqrt{a^2 + r_n^2} + \sqrt{b^2 + r_n^2} = a + b + nm\lambda/2. \tag{10}$$

When used in a microscope, we can simplify Eq. (10), by assuming $a \to \infty$ and $b \to f$ (focal length), to

$$r_n^2 = mn\lambda(f + mn\lambda/4). \tag{11}$$

For x-ray focusing we can further assume $f \gg mN\lambda/4$, so that

$$r_n^2 = mn\lambda f. \tag{12}$$

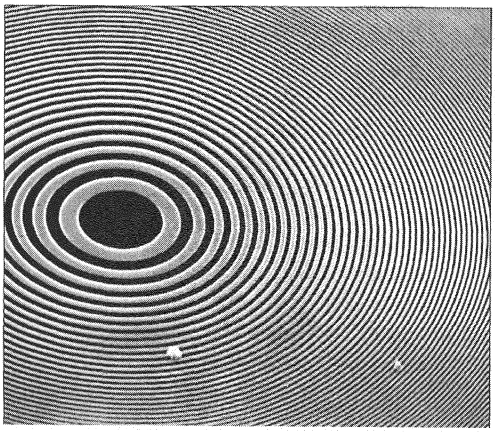

FIG. 4. Electron-microscope image of zone-plate lens with an outermost zone width of 35 nm. The zones are made of nickel and are 150 nm thick. Note: the zones appear elliptical because the lens is viewed at an angle. (Courtesy of E. Anderson.)

The zone width is $\Delta r_n \approx \mathrm{d}r_n/\mathrm{d}n = m\lambda f/2r_n$ and the width of the outermost zone is

$$\Delta r_N = m\lambda f/2r_N. \tag{13}$$

The diffraction pattern of zone plates approaches the Airy pattern of an ideal lens in the limit of many zones. For N bigger than 100, differences can be neglected. The transverse resolution of zone plate lenses is therefore in first order (Born and Wolf 1980)

$$\delta_t = 0.61\lambda/\mathrm{NA} \approx 1.22\Delta r_N/m \tag{14}$$

with the numerical aperture $\mathrm{NA} = \sin(r_N/f) \approx r_N/f$ (see Optical Microscopy). Equation (14) shows that the spatial resolution that can be achieved with zone-plate lenses in first order is approximately equal to the width of the outermost zone, independent of the wavelength. As with thin lenses, the focal tolerance of zone plate lenses in first order is (Born and Wolf 1980)

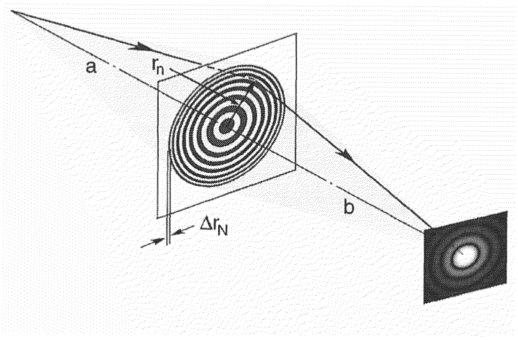

FIG. 5. Geometry of a zone-plate lens.

$$\delta_z = \pm 0.5\lambda/NA^2 = \pm 2(\Delta r_N)^2/m^2\lambda. \tag{15}$$

Zone plates work best with monochromatic radiation. Chromatic aberration is considered negligible if

$$\lambda/\Delta\lambda > mN, \tag{16}$$

with the full bandwidth of the radiation $\Delta\lambda$. A certain amount of chromatic aberration beyond the limit stated in Eq. (16) results in only a modest loss in resolution and is therefore in practice often tolerated.

The efficiency of zone-plate lenses is the fraction of incident radiation that is diffracted into the desired order. The efficiency of an ideally made zone plate with a zone thickness of t made from a material with δ and β as in Eq. (4) is (Kirz, 1974)

$$\begin{aligned} \eta &= \frac{1}{(m\pi)^2}[1+\mathrm{e}^{-2(\beta/\delta)\phi} - 2\mathrm{e}^{-(\beta/\delta)\phi}\cos\phi], \\ \phi &= (2\pi/\lambda\delta)t. \end{aligned} \tag{17}$$

Perfectly made alternating opaque ($\beta \to \infty$) and transparent zones have an efficiency of $1/\pi^2$ or about 10% in the first order. By selection of a proper material and thickness, the efficiency can be improved for a specific wavelength. Efficiencies on the order of 20% can be achieved for soft x rays, and even higher values are possible for hard x rays.

Zone-plate lenses with an outermost zone width of about 20 nm have been under development since 1993 and are used in x-ray microscopes with the highest spatial resolution (Schneider *et al.*, 1995). Similar lenses produce the finest focused spot of electromagnetic radiation of any wavelength (Spector *et al.*, 1997). Present developments focus on the improvement of the fine zone structures required for best resolution and diffraction efficiency. New electron-beam writing machines (Anderson *et al.*, 1995) and nanofabrication tools are expected to provide zone plates that provide 20 nm or better spatial resolution in the future.

3. X-RAY MICROSCOPES

Various instrumental setups are used to take advantage of the contrast mechanisms provided with soft x rays. These include direct-imaging, scanning, contact, projection, and holographic methods. In a scanning transmission x-ray microscope (STXM) the sample is scanned through a small focal spot. The transmitted photons are detected and used to build up a digital image in a computer. The smallest focal spot requires a spatially coherent illumination of the zone-plate lens and nanometer-precision scanning capabilities.

In a conventional transmission x-ray microscope (TXM) setup, the sample is illuminated with incoherent x rays. The transmitted photons are enlarged by a zone-plate lens to form an image on an x-ray–sensitive area detector.

The image formation, which means the relation between the transmission of the sample and the image detected, is partially coherent for both scanning and conventional x-ray microscopes and therefore nonlinear (Jochum and Meyer-Ilse, 1995). The nonlinearity is especially noticeable for feature sizes approaching the diffraction limit. Figure 6 shows the comparison of the intensity across a knife edge, from a theoretical calculation that included partial coherence effects, with experimental data. Indicated is the width of the knife edge from 10 to 90% intensity, which coincides with the Raleigh δ_t in an incoherent imaging system.

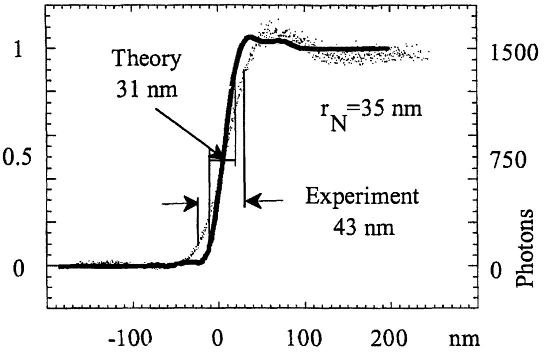

FIG. 6. Spatial resolution of the x-ray microscope XM-1 in Berkeley experimentally determined with a knife-edge test compared with a theoretical calculation that included partial coherence and chromatic aberrations.

3.1 X-ray Sources and Detectors

X-ray microscopes require bright x-ray sources and sensitive detectors for best performance. Synchrotron radiation sources (see Synchrotron Radiation) are used for the most advanced systems. Further development of plasma sources and more common availability of x-ray lasers (DaSilva *et al.*, 1992) might change that in the future.

The most important criterion for detectors in x-ray microscopy is the detection efficiency. A high detection efficiency reduces the dose to the sample and makes best use of the photons available from the source.

Zone-plate microscopes utilize monochromatic radiation, which has to be provided by a suitable monochromator. In this case, the detector will not have to provide information on the energy of the photons detected, because that is already known. With focusing x-ray microscopes, the magnification of the instruments can be designed to match the spatial resolution of the detector. Contact microscopy and x-ray holography, however, require a formidable spatial resolution, generally achieved with x-ray–sensitive photoresist, which is often read out with scanning-probe microscopes.

3.2 Focusing Transmission X-Ray Microscopes

3.2.1 Conventional X-Ray Microscopes Schmahl and collaborators at the University of Göttingen (Niemann *et al.*, 1976; Rudolph *et al.*, 1984; Niemann *et al.*, 1994; see also description in Optical Microscopy) pioneered conventional transmission x-ray microscopes (TXM) with zone-plate optics. The Göttingen microscope is installed at the BESSY synchro-

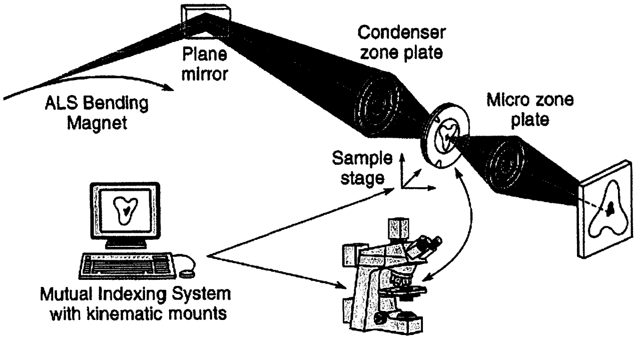

FIG. 7. Optical layout of the conventional x-ray microscope XM-1 at the Advanced Light Source (ALS) in Berkeley. This instrument features a mutual indexing system, which allows one to examine, select, and precisely adjust samples into the focal plane with a visible-light microscope before and after examination with x rays.

tron light source in Berlin. The optical layout is now used in several other places. Figure 7 shows the layout of the microscope at the Advanced Light Source in Berkeley, which also follows the optical design of the Göttingen x-ray microscope.

Conventional x-ray microscopes of the Göttingen type use two zone-plate lenses: a condenser and a micro zone-plate objective. The condenser focuses x rays from a bending-magnet source of a synchrotron facility onto the sample. The chromatic aberrations of the condenser are used to monochromatize the radiation through a pinhole placed in close proximity of the sample. The field that can be imaged is limited to about the size of this pinhole; larger images can be tiled together as shown in Fig. 8. The radiation transmitted through the sample is then enlarged with the objective zone-plate lens onto an x-ray–sensitive area detector. Usually this is a CCD detector (see CHARGE COUPLED DEVICES), which is connected to a computer to generate a digital image.

Conventional x-ray microscopes are relatively uncomplicated as compared with scanning microscopes, but still provide the highest possible resolution achievable with the objective zone-plate lens. These instruments make effective use of the radiation provided, and exposure times are on the order of seconds when used at a synchrotron-radiation facility.

A conventional x-ray microscope setup is also ideally suited for x-ray pulse sources. Both the laboratory x-ray microscope built by the Göttingen group (Niemann *et al.*, 1990) and the first x-ray microscope using an x-ray laser (DaSilva *et al.*, 1992) utilize the conventional setup.

Zernike-type phase contrast (see OPTICAL MICROSCOPY) is relatively easy to do with conventional x-ray microscopes. Figure 9 shows the phase-contrast setup pioneered by Schmahl *et al.* (1993). An example of a alga, *Euglena gracilis*, imaged in absorption and phase contrast is shown in Fig. 10. These images were acquired at low temperature from samples frozen in liquid ethane (see Cryogenic Fixation, Sec. 3.6).

There are two drawbacks of conventional x-ray microscopes. First, the radiation dose to

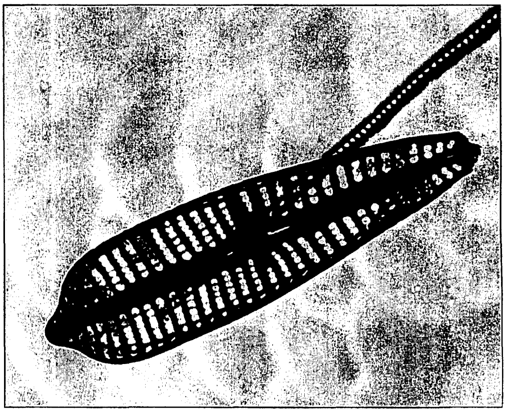

FIG. 8. High-resolution image of a diatom imaged with the conventional x-ray microscope XM-1 in Berkeley. This image contains approximately 100 million pixels of 20 nm size and was tiled together from 36 individual images obtained with a zone plate with 35-nm outermost zone width. The size of this image is 27 by 22 μm. (Courtesy of A. Nair.)

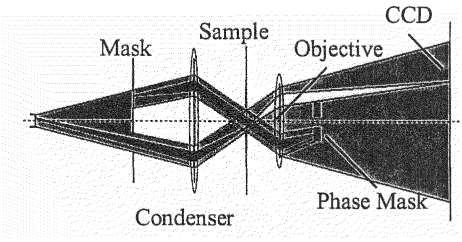

FIG. 9. Zemike phase-contrast setup used in the Göttingen x-ray microscope at BESSY. The condenser illuminates the sample by focusing the source into the sample plane. An annular mask is placed approximately one focal length in front of the condenser. The condenser and objective together concentrate all radiation unperturbed by the sample into the back focal plane of the objective, where it is phase shifted with a phase mask. This leads to image contrast from the phase shift in the sample. In contrast to visible-light microscopes, the phase mask will be absorbent as well and the setup can be optimized for one wavelength only.

the sample is about an order of magnitude higher with a conventional x-ray microscope than with a scanning x-ray microscope. This is because the objective zone-plate lens with its limited efficiency is downstream of the sample. The limitations from such increased dose can be greatly reduced by cryogenic fixation (see below). The second drawback is that the luminescent mode of operation (see below) is not possible at a resolution exceeding that of visible light microscopes.

3.2.2 Scanning X-Ray Microscopes The group of Kirz at the State University of New York at Stony Brook pioneered scanning x-ray microscopy. Their instrument is operated at the National Synchrotron Light Source (NSLS) in Brookhaven at a beam line called X1A. Figure 11 shows the principal setup of a scanning transmission x-ray microscope (STXM). Scanning microscopes were the first to apply spectromicroscopy techniques. The radiation dose to the sample is about an order of magnitude less than with a conventional TXM, because the zone-plate lens is located upstream of the sample. There are several practical advantages of scanning x-ray microscopes. For instance, the field of view can be varied easily, permitting low-resolution overview scans as well as high-resolution scans through computer control.

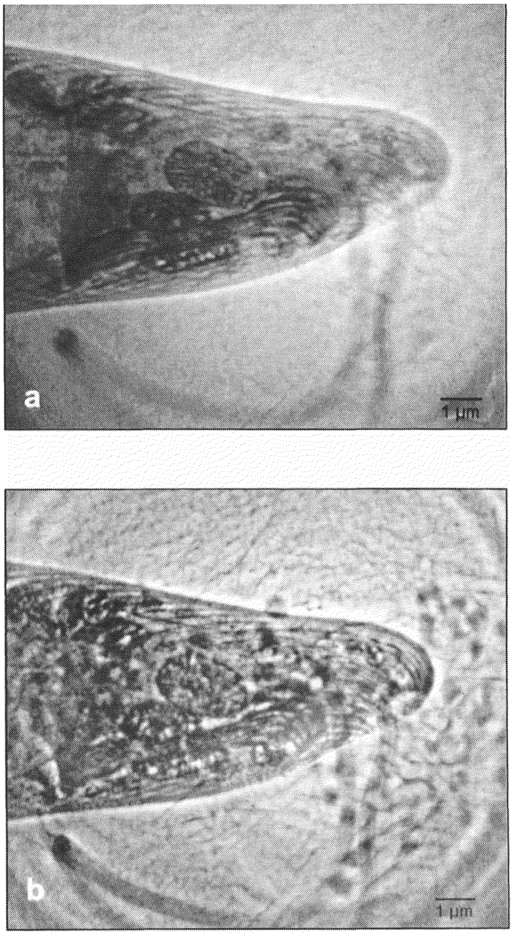

FIG. 10. X-ray micrographs acquired at 517 eV photon energy ($\lambda = 2.4\,\mathrm{nm}$) in (a) amplitude contrast and (b) phase contrast of an alga *Euglena gracilis*, which was shock frozen in liquid ethane. With cryogenic fixation, the organization of the chloroplasts —which are important organelles of photosynthesis—can be analyzed in the intact alga by using solely the natural contrast. Exposure times: (a) 3 s, (b) 15 s. The dose accumulated in vitrified water was about 3×10^6 Gy for either image. These images were acquired with the Göttingen x-ray microscope at BESSY in Berlin. [Adapted from Schneider *et al.* (1998), courtesy of G. Schneider.]

In addition to transmission mode, STXMs are also used to excite visible-light luminescence for high-resolution specific labeling experiments (Jacobsen *et al.*, 1993; Irtel von Brenndorff *et al.*, 1994), and dark-field mode is used to detect silver-enhanced immono-gold labels (Chapman *et al.*, 1996).

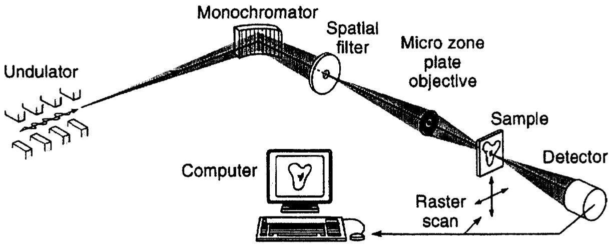

FIG. 11. Optical layout of a scanning transmission x-ray microscope (STXM). Scanning photoemission microscopes (SPEM) and scanning luminescence microscopes differ in the detector arrangement but are similar otherwise.

Scanning microscopes are more complex and demanding on the light source than are conventional microscopes, because they require spatially coherent illumination of the objective zone plate for highest spatial resolution. Exposure times are therefore long, even with today's brightest undulator sources (see SYNCHROTRON RADIATION). Despite this complexity, it is easier to implement spectromicroscopy techniques with scanning x-ray microscopes than with conventional microscopes. This is because high–spectral-resolution monochromators can be realized more easily for the spatially coherent radiation used with scanning microscopes, as compared with the incoherent radiation used with conventional microscopes.

3.3 Examples of Other Focusing Microscopes

The transmission microscopes described above provide the highest–spatial-resolution x-ray images available today. There are other x-ray focusing systems, in general with reduced resolution, that have properties relevant in the context of x-ray microscopy.

3.3.1 X-Ray Fluorescence Microprobes In an x-ray fluorescence microprobe, x-ray optics are used to focus a bright source of x-rays onto a sample of interest. These x-rays excite the characteristic fluorescent x-rays of the elements in the sample, which are detected and analyzed by an energy-dispersive detector such as a Si–Li detector (see X-RAY DETECTORS), allowing an assay of the elements in the sample. Scanning the focal spot over the sample allows the spatial distribution of the elements in the sample to be determined. With a very bright source such as a third-generation synchrotron-radiation storage ring, this method can be very sensitive.

The optics in such an instrument may be regarded as inverted microscopes that form a highly demagnified real image of the source on the sample. Various types of optical element, including zone plates, can be used to achieve this focusing. Another possibility is to use glancing-incidence mirrors in the Kirkpatrick–Baez configuration (Underwood *et al.*, 1988). Coating the surfaces with x-ray–reflecting multilayers (see also X-RAY OPTICS) allows an increase in the glancing angle (and hence a decrease in the aberrations), while at the same time acting as a low-resolution monochromator for the x rays; thus a separate monochromator is not required. The resultant spectral filtering allows selective excitation of the elements of interest and increases the sensitivity of the fluorescence analysis. Figure 12 is a schematic diagram showing the principle of an x-ray fluorescence beam line (Underwood *et al.*, 1996).

This beam line uses a pair of multilayer coated mirrors, shaped into an ellipse by precisely controlled bending, to demagnify the x-ray source (at the Advanced Light Source). With this scheme a spot size of 1 μm is achieved. Using mirrors optimized to focus 12.5-keV x rays, the instrument is capable of analyzing elements with atomic numbers ranging from $Z = 16$ (sulfur) to $Z = 33$ (arsenic) with a sensitivity of tens of femtograms (10^{-14} g). This capability allows the study of a wide range of samples in the fields of materials science, geology, geochemistry, and the life sciences. As an example of the capability of this instrument, Fig. 13 shows a scan of the poison-injecting fang (chelicera) of a garden spider. Interesting features of this scan are the

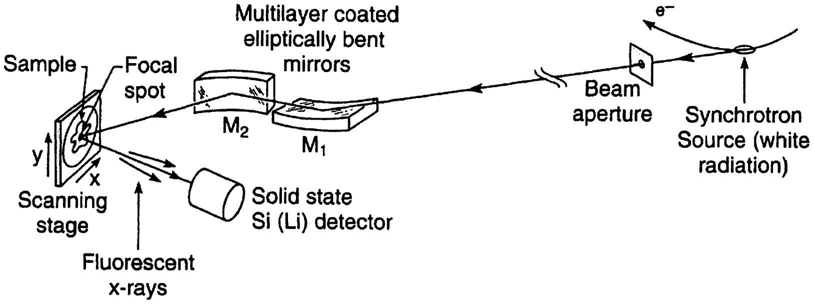
Multilayer coated elliptically bent mirrors
Sample
Focal spot
M2
M1
y
x
Scanning stage
Fluorescent x-rays
Solid state Si (Li) detector
Beam aperture
e⁻
Synchrotron Source (white radiation)

FIG. 12. X-ray fluorescence microprobe beam line at the Advanced Light Source. (Courtesy of J. Underwood.)

extremely high concentration of zinc at the tip of the fang and the high concentration of arsenic in the venom duct of the fang. This latter observation is particularly interesting since spider venoms are not inorganic poisons but large proteins similar to snake venom.

Zone plates for hard x rays have been demonstrated as well (Yun *et al.*, 1992) and produce submicrometer size x-ray spots at about 10 keV photon energy. As mentioned above, zone plates have relatively high diffraction efficiencies for hard x rays. This advantage, however, is balanced by the requirement of monochromatic illumination as compared with a microprobe that uses Kirkpatrick–Baez mirrors.

3.3.2 Photoemission X-Ray Microscopes

Photoelectrons generated by x-ray excitation carry information on the electronic structure of the materials involved (see PHOTOEMISSION AND PHOTOELECTRON SPECTRA). Photoemission microscopes use either a small x-ray focal spot to excite the photoemission or electron imaging to provide spatial information.

A variety of photoemission microscopes are installed at several synchrotron-radiation facilities around the world. A review of these is beyond the scope of this article. More information on these systems can be found the proceedings of the x-ray microscopy conference series (Thieme *et al.*, 1998; Aristov and Erko 1994; Michette *et al.*, 1992). The instruments below are described to illustrate the principles of scanning photoemission microscopes.

Two scanning x-ray microscopes are installed at beam line 7.0 of the Advanced Light Source (ALS) in Berkeley (Warwick *et al.*, 1998). One of these instruments is a scanning transmission microscope (STXM); the other is a photoemission microscope (SPEM). Figure 14 shows an image from the SPEM with corresponding energy spectra of the emitted electrons. This SPEM uses a zone-plate lens to

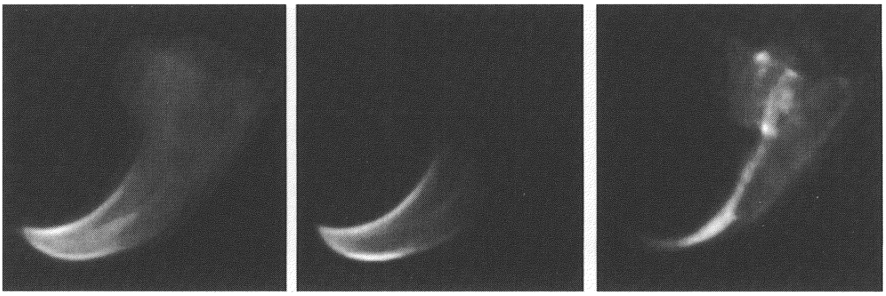

FIG. 13. X-ray microprobe image of a spider fang. The left image shows the x-ray absorption images; the middle and right images show the distribution of zinc and arsenic. (Courtesy of A. Thompson.)

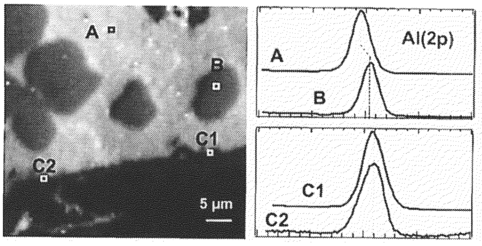

FIG. 14. Scanning photoemission microscopy (SPEM) image and corresponding Al x-ray photoemission spectra (XPS) from a study of Al–Ti melt interactions with graphite. These data were obtained with the SPEM at beam line 7.0 of the Advanced Light Source in Berkeley. (Adapted from Warwick *et al.* (1998), courtesy of T. Warwick.)

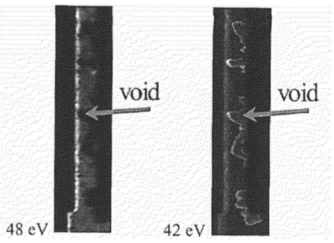

FIG. 15. Photoemission images of an Al–Cu wire on a substrate (4 wt% Cu, 5 μm width) showing electromigration after failure caused by an over-current experiment performed *in situ*. The images are at 42 and 48 eV electron energy. The reverse contrast at 42 eV indicates the presence of charging Al in the voids. The size of the images is $13 \times 60\,\mu m^2$. (Courtesy of J. F. Lorusso.)

excite photoelectrons from a small spot at the sample. The energy of the electrons is analyzed by an electron spectrometer, the image is formed by raster scanning of the x-ray spot. The spatial resolution of this instrument is currently about 0.3 μm; and the exciting photon energy can be varied from 100 eV to 1 keV.

Another scanning photoemission microscope installed at the ALS is called MAXIMUM (Multiple Application X-ray Imaging Undulator Microscope; Ray-Chaudhuri *et al.*, 1993). It uses a multilayer-coated Schwarzschild objective to produce the scanning spot. The multilayer coating (Ru/B_4C) has its maximum of reflectivity at 130 eV photon energy; the emitted photoelectrons are analyzed by means of a cylindrical-mirror analyzer (CMA). This microscope was originally developed at the Synchrotron Radiation Center in Wisconsin and has demonstrated a spatial resolution of 0.1 μm and an energy resolution of 0.3 eV. See Fig. 15 for an example of results obtained with this microscope.

3.4 Nonfocusing X-Ray Microscopes

3.4.1 Contact X-Ray Microscopy Contact x-ray microscopes use laser-produced plasmas to generate short x-ray pulses that are used to illuminate the sample, which is placed directly onto the detector. With short pulse duration, sample damage only becomes apparent after the image has been recorded, so that initially living material can be imaged without any fixation. The spatial resolution however, is limited by diffraction.

3.4.2 Holography and Diffraction When a sample is illuminated with a coherent wave front, the radiation passing through it carries all the information on the absorption and phase shifts from the sample. This is the basis of holography and diffraction methods (see HOLOGRAPHY, OPTICAL; X-RAY DIFFRACTION).

3.5 Microtomography

Information on the three-dimensional structure of the sample can be obtained with stereo and tomographic methods (see X-RAY IMAGING: RADIOGRAPHY AND TOMOGRAPHY; Lehr, 1997).

3.6 Cryogenic Fixation

Cryogenic fixation is the preferred method to prevent sample damage in x-ray microscopes. The first group that built a cryogenic sample stage for x-ray microscopy was the Göttingen group. Their system uses cold dry nitrogen gas to keep the samples at low temperatures. The samples do remain at atmospheric pressure for easy access with visible light microscopy. The amplitude- and phase-contrast images in Fig. 10 were taken at low temperatures. A total dose of 6×10^6 Gy was

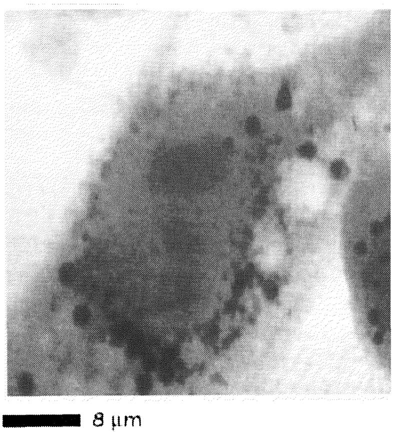

FIG. 16. Frozen hydrated 3T3 fibroblast imaged at 110 K in a cryogenic scanning transmission x-ray microscope. The cultured cell was grown on a Formvar-coated gold electron-microscope grid and plunge frozen from culture medium into liquid ethane. This image was acquired with use of a 60-nm zone plate fabricated by Stony Brook and Bell Labs. [After Maser *et al.*, (1998), courtesy of C. Jacobsen.]

needed to acquire those images. Tests with other frozen algae by the same group showed that an at least three orders of magnitude higher dose could be tolerated without visible damage.

Other groups adapted the cryogenic sample stages available for electron microscopy for their x-ray instruments (Maser *et al.*, 1998 and Lindaas *et al.*, 1998). Figure 16 shows an image obtained at X1A.

3.7 Antibody Labeling

Antibodies that recognize specific proteins can be used to detect and locate those proteins in cells. To do this, either the antibody that recognizes the specific protein is directly tagged with a marker that can be detected in the microscope, or a tagged secondary antibody is subsequently utilized to locate the primary antibody/protein complex. The important point is that in either case the marker selected must be detectable by the type of microscope used. Fluorescent markers are commonly used with visible-light microscopes, and colloidal gold, sometimes enlarged with silver, is primarily used as a marker with electron microscopes.

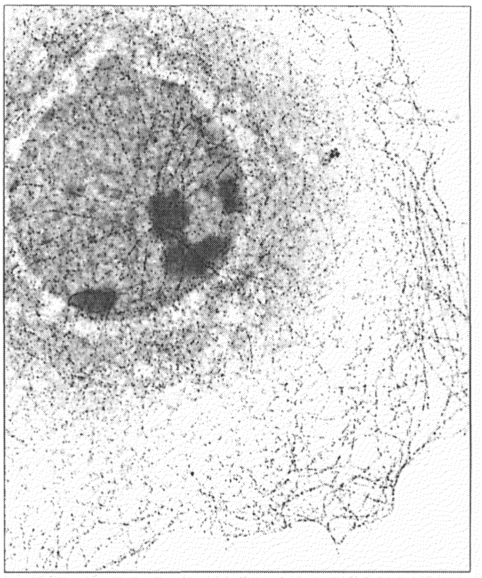

FIG. 17. Part of a hydrated mouse epithelial cell labeled with an antibody against tubulin and 1.4-nm gold markers that were enhanced with silver to about 50–100 nm size. The nucleus and its nucleoli as well as the cell structures in the cytoplasm are visible through natural contrast. Note the strands of labeled tubulin that cross either above or below the nucleus. The images were taken with the transmission x-ray microscope XM-1 at the Advanced Light Source in Berkeley. The size of the image is 26 by 31 μm^2. (Courtesy of C. Larabell, D. Hamamoto, S. Lelièvre, and W. Meyer-Ilse, unpublished.)

Lanthanide-based luminescent markers that can be used in x-ray microscopy are currently under development. Chapman *et al.*, (1996) pioneered the use of silver-enhanced gold markers with x-ray microscopes. The high penetrating power of x rays allows for the visualization of the labels even through or inside the nuclei of unsectioned cells (Fig. 17).

4. APPLICATIONS

X-ray microscopy application cover a wide range of fields. All these fields take advantage of the ability to study hydrated, relatively thick, and often unstained samples with high spatial resolution. The examples shown below were selected to represent the diversity of the field and are not meant to be complete.

4.1 Biology

Biological applications of x-ray microscopy include the investigations of mammalian cells, algae, sperm, spores, chromosomes, parasites, and bone structure (Thieme *et al.*, 1998). Studies of malaria and sperm are given to illustrate the use of high-resolution transmission imaging and elemental mapping, respectively.

4.1.1 Malaria Intracellular parasites like those of malaria are small; hence their morphology is challenging for visible-light microscopy studies. X-ray microscopy combines the advantage of high resolution with penetration of any enclosing structure for studies of intact parasites and their interaction with the host red blood cells.

The development of malaria (*Plasmodium falciparum*) in human red blood cells was intensively studied by Magowan *et al.* (1997). Thousands of x-ray images from malaria parasites were analyzed. In infected red blood cells, the mass redistribution from parasite activity was measured and newly elaborated structures in the cytosol were detected. The changes of the parasite morphology due to potential chemotherapeutic agents (Fig. 18) and from skeletal mutations of the host red blood cells were extensively studied as well. The findings demonstrate conclusively for the first time that constituents of the red blood cell membrane play a role in parasite development and extend our understanding of the relationship between the host blood cell membrane and the parasite.

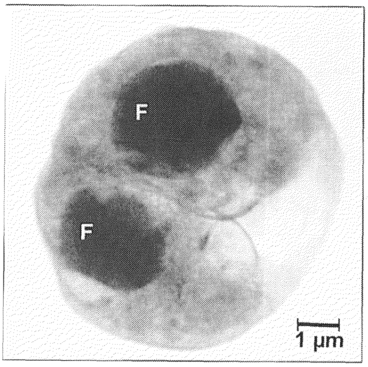

FIG. 18. Two malaria parasites in a human red blood cell. The parasites contain abnormally dense food vacuoles (F) from treatment with a protease inhibitor, a possible chemotherapeutic agent. The images were taken with the transmission x-ray microscope XM-1 at the Advanced Light Source in Berkeley. [Courtesy of C. Magowan, J. T. Brown, and W. Meyer-Ilse; see also Magowan *et al.* (1997).]

4.1.2 Sperm The small size of sperm cells makes them, like the malaria parasite, challenging for visible-light microscopy. Electron microscopy has revealed that DNA is packed so densely inside the nucleus of the mature sperm that electrons cannot penetrate it. X-ray microscopy, combined with atomic-force microscopy, obtained reasonably accurate estimates for the water content of air-dried rat sperm chromatin (DaSilva *et al.*, 1992) and provided the first clear evidence that sperm chromatin must be extensively hydrated, even in its highly compacted state.

Differences between near-edge x-ray absorption spectra of protein, bovine serum albumin (BSA), and DNA make it possible to map these components separately in bull sperm, on the basis of images taken at several energies. Figure 19 shows the spectra of DNA and BSA and Fig. 20 shows the distribution of protein and DNA in bull sperm, based on a series of images taken at different photon energies.

4.2 Environmental Sciences

The ability to study hydrated systems with high spatial resolution combined with the ability to map certain elemental and chemical properties of the sample makes x-ray techniques very valuable for environmental sciences. Notably, reactions that require a certain *p*H to occur cannot be studied with techniques that require introduction of the sample into vacuum.

Of great interest are, for instance, redox reactions that change the solubility of contaminants in our environment. Many species of bacteria can oxidize or reduce transition-metal ions, especially iron and manganese. Imaging with photon energies either above the absorption edges of the transition metal or for carbon-rich microorganisms in the water window provides specific contrast of the metal-containing particles or the organic material. Figure 21 shows x-ray microscope images

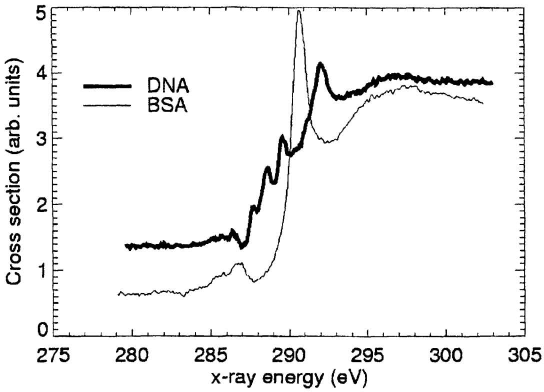

FIG. 19. X-ray absorption spectra of protein, bovine serum albumin (BSA) and DNA make it possible to map these components separately [adapted from Zhang *et al.* (1996), courtesy of R. Balhom].

taken in the same area of a sample, at 646 eV, where x rays are strongly absorbed by the manganese atoms in the needle-shaped particles of manganite (MnOOH), and at 517 eV where x rays are absorbed by the carbon atoms in the bacteria. Figure 22 shows the absorption spectrum of manganese at the L edge. The specific shape of this spectrum indicates that the sample predominantly contains Mn^{3+}. The ability to use spectral analysis illustrates the spatial arrangement of the reactant particles in a nearly natural hydrated state with unprecedented resolution.

The morphology of soil in its hydrated state is of genuine interest in the geosciences. Penetration of water through pores in the soil is responsible for many transport and reaction mechanisms of nutrients and toxicants. Techniques exist to characterize these processes for pore sizes down to about 10 µm. X-ray microscopy is used to study the porous inner structure of soils at a submicrometer scale (Niemeyer and Thieme 1998). Figure 23 shows the microstructure of a dystric cambisol as an example.

4.3 Materials Sciences

4.3.1 Chemistry, Civil Engineering X-ray microscopy has been applied to several problems important to our infrastructure and for industry.

4.3.2 Cement and Concrete In concrete, alkali-metal ions and hydroxyl ions contributed by the cement and reactive silicates present in aggregate can participate in a destructive alkali–silica reaction. This reaction of the alkalis with the silicates produces a gel that tends to imbibe water found in the concrete pores, leading to swelling of the gel and eventual cracking of the affected concrete member. Over 104 cases of alkali–aggregate reaction in dams and spillways have been reported around the world. At present, no method exists to arrest the expansive chemical reaction that generates significant distress in the affected structures. X-ray microscopy has been used to study these alkali–silica reactions *in situ* with high spatial resolution (Kurtis *et al.*, 1998) (Fig. 24).

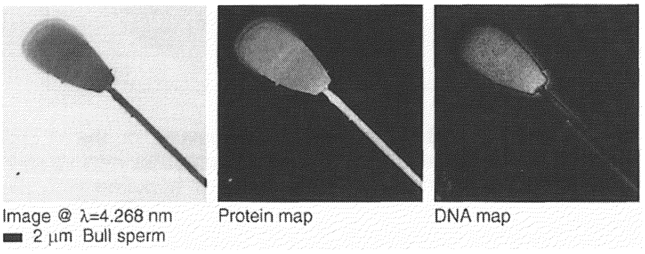

FIG. 20. X-ray microscopy image of a bull sperm taken at 290.5 eV and the distribution of protein (middle) and DNA (right) based on a series of images taken at different photon energies [from X. Zhang *et al.* (1996), courtesy of R. Balhom].

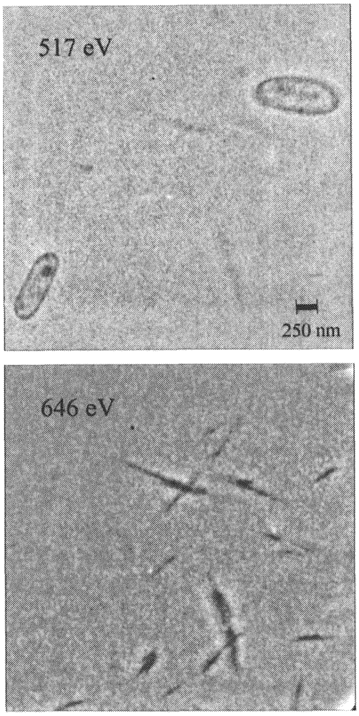

FIG. 21. Hydrated manganese particles and bacteria imaged in the water window (top) and above the manganese L edge (bottom). The images were taken with the transmission x-ray microscope XM-1 at the Advanced Light Source in Berkeley. (Courtesy of B. Tonner and K. Nealson.)

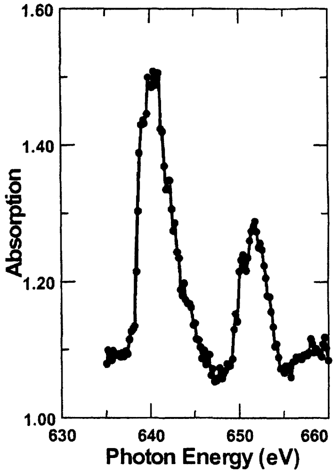

FIG. 22. Spectrum at the Mn L edge from manganese particles as shown in Fig. 21. This spectrum was taken at the STXM at beam line 7.0 of the Advanced Light Source in Berkeley. (Courtesy of B. Tonner.)

4.3.3 Polymers High–spatial-resolution near-edge spectroscopy at the carbon edge is of importance for polymer manufacturing industry. An early example of NEXAFS microscopy is shown in Figs. 25 and 3. The STXM at beam line X1A at the NSLS was used to image the phases from a polycarbonate–poly(ethylene terephthalate) (PC–PET) blend without resorting to heavy-element staining (Ade *et al.*, 1995). The chemical composition of the phases is selectively highlighted with small changes in photon energy. The observed contrast reflects the presence or absence of an electronic resonance between carbonyl groups and the phenylene in the polymers. The spectra of PC and PET as acquired with the X1A-STXM are shown in Fig. 3, along with the chemical structure of the respective polymers. Similar functional groups are present in these polymers, yet the spectra are strikingly different.

4.4 Future Applications

Though it is almost impossible to foresee future applications of x-ray microscopy, it is expected that the techniques described here will be used in a variety of fields not mentioned here. The semiconductor industry, for in-

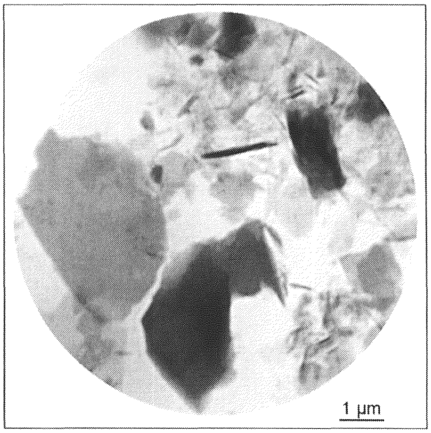

FIG. 23. X-ray microscopy image of the microstructure formed by colloidal particles within a dystric cambisol. This image was acquired with the Göttingen x-ray microscope at BESSY. [See also Niemeyer *et al.* (1998), courtesy of J. Thieme.]

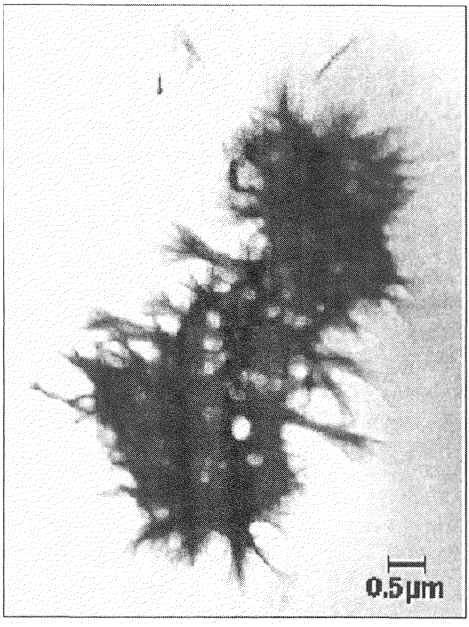

FIG. 24. Product of the reaction of the alkali–silica reaction gel in a saturated calcium hydroxide solution imaged *in situ* with the x-ray microscope XM-1 at the Advanced Light Source in Berkeley. [Adapted from Kurtis *et al.* (1998).]

stance, is developing new lithography techniques to make ever smaller chips. This will require the use of extreme-ultraviolet or soft–x-ray wavelength, as well as the assisting microscopy and microanalysis methods.

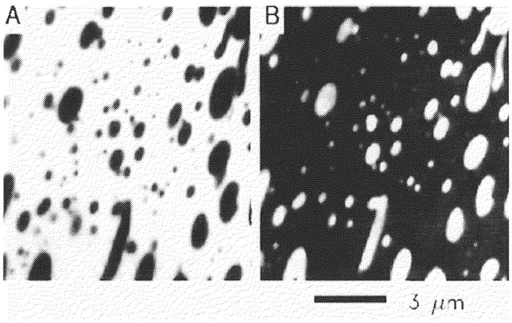

FIG. 25. Micrographs of PC–PET blend acquired at (A) 284.8- and (B) 285.1 eV photon energy. The wavelengths of the images shown here are also indicated in Fig. 3. [Adapted from Ade *et al.* (1995), courtesy of H. Ade.]

GLOSSARY

Areal Density: Mass per area, often measured in g cm^{-2}. Transmission experiments see the areal density.

Advanced Light Source (ALS): Synchrotron radiation facility at the Lawrence Berkeley National Laboratory, California, U.S.A.

BESSY: Synchrotron radiation facility in Berlin, Germany.

Electron Writing Tool: Instrument similar to a scanning electron microscope, which is used to write very small patterns for lithographic processing. Used to fabricate zone plates.

Fixation: Method to stabilize sensitive samples to prevent structural damage from processing or investigations. Chemical fixatives cross-link sample molecules. Cryogenic fixation means deep freezing the sample.

Kirkpatrick–Baez Mirror: Arrangement of two curved mirrors, used to focus x radiation (see X-RAY OPTICS).

Multilayer: Multiple layers of coating to enhance significantly the reflectivity of extreme ultraviolet and x radiation (see X-RAY OPTICS).

National Synchrotron Light Source (NSLS): Synchrotron radiation facility at Brookhaven National Laboratory, Long Island, New York, U.S.A.

NEXAFS: Near-edge x-ray–absorption fine structure. Spectroscopic method to determine chemical properties of the sample.

SPEM: Scanning photoemission microscope (see XPS).

Synchrotron Radiation: Electromagnetic radiation with very high brightness, produced by accelerated charged particles, often electrons (see also SYNCHROTRON RADIATION).

Water Window: X rays of photon energies between the carbon and oxygen K-shell absorption edges. Those x rays provide natural contrast from hydrated organic materials.

XANES: X-ray absorption near-edge spectroscopy (see NEXAFS).

XPS: X-ray photoemission spectroscopy. Spectroscopic analysis of photoelectrons generated by x-ray photon excitation.

Zone-plate Lens: Circular grating structure that acts like a lens, bases on diffraction. Zone plates are used for highest–spatial-resolution x-ray microscopes.

Works Cited

Ade, H., Zang, X., Cameron, S., Costello, C., Kirz, J., Williams, S. (1992), *Science* **258**, 972–975.

Ade, H., Hsiao, B. (1993), *Science* **262**, 1427–1429.

Ade, H., Smith, A., Cameron, S., Cieslinski, R., Costello, C., Hsiao, B., Mitchell, G., Rightor, E. (1995), *Polymer* **36**, 1843–1848.

Anderson, E. H., Kern, D. (1992), in: A. Michette, G. R., Morrison, C. J. Buckley (Eds.), *X-Ray Microscopy III*, Heidelberg: Springer, pp. 75–78.

Anderson, E. H., Boegli, V., Muray, L. P. (1995), *J. Vac. Sci. Technol. B*, **13**, 2529–2534.

Aristov, V. V., Erko, A. I. (Eds.) (1994), *X-Ray Microscopy IV*, Chernogolovka, Russia: Bogorodskii Pechatnik.

Baez, A. V. (1952), *J. Opt. Soc. Am.* **42**, 756–762.

Born, M., Wolf, E. (1980), *Principles of Optics*, Oxford: Pergamon.

Chapman, H., Fu, J., Jacobsen, C., Williams, S. (1996), *J. Microsc. Soc. Am.* **2**, 53.

Ray-Chaudhuri, A. K., Ng, W., Liang, S., Singh, S., Welnak, J. T., Wallace, J. P., Capasso, C., Cerrina, F., Margaritondo, G., Underwood, J. H., Kortright, J. B., Perera, R. C. C. (1993), *J. Vac. Sci. Technol*, **A 11**(4), 2324.

Coslett, V. E., Nixon, W. C. (1960), *X-Ray Microscopy*, Cambridge: Cambridge University Press.

DaSilva, L. B., Trebes, J. E., Balhorn, R., Mrowka, S., Anderson, E., Attwood, D. T., Barbee, Jr., T. W., Brase, J., Corzett, M., Gray, J., Koch, J. A., Lee, C., Kern, D., London, R. A., MacGowan, B. J., Matthews, D. L., Stone, G. (1992), *Science* **258**, 269–271.

Henke, B. L., Gullikson, E. M., Davis, J. C. (1993), *At. Data Nucl. Data Tables* **54**, 181–342. See also http://www-cxro/lbl.gov/optical_constants.

Irtel von Brenndorff, A., Moronne, M. M., Larabell, C., Selvin, P., Meyer-Ilse, W. (1994), in: V. V., Aristov, A. I. Erko (Eds.) *X-Ray Microscopy IV*, Chernogolovka, Russia: Bogorodskii Pechatnik, pp. 338–344.

Jacobsen, C., Lindaas, S., Williams, S., Zhang, X. (1993), *J. Microsc.* **172**, 121–129.

Jochum, L., Meyer-Ilse, W. (1995), *Appl. Opt.* **34**, 4944–4950.

Kirkpatrick, A., Baez, V. J. (1948), *J. Opt. Soc. Am.* **38**, 766–774.

Kirz, J. (1974), *J. Opt. Soc. Am.* **64**, 301–309.

Lehr, J. (1997), *Optik* **104**, 166–170.

Lindaas, S., Howells, M., Pinkas, D., Magowan, C., Downing, K., Jacobsen, C. (1998), in: J. Thieme, G. Schmahl, E. Umbach, D. Rudolph (Eds.), *X-Ray Microscopy and Spectromicroscopy*, Heidelberg: Springer.

Kurtis, K., Monteiro, P., Brown, J. T., Meyer-Ilse, W. (1998), *Cem. Concr. Res.* **28**, 411–421.

Magowan, C., Brown, J. T., Liang, J., Heck, J., Coppel, R. L., Narla, M., Meyer-Ilse, W. (1997), *Proc. Natl. Acad. Sci.* **94**, 6222–6227.

Maser, J., Jacobsen, C., Osanna, A., Wang, S., Kalinovsky, A., Kirz, J., Spector, S., Warnking, J. (1998), in: J. Thieme, G. Schmahl, E. Umbach, D. Rudolph (Eds.), *X-Ray Microscopy and Spectromicroscopy*, Heidelberg: Springer, pp. I-35–I-44.

Michette, A., Morrison, G. R., Buckley, C. J. (Eds.) (1992), *X-Ray Microscopy III*, Heidelberg: Springer.

Niemann, B., Rudolph, D., Schmahl, G. (1976), *Appl. Opt.* **15**, 1883.

Niemann, B., Rudolph, D., Schmahl, G., Diehl, M., Thieme, J., Meyer-Ilse, W., Neff, W., Holz, R., Lebert, R., Richter, F., Herziger, G. (1990), *Optik* **84**, 35–36.

Niemeyer J., Thieme J., Machulla, G. (1998), in: J. Thieme, G. Schmahl, E. Umbach, D. Rudolph (Eds.), *X-Ray Microscopy and Spectromicroscopy*, Heidelberg: Springer, pp. II-3–II-10.

Rudolph, D., Niemann, B., Schmahl, G., Christ, O. (1984), in: G. Schmahl, D. Rudolph (Eds.), *X-Ray Microscopy*, Heidelberg: Springer, pp. 192–202.

Niemann, B., Schneider, G., Guttmann, P., Rudolph, D., Schmahl, G. (1994), in: Aristov, V. V., Erko, A. I. (Eds.), *X-Ray Microscopy IV*, Chernogolovka, Russia: Bogorodskii Pechatnik, pp. 66–75.

Sayre, D., Kirz, J., Feder, R., Kim, D. M., Spiller, E. (1997), *Ultramicroscopy*, **2**, 337–349.

Sayre, D., Howells, M., Kirz, J., Rarback, H. (Eds.) (1988), *X-Ray Microscopy II*, Heidelberg: Springer.

Schmahl, G., Rudolph, D. (1969), *Optik* **29**, 557–585 (in German).

Schmahl, G., Rudolph, D. (Eds.) (1984), *X-Ray Microscopy*, Heidelberg: Springer.

Schmahl, G., Rudolph, D., Niemann, B., Guttmann, P., Thieme, J., Schneider, G., David, C., Diehl, M., Wilhein, T. (1993), *Optik* **93**, 95–102.

Schmahl, G., Guttmann, P., Raasch, D., Fischer, P., Schütz, G. (1996), *Synchro. Rad. News* **9**, 35–39.

Schneider, G., Schliebe, T., Aschoff, H. (1995), *J. Vac. Sci. Technol. B* **13**, 2809.

Schneider, G., Schmahl, G., Schliebe, T., Peuker, M., Guttmann, P. (1998), in: J. Thieme, G. Schmahl, E. Umbach, D. Rudolph (Eds.), *X-Ray Microscopy and Spectromicroscopy*, Heidelberg: Springer, pp. I-111–I-116.

Spector, S. J., Jacobsen, C. J., Tennant, D. M. (1997), *J. Vac. Sci. Technol. B*, **15**, 2872–2876.

Thieme, J., Schmahl, G., Umbach, E., Rudolph, D. (Eds.) (1998), *X-ray Microscopy and Spectromicroscopy*, Heidelberg: Springer.

Underwood, J. H., Thompson, A. C. Wu, Y., Giauque, R. D. (1988), *Nucl. Instrum Meth. Phys. Res.* **A 266**, 318–323.

Underwood, J. H., Thompson, A. C., Kortright, J. B., Chapman, K. C., Lunt, D. (1996), *Rev. Sci. Instrum.* **67** (9), 1–5, (available on CD-ROM only).

Warwick, T., Ade, H., Cesari, S., Denlinger, J., Frank, K., Garcia, A., Hayakawa, S., Hitchcock, A., Kikuma, J., Klingler, S., Kortright, J., Morrison, G., Moronne, M., Rightor, E., Rotenberg, E., Seal, S., Shin, H.-J., Steele, W., Tonner, B. P. (1998), *J. Synchrotron Rad.* **5**, 1090–1092.

Yun, W. B., Viccaro, P. J., Chrzas, J., Lai, B. (1992), *Rev. Sci. Instrum.* **63**, 582–585.

Zhang, X., Balhorn, R., Mazrimas, J., Kirz, J. (1996), *J. Struct. Biol.* **116**, 335–344.

Further Reading

Kirz, J., Jacobson, C., Howells, M. (1995), *Quart. Rev. Biophys.* **28**, 33–130.

Michette, A. G. (1986), *Optical Systems for Soft X-rays*, New York: Plenum Press.

ACOUSTICAL TOMOGRAPHY

(An Addendum to *Encyclopedia of Applied Physics*, Volume 1, pages 89–108.)

Bruce M. Howe and Daniel Rouseff, *Applied Physics Laboratory, College of Ocean and Fishery Sciences, University of Washington, Seattle, Washington, U.S.A.*

Introduction 345

1. **Nonlinear Diffraction Tomography** 345

2. **Vector Tomography** 346

3. **Applications of Acoustical Tomography** 347

3.1 Ocean Tomography 347

3.2 Sediment Tomography 349

3.3 Atmospheric Tomography 349

3.4 Seismic Tomography 349

4. **Concluding Remarks** 349

Works Cited 349

Further Reading 350

INTRODUCTION

Since the publication of our original article, several noteworthy advances have occurred in acoustical tomography. Improvements in computer technology have made nonlinear diffraction tomography an active research topic. A theoretical foundation has been developed for using acoustical tomography to reconstruct vector quantities. Acoustical tomography has become a mature technology for studying deep-water oceanography. Tomography experiments have produced new insights into deep convection, internal tides, large-scale circulation, and basin-scale temperature variability. More recently, techniques have been developed to study both the water column and the sediment layers in shallow water. Acoustical tomography has been demonstrated in the atmosphere for measuring near-surface temperature and wind velocity.

1. NONLINEAR DIFFRACTION TOMOGRAPHY

In acoustical tomography, a medium of interest is ensonified from several different directions, and the acoustic field is measured for each view. By processing and combining the data in an inversion algorithm, an image of the medium is developed. The inversion algorithm will necessarily depend on a model for the interaction between the sound and the medium. If the model is based on an acoustic wave equation, the process of making the measurements and producing an image is called diffraction tomography.

Mathematically, the scattered field $\psi_S(\mathbf{x})$ is related to the scattering function $F(\mathbf{x})$ describing the medium by an integration over the scattering volume V. Equation (4) from the original article is repeated here:

$$\psi_S(\mathbf{x}) = \int_V F(\mathbf{x}')\psi(\mathbf{x}')G_f(\mathbf{x},\mathbf{x}')\,d\mathbf{x}'.$$

Here, G is the free-space Green's function, or impulse response, that maps scattering from a point $\mathbf{x}'$ in the volume to the observation point $\mathbf{x}$. The total field $\psi(\mathbf{x}')$ is the sum of the known incident field and the unknown scattered field $\psi_S(\mathbf{x}')$. Consequently, the equation is nonlinear; the scattered field appears in a complicated way on both sides of the equal sign. Most algorithms used in diffraction tomography make simplifying approximations to linearize the equation. In the Born approximation, the unknown total field inside the integral is replaced by the known incident field. This is a weak-scattering approximation equivalent to ignoring multiple scattering. Strong scattering is the norm, however, in medical and many other prospective applications. Consequently, tomographic algorithms based on the linearizing Born approximation often produce unsatisfactory results.

In recent years, researchers have proposed various ways to take diffraction tomography beyond weak-scattering approximations. Since

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

linearizing approximations are avoided, the methods are often called *nonlinear diffraction tomography*. In one way or another, these methods rely on numerical solutions to the scattering integral. These methods are usually iterative. Typically, an initial guess is made for the scattering function and the internal field. The integral is then solved numerically for the scattered field that would be measured at the receivers. The difference between the calculation and the observation is used to update the estimates. The procedure is then repeated until the error is below a specified tolerance.

Clearly, nonlinear diffraction tomography is a computationally intensive procedure. At each iteration, the forward-scattering calculation must be repeated. The information from all scattering directions, from all incident field directions, and at all frequencies must be included. When the dimension of the scattering regime is many acoustic wavelengths, the resulting system of equations that must be solved at each iteration can be enormous. Even then, there is no guarantee of convergence.

Borup *et al.* (1992) examined numerical solutions to the scattering integral based on the method of moments. Chew *et al.* (1996) reviewed several other nonlinear diffraction tomography algorithms. These include the distorted Born iterative method. In this procedure, the background medium is not constrained to be homogeneous. This means that a simple analytical representation for the Green's function is no longer possible. A numerical representation is required and must be updated at each iteration. Chew *et al.* (1996) report that the method has improved convergence and can handle objects with a large acoustic contrast.

Nonlinear diffraction tomography remains the subject of active research. As computational power continues to increase, and as more sophisticated algorithms are developed, it is likely that nonlinear diffraction tomography will have a significant influence on measurement technology.

2. VECTOR TOMOGRAPHY

Sound traveling through a moving fluid will be affected by the motion. Recently, tomographic methods have been proposed for characterizing fluid flow. Since the fluid motion has both a local magnitude and a local direction, tomographic methods have been generalized to reconstruct vector fields. Attention is restricted here to cases where the propagation paths are sufficiently short to allow neglect of refraction. This is typically the case in industrial or small-scale oceanographic applications.

Tomographic methods for characterizing fluid flow operate primarily in the transmission mode. As usual, the acoustic travel time from source to receiver will depend on the scalar sound-speed perturbation $\delta C(\mathbf{x})$ of the intervening fluid.In addition, the travel time is also influenced by fluid motion, which can be described by the vector flow velocity field $\mathbf{u}(\mathbf{x})$. The travel time of an acoustic ray propagating along path $\boldsymbol{\tau}$ is a function of $\mathbf{u} \cdot \boldsymbol{\tau}$ integrated along the propagation path. To separate the contribution of fluid flow to travel time from that of sound speed, a reciprocal transmission strategy is used. The travel time is measured along a given path in both directions. Averaging the two measurements removes the effects of fluid motion and yields an estimate of the sound speed. Consequently, when refraction is negligible, the reconstruction of $\delta C(\mathbf{x})$ for a moving fluid reduces to the classic tomography problem detailed in the original article. The projection-slice theorem relates the measurements to the unknown field. The averaged travel-time data from several different directions are combined using the filtered back-projection algorithm to generate the reconstruction.

Taking the difference of the measurements isolates a projection of $\mathbf{u} \cdot \boldsymbol{\tau}$. In contrast to the scalar problem, the reconstruction of vector fields has received sustained attention only in recent years. Initially, the differential travel time was fed directly into the standard reconstruction algorithm. The resulting images, however, were incomplete because components of the flow were invisible. The difficulty is that the differential travel time is sensitive only to that component of the flow vector that lies along the acoustic ray path. Transverse components of the flow are unsampled.

Norton (1988) put vector inversion on a firmer theoretical footing. By means of a vector projection-slice theorem, he showed that only the divergence-free (solenoidal) component of a two-dimensional vector field could be recovered from measured projections. This still gives considerable information about the flow.

The solenoidal component is directly related to the fluid vorticity, itself an important quantity in many areas of fluid dynamics. Instabilities that cause the transition from laminar flow to turbulent flow often exhibit characteristic vorticity structure. A fluid traveling through closely spaced elbows in a pipe, for example, can acquire an unwanted swirling component of motion that can be described in terms of vorticity. Winters and Rouseff (1990) developed a filtered back-projection algorithm for vorticity. Braun and Hauck (1991) successfully imaged flow fields in laboratory experiments.

Rouseff and Winters (1994) proposed using linear diffraction tomography to image flow fields. An appropriately modified wave equation was used to describe acoustic propagation through the flow. They showed that with diffraction tomography it was possible to recover a low-pass filtered version of the complete vector field $\mathbf{u}(\mathbf{x})$. This illustrates an important concept for inverse problems in general and acoustical tomography in particular: to generate a satisfactory reconstruction, it is imperative to include the proper physics in the forward propagation model. To reconstruct the full two-dimensional vector velocity, a ray-based forward model is inadequate whereas a wave-based model provides a more complete inversion.

3. APPLICATIONS OF ACOUSTICAL TOMOGRAPHY

Strictly speaking, most acoustical imaging systems currently used in medicine and nondestructive testing are not tomographic. These systems usually operate in the pulse-echo mode and do not coherently combine data taken from several different directions to form an image. Computer capability and often the underlying science are still not sufficiently advanced to permit practical tomographic imaging systems to be built. Acoustical tomography has found utility primarily in oceanography.

3.1 Ocean Tomography

The results of several recent ocean tomography experiments are summarized here. A review of this field has recently been published by Munk *et al.* (1995).

In the Greenland Sea tomography experiment (1987–1988) the time-dependent three-dimensional ocean temperature was mapped using the tomographic technique. The forward model used to describe the acoustic propagation was based on mode theory. This was the first time that the travel times of acoustic normal modes were measured and used in the inversion process. Given the weakly range-dependent (adiabatic) sound-speed profile, late-arriving acoustic modes sample the near-surface ocean whereas rays sample the deeper ocean. The inversions incorporated point and profile measurements as well as the acoustic measurements. The results provided the first four-dimensional view of oceanic deep convection (Morawitz *et al.*, 1996).

In the Acoustic Mid-Ocean Dynamics Experiment—Moving Ship Tomography (AMODE-MST) project, six transceiver moorings were installed in a pentagonal array south of Bermuda in the Sargasso Sea, where they then "talked" to each other over ranges of 300 to 700 km. During 50 days of the 1-yr experiment, a ship circumnavigated the moored array two and a half times, lowering a vertical receiving array every 3 hours and every 25 km. Although laborious, this provided the large number of ray paths crossing at a multitude of angles required for high-resolution imaging (cf. the projection-slice theorem). One resulting map over a $1000 \times 1000\text{-km}^2$ area is shown in Fig. 1 (AMODE-MST Group, 1994). Because the ocean was changing during this time, the inversion must be explicitly cast in a time-dependent model.

The Acoustic Thermometry of Ocean Climate (ATOC) project has been a major new effort. The principal idea behind ATOC is to use basin-scale acoustic transmissions to measure large-scale variability in ocean temperature (heat content). A prerequisite to doing this is to understand the forward problem. To ranges of 5 Mm, it has been determined that steep ray arrivals are stable, resolvable, and identifiable with predicted rays. The shallow-angle energy described by acoustic modes is more affected by nonlinear mode coupling caused by scattering from ocean internal waves; this has so far made the interpretation of the data difficult. However, the final (lowest mode) cutoff arrival does appear to be stable. Furthermore, a recent experiment using a 28-Hz source (ATOC uses a 75-Hz source) shows

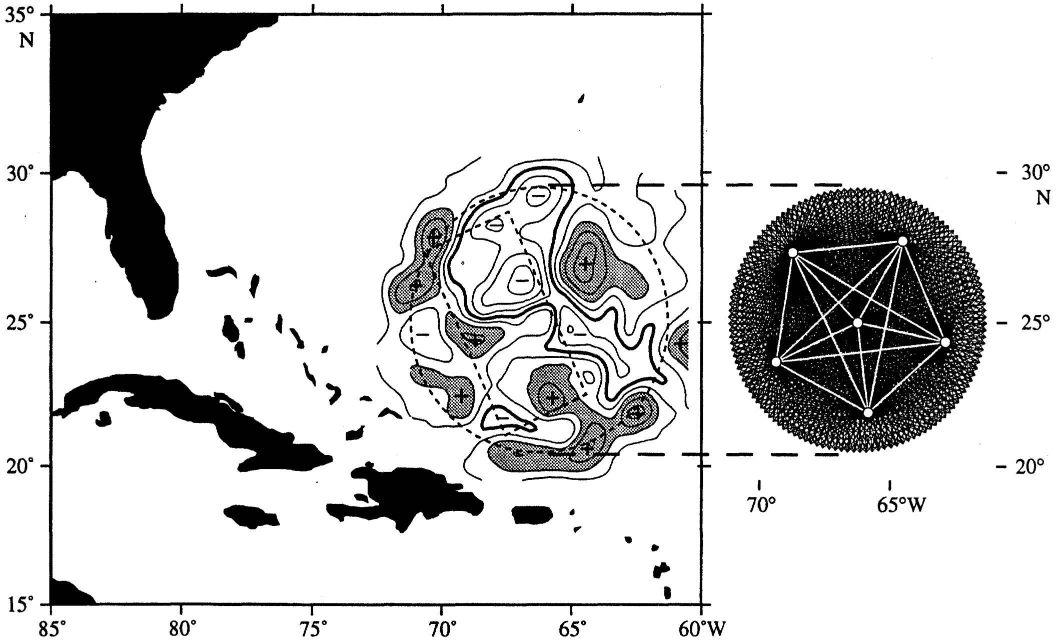

FIG. 1. Sound-speed perturbations from the climatological mean in 1-m/s (0.25 °C) contours at 700-m depth, as measured during the Moving Ship Tomography experiment, 15–30 July, 1991 (AMODE-MST Group, 1994). Five of the moorings are in a pentagon 700 km in diameter, with the sixth mooring in the center. The receiving ship moved around a 1-Mm circle, making 135 receiving stops.Rays from each of the moorings to the receiving stops are shown on the right. The dashed rectangle shows the area surveyed with air-expendable temperature sensors for comparison purposes; the two methods agreed within the expected errors (from Munk et al., 1995; © 1995, Cambridge University Press).

a significant improvement in the acoustic arrival structure, because the lower-frequency, longer-wavelength sound is less affected by the internal waves. The routine use of this and even lower frequencies awaits the development of a more practical transducer.

An acoustic source on Pioneer Seamount off California has been transmitting since the beginning of 1996 to various receivers around the Pacific Basin at ranges up to 10^4 km. The travel times change by as much as 1 s, with clear seasonal cycles (Howe *et al.*, 1996).

A similar, but briefer, experiment was carried out in the Arctic, with a source transmitting to several receivers around the Arctic Basin (Mikhalevsky *et al.*, 1995). As in the Greenland Sea work, the data are most naturally interpreted in terms of acoustic modes. A clear change in mode travel times relative to those that were predicted by use of climatological data indicated a significant warming, which was later verified by intensive *in situ* temperature measurements.

There have recently been several shallow-water tomography experiments. These must by necessity take into account bottom interaction, and therefore the data are much more difficult to interpret; see for instance Jin *et al.* (1996). A moving-source tomography experiment was recently conducted over 2-km ranges in the Haro Straits (Bellingham *et al.*, 1996). Velocity over ranges of 7 km has been determined in water 70 m deep by means of reciprocal transmissions (Kaneko *et al.*, 1996).

Send *et al.* (1995) describe an experiment to measure deep convection in the Mediterranean and are working toward establishing a permanent monitoring system. Groups in Japan are developing new hardware as well as using commercially available systems to begin work off Japan, in the Philippine Sea, and in the equatorial Pacific.

3.2 Sediment Tomography

Tomographic methods have been applied to study the properties of ocean sediment. Yamamoto (1995) adapted seismic cross-borehole techniques for use with propagation paths on the order of hundreds of meters to infer sediment properties like density, porosity, and shear strength. To image smaller scales, Tang *et al.* (1996) developed a U-shaped frame, each leg 1 m long, with 20 transceiver elements on each leg. The two legs are inserted into the sediment and transmission measurements are made. These applications have, to date, been limited to two-dimensional (single-slice) geometries.

3.3 Atmospheric Tomography

In the atmosphere, acoustic tomography has recently been demonstrated by Wilson and Thomson (1994) as a method of measuring the near-surface temperature and wind velocity. Attenuation limits the possible acoustic frequency bands and the range. Their demonstration system consisted of three sources and four receivers with a maximum range of 240 m (their high signal-to-noise ratio implies that longer ranges are possible). An impediment to further development and routine use of such a system is the lack of robust software to pick and track ray-arrival peaks. (This is a problem shared with ocean tomography, but in the latter case, there are presently no requirements for real-time processing.)

3.4 Seismic Tomography

Seismology is concerned with determining the mechanical properties and the dynamics of the interior of the Earth. Because we are limited to only (near) surface measurements of Earth motion, tomography is a natural technique for using the available data to infer the structure of the Earth's interior. Seismic tomography is based on the analysis of body (compressional P and shear S) waves, surface (Rayleigh and Love) waves, and free oscillations (normal modes) as typically recorded on seismographs. The P waves are the "usual" acoustic waves, and thus the connection to acoustical tomography; the presence of the other types of waves complicates (even more) the seismic forward and inverse problems. The literature on seismology and seismic tomography is vast; only a few references are provided here to guide the reader (see also SEISMOLOGY). Nolet (1990) showed how the partitioning of the inverse problem into linear and nonlinear steps can greatly facilitate the interpretation of the large data sets that are commonly used. This work has been extended to include diffraction (Meier *et al.*, 1997) and mode coupling (Marquering *et al.*, 1996). The use of free oscillations to determine the largest scales of the Earth's interior is described by Masters and Widmer (1995).

4. CONCLUDING REMARKS

Many prospective applications of tomography in medical and nondestructive testing require accurate and efficient techniques for modeling the forward problem in highly diffractive or strongly scattering regimes. Progress is being made toward solving this problem, with the most gain expected to come from the ever-increasing computational capability. Deep-water ocean tomography is now a mature technique; applications involving bottom interaction and long ranges are probing the limits of the method. Atmospheric tomography has been demonstrated, but further implementation awaits an efficient method of tracking ray-arrival peaks. It is expected that acoustical tomography will prove to be an elegant and efficient method of measuring ocean sediments in a minimally intrusive manner, particularly at smaller scales. Lastly, acoustical tomography will most likely benefit from the theoretical and numerical development being made in seismic tomography.

List of Works Cited

AMODE-MST Group (1994), *Eos Trans. AGU* **75**, 71, 21, 23.

Bellingham, J. G., Schmidt, H., Deffenbaugh, M.(1996), *J. Acoust. Soc. Am. Suppl.* **100**, 2612.

Borup, D. T., Johnson, S. A., Kim, W. W., Berggren, M. J. (1992), *Ultrason. Imaging* **14**, 68–85.

Braun, H., Hauck, A. (1991), *IEEE Trans. Sig. Proc.* **39**, 464–471.

Chew, W. C., Otto, G. P., Weedon, W. H., Lin, J. H., Lu, C. C., Wang, Y. M., Moghaddam, M. (1996), *Int. J. Imaging Sys. Tech.* **7**, 16–24.

Howe, B. M., Dushaw, B. D., Mercer, J. A., Spindel, R. C., Worcester, P. F., Cornuelle, B. D., Colosi, J. A., Dzieciuch, M. A., Munk, W. H., Metzger, K. (1996), *J. Acoust. Soc. Am. Suppl.* **100**, 2581.

Jin, G., Lynch, J. F., Chiu, C. S., Miller, J. H. (1996), *J. Acoust. Soc. Am.* **100**, 193–205.

Kaneko, A., Zheng, H., Noguchi, H. (1996), *J. Acoust. Soc. Am. Suppl.* **100**, 2565.

Marquering, H., Sneider, R., Nolet, G. (1996), *Geophys. J. Int.* **124**, 258–278.

Masters, T. G., Widmer, R. (1995), "Free Oscillations: Frequencies and Attenuations," in: T. J. Ahrens (Ed.), *Global Earth Physics: A Handbook of Physical Constants*, Washington, D.C.: American Geophysical Union.

Meier, T., Lebedev, S., Nolet, G., Dahlen, F. A. (1997), *J. Geophys. Res.* **102**, 8255–8267.

Mikhalevsky, P. N., Garilov, A., Baggeroer, A. B. (1995), *J. Acoust. Soc. Am. Suppl.* **97**, 3234.

Morawitz, W. M. L., Cornuelle, B. D., Worcester, P. F. (1996), *J. Atmos. Oceanic Technol.* **13**, 659–679.

Munk, W., Worcester, P., Wunsch, C. (1995). *Ocean Acoustic Tomography*, New York: Cambridge University Press.

Nolet, G. (1990), *J. Geophys. Res.* **95**, 8499–8512.

Norton, S. J. (1988), *Geophys. J. Royal Astron. Soc.* **97**, 162–168.

Rouseff, D., Winters, K. B. (1994), *Inverse Problems* **10**, 687–697.

Send, U., Schott, F., Gaillard, F., Desaubies, Y. (1995), *J. Geophys. Res.* **100**, 6927–6941.

Tang, D. J., Chu, D., Auston, T. (1996), *J. Acoust. Soc. Am. Suppl.* **100**, 2712.

Wilson, D. K., Thomson, D. W. (1994), *J. Atmos. Oceanic Technol.* **11**, 751–769.

Winters, K. B., Rouseff, D. (1990), *Inverse Problems* **6**, L33–38.

Yamamoto, T. (1995), *J. Acoust. Soc. Am.* **98**, 2235–2248.

Further Reading

Munk, W., Worcester, P., Wunsch, C. (1995), *Ocean Acoustic Tomography*, New York: Cambridge University Press.

COMPUTER DATABASES

(An Addendum to *Encyclopedia of Applied Physics*, Volume 4, pages 103–126.)

H. BEHRENS AND J. W. TEPEL, *Fachinformationszentrum Karlsruhe, Eggenstein-Leopoldshafen, Germany*

1. Recent Developments 351
1.1 Technological Advances 351
1.2 Databases 352
2. Databases and the Internet 352
3. Preprint Servers and Electronic Journals 353
Works Cited 353

1. RECENT DEVELOPMENTS

Since the original article COMPUTER DATABASES was published, significant advances have been made in computer technology. These include the availability of better hardware at lower cost and the appearance of new and more sophisticated software. The overwhelming success of the World-Wide Web (WWW) has revolutionized the way we do business in information. In addition, some changes have occurred as far as the databases described in the original article are concerned.

1.1 Technological Advances

The dramatic changes which have taken place in the distribution, transmission, and access of data of all kind in the last five years were driven by technological advances. In the first instance, the price of mass storage devices for personal computers, work stations, and even mainframes all but collapsed. Whereas personal computers in the beginning of the nineties were fitted with hard discs of about 200 MB, nowadays PCs "off the shelf" have capacities in excess of 1 or 2 GB. Secondly, the prices of random-access memory (RAM) storage chips have dropped considerably. Nowadays, a standard PC is fitted with 16 MB or more of RAM memory compared to 2 or 4 MB a few years ago. Thirdly, hand in hand with these developments, microprocessor speeds increased tremendously—doubling in about every 18 months.

Needless to say, software expanded to these new realities. New software packages invariably use up more hard disc space and need more RAM memory and higher processor speeds in order to perform satisfactorily. Fortunately, their functionality has increased too. Terminals connected to mainframe computers or work stations have been replaced by intelligent front ends (clients) with graphical user interfaces (GUIs). Lately, the idea of "network computers" without hard discs (NCs) was born, with more limited intelligence but running software downloaded from the Internet or a central in-house platform. This greatly facilitates software maintenance, since only a single copy of some product has to be kept up to date.

A few years ago, CD-ROMs were considered somewhat exotic. Drives were slow, and the availability of software or data collections on CDs was low. In the meantime, software and data on CDs have become the rule rather than the exception. It has become possible to produce CDs in small quantities by recording or "burning" them directly in low-cost devices. Master copies can be produced inexpensively for pressing CDs on a large scale.

Furthermore, external magneto-optical storage devices scarcely larger than the well-known $3\frac{1}{2}$-in. magnetic discs have appeared, with storage capacities ranging from 100 MB to 1 GB and good accession times. These new discs have come in handy for information exchange and as storage media for data backups. Being direct-access devices, they have some advantages over the traditional magnetic tape streamers.

Perhaps the most spectacular development happened in the area of communication networks. This is related to the explosive development of the Internet, which consists of a set of linked networks talking the same "language", viz. TCP/IP, the Transmission Control Protocol/Internet Protocol, and its technologies. The Internet is conquering the world

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

and is now commonly used for disseminating and accessing data from places all over the world. This global information infrastructure that has been realized by the creation of the Internet has resulted in a revolution of information behavior in the 1990s. Far from making the traditional hosts superfluous, the Internet offers other and numerous new information possibilities. More and more of the traditional information providers are accessible via the Internet. Almost all researchers now have access to the Internet and are using it daily. In the form of the Intranet the Internet technology is also applied to in-house networks.

1.2 Databases

With the end of 1994 the production of the PHYS database was stopped. The PHYS database was merged with the INSPEC database, resulting in a high-quality database containing all areas of physics and the additional subject fields of electrical engineering, electronics, control, computing, and information technology. The physics subset of INSPEC is now cooperatively produced by the Institution of Electrical Engineers (IEE) and Fachinformationszentrum (FIZ) Karlsruhe. The dissemination of the printed abstract journal *Physics Briefs/Physikalische Berichte*, the content of which was based on the electronic database PHYS, was also terminated and incorporated into the abstract journal *Physics Abstracts*, which is the printed analog of INSPEC for physics. The total number of citations in INSPEC as of September 1996 was about 5.3×10^6 (50% to 60% physics).

The database C13 NMR/IR is not offered anymore. It has been replaced by the database SpecInfo which contains the same data, but in addition more and other types of spectral data as well as more and better retrieval components. It is distributed as an in-house version by Chemical Concepts and online via STN International.

2. DATABASES AND THE INTERNET

The Internet and its associated services, i.e. electronic mail (e-mail), file-transfer protocol (FTP), and in particular the world-wide web (WWW), have revolutionized communications with the outside world. Under the HyperText Transfer Protocol (HTTP), multimedia documents written in HTML, the HyperText Markup Language, are retrieved from special servers (web servers) and displayed locally by use of a web browser such as Netscape. With knowledge of the symbolic address of a web server, or more precisely the so-called URL or Uniform Resource Locator of its HTML files, information may be requested and received world-wide. For example, FIZ Karlsruhe can be reached via the symbolic address http://WWW.fiz-karlsruhe.de.

HTML documents are static documents stored as files on host computers. However, web pages can be made interactive with the user in a number of different ways. First, most HTML documents contain references or links (URLs) to other documents. Selecting highlighted sections of text on a web page is sufficient to obtain access to these documents, which on their own may reside at any location within the Internet. Following these links a user can "surf" through the Internet, starting for example with the Library of Congress and ending in the Vatican while looking at pictures by some famous painter. Second, web pages of HTML documents can be made dynamic with the aid of the CGI, the Common Gateway Interface, which allows user programs to be executed on the web server itself. Third, Plug-Ins and/or Java Applets may be downloaded from the web site and executed on the client computer itself. In this manner web pages can be animated—for example, by rotating figures and graphical elements.

Search engines such as Lycos or Hotbot may be used to find information in the WWW by matching one or more search strings entered by the user. The retrieved documents are ranked according to the degree of overlap with the search terms.

Retrieval of data from centralized databases via the WWW needs HTML forms in which the user can type his requests, and some process on the web server to parse the ASCII strings and pass them through to a database server. The answers from the database have to be passed back via the web server to the end user, who sees the result in the form of a dynamically generated web page. The technical realization of such a system encompasses a number of points that cannot be treated here.

Table 1. Journals on-line.

Journal	URL
Nature	http://www.nature.com
Nuclear Physics Electronic	http://www.nucphys.nl/www/pub/nucphys/npe.html
Physical Review Letters	http://prlo.aps.org/
Applied Physics Letters	http://ojps.aip.org/aplo/
Physical Review C	http://publish.aps.org/
Physical Review D	http://publish.aps.org/

Needless to say, the overall performance loss of such a system with many intervening and interdependent processes can be quite significant when compared to classical in-house database management systems. The situation is improved when the web and database servers are combined in a single unit as, for example, in the Oracle WebSystem. In addition to delays in treating database requests at the host site, the Internet itself may be the cause of poor response times, particularly at certain peak hours of the day. These apparent disadvantages are more than offset by the flexibility and availability of the WWW, so that the number of database providers within the WWW is expected to increase enormously within the next few years. Access to large online databases has recently been demonstrated by STN-International with STN-Easy and EI-Village. Here HTML forms are used for requesting and retrieving bibliographic and other information.

3. PREPRINT SERVERS AND ELECTRONIC JOURNALS

Most physicists are using computers at work today, for writing articles and for communication. The progress in network technology now facilitates the transmission of full-text documents including mathematical formulas, figures, and graphs. Thus, it is no problem to distribute preprints electronically all over the world today. The inexpensive disc storage allows also storage of all these preprints at one place. The first example of a wide usage in physics was the Los Alamos preprint server, which was established by Paul Ginsparg at Los Alamos in 1991 (see Taubes, 1993). The physicist submits papers electronically by e-mail, via FTP, or by using WWW uploaded pages in a prescribed specific format (e.g. TEX) to this server; the system automatically files these papers and sends out abstracts on demand or to a list of subscribers. The physicist can then scan the abstracts, decide which documents are of interest, and download them or require that the preprints be sent to him. This server system is located within the WWW, so that the searching and browsing capabilities of the WWW can be applied. At present examples of such preprint servers are the Los Alamos e-print archive, URL = http://xxx.lanl.gov/, and the CERN Preprint Server, URL = http://preprints.cern.ch/.

This preprint database system allows scientists to publish their results without having to go through a peer review and publication process. This has given the term "desktop publishing" a completely new meaning.

In the meantime also a number of publishers have placed their journals on line with full text and graphics and are offering their printed journals also in electronic form as so-called electronic journals (see Durrani, 1996). In most cases this is done in parallel with the paper version. In this case the usual journal concept, whereby articles are judged and selected within the peer review system and made accessible under subscription, is maintained. Compared with the printed versions these electronic journals have the advantage of allowing full-text searching and hypertext links as well as the printing of an article on demand. Access is possible via Internet or other telecommunication lines. Important examples are given in Table 1.

List of Works Cited

Durrani, M. (1996), "Terminal Future for Paper Journals," *Phys. World* **9** (2), 7.

Taubes, G. (1993), "Publication by Electronic Mail Takes Physics byStorm," *Science* **259**, 1246–1248.

COMPUTER PROGRAMMING LANGUAGES

(An addendum to *Encyclopedia of Applied Physics*, Volume 4, pages 163–188.)

HERBERT J. BERNSTEIN, *Bernstein + Sons, Bellport, New York, U.S.A.*

Introduction 355
1. **The Evolution of FORTRAN** 355
2. **C and C++** 356
3. **HTML and JAVA™** 357
4. **Document-Preparation and Data-Description Languages** 358

Acknowledgments 358
Glossary 358
Works Cited 358
Further Reading 359

INTRODUCTION

Computer programming languages are under active development in many areas. For physicists, the most important recent changes in computer programming languages have been extensions to FORTRAN and increased use and availability of C and C++. Increased access to computer networks, high-performance graphics, and databases has led to the increased use of HyperText Markup Language (HTML) and JAVA™ (a special dialect of C++) to create Web pages and forms, use of languages such as TEX™, Postscript™, and Rich Text Format (RTF) for document preparation, and use of languages such as STAR and CIF for database entry preparation. For a more general perspective on recent developments in programming languages, see Bruce (1996), Cardelli (1996), Cohen (1996), Goldberg (1996), Hirschfield and Ege (1996), Schmidt (1996), Wolfe (1996). For a list of freely available language processors, see Robenalt and Sharnoff (1996).

1. THE EVOLUTION OF FORTRAN

FORTRAN remains the dominant language in which scientific computing applications are written. As of this writing, the most popular dialect is the one from 1978 called FORTRAN 77, with gradually increasing use of features from the dialect called FORTRAN 90 adopted by the International Standards Organization in 1991. The language continues to evolve, with a dialect called FORTRAN 95 now available, and the general features of FORTRAN 2000 have been proposed. A guide to the literature can be found in Metcalf (1996), and a detailed description of the current dialects in Metcalf and Reid (1996).

When code portability is an important consideration, it is still best to restrict FORTRAN programs to the subset of FORTRAN 77 described in COMPUTER PROGRAMMING LANGUAGES, but, eventually, many new features from the more recent dialects should be widely available.

FORTRAN was created when the most common computer input medium was the 80-column punch card on which it was simple to keep track of column numbers. Programs are now written with use of interactive editors, many of which do not provide simple mechanisms to track column numbers. Therefore, newer dialects of FORTRAN permit programs to be written in free format, in which statements may be begun in the first character position of a line and use up to 132 character positions of a line. When an exclamation point appears on a line, the rest of the line is treated as a comment. Multiple statements may appear on the same line separated by semicolons. When the portion of a line that is not a comment ends with an ampersand, the FORTRAN statement continues on the next line, which may optionally begin with an ampersand In most cases, blanks are optional, but for most keywords, such as "READ", "WRITE", etc., a blank is required. Full use of free format can cause problems with portability of code. With some care, it is possible to

© 1999 WILEY-VCH Verlag GmbH

ISBN 3-527-29307-8

adopt a style that is accepted by most compilers as either free format or the older fixed-field format by restricting statements to the first 72 columns of a line, reserving columns 1 through 5 for statement labels, reserving columns 6 and 73 for continuation, separating all keywords with blanks, and avoiding multiple statements on one line. Most compilers now accept the exclamation point as a comment character.

Newer dialects accept pointers that allow indirect references to data. This brings to FORTRAN much of the power that languages such as C have in manipulating complex dynamically allocated data structures, but also brings the attendant difficulties in debugging and optimization seen when pointers are used. When code to manipulate linked lists or trees of information is being written, pointers can greatly improve the clarity and reliability of the code produced, but in the writing of code that can be done with equal clarity with arrays, pointers are best avoided.

Newer dialects also allow arrays to be manipulated as easily as scalar variables. All the elements of an array may be moved or manipulated at one time by using the name of the array, instead of having to write an element-by-element loop. An entire array may be given as the value of a function. When a compiler provides such features, not only can the clarity and reliability of the code be improved by their use, but efficiency can be improved as well. When array operations are specified element by element, the compiler is not as free to create optimal code as it is when only the higher-level specification in terms of arrays is given.

There are many other improvements to the language, but one of the most important is the definition of modules, which allows the grouping of subroutines and related data into clearly defined packages. This is a step towards the eventual inclusion of a full range of object-oriented features into FORTRAN (see COMPUTER PROGRAMMING LANGUAGES for more information on "object-oriented" programming).

An important change in the environment for the writing of portable FORTRAN code has been the increasing use of F2C (Feldman *et al.*, 1990), and the availability of G77, a FORTRAN compiler integrated with the GNU project C compiler (Stallman *et al.*, 1996). F2C translates FORTRAN code into C and provides a standard library of operating system interface routines. G77 is a fully operational compiler that produces object code very similar to that produced by F2C but bypassing the need for intermediate C code. G77 and F2C are very popular for FORTRAN compilation on personal computers running variants of the UNIX™ operating system. Subtle differences in the interpretation of language standards (such as in the handling of string concatenation) and in operating-system interfaces and interfaces to routines written in other languages, such as C, has made the creation of portable FORTRAN code difficult, requiring the use of very restricted language subsets and special platform-dependent variations. F2C and G77 have greatly simplified the process of porting codes by providing virtually identical language processors on a wide variety of platforms and clearly defined interfaces to the operating system and to code written in other languages.

2. C AND C++

C has become the standard language for most operating-system programming. With various extensions, C is now the language of choice for building basic systems-level tools, displacing both assembly language and PASCAL. Object-oriented extensions to C have been defined. The most popular is C++ (Ellis and Stroustrup, 1990). Originally provided as a translation program that converted an object-oriented extension of C into C, C++ is now processed as a full-fledged compiled language, and C has evolved to become a compatible subset of C++. The GNU C compiler GCC (Stallman *et al.*, 1996) processes C, C++, and another object-oriented dialect, called "Objective C". C++ supports polymorphism and overloading (see SOFTWARE ENGINEERING), and emphasizes static binding that allows a wide range of potential errors to be detected at compile time. Objective-C emphasizes dynamic binding that simplifies the management of complex code libraries.

An important part of static checking in C or C++ is the use of function prototypes, which carry the important information about how to call a routine without having to provide all the code. Consider the simple FORTRAN function "nblen," which returns the length of a string to the last non-blank character:

```
   function nblen (str)
C  non-blank length of a string
   character*(*) str
   integer nblen,jj,ii
   jj = len(str)
   do ii = 1, jj
     nblen = 1 + jj - ii
     if(str(nblen:nblen).ne.' ')return
   enddo
   nblen = 0
   return
   end
```

When this subroutine is converted by F2C to C or C++, a file called "nblen.P" is created with the statement

```
extern integer
nblen_ (char *str, ftnlen str_len);
```

which says that there is an integer function named "nblen_" (the underscore was added by F2C) with two arguments, the address of a string called "str" and its length "str_len". Programs that use nblen can be compiled using the information in nblen.P to check that the calls that they make are correctly structured. The C or C++ code generated by the loop in the middle of the function looks very much like the original FORTRAN code:

```
jj = i_len(str, str_len);
i__1 = jj;
for (ii = 1; ii <= i__1; ++ii) {
ret_val = jj + 1 - ii;
if (*(unsigned char *)
  &str[ret_val - 1]
  != ' ') {
   return ret_val;
}
}
```

For simple procedural algorithms, there is no special advantage in converting from FORTRAN to C or C++, but in the writing of applications involving complex graphical interfaces and real-time interactions, access to powerful, well-debugged "class-libraries" (predefined libraries of C++ routines and data structures) to perform such tasks as drawing images in windows and passing messages to motor-drivers makes use of C++ highly desirable.

3. HTML AND JAVA™

Much of science is now dependent on the communications infrastructure known as the Internet. Experimental scientists can now start an experiment at a major facility and control it from local or remote terminals using the Internet. The HyperText Markup Language (HTML), augmented by a dialect of C++ called JAVA, is an effective programming language for the creation of applications distributed across a network that can be run on a wide variety of platforms. HTML describes documents. In its most common application HTML documents are presented for viewing in terminal windows by a Web browser. A raw HTML document consists of lines of text marked up by special instructions delimited by angle brackets. The special instructions may flag new paragraphs, headings, lists or other text formatting, may cause images to be fetched, or may cause programs to be executed. The image presented can have areas that respond to mouse clicks to cause another page to be found or, by use of program fragments ("applets") written in JAVA, cause a wide variety of processing actions to occur. There are many competing, evolving, somewhat incompatible dialects of HTML. The most recent "official" version of HTML is version 3.2 (Raggett, 1996). This is a well-supported version to use in writing Web pages. HTML Version 4.0 has been released in draft form and is supported by many Web browsers, but many other Web browsers in use fail to support features of the most recent HTML versions. See SOFTWARE ENGINEERING for more on HTML.

JAVA defines a dialect of C++ that is highly unlikely to cause problems for a computer that downloads an applet from an unknown source. The applet is not allowed general access to memory through pointers, nor to arbitrary files on the host machine. The applet is downloaded in very compact "byte-code" form (*i.e.*, partially compiled to the level of numeric instructions to a JAVA virtual machine). Much of the work of creating powerful applets is taken care of by the existence of comprehensive class libraries and collections of sample applets. For more on JAVA, see Gosling *et al.* (1996).

The process of creating applications to interact with Web pages is greatly simplified by the use of scripting languages, which function interpretively. Two popular scripting languages are TCL and Perl. Perl combines pattern-matching capabilities with C-like programming. TCL allows powerful stand-alone or browser-embedded graphic applications to be built quickly. For more on Perl, see Wall and Schwartz (1990). For more on TCL, see Ousterhout (1996).

4. DOCUMENT-PREPARATION AND DATA-DESCRIPTION LANGUAGES

The transition to electronic presentation of scientific information and the decentralization of computing and document preparation has changed our view of what constitutes a scientific publication. We now need to separate the information to be presented from the mode of presentation, and the languages for the presentation of information have become increasingly important as programming languages. HTML is effective for presentation on a variety of platforms, but does not allow for precise control of the style of presentation in terms of fonts, character sizes, and other layout issues. Currently, the most popular "page description languages" (languages for precise document layout) are TEX™ (Knuth, 1986), which is used internally by many scientific journals, Postscript™ (Adobe Systems, 1990), which is used to create on-line distribution files of many scientific documents with full formatting, and Rich Text Format (RTF) (Microsoft, 1995), which is used for interchange of word-processing documents across systems. An effort is under way to combine precise document layout with HTML by means of "Cascading Style Sheets" (CSS) (Lie and Bos, 1996).

As databases and data archives have become a critical part of scientific research, the need for common languages for capturing data and descriptive information about the data has become clear. One of the most successful efforts in this direction has been the work at the International Union of Crystallography on a Self-Defining Text Archive and Retrieval (STAR) language and a Crystallographic Information File (CIF) (Hall *et al.*, 1991). Based on simple tag-value constructs, these languages have made it possible to gather all the information needed for both a scientific publication and the archival data upon which the publication is based into a common data format. When complex data structures must be described in a precise but machine-independent manner, Abstract Syntax Notation One (ASN.1) is often used. See Gaudette *et al.* (1989, Chapter 2) for an excellent tutorial and references on ASN.1.

It has become clear that these languages are all more than means for expressing the detailed layout of documents or making entries in databases and archives. They are intrinsic parts of an evolving composite language for the control of computers in the transformation of information among many possible representations.

ACKNOWLEDGMENTS

Our thanks to Daniel J. Bernstein and Frances C. Bernstein for helpful comments and suggestions.

GLOSSARY

GNU ("GNU's Not UNIX"): A "complete integrated software system ... that will be upwardly compatible with UNIX" (statement by Free Software Foundation). We regret the necessity of presenting the Stallman *et al.* (1996) citation for GCC without the full list of authors. The full list of contributors can be found within the GCC documentation.

JAVA: A trademark of Sun Microsystems.

Postscript: A trademark of Adobe Systems Incorporated.

TEX: A trademark of the American Mathematical Society.

UNIX: A trademark of AT&T.

List of Works Cited

Adobe Systems (1990), *Postscript®* Language Reference Manual, 2nd ed., Reading, MA: Addison-Wesley.

Bruce, K. B. (1996), *ACM Comput. Surveys* **28**, 245–247.

Burley, J. C. (1996), *GNU Fortran Compiler*, ftp://prep.ai.mit.edu/pub/gnu/g77-0.5.19.tar.gz, Boston MA: Free Software Foundation.

Cardelli, L. (1996), *ACM Comput. Surveys* **28**, 263–264.

Cohen, J. (1996), *ACM Comput. Surveys* **28**, 257–259.

Ellis, M. A., Stroustrup, B. (1990), *The Annotated C++ Reference Manual*, Reading, MA: Addison-Wesley.

Feldman, S. I., Gay, D. M., Maimone, M. W., Schryer, N. L. (1990), *A Fortran-to-C Converter*, Computing Science Technical report No. 149, Murray Hill, NJ: AT&T Bell Laboratories.

Gaudette, P., Trus, S., Collins, S. (1989), *The Free Value Tool For ASN.1*, Technical Report NCSL/SNA-89/1, Gaithersburg, MD: National Institute of Standards and Technology.

Gosling, J., Joy, B., Steele, G. (1996), *The Java™ Language Specification*, Reading, MA: Addison-Wesley.

Goldberg, B. (1996), *ACM Comput. Surveys* **28**, 249–251.

Hall, S. R., Allen, F. H., Brown, I. D. (1991), *Acta Cryst.* **A47**, 655–685.

Hirshfield, S., Ege, R. E. (1996), *ACM Comput. Surveys* **28**, 253–255.

Knuth, D. E. (1986), *The TEX book*, Reading, MA: Addison-Wesley.

Lie, H. W., Bos, B. (1996), *Cascading Style Sheets, level 1*, W3C Recommendation REC-CSS1-96127, Cambridge, MA: World Wide Web Consortium.

Metcalf, M. (1996), *Fortran Forum* **15**, 32–35.

Metcalf, M., Reid, J. (1996), *Fortran 90/95 Explained*, Oxford, UK.: Oxford University Press.

Microsoft (1995), *Rich Text Format (RTF) Specification and Sample RTF Reader Program*, RTF Version 1.4, GC0165, Redmond, WA: Microsoft Corporation.

Ousterhout, J. (1996), *The TCL 8.0 and Tk 8.0 releases*, ftp://ftp.sunlabs.com/pub/tcl, Mountain View, CA: Sun Microsystems.

Raggett, D. (1996), *HTML 3.2 Reference Specification*, http://www.w3.org/pub/WWW/TR/PR-html32-961105, Cambridge, MA: World Wide Web Consortium.

Robenalt, S. A., Sharnoff, D. M. (1996), *Catalog of Free Compilers and Interpreters*, ftp://ftp.idiom.com/pub/compilers-list/free-compilers or http://www.idiom.com/free-compilers, Oakland, CA: Idiom Consulting.

Schmidt, D. A. (1996), *ACM Comput. Surveys* **28**, 265–267.

Stallman, R., *et al.* (1996), *GNU CC*, ftp://prep.ai.mit.edu/pub/gnu/gcc-2.7.2.1.tar.gz, Boston, MA: Free Software Foundation.

Wall, L., Schwartz, R. L. (1990), *Programming Perl*, Sebastopol, CA: O'Reilly & Associates.

Wolfe, M. (1996), *ACM Comput. Surveys* **28**, 261–262.

Further Reading

Budd, T. (1997), *An Introduction to Object-Oriented Programming*, 2nd ed., Reading, MA: Addison-Wesley.

Eckel, B. (1989), *Using* C++, Berkeley, CA: Osborne McGraw-Hill.

Gehani, N. (1986), *Document Formatting and Typesetting on the Unix™ System*, Summit, NJ: Silicon Press.

Graham, I. S. (1995), *HTML Sourcebook*, New York, NY: John Wiley & Sons.

Nielson, J. (1990), *Hypertext and Hypermedia*, San Diego, CA: Academic Press.

Olagnon, M. (1996), *F90 FAN's: Fortran 90 Frequently Asked about News*, ftp://ftp.ifremer.fr/ifremer/ditigo/fortran90, Brest, France: Institut Français de Recherches pour l'Exploitation de la Mer.

Sams.net (1996), *Java™ Unleashed*, Indianapolis, IN: Sams.net Publishing.

van Herwijnen, E. (1990), *Practical SGML*, Dordrecht: Kluwer Academic Publishers.

ELECTROLUMINESCENCE

(An Addendum to *Encyclopedia of Applied Physics*, Volume 5, pages 295–325.)

YOSHIMASA A. ONO, *Advanced Research Laboratory, Hitachi, Ltd., Hatoyama, Saitama, Japan*

Introduction 361
1. **Color EL Phosphors for Inorganic EL Displays** 361
1.1 Color Phosphors for the Patterned-Phosphor Structure... 362
1.1.1 Red Phosphors................. 362
1.1.2 Green Phosphors............... 363
1.1.3 Blue Phosphors 363
1.2 Color Phosphor for the Filtered White (Color-by-White) Structure 363
2. **Inorganic Color Thin-Film EL Displays** 364
2.1 Multicolor Displays............ 364
2.2 Full-Color Displays............ 365
2.2.1 Filtered White (Color-by-White) Scheme in the Inverted Structure 365
2.2.2 Patterned Phosphor Scheme in the Dual-Substrate Structure ... 365
3. **Organic Color Thin-Film EL Displays**........................ 366
4. **Conclusions and Future Prospects**....................... 367
Works Cited.................... 368
Further Reading 369

INTRODUCTION

This Addendum discusses developments in inorganic and organic electroluminescence since the publication of the original article ELECTROLUMINESCENCE.

The biggest challenge in inorganic thin-film EL technology today is producing multicolor and full-color displays. Manufacturing of practical color displays was long delayed because of the insufficient luminance of the primary colors (red, green, and blue). Recently, however, progress in the development of the multicolor display structures and color phosphors has made it possible to fabricate commercial multicolor EL displays fulfilling customer's requirements (Cramer *et al.*, 1993), and commercialization of full-color EL displays is close to being a reality (King, 1996a).

Organic thin-film EL devices have been extensively studied since an efficient green emission was demonstrated in 1987 (Tang and VanSlyke, 1987). Substantial improvements in luminous efficiency, color, and stability have been made (Tang, 1996a,b). To a large extent, they have been made possible by use of a multilayer device structure prepared by using vapor deposition and by the discovery of several important classes of organic EL materials that have excellent transport and luminescent properties. These advances have stimulated commercial interest in developing new color displays.

The present review discusses new inorganic thin-film EL phosphor materials with blue and broad-band emissions and their applications to commercial multicolor and full-color EL displays. In addition, recent developments of organic color thin-film EL are discussed.

1. COLOR EL PHOSPHORS FOR INORGANIC EL DISPLAYS

The development of color phosphor for inorganic EL displays has divided into two different material approaches that are based on the structure of the thin-film EL panel. The two basic color thin-film EL structures are the patterned-phosphor structure and the filtered-white structure, also known as the color-by-white structure (see, for example, Ono, 1995, 1996, 1997).

Figure 1 shows practical examples of these two structures. In the patterned-phosphor structure, the development of efficient red, green, and blue primary-color EL phosphors is necessary, and in the filtered-white structure, the development of an efficient white (or broad-band) phosphor that can be filtered to produce a red, green, and blue (RGB) display is necessary. Improved phosphor performances have been the key technological advancement lead-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

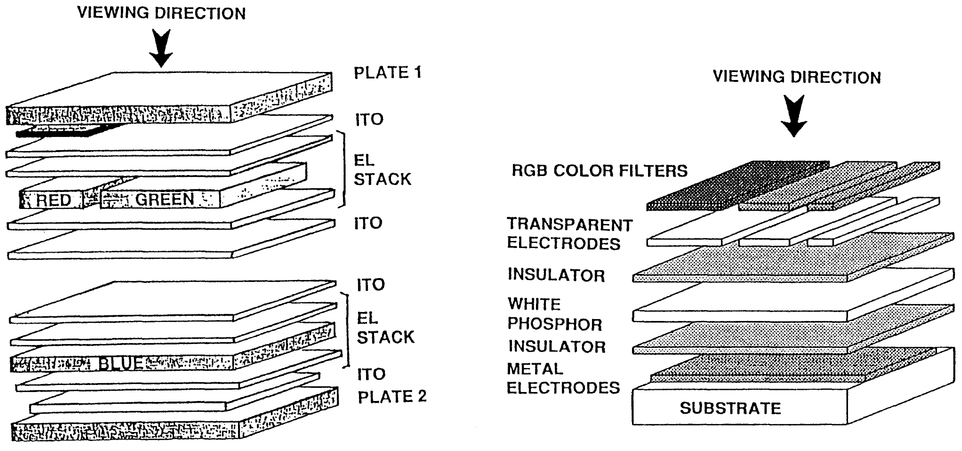
VIEWING DIRECTION
PLATE 1
ITO
EL STACK
RED
GREEN
ITO
ITO
EL STACK
BLUE
ITO
PLATE 2
VIEWING DIRECTION
RGB COLOR FILTERS
TRANSPARENT ELECTRODES
INSULATOR
WHITE PHOSPHOR
INSULATOR
METAL ELECTRODES
SUBSTRATE

(a) Patterned-phosphor scheme in a dual-substrate structure

(b) Filtered-white or color-by-white scheme in an inverted structure

FIG. 1. Two types of inorganic color thin-film EL structures: (a) patterned-phosphor scheme in a dual-substrate structure; (b) filtered-white or color-by-white scheme in an inverted structure.

ing to the recent successes in both approaches to develop color thin-film EL displays.

In this section, recent developments in red, green, and blue phosphors for the patterned-phosphor structure and broad-band and white phosphors for the filtered-white (color-by-white) structure are discussed.

1.1 Color Phosphors for the Patterned-Phosphor Structure

1.1.1 Red Phosphors. The most efficient EL red emission is produced by the CdS_xSe_{1-x}-filtered ZnS : Mn phosphor (Tuenge and Kane, 1991), whose emission spectrum is very broad, ranging from green to red. The Commission Internationale de l'Eclairage (CIE) color coordinates of $x = 0.65$ and $y = 0.35$ are very close to the red coordinates for color CRTs ($x = 0.624$, $y = 0.337$). Luminance of 70 cd/m^2 at 60 Hz is sufficient for color display applications. These data are summarized in Table 1, which shows a representative list of the best reported EL phosphor performance data. In the table, L_{40} indicates the luminance measured at 40 V above the threshold voltage.

Table 1. Luminance and luminous efficiency of inorganic color EL phosphors.

Phosphor material	Emission color	CIE coordinates x	y	Luminance L_{40} at 60 Hz (cd/m^2)	Luminous efficiency η (lm/W)
ZnS : Mn	Yellow	0.50	0.50	300	3–6
ZnS : Mn/Filter	Red	0.65	0.35	70	0.8
ZnS : TbOF	Green	0.30	0.60	100	1.3
ZnS : Tb,F	Green	0.31	0.60	90	0.6
ZnS : Mn/Filter	Yellow-green	0.47	0.53	160	1.0
SrS : Ce	Blue-green	0.26	0.48	140	1.2
SrS : Ce/Filter	Blue	0.10	0.16	14	0.12
$SrGa_2S_4$: Ce	Blue	0.14	0.10	5	0.02
$CaGa_2S_4$: Ce	Blue	0.14	0.19	10	0.03
$Ca_{0.5}Sr_{0.5}Ga_2S_4$: Ce	Blue	0.14	0.13	4.3	—
ZnS : Mn/SrS : Ce	Yellowish-white	0.49	0.48	470	1.3–1.6

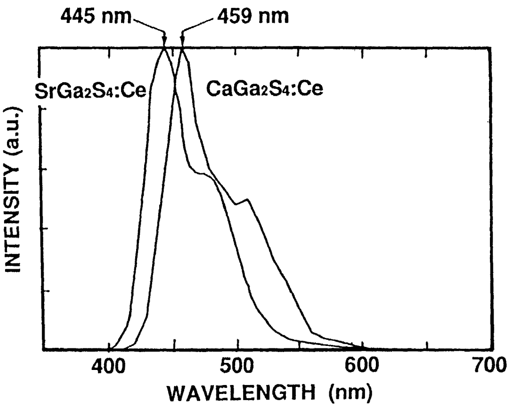

FIG. 2. Inorganic EL emission spectra of $CaGa_2S_4:Ce$ and $SrGa_2S_4:Ce$.

1.1.2 Green Phosphors. For a pure green emission, ZnS:TbOF has the highest luminance and luminous efficiency (see Table 1). The CIE color coordinates of $x = 0.30$ and $y = 0.60$ are close to the green coordinates for color CRTs ($x = 0.312$, $y = 0.597$). Green emission can also be obtained by filtering out the green component of the ZnS:Mn spectrum, which results in a yellow-green luminance.

1.1.3 Blue Phosphors. The first material to show promise as a blue EL phosphor was SrS:Ce (Barrow *et al.*, 1984). However, the broad-band blue-green emission of SrS:Ce requires filtering in order to achieve a blue chromaticity that is acceptable for RGB color displays. The performance of the SrS:Ce phosphor has recently been improved: molecular-beam-epitaxy (MBE)-deposited SrS:Ce was used to achieve a filtered luminance of 14 cd/m^2 with the CIE color coordinates $x = 0.10$ and $y = 0.16$ (Törnqvist, 1997).

Barrow *et al.* (1993) reported a significant breakthrough in blue phosphors. They showed that a group of alkaline-earth thiogallates MGa_2S_4 (M: Ca, Sr) were efficient host materials for the blue EL phosphor when activated with Ce. Figure 2 shows the EL emission spectra for calcium and strontium thiogallate films activated with 1 mol% cerium. Luminance and luminous efficiencies of these thiogallate phosphors are given in Table 1.

The data in Table 1 clearly show that the alkaline-earth thiogallates provide a much deeper blue emission than that of the SrS:Ce phosphor. The $SrGa_2S_4:Ce$ phosphor achieves the deepest blue color with the CIE color coordinates of $x = 0.14$ and $y = 0.10$ and provides a color gamut close to that of the standard CRT display ($x = 0.157$, $y = 0.069$). However, the eye response for the $SrGa_2S_4:Ce$ phosphor emission is nearly a factor of two less than that of the $CaGa_2S_4:Ce$ emission. Thus for equivalent energy efficiency, the $CaGa_2S_4:Ce$ would achieve twice the photopic brightness. The best luminance of 10 cd/m^2 at 60 Hz was obtained for sputtered $CaGa_2S_4:Ce$ using high–dielectric-constant insulators. Very good EL aging characteristics have been demonstrated for the thiogallates (Barrow *et al.*, 1993). Therefore, the thiogallate blue EL phosphors demonstrate the sufficient stability, luminance, and blue chromaticity needed to build a full-color thin-film EL display.

Recently, blue purity has been further improved by mixing Ca and Sr in the host (Tuenge, 1995a, 1995b). The cerium-emission peak wavelength can be continuously shifted from 445 to 459 nm in a mixed calcium and strontium host because these two alkaline-earth thiogallates are miscible over the entire composition range. By optimizing the composition of the mixed host, $Ca_{0.5}Sr_{0.5}Ga_2S_4:Ce$ was found to give a good result with the CIE color coordinates of $x = 0.14$ and $y = 0.13$ and the luminance $L_{40} = 4.3$ cd/m^2 at 60-Hz drive.

1.2 Color Phosphors for the Filtered-White (Color-by-White) Structure

A white-emitting phosphor may be designed either for monochromatic white thin-film EL displays or for multicolor or full-color thin-film EL displays achieved by filtering white light through color filters.

Recently, white emission with the luminance level required for filtered color thin-film EL displays has been achieved by using a stacked ZnS:Mn/SrS:Ce phosphor. The emission spectrum is shown in Fig. 3. This broad-band multilayer phosphor has been investigated by several groups using different deposition processes including multisource deposition (Nire *et al.*, 1992), reactive evaporation (Mauch *et al.*, 1993; Mauch, 1994; Velthaus *et al.*, 1994), and atomic-layer epitaxy (ALE) (Soininen *et al.*, 1993; Soininen *et al.*, 1994; Törnqvist, 1996). The luminance

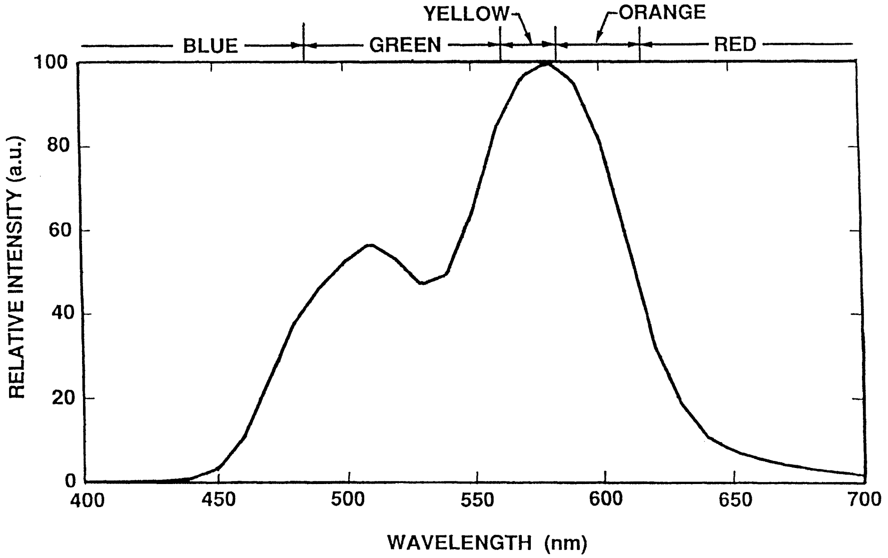

FIG. 3. White EL emission spectrum of ZnS : Mn/SrS : Ce phosphor.

performance of the state-of-the-art ZnS:Mn/SrS:Ce grown by ALE is as follows (Törnqvist, 1997): The filtered red, green, and blue luminances L_{50} at 60-Hz driving are 97, 220, and 11 cd/m^2, respectively. These values are sufficient for RGB multicolor EL display applications. The blue-filtered color coordinates of $x = 0.13$ and $y = 0.36$ need further improvement in order to extend this approach to full-color capability.

The ALE-deposited SrS:Ce films have shown good luminance stability (Soininen *et al.*, 1994), and recently a significant improvement in the stability of evaporated SrS:Ce, Cl films has been demonstrated by codoping them with Zn or Mn (Mauch *et al.*, 1995). These results indicate that by careful control of the defects in the SrS host lattice, the aging characteristics of this material will be sufficient for use in color thin-film EL display panels.

2. INORGANIC COLOR THIN-FILM EL DISPLAYS

2.1 Multicolor Displays

The first RG (red/green) multicolor display (Cramer *et al.*, 1993) employs the filtered broad-band emission from ZnS:Mn in the inverted EL structure [see Fig. 1(b)]. Red and green pixels are generated by use of a patterned color filter plate that is in direct contact with the thin-film stack, thereby eliminating any parallax effects. This structure has several advantages. A very good contrast can be achieved without the use of the traditional circular polarizer, and contrast ratios of 20 : 1 at 500 lx ambient and 10 : 1 at 1000 lx ambient are typical results. In addition, the fabrication of the active thin-film stack can be done with the same equipment line that is now used to fabricate monochrome ZnS:Mn panels, producing very good yields.

In 1994 an eight-color, 10-in. diagonal VGA display (640 × 480 pixels) was brought to the market (Soininen *et al.*, 1994). This multicolor EL panel is an ideal choice for display applications such as industrial control systems and medical electronics, where color highlighting and differentiation of graphics information are needed.

Recently the green color purity of the RG multicolor display has been improved by using $Zn_{0.9}Mg_{0.1}S$:Mn as a broad-band phosphor (Mikami *et al.*, 1996; Mikami, 1997), which has exhibited a pure yellow emission with a peak wavelength of 575 nm. Based on this technology, 4.7- and 9.4-in. diagonal prototype RG multicolor EL displays with a wider color range were fabricated.

2.2 Full-Color Displays

2.2.1 Filtered-White (Color-by-White) Scheme in the Inverted Structure. The simple filtered, inverted multicolor panel structure can be extended to full-color capability by replacing the ZnS:Mn EL phosphor with a ZnS:Mn/SrS:Ce white phosphor. In 1995 a 9-in. diagonal 512 × 256-pixel RGB multicolor panel was fabricated (Haaranen *et al.*, 1995) that demonstrated the performance that meets product requirements. The typical white areal luminance was 21 cd/m^2. The contrast ratio was 40 : 1 in normal office ambient illumination of 500 lx and was 5 : 1 in bright ambient illumination of 10 000 lx. In 1997 a much brighter 5-in. diagonal quarter-VGA (320 × 240 pixels) display was developed (Harju *et al.*, 1997). A white areal luminance of 70 cd/m^2 at 350-Hz driving and contrast ratios of more than 100 : 1 at 500 lx and 7 : 1 at 10 000 lx have been achieved. The CIE color coordinates of the red, green, and blue emissions are ($x = 0.61$, $y = 0.38$), ($x = 0.46$, $y = 0.51$), and ($x = 0.28$, $y = 0.41$), respectively. The color gamut is adequate for applications where the three primary colors must be present but full color is not needed, such as instrumentation applications and automotive displays.

This filtered-white scheme has been applied to develop a 1.5-in. diagonal color VGA active-matrix EL (AMEL) display that is fabricated on single-crystal silicon-on-insulator IC wafers (Aguilera *et al.*, 1996; King, 1996b; Törnqvist, 1997). A filtered white luminance of 555 cd/m^2 was obtained at 4.5-kHz AMEL drive, and filtered luminances were 110, 410, and 35 cd/m^2 for red ($x = 0.63$, $y = 0.36$), green ($x = 0.29$, $y = 0.68$), and blue ($x = 0.12$, $y = 0.20$) emissions, respectively (Törnqvist, 1997). The color gamut of this display should be sufficient for video applications. In 1997 a lower-power, larger–color-gamut, 0.7-in. diagonal full-color VGA display was developed (Tuenge *et al.*, 1997). With field-sequential filtering using a liquid-crystal color shutter, 256 colors have been achieved. The power consumption is typically 2 W. These color AMEL displays are expected to provide a low-cost solution for compact, high-resolution miniature color displays that are suitable for a variety of military and commercial head-mounted and personal viewer applications.

2.2.2 Patterned-Phosphor Scheme in the Dual-Substrate Structure. As shown in Fig. 1(a), the dual-substrate panel structure optimizes the net blue luminance. Blue light is obtained from an unpatterned cerium-doped alkaline-earth thiogallate phosphor. Green light is obtained from a patterned ZnS:Tb phosphor film, and the red component is obtained by filtering the ZnS:Mn emission with a CdS_xSe_{1-x} thin-film filter. The dual-substrate panel structure also allows the drive electronics to be independently optimized for each substrate. Therefore, a higher refresh rate can be used for the blue substrate to increase blue luminance.

In 1993 using the dual-substrate panel structure the first full-color EL display, a 10-in. diagonal VGA prototype display, was reported (Barrow *et al.*, 1993). In 1994 pursuing the same structural approach, a quarter-VGA (320 × 256 pixels) full-color display (FC2) that has a performance sufficient for the first commercial full-color products was introduced (Barrow *et al.*, 1994; King, 1994). In the VGA prototype, $CaGa_2S_4$:Ce was used, and in the commercial FC2 panel, the improved blue phosphor $Ca_{0.5}Sr_{0.5}Ga_2S_4$:Ce was used (Tuenge, 1995a). The color gamut of these full-color panels (VGA panel and FC2) is shown in Fig. 4.

This figure shows that the color ranges of the full-color panels are very close to those of the color CRT. The EL red and green primaries are more saturated than those of the color CRT, whereas the EL blue approaches the CRT blue. A white areal luminance of 30 cd/m^2 with the CIE color coordinates ($x = 0.37$, $y = 0.41$) is achieved at a 180-Hz drive. Filtered areal luminances are 9, 17, and 4 cd/m^2 for red ($x = 0.65$, $y = 0.35$), green ($x = 0.31$, $y = 0.61$), and blue ($x = 0.14$, $y = 0.13$) emissions, respectively. A high contrast of 20 : 1 is realized in a 500-lx ambient condition. This quarter-VGA full-color display is expected to be competitive in price with TFT LCDs and will offer EL's traditional advantages over LCDs, which include undegraded performances over extremes of temperature and a very wide viewing angle.

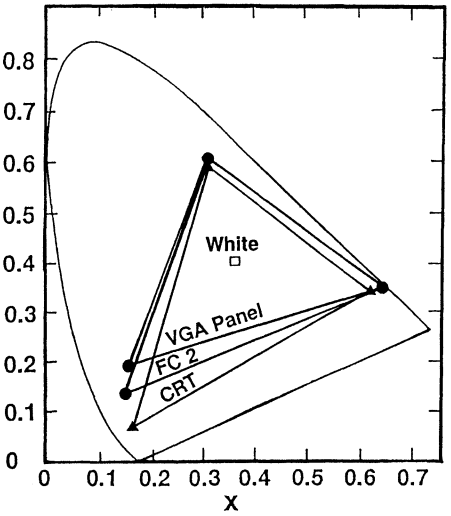

FIG. 4. Color gamut of inorganic full-color EL displays (VGA panel and FC2) and color CRT.

3. ORGANIC COLOR THIN-FILM EL DISPLAYS

The efficient organic EL devices generally have the three-layer structure with electrodes: anode/hole-transport layer (HTL)/emitting layer (EML)/electron-transport layer (ETL)/cathode (Adachi *et al.*, 1988). The most efficient EL devices use an EML layer consisting of a host material doped with a small amount of fluorescent molecules (Tang *et al.*, 1989). For example, efficient green and orange-red ELs have been obtained from an 8-hydroxyquinoline aluminum (Alq) host using dopants such as coumarin 6, QA, and DCJT (Tang, 1996a, 1996b). Other host materials include distyrylarylene derivatives (DSA), such as 1,4-bis(2,2-diphenylvinyl)biphenyl (DPVBi) (Hosokawa *et al.*, 1995), which are excellent blue emitters. The chemical structures of these organic emitting layer (EML) materials and dopants are shown in Fig. 5.

The EL spectra of a set of organic EL devices emitting in blue, green, and red are shown in Fig. 6 and their luminance characteristics are summarized in Table 2 (Tang, 1996a, 1996b). These EL devices used an EML layer based on doped Alq.

The chromaticity of the green EL is close to that of the CRT green, but the blue and red colors are less pure. Further improvement in the purity of the blue and red emissions is necessary.

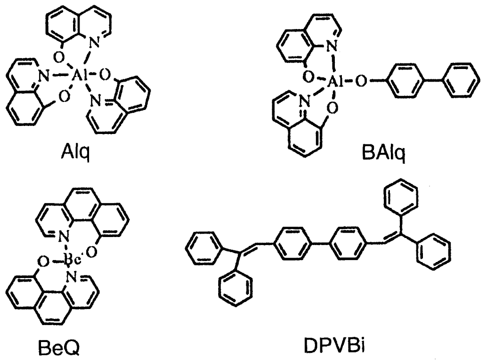

(a) Emitting Layer (EML) Materials

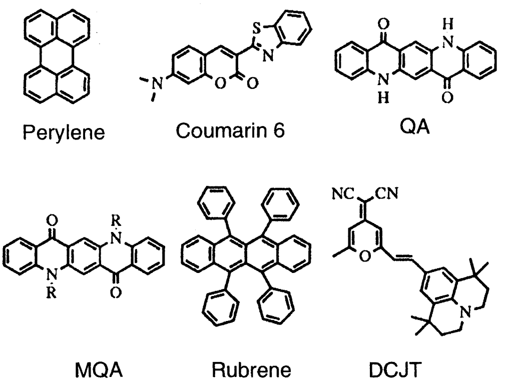

(b) Dopants

FIG. 5. Chemical structures of representative organic EL materials: (a) emitting-layer (EML) materials; (b) dopants.

Other combinations of organic materials and dopants were reported to produce excellent efficiencies. The quinacridone-doped Alq device that had green emission exhibited an efficiency 2.5 times higher than that of the undoped Alq device (Wakimoto *et al.*, 1996). For blue EL, the DPVBi class of materials was found to be one of the most efficient. A luminous efficiency of 1.5 lm/W was reported for a DPVBi-doped DSA device (Hosokawa *et al.*, 1995).

Remarkably stable EL devices—in particular, those with emitters based on doped Alq

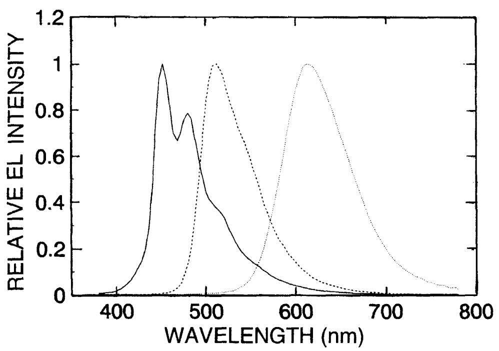

FIG. 6. Organic EL emission spectra of RGB emitters.

and doped DPVBi devices—have been demonstrated. For example, for a green EL device with a modified quinacridone MQA-doped Alq, the half-life exceeded 6000 h (Tang, 1996a,b) and for a blue EL device with DPVBi-doped DSA, the half-life was 5000 h (Wakimoto *et al.*, 1996).

The only approach to have achieved multicolor and full-color displays has been to develop a bright blue-emitting thin-film layer and to obtain green and red emissions by using color-changing media (CCM) (Matsuura *et al.*, 1995; Hosokawa *et al.*, 1997). This has been done by using the CCM made of organic fluorescent materials. The conversion efficiencies of luminance from blue to green emissions and blue to red emissions were 120% and 30%, respectively. A distyrylarylene (DSA) derivative, DPVBi, was used as a blue-emitting layer and an amino-substituted DSA (DSA amine) was used as the fluorescent dopant. These materials showed excellent EL performances. The luminous efficiency was improved to 6 lm/W and the half-life was more than 10,000 hours. Luminances were 40, 150, and 65 cd/m^2 for red (0.63, 0.32), green (0.23, 0.65), and blue (0.12, 0.18) emissions, respectively. The color coordinates of RGB emissions are very close to those of the color CRT.

In June 1997 a prototype 5-in. diagonal quarter-VGA (320 × 240 pixels) full-color EL display, which can display 16.8 million colors, was reported (Yomogida, 1997). The half-life was extended to 20 000 h and the power consumption was only 3.9 W. Furthermore, just six months later, a prototype 10-in. diagonal VGA (640 × 480 pixels) full-color display was demonstrated at the Japan Electronics Show '97 (Komori, 1997).

4. CONCLUSIONS AND FUTURE PROSPECTS

Inorganic color thin-film EL phosphors have been developed to a level that is sufficient for use in initial multicolor and full-color EL display products. The most efficient red and green phosphors, Mn- and Tb-doped ZnS, have been optimized for the manufacturing of patterned color displays. The Ce-doped alkaline-earth thiogallates are the first EL phosphors to provide the unfiltered blue color essential for full-color capability. Improvements in the luminance and aging behavior of the SrS : Ce blue-green phosphor have enabled an efficient and stable broad-band emission that will provide the basis for filtered multicolor EL displays.

A future challenge in the development of patterned color phosphors will be to improve the blue luminance further to enable larger, lower-power displays. In order to reduce panel cost, it will also be important to be able to reduce the blue-phosphor processing temperature. Enhanced emission in the blue spectral region will be required in order to extend the filtered broad-band structure approach to full-color capability.

Table 2. Luminance and luminous efficiency of organic color EL phosphors.

Emisson color	Host material	Dopant material	CIE coordinates x	y	Luminance at 20 mA/cm^2 (cd/m^2)	Luminous efficiency η (lm/W)
Red	Alq	DCJT	0.616	0.381	770	1.3
Green	Alq	Coumarin 6	0.263	0.619	1980	3.9
Blue	BAlq	Perylene	0.163	0.194	355	0.56

Thanks to the improvements in the performance of blue and white phosphors, color thin-film EL displays are now on the market. Based on the color-by-white scheme, the 9-in. diagonal RGB display suitable for instrumentation applications was commercialized and a 1.5-in. diagonal color VGA active-matrix EL display for use in head-mounted displays was developed. The patterned-phosphor scheme in the dual-substrate structure provided the basis for the VGA and quarter-VGA full-color EL displays to be developed, and these are expected to fulfill new flat-panel display applications.

There has been considerable progress in organic thin-film EL devices that are based on a multilayer structure, and it has become possible to make color EL devices with luminous efficiencies much greater than 1 lm/W throughout the visible spectrum. Long operational life has also been demonstrated, and a display with a half-life on the order of 10^4 h has been made. Organic color EL has thus emerged as a new emissive display technology with the potential for a wide range of applications.

List of Works Cited

Adachi, C., Tokito, S., Tsutsui, T., Saito, S. (1988), *Jpn. J. Appl. Phys.* **27**, L267–L271.

Aguilera, M., Aitchison, B., Basham, D., Moehnke, S., Ping, K., Tuenge, R., Vetanen, W., King, C. N. (1996), in: *Digest of the 1996 SID International Symposium*, Santa Ana, CA: Society for Information Display, pp. 125–128.

Barrow, W. A., Coovert, R. E., King, C. N. (1984), in: *Digest of the 1984 SID International Symposium*, Los Angeles: Society for Information Display, pp. 249–250.

Barrow, W. A., Coovert, R. C., Dickey, E., King, C. N., Laakso, C., Sun, S. S., Tuenge, R. T., Wentross, R., Kane, J. (1993), in: *Digest of the 1993 SID International Symposium*, Playa del Rey, CA: Society for Information Display, pp. 761–764.

Barrow, W. A., Coovert, R. C., Dickey, E., Flegal, T., Fullman, M., King, C. N., Laakso, C. (1994), in: *Conference Record of the 1994 International Display Research Conference*, New York: IEEE, and Playa del Rey, CA: Society for Information Display, pp. 448–451.

Cramer, D., Haaranen, J., Törnqvist, R., Pitkänen, T., (1993), in: *Application Digest of the 1993 SID International Symposium*, Playa del Rey, CA: Society for Information Display, pp. 57–60.

Haaranen, J., Harju, T., Heikkinen, P., Härkönen, G., Leppänen, M., Lindholm, T., Maula, J., Määttänen, J., Pakkala, A., Soininen, E., Sonninen, M., Törnqvist, R., Viljanen, J. (1995), in: *Digest of the 1995 SID International Symposium*, Santa Ana, CA: Society for Information Display, pp. 883–886.

Harju, T., Härkönen, G., Lahonen, M., Määttänen, J., Soininen, E., Törnqvist, R., Vasama, K., Viljanen, J., Flega, T. (1997), in: *Digest of the 1997 SID International Symposium*, Santa Ana, CA: Society for Information Display, pp. 859–861.

Hosokawa, C., Higashi, H., Nakamura, H., Kusumoto, T. (1995), *Appl. Phys. Lett.* **67**, 3853–3855.

Hosokawa, C., Eida, M., Matsuura, M., Fukuoka, K., Nakamura, H., Kusumoto, T. (1997), in: *Digest of the 1997 SID International Symposium*, Santa Ana, CA: Society for Information Display, pp. 1073–1076.

King, C. N. (1994), in: *Conference Record of the 1994 International Display Research Conference*, New York: IEEE, and Playa del Rey, CA: Society for Information Display, pp. 69–77.

King, C. N. (1996a), *J. Soc. Inf. Display* **4**, 1–8.

King, C. N. (1996b), in: R. H. Mauch, H.-E. Gumlich (Eds.), *Inorganic and Organic Electroluminescence, EL 96 Berlin*, Berlin: Wissenschaft und Technik Verlag, pp. 375–380.

Komori, K. (1997), *Nikkei Sangyo Shinbun (Nikkei Industrial Newspaper)*, October 12, 7 (in Japanese).

Matsuura, M., Tokailin, H., Eida, M., Hosokawa, C., Hironaka, Y., Kusumoto, T. (1995), in: *Proceedings of the Fifteenth International Display Research Conference (Asia Display '95)*, Tokyo: The Institute of Television Engineers of Japan, and Santa Ana, CA: Society for Information Display, pp. 269–272.

Mauch, R. H. (1994), in: X. Xurong (Ed.) *Proceedings of 1994 International Workshop on Electroluminescence*, Beijing: Science Press, pp. 86–96.

Mauch, R. H., Velthaus, K. O., Hüttl, B., Schock, H. W. (1993), in: *Digest of the 1993 SID International Symposium*, Playa del Rey, CA: Society for Information Display, pp. 769–772.

Mauch, R. H., Velthaus, K. O., Hüttl, B., Troppenz, U., Herrmann, R. (1995), in: *Digest of the 1995 SID International Symposium*, Santa Ana, CA: Society for Information Display, pp. 720–723.

Mikami, A. (1997), in: *Digest of the 1997 SID International Symposium*, Santa Ana, CA: Society for Information Display, pp. 851–854.

Mikami, A., Yashima, I., Kajikawa, F. (1996), in: R. H. Mauch, H.-E. Gumlich (Eds.), *Inorganic and Organic Electroluminescence, EL 96 Berlin*, Berlin: Wissenschaft und Technik Verlag, pp. 369–374.

Nire, T., Matsuno, A., Wada, F., Fuchiwaki, K., Miyakoshi, A. (1992), in: *Digest of the 1992 SID International Symposium*, Playa del Rey, CA: Society for Information Display, pp. 352–355.

Ono, Y. A. (1995), *Electroluminescent Displays*, Singapore: World Scientific.

Ono, Y. A. (1996), in: R. H. Mauch, H.-E. Gumlich (Eds.), *Inorganic and Organic Electroluminescence, EL 96 Berlin*, Berlin: Wissenschaft und Technik Verlag, pp. 273–278.

Ono, Y. A. (1997), *Annu. Rev. Mater. Sci.* **27**, 283–303.

Soininen, E., Leppänen, M., Pakkala, A. (1993), in: *Proceedings of the 13th International Display Research Conference, Eurodisplay 93*, Playa del Rey, CA: Society for Information Display, and London: Institute of Physics, pp. 233–236.

Soininen, E., Leppänen, M., Törnqvist, R., Viljanen, J. (1994), in: X. Xurong (Ed.), *Proceedings of 1994 International Workshop on Electroluminescence*, Beijing: Science Press, pp. 97–109.

Tang, C. W. (1996a), in: *Digest of the 1996 SID International Symposium*, Santa Ana, CA: Society for Information Display, pp. 181–184.

Tang, C. W. (1996b), *Inf. Display* **12**, 16–19.

Tang, C. W., VanSlyke, S. A. (1987), *Appl. Phys. Lett.* **51**, 913–915.

Tang, C. W., VanSlyke, S. A., Chen, C. H. (1989), *J. Appl. Phys.* **65**, 3610–3616.

Törnqvist, R. O. (1996), in: R. H. Mauch, H.-E. Gumlich (Eds.), *Inorganic and Organic Electroluminescence, EL 96 Berlin*, Berlin: Wissenschaft und Technik Verlag, pp. 279–284.

Törnqvist, R. O. (1997), in: *Digest of the 1997 SID International Symposium*, Santa Ana, CA: Society for Information Display, pp. 855–858.

Tuenge, R. T., Kane, J. (1991), in: *Digest of the 1991 SID International Symposium*, Playa del Rey, CA: Society for Information Display, pp. 279–282.

Tuenge, R. T. (1995a), in: *Proceedings of the Fifteenth International Display Research Conference, Asia Display 95*, Tokyo: The Institute of Television Engineers of Japan, and Santa Ana, CA: Society for Information Display, pp. 279–283.

Tuenge, R. T. (1995b), in: *Proceedings of the 2nd International Display Workshops (IDW '95)*, Tokyo: Phosphor Research Society, and Tokyo: Japan Chapter, Society for Information Display, pp. 89–92.

Tuenge, R. T., Larsson, T., Moehnke, S., Wald, S., Sproull, W., King, C., Sharp, G. D., Johnson, K. M., Gilman, S. (1997), in: *Digest of the 1997 SID International Symposium*, Santa Ana, CA: Society for Information Display, pp. 862–865.

Velthaus, K. O., Troppenz, U., Hüttl, B., Hermann, R., Mauch, R. H. (1994), in: *Conference Record of the 1994 International Display Research Conference*, New York: IEEE, and Playa del Rey, CA: Society for Information Display, pp. 346–349.

Wakimoto, T., Murayama, R., Nagayama, K., Okuda, Y., Nakada, H., Tohma, T. (1996), in: *Digest of the 1996 SID International Symposium*, Santa Ana, CA: Society for Information Display, pp. 849–852.

Yomogida, H. (1997), *Nikkei Electronics* **695** (July 28), 13–14 (in Japanese).

Further Reading

Mauch, R. H., Gumlich, H.-E. (Eds.) (1996), *Inorganic and Organic Electroluminescence, EL 96 Berlin*, Berlin: Wissenschaft und Technik Verlag.

Ono, Y. A. (1995), *Electroluminescent Displays*, Singapore: World Scientific.

Ono, Y. A. (1997), *Annu. Rev. Mater. Sci.* **27**, 283–303.

EXPLOSIVES

(An Addendum to *Encyclopedia of Applied Physics*, Volume 6, pages 327–357.)

Ray Engelke and Stephen A. Sheffield, *Detonation Science and Technology Group, Los Alamos National Laboratory, Los Alamos, New Mexico, U.S.A.*

Introduction 371
1. **Pure Nitrogen Explosives** 373
2. **Reaction-Zone Chemistry** 374
3. **Detonation Reaction-Zone Measurements** 375
4. **Detonation-Products Equation of State** 376
5. **Initiation Measurements** 377
Works Cited 378
Further Reading 379

INTRODUCTION

The technology of explosives is advancing in a number of areas. The impetus for this work is to improve understanding of

1. how and why explosive materials initiate and detonate,
2. how to make more powerful and safe explosives,
3. what the properties of known explosives are, and
4. how to compute, in detail, the initiation, detonation, and action of explosives on adjacent materials.

Because of space limitations, we only mention some of the more important developments and then concentrate on five specific topics that are representative of recent work.

The five areas that will be discussed in some detail are

1. pure nitrogen explosives,
2. molecular-level measurements of the high-pressure chemistry at detonation conditions,
3. mechanical measurements of reaction-zone properties,
4. detonation-products equation of state (EOS), and
5. initiation measurement techniques and examples of the data obtained.

As in the original article (we assume that the concepts discussed in that article are familiar to the reader), we concentrate on condensed-phase, molecular (rather than mixture) explosives. They are the simplest type and hold the most promise for improved understanding, although mixture explosives are of great practical importance in the mining and blasting area. Because of the short time scales and small space scales under which the chemical reactions in condensed-phase explosives take place, much of our understanding remains phenomenological and is based on studies of one-dimensional (1-D) detonation waves and their interaction with the surrounding "confinement" material. There are many areas in which research on condensed-phase high explosives is being done. Much of the work being published relates to weapon or specialty high-explosive applications.

Research on the synthesis of new explosive materials continues. A current emphasis of this work is the synthesis of explosive molecules that contain a high proportion of nitrogen atoms or that are very insensitive to initiation. A pure-nitrogen explosive, N_8 cubane, is discussed in Sec. 1. Both theoretical and chemical synthesis work is aimed at making and understanding this material, as well as other high-nitrogen explosives.

Understanding the chemical reactions that occur in initiating and detonating explosives continues to be an important area of research. The experiments are costly on account of the extreme pressures and temperatures generated in a detonating sample; this results in the complete destruction of the test assembly and a portion of the experimental probes. The very short time scales involved in detonation phenomena also present difficulties. Additionally, detonating explosives are often opaque to the normal electromagnetic radiations used as spectroscopic probes. Because of these factors,

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

the work is challenging and progress is slow. In order to produce significant progress a research program must be sustained for 5 to 10 years.

Several recent studies at universities have been done under conditions designed to provide information relating to an initiation or detonation environment. Examples of this are the ring-up shock experiments of Gupta and coworkers (see Constantinou *et al.*, 1994; Gupta *et al.*, 1995), the temperature-jump experiments of Brill and coworkers (see Brill *et al.*, 1992), and the picosecond-spectroscopy experiments of Dlott and coworkers (see Tas *et al.*, 1997). These studies are helpful in bridging the gap between ambient and initiation and detonation reaction conditions but do not completely simulate the high-pressure/temperature regimes. Other workers continue to use gas-phase measurements to calibrate reaction mechanisms, which are then applied to condensed-phase detonation processes; this is almost certainly a gross oversimplification. Some workers believe that the first reactions in a detonation are of the condensation and ionization types (e.g., Blais *et al.*, 1997). Others think that there are "anomalous" phenomena present (Graham, 1979) that cannot be explained by ordinary chemical principles (see Sec. 2). French workers (see Soulard, 1996) have shown that small changes in the chemical makeup of a material can lead to large changes in performance, e.g., deuterated nitromethane has a critical diameter that is roughly twice that of protonated nitromethane, when fired under the same conditions.

Known high-pressure, condensed-phase chemical principles (le Noble, 1988; Issacs, 1981) have not been adequately incorporated into many explosive-chemistry studies, primarily because relatively few explosive workers have expertise in this area. Important research relating to this topic is discussed in Sec. 2.

Safety concerns have motivated many studies. Low-level insults, such as those from impact, flame, heat, etc., can lead to violent explosions and/or detonations. Various experimental setups have been devised to introduce such inputs into explosives, but these are usually not well instrumented and are difficult to interpret. Complex phenomena such as crushing, cracking, heating, shearing, etc. (or combinations thereof) can occur as a result of such insults, and they are difficult to model and understand. Even "cook-off" experiments, in which a material is controllably heated, are not sufficiently reproducible to yield detailed understanding that can be incorporated into computer models. In cook-off situations, heat generated by chemical reaction is competing with heat loss by thermal conduction to produce a runaway heating situation; under these circumstances the size of the explosive charge is important. Nevertheless, work is progressing in all these areas toward the goal of being able to model these phenomena.

Research continues in the area of homogeneous/heterogeneous initiation and detonation propagation. Understanding of the thermodynamic fluctuations in the flow (hot spots) and what causes them is at the root of this work. Again, because of difficulties in separating out the effects of various types of hot-spot phenomena, progress is difficult. Recently, even crystal imperfections in an explosive have been shown to affect initiation behavior (Borne, 1996), as has the orientation of the crystal axes (in perfect crystals) to the direction of motion of an initiating shock wave (Dick, 1997).

Measuring the detonation reaction zone of various explosives is an active area of work. Ideal explosives (molecular explosives in which the full exothermic chemistry occurs in the detonation reaction zone) are usually the explosives of interest. Some work in this area is discussed in Sec. 3. There is considerable current interest in nonideal explosives (usually mixtures of explosives, fuels, and oxidizers). These present an interesting area because the diameter-effect curves of such materials can exhibit very large decreases in detonation velocity before the failure diameter is reached. Furthermore, the extrapolated infinite-diameter detonation velocity is well below what is expected based on equilibrium thermodynamics (Kennedy, 1993).

Computational techniques are rapidly advancing as available computing capabilities increase. This is particularly true in the areas of parallel processing and adaptive meshing techniques. In problems where the explosive is only part of a much larger problem (several tens of centimeters in size) and it is important to resolve the detonation reaction zone (micron-size zoning), these techniques hold hope for meaningful calculations. Such calculations will probably be possible in the next few years.

In numerical calculations, reactivity is generally implemented by using a global rate form to calculate the state that lies between the unreacted explosive and the fully reacted products. This global rate model is usually a mathematically simple form based on initial shock parameters (such as pressure) and contains parameters that are adjusted to match explosive experiments. Reaction models have been developed by a number of workers, including, e.g., Arrhenius rates, Forest Fire (Mader and Forest, 1976), DAGMAR (Wackerle *et al.*, 1978), Ignition and Growth (Lee and Tarver, 1980), HVRB (History Variable Reactive Burn) (Kerley, 1992), and JTF (Johnson, Tang, Forest) (Johnson *et al.*, 1985). Since these are global rates, usually with one forward reaction, they cannot be expected to reproduce the detailed chemical-reaction mechanisms that are present in reacting explosives. Use of such global reaction models in a wave-propagation code often does a credible job of calculating the mechanical phenomena that occur when explosives are shock initiated or detonating explosives push inert materials.

More complicated multiphase flow models have been developed in order to calculate deflagration-to-detonation transitions (Baer and Nunziato, 1989; Bdzil *et al.*, 1997, 1999a,b), and there is a considerable effort to make these models more accurate. Because of the advent of parallel-processor computers, it is now becoming possible to numerically model discrete particles in three dimensions and to compute (and visualize) the particle crushing and hot spots produced by a passing pressure wave (Baer *et al.*, 1997). This type of modeling will undoubtedly become more important as time goes by.

To do credible modeling, accurate high-pressure equations of state (EOS) are required for both the unreacted explosive and the reaction products. Generally a portion of the unreacted Hugoniot is obtained from experiments done well below the detonation pressure; it is then extrapolated to the von Neumann spike point of the unreacted explosive. Reaction-zone profile measurements are helpful in validating unreacted Hugoniot forms at high pressures. Because of safety concerns, there is also considerable interest in the low-pressure (below 1 GPa) EOS of explosives. Work is under way to study the elastic/plastic nature of explosive materials and to determine the ways in which they crack and deform under the action of low-level insults.

The reaction-product EOS is often modeled by use of an empirical form called the JWL (after Jones, Wilkins, Lee) EOS (Lee *et al.*, 1968). It has a number of parameters that are determined by calibration to metal-pushing experiments. Recently, overdriven detonation experiments have been performed on a plastic bonded explosive (PBX 9501) in which the sound speed at the supported detonation state was measured. Such experiments help constrain the EOS form. These experiments are discussed in Sec. 4.

Measurement techniques are continually being refined and new ones developed, thus improving the experimental data available. For example, shock initiation of explosives can be studied by using multiple-gauge techniques in which gauges are built into the explosive system and used to make 1-D measurements. Pressure or particle-velocity measurements can be made (continuously as a function of time) at from one to as many as twelve locations in the explosive material. Pressure measurements are made with manganin and polyvinylidene-difluoride (PVDF) gauges. Particle-velocity measurements are made with magnetic gauges; this technique is briefly discussed in Sec. 5.

Laser velocity-interferometer measurements are made to study reaction zones (see Sec. 3), and these have, in some cases, subnanosecond time resolution. Techniques for measuring detonation velocity at high and low temperatures are being refined to account for the density and internal-energy changes of the explosive as functions of temperature (Hill, 1997). Proton radiography is used at Los Alamos National Laboratory to visualize the reaction zone and to study detonation corner-turning effects (Sarracino, 1997). Sound-speed measurements of detonated explosive at pressure are being made by optical emission techniques (see Sec. 4) to provide additional product EOS information (Fritz *et al.*, 1996).

1. PURE NITROGEN EXPLOSIVES

There is a proposed new class of chemical explosives that has excited considerable theoretical attention in recent years. The unusual feature of this class is that the molecules are composed entirely of nitrogen atoms. Some members of the class are predicted to have markedly superior explosive performance compared to any currently used chemical explosive.

The primary useful feature of condensed-phase explosives is the transient extremely high pressures that they generate and their concomitant ability to do work rapidly on the materials that surround them. Three factors determine a condensed-phase explosive's ability to do work. These are

1. the number of mols of gas per unit mass (G) produced when detonated,
2. its initial mass density (ρ_0), and
3. the amount of heat released per unit mass when detonated (ΔH_{det}).

The most powerful explosive in current use is HMX (cyclotetramethylene-tetranitramine–$C_4H_8N_8O_8$). For this explosive, the values of the three factors listed above are $G = 40.5\,\text{mol/kg}$, $\rho_0 = 1.90\,\text{g/cm}^3$ (crystal density), and $\Delta H_{\text{det}} = -1.5\,\text{kcal/g}$. The corresponding (Chapman-Jouguet—CJ) detonation pressure and velocity at crystal density are 40.5 GPa and 9.14 mm/μs, respectively (Engelke and Stine, 1990).

Property predictions have been made for various pure-nitrogen structures; e.g., the aza forms of tetrahedrane, cubane, cyclooctatetraene, and various linear molecular forms. The most interesting structure is N_8 cubane because of its

1. predicted very good performance as an explosive,
2. being isoelectronic with the experimentally known, very kinetically stable, carbon/hydrogen cubane, and
3. dissociation to N_2 molecules is a Woodward–Hoffmann symmetry-forbidden reaction, which argues for its kinetic metastability.

For N_8 cubane, the predicted values of G, ρ_0 and ΔH_{det} are $G = 35.7\,\text{mol/kg}$, $\rho_0 = 2.45\,\text{g/cm}^3$ (predicted crystal density), and $\Delta H_{\text{det}} = -4.3\,\text{kcal/g}$. The corresponding predicted detonation pressure and velocity are 131 GPa and 14.5 mm/μs, respectively. Note the very substantial improvement of these values over those of HMX. N_8 cubane's predicted ρ_0 value is obtained from accepted phenomenological methods (Stine, 1981). The G and ΔH_{det} values were obtained by assuming that the detonation produces the chemical reaction $N_8 \rightarrow 4N_2$. High-level quantum-chemical techniques were used to compute ΔH_{det} within this assumption (Engelke and Stine, 1990). Perhaps the most important measure of an explosive's performance is its ability to do work on its surroundings. If this criterion is used as the measure of performance, N_8 cubane is predicted to be a ca. 100% better explosive than HMX.

Reliable quantum-theoretical calculations of N_8 cubane's harmonic vibrational spectrum indicate that it is a stable, fairly rigid, bound structure. Therefore, given the right experimental conditions, it is highly probable that it could be produced as a metastable molecular form.

The salient current question is, "Is N_8 cubane sufficiently kinetically stable to be of practical interest?" The search for an answer to this question is a topic of ongoing research. Earlier computations along a high-symmetry reaction path to dissociation indicate a barrier to dissociation of ca. 40 kcal/mol. Improved more recent calculations indicate the actual value may be closer to 20 kcal/mol (Gagliardi *et al.*, 1997). If the true value of the lowest-energy activation barrier to dissociation or reaction of N_8 is much lower than 20 kcal/mol, the material may be uninteresting for practical uses. If it is near this value or higher, then N_8 cubane could be practically used and would bring about a qualitative advance in the performance of available chemical explosives.

2. REACTION-ZONE CHEMISTRY

There has been some recent progress in experimentally examining chemical kinetic processes in materials shocked into the pressure regime characteristic of detonating condensed-phase explosives.

Approximately fifteen years ago, a number of workers suggested that "anomalous" chemical processes occur at or near the shock front in shocked materials (Graham, 1979); supposedly these processes could not be explained by standard chemical theory. This idea has recently been shown to be incorrect. By analysis of shock recovery experiments on a diverse set of materials that undergo a variety of chemical reactions that are pressure hindered, enhanced, or neutral, Davis and Brower (1996) have shown that Eyring transition-state theory suitably augmented with an activation volume term is capable of explaining all their observations. The pressure and temperature range in which their experiments were conducted was 6 to 16 GPa and 900 to 1400 K.

There have also been recent very direct microscopic observations of chemical kinetic processes in detonating condensed-phase explosives and in surrogate materials shocked into the same pressure regime. Two of these efforts used a time-of-flight mass spectrometer, within which one can detonate or strongly shock energetic materials (Blais *et al.*, 1993). Using this apparatus, it has been shown that the cusp in anthracene's principal Hugoniot (at ca. 17 GPa) is the result of a pressure-driven dimerization reaction in which two anthracene molecules cross link (Engelke and Blais, 1994). It is thought that this cross linking is a Diels–Alder process. This same type of process has also been observed (using the same apparatus) in detonating solid TATB and TNT explosives (Greiner *et al.*, 1993). This indicates that an endothermic pressure-driven reaction is occurring in the early detonation kinetics of explosives containing aromatic rings. It is not known whether such reactions are a dominant feature of the detonation chemistry of the explosives containing aromatic rings or are an irrelevant side channel.

A similar mass-spectroscopic study has been made of detonating liquid nitromethane (NM) (Blais *et al.*, 1997). This material is found to have a very short chemical reaction zone; it is estimated to be ca. 50 μm and 7 ns in space and time duration. The most important result obtained was that the new chemical species found within the detonation reaction zone are the result of condensation reactions—i.e., two or more nitromethane molecules are combining to form more massive molecules in the chemical processes that eventually lead to the detonation products. This is completely different from the low-pressure chemistry of liquid nitromethane, where the decomposition chemistry begins with the unimolecular breaking of the CN bond.

In other work relating to the high-pressure chemistry of NM, Winey (1995) found, via optical spectroscopy experiments, that chemically pure NM shows irreversible changes when rung up to 17 GPa and 725 °C. In similar experiments, Constantinou *et al.* (1994) found similar irreversibilities at 10 GPa and 262 °C in base-sensitized NM. Neither of these sets of experiments took the NM explosives into the same temperature regime as occurs in detonating liquid NM.

There is a current hypothesis that ionic processes may be very important in the early detonation chemistry of some explosives. As a test case to examine how such processes occur, shock experiments were carried out to determine how rapidly the equilibrium

$$\mathrm{CoCl_2}X_2 + 4X \rightleftharpoons \mathrm{Co}X_6^{2+} + 2\mathrm{Cl}^- \qquad (1)$$

is reestablished when the material is strongly shocked (Engelke *et al.*, 1997); in Eq. (1), $X =$ ethanol. The pressures utilized in the experiments covered the range 0.8 to 6.7 GPa. The shocked materials were interrogated with time-resolved visible light spectroscopy. In the experiments at the 6- to 7-GPa pressure range, the equilibrium indicated in Eq. (1) was found to be shifted almost completely to the right, indicating the strong preference for the molecules to be ionized at such pressures. Furthermore, even with ligands as bulky as ethanol molecules, it was found that the equilibrium is reestablished in ca. 500 ns. The hydrogen or proton abstraction processes thought to be important in the detonation of some explosives are undoubtedly much faster than this and have not yet been experimentally explored.

3. DETONATION REACTION-ZONE MEASUREMENT

Detonation-wave profiles can be measured by *in situ* particle-velocity or pressure gauges. Interface velocimetry measurements of the particle velocity at the detonating-explosive/window interface provide the highest time-resolution reaction-zone data. Reaction-zone perturbations resulting from this measurement technique are discussed below. Detonation-wave profile experiments, coupled with computer modeling, provide estimates of global reaction-rate parameters.

There are several types of laser interferometry used to make measurements in both inert and energetic materials. These are VISAR (Barker and Hollenbach, 1972; Hemsing, 1979), Fabry-Pérot (Durand *et al.*, 1977), and ORVIS (Bloomquist and Sheffield, 1983); each of them has undergone improvements over time. Time resolution of streak VISAR (Hemsing, 1985) and ORVIS can be as fast as 200–300 picoseconds, while that of VISAR (with photomultiplier tubes and digitizers for recording the fringes) has a time resolution of 2–3 ns. Fabry-Pérot has a time resolution of 5–10 ns.

All these interferometers operate on the same principle: laser light (usually 514.5- or 532-nm wavelength) is Doppler shifted by a moving surface. This Doppler-shifted light is transduced by the interferometer into fringes that change as the velocity of the surface reflecting the laser light changes. The fringe changes are recorded and then analyzed to obtain the velocity–time profile. Such measurements have been made on nitromethane, PBX 9404, PBX 9501, TATB, PETN, and other explosives, although some of these data have not yet been published (Sheffield *et al.*, 1984; Seitz *et al.*, 1989; Tarver *et al.*, 1997).

A difficulty with an interface measurement of the reaction zone is that the wave profiles are perturbed by the mechanical impedance mismatch at the window through which the observations are made. This is not easily correctable because the impedance of the unreacted explosive is different from that of the reaction products; this makes it impossible to have a window impedance match through the entire reaction zone. Interpretation of the measured profiles require impedance matching to the window and/or computer modeling of the interaction process to obtain an estimate of reaction-zone properties. A recent paper used this technique to estimate the reaction zone of PETN to be less than 5 ns (Tarver *et al.*, 1997).

4. DETONATION-PRODUCTS EQUATION OF STATE

A detonation-products EOS is usually developed for a particular explosive by using the JWL form (Lee *et al.*, 1968) and calibrating it to a copper cylinder expansion experiment (using a wave code calculation). The cylinder test consists of a precisely machined copper cylinder filled with the explosive and detonated at one end. The wall expansion at a particular location is accurately measured as the detonation propagates up the tube. Measuring the wall expansion speed vs. time gives an estimate of the detonation products' ability to do work on the cylinder.

Recently, Fritz *et al.* (1996) used explosively driven metal flyer plates to overdrive a detonation in PBX 9501. An overdriven condition results when the flyer plate sustains the explosive at a pressure state higher than the CJ pressure. In the case of PBX 9501, the CJ pressure is 34.8 GPa, and several experiments were done with the support pressure ranging from 30 to 90 GPa. The velocity of the propagating wave was measured and the particle velocity determined by impedance matching to the standard materials in contact with the explosive. In addition, the explosive was in contact with either liquid bromoform or high-density glass, both of which emit light when they are shocked. The amount of emitted light is directly related to the shock pressure (McQueen *et al.*, 1982); higher pressures give more light.

In these experiments the flyer plate produces a pressure pulse in the explosive, creating an overdriven condition that persists until the rarefaction wave from the back surface of the plate moves into the detonation products and relieves the pressure. By use of fiber-optic probes and photomultiplier tubes to measure the light output as a function of time and making the bromoform or glass thickness vary at different positions on the experiment, the overtake time of the rarefaction overtaking the first shock was obtained. From these data, the sound speed at pressure in the detonation products was measured. In these highly overdriven states, chemical equilibrium was assumed, and thus the measured sound speed is taken to be that characteristic of the equilibrium detonation products.

The sound speed ($= c$) at pressure is an important parameter because it is an EOS derivative [i.e., $c^2 = (\partial P/\partial \rho)_s$]. When the experimentally measured sound-speed data are compared to the JWL sound-speed predictions, there is a considerable discrepancy. This method provides a tool that will help make calibrations of the product EOSs more accurate in the future.

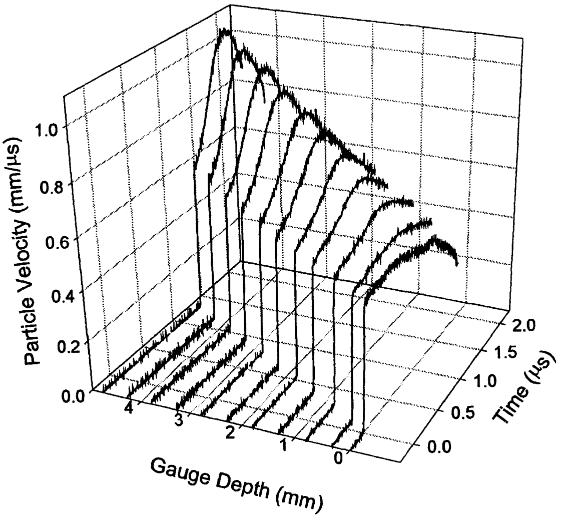

FIG. 1. *In situ* particle-velocity wave forms from a shock initiation in PBX 9501. The first gauge was at the shock input interface and the other ten gauges were embedded in the sample at positions ranging from $\frac{1}{2}$ to 6 mm into the explosive. (Unpublished data of R. L. Gustavsen and S. A. Sheffield, Los Alamos National Laboratory, 1997.)

Measurements on high-carbon–containing molecular explosives have led to a number of predictions that detonation conditions are such that carbon in the reaction products is converted to diamond. In fact, micron-sized diamonds have been recovered from the solid detonation products from confined detonations in some explosives (Greiner and Hermes, 1989). Since the carbon- or graphite-to-diamond transition is an endothermic process, the presence of diamond would be expected to manifest itself in some of the measured mechanical parameters. Indeed, there are slight changes in the product isentropes and other parameters in TNT and TATB that support this prediction. Consequently, this is a current area of both theoretical and experimental activity.

5. INITIATION MEASUREMENTS

Shock-to-detonation (SDT) processes are best measured by *in situ* gauging techniques, which can track the reactive wave growth as a function of time. This provides valuable information to those doing reactive modeling, because it is a rigorous test of the particular global reaction-rate model used to predict the wave growth. Cowperthwaite and Rosenberg (1981) used a multiple-magnetic particle-velocity gauge technique to study nonideal explosives. Both magnetic particle-velocity and manganin pressure measurements are valuable in such studies.

In homogeneous initiation, reactive wave growth occurs well behind the front. In heterogeneous initiation, the wave growth occurs much nearer the front. Most high-density explosives (at 98% or more of theoretical maximum density) exhibit a combination of both types of reactive wave growth.

Particle-velocity wave forms from an *in situ* magnetic gauge experiment (in which PBX 9501 was shock initiated in a gas-gun experiment) are shown in Fig. 1. In this experiment eleven gauges were embedded at different Lagrangian positions in the flow. The wave forms are plotted in 3-D space to allow unambiguous viewing of the details of each gauge measurement. Detailed analysis of the wave forms supports the premise that this high-density explosive exhibits a hybrid homogeneous/heterogeneous nature in the reactive wave growth. The analysis shows that the wave front grows by 15% from the first to the eleventh gauge, and a reactive wave develops behind the front with the intensity increasing as the wave moves into the explosive sample. In fact, the particle velocity of the reactive wave increases until it is double that of the wave front. The maximum in the reactive wave catches the front by about 0.5 μs during the measurement time. The wave developed into a detonation about 10 mm into the sample, well past the last gauge, which was at a depth of 6 mm. These wave forms illustrate the usefulness of *in situ* gauging for examining reactive flow.

List of Works Cited

Baer, M. R., Nunziato, J. W. (1989), in: *Proceedings of the Ninth Symposium (International) on Detonation*, Office of Naval Research Report No. OCNR 113291-7, pp. 293–305. See also Baer, M. R. (1997), in: L. Davison, Y. Horie, M. Shahinpoor (Eds.), *High-Pressure Shock Compression of Solids IV*, New York: Springer-Verlag, pp. 63–82.

Baer, M. R., Graham, R. A., Anderson, M. U., Sheffield, S. A., Gustavsen, R. L. (1997), presented at 1996 Joint Army-Navy-Air Force (JANAF) Combustion Subcommittee & Propulsion Systems Hazards Subcommittee Joint Meeting, Monterey, CA, Nov. 4–8, 1996.

Barker, L. M., Hollenbach, R. E. (1972), *J. Appl. Phys.* **43**, 4669–4675.

Bdzil, J. B., Son, S. F., Menikoff, R., Kapila, A. K., Stewart, D. S. (1997), "Two-phase Models of DDT, Part II," *Phys. Fluids* **9**, 3885–3897.

Bdzil, J. B., Son, S. F., Menikoff, R., Kapila, A. K., Stewart, D. S. (1999a), "Two-phase Models of DDT, Part I," *Phys. Fluids* **11**, 378–402.

Bdzil, J. B., Son, S. F., Menikoff, R., Kapila, A. K., Stewart, D. S. (1999b), "Two-phase Models of DDT, Part III," submitted to *Phys. Fluids*.

Blais, N. C., Fry, H. A., Greiner, N. R. (1993), *Rev. Sci. Instrum.* **64**, 174–183.

Blais, N. C., Engelke, R., Sheffield, S. A. (1997), *J. Phys. Chem. A.* **101**, 8285–8295.

Bloomquist, D. D., Sheffield, S. A. (1983), *J. Appl. Phys.* **54**, 1717–1722.

Borne, L. (1996), French-German Research Institute of Saint-Louis, St. Louis, France, private communication of paper on "Microstructure effect on the shock sensitivity of cast plastic bonded explosives."

Brill, T. B., Brush, P. J., James, K. J., Shepherd, J. E., Pfeiffer, K. J. (1992), *Appl. Spectrosc.* **46**, 900–911, and references therein.

Constantinou, C. P., Winey, J. M., Gupta, Y. M. (1994), *J. Phys. Chem.* **98**, 7767–7778, and references therein.

Cowperthwaite, M., Rosenberg, J. T. (1981), in: *Proceedings of the Seventh Symposium (International) on Detonation*, Naval Surface Weapons Center Report No. NSWC MP 82-334, pp. 1072–1083.

Davis, L. L., Brower, K. R. (1996), *J. Phys. Chem.* **100**, 18775–18783.

Dick, J. J. (1997), *J. Appl. Phys.* **81**, 601–612, and references therein.

Durand, M., Laharrague, P., Lalle, P., Le Bihan, A., Morvan, J., Pujols, H. (1977), *Rev. Sci. Instrum.* **48**, 275–278.

Engelke, R., Stine, J. R. (1990), *J. Phys. Chem.* **94**, 5689–5694.

Engelke, R., Blais, N. C. (1994), *J. Chem. Phys.* **101**, 10961–10972.

Engelke, R., Pettit, D. R., Sheffield, S. A. (1997), *J. Phys. Chem. A* **101**, 1696–1704.

Fritz, J. N., Hixson, R. S., Shaw, M. S., Morris, C. E., McQueen, R. G. (1996), *J. Appl. Phys.* **80**, 6129–6141.

Gagliardi, L., Evangelisti, S., Widmark, P. O., Roos, B. O. (1997) *Theor. Chem. Acc.* **97**, 136–142.

Graham, R. A. (1979), *J. Phys. Chem.* **83**, 3048–3056.

Greiner, N. R., Fry, H. A., Blais, N. C., Engelke, R. (1993), in: *Proceedings of the Tenth International Detonation Symposium*, Office of Naval Research Report No. ONR 33395-12, pp. 665–674.

Greiner, N. R., Hermes, R. (1989), in: *Proceedings of the Ninth Symposium (International) on Detonation*, Office of the Chief of Naval Research Report No. OCNR 113291-7, pp. 1170–1184.

Gupta, Y. M., Pangilinan, G. I., Winey, J. M., Constantinou, C. P. (1995), *Chem. Phys. Lett.* **232**, 341–345, and references therein.

Hemsing, W. F. (1979), *Rev. Sci. Instrum.* **50**, 73–78.

Hemsing, W. F. (1985), in: *Proceedings of the Eighth Symposium (International) on Detonation*, Naval Surface Weapons Center Report No. NSWC MP 86-194, pp. 468–472.

Hill, L. G. (1997), Los Alamos National Laboratory, private communication about work in progress.

Issacs, N. S. (1981), *Liquid Phase High-Pressure Chemistry*, New York: Wiley.

Johnson, J. N., Tang, P. K., Forest, C. A. (1985), *J. Appl. Phys.* **57**, 4323–4334.

Kennedy, D. L., Jones, D. A. (1993), in: *Proceedings of the Tenth International Detonation Symposium*, Office of Naval Research Report No. ONR 33395-12, pp. 665–674.

Kerley, G. I. (1992), *CTH Equation of State Package: Porosity and Reactive Burn Models*, Sandia National Laboratories Report No. SAND92-0553, Albuquerque, NM.

Lee, E. L., Hornig, H. C., Kury, J. W. (1968), *Adiabatic Expansion of High Explosive Detonation Products*, Technical Report No. UCRL-50422, Lawrence Livermore National Laboratory.

Lee, E. L., Tarver, C. M. (1980), *Phys. Fluids* **23**, 2362–2372.

le Noble, W. J. (1988), *Organic High-Pressure Chemistry*, New York: Elsevier.

Mader, C. L., Forest, C. A. (1976), *Two-Dimensional Homogeneous and Heterogeneous Detonation Wave Propagation*, Los Alamos Scientific Laboratory Report No. SA-6259, p. 60 (App. C).

McQueen, R. G., Hopson, J. W., Fritz, J. N. (1982), *Rev. Sci. Instrum.* **53**, 245–250.

Sarracino, J. S. (1997), private communication about work in progress with coworkers at Los Alamos National Laboratory.

Sheffield, S. A., Bloomquist, D. D., Tarver, C. M. (1984), *J. Chem. Phys.* **80**, 3831–3844.

Seitz, W. L., Stacy, H. L., Engelke, R., Tang, P. K., Wackerle, J. (1989), in: *Proceedings of the Ninth Symposium (International) on Detonation*, Office of the Chief of Naval Research Report No. OCNR 113291-7, pp. 657–669.

Soulard, L. (1996), private communication with Soulard and coworkers at Commissariat à l'Energie Atomique, CEV-M, BP.7, Courtry, France.

Stine, J. R. (1981), *Prediction of Crystal Densities of Organic Explosives by Group Additivity*, Los Alamos National Laboratory Report No. LA-8920.

Tarver, C. M., Breithaupt, D. R., Kury, J. W. (1997), *J. Appl. Phys.* **81**, 7193–7202.

Tas, G., Hambir, S. A., Franken, J., Hare, D. E., Dlott, D. D. (1997), *J. Appl. Phys.* **82**, 1080–1087, and references therein.

Wackerle, J., Rabie, R. L., Ginsberg, M. J., Anderson, A. B. (1978), in: *Proceedings of the Symposium on High Dynamic Pressures: Behavior of Dense Media Under High Dynamic Pressures*, Paris: Editions du Commissariat à l'Energie Atomique, Centre d'Etudes Nucléaires de Saclay, pp. 127–138.

Winey, J. M. (1995), "Time-Resolved Optical Spectroscopy to Examine Shock-Induced Decomposition in Liquid Nitromethane," Ph.D. Dissertation, Washington State University, Pullman, WA.

Further Reading

APS Topical Conferences—*Proceedings of the Conference of the American Physical Society Topical Group on Shock Compression of Condensed Matter.* Published by: Plenum Press (1985); American Institute of Physics (AIP) Conference Proceedings (1981, 1993, 1995); North-Holland (1983, 1987, 1989, 1991).

Cheret, R. (1993), *Detonation of Condensed Explosives*, New York: Springer-Verlag.

Cooper, P. W. (1996), *Explosives Engineering*, New York: VCH.

Cooper, P. W., Kurowski, S. R. (1996), *Introduction to the Technology of Explosives*, New York: VCH.

Detonation Symposia—*Proceedings of the Third (1960) to Eleventh (1998) International Symposia on Detonation*, sponsored by the Office of Naval Research, Washington: U.S. GPO.

Graham, R. A. (Ed. in Chief), *High Pressure Shock Compression of Solids* (1993 to present), New York: Springer-Verlag. This is a series of books that have chapters written by different authors. In each of the books there is one or more articles relating to explosives.

MESOSCOPIC SYSTEMS

(An Addendum to *Encyclopedia of Applied Physics*, Volume 10, pages 81–111.)

YOSHIMASA MURAYAMA, *Advanced Research Laboratory, Hitachi, Ltd., Hatoyama, Saitama, Japan*

Introduction 381
1. **Atomic Surfaces** 381
2. **Atomic Wires** 383
3. **Atomic Clusters** 385
4. **Future Prospects** 386
Works Cited 387

INTRODUCTION

Mesoscopic systems cover all types of extremely tiny devices made of semiconducting or metallic materials. The original definition of the mesoscopic device comes, not from its tininess, but from whether electrons in the device behave coherently. Coherence assures us that the device is ruled by quantum mechanics. As was described in the original article, any system will suffer, in principle, more and more from *decoherence* when the number of its degrees of freedom increases. Here, decoherence means processes where quantum systems decohere, *i.e., lose their coherence*, and eventually become *incoherent* systems. In that sense, these tiny devices are expected to behave partially *coherently*.

In condensed matter, any environment of a device is considered to have a variety of degrees of freedom other than those of the electron of concern. All such degrees, more or less, cause the electron to decohere, as long as they subject the electron to *inelastic* scattering, because every electron keeps its coherence if and only if its energy is not changed. These degrees belong to phonons, photons, and even other electrons. Crystalline imperfections such as impurities, grain boundaries, surfaces, and interfaces are considered to cause only elastic scattering.

The original article discussed many possible decohering mechanisms in tiny 3-, 2-, 1- and 0-dimensional devices. However, it did not discuss atomic-scale 2-, 1- and 0-dimensional systems, which have appeared very recently and are attracting increased attention thanks to the use of the scanning tunneling microscope (STM) and related techniques. Readers are referred to TUNNELING MICROSCOPY AND SPECTROSCOPY for details on the observation and fabrication techniques applied for atom-order samples. This Addendum covers atomically small surfaces, wires, and clusters, whose study has become possible only recently.

We can now actually observe images of arrays of atoms in 0, 1, and 2 dimensions using an STM. In one sense these images are related to the spatial distribution of electrons whose energy is situated near the Fermi level. In most cases, the images vary depending on the applied electric potential, even when the gap between the STM probe tip and the sample surface is kept constant. Thus, an STM can be used to perform scanning tunneling spectroscopy (STS). Theories to interpret the observed images are already developed. However, subtle questions still remain as to whether we can observe the electronic wave function itself, including its phase, by means of tunneling microscopy.

An STM can also be used to manipulate atoms one by one (or at least, as very tiny atomic clusters), for example by extracting, moving, or depositing them. Atoms on a sample surface can be extracted from the surface, and similarly atoms on the surface of the probe tip can be deposited onto the surface of a sample by applying a high electric voltage between the tip and sample. This manipulator is often applied to fabricate atomic arrays (1-dimensional quantum systems), atomically small terraces (2-dimensional quantum systems), and atomically small clusters (0-dimensional quantum systems).

These atomic-order–sized electronic devices are often called *atomics*; however, more plausibly, they should be called *atomic electronics* or *atom-electronics*, since the main role for the phenomena of interest is played by electrons, not by atoms.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

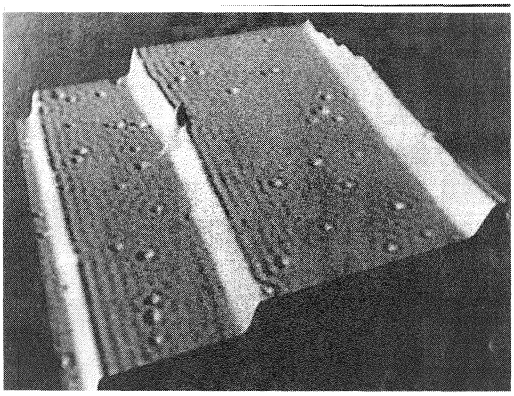

FIG. 1. A 500 × 500-Å^2 constant-current image of a Cu(111) surface. Three monoatomic steps and a few tens of point defects are visible. Spatial oscillations with a period of ~15 Å are evident. The small circles surrounded by circular ripples are point defects on the surface, whose origin is not known. (After Crommie *et al.*, 1993a.)

1. ATOMIC SURFACES

It is well recognized that the Bloch theorem holds formally in a perfect crystal. However, for the Bloch theorem to work, it is implicitly assumed that the crystal system assures that the Bloch states are coherent throughout the system. To have coherent Bloch states, an infinite long-range crystalline order and no decohering mechanisms are needed. Regarding the long-range order of the surface, the situation is the same. Coherence must exist for all the surface Bloch states.

In ordinary bulk crystalline systems, electronic states have the same periodicity as the crystal. In atomically tiny systems, the states are specifically modulated by means of the envelope wave functions according to the boundary conditions the systems are subject to. This situation is very similar to the *charge-density wave* (CDW), where the period is ordinarily incommensurate with that of the crystal, although the CDW is constituted by the Bloch states themselves.

The first astonishing observation of atom-electronics properties on an atomic surface was made by IBM scientists in 1993 (Crommie *et al.*, 1993a; Fig. 1). They used an STM to scan a surface with a few atomic steps (terraces) and obtained an image of electronic states straightforwardly. The surface monatomic step creates a potential step for the surface electrons. The potential step imposes a new boundary condition on Bloch electrons and modulates the wave function so that the distribution probability is affected by this boundary condition; that is, a CDW is created perpendicularly to the potential step, and the electronic charge is distributed in the STM image as a standing ripple wave. The ripple is obviously caused by the interference between the back- and forth-going waves.

Afterwards, the IBM scientists proceeded to create a do-it-yourself boundary condition by utilizing an STM-related technique. Their man-made corral was made of Fe atoms that encircled a certain Cu surface area, and the generated CDW was confined within the corral. Although not displayed here, the co-centered circular CDW ripple is well worth seeing (Crommie *et al.*, 1993b).

Several documents refer to these observed images as a "wave function". Aharonov *et al.* (1993) proposed that it could be possible to observe a wave function itself, if a "protective measurement" is applied to it. However, it is commonly recognized that it is difficult to observe a wave function itself by using an ordinary measurement process. Thus, it is safe to say that what we observe is only a squared wave function that represents the probability distribution of electrons over various states. To establish this ripple, it is necessary to have coherence in two senses. The first is the coherence for a single electron. This assures that wave functions are coherent in their initial and reflected directions. The other means that an ensemble of electrons must be coherent in the same sense as the people in the electron microscope field tell that an electron source is "coherent". This meaning ensures that all the electrons involved in the CDW should be the same as one another, *i.e.*, under the same or very similar quantum conditions. Otherwise, all waves composing a CDW must be randomly added and the ripple would be smeared out.

All these features represent the local density of states (LDOS) defined by

$$\rho(E,\boldsymbol{r}) = \sum_i |\psi_i(\boldsymbol{r})|^2 \delta(E - E_i), \tag{1}$$

where $\psi_i(\boldsymbol{r})$ is the wave function of state i with its eigenenergy E_i.

The atomic reconstructed surface is another interesting example of a 2-dimensional atom-

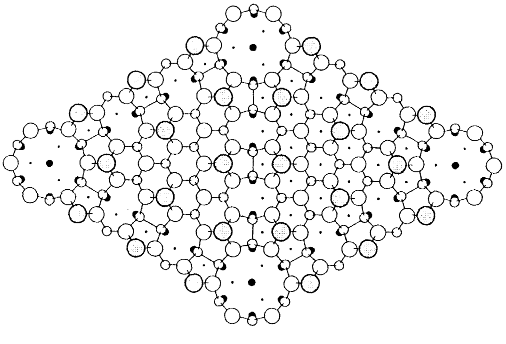

FIG. 2. Schematic representation of Takayanagi's DAS model (after Takayanagi *et al.*, 1985). Four Si atoms on the corners of the lowest layer 7 × 7 rhombus are lacking. The small open circles neighboring small solid circles combine pairwise to form dimers. The next stacking-fault-layer atoms are denoted with larger open circles. Their configurations are faulted between the right and left half rhombus. The darkened large circles are for adatoms on the topmost layer.

electronic system. Under reconstruction, atoms are displaced to construct a new periodicity and the surface Bloch electrons follow the same period.

A marvelous example is related to the complicated reconstruction of Si atoms on the Si(111) surface, as Binnig and Rohrer observed by STM (Binnig *et al.*, 1983). (For more details, see, for example, SILICON, CRYSTALLINE.) It had long been known that the surface constitutes a 7 × 7 structure, and this was later modelled by Takayanagi *et al.* (1985) as a DAS (dimer-adatom-stacking fault) model based on low-energy electron diffraction (LEED) observation. The schematic structure is shown in Fig. 2. This 7 × 7 structure was precisely investigated, again by STM, later on and the structure proposed by Takayanagi was confirmed. Such surface reconstruction is realized as a very long-range order, which tells us that the coherence of the surface Bloch states is also long.

To further investigate and apply the reconstructed surfaces to actual devices, though, it is necessary to manipulate or decorate the surfaces. Hashizume *et al.* (1993) first reported the surface of a Cu(111)-(1×1) sample decorated with C_{60} molecules. These molecules were neatly arrayed forming a triangular structure as shown in Fig. 3.

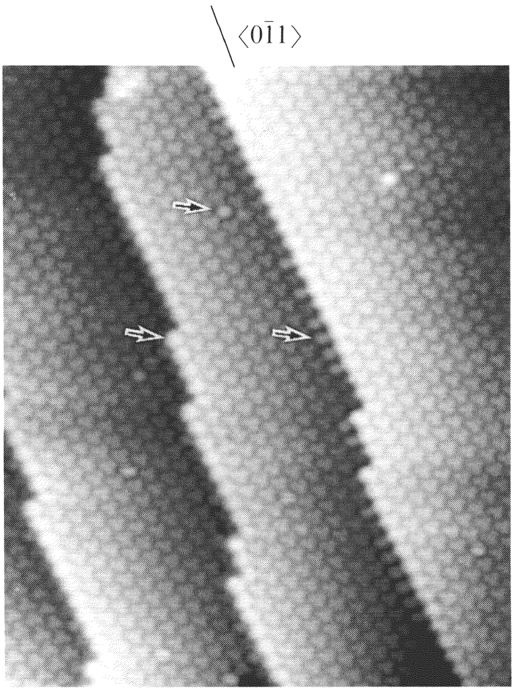

FIG. 3. An STM image of the Cu(111) surface covered completely with a monolayer C_{60} film (area: 300 × 380 Å²). Individual molecules are imaged in a three-lobe clover shape. One lobe of each molecule is oriented either upward or downward along the $\langle 0\bar{1}1 \rangle$ direction. The arrows indicate defects in alignment. (After Hashizume *et al.*, 1993.)

The C_{60} molecules themselves have interesting features. They form an insulating molecular fcc crystal (Kroto, 1985), but if doped with alkali-metal atoms, they can be superconducting. Small alkali-metal atoms are believed to situate themselves between the molecules, and the electronic states are sufficiently modified to force the crystal to be superconductive (Haddon *et al.*, 1991; Hebard *et al.*, 1991; Rosseinsky *et al.*, 1991; Holzer *et al.*, 1991; Chen *et al.*, 1991; Tanigaki *et al.*, 1991).

2. ATOMIC WIRES

An atom is very stable and keeps its coherence, as long as it does not interact with any other degrees of freedom. For example, a hydrogen atom in its ground state remains almost forever in that state, unless it absorbs photons or collides with other high-energy particles, atoms, or molecules. If one goes from an atom to a linear chain of atoms, what happens? A long linear molecule is still coher-

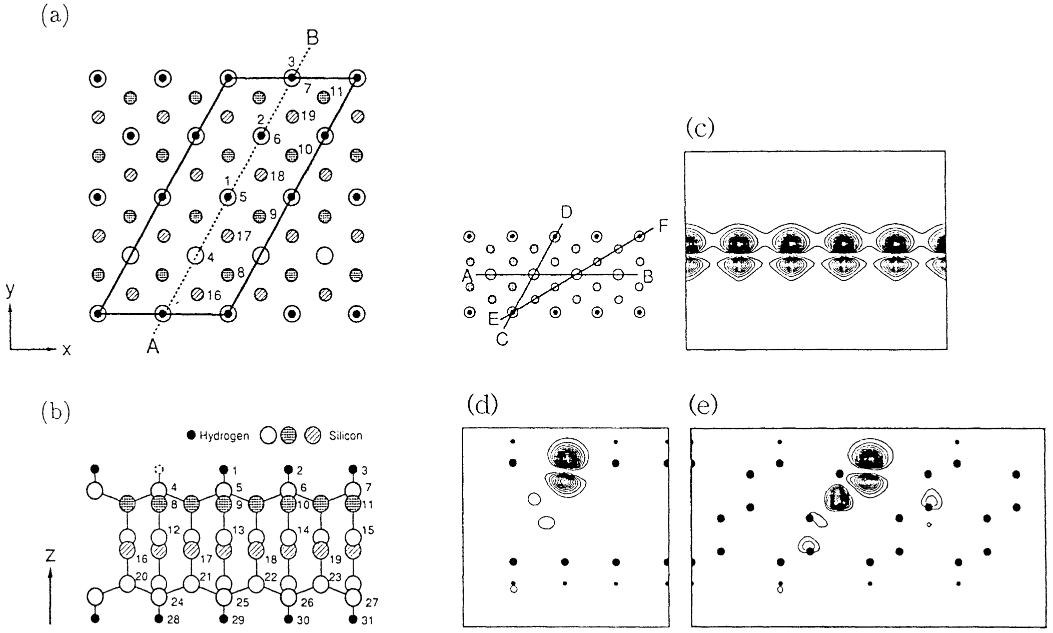

FIG. 4. An atomic wire on the H-terminated Si(111) surface. (a) Top view. A horizontal array of open circles through No. 4 is expected to form a quantum wire of depassivated Si atoms. A rhombus is the 2 × 4 supercell to calculate. (b) Side view along the A-B line in (a). Small solid and larger circles represent, respectively, H and Si atoms. A dotted circle shows the depassivated H site. (c), (d), (e) The charge-density contour maps of the midgap state at the symmetry Γ point (c) along line A-B, (d) along line C-D, and (e) along line E-F. (After Watanabe *et al.*, 1995.)

ent and stable. However, a problem occurs if this system is lying on a substrate, in that a variety of interactions are likely to occur between the chain of atoms and its substrate.

1-dimensional atomic (quantum) wires have been investigated, so far, mostly theoretically, because the fabrication technique is far beyond the present attainable level. In studies on practical quantum wires, the effect of the interaction with the substrate is very important. Readers are recommended to refer to the review paper by Aono *et al.* (1993) and the conference proceedings of Serena and Garcia (1997) and Joachim and Roth (1997).

Quantum wires are most easily fabricated by removing H atoms in one dimension from a fully passivated Si surface. This can be done by applying a high voltage between the STM tip and the passivated surface. This sort of atomic wire is sometimes called a dangling-bond (DB) wire or a depassivated (DP) wire. Watanabe *et al.* (1995) tried a pseudopotential, density-functional approximation calculation for a DB wire on an Si(111) surface and showed that the wire becomes truly conductive (Fig. 4).

How to obtain a p–n junction utilizing these quantum wires, something not yet achieved, is an important question from a practical viewpoint. It is also very interesting from a device-physics viewpoint. In an earlier study on what dopant concentrations are needed, to realize n- or p-type atomic-scale conductors (Murayama, 1995), the author reasoned as follows. Doping is effective in conventional bulk semiconductors when its level is something like $(a/a_B)^3$, where a is the lattice constant and a_B is the appropriate Bohr radius of a hydrogenlike dopant atom, which is on the order of several tens of angstroms. This means that 1 ppm or a slightly larger dopant concentration works quite well in the bulk case. However, in 1-dimensional wires, a/a_B means approximately 1%, so that a great quantity of dopants is needed for the wire to work well. This large quantity may lead to a severe stability problem in such heavily doped semiconducting substances.

3. ATOMIC CLUSTERS

The original article already reviewed 0-dimensional quantum systems, where conductance is characterized by the Coulomb blockade, the single-electron tunneling (SET) oscillation, the turnstile effect, and so on. Studies that should be referred to in this field include Fulton and Dolan (1987), Delsing *et al.* (1989), Geerligs *et al.* (1990), and Averin and Likharev (1991). The main point of the underlying physics is that electrons can be transferred via tunneling only under the correlation effect, which is efficient for ultrasmall tunnel junctions. In larger tunnel junctions, the electron wave function spreads, and then the correlation effect is diminished in the same manner that makes it almost insignificant in bulk metals.

A unit of charge q is defined by CV, where V is the applied electric potential and C is the capacity of the tunnel junction. In ordinary capacitors, the charge q or $-q$ on each electrode is kept fixed. In contrast, in tunnel junctions, one unit of charge e can tunnel and eventually annihilate to give $q-e$ and $-(q-e)$. However, if a charge q between $-e/2$ and $e/2$ is diminished by one unit e via tunneling through the junction, the electrostatic energy of the capacitor $(q-e)^2/2C$ becomes larger than before, say $q^2/2C$. That is, the energy is increased rather than decreased by tunneling. This is in accordance with the Coulomb repulsion energy, as first suggested by Likharev (see Averin and Likharev, 1991). In other words, tunneling conductance is suppressed. This phenomenon is known as the Coulomb blockade.

From a device-application viewpoint, the Coulomb blockade effect applies to few actual nonlinear quantum devices. This is mainly because the electrostatic energy is too far below the ambient thermal energy kT (25 meV at room temperature). If kT is equated to $e^2/2C$, the capacity C must be as small as 3 aF ($= 3 \times 10^{-18}$ F). This value corresponds to a $3 \times 3\text{-nm}^2$ capacitor with a 3-nm gap and a dielectric constant $\varepsilon = 4\varepsilon_0$ (ε_0 denotes the permittivity of vacuum). Such a small capacitor cannot be fabricated with conventional photolithographic techniques. For this reason, plausible devices could initially only be operated at low temperatures (Nakazato *et al.*, 1993).

However, Yano *et al.* (1994) first observed room-temperature operation of the Coulomb blockade phenomena within granular polycrystalline Si ultrathin films. The device structure they examined was similar to the conventional electrically alterable MOS (EAMOS), or a floating-gate nonvolatile-memory device, since the thickness of the granular Si film was so thin that the potential produced by the gate voltage randomly varied from grain to grain. When the gate voltage was high enough, percolation channels formed. Otherwise, the channels were broken into pieces of quantum dots. In the latter case, every dot would have acted as an island for the Coulomb blockade capacitor. The size of these islands seems to have been sufficiently small compared to the size of the above mentioned 3-aF capacitor.

Figure 5(a) shows a model for an ultrasmall MOS with a floating island embedded inside the gate insulator. In this device, an increase in the gate voltage V_g causes the "write-in" into the floating island (the floating gate) to be changed. As suggested in Fig. 4(b), the floating charge shifts the effective gate voltage by a specific amount. This shift in V_g is given by the relationship

$$\Delta V_g \sim \frac{q}{C_{gc}}, \tag{2}$$

where C_{gc} is the capacitor specific to the circuit. Thus, the original V_g-vs-I_d characteristic where I_d is the drain current becomes a shifted $(V_g + \Delta V_g)$-vs-I_d characteristic as shown in Fig. 5(c). The most significant part of this phenomenon is that the shift in V_g is quantized. Certainly, a single floating charge seems to block a single conducting channel. As described in the original article, the Landauer formula gives the conductance in terms of the number of channels. With more channels, the device becomes more conductive. A further increase in V_g may reduce the number of channels one by one, depending on whether the islands are charged in a floating manner. The observed data are shown in Fig. 6.

A "single-electron box" with a shunting

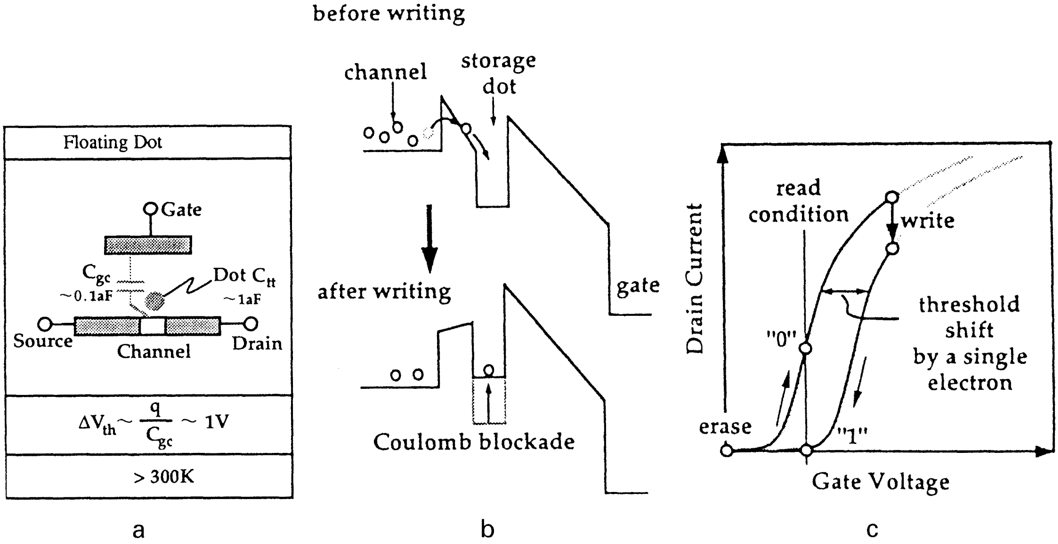
Floating Dot
Gate
C_{gc}
~0.1aF
Dot C_{tt}
~1aF
Source
Channel
Drain
$\Delta V_{th} \sim \frac{q}{C_{gc}} \sim 1V$
> 300K
before writing
channel
storage dot
after writing
gate
Coulomb blockade
Drain Current
read condition
write
"0"
threshold shift by a single electron
erase
"1"
Gate Voltage
a
b
c

FIG. 5. (a) Structure model of a MOSFET with a floating-island gate embedded inside the gate insulator. (b) A schematic presentation of the MOSFET with and without the floating gate charged. If charged, the side potential causes the channel to conduct less current. (c) Two V_g-vs-I_d's are contrasted. One is with the floating gate uncharged and the other is with V_g shifted according to a charged floating gate. (After Yano *et al.*, 1994, © 1994 IEEE.)

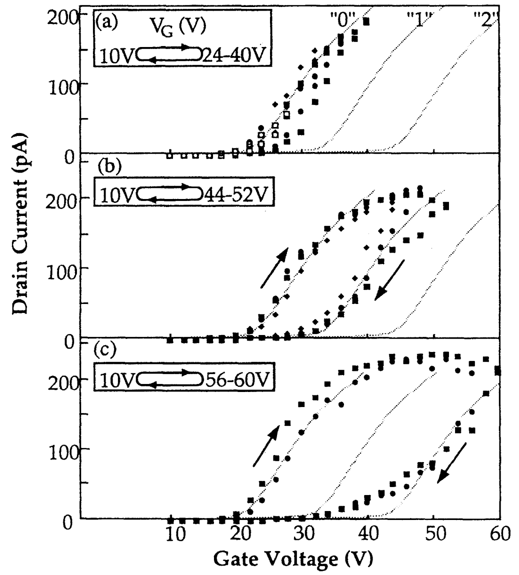

FIG. 6. Measured V_g vs. I_d. Each group of the data corresponds to a single floating-gate condition. Lightly drawn continuous lines trace the characteristics for different numbers of charges inside the islands. (After Yano *et al.*, 1994. © 1994 IEEE.)

linear resistor and "single-electron memory" with a *nonlinear* shunting resistor are compared in Fig. 7. The voltage at the floating dot exhibits a hysteretic behavior as shown in the figure, indicating that operation as a single-electron memory device is possible. This type of memory operation was first observed by Nakazato *et al.* (1993), but at a low temperature.

The device introduced here has a strong advantage in that it works at room temperature. However, how the gate voltage can be precisely controlled is another formidable problem. In the experiment, the ultrathin Si film unexpectedly produced a network of floating islands. Therefore, the gate voltage itself varied a great deal from sample to sample. In contrast, the shift in the gate voltage was definite as a result of the Coulomb blockade effect with an integral number of floating charges, which is the biggest advantage of this device.

4. FUTURE PROSPECTS

This Addendum has reviewed recent trends in atom-electronic research. From the device application point of view, this area seems

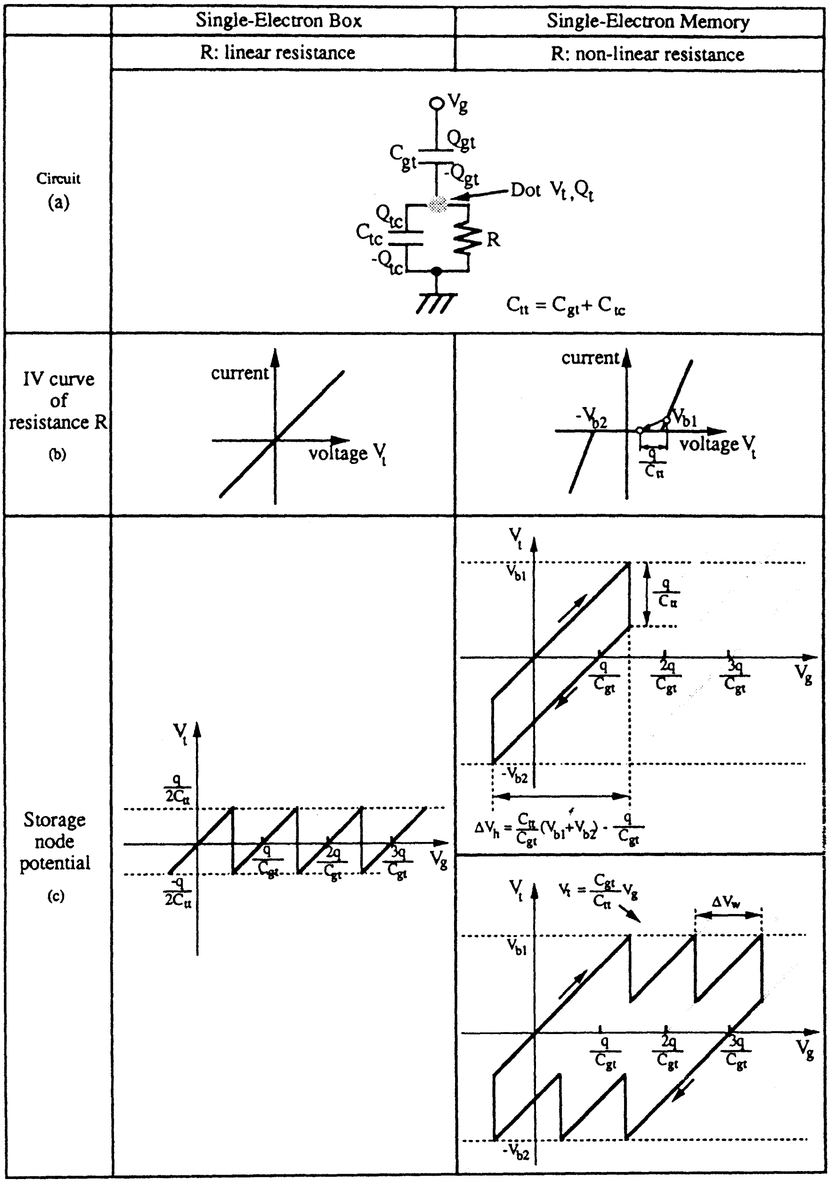
Single-Electron Box
Single-Electron Memory
R: linear resistance
R: non-linear resistance
Circuit
(a)
Vg
Qgt
Cgt
-Qgt
Dot Vt,Qt
Qtc
Ctc
-Qtc
R
Ctt = Cgt+ Ctc
IV curve of resistance R
(b)
current
voltage Vt
-Vb2
Vb1
q/Ctt
Storage node potential
(c)
Vt
q/2Ctt
-q/2Ctt
q/Cgt
2q/Cgt
3q/Cgt
Vg
ΔVh = Ctt/Cgt(Vb1+Vb2) - q/Cgt
Vt = Cgt/Ctt Vg
ΔVw

FIG. 7. Comparison between the single-electron box and the single-electron memory. The two are differentiated by a shunting linear or nonlinear resistor. (After Yano *et al.*, 1994, © 1994 IEEE.)

likely to flourish in the next century, thanks to the STM-related fabrication technologies.

The size of the devices described so far seems to be just between the photolithographically processable size and the atomically maneuverable size. The size of an atom is on the order of 0.1 nm, while an approach from a macroscopic size using photolithography is possible only down to a few hundred nanometers. The two approaches have a gap in size amounting to a ratio of about a hundred, since actual devices are expected to be, at least, on the order of ~10 nm. On a smaller scale, the devices may not work stably, even though the packing density of the devices would be increased.

For example, an array of single atoms in a chain would be too fragile to manipulate without damaging it by, for example, electrostatic discharging. It must work as if it is a "quantum fuse". This effect was a particularly annoying nuisance in conventional tiny MOSFETs. This reliability problem will be the last, but may be the most serious, obstacle in the development of atom-electronic devices.

How we can access these devices is also a serious problem. Maneuverable leads to devices are on the order of ~1 μm. However, actual devices are sized to only ~10 nm in this case. The ratio of the two scales is again ~100. Thus, we must gradually expand the scale of the leads from ~10 nm to ~1 μm.

When atomic clusters are intended to be used as a 2-dimensional memory array, how to achieve rapid access and how long is needed to "write-in" and "read-out" are the critical engineering bottlenecks.

In the early years of the next century, many nanoscale devices will be proposed. However, the study will have to continue mainly from a purely scientific view for a while, rather than from actual application-oriented interests. Nevertheless many breakthroughs are still expected, because today's prevailing Si devices are so tough and seem likely to survive longer than once expected. The realization of practical nanoscale devices may be considerably later than had once been hoped.

This atomic-scale device technology, however, has a promising future and an important role to play in human technological development. The ultimate goal is to mimic the human brain with a network of nanoscale devices on the order of synapse size! In addition, although nobody knows whether these devices will work *coherently* or not, coherent operation could make the "quantum computer" attainable.

List of Works Cited

Aharonov, Y., Anandan, J., Vaidman, L. (1993), *Phys. Rev. A* **47**, 4616–4626.

Aono, M., Kobayashi, A., Grey, F., Uchida, H., Huang, D.-H. (1993), *Jpn. J. Appl. Phys.* **32**, 1470–1477.

Averin, D. V., Likharev, K. K. (1991), in: B. L. Al'tshuler, P. A. Lee, R. A. Webb (Eds.), *Mesoscopic Phenomena in Solids*, Amsterdam: North-Holland, pp. 173–272.

Binnig, G., Rohrer, H., Gerber, C., Weibel, E. (1983), *Phys. Rev. Lett.* **50**, 120–123.

Chen, C.-C., Kelty, S. P., Lieber, C. M. (1991), *Science* **253**, 886–888.

Crommie, M. F., Lutz, C. P., Eigler, D. M. (1993a), *Nature* **363**, 524–527.

Crommie, M. F., Lutz, C. P., Eigler, D. M. (1993b), *Science* **262**, 218–220.

Delsing, P., Likharev, K. K., Kuzmin, L. S., Claeson, T. (1989), *Phys. Rev. Lett.* **63**, 1861–1864.

Fulton, T. A., Dolan, G. J. (1987), *Phys. Rev. Lett.* **59**, 109–112.

Geerligs, L. J., Anderegg, V. F., Holweg, P. A. M., Mooij, J. E. (1990), *Phys. Rev. Lett.* **64**, 2691–2694.

Haddon, R. C., Hebard, A. F., Rosseinsky, M. J., Murphy, D. W., Duclos, S. J., Lyons, K. B., Miller, B., Mosamilia, J. M., Fleming, R. M., Kortan, A. R., Glarum, S. H., Makhija, A. V., Muller, A. J., Eick, R. H., Zahurak, S. M., Tycko, R., Dabbagh, G., Thiel, F. A. (1991), *Science* **350**, 320–322.

Hashizume, T., Motai, K., Wang, X. D., Shinohara, H., Saito, Y., Maruyama, Y., Ohno, K., Kawazoe, Y., Nishina, Y., Pickering, H. W., Kuk, Y., Sakurai, T. (1993), *Phys. Rev. Lett.* **71**, 2959–2962.

Hebard, A. F., Rosseinsky, M. J., Haddon, R. C., Murphy, M., Glarum, V., Palstra, T. T. M., Ramirez, A. P., Koratan, A. R. (1991), *Nature* **350**, 600–601.

Holzer, K., Klein, O., Huang, S.-M., Kaner, R. B., Fu, K.-J., Whetten, R. L., Diederich, F. (1991), *Science* **252**, 1154–1157.

Joachim, C., Roth, S. (Eds.) (1997), *Nanowires*, Dordrecht: Kluwer Academic.

Kroto, H. W., Heath, J. R., O'Brien, S. C., Curl, R. F., Smalley, R. E. (1985), *Nature* **318**, 162–163.

Murayama, Y. (1995), US and Japan patents pending.

Nakazato, K., Blaikie, R. J., Cleaver, J. R. A., Ahmed, H. (1993), *Electron. Lett.* **29**, 384–385.

Rosseinsky, M. J., Ramirez, A. P., Glarum, S. H., Murphy, D. W., Haddon, R. C., Hebard, A. F., Palstra, T. T. M., Kortan, A. R., Zahurak, S.M., Makhija, A. V. (1991), *Phys. Rev. Lett.* **66**, 2830–2832.

Serena, P. A., Garcia, N. (Eds.) (1997), *Atomic and Molecular Wires*, Dordrecht: Kluwer Academic.

Takayanagi, T., Tanishiro, Y., Takahashi, M., Takahashi, S. (1985), *J. Vac. Sci. Technol.* **A3**, 1502–1506.

Tanigaki, K., Ebbersen, T. W., Saito, S., Mizuki, J., Tsai, J. S., Kubo, Y., Kuroshima, S. (1991), *Nature* **352**, 222–223.

Watanabe, S., Ono, Y. A., Hashizume, T., Wada, Y., Yamauchi, J., Tsukada, M. (1995), *Phys. Rev.* **B52**, 10768–10771.

Yano, K., Ishii, T., Hashimoto, T., Kobayashi, T., Murai, F., Seki, K. (1994), *IEEE Trans. Electron Devices* **41**, 1628–1638.

MUONIC, MESONIC, AND BARYONIC ATOMS AND MOLECULES

(An Addendum to *Encyclopedia of Applied Physics*, Volume 11, pages 55–78.)

F. J. HARTMANN, *Physik-Department, Technische Universität München, Garching, Germany*

Introduction 389

1. **Energy Levels in Exotic Atoms** .. 389

1.1 Electromagnetic Binding Energies 389

1.1.1 Muonic Hydrogen 390

1.1.2 Metastable Antiprotonic Helium 390

1.2 Hadronic Atoms: the Strong Interaction 392

1.2.1 Pionic Atoms 392

1.2.2 Antiprotonic Atoms 393

2. **Formation of Exotic Atoms** 394

2.1 Kinetic Energy of the Exotic Particle Before Capture 394

2.2 Kinetic Energy of the Exotic Atom After Capture 395

2.3 The First-Populated States in the Exotic Atom 395

3. **The Exotic-Atom Cascade** 396

3.1 The Exotic-Atom Cascade in Elements with $Z > 2$ 396

3.2 Acceleration During the Cascade 396

3.3 Transfer Processes 397

4. **Application of Exotic Atoms** 397

4.1 Application of Muonic X-Rays in Diagnostics 397

5. **Muon-Catalyzed Fusion** 397

5.1 Basic Knowledge 397

5.1.1 Atomic Capture 398

5.1.2 Slowing Down 398

5.1.3 Transfer 398

5.1.4 Molecular Formation 398

5.1.5 Nuclear Fusion and Sticking ... 399

5.2 Applications of Muon-Catalyzed Fusion 399

5.2.1 Generation of 14-MeV Neutrons 399

5.2.2 Generation of Low-Energy Muons with the Help of Muon-Catalyzed *dt* Fusion 399

Works Cited 399

INTRODUCTION

The last years saw some highlights in the research on muonic, pionic, and antiprotonic atoms and on the muon-catalyzed-fusion cycle. Especially in the field of exotic-atom binding energies and of the formation of exotic atoms some completely new results emerged. They are here briefly described. Reference to the original article will be marked by a roman I.

1. ENERGY LEVELS IN EXOTIC ATOMS

1.1 Electromagnetic Binding Energies

Progress continues both in theory *and* in experiments on the electromagnetic binding energies in exotic atoms. Attention in this field has turned more and more to the simplest systems, muonic protium (protium is the most abundant isotope of hydrogen, with only one proton as nucleus) and antiprotonic He. This has to do partly with the fact that laser-resonance techniques have been applied or are planned to be employed to determine binding-energy differences in the atoms with unprecedented accuracy. All experiments of this kind proceed in roughly the same way: An exotic atom is formed in a sufficiently long-lived state; the muonic protium atom $p\mu$ in its $2s$ state (principal quantum number $n = 2$, angular-momentum quantum number $l = 0$) is used as an example (cf. Fig. 1). Afterwards a light pulse from a laser system is applied. If the wavelength corresponds to the energy of the transition [in the case of $p\mu$, the transition from $2s$ to $2p$ ($n = 2, l = 1$)], the laser pulse is long enough, and the laser power is sufficient, a transition to the final state is induced with high probability. In order to detect that the

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

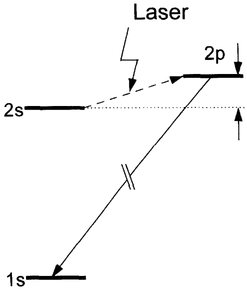

FIG. 1. Scheme of the lowest levels in the $p\mu$ atom.

resonance condition has been met and hence a transition has occurred, the daughter state has to be short-lived and must be depopulated by radiation that can be easily detected. In our example the $p\mu(2p)$ state has a lifetime of only a few picoseconds and is depopulated to the $p\mu(1s)$ state by quantum radiation of about 1.9 keV energy, not too hard to detect with low-energy semiconductor detectors. If the resonance condition is fulfilled a detector signal is registered in coincidence with the laser pulse. Relative errors for the transition energies as low as a few ppm (parts per million) may be achieved.

1.1.1 Muonic Hydrogen. In the case of the lowest levels of muonic hydrogen the uncertainty of all QED corrections turned out to be at the few-ppm level. QED, quantum electrodynamics, is the quantum theory of the electromagnetic interaction, the most important of the four interactions in the universe known up to now: strong, weak, and electromagnetic interactions, and gravity. The interest in muonic hydrogen is easily explained: In electronic hydrogen the accuracy of the $2s - 2p$ energy difference, a very stringent test of QED, is limited at the moment not by a lack of experimental accuracy but by the uncertainty of the proton radius $r_p = 0.862 \pm 0.012\,\mathrm{fm}$, which strongly influences the theoretical interpretation of the experimental results (Pachucki, 1996). In the muonic counterpart of the hydrogen atom the uncertainty in r_p plays a much larger role—the muon orbits much closer to the proton than the electron in the electronic hydrogen atom and therefore probes its dimension much more thoroughly. A precision determination of the $2s - 2p$ *difference* in muonic protium allows a much better determination of r_p and hence a test of QED on the normal hydrogen atom with an orders of magnitude better accuracy.

A direct test on the QED corrections in *muonic* atoms aiming at a precision much better than the present 10^{-3} has been proposed by an ENEA–Neuchatel–Paul Scherrer Institute–Trieste collaboration (Chatellard *et al.*, 1993). In this case the energy difference $3p - 3d$ in muonic hydrogen is to be determined with very high precision; this will be achieved by stimulating the $3p \rightarrow 3d$ transition during the muon cascade in μp with a laser and thus—in case the laser is on resonance—changing the intensity pattern of the x-ray transitions in the Lyman series compared to nonresonance. As no s level is involved, the influence of the proton radius is drastically reduced. A comparison with the results from electronic hydrogen allows one to draw conclusions about the μ–e universality.

1.1.2 Metastable Antiprotonic Helium. Metastable antiprotonic He, $^{3,4}\mathrm{He}^{++}\bar{p}e^-$, is a very special exotic atom, half-way between atom and molecule. Some years ago it was shown (Iwasaki *et al.*, 1991) that about 3% of all $\bar{p}$He atoms formed exist in such metastable states (lifetime several microseconds), in total contrast to conventional wisdom, which predicts the cascade in this exotic atom from the states populated immediately after atomic capture to the states from which annihilation occurs to be completed within picoseconds to nanoseconds, even in gaseous targets at moderate pressures. The metastability of the $\bar{p}$He atoms (their lifetime in liquid helium is more than *five* orders of magnitude larger than expected) may be explained by the formation of circular $(l = n - 1, v = n - l - 1 = 0)$ or nearly circular states $(v \leq 3)$ of the $\mathrm{He}^{++}\bar{p}e^-$ system (cf. Fig. 2) which can be depopulated neither by Auger effect with moderate angular-momentum transfer nor via Stark mixing with lower-l states, from which fast deexcitation may occur. Only radiative transitions are allowed, which are slow. Irradiation of the system with laser light of suitable wavelength, however, leads to stimulated deexcitation of some metastable states to short-lived ones, which is followed by the usual fast cascade of the $\bar{p}$ to states from which annihilation occurs. In this case the signal for a resonant transition is the prompt emission of charged and neutral pions following the annihilation of the $\bar{p}$ (Ya-

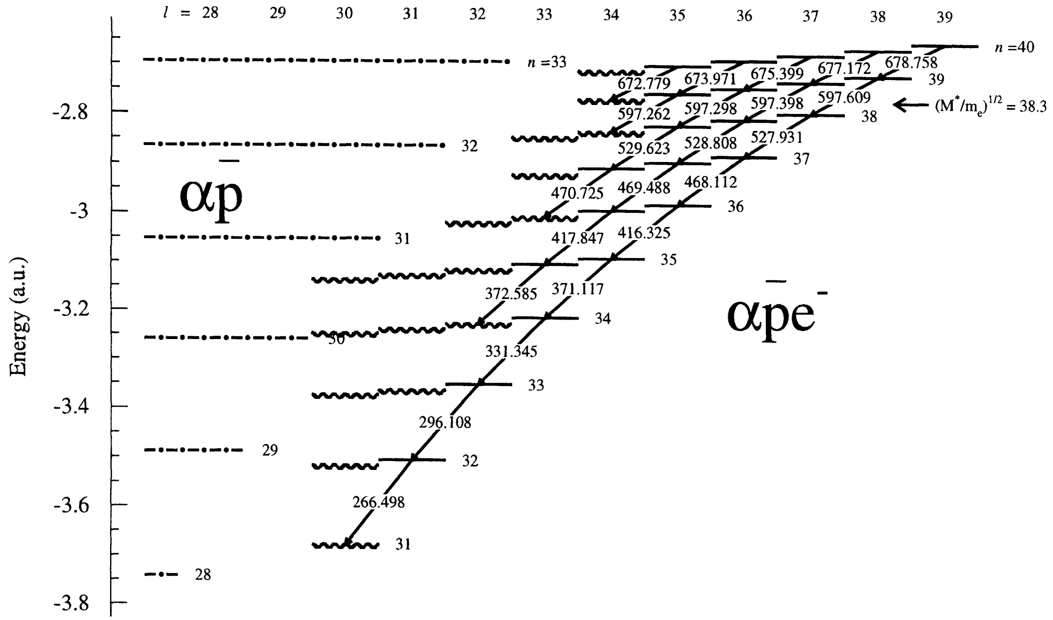
l = 28 29 30 31 32 33 34 35 36 37 38 39
Energy (a.u.)
-2.8
-3
-3.2
-3.4
-3.6
-3.8
n =33
32
31
30
29
28
αp̄
αp̄e⁻
n =40
39
38
37
36
35
34
33
32
31
(M*/mₑ)^{1/2} = 38.3
672.779
673.971
675.399
677.172
678.758
597.262
597.298
597.398
597.609
529.623
528.808
527.931
470.725
469.488
468.112
417.847
416.325
372.585
371.117
331.345
296.108
266.498

FIG. 2. Scheme of highly excited states in $\bar{p}^4$He. Full lines: metastable states; wavy lines: short-lived states; dot-dashed lines: states in the $\alpha\bar{p}$ system. The numbers at the arrows indicating the transitions are transition wavelengths in nanometers.

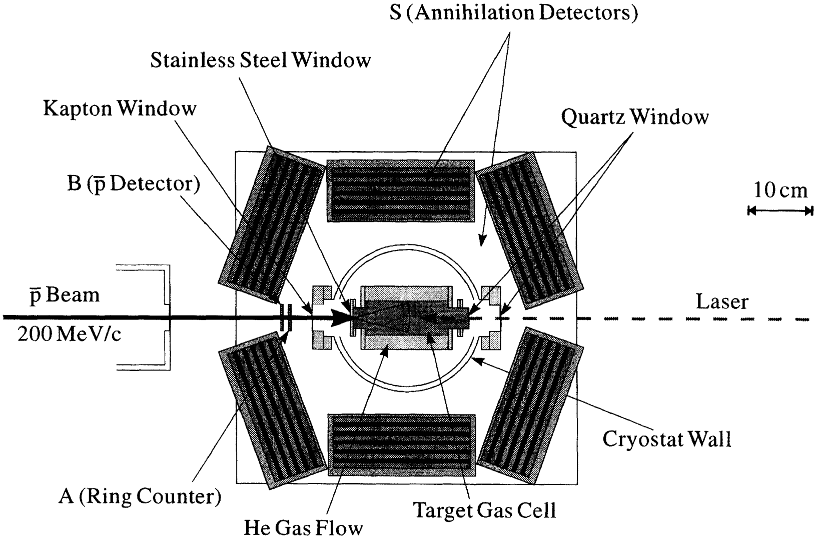
S (Annihilation Detectors)
Stainless Steel Window
Kapton Window
Quartz Window
B (p̄ Detector)
10 cm
p̄ Beam
200 MeV/c
Laser
Cryostat Wall
A (Ring Counter)
Target Gas Cell
He Gas Flow

FIG. 3. Experimental arrangement for the laser spectroscopy of metastable antiprotonic He. The target cell is kept at temperatures between 5 and 80 K; the helium pressure was kept in the 1–8-bar range. A, B: $\bar{p}$ beam counters. One out of the seven annihilation detectors is below the target.

Table 1. Experimental and calculated wavelengths for transitions in antiprotonic He.

Transition $n_i, \ell_i \Rightarrow n_f, \ell_f$	v	Wavelength(nm) Experiment	 Theory
$37,34 \Rightarrow 36,33$	2	470.724 (2)	470.725
$38,35 \Rightarrow 37,34$	2	529.622 (3)	529.623
$39,35 \Rightarrow 38,34$	3	597.259 (2)	597.262
$38,37 \Rightarrow 39,38$	0	597.607 (2)	597.609
$37,36 \Rightarrow 38,37$	0	527.930 (2)	527.931
$38,36 \Rightarrow 39,37$	1	597.398 (2)	597.398
$37,35 \Rightarrow 38,36$	1	528.81 (3)	528.808
$38,35 \Rightarrow 39,36$	2	597.298 (2)	597.298
$38,37 \Rightarrow 39,38$	0	597.607 (2)	597.609

mazaki *et al.*, 1993). A typical setup for such an experiment is shown in Fig. 3.

The laser spectroscopy of highly excited, metastable states of the $\mathrm{He}^{++}\bar{p}\mathrm{e}^-$ system—besides offering other benefits to be described later—opened the way to investigate the binding energies in this simple three-body system with a He nucleus, an antiproton, and an electron. As already pointed out, this special system may be viewed either as an antiprotonic atom with the nuclear charge partly shielded by the electron or as a special electronic molecule with one partner having negative charge. Such a system is perfectly suited to test the existing techniques for the calculation of atomic and molecular wave functions and energies. The most recent calculations are in agreement with the outcome of the experiment at a 5-ppm level. Table 1 shows the transitions detected by laser spectroscopy and their experimental and calculated wavelengths (Morita *et al.*, 1994; Hayano *et al.*, 1994; Maas *et al.*, 1995; Ketzer *et al.*, 1997).

1.2 Hadronic Atoms: the Strong Interaction

1.2.1 Pionic Atoms. The successful development of new crystal-spectrometer techniques [cylindrically bent quartz (110) crystals, position-sensitive CCD detectors] has stimulated research on the strong interaction of pions in low-lying states of pionic hydrogen (protium or deuterium). A measurement of shifts and widths of transitions to the 1*s* state gives direct access to the fundamental pion–nucleon scattering length, a quantity governing the pion–nucleus strong-interaction potential (cf. I, Sec. 4). An improvement of the energy resolution of these spectrometers by a factor of six to about 0.7 eV at 2.8 keV and an increase in the beam current in the modern meson factories (e.g., at Paul Scherrer Institute, Villigen, Switzerland) made it possible to determine the strong-interaction shifts and widths of the $3p \rightarrow 1s$ transitions in π^-p (Sigg *et al.*, 1995) and π^-d (Chatellard *et al.*, 1995) with hitherto unknown accuracy. The values are listed in Table 2, together with deduced experimental and theoretical (Bernard *et al.*, 1995) scattering lengths. The agreement is good.

Successes have been reported also from the very heavy pionic atoms. The lowest levels in these atoms are inaccessible to research via formation of and cascade in the pionic atom, because the pion is absorbed by the nucleus long before it reaches these levels. Knowledge about the deeply bound atomic states (cf. Fig. 4), however, is extremely interesting because there the pion is practically inside the nucleus and probes the nuclear matter density as well as changes in the strength of the pion–nucleon interaction caused by the nuclear medium. It has been argued that the strong-interaction potential in heavy nuclei should become repulsive (*i.e.*, decrease binding) on account of the action of the nuclear medium. This would result not only in a reduction of the binding energy of the pion but also in a smaller width ΔE of the lowest π-atomic levels: the pion is pushed outward by the repulsive force, so that its interaction with the nucleus decreases and with it the width of the level.

Table 2. Shifts and widths of the $3p \rightarrow 1s$ transition in pionic protium and deuterium, together with deduced experimental and recently calculated π-nucleon s-wave scattering lengths. Experiment and theory agree quite well now.

Pionic atom	Energy shift ε_{1s} (eV)	Linewidth Γ_{1s} (eV)	Scattering length a (fm)
$p\pi$	-7.13 ± 0.05	0.97 ± 0.11	$(0.086 \pm 0.002)m_\pi^{-1}$ (Experiment)
$d\pi$	2.48 ± 0.07	1.02 ± 0.21	$0.088 \leq am_\pi \leq 0.096$ (Theory)

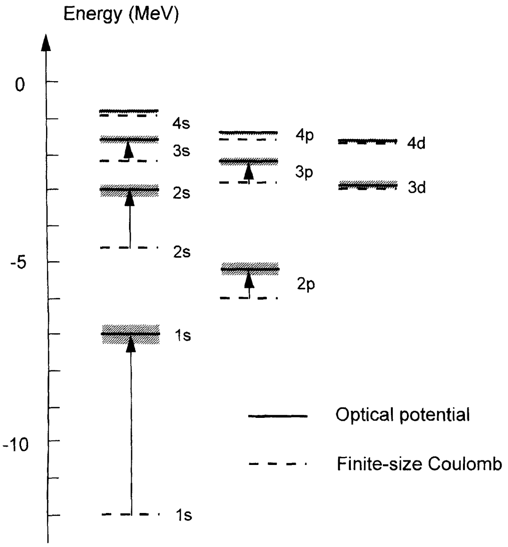

FIG. 4. Scheme of the lowest bound levels in pionic ^{207}Pb. Full lines: level energies calculated taking electromagnetic and strong interactions into account; dashed lines: level energies calculated taking only the electromagnetic interaction into account.

For the first time ever an experiment succeeded in populating directly the $2p$ level in $\pi^{-}\,{}^{207}\mathrm{Pb}$ via the high-energy reaction $^{208}\mathrm{Pb}(d, {}^{3}\mathrm{He})\pi^{-}\,{}^{207}\mathrm{Pb}$ (Yamazaki *et al.* 1996). The pion is created by the reaction of the deuteron with the lead nucleus and at the same time a neutron is picked up from this nucleus to make the momentum and energy balance favorable for the pion to stay bound. From the experiment the binding energy of the pion in the $2p$ level was extracted to be only $5.4 \pm 0.2\,\mathrm{MeV}$, much smaller than expected from purely electromagnetic binding (cf. Fig. 4). This experimental finding corroborated the suspicion that the strong low-energy pion–nucleus interaction indeed becomes repulsive towards the center of the nucleus in heavy nuclei. The small value of the level width (estimated to be $\Delta E < 0.8\,\mathrm{MeV}$) results from the pion being driven out of the center of the nucleus, markedly weakening its interaction with the nucleons. More experiments are planned for the coming years, which aim at detecting other low-lying levels, like $2s$ and even $1s$.

1.2.2 Antiprotonic Atoms. The experimental basis for levels and shifts in antiprotonic atoms was only slightly enlarged and no new data were taken for other exotic atoms with particles heavier than antiprotons. A new attempt was made (Batty *et al.*, 1995) to reproduce a representative subset of the $\bar{p}$-atomic data (for elements between carbon and praseodymium) by the full nonlocal potential used for pionic atoms (cf. I, Sec. 4) and a *density-dependent* scattering length:

$$(\mu/2\pi)\,V_{Strong}(\boldsymbol{r}) = (1+\mu/M)\left\{\left[b_0+B_0\left(\frac{\rho(r)}{\rho_0}\right)^{\gamma}\right]\rho(\boldsymbol{r})+b_1\delta\rho(\boldsymbol{r})\right\} - \boldsymbol{\nabla}\alpha(\boldsymbol{r})/[1-(4/3)\pi\zeta\alpha(\boldsymbol{r})]\boldsymbol{\nabla}, \qquad (1)$$

where $\delta\rho(\boldsymbol{r}) = \rho_{\mathrm{n}}(\boldsymbol{r}) - \rho_{\mathrm{p}}(\boldsymbol{r})$. Such a density-dependent parametrization of the scattering length was triggered by the observation that the $\bar{p}$–nucleus potential is repulsive for the free-antiproton–nucleon interaction whereas it becomes strongly attractive inside a nucleus even at the low nucleon densities that prevail at its periphery.

Antiprotonic atoms were also recently applied to probe espeically the difference between neutron and proton densities at the nuclear periphery. To this aim antiprotons were shot into isotopically pure samples. They form antiprotonic atoms and are annihilated. As the annihilation is very peripheral, it may happen that all annihilation products (mostly pions) miss the nucleus. In this case the nucleus with mass number $A-1$ that is produced after annihilation has either $N-1$ neutrons (annihilation of the $\bar{p}$ on a neutron) or $Z-1$ protons (annihilation on a proton). Here Z and N are the atomic number and the neutron number, respectively, of the original nucleus. If the nuclei generated in this way are radioactive they may be detected with high sensitivity via their decay. If the neutron distribution extends to larger radii than the proton distribution more annihilations take place on neutrons than if both nucleon distributions are the same. Such an effect has been detected for a number of nuclei (Lubiński *et al.*, 1994). Figure 5 shows the correlation of this enhancement with the binding energy of the most loosely bound neutron. Apparently a smaller neutron binding energy means a neutron localized more at the

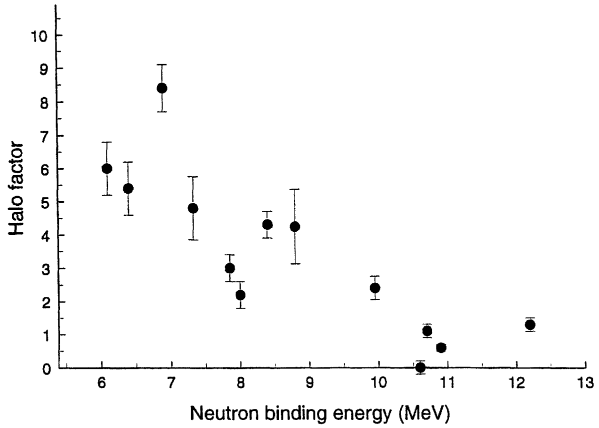

FIG. 5. Halo factor as a function of the binding energy. The halo factor is defined by

$$h = \frac{n(N_T - 1)}{n(Z_T - 1)} \frac{\mathrm{Im}(a_p)}{\mathrm{Im}(a_n)} \frac{Z_T}{N_T}.$$

Here N_T and Z_T are the numbers of neutrons and protons, respectively, in the target nucleus. $n(N_T - 1)$ denotes the number of nuclei generated with one neutron missing, $n(Z_T - 1)$ the number of nuclei generated with one proton missing. Finally, $\mathrm{Im}(a_n)$ and $\mathrm{Im}(a_p)$ are the scattering lengths for annihilation on a neutron and a proton, respectively.

periphery of the nucleus and hence a larger probability for peripheral annihilation on the *neutron.*

2. FORMATION OF EXOTIC ATOMS

Three fundamental questions may be raised about the formation of exotic atoms:

1. What is the kinetic energy the exotic particle has before it enters the atom it is finally captured by?
2. What is the energy the exotic atom gains immediately after capture?
3. Which are the bound states in the exotic atom populated immediately after Coulomb capture? The phrase "Coulomb capture" indicates that only Coulomb-like forces are acting during atomic capture.

2.1 Kinetic Energy of the Exotic Particle Before Capture

The simplest exotic atom one could envisage is the muonic hydrogen atom: only nucleus and muon. It is formed in the following way: The muon approaches a hydrogen atom; its charge screens the nuclear charge until finally the electron becomes free. Proton and muon start orbiting around their common center of gravity if the muon kinetic energy is low enough. If the energy of the emitted electron energy stays low—and there is good reason to assume this, as the momentum transfer of the comparatively heavy muon to the light electron should be small—the momentum of the muon (mass m_μ) before capture and of the $p\mu$ atom generated (mass $m_{p\mu}$) are roughly the same. Then the $p\mu$ kinetic energy is given by $T_{p\mu} = T_\mu(m_\mu/m_{p\mu})$, with T_μ the kinetic energy of the muon before capture. The kinetic energies of the exotic particle before and the exotic atom after capture are hence strongly correlated for this especially simple case.

The situation is somewhat complicated by the molecular structure of hydrogen and the measurement is made more difficult by the cascade processes taking place after capture (cf. Sec. 3.2). The *density-dependent* collision processes (Coulomb deexcitation, Auger effect and elastic collisions, cf. I, Sec. 3) may change

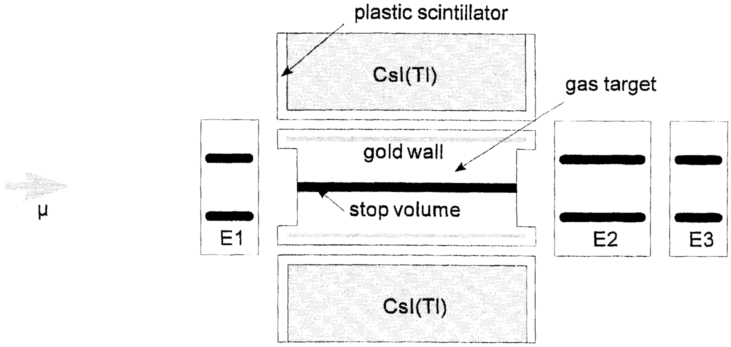

FIG. 6. Setup for the determination of the velocity of muonic hydrogen atoms after formation and cascade. The negative muons coming from the left are forced to travel to and fro between the electrodes E1, E2, E3 which are switched to high voltage at the right moment. The particles are finally stopped in the "stop volume". The $p\mu$ atoms formed travel to the walls of the cylindrical target; there they transfer the muon to the atoms of the "gold wall". Muonic X-rays from the gold walls are detected by the "CsI(Tl)" scintillation counter. Electrons from muon decay which also hit the CsI(Tl) counter are vetoed by the "plastic scintillator".

the kinetic energy of the $p\mu$ system. By working at the lowest pressures (below 0.1 hPa) they may be made small enough not to disturb the measurement of $T_{p\mu}$.

2.2 Kinetic Energy of the Exotic Atom After Capture

A series of experiments has been performed to study specifically the diffusion of the $p\mu$ atom after capture without any side effect due to collisions with hydrogen. To this end the pressure was lowered down to 0.062 hPa (which corresponds to a fraction $\varphi = 8.1 \times 10^{-5}$ of the atomic density of liquid hydrogen, 4.26×10^{22} atoms/cm^3); at these pressures collisional processes do not play a role for flight paths in the centimeter region: if we assume the reaction cross-section to be the geometrical cross-section $\sigma = \pi a_\mu^2 n^2$ the probability of a collision for a 1-cm flight path is below 0.001, even for $n = 14$. The velocity was determined via a time-of-flight measurement of the $p\mu$ system (Hartmann *et al.*, 1996) with a setup shown in Fig. 6. At this low pressure the $p\mu$ energy was found to be significantly smaller than 1 eV. From this the energy of the muon *before* capture can be estimated to be below 10 eV, in accordance with predictions from semiclassical calculations (Cohen, 1983).

2.3 The First-Populated States in the Exotic Atom

New knowledge about the characteristics of the first bound states in the simplest atoms came as a highly welcome by-product of the laser experiment on metastable antiprotonic He atoms, described in Sec. 1.1.2. By comparing experimental and calculated transition energies in these atoms more than a dozen transitions in $\bar{p}$-^{4}He and $\bar{p}$-^{3}He were unambiguously identified to proceed between levels with $n \approx 38$. Even the populations of some levels were deduced. The $\bar{p}$ really is captured into orbits around $n = 38$. The finding is the very first experimental proof for the long-standing assumption that the exotic particle is preferentially captured into those states for which the dimensions of the orbit of the ejected electron and that of the exotic particle after capture are roughly the same.

Experiment and latest calculations (Beck *et al.*, 1993) agree in their predictions about the circular or nearly circular character of the formed metastable states. Disagreement, however, exists about the percentage of the metastable atoms formed: whereas the experiment found a fraction of only 3%, theory estimates this fraction to be as high as 30% with most of the $\bar{p}$ initially populating the levels with $n \approx 40$. From these more loosely bound levels, however, the antiproton is predicted to be immediately quenched by collisions with other He atoms (Beck *et al.*, 1993). Mechanisms for this quenching process have been elaborated; quantitative results are still missing. As, naturally, this process is a collisional process it may be efficiently suppressed by going over to low pressures or even by storing the atoms in a trap. In these cases the probability for forming metastable antiprotonic atoms may be drastically increased, making these atoms a perfect laboratory for the research on electromagnetic interaction (three-body Coulomb interaction $He^{++}\bar{p}e^{-}$), fundamental physics (magnetic moment of the $\bar{p}$), and atomic collision processes.

3. THE EXOTIC-ATOM CASCADE

Knowledge about the exotic-atom cascade has developed mostly in the field of the lightest species, exotic hydrogen and helium. In these atoms no electrons are present from formation on or very soon afterwards, and some special effects already described earlier are present that make e.g. muonic hydrogen interesting for muon-catalyzed fusion. Before turning to the cascade in muonic and pionic hydrogen we say a few words about the progress in the understanding of the cascade for exotic atoms with $Z > 2$.

3.1 The Exotic-Atom Cascade in Elements with $Z > 2$

Absolute intensities of x rays from pionic atoms with $Z > 2$ have been measured recently for elements between $Z = 3$ and $Z = 83$ (Shinohara *et al.*, 1994). From the data "initial populations" of the angular-momentum sublevels at $n = 14$ of the form $p(l) \propto (2l + 1)\exp(\alpha l)$ have been deduced. Please keep in mind that these distributions at $n = 14$ have little to do with those at formation time but are rather a snapshot of the cascade pretty well towards its end. Nevertheless a remarkable correlation of α with the maximum angular momentum $\boldsymbol{L}_{\max} = \boldsymbol{r}_{\max} \times \boldsymbol{p}$ brought in by the exotic particle was found. Here $\boldsymbol{r}_{\max}$ is half the next-neighbor-atom distance in the solid and $\boldsymbol{p}$ the pion momentum corresponding to a kinetic energy equal to the ionization potential of the respective element (this is a rough estimate for the energy at which the exotic particle is Coulomb-captured into the atom).

The large number of Auger transitions during the exotic-atom cascade leads to a peculiar phenomenon: The exotic atom may be completely stripped of electrons during the cascade if electron refilling from the outside is suppressed by a lack of collisions (cf. Sec. 2.1). Such an exotic atom is an ideal laboratory for research on the electromagnetic and, if present, strong interaction between nucleus and exotic particle: No electrons are around disturbing this interaction. That those naked exotic atoms can be really formed with gases at low pressure has been recently shown by measuring the antiprotonic x-ray cascade in noble gases at pressures around 20 hPa (Simons *et al.*, 1994). Only transitions between circular orbits were observed and all lines had, after correction for the detection efficiency of the measuring device (a Ge detector) the same intensity: Because lack of electrons allows only radiative transitions with maximum $|\Delta n|$ (see I, Sec. 3.1), circular orbits are populated with high probability, and the antiproton cascades down the ladder of these orbits by radiative transitions only.

3.2 Acceleration During the Cascade

Even more than in the velocity of the exotic atom at formation time, one is interested in the velocity at the moment when the particle has reached the ground state or some might-be long-lived excited state (like the $2s$ state; cf. Sec. 1). The reason is that the velocity determines the chance to prepare certain states of the $p\mu$ system in pure form [e.g., the singlet hyperfine state of the $p\mu(1s)$ level (in this state the spins of proton and muon are antiparallel)] or the speed of some process (e.g., the transfer of the muon from deuterium to tritium in the muon-catalyzed dt-fusion cycle). Calculations (Markushin, 1994b) and experiments (Kraiman *et al.*, 1989; Aschenauer *et al.*, 1995; Hartmann *et al.*, 1996; Abbott *et al.*, 1997) give a roughly consistent picture: At pressures between about 1 kPa and 78 MPa the *mean* kinetic energy of the $p\pi$ and $p\mu$ system after reaching its ground state is several electronvolts. This kinetic energy comes from acceleration effects during the cascade (preferentially Coulomb deexcitation; cf. I, Sec. 3.5). The velocity is thus increased compared to the energy immediately after atom formation (cf. Sec. 2.2). Such an energy is, e.g., high enough to allow transitions between the $p\mu(1s)$ singlet and triplet (spin of proton and muon parallel) states by collision to occur (the triplet state lies 182 meV higher in energy than the singlet state). Precision measurements of the rate of nuclear capture (a weak-interaction process) from the $p\mu$ singlet state are thus hindered by the existence of $p\mu$ atoms in the triplet state with much smaller nuclear-capture rate.

3.3 Transfer Processes

During the last few years the *excited-state* transfer from muonic hydrogen to atoms with higher Z has gained more and more interest. The cross section for such transfers is considerably larger than for ground-state transfer, as the radius of the muon orbit increases with the square of the principal quantum number n, $r_n = a_\mu n^2/Z$; hence transfer, which has to proceed via level crossings in the correlation diagram (cf. I, Sec. 3.4) in order to become really fast, is possible already for larger impact parameters. Furthermore the higher polarizability of the $p\mu$ atom in the excited state gives rise to an additional attractive potential. Transfer from the $p\mu$ atom to argon, as an example, was calculated to be more than 20 times faster from the 2s level than from the $1s$ level. Such a fast transfer, together with effects of the deceleration of the $p\mu$ atom, served to explain the unusual time spectra observed for transfer from hydrogen to oxygen in a 15-bar hydrogen-SO_2 mixture. Such time distributions are obtained by registering muonic x rays from the element to which transfer occurs (the cascade in this element is fast). In normal cases transfer proceeds at one constant rate, and the distribution of transfer times follows the simple exponential law known from radioactive decay. In the mixture mentioned above, however, the time distribution of transfers to oxygen had a double-exponential shape, whereas for transfer to sulfur it was single exponential with a rate roughly corresponding to the smaller of the rates valid for oxygen. This may be explained (Schneuwly, 1994; Werthmüller *et al.*, 1996) by the population of higher $p\mu$ states, from which (faster) transfer is possible to oxygen but not to sulfur (because of a lack of level crossings).

4. APPLICATION OF EXOTIC ATOMS

4.1 Application of Muonic X Rays in Diagnostics

An experiment was reported only recently which applied muonic x rays to the nondestructive analysis of vertebral body trabecular bone (Hosoi *et al.*, 1995). It showed that quantitative information may be obtained about the elemental composition of deep-lying well-localized regions inside the human body. Muons were stopped in the lumbar vertebral bones (vertebra L3–L4) from patients of different ages (16 to 72 years), who had died from various diseases. The results for the elemental composition (Ca and P) found from the intensity of muonic x rays and from atomic absorption analysis were found to be strongly correlated, showing the x-ray method to be applicable for nondestructive analysis of the human body. Multiwire proportional chambers were again proposed (cf. I, Sec. 6.2) to trace muon-decay electrons back to their creation point and to localize better the place where the muons stop, but an experimental verification of this method is still missing.

5. MUON-CATALYZED FUSION

5.1 Basic Knowledge

The progress in our knowledge about muon-catalyzed fusion has slowed down somewhat during the past three years. Muon-catalyzed *deuterium-tritium* fusion is still in the center of research (cf. I, Sec. 7.6). Calculations and measurements, even if they aim at clarifying the processes in other types of fusion, are undertaken with the goal of obtaining further insight into the muon-catalyzed *dt*-fusion cycle.

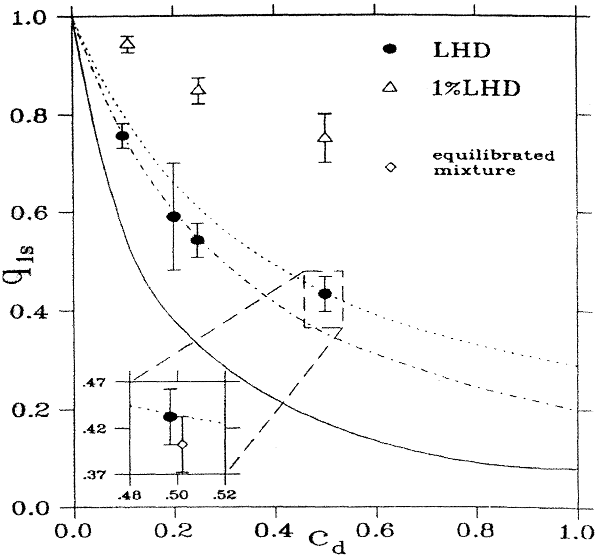

FIG. 7. Probability for excited-state transfer (Lauss *et al.*, 1996): q_{1s} for a protium-deuterium mixture as function of the deuterium concentration c_d at different target densities. LHD = liquid-hydrogen density. Equilibrated mixture: the concentration of HD molecules has been reached according to equilibrium in the reaction $H_2 + D_2 \rightleftharpoons 2HD$, which naturally depends on c_d. The curves show the results of calculations on q_{1s}. Full line: Markushin (1994b); dotted line: Czaplinski *et al.*, (1994); dash-dotted line: Markushin (1994a).

5.1.1 Atomic Capture. The latest results on Coulomb capture have already been presented.

5.1.2 Slowing Down. The complex interplay between acceleration and slowing down during the $p\mu$-atom cascade has already been described in Sec. 3.2. As soon as the muonic hydrogen atom reaches its ground state the only reactions capable of changing its velocity are elastic collisions. Here the structure of the collision partner (atomic or molecular) has a strong influence on the cross-sections. Especially at low $p\mu$ energies the molecular structure of hydrogen plays a more and more important role (Adamczak *et al.*, 1996). The cross section for interaction with a hydrogen molecule is not just twice the atomic cross section, as one might expect: Depending on the orientation, one of the atoms in the molecule may lie more or less in the shade of the other one. More important, the electron cloud between the covalently bound partners has a strong influence on the interaction.

5.1.3 Transfer. The implication of excited-state transfer has already been indicated in Sec. 3.3. In the case of hydrogen mixtures it tends to speed up the fusion cycle. It has become customary to characterize the excited-state transfer in hydrogen mixtures by the quantity q_{1s}, which simply means the fraction of muons captured in the lighter isotope (deuterium for instance) that reach the ground state without being transferred to the heavier isotope (e.g., tritium). As soon as the muons reach $\mu d(1s)$ they are trapped because the transfer from there is much slower than that from the excited states. At the lowest concentrations of the heavier isotope q_{1s} approaches unity as the number of collisions goes to zero. With increasing concentration the competition between cascade transitions and transfer tends more and more towards transfer, which makes q_{1s} decrease. An increase in density naturally has the same effect. Figure 7 shows results of a calculation for q_{1s} in HD mixtures together with experimental values taken only recently by just comparing the intensity of muonic x rays from $p\mu$ and $d\mu$ (they may be disentangled because their energy differs because of the difference in the muon reduced mass; cf. I, Sec. 1).

5.1.4 Molecular Formation. In the field of molecular formation, the main bottleneck in the dt-fusion cycle, attention has turned more and more to the possibility of epithermal resonant molecule formation. In this process the resonance condition for molecule formation [cf. I, Eq. (11)], playing an important role in dt fusion, is met for a large number of energies beyond thermal energy through which the tμ atom passes on its way down from the large kinetic energies attained after transfer. This is true for different excited states of the electronic (i.e., nonmuonic) hydrogen molecule that the *dt* molecule is part of and may considerably speed up molecular formation and hence the whole fusion cycle (Jeitler *et al.*, 1995).

5.1.5 Nuclear Fusion and Sticking. No new results are to be reported in the case of nuclear fusion rates. As this process is very fast in the *dt* cycle it is of restricted interest there, anyhow. Much more interesting is the fraction of muons finally sticking to the ^{4}He nucleus after fusion. Three different kinds of measurements have been performed:

1. Measurement of the number of neutrons emitted per muon. At high density, when the fusion cycle becomes faster, this number is limited by the loss due to sticking (Petitjean *et al.*, 1993).
2. Detection of muonic x rays from ^{4}Heμ, the fusion reaction product. The smaller this intensity the larger the number of cycles and the lower sticking (Bossy *et al.*, 1987).
3. Direct comparison of the He^{++} (no sticking) and Heμ^+ (sticking) ions generated. Such a measurement has been performed at low tritium concentrations (Petitjean *et al.*, 1993).

Combining these methods the sticking coefficient ω_s was evaluated to be $\omega_s = (0.51 \pm 0.05)\%$. Sticking really limits the possible number of catalyzed *dt*-fusion cycles to less than 200 (cf. I, Sec. 7.6). Schemes to revitalize the muon *after* sticking are highly desirable. Some proposals have been made, which, however, could not yet be tested.

5.2 Applications of Muon-Catalyzed Fusion

5.2.1 Generation of 14-MeV Neutrons. A design study has been finished for the μCF-based neutron source sketched earlier (cf. I, Sec. 7.8). Such a source is needed to test the neutron robustness of materials to be used in a *thermal-fusion* reactor (e.g. ITER). The study (Petitjean *et al.*, 1994) gave the following results: With a 12-mA beam of 1.5-GeV deuterons on a carbon target one may expect $10^{16}\,\pi^-$/s, which results in about $10^{15}\mu^-$/s in the DT mixture (volume 800 cm^3) and 0.5×10^{17} to 1×10^{17} 14-MeV neutrons per second. The test surface in this case would be 800 cm^2, resulting in roughly 10^{14}n/cm$^2 \cdot$ s. This is expected to be about the flux of neutrons hitting the walls of a future fusion reactor. The advantage of using the same fusion reaction as in the thermal fusion reactor would be to have the same neutron-energy spectrum in the test facility *and* in the future reactor. The technological difficulties for such a source, however, are large.

5.2.2 Generation of Low-Energy Muons with the Help of Muon-Catalyzed *dt* Fusion. The muon-catalyzed *dt*-fusion process was proposed to be employed to generate low-energy muons (i.e. muons with energies around and below 10 keV). In this process the muon is, in more than 99% of the case, only a spectator and is released with about 10 keV to catalyze another fusion (see above). In the proposed scheme a fraction of the MeV-muon beam is stopped in a thin layer of a solid protium-tritium mixture with small tritium content. The muons captured by protium are transferred to tritium, and the muonic tritium atoms (energy about 45 eV) travel into a second layer made of deuterium and tritium. Here fusion takes place and a fraction of the spectator muons leaves this layer with low energy. Those muons that are stopped in the layer catalyze another fusion and again have the chance to escape from the DT layer, making the conversion efficiency rather large. In an experiment at TRIUMF, Vancouver, the feasibility of the method was shown using protium-deuterium mixtures, which are much easier to handle than those with deuterium and tritium (Strasser *et al.*, 1996).

List of Works Cited

Abbott, D. J. *et al.* (1997), *Phys. Rev.* **A 55**, 214–229.

Adamczak, A., Faifman, M. P., Ponomarev, L. I., Korobov, V. I., Meleshik, M. S., Siegel, R. T., Wozniak, J. (1996), *At Data Nucl. Data Tables* **62**, 255–344.

Aschenauer, E. C., *et al.* (1995), *Phys. Rev.* **A 51**, 1965–1979.

Batty, C. J., Friedman, E., Gal, A. (1995), *Nucl. Phys.* **A 592**, 487–512.

Beck, W. A., Wilets, L., Alberg, M. A. (1993), *Phys. Rev.* **A 48**, 2779–2785.

Bernard, V., Kaiser, N., Meißner, U.-G. (1995), *Phys. Rev.* **C 52**, 2185–2187.

Bossy, H., *et al.* (1987), *Phys. Rev. Lett.* **59**, 2864–2867.

Chatellard, D. *et al.* (1993), in: L. A. Schaller, C. Petitjean (Eds.), *Muonic Atoms and Molecules*, Basel: Baltzer, p. 53.

Chatellard, D. *et al.* (1995), *Phys. Rev. Lett.* **74**, 4157–4160.

Cohen, J. (1983), *Phys Rev.* **A 27**, 167–179.

Czaplinski, W., Gula, A., Kravtsov, A., Mikhailov, A., Popov, N. (1994), *Phys. Rev.* **A 50**, 525–539.

Hartmann, F. J., *et al.* (1996), *Hyper. Int.* **101/102**, 623–632.

Hayano, R., *et al.* (1994), *Phys. Rev. Lett.* **73**, 1485–1488.

Hosoi, Y., Watanabe, Y., Sugita, R., Tanaka, Y., Nagamine, K., Ono, T., Sakamoto, K. (1995), *Brit. J. Radiol.* **68**, 1325–1331.

Iwasaki, M., *et al.* (1991), *Phys. Rev. Lett.* **67**, 1246–1249.

Jeitler, M., *et al.* (1995). *Phys. Rev.* **A 51**, 2881–2898.

Ketzer, B., *et al.* (1997), *Phys. Rev. Lett.* **78**, 1671–1674.

Kraiman, J. B., *et al.* (1989), *Phys. Rev. Lett.* **63**, 1942–1945.

Lauss, B., *et al.* (1996), *Hyper. Int.* **101/102**, 285–291.

Lubiński, P., *et al.* (1994), *Phys. Rev. Lett.* **73**, 3199–3202.

Maas, F., *et al.* (1995), *Phys. Rev.* **A 52**, 4266–4269.

Markushin, V. (1994a), *Phys. Rev.* **A 50**, 1137–1143.

Markushin, V. (1994b), in: *Abstracts of the International Workshop on Low Energy Muon Science, LEMS '93*, Los Alamos National Laboratory Report No. LA-12698-C.

Morita, N., *et al.* (1994), *Phys. Rev. Lett.* **72**, 1180–1183.

Pachucki, K. (1996), *Phys. Rev.* **A 53**, 2092–2100.

Petitjean, C., *et al.* (1993), *Hyper. Int.* **82**, 273–293.

Petitjean, C., *et al.* (1994), *Fusion Technol.* **25**, 437–450.

Schneuwly, H. (1994), *Phys. Lett.* **A 191**, 416–421.

Shinohara, A., Furukawa, M., Saito, T., Baba, H., Miura, T., Imanishi, N. (1994), *Nucl. Instr. Methods Phys. Res.* **B 84**, 14–22.

Sigg, D., *et al.* (1995), *Phys. Rev. Lett.* **75**, 3245–3248.

Simons, L., *et al.* (1994), *Nucl. Instrum. Methods Phys Res.* **B 87**, 293–300.

Strasser, P., Ishida, K., Sakamoto, S., Shimomura, K., Kawamura, N., Torikai, E., Iwasaki, M., Nagamine, K. (1996), *Hyper. Int.* **101/102**, 539–546.

Werthmüller, A., *et al.* (1996), *Hyper. Int.* **101/102**, 271–275.

Yamazaki, T., *et al.* (1993), *Nature* **361**, 238–240.

Yamazaki, T., *et al.* (1996), *Z. Phys.* **A 355**, 219–221.

NEUROBIOPHYSICS

(An Addendum to *Encyclopedia of Applied Physics*, Volume 11, pages 297–322.)

Harold Sonnier and Andrew A. Marino, *Department of Orthopaedic Surgery, Louisiana State University Medical Center, Shreveport, Louisiana, U.S.A.*

Introduction 401
1. **Structure and Function** 401
2. **Ion Channels** 401
3. **Neuronal Systems** 402
3.1 *Dendrites* 402
3.2 Neural Networks 402
3.3 Volume Transmission 402
3.4 Memory 403
Glossary 404
Works Cited 404
Further Reading 405

INTRODUCTION

The distinctions between neurobiophysics and other areas of neuroscience are becoming progressively more difficult to discern. Attempts to understand how the nervous system functions have been made from a variety of different scientific perspectives, and a consensus approach appears to be developing that involves some mathematical modeling and attempts to elucidate physical principles, but which is more focused on elucidating physical mechanisms.

1. STRUCTURE AND FUNCTION

Recent evidence suggests that gap junctions play an important role in regulation of the nervous system. Gap junctions are composed of transmembrane protein subunits called *connexins*, 4 to 6 of which form a pore that constitutes one half of a gap junction, the other half of which is formed by a similar structure on an adjacent cell. Heterology among the connexin subunits allows for differences in permeability, conductance, and gating of gap junctions by voltage or various ligands (Kumar and Gilula, 1996).

Flux through gap junctions in retinal cells was visualized in the form of Ca^{2+} waves induced by electrical, chemical, or mechanical stimuli that caused no change in cell-membrane potential (Newman and Zahs, 1997). The Ca^{2+} waves, which could propagate for more than 10 s and as far as 180 μm from the initiation site, may constitute a previously unrecognized form of neural communication.

A form of Charcot-Marie-Tooth disease (a hereditary peripheral neuropathy) has been attributed to the production of nonfunctional connexin subunits (Spray and Dermietzel, 1995), thereby underscoring the importance of electrotonic conduction between cells in the functioning of the nervous system.

2. ION CHANNELS

X-ray diffraction of most types of ion-channel proteins is presently not possible because of the difficulty in producing suitable crystals. *Site-directed mutagenesis* studies, however, have provided insight into structural features of ion channels such as the location of the ion-selective filter for voltage-gated Na^+, Ca^{2+}, and K^+ channels (Guy and Durell, 1995). The filter appears to lie in the channel's pore-lining segments (P segments). By changing specific amino acids in the P segments, the permeant-ion profile of Na^+ and Ca^{2+} channels can be interchanged. Even though the primary structures of K^+ channels are more heterogeneous, location of the selectivity filter has been narrowed down to a three-amino-acid motif in the P segment where mutations have invariably disrupted function (Navarro *et al.*, 1996).

Just as ion channels seem to share a common structural and functional domain for ion

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

selectivity, all known voltage-gated channels sense voltage changes with a protein subunit called the S4 segment. This segment has been identified as the voltage sensor of K^+ channels on the basis of measurements of the gating current (Mannuzzu *et al.*, 1996). A span of at least seven S4 amino acids was exposed extracellularly during activation of K^+ channels in a *Xenopus* oocyte model system. This suggests that during activation either the S4 amino acids moved extracellularly, or they remained static while the rest of the channel protein moved intracellularly.

The basis of the gating current in Na^+ channels was studied using hydrophilic cysteine-modifying reagents (Yang *et al.*, 1996). The procedure (systematic replacement of each S4 residue with cysteine followed by use of the membrane-impermeant reagents to modify the cysteines and detect any changes during activation) permitted a determination of precisely how many S4 amino acids and their accompanying charges moved from the intracellular to the extracellular compartment during channel activation, and the distance through which they moved. The total gating current was found to consist of the movement of about 10 positive charges through a distance of 4.5–11 Å. Since the cell membrane is normally 40 Å thick, the observation of (at most) only an 11-Å displacement between the extracellular and intracellular compartments suggests that the geometry of the cell membrane is radically altered in the vicinity of ion channels.

3. NEURONAL SYSTEMS

3.1 Dendrites

The role of dendrites was traditionally thought to be limited to that of supporting the passive unidirectional propagation of current, but new evidence suggests that dendrites can also integrate and process bidirectional information. Back-propagation of axon-derived action potentials (*antidromic action potentials*) was demonstrated in the hippocampus (Magee and Johnston, 1997) and neocortex (Markram *et al.*, 1997), and shown to be necessary for long-term potentiation. Subthreshold synaptic inputs that were followed within about 100 μs by the arrival of back-propagating action potentials facilitated amplification of subsequent subthreshold synaptic inputs. Timing differences above the window fundamentally altered the plasticity of the synapses (Sejnnowski, 1997). Thus, it appears that dendrites may have an important and previously unrecognized role in synaptic plasticity.

3.2 Neural Networks

Neural networks have been developed as useful interpretative tools. For example, they have been used to determine secondary structure of proteins (Rost *et al.*, 1995), putative coding sites in DNA, and chromosome classification (Leon *et al.*, 1996). Networks can even outperform human beings in some diagnostic judgment tasks, including diagnosis of acute myocardial infarction under the high stress and time constraints of the emergency room (Baxt and Skora, 1996) and interpretation of breast biopsies (Baker *et al.*, 1996). However, the original promise of neural networks, that they could accurately model the actual mechanism of biological learning, has not been fulfilled. It is becoming increasingly more clear that information processing within the central nervous system is more complicated than that embodied in the neural-network approach.

3.3 Volume Transmission

Movement of information in the central nervous system can occur via nonsynaptic mechanisms (volume transmission). One type of such information flow occurs when potentials produced by groups of cells propagate in the electrically conductive extracellular fluid and induce effects in distant cells—for example, synchronizing firing or altering firing probability (Bullock, 1997). The physiological significance of signalling by means of volume electrical conduction is still problematical, but diffusion of neuroactive substances through brain tissue from their point of secretion to distant binding sites is increasingly being recognized as an important mechanism for regulating cell activity.

Neuropeptides (a subset of neurotransmitters) are usually cosecreted with a classic neurotransmitter (biogenic amines, amino acids, and acetylcholine), but at lower levels. Neuropeptides are good candidates for volume transmitters because they are present in the extracellular fluid for relatively long periods and have relatively high affinity for their receptors. For example, when mossy fibers of the cerebellum were stimulated, dynorphin (a neuropeptide that was cosecreted with glutamate) inhibited the effects of glutamate in unstimulated mossy fiber tracts as a result of diffusion through the extracellular fluid (Weisskopf *et al.*, 1993). Comparable results were found in other studies (Chen *et al.*, 1995; Kolaj *et al.*, 1995).

Nitric oxide (NO) is also a possible signaling agent because it is produced in both neurons and glia and can quickly diffuse through the extracellular fluid. NO depressed voltage- and Ca^{2+}-dependent currents in hippocampal slices from Sprague-Dawley rats (Erdemli and Krnjevic, 1995), again suggesting a mechanism whereby distant groups of cells could interact with one another.

In addition to isotropic diffusion, evidence has been found for the occurrence of preferential extracellular diffusion along fiber bundles within the brain. Channel-like pathways that paralleled myelinated fiber tracts in rats were demonstrated using fluorescent-labeled dextrans and confocal microscopy (Bjelke *et al.*, 1995). The label reached the contralateral side of the neostriatum, indicating that long-range routes were possible.

Classically, it has been taught that neurotransmitters are released through calcium-dependent fusion of neurotransmitter-laden vesicles at neuronal presynaptic membranes (highly specific regions that form one of the boundaries of the synaptic cleft). Atwell *et al.* (1993) showed that *uptake carriers* can reverse direction and actually secrete neurotransmitter (Fig. 1). This secretion is nonvesicular, usually occurs away from the synaptic cleft, does not require Ca^{2+}, and can be demonstrated in glial cells as well as neurons. Thus, the phenomenon of reverse uptake provides a possible source of diffusable neuroactive agents.

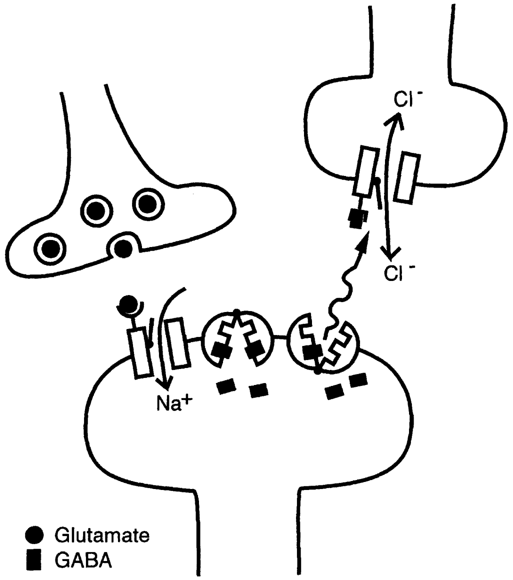

FIG. 1. Reverse uptake and volume diffusion of GABA (Atwell *et al.*, 1993). Glutamate is released by a presynaptic neuron and stimulates its receptors on either glial or neuronal cells, resulting in an Na^+ influx. The increased intracellular Na+ causes GABA uptake carriers (which normally couple the influx of GABA and Na^+) on the postsynaptic cell to operate in reverse and extrude GABA from intracellular stores. The extracellular GABA can then diffuse and affect other cells having appropriate receptors.

3.4 Memory

Memory must somehow be mediated by a relatively permanent physical change in the nervous system, but the responsible mechanisms have not been delineated. One possible explanation involves the irreversible activation of a calcium-calmodulin-dependent kinase (CaMKII), which is abundant beneath presynaptic and postsynaptic membranes. After Ca^{2+} activation, CaMKII can phosphorylate and regenerate itself independent of Ca^{2+} (Fig. 2). Mathematical models have predicted that spontaneous reversion to the inactive state would be highly unlikely, and the active Ca^{2+}-independent state could potentially last a lifetime (Lisman and Goldrign, 1988). The kinase could, therefore, mediate long-term changes in synaptic function by phosphorylating another protein, for example synapsin I (Nayak *et al.*, 1996).

The role of the kinase in mediating memory was demonstrated in animal studies. Mice were genetically engineered to express a permanently activated form of CaMKII (CaMKIIα) whose expression was regulated by a promoter in the forebrain and by a transactivator (tetracycline) (Mayford *et al.*, 1996). Thus, expression of the transgene (the foreign-DNA gene) was con-

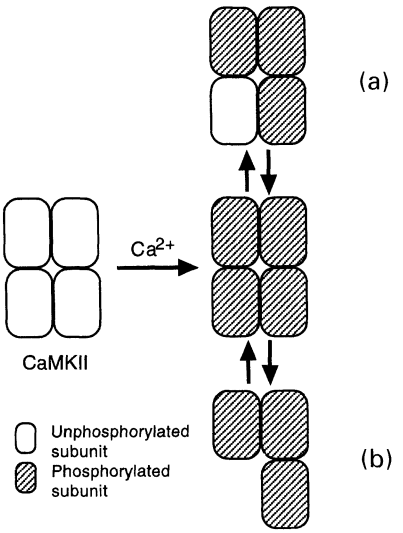

FIG. 2. Irreversible activation of CaMKII (Hanson and Schulman, 1992). In the inactivated state. CaMKII's subunits are unphosphorylated. Stimuli that induce memory activate CaMKII by Ca^{2+}-dependent phosphorylation of the subunits, after which autophosphorylation occurs independent of Ca^{2+}. Thus, if (a) a phosphatase dephosphorylates a subunit, Ca^{2+}-independent activity rephosphorylates it and returns CaMKII to full activity. Furthermore, when (b) a subunit becomes unbound or degraded by normal protein removal processes, the subunit is replaced and phosphorylated independent of Ca^{2+}, resulting in fully activated CaMKII.

fined to the forebrain (the only location of promoter) and was turned off only when tetracycline (T) was added. Tetracycline regulation of the transgene allowed some important observations. The transgenic mice learned and retained memory the same as normal mice when the transgene was inhibited (+T). Expression of the transgene (−T) impaired memory. The impaired memory was reversible by inhibiting the transgene (+T), and memories acquired while the transgene was inhibited could be extinguished by expressing CaMKIIα (−T).

Other studies in which CaMKII was inactivated also showed impairment of various aspects of memory (Glazewski *et al.*, 1996). Thus, inhibiting *or* activating CaMKII results in significant changes in memory storage, indicating that the kinase is at least partially responsible for information storage in the central nervous system.

GLOSSARY

Antidromic Action Potential: Movement of an action potential in a direction opposite to that of normal propagation. Backpropagation.

Connexin: Protein subunit of a gap junction.

Long-Term Potentiation: A relatively permanent change in a neuronal system.

Site-Directed Mutagenesis: A form of genetic engineering that permits changes in specific nucleic acids in a given gene.

Transgene: A foreign-DNA gene incorporated into an organism's genome early in development, and inherited by offspring in an Mendelian fashion.

Uptake Carrier: Membrane protein that facilitates uptake or release of ions and other substances.

List of Works Cited

Atwell, D., Barbour, B., Szatkowski, M. (1993), "Nonvesicular release of neurotransmitters," *Neuron* **11**, 401–407.

Baker, J. A., Kornguth, P. J., Lo, J. Y., Floyd, C. E. (1996), "Artificial neural network: Improving the quality of breast biopsy recommendations," *Radiology* **198**, 131–135.

Baxt, W. G., Skora, J. (1996), "Prospective validation of artificial neural network trained to identify acute myocardial infarction," *Lancet* **347**, 12–15.

Bjelke, B., England, R., Nicholson, C., Rice, M., Lindberg, J., Zoli, M., Agnati, L., Fuxe, K. (1995), "Long distance pathways of diffusion for dextran along fibre bundles in the brain. Relevance for volume transmission," *Neuroreport* **6**, 1005–1009.

Bullock, T. H. (1997), "Signals and signs in the nervous system: The dynamic anatomy of electrical activity is probably information rich," *Proc. Natl. Acad. Sci.* **94**, 1–6.

Chen, L., Gu, Y. P., Huang, L. Y. M. (1995), "The mechanism of action for the block of NMDA receptor channels by the opioid peptide dynorphin," *J. Neurosci.* **15**, 4602–4611.

Erdemli, G., Krnjevic, K. (1995), "Nitric oxide tonically depresses a voltage- and Ca-dependent outward current in hippocampal slices," *Neurosci. Lett.* **201**, 57–60.

Glazewski, S., Chen, C., Silvia, A., Fox, K. (1996), "Requirement for α-CaMKII in experience-dependent plasticity of the barrel cortex," *Science* **272**, 421–423.

Guy, H. R., Durell, S. R. (1995), "Structural models of Na^+, Ca^{2+}, and K^+ channels," *Soc. Gen. Physiol. Ser.* **50**, 1–16.

Hanson, P. I., Schulman, H. (1992), "Neuronal Ca^{2+}/calmodulin-dependent protein kinases," *Annu. Rev. Biochem.* **61**, 559–601.

Kolaj, M., Cerne, R., Randic, M. (1995), "The opioid dynorphin modulates AMPA and kainate responses in acutely isolated neurons from the dorsal horn," *Brain Res.* **671**, 227–244.

Kumar, N. M., Gilula, N. B. (1996), "The gap junction communication channel," *Cell* **84**, 381–388.

Leon, M. A., Gader, P., Keller, J. (1996), "Multiple neural network response variability as a predictor of neural network accuracy for chromosome recognition," *Biomed. Sci. Instrum.* **32**, 31–37.

Lisman, J. E., Goldrign, M. A. (1988), "Feasibility of long-term storage of graded information by the Ca^{2+}/calmodulin-dependent protein kinase molecules of the postsynaptic density," *Proc. Natl. Acad. Sci. USA* **85**, 5320–5324.

Magee. J. C., Johnston, D. (1997), "A synaptically controlled, associative signal for hebbian plasticity in hippocampal neurons," *Science* **275**, 209–213.

Mannuzzu, L. M., Moronne, M. M., Isacoff, E. Y. (1996), "Direct physical measure of conformational rearrangement underlying potassium channel gating," *Science* **271**, 213–216.

Markram, H., Lubke, J., Frotscher, M., Sakmann, B. (1997), "Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs," *Science* **275**, 213–215.

Mayford, M., Baranes, D., Posypanina, K., Kandel, E. R. (1996), "Control of memory formation through regulated expression of a CaMKII transgene," *Science* **274**, 1678–1683.

Navarro, B., Kennedy, M. E., Velimirovic, B., Bhat, D., Peterson, A. S., Clapham, D. E. (1996), "Nonselective and G betagamma-insensitive weaver K^+ channels," *Science* **272**, 1950–1953.

Nayak, A. S., Moore, C. I., Browning, M. D. (1996), "Ca^{2+}/calmodulin-dependent protein kinase II phosphorylation of the presynaptic protein synapsin I is persistently increased during long term potentiation," *Proc. Natl. Acad. Sci.* **93**, 15451–15456.

Newman, E. A., Zahs, K. R. (1997), "Calcium waves in retinal glial cells," *Science* **275**, 844–847.

Rost, B., Casadio, R., Fariselli, P., Sander, C. (1995), "Transmembrane helices predicted at 95% accuracy," *Protein Sci.* **4**, 521–533.

Sejnnowski, T. J. (1997), "The year of the dendrite," *Science* **275**, 178.

Spray, D. C., Dermietzel, R. (1995), "X-linked dominant Charcot-Marie Tooth disease of the nervous system," *Trends Neurosci.* **18**, 256–262.

Weisskopf, M. G., Zalutski, R. A., Nicoll, R. A. (1993), "The opioid peptide dynorphin mediates heterosynaptic depression of hippocampal mossy fibre synapses and modulates long term potentiation," *Nature* **362**, 423–437.

Yang, N., George, A. L., Horn, R. (1996), "Molecular basis of charge movement in voltage-gated sodium channels," *Neuron* **16**, 113–122.

Further Reading

Bruzzone, R., White, T. W., Goodenough, D. A. (1996), "The cellular internet: On-line with connexins," *BioEssays* **18**, 709–718.

Yuste, R., Tank, D. W. (1996), "Dendritic integration in mammalian neurons, a century after Cajal," *Neuron* **16**, 701–716.

Zoli, M., Agnati, L. F. (1996), "Wiring and volume transmission in the central nervous system: The concept of closed and open synapses," *Prog. Neurobiol.* **49**, 363–380.

24

Raman Scattering for Speciation and Analysis

Karen Esmonde-White, Mekhala Raghavan and Michael Morris

24.1 **Introduction** **835**
24.2 **Raman Spectroscopy Basics** **835**
24.2.1 Theory 835
24.2.2 Enhancement 836
24.2.3 Subsurface Measurements – Spatially Offset Raman Spectroscopy (SORS) 837
24.2.4 Instrumentation 837
24.2.4.1 Laser Safety 837
24.2.4.2 Basic Instrumentation 838
24.2.4.3 Fiber Optic Probes 839
24.2.4.4 Microprobes 839
24.2.5 Data Reduction 841
24.3 **Identification and Quantification of Inorganic Materials** **842**
24.3.1 Introduction 842
24.3.2 Minerals 842
24.3.3 Art, Archaeology, and Architecture 842
24.3.4 Forensics 844
24.3.5 Nanoscale Semiconductors 845
24.3.6 Industrial Process Control 848
24.4 **Identification and Quantification of Organic Materials** **848**
24.4.1 Pharmaceutical Applications 848
24.4.2 Polymer Chemistry and Physics 850
24.4.3 Forensic and Homeland Security Applications 850
24.5 **Biological Applications of Raman Spectroscopy** **851**
24.5.1 Subcellular Level 851
24.5.2 Tissue Level 852
24.5.2.1 Epithelial Tissue 852
24.5.2.2 Muscular Tissue 854
24.5.2.3 Nervous Tissue 854
24.5.2.4 Connective Tissue 855
24.5.3 Microorganisms 855

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

Glossary 858
References 859
Further Reading 863

24.1
Introduction

The basic principles and instrumentation of Raman spectroscopy are reviewed. Emphasis is placed on fiber-optic probes and microprobes. Applications to a wide range of problems in several industries, as well as to forensic applications are surveyed. Applications to the arts and sciences, including the life sciences, are also reviewed. Literature references provide the reader with links to more detailed discussion of the theory of Raman spectroscopy, to the methodology used in each of the surveyed areas, and to the details of the Raman spectra of many materials.

24.2
Raman Spectroscopy Basics

24.2.1
Theory

Raman spectroscopy is a widely employed tool in physical science, life science, and engineering. Consequently, there are numerous monographs and chapters in monographs that address the fundamentals and applications for different audiences. Good introductory surveys are available (Smith and Dent, 2005; Lewis and Edwards, 2001).

Raman scattering is a form of inelastic light scattering. In general, the lasers that are used to excite Raman scatter have photon energies lesser than the energy of the first electronic excited state. The weak interaction between the exciting laser photons and the molecule results in scattering, with light scattered at a new, lower frequency (expressed in wave numbers) according to Eq. (24.1) as follows:

$$\overline{\nu}_{\text{scattered}} = \overline{\nu}_{\text{exciting}} - \overline{\nu}_{\text{vibration}} \qquad (24.1)$$

Where $\overline{\nu} = 1/\lambda$. The scattered light is red-shifted and is called *Stokes-scattered*. It is assumed that all of the molecules are in the ground vibrational state. That is not quite true, especially for low frequency vibrations. If the scatter originates from a molecule in an excited vibrational state, then light is blue-shifted by the same frequency. Because this effect called *anti-Stokes Raman scattering* is weak, it is rarely used. Generally, most workers report Raman spectra as plots of intensity vs. $\overline{\nu}_{\text{vibration}}$ of scattered light.

The quantum efficiency for Raman scattering is low. Generally, it is between 10^{-7} and 10^{-9}. Although this number

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

sounds low, it should be noted that a 1-W laser operating at 800 nm outputs about 4×10^{18} photons per second. Raman spectroscopic instruments use lasers that provide 0.01–0.5 W at the sample. Therefore, even with 0.01-W laser power and 10^{-9} quantum efficiency, there should be about 10^7 Raman-scattered photons generated per second. Even with very low collection efficiency, the experimental problem is often not the absolute intensity of Raman scatter but rather the presence of an intense background signal from fluorescing molecules. Because fluorescence quantum efficiencies are often at least 0.01 and can approach 1, fluorescence intensity even from impurities in a matrix can exceed Raman scatter intensity by 2 or 3 orders of magnitude.

The wavelength dependence of the intensity of Raman scatter is approximated by Eq. (24.2).

$$\overline{\nu}^4_{\text{exciting}} = \frac{1}{\lambda^4_{\text{exciting}}} \tag{24.2}$$

If only the intensity of the Raman signal governed the choice of exciting wavelength, spectroscopists would choose as blue a wavelength as possible. In fact, the choice of wavelength is usually governed by the need to circumvent fluorescence background. If the material to be analyzed or its matrix is fluorescent, then red or near-infrared lasers are used. This is usually the case for most industrial organic compounds and most polymers and virtually all tissue specimens and human, animal, or bacterial cells. On the other hand, many inorganics, including ceramics and semiconductors, have little or no fluorescence. In this case, green lasers are usually used. Ultraviolet lasers are used only in special cases, both to avoid photochemical damage and to preserve compatibility with glass optics and sample containers.

The selection rules for Raman scatter are different from those for infrared absorption. A vibration is Raman active if the vibration causes a change in the dipole moment of the molecule induced in an external electric field. This quantity is proportional to a parameter called *polarizability*, *P*. By contrast, a vibration is infrared active if there is a change in the dipole moment of the molecule when the vibration is excited. These different selection rules mean that different types of vibrations are intense in the Raman and in the infrared. In many cases, a vibration that has an intense band in one mode of observation will have a weak band in the other. For example, stretches and bends of highly polar functional groups are intense in the infrared and weak in Raman scatter. On the other hand, stretches of unsaturated moieties are intense in Raman scatter and weak in infrared absorption. Significantly, the OH stretching and bending modes of water are strong in the infrared, but weak in Raman scatter. The stretching and bending modes of Si–O moieties are also intense in infrared absorption and weak in Raman scatter. Indeed, compatibility with aqueous solutions, glass, or fused silica containers and optics has long been major reasons for use of Raman spectroscopy rather than infrared spectroscopy.

24.2.2
Enhancement

There are two general enhancement mechanisms that result in great increases in Raman intensity, resonance enhancement, and surface enhancement. The corresponding methodologies are usually called *resonance Raman spectroscopy* (*RRS*) and surface-enhanced Raman spectroscopy (SERS) (Smith and Dent, 2005; Lewis and Edwards, 2001).

Resonance enhancement occurs when the exciting laser wavelength is coincident or nearly coincident with the electronic absorption band of a molecule. For colored molecules, resonance enhancement occurs at visible wavelengths. Examples of practical interest include hemoglobin, and dyes, especially those that are red or orange and absorb 532 nm light, the frequency-doubled wavelength of the very common Nd : YAG and Nd : VO_4 lasers. Of course, several molecules absorb in the ultraviolet. Resonance enhancement occurs at UV wavelengths; but for practical reasons, only a small number of specialist laboratories make use of this effect.

Surface enhancement occurs when the wavelength of the exciting laser is coincident with the wavelength of the surface plasmon of a metal on which the molecule is absorbed. For specially structured surfaces of silver or gold, the surface plasmon wavelengths are in the visible or near infrared. In theory, enhancement factors can exceed 10^{10}. In practice, enhancement factors of 10^3-10^5 are observed. Although surface enhancement can be observed with easily prepared gold or silver colloids, current practice has been moving toward structured surfaces prepared by photolithography because these provide more reproducible enhancement. Increasingly, gold is preferred over silver because gold surfaces are more stable than silver surfaces. The potential user is cautioned that the technique is limited to the monolayer of molecules that are actually in contact with the surface or, but with greatly reduced enhancement, to the monolayer and one or two layers directly above it. The dynamic range of SERS is, therefore, limited and it is largely a trace or an ultratrace analytical method.

24.2.3
Subsurface Measurements – Spatially Offset Raman Spectroscopy (SORS)

An important recent development has been the use of spatially separated laser illumination and collection regions to probe beneath the surface of a turbid medium, such as a polymer, a powdered material, or tablet or a human or animal tissue (Matousek, 2007). Generally, delivery of laser and collection of Raman scatter is by optical fibers, which may be arranged in any number of different geometries. The spatial separation of illumination and collection regions allows the collection of light that has been multiply scattered as it propagates through the turbid medium. Low definition images have been acquired with SORS-type systems (Schulmerich *et al.*, 2006; Schulmerich *et al.*, 2007) and even Raman tomography has been reported (Schulmerich *et al.*, 2008).

24.2.4
Instrumentation

24.2.4.1 Laser Safety

Standard laser safety protocols should always be enforced. In particular, users should wear appropriate laser safety glasses. Depending on the laser power, door interlocks and warning signs may be necessary. American laboratories follow the guidelines of the ANSI Z136.1–2007 standard (Laser Institute of America, 2007). Similar standards for laser safety apply in other countries.

24.2.4.2 Basic Instrumentation

A standard Raman spectroscopy system comprises a laser source, a sample illumination and collection system, a single stage spectrograph with a holographic or dielectric prefilter to reject laser light, and a charge-coupled device (CCD) imaging detector. Many types of devices are available to illuminate the sample and to collect the scattered light. The majority of users employ either fiber-optic probes or microscope-based microprobes. A common accessory is a motorized stage to facilitate spectroscopy of multiple samples, including samples in microliter plates and similar systems.

Frequency-doubled Nd : YAG and Nd : VO_4 lasers, which operate at 532 nm, are preferred for nonfluorescent or weakly fluorescent samples. They are compact, stable, and have no special electrical service requirements. For fluorescent materials, diode lasers operating at 785 nm, or for many biomedical applications, those operating at 830 nm are preferred. Where the excited fluorescence is low enough, 785 nm is preferred. It is the deepest red standard wavelength diode laser for which Raman shifts of C–H stretches (i.e., below about 3100 cm^{-1}) can be observed with a CCD or other silicon photodetector. The recent development of InGaAs-based imaging detectors has made Nd : YAG 1064 nm lasers a viable choice.

Several different spectrograph configurations are used. Czerny–Turner systems, usually with toroidal mirrors for astigmatism correction, are common. On-axis designs are frequently used for their superior imaging performance. The axial transmissive spectrograph (Battey *et al.*, 1993), which uses lenses instead of mirrors for collimation and focusing and a transmission grating instead of a reflection grating, is the most prevalent.

The detector in most Raman spectroscopy systems is a slow-scan (also called *scientific*) back-illuminated CCD (Janesick, 2001). The deep depletion version, which is fabricated from high resistivity silicon, has superior quantum efficiency above 800 nm, and is used in conjunction with 785 and 830 nm lasers. Replacement of conventional output amplifiers with a chain of avalanche breakdown stages yields an electron-multiplying charge-coupled device (EMCCD) (Coates *et al.*, 2003; Denvir and Conroy, 2003). The EMCCD has high gain (up to 1000 X), is capable of single photon detection and can be operated at high read-out rates (greater than 1 MHz) with very little excess noise. The electron-multiplying technology has not yet been successfully implemented on deep depletion CCD's and so has not been useful with deep red or NIR diode lasers.

As the size of solid state lasers and CCD detectors has decreased, so has the size of laboratory bench-top Raman spectroscopy systems. Modern systems generally have a footprint that is slightly higher than 0.5×0.5 m and many are smaller. Of course, the addition of a microscope or a stand for a fiber-optic probe or some other specialized accessory may increase the overall size of the instrument.

The increasing use of Raman spectroscopy for chemical process control has encouraged vendors to develop versions of their instruments for use in the factory environment. These instruments are usually packaged in enclosures for high dust, vibration, and electrical noise environments and may contain rugged versions of some components and subsystems.

Recently, several companies have introduced handheld, battery-powered instruments for off-site use in incoming materials inspection and for geological, environmental, homeland security and forensic applications. These instruments can operate for several hours without recharging. They typically are low-resolution (8–10 cm^{-1}) instruments and may have on-board databases and software for matching spectra to commonly encountered materials.

24.2.4.3 **Fiber Optic Probes**

An obvious advantage of the fiber optic probes is that the laser, spectrograph, and CCD can be some distance away from the sample. In the laboratory, 3–5 m is considered as the typical fiber length. Often, longer lengths are found in industrial process control to minimize the effects of noise and vibration on the instrumentation. Two general designs of fiber-optic Raman probes are in widespread use: the N-around-1 probe and the bifurcated filtered probe.

In the N-around-1 design laser, light is transmitted through a central optical fiber and Raman scatter is collected from a circular bundle of fibers that surrounds it. These are cemented into a metal, ceramic, or plastic holder and the ends are cut and polished. However, as the laser light propagates through the excitation fiber, it generates a silica Raman signal. This usually causes no problem in measurements of clear liquids. If the sample is a turbid fluid or a solid, some silica Raman scatter is reflected back into the collection fibers and becomes a background, which increases with the length of the fiber used. The background problem can be reduced if filters are placed at the ends of the optical fibers. An example is shown in Figure 24.1 (Motz *et al.*, 2004). Of course, the filters add cost and complexity to the probe.

In the bifurcated filtered probe, a single excitation fiber and a single collection fiber are used. A dichroic mirror, notch filter, or other wavelength-selective optic are used to combine the excitation and collection light. Delivery of laser light to the sample and collection of Raman-scattered light from the sample are through a single lens. Auxiliary filters may be used to remove silica Raman scatter from the excitation light and to prevent laser light from entering the collection fiber, resulting in very low silica background. In addition to the basic, general purpose designs, available probes include models with sampling optics designed for immersion in liquids or for operation at elevated or low temperatures.

24.2.4.4 **Microprobes**

Raman microprobes are used because they are applicable to opaque materials and water-containing materials (such as plant or animal tissue or cells) and offer lateral spatial resolution that is similar to resolution obtainable with light microscopes, about 0.5–1 µm. Operation in this wavelength range gives the Raman microscope better spatial resolution than is available with infrared microscopes. Additionally, most employ a confocal geometry to improve axial resolution and reduce background scatter and fluorescence. Depth profiling by Raman microscopy is well established (Reinecke *et al.*, 2001; Adar, 2001; Fleming and Kazarian, 2004). It should be cautioned that the spherical aberration induced by refractive index changes can distort

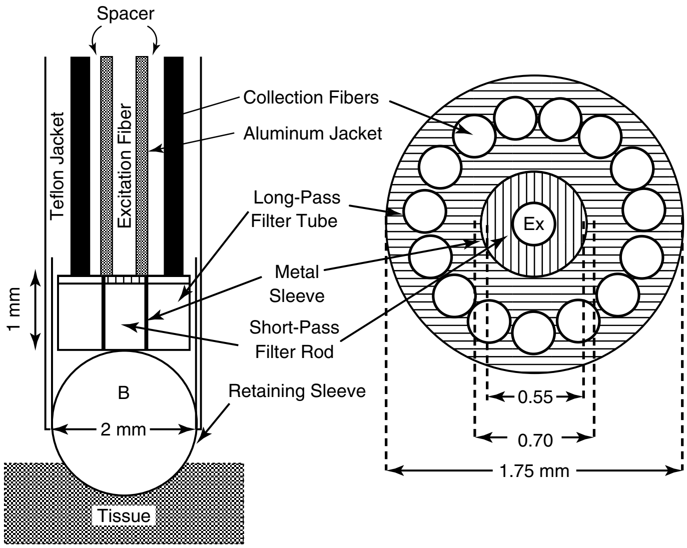

Fig. 24.1 An N-around-1 fiber-optic Raman probe. In this probe design, dielectric filters are used to block silica Raman scatter generated in the excitation (laser delivery) fiber and to block backscattered laser light from entering the collection fibers. A ball lens is used to focus the sample onto the specimen, which is skin tissue. (Reprinted by permission from reference (Motz *et al.*, 2004).)

depth measurements (Everall, 2000a,b). Index matching and the use of immersion objectives are needed to minimize this problem (Everall *et al.*, 2007).

Almost all commercial systems are constructed around standard epi-fluorescence microscope frames made by the major microscope makers. Coupling optics are added, but the microscope hardware is usually unchanged. For safety reasons, it is not common for a Raman microprobe to include binocular eyepieces. A video camera is used for viewing the sample and focusing the instrument. Because Raman microprobes operate in the visible or near-infrared, the spectroscopist has ready access to the wide range of specialized objectives used for light microscopy. Temperature-controlled stages and many other accessories made for the microscope frame can also be used.

A schematic Raman microprobe is shown in Figure 24.2. Coupling optics bring laser light to the microscope objective and direct Raman scatter to the spectrograph. Direct (i.e., with lenses and mirrors) coupling is common. Some vendors offer fiber optic coupling, which provides greater flexibility in component positioning and facilitates using the same spectroscopy optics with other sampling systems. However, to retain the spatial resolution of the direct-coupled instrument, it is necessary to use a single mode optical fiber to deliver the laser light to the microscope. A multimode fiber can be used to direct the Raman scatter to the spectrograph.

A basic instrument operates as a single point microprobe. Mapping is

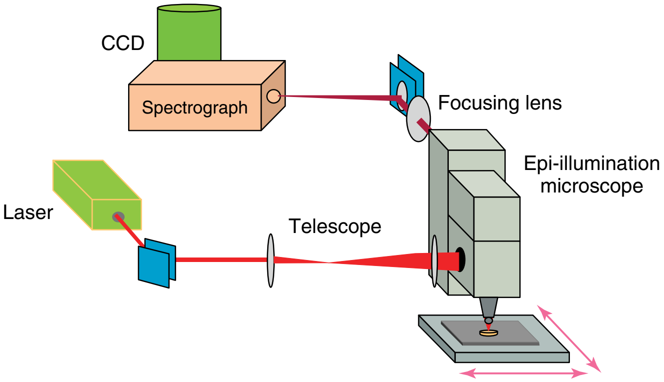

Fig. 24.2 Schematic of a Raman microprobe. Light from the laser is shaped by the telescope and focused on the sample by a microscope objective. Raman scatter is collected through the objective and focused onto the entrance slit of the spectrograph. In this simple design, a manual or motor-driven stage can be used to build up a Raman map pixel by pixel.

provided by motor-driven or piezoelectric *XY* stages, or by scanning mirror systems similar to those used in confocal microscopy. Motor-driven *XY* stages have a settling time that can approach 1 second, and this dead time can limit the speed at which point–point images can be acquired. Piezoelectric stages settle in less than 1 ms. The combination of a piezoelectric stage and an EMCCD allows very fast mapping, at least with green laser excitation. Line-focusing of the exciting laser can be used to acquire an entire line of spectra simultaneously, while increasing the total power that can be safely focused on the sample. In this case, only one-dimensional scanning is needed to generate an image, and the total image acquisition time can be much shorter than if point-point scanning is used.

24.2.5
Data Reduction

Raman spectroscopy is often used for material or contaminant identification. Visual inspection or computer-aided matching to a spectrum in a database of Raman spectra is often all that is needed. Because the intensity of Raman scatter is directly proportional to the number of molecules interrogated, quantitative analysis using band heights or areas and standards of known concentration is satisfactory. However, inaccurate subtraction of fluorescence or background can introduce errors into this seemingly simple operation. The merits and problems of background subtraction protocols remain contentious.

In industrial process control, small changes in complex materials are frequently sought. Tracking the change in the height or area of a single band or of a band height or area ratio is often not satisfactory. In such cases, a multivariate model, typically a partial least squares model, is used instead. Models contain contributions from throughout the Raman spectrum and thus strongly depend on reproducible instrument and process conditions. They can yield erroneous results if the process conditions or instrument parameters change, even if the change is small.

24.3
Identification and Quantification of Inorganic Materials

24.3.1
Introduction

Characterization of inorganic materials by Raman spectroscopy has been reported in topics ranging from surfaces of nanocrystals to characterization of bulk geological minerals. Since the late 1990s, there have been many applications of Raman to new classes of materials and novel approaches to examining existing classes of materials. More recently, fundamental studies of phonons in reduced dimensions have been investigated using Raman spectroscopy. Most reports fall under three broad topics: chemical identification, quantification, and spectroscopic response to mechanical or chemical perturbation (Lyon *et al.*, 1998; Pitt *et al.*, 2005). A comprehensive review of the theory of molecular vibrational spectroscopy and Raman spectroscopy of small inorganic molecules provides both a theoretical and experimental basis for examining a wide variety of inorganic molecules (Nakamoto, 1997). A wide variety of materials, including silicate glasses, cements, catalysts, clay minerals and ceramics (Potgieter-Vermaak, Potgieter and Grieken, 2006; McMillan, 1984; Bartlett and Cooney, 1987; Butler and Frost, 2006), have been examined using Raman spectroscopy. Mineral-based inorganic pigments, forensic physical evidence, nanoscale semiconductors and carbon nanotubes have been recently explored using Raman. In this chapter, Raman identification and quantification in these broad classes of inorganic materials are reviewed. Finally, a case study using Raman spectroscopy as a real-time industrial process control tool is presented.

24.3.2
Minerals

Examination of geological minerals is one of earliest applications of Raman spectroscopy, and it remains a vibrant field of study. A review of the major Raman bands from a wide variety of geology samples was provided by (Griffith, 1987). Various modes of Raman (resonance Raman, Raman microscopy, isotopic substitution) were reviewed and spectroscopic data was presented from two broad families: nonsilicates and silicates. A wide variety of nonsilicate minerals, such as diamond, carbonate minerals, and gypsum, and silicate minerals, such as clay minerals, micas, and olivine, were included in this chapter. Modern Raman applications of geology minerals include astrobiology remote sensing and planetary exploration (Villar and Edwards, 2006; Lipschutz *et al.*, 2007; Wang, Haskin and Cortez, 1998; Sharma *et al.*, 2002; Haskin *et al.*, 1997). Effects of high temperature and/or high pressure on Raman spectra of minerals have been extensively studied (Gillet *et al.*, 1993; Butler and Gilson, 1997; Gillet, Hemley and McMillan, 1998; Butler and Frost, 2006).

24.3.3
Art, Archaeology, and Architecture

Raman spectroscopy is used to examine mineral-based inorganic pigments in works of art, architectural buildings, and archaeology samples. Recent reviews highlighted the many applications of Raman spectroscopy in art and archaeology (Vandenabeele, Edwards and Moens, 2007; Edwards, 2004). The primary usage of Raman spectroscopy is the identification of pigments, corrosion products, or contaminants in manuscript and paintings. Ceramics, glasses, enamels, glazes, minerals, and gem artifacts are also examined with Raman spectroscopy. Spectral libraries of reference inorganic pigments are available in journals (Bell, Clark and Gibbs, 1997; Vandenabeele, Edwards and Moens, 2007) and online. Table 24.1 presents some spectral libraries available online.

Raman spectra of artifact pigment samples are typically collected in a confocal microscopy mode, because of sample heterogeneity and minute sample sizes. The choice of laser wavelength is influenced by sample fluorescence and the nature of the artifact. Many studies use multiple wavelengths to examine a single artifact. Pigments are typically examined using a near-infrared laser while metal corrosion is examined using 514.5 or 532 nm lasers. White pigments were commonly used as wall paint and as a base for the addition of colored pigments. Figure 24.3 shows the commonly used white pigments: (a) gypsum, (b) calcite, (c) barium white, and (d) bone white, adapted from UCL spectra library as described by Bell *et al.* (Bell, Clark and Gibbs, 1997).

Mixtures of mineral particles, resulting from artist blending pigments in the original composition or subsequent restoration efforts, complicate the interpretation of Raman spectra taken from paintings. Carbonaceous particles from soot or dirt contaminate artifacts and are a source of fluorescence. However, they are easily distinguishable in light microscopy images because they are often black particles and Raman spectra of carbon particles do not interfere with the identification of mineral-based inorganic pigments. A complicating issue is pigment degradation, either through environmental-induced or laser-induced pathways. Mixtures of lead corrosion products, such as lead oxide, lead sulfate and lead nitrate, were quantified using Raman spectroscopy (Black, Allen and Frost, 1995). Raman spectra of lead carbonates identified the presence of

Tab. 24.1 Online libraries of Raman spectra for inorganic minerals, pigments, and reference materials.

Source	URL site
University of Catania	http://www.ct.infn.it/~archeo/
Universita di Siena	http://www.dst.unisi.it/geofluids/raman/spectrum_frame.htm
University of Parma	http://www.fis.unipr.it/phevix/ramandb.html
University of Arizona (RRUFF)	http://rruff.info/
Infrared and Raman users group	http://www.irug.org/default.asp
Advanced Industrial Science and Technology (AIST)	http://riodb.ibase.aist.go.jp/rasmin/E_index.html
University College London	http://www.chem.ucl.ac.uk/resources/raman/index.html

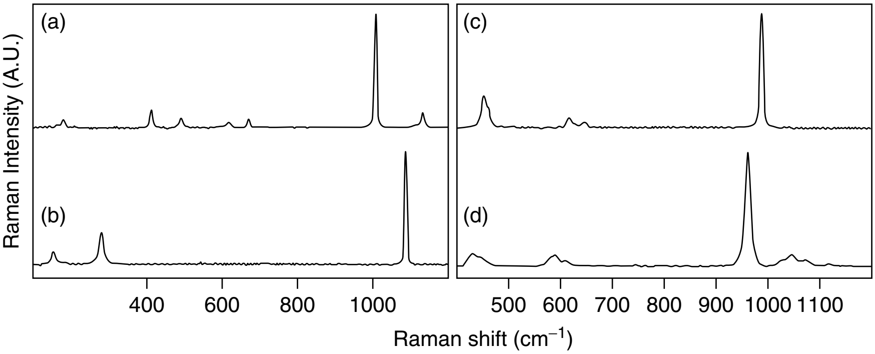

Fig. 24.3 Spectra of white-colored pigments used in paints. Figure 24.3(a) and (b) show gypsum and calcite, and Figure 24.3(c) and (d) are barium white and bone white. Spectra 3a–3d were published previously and reproduced from the University College London online spectral library (Bell, Clark and Gibbs, 1997). (http://www.chem.ucl.ac.uk/resources/raman/index.html.)

adducts and unstable corrosion products (Brooker *et al.*, 1983; Ciomartan *et al.*, 1996).

Raman spectra of paint layers complement historical data on the pigment. These data can be used to estimate the painting's age or direct restoration strategies (Edwards, Farwell and Brooke, 2005). In a study of architectural paint samples, cross-sections of eight paint layers were examined using Raman spectroscopy, light microscopy, and scanning electron microscopy with energy dispersive spectroscopy (SEM–EDS) (Kendix, Nielsen and Christensen, 2004). Raman spectra indicated that each paint layer had its own mixture of pigments, and the Raman data correlated with results from SEM-EDS. The dominant white pigments were calcium carbonate, white lead, titanium dioxide, and barites.

In another study, layered samples paint and stucco were recovered from a preserved Mayan temple found within the Rosalila building (Goodall *et al.*, 2006). Paint/stucco samples from the temple and a stucco mask were examined using Raman, micro attenuated total reflectance infrared spectroscopy (ATR-IR and scanning electron microscopy (SEM) (Goodall *et al.*, 2006). The nature of the green paint was a key finding. Raman bands associated with plagioclase and celadonite were found in green particles, but the bands were too broad and noisy to be definitive. ATR-IR of the same particles reveal a celadonite-based mixture (also called *Terra Verte*), which was confirmed by SEM and micro X-ray. Spectroscopic studies of the Rosalila samples revealed the use of the green Terra Verte pigment, and this pigment was not found in samples collected from another Mayan temple (Vandenabeele *et al.*, 2005).

24.3.4
Forensics

Applications of Raman spectroscopy for nonbiological forensic evidence primarily fall under sample identification. Physical evidence in criminal cases is often in trace amounts, or on a large support structure such as a car panel. Many reports cite flexibility in sampling modes (e.g., microscopy or fiber optics systems), small sample requirements and the nondestructive nature of Raman spectroscopy as an advantage of the method for analysis of trace-level evidence. A 2007 applications review on forensic science highlighted the use of Raman spectroscopy to identify automobile paint, pen ink, and explosives (Brettell, Butler and Almirall, 2007). Raman spectra are primarily used to distinguish physical evidence found at a crime scene from evidence collected from a suspect. Raman spectra of automobile paints identified inorganic fillers or binders, such as magnesium silicate or talc, and inorganic pigments, such as titanium dioxide, lead chromate, or calcite (Buzzini, Massonnet and Sermier, 2006; Zieba-Palus and Borusiewicz, 2006; Gelder *et al.*, 2005). Pigments from spray paint or house paint were also examined by Raman spectroscopy. In one case study, Raman spectra collected from red paint splattered on a suspect's shoe was undifferentiated from spectra of red paint stain on a wall. Raman spectra identified two red paints: (i) Iron oxide using 785 nm excitation and (ii) either chrome orange (lead (II) chromate) or molybdate orange (lead molybdate chromate) in stains at the crime scene. Raman spectra of paint droplets collected from the suspect's shoe also contained bands from the two paint pigments and it was concluded that the shoe was in contact with the paint used to damage the wall (Buzzini, Massonnet and Sermier, 2006). As seen in Figure 24.4, Raman spectra collected from a suspect's shoe indicated a mixture of two red paints, which were similar to the spectra collected from the crime scene.

Proof of concept for metals and anions identification in gunshot residue was shown by (Stich *et al.*, 1998). Gunshot residues were produced from 10, 20, or 30 cm from the paper substrate. Residues left on a paper substrate were examined under a Raman microscope ($\lambda = 633$ nm) in three domains surrounding the impact point. Raman spectra showed heterogeneity in the chemical composition of residues, which was dependent on the distance of the particle from the impact site. In the region closest to the impact site, barium carbonate was the dominant component. In regions farther from the impact site, a mixture of lead-based and barium-based particles were found, and the oxidation state (carbonate or oxide) also varied.

24.3.5
Nanoscale Semiconductors

A review of Raman studies from 1995 to 1997 that characterized semiconductor surfaces was provided by (Lyon *et al.*, 1998). Later reviews in 2003 and 2004 provide a theoretical basis of analyzing fundamental semiconductor properties such as phonon interactions, doping, strain, crystallinity, and surface composition; resonance Raman and polarized Raman were also reviewed (Jimenez, Wolf and Landesman, 2003; Nakashima, 2004). As seen in Figure 24.5, Raman bands are sensitive to local environments and differences in band shape, intensity, or width can be used to differentiate regions on a semiconductor surface.

Since 1998, two nanoscale materials have emerged that have unique electronic properties: single-wall nanotubes (SWNTs) made from carbon and quantum dots (QDs) composed of semiconductor materials. QDs are nanoparticles composed of semiconductor materials, typically with a

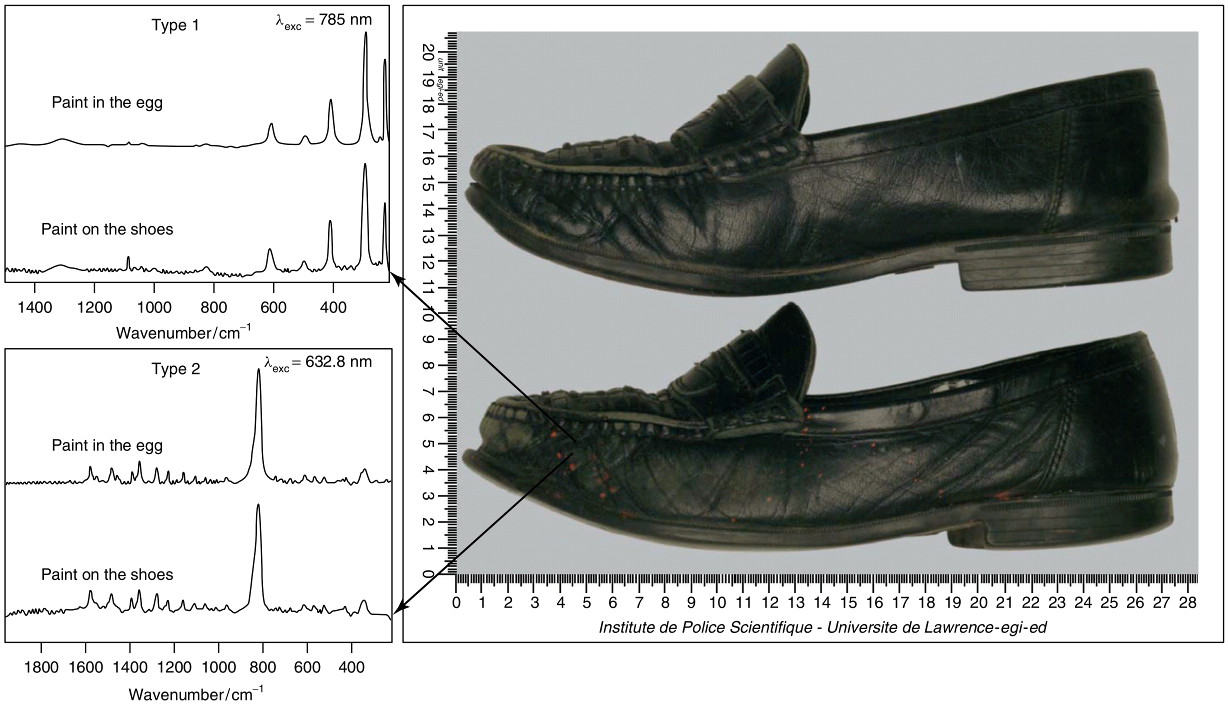

Fig. 24.4 Raman spectroscopy used in forensic evidence analysis shows correlations between evidence collected at a crime scene and evidence collected from a suspect. Paint was thrown onto the wall of a home in small ''eggs,'' and Raman spectra of the splattered paint showed a unique mixture of two red/orange pigments. Raman spectra collected from paint evidence on a suspect shoe were undifferentiated from spectra collected from a crime scene. (Reprinted by permission from reference (Buzzini, Massonnet and Sermier, 2006).)

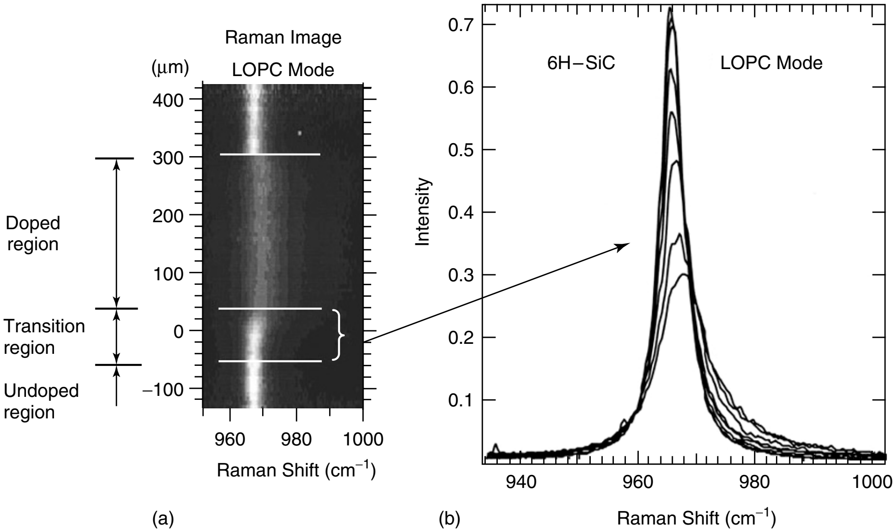

Fig. 24.5 (a) shows a Raman image of a modulation-doped 6H-SiC semiconductor surface and (b) shows Raman spectra of the longitudinal optical phonon plasmon coupled (LOPC) mode in the transition region. Raman band position and band width varied in the transition zone as a result of starting and stopping of carrier gas used in modulated doping experiments. (Figure is reproduced with permission from reference (Nakashima, 2004), copyright Institute of Physics.)

cadmium/selenium core, with potential applications as laser sources, storage materials, fluorescent dyes, and sensors (Voronin, 2006; Cottingham, 2005; Murphy, 2002). Exciting applications in diffusion, flow cytometry, and biomechanics experiments show the broad applicability of QDs in biological fluorescence or luminescence imaging (Dahan *et al.*, 2003; Gao and NIe, 2004; Golcuk *et al.*, 2007).

Raman spectroscopy has proven valuable in characterizing the chemical and electronic properties of QD's. Raman spectroscopy was used to understand the fundamental phonon behavior in reduced dimensions, characterize QD surfaces after chemical doping or mechanical strain, and measure the effects of temperature on longitudinal optical (LO) or transverse optical (TO) phonon frequency. Rolo and Vasilevskiy reviewed theoretical and experimental results of optical phonon modes using Raman spectroscopy (Rolo and Vasilevskiy, 2007). Resonance Raman spectroscopy and nuclear magnetic resonance (NMR) spectroscopy monitored exciton-phonon interactions; results showed improved sensitivity at the molecular level for single QDs and QDs in an ensemble (Gammon *et al.*, 1997).

Carbon nanotubes, multiwalled or single-walled, are frequently investigated by Raman spectroscopy because graphitic structures have intense Raman bands (Barbarossa *et al.*, 1991; Everall and Lumsdon, 1991; Lespade *et al.*, 1984; Gibson, Ayorinde and Wen, 2007). These nanotubes are proposed as reinforcing materials in polymer composites (Coleman *et al.*, 2006; Bokobza, 2007). Raman spectroscopy has emerged as a preferred method for characterizing these materials and has been suggested as a method for monitoring their manufacture (Degamber and Fernando, 2002). In particular, single-walled carbon nanotubes (SWCNTs) have unique electronic, mechanical, and chemical properties; their uses as a composite material or electronic device have been reviewed (Popov, 2004). Raman spectra of SWCNTshave provided valuable insight into the electronic and the phonon properties of SWCNTs. Raman spectra of SWCNTs are sensitive to tube diameter, functionalization, and mechanical strain (Dresselhaus *et al.*, 2002, 2005; Graupner, 2007). A recent study of whole-body Raman imaging using surface-enhancing carbon nanotubes indicates exciting applications of SWCNTs in biomedical imaging (Keren *et al.*, 2008).

24.3.6
Industrial Process Control

High specificity, spatial resolution, molecular identification, and ease of use are features that make Raman an attractive technology for monitoring inorganic chemical production (Workman *et al.*, 2001; Workman, Koch and Veltkamp, 2003). Raman has been used to examine chlorosilane streams as a complementary technique to gas chromatography (GC) analysis (Lipp and Grosse, 1998). A fiber-optics Raman system has been used for direct monitoring of the corrosive stream, where several species were identified and quantified. Raman PAT monitored fluctuations in stream components that indicated poor performance of distillation columns.

Direct monitoring of titanium dioxide production using a fiber-optic Raman spectroscopy instrument was reported by (Clegg *et al.*, 2001). Two forms of TiO_2, anatase and rutile, are made during one step (calciner discharge) of the synthetic route. However, the amount of anatase must be controlled so that the final product retains good optical properties. Powder X-ray diffraction (XRD) had been used as an off-line measurement of quality; however, measurements were performed infrequently. More frequent quality measurements were hypothesized to enable a more efficient calciner discharge step, and Raman spectroscopy was chosen as the analytical tool. Raman spectra of anatase and rutile in the 150–800 cm^{-1} region have unique bands. Raman spectra of anatase/rutile mixtures show that the intensity ratio of 142 : 610 cm^{-1} bands correlated to ratios measured by XRD. Effects of temperature, crystal size, and window cleanliness on anatase measurement accuracy were also discussed.

24.4
Identification and Quantification of Organic Materials

24.4.1
Pharmaceutical Applications

Raman spectroscopy is a major component of process analytical technology (PAT) (Bakeev, 2005), the pharmaceutical industry's term for process control analytical chemistry and related topics. The most important PAT application of Raman spectroscopy is crystallization monitoring and control (Févotte, 2007; Yu *et al.*, 2004, 2007). The application is important because the polymorph determines such important properties as

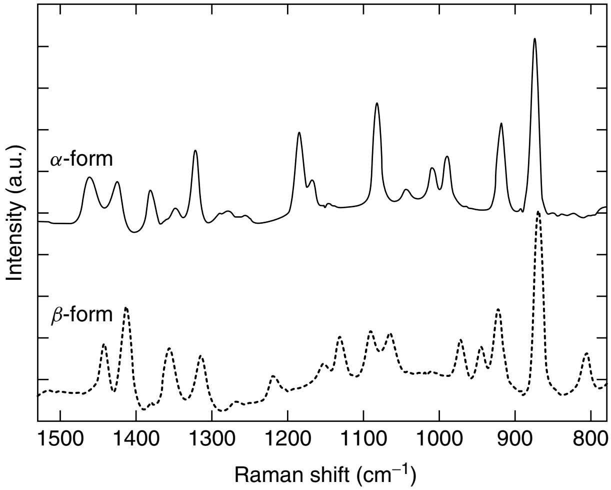

Fig. 24.6 Raman spectra of the metastable α-form and the stable β-form of L-glutamic acid. (Reprinted with permission from reference (Scholl *et al.*, 2006).)

solubility and rate of dissolution. Establishment of the theoretical and the experimental frameworks that define conditions under which the desired polymorph is the only or at least the dominant form are the major goals of researchers (Rodrìguez-Spong *et al.*, 2004; Llinàs and Goodman, 2008).

Crystallization monitoring depends on the differences in the Raman spectra of different polymorphs. L-Glutamic acid, which has two polymorphs, offers a dramatic example (Figure 24.6) (Scholl *et al.*, 2006). The metastable α-form is kinetically favored and is slowly converted to the thermodynamically more stable β-polymorph. In some cases, the differences in the spectra may be more subtle than those for L-glutamic acid, but they will be easily found by PCA, if not by visual inspection (Falcon and Berglund, 2004; Okumura and Otsuka, 2005). However, some authors consider that polymorph spectra differ sufficiently that univariate calibration is almost always adequate (Rantanen *et al.*, 2005).

Because of the ease of use and the ability to monitor changes online (O'Sullivan *et al.*, 2003), Raman spectroscopy has increasingly displaced powder X-ray diffraction in crystallization monitoring. There are also differences in the spectra of amorphous materials and crystalline forms, enabling on-line measurement of crystallinity (Nakano, Tokumura and Umakoshi, 2002). However, there are complications caused by presence of highly scattering materials, whose number and size are changing throughout the process. Sampling a large volume through use of a nonconfocal probe with large spot size, such as the Kaiser Optical Systems PhAT probe is an effective approach to this problem (Wikström, Lewis and Taylor, 2005). Another approach to improved quantification has been to combine Raman spectroscopy with image analysis using a conventional fiber probe and a small video camera simultaneously (Caillet *et al.*, 2007).

Other major pharmaceutical applications include monitoring of homogenization (De Beer *et al.*, 2006), inspection of incoming materials, and inspection of tablets for composition and uniformity. Transmission Raman spectroscopy may be nearly ideal for composition monitoring in formulated products, because it is rapid and provides accurate measures of average composition of a tablet (Matousek and Parker, 2007; Eliasson *et al.*, 2008).

24.4.2
Polymer Chemistry and Physics

Raman spectroscopy plays major roles in polymer processing and analysis. Even today, the technique remains a major tool in the characterization of entire classes of polymers (Billes and Mohammed-Ziegler, 2007). Most commercially important polymers do not themselves fluoresce under visible light, but many formulations contain additives or impurities that do. Excitation with deep red or near-infrared lasers is often needed to circumvent a fluorescence background. Many major applications are to measurement of physical properties such as crystallinity, porosity, stress, and film or fiber orientation. These are the properties that must be adjusted to obtain the desired application performance. Except in contamination studies, a known chemical composition is usually the starting point of the measurement.

It has long been known that the Raman spectra of crystalline and amorphous forms of most polymers are different (Stuart, 1996). The polymer conformation usually determines crystallinity, and markers can include intensity, position, and band width changes. Applications continue to appear (Rasha, 2006). Although most applications use single point spectroscopy only, high definition imaging has been used to map crystallinity in syndiotactic polystyrene (Zhang *et al.*, 1998).

Because polymer bands undergo small shifts when a polymer is stretched or bent, Raman spectroscopy can be used to probe the conformational changes associated with these deformations (Colomban, 2002; Young and Eichhorn, 2007). Polarized spectroscopy is generally used, because more detailed information is available (Sourisseau and Talaga, 2006). The technique has proven useful in measurement of orientation in films (Sourisseau, 2004; Reinecke *et al.*, 2001) as well as in fibers. The same techniques that have been developed for polymers and polymer blends can be applied to composites (Colomban, 2002; Tomba, Carella and Pastor, 2006).

24.4.3
Forensic and Homeland Security Applications

The Raman microprobe has long been used for the analysis of small particles, such as fibers, paint chips, and hairs (Adar, 1988). It is used primarily with samples for which the spatial resolution of an infrared microprobe is inadequate. Recently, there has been extensive application to forensic analysis, for example (Buzzini, Massonnet and Sermier, 2006; Massonnet *et al.*, 2003, 2005; Thomas *et al.*, 2003a,b). Analysis of paint chips from automobile accidents and from burglaries has been especially important. Buzzini and coworkers have presented case studies that illustrate the range of problems that are encountered in these two important areas (Buzzini, Massonnet and Sermier, 2006). Identification of illicit drugs (Hodges *et al.*, 1989; Huong, 1986) is a closely related example.

With increasing interest in detection of explosives, Raman spectroscopy is a candidate for rapid detection (Carter *et al.*, 2005; Cheng *et al.*, 1995; Hayward *et al.*, 1995; Lewis *et al.*, 1997). If low duty cycle pulsed lasers and gated intensified CCD detectors are used, stand-off systems that can operate in ambient light are feasible (Carter *et al.*, 2005). Recently developed fiber-optic spatial offset techniques that can retrieve spectra through translucent containers may provide a convenient means for the detection of concealed explosive materials or component materials for making explosives (Eliasson, Macleod and Matousek, 2007).

24.5
Biological Applications of Raman Spectroscopy

Raman spectroscopy has evolved as a versatile biochemical analysis tool with applications ranging from the analysis of subcellular components to the diagnosis of diseases and *in situ* evaluation of whole tissues. The positions and relative intensities of various Raman spectral bands can be used to quantitatively probe the structure and function of important biological molecules.

24.5.1
Subcellular Level

Raman band assignments, secondary structure signatures (Bandekar, 1992), and markers of side-chain environment (George, 2002) have been established for proteins. The Raman spectrum of a protein is largely dominated by bands associated with the peptide main chain, aromatic side chains, acidic residues, and sulfur-containing side chains. Amino acids that are important in primary structure are tyrosine with peaks at 830 and 850 cm^{-1}, tryptophan with peaks at 1014, 1338, 1361, and 1553 cm^{-1}, and phenylalanine with a peak at 1006 cm^{-1}. Amide I (C=O stretching) and amide III (C–H stretching and N–H in-plane bending) modes are usually used to characterize the secondary structure of the peptide backbone. The wavenumbers at which the amide I and amide III Raman bands occur for α-helical, β-sheets, and random-coiled structures are given in Table 24.2. Apart from the study of folded active proteins, Raman spectroscopy has the ability to provide a variety of probes into the structure of natively unfolded proteins (see Figure 24.7) (Maiti *et al.*, 2004). Protein-based enzymes have been studied in real time by Raman crystallography (Carey, 2006).

The characteristic Raman bands of phospholipids arise from the vibrations of the hydrocarbon chains. The hydrocarbon chains can exist with a high degree of ordering (trans form) or in relative disorder (gauche form)

Tab. 24.2 Raman wavenumbers (per centimeter) for amide I and amide III bands for various polypeptide conformations.

Protein secondary structure	Amide I	Amide III
α-Helix	1645–1660	1265–1300
β-Sheet	1665–1680	1230–1240
Random coil	1660–1670	1240–1260

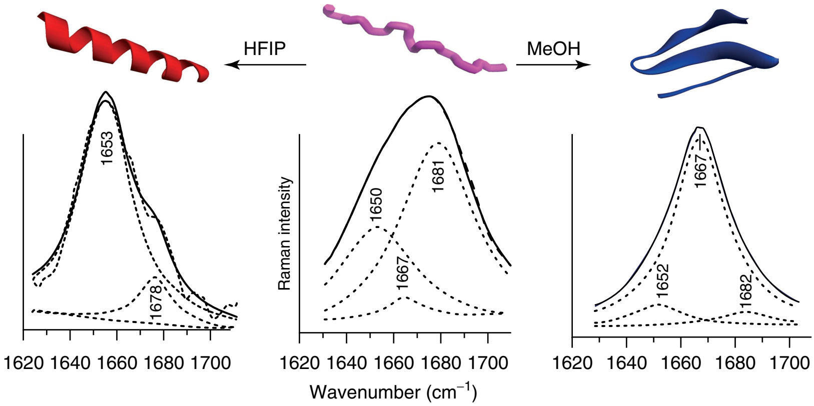

Fig. 24.7 Raman spectra of α-synuclein, a natively unfolded protein, in the presence of hexafluoro-2-propanol (HFIP) and methanol (MeOH). The secondary structure becomes largely α-helical in HFIP (1653 cm^{-1}) and predominantly β-sheet (1667 cm^{-1}) in 25% MeOH in water. (Reprinted with permission from reference (Maiti *et al.*, 2004), copyright American Chemical Society.)

(Mendelsohn and Moore, 1998). The phospholipids membrane conformation is temperature sensitive and can be probed with Raman spectroscopy. Sample Raman spectrums of cholesterol, a lipid, and lecithin, a phospholipid, are shown in Figure 24.8 (Tantipolphan *et al.*, 2006).

The Raman spectrum of a nucleic acid is dominated primarily by bands caused by vibrations localized either within the heterocyclic bases (500–800 cm^{-1}) or in backbone phosphate groups (800–1100 cm^{-1}). These Raman bands are employed as sensitive indicators of local structure, global conformation, intermolecular interaction, and molecular dynamics. Table 24.3 indicates the location of various Raman peaks in the spectra of different conformations of DNAs (Mahadevan-Jansen and Richards-Kortum, 1996). The Raman spectra of DNA–protein complexes yield information about the composition, secondary structure, and interactions of these molecules, including the chemical microenvironment of molecular subgroups.

24.5.2
Tissue Level

In recent years, Raman spectroscopy has seen a surge in applications to the study and diagnosis of diseases. The main aspect of Raman spectroscopy of cells and tissue is that it is capable of probing the subcellular milieu of intact cells. In this section, we describe the applications of Raman spectroscopy by tissue type. For a review on surface-enhanced Raman scattering and its biological applications, refer to these articles (Katrin *et al.*, 2002; Cotton, Kim and Chumanov, 1991).

24.5.2.1 Epithelial Tissue

The majority of the cancers originate in the epithelial tissue, such as esophagus, colon, larynx, bladder, prostate, cervix, breast, skin, lung, and lymph nodes. As these are readily accessible to fiber-optic probes in many cases,

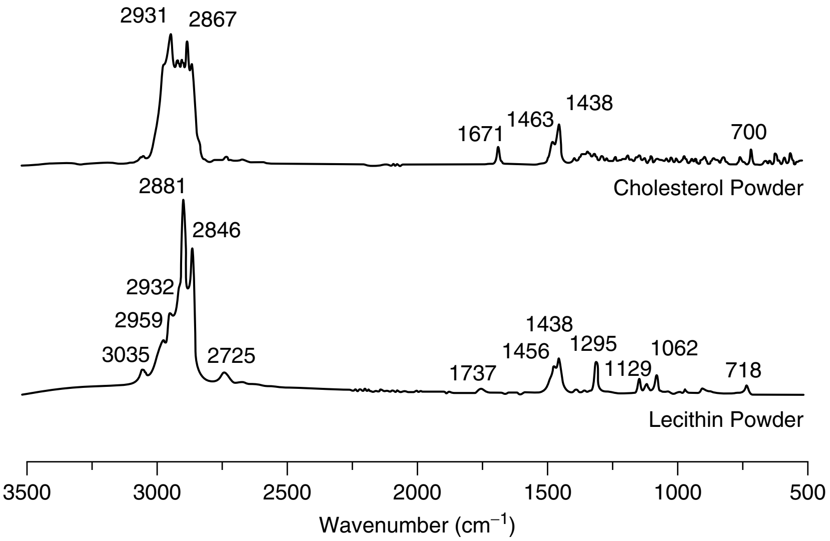

Fig. 24.8 Raman spectra of soybean lecithin and cholesterol powder. (Reprinted with permission from reference (Tantipolphan *et al.*, 2006), copyright Elsevier.)

Tab. 24.3 Raman bands (per centimeter) for the A, B, C, and Z backbone configurations of DNA.

Nucleic acid (form)	A	C	G	T	OPO	PO_2^-	Deoxy-ribose
DNA (A)	727	780	666	642	806–813	1099	
	1335		1318	777			
				1239			
DNA (B)	727	782	682	665	825–842	1091	917
	1339	1250	1333	748			975
				1208			1448
							1462
DNA (C)			670		785	1090	
DNA (Z)	624		640		742–748	1095	
	729		1316				

Mahadevan-Jansen and Richards-Kortum (1996).

Raman spectroscopy has the potential to identify molecular and cellular changes associated with cancer cell proliferation (Short *et al.*, 2005). The use of multivariate spectral predictive models has enabled discrimination between benign and malignant tissue to a high degree of accuracy both *in vitro* and *in vivo*. In esophageal cancer, Raman spectroscopy has demonstrated higher levels of guanine, adenine, cytosine, uracil, O–P–O, β-sheet and disordered proteins and lower concentrations of glycogen, tryptophan, proline, and α-helix proteins in cancerous tissue as compared to benign tissue (Kendall *et al.*, 2003). In colon cancer, Raman spectroscopy has reported higher levels of guanine, adenine, cytosine, uracil, O−P−O, lipids, disordered proteins, phenylalanine, and lower levels of tyrosine and carotenoids than in benign tissue (Boustany *et al.*, 1999). In breast cancer, malignant tissue has shown greater concentrations of hydroxyproline and collagen and lower levels of glucose, lipid, and carotenoids than benign tissue (Redd *et al.*, 1993). Neoplastic prostate tissue has been demonstrated to have higher concentrations of guanine, adenine, cytosine, uracil, O−P−O, beta-sheet proteins, lipids, porphyrins, and lower levels of tryptophan, tyrosine, α-helix proteins, phenylalanine, and water compared to benign tissue (Crow *et al.*, 2003). In malignant lung tumor tissue, Raman spectroscopy has identified higher levels of nucleic acid, tryptophan and phenylalanine and lower levels of phospholipids, proline, and valine compared to normal tissue (Huang *et al.*, 2003). Raman spectroscopic studies on skin have looked at chemical enhancing agents for the percutaneous absorption of drugs (Williams *et al.*, 2006), the effect of UV-radiation protective agents in sunscreens for skin (Schallreuter and Wood, 2001) and the effect of aging on the water and protein structures in skin (Williams, Edwards and Barry, 1995).

24.5.2.2 **Muscular Tissue**

Raman spectroscopy has been applied to the study of contraction-induced conformational changes in muscle fibers and proteins of the skeletal muscles. Studies on the effect of Ca^{2+}, Mg^{2+}, and ATP on the contraction of muscle fibers showed that the contraction is characterized by a decrease in the intensity of Raman bands of protein side chains and an increase in the intensity of amide I and C−C stretch bonds (Caillé, Pigeon-Gosselin and Pézolet, 1983). A recent Raman spectroscopic study monitored changes in the relative levels of NADH and NAD in the bicep muscles of mice under oxidative stress (Sriramoju *et al.*, 2008). This approach could be useful in the study of age- and stress-induced changes in muscles and disorders such as muscular dystrophy and ischemia. Raman spectroscopy has been used to identify and classify different layers of tissue in bladder pathology and bronchial studies. Studies on bladder walls showed that the spectra of urothelium (lipids), lamina propria (collagen), and smooth muscle layer (myosin and actin) are easily distinguishable (de Jong *et al.*, 2002). Raman spectroscopy of bronchial tissue distinguishes between fibrocollagenous stroma tissue and smooth muscle, and could be used to diagnose histological changes in the lung tissue due to lung cancer (Koljenovic *et al.*, 2004).

24.5.2.3 **Nervous Tissue**

Raman spectroscopy has been used to study the human brain and some of the associated tumors. In normal brain tissue, the spectra of white matter show a greater contribution from lipids, proteins, and cholesterol, whereas the spectra of gray matter exhibit a large contribution from water (Mahadevan-Jansen and Richards-Kortum, 1996). In general, both gray and white matter spectra exhibit peaks from proteins and lipids (Mizuno *et al.*, 1994). Intracranial tumors such as meningeoma, glioma, and schwannoma, which originate from the meninges, glial cells, and Schwann cells, respectively, have been studied (Krafft *et al.*, 2005). Lipid concentration and composition in the brain have been observed to change indicative of brain tumors. Calcification of tissue has also been reported in the case of brain tumors (Mizuno *et al.*, 1994). A Raman map from a human brain tumor section is shown in Figure 24.9 (Krafft, 2004). Raman spectroscopy has been proposed as a water concentration monitoring tool for brain edema in addition to the intracranial pressure measurements using MRI and CT scans (Wolthuis *et al.*, 2001). Because protein misfolding underlies the pathology of several neurodegenerative conditions, Raman microscopy has been used to characterize the biochemistry of transmissible spongiform encephalopathies (Carmona *et al.*, 2004) and Alzheimer's disease (Sudworth *et al.*, 2006).

24.5.2.4 **Connective Tissue**

Raman spectroscopy has the ability not only to characterize tissue but also to offer quantitative information about tissue components. Quantitative measurements on analytes in blood serum (Rohleder, Kiefer and Petrich, 2004) and whole blood (Enejder *et al.*, 2002) have been made. Raman spectroscopy has also been used to identify microbial pathogens (Maquelin *et al.*, 2003) in blood. Further, it can accurately quantify the relative amounts of cholesterol, calcium salts, triglycerides, and phospholipids in homogenized arterial tissue (Brennan *et al.*, 1997) and hence, can be used to identify atherosclerotic plaque (van de Poll *et al.*, 2002).

Raman spectroscopy has the advantage of being sensitive to both the mineral and the organic components of mineralized tissue and is applicable to wet, thick solid specimens of bone. Raman spectroscopy has been used to study the composition (Penel *et al.*, 2005), chemical changes accompanying mechanical deformation (Morris *et al.*, 2004), aging (Freeman *et al.*, 2001; Ager *et al.*, 2005), and disorders (McCreadie *et al.*, 2006) of bone tissue and to characterize interactions between prosthetic implants and bone tissue (Bertoluzza *et al.*, 1994; Dippel *et al.*, 1998). Raman spectroscopy has also been used to study cartilage and its associated disorders (McCreadie *et al.*, 2006). Dental enamel has a similar composition as that of bone mineral. Polarized Raman spectroscopic studies have been performed to study the orientation of crystallites in dental enamel (Tsuda and Arends, 1994) and in dental caries (Ko *et al.*, 2006). Noninvasive Raman spectroscopy and imaging of biological specimens is a nascent, rapidly advancing field. Transcutaneous measurements on cadaveric and *ex vivo* bone specimens using spatially separated optical fibers were first reported in 2006 (Schulmerich *et al.*, 2006). In these measurements, depths of 3–4 mm below the skin were reached. Recently, diffusive Raman tomography has been demonstrated on canine bone tissue, also showing that Raman tomography can be integrated with CT and MRI (see Figure 24.10) (Schulmerich *et al.*, 2008).

24.5.3
Microorganisms

Raman spectroscopy has been developed as a simple and rapid identification and characterization tool for pathology. Single bacteria (Rösch *et al.*, 2005), yeast, (Rösch *et al.*, 2005) and cellular components of single spores (Esposito *et al.*, 2003) have been studied. Raman spectroscopy has been able to detect trace levels of microorganisms in the presence

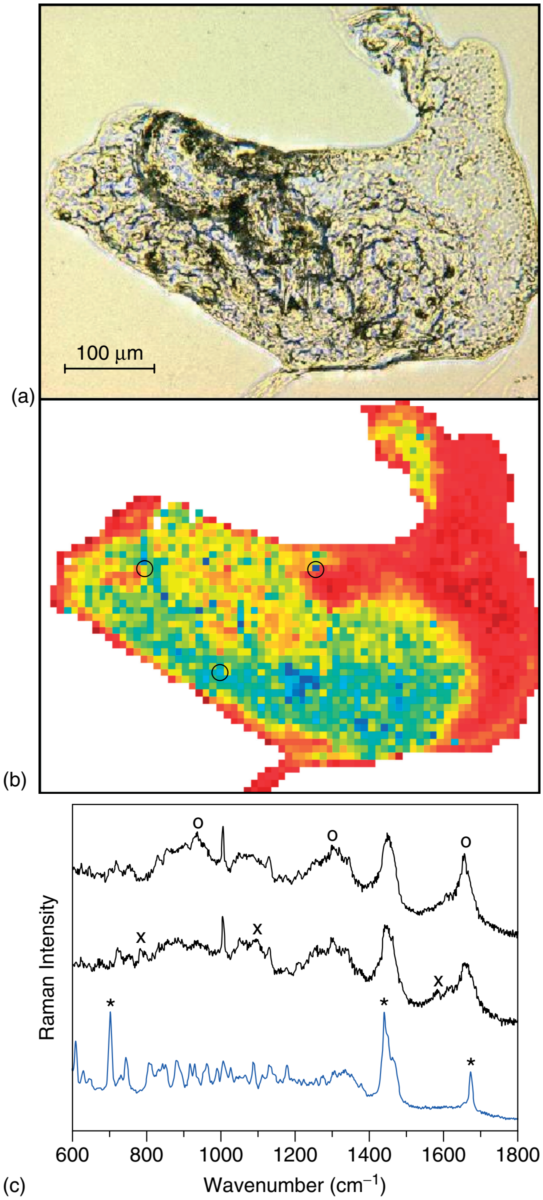
100 μm
(a)
(b)
o
o
o
x
x
x
*
*
*
Raman Intensity
600
800
1000
1200
1400
1600
1800
Wavenumber (cm^{-1})
(c)

Fig. 24.9 (a) Photomicrograph of dried astrocytoma WHO grade 2 human brain tumor section. (b) Color scaled plot based on a Raman map of 74 × 57 pixels acquired with a step width of 7.5 μm and 1-minute exposure time per spectrum. (c) Selected Raman spectra from B (indicated by circles) in the wavenumber range 600–1800 cm^{-1}. Protein bands characteristic of α-helix or collagen helix (top spectrum) are marked by open dots, DNA bands (middle spectrum) by crosses, and cholesterol bands (bottom spectrum) by asterisks. (Reprinted with permission from reference (Krafft, 2004), copyright Springer-Verlag.)

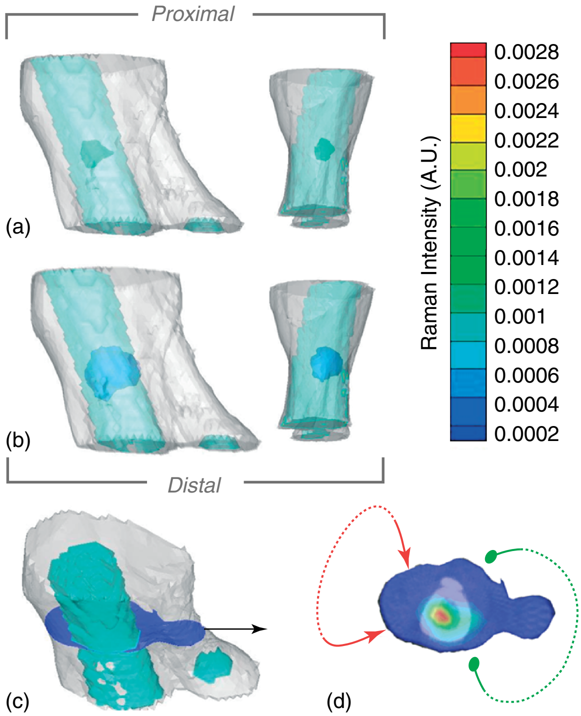

Fig. 24.10 Raman tomographic images of canine bone tissue. (a) Medial and anterior views of soft tissue mesh (white) and bone surface mesh of tibia and calcaneus (turquoise) overlaid with 50% contrast isosurface of the reconstructed three-dimensional (3-D) Raman image of bone (green). (b) Same view as (a) overlaid with 10% contrast isosurface of the reconstructed Raman image of bone (blue). (c) 3-D mesh of limb section (white), including bone (turquoise), illustrating location of the cross section (blue) containing the highest Raman scatter intensity. (d) Raman intensity at cross section in (c) in pseudocolor overlaid on the micro-CT image of the bone, showing range of illumination (red arrows) and collection (green dots) positions. (Reprinted with permission from reference (Schulmerich *et al.*, 2008), copyright Society of Photo-Optical Instrumentation Engineers.)

of complex environmental backgrounds when used in combination with fluorescence spectroscopy and digital imaging (Kalasinsky *et al.*, 2007). The spatial distribution of different molecular species inside a living yeast cell has been probed (see Figure 24.11) (Huang *et al.*, 2005). In another interesting study, gram levels of anthrax spores in envelopes were detected on a mail sorting system using Raman spectroscopy (Farquharson *et al.*, 2004). Building Raman spectral reference libraries of all microorganisms of interest has great potential in the study

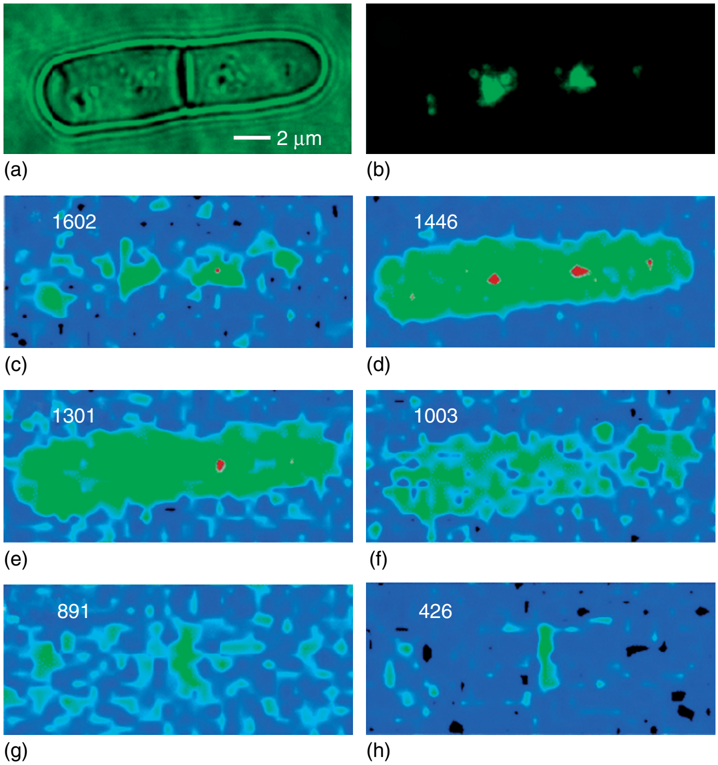

Fig. 24.11 Raman mapping of a living yeast cell – *S. pombe* cell in the G1/S phase: (a) microscopic optical image, (b) GFP image, (c) 1602 cm^{-1}, (d) 1446 cm^{-1}, (e) 1301 cm^{-1}, (f) 1003 cm^{-1}, (g) 891 cm^{-1}, and (h) 426 cm^{-1}. (Reprinted with permission from reference (Huang *et al*., 2005), copyright American Chemical Society.)

of surface contaminants, bioaerosols, water-borne, and food-borne pathogens.

Glossary

Epi-Fluorescence: A microscopy technique in which the same objective is used for the illumination of a specimen and collection of the excited fluorescence.

Exciton: A bound state of an excited electron and a hole in the lattice of a semiconductor or insulator. Excitons can move around the crystal lattice of the material in which it is formed.

Holographic Filter: An optical filter in which a three-dimensional diffraction pattern is formed in a photosensitive material. Depending on the pattern, the holographic filter can function as a wavelength sensitive mirror (also called a *notch filter*), diffraction grating or other type of filter.

Phonon: A vibrational mode of a crystal lattice.

Surface Plasmon: A surface electromagnetic wave that propagates parallel to the surface of a metal. Surface plasmons are excited at wavelengths that depend on the metal and the three-dimensional surface structure. Silver plasmons can be excited with visible radiation. Gold and copper plasmons can be excited with deep red and near-infrared radiation.

Tomography: A three-dimensional imaging technique in which the image is constructed from a series of sections obtained at different points in a solid material.

References

Adar, F. (1988) *Microchem. J.*, **38**, 50–79.

Adar, F. (2001) Evolution and revolution of raman instrumentation – application of available technologies to spectoscopy and microscopy, in *Handbook of Raman Spectroscopy* (eds I.R. Lewis and H.G.M. Edwards), Marcel Dekker, Inc, New York.

Ager, J.W., Nalla, R.K., Breedon, K.L., and Ritchie, R.O. (2005) *J. Biomed. Opt.*, **10**(3), 034012.

Bakeev, K.A. (ed.) (2005) *Process Analytical Technology*, Blackwell, Oxford.

Bandekar, J. (1992) *Biochim. Biophys. Acta: Protein Struct. Mol. Enzymol.*, **1120**(2), 123–143.

Barbarossa, V., Galluzzi, F., Tomaciello, R., and Zanobi, A. (1991) *Chem. Phys. Lett.*, **185**, 53–55.

Bartlett, J.R. and Cooney, R.P. (1987) Raman spectroscopic studies in chemisorption and catalysis, in *Spectroscopy of Inorganic-Based Materials* (eds R.J.H. Clark and R.E. Hester), John Wiley & Sons, Inc, New York.

Battey, D.E., Slater, J.B., Wludyka, R., Owen, H., Pallister, D.M., and Morris, M.D. (1993) *Appl. Spectrosc.*, **47**, 1913–1919.

Bell, I.M., Clark, R.J.H., and Gibbs, P.J. (1997) *Spectrochim. Acta A Mol. Biomol. Spectrosc.*, **53**, 2159–2179.

Bertoluzza, A., Fagnano, C., Tinti, A., Morelli, M.A., Tosi, M.R., Maggi, G., and Marchetti, P.G. (1994) *J. Raman Spectrosc.*, **25**(1), 109–114.

Billes, F. and Mohammed-Ziegler, I. (2007) *Appl. Spectrosc. Rev.*, **42**(4), 369–441.

Black, L., Allen, G.C., and Frost, P.C. (1995) *Appl. Spectrosc.*, **49**(9), 1299–1304.

Bokobza, L. (2007) *Polymer*, **48**(17), 4907–4920.

Boustany, N.N., Crawford, J.M., Manoharan, R., Dasari, R.R., and Feld, M.S. (1999) *Lab. Invest.*, **79**(10), 1201–1214.

Brennan, J.F. III, Romer, T.J., Lees, R.S., Tercyak, A.M., Kramer, J.R. Jr., and Feld, M.S. (1997) *Circulation*, **96**(1), 99–105.

Brettell, T.A., Butler, J.M., and Almirall, J.R. (2007) *Anal. Chem.*, **79**, 4365–4384.

Brooker, M.H., Sunder, S., Taylor, P., and Lopata, V.J. (1983) *Can. J. Chem.*, **61**, 494–501.

Butler, I.S. and Frost, R.L. (2006) *Appl. Spectrosc. Rev.*, **41**, 449–471.

Butler, I.S. and Gilson, D.F.R. (1997) *J. Mol. Struct.*, **408/409**, 39–45.

Buzzini, P., Massonnet, G., and Sermier, F.M. (2006) *J. Raman Spectrosc.*, **27**, 922–931.

Caillé, J.P., Pigeon-Gosselin, M., and Pézolet, M. (1983) *Biochim. Biophys. Acta*, **758**(2), 121–127.

Caillet, A., Rivoire, A., Galvan, J.M., Puel, F., and Fevotte, G. (2007) *Cryst. Struct. Des.*, **7**(10), 2080–2087.

Carey, P.R. (2006) *Annu. Rev. Phys. Chem.*, **57**(1), 527–554.

Carmona, P., Monleón, E., Monzón, M., Badiola, J.J., and Monreal, J. (2004) *Chem. Biol.*, **11**(6), 759–764.

Carter, J.C., Angel, S.M., Lawrence-Snyder, M., Scaffidi, J., Whipple, R.E., and Reynolds, J.G. (2005) *Appl. Spectrosc.*, **59**(6), 769–775.

Cheng, C., Kirkbride, T.E., Batchelder, D.N., Lacey, R.J., and Sheldon, T.G. (1995) *J. Forensic Sci.*, **1**, 31–37.

Ciomartan, D.A., Clark, R.J.H., McDonald, L.J., and Odlyha, M. (1996) *J. Chem. Soc., Dalton Trans.*, **1996**, 3639–3645.

Clegg, I.M., Everall, N.J., King, B., Melvin, H., and Norton, C. (2001) *Appl. Spectrosc.*, **55**(9), 1138–1150.

Coates, C.G., Denvir, D.J., Conroy, E., McHale, N., Thornbury, K., and Hollywood, M. (2003) *Proc. SPIE*, **4962**, 319–328.

Coleman, J.N., Khan, U., Blau, W.J., and Gun'ko, Y.K. (2006) *Carbon*, **44**(9), 1624–1652.

Colomban, P. (2002) *Adv. Eng. Mater.*, **4**(8), 535–542.

Cottingham, K. (2005) *Anal. Chem.*, **77**, 354A–357A.

Cotton, T.M., Kim, J.-H., and Chumanov, G.D. (1991) *J. Raman Spectrosc.*, **22**(12), 729–742.

Crow, P., Stone, N., Kendall, C.A., Uff, J.S., Farmer, J.A., Barr, H., and Wright, M.P. (2003) *Br. J. Cancer*, **89**(1), 106–108.

Dahan, M., Lévi, S., Luccardini, C., Rostaing, P., Riveau, B., and Triller, A. (2003) *Science*, **302**, 442–445.

De Beer, T.R.M., Baeyens, W.R.G., Ouyang, J., Vervaet, C., and Remon, J.P. (2006) *Analyst*, **131**(10), 1137–1144.

Degamber, B. and Fernando, G.F. (2002) *MRS Bull.*, **27**(5), 370–380.

Denvir, D. and Conroy, E. (2003) *Proc. SPIE*, **4877**, 55–68.

Dippel, B., Mueller, R.T., Pingsmann, A., and Schrader, B. (1998) *Biospectroscopy*, **4**(6), 403–412.

Dresselhaus, M.S., Dresselhaus, G., Jorio, A., Souza Filho, A.G., and Saito, R. (2002) *Carbon*, **40**(12), 2043–2061.

Dresselhaus, M.S., Dresselhaus, G., Saito, R., and Jorio, A. (2005) *Phys. Rep.*, **409**(2), 47–99.

Edwards, H.G.M. (2004) *Analyst*, **129**(10), 870–879.

Edwards, H., Farwell, D., and Brooke, C. (2005) *Anal. Bioanal. Chem.*, **383**(2), 312–321.

Eliasson, C., Macleod, N.A., Jayes, L.C., Clarke, F.C., Hammond, S.V., Smith, M.R., and Matousek, P. (2008) *J. Pharm. Biomed. Anal.*, **47**(2), 221–229.

Eliasson, C., Macleod, N.A., and Matousek, P. (2007) *Anal. Chem.*, **79**(21), 8185–8189.

Enejder, A.M.K., Koo, T.-W., Oh, J., Hunter, M., Sasic, S., Feld, M.S., and Horowitz, G.L. (2002) *Opt. Lett.*, **27**(22), 2004–2006.

Esposito, A.P., Talley, C.E., Huser, T., Hollars, C.W., Schaldach, C.M., and Lane, S.M. (2003) *Appl. Spectrosc.*, **57**(7), 868–871.

Everall, N.J. (2000a) *Appl. Spectrosc.*, **54**(6), 773–782.

Everall, N.J. (2000b) *Appl. Spectrosc.*, **54**(10), 1515–1520.

Everall, N., Lapham, J., Adar, F., Whitley, A., Lee, E., and Mamedov, S. (2007) *Appl. Spectrosc.*, **61**, 251–259.

Everall, N.J. and Lumsdon, J. (1991) *J. Mater. Sci.*, **26**, 5269–5274.

Falcon, J.A. and Berglund, K.A. (2004) *Cryst. Growth Des.*, **4**(3), 457–463.

Farquharson, S., Grigely, L., Khitrov, V., Smith, W., Sperry, J.F., and Fenerty, G. (2004) *J. Raman Spectrosc.*, **35**(1), 82–86.

Févotte, G. (2007) *Chem. Eng. Res. Des.*, **85**(7), 906–920.

Fleming, O. and Kazarian, S.G. (2004) *Appl. Spectrosc.*, **58**(4), 390–394.

Freeman, J.J., Wopenka, B., Silva, M.J., and Pasteris, J.D. (2001) *Calcif. Tissue Int.*, **68**(3), 156–162.

Gammon, D., Brown, S.W., Snow, E.S., Kennedy, T.A., Katzer, D.S., and Park, D. (1997) *Science*, **277**(5322), 85–88.

Gao, X. and NIe, S. (2004) *Anal. Chem.*, **76**(8), 2406–2410.

Gelder, J.D., Vandenabeele, P., Govaert, F., and Moens, L. (2005) *J. Raman Spectrosc.*, **36**, 1059–1067.

George, J.T. Jr. (2002) *Biopolymers*, **67**(4–5), 214–225.

Gibson, R.F., Ayorinde, E.O., and Wen, Y.-F. (2007) *Compos. Sci. Technol.*, **67**(1), 1–28.

Gillet, P., Biellmann, C., Reynard, B., and McMillan, P. (1993) *Phys. Chem. Miner.*, **20**(1), 1–18.

Gillet, P., Hemley, R.J., and McMillan, P.F. (1998) Vibrational properties at high pressures and temperatures, in *Ultrahigh-Pressure Mineralogy: Physics and Chemistry of the Earth's Deep Interior* (ed. R.J. Hemley), Mineralogical Society of America, Washington, DC.

Golcuk, K., Vanasse, T.M., Morris, M.D., and Goldstein, S.A. (2007) *Proc. SPIE*, **6448**, 644815.

Goodall, R.A., Hall, J., Viel, R., Agurcia, F.R., Edwards, H.G.M., and Fredericks, P.M. (2006) *J. Raman Spectrosc.*, **37**(10), 1072–1077.

Graupner, R. (2007) *J. Raman Spectrosc.*, **38**(6), 673–683.

Griffith, W.P. (1987) Advances in the Raman and infrared spectroscopy of minerals, in *Spectroscopy of Inorganic-Based Materials* (eds R.J.H. Clark and R.E. Hester), John Wiley & Sons, Inc, New York.

Haskin, L.A., Wang, A., Rockow, K.M., Joliff, B.L., Korotev, R.L., and Viskupic, K.M. (1997) *J. Geophys. Res.*, **102**, 19293–19306.

Hayward, I.P., Kirkbride, T.E., Batchelder, D.N., and Lacey, R.J. (1995) *J. Forensic Sci.*, **40**, 883–884.

Hodges, C.M., Hendra, P.J., Willis, H.A., and Farley, T. (1989) *J. Raman Spectrosc.*, **20**, 745–749.

Huang, Y.S., Karashima, T., Yamamoto, M., and Hamaguchi, H. (2005) *Biochemistry*, **44**(30), 10009–10019.

Huang, Z., McWilliams, A., Lui, H., McLean, D.I., Lam, S., and Zeng, H. (2003) *Int. J. Cancer*, **107**(6), 1047–1052.

Huong, P.V. (1986) *J. Pharm. Biomed. Anal.*, **4**, 811–823.

Janesick, J.R. (2001) *Scientific Charge-Coupled Devices*, SPIE–The International Society for Optical Engineering, Bellingham.

Jimenez, J., Wolf, I.D., and Landesman, J.P. (2003) Micro-Raman spectroscopy of semiconductors: principles and applications, in *Microprobe Characterizations of Optoelectronic Materials* (ed. J. Jimenez), Taylor & Francis Books, Inc, New York.

de Jong, B.W.D., Bakker Schut, T.C., Wolffenbuttel, K.P., Nijman, J.M., Kok, D.J., and Puppels, G.J. (2002) *J. Urol.*, **168**(4), Suppl 1, 1771–1778.

Kalasinsky, K.S., Hadfield, T., Shea, A.A., Kalasinsky, V.F., Nelson, M.P., Neiss, J., Drauch, A.J., Vanni, G.S., and Treado, P.J. (2007) *Anal. Chem.*, **79**(7), 2658–2673.

Katrin, K., Harald, K., Irving, I., Ramachandra, R.D., and Michael, S.F. (2002) *J. Phys.: Condens. Matter*, **18**, 597–624.

Kendall, C., Stone, N., Shepherd, N., Geboes, K., Warren, B., Bennett, R., and Barr, H. (2003) *J. Pathol.*, **200**(5), 602–609.

Kendix, E., Nielsen, O.F., and Christensen, M.C. (2004) *J. Raman Spectrosc.*, **35**(8–9), 796–799.

Keren, S., Zavaleta, C., Cheng, Z., de la Zerda, A., Gheysens, O., and Gambhir, S.S. (2008) *Proc. Natl. Acad. Sci. U.S.A.*, **105**(15), 5844–5849.

Ko, A.C.T., Choo-Smith, L.-Pi., Hewko, M., Sowa, M.G., Dong, C.C.S., and Cleghorn, B. (2006) *Opt. Express*, **14**(1), 203–215.

Koljenovic, S., Schut, T.C.B., van Meerbeeck, J.P., Maat, A.P.W.M., Burgers, S.A., Zondervan, P.E., Kros, J.M., and Puppels, G.J. (2004) *J. Biomed. Opt.*, **9**(6), 1187–1197.

Krafft, C. (2004) *Anal. Bioanal. Chem.*, **378**(1), 60–62.

Krafft, C., Neudert, L., Simat, T., and Salzer, R. (2005) *Spectrochim. Acta A: Mol. Biomol. Spectrosc.*, **61**(7), 1529–1535.

Laser Institute of America (2007) *American National Standard for Safe Use of Lasers (ANSI Z136.1–2007)*, Laser Institute of America, Orlando.

Lespade, P., Marchand, A., Couzi, M., and Cruege, F. (1984) *Carbon*, **22**, 375–385.

Lewis, I.R., Daniel, J., Nelson, W., and Griffiths, P.R. (1997) *Appl. Spectrosc.*, **51**(12), 1854–1867.

Lewis, I.R. and Edwards, H.G.M. (eds) (2001) *Handbook of Raman Spectroscopy*, Marcel Dekker, Inc, New York.

Lipp, E.D. and Grosse, R.L. (1998) *Appl. Spectrosc.*, **52**(1), 42–46.

Lipschutz, M.E., Wolf, S.F., Culp, F.B., and Kent, A.J.R. (2007) *Anal. Chem.*, **79**, 4249–4274.

Llinàs, A. and Goodman, J.M. (2008) *Drug Discov. Today*, **13**(5–6), 198–210.

Lyon, L.A., Keating, C.D., Fox, A.P., Baker, B.E., He, L., Nicewarner, S.R., Mulvaney, S.P., and Natan, M.J. (1998) *Anal. Chem.*, **70**, 341R–361R.

Mahadevan-Jansen, A. and Richards-Kortum, R.R. (1996) *J. Biomed. Opt.*, **1**(1), 31–70.

Maiti, N.C., Apetri, M.M., Zagorski, M.G., Carey, P.R., and Anderson, V.E. (2004) *J. Am. Chem. Soc.*, **126**(8), 2399–2408.

Maquelin, K., Kirschner, C., Choo-Smith, L.P., Ngo-Thi, N.A., van Vreeswijk, T., Stammler, M., Endtz, H.P., Bruining, H.A., Naumann, D., and Puppels, G.J. (2003) *J. Clin. Microbiol.*, **41**(1), 324–329.

Massonnet, G., Buzzini, P., Jochem, G., Staube, M., Coyle, T., Roux, C., Thomas, J., Leijenhorst, H., van Zanten, Z., Griffin, R., Wiggins, K., and Chabli, S. (2003) *Forensic Sci. Int.*, **136**, 124.

Massonnet, G., Buzzini, P., Jochem, G., Stauber, M., Coyle, T., Roux, C., Thomas, J., Leijenhorst, H., Van Zanten, Z., Wiggins, K., Russell, C., Chabli, S., and Rosengarten, A. (2005) *J. Forensic Sci.*, **50**(5), 1028–1038.

Matousek, P. (2007) *Chem. Soc. Rev.*, **36**(8), 1292–1304.

Matousek, P. and Parker, A.W. (2007) *J. Raman Spectrosc.*, **38**(5), 563–567.

McCreadie, B.R., Morris, M.D., Chen, T.-C., Sudhaker Rao, D., Finney, W.F., Widjaja, E., and Goldstein, S.A. (2006) *Bone*, **39**(6), 1190–1195.

McMillan, P. (1984) *Am. Mineral.*, **69**, 622–644.

Mendelsohn, R. and Moore, D.J. (1998) *Chem. Phys. Lipids*, **96**(1–2), 141–157.

Mizuno, A., Kitajima, H., Kawauchi, K., Muraishi, S., and Ozaki, Y. (1994) *J. Raman Spectrosc.*, **25**(1), 25–29.

Morris, M.D., Finney, W.F., Rajachar, R.M., and Kohn, D.H. (2004) *Faraday Discuss.*, **126**, 159–168.

Motz, J.T., Hunter, M., Galindo, L.H., Gardecki, J.A., Kramer, J.A., Dasari, R.R., and Feld, M.S. (2004) *Appl. Opt.*, **43**(5), 542–554.

Murphy, C.J. (2002) *Anal. Chem.*, **74**(19), 520A–526A.

Nakamoto, K. (1997) *Infrared and Raman Spectra of Inorganic and Coordination Compounds Part A: Theory and Applications in Inorganic Chemistry*, 5th edn, John Wiley & Sons, Inc, New York.

Nakano, T., Tokumura, A., and Umakoshi, Y. (2002) *Metallurg. Mater. Transact. A*, **33A**, 521–528.

Nakashima, S. (2004) *J. Phys.: Condens. Matter*, **16**(2), S25–S37.

Okumura, T. and Otsuka, M. (2005) *Pharm. Res.*, **22**(8), 1350–1357.

O'Sullivan, B., Barrett, P., Hsiao, G., Carr, A., and Glennon, B. (2003) *Org. Process Res. Dev.*, **7**(6), 977–982.

Penel, G., Delfosse, C., Descamps, M., and Leroy, G. (2005) *Bone*, **36**(5), 893–901.

Pitt, G.D., Batchelder, D.N., Bennett, R., Bormett, R.W., Hayward, I.P., Smith, B.J.E., Williams, K.P.J., Yang, Y.Y., Baldwin, K.J., and Webster, S. (2005) *IEE Proc. Sci. Meas. Tech.*, **152**(6), 241–318.

Popov, V.N. (2004) *Mater. Sci. Eng. R: Rep.*, **43**(3), 61–102.

Potgieter-Vermaak, S.S., Potgieter, J.H., and Grieken, R.V. (2006) *Cement Concr. Res.*, **36**, 656–662.

Rantanen, J., Wikström, H., Rhea, F.E., and Taylor, L.S. (2005) *Appl. Spectrosc.*, **59**(7), 942–951.

Rasha, M.K. (2006) *J. Polym. Sci., part B: Polym. Phys.*, **44**(15), 2173–2182.

Redd, D.C.B., Feng, Z.C., Yue, K.T., and Gansler, T.S. (1993) *Appl. Spectrosc.*, **47**, 787–791.

Reinecke, H., Spells, S.J., Sacristan, J., Yarwood, J., and Majangos, C. (2001) *Appl. Spectrosc.*, **55**(12), 1660–1664.

Rodrìguez-Spong, B., Price, C.P., Jayasankar, A., Matzger, A.J., and Rodrìguez-Hornedo, N. (2004) *Adv. Drug Deliv. Rev.*, **56**(3), 241–274.

Rohleder, D., Kiefer, W., and Petrich, W. (2004) *Analyst*, **129**(10), 906–911.

Rolo, A.G. and Vasilevskiy, M.I. (2007) *J. Raman Spectrosc.*, **38**(6), 618–633.

Rosch, P., Harz, M., Schmitt, M., Peschke, K.-D., Ronneberger, O., Burkhardt, H., Motzkus, H.-W., Lankers, M., Hofer, S., Thiele, H., and Popp, J. (2005) *Appl. Environ. Microbiol.*, **71**(3), 1626–1637.

Rösch, P., Harz, M., Schmitt, M., and Popp, J. (2005) *J. Raman Spectrosc.*, **36**(5), 377–379.

Schallreuter, K.U. and Wood, J.M. (2001) *J. Photochem. Photobiol. B: Biol.*, **64**(2–3), 179–184.

Scholl, J., Bonalumi, D., Vicum, L., Mazzotti, M., and Muller, M. (2006) *Cryst. Growth Des.*, **6**(4), 881–891.

Schulmerich, M.V., Finney, W.F., Fredericks, R.A., and Morris, M.D. (2006) *Appl. Spectrosc.*, **60**(2), 109–114.

Schulmerich, M.V., Morris, M.D., Vanasse, T.M., and Goldstein, S.A. (2007) Paper read at Advanced Biomedical and Clinical Diagnostic Systems V, San Jose, CA.

Schulmerich, M.V., Srinivasan, S., Cole, J.H., Kreider, J., Dooley, K.A., Goldstein, S.A., Pogue, B.W., and Morris, M.D. (2008) *J. Biomed. Opt.*, **13**, 020506.

Sharma, S.K., Angel, S.M., Ghosh, M., Hubble, H.W., and Lucey, P.G. (2002) *Appl. Spectrosc.*, **56**(6), 699–705.

Short, K.W., Carpenter, S., Freyer, J.P., and Mourant, J.R. (2005) *Biophys. J.*, **88**(6), 4274–4288.

Smith, E. and Dent, G. (2005) *Modern Raman Spectroscopy : A Practical Approach*, John Wiley & Sons, Ltd, Chichester.

Sourisseau, C. (2004) *Chem. Rev.*, **104**(9), 3851–3892.

Sourisseau, C. and Talaga, D. (2006) *Appl. Spectrosc.*, **60**(12), 1368–1376.

Sriramoju, V., Alimova, A., Chakraverty, R., Katz, A., Gayen, S.K., Larsson, L., Savage, H.E., and Alfano, R.R. (2008) Paper read at biomedical optical spectroscopy, San Jose, CA.

Stich, S., Bard, D., Gros, L., Wenz, H.W., Yarwood, J., and Williams, K. (1998) *J. Raman Spectrosc.*, **29**, 787–790.

Stuart, B.H. (1996) *Vib. Spectrosc.*, **10**(2), 79–87.

Sudworth, C.D., Archer, J.K.J., Black, R.A., and Mann, D. (2006) Paper read at biomedical vibrational spectroscopy III: advances in research and industry, San Jose, CA.

Tantipolphan, R., Rades, T., Strachan, C.J., Gordon, K.C., and Medlicott, N.J. (2006) *J. Pharm. Biomed. Anal.*, **41**(2), 476–484.

Thomas, J., Buzzini, P., Massonnet, G., and Roux, C. (2003a) *Forensic Sci. Int.*, **136**, 118.

Thomas, J., Buzzini, P., Roux, C. and Reedy, B. (2003b) *Forensic Sci. Int.*, **136**, 125.

Tomba, J.P., Carella, J.M., and Pastor, J.M. (2006) *Appl. Spectrosc.*, **60**(2), 115–121.

Tsuda, H. and Arends, J. (1994) *J. Dent. Res.*, **73**(11), 1703–1710.

van de Poll, S.W., Sweder, W.E., Römer, T.J., Tjeerd, J., Puppels, G.J., and van der Laarse, A. (2002) *J. Cardiovasc. Risk*, **9**(5), 255–261.

Vandenabeele, P., Bodé, S., Alonso, A., and Moens, L. (2005) *Spectrochim. Acta A: Mol. Biomol. Spectrosc.*, **61**(10), 2349–2356.

Vandenabeele, P., Edwards, H.G.M., and Moens, L. (2007) *Chem. Rev.*, **107**(3), 675–687.

Villar, S.E.J. and Edwards, H.G.M. (2006) *Anal. Bioanal. Chem.*, **384**, 100–113.

Voronin, M.L. (ed.) (2006) *Frontiers in Quantum Dots Research*, Nova Science Publishers, Inc, New York.

Wang, A., Haskin, L.A., and Cortez, E. (1998) *Appl. Spectrosc.*, **52**(4), 477.

Wikström, H., Lewis, I.R., and Taylor, L.S. (2005) *Appl. Spectrosc.*, **59**(7), 934–941.

Williams, A.C., Edwards, H.G.M., and Barry, B.W. (1995) *Biochim. Biophys. Acta Protein Struct. Mol. Enzymol.*, **1246**(1), 98–105.

Williams, A.C., Edwards, H.G.M., Lawson, E.E., and Barry, B.W. (2006) *J. Raman Spectrosc.*, **37**(1–3), 361–366.

Wolthuis, R., van Aken, M., Fountas, K., Robinson, J.S. Jr., Bruining, H.A., and Puppels, G.J. (2001) *Anal. Chem.*, **73**(16), 3915–3920.

Workman, J., Creasy, K.E., Doherty, S., Bond, L., Koch, M., Ullman, A., and Veltkamp, D.J. (2001) *Anal. Chem.*, **73**(12), 2705–2718.

Workman, J., Koch, M., and Veltkamp, D.J. (2003) *Anal. Chem.*, **75**(12), 2859–2876.

Young, R.J. and Eichhorn, S.J. (2007) *Polymer*, **48**(1), 2–18.

Yu, Z.Q., Chew, J.W., Chow, P.S., and Tan, R.B.H. (2007) *Chem. Eng. Res. Des.*, **85**(7), 893–905.

Yu, L.X., Lionberger, R.A., Raw, A.S., D'Costa, R., Wu, H., and Hussain, A.S. (2004) *Adv. Drug Deliv. Rev.*, **56**(3), 349–369.

Zhang, S., Pezzuti, J.A., Morris, M.D., Appawedula, A., Hsiung, C.M., Leugers, M.A., and Bank, D. (1998) *Appl. Spectrosc.*, **52**(10), 1264–1268.

Zieba-Palus, J. and Borusiewicz, R. (2006) *J. Mol. Struct.*, **792–793**, 286–292.

Further Reading

Aroca, R. (2006) *Surface-Enhanced Vibrational Spectroscopy*, John Wiley & Sons.

Chalmers, J.M. and Griffiths, P.R. (eds) (2002) *Handbook of Vibrational Spectroscopy*, John Wiley & Sons, Inc, New York.

Edwards, H.G.M. and Chalmers, J.M. (eds) (2005) *Raman Spectroscopy in Archaeology and Art History, RSC Analytical Spectroscopy Monographs*, Royal Society of Chemistry, Cambridge.

Everall, N.J., Chalmers, J.M., and Griffiths, P.R. (eds) (2007) *Vibrational Spectroscopy of Polymers: Principles and Practice*, John Wiley & Sons.

Kneipp, K., Moskovits, M., and Kneipp, H. (eds) (2006) *Surface-Enhanced Raman Scattering: Physics and Applications, Topics in Applied Physics*, Springer.

Lasch, P. and Kneipp, J. (eds) (2008) *Biomedical Vibrational Spectroscopy*, John Wiley & Sons.

Slobodan, S. (ed.) (2007) *Pharmaceutical Applications of Raman Spectroscopy, Wiley Series on Technologies for the Pharmaceutical Industry*, John Wiley & Sons.

Smith, E. and Dent, G. (2005) *Modern Raman Spectroscopy: A Practical Approach*, John Wiley & Sons.

SEISMOLOGY

(An Addendum to *Encyclopedia of Applied Physics*, Volume 17, pages 155–174.)

CARL KISSLINGER, *Cooperative Institute for Research in Environmental Sciences and Department of Geological Sciences, University of Colorado, Boulder, Colorado, U.S.A.*

Introduction 415
1. **The Seismic Moment Tensor** 416
1.1 Body Force—Dislocation Equivalence 416
1.2 Moment-Tensor Representation of Basic Fault Types 417
2. **Computing the Moment Tensor from Seismograms** 417
2.1 Ground Motion in Terms of Moment-Tensor Components ... 417
2.2 Inversion of Ground Motion for the Moment Tensor 418
3. **Complex Sources** 418
4. **The Stress Field Driving the Rupture** 419
Glossary 421
Works Cited 421
Further Reading 422

INTRODUCTION

An important advance in seismology during the past two decades has been the expansion and improvement of techniques for determining the attributes of an earthquake source from the information contained in the wave forms of the ground motion generated by that event. It is generally accepted that an earthquake is caused by the sudden slip of rocks along a geologic fault (see the original article). This mechanism can be characterized by a number of parameters: the orientation of the fault in space (dip and strike, which may vary along the fault); the geometry and dimensions of the rupture surface (length, width, area); the distribution of the slip vector specifying the motion of the rocks on one side of the fault relative to the other side (magnitude and direction, which may vary with position on the fault); the point of initiation of the rupture and the speed and direction of the subsequent rupture propagation; the orientation of the stresses driving the rupture; and some measure of the energy released.

With few exceptions, the volumes within which earthquakes occur are too deep within the Earth to be accessible to direct examination and instrumentation. The enhanced ability to derive the source parameters from seismograms has led to major advances in understanding the physics of the earthquake process and the relation of this physics to observable geological features, material properties, and Earth dynamics (Scholz, 1990). A succinct statement of the current status of source analysis from seismograms is offered by Tanioka and Ruff (1997).

In the original article, the classical method of determining basic characteristics of the focal mechanism (strike and dip of two planes, one of which is the fault; the direction of the slip vector on each of these planes; and the directions of the corresponding principal axes) is explained. The original method, still widely used, is an interpretation of the pattern of polarities of the first motion of the longitudinal (P) waves as recorded at a distribution of seismograph stations. This analysis assumes pure tangential slip on a planar surface (the fault–slip model of earthquakes) and gives the initial focal mechanism in cases where fault geometry changes during rupture. This method was supplemented by including the polarizations of the shear (S) waves (Stauder, 1960), which helped to settle an old debate about the correct body-force representation of an earthquake source.

In this era of broad-band, high–dynamic-range instrumentation (see SEISMOGRAPHS) the records of ground motion contain much more information than the arrival time and direction of vibration at the onset of the fundamental wave types. At the same time, the theory of seismic wave generation has advanced suffi-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ciently that much of this information can be extracted and analyzed in terms of the properties of the source.

A parameter that combines important properties of a simple point shear source is the *seismic moment*, the moment of one of a pair of equal but opposite orthogonal couples, i.e. there is zero net force couple (the *double couple source*; see the original article). In terms that relate this representation to the equivalent discontinuity in tangential displacement across a surface (shear dislocation), the seismic moment is given by $M_0 = \mu A D$, where μ is the shear modulus of the material in which the slip occurs. A is the area of the rupture surface, and D is the average displacement discontinuity across this surface. The seismic moment, expressed in newton-meters, is a useful way of stating the strength of an earthquake in terms of well-defined properties of the source. The introduction of the seismic moment as a measure of earthquake strength (Aki, 1966) was one step toward to utilization of a more complete description of the source, the *moment tensor*.

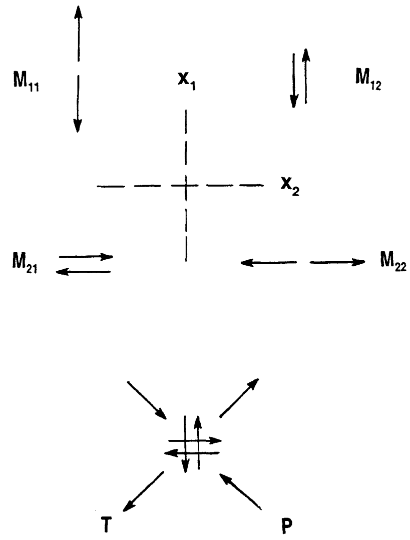

FIG. 1. (Top) Representative elements of the moment tensor, shown for the 1–2 plane only; X_3 is positive into the page. The tensor is symmetric, e.g., $M_{12} = M_{21}$, and so there is no net rotation about any axis; and the trace is usually assumed to be zero, so that there is no net volume change. (Bottom) The double couple equivalent to slip on a fault plane. If the fault plane is either the 1–3 plane or the 2–3 plane, the couples lie in the 1–2 plane. The eigenvectors of this tensor are the P and T axes of the focal-mechanism solution, the combination of dipoles without moment that generate the same movement of the surrounding medium.

1. THE SEISMIC MOMENT TENSOR

1.1 Body Force—Dislocation Equivalence

Much of the recent progress in source representation developed from papers published in the late 1950s and the 1960s (Honda, 1957; Kasahara, 1957, 1958; Steketee, 1958; Knopoff and Gilbert, 1960; Maruyama, 1963; Burridge and Knopoff, 1964). These contributions grew out of a classic work by Nakano (1923) on the forces giving rise to earthquake motion. One of the chief contributions of this era was understanding of the equivalence, for wave generation and the final static displacement and strain fields, of a jump in displacement across a surface in an elastic continuum (elastic dislocation theory; Love, 1944, pp. 221–228), and an appropriate distribution of body forces acting at the same point.

The double-couple model of fault slip is based on simplifying assumptions and represents the source as a pair of equal but opposite orthogonal couples about an axis, with no net moment. The general seismic source can be represented by body forces as a combination of double couples about three orthogonal axes (M_{ij} about the k axis) and dipoles without moment along these three axes, M_{kk}, Fig. 1 (Steketee, 1958; Aki and Richards, 1980, Chaps. 3 and 4). We further require that there is no net rotation produced at the source, so that $M_{ij} = M_{ji}$, and we usually assume no net volume change, so that $\sum M_{kk} = 0$. Debate continues as to whether there may be volume changes associated with the sources, in particular for deep (subcrustal) earthquakes.

The set of orthogonal axes may be chosen conveniently as either the set naturally associated with the fault (the normal **n** to the fault, the unit slip vector **d**, and the vector orthogonal to these), or geographic coordinates (vertical, north-south, east-west, the axes with respect to which three components of ground motion are usually measured at standard seismograph stations). The six elementary source elements are taken as the components of a symmetric tensor **M**, the *seismic moment* *tensor*. For a simple double-couple source, $\mathbf{M} = M_0(\mathbf{nd} + \mathbf{dn})$. The interchangeability of **n** and **d** in this expression shows that there are two orthogonal planes that are equivalent seismic sources, hence the inherent ambiguity of focal-mechanism solutions (see the original article).

In an arbitrary orthogonal coordinate system $\mathbf{e}_k$, $k = 1, 2, 3$, we have $\mathbf{M} = \sum\sum M_{ij}\mathbf{e}_i\mathbf{e}_j$, symmetric with zero trace. Because the moment tensor is symmetric, it can always be diagonalized, with orthogonal eigenvectors and real eigenvalues. Thus, the general source can always be represented by three orthogonal force dipoles. The directions of these dipoles give the orientations of the principal axes corresponding to the moment tensor, the so-called T (tension, minimum compression), P (pressure, maximum compression), and B (intermediate) axes. These axes can be related to the orientation of the stress field that produced the fracture by application of the theory of fracture appropriate for the medium, usually assumed in seismology to be brittle fracture, with slip governed by Coulomb friction (e.g., Jaeger, 1956, Chap. 22; also see Scholz, 1990).

1.2 Moment-Tensor Representations of Basic Fault Types

A pure double couple will be equivalent to two of these eigenvalues being equal in magnitude and opposite in sign, with the third one zero. The directions of the two dipoles (T and P axes) are at 45° to the forces forming the two couples (Fig. 1), or, equivalently, to the two planes, one of which is the fault plane. The intermediate axis (B axis) always lies in the fault plane. For a pure double-couple source in the p-q plane, $\mathbf{M} = M_{pq}\mathbf{e}_p\mathbf{e}_q - M_{pq}\mathbf{e}_q\mathbf{e}_p$ $(p \neq q)$. The seismic moment is M_{pq}. For real earthquake data, the eigenvalue with the smallest absolute value is usually much smaller than the other two, which are almost equal in magnitude (i.e., the double couple is often a good model), and the seismic moment is taken as the average of the absolute values of the larger two. Once the eigenvectors of the moment tensor are calculated, the strike, dip, and slip directions of the planes corresponding to the best double couple from the data are found by simple geometric relations (Aki and Richards, 1980, pp. 117–118).

The general moment tensor may be decomposed in a number of ways into tensors that can be interpreted in terms of the physical properties of the source (Lay and Wallace, 1995, Chap. 8). The moment tensor for faults of given orientation and mode of slipping may be calculated from the relations given in Aki and Richards (1980, p. 117). Thus, for a vertical, strike-slip fault, oriented north-south (dip = 90°, strike = 0°, slip direction = 0°)

$$\mathbf{M} = M_0 \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

while for a reverse north-south striking fault (dip = 45°, strike = 0°, slip direction = 90°)

$$\mathbf{M} = M_0 \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

2. DERIVING THE MOMENT TENSOR FROM SEISMOGRAMS

2.1 Ground Motion in Terms of Moment-Tensor Components

The components of the moment tensor are derived from the recorded ground motion by treating the seismogram as the convolution of that tensor (assumes that all tensor components have the same time history) with linear filters that represent the action at the source (source spectrum, radiation pattern), the effects of propagation through the Earth to the station for the particular wave type, and the response characteristics of the seismograph (see SEISMOGRAPHS). In modern practice, the instruments are well calibrated, and so the third filter is known. The general structure of the Earth and the relevant material properties (elastic and inelastic) are adequately determined to permit the calculation of a satisfactory representation of the propagation effects. Thus, the radiation pattern for each wave type is known, and the propagation and instrumentation effects can be accounted for, leaving the source time function and the moment-tensor components to be recovered by inversion of the signal corrected for instrument response (Langston, 1981). If the source can be treated as a point (the wavelengths involved are sufficiently longer than the source dimensions), the displacement components (instrumental effects removed) are given by the convolution (Aki and Richards, 1980, p. 53)

$$u_n(\mathbf{x}, t) = \sum\sum M_{pq} G_{np,q}(\mathbf{x}, \mathrm{t}; \xi, \tau) \qquad (1)$$

where u_n is the n component of the displacement at position $\mathbf{x}$ and time t, and the second factor is the derivative with respect to ξ_q of G_{np}, the Green's function for the system, i.e., the n component of the displacement at $(\mathbf{x},t)$ due to a unit impulse in the p direction applied at position ξ at time τ. The Green's function is calculated from the known structure and material properties in the Earth along the propagation path. In studies on local or regional scales, empirical Green's functions, the recorded wave forms generated by earthquakes small enough to be taken as delta functions in space and time, are often used.

2.2 Inversion of Ground Motion for the Moment Tensor

Equation (1) may be used to calculate synthetic seismograms corresponding to a selected Earth model. Conversely, with u_n known from the seismograms and the Green's function calculated from known Earth structure, or through use of the empirical Green's functions, Eq. (1) can be inverted to yield the moment-tensor components. The seminal paper from which many contemporary computational schemes for this purpose have been developed is Dziewonski *et al.* (1981). The Harvard *Centroid Moment Tensor* analysis was developed directly from this paper and has produced source characteristics of many thousands of earthquakes as products of a routine procedure.

The general approach to inversion for source properties is to vary the source characteristics (moment-tensor components and source depth) systematically to obtain the best fit between calculated and observed portions of the wave train at a number of seismograph stations. Many computational schemes have been developed to implement the inversion process. Some algorithms use only the *P*-wave portion of the seismograms; others use the entire body-wave train up to the arrival of surface waves; still others incorporate surface waves as well.

The moment-tensor solutions for many earthquakes strong enough to produce good seismograms at a number of stations are now calculated routinely in almost-real time and are included as part of the information included in some earthquake bulletins and catalogs. Several centers offer world-wide distribution of their solutions by e-mail within hours of an event. Summaries of the methods used in their centers, with useful lists of references, are offered by Sipkin (1986, 1994) of the U.S. Geological Survey, and Kawakatsu (1995) of the Earthquake Research Institute, Tokyo.

3. COMPLEX SOURCES

The discussion of source analysis to here has been on the simplifying assumption of a point source of seismic energy release. It is obvious from observed fault lengths of a few kilometers for moderate earthquakes to hundreds of kilometers for great earthquakes, and from examination of the often complex early portion of seismograms, that, except for the very smallest, earthquakes are not well approximated by point sources. The slip is distributed, often in a patchy fashion, over a surface, and the source time function, essentially the rate function of the seismic moment release (Tanioka and Ruff, 1997), may be quite complicated. The complexity may be due to variations in fault strength and/or stresses along the rupture surface that lead to pauses and acceleration in the slip process.

Kikuchi and Kanamori (1982) and Kikuchi and Fukao (1985) describe in detail an analysis of the source time function and spatial distribution of moment release associated with complex body waves often observed from strong earthquakes. They show that the far-field source time function is the convolution of the velocity of the relative slip across the dislocation surface and the rate of change of the rupture area:

$$S(t) = \mu \int_0^\infty \dot{D}(t - t')\dot{A}(t')\,dt' \qquad (2)$$

where $S(t)$ is the far-field source time function, $D(\xi, t)$ is the relative displacement across the dislocation surface, $t'(\xi)$ is the arrival time of the rupture front at a point ξ of the surface, $A(t)$ is the area of the dislocation at time t, and the dot represents time differentiation. The integrand $\mu\dot{D}\,dA$ is the moment rate at the grid point. The fault is modeled by assigning a seismic moment value to point double couples distributed on a grid covering the fault surface. All of these point sources are assumed to have the same focal mechanism and time history. The sum of the moments at the grid points is the total seismic moment M_0 of the event. The orientation of the fault is taken from moment-tensor analysis. The dimensions of the fault may be taken form the aftershock distribution if there is no urgency about getting the result, or may be taken as big enough to accommodate a very large earthquake (Mendoza, 1996), so that the result will yield zero moment release over large areas of this assumed fault if it is bigger than the actual fault. The moments at the point sources are determined iteratively by application of the appropriate form of Eq. (1) together with least-squares adjustment. An example showing the moment release distribution on the fault, the source time function, and the focal mechanism corresponding to the best double couple is shown in Fig. 2.

A method based on the work of Hartzell and Helmberger (1983) for rapid inversion of P waves to derive the rupture history for large earthquakes is outlined in Mendoza (1996). An example of the slip distribution for a large earthquake is shown in Fig. 3. The displacements at each station are computed as described above, and the amount of slip at each grid point is varied to give the best fit of the calculated wave forms. The process may then be repeated with a new rupture velocity to search for the velocity that produces the smallest total difference between the observed and calculated wave forms. Hartzell and Heaton (1983) reached the encouraging conclusion that the inversion results using teleseismic data alone and combining teleseismic and local strong-motion data are very similar.

4. THE STRESS FIELD DRIVING THE RUPTURE

A prime purpose of seismic source studies is the acquisition of information about the regional or local stress field that drives the fault slip. This information provides the link between dynamic geological processes and the occurrence of earthquakes. As indicated previously, the eigenvectors of the moment tensor give the directions of the principal axes corresponding to moment tensor that describes the rupture. These directions can be related by fracture theory to the true ambient stress orientation only for fracture in an infinite medium that is intact, homogeneous, and isotropic with regard to strength, with known coefficient of Coulomb friction on the fracture surface, or for slip on a pre-existing fault for which the orientation of the fault plane and the ambient stress field have not changed since the fault was formed. In the real case, surfaces of weakness exist often within rock masses, mostly due to previous fractures, and the great majority of earthquakes occur on pre-existing faults. Planes of weakness may also be the result of the processes by which the rocks were emplaced. A new fault in homogeneous material will have the expected orientation with respect to the ambient stress field at the time of the initial fracture. Because of dynamic tectonic processes, the orientation of the stress field and/or the fault surface may change with time, so that later slip on that fault may not have a simple relation to the contemporary stress field.

Failure will occur preferentially along pre-existing surfaces of weakness. The focal-mechanism solution of subsequent earthquakes will reveal the mode of faulting (strike-slip, normal, reverse, or some combination) and the quadrant within which the axes of maximum and minimum compression lie. Gephardt and Forsyth (1984) developed a technique for searching for the most probable orientation of the axes of the ambient tectonic stress field in a region in which a large number of focal mechanisms of earthquakes have been determined.

In addition to the contemporary state of tectonic stresses in a region, the evolution of this stress state and the current rate of stress changes are of fundamental importance. Deng and Sykes (1997) have calculated the alteration of the stress field in southern California due to the occurrence of 36 earthquakes since the early 19th century and have related the post-seismic static stress patterns to the occurrence of subsequent moderate to micro earthquakes. The stress patterns of old, strong earthquakes

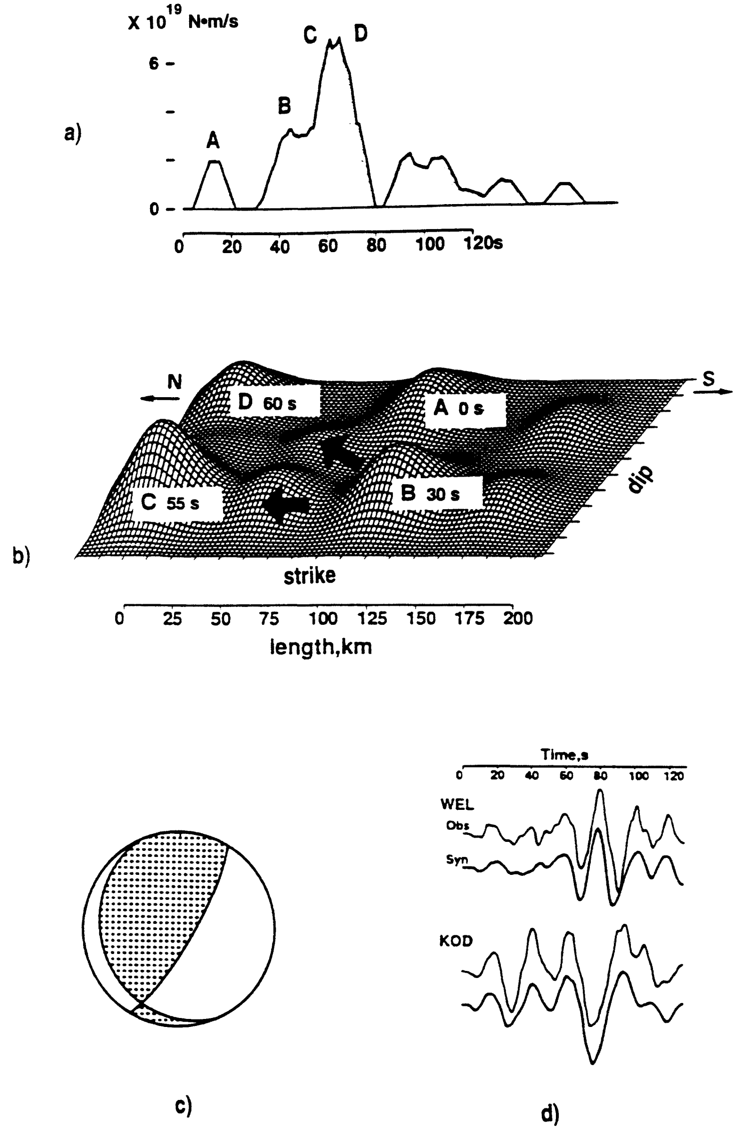

FIG. 2. Moment release in time and space for the great Tokachi-oki, Japan, earthquake of 16 May 1968. The moment magnitude was $M_w = 8.2$; total moment release as calculated by Kikuchi and Fukao (1987), 2.3×10^{21} N · m. (a) The source-time function (moment release rate) of this complex event. (b) The pattern of moment release over the fault plane. The largest subevents occurred 50 to 60 s after the origin time, and the moment release was essentially complete after 70 s. The fault plane strikes to the southeast at about 150°, and dips to the southwest at about 20°, as selected by those authors from the best double-couple solution, shown in (c). (d) Examples of the agreement between the observed wave forms and those synthesized from the final moment-tensor solution for two representative seismograph stations: Wellington, New Zealand, and Kodaikanal, India. (Courtesy of M. Kikuchi, modified by him from Fig. 5 in Kikuchi and Fukao, 1987.)

persist, and the evidence suggests that they have a major influence on the locations of later events. Gross and Kisslinger (1997) have suggested a technique based on theory by Dieterich (1994) for estimating the state and rate of tectonic stresses from the temporal and spatial behavior of aftershocks of strong earthquakes. Both of these recent studies produced results

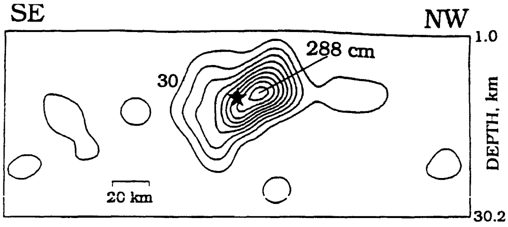

FIG. 3. The distribution of slip across the large assumed fault surface for the Guerrero, Mexico, earthquake of 14 September 1995, surface-wave magnitude 7.2, as calculated by Mendoza (1996) using his rapid finite-fault inversion procedure. The line of sight is perpendicular to the fault plane from above. Slip greater than 30 cm is contoured at 30-cm intervals. The star marks the hypocenter.

that can be compared with tectonic stress rates estimated from geodetic observations. Much research remains to be done in this emerging area of seismology.

In this Addendum, methods for recovering information about the kinematic properties of the earthquake source have been examined. The difficult problem of the dynamics of earthquake source processes is a subject of active investigation. While progress is being made, there are still major uncertainties about almost every aspect of this problem, including such fundamental questions as the appropriate constitutive relation for the fault plane. The energy budget of an earthquake is known only qualitatively, and the distribution of released elastic strain energy among processes like creation of new fracture surfaces, frictional heating of fault surfaces, and generation of propagating waves is still to be determined on a rigorous, quantitative basis.

GLOSSARY

Double-Couple Source: Body-force representation of shearing slip on a surface, consisting of two orthogonal force couples about an axis. The two couples have equal moments but opposite senses of rotation, so that the net moment is zero. This source is equivalent to two orthogonal dipoles without moment in the plane of the two couples, oriented at 45° to the forces constituting the couples, one dipole directed outward from the origin, the other inward.

Geologic Fault: Surface separating two blocks of rock, along which there has been tangential relative motion of the blocks.

Seismic Moment: A measure of earthquake strength, proportional to the area over which slip has occurred and the mean slip across that area.

Seismic Moment Tensor: A tensor, the components of which are the elementary dipoles (three without moment, six with moment) that together represent the general seismic source. The tensor is symmetric and usually assumed to have zero trace.

Source Time Function: The function expressing the rate of seismic moment release.

List of Works Cited

Aki, K. (1966), *Bull. Earthquake Res. Ins. Univ. Tokyo* **44**, 23–88.

Aki, K, Richards, P. G. (1980), *Quantitative Seismology*, San Francisco: W. H. Freeman.

Burridge, R, Knopoff, L. (1964), *Bull. Seismol. Soc. Am.* **54**, 1875–1888.

Deng, J., Sykes, L. R. (1997), *J. Geophys. Res.* **102**, 24411–24435.

Dieterich, J. H. (1994), *J. Geophys. Res.* **99**, 2601–2618.

Dziewonski, A. M., Chou, T.-A. Woodhouse, J. H. (1981), *J. Geophys. Res.* **86**, 2825–2852.

Gephardt, J. W., Forsyth, D. W. (1984), *J. Geophys. Res.* **89**, 9305–9320.

Gross, S., Kisslinger, C. (1997), *J. Geophys. Res.* **102**, 7603–7612.

Hartzell, S. H., Heaton, T. H. (1983), *Bull. Seismol. Soc. Am.* **73**, 1553–1583.

Honda, H. (1957), in: J. H. Hodgson (Ed.), *The Mechanics of Faulting, with Special Reference to the Fault-Plane Work*, Publications of the Dominion Observatory Ottawa Vol. 22, Ottawa: HMSO, pp. 295–340.

Jaeger, J. C. (1956), *Elasticity, Fracture and Flow*, London: Methuen.

Kasahara, K. (1957), *Bull. Earthquake Res. Inst. Univ. Tokyo* **35**, 473–532.

Kasahara, K. (1958), *Bull. Earthquake Res. Inst. Univ. Tokyo* **36**, 21–53.

Kawakatsu, H. (1995), *Geophys. Res. Lett.* **22**, 2569–2572.

Kikuchi, M., Kanamori, H. (1982), *Bull. Seismol. Soc. Am.* **72**, 491–506.

Kikuchi, M., Fukao, Y. (1985), *Phys. Earth Plan. Inter.* **37**, 235–248.

Kikuchi, M, Fukao, Y. (1987), *Tectonophysics* **144**, 231–247.

Knopoff, L., Gilbert, F. (1960), *Bull. Seismol. Soc. Am.* **50**, 117–134.

Langston, C. (1981), *Bull. Seismol. Soc. Am.* **71**, 1–24.

Lay, T., Wallace, T. C. (1995), *Modern Global Seismology*, San Diego: Academic Press.

Love, A. E. H. (1944), *A Treatise on the Mathematical Theory of Elasticity*, New York: Dover Publications.

Maruyama, T. (1963), *Bull. Earthquake Res. Inst. Univ. Tokyo* **41**, 467–486.

Mendoza, C. (1996), *Seismol. Res. Lett.* **67** (6), 19–26.

Nakano, H. (1923), *Seismol. Bull. Center Meteor. Obs. Jpn.* **1**, 92–120.

Scholz, C. H. (1990), *The Mechanics of Earthquakes and Faulting*, Cambridge, U.K.: Cambridge University Press.

Sipkin, S. (1986), *Bull. Seismol. Soc. Am.* **76**, 1515–1541.

Sipkin, S. (1994), *Geophys. Res. Lett.* **21**, 1667–1670.

Stauder, W. (1960), in: J. H. Hodgson Ed. *A Symposium of Earthquake Mechanism*, Publications of the Dominion Observatory Ottawa Vol. 24, Ottawa: HMSO, pp. 343–353.

Steketee, J. A. (1958), *Can. J. Phys.* **36**, 1168–1197.

Tanioka, Y., Puff, L. J. (1997), *Seismol. Res. Lett.* **68**, 386–397.

Further Reading

Droste, S, Teisseyre, R. (1959), *Sci. Rep. Tohoku Univ. Ser. 5, Geophysics* **11**, 55–71.

Knopoff, L (1960), *Bull. Seismol. Soc. Am.* **50**, 117–134.

Knopoff, L., Randall, M. J. (1970), *J. Geophys. Res.* **75**, 4957–4963.

Kostrov, B. V., Das, S. (1988), *Principles of Earthquake Source Mechanics*, Cambridge, U.K.: Cambridge University Press.

Scholz, C. H., Mikumo, T. (Eds.) (1991), special issue on "Modelling the Earthquake Source," *Tectonophysics* **197**.

CAPILLARY ELECTROPHORESIS

Curtis D. Mowry and Michael J. Kelly, *Sandia National Laboratories, Albuquerque, New Mexico, U.S.A.*

Introduction 3
1. **Basic Concepts** 4
1.1 Electrophoresis 4
1.1.1 Electrophoretic Mobility 5
1.1.2 Electroosmotic Flow 6
1.1.3 Separation in Capillary Electrophoresis 6
1.2 Resolution 7
1.3 Qualitative and Quantitative Analysis 9
2. **Instrumentation** 10
2.1 Sample Introduction 10
2.2 Capillaries 12
2.3 Buffers and Other Reagents 13
2.4 High-Voltage Power Supplies 14
2.5 Detectors 14
2.5.1 Ultraviolet–Visible Absorbance 15
2.5.2 Fluorescence 15
2.5.3 Mass-Spectrometric 16
2.5.4 Electrochemical 17
2.5.5 Other Detectors 18
2.6 Integrated Capillary Electrophoresis Microdevices 18
3. **Modes of Separation** 18
3.1 Capillary Zone Electrophoresis (CZE) 18
3.2 Capillary Isoelectric Focusing (CIEF) 22
3.3 Capillary Gel Electrophoresis (CGE) 22
3.4 Micellar Electrokinetic Capillary Chromatography (MECC) 25
3.5 Capillary Isotachophoresis (CITP) 27
3.6 Capillary Electrochromatography (CEC) 28
Acknowledgments 30
Glossary 30
Works Cited 32
Further Reading 32

INTRODUCTION

In 1809, the Russian physicist Reuss performed the first recorded experiment of what would come to be known as electrophoresis (Reuss, 1809). Electrophoresis is defined as the migration, or movement, of electrically charged particles or dissolved species (solutes) in a conductive liquid or gel under the influence of an applied electric field. Reuss observed the electrophoretic migration of colloidal clay particles at an anode placed in a glass tube containing a bottom layer of wet clay, a second layer of wet sand, and a top layer of water. At the cathode, constructed in a similar fashion and ionically connected to the anode by a glass tube filled with water, the water remained clear and increased in volume, demonstrating the principle of electroosmotic flow (EOF). Also known as electroendosmotic flow or electroendosmosis, EOF is defined as the transport of neutral molecules or particles by a conducting solution under the influence of an electric field.

Numerous other experiments advanced the science of electrophoresis. Quincke (1861) showed that the migration rate of particles in an electric field is a linear function of potential gradient, while Hardy (1899) demonstrated the principle of the isoelectric point, which is defined as the *p*H at which a zwitterionic solute has zero electrophoretic mobility in an electric field. It was Michaelis (1909) who actually coined the phrase "electrophoresis" to describe the migration of ionic species in electric fields, and determined the isoelectric points of some enzymes in a U-shaped electrophoresis cell by measurements of electrophoretic mobilities at various *p*H values. Tiselius (1937) performed free-solution electrophoresis for protein separations and was the first to show that serum is composed of albumin and α-, β-, and γ-globulins, for which he won the Nobel prize in 1948. In Tiselius's experiment, the arms of a U-tube cell were filled with a *p*H buffer and a small amount of the protein under study, and then connected to nonpolarizable electrodes. The motion of the protein–buffer boundaries

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

was observed under the influence of a known potential gradient using a schlieren optical detection method to observe the boundaries between proteins.

Despite some significant instrumental improvements in electrophoresis experiments conducted with U-tube cells or on rectangular-shaped gel slabs, electrophoresis techniques were cumbersome and time-consuming, and produced separations with somewhat poor resolution among the separated components. Dramatic improvements were on the horizon, beginning in 1967 with the construction of the first experimental capillary electrophoresis (CE) system by Hjerten (1967), who then used the instrument to achieve high-efficiency separations of proteins, nucleic acids, and inorganic ions. In CE, the glass tubes or gel slabs used previously to separate solutes based on their electrophoretic mobilities were replaced by narrow-bore capillaries as the separation device. In 1979, Mikkers *et al.* (1979) performed electrophoresis in 200-µm inside diameter (i.d.) capillaries, and Jorgenson and Lukacs began electrophoresis experiments in 75-µm i.d. glass capillaries that resulted in seminal papers (Jorgenson and Lukacs, 1981, 1983) that showed the full potential of the technique for high-resolution separations. An intense period of research and development on CE science, separation methods, and instrumentation ensued, culminating in the introduction of the first commercial instrument in 1988. The first international scientific symposium devoted to CE was held in 1989, and the field has experienced exponential growth to the present day.

Today, capillary electrophoresis is one of the fastest growing techniques in analytical chemistry. Approximately 1500 scientific papers are published each year on CE, covering not only advances in instrumentation and methodology, but also applications to important problems in a variety of scientific disciplines. Among the important applications of CE are high-resolution separations in life science [DNA, polymerase chain reaction (PCR) products, oligonucleotides, purines, proteins, polypeptides, amino acids, carbohydrates, catecholamines, vitamins, chiral isomers, pharmaceuticals], environmental science (polyaromatic hydrocarbons, pesticides, pollutants), analytical chemistry (polymers, inorganic and organic ions, organic acids and bases, dyes and their precursors) and forensic and anti-terrorist science (illicit drugs, gunshot residues, DNA, explosive residues). Clearly, it is not possible to do justice to all areas of such a large field in a short article such as this. Therefore, in the sections that follow, the emphasis is on the most important theoretical, chemical, and procedural aspects of capillary electrophoresis.

The outline of the remainder of the article is as follows. The first section is devoted to a description of basic concepts that apply to all embodiments of capillary electrophoresis. Next, the instrumentation section describes the five essential components of a CE instrument, also known as a capillary electropherograph: the sample introduction system, the capillary (where the separation of complex mixtures into individual components occurs), the buffer system, the high-voltage power supply, and the detector. In the section entitled "Modes of Separation", a discussion of separation principles (*i.e.*, the physical and chemical mechanisms for resolving a mixture into its individual components) is combined with a description of diverse applications from the current scientific literature to illustrate the usefulness and limitations of the different CE separation techniques.

1. BASIC CONCEPTS

1.1 Electrophoresis

The objective in an electrophoresis experiment is to move individual solutes in a mixture through a separation medium at different velocities so that they exit from the medium at different times or, alternatively, reach different locations on the medium at a given time. In one popular embodiment of conventional electrophoresis, the medium is a gel slab either in the form of a free-standing membrane or as a gel supported on a planar stationary support, *e.g.*, a rectangular glass plate. Typical gels used in this application are two-phase colloidal systems (paper, starch, cellulose acetate, agarose) in an aqueous solvent or polymeric materials such as polyacrylamide. The slab is immersed in a buffer solution in an electrophoresis chamber and an electrical potential is applied across the length of the gel to set up an electric field (*i.e.*, a gradient of electrical potential). At the end of the experiment, the individual components will have migrated to different points along the direction of the gradient by processes that are described in Sec. 3. The locations of the components on the slab can be determined by a number of methods, including staining (followed by optical detection), radiography (for radioactive or radioisotopically tagged solutes), or detection with immunoreagents. This electrophoresis method is slow, labor intensive, not amenable to full automation, and prone to relatively poor reproducibility. The last results from a number of factors, including difficult-to-control solute adsorption effects on the gel and thermal contributions to solute migration (due to poor dissipation of Joule heat) that vary across the slab and result in broadening of the zones of the separated components. Improvements in the separation of mixtures are achieved by two-dimensional techniques, where successive electrophoresis experiments are performed on the same gel slab in perpendicular directions *via* two different separation modes.

Despite the problems described above, gel slab electrophoresis is an extremely useful and versatile method that is widely used today, principally in the separation and identification of complex mixtures of biological macromolecules. However, these problems have been addressed by capillary electrophoresis, and dramatic improvements in separations have been achieved. CE grew out of a remarkable convergence of scientific and technological advances in analytical chemistry. The principal advance came in the recognition that performing electrophoretic separations in small i.d., open-tubular capillaries, such as those used in gas chromatography (GC; see CHROMATOGRAPHY), reduces the problems associated with adsorption on gel media and thermally induced convective dispersion (Mikkers *et al.*, 1979). Separations in small i.d. capillaries also increases the degree of interaction between solutes and the capillary surface, which we will see has a large effect on separation efficiency. Further, detectors developed for chromatography instrumentation (principally liquid chromatography, LC) could be adapted, with certain modifications, to automatic detection of components eluting from the electrophoresis capillary. With these advances, it became possible to develop extremely fast, high-resolution separations for small quantities of materials. Capillary electrophoresis complements chromatography because separation selectivities are based on different chemical and physical principles and because it provides high-resolution separations of materials that are not amenable to chromatographic separations (*e.g.*, some large biological macromolecules). In fact, one approach for certain difficult samples takes advantage of their complementary nature by coupling CE to LC in a serial fashion to get two-dimensional separations with resolutions that exceed the capabilities of either method alone.

CE has a number of dramatic advantages compared with LC. These include faster analyses, higher efficiencies (defined in Sec. 1.3), smaller sample size requirements ($\leq \approx 10\,\mathrm{nL}$ *versus* $\approx 10\,\mu\mathrm{L}$), greater mass sensitivity (attomole *versus* picomole), lower flow rates ($\approx 100\,\mathrm{nL\,min^{-1}}$ *versus* $\approx 1\,\mathrm{mL\,min^{-1}}$), which reduce reagent consumption, and fewer column inventory requirements (one CE capillary suits a variety of samples so that different columns for different sample types are not required as in LC). These are among the key factors that account for the increasing popularity of CE for separations in liquid media.

1.1.1 Electrophoretic Mobility Electromigration is defined as the movement of a charged species under the influence of an electric field. In electrophoresis, however, it is more generally used to describe the movement of all solutes, whether they be neutral or charged. In this article, the term migration is used for brevity. In the case of a charged solute, a major contributor to migration results from its electrophoretic mobility, μ_{EP}, defined as

$$\mu_{\mathrm{EP}} = q/6\pi\eta r, \tag{1}$$

where q is the solute's electrical charge, η is the viscosity of the medium through which the solute is migrating, and r is the solute radius. Electrophoretic mobility is essentially a measure of the velocity of a charged species in an electric field of unit strength, in units of $\mathrm{cm^2\,V^{-1}\,s^{-1}}$. The linear dependence of μ_{EP} on the charge/size ratio is exploited in many modes of CE separations. Further, η is a strong function of such experimental parameters as *p*H, temperature, and ionic strength (a quantity related to the sum of the products of concentration and squared charge for all the ionic species in solution, expressed in terms of concentration units, *e.g.*, molarity). This provides additional experimental control of μ_{EP} in capillary electrophoresis. The effect of some parameters can be predicted *a priori*, *e.g.*, the dependence of the electrophoretic mobility of a solute with a known acid dissociation constant (K_a) as a function of *p*H. In the case of a weak acid, for example, the equilibrium concentration of the charged (unprotonated) and neutral (protonated) forms of the solute as a function of *p*H will control its mobility.

1.1.2 Electroosmotic Flow As described earlier, electroosmotic flow (EOF) is the transport of neutral molecules by a conducting solution under the influence of an electric field. In capillary electrophoresis, EOF results directly from the physical and chemical properties of the materials and chemicals used in the experiment. In the most common CE configuration, a fused-silica capillary is used with an aqueous buffer at *p*H >3 (see Secs. 2.2 and 2.3). In this case, the surface of the capillary is composed of ionized silanoate groups ($\mathrm{SiO^-}$). In the buffer, the layer closest to the capillary wall contains specifically adsorbed (tightly held) buffer cations, while a second layer consists of nonspecifically adsorbed (loosely held), hydrated buffer cations. This arrangement is known as an electrical double layer, and the electrical potential difference across the shear plane between these layers is termed the zeta potential (ζ). Under the influence of an applied electric field, the loosely held cations migrate toward the negatively charged cathode and drag other buffer constituents (water, cations, anions, solutes) along with them. This is known as EOF, and it results in the transport of neutral molecules and cations in the direction of the cathode. In addition, even though anions have a migration component in the direction of the anode due to their electrophoretic mobilities, they too will be transported toward to the cathode if the electroosmotic mobility (defined below) is greater than their electrophoretic mobility. This is typically the case at alkaline *p*H for most solutes of interest.

The electroosmotic mobility μ_{EOF} is defined as

$$\mu_{\mathrm{EOF}} = \varepsilon\zeta/4\pi\eta, \qquad (2)$$

where ε is the dielectric constant of the buffer and ζ is the zeta potential. Electroosmotic mobility is essentially a measure of the velocity of a neutral species in an electric field of unit strength, in units of $\mathrm{cm^2\,V^{-1}\,s^{-1}}$.

There are CE applications where it is desirable to eliminate EOF or reverse its direction. Coating the capillary with a hydrophobic material that prevents access of the buffer to the surface silanoate groups would eliminate the electrical double layer described above that is a requirement for EOF. In this ideal case, there is no EOF, and only charged species in the capillary will migrate in the electric field, and they will do so according to their electrophoretic mobilities. The simplest way to reverse EOF is to reverse the polarity of the two electrodes; however, this does not reverse the relative directions of EOF and the electrophoretic mobility of a given solute. To accomplish this type of EOF reversal, a flow modifier such as cetyltrimethylammonium bromide (CTAB) or cetyltrimethylammonium chloride (CTAC) is added to the buffer at concentrations above a threshold required to effect flow reversal. These molecules attach to the silanoates on the capillary in a manner that creates a modified capillary surface with a net positive charge. The electrical double layer is then composed of anions and hydrated anions from the buffer, in contrast to the example described earlier, and EOF will be in the direction of the anode. In general, the magnitude of μ_{EOF} is dependent on experimental parameters such as *p*H, temperature, ionic strength, and viscosity. In special cases (*e.g.*, ampholytic flow modifiers, see Sec. 3.2), μ_{EOF} magnitude and EOF direction are both strong functions of *p*H.

1.1.3 Separation in Capillary Electrophoresis In a capillary electrophoresis separation, the net mobility of a solute being analyzed (an analyte) is the sum of its electrophoretic and electroosmotic mobilities. The velocity of the analyte (in $\mathrm{cm\,s^{-1}}$) through the capillary is simply the product of the mobility and the electric field strength. Therefore,

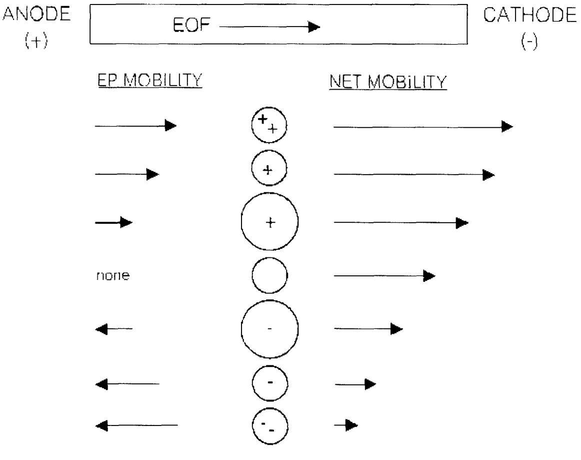

FIG. 1. Schematic representation of the electrophoretic migration of a mixture of anions (−), cations (+), and neutral molecules in capillary electrophoresis. The sizes of the molecules are represented by the sizes of the circles. The magnitude of the ionic charge is represented by the number of + and − symbols in the circles. In this example, the mixture was introduced at the anode end of the capillary and the direction of EOF is towards the cathode. The net mobility of a species is the sum of its electrophoretic and electroosmotic mobilities, and the magnitude and direction of each of these are indicated by the length and direction of the arrows, respectively.

the fastest separations can be performed with the highest field strengths. In practice, however, separations are performed at moderate field strengths to reduce Joule heating of the buffer, thereby preventing such undesirable consequences as vaporization of the buffer and thermal degradation of temperature-sensitive samples.

In the absence of specific analyte interactions with the capillary (*e.g.*, adsorption on silanoate groups), with materials added to the capillary (*e.g.*, a gel or a chromatographic stationary phase), or with a buffer modifier, a mixture of analytes will exit from the capillary (elute) at the cathode in the order of their net mobility. As an example, consider the electrophoretic separation of a mixture of anions and cations (of a variety of sizes and charges) and neutral molecules in an uncoated capillary. The analytes will elute at the cathode in the following order (see Fig. 1):

1. small, multiply charged cations;
2. small, singly charged cations;
3. large, singly charged cations;
4. neutral molecules (unresolved, *i.e.*, all elute at the same time);
5. large, singly charged anions;
6. small, singly charged anions;
7. small, multiply charged anions.

If the mixture being analyzed contains only anions, the direction of EOF can be reversed so that EOF and anion electrophoretic mobility are in the same direction. This causes the anions to migrate through the capillary more quickly and results in a faster separation. In this case, small, multiply charged anions would elute prior to large, singly charged anions.

In Sec. 3, a variety of CE separation modes that include special capillary and buffer arrangements (designed to promote specific analyte interactions with the capillary or the buffer modifier) are described. It will be seen that additional levels of selectivity in CE separations can be added to electrophoretic mobility and electroosmotic flow.

1.2 Resolution

The objective in a capillary electrophoresis experiment is to move individual analytes in a mixture through the capillary at different velocities so that they elute at different times. The analytes then move into an on-line detector that provides a real-time plot of detector response as a function of migration time. This plot is known as an electropherogram, and is a direct analog of the chromatogram (see CHROMATOGRAPHY).

The statistical treatment of electropherograms and that of chromatograms to determine important qualities of separations (efficiency, resolution, separation factor, and capacity factor) are fairly similar. The statistical methods for modes of CE that involve partitioning of the analyte between the aqueous phase and another phase (gel, micelle, chromatographic stationary phase, *etc.*) are strictly analogous to those used for chromato-

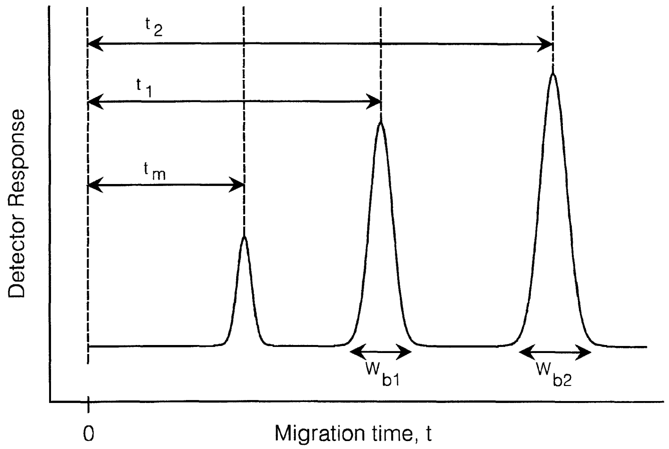

FIG. 2. Hypothetical electropherogram showing the response of a detector as a function of migration time t. The migration times for analyte 1, analyte 2, and a neutral marker are t_1, t_2, and t_m, respectively. Electrophoretic peak with is W_b.

graphic data (see CHROMATOGRAPHY), and are not included in this article. However, the treatment for modes of CE based on electrophoretic and electroosmotic mobilities is briefly discussed here.

The hypothetical electropherogram shown in Fig. 2 displays the response of a detector as a function of migration time t. The sample (composed of analytes 1 and 2) is introduced into the capillary at time zero as a narrow zone. The molecules of analytes 1 and 2 migrate through the capillary and elute at their characteristic migration times t_1 and t_2. As a way to normalize the time axis of the electropherogram, the elution time t_m of a "neutral marker", a molecule that is transported only by EOF, is used. While the components in the sample are being separated and eluted, the once-narrow zones broaden. This is primarily the result of diffusion under the influence of both temperature and concentration gradients. As a result of this zone spreading, the components elute as Gaussian-like peaks rather than as sharp bands. The degree of spreading is indicated by the magnitude of the peak width W_b. A fundamental chromatographic parameter that encompasses migration time and band spreading is the number of theoretical plates N:

$$N = 16(t_l/W_{bl})^2. \tag{3}$$

In terms of electrophoretic variables, N is expressed as

$$N = (\mu_{\mathrm{EPl}} + \mu_{\mathrm{EOF}})V/2D_l, \tag{4}$$

where V is the applied electrical potential (commonly referred to as the applied voltage) and D_l is the diffusion coefficient of the analyte. Theoretical plates were first used to describe separation mechanisms in distillation processes, and were later adapted to chromatographic and electrophoretic separations. Briefly, a theoretical plate is an imaginary unit in the capillary wherein one complete mechanistic separation process step occurs. For example, in micellar electrokinetic capillary chromatography (MECC, Sec. 3.4), this would be a single equilibrium partitioning of the analyte between the micellar phase and the aqueous phase. In the entire column, there are N of these imaginary units. CE practitioners routinely calculate N values from electropherograms because it is a good indication of separation quality and the ability of a method to provide narrow peaks (high values of N being more desirable). This measure of CE efficiency is frequently normalized to the capillary length L to give the height equivalent of a theoretical plate,

$$H = L/N. \tag{5}$$

In the CZE of large biomolecules, wall adsorption can adversely affect separation efficiency, and models have been developed to predict its effects on plate height (Minarik *et al.*, 1995).

Other fundamental parameters are used as measures of the quality of a separation scheme. The separation factor α is a measure of the ability of a particular separation scheme to separate two components, and is defined in terms of electrophoretic variables as follows (Karger and Foret, 1993):

$$\alpha = \mu_{\mathrm{EP1}}/\mu_{\mathrm{EP2}}. \tag{6}$$

Therefore, the separation of two analytes depends on their relative electrophoretic mobilities. The extent of separation depends on the difference in migration times and the sharpness of the peaks, and is commonly described in terms of the resolution R:

$$R = 2(t_2 - t_1)/(W_{\mathrm{b1}} + W_{\mathrm{b2}}). \tag{7}$$

Narrow peaks with very different migration times are said to be separated with high resolution. A well-resolved peak arising from a pure substance is very desirable from both an analytical and a preparative viewpoint. The resolution can also be expressed in terms of electrophoretic variables (Jorgenson and Lukacs, 1983):

$$R = 0.177(\mu_{\mathrm{EP2}} - \mu_{\mathrm{EP1}})\{\mathrm{V}/[(\mu_{\mathrm{EP}} + \mu_{\mathrm{EOF}})\mathrm{D}]\}^{1/2}, \tag{8}$$

where μ_{EP} is the average electrophoretic mobility and D the average diffusion coefficient of the two analytes. It can be seen that resolution increases as $\mu_{\mathrm{EP}} + \mu_{\mathrm{EOF}}$ approaches zero, but this requires increasingly longer separation times. For example, if μ_{EP} equaled $-\mu_{\mathrm{EOF}}$, resolution would be infinite; however, the solutes would be moving at equal but opposite velocity to the EOF and therefore would never elute. Therefore, a compromise between resolution and analysis time is usually required.

CE separations are typically faster and more efficient than LC separations. This advantage arises in great part from the characteristics of EOF and the fast migration times. Capillary EOF is characterized by a flat flow profile across the cross-sectional diameter of the capillary, with smaller flow rates at locations only in extremely close proximity to the capillary wall. This flat flow profile results in a narrow electrophoretic zone for a given solute, and gives narrow, well-resolved peaks. In contrast, pumped flow in LC is characterized by the parabolic profile of laminar flow, with the flow velocity at the column walls much smaller than the flow at the center of the column. This results in a broadening of the chromatographic zone for a given solute, and gives rise to broader peaks with lower resolution compared with CE.

CE separations with $N = 400\,000$–$2\,700\,000$ have been reported on 80–100-cm-long capillaries for separations that require only a few minutes. This is approximately 100 times better than the efficiencies of LC separations (as defined by the number of theoretical plates, N), which typically take about 10 times as long. For comparison, CE efficiencies are also approximately 2–10 times better than that of capillary GC separations. CE has the advantage that, with minimized injection and detection volumes and properly dissipated Joule heat, the largest contributor to band broadening is longitudinal diffusion of the analyte. The effects of longitudinal diffusion are reduced by keeping analysis times short.

A fundamental knowledge of the physical and chemical processes occurring in the capillary is critical to the successful design and implementation of a separation. The experimental variables that control N, α, and R are different for the various modes of CE, and are discussed in Sec. 3.

1.3 Qualitative and Quantitative Analysis

Qualitative analysis is the identification of a component responsible for a peak in the electropherogram. It is not possible to predict migration times exactly on the basis solely of the characteristics of the sample and the CE separation scheme. Several empirical expressions have been developed to calculate effective mobilities better on the basis of operational parameters. One such expression allows the effective mobilities of singly and multiply charged ions to be predicted over a wide range of buffer ionic strength (Friedl *et al.*, 1995). However, from an operational standpoint, the ability to predict migration times is generally not required. Correct peak assignments for an unknown analyte can be made by comparison with a literature electropherogram run under identical conditions, but this is not a very reliable approach. More reliable peak assignments are made by comparison with an electropherogram from injection of a known standard, by the use of an internal standard, or by identification using an on-line or off-line analytical method. Identification of unknown materials can sometimes be done on line with detectors based on ultraviolet–visible absorbance, fluorescence, mass spectroscopy, or electrochemistry (Sec. 2.5). Alternatively, it can be done by collection of effluent fractions (corresponding to individual components) followed by off-line spectroscopic, chemical, or biochemical analysis.

Quantitative analysis of analyte concentration is usually based on measurement of the height or area of the electrophoretic peak and comparison with peaks resulting from CE of a known standard. Modern CE instrumentation has data-acquisition and -processing equipment (*e.g.*, computers with integration and calibration software) for this purpose. However, CE peak areas are a function not only of analyte concentration, but also of analyte migration time. Since the analyte zone velocity in CE is a function of both μ_{EP} and μ_{EOF}, the separated components do not all migrate to the detector at the same velocity. The peak areas for two components of equal concentration and detector response will therefore not be equal, as the later eluting component travels more slowly through the detector. In contrast, for high-performance liquid chromatography (HPLC) the later eluting peak would be shorter and broader, but the two peak areas would be the same (see CHROMATOGRAPHY). This velocity effect in CE must be factored into the calibration calculations if the goal is the determination of analyte concentrations based on peak areas. High accuracy and precision in quantitative analysis by CE requires precise control of many factors, including EOF, capillary surface chemistry, temperature, buffer composition (*p*H, ionic strength, concentration of modifiers), and sample introduction volumes.

2. INSTRUMENTATION

Modern CE instrumentation (see schematic diagram in Fig. 3) has reached a very sophisticated state of development. It is highly automated, with computer control of the sample introduction, separation, and detection parameters and of the data acquisition and analysis. Special software for issues specific to CE separations, such as shifting migration times, correction of peak areas for analyte velocities, and estimation of analyte isoelectric points and molecular weights is also commercially available. At present, there are more than ten suppliers of capillary electropherographs in the U.S.

2.1 Sample Introduction

There are several sample introduction methods commonly used in capillary electrophoresis. The specific method used depends upon parameters such as separation mode and sample matrix. In each method, however, the inlet end of the capillary is physically moved from the buffer inlet vial to a sample vial for sample introduction, then returned to the inlet vial for electrophoresis. The length of the sample plug is usually more important than the actual volume, and only picoliters or nanoliters of sample solution are actually injected or consumed. Such a small volume requires care in handing, and it is typically contained in microliter or milliliter sample vials. The sample is often dissolved in the same buffer solution used during CE separation (Sec. 2.3). The sample concentration should be significantly lower than the buffer concentration, since concentrated samples may distort the electric field within the capillary, reducing separation efficiency and resolution. Sample introduction methods can be divided into three types: hydrodynamic, electrokinetic, and sample concentration.

Hydrodynamic sample introduction methods use flow generated by pressure, vacuum, or gravity to transport the sample solution into the capillary. In all cases, important factors for reproducible injections include buffer viscosity, sample viscosity, capillary radius, capillary length, and injection time. Sample temperature is also important because it affects buffer and sample viscosities. When these factors are reproducibly controlled, the volume of sample introduced is reproducible even if solution composition or *p*H is not well controlled.

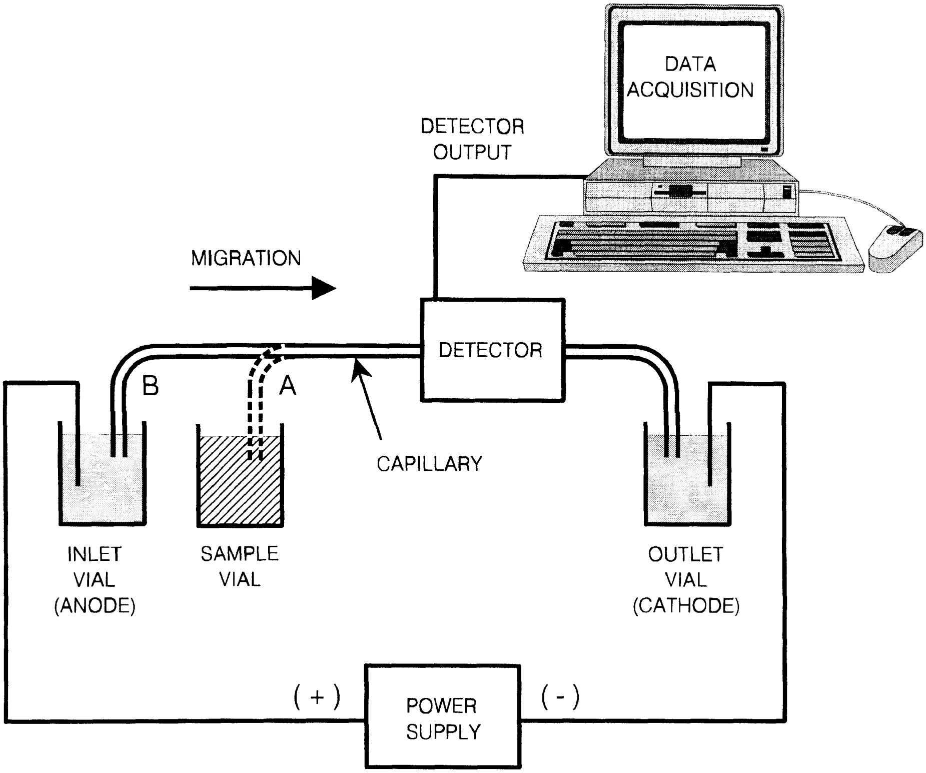

FIG. 3. Schematic diagram of capillary electrophoresis instrument, also known as a capillary electropherograph. During sample introduction (A), the capillary is immersed in the sample vial. During electrophoresis (B), the capillary is immersed in the inlet vial. This schematic diagram shows the use of an on-column detector (see Sec. 2.5).

Hydrodynamic methods require more instrumentation than electrokinetic methods and do not work well with packed capillaries (*e.g.*, capillary gel electrophoresis, CGE), yet have the advantages that sampling bias is not created and matrix effects are minimal. For aqueous sample solutions, the sample plug volume or length can be easily calculated from the experimental parameters. Pressure-based hydrodynamic injection requires that pressure be applied to the sample vial or, alternatively, a vacuum applied at the detector side of the capillary. Pressure leaks affect reproducibility, and the time period that pressure is applied must be carefully controlled. This type of sample introduction works particularly well with capillary isoelectric focusing (CIEF, Sec. 3.2), since the entire capillary can be filled in a short time. Gravity-based hydrodynamic injection, also known as siphoning injection, is performed by inserting the inlet end of the capillary into the sample vial and then raising this vial above the capillary outlet vial. The height and time required are on the order of millimeters and seconds, both of which must be carefully controlled, yet this type of sample introduction remains instrumentally simpler than pressure-based methods. Electrophoresis is initiated after the sample has been introduced and the capillary is returned to the inlet vial. It is important to keep the inlet and outlet vials at the same level to prevent siphoning during electrophoresis, which would induce laminar flow in the capillary and resultant degradation of separation efficiency and resolution.

Electrokinetic (or electromigration) sample introduction uses an electric field to transport analytes into the capillary. The amount of analyte injected depends upon the applied voltage, capillary radius, μ_{EOF}, analyte concentration, analyte μ_{EP}, viscosity, and injection time. Sample solution temperature, buffer temperature, conductivity, and *p*H will affect μ_{EOF} and μ_{EP} and therefore injection reproducibility. In some cases, it is necessary to adjust the ionic strength of the sample to obtain good injection precision. Electrokinetic introduction is instrumentally simpler than pressure-based methods, as the high-voltage supplies and controls are already in place on the instrument. The main disadvantage of electrokinetic introduction is that because analyte transport onto the column is based upon both μ_{EP} and μ_{EOF}, a sample loading bias is created (*i.e.*, a higher analyte μ_{EP} results in greater loading). However, this bias can be corrected for mathematically and can actually be used for an additional level of selectivity. For example, a capillary treated to eliminate electroosmotic flow injects cations or anions exclusively, and neutrals will not be injected.

Sample concentration methods for sample introduction include techniques known as sample stacking and field-amplified sample injection. These methods are similar to electrokinetic sample introduction and also concentrate the analyte in the capillary, often leading to signal enhancements of 100–1000-fold. Sample stacking utilizes a sample solvent (*e.g.*, water) with lower conductivity than the electrophoretic buffer. This results in higher electric fields when the capillary is placed in the sample vial and voltage is applied. Under the higher electric field, analytes migrate to the boundary between the low-conductivity sample plug and the high-conductivity buffer. At this interface the migration is slowed on account of the (relatively) lower electric field. This serves to focus or decrease the width of the analyte plug; however, it must be noted that neutrals are not focused. During sample introduction, cations are stacked at the front of the lower-conductivity plug, whereas anions are at the back. Laminar flow and Joule heat can degrade performance, however, and so typical voltages are lower than those used during CE separations. Field-amplified sample injection utilizes similar principles to focus the analytes; however, a small plug of very low-conductivity solution that does not contain analyte is injected first. This results in even higher electric fields than generated in sample stacking and smaller injected sample volumes.

Modern CE instrumentation includes autosampling accessories for automatic sample introduction by the techniques just described. Automatic sample introduction typically increases the precision of CE analyses, and also increases the efficiency of laboratory analyses since it does not require the constant attention of an operator.

2.2 Capillaries

The capillaries used in CE are made of fused silica, Teflon, or Pyrex, with fused silica being used for the vast majority of separations. They range in length from 3 to 100 cm (with most separations performed using 30–100-cm-long capillaries), in i.d. from 25 to 200 µm (with 50- and 75-µm i.d. capillaries being fairly typical) and in outside diameter (o.d.) from 200 to 375 µm (with 350–375 µm being typical). The fused-silica capillaries have a 10–20-µm-thick external polyimide cladding to impart strength and flexibility.

Fused-silica capillaries are commonly available in untreated, surface-modified (coated), and gel-filled forms. Capillary choice is affected by the CE separation mode used, compatibility with the buffer or reagent choices, and cost. Careful control of the surface of untreated capillaries is important for reproducible control of electrophoretic and electroosmotic mobilities and can be easily achieved by selection of a buffer *p*H that sets the silanol/silanoate equilibrium chemistry at the desired point. The capillary surface can also be modified to reduce or eliminate EOF and to reduce analyte adsorption or interaction with the column. This was first demonstrated by Hjerten (1985). Analyte–column interactions can cause significant problems such as band broadening, irreproducibility, and irreversible analyte binding, especially in the analysis of large biopolymers such as peptides, proteins and DNA. Modifiers include covalently bonded molecules, polymers, and groups with either anionic or cationic functionality. Non-covalently bonded modifiers can also be added to the buffer system for temporary, removable surface modification. Modifier characteristics to consider include stability with the buffer during long-term use and inertness to analytes and buffers. Commonly used modifiers include bonded alkyl chains, polymers such as polyacrylamide (PA) or poly(vinyl alcohol) (PVA), and zwitterionic buffer additives. PA is the most common polymer used in this application and is also frequently used as the packing material in gel-filled capillaries for CGE. These CGE capillaries are prepared by performing the polymerization directly in the capillary to produce a cross-linked polymeric gel with a porous network having narrow, well-defined pore size distributions. Finally, surface-modified capillaries ordinarily used in gas chromatography and cyclodextrin-modified capillaries (for chiral separations) have also found limited use.

Precise control of capillary temperature is essential for reproducible CE, since migration time, electrophoretic and electroosmotic mobilities, injection volume, and separation efficiency and resolution are all functions of temperature. In addition, temperature control helps dissipate the Joule heat generated by the electrical power applied to the capillary, thereby preventing such undesirable consequences as vaporization of the buffer and thermal degradation of temperature-sensitive samples. Temperature control of the capillary (at temperatures near ambient, to approximately ±0.1 °C) is achieved by a variety of methods, including forced-air cooling, thermostatted liquid cooling, and the use of Peltier cooling devices. Finally, pressurization of the capillary at the inlet and outlet is sometimes required to prevent solvent outgassing (especially in capillary electrochromatography, CEC), which can perturb the flat flow profile required for high-resolution CE separations.

In capillary zone electrophoresis (CZE) and MECC, uncoated capillaries are used; in CZE, capillary ion electrophoresis (CIE), CIEF, and dynamic sieving capillary electrophoresis (DSCE) applications, coated capillaries are used; and in CGE and CEC, capillaries packed with sieving gels and LC stationary phases, respectively, are used. More details on these separation modes are presented in Sec. 3.

2.3 Buffers and Other Reagents

CE separations can be performed over a wide *p*H range (approximately 2–10.5). A variety of buffer systems are used in CE, including phosphate, tetraborate, carbonate, phthalate (for indirect detection methods, discussed below), and tris(hydroxymethyl) aminomethane–boric acid ("Tris–borate buffer"). Buffers are used individually and in combination, at typical concentrations of 0.01–0.1 *M*. Precise control of the buffer composition in the inlet vial, outlet vial, and capillary is essential for reproducible CE. In practice the buffer solution is replaced after a few experiments to prevent ionic depletion, which causes irreproducibility. The primary purposes of the buffer are to provide a conducting medium for electrophoretic and electroosmotic movement of analytes, to control the surface chemistry of fused silica capillaries, and to control the ionization and partitioning equilibria of analytes. These are key factors in adjusting the resolution of a CE separation and can be achieved by controlling the *p*H or by the use of buffer modifiers.

Increasing buffer *p*H will increase μ_{EOF}, since the zeta potential (with silica capillaries) increases with increasing *p*H. In contrast, increasing the ionic strength of the buffer will decrease μ_{EOF}, since the zeta potential decreases with increasing ionic strength. It is also important to know the isoelectric point (*p*I) or the acid dissociation constant K_a (typically expressed as $pK_a = -\log K_a$) of the analytes being separated, since they play important roles in determining their electrophoretic migration properties. Many CE separations are designed around the fact that at high *p*H, most biological molecules are anionically charged and are repelled from the negatively charged walls (in the case of an uncoated capillary); this reduces sample–wall interactions and sample losses due to adsorption on the wall. This approach allows uncoated capillaries to be used in the separation of proteins, as first shown by Lauer and McManigill (1986).

A variety of additives can be added to the buffer to modify the resolution of CE separations. For example, organic solvents are added to improve the solubility of analytes with limited solubility in water. In some cases, even 100% nonaqueous buffers have been successfully used in CE separations. Other materials (*e.g.*, potassium sulfate, ammonium sulfate) are added to prevent analyte adsorption on the capillary. Cyclodextrins, which are donut-shaped, water-soluble oligosaccharides with hydrophobic cavities and hydrophilic exteriors, can be added as additional partitioning phases to help separate analytes on the basis of their relative sizes and hydrophobic/hydrophilic characteristics. Another important cyclodextrin application is in the separation of stereochemical isomers, or enantiomers (see Sec. 3.1). In MECC (Sec. 3.4), detergents (*e.g.*, sodium dodecyl sulfate, SDS) are added as micellar partitioning phases. Soluble ion exchangers and charged colloidal particles have also been used in partition-assisted CE separations.

Additives can also be used to measure important CE parameters or to improve detector sensitivity. A small molecule, soluble and neutral at the *p*H used for analysis, can be added as a neutral marker to provide a measure of EOF velocity, since it has zero electrophoretic mobility. In ultraviolet–visible absorbance detection (Sec. 2.5.1), highly absorbing modifiers can be added to the buffer to perform indirect detection, where the decrease in measured absorbance due to the elution of a nonabsorbing analyte is detected as a negative peak. A similar approach can also be used in fluorescence (Sec. 2.5.2) and electrochemical (Sec. 2.5.4) detection.

In CZE, MECC, and CGE, the same buffer is used in the inlet and outlet vials; in CIEF, different solutions are used in the inlet and outlet vials, while an ampholyte is used in the capillary; and in capillary isotachophoresis (CITP), leading and trailing buffers are used to segregate the analytes into discrete zones. More details on these separation modes are presented in Sec. 3.

2.4 High-Voltage Power Supplies

High electric fields are required to generate the electrophoretic and electroosmotic velocities needed for high-resolution CE separations. Electric field strengths of 100–700 V cm^{-1} are generated across capillaries using high-voltage power supplies rated for use between 10 and 70 kV, with a typical experiment being performed at 30 kV. The current requirements on the power supply are fairly low, typically $<300\,\mu$A, and so the power ratings for these supplies are on the order of 10 W. The voltage and current outputs should be very stable for reproducible separations. Capillary electropherographs have interlocks to prevent electrical shocks to operators *via* accidental exposure to high voltages. Most CE modes are performed with constant-voltage power supplies, with the exception of CITP (Sec. 3.5), which is done in a constant-current mode.

Additional levels of efficiency and sensitivity in CE separations can be obtained with programmable power supplies. ac voltages, dc voltage pulses, and voltage ramps ("gradients") have all been used to modify analyte velocities in the capillaries in ways that improve resolution. The use of voltage gradients, where the applied voltage is changed linearly with time during the electrophoresis experiment, has been particularly successful in high-resolution separations of DNA fragments by CGE. Reversible-polarity options can allow the direction of migration to be reversed for applications where this is desired. Improved sensitivity in radiometric detection (Sec. 2.5.5) can be realized by increasing analyte residence time in the detector by voltage control of the analyte migration velocity. Finally, the use of a second power supply to apply a radial voltage to the capillary has been used to change the zeta potential and, therefore, change μ_{EOF} [Eq. (2)].

2.5 Detectors

A wide variety of detectors are available for use with CE. Many of these are common to LC (see CHROMATOGRAPHY) and have been modified to accommodate the unique constraints of CE. Requirements common to all CE detectors include issues related to connection to the capillary, compatibility with the buffer chemicals, low flow rates, high electrical fields at the capillary outlet, and sensitivities for the small amounts of analytes used in CE ($\leq 10^{-15}$ mol). Real-time, on-line detection is the preferred approach because of the problems inherent in handling small amounts of analytes for post-CE detection by an off-line analysis method.

The choice of a suitable detector from among the many available types is based on the requirements of the separation problem as well as detector availability and cost. General principles and CE applications of the most common detectors are discussed briefly in the rest of this section. Nondestructive detectors do not chemically or physically destroy the sample and are well suited to preparative applications where it is desirable to collect pure fractions of an individual component eluting from the capillary. Preparative CE is often performed to collect a pure sample of an analyte for subsequent off-line chemical, biochemical, or spectrometric analysis. In addition, nondestructive detectors can be used as the first detector in a tandem detector arrangement to improve the selectivity of on-line analyte detection.

In the sections that follow, detection limits (the minimum amount of detectable material) are given for each detector, although it must be emphasized that these are approximate values that depend on many variables, particularly the chemical and physical properties of the material being detected. The detection limits are given in terms of moles of analyte in the injected sample.

2.5.1 Ultraviolet–Visible Absorbance Ultraviolet–visible (UV–vis) absorbance detectors simply measure the absorption of light (in a particular wavelength range) by the analytes, and are the most widely used detectors in CE. This is primarily a result of two factors. First, optical absorbance detectors are amenable to integration with optically transparent fused-silica capillaries. Second, most compounds of interest absorb 190–800-nm radiation, while many other compounds that do not absorb can be chemically complexed or derivatized (*e.g.*, in a post-capillary column reactor) to form highly absorbing species. An example of this is the absorbance detection at 520 nm of the complexes formed between 4-(2-pyridylazo)resorcinol (PAR) and transition-metal ions such as Fe^{II}, Fe^{III}, Cu^{II}, Ni^{II}, Zn^{II}, Co^{II}, and Pb^{II}. If a nonabsorbing analyte is not amenable to complexation, CE can be performed in a buffer containing a highly absorbing modifier or in a light-absorbing buffer (*e.g.*, phthalate), and the analyte can be detected as a negative peak in a high-absorbance background (Small and Miller, 1982). This is sometimes referred to as indirect UV–vis absorbance detection. Monochromator-based scanning spectrometers and photodiode array detectors can provide an added level of identification, as the UV–vis absorbance spectrum can sometimes be used to confirm the identity of an analyte giving rise to a CE peak. The current generation of fast-scanning monochromator detectors typically have better signal-to-noise ratios than photodiode array detectors, and are preferred for applications requiring greatest sensitivity.

The capillary itself is typically used as the flow-through absorbance cell ("on-column" detection). Here, a section of the external polyimide cladding is removed from the capillary, and the UV–vis radiation is directed through the capillary. The absorbance A of an analyte is given by Beer's law:

$$A = \varepsilon b C, \tag{9}$$

where ε is the analyte molar absorption coefficient, b is the optical path length, and C is the analyte concentration. The small path lengths offered by the capillary i.d. have been largely offset by operation at low wavelengths (190–220 nm) where absorption coefficients are typically greatest, by sample-concentration injection methods (Sec. 2.1), which increase analyte concentrations in the capillaries by 100–1000-fold, and by the use of extended–path-length absorbance cells. An extended–path-length capillary known as a bubble cell is formed by a proprietary process that increases capillary i.d. (and therefore optical path length) in the detector by 3–5-fold. A second type of extended–path-length cell ("Z-cell") is formed by making two right-angle bends in the capillary and then directing the sample beam down the length of the middle leg of the "Z" instead of through the cross-sectional diameter of the capillary. This results in approximately a 20-fold improvement in sensitivity. Finally, optical detector designs are very efficient at reducing the amount of scattered light that reaches the photodiode, which also helps sensitivity.

UV–vis absorbance detectors are nondestructive, and for strongly absorbing analytes the detection limit is about 10^{-13}–10^{-16} mol. Further details on UV–vis absorbance can be found in OPTICAL SPECTROMETERS and MOLECULAR SPECTROSCOPY.

2.5.2 Fluorescence Fluorescence detectors measure the light emitted by the analyte after optical excitation by either a lamp (deuterium, quartz–tungsten–halogen, xenon, mercury) or a laser (laser-induced fluorescence, LIF). When a broad-band radiation source such as a lamp is used, the excitation of the analyte is limited to a narrow wavelength range by use of a monochromator. Detection of the fluorescent emission is performed with a photodiode or a photomultiplier tube located at a right angle from the excitation source. As in UV–vis absorbance detection, the capillary itself can be used as a flow-through detection cell, but improvements in sensitivity have been achieved by coupling the capillary to a cell with planar quartz windows, which reduces the amount of light scattering (Chen and Dovichi, 1994).

Fluorescence detectors are more selective than UV–vis absorbance detectors because not all compounds that absorb light subsequently fluoresce. Since only a small percentage of compounds natively fluoresce, analytes are usually derivatized with fluorescent tags. Derivatization can be performed either prior to the separation, in the capillary using a fluorescent tag added as a buffer modifier, or after elution from the capillary in a post-column reactor. Novel, analyte-specific fluorescent tags form an intense area of research, and have resulted in tremendous improvements in sensitivity for important analytes. For example, the fluorogenic reagent 3-(4-carboxybenzoyl)-2-quinoline-carboxaldehyde produces highly fluorescent isoindole derivatives with amino acids and peptides (Liu *et al.*, 1991). Nonfluorescing analytes can also be detected by indirect detection, whereby the CE separation is performed in a buffer containing a highly fluorescing modifier and the analyte is detected as a negative peak in a high-fluorescence background.

LIF is used when low detection limits are required, since the intensity of the emitted fluorescence is proportional to the number of excited molecules, which in turn is proportional to the intensity of the excitation radiation. LIF has been particularly useful for DNA analyses with intercalating dyes, which has eliminated the need for derivatization of the DNA. The increasing number and types of semiconductor and solid-state laser sources has increased the popularity of fluorescence detection and made it applicable to a wider range of analytes. Fluorescence detectors can be nondestructive (*e.g.*, if the analyte is not derivatized) and they have a detection limit of about 10^{-14}–10^{-21} mol. Further details on molecular fluorescence can be found in OPTICAL SPECTROMETERS and MOLECULAR SPECTROSCOPY.

2.5.3 Mass-Spectrometric Mass-spectrometric detectors measure the abundance and mass-to-charge (m/z) ratio of ions produced from analytes eluting from the capillary. For this measurement, the molecules are ionized (singly or multiply) after they elute and then introduced into the vacuum chamber of a mass analyzer. Ionization and subsequent mass measurement are accomplished with a variety of instrumentation. In order to do on-line coupling of CE with mass spectrometry (MS), the mass spectrometer must accommodate conditions unique to CE, including high–salt-content buffer systems, high electrical fields on the capillary outlet (which is interfaced to the inlet of the MS), and low flow rates. Although the low flow rates characteristic of CE make coupling to MS easier than is the case for LC, the other conditions constrain the ionization methods and mass analyzers that are used in CE.

Ionization methods coupled on-line with CE include electrospray ionization, ion-spray ionization, atmospheric-pressure ionization, and inductively coupled plasma ionization. Matrix-assisted laser desorption ionization (MALDI) has been used successfully off line in CE analyses. Electrospray ionization interfaces were first developed by Smith *et al.* (1988) and are the most popular devices for coupling capillaries to mass spectrometers. Typically, electrospray ionization uses buffer flow through a steel needle (on the end of the capillary) held at a high electrical potential to ionize analytes that are amenable to either negative or positive ionization modes. For optimum ionization efficiency, and therefore optimum detection limits, the design of the interface between the capillary outlet and the electrospray needle is critical. Combinations of sheath flow liquids (Wahl and Smith, 1994) with capillary outlet tip modifications have been used to enhance detection sensitivity.

Mass analysis brings an additional level of information to CE analysis. The mass spectrum of a detected peak, which is a plot of ion intensity versus m/z, contains molecular-weight, formula, and structural information. Mass analyzers utilized include quadrupole ion trap, quadrupole filter, time-of-flight, and ion-cyclotron-resonance (ICR) mass spectrometers. Each of these can also be used to perform tandem experiments ("MS–MS") if the analyzers are properly equipped. MS–MS experiments provide detailed structural information about an analyte and have been applied to polypeptide sequence determination. Time-of-flight analyzers are useful for the measurement of very high-mass analytes ($>3\,000$ atomic mass units, u) and for rapid CE (in which eluted peaks may be only a few seconds wide), while ICR instruments are used in applications requiring the highest mass resolution. Mass analyzers can be operated either as universal detectors (by monitoring the intensity of all m/z regions of the mass spectrum) or as selective detectors (by monitoring selected m/z regions). Mass-spectrometric detectors are destructive and have typical detection limits of about 10^{-15}–10^{-18} mol. Further details on mass spectrometers can be found elsewhere in this Encyclopedia (see MASS SPECTROMETERS).

2.5.4 Electrochemical Electrochemical detectors measure changes in current or conductivity between a set of electrodes in a detection cell. It must be noted that these are not the same electrodes that are used to generate the high electric fields needed for electrophoretic migration. Electrochemical detectors are seeing increasing use in biochemical analyses because many biochemical compounds are electroactive since they are easily oxidized. Also, several novel detector designs have been reported that are free from electrophoretic voltage interference, making this a reliable detection method.

Amperometric detectors measure the current produced by a controlled-potential electrolysis of the analyte in the buffer. The analytes from the column elute past a working electrode (one of three electrodes in the detector cell) set at a specific potential with respect to a reference electrode. If an analyte can be oxidized or reduced at this potential, the transfer of electrons at the electrode–buffer interface serves as the amperometric detection signal. Some recent designs use on-column detection with 5–25-μm-diameter working electrode wires either placed over the capillary outlet (Zhong and Lunte, 1996) or inserted directly into the end of the capillary (Swanek *et al.*, 1996). An additional level of information can be obtained with voltammetric detectors, wherein the potential of the working electrode is changed with respect to time as the analyte passes through the cell. Voltammetric detectors provide additional measures of the thermodynamic and kinetic properties of the electron-transfer reactions of the analytes, which can be used to aid in the identification of unknown analytes. As with UV–vis and fluorescence detectors, indirect methods of detection are possible with amperometric detectors, whereby a nonelectroactive analyte is detected as a negative peak when the CE separation is performed in a buffer containing a high concentration of an electroactive modifier.

In conductivity detectors, the electrical conductivity of the eluting buffer is measured by applying an alternating voltage between two electrodes and measuring the resulting current. Conductivity is related to the current/voltage ratio and represents the contribution of the individual equivalent conductivities of all ions in solution, including the analyte ions. Measuring the small conductivity due to an eluting ion in the presence of the high background conductivity of the buffer is difficult. Suppression of the conductivity of the buffer (Small *et al.*, 1975) dramatically improves the sensitivity. For example, consider the case where a sodium carbonate buffer is used in anionic separation by CE. As the buffer exits from the capillary, it passes through a cation exchanger (suppressor) loaded with hydrogen ions. Sodium ions from the buffer displace hydrogen ions from the ion-exchange sites on the suppressor, with two beneficial results. First, carbonic acid is formed by the ion-exchange reaction, so that detection is carried out with a background characterized by the low conductivity of carbonic acid, as opposed to the higher conductivity of sodium carbonate. Second, the analyte anions are paired with hydrogen ions instead of sodium ions, so that detector sensitivity is increased because the equivalent conductivity of the hydrogen ion is approximately seven times as large as that of the sodium ion. Several on-column conductivity-cell designs have been developed that incorporate the electrodes and the suppressor directly on the end of the capillary. Conductivity detection is usually described as suppressed or nonsuppressed to indicate whether ion exchangers are incorporated.

Amperometric detectors are destructive and have typical detection limits of about 10^{-18}–10^{-19} mol. Conductivity detectors are destructive if a suppressor is used, and have typical detection limits of about 10^{-15}–10^{-16} mol. Further details on electrochemistry can be found in ELECTROCHEMISTRY.

2.5.5 Other Detectors Other detectors use optical, radioactive, or other analyte properties as the basis of detection, but are not as widely used as those already discussed. Detectors based upon analyte optical properties (or on the properties of the products of post-separation reactions involving the analyte) include refractive-index, thermooptical, and chemiluminescence detectors. Radiometric detectors detect radioactive decay from commonly used isotopes of phosphorus, sulfur, or hydrogen that have been incorporated into the analyte. Detectors based on other analyte characteristics (*e.g.*, nuclear magnetic resonance spectrum, Raman spectrum, proton-induced x-ray emission, particle light scattering) and capillary characteristics (laser-induced capillary vibration) have found special applications as on-line detectors in capillary electrophoresis.

2.6 Integrated Capillary Electrophoresis Microdevices

An exciting development in capillary electrophoresis is the fabrication of miniaturized capillary electropherographs on planar substrates such as silicon, quartz and glass. These CE instruments were first introduced in 1991 (Fan and Harrison) with CE separation channels (the planar analog to conventional capillaries) built on the surface of silicon substrates. These channels were fabricated by use of a combination of masking, photolithographic, and etching techniques that had been primarily developed for the fabrication of silicon-based integrated circuits and microelectromechanical structures (MEMS). Since then, sample-introduction, separation-channel, and detector components have been fabricated on glass and quartz by modifications of silicon micromachining methods. While some detectors, such as electrochemical detectors, have been integrated into these devices (Ewing and Gavin, 1996), other detection schemes, such as laser-induced fluorescence, are being performed with some integrated components (*e.g.*, post-column reactors) and some external components (*e.g.*, the laser and photodiode) not integrated onto the substrate (Fluri *et al.*, 1996). On account of the extremely small size of the separation channels, which are typically $5 \times 50\,\mu m^2$ in cross-section, the efficiency of these separations is excellent, with 70 000–160 000 theoretical plate separations reported on 5–50-mm-long channels with 0.1-nL sample volumes and 10–20-s migration times (Manz *et al.*, 1993). In addition, these devices can perform multichannel separations of a single sample by multiple, complementary modes of separation. This is achieved by voltage control of the direction and speed of sample electrophoretic migration in a grid array of perpendicular channels, each of which can be tailored to provide a particular mode of separation (Sec. 3).

3. MODES OF SEPARATION

A wide range of separation selectivity can be obtained by separating analytes not only on the basis of differences in their electrophoretic mobilities, but also on differences in their isoelectric point, molecular size, hydrophobic and hydrophilic nature, molecular structure, polarity, and stereochemistry. In this section, CE modes of separation are discussed that exploit these analyte properties. In addition, applications from the current scientific literature are used to illustrate the usefulness and limitations of the different CE separation techniques. It will be seen that CIEF, CGE, and CITP are capillary analogs of slab gel electrophoresis, while CZE, CIE, MECC, and CEC are uniquely capillary techniques.

3.1 Capillary Zone Electrophoresis (CZE)

Capillary zone electrophoresis (CZE), also known as free-solution capillary electrophoresis (FSCE), separates analytes primarily on the basis of differences in their electrophoretic migration rates. When CZE is performed in untreated capillaries, migration is related to the sum of the electrophoretic and electroosmotic mobilities. In treated (*i.e.*, wall-coated) capillaries, migration is due solely to the electrophoretic mobility of the analytes. The capillaries are not packed with a gel or a sta-

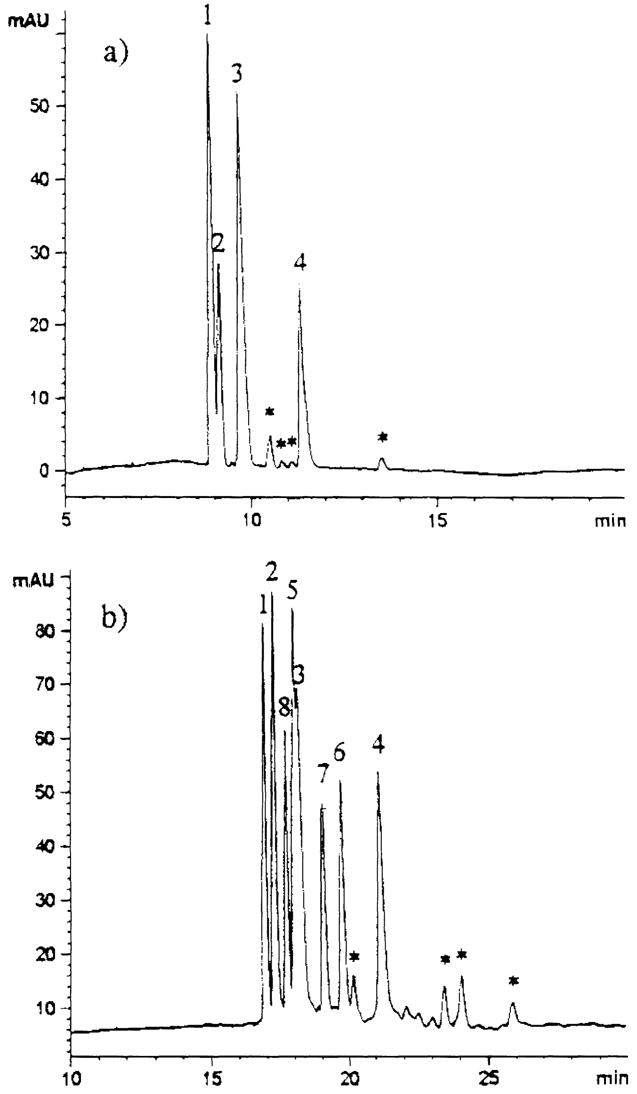

FIG. 4. CZE separation of all eight stereoisomers of Tyr-Lys-Trp (a) without and (b) with 18-crown-6 tetracarboxylic acid added to the buffer as a chiral selector. Sample introduction: hydrodynamic (pressure) injection, 50 mBar, 2 s. Capillary: fused silica, untreated, 50 cm long × 75 μm i.d., 30 °C, 30 kV applied voltage. Buffer: *p*H 2 Tris-borate buffer, with or without 0.01 *M* chiral selector. Detection: UV–vis absorbance at 210 nm. Elution order of analytes in (a): 1, LLL/DDD; 2, LDL/DLD; 3, LLD/DDL; 4, DLL/LDD. Elution order of analytes in (b): 1, LLL; 2, LLD; 3, LDL; 4, DLL; 5, DDL; 6, DLD; 7, LDD; 8, DDD. Asterisks indicate peaks due to side products of the crude peptides. Reprinted (in part) from Riester *et al.* (1996). Copyright (1996) American Chemical Society.

tionary phase material, and the same buffer is used in the capillary and the inlet, outlet, and sample vials. This mode of CE has found particularly good applications in the determination of amino acids, peptides, proteins, nucleosides, nucleotides, and vitamins.

Chiral compounds can be separated by addition of an appropriate modifier (*e.g.*, cyclodextrin) to the buffer. These compounds are molecules that are not superimposable with their mirror images; *e.g.*, in organic chemistry, they are molecules that contain a carbon atom that is covalently bonded to four different functional groups. They are also known as stereoisomers, optical isomers, optically active isomers, diastereomers, and enantiometers. In many pharmaceutical, agrochemical, and biological applications, where optical purity can influence effectiveness, stereoisomers can be difficult to separate in HPLC. The ability of CZE to separate stereoisomers is demonstrated in Fig. 4, showing separation of the eight stereoisomers produced in the solid-phase synthesis of the (target) tripeptide D-Tyr-L-Lys-L-Trp (Riester *et al.*, 1996). The prefixes D and L refer to the relative configuration of the atoms on the enantiomeric forms of each peptide group. In solid-phase peptide synthesis the polypeptide is produced stepwise from individual amino acids in an automated reactor. Side reactions, deletion errors due to an incomplete coupling cycle, and isomerization can all lead to undesired stereoisomers. The eight different tripeptide stereoisomers are separated into four pairs that each elute together in the absence of a chiral selector [Fig. 4(a)], but are completely resolved when a chiral selector is added as a buffer modifier [Fig. 4(b)]. This CZE method was used to investigate the rate of racemization (a reaction whereby a pure L or D form of a stereoisomer is converted into a mixture of equal parts of L and D forms). This work demonstrated the ability of CZE to measure the optical purity of the derivatized amino acid starting materials used in solid-phase peptide synthesis and was the first separation of all stereoisomers of a three–chiral-center tripeptide by CZE.

Figure 5 shows the electropherogram for the separation of 5-(dimethylamino)naphthalene-1-sulfonic acid (Dns-OH) and four L-amino acids using a capillary electropherograph microfabricated on quartz and Pyrex glass substrates (Fluri *et al.*, 1996). The analytes were derivatized with *o*-phthaldialdehyde in a microfabricated post-column reactor prior to fluorescence detection. The remarkable speed and efficiency of separations performed on integrated CE microdevices is demonstrated by the fact that all five analytes elute within 20 s, with N ranging between 23 000 and 94 000 theoretical plates based on analysis of the peaks (Sec. 1.2).

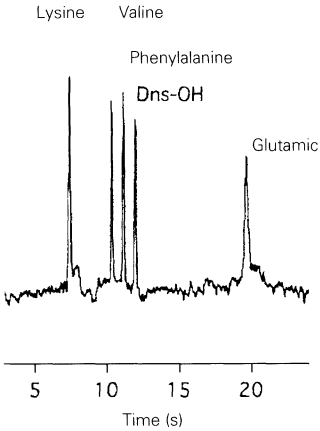

FIG. 5. CZE separation of a mixture of of 5-(dimethylamino)naphthalene-1-sulfonic acid (Dns-OH) and four L-amino acids using a capillary electropherograph microfabricated on quartz and Pyrex glass substrates. Sample introduction: electrokinetic injection, −1.8 kV, 5 s. Separation channel: Pyrex glass, 4.5 cm long × 45 μm wide × 15 μm deep, −11 kV. Buffer: 0.027 *M* sodium hydrogen carbonate and 0.012 *M* sodium carbonate, adjusted to *p*H 9.7. Detection: LIF following post-column reaction with 0.004 *M* *o*-phthaldialdehyde, He/Cd excitation (9.3 mW, 325 nm), photomultiplier tube detection. Reprinted (in part) from Fluri *et al.* (1996). Copyright (1996) American Chemical Society.

The speed and efficiency of CZE illustrated in the determination of small polypeptides and amino acids has also been applied in the determination of proteins. For large molecules such as proteins, the secondary and tertiary structures of the molecule have an effect on the net charge and its size or hydrodynamic radius. The effect of these parameters on the electrophoretic mobility of some proteins has been studied (Ma *et al.*, 1995). Protein modifications such as degradation, glycosylation, deamidation, or sequence changes will result in altered CZE migration times. These effects have allowed CZE to be successfully applied to confirmation of protein identities and detection of purity, heterogeneity and post-translational modifications in both native and synthetic proteins (see Sec. 3.2). Alternatively, protein sequence can be determined by peptide mapping, as reviewed for recombinant proteins by Ganzler *et al.* (1995). Peptide mapping involves enzymatically breaking a protein into smaller polypeptides. Analysis by CZE provides a map or fingerprint of the constituent polypeptides. Different enzymes are used to produce a variety of maps, and each map can be used for comparison or as a quality check for recombinant methods. Further, using a mass spectrometric detector, molecular weight can be measured and in many cases the sequence of each individual polypeptide can be determined.

Capillary ion electrophoresis (CIE), also sometimes referred to as capillary ion analysis (CIA), is a special name given to CZE separations that are optimized for the separation of small ionic species. UV–vis absorbance (Sec. 2.5.1) is the most common detection mode used in CIE; however, since many of these ions do not absorb UV–vis radiation, indirect de-

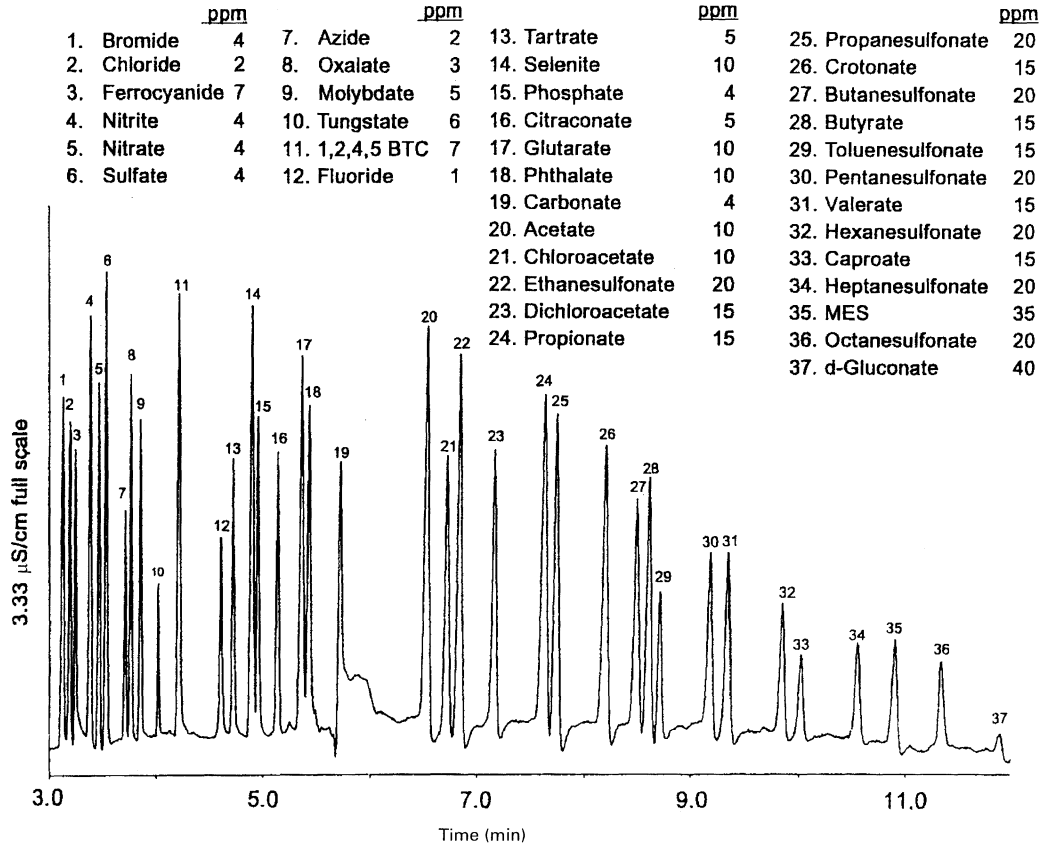

FIG. 6. CIE separation of 37 anions. Sample introduction: hydrodynamic (pressure) injection, 25 mbar, 12 s. Capillary: fused silica, 60 cm long × 50 μm i.d., room temperature, 30 kV applied voltage. Buffer: 0.05 *M* 2-{*N*-cyclohexylamino}-ethanesulfonic acid, 0.02 *M* lithium hydroxide, 0.03% Triton® X-100 surfactant (Rohm & Haas, Philadelphia, PA, U.S.A.). Detection: nonsuppressed conductivity, with detector response shown as 33 μS cm^{-1}. Cetyltrimethylammonium bromide (CTAB), 0.001 *M*, was flushed through the capillary prior to analysis. Peak assignments for the 37 anions are shown along with the concentrations of the individual anions in the mixture in parts per million (ppm). MES represents 2-[*N*-morpholino]ethanesulfonic acid. Reprinted (in part) from Jones *et al.* (1996) with permission from *American Laboratory*. Copyright (1996) International Scientific Communications, Inc.

tection schemes are typically used. Conductivity detectors, both suppressed and nonsuppressed, are also finding increasing use in CIE. When cations and anions are both present in the sample mixture, untreated fused-silica capillaries are used. However, when the sample contains only anions, wall-coated capillaries can be used to decrease the migration time for the analyte anions.

In Fig. 6, the CIE separation of 37 anions is shown (Jones *et al.*, 1996). To decrease the electrophoretic migration time of the anions, the direction of electroosmotic flow was reversed by treating the fused-silica capillary with CTAB prior to analysis. The elution order is determined by the electrophoretic mobility and, therefore, the charge/size ratio of the anions. As an example, consider the order of elution for the Group VIIA anions in the sample (bromide, chloride and fluoride). These anions are singly charged and their hydrated ionic radii increase in the order bromide < chloride < fluoride. Therefore, bromide elutes first, followed in order by chloride and fluoride. The excellent speed and resolution of capillary electrophoresis is also demonstrated by this example, as nearly all of the peaks are separated with baseline resolution in a separation that takes less than 12 min and uses only 12 nL of sample solution.

3.2 Capillary Isoelectric Focusing (CIEF)

In capillary isoelectric focusing (CIEF), analytes are separated on the basis of differences in their isoelectric points. As opposed to the other CE zonal modes of separation, CIEF is, as the name implies, a focusing technique. Further discussion of CIEF requires some definitions of terms. An ampholyte (or amphotere) is a compound that can exist as either a cation or an anion depending on the buffer *p*H. Ampholytes can take many forms, but the type most commonly used in CIEF comprises organic compounds with both amino (basic) and carboxylic acid functional groups. In its basic form, an ampholyte is an anion, while in its acidic form, it exists as a cation. The buffer *p*H at which the ampholyte is neutral is the isoelectric point, *p*I.

CIEF is performed by a three-step process on capillaries that have been treated to eliminate EOF. First, the sample is injected into the capillary along with a mixture of ampholytes. This is typically done by hydrodynamic injection of relatively large sample volumes, since the sample can fill the entire capillary. Second, a focusing step is performed with high electric fields (typically 400–700 $\mathrm{V\,cm^{-1}}$) across the length of the capillary. The inlet (anode) vial contains a dilute, acidic anolyte (*e.g.*, phosphoric acid), and the outlet (cathode) vial contains a dilute, basic catholyte (*e.g.*, sodium hydroxide). Focusing is performed until the current drops to a steady value, and usually takes 3–5 min. During this step, a *p*H gradient is established along the length of the capillary, and sample analytes migrate to their individual isoelectric points in the capillary and are focused into sharp zones. These zones are the narrowest when there are a large number of ampholytes with evenly spaced isoelectric points introduced during the first step. In the third and final step, the analyte zones are mobilized so that they move to the CE detector. Mobilization can be achieved by an electrophoretic process, whereby the high voltage is first turned off, the compositions of the anolyte and/or the catholyte are changed (*e.g.*, by addition of a salt), and the voltage is then turned back on. When the voltage is turned on, a gradual shift in the *p*H gradient in the capillary occurs, the analytes acquire electric charges, and the analytes then move towards the detector as a result of electrophoresis. Adding a salt containing a cation other than the hydrogen ion to the anolyte will produce mobilization towards the anode. Conversely, adding a salt containing an anion other than the hydroxyl ion to the catholyte will produce mobilization to the cathode. Mobilization can also be performed by hydrodynamic means. Here, hydrostatic pressure caused by pressurization of the catholyte or anolyte vial, or by changing the level of either the catholyte or anolyte vial, causes buffer flow through the capillary.

CIEF is used for a variety of separation problems, such as proteins, monoclonal antibodies, and hemoglobins. In Fig. 7, the CIEF separation of recombinant human growth hormone (rhGH) variants is shown (Shimura and Karger, 1994). The analytes are detected as a complex formed by post-column reaction with a reagent (TR-Fab') synthesized by the reaction of a purified mouse monoclonal antibody with tetramethylrhodamine-5-iodoacetamide. CIEF was a particularly good mode of separation for this analysis on account of the fairly large sample volume that could be loaded into the capillary. In CIEF, the order of elution is related to the *p*I's of the individual analytes. With appropriate *p*I standards, a calibration curve of *p*I versus mobilization time can be used to determine the *p*I of an unknown analyte with an absolute accuracy of 0.05 pI units.

3.3 Capillary Gel Electrophoresis (CGE)

In capillary gel electrophoresis (CGE), separation is enhanced by size exclusion (*i.e.*, sieving) in polymeric gels containing a network of pores with narrow, well-defined size distributions. Small analytes permeate through the pores in the polymer and appear early in the electropherogram. Large analytes are slowed down by the gel; therefore, they have longer migration times and appear later in the electropherogram. This elution order is the

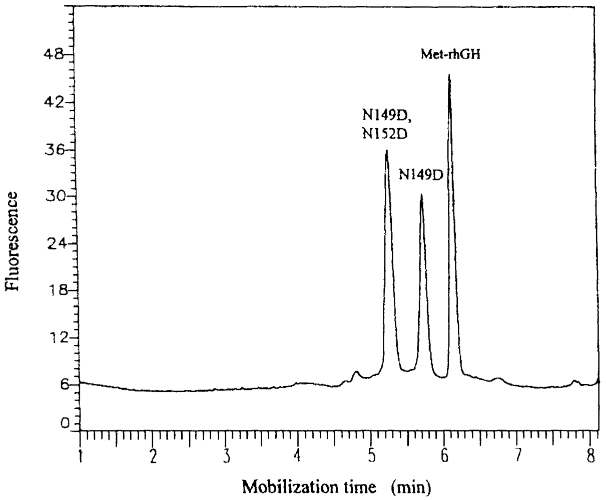

FIG. 7. CIEF separation of a mixture of recombinant human growth hormone (rhGH) variants. Mixture contains met-rhGH, the monodeamidated variant (N149D), and the dideamidated variant (N149D, N152D), each at a concentration of 10 $\mu g\,L^{-1}$. Sample introduction: hydrodynamic (vacuum) injection, 10 μL syringe. Capillary: polyacrylamide-coated fused silica, 15 cm long × 75 μm i.d. Ampholyte: 6% Pharmalyte 3–10 (Pharmacia LKB Biotechnology, Piscataway, NJ, U.S.A.). Catholyte: 0.02 *M* sodium hydroxide. Anolyte: 0.01 *M* phosphoric acid. Focusing: 1 min at 200 $V\,cm^{-1}$ and 4 min at 400 $V\,cm^{-1}$. Mobilization: change anolyte to 0.01 *M* phosphoric acid containing 0.1 *M* sodium chloride, 400 $V\,cm^{-1}$. Detection: LIF at 580 nm following post-column reaction with TR-Fab' complexing reagent, Ar^+ excitation (2 mW, 488 nm). Reprinted (in part) from Shimura and Karger (1994). Copyright (1994) American Chemical Society.

opposite of that in size-exclusion chromatography (SEC; see CHROMATOGRAPHY), where molecules that are too large to permeate the pores in the stationary-phase particles flow around the particles and appear first in the chromatogram, while small molecules that permeate the pores travel a long and tortuous path through the particles and elute last. In CGE, capillary walls are typically treated to enhance gel adhesion and stability. EOF is therefore eliminated and the analytes migrate through the gel as a result of their electrophoretic mobilities, which can add a second degree of separation selectivity to that based on size exclusion.

Typical gels in CGE include polyacrylamide (PA), agarose, and poly (ethyleneoxide), with PA being the most common. CGE with PA gels is frequently referred to as capillary polyacrylamide gel electrophoresis (C-PAGE). The gel is usually either a rigid, cross-linked polymer that is attached to the capillary wall or a linear polymer solution that is not cross-linked and is not fixed to the capillary. CGE capillaries with cross-linked gels occasionally have problems with gel stability (which can limit their usefulness to 100–200 analyses) and with bubble formation during electrophoresis (which can be addressed by careful control of experimental parameters). Also, electrokinetic injection is typically used with these types of gels because hydrodynamic injection is not usually capable of introducing the sample into the gel-filled capillary. In contrast, when the gel is a polymer solution composed of non-cross-linked polymers (*e.g.*, a linear polyacrylamide gel), either hydrodynamic or electrokinetic injection can be used. Also, capillary lifetimes are much longer than for capillaries with cross-linked polymers, as it is possible to introduce a new polymer solution into a capillary by a simple pressure loading technique prior to a given set of CGE experiments. With this mode of CGE, sometimes referred to as dynamic sieving capillary electrophoresis (DSCE), the capillary wall is treated to prevent EOF transport of the sieving gel out of the capillary. The polymer type, chain length, and concentration are all selected to give the desired size-exclusion properties for the separation problem at hand.

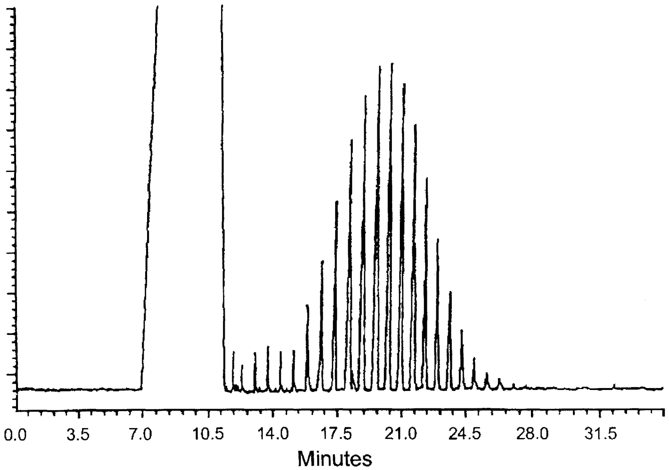

FIG. 8. CGE separation of phthalate-derivatized poly (ethylene glycol) oligomers. Sample introduction: electrokinetic injection, −10 kV, 60 s, sample concentration 1300 mg L^{-1}. Capillary: fused silica, 45 cm long× 75 μm i.d., filled with cross-linked polyacrylamide gel, −10.9 kV applied voltage. Buffer: Tris–borate, pH 8.3. Detection: UV–vis absorbance at 275 nm. Reprinted (in part) from Wallingford (1996). Copyright (1996) American Chemical Society.

In the CGE electropherogram shown in Fig. 8, the separation of phthalate-derivatized poly(ethylene glycol) oligomers is shown (Wallingford, 1996). An oligomeric set is a collection of polymers, each composed of one more monomer unit than the previous polymer. It is important to analyze these because the distribution of polymers within the set and the average molecular weight of the set determine key physical properties of the polymer. The oligomers elute in order of increasing size, and a smooth curve drawn so as to connect the individual oligomeric peak maxima can be related to the molecular-weight distribution (polydispersity) of the oligomers. With appropriate molecular-weight standards, a calibration curve of log(molecular weight) *versus* migration time can be used to estimate the molecular weight of an unknown polymer by CGE analysis. The benefits of CGE in the example shown in Fig. 8 were higher-accuracy measurements relative to gel permeation chromatography (GPC) and higher separation efficiencies (100 000–300 000 theoretical plates) than is possible with HPLC. Differences in the average weight and polydispersity between two vendors of the "same" polymer were also observed by the author.

Gel-filled capillaries were introduced by Cohen and Karger (1987), who first used them for the separation of polynucleotides by CGE. Today, CGE is widely used for many important separations in biochemistry, including nucleic acids, oligonucleotides, PCR products, DNA molecules (with as many as 5000 base pairs), DNA sequencing reaction products, and proteins. Native proteins can be separated by CGE as uncomplexed species, but molecular size determination becomes difficult because of mixed electrophoretic and size-exclusion contributions to the net mobilities of the proteins. In one very popular CGE method, proteins are treated with sodium dodecyl sulfate (SDS). The SDS denatures and interacts with each protein to produce adducts with virtually the same charge/size ratio, yet the size differences are retained. Similarly, some oligonucleotides and DNA fragments differ from one another in size but not in their charge/size ratios. None of these samples can be resolved by CZE (since they all have the same charge/size ratio), but the size discrimination properties of CGE allow separation with high resolution. CGE can also therefore be used to determine molecular weights.

The application of CGE to DNA sequencing analyses, first demonstrated by Swerdlow and Gesteland (1990), has provided significant advances in terms of speed and resolution compared with traditional slab gel electrophoresis. Standard DNA sequencing methodologies are used to produce single strands of DNA that terminate after each nucleotide type (deoxyadenosine, deoxycytidine, deoxyguanosine, and deoxythymidine, or A, C, G, T). For a DNA strand that contains G at base

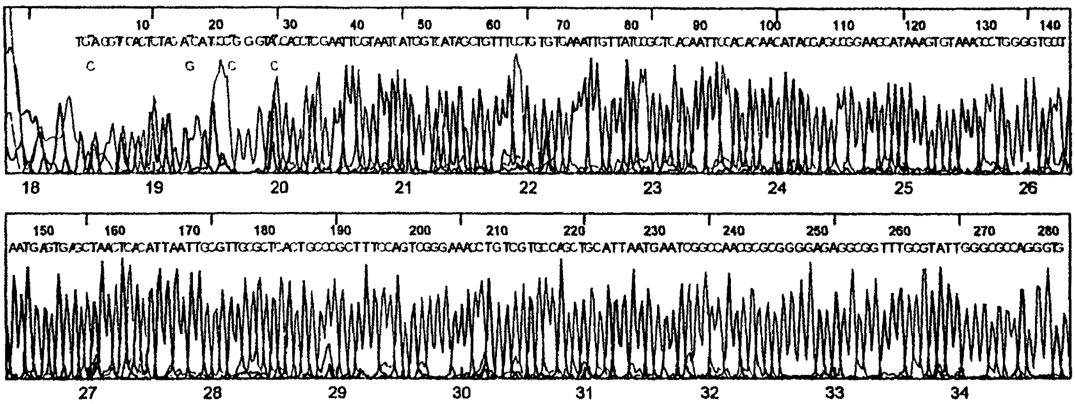

FIG. 9. Capillary gel electrophoresis of DNA sequencing products: partial electropherograph (fluorescence intensity *versus* time) showing separation of products of length 1 through 282 bases. Upper numbers show length (in bases), letters indicate terminal base. Sample introduction: electrokinetic injection, $-200\,\mathrm{V\,cm^{-1}}$, 5 s. Capillary: fused silica, poly (vinyl alcohol) coated, filled with 2% *T* (*T* = total monomer concentration) linear polyacrylamide, 30 cm long × 75 μm i.d., $-150\,\mathrm{V\,cm^{-1}}$ applied voltage. Buffer: 100 m*M* Tris-50 m*M* TAPS-2 m*M* EDTA-7 *M* urea. Detection: LIF, argon-ion laser (5 m*W*, 514 nm) excitation, photodiode array detection. Reprinted (in part) from Carrilho *et al.* (1996). Copyright (1996) American Chemical Society.

number 54 and 87, for example, two individual strands (sequencing products) of lengths 54 and 87 are produced with a terminal G. Ordinarily, the four groups of sequencing products would be labeled with radioactive isotopes and determined individually in a single slab gel experiment, with the analysis alone taking many hours. With CGE, however, each group of sequencing products is labeled with a different fluorescent tag and the groups are combined and analyzed by CGE with a significantly shorter analysis time.

An example of this type of analysis, using a replaceable linear polyacrylamide, low voltages, and laser-induced fluorescence detection, is shown in Fig. 9 (Carrilho *et al.*, 1996). By identifying the terminal base on each sequenced product, the sequence of the entire polynucleotide was determined. The experiment shown in Fig. 9 was used to sequence a polynucleotide (ssM13mp18 template) with greater than 1000 bases (only a portion is shown) in less than 90 min, whereas comparable gel-slab analysis would take from 8 to 18 h. The ability to sequence large polynucleotides by CE provides significant improvements relative to slab gel electrophoresis. CE is faster, uses smaller reagent and sample quantities, generates less waste, and is easier to automate.

3.4 Micellar Electrokinetic Capillary Chromatography (MECC)

As the name implies, micellar electrokinetic capillary chromatography (MECC) takes advantage of both electrophoretic and chromatographic separation mechanisms. The detector output is therefore called an electrokinetic chromatogram. MECC can separate charged and neutral analytes, but its most important application is for separation of neutral species, which cannot be separated by CZE. MECC is performed in untreated capillaries and analytes are partially solubilized by micelles added to the buffer, *i.e.*, they partition between the micellar and buffer phases. The degree to which each analyte partitions between the micelles and the buffer is the basis for separation. Electroosmotic mobility and micellar electrophoretic mobility are typically controlled so that the time an analyte spends in the micelle phase is directly related to its migration time. The micelles are analogous to the stationary phase in reversed-phase LC (see CHROMATOGRAPHY) and are referred to as a "pseudophase" because they migrate through the capillary. The use of a micellar phase to enhance electrophoretic separation was first demonstrated by Terabe *et al.* (1984). Drugs, vitamins, and

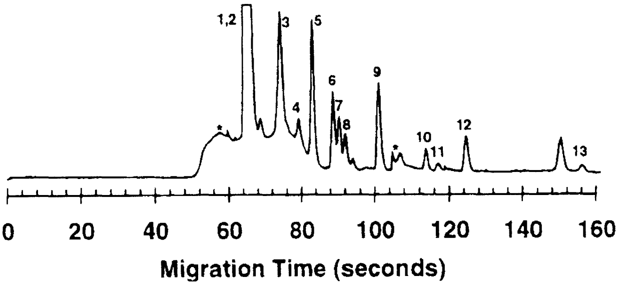

FIG. 10. MECC analysis of fifteen amino acids, allowing quantitation of ten. Sample introduction: electrokinetic injection, −100 V, 2 s. Capillary: fused silica, untreated, 10 cm long × 25 µm i.d., −12 kV applied voltage. Buffer: 0.175 *M* 2-(*N*-cyclohexylamino) ethanesulfonic acid adjusted to pH 9; containing 0.10 *M* sodium dodecyl sulfate (SDS) as the micellar additive. Detection: LIF following pre-column on-line reaction with 0.110 *M* *o*-phthaldialdehyde and 0.220 *M* β-mercaptoethanol in 0.025 *M* borate buffer at pH 9.5, He/Cd excitation (2 mW, 354 nm), photomultiplier tube detection. Elution order of analytes: 1 and 2, glutamine and serine and threonine; 3, alanine and glycine; 4, tyrosine; 5, taurine; 6, valine; 7, glutamate; 8, methionine; 9, aspartate; 10, leucine; 11, phenylalanine; 12, isoleucine; 13, lysine. Asterisks indicate peaks due to impurities and/or reagents. Reprinted (in part) from Lada and Kennedy (1996). Copyright (1996) American Chemical Society.

large peptides are among the compounds commonly determined by MECC.

Micelles are created in the buffer solution by the addition of surfactant molecules. These straight-chain molecules have a hydrophilic end and a hydrophobic end, and, above a critical micelle concentration (typically 10^{-3} *M*), they aggregate to form a micelle. The number of molecules that form a micelle is called the aggregation number and can vary from 2 to over 100. Different surfactants will have different critical concentrations and aggregation numbers. Surfactant molecules can be nonionic, anionic, cationic, or zwitterionic (*e.g.*, having both cationic and anionic functional groups). SDS, the most common surfactant used in MECC, is an anionic molecule that forms micelles with an aggregation number of 162, an average charge of −63, and an average molecular weight of 18 000. In aqueous buffers, SDS molecules orient their hydrophobic ends inward and hydrophilic ends outward. Surfactant concentration and structure determine the degree of partitioning for each analyte and therefore affect resolution and migration time. The charge/size ratio determines micelle electrophoretic mobility, and buffer *p*H control is used to keep the surfactant molecules in the correct ionic state. Hydrophilic analytes that remain in the buffer will elute first, while hydrophobic analytes remaining in the micelle will elute last. Surfactant molecules are chosen for their hydrophilic/hydrophobic balance, solubility, purity, and cost. The buffer is then chosen so that the electrophoretic mobility of the micelles is opposite to, and of lower magnitude than, the electroosmotic flow. In this manner the micelles are attracted to the inlet vial (based on electrophoretic mobility) but ultimately migrate to the detector. Finally, it should be noted that separation of chiral compounds can be achieved by the use of a chiral surfactant (such as deoxycholate or taurodeoxycholate).

Separations can be enhanced in MECC by the use of buffer modifiers. Molecules added to the buffer can change the charge of the micelle, change the solubility of the analytes in the buffer, or act as a secondary pseudophase into which the analytes partition. These can all affect analyte solubility in the micelle and therefore MECC separation characteristics. Common modifiers include organics such as methanol and isopropanol, and metal ions, which attach to the surface of the micelle. Cyclodextrins are also used as modifiers, and analytes can partition into cyclodextrin molecules on the basis of the relative sizes of the analyte and the cyclodextrin cavity and on hydrophobic/hydrophilic interactions.

Figure 10 shows an electrokinetic chromatogram of a mixture of amino acids detected in a microdialysis fluid (Lada and Kennedy, 1996). Continuous dialysate flow was mixed on line with a derivatization reagent and introduced into a flowing buffer solution (orthogonal to the capillary inlet). The buffer flow was paused periodically to allow electrokinetic injection into the capillary. Most of the derivatized analytes were unresolved under CZE conditions. In contrast, MECC using SDS micelles allowed the separation and quantitation of ten amino acids, including leucine and isoleucine. The rapid analysis time allowed for greater temporal resolution during *in vivo* monitoring. Compounds were identified by matching migration times to known standards, while unidentified peaks and the broad peak early in the electropherogram were products of side reactions or minor products detected as a result of *in vivo* conditions. The rapid analysis and high number of theoretical plates for the MECC method permitted a fast sampling rate and simultaneous quantitative monitoring of ten analytes.

3.5 Capillary Isotachophoresis (CITP)

Capillary isotachophoresis (CITP) can be used either as a separation method or as a sample preconcentration method prior to another separation mode. In CITP, the sample solution is sandwiched between two buffers, a leading buffer and a trailing (or terminating) buffer. The leading buffer contains an ion with a higher electrophoretic mobility than any analyte ions in the sample. Similarly, the trailing buffer contains an ion with a lower mobility than all those in the sample. A constant current is passed through the capillary from the high-voltage power supply, and the analyte ions segregate into individual, discrete zones. Analyte ions having the highest mobilities form a zone at the boundary near the leading buffer, whereas those with the lowest mobilities form a zone nearest the trailing buffer. In this way the analytes are "stacked" according to their mobilities, with the length of each zone determined by the total amount of each analyte present. An analyte present in a large amount in the sample, for example, will form a wide zone, while a trace component will form a narrow zone. In addition, the concentration of each analyte in its particular zone is proportional (based upon mobility factor) to the concentration of the leading buffer. Thus, CITP can be used to calibrate many analytes with only a single component concentration standard to determine the relationship between zone length and the amount of analyte.

Once the analyte zones come to equilibrium, both buffers and all analyte zones begin to migrate together to the detector, and they do so at the migration velocity of the leading buffer. The zones do not separate as in other CE modes, but remain adjacent to one another. The result is that a CITP electropherogram has an unusual step-function appearance (each step a different analyte zone) rather than the series of Gaussian-like peaks observed in the electropherograms for other CE modes. CITP is usually carried out in coated capillaries to eliminate electroosmotic flow; thus, CITP is typically used to determine anions and cations. Migration time cannot be used for analyte identification in CITP because it is a function of the number of analyte ions in the sample.

The ability of CITP to focus a trace component into a narrow, concentrated zone at a concentration level near that of the leading buffer makes it an ideal candidate for analysis of dilute samples, since they are, in effect, preconcentrated by this method. Once the analytes have been focused, additional leading buffer is loaded behind the terminating buffer to enable electrophoretic migration and analyte separation to occur. The time at which electrophoresis is initiated is important because the analytes begin to migrate after focusing occurs, and enough capillary length must remain for adequate separation. This method of focusing followed by normal electrophoresis is termed transient CITP. Focusing the analytes in this way allows very large sample volumes (>100-fold larger than other methods, filling 20–50% of the capillary volume) to be injected without loss of electrophoretic performance.

The use of transient CITP to focus and separate analytes is illustrated in Fig. 11, where the separation of various length oligomers of poly(deoxyadenylic acid) is demonstrated under varying CITP focusing conditions (Auriola *et al.*, 1996). The decrease in migration time and resolution observed in Fig. 11(d) occurred because the focusing voltage was applied beyond the point in time at which the analytes were focused. As a result, the focused analytes migrated through the capillary prior to sepa-

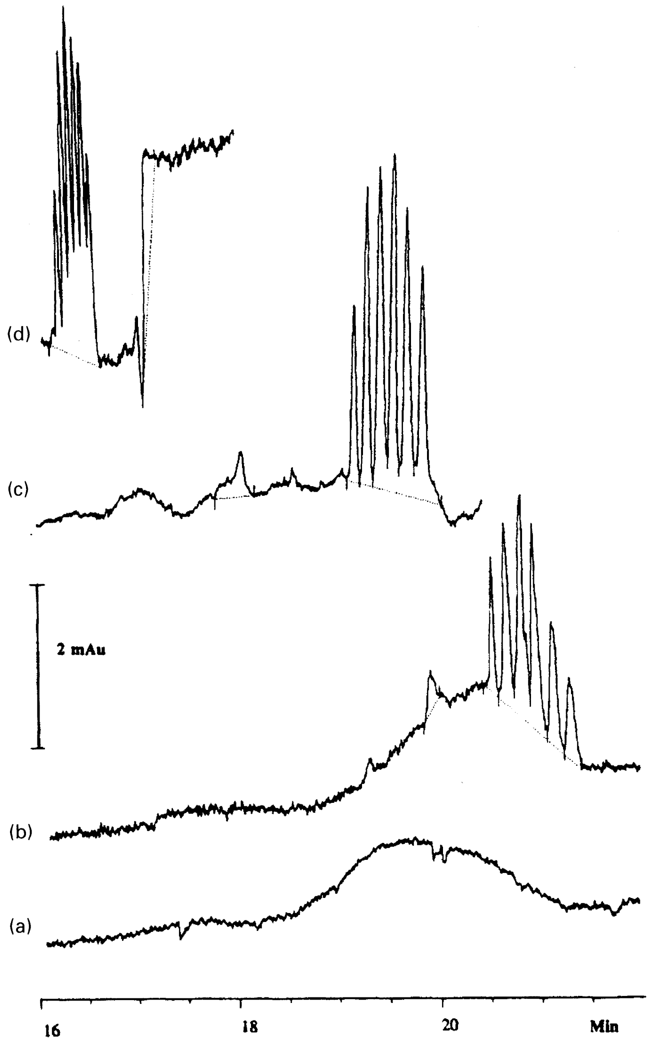

FIG. 11. Separation of poly(deoxyadenylic acid) pd(A)$_{19-24}$ oligomers with lengths of 19 through 24 units, using transient CITP. The effects of the following focusing times and voltages are shown: (a), no focusing; (b), 0.25 min; −15 kV; (c), 0.5 min, −30 kV; (d), 1.5 min, −30 kV. Sample introduction: hydrodynamic (pressure) injection, 48 s (10% of column length). Capillary: fused silica, untreated, 60 cm long ×50 μm i.d., 23 °C, −30 kV applied voltage for electrophoresis. Buffer: leading buffer 0.100 *M* ammonium formate, *p*H 4.5, containing 12% polyethylene glycol (average molecular weight 20 000) as a molecular sieving agent and to reduce EOF; terminating buffer 2-(*N*-morpholino)ethanesulfonic acid (MES) adjusted to *p*H 4.5. Detection: UV–vis absorbance at 254 nm. Elution order of analytes: pd(A)$_{19}$ through pd(A)$_{24}$ in order of increasing molecular weight. Reprinted (in part) from Auriola *et al.* (1996). Copyright (1996) American Chemical Society.

ration. This reduced the effective electrophoretic capillary length, adversely affecting electrophoretic efficiency.

3.6 Capillary Electrochromatography (CEC)

Capillary electrochromatography (CEC) is another separation method that utilizes both electrophoretic and chromatographic separation mechanisms. Similar to LC, separation in CEC is provided by the partitioning of analyte molecules between a mobile phase and a stationary phase, and it is well suited to neutral compounds. CEC uses electroosmotic flow rather than pressure to drive the mobile phase through the column. This results in better resolution than LC because it eliminates zone broadening due to parabolic flow profiles. Further, the small stationary-phase particles (1.5 μm diameter) required in LC to achieve resolutions approaching those obtained in CE require extraordinarily high pumping pressures, which are difficult to obtain in practice. Currently, CEC is only slightly (5–10 times) superior to traditional pumped-flow LC in terms of separation efficiency (N), and the expected narrow peaks are not observed on a routine basis. This has been theorized (Wan, 1997) to result from the effects of overlapping electrical double layers (on neighboring silica particles), which disturb the EOF profile. The magnitude of this perturbation decreases with increasing interparticle porosity, increasing particle diameter, and increasing buffer concentration.

In CEC, stationary phases consist of untreated or surface-modified silica particles that are typically 1.5 µm (or greater) in diameter. Untreated silica particles can contain ionizable surface groups that contribute to generation of EOF, while surface-modified particles reduce EOF in interparticle regions. The most common stationary phases are those based on siloxanes (see Chromatography), including particles that have been treated with long-chain–hydrocarbon functional groups to form $Si{-}O{-}Si(CH_3)_2{-}(CH_2)_{17}{-}CH_3$. These octadecylsilica particles, also known as C_{18} particles, are the most common stationary phase in CEC, although other particles where the C_{18} hydrocarbon group has been replaced with a more polar functional group (*e.g.*, phenyl, cyano, amino) can be used in the separation of more polar molecules. Particles can be loaded into the capillary as a dry powder or as a slurry, using either high-pressure liquid, high-pressure gas, electroosmosis, or electrophoresis. The stationary phases are retained in the capillaries by column-end frits of various designs. Common problems encountered with frits include bubble formation and zone broadening. Some research has been directed toward creating packing material by *in situ* polymerization, and such stationary phases do not require column-end frits. Pressurization of the capillary at the inlet and outlet is sometimes required to prevent solvent outgassing in CEC, which can perturb the EOF flow profile. Finally, electrokinetic injection is typically used in CEC because hydrodynamic injection is not usually capable of introducing the sample into the packed capillary.

Mobile phases commonly used in CEC are similar to those used in reversed-phase LC (see Chromatography), such as acetonitrile–water mixtures; however, a buffer salt is added to support electroosmosis and electrophoresis. Solvent gradients can be created by pressure programming, or by using a T-shaped inlet capillary in combination with two programmable power supplies to combine two different mobile phases dynamically (Yan *et al.*, 1996).

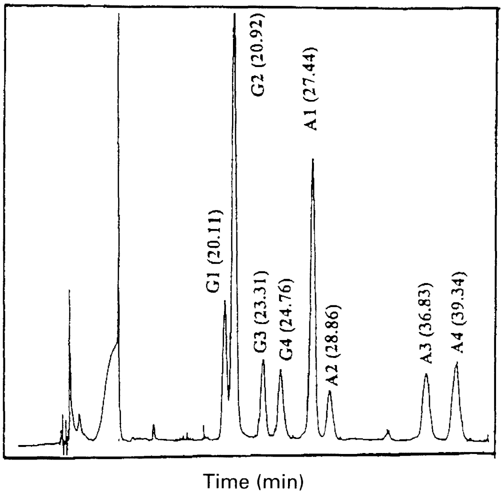

FIG. 12. CEC separation of eight diastereoisomeric deoxyguanosine and deoxyadenosine adducts of *anti*-benzo[g]chrysene. Sample introduction: electrokinetic injection. Capillary: fused silica, untreated, 40 cm long × 75 µm i.d., packed with 5 µm octadecylsilica particles, 15 kV applied voltage. Buffer: sample buffer 0.005 *M* ammonium acetate; chromatographic buffer 0.0085 *M* ammonium acetate in 29:71 acetonitrile:water. Detection: UV absorbance at 261 nm. Elution order of analytes: deoxyguanosine adduct diastereomers G1–G4 followed by deoxyadenosine adduct diastereomers A1–A4. Numbers in parentheses next to analyte peak labels are migration times in minutes. Reprinted (in part) from Ding and Vouros (1997). Copyright (1997) American Chemical Society.

CEC is often used to separate analytes that are difficult to resolve by LC, for neutral analytes not resolved by CZE, and for hydrophobic analytes not separated well by MECC. Here, CEC separations performed without surfactants (as are needed in MECC) make interfacing the capillary to some detectors, such as a mass spectrometer, less troublesome. Figure 12 shows the application of CEC to the separation of enantiomeric adducts of deoxynucleosides and anti-dihydrodiol epoxides of benzo[g]chrysene (Ding and Vouros, 1997). The relative amounts of the stereoisomers produced in the reaction between the deoxynucleoside and the epoxide provides information useful for mutagenic analysis. In this example, CEC separation results in nearly baseline resolution of all eight possible stereoisomers in the standard mixture.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy under Contract DE-AC04-94AL85000. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy.

GLOSSARY

Ampholyte, Amphotere: A compound that can exist as either a cation or an anion depending on buffer *p*H. In its basic form, an ampholyte is an anion, while in its acidic form, it exists as a cation.

Analyte: In capillary electrophoresis, the solute being determined.

Anode: In capillary electrophoresis, a positively biased electrode used in applying high electrical potentials across capillaries; typically immersed in the inlet buffer vial of a capillary elelctropherograph.

Anolyte: In capillary isoelectric focusing, buffer solution in the inlet (anode) vial of a capillary electropherograph.

Buffer: A solution whose composition is selected to minimize changes in *p*H that would otherwise occur as a result of chemical reactions. The *p*H is controlled by the relative concentrations of acidic and basic forms of ionizable solutes.

Capillary, Capillary Column: In capillary electrophoresis, the narrow-bore column of fused silica, Pyrex, or Teflon in which separation occurs.

Capillary Electrochromatography (CEC): Capillary electrophoresis mode whereby analytes are separated on the basis of differences in both electrophoretic migration rates and partitioning between the buffer and a chromatographic stationary phase.

Capillary Electropherograph: Instrumentation used in capillary electrophoresis.

Capillary Electrophoresis (CE): Technique for high-resolution separations of complex mixtures based on electrophoresis within small-bore capillaries.

Capillary Gel Electrophoresis (CGE): Capillary electrophoresis mode whereby analytes are separated on the basis of differences in both electrophoretic migration rates and molecular size.

Capillary Ion Analysis (CIA), Capillary Ion Electrophoresis (CIE): Capillary zone electrophoresis mode that is optimized for the analysis of small ionic species.

Capillary Isotachophoresis (CITP): Capillary electrophoresis mode used to separate analytes in dilute solutions on the basis of differences in their electrophoretic migration rates.

Capillary Isoelectric Focusing (CIEF): Capillary electrophoresis mode used to separate analytes on the basis of differences in isoelectric points.

Capillary Zone Electrophoresis (CZE): Capillary electrophoresis mode performed in open-tubular, uncoated capillaries whereby analytes are separated on the basis of differences in electrophoretic migration rates.

Cathode: A negatively biased electrode used in applying high electrical potentials across capillaries; typically immersed in the outlet buffer vial of a capillary electropherograph.

Catholyte: In capillary isoelectric focusing, the buffer solution in the outlet (cathode) vial of a capillary electropherograph.

Chromatography: Chemical separation and analysis technique based on differences in partitioning of an analyte between a solid or liquid stationary phase and a gas, liquid, or supercritical fluid mobile phase.

Dynamic Sieving Capillary Electrophoresis (DSCE): Capillary electrophoresis mode whereby analytes are separated on the basis of differences in both electrophoretic migration rates and molecular size.

Electroendosmosis, Electroendosmotic Flow, Electroosmotic Flow (EOF): Transport of neutral molecules or particles by a conducting solution under the influence of an electric field.

Electromigration: Movement of a charged species under the influence of an electric field. In electrophoresis, however, more generally used to describe the movement of all solutes, whether they be neutral or charged.

Electroosmotic Mobility: A measure of the velocity of a neutral species in an electric field of unit strength.

Electropherogram: In capillary electrophoresis, a plot of detector response as a function of migration time.

Electrophoresis: Migration, or movement, of electrically charged particles or solutes in a conductive liquid or gel under the influence of an applied electric field.

Electrophoretic Migration: See **Migration**.

Electrophoretic Mobility: A measure of the velocity of a charged species in an electric field of unit strength.

Elution: In capillary electrophoresis and chromatography, a term used to describe the process whereby the analyte exits from the separation column in a flowing medium.

Free-solution Capillary Electrophoresis (FSCE): See **Capillary Zone Electrophoresis**.

Gas Chromatography (GC): Chromatography with a gaseous mobile phase.

Gel: In slab gel electrophoresis, a two-phase colloidal system or a polymer used as a partitioning media. In capillary electrophoresis, a polymer containing a network of pores with a narrow, well-defined size distribution that is incorporated into the capillary either as a fixed, cross-linked polymer or as a non-cross-linked polymer solution.

Gel Electrophoresis: Separation and analysis technique, performed on gel slabs, where analytes are separated on the basis of differences in electrophoretic migration rates and then determined as separated zones on the slab.

Height Equivalent of a Theoretical Plate: In capillary electrophoresis and chromatography, a quantity used to describe the efficiency of a separation.

Ionic Strength: A term equal to one-half the sum of the products of concentration and squared charge for all the ionic species in solution. Denoted by the symbol I and expressed in terms of concentration units, *e.g.*, molarity.

Isoelectric Point: The *p*H of a solution in which an ampholyte or some other ionizable species is neutral and has zero electrophoretic mobility in an electric field. Denoted by the symbol *p*I.

Liquid Chromatography (LC): Chromatography with a liquid mobile phase. Also referred to as high-performance liquid chromatography.

Micellar Electrokinetic Capillary Chromatography (MECC): Capillary electrophoresis technique in which the analytes are partially solubilized by micelles added to the buffer, *i.e.*, they partition between the micellar and buffer phases. The degree to which each analyte partitions between the micelles and the buffer is the basis for separation.

Micelle: Aggregation of straight-chain molecules, each of which has both a hydrophilic end and a hydrophobic end. The aggregation occurs above a certain concentration, called the critical micelle concentration. Used in micellar electrokinetic capillary chromatography to add separation selectivity based on the degree of analyte partitioning between the buffer and micellar phases.

Migration: In electrophoresis, an abbreviated version of the term electromigration, used to describe the movement of both neutral and charged species under the influence of an electric field. Also referred to as electrophoretic migration.

Migration Time: In capillary electrophoresis, the length of time it takes an analyte to travel through the capillary.

Neutral Marker: In capillary electrophoresis, a molecule that is transported only by electroosmotic flow and can therefore be used to measure EOF mobility.

Number of Theoretical Plates: In capillary electrophoresis and chromatography, a measure of the efficiency of a particular separation scheme.

Resolution: In capillary electrophoresis and chromatography, a measure of the extent of separation achieved between two analytes.

Separation Factor: In capillary electrophoresis, the ratio of the electrophoretic mobilities of the two analytes. In chromatography, the ratio of the capacity factors (or partition coefficients) of the two analytes.

Size Exclusion Chromatography (SEC): Liquid chromatography technique that separates chemical compounds on the basis of their physical size.

Slab Gel Electrophoresis: See **Gel Electrophoresis**.

Solute: A species dissolved in a liquid.

Theoretical Plate: In capillary electrophoresis and chromatography, an imaginary unit in the capillary or column wherein one complete mechanistic separation process step occurs. For example, in micellar electrokinetic capillary chromatography, this would be a single equilibrium partitioning of the analyte between the micellar phase and the buffer phase.

Zeta Potential: In capillary electrophoresis, the potential difference across the shear plane between the layers of the electrical double layer formed at the capillary/buffer interface.

Zone: In capillary electrophoresis, a term used to describe the width and profile of an analyte band as it migrates through the capillary.

Zwitterion: A molecule containing both cationic and anionic functional groups.

Works Cited

Auriola, S., Jääskeläinen, I., Regina, M., Urtti, A. (1996), *Anal. Chem.* **68**, 3907–3911.

Carrilho, E., Ruiz-Martinez, M. C., Berka, J., Smirnov, I., Goetzinger, W., Miller, A. W., Brady, D., Karger, B. L. (1996), *Anal. Chem.* **68**, 3305–3313.

Chen, D. Y., Dovichi, N. J. (1994), *J. Chromatogr. B: Biomed. Appl.* **657**, 265–269.

Cohen, A. S., Karger, B. L. (1987), *J. Chromatogr.* **397**, 409–417.

Ding, J., Vouros, P. (1997), *Anal. Chem.* **69**, 379–384.

Ewing, A. G., Gavin, P. F. (1996), *J. Am. Chem. Soc.* **118**, 8932–8936.

Fan, Z., Harrison, D. J. (1991), *Anal. Chem.* **66**, 177–184.

Fluri, K., Fitzpatrick, G., Chiem, N., Harrison, D. J. (1996), *Anal. Chem.* **68**, 4285–4290.

Friedl, W., Reijenga, J. C., Kenndler, E. (1995), *J. Chromatogr. A.*, **709**, 163–170.

Ganzler, K., Warne, N., Hancock, W. S. (1995), in: P. G. Rhigetti (Ed.), *Capillary Electrophoresis in Analytical Biotechnology*, Boca Raton, FL: CRC Press.

Hardy, W. B. (1899), *J. Physiol.* **24**, 288.

Hjerten, S. (1967), *Chromatogr. Rev.* **9**, 122.

Hjerten, S. (1985), *J. Chromatogr.* **347**, 191–198.

Jones, W. R., Soglia, J., McGlynn, M., Haber, C., Reineck, J., Krstanovic, C. (1996), *Am. Lab.* **28** (5), 25–33.

Jorgenson, J. W., Lukacs, K. D. (1981), *Anal. Chem.* **53**, 1298–1302.

Jorgenson, J. W., Lukacs, K. D. (1983), *Science* **222**, 266–272.

Karger, B. L., Foret, F. (1993) in: N. A. Guzman (Ed.), *Capillary Electrophoresis Technology*, New York: Marcel Dekker, Inc., p. 17.

Lada, M. W., Kennedy, R. T. (1996), *Anal. Chem.* **68**, 2790–2797.

Lauer, H. H., McManigill, D., (1986), *Anal. Chem.* **58**, 166–170.

Liu, J. P., Hsieh, Y. Z., Wiesler, D., Novotny, M. (1991), *Anal. Chem.* **63**, 408–412.

Ma, S., Hodel, A., Fox, R. O., Horvath, C. (1995), *Electrophoresis* **16**, 595.

Manz, A., Verpoorte, E., Effenhauser, C. S., Burggraf, N., Raymond, D. E., Harrison, D. J., Widmer, H. M. (1993), *J. High Resolut. Chromatogr.* **16**, 433–436.

Michaelis, L. (1909), *Biochem. Z.* **16**, 81.

Mikkers, F. E. P., Everaerts, F. M., Verheggen, Th. P. E. M. (1979), *J. Chromatogr.* **169**, 11–20.

Minarik, M., Gas, B., Rizzi, A., Kenndler, E. (1995), *J. Cap. Electrophor.*, **2**, 89–96.

Quincke, G. (1861), *Ann. Phys. (Leipzig)* **113**, 513.

Reuss, F. F. (1809), *Mem. Soc. Imp. Natur. Moskau* **2**, 327.

Riester, D., Wiesmuller, K. H., Stoll, D., Kuhn, R. (1996), *Anal. Chem.* **68**, 2361–2365.

Shimura, K., Karger, B. L. (1994), *Anal. Chem.* **66**, 9–15.

Small. H., Miller, T. E. (1982), *Anal. Chem.* **54**, 462–469.

Small, H., Stevens, T. S., Bauman, W. C. (1975), *Anal. Chem.* **47**, 1801–1809.

Smith, R. D., Barinaga, C. J., Udseth, H. R. (1988), *Anal. Chem.* **60**, 1948–1952.

Swanek, F. D., Chen, G., Ewing, A. G. (1996), *Anal. Chem.* **68**, 3912–3916.

Swerdlow, H., Gesteland, R. (1990), *Nucl. Acids Res.* **18**, 1415.

Terabe, S., Otsuka K., Ichikawa, K., Tsuchiya, A., Ando, T. (1984), *Anal. Chem.* **56**, 111.

Tiselius, A. (1937), *Trans. Faraday. Soc.* **33**, 524.

Wahl, J. H., Smith. R. D. (1994), *J. Cap. Electrophor.* **1**, 62–71.

Wallingford, R. A. (1996), *Anal. Chem.* **68**, 2541–2548.

Wan, Q. H. (1997), *Anal. Chem.*, **69**, 361–363.

Yan, C., Dadoo, R., Zare, R. N., Rakestraw, D. J., Anex, D. S. (1996), *Anal. Chem.* **68**, 2726–2730.

Zhong, M., Lunte, S. M. (1996), *Anal. Chem.* **68**, 2488–2493.

Further Reading

Baker, D. R. (1995), *Capillary Electrophoresis*, New York: John Wiley and Sons.

Camilleri, P. (Ed.) (1993), *Capillary Electrophoresis: Theory and Practice*, Boca Raton, FL: CRC Press.

Coleman, D. (Ed.) (1994), *Directory of Capillary Electrophoresis*, Amsterdam: Elsevier.

Foret, F., Krivankov, L., Bocek, P. (1994), *Capillary Zone Electrophoresis*, New York: VCH Publishers, Inc.

Guzman, N. A. (Ed.) (1993), *Capillary Electrophoresis Technology*, New York: Marcel Dekker, Inc.

Hartwick, R. A. (1994), *Introduction to Capillary Electrophoresis*, Boca Raton, FL: CRC Press.

Landers, J. P. (Ed.) (1997), *Handbook of Capillary Electrophoresis*, 2nd ed., Boca Raton, FL: CRC Press.

St. Claire, R. L. (1996), "Capillary Electrophoresis," *Anal. Chem.* **68**, 569R–586R.

Weinberger, R. (1993), *Practical Capillary Electrophoresis*, Boston: Academic Press.

Yang, Q., Hidajat, K., Li, S. F. Y., (1997) "Trends in Capillary Electrophoresis," *J. Chromatogr. Sci.* **35**, 358–373.

DATING TECHNIQUES

Alice C. Mignerey, *Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland, U.S.A.*

Introduction 35
1. **Basic Principles** 36
1.1 Radioactive Decay 36
1.2 Sources of Radionuclides 37
1.2.1 Primordial Radionuclides 37
1.2.2 Cosmogenic Radionuclides 37
1.2.3 Noncosmogenic *In Situ* Production 38
1.2.4 Anthropogenic Radionuclides 38
1.3 Detection Techniques 38
1.3.1 Traditional Decay Counting 38
1.3.2 Traditional Mass Spectrometry 40
1.3.3 Accelerator Mass Spectrometry 40
1.4 Thermoluminescence and Electron-Spin Resonance 41
2. **Techniques Based on Primorial Radionuclides** 42
2.1 Equilibrium Techniques 42
2.1.1 Rb–Sr and the Isochron Method 42
2.1.2 Sm–Nd 44
2.1.3 Lu–Hf 44
2.1.4 Re–Os 44
2.1.5 K–Ar, Ar–Ar 44
2.1.6 U–Pb, Th–Pb, Pb–Pb 45
2.2 Nonequilibrium Techniques and the U–Th Decay Chains 47
2.2.1 ^{238}U–^{234}U–^{230}Th System: Dating of Corals and Ocean Sediments 47
2.2.2 ^{235}U–^{231}Pa 48
2.2.3 ^{210}Pb 48
2.3 Extinct Radionuclides 49
2.4 Fission-track methods 49
2.5 (U–Th)/He Thermochronometry 50
2.6 Thermoluminescence (TL) Dating Applications 50
3. **Techniques Based on Cosmogenic Nuclides** 51
3.1 Carbon-14 51
3.2 *In Situ* Dating of Geological Formations 54
3.3 Ice Cores, Ocean and Soil Studies 57
3.3.1 Atmospheric Beryllium-10 in the Lithosphere and Hydrosphere 57
3.3.2 Aluminum-26 in the Oceans 57
3.4 Groundwater Dating 57
3.4.1 Natural Sources 57
3.4.2 Anthropogenic Sources 58
3.5 Silicon-32: The Potential 59
4. **Paleomagnetism** 59
5. **Aminostratigraphy and Aminochronology** 59
6. **Chlorofluorocarbons** 60
7. **Conclusion** 61
Glossary 62
Works Cited 63
Further Reading 65

INTRODUCTION

The ability to assign a time to an object's formation or to an event on historic, prehistoric, or geologic timescales relies on a set of techniques that are commonly termed dating. In practice, this can mean a variety of things depending on the field of investigation, but it usually refers to the chronological date before the present. The most widely used dating tool is the transformation of elements via radioactive decay. Depending on the choice of radioactive nuclides, times from a few days or years to the age of the solar system can be explored. Techniques based on chemical tracers and atomic transformations have also proven important as dating tools in specific applications.

The science of dating techniques is central to a broad range of disciplines, including geomorphology, geochemistry, tectonics, cosmology, art history and archaeology, paleontology, hydrology, oceanography, environmental

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

science, and even criminology. The field has expanded rapidly in recent years with advances in the ability to quantify very small amounts of even the rarest of nuclides. This article attempts to give a broad overview of the field as it is currently practiced, with specific references to some of the more recent work.

The first section contains brief discussions of the basic principles of radioactive decay, sources of radionuclides, and the techniques of mass spectrometry, thermoluminescence, and electron-spin resonance. The subsequent sections describe specific applications in a variety of fields. Section 2 is devoted to the uses of the primordial radionuclides. The text briefly presents each nuclide or system of nuclides with regard to its origin, uses, and limitations. Section 3 focuses on the applications of cosmogenic and anthropogenic nuclides and is aligned according to application, since a number of nuclides can be applied to a given problem. Each section has been augmented with suggested reading to provide more in-depth treatments in specific areas. Summary tables are provided to correlate nuclides with various areas of research. The interested reader is encouraged to explore the titles in the Further Reading list for discussions on the wide variety of applications.

1. BASIC PRINCIPLES

1.1 Radioactive Decay

The general term radioactive decay is applied to any process by which one nuclide transforms into another via the emission of an α particle or a β particle (positive or negative), fission, electron capture, and even, in rare cases, emission of a proton (see RADIOACTIVITY). This transformation follows a first-order rate law, with the rate of change of the number of radionuclide atoms N given by

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \lambda N, \tag{1}$$

where the decay constant λ is ln2 divided by the nuclide half-life, $t_{1/2}$. In Eq. (1), $\mathrm{d}N/\mathrm{d}t$ is termed the activity and expressed in disintegrations per unit time. Integrating Eq. (1) with respect to time gives the relationship between the number of initial atoms N_0 and the number remaining after time t,

$$N = N_0 \mathrm{e}^{-\lambda t}. \tag{2}$$

Knowing the number of atoms N and with an estimate of the initial N_0, the time t over which the system has been decaying can be calculated. For example, after a time $t = t_{1/2}$, N will be equal to $\frac{1}{2}N_0$.

Alternatively, if the daughter nuclide is stable, the number of daughter atoms present at time t, D_t, is given by the number of parent atoms which have decayed in time t, $P_0 - P_t$, plus any initial daughter atoms D_0 present at time $t = 0$:

$$D_t = D_0 + (P_0 - P_t) = D_0 + P_t(e^{\lambda t} - 1). \tag{3}$$

The time t can then be obtained as

$$t = \frac{1}{\lambda}\ln\left(1 + \frac{D_t - D_0}{P_t}\right). \tag{4}$$

Equations (3) and (4) are valid only for closed systems (with no loss of parent or daughter or intermediate nuclides.)

Sometimes the parent is connected to the stable daughter through a series of shorter-lived radionuclides, such as in the ^{238}U–^{206}Pb series shown in Fig. 1. Here the parent ^{238}U decays to ^{234}Th via the emission of an α particle. The ^{234}Th goes on to decay to ^{234}Pa via β^- emission, and so on until the stable end product ^{206}Pb is reached. This chain involves a total of eight α-particle and six β^- decays. Provided the decay time of the parent is long compared with those of the intermediate products, Eqs. (3) and (4) can still be used to connect the initial parent and final stable daughter nuclides.

In the case where the measured daughter is itself decaying, the parent–daughter relationship becomes

$$D_t = D_0 \mathrm{e}^{-\lambda_2 t} + \frac{\lambda_1}{\lambda_2 - \lambda_1} P_0(e^{-\lambda_1 t} - e^{-\lambda_2 t}), \tag{5}$$

with λ_1 and λ_2 being the parent and daughter decay constants, respectively. If the parent half-life is longer than the daughter half-life, an equilibrium situation is reached after a

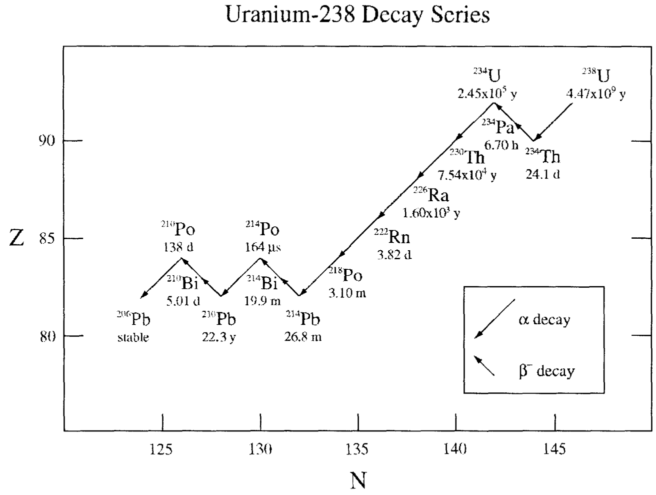

FIG. 1. Schematic diagram of the the ^{238}U decay series, displayed on a grid of proton number Z *versus* neutron number N.

few daughter half-lives, where the activity ($A = \lambda N$) ratio of the parent to the daughter, A_1/A_2, is given by $(\lambda_2 - \lambda_1)/\lambda_2$. In the case of a very long-lived parent (such as ^{238}U compared with ^{234}U in Fig. 1), $\lambda_2 \gg \lambda_1$ and $A_1 = A_2$. This is called secular equilibrium and can be extended to the activity of any nuclide in the chain, $A_1 = A_2 = A_3 \ldots.$

1.2 Sources of Radionulides

1.2.1 Primordial Radionuclides At the condensation of the solar system from the solar nebula there were a large number of unstable isotopes of the elements that are present on earth today, and even a few higher-Z elements no longer present. Most of these have since decayed to the stable nuclides that make up our natural isotopic abundances. The few exceptions are those with half-lives on the order of, or greater than, the age of the Earth (about 4.5×10^9 y). These long-lived radionuclides form the group called the primordial or primary radionuclides, because they have been around since the beginning of the solar system. Those commonly used for dating are given in Table 1, along with their half-lives.

1.2.2 Cosmogenic Radonuclides Cosmogenic nuclides are constantly being formed through nuclear reactions between the cosmic-ray flux and various constituent atoms of the Earth's atmosphere, hydrosphere, and lithosphere. Cosmic rays consist mostly of high-energy protons, about 10% α particles and 1% heavier nuclides. They originate from the Sun and from outside our solar system. The portion from the Sun shows a definite modulation, which has been correlated with sunspot activity. The Earth's magnetic field tends to focus the cosmic rays at the poles, giving a latitudinal variation in intensity. The energy distribution of the incident or primary cosmic rays is very broad and peaks around 300 MeV. This is sufficient energy to produce numerous secondary products from a single primary collision, in what is commonly termed a shower. These secondaries also include neutrons, which can interact even at very low energies via capture reactions. Lal and Peters (1967) contains an in-depth characterization of the cosmic-ray flux, its spatial distribution in height and latitude, and the resulting reactions.

The atmosphere is quite efficient at reacting with cosmic rays, with about $\frac{2}{3}$ of the atmospheric cosmogenic nuclides produced in the stratosphere and $\frac{1}{3}$ in the troposphere. Those produced in the stratosphere must first make it to the troposphere before they are removed via both wet and dry deposition.

The cosmic-ray flux that reaches the Earth's surface is greatly reduced in intensity by interactions with the atmosphere. Nevertheless, significant cosmogenic nuclides are produced by interactions of cosmic-ray components with target nuclei in the oceans, and in

Table 1. Primordial radionuclides: parent–stable daughter systems.

Parent–daughter	Parent half-life[a] (y)	Parent natural abundance (%)[a]	Applications
^{40}K–$^{40}Ar(10.72\%)$–$^{40}Ca(89.28\%)$	1.277×10^9	0.0117	Micas, volcanic rocks, glauconite
^{87}Rb–^{87}Sr	4.75×10^{10}	27.835	Sedimentary rocks, lunar rocks, volcanic and plutonic rocks, authigenic clays, meteorites
^{147}Sm–^{143}Nd	1.06×10^{11}	15.0	Meteorites, lunar rocks, archean volcanites, zircons, mantle evolution
(^{148}Sm–^{144}Nd)	7×10^{15}	11.3	
(^{144}Nd–^{140}Ce)	2.29×10^{15}	23.80	
^{176}Lu–^{176}Hf	3.73×10^{10}	2.59	Meteorites, apatites, garnets, monazites
^{187}Re–^{187}Os	4.35×10^{10}	62.60	Mantle separations, sulfides, iron meteorites
^{232}Th–^{208}Pb	1.405×10^{10}	100	calcite, zircon and rutile granites, meteorites, crustal evolution
^{235}U–^{207}Pb	7.04×10^8	0.720	
^{238}U–^{206}Pb	4.468×10^9	99.2745	

[a] From Nuclear Wallet Cards, July 1995, National Nuclear Data Center, Brookhaven National Laboratory, P. O. Box 5000, Upton, NY 11973–5000, U.S.A.

surface soils and rocks. Since cosmic rays are strongly absorbed in the surface, these *in situ* produced nuclides provide important means of studying erosion rates and exposure histories for a variety of geomorphic surfaces (Cerling and Craig, 1994; Bierman *et al.*, 1995). Figure 2 shows a representative depth profile for the neutron flux generated by cosmic-ray interactions in concrete (Liu *et al.*, 1994). Note that the flux peaks slightly below the surface. Only reactions induced by muons are important at significant depths.

1.2.3 Noncosmogenic *In Situ* Production A low flux of slow neutrons can also be generated by the reaction of α particles from U and Th with light elements such as Ca and K in the Earth via (α, n) reactions. These neutrons can be captured by ^{35}Cl to produce ^{36}Cl or by ^{235}U to induce fission. These are *in situ* processes that are not cosmogenic in origin, and are important in U- and Th-bearing deposits.

1.2.4 Anthropogenic Radionuclides The advent of the nuclear age has provided a new source of radionuclides. The atmospheric nuclear tests of the late 1950s and early 1960s spread neutron-activation and fission products around the globe that are now serving as tracers for processes on the order of 30–40 y. Examples include tritium (^3H), ^{36}Cl, and ^{129}I. This source is termed the bomb pulse and produced concentrations of tritium and ^{36}Cl in rainwater and ice cores several orders of magnitude above the prebomb background. This effect is clearly shown in Fig. 3. Man's continued activities with nuclear power reactors and nuclear fuel reprocessing have significantly elevated the global inventory of ^{129}I, producing point sources for tracing modern groundwaters and ocean circulation patterns (Rucklidge, 1995; Raisbeck *et al.*, 1994). There is also evidence for local elevated concentrations of ^{36}Cl in surface and shallow ground water from nuclear fuel reprocessing plants and reactor emissions.

1.3 Detection Techniques

1.3.1 Traditional Decay Counting The ability to determine the number of parent and daughter nuclides at a given time, or the activity due to their decay, is a necessary requirement of most nuclear dating methods. This can be achieved by measuring the nuclide directly by mass spectrometric methods or by counting the number of β or α particles emitted (or the associated γ-ray emission) in the decay and using Eq. (1) to obtain the activity. Activity can then be converted back to the number

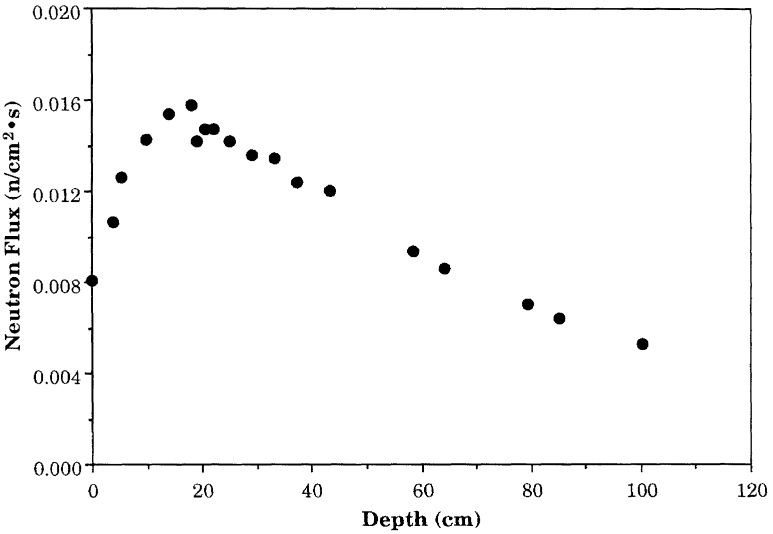

FIG. 2. Measured depth profile of neutrons produced in concrete (data from Lui *et al.*, 1994).

of atoms N or, in some cases, used directly. For example, the widely used radionuclide ^{14}C ($t_{1/2} = 5730\,y$) is traditionally quantitated by counting its β^- decays. This is possible because the average activity of "modern" prebomb carbon in the biosphere is about 13.6 decays per minute per gram carbon; counting rates above background are obtained when sufficient carbon is available. If the nuclide is very long-lived, but has a measurable abundance, then the activity is also measurable by virtue of the much greater N. This is the case with many of the nuclides in the uranium and thorium decay series. However, decay counting becomes less feasible, if not impossible, for much rarer nuclides with longer half-lives. Reviews of nu-

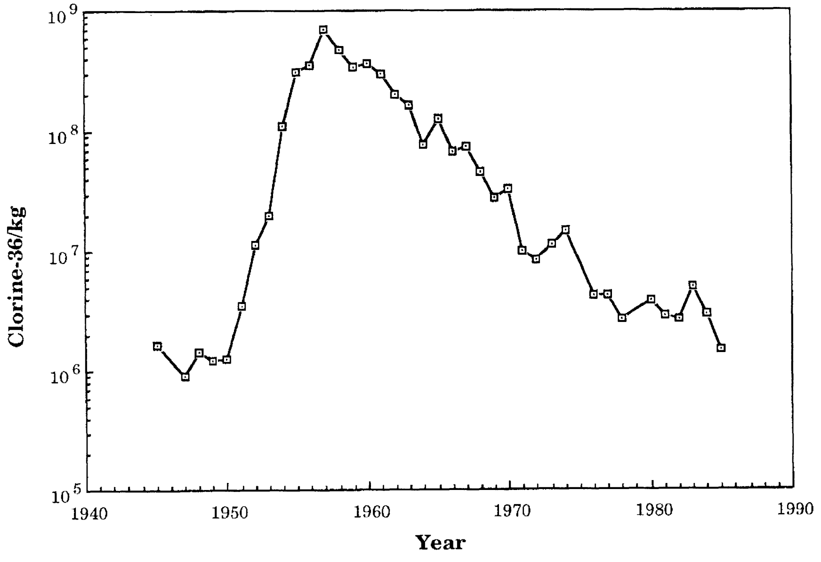

FIG. 3. Measured ^{36}Cl concentration in ice cores from Dye-3 (data from Synal *et al.*, 1990).

clear detection techniques can be found in this Encyclopedia and in Knoll (1979).

1.3.2 Traditional Mass Spectrometry Modern mass spectrometry has revolutionized the field of nuclear dating. While the ability to measure time comes from the radioactive decay clock, it is no longer necessary to count these decays to obtain an accurate estimate of the number N in Eq. (2). Only a brief overview of the technique is given here. For a more complete discussion of traditional mass spectrometric techniques the reader is referred to MASS SPECTROMETRY.

The basic principle of a mass spectrometer is the fact that accelerated ions in a uniform magnetic field will bend with radii of curvature squared, r^2, proportional to their mass to charge ratio m/q. If a sample of an element is ionized to a single charge state, then the individual isotopes of that element will appear at different positions in the focal plane of the magnet and can be individually identified. Dickin (1995) discussed the requirements as applied to measurements of geological samples, including chemical separations necessary prior to analysis and preferred means of ionization and detection. Thermal-ionization mass spectrometry (TIMS) is the most widely used means of ionization today.

The analytical quantification of the individual isotope concentration is usually achieved through isotope dilution. In this technique the sample is measured with a spike of enriched isotope of known concentration. Comparison of the results for the pure sample, the pure spike and mixtures of the two allows the initial isotope concentration to be determined. This has the advantage that any matrix effects cancel because two isotopes of the same element are compared directly.

1.3.3 Accelerator Mass Spectrometry The technique of accelerator mass spectrometry (AMS) follows the basic principles of mass spectrometry, except that in AMS the ions are accelerated to much higher energies, as high as tens of MeV. This is most often carried out in the two stages of a tandem accelerator, similar to that depicted in Fig. 4. In the first stage

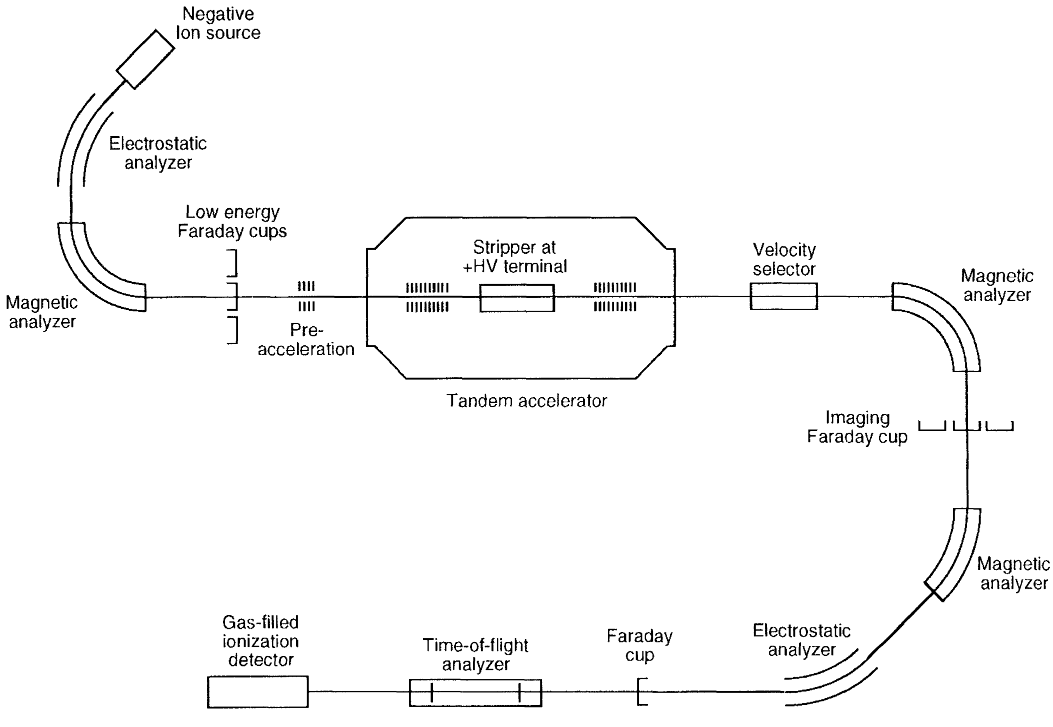
Negative
Ion source
Electrostatic
analyzer
Low energy
Faraday cups
Magnetic
analyzer
Pre-
acceleration
Stripper at
+HV terminal
Tandem accelerator
Velocity
selector
Magnetic
analyzer
Imaging
Faraday cup
Magnetic
analyzer
Electrostatic
analyzer
Faraday
cup
Time-of-flight
analyzer
Gas-filled
ionization
detector

FIG. 4. Schematic diagram for an accelerator system used for accelerator mass spectrometry (AMS).

negative ions from the ion source are electrically and magnetically selected for the desired mass (for example, 14 for ^{14}C determination). In some applications the negative-ion injection has an advantage in that interfering isobars (nuclides with the same mass but different nuclear charge Z) may not make negative ions (as is the case for ^{14}N in the ^{14}C determination) However, this restriction can also prove problematic for the determination of very electropositive species, such as ^{41}Ca, or noble gases such as Ar.

The negative ions are accelerated through a potential gradient to energies of 3–10 MeV (depending on the terminal voltage) at the central terminal of the accelerator. At this point the ions are passed through a stripper foil or gas and transformed into positive ions. The result is a distribution of positive charge states, which is a function of the initial ion identity and ion velocity. At this stage any residual molecular species which may have survived the initial ionization process are destroyed. The positive ions are then further accelerated back to ground potential at the exit of the accelerator.

The resultant ions exiting the accelerator have a spread in energies due to multiple charge states produced in the central stripper. These are separated in an electrostatic analyzer and/or velocity filter. An additional magnetic analyzer further reduces the background, selecting only those ions of a given mass (which now ideally have the same charge state.) This set is not necessarily unique and may still have a mix of isobars. The next major advantage of AMS is the detection of the product ions using nuclear detection techniques. These techniques are capable of counting individual atoms and separating them by their elemental Z. For example, modern AMS systems can separate ^{36}Cl from the much more common ^{36}S and measure 1 atom of ^{36}Cl in the presence of 10^{15} atoms of stable chlorine atoms (^{35}Cl and ^{37}Cl). A more thorough discussion of the technique of AMS can be found in Elmore and Phillips (1987) and Tuniz *et al.* (1998).

Accelerator mass spectrometry has opened up the area of cosmogenic nuclides to a broad range of applications, making possible studies with isotopes which are too rare or in concentrations too low to have been measured by decay counting. In the case of the widely used ^{14}C, AMS now allows measurements of much smaller samples and extends the technique to much older dates. The main disadvantage of AMS is the cost involved. Historically developed on machines built for basic nuclear physics research, the field has now come into its own with many laboratories acquiring much smaller accelerators that have been optimized for the purpose of AMS. One such facility is the National Oceans Science AMS facility (NOSAMS) at Woods Hole Oceanographic Institute, Massachusetts, U.S.A., dedicated to the measurement of ^{14}C. There are currently over 30 facilities worldwide with AMS capabilities.

1.4 Thermoluminescence and Electron-Spin Resonance

Nuclear radiation is typically called ionizing radiation because it is of sufficient energy to eject bound electrons from substances, producing free electrons and positive ions. The techniques of thermoluminescence (TL) and electron-spin resonance (ESR) rely on the trapping of these electrons by defects in the crystal structures of nonconducting solids. These defects can be vacancies, displacements (sometimes caused by the ionizing radiation itself), or impurities in the lattice, which provide sites for free electrons. The electrons usually remain trapped until dislodged by thermal vibrations (heating) or photoexcitation by ultraviolet light (as in solar bleaching). Once dislodged the electrons can diffuse through the crystal until they either become retrapped in a stable (deeper) site or recombine with a positive ion. If the recombination is radiative, light is emitted whose wavelength is characteristic of the type of luminescence center or crystal impurity. This process is called thermoluminescence (TL).

Potassium-, uranium-, and thorium-bearing minerals in the surrounding soils, or in the sample itself, provide a relatively constant flux of ionizing radiation. Therefore, the object to be dated has been accumulating dose over the period of time since it was last heated to a temperature (or exposed to the Sun's UV radiation) that released the electrons from any trapping sites and reset the TL clock. If we assume that the TL light obtained is propor-

Table 2. Primordial radionuclides: disequilibrium systems.

Nuclides	Half-lives[a]	Applications
^{238}U–^{234}U–^{230}Th	4.468×10^9 y–2.455×10^5 y–7.538×10^4 y	Carbonates: corals, shells; sediments; ferromanganese nodules/crusts
^{235}U–^{231}Pa	7.038×10^8 y–3.276×10^4 y	Carbonates, sediments
^{210}Pb–^{210}Bi–^{210}Po–^{206}Pb	22.3 y–5.013 d–138.376 d–stable	Deposition rates: snow, glaciers, sediments

[a] From Nuclear Wallet Cards, July 1995, National Nuclear Data Center, Brookhaven National Laboratory, P.O. Box 5000, Upton, NY 11973–5000, U.S.A.

tional to the accumulated paleodose dose D_P, the time t since the last resetting of the TL clock can be calculated by

$$t = D_P/D_R. \tag{6}$$

The annual dose rate D_R is a sum of the doses due to the α, β, and γ radiations from the nuclear decays. If the sample has been near the surface for any length of time, there will also be a background flux from cosmic-ray reactions. The specific contributions of each component in a sample depend on a number of factors, including burial history and soil and sample water content (which attenuates the radiation.) When possible the γ flux is measured at the excavation site.

In practice, a material to be TL dated is sequentially heated. Different types of traps release their electrons at different temperatures, termed glow peaks, those with the highest temperatures being the most stable. By measuring the number of photons of each type emitted as a function of temperature, the total radiation dose D_P can be determined if the trapping efficiencies are known. This is usually calibrated by making TL measurements with γ-ray sources at known doses.

Since TL requires heating the sample to high temperatures, it is unsuitable for a number of materials which decompose upon heating. An alternative is the method of electron-spin resonance (ESR), which measures the microwave absorption by unpaired electrons in the material when it is placed in a varying magnetic field. Although not as sensitive as TL, ESR has the advantage of measuring the trapped electrons in their trapping centers, and thus does not require luminescence centers or annealing the material. This means that the same sample can be measured numerous times and used to calibrate the dose by adding additional radiation through exposure to γ sources. Applications include tooth enamel, bone, and calcite (Aitken, 1990). The reader is referred to PHYSICS IN ARCHAEOLOGY for additional discussions of TL and ESR.

2. TECHNIQUES BASED ON PRIMORDIAL RADIONUCLIDES

This section contains an overview of the major applications of primordial radionuclides in use today. The parent–daughter systems are summarized in Tables 1 and 2. Many of these utilize what is called the isochron method for age determination. This method is presented in Sec. 2.1.1 in relation to the Rb–Sr system. Since it is not possible to discuss each system in great depth, the reader is referred to Faure (1986), Titayeva (1994), and Dickin (1995) for details on applications of interest.

2.1 Equilibrium Techniques

2.1.1 Rb–Sr and the Isochron Method The element rubidium has two naturally occurring isotopes, ^{85}Rb and ^{87}Rb, with abundances of 74.17 and 27.83%, respectively. The dating capabilities come from the radioactive ^{87}Rb, which decays with a half-life of 4.75×10^{10} y by β^- emission to ^{87}Sr (natural abundance 7.00%). While rubidium has no minerals of its own, the ionic radius of the Rb^+ ion is close to that for the K^+ ion. Since both are alkali metals, Rb^+ replaces K^+ in all potassium-bearing minerals. The product Sr isomorphously replaces Ca in some minerals, but differs in coordination number in others.

This results in considerable variation in the Rb/Sr ratio between formations with different histories and even between different minerals in the same formation. The differentiation of Rb and Sr in magmatic melts produces Rb/Sr ratios which vary from 0.06 in basalts to 0.53 in granite, with an average value of 0.23 for the Earth's crust (Titayeva, 1994).

In principle, the time since the formation of a rock or mineral can be found by Eq. (3). In the Rb–Sr case, D_t = number of ^{87}Sr atoms in the sample, D_0 = number of initial ^{87}Sr atoms and P_t = number of ^{87}Rb atoms in the sample. However, measuring absolute abundances is very difficult. A way around this problem is to normalize everything to a nearby isotope with the same approximate abundance, and measure all three by mass spectrometry. In this case, ^{86}Sr (natural abundance of 9.86%) is chosen as the normalizing isotope, and Eq. (3) becomes

$$\left(\frac{^{87}\mathrm{Sr}}{^{86}\mathrm{Sr}}\right)_{\mathrm{sample}} = \left(\frac{^{87}\mathrm{Sr}}{^{86}\mathrm{Sr}}\right)_{\mathrm{initial}} + \frac{^{87}\mathrm{Rb}}{^{86}\mathrm{Sr}}(e^{\lambda t} - 1) \quad (7)$$

The use of Eq. (7) requires knowledge of the initial $^{87}\mathrm{Sr}/^{86}\mathrm{Sr}$ ratio. This is difficult to estimate on its own, and today determination through the construction of isochrons is the method of choice. Inspection of Eq. (7) shows that it has the form of a straight line. If a number of samples with the same age but different Rb/Sr ratios are plotted on a graph of $^{87}\mathrm{Sr}/^{86}\mathrm{Sr}$ versus $^{87}\mathrm{Rb}/^{86}\mathrm{Sr}$, the slope $e^{-\lambda t} - 1$ will give the age and the intercept gives the initial $^{87}\mathrm{Sr}/^{86}\mathrm{Sr}$ ratio of the sample, a quantity of geological interest in itself. In this case the resulting line is a true isochron. Because of the large natural abundance of the decay product ^{87}Sr and the long half-life of the parent ^{87}Rb, this technique works best when applied to samples with high Rb/Sr ratios.

Just because the data plot on a line does not necessarily mean that they are derived from samples which are the same age. One test of a true isochron is to calculate the mean square weighted deviates (described in Dickin, 1995). The MSWD, as it is commonly referred to, gives a measure of the scatter of the data points about the straight line. If the scatter is larger than predicted by the analytical uncertainties of the measurements, then the MSWD > 1. The larger the MSWD, the less likely it is that the line formed is a true isochron. This pitfall is demonstrated in Fig. 5 from Innocent *et al.* (1997), showing Rb/Sr isochrons from the dating of celadonites, green clays close to glauconite. The data in Fig. 5(a) are derived from a group of whole minerals, giving an effective age for the group of 96 ± 1 million years. However, the large MSWD of 44 suggests that there are large variations in the individual ages of the minerals, even though the group defines a single line. An alternative method is to construct an internal isochron, derived from the leachates and residues of a single mineral. Figure 5(b) is an example where the ratios for two leachates (enriched in ^{86}Sr) and the residue (depleted in ^{86}Sr, but enriched in radiogenic ^{87}Sr) are plotted along with the whole-mineral ratios. The

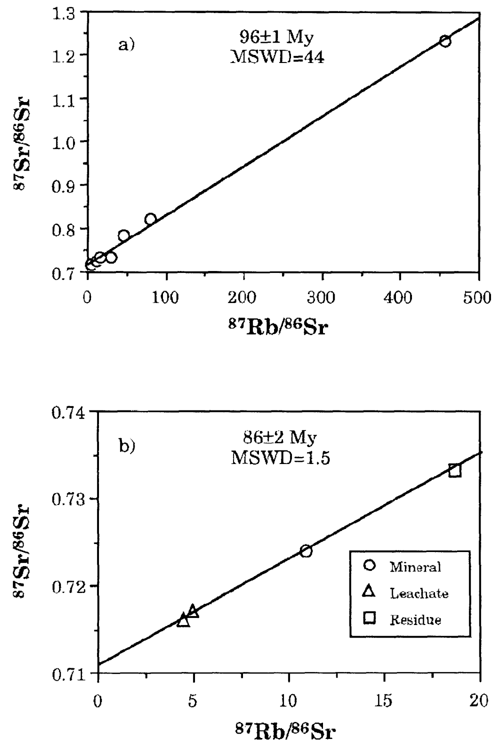

FIG. 5. Rb/Sr isochrons for (a) whole-mineral samples and (b) leachates, whole mineral, and residues for a single sample (data from Innocent *et al.*, 1997).

data clearly define a line with an extracted age of 86 ± 2 millions years for the sample (with a corresponding MSWD of only 1.5).

2.1.2 Sm–Nd There are seven naturally occurring isotopes of samarium, three of which are radioactive (see Table 1). The shortest-lived ^{147}Sm is the only one useful for dating, with a half-life of 1.06×10^{11} y and a natural abundance of 15.0%. It decays by α-particle emission to ^{143}Nb (abundance 12.18%). The isotope of choice for normalization in an equation analogous to Eq. (7) is ^{144}Nd, with a natural abundance of 23.80%. (Although ^{144}Nd is itself radioactive, its half-life is so long that is can be considered stable in this context.) Since both the parent and daughter are lanthanides, they have very similar chemistries and are not differentiated by secondary processes. This gives much more uniform natural Sm/Nd ratios in samples than is seen in Rb–Sr dating, and isochrons tend to cover a very small range in ^{143}Nd/^{144}Nd ratio. The Sm–Nd technique is particularly well suited to dating the origin of very old rocks and sediments and meteorites.

2.1.3 Lu–Hf Lutetium exists in nature as two isotopes, the stable ^{175}Lu (natural abundance 97.41%) and ^{176}Lu (natural abundance 2.59%), which has a half-life of 3.78×10^{10} y and decays by β^- emission to ^{176}Hf (natural abundance 5.206%). Standard isochron analysis can be performed using 18.606% abundant ^{177}Hf as the normalizing isotope.

One of the principal applications is in dating zircon. Since hafnium is a chemical analogue of zirconium, it is enriched in zircons and the chemical resistance of zircon to weathering makes it a good candidate for dating ancient sediments. However, ^{176}Lu/^{177}Hf is strongly fractionated in sedimentary systems. This may actually prove useful in differentiating between sediment reservoirs and recycling of basalts to the mantle.

2.1.4 Re–Os Rhenium-187 has a natural abundance of 62.60%. It decays with a half-life of 4.35×10^{11} y via β^- emission to stable ^{187}Os (natural abundance 1.6%). The two elements are strongly fractionated between mantle and crustal material, the Re being partitioned into magmatic liquids and the Os staying with the mantle material. This has the potential for studying such processes as ore formation (Freydier *et al.*, 1997), magma genesis, and mantle evolution (Esperanca *et al.*, 1997). The Re–Os system is also well suited to the study of iron meteorites, which contain relatively high concentrations of Re and Os and have large variations in the Re/Os ratios, making it easy to construct isochrons with high precision (Smoliar *et al.*, 1996). The normalizing isotope of choice is ^{188}Os, with a natural abundance of 13.3%.

The potential of Re–Os dating has only begun to be realized through advances in ionization techniques in mass spectrometry, most notably negative thermal ionization mass spectrometry (NTIMS) (Creaser *et al.*, 1991). Rhenium–osmium isochrons in iron meteorites have also been measured with AMS by Ding *et al.* (1997). The two methods compare favorably, with AMS having the advantage of being able to make measurements on single grains with no chemical preparation of the sample.

2.1.5 K–Ar, Ar–Ar The element potassium has three naturally occurring isotopes: ^{39}K, ^{40}K, and ^{41}K. The radioactive ^{40}K has a natural abundance of only 0.0117% and a half-life of 1.227×10^9 y. However, since potassium is prevalent in minerals, there is sufficient ^{40}K in the Earth's crust to be a useful dating tool. The decay of ^{40}K has two branches; β^- decay produces ^{40}Ca in 89.33% of the decays and electron capture (EC) leads to ^{40}Ar in the other 10.67%. The product ^{40}Ca is of limited use because it comprises 96.94% of natural calcium, which is very abundant in most rocks. This makes identifying the small radiogenic component essentially impossible. Argon, on the other hand, is a noble gas and rarely found naturally in rock formations. The main contaminant is atmospheric Ar, which is 99.60% radiogenic ^{40}Ar. This is overcome by using the 0.337% abundant, non-radiogenic ^{36}Ar as an indicator of atmospheric contamination, since purely radiogenic ^{40}Ar in rocks should be free of ^{36}Ar. Isochrons can also be constructed for the K–Ar system, in this case adding a separate term for the atmospheric contribution and accounting for the branching ratio for the EC decay:

$$\left(\frac{^{40}\mathrm{Ar}}{^{36}\mathrm{Ar}}\right)_{\mathrm{sample}} = \left(\frac{^{40}\mathrm{Ar}}{^{36}\mathrm{Ar}}\right)_{\mathrm{atm}} + \left(\frac{^{40}\mathrm{Ar}}{^{36}\mathrm{Ar}}\right)_{\mathrm{initial}} + \frac{^{40}\mathrm{K}}{^{36}\mathrm{Ar}}\frac{\lambda_{\mathrm{EC}}}{\lambda_{\mathrm{tot}}}(e^{\lambda_{\mathrm{tot}}t} - 1). \qquad (8)$$

The analytical techniques commonly used to obtain the individual concentrations of Ar and K are discussed in Dickin (1995). The potassium content of the sample can be found by a number of means, including x-ray fluorescence, atomic absorption, and isotope dilution mass spectrometry. The argon is determined by isotope dilution mass spectrometry using $^{38}\mathrm{Ar}$ as the known spike.

A variation on the K–Ar dating technique is Ar–Ar dating (sometimes referred to as $^{40}\mathrm{Ar}/^{39}\mathrm{Ar}$ dating). In this application the sample is placed in a fast-neutron flux and $^{39}\mathrm{K}$ (natural abundance 93.258%) is converted into $^{39}\mathrm{Ar}$ via the (n, p) reaction. The amount of potassium in a sample can then be determined by mass spectrometry of $^{39}\mathrm{Ar}$, at the same time as the other Ar determinations are being made. Since the number of $^{39}\mathrm{Ar}$ atoms formed is a function of the reactor flux Φ and the energy distribution of the neutrons, standards of known age are used as flux monitors and irradiated along with the sample. One problem with this technique is the possibility of other nuclear reactions occurring during irradiation that also produce argon isotopes, the principal ones being $^{40}\mathrm{K(n, p)}^{40}\mathrm{Ar}$, $^{40}\mathrm{Ca(n, n\alpha)}^{36}\mathrm{Ar}$ and $^{40}\mathrm{Ca(n, \alpha)}^{39}\mathrm{Ar}$. Loss of $^{39}\mathrm{Ar}$ from mineral grains due to recoil during the irradiation is also a concern.

The assumption of a closed system is often difficult to fulfill when the species is a trapped noble gas. The Ar–Ar method allows the sample to be heated incrementally and the ages determined from the argon released in each step. This is an important test of alteration in the ages due to loss of argon, with the first ages often younger than those obtained later in the heating process (Ortega-Rivera *et al.*, 1997).

Another advance in Ar–Ar dating is the application of laser microprobe ablation. This allows single grains to be analyzed and selection of specific inclusions can then be made. A portion of the September 1997 issue of *Geochimica et Cosmochimica Acta* (Vol. 61, No. 18) is devoted to papers on Innovative Applications of $^{40}\mathrm{Ar}/^{39}\mathrm{Ar}$ Microanalytical Research.

2.1.6 U–Pb, Th–Pb, Pb–Pb Lead has four naturally occurring isotopes: $^{204}\mathrm{Pb}$, $^{206}\mathrm{Pb}$, $^{207}\mathrm{Pb}$, and $^{208}\mathrm{Pb}$. Only the $^{204}\mathrm{Pb}$ is non-radiogenic. The other three are the final stable decay products of $^{238}\mathrm{U}$, $^{235}\mathrm{U}$, and $^{232}\mathrm{Th}$, respectively (see Fig. 1 and Table 1). If the time scale is long enough, each parent–daughter system satisfies Eq. (3), with $^{204}\mathrm{Pb}$ as the normalizing isotope. For example, the equation for the $^{238}\mathrm{U}$–$^{206}\mathrm{Pb}$ system is

$$\left(\frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_{\mathrm{sample}} = \left(\frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_{\mathrm{initial}} + \frac{^{238}\mathrm{U}}{^{204}\mathrm{Pb}}(e^{\lambda t} - 1), \qquad (9)$$

with analogous equations for $^{235}\mathrm{U}$–$^{207}\mathrm{Pb}$ and $^{232}\mathrm{Th}$–$^{208}\mathrm{Pb}$. In principle, isochrons can be constructed for each of these three systems, but in practice the requirement of a closed system is usually difficult to meet on account of the mobility of Pb, Th, and U. Nevertheless, direct isochron dating has been successful in certain applications, such as marine sediments (Smith and Farquhar, 1989) and ore-stage calcite (Brannon *et al.*, 1996).

Lead usually exists as Pb^{2+} with an ionic radius of 0.12 nm, which makes it the wrong charge and too large to fit into the lattice of zirconium, uranium, or thorium minerals. This means that the only lead present in these minerals can usually be considered radiogenic. [Corrections for any initial lead are made by determining the $^{204}\mathrm{Pb}$ content (natural abundance 1.4%).] For these cases the $^{235}\mathrm{U}$–$^{207}\mathrm{Pb}$ and $^{238}\mathrm{U}$–$^{206}\mathrm{Pb}$ decay equations then simplify to

$$^{207}\mathrm{Pb}_{\mathrm{radiogenic}} = {}^{235}\mathrm{U}(e^{\lambda_{235}t} - 1) \qquad (10)$$

and

$$^{206}\mathrm{Pb}_{\mathrm{radiogenic}} = {}^{238}\mathrm{U}(e^{\lambda_{238}t} - 1) \qquad (11)$$

The applicability of U–Pb dating is greatly

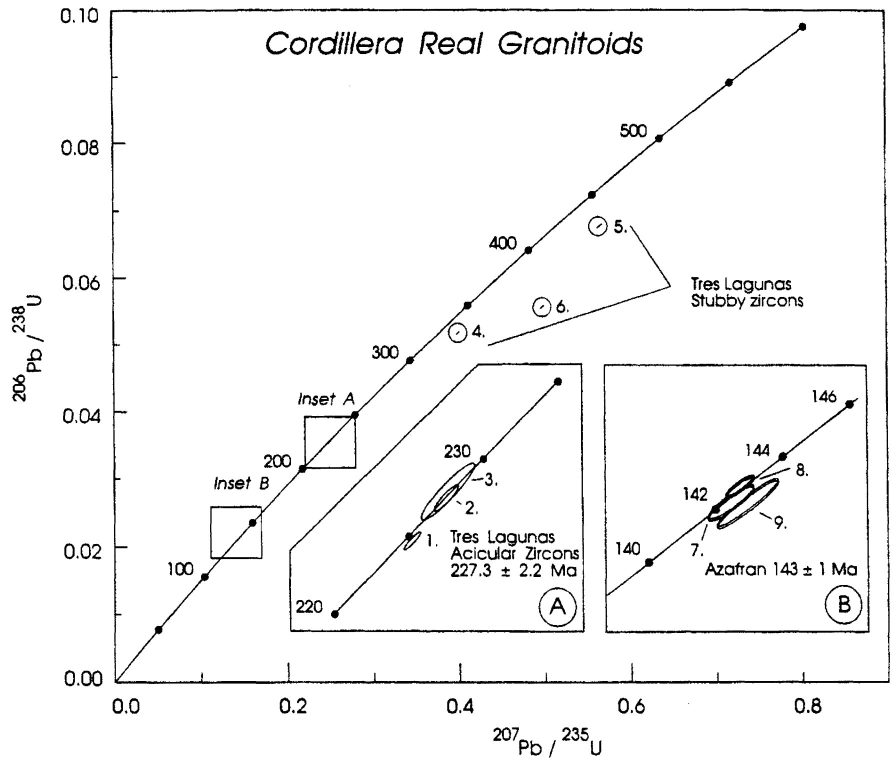

FIG. 6. Concordia plot of $^{206}\text{Pb}/^{238}\text{U}$ as a function of $^{207}\text{Pb}/^{235}\text{U}$ (from Noble *et al.*, 1997, reproduced by permission of Geological Society of America). The ages along the curve are indicated by the numbers, in units of Ma (My). Samples labeled 4, 5 and 6 are clearly discordant.

expanded by creating what are termed concordia plots. Since the decay constants are known, plots of $^{206}\text{Pb}/^{238}\text{U}$ as a function of $^{207}\text{Pb}/^{235}\text{U}$ for varying times should give a unique curve called a concordia. If there has been loss of radiogenic lead, the data will be discordant and will fall off the curve. Since the mechanisms for lead loss should be the same for the two uranium isotopes, the discordant points fall on a straight line below the concordia. Extrapolation of this discordant line to the concordia curve gives the "true" age of the sample. Figure 6 is an example of a concordia for zircons taken from Noble *et al.* (1997).

Alternatively, the ratio of Eqs. (10) and (11) gives

$$\left(\frac{^{207}\text{Pb}}{^{206}\text{Pb}}\right)_{\text{radiogenic}} = \frac{^{235}\text{U}(e^{\lambda_{235}t}-1)}{^{238}\text{U}(e^{\lambda_{238}t}-1)}, \qquad (12)$$

with the $^{235}\text{U}/^{238}\text{U}$ ratio well known to be 1/137.88. Application of Eq. (12) is called Pb–Pb dating, since it only requires knowledge of the ratio of two isotopes of lead. The Pb–Pb ratios are a reflection of the Pb/U ratios; however, they have the advantage of being quite insensitive to recent conditions that might produce uranium loss, since it is the radiogenic daughter lead isotope which is measured. Taking into account any initial ^{206}Pb and ^{207}Pb and using the isochron technique of normalizing to the nonradiogenic ^{204}Pb gives

$$\frac{\left(\frac{^{207}\text{Pb}}{^{204}\text{Pb}}\right)_{\text{sample}} - \left(\frac{^{207}\text{Pb}}{^{204}\text{Pb}}\right)_{\text{initial}}}{\left(\frac{^{206}\text{Pb}}{^{204}\text{Pb}}\right)_{\text{sample}} - \left(\frac{^{206}\text{Pb}}{^{204}\text{Pb}}\right)_{\text{initial}}} = \frac{1}{137.88}\frac{(e^{\lambda_{235}t}-1)}{(e^{\lambda_{238}t}-1)}. \qquad (13)$$

While this cannot be solved in closed form for the time t, computer integration yields isochrons as a function of time, which can be matched to the data. This technique is used when recent loss of uranium is suspected.

The Pb–Pb technique was first applied to meteorites. The Diablo Canyon meteorite produced an age of 4.55×10^9 y. Subsequent studies using Pb–Pb dating have shown that some deep ocean sediments lie on the Diablo Canyon isochron. This isochron is considered characteristic of the age of the Earth and now called the geochron. Numerous other special applications of U–Pb dating have been used to study the evolution of the Earth. For further variations and applications the reader is referred to Titayeva (1994) and Dickin (1995).

2.2 Nonequilibrium Techniques and the U–Th Decay Chains

The methods discussed in Sec. 2.1 all utilize simple parent–daughter decay, where the daughter is stable. However, inspection of Fig. 1 for the decay of ^{238}U shows that the final stable product ^{206}Pb is the result of a complex sequence of decays. Similar decay schemes connect ^{235}U to ^{207}Pb and ^{232}Th to ^{208}Pb. If conditions are right to remove and isolate one of the intervening long-lived decay products, this will then be the parent for its own decay chain. If an unstable daughter nuclide is monitored, the daughter formation and decay follows Eq. (5). Deviations from secular equilibrium also provide opportunities to gain information by monitoring the intermediate radionuclides. There are numerous possibilities, a few of which will be described here. The reader is referred to Faure (1986) and Dickin (1995) for more extensive treatments.

2.2.1 ^{238}U–^{234}U–^{230}Th System: Dating of Corals and Ocean Sediments One example of a non-equilibrium situation is the ^{238}U–^{234}U–^{230}Th parent–daughter system. In a marine environment, uranium exists as soluble carbonates which are readily incorporated into corals and shells along with calcium carbonate. In contrast, thorium is insoluble and is absorbed onto detrital particles, precipitates, and is preferentially incorporated into sediments. Therefore, the ^{232}Th/^{238}U ratio in seawater is very low (on the order of 10^{-5}) and the ^{230}Th/^{238}U ratio in seawater is about 10^{-5} times what would be obtained from ^{238}U equilibrium decay (Edwards *et al.*, 1986). Growing coral incorporates ^{234,238}U along with the carbonate, but will have very little of the daughter ^{230}Th initially. Monitoring the subsequent decay of ^{234}U and in-growth of ^{230}Th provides an important means of dating coral formations as old as 500 ky, a timescale inaccessible using radiocarbon dating.

In the ^{238}U–^{234}U–^{230}Th system, with the assumption of no initial ^{230}Th, Eq. (5) leads to

$$1 - \left[\frac{^{230}\mathrm{Th}}{^{234}\mathrm{U}}\right]_{\mathrm{act}} = e^{-\lambda_{230}t} - \left[\frac{\delta^{234}\mathrm{U}}{1000}\right] \times \left(\frac{\lambda_{230}}{\lambda_{230} - \lambda_{234}}\right)[1 - e^{(\lambda_{234}-\lambda_{230})t}], \qquad (14)$$

where the subscript "act" denotes activity ($A = \lambda N$). The δ^{234}U is a measure of the deviation of the ^{234}U/^{238}U activity ratio from unity (secular equilibrium) and is defined as

$$\delta^{234}\mathrm{U} = \left[\frac{\left(\frac{^{234}\mathrm{U}}{^{238}\mathrm{U}}\right)}{\left(\frac{^{234}\mathrm{U}}{^{238}\mathrm{U}}\right)_{\mathrm{equil}}} - 1\right] \times 1000, \qquad (15)$$

and

$$\delta^{234}\mathrm{U}(t) = \delta^{234}\mathrm{U}(0)e^{-\lambda_{234}t}. \qquad (16)$$

If the ^{234}U and ^{238}U were in secular equilibrium is seawater, the δ^{234}U would be zero and the second term on the right of Eq. (14) would vanish. However, this is not the case, and in practice isochrons are constructed of the activity ratios of ^{234}U/^{238}U as a function of ^{230}Th/^{234}U.

The dating of corals has been important in studies of sea-level stance and climatic fluctuation of the Pleistocene era. The use of TIMS has made coral dating possible with samples as small as 200 mg and analysis of growth rings. This now allows extension of the ^{14}C calibration curve to ages older than accessible by dendochronology (Bard *et al.*, 1990). Applications can also be made to shells, but in this case the possibility of secondary impurities is dealt with by successive stripping and measuring the individual fractions.

A second simplifying case occurs for ocean sediments and ferromanganese nodules (thorium coprecipitates with the Fe and Mn hydroxides), where there is essentially no parent ^{234}U to feed the ^{230}Th; this is called unsupported ^{230}Th. For example, the decay of the ^{230}Th can be followed as a measure of sedimentation rate R, with $^{230}\mathrm{Th} = {}^{230}\mathrm{Th}_{\mathrm{initial}}e^{-\lambda_{230}t}$. A graph of log(^{230}Th activity) versus depth should give a straight line with a slope $-\lambda_{230}/R$. This is usually true for shallow depths, but the effect of U-supported ^{230}Th usually shows up as depth increases. This can be corrected for by determining the ^{234}U content (which is a measure of the ^{230}Th activity in secular equilibrium with any uranium) and subtracting this from the total ^{230}Th activity.

2.2.2 ^{235}U–^{231}Pa

The radionuclide ^{231}Pa is the first long-lived nuclide in the ^{235}U decay chain

$$\begin{aligned}&^{235}\mathrm{U}(t_{1/2} = 7.04 \times 10^6\,\mathrm{y})\\ &\quad\rightarrow {}^{231}\mathrm{Th}(t_{1/2} = 25.5\,\mathrm{h})\\ &\quad\rightarrow {}^{231}\mathrm{Pa}(t_{1/2} = 3.28 \times 10^4\,\mathrm{y}) \rightarrow \ldots .\end{aligned}$$

Hence, in a closed system ^{231}Pa can be considered in equilibrium with its parent ^{235}U. However, in seawater ^{231}Pa is similar to ^{230}Th and is absorbed onto particulate matter and rapidly incorporated into sediments. This leaves the seawater rich in ^{235}U, but almost devoid of the daughter ^{231}Pa, allowing dating of corals by following the ingrowth and decay of ^{231}Pa (Edwards *et al.*, 1997). Under conditions where the system had no initial ^{231}Pa and was closed with respect to ^{231}Pa and ^{235}U, the time elapsed since closure is given by

$$t = \frac{-\ln\left(1 - \left[\frac{^{231}\mathrm{Pa}}{^{235}\mathrm{U}}\right]_{\mathrm{act}}\right)}{\lambda_{231}}, \tag{17}$$

where "act" denotes the ratio of the activities of ^{231}Pa and ^{235}U. It has been proposed that since both ^{231}Pa and ^{235}U are daughters of uranium isotopes, concordia could be constructed and used in a manner similar to the U–Pb dating of minerals. (See the discussion in Sec. 2.1.6.)

2.2.3 ^{210}Pb

Lead-210 is a short-lived ($t_{1/2} = 22.3$ y) intermediate in the decay chain of ^{238}U (see Fig. 1). It is produced in the atmosphere by the decay of ^{222}Rn, which is a noble gas and escapes from the earth at a rate of about 42 atoms/min cm^2 (Faure, 1986). With a residence time in the atmosphere of only 10 d, the ^{210}Pb is quickly removed by precipitation and becomes an excellent dating tool for snow pack and glaciers. In water columns ^{210}Pb has a residence time of 1–2 y before it becomes attached to sediment particles. Since the lead is now removed from its radioactive precursors, it decays away unsupported with a half-life of 22.3 y, making it useful for dating sediments over a period of several hundred years. Under the assumption of a constant deposition rate, the age of a sample can be found by solving Eq. (2) as applied to ^{210}Pb in units of atoms or activity,

$$^{210}\mathrm{Pb} = {}^{210}\mathrm{Pb}_{\mathrm{initial}}e^{-\lambda_{210}t}, \tag{18}$$

where $^{210}\mathrm{Pb}_{\mathrm{initial}}$ is the value at the surface. In a plot of log (^{210}Pb) versus depth, the slope gives the sedimentation or accumulation rate. The major limitation is the contribution from ^{210}Pb generated from ^{226}Ra in the sediments. This contribution is termed supported, since it should be in equilibrium with the long-lived ^{226}Ra and any intermediate decay products. The supported ^{210}Pb can be corrected for by the measurement of ^{226}Ra (or the short-lived intervening daughters).

While the ^{210}Pb itself emits a low-energy β particle that is difficult to detect, it has traditionally been quantitated by measuring the β decay of its daughter ^{210}Bi or the α decay of the ^{210}Po daughter. However, advances in γ-ray spectroscopy have made the direct measurement of ^{210}Pb possible through detection of its 46.5-keV γ ray (Böllhofer *et al.*, 1994). This has the advantage of also allowing detection of any ^{226}Ra ($E_\gamma = 186.1$ keV) or its daughters ^{214}Pb ($E_\gamma = 352.0$ keV) and ^{214}Bi ($E_\gamma = 609.3$ keV), important in correcting for any supported ^{210}Pb. Lead-210 dating is now

Table 3. Extinct radionuclides (adapted from Dickin, 1995).

Parent–daughter	Parent half-life[a]	Daughter natural abundance (%)[a]
^{244}Pu–fission products including ^{132}Xe, ^{134}Xe, ^{136}Xe	8.08×10^{7} y	26.9, 10.4, 8.9
^{129}I–^{129}Xe	1.57×10^{7} y	26.4
^{107}Pd–^{107}Ag	6.5×10^{6} y	51.839
^{53}Mn–^{53}Cr	3.74×10^{6} y	9.501
^{60}Fe–(^{60}Co)–^{60}Ni	1.5×10^{6} y (1.9251×10^{3} d)	26.223
^{26}Al–^{26}Mg	7.4×10^{5} y	11.01
^{41}Ca–^{41}K	1.03×10^{5} y	6.7302

[a] From Nuclear Wallet Cards, July 1995, National Nuclear Data Center, Brookhaven National Laboratory, P.O. Box 5000, Upton, NY 11973–5000, U.S.A.

an established method for dating marine and lake sediments. It has recently become important in supplying dates that can be correlated with anthropogenic activities, such as lead and other heavy-metal deposition and radioactive fallout.

2.3 Extinct Radionuclides

The nucleosynthetic processes that produced the naturally occurring nuclides found on Earth today also produced numerous unstable nuclides that have long since decayed away to their stable daughter products. In fact, many of our stable nuclides are the ultimate end points of the decay of short-lived radioactive species. Only the longest-lived ones are still around for us to study today. However, the evidence for now extinct radionuclides can be found in meteorites, which preserve remnants of the original solar nebula from which our solar system condensed. This evidence is in the form of the anomalous isotopic abundances of daughter nuclides. If the radiogenic component can be isolated from that which was directly produced, limits can be set on the time between the nucleosynthesis event and the formation of the meteoritic material in which the isotope was found. However, this requires a number of assumptions on the initial nucleosynthesis-produced isotopic ratios, the time over which the production occurred, and the possible mixture of sources.

Table 3 (adapted from Dickin, 1995) lists a number of extinct radionuclide systems that have been studied in meteorites. They cover a range of half-lives from 10^{5} to 10^{7} y. Combining time estimates for a number of systems provides constraints on production scenarios that cannot be obtained with a single system alone. For example, times derived from measuring ^{129}Xe and xenon isotopes in ^{244}Pu fission fragments are consistent with a delay of on the order of 100 My between nucleosynthesis and closure of the system to xenon loss. However, studies of the shorter lived radionuclides in Table 3 also provide evidence for a late injection of material into the solar nebula between 2 and 20 My before condensation.

2.4 Fission-track Methods

The kinetic energy released when a heavy nucleus such as uranium undergoes fission is on the order of 200 MeV. This large amount of energy is divided between the two fission products (and any neutrons) according to the law of conservation of momentum. If the fission were symmetric (giving two equal-size fragments), then each would carry away about 100 MeV of kinetic energy. These heavy, high-Z, energetic fragments are highly ionizing and produce damage tracks in their stopping material.

The only naturally occurring radionuclide that has a significant spontaneous fission branch is ^{238}U. Even this is very small, accounting for only 0.0001% of all ^{238}U decays. However, given the moderately large concentrations of uranium in many minerals, the effects of the stopping of fission products in the material are clearly observable. If the damage

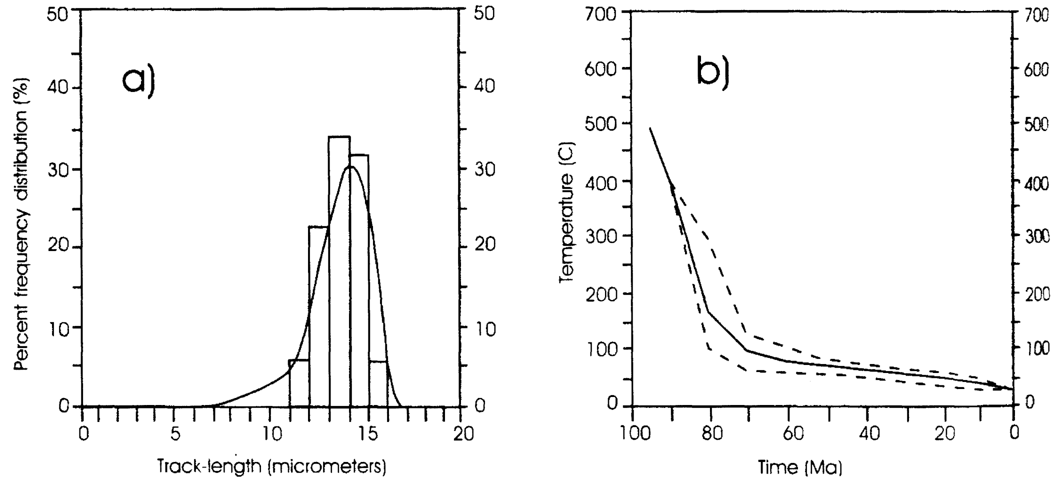

FIG. 7. Results of modeling fission track lengths (from Ortega-Rivera *et al.*, 1997, reproduced by permission of Geological Society of America). (a) Comparison of measured track-length distributions (histogram) and model length distributions (solid curve). (b) Time–temperature ranges for 250 acceptable solutions generated by the model (dashed curve.) The solid line represents the exponential-mean solution.

tracks are enhanced by etching with acid or alkaline solution, they can be counted and used as a measure of the integrated dose received by the material. The dose can be calibrated by exposing like materials to known fluxes of induced fission fragments and comparing the number of etched tracks. A thorough discussion of the analytical techniques can be found in Dickin (1995) and Ravenhurst *et al.* (1994).

How long fission tracks stay preserved depends on the material type and the thermal history. In general, the annealing process can be described by a simple Boltzmann distribution, where the annealing time can be written as a function of the material specific activation energy U and the absolute temperature T:

$$t = Ae^{U/kT}, \tag{19}$$

with A a calibration constant and k the Boltzmann constant.

Considerable work has been done on the effects of temperature on track annealing in apatites (Laslett *et al.*, 1987). As tracks anneal they decrease in length, producing a length distribution that can be modeled in terms of a time and temperature profile. The partial annealing zone (PAZ) is sensitive to temperatures as low as 60–110 °C. Figure 7(a), from Ortega-Rivera *et al.* (1997), shows a track length profile from an apatite from the Sierra San Pedro Martir pluton in Baja California. A model fit to the track distribution gives the time–temperature relationship constraints shown in Fig. 7(b). The fission-track technique is well suited to studies of cooling, exhumation, and denudation histories of tectonic formations and the effects of erosion on these formations (Johnson, 1997).

2.5 (U–Th)/He Thermochronometry

The α decays in the uranium and thorium parent–daughter systems produce radiogenic helium, which can be retained in rock masses. Recent work in apatite is providing a new means of tracking thermal histories of low-temperature processes by following the diffusive loss of He. The closure temperature in apatite is around 75 °C, with partial retention occurring between 40 and 75 °C (Wolf *et al.*, 1997). This is analogous to the partial annealing zone in fission-track dating (see Sec. 2.4) and can provide a means of extending studies on apatites to lower temperatures than accessible via fission-track analysis. The analytical details of the technique can be found in Wolf *et al.* (1996).

2.6 Thermoluminescence (TL) Dating Applications

Dating artifacts via TL has been an important part of art history and archaeology. The technique was first pioneered for dating pottery and other baked clays, being applied to both the clays and glazes. In this usage the TL light typically comes from the quartz and feldspars in the clay, with the clock reset to zero upon firing of the pot. Both bulk pottery fragments and mineral grains can be dated. In alkali feldspars with a high potassium content, the paleodose is mainly internal, coming from the decay of ^{40}K. This has the advantage of being less sensitive to the moisture content of the surroundings than the external dose.

The widespread availability of fired clay at most archaeological sites makes TL a particularly useful dating tool. While TL dating is not as accurate as radiocarbon dating for samples with ages between 1000 and 10 000 y, it is often the method of choice for ages older than 10 000 y, where the radiocarbon time scale is poorly calibrated, or for samples younger than 1000 y, where radiocarbon can give ambiguous ages (see Sec. 3.1). The ability of TL to extend to ages older than those accessible by radiocarbon is particularly important for paleontological artifacts such as burnt flint and cooking stones.

The TL dating limit depends on the lifetimes of the trapping sites and the saturation light sum. Quartz has proven to be an excellent paleodosimeter for Quarternary deposits, since it does not reach its saturation light sum until a total dose of 10^4 Gy, which means that it can be sensitive to times as long as 10^6 y, depending on the dose rate. This is in contrast to radiocarbon's practical limit of around 50 000 y. Geological applications of TL include lunar material and meteorites, volcanic eruptions, sediments, and stalagmitic calcite. Specific applications and techniques can be found in Aitken (1985) and Aitken (1990).

3. TECHNIQUES BASED ON COSMOGENIC NUCLIDES

Cosmogenic nuclides provide access to time scales much shorter than the primordial radionuclides. Since they are constantly being produced, radionuclides with half-lives as short as 53 d (as in the case of ^{7}Be) can prove useful. Table 4 lists the main cosmogenic nuclides used as dating tools today, along with their production modes, half-lives, and relevant applications. Until the advent of accelerator mass spectrometry (AMS), relatively large sample sizes were necessary to obtain nuclide concentrations, using low-level decay counting. However, not all of the nuclides in Table 4 are currently measurable by AMS at levels low enough to be useful, and decay counting is still an important analytical technique. Today cosmogenic nuclides are finding uses in a broad range of research areas, from archaeology and paleontology, to paleoclimate and ocean circulation studies, to hydrology and geomorphology, and even cosmology. The following subsections highlight a few of these, with a separate section devoted to ^{14}C.

3.1 Carbon-14

The most widely used cosmogenic nuclide is probably ^{14}C, commonly termed radiocarbon. Carbon-14 is produced in the atmosphere primarily through the reaction ^{14}N(n,p). It is rapidly incorporated into atmospheric CO_2 and becomes thoroughly mixed with the global carbon pool, showing little variation with latitude. When a plant or animal dies it is removed from active exchange with the carbon pool, and its radiocarbon content will steadily decrease as the ^{14}C decays. Carbon-14 has been a mainstay of archaeology since Libby first proposed using it as a dating tool in the 1950s (Libby, 1955), and many laboratories still specialize in radiocarbon dating through decay counting. The archaeological applications of ^{14}C are discussed in PHYSICS IN ARCHAEOLOGY.

One of the biggest challenges to obtaining accurate radiocarbon dates is the calibration of the time dependence of the atmospheric equilibrium ^{14}C/C ratio. This varies with the cosmogenic production rate through the effect of solar cycles and the Earth's magnetic field on cosmic ray flux. An additional effect on atmospheric ^{14}C has been the increasing use of fossil fuels. This has been diluting the natural radiocarbon reservoir with nonradioactive

Table 4. Cosmogenic nuclides.

Nuclide	Production modes	Half-life[a]	Applications
^{3}H	^{14}N(n, ^{3}H)^{12}C[b,c], ^{6}Li(n, α)^{3}H[d]	12.33 y	Post-bomb groundwater dating
^{3}He	^{14}N, ^{16}O(spallation)[c,e] ^{3}H(β^-decay)[b,e]	Stable	Exposure dating
^{7}Be	^{14}N(n, 3p5n)[c], ^{14}N(p, 4p4n)[c] ^{16}O(p, 5p5n)[c]	53.29 d	Soils and sediments
^{10}Be	^{14}N(p, 4pn)[c] ^{16}O(spallation)[c,e]	1.51×10^{6} y	Soils and sediments, erosion rates and exposure dating, glacial ice
^{14}C	^{14}N(n, p)[b,c] ^{16}O(p, ^{3}H)[c,e] ^{17}O(n, α)[d]	5730 y	Erosion rates and exposure dating, ocean studies, groundwater studies, soil and sediments, glacial ice, paleodating, archaeological artifacts
^{21}Ne	^{24}Mg(n, α)[e], ^{23}Na(n, p2n)[e]	Stable	Exposure dating
^{26}Al	^{40}Ar(spallation)[c] ^{26}Mg(p, n)[e] ^{28}Si(p, 2pn)[e]	7.4×10^{5} y	Soils and sediments, erosion rates and exposure dating, meteorites
^{32}Si	^{40}Ar(spallation)[c]	140 y[f]	Sedimentation rates, glacial ice
^{36}Cl	^{40}Ar(p, 2p3n)[c] ^{35}Cl(n, γ)[b,d,e], ^{40}Ar(n, 4np)[b] ^{39}K, ^{40}Ca(spallation)[e] ^{39}K(n, α)[d], ^{39}K(μ, p2n)[e] ^{40}Ca(μ, α)[d]	3.01×10^{5} y	Erosion rates and exposure dating, groundwater studies, glacial ice, paleodating of organic fossil material
^{39}Ar	^{40}Ar(n, 2n)[c] ^{39}K(n, p)[d,e]	269 y	Erosion rates and exposure dating, groundwater studies
^{53}Mn	^{53}Fe(p, 2p)[e], ^{53}Fe(p, α)[e]	3.74×10^{6} y	Meteorites
^{59}Ni	^{59}Co(p, n)[e], ^{60}Ni(p, pn)[e]	7.6×10^{4} y	Meteorites
^{81}Kr	^{82}Kr(n, 2n)[c], ^{80}Kr(n, γ)[c,d] ^{238}U(spontaneous fission)[d]	2.29×10^{5} y	Groundwater dating
^{85}Kr	^{238}U(spontaneous fission)[d] ^{235}U(n, fission)[d], ^{84}Kr(n, γ)[c], ^{86}Kr(n, 2n)[c]	3934.4 d	Groundwater dating
^{129}I	^{129}Xe(n, p)[c] ^{131}Xe(spallation)[c] ^{235}U(n, fission)[b,e] ^{238}U(spontaneous fission)[c]	1.57×10^{7} y	Oil formation and salt brine dating, groundwater studies, ocean circulation

[a] From Nuclear Wallet Cards, July 1995, National Nuclear Data Center, Brookhaven National Laboratory, P.O. Box 5000, Upton, NY 11973-5000, U.S.A.

[b] Anthropogenic.

[c] Atmospheric cosmogenic.

[d] Subsurface.

[e] *In situ*.

[f] Silicon-32 half-life from Morgenstern *et al.*, 1996.

carbon (termed dead carbon), causing a steady decrease in the equilibrium atmospheric ^{14}C/C ratio, called the Suess effect after the individual who first observed it. Even more recently, nuclear bomb tests have elevated the atmospheric ^{14}C/C ratio.

The unit usually used to express ^{14}C in a sample is Δ^{14}C(‰), which is defined using a

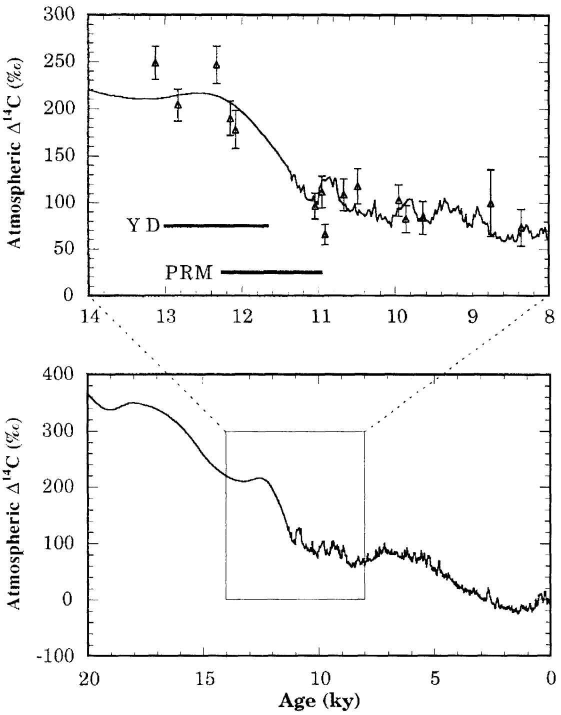

FIG. 8. Atmospheric radiocarbon expressed as Δ^{14}C [actually plotted as the negative of Eq. (20)] as a function of age, derived from bidecadal tree-ring data and U/Th coral data (obtained from CALIB program as described in Stuiver and Reimer, 1993). The triangles are the coral data from Edwards *et al.* (1993), plotted against ^{230}Th date.

prebomb-era oxalic acid carbon standard as

$$\Delta^{14}\mathrm{C}‰ = \frac{\left(\frac{^{14}\mathrm{C}}{^{12}\mathrm{C}}\right)_{\mathrm{sample}} - 0.95\left(\frac{^{14}\mathrm{C}}{^{12}\mathrm{C}}\right)_{\mathrm{standard}}}{0.95\left(\frac{^{14}\mathrm{C}}{^{12}\mathrm{C}}\right)_{\mathrm{standard}}} \times 1000. \tag{20}$$

The 0.95 is a correction factor to modern ratios. A Δ^{14}C value of zero is defined at 1950 as modern carbon. The more negative the value, the older the radiocarbon age. Positive values reflect input from bomb-produced carbon and are termed post-modern. [Note that in some cases the negative of Eq. (20) is defined as Δ^{14}C, as in Fig. 8.]

The field of dendochronology (tree-ring dating) has been a key to understanding the variation in Δ^{14}C and its correlation with absolute dates, but is limited to around 10 000 years. *Calibration* dates have recently been extended by the study of corals (which grow in layers like trees), correlating radiocarbon dates with those obtained from independent U–Th dating (Bard *et al.*, 1990; Edward *et al.*, 1993). (See Sec. 2.2.1.) Figure 8 shows Δ^{14}C as a function of age as derived from bidecadal tree-ring data and U–Th coral dates (Stuiver and Reimer, 1993). It is clear that there are regions of the curve where a single ^{14}C value does not correspond to a unique age. As the expanded region shows, there can be considerable variation over both long and relatively short times due to fluctuations in the cosmic-ray flux (see Sec 1.2.2). In cases where a series of Δ^{14}C measurements have been made at what are interpreted as regular time intervals, the ambiguity in the age determination can be reduced by matching the trend in the sequence of Δ^{14}C values. This technique is referred to as wiggle matching.

Applications of ^{14}C have expanded tremendously since AMS made measurements of small samples feasible. For example, the coral records in the Huon Peninsula, Papua, New Guinea, in conjunction with the ^{230}Th *dating*, show interesting signatures of changes in the atmospheric–oceanic exchange of ^{14}C during

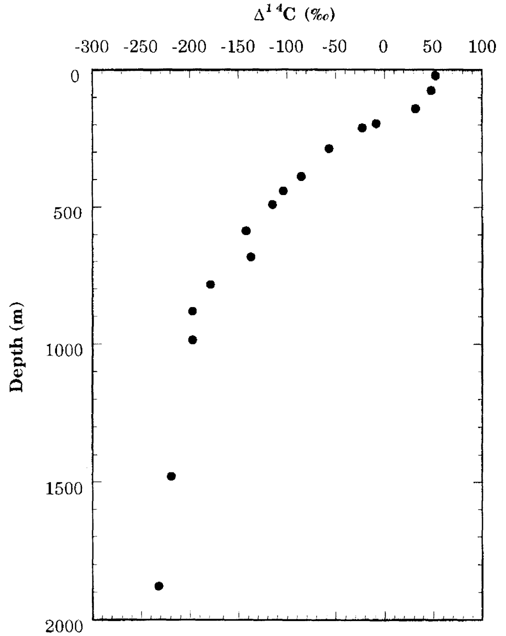

FIG. 9. Depth profile of the Δ^{14}C in ocean water, measured for a site in the Pacific Basin collected as part of the World Ocean Circulation Experiment (WOCE) (data from von Reden *et al.*, 1997).

the Younger Dryas (YD) climate event and the period of reduced melting (PRM) since the last major ice age (Edward *et al.*, 1993). These data are shown as the triangles in the expanded graph in Fig. 8. The times of the YD and PRM are marked by the horizontal bars and correspond to a rapid change in the inferred atmospheric Δ^{14}C.

Studies of radiocarbon in the oceans were carried out long before AMS was available, using large-volume samples. Now small-volume samples are being measured at the National Ocean Sciences AMS facility (NOSAMS) at Woods Hole Oceanographic Institute with a precision of better than 0.4%, rivaling that obtained using decay counting (von Reden *et al.*, 1997). Figure 9 shows a ^{14}C depth profile for a site in the Pacific basin, which is part of the data set collected in the World Ocean Circulation Experiment (WOCE). The surface waters clearly show the influence of bomb carbon. The deep waters have reached an equilibrium value. Similar results have been obtained for the Canadian Basin of the Arctic Ocean (Schlosser *et al.*, 1997). The differences in radiocarbon content of surface and deep waters provide a tool for studying ocean currents and mixing.

Another application of ^{14}C dating is the ages of sediments in lakes and peat bogs. Figure 10 shows the results of dating a sediment core from Germany and its correlation to the pollen analysis showing the history of the prevailing vegetation at the site (Kretschmer *et al.*, 1997). The time scale for ^{14}C also makes it perfect for dating many groundwater aquifer systems. This has been attempted with varying degrees of success using both the dissolved CO_2 (inorganic carbon, DIC) and the dissolved organic carbon (DOC) (Murphy *et al.*, 1989; Drimmie *et al.*, 1991). In the case of DIC, corrections must be made for any dissolution of carbonate minerals. Using DOC to obtain an age requires knowledge of the radiocarbon age of the initial input and the assumption of no addition of dead carbon by subsurface bacterial activity producing organic compounds from kerogen sources. The input function can be either contaminated with anthropogenic ^{14}C or much older than modern by reason of the long residence time of organic matter in soils. Dating of soil organic matter has been shown to be quite complex, and a detailed understanding of the maturation process at a given site is needed before radiocarbon ages can be interpreted (Y. Wang *et al.*, 1996).

3.2 *In Situ* Dating of Geological Formations

The production of radionuclides inside mineral grains from cosmic-ray interactions has been used to estimate erosion rates and exposure ages of surface land forms. This is a multidimensional problem, with the radionuclide concentration changing with the time exposed to cosmic rays and the flux of cosmic rays. The cosmic-ray flux is a function of depth (see Sec. 1.2.2) and can change over time by either surface erosion (which will increase the flux) or burial (which decreases the flux). An additional effect is the change in tilt of the surface, which changes the path length of the cosmic rays and the attenuation. Add to this the fact that the flux of cosmic rays is a function of latitude and can vary as a result of solar and geomagnetic factors, and the complexity of the problem becomes apparent.

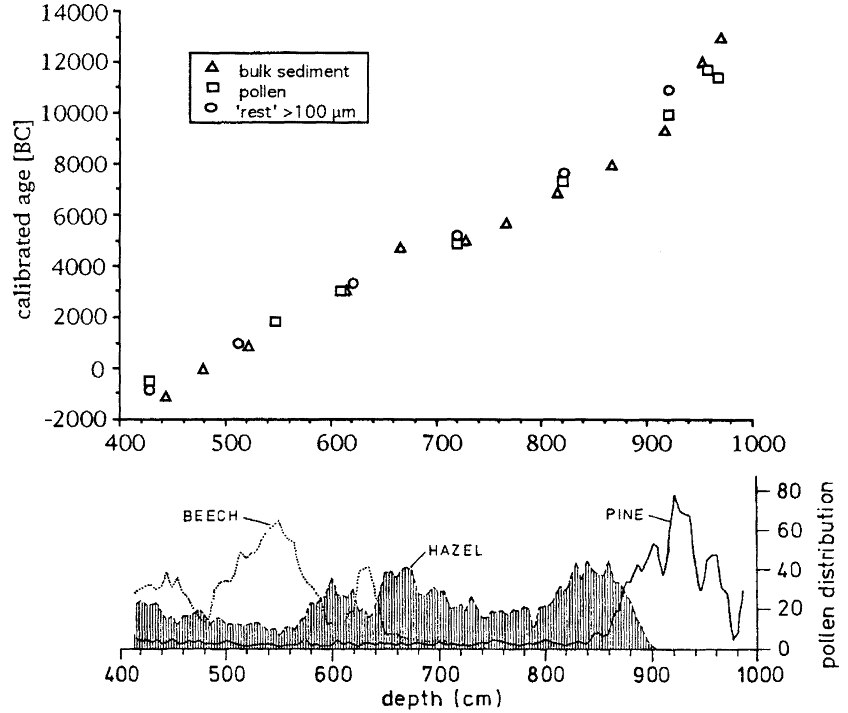

FIG. 10. Calibrated radiocarbon ages of sediment constituents as a function of depth, compared with the vegetation type determined from pollen analysis, for a site in Germany (data from Kretschmer *et al.*, 1997, drawing reproduced by permission of North-Holland Publishing Co.).

In the simplified case of a flat surface exposed to a constant flux, the concentration of the cosmogenic nuclide N_i in atoms g^{-1} is given by

$$N_i = \frac{P_i}{\lambda_i + \varepsilon/k}\left\{1 - \exp\left[-\left(\lambda_i + \frac{\varepsilon}{k}\right)t\right]\right\} + N_0 e^{\lambda t}, \qquad (21)$$

where P_i = production rate, atoms (g-ky)$^{-1}$; λ_i = decay constant, ky^{-1}; ε = actual erosion rate, cm ky^{-1}; k = scale depth, calculated as cosmic-ray attenuation coefficient Λ divided by the density ρ; t = actual time, ky; and N_0 = initial concentration, atoms g^{-1} (notation after Gillespie and Bierman, 1995). In Eq. (21) the attenuation of the cosmic-ray flux is assumed exponential with depth. However, this is not strictly true of the neutron flux generated, which peaks at a depth of about 20 cm before displaying exponential behavior (see Fig. 2).

In the case of negligible erosion, with no initial concentration N_0, Eq. (21) reduces to

$$N_i = (P_i/\lambda_i)(1 - e^{-\lambda_i t}). \qquad (22)$$

In this limiting case, the concentration of a single isotope can give the exposure age. The solutions to Eq. (22) for some commonly used cosmogenic isotopes are displayed graphically in Fig. 11. Alternatively, if an independent estimate can be found for either erosion rate or exposure age, other can be found. In either case, a steady state is eventually reached. This is shown in Fig. 12(a) for the buildup of ^{10}Be in quartz for various erosion rates. When the cosmogenic isotope is stable, a steady state is still reached as a result of the erosion rate, as shown in Fig. 12(b).

If both erosion rate and exposure time need to be determined, this can be achieved by measuring two isotopes in the same sample. Isotopes usually paired are ^{10}Be, ^{26}Al, and ^{36}Cl. Readers are referred to Lal (1991), Cerling and Craig (1994), Gillespie and Bierman

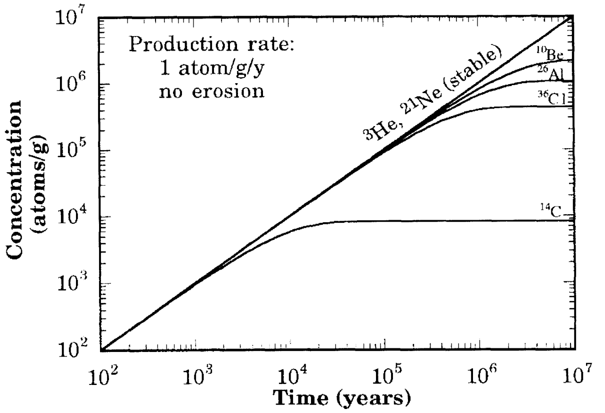

FIG. 11. Buildup of various nuclides with time for the case of no erosion [Eq. (22)].

(1995), and Clark *et al.* (1995) for discussion of the many considerations needed to obtain accurate results from this technique. An interesting application is the study of the Canyon Diablo meteorite. Surface-exposure dating of the surrounding crater using ^{10}Be/^{26}Al (Nishizumi *et al.*, 1991) and ^{36}Cl (Phillips et al., 1991) agreed with the thermoluminescence date of 49 000 ± 3000 y (Sutton, 1985). Measurements of the *in situ* ^{10}Be, ^{26}Al, and ^{36}Cl in the meteorite itself were used to estimate the production rates before the meteorite's fall to Earth. Because of the relatively short half-lives of the radionuclides, it could be assumed that

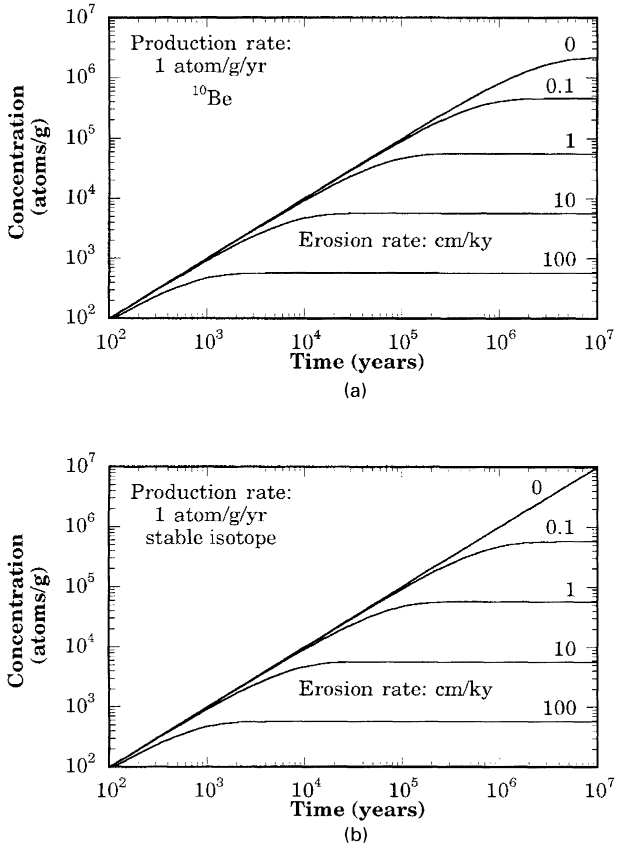

FIG. 12. Build-up of (a) ^{10}Be and of (b) a stable nuclide with time for different erosion rates, calculated with Eq. (21) using $k = 57$ cm.

they had reached equilibrium at the time of the fall, with production equaling decay. After correcting for the time on Earth, the production rate as a function of preatmospheric depth could be found. Using the ratio of the radionuclide activity to a stable noble-gas concentration ($^{10}Be/^{21}Ne$, $^{26}Al/^{21}Ne$, and $^{36}Cl/^{36}Ar$) as an integrator of the exposure period, the cosmic-ray exposure ages of six samples were found to give means of 560 ± 130 My for $^{10}Be/^{21}Ne$, 550 ± 170 My for $^{26}Al/^{21}Ne$, and 470 ± 120 My for $^{36}Cl/^{36}Ar$ (Michlovich *et al.*, 1994). Differences among samples were attributed to variations in production rates due to matrix effects (elemental composition) and irradiation geometries.

3.3 Ice Cores, Ocean and Soil Studies

3.3.1 Atmospheric Beryllium-10 in the Lithosphere and Hydrosphere Beryllium-10 has a very short residence time in the atmosphere, being strongly adsorbed onto aerosol particles and removed by wet and dry deposition. Both processes provide pathways for entrance into the hydrosphere. Profiles of ^{10}Be in ice cores and snow packs have been used as a measure of cosmic-ray flux variations and correlated with other cosmogenic isotopes, such as ^{14}C and ^{36}Cl. Arctic and Antarctic ice cores show clear evidence of the effects of the 11 y sunspot cycle (Beer *et al.*, 1990; Steig *et al.*, 1996). The influence of geomagnetic variations is less evident, with conflicting reports on the importance of latitudinal mixing (Raisbeck *et al.*, 1985; Mazaud *et al.*, 1994; Steig *et al.*, 1996).

Under the right conditions, ^{10}Be can be used as a dating tool for sediments and soil cores. Beryllium becomes attached to surface soils and can produce ^{10}Be-rich soils. If the mobility and initial deposition rate can be approximated, then a depth profile can give an age estimate (Graham *et al.*, 1997). More often, however, the concentration profile is affected by climatic variations, which can produce decreases by increased dissolution (wetter) or enrichment from glacial melt or erosion runoff. Dry climates produce more windblown soils, which are usually enriched in ^{10}Be (Gu *et al.*, 1996). Graham *et al.* (1997) studied two cores from the Wanganui region of New Zealand that had been previously characterized by magnetic analysis (Turner and Kamp, 1990). A large sixfold enhancement in ^{10}Be during the Matuyama–Brunhes magnetic transition has been interpreted as only partly due to the reversal. Since the date coincides with the beginning of an interglacial period of warming, the enhancement could be due to increased erosion and soil input and/or melting of adjacent ice sheets.

In the oceans ^{10}Be both attaches to particles and has a soluble phase, which lengthens the residence time to hundreds of years. Beryllium-10 concentration profiles in deep ocean waters are consistent with that of deep sea sediments. L. Wang *et al.* (1996) have shown that the direct atmospheric contribution (as separated by leaching with NaOH) is relatively constant, while that derived from the lithosphere is influenced by climate factors that increase erosion and runoff and by changes in ocean circulation patterns.

3.3.2 Aluminum-26 in the Oceans Another potential cosmogenic nuclide for ocean studies is ^{26}Al. The majority of the ^{26}Al in the oceans is atmospheric in origin. Once it enters the ocean, it is rapidly scavenged and locally deposited. This is in contrast to ^{10}Be, which can travel considerable distances from its source before being deposited. While the atmospheric $^{26}Al/^{27}Al$ ratio is readily measured with AMS, dilution from ^{27}Al derived from eroded crustal material has put determination of ^{26}Al in bulk sediment cores close to the AMS detection limit. A new leaching technique using NaOH now allows separation of the ^{26}Al-enriched atmospheric component and brought the $^{26}Al/^{27}Al$ ratios well within the capabilities of AMS (Wang *et al.*, 1996). When compared with ^{10}Be it may be possible to distinguish the effects of ocean dynamics and climate from variations in the cosmogenic input.

3.4 Groundwater Dating

3.4.1 Natural Sources Naturally occurring (nonanthropogenic) cosmogenic radionuclides that have the potential of dating groundwaters include ^{14}C, ^{36}Cl, ^{39}Ar, ^{81}Kr, and ^{129}I. All of these are produced in the atmosphere and enter groundwaters through wet and dry deposition, either dissolved in water droplets or attached to aerosol particles. The half-life of ^{14}C (5730 y) makes it well suited for dating waters up to around 50 000 y. The potential uses of ^{14}C and their associated complications have been discussed in Sec. 3.1. Argon-39 is relatively short-lived ($t_{1/2} = 269$ y) and can be used as an indicator of more modern water. Krypton-81, with a half-life comparable to that of ^{36}Cl, has not found widespread use on account of the analytical difficulties of measurement. Should these be overcome, ^{81}Kr may prove to be a nuclide with the fewest interfering contributions as a tool for groundwater dating (Lehmann *et al.*, 1993). The remaining nuclides, ^{36}Cl and ^{129}I, have the potential to provide ages for groundwaters that are well beyond the reach of ^{14}C. They are readily measured via AMS as ratios to their stable isotopes, $^{36}Cl/Cl$ and $^{129}I/I$. Background atmospheric levels are well above the detection limits of 1×10^{-15} for $^{36}Cl/Cl$ and 1×10^{-14} for $^{129}I/I$, providing the ability to follow their decays through several half-lives. With half-lives of 3.01×10^5 y for ^{36}Cl and 1.57×10^7 y for ^{129}I, these two nuclides could provide a range of ages from 0.1 to 80 My (Fabryka-Martin *et al.*, 1987; Frölich *et al.*, 1991).

Chloride is soluble in aqueous systems and is generally considered to be a conservative tracer in groundwaters, although there may be evidence of retention of the bomb pulse in surface soils. Aside from radioactive decay, factors that influence the $^{36}Cl/Cl$ ratio are changes in the chloride and cosmogenic ^{36}Cl of the recharge water, introduction of salt from ancient trapped brines or formation waters, diffusion of salts from the aquifer material, and subsurface production of ^{36}Cl (Fabryka-Martin *et al.*, 1989; Nolte *et al.*, 1991). Much the same is true for ^{129}I, except that the atmospheric input is expected to vary less with time by reason of an additional atmospheric contribution from volcanoes emitting ^{129}I from ^{238}U fission and the large buffering capacity of the oceans for ^{129}I. The residence time of ^{129}I in the oceans is long enough for mixing to occur with little decay. This is in contrast to ^{36}Cl, which is negligible in the oceans, because of its shorter half-life, and which has a very pronounced temporal dependence on the cosmic-ray production.

The subsurface production mechanism for ^{129}I is directly as a fission product of ^{238}U. In the case of ^{36}Cl, the main subsurface production mode is from the $^{35}Cl(n,\gamma)$ reaction on chloride in the water itself. The neutrons come from the interaction of α particles, emitted in the U–Th decay chain, with light elements such as K and Ca in the surrounding minerals, via (α, n) reactions. In some settings, as in the Stripa mine in Sweden, these subsurface sources dominate and the $^{36}Cl/Cl$ and $^{129}I/I$ ratios increase with depth, limiting their direct use as dating tools (Andrews *et al.*, 1986, 1989). It may actually be possible to use the buildup of these isotopes to obtain residence-time estimates.

While subsurface production will increase isotopic ratios, the introduction of old (dead) chloride or iodide from the aquifer material or from salt in the initial recharge water will decrease the ratio. It is well documented that the chloride concentration in rainwater decreases with distance from the coast as the influence of marine chloride decreases (Bentley *et al.*, 1986; NADP/NTN, 1990). In a study of the Aquia aquifer in southern Maryland, U.S.A., the chloride profile was found to vary with flow distance in a manner consistent with the amount of marine chloride in the recharge water, as governed by changing distance from the sea in the period covering the last glacial maximum (Purdy *et al.*, 1996). The variation in total chloride concentration can be corrected for by multiplying by the $^{36}Cl/Cl$ ratio to obtain the concentration of ^{36}Cl directly, usually expressed as atoms L^{-1}. However, this quantity is now sensitive to dilution and/or concentration effects caused by changing climatic conditions at recharge (changes in rainfall and/or evapotransporation), as well as variations in the cosmic-ray flux. Unfortunately, the time scale of interest is outside the range of good flux monitors; otherwise, ^{36}Cl concentration in confined aquifers might be a good measure of paloeclimate.

The initial promise of a simple technique for dating groundwater has not been fulfilled. However, cosmogenic isotopes have provided another experimental dimension to test models of groundwater systems and to provide constraints on ages within a given set of conditions.

3.4.2 Anthropogenic Sources The activities of man in the nuclear age have produced additional sources of nuclides commonly identified with cosmogenic processes. The bomb tests in the late 1950s and 1960s produced large pulses of tritium (^3H), ^{14}C, ^{36}Cl, and ^{129}I. Atmospheric levels of tritium and ^{36}Cl have mostly returned to prebomb levels (Bentley *et al.*, 1982 and Fig. 3), with the notable exception of nuclear reactor and fuel reprocessing sites (Milton *et al.*, 1994). This is not true of ^{129}I, which is being continually released by the nuclear power industry in quantities sufficient to have maintained elevated levels. All except tritium have half-lives long enough to serve as tracers of modern input for the far future. Even tritium, with its 12-y half-life, is still readily detectable (directly or as its decay product ^3He), and its short half-life can be used for modern age determinations (Schlosser *et al.*, 1988). The so-called bomb pulse can therefore serve as a dating tool for modern input. This might better be called a tracer, but its detection puts a definite time frame to a process, be it the advance of a contaminant plume in groundwaters (Solomon *et al.*, 1995) or the movement of currents in the ocean (Santschi *et al.*, 1996).

3.5 Silicon-32: The Potential

The widespread use of ^{32}Si as a dating tool has been hampered by the difficulty in achieving $^{32}Si/Si$ ratios below modern levels through AMS (Morgenstern *et al.*, 1996). This has limited studies to events that have elevated ratios or to the use of laborious processing for decay counting, requiring large sample sizes. The half-life of ^{32}Si is also poorly known, but recent work using AMS may improve this. The great lure of ^{32}Si is that, with a half-life of around 140 y, it fills a gap in dating that is left between tritium and ^{14}C. Exploratory studies (Morgenstern *et al.*, 1995, 1996) show promise as a dating tool in glacial ice, groundwaters, and sediments, and as a tracer for ocean circulation (Somayajulu *et al.*, 1987).

4. PALEOMAGNETISM

Paleomagnetism as a dating technique relies on the fact that the Earth's molten core is not static but dynamic, causing the resultant magnetic field of the earth to vary over time. When heated to a high enough temperature, usually 500–700 °C, certain minerals, such as hematite and magnitite, acquire magnetic alignments to the Earth's magnetic field, which persist upon cooling. These remnant fields can be measured and compared with the known reversals in the Earth's magnetic field. If a single date can be fixed by an independent technique, then the correspondence of the changing fields allows dating of the remaining samples. Details of magnetic analysis and its use as a dating tool are presented in PHYSICS IN ARCHAEOLOGY and in Aitken (1990).

5. AMINOSTRATIGRAPHY AND AMINOCHRONOLOGY

Amino acids are the building blocks of proteins. Once an organism dies, chemical changes due to hydrolysis take place, which gradually break the proteins into smaller groups called peptides. Amino acids may exist as internal constituents of a peptide chain, as terminal amino acids at either end of a protein or peptide, as a cyclic structure formed from two small peptides, or as free amino acids. The specifics of the diagenetic process are dependent on protein composition and thus the organism involved. After fossilization of shell material as much as 50% of the original protein content may be lost. However, that which remains is relatively stable and can be used as a relative dating tool.

Natural amino acids (except glycine) in proteins from living organisms exist in a form which has a unique stereospecific configuration due to the spatial arrangements of their atoms. This is termed L and is distinct from another spatial configuration of the same molecule (termed D). When synthesized in a laboratory, equal amounts of L and D are formed, but nature produces only L to give a ratio of D/L of 0.0. Over time some amino acids change their configurations from L to D, such that the equilibrium ratio of D/L is equal to 1.0 in very old fossils. This process of conversion is called racemization, and the 50–50% mixture is termed racemic. When an amino acid contains two groups that have this symmetry difference, but only one changes, the term epimerization is used. This is true of leucine, which can exist

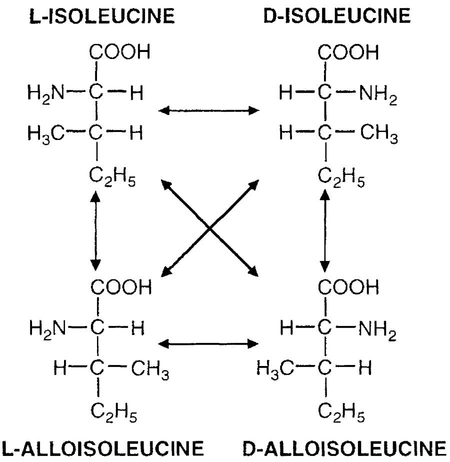

FIG. 13. Stereochemical configurations of isoleucine (after Mitterer *et al.*, 1993).

as epimers isoleucine or alloisoleucine, each of which also has L and D configurations. This difference is demonstrated in Fig. 13 (after Mitterer, 1993). The general term racemization is usually also applied to epimerization.

The rate at which racemization occurs is a very sensitive function of temperature. For this reason, it is most often used as a relative measure in aminostratigraphy, the ordering of what are termed aminozones by their D/L ratios for a specific amino acid. These zones can be dated if a specific group or zone can be correlated with a known age, as determined by an independent technique such as radiometric dating. Aminochronology is the application of these calibrated D/L ratios to predict ages for uncalibrated samples. This involves modeling the time dependence of the racemization for the appropriate temperature conditions (Wehmiller, 1993).

Racemization rates are also dependent on the specific type of organism studied. For example, under the same conditions of temperature, floraminifera from the genus *G. tumida* has a different rate of racemization from that of the genus *O. universa*. This effect is shown in Fig. 14(a), where the ratio of D-alloisoleucine to L-isoleucine is plotted as a function of age for the two genera, obtained from the same deep-sea sediment core (Müller, 1984). In some cases concordia plots for two genera can be constructed and used to determine age in the same manner as those for radiogenic nuclides (see Sec. 2.1.6). Figure 14(b) shows this for the data displayed in Fig. 14(a).

The very sensitive temperature dependence of the rate of racemization can be used in the framework of a kinetic model to predict temperature variations in Quaternary deposits during the last ice age. This technique has been applied to mollusk shells in the Mississippi Valley. Radiocarbon and thermoluminescence dating were used to obtain the time scale. Figure 15 shows the effective diagenetic temperatures (EDT) derived from modeling the rates of epimerization of isoleucine in five genera of gastropods (Oches *et al.*, 1996). The results are plotted along with the mean annual air temperature from 1951 to 1980.

Paleontological applications of aminostratigraphy and aminochronology can be found in Wehmiller (1993) and the references therein. Specific applications in archaeology are discussed in Aitken (1990).

6. CHLOROFLUOROCARBONS

An example of an anthropogenic source that can be turned into a modern dating tool is the release of compounds classified as chlorofluorocarbons (CFCs). First manufactured in the 1930s, these compounds were valued for their stability and used extensively as refrigerants and propellants until they became implicated in the destruction of stratospheric ozone. They are highly problematic because they have been accumulating rapidly in the atmosphere and have residence times of 120 y for CCl_3F (trade name F-11) and 60 y for CCl_2F_2 (trade name F-12) (Derra, 1990). It is just this stability which makes them excellent tracers of modern processes.

The detection of CFCs in groundwaters indicates post-1945 input, and studies have correlated CFC concentrations with tritium dating. Its use as an actual dating technique requires accurate profiles of the change in atmospheric CFC concentrations and the temperature of the recharge zone over time. Sorption onto soils is also a consideration. Details of the use of CFCs and the analytical techniques required can be found in Busenberg and Plummer (1992).

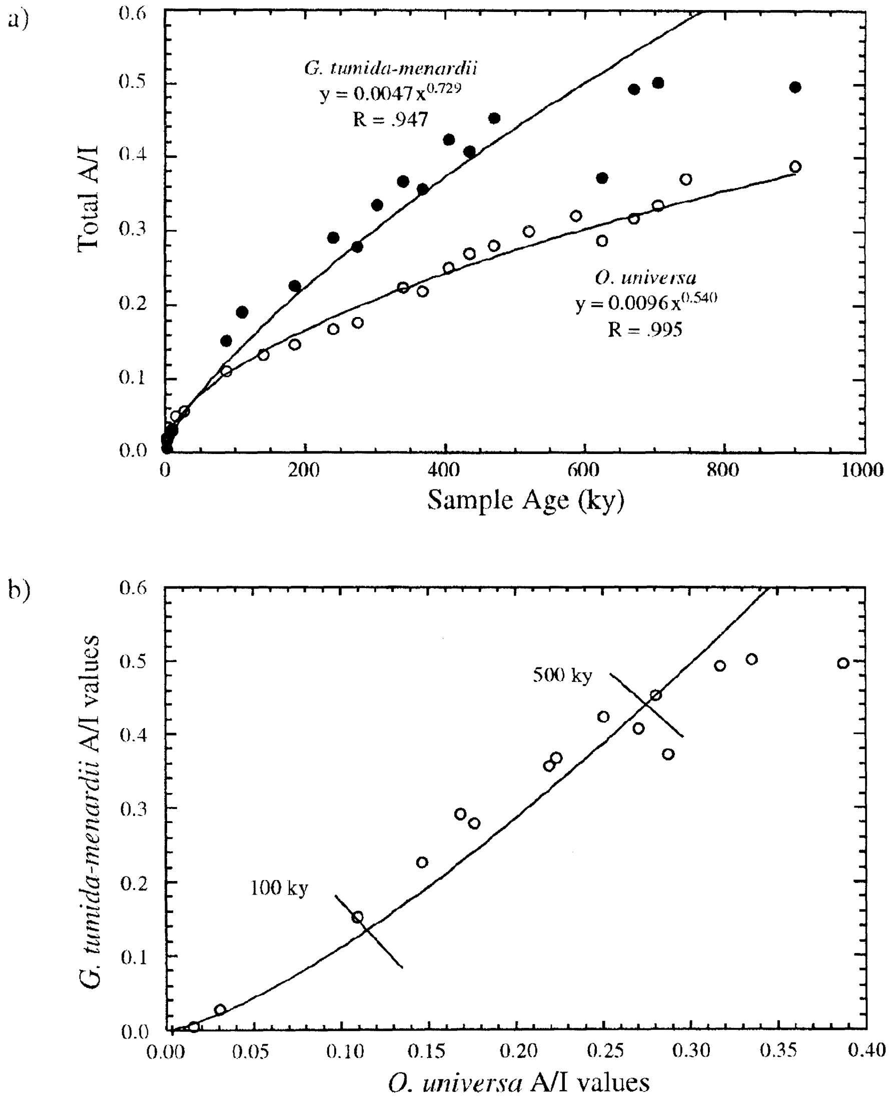

FIG. 14. (a)Ratio of D-alloisoleucine to L-isoleucine (A/I) as a function of time for floraminifera *G. tumida-menardii* (solid circles) and *O. universa* (open circles), measured in the same deep-sea sediment. (b) Concordance in the (A/I) ratios for *O. universa versus G. tumida-menardii*, plotted from the data displayed in part (a). Data from Müller (1984).

7. CONCLUSION

The application of techniques that can provide either absolute or relative chronologies is central to a wide range of research in the fields of geology, hydrology, oceanography, archaeology, paleontology, cosmology, and the environmental sciences. This article has provided an overview of techniques used today and some of the possible applications. Each has its strengths and weaknesses, requiring different sets of assumptions for validity. Hence, the study of a specific process or system is best served when a range of dating tools is

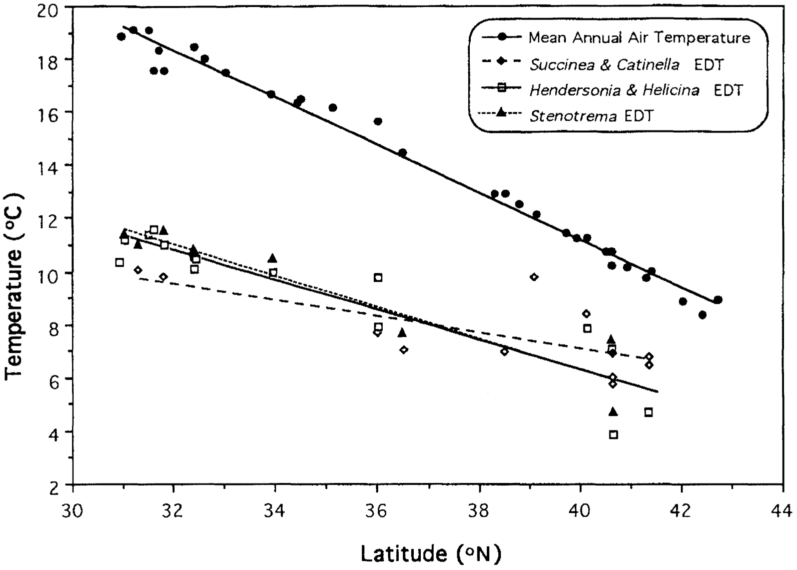

FIG. 15. Temperature profile for ice-age gastropods as a function of latitude in Mississippi Valley loess, as derived from modeling the epimerization rates of leucine. The mean annual air temperature for the period 1951–1980 is shown in the upper curve for comparison (from Oches *et al.*, 1996).

drawn upon to corroborate results. As in the cases of AMS and TIMS, analytical advances are constantly making these tools accessible to new areas of research.

GLOSSARY

Alpha Particle (α): Nucleus of a helium atom, consisting of two protons and two neutrons.

Anthropogenic: Influenced by the impact of man.

Apatite: A group of hexagonal minerals consisting of calcium phosphate together with fluorine, chlorine, hydroxyl, or carbonate.

Basalt: A dark-colored igneous rock, commonly extrusive, composed primarily of calcic plagioclase and pyroxene. Apatite and magnetite are almost always present.

Beta particle (β): A positive or negative electron that is emitted from the nucleus during radioactive decay.

Calcite: A common rock-forming mineral, $CaCO_3$, a chief constituent of limestome and most marble.

Connate: Water trapped in the interstices of sedimentary rock at the time the rock was deposited.

Dose: A measure of energy deposited in a material by ionizing radiation (α, β, γ).

Feldspar: A group of abundant rock-forming minerals of the general formula $M\mathrm{Al(Al,Si)_3O_8}$, where M can be K, Na, Ca, Ba, Rb, Sr or Fe.

Floraminifera: Large single-celled animals with a calcareous shell of calcium carbonate or calcite.

Gamma ray (γ): High-energy electromagnetic radiation emitted when a nucleus converts from a higher energy state to a lower.

Gastropod: A class of mollusks that includes snails and slugs.

Glauconite: A green mineral closely related to micas and essentially a hydrous potassium iron silicate.

Hematite: A common mineral containing Fe_2O_3, found in igneous, sedimentary and metamorphic rocks.

Hydrosphere: The waters of the Earth, as distinguished from the rocks (lithosphere), living things (biosphere) and the air (atmosphere).

Igneous: A rock or mineral that has solidified from molten material, e.g., magma.

Lithosphere: The solid portion of the earth.

Loess: A blanket deposit of buff-colored calcareous silt, homogeneous, nonstratified, weakly coherent, and porous; considered to be windblown dust of the Pleistocene age.

Magnetite: A black, opaque, strongly magnetic mineral of the group $(Fe,Mg)Fe_2O_4$.

Muon: A weakly interacting particle, similar to an electron but more massive.

Nuclide: A specific type of nucleus, with a given number of protons and neutrons.

Paleo: Old or ancient.

Pleistocene: The early epoch of the Quaternary period, which ended around 8000 y ago.

Pluton: An igneous intrusion.

Quarternary: Geologic time period from present back to 1.8 My ago.

Quartz: Crystalline silica, an important rock-forming mineral, SiO_2.

Radiogenic: Formed as the result of radioactive decay.

Radionuclide: A nuclide that is unstable and decays to another nuclide.

Tectonic: Pertaining to the forces involved in or the resulting structures of tectonics.

Tectonics: Branch of geology dealing with the broad architecture of the outer part of the earth.

Zircon: A mineral, $ZrSiO_4$.

Works Cited

Aitken, M. J. (1985), *Thermoluminescence Dating*, London: Academic Press.

Aitken, M. J. (1990), *Science-Based Dating in Archaeology*, London: Longman.

Andrews, J. N., Fontes, J.-C., Michelot, J.-L., Elmore, D. (1986), *Earth Planet. Sci. Lett.* **77**, 49–58.

Andrews, J. N., Davis, S. N., Fabryka-Martin, J., Fontes, J.-C., Lehmann, B. E., Loosli, H. H., Michelot, J.-L., Moser, H., Smith, B., Wolf, M. (1989), *Geochim. Cosmochim. Acta* **53**, 1803–1815.

Bard, E., Hamelin, B., Fairbanks, R. G., Zindler, A. (1990), *Nature* (*London*) **345**, 405–410.

Beer, J., Blinov, A., Bonani, G., Finkel, R. C., Hofmann, H. J., Lehmann, B., Oeschger, H., Sigg, A., Schwander, J., Staffelbach, T., Stauffer, B. R., Suter, M., Wolfli, W. (1990), *Nature* (*London*) **347**, 164–166.

Bentley, H. W., Phillips, F. M., Davis, S. N., Gifford, S., Elmore, D., Tudds, L. E., Gove, H. E. (1982), *Nature* (*London*) **300**, 737–740.

Bentley, H. W., Phillips, F. M., Davis, S. N., Habermehl, M. A., Airey, P. L., Calf, G. E., Elmore, D., Gove, H. E., Torgerson, T. (1986), *Water Resour. Res.* **22**, 1991–2002.

Bierman, P., Gillespie, A., Caffee, M., Elmore, D. (1995), *Geochim. Cosmochim. Acta* **59**, 3779–3798.

Bollhöfer, A., Mangini, A., Lenhard, A., Wessels, M., Giovanoli, F., Schwarz B. (1994), *Environ. Geol.* **24**, 267–274.

Brannon, J. C., Cole, S. C., Podosek, F. A., Ragan, V. M., Coveney Jr., R. M., Wallace, M. W., Bradley, A. J. (1996), *Science* **271**, 491–493.

Busenberg, E., Plummer, N. (1992), *Water Resour. Res.* **28**, 2257–2283.

Cerling, T. E., Craig, H. (1994), *Ann. Rev. Planet. Sci.* **22**, 273–317.

Clark, D. H., Bierman, P. R., Larsen, P. (1995), *Quarter. Res.* **44**, 367–377.

Creaser, R. A., Papanastassiou, D. A., Wasserburg, G. J. (1991), *Geochim. Cosmochim. Acta* **55**, 397–401.

Derra, S. (1990), *Res. Develop.* **32**, 54–66.

Dickin, A. P. (1995), *Radiogenic Isotope Geology*, Cambridge, UK: Cambridge University Press.

Ding, G.-J., Kilius, L. R., Wilson, G. C., Zhao, X.-L., Rucklidge, J. C. (1997), *Nucl. Instrum. Methods Phys. Res.* **B123**, 424–430.

Drimmie, R. J., Aravena, R., Wassenaar, L. I., Fritz, P., Hendry, M. J., Hut, G. (1991), *Appl. Geochem.* **6**, 381–392.

Edwards, R. L., Chen, J. H., Wasserburg, G. J. (1987), *Earth Planet. Sci. Lett.* **81**, 175–192.

Edwards, R. L., Beck, J. W., Burr, G. S. (1993) *Science* **260**, 962–967.

Edwards, R. L., Cheng, H., Murrell, M. T., Goldstein, S. J. (1997), *Science* **276**, 782–786.

Elmore, D., Phillips, F. M. (1987), *Science* **236**, 543–550.

Esperança, S., Carlson, R. W., Shirey, S. B., Smith, D. (1997), *Geology* **25**, 651–654.

Fabryka-Martin, J. T., Davis, S., Elmore, D. (1987), *Nucl. Instrum. Methods Phys. Res.* **B29**, 361–371.

Fabryka-Martin, J. T., Davis, S. N., Elmore, D., Kubik, P. W. (1989), *Geochim. Cosmochim. Acta* **53**, 1817–1823.

Faure, G. (1986), *Principles of Isotope Geology*, 2nd ed., New York: John Wiley & Sons Inc.

Freydier, C., Ruiz, J., Chesley, J., McCandless, T., Munizaga, F. (1997), *Geology* **25**, 775–778.

Fröhlich, K., Ivanovich, M., Andrews, J. N., Davis, S. N., Drimmie, R. J., Fabryka-Martin, J., Florkowski, T., Fritz, P., Lehmann, B., Loosli, H. H., Nolte, E. (1991), *Appl. Geochem.* **6**, 465–472.

Gillespie, A. R., Bierman, P. R. (1995), *J. Geophys. Res.* **100**, 24 637–24 649.

Graham, I. J., Ditchburn, R. G., Sparks, R. J., Whitehead, N. E. (1997), *Nucl. Instrum. Methods Phys. Res.* **B123**, 301–318.

Gu, Z. Y., Lal, D., Liu, T. S., Southon, J., Caffee, M. W., Guo, Z. T., Chen, M. Y. (1996), *Earth Planet. Sci. Lett.* **144**, 273–287.

Innocent, C., Parron, C., Hamelin, B. (1997), *Geochim. Cosmochim. Acta* **61**, 3753–3761.

Johnson, C. (1997), *Geology* **25**, 623–626.

Knoll, G. F. (1979), *Radiation Detection and Measurement*, New York: John Wiley & Sons, Inc.

Kretschmer, W., Anton, G., Bergmann, M., Finckh, E., Kowalzik, B., Klein, M., Leigart, M., Merz, S., Morgenroth, G., Piringer, I., Küster, H., Low, R. D., Nakamura, T. (1997), *Nucl. Instrum. Methods Phys. Res.* **B123**, 455–459.

Lal, D., Peters, B. (1967), in: K. Sitte (Ed.), *Handbuch der Physik*, Vol. 46/2, Berlin: Springer, pp. 551–612.

Lal, D. (1991), *Earth Planet. Sci. Lett.* **104**, 424–439.

Laslett, G. M., Green, P. F., Duddy, I. R., Gleadow, A. J. W. (1987), *Chem. Geol.* **65**, 1–13.

Lehmann, B. E., Davis, S. N., Fabryka-Martin, J. T. (1993), *Water Resour. Res.* **29**, 2027–2040.

Libby, W. F. (1955), *Radiocarbon Dating*, Chicago: Univ. of Chicago Press.

Lui, B., Phillips, F. M., Fabryka-Martin, J. T., Fowler, M. M., Stone, W. D. (1994), *Water Resour. Res.* **30**, 3115–3125.

Mazaud, A., Laj, C., Bender, M. (1994), *Geophys. Res. Lett.* **21**, 337–340.

Michlovich, E. S., Vogt, S., Masarik, J., Reedy, R. C., Elmore, D., Lipschutz, M. E. (1994), *J. Geophys. Res.* **99**, 187–194.

Milton, G. M., Andrews, H. R., Causey, S. E., Chant, L. A., Cornett, R. J., Davies, W. G., Greiner, B. F., Koslowsky, V. T., Imahori, Y., Kramer, S. J., McKay, J. W., Milton, J. C. D. (1994), *Nucl. Instrum. Methods Phys. Res.* **B92**, 376–379.

Mitterer, R. M. (1993), in: M. H. Engel, S. A. Macko, (Eds.), *Organic Geochemistry: Principles and Applications*, New York: Plenum, Chap. 35.

Morgenstern, U., Gellerman, R., Hebert, D., Börner, I., Stolz, W., Vaikmäe, R., Rajamäe, R., Putnik, H. (1995), *Chem. Geol.* **120**, 127–134.

Morgenstern, U., Taylor, C. B., Parrat, Y., Gäggeler, H. W., Eichler, B. (1996), *Earth Planet. Sci. Lett.* **144**, 289–296.

Müller, P. J. (1984), *"Meteor" Forsch.-Ergebnisse, Reihe C.* **38**, 25–47.

Murphy, E. M., Davis, S. N., Long, A., Donahue, D., Jull, A. J. T. (1989), *Water Resour. Res.* **25**, 1893–1905.

NADP/NTN (1990), National Atmospheric Deposition Program, *Annual Data Summary: Precipitation Chemistry in the United States*, Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO, U.S.A.

Nishiizumi, K., Kohl, C. P., Shoemaker, J. R., Arnold, J. R., Klein, J., Fink, D., Middleton, R. (1991), *Geochim. Cosmochim. Acta* **55**, 2699–2703.

Noble, S. R., Aspden, J. A., Jemielita, R. (1997), *GSA Bull.* **109**, 789–798.

Nolte, E., Krauthan, P., Korschinek, G., Maloszewski, P., Fritz, P., Wolf, M. (1991), *Appl. Geochem.* **6**, 435–445.

Oches, E. A., McCoy, W. D., Clark, P. U. (1996), *GSA Bull.* **108**, 892–903.

Ortega-Rivera, A., Farrar, E., Hanes, J. A., Archibald, D. A., Gastil, R. G., Kimbrough, D. L., Zentilli, M., Lopez-Martinez, M., Feraud, G., Ruffet, G. (1997), *GSA Bull.* **109**, 728–745.

Phillips, F., Zreda, M. G., Smith, S. S., Elmore, D., Kubik, P. W., Dorn, R. I., Roddy, D. R. (1991), *Geochim. Cosmochim. Acta* **55**, 2695–2698.

Purdy, C. B., Helz, G. R., Mignerey, A. C., Kubik, P. W., Elmore, D., Sharma, P., Hemmick, T. (1996), *Water Resour. Res.* **32**, 1163–1171.

Raisbeck, G. M., Yiou, F., Bourles, D., Kent, D. V. (1985), *Nature* (*London*) **315**, 315–317.

Raisbeck, G. M., Zhou, Z. Q., Kilius, L. R. (1994), *Nucl. Instrum. Methods Phys. Res.* **B92**, 436–439.

Ravenhurst, C. E., Willet, S. D., Donelick, R. A., Beaumont, C. (1994), *J. Geophys. Res.* **99**, 20023–20041.

Santschi, P. H., Schlink, D. R., Oktay-Marshall, S., Corapcioglu, O., Fehn, U., Sharma, P. (1996), *Deep Sea Res.* **43**, 259–265.

Schlosser, P., Stute, M., Dörr, H., Sonntag, C., Münnich, K. O. (1988), *Earth Planet. Sci. Lett.* **89**, 352–363.

Schlosser, P., Kromer, B., Ekwurzel, B., Bönisch, G., McNichol, A., Schneider, R., von Reden, K., Ostlund, H. G., Swift, J. H. (1997) *Nucl. Instrum. Methods Phys. Res.* **B123**, 431–437.

Smith, P. E., Farquhar, R. M. (1989), *Nature* (*London*) **341**, 518–521.

Smoliar, M. I., Walker, R. J., Morgan, J. W. (1996), *Science* **271**, 1099–1102.

Solomon, D. K., Poreda, R. J., Cook, P. G., Hunt, A. (1995), *Ground Water* **33**, 988–996.

Somayajulu, B. L. K., Rengarajan, R., Lal, D., Weiss, R. F., Craig, H. (1987), *Earth Planet. Sci. Lett.* **85**, 329–342.

Steig, E. J., Polissar, P. J., Stuiver, M., Grootes, P. M., Finkel, R. C. (1996), *Geophys. Res. Lett.* **23**, 523–526.

Stuiver, M., Reimer, D. J. (1993), *Radiocarbon* **35**, 215–230.

Sutton, S. R. (1985), *Geophys. Res.* **90**, 3690–3700.

Synal, H.-A., Beer, J., Bonani, G., Suter, M., Wölfli, W. (1990) *Nucl. Instrum. Methods Phys. Res.* **B52**, 483–488.

Taylor, R. E. (1987), *Anal. Chem.* **59**, 317A–331A.

Titayeva, N. A. (1994), *Nuclear Geochemistry*, Moscow: Mir Publishers.

Turner, G. M., Kamp, P. J. J. (1990), *Earth Planet. Sci. Lett.* **100**, 42–50.

von Reden, K. F., McNichol, A. P., Peden, J. C., Elder, K. L., Gagnon, A. R., Schneider, R. J. (1997), *Nucl. Instrum. Methods Phys. Res.* **123**, 438–442.

Wang, L., Ku, T. L., Luo, S., Southon, J. R., Kusakabe, M. (1996), *Geochim. Cosmochim. Acta* **60**, 109–119.

Wang, Y., Amundson, R., Trumbore, S. (1996), *Quater. Res.* **45**, 282–288.

Wehmiller, J. F. (1993), in: M. H. Engel, S. A. Macko, (Eds.), *Organic Geochemistry: Principles and Applications*, New York: Plenum, Chap. 36.

Wolf, R. A., Farley, K. A., Silver, L. T. (1996), *Geochim. Cosmochim. Acta* **60**, 4231–4240.

Wolf, R. A., Farley, K. A., Silver, L. T. (1997), *Geology* **25**, 65–68.

Further Reading

Aitken, M. J. (1985), *Thermoluminescence Dating*, London: Academic Press.

Aitken, M. J. (1990), *Science-Based Dating in Archaeology*, London: Longman.

Choppin, G., Rydberg, J., Liljenzin, J. O. (1995), *Radiochemistry and Nuclear Chemistry*, 2nd ed., Oxford: Butterworth-Heinemann Ltd.

Dickin, A. P. (1995), *Radiogenic Isotope Geology*, Cambridge, UK: Cambridge University Press.

Faure, G. (1986), *Principles of Isotope Geology*, 2nd ed., New York: John Wiley & Sons Inc.

Knoll, G. F. (1979), *Radiation Detection and Measurement*, New York: John Wiley & Sons, Inc.

Tuniz, C., Bird, J. R., Fink, D., Herzog, G. F. (1998), *Accelerator Mass Spectrometry, Ultra Sensitive Analysis for Global Science*, Boca Raton, FL: CRC Press.

INTERFEROMETRIC TECHNIQUES

Joanna Schmit, *Veeco Instruments Incorporated, Tucson, Arizona, U.S.A.*

Katherine Creath, *Creath Optineering Services, Tucson, Arizona, U.S.A.*

Introduction 67
1. **Phase Across the Plane** 68
1.1 Fringe Tracking 69
1.2 Interferogram Plus Intensities 69
1.3 Temporal Phase Measurement—Temporal Heterodyning 69
1.4 Single Interferogram Analysis—Spatial Heterodyning 71
1.4.1 Fourier-Transform Technique 71
1.4.2 Synchronous Detection 71
1.5 White-Light Interferogram Analysis 71
1.6 Summary 72
2. **Measurement Sensitivity and Limitations** 72
2.1 Measurement Sensitivity 72
2.2 Removal of Systematic Errors 72
2.3 Reduction of Nonsystematic Errors 73
3. **Form Measurement** 73
3.1 Optical-System and -Component Testing 73
3.1.1 Flats 73
3.1.2 Spherical and Aspheric Surfaces 73
3.1.3 Prisms 74
3.1.4 Light Sources 74
3.1.5 Homogeneity 74
3.2 Testing Engineering Surfaces 74
3.2.1 Fringe Projection 74
3.2.2 Moiré 75
3.2.3 Interference Microscopy 75
3.2.4 Fringes of Equal Chromatic Order Interferometry 75
4. **Phase at One Point** 75
4.1 Heterodyne Interferometry 76
4.2 Fringe Counting 76
4.3 Phase Locking 77
5. **Applications of Phase at One Point** 78
5.1 Length Measurement 78
5.2 Gravity-Wave Detection 78
5.3 Other 78
6. **Interference Spectroscopy** 79
6.1 Fabry–Pérot Spectrometer 79
6.2 Fourier-transform Spectroscopy 79
6.3 Spectral Imaging 80
7. **Interference Imagery** 80
7.1 Michelson Stellar Interferometry 81
7.2 Stellar Intensity Interferometry 82
7.3 Radio Interferometry 82
7.4 Stellar Speckle Interferometry 82
8. **Other Applications** 82
8.1 Nondestructive Testing with Interferometry 82
8.1.1 Holographic Interferometry 82
8.1.2 Speckle Interferometry 83
8.1.3 TV Holography 83
8.1.4 Grating Interferometry 83
8.2 Optical-Fiber Interferometry 83
8.3 Refractometry 84
8.4 Ellipsometry 84
8.5 Atom Interferometry 84
Glossary 84
Works Cited 84
Further Reading 85

INTRODUCTION

This article focuses on interferometric techniques, which are various methods for obtaining and analyzing an interferogram (fringe pattern) in an interferometer (see Interferometers and Interferometry). An interferogram is the recorded interference signal between two beams of light exiting from the same radiation source. An interferogram carries information about the phase (the difference between the two wave fronts of the beams), the spectral content of the source, and its spatial distribution. These three phenom-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ena, and the major interferometric techniques used to collect these data, provide the general topics discussed in this article. First is a discussion of phase measurement (see Secs. 1 and 4); next is a brief examination of the spectral distribution of the source (see Sec. 6), followed by a look at the spatial distribution of the source (see Sec. 7). By combining techniques it is possible to measure more than one of these phenomena at a time (see Sec. 6.3).

In the discussion of phase measurement a distinction between techniques that detect phase at a single point and those that detect across the plane of an interferogram is made. Techniques that measure phase are used to image test-object characteristics such as surface topography, optical-system quality, homogeneity of media, relative position of targets, and existence of gravity waves. Measuring the spectral distribution of radiation is one of the oldest applications of interferometry; a description of two basic interference spectrometers is provided in this article. The techniques that astronomy has developed for testing the spatial distribution of source radiation in the measurement of stellar diameters are also briefly described. Finally, because phase can represent so many characteristics, a number of techniques adapted from those based on phase measurement are mentioned. Experimental mechanics tests for object displacements; refractometry and ellipsometry measure the material refractive index and thin-film thickness; and optical-fiber interferometers sense temperature and rotation. All these are applications based on interferometry. This article covers neither interference phenomena nor basic types of interferometers, as these topics were discussed in Interferometers and Interferometry and Optics, Linear.

1. PHASE ACROSS THE PLANE

The techniques that measure the phase across the plane employ a two-beam interferometer to record an interference signal in the form of an interferogram (fringe pattern). These interferograms, consisting of fringes of equal thickness called Fizeau fringes (see Interferometers and Interference), carry information about the phase, which corresponds to the relative difference between the test and the reference wave front. Phase variation across the plane can represent surface shape, material homogeneity or thickness, or optical-component quality. A simple way to approximate the maximum deviation of phase from the plane involves introducing tilt between the test and reference wave fronts. In this way almost straight and parallel fringes are introduced into the interferogram, Fig. 1(a). Phase deviation from the plane is seen as a departure from straight lines in the fringe pattern, Fig. 1(b). In the two interferometers most commonly used for surface contouring, the Fizeau and Twyman–Green (see Interference and Interferometry), the spacing between two neighboring fringes corresponds to a surface height difference of 1/2 of the source wavelength used for testing (1 fringe $= \lambda/2$). The maximum surface height deviation from the plane, Fig. 1(b), can be found by measuring the maximum fringe deviation from a straight line $[\Delta(x, y)]$ and the average spacing between two neighboring fringes (S). However, the relationship between fringe spacing and surface height depends not only on the wavelength used for measurement but also the interferometric configuration and the indices of refraction of the media the beam passes through. Interferometric techniques used to measure the phase across the plane typically use a coherent light source such as a laser so as to achieve good-contrast fringes; these techniques may involve single or multiple interferogram analysis to retrieve the phase.

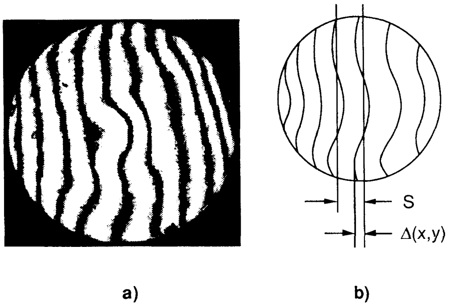

FIG. 1. (a) Typical interferogram for surface contouring. (b) Surface height deviation from a plane $= (\lambda/2)[\Delta(x, y)/S]$.

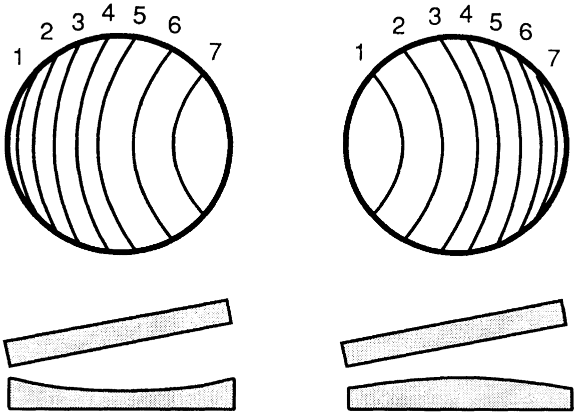

FIG. 2. Interferograms for (a) concave and (b) convex test surfaces relative to a flat reference surface.

1.1 Fringe Tracking

Until the late 1970s, fringe tracking (Yatagai, 1993) was the main way to analyze interferograms quantitatively, and although this method required only a single interferogram, it was relatively imprecise; the accuracy obtained was only one tenth of the fringe spacing. Fringe tracking works by recording the positions of the fringe intensity maxima and minima; the rest of the data are interpolated in order to create a phase map across the plane. Fringe tracking requires that the fringes be ordered so as to reflect changes in the phase between fringes, a process simplified by introducing tilt between the reference and test wave fronts. However, the direction of the tilt must be known beforehand in order to determine the shape of the tested surface correctly, *i.e.*, concave or convex (Fig. 2). If closed fringes are present in the interferogram, then fringe ordering also requires some knowledge about the tested element.

1.2 Interferogram Plus Intensities

The phase across the plane can be determined at each point without interpolation if the intensity distribution of each beam is recorded separately and one interferogram is taken. After these three intensity distributions have been determined, the phase variation can be calculated from the following equation that describes the intensity distribution I in an interferogram:

$$I = I_1 + I_2 + 2(I_1 I_2)^{1/2} \cos \varphi(x, y), \qquad (1)$$

where I_1 and I_2 are the intensities of the two beams, $2(I_1 I_2)^{1/2}$ is the interference visibility (modulation) and $\varphi(x, y)$ is the measured phase. Michelson or Mach–Zehnder interferometers (see INTERFERENCE AND INTERFEROMETRY) can use this method since the beams in the reference and test arms can be blocked separately. However, blocking each beam separately is not always possible, as when using a Fizeau interferometer, and alternative methods for retrieving the phase must be used.

1.3 Temporal Phase Measurement—Temporal Heterodyning

Temporal phase-measurement techniques (also known as temporal heterodyning, phase stepping, or phase shifting) introduce multiple times a known increment in the relative difference between the test and reference beams, called a phase shift. At least three interferograms are necessary to determine the phase encoded in the intensity distribution, Eq. (1), that can be rewritten as

$$I = I_1 + I_2 + 2(I_1 I_2)^{1/2} \cos[\varphi(x, y) + m\tau] \qquad (2)$$

where τ is the phase shift induced m times. The

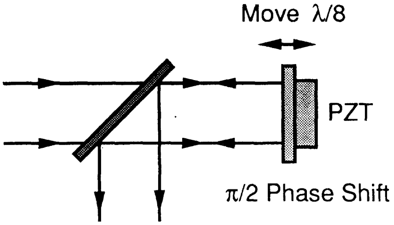

FIG. 3. Phase shift generated using a mirror moved by a piezo-electric transducer (PZT).

most common way to accomplish this phase shift is by changing the optical path difference between the interfering beams through a shift of the reference mirror along the optical axis, Fig. 3, although other ways exist—for example, by tilting a glass plate, moving a grating, rotating a half-wave plate or analyzer, or using an acousto-optic modulator (Creath, 1993). While some of these ways of shifting the phase alter the optical path between the interfering beams, others change the optical phase of one of the beams; nevertheless, both methods lead to an incremental difference (phase shift) between beams. Figure 4 shows three typical interferograms registered in the temporal phase measurement technique.

The measurement precision can be improved using more samples. An example of a common algorithm employing five intensity samples $I_1, \ldots, I_5$ with a relative $\pi/2$ phase shift to retrieve the phase $\varphi(x, y)$ is (Creath, 1993)

$$\varphi(x, y) = \arctan\left(\frac{I_1 + I_5 - 2I_3}{2I_2 + 2I_4}\right) \tag{3}$$

Because an arctangent function is not continuous and gives multiple solutions, an unwrapping procedure is employed to remove the discontinuities in the retrieved phase and essentially automatically orders the fringes (Robinson, 1993). Unwrapping procedures assume that the maximum phase discontinuity equals the difference between two neighboring solutions (2π, which is a property of the arctangent function). The values of the phase, which are measured in radians, are then converted into the values that the phase represents—for example, surface height or thickness variations. Temporal phase measurement is 10–100 times more precise than tracking the position of the fringe maxima and minima, since the repeatability of temporal phase measurement is 1/100–1/1000 of a wavelength.

Temporal phase-measurement techniques generally require that each interference fringe be sampled at least twice per fringe spacing; this constraint limits the slope of the phase that can be measured. However, a few techniques work around this limitation (Greivenkamp and Bruning, 1992). An often-used solution, white-light interferometry (see Sec. 1.5), is suitable when measuring rough surfaces or surfaces with discontinuities where sampling twice per fringe is not possible (Larkin, 1996).

Finally, although temporal phase-measurement techniques usually capture interferograms sequentially in time, and only phase not varying in time can be analyzed, a few techniques modify the configuration of the interferometer so the that *time-varying phase* can be measured, *e.g.*, gas flow. These techniques record interferograms simultaneously using multiple cameras. However, accomplishing

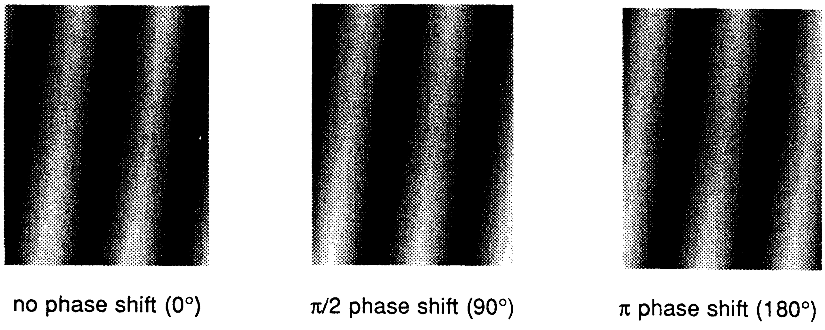
no phase shift (0°) π/2 phase shift (90°) π phase shift (180°)

FIG. 4. Interferograms showing sequence of phase shifts.

this end requires complicated hardware and alignment is critical (Kujawinska, 1993).

1.4 Single Interferogram Analysis—Spatial Heterodyning

Measuring time-varying phase can also be accomplished through the analysis of a single interferogram. Two common techniques discussed below require that a large tilt be induced between the reference and test wave fronts and that the direction of the tilt or knowledge about the tested object be known beforehand. In addition, like the temporal phase-measurement techniques described above, these techniques also require a phase unwrapping procedure in order to obtain a continuous phase map, and each fringe must be sampled at least twice per fringe spacing.

1.4.1 Fourier-Transform Technique In the Fourier-transform technique (Takeda, 1990) the discrete Fourier transform of a single interferogram is calculated revealing three distinct spectral orders, typical for a consinusoidal function, which also represents fringes (see FOURIER AND OTHER MATHEMATICAL TRANSFORMS; Ramirez, 1985). The number of almost parallel fringes from the induced tilt between the reference and test wave fronts must be large enough to separate the spectral orders to enable filtering one of these orders at the spatial frequency of the fringes. The inverse Fourier transform is then performed and the phase encoded in the interferogram is retrieved from the arctangent function of the real and imaginary parts of the inverse Fourier transform.

1.4.2 Synchronous Detection Synchronous detection (Womack, 1984) calculates the phase by sequentially applying a temporal phase-measurement technique algorithm to a few consecutive intensity samples across the plane of a single interferogram. A known phase shift is induced between adjacent sampling points, rather than between interferograms as discussed above, by tilting one wave front with respect to the other. This phase shift manifests itself as a set of straight parallel fringes. Fringe spacing must be chosen carefully so that the phase difference between adjacent sampling points is close to the phase shift required by the algorithm employed. Uniform sample reflectivity and intensity of both beams are assumed in this technique.

1.5 White-light Interferogram Analysis

Interference fringes can also be observed by use of a light source with a low degree of coherence such as a halogen lamp. Interferograms obtained with these "white light" sources show fringes of good modulation only when the optical paths of the two beams match closely. This modulation characteristic becomes important when one is trying to determine large discontinuities in phase where temporal phase-shifting interferometry is unable to assign the correct order to the fringes. *White-light interferometry* (also known as *low-coherence interferometry*) allows for the easy identification of the fringe with the best visibility because the contrast falls off so quickly. For example, Figs. 5(a) and 5(b) show interferograms of a step object; the interferogram in Fig. 5(a) is achieved with a laser as a source, and in Fig. 5(b) a white light source is used. In Fig. 5(a) there are multiple solutions to the ordering of the fringes. Thus determining the height of the step (size of the phase discontinuity) is impossible without additional information. In Fig. 5(b) the fringes can be ordered fairly simple by tracking the highest visible fringe, and the step height can then be easily determined. White-light interferometry is generally used for measuring rough surfaces or any surface with discontinuities where sampling twice per fringe is not possible.

To measure a surface using white-light interferometry, the surface is scanned in height relative to the interferometer while frames of data are acquired. As the separation between the surface and the interferometer changes, the interference fringes will intersect heights on the test surface where the optical paths are equal. A complete scan conceptually includes a three-dimensional volume of data corresponding to the location of the white-light fringes in space. The highest-contrast fringe or the peak of the fringe modulation can be used to determine the height at a single point on the test surface. This simple approach has spurred a number of new techniques for determining surface structure (Larkin, 1996). Some techniques follow

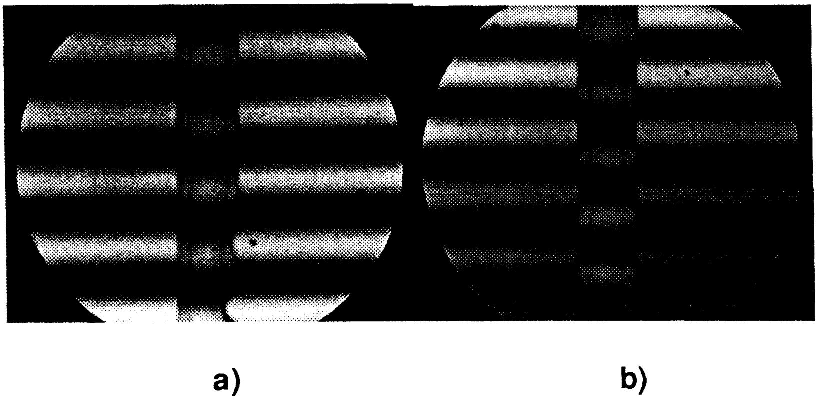

FIG. 5. Interferogram of a line feature with (a) a laser source and (b) a white-light source.

the highest-contrast fringe and others look at the peak of the fringe modulation. The relative position of these two quantities depends upon the phase change upon reflection of the light from the test surface.

1.6 Summary

The various interferogram techniques described in this section complement one other. The temporal technique and the single-interferogram techniques are used for high-precision measurement of smooth surfaces, while the fringe-modulation peak-sensing technique can examine very rough and discontinuous surfaces. Moreover, interferometric techniques analyze not only static but also time-varying phase. Techniques such as synchronous detection, Fourier transform, and temporal-phase measurement employing a spatial separation of interferograms can measure time-varying phase.

2. MEASUREMENT SENSITIVITY AND LIMITATIONS

2.1 Measurement Sensitivity

The sensitivity of an interferometric measurement can be changed by altering the test wavelength, the angle of incidence of the test beam illumination, and the number of passes through or reflections from the tested element. Varying the sensitivity of the interferometer allows for the measurement of rougher or smoother test objects. For example, a source that emits a shorter wavelength will produce a greater number of fringes in an interferogram; this choice of source wavelength increases the sensitivity of the interferometer and allows for the testing of smoother objects with high precision. A large angle of incidence of test beam illumination decreases sensitivity by producing fewer fringes in the interferogram and allows for rougher test-object measurement.

2.2 Removal of Systematic Errors

The reference wavefront in an interferometric setup is not ideal, for it carries information about imperfections in both the reference mirror and the optical system of the interferometer. If an ideal surface were available, all systematic errors could be measured and then subtracted. However, because such a surface does not exist, a set of multiple measurements is necessary in order to subtract these systematic errors (Schultz and Schwider, 1976). In addition, because various interferogram analysis techniques contain systematic sources of error, ways have been developed to decrease this error. For example, phase-shifting interferometry is vulnerable to phase-shifter miscalibration. Methods of calibrating the phase shifter precisely and the use of error-reducing algorithms help reduce any systematic errors (Creath, 1993).

2.3 Reduction of Nonsystematic Errors

High-precision interferometric measurements require special laboratory conditions. Care needs to be taken to ensure vibration and acoustic isolation as well as temperature stability and controlled air flow within the exposed optical path. If the laboratory conditions are not well controlled, nonsystematic errors may dominate a measurement. However, as long as these errors are random, statistically independent measurements can be averaged to reduce their effects and improve measurement precision and accuracy.

3. FORM MEASUREMENT

Testing form requires the measurement of phase across the plane. Form is defined as object shape or optical-system quality, in order to differentiate it from other measurable quantities such as object displacement and vibration. This section divides the examination of form measurement into optical-component testing and engineering-surface testing. The optical-component testing section is broken down into the types of components tested because similar components are tested in similar interferometric configurations. The engineering-surface testing section is organized around measurement techniques because the wide range of engineering surfaces requires a variety of different testing methods. The interferometric technique used to measure form depends on both the features of the tested object, such as its size, shape, or roughness, and the type of data that needs to be obtained.

3.1 Optical-System and -Component Testing

Interferometric techniques have proven extremely useful in testing optical systems and components, especially with the strict tolerances imposed on high-precision components and the smooth wave fronts generated by optical systems. The techniques detailed below for testing optical elements are applicable to most types of interferometers but are mainly used with the Fizeau or Twyman–Green interferometers (see INTERFEROMETERS AND INTERFEROMETRY, Secs. 2.1 and 2.4). The two interferometers have different strengths. The Fizeau interferometer is less sensitive to vibration than the Twyman–Green because the two optical paths of the Fizeau interferometer are virtually common. However, because the test and reference optical paths differ in length, the Fizeau requires a source with higher coherence than a Twyman–Green does.

3.1.1 Flats Optical flats are typically measured against a reference flat in a collimated beam. Often the flatness is specified as the maximum surface irregularity over a given surface area. Commercially available instruments can routinely measure surface irregularities as small as 1/100 of the source wavelength. If more accurate measurements are needed, flat surfaces can be tested relative to a liquid surface (Bünnagel, 1956); the surface of the liquid, which must be clear and highly viscous, works as a reference flat because its radius equals that of the Earth's and for a diameter of 0.5 m the maximum error is $\lambda/100$ for $\lambda = 500\,\text{nm}$. However, practicalities such as the influence of vibration and dust settling on the surface are problems. If a good flat is not available to test against, three flats can be tested in triple combinations in order to obtain absolute line profiles of each flat (Dew, 1966; Fritz, 1984).

3.1.2 Spherical and Aspheric Surfaces Spherical surfaces may be tested for their departure from spherical shape by comparing them with an ideal spherical reference wave front. When a spherical surface is tested in a Fizeau interferometer, Fig. 6, a transmission reference sphere is placed in the common path of the beams and the radius of the beam incident on the tested spherical surface must fill and match the radius of that surface.

Aspheric surfaces are difficult to test against a spherical wave front because they may depart too much from a spherical shape and generate too many fringes to resolve the interferogram. Null lenses and computer-generated holograms or real holograms (Offner and Malacara, 1992) are used to produce a wave front matching a perfect aspheric wave front, thereby reducing the number of fringes in the interferogram. Shearing interferometry reduces the number of fringes by interfering the tested wave front with its own replica, which can be

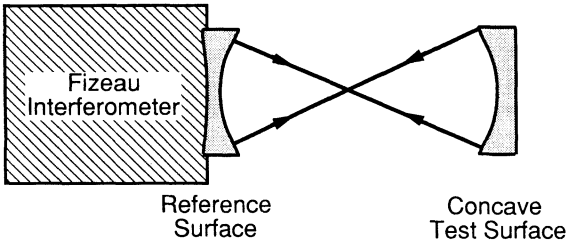

FIG. 6. Configuration for testing a concave spherical surface in a Fizeau interferometer.

displaced laterally, rotated, radially expanded or contracted. The number of fringes varies with the amount of shear, and fringes do not represent the shape of the aspheric surface directly (Malacara, 1992; Mantravadi, 1992).

3.1.3 Prisms Although prism angles may be measured with an autocollimator and a goniometer, interferometers are the most accurate method for measuring angle deviations. The procedure simply involves inserting the test prism and return flat into the test arm of the interferometer. Prisms with roof angles of 90° may be put into the test arm with (Fig. 7) or without the return mirror, as the beam can be retroreflected back to the system. The maximum deviation from 90° that a phase measuring interferometer can measure is about 1 min of arc.

3.1.4 Light Sources The quality of an optical wave front exiting from a light source may be tested using a point-diffraction interferometer or a special configuration of a Mach–Zehnder interferometer (Greivenkamp and Bruning, 1992). Both setups induce a spherical reference wave front through the use of a pinhole in the reference arm.

3.1.5 Homogeneity Homogeneity can be measured by preparing a plane-parallel sample and placing it in the test arm of an interferometer. A sample does not have to be specially prepared if it is submerged in a refractive-index–matching oil. Systematic errors can be subtracted by making a measurement before placing the sample in the test arm. Surface-independent measurements (Schwider, 1990) require multiple takes to remove variations due to the test sample's surface; this is done to isolate the refractive-index variations in the sample.

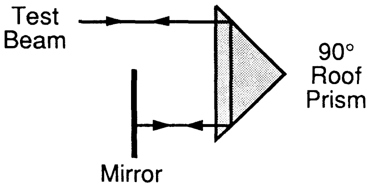

FIG. 7. Configuration for testing a 90° roof prism.

3.2 Testing Engineering Surfaces

Interferometry can also be used to measure the shape and properties of surfaces not used as optical components, such as machined metal or molded plastics. Because testing engineering surfaces is rather a specialized application of interferometric techniques, this section focuses on the methods developed to test these surfaces. The types of surfaces to be tested vary from very large ones that require overall testing of shape to very small ones that require a high resolving power to measure their microroughness. Speckle interferometry or holography may also be used for testing engineering surfaces, but since these methods can also be used for determining object displacement and vibration, two parameters important in experimental mechanics, they are discussed in the nondestructive testing section.

3.2.1 Fringe Projection Many engineering surfaces are too complex to be measured with standard interferometric techniques. Rowe and Welford (1967) devised a simple approach for measuring surface contours by projecting interference fringes or a grating onto an object and then viewing the fringes from another angle. Fringe projection is related to optical triangulation, which uses a single point of light and light sectioning; a single line is projected onto an object and then viewed from a different direction.

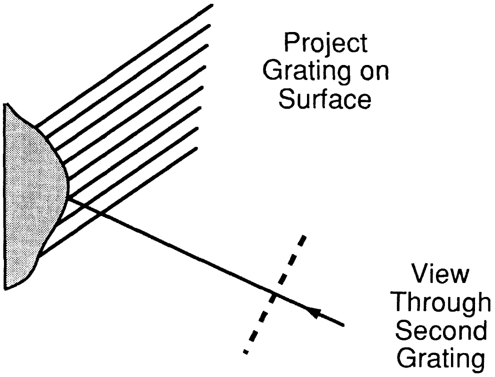

FIG. 8. Moiré technique for testing engineering surfaces.

3.2.2 Moiré In optics the term moiré (Rayleigh, 1874) refers to a pattern arising when two gratings of approximately equal spacing are superimposed. Interferometry using moiré (Patorski, 1993) can be implemented in a number of ways. A common approach is to project interference fringes (or a grating) onto an object and view them through a second grating placed in front of the viewer; see Fig. 8.

These techniques produce fringes corresponding to contours of equal height on the object. The contour interval is determined by the fringe (grating line) spacing and the angle between the illumination and viewing directions. Phase shifting can be introduced by shifting one grating or the projected fringes. Moiré and fringe-projection interferometry complement conventional interferometry, as they can contour objects with an effective contour interval of 10 μm and larger.

3.2.3 Interference Microscopy Interference microscopy combines an interferometer and a microscope into one instrument and is used for measuring engineering surfaces that demand testing with high resolving power. Interference microscopy uses three types of interferometric objectives: Michelson, Mirau, and Linnik (all are modifications of the Michelson interferometer). In a Michelson interferometer objective, which is used for small magnifications, a beam-splitter cube and reference mirror are inserted between the objective and the tested surface. At magnifications above five times, the working distance, which refers to the space between the last surface of the lens and the tested surface, becomes too small to squeeze in a beam splitter cube; instead a Mirau interferometric objective (described in OPTICAL INSTRUMENTATION, Sec. 3.2.1) is employed. For magnifications of one hundred times and above, a Linnik interferometric objective (Linnik, 1933) is used since the working distance for such high magnification is even smaller and the Mirau design is impossible to implement. In a Linnik configuration a beam-splitter cube placed before the light reaches the objective directs the beam onto two objectives; one beam is directed to the reference mirror and the other is directed to the test surface. Any type of phase-across-the-plane analysis technique discussed in this article may be used with these microscopes. Other interference microscopes are reviewed by Steel (1983).

3.2.4 Fringes of Equal Chromatic Order Interferometry Fringes of equal chromatic order (FECO) interferometry (Fig. 9) is based on multiple-beam interference between two highly reflective and nearly parallel plates and results in very sharp fringes (Tolansky, 1970). Multiple-beam interference magnifies the extent of the deformation, thereby making it easier to detect small imperfections on the tested surface. FECO uses a white-light source and is combined with a spectrometer to test thickness variation and surface roughness, which are determined from the width of the fringes. Small sections of the test surface are observed through a spectrometer that displays fringes of equal chromatic order. A FECO interferometer can achieve a precision level of $\lambda/500$ and is capable of distinguishing hills from valleys by use of the direction of the bending of the colored fringes. However, the test surface must be overlaid with a highly reflective coating and the reference surface rms (root mean square) roughness should be better then $\lambda/20$.

4. PHASE AT ONE POINT

Phase-at-one-point interferogram detection is commonly used to measure length, the spectral content of a source (see Sec. 6), or its polarization (see Sec. 8.4). For length measurement these interferometric techniques

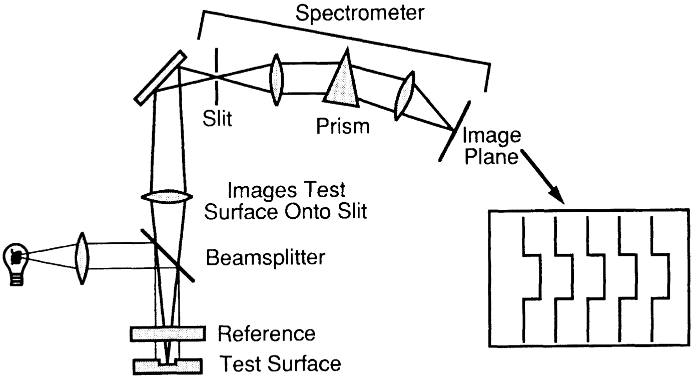

FIG. 9. Interferometer for measuring fringes of equal chromatic order (FECO).

(Hariharan, 1985; Steel, 1983) generally utilize a single detector to measure the phase at a single point of the interferogram where the fringe spacing is much larger than the size of the detector. The interferometric techniques that test the phase at one point measure length with very high precision, up to $1/10^4$ of a fringe, and are extensively used to measure relative and absolute distance. A few common single-point–phase detection techniques are described below. While measuring length is an important application of these techniques, the phase at a single point may represent other parameters, such as changes in pressure, temperature, or rotation.

4.1 Heterodyne Interferometry

Heterodyne interferometry observes interference between two beams with slightly different optical frequencies. This frequency difference is usually achieved by introducing a frequency shift into one of the beams. This frequency shift may be a result of a linear change in the phase difference and is achieved by rotating a radial grating, combining stationary and rotating quarter-wave plates, or moving a mirror in one of the arms of the interferometer. When the two beams interfere, a photodetector records a traveling interference signal at the beat frequency, which is equal to the difference in the two optical frequencies of the beams.

Heterodyne interferometry is often used to detect small changes in length. This detection is accomplished by evaluating changes in the beat frequency between the stable laser and the test (slave) laser frequency. Heterodyne interferometry measures changes in length by measuring changes in frequency.

4.2 Fringe Counting

These techniques essentially count the changes in intensity as they scan across a point detector while increasing the optical path difference between two beams. If the signal is lost for any reason, the measurement must be repeated. Two basic approaches to phase detection using fringe counting are often used. The first is based on observing the intensity changes in two interferograms. Into one interferogram an additional $\pi/2$ phase shift is introduced between interfering beams in order to determine the direction of the phase-difference changes (signals are in quadrature). The $\pi/2$ phase shift is induced by use of a special beam splitter or through polarization effects. The phase change can be detected to within a fraction of the fringe by observing the circular pattern formed by the two intensity signals mapped onto the horizontal and vertical axes on an oscilloscope. A full circle is traced each time one full fringe is scanned. The drawback to this method is that any changes in the intensity signal (such as source intensity variations) could be interpreted as a fringe.

The second approach avoids this problem; it is based on a heterodyne system that detects the interference of beams of different

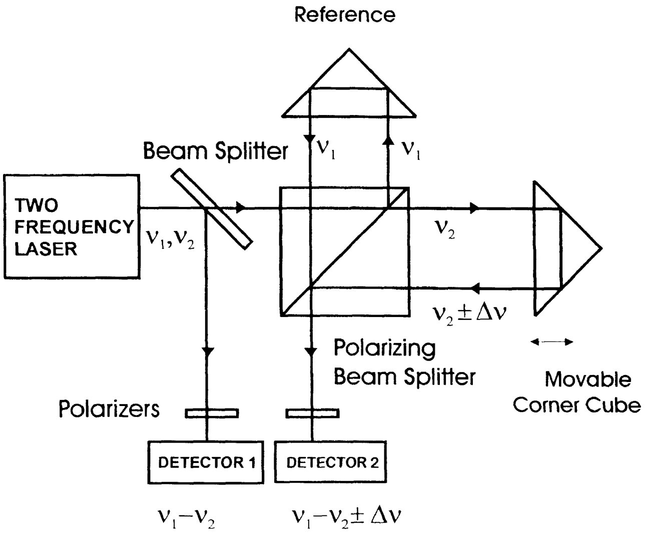

FIG. 10. Distance-measuring interferometer based on heterodyne system.

optical frequencies. Figure 10 shows a distance-measuring interferometer based on a heterodyne system that uses fringe counting. The differing frequencies can be obtained by using either a laser with two output frequencies or a single-frequency laser with two acousto-optic modulators. Either way the beams need to be orthogonally polarized. The first detector records the traveling interference reference signal at the beat frequency, which is equal to the difference in the two optical frequencies of the source. A polarizing beam splitter directs the beam of ν_1 into the reference arm and ν_2 into the test arm of the interferometer. Corner cubes return these beams to the second detector where the polarizer allows them to interfere. When the test corner cube is stationary, an intensity signal at the reference beat frequency is observed; when the test corner cube is in motion, it introduces a Doppler frequency shift in the returning beam and a different beat frequency is observed at the second detector. Signals from both detectors are forwarded to a differential counter and the change in the length of the test arm is determined.

4.3 Phase Locking

Phase-locking interferometry measures small changes in phase by locking the phase difference between the beams while detecting changes in the intensity signal (Moore, 1979). Locking the phase difference is accomplished by offsetting the phase of one of the beams and then subjecting that phase to sinusoidal modulation introduced by an oscillating mirror. This procedure induces intensity modulation that can be expressed as

$$I = I_1 + I_2 + 2(I_1 I_2)^{1/2} \cos(\varphi + \delta + a \sin \omega t). \tag{4}$$

The amplitude value a of the phase oscillation $\sin \omega t$ is much smaller than the source wavelength. The offset phase δ is set so that the average phase difference satisfies

$$\varphi + \delta = m\pi, \tag{5}$$

where m is an integer. Under this condition the intensity signal goes to zero and there is no intensity modulation due to mirror oscillation. When the average phase difference changes because of changes in tested phase φ, the intensity signal starts to modulate with its magnitude being proportional to the changes in the phase. This signal is then filtered and used to produce a feedback signal that tells the phase modulator how to alter the phase offset δ so that the condition $\varphi + \delta = m\pi$ is satisfied and the intensity modulation due to mirror oscillation is eliminated.

5. APPLICATIONS OF PHASE AT ONE POINT

5.1 Length Measurement

Phase-at-one-point detection in interferometry is very often applied to the precise measurement of length (Steel, 1983). In fact, interferometric techniques really established their value when they were applied to the measurement of material standards such as length, wavelength, line, and gauging.

5.2 Gravity-Wave Detection

Single-point phase detection has recently been employed to detect gravity waves. Much as accelerating charges produce electromagnetic waves, accelerating masses emit gravity waves. Gravity waves produced by collapsing supernovas, binary systems of neutron stars, and black holes affect the dimensions and spacing of all material objects. The movements that signal the existence of gravity waves can be detected by observing the small changes in phase at a single point in an interferometer. To register these changes requires an interferometer of very high sensitivity. A simple Michelson interferometer would need unrealistically long arms or high laser power to achieve the required accuracy. A number of new interferometers have been devised that should have sufficient sensitivity to detect gravity waves (Saulson, 1994). One of these has been coined the Laser Interferometric Gravitational-wave Observatory (LIGO). Its interferometric setup (Fig. 11) places identical Fabry–Pérot resonant optical cavities into each arm of a Michelson interferometer. The Fabry–Pérot arm cavities effectively increase the optical path length of each arm on account of the multiple reflection of light inside the cavities. The Fabry–Pérot mirrors act as inertial test masses suspended like pendulums from seismically isolated platforms. Passing gravity waves change the relative length of the resonant Fabry–Pérot cavities; this change alters the phase of the light and the phase change is picked up by the detector, thus revealing the existence of passing gravity waves. In order to detect the phase, the frequency of the laser is locked so that it is in resonance with the length of one cavity and then the length of the other cavity is adjusted to achieve the same resonance. The system detects the dark interference fringe, which allows most of the light to be reflected back to the source and back out again into the system with a recycling mirror. This additional mirror and the first surfaces of the test masses create the cavities where the light builds up, which effectively increases the optical path length and the illuminating power of the laser. The LIGO interferometer allows the detection of movements as slight as 1/1000 the diameter of a proton (10^{-18} m) and enables the identification of the presence of matter at distances of more than 70×10^6 light years from Earth. The distance separating the mirrors, more than 4 km, makes the LIGO interferometer the world's largest precision optical instrument and is pushing the limits of technology.

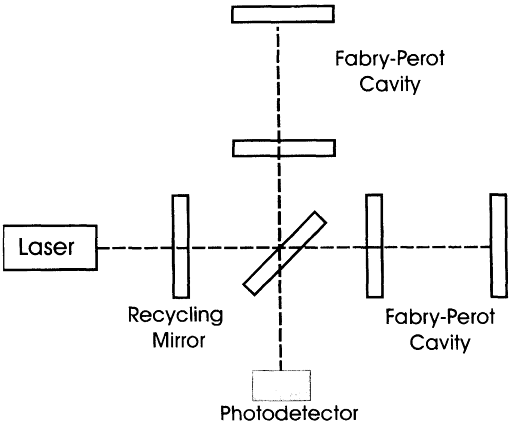

FIG. 11. Gravity-wave detection interferometer.

5.3 Other

Interferometry is used for monitoring processes such as plasma etching, film etching rate, uniformity of etching, and substrate temperature (see Plasma Etching). Phase-at-one-point detection can be also used for measuring refractive index and electron density of a plasma. Temperature, pressure, strain, and rotation can be detected by optical-fiber sensors (see Sec. 8.2). Phase detection has also been used to measure the speed of light (Steel, 1983).

6. INTERFERENCE SPECTROSCOPY

Interferometric spectrometers analyze the spectral distribution of a radiating source rather than examining the phase difference as the techniques discussed previously do. Generally, spectrometers measure not just the spectrum of radiation but rather the combination of the radiating source, the detector sensitivity, and the transmittance of the interferometer. In addition, a perfectly monochromatic spectral line will result in different spectral energy distributions when detected by different types of spectrometers. This characteristic of spectrometers is called the response of the system, and it is dependent on the spectrometer's instrumental function, which determines when two spectral lines of frequencies ν_1, ν_2 (or λ_1, λ_2 in terms of wavelength) can be still resolved by the instrument. The resolving power of the spectrometer is then determined by

$$\Re = \nu/\Delta\nu = \lambda/\Delta\lambda \tag{6}$$

An important characteristic of a spectrometer, especially when faint spectra are studied, is its "light-gathering power", which is defined by the invariant known as *étendue*. For example, dispersive spectrometers, like grating or prism spectrometers, generally use a narrow slit; these systems have smaller étendue than those spectrometers that do not use this type of slit. Two types of interferometric spectrometers and their techniques for recovering the spectral content of a radiating source are discussed below, a Fabry–Pérot etalon and a version of a Michelson interferometer.

6.1 Fabry–Pérot Spectrometer

The Fabry–Pérot interferometric spectrometer (Fabry and Pérot, 1899; Vaughan, 1989) employs multiple interference between two high-quality and highly reflective plates, which can be either fixed (etalon) or variably separated. Fixed plates yield a set of circular (spatial) sharp fringes for different wavelengths of the source. This setup is often used to measure the separation of two planes in etalons (standards), the thickness of a thin film, or the wavelength of a source (Hariharan, 1985).

The instrumental function of Fabry–Pérot plates separated by distance d is the *Airy function*

$$A(\nu) = T^2/[(1-R)^2 + 4R\sin^2(\psi/2)], \tag{7}$$

where R and T are the reflectance and transmittance of the plates and $\psi = (4\pi\nu nd/c)\cos\theta$, where n is index of refraction of the medium between the plates and θ is the angle of incidence. The Airy function has a set of narrow peaks that are separated by a frequency difference equal to $c/2dn$; this frequency difference describes the *free spectral range* where the fringes will not overlap. The half-width of the peaks $\Delta\nu$ is described by

$$\Delta\nu = c/2ndF, \tag{8}$$

where F is the *finesse* of the system,

$$F = \pi\sqrt{R}/(1-R). \tag{9}$$

The limited spectral range of the etalon can be overcome by separating the ranges in order to determine the frequency they represent without any ambiguity; however, this procedure is accomplished by imaging fringes onto a slit of the spectrometer, which then limits the étendue of the system.

The scanning Fabry–Pérot interferometric spectrometer has significantly higher étendue and is often used as a spectrometer for testing faint light sources. In this system the optical separation between the two plates changes linearly and the transmitted intensity signal (section of spectrum) is recorded by a point detector placed in the focal plane of the lens, which is placed behind the interferometer. Many modifications can be introduced to increase the performance of this spectrometer. For example, spherical mirrors may replace plates and thus increase light usage efficiency.

6.2 Fourier-Transform Spectroscopy

Interferometry may also be used to obtain the spectral energy distribution of a source through a technique called Fourier-transform spectroscopy (Michelson, 1902; Vanasse and Sakai, 1967; Bell, 1972). Fourier spectroscopy records the variation in intensity at a single point while varying the optical path difference between interfering wavefronts in a two-beam interferometer, which is typically a Michelson with the effective source at infinity. The optical path difference is realized by moving one of the mirrors, as in phase-measurement techniques. The detected intensity of the interference signal is the sum of independent interference signals for each spectral component and differs with the spectral content of the tested source. This intensity can be represented by

$$I(\tau) = \int G(\nu)[1 + \cos(\tau 2\pi\nu)]\,\mathrm{d}\nu \qquad (10)$$

where τ is the introduced optical path difference and $G(\nu)$ is the spectral distribution of the source. Since the recorded intensity signal is not the direct spectrum signal as in the Fabry–Pérot interferometer but rather the cosine Fourier transform of the spectrum, the Fourier transformation of the recorded signal must be calculated to achieve the spectrum of the source. The intensity of the interference signal must be sampled at least twice per fringe (as in phase-detection techniques). Because the fringe spacing in an interference signal corresponds to the wavelengths of the source, the minimum wavelength that can be detected is limited (as the slope is limited in phase-across-the-plane detection). Unequal sampling intervals and finite changes in optical path difference will introduce errors into the calculated spectrum. A process called *apodization*, where a measured interference signal is multiplied by a symmetrical signal with gradually decreasing values, like a triangle function, reduces the magnitude of errors in the calculated spectrum.

Fourier spectroscopy may be relatively difficult to implement; but when compared with prism or grating spectroscopy, Fourier spectroscopy has a very high étendue and better resolution than either of the other two methods. Another advantage of Fourier spectroscopy is its *multiplex advantage*: the signal-to-noise ratio is significantly improved as a result of simultaneous registration of all spectral components of radiation. Major applications of this technique include detection of emission spectra and absorption spectra of aqueous solutions, gases, and transient species. Because of its efficient use of light, Fourier spectroscopy is used often in astronomy to record spectra (usually infrared—see SPECTROMETERS, INFRARED) from very faint sources.

Fourier transformation is a powerful tool in interferometry; it can be used with either a coherent or an incoherent spectral source to analyze the phase across the plane. In addition, Fourier techniques are used for analyzing both the spatial and spectral distribution of a radiating source.

6.3 Spectral Imaging

The interferometric instruments described so far have been able to acquire either phase or spectral information about an object. However, many disciplines such as astronomy, robotic vision, or remote sensing are interested in obtaining both types of information at once. Recent progress in optics and electronics has allowed researchers to register multiple spectral images, thus spurring the development of new spectral imaging techniques (Itoh, 1996). Many of these techniques combine interferometric methods or couple interferometry with other optical testing procedures. For example, a double Fourier spatio-spectral interferometer merges a Michelson interferometer for Fourier spectrometry with a Michelson stellar interferometer for star diameter determination.

7. INTERFERENCE IMAGERY

Interference imagery (Steel, 1983) is an interferometric technique for testing the spatial distribution of a radiating source and is often used in astronomy to measure stellar diameters that are nearly impossible to resolve with conventional telescopes. In 1921 Michelson first developed a method for testing stellar diameters; his stellar interferometer could resolve distant objects with angular diameters on the order of 10^{-2} arc s (Michelson, 1921; see also INTERFEROMETERS AND INTERFERENCE, Sec. 1.3.2). Since that time, interferometric techniques have become more precise, using visible light as well as radio waves to measure spatial distributions. The resolution of stellar interferometers depends on both the distance between the mirrors (antennas) and the number of mirrors (antennas) used. Stellar interferometers are sometimes referred to as long-baseline or very-long-baseline (VLBI) interferometers or aperture-synthesis arrays.

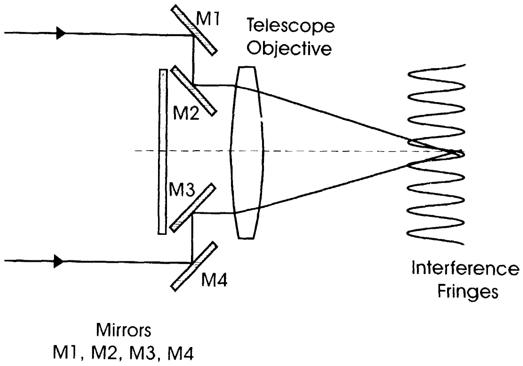

FIG. 12. Michelson stellar interferometer for measuring stellar dimensions.

Stellar interferometers today test many parameters in order to achieve maximum precision, including the phase across the plane for fringe tracking and heterodyne techniques for delay.

7.1 Michelson Stellar Interferometry

A Michelson stellar interferometer is a two-beam interferometer that uses wave-front division rather than amplitude division as the interferometers and techniques described thus far do. A Michelson stellar interferometer (Fig. 12) consists of two widely spaced mirrors (whose separation can be changed) pointed at the same stars. Light from the two mirrors is combined to form interference fringes on the image plane of a telescope. If the interferometer is pointing at two stars, two point sources, then two slightly shifted interference patterns of mean wavelength λ, one from each star, are formed. Because the source radiation from the stars is spatially incoherent, the sum of the interference patterns' intensities observed on the detector forms a new fringe pattern. The visibility of these new fringes depends on the shift between the interference patterns from each star. This shift is dependent on both the angular separation θ of the stars, which is fixed, and the separation d between the mirrors, which is referred to as the baseline and varies. The visibility of the fringes versus the mirror separation is plotted in Fig. 13(a). This visibility of the fringes is an envelope function of the Fourier transform of the spatial distribution of the source. Thus, if the mirror separation d is found for which the visibility of fringes equals zero, then the distance between the stars can be calculated.

When one is measuring the diameter of a star, the star is treated not as a point source but as a set of incoherent point sources that fills the star's diameter. The visibility of the fringes, Fig. 13(b), changes in a different fashion than for two point sources. By adjusting the separation of the two input mirrors until the fringe visibility reaches zero, which occurs when the spacing between the two mirrors is approximately $1.22\lambda\theta$, where θ is the stellar diameter, the diameter of the star can be determined but only in the direction of the mirror separation. Another procedure records the fringe pattern and then calculates the zero fringe visibility in order to determine the stellar diameter. In practice stellar interferometers are difficult to build, since the optical paths between the arms of the interferometer need to match as stars emit incoherent light. Stabilizing a system such as this requires very stable mechanical construction of the interferometer

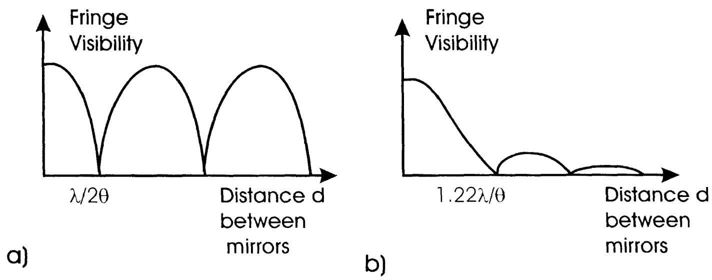

FIG. 13. Fringe visibility in a Michelson stellar interferometer for (a) two point sources and (b) a disk source.

and correction for atmospheric turbulence. In addition, if a star is not radially symmetrical, the mirrors need to be rotated or an additional set of mirrors implemented in order to determine the diameter of the star in a different direction. Finally, if radiation from the stellar source is not uniform, measurements that employ multiple separations of the mirrors need to be taken.

7.2 Stellar Intensity Interferometry

A major step in the development of stellar interferometers was the invention of the intensity interferometer by Hanbury Brown and Twiss (1954). The intensity interferometer is essentially a Michelson stellar interferometer that observes the intensity variation coming from each mirror separately; these intensity variations are then electronically correlated. The detectors in the intensity interferometer can be separated by thousands of kilometers, thereby significantly increasing the resolution of the system.

7.3 Radio Interferometry

Radio astronomy uses radio waves with wavelengths ranging from millimeters to hundreds of meters for observations in a different spectral range. Similarly, as stellar interferometry increases the resolution of visible light telescopes, radio interferometry, which is based on stellar interferometry (Steel, 1983), improves the resolution of radio telescopes (see RADIO TELESCOPES). Radio interferometry may use antennas for radiation collection instead of lenses and mirrors, but its principle of operation is still based on interference phenomena. Radio interferometry has many uses, ranging from listening for signs of earthquakes and volcanic activity to measuring the thickness of ice sheets and monitoring sea levels (see GEODESY).

7.4 Stellar Speckle Interferometry

Stellar diameters can be determined using a single telescope and a technique developed by Labeyrie (Labeyrie, 1976; Dainty, 1984). This technique allows a telescope to measure angular distances at its diffraction limit regardless of the light conditions. Individual speckles formed at the focus of the telescope have angular sizes determined by the diffraction limit of the telescope. When short exposures are made, the speckle pattern can be frozen in time. By taking the square modulus of the Fourier transform of a speckle pattern and averaging a number of these, a fringe pattern will be seen whose spacing is proportional to the stellar diameter or separation.

8. OTHER APPLICATIONS

In addition to the applications discussed so far in this article, there are many other applications that do not fit into the previous categories. These include the nondestructive testing of mechanical displacement and strain, the use of optical fibers in interferometers, the measurement of polarization and material properties, and interference between particles rather than waves. Since it is not possible to describe every application, this section highlights a few of these other applications.

8.1 Nondestructive Testing with Interferometry

Nondestructive evaluation of the mechanical properties of technical components is increasingly important for process control during manufacturing. Information about components' deformation due to an applied force or a change in temperature can be delivered by different interferometric techniques. As these techniques are applied to experimental mechanics, changes in phase can represent the displacement both in and out of the plane of the object.

8.1.1 Holographic Interferometry Holographic interferometry is an often-used technique because both amplitude and phase and in- and out-of-plane displacement information can be obtained about a surface as a known force is applied to it. These results can be compared with simulations generated by finite-element analysis to see whether mechanical parts behave as they are modeled. For example, holographic interferometry can provide information about whether a part in an airplane motor will perform as expected when it is stressed and loaded while in operation. The mathematical theory and basics of holographic (hologram) interferometry are outlined in OPTICAL HOLOGRAPHY, Sec. 4.3.

Recent developments in nondestructive testing using holographic interferometry have enabled holograms to be made in media that are erasable and reusable (Hariharan, 1994). Phase-measurement techniques can be applied to holographic interferometers to make quantitative analysis possible. Quantitative data enable a better comparison with simulation data for structural analysis. Applications include the testing of airplane tires, audio speakers, turbine blades, satellite fuel tanks, musical instruments, and infrastructure such as pipes and bridges.

8.1.2 Speckle Interferometry Speckle interferometry is similar in many ways to holographic testing; the difference is that information about object displacement is derived from the speckle patterns. Correlating speckle patterns before and after a force is applied yields information about the surface displacements. When speckle techniques were first developed, they relied on film as the recording medium. Sensor arrays have replaced film as the recording media, and advances in signal processing have allowed for fast quantitative analysis of object displacements (Jones and Wykes, 1983).

8.1.3 TV Holography By combining holographic and speckle interferometry, it is possible to record holograms on a CCD (charge-coupled device) sensor array and do image processing essentially to create digital holograms (Creath and Wyant, 1992; Kreis, 1996; Rastogi, 1997). Because the hologram must be resolvable with a TV camera, the frequency information content has to be within the Nyquist sampling limit of the sensor. Another consideration is that image processing must remove the self-interference terms from the hologram so that the cross-interference term is not corrupted. Many different approaches have been developed to perform this processing. Displacement measurements are easily obtained by computationally subtracting two holograms. Vibration analysis can be carried out stroboscopically or by adding a known vibration to the system and looking at the difference in frequency. Instruments using these techniques can be made to be quite insensitive to vibration. These instruments are often portable so that measurements are not limited to laboratory settings.

8.1.4 Grating Interferometry High-sensitivity grating interferometry (Patorski, 1993; Post *et al.*, 1994), has become an important method in experimental mechanics to measure, with submicrometer sensitivity, the in-plane displacements of nearly flat objects under load. In conventional high-sensitivity grating interferometry, the object under test with a reflection-type grating adhered to it is illuminated by two coherent beams symmetrical to the grating normal. The interference between the two diffraction orders reflected from the grating is recorded, resulting in a map of the in-plane displacements. Polarization techniques can be used in this system for phase shifting.

8.2 Optical-fiber Interferometry

Optical-fiber interferometers do not differ in principle from conventional two-beam interferometers; the difference is that in optical-fiber interferometers the beams travel through single-mode optical fibers, allowing for compact construction of an interferometer (Hariharan, 1985; also see SENSORS, OPTICAL, Sec. 2.8).

Optical-fiber interferometers are often used as optical sensors to detect changes in pressure or temperature, as these effects alter the optical path length of the fiber. The sensitivity of the interferometric system increases as longer fibers are used; thus optical-fiber interferometers find their advantage in combining compactness with sensitivity, since long optical paths can be contained in compact areas.

The optical-fiber interferometer can use a Sagnac-type (Vali and Shorthill, 1976) interferometric configuration to sense rotation (see INTERFEROMETERS AND INTERFEROMETRY). This is done by replacing the ring cavity of a conventional Sagnac interferometer with a multiturn loop made of single-mode optical fiber. The phase of the counterpropagating beams can then be detected using a phase-sensitive detector (Ezekiel, 1984).

8.3 Refractometry

Refractometers use interferometric techniques to measure changes in the refractive index of solids, liquids, or gases more precisely than any other method (see REFRACTOMETERS, OPTICAL). The first interferometer to measure a refractive index was devised by Michelson. Later Rayleigh and Jamin interferometers were developed specifically for this purpose (INTERFEROMETERS AND INTERFEROMETRY).

8.4 Ellipsometry

Ellipsometry (Azzam and Bashara, 1987; ELLIPSOMETERS) employs light polarization to characterize surfaces, interfaces, and thin films. In an ellipsometer linearly polarized incident light reflects from the test surface as elliptically polarized light. By superimposing two orthogonal components of an electric field, the elliptically polarized light can then be measured. Ellipsometry is a variation of common-path polarization interferometry where one of two copropagating orthogonally polarized waves acts as the reference for the other. Some ellipsometers use interferometric arrangements (mainly Michelson) to measure the index of refraction or film thickness over an entire surface, not just at a single point.

8.5 Atom Interferometry

Waves are characterized by phase and amplitude while particles are characterized by position and momentum. Both atoms and light can be treated as waves and particles, and both can be described by similar sets of equations. However, atoms have mass; therefore atomic waves obey a different dispersion relationship than does light. As the interference of wave fronts is an inherent feature of interferometry, atom interferometry (Adams *et al.*, 1994; Berman, 1996; see also INTERFERENCE AND INTERFEROMETRY) observes atom interference patterns. (Similarly, electron and neutron interference patterns are observed in electron and neutron interferometry.) Using methods similar to classical interferometry, atomic interference patterns are analyzed for modulation and fringe spacing and position. With much better precision than other interferometric methods, atom interferometry is able to detect and test gravitational and inertial effects, atomic properties, and some effects in quantum mechanics. Atom interferometry replaces the optical elements from conventional interferometers with microfabricated elements that diffract or split thermal atom beams, and tunable lasers that manipulate the trajectories of neutral atoms.

GLOSSARY

Fringe Order: The number assigned to a fringe in a single interferogram; it represents the direction of the changes in the recorded phase wavefront.

Optical Path: The length of the path that light takes through the medium multiplied by the index of refraction of the medium.

Optical Phase: Optical path of the light modified by phase changes on reflection of the light from the medium, where the phase changes depend on angle of incidence of the light and on the medium's complex index of refraction.

Phase Shift: The controlled change of the optical phase between interfering wavefronts.

Phase: Optical phase difference between test and reference wavefronts. The phase retrieved from an interferogram may represent the shape of the tested surface, its slope, or its displacement, among other things.

Works Cited

Adams, C. S., Sigel, M., Mlynek, J. (1994), *Phys. Rep.* **240**, 143–210.

Azzam, R. M. A., Bashara, N. M. (1987), *Ellipsometry and Polarized Light*, Amsterdam: Elsevier Science.

Bell, R. J. (1972), *Introductory Fourier Transform Spectroscopy*, New York: Academic Press.

Berman, P. R. (Ed.) (1996). *Atom Interferometry*, New York: Academic Press.

Bünnagel, R. (1956), *Z. Angew. Phys.* **8**, 342–350.

Creath, K., Wyant, J. C. (1992), in: D. Malacara (Ed.), *Optical Shop Testing*, New York: John Wiley and Sons, Chap. 15.

Creath, K. (1993), in: D. W. Robinson, G. T. Reid (Eds.), *Interferogram Analysis: Digital Fringe Pattern Measurement Technique*, Bristol: IOP Publishing, Chap. 4.

Dainty, J. C. (Ed.) (1984), *Laser Speckle and Related Phenomena*, Topics in Applied Physics Vol. 9, 2nd ed., Berlin: Springer.

Dew, G. D. (1966), *J. Sci. Instrum.* **43**, 409–415.

Ezekiel, S. (1984), in: S. F. Jackobs (Ed.), *Physics of Optical Ring Gyros: 7–10 January 1987, Snowbird, Utah*, SPIE Proceedings No. 487, Bellingham, WA: SPIE, pp. 13–20.

Fabry, C., Pérot, A. (1899), *Ann. Chim. Phys.* **7**, 115.

Fritz, B. S. (1984), *Opt. Eng.* **23**, 379–383.

Greivenkamp, J. E., Bruning, J. H. (1992), in: D. Malacara, (Ed.), *Optical Shop Testing*, New York: John Wiley and Sons, Chap. 14.

Hanbury Brown, R., Twiss, R. Q. (1954), *Philos. Mag.* **45**, 663.

Hariharan, P. (1985), *Optical Interferometry*, Sydney: Academic Press.

Hariharan, P. (1994), in: P. K. Rastogi (Ed.), *Holographic Interferometry*, Berlin: Springer-Verlag, Chap. 2.

Itoh, K. (1996), in: E. Wolf, (Ed.), *Progress in Optics*, Vol. 35, Amsterdam: North-Holland, pp. 147–196.

Jones, R., Wykes, C. (1983), *Holographic and Speckle Interferometry*, Cambridge: Cambridge University Press.

Kreis, T. (1996), *Holographic Interferometry: Principles and Methods*, Berlin: Academie Verlag.

Kujawinska, M. (1993), in: D. W. Robinson, G. T. Reid (Eds.), *Interferogram Analysis: Digital Fringe Pattern Measurement Technique*, Bristol: IOP Publishing, Chap. 5.

Labeyrie, A. E. (1976), in: E. Wolf (Ed.), *Progress in Optics*, Vol. **14**, Amsterdam: North-Holland, pp. 47–87.

Larkin, K. G. (1996). *J. Opt. Soc. Am. A* **13**, 832–843.

Linnik, W. (1933), *C. R. Acad. Sci. URSS* **1**, 18.

Malacara, D. (1992), in: D. Malacara (Ed.), *Optical Shop Testing*, New York: John Wiley and Sons, Chap. 5.

Mantravadi, M. V. (1992), in: D. Malacara (Ed.), *Optical Shop Testing*, New York: John Wiley and Sons, Chap. 4.

Michelson, A. A. (1902, 1961), *Light Waves and their Uses*, Chicago: Chicago University Press.

Michelson, A. A. (1921), *Astrophys. J.* **53**, 249.

Moore, D. T. (1979), *Appl. Opt.* **18**, 91.

Offner, A., Malacara, D. (1992), in: D. Malacara (Ed.), *Optical Shop Testing*, New York: John Wiley and Sons, Chap. 12.

Patorski, K. (1993), *Handbook of the Moiré Fringe Technique*, Amsterdam: Elsevier.

Post, D., Han D., Ifju, P. (1994), *High Sensitivity Moiré*, New York: Springer.

Ramirez, R. W. (1985), *The FFT: Fundamentals and Concepts*, Englewood Cliffs, NJ: Prentice-Hall, Inc.

Rastogi, P. K. (1997), *Optical Measurement Techniques and Applications*, Cambridge, MA: Artech House.

Rayleigh, L. (1874), *Philos. Mag.* **47**, 81–93 and 193–205.

Robinson, D. W. (1993), in: D. W. Robinson, G. T. Reid (Eds.), *Interferogram Analysis: Digital Fringe Pattern Measurement Technique*, Bristol: IOP Publishing, Chap. 6.

Rowe, S. H., Welford, W. T. (1967), *Nature* **216**, 786–787.

Saulson, P. R. (1994), *Fundamentals of Interferometric Gravitional Wave Detectors*, Singapore: World Scientific.

Schultz, G., Schwider, J. (1976), in: E. Wolf (Ed.), *Progress in Optics*, Vol. 13, Amsterdam: North-Holland, pp. 93–167.

Schwider, J. (1990), in: E. Wolf (Ed.), *Progress in Optics*, Vol. 29, Amsterdam: North-Holland, pp. 271–359.

Steel, W. H. (1983), *Interferometry*, 2nd ed., Cambridge: Cambridge University Press.

Tolansky, S. (1970), *Multiple Beam Interferometry of Surfaces and Thin Films*, Oxford: Oxford University Press.

Vali, V., Shorthill, R. W. (1976), *Appl. Opt.* **15**, 1099–1100.

Vanasse, G. A., Sakai, H. (1967), in: E. Wolf (Ed.), *Progress in Optics*, Vol. 6, Amsterdam: North-Holland, pp. 261–330.

Vaughan, J. M. (1989), *The Fabry–Pérot Interferometer*, Bristol: Adam Hilger.

Takeda, M. (1990), *Industrial Metrology* **1**, 79–99.

Womack, K. H. (1984), *Opt. Eng.* **23**, 391–395.

Yatagai, T. (1993), in: D. W. Robinson, G. T. Reid (Eds.), *Interferogram Analysis: Digital Fringe Pattern Measurement Technique*, Bristol: IOP Publishing, Chap. 3.

Further Reading

Bass, M. (Ed.) (1978), *Handbook of Optics*, New York: McGraw-Hill, Inc.

Born, M. Wolf. E. (1970), *Principles of Optics*, Oxford: Pergamon.

Françon, M., Mallick, S. (1971), *Polarization Interferometers*, New York: Wiley Interscience.

Gasvik, K. J. (1987), *Optical Metrology*, New York: Wiley.

Hariharan, P. (1985), *Optical Interferometry*, Sydney: Academic Press.

Hariharan, P., Malacara, D. (Eds.) (1995), *Selected Papers on Interference, Interferometry, and Interferometric Metrology*, SPIE Milestone Series No. MS 110, Bellingham, WA: SPIE.

Malacara, D. (Ed.) (1978, 1992), *Optical Shop Testing*, New York: John Wiley and Sons.

Steel, W. H. (1983), *Interferometry*, 2nd ed., Cambridge: Cambridge University Press.

Udd, E. (Ed.) (1991), *Fiber Optic Sensors: An Introduction for Engineers and Scientists*, New York: John Wiley and Sons.

Udd, E., Tatam, R. P. (Ed.) (1994), *Interferometry '94: Interferometric Fiber Sensing: 16–20 May 1994, Warsaw, Poland*, SPIE Proceedings No. 2341, Bellingham, WA: SPIE.

NEGATIVE HYDROGEN-ION SOURCES

J. H. WHEALTON, *Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A.*

Introduction 87
1. **Physics of Hydrogen Negative-ion Production** 88
1.1 Cesiated-Surface Production (Dudnikov, Belchenko, Dimov) 88
1.1.1 Cold-Cathode Penning (Dudnikov Source) 88
1.1.2 Semiplanotron (Belchenko Source) 88
1.1.3 Hot-Cathode Semiplanotron SITEX (Stirling, Soloshenko Source) 89
1.1.4 Planotron/Magnetron 91
1.1.5 Field-Free Surface Source 91
1.2 Charge-Exchange (Non-Liouville) Beam Production 93
1.2.1 Dudnikov Source 94
1.2.1.1 Cold-Cathode Penning (Dudnikov Source) 94
1.2.2 Semiplanotron and SITEX Variation 97
1.2.3 Elizarov Source 98
1.3 Excited-State Volume Production (Bacal) 99
1.3.1 Early Versions 100
1.3.2 Variations 100
1.3.3 rf Plasma Generation 104
1.3.4 Penning Source 107
1.3.5 Toroidal Discharge 107
1.3.6 Adding Cesium 107
2. **Physics of Hydrogen Negative-Ion Extraction** 107
2.1 Ion-Extraction Physics for Surface Sources 107
2.1.1 Low-Field Surface Sources 107
2.1.2 High-Field Surface Sources 109
2.1.3 Dudnikov Source 110
2.2 Ion-Extraction Physics for Volume Sources 111
2.2.1 Asymmetries Between Positive- and Negative-Ion Extraction 112
2.2.2 Saddle-Point Extraction 114
2.2.3 Double-Vlasov Model 119
2.2.4 Transverse Space-Charge Limits 120
2.2.5 Effect of Cesium 121
2.2.6 Increasing the Transverse Space-Charge Limit (TSCL) 122
3. **Applications** 123
3.1 Beam-Storage Rings—Pulsed Spallation Neutron Sources 123
3.2 High-Energy Neutral Beam for Fusion-Plasma Heating 123
3.3 High-energy Neutral Beams for the Strategic Defense Initiative 124
3.4 Ion-Projection Lithography 124
4. **Conclusions** 124
Acknowledgements 125
Glossary 125
Works Cited 126
Further Reading 130

INTRODUCTION

A number of low-energy collision processes can result in electron attachment to a hydrogen atom, producing a negatively charged ion. Being loosely bound, the electron can easily be removed, making a neutral atom once again. An important objective of negative-ion sources is how to attach this extra electron onto a hydrogen atom and keep it there throughout the extraction process. Early attempts to extract negative ions yielded currents in the microampere range. Two significant discoveries allowed the extracted current to increase by many orders of magnitude. These are covered in Secs. 1.1, 1.2, and 1.3.

The negative-ion source development, at the Budker Institute of Nuclear Physics, was motivated by the invention of charge-exchange injection into proton storage rings (Budker *et al.*, 1967). However, one inconvenience of the charge-exchange method is that it required negative-ion beams to start with. Ironically, one of the preferred embodiments for negative-ion sources also used a charge-exchange principle. Another negative-ion source develop-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ment at Ecole Polytechnique found an increase in beam current thousands of times more than expected through discovery of new atomic processes, brought about by a series of insightful experiments.

Originally needed for high-energy physics, and then for the Strategic Defense Initiative, application spread to high-energy neutral-beam heating and current drive for fusion plasma-confinement experiments. These activities led to a massive effort to develop and optimize these sources. Results of this development enabled applications, such as nuclear-physics materials studies, using the spallation neutron source and the next generation of semiconductor chip manufacture.

All these negative-ion sources involve the generation and maintenance of a plasma, consisting of positive ions, negative ions, and electrons, while optimizing the actual negative-ion generation and extraction. In each case there are several significant parameters which must be jointly optimized. We discuss many of the issues and some of the laboratory data and the theory necessary to guide the development of these negative-ion sources.

1. PHYSICS OF HYDROGEN NEGATIVE-ION PRODUCTION

In this section several types of hydrogen negative-ion production mechanisms that have found use in ion sources are discussed. Embodiments that illustrate the principles are presented along with some degree of comparison. Surface production from cesiated surfaces (Dudnikov and Derevyankin 1994; Belchenko *et al.*, 1973, 1997) is considered in some detail in Sec. 1.1. Volume sources are of two forms:

1. dissociative attachment of vibrationally excited hydrogen molecules, considered in Sec. 1.3, and
2. charge exchange, considered in Sec. 1.2.

1.1 Cesiated-Surface Production (Dudnikov, Belchenko, Dimov)

Positive ions hitting a molybdenum surface convert into negative ions with a significant probability (Greer and Seidl, 1984; Seidl *et al.*, 1980, 1982; Melnychuk *et al.*, 1991). As many as a third of the positive ions hitting a suitably cesium-coated surface will result in a negative ion being produced on or near this surface (Seidl *et al.*, 1984). Since positive ions are generally easier to create than negative ions, this surface process represents a significant advantage. Moreover, since there is a force at the surface due to the plasma sheath fields, the negative ions once generated are accelerated approximately normal to the surface, which means that the shape of the cesiated surface itself may have a significant influence on the ion optics. This allows for the possibility of focusing the negative-ion beam and in some cases amplifying the beam current. Many surface sources make use of this feature. If it were not for the fact that the negative-ion "temperature" or transverse energy can be fairly high (7 eV for a 150-V plasma-sheath potential drop) these sources would probably lead the other concepts of negative-ion sources, since the current density can be as high as 1 A cm^{-2} of H$^-$, and early results by Belchenko *et al.* (1994), of the Institute of Nuclear Physics (INP), yielded 11 A of total current.

1.1.1 Cold-Cathode Penning (Dudnikov Source) While this is apparently the first source using surface production from cesiated surfaces, the actual beam formation occurs as a result of charge exchange and is therefore considered in Sec. 1.2.2.

1.1.2 Semiplanotron (Belchenko Source) This form was developed by Belchenko, Dimov, and Dudnikov at the Institute of Nuclear Physics at Novosibirsk (Belchenko *et al.*, 1973), and two views of an embodiment (Belchenko and Kupriyanov, 1990) can be seen in Fig. 1.

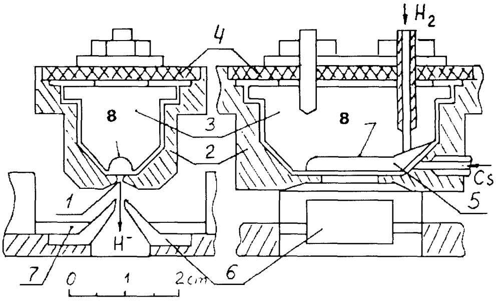

FIG. 1. Two views of semiplanotron surface source showing 1, emission slit (0.5 × 10 cm^2); 2, anode; 3, cathode; 4, insulator; 5, cathode groove; 6, extraction electrode; 7, mild steel inserts; and 8, converter.

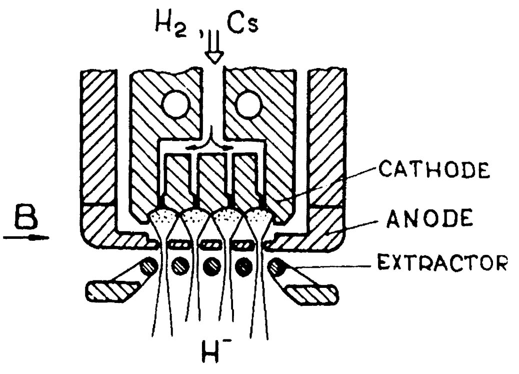

FIG. 2. Multibeamlet variation of semiplanotron.

A Penning discharge is established in region 5; this discharge consists of H^+ ions and electrons formed adjacent to a cesiated plate, 3, biased negatively by 100–300 V. If the cesium coating on the converter 8 is very thin (Seidl *et al.*, 1982), a significant amount of negative ions will result from double electron attachment to the incident positive ions or from electron attachment to sputtered hydrogen atoms. These H^- ions will accelerate through the 100–300-V sheath, go through the Penning discharge, 5, and be ready for extraction at 1. Since the H^- ions have about 100–300 V of energy, the inevitable plasma sheaths will have less of an effect than if their energy were smaller. This is one of the advantages of this mode of operation; another is that an amplification of beam current density occurs at 1 because of the concave shape of 8. However, the H^- ions, upon formation at the surface, acquire a transverse energy of a few electronvolts (Ehlers and Leung, 1980; York *et al.*, 1983), which results in a high beam emittance.

This configuration was used in multiple-beamlet configurations as illustrated in Fig. 2. 11 A of H^- were extracted from a 600-beamlet configuration with geometric focusing (Belchenko *et al.*, 1984, 1993). This multibeamlet configuration could be useful in certain applications, such as neutral injection into magnetic-confinement fusion experiments where high beam currents are needed without a concomitant increase in brightness.

1.1.3 Hot-Cathode Semiplanotron (Stirling, Soloshenko Source) This source was developed at ORNL (Dagenhart *et al.*, 1980, 1983, 1984; Stirling *et al.*, 1984), at Kiev (Soloshenko 1997; Goretsky 1996), and at BINP (Bashkeev *et al.*, 1990). It is illustrated in Figs. 3 (end view), 4 (top view) and 5 (oblique view). Referring to the end view, electrons are emitted from a filament and ionize hydrogen gas in a Penning discharge, which is confined in a strong magnetic field (see regions labeled "positive ion arc column" in Fig. 4, or the region labeled "discharge" in Fig. 3). The discharge starts at an anode aperture and passes

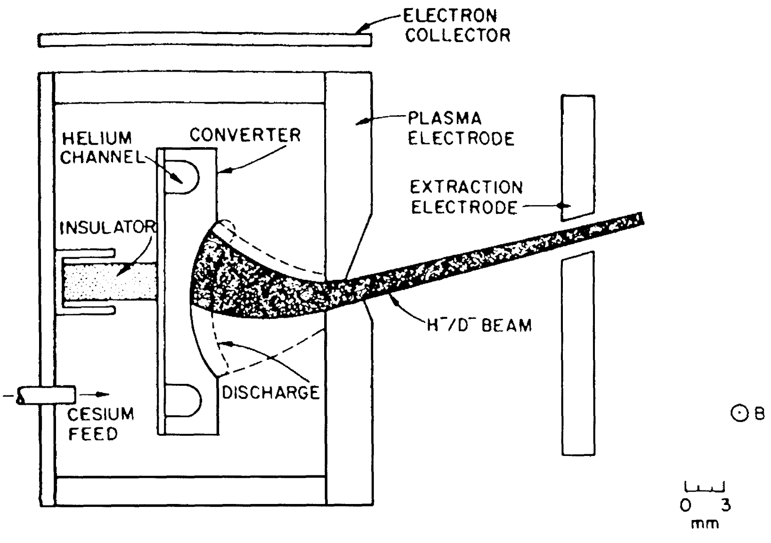

FIG. 3. Hot-cathode semiplanotron—SITEX end view.

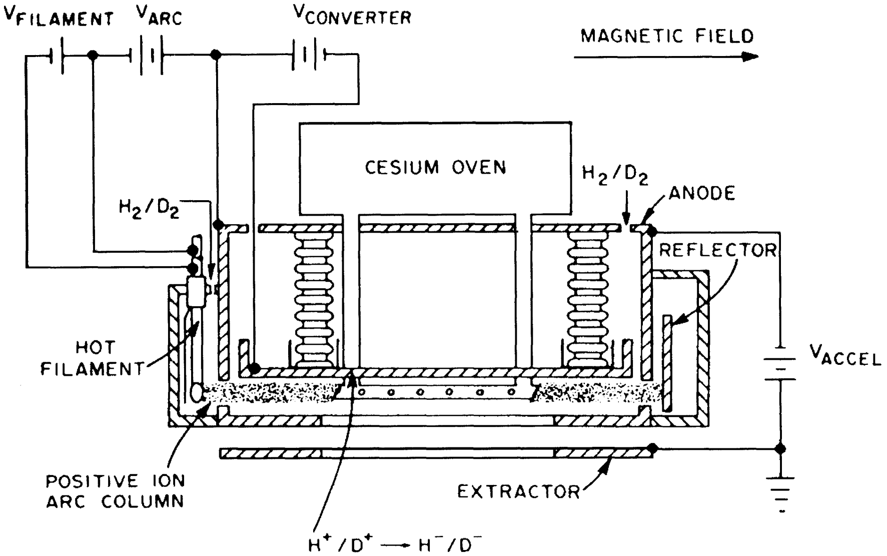
$V_{FILAMENT}$
V_{ARC}
$V_{CONVERTER}$
MAGNETIC FIELD
CESIUM OVEN
H_2/D_2
H_2/D_2
ANODE
REFLECTOR
V_{ACCEL}
HOT FILAMENT
POSITIVE ION ARC COLUMN
EXTRACTOR
$H^+/D^+ \rightarrow H^-/D^-$

FIG. 4. Hot-cathode semiplanotron—SITEX top view.

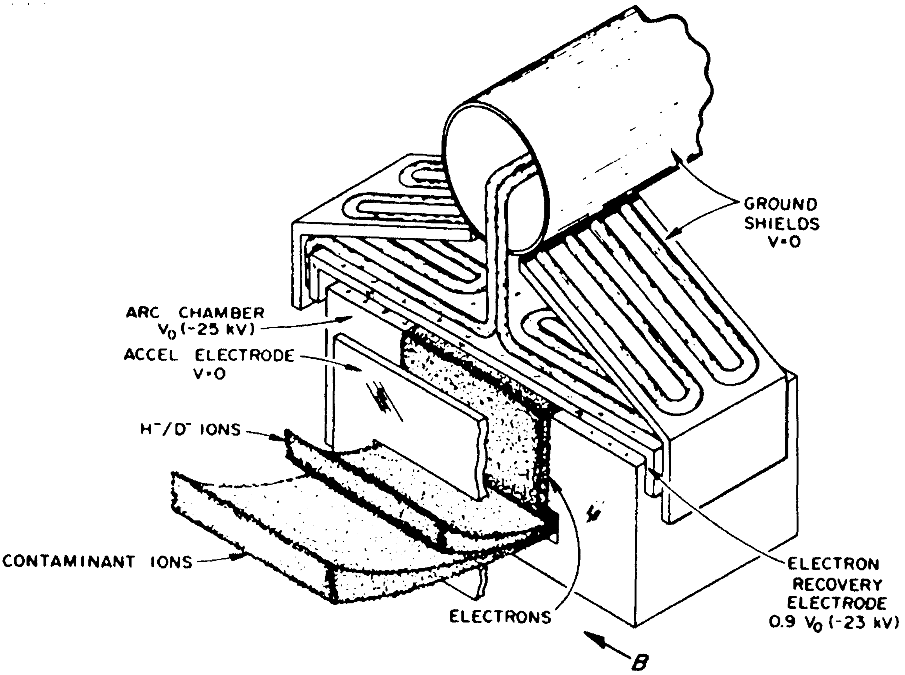
GROUND SHIELDS V=0
ARC CHAMBER V_0 (−25 kV)
ACCEL ELECTRODE V=0
H^-/D^- IONS
CONTAMINANT IONS
ELECTRONS
ELECTRON RECOVERY ELECTRODE 0.9 V_0 (−23 kV)
B

FIG. 5. SITEX—oblique view showing especially the dispatch of the electrons.

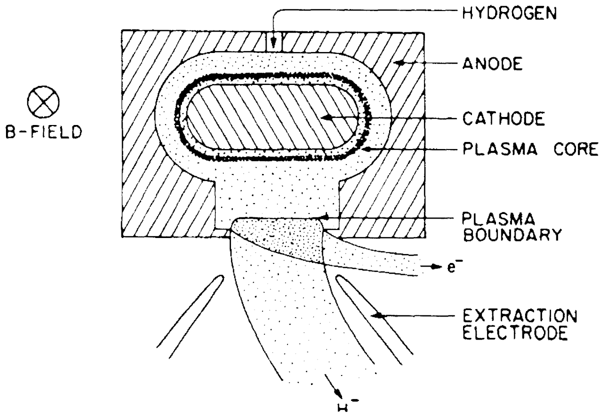

FIG. 6. Magnetron negative-ion surface source.

next to the converter surface shown most clearly in Fig. 3. This anode aperture is intended to reflect the electrons back into the main part of the column in front of the converter where they can continue to ionize the gaseous hydrogen. Since the plasma is so dense (electron density up to $10^{14}\,\mathrm{cm}^{-3}$), it requires a thick anode to provide field penetration into the column. This device has produced 650 mA of H^- for 10-s-long pulses (Dagenhart *et al.*, 1983). An oblique view is shown in Fig. 5, which shows how the electrons are captured at only a small loss. As in common with the Belchenko source, Sec. 1.1.2, positive H^+ from the plasma is attracted to the converter, which is biased relatively negative. The positive ions are converted into H^- at the surface, and these H^- are accelerated back through the sheath ($\approx$100–300 V) and continue on through the extraction aperture. An advantage of this source over that of the cold-cathode semiplanotron is in long-pulse, or cw, applications. A disadvantage is lower current densities.

1.1.4 Planotron/Magnetron A Planotron, whose principle of operation is depicted in Fig. 6 (Sluyters, 1979; Prelec, 1977a,b, 1980), differs from the semiplanotron in that the electrons in the primary discharge can go around the cathode in an $\boldsymbol{E} \times \boldsymbol{B}$ cycloidal drift determined by the applied magnetic field and the electric field that develops between the plasma and the cathode. The extra path made by the electrons, compared with the semiplanotron, in many cases provides less loss than the reflections made on a negatively biased anode Penning configuration (shown in Fig. 4). Other than providing an $\boldsymbol{E} \times \boldsymbol{B}$ recirculating electron stream, the source operates similar to the semiplanotron. An improvement made by Alessi and Sluyters (1980) is to increase the space in the back of the magnetron channel. This is depicted in Fig. 7 and has resulted in a 50% increase of H^- ion current to 600 mA. The increased negative-ion extracted current is depicted in Fig. 8. Also shown is the amplification (×5) of the beam current effected by the focusing due to the groove. This source was also developed at the Fermi National Laboratory (Schmidt *et al.*, 1977, 1979, 1980).

The magnetron geometry is also scalable to multiple apertures for high current (250 mA for several hours) (Prelec 1984). An example is shown in Fig. 9.

1.1.5 Field-Free Surface Source Developed at Lawrence Berkeley Laboratory (LBL) (Ehlers and Leung, 1980; Leung and Ehlers, 1980), Los Alamos Scientific Laboratory (LASL) (York *et al.*, 1983; York and Stevens, 1984), and the National Laboratory for High Energy Physics, Japan (KEK) (Takagi *et al.*, 1984; Mori 1987), the concept is shown in Fig. 10. Ions are generated on a spherical cesium-coated converter surface as shown and, as a result of the presence of a plasma, immediately accelerate down the 100–300-V sheath normal to the surface. These ions "self-extract" by moving through the intervening plasma at high speed to an awaiting accelerator, shown at the right of the figure. The electrons are contained in the plasma by cusp-field confinement pro-

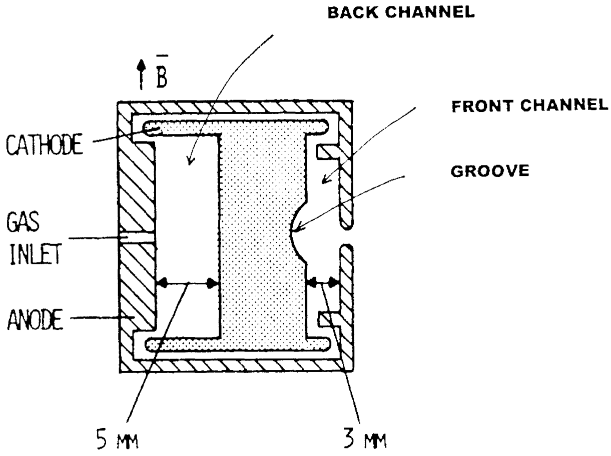

FIG. 7. The asymmetric magnetron configuration.

vided by many small magnets on the ion source chamber. 20 mA of H$^-$ were extracted at 10% duty cycle (York and Stevens, 1984; Takagi *et al.*, 1984); 400 mA of H$^-$ were extracted from the LBL source (Ehlers and Leung, 1980). A variation (Kwan *et al.*, 1993) produced 35 mA of D$^-$ current.

This source is called field-free because the magnetic fields in the region of negative-ion production are small. An example of the field penetration (Lietzke *et al.*, 1984) is shown in Fig. 11. The cusp field enables confinement of plasma electrons because electrons approaching them start spiraling towards the cusp and reach a point where their longitudinal energy gets totally converted into transverse energy, and so they are reflected. Without such cusp fields the wall of the source would have to be biased highly negative ($3kT_e$), repelling the electrons electrostatically. The trouble with this arrangement is that all the positive ions are attracted to the wall and are lost (albeit at a smaller rate than the electrons would be lost if the walls were not biased). However, when the electrons are trapped by the cusp field the wall can be biased positively, or at least less negatively, and thus electrostatically confine the positive ions as well (Stirling *et al.*, 1979) for volume sources discussed in the next section.

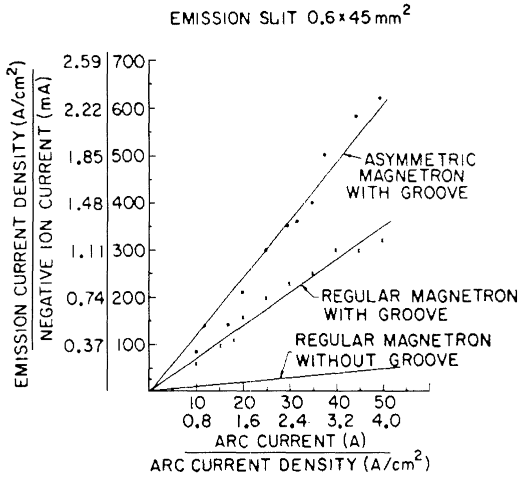

FIG. 8. H$^-$ current versus arc current for the normal magnetron with flat cathode, normal magnetron with grooved cathode, and asymmetric magnetron with grooved cathode.

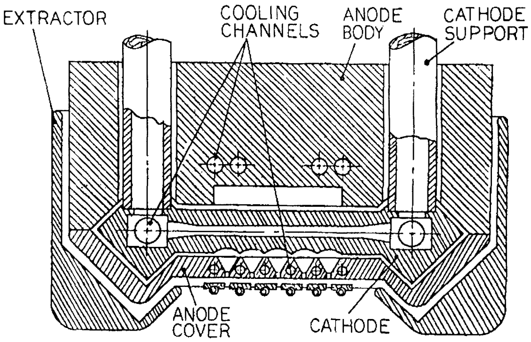

FIG. 9. Cross section of a steady-state multi-aperture magnetron source.

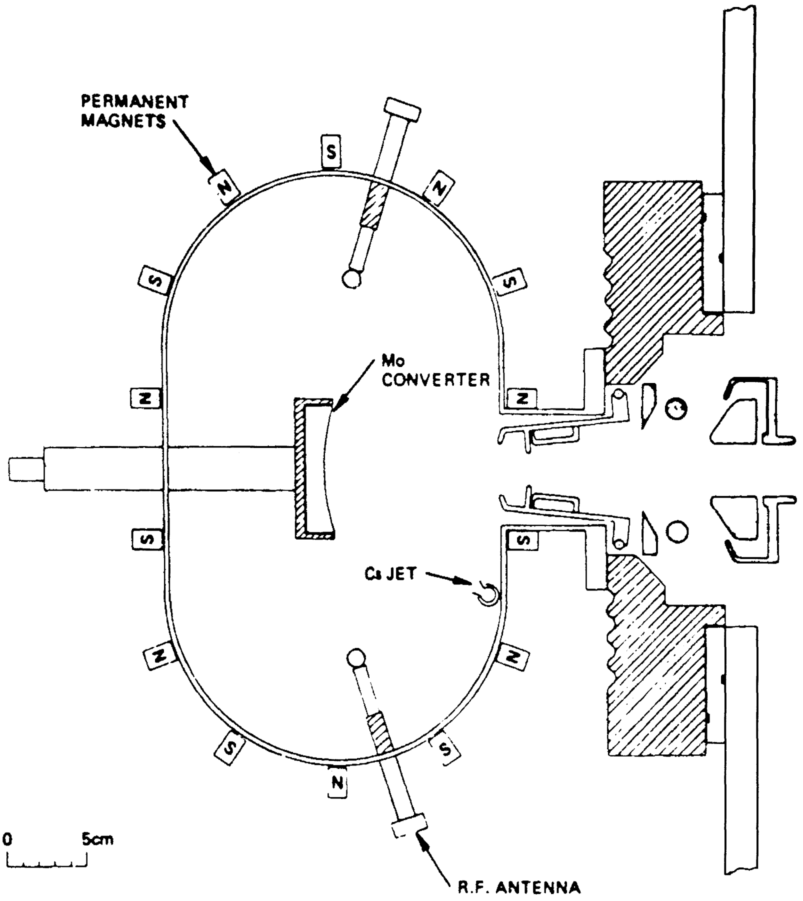

FIG. 10. Schematic diagram of the LBL self-extraction negative-ion source.

1.2 Charge-Exchange (Non-Liouville) Beam Production

One of the earliest negative-ion sources consisted of generating first a positive-ion beam and accelerating it by at least a few kilovolts and then, by running it through a gas cell, causing the positive-ion beam to undergo double charge exchange, turning it into a negative-ion beam (Semashko *et al.*, 1987). This process is inefficient compared with today's ion sources and resulted in a rather poor-quality beam. What we discuss here are the modern varieties, starting with the Penning source of Dudnikov, in which the negative ions are produced directly on a surface within the source and then in turn undergo charge exchange before extraction. This effectively lowers the ion temperature by an order of magnitude from that of the originating surface-produced negative ions.

Unfortunately, an unlucky accident occurred, which inhibited development of the low-energy, charge-exchange negative ion source—especially in the West. An inadvertent error in what is perhaps the most widely used data compilation (Barnett *et al.*, 1977) mistakenly tabulated a charge-exchange cross-section for the process

$$H^0 + H^- \rightarrow H^- + H^0,$$

that was too low by as much as a factor of eight at energies around 100 eV (Fig. 12). This was corrected in the next edition of the work (Barnett *et al.*, 1990). Anyone using these

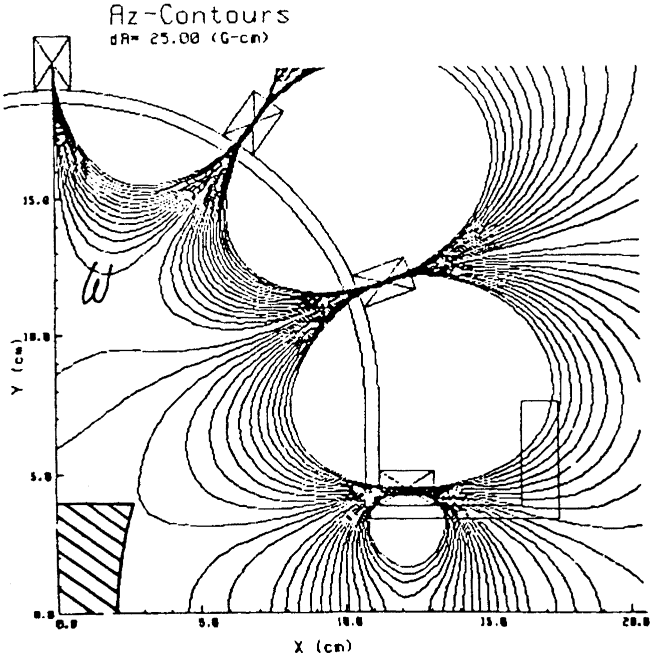

FIG. 11. Calculated magnetic field structure in one quadrant of the plasma chamber showing the filament filter and exit filter in relation to the filaments and the converter.

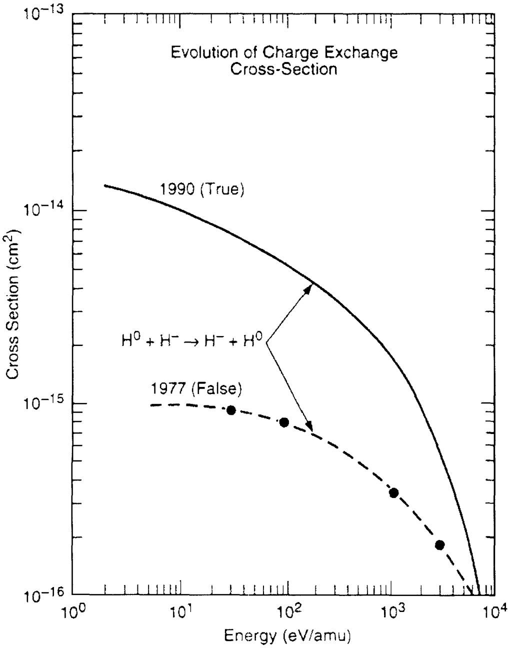

FIG. 12. Charge-exchange cross section, as published in 1977 and 1990, showing a large discrepancy especially at low energy.

erroneous data would underestimate the charge-exchange probability at these energies, resulting in the possible rejection of viable concepts. An illustration of the confusion that can arise is recorded in the comments published after the paper of Alessi and Sluyters (1980). Indeed, the discovery of the low-energy charge-exchange source was made on the other side of the world (Siberia), where the Western database might have received less attention.

1.2.1 Dudnikov Source

1.2.1.1 Cold-Cathode Penning (Dudnikov Source) The earliest negative-ion source using cesium, this configuration (see Figs. 13 and 14) is capable of prodigious output on account of a very interesting combination of brilliant inventions (Dudnikov, 1972; Belchenko et al., 1973; Dudnikov, 1980; Derevyankin and Dudnikov, 1984). This embodiment starts with a cesiated surface converter labeled as shown in Fig. 14 for initial production, but there the similarity ends; a significant variation allows a large non-Liouvillian increase in beam brightness. The negative ions once created undergo charge exchange (see Fig. 15) and may be extracted (sideways). The advantage of this

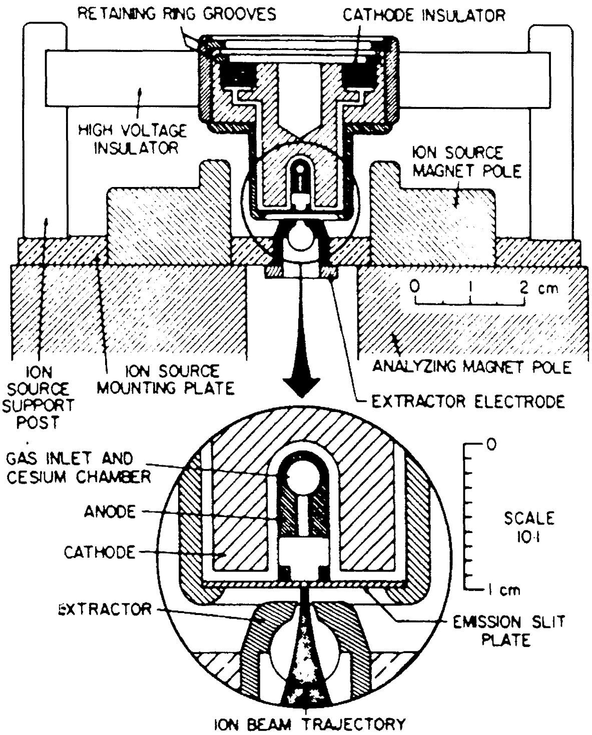
RETAINING RING GROOVES
CATHODE INSULATOR
HIGH VOLTAGE INSULATOR
ION SOURCE MAGNET POLE
0 1 2 cm
ION SOURCE SUPPORT POST
ION SOURCE MOUNTING PLATE
ANALYZING MAGNET POLE
EXTRACTOR ELECTRODE
GAS INLET AND CESIUM CHAMBER
ANODE
CATHODE
EXTRACTOR
0
SCALE 10:1
1 cm
EMISSION SLIT PLATE
ION BEAM TRAJECTORY

FIG. 13. Cross section of LASL version of Dudnikov source.

source over the direct surface extraction described in the previous section is that the charge-exchanged ions have a "temperature" that is likely to be an order of magnitude lower than that of the surface-produced ions.

An embodiment of one of the early versions of this source is illustrated in Fig. 13 showing the placement of magnets, hydrogen feed, cesium chamber, anode, and cathode. The region bounded by the anode and cathode is depicted in Fig. 14 showing the anode, cathode, converter, direction of magnetic field, and special anode chamber between the central discharge region and the accelerator. Some of the ion

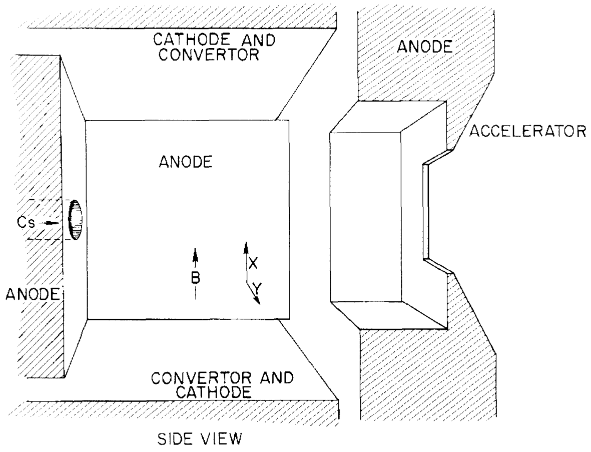
CATHODE AND CONVERTOR
ANODE
ACCELERATOR
ANODE
Cs
ANODE
B
X
Y
CONVERTOR AND CATHODE
SIDE VIEW

FIG. 14. Blowup of plasma-formation area in Dudnikov source.

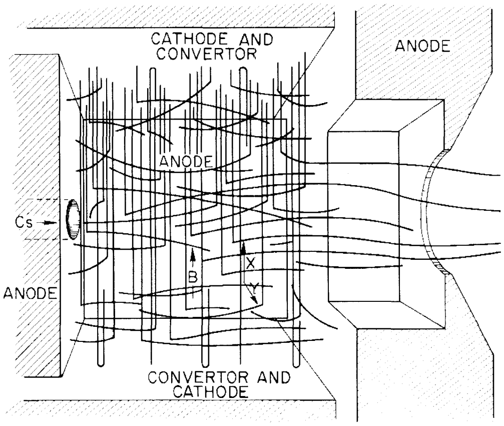

FIG. 15. Ion motion due to surface-produced ions and charge-exchanged ions within the Dudnikov source.

motion ensuing in this device is depicted in Fig. 15. Within this region, first positive ions (mostly H^+) and electrons are generated from a striker arc forming a dense, almost neutralized plasma (up to $10^{14}\,cm^{-3}$) with a plasma potential positive with respect to the cathode. In the absence of a magnetic field, the plasma potential would also be positive compared with the anode (which in turn is biased positive with respect to the cathode). However, the strong magnetic field inhibits the flux of electrons to the anode by means of the Lorentz force ($\boldsymbol{F} = q\boldsymbol{v} \times \boldsymbol{B}$), which is perpendicular to the direction of motion. The positive plasma ions are attracted to the cathode, and they intercept at an energy on the order of a hundred volts. Then, through surface production processes described above, a negative ion may be emitted [as many as one for every three incident positive ions (Seidl *et al.*, 1982)] and accelerate back down the plasma sheath, acquiring an energy of about a hundred electron volts. Because of the magnetic field the ions move largely perpendicular to the cathode (in helices). The hole in the anode shown for acceleration of these ions (*e.g.*, right side of Fig. 14) is not generally reachable by these negative ions, since they are largely confined along magnetic field lines to the region between the cathodes. These surface-produced ions would not be especially desirable for beam formation, since they have several electronvolts of random transverse energy upon leaving the converter surface (cathode) (Seidl, 1994). What is done instead is to let the surface-generated H^- ions undergo charge exchange with the gas—largely atomic H^0 under these conditions:

$$H^- + H^0 \rightarrow H^0 + H^-.$$

Some of these charge-exchanged ions do get out into the accelerator region via the anode hole depicted in Fig. 15. The virtue of this process is that the charge-exchanged ions are expected to have a much lower ion temperature than the surface-produced ones, since they originated from the much slower moving neutral atomic hydrogen gas.

Another aspect to this revolutionary source is the anode web, now known as a collar, shown in Figs. 14 and 16. Some of the effects of this web are still being explored, but a preliminary study by Allison *et al.* (1980) showed that there was an optimum web thickness whereby the extracted beam could be increased by over a factor of two (see Fig. 16). Additional meaning will be discussed in Sec. 2.2. At LANL a large effort was undertaken to vali-

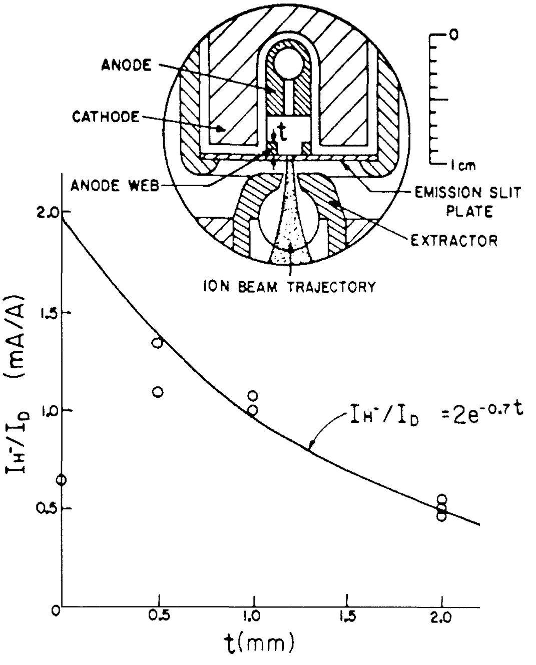

FIG. 16. H^- production efficiency (I_{H^-}/I_D) versus anode web thickness t. A curve with exponential dependence on t has been fitted to the data.

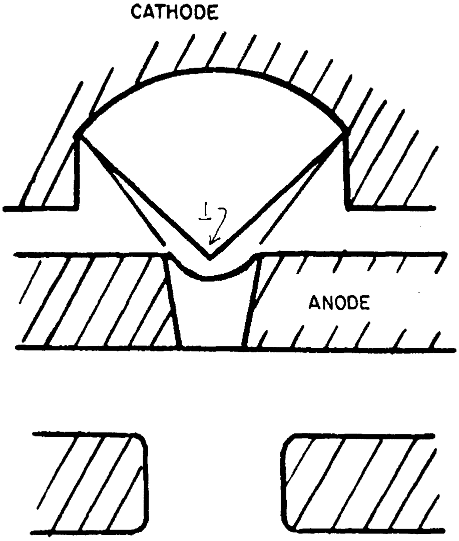
CATHODE
ANODE

FIG. 17. The rotating ion source.

date, scale up, and deploy this Dudnikov source (Allison, 1977; Allison *et al.*, 1980) for various purposes; the 4X (Smith *et al.*, 1984, 1985), 8X (Smith *et al.*, 1993), and 16X sources and the ORNL-produced "VITEX" (Dagenhart *et al.*, 1987) are examples of such scaleups.

Additional attempts to get to a higher duty factor operation include the rotating Dudnikov source (Smith *et al.*, 1980) (Fig. 17); 160 mA of H^- were reported in the 1X source.

1.2.2 Semiplanotron and SITEX Variation

By focusing the ions ahead of the extraction aperture (Dudnikov, 1980; Fig. 18), the charge-exchange reaction can be made to dominate. This was possibly achieved inadvertently in some SITEX experiments (Dagenhart *et al.*, 1984; Stirling *et al.*, 1984), because the beam-profile measurements correspond to a significantly lower ion temperature at the source as deduced from Liouville's theorem—although the charge-exchange mode

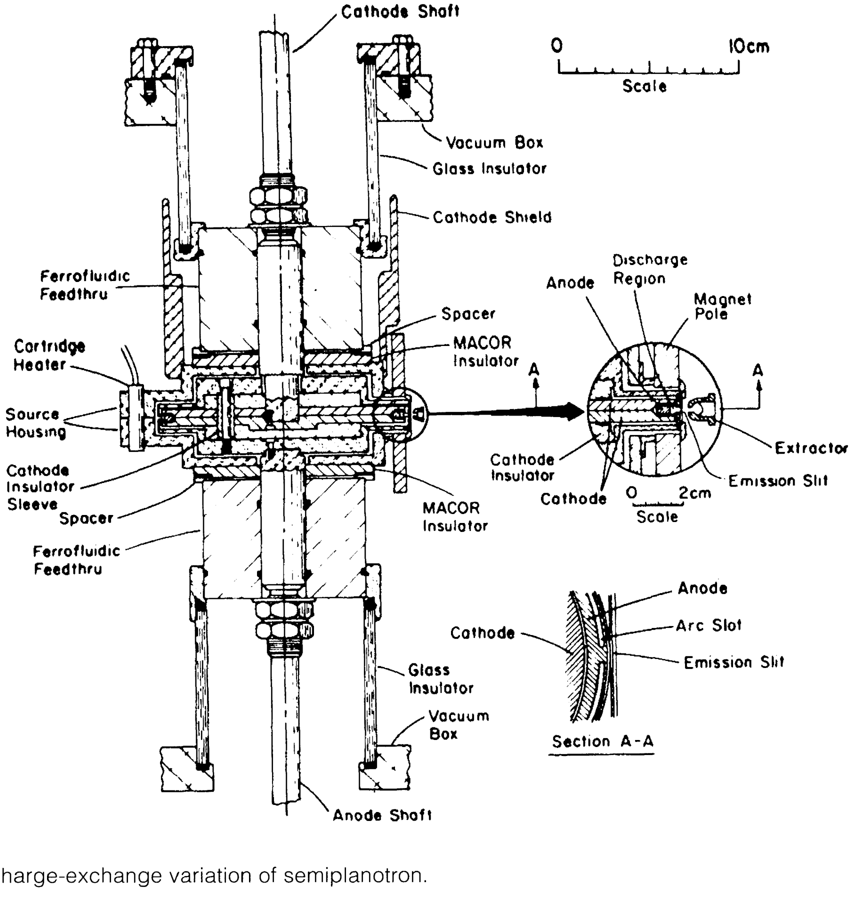
Cathode Shaft
0
10cm
Scale
Vacuum Box
Glass Insulator
Cathode Shield
Ferrofluidic Feedthru
Cartridge Heater
Source Housing
Cathode Insulator Sleeve
Spacer
Ferrofluidic Feedthru
Spacer
MACOR Insulator
MACOR Insulator
A
A
Anode
Discharge Region
Magnet Pole
Extractor
Emission Slit
Cathode Insulator
Cathode
0
2cm
Scale
Anode
Arc Slot
Emission Slit
Cathode
Section A-A
Glass Insulator
Vacuum Box
Anode Shaft

FIG. 18. Charge-exchange variation of semiplanotron.

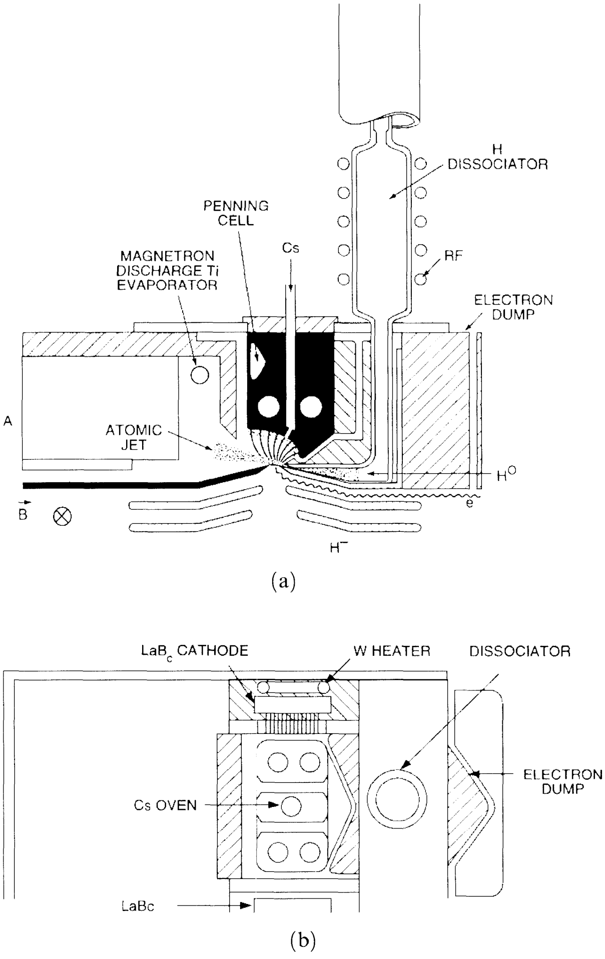

FIG. 19. Optimized version of SPS of surface plasma source. (a) Elevation view; (b) end view; (c) front view; (d) blowup of the charge-exchange region showing the configuration of the surface-generated negative-ion beam, the supersonic gas jet, and the charge-exchange region. Other features shown are the pumping, in (a), the gas dissociator for the atomic jet, in (a), the Penning discharge, in (a–c) and the cesium feed, in (a).

was not anticipated by the authors. An optimized configuration is shown in Fig. 19 (Dudnikov *et al.*, 1997). The benefit of this variation is that the advantage of high–current-density surface production can be combined with the current amplification due to geometrical focusing to produce high-density, low-temperature, charge-exchanged volume production. The focused surface-produced ions are focused to a region just before the extraction aperture where they undergo charge exchange with a gas (a supersonic jet may enhance the charge-exchange reaction rate). This is shown in Fig. 19. This source has the potential to deliver an H^- beam with a virtual source size of 10 μm, an emission current density of $3\,\mathrm{A\,cm^{-2}}$, an ion temperature of $\frac{1}{2}$eV, and a normalized brightness of $10^{13}\,\mathrm{A\,mrad^{-2}}$, which is very high (Dudnikov, 1993; Gudharay *et al.*, 1994).

1.2.3 Elizarov Source The Elizarov ion source (Antipov *et al.*, 1990) is taken from the positive-ion sources used in the Russian ion-thruster rocket-engine space program. A current of 1 A of H^- was extracted (Elizarov, 1972). The source (Fig. 20) consists of a hollow cathode (1) with an abundance of Cs. A cesium plasma is formed in region 9. Holes drilled in the outside of this region's cylindrical hollow cathode (2) allow plasma to escape initially into region 3, in which a radial magnetic field $\boldsymbol{B}$ is present. A series of Penning-discharge spokes is obtained. The H^- ions

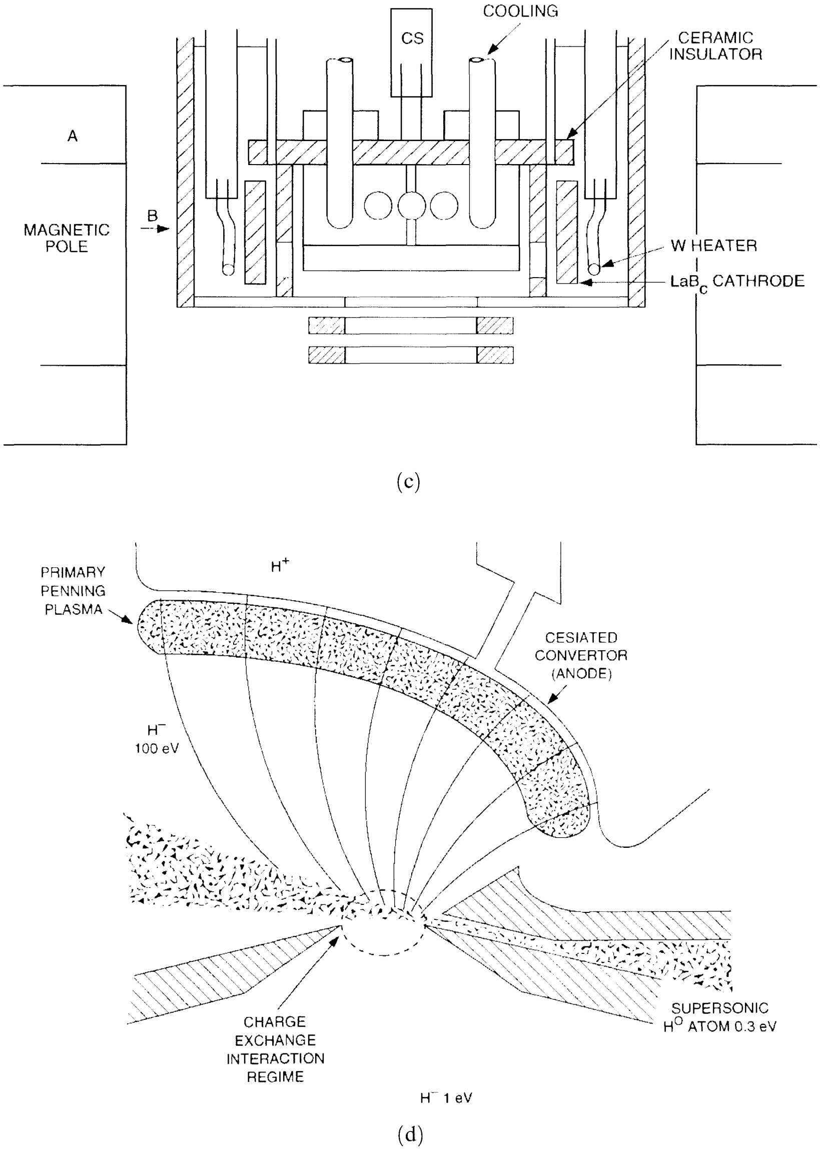
COOLING
CS
CERAMIC INSULATOR
A
MAGNETIC POLE
B
W HEATER
LaB$_C$ CATHRODE
PRIMARY PENNING PLASMA
H$^+$
CESIATED CONVERTOR (ANODE)
H$^-$ 100 eV
CHARGE EXCHANGE INTERACTION REGIME
SUPERSONIC H^0 ATOM 0.3 eV
H$^-$ 1 eV

(c)

(d)

FIG. 19. (continued)

drift through charge exchange toward the extraction electrodes toward region 8, with anode 1 in combination with the externally applied magnetic field providing an intraplasma field. A source built as in Fig. 20 produced 1000 mA. The Elizarov source may also be configured to produce enhanced current density in a non-Liouvillian manner. This is achieved by forming a conical channel with resistive walls or separate wall electrodes to produce the intraplasma electric fields (in combination with the transverse magnetic field). An embodiment is shown in Fig. 21.

1.3 Excited-State Volume Production (Bacal)

Researchers at Ecole Polytechnique (Bacal *et al.*, 1977) invented a pure hydrogen hybrid multicusp volume negative-ion source. It con-

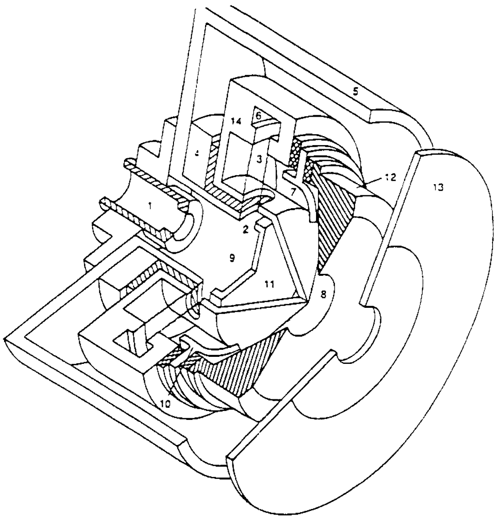

FIG. 20. Compression variation of Elizarov ion source. Shown is the cesium feed, 1; the cesium-hydrogen plasma, 9; the extraction, 2, of the cesium hydrogen plasma into the weak radial Penning discharge, 3, generated by cathode, 6; magnetic pole pieces 5, 11, 12 abide 7; funnelling electrodes 11, 12; beam extraction 8; and accelerator electrode 13.

sisted of two distinct plasmas. One, called a driver plasma (in most configurations confined by cusp field magnets; see Sec. 1.1.4) consists of a conventional discharge where hot electrons are created, causing ionization of the gaseous hydrogen and generation of vibrationally excited hydrogen molecules:

$$H_2 + e^- \rightarrow H_2^* + e^-. \qquad (1)$$

These excited hydrogen molecules drift into a second plasma region, called the target plasma, where the electron energy is arranged to be much lower (≈ 1 eV). Then the reaction

$$e^- + H_2^* \rightarrow H^- + H^0 \qquad (2)$$

has a high probability.

For the 7th vibrational level of H_2^*, for example, the reaction is 10^5 times greater than the corresponding reaction for ground-state H_2. Absence of high-energy electrons in the target plasma region reduces the rate of H^- destruction processes. An illustration of this separation in regions for energetic electrons (>3 eV) and cold electrons (≈ 1 eV) is shown in Fig. 22. The isolation of the target plasma, where cold electrons dominate, from the driver plasma, with its dominantly energetic electrons, occurs because of the presence of filter fields. Electrons going from the driver plasma to the target plasma need to undergo hundreds of collisions (some of which are inelastic) with neutral H_2, which reduces their energy so that when (and if) they finally arrive at the target plasma they are cooled. The filter fields provide a path length for the electrons that is several orders of magnitude higher than what would obtain without it, and thus increase the probability of energy-losing collisions.

1.3.1 Early Versions Two early embodiments are shown in Figs. 23 and 24 (Bacal *et al.*, 1984, 1986, 1987, 1988). In the first concept, the hot electrons are confined to the larger radii by direct insertion of multiple magnets between the filament region (driver plasma) and the target plasma. The extraction hole is shown in the bottom. In this configuration there is an opportunity for geometrical amplification of excited molecules into this excited region. The magnetic filter between the driver plasma and the target plasma is furnished by the external cusp fields, with the filaments shown being in the fringe of these cusp fields. In this case the target plasma is actually divided in turn by the fringe fields of the confining cusp fields, which serve to decrease the electron temperature in the post-extractor region, shown explicitly in Fig. 24 (Bruneteau 1986; Bacal *et al.*, 1988) and is described in the extraction section (Sec. 2). In any event, the fields further retard transport of high-energy electrons, which would otherwise quickly destroy the newly formed negative ions. The main driving force for this flurry of activity in volume sources was the requirements of the neutral particle-beam (NPB) "Star Wars" program of the 1980s where low beam emittance was critical.

1.3.2 Variations More recent embodiments of the volume source use cusp-field confinement of the driver plasma and are typified by the devices produced in the UK (Holmes and Green, 1984, 1987; McAdams *et al.*, 1984, 1987), France (Bacal and Leung, 1984), Japan (Okumura *et al.*, 1987), and the United States (Leung and Ehlers, 1984). In

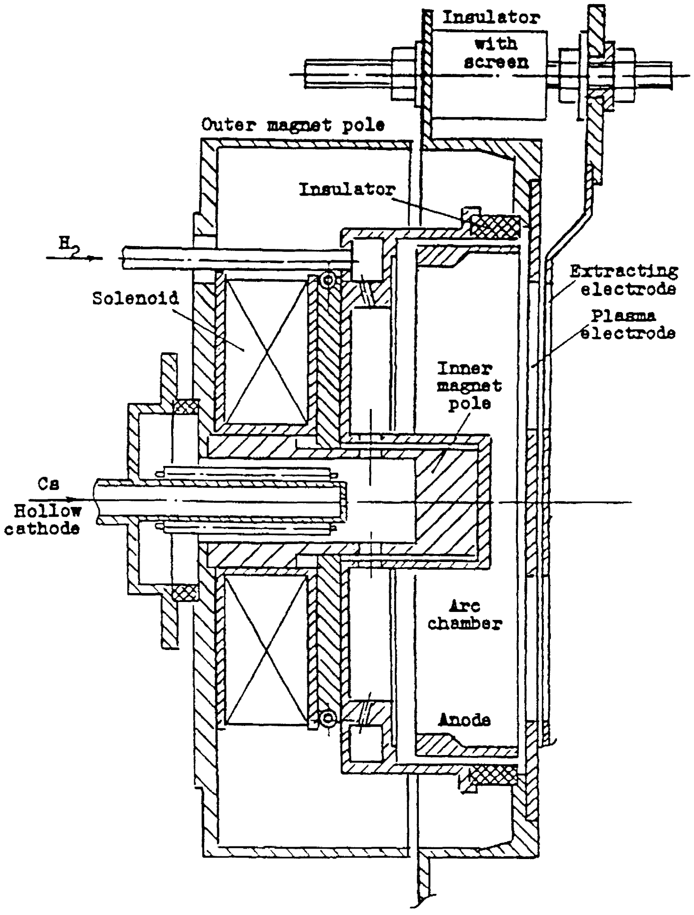

FIG. 21. Schematic diagram of Elizarov ion source.

these sources the filaments are kept totally inside the cusp field for maximum efficiency of gas ionization. The cusp fields surrounding the driver plasma enable superior confinement, since all but the lowest-energy electrons can be confined by mirror trapping resulting in their being bounced back into the plasma. The need for electrostatic electron confinement is thus reduced or even eliminated. The magnetic cusp fields also allow some reflection of the positive ions back into the driver plasma. This effect can be enhanced by biasing the containment wall less negative (or even positive) relative to the plasma.

Several more recent embodiments of this source have been vigorously pursued. One source pursued at Culham (McAdams *et al.*, 1987) (Fig. 25) employs cusp-field confinement on all but one side, as shown. The magnetic filter in this approach is provided by the two rows of linear cusp-field magnets that provide a transverse magnetic field throughout the device. A plasma with hot electrons (driver plasma) is situated on the filament side of these filter fields, while the negative-ion generation (target) plasma is shown on the extraction side.

Large-scale, multiaperture sources of this general type are being designed and built at the Japan Atomic Energy Research Institute (JAERI) (Okumura *et al.*, 1986). The electrons have to be collected locally in these sources, since the apertures are very close together. In addition, the proximity of apertures precludes the use of collars and filter-field tailoring, both of which may be useful, as discussed in Sec. 2. These sources have produced 3 A of H^- (by 1991). On a $\frac{1}{5}$-scale cesiated hydrogen source, 3.9 A of H^- at 73 keV were extracted

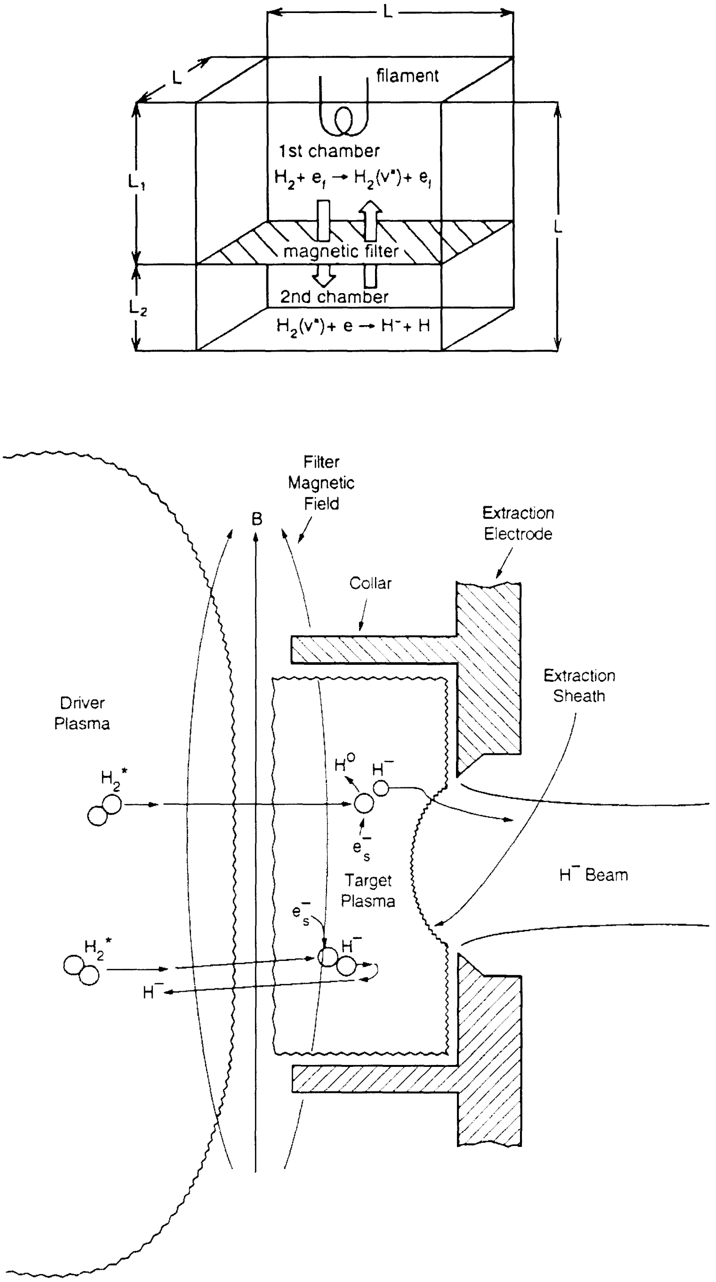
L
L
filament
1st chamber
H2+ ef → H2(v")+ ef
magnetic filter
2nd chamber
H2(v")+ e → H-+ H
L1
L2
L
Filter
Magnetic
Field
B
Extraction
Electrode
Collar
Extraction
Sheath
Driver
Plasma
H2*
H0
H-
es-
Target
Plasma
H- Beam
es-
H2*
H-
H-

FIG. 22. Top: Illustration of principle of dissociative-attachment ion source. Bottom: Actual embodiment illustrating the role of collar and magnetic filter.

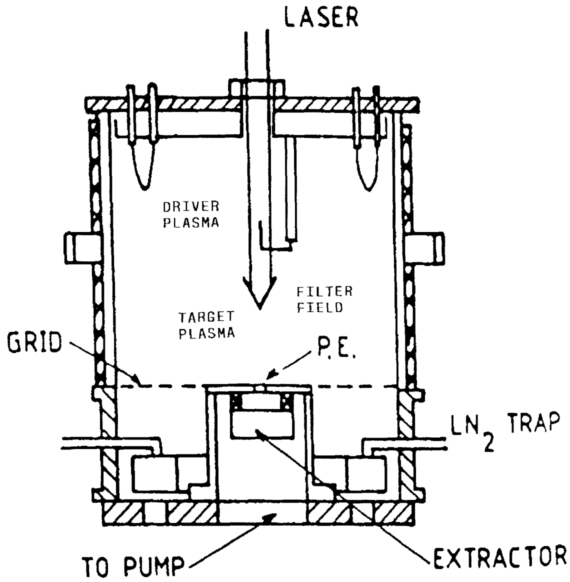

FIG. 23. Early Bacal source.

at NIFS (Kuroda *et al.*, 1996) and 14 A of H^- were extracted and accelerated to 125 keV in a cesium-seeded source. At JAERI, 10 A of H^- were produced (Inoue 1993); 13.6 A of H^- were extracted without cesium and accelerated to 125 keV (Takeiri *et al.*, 1996); and 5.9 A of D^- were extracted to 400 keV, with cesium (1.7 A of D^- without cesium) (Okumura *et al.*, 1996).

Multibeamlet ion sources of this type are also being actively developed at JAERI as shown in Figs. 26 and 27. The former shows the arrangement of edge magnets, which provide the plasma configuration and the magnetic filter. Also shown is the electrostatic accelerator including a thick extraction grid. This thick electrode is intended to catch the tangentially extracted electrons, assuming that no secondary electrons are produced from this surface being bombarded with 10-keV electrons, tangentially. These volume sources have been found to work better with the injection of Cs, as shown, which usually results in a significant (frequently a factor of 3) increase in the negative-ion output accompanied with a decrease in extracted-electron output (sometimes a factor of 1/100). Possible explanations will be discussed in Sec. 2.2. Another version is named the Kamaboko source (shown in Fig. 27) because the plasma containment vessel is said to be shaped like a fish. This source has been developed for thermonuclear fusion experiments and has delivered 18.4 amps of H^- current at 350 keV energy. The current density at the source is about 10 mA cm^{-2} (Ohara, 1998).

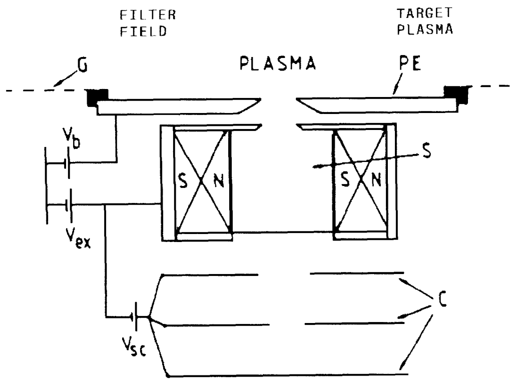

FIG. 24. Blowup of extractor region.

Development of simple beamlet sources, now mostly for accelerator applications for high-energy physics and materials science (spallation neutron sources), has proceeded along several lines. One of those is the spallation-neutron source design initiated at Berkeley (Leung, 1998).

In 1993, a variant of the Bacal source employing cusp fields, a magnetic filter, and a

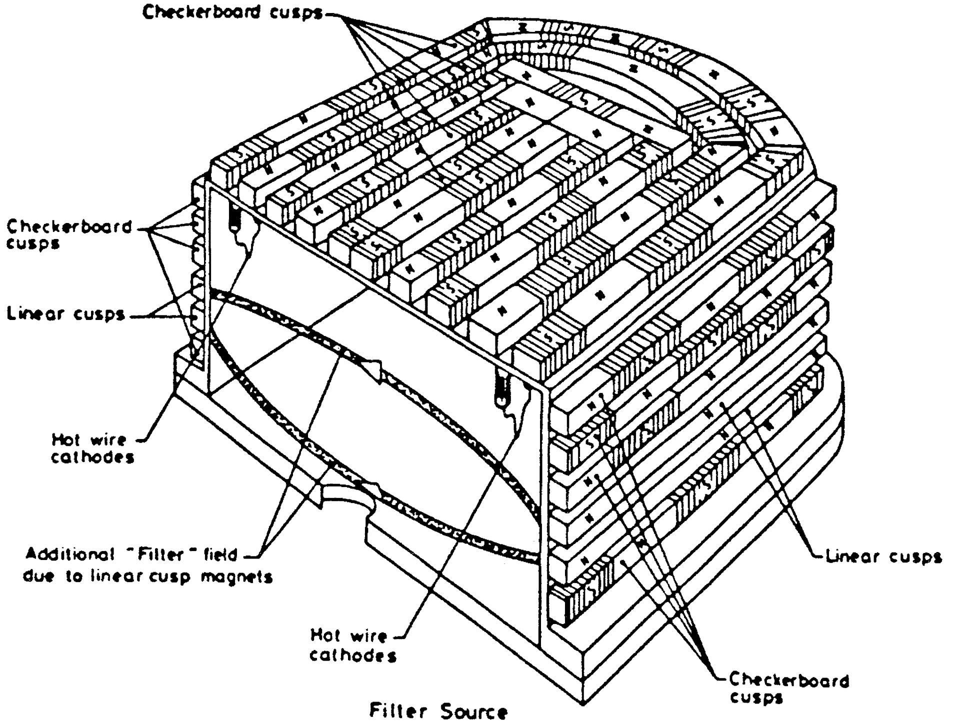

FIG. 25. Culham volume source showing a variation in the magnetic-filter configuration and cusp-field confinement on the end wall as well.

collar electrode (introduced earlier in the Dudnikov source) was produced at LBL (Leung 1994). Current densities of 250 mA cm^{-2} have been extracted for small apertures with gas pressures of 50 m Torr and $e^-/H^- \geq 100$. Ion currents of 30 mA have been extracted with a 5.4-mm-diam aperture (Leung, 1991).

An illustration of a collar arrangement is shown in Fig. 28. Notice that the target and driver plasmas are isolated from each other by the "magnetic" insulation provided by the filter fields. This insulation derives from the lower mobility of the electrons across the magnetic field, enabling the buildup of potential differences. This also means that the target plasma may be biased somewhat independently with respect to the driver plasma. The driver-plasma potential is usually determined largely by the outer walls of the ion source, which supply the arc current. The electrons drift into the target plasma via collisions. Each electron undergoes on the order of 100 collisions between the electrons and the H_2 within the filter fields and 200 collisions within the collar region. Each collision serves to cool the 8-eV electrons with the cold H_2. In addition, as depicted by the bottom half of Fig. 29, the electrons bounce off the collar electrode many hundreds of times. The probability of being captured by the collar electrode is very small per approach, but is magnified by as much as two orders of magnitude during the electron transport from the driver region to the extraction sheath. This probability is very sensitive to the plasma self bias Φ_0, so that any reduction in Φ_0 would increase the absorption per pass exponentially. A conceptual diagram is shown in Fig. 28 and a "how it goes together" diagram is shown in Fig. 30. This source uses a collar electrode, which is known to increase H^- output, especially with the addition of cesium. Discussion of this effect appears in Sec. 2.2.6.

1.3.3 rf Plasma Generation In this type of source, rf energy is injected and used to generate and maintain the plasma, instead of a filament, hollow cathode, or cold cathode.

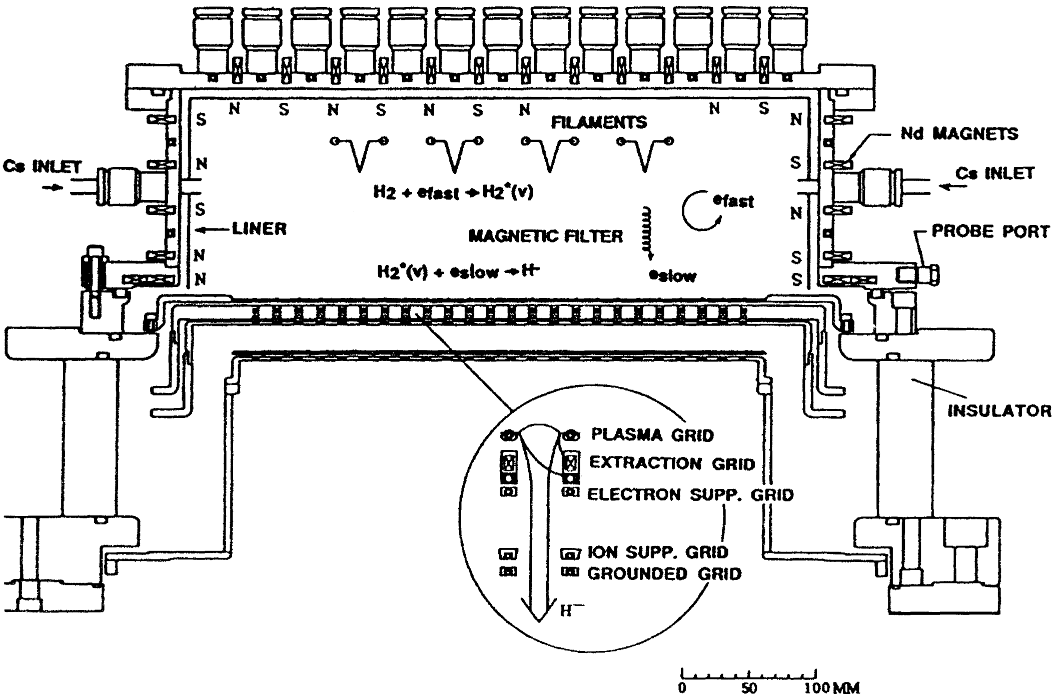
N S N S N S N
FILAMENTS
N S N
S
N
S
N
N
S
N
S
S
Cs INLET
Nd MAGNETS
Cs INLET
LINER
PROBE PORT
H2 + efast → H2*(v)
MAGNETIC FILTER
efast
H2*(v) + eslow → H−
eslow
INSULATOR
PLASMA GRID
EXTRACTION GRID
ELECTRON SUPP. GRID
ION SUPP. GRID
GROUNDED GRID
H−
0 50 100MM

FIG. 26. A schematic diagram of JAERI multiampere H^- ion source. Cesium oven is attached in the present experiment.

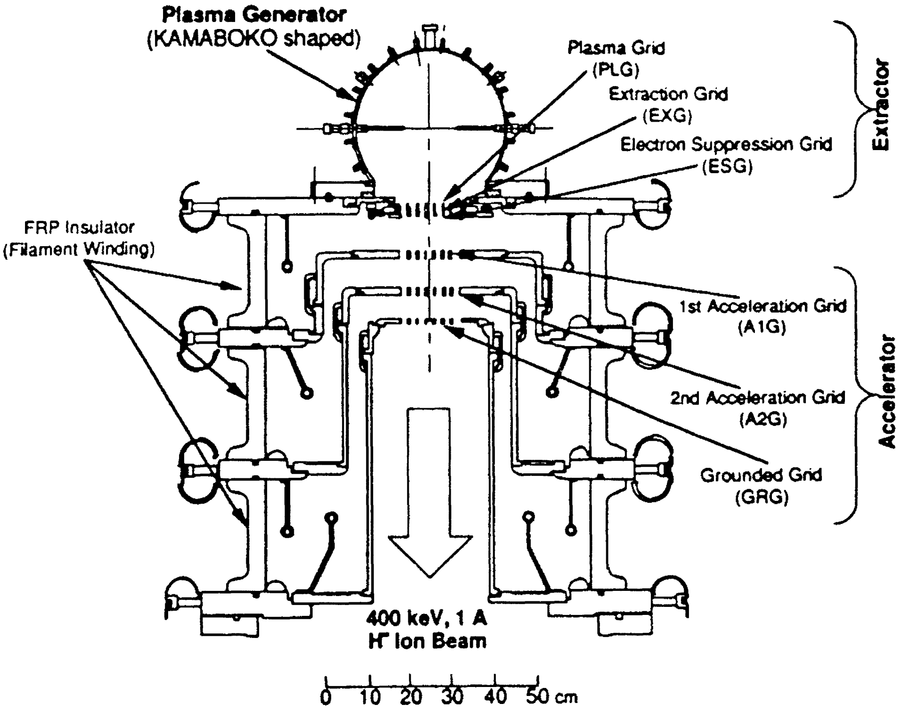
Plasma Generator
(KAMABOKO shaped)
Plasma Grid
(PLG)
Extraction Grid
(EXG)
Electron Suppression Grid
(ESG)
Extractor
FRP Insulator
(Filament Winding)
1st Acceleration Grid
(A1G)
2nd Acceleration Grid
(A2G)
Grounded Grid
(GRG)
Accelerator
400 keV, 1 A
H− Ion Beam
0 10 20 30 40 50 cm

FIG. 27. A cross-sectional view of the 400-keV negative-ion source.

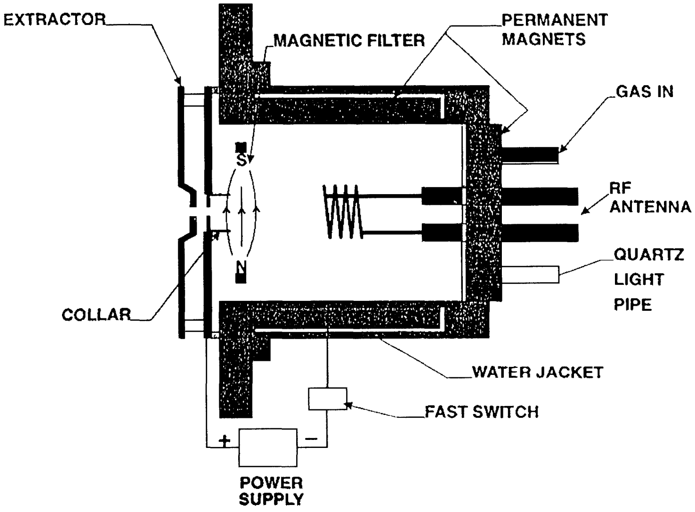
EXTRACTOR
MAGNETIC FILTER
PERMANENT MAGNETS
GAS IN
S
N
RF ANTENNA
QUARTZ LIGHT PIPE
COLLAR
WATER JACKET
FAST SWITCH
+
−
POWER SUPPLY

FIG. 28. Preliminary version of Spallation Neutron Source H$^-$-ion source.

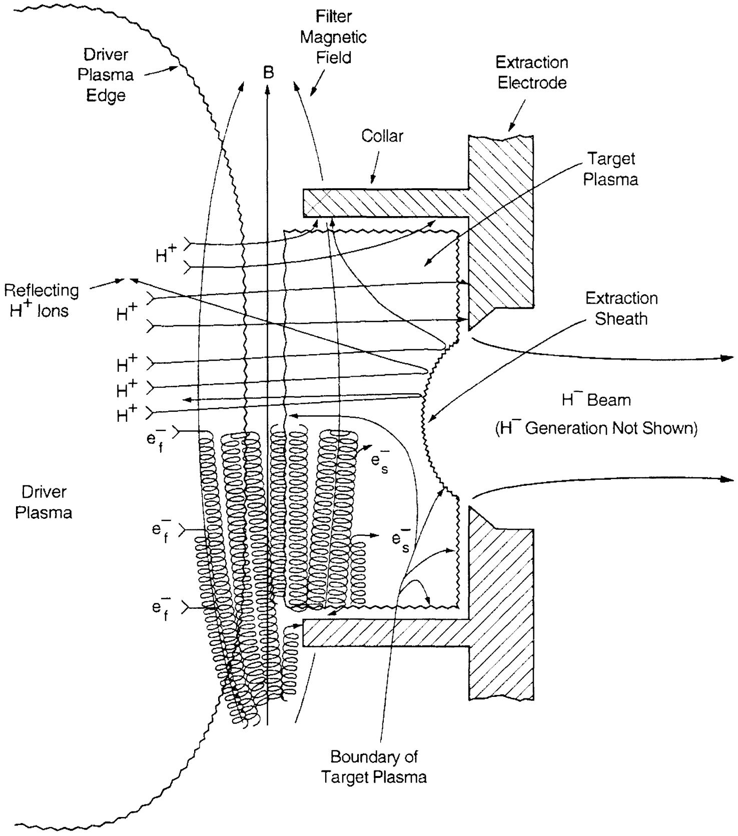
Driver Plasma Edge
B
Filter Magnetic Field
Extraction Electrode
Collar
Target Plasma
H⁺
Reflecting H⁺ Ions
H⁺
H⁺
H⁺
H⁺
Extraction Sheath
H⁻ Beam
(H⁻ Generation Not Shown)
e⁻f
e⁻s
Driver Plasma
e⁻f
e⁻s
e⁻f
Boundary of Target Plasma

FIG. 29. Charged-particle flow in volume negative-ion source.

Antenna lifetime has been studied by Peters (1998), and so far in a particular configuration the mean lifetime is somewhat under a week with continuous operation. The magnetic filter shown in Figs. 28 and 30 allows tailoring of the magnetic field to the collar geometry, which could be important. An rf-driver volume ion source was pursued at LBL (Leung 1991, 1994; Kwan *et al.*, 1993). Development in antenna design is continuing to achieve long-pulse operation at high power (Lee *et al.*, 1998). For long-term reliable operation without injection of impurities, filaments might be replaced by rf injection by antennas or horns. Several sources of this variety have been built (Fig. 28). Recent data are questioning the reliability of present development (Peters, 1998).

1.3.4 Penning Source A Penning configuration for the driver plasma is possible with a natural insulation of the driver and target plasmas taking place through the configuration of the end walls (see Fig. 31). This configuration was explored at ORNL and was named the VITEX, for Volume Ion Transverse EXtraction (Dagenhart *et al.*, 1987).

1.3.5 Toroidal Discharge One embodiment (Alessi and Sluyters, 1990), which combines the geometry of the driver plasma indicated in Sec. 1.3.1 with magnetic cusp confinement, is shown in Fig. 32. It produces 35 mA with an $e^-/\mathrm{H}^- = 5$ and 20 mA with $e^-/\mathrm{H}^- = 3$ (Alessi and Prelec, 1991). The normalized rms emittance of this source is measured to be 0.32π mm-mrad for a 13-mA beam, which corresponds to an ion temperature of 0.6 eV, an unusually low value (Alessi and Sluyters, 1990). This source generates a plasma in region 1 by electron-impact ionization from the electrons generated by the filament and accelerated through the plasma sheath surrounding the filament. This driver plasma is confined by the cusp fields on the edge of the container as shown. By drifting through the magnetic fields within the plasma, low-energy electrons arrive in the target region 2. The device was subsequently embellished (Alessi *et al.*, 1994) by providing a stronger conical intraplasma magnetic field (Fig. 33). This increased strength provides fewer unwanted high-energy electrons in region 2. The conical shape ensures that the maximum pos-

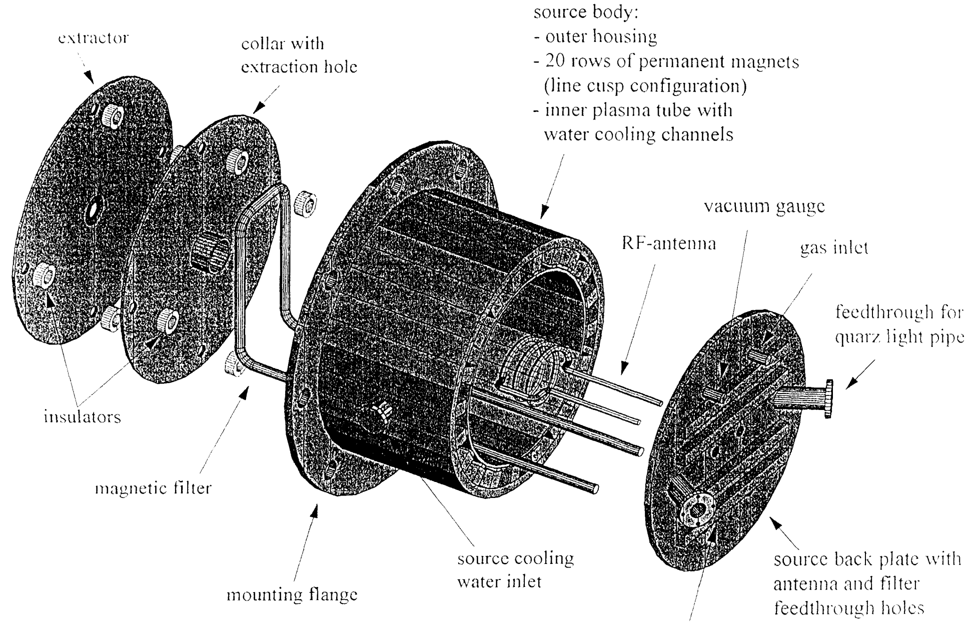
extractor
collar with extraction hole
source body:
- outer housing
- 20 rows of permanent magnets (line cusp configuration)
- inner plasma tube with water cooling channels
vacuum gauge
RF-antenna
gas inlet
feedthrough for quarz light pipe
insulators
magnetic filter
mounting flange
source cooling water inlet
source back plate with antenna and filter feedthrough holes

FIG. 30. Details of a preliminary version of the SNS ion source.

sible flux of low-energy electrons appears in the dissociative attachment regions. Also, because of the diverging magnetic fields near the extraction aperture 3, electrons are discouraged from being extracted. This is very important as indicated in Sec. 2.2.4. This concept has been subsequently explored at LANL (York *et al.*, 1993) with some variations in the filter-field configuration. H^- ion currents of $20\,\mathrm{mA\,cm^{-2}}$ (without cesium) and $40\,\mathrm{mA\,cm^{-2}}$ (with cesium) have been achieved.

1.3.6 Adding Cesium The effect of adding cesium to a volume source is generally that the negative-ion current increases by as much as a factor of 3. Also, and no less important, the extracted electron current decreases by a factor of as much as a hundred. This is shown in Fig. 34. Adding cesium in the compact multicusp source described in Sec. 1.3.2 increased the current from 7 to 34 mA. The current density with cesium was as much as $1.2\,\mathrm{A\,cm^{-2}}$ (Leung *et al.*, 1989). The addition of cesium in the toroidal filter source being explored at Brookhaven National Laboratory (BNL) (York *et al.*, 1993) provided a doubling of the negative-ion current and a reduction of e^-/H^- extracted from 5:1 to 2:1.

2. PHYSICS OF HYDROGEN NEGATIVE-ION EXTRACTION

It is one matter to produce negative ions as discussed in Sec. 1 but quite another to extract them into a beam with low emittance. In this section we discuss two classes of extraction: direct extraction from surface sources and extraction from volume sources. Each has its peculiarities, and the physics and computational issues of each are discussed.

2.1 Ion-Extraction Physics for Surface Sources

Surface sources are divided into two categories depending on whether the magnetic field is large or small.

2.1.1 Low-Field Surface Sources For direct-extraction sources, the extraction physics may seem relatively straightforward, since the negative ions travel through the intervening plasma at a high energy of 100 eV. However, for the cases where there is a large distance between the converter and the extraction region, finite-T_e effects may be important and need to be treated with care.

For volume extraction, all situations need special treatment since the analysis used for positive-ion extraction is not appropriate. Lawrence Berkeley Laboratory proposed a direct-extraction negative-ion source (Ehlers and Leung, 1980), which is shown in Fig. 35. Measured beam widths are shown in Fig. 36. The common interpretation of the large beam width at the focal point is that it is due to the high ion temperature ($\approx 7\,\mathrm{eV}$) coming off the converter. While this is a significant cause for this beam divergence, another one is the high space-charge forces within an imperfect ($T_e > 0$) plasma in which they are transported. An illustration of this latter effect is provided in Fig. 37, which shows the ion trajectories for three different plasma electron temperatures. At higher electron temperatures, deviations from perfect plasma neutrality occur, causing the beam optics to deviate from the idealistic case of Fig. 37 (top) with $T_e = 0.25\,\mathrm{eV}$. The solution of the Vlasov–Poisson equations for this type of source is usually done with the assumption (Whealton *et al.*, 1984, 1987, 1988) that the sum of the positive ions and electrons is a Boltzmann distribution. The temperature, denoted as the electron temperature, is actually that of the Boltzmann distribution. This type of analysis is expected to give a lower bound to various optical aberrations, since it neglects any Bohm ion-acoustic instabilities at the beam edge and also neglects finite negative-ion temperature (as much as 6 eV) coming off the converter. Finite ion temperatures in the limit of zero plasma temperature were considered separately (Anderson, 1984). For finite plasma temperature, $T_e = 4\,\mathrm{eV}$, which is the range commonly expected, noticeable but perhaps not drastic deviations occur; for $T_e = 16\,\mathrm{eV}$, which is higher than expected, the crossover essentially disappears, and the optics are seriously affected for this geometry. Other geometries could produce better options for these high–electron-temperature situations. The fraction of beam transmitted and beam diver-

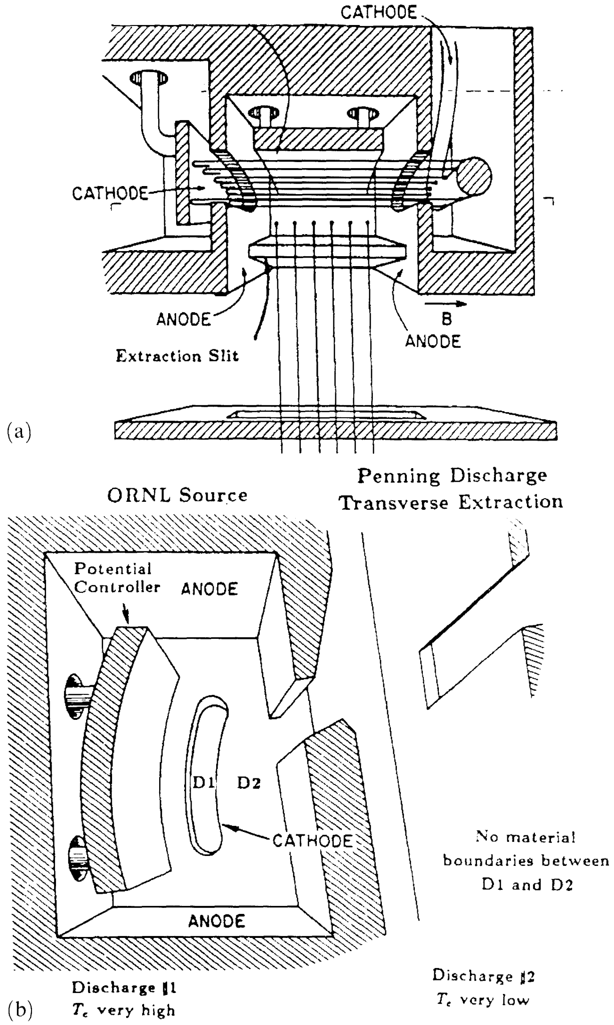
CATHODE
CATHODE
ANODE
Extraction Slit
B
ANODE
(a)
ORNL Source
Penning Discharge
Transverse Extraction
Potential
Controller
ANODE
D1
D2
CATHODE
ANODE
No material
boundaries between
D1 and D2
Discharge #1
T_e very high
Discharge #2
T_e very low
(b)

FIG. 31. (a) Top view and (b) elevation of VITEX source. Two plasmas are defined by the cathode slot, discharge DI being the Driver plasma and DZ being the target plasma.

gence as functions of electron temperature in eV, derived by the accelerator potential, are given in Fig. 38, showing the regions of single vignetting and multiple vignetting.

A more extreme situation of imperfect transport is shown for the Los Alamos Meson Pion Factory (LAMPF) injector (Whealton *et al.*, 1984), on account of a much more elongated region between the converter and extraction. The idealistic case of 0.0009 eV shows the optics as intended (York, 1983). However, even for $T_e = 0.1$ eV, which is far lower than expected, deviations from a perfect crossover point are expected. For the more realistic electron temperature of 10 eV, as shown by Whealton *et al.* (1984), the optics are significantly affected.

2.1.2 High-Field Surface Sources The long transport region that was at least conceivable in the field-free sources discussed in the previous section is not possible in the high (magnetic)-field sources where the transport region must be short compared with an ion gyroradius. For the SITEX source, Fig. 39, the negative-ion trajectories from the converter in two dimensions (2D) are shown in Fig. 40 (Whealton *et al.*, 1984), and a scheme to use a finite ellipsoidal segment converter to obtain an approximately round beam is shown by the

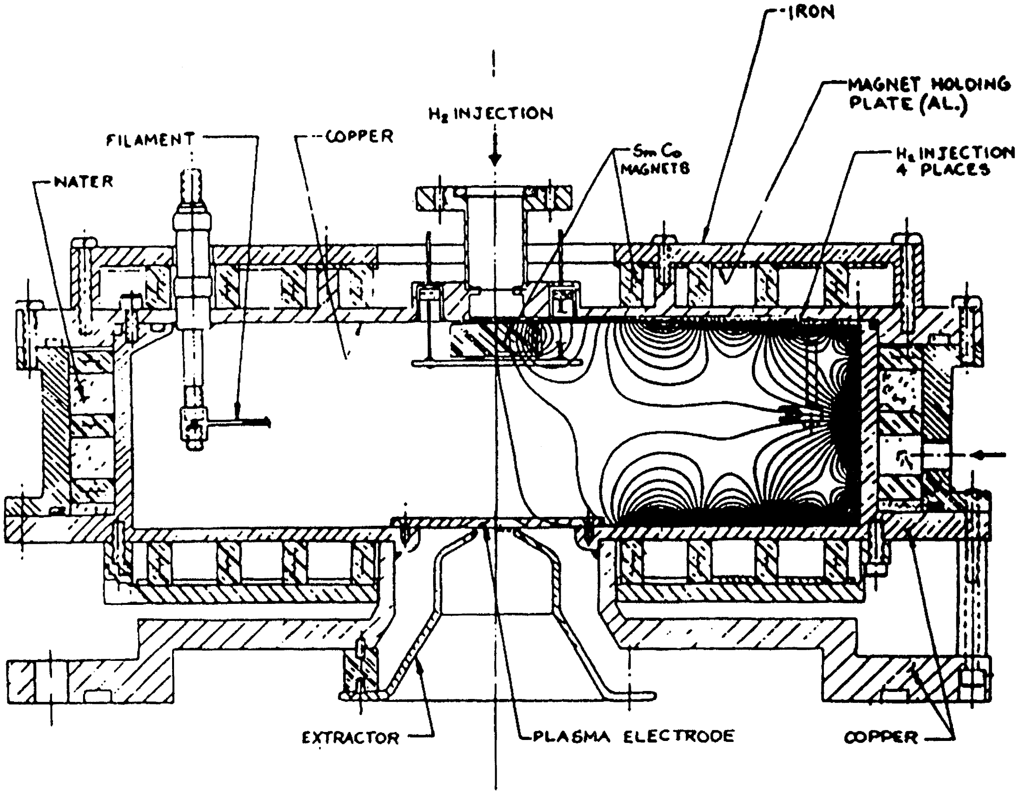
IRON
MAGNET HOLDING PLATE (AL.)
H₂ INJECTION
FILAMENT
COPPER
WATER
Sm Co MAGNETS
H₂ INJECTION 4 PLACES
EXTRACTOR
PLASMA ELECTRODE
COPPER

FIG. 32. BNL volume negative-ion source.

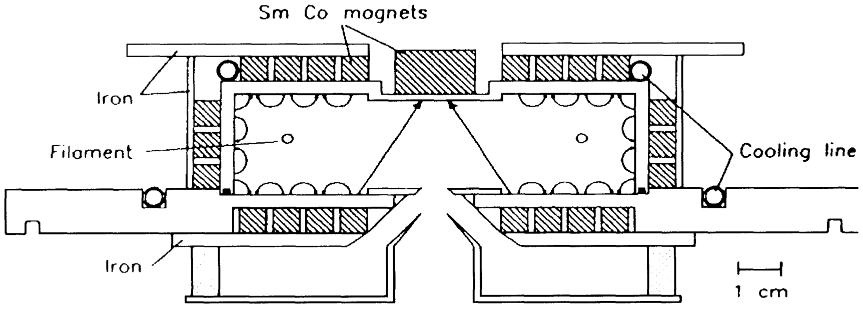
Sm Co mognets
Iron
Filament
Iron
Cooling line
1 cm

FIG. 33. Improvement on toroidal volume ion source.

3D calculation of Fig. 41 (Whealton *et al.*, 1985). The ellipsoidal connector ion source was designed such that the magnetic field was parallel to the surface and a Penning discharge would feed positive ions to the converter. A better scheme is to use the magnetron architecture described previously and also to use the charge-exchange focusing technique. An optimized version of all this is discussed in Sec. 1.2.2.

2.1.3 Dudnikov Source The Dudnikov source is considered to be a volume-production source since the actual ions extracted are

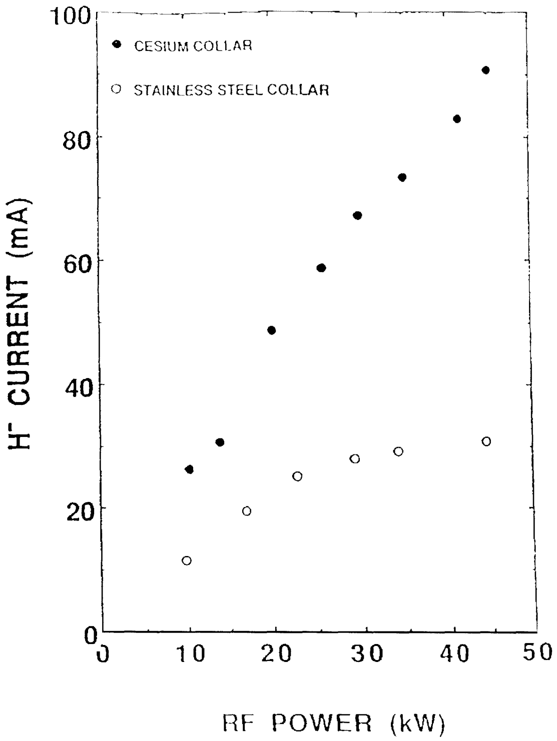

FIG. 34. Effect of cesium on negative-ion production.

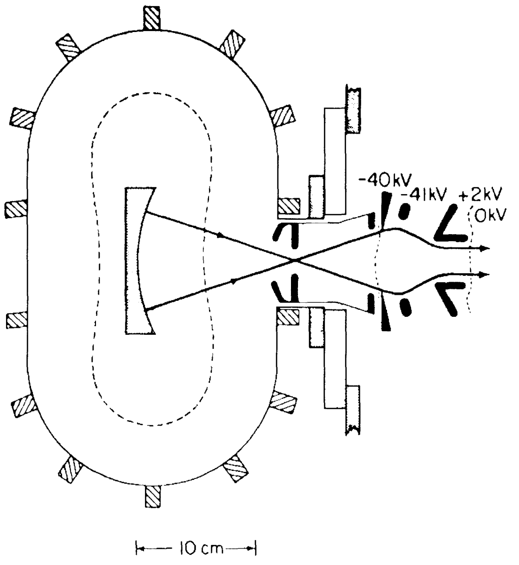

FIG. 35. Field-free converter source considered at LBL showing converter, ideal ion envelope, various vignetting and accelerating electrodes and cusp-field plasma containment.

produced through charge exchange from surface-produced H^-. Initial attempts to model this source assumed perfect symmetry with positive-ion extraction, showing something like Fig. 42 (Smith *et al.*, 1984) in 2D, and Fig. 43 (Bell *et al.*, 1985) in 3D. This approach generally was at significant variance with experiments that showed a variation of ion current with extraction voltage. The oblique sheath produced by the $E \times B$ drifting electrons, which also must share the transverse space-charge limit, is important but was not considered at the time. The existence of a drift region just before the extraction plane was suggested by Dudnikov and has an optimum thickness that is the balance between negative-ion depletion and extracted-electron depletion. The electron current decreases much faster with thickness than the ion depletion (Allison *et al.*, 1980).

2.2 Ion-Extraction Physics for Volume Sources

Negative-ion extraction from volume sources is less robust than extraction from

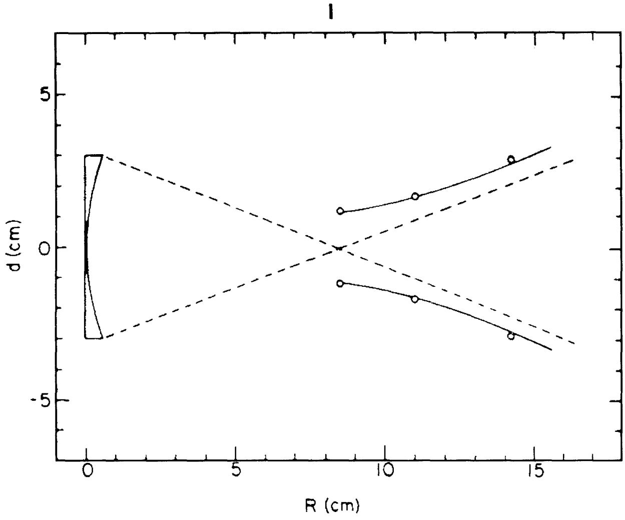

FIG. 36. Measured beam width (FWHM) compared with the idealistic zero-ion-temperature envelope trajectory.

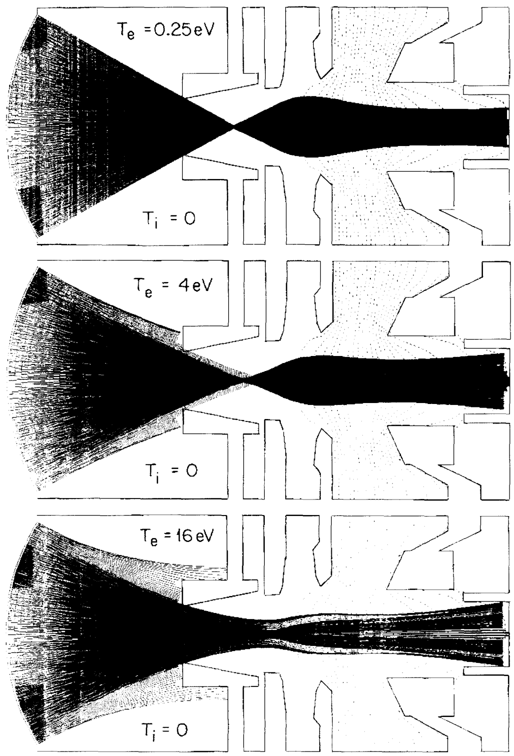

FIG. 37. Ion trajectories, still assuming zero ion temperature but considering a finite electron temperature.

surface sources, where the ions are rocketing toward the extraction aperture at energies of 100 eV or more. In fact, volume negative-ion extraction is less robust than volume positive-ion extraction for a number of reasons. For volume sources, creation of initial energies on the order of only 1 eV are expected. If the intraplasma electric fields accelerated the negative ions toward the extraction aperture, as is largely the case with positive ions in positive-ion sources, then the low initial kinetic energy of the negative ions would not be so important. The ionic flow is illustrated in Fig. 44(a) for a steady-state positive-ion source. For cases where ion-acoustic plasma-sheath instabilities develop, the situation is typified by Fig. 44(b). These positive ions, whether stable or otherwise, are electrostatically attracted toward the extraction region and either are extracted or intercept the plasma electrode. Unfortunately, these very intraplasma electric fields that tend to accelerate the positive ions toward the extraction region accelerate most of the newly formed negative ions in the opposite direction, toward the source plasma (Whealton *et al.*, 1987).

2.2.1 Asymmetries Between Positive- and Negative-Ion Extraction

For positive-ion extraction, the analysis is well developed (Whealton, 1981) using the following form of the Vlasov–Poisson equations:

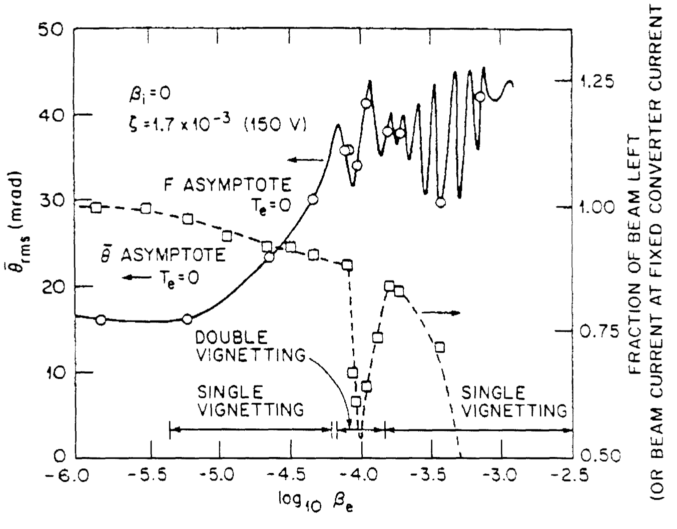

FIG. 38. Beam divergence and transmission as functions of electron temperature assuming an ion temperature of zero.

$$\nabla^2\varphi(\mathbf{r},t) = -\int f(\mathbf{r},\mathbf{v},t)\,\mathrm{d}\mathbf{v} + \exp[-\varphi(\mathbf{r},t)], \tag{3}$$

$$\frac{\partial f(\mathbf{r},\mathbf{v},t)}{\partial t} + \mathbf{v}\cdot\nabla f(\mathbf{r},\mathbf{v},t) + \{\mathbf{v}\times\mathbf{B} - \nabla\varphi\}\cdot\nabla_v f(\mathbf{r},\mathbf{v},t) = 0, \tag{4}$$

where φ is the electrostatic potential; f is the ion distribution function, which is a function of position $\mathbf{r}$, velocity $\mathbf{v}$, and time t; $\mathbf{B}$ is the magnetic field.

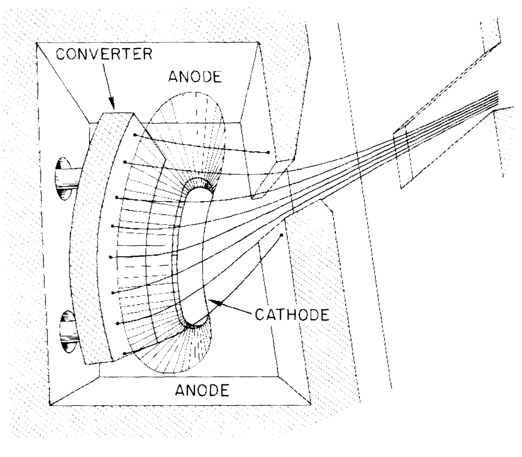

FIG. 39. Elevation view of the ORNL SITEX surface negative-ion source showing the converter, surface negative-ion trajectories, Penning discharge as projected from anode slot and extractor slot.

The first convergent solution to these equations in 2D was obtained by Cooper *et al.* (1974), for a partial sheath, and by Whealton *et al.* (1978) and Whitson *et al.* (1978) for a full sheath in 2D. The three-dimensional solution came later (Wooten *et al.*, 1979, 1981; Whealton *et al.*, 1986) and ion-acoustic time scales yet later (Whealton 1985).

The first asymmetry between negative-ion sources and a reversed-potential positive-ion source is that the electrons cancel charge imbalance excursions in positive-ion sources much better than positive ions do in negative-ion sources. This observation led to the heuristic model of Whealton *et al.* (1989), which appears to be in agreement with the laboratory data of Bacal and Stevens; see Figure 45. An example of the actual trajectory computations, at points A, B, C, and D in Fig. 45(c), is shown in Fig. 46. The plasma sheath condition in these cases varies from extremely underdense (A) to extremely over-dense (D), all falling within the laboratory observations at Ecole Polytechnique. Similar results were obtained for the LASL data [Fig. 45(a)]. It is common to model negative-ion extraction from volume sources using the above model (Tanaka *et al.*, 1998; Watanabe *et al.*, 1996; McAdams, 1994; Cowan, 1994; Bell *et al.*, 1985; McGaffey *et al.*, 1984; Smith *et al.*, 1984; Gudharay *et al.*, 1994; Chan, 1993, 1994; Maeser *et al.*, 1996; Whealton *et al.*, 1986, 1988), an example of which is shown in Fig. 47. How-

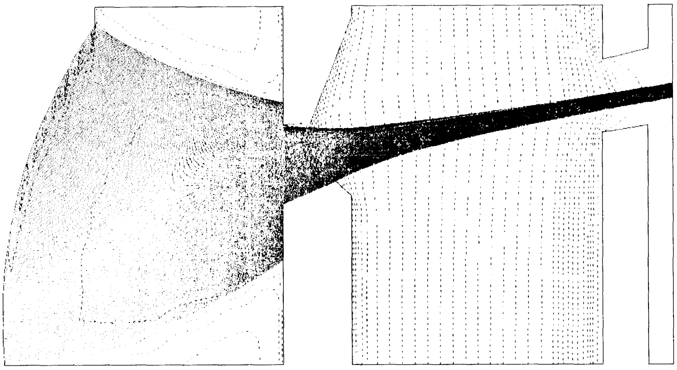

FIG. 40. Computer simulation (2D) of trajectories of negative ions from converter and extracted into slot beam.

ever, this procedure has little meaning, since it assumes that the positive ions are in equilibrium (which is not true) and neglects any explicit consideration of plasma electrons, which could be very important. Furthermore, the model assumes that the negative ions enjoy the same downhill run from plasma to extraction that positive ions have in positive-ion sources.

2.2.2 Saddle-Point Extraction The second asymmetry between negative-ion sources and reversed-potential positive-ion sources arises from the observation that the plasma electrode is biased negative to the local plasma potential for negative-ion sources just like in positive-ion sources and for the same reason—to contain the plasma electrons electrostatically. Therefore, in a positive-ion source, there is a continuous, monotonic downhill run for the positive ions generated deep in the plasma from formation through extraction. For negative ions, there is a saddle point formed within the plasma, since the plasma potential lies intermediate between those of the plasma electrode and the acceleration electrode. This saddle point is shown in Fig. 48, which for illustrative purposes is calculated in the limit of zero space charge (vacuum fields). For the positive ions [Fig. 48(a)], there is acceleration toward extraction in the region where $A_z < 0$ until they reach the saddle range ($A_z = 0$), after which they decelerate, turn around, and either return to the plasma or intercept the plasma electrode. The created negative ions, shown in Fig. 48(b), are electrostatically repelled from the extractor in regions labeled $A_z < 0$ and are electrostatically attracted toward the extractor in the region $A_z > 0$. A large volume of negative ions created is repelled from the extraction region. There is no extraction sheath in this case, since the plasma density is zero. This observation of a potential hill was made ten years ago (Whealton *et al.*, 1987) but has attracted scant attention. Recent notice has resulted in a heuristic model of a virtual cathode (Becker *et al.*, 1998), which may eventually provide a useful approach. Another approach in order to avoid the complex nonlinear theoretical problem is to use steady-state intraplasma fields derived from experiments and then plot trajectories

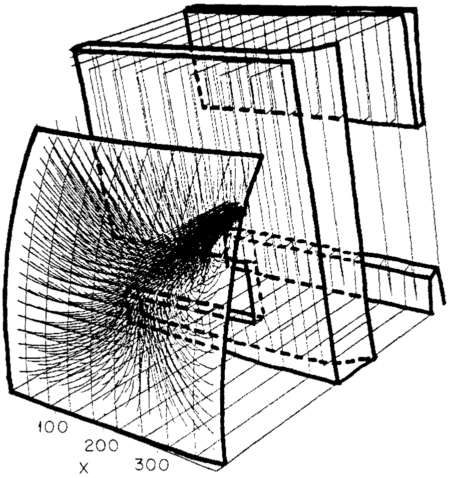

FIG. 41. Computer simulation (3D) of trajectories of negative ions from connecter and extracted into finite slot beam. The slot ends have been tailored as shown to minimize aberrations from slot end effects.

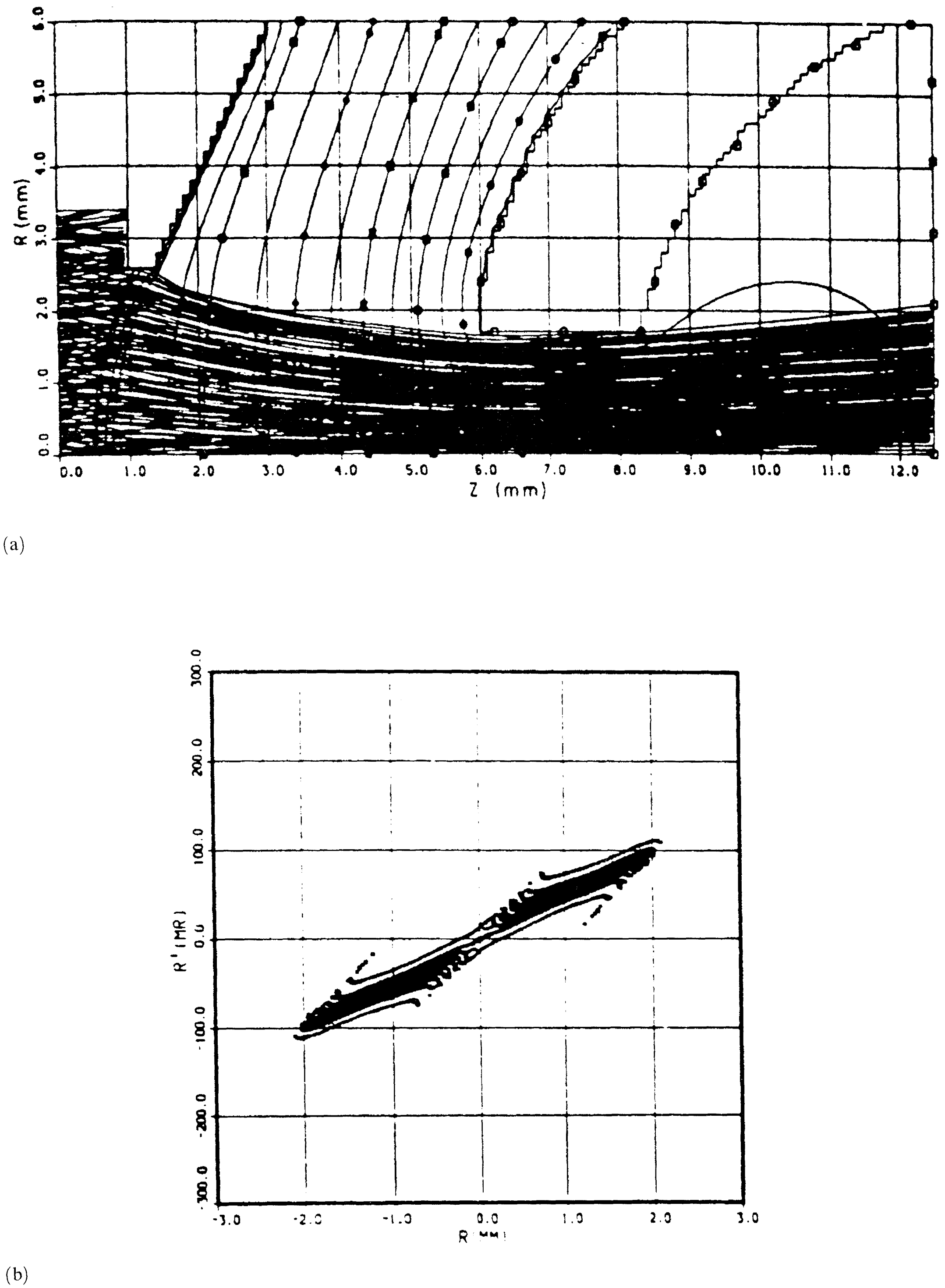

FIG. 42. (a) Snow-code calculation of the H^- ion trajectories for the 4X (Dudnikov) source extraction optics; (b) phase-space diagram at the maximum axial position shown in (a).

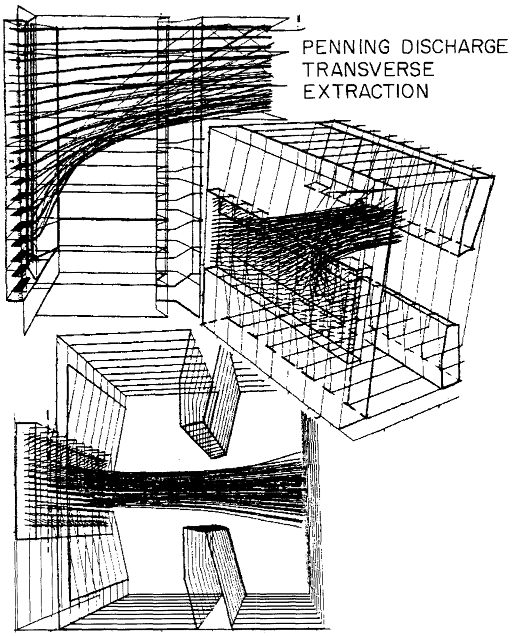

FIG. 43. A 3D calculation of the H^- ion trajectories for the Dudnikov negative-ion source, in three perspective views, showing slot end effects.

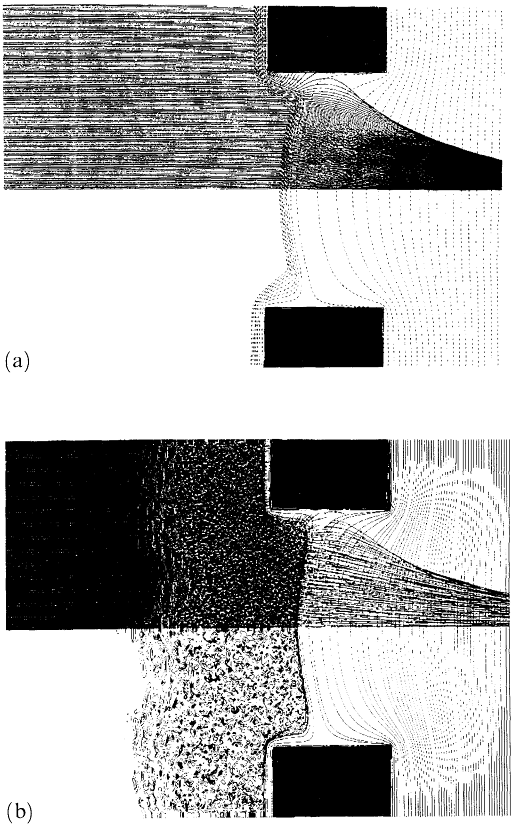

FIG. 44. (a) Plot of ion trajectories (solid lines) and equipotentials (dotted lines) in the extraction region bounded by an electrode (solid) with an aperture for ions. Beam current due to accelerator electrode interception. (b) Same as (a), except that an ion time-scale solution to the Vlasov-Poisson equations is shown for positive-ion extraction (electrons are still Boltzmann) in a region where the Bohm sheath stability criteria is not satisfied, resulting in ion-acoustic waves in the presheath and concomitant rms emittance growth.

numerically through these fields (Riz and Pamela, 1996). This makes the theoretical problem linear and straightforward. Unfortunately, the experimental measurements of plasma potential have, so far, not been resolved to provide reliable fields. Furthermore, ion-acoustic instabilities may further complicate the measurement so that a steady-state treatment of ion kinetics is questionable. Negative ions formed on one side of the ridge go toward the source plasma instead of the extraction aperture. In these cases, the location and control of such ridges could be important.

A third asymmetry results in electrons being extracted along with the negative ions, unlike the case for positive-ion sources. A fourth asymmetry is that the Bohm sheath-stability criterion is expected to be violated more extensively within certain regions for negative-ion sources than is the case for positive ion sources, at least in the absence of cesium. An illustration of the consequences of the violation of the Bohm stability criterion is shown in Fig. 44(b) [although this example is for a single Vlasov positive-ion model (Whealton *et al.*, 1987)]. An example of an actual situation where ion-acoustic instabilities may occur is illustrated in Fig. 49, where the ion temperature appears to be increasing with plasma density. This instability occurs when any process that increases the negative space charge in the pre-extraction region causes the curvature of the potential to become positive (as evidenced from the structure of the Poisson equation). Evidence for the existence of ion-acoustic instabilities in the extraction sheath is present in the disparity between ion temperatures measured in the beam, normally 1–7 eV, and within the plasma (Bacal *et al.*, 1990) of

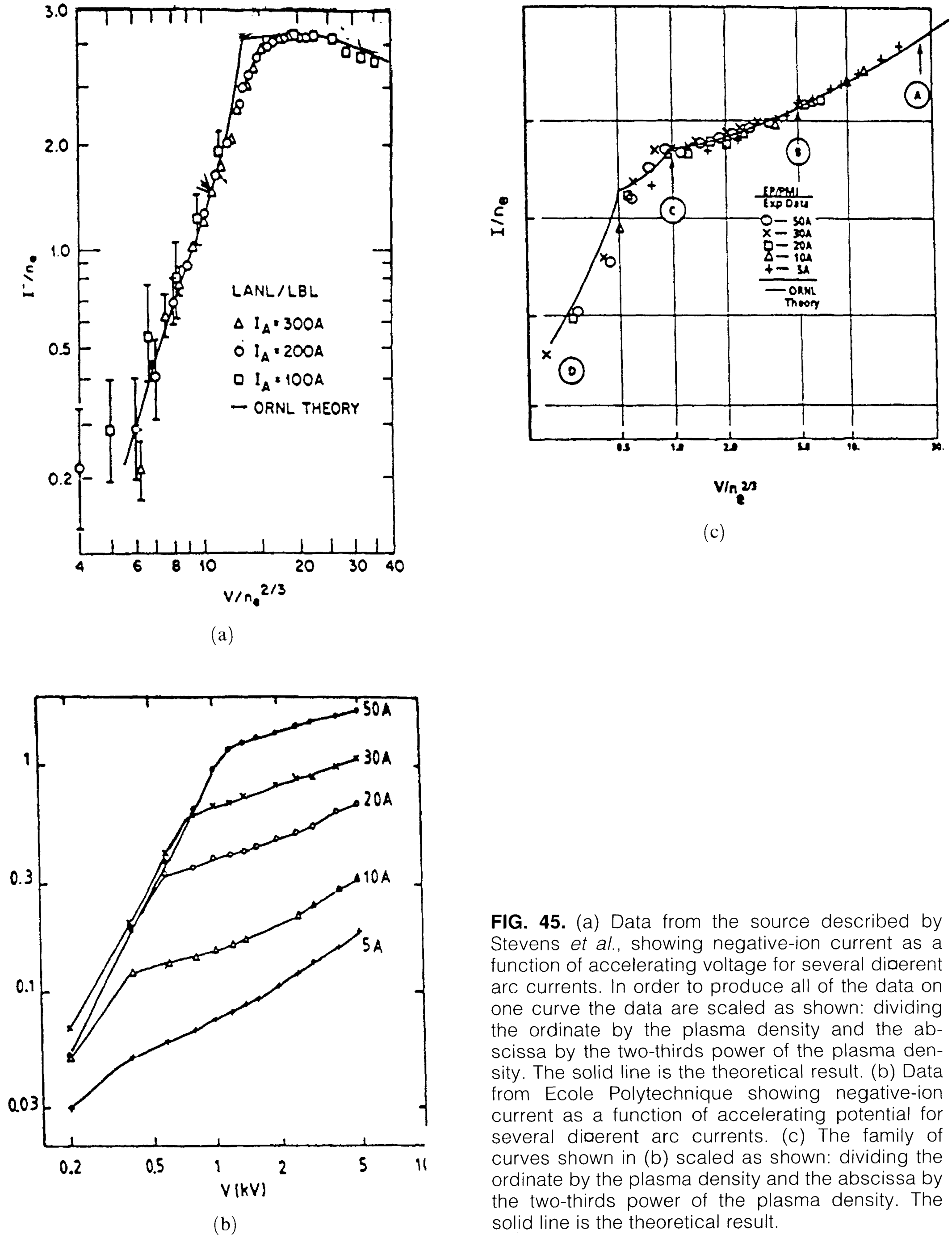

(a)

(c)

(b)

FIG. 45. (a) Data from the source described by Stevens *et al.*, showing negative-ion current as a function of accelerating voltage for several different arc currents. In order to produce all of the data on one curve the data are scaled as shown: dividing the ordinate by the plasma density and the abscissa by the two-thirds power of the plasma density. The solid line is the theoretical result. (b) Data from Ecole Polytechnique showing negative-ion current as a function of accelerating potential for several different arc currents. (c) The family of curves shown in (b) scaled as shown: dividing the ordinate by the plasma density and the abscissa by the two-thirds power of the plasma density. The solid line is the theoretical result.

0.1–0.7 eV. Analysis capable of considering sheath-produced ion-acoustic waves, as a result of the Bohm instability, will be a necessary future step in the accurate modeling of negative-ion sources. These instabilities may actually increase the availability of negative ions for extraction. Still another process, known as stochastic resonance, can occur whereby some negative ions are pushed over the hill by the fluctuating fields.

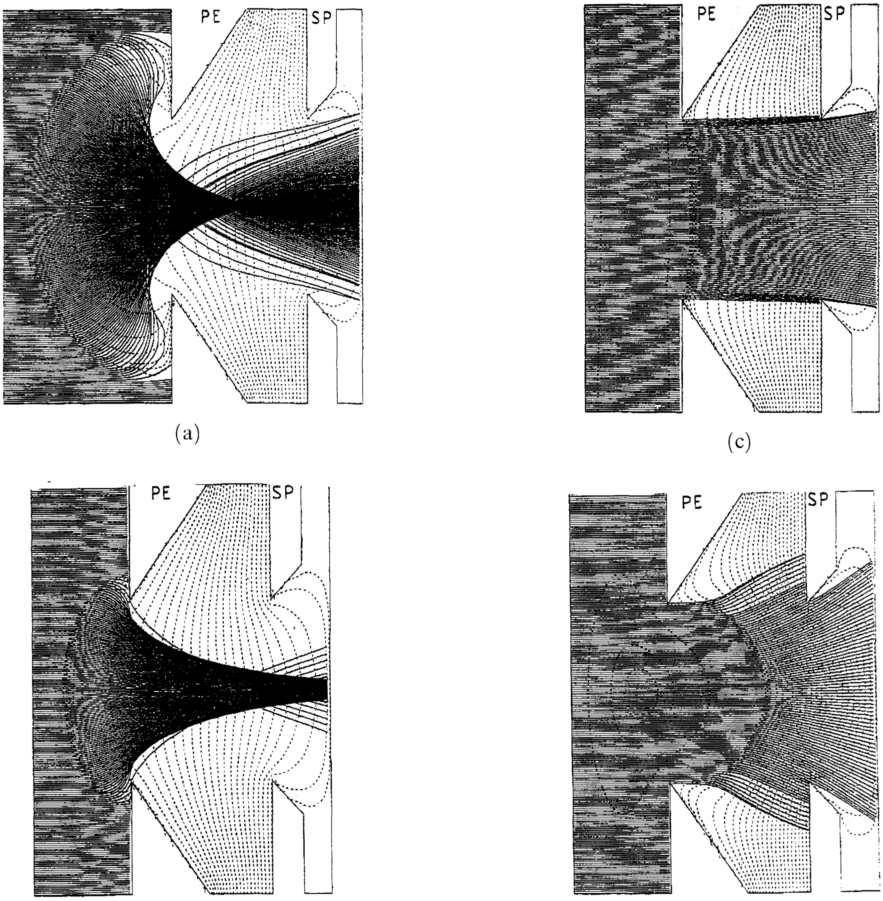
PE
SP
PE
SP
PE
SP
PE
SP

(a)

(c)

(b)

(d)

FIG. 46. (a)–(d) Calculated potentials and H^- ion orbits for the perveances marked A–D, respectively, in Fig. 45 (c).

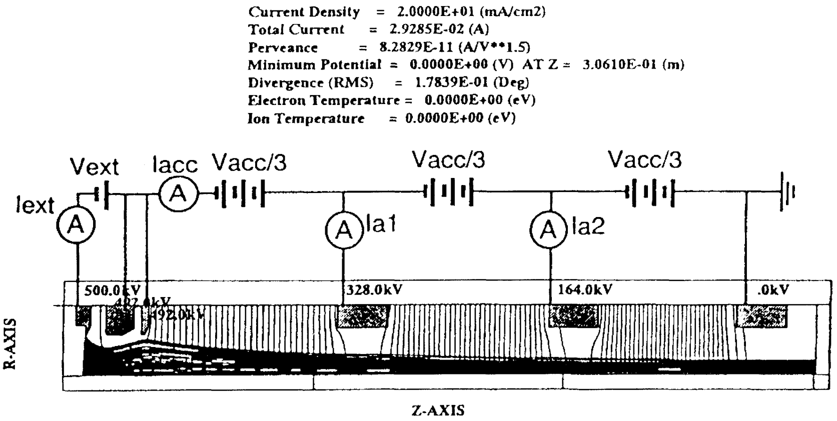

```
Current Density  =  2.0000E+01 (mA/cm2)
Total Current    =  2.9285E-02 (A)
Perveance        =  8.2829E-11 (A/V**1.5)
Minimum Potential =  0.0000E+00 (V)  AT Z =  3.0610E-01 (m)
Divergence (RMS)  =  1.7839E-01 (Deg)
Electron Temperature =  0.0000E+00 (eV)
Ion Temperature   =  0.0000E+00 (eV)
```

Vext
Iacc
Vacc/3
Vacc/3
Vacc/3
Iext
A
A
A
Ia1
A
Ia2
500.0kV
492.0kV
328.0kV
164.0kV
.0kV
R-AXIS
Z-AXIS

FIG. 47. The 500-keV accelerator and beam orbit simulation.

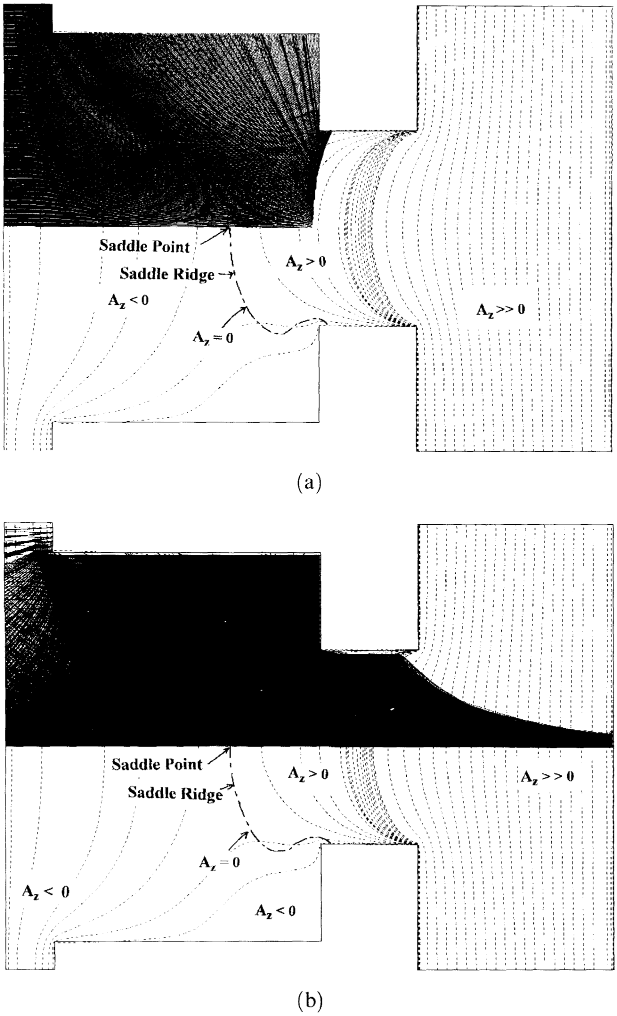

FIG. 48. (a) Positive-ion trajectories shown for negative-ion source with geometry similar to that shown in Fig. 28 or 30. Equipotentials are shown by dashed lines. In this very low-density case (the length of the figure is one Debye length), a saddle point in the potential is clearly shown; the region where volume-produced negative ions are kicked in the direction of the extraction are denoted by $A_z > 0$; regions denoted by $A_z > 0$ indicate that created negataive ions are given a kick back into the plasma. The positive-ion trajectories shown all end up hitting the electrode since they are repulsed by the beam accelerating fields. (b) Volume-created negative ion trajectories for the same case as discussed as above.

2.2.3 Double-Vlasov Model For negative ions it is useful to consider the plasma positive ions using a Vlasov description. It is necessary to consider these plasma positive ions kinetically, Eq. (5), because they are far from equilibrium, executing at most two passes through the region. It is still appropriate to consider the electron density in the presheath to be a Boltzmann distribution since the electrons are collision dominated and generally have to transverse a magnetic field in the pre-extraction region. A description is as follows:

$$\frac{\partial f_+}{\partial t} + \boldsymbol{v} \cdot \nabla f_+ + [\boldsymbol{v} \times \boldsymbol{B} - \nabla\varphi] \cdot \nabla_v f_+ = \delta(S_1), \tag{5}$$

$$\frac{\partial f_-}{\partial t} + \boldsymbol{v} \cdot \nabla f_- - [\boldsymbol{v} \times \boldsymbol{B} - \nabla\varphi] \cdot \nabla_v f_- = G(\boldsymbol{r}), \tag{6}$$

$$\frac{\partial f_e}{\partial t} + \boldsymbol{v} \cdot \nabla f_e - \frac{M}{m_e}[\boldsymbol{v} \times \boldsymbol{B} - \nabla\varphi] \cdot \nabla_v f_e = \delta(S_2), \tag{7}$$

$$\nabla^2\varphi = \int [f_+(\boldsymbol{v}) - f_-(\boldsymbol{v}) - f_e(\boldsymbol{v})]\,\mathrm{d}^3v - \mathrm{e}^{-\varphi}. \tag{8}$$

A case with significant plasma density (which corresponds to 1000 Debye lengths λ_D for the length of the device) is shown in Fig. 50. At these densities, positive ions falling downhill from the center of the plasma, as shown in Fig. 50, are accelerated until they reach the saddle

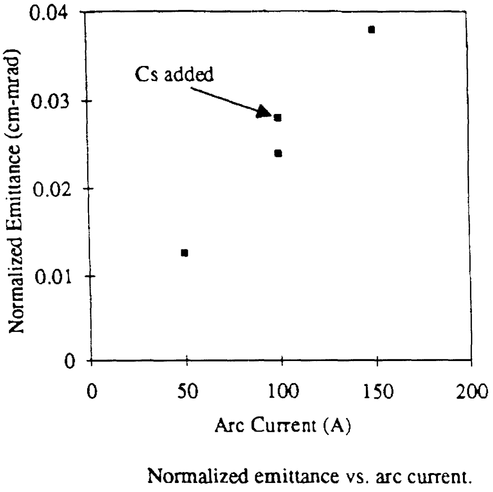

FIG. 49. Normalized emittance as a function of arc current. This indicates that the ion "temperature" depends on plasma density. Increasing ion-acoustic waves in the presheath is a likely explanation for this effect.

ridge. Then the positive ions are repelled by the accelerator fields and are attracted into the plasma electrode as shown. This is in contrast to the case for positive ions, shown in Fig. 44(a), some of which can get extracted. For volume-produced negative ions, the trajectories are shown in Fig. 50(b). Here, only a fraction of the negative ions produced are extracted, and the rest are attracted to the body of the plasma since they are repulsed by the plasma electrode.

2.2.4 Transverse Space-Charge Limits

When the extracted current depends on extraction voltage [see Fig. 9a of Ando (1994), or Fig. 9 of McAdams *et al.* (1987), or Fig. 45], which is usually the case at high currents, the beam is said to be transverse space-charge limited (TSCL). This is because the extraction voltage has very little effect on the negative-ion generation rate, and so the beam is limited by interception on one or more electrodes. This interception comes about as a result of poorly matched optics because of an oversupply of space charge (like charges repel—Coulomb's law). An example of experimental results is shown in Fig. 45, which illustrates this space-charge–limited flow. This is further substantiated by noting that the separate curves

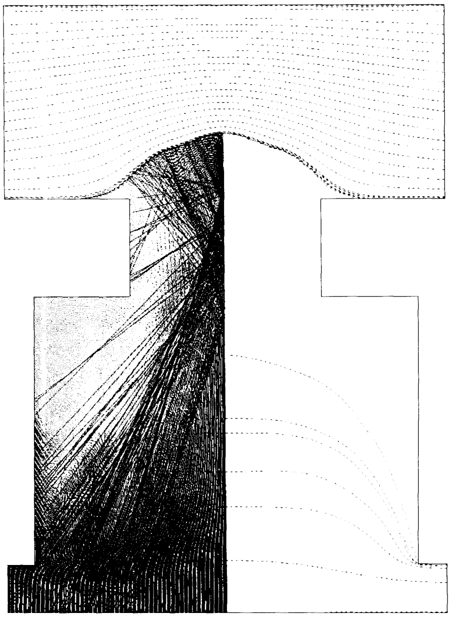

(a)

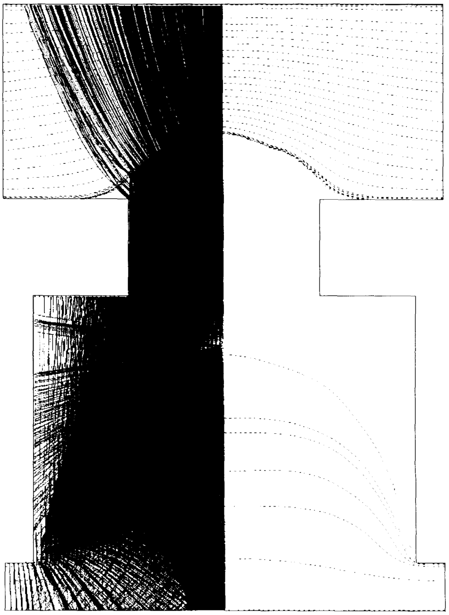

(b)

FIG. 50. (a) Same as Fig. 48(a), except that the plasma density is higher. Instabilities have been suppressed by space charge under relaxation. (b) Same as Fig. 48(b), except that the plasma density is higher.

in Fig. 45(b) actually scale by the Child–Langmuir proportionality [a consequence of dimensional analysis for positive ions (Whealton and Jaeger, 1977)], as proven by the coalescence of the curves in Fig. 45(b) into one curve, Fig. 45(c). The very nature of a transverse space-charge limit means that if the generation rate is increased (beyond any decrease in electron space charge), the beam current actually decreases. This is seen for example in Fig. 45(b), where at 200 V, the extracted H^- current for an arc current of 50 A is less than that for 30 A. The reason for this decrease is that the excess negative-ion beam generation not only gets intercepted by the electrodes but causes some of the formerly transmitted ions to be intercepted by the accelerator structure.

2.2.5 Effect of Cesium In volume sources the addition of cesium can cause an increase in the negative-ion current by up to a factor of 3 and a decrease in electrons extracted by a factor of as much as 100. One reason for the increase in negative-ion current extracted is that cesium converts positive ions into negative ions at a suitably coated surface. However, since the source typically operates at the space-charge limit, addition of more negative ions by surface production does not account for the increase in negative-ion beam current. Alternatively, the concomitant decrease in electrons extracted (frequently a factor of 100) may increase the partition of the space-charge limit occupied by the negative ions. If, for example, the electrons extracted were 100 times the ion current, not unusual for a negative-ion source operating without cesium, then the electrons could be taking up as much as two-thirds of the space-charge limit. The elimination of most of these electrons frees up two-thirds of the transverse space-charge limit, enabling the ion current to be 3 times higher, if one assumes that no emission limit exists at these higher currents.

An important question is how the presence of cesium so drastically decreases the number of extracted electrons. The effect of cesium on negative-ion output and electrons extracted varies from source to source, but the cases where the effect is greatest usually involve the application of magnetic fields and frequently such accoutrements as a collar electrode (see Sec. 1.3). What occurs is that the self-biasing of the plasma with respect to the extraction electrode can change with the presence of cesium from that which it would have without it. What determines the self-bias is the flux of positive and negative charges from the electrode. For a noncesiated electrode, in the simplest case, positive ions are drawn into the electrode and are balanced by the flux of the much more mobile electrons shown simplistically in Fig. 51(a). In this simple case, the positive charge flux is balanced by retarding the source-plasma electrons electrostatically. This retardation of electrons is done by a self-biasing of the plasma positive with respect to the electrode, as shown in Fig. 52. As an example for H^+ and a Boltzmann equilibrium electron distribution, a bias V_b of $3.6\,kT_e/e$, as shown in Fig. 52, will effect a balance between the electron and ion fluxes.

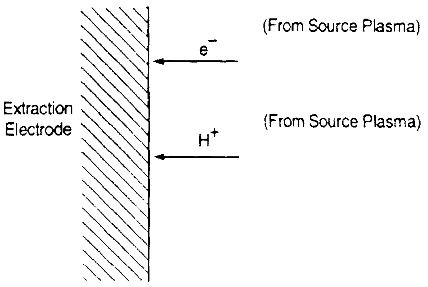

(a) Without Cesium

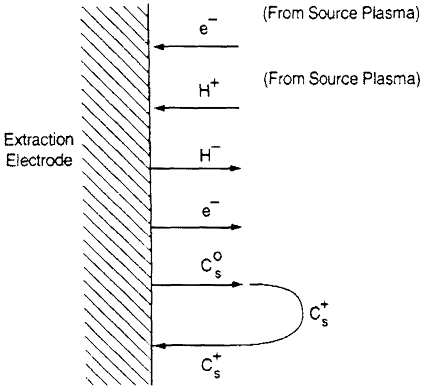

(b) With Cesium

FIG. 51. Charge balance in ion source (a) without and (b) with cesium.

A cesium coating, with its lower work

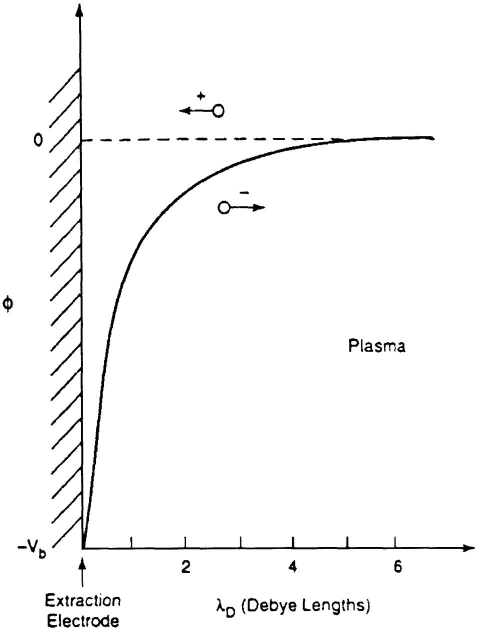

FIG. 52. Positive and negative charge flow near negatively biased electrode.

function, allows other processes to occur. One process is the conversion into negative ions of the positive ions intercepting the surface. Another process is the ejection of electrons from the surface. Finally, from neutral sputtered cesium (due to positive-ion bombardment) come positive ions, Cs^+, due to ionization by electrons and H^+; these ions return to the electrode. The Cs^+ has been observed to be a dominant species of positive ions (Ogasawara *et al.*, 1998). All three of these new processes, due to the cesium coating, cause an additional net flux of positive ions, as shown in Fig. 51(b). All of these additional processes tend to lower the magnitude of the negative wall bias so that more plasma electrons are absorbed by the extraction electrode.

To see what a profound effect this has on the electrons extracted, we need to have the transport configuration typified by Fig. 29. This is the simplest case without cesium, with only the positive ions and electrons shown. The electrons execute multiple bounces with the extraction collar electrode. The increased probability of absorption on each pass, due to decreased magnitude of bias, will dramatically decrease the number of electrons arriving at the extraction sheath. An illustration of the various additional surface mechanisms due to cesium is shown in Fig. 53.

Therefore, the reasons why a cesium coating can reduce the electrons extracted are twofold:

1. The self bias of the target plasma is diminished, allowing increased absorption of the electrons drifting in from the driver plasma; and
2. the additional flux of surface-produced negative ions partially goes back to the driver plasma, which reduces the electron flux drifting in from the driver plasma since fewer are needed for charge neutrality.

2.2.6 Increasing the Transverse Space-Charge Limit (TSCL) Now that it has been shown that cesium allows the negative ions to occupy virtually the entire transverse space-charge–limit budget, it remains to consider how the space-charge limit itself can be increased. The transverse space-charge limit is usually thought to depend on the geometry of the electrodes, the potentials applied, the beam current density, and the mass and charge of the ions. Actually, the transverse space-charge limit also depends on the shape of the extraction sheath, which by artificial means can be further manipulated. For example, the deliberate introduction of an additional positive charge into the pre-extraction sheath region would have the effect of changing the curvature of the sheath, effecting a *higher* space-charge limit (Whealton *et al.*, 1998). (Actually, this could be done also for positive-ion sources by the introduction of negative space charge in the presheath region.) Ionized sputtered cesium is just such a source of positive ions. There is an abundance of experimental evidence indicating the increase in space-charge limit for negative ions. The Dudnikov source, Sec. 1.2.1 (see Fig. 13), is perhaps the first source to use a collar with significant improvement (Allison 1980). Therefore, a cesium addition could increase the beam current beyond the transverse space-charge limit.

The Cs^+ space charge in the region just before extraction not only serves to increase the transverse space-charge limit by causing the extraction sheath to be more concave (Fig. 54 compared with 50), but can also reduce Bohm instabilities, through a reduction of the regions

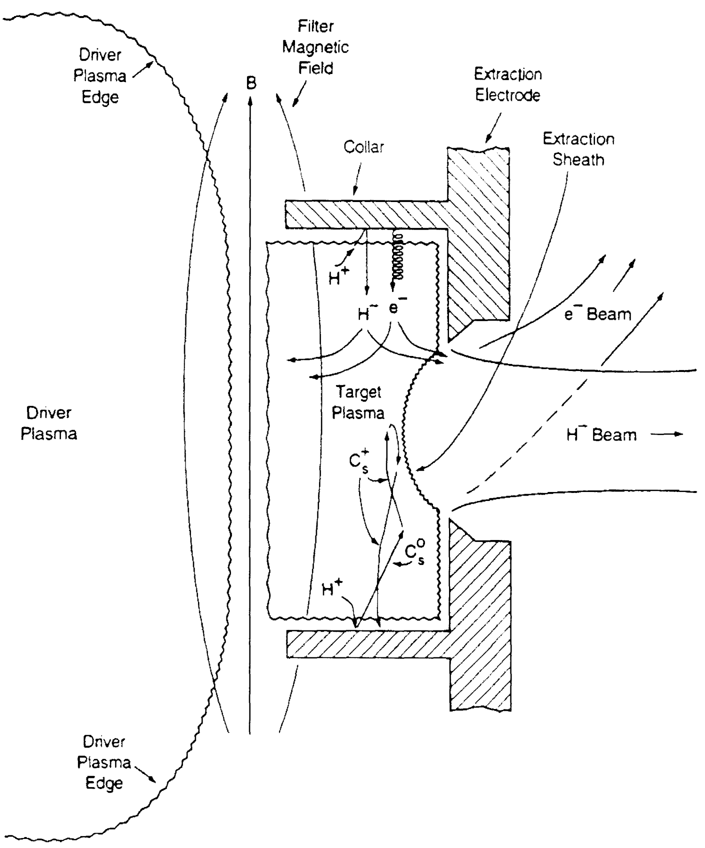
Driver Plasma Edge
Filter Magnetic Field
B
Extraction Electrode
Collar
Extraction Sheath
H^+
H^-
e^-
e^- Beam
Target Plasma
Driver Plasma
H^- Beam
C_s^+
C_s^0
H^+
Driver Plasma Edge

FIG. 53. Surface production of negative ions and production of Cs^+.

where there is an excess of negative space charge in the pre-extraction volume.

3. APPLICATIONS

Several applications require beams of H^-. They are in the areas of material sciences, semiconductor processing, high-energy physics, nuclear physics, energy technology, and weapons.

3.1 Beam-Storage Rings—Pulsed Spallation Neutron Sources

Pulsed spallation neutron sources require proton energies of 1 GeV or so and as high a current as possible. This is usually done by storing GeV protons in an accumulator ring. To reduce the emittance of the ring current, negative-ion injections are used to avoid the consequences of Liouville's theorem. The negative ions are produced in a single-aperture source and accelerated with an rf LINAC. For a review, see Alonso (1996).

3.2 High-Energy Neutral Beam for Fusion-Plasma Heating

Neutral beams (300 keV) of hydrogen, made by stripping an H^- beam, that pass through ambient magnetic fields become ionized (positive), which puts a positive potential in the end-plug plasmas. Positive ions escaping from the mirror device are repelled by these intraplasma fields, thus enhancing plasma confinement in the mirror field.

Neutral beams of high energy, 0.5–2 MeV, are required to heat, fuel, and provide a current driver for Tokamak fusion experiments. These beams need to be neutral to get past the intense magnetic fields at the tokamak's edge. The best way to produce significant H^0 currents is to start them as negative ions and accelerate them to the required energy, then pass the beam through a gas (or plasma or perhaps a laser beam), stripping off the excess electron and making them neutral. It is far easier to remove the electron than to add an electron to a positive ion at such energies.

3.3 High-Energy Neutral Beam for the Strategic Defense Initiative

In order to fly through space unaffected by the Earth's magnetic field, neutral beams are desired. These neutral beams need to be on the order of 10% of the speed of light, and so the best way they can be produced is by generation from H^- (or D^-). These negative ions are accelerated, and then the excess electron is stripped off, leaving a neutral beam. Of course, these neutral beams need to be steered in a specific direction without significant aberrations.

3.4 Ion-Projection Lithography

Advantages of using negative ions for ion projection lithography are that the loosely bound electrons fly off as secondary-emission electrons, preventing charging (see Fig. 55), a major limitation in reliability for small feature size (Gudharay *et al.*, 1994, 1996; Ishikawa, 1994, 1996).

4. CONCLUSIONS

Negative-ion sources during the last quarter century have advanced from microampere to ampere current production. This advance has been associated with many measurements; however, the improvement has been largely empirical. A deep understanding of the beam formation mechanism has yet to be demonstrated. The half dozen or so nonlinear processes simultaneously occurring have prevented a self-consistent synthesis.

The material presented and interpreted here summarizes the major components of our understanding. Using negative-ion current as a marker, six orders of magnitude improvement

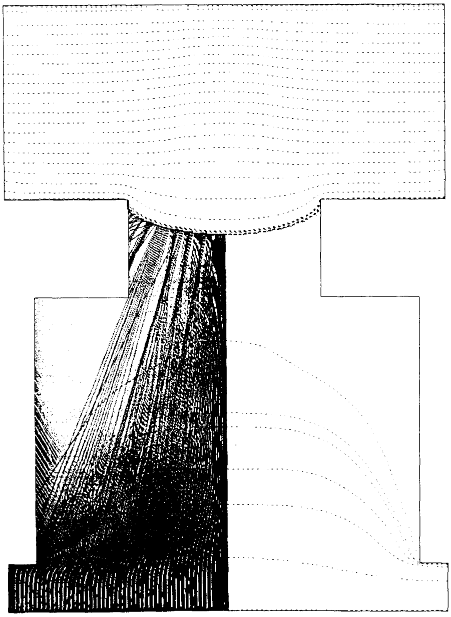
(a)

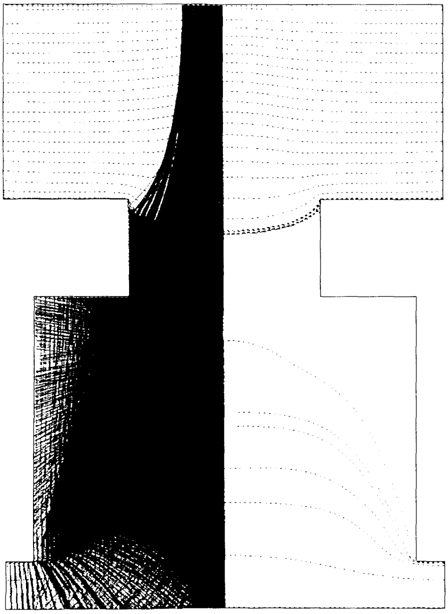
(b)

FIG. 54. (a) Same as Fig. 50(a), but positive space charge added in the presheath. (b) Same as Fig. 50(b), but positive space charge added in the presheath region.

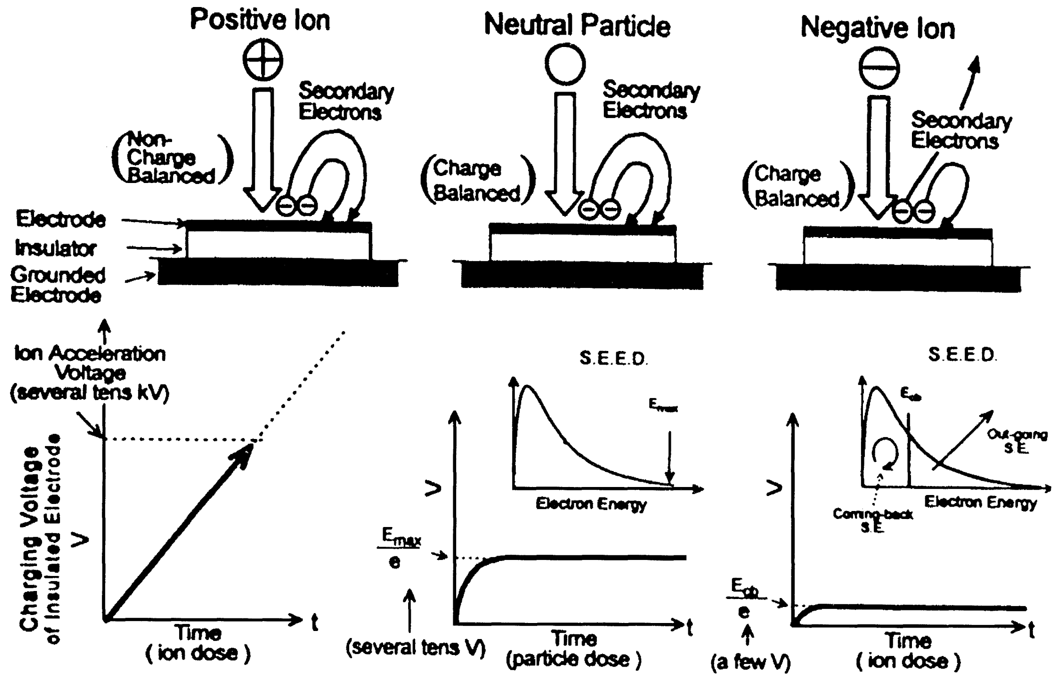
Positive Ion
Neutral Particle
Negative Ion
Secondary Electrons
Non-Charge Balanced
Charge Balanced
Electrode
Insulator
Grounded Electrode
Ion Acceleration Voltage (several tens kV)
Charging Voltage of Insulated Electrode
V
Time (ion dose)
t
S.E.E.D.
Electron Energy
E_max
E_max/e
(several tens V)
Time (particle dose)
E_ch/e
(a few V)
Out-going S.E.
Coming-back S.E.

FIG. 55. Schematic representation of particle trajectories and charging potentials for the case of positive-ion, neutral-particle, and negative-ion implantation.

has taken place, opening up applications otherwise unfeasible. We can expect that additional applications will keep pace with further improvements in ion sources.

ACKNOWLEDGMENTS

The authors would like to thank Jeanie Shover and Al Akerman for their help in preparing this article.

GLOSSARY

Brightness: The specific current of particles emitted by unit surface into unit solid angle.

Calutron: A powerful plasma ion source with hot cathodes, gas discharge with electron oscillation in a strong magnetic field, and transverse extraction of the ion beam, developed by E. O. Lawrence for isotope separation.

Cesium Catalysis: The significant enhancement of negative-ion formation in a gas discharge by the addition of a small admixture of cesium or other substances with low ionization potential, decreasing the electrode work function.

Child–Langmuir Law: The equation for space-charge–limited current density in a vacuum diode.

Electron Affinity: The binding energy of the additional electron in a negative ion.

Electron Cyclotron Resonance (ECR): Resonance absorption of electromagnetic radiation by electrons in a static magnetic field when the electromagnetic frequency is equal to the cyclotron frequency of the electrons. It is a most powerful method of heating electrons to a high temperature.

Emittance: A measure for the volume in phase space occupied by a beam, indicative of the reciprocal of the quality of the beam.

Ionization Potential: The smallest potential difference necessary for electron acceleration to produce atom or molecule ionization; equal to the binding energy of an electron in the atom or molecule.

Penning Discharge: A low-pressure gas discharge with oscillation of electrons along the anode tube in a magnetic field between two cathodes.

Perveance: Beam current divided by square root of the acceleration potential; used as a measure of the ratio of current to the transverse space-charge–limited current.

Secondary Emission: The emission of electrons or ions initiated by bombardment with electrons, ions, excited or fast atoms, *etc.*

Surface-Plasma Sources (SPS): Ion sources with negative-ion formation by the interaction of plasma particles with electrode surfaces in gas discharges.

Thermionic Emission: The emission of electrons or ions from a solid heated to a high temperature, taking place by the "evaporation" of the particles.

Thermionic-Emission Converter: A power supply for conversion of heat into electricity by emission of electrons from a cathode with a higher work function and collection of them by an anode with a lower work function.

Work Function: The energy difference between the vacuum energy level of an electron and the Fermi level in a solid: the evaporation energy of an electron.

Works Cited

Alessi, J. G. (1990), in: A. Hershcovitch (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 210, New York: American Institute of Physics, pp. 526–533.

Alessi, J. G., Prelec, K. (1991), in: *Proceedings of the 1991 Particle Accelerator Conference, Washington D. C.*, pp. 1913–1915.

Alessi, J. G., Prelec, K., McCafferty, D. (1994), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 387–396.

Alessi, J. G., Sluyters, T. (1980), *Rev. Sci. Instrum.* **51**, 1630–1635.

Allison, P. W. (1977), in: K. Prelec (Ed.), *Proceedings of the Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No. BNL-50727, 119–122.

Allison, P. W., Smith, H. V., Sherman, J. D. (1980), in: T. Sluyters (Ed.), *Proceedings of the Second International Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No. BNL-51304, pp. 171–177.

Alonso, J. R. (1996), *Rev. Sci. Instrum.* **67**, 1308–1313.

Anderson, O. A. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 473–488.

Ando, A. (1994), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 339–352.

Antipov, S. P., Elizarov, L. I., Martynov, M. I., Chesnokov, V. M. (1990), in: A. Hershcovitch (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 210, New York: American Institute of Physics, pp. 184–197.

Bacal, M., Leung, K-N. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 105–112.

Bacal, M., Nicolopoulou, E., Doucet, H. J. (1977), in: K. Prelec (Ed.), *Proceedings of the Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No. BNL-50727, pp. 26–34.

Bacal, M., Bruneteau, A. M., Doucet, H. J., Graham, W. G., Hamilton, G. W. (1980), in: T. Sluyters (Ed.), *Proceedings of the Second International Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No. BNL-51304, pp. 95–105.

Bacal, M., Hillion, F., Nachman, M., Steckelmacher, W. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 418–428.

Bacal, M., Bruneteau, A. M., Bruneteau, J., Devynck, P., Hillion, F. (1986), in: *Proceedings of the 2nd European Negative Ion Source Conference, Ecole Polytechnique*, pp. 17–24.

Bacal, M., Bruneteau, J, Devynda, P., Hillion, F. (1987), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 158, New York: American Institute of Physics, pp. 246–258.

Bacal, M., Bruneteau, J., Devynck P. (1988), *Rev. Sci. Instrum.* **59**, 2152.

Bacal, M., Berlemont, P., Bruneteau, J., Devynck, P., Konieczny, C., Leroy, R., Stern, R. A. (1990), in: A. Hershcovitch (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 210, New York: American Institute of Physics, pp. 489–503.

Barnett, C. F., Ray, J. A., Ricci, E., Wilker, M. I., McDaniel, E. W., Thomas, E. W., Gilbudy, H. B.

(1977), Oak Ridge National Laboratory Report No. ORNL-5206 (Vol. 1) p. A.4.3.

Barnett, C. F., Hunter, H. T., Kirkpatrick, M. I., Alvarez, I., Cisneros, C., Phaneuf, R. A. (1990), Oak Ridge National Laboratory Report No. ORNL-6086 (Vol. 1) pp. A-21.

Bashkeev, A. A., Dudnikov, V. G. (1990), in: A. Hershcovitch (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 210, New York: American Institute of Physics, pp. 329–339.

Becker, R., Leung, K. N., Kunkel, W. (1998), *Rev. Sci. Instrum.* **69**, 1107–1109.

Belchenko, Y. I., Dimov, G. I. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 363–375.

Belchenko, Y. I., Kupriyanov, A. S. (1990), in: A. Hershcovitch (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 210, New York: American Institute of Physics, pp. 198–213.

Belchenko, Y. I., Kupriyanov, A. S. (1994), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 255–270.

Belchenko, Y. I., Dimov, G. I., Dudnikov, V. G., Ivanov, A. A. (1973), *Dokl. Akad. Nauk. SSSR* **213**, 1283–1786.

Belchenko, Y. I., Dimov, G. I., Dudnikov, V. G. (1977), in: K. Prelec (Ed.), *Proceedings of the Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No. BNL-50727, pp. 79–96.

Bell, M. A., Whealton, J. H., Raridon, R. J., McGaffey, R. W., Wooten, D. E. (1985), *Bull. Am. Phys. Soc.* **30**, 1609.

Bruneteau, J. (1986), in: *Proceedings of the 2nd European Negative Ion Source Conference, Ecole Polytechnique*, pp. 81–86.

Budker, G. I., Dimov, G. I., Dudnikov., V. G. (1967), *Sov. At. Energ.* **22**, 441–449.

Chan, C. F., Leung, K.-N. (1993a), in: *Proceedings of the 1993 Particle Accelerator Conference, Washington, D. C.*, pp. 3160–3162.

Chan, C. F., Leung, K. N. (1994), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 762–769.

Cooper, W. S., Halbach K. (1974), in: C. Pezzoti (Ed.), *Proceedings of the 2nd Symposium on Ion Sources and Formation of Ion Beams*, Lawrence Berkeley Laboratory Report No. LBL-3399. II-1-1.

Cowan, R. G., Niemel, J., Campbell, R. L., Raridon, R. J., Whealton, J. H., Michaut, C., Bacal, M. (1994), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 621–633.

Dagenhart, W. K., Stirling, W. L., Haselton, H. H., Kelley, G. G., Kim, J., Tsai, C. C., Whealton, J. H. (1980), in: T. Sluyters (Ed.), *Proceedings of the Second International Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No. BNL-51304, pp. 217–224.

Dagenhart, W. K., Gardner, W. L., Stirling, W. L., Whealton, J. H. (1983), *Nucl. Technol./Fusion* **4**, pp. 1430–1442.

Dagenhart, W. K., Stirling, W. L., Barric, G. M., Barber, G. C., Ponte, N. S., Whealton, J. H. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 353–362.

Dagenhart, W. K., Tsai, C. C., Stirling, W. L., Ryan, P. M., Schechter, D. E., Whealton, J. H., Donaghy, J. J. (1987), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 158, New York: American Institute of Physics, pp. 366–377.

Derevyankin, G. E., Dudnikov, V. G. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 376–397.

Dudnikov, V. G. (1972), SV Autor Certificat., No. 411542.

Dudnikov, V. G. (1980), in: T. Sluyters (Ed.), *Proceedings of the Second International Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No. BNL-51304, pp. 137–144.

Dudnikov, V. G. (1993), *Rev. Sci. Instrum.* **63**, pp. 2660–2662.

Dudnikov, V. G., Derevyankin, G. E. (1994), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 239–254.

Dudnikov, V. G., Stirling, W. L., Whealton, J. H. (1997), in: *Proceedings of the 1997 Particle Accelerator Conference, Vancouver, B.C.*, pp. 2714–2716.

Ehlers, K. W., Leung, K. N. (1980), *Rev. Sci. Instrum.* **51**, 721–730.

Ehlers, K. W., Leung., K. N. (1981), *Appl. Phys. Lett.* **38**, 287–315.

Elizarov, L. I., Chesnokov, V. M. (1994), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 305–315.

Goretsky, V. P., Ryabtsev, A. V., Soloshenko, I. A., Tarasenko, A. F., Schedrin, A. I. (1996), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 384, New York: American Institute of Physics, pp. 142–152.

Greer, J. A., Seidl, M. V. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 220–238.

Gudharay, S. K., *et al.* (1994), *Rev. Sci. Instrum.* **65**, 1745–1748.

Gudharay, S. K., *et al.* (1996), *J. Vac. Sci. Technol.* **B14**, 3907–3716.

Gudharay, S. K., Allen, C. K., Reiser, M., Saadatmaud, K., Chang, C. R. (1994), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 672–685.

Holmes, A. J. T., Green, T. S. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 429–437.

Inoue, T., Hanada, M., Mizuno, M., Ohara, Y., Okumura, Y., Suzuki, Y., Tanaka, M., Watenabe, K. (1994), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 316–325.

Isenberg, J. D., Kwon, H. J., Seidl, M. (1994), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 38–47.

Ishikawa, J. (1994), *Rev. Sci. Instrum.* **65**, 1290–1292.

Ishikawa, J. (1996), *Rev. Sci. Instrum.* **67**, 1410–1415.

Kuroda, T., *et al.* (1996), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 380, New York: American Institute of Physics, pp. 201–213.

Kwan, J. W., *et al.* (1993), in: *Proceedings of the 1993 Particle Accelerator Conference, Washington, D. C.*, pp. 3169–3171.

Kwan, J. W., Gough, R. A., Leung, K. N., Perkins, L. T., Pickard, D. S., Vujic, J., Wu, L. K., Olivo, M., Einenkel, H. (1994), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 295–304.

Lee, Y., Gough, R. A., Leung, K. N., Perkins, L. T., Pickard, D. S., Vujic, J., Wu, L. K., Olivim, M., Einenkel, H. (1998), *Rev. Sci. Instrum.* **69**, 1023–1025.

Leung, K.-N. (1998), *Rev. Sci. Instrum.* **69**, 998–1002.

Leung, K.-N., Ehlers, K. W. (1980), in: T. Sluyters (Ed.), *Proceedings of the Second International Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No, BNL-51304, pp. 65–73.

Leung, K.-N., Ehlers, K. W. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 67–81.

Leung, K.-N., Ehlers, K. W., Bacal, M. (1983), *Rev. Sci. Instrum.* **54**, 56–59.

Leung, K.-N., Hauck, C. A., Kunkel, W. B., Walther, S. W. (1989), *Rev. Sci. Instrum.* **60**, 531–537.

Leung, K.-N., DeVries, G. J., Divergilio, W. F., Hamm, R. W., Hauck, C. W., Kunkel, W. B., McDonald, D. S., Williams, M. D. (1991), *Rev. Sci. Instrum.* **62**, 100–108.

Leung, K.-N., Bachman, D. A., McDonald, D. S. (1994), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 368–374.

Lietzke, A. F. K., Ehlers, W., Leung, K. N. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 344–352.

McAdams, R., Holmes, A. T. J., Nightingale, M. P. S., Lee, L. M., Hinton, M. D., Neumann, A. F., Green, T. S. (1987), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 158, New York: American Institute of Physics, pp. 298–308.

McAdams, R., King, R. F., Proudfoot, G., Holmes, A. T. J. (1994), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 353–367.

McGaffey, R. W., Meszaros, P. S., Whealton, J. H., Raridon, R. J., McCollough, D. H. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 533–543.

Maesser, A., Volk, K., Weber, M., Klein, H. (1996), *Rev. Sci. Instrum.*, **67**, 1054–1056.

Melmychuk, S. T., *et al.* (1991), *J. Vac. Sci. Technol.* **A9**, 1650–1655.

Mori, Y., Takagi, A., Ikegami, K., Fukumoto, S. (1987), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 158, New York: American Institute of Physics, pp. 378–383.

Ogasawara, M., Marishita, T., Miura, M., Shibayama, N., Hatayama, A. (1998), *Rev. Sci. Instrum.* **69**, 1132–1137.

Ohara, Y. (1998), *Rev. Sci. Instrum.* **69**, 908–913.

Okumura, Y., Horiike, H., Inone, T., Kurashima, T., Matsuda, S., Ohara, Y., Tanaka, S. (1987), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 158, New York: American Institute of Physics, pp. 309–318.

Okumura, Y., *et al.* (1996), *Rev. Sci. Instrum.* **67**, 1018–1020.

Peters, J. (1998), *Rev. Sci. Instrum.* **69**, 992–994.

Prelec, K. (1977a), in: K. Prelec (Ed.), *Proceedings of the Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No. BNL-50727, p. 111.

Prelec, K. (1977b), *Nucl. Instrum. Meth.* **144**, 413–422.

Prelec, K. (1980), in: T. Sluyters (Ed.), *Proceedings of the Second International Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No. BNL-51304, pp. 145–152.

Prelec, K. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 333–343.

Riz, D., Pamela, J. (1996), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 380, New York: American Institute of Physics, pp. 315–323.

Riz, D., Pamela, J. (1998), *Rev. Sci. Instrum.* **69**, 914–919.

Schmidt, C. W. (1980), in: T. Sluyters (Ed.), *Proceedings of the Second International Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No. BNL-51304, pp. 189–196.

Schmidt, C. W., Curtis, C. D. (1977), in: K. Prelec (Ed.), *Proceedings of the Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No. BNL-50727, pp. 123–128.

Schmidt, C. W., *et al.* (1979), *IEEE Trans. Nucl. Sci.*, **NS-26**, 4120.

Seidl, M., Pargellis, A. N., Greer, J. (1980), in: T. Sluyters (Ed.), *Proceedings of the Second International Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No. BNL-51304, pp. 111–118.

Seidl, M., Ciu, H. L., Isenberg, J. D., Kwan, H. J., Lee, B. S. (1997), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 25–37.

Seidl, M., *et al.* (1982), *Phys. Rev. B*, **26**, 1.

Semashko, N. N., Kuzltsov, V. V., Krylov, A. I., Firsov, P. S. (1987), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 158, New York: American Institute of Physics, pp. 334–345.

Simonin, A., *et al.* (1996), *Rev. Sci. Instrum.* **67**, 1102–1107.

Sluyters, T. (1974), Lawrence Berkeley Laboratory Report No. LBL-3399, p. II-1-1.

Smith, H. V., Allison, P., Sherman, J. D. (1980), in: T. Sluyters (Ed.), *Proceedings of the Second International Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No. BNL-51304, pp. 178–183.

Smith, H. V., Allison, P., Sherman, J. D. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 458–462.

Smith, H. V., Allison, P., Sherman, J. D. (1985), *IEEE Trans. Nucl. Sci.* **NS32**, 1797–1800.

Smith, H. V., Allison, P., Geisik, C., Schmitt, D. R. (1993), in: *Proceedings of the 1993 Particle Accelerator Conference, Washington, D. C.*, pp. 3172–3174.

Smith, H. V., Allison, P., Geisik, C., Orbesen, S. D., Schmitt, R. D., Schneider, J. D., Stelzer, J. E. (1994), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 287, New York: American Institute of Physics, pp. 282–294.

Soloshenko, I. A., Shchedrin, A. I., Ryabtev, A. V. (1998), in: C. Jacquot (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 439, New York: American Institute of Physics, pp. 139–157.

Stevens, R., private communication, 1987.

Stirling, W. L., *et al.* (1979), *Rev. Sci. Instrum.* **50**, 523.

Stirling, W. L., Dagenhart, W. K., Whealton, J. H., Donaghy, J. J. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 450–457.

Takagi, A., Mori, Y., Igarashi, Z., Ikegami, K., Kubota, C., Fukumoto, S. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 520–523.

Takeiri, Y., *et al.* (1996), *Rev. Sci. Instrum.* **67**, 1021–1023.

Tanaka, M., Ose, Y., Koizumi, M., Yamashita, Y., Kawakami, H., Takeiri, K., Kaneko, Oka, Y., Tsumori, K., Osakabe, M. (1998), *Rev. Sci. Instrum.* **69**, 1144–1146.

Watanabe, K., *et al.* (1996), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 380, New York: American Institute of Physics, pp. 309–319.

Whealton, J. H. (1981), *Nucl. Instrum. Methods.* **189**, 55–77.

Whealton, J. H. (1985), *Bull. Am. Phys. Soc.* **30**, 1513.

Whealton, J. H. (1990), in: A. Hershcovitch (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 210, New York: American Institute of Physics, pp. 539–556.

Whealton, J. H. (1987), in: J. G. Alessi (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 158, New York: American Institute of Physics, pp. 482–506.

Whealton, J. H., Jaeger, E. F. (1977), *Rev. Sci. Instrum.* **48**, 829.

Whealton, J. H., Jaeger, E. F., Whitson, J. C. (1978), *J. Comput. Phys.* **27**, 32.

Whealton, J. H., Raridon, R. J., McGaffey, R. W., McCollough, D. H., Stirling, W. L., Dagenhart, W. K. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 524–532.

Whealton, J. H., *et al.* (1985), *IAEA Negative Ion Conference, Grenoble.*

Whealton, J. H., McGaffey R. W., Meszaros, P. S. (1986), *J. Comput. Phys.* **63**, 20–33.

Whealton, J. H., Bell, M. A., Raridon, R. J., Rothe, K. E., Ryan, P. M. (1988), *J. Appl. Phys.* **64**, 6210–6224.

Whealton, J. H., Meszaros, P. S., Raridon, R. J., Rothe, K. E., Bacal, M., Bruneteau, J., Devynda, P. (1989), *Rev. Phys. Appl.* **24**, 945–951.

Whealton, J. H., Olsen, D. K., Raridon, R. J. (1998), *Rev. Sci. Instrum.* **69**, 1103–1105.

Whitson, J. C., Smith, J., Whealton, J. H. (1978), *J. Comput. Phys.* **28**, 408–415.

Wooten, J. W., *et al.* (1979), Oak Ridge National Laboratory Report No. ORNL/TM-6740.

Wooten, J. W., Whealton J. H., McCollough, D. H. (1981), *J. Comput. Phys.* **43**, 95.

York, R. L., *et al.* (1983), *IEEE Trans. Nucl. Sci.*, **NS30**, 2705–2707.

York, R. L., Stevens, R. (1984), in: K. Prelec (Ed.), *Production and Neutralization of Negative Ions and Beams*, AIP Conference Proceedings No. 111, New York: American Institute of Physics, pp. 410–417.

York, R. L., Tupa, D., Swenson, D. R., Damjanovich, R. (1993), in: *Proceedings of the 1993 Particle Accelerator Conference, Washington, D. C.*, pp. 3175–3177.

Further Reading

Alessi, J. G. (Ed.) (1987), *Production and Neutralization of Negative Ions and Beams (1986)*, AIP Conference Proceedings No. 158, New York: American Institute of Physics.

Alessi, J. G. (Ed.) (1993), *Production and Neutralization of Negative Ions and Beams (1992)*, AIP Conference Proceedings No. 287, New York: American Institute of Physics.

Bacal, M., *et al.* (1996), *Rev. Sci. Instrum.* **67**, 1138–1143.

Belchenko, Y. I. (1996), *Rev. Sci. Instrum.* **67**, 1108–1110.

Forrester, A. T. (1988), *Large Ion Beams*, New York: Wiley, Chap. 10.

Hershcovitch, A. (Ed.) (1990), *Production and Neutralization of Negative Ions and Beams (1989)*, AIP Conference Proceedings No. 210, New York: American Institute of Physics.

Jacquot, C. (Ed.) (1998), *Production and Neutralization of Negative Ions and Beams (1997)*, AIP Conference Proceedings No. 430, New York: American Institute of Physics.

Pezzoti, C. (Ed.) (1974), *Proceedings of the 2nd Symposium on Ion Sources and Formation of Ion Beams*, Lawrence Berkeley Laboratory Report No. LBL-3399.

Prelec, K. (Ed.) (1977), *Proceedings of the Symposium on the Production and Neutralization of Negative Ions and Beams*, Brookhaven National Laboratory Report No. BNL-50727.

Prelec, K. (Ed.) (1984), *Production and Neutralization of Negative Ions and Beams (1983)*, AIP Conference Proceedings No. 111, New York: American Institute of Physics.

Prelec, K. (Ed.) (1996), *Production and Neutralization of Negative Ions and Beams (1995)*, AIP Conference Proceedings No. 380, New York: American Institute of Physics.

Sluyters, T. (Ed.) (1971), *Proceedings of the Symposium on Ion Sources and Formation of Ion Beams*, Brookhaven National Laboratory Report No. BNL-50310.

Sluyters, T. (Ed.) (1980), *Proceedings of the Second International Symposium on the Production and Neutralization of Negative Hydrogen Ions and Beams*, Brookhaven National Laboratory Report No. BNL-51304.

Whealton, J. H., Raridon, R. J., Bell, M. A., Rothe, K. E., Wooten, D. W., McGaffey, Wooten, J. W., Whitson, J. C. (1988), *J. Appl. Phys.* **64**, 6210–6221.

SAFETY AND HEALTH IN THE PHYSICAL SCIENCE LABORATORY*

Daniel T. Crane, *Salt Lake Technical Center, Occupational Safety and Health Administration, United States Department of Labor, Salt Lake City, Utah, U.S.A.*

Introduction 189

1. **Responsibility for Safety** 190
1.1 Management Commitment 190
1.1.1 Occupational Safety and Health Act 190
1.1.2 Responsibility to Inform........ 190
1.1.3 Management Responsibility 191
1.1.4 Empowerment of Safety and Health Professional 191
1.1.5 Safety and Health Professional Responsibilities 191
1.2 Responsibility of Employees 191

2. **Planning for Safety and Health** 191
2.1 Hazard and Risk 191
2.1.1 Hazard 191
2.1.2 Risk 192
2.1.3 The Presidential/Congressional Commission on Risk Assessment and Risk Management 192
2.2 Written Safety and Health Plan 192
2.2.1 Chemical Hygiene Plan......... 192
2.2.2 Emergency Response Plan...... 192
2.3 Monitoring Safety and Health 193
2.3.1 Monitoring and Sampling 193
2.3.2 Medical Examinations for Routine Exposures 193
2.3.3 Medical Examinations After an Accident 193
2.4 Novel Hazards 193
2.4.1 Substitution 193
2.4.2 Isolation 193
2.4.3 Ventilation 194
2.4.4 Personal Protective Equipment.. 194

3. **Recognition and Remediation of Safety and Health Hazards in the Physical Laboratory** 194
3.1 Planning for Safety 194
3.1.1 Exposure Routes 194
3.1.2 Looking for Hazards 194
3.2 Common Hazards Found in the Physical Science Laboratory.... 195
3.2.1 Ergonomic Hazards 195
3.2.2 Chemical Exposure 195
3.2.3 Ignition Sources 195
3.2.4 Cryogens 195
3.2.5 Physical Hazards 196
3.2.6 Inhalation 196
3.2.7 Radiation and Radioactive Materials 196
3.2.8 Noise 196
3.2.9 Biological and Nontraditional Hazards 196
3.2.10 Heat Stress or Exposure to Temperature Extremes......... 196

4. **What To Do When an Accident Happens** 197
4.1 Emergency Response 197
4.1.1 Responsibility of the Safety and Health Professional 197
4.1.2 Injuries 197
4.2 Notification Requirements 197
4.2.1 Federal Government Notification 197
4.2.2 Local Government Notification 197
4.3 Evaluation and Remediation ... 198

5. **Conclusion** 198

Glossary 198
Works Cited 198
Further Reading 199

INTRODUCTION

Most accidents can be foreseen and avoided. Accidents usually occur either because a known hazard went unabated or because a new, unforseen hazard is encountered. Understanding the nature of common hazards

---

*The opinions and statements of the author are completely his own and do not establish policy of the United States of America or necessarily establish interpretations of federal regulations. Any mention of a specific product name does not imply endorsement or preferred use.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

in the physical science laboratory is key to managing safety and health. The purpose of this article is to examine the role of the safety and health professional in activities and processes that may exist or occur in physical science laboratories. These professionals are known variously as industrial hygienists, safety officers, or chemical hygiene officers, or by other suitable titles. They may be certified by a professional organization.

1. RESPONSIBILITY FOR SAFETY

Safety is the responsibility of every person in a facility, beginning with top management and extending even to visitors. It is therefore important that potential hazards be communicated to all involved individuals at a facility. The control of the processes within physical science laboratories lies with the management of those facilities. While the laboratory workers perform the work directly, management ultimately has the basic responsibility to ensure that the workers have a workplace free of recognized hazards [OSHA, 1970, Section 5(a)].

1.1 Management Commitment

No program of safety and health can be effective without active support and impetus from the highest management officials. This support must be sincere and adequate. Safety and health is not the sole responsibility of the safety and health professional. All levels of management must be committed to the safety and health of all the employees whether they are researchers, graduate students, or custodial staff. The reason for this is clear. The response of employees in the work environment will follow immediately from what supervisory and management staff emphasize. Top laboratory management must actively pursue safety and health in a visible manner such that middle and first-line managers and supervisors understand that worker safety cannot be sacrificed for production or research. No influence from independent researchers must interfere with the safe operation of the facility. The safety and health professional is usually charged with the direct responsibility to provide information to laboratory personnel so that they may be aware of potential hazards and workable remediations.

1.1.1 Occupational Safety and Health Act Safety and health in the workplace is an international concern. Many nations have instituted particular regulations to limit exposure to health and physical hazards. Efforts in the United States of America are illustrative, in general, of international activities. Although the specific regulations will vary by nation, the concepts used by safety and health professionals are not tied to specific regulations. In 1970, in the United States, the Congress passed the Occupational Safety and Health Act, which states that every employer has the responsibility to provide a workplace free of recognized health or safety hazard [OSHA, 1970, Section 5(b)]. Other countries have established similar requirements (Kortsha, 1991). This article uses the United States regulations to illustrate particular concepts used by safety and health professionals.

1.1.2 Responsibility to Inform It is an important function of management, through the safety and health professional, to ensure that employees understand the hazards of the processes and materials that are present in the laboratory. The Occupational Safety and Health Administration (OSHA) has established guidelines for employee right to know (OSHA, 1997, Subpart Z, Sections 1200 and 1450). Specifically, this standard requires a number of things in the laboratory. Labels on containers of hazardous chemicals may not be removed or defaced. Material Safety Data Sheets (MSDS) received with incoming shipments of hazardous chemicals must be maintained by the laboratory and be readily accessible to laboratory personnel. Laboratory personnel must be appraised of the hazards of the chemicals and processes in the workplace through information and training at the time of their initial work assignments and whenever a new hazard is introduced into the work area. Employees must be made aware of and trained about the measures they can take to protect themselves from these safety and health hazards. This would include such measures as appropriate work practices, emergency procedures, and personal protective equipment. The safety and health officer is generally charged with making certain that such training occurs.

1.1.3 Management Responsibility Laboratory management must have an active involvement with safety. An effective means to establish and maintain a working safety and health program is to provide an empowered safety office with safety and health professionals who understand the particular issues of the physical science laboratory. The most effective safety programs are administered by joint labor and management personnel. Small operations may have a single employee serving part-time in this capacity, while in large operations the safety office or officer may have literally hundreds of employees devoted to maintaining safety in the extensive operations of the laboratory system, such as within the United States Department of Energy (Lawrence Livermore National Laboratory, 1997).

1.1.4 Empowerment of Safety and Health Professional The safety and health professional must be empowered to make administrative decisions and take immediate action regarding safety. This is why management participation is essential. He or she has the direct responsibility (with management support) to determine the content and provide the training required in each work area. Employee input is essential for determining the hazards that are present in the various work areas. Top management may be unaware of some of the hazards and must therefore be responsive to employee suggestions.

1.1.5 Safety and Health Professional Responsibilities The safety and health professional has the responsibility to provide initial training for new employees and for transferred employees before they are exposed to potential hazards. The safety office or officer has the responsibility to maintain records of all training offered and to whom it was given. The safety office or officer has the responsibility, with management, to conduct inspections to identify and correct hazards or see that the inspections are done. Effective internal safety inspections are made with a team composed of both management and labor, including employees from within the area being inspected and others who have no direct involvement so that they can take an unbiased view of the operations being inspected.

1.2 Responsibility of Employees

While management has the primary responsibility to provide a workplace free of recognized hazard, it remains with the employees of the laboratory to follow safe work practices and to work with management to abate unsafe conditions. It is not appropriate for employees purposely to put themselves at unreasonable risk to meet institutional deadlines or demands. In the course of work, employees at the work-detail level frequently are the first to become aware of new hazards. They are certainly the ones who are most often injured by not following accepted work practices. Employees must take a personal interest in using existing and appropriate personal protective equipment and engineering controls and following established safety guidelines and regulations. Employees must follow guidelines such as the wearing of safety glasses, protective clothing, or other personal protective equipment including respirators, hearing protection, face shields, *etc.*

2. PLANNING FOR SAFETY AND HEALTH

One of the best ways to avoid accident and injury is to plan ahead for foreseeable hazards. Management, along with the safety and health professional, has the first responsibility to assess the presence of potential hazards.

2.1 Hazard and Risk

To plan effectively for safety, it is necessary first to understand the difference between hazard and risk.

2.1.1 Hazard A hazard exists if a particular substance, process or situation has the potential to cause harmful effects. Given such a broad definition, there are few materials or processes which do not have some potential to cause harm. It is necessary to establish a system for determining the likelihood that a particular hazard might cause a harmful effect—that is, to determine risk (Newsom, 1996).

2.1.2 Risk Risk is defined as the probability that a substance or situation will produce harm under specified conditions. Risk is a combination of two factors: the probability that an adverse event will occur (such as a specific disease or type of injury) and the consequences of the adverse event. Risk encompasses impacts on public health and on the environment, and arises from exposure and hazard. Risk does not exist if exposure to a harmful substance or situation does not or will not occur. Hazard is determined by whether a particular substance or situation has the potential to cause harmful effects.

2.1.3 The Presidential/Congressional Commission on Risk Assessment and Risk Management This commission has published effective guidelines for determining the extent of risk. It is beyond the scope of this article to explore fully the intricacies of risk management. However, the Commission has established a six-stage framework for risk management (Newsom, 1996):

1. Define the problem and put it in context.
2. Analyze the risks associated with the problem in context.
3. Examine options for addressing the risks.
4. Make decisions about which options to implement.
5. Take actions to implement the decisions.
6. Conduct an evaluation of the actions results.

2.2 Written Safety and Health Plan

Safety and health professionals aid employers by providing written safety and health plans that are required if there is a potential for release of hazardous materials (OSHA, 1997, Subpart Z, Section 120). Such a plan assures that there is a coherent and controlled approach to the safety and health in the laboratory. A written plan provides a framework against which to build effective prevention and response.

2.2.1 Chemical Hygiene Plan The safety and health professional is instrumental in writing and maintaining the chemical hygiene plan, which is a comprehensive document that covers all aspects of safe laboratory operation. It sets forth procedures, equipment, laboratory equipment, personal protective equipment, and work practices that are capable of protecting employees from the safety and health hazards presented by hazardous chemicals and processes used in that particular workplace (OSHA, 1997, Subpart Z, Section 1450). It prescribes the proper storage and handling of chemicals, radionuclides, and physical agents. It prescribes the appropriate handling of lasers, magnetic fields, and the specific processes and hazards present in the laboratory. It is beyond the scope of this article to prescribe specific plans. However, the plan must be capable of protecting employees from safety and health hazards associated with hazardous chemicals and processes in that laboratory and capable of keeping exposures below the limits specified by OSHA or other governing national standards organization. The Chemical Hygiene Plan must also be readily available to employees, employee representatives, and, upon request, government regulators. The Chemical Hygiene Plan includes specific measures that the employer must take to ensure laboratory employee protection. These include written standard operating procedures, special engineering controls, personal protective equipment and its use, provisions for employee information and training, required medical monitoring, and emergency response procedures as well as specific instructions on how to revise the plan as needed. The safety and health professional is very familiar with all aspects of the Chemical Hygiene Plan and keeps it current and effective.

2.2.2 Emergency Response Plan The Emergency Response Plan establishes the procedures to follow in case of emergency. The plan outlines the steps to be taken following an accident. It will include who is to be notified in the laboratory organization, the local and state governments, and the national government along with required notification deadlines. It will include any specific guidelines that must be given to company or local emergency personnel in order for them to carry out their emergency response. This might include certain areas where not to fight a fire because of explosion hazard, or areas in which radiation is a concern. These instructions must be shared with the expected emergency providers before an emergency occurs to avoid any unnecessary confusion. Knowing ahead of time what hazards are present can allow professional emergency personnel to develop appropriate action plans and procedures.

2.3 Monitoring Safety and Health

It is necessary to know when and if laboratory personnel or others are affected by operations in the laboratory. For this reason, laboratory management must monitor potential exposures to hazards and then maintain for each employee an accurate record of any measurements taken to monitor employee exposures and any medical consultation and examinations, including tests or written opinions.

2.3.1 Monitoring and Sampling Generally, the safety and health official conducts or oversees monitoring or sampling to determine whether or not an employee is exposed to hazards. There are various devices for this purpose, ranging from filters to sorbent tubes to passive monitors. In some cases, there are direct-reading instruments to give an estimate of the hazard. Filters, tubes, and passive devices usually need to be analyzed by an industrial hygiene laboratory, while direct-reading instruments give an immediate indication of the exposure. Other devices are used to monitor exposures to radiation or noise. Again, some devices give an immediate reading while others such as film badges must be sent to a laboratory for interpretation. It is the responsibility to the safety and health officer to understand the proper use of these devices and techniques.

2.3.2 Medical Examinations for Routine Exposures Whenever an employee develops signs or symptoms associated with a hazardous chemical or process to which the employee may have been exposed in the laboratory, the employee must be provided an opportunity to receive an appropriate medical examination. If exposure monitoring reveals an exposure level routinely above the action level [or, in the absence of an action level, the permissible exposure limit (PEL)] for an OSHA-regulated substance for which there are exposure monitoring and medical surveillance requirements, medical surveillance must be provided for the affected employee.

2.3.3 Medical Examinations After an Accident Whenever an event takes place in the work area such as a spill, leak, explosion, or other occurrence resulting in the likelihood of a hazardous exposure, affected employees must be provided an opportunity for medical consultation.

2.4 Novel Hazards

The nature of a laboratory is generally to investigate processes that are not routine, and may be new, novel, and undocumented outside the immediate work environment. Such situations must be examined for the potential to cause harm to any exposed personnel. The safety and health professional helps researchers control hazards through the systematic use of industrial hygiene controls. Engineering controls are the first line of protection followed by good work practices, administrative controls, and, where no other mode of protection is feasible, personal protective devices. The principles of engineering control are substitution, isolation, and ventilation (Soule, 1991).

2.4.1 Substitution If a health or safety problem is presented by the use of a specific material, wavelength, or process, substitution of a more benign material, wavelength, or process should be investigated as a remediation of a hazard. For example, enclosures might be made of plastic or safety glass rather than regular glass, or cleaning chemicals might be chosen that are safer. An externally vented hood might be substituted for a stand-alone unit. Care must be taken when substituting one chemical or process for another so as not to cause further hazard.

2.4.2 Isolation The second principle of engineering control is isolation. This is frequently some barrier such as a blast shield, or may consist of remotely conducting the operation in a separate room. Also, simply putting distance between the operator and the operation can be considered isolation. In addition, compartmentalization such as is done for tank farms is a form of isolation that limits the involvement of an entire facility in the event of an accident. It may be possible to separate different procedures, isolating the most dangerous from other, more benign ones.

2.4.3 Ventilation The third principle of engineering control is ventilation. Ventilation provides control of the environment by varying the airflow. The airflow can be used to control the temperature of the space, remove a contaminant, or dilute the concentration of the contaminant to below permissible exposure levels. Control of ventilation requires a detailed knowledge of air flow and the materials or processes to be controlled. Such information is beyond the scope of this article. Information geared to industrial hygiene can be found in standard industrial hygiene references (Soule, 1991).

2.4.4 Personal Protective Equipment Personal protective equipment includes safety glasses, safety shoes, special clothing, respiratory protection, and any other device or system that is designed to be worn or used by an individual. Personal protective equipment should never be considered an acceptable alternative to a well-designed engineering approach to safety except where such is shown to be either unfeasible or more unsafe than the use of personal protective equipment. Certain pieces of personal protective equipment such as safety glasses should be used regardless of the presence of engineering controls. Where respirators are used, a respiratory protection program is required to be in place. Certain operations require specialized personal protective equipment. For example, when working with lasers, special eye protection is required. The safety and health professional must be knowledgeable as to the proper use of this type of personal protective equipment.

3. RECOGNITION AND REMEDIATION OF SAFETY AND HEALTH HAZARDS IN THE PHYSICAL SCIENCE LABORATORY

Awareness of recognized hazards is the primary tool used in recognition of the potential hazards. Management along with the safety and health professional, as well as involved workers, must review the processes used or expected to be used in the laboratory. Forethought and planning to perform the work safely knowing recognized hazards will prevent many accidents.

3.1 Planning for Safety

Looking for hazards in the physical science laboratory begins with the planning stages of a project or process by looking for recognized hazards in the processes to be used. For example, if gas cylinders are to be used, the cylinders must be properly secured and vented. If a process involves ionizing radiation, the proper safeguards, such as shielding, must be established. Laboratory personnel must continue to look for hazards after the project begins and throughout the project to remediate any significant hazard which may appear. Looking for hazards must not be limited to items which are recognized, codified into regulation, or found on a list.

3.1.1 Exposure Routes The routes of exposure to chemical and physical hazards are dermal, respiration, and ingestion.

Dermal includes any skin or eye exposure which would result in

1. direct absorption of a chemical into the skin or eye;
2. radiation exposure of the skin; or
3. temperature extreme.

Exposure through respiration is by

1. inspiration of a chemical or radiation hazard;
2. extremely hot or cold air;
3. inhalation of toxic gases;
4. exposure to atmospheres deficient in oxygen;
5. in hyperbaric conditions involving too much oxygen; or
6. the absence of air or exposure to vacuum.

Ingestion hazards are from anything that might be taken by mouth intentionally or unintentionally, by transmission from hand to mouth, or otherwise transmitted to the mouth.

3.1.2 Looking for Hazards Safety personnel, along with representatives from the work area being inspected, should look for any process, employee action, chemical storage, electrical wiring, virtually anything which might cause an accident. This inspection should not be limited to recognized hazards. Rather, the safety and health professional should look for any process or situation which *may* present a hazard. Immediately dangerous situations should be remediated as rapidly as possible. Once the team has identified potential hazards, those for which there are specific rules or regulations should be remediated. New or novel potential hazards should be evaluated for risk and remediated as appropriate.

3.2 Common Hazards Found in the Physical Science Laboratory

There are a number of common hazards to be found in the physical science laboratory. Most accidents occur as a result of slips, trips, and falls (OSHA, 1997, Subpart E, Section 38). Many other hazards are a result of environmental exposures such as extreme heat or cold (Mutchler, 1991; OSHA, 1997, Subpart D, Section 22). Confined spaces may present a hazard in breathing air quality. These spaces may be too small to contain sufficient air for the duration of the task, or a concentration of explosible gases may easily be reached when welding gases or other generated gases are being used. Certain spaces will have insufficient oxygen for breathing, concentrated toxic gases, or vacuum. (OSHA, 1997, Subpart J, Section 146).

3.2.1 Ergonomic Hazards Ergonomic hazards are generally the result of three factors: force, frequency, and posture. In the physical science laboratory, there are many tasks involving one or more of these recognized stressors that may result in repetitive motion injury, back injury, or other upper extremity cumulative trauma disorder. Back injuries can usually be limited by implementing training programs that teach employees the proper techniques for lifting and by limiting the amount of weight that individual employees are required to lift (Waters, 1993). Often, the weight involved is very little, but the frequency of repetition may reach very high levels, as in manual data input on a keyboard. Examples of this sort of stress may be seen in keyboard entry personnel and other employees in related tasks. Posture is a strong factor in cumulative trauma disorder especially when developing new tasks. Employees must not be allowed to let their dedication to the project overshadow good posture, or other good work habits. Thought must be given to service access when designing test enclosures. Proper access allows the employee to address the task with good posture. Repetitive tasks should be evaluated for good ergonomic practice, and employees must be trained in the principles of biomechanics and how good work practice will prevent ergonomic injury.

3.2.2 Chemical Exposure Much of the health hazard in the physical science laboratory comes from exposure to chemicals. Common problems arise from improper use and storage of chemicals. Incompatible chemicals should never be stored in the same cabinets. Before using any chemical in the laboratory, the employee should review the material safety data sheet, which must be available for all employees. Particular attention should be paid to the physical properties of the chemical, to ensure that the chemical will be used within safe environmental limits and stored safely and that the proper safety and spill clean-up materials are on hand. Employees handling chemicals must wear appropriate personal protective gear as necessary. This may include single or double gloves, safety glasses, respirators, or special clothing prescribed by regulation or by the producer of such chemicals.

3.2.3 Ignition Sources Generally, as part of the chemical safety requirements, the laboratory must assure that equipment used in the presence of chemicals that may form an explosible atmosphere are certified and maintained as explosion resistant (OSHA, 1997, Subpart H, Section 106 and Subpart S, Sections 304, 305 and 307).

3.2.4 Cryogens Cryogenic liquids or solids at very low temperatures pose several hazards. The first and most obvious is the low temperature itself, which is capable of freezing flesh. Appropriate personal protective equipment must be used when working with cryogens. Transfer pipes, hoses, and vessels often provide a location for condensation of air or other gas in contact with the vessel surface.

This condensate may be enriched with oxygen. As a result, such pipes and vessels should not be in contact with or suspended above asphalt surfaces or other combustible surface materials (OSHA, 1997, Subpart Z, Section 1051).

3.2.5 Physical Hazards Physical hazards may include loose objects on high ledges or shelves, improperly guarded machines, or even loose tools near large magnets. Physical hazards may also include unguarded pulleys, swinging arms, low ceilings, or stacks of unused equipment or boxes that might fall on a worker. Physical hazards might also include loose items that may fall or be thrown during an earthquake (in areas prone to earthquakes). Physical hazards might also include objects that are ejected from experimental devices, or a whole variety of electrical hazards. Some of these may include ungrounded wiring, loose cables on the floor, unshielded cables in wet areas, and too many plugs for a given circuit rating.

3.2.6 Inhalation When looking for inhalation hazards, it must be remembered that many substances when inhaled are toxic but may not necessarily present a hazard otherwise. Any dust that can be inhaled must be controlled to acceptable levels, preferably by the use of engineering controls such as fume hoods, local exhaust, barriers, or remote operation (OSHA, 1997, Subpart Z, Section 1000). Protection from exposure to gases to prevent toxic exposure or suffocation must be provided. Again, engineering controls are preferred to personal protective devices such as respirators. Survey the materials being used and check for any applicable standards such as OSHA standards, HSE standards, ISO standards, EPA standards where appropriate, DOE standards, DOT regulations where appropriate, or special industry standards. In certain emergency situations, as specified in OSHA standards, respirators may be used when

1. they meet use criteria;
2. there is a respirator program in place to assure safe use of the respirator; and
3. the safe capacity of the particular respirator meets or exceeds the expected concentration of contaminant.

Other criteria may apply and the laboratory should consult current OSHA requirements on this issue (OSHA, 1997, Subpart I, Sections 132–136).

3.2.7 Radiation and Radioactive Materials Most generally, radiation is divided into two classes, ionizing and nonionizing. Ionizing radiation includes particulate, beta, alpha, and gamma radiation. Nonionizing radiation generally includes ultraviolet (UV), infrared (IR), radio frequency (rf), extremely low-frequency (ELF), and sonic energy. The distinction between the two classes is generally made by considering whether or not a particular type of radiation can cause ionization in human skin. This energy level is often taken to be 10 eV. Further specific information on appropriate levels of exposure will be found in several of the references given in the Further Reading section. See also Wilkening (1991); OSHA (1997), Subpart G, Section 97.

3.2.8 Noise Noise exposures may come from a variety of sources including equipment, tools, radios, and explosions. Care must be taken to assure that exposures to noise are not excessive. Ear damage due to noise exposure is generally permanent. OSHA has established permissible exposure levels based upon the A-weighted scale of a standard noise meter (OSHA, 1997, Subpart G, Section 95).

3.2.9 Biological and Nontraditional Hazards Physical science laboratories are becoming more interdisciplinary in nature, with equipment being developed for medical diagnostics and other nontraditional markets. It is therefore necessary to be aware of and follow requirements for safe handling of biological materials. It is possible that biological samples such as microorganisms, including viruses, may be present in the physical laboratory. Laboratory animals or pathogenic tissue samples may be found in some laboratories. OSHA regulates the handling of biological materials primarily through the bloodborne pathogens regulations, which are aimed at HIV and HIB (OSHA, 1997, Subpart Z, Section 1030).

3.2.10 Heat Stress or Exposure to Temperature Extremes Precautions must be taken to ensure that employees do not suffer either hyperthermia (high temperature) or hypothermia (low temperature). Hyperthermia is generally considered to exist when the body core temperature exceeds 38 °C and there is no underlying physical condition to cause the high temperature. Hypothermia is generally considered to exist when the core body temperature falls below about 32 °C.

4. WHAT TO DO WHEN AN ACCIDENT HAPPENS

The definition of an accident includes incidents involving spills of hazardous materials, bomb threats, civil disturbance, explosion, fire, natural disaster such as an earthquake, terrorism, or any other act or happenstance that could put laboratory personnel or the community at risk. Specific requirements vary by locality and nation. Many community governments and most national governments can provide contacts and specific information necessary to meet legal requirements in their respective venues.

4.1 Emergency Response

It is beyond the scope of this article to describe procedures for every possible accident. A laboratory chemical response plan such as that required by OSHA will establish the levels of response based upon the relative severity of the accident. It includes whom to call for particular events, what to do, and whom to inform. All employees in a facility must be trained in the contents of this plan and how to respond if an accident occurs.

4.1.1 Responsibility of the Safety and Health Professional Depending upon the size of the facility, the safety and health professional and one or more employees may be specially trained to respond to medical emergencies, chemical spills, radiation releases, or other hazards. When such employees or teams are used, they must be appropriately trained according to applicable local and federal regulations.

4.1.2 Injuries When an injury occurs, immediate notification of in-laboratory first aid responders or local emergency medical providers is imperative. For other emergencies such as fire or other large incidents, an automatic call system may be used to contact emergency responders. Such procedures will be set out clearly in the written safety and health plan for the laboratory. Generally, a good relationship should be established between the laboratory and the local emergency responders. Where there are security areas or special hazards, it is important for the local emergency responders and laboratory management to work out agreeable procedures for emergency response. The safety and health official is usually the point of contact with local service providers in setting up such procedures.

4.2 Notification Requirements

Every incident will be reported to immediate supervisors, while more severe incidents will be required to be reported to local, state, and national agencies.

4.2.1 National Government Notification In the United States, OSHA requires employers to maintain a log of injury and illness that is occupationally related. This is kept on the OSHA 200 log (OSHA, 1997, Subpart F, Section 67). Moreover, OSHA must be notified within 8 h of any fatal accident or any work-related incident requiring the hospitalization of three or more employees (OSHA, 1997, Subpart Z, Section 8). The Environmental Protection Agency requires notification of spills through the Resource Conservation and Recovery Act (RCRA) (EPA, 1976a), Toxic Substances Control Act (TSCA) (EPA, 1976b), Comprehensive Environmental Response Compensation and Liability Act (CERCLA) (EPA, 1977), and Superfund Amendments and Recovery Act (SARA) (EPA, 1986). This response is required within 24 h of the incident to the closest EPA office. Other nations may require differing levels of reporting. The safety and health professional will be familiar with all requirements for reporting.

4.2.2 Local Government Notification Many local governments have regulations requiring notification in the event of an accident. As it is a good practice to involve local service providers in safety planning, these same local providers should prove to be valuable sources for specific local reporting requirements. Along with these providers the local air-quality board or its equivalent, the waste-management authority, the EPA, and the OSHA office or their equivalents are excellent sources of specific and current information.

4.3 Evaluation and Remediation

Once an incident has occurred, it is important to evaluate the causes and ramifications of the incident to avert further incidents. Management, along with the safety team and the principal investigators, should evaluate all pertinent information to establish appropriate remediations in whatever form necessary to assure safe working conditions for laboratory personnel, visitors, and the community in which the laboratory is located.

5. CONCLUSION

Safety and health in the physical science laboratory is the responsibility of everyone in the facility. The safety and health professional acts as an integral part of the laboratory to help each employee to understand his or her role in managing safety in the laboratory. There may be specific requirements to be followed by the laboratory imposed by national or local government as well as by contracting authorities. No element of a sound safety plan can be abrogated by the pressure of project completion deadlines or cavalier attitudes of investigators. It is the responsibility of management to assure that laboratory employees and invited investigators adhere to established procedures and that where new processes are being developed, appropriate planning and care is taken to evaluate potential hazards and risk for those processes. Because of the intricacy of the processes and governmental requirements, it is suggested that competent professional safety and health consultants be retained either on an as-needed basis or on staff. In addition, all the regulatory agencies provide help for specific processes and can be very helpful in setting up and maintaining effective safety and health programs.

GLOSSARY

Accident: Any unexpected happening that interrupts the work sequence or process and that may result in injury, illness, or property damage to the extent that it causes loss.

Ergonomics: The science of fitting the job to the worker in a manner so as to lessen the risk of injury from operations involving repetitious, forceful or awkward postures.

Hazard: A situation or process that has the potential to cause harm.

MSDS: Material safety data sheet. A document that communicates the physical and chemical hazards of chemicals to the employee using or handling the chemical. The MSDS contains detailed information about handling, storage and use of a chemical and how to control unexpected events.

OSHA: Occupational Safety and Health Administration.

PEL: Permissible exposure limit. An exposure level to a chemical or other hazard, established by OSHA, above which employees may not be exposed.

PPE: Personal protective equipment. Items such as safety glasses, gloves, respirators, or similar apparel that are used by individuals as protection from hazard.

Risk: The likelihood that a particular hazard will result in an accident.

Works Cited

EPA (Environmental Protection Agency) (1976a), *Resource Conservation and Recovery Act* (RCRA), 42 USC 321 *et seq.*, Washington: GPO.

EPA (Environmental Protection Agency) (1976b), *Toxic Substances Control Act* (TSCA), 15 USC 2601 *et seq.*, Washington: GPO.

EPA (Environmental Protection Agency) (1977), *Comprehensive Environmental Response, Compensation, and Liability Act* (CERCLA), 42 USC 9601 *et seq.*, Washington: GPO.

EPA (Environmental Protection Agency) (1986), *Superfund Amendments and Reauthorization Act* (SARA), 42 USC 9601 *et seq.*, Washington: GPO.

Kortsha, G. X. (1991) in: G. D. Clayton, F. E. Clayton (Eds.), *Patty's Industrial Hygiene and Toxicology, General Principles*, Vol. 1, Part A, New York: John Wiley and Son, Chap. 3.

Lawrence Livermore National Laboratory (1997), *The Environment, Safety, and Health Program at the Lawrence Livermore National Laboratory*, Livermore, CA: www.llnl.es_and_h/esh_prog.

Mutchler, J. E. (1991) in: G. D. Clayton, F. E. Clayton (Eds.), *Patty's Industrial Hygiene and Toxicology*, Vol. 1, Part A, New York: Wiley and Sons, p. 763 *et seq.*

Newsom, S. (Ed.) (1996), *Risk Assessment and Risk Management in Regulatory Decision Making*, The Presidential/Congressional Commission on Risk Assessment and Risk Management, Vols. 1, 2, Washington: GPO.

OSHA (Occupational Safety and Health Administration) (1970), *OSHA Act of 1970*, amended on November 5, 1990, Washington: GPO.

OSHA (Occupational Safety and Health Administration) (1997), *Code of Federal Regulations*, Title 29, Washington: Office of the Federal Register, National Archives and Records Administration.

Soule, R. D. (1991) in: G. D. Clayton, F. E. Clayton (Eds.), *Patty's Industrial Hygiene and Toxicology*, Vol. 1, Part B, New York: Wiley and Sons, p. 137 *et seq.*

Waters, T., Putz-Anderson, V., Garg, A., Fine, L. J. (1993), *The Revised National Institute for Occupational Safety and Health Lifting Equation*, London: Taylor and Francis.

Wilkening, G. M. (1991) in: G. D. Clayton, F. E. Clayton (Eds.), *Patty's Industrial Hygiene and Toxicology*, Vol. 1, Part B, New York: Wiley and Sons, p. 599 *et seq.*

Further Reading

American Conference of Governmental Industrial Hygienists (ACGIH) (1998), *Threshold Limit Values for Chemical Substances and Physical Agents, Biological Exposure Indices*, Cincinnati, OH: ACGIH.

Clayton, G. D., Clayton, F. E. (Eds.), 1991, *Patty's Industrial Hygiene and Toxicology*, New York: Wiley and Sons.

Earley, M. W. (Ed.) (1996), *National Electrical Code Handbook*, Quincy, MA: National Fire Protection Association.

Furr, A. K. (Ed.) (1990), *Handbook of Laboratory Safety*, 2nd ed., Boca Raton: Chemical Rubber Company.

Newsom, S. (Ed.) (1996), *Risk Assessment and Risk Management in Regulatory Decision Making*, The Presidential/Congressional Commission on Risk Assessment and Risk Management, Vols. 1 and 2, Washington: GPO.

Occupational Safety and Health Administration, *OSHA Computerized Information System*, CD-ROM, 729-013-00000-5, Washington: GPO.

Occupational Safety and Health Administration, *OSHA Computerized Information System*, Internet Web site, Washington, D.C. *http://www.osha.gov.*

Waters, T. R., *et al.* (1993), *The Revised NIOSH Lifting Equation*, London: Taylor and Francis.

SILICIDES

Jeffrey P. Gambino, *IBM Microelectronics, Hopewell Junction, New York, U.S.A.*

Evan G. Colgan, *IBM Research, Yorktown Heights, New York, U.S.A.*

Introduction 201
1. **Properties** 202
1.1 Structure 202
1.1.1 Electronic Structure 202
1.1.1.1 Bulk Electronic Structure 202
1.1.1.2 Interfacial Electronic Structure 203
1.1.2 Atomic Structure 204
1.1.2.1 Crystal Structure 204
1.1.2.2 Interfacial Atomic Structure: Epitaxial Silicide–Silicon Interface 205
1.1.2.3 Phonons 206
1.2 Reactions and Stability 207
1.2.1 Metal–Si Reactions 207
1.2.1.1 Silicide Phases 207
1.2.1.2 Kinetics 209
1.2.1.3 Results on Metal–Si Reactions 210
1.2.2 Metal Reactions with Si-Based Compounds 211
1.2.2.1 Metal–SiGe Reactions 212
1.2.2.2 Metal–SiO_2 Reactions 212
1.2.2.3 Metal–Si_3N_4 Reactions 213
1.2.2.4 Metal–SiC Reactions 214
1.2.3 Thermal Stability 214
1.2.3.1 Melting Point 214
1.2.3.2 Stability of Thin Films 215
1.2.4 Chemical Stability 217
1.2.4.1 Oxidation 217
1.2.4.2 Silicide Reactions with Dopants (As, P, B) 218
1.2.4.3 Chemical Reactivity 220
1.3 Properties 220
1.3.1 Mechanical Properties 220
1.3.2 Optical Properties 223
1.3.3 Electrical Properties 225
1.3.3.1 Resistivity 225
1.3.3.2 Ohmic, Rectifying, and Gate Contacts 226
1.3.4 Magnetic Properties 227
2. **Fabrication** 228
2.1 Salicide Process 228
2.2 Polycide Process 230
2.3 Physical Vapor Deposition 231
2.4 Chemical Vapor Deposition 232
2.5 Ion Implantation 232
2.6 Melt Growth 233
2.7 Sintering 233
2.8 Plasma Spray Deposition 234
3. **Applications** 234
3.1 Microelectronics 234
3.1.1 MOS Technology 234
3.1.2 Bipolar Technology 236
3.1.3 Infrared Detectors 236
3.1.4 Permeable-Base Transistors 238
3.2 Structural Materials 238
3.3 Oxidation-Resistant Coatings 239
3.4 Thermoelectric Materials 240
3.5 Magnetic Materials 240
3.6 Optical Materials 240
4. **Summary** 240
Acknowledgement 241
Glossary 241
Works Cited 242
Further Reading 247

INTRODUCTION

Silicon forms a wide variety of compounds with metals; these compounds are called silicides. Silicides have a number of useful properties including high thermal stability, high mechanical hardness, and low resistivity. Silicides are used in a wide variety of applications. In microelectronics, silicides are used as Ohmic contacts, Schottky contacts, and interconnects. In these applications, the high thermal stability of the silicides and the low resistivity provide an advantage compared with other conductors, such as Al, which has much lower thermal stability, or polycrystalline silicon, which has much higher resistivity. Silicides are often used as furnace elements, on account of their thermal stability in oxidizing ambients. Because of their high strength, silicides are also being considered for use as high-temperature structural materials, in aircraft engines for example. For these applications, silicides have an advantage of higher thermal stability and lower density compared with Ni-based super-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

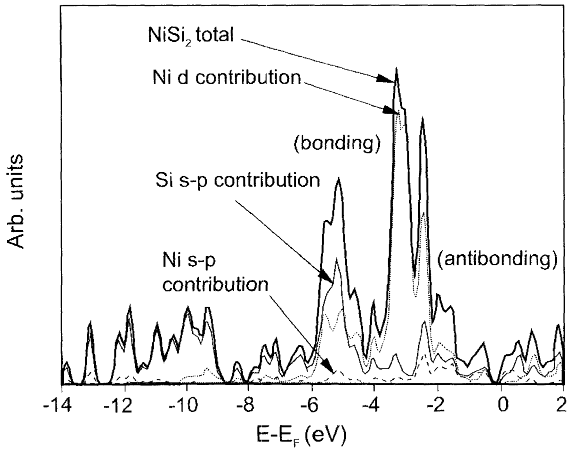

FIG. 1. Theoretical density of states and partial density of states for $NiSi_2$ (Bisi and Calandra, 1981).

alloys, and higher ductility and lower cost compared with Si-based ceramics. Other applications where silicides are being considered include thermoelectric devices, magnets and superconductors.

This article surveys the properties, processing, and applications of transition-metal silicides. The first section focuses on the properties of silicides, including the electronic and atomic structure, reactions and stability, and the mechanical, optical, electrical, and magnetic properties of these materials. The second section describes processes for forming silicides, such as the salicide process, the polycide process, physical vapor deposition, chemical vapor deposition, ion implantation, melt growth, sintering, and plasma-spray deposition. The final section discusses applications of silicides, including microelectronics, structural materials, coatings, thermoelectric materials, magnetic materials, and optical materials. Silicides formed from alkali metals, alkaline-earth metals, and rare-earth metals (Samsanov and Vinitskii, 1980) are not discussed in detail here, because there are very few potential applications for these materials at present.

1. PROPERTIES

1.1 Structure

1.1.1 Electronic Structure

1.1.1.1 Bulk Electronic Structure Many of the useful properties of silicides, such as high melting point and high mechanical hardness, arise from the formation of strong covalent bonds. Despite these strong covalent bonds, in most silicides there is a continuous density of states across the Fermi level, which provides high conductivity. Knowledge of the electronic structure of silicides is also useful for understanding properties such as Schottky-barrier heights (see Sec. 1.1.1.2) and chemical bonds and compound stability (see Sec. 1.1.2.1).

The main factor in determining the electronic structure of transition-metal silicides is the formation of hybrid orbitals between d states in the transition metal and 3p states in Si, resulting in strongly covalent metal d–Si p bonds (Lange, 1996; Bisi and Calandra, 1981; Speier *et al.*, 1989; Mattheiss, 1992; Reader *et al.*, 1992). These bonds produce bonding and antibonding states above and below the Fermi level (Fig. 1). In addition, there are some nonbonding d states with energies close to those of the pure metals, but of reduced width as a result of the greater separation between metal atoms in the silicides compared with the pure metal. The Si s states are only weakly involved in chemical bonding and form a relatively broad band several electronvolts below the Fermi level. The sequence of these levels with increasing energy is as follows: Si s, Si p/metal d bonding, metal d nonbonding, and Si p/metal d antibonding. The Fermi level is generally between the nonbonding and the antibonding states; hence the nonbonding states are usually occupied while the antibonding states are usually empty (Derrien, 1995).

Some trends in the band structure of silicides are observed with increasing atomic

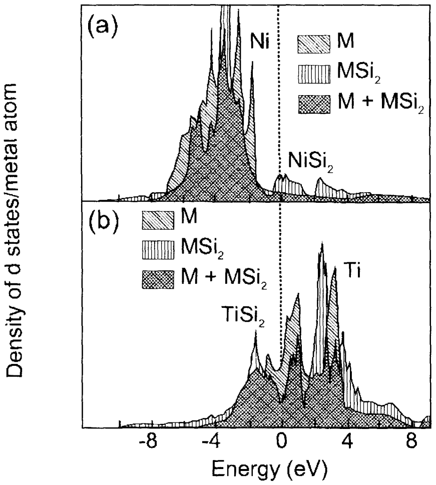

FIG. 2. Comparison of the calculated metal d-state densities in the pure metal and the disilicides for (a) Ni and (b) Ti. The pure metal bands have been shifted so that the centroids of the d band match those of the disilicides. (Reprinted from Weijs *et al.*, 1990, with permission from Springer-Verlag. Copyright 1990.)

number, moving through a series in the periodic table (Derrien, 1995; Reader *et al.*, 1992). First, the metal d states move through the Fermi level with increasing atomic number. In addition, the shape of the d states varies within a series, starting as a broad band of mostly empty states and finishing as a narrow band of mostly filled states. These trends are apparent when one compares the band structures of $NiSi_2$ and $TiSi_2$ (Weijs *et al.*, 1990; Reader *et al.*, 1992) (Fig. 2). The reduction in band width of the nonbonding metal d states compared with the elemental metal is greater for $NiSi_2$ than for $TiSi_2$, because of a stronger interaction with the Si p states in the case of $TiSi_2$. In addition, compared with $TiSi_2$, the Ni d character in the band structure of $NiSi_2$ is relatively weak in the bonding–antibonding regions above and below the metal d bands. The difference is due to the smaller radial extent of the Ni 3d wave function compared with that in Ti, which results in less mixing of the Si p and metal d states.

In most of the transition-metal silicides, there is a continuous density of states across the Fermi level, resulting in a metallic conduction. However, some silicides are semiconductors (see Sec. 1.3.3.1), with the minimum in the density of states at the Fermi level opening up into an energy gap. For a given metal, the density of states at the Fermi level increases as one goes from the Si-rich phases to the metal-rich phases, as a result of an increasingly strong interaction between metal d states (Speier *et al.*, 1989).

1.1.1.2 Interfacial Electronic Structure

One of the major applications for silicides is as contacts to semiconductors (see Sec. 3.1). The electronic structure of interfaces is also of interest for determining the adhesion strength of metal–silicide interfaces in structural materials (Hong *et al.*, 1993).

The electrical properties of the contacts are determined by the electrical barrier that forms at the metal–silicon interface. The height and width of the barrier determine the ease of electron transport across the interface, and hence determine whether the contact is Ohmic or rectifying, and whether it has high or low resistance. The barrier height depends on the work-function difference between the metal and semiconductor, the interface state density, the atomic structure at the interface, and the doping level in the semiconductor.

The mechanism of Schottky-barrier formation in covalent semiconductors is still not well understood, despite years of study. The barrier height ϕ_b is determined by the requirement that in equilibrium, the Fermi levels of the metal and semiconductor must be equal. This is achieved by creation of an electric dipole at the interface. In the absence of interface states, the charge at the interface is related to the difference in Fermi levels between the metal and the semiconductor, and is given by

$$\Phi_b = \varphi_m - \chi_s, \qquad (1)$$

where φ_m is the metal work function and χ_s is the electron affinity of the semiconductor. This model works fairly well for ionic semiconductors, where the interface state densities are low, and the barrier height does in fact show a strong dependence on the electronegativity of the metal (which is related to the work function of the metal; see Louie *et al.*, 1977; Schluter, 1977; Kurtin *et al.*, 1969). However, for covalent semiconductors, there is much less dependence of the barrier height on the electronegativity of the metal, as a result of the presence of interface states that pin the

Fermi level (Cowley and Sze, 1965; Heine, 1965). For covalent semiconductors, an extra term is added to Eq. (1), corresponding to the electrostatic interface dipole, that is determined by the electron distribution at the interface (Flores and Ortega, 1992). In fact, interface states associated with the silicide have been postulated from recent band-structure calculations of the silicide–Si interface (Fujitani and Asano, 1994). Some of these interface states have energies within the Si band gap and spatially extend a few atomic layers into the Si.

Recently it has been shown that Schottky-barrier heights depend on the local atomic structure of the interface (Tung, 1992). For example, the barrier height of epitaxial $NiSi_2$ on *n*-type Si can range from 0.40 eV [on (100) Si] to 0.79 eV [for type B $NiSi_2$ on (111) Si] (Tung, 1993). (For the type A interface, the silicide has the same orientation as the substrate, whereas for the type B interface, the silicide also has a ⟨111⟩ orientation, but is rotated by 180° about the ⟨111⟩ surface normal with respect to the substrate.) Fujitani and Asano (1994) have proposed that the atomic-structure dependence of the barrier height is due to bond bending at the interface, though more studies are required to determine if this hypothesis is correct.

Most of the Schottky-barrier heights on Si reported in the literature have been measured on polycrystalline silicides (Table 1). However, the different orientations in the polycrystalline silicide are likely to have different barrier heights, so that the contacts are electrically nonhomogeneous. The Schottky-barrier heights measured from polycrystalline silicides will be determined by the silicide grains with the lowest barrier heights (Tung, 1992).

The Schottky-barrier heights of silicides have been characterized on a number of compound semiconductors (Liu *et al.*, 1997; Shih *et al.*, 1988, Yokoyama *et al.*, 1986; Porter *et al.*, 1995a,b,c) including SiC, GaAs and GaN (Table 2). GaN and SiC have a greater degree of ionic bonding than Si or GaAs; hence, the barrier heights are expected to show a greater dependence on the type of metal or silicide.

Table 1. Schottky-barrier heights for various silicides on *n*-type Si (Tung, 1993; Sullivan *et al.*, 1993).

Silicide	Silicon substrate	Schottky barrier (eV)
IrSi (polycrystalline)	(100)	0.90
PtSi (polycrystalline)	(100)	0.86
$OsSi_2$ (polycrystalline)	(100)	0.85
Pt_2Si (polycrystalline)	(100)	0.77–0.85
MnSi (polycrystalline)	(100)	0.76
Pd_2Si (polycrystalline)	(100)	0.72
$ReSi_2$ (polycrystalline)	(100)	0.68
NiSi (polycrystalline)	(100)	0.65
$NiSi_2$ (type B, epitaxial)	(111)	0.79
$NiSi_2$ (type A, epitaxial)	(111)	0.65
$NiSi_2$ (epitaxial)	(100)	0.40
$CoSi_2$ (polycrystalline)	(100)	0.65
$CoSi_2$ (type B, epitaxial)	(111)	0.69
WSi_2 (polycrystalline)	(100)	0.65
$MoSi_2$ (polycrystalline)	(100)	0.64
RhSi (polycrystalline)	(100)	0.63
VSi_2 (polycrystalline)	(100)	0.63
FeSi (polycrystalline)	(100)	0.62
$TiSi_2$ (polycrystalline)	(100)	0.60
$CrSi_2$ (polycrystalline)	(100)	0.60
$TaSi_2$ (polycrystalline)	(100)	0.59
$NbSi_2$ (polycrystalline)	(100)	0.59
$ZrSi_2$ (polycrystalline)	(100)	0.55
HfSi (polycrystalline)	(100)	0.53
YSi_2 (polycrystalline)	(100)	0.39
$GdSi_2$ (polycrystalline)	(100)	0.37

1.1.2 Atomic Structure

1.1.2.1 Crystal Structure The crystallographic structures of transition-metal silicides are varied and complex (Fig. 3). Some structures form over significant regions; an MSi compound with the FeSi prototype is formed for nine different transition metals. Typically, an identical or related structure is found for silicides of the same composition and group (Murarka, 1983), though this is not without

Table 2. Schottky-barrier heights for various silicides on n-type compound semiconductors. In all cases, the silicide is polycrystalline.

Silicide	Substrate	Schottky barrier (eV)	Reference
Ti_5Si_3	α-SiC (1000)	0.90	Porter *et al.*, 1995b
CoSi	α-SiC (1000)	1.05	Porter *et al.*, 1995a
Pt_2Si	α-SiC (1000)	1.26	Porter *et al.*, 1995c
$WSi_{0.6}$	GaAs	0.70	Shih *et al.*, 1988
$MoSi_{0.7}$	GaAs	0.70	Yokoyama *et al.*, 1986
PtSi	GaN	0.87	Liu *et al.*, 1997

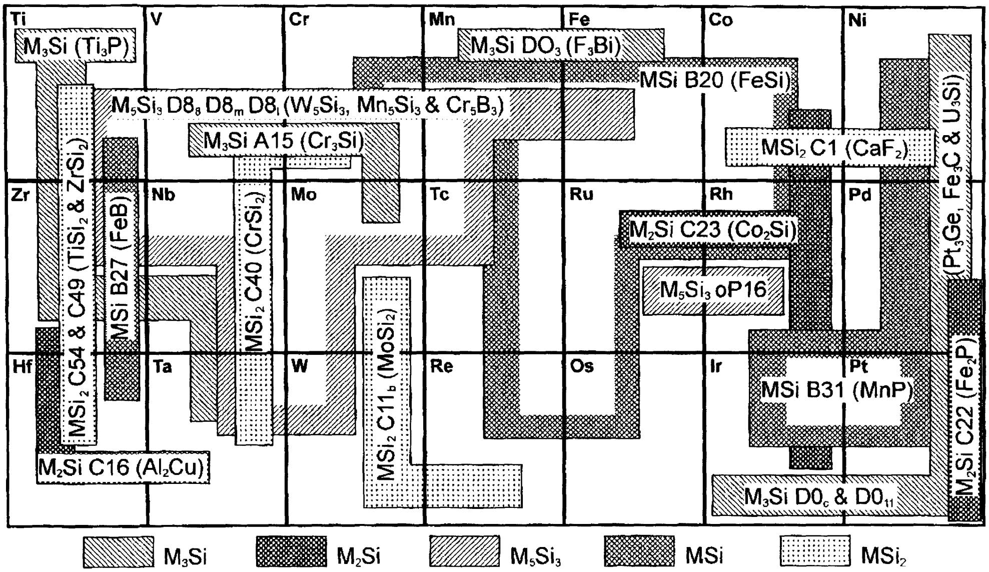
Ti
V
Cr
Mn
Fe
Co
Ni
Zr
Nb
Mo
Tc
Ru
Rh
Pd
Hf
Ta
W
Re
Os
Ir
Pt
M₃Si (Ti₃P)
M₃Si DO₃ (F₃Bi)
MSi B20 (FeSi)
M₅Si₃ D8₈ D8ₘ D8ₗ (W₅Si₃, Mn₅Si₃ & Cr₅B₃)
M₃Si A15 (Cr₃Si)
MSi₂ C1 (CaF₂)
MSi₂ C54 & C49 (TiSi₂ & ZrSi₂)
MSi B27 (FeB)
MSi₂ C40 (CrSi₂)
M₂Si C23 (Co₂Si)
M₅Si₃ oP16
(Pt₃Ge, Fe₃C & U₃Si)
MSi₂ C11ᵦ (MoSi₂)
MSi B31 (MnP)
M₂Si C22 (Fe₂P)
M₂Si C16 (Al₂Cu)
M₃Si DO꜀ & DO₁₁
M₃Si
M₂Si
M₅Si₃
MSi
MSi₂

FIG. 3. Fields of structures for transition-metal silicides (from Goldschmidt, 1967; Massalski, 1990).

exception (for example, V_3Si and $CrSi_2$; see Fig. 3 and Table 3). Further details on crystal structure are available in standard references (Pearson, 1972).

The $MoSi_2/CrSi_2/TiSi_2$ prototypes are closely related and are just variations in stacking sequences of close-packed triangular layers (Fig. 4). The tetragonal $MoSi_2$ structure results from a twofold repeat, the hexagonal $CrSi_2$ structure results from a threefold repeat, and the orthorhombic $TiSi_2$ results from a fourfold repeat. The $ZrSi_2$ structure is also related to these structures. The structural differences correlate with d-band filling, with C40 being more stable than C54, and $C11_b$ being more stable than C40, as the number of d electrons in the transition metal increases. This suggests that the structural energies of transition metal disilicides are mainly determined by electronic band effects rather than atomic size effects (Carlsson and Meschter, 1991).

1.1.2.2 Interfacial Atomic Structure: Epitaxial Silicide–Silicon Interface In Sec. 1.1.1.2, it was noted that the Schottky-barrier height depends on the atomic structure of the silicide–silicon interface. In addition, novel device structures are made possible by growing silicides epitaxially on Si (see Sec. 3.1.4). Hence, there has been much interest in the atomic structure of epitaxial silicide–Si interfaces.

Factors that affect epitaxial growth are the lattice mismatch between the two materials and the chemical bonding at the interface (Chen and Tu, 1991; van den Hoek *et al.*, 1988; Zur *et al.*, 1985). Epitaxial growth is often observed in systems where the two lattices have identical or quasi-identical forms and have similar lattice spacings. The interfacial chemistry also plays an important role in epitaxial growth; epitaxial growth can occur if the interfacial energy is sufficiently low, even if the lattice mismatch is relatively high (an example is C54-$TiSi_2$ on Si).

Recently, the differences in interface structure between epitaxial $CoSi_2$ and epitaxial $NiSi_2$ on (111) Si have been explained in terms of the interfacial bonds (van den Hoek *et al.*, 1988). For $NiSi_2$, the metal atoms closest to the interface are seven-fold coordinated (seven nearest-neighbor Si atoms), and the Si atoms of the silicide form bonds to the Si substrate, whereas for $CoSi_2$, the metal atoms closest to the interface are eight-fold coordinated (eight nearest neighbor Si atoms) and form bonds to the Si substrate. The most energetically favorable structure depends on the number of electrons in bonding orbitals. In $CoSi_2/(111)$, only

Table 3. Crystal structures of some transition-metal silicides (from Pearson, 1972).

Composition	Prototype	Strukturbericht designation	Pearson symbol	Space group
M_3Si	Ti_3P	—	tP32	$P4_2/n$
	Cr_3Si	A15	cP8	$Pm\bar{3}n$
	Pt_3Ge	—	mC16	$C2/m$
	Fe_3C	DO_{11}	oP16	$Pnma$
	U_3Si	DO_c	tI16	$I4/mcm$
	Fe_3Bi	DO_3	cF16	$Fm\bar{3}m$
M_2Si	Co_2Si	C23	oP12	$Pnma$
	Al_2Cu	C16	tI12	$I4/mcm$
	Fe_2P	C22	hP9	$P\bar{6}2m$
M_5Si_3	W_5Si_3	$D8_m$	tI38	$I4/mcm$
	Mn_5Si_3	$D8_8$	hP16	$P6_3/mcm$
	Cr_5B_3	$D8_l$	tI32	$I4/mcm$
	Ru_5Si_3, Rh_5Si_3[a]	—	oP16	$Pbam$
MSi	FeSi	B20	cP8	$P2_13$
	MnP	B31	oP8	$Pnma$
	FeB	B27	oP8	$Pnma$
MSi_2	$CrSi_2$	C40	hP9	$P6_222$
	$MoSi_2$	$C11_b$	tI6	$I4/mmm$
	$ZrSi_2$	C49	oC12	$Cmcm$
	$TiSi_2$	C54	oF24	$Fddd$
	CaF_2	C1	cF12	$Fm\bar{3}m$

[a] Prototype not known for Ru_5Si_3 and Rh_5Si_3.

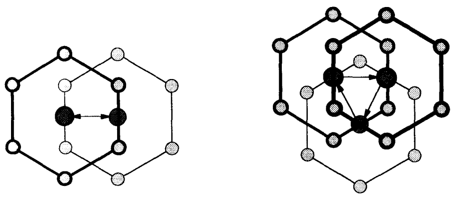

a) $C11_b$ $MoSi_2$ tetragonal b) C40 $CrSi_2$ hexagonal

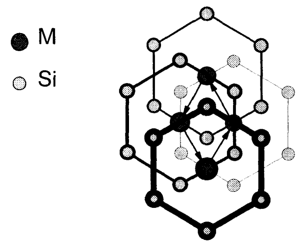

c) C54 $TiSi_2$ orthorhombic

FIG. 4. Stacking relationships between $MoSi_2$, $CrSi_2$, and $TiSi_2$ structures.

the metal–Si bonding orbitals are filled with electrons; the Si–Si bonding orbitals have no electrons, and hence are energetically unfavorable. For $NiSi_2$/(111), the extra d electron from Ni occupies the Si–Si bonding orbital; this is energetically more favorable than metal–Si bonding orbitals with two electrons, because of the large overlap of the Si–Si bonds (Fig. 5).

1.1.2.3 Phonons There have been relatively few studies on phonon spectra (atomic vibrations) in silicides. Most of the studies have focused on semiconducting silicides, where the phonon spectra can be used to analyze phases and crystalline quality (Fenske *et al.*, 1996). A number of studies have also been made of specific heat and low temperature resistivity of silicides, which are related to phonons (Nava *et al.*, 1993). Infrared spectroscopy or Raman scattering has been used to study phonons in PtSi, $TiSi_2$, WSi_2, $CoSi_2$, $NiSi_2$, and β-$FeSi_2$ (Miglio and Meregalli, 1996; Guizetti and Marabelli, 1995).

The phonon spectra of semiconducting silicides show high complexity (Fig. 6), due to the

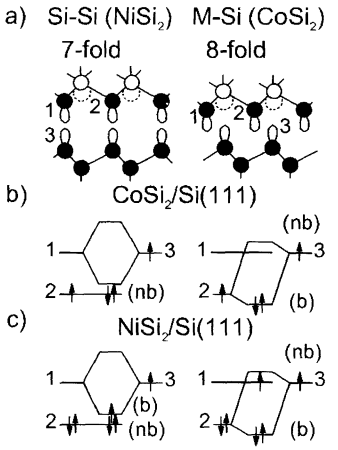

FIG. 5. Energy-level diagrams of interface bonds for (a) seven-fold (left) and eight-fold (right) structures, showing a comparison between (b) $CoSi_2$/Si (111) and (c) $NiSi_2$/Si (111). (Reprinted with permission from van den Hoek *et al.*, 1998. Copyright 1988 by the American Physical Society.)

large number of atoms per unit cell (Miglio and Meregalli, 1996). The phonon spectra for these silicides show a high absorption coefficient, which is a typical feature of ionic bonding, despite the fact the these silicides have strong covalent bonds. A possible explanation is that dynamical charges arise when the atoms vibrate on account of their large polarizability.

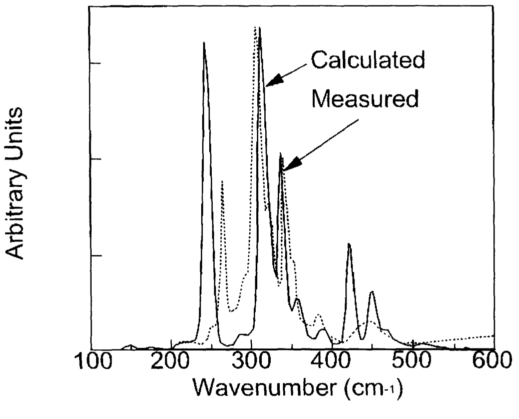

FIG. 6. Comparison between simulation (solid line) and experiment (dotted line) for infrared spectra of β-$FeSi_2$. The peaks in the spectra are probably related to lattice vibrations in the orthorhombic CaF_2 structure. (Reprinted with permission from Miglio and Meregalli, 1996.)

In addition, the phonon spectra for the semiconducting silicides are in nearly the same frequency interval, despite the large mass differences between the metals (for example, Fe and Ir). This suggests that the spectra are mainly determined by the bonding properties, which are similar in these silicides (Fenske *et al.*, 1996; Lange, 1996). Finally, phonons may be responsible for the low mobilities observed in semiconducting silicides (Christensen, 1990).

For silicides with cubic CaF_2 structures, such as $NiSi_2$ and $CoSi_2$, two phonon modes are expected, one Raman-active and the other (at higher frequency) IR-active (Bocelli *et al.*, 1995). The Raman-active mode consists of counterphase motion of the Si atoms along the ⟨111⟩ direction, while the metal atom is at rest (*i.e.*, only the Si atoms move), whereas for the IR-active mode, there is counterphase motion of the Si atoms with respect to the metal along the ⟨111⟩ direction (*i.e.*, both Si and metal atoms move) (Malegori and Miglio, 1993).

1.2 Reactions and Stability

1.2.1 Metal–Si Reactions One of the most common methods to form silicides is by reacting a thin film of metal with Si. This approach is used in microelectronics manufacturing to form silicides on polycrystalline Si gates or single crystal Si contacts (see Sec. 3.1). Reactions between metals and Si are also used in diffusion bonding of Si-based ceramics to metals (see Sec. 3.3). In addition, metal–Si reactions are of scientific interest and have revealed much information on mechanisms of thin-film reactions.

1.2.1.1 Silicide Phases In general, when Si comes in contact with a metal it is not thermodynamically stable and will react to form silicides, with the reaction depending on the annealing temperature, the metal–Si phase diagram, and kinetics (Gas and d'Heurle, 1993; Zhang and d'Heurle, 1994). For transition metals, a series of intermediate silicide phases will form. In principle, all the stable intermediate phases in the phase diagram at the annealing temperature should form. However, often for thin-film reactions some stable phases are not observed on account of kinetics factors (see below).

A number of general observations can be made about the binary phase diagrams of

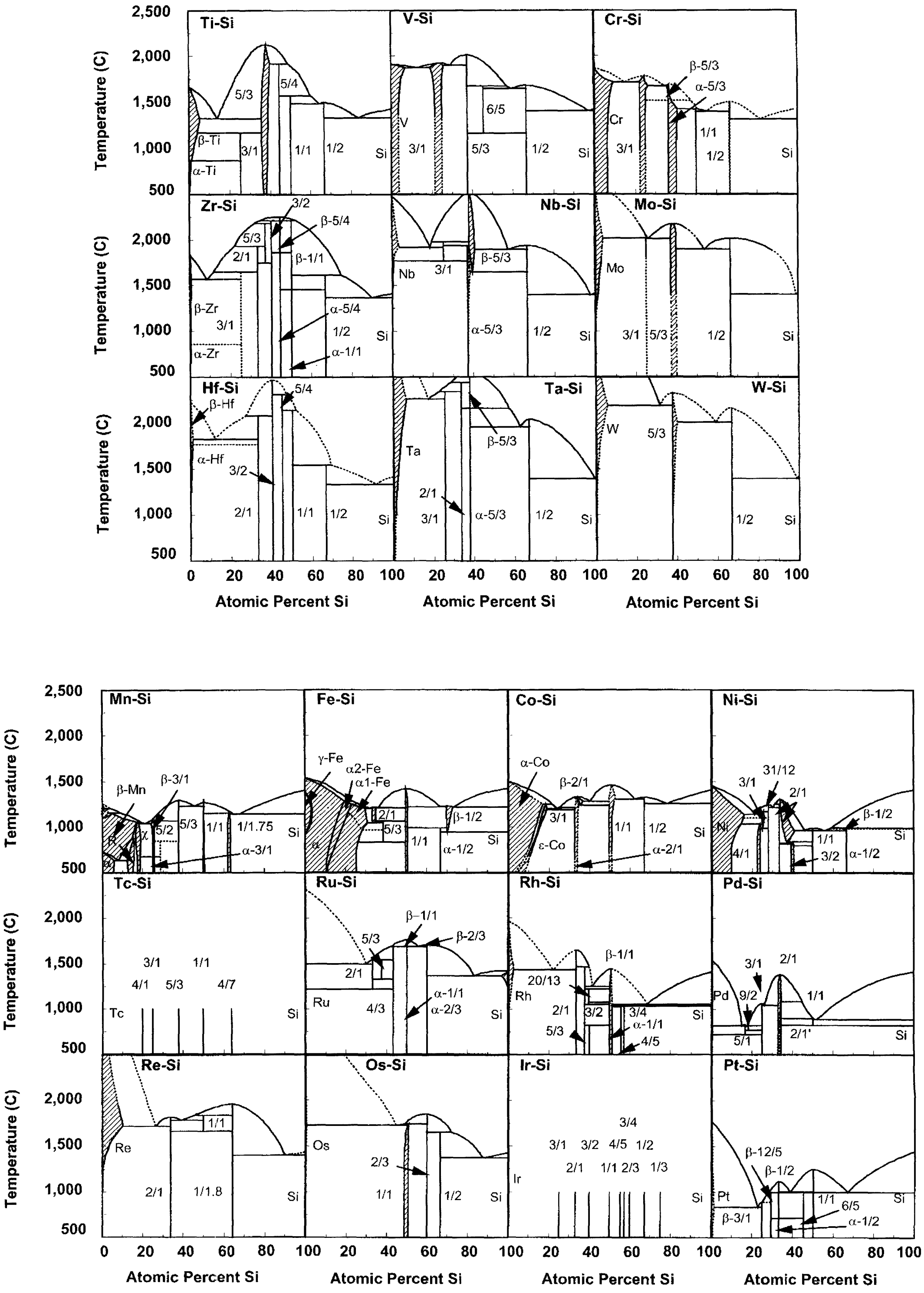

FIG. 7. Phase diagrams of transition-metal silicides (from Massalski, 1990).

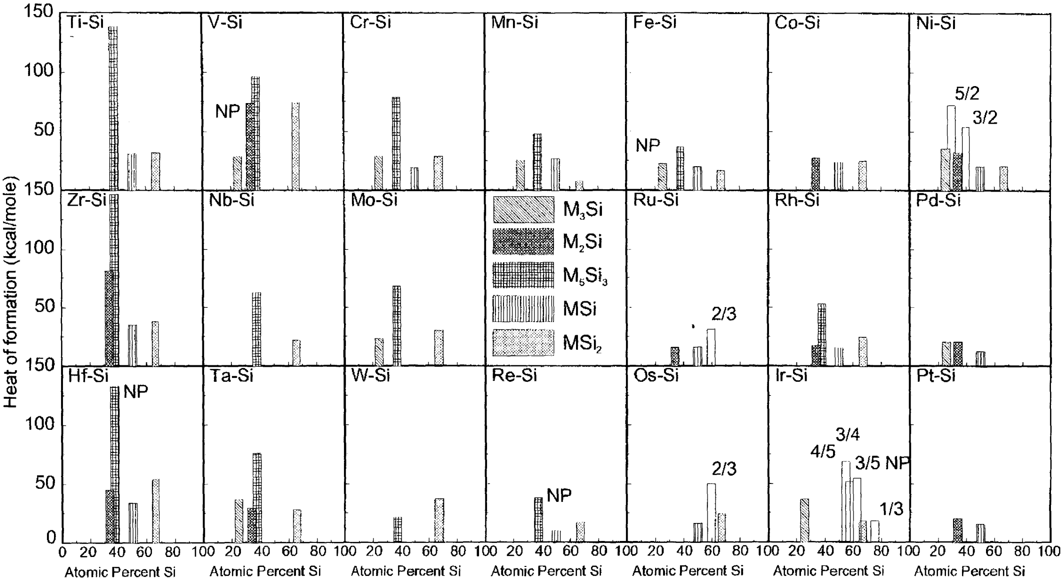

FIG. 8. Heats of formation for transition metal silicides. The M/Si composition is indicated by shading for common stoichiometries or by text. The NP designation is for compositions not present in the equilibrium phase diagram (from Mann and Clevenger, 1995).

transition-metal silicides (Massalski, 1990) (Fig. 7). As noted above, the Si-rich phase is typically MSi_2 except in the case of Ir where a more Si-rich phase, $IrSi_3$, is formed. The most metal-rich phase is generally M_4Si or M_3Si with a number of exceptions. Typically, about five silicide phases are formed for each element, with a minimum of two for W–Si or a maximum of almost ten for Ni, Rh, or Ir. There is significant Si solubility for the first row of transition metals (Ti to Ni), about 3–20 at.% Si, and for Nb, Mo, Ta, W, and Re in the second and third rows, about 3–10 at.% Si. In general, silicides have very limited composition ranges. Significant deviations from stoichiometry are found in Ti_5Si_3, Nb_5Si_3, Cr_5Si_3, Mo_5Si_3, W_5Si_3 (W_5Si_3 and Mn_5Si_3 are prototype structures; see Fig. 3), V_3Si, Cr_3Si (Cr_3Si prototype), FeSi, OsSi, CoSi (FeSi prototype), and a number of other silicides.

Typical values of the transition-metal silicide heats of formation are arranged according to the periodic table (Fig. 8) (Mann and Clevenger, 1995). The heat of formation is related to the d-band occupancy of the silicide. In refractory-metal silicides, only the bonding orbitals are occupied, so that the heat of formation directly reflects the bonding energy. However, in near-noble–metal silicides, nonbonding and antibonding orbitals are occupied, which reduces the bonding energy. Hence, the heat of formation is lower for near-noble–metal silicides compared with refractory-metal silicides (Hara and Ohdomari, 1988).

1.2.1.2 Kinetics Silicides formed by reacting a metal with Si require nucleation and growth of the new phase. If the growth rate is faster than the nucleation rate, then the silicide formation is nucleation controlled. If the nucleation rate is faster than the growth rate, then the silicide formation is either diffusion limited or interface reaction-rate limited (Mayer and Lau, 1989).

Diffusion-limited reactions generally exhibit laterally uniform growth with well-defined kinetics, whereas nucleation-limited reactions exhibit laterally nonuniform growth and a critical temperature dependence (no reaction below the critical temperature and extremely rapid reaction above the critical temperature). In most applications (and especially for integrated circuits), a uniform reaction is required; hence diffusion-limited kinetics are preferred.

Nucleation effects are often observed during formation of new phases in thin films (d'Heurle, 1988). Nucleation-limited reactions are expected when a new phase is forming at an interface and the free-energy change for the transformation is small with respect to the increase in surface energy. Nucleation is generally not rate limiting in reactions between transition metals and Si, because the free-energy change is usually large. Nucleation-limited reactions are often observed during the formation of the Si-rich silicide phases or when the transition results in only a small change in composition or structure (for example, the transformation of C49-$\mathrm{TiSi_2}$ to C54-$\mathrm{TiSi_2}$). The films produced from nucleation-limited reactions are often rough; the film nucleates in isolated regions, then grows rapidly both laterally and vertically.

After nucleation, growth of the new phase is limited either by the interface reaction or by diffusion (Gas and d'Heurle, 1993). As the thickness increases, the growth of the new phase becomes limited by the flux of atoms through the new phase to the growing interface(s). The growth of the new phase will be diffusion limited and will increase with the square root of time,

$$L(t)^2 = K_{\mathrm{d}}t \tag{2}$$

where K_{d} is related to the diffusivity of the dominant moving species through the new phase. The analysis is much more complicated if two or more phases grow simultaneously (Zhang and d'Heurle, 1994).

For thin-film reactions such as those used in CMOS processes (*i.e.*, formation of NiSi from $\mathrm{Ni_2Si}$ and Si), it is generally observed that the silicide phases form sequentially rather than simultaneously. The new phase (NiSi) only begins to grow after the most metal-rich phase (Ni) is completely consumed. In contrast, in bulk diffusion couples the $\mathrm{Ni_2Si}$ phase can reach the critical thickness such that $\mathrm{Ni_2Si}$ and NiSi (as well as other phases) can exist simultaneously. Of course, if the diffusion of reactants is altered (by impurities in the films, for example) then multiple silicide phases may be observed simultaneously even in the thin-film case.

1.2.1.3 Results on Metal–Si Reactions Diffusion-limited kinetics have been observed for both bulk and thin-film reactions between metals and Si (Lien *et al.*, 1986; d'Heurle and Gas, 1986; Mayer and Lau, 1989; Gas and d'Heurle, 1997) (Table 4). The reaction kinetics are similar for bulk and thin-film couples. However, the diffusivities are higher in the thin-film case as a result of the contribution from grain-boundary diffusion (Barge *et al.*, 1995). The kinetics can be analyzed by plotting the square of thickness versus time for different annealing temperatures. For diffusion-limited kinetics and assuming that only one phase forms at a time, a series of straight lines is ob-

Table 4. Silicide phase sequence for thin metal films on single-crystal Si. The silicides are listed in the order of their appearance from the top down. The diffusing species during silicide formation is italicized. Where known, the reaction kinetics, t or $\sqrt{t}$, are indicated. (Reprinted from Colgan *et al.*, 1996, copyright 1996, with permission from Elsevier Science.)

Ti[a] $\mathrm{Ti_5Si_3}$ (?) $\mathrm{Ti}\mathit{Si}$ $\mathrm{Ti}\mathit{Si}_2$ t, $\sqrt{t}$(?)	**V** $\mathrm{V}\mathit{Si}_2$ t, $\sqrt{t}$	**Cr** $\mathrm{Cr}\mathit{Si}_2$ t
Zr $\mathrm{ZrSi_2}$	**Nb** $\mathrm{Nb}\mathit{Si}_2$	**Mo** $\mathrm{Mo}\mathit{Si}_2$ t
Hf $\mathrm{Hf}\mathit{Si}$ $\sqrt{t}$ $\mathrm{HfSi_2}$[b]	**Ta** $\mathrm{Ta}\mathit{Si}_2$ t	**W** $\mathrm{W}\mathit{Si}_2$ t, $\sqrt{t}$

Co $\mathit{Co}_2\mathrm{Si}$ $\sqrt{t}$ $\mathrm{Co}\mathit{Si}$ $\sqrt{t}$ $\mathit{Co}\mathrm{Si}_2$ t[b]	**Ni** $\mathit{Ni}_2\mathrm{Si}$ $\sqrt{t}$ $\mathit{Ni}\mathrm{Si}$ $\sqrt{t}, t$ $\mathit{Ni}\mathrm{Si}_2$ t[b]
Rh $\mathrm{Rh}\mathit{Si}$ $\sqrt{t}$ $\mathrm{Rh}_4\mathit{Si}_5$[b] $\mathrm{Rh}_3\mathit{Si}_4$[b]	**Pd**[c] $\mathit{Pd}_2\mathit{Si}$ $\sqrt{t}$ PdSi[b]
Ir $\mathrm{Ir}\mathit{Si}$ $\sqrt{t}$ $\mathrm{IrSi_{1.75}}$[b]	**Pt** $\mathit{Pt}_2\mathrm{Si}$ $\sqrt{t}$ $\mathit{Pt}\mathrm{Si}$ $\sqrt{t}$

[a] Whether $\mathrm{Ti_5Si_3}$ and TiSi form before $\mathrm{TiSi_2}$ in a clean system is not clear. The growth kinetics of $\mathrm{TiSi_2}$ are linear on single crystal Si and $\sqrt{t}$ on α-Si.

[b] Silicide formation occurs in a laterally nonuniform manner.

[c] The diffusing species during $\mathrm{Pd}_2\mathit{Si}$ formation is probably both Pd and Si.

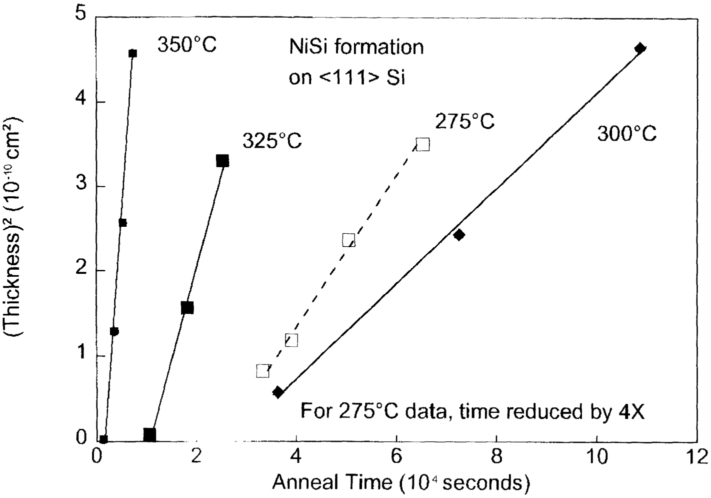

FIG. 9. The thickness of NiSi squared as a function of time, showing diffusion-limited kinetics. (Reprinted with permission from d' Heurle *et al.*, 1984. Copyright 1984 American Institute of Physics.)

tained (d'Heurle *et al.*, 1984) (Fig. 9). The slopes of the lines are related to the diffusivity of the dominant diffusing species, while an Arrenhius plot of the slopes is related to the activation energy for diffusion.

In most cases (Co in Co_2Si, Ni in Ni_2Si, Si in $TiSi_2$), the majority atoms in the silicide are the most mobile (d'Heurle and Gas, 1986). In general, these atoms are next-nearest neighbors to one another, so that a vacancy can propagate along the network of majority atoms with only small perturbations of the lattice. In contrast, the motion of minority atoms requires the coexistence of either two vacancies (minority and majority) or local destruction of the lattice, resulting in a high activation energy.

Nucleation-limited reactions have been observed in a number of silicides where the free energy of formation is small (d'Heurle, 1988). These nucleation effects are eliminated when these phases are formed on amorphous Si, because of the additional free energy in the system associated with crystallization of amorphous Si. Similarly, there is evidence that grain boundaries and grain-boundary triple points are preferred sites for the nucleation of both $CoSi_2$ (Applebaum *et al.*, 1985) and C54-$TiSi_2$ (Ma and Allen, 1994), presumably because of the additional gain in free energy produced by eliminating grain boundaries. The nucleation of C54-$TiSi_2$ is an important problem in microelectronics processing (see Sec. 3.1).

Metal–Si reactions can also be affected by impurities, which generally inhibit silicide formation. The impurities can be films at the metal–Si interface, such as native oxides or surface contamination, contaminants in the metal film from either the metal deposition or the anneal ambient, or dopants in the Si substrate. The most common contaminants during processing of silicides for integrated circuits are oxygen and nitrogen, as well as dopants such as As, P, and B (Gambino and Colgan, 1998).

1.2.2 Results on Metal Reactions with Si-Based Compounds

Metals can react with a number of Si-based compounds to form silicides, including SiO_2, Si_3N_4, SiC, and SiGe. For Si integrated-circuit processing, SiO_2 and Si_3N_4 are used as insulators and often react with transition metals used for contacts and interconnects, such as Ti or Co. These reactions can be beneficial, for example by providing good adhesion between the metal and the insulator, or they can be detrimental, for example by leaving conducting filaments on insulating regions. For structural materials, SiO_2, Si_3N_4, or SiC is used either as reinforcing fibers or as the matrix in composites, and often reacts with transition metals that are in the composites. These reactions are usually beneficial, providing strong adhesion between the different components of the composites. Reactions between metals and ceramics are also used to join structural metal alloys to structural ceramics, by diffusion bonding. Finally, SiC or SiGe is used in novel microelectronic devices, and reacts with metals in contacts to form Ohmic or Schottky contacts.

1.2.2.1 Metal–SiGe Reactions There are many studies of reactions of metals with SiGe alloys, including Ti, Zr, W, Co, Ni, Fe, Cr, Mn, Pd, and Pt (Aldrich *et al.*, 1996; Aldrich *et al.*, 1995; Hong and Mayer, 1990; Luo *et al.*, 1997; Carmondy *et al.*, 1996; Boutarek and Madar, 1993). Si and Ge are very similar chemically; hence many of the silicide phases observed in metal–Si reactions are also observed in metal reactions with SiGe alloys. However, there are a number of notable differences, including formation of metal–Ge compounds or metal–SiGe solid solutions, and segregation of Ge.

A number of general trends are observed during reactions of metals with SiGe alloys. Silicide formation is favored over germanide formation, because of the higher heat of formation for silicides (Aldrich *et al.*, 1996). For the same reason, the silicide formation temperatures are higher on SiGe alloys compared with pure Si (Aldrich *et al.*, 1995; Luo *et al.*, 1997). The phase sequence is silicide-like for Si-rich SiGe alloys and germanide-like for Ge-rich SiGe alloys (Aldrich *et al.*, 1995). The dominant diffusing species is usually the same as during silicide formation (*i.e.*, generally metal for near-noble metals and Si/Ge for refractory metals).

The phases that form depend on the miscibility of the silicides and germanides of a given metal (Boutarek and Madar, 1993). For Ti (and Zr), the C54 phases of $TiSi_2$ and $TiGe_2$ have the same structure and very similar unit cell dimensions, and therefore are completely miscible, forming a $Ti(Si_{1-y}Ge_y)_2$ solid solution. Metals where the silicides and germanides are partially miscible, such as V, form solid solutions over a limited range of SiGe alloy compositions. Finally, metals for which the silicides and germanides are immiscible, such as Fe, Cr, and Mn, form separate silicide and germanide phases (Fig. 10). For some metals both types of behavior are observed, depending on the phase that forms; CoSi and CoGe form solid solutions whereas $CoSi_2$ and $CoGe_2$ do not.

1.2.2.2 Metal–SiO_2 Reactions The transition metals exhibit a wide range of reactivities with SiO_2. Titanium and Ta will reduce SiO_2, forming metal oxide and silicide compounds, whereas Co, Mo, and W do not react significantly with SiO_2. These differences can be

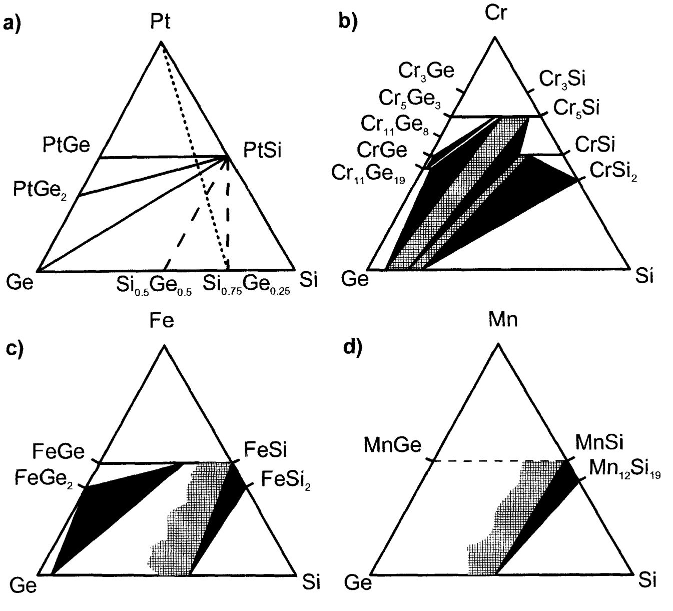

FIG. 10. Phase diagrams for ternary metal–Si–Ge systems; (a) Pt–Si–Ge (Hong and Mayer, 1990), (b) Cr–Si–Ge, (c) Fe–Si–Ge, and (d) Mn–Si–Ge (reprinted from Boutarek and Madar, 1993, copyright 1993, with permission from Elsevier Science).

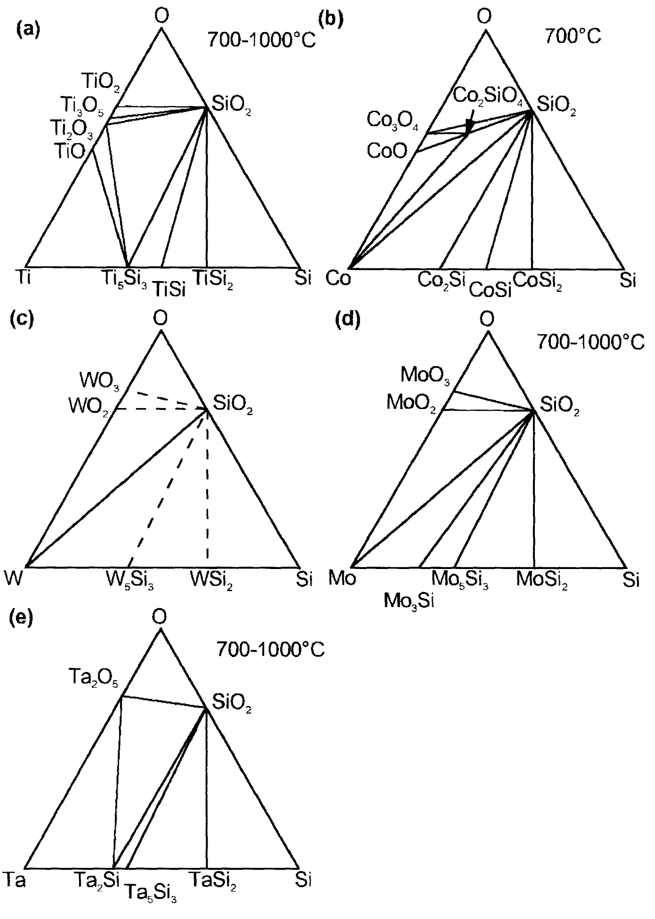
(a)
O
700-1000°C
TiO2
Ti3O5
Ti2O3
TiO
SiO2
Ti
Ti5Si3
TiSi
TiSi2
Si
(b)
O
700°C
Co2SiO4
SiO2
Co3O4
CoO
Co
Co2Si
CoSi
CoSi2
Si
(c)
O
WO3
WO2
SiO2
W
W5Si3
WSi2
Si
(d)
O
700-1000°C
MoO3
MoO2
SiO2
Mo
Mo5Si3
MoSi2
Si
Mo3Si
(e)
O
700-1000°C
Ta2O5
SiO2
Ta
Ta2Si
Ta5Si3
TaSi2
Si

FIG. 11. Ternary phase diagrams for (a) Ti–Si–O, (b) Co–Si–O, (c) W–Si–O, (d) Mo–Si–O and (e) Ta–Si–O. [Parts (a) and (c)–(e) reprinted with permission from Bayers, 1984. Copyright 1984 American Institute of Physics. Part (b) from Maex, 1993, with permission from Elsevier Science.]

understood in terms of the M–Si–O phase diagrams (Fig. 11) (Beyers, 1984; Maex 1993); no reaction will occur between the metal and SiO_2 if there is a tie line between the metal and SiO_2.

Titanium is a good example of a metal that reacts with SiO_2. Reactions between Ti thin films and SiO_2 result in the formation of Ti_5Si_3 at the interface and TiO at the surface (Wang and Mayer, 1990). Oxygen from the consumed SiO_2 is pushed into the unreacted Ti and eventually forms TiO on top of the Ti_5Si_3. A metal-rich silicide (Ti_5Si_3) forms rather than a Si-rich silicide ($TiSi_2$) because the supply of Si is limited (since it is bonded to SiO_2) whereas the supply of Ti is relatively large. Similarly, a Ti-rich oxide is formed (TiO) rather than an oxygen-rich oxide (TiO_2), because the supply of oxygen is limited. Cobalt is a good example of a metal that is relatively unreactive with SiO_2. In fact, no reaction is observed at temperatures below 700 °C (Nguyen *et al.*, 1996).

1.2.2.3 Metal–Si_3N_4 Reactions There have been relatively few studies of metal reactions with Si_3N_4. However, the results are expected to be qualitatively similar to metal reactions with SiO_2: metals that react with SiO_2 will generally also react with Si_3N_4. Reactions between Ti and Si_3N_4 have been studied in detail. The ternary phase diagram (Fig. 12) shows that reactions are expected between Ti and Si_3N_4 (Sambasivan and Petuskey, 1994). At low temperatures (500 °C), the reaction between Ti and Si_3N_4 results in Ti_5Si_3 at the interface, with N dissolving into the unreacted Ti, analogous to the case of Ti on SiO_2. However, at higher temperatures (800 °C), $TiSi_2$ forms at the interface and TiN

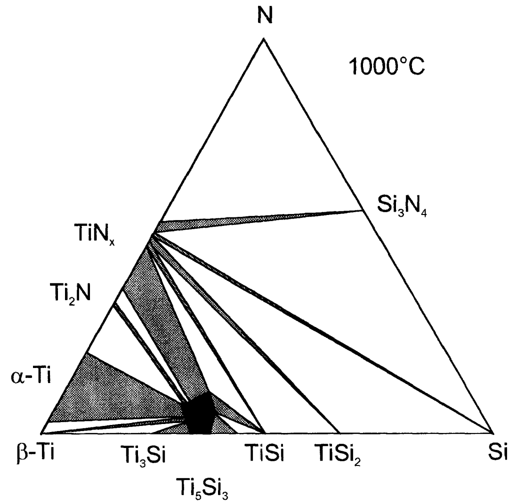

FIG. 12. Ternary phase diagrams for Ti–Si–N (reprinted with permission from Sambasivan and Petuskey, 1994, Fig. 2).

forms on the surface (Barbour *et al.*, 1987), presumably because the reaction is no longer limited by the supply of Si.

1.2.2.4 Metal–SiC Reactions There are many studies of reactions of metals with SiC, including Ti, Hf, Mo, Co, Ni, and Pt. At low temperatures (<750 °C) the results are qualitatively similar to studies on metal reactions with SiO_2 or Si_3N_4. The Ti–Si–C phase diagram (Fig. 13) shows that reactions are expected between Ti and SiC (Taubenblatt and Helms, 1986). Experiments show the Ti_5Si_3 forms at the interface and TiC forms at the surface (Porter *et al.*, 1995a,b). No reactions are observed between Pt, Ni, or Co and SiC at low temperatures. However, at high temperatures (>900 °C), extensive reactions are observed associated with interfacial melting between the metal and SiC. The most likely mechanism is formation of a thin molten silicide layer at the interface (*i.e.*, if the anneal temperature is above the eutectic temperature of the silicide), followed by decomposition of the SiC, with the decomposition of SiC expected to be the rate-limiting step. Because of the limited supply of Si, the reacted layer generally consists of a metal-rich silicide, such as Pt_2Si, Co_2Si, or Ni_3Si (Chou *et al.*, 1991; Porter *et al.*, 1995a,b), though Si-rich silicides are sometimes observed at the highest reaction temperatures. Carbon is also observed, rather than metal carbides, probably in consequence of the low heats of formation for these metal carbides.

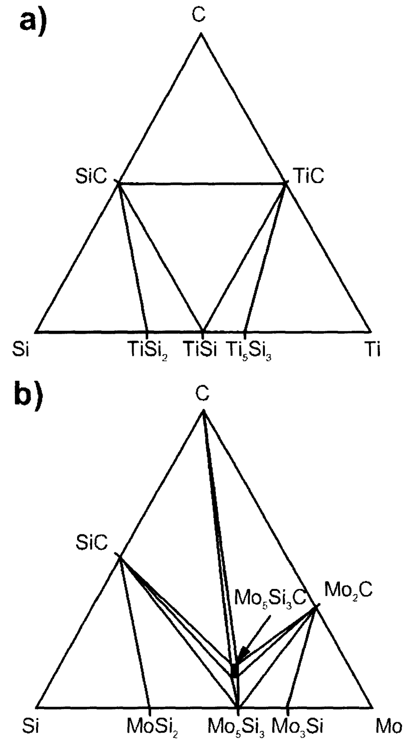

FIG. 13. Ternary phase diagrams for (a) Ti–Si–C (reprinted with permission from Taubenblatt and Helms, 1986. Copyright 1986 American Institute of Physics) and (b) Mo–Si–C (from van Loo *et al.*, 1982, Fig. 2, reproduced with permission).

1.2.3 Thermal Stability The thermal stability of silicides is one of the main reasons that they are used in so many different applications. In microelectronics, the silicide must be stable with respect to the underlying semiconductor during annealing steps used in integrated-circuit manufacturing (such as glass reflow). In structural materials, the silicide must be stable in high-temperature environments, such as the turbines in jet engines.

1.2.3.1 Melting Point Using the traditional definition of "refractory" (stable at >1538 °C, the melting point of Fe), the majority of transition-metal silicides are refractory compounds (Fig. 7). The melting points of the transition metals range from 1246 °C for Mn to 3422 °C for W (Massalski, 1990) and

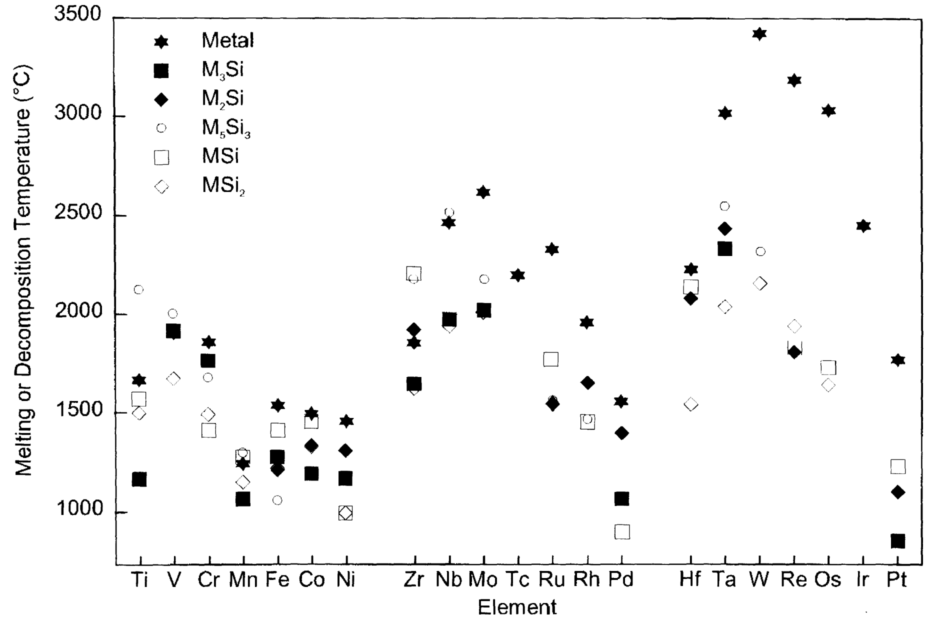

FIG. 14. Melting or decomposition temperature for transition metals and some transition metal silicides by period and group in the periodic table (from Massalski, 1990).

the decomposition temperature of the most stable transition-metal silicides range from 1229 °C for PtSi to 2550 °C for Ta_5Si_3. For comparison, the melting temperature of Si is 1414 °C. The melting or decomposition temperatures increase with the period number (Fig. 14), reaching their maximum values for groups V or VI (V, Cr, Nb, Mo, Ta and W), and are at their minimum values for the near-noble metals (Ni, Pd and Pd). Interestingly, there are silicide phases that have higher melting points than that of the parent metal for Ti, Zr, Hf, V, Nb, and Mn (Figs. 7 and 14). The thermal stability of the near-noble–metal silicide thin films on silicon is also limited by the presence of low-temperature (< 1000 °C) Si-rich eutectics (Fig. 7). The metal melting point and the melting point of the most stable silicide are somewhat correlated (Fig. 15).

1.2.3.2 Stability of Thin Films The stability of silicide thin films is of interest in microelectronics applications. In Si integrated circuits, silicides must be stable on both single-crystal Si and polycrystalline Si. High-temperature annealing can cause agglomeration of silicides on either single-crystal or polycrystalline Si, and can cause silicide-enhanced grain growth in polycrystalline Si (Colgan *et al.*, 1996a).

Agglomeration of silicides on single-crystal Si is often observed after high-temperature annealing, especially for near-noble–metal silicides. There have been many studies of the agglomeration of silicide thin films on Si, including $TiSi_2$, $CoSi_2$, and $NiSi_2$ (Colgan *et al.*, 1996a,b). Agglomeration starts with grain-boundary grooving in the silicide, followed by grain separation and finally the formation of silicide islands. The driving force for grain-boundary grooving is the reduction of grain-boundary energy in the system (Nolan *et al.*, 1992; Mullins, 1957; see also GRAIN BOUNDARIES).

Agglomeration depends on a number of factors, including film thickness, grain size, grain-boundary energy, surface energy (of the film and the substrate), and interface energy (of the film–substrate interface). Agglomeration will increase as the groove depth increases; hence, agglomeration increases with decreasing film thickness, increasing grain size, and increasing grain-boundary energy. Agglomeration is probably limited by the slowest-diffusing species, because there must be transport of

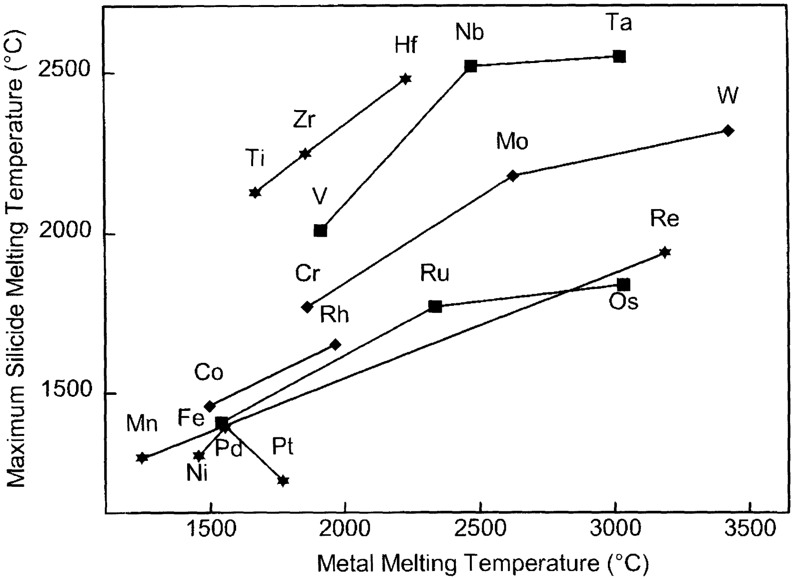

FIG. 15. Maximum silicide melting temperature *versus* parent metal melting temperature. Elements in the same group are connected by lines (Massalski, 1990).

both Si and the metal. Hence, it is expected that the agglomeration temperature for a given silicide will scale with the melting point of the silicide.

Silicide thin films are less stable on polycrystalline Si than on single-crystal Si, because there is an additional driving force for morphological changes: the grain-boundary energy of the polycrystalline Si. Grain growth in the polycrystalline Si is enhanced by the silicide, which provides a fast diffusion path for transport of material. The grain growth in the polycrystalline Si can cause severe roughening of the silicide–polycrystalline Si interface, and is undesirable in integrated circuits (Colgan *et al.*, 1996a; Gambino *et al.*, 1998). On the other hand, similar silicide-enhanced grain growth has been used to form large-grain Si films, which are desirable for thin-film transistors (Lee *et al.*, 1996). The temperature at which silicide-enhanced grain growth occurs for a given silicide scales with the deformation temperature of the silicide (Fig. 16). The growing Si grains protrude into the silicide; the growth

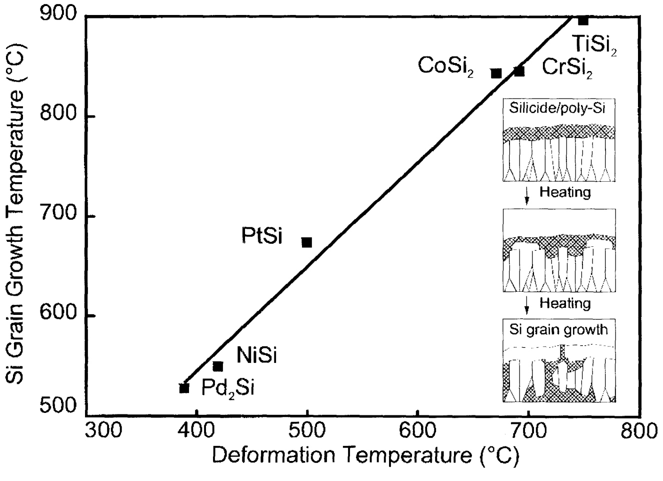

FIG. 16. Silicide degradation temperature as a function of silicide deformation temperature for various silicides (from Hong *et al.*, 1994).

of the grains will be constrained if the silicide is difficult to deform (Hong *et al.*, 1994).

1.2.4 Chemical Stability The chemical stability of silicides is important in a number of applications. In structural materials, it is desirable that the silicides withstand corrosive environments such as high-temperature oxidation. The oxidation resistance of silicides is also of interest in microelectronics applications, where silicides are exposed to high-temperature oxidations used to fabricate devices. However, during integrated-circuit fabrication it is also necessary to pattern silicides; in this case etching of the silicides is necessary and desirable.

1.2.4.1 Oxidation The oxidation of silicides has been widely studied for both microelectronics applications and structural materials. Oxidation can result in either SiO_2 formation or metal oxides, depending on the availability of Si at the oxide–silicide interface. In general, it is preferred to form SiO_2 rather than metal oxides, because SiO_2 is a stable oxide that grows relatively slowly without affecting the integrity of the underlying silicide. In contrast, many metal oxides (such as W oxides and Mo oxides) are volatile and can produce cracks or blisters in the oxide film.

Metal-free oxides can be formed on most of the silicides (including WSi_2, $MoSi_2$, $CoSi_2$, $TiSi_2$, $NiSi_2$ and PtSi) if there is an adequate supply of Si. This can be seen from the ternary phase diagrams (Fig. 11); metal oxides will not form if the system stays in the MSi_2–SiO_2–Si triangle (Beyers *et al.*, 1984). The supply of Si at the oxide–silicide interface can be analyzed in terms of the activity of Si at the interface, which is related to the chemical potential (Zhang and d'Heurle, 1992; Lee, 1992). If the silicide is saturated with Si, the activity of Si at the oxide–silicide interface approaches unity and SiO_2 is formed. However, if the supply of Si is limited, the activity at the oxide–silicide interface will decrease, and may become low enough to allow oxidation of metals to occur.

For refractory-metal silicides (such as $TiSi_2$, $MoSi_2$, and WSi_2), Si diffusion through the silicide provides the excess Si required for SiO_2 formation. The excess Si must be supplied from the underlying Si (in the case of a thin silicide film) or from excess Si in the bulk of the silicide. In either case, the diffusion of Si to the oxide–silicide interface must be unhindered, to ensure that there is an adequate supply for SiO_2 formation. Because the activation energy for Si diffusion throught the silicide is higher than that for the growth of SiO_2, high oxidation temperatures (800 °C or above) are required to form SiO_2. At low temperatures, the supply of Si will be limited and metal oxides will form (Zhang and d'Heurle, 1992).

For near-noble–metal silicides (such as $NiSi_2$, $CoSi_2$, and PtSi), the metal diffusivity is high and oxidation occurs by a different mechanism. During oxidation, the silicide decomposes at the oxide–silicide interface, providing Si for oxidation. The free metal then diffuses to the underlying Si, forming more silicide. During oxidation of $NiSi_2$ and $CoSi_2$, there is also diffusion of excess Si away from the oxide–silicide interface, due to the volume increase (and hence compressive stress) associated with SiO_2 formation (d'Heurle *et al.*, 1987). In either case, the diffusivity of the metal must be high enough to ensure an adequate supply of Si at the oxide–silicide interface.

The formation of SiO_2 on silicides and on Si occur by similar mechanisms (Zhang and d'Heurle, 1992). During the initial stages of the reaction, oxidation is limited by the interfacial reaction between Si and oxygen, and oxide growth will be linear with time. As the oxide grows, the reaction will eventually be limited by diffusion of oxidizing species through the SiO_2, and the oxide thickness will increase with the square root of time. For dry oxidations, the diffusing species is molecular oxygen, whereas for wet oxidations (*i.e.*, steam), the diffusing species is H_2O (Irene, 1988). Although the kinetic processes are similar, the oxidation rates of silicides are generally higher than that of Si, with near-noble–metal silicides usually having higher oxidation rates than refractory metal silicides (Frampton *et al.*, 1987) (Fig. 17).

The oxidation of WSi_2 and $MoSi_2$ can result in the formation of volatile metal oxides (such as WO_3 and MoO_3) if there is insufficient Si at the surface. The formation of these volatile metal oxides causes volume expansion and blistering of the protective SiO_2 films, resulting in futher oxidation of the silicide. The formation of metal oxides can cause peeling of silicide thin films (Yoo *et al.*, 1990) and can cause disintegration of bulk silicides; this is

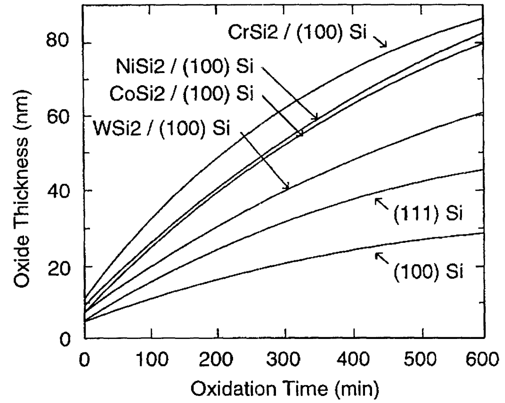

FIG. 17. The kinetics of SiO_2 formation for several silicides at 800°C in nominally dry O_2 (reprinted with permission from d'Heurle *et al.*, 1987, and Taylor & Francis, Ltd.).

often referred to as a "pest" reaction (Fig. 18) (Grabke and Meier, 1995; Chou and Nieh, 1993).

A number of methods have been investigated for preventing metal oxidation. For thin films used in microelectronics, a Si-rich silicide is used to ensure that there is an adequate supply of Si at the surface during oxidation. For bulk $MoSi_2$ structural materials, one approach that has been used is to alloy the silicide with Ge. Germanium forms a glass (GeO_2) at lower temperatures than SiO_2; the GeO_2 glass protects the silicide at low temperatures. The GeO_2 also lowers the viscosity of SiO_2 at high temperatures, making it easier to seal cracks if they form. In addition, GeO_2 increases the coefficient of thermal expansion, which reduces the expansion mismatch between the $MoSi_2$ and the protective SiO_2 layer (Mueller *et al.*, 1992).

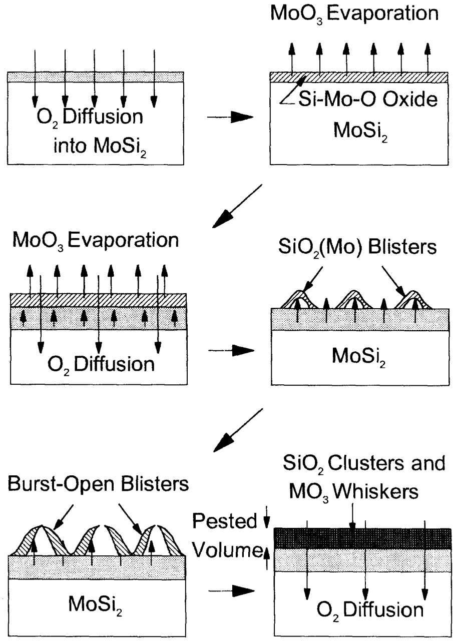

FIG. 18. Schematic of pest reaction in single crystal $MoSi_2$ (reprinted with permission from Chou and Nieh, 1993, Fig. 17).

1.2.4.2 Silicide Reactions with Dopants (As, P, B) In microelectronics applications, silicides are often in contact with heavily doped Si (*i.e.*, Si doped with As, P, or B). For these applications, it is desirable to maximize the active dopant concentration in the Si. Hence, the thermodynamic and kinetic stability of the silicide on doped Si is important. Diffusion of dopants into the silicide can lower the dopant concentration in the Si, causing a higher contact resistance for contacts to diffusions and a threshold-voltage shift for contacts to gates. At the same time, dopant diffusion from the silicide can be beneficial, making it possible to form shallow junctions. Interactions between dopants and silicides are also relevant for structural materials; B doping in Ni_3Si improves the ductility of the alloy (Takasugi and Yoshida, 1991). Hence, an understanding of the interactions between silicides and dopants is required, including compound formation, solubility, diffusivity, and segregation (Maex *et al.*, 1989; Lien and Nicolet, 1984; Murarka and Williams, 1987).

There have been relatively few studies on the solubility of dopants in silicides (Shone *et al.*, 1986; Jahnel *et al.*, 1982); most of the data on solubilities are from studies on dopant diffusion in polycrystalline silicides. In most cases, the solubilities were estimated on the basis of the average dopant concentration in the silicide. It is expected that dopants segregate to grain boundaries in many cases, and so the actual solubility (*i.e.*, in the grains) will be less than that estimated from measurements on polycrystalline thin films.

Silicides can react with dopants to form metal–dopant compounds. Most silicides are

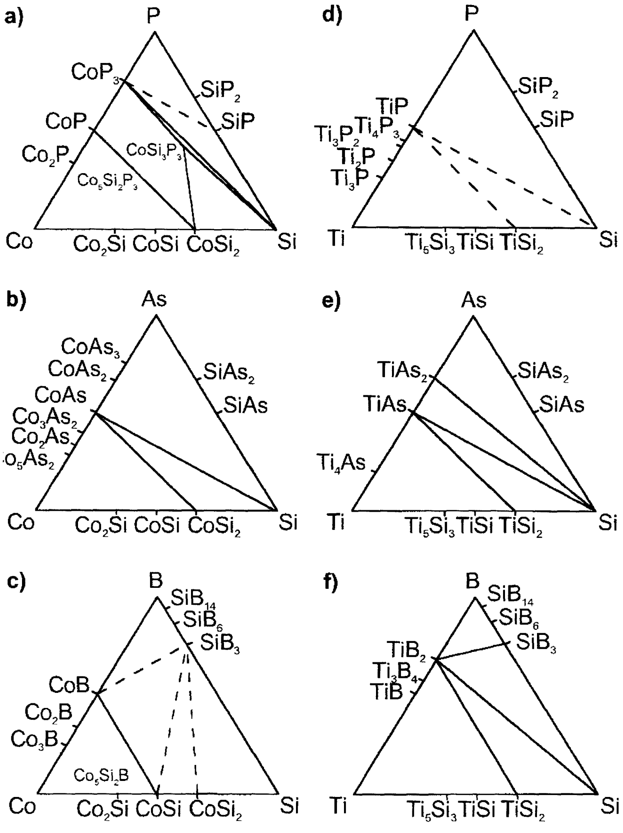
a)
P
CoP3
SiP2
CoP
SiP
Co2P
CoSi3P3
Co5Si2P3
Co
Co2Si CoSi CoSi2
Si
d)
P
TiP
Ti3P2Ti4P3
SiP2
SiP
Ti2P
Ti3P
Ti
Ti5Si3TiSi TiSi2
Si
b)
As
CoAs3
CoAs2
SiAs2
CoAs
SiAs
Co3As2
Co2As
o5As2
Co
Co2Si CoSi CoSi2
Si
e)
As
TiAs2
SiAs2
TiAs
SiAs
Ti4As
Ti
Ti5Si3TiSi TiSi2
Si
c)
B
SiB14
SiB6
SiB3
CoB
Co2B
Co3B
Co5Si2B
Co
Co2Si CoSi CoSi2
Si
f)
B
SiB14
SiB6
SiB3
TiB2
Ti3B4
TiB
Ti
Ti5Si3TiSi TiSi2
Si

FIG. 19. Ternary phase diagrams for metal–Si–dopant phases: (a) Co–Si–P, (b) Co–Si–As, (c) Co–Si–B, (d) Ti–Si–P, (e) Ti–Si–As and (f) Ti–Si–B (reprinted with permission from Madar, 1995).

expected to form binary or ternary compounds with dopants at sufficiently high dopant concentrations (Fig. 19, from Madar, 1995). However, these compounds are often not observed in microelectronics applications (Maex, 1993), perhaps because dopant concentrations are below the solubility limit in the silicide. In contrast, binary and ternary dopant compounds are often observed in structural silicides (Briant and Taub, 1988).

There have been many studies on dopant diffusion in polycrystalline silicides (Gas *et al.*, 1995). With the exception of a few studies where silicides with large grain sizes were used, diffusivities measured in most studies are probably dominated by grain-boundary diffusion. This is directly relevant for microelectronics applications, where the silicides are polycrystalline thin films. A comparison of dopant diffusion studies (Fig. 20) indicates that grain-boundary diffusion greatly enhances dopant transport in silicides with small grains

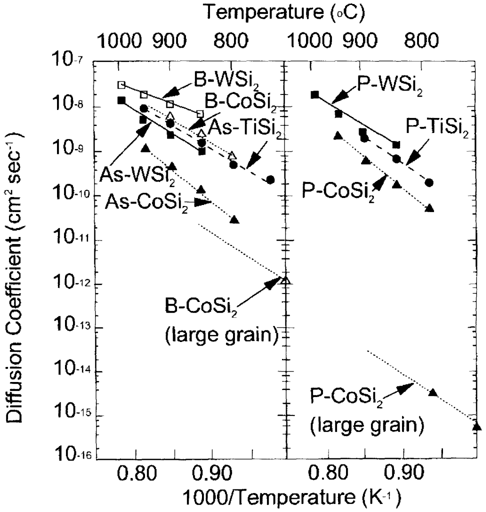

FIG. 20. Diffusion coefficients of B, As and P in various silicides (reprinted from Gambino and Colgan, 1998, with permission from Elsevier Science).

(Chu *et al.*, 1992) compared with silicides with large grains (Thomas *et al.*, 1988b; Zaring *et al.*, 1990). In most cases, the diffusion coefficients of the dopants in silicides are higher than those in Si. The activation energies for volume diffusion are typically 2–3 eV, and are similar to those for silicide self-diffusion, suggesting a substitutional diffusion mechanism (Gas *et al.*, 1995).

1.2.4.3 Chemical Reactivity The chemical reactivity of silicides is important in microelectronics applications. When silicides are patterned by dry etching methods a high reactivity is required. However, during subsequent wet or dry etch steps, a low reactivity is required. Low reactivity is also required for silicides used for coatings in corrosive environments.

Vijh (1994) has studied the hydrogen evolution reaction ($H_2O \rightarrow H_2 + \frac{1}{2}O_2$) in a number of silicides; the hydrogen-evolution reaction is of relevance for corrosion and electrolysis reactions. In general, silicides show high chemical and electrochemical stability in electrolyte solutions, especially at cathodic potentials. At anodic potentials, some oxide growth and dissolution of the silicide can occur. In addition, a variety of electrode reactions, including the hydrogen-evolution reaction, can be sustained at high rates on silicide electrode materials. Silicides formed from a catalytic metal (such as PtSi) show higher rates for the hydrogen-evolution reaction in acidic solutions compared with alkaline solutions; the reverse is true for silicides formed from noncatalytic metals (such as Zr).

In microelectronics applications, silicides must withstand a number of wet etches used for cleaning the substrates or for removing excess metal (see Sec. 2.1), including HF, $H_2O_2 + H_2SO_4$, $H_2O_2 + NH_4OH + H_2O$ (SC-1 solution), and $H_2O_2 + HCl + H_2O$ (SC-2 solution). Most commonly used silicides etch slowly if at all in $H_2O_2 + H_2SO_4$ cleans and SC-1 or SC-2 cleans. One exception is $TiSi_2$, which can be etched in $H_2O_2 + NH_4OH + H_2O$ solutions. Another exception is As-doped $CoSi_2$, which can be etched and oxidized in $H_2O_2 + HCl + H_2O$ solutions (Gambino *et al.*, 1990). Most silicides etch to some degree in HF. A common method for patterning silicide thin films formed by metal-Si reactions is to use a wet etch that attacks the nonsilicided metal over the insulator regions, but does not attack the silicide that forms in the contacts (*i.e.*, salicide process; see Sec. 2.1). A number of etches have been developed for $TiSi_2$, $CoSi_2$, NiSi, and PtSi (Gambino and Colgan, 1998).

Reactive ion etching (RIE) is commonly used for patterning codeposited silicide thin films for microelectronics applications (*i.e.*, polycide process, see Sec. 2.2) (Flamm *et al.*, 1984). Anisotropic etching is achieved by using ion bombardment to accelerate the etch reactions. Reactive ion etching can occur if the material to be etched forms volatile compounds with the etch gases. Tungsten, Ti, and Si all form relatively volatile fluorides, chlorides, and bromides, and therefore can be etched with either F-, Cl-, or Br-based chemistries (Chow *et al.*, 1984). In contrast, Ni, Pt, and Co fluorides, chlorides and bromides have high boiling points; hence $NiSi_2$, $CoSi_2$, and PtSi are difficult to pattern by RIE.

1.3 Properties

1.3.1 Mechanical Properties The mechanical properties of silicides, such as creep rate, hardness, yield strength, and brittle–ductile transition temperature, are obviously relevant for structural materials. Because of this, there have been many studies on the mechanical properties of bulk silicides (Kumar, 1994). Mechanical properties are also important for microelectronics applications, where high stresses can be detrimental to underlying devices.

A high value of hardness is required for applications such as coatings, and is also a measure of strength. In general, hardness increases with increasing melting point of the material (Fig. 21), and so most silicides have high values of hardness (Fleischer, 1994; Ostling and Zaring, 1995). The hardness and fracture toughness of silicides can often be increased by alloying (Fig. 22) (Maloy *et al.*, 1991).

The yield strength of a material is the minimum stress that must be applied to induce plastic deformation of the material, and is usually proportional to the elastic modulus (Young's modulus) of the material (Barrett *et al.*, 1973). The yield strengths and elastic moduli for silicides are generally high compared with other materials (Fig. 23) (Shah *et al.*, 1992), Ostling and Zaring, 1995).

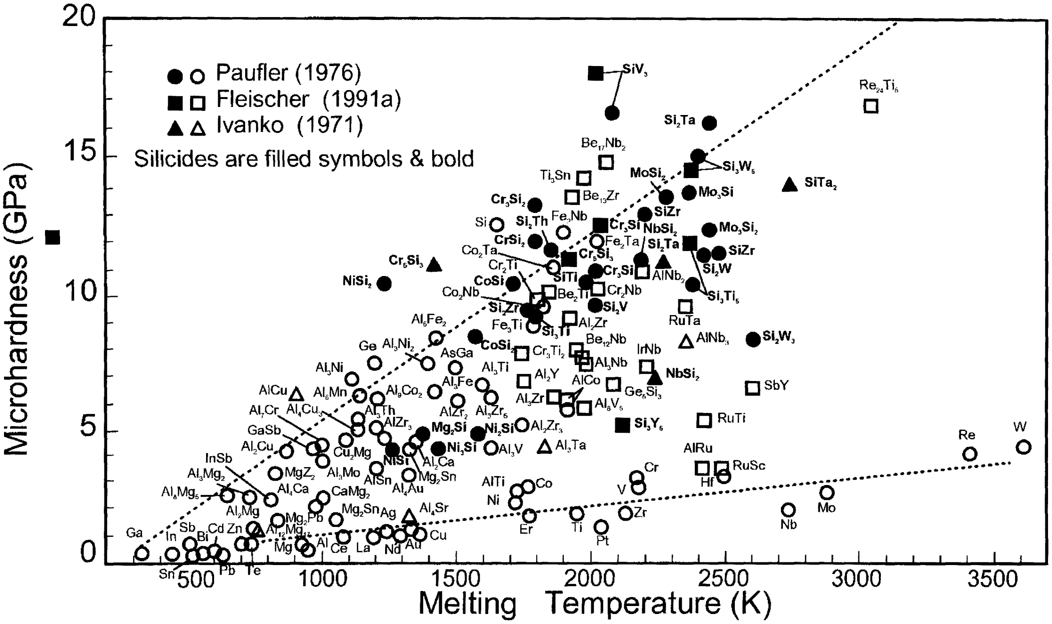

FIG. 21. Hardness at room temperature of metals and intermetallic compounds. (From Fleischer, 1994. Copyright John Wiley & Sons Limited. Reproduced with permission.)

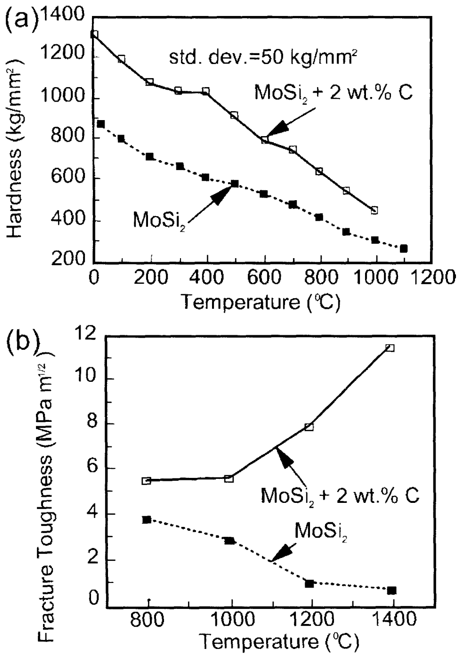

FIG. 22. (a) Microhardness and (b) fracture toughness versus temperature for $MoSi_2$ and $MoSi_2 + 2wt\%$ C. (From Maloy *et al.*, 1991, reprinted with permission of the American Ceramic Society, P.O. Box 6136, Westerville, OH 43086-6136. Copyright 1991 by The American Ceramic Society. All rights reserved.)

The yield strength and ductility of a material are related to the slip systems that are available in the structure for dislocation motion. Slip is difficult in most silicides, because of complicated crystal structures and covalent bonding. The mobility of dislocations increases as the ratio of the lattice spacing to the Burgers-vector length (*i.e.*, the width of the dislocation) decreases. In silicides, the lattice spacing for the slip planes is often large, on account of the complicated crystal structures. In addition, the Burgers vectors are often small, as a result of dissociation of dislocations caused by the highly directional covalent bonds. Hence, dislocations in silicides generally have low mobility. As a result, silicides have high yield strength but low ductility. Silicides are generally brittle at low temperatures, which is a disadvantage for structural materials, because they are weak against fracture. Slip becomes easier as temperature increases, because dislocation motion, which requires an atom at the core of the dislocation to jump a distance on the order of the Burgers vector length, increases with temperature (Hertzberg, 1976).

The temperature of the brittle-to-ductile transition generally increases with increasing

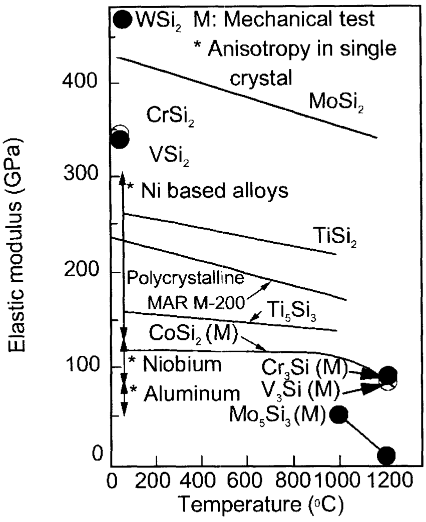

FIG. 23. Comparison of average elastic modulus as a function of temperature (reprinted from Shah *et al.*, 1992, copyright 1992, with permission from Elsevier Science).

melting point of the material (Fig. 24) (Fleishcher, 1994). Because of their low temperature brittleness, silicides by themselves cannot be used as structural materials. However, this problem can be overcome by forming a composite, adding a second phase to the silicide that improves the low temperature ductility. The fracture toughness is a measure of the work expended at a crack tip during crack propagation, and gives a measure of brittleness (*i.e.*, brittle materials have low fracture toughness). The fracture toughness of silicide composites is much greater than that of the silicide alone (Vasudevan and Petrovic, 1992).

The high-temperature (>900 °C) yield strength of MoSi$_2$ is anisotropic (Fig. 25) and is much higher at orientations near [001] compared with other orientations (Umakoshi *et al.*, 1990; Kimura *et al.*, 1990). The dominant high-temperature slip planes in MoSi$_2$ (and also WSi$_2$, which has the same structure) are the {110} and {013} planes, with the {110}

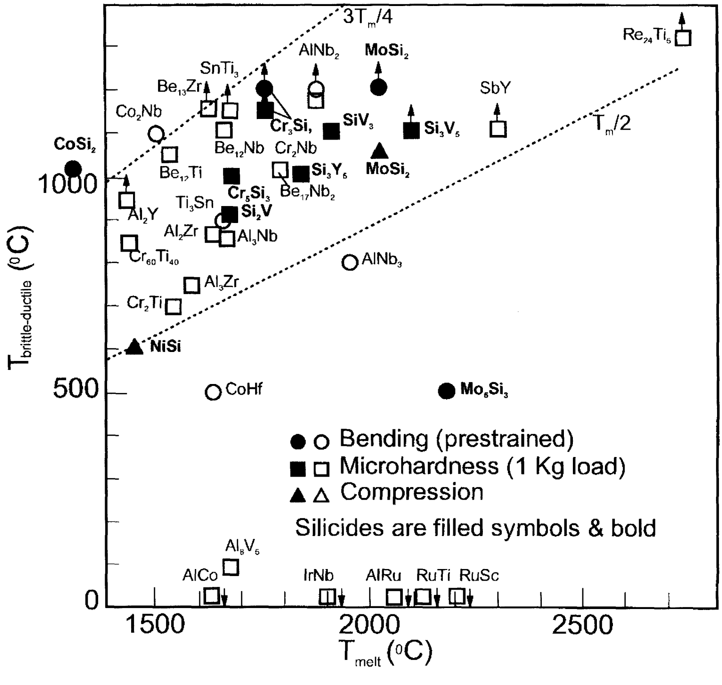

FIG. 24. Brittle-to-ductile transistion temperatures as a function of melting points of the material. Data are from bending, microhardness, and compression experiments. (From Fleischer, 1994. Copyright John Wiley & Sons Limited. Reproduced with permission.)

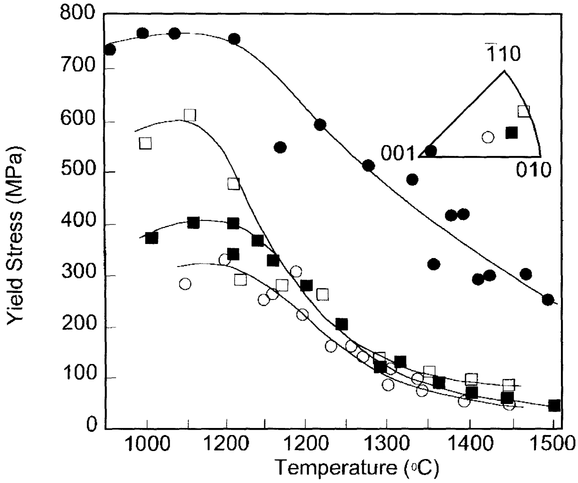

FIG. 25. Temperature dependence of yield strength for several orientations of $MoSi_2$ single crystals (from Umakoshi *et al.*, 1990, copyright 1990, with permission from Elsevier Science).

planes being dominant; the slip direction is ⟨331⟩ in both cases. For $MoSi_2$, the critical resolved shear stress for {013} planes is higher than that for {110} planes, perhaps because the cores of dislocations on {013} planes are nonplanar and spread onto several planes (the {110} planes have a simple ABAB stacking sequence, whereas the {013} planes have a long-period stacking sequence). Orientations near [001] will deform on the [013] slip system, and hence will have a higher yield strength. In contrast, for WSi_2, the critical resolved shear stress for {110} planes is higher than that for {013} planes, perhaps due to a large number of stacking faults on {110} planes.

$CoSi_2$ is more ductile than $MoSi_2$, on account of a more active slip system. $CoSi_2$ has the fluorite structure, and has an fcc crystal symmetry (compared with the more complicated tetragonal $C11_b$ structure of $MoSi_2$). Possible slip systems include {110}, {111}, and {001} planes, all in the ⟨110⟩ direction, as well as {001} planes in the ⟨100⟩ direction. The {111} planes are probably the dominant slip planes at high temperatures (>800 °C), as in fcc metals (Takeuchi *et al.*, 1992). However, at room temperature the {001} planes are the dominant slip planes, in the ⟨100⟩ direction (Ito *et al.*, 1992). Orientations near [100] will deform on the [110] slip system, and hence will have a higher yield strength.

The yield strength of $MoSi_2$ decreases abruptly at high temperatures (between 1100 and 1200 °C, see Fig. 25), indicating an increase in ductility (Umakoshi *et al.*, 1990). The improvement in ductility is due to activation of the {110} and {013} slip systems. At these temperatures, most dislocations have ⟨331⟩ Burgers vectors. At higher tempertures (above 1300 °C), ⟨100⟩ and ⟨110⟩ dislocations are activated, and the ductility is greatly improved.

At high temperatures (greater than half the melting point), the plastic deformation of a material under an applied stress varies with time, independent of whether the applied stress is greater than or less than the yield strength. This deformation is called creep. Many silicides have low creep rates, which are desirable in structural materials (Fig. 26) (Shah *et al.*, 1992). The creep rates of $MoSi_2$ can be further improved by forming composites (Bose, 1992). The activation energy for creep in $MoSi_2$ is high in comparison with lattice diffusion (4.48 eV compared with 2.4 eV for Si diffusion and 3.6 eV for Mo diffusion in $MoSi_2$). Possible creep mechanisms include grain-boundary sliding and dislocation climb at higher applied stresses, and grain-boundary sliding at lower applied stresses (Vasudevan and Petrovic, 1992; Sadananda *et al.*, 1992). Dislocation climb is expected to be limited by diffusion of Mo, which is the slower diffusing species.

1.3.2 Optical Properties The optical reflectivity of metallic silicides can be used to examine the electronic structure such as the

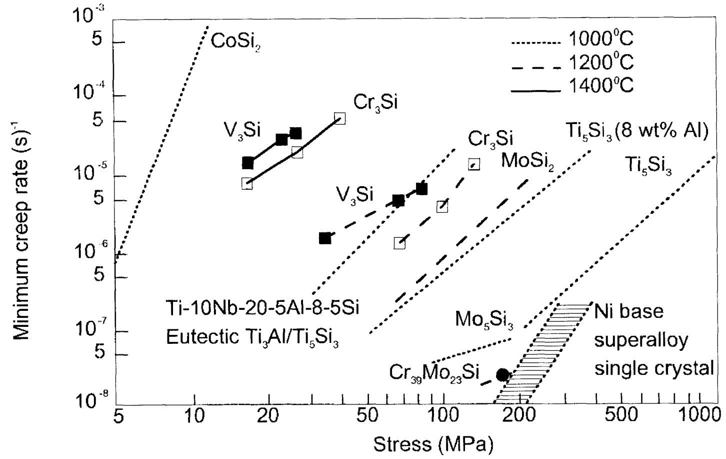

FIG. 26. Comparison of minimum creep rates in compression *versus* stress for several silicides and alloys (reprinted from Shah *et al.*, 1992, copyright 1992, with permission from Elsevier Science).

density of states and character of the bands through the complex dielectric function. The complex dielectric function, $e(\omega) = \varepsilon_1(\omega) + i\varepsilon_2(\omega)$, is related to the complex refractive index, $n = n + ik$ so that $\varepsilon_1 = n^2 + k^2$ and $\varepsilon_2 = 2nk$ and can be calculated from reflectance or other optical data. A recent review (Guizetti and Marabelli, 1995) summarizes information on the complex refractive index for silicides. Examples of the interpretation of the optical response of metallic silicides in terms of the electronic structure can be found in the work of Nava *et al.* (1993). Following Nava *et al.*, for the free-electron region at low photon energies, the classical Drude model (Ashcroft and Mermin, 1976) can be used to determine the plasma frequency (ω_p) and the relaxation time (τ) from the complex refractive index:

$$\varepsilon_1(\omega) = 1 - \omega_p{}^2/[\omega^2 - (1/\tau)^2] \tag{3}$$

$$\varepsilon_2(\omega) = 1 - (\omega_p{}^2/\tau)/\omega[\omega^2 - (1/\tau)^2] \tag{4}$$

The optical resistance at zero frequency, $\rho_{\mathrm{opt}} = 4\pi/\omega_p{}^2\tau$, was also calculated and compared well with the dc resistance measurements. The optical reflectivity and dielectric functions for VSi$_2$, NbSi$_2$, and TaSi$_2$ thin films are shown in Fig. 27 (Nava *et al.*, 1993). At low photon energy, the reflectivity is very high and rapidly decreases for energies around 1 eV. The value of ε_2 drops from a large free-elec-

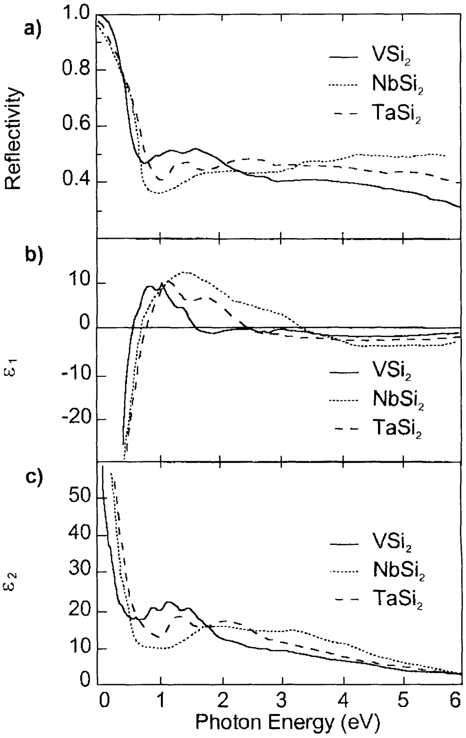

FIG. 27. (a) Optical reflectivity, (b) real part of the dielectric constant, and (c) imaginary part of the dielectric constants for VSi$_2$, NbSi$_2$ and TaSi$_2$ (reprinted from Nava *et al.*, 1993, copyright 1993, with permission from Elsevier Science).

tron value at low photon energy and passes through a minimum around 1 eV.

The optical properties of semiconducting silicides can be used to determine the band gap and whether the transition is direct or indirect. The absorption coeffient, $\alpha = 4\pi k/\lambda$, is calculated from k, the extinction coefficient, where λ is the wavelength. For an indirect transition,

$$\alpha = 0 \quad \text{for } h\nu < E_g - E_p, \tag{5}$$

$$\alpha = A(h\nu - E_g + E_p)^2 \quad \text{for} \quad E_g + E_p > h\nu > E_g - E_p, \tag{6}$$

$$\alpha = A(h\nu - E_g + E_p)^2 + B(h\nu - E_g - E_p)^2 \quad \text{for } h\nu > E_g + E_p, \tag{7}$$

where E_g is the energy gap, E_p is the phonon energy, h is Planck's constant, ν is the frequency of the absorbed light, and A and B are constants. For a direct transition,

$$\alpha = 0 \quad \text{for } h\nu < E_d, \tag{8}$$

$$\alpha = C(h\nu - E_d)^{1/2} \quad \text{for } h\nu > E_d, \tag{9}$$

where E_d is the energy of the direct transition and C is a constant. A good example of a semiconducting silicide with a direct energy gap is β-FeSi$_2$, with a direct band-gap energy of 0.89 eV (Fig. 28) (Bost and Mahan, 1988; Guizetti and Marbelli, 1995).

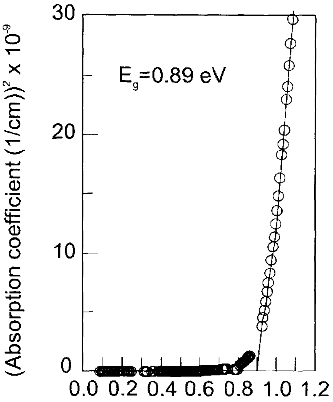

FIG. 28. Plot of the β-FeSi$_2$ absorption coefficient squared versus energy. The solid line is a model based on direct transistions of 0.89 eV and 1.1 eV. (Reprinted with permission from Bost and Mahan, 1988. Copyright 1988 American Institute of Physics.)

1.3.3 Electrical Properties The electrical properities of silicides have been studied extensively, since the resistivity is of critical importance for many applications (Nava *et al.*, 1993; Gottlieb *et al.*, 1995). The electrical transport properties of silicide–semiconductor interfaces have also been widely studied.

1.3.3.1 Resistivity In general, silicides are metallic conductors with room-temperature resistivities on the order of 10^{-6}–10^{-3} Ωcm, but a number of silicides are superconducting or semiconducting (Fig. 29). For the metallic silicides, the resistivity increases with temperature and can be approximated by Matthiessen's rule (Kittel, 1976), where the temperature-dependent resistivity is equal to an electron-impurity scattering component that is temperature independent plus temperature-dependent terms that include electron–phonon scattering, electron–electron scattering, and other temperature-dependent effects. The temperature dependence of the resistivity for a number of silicide single crystals has been modeled by Nava *et al.* (1993), and it was found that in several cases there were well defined temperature ranges in which a specific electron–phonon scattering mechanism dominated.

The resistivity of noncubic single crystals exhibits anisotropy due to the shape of the Fermi surface (Fig. 30 and Table 5) (Nava *et al.*, 1993). The room-temperature anisotropy is nearly 2 for the hexagonal silicides (VSi$_2$, NbSi$_2$, TaSi$_2$ and Pd$_2$Si) and almost undetectable for the higher-symmetry orthorhombic TiSi$_2$.

The semiconducting silicides (Fig. 29) are generally formed from the middle groups of transition metals. As mentioned in Sec. 1.1.1.1, for the semiconducting silicides, a minimum in the density of states at the Fermi level opens into a gap. In moving from NiSi$_2$ to TiSi$_2$, the density of states due to the metal d electrons moves to lower energy, passing through the Fermi level, with increasing atomic number (see Fig. 2). The shift of the minimum in the density of states relative to the Fermi level changes the conduction in the disilicide. The conduction in NiSi$_2$ and CoSi$_2$ is metallic, but the majority carriers are holes (Nicolet and Lau, 1983). The next silicides, CrSi$_2$, MnSi$_{1.7}$,

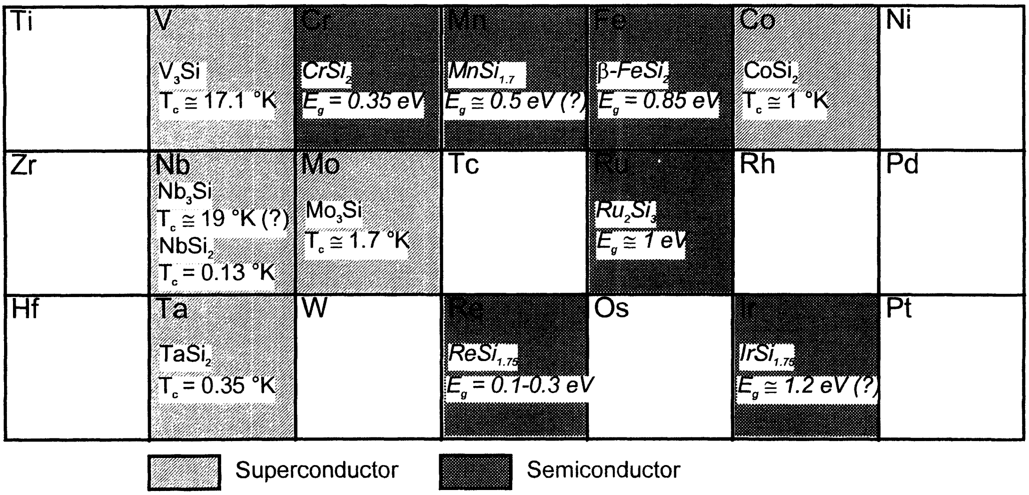

Ti	V V_3Si $T_c \cong 17.1$ °K	Cr $CrSi_2$ $E_g = 0.35$ eV	Mn $MnSi_{1.7}$ $E_g \cong 0.5$ eV (?)	Fe β-$FeSi_2$ $E_g = 0.85$ eV	Co $CoSi_2$ $T_c \cong 1$ °K	Ni
Zr	Nb Nb_3Si $T_c \cong 19$ °K (?) $NbSi_2$ $T_c = 0.13$ °K	Mo Mo_3Si $T_c \cong 1.7$ °K	Tc	Ru Ru_2Si_3 $E_g \cong 1$ eV	Rh	Pd
Hf	Ta $TaSi_2$ $T_c = 0.35$ °K	W	Re $ReSi_{1.75}$ $E_g = 0.1$-0.3 eV	Os	Ir $IrSi_{1.75}$ $E_g \cong 1.2$ eV (?)	Pt

▨ Superconductor ■ Semiconductor

FIG. 29. Superconducting and semiconducting transition-metal silicides (from Gottlieb *et al.*, 1995; Mantl, 1994).

and b-$FeSi_2$, are semiconductors. Finally, $TiSi_2$ is a metallic conductor with electrons as the majority carriers (the carrier type for VSi_2 has not been reported). "Undoped" semiconducting silicides (*i.e.*, no intentional dopants) usually have holes as the majority carriers. They can be made *n*-type by the addition of transition metals with a larger number of d electrons (Co, Ni, Pd, or Pt has been used with b-$FeSi_2$), or *p*-type by the addition of transition metals with a smaller number of d electrons (Mn, Cr, V, or Ti has been used with b-$FeSi_2$), than the metal atom they replace (Lange, 1996). The activation energies for most dopants vary between about 0.08 and 0.14 eV.

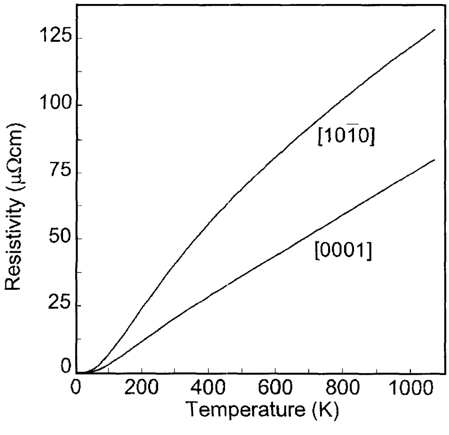

FIG. 30. Temperature dependence of single-crystal $TaSi_2$ resistivity with the current parallel to the [0001] and [10$\bar{1}$0] crystallographic directions (reprinted from Nava *et al.*, 1993, copyright 1993, with permission from Elsevier Science).

There have been relatively few studies on the superconducting properties of silicides, probably because the critical temperature (T_c) is low for most superconducting silicides (see Fig. 29). One exception is V_3Si, which has a relatively high critical temperature (≈ 17 K) compared with other intermetallic superconductors. However, all of the intermetallic superconductors have much lower critical temperatures than oxide superconductors (Bednorz and Müller, 1986), and therefore have been less studied in recent years.

V_3Si, like other intermetallic superconductors with high T_c, has the A15 crystal structure. The Si atoms are arranged on a body-centered cubic (bcc) lattice, and the V atoms are in pairs that straddle the center of a face of the cube. This results in orthogonal chains of closely spaced V atoms running through the crystal structure. It is believed that these chains lead to the good superconducting properties of these materials (Stekly and Gregory, 1994).

1.3.3.2 Ohmic, Rectifying, and Gate Contacts

Metal contacts on lightly doped Si form rectifying contacts (*i.e.*, nonlinear current–voltage

Table 5. Electrical properties of some transition-metal silicides (from Nava *et al.*, 1993).

Silicide	Structure/prototype	Orientation	Resistivity (RT) ($\mu\Omega$ cm)
$TiSi_2$	Orthorhombic/$TiSi_2$	[100]	11.0
		[010]	10.0
		[001]	12.0
VSi_2	Hexagonal/$CrSi_2$	[0001]	34.1
		$[10\bar{1}0]$	59.6
$NbSi_2$	Hexagonal/$CrSi_2$	[0001]	22.0
		$[10\bar{1}0]$	39.2
$TaSi_2$	Hexagonal/$CrSi_2$	[0001]	20.2
		$[10\bar{1}0]$	40.1
$MoSi_2$	tetragonal/$MoSi_2$	[001]	12.6
		[110]	16.9
WSi_2	tetragonal/$MoSi_2$	[001]	12.9
		[110]	11.9
Pd_2Si	Hexagonal/Fe_2P	I $\parallel$ to c-axis	13.7
		I $\perp$ to c-axis	18.9

characteristics, with high current in one polarity and low current for the opposite polarity), which are used to make Schottky diodes. For Schottky diodes, the most important property is the barrier height, which depends on the properties of the metal and the semiconductor surface. Metal contacts on heavily doped Si can be Ohmic (*i.e.*, linear current–voltage characteristics) and are used to make contacts to diffusions. The most relevant parameter for Ohmic contacts is the contact resistance, which depends on the metal and the doping in the Si. For gate contacts to metal–oxide–semiconductor (MOS) transistors, the most relevant parameter is the threshold voltage, which determines when the device "turns on" (*i.e.*, when the MOS transistor conducts current). Silicide or metal contacts on top of polycrystalline Si in polycide structures (see Sec. 2.2) do not directly affect the threshold voltage of the device. However, the silicide can indirectly affect the threshold voltage by depleting dopant from the underlying polycrystalline Si.

The current–voltage characteristics of Schottky diodes are determined by the barrier height at the metal–semiconductor interface. For lightly doped Si at room temperature, the current transport across the interface is dominated by thermionic emission of majority carriers over the barrier (Sze, 1981).

For heavily doped Si, the potential barrier at the metal–Si interface becomes very narrow, and tunneling of electrons from the conduction band of the Si into the metal is the dominant transport mechanism. A critical parameter for characterizing Ohmic contacts is the specific contact resistance, which depends strongly on the both the barrier height and the active dopant concentration (Sze, 1981).

Most Ohmic contacts in microelectronics applications are made with Ti or $TiSi_2$. The barrier height of Ti on Si is between 0.5 and 0.6 eV, depending on the annealing temperature after metal deposition (Aboelfotoh and Tu, 1986). Hence, the resistance for these contacts is mainly dependent on the doping level. In order to minimize contact resistance, the contacts are often doped to above the solubility limit. A lower contact resistance is expected for *n*-type contacts than *p*-type contacts, because the *n*-type dopants such as As and P have a higher solubility than *p*-type dopants such as B (Nobili, 1990). In fact, data on $TiSi_2$ contacts (Hui *et al.*, 1985) show a 2× higher contact resistance for *p*-type contacts compared with *n*-type contacts (Fig. 31).

1.3.4 Magnetic Properties The magnetic properties of some silicides are of interest for permanent magnets (Buschow, 1991). There are very few reports on the magnetic properties of silicides. All the studies on the magnetic properties of silicides involve ferromagnetic compounds, where Si is combined with Fe, Co, or rare earths. One type of ferromagnetic material consists of ternary Fe–Al–Si alloys, which exhibit high magnetic permeability (Kumar, 1994). More recent studies have focused on the magnetic properties of ternary

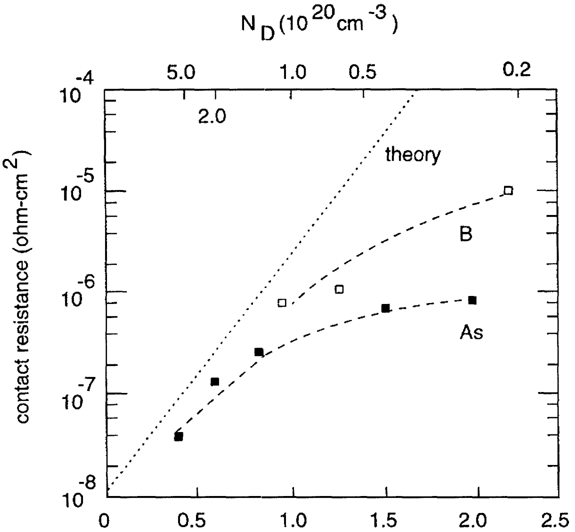

FIG. 31. Specific contact resistance *versus* dopant concentration for $TiSi_2$ on B-doped or As-doped Si (from Hui *et al.*, 1985). Theoretical curve is from Sze (1981), assuming a barrier height of 0.6 eV. (Reprinted from Gambino and Colgan, 1998, copyright 1998, with permission from Elsevier Science.)

silicides with the tetragonal $ThMn_{12}$- or $ThCr_2Si$-type structures (Buschow, 1991). The tetragonal structures provide magnetic anisotropy, which is desirable for permanent magnets. The $ThMn_{12}$ structures also possess high Curie temperatures.

The $ThMn_{12}$ compounds have a composition $RFe_{12-x}Si_x$, where R is a rare-earth element, such as Y, Nd, or Sm (note that other elements can be used instead of Si, including Al, Ti, V, Cr, Mo, and W). The role of the Si is to stabilize the structure, while the rare earths and the Fe provide the permanent magnetic moment. The magnetic properties can be described by combining the $Fe_{12-x}Si_x$ sublattice magnetization with the R sublattice magnetization. For light rare-earth elements, there is a ferromagnetic coupling between the two sublattices, whereas for heavy rare-earth elements there is an antiferromagnetic coupling (Buschow, 1991; Matar *et al.*, 1994).

2. FABRICATION

2.1 Salicide Process

Silicides in Si integrated ciruits are often formed by a solid-state reaction between the metal and Si (Fig. 32). The metal is deposited by physical vapor deposition (see Sec. 2.3) over a patterned wafer, where the pattern consists of insulators (used for isolating de-

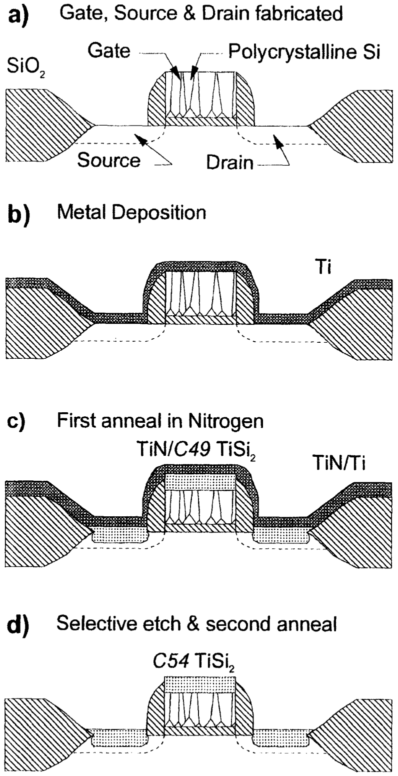

FIG. 32. Schematic of salicide process on an MOS transistor to form $TiSi_2$: (a) starting structure, (b) Ti deposition, (c) first anneal in N_2 to form C49 $TiSi_2$ (and TiN), (d) selective wet etch to remove Ti/TiN and second anneal to form C54-$TiSi_2$. (Reprinted from Gambino and Colgan, 1998, copyright 1998, with permission from Elsevier Science.)

vices) and exposed Si regions (where junctions or gates are to be formed). The wafers are then annealed, with the metal only forming silicide in regions where the Si or polycrystalline Si is exposed. For the appropriate anneal conditions (*i.e.*, sufficiently low temperatures) there is no significant silicide formed in regions where there are insulators, such as Si_3N_4 or SiO_2, since metals react much more slowly with these materials compared with Si. Hence, the silicide is self-aligned to the contacts and gates of the device. The nonsilicided metal is then removed by a wet etch, which etches the metal much faster than the silicide (see Sec. 1.2.4.3). This process is often called a self-aligned silicide or salicide process (Colgan *et al.*, 1996a).

There are a number of advantages to using the salicide process compared with other metallizations. Silicides formed by solid-state reactions often consume contamination and defects at the original metal–Si interface, resulting in low contact resistance for Ohmic contacts and reproducible rectifying behavior for Schottky diodes. Relatively planar metal–Si reactions allow metallization of shallow junctions without degradation of the junction electrical properties (*i.e*, without metal penetrating the junctions and causing high leakage currents). Finally, the self-aligned formation allows a smaller spacing between conductors or contacts than can be achieved with lithography, resulting in higher circuit densities.

There are also problems associated with the salicide process. Lateral silicide formation can result in shorting of contacts (Fig. 33); hence, the anneal temperature and time must be minimized to limit lateral silicide formation. Excessive surface contamination or doping can inhibit silicide formation, resulting in nonuniform silicide thickness and high sheet resistance (see Sec. 1.2.1.3). Nonplanar silicide formation or silicide agglomeration (see Sec. 1.2.3.2) during high-temperature annealing can result in silicide protrusions, junction leakage or high sheet resistance. For silicides on polycrystalline Si, silicide-enhanced grain growth can also occur during high-temperature annealing (see Sec. 1.2.1.3), resulting in threshold-voltage shifts or gate oxide leakage (Fig. 33). Finally, phase transformations (such as C54-$TiSi_2$) can be inhibited on narrow lines, resulting in high sheet resistance (Fig. 34).

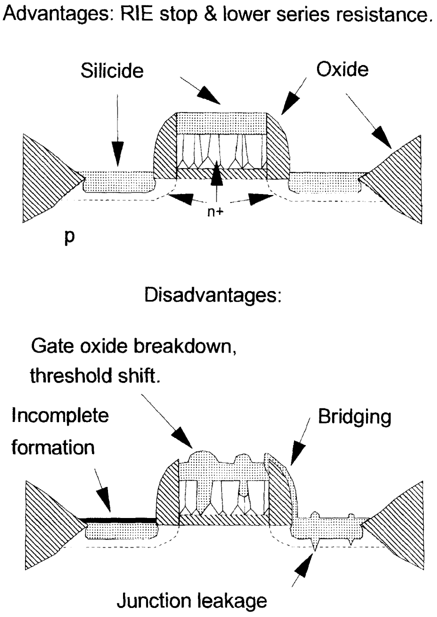
Advantages: RIE stop & lower series resistance.
Silicide
Oxide
n+
p
Disadvantages:
Gate oxide breakdown, threshold shift.
Incomplete formation
Bridging
Junction leakage

FIG. 33. Schematic showing advantages and disadvantages of salicide structure. (Reprinted from Gambino and Colgan, 1998, copyright 1998, with permission from Elsevier Science.)

A number of techniques have been used to avoid the problems described above, including integrated processing, capping layers, preamorphization, ion-beam mixing, epitaxial growth and nucleation enhancement. Integrated processing combines a number of steps, such as the preclean, metal deposition, and silicide formation anneal, in one cluster tool (Nulman, 1990; Aoki *et al.*, 1994). This minimizes the exposure of the Si surface and metal films to air, thereby improving silicide formation. Another approach to prevent oxygen contamination of the metal used in the salicide process (*i.e.*, during exposure to air or the anneal ambient) is to use capping layers such as TiN or Si on top of the metal film (Kaplan *et al.*, 1992; Berti and Bolkhovsky, 1992). For nucleation-limited reactions, such as $TiSi_2$ or $CoSi_2$ formation, amorphization of the substrate can enhance the nucleation of the new phase (Mogami *et al.*, 1994; Kuwano *et al.*, 1990). The free energy of the system is increased, resulting in an increased driving force for the reaction. Amorphization is usually achieved by implantation of Si, Ge, or As prior to metal deposition. Another approach to improve nucleation is ion-beam mixing (Mayer *et al.*, 1981; Maex *et al.*, 1986). In this

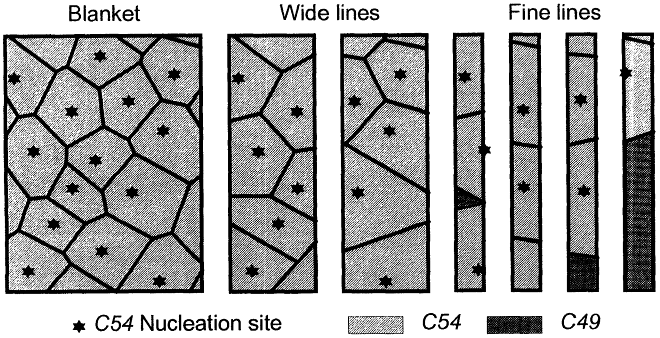

FIG. 34. Schematic of nucleation of C54-$TiSi_2$ as a function of linewidth. For narrow lines, the average distance between nuclei will increase compared with that for blanket films as a result of geometric effects. (Reprinted from Gambino and Colgan, 1998, copyright 1998, with permission from Elsevier Science.)

case, ion implantation is performed after metal deposition, to intermix the metal and Si atoms at the interface. A third approach to improve nucleation is to add impurities that aid in the nucleation of the new phase. This method has been used to improve the nucleation of C54-$TiSi_2$ (Clevenger *et al.*, 1996). The formation of epitaxial silicides has been studied as a method to improve the uniformity of the silicide–Si interface (Hsia *et al.*, 1991; Ogawa *et al.*, 1994; Tung, 1997). Epitaxial growth can be achieved in a salicide process by depositing a thin diffusion barrier (SiO_2 or a refractory metal, such as Ti, Ta, Zr, W) between the metal to be reacted (such as Co) and the Si substrate. If the diffusion barrier is sufficiently thin, the overlying metal can diffuse through it and react with the substrate to form a silicide. The reaction is slowed down, allowing the silicide to grow epitaxially.

2.2 Polycide Process

Silicides in integrated circuits are often formed by codeposition of the silicide, followed by reactive ion etching (RIE) to pattern it (Fig. 35). This process is often used to fabricate MOS transistors (see Sec. 3.1.1) and is referred to as a polycide process. In the polycide process, the polycrystalline Si for the gate is deposited and doped; doping can occur *in situ* during the deposition of the polycrystalline Si or can be done after the deposition by ion implantation (Colgan et al., 1996a). The silicide is then deposited by PVD or CVD (see Secs. 2.3 and 2.4), followed by deposition of an insulator such as SiO_2 or Si_3N_4. The gate stack

a) Deposit gate stack & pattern.

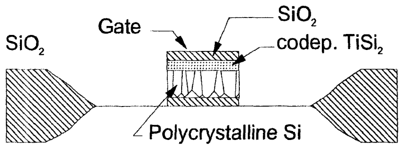

b) Thermal oxidation, fabricate sidewall spacer.

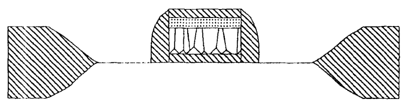

c) Ion implantation & annealing.

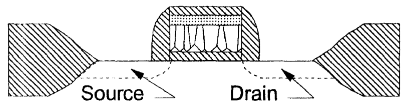

d) Salicide process for Source & Drain.

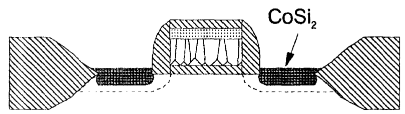

FIG. 35. Schematic of polycide process on an MOS transistor to form $TiSi_2$ on gates: (a) deposit polycrystalline Si, $TiSi_2$, and SiO_2, then pattern using lithography and reactive ion etching, (b) thermally oxidize sidewalls and form sidewall spacers using conformal deposition of an insulator and RIE, (c) form junctions by ion implantation and annealing, and (d) form $CoSi_2$ on the junctions using a salicide process. (Reprinted from Gambino and Colgan, 1998, copyright 1998, with permission from Elsevier Science.)

(*i.e.*, insulator, silicide, polycrystalline Si, and gate oxide) is patterned by lithography and RIE. Next, the structure is exposed to a brief thermal oxidation to remove etch damage at the perimeter of the gate oxide. Insulating sidewalls are formed by conformally depositing an insulator (by CVD), then removing the insulator on horizontal surfaces using RIE. Finally, junctions are formed by ion implantation and annealing. The silicide patterning is a self-aligned process because the silicide is aligned to the polycrystalline Si using no additional masks (Colgan *et al.*, 1996a).

There are a number of advantages to using codeposited silcides compared with other metalizations. Silicon-rich silicides are thermodynamically stable on Si at high temperatures (800–1000 °C) (see Sec. 1.2.1.1), and therefore are compatible with high-temperature processes such as junction anneals and glass-reflow anneals. Silicides are also generally stable during thermal oxidation (see Sec. 1.2.4.1), which is useful during the sidewall oxidation of gates in MOS transistors. The silicide–Si interface is more planar than that resulting from solid state metal–Si reactions, and no Si is consumed. Hence, the formation of silicide protrusions into the polycrystalline Si is reduced.

There are also problems associated with the polycide process. Patterning of some silicides (such as $CoSi_2$, $NiSi_2$, and PtSi) by RIE is difficult because of the relatively low volatility of the metal chlorides and bromides, resulting in incomplete removal of the silicide (see Sec. 1.2.4.3). In addition, the silicide must be defined by lithography; this is not a problem for gate conductors, but is a problem for junctions, where the silicide must be as closely aligned to the Si as possible to avoid problems with leakage or high series resistance. High-temperature annealing can lead to enhanced grain growth in the polycrystalline Si (see Sec. 1.2.3.2), resulting in threshold-voltage shifts or gate-oxide leakage. High-temperature annealing can also lead to dopant diffusion into the silicide and, hence, dopant depletion from the underlying Si, resulting in high contact resistance or threshold-voltage shifts. Finally, oxidation of the silicide can lead to localized consumption and thinning of the polycrystalline Si or oxidation of the metal in the silicide (see Sec. 1.2.4.1).

2.3 Physical Vapor Deposition

Physical vapor deposition (PVD) techniques include evaporation and sputtering, where material is removed from a target by vaporization or ion bombardment and deposited on a substrate. Sputtering (*q.v.*) is often used in integrated-circuit fabrication for depositing elemental metals for silicide reactions or for codepositing the metal and Si, because it is a simple and relatively low-cost method (Table 6). For integrated circuits, sputtering is preferred over evaporation because it is easier to obtain uniform films on large (200-mm diameter) wafers and because the deposition is more conformal over topography on the wafers. The one application where evaporation is used (at least for research) is the epitaxial growth of silicides on Si by molecular-beam epitaxy (MBE). For MBE of silicides (Table 6), metal and Si atoms are evaporated under ultrahigh vacuum on a clean Si surface at high temperatures (von Känel, 1992). PVD methods are also used to make coatings in structural materials (Mashayekhi *et al.*, 1992; Arps, 1996).

In advanced integrated circuits (sub-0.25-µm feature sizes), the metals are often deposited into features with high aspect ratios, such as contact holes. The sputtering process

Table 6. Physical vapor deposition of silicides (poly = polycrystalline silicide, epi = epitaxial silicide).

Silicide	Deposition method	Temperature (°C)	Reference
WSi_2 (poly)	Magnetron sputtering	20, 450	Sundahl *et al.*, 1995
$TiSi_2$ (poly)	Magnetron sputtering	20	Broadbent *et al.*, 1987
$TaSi_2$ (poly)	Magnetron sputtering	20	Goldsmith *et al.*, 1984
$NiSi_2$ (epi)	Molecular-beam epitaxy	250–600	von Känel, 1992
$CoSi_2$ (epi)	Molecular-beam epitaxy	250–600	von Känel, 1992
β-$FeSi_2$ (epi)	Molecular-beam epitaxy	650–700	von Känel *et al.*, 1994

Table 7. Chemical vapor deposition of silicides.

Silicide	Deposition method	Temperature (°C)	Reference
WSi_2 (blanket)	$WF_6 + SiH_4$	300–450	Saraswat *et al.*, 1983
WSi_2 (blanket)	$WF_6 + Si_2H_6$	225–425	Shioya *et al.*, 1987
WSi_2 (blanket)	$WF_6 + SiH_2Cl_2$	510–600	Telford *et al.*, 1993
$TiSi_2$ (selective)	$TiCl_4 + SiH_2Cl_2$	720	Maury *et al.*, 1997
$TiSi_2$ (selective)	$TiCl_4 + SiH_4$	650–950	Southwell & Seebauer., 1997
$TiSi_2$ (blanket)	$TiCl_4 + SiCl_4$	1000–1100	Singheiser *et al.*, 1982
$TaSi_2$ (selective)	$TaCl_5 + SiH_2Cl_2$	670	Carlsson *et al.*, 1990
$MoSi_2$ (blanket)	$MoCl_5 + SiH_4$	550–670	Bobet *et al.*, 1995
$MoSi_2$ (blanket)	$MoF_6 + SiH_4$	140–550	West + Beeson, 1987
β-$FeSi_2$ (blanket)	$Fe_2Cl_6 + SiH_4$	900	Morand *et al.*, 1994
$ReSi_2$ (blanket)	$ReCl_5 + SiH_4$	650–900	Dutron *et al.*, 1995

must be modified to ensure that there is adequate step coverage at the bottom of the hole. One approach is to use a collimator in between the target and the wafers (*i.e.*, collimated sputtering), so that only sputtered atoms with nearly normal incidence reach the substrate. Another approach is to ionize the sputtered atoms (*i.e.*, ionized PVD), so that highly directional deposition of the metal ions can be achieved (Rossnagel, 1995).

2.4 Chemical Vapor Deposition

Chemical vapor deposition (CVD) (*q.v.*) is often used to form WSi_2 in integrated circuits, providing improved step coverage and lower resistivity compared to PVD films (Domenicucci *et al.*, 1997). Recently, there has also been interest in selective CVD of $TiSi_2$ for contacts to Si devices (Maury *et al.*, 1997). Silicides formed by the selective CVD process potentially have lower resistivity and lower Si consumption compared with silicides formed by metal–Si reactions. Another advantage of CVD processes is that the as-deposited silicide is polycrystalline with a relatively large grain size, as opposed to codeposited PVD films which are an amorphous mixture of the metal and Si. CVD has also been used to form silicide coatings on structural materials, by depositing Si or silicides on metal alloys at high temperatures (Singheiser *et al.*, 1982; Nicoll *et al.*, 1979; Southwell *et al.*, 1987).

For CVD of silicides, a metal precursor (WF_6, $TaCl_5$, $MoCl_4$, or $TiCl_4$) is reacted with a Si precursor (SiH_4, Si_2H_6, or SiH_2Cl_2) at elevated temperatures (Table 7). There have been few detailed studies on the kinetics of silicide formation by CVD. Kinetics studies on CVD of WSi_2 and $TiSi_2$ suggest that there are at least two competing reactions that are depositing the metals and/or silicon (Srinivas *et al.*, 1990; Southwell and Seebauer, 1997). In both cases, Si deposition occurs by dissociation of an Si precursor, such as SiH_4, on the Si surface. During CVD of $TiSi_2$, Ti is deposited by adsorption and dissociation of $TiCl_4$. During CVD of WSi_2, W is probably deposited by adsorption and dissociation of WF_6 on the surface of the substrate, analogous to the case of W CVD (Kobayashi *et al.*, 1991; Colgan and Chapple-Sokol, 1992; Hsieh, 1993). In the case of $TiSi_2$ CVD, the rate-limiting step is believed to be the arrival of $TiCl_4$ at the surface, whereas for WSi_2 CVD, the rate-limiting step may be related to the dissociation of SiH_4 at the surface. The byproducts of these reactions are F- and Cl-based species that can etch Si. These byproducts should be minimized to avoid damaging underlying Si devices.

2.5 Ion Implantation

Ion implantation is the introduction of ionized atoms into a substrate with enough energy to penetrate beyond the surface (Seidel, 1983). The technique is widely used for doping semiconductors, but can also be used to form buried polycrystalline or epitaxial silicide layers in Si (Mantl, 1992). Formation of silicides in Si requires implanting a high dose of metal ($>10^{17}\,cm^{-3}$) followed by a high-temperature anneal (>600 °C). Ion implantation has been used to form a variety of thin silicide films, including $CoSi_2$, $TiSi_2$, $NiSi_2$, $CaSi_2$, $CrSi_2$, and $FeSi_2$ (Mantl, 1992; Audet *et al.*, 1992; Omura *et al.*, 1991). Although the technique has been widely studied, it has not been used in manufacturing on account of high cost and residual implant damage in the substrate.

Some unique mechanisms occur when silicides are formed by reacting Si with implanted metals, rather than deposited metals (Mantl, 1992). Silicide precipitates form in the Si during ion implantation. During the early stages of thermal annealing, there is competive growth of the precipitates, with larger precipitates growing at the expense of smaller precipitates (*i.e.*, Ostwald ripening). In order for a continuous silicide layer to form, there must be an interconnected network of precipitates; otherwise, the Ostwald ripening will result in large isolated grains. The coalescence of adjacent silicide precipitates is driven by a reduction of interface energy, and results in planarization of the silicide–Si interface (Mantl, 1992).

2.6 Melt Growth

Growth by solidification of a molten alloy is often used to form structural materials, including silicides. Growth from a melt is also used to form single-crystal silicides for experimental studies, using Czochralski or float zone techniques.

Bulk polycrystalline silicides can be formed by traditional metallurgy techniques such as arc melting and casting (Shah *et al.*, 1992). Novel composites can be fabricated by solidification of melt near the eutectic composition. This technique has been used to fabricate Nb/Nb_5Si_3 composites (Nb_5Si_3 matrix) and SiC/$MoSi_2$ composites ($MoSi_2$ matrix) (Mendiratta *et al.*, 1991; Tilly *et al.*, 1992).

Single crystals of silicides have been grown by either Czochralski or floating-zone techniques (Nava *et al.*, 1993). In both techniques, single crystals are grown on seed crystals from a melt. In the Czochralski method, a single crystal is pulled from a melt of the appropriate material, whereas in the float-zone method, a small region in a rod of polycrystalline material is melted, then resolidified (Ghandi, 1983). A problem with applying these methods to refractory-metal silicides is that Si has a high vapor pressure at the melting point of the silicide, resulting in loss of Si and the formation of metal-rich silicide phases (Lograsso, 1992). Crystal growth is performed in an inert ambient to minimize the concentration of oxygen or nitrogen impurities in the melt. A problem with applying the Czochralski technique to silicides is that there is no crucible material that can contain liquid metal silicides without contaminating them. This problem can be overcome by using an rf-levitated melt, and has been used to grow single crystals of $TiSi_2$, VSi_2, $CrSi_2$, $NbSi_2$, $MoSi_2$, $TaSi_2$, WSi_2, and Pd_2Si (Nava *et al.*, 1993). The floating-zone technique produces purer crystals than the Czochralski technique because no crucible is required, and has been used to grow single crystals of of $TiSi_2$, VSi_2, $CrSi_2$, $CoSi_2$, $MoSi_2$, α-$FeSi_2$, and V_3Si (Nava *et al.*, 1993).

2.7 Sintering

Sintering is often used to form high-temperature structural materials, including silicides and silicide composites (Table 8). Sintering is generally used when melt growth is not practical, because of the high melting point of the materials, or for composites, where a two- (or more-) phase structure is desired. Sintering is a process where particles are pressed into a die that has the shape of the desired part. The compact is then heated at a high enough temperature so that the particles bond together, with the driving force being the reduction in surface energy of the particles (Barrett *et al.*, 1973; Kingery *et al.*, 1976). Although most of the porosity of the original powder is eliminated, some porosity often remains. Some special types of sintering that have been applied to silicides include hot pressing or hot isostatic pressing, where the compact is sintered under high pressure and high temperature, and reaction sintering, where reactive powders (such as a metal and SiC) are mixed and then annealed.

Oxygen contamination must be minimized during the sintering of $MoSi_2$, because the presence of an SiO_2 phase at grain boundaries in the silicide reduces the high-temperature strength (see Sec. 1.3.1). Hence, sintering must

Table 8. Formation of structural silicides. The table includes the powders used, the method (HIP = hot isostatic pressing), the maximum annealing temperature, the density (% of theoretical density), and the reference.

Silicide	Powders	Fabriction method	Temperature (°C)	Density (%)	Reference
$MoSi_2$	Mo + Si	Hot pressing	1500	96.6	Schwarz *et al.*, 1992
$MoSi_2$	$MoSi_2$	Plasma spray	—	97	Castro *et al.*, 1992
$MoSi_2$–Mo_5Si_3	Mo + Si	Hot pressing	1500	97.7	Schwarz *et al.*, 1992
$MoSi_2$–WSi_3	Mo + W + Si	Hot pressing	1500	96.8	Schwarz *et al.*, 1992
$MoSi_2$–SiC	$MoSi_2$ + SiC	Hot pressing	1900	—	Sadananda *et al.*, 1992
$MoSi_2$–SiC	Mo_2C + Si	Hot pressing	1700	—	Henager *et al.*, 1992
$MoSi_2$–SiC	$MoSi_2$ + C	Hot pressing	1325	—	Maloy *et al.*, 1991
$MoSi_2$–SiC	$MoSi_2$ + SiC	Plasma spray	—	—	Tiwari *et al.*, 1992
$MoSi_2$–Al_2O_3	$MoSi_2$ + Al_2O_3	HIP	1500	—	Alman *et al.*, 1992
$MoSi_2$–Al_2O_3	$MoSi_2$ + Al_2O_3	Plasma spray	—	—	Alman *et al.*, 1992
$MoSi_2$–ZrO_2	$MoSi_2$ + ZrO_2	Hot pressing	1700	95–96	Petrovic *et al.*, 1992
$MoSi_2$–CaO	$MoSi_2$ + CaO	Hot pressing	1700	95	Gibala *et al.*, 1992
$MoSi_2$–TiC	$MoSi_2$ + TiC	Hot pressing	1700	95	Gibala *et al.*, 1992
$MoSi_2$–TiB_2	$MoSi_2$ + TiB_2	Plasma spray	—	—	Tiwari *et al.*, 1992
$MoSi_2$–Nb	$MoSi_2$ + Nb	HIP	1350	—	Alman *et al.*, 1992
$MoSi_2$–Ta	$MoSi_2$ + Ta	Plasma spray	—	—	Castro *et al.*, 1992

take place in an oxygen-free ambient, such as Ar. In addition, the starting powders must be free of oxygen contamination. This can be achieved by mixing powders in an oxygen-free environment (Schwarz *et al.*, 1992).

Reaction sintering (also referred to as solid-state displacement reactions) uses diffusional phase transformations to react two or more elements or compounds to form new compounds that are more thermodynamically stable (Hanager *et al.*, 1992). Interwoven or dispersed microstructures can be fabricated with this method, making it especially useful for making composites.

2.8 Plasma Spray Deposition

Plasma spray deposition can be used to deposit large thicknesses of materials (>30 μm) on a substrate, and has been used to form silicides and silicide composites for structural materials (Table 8). Plasma spray deposition uses a thermal plasma (plasma flame) formed by an electric arc. A powder of the material to be deposited is fed into the flame in a carrier gas (typically Ar). The injected powder accelerates, melts, and is carried to the substrate, where it solidifies rapidly, forming a thick coating (Herman, 1988). Some of the advantages of using plasma spraying for silicides include dense, oxide-free deposits, small grain size, and the possibility of making laminated composites (Tiwari *et al.*, 1992; Castro *et al.*, 1992; Alman *et al.*, 1992).

3. APPLICATIONS

3.1 Microelectronics

3.1.1 MOS Technology Most Si integrated circuits are fabricated by use of complementary metal–oxide–semiconductor (CMOS) technology. The advantages of CMOS technology compared with bipolar technology are lower power consumption, higher circuit densities, and lower manufacturing costs (Ning, 1992; Masaki, 1992).

Silicides are often used on the gates and diffusions of MOS transistors to reduce the series resistance of the device, and hence increase the circuit speed. In memory chips such as dynamic random-access memories (DRAMs), silicides are mainly used on the word lines (gates) and bit lines (first interconnect level) (Fig. 36) (Winnerl, 1994). For these devices, there is little performance gained by using silicides on diffusions since the series resistance of the device is usually dominated by the contact to the storage node. In addition, silicides on the diffusions increase cost and potentially

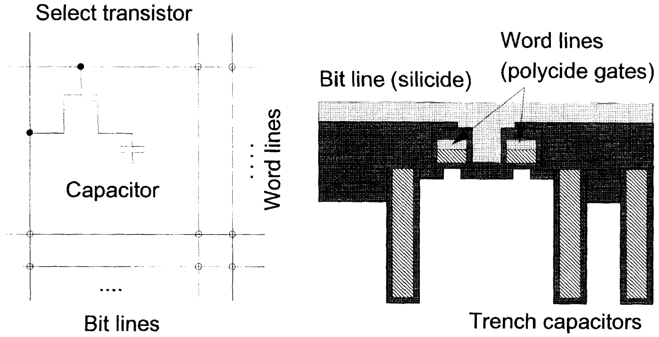

FIG. 36. Schematic of DRAM with silicide on word lines and bit lines (reprinted from Gambino and Colgan, 1998, copyright 1998, with permission from Elsevier Science.)

could decrease yield (through junction leakage associated with metal contamination). Polycides are generally used on the gates because there is a large gain in performance (gates run along the length of the array and so low resistance is required to minimize the *RC* delay) and there is less extra processing compared with a salicide process. In addition, polycide structures can be easily capped with an insulator such as Si_3N_4, which is useful for fabricating borderless contacts (a 30–40% increase in density can be achieved by using contacts that are borderless to gates and isolation; see Gambino *et al.*, 1995). Polycides are often used for bit lines as well, with the polycrystalline Si underlayer serving to act as a contact to diffusions. If lower resistance is required, a W bit line could be used instead of a polycide, with either polycrystalline Si or W studs as contacts to gates and diffusions. The most commonly used silicide in polycide structures is WSi_2, because it has good high temperature stability, is relatively easy to pattern by RIE, and can be deposited by either PVD or CVD.

In logic chips, silicides are often used on source/drain junctions as well as gates and local interconnects (Fig. 37). The improvement in performance is great enough to justify the extra cost and yield loss. In addition, contacts that are borderless to gates are seldom used in logic devices, and so insulator-capped polycide structures are not required. Silicided local interconnects can be fabricated at the same time as the gate and source/drain silicides. Alternatively, W local interconnects can be fabricated along with W studs for contacts. For logic applications, the silicide is usually formed by a salicide process using $TiSi_2$. $TiSi_2$ has a number of advantages compared with other silicides, including low resistivity, relatively high thermal stability, and relatively low sensitivity to interfacial contamination (because Ti reduces native oxides).

As with DRAMs, the switching speed of the device is increased by lowering the resistance of the gates (Fig. 38). The silicide on the source/drain junctions increases the drive current of the MOS transistor (Fig. 39). The silicide reduces the voltage drop at the source end of the device, which increases the effective applied gate potential. The reduced voltage drop is mainly the result of a reduction in the effective contact resistance, due to an increase in the contact area (Ng and Lynch, 1987). The low sheet resistance of the silicide is important for devices that have long lengths of diffusions and for wide devices that have a low channel resistance (Osburn *et al.*, 1991). Silicides can

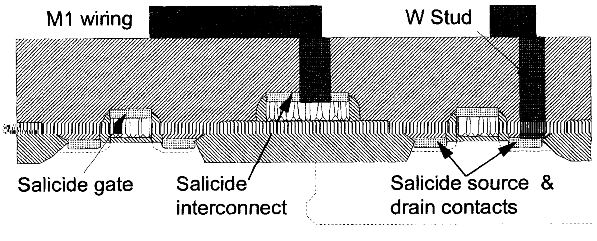

FIG. 37. Schematic of logic device showing silicides on gates and diffusions (reprinted from Gambino and Colgan, 1998, copyright 1998, with permission from Elsevier Science.)

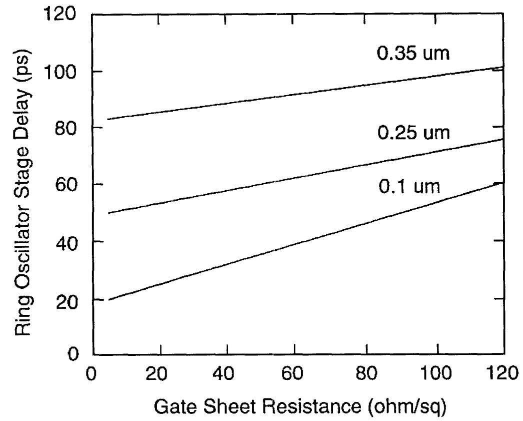

FIG. 38. Performance of ring oscillator as a function of gate resistance (reprinted with permission from Mann *et al.*, 1995).

also be used to increase the packing density of devices. For example, butting n^+ and p^+ diffusions can be connected by a silicide rather than the first metal level. The area of the diffusions can be reduced, since an extra contact hole can be eliminated.

Although silicides can enhance device performance, they can also degrade yield. Junction leakage can be increased (Osburn, 1990; Maex, 1993; Nauka *et al.*, 1986), and gate-oxide breakdown voltage decreased (Koburger *et al.*, 1982; Gambino *et al.*, 1992; Tanielian *et al.*, 1985), if too much Si or polycrystalline Si is consumed in the junctions or the gates, respectively. Interface state densities can be increased by rapid thermal anneals (Hsieh *et al.*, 1993; Liu *et al.*, 1995) used to form silicides or due to gettering of hydrogen (Chang and Chiu, 1989; Hirade *et al.*, 1995) by Ti during post-metallization anneals. Dopant depletion or counterdoping can occur during high-temperature anneals after silicide formation, resulting in high contact resistance (Scott *et al.*, 1987; Chittipeddi *et al.*, 1993) or threshold-voltage shifts (Matsuoka *et al.*, 1994; Hillenius *et al.*, 1986). Finally, excessive lateral silicide formation in a salicide process can cause leakage between junctions and gates.

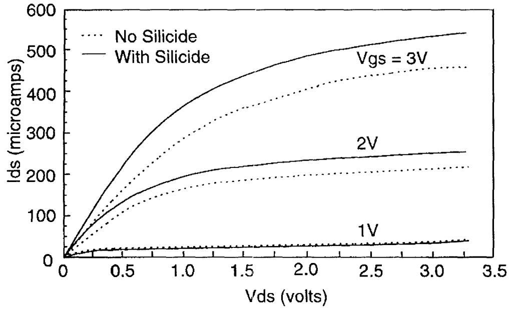

FIG. 39. Device characteristics with and without silicide on source/drain junctions (reprinted from Osburn *et al.*, 1991, copyright 1991, with permission from Elsevier Science).

3.1.2 Bipolar Technology The main use of silicides in bipolar transistors is as Schottky diodes. There have also been proposals to use silicides on base contacts (Fig. 40) or collector contacts, but these processes are not currently used in manufacturing (Shiba *et al.*, 1996; Yallup *et al.*, 1995).

Schottky diodes are used as clamping diodes between the base and collector to prevent excess minority-carrier charge storage that would limit switching speed (Sze, 1981; Tada and Laraya, 1967). PtSi is the most commonly used silicide for bipolar transistors (Li *et al.*, 1987), on account of a high barrier height and ease of patterning by a salicide process. Some problems associated with PtSi formation on bipolar transistors are incomplete silicide formation (Moy *et al.*, 1988), because Pt does not reduce native oxides, and nonuniform reactions between Pt and polycrystalline Si, which can cause junction leakage (Gambino *et al.*, 1989).

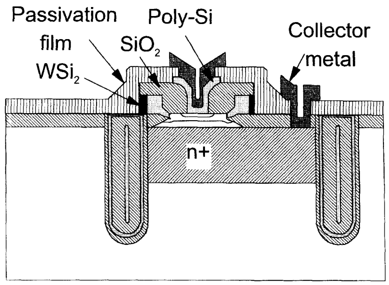

FIG. 40. Schematic of bipolar transistor with silicide on base contact to reduce series resistance (reprinted with permission from Shiba *et al.*, 1996, © 1996 IEEE).

3.1.3 Infrared Detectors Another application for silicide Schottky diodes is as infrared detectors (Kimata and Tubouchi, 1995). Schottky diodes have a number of advantages compared with other infrared detectors, including monolithic construction using standard Si processing and high uniformity. The silicides for Schottky diodes are formed by a salicide process.

The operation of Schottky-diode photodetectors is based on internal photoemission (Kimata and Toubouchi, 1995; Jimenez, 1996). In the first step (Fig. 41), incident photons are absorbed in the silicide and generate electron–hole pairs. Some of the photoexcited holes are then emitted over the Schottky barrier, becoming detected carriers. Hence, the Schottky barrier height defines a minimum energy or cutoff wavelength for detection. Because the energy of infrared radiation in the atmospheric window is in the range of 0.1–0.4 eV, a low Schottky-barrier height is required; this can be most easily achieved on p-type Si.

The quantum efficiency of Schottky-diode infrared detectors depends on the photon energy, because of the strong dependence of photoemission on the energy of the excited holes (Kimata and Tubouchi, 1995). The quantum efficiency η is given by

$$\eta = C_1(h\nu - \phi_B)^2/h\nu \tag{10}$$

where C_1 is the quantum efficiency coefficient, h is Planck's constant, ν is the wavelength of the incident radiation, and ϕ_B is the barrier height. Hence, the Schottky-barrier detector has no response to photons with energy lower than the barrier height.

There are five silicides used for Schottky-barrier infrared detectors: Pd_2Si, PtSi, IrSi, Co_2Si, and NiSi (Table 9 and Fig. 42). PtSi on *p*-type Si is most commonly used, on account of a spectral response to radiation in the range 3–5 μm. IrSi on *p*-type Si is expected to have the lowest barrier height, and therefore should be able to detect radiation at the longest wavelengths (8.2 μm).

Table 9. Properties of silicides used in infrared detectors (from Kimata and Tubouchi, 1995).

Silicide	Silicon	Barrier height (eV)	Cutoff wavelength (μm)
Pd_2Si	*n*-type	0.6	1–2
$CoSi_2$	*p*-type	0.44	2.8
NiSi	*p*-type	0.40	3.1
Pd_2Si	*p*-type	0.34	3.5
PtSi	*p*-type	0.22	6
IrSi	*p*-type	0.152	8.2

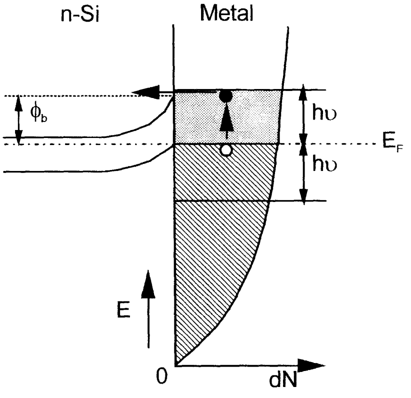

FIG. 41. Schematic of internal photoemission in a Schottky-barrier infrared detector (used with permission from Kimata and Tubouchi, 1995).

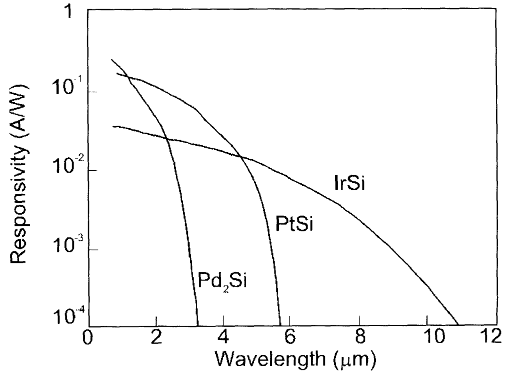

FIG. 42. Plot of $(\eta h\nu)^{1/2}$ *versus* $h\nu$ for a Schottky-barrier infrared detector. The gradient of the plot gives the square root of the quantum efficiency while the intersection with the *x* axis gives the barrier height (used with permission from Kimata and Tubouchi, 1995).

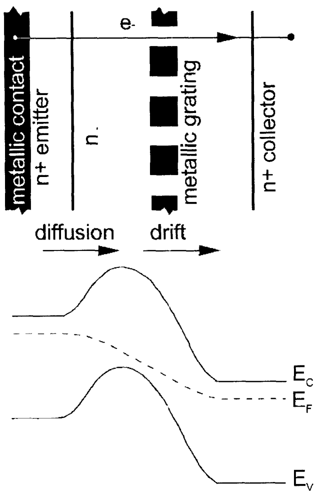

FIG. 43. Schematic cross section of permeable-base transistor with corresponding energy diagram (from Reader *et al.*, 1992).

3.1.4 Permeable-Base Transistors In permeable-base transistors (Reader *et al.*, 1992), a metal grid is used to modulate current flow from an emitter to a collector (Fig. 43). The metal grid is embedded in an n region, resulting in a Schottky diode and a depletion layer around the metal. The extension of the Schottky diode is controlled by the voltage on the metal grid, which in turn controls the flow of carriers through the grid.

There are a number of advantages of permeable-base transistors compared with bipolar or MOS transistors. The main advantage is that the width of the control electrode is determined by the thickness of the metal layer, rather than by lithography (as is the case in MOS transistors), so that the transit time for charge carriers can be very fast, resulting in high switching speeds. The main problem with this device is the difficulty of fabricating buried silicide layers in defect-free Si. Defects in the Si associated with the buried silicide can cause high leakage currents. Because of these problems, permeable-base transistors are still in the development stage and are not being manufactured.

3.2 Structural Materials

Silicides are being considered for a number of applications as structural materials in the aerospace and automotive industries. Aerospace applications include aircraft-engine turbine components, such as blades, vanes, combustors, nozzles, and seals. Automotive applications include turbocharger rotors, valves, glow plugs, and advanced turbine engine parts (Petrovic, 1993).

Structural materials can be broadly classified into two groups of applications, depending on the use temperature. For applications in the range of 800–1000 °C, nickel-based superalloys are being used. Above 1000 °C, silicon-based ceramics are being considered, because of their excellent oxidation resistance and low density. However, these materials are brittle over the entire temperature range, and the cost of fabrication is high. Hence alternative materials are being investigated for high-temperature structural materials, including silicides (Vasudevan and Petrovic, 1992). The use of high-temperature materials can provide a number of advantages, including minimized air cooling, fewer engine parts, and higher operating temperatures (which can improve fuel efficiency).

Silicides have a number of advantages compared with other structural materials, including higher melting points and lower densities compared with other intermetallics, and lower costs and greater ductility (at high tempertures) compared with Si-based ceramics. In addition, silicides can be alloyed with other materials such as SiC or other silicides, to improve their properties. This is important because the mechanical properties of the silicides by themselves are inadequate for these applications; improvements are required in the high-temperature strength and creep resistance, and in the low-temperature fracture toughness. Composite reinforcement of a silicide matrix can significantly improve these mechanical properties (see Figs. 19–21). These composites can be fabricated using a wide variety of techniques (Table 8), including melt casting and sintering.

Composites are classified into the following categories based on the morphologies, orientations, and distributions of the strengthening or toughening components in the intermetallic (silicide) matrix: discontinuous reinforcements, continuous reinforcements, and hybrids (Miracle and Mendiratta, 1994). In discontinuous composites, an intermetallic matrix contains particulates, whiskers, or chopped fibers, such as SiC whiskers, in an $MoSi_2$ matrix. At high temperatures, where there is significant plasticity of the silicide matrix, the morphology and aspect ratio of the discontinuous reinforcements strongly influences the strength and creep resistance of the composite; needle-like reinforcements with a high aspect ratio provide the greatest strengthening and toughening. Continuous composites contain continuous fibers or filaments, typically distributed in a uniaxial direction in a silicide matrix. Fibers with high strength and stiffness enhance the strength and creep resistance of the silicide. Since the mechanical properties are highly anisotropic, applications as structural materials are limited to components where transverse properties are not critical (*e.g.*, turbine blades, ring components, and structural rods). Both refractory-metal and ceramic fibers have been used to reinforce $MoSi_2$ (Maloney and Hecht, 1992). The refractory-metal fibers are ductile, and therefore increase the damage tolerance of the brittle silicide matrix. Although the ceramic fibers are brittle, the composite can have improved fracture toughness if there is a weak fiber–matrix interface; a weak interface when present in the path of a crack will debond locally and blunt the crack (Shaw *et al.*, 1995). Hybrid composites incorporate multiple types of reinforcement, designed to enhance specific mechanical properties. For example, continuous uniaxial fibers may primarily enhance high-temperature strength and creep resistance, whereas a dispersion of whiskers in the matrix may enhance toughness and transverse mechanical properties.

There are a number of other considerations in the fabrication of composites (Maloney and Hecht, 1992). The thermal expansion coefficient must be closely matched between the fiber and the matrix, in order to minimize thermal stress. For $MoSi_2$, sapphire fibers are closest in terms of thermal expansion coefficient, whereas SiC is significantly lower. This problem can be alleviated by adding a phase with a low thermal expansion coefficient to the silicide matrix (such as SiC platelets or particles). Another consideration when using refractory metal fibers is the reactivity of the metal with the silicide. Minimal reactivity is desirable, so that the refractory metal remains a metal, rather than being converted into a (brittle) silicide. For $MoSi_2$, Ta exhibits the lowest reactivity compared with other refractory metals.

3.3 Oxidation-Resistant Coatings

Silicides have been used as oxidation-resistant coatings on a number of materials, including Ni-based alloys, steels, and refractory metals and alloys (Kumar, 1994). The SiO_2 layers that grow on silicides in oxygen ambients show a low permeation rate to oxygen compared with other materials, especially above 1500 °C. The SiO_2 layers are also stable in the alkaline sulfates that are often found in gas turbines. More recently, silicides have been examined as coatings to protect Ni-based superalloys from carburization in high-temperature reactors (Singheiser *et al.*, 1982).

Silicides that are used as coatings must be "matched" to the underlying substrate, to prevent degradation by cracking, spalling, or interdiffusion. The brittleness of silicides and the thermal expansion mismatch between the coating and the substrate can lead to cracking and spalling during thermal cycling. Alloying can be used to adjust the thermal expansion mismatch of the protective SiO_2 layer, and hence reduce cracking. B_2O_3 or GeO_2 additions to the coating not only decrease the thermal expansion mismatch but also decrease the viscosity of the SiO_2 layer, making it more resistant to cracking. Interdiffusion between the substrate and the coating can alter the protective properties of the coating, by depleting Si from the silicide. Metal-rich silicides generally oxidize much faster than silicon-rich silicides, resulting in a less effective coating (Kircher and Courtright, 1992). Interdiffusion can be controlled by using an appropriate diffusion barrier between the substrate and the coating; for example, a TiC diffusion barrier can be used between a graphite substrate and an $MoSi_2$ coating (Grunling and Bauer, 1982).

Most of the applications for silicide coatings involve protecting the underlying material from oxidation. Some examples include Ni–Cr silicide coatings on Ni-based superalloys, Cr–Ti silicide coatings on Nb-based alloys, and $MoSi_2$ coatings on refractory metals. Silicide coatings have been fabricated with a wide variety of methods, including PVD, CVD, plasma spray deposition, or diffusion reactions (pack coating) (Nicholls and Stephenson, 1994).

3.4 Thermoelectric Materials

One of the earliest applications for silicides was as heating elements in high-temperature furnaces (Kumar, 1994). More recently, semiconducting silicides have been investigated as thermoelectric devices, for converting heat into electricity (Verdnikov, 1994).

Furnace elements made from silicides consist of $MoSi_2$ with about 20 vol% of a glassy phase. $MoSi_2$ has a number of advantages as a heating element compared with other materials. First, the resistivity of $MoSi_2$ increases with temperature; if a new batch of material is introduced into the furnace (causing the temperature to drop), the power output of the heating elements rises rapidly and automatically compensates. In addition, the heating elements will accept high power loads so that large amounts of power can be concentrated in a small area. Finally, a protective SiO_2 coating forms on the silicide, resulting in a long life span for the heating elements, even in oxidizing ambients.

Thermoelectric devices convert heat into electricity, and are often used in deep-space vehicles with a radioactive heat source. At present, these devices are made from SiGe alloys. However, several semiconducting silicides are promising as thermoelectric devices. The themoelectric figure of merit for $MnSi_{1.7}$ is comparable to that for SiGe alloys, and calculations suggest that the figure of merit for Ru_2Si_3 may be higher than that for SiGe alloys (Vedernikov, 1994; Vining, 1992).

3.5 Magnetic Materials

There have been two different approaches to using silicides as magnetic materials. One approach is to add Si to an Fe-based alloy to improve the wear resistance and corrosion resistance of the alloy. Ternary Fe-Al-Si alloys are an example of this approach, and have been used in magnetic heads (Kumar, 1994). The other approach is to use Si-stabilized Fe-based compounds for permanent magnets (see Sec. 1.3.4). These materials have not yet been used commercially.

3.6 Optical Materials

Silicides have been used as optical materials in attenuated phase-shift masks. Phase-shift masks are used to improve the resolution of optical lithography during fabrication of integrated circuits. An attenuated phase-shift mask allows some transmission of light through the "opaque" regions of the mask, thereby changing the phase and improving the resolution of the image (Hashimoto *et al.*, 1994). Recently, MoSi alloys have been used as the "opaque" material in attentuated phase-shift masks, instead of Cr. The MoSi alloys are easier to pattern than Cr, thus making it easier to fabricate defect-free masks (Jonckheere *et al.*, 1994).

4. SUMMARY

This article has reviewed the properties, processing, and applications of silicides. The combination of high thermal stability, high mechanical hardness, and low electrical resistivity has made the use of silicides attractive for a wide range of applications. Already, there is widespread use of silicides in the microelectronics industry. With continued studies on these materials, it is likely that silicides will also be used in other applications such as structural materials.

There are many areas of study that could be useful in terms of improving these materials. A better understanding of the atomic and electronic structures of interfaces between silicides and semiconductors could be useful for microelectronics applications. Much has been learned about metal–Si reactions, but further work is required to understand the effects of nucleation, impurities, and grain boundaries on these reactions. More work is needed to understand reactions between metals and Si compounds, for both microelectronics and structural materials applications. More details on the surface and interface energies would be useful in understanding the thermal stability of silicide thin films. Information on the effect of defects on oxidation of silicides could be used to improve the properties of silicide coatings. Detailed studies of the phase equilibria between dopants and silicides could be useful for creating new structural materials. Much information is still needed to understand the etching and passivation reactions for silicides in aqueous solutions. Additional studies are needed on the basic mechanical properties of silicides, as well as processes for improving these properties. More work is also needed to improve the electrical properties of silicide thin films used in microelectronics applications. The initial work on the optical and magnetic properties of silicides shows that there may be new opportunities to use silicides for these applications.

Because of the many potential applications, there are likely to be many new studies on silicides in the coming years.

ACKNOWLEDGEMENT

The authors thank F. M. d'Heurle and J. M. E. Harper for helpful discussions.

GLOSSARY

Agglomeration: The breaking up of a thin film on a substrate into islands, due to a reduction in surface and interface energies.

Burgers Vector: The unit slip distance for the dislocation, parallel to the direction of slip.

Composite: A polyphase material with a large fraction of a minor phase (generally at least 10%).

CVD: Chemical vapor deposition.

DRAM: Dynamic random-access memory.

Electron Affinity: The energy released when an electron is attached to an atom or molecule.

Epitaxy: Growth of a thin layer of material with the same orientation and crystal structure as the underlying single crystal substrate.

Eutectic: A solid solution of two or more components (eutectic composition), having the lowest freezing point (eutectic temperature) of all possible mixtures of the components.

Ferromagnetic: Materials that have a permanent magnetic field even in the absence of an applied field.

Fracture Toughness: Work expended at a crack tip during crack propagation.

Fermi Level: The energy at which the probability of finding an electron in an energy band in a solid is 1/2.

MBE: Molecular beam epitaxy.

MOS Transistor: Metal–oxide–semiconductor transistor.

Nucleation: The initial formation of a new phase within a material or at an interface or surface.

Phase: A homogeneous, physically distinct, and mechanically separable portion of a material with a given chemical composition and structure.

Phase Diagram: A phase equlibria diagram between two or more components (binary = two components, ternary = three components) at a constant pressure, with temperature and composition as variables.

Phonons: Lattice vibrations in a crystal.

Photoemission: Emission of electrons as a result of bombardment by photons.

Polycide: A self-aligned silicide process used in microelectronics applications, where the silicide is deposited on a polycrystalline Si gate stack and patterned by RIE.

PVD: Physical vapor deposition, such as evaporation or sputtering.

Raman Scattering: A scattering effect where the scattered light differs in wavelength from the incident light.

Refractory Materials: Materials that have a high melting point (>1538 °C, the melting point of iron).

RIE: Reactive ion etching; a directional dry etching method used in microelectronics processes.

Salicide: A self-aligned silicide process used in microelectronics applications, where the silicide is formed selectively on Si using a metal–Si reaction and a selective etch.

Schottky Barrier: Potential barrier for electronic carriers at a metal–semiconductor or metal–insulator interface.

Sintering: Compression of particles into a solid body using heat and pressure, at a temperature below the melting point of the solid.

Slip Plane: The plane on which a dislocation moves. A slip system describes all the different slip planes in a given crystal structure.

Thermoelectric Effect: The generation of electrical currents in a material due to a temperature gradient in the material. Also, the production or absorption of heat when electronic carriers are forced across a temperature gradient in a material.

Threshold Voltage: The voltage that must be applied to the gate of an MOS transistor to "turn on" the transistor.

Transition Metals: The metals between II and III in the periodic table; the subgroups starting with Sc and ending with Zn.

Work Function: The difference in energy between the Fermi level of a solid and the vacuum energy.

Yield Strength: The minimum stress that must be applied to induce plastic deformation in a material.

Works Cited

Aboelfotoh, M. O., Tu, K. N. (1986), *Phys. Rev. B.* **34**, 2311–2318.

Aldrich, D. B., Chen, Y. L., Sayers, D. E., Nemanich, R. J., Ashburn, S. P., Ozturk, M. C. (1995), *J. Mater. Res.* **10**, 2849–2863.

Aldrich, D. B., d'Heurle, F. M., Sayers, D. E., Nemanich, R. J. (1996), *Phys. Rev. B* **53**, 16279–16282.

Alman, D. E., Shaw, K. G., Stoloff, N. S., Rajan, K. (1992), *Mater. Sci. Eng.* **A155**, 85–93.

Aoki, A., Yamaguchi, H. Saito, T., Otsuka, F., Owada, N. (1994), in: *Proceedings of the Eleventh International VLSI Multilevel Interconnection Conference* (*VMIC*), Santa Clara, CA: Eleventh International VLSI Multilevel Interconnection Conference.

Appelbaum, A., Knoell, R. V., Murarka, S. P. (1985), *J. Appl. Phys.* **57**, 1880–1886.

Arps, J. H., Page R. A., Dearnaley, G. (1996), *Surf. Coat. Technol.* **84**, 579–583.

Ashcroft, N. W., Mermin, N. D. (1976), *Solid State Physics*, New York: Holt, Rinehart and Winston, Chap. 1.

Audet, S. A., White, A. E., Short, K. T., Hsieh, Y.-F., Ross, F. M., Rafferty, C. S. (1992), *Appl. Phys. Lett.* **61**, 2311–2313.

Barbour, J. C., Kuiper, A. E. T., Willemsen, M. F. C., Reader, A. H. (1987), *Appl. Phys. Lett.* **50**, 953–955.

Barge, T., Gas, P., d'Heurle, F. M. (1995), *J. Mater. Res.* **10**, 1134–1145.

Barrett, C. R., Nix, W. D., Tetelman, A. S. (1973), *The Principles of Engineering Materials*, Englewood Cliffs, N. J: Prentice-Hall, Chaps. 6 and 9.

Bednorz, J. G., Muller, K. A. (1986), *Phys. Rev. B*, **64**, 189–193.

Berti, A. C., Bolkhovsky, V. (1992), in: *Proceedings of the Tenth International VLSI Multilevel Interconnection Conference* (*VMIC*), Santa Clara, CA: Tenth International VLSI Multilevel Interconnection Conference.

Beyers, R. (1984), *J. Appl. Phys.* **56**, 147–152.

Beyers, R., Sinclair, R., Thomas, M. E. (1984), *J. Vac. Sci. Tech.*, **B2**, 781–784.

Bisi, O., Calandra, C. (1981), *J. Phys. C: Solid State Phys.* **14**, 5479–5494.

Bobet, J. L., Naslain, R., Bernard, C. (1995), *J. Chem. Vap. Dep.* **3**, 223–253.

Bocelli, S., Guizzetti, G., Marabelli, F., Thungstrom, G., Petersson, C. S. (1995), *Appl. Surf. Sci.* **91**, 30–33.

Bose, S. (1992), *Mater. Sci. Eng.* **A155**, 217–225.

Bost, M. C., Mahan, J. E. (1985), *J. Appl. Phys.* **58**, 2696–2703.

Bost, M. C., Mahan, J. E. (1988), *J. Appl. Phys.* **64**, 2034–2037.

Boutarek, N., Madar, R. (1993), *Appl. Surf. Sci.* **73**, 209–213.

Briant, C. L., Taub, A. I. (1988), *Acta Metall.* **36**, 2761–2770.

Broadbent, E. K., Morgan, A. E., Coulman, B., Huang, I.-W., Kuiper, A. E. T. (1987), *Thin Solid Films* **151**, 51–63.

Buschow, K. H. J. (1991), *Rep. Prog. Phys.* **54**, 1123–1213.

Carlsson, J.-O. (1990), *CRC Crit. Rev. Solid State Mater. Sci.* **16**, 161–212.

Carlsson, A. E., Meschter, P. J. (1991), *J. Mater. Res.* **6**, 1512–1517.

Carmondy, M. W., Johnson, A. S., Kvan, E. P. (1996), in: R. T. Tung, K. Maex, P. W. Pellegrini, L. H. Allen (Eds.), *Silicide Thin Films—Fabrication, Properties, and Applications*, MRS Symposium Proceedings Vol. 402, Pittsburgh: Materials Research Society, pp. 399–404.

Castro, R. G., Smith, R. W., Rollett, A. D., Stanek, P. W. (1992) *Mater. Sci. Eng.* **A155**, 101–107.

Chang, S.-T., Chiu, K. Y. (1989), *IEEE Trans. Elec. Dev.* **36**, 145–147.

Chen, L. J., Tu, K. N. (1991), *Mater. Sci. Rep.* **6**, 53–140.

Chittipeddi, S., Dziuba, C. M., Kannan, V. C., Kelly, M. J., Cochran, W. T., Rambabu, B. (1993), *J. Elec. Mat.* **22**, 785–791.

Chou, T. C., Joshi, A., Wadsworth, J. (1991), *J. Mat. Res.* **6**, 796–809.

Chou, T. C., Nieh, T. G. (1993), *J. Mat. Res.* **8**, 214–226.

Chow, T. P., Saxena, A. N., Ephrath, L. M., Bennett, R. S. (1984), in: R. A. Powell (Ed.), *Dry Etching for Microelectronics*, New York: Elsevier, Chap. 2.

Christensen, N. E. (1990), *Phys. Rev. B* **42**, 7148–7153.

Chu, C. L., Saraswat, K. C., Wong, S. S. (1992), *IEEE Trans. Elec. Dev.* **39**, 2333–2340.

Clevenger, L. A., Cabral, C., Roy, R. A., Lavoie, C., Viswanathan, R., Saenger, K. L., Jordan-Sweet, J., Morales, G., Ludwig, K. L., Stephenson, G. B. (1996), in: R. T. Tung, K. Maex, P. W. Pellegrini, L. H. Allen (Eds.), *Silicide Thin Films—Fabrication, Properties, and Applications*, MRS Symposium Proceedings Vol. 402, Pittsburgh: Materials Research Society, pp. 257–268.

Colgan, E. G., Chapple-Sokol, J. D. (1992), *J. Vac. Sci. Tech.* **B10**, 1156–1166.

Colgan, E. G., Gambino, J. P., Hong, Q. Z. (1996a), *Mater. Sci. Eng. Rep.* **R16**, 43–96.

Colgan, E. G., Gambino, J. P., Cunningham, B. (1996b), *Mater. Chem. Phys.* **46**, 209–214.

Cowley, A. M., Sze, S. M. (1965), *J. Appl. Phys.* **36**, 3212–3220.

Derrien, J. (1995), in: K. Maex, M. van Rossum (Eds.), *Properties of Metal Silicides*, London: INSPEC, p. 155.

d'Heurle, F. M., (1988), *J. Mater. Res.* **3**, 167–195.

d'Heurle, F. M., Gas, P. (1986), *J. Mater. Res.*, **1**, 205–221.

d'Heurle, F. M., Petersson, C. S., Baglin, J. E., LaPlaca, S. J., Wong, C. Y. (1984), *J. Appl. Phys.* **55**, 4208–4218.

d'Heurle, F. M., Cros, A., Frampton, R. D., Irene, E. A. (1987), *Philos. Mag.*, **B55**, 291–308.

Domenicucci, A., Dehm, C., Loh, S., Clevenger, L., Dziobkowski, C., Cabral, C., Lavoie, C., Jorden-Sweet, J. (1997), in: S. C. Moss, D. Ila, R. C. Cammarata, E. H. Chason, T. L. Einstein, E. D. Williams (Eds.), *Thin Films—Structure and Morphology*, MRS Symposium Proceedings Vol. 441, Pittsburgh: Materials Research Society, pp. 3–8.

Dutron, A.-M., Blanquet, E., Bourhila, N., Madar, R., Bernard, C. (1995), *Thin Solid. Films*, **259**, 25–31.

Fenske, F., Lange, H., Oertel, G., Reinsperger, G.-U., Schumann, J., Selle, B. (1996), *Mater. Chem. Phys.* **43**, 238–242.

Flamm, D. L., Donnelly, V. M., Ibbotson, D. E. (1984), in: N. G. Einspruch, D. M. Brown (Eds.), *Plasma Processing for VLSI*, New York: Academic Press, p. 189.

Fleischer, R. L. (1991), in: O. Izumi (Ed.), *Proceedings of the International Symposium on Intermetallic Compounds*, Sendai: The Japan Institute of Metals, p. 157.

Fleischer, R. L. (1994), in: J. H. Westbrook, R. L. Fleischer (Eds.) *Intermetallic Compounds: Vol. 2, Practice*, New York: John Wiley, Chap. 11.

Flores, F., Ortega, J. (1992), *Appl. Surf. Sci.* **56–58**, 301–310.

Frampton, R. D., Irene, E. A., d'Heurle, F. M. (1987), *J. Appl. Phys.* **62**, 2972–2980.

Frommeyer, G., Rosenkranz, R., Ludecke, C. (1990), *Z. Metallk. (Germany)* **81**, 307–313.

Fujitani, H., Asano, S. (1994), *Phys. Rev. B* **50**, 8681–8698.

Gambino, J. P., Colgan, E. G. (1998), *Mater. Chem. Phys.* **52**, 99–146.

Gambino, J. P., Monkowski, M. D., Tsang, P. J., Shepard, J. F., Ransom, C. M., Wong, C. Y. (1989), *J. Electrochem. Soc.* **136**, 2063–2067.

Gambino, J. P., Kastl, R. H., Cunningham, B., Turene, F. E. (1990), *Ext. Abstr. Electrochem. Soc.* **90–1**, Abs. No. 170.

Gambino, J. P., Cunningham, B., Buchanan, D. A. (1992), in: A. Katz, S. P. Murarka, Y. I. Nissim, J. M. E. Harper (Eds.), *Advanced Metallization and Processing for Semiconductor Devices and Circuits—II*, MRS Symposium Proceedings Vol. 260, Pittsburgh: Materials Research Society, pp. 181–186.

Gambino, J. P., Dobuzinsky, D., Armacost, M., Cunningham, B., Turene, F. E. (1995), in: *Proceedings of the Twelfth International VLSI Multilevel Interconnection Conference (VMIC)*, Santa Clara, CA: Twelfth International VLSI Multilevel Interconnection Conference.

Gambino, J. P., Colgan, E. G., Domenicucci, A. G., Cunningham, B. (1998), *J. Electrochem. Soc.* **145**, 1384–1389.

Gas, P., d'Heurle, F. M. (1993), *Appl. Surf. Sci.* **73**, 153–161.

Gas, P., d'Heurle, F. M. (1997), in: D. L. Beke (Ed.), *Diffusion in Semiconductors and Non-Metallic Solids*, New York: Springer Verlag.

Gas, P., Thomas, O., d'Heurle, F. M. (1995), in: K. Maex, M. van Rossum (Eds.), *Properties of Metal Silicides*, London: INSPEC, p. 298.

Ghandi, S. K. (1983), in *VLSI Fabrication Principles*, New York: John Wiley & Sons, p. 81.

Gibala, R., Ghosh, A. K., Van Aken, D. C., Srolovitz, D. J., Basu, A., Chang, H., Mason, D. P., Yang, W. (1992), *Mater. Sci. Eng.* **A155**, 147–158.

Goldschmidt, H. J. (1967), *Interstitial Alloys*, London: Butterworth, Chap. 7.

Goldsmith, C. C., Campbell, D. R., Mader, S. (1984), in J. E. E. Baglin, D. R. Campbell, W. K. Chu (Eds.), *Thin Films and Interfaces II*, New York: Elsevier, p. 569.

Gottlieb, U., Lasjaunias, J. C., Laborde, O., Thomas, O., Madar, R. (1993), *Appl. Surf. Sci.* **73**, 232–236.

Gottlieb, U., Nava, F., Affronte, F., Laborde, O., Madar, R. (1995), in: K. Maex, M. van Rossum (Eds.), *Properties of Metal Silicides*, London: INSPEC, p. 205.

Grabke, H. J. Meier, G. H. (1995), *Oxid. Met.* **44**, 147–176.

Grunling, H. W., Bauer, R. (1982), *Thin Solid. Films* **95**, 3–20.

Guizetti, G., Marabelli, F. (1995), in: K. Maex, M. van Rossum (Eds.), *Properties of Metal Silicides*, London: INSPEC, p. 227.

Hara, S., Ohdomari, I. (1988), *Phys. Rev.* **B38**, 7554–7557.

Hashimoto, K., Samuels, D. J., Farrell, T. R., Moy, D., Martino, R. M., Ferguson, R. A., Sato, T., Maurer, W. (1994), *Jpn. J. Appl. Phys.* **33**, 6823–6830.

Heine, V. (1965), *Phys. Rev.* **A138**, 1689–1696.

Henger, C. H., Brimhall, J. L., Hirth, J. P. (1992), *Mater. Sci. Eng.* **A155**, 109–114.

Hertzberger, R. W. (1976), *Deformation and Fracture Mechanics of Engineering Materials*, New York: John Wiley, Chap. 2.

Herman, H. (1988), *Mat. Res. Soc. Bull.* **13** (12), 60–67.

Hillenius, S. J., Liu, R., Georgiu, G. E., Field, R. L., Williams, D. S., Kornblit, A., Boulin, D. M., Johnston, R. L., Lynch, W. T. (1986), in: *IEEE International Electron Devices Meeting* (*IEDM*), *Technical Digest*, Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Hirade, S., Inoue, Y., Yamaha, T., Hangasaki, O., Hotta, T. (1995), in: *Proceedings of the Twelfth International VLSI Multilevel Interconnection Conference* (*VMIC*), Santa Clara, CA: Twelfth International VLSI Multilevel Interconnection Conference.

Hong, Q. Z., Hong, S. Q., d'Heurle, F. M., Harper, J. M. E. (1994), *Thin Solid. Films* **253**, 479–484.

Hong, Q. Z., Mayer, J. W. (1990), in: S. Katz, S. P. Murarka, A. Appelbaum (Eds.), *Advanced Metallization in Microelectronics*, MRS Symposium Proceedings Vol. 181, Pittsburgh: Materials Research Society, pp. 35–38.

Hong, T., Smith, J. R., Srolovitz, D. J. (1993), *Inter. Sci.* **1**, 223–235.

Hsia, S. L., Tan, T. Y., Smith, P., McGuire, G. E. (1991), *J. Appl. Phys.* **70**, 7579–7587.

Hsieh, J. J. (1993), *J. Vac. Sci. Tech.* **A11**, 3040–3046.

Hsieh, J. C., Fang, Y. K., Chen, C. W., Tsai, N. S., Lin, M. S., Tseng, F. C. (1993), *J. Appl. Phys.* **73**, 5038–5042.

Hui, J., Wong, S., Moll, J. (1985), *IEEE Trans. Elec. Dev. Lett.*, **EDL-6**, 479–481.

Irene, E. A. (1988), *CRC Crit. Rev. Solid. State Mater. Sci.* **14**, 175–223.

Ito, K., Inui, H., Hirano, T., Yamaguchi, M. (1992), *Mater. Sci. Eng.* **A152**, 153–159.

Ivanko, A. A. (1968), in: G. V. Samsonov (Ed.), *Handbook of Hardness Data*, Kiev: Academy of Science of the Ukrainian SSR [English translation (1971), Jerusalem: Israel Program of Scientific Translations].

Jahnel, F., Biersack, J., Crowder, B. L., d'Heurle, F. M., Fink, D., Isaac, R., Lucchese, C. J., Petersson, C. S. (1982), *J. Appl. Phys.* **53**, 7372–7378.

Jimenez, J. R. (1996), in: R. T. Tung, K. Maex, P. W. Pellegrini, L. H. Allen (Eds.), *Silicide Thin Films—Fabrication, Properties, and Applications*, MRS Symposium Proceedings Vol. 402, Pittsburgh: Materials Research Society, pp. 419–430.

Jonckheere, R., Ronse, K., Popa, O., Van den Hove, L. (1994), *J. Vac. Sci. Technol.* **B12**, 3765–3772.

Kaplan, W., Zhang, S.-L., Norstrom, N., Ostling, M., Lindberg, A. (1992), *Microelectron. Eng.* **19**, 661–664.

Kimata, M., Tubouchi, N. (1995), in: A. Rogalski (Ed.), *Infrared Photon Detectors*, Bellingham, WA: SPIE Optical Engineering Press, Chap. 9.

Kimura, K., Nakamura, M., Hirano, T. (1990), *J. Mater. Sci.* **25**, 2487–2492.

Kingery, W. D., Bowen, H. K., Uhlmann, D. R. (1976), *Introduction to Ceramics*, New York: John Wiley, Chap. 10.

Kircher, T. A., Courtright, E. L. (1992), *Mater. Sci. Eng.* **A155**, 67–74.

Kittel, C. (1976), *Introduction to Solid State Physics*, New York: John Wiley, Chaps. 6, 14 and 15.

Kobayashi, N., Goto, H., Suzuki, M. (1991), *J. Appl. Phys.* **69**, 1013–1019.

Koburger, C., Ishaq, M., Geipel,, H. J. (1982), *J. Electrochem. Soc.* **129**, 1307–1312.

Kumar, K. S. (1994), in: J. H. Westbrook, R. L. Fleisher (Eds.), *Intermetallic Compounds: Vol. 2, Practice*, New York: John Wiley, Chap. 10.

Kurtin, S., McGill, T. C., Mead, C. A. (1969), *Phys. Rev. Lett.* **22**, 1433–1436.

Kuwano, H., Phillips, J. R., Mayer, J. W. (1990), *Appl. Phys. Lett.* **56**, 440–442.

Lange, H. (1996), in: R. T. Tung, K. Maex, P. W. Pellegrini, L. H. Allen (Eds.), *Silicide Thin Films—Fabrication, Properties, and Applications*, MRS Symposium Proceedings Vol. 402, Pittsburgh: Materials Research Society, pp. 307–318.

Lee, H. G. (1992), *Thin Solid Films* **216**, 230–234.

Lee, S.-W., Ihn, T.-H., Joo, S.-K. (1996), *IEEE Elec. Dev. Lett.* **17**, 407–409.

Li, F., Lustig, N., Klosowki, P., Lannin, J. S. (1990), *Phys. Rev.* **B41**, 10210–10213.

Li, G. P., Ning, T. K., Chuang, C. T., Ketchen, M. B., Tang, D. D., Mauer, J. (1987), *IEEE Trans. Elec. Dev.* **ED-34**, 2246–2253.

Lien, C. D., Nicolet, M.-A., Lau, S. S. (1986), *Thin Solid Films* **143**, 63–72.

Lien, C. D., Nicolet, M.-A. (1984), *J. Vac. Sci. Technol.*, **B2**, 738–747.

Liu, Q. Z., Yu, L. S., Lau, S. S., Redwing, J. M., Perkins, N. R., Kuech, T. F. (1997), *Appl. Phys. Lett.* **70**, 1275–1277.

Liu, R., Lu, C.-Y., Sung, J. J., Pai, P.-S., Tsai, N.-S. (1995), *Solid State Electron.*, **38**, 1473–1477.

Lograsso, T. A. (1992), *Mater. Sci. Eng.* **A155**, 115–119.

Louie, S. G., Chelikowsky, J. R., Cohen, M. L. (1977), *Phys. Rev. B* **15**, 2154–2162.

Luo, J.-S., Lin, W.-T., Chang, C. Y., Tsai, W. C., Wang, S. J. (1997), *Mater. Chem. Phys.* **48**, 140–144.

Ma, Z., Allen, L. H. (1994), *Phys. Rev. B* **49**, 13501–13511.

Madar, R. (1995), in: K. Maex, M. van Rossum (Eds.), *Properties of Metal Silicides*, London: INSPEC, The Institute of Electrical Engineers, p. 109.

Maex, K. (1993), *Mat. Sci. Eng. Rev.* **R11**, 53–153.

Maex, K., Ghosh, G., Delaey, L., Probst, V., Lippens, P. Van den hove, L., DeKeersmaecker, R. F. (1989), *J. Mater. Res.* **4**, 1209–1217.

Maex, K., Van den hove, L., DeKeersmaecker, R. F. (1986), *Thin Solid Films* **140**, 149–161.

Malegori, G., Miglio, L. (1993), *Phys. Rev. B* **48**, 9223–9230.

Maloney, M. J., Hecht, R. J. (1992), *Mater. Sci. Eng.* **A155**, 19–31.

Maloy, S., Heuer, A. H., Lewandowski, J., Petrovic, J. (1991), *J. Am. Ceram. Soc.* **74**, 2704–2706.

Mann, R. W., Clevenger, L. A. (1995), in: K. Maex, M. van Rossum (Eds.), *Properties of Metal Silicides*, London: INSPEC, The Institute of Electrical Engineers, p. 55.

Mann, R. W., Clevenger, L. A., Agnello, P. D., White, F. R. (1995), *IBM J. Res. Dev.* **39**, 403–417.

Mantl, S. (1992), *Mater. Sci. Rep.* **8**, 1–95.

Mantl, S. (1994), *Nucl. Instrum. Meth. B*, **84**, 127–134.

Mashayekhi, A., Parfitt, L., Kalnas, C., Jones, J. W., Was, G. S., Hoffman, D. W. (1992), *Mater. Res. Soc.*, **273**, 275–280.

Masaki, A. (1992), *IEEE Circuits Dev.*, Nov., p. 18.

Massalski, T. B. (1990), *Binary Alloy Phase Diagrams*, Materials Park, OH: ASM International.

Matar, S. F., Chevalier, B., Etourneau, J. (1994), *J. Magn. Magn. Mater.* **137**, 293–304.

Matsuoka, F., Ishimaru, K., Gojohbori, H., Koike, H., Unno, Y., Sai, M., Kondo, T. Ichikawa, R., Kakumu, M. (1994), *IEICE Trans. Elec.* **E77-C**, 1385.

Mattheiss, L. F. (1992), *Phys. Rev. B* **45**, 3252–3259.

Maury, D., Rostoll, M. L., Gayet, P., Regolini, J. L. (1997), *J. Vac. Sci. Technol.* **B15**, 133–137.

Mayer, J. W., Lau, S. S. (1989), *Electronic Materials Science: For Integrated Circuits in Si and GaAs*, New York: Macmillan, p. 307.

Mayer, J. W., Tsaur, B. Y., Lau, S. S., Hung, L.-S. (1981), *Nucl. Instrum. Meth.* **182/183**, 1–13.

Mendiratta, M. G., Lewandowski, J. L., Dimiduk, D. M. (1991), *Metall. Trans.* **22A**, 1573–1583.

Miglio, L., Meregalli, V. (1996), in: R. T. Tung, K. Maex, P. W. Pellegrini, L. H. Allen (Eds.), *Silicide Thin Films-Fabrication, Properties and Applications*, MRS Symposium Proceedings Vol. 402, Pittsburgh: Materials Research Society, 367–372.

Miracle, D. B., Mendiratta, M. G. (1994), in: J. H. Westbrook, R. L. Fleisher (Eds.), *Intermetallic Compounds: Vol. 2, Practice*, New York: John Wiley, Chap. 13.

Mogami, T., Wakabayashi, H., Saito, Y., Matsuki, T., Tatsumi, T., Kunio, T. (1994), *IEEE Int. Elec. Dev. Meeting Proc.*, p. 687.

Morand, Y., Blanquet, E., Bourhila, N., Thomas, N., Bernard, C., Madar, R. (1994) in: R. W. Fathauer, S. Mantl, L. J. Schowalter, K. N. Tu (Eds.), *Silicides, Germanides, and Their Interfaces*, MRS Symposium Proceedings Vol. 320, Pittsburgh: Materials Research Society, pp. 91–96.

Moy, D., Basavaiah, S., Chuang, C. T., Li, G. P., Hackbarth, E., Brodsky, S. B., Polcari, M. R. (1988), *Solid State Electron*, **31**, 843–849.

Mueller, A., Wang, G., Rapp, R. A., Courtright, E. L., Kircher, T. A. (1992), *Mater. Sci. Eng.* **A155**, 199–207.

Mullins, W. W. (1957), *J. Appl. Phys.* **28**, 333–339.

Murarka, S. P. (1983), *Silicides for VLSI Applications*, New York: Academic Press, p. 50.

Muraraka, S. P., Williams, D. S. (1987), *J. Vac. Sci. Technol.* **B5**, 1674–1688.

Nauka, K., Amano, J., Scott, M. P., Weber, E. R., Turner, J. E., Tsai, R. (1986), in: M. Wittmer, J. Stimmell, M. Strathman (Eds.), *Materials Issues in Silicon Integrated Circuit Processing*, MRS Symposium Proceedings Vol. 71, Pittsburgh: Materials Research Society, pp. 319–324.

Nava, F., Tu, K. N., Thomas, O., Senateur, J. P., Madar, R., Borghesi, A., Guizzetti, G., Gottlieb, U., Laborde, O., Bisi, O. (1993), *Mater. Sci. Rep.* **9**, 141–200.

Ng, K. K., Lynch, W. T. (1996), *IEEE Trans. Elec. Dev.* **ED-33**, 965–972.

Nguyen, T., Ho, H. L., Kotecki, D. E., Nguyen, T. D. (1996), *J. Appl. Phys.* **79**, 1123–1128.

Nicholls, J. R., Stephenson, D. J. (1994), in: J. H. Westbrook, R. L. Fleisher (Eds.), *Intermetallic Compounds: Vol. 2, Practice*, New York: John Wiley, Chap. 22.

Nicolet, M.-A., Lau, S. S. (1983), in: N. G. Einspruch, G. B. Larrabee (Eds.), *VLSI Microelectronics Science*, Vol. 6, New York: Academic Press, Chap. 6.

Nicoll, A. R., Hildebrandt, U. W., Wahl, G. (1979), *Thin Solid Films* **64**, 321–326.

Ning, T. H. (1992), in: *Proceedings of the Thirteenth IEEE/CHMT International Electronics Manufacturing Technology Symposium*, Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Nobili, C. (1990), in: H. R. Huff, K. G. Barraclough, J. Chikawa (Eds.), *Semiconductor Silicon 1990*, Pennington, N. J.: Electrochemical Society, p. 550.

Nolan, T. P., Sinclair, R., Beyers, R. (1992), *J. Appl. Phys.* **71**, 720–724.

Nulman, J. (1990), in: Katz, S. P. Murarka, A. Appelbaum (Eds.), *Advanced Metallization in Microelectronics*, Pittsburgh: Materials Research Society Proc. **181**, pp. 123–132.

Ogawa, S., Fair, J. A., Kouzaki, T., Sinclair, R., Jones, E. C., Cheung, N. W., Fraser, D. B. (1994), in: R. W. Fathauer, S. Mantl, L. J. Schowalter, K. N. Tu (Eds.), *Silicides, Germanides, and Their Interfaces*, MRS Symposium Proceedings Vol. 320, Pittsburgh: Materials Research Society, pp. 355–360.

Omura. T., Inokawa, H., Izumi, K. (1991), *J. Mater. Res.* **6**, 1238–1247.

Osburn, C. M. (1990), *J. Electron. Mater.* **19**, 67–88.

Osburn, C. M., Wang, Q. F., Kellam, M., Canovai, C., Smith, P. L., McGuire, G. E., Xiao, Z. G., Rozgonyi, G. F. (1991), *Appl. Surf. Sci.* **53**, 291–312.

Ostling, M., Zaring, C. (1995), in: K. Maex, M. van Rossum (Eds.), *Properties of Metal Silicides*, London: INSPEC, The Institute of Electrical Engineers, p. 15.

Paufler, P. (1976), in: *Intermetallische Phasen*, Leipzig: VEB Deutscher Verlag für Grundstoffindustrie, pp. 165–187.

Pearson, W. B. (1972), *The Crystal Chemistry and Physics of Metals and Alloys*, New York: Wiley-Interscience.

Petrovic, J. J. (1993), *Mater. Res. Soc. Bull.*, **18** (7), 35–40.

Petrovic, J. J., Bhattacharya, A. K., Honnell, R. E., Mitchell, T. E., Wade, R. K., McClellan, K. J. (1992), *Mater. Sci. Eng.* **A155**, 259–266.

Porter, L. M., Davis, R. F., Bow, J. S., Kim, M. J., Carpenter, R. W. (1995a), *J. Mater. Res.* **10**, 26–33.

Porter, L. M., Davis, R. F., Bow, J. S., Kim, M. J., Carpenter, R. W., Glass, R. C. (1995b), *J. Mater. Res.* **10**, 668–679.

Porter, L. M., Davis, R. F., Bow, J. S., Kim, M. J., Carpenter, R. W. (1995c), *J. Mater. Res.* **10**, 2336–2342.

Reader, A. H., van Ommen, A. H., Weijs, P. J. W., Wolters, R. A. M., Oostra, D. J. (1992), *Rep. Prog. Phys.* **56**, 1397–1467.

Rossnagel, S. (1995), in: D. C. Jacobson, D. E. Luzzi, T. F. Heinz, M. Iwaki (Eds.) *Beam-Solid Interactions for Materials Synthesis and Characterization*, MRS Symposium Proceedings Vol. 354, Pittsburgh: Materials Research Society, pp. 503–510.

Sadananda, K., Feng, C. R., Jones, H., Petrovic, J. (1992), *Mater. Sci. Eng.* **A155**, 227–239.

Sambasivan, S., Petuskey, W. T. (1994). *J. Mater. Res.* **9**, 2362–2369.

Samsanov, G. V., Vinitskii, I. M. (1980), *Handbook of Refractory Compounds*, New York: IFI/Plenum Data Company.

Saraswat, K. C., Brors, D. L., Fair, J. A., Monnig, K. A., Beyers, R. (1983), *IEEE Trans. Elec. Dev.* **ED-30**, 1497–1505.

Schluter, M. (1977), *Phys, Rev. B.* **17**, 5044–5047.

Schwarz, R. B., Srinivasan, S. R., Petrovic, J. J., Maggiore, C. J. (1992), *Mater. Sci. Eng.* **A155**, 75–83.

Scott, D. B., Chapman, R. A., Wei, C.-C., Mahant-Shetti, S. S., Haken, R. A., Holloway, T. C. (1987), *IEEE Trans. Electron. Dev.* **ED-34**, 562–573.

Seidel, T. E. (1983), in: S. M. Sze (Ed.), *VLSI Technology*, New York: McGraw-Hill, Chap. 6.

Shah, D. M., Berczik, D., Anton, D. L., Hecht, R. (1992), *Mater. Sci. Eng.* **A155**, 45–57.

Shaw, L., Miracle, D., Abbaschian, R. (1995), *Acta Metal.* **43**, 4267–4279.

Shiba, T., Tamaki, Y., Onai, T., Kiyota, Y., Kure, T., Nakamura. T. (1996), *IEEE Trans. Electron. Dev.* **43**, 1357–1363.

Shih, Y.-C., Callegari, A., Murakami, M., Wilkie, E. L., Hovel, H. J., Parks. C. C., Childs, K. D. (1988), *J. Appl. Phys.* **64**, 2113–2121.

Shioya, Y., Ikegami, K., Kobayashi, I., Maeda, M. (1987), *J. Electrochem. Soc.* **134**, 1220–1224.

Shone, F. C., Hansen, S. E., Kao, D. B., Saraswat, K. C., Plummer, J. D. (1986), *Proc. IEEE Inter. Elec. Dev. Mett.* p. 534.

Singheiser, L., Wahl, G., Thiele, W. (1982), *Thin Solid Films* **95**, 35–45.

Southwell, G., MacAlpine, S., Young, D. J. (1987), *Mater. Sci. Eng.* **88**, 81–87.

Southwell, R. P., Seebauer, E. G. (1997), *J. Electrochem. Soc.* **144**, 2122–2137.

Speier, W., Kumar, L., Sarma, D. D., de Groot, R. A., Fuggle, J. C. (1989), *J. Phys.: Condens. Matter* **1**, 9117–9129.

Srinivas, D., Raupp, G. B., Hillman, J. (1990), in: S. S. Wong, S. Furukawa (Eds.), *Tungsten and Other Advanced Metals for VLSI/ULSI Applications*, Pittsburgh: Materials Research Society Proceedings Vol. V-6, pp. 407–413.

Stekly, Z. J. J., Gregory, E., (1994), in: J. H. Westbrook, R. L. Fleisher (Eds.), *Intermetallic Compounds: Vol. 2, Practice*, New York: John Wiley, Chap. 16.

Sullivan, J. P., Tung, R., Eaglesham, D. J., Schrey, F., Graham, W. R. (1993), *J. Vac. Sci. Technol.* **B11**, 1564–1570.

Sundahl, M. J., Irwin, R. B., Nanda, A., Wilkins, C. W., Rambabu, B. (1995), *Phys. Status Solidi* **A147**, 477–490.

Sze, S. M. (1981), *Physics of Semiconductor Devices*, New York: John Wiley & Sons, Chap. 5.

Tada, K., Laraya, J. L. R. (1967), *Proc. IEEE* **55**, 2064–2065.

Takasugi, T., Yoshida, M. (1991), *J. Mater. Sci.* **26**, 3032–3040.

Takeuchi, S., Hashimoto, T., Shibuya, T. (1992), *J. Mater. Sci.* **27**, 1380–1384.

Tanelian, M., Lajos, R., Blackstone, S., Pramanick, D. (1985), *J. Electrochem. Soc.* **132**, 1456–1460.

Taubenblatt, M. A., Helms, C. R. (1986), *J. Appl. Phys.* **59**, 1992–1997.

Telford, S. G., Eizenberg, M., Chang, M., Sinha, A. K., Gow, T. R. (1993), *J. Electrochem. Soc.* **140**, 3689–3701.

Thomas, O., Gas, P., Charai, A., LeGoues, F. K., Michel, A., Scilla, G., d'Heurle, F. M. (1988a), *J. Appl. Phys.* **64**, 2973–2980.

Thomas, O., Gas, P., d'Heurle, F. M., LeGoues, F. K., Michel, A., Scilla, G., (1988b), *J. Vac. Sci. Technol.* **A6**, 1736–1739.

Tilly, D. J., Lofvander, J. P. A., Levi, C. G. (1992), in: D. B. Miracle, D. L. Anton, J. A. Graves (Eds.), *Intermetallic Matrix Composites II*, MRS Symposium Proceedings Vol. 273, Pittsburgh: Materials Research Society, pp. 295–300.

Tiwari, R., Herman, H., Sampath, S. (1992), *Mater. Sci. Eng.* **A155**, 95–100.

Tung, R. (1992), in: D. Wolf, S. Yip (Eds.), *Materials Interfaces: Atomic Level Structure and Properties*, New York: Chapman & Hall, p. 550.

Tung, R. (1993), *J. Vac. Sci. Technol.* **B11**, 1546–1552.

Tung, R. (1997), *Jpn. J. Appl. Phys.* **36**, 1650–1654.

Umakoshi, Y., Sakagami, T., Hirano, T., Yamane, T. (1990), *Acta Metallogr.*, **38**, 909–915.

van den Hoek, P. J., Ravenek, W., Baerends, E. J. (1988), *Phys. Rev. Lett.* **60**, 1743–1746.

van Loo, F. J. J., Smet, F. M., Rieck, G. D., Verspui, G. (1982), *High Temp. High Press.* **14**, 25–31.

Vasudevan, A. K., Petrovic, J. J. (1992), *Mater. Sci. Eng.* **A155**, 1–17.

Vedernikov, M. V. (1994), in: J. H. Westbrook, R. L. Fleisher (Eds.), *Intermetallic Compounds*, Vol. 2, *Practice*, New York; John Wiley, Chap. 20.

Vijh, A. K. (1994), in: J. H. Westbrook, R. L. Fleisher (Eds.), *Intermetallic Compounds*, Vol. 2, *Practice*, New York: John Wiley, Chap. 23.

Vining, C. B. (1992), in: C. B. Vining (Ed.), *Proceedings 9th International Conference on Thermoelectronics*, Pasadena: California Institute of Technology, p. 249.

von Känel, H. (1992), *Mater. Sci. Rep.* **8**, 193–269.

von Känel, H. Kafader, U., Sutter, P., Onda, N., Sirringhaus, H., Muller, E., Kroll, U., Schwarz, C., Goncalces Conto, S. (1994), in: R. W. Fathauer, S. Mantl, L. J. Schowalter, K. N. Tu (Eds.), *Silicides, Germanides, and Their Interfaces*, MRS Symposium Proceedings Vol. 320, Pittsburgh: Materials Research Society, pp. 73–84.

Wang, S. Q., Mayer, J. W. (1990), *J. Appl. Phys.* **67**, 2932–2938.

Weijs, P. J. W., Czyzyk, M. T., Fuggle, J. C., Speier, W., Sarma, D. D., Buschow, K. H. J. (1990), *Z. Phys.* **B78**, 423–430.

West, G. A., Beeson, K. W. (1987), in: G. W. Cullen (Ed.), *Proceedings 10th International Conference on CVD*, Pennington NJ: Electrochemical Society, p. 720.

Winnerl, J. N. (1994), in: R. W. Fathauer, S. Mantl, L. J. Schowalter, K. N. Tu (Eds.), *Silicides, Germanides, and Their Interfaces*, MRS Symposium Proceedings Vol. 320, Pittsburgh: Materials Research Society, pp. 37–46.

Yallup, K., Wilson, R., Quinn, C., McDonnell, B., Blackstone, S. (1995), *Proc. IEEE Int. SOI Conf.*, p. 137.

Yokoyama, N., Ohnishi, T., Nakamura, T., Nishi, H. (1986), in: R. J. Nemanich, P. S. Ho, S. S. Lau (Eds.), *Thin Films—Interfaces and Phenomena*, MRS Symposium Proceedings Vol. 54, Pittsburgh: Materials Research Society, pp. 401–408.

Yoo, C.-S., Lin, T.-H., Tsai, N.-S., Ijsendoom, L. J. (1990), *Jpn. J. Appl. Phys.* **29**, 2535–2540.

Zaring, C., Gas, P., Svensson, B. G., Ostling, M., Whitlow, H. F. (1990), *Thin Solid Films*, **193/194**, 244–247.

Zhang, S.-L., d'Heurle, F. M. (1992), *Philos. Mag.* **66**, 415–424.

Zhang, S.-L., d'Heurle, F. M. (1994), *Mater. Sci. Forum* **155–156**, 59–70.

Zheng, L. R., Hung, L. S., Mayer, J. W. (1985), *J. Appl. Phys.* **58**, 1505–1514.

Zur, A., McGill, T. C., Nicolet, M. A. (1985), *J. Appl. Phys.* **57**, 600–603.

Further Reading

Fathauer, R. W., Mantl, S., Schowalter, L. J., Tu, K. N. (Eds.) (1994), *Silicides, Germanides, and Their Interfaces*, MRS Symposium Proceedings Vol. 320, Pittsburgh: Materials Research Society.

Gambino, J. P. Colgan, E. G. (1998), *Mater. Chem. Phys.* **52**, 99–146.

Maex, K., van Rossum, M. (Eds.) (1995), *Properties of Metal Silicides*, London: INSPEC, The Institute of Electrical Engineers.

Murarka, S. P. (1983), in: *Silicides for VLSI Applications*, New York: Academic Press.

Nicolet, M.-A., Lau, S. S. (1983), in: N. G. Einspruch, G. B. Larrabee (Eds.), *VLSI Microelectronics Science*, Vol. 6, New York: Academic Press, Chap. 6.

Samsanov, G. V., Vinitskii, I. M. (1980), in: *Handbook of Refractory Compounds*, IFI/Plenum Data Company.

Tung, R. T., Maex, K., Pellegrini, P. W., Allen, L. H. (Eds.) (1996), *Silicide Thin Films—Fabrication, Properties, and Applications*, MRS Symposium Proceedings Vol. 402, Pittsburgh: Materials Research Society.

Westbrook, J. H., Fleisher, R. L. (Eds.) (1994), *Intermetallic Compounds*, Vol. 2, *Practice*, New York: John Wiley.

SILICON: POROUS

PHILIPPE M. FAUCHET, *Department of Electrical and Computer Engineering, The Institute of Optics, and Laboratory for Laser Energetics, University of Rochester, Rochester, New York, U.S.A.*

Introduction 249

1. **Preparation/Materials Science** 250

1.1 Electrochemistry: Electropolishing and Porous-Silicon Formation 250

1.2 Other Techniques to Produce Nanocrystalline Silicon 252

1.3 Post-Preparation Treatments 253

1.3.1 Supercritical Drying 253

1.3.2 Chemical Quenching 254

1.3.3 Aging and Oxidation 254

2. **Properties** 255

2.1 Structural 255

2.1.1 X-Ray Scattering 255

2.1.2 Electron Microscopy 255

2.2 Mechanical 256

2.3 Thermal 257

2.4 Vibrational 257

2.5 Electrical 258

2.6 Optical: Experiments 259

2.6.1 Absorption/Transmission 259

2.6.2 Dielectric Function and Index of Refraction 260

2.6.3 Luminescence 260

2.7 Optical: Theory 263

2.7.1 Band Gap 263

2.7.2 Luminescence 264

3. **Applications** 264

3.1 Light-Emitting Devices 264

3.2 Solar Cells 266

3.3 Mirrors and Photonic Band Gap 268

3.4 Waveguides 268

3.5 Sensors 268

3.6 SOI Technology 269

3.7 Micromachining 269

3.8 Biomedical Applications 269

3.9 Capacitors 269

Glossary 269

Works Cited 270

Further Reading 272

INTRODUCTION

Silicon is the second most abundant element in the Earth's crust. One can argue that in fact it is the second most important element in today's technological world, just after carbon, which is the essential ingredient of life. This is because microelectronics relies almost exclusively on silicon to make the chips that are the soul of every microprocessor and intelligent system. Silicon has taken such a leadership position in high technology and consumer electronics alike for two main reasons: it is cheap and abundant, and its oxide can be prepared with excellent properties.

The wafers that are used in microelectronics are made of bulk, crystalline silicon (c-Si). The properties of this form of silicon have been extensively investigated for more than 40 years and, as a result, are well known. But there are many other forms of silicon. Polycrystalline and microcrystalline silicon films are made of crystalline silicon grains with sizes ranging from ≈1 μm to a few nanometers. They are routinely prepared by a variety of techniques such as chemical vapor deposition or laser recrystallization of yet another form of silicon, called amorphous silicon. The boundaries between the grains, which act as a source of traps and scattering centers, control many of the electrical properties. Amorphous silicon is a form of silicon that lacks long-range order but maintains the local tetrahedral coordination. Amorphous silicon itself is useless as an electronic material because it contains a huge density of traps, called dangling bonds. These dangling bonds can be passivated by hydrogen, which removes them from the band-gap region. Thus alloying amorphous silicon with several atomic percent of hydrogen makes it useful for photovoltaics, for example. Silicon is also present in a large variety of polymeric and organic compounds.

Porous silicon is a form of silicon with

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

unique properties, distinct from those of crystalline, microcrystalline, or amorphous silicon. It was first prepared in 1956, but its true nature was not understood until more than one decade later. The most common preparation procedure consists of etching a silicon wafer in an electrochemical cell in the presence of HF. Under the appropriate conditions, a large density of pores is produced on the surface of the wafer, which is used as the anode. These pores, which can vary in size from ≈ 1 nm to >1 μm, can run straight perpendicular to the wafer surface or form a complex mesh of interconnected pores. They can penetrate tens of micrometers below the wafer surface. The silicon material left between the pores usually retains its crystalline nature and is made of structures whose sizes can vary from ≈ 1 nm to >1 μm.

It is clear that porous silicon can be prepared with a wide range of properties, which makes it a very versatile material. All forms of porous silicon share a very large internal surface area, which can be as large as an Olympic stadium per gram of material. Thus, most properties of porous silicon are affected not only by its crystalline core but also by its surface properties. In addition, the properties of the crystalline core may differ from those of bulk c-Si when the size of the silicon structures drops below 10 nm, a size regime in which quantum mechanical effects modify its electronic states.

Although porous silicon has been studied and used for more than 40 years, it is only starting in 1990 that it attracted a lot of attention. This interest was triggered by the discovery that appropriately prepared porous silicon can emit bright visible light that can be tuned from the near infrared throughout most of the visible spectrum. Since bulk crystalline Si has an indirect band gap near 1.1 eV (or 1.1 μm), both the brightness and the color range were unexpected. For the first time, silicon light emitters appeared possible. Since then, we have learned much about the properties of this unusual form of silicon. This article reviews our understanding of many of its properties. It starts with a discussion of the preparation conditions and the post-preparation treatments that stabilize its properties. The article then moves to a review of many of its properties. Its optical properties are highlighted, since they are responsible for most of the interest in porous silicon since 1990. Finally, several applications of porous silicon are discussed, in the areas of light-emitting devices, solar cells, integrated optics, sensors, micromachining, and biomedicine.

1. PREPARATION/MATERIALS SCIENCE

1.1 Electrochemistry: Electropolishing and Porous-Silicon Formation

Porous silicon (PSi) is a form of silicon that is produced chemically or electrochemically. The electrochemical route is more flexible, in the sense that it leads to a wider range of porosities, thicknesses, *etc.* PSi is prepared by anodic etching of silicon in a solution containing HF. It was first made in 1956 (Uhlir, 1956) and later recognized as etched silicon, as opposed to a deposit on the silicon wafer. In the electrochemical cell, the anode is the silicon wafer and the cathode is a platinum wire. When a constant current is passed between the electrodes, one can distinguish three regions in the curve of current (I) *versus* polarity (V) (Fig. 1). In reverse bias, I is zero and the dissolution rate of the silicon wafer is very slow. In forward bias, at low I, the I-V curve is reminiscent of a diode; this is the region where PSi is formed. As V is increased further, a region of instability is entered, beyond which the increase of I with V becomes minimal; this is the regime of electropolishing. In this regime, the Si is dissolved layer by layer, whereas in the PSi regime, pores of different sizes, depending on several parameters, are produced (Smith and Collins, 1992). There is a

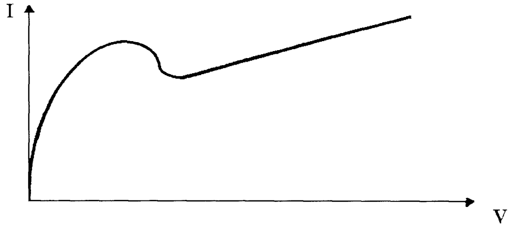

FIG. 1. Current–voltage relationship corresponding to the anodic oxidation of a p-type Si wafer in HF.

critical current density below which PSi is formed and above which electropolishing takes place. This demarcation exists because the formation of pores requires that the transport of holes from the bulk to the surface be the limiting step.

It is also possible to produce PSi starting from an *n*-type Si wafer. In this case, there are two interesting regimes in the *I*-*V* curve. In forward bias, *I* increases as in a diode and PSi can be formed. In reverse bias, no current flows except when the liquid/semiconductor interface is flooded by light. In this case, the *I*-*V* curve resembles that for *p*-type Si, except with an offset near 0 V. Note that porous silicon can also be prepared starting from polycrystalline, microcrystalline, or even heavily doped amorphous silicon. In addition, several other semiconductors have been made porous electrochemically.

The pore morphology and achievable porosities are strongly dependent on the current density, the HF concentration, the use of light during anodization, and the substrate type and doping density. Additionally, they are influenced by the *p*H of the solution, the use of alcohol in the solution, and other parameters such as the temperature. As the porosity increases,

1. the pores lose their directionality and the structure acquires a spongy character, and
2. the average size of the silicon skeleton decreases to a few nanometers.

The terminology used to described PSi has traditionally focused on the pore size (Table 1). This classification is convenient when the pores are of interest, but less useful for studies of the properties of electrons and holes in the silicon material left between the pores (Fauchet, 1997). Since Canham's work in 1990 (Canham, 1990), most of the interest in PSi has come from its strong, visible luminescence. Luminescence has been associated with the presence of silicon wires or nanocrystallites with typical dimensions in the nanometer size. Thus, a new term has been introduced, nanoporous silicon. Here, the prefix nano is used to describe the size of the Si objects. The porosity and the nanocrystallite size can be correlated for various types of samples. Because PSi is a very complex material, which can contain either Si wires or Si dots or even both types of structure simultaneously, the words nanocrystallites, dots, wires, and columns will be considered interchangeable. In a given section, one word will be preferred if this choice helps conveys the concept under examination.

Table 1. Traditional porous silicon classification.

Classification	Diameter of pores (nm)	Surface area ($m^2\,cm^{-3}$)
Macro	>50	10–100
Meso	2–50	100–300
Micro	≤ 2	300–800

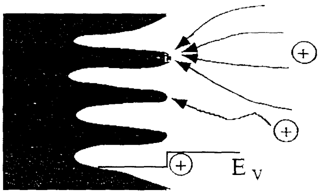

FIG. 2. Three mechanisms promote the formation of pores. From top to bottom, they are the concentration of the electric field lines at the pore's tip; random walk, which makes it easier for the holes to reach the pore's tip; and the potential barrier to hole transport from the bulk to the Si column's tip due to quantum confinement.

The pore initiation and growth mechanisms are qualitatively understood. Pore growth can be explained by several models, each one of which is more relevant for a specific regime of porosity and pore size (Smith and Collins, 1992). Assume first that shortly after anodic etching has started, the silicon/electrolyte interface has become irregular (Fig. 2). The key question is whether the fluctuations of the Si/electrolyte interface will grow (PSi formation) or disappear (electropolishing). Si etching requires the injection of holes from the substrate. A depletion region is formed in the thin PSi layer itself and also in a region of the Si wafer near the PSi/c-Si interface. In forward bias, the holes can still reach the Si/electrolyte interface by one of two mechanisms. First, the electric field lines are focused at the pore's tip. Thus, holes will preferentially reach the Si/electrolyte interface deep in the pores, where etching will proceed rapidly. In contrast, no holes will reach the end of the Si rods, effectively stopping the etching there. Second, if this electrostatic effect is not strong enough, random walk of the holes toward the Si/electrolyte interface makes it more likely that they reach it at or near the pore's tip, resulting in an effect similar to the electrostatic one. Another important mechanism becomes predominant if the Si rods are narrow enough (typically less than 10 nm). In this size regime, the electronic states start to differ from those of bulk silicon. When the motion of carriers is restricted in one or more dimensions, their minimum energy can no longer be zero. As a result of this effect, which is called quantum confinement, the holes in the valence band of the Si rods are pushed to lower energy, which produces a potential barrier to hole transport from the wafer to the Si rods. The holes can no longer drift or diffuse into the Si rods and further etching stops (Lehmann *et al.*, 1992).

When the top layer of the wafer is amorphized—for example, by ion implantation at high doses—hole transport is suppressed and porous-silicon formation is inhibited. In contrast, when the wafer surface is lightly damaged—for example, by a light ion milling—a higher density of larger nanopits or nano-irregularities is produced on the surface, the electric field lines are redistributed, and porous silicon is produced at a much increased rate (Duttagupta *et al.*, 1995). When light of increasing photon energy is used during anodization, the resulting peak luminescence energy also increases. This result can be explained as follows: In the dark, the etching of the Si rods stops when their valence-band edge drops by an amount that is sufficient to present an insurmountable barrier to hole transport from the Si wafer. With light, holes continue to be generated by absorption of the incoming photons, resulting in further etching and thinning of the Si rods. The valence-band edge continues to drop and the conduction-band edge continues to increase. The band-gap energy of the Si rods, which is given by the sum of the band-gap energy of bulk Si, the drop in the valence-band edge, and the increase in the conduction-band edge, increases further. This increase in the band-gap energy should be reflected by an increase of the luminescence energy. Similar experiments performed in the electrolyte in the open-circuit configuration just after the end of anodization yield similar result and thus confirm the interpretation (Koyama and Koshida, 1993).

1.2 Other Techniques to Produce Nanocrystalline Silicon

Since the discovery that PSi can emit light efficiently in the visible range, there has been a renewed interest in the manufacture of silicon quantum dots and wires. Spark erosion (Hummel *et al.*, 1993) consists of producing an electric discharge between a silicon wafer and an electrode, either in vacuum or in the presence of a gas. The resulting Si surface is covered by pits and hills of various sizes in the micrometer to nanometer scale. Spark-processed silicon is found to be highly luminescent in the visible range. Its color depends on the processing conditions, ranging from the red to the blue.

Lithographic techniques have also been used produce nanometer-size Si pillars or wires. Optical or electron-beam lithography is first used to produce submicron features. The Si stuctures are then thinned—for example, by an anisotropic etch in KOH and/or by oxidation—until the remaining Si nanostructure is on the order of 10 nm or less. Vertical Si pillars (Nassiopoulos *et al.*, 1995) and horizontal Si wires (Zaidi *et al.*, 1995) with dimensions below 10 nm have been fabricated. In many cases, luminescence has been reported with these structures, including at least one report of electroluminescence. However, the density of the Si nanostructures is low, which severely limits the brightness.

Another interesting method to produce Si nanocrystals that emit light is to start with a microcrystalline film and then to oxidize the film such that the diameter of the Si crystallites decreases to the appropriate size (less than 5 nm) and the surface of each nanocrystallite is passivated by silicon dioxide (Tamura *et al.*, 1994). This procedure has been shown to produce films with easily detectable visible luminescence. The advantages of this technique are that it is simple and compatible with Si microelectronic technology. The highest reported efficiency to date remains more than one order of magnitude below that of a typical PSi film. In addition, the crystallite size distribution is not well controlled.

Theory predicts that it should be possible to achieve visible luminescence even if the carriers are confined in only one direction, in Si slabs. These Si slabs are very similar to the familiar quantum wells grown by molecular-beam epitaxy (MBE) or chemical vapor deposition (CVD). Si/CaF_2 films grown by MBE exhibit visible luminescence when the Si layer thickness drops below 2 nm (Vervoot *et al.*, 1995). However, the Si films are microcrystalline because the large lattice mismatch and difference in thermal expansion coefficient between Si and CaF_2 make it impossible to grow epitaxial films free of dislocations. Thus, quantum confinement probably takes place in three directions, not just one, which triples the increase in the band-gap energy. It has also been shown that amorphous silicon layers sandwiched between silicon dioxide barriers luminesce in the visible range (Lu *et al.*, 1995). Once again, there is a critical thickness of the a-Si layers above which no luminescence is detected. The maximum luminescence efficiency of these structures remains at least one order of magnitude below that of a good-quality PSi film.

Another promising technique for the production of highly luminescent Si nanostructures involves the recrystallization of an amorphous Si/SiO_2 superlattice, made of many alternating layers of amorphous Si and SiO_2 (Tsybeskov *et al.*, 1998). These recrystallized Si layers are made of identical Si crystallites whose diameter is determined by the thickness of the original Si layer. These films exhibit an external photoluminescence quantum efficiency between 0.1 and 1% and a luminescence spectrum including all the well-known phonon replicas, which coincides with that of bulk c-Si for diameters in excess of 10 nm.

Si quantum dots can be prepared by several other techniques. One example consists of implanting SiO_2 with Si and then annealing the film to form Si clusters that then recrystallize. The luminescence of these structures can be tuned throughout the visible part of the spectrum by controlling the processing steps (Petrova-Koch *et al.*, 1996). However, the control on the size of the crystallites and the crystallite packing density is limited. Another example consists of preparing clusters or crystallites in the gas phase, which are then collected on a glass plate or in a matrix (Littau *et al.*, 1993). Gas-phase synthesis of Si quantum dots, followed by gas-phase oxidation and size selection, has produced the most luminescent material to date (quantum efficiency of 50% at low temperatures). Unfortunately, the production rate of these crystallites is too low to make them technologically useful.

1.3 Post-Preparation Treatments

1.3.1 Supercritical Drying One of the major problems in the manufacture of PSi involves the removal of the etching solution from the pores after anodization. The most straightforward technique consists of letting the solution evaporate in air. This procedure can lead to the collapse of the Si rods in highly porous samples. This collapse is a consequence of the large capillary-induced tensile stresses encountered during evaporation at the vapor–liquid interface in the pores. To avoid this problem, various methods have been proposed, including the use of low–surface-tension fluids such as pentane, freeze drying, and supercritical drying, a technique well known from silica aerogels. The use of supercritical fluids as drying agents has led to the best results (Canham *et al.*, 1994). Their application requires knowledge of the phase boundary and critical point of the confined fluid in the porous material, which are usually estimated from bulk fluid properties. The phase-coexistence boundary and critical point of a confined fluid can, however, be significantly different from its bulk values.

Figure 3 shows a typical trajectory followed during supercritical drying and contrasts it to a trajectory corresponding to normal drying in air. Trajectory 1–2–3 illustrates the supercritical drying of PSi with carbon dioxide (bulk phase $T_c = 31.1\,°C$, $P_c = 73.8\,bar$). The ethanol is first replaced by liquid CO_2 for 3 h at point 1. This is followed by the conversion of liquid CO_2 into a supercritical state by raising its temperature to 40 °C during the step 1–2. Finally, the CO_2 is vented at 40 °C for a period of 3 h during step 3. Similar results are obtained if another fluid is used (for example, C_2H_4). Whereas most samples prepared without supercritical drying often break for porosities greater than 80%, PSi films of porosities greater than 95% survive intact if they are prepared by supercritical drying. This is the case even for free-standing films lifted off the Si substrate (von Behren *et al.*, 1997).

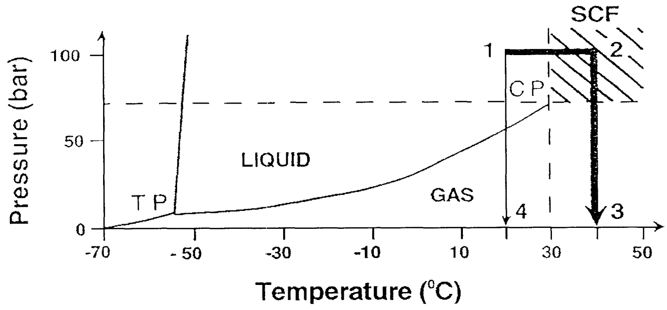

FIG. 3. Pressure–temperature diagram of bulk CO_2. Supercritical drying (trajectory 1–2–3) avoids the liquid–gas interface by converting liquid CO_2 into a supercritical fluid (SCF). Conventional drying is shown by trajectory 1–4.

1.3.2 Chemical Quenching Because of its very large internal surface, PSi is an ideal laboratory for the study of the chemistry of silicon surfaces. The surface of PSi, which is passivated by Si–H_n bonds, can be modified by chemical reactions (Sailor and Lee, 1997). The chemical compounds or molecules that become attached to the PSi surface can act as surface-state traps or sites for interfacial energy transfer and charge transfer, resulting in the quenching of the photoluminescence (PL). Reaction of PSi with iodine completely quenches the PL, on account of the production of Si–I surface traps that act as nonradiative recombination centers. This type of reaction is not fully reversible. In contrast, exposure to liquid or vapor methanol also quenches the PL, but this quenching is reversible because no net chemical reaction has taken place. Exposure to electron acceptor or donor components can also quench the PL reversibly (Rehm *et al.*, 1996), as in the case of a toluene solution of anthracene. However, if the electron acceptor (donor) state has a higher (lower) energy than the conduction- (valence-) band edge of PSi, electron (hole) transfer from PSi to the component is energetically forbidden and the PL is not quenched (*e.g.*, naphthalene). The quenching ability of a chemical species is also a function of the initial PSi surface termination. For example, reversible quenching is observed when Si–H–passivated PSi is exposed to liquid methanol. When SiO_2-passivated PSi is exposed to the same solution, no quenching is observed (Rehm *et al.*, 1995).

1.3.3 Aging and Oxidation After anodization, the internal surface of PSi is covered by silicon hydride bonds that remove any surface dangling-bond states from the band gap and thus provide excellent surface passivation. Excellent passivation is a requirement for achieving a high luminescence efficiency, since the radiative lifetime of PSi is in the microsecond to millisecond time domain because of the indirect gap of porous silicon, which is ample time for trapping in any nonradiative recombination center. Experimentally, it is observed that the PL efficiency is decreased by many factors (Collins *et al.*, 1992), including illumination with short-wavelength light, heating, exposure to various gases and liquids, and the presence of a high electric field. This PL degradation is intrinsic to the "fragile" Si–H_n bond and thus unavoidable. Hence, it is necessary to find a substitute for the Si–H bond, and the most obvious candidate is silicon dioxide. Silicon dioxide and its interface with crystalline silicon can be of very good quality, and in fact they are in large part responsible for the dominant role of silicon in microelectronics.

The stability of the luminescence efficiency of porous silicon passivated by silicon dioxide (SiO_2) has been shown to be excellent (Fauchet *et al.*, 1997). Oxide-passivated porous silicon can be prepared by different approaches: anodization at a low current density with light assistance, furnace oxidation, rapid thermal oxidation, or even aging the sample in air. A marked improvement in stability is observed when the Si–H bonds are replaced by Si–O bonds by using anodization with light assistance. After rapid thermal oxidation at 900 °C, the PL efficiency is comparable to that of Si–H–passivated PSi because the densities of dangling bonds present in as-anodized and oxidized samples are comparable. The difference is that stability of the Si/SiO_2 interface is

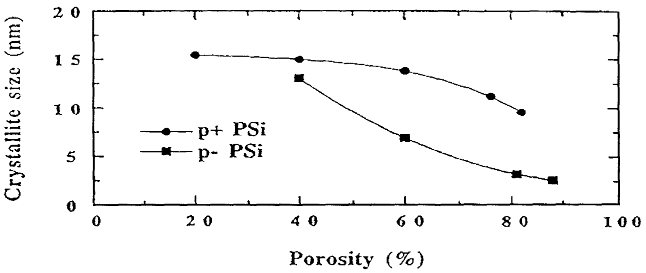

FIG. 4. The crystallite size in PSi depends on the porosity and the doping level of the starting Si wafer.

such that no additional dangling bonds are produced even under most adverse conditions. Annealing at high temperatures is necessary for forming a high-quality oxide and a high-quality Si/SiO_2 interface. To avoid the complete oxidation of the Si skeleton and the transformation of the PSi layer into a porous glass, the high-temperature oxidation is performed in an atmosphere poor in oxygen and for a short time.

2. PROPERTIES

2.1 Structural

2.1.1 X-Ray Scattering X-ray scattering from porous silicon (Lehmann *et al.*, 1993) shows that PSi simultaneously exhibits the characteristics of a monocrystalline solid (long range order) and an inhomogeneous solid (short range order) (Naudon *et al.*, 1995). The average size of the Si nanocrystallites present in PSi films can be calculated from an analysis of the x-ray lineshape by applying the Scherer equation (Cullity, 1978) to the broad peak profile. The variation of size with porosity for p^- and p^+ PSi films is plotted in Fig. 4 (Fauchet *et al.*, 1997). It should be noted that this technique is most useful when applied to samples with a single particle size. X-ray diffraction also shows that the lattice constant of PSi increases with the porosity and depends on whether the surface is covered by SiO_2 or Si–H_n. A value of the strain as large as 0.1% is possible.

2.1.2 Electron Microscopy Although techniques such as scanning-electron and atomic-force microscopies reveal some of the features of PSi, its nanostructure is best examined by transmission-electron microscopy (Berbezier, 1995; Cullis *et al.*, 1997). TEM, which is capable of a resolution down to ≈ 1 Å, has provided very detailed images of PSi. Crystallites as small as 1 nm and their lattice planes have been imaged. Scanning-electron microscopy and atomic-force microscopies do not provide the same atomic scale resolution and are better suited to the study of mesoporous and macroporous PSi.

Figure 5 shows cross-sectional TEM images of PSi prepared from a p^+ Si wafer (Beale *et al.*, 1985). In these samples, the pores are ≈ 10 nm wide and run perpendicular to the Si wafer surface (along the ⟨100⟩ direction, a general trend for all substrate orientations). For the highest porosity (79%) shown in Fig. 5, the structure becomes open, the pores are interconnected, and it is difficult to identify individual pores. In contrast, the TEM image of Fig. 6 shows that in microporous PSi formed from a p^- Si wafer, the nanostructure is coral-like or sponge-like, with characteristic dimensions on the 1–3-nm scale (Cullis *et al.*, 1992). This is the type of PSi that produces the brightest luminescence. In addition, we note that PSi can be produced with pores on two widely different length scales ($\approx$nm and sub-μm) by using light during the anodization of n^+ Si wafers.

Lattice-plane imaging shows that the crystallographic orientation of the substrate is maintained after anodization, with a trend to a small misorientation at the highest porosities. Transmission electron diffraction patterns also indicate that, in general, PSi maintains the crystalline nature of the substrate, although the presence of some disordered material has sometimes been reported.

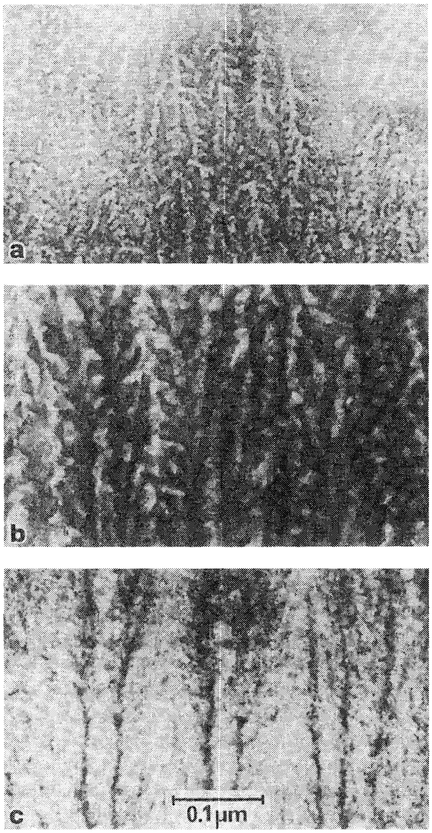

FIG. 5. Cross-sectional transmission-electron microphotographs of PSi samples prepared from a p^+ Si wafer. The porosities are 31%, 51%, and 79%, from top to bottom.

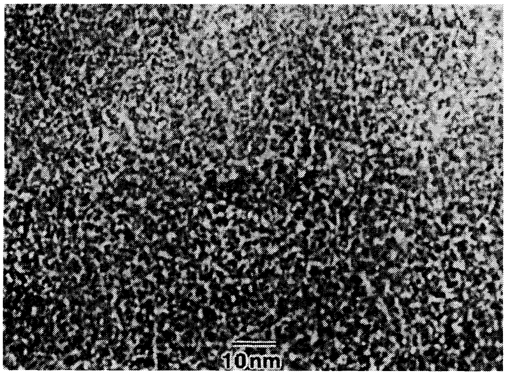

FIG. 6. Cross-sectional transmission-electron microphotographs of PSi prepared from a p^- Si wafer. The spongy structure is made of highly interconnected micropores.

2.2 Mechanical

The mechanical properties of PSi depend critically on the porosity. The open structure of high-porosity PSi makes it a very fragile material. One convenient characterization technique, which can provide useful information regarding material strength, crystallite size, and surface quality, is a hardness measurement. Hardness is defined as the resistance of a solid to plastic deformation. The method most widely used for measuring hardness is indentation. In the Vickers test, a square-based pyramidal indenter produces a diamond-shaped indent on the surface and the hardness H is related to the applied load P_L (in kg f) and the diagonal of the indent d (in mm) by

$$H = 1.8544(P_L d^{-2}) \tag{1}$$

The load dependence of hardness (indentation size effect) provides useful information regarding the nature of a material (brittle, ductile). A power law relationship exists between the applied load and the indentation diagonal:

$$P_L = K_L d^n, \tag{2}$$

where n is called the Meyer index.

Figure 7 compares the hardness of different PSi films, as a function of porosity (Duttagupta *et al.*, 1997a). The hardness of the p^+ PSi films is comparable to that of the p^- PSi films. At low porosities, there is a moderate decrease of hardness with porosity. This is expressed by the full line in Fig. 7 as $H = H_0(1-P)^m$ ($m = 0.67$), where H_0 is the hardness of crystalline Si (11.5 GPa). However, in the porosity range of 70–80%, the hardness drops appreciably and there is considerable deviation from the previous $(1-P)^{2/3}$ proportionality. This deviation indicates the change in the morphology of the films from isolated pores to isolated Si columns near that porosity. The trend of decreasing hardness with increasing porosity in confirmed by measurements of the Young's modules *versus* porosity in p^+ PSi films using nanoindentation (Bellet *et al.*, 1996).

The Meyer's index may be used to gauge the nature of a material. For Si, the value of n is 1.7 for c-Si (brittle, ordered), while that for

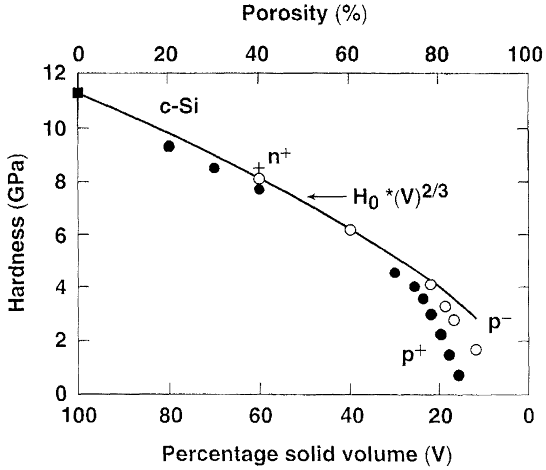

FIG. 7. The hardness of PSi depends on the porosity and the doping level of the starting Si wafer. The open circles are data from PSi prepared from p^- Si wafers, the closed circles are data from p^+ Si wafers and the cross is from n^+ Si wafers.

amorphized Si (softer, highly disordered) is 2. In PSi even at 80% porosity, $n = 1.75$ for p^- PSi and $n = 1.72$ for p^+ PSi. This is a further confirmation of the crystalline nature of porous silicon.

2.3 Thermal

The thermal conductivity of PSi depends strongly on porosity and is expected to be much below that of c-Si as the porous structure becomes less dense. At very high porosities, the preferred heat transfer could switch from conduction to radiation. The thermal conductivity of PSi has been measured by several techniques (Lang *et al.*, 1995; Duttagupta, 1997b). It depends not only on the porosity but also on the pore nanostructure (whether the material is macro-, meso-, or microporous) and on whether oxidation has been performed or not. The thermal conductivity of several PSi samples prepared from a p^+ Si wafer is shown in Fig. 8 (Duttagupta, 1997b). At high porosities, the thermal conductivity of PSi drops below that of glass.

2.4 Vibrational

The Si crystallite size in PSi films can be measured by Raman spectroscopy (Fauchet and Campbell, 1988). In unstrained bulk crystalline silicon the Si—Si peak is narrow, symmetric, and centered at 520 cm^{-1} with a full width at half maximum of approximately 4 cm^{-1}. This peak broadens, becomes asymmetric, and is downshifted with increasing porosities. In addition, a broad peak often arises around 480 cm^{-1}, which is the signature of amorphous (disordered) Si. The changes in the 520-cm^{-1} peak can be used to deduce the average crystallite size (and in some cases, its shape as well). Figure 9 shows that the crys-

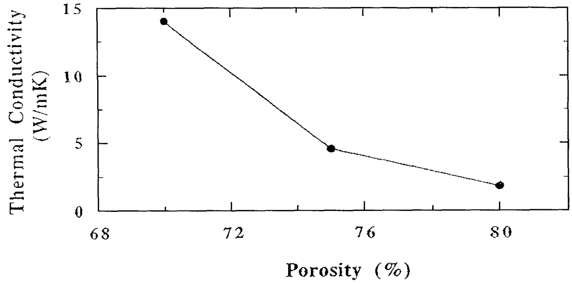

FIG. 8. The thermal conductivity of PSi decreases with increasing porosity and is much lower than that of bulk c-Si (150 W m^{-1} K^{-1}). The samples are prepared from a p^+ Si wafer.

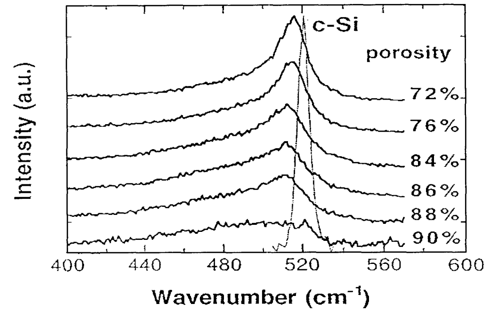

FIG. 9. Dependence of the Raman spectrum of PSi on porosity for samples prepared from a *p* Si wafer. The optic phonon Raman spectrum of bulk c-Si is shown for comparison.

tallite size decreases with increasing porosity (von Behren *et al.*, 1998). Below 25 Å, Raman spectroscopy can no longer be used as a reliable tool to measure the nanocrystal size. Additionally, stress affects the phonon line shape of nanocrystalline films, resulting in a shift of the Raman line for a uniform, homogeneous stress and a broadening of the Raman line for nonuniform, inhomogeneous stress.

2.5 Electrical

The current density (J)–voltage (V) characteristic of a simple metal/PSi/c-Si device is shown in Fig. 10 (Peng *et al.*, 1996). The J-V curve is rectifying and follows, in forward bias, a power law relationship $J = KV^m$ where $m \approx 2$. This power-law relationship is typical of a space-charge–limited current and shows that the current is dominated by carrier transport in the high-resistivity ($\approx 10^{10}\,\Omega\,\mathrm{cm}$) PSi layer. The entire device can be modeled as an insulator sandwiched between two conductors (the top metal contact and the Si substrate) with negligible band bending at the two interfaces. When the injected-carrier concentration exceeds the thermal-carrier concentration, a departure from Ohmic behavior is observed, and for double-carrier injection, the J-V relationship is given by

$$J = \varepsilon\varepsilon_0\mu_{\mathrm{eff}}V^2/d^3 \tag{3}$$

where $\varepsilon\varepsilon_0$ is the dielectric function of PSi (in $\mathrm{F\,m^{-1}}$), d is the thickness of the PSi layer, and μ_{eff} is the effective carrier mobility. The measured dependence of J on both V and d satisfies Eq. (3).

Two methods have been employed to measure the carrier mobility. In the first method, the response of the electroluminescence (EL) to a small alternating (ac) bias is measured as a function of frequency in a device held under a large forward bias (Peng and Fauchet, 1995). The injected carriers drift through the PSi layer where they recombine radiatively, and the transit time through the device (τ_t) is given by

$$\tau_t = d^2/\mu_{\mathrm{eff}}V. \tag{4}$$

When the modulation frequency of the ac bias becomes large, the electric field reverses itself in a time shorter than the transit time and the magnitude of the ac EL decreases. Figure 11 plots the 3 dB frequency f_0 (the frequency

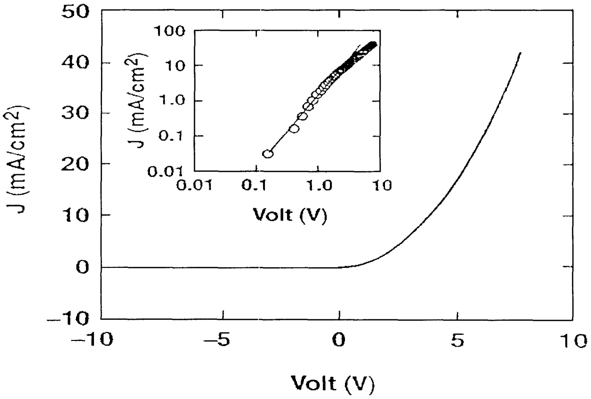

FIG. 10. Current density–voltage relation for a Au/PSi/c-Si device structure. The inset shows that the current density increases as the square of the voltage.

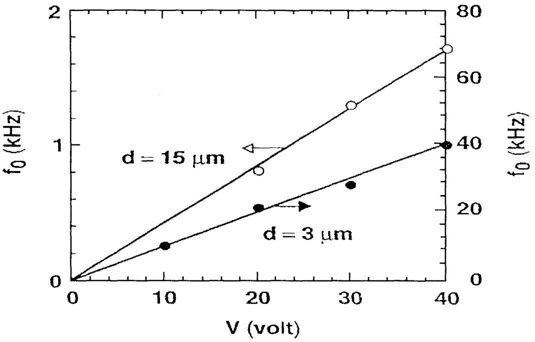

FIG. 11. The 3dB frequency of the electroluminescence response of device structures similar to that of Fig. 10 depends on the applied voltage and the thickness of the PSi layer. The solid lines are a fit using Eq. (4) and a mobility of $10^{-4}\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$.

at which the magnitude of the ac EL drops by 50%) *versus* dc bias for two devices having PSi layers of 3 and 15 μm, respectively. Other measurements have confirmed that f_0 scales with d^{-2} and with V. Thus, f_0 is equal to the inverse of τ_t, and the effective mobility obtained in PSi of ≈75% porosity is $10^{-4}\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$. This measurement does not indicate whether the mobility is that of the electrons, the holes, or a combination of both.

Another method to measure the mobility is the time-of-flight technique. Here, the device is held in reverse bias, and a short laser pulse injects electrons and holes on one side of the PSi layer. One type of carrier is then swept through the PSi layer under the effect of the electric field, producing a current that is detected as a function of time after injection. These experiments performed on PSi layers of ≈75% porosity and passivated by ≈2 monolayers of SiO_2 have shown that the electron drift mobility is between 10^{-4} and $10^{-5}\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$ and is weakly dependent on temperature from 400 to 77 K (Rao *et al.*, 1997). Electron transport is found to be dispersive and not Gaussian. This behavior has implications for the nature of electronic transport is PSi, which have not yet been explored throroughly.

2.6 Optical: Experiments

2.6.1 Absorption/Transmission The influence of the porosity (P) on the optical properties is most easily seen in transmission experiments performed on free-standing PSi films separated from the c-Si substrate by a lift-off step. The transmission spectrum of high-quality thin films is modulated by Fabry–Pérot interference fringes in the transparency region. The transmission can reach 100%, which is indicative of high-quality films having parallel front and back surfaces and negligible light scattering. The optical transmission spectra of thicker free-standing porous silicon films of various P are shown in Fig. 12 and compared with that of c-Si (von Behren *et al.*, 1998). The absorption edge shifts by more than 1.5 eV toward the blue when the porosity increases from 0% (bulk c-Si) to 92%. Since the index of refraction is close to unity for $P \geq 80\%$ (Fauchet *et al.*, 1998), the optical thicknesses of these films are equal, and the changes in transmission with porosity are a direct proof of the opening of the band gap by quantum confinement. An additional contribution comes from the fact that as the porosity increases, a smaller fraction of the volume of film is made of Si. The absorption coefficient of Si nanocrystals of average sizes from 43 to

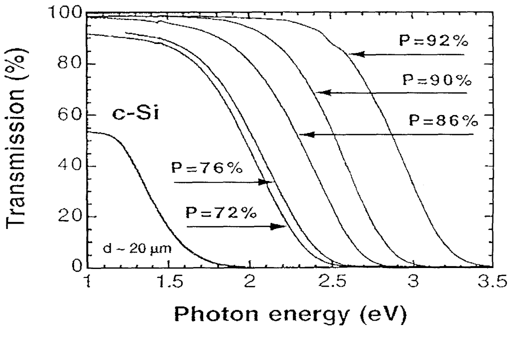

FIG. 12. The transmission spectra of 20-μm-thick PSi films depend on the porosity. The samples are prepared from a *p*-Si wafer and the spectrum of bulk c-Si is shown for comparison.

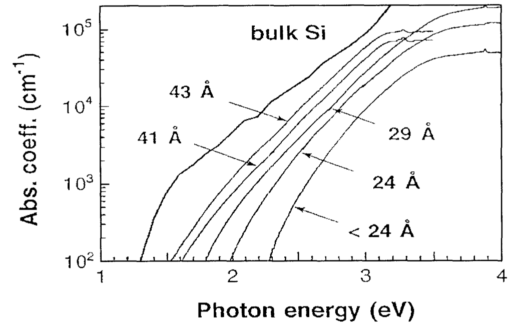

FIG. 13. The absorption spectra of the nanocrystallites present in PSi depend on the crystallite size. The samples are those of Fig. 12.

24 Å is plotted in Fig. 13 (von Behren *et al.*, 1998). The shape of the spectra shows that PSi remains an indirect-gap semiconductor even at the highest porosities/smallest crystallite sizes.

2.6.2 Dielectric Function and Index of Refraction The dielectric function ε of PSi maintains the spectral dependence of bulk crystalline Si at low porosities. With increasing porosity, the peaks in the imaginary part of ε at the E_1 and E_2 critical points broaden and decrease in amplitude. Finally, the imaginary part of ε approaches that of amorphous silicon (Koshida *et al.*, 1993), consistent with the loss of infinite long-range order. The index of refraction of PSi at or below the band gap is strongly dependent on the porosity. Figure 14 shows that the index of refraction decreases to near 1.1 at porosities of 90% and above (Fauchet *et al.*, 1998). This decrease is predicted by effective-medium theories.

2.6.3 Luminescence Radiative recombination of an electron with a hole across the band gap of a semiconductor produces luminescence at an energy equal or close to the band-gap energy. In direct-gap semiconductors such as GaAs, the electron and hole have the same momentum, the radiative recombination rate is large, and the radiative lifetime is short (a few nanoseconds). To obtain a large PL efficiency, the nonradiative recombination rate must be smaller than the radiative recombination rate. The quantum efficiency of the luminescence is defined by

$$\eta = \tau_{\text{nonrad}}/(\tau_{\text{nonrad}} + \tau_{\text{rad}}), \tag{5}$$

where η is the quantum efficiency, τ_{rad} is the radiative lifetime and τ_{nonrad} is the nonradiative lifetime. Nonradiative recombination occurs at the surface (surface recombination) and in the bulk (Auger recombination, recombination at defects).

In silicon, an indirect-gap semiconductor (see SILICON, CRYSTALLINE), the radiative recombination of an electron and a hole is possible only if another particle capable of carrying a large momentum is involved. The participation of a third particle (*e.g.*, a phonon) makes the radiative rate much lower and pushes the radiative lifetime into the millisecond regime. The quantum efficiency decreases by several orders of magnitude because the competing nonradiative processes take over, even for high-purity materials and good surface passivation. At room temperature, the quantum efficiency of bulk c-Si is usually on the order of 10^{-6}–10^{-7}. To improve this number, one can either increase the radiative rate

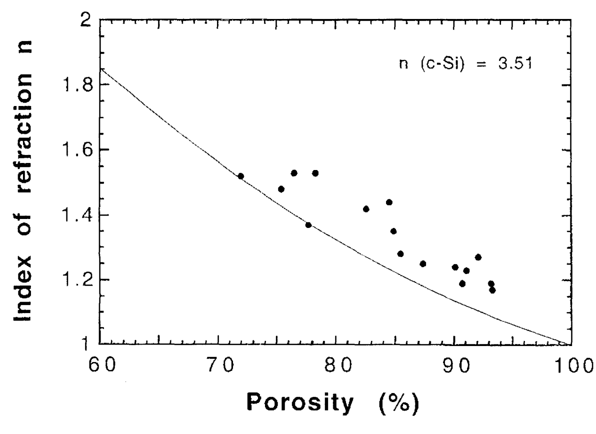

FIG. 14. The index of refraction of PSi depends on the porosity, in qualitative agreement with the prediction of an effective-medium theory, shown by the solid line.

or decrease the nonradiative rate. In PSi, both occur, although the large quantum efficiency is in large part due to the decrease in the nonradiative rate by a proper surface passivation.

The most common PL band is in the red/near-infrared part of the spectrum (Fauchet, 1996). This "red" band (or "slow" band) extends from slightly above the band gap of crystalline silicon to the blue-green part of the spectrum. The PL characteristics are wide spectrum (100–500 meV), large efficiency under UV pumping (1–10% at room temperature, greater at cryogenic temperatures), and long decay time (< 1–100 μs at room temperature, up to 10 ms at cryogenic temperatures). Figure 15 shows the evolution of a typical PSi spectrum after pulsed excitation and the temperature dependence of the PL lifetime and intensity (Vial *et al.*, 1993). In general, the shorter the emission wavelength, the shorter the PL lifetime. The PL decay does not follow a simple law and can be fitted very well using a stretched exponential function (Kanemitsu, 1993). Tuning of the red PL band can be achieved by many parameters, such as the porosity/crystallite size (by changing the HF concentration, current density, wafer doping), the use of light during anodization, or post-anodization treatments (Fauchet, 1996).

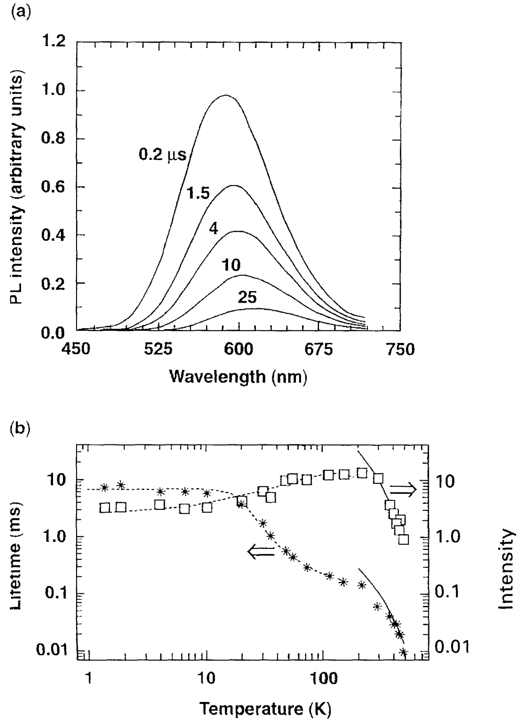

FIG. 15. (a) Temporal evolution of the PL spectrum of red-emitting PSi after pulsed UV excitation. (b) The PL intensity and the PL lifetime depend on the measurement temperature.

Although ultrahigh-porosity PSi that has not been exposed to air can luminesce weakly in the blue, strong blue PL follows high-temperature oxidation ($T_{ox} \geq 1000\,^{\circ}\mathrm{C}$) of PSi. Figure 16 shows the PL spectrum and dynamics in PSi oxidized at 1100 °C for 2 min (Tsybeskov *et al.*, 1993). The broad PL band (FWHM > 0.5 eV) with a peak near 2.6 eV and the decay on a nanosecond time scale are typical of these samples. The decay is non-exponential, as for the red band, but in contrast to the red band, no significant wavelength dependence of the "blue" (or "fast") PL decay is observed from 440 to 650 nm, the decay dynamics does not change appreciably when the measurements are performed at cryogenic temperatures, and the decay dynamics is not fitted well by a stretched exponential. The highest quantum efficiency of the blue PL can approach that of the red PL, especially after long-term exposure to water vapor and contaminants containing carbon (Loni *et al.*, 1995a).

Theory indicates that for crystallites of ≈ 1 nm diameter, the band gap should be in the blue/near UV (Wang and Zunger, 1996), and phononless, fast (approximately nanosecond) radiative recombination should dominate (Hybertsen, 1994). This model, however, cannot explain the reversible quenching of the blue PL in methanol after high-temperature oxidation of PSi (Rehm *et al.*, 1995), since the very small Si nanocrystallites, if they exist at all, are protected by a very thick oxide layer. Silicon dioxide itself is known to luminesce efficiently in the visible range under appropriate conditions, and a nanosecond blue PL has been reported after excitation by UV photons in high-purity, wet synthetic silica (Anedda *et al.*, 1993). However, the fact that the blue PL increases with exposure to water vapor and car-

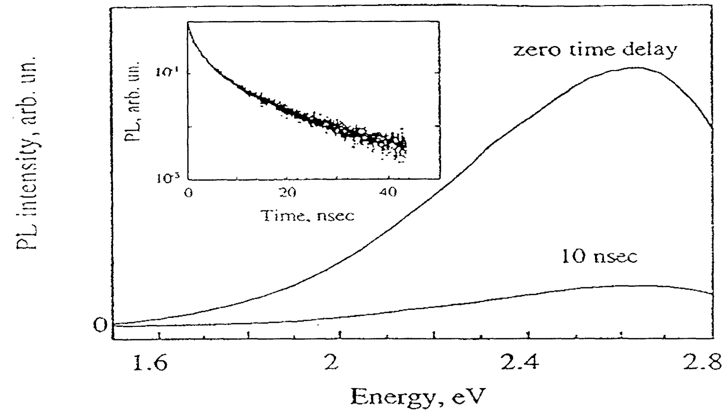

FIG. 16. Temporal evolution of the PL spectrum of blue-emitting PSi after UV excitation. The inset shows the temporal decay of the PL intensity.

bon compounds (Loni *et al.*, 1995a) strongly suggests that a chemical species adsorbed on the inner surface of the porous oxide is responsible for the blue PL. Silanol has been suggested as one candidate, but no conclusive identification has been made so far.

PSi can also emit light in the infrared. Figure 17 shows the normalized PL spectra after annealing in vacuum at different temperatures (Fauchet *et al.*, 1993). This PL band is tunable from ≈0.8 eV to past 1.2 eV, depending on the sample preparation and processing conditions, and has a nonexponential lifetime in the ≈1-μs range. Its intensity increases after annealing in ultrahigh vacuum at a temperature as high as ≈500 °C, when most of the hydrogen that passivates PSi has been desorbed and the red PL has disappeared. The tunable infrared band has been associated with dangling bonds. Figure 18 shows that the peak energies of the red and infrared bands vary linearly with each other (Petrova-Koch *et al.*, 1995). This can be explained if the infrared band corresponds to recombination of a near–band-edge electron with a hole trapped at a dangling bond whose absolute energetic position is insensitive to size, ≈0.3 eV above the bulk silicon valence-band edge.

PSi containing large crystallites also produces a PL near 1.1 eV, which coincides with

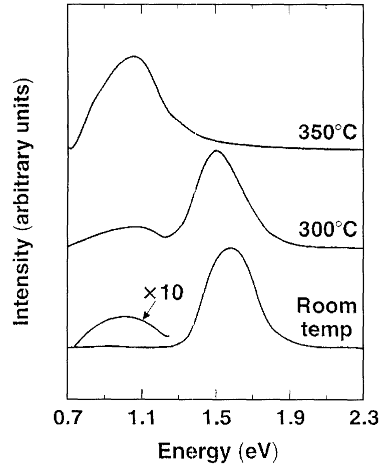

FIG. 17. Normalized PL spectra of PSi before and after annealing in vacuum at two temperatures.

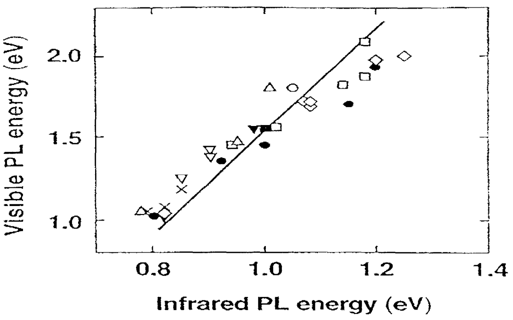

FIG. 18. Correlation between the red PL peak energy and the infrared PL peak energy over a wide range of samples prepared under different conditions. The line has a slope of 3.

the PL spectrum observed in bulk c-Si (Tsybeskov *et al.*, 1996a), including all the phonon replicas. As the temperature increases toward 300 K, the PL spectrum widens, but the PL efficiency remains unchanged near or above 0.1%. The PL is due to radiative recombination in well-passivated, large crystallites ($\gg 10$ nm).

2.7 Optical: Theory

2.7.1 Band Gap The band gap of low-dimensional Si nanostructures is larger than that of bulk Si because of quantum confinement in one, two or three dimensions. In the simple particle-in-a-box model, when electrons in the conduction band and holes in the valence band are confined spatially by potential barriers, the lowest energy for optical transitions increases by

$$\Delta E = h^2/8[1/m_c + 1/m_v] \sum 1/L_i^2 \qquad (6)$$

where h is Planck's constant, m_c and m_v are the effective masses in the conduction and valence bands respectively, L_i is the dimension of the confined region in direction i, and the number of confined directions may be 1, 2, or 3 corresponding to a quantum well, wire, or dot, respectively (Collins *et al.*, 1997). This expression assumes that the bands are parabolic, there is only one conduction (valence) band, and the potential barrier is infinitely large. Thus, it can only be used as a guide, since none of these assumptions is entirely valid.

Various techniques have been employed to calculate the band gap of Si quantum wells, wires, and dots. Calculations have been performed using different types of tight-binding and local-density approximations, mostly using a (semi)empirical approach (Delerue *et al.*, 1993; Wang and Zunger, 1994; Delley and Steigmeier, 1995; Ogut *et al.*, 1997). Figure 19 compares the band gap for wells, wires, and dots calculated by the effective-mass approximation and by a pseudopotential approach (Wang and Zunger, 1996). The band gap increases faster with decreasing size when the number of confined directions increases. The effective-mass approximation overestimates the band gap by a very large amount (≈ 1 eV for ≈ 1 nm wells, ≈ 2 nm wires, and ≈ 3 nm dots). This overestimate occurs because theory shows that ΔE scales with size as $L^{-1.3}$ instead of L^{-2} (Delerue *et al.*, 1993). Furthermore, theory (Ren and Dow, 1992) and experiments (van Buuren *et al.*, 1993) indicate that the quantization energy in the valence band is larger than in the conduction band [$\Delta E_v \approx (1.5\text{–}2) \times \Delta E_c$], also in contrast with the effective-mass approximation (which predicts that $\Delta E_v \sim 1/m_v < \Delta E_c \sim 1/m_c$).

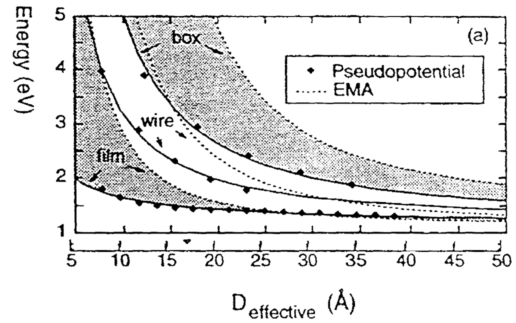

FIG. 19. Comparison of the band gap calculated by the effective-mass approximation [Eq. (6)] and a pseudo-potential approach, for Si quantum dots (box), wires, and wells (films).

Most of the calculations have been performed for H-terminated Si nanostructures. It is expected that the exact numerical answers will be affected by a change in the surface termination, especially a change from H to O passivation following partial oxidation. At this time, the models are not sophisticated enough to account properly for effects such as strain that exist at such an interface.

Despite the profound effect that quantum confinement has on the magnitude of the band gap of the Si nanocrystals present in porous silicon, the nature of the optical transition remains identical to that of bulk c-Si, in the sense that phonon-assisted optical transitions continue to predominate over transitions involving no phonons. Theory suggests that the oscillator strength for phonon-assisted transitions continues to exceed that for phononless transitions down to sizes of ≈ 1 nm (Hybertsen, 1994). Thus porous silicon behaves as an indirect-gap semiconductor, as confirmed by the presence of pronounced phonon steps in resonant PL measurements performed at low temperatures (Calcott *et al.*, 1993).

2.7.2 Luminescence The models that attempt to explain the luminescence in porous silicon can be divided into two broad categories (Fauchet 1996):

1. The PL occurs in quantum-confined crystalline silicon nanostructures, and
2. the PL occurs elsewhere—for example, in a disordered tissue material, in specific bonds or molecules (such as siloxene) that would be present on or attached to the surface, or in the nonbridging oxygen hole center present in silicon dioxide.

The overwhelming experimental evidence indicates that carriers confined in the Si nanocrystals produce the strong, tunable PL. The remaining question is whether the PL energy coincides with the band-gap energy. If recombination involves carriers confined to the nanocrystal, the PL energy should equal the band gap minus the binding energy of the exciton. If, on the other hand, the PL energy is less than the band gap, one or both carriers may be trapped, either in surface states (Koch, 1993) or as self-trapped excitons (Allan *et al.*, 1996). Because the nanocrystals present in PSi do not all have the same size, the absorption edge reflects the presence of large crystallites while the PL is dominated by the smaller crystallites, which are less likely to contain a defect or to suffer from Auger recombination. Thus, the PL and absorption measurements may not sample the same fraction of the nanocrystal population, which makes a comparison difficult.

Changes in the surface chemistry and in the chemical environment can lead to modifications in the PL properties. Although this has been invoked to support recombination involving trapped carriers, changes in surface properties can also modify the electronic states in the interior of the nanocrystals. Despite the difficulties with a comparison of the band-gap and PL energies as mentioned above, this is the only type of experiment that can directly address the question.

The band gap measured from absorption measurements should be equal to the one-electron band gap minus the binding energy of the free exciton confined to the nanocrystal. The free-exciton binding energy increases with decreasing radius size (Wang and Zunger, 1996) for two major reasons: namely, because the Coulombic attraction increases when the exciton is confined to a radius smaller than the bulk-exciton Bohr radius, and because the dielectric constant decreases dramatically for sizes below 2–3 nm. Realistic calculations of the exciton binding energy indicate that the binding energy approaches 0.5 eV for sizes of ≈2 nm. The calculated excitonic band gap, defined as the one-electron band gap minus the exciton binding energy, is in agreement with measurements of the absorption edge, in the size regime below 4.5 nm (Fig. 20) (von Behren *et al.*, 1998). The measured PL peak energy is consistently lower than the band-gap energy, especially for smaller sizes. This Stokes shift between band-gap and PL energies increases with decreasing size, as shown in Fig. 21 (Fauchet, 1998). The origin of this Stokes shift is still under discussion. A possible candidate is that the exciton becomes self-trapped at the surface. Model calculations of self-trapping at the Si–H bond show that the nanocrystal size necessary for self-trapping is much too small, below 1 nm (Allan *et al.*, 1996). However, trapping of an electron, or a hole, or even an exciton, could occur for more reasonable sizes with oxide-covered nanocrystals. This would produce, for smaller nanocrystallites, an increase of the PL energy that is much smaller than the increase of the band-gap energy.

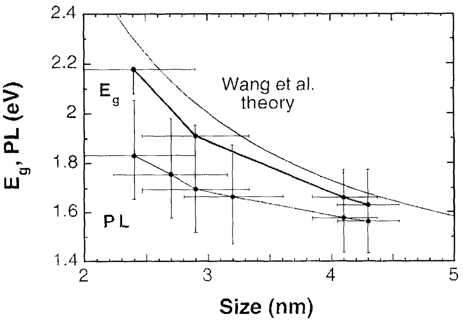

FIG. 20. Measured and calculated band gaps (E_g) for Si nanocrystallites of different sizes. The PL peak energy in samples exposed to air is also shown for comparison.

3. APPLICATIONS

3.1 Light-emitting Devices

Electroluminescence (EL) was observed in PSi shortly after Canham's report of strong

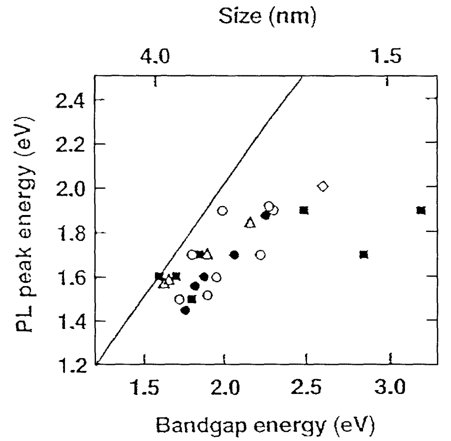

FIG. 21. Dependence of the PL peak energy on the band gap and crystallite size for the Si nanocrystals found in PSi. The data would fall on the solid line if the two energies were identical.

room-temperature PL—first in solution during anodic oxidation (Halimaoui *et al.*, 1991) and then in solid-state devices (Richter *et al.*, 1991; Kalkhoran *et al.*, 1992; Bassous *et al.*, 1992; Koshida and Koyama, 1992). Many different device structures have been demonstrated, and strong (0.1%–1% efficiency), voltage-tunable EL has even been reported under cathodic polarization of n-type PSi using liquid contacts (Bsiesy *et al.*, 1993). Research has focused on solid-state devices because of their technological importance and because liquid contacts are inherently unstable. A typical PSi light-emitting device (LED) consists of a transparent or semitransparent contact (metal, ITO, or conducting polymers) and a 1- to 10-μm-thick PSi layer on a crystalline silicon (c-Si) substrate (*p* or *n* type). Typical threshold conditions for EL are $V \approx 10\,\mathrm{V}$ and $J \approx 10\,\mathrm{mA\,cm^{-2}}$, although for the best devices V is now $\approx 2\,\mathrm{V}$ and $J \ll 1\,\mathrm{mA\,cm^{-2}}$ (Loni *et al.*, 1995b). The EL external quantum efficiency has also improved to >0.1%, and the irreversible degradation that plagued the early devices has been eliminated.

Initially, the stability of the early PSi LEDs was poor because the Si–H bonds that passivate the Si nanocrystal surfaces are very fragile and can be easily broken by exposure to light, ambient air, moderate temperatures, and large electric fields. To improve the stability, the fragile Si–H bonds have been replaced by the stronger Si–O bonds and the devices have been engineered to provide better heat sinking and greater mechanical stability, leading to no degradation after weeks of operation under ambient conditions. Figure 22 shows the results of stability tests, performed under pulsed drive conditions well above EL threshold for a PSi LED (Tsybeskov *et al.*, 1996b). No degradation is seen for several weeks. A key factor in achieving stable devices is to minimize the density of states at the interfaces between the different layers present in the structure. Devices that contain too many interface states suffer from reversible degradation, produced by the charging of the interface states (Fauchet *et al.*, 1997).

The two main factors that keep the EL efficiency one to two orders of magnitude below the PL efficiency are the difficulties in injecting carriers from the contact into the PSi and the poor transport properties in PSi. By improving the contacts and keeping the PSi layer thickness below 1 μm, the power efficiency of the

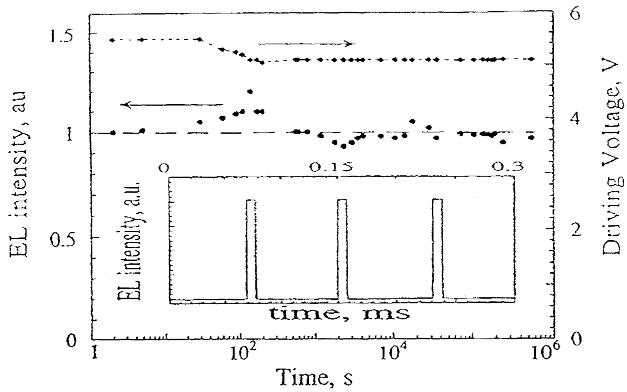

FIG. 22. Stability of the electroluminescence intensity under pulsed current excitation conditions for a PSi LED.

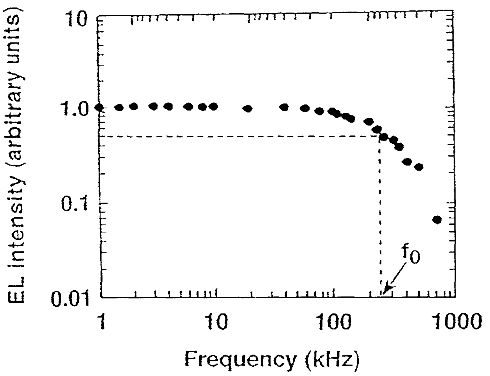

FIG. 23. The amplitude of the ac electroluminescence depends on the modulation frequency. The measurement was performed on a p–n PSi junction LED. The 3-dB frequency is indicated.

best PSi LEDs has been increased to greater than 0.1%. The internal quantum efficiency, defined as the ratio of the number of photons generated inside the PSi layer to the number of carriers injected inside the PSi layer, is estimated to be $>1\%$, which is comparable to the PL efficiency. The threshold voltage (2 V) and current density ($1\,\mu\mathrm{A\,cm^{-2}}$) of some of these devices makes them compatible with microelectronic technology.

The response time of PSi LEDs is an important parameter that determines their suitability in applications such as optical interconnects and real-time optical signal processing. The speed of the LEDs can be obtained by measuring the frequency response in the small-signal limit. The device is held at a large forward bias, well above the threshold for detectable EL. A small ac voltage is applied, which modulates the EL. The amplitude of the ac EL is monitored as a function of frequency, as already mentioned in Sec. 2.5. Figure 23 shows that the magnitude of the ac EL signal remains constant up to a critical frequency beyond which it drops quickly (Peng and Fauchet, 1995). The 3-dB frequency, which is defined as the frequency at which the ac EL has dropped by a factor of 2, has been found in Sec. 2.5 to be the inverse of the transit time. The 3-dB frequency determines the maximum frequency response of the diode in this small-signal analog test. Thus, the response time of the EL is limited by the time it takes carriers to cross the PSi layer, and not the PL lifetime, which is usually shorter. By making the PSi layer thinner than 1 μm or by using a PSi *p-n* junction, the 3-dB frequency can approach 1 MHz, when it is limited by the luminescence lifetime. Clearly, the small-signal ac modulation speed of these PSi LEDs is not likely to approach 1 GHz, unless they can be made using blue-emitting PSi, whose PL lifetime is ≈ 1 ns.

Since the PL can span the spectrum from the infrared to the blue, it is reasonable to expect that EL can cover the same range of wavelengths. As early as 1992, LEDs with peak wavelengths ranging from the deep red to the blue were fabricated. The EL efficiency of these devices was between 0.005% and 0.01%. LEDs operating in the infrared are also attractive. For example, 1.5-μm radiation is eye-safe and compatible with fiber communication systems. In addition, it is not absorbed by the c-Si wafer, which may be an advantage for the integration of PSi LEDs in close proximity to microelectronic circuits. PSi LEDs emitting 1.1-μm light and 1.5-μm light have been demonstrated, using oxidized mesoporous Si and Er-doped oxidized PSi, respectively (Fig. 24) (Fauchet *et al.*, 1997).

Ultimately, the success or failure of PSi-based technology rests on the compatibility of PSi LEDs with standard microelectronic processing. An integrated bipolar transistor/PSi LED structure has been demonstrated (Hirschman *et al.*, 1996). The complete structure, shown in Fig. 25, was fabricated using accepted silicon microelectronic fabrication procedures. The LED could be turned on and off by applying a small current pulse to the base of the bipolar transistor. Arrays of such integrated structures can easily be fabricated. This represents a critical first step toward the integration of silicon-based LEDs into VLSI circuits.

3.2 Solar Cells

Solar cells have been made using PSi, either as the photovoltaic material or as a passive layer intended to improve the performance of the c-Si solar cell. The use of PSi in photovoltaic devices has been reported as early as 1982 (Prasad *et al.*, 1982). This involved the anodization of p^+/n c-Si solar cells to form a

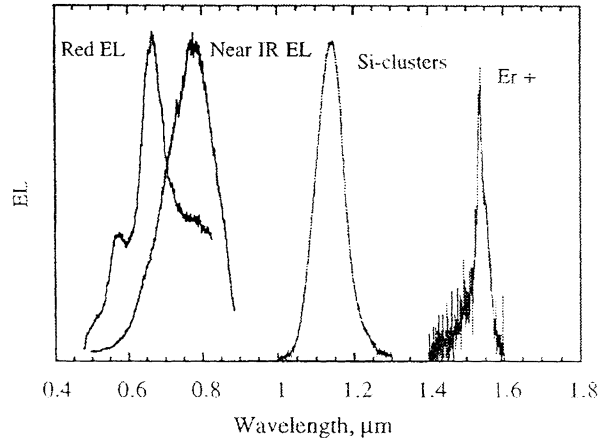

FIG. 24. Electroluminescence spectra for four PSi LEDs. The red/near-infrared LEDs were prepared with a nanoporous light-emitting layer. The two infrared LEDs were prepared with a mesoporous light-emitting layer, which was doped with erbium for the 1.5-µm device.

thin porous layer, which was oxidized at 500–600 °C. As a result, the reflectance of the solar cell was reduced from 37% to 8% in the 300- to 800-nm region, which increased the short-circuit current density (J_{sc}) by 25%. The potential advantages of PSi-based photovoltaic devices include the following:

1. Its highly textured surface increases light trapping and reduces loss due to reflection.
2. The band gap of PSi may be optimized for the absorption of sunlight (the maximum theoretical solar-cell efficiency vs. band gap curve peaks at around 1.5 eV).
3. The efficient photoluminescence of PSi might be used to down-convert higher-energy solar radiation (blue/UV) into longer-wavelength light (red/IR), which is absorbed more efficiently in bulk silicon.

Solar cells in which PSi is the photovoltaic material have been demonstrated with an efficiency in excess of 10% (Duttagupta and Fauchet, 1998). A µc-Si film deposited on *n*-type (100) c-Si substrates (1 Ω cm) at 580 °C using low-pressure chemical vapor deposition was doped *in situ* with phosphorus (SiH_4/PH_3) and boron (SiH_4/B_2H_6). The p^+-n films were anodized, and the average porosity *P* was varied to achieve maximum efficiency. Under illumination, an ideal solar cell is equivalent to a constant-current generator in parallel to a *p*-*n* junction diode. With increasing porosity, the series resistance first increases slightly and

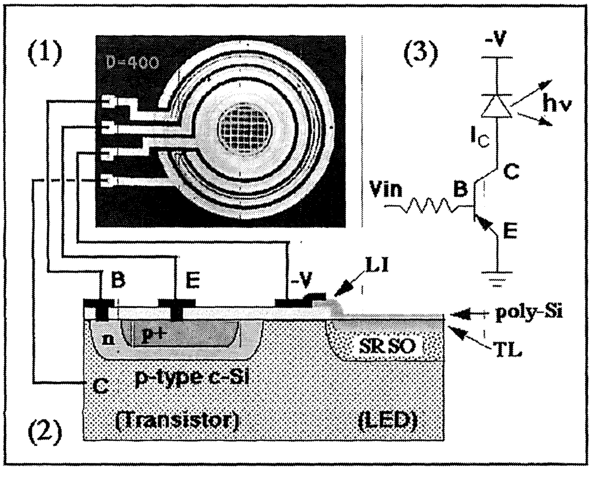

FIG. 25. Photograph, cross section, and equivalent circuit of a PSi LED integrated with a bipolar transistor. The LED surface is covered by the cross-hatched Al contacts. The transistor surrounds the LED. The cross section starts on the right at the center of the LED. (TL is a mesoporous layer and LI is the contact to the metal electrodes.)

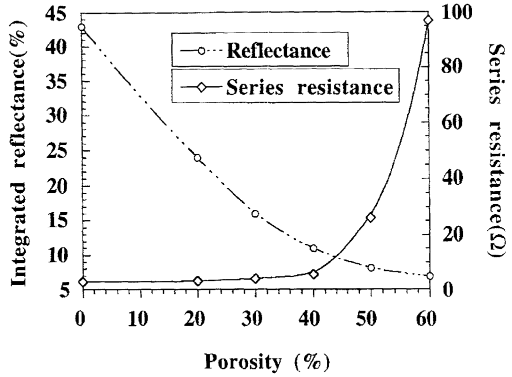

FIG. 26. Dependence of the reflectance and the series resistance on porosity for a solar-cell device made of porous microcrystalline silicon.

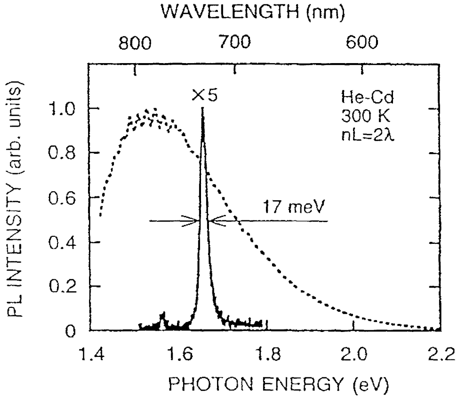

FIG. 27. Narrowing of the PL spectrum when porosity superlattices are used to shape the emission spectrum of PSi.

then, above a porosity of 50%, increases by several orders of magnitude. In contrast, the reflectance decreases monotonically with increasing porosity (Fig. 26). As a result, the short-circuit current has a maximum around 40% porosity. The best solar energy efficiency of 10.2% has been achieved for devices with a 40% average porosity, which also exhibit an enhanced blue response Note that photodetectors, which are devices closely related to solar cells, have also been made using PSi (Zheng *et al.*, 1992).

3.3 Mirrors and Photonic Band Gap

High-reflectivity mirrors made of PSi can be fabricated with reflectivities in excess of 99%. They consist of a stack of alternating high-porosity/low-porosity layers (Berger *et al.*, 1994), produced by changing the current density periodically during anodization or by anodizing under a constant current density alternating c-Si layers having high and low doping concentrations. These porosity superlattices, which can be viewed as Fabry–Pérot resonators or Bragg reflectors, have a high reflectivity in specific wavelength regions only. They have been inserted in PSi structures to narrow the spectral width of the luminescence by more than one order of magnitude, as shown by Fig. 27 (Araki *et al.*, 1996).

PSi containing periodically spaced macropores (initiated by means of a mask on the wafer) has been used as a two-dimensional photonic band gap structure (Gruning *et al.*, 1996). A photonic band gap material is to photons what a semiconductor crystal is to electrons: it prevents the propagation of light in a range of photon energies by the destructive interference produced by the alternating regions of high index of refraction (Si) and low index of refraction (pores), provided that the periodicity is on the order of the wavelength in PSi. Attenuation of electromagnetic waves by more than two orders of magnitude has been observed near 5 μm using a triangular lattice of 75-μm-deep pores spaced by 2.3 μm.

3.4 Waveguides

Optical waveguides require that a region of high index of refraction be sandwiched between regions of low index of refraction. Because the index of refraction of PSi can easily be controlled by changing the porosity, PSi waveguides can be readily fabricated. PSi waveguides can be useful in Si optoelectronic integrated circuits (OEICs), where they could manipulate optical signals. PSi waveguides with losses as small as $\leq 1\,\mathrm{dB\,cm^{-1}}$ have been demonstrated (Bondarenko *et al.*, 1995). They have also been used to collect efficiently the electroluminescence produced by PSi LEDs.

3.5 Sensors

Because the properties of PSi are very sensitive to the state of its internal surface, it should be a useful material in detecting the presence of chemical compounds, humidity, *etc.* PSi sensors could operate on the basis of a reversible or irreversible change in optical properties (PL), electrical properties (conductivity), or even structural properties. PSi sensors have been developed for detecting humidity in integrated circuits. They have also been proposed for detecting environmentally important chemicals (smokestack gases) and alcohols (breathalyzers) (Sailor and Lee, 1997). Although PSi sensors with promising properties have been demonstrated, one common problem is the lack of specificity.

3.6 SOI Technology

In the silicon-on-insulator (SOI) technologies, the electronic devices are manufactured on a thin Si film on an insulating substrate, which minimizes problems associated with parasitic capacitances, latch up, and radiation damage. PSi was first used in the FIPOS (full isolation by porous oxidized silicon) process and then in a variety of SOI applications (Tsao, 1987). In these applications, the oxide layer was formed by the full oxidation of PSi usually formed from heavily doped layers. This approach has been abandoned in favor of alternative techniques.

3.7 Micromachining

Micromachining of Si is a technological area of rapidly growing importance because it allows the development of micro-electro-mechanical systems (MEMS) devices that can be used in a variety of applications, such as pressure sensors. PSi, as a sacrificial layer, is well suited to micromachining (Lang, 1996) because the etching rate of Si depends strongly on the doping density and the dopant type and on the current density. An etching hierarchy (from most easily etched to less easily etched) has been introduced: n^+ Si under illumination, p^+ Si under illumination, p and n Si under illumination, p Si in the dark, n Si in the dark and amorphized Si. A combination of etch stops (*e.g.*, using layers of different resistivities) and masking techniques (using photoresists and/or silicon nitride or carbide protective layers) permits the micromachining of Si in three dimensions. Examples of integrated microsensors fabricated using the PSi technology include bolometers.

3.8 Biomedical Applications

Electronic components are increasingly used inside the human body (*e.g.*, pacemaker, drug delivery systems). In these applications, the Si chip is isolated from the biological medium by a package that is more bio-compatible or has the desired bioactivity. Unlike bulk Si, which is mostly bio-inert, PSi has been shown to induce the nucleation and growth of calcium phosphate (Canham, 1995). This implies that PSi can naturally coat to bone and other soft tissues in the human body. Interestingly, the kinetics of calcification can be controlled electrically. Thus, chips containing PSi regions could be used to bind to specific sites on demand. This field is just in its early infancy, and future developments (Canham *et al.*, 1997) will tell whether the early promises transform into real applications.

3.9 Capacitors

Capacitors have been manufactured using PSi technology (Lehmann *et al.*, 1995). Macroporous PSi was formed by seeding the pores with a 3.5-μm pitch. The resulting gain in surface area over a bulk sample was 85. The PSi was then doped heavily n-type and a 30-nm-thick dielectric layer was grown inside the pores. Such devices, which showed excellent characteristics, can be integrated on the back side of Si chips, thus saving valuable real estate and avoiding parasitic effects in high-quality RC timing circuits.

GLOSSARY

Hole: An empty energy level in the valence band of a semiconductor; it behaves as a particle carrying a positive charge.

Luminescence: Emission of a photon during the recombination of an electron with a hole. If the electron and hole have been injected optically, it is called photoluminescence; if they have been injected electrically, it is called electroluminescence.

Pore: Region from which the silicon atoms have been removed by the etching process.

Works Cited

Allan, G., Delerue, C., Lannoo, M. (1996), *Phys. Rev. Lett.* **76**, 2961–2964.

Anedda, A., Bongiovanni, G., Cannas, M., Congiu, F., Mura, A., Martini. M. (1993), *J. Appl. Phys.* **74**, 6993.

Araki, M., Koyama, H., Koshida, N. (1996), *Jpn. J. Appl. Phys.* **35**, 577–580.

Bassous, E., Freeman, M., Halbout, J.-M., Iyer, S. S., Kesan, V. P., MMungia, P., Pesarcik, S. F., Williams, B. L. (1992), in: S. S. Iyer, R. T. Collins, L. T. Canham (Eds.), *Light Emission from Silicon*, MRS Symposium Proceedings Vol. 256, Pittsburgh: Materials Research Society, p. 23.

Beale, M. I. J., Benjamin, J. D., Uren, M. J., Chew, N. G., Cullis, A. G. (1985), *J. Cryst. Growth* **73**, 622.

Bellet, D., Lamagnere, P., Vincent, A., Brechet, Y. (1996), *J. Appl. Phys.* **80**, 3772.

Berbezier, I. (1995), in: J.-C. Vial, J. Derrien (Eds.), *Porous Silicon Science and Technology*, Berlin: Springer-Verlag, p. 207.

Berger, M. G., Dieker, C., Thonissen, M., Vescan, L., Luth, H., Munder, H., Thiess, W., Wernke, M., Grosse, P. (1994), *J. Phys. D: Appl. Phys.* **27**, 1333.

Bondarenko, V. P., Dorofeev, A. M., Kazuchits, N. M. (1995), *Microelectron. Eng.* **28**, 447.

Bsiesy, A., Muller, F., Ligeon, M., Gaspard, F., Herino, R., Romestain, R., Vial, J.-C. (1993), *Phys. Rev. Lett.* **71**, 637–641.

Calcott, P. D. J., Nash, K. J., Canham, L. T., Kane, M. J., Brumhead, D. (1993), *J. Phys.: Cond. Matter* **5**, L91.

Canham, L. T. (1990), *Appl. Phys. Lett.* **57**, 1046–1048.

Canham, L. T. (1995), *Adv. Mater.* **7**, 1033.

Canham, L. T., Cullis, A. G., Pickering, C., Dosser, O. D., Cox, T. I., Lynch, T. P. (1994), *Nature (London)* **368**, 133.

Canham, L. T., Reeves, C. L., Wallis, D. J., Newey, J. P., Houlton, M. R., Sapsford, G. J., Godfrey, R. E., Loni, A., Simons, A. J., Cox, T. I., Ward, M. C. L. (1997), in: R. W. Collins, P. M. Fauchet, I. Shimizu, J.-C. Vial, T. Shimada, A. P. Alivisatos, (Eds.), *Advances in Microcrystalline and Nanocrystalline Semiconductors—1996*, MRS Symposium Proceedings Vol. 452, Pittsburgh: Materials Research Society, pp. 579–590.

Collins, R. T., Tischler, M. A., Stathis, J. H. (1992), *Appl. Phys. Lett.* **61**, 1649–1651.

Collins, R. T., Fauchet, P. M., Tischler, M. A. (1997), *Phys. Today* **50**, 24–31.

Cullis, A. G., Canham, L. T., Dosser, O. D. (1992), *Mat. Res. Soc. Symp. Proc.* **258**, 7.

Cullis, A. G., Canham, L. T., Calcott, P. D. J. (1997), *J. Appl. Phys.* **82**, 909–965.

Cullity, B. D. (1978), *Elements of X-Ray Diffraction*, 2nd ed., Reading, MA: Addison-Wesley.

Delerue, C., Allan, G., Lannoo, M. (1993), *Phys. Rev. B* **48**, 11024.

Delley, B., Steigmeier, E. F. (1995), *Appl. Phys. Lett.* **67**, 2370.

Duttagupta, S. P., Chen, X. L., Jenekhe, S. A., Fauchet, P. M. (1997a), *Solid State Commun.* **101**, 33–37.

Duttagupta, S. P., Fauchet, P. M., Chen, X. L., Jenekhe, S. A. (1997b), in: R. W. Collins, P. M. Fauchet, I. Shimizu, J.-C. Vial, T. Shimada, A. P. Alivisatos, (Eds.), *Advances in Microcrystalline and Nanocrystalline Semiconductors—1996*, MRS Symposium Proceedings Vol. 452, Pittsburgh: Materials Research Society, pp. 473–478.

Duttagupta, S. P., Fauchet, P. M. (1998), unpublished.

Duttagupta, S. P., Peng, C., Fauchet, P. M., Kurinec, S. K., Blanton, T. N. (1995), *J. Vac. Sci. Technol. B* **13**, 1230.

Fauchet, P. M. (1996), *J. Luminesc.* **70**, 294–309.

Fauchet, P. M. (1997), in: *Pits and Pores: Formation, Properties and Significance for Advanced Luminescent Materials*, Pennington, NJ: The Electrochemical Society, pp. 27–60.

Fauchet, P. M. (1998), in: *Light Emission in Silicon: From Physics to Devices*, Semiconductors and Semimetals, Vol. 49, San Diego, CA: Academic Press, pp. 206–252.

Fauchet, P. M., Campbell, I. H. (1988), *Crit. Rev. Solid State Mater. Sci.* **14**, S79–S101.

Fauchet, P. M., Ettedgui, E., Raisanen, A., Brillson, L. J., Seioferth, F., Kurinec, S. K., Gao, Y., Peng, C., Tsybeskov, L. (1993), in: M. A. Tischler, R. T. Collins, M. L. W. Thewalt, G. Abstreiter (Eds.), *Silicon-Based Optoelectronic Materials*, MRS Symposium Proceedings Vol. 298, Pittsburgh: Materials Research Society, p. 271.

Fauchet, P. M., Tsybeskov, L., Duttagupta, S. P., Hirschman, K. D. (1997), *Thin Solid Films* **297**, 254–260.

Fauchet, P. M., von Behren, J., Hirschman, K. D., Tsybeskov, L., Duttagupta, S. P. (1998), *Phys. Status Sol.* **165**, 3–13.

Gruning, U., Lehmann, V., Engelhardt, C. M. (1996), *Appl. Phys. Lett.* **66**, 3254–3256.

Halimaoui, A., Oules, C., Bomchil, G., Bsiesy, A., Gaspard, F., Herino, R., Ligeon, M., Muller, F. (1991), *Appl. Phys. Lett.* **59**, 304.

Hirschman, K. D., Tsybeskov, L., Duttagupta, S. P., Fauchet, P. M. (1996), *Nature (London)* **384**, 338–340.

R. Hummel, R., Morrone, A., Ludwig, M., Chang, S. S. (1993), *Appl. Phys. Lett.* **63**, 2771.

Hybertsen, M. S. (1994), *Phys. Rev. Lett.* **72**, 1514.

Kalkhoran, N. M., Namavar, F., Maruska, H. P. (1992), in: S. S. Iyer, R. T. Collins, L. T. Canham (Eds.), *Light Emission from Silicon*, MRS Symposium Proceedings Vol. 256, Pittsburgh: Materials Research Society, p. 84.

Kanemitsu, Y. (1993), *Phys. Rev. B* **48**, 12,357.

Koch, F. (1993), in: M. A. Tischler, R. T. Collins, M. L. W. Thewalt, G. Abstreiter (Eds.), *Silicon-Based Optoelectronic Materials*, MRS Symposium Proceedings Vol. 298, Pittsburgh: Materials Research Society, p. 319.

Koshida, N., Koyama, H. (1992), *Appl. Phys. Lett.* **60**, 347.

Koshida, N., Koyama, H., Suda, Y., Yamamoto, Y., Araki, M., Saito, T., Sata, K., Sata, N., Shin, S. (1993), *Appl. Phys. Lett.* **63**, 2774–2776.

Koyama, H., Koshida, N. (1993), *J. Appl. Phys.* **74**, 6365.

Lang, W. (1996). *Mater. Sci. Eng.* **R17**, 1–55.

Lang, W., Drost, A., Steiner, P., Sandmaier, H. (1995), in: R. W. Collins, C. C. Tsai, M. Hirose, F. Koch, L. Brus (Eds.), *Microcrystalline and Nanocrystalline Semiconductors*, MRS Symposium Proceedings Vol. 358, Pittsburgh: Materials Research Society, p. 561.

Lehmann, V., Cerva, H., Gosele, U. (1992), in: S. S. Iyer, R. T. Collins, L. T. Canham (Eds.), *Light Emission from Silicon*, MRS Symposium Proceedings Vol. 256, Pittsburgh: Materials Resarch Society, p. 3.

Lehmann, V., Jobst, B., Muschik, T., Kux, A., Petrova-Koch, V. (1993), *Jpn. J. Appl. Phys.* **32**, 2095.

Lehmann, V., Honlein, W., Reisinger, H., Spitzer, A., Wendt, H., Willer, J. (1995), *Solid State Technol.*, November 1995, 99.

Littau, K. A., Szajowski, P. J., Miller, A. J., Kortan, A. R., Brus, L. E. (1993), *J. Phys. Chem.* **97**, 1224.

Loni, A., Simons, A. J., Calcott, P. D. J., Canham, L. T. (1995a), *J. Appl. Phys.* **77**, 3557.

Loni, A., Simons, A. J., Cox, T. I., Calcott, P. D. J., Canham, L. T. (1995b), *Electron. Lett.* **31**, 1288.

Lu, Z. H., Lockwood, D. J., Baribeau, J.-M. (1995), *Nature (London)* **378**, 258.

Nassiopoulos, A. G., Grigoropoulos, S., Papadimitrou, D., Gogolides, E. (1995), *Phys. Status Sol. (b)* **190**, 91.

Naudon, A., Goudeau, P., Vezin, V. (1995), in: J.-C. Vial, J. Derrien (Eds.), *Porous Silicon Science and Technology*, Berlin: Springer-Verlag, p. 255.

Ogut, S., Chelikowsky, J. R., Louie, S. G. (1997), *Phys. Rev. Lett.* **79**, 1770–1773.

Peng, C., Fauchet, P. M. (1995), *Appl. Phys. Lett.* **67**, 2515–2517.

Peng, C., Hirschman, K. D., Fauchet, P. M. (1996), *J. Appl. Phys.* **80**, 295–300.

Petrova-Koch, V., Muschik, T., Polisski, G., Kovalev, D. (1995), in: R. W. Collins, C. C. Tsai, M. Hirose, F. Koch, L. Brus (Eds.), *Microcrystalline and Nanocrystalline Semiconductors*, MRS Symposium Proceedings Vol. 358, Pittsburgh: Materials Research Society, p. 483.

Petrova-Koch, V., Fischer, T., Sheglov, K., Koch, F. (1996), in: D. J. Lockwood, P. M. Fauchet, N. Koshida, S. R. J. Brueck (Eds.), *Advanced Luminescent Materials*, Perrington, NJ: The Electrochemical Society, p. 382.

Prasad, A., Balakrishnan, S., Jain, S. K., Jain, G. C. (1982), *J. Electrochem. Soc.*, **129**, 596.

Rao, P., Schiff, E. A., Tsybeskov, L., Fauchet, P. M. (1997), in: R. W. Collins, P. M. Fauchet, I. Shimizu, J.-C. Vial, T. Shimada, A. P. Alivisatos, (Eds.), *Advances in Microcrystalline and Nanocrystalline Semiconductors—1996*, MRS Symposium Proceedings Vol. 452, Pittsburgh: Materials Research Society, pp. 613–618.

Rehm, J. M., McLendon, G. L., Tsyebskov, L., Fauchet, P. M. (1995), *Appl. Phys. Lett.* **66**, 3669–3671.

Rehm, J. M., McLendon, G. L., Fauchet, P. M. (1996), *J. Am. Chem. Soc.* **118**, 4490–4491.

Ren, S. Y., Dow, J. D. (1992), *Phys. Rev. B* **45**, 6492.

Richter, A., Steiner, P., Kozlowski, F., Lang, W. (1991), *IEEE Electron Dev. Lett.* **12**, 691.

Sailor, M. J., Lee, E. J. (1997), *Adv. Mater.* **9**, 783–793.

Smith, R. L., Collins, S. D. (1992), *J. Appl. Phys.* **71**, R1.

Tamura, H., Ruckschloss, M., Wirschem, T., Veprek, S. (1994), *Appl. Phys. Lett.* **65**, 1537.

Tsao, S. S. (1987), *IEEE Circuit Dev. Mag.* **3**, 8.

Tsybeskov, L., Vandyshev, J. V., Fauchet, P. M. (1993), *Phys. Rev. B* **49**, 7821–7824.

Tsybeskov, L., Moore, K. L., Hall, D. G., Fauchet, P. M. (1996a), *Phys. Rev. B* **54**, R8361–R8364.

Tsybeskov, L., Duttagupta, S. P., Hirschman, K. D., Fauchet, P. M. (1996b), *Appl. Phys. Lett.* **68**, 2058–2060.

Tsybeskov, L., Hirschman, K. D., Duttagupta, S. P., Fauchet, P. M., Zacharias, M., McCaffrey, J. P., Lockwood, D. J. (1998), *Phys. Status Sol. (a)* **165**, 69–77.

Uhlir, Jr., A. (1956), *Bell Syst. Tech. J.* **35**, 333.

Vervoort, L., Bassani, F., Mihalcescu, I., Vial, J.-C., Arnaud d'Avitaya, F. (1995), *Phys. Status Sol. (b)* **190**, 123.

Vial, J.-C., Bsiesy, A., Fishman, G., Gaspard, F., Herino, R., Ligeon, M., Muller, F., Romestain, R., MacFarlane, R. M. (1993), in: P. M. Fauchet, C. C. Tsai, L. T. Canham, I. Shimizu, T. Aoyagi

(Eds.), *Microcrystalline Semiconductors: Materials Science & Devices*, MRS Symposium Proceedings Vol. 283, Pittsburgh: Materials Research Society, p. 241.

van Buuren, T., Tiedje, T., Dahn, J. R., May, B. M. (1993), *Appl. Phys. Lett.* **63**, 2911–2913.

von Behren, J., Chimowitz, E. H., Fauchet, P. M. (1997), *Adv. Mater.* **9**, 921–926.

von Behren, J., van Buuren, T., Zacharias, M., Chimowitz, E. H., Fauchet, P. M. (1998), *Solid State Commun.*, **105**, 317–32.

Wang, L. W., Zunger, A. (1994), *J. Phys. Chem.* **98**, 2158.

Wang, L.-W., Zunger, A. (1996), in: P. V. Kamat, D. Meisel (Eds.), *Nanocrystalline Semiconductor Materials*, Amsterdam: Elsevier Science.

Zaidi, S. H., Chu, A.-S., Brueck, S. R. J. (1995), in: R. W. Collins, C. C. Tsai, M. Hirose, F. Koch, L. Brus (Eds.), *Microcrystalline and Nanocrystalline Semiconductors*, MRS Symposium Proceedings Vol. 358, Pittsburgh: Materials Research Society, p. 957.

Zheng, J. P., Jiao, K. L., Shen, W. P., Anderson, W. A., Kwok, H. S. (1992), *Appl. Phys. Lett.* **61**, 459–461.

Further Reading

Canham, L. T. (1997), *Properties of Porous Silicon*, London: The Institution of Electrical Engineers.

Collins, R. W., Tsai, C. C., Hirose, M., Koch, F., Brus, L. (1995), *Microcrystalline and Nanocrystalline Semiconductors*, MRS Symposium Proceedings Vol. 358, Pittsburgh: Materials Research Society.

Collins, R. W., Fauchet, P. M., Shimizu, I., Vial, J.-C., Shimada, T., Alivisatos, A. P. (1997), *Advances in Microcrystalline and Nanocrystalline Semiconductors—1996*, MRS Symposium Proceedings Vol. 452, Pittsburgh: Materials Research Society.

Fauchet, P. M., Tsai, C. C., Canham, L. T., Shimizu, I., Aoyagi, Y. (1993), *Microcrystalline Semiconductors: Materials Science & Devices*, MRS Symposium Proceedings Vol. 283, Pittsburgh: Materials Research Society.

Lockwood, D. J. (1998), *Light Emission in Silicon: From Physics to Devices*, San Diego: Academic Press.

Lockwood, D. J., Fauchet, P. M., Koshida, N., Brueck, S. R. J. (1996), *Advanced Luminescent Materials*, Pennington, NJ: The Electrochemical Society.

Schmuki, P., Lockwood, D. J., Isaacs, H., Bsiesy, A. (1997), *Pits and Pores: Formation, Properties, and Significance for Advanced Luminescent Materials*, Pennington, NJ: The Electrochemical Society.

Tsu, R., Feng, Z. C. (1994), *Porous Silicon*, Singapore: World Scientific.

Vial, J.-C., Derrien, J. (1995), *Porous Silicon Science and Technology*, Berlin: Springer-Verlag.

SUN, STRUCTURE OF

Marc H. Pinsonneault, *Department of Astronomy, The Ohio State University, Columbus, Ohio, U.S.A.*

Introduction 273

1. **Global and Surface Properties** 274
1.1 Mass, Radius, Luminosity, Surface Temperature, and Age 274
1.2 Composition—Photospheric and Meteoritic 274
1.3 Surface Rotation and Oblateness 275
1.4 The Solar Sunspot Cycle and Magnetic Fields 276
1.5 Solar Chromosphere, Corona, and Wind 276
2. **Theory of Stellar Structure and Evolution** 276
2.1 Basic Equations and Assumptions 276
2.2 Input Physics 277
2.2.1 Nuclear-Reaction Roles 277
2.2.2 Energy Transport 278
2.2.3 Equation of State 279
2.3 Constructing a Calibrated Solar Model 279
2.4 Physical Processes Not in the Standard Solar Model 279
2.4.1 Rotation 279
2.4.2 Microscopic Diffusion 280
3. **Diagnostics of the Internal Properties of the Sun** 280
3.1 Solar-Neutrino Fluxes 280
3.2 Solar Oscillations 281
3.2.1 Oscillations as a Diagnostic of the Thermal Structure 281
3.2.2 Internal Rotation 282
4. **Internal Structure of the Sun** 282
4.1 Mass and Luminosity Distributions 282
4.2 Thermal Structure 282
4.3 Abundance Profiles 283
4.4 Convection Zone 285
4.4.1 Superadiabatic Layer 286
4.4.2 Surface versus Initial Abundances 286
4.4.3 Rotation as a Function of Depth and Latitude 287
4.5 Radiative Mantle 288
4.5.1 Evidence for Microscopic Diffusion and Mixing 288
4.5.2 Rotation as a Function of Depth 289
4.6 Nuclear Burning Core 290
4.6.1 Thermal Structure 290
4.6.2 Abundances and Energy Generation as Functions of Depth 291
4.6.3 Predicted Neutrino Fluxes 291
4.6.3.1 Kamiokande 291
4.6.3.2 Homestake 291
4.6.3.3 GALLEX/SAGE 291
5. **Summary** 291

Glossary 292

Works Cited 292

Further Reading 292

INTRODUCTION

The Sun is the foundation for the study of stars. Its surface and global properties can be determined to much higher precision than for other stars; traditionally these have been combined with theoretical models to deduce the internal properties of the Sun. The study of the Sun has been revolutionized, however, by the availability of diagnostics of its internal properties. Neutrino experiments provide direct evidence for nuclear reactions in the core of the Sun; the frequencies of acoustic waves in the Sun provide a means of determining the speed of sound and the rotation rate as functions of depth. In addition, the relative abundances of elements in meteorites can be measured precisely. Meteoritic data can provide information about the composition of the material that the Sun was formed from, and a comparison of the current surface abundances with the initial abundances can be used to test for changes in the surface abundances during the lifetime of the Sun. This wealth of data has

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

Table 1. Global properties of the sun.

Mass	1.9891×10^{33} g
Radius	6.9598×10^{10} cm
Luminosity	$3.844(1 \pm 0.004) \times 10^{33}$ erg s^{-1}
Age	4.57 ± 0.02 Ga
T_{eff}	5770 K

provided astronomers with a stringent test of the standard models of the structure of stars, as well as evidence that allows an assessment of the importance of physical processes usually neglected in theoretical stellar models.

Section 1 of this article summarizes the global and surface properties of the Sun. Section 2 introduces the main components of the theoretical models. In Section 3, the diagnostics of the internal properties of the Sun are discussed. The inferred internal structure of the Sun is discussed in Sec. 4. The results are summarized in Sec. 5.

1. GLOBAL AND SURFACE PROPERTIES

1.1 Mass, Radius, Luminosity, Surface Temperature, and Age

Table 1 summarizes some of the important global properties of the Sun (Lang, 1991). Note that these properties, especially age, are determined far more accurately than for stars in general. For example, the solar mass is inferred from the orbital periods of the planets, and the primary uncertainty is the value of the gravitational constant *G*. It is extremely difficult to infer directly the age of a single star such as the Sun. However, the time scale for the formation of the Sun, planets, and smaller bodies of the solar system is short compared with the age of the Sun. The age of the Sun is therefore most accurately determined by radioactive age dating of meteorites, and the primary uncertainty is the possibility of a small difference in age between the meteorites and the Sun (see Bahcall *et al.*, 1995; this article also discusses updated values for the solar luminosity and other properties of the Sun).

In general the radius and luminosity (total energy output per second) of the Sun will change over both long and short time scales. The effective temperature (T_{eff}) of the Sun is the surface temperature that a blackbody emitting the same flux per unit area would have. Radius (R), luminosity (L), and the effective surface temperature are related by

$$L = 4\pi R^2 \sigma {T_{\mathrm{eff}}}^4, \tag{1}$$

where σ is the Stefan–Boltzmann constant. In Table 2 we show the time dependence of the solar radius and luminosity inferred from theoretical models (see Sec. 2). The long-term changes in the solar energy output are of interest for understanding the time evolution of the climate of the Earth. Shorter-term changes in brightness and radius can arise from phenomena such as the solar sunspot cycle (Sec. 1.4; see also SOLAR RADIATION).

Table 2. Changes in global solar properties as functions of time.

Time (10^9 a)	Radius ($R_\odot$)	Luminosity ($L_\odot$)	Convection-zone depth ($M_\odot$)
0.0	0.868	0.673	0.0329
0.04	0.874	0.698	0.0310
0.27	0.883	0.724	0.0300
0.5	0.889	0.736	0.0291
1.0	0.900	0.763	0.0282
1.5	0.912	0.792	0.0282
2.5	0.937	0.850	0.0268
3.5	0.965	0.916	0.0252
4.57	1.000	1.000	0.0240

1.2 Composition—Photospheric and Meteoritic

The composition of a star exerts a strong influence on its structure and evolution. The Sun is primarily composed of hydrogen (≈ 0.70 by mass fraction), with most of the rest of the material in the form of helium (≈ 0.28) and only about 0.02 in the form of heavier elements. Meteorites provide the most accurate means of measuring the relative abundances of heavy elements in the protosolar nebula. They do not, however, provide a means of determining the mass fractions of abundant volatile species, such as H, He, C, N, O, and Ne. In the solar photosphere we can measure the current surface mass fractions of different species, with the notable exception of ^{4}He, relative to hy-

Table 3. Abundances in the solar photosphere ($\log N_{\mathrm{H}} = 12.0$).

Element	Photosphere	Meteorites
H	12.0	—
He	[10.99 ± 0.035]	—
Li	1.16 ± 0.1	3.31 ± 0.04
Be	1.15 ± 0.1	1.42 ± 0.04
B	2.6 ± 0.3	2.88 ± 0.04
C	8.55 ± 0.05	—
N	7.97 ± 0.07	—
O	8.87 ± 0.06	—
Ne	[8.08 ± 0.06]	—
Na	6.33 ± 0.03	6.31 ± 0.03
Mg	7.58 ± 0.05	7.58 ± 0.02
Al	6.47 ± 0.07	6.48 ± 0.02
Si	7.55 ± 0.05	7.55 ± 0.02
P	5.45 ± 0.04	5.57 ± 0.04
S	7.21 ± 0.06	7.27 ± 0.05
Cl	5.5 ± 0.3	5.27 ± 0.06
Ar	[6.52 ± 0.10]	—
K	5.12 ± 0.13	5.13 ± 0.03
Ca	6.36 ± 0.02	6.34 ± 0.03
Ti	5.02 ± 0.06	4.93 ± 0.02
Cr	5.67 ± 0.03	5.68 ± 0.03
Mn	5.39 ± 0.03	5.53 ± 0.04
Fe	7.50 ± 0.04	7.51 ± 0.01
Ni	6.25 ± 0.04	6.25 ± 0.02

drogen. These measurements, however, are not as accurate as those in the meteorites.

To infer the initial composition of the Sun, data from the two sources are combined. Meteorites are used to set the relative initial abundances of the species when possible, on a logarithmic scale relative to Si, where log(Si) = 6. The photospheric abundances are measured relative to hydrogen, on a logarithmic scale where log(H) = 12. The meteoritic scale is converted into the photospheric scale by measuring the ratios of the abundances on the two different scales for well-measured species (see Anders and Grevesse, 1989). The mass fractions of abundant species in the Sun from Anders and Grevesse (1989) are given in Table 3.

The initial ^{4}He abundance of the Sun is determined indirectly. In astronomy, the mass fractions of hydrogen, helium, and all other elements combined (referred to as "metals") are denoted by X, Y, and Z, respectively, and satisfy $X + Y + Z = 1$. The surface abundances give the ratio Z/X. Because the luminosity of a star depends on its initial composition, the requirement that a solar model has the luminosity of the Sun at the age of the Sun can be used to infer the initial ^{4}He mass fraction (Y). In addition, helioseismic data can be used to infer the current surface Y (Sec. 3.2).

The light elements Li and Be are significantly more abundant in meteorites than in the solar photosphere (by factors of 100 and 2, respectively). This constitutes evidence for mild mixing in the solar envelope, as these species burn at temperatures of 2.5 and 3.5 MK, respectively (Sec. 2.4.1). The current surface Y is also lower than the initial Y, which is evidence for the gravitational settling of helium out of the surface layers of the Sun (Sec. 2.4.2). Both theoretical models and the sound speed inferred from helioseismology require the burning of helium in the core of the Sun (Sec. 2, Sec. 3.2).

1.3 Surface Rotation and Oblateness

The Sun rotates at a rate that depends on the latitude at the surface; Libbrecht and Morrow (1991) give the surface rotation frequency as a function of latitude as

$$\nu(R_\odot, \phi) = 462 - 75 \sin^2\phi - 50 \sin^4\phi \ \mathrm{nHz}, \tag{2}$$

which implies an equatorial rotation period of about 25.4 d and a polar period of order 34 d. Helioseismic data can be used to infer the internal rotation as a function of depth (Sec. 3.2.2).

The measured solar oblateness is

$$\varepsilon = 8.63 \pm 0.88 \times 10^{-6} \tag{3}$$

where $\varepsilon = 2(R_{\mathrm{eq}} - R_{\mathrm{pole}})/(R_{\mathrm{eq}} + R_{\mathrm{pole}})$. This provides not only a measurement of the local distortion of the surface of the Sun, but an indication of the quadrupole moment and a constraint on the internal solar rotation. The distortion from spherical symmetry induces small corrections to the equations of stellar structure. We have evidence that the Sun is experiencing angular-momentum loss (Sec. 1.5), and young solar analogs rotate more rapidly than the present-day Sun. This distortion could therefore have been larger in the past. Rotation can also induce circulation currents and mixing, which could have a more significant impact on the structure of the Sun.

1.4 The Solar Sunspot Cycle and Magnetic Fields

The existence of sunspots has been known for centuries; they are regions with high magnetic field strengths of 2–3 kG and cover a fraction of the solar surface ranging up to about 0.001. Sunspots appear dark relative to the rest of the surface because they are significantly cooler than the rest of the photosphere; they are also surrounded by active regions that are hotter than the mean surface temperature (and therefore brighter). We have extensive records that indicate that the number of such spots varies with a period of order 22 years (including a reversal in polarity; the number of sunspots varies over an 11-year cycle). The flux per unit area is different in and around sunspots than for the rest of the solar surface; as a result, these changes in spot coverage can cause changes in the energy output of the Sun. The amplitude of the modulation of the total solar irradiance during the sunspot cycle is 0.08%. In some spectral regions the variation in the amplitude is higher (see SOLAR RADIATION). For an extensive discussion of the solar activity cycle, see the chapters in Sec. IV of Cox *et al.* (1991).

The sunspot cycle is intimately linked with the generation of the solar magnetic field. The dynamo theory postulates that differential rotation in the solar interior leads to an amplification of the internal magnetic field strength. At some point the fields become strong enough to be subject to magnetic instabilities, and magnetic flux can then rise to the surface. In the outer layers ($R > 0.713 R_\odot$) of the Sun, energy is transported by mass motions (convection), with an overturn time scale of the order of a month. The most likely site of the solar dynamo is therefore below the surface convection zone, where the magnetic field can grow over a longer time scale than is permitted in the convection zone.

1.5 Solar Chromosphere, Corona, and Wind

Above the solar photosphere there are three different regions. The Sun has a chromosphere hotter than the photosphere, a very hot and extended corona, and a steady stream of material escaping from the Sun referred to as the solar wind. The measured solar mass loss rate is $3 \times 10^{-14} M_\odot/\mathrm{a}$, which implies that only a small fraction of the Sun's mass has been lost over the 4.5 Ga of its lifetime.

The heating mechanism for the chromosphere and corona has been a subject of some controversy. Turbulent motions in the solar atmosphere can generate acoustic waves, which increase in velocity as the density decreases outwards until they become shock waves. Alfvén waves may also contribute to the heating of the chromosphere. The temperature in the chromosphere rises from the minimum at the top of the photosphere of 4500 K to a maximum temperature of 20 000 K. There is then a transition region above the chromosphere where the temperature rises to of order 10^6 K in the corona. The properties of the chromosphere and corona change significantly during the course of the solar sunspot cycle.

2. THEORY OF STELLAR STRUCTURE AND EVOLUTION

Much of our understanding of the interior of the Sun comes from theoretical models of its structure constrained by the surface and global properties outlined in Sec. 1. In this section the ingredients of a theoretical stellar model are described. In Sec. 4 the internal structure of a theoretical solar model is described and compared with the diagnostics of the internal properties of the Sun described in Sec. 3. We begin by outlining the equations of stellar structure and then comment on the input physics, method for constructing a solar model, and some additional physical processes not usually included in standard solar models.

2.1 Basic Equations and Assumptions

The basic equations of stellar structure are expressions of the conservation of mass, energy, and momentum, along with a prescription for the transport of energy. The continuity equation gives

$$\frac{\mathrm{d}M}{\mathrm{d}r} = 4\pi\rho r^2 \tag{4}$$

where $M(r)$ is the mass, r is the radius, and

$\rho(r)$ is the density. Hydrostatic equilibrium can be expressed as

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\rho g \tag{5}$$

where $P(r)$ is the pressure and $g = GM/r^2$ neglecting departures from spherical symmetry. The high mass of the Sun implies that a very large internal pressure gradient is required to balance the force of gravity. Energy generation requires

$$\frac{\mathrm{d}L}{\mathrm{d}r} = 4\pi\rho r^2 \varepsilon, \tag{6}$$

where $L(r)$ is the luminosity ($\mathrm{erg\,s^{-1}}$) and $\varepsilon(r)$ is the energy generation rate (which includes nuclear energy generation, energy loss from neutrino emission, and changes in gravitational potential energy).

The temperature gradient is the minimum of that for energy transport by radiation and that for energy transport by mass motions (convection); for transport by radiation,

$$\frac{\mathrm{d}T}{\mathrm{d}r} = -\frac{3}{16\pi ac}\frac{\kappa L}{r^2 T^3}, \tag{7}$$

where $T(r)$ is the temperature, c is the speed of light, a is the radiation constant, and $\kappa(T, \rho, \text{composition})$ is the opacity of matter to radiation.

These equations must be supplemented by an equation of state that relates the pressure, temperature, and density; a convection theory that determines the temperature gradient in regions where energy is carried by mass motions; and a prescription for changes in abundance induced by nuclear reactions as a function of time. The standard model assumes spherical symmetry, neglects magnetic fields and mass loss, and does not include either mixing processes linked with rotation or gravitational settling of heavy elements with respect to lighter ones.

2.2 Input Physics

In Fig. 1 the energy generation rate ε and the opacity κ are shown as functions of fractional radius in a model of the present-day Sun. Energy generation is primarily through nuclear reactions in the Sun and is highly temperature sensitive. The final luminosity of the Sun, however, is set by the balance between energy generation and energy transport. In the inner part of the Sun, the opacity of matter to radiation is low and energy transport is primarily by radiation. When the opacity is too high material becomes convectively unstable; the strong increase in opacity in the outer layers of the Sun indicates that energy transport is primarily through convection above $0.713R_\odot$. The nuclear-reaction rates, equation of state and energy transport in the solar interior are reviewed below.

2.2.1 Nuclear-Reaction Rates The conversion of hydrogen into helium is the primary energy source for the Sun [see Bahcall (1989) for a discussion of nuclear-reaction rates relevant for the Sun; cross sections are summarized in Bahcall and Pinsonneault (1992) and Bahcall *et al.* (1995)]. In most cases the cross sections for the relevant reactions can be measured experimentally; because the typical energies in the Sun ($\approx 1\,\mathrm{keV}$ at the center) are much lower than the regime which is accessible experimentally ($> 100\,\mathrm{keV}$), the experimental results are extrapolated to the low-temperature regime. There are two principal mechanisms for hydrogen burning. In the proton–proton (pp) chain,

$$^1\mathrm{H} + {}^1\mathrm{H} \rightarrow {}^2\mathrm{H}, \quad {}^1\mathrm{H} + {}^2\mathrm{H} \rightarrow {}^3\mathrm{He}, \tag{8}$$

and either

$$^3\mathrm{He} + {}^3\mathrm{He} \rightarrow {}^4\mathrm{He} + 2\,{}^1\mathrm{H} \text{ (85\% of the time)} \tag{9}$$

or

$$^3\mathrm{He} + {}^4\mathrm{He} \rightarrow {}^7\mathrm{Be} \text{ (15\% of the time).} \tag{10}$$

The ^{7}Be then either experiences an electron capture or a proton capture, and is ultimately processed into two ^{4}He nuclei. Some of these reactions produce neutrinos that can be detected experimentally. This chain accounts for 98% of the energy output of the Sun.

In the carbon–nitrogen–oxygen (CNO) cycle the C, N, and O nuclei are used as cata-

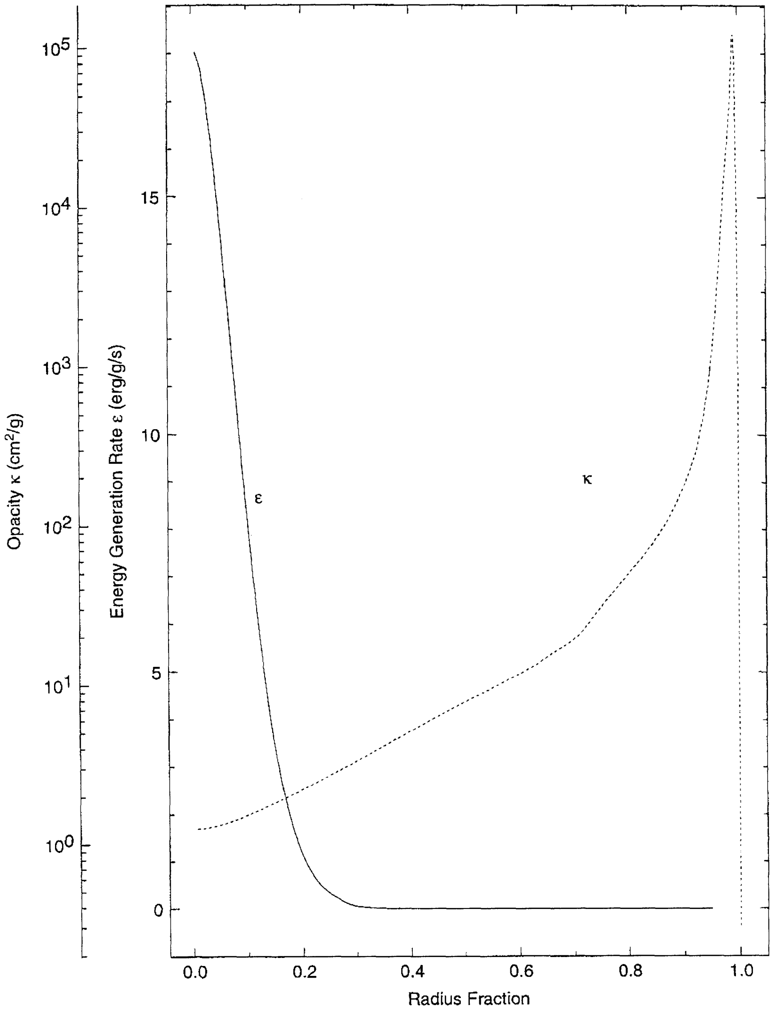

FIG. 1. The energy generation rate (solid line, first linear scale) and opacity (dashed line, second logarithmic scale) as functions of the fractional radius in a theoretical solar model.

lysts, and hydrogen is converted into helium through a series of proton captures. These reactions are more temperature sensitive than the pp chain and predominate in stars more massive than the Sun (accounting for only 2% of the solar luminosity). They produce some neutrinos that can be detected experimentally. The Sun is not hot enough to burn He and heavier elements.

2.2.2 Energy Transport The mean free path of a photon in the solar interior is of the order of 1 cm. Because the Sun is much hotter at the center ($\approx 15.6\,\mathrm{MK}$) than it is at the surface ($\approx 5770\,\mathrm{K}$), the photons produced at a given radius will on average be more energetic than those produced at a slightly larger radius; when these photons interact with matter there will be a net transport of energy outwards. As a result of the small length scale, the transport of energy by radiation can be treated as a diffusion process. Theoretical models rely on quantum mechanical calculations of the opacity of matter to radiation, such as the recent work of Iglesias and Rogers (1996), to compute the temperature gradient needed to carry the solar flux via radiation. If we denote the actual temperature gradient by $d \ln T / d \ln P =$

∇, then convection takes over as the dominant mode of energy transport whenever the radiative temperature gradient ∇_{rad} is greater then the adiabatic temperature gradient ∇_{ad}. The mode of energy transport switches from radiative to convective near $0.713 R_{\odot}$; there is a discontinuity in the speed of sound as a function of depth at this point that can be detected through the study of solar oscillations (Sec. 3.2). The temperature gradient in the convective region of the Sun is determined using mixing-length theory. In the deeper layers of the solar convection zone, the actual temperature gradient is close to the adiabatic gradient; it can depart significantly from the adiabatic gradient in the outermost layers of the Sun.

2.2.3 Equation of State The bulk of the mass of the Sun is in the form of hydrogen and helium, which ionize at low temperatures; most of the heavier elements are in the form of C, N, and O, which are fully ionized at modest temperatures of order 1 MK. As a result, the interior of the Sun is essentially fully ionized and is also close to an ideal gas. An important quantity for the solar equation of state is the mean molecular mass μ, defined by

$$P = \rho \Re T / \mu \tag{11}$$

where $\Re$ is the gas constant.

The mean molecular mass in the solar interior changes in regions of partial ionization. For the purposes of computing the pressure, free electrons are as important as ions; for example, there will be twice as many particles per gram in an ionized hydrogen gas as in a neutral hydrogen gas. There are more particles per gram in fully ionized hydrogen (2) than there are in fully ionized helium (3/4) or heavier elements ($\approx 1/2$), and so composition changes can affect the equation of state; nuclear burning and gravitational settling in the solar interior will therefore also cause changes in μ as a function of radius and time. Modern equations of state include relativistic effects, electron degeneracy, radiation pressure, and electrostatic corrections. In the outer layers, the partial ionization of H, He, and metals needs to be computed. Partial ionization is important for the thermal structure in the outer layers because of the strong dependence of the number of free particles per gram on the ionization state of the matter. Other thermodynamic properties, such as the adiabatic temperature gradient (see Sec. 2.2.2), are also strongly affected by partial ionization. Most modern solar models use tabulated equations of state that include the above effects; examples include the OPAL equation of state (Rogers *et al.*, 1996).

2.3 Constructing a Calibrated Solar Model

Stellar evolution is an initial-value problem. A model with the mass of the Sun is evolved to the age of the Sun. A calibrated model reproduces the solar surface Z/X, radius, and luminosity at the age of the Sun. The initial Y of the Sun is adjusted to reproduce the solar luminosity, while the mixing length α (the distance in pressure scale heights that a fluid element travels before releasing its heat) in the convective portion of the Sun is adjusted to reproduce the solar radius. The inferred initial Y depends on the other input physics, but is typically in the range 0.27–0.29 by mass fraction.

2.4 Physical Processes not in the Standard Solar Model

Standard solar models do not incorporate rotation and element-separation processes. However, there is evidence that these phenomena have a measurable impact on the observed properties of the Sun; solar and stellar models that include rotation and element separation have been constructed.

2.4.1 Rotation Rotation has a direct effect on the structure of the Sun; in addition, it can induce mixing from meridional circulation and hydrodynamic instabilities triggered by differential rotation with depth. A full treatment of rotation in the Sun includes the effect of rotation on the structure of the Sun, the initial angular momentum in the protostellar phase, angular-momentum loss via the solar wind, and internal angular-momentum transport. Solar models with rotation can explain the surface light-element abundances of the Sun, but rotation has only a modest impact on the thermal structure of the solar models (Richard *et al.*, 1996). In other areas of stellar evolution rotational mixing can produce dramatic observational effects.

2.4.2 Microscopic Diffusion Heavy elements sink with respect to lighter ones in the presence of gravity; in addition there is a net downward flux of heavier elements in the presence of a temperature gradient. These diffusion effects are included in recent solar models; they result in a lowering of the surface abundances of elements heavier than hydrogen, a deepening of the surface convection zones in the models, and an increase in the inferred central temperature and the expected solar neutrino fluxes.

3. DIAGNOSTICS OF THE INTERNAL PROPERTIES OF THE SUN

The global and surface properties of the Sun can be used to infer its internal properties in conjunction with the theoretical models described in Sec. 2. However, there are also diagnostics of the internal properties of the Sun. Neutrinos are produced in nuclear reactions in the solar core; because of their weak interaction with matter, they can pass through the Sun and be detected on the Earth. Acoustic waves in the Sun also provide information about the speed of sound and rotation as a function of depth.

3.1 Solar-Neutrino Fluxes

Some of the nuclear reactions that occur in the Sun emit neutrinos. Because the cross section of neutrinos in matter is extremely low, these neutrinos escape from the Sun and can be detected in terrestrial experiments. Different reactions emit neutrinos with different energies, and there are now four experiments, which are sensitive to neutrinos with different energies. The resulting neutrino spectrum can be compared with that predicted by theoretical models. The detection of solar neutrinos constitutes evidence that nuclear reactions from hydrogen burning are the predominant solar energy source.

However, the observed fluxes are significantly less than the predicted rates. This persistent discrepancy between theory and experiment is the *solar neutrino problem*, and its solution involves either significant modifications of the theoretical solar models or new neutrino physics. In Sec. 4.6, the observed and predicted fluxes are compared. Here we briefly review the processes that produce solar neutrinos. Bahcall (1989) contains a detailed description of the different solar neutrino experiments.

The different neutrino experiments currently under way include the following:

1. The Homestake experiment is a radiochemical experiment that observes neutrinos via the reaction

$$\nu_e + \mathrm{Cl}^{37} \rightarrow \mathrm{e}^- + \mathrm{Ar}^{37}. \tag{12}$$

   The radioactive Ar^{37} is extracted from a C_2Cl_4 fluid. The threshold energy is 0.8 MeV.
2. The Kamiokande experiment is based on neutrino–electron scattering in water. The direction of the neutrinos can be inferred in this experiment, but it has a very high threshold neutrino energy of 7.5 MeV. A scaled-up version of the Kamiokande experiment (Super-Kamiokande) is currently in progress; it will permit far better statistics and a detailed analysis of the neutrino energy spectrum.
3. The GALLEX and SAGE experiments detect the absorption of neutrinos in gallium via the reaction

$$\nu_e + \mathrm{Ga}^{71} \rightarrow \mathrm{e}^- + \mathrm{Ge}^{71}. \tag{13}$$

   The threshold energy for this process is low (0.23 MeV), making it possible to detect the majority of the neutrinos produced in the Sun. These are radiochemical experiments, similar to the Homestake experiment.
4. The SNO experiment, currently underway, uses heavy water to observe Cherenkov light from electrons interacting with deuterium in addition to the neutrino–electron scattering to which the Kamiokande experiment is sensitive. The charged-current reaction

$$\nu_e + \mathrm{H}^2 \rightarrow \mathrm{e}^- + \mathrm{p} + \mathrm{p} \tag{14}$$

   is sensitive to the electron neutrinos produced by the Sun, while the neutral-current reaction

$$\nu + H^2 \rightarrow \nu + p + n \qquad (15)$$

is sensitive to other neutrino flavors. This could provide an important test of some theories of new neutrino physics, which explain the solar neutrino deficit by postulating that some of the electron neutrinos produced in the Sun are converted into other flavors as they travel from the center of the Sun to the Earth.

The major processes that produce neutrinos in the Sun are as these:

1. The pp reaction produces a low-energy neutrino. The GALLEX and SAGE experiments are sensitive to this neutrino, as well as those from the other reactions listed below.
2. The pp reaction can be mediated with an electron (the pep reaction). All of the experiments except Kamiokande can detect pep neutrinos.
3. After ^{3}He is produced, there are two branches of the pp chain. ^{3}He + ^{3}He occurs 85% of the time and does not produce a neutrino; ^{3}He + ^{4}He occurs 15% of the time and has two possible terminations, each of which produces a neutrino (^{7}Be and ^{8}B). The ^{8}B neutrino is highly energetic and can be detected in all of the experiments; Kamiokande and SNO are sensitive only to the ^{8}B neutrinos. The other experiments can also detect some of the lower-energy ^{7}Be neutrinos.
4. The Sun can also process hydrogen into helium by using the C, N, and O nuclei as catalysts; these reactions are responsible for about 2% of the Sun's luminosity and produce neutrinos that can be detected by GALLEX, SAGE, and the Homestake experiment.

The cross section for neutrino capture is extremely small; for the Homestake experiment the measured capture rate is $2.55 \pm 0.17(\text{stat}) \pm 0.18(\text{syst})$ solar neutrino units (SNUs), which are units of 10^{-36} events per target atom per second. Detecting such rare events is technically challenging (see Bahcall, 1989, for a discussion of the experimental techniques). By comparison, the theoretical fluxes are of the order of 8–9 SNUs (Sec. 4.3.3). Different experiments are sensitive to neutrinos of different energies and the results of different experiments can therefore be combined to form a neutrino energy spectrum of the Sun. One of the most interesting developments in the solar neutrino problem has been the apparent strong dependence of the discrepancy between theory and observation on neutrino energy; this feature is difficult to explain by changing the theoretical fluxes alone. The comparison between theory and observation is discussed in Sec. 4.6.

3.2 Solar Oscillations

Solar oscillations can be observed through either intensity fluctuations or Doppler shifts on the solar surface as a function of position and time. The amplitudes are at the centimeter per second level, with characteristic periods of the order of 5 min. These fluctuations arise because internal waves in the Sun can be refracted by the strong density gradient in the outer layers of the Sun. Therefore, the Sun acts as an acoustic cavity, and there will be constructive interference for certain characteristic frequencies. The frequencies ($\approx 10^7$ modes of oscillation) with which it oscillates can be used to deduce a variety of internal properties. The restoring force for the solar oscillations detected to date is pressure, and so they are sometimes referred to as p-mode oscillations. These modes have their maximum amplitude in the outer layers, making them an ideal diagnostic of the outer layers of the Sun. Waves where the restoring force is gravity (g modes) have not yet been unambiguously identified; because they have their maximum amplitude in the core, a detection of solar g modes would provide important tests of the conditions at the center of the Sun. An extensive discussion of the observational and theoretical studies of the solar oscillations, referred to as helioseismology, can be found in Part II of Cox *et al.* (1991).

3.2.1 Oscillations as a Diagnostic of the Thermal Structure Expressed in terms of spherical harmonics (n, l, m), the depth to which a given p mode will penetrate depends strongly on l: lower-l modes can penetrate further into the Sun. Information from these modes can be combined to form a map of the speed of sound as a function of depth in the solar interior. The errors are largest in the deep interior because there are few p modes that penetrate very deeply, and the frequencies of these low-l modes have proven difficult to measure. However, high-quality data for the low-l p modes (and therefore good information on the physical conditions in the deep interior of the Sun) have recently become available (Tomczyk *et al.*, 1995). The sound speed inferred in the solar interior is compared with theoretical predictions in Sec. 4.

3.2.2 Internal Rotation In the presence of rotation, there is a splitting of the p-mode frequencies that can be used to infer the internal solar rotation as a function of latitude and depth. The rotation in the deep solar interior is currently uncertain, but there is good information on the rotational properties outside the core of the Sun.

4. INTERNAL STRUCTURE OF THE SUN

The internal properties of the Sun that are described in this section are inferred from a theoretical solar model (Sec. 2) that satisfies the global properties of the Sun as described in Sec. 1. Throughout this section theory will be compared with data on the internal properties of the Sun (Sec. 3); current theoretical models are in good agreement with solar oscillation data but conflict with solar-neutrino data. We begin with an overview of the major components of the internal structure of the Sun in Secs. 4.1–4.3; a detailed discussion of the different layers of the solar interior is then presented in Secs. 4.4–4.6.

4.1 Mass and Luminosity Distributions

The mass fraction M and luminosity L of the Sun as functions of radius are shown in Fig. 2. The Sun is highly centrally concentrated; 60% of its mass is contained within $0.3R_\odot$, 90% of the mass is contained within $0.5R_\odot$, and 98% of the mass is contained within $0.7R_\odot$. The central density of $156\,\mathrm{g\,cm^{-3}}$ is far higher than the mean density of $1.4\,\mathrm{g\,cm^{-3}}$. The energy generation of the Sun is concentrated in the deep interior; 46% of the luminosity is generated within $0.1R_\odot$, 95% of the energy is generated within $0.2R_\odot$, and effectively all of the energy (with the exception of energy from gravitational contraction) is generated within $0.4R_\odot$. Significant conversion of hydrogen into helium is also concentrated in the inner $0.2R_\odot$ (see Sec. 4.3 below). In the sections that follow, this will be referred to as the solar core.

4.2 Thermal Structure

The pressure, temperature, and density of the Sun as functions of radius are shown in Fig. 3. Note that the temperature declines much less rapidly than the density, which in turn declines much less rapidly than the pressure. The speed of sound as a function of depth inferred from seismology is shown in Fig. 4(a); the fractional difference between the sound speed of the Sun and that of the theoretical model discussed in this article is shown in Fig. 4(b). Theory agrees with observation to an accuracy of 0.2% throughout the solar interior. The deviations from the predictions of the model are interesting and will be discussed below.

The actual, adiabatic, and radiative temperature gradients are compared in Fig. 5. Energy is carried by radiation in the interior; in the outer layers energy is carried by convection (conduction is not an important mode of energy transport within the Sun). The time scale for convection is short compared with the solar age, which implies that the outer layers of the Sun are fully mixed; the interior can experience composition gradients due to nuclear burning and gravitational settling, and these gradients can be reduced by mixing. We can therefore divide the outer layers of the Sun into a surface convection zone and a radiative mantle above the nuclear burning core.

There is a change in the derivative of the speed of sound as a function of depth at the base of the surface convection zone. The base of the convection zone inferred from helioseismology is $0.713 \pm 0.003R_\odot$. Theoretical models that include the latest radiative opac-

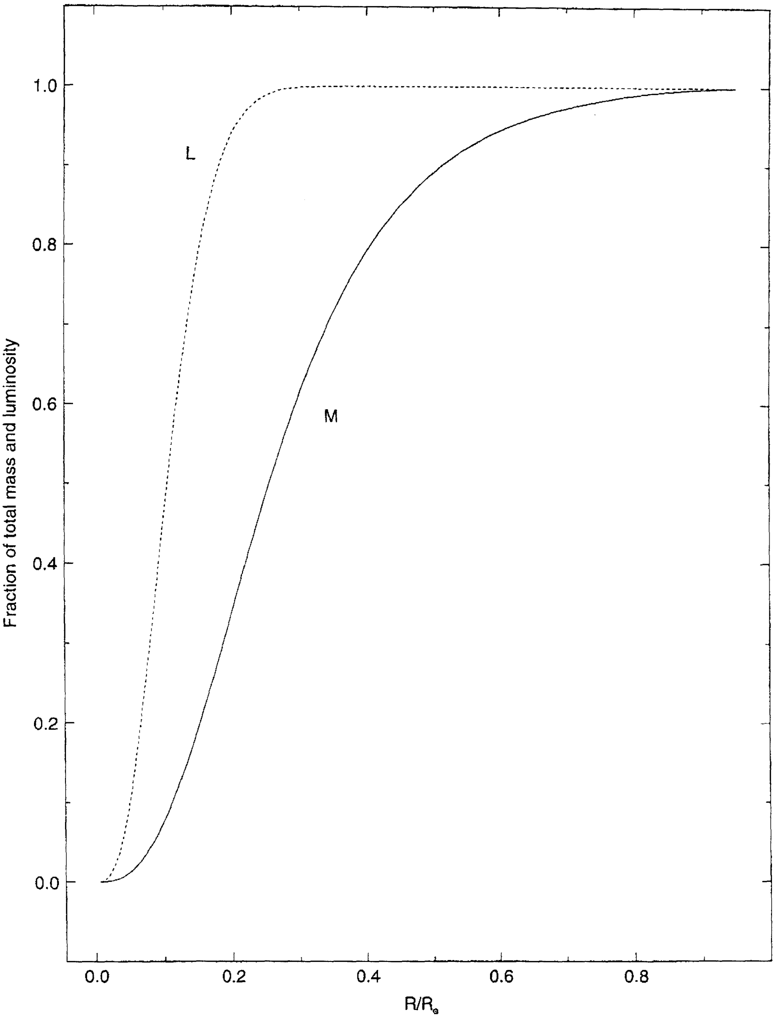

FIG. 2. The fractions of the total luminosity (dashed line) and mass (solid line) as functions of radius fraction in a theoretical solar model.

ities and the gravitational settling of helium are in good agreement with the measured depth of the surface convection zone.

4.3 Abundance Profiles

The hydrogen and metal abundances as functions of radius are shown in Fig. 6. The strong decrease in the central hydrogen abundance within $0.2R_\odot$ is caused by hydrogen burning; the increase in hydrogen in the surface layers is a consequence of the gravitational settling of ^{4}He. Because the Sun is not burning elements heavier than ^{4}He, the radial profile of the metal abundance shows the effect of gravitational settling of heavy elements from an initial mass fraction of 0.02. There is a 10% drop in the surface metal abundance and a 5% increase in the central metal abundance that result from settling.

Different species burn at different temperatures; the abundances as functions of radius of ^{3}He and the major isotopes involved in the CNO cycle are shown in Fig. 7. The small increases in abundance of C, N, and O below the convection zone are caused by diffusion (diffusion of ^{3}He was not included in this calcula-

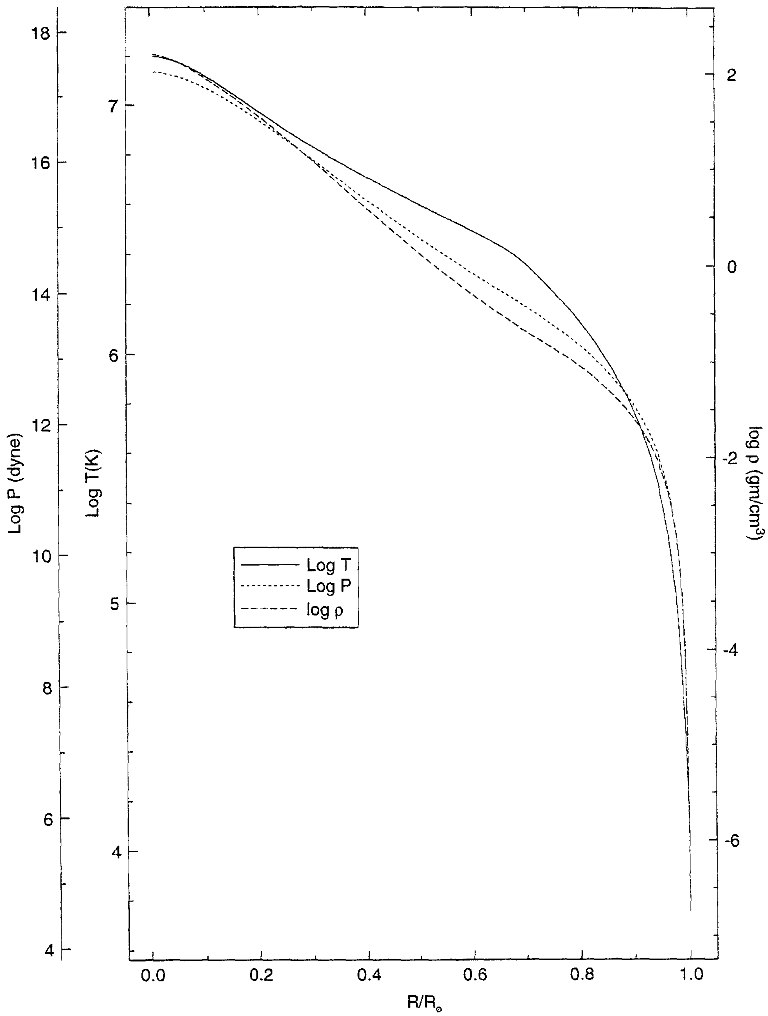

FIG. 3. Pressure, temperature, and density on a logarithmic scale as functions of radius fraction in a theoretical solar model.

tion). The peak in the abundance of ^{3}He at $0.26R_\odot$ and the peak in the abundance of ^{13}C at $0.18R_\odot$ are caused by nonequilibrium nuclear burning. ^{12}C burns into ^{13}C at a lower temperature than ^{13}C burns into ^{14}N; there will therefore be a region where ^{13}C is produced more efficiently than it is destroyed. A similar phenomenon occurs for ^{3}He. The strong increase in N and decrease in C interior to $0.15R_\odot$ occurs because the surface C/N ratio (3) is different from the C/N nuclear equilibrium ratio of 1/200. As the temperature rises, the rate of the reactions that destroy ^{3}He rise faster than the rate of the reaction that produces ^{3}He; as a result, the ^{3}He abundance decreases dramatically in the solar core. There is a slight depletion of O and increase in N in the very central region from partial processing of oxygen into nitrogen, which occurs only at relatively high temperatures.

We discuss the detailed properties of the three major layers of the Sun (convective envelope, radiative mantle, and nuclear burning core) separately below.

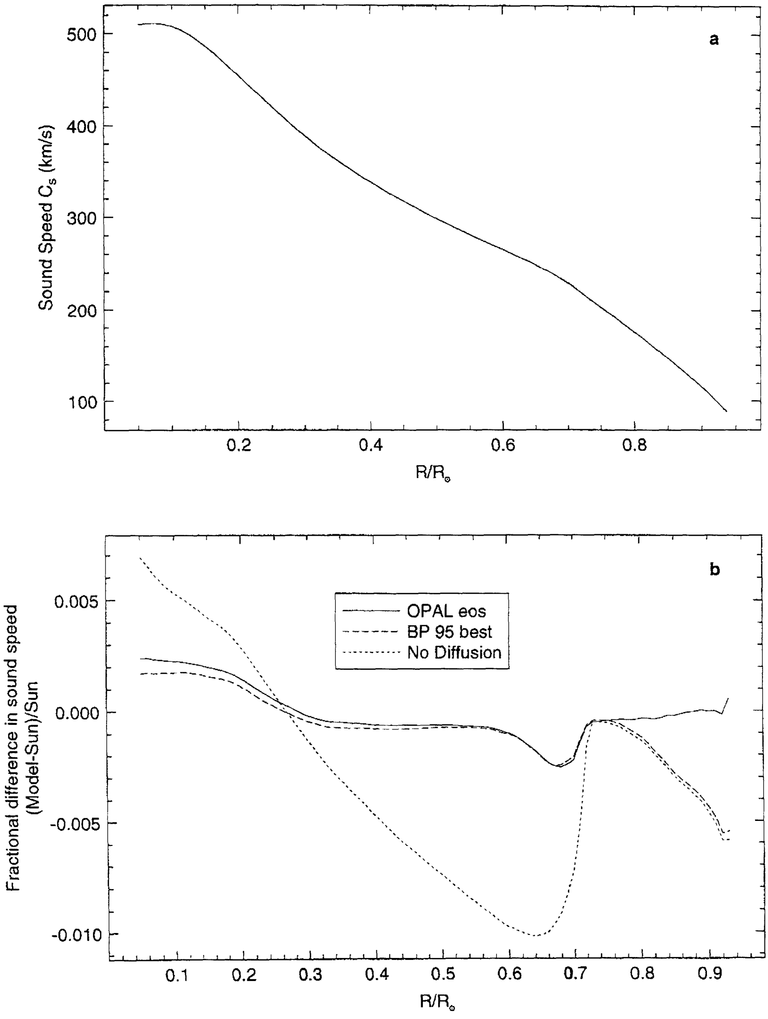

FIG. 4. (a) The sound speed as a function of radius fraction that is inferred from helioseismology; the outer layers ($R > 0.94R_\odot$) are not shown. Data are from S. Basu and J. Christensen-Dalsgaard (personal communication). (b) The difference between the real Sun and different theoretical models. The short-dashed line is the standard model of Bahcall and Pinsonneault (1992), which does not include gravitational settling; the long-dashed line is the model of Bahcall *et al.* (1995), which includes gravitational settling; the solid line is the same model with the improved OPAL equation of state (Rogers *et al.*, 1996).

4.4 Convection Zone

The opacity of matter to radiation is high at low temperatures and in regions where ionization is occurring; both conditions characterize the outer layers of the Sun. The ionization structure of the outer layers of the Sun is illustrated in Fig. 8. In this region, energy is transported through mass motions (convection). Because the time scale for energy transport via convection is short (of the order of a month) compared with the lifetime of the Sun (4.57 Ga) the convection zone is fully mixed. Typical convective velocities are of order 1 km s^{-1} and gradually decline with increased depth.

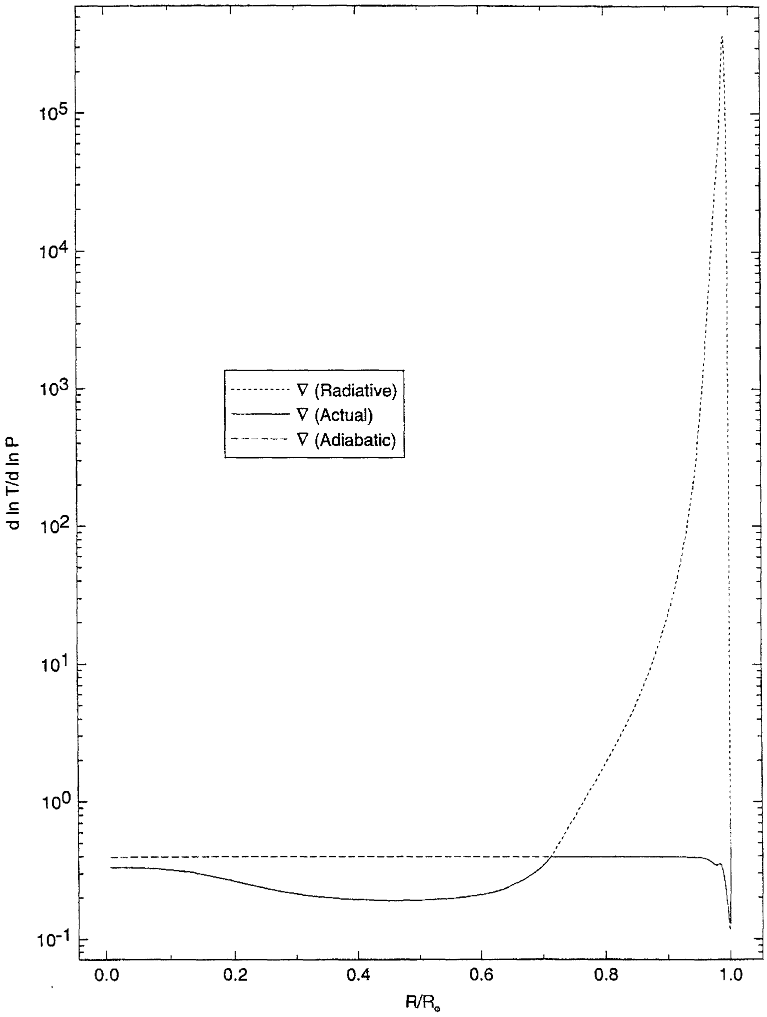

FIG. 5. Actual, radiative, and adiabatic temperature gradients as functions of radius fraction in a theoretical solar model. The actual and radiative gradients are identical for $R < 0.713R_\odot$; the actual gradient is very close to the adiabatic gradient for $R > 0.713R_\odot$ except for a thin layer near the surface (see Fig. 9).

4.4.1 Superadiabatic Layer In the outer layers of the Sun there is a transition from mostly neutral H and He at the surface to full ionization. In this region the adiabatic temperature gradient drops and the actual temperature gradient differs significantly from the adiabatic value (Fig. 9). In this superadiabatic region the model properties are sensitive to the treatment of convection. In the lower portion of the convection zone the superadiabatic temperature gradient is small.

4.4.2 Surface versus Initial Abundances The surface mass fraction of ^{4}He inferred from helioseismology is 0.242 ± 0.003, while the initial Y found from solar models ranges from 0.27 to 0.29. This is evidence for mild gravitational settling of ^{4}He; theoretical calculations predict a comparable degree of settling for heavier elements ($\approx 10\%$).

The surface abundances of both Li and Be are significantly below their meteoritic values, and the base of the convection zone is not deep

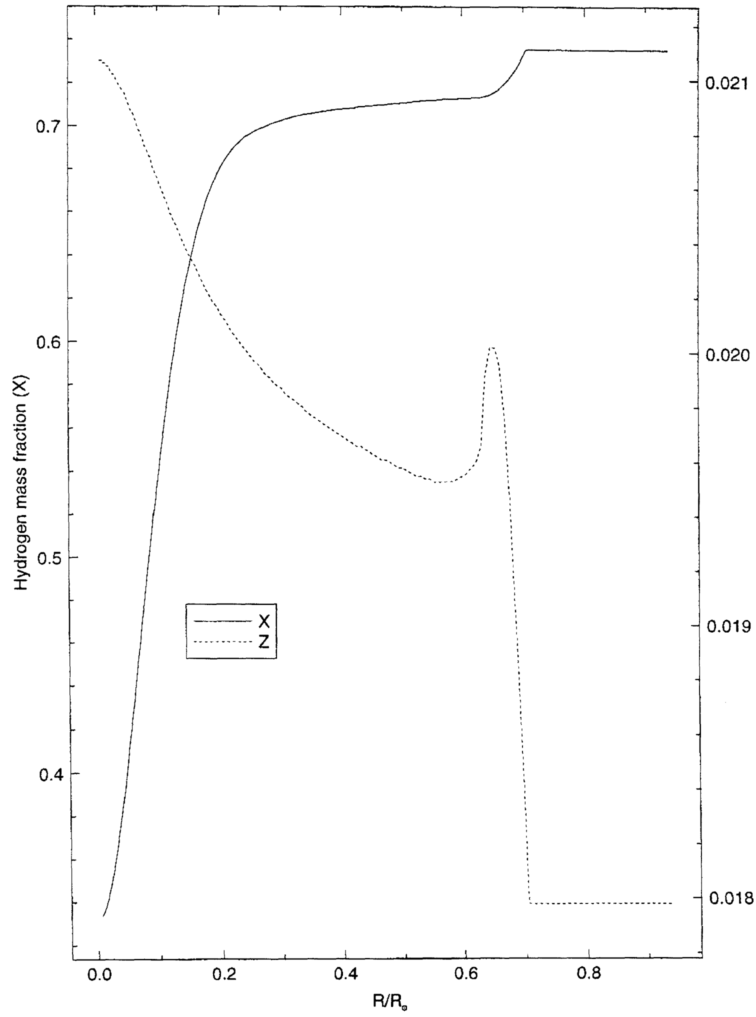

FIG. 6. Mass fractions of hydrogen (solid line, left axis) and all elements heavier than helium (dashed line, right axis) as functions of radius fraction in a theoretical solar model.

and hot enough to burn either. Although stars like the Sun do experience mild Li depletion early in their lifetime, they are observed to have abundances much higher than the current solar photospheric value. The expected degree of gravitational settling for Li is of of order of that found for He, and the degree expected for Be is smaller. A mild mixing mechanism, in addition to convection, is therefore required to explain the surface light element abundances, Models that include mixing induced by rotation are capable of reproducing the observed surface abundances of both Li and Be.

4.4.3 Rotation as a Function of Depth and Latitude The seismic inversions indicate that the surface differential rotation as a function of latitude persists throughout the surface convection zone. The rotation as a function of depth at a given latitude is nearly constant through the surface convection zone. There is a narrow region at the base of the surface

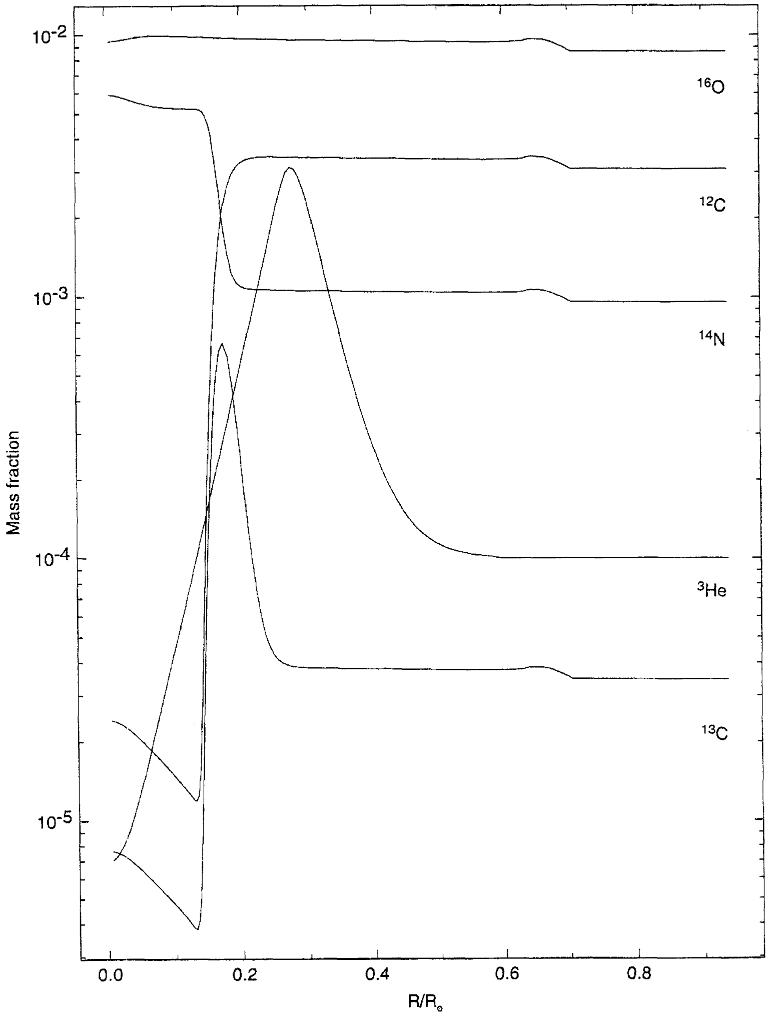

FIG. 7. Mass fractions of ^{3}He and the important C, N, and O isotopes on a logarithmic scale as functions of radius fraction in a theoretical solar model.

convection zone where differential rotation as a function of latitude vanishes; the depth of this region is currently unresolved. The layer below the surface convection zone is the likely seat of the solar dynamo, responsible for the sunspot cycle and the generation of the Sun's magnetic field.

4.5 Radiative Mantle

Below $0.713R_\odot$ the Sun is stable against convection. Nuclear energy generation does not become significant until about $0.2R_\odot$; the intervening region can be considered to be a radiative mantle surrounding the nuclear-burning core.

4.5.1 Evidence for Microscopic Diffusion and Mixing The depth of the solar convection zone and the surface He abundance inferred from helioseismology both require gravitational settling of helium in the solar envelope. Models with only *diffusion*, however, predict a sharp change in the mean molecular mass below the surface convection zone, which is not seen; this produces the discrepancy between the model and the Sun seen

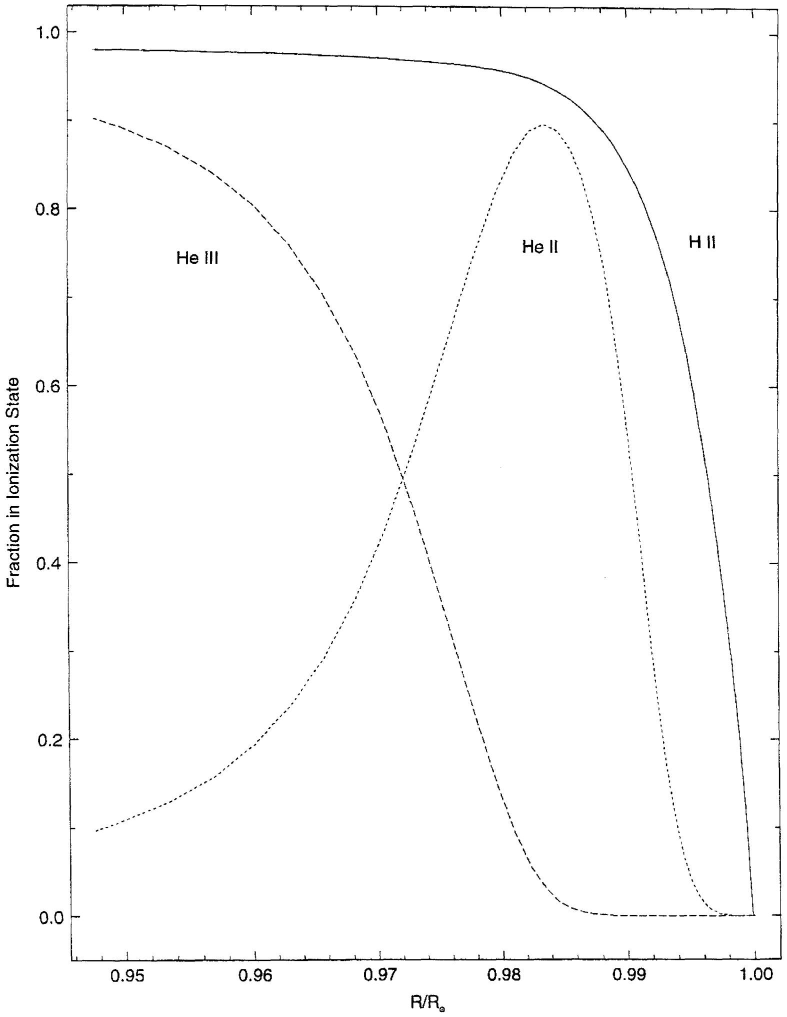

FIG. 8. The fractions of hydrogen that is fully ionized (H II, solid line), helium that is singly ionized (He II, short-dashed line) and helium that is doubly ionized (He III, long-dashed line) as functions of radius fraction in the outer layers of a theoretical solar model.

in Fig. 4(b). In addition, the surface abundances of lithium and beryllium require a slow mixing mechanism. The survival of these light elements places strong constraints on how much mixing can have occurred, and calculations indicate that mixing consistent with the surface abundances cannot entirely remove the effects of gravitational settling. Combined with the data on the solar rotation curve, this implies that the mantle of the Sun is the site of both slow separation and mixing processes not present in standard stellar models.

4.5.2 Rotation as a Function of Depth

There is general agreement that the rotation of the Sun depends neither on latitude nor on depth between a transition region near the base of the convection zone and $\approx 0.4R_\odot$. The shear at the base of the solar convection zone is a likely site for the solar dynamo that gives rise to the Sun's magnetic field. There have been contradictory claims, ranging from a mild decrease to rigid rotation to increased rotation in the region between $0.2R_\odot$ and $0.4R_\odot$. The most recent inversions of seismic data

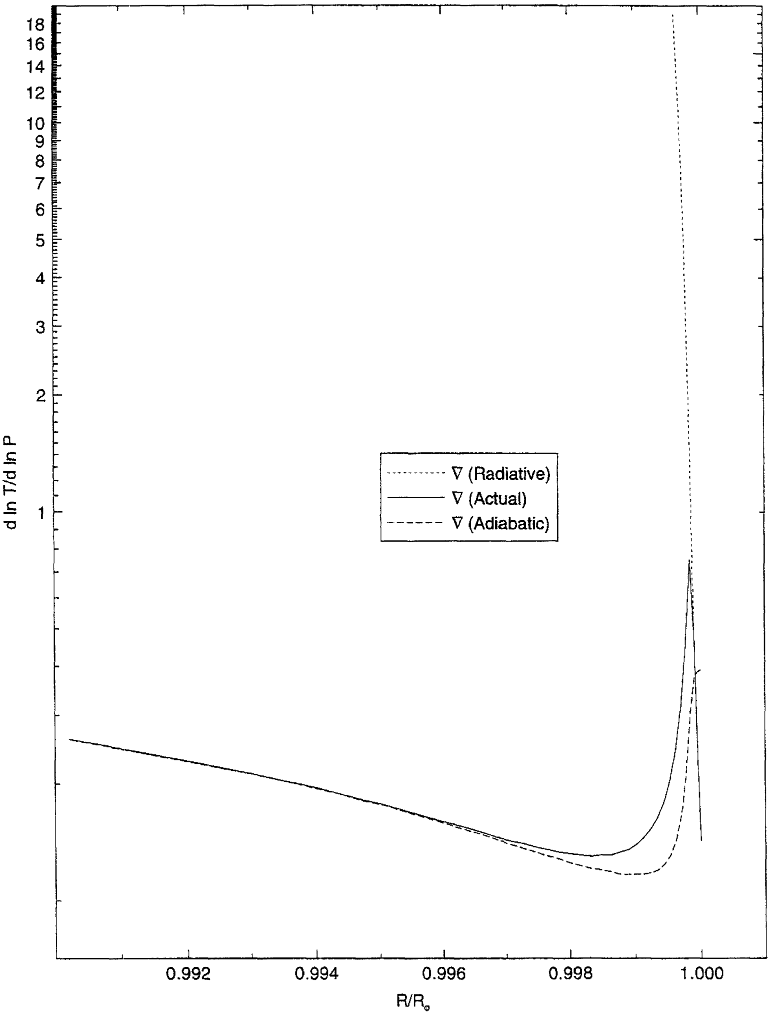

FIG. 9. An enlarged portion of Fig. 5, showing the behavior of the adiabatic, radiative, and actual temperature gradients near the solar surface as functions of radius fraction. The actual temperature gradient starts to depart significantly from the adiabatic gradient at 0.996; the radiative gradient is off the scale before 0.9985 and drops sharply with increased radius, becoming less than the adiabatic gradient just below the solar atmosphere.

favor essentially solid-body rotation down to $0.2R_\odot$ (Tomczyk *et al.*, 1995). Because the Sun is experiencing angular-momentum loss via the solar wind, this implies that the time scale for angular-momentum transport in the radiative core is less than the age of the Sun.

4.6 Nuclear-Burning Core

The deep core is the site of the Sun's energy generation. Conditions in this region are tested directly by the solar neutrino experiments, which are in disagreement with theory. The helioseismic data, however, are in much better agreement with the theoretical models in the core. Because large changes in the solar-model fluxes would required large changes in the thermal structure of the solar core, this severely restricts the allowed classes of solar-model solutions for the solar neutrino problem.

4.6.1 Thermal Structure Nuclear burning has altered the composition and temperature of the core of the Sun since its birth. To first order, the local sound speed scales as $(T/\mu)^{1/2}$. The agreement between observation and theory is at the 0.2% level in the core, which would require finely tuned changes in the temperature and composition to preserve the close observed match between the two.

4.6.2 Abundances and Energy Generation as Functions of Depth A series of different nuclear reactions produce energy in the Sun. Because nuclear reactions are strongly temperature sensitive, the reaction rates are usually a maximum at the center of the Sun. The differing temperature dependences of the individual reactions, however, imply that the degree of central concentration of the energy generation will differ from one reaction to the next. This in turn is reflected in the detailed abundance profiles seen in Figs. 6 and 7. In later phases of evolution the convection zone of Sunlike stars becomes much deeper and incorporates partially processed material from the core into their surface convection zone where it can be seen. Surface abundance changes from this dredge-up have been seen and are in fairly good agreement with theoretical predictions.

4.6.3 Predicted Neutrino Fluxes The neutrino fluxes from the different experiments are in contradiction with those predicted by theoretical models. The measured thermal structure is in good agreement with theoretical models; this implies that a large reduction in the neutrino fluxes predicted by these models is unlikely. The most likely explanation is that the Sun is producing neutrinos in the amount expected from the theoretical models, but they are being converted from electron neutrinos (which can be detected) into other kinds of neutrinos, which cannot be measured by the current generation of experiments. If there is a nonzero neutrino mass, electron neutrinos can be converted into other flavors between the region where they are produced and the Earth. The SNO experiment, which is currently under way, is sensitive to both neutral- and charged-current reactions. If the electron neutrino oscillates into a μ or τ neutrino, SNO should be able to provide direct experimental confirmation of this phenomenon. Below we compare the theoretical predictions and the observational data.

4.6.3.1 Kamiokande. The measured flux of ^{8}B neutrinos (with energies above 7.5 MeV) is $[3.0 \pm 0.41(\mathrm{stat}) \pm 0.35(\mathrm{syst})] \times 10^6\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. This compares with a theoretical prediction of $6.6(1.0^{+0.14}_{-0.17}) \times 10^6\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. The observed flux is thus only about 45% of the predicted value.

4.6.3.2 Homestake. The Homestake experiment is sensitive to both the higher-energy ^{8}B neutrinos and those from other sources, such as ^{7}Be and the pep reaction. It is not sensitive to the lowest-energy neutrinos arising directly from the pp reaction. The measured flux (in units of 1 SNU, or 10^{-36} interaction per target atom per second) is $2.55 \pm 0.17(\mathrm{stat}) \pm 0.18(\mathrm{syst})$ SNU; the theoretical prediction is $9.3^{+1.2}_{-1.4}$ SNU. This discrepancy is particularly difficult to explain with standard models in conjunction with the Kamiokande data discussed above. The measured Kamiokande flux can be converted into a minimum flux expected for the Homestake experiment, which leaves little or no room for any contribution from neutrinos produced by other reactions to which Homestake is sensitive.

4.6.3.3 GALLEX/SAGE. The GALLEX and SAGE experiments are sensitive to all of the neutrinos detected by the Homestake and Kamiokande experiments, plus the low-energy pp neutrinos. The measured fluxes from GALLEX and SAGE are, respectively, $79 \pm 10(\mathrm{stat}) \pm 6(\mathrm{syst})$ SNU and $69 \pm 10(\mathrm{stat}) \pm 6(\mathrm{syst})$ SNU. This compares with a theoretical prediction of 137^{+8}_{-7} SNU, which is relatively weakly model dependent. Of this total about 70 SNU arises directly from the pp reaction; the magnitude of this component is fixed by the measured energy output of the Sun.

5. SUMMARY

The Sun can be divided up into three distinct layers: the nuclear-burning core, a radiative mantle, and a surface convection zone. The speed of sound as a function of depth provides strong constraints on theoretical models, and the overall agreement between theory and observations is at a level (0.2% in the sound speed) not typically found in astronomy. The observed solar neutrino fluxes do not agree with those expected from theoretical models; the most likely explanation is a nonzero neutrino mass. A detailed comparison of model properties with the solar data provides evidence for element-separation processes and mild envelope mixing below the surface convection zone.

GLOSSARY

Adiabatic Temperature Gradient (∇_{ad}): The logarithmic derivative of temperature with respect to pressure for a fluid element that does not exchange heat with its surroundings. If the radiative temperature gradient exceeds the adiabatic temperature gradient in a given region, the region is unstable with respect to convection and energy is transported by mass motions rather than by photon diffusion.

CNO Cycle: A series of nuclear reactions in which carbon, nitrogen, and oxygen nuclei are used as catalysts for the conversion of hydrogen into helium. These reactions account for about 2% of the solar energy output.

Convection Zone: A region where energy is transported by mass motions; this occurs in the outer layers of the Sun.

g Mode: A mode of oscillation in the Sun where the restoring force of the wave is gravity; these modes have not been definitively identified in the Sun, but are predicted to have a characteristic period of 3 h.

Helioseismology: The study of the internal structure of the Sun deduced from the properties of nonradial oscillations observed at the solar surface.

Mixing-Length Theory: A simplified theory of convective energy transport used extensively in theoretical solar models, in which a characteristic length scale α is assumed and the characteristic velocities of fluid elements and the temperature gradient are then calculated. The value of the mixing length α is usually adjusted to reproduce the solar radius at the solar age.

Opacity: A measure of the mean free path of photons in the presence of matter; a typical length scale is 1 cm in the Sun. Astrophysical calculations use a weighted average over photon frequency, called the Rosseland mean opacity, and solve a diffusion equation to infer the temperature gradient needed to carry a given flux.

p Mode: An acoustic mode of oscillation in the Sun where the restoring force of the wave is pressure; these modes have a characteristic period of 5 min.

pp Chain: A series of nuclear reactions in which protons combine directly to form the intermediate species deuterium and helium-3, followed by three different branches that produce helium-4. These reactions produce 98% of the solar energy output.

Radiative Core: A region where energy is transported by photon diffusion; this occurs in the inner layers of the Sun.

Radiative Temperature Gradient (∇_{rad}): The logarithmic derivative of temperature with respect to pressure that would be required to carry the local flux by the diffusion of photons.

Works Cited

Anders, E., Grevesse, N. (1989), *Geochim. Cosmochim. Acta* **53**, 197–214.

Bahcall, J. N. (1989), *Neutrino Astrophysics*, Cambridge, UK: Cambridge Univ. Press.

Bahcall, J. N., Pinsonneault, M. H. (1992), *Rev. Mod. Phys.* **64**, 885–926.

Bahcall, J. N., Pinsonneault, M. H., Wasserburg, G. J. (1995), *Rev. Mod. Phys.* **67**, 781–808.

Cox, A. N., Livingston, W. C., Matthews, M. S. (Eds.) (1991), *Solar Interior and Atmosphere*, Tucson: Univ. of Arizona Press.

Iglesias, C. A., Rogers, F. J. (1996); *Astrophys. J.* **464**, 943–953.

Lang, K. R. (1991), *Astrophysical Data: Planets and Stars*, New York: Springer-Verlag.

Libbrecht, K. G., Morrow, C. A. (1991), in: A. N. Cox, W. N. Livingston, M. S. Matthews (Eds.), *Solar Interior and Atmosphere*, Tucson: Univ. of Arizona Press.

Richard, O., Vauclair, S., Charbonnel, C., Dziembowski, W. A. (1996), *Astron. Astrophys.* **312**, 1000–1011.

Rogers, F. J., Swenson, F. J., Iglesias, C. A. (1996), *Astrophys. J.*, **456**, 902–908.

Tomczyk S., Schou J., Thompson M. J., (1995). *Astrophys. J. Lett.*, **448**, L57–L60.

Further Reading

Bahcall, J. N. (1989), *Neutrino Astrophysics*, Cambridge, UK: Cambridge Univ. Press. A discussion of the solar neutrino problem, including the experiments and nuclear physics relevant to the solar

problem. Updated review articles on the current status of the solar models are listed in Bahcall and Pinsonneault (1992) and Bahcall *et al.* (1995).

Clayton, D. (1983), *Principles of Stellar Evolution and Nucleosynthesis*, Chicago: Univ. of Chicago Press. A discussion of the physics appropriate for the interiors of stars.

Cox, A. N., Livingston, W. C., Matthews, M. S. (1991), *Solar Interior and Atmosphere*, Tucson: Univ. of Arizona Press. An extensive compilation of review articles on all aspects of the Sun, both observational and theoretical.

Kippenhahn, R., Weigart, A. (1990), *Stellar Structure and Evolution*, New York: Springer-Verlag. An introduction to stellar interiors, along with a discussion of the evolution of stars. A resource for putting the Sun in the broader context of astronomy.

SUPERFLUIDITY: LIQUID HELIUM SYSTEMS

Shaun N. Fisher and George R. Pickett, *School of Physics and Chemistry, Lancaster University, Lancaster, United Kingdom*

Introduction 295
1. **What is Superfluidity?** 296
1.1 Bosons, Fermions 296
1.2 Bose–Einstein Condensation 297
1.3 Cooper Pairing 298
2. **Superfluid ^{4}He** 298
2.1 Cooling Methods 298
2.2 The Superfluid Mechanism 299
2.3 The Lambda Transition 299
2.4 The Phase Diagram 300
2.5 The Two-Fluid Model 300
2.5.1 Hydrodynamic Heat Flow 301
2.5.2 Thermomechanical Effect 301
2.5.3 Second Sound 302
2.6 Excitations and Superflow 302
2.7 Persistent Currents 303
2.8 The Wave Function and Phase Coherence 304
2.9 Vortices 304
2.9.1 Rotating Superfluid ^{4}He 305
2.9.2 Measurement of Absolute Rotation with Superfluid ^{4}He 306
3. **Superfluid ^{3}He** 306
3.1 Cooling Techniques 306
3.1.1 Pomeranchuk Cooling 307
3.1.2 Nuclear Cooling 307
3.2 A Multiplicity of Possible Phases 308
3.2.1 The A Phase 309
3.2.2 The B Phase 310
3.3 Broken Symmetries of the Superfluid Phases 311
3.3.1 The A phase 311
3.3.2 The B phase 311
3.4 The Phase Diagram 311
3.5 Properties of Superfluid ^{3}He 312
3.5.1 Thermal Excitations and Superflow 312
3.5.2 Thermal Properties of Superfluid ^{3}He 313
3.5.3 Textures 313
3.5.4 Vortices 314
3.5.5 Magnetic Properties 315
3.5.5.1 Nuclear Magnetic Resonance 315
3.5.5.2 Coherent Excited States in ^{3}He-B 316
4. **Other Similar Superfluids** 317
Glossary 318
Works Cited 319
Further Reading 319

INTRODUCTION

The two common isotopes of helium, ^{3}He and ^{4}He, are unique in remaining liquid down to absolute zero temperature. At low temperatures the liquid phases of both isotopes become superfluid. The most apparent property of a superfluid is its ability to flow without friction. This leads to the possibility of non-decaying persistent currents flowing indefinitely around loops. However, on a more fundamental level, the really unique aspects of superfluidity lie in the influence of the underlying quantum mechanics on the macroscopic behavior of the liquid.

Superfluidity was first observed in ^{4}He just before the second World War, and the effect was elucidated and named by Kapitza in Moscow in 1942. The superfluid transition temperature in liquid ^{4}He is close to 2 K. The superfluid transition in liquid ^{3}He does not occur until much lower temperatures, about 1 mK, and was first observed in 1971 by Osheroff, Richardson, and Lee, who received the Nobel Prize for this work in 1996.

The properties of the two helium isotopes are almost identical at ordinary temperatures; they are inert gasses with very weak interatomic forces. Nevertheless, their superfluid states are very different. The differences arise as a direct consequence of the nuclear spin, which is zero for ^{4}He and $\frac{1}{2}$ for ^{3}He. Even though the interatomic forces are almost identical for the two isotopes, quantum mechanics

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

dictates that liquids at very low temperatures will behave very differently depending on the net intrinsic spin of the constituent particles.

Superfluidity, unlike the analogous superconductivity seen in some metals and conducting ceramics, has as yet found very few practical applications outside low-temperature research. The real impact of the superfluid state is the simplicity with which it can be described. Quantum mechanics describes a superfluid, consisting of countless numbers of atoms, as a single macroscopic object (wave function). These systems therefore provide a very rich laboratory for testing ideas from other areas of physics, even extending as far as particle physics and cosmology.

1. WHAT IS SUPERFLUIDITY?

The two isotopes of helium, the common isotope ^{4}He and the lighter isotope ^{3}He, are so light and so weakly interacting that their behavior is dominated by quantum mechanics much more than that of heavier elements. As a result, helium in the liquid phase is often described as a quantum fluid.

The laws of thermodynamics dictate that a system becomes more ordered as the temperature is reduced. A typical example is the solidification of a substance (such as the freezing of water) as the temperature is reduced below the melting point. The level of order of the system is increased on solidification, since in the liquid the component atoms or molecules are randomly distributed in space, whereas in the solid they take up fixed positions relative to each other and in general form a regular ordered crystalline structure.

The property of superfluidity in helium follows from the fact that helium is unique in remaining liquid down to absolute zero temperature and will only solidify if a pressure of some 30 atm is applied. It is the unusually strong influence of quantum mechanics on helium that prevents solidification. To form a solid, each atom must be confinable in its place in the crystal. However, a particle may only be confined within a region of space at the expense of increasing its energy (the zero-point energy). For helium, which has weak interatomic forces, this energy is large enough to overcome the attractive forces between the atoms that would otherwise lead to solidification at a sufficiently low temperature.

While liquid helium cannot increase its order by becoming solid, there are other ways in which it may order. In fact, at some low temperature the atoms in the liquid begin to fall into the same-lowest energy quantum mechanical state. It is this process that leads to superfluidity. The term was first coined by Kapitza to describe the properties of liquid ^{4}He.

Once a significant fraction of the atoms occupy the same quantum mechanical state, the behavior becomes quite different from what we intuitively expect for normal liquids. For example, there is no viscosity. Viscosity implies the process whereby atoms are mutually scattered (with each atom changing its state) and information about velocity inhomogeneities is gradually transmitted through the liquid. If all the atoms are in the same state in the first place, then such scattering processes cannot occur; *i.e.* the viscosity is zero. With zero viscosity, flows of the liquid can be set up without any pressure gradient. Furthermore, in the absence of scattering processes, these currents may, in principle, continue indefinitely. It is this behavior that gave rise to the term superfluidity in analogy with the term superconductivity, which describes the closely related behavior of the electron gas in certain metals.

1.1 Bosons, Fermions

While both isotopes of helium (^{4}He and ^{3}He) become superfluid, the mechanism by which the ordering takes place is very different for the two isotopes.

All particles have an intrinsic angular moment or spin. Quantum mechanics distinguishes between two types of particle depending on the magnitude of their spin. These are known as fermions and bosons and have very different properties. Fermions are particles having a half-integral spin quantum number, whereas bosons have an integral spin quantum number. The building blocks of matter (neutrons, protons and electrons) are all fermions, whereas particles that carry information about forces—for example, photons (telling about the motion of charges)—are bosons. The distinction applies also to composite objects. An object made up of an even number of

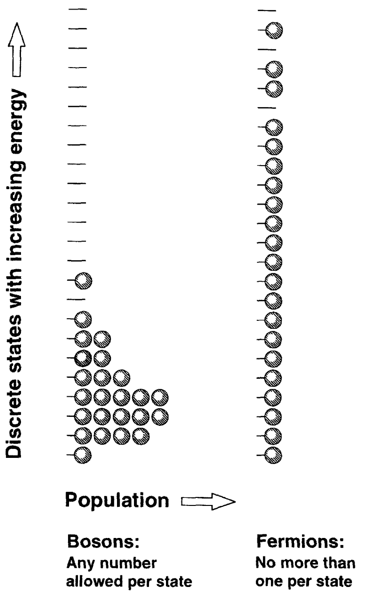
Discrete states with increasing energy
Population
Bosons:
Any number allowed per state
Fermions:
No more than one per state

FIG. 1. The filling of energy states by bosons and fermions. Any number of identical bosons may fill any one state whereas only one fermion is allowed per state.

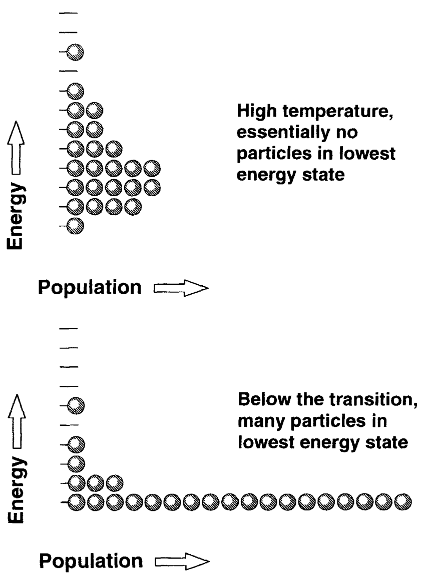
Energy
High temperature, essentially no particles in lowest energy state
Population
Energy
Below the transition, many particles in lowest energy state
Population

FIG. 2. Bose–Einstein condensation; below a critical temperature a macroscopic number of particles begins to populate the lowest energy state.

fermions becomes a boson (because the total spin in now integral), while an object constructed of an odd number of fermions remains a fermion.

A ^{4}He atom is made up of two electrons, two protons and two neutrons. The ^{4}He atom is therefore a boson (with spin 0). In contrast, ^{3}He is made from two electrons, two protons but only one neutron, and therefore the atom (with spin $\frac{1}{2}$) is a fermion.

The quantum mechanical distinction between these classes of particle comes into play when we have several particles occupying the same space. As shown schematically in Fig. 1, fermions are only allowed to occupy a quantum mechanical state singly, *i.e.*, no more than one particle per state. Bosons, on the other hand, face no such restriction, and there is no limit to how many particles may be in the same state with identical energies and momenta.

1.2 Bose–Einstein Condensation

At high temperatures, an assembly of bosons will be distributed over a large number of states. As the temperature is reduced the particles take up a distribution centered on lower and lower energies until at some temperature particles start filling the lowest state or ground state. This process is known as the Bose–Einstein condensation, and the temperature at which it begins is the Bose–Einstein condensation temperature (T_{BE}). This process is illustrated in Fig. 2. When a macroscopically large number of the particles occupy a single state the level of order in the system increases. In this case, unlike in a crystal, the order is not in terms of position but in terms of momentum, since the particles collect in the lowest state where the momentum is zero.

A system of bosons which are all in the same state displays some very unusual properties. If all the bosons are identical in every sense then they must behave as a single "object". It is meaningless to think of the system in terms of distinct individual particles since (according to quantum mechanics) it is fundamentally impossible to distinguish one particle from another. In consequence, while the bosons are governed by the laws of quantum mechanics only on the *microscopic scale for* temperatures higher than T_{BE}, at lower temperatures, once there is a macroscopic occupation of the lowest state, the bosons coalesce into a single macroscopic "object" represented by a single wave function. The system is still governed by quantum mechanics, but now operating on the entire macroscopic sample. Such an assembly of particles described by a common wave function is known as a coherent state.

1.3 Cooper Pairing

Direct Bose–Einstein condensation is impossible for a system of independent fermions since no more than one fermion may occupy a single state. Indeed, at moderately low temperatures a system of weakly interacting fermions has all the low-lying single-particle states singly occupied up to a cutoff energy called the Fermi energy. Nevertheless, liquid ^{3}He undergoes a transition into a superfluid state and also the conduction electrons of many metals undergo an analogous transition into a superconducting state. The superfluid state clearly exhibits properties associated with a coherent quantum state and therefore must represent an ordering related to Bose–Einstein condensation but in a system of fermions.

The solution to the problem was provided by Bardeen, Cooper, and Schrieffer (BCS) in 1957 (Bardeen *et al.*, 1957) who found that an attraction between two fermions with energies close to the Fermi energy will always lead to the formation of bound pairs at sufficiently low temperatures. These pairs are known as Cooper pairs. While the work of BCS was in the context of superconductivity, their ideas are equally applicable to ^{3}He atoms. The pairs prefer to form from fermions with equal and opposite momenta such that the combined momentum of the pair is zero. Furthermore, since an even number of fermions is involved, each pair is itself a boson. Therefore at sufficiently low temperatures, any attractive force between the fermions will lead to the production of Cooper-pair bosons, which are free to form a coherent ground state. This is, of course, a variant of Bose–Einstein condensation in which the bosons are formed from pairs of fermions.

The binding energy of the Cooper pairs ensures that the pairs have a lower energy than the constituent unbound particles. This means that there must be a minimum energy difference between an unpaired particle and those particles which are paired. This is known as the energy gap, usually represented by Δ. When a Cooper pair is broken, at least two unpaired particle excitations must be produced, needing a minimum energy of 2Δ.

2. SUPERFLUID ^{4}He

2.1 Cooling Methods

To reach the temperature at which liquid ^{4}He becomes superfluid at around 2 K requires only quite modest equipment. All that is needed is a simple insulating glass Dewar flask that may be filled with liquid ^{4}He and a modest vacuum pump to reduce the pressure of the vapor phase to reduce the temperature. A suitable system is shown in Fig. 3. It is particularly convenient for experimentalists that the superfluid transition can be reached simply by the reduction of the pressure over the liquid phase.

Liquid ^{4}He in equilibrium with its vapor at

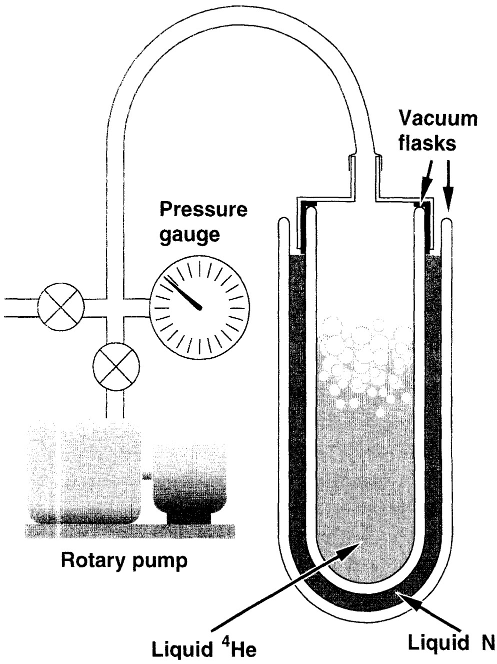

FIG. 3. A simple arrangement for reaching the superfluid transition in liquid ^{4}He.

one atmosphere sits at a temperature of 4.2 K. If the pressure above the liquid is slowly reduced the liquid will cool from the extraction of the latent heat of evaporation. At the modest vapor pressure of around 0.05 atm the temperature of the liquid–vapor mixture has fallen to 2.2 K, at which point the transition to superfluidity occurs.

The unique behavior of the superfluid state can already be appreciated with the naked eye even with such a modest arrangement. As the liquid is cooled to the transition temperature, it boils violently. A striking feature of the phase transition is the abrupt cessation of boiling once the superfluid state has been entered. The process of boiling arises from temperature gradients in the fluid. In a pan of boiling water, vapor bubbles will form in the hottest regions of the water at the heated surface of the pan. The abrupt disappearance of boiling in liquid ^{4}He below the transition suggests that temperature gradients must be greatly reduced in the superfluid phase. Clearly something very unusual has happened to the thermal properties, since the thermal conductivity of the superfluid phase must be extremely high. Vapor must still evaporate from the superfluid when the pressure is further reduced, but the vapor now forms at the free surface of the liquid rather than on the walls of the container. Many properties of superfluids are similarly unusual and quite at variance with what would be expected for a normal liquid.

2.2 The Superfluid Mechanism

The basic mechanism for superfluidity in ^{4}He is the Bose–Einstein condensation of the ^{4}He atoms. The calculated "ideal" Bose–Einstein condensation temperature falls within a factor of two of the measured superfluid transition temperature. However, the real situation is more complex than the ideal case. The ^{4}He atoms cannot be considered weakly interacting, since the superfluidity occurs in the liquid phase where interactions are obviously important. They are in any case responsible for the liquid state itself. Therefore it is not surprising that an ideal Bose–Einstein condensation provides only a very qualitative description of superfluidity in ^{4}He.

Nevertheless, the superfluid properties of ^{4}He are certainly due to a boson condensate. The nonideality of the superfluid ^{4}He condensate is, in many ways, responsible for many of the more interesting superfluid properties. For example, neutron-scattering experiments (Campbell, 1983; Sears *et al.*, 1982) show that even at very low temperatures, the fraction of ^{4}He atoms in the zero-momentum state is only about 15% compared with the ideal Bose-gas prediction of 100%. However, as one consequence of the interatomic forces, the superfluid ground state incorporates a mixture of momentum states.

2.3 The Lambda Transition

Since the superfluid state is highly ordered, the entropy of the liquid must fall rapidly below the transition, and this should give rise to an anomaly in the heat capacity. The measured heat capacity plotted against temperature is shown in Fig. 4. The form of the heat capacity is reminiscent of the Greek letter λ.

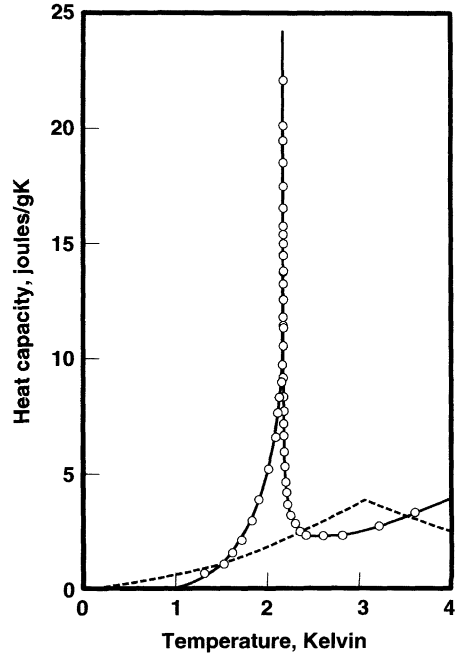

FIG. 4. The heat capacity of liquid ^{4}He showing the extremely steep lambda transition. The heat capacity for an ideal Bose–Einstein condensation in a gas of similar density is shown by the dashed line. (After Buckingham and Fairbank, 1961.)

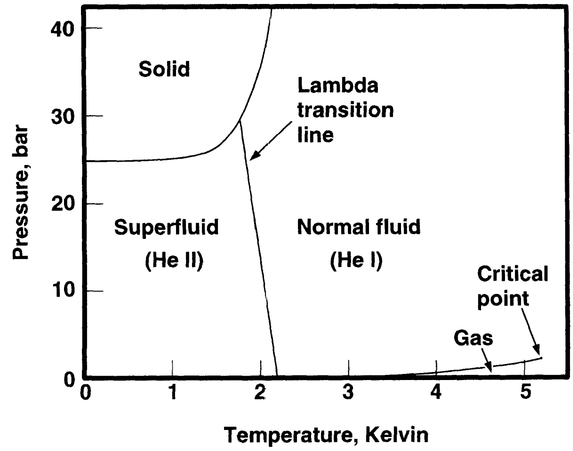

FIG. 5. The phase diagram for ^{4}He. Since liquid extends to zero temperature, there is no triple point. The designations He I for the normal liquid above the transition, He II for the superfluid phase, are often used.

Consequently the superfluid transition in ^{4}He is commonly known as the lambda transition. There is a logarithmic singularity at the transition in which the heat capacity rises dramatically and approaches infinity. This indicates the differences from the ideal Bose-gas case as indicated in Sec. 2.2 above, since the ideal Bose-gas heat capacity should display a cusp at the transition temperature as shown by the dashed curve in Fig. 4.

2.4 The Phase Diagram

The phase diagram for ^{4}He at low temperatures is shown in Fig. 5. Quite apart from the superfluidity itself, this is a very unusual system. We see, for example (as discussed in Sec. 1), that the liquid state extends all the way to zero temperature. This being so, there is clearly no triple point. Furthermore, the level of order in both the liquid and solid phases at low temperatures is virtually complete. In other words, the entropy of both phases is very small. The superfluid phase is often referred to as helium II, in contrast to the normal liquid, helium I.

2.5 The Two-Fluid Model

At absolute zero temperature, all the ^{4}He atoms may be considered to be in their lowest-energy state and therefore contribute to the superfluid properties, since, by definition, at absolute zero there are no thermal excitations. At higher temperatures there will be a finite number of excitations. In an ideal noninteracting boson system, these excitations simply represent those particles not in the ground state. However, in the case of liquid ^{4}He, which is strongly interacting, the excitations represent collective modes of the liquid involving many atomic states. These are phonons and, at higher energies, rotons, which are unique to liquid ^{4}He. At a microscopic level we cannot distinguish between atoms in the ground state and those contributing to the excitations. However, that having been said, it is very convenient to consider the liquid below the transition as consisting of a mixture of two freely interpenetrating fluid components, namely the superfluid and a normal fluid. The superfluid component has zero entropy and zero viscosity, whereas the normal component has a finite entropy and viscosity. This is the two-fluid model. The density (ρ) of the fluid can be separated into two terms:

$$\rho = \rho_s + \rho_n, \tag{1}$$

where ρ_s represents the superfluid density and ρ_n the normal-fluid density. At the lambda temperature, ρ_s is zero while $\rho_n = \rho$. As the temperature is reduced ρ_s increases and ρ_n decreases until at absolute zero $\rho_s = \rho$ and $\rho_n = 0$. Although we have noted that it is meaningless to label any particular atom as belonging to either component, nevertheless the two-fluid model, as we show below, is very successful in explaining many of the superfluid

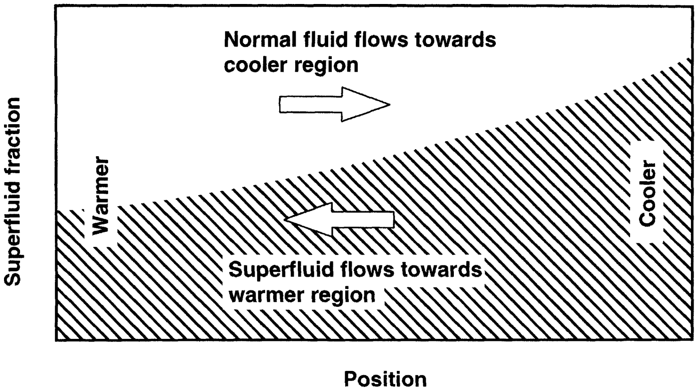
Normal fluid flows towards cooler region
Superfluid flows towards warmer region
Warmer
Cooler
Superfluid fraction
Position

FIG. 6. The counterflow of normal and superfluid ^{4}He in a temperature gradient, giving a high thermal conductance just below the transition.

properties and is a very convenient concept for describing the behavior.

2.5.1 Hydrodynamic Heat Flow We can now explain the virtually infinite thermal conductivity of the superfluid state as a natural consequence of two-fluid behavior. The superfluid component has zero entropy and therefore cannot conduct heat. Since the ratio of the amounts of normal fluid and superfluid depends on temperature, a temperature gradient imposed on the superfluid phase must be associated with a gradient in the superfluid-to-normal fluid ratio such that there is a higher density of the normal fluid towards the warmer regions of the sample. The concentration gradient gives rise to an osmotic pressure, which drives the normal fluid towards lower temperatures and the superfluid towards higher temperatures (*i.e.*, the osmotic pressure acts to restore thermal equilibrium), as shown in Fig. 6. The counterflow of the two components leads to an extremely effective "hydrodynamic" heat flow in which the heat is carried bodily by the flow of normal fluid. The effective thermal conductivity is thus extremely high, in analogy with the heat transfer by the two-phase (but much less efficient) counterflow in a heat pipe.

2.5.2 Thermomechanical Effect The high thermal conductance is just one manifestation of an important consequence of two-fluid behavior, which is the linking of mechanical and thermal behavior. This has no analogies with the behavior of normal fluids. In these thermomechanical (and mechanothermal), effects a temperature gradient gives rise to a pressure gradient (and *vice versa*). Because of the very large thermal conductivity it is difficult to maintain a significant temperature gradient in the bulk fluid. However, it is easy to establish a temperature difference between two volumes of fluid that are connected by a device that only allows the passage of the superfluid component (analogous to a semipermeable membrane). Such a device, called a "superleak", consists of an array of very narrow passages, usually those threading a porous medium, such that the normal component is very effectively clamped by its finite viscosity, while the zero-viscosity superfluid component is free to pass. When a temperature difference is maintained between the two volumes, there must be a corresponding difference in the concentrations of the superfluid component on the two sides of the superleak. The osmotic pressure gradient will force some of the superfluid component through the superleak into the hotter of the two volumes leading to a hydrostatic pressure difference. The flow will continue until in equilibrium the hydrostatic pressure difference balances the osmotic pressure difference. Conversely, if a hydrostatic pressure gradient is applied across the superleak, superfluid is forced through the superleak thus "diluting" the normal component in one volume and concentrating it in the other leading to a temperature gradient.

A rather striking demonstration of such

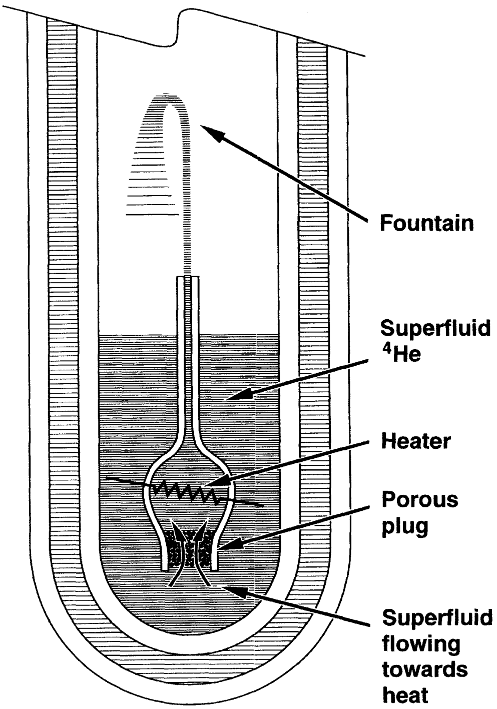

FIG. 7. The fountain effect. On heating the liquid in the container, superfluid flows in through the porous plug from the surrounding bath. The normal fluid cannot flow out to compensate. The excess liquid is therefore ejected upwards through the narrow tube as a fountain.

thermomechanical properties is provided by the fountain effect illustrated in Fig. 7. Here, a container with a narrow open neck at the top and a porous plug below is lowered into a reservoir of superfluid, such that the neck is above the liquid level. If the fluid inside the flask is then heated, the superfluid component is forced to flow from the reservoir into the flask via the pores which impede the flow of normal fluid. Since the only other outlet is the neck, the excess liquid is ejected from the neck of the flask in the form of a fountain. The effect is very dramatic: it is relatively easy to produce fountains more than 30 cm high. By this process we can pump superfluid ^{4}He around an apparatus simply by heating it.

2.5.3 Second Sound Ordinary sound consists of propagating density (or pressure) waves. Two-fluid behavior gives rise to the possibility of temperature waves in which the normal fluid and superfluid density oscillate in antiphase to each other, but such that the total density remains constant. This is called second sound and can be generated and detected in superfluid ^{4}He with a simple heater and thermometer arrangement.

2.6 Excitations and Superflow

Since the superfluid has zero viscosity, flow of the liquid must occur without the production of any excitations. The reason for this can be inferred from the energy-momentum spectrum (dispersion relation) for the excitations as shown in Fig. 8. In the superfluid phase there are two types of thermal excitation. Phonons are collective density oscillations of the liquid (sound waves) travelling at the speed of sound c_1, which is approximately constant at about

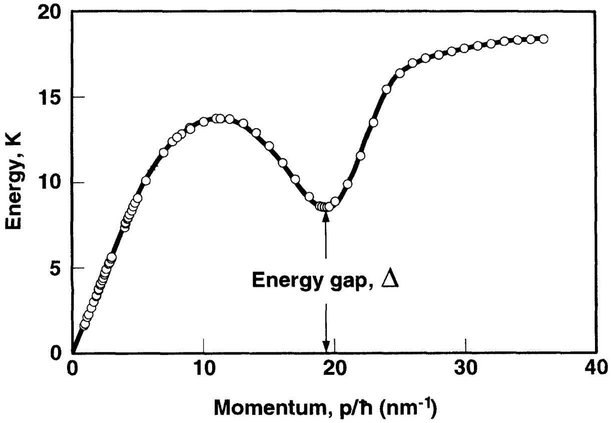

FIG. 8. The energy–momentum relationship (dispersion curve) for the excitations in superfluid ^{4}He. Phonons lie on the straight part of the curve near the origin, rotons around the minimum. The energy gap Δ to the minimum is indicated. (After Donnelly *et al.*, 1981.)

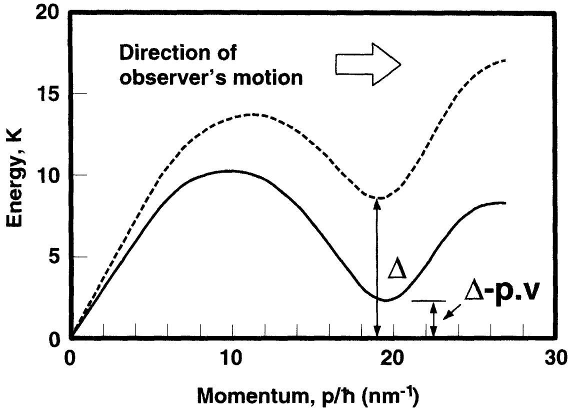

FIG. 9. The dispersion curve of Fig. 8 as seen in a reference frame moving at velocity $-\boldsymbol{v}$ relative to the superfluid. In this frame the roton minimum energy falls by $\boldsymbol{p} \cdot \boldsymbol{v}$. At the Landau velocity, when the minimum falls to zero energy, a scatterer moving at this velocity can freely create excitations.

$240\,\mathrm{m\,s^{-1}}$. Their dispersion relation is therefore approximately linear and described by the expression $E = c_1 p$. At higher energies and momenta the phonon excitations are gradually transformed into excitations called rotons, which have a minimum in the dispersion curve at an energy Δ and at a finite momentum p_0. In consequence, rotons that have momenta less than p_0 have a group velocity ($v_g = \mathrm{d}E/\mathrm{d}p$) which is in the opposite direction to their momenta; in other words they are "hole"-like excitations. Rotons with momenta greater than p_0 have group velocities in the same direction as their momenta and may be thought of as "particle"-like excitations. While the dispersion relation for rotons is well known, their microscopic nature remains a matter of conjecture.

The dispersion curve shown in Fig. 8 is valid in the stationary reference frame of the superfluid. Let us assume we have flow and that the superfluid is moving with velocity v_s with respect to some boundary wall. To a stationary observer in the rest frame of the wall the excitation energies will be shifted by an amount $-\boldsymbol{p} \cdot \boldsymbol{v}_s$ in accordance with the appropriate Galilean transformation. In other words, excitations with momenta approaching the observer, looking into the flow direction, are seen to have a higher energy and those retreating a lower energy. The resulting dispersion curve becomes canted as shown in Fig. 9. Since any scattering process must conserve energy in the rest frame of the scatterer (in this case the wall), new excitations can only be created at zero energy. As the figure shows, there are no such possibilities until the superflow exceeds the value $v_L = \Delta/p_0 \simeq 50\,\mathrm{ms^{-1}}$, at which velocity the dispersion-curve minimum reaches zero energy and excitations may be freely created. This process clearly extracts energy from the flow, *i.e.*, is dissipative. The velocity at which this process begins is known as the Landau critical velocity (Landau, 1941). From these arguments, superfluid ^{4}He should exhibit nondissipative superflow for velocities less than the critical value. It is interesting to note here that the dispersion curve for an ideal Bose gas follows the classical relation $E = p^2/2m$. In this case the Landau critical velocity is zero and there would be no regime of zero viscosity.

In practice it is found that the critical flow velocity for dissipation is very much smaller than the Landau velocity. Measured critical velocities vary according to the precise experimental geometry, but are generally of order $10\text{–}100\,\mathrm{mm\,s^{-1}}$, or some three orders of magnitude less than the Landau value. These velocities correspond to those needed to generate quantized vortices in the superfluid (superfluid turbulence), which we describe in Sec. 2.9.

2.7 Persistent Currents

A firm experimental test of nondissipative flow is given by the demonstration of persistent currents. Persistent currents have been observed in "superfluid gyroscopes" (Clow and Reppy, 1972) consisting of a toroidal flow channel packed with fine powder. The powder is required to clamp the normal component and to prevent the motion of vortices, which would otherwise lead to mutual friction between the superfluid and normal components and thus to dissipation of the superflow. The superfluid is set into motion by rotating the torus about its axis of symmetry above a certain critical angular velocity. The torus is then held stationary while the superfluid continues to circulate around the torus with negligible dissipation. The superflow can be detected by tipping the torus, which induces a precession of the symmetry axis by the gyroscope effect. In such gyroscopes supercurrents have been observed to persist without noticeable loss for several months, demonstrating that the viscosity of the superfluid component is immeasurably small.

2.8 The Wave Function and Phase Coherence

As discussed in Sec. 2.5, the ^{4}He superfluid component corresponds to the macroscopic occupation of the ground-state energy level in the fluid. All particles contributing to the superfluid component can thus be described by a single wave function of the form $\psi = \psi_0 \exp(i\theta)$. The magnitude of the wave function is determined by the density of the superfluid component, $\psi_0 = \sqrt{\rho_s}$. The value of the phase θ of the wave function is arbitrary; however, gradients of this phase determine the velocity of the superfluid component via the relationship $\hbar\nabla\theta = m_4 v_s$ where m_4 is the mass of a ^{4}He atom. The fact that a single phase factor describes all the atoms associated with the superfluid component is known as phase coherence.

2.9 Vortices

The existence of quantized vortex lines in superfluid ^{4}He (and the analogous flux lines in superconductors) is a direct consequence of the existence of a single macroscopic wave function describing the entire sample. Since a gradient in the phase of the wave function implies a flow of the superfluid, which must have some kinetic energy, the most energetically favorable configuration for a given volume of superfluid will be one with a slowly varying phase (small superflows) throughout; and for the special case of a stationary sample, the phase should be constant everywhere. However, in practice it is very difficult to realize this ideal configuration in any given experiment.

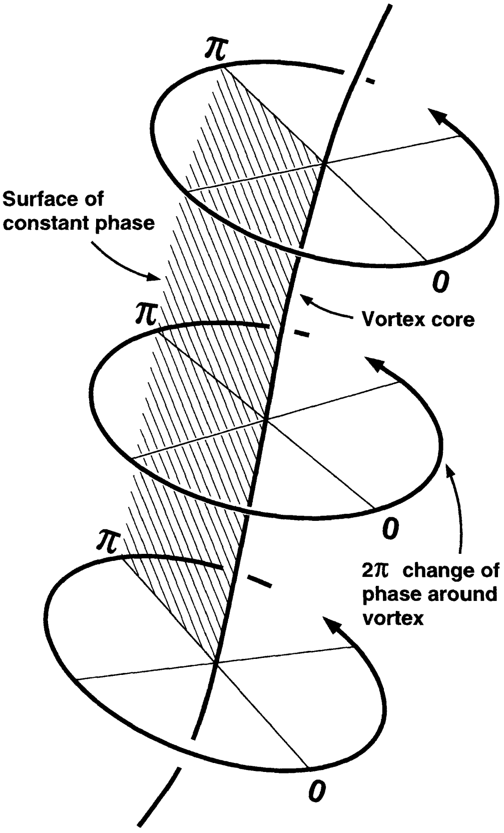

FIG. 10. The 2π change of the phase of the wave function around a vortex. Along the axis there must be a singularity where the phase cannot be defined.

Since the wave function must be single-valued throughout the superfluid, the change in the phase around any closed path has to be an integer multiple of 2π. If we imagine such a path with a phase change of 2π around it, then we can shrink the path arbitrarily while still maintaining the 2π phase change, until the path becomes a single point. From Fig. 10 we see that there must be a line of such points through the liquid defining a linear topological defect. This represents the core of a vortex. The 2π phase change around the core leads to a circulating superflow with a velocity described by the expression

$$v_s = \hbar/m_4 r \tag{2}$$

where r is the distance from the core. The velocity thus decreases inversely with distance. The circulation of liquid, defined as $\kappa = \int v_s \cdot dl$ around a loop surrounding the core, is thus quantized in units of h/m_4.

We now appear to face a contradiction since, first, the phase of the wave function cannot be defined along the actual core and, second, adjacent to the line the gradient of the phase approaches infinity implying an infinite velocity of the superfluid. There is only one way of avoiding these contradictions: the magnitude of the wave function must vanish at the core of the vortex; *i.e.*, the core comprises normal fluid. The length scale over which the wave function is suppressed at the vortex core is called the coherence length and is of the order of an interatomic spacing. Since the velocity field extends to infinite distance from the core these vortices do not have a unique energy per unit length since the kinetic energy stored in the flow field depends on the extent of the container.

It is worth noting at this point that similar difficulties with infinite velocities are a feature of most types of vortex and are invariably resolved in the same way by the exclusion of the motion from the axis. For example, the vortex formed by water running out of the plughole of a bath has a cylindrical air core, which excludes the velocity field of the water from the axis.

2.9.1 Rotating Superfluid ^{4}He

If the wave function of the superfluid is continuous (*i.e.*, contains no voids or vortex cores) then the change in phase of the wave function around any loop must be zero. This means that the superflow is irrotational ($\nabla \times \boldsymbol{v}_s = 0$). This has the important consequence that solid-body rotation of the superfluid is impossible. Furthermore, in the absence of vortices, the net rotation of the superfluid must be zero with respect to the fixed stars. If a bucket of normal liquid ^{4}He is set into rotation (with sufficiently small angular velocity) and then cooled below the lambda transition, the superfluid component will be stationary. This could in principle be seen from the shape of the free liquid surface. the surface would take up a parabolic profile in the normal state in response to the centrifugal force acting on the rotating helium. However as the liquid is cooled well below the transition temperature the surface should become flat since the superfluid component cannot rotate.

As the rotation speed of the container is increased, then eventually the relative velocity between the stationary superfluid and the moving walls will exceed the critical velocity for vortex creation. A small loop of vortex is then created which is immediately dragged around by the friction force of the rotating normal fluid. Here the Magnus force on the vortex from the moving superfluid comes into play. (The Magnus force is the sideways force acting on a rotating cylinder moving transversely in a fluid). The Magnus force causes the vortex line to grow outwards from the container wall and eventually take up a position along the rotation axis of the container as illustrated in Fig. 11. As the rotation is further increased, more vortex lines are produced, which move towards the center of the cell forming a regular lattice of vortex lines parallel to the rotation axis. Photographic images of the vortex line lattice have been produced by Yarmchuk *et al.* (1979). The vortex lattice is analogous to the flux-line lattice produced by a magnetic field applied to a type II superconductor. In this sense, rotation applied to a superfluid is exactly analogous to a magnetic

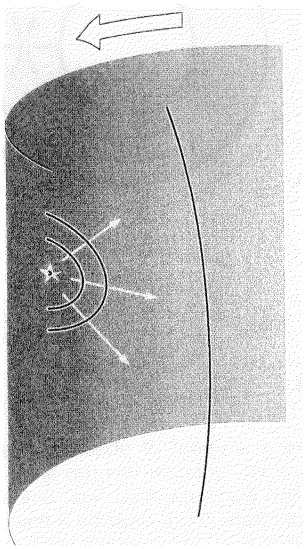

FIG. 11. In a rotating container, microscopic vortex loops created at the moving wall grow, under the influence of the Magnus force, and take up a position parallel to the axis of rotation.

field applied to a superconductor, since in the former case the rotation is quantized (circulation of mass) and in the latter flux is quantized (circulation of charge). The analogy may be taken further since another important property of a superconductor is the expulsion of magnetic field (the Meissner effect), which is analogous to the expulsion of solid-body rotation in a superfluid.

2.9.2 Measurement of Absolute Rotation with Superfluid ^{4}He Since superfluid ^{4}He is irrotational with respect to the fixed stars, a measurement of the superfluid rotation with respect to its container constitutes a measurement of the rotation of the container with respect to the fixed stars. In this manner, superfluid ^{4}He can be used to measure absolute rotation. This principle has already been tested by Schwab *et al.* (1997), who have recently measured the rotation of the Earth. The device they used is shown schematically in Fig. 12. A toroidal container incorporating a small aperture is filled with superfluid ^{4}He. The container is stationary with respect to the Earth and thus rotating with respect to the fixed stars. Since the superfluid cannot rotate, there must be a superflow with respect to the container which is magnified as the superfluid flows through the small aperture. The flow through the aperture may be measured, from which the rotation of the container (the Earth) is inferred. The device shown in Fig. 12 is actually the superfluid analog of a superconducting SQUID, which measures magnetic fields in a similar manner.

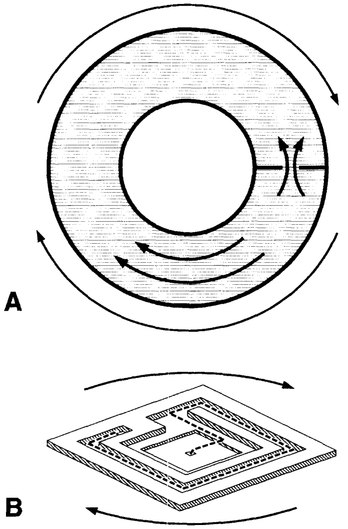

FIG. 12. Measurement of absolute rotation. (a) A rotating toroidal container incorporating a small aperture is filled with superfluid ^{4}He. Since the superfluid cannot rotate, there must be a flow through the aperture. This flow can be measured and the rotation of the toroid inferred. (b) In the actual device, the channel is formed in a thin slab. The flow loop, indicated by the dashed line, passes through the small aperture sited at the center. (After Schwab *et al.*, 1997.)

3. SUPERFLUID ^{3}He

Superfluid ^{3}He was first observed by Osheroff, Richardson, and Lee in 1971 (Osheroff *et al.*, 1972). Although the superfluid transition in ^{3}He had long been predicted, the observation of the superfluid transition was at first mistakenly associated with the solid phase. However, soon afterwards, the experiment was repeated with the use of nuclear magnetic resonance (NMR) techniques with spatial resolution that allowed the identification of two superfluid phases, designated the A phase and the B phase.

The arguments of Bardeen, Cooper, and Schrieffer for Cooper pairing discussed in Sec. 1.3 above were originally proposed in the context of superconductivity. The Cooper pairs that form in conventional superconductors are relatively simple, having zero spin, zero orbital angular momentum, and therefore no internal structure. (The form of the pairs in high-T_c materials is still under debate at the time of writing.) In contrast, the Cooper pairs in superfluid ^{3}He have spin unity and orbital angular momentum unity. The ^{3}He pairs thus have a more complex internal structure, which has a dramatic effect on the superfluid properties, not least in ensuring that there are several distinct superfluid phases.

3.1 Cooling Techniques

The superfluid transition temperature for ^{3}He varies between about 1 and 3 mK depending on the pressure. Even with current cooling techniques these temperatures are not straightforward to attain. The most usual methods in use rely on a continuously operating dilution refrigerator to produce a steady temperature in the millikelvin regime from which base a further cooling stage to reach the superfluid region is used. For this final stage nuclear cooling is currently almost universally used, but in the past Pomeranchuk cooling (see Sec. 3.1.1) has also been used successfully. A dilution refrigerator may be obtained commercially. However, while a modern commercial dilution refrigerator will typically have a base temperature of around 5 mK this is not as good as the best in-house produced machines, which currently can reach around 2 mK.

3.1.1 Pomeranchuk Cooling During the early 1970s before nuclear cooling had become established, the only other practical methods for cooling ^{3}He to low millikelvin temperatures were by the Pomeranchuk effect and, for work just reaching the superfluid transition, magnetic cooling using cerium magnesium nitrate. In fact Pomeranchuk cooling was the method employed by Osheroff *et al.* (1972) when in 1971 they first observed superfluidity in ^{3}He. The technique relies on an unusual property of the ^{3}He phase diagram shown in Fig. 13. At low temperatures, the pressure at which solidification occurs in liquid ^{3}He falls as the temperature rises. This follows from the unusual property of ^{3}He that the entropy of the solid phase is actually higher than that of the liquid phase below 0.3 K. In other words, the solid phase in more disordered than the liquid phase. While Fermi statistics ensure that the entropy of the liquid is low, the solid has almost complete disorder of the nuclear spin system above 1 mK. Consequently, an increase in pressure that causes solidification of the liquid will produce cooling from the latent heat of solidification, *i.e.*, will force the liquid–solid mixture to follow the solidification line to lower temperatures as the pressure is increased. This is the principle utilized in Pomeranchuk cooling.

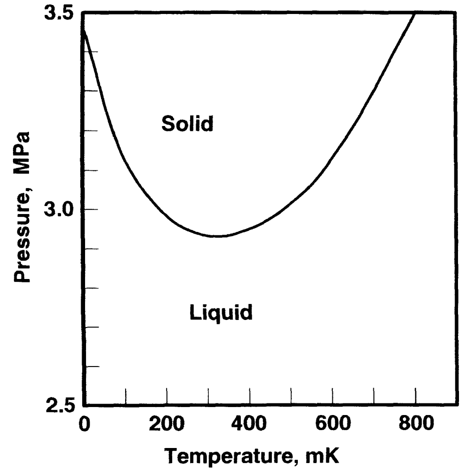

FIG. 13. The liquid–solid region of the ^{3}He phase diagram.

Pomeranchuk cooling has the advantage that the ^{3}He is itself the refrigerant, thus overcoming any problems of establishing thermal contact to the specimen at low temperatures. The disadvantages of this technique are

1. that the specimen will always contain a mixture of solid and liquid ^{3}He;
2. that the specimen is always at melting pressure, thus preventing studies at lower pressures; and
3. that the minimum temperature that can be achieved is limited by the magnetic ordering temperature of the solid (about 1 mK).

These limitations have meant that the Pomeranchuk technique has largely been replaced as a cooling method by the far more versatile technique of nuclear cooling.

3.1.2 Nuclear Cooling The principle of nuclear cooling is shown schematically in Fig. 14. A material containing weakly interacting spins (magnetic moments) is placed in a high magnetic field and then precooled to some temperature T_i. The high magnetic field B_i produces Zeeman splitting of the magnetic energy levels, and the precooling ensures a relatively high population of the lower energy levels as shown in the figure. The magnetic entropy of the system depends only on the relative populations of the energy levels, which in turn depend only on the ratio $\mu B_i/kT_i$ where μ is the magnetic moment associated with the spin. Now suppose that we can reduce the magnetic field adiabatically, *i.e.*, such that no thermal energy can be exchanged between the outside world and the spin system. Since no heat is exchanged the populations of the various levels must remain the same. Therefore, during "adiabatic demagnetization" the entropy of the spin system must remain constant,

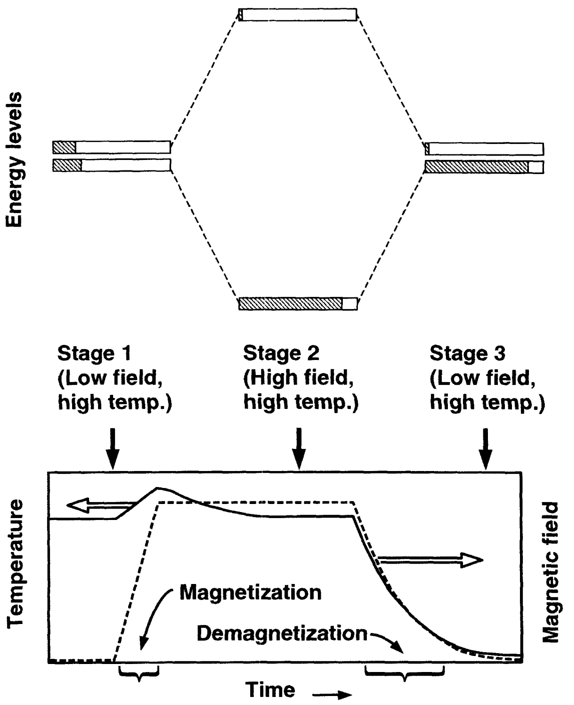

FIG. 14. Nuclear cooling. At the top is a schematic illustration of the energy levels and populations as the field is applied and removed. Below is shown schematically the variations of the magnetic field and the temperature during the process.

ensuring that the ratio $\mu B/kT$ is also constant. Therefore a reduction of the field to some final value B_f results in a lower temperature T_f as given by:

$$T_f = (B_f/B_i)T_i.$$

In principle the technique of nuclear cooling could not be simpler. When the field applied to a material containing nuclear magnetic moments is reduced the temperature falls in the same ratio. In practice, of course, it is not so straightforward.

The minimum temperature possible with nuclear cooling is ultimately limited to the temperature at which the nuclear spins spontaneously order as a result of the interactions between spins, which we neglected above. However, in the case of copper, the most widely used nuclear refrigerant, this ordering temperature lies in the nanokelvin regime and does not present any drawback for cooling liquid ^{3}He to the millikelvin and microkelvin regimes. In practice the minimum ^{3}He temperature achievable by this method is limited by a combination of heat leaks into the specimen and the very poor thermal contact between the specimen and the nuclear coolant. Since the nuclear heat capacity of copper is not very large compared with that of liquid helium, a volume of copper comparable to that of the liquid helium specimen is necessary to generate enough cooling power. Typical starting conditions for demagnetization are a starting temperature of 5–10 mK (well within the range of a dilution refrigerator) in a field of 6–8 T (readily generated by a commercial superconducting solenoid). By reducing the magnetic field by a factor of 1000, temperatures around 100 μK are attainable in the superfluid ^{3}He. This is already low enough that to achieve significantly lower temperatures we may need to consider shielding the specimen against the heating from cosmic rays. A practical system is shown schematically in Fig. 15.

3.2 A Multiplicity of Possible Phases

Since the Cooper pairs in superfluid ^{3}He have spin angular momentum $\boldsymbol{S} = 1$ and orbital angular momentum $\boldsymbol{L} = 1$, they have in-

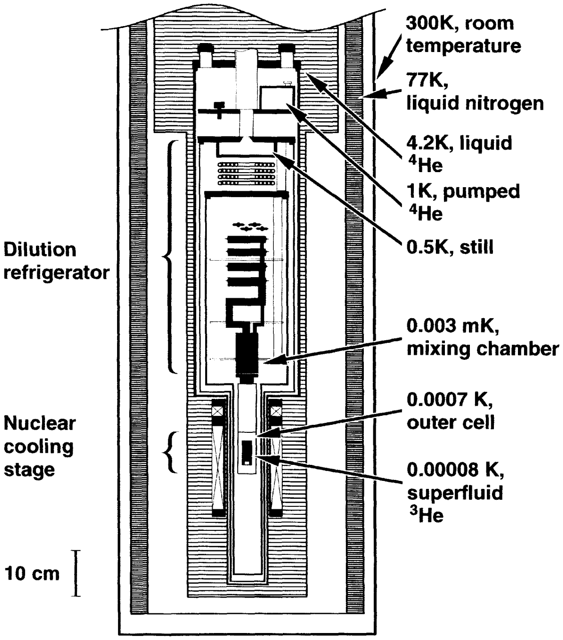

FIG. 15. A nuclear cooling cryostat for achieving superfluid temperatures in liquid ^{3}He.

ternal structure in the sense that there are two axes defined by the directions of the spin and orbital angular momenta. The internal degrees of freedom are specified by the spin projection quantum numbers $m_S = +1, 0, -1$ and orbital angular momentum projection $m_L = +1, 0, -1$. Therefore there are $3 \times 3 = 9$ different types of Cooper pair allowed in superfluid ^{3}He, and correspondingly the superfluid wave function has 9 complex components (18 parameters). This complex structure places superfluid ^{3}He in a class of its own. In contrast, superfluid ^{4}He can be described by a simple two-parameter wave function $\psi = \psi_0 \exp(i\theta)$ (see Sec. 2.8), which is also the case for conventional superconductors where the Cooper pairs have zero spin and zero orbital angular momentum and therefore have no internal structure.

One immediate consequence of the multiplicity of possible Cooper pair states in superfluid ^{3}He is the possibility of several distinct superfluid phases. It is a tribute to modern theoretical physics that several of the possible phases of systems with $\boldsymbol{S} = 1$ and $\boldsymbol{L} = 1$ Cooper pairs had been studied theoretically (albeit in the context of superconductivity) long before the discovery of superfluid ^{3}He. There are 13 possible phases but only three occur under normal experimental conditions. Which phase is preferred depends on temperature, pressure, and magnetic field. The phase diagram for the two phases that occur in zero magnetic field is shown in Fig. 16.

3.2.1 The A Phase

As shown in Fig. 16, the preferred phase at high temperatures and pressures is known as the *A* phase. The Cooper-pair structure of this phase was first considered theoretically by Anderson, Brinkman, and Morel (ABM) (Anderson and Morel, 1961; Anderson and Brinkman, 1973). In this phase the Cooper-pair spin projection may only take the values $m_S = +1$ and $m_S = -1$. Furthermore, the orbital angular momentum of every pair points along a common direction called the orbital angular momentum vector (or $\boldsymbol{l}$ vector). The pairs are constructed from ^{3}He atoms close to the Fermi surface. If the component atoms are to couple to form pairs with angular momentum $\boldsymbol{L} = 1$ pointing in, say, the polar direction, then atoms with momenta lying in the equatorial plane can achieve this easily by mutually orbiting but atoms with momenta in the polar direction cannot mutually orbit to create an angular momentum also in this direction. In consequence this phase

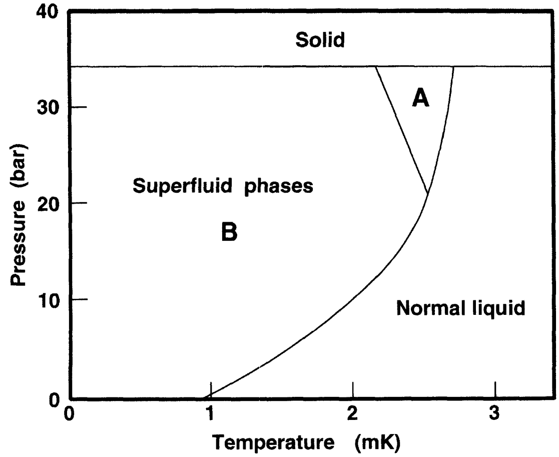

FIG. 16. Phase diagram of the superfluid phases of ^{3}He in zero magnetic field.

acts as an excellent superfluid for atoms with momenta in the equatorial plane but cannot at all form a superfluid for atoms with momenta along the polar direction. The superfluid energy gap Δ, which tells us the strength of the pairing, is thus anisotropic, vanishing along the $\boldsymbol{l}$ axis as shown in Fig. 17.

Since the pairs all orbit about a common axis, defined by the direction of the $\boldsymbol{l}$ vector, the nuclear magnetic moment associated with each ^{3}He nucleus experiences a changing magnetic field from its partner. This is equivalent to two parallel bar magnets orbiting about a common center. The magnetic (dipole–dipole) energy of such a system is reduced when the magnets are oriented perpendicular to the rotation axis. Therefore, in the absence of any other disturbing features such as walls or flow, this will align the spin perpendicular to the $\boldsymbol{l}$ vector.

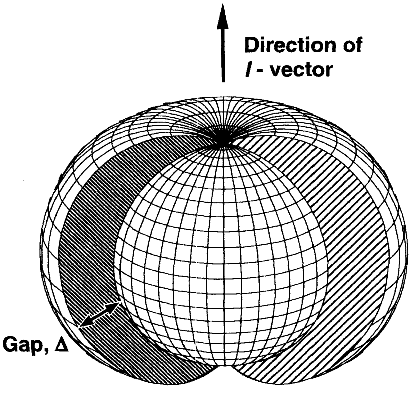

FIG. 17. The energy gap in the *A* phase. The gap is zero along the *l*-vector axis and is at its maximum in the perpendicular directions.

For an isolated pair this energy is miniscule and would not cause any orientation at millikelvin temperatures. However, all the pairs are in the same quantum ground state acting as a single entity such that all the spin vectors acting together are orienting themselves relative to all the orbital moments acting together. The net result is a greatly enhanced spin–orbit coupling. This demonstrates that indeed we are dealing with a coherent ground state.

3.2.2 The B Phase The structure of the *B* phase of superfluid ^{3}He is somewhat more subtle than that of the A phase. In low magnetic fields the system shows a superposition of all possible m_S values and all possible m_L values, with the result that there is no net preferred spin or angular momentum direction in the liquid. As a consequence of this apparent higher degree of symmetry, the B-phase energy gap is isotropic. To first order, each pair attempts to make its total angular momentum (spin plus orbital) zero, $\boldsymbol{J} = \boldsymbol{S} + \boldsymbol{L} = 0$. Therefore pairs consisting of ^{3}He atoms with momenta parallel to the field are in the $m_S = 0$ state. Those formed from atoms with momenta perpendicular to the field have $m_S = +1$ and -1, with all three states contributing for other angles. However, the same magnetic dipole forces that orient the A phase also act here, forcing the spin direction to be rotated with respect to the orbital direction. However, unlike in the A phase where the total spin system of the pairs is rotated with respect to the total orbital momentum, in the B phase the rotation takes place for each pair individually, but about an axis common for all pairs, specified by a vector ***n***. The dipole–dipole coupling energy is minimized when this rotation angle is 104° (known as the Leggett angle). This leads to quite a subtle structure but does have the effect of producing a simple isotropic energy gap, at least in low magnetic fields.

The B phase corresponds to the theoretical BW state first examined by Balian and Werthamer (1963), long before this behavior was realized in superfluid ^{3}He. The original BW state had Cooper pairs with total angular momentum $\boldsymbol{J} = \boldsymbol{L} + \boldsymbol{S} = 0$, but the relative population of the m_S states was distributed around the Fermi surface as described above.

3.3 Broken Symmetries of the Superfluid Phases

3.3.1 The A Phase In Sec. 3.3.1 we introduced the ***l*** vector, which defines the axis along which the energy gap vanishes in the A phase. In the absence of any external fields and "foreign" objects such as container walls, it is clear that the direction of ***l*** must be arbitrary. Above the transition the direction of the ***l*** vector has no meaning. Thus as we pass through the transition the symmetry of the liquid is reduced, as there is now a definite (but arbitrary) direction for ***l*** whereas above the transition all directions were equivalent. This behavior is known as spontaneous symmetry breaking. Conventional superconductors and superfluid ^{4}He also exhibit spontaneous symmetry breaking, although here a more abstract symmetry is broken, known as gauge symmetry, which leads to the definite (but again arbitrary) phase factor in the superfluid/superconductor wave function. Since spatial gradients of the ***l*** vector give rise to superflows in a similar manner as spatial gradients of the phase factor give rise to superflows in superfluid ^{4}He, the broken gauge and orbital symmetries are linked in the A phase. This is described as a broken relative gauge-orbit symmetry. In addition, the directionality of the spin part of the wave function leads to broken spin rotation symmetry in ^{3}He-A. The spin part of the wave function in the A phase is described by a vector ***d***, which is defined to point along the direction of zero spin projection. In other words the ***d*** vector points perpendicular to the spin ***S*** and is common for all pairs.

3.3.2 The B Phase In superfluid ^{3}He-B gauge symmetry is broken (*i.e.*, the phase of the wave function takes a particular value). As discussed in Sec. 3.2.2, there is no independent directionality for the spin and orbital parts of the wave function as in the A phase. However, the orbital and spin vectors are coupled for each pair such that all the spin vectors may be rotated with respect to all the orbital vectors around a common axis specified by the vector ***n***. Thus all the directionality of the B phase (in low magnetic fields) is specified by this spin–orbit rotation axis and therefore the B phase exhibits a broken relative spin–orbit rotation symmetry.

3.4 The Phase Diagram

The phase diagram of superfluid ^{3}He has already been introduced in Fig. 16. In low magnetic fields, the B phase covers the majority of the *P-T* space while the A phase is confined to a region of high pressure and temperature. The phase transition between the normal state and the superfluid state is second order with a discontinuity in the heat capacity (in this sense very similar to conventional superconductors).

The phase transition between the A and B phases is, however, first order, and very substantial supercooling of the A phase has been observed (Hakonen *et al.*, 1985). The mechanism for the nucleation of the B phase from the A phase has been a matter of much debate. The first theoretical estimates of the nucleation time for the B phase suggested that it should be of order the age of the universe. The precise nucleation mechanism is as yet still uncertain; however, there is strong evidence that cosmic rays may be playing an important role (Leggett, 1984).

The phase diagram is transformed by the application of a magnetic field. Another phase, called the A1 phase, appears in a narrow temperature range separating the normal state from the A phase. The A phase becomes stable over a wider range of temperatures and pressures. For magnetic fields larger than about 0.5 T, the B phase is removed completely from the phase diagram, and the A phase becomes the low-temperature superfluid phase. The A1 phase is very similar to the A phase, the main difference being that the A1 phase contains only $m_S = +1$ (up spin) pairs; that is, the A1 phase is a magnetic superfluid. The transition from the A1 phase to the A phase simply corresponds to the phase transition for the $m_S = -1$ (down spin) Cooper pairs. The A phase is often considered as consisting of two interpenetrating superfluids, one for the down-spin pairs and one for the up-spin pairs.

3.5 Properties of Superfluid ^{3}He

Not surprisingly, the high level of structure of the Cooper pairs in superfluid ^{3}He leads to rather unusual properties. In spite of the complexity of the superfluid wave function, many of the superfluid properties were found theoretically prior to their discovery. The reasons for this are that

1. the basic interactions between ^{3}He atoms are known rather precisely;
2. the entire superfluid sample (typically consisting of some 10^{23} atoms) is described by a single, albeit rather complicated, wave function with a well defined structure in each superfluid phase; and
3. there are virtually no impurities in liquid ^{3}He at low temperatures.

The only element that has appreciable solubility in ^{3}He is ^{4}He, and the extrapolated solubility of ^{4}He at sub-millikelvin temperatures is around one ^{4}He atom in an astronomical volume of liquid ^{3}He. Superfluid ^{3}He is therefore an exceptionally clean medium, for which theories should be applicable with extreme accuracy.

Nevertheless, there are still new phenomena being found in superfluid ^{3}He that have, as yet, no theoretical explanation. The reason for this is that the system is so complex that even though the fundamental equations are known, all the solutions are not.

3.5.1 Thermal Excitations and Superflow

In order to describe the thermal excitations in the superfluid phases, it is first necessary to describe those in the normal phase of liquid ^{3}He. The normal phase is described well by Landau's phenomonological Fermi-liquid theory (Landau, 1956). Here, the main effect of the interatomic forces can be simply accounted for by introducing an effective ^{3}He mass, m^*. The modified atoms, called quasiparticles, can then be considered to be almost noninteracting, with energy levels given by $E = p^2/2m^*$. The main thermal excitations of the system correspond to either the occupation of energy levels above the Fermi energy (quasiparticle excitations) or empty energy levels below the Fermi surface (quasihole excitations). The excitation energies, given by $\xi = |p^2/2m^* - E_F|$, are shown by the dashed line in Fig. 18.

At absolute zero temperature, all the ^{3}He quasiparticles are paired and can be described by the superfluid wave function. At finite temperatures, some of these pairs will be broken by thermal agitation. The unpaired quasiparticle excitations are still influenced by the pairing interaction and their energies are raised with respect to the normal-state excitations as shown by the solid line in Fig. 18. The minimum quasiparticle excitation energy is equal to the superfluid energy gap Δ, as shown in the figure. The Cooper pairs may be considered to lie at the origin in Fig. 18, and in order to break a pair, two excitations must be created, requiring a minimum of 2Δ of energy.

The arguments presented in Sec. 2.6 relating dissipationless flow with the excitation dispersion curve for superfluid ^{4}He are equally applicable to superfluid ^{3}He. The dispersion curve shown in Fig. 18 will clearly allow the existence of dissipationless superflow below the Landau critical velocity $v_L = \Delta/p_F$. In the B phase the measured critical velocity of the superfluid is indeed limited by the Landau velocity even in the presence of strong magnetic fields (Fisher *et al.*, 1991). (This is in contrast to the case of superfluid ^{4}He discussed in Sec. 2.6.) In the A phase the situation is far more complicated as a result of the extreme anisotropy of the energy gap. The Landau critical velocity for the A phase is dependent on direction, falling to zero for superflow along the **l** axis. The critical flow velocity for the A phase is in practice dependent on geo-

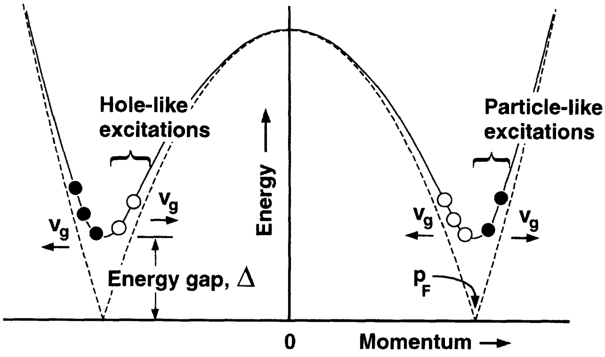

FIG. 18. The energy–momentum relations (dispersion curves) for the excitations in superfluid (solid line) and normal liquid ^{3}He (dashed line). The solid circles represent quasiparticles and the open circles quasiholes. The Fermi momentum p_F is indicated. The excitation energies plotted are measured from the Fermi energy.

metrical factors on account of the ***l*** textures which we describe in Sec. 3.5.3.

3.5.2 Thermal Properties of Superfluid ^{3}He The heat capacity of superfluid ^{3}He as measured by Greywall (1986) is shown in Fig. 19. The behavior at the transition is quite unlike that of ^{4}He. There is a clear second-order phase transition between the normal phase and the superfluid phases with no logarithmic singularity. The B-phase heat capacity is very similar to that of a conventional superconductor and is well described by the BCS theory with a slightly enhanced energy gap due to strong coupling corrections (the BCS theory is only applicable in the limit of weak pairing interactions). At low temperatures the heat capacity of the B phase falls exponentially with decreasing temperature because of the dominance of the Boltzmann factor $\exp(-\Delta/kT)$ governing the number of thermal quasiparticle excitations. The extremely small heat capacity of ^{3}He-B at temperatures of around 100 μK has been utilized in superfluid-^{3}He particle detectors (Bradley *et al.*, 1995) capable of measuring energy depositions of less than 1 keV.

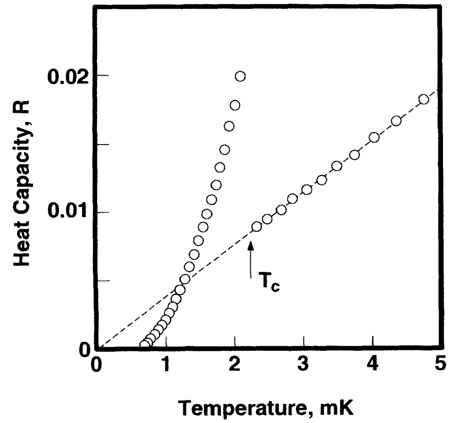

FIG. 19. The heat capacity of superfluid ^{3}He at 20.3 bar (after Greywall, 1986).

In contrast, the heat capacity of ^{3}He-A is expected to vary as T^3 at low temperatures on account of the nodes in the energy gap (see Sec. 3.3.1). However, as a significant field is needed to stabilize the A phase to low temperatures this behavior has not been experimentally verified.

The thermal transport properties of the superfluid phases are qualitatively similar to those of superfluid ^{4}He in that at high temperatures a counterflow of normal fluid and superfluid is set up by a temperature gradient and the thermal conductivity is very high. The situation is particularly interesting at very low temperatures where the normal-fluid density is very low, since the quasiparticle excitations making up the normal fluid travel ballistically. The theoretical mean free path of a quasiparticle in ^{3}He-B at a temperature of around $0.1\,T_c$ (the lowest temperature presently achievable) is about 1 km. This property has been utilized in quasiparticle beam experiments (Fisher *et al.*, 1992).

3.5.3 Textures The directional properties of the superfluid give rise to properties very reminiscent of those seen in liquid crystals. The directions of the ***l*** and ***d*** vectors in the A phase and the ***n*** vector in the B phase, which determine the dynamic behavior of the phases, are influenced both by external fields and by interfaces (including solid boundaries and superfluid phase boundaries). The arrangement of the various directionalities in a given situation is known as the texture.

In the A phase, a solid–superfluid boundary forces the ***l*** vector to lie normal to the surface. Put in simple terms, the pairs at the interface may orbit in the plane of the surface, whereas other directions will involve collisions of the component ^{3}He atoms with the boundary. In order to minimize bending energies, a slow variation in the direction of the ***l*** vector is preferable; however, in a simply connected volume this condition cannot be satisfied everywhere. There has to be at least one singularity (or textural defect). This is illustrated for a spherical cell in the first part of Fig. 20. If, on the other hand, we confine the A phase in a torus then the singularity can be avoided, as shown in the lower part of the figure. Since the spin direction can be influenced by a magnetic field, there is an interesting region near walls where there is a competition between the orienting effect of the wall on the orbital angular momentum and the orienting effect of the field on the spin angular momentum.

The effect of a solid boundary in the B phase is more complex. In the absence of walls the ***n*** vector prefers to lie parallel to the field. However, if we have a wall nearby, then the direction of ***n*** depends on the orientations of both the field and the wall.

Other orienting effects include electric fields, thermal conduction, and superflow. The combination of all these orienting effects in practice leads to very messy textures with many defects (somewhat in analogy with the structure of a deformed polycrystalline solid).

Textural effects in the A phase are obviously very strong as a result of the large anisotropy of the energy gap. In consequence virtually all the dynamic properties of the A phase (transport properties and NMR properties) are strongly affected by the texture. For example, at the lowest temperatures virtually all the quasiparticle excitations of the A phase have momenta along the ***l*** axis (where the energy gap is zero), so that in this sense the excitation gas is almost one dimensional. Spatial variations of the direction of ***l*** lead to confinement of the excitations, thus reducing all excitation transport properties by an enormous factor.

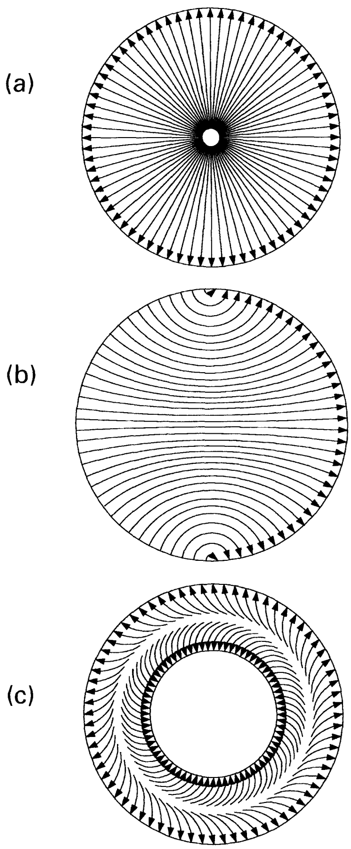

FIG. 20. (a), (b) The A-phase texture (the direction of the *l*-vector) in a spherical volume. As can be seen, for the direction of *l* to vary only slowly over the volume there must be either (a) a point singularity in the bulk or (b) a line singularity at the boundary. (c) The texture in a toroidal container can be accommodated with no singularities.

For low magnetic fields, the textures in the B phase are far weaker as far as excitations are concerned since the energy gap is isotropic. In higher magnetic fields (above 0.1 T, say) the energy gap becomes distorted thus strengthening the influence of the texture. The B-phase texture does, however, have an enormous effect on the NMR properties even in small magnetic fields (see Sec. 3.5.5.1).

3.5.4 Vortices The complexity of the superfluid ^{3}He wave function also reveals itself in the variety of superfluid ^{3}He vortices, which have been methodically studied and categorized by the Helsinki ultralow-temperature group (Krusius, 1993).

There is only one type of vortex possible in superfluid ^{4}He because of the simplicity of the wave function (which reflects the simplicity of the ordering). Put simply, in superfluid ^{4}He the phase of the wave function is allowed to change by a multiple of 2π around a vortex core. In superfluid ^{3}He, however, there are several properties that may change as we circle the core. The phase may change, but also the directions of the orbital and spin angular momentum vectors may change, and since the latter are coupled the number of possible structures is quite large.

The circulation κ of vortices in superfluid ^{4}He is quantized in units of h/m_4, as discussed in Sec. 2.9. Similarly, in superfluid ^{3}He vortices are quantized in units of $h/2m_3$ where m_3 is the bare ^{3}He atomic mass, the reason being that the superfluid is composed of Cooper-pair bosons consisting of two ^{3}He atoms (analogously, the flux carried by a flux-line vortex in a superconductor is quantized in units of $h/2e$).

In the B phase, the two most common types of vortex are the axisymmetric (V1) vortex and the nonaxisymmetric (V2 or dumbbell) vortex. Both of these are hard-core vortices, meaning that the structure of the wave function is modified within the vortex core over a length scale of the order of the superfluid coherence length $\xi \approx 10\text{–}100\,\text{nm}$. The V1 vortex occurs at higher temperatures and carries a single quantum of circulation. The B phase is gradually transformed into the A phase on approaching the vortex axis. The V2 vortex has an exotic double-core structure consisting of two topologically confined half-quantum vortex cores, so that the core carries, in total, a single quantum of circulation. Single half-quantum vortices are topologically forbidden; their confinement within the V2 core is analogous to the confinement of quarks within nucleons.

The predominant vortices in the A phase are soft-core vortices, meaning that the structure of the superfluid is unchanged in the vortex core. They are in fact textural vortices in the sense that they are defined by the $\boldsymbol{l}$ texture (as mentioned in Sec. 3.3.1, a spatially varying $\boldsymbol{l}$ texture produces superflow). The size of the soft core, typically a few micrometers, is much larger than that of hard cores, and they carry two quanta of circulation. Other vortex types in the A phase include hard-core vortices and, remarkably, a vortex sheet consisting of multiple soft-core vortices locked together side by side.

3.5.5 Magnetic Properties Since the Cooper pairs in superfluid ^{3}He have spin $S = 1$ arising from the nuclear spin of the constituent ^{3}He atoms, they have a magnetic moment. This is the only superfluid/superconducting system which is currently known to have magnetic Cooper pairs. The absence of electrical charge in the ^{3}He pairs makes certain types of measurements (such as the measurement of superflow) rather difficult as compared with a similar measurement in a superconductor. However, this is more than compensated by the magnetic moment of the pairs, which allows them to be probed using the powerful technique of NMR, which is unique in that the technique allows us to look very directly at the Cooper-pair wave function. A part of the wave function concerns the direction of the nuclear spin (or magnetic moment). Not only can we see what this part of the wave function is doing by NMR, but we can also use NMR to manipulate it. In consequence, new phenomena can be observed such as spin supercurrents, spin vortices and coherent excited states of the Cooper-pair condensate.

3.5.5.1 Nuclear Magnetic Resonance Historically, the discovery of superfluid ^{3}He was made using NMR techniques, and the A and B phases of the superfluid were identified and then associated with the theoretical ABM and BW states primarily because of their NMR signatures.

The magnetic susceptibility of normal liquid ^{3}He is temperature independent at millikelvin temperatures. Since the A-phase Cooper pairs are formed with parallel spins the A-phase susceptibility is identical to that of the normal state. However, as mentioned above in Sec. 3.2.1, the pairs have the lowest magnetic (dipole) energy when the orbital and spin directions are perpendicular. In NMR we apply a magnetic field and then with a perpendicular rf field tip the magnetization from the equilibrium direction and follow the precession by observing the voltage output of a surrounding pickup coil. In the normal phase, the spins precess with a frequency determined only by the field and the value of the nuclear moment. However, in the A phase, since the spin direction is influenced by the orbital moment, if we tip the spins from their equilibrium direction there is a restoring force from the Cooper pairs themselves trying to return the spin direction to the perpendicular state. This gives an extra restoring force above the "bare" magnetic value, which means that the precession frequency is greater than that for ^{3}He atoms in the normal liquid. This is a unique signature of the A phase. In contrast, the B phase shows no such effect in the absence of other textural orienting forces, and the NMR frequency is similar to that of the normal fluid. However the susceptibility falls rapidly with decreasing temperature close to T_c and reaches a low-temperature value equal to about a third of the normal-state susceptibility. This is because in the normal liquid all the atoms have spin $\frac{1}{2}$, whereas in the B-phase superfluid the pairs have $S = 1$, $S = -1$, and $S = 0$. The $S = 0$ pairs do not contribute to the susceptibility, which therefore falls with the increasing superfluid fraction.

The dynamic magnetic properties of the superfluid phases are more interesting. Since the direction of the spins in the pairs is influenced not only by the magnetic forces but also the dipole–dipole interaction from the influence of the orbital moment, the spin dynamics becomes very complex. First, unlike the situation with free spins, we see the phenomenon of longitudinal NMR in which there are oscillations of the magnetization parallel to the steady field. The extra dipole–dipole restoring forces also give rise to subtle feedback mechanisms that lock the whole precessing system into one coherent state, as discussed in the next section.

3.5.5.2 Coherent Excited States in ^{3}He-B Any wave function has the property that a spatial gradient in any of its quantities is associated with an appropriate flow. Such gradients are responsible for some of the most fundamental properties of superfluids and superconductors. For example, gradients in the phase θ of the wave function in superfluid ^{4}He are associated with dissipationless mass flow (see Sec. 2.8). This can be thought of as the operation of Lenz's law. In an attempt to nullify the gradient the superfluid responds with a supercurrent. In superfluid ^{3}He-B we can set up (by NMR) gradients across a sample in the spin part of the wave function, by arranging for the spins to precess at different frequencies in different parts of the liquid. In this case the flow set up is a flow of spin, a spin supercurrent. Spin supercurrents are damped by thermal excitations (or the normal-fluid component), but nevertheless play a vital role in the spin dynamics and can give rise to extraordinary phenomena at very low temperatures where the normal-fluid damping becomes negligible.

The first coherent excited state discovered in ^{3}He-B is called the homogeneous precession domain (HPD) (Borovik-Romanov *et al.*, 1984; Fomin, 1984). It can be excited by both pulsed and continuous-wave (cw) NMR methods. The most striking property of the HPD, as observed with pulsed NMR, is that it is a very long-lived signal even in the presence of large gradients of the external field. In fact, a magnetic field gradient is required in order to stabilize the HPD signals.

The mechanism behind the HPD formation is illustrated in Fig. 21. Before the NMR tipping pulse, the net magnetization (spin) induced by the vertical magnetic field H is vertical throughout the whole volume. On application of a 90° NMR pulse, the net magnetization (spin) is tipped into the horizontal plane and then precesses around the steady field direction at the Larmor frequency $f_L = \gamma H/h$, where γ is the gyromagnetic ratio for ^{3}He and h is Planck's constant. If the cell is exposed to a vertical field gradient, the precession frequency is higher at the top of the cell than at the bottom. In normal liquid ^{3}He we would see the signal picked up by the NMR coil (reflecting the net transverse magnetization of the whole volume) rapidly decay as the spins precessing at different rates in different parts of the cell become out of phase. Under conditions we consider here this would limit the NMR signal time to about 1 ms.

In the superfluid B phase, however, the different precession frequencies of the spins in different parts of the cell leads to a gradient in the direction of the spin, which induces a spin supercurrent acting to counter the phase gradient. The vertical spin supercurrent so induced transports down-spin to the lower field regions and up-spin to the higher field regions.

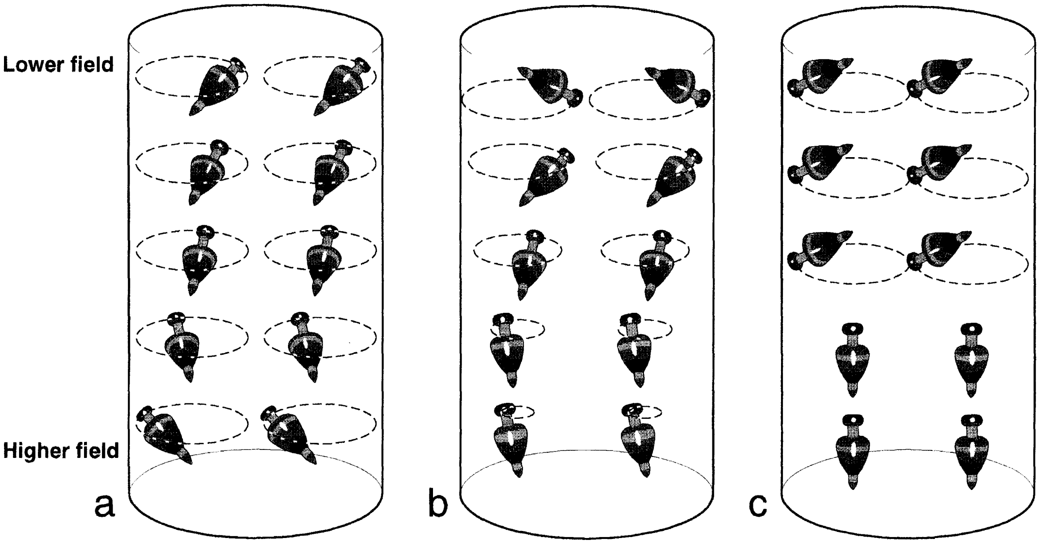

Just after pulse, spins slowly get out of phase.

Spin supercurrent flows. Spins at top tip further over, spins at bottom straighten up.

Precessing domain is formed.

FIG. 21. Formation of the homogeneously precessing domain.

Thus the spin deflection gradually increases in the low-field region and decreases in the higher-field region. This continues until the angle of deflection reaches the critical deflection of 104°. Once this angle is exceeded the extra dipole–dipole restoring force cuts in. This has the effect of increasing the precession frequency above the local Larmor frequency.

Thus as the spin deflection in the low-field region (where the Larmor precession frequency is at its lowest) increases beyond 104°, the frequency increases. This then matches the frequency in the next highest field region; the gradient thus disappears, the supercurrent stops, and the two regions are locked together in frequency. After a short period all the spin deflection of the high-field region of the cell has been transported to the lower-field region and two domains remain, one with coherently precessing spins in the low-field region and the other of completely undeflected spin in the high-field region. This situation can be maintained for hundreds of milliseconds, and if the precessing domain can be trapped in a minimum in the texture away from the disturbing effect of walls, it can continue for tens of seconds (Bunkov *et al.*, 1992). This is in stark contrast to the 1-ms signal seen under the same conditions in the normal state.

The importance of these precessing states lies in the fact that they represent an excited but coherent state. The fact that we can study their evolution by NMR makes them not only interesting but very accessible to study.

4. OTHER SIMILAR SUPERFLUIDS

The superfluids are largely unique in that superfluidity is in general a low-temperature property and relies on the system being liquid. There are other potential superfluids, but phase transitions to solids intervene. However, superfluidity can occur in exotic circumstances. It is possible that some of the heavy-fermion metals that show superconductivity (see HEAVY-FERMION PHENOMENA) do so with Cooper pairs with spin 1. The exact structure of the pairs in high-temperature superconductors at present is a matter of conjecture.

The closest analogies to the helium superfluids may lie among the neutron and proton liquids that are believed to exist inside neutron stars. These exotic objects give rise to the very regular radio pulses that suggested the original name pulsars. They are the stellar bodies that remain after a supernova explosion. Although the temperature of these liquids in absolute terms is very high, they are actually low-temperature systems in the sense that the temperature is much lower than the appropriate Fermi temperatures ($T_F = E_F/k$). This suggests that the liquids are likely to be superfluid with neutron–neutron and proton–proton Cooper pairs as in superfluid ^{3}He. Since neutron stars rotate very rapidly, up to revolution rates approaching 1000 Hz, any internal superfluid must be traversed by a very dense lattice of vortices. It is possible that the starquakes observed when the very slowly decaying rotation rate makes very small (but easily detected) jumps are the result of sudden reorganizations of this vortex lattice.

The dilute gases of trapped cooled atoms in which Bose–Einstein condensations have recently been observed must also be superfluid below the condensation temperature. These systems are being intensively studied at present, and we expect a whole new range of superfluid phenomena to be observed. Superfluidity in these systems is quite different from that in a liquid in that the gas of particles is dilute and interactions are very small.

Finally, one further related system of great interest is provided by dilute solutions of ^{3}He in ^{4}He. At low temperatures, superfluid ^{4}He will dissolve ^{3}He up to around 10% depending on pressure. It is an intriguing thought to imagine what might happen if we took such a mixture to a temperature low enough for the dilute ^{3}He fraction to become superfluid while dissolved in an already superfluid ^{4}He. Such a mixture of two interpenetrating superfluids would certainly show unusual behavior. Interesting questions arise as, for example, in rotation how would the vortices in the two superfluids interact? Unfortunately, at the densities possible in the solution this ^{3}He transition may well need a temperature of fractions of a microkelvin, which is several orders of magnitude below our present capability.

GLOSSARY

Absolute Zero: The lowest possible temperature, at which the thermal energy and entropy are zero.

Adiabatic Demagnetization: Cooling method relying on the temperature reduction undergone by paramagnetic magnetic moments when the ambient magnetic field is reduced.

Boson: Any fundamental or composite particle whose total intrinsic spin is an even multiple of $\hbar/2$.

Bose–Einstein Condensation: The process which occurs at low temperatures in a collection of bosons when large numbers of the particles enter the ground state.

Circulation: The integrated scalar product of the velocity and distance around a closed loop.

Coherence Length: Minimum length scale on which the superfluid wave function may vary.

Cooper Pairs: The boson particles formed from a pair of fermions that allow fermion systems to become superfluid.

Dilution Refrigerator: Refrigerator operating in the millikelvin region, whose working fluid is ^{3}He dissolved in liquid ^{4}He.

Dipole–Dipole Interaction: The nonisotropic interaction between two magnetic moments. In superfluid ^{3}He this coupling determines the orientation of the (magnetic) nuclear spin relative to the (nonmagnetic) orbital angular momentum of the Cooper pairs.

Dispersion Relation: The relationship between the momentum and the energy of a particle.

Energy Gap: The minimum energy of a quasiparticle excitation in superfluid ^{3}He.

Fermion: Any fundamental or composite particle whose total intrinsic spin is an odd multiple of $\hbar/2$. Obeys Pauli's exclusion principle.

Fermi Momentum p_F, Fermi Energy E_F: Momentum and energy of a particle at the Fermi surface.

Fermi Surface: The surface marking the limit of filled states in momentum space for a system of fermions.

Gauge Symmetry: The invariance of a system in condensed matter physics under a global change of the phase of the wave function.

Homogeneous Precessing Domain: Domain that can be set up in superfluid ^{3}He in which the magnetizations of all the Cooper pairs precess coherently.

Landau Velocity: The critical velocity in a superfluid at which excitations may be freely created and the superfluidity breaks down.

Longitudinal NMR: Exotic nuclear magnetic resonance in which the exciting radiofrequency field is parallel to the static field.

Nuclear Cooling: Cooling method for the microkelvin region based on the adiabatic demagnetization of nuclear spins.

Persistent Current: A current in a superfluid that flows indefinitely because of the lack of dissipation.

Phonons, Rotons: Thermal excitations in superfluid ^{4}He.

Quasiparticle Excitations: Excited states of the system whose behavior is similar to propagating particles. These excitations may be either bosons or fermions. In the present article the term is used for the fermionic excitations, which may be particle-like or hole-like in superfluid ^{3}He.

Quantized Vortices: The vortices in a superfluid whose circulation is quantized since the wave function is limited to making a multiple of 2π change around the vortex axis.

Quantum Fluid: Fluid whose behavior is dominated on a macroscopic scale by quantum mechanics.

Spin Supercurrent: A flow of spin in superfluid ^{3}He in response to a gradient in the spin part of the wavefunction.

Spontaneous Symmetry Breaking: The lifting of a symmetry arising from ordering at a phase transition.

Superfluid: Fluid able to flow without dissipation; frictionless liquid.

Superleak: A restricted passage that allows the passage of superfluid but restricts the passage of normal fluid.

Texture: The "grain" in the structure of superfluid ^{3}He, governed ultimately by the directions of the orbital and spin angular momentum directions of the Cooper pairs.

Thermomechanical, Mechanothermal: Terms describing the linking of thermal and mechanical effects which occurs in superfluids.

Works Cited

Anderson, P. W., Brinkman, W. F. (1973), *Phys. Rev. Lett.* **30**, 1108–1110.

Anderson, P. W., Morel, P. (1961), *Phys. Rev.* **123**, 1991–1934.

Balian, R., Werthamer, N. R. (1963), *Phys. Rev.* **131**, 1553–1564.

Bardeen, J., Cooper, L. N., Schrieffer, J. R. (1957), *Phys. Rev.* **108**, 1175–1204.

Borovik-Romanov, A. S., Bunkov, Yu. M., Dmitriev, V. V., Mukharskiy, Yu. M. (1984), *Pis'ma Zh. Eksp. Teor. Fiz.* **40**, 256–259 [*JETP Lett.* **40**, 1033–1037.

Bradley, D. I., Bunkov, Yu. M., Cousins, D. J., Enrico, M. P., Fisher, S. N., Follows, M. R., Guénault, A. M., Hayes, W. M., Pickett, G. R., Sloan, T. (1995), *Phys. Rev. Lett.* **75**, 1887–1890.

Buckingham, M. J., Fairbank, W. M. (1961) *Prog. Low Temp. Phys.* **3**, 80–112.

Bunkov, Yu. M., Fisher, S. N., Guénault, A. M., Pickett, G. R. (1992), *Phys. Rev. Lett.* **69**, 3092–3095.

Campbell, L. J. (1983), *Phys. Rev. B* **27**, 1913–1915.

Clow, J. R. Reppy, J. D. (1972), *Phys. Rev. A* **5**, 424–438.

Donnelly, R. J., Donnelly, J. A., Hills, R. N. (1981) *J. Low Temp. Phys.* **44**, 471–489.

Fisher, S. N., Guénault, A. M., Kennedy, C. J., Pickett, G. R. (1991), *Phys. Rev. Lett.* **67**, 1270–1273.

Fisher, S. N., Guénault, A. M., Kennedy, C. J., Pickett, G. R. (1992), *Phys. Rev. Lett.* **69**, 1073–1076.

Fomin, I. A. (1984), *Pis'ma Zh. Eksp. Teor. Fiz.* **40**, 260–262 [*JETP Letters* **40**, 1037–1040].

Greywall, D. S. (1986), *Phys. Rev. B* **33**, 7520–7538.

Hakonen, P. J., Krusius, M., Salomaa, M. M., Simola, J. T. (1985), *Phys. Rev. Lett.* **54**, 245.

Krusius, M. (1993) *J. Low Temp. Phys.* **91**, 233–273.

Landau, L. D. (1941), *Zh. Eksp. Teor. Fiz.* **11**, 592–614.

Landau, L. D. (1956), *Zh. Eksp. Teor. Fiz.* **30**, 1058–1064 [(1957) *Sov. Phys. JETP* **3**, 920–925].

Leggett, A. J. (1984), *Phys. Rev. Lett.* **53**, 1096–1099.

Osheroff, D. D., Richardson, R. C., Lee, D. M. (1972), *Phys. Rev. Lett.* **28**, 885–888.

Schwab, K., Bruckner, N., Packard, R. E. (1997), *Nature (London)* **386**, 585–587.

Sears, V. F., Svensson, E. C., Martel, P., Woods, A. D. B. (1982), *Phys. Rev. Lett.* **49**, 279–282.

Yarmchuk, E. J., Gordon, M. J. V., Packard, R. E. (1979), *Phys. Rev. Lett.* **43**, 214–217.

Further Reading

Tilley, D. R., Tilley, J. (1974), *Superfluidity and Superconductivity*, London: Van Nostrand.

Wilks, J., Betts, D. S. (1987), *An Introduction to Liquid Helium*, Oxford: Oxford University Press.

SURFACES AND INTERFACES OF SOLIDS, STRUCTURE OF

Feng Liu, M. Hohage and M. G. Lagally, *University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.*

	Abstract	321
	Introduction	321
1.	**Surface Thermodynamics**	322
1.1	Gibbs Surface Model and Surface Tension	322
1.2	Surface Stress and its Relation to Surface Tension	324
2.	**Surface Crystallography**	325
2.1	Orientation of Crystalline Surfaces	325
2.2	Surface Periodicity and Symmetry	326
2.3	Reciprocal-Space Representation and Diffraction	327
3.	**Surface Relaxation and Surface Reconstruction**	329
3.1	A General Discussion	329
3.2	Semiconductor Surfaces	330
3.2.1	The {100} Surfaces	331
3.2.1.1	Dimer Buckling in the {100} Surfaces of Elemental Semiconductors	331
3.2.1.2	The {100} Surfaces of Binary Compound Semiconductors and Electron Counting	332
3.2.2	The {111} Surfaces	334
3.2.2.1	The (2 × 1) Reconstruction on Cleaved Si(111) and Ge(111) and on Clean C(111)	334
3.2.2.2	The Si(111)-(7 × 7) and Ge(111)-c(2 × 8) Reconstructions	336
3.2.2.3	The {111} Surfaces of Compound Semiconductors	337
3.2.3	The {110} Surfaces	338
3.3	Metal Surfaces	339
3.3.1	The fcc (110) Surface	340
3.3.2	The fcc (100) Surface	342
3.3.3	The fcc (111) Surface	343
4.	**Morphological Features of Surfaces, Interfaces, and Thin Films**	343
4.1	The Terrace–Step–Kink (TSK) Model	343
4.2	Equilibrium Step Configurations	344
4.2.1	Stress-Domain Structure	344
4.2.2	Steps on Vicinal Surfaces	344
4.3	Surface/Interface Morphology of Epitaxial Thin Films	345
4.3.1	Equilibrium Growth Modes	345
4.3.2	Kinetics-Limited Growth Morphology	346
4.3.3	Interface Morphology	347
	Glossary	348
	Works Cited	349
	Further Reading	352

INTRODUCTION

Among the properties of surfaces and interfaces, structure is fundamental, as it determines or influences chemical, electrical, mechanical, and even magnetic and optical behavior. In describing structure, it is essential that a scale length be defined. For example, a surface may feel smooth or rough or it may scatter light specularly or diffusely, which are clearly macroscopic or mesoscopic measures of surface structure. On the other extreme, the atomic structure of the surface describes where the atoms are located relative to their positions in the bulk of the crystal. In the past half century, remarkable progress has been made in determining the atomic structure of the surface, driven initially by the development of ultrahigh vacuum, which permitted the preparation of clean surfaces and maintenance of the cleanliness for sufficient time to make detailed structural measurements. The subse-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

quent invention of numerous structural characterization methods and the development of supercomputers and advanced computational algorithms have led to a broad understanding of surface atomic structure.

"Surface structure" is not just "where are the atoms". There are also larger-scale features for whose description the detailed positions of the atoms need not be known. These features are typically affected or determined by extended structural entities such as steps, terraces, dislocations, grain boundaries, and voids. The term "surface morphology" has come into use to describe such larger-range features. Their influence on surface properties can be significantly greater than that of the atomic structure. The analogous term for the bulk of materials is "microstructure". Looking down on a surface to observe its form and topology in the way one looks at the ground from an airplane illustrates why "morphology" is appropriate.

This article discusses surface atomic structure and surface morphology under the umbrella of "surface structure". It aims to provide an overview of basic concepts and fundamental principles related to structural properties of solid surfaces and interfaces and to illustrate systematic trends in structures of surfaces and interfaces in different classes of materials. Section 1 reviews surface thermodynamics, as a basis and a driving force for formation of surface structure. Section 2 introduces concepts and notations required to describe the structures of single-crystal surfaces. Section 3 overviews intrinsic structures of single-crystal surfaces and discusses surface relaxation and reconstruction, with examples chosen from semiconductor and metal surfaces. Section 4 deals with the morphology of real surfaces and interfaces, and the development of morphology through kinetic limitations or thermodynamic driving forces encountered during thin-film growth or surface treatment.

1. SURFACE THERMODYNAMICS

1.1 Gibbs Surface Model and Surface Tension

J. W. Gibbs (1928) laid the groundwork for surface thermodynamics. He introduced a notion of a *dividing surface* and derived all surface thermodynamic properties by an *excess procedure*. A real system consisting of two homogeneous phases separated by a finite inhomogeneous transition region [Fig. 1(a)] is idealized as two homogeneous phases separated by a hypothetical geometrical surface, the dividing surface [Fig. 1(b)]. The dividing surface is chosen to be near the macroscopic observable physical interface and everywhere parallel to it; different choices of dividing surface may be used that are parallel to each other but displaced from each other in the direction of the surface normal. The excess procedure expresses the thermodynamic properties of the real system in terms of the thermodynamic properties of the ideal system and excess properties, which define the difference between the real and ideal values. Thus, all extensive surface thermodynamic functions can be expressed in terms of the dividing surface through the excess procedure. For the two-phase system that is represented by two homogeneous phases, phases 1 and 2, and a dividing surface, as shown in Fig. 1(b), any extensive property can be divided unambiguously into contributions from phase 1, phase 2, and the surface as

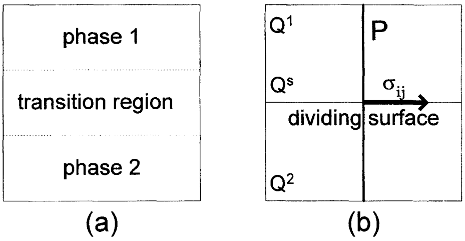

FIG. 1. The Gibbs surface model: (a) the real system; (b) the idealized system. The thin horizontal line in the middle marks the dividing surface; the thick vertical line marks the imaginary plane P used to derive the surface stress tensor σ_{ij}.

$$Q = Q^1 + Q^2 + Q^s, \qquad (1)$$

where Q is the extensive property of the whole system, which is the integration of the specific value of Q over the whole system including the inhomogeneous transition region; Q^1 and Q^2 are, respectively, the same extensive property for phases 1 and 2 that would apply as if they are homogeneous up to the (infinitely narrow) dividing surface; Q^s is the excess extensive property, *i.e.*, the difference between Q and $Q^1 + Q^2$, assigned to the surface. One intriguing problem with the excess extensive property is that it is usually not unique, depending on the choice (or location) of the dividing surface (see discussion below).

One defines surface tension γ as the reversible work required to create unit area of surface (by cleavage), at constant temperature (T) and pressure (p); this function is the partial derivative of Gibbs free energy (G) of the whole system with respect to area of surface formed (A), at constant T, p, and mole concentration of each component (n_i):

$$\gamma = \left(\frac{\partial G}{\partial A}\right)_{T,p,n_i}. \tag{2}$$

As $G = G(T, p, n_i, A)$, the differential of Gibbs free energy can be written as

$$dG = -S\,\mathrm{d}T + V\,\mathrm{d}p + \Sigma\mu_i\,\mathrm{d}n_i + \gamma\,\mathrm{d}A, \tag{3}$$

where S is the entropy, V the volume, and μ_i the chemical potential of molecular species i.

Alternatively, γ can be defined as the partial derivative of Helmholtz free energy (F) of the whole system with respect to A, at constant T, V, and n_i,

$$\gamma = \left(\frac{\partial F}{\partial A}\right)_{T,V,n_i}; \tag{4}$$

and the differential of Helmholtz free energy, $F = G - pV$, is

$$\mathrm{d}F = -S\,\mathrm{d}T - p\,\mathrm{d}V + \sum \mu_i\,\mathrm{d}n_i + \gamma\,\mathrm{d}A. \tag{5}$$

(Similarly, γ can be defined in terms of internal energy, $E = F + TS$, and grand canonical potential, $\Omega = F - \sum \mu_i n_i$.)

From Eqs. (1) and (3) and applying the first and second law of thermodynamics to the two homogeneous phases 1 and 2, the differential of surface Gibbs free energy is derived as

$$\mathrm{d}G^s = \sum \mu_i\,\mathrm{d}n_i{}^s + \gamma\,\mathrm{d}A. \tag{6}$$

Here $n_i{}^s$ is the excess mole concentration of the ith species assigned to the surface. According to Euler's theorem, the surface Gibbs free energy is

$$G^s = \sum \mu_i n_i{}^s + \gamma A, \tag{7}$$

and similarly, the surface Helmholtz free energy is

$$F^s = pV^s + \sum \mu_i n_i{}^s + \gamma A \approx \sum \mu_i n_i{}^s + \gamma A, \quad \text{for } V^s \to 0. \tag{8}$$

Equations (7) and (8) give another definition for surface tension:

$$\gamma = g^s - \sum \mu_i \Gamma_i{}^s \approx f^s - \sum \mu_i \Gamma_i{}^s, \tag{9}$$

where g^s (f^s) is the Gibbs (Helmholtz) free energy per unit area of surface, *i.e.*, the specific surface Gibbs (Helmholtz) free energy, and $\Gamma_i{}^s$ is the surface density $n_i{}^s/A$ of the ith species.

In general, surface tension γ, which measures the change in Gibbs or Helmholtz free energy for the whole system in creating the surface, has two contributions: the change in Gibbs or Helmholtz free energy per unit area for the surface "phase", g^s or f^s, and the change per unit area of surface formed for the surrounding bulk phases. Note that g^s, f^s, and $\Gamma_i{}^s$ are all *excess* surface properties that are dependent on the position of the dividing surface, but γ, a quantity associated with the whole system, is not, because the position of the dividing surface affects the Gibbs (or Helmholtz) free energy in such a way that the change in the surface term g^s (or f^s) is always balanced by the change in the bulk term $\Sigma\mu_i\Gamma_i{}^s$. For a one-component system, it is possible to choose the dividing surface so that $n_1{}^s$ and, hence, $\Gamma_1{}^s$ vanish; the specific surface Gibbs or Helmholtz free energy then becomes equal to surface tension. Consider a one-component system with solid and gas phases;

the atomic density assumes a constant value of n_{bulk} in the solid phase and a value of zero in the gas phase. In the interface (surface) region, the atomic density decreases gradually from the bulk value to zero as one goes from the bulk phase toward the gas phase. If the dividing surface is located inside the transition region, in the vicinity of the surface, there will be a thin region of "extra" atomic density (>0) in the gas phase and a thin region of "lost" atomic density ($<n_{\mathrm{bulk}}$) in the solid phase. By choosing the position of the dividing surface so that the extra density in the gas phase equals exactly the lost density in the solid phase, there will be no excess surface density, *i.e.*, n_1^{s} and hence Γ_1^{s} vanish. For a multi-component system, however, one choice of the dividing surface that makes n_i^{s} vanish will not make other $n_{j\neq 1}^{\mathrm{s}}$ vanish. Therefore, in general, surface tension γ should not be confused with the specific Gibbs (or Helmholtz) surface free energy, g^{s} (or f^{s}).

Surface tension is a very difficult quantity to measure experimentally, but it can in principle be calculated with a reliable energy functional, such as those based on first-principles (or *ab initio*) computational techniques. In an empirical way, surface tension (surface energy) of a solid surface has two contributions: the formation energy and the relaxation energy. The former reflects the breaking of bonds to make a solid surface (at the ideal bulk terminations); the latter reflects the tendency of a solid surface to distort because it is a quasi two-dimensional (2D) system and hence would like to assume an atomic structure and bonding configuration different from that of the bulk. To create a surface, it costs energy to break bonds, and the cost is partly recovered by the relaxation process involving rearrangement of atoms and bonds at the surface. The formation energy (positive) dominates the relaxation energy (negative). The energy of a single-crystal surface scales approximately with the cohesive energy of the bulk crystal. The greater the bulk cohesion, the stronger the interatomic bonds, the higher the surface energy. The weakly van der Waals–bonded rare-gas solids have the lowest surface energy; the strongly bonded semiconductors and metals have high surface energy. In contrast to that of an isotropic liquid surface, the energy of a solid surface is anisotropic, depending on the orientation of the surface, because a different number of bonds is broken to create different surface orientations.

1.2 Surface Stress and its Relation to Surface Tension

Imagine that there exists a plane P normal to the dividing surface [Fig. 1(b)]. One may derive the surface force acting on P by the same excess procedure discussed above. The total force acting across P, from the material on one side of P to the material on the other side, can be divided into contributions from the two homogeneous bulk phases and from the surface phase. The surface force per unit length of the line intersection of the plane P with the dividing surface plane is then defined as the surface stress.

Gibbs (1928) first pointed out the distinction, for the case of a solid, between surface tension and surface stress: the former measures the energy cost to create unit area of new surface; the latter measures the energy cost to deform the surface. The relationship between surface stress and surface tension can be derived as follows.

Suppose we deliberately deform the surface, at constant temperature and total amount of surface species, introducing an infinitesimal strain ε_{ij} to the surface. The work required to do so is

$$\delta W = A\sum \sigma_{ij}\varepsilon_{ij} \quad (i,j=1,2). \tag{10}$$

Here σ_{ij} denotes the surface stress tensor. This work is equal to the infinitesimal change of surface Gibbs or Helmholtz free energy, G^{s} or F^{s}, at constant T and n_i^{s},

$$\begin{aligned}\delta G^{\mathrm{s}} &= \delta F^{\mathrm{s}} = \delta(\gamma A) = \gamma\delta A + A\delta\gamma \\ &= \gamma A\sum \varepsilon_{ii} + A\sum \frac{\partial\gamma}{\partial\varepsilon_{ij}}\varepsilon_{ij} \quad (i,j=1,2).\end{aligned} \tag{11}$$

Combining Eqs. (10) and (11), we have

$$\sigma_{ij} = \gamma\delta_{ij} + \frac{\partial\gamma}{\partial\varepsilon_{ij}} \quad (i,j=1,2), \tag{12}$$

where the Kronecker delta $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. Surface stress, σ_{ij}, is independent of the location of dividing surface, although it is an excess surface property, as indicated by its relation to surface tension γ in Eq. (12).

Equation (12) also implies that the surface stress tensor σ_{ij} is defined as the derivative of surface Gibbs or Helmholtz free energy (G^s or F^s) with respect to surface strain ε_{ij}, at constant temperature and total amount of surface species,

$$\sigma_{ij} = \frac{1}{A}\left(\frac{\mathrm{d}G^s}{\mathrm{d}\varepsilon_{ij}}\right)_{T,n_i^s} = \frac{1}{A}\left(\frac{\mathrm{d}F^s}{\mathrm{d}\varepsilon_{ij}}\right)_{T,n_i^s} \quad (i,j = 1,2). \tag{13}$$

Equivalently, surface stress may be defined in terms of other thermodynamic energy functions.

For a liquid, surface stress is equal to surface tension because liquid is a purely plastic medium and its surface tension is independent of strain, *i.e.*, the second strain-derivative term on the right-hand side of Eq. (12) vanishes. (For a one-component liquid, surface tension also equals the specific Gibbs or Helmholtz surface free energy. So, for this special case, the values of all three quantities, surface stress, surface tension and specific surface free energy, are the same, which has often caused confusion.) When a liquid surface is stretched, atoms rapidly enter the surface from the bulk; the existing surface atoms stay where they are and the new surface is formed of atoms from the bulk now occupying surface positions. Thus, expanding a liquid surface is equivalent to creating a new liquid surface. In contrast, solids are a purely elastic medium, at least at low deformations. When a solid surface is stretched, surface atoms are displaced from their minimum-energy positions, sitting at new sites for which the surface as a whole has a different energy. Thus, the surface tension of a solid surface depends on surface strain. While a liquid surface is always under tensile stress for γ to be positive, a solid surface can be under an overall stress that is either tensile or compressive because the second term in Eq. (12), the strain derivative of tension, can be either positive or negative and may have a magnitude larger than the first term, the tension contribution. Also, the strain-derivative term often makes the surface stress of a solid surface anisotropic, while the surface stress of a liquid surface is always isotropic.

2. SURFACE CRYSTALLOGRAPHY

2.1 Orientation of Crystalline Surfaces

Given a single crystal of a particular Bravais lattice (see CRYSTALLOGRAPHY), a lattice plane is defined to be any plane containing at least three noncollinear Bravais lattice points; it is denoted by Miller indices, which are the coordinates of the shortest reciprocal-lattice vector normal to the plane. A common way of creating a crystalline surface is to cut a bulk single crystal parallel to one of its low-index atomic planes (*e.g.*, {100}, {110} or {111} planes in a cubic single crystal). An ideal surface (one that has all atoms in bulk positions) is made by removing all the atoms lying on one side of a chosen atomic plane and keeping the positions of all remaining atoms intact. The orientation of this surface is then well defined, with its surface normal specified by the directional indices, [*hkl*], defined in the conventional way with respect to the three-dimensional (3D) unit cell of the bulk crystal (Ashcroft and Mermin, 1976). The surface is denoted as an (*hkl*) surface, using the Miller indices of the corresponding parallel set of atomic planes in the bulk crystal. The notation {*hkl*} denotes a set of equivalent planes or surfaces.

Low–Miller-index surfaces are widely used in surface science research and epitaxial growth of thin films because of their relatively greater stability and high symmetry. A real surface generally consists of a number of terraces of low-index surface planes of the same orientation separated by steps. In *nominal* {*hkl*} surfaces the normal to the average surface orientation coincides with the normal to the *singular* {*hkl*} terraces that constitute the surface. To achieve this condition, there must be on the average an equal number of "up" steps and "down" steps in a nominal {*hkl*} surface [see

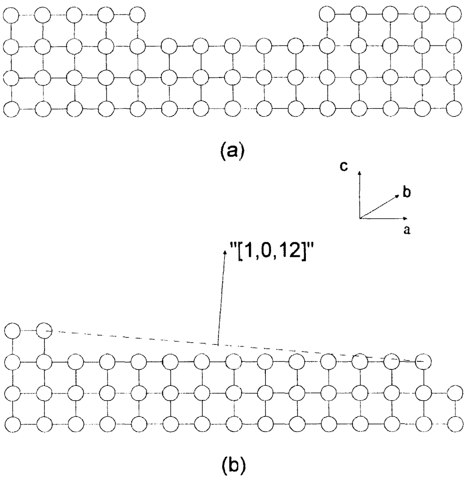

FIG. 2. Side views ([0$\bar{1}$0] projection) of (001) surfaces of a simple cubic lattice. (a) A nominal surface consisting of three singular (001) terraces separated by one up and one down single-atomic-height step. (b) A vicinal surface miscut by ≈4.8° toward [100]. The normal (arrow) to the average surface orientation (dashed line) is [1 0 12].

Fig. 2(a)]. In practice, it is very difficult to cut a crystal exactly parallel to a chosen low-index plane. Even if there is no average miscut, the uncertainties in cutting and polishing lead to a wavy surface, giving only a nominally low-index surface. A vicinal surface is made by a deliberate cut up to a few degrees away from a low-index plane toward a specific direction. For example, Fig. 2(b) shows a vicinal surface of a simple cubic lattice formed by a miscut $\approx 4.8^0$ away from [001] toward [100]. A vicinal surface may be specified either by the normal to its average surface orientation, which generally produces unreasonably large Miller indices [*e.g.*, the surface in Fig. 2(b) would be labeled as (1 0 12)], or by the indices of the constituent terraces plus the miscut angle and miscut orientation {*e.g.*, Si(001) miscut 2^0 toward [110]}. The latter choice makes it convenient to understand and discuss the properties of a small-miscut vicinal surface in terms of its low-index singular terraces and the effects of defects (steps). If the miscut angle is not too small and the index of the normal to the average surface orientation does not contain numbers that are too large, the surface is named with its appropriate Miller indices.

2.2 Surface Periodicity and Symmetry

The structure of a single-crystal solid is uniquely defined by its intrinsic symmetry (Ashcroft and Mermin, 1976), which includes translational symmetry (periodicity), symmetry of the Bravais lattice (primitive unit cell), and symmetry of the unit cell basis (atoms or molecules) in the unit cell. The translational symmetry in 3D is realized by 7 different *crystal systems*: cubic, tetragonal, orthorhombic, monoclinic, triclinic, trigonal, and hexagonal. Allowing more than one grid point per unit cell leads to 14 *Bravais lattices*, including types of primitive (p), body-centered (bc), face-centered (fc), and base-centered unit cells. Each lattice point may have more than one atom associated with it, reducing the symmetry of the original crystal system. There are 32 *crystallographic point groups* that a crystal structure can have, covering symmetry operations of rotation, reflection, and inversion, and combinations of them. The full symmetry of the crystal (including atomic arrangement) is described completely by 230 *crystallographic space groups*, with the addition of symmetry operations of screw axes and glide planes.

As a crystalline solid is truncated to make a surface, the translational symmetry of the solid perpendicular to the surface is removed, while the periodicity parallel to the surface remains. Because in the surface region each atom layer becomes intrinsically inequivalent to other layers, all the symmetry properties of a solid surface are two-dimensional. Consequently, surface structures, in analogy to 3D crystal structures, are classified by the 2D Bravais lattice (net) and symmetry groups of the net and of the atomic basis associated with the net. There are five different 2D Bravais nets: oblique, primitive (p) rectangle, centered (c) rectangle, square, and hexagonal, as depicted in Fig. 3; ten 2D point groups; and seventeen possible 2D space groups (Woodruff, 1981).

The 2D Bravais net consists of a grid of equivalent points; the atomic basis represents the arrangements of atoms at each net point. The translational symmetry of a real surface usually differs from that of an ideal bulk termination of the solid, *i.e.*, that of the corresponding atomic plane in the bulk, because of surface reconstruction (see discussion below). If $(\boldsymbol{a}_s, \boldsymbol{b}_s)$ and $(\boldsymbol{a}_i, \boldsymbol{b}_i)$ are the unit net vectors

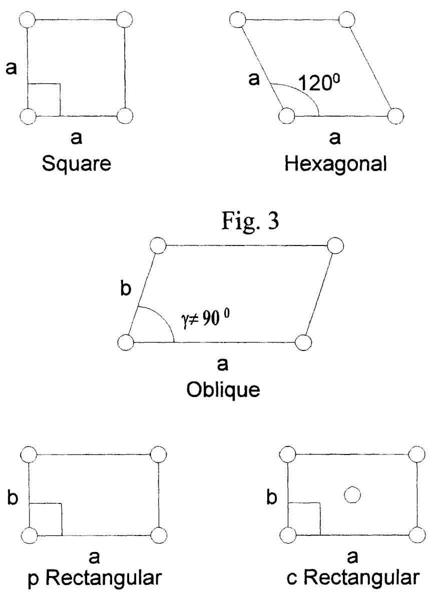

FIG. 3. The five two-dimensional Bravais nets: oblique, square, hexagonal, and primitive (p) or centered (c) rectangle.

(the primitive translational vectors) of the real surface and of the ideal surface, respectively, the general translational vectors are

$$\boldsymbol{T}_s = n\boldsymbol{a}_s + m\boldsymbol{b}_s \tag{14}$$

in the real surface and

$$\boldsymbol{T}_i = n\boldsymbol{a}_i + m\boldsymbol{b}_i \tag{15}$$

in the ideal surface. Here n and m are integers. The relationship between the real surface net and ideal surface net is uniquely defined by a (2 × 2) matrix, $\boldsymbol{G}$ (Park and Madden, 1968),

$$\begin{pmatrix} \boldsymbol{a}_s \\ \boldsymbol{b}_s \end{pmatrix} = \boldsymbol{G}\begin{pmatrix} \boldsymbol{a}_i \\ \boldsymbol{b}_i \end{pmatrix}. \tag{16}$$

The determinant of $\boldsymbol{G}$, $|\boldsymbol{G}|$, equals the ratio of the areas of the real unit net and the ideal unit net. If $|\boldsymbol{G}|$ is an integer, the nets are *simply* related; if $|\boldsymbol{G}|$ is a rational number, the two nets are *rationally* related. When $|\boldsymbol{G}|$ is irrational, the real surface has an incommensurate structure relative to the ideal surface (substrate).

Although the matrix notation is exact, it is seldom used. Instead, a notation introduced by

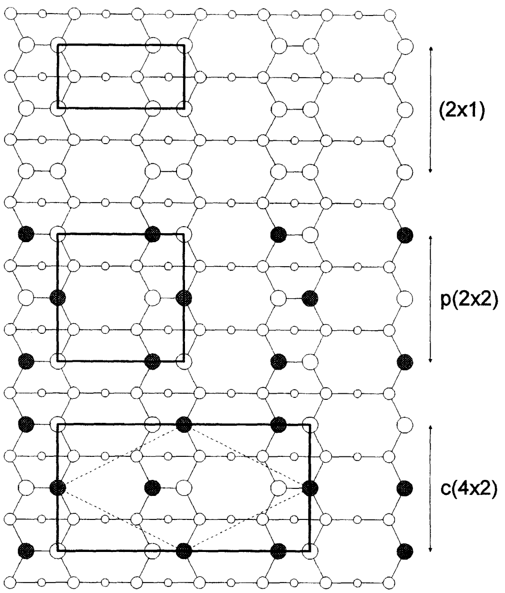

FIG. 4. Schematic top view of three possible reconstructions on Si(001). Top region: (2 × 1); middle: p(2 × 2); bottom: c(4 × 2). Open and solid circles mark atom positions; their size indicates different layers of atoms, with the largest circles in the outermost layer. In the middle and bottom regions, the slightly different heights of the two atoms in a dimer due to buckling are indicated by solid and open circles. The unit mesh of each reconstruction is depicted by a dark rectangle. For the c(4 × 2) structure, the primitive unit mesh is depicted by a dashed-line rhombus.

Wood (1964) has been widely adopted and is convenient for small commensurate unit meshes. This simple notation uses the ratio of the lengths of the real primitive net vectors to those of ideal primitive vectors plus an angle through which the real net is rotated with respect to the ideal net. Figure 4 illustrates the use of this notation to describe several possible reconstructions on Si(001). A detailed discussion of these reconstructions will be presented later.

2.3 Reciprocal-Space Representation and Diffraction

Conventionally, surface structures are detected through their representation in reciprocal space using diffraction-based surface probes. An important consequence of periodicity is the modification of the law of momentum conservation. As a wave (*e.g.*, electrons or photons) interacts with a periodic structure, its momentum is conserved plus or minus any reciprocal-lattice vector (lattice momentum). For a surface, the periodicity only exists parallel to the surface, so the modified law of momentum conservation applies only to the components of the wave vectors parallel to the surface,

$$\boldsymbol{k}_{\|} = \boldsymbol{k}_{0\|} + \boldsymbol{g}_{hk}. \tag{17}$$

$\boldsymbol{k}_{0\|}$ and $\boldsymbol{k}_{\|}$ are parallel to surface components of incident and diffracted-wave vectors. $\boldsymbol{g}_{hk}$ is a surface reciprocal-net vector (*i.e.*, a general translational vector in $\boldsymbol{k}$ space), which is expressed as

$$\boldsymbol{g}_{hk} = h\boldsymbol{a}^* + k\boldsymbol{b}^*, \tag{18}$$

where h and k are integers; $\boldsymbol{a}^*$ and $\boldsymbol{b}^*$ are primitive unit vectors of the reciprocal net, defined relative to the real-space net as

$$\boldsymbol{a}^* = (2\pi\boldsymbol{b}_s \times \boldsymbol{n})/A, \quad \boldsymbol{b}^* = (2\pi\boldsymbol{a}_s \times \boldsymbol{n})/A,$$

$$A = \boldsymbol{a}_s \cdot \boldsymbol{b}_s \times \boldsymbol{n}, \tag{19}$$

where $\boldsymbol{n}$ is the unit vector normal to the surface.

The conditions for diffraction are totally defined by the momentum conservation expressed in Eq. (17) together with an energy conservation equation,

$$k^2 = k_0{}^2 \text{ or } k_{\|}{}^2 + k_{\perp}{}^2 = k_{0\|}{}^2 + k_{0\perp}{}^2. \tag{20}$$

These conditions can be conveniently visualized using the Ewald sphere construction in $\boldsymbol{k}$ space, as illustrated in Fig. 5. One first draws a vector of the incident wave, $\boldsymbol{k}_0$, to the origin of the reciprocal space, and then draws a sphere of radius k_0 about the tail of the vector $\boldsymbol{k}_0$. In three dimensions, this (Ewald) sphere passes through certain reciprocal-lattice points in addition to the origin, and the diffracted beam $\boldsymbol{k}$ is simply given by a vector from the tail of $\boldsymbol{k}_0$ to those reciprocal-lattice points lying on the Ewald sphere. A surface can be approximated by a 2D net. It is straightforward to show that for a 2D net, the complete reciprocal space consists of a set of lines or "rods" that are infinite in extent, are perpendicular to the 2D net, and pass through the reciprocal-net points. Because those rods inside the Ewald sphere will always intersect the sphere, the conditions for the appearance of diffracted beams are greatly relaxed, relative to those for a 3D crystal, for which the reciprocal lattice consists of points. Therefore, a small change in energy of the incident wave, *i.e.*, in the length of $\boldsymbol{k}_0$ and the radius of Ewald sphere, will totally extinguish one set of diffracted beams in three dimensions. The same change for a surface will only shift slightly the direction of the diffracted beams.

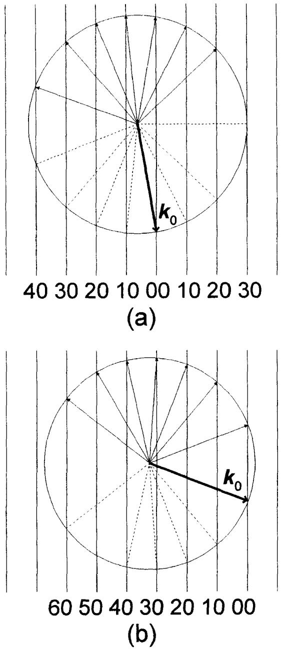

FIG. 5. Ewald sphere construction for a two-dimensional system, *e.g.*, a single layer. The incident beam is denoted by the vector k_0 and all other radial vectors ending on the Ewald sphere correspond to diffracted beams, k. (a) Near-normal incidence for LEED; (b) low-angle incidence for RHEED. Dashed lines indicate those diffracted beams penetrating into the bulk. For RHEED, k_0 is generally much larger, relative to the spacing of the reciprocal-lattice rods, than is shown in (b); for LEED, the relative magnitudes shown in (a) are appropriate.

Electron beams and x rays are typically used in diffraction-based surface probes. Figure 5 illustrates two common setups for electron diffraction: low-energy electron diffraction (LEED), generally performed at near-normal incidence, and reflection high-energy electron diffraction (RHEED), at glancing incidence. Most forms of radiation used for diffraction (electrons, x rays) have a mean free path in solids sufficiently large to penetrate some distance into the crystal. For example, for electrons this mean free path varies from approximately 5 Å up to thousands of ångströms, as the energy is changed. The actual penetration into the crystal can be limited by changing the angle of incidence (compare LEED and RHEED). The finite penetration of radiation into the crystal implies that a 2D net and perfect reciprocal-lattice rods to describe scattering from a surface form only a limiting approximation, *i.e.*, there is some 3D character to the reciprocal lattice. Thermal He and other noble-gas atoms, on the other hand, are scattered above the surface without penetrating the crystal because the energies for the wavelengths appropriate for crystal diffraction are much lower. For He scattering, a 2D net and pure reciprocal-lattice rods are a good approximation; consequently, of course, only the structure of the outer surface layer can be detected. Helium scattering is also very useful in studying "unstable" surface structures such as adsorbate layers because of the very low energy of the probe.

In addition to determining the periodicity (*i.e.*, long-range order) of the surface structure, diffraction can in principle also determine the positions of the atoms in the structure. This sensitivity to the atomic basis associated with the lattice occurs through the intensities of the diffracted beams. For electron diffraction, determination of atomic positions is complicated by the strong interaction of electrons with atoms, causing multiple scattering. The use, for example, of x rays (for which the interaction is much weaker) at grazing incidence avoids this problem.

As real-space surface structures are derived indirectly from their representations in reciprocal space with diffraction techniques, controversies often arise when different structural models are fitted to the diffraction data, especially when a large atomic basis is involved or when the relaxation from the ideal (bulk-termination) surface is small. The advance of a new generation of surface probes using scanning tunneling and force microscopies has allowed the direct observation of surfaces in real space, often with atomic resolution. In many ways, the "old" diffraction probes and the new scanned probes compensate each other: the former, working in reciprocal space, are sensitive to the long-range order, and the latter, working in real space, give local order with atomic detail. It is frequently much easier to construct a correct model to fit the diffraction data using scanned-probe real-space images as input. In modern laboratories, diffraction and real-space techniques are often combined to monitor and determine surface structures *in situ* and in real time.

3. SURFACE RELAXATION AND SURFACE RECONSTRUCTION

3.1 A General Discussion

Real surfaces do not retain the ideal bulk termination. Atoms of surface and near-surface layers are generally displaced from their ideal bulk positions, residing at new positions of minimum energy in response to a change of atomic and electronic environments in the surface regions. The atomic rearrangements can take place a few layers deep into the bulk, but the most dramatic structural changes usually occur at only the true surface layer. The structure of this outer layer dominates most surface properties. Surface *relaxation* involves only atomic displacements that do not change the translational symmetry of the surface (*i.e.*, the periodicity of the surface), *e.g.*, collective equal displacements of all atoms in the surface layer; surface *reconstruction* describes atomic displacements that change the periodicity of the surface. Surface relaxation should not be confused with atomic relaxations at defects (vacancies, adatoms, steps, *etc.*), at which the periodicity is already disturbed. A change in surface periodicity can also be caused by a change of atom number density in the surface layers. Because atoms in the surface layer are undercoordinated, missing all neighbors on the vacuum side of the surface, atoms may be added or deleted from the top layer in order to minimize the surface free energy, as has been discussed in Sec. 1.

The driving force for surface relaxation and reconstruction is minimization of surface (free) energy. Surfaces with high surface energy have a stronger tendency to relax and/or reconstruct than surfaces with low surface energy. For a given solid, higher-energy (generally higher-index) orientations of a surface usually relax and/or reconstruct more than lower-energy (lower-index) orientations [see, *e.g.*, the discussion in Sec. 3.3 of fcc metal (110), (100), and (111) surfaces]. It is, however, the reduction of surface energy rather than the surface energy itself that controls the tendency for a surface to relax and/or reconstruct. The nature of interatomic interactions in different classes of solids is the key factor controlling their respective surface structural properties. The surfaces of rare-gas solids, in which atoms are bonded with the weak van der Waals interaction, are most inert, exhibiting only small relaxations perpendicular to the surface; surfaces of metals that have isotropic metallic bonds most often do not reconstruct but display large relaxations; semiconductor surfaces with broken highly directional covalent bonds have the strongest tendency to reconstruct.

The lowering of surface energy can be achieved by reducing surface chemical energy and/or surface strain energy; the two often compete. The reduction of chemical energy, which dominates at the atomic level by optimizing bonding and/or electron density, often occurs at the expense of strain energy, which is caused by bond distortion [see, *e.g.*, the discussion of reconstruction on Si(001) in Sec. 3.2]. The reduction of strain energy becomes dominant in determining morphologies at the mesoscopic scale because elastic interactions extend to a much longer range than do chemical interactions.

The fundamental mechanisms for reducing surface chemical energy through surface relaxation/reconstruction depend strongly on the nature of chemical bonding in the underlying solid. In semiconductors, atoms bond with each other by sharing electron pairs in covalent bonds. The major effect of creating a surface is to break covalent bonds, introducing a large number of dangling bonds in the surface. Consequently, there exists a strong driving force for reconstructing semiconductor surfaces to remove dangling bonds on surface atoms. In metals, atoms bond with each other through the effective medium of an electron gas, and the major effect of creating a surface is the depletion of electron density in the vicinity of the surface. Surface atoms tend to rearrange themselves to reoptimize their surrounding electron density. Such atomic rearrangement in both semiconductor and metal surfaces involves, inevitably, changes of interatomic spacings and distortion of bond angles, so that the gain in chemical energy is usually achieved at the expense of strain energy.

The surface strain energy can be reduced by morphological changes at the mesoscopic scale, such as creation and modulation of steps and dislocations (see Sec. 4.2). Such structural and morphological changes give rise to a variation of surface stress over the surface, which generates a distribution of elastic force density. The strain relaxation energy equals the integral of force density times displacement over the whole surface area. Thus, the strain relaxation process is always accompanied by a redistribution of the surface stress field.

A large portion of research in surface science, especially in earlier years, focused on determination of the intrinsic surface structure—in particular, surface relaxation and surface reconstruction of clean crystalline solid surfaces. In the following we discuss some general features of relaxation and reconstruction in semiconductor and metal surfaces and present a few typical examples.

3.2 Semiconductor Surfaces

Semiconductors are generally made of atoms with s and p valence orbitals each containing on average two electrons (the half-filled s and p shells of an average atomic configuration yield the filled valence band and empty conduction band of the semiconducting bulk). In those semiconductors that crystallize in diamond (elemental semiconductors) and zincblende (binary compound semiconductors) structures, the s and p orbitals at each atomic site hybridize into an sp^3 configuration, which is characterized by four equivalent hybrid orbitals pointing to the corners of a tetrahedron with the atom at the center. Therefore, each atom is tetrahedrally coordinated, forming four covalent bonds with its four nearest neighbors, and each bond shares an electron pair, amounting effectively to a total of eight electrons, a very stable closed s and p shell, on each atom. In a Group-IV elemental semiconductor, each atom contributes one electron per covalent bond. In a binary compound semiconductor, each atom contributes $f = n_V/4$ electrons per covalent bond, where n_V is the number of valence electrons in a single atom; $f_c + f_a = 2$, where subscripts c and a denote cations and anions, respectively.

In creating a surface, some covalent bonds must be broken, leaving behind dangling bonds, hybrid orbitals with a single electron. Because of the high energy cost of dangling bonds, the surface energy of semiconductor surfaces is usually high, and there exists a strong tendency for semiconductor surfaces to relax and/or to reconstruct to reduce the number of dangling bonds or to reconfigure the dangling bonds into a lower-energy state.

3.2.1 The {100} Surfaces The {100}-oriented surfaces of elemental and compound semiconductors have been the most extensively studied surfaces because of their technological importance. In particular, the Si(100) surface serves as the foundation of modern semiconductor technology. In an ideal {100} surface of a diamond or zincblende structure, each atom has two dangling bonds, caused by the loss of two of its nearest neighbors [see Fig. 6(a)]. The primary feature of a reconstructed {100} surface is dimerization, the rebonding of two neighboring surface atoms to form a dimer to eliminate one dangling bond per atom [Fig. 6(b)]. The rebonding, which bonds two second-nearest-neighbor atoms into a distance almost equal to the first-nearest-neighbor separation, requires a rather large distortion of bond angles. Thus, the decrease in electronic energy by rebonding is partly offset by an increase in strain energy. The minimization of strain energy drives the dimers to order into parallel rows, leading to a (2 × 1) reconstruction (see top region of Fig. 4). The dimerization of Si(001) and Ge(001) was first proposed by Schlier and Farnsworth (1959) as an explanation of their (2 × 1) LEED patterns observed at room temperature. The dimer rows have been "seen" directly by scanning tunneling microscopy (STM) on both Si(001) (Tromp *et al.*, 1985, Hamers *et al.*, 1986a) and Ge(001) (Kubby *et al.*, 1987), as well as on (001) surfaces of compound semiconductors, such as GaAs(001) (Biegelsen *et al.*, 1990a, 1990b).

Using a typical dimer bond length of 2.5 Å, it is estimated that the gain in electronic energy is about 4 eV per dimer, based on a tight-binding model (Harrison, 1981), and the cost in strain energy is about 2 eV per dimer, based on the Keating model (Keating, 1966); so, overall, the dimerization, or more precisely the (2 × 1) reconstruction, lowers the surface energy by ≈2 eV per dimer. This estimate is roughly the same for all the elemental and binary compound semiconductor {100} surfaces, and it is in very good agreement with *ab initio* calculations (Ramstad *et al.*, 1995) for Si(001).

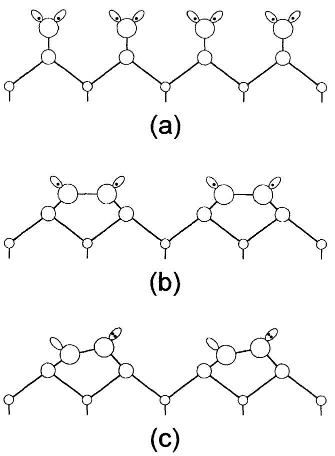

FIG. 6. Dangling-bond configurations on {100} surfaces of elemental semiconductors: (a) two dangling bonds (hybrids with a single electron) on each surface atom in an ideal bulk-terminated surface; (b) one dangling bond on each atom in a dimer. Each dangling bond is half-filled, leading to a metallic surface layer. (c) Charge transfer from downward atom to upward atom in a tilted ("buckled") dimer. The dangling bond on the lower atom is empty and the one on the higher atom is filled, leading to a semiconducting surface layer.

3.2.1.1 Dimer Buckling in the {100} Surfaces of Elemental Semiconductors The energies of two dangling bonds are degenerate in a dimer reconstruction if the dimers consist of two identical atoms in identical positions. Such dimers are called symmetric or "unbuckled" dimers. In the symmetric configuration, the two degenerate dangling bonds are both half-filled, resulting in a metallic surface. Because such energy degeneracy is not a spin degeneracy, the symmetric structure is unstable against Jahn–Teller distortion, which will lower the free energy by lowering the symmetry of the dimer, *i.e.*, by removing the degeneracy of the dangling bonds in the dimer and changing the surface from metallic to semiconducting. An asymmetric dimer can be most easily achieved by a simple tilt, or buckling, as shown in Fig. 6(c).

The buckling of dimers induces rehybridization of surface bonds: the atom moving downward changes its back bonds more to an sp^2-like bonding (about equally bonded to its three nearest neighbors within a plane) and hence its dangling bond to more p-like; the atom moving upward changes its back bonds to more p-like (bond angles are reduced toward 90°) and its dangling bonds to more s-like. The two reconfigured dangling-bond states result in a charge transfer from the downward atom to the upward atom and a lowering of electronic energy. In other words, the buckling of a dimer splits two degenerate dangling-bond energy levels. The dangling-bond bonding level, of s character, lies below the maximum of the bulk valence band and will then be occupied; the dangling-bond anti-bonding level, of p character, lies above the maximum of the valence band and will be empty, making a metallic surface layer semiconducting. Chadi (1979), using tight-binding calculations, first showed that the symmetric dimers are unstable against buckling on Si(001). *Ab initio* calculations (for a list of these calculations, see, *e.g.*, Ramstad *et al.*, 1995) later confirmed this conclusion. The energy gain from buckling is $\approx 0.2\,\text{eV}$ per dimer (Ramstad *et al.*, 1995). Similar results are obtained for Ge(001) (Needels *et al.*, 1987).

The buckling, like the dimerization, decreases electronic energy at the expense of increasing strain energy. The minimization of strain energy causes periodic arrays of buckling patterns that give rise to $c(4 \times 2)$ and $p(2 \times 2)$ reconstructions, which have been observed at low temperatures on Si(001) and Ge(001) by LEED (Kevan and Stoffel, 1984; Tabata *et al.*, 1987) and by x-ray diffraction (Lucas *et al.*, 1993). Low-temperature STM (Wolkow, 1992) of Si(001) shows that surface dimer buckling begins to "lock in" below $\approx 200\,\text{K}$, the number of buckled dimers that are locked into the buckled position increases with decreasing temperature, and the ordering of buckled dimers forms local $c(4 \times 2)$ and $p(2 \times 2)$ domains, with the $c(4 \times 2)$ domains dominating.

Although buckled (tilted) dimers are more stable than unbuckled (untilted) dimers, there seems to be a rather low energy (kinetic) barrier for a buckled dimer to switch from one buckling configuration (*e.g.*, with the left-side atom up) to another (with the left-side atom down). At sufficiently high temperature, surface dimers can switch their orientations so rapidly and independently of each other that the buckling configurations are in a dynamic disorder, leading to an averaged symmetric appearance in STM and to a (2×1) diffraction pattern. LEED (Kevan and Stoffel, 1984; Tabata *et al.*, 1987) and x-ray diffraction (Lucas *et al.*, 1993) have revealed a structural phase transition from the (2×1) to the $c(4 \times 2)$ phase as the temperature is decreased, corresponding to the freezing in of rocking dimers, in the temperature range between 150 and 250 K. Theoretical calculations (Ihm *et al.*, 1983; Saxena *et al.*, 1985; Zubkus *et al.*, 1991) predict a second-order phase transition from an ordered $c(4 \times 2)$ or $p(2 \times 2)$ structure to a disordered (2×1) structure at approximately 200–250 K, in good agreement with experiments.

3.2.1.2 The {100} Surfaces of Binary Compound Semiconductors and Electron Counting Covalent bonds in a binary compound semiconductor are partially ionic because the electronegativities are smaller for the metal atoms (cations) than for the nonmetal atoms (anions). In a zincblende structure, bulk {100} planes are alternately occupied by cations and anions, and so an ideal {001} surface will be completely terminated by either cations or anions, leading to a polar surface. Polar surfaces, especially the {100}-oriented ones, have been widely used in homo- and hetero-epitaxial growth of compound semiconductors because the sticking coefficient of anion molecules depends strongly on the cation surface concentration. For example, the sticking coefficient of As_2 molecules on a GaAs(001) surface is unity if the surface is Ga-terminated but zero if it is As-terminated (Arthur, 1974; Foxon and Joyce, 1977). During MBE growth, the surface composition can be controlled from anion- to cation-rich by adjusting growth conditions, *e.g.*, the substrate temperature, the growth rate, and the ratio of the respective fluxes. All

{100} surfaces of III–V and II–VI compound semiconductors display reconstructions depending on surface composition.

The GaAs(001) surface has attracted much attention because of its potential importance in optoelectronic devices. As the As-rich GaAs(001) surface is heated from room temperature to above 450 °C, the following sequence of reconstructions is observed (Cho, 1971; Drathen *et al.* 1978; Massies *et al.*, 1980):

$$\begin{aligned} \mathrm{c}(2\times 8) &\xrightarrow{\approx 475\,^{\circ}\mathrm{C}} 1\times 6 \longrightarrow \mathrm{c}(6\times 4) \\ &\xrightarrow{\approx 575\,^{\circ}\mathrm{C}} 3\times 1 \longrightarrow 4\times 1 \\ &\xrightarrow{\approx 650\,^{\circ}\mathrm{C}} \mathrm{c}(8\times 2). \end{aligned} \tag{21}$$

The sequence is correlated with the continuous decrease of the As concentration in the surface due to thermal desorption of As_2 molecules without evaporation of Ga (Arthur, 1974). A reverse sequence is observed (van Bommel *et al.*, 1978) when cooling down the Ga-rich surface:

$$\mathrm{c}(8\times 2) \xrightarrow{\approx 450\,^{\circ}\mathrm{C}} \mathrm{c}(6\times 4) \xrightarrow{\approx 350\,^{\circ}\mathrm{C}} \mathrm{c}(2\times 8), \tag{22}$$

which is attributed to surface segregation of As from inside the bulk (Arthur, 1974; Neave and Joyce, 1978). In the As-rich surfaces, three different forms of (2 × 4) and/or c(2 × 8) reconstructions may occur depending on growth conditions, and they are denoted as (2 × 4)-α, -β, and -γ structures (Farrell and Palmstrøm, 1990). If an excess amount of As is present on the surface beyond one monolayer (ML) coverage, the surface displays a c(4 × 4) reconstruction (Neave and Joyce, 1978) and eventually becomes (1 × 1) when a thin film of As of a few monolayers thick is formed. All {100} surfaces of III–V and most of the II–VI compound semiconductors except ZnSe behave like GaAs(001) (Cornelissen *et al.*, 1988).

STM images of As-rich GaAs(001) surfaces (Biegelsen *et al.*, 1990a, 1990b) reveal that the basic features of the surface reconstructions are dimers and dimer vacancies (missing dimers). At different coverages, different reconstructions arise from the specific density, the arrangement, and the orientation of dimers and dimer vacancies. The three forms of (2 × 4)/c(2 × 8) reconstruction on the As-rich surface correspond to coverages of 0.5, 0.75, and 1 ML of As, respectively; the c(4 × 4) reconstruction corresponds to a As coverage of 1.75 ML.

The formation of dimers and dimer vacancies, and hence the reconstructions, can be understood by an electron-counting model (Chadi, 1987a; Pashley, 1989). As for elemental semiconductors, the dimerization on {100} surfaces of compound semiconductors reduces the number of dangling bonds. The energy of dangling-bond states can be reduced by charge transfer between two dangling bonds to fill one dangling bond with a lone electron pair while emptying the other one. On elemental {100} surfaces, this energy redistribution is achieved by intra-dimer charge transfer through dimer buckling, while on compound {100} surfaces, it is achieved by charge transfer between the second-layer atoms and surface dimers through the formation of dimer vacancy rows. Because dangling-bond energies are lower on anions than on cations, charge transfer from dangling bonds on cations to those on anions is greatly favored over the reverse. On a cation-terminated surface, all the dangling bonds on surface cation dimers can be emptied by transferring their electrons to the second-layer anions that are exposed at the dimer-vacancy sites; on an anion-terminated surface, all the dangling bonds on surface anion dimers can be filled by obtaining extra electrons from the second-layer cations that are exposed at the dimer-vacancy sites. Consequently, all compound-semiconductor {100} surfaces exhibit a general (2 × N) reconstruction; the twofold periodicity results from the dimerization, and the N-fold periodicity results from the dimer vacancy row formation. The periodicity N defining the optimal density of dimer vacancies is determined by an electron-counting rule (Mönch, 1995).

Consider an anion-stabilized (2 × N) structure with M_a anion dimers and $N - M_a$ missing dimers per unit mesh. Each dimer will contain six electrons: two in the dimer bond itself and two each in the dangling bonds at each of the dimer atoms. These electrons are supplied by the surface anions as well as by the second-layer cations that are exposed to the surface at the vacancy sites. Because each cation and anion contributes, respectively, f_c and f_a electrons per bond, counting the total number of electrons in the unit mesh, we have

$$6M_a = 4M_a f_a + 4(N - M_a)f_c. \tag{23}$$

Since $f_c = 3/4$ and $f_a = 5/4$ for III–V semiconductors, and $f_c = 2/4$ and $f_a = 6/4$ for II–VI semiconductors, we obtain from Eq. (23)

$$M_a(\text{III–V}) = 3N/4; \quad M_a(\text{II–VI}) = N. \tag{24}$$

Because the value of M_a, *i.e.*, the number of dimers in the unit cell, must be an integer, the relations in Eq. (24) predict that the smallest unit mesh on anion-stabilized {100} surfaces is (2 × 4) for III–V systems and (2 × 1) for II–VI systems. Figure 7 shows a schematic diagram of the (2 × 4) reconstruction on As-terminated GaAs(001). Similarly, for a cation-stabilized ($N \times 2$) structure with M_c cation dimers, we have

$$2M_c + 8(N - M_c) = 4M_c f_c + 4(N - M_c)f_a, \tag{25}$$

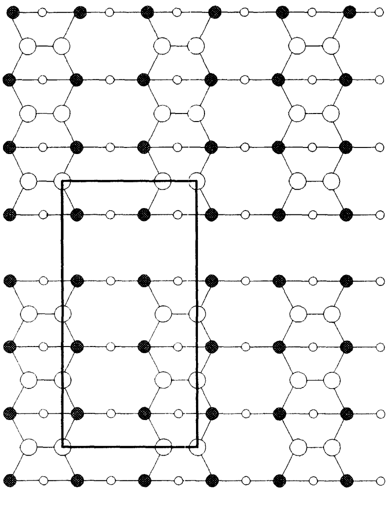

FIG. 7. Schematic top view of the (2 × 4)-β reconstruction on As-stabilized GaAs(001), induced by As vacancies. The large and small open circles denote As atoms in the first and third layers, respectively. The solid circles denote Ga atoms in the second layer. The solid rectangle marks the unit mesh.

and

$$M_c(\text{III–V}) = 3N/4; \quad M_c(\text{II–VI}) = N. \tag{26}$$

The smallest unit mesh on cation-stabilized {100} surfaces is (4 × 2) for III–V systems and (1 × 2) for II–VI systems. The acceptable values of N clearly depend on surface composition, which alters the number of electrons in the surface. So far, all the known reconstructions on the {100} surfaces of binary compound semiconductors satisfy the electron-counting rule (Mönch, 1995).

3.2.2 The {111} Surfaces The {111} surfaces of semiconductors, like the {100} surfaces, have also attracted much attention, the interest stemming from their complex structures, such as the well-known Si(111)-(7 × 7) reconstruction. In an ideally bulk-terminated {111} surface of diamond or zincblende structures, surface atoms have three nearest neighbors in the second layer and one half-filled dangling bond perpendicular to the surface [see Fig. 8(a)]. The elemental semiconductors Si and Ge cleave along {111} planes, and both as-cleaved surfaces exhibit a metastable (2 × 1) reconstruction. The ground-state structure of Si(111) is a (7 × 7) reconstruction and that of Ge(111) is a c(2 × 8) reconstruction. A simple rebonding (dimerization) mechanism like the one on Si(001) to reduce the number of dangling bonds is unlikely here because there is no easy direction for surface bonds to rotate. However, an adatom can effectively reduce the number of dangling bonds, two per adatom, by saturating three dangling bonds but introducing only one, a key ingredient on both the Si(111)-(7 × 7) and Ge(111)-c(2 × 8) reconstructions. For binary compound semiconductors, the {111} surfaces are fully terminated by either cations or anions, and the same (2 × 2) reconstructions are observed on different terminations, but with different structural origins.

3.2.2.1 The (2 × 1) Reconstruction on Cleaved Si(111) and Ge(111) and on Clean C(111) Both Si{111} and Ge{111} display a (2 × 1) reconstruction upon cleavage. The doubling of surface periodicity relative to the ideal surface (bulk {111} planes) occurs in

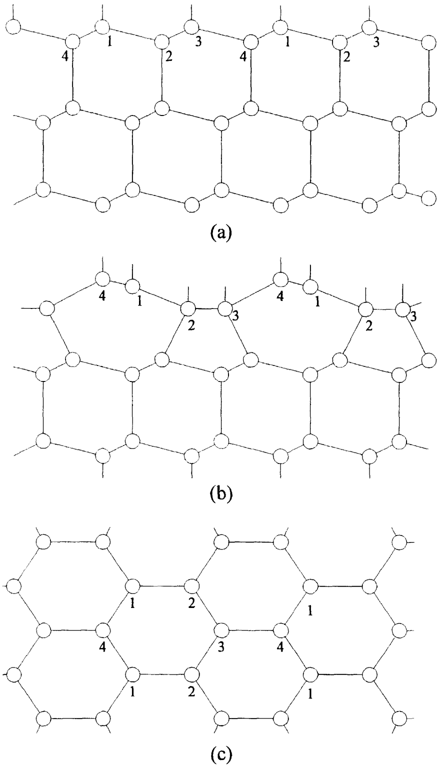

FIG. 8. π-bonded chain model of the (2 × 1) reconstruction on {111} surfaces of solids with the diamond structure. (a) Side view of ideally terminated surface. (b) Side view of the reconstructed surface. (c) Top view of the reconstructed surface.

⟨211⟩ directions, and the three equivalent ⟨211⟩ directions in the diamond lattice may give rise to three different orientations of (2 × 1) domains. In an early model (Haneman, 1961), the doubling of periodicity was thought to result from alternating rows of depressed and raised surface atoms, which was expected to lower surface energy by rehybridization of sp^3 orbitals on the surface atoms, inducing a charge transfer from the downward atoms with p-like dangling bonds to the upward atoms with s-like dangling bonds. This simple model, accepted until 1980, was found to be inconsistent with the measurements of surface electronic properties—in particular, the core-level shifts of surface atoms measured by photoemission (Himpsel *et al.*, 1980; Brennan *et al.*, 1980). A model proposed by Pandey (1982), which consists of zigzag π-bonded chains along ⟨110⟩ directions [Figs. 8(b) and 8(c)], was found to be consistent with these data. These chains are joined to the underlying bulk by five- and seven-member rings instead of the six-member rings that are characteristic of the diamond structure [see Fig. 8(b)]. The arrangement of surface atoms into such zigzag chains reduces surface energy because it conserves the total number of dangling bonds in the surface but makes them nearest neighbors and hence promotes some π-bonding among them rather than having them as second-nearest neighbors, as in the ideal or the earlier model surfaces. Surface energy is further reduced by tilting (or buckling) of chains, similar to buckling of dimers on Si(001) and Ge(001), as shown by both *ab initio* theories and various experiments (for a collection of theories and experiments, see e.g., Schlüter, 1988; Mönch, 1995).

The chains can be easily formed by simultaneously depressing a surface atom [atom 3 in Fig. 8(b)] into the second layer and raising a second-layer atom [atom 4 in Fig. 8(b)] into the surface layer, with a very small barrier of $\leq 0.03\,\text{eV}$ (Northrup and Cohen, 1982). The reconstructed surface is about 0.3–0.4 eV per atom more stable than the ideal surface (Pandey, 1982; Northrup and Cohen, 1982). The small formation barrier explains why the chain structure can be readily formed during cleavage. The formation process may be mediated by generation of stacking faults that involves formation of five- and seven-member rings (Reichardt, 1991).

Clean diamond {111} surfaces, which are prepared by annealing polished samples in ultrahigh vacuum in the temperature range of 950–1200 °C, show a (2 × 2) LEED pattern (Marsh and Farnsworth, 1964; Lander and Morrison, 1966). The (2 × 2) pattern is interpreted as the superposition of diffraction patterns from more than one of the three possible (2 × 1) domains present in the surface. [Occasionally, a (2 × 2) pattern also occurs on Si(111) and Ge(111).] So, similar to the {111} surfaces of Si and Ge, the {111} surfaces of C display a (2 × 1) reconstruction, which can be explained by Pandey's chain model (Pandey, 1982). However, differences exist between the {111} surfaces of C and those of Si or Ge. In C(111), the chains are flat, as shown by total-energy calculations (for a list of theories, see

Mönch, 1995), and are purely stabilized by strong π-bonding between C dangling bonds. In Si(111) [or Ge(111)], however, the π-bonding between dangling bonds is much weaker, and it alone is not sufficient to stabilize the chain structure. Consequently, the chains are buckled and the structure is primarily stabilized by the buckling-induced charge transfer between surface atoms rather than by π-bonding (Badziag and Verwoerd, 1988).

3.2.2.2 The Si(111)-(7 × 7) and Ge(111)-c(2 × 8) Reconstructions The (2 × 1) reconstructions on the as-cleaved Si(111) and Ge(111) surfaces transform irreversibly into (7 × 7) and c(2 × 8) structures, respectively, after annealing at elevated temperature (Lander *et al.*, 1963), confirming the metastability of the (2 × 1) reconstruction. The calculated surface energy of Si(111)-(7 × 7) is indeed lower than that of Si(111)-(2 × 1) (Brommer *et al.*, 1992). The determination of atomic structure involving large surface unit meshes is generally difficult. It took almost thirty years to solve the Si(111)-(7 × 7) reconstruction after it was first observed (Schlier and Farnsworth, 1959). After many different experimental techniques had been applied and a variety of models had been proposed, a complete description of the structure was finally achieved by Takayanagi *et al.* (1985a,b), who used a comprehensive analysis of transmission electron-diffraction patterns and all the prior knowledge developed for this surface. They introduced the dimer–adatom–stacking fault (DAS) model, showing how three key ingredients involved in the reconstruction that had each been independently suggested earlier, the adatom, the stacking fault, and the dimer, contribute to the overall structure shown in Fig. 9.

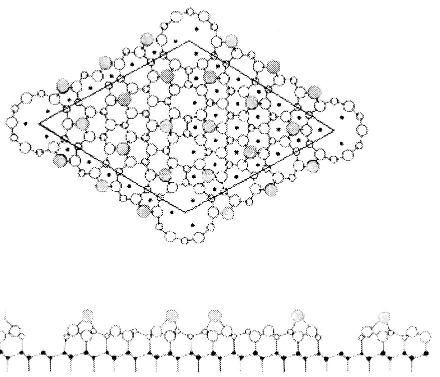

FIG. 9. (Top) Top view and (bottom) side view, taken along the long diagonal of the unit cell, of the dimer–adatom–stacking fault (DAS) model of the Si(111)-(7 × 7) reconstruction. The adatom layer (large shaded circles) and two double layers (open and solid circles) are indicated. The stacking fault is in the left half of the unit cell. (After Takayanagi *et al.*, 1985b.)

The (7 × 7) unit cell consists of two triangular subunits; one of them (the left half in Fig. 9) contains a stacking fault, producing wurtzite-type stacking of the outermost two double layers. There are six adatoms in each of the two subunits, in a (2 × 2) arrangement. They are located directly over the second-layer atoms and bind to three first-layer atoms. The faulted and unfaulted subunits are separated by a triangular network of partial dislocations, and dangling bonds along these dislocations are partially saturated by the formation of dimers (three dimers at each triangle edge), which are linked by eight-member rings. At the corners of the unit cell, *i.e.*, the crossing points of the dislocations, there are holes (corner holes) exposing large portions of the second double layer. Around the corner holes, atoms are connected by twelve-member rings.

The DAS model combines all the structural ingredients that had emerged from previous experimental studies. The adatom is the most effective way to reduce the dangling-bond density on a {111} surface of diamond structures (Harrison, 1976). However, incorporation of adatoms also introduces large strain in the surface. *Ab initio* theories (Northrup and Cohen, 1984; Northrup, 1986; Meade and Vanderbilt, 1989) show that the balance between the gain in electronic energy and the cost in strain energy produces, as the most stable arrangement of adatoms on Si(111), a (2 × 2) lattice in which adatoms sit on top of the second-layer atoms (in T_4 sites), with one rest atom (unsaturated surface atom) per (2 × 2) unit mesh. This arrangement has been seen directly by STM (Binnig *et al.*, 1983a; Becker *et al.*, 1985a; Hamers *et al.*, 1986b) in both triangle subunits of the (7 × 7) unit cell of Si(111). However, for Si(111), adatoms alone are unable to stabilize the surface. Additional structural changes must be incorporated; in particular, stacking faults (Bennett *et al.*, 1983) are introduced to create dimers at

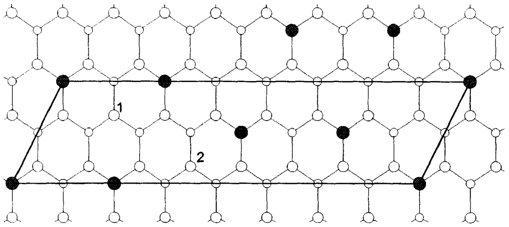

FIG. 10. Adatom model of the Ge(111)-c(2 × 8) reconstruction. Large and small open circles mark the first-layer and second-layer atoms, respectively. Shaded circles are Ge adatoms sitting on top of the second-layer atoms. Numbers 1 and 2 denote two inequivalent sites of rest atoms in the first layer. (After Becker *et al.*, 1989.)

the boundaries between faulted and unfaulted subunits (McRae, 1983; Himpsel, 1983) to reduce the dangling-bond density further. STM, equipped with spectroscopy (Tromp *et al.*, 1986), also reveals details of other structural features, such as the rest atoms, the corner holes, and the asymmetry between the two triangle subunits. Consequently, in the (7 × 7) structure, the overall electronic-energy gain (from the reduction of dangling bonds by a factor of 2.6 by creation of adatoms and dimers) offsets the strain-energy cost associated with creation of adatoms and with the formation of dimers and stacking faults, as shown by theoretical calculations (Qian and Chadi, 1987; Štich *et al.*, 1992; Brommer *et al.*, 1992).

The structure of the Ge(111)-c(2 × 8) surface, which involves only adatoms, is much simpler than that of the Si(111)-(7 × 7) surface. STM images (Becker *et al.*, 1985b, 1989) show that there are four adatoms and four rest atoms in the unit cell, as shown in Fig. 10. All adatoms are located at T_4 sites (van Silfhout *et al.*, 1990) while the rest atoms occupy two inequivalent positions (Hirschorn *et al.*, 1991). There is one dangling bond on each adatom and on each rest atom. The dangling bonds on adatoms are empty while those on rest atoms are filled (Becker *et al.*, 1989), indicating charge transfer from adatoms to rest atoms to reduce surface energy. The Ge dangling-bond energy is lower than that of Si. Also, Ge adatoms introduce less surface strain on Ge(111) than do Si adatoms on Si(111) because Ge is softer than Si (the elastic constants are approximately 40% larger for Si than for Ge). Consequently, for Ge(111), the introduction of adatoms alone is energetically favorable because the energy gain by reducing dangling-bond density by a factor of 2 more than offsets the adatom-induced energy cost in strain. The balance between the electronic-energy gain and the strain-energy cost is rather delicate in the (7 × 7) and c(2 × 8) reconstructions. Applying a compressive strain to a Ge(111) surface will convert the c(2 × 8) structure into the (7 × 7) structure, as demonstrated by the growth of thin Ge films on Si(111) (Gossmann *et al.*, 1985). Also, there exists a family of $[(2n-1)\times(2n-1)]$ DAS structures that employ the same underlying principle of energy balance as in the (7 × 7) structure. Depending on sample conditions and annealing temperatures, different domains with (5 × 5), (7 × 7), and (9 × 9) unit cells may appear on Si(111) (Becker *et al.*, 1986). The detailed mode of reconstruction can be controlled by varying the lateral stress (strain) in the surface. For example, the (5 × 5) reconstruction appears in the SiGe alloy surface under compressive stress as a thin SiGe alloy film is grown on a Si(111)-(7 × 7) substrate (Gossmann *et al.*, 1984), as well as in the Si(111) surface under tensile stress, obtained from a Si–Ge thin-film sandwich structures (Ourmazd *et al.*, 1986).

3.2.2.3 The {111} Surfaces of Compound Semiconductors The ideal {111}-(1 × 1) surfaces of binary compound semiconductors terminated by either cations or anions are conventionally denoted as (111) and $(\bar{1}\bar{1}\bar{1})$ surfaces, respectively. For III–V as well as II–VI compound semiconductors, both cation-terminated (111) and anion-terminated $(\bar{1}\bar{1}\bar{1})$ surfaces exhibit a (2 × 2) reconstruction (Ebina and Takahashi, 1982), but with different structural origins. The (2 × 2) structure of (111) surfaces arises from the formation of cation vacancies while that of $(\bar{1}\bar{1}\bar{1})$ surfaces arises from the formation of a regular array of anion trimers on top of the anion termination layer.

Most studies have been carried out on GaAs (111) and $(\bar{1}\bar{1}\bar{1})$ surfaces. In GaAs(111), each (2 × 2) unit mesh contains one Ga vacancy, exposing three As atoms in the second layer below the surface. Consequently, there are equal numbers (three) of Ga and As atoms in the unit mesh, and Ga dangling bonds are empty and As ones are filled as a result of the charge transfer from Ga to As. This structural and electronic configuration has been directly seen by STM (Haberern and Pashley, 1990). A LEED study (Tong *et al.*, 1984) shows that the exposed second-layer As atoms relax toward the center of Ga vacancies, leading to almost planar (on Ga atoms) and pyramidal (on As atoms) surface bonding configurations. The resultant atomic arrangement becomes similar to the tilted chains on nonpolar (110) surfaces of compound semiconductors (see discussion below). Total-energy calculations (Chadi, 1984) reveal that the formation of Ga vacancies on GaAs(111) is exothermic but the formation of As vacancies on GaAs($\bar{1}\bar{1}\bar{1}$) is endothermic. The vacancy-induced (2×2) reconstruction is also observed on similar systems, such as GaP(111), InSb(111) and GaSb(111) (Xu *et al.*, 1985; Bohr *et al.*, 1985; Feidenhans'l *et al.*, 1987).

Different models have been proposed for the GaAs($\bar{1}\bar{1}\bar{1}$)-(2×2) reconstruction. It cannot originate from As vacancies because of their large positive formation energy (Chadi, 1984). An anion-cluster model that consists of As trimers sitting on top of a complete terminating As layer was suggested from STM (Biegelsen *et al.*, 1990c). This structural model satisfies the electron-counting rule (Pashley, 1989) discussed in Sec. 3.2.1.2, and it possesses a lower surface energy than many other proposed models, as shown by an *ab initio* calculation (Kaxiras *et al.*, 1986). The anion trimer model is also supported by a TEM study of InSb($\bar{1}\bar{1}\bar{1}$) (Nakada and Osaka, 1991).

3.2.3 The {110} Surfaces The {110} surfaces of elemental semiconductors have complex structures, most notably superstructures involving steps. For example, a (16×2) reconstruction, which consists of 25-Å-wide stripes of singular terraces separated by up and down monatomic-height steps running along ⟨112⟩ directions, has been observed on both Si(110) (Yamamoto, 1994) and Ge(110) (Ichikawa *et al.*, 1995). A similar reconstruction with all the steps in the same sense (*i.e.*, all steps up or down) also forms on Si(110) (Yamamoto, 1994), leading to a vicinal facet of (17, 15, 1) structure.

Both III–V and II–VI semiconductors that crystallize in the zincblende structure cleave naturally along {110} planes, because these planes contain an equal number of cations and anions and thus are intrinsically neutral. The ideal surface consists of planar zigzag chains of alternating cations and anions along the [$\bar{1}$10] direction; each surface atom remains threefold coordinated, with one dangling bond [Fig. 11(a)]. The surface displays no reconstruction but sizable relaxations. The simplest relaxation at the {110} surfaces of a zincblende crystal is the tilting (buckling) of the zigzag chains (MacRae and Gobeli, 1966; Levine and Freeman, 1970), as shown in Fig. 11(b). Because bond energies change less with changing bond angle than with changing bond length, the buckling of the chain involves mainly bond rotations with bond length being approximately conserved, as shown by a collection of experimental results on {110} surfaces of III–V, II–VI, and even I–VII compounds (Mönch, 1995). Similar to buckled dimers on Si(100) or Ge(100), the buckled chains reduce surface energy by charge transfer from the downward atoms to the upward atoms because the dangling bond on the lower atoms becomes more p-like (with backbonds more sp^2-like) while that on the upper atoms becomes more s-like (with backbonds more p-like). Also, the dan-

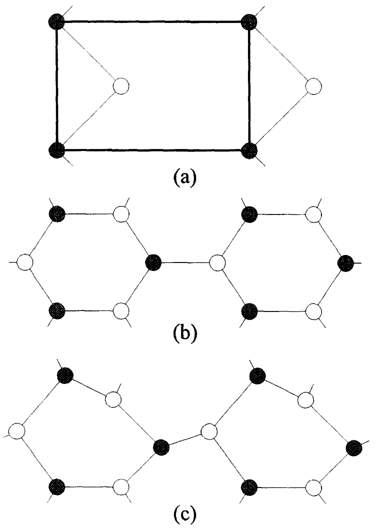

FIG. 11. Schematic atomic arrangements at {110} surfaces of zincblende-structure compound semiconductors. Open and shaded circles denote cations and anions, respectively. (a) Top view. (b) Side view of ideal bulk-terminated surface. (c) Side view of the relaxed surface.

gling-bond energy is lower on anions than on cations (see Sec. 3.2.2.3), so that all the cations on the {110} surfaces of compound semiconductors move downward and all the anions move upward [Fig. 11(b)], leading consistently to a charge transfer from the downward cations to upward anions. A correlation exists between buckling angle and ionicity of the underlying compound. For compounds with large ionicity, substantial charge transfer can be induced by the large difference in electronegativities between surface cations and anions, without a need for much buckling of chains. Thus, in general, the angle of buckling is found to decrease with increasing ionicity (Mönch, 1995).

3.3 Metal Surfaces

The nature of chemical bonding in metals differs from that in semiconductors. Covalent bonds are directional, while metallic bonds are more or less isotropic. Consequently, atoms in a semiconductor have to bond with each other in specific orientations, leading to a tetrahedral network and the diamond and zincblende structures; atoms in a metal, in contrast, bond as much as possible closely packed, forming face-centered cubic (fcc), hexagonal close-packed (hcp), or body-centered cubic (bcc) crystal structures, the three simple 3D lattices with highest atom number density. The fcc and hcp structures represent the closest packing of spheres in three dimensions, both having twelve nearest neighbors, while the bcc structure has eight nearest neighbors but six nearly as close second-nearest neighbors, with a separation only slightly larger than the first-nearest-neighbor distance. Such close packing maximizes the electron-density overlap. Different packing sequences result from different electronic structures of the constituent metal atoms.

The difference in bonding between semiconductors and metals also strongly influences the structural properties of the respective surfaces. As just described, in semiconductor surfaces, removal of atoms at the vacuum side produces a great number of dangling bonds, and thus a strong tendency for reconstruction to reduce their number. This driving force for reconstruction disappears in metal surfaces because there are no dangling bonds. Metal surfaces generally exhibit less tendency to reconstruct than semiconductor surfaces.

An empirical estimate of the magnitude of surface relaxation and the tendency to reconstruct for systems of close-packed spheres can be obtained by comparing the interatomic spacing to the diameter of a free atom. For example, rare-gas crystals can be approximated as the close packing of 'hard' spheres. The atoms bond with each other *via* the weak van der Waals interaction, and the atomic radii hardly change in the solid. Consequently, there is very little relaxation (less than a few percent) and no reconstruction of rare-gas crystal surfaces. Metallic bonds are much stronger than the van der Waals interaction; consequently interatomic spacings in a metal are less than the diameters of free metal atoms and metal surfaces have a stronger tendency to relax. Some metal surfaces also reconstruct, optimizing the interatomic spacings and/or the coordination numbers of surface atoms in order to minimize the surface free energy.

A more detailed understanding of the structure of metal surfaces than the sphere packing model requires knowledge of the electronic properties of metals. They can be described in the simplest limit by a jellium model, in which a periodic array of positive ions is embedded in the uniform density of a free-electron gas. Two theoretical approaches in the spirit of the jellium model, the embedded-atom method (EAM; see, *e.g.*, Foiles, 1987) and the effective-medium theory (EMT; see, *e.g.*, Jacobson and Nørskov, 1988) have been widely used to determine many properties of metals and metal surfaces. At equilibrium, metal atoms are located at sites of most favorable electron density to minimize their free energy. At the surface the electron density is reduced because of the removal of neighbors on the vacuum side. The most intuitive (but not always correct) conclusion one draws is that the atoms in the outermost surface layer can reduce their energy by relaxing inward to a higher electron density. Such inward relaxation frequently does occur. For example, the spacing of the two outermost layers of unreconstructed fcc (110) metal surfaces is reduced between 5.1% in Pd(110) (Schottke *et al.*, 1987) and 17.1% in Pb(110) (Breuer *et al.*, 1990) relative to the bulk value.

A relaxation in the opposite direction, *i.e.*, an increase in the separation of the two outer layers, is also possible, as observed, *e.g.*, in the Pt(111) surface (Materer *et al.*, 1995). Metals are frequently not well described by a jellium model, *i.e.*, the electron density does not decrease uniformly at the surface. Electronic surface states can introduce a local maximum in the electron density at a position outside the regular 3D lattice position of the outermost layer of atoms. The lowest-energy state of the surface may then involve an outward relaxation.

The electron density at the surface can also be optimized by reducing the interatomic spacing within the surface layer or by increasing the coordination number of the surface atoms, either by incorporating extra atoms into the surface layer or by rearrangement of the atoms. These effects produce reconstructions. A reconstruction may be caused by large atomic displacements or by atomic displacements so small that the reconstruction is difficult to observe. We restrict ourselves here to reconstructions that are unambiguously identified.

The desire of the surface atoms to reduce their lateral spacing leads to tensile stress in not yet reconstructed surfaces (this means that the surface energy would be reduced if the crystal were compressed). For example, *ab initio* calculations (Feibelman, 1995) have shown an excess surface stress in Pt(111) of about $0.25\,\mathrm{eV}\,\mathrm{\AA}^{-2}$. Whether the excess surface stress is strong enough to induce a reconstruction depends on its value and on the energy cost in creating the reconstruction. For example, atomic rearrangement, including the incorporation of additional atoms within the surface layer, displaces surface atoms from their regular lattice positions. Because atoms in the topmost layer will lose some of their nearest neighbors in the second layer, the interaction between the topmost layer and the second layer is weakened. Whether the surface shows a reconstruction depends on the balance between the energy gain from increasing the atom number density in the topmost layer and the energy loss from weakening of the interaction between the topmost and second layer.

Most metal surfaces exhibit large relaxation but only some surfaces, including some orientations of the bcc metals W and Mo and of the fcc metals Au, Pt, and Ir, reconstruct. Intuitively, the higher the atom number density in a surface layer, the lower is the tendency for the surface to reconstruct, because a higher coordination implies a more bulk-like environment. The atom number density of a surface plane depends on the surface orientation (see Fig. 12). Nevertheless, there are surfaces with highest atom number densities [*e.g.*, Au(111)] that do reconstruct.

Among bcc metals, the W(100) surface is unreconstructed at room temperature, but during cooling to 250 K W(100) reconstructs spontaneously to a $\sqrt{2} \times \sqrt{2}$ R45° structure (Fig. 13) (see, *e.g.*, Pendry *et al.*, 1988). Figure 13(b) shows that the lateral displacement of the atoms in the first layer leads to zigzag chains. The atoms in the second layer are also slightly displaced, by about 20% of the displacement of the atoms in the topmost layer (Altmann *et al.*, 1988). Additionally the topmost layer shows an inward relaxation of about 6%. *Ab initio* calculations (Singh and Krakauer, 1988) have identified the gain in energy due to an increase in the effective coordination of the surface atoms as the driving force towards the reconstruction. Mo(100) exhibits a similar reconstruction at lower temperatures, but with a seven times larger unit cell (Daley *et al.*, 1993).

The fcc 5d transition metals (Au, Pt, Ir) display a variety of reconstructions on different low-index surfaces. In the following we discuss, as an example, the general trend of surface relaxation and reconstruction on these 5d metals.

3.3.1 The fcc (110) Surface

Many experiments (for a list of these experiments, see, *e.g.*, Rous, 1995) have shown that the (110) surface of fcc 5d metals has a strong tendency to reconstruct. Figure 14(a) shows a schematic diagram of the unreconstructed surface. Atoms in the topmost layer have only seven nearest neighbors, a great reduction from the bulk value of twelve. Atoms in the second layer also are missing one nearest neighbor, which would sit on top of them. Even more importantly, an atom located in the (110) plane has only two in-plane nearest neighbors along the [110] direction. In the perpendicular direction there are no in-plane nearest neighbors, leading to a very weak interaction between atoms in different [110] chains. Chains of atoms in such a configuration can be rearranged easily without losing nearest-neighbor bonds.

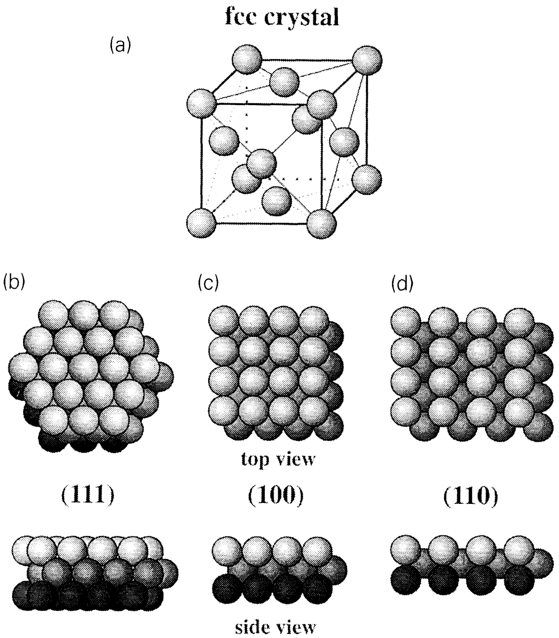

FIG. 12. Surfaces of fcc crystals. (a) Model of the fcc crystal. (b)–(d) Top and side views of the (111), (100), and (110) surfaces, respectively.

The large potential for reconstruction leads to the missing-row structure, observed, *e.g.*, on Au(110) and Pt(110). Figure 14(c) shows a schematic view of the missing-row reconstruction. Counting all atoms exposed to vacuum in the top, second, and third layers as surface atoms, the reconstruction changes neither the average coordination of surface atoms nor the surface atom-number density. However, the distribution of coordination of surface atoms has changed. In the unreconstructed surface [Fig. 14(b)], half of the surface atoms (in the topmost layer) lose five neighbors, and the other half (in the second layer) lose one neighbor; in the reconstructed surface [Fig. 14(d)], a fourth of the surface atoms (in the topmost layer) lose five neighbors, one half (in the second layer) lose three neighbors, and another fourth (in the third layer) lose one neighbor. The missing-row reconstruction transforms half of the atoms missing five or one neighbors into ones missing three neighbors by breaking some second-nearest-neighbor bonds between rows. Calculations have shown that this more uniform coordination leads to a reduction in free energy. The free-energy reduction for the missing-row reconstruction can be visualized by recognizing that the reconstruction is equivalent to forming small (111) facets (Binnig *et al.*, 1983b), the most stable surface of fcc crystals.

Surfaces of fcc (110) metals exhibit a systematic correlation between formation of a missing-row reconstruction and electronic properties. Whereas all 5d fcc transition metals reconstruct, the 4d fcc metals and the 3d transition metals show no tendency to reconstruct. The trend can be understood by comparing the energy cost of forming the missing-row reconstruction with the energy gain. The cost of

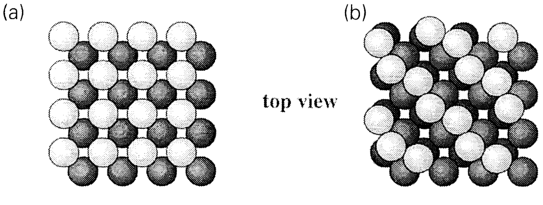

FIG. 13. Top view of the W(100) surface: (a) unreconstructed and (b) $\sqrt{2} \times \sqrt{2}$ reconstructed.

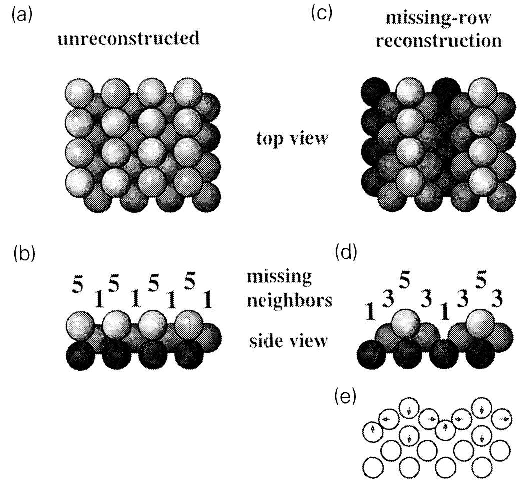

FIG. 14. Reconstruction of fcc (110) surfaces. (a), (b) Top and side views of the unreconstructed surface same as Fig. 12(d). (c), (d) Top and side views of the 2×1 missing-row reconstructed surface. (e) A schematic side view of the relaxation of the Au(110)-(2×1) missing-row reconstruction as proposed by Moritz and Wolf (1985).

breaking the second-nearest-neighbor bonds between neighboring rows is balanced by the gain produced by a more uniform distribution of the number of nearest neighbors. Thus, the gain is related to a more localized effect (the nearest-neighbor bonds) than is the cost (the second-nearest-neighbor bonds). The electrons of the 5d metals are very localized, so that the first-nearest-neighbor effect dominates over the second-nearest-neighbor effect, favoring the reconstruction. As one moves from 5d to 4d and 3d metals, electrons become less and less localized. As a result, the second-nearest-neighbor effect becomes dominant, preempting the reconstruction. This trend is reproduced qualitatively by many calculations (see Woodruff, 1994, for a selection of theoretical publications).

An interesting question is why the (110) surface does not form larger (111) facets, leading to ($N \times 1$) ($N > 2$) instead of (2×1) missing-row reconstructions. The simple model described above would lead to larger facets because the energy gain would be greater. The cause lies in the atomic relaxation. As shown in Fig. 14(e), the relaxation involves atomic displacements both perpendicular and parallel to the surface (Fery *et al.*, 1988). For larger (111) facets the relaxation would have to be mainly a uniform displacement of atoms perpendicular to the (111) facet. In certain cases, such as Au(110), calculations have shown that there is only a slight energy advantage for the 2×1 missing-row reconstruction relative to ($N \times 1$) structures with larger (111) facets [about 10 meV per atom (Ho and Bohnen, 1987)]. In agreement with these calculations, the Au(110) surface has an intrinsic amount of disorder involving (3×1) and (4×1) reconstructed regions (Binnig *et al.*, 1983b; Gritsch *et al.*, 1991). On Ir(110) a (3×1) reconstruction has been observed (Naumovets *et al.*, 1975).

3.3.2 The fcc (100) Surface In the (100) surface, atoms have eight nearest neighbors, four of them within the surface plane. The (100) surface of metal crystals is expected to have a smaller tendency to reconstruct than the (110) surface. Nevertheless all three 5d transition metals have reconstructed (100) surfaces (for a list of relevant experiments, see, *e.g.*, Rous, 1995). In contrast to the reconstruction of the (110) surface, the reconstruction of the (100) surface is controlled by gain and cost both related to nearest-neighbor bonds. The details of the underlying mechanisms are not yet known.

Both Pt(100) and Au(100) form a hexagonal surface layer on top of the square (100) lattice (Fig. 15) (Abernathy *et al.*, 1992). The

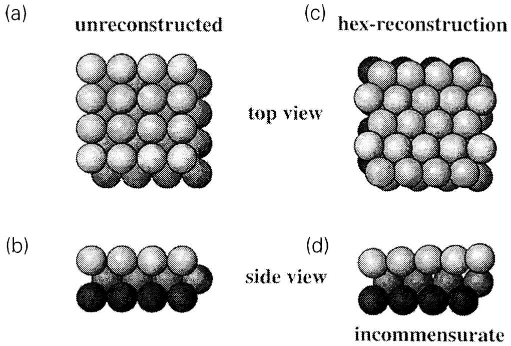

FIG. 15. Reconstruction of fcc (100) surfaces. (a), (b) Top and side views of the unreconstructed surface same as Fig. 12(c). (c), (d) Top and side views of the hexagonally reconstructed surface. The hexagonal layer has a periodicity incommensurate with that of the substrate lattice.

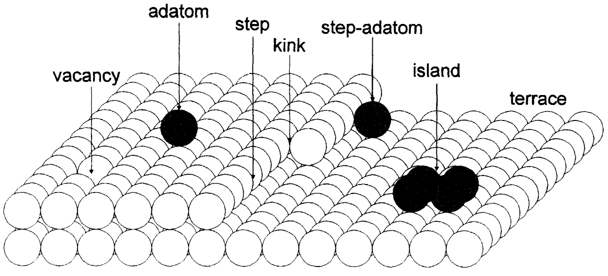

FIG. 16. The TSK model of a surface defined for a simple cubic crystal. Open circles are substrate atoms. Shaded circles are adatoms deposited on the substrate. There are two terraces (the upper terrace is on the left) separated by a step containing a single-atom kink.

atom number density in this hexagonal outermost layer is 8% higher than in a regular (111) plane. This enhanced atom number density lowers the surface free energy. Such a layer is generally higher-order commensurate or incommensurate with the underlying bulk. The energy cost of this type of reconstruction is the reduced interaction with the second layer, because the incommensuration leads to a lower average coordination with the second layer.

Ir(100) reconstructs by forming a quasihexagonal (1×5) reconstruction (Lang *et al.*, 1983). In one close-packed direction, $\langle 110 \rangle$, the interatomic spacing remains unchanged, whereas along the other direction the structure is highly compressed. The surface layer remains commensurate with the second layer. Because more atoms are located at (or near) their regular lattice sites, the weakening of interactions between the topmost layer and the second layer is not as great as in the incommensurate structures of Pt(100) and Au(100). On the other hand, the bonds within the outermost layer are not as optimized as those in an incommensurate structure.

3.3.3 The fcc (111) Surface The (111) surface is the most stable and inert surface of fcc crystals, in which surface atoms have nine nearest neighbors, three fewer than in the bulk. The reconstructions on Au(100) and Pt(100), with their hexagonal outer layers having about 8% higher atom number density than a regular (111) plane, suggest that a (111) surface reconstruction that increases the outer-layer atom number density would lower the free energy.

Au(111) reconstructs by forming the $(22 \times \sqrt{3})$ herringbone structure that allows the surface to increase the atom number density by about 5% (Perderau *et al.*, 1974; Barth *et al.*, 1990). Pt(111) reconstructs by forming a similar structure in equilibrium at high temperatures (≈ 1300 K) (Sandy *et al.*, 1992). The reconstruction of Pt(111) is also observed during homoepitaxial growth at lower temperatures (Hohage *et al.*, 1995). Calculations (Needs *et al.*, 1991) show that the excess surface stress causes a general tendency for the fcc 5d metals to reconstruct in this manner.

4. MORPHOLOGICAL FEATURES OF REAL SURFACES, INTERFACES AND THIN FILMS

4.1 The Terrace–Step–Kink (TSK) Model

Real surfaces are not perfect but contain defects, such as steps, kinks at steps, adatoms, and vacancies, which are generally described in terms of the terrace–step–kink (TSK) model (Burton *et al.*, 1951), as shown in Fig. 16. These morphological features not only influence surface properties but also play a fundamental role in thin-film growth (see THIN-FILM DEPOSITION). The STM has allowed the direct visualization of the TSK model (Swartzentruber *et al.*, 1990; Wang *et al.*, 1990) and has made possible the quantitative determination of kinetic and thermodynamic properties of morphological features ranging from a single adatom to micrometer-sized terraces.

At thermal equilibrium, the concentration of surface defects is determined by minimization of free energy, *i.e.*, the competition between their formation energy and their configurational entropy. The formation energy of a particular type of defect can be determined by measuring its concentration as a function of temperature. For example, surface steps meander by forming kinks. Step energies and kink energies on Si(001) have been determined by STM from the distribution function of kinks along the steps (Swartzentruber *et al.*, 1990). Beyond a critical temperature, the surface roughening temperature, surface steps form spontaneously, as the entropy term, which increases linearly with temperature, becomes equal to the step formation energy.

Surface defects may induce both local structural and long-range morphological changes. For example, steps and dimer vacancies on Si(001) "freeze" the motion of buckled dimers (see Sec. 3.2.1.1) between the two equivalent positions at room temperature, as a means to release surface strain energy (Wolkow, 1992). Step–step interactions (*e.g.*, entropic repulsion) influences kink formation on steps; the meandering of an individual step is independent of that of other steps only when steps are far separated (Swartzentruber *et al.*, 1990; Wang *et al.*, 1990; Zandvliet *et al.*, 1992). The kinetics of step motion (Kitamura *et al.*, 1993a) and vacancy diffusion (Kitamura *et al.*, 1993b) have also been measured with STM at high temperatures. A novel mechanism for the vacancy diffusion, involving a rolling motion at surface and subsurface atoms, has been proposed (Zhang *et al.*, 1993).

4.2 Equilibrium Step Configurations

4.2.1 Stress-Domain Structure Surface stress plays an important role in determining surface structure and morphology. Theories (Marchenko, 1981; Alerhand *et al.*, 1988) predict that if the intrinsic surface stress is anisotropic, a single-domain surface is unstable against formation of alternating degenerate stress domains. For example, the (2×1) reconstruction of Si(001) introduces a highly anisotropic surface stress: the stress $\sigma_{||}$ along the dimer bond is tensile (*i.e.*, surface atoms would like to be closer together along this direction than they are); the stress $\sigma_{\perp}$ along the dimer row is consequently compressive (or at least less tensile than $\sigma_{||}$). Because of the tetrahedral bonding configuration in the diamond structure, the dimer directions are orthogonal on terraces separated by monatomic steps, giving rise to orthogonal (2×1) domains on adjacent terraces, which are referred to as (2×1) and (1×2). The intrinsic stress anisotropy, $F = \sigma_{||} - \sigma_{\perp}$, causes a morphological instability (Marchenko, 1981; Alerhand *et al.*, 1988): a flat Si(001) surface is unstable against formation of monatomic steps as domain boundaries separating alternating (2×1) and (1×2) domains. Such stress domains are quite generally possible in systems that lower their surface energy by reconstruction.

The discontinuity of stress at the step introduces a force monopole, giving rise to an elastic step–step interaction that lowers the surface energy with a logarithmic dependence on step separation (terrace width, L) (Marchenko, 1981; Alerhand *et al.*, 1988), $E_s = C_1 - C_2 \ln(L/a)$, where C_1 denotes the step-formation energy per unit length, C_2 reflects the strength of the interaction, which is related to the stress anisotropy and elastic constants, and a is a microscopic cutoff length on the order of a lattice constant. The surface assumes the lowest-energy configuration with an optimal step separation of $L_0 = \pi a \exp(C_1/C_2 + 1)$. Therefore, a surface with sufficiently low step density could reduce its energy by introducing extra steps. However, the spontaneous formation of extra steps to reduce the size of the stress domain predicted by theory is not observed on nominally singular Si(001). Instead, the existing steps are observed to become wavy in the low–step-density regions (Tromp and Reuter, 1992). Step undulations lower the surface energy, just as creation of new steps does, by effectively reducing the size of the stress domains (Tersoff and Pehlke, 1992). Step undulations preempt the formation of extra steps because the undulations are kinetically preferred and are compatible with step flow.

4.2.2 Steps on Vicinal Surfaces Vicinal surfaces—in particular, vicinal Si(001)—play important roles in epitaxial growth and device fabrication, as templates for growing smooth thin films. The step morphology is frequently related to the vicinality of the surface. The behavior of steps on vicinal Si(001) demonstrates the range of morphological phenomena related to steps that can be encountered. Most effort has been dedicated to surfaces miscut toward a $\langle 110 \rangle$ direction, because in these directions, the dimer rows on the surface are either parallel or perpendicular to the steps. On vicinal

Si(001) surfaces with a small miscut ($<1.5°$) toward $\langle 110 \rangle$, the misorientation produces only monatomic steps, which separate alternating (2×1) and (1×2) domains constituting the stress-domain structure. By convention, the (2×1) notation is chosen to describe the domains that have the dimer rows perpendicular to the steps. If the surface is nominally singular (miscut by less than a few tenths of a degree), the equilibrium populations of these two domains are about equal (Men *et al.*, 1988). When an external stress is applied to such a surface, the relative populations of the two domains can be changed by straining the surface at a temperature sufficiently high so that the step mobility is high (Men *et al.*, 1988). The domain compressed along the dimer bond is favored. The experiment is well understood by the theoretical model (Alerhand *et al.*, 1988) mentioned in the last section. The response of the relative domain population to an external stress provides a unique way to determine the intrinsic surface stress anisotropy (Webb *et al.*, 1990, 1991). For Si(001), the anisotropy is $0.07 \pm 0.01\,\mathrm{eV\,Å^{-1}}$, as obtained from LEED and STM measurements (Webb *et al.*, 1990, 1991). *Ab initio* calculations (Garcia and Northrup, 1993; Dabrowski *et al.*, 1994) have produced a value in good agreement with the experiment.

Step configurations on high-miscut vicinal Si(001) surfaces are more complicated and their behavior is less well understood. As the miscut angle toward $\langle 110 \rangle$ increases, the populations of (2×1) and (1×2) domains begin to differ, with the population of (2×1) domains increasing gradually with increasing miscut angle (Tong and Bennett, 1991; Pehlke and Tersoff, 1991; Swartzentruber *et al.*, 1992). At sufficiently high miscut ($\approx 4°$) double-atomic-height steps start to form (Wierenga *et al.*, 1987; Swartzentruber *et al.*, 1992). The formation energy of double-atomic-height steps is lower than that of the monolayer-high steps (Chadi, 1987b), but, at a sufficiently small miscut angle, single-atomic-height steps may be stabilized by their elastic step–step interactions (Vanderbilt *et al.*, 1989). Several model calculations (Alerhand *et al.*, 1990; Poon *et al.*, 1990; Pehlke and Tersoff, 1991) predict the existence of a first-order phase transition from the single-atomic-height-step surface to the double-atomic-height-step surface as the miscut angle increases. Experimentally, however, no indication of a first-order phase transition has been observed. Instead, the step concentration changes continuously with both miscut angle (Swartzentruber *et al.*, 1992) and temperature (de Miguel *et al.*, 1991). Experiments (de Miguel *et al.*, 1992) also indicate that at high temperatures and/or large miscut angles, a simple 1D elastic model becomes unsatisfactory.

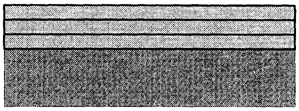
Frank-van der Merwe

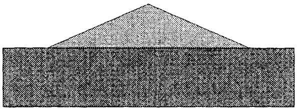
Volmer-Weber

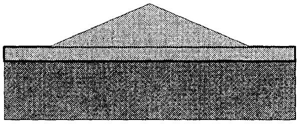
Stranski-Krastanov

FIG. 17. Three equilibrium growth modes in heteroepitaxy.

4.3 Surface/Interface Morphology of Epitaxial Thin Films

4.3.1 Equilibrium Growth Modes The surface morphology of a thin film grown on a substrate by vapor deposition is described, at thermal equilibrium, by three different growth modes, as depicted in Fig. 17. In the Frank–van der Merwe mode, growth proceeds layer by layer, maintaining a smooth film; in the Volmer–Weber mode, 3D islands form immediately and the film does not cover the whole substrate surface; in the Stranski–Krastanov mode, the film first grows layer by layer—*i.e.*, "wets" the surface—followed by the formation of 3D islands. The realization of different growth modes, and hence different thin-film morphologies, is determined again by minimization of surface free energy, in this case for the combined substrate/film system. The surface energy of the substrate ($\gamma_{\mathrm{substrate}}$), the surface energy of the film (γ_{film}), and the interface energy between the film and substrate ($\gamma_{\text{interface}}$) (Bauer, 1958) enter into this minimization. For heteroepitaxy, *i.e.*, the growth of material A on a different substrate B, the surface energy of the film (A) will differ from that of the substrate (B). If $\gamma_{\text{substrate}} < \gamma_{\text{film}} + \gamma_{\text{interface}}$, the deposited film tends to nucleate 3D islands, leaving the low-energy substrate exposed (the Volmer–Weber mode). The condition of $\gamma_{\text{substrate}} > \gamma_{\text{film}} + \gamma_{\text{interface}}$ defines the layer-by-layer or Frank–van der Merwe growth mode. However, γ_{film} increases with increasing film thickness because of misfit strain, which leads to Stranski–Krastanov growth: the film first grows layer by layer as long as $\gamma_{\text{substrate}} > \gamma_{\text{film}} + \gamma_{\text{interface}}$, forming a finite thickness of wetting layer, followed by formation of 3D islands as $\gamma_{\text{substrate}} < \gamma_{\text{film}} + \gamma_{\text{interface}}$ to relieve the misfit strain built up in the film. In general, if A wets B, then B will not wet A. For homoepitaxy, *i.e.*, the growth of material A on a substrate of the same material A, $\gamma_{\text{substrate}} = \gamma_{\text{film}} + \gamma_{\text{interface}}$. The equilibrium film grows layer by layer in the Frank–van der Merwe mode.

Stranski–Krastanov growth has attracted considerable interest recently because of its potential for fabricating 3D spatially confined structures (Liu and Lagally, 1997a). One typical example is the growth of Ge or SiGe alloy on Si. Because Ge has a lower surface energy than Si, Ge initially grows layer by layer, wetting the Si substrate. As the film gets thicker, the strain energy in the film increases, and eventually 3D islands start to form to relieve the 4% lattice mismatch between Ge and Si. The equilibrium surface morphology of a thin Ge film grown on Si(001) is rather complex (Liu *et al.*, 1997). For pure Ge films grown on Si(001), the thickness of the wetting layer is extended up to 3 ML by strain-relaxation mechanisms involving only 2D surface roughening processes before 3D islanding takes place (Mo *et al.*, 1990; Tersoff, 1991). The misfit strain is partially relaxed by the formation of dimer vacancy lines, the $(2 \times N)$ reconstruction (Chen *et al.*, 1994). The reconstruction also induces changes in surface (step) morphology and stress field, reversing the relative roughness of the two types of steps on the surface (Wu *et al.*, 1995) as well as the surface stress anisotropy (Wu and Lagally, 1995; Liu and Lagally, 1996; Liu *et al.*, 1997). The 3D islands form in two stages. Initial islands are small and coherent with the substrate lattice, *i.e.*, without dislocations (Eaglesham and Cerullo, 1990; Mo *et al.*, 1990). Larger dislocated islands form later.

4.3.2 Kinetics-Limited Growth Morphology

Equilibrium thin-film morphologies for different film systems can be achieved by growth at high temperature and low deposition rate. Very often, growth is carried out far from equilibrium, and consequently, the evolving surface morphology during thin-film growth is kinetically limited (Liu and Lagally, 1997b). The ability to control kinetic parameters to manipulate thin-film morphology allows us to grow novel artificial film structures with desirable properties. Surface diffusion is the most important kinetic parameter controlling growth morphology. Because the rate of diffusion is an exponential function of temperature, different regimes of growth that lead to different thin-film morphologies can be achieved by tuning the growth temperature, even for systems that belong to the Frank–van der Merwe growth mode, e.g., homoepitaxial growth.

Real surfaces have imperfections such as steps. As adatoms are deposited onto a surface, they are preferentially incorporated at steps, particularly at kink sites where they can bind to the most neighbors (Burton *et al.*, 1951). If all deposited adatoms are attached at steps, growth will proceed with advancement of steps, so called step-flow growth. Step-flow growth requires a high mobility of the adatoms so that they reach the existing steps before meeting with each other to nucleate islands. It is achieved, in homoepitaxy, by growing at high temperature, at low deposition rates, or on high–step-density surfaces. Step-flow growth, representing a situation close to thermal equilibrium, produces smooth films for homoepitaxial film growth. However, morphological instabilities, *e.g.*, the bunching of substrate steps, may arise from the additional kinetic barrier for adatoms crossing the steps (Ehrlich and Hudda, 1966; Schwoebel and Shipsey, 1966) or from complex dynamics of step flow (Mullins and Hirth, 1963; Kandel and Weeks, 1995). For step-flow growth of a strained film, a step-bunching instability caused by the elastic attractions between steps (Tersoff *et al.*, 1995) or by strain-modified adatom diffusion (Duport *et al.*, 1995) can occur.

At lower temperature, at high deposition rates, or on surfaces with a low step density, adatoms may meet each other in the middle of terraces before reaching the existing steps and nucleate stable 2D islands. As the deposition continues, the nucleated islands grow by incorporating new arriving adatoms, and more islands form by new nucleation events. Eventually, the 2D islands meet and coalesce. This regime is referred to as two-dimensional island growth. The 2D island growth mode is used for fundamental studies of kinetic mechanisms of homoepitaxial growth and to calibrate deposition rates in layer-by-layer growth, but serves little technological function.

The nucleation and growth of 2D islands can be described by phenomenological rate equations (Venables *et al.*, 1984). The properties of islands (*e.g.*, island number density, island compactness, and island shape) are largely determined by surface diffusion. The surface diffusion coefficient can be quantitatively determined by STM measurement of the island number density as a function of temperature at the initial stage of island growth (Mo *et al.*, 1991; Bott *et al.*, 1996; Liu and Lagally, 1997b). The compactness of islands (fractal-like versus close-packed) is controlled by how fast an adatom diffuses along island edges and crosses corners where two edges meet (Zhang, *et al.*, 1994; Hohage *et al.*, 1996; Zhang and Lagally, 1997). In general, growth at relatively lower temperatures produces less compact (more fractal-like) islands, because adatoms diffuse more slowly and thus do not reach their equilibrium positions before they become trapped by newly arriving atoms. On reconstructed surfaces, 2D island growth becomes more complex. For example, on (2×1) reconstructed Si(001), the large anisotropy in adatom sticking coefficients to two different edges (steps) of an island creates kinetics-limited islands with long and thin shapes. Reconstruction can also introduce additional roughness to the growth front. Two islands nucleating at different locations in the same terrace can form antiphase domain boundaries when they meet. For example, in Si(001) islands have a 50% chance to align their dimer rows out of phase (Hamers *et al.*, 1989), roughening the growth front. The GaAs(001) (4×2) structure has four different possibilities of island positions with only 1/4 of the islands meeting without antiphase boundaries. The larger the unit mesh, the more possibilities that exist.

Islands may nucleate not only on terraces but also on top of those islands that are already formed, when these become large enough so that atoms falling on them can no longer reach the edge. In addition to surface diffusion, the barrier for an adatom to jump off an island (*i.e.*, crossing a step) now becomes a key kinetic parameter in determining the film roughness, even for homoepitaxial films. The roughness of the film increases as the film grows thicker if atoms cannot cross the steps.

In heteroepitaxy, misfit strain can lead to various 2D morphological instabilities that roughen the surface, even in the absence of kinetic roughening, and are due to energetic (thermodynamic) driving forces. For example, as Ge is deposited on Si(001), the surface first roughens by formation of dimer vacancies, giving rise to the $(2 \times N)$ reconstruction. The dimer vacancy lines reduce adatom diffusion (Lagally, 1993), leading subsequently to additional roughness induced by new kinetic limitations (Mo and Lagally, 1991; Lagally, 1993; Wu and Lagally, 1996).

4.3.3 Interface Morphology The quality of the interface—in particular, the smoothness of the interface of a thin-film structure—frequently influences other properties, and may determine the usefulness of the film, for example, for electronic and magnetic applications. A strained film is inherently unstable; strain energy increases as the film grows thicker. Upon reaching the "equilibrium critical thickness" (van der Merwe, 1963; Matthews and Blakeslee, 1976), a strained film will relax back to its equilibrium lattice constant, forming dislocations at the interface, if the kinetics permit it. Films can be grown beyond the equilibrium critical thickness by growing at low enough temperatures so that the kinetics are so small that dislocation formation is suppressed. For homogeneous nucleation of dislocations, the kinetic barrier for nucleating a dislocation decreases with increasing strain (LeGoues *et al.*, 1993), so that eventually, as the misfit strain increases, dislocations always will form. A structural defect at the surface or interface may activate the nucleation of dislocations.

It is difficult to create a smooth interface, even without forming dislocations. Interdiffusion roughens the interface by creating a transition region in which a compositional gradient exists. In general, interdiffusion lowers the interface energy by increasing the configurational entropy. In those cases when $2E_{A-\mathrm{B}} > E_{A-A} + E_{\mathrm{B-B}}$, where E denotes the bond energy, there will also be a strong tendency for interdiffusion to form interface alloy compounds. Substantial interdiffusion can occur even at relatively low temperatures, at which bulk diffusion is negligible, through surface diffusion-assisted atom place exchange in the outer few layers, driven by surface segregation of the thermodynamically more stable component and/or by preferential occupation of different components at different lattice sites induced by surface reconstruction. For example, Ge surface segregation causes a decrease of Ge concentration at the interface when a SiGe alloy film is grown on Si and a nonabrupt interface when Si is grown on Ge. The (2 × 1) reconstruction on Si(001) introduces a nonuniform stress distribution in the surface layers (LeGoues *et al.*, 1990; Liu and Lagally, 1996), leading to preferential occupations of Si or Ge at different sites and formation of an ordered SiGe phase at the interface when SiGe alloy is deposited on Si(001) (LeGoues *et al.*, 1990).

Pre-existing surface roughness generally remains at the interface after overgrowth. In the growth of a multilayer thin film, the situation is more complex because the growth of B on A can be quite different from that of A on B. As stated earlier, if B wets A, A will not wet B. In general, the roughness of a surface or interface can be quantitatively characterized by its rms roughness, lateral correlation length, and fractal exponent (Sinha *et al.*, 1988). An x-ray diffraction technique has been developed to measure the magnitude and correlation of interfacial roughness in a multilayer film (Savage *et al.*, 1991; Phang *et al.*, 1994). The interfacial roughness is frequently correlated among different layers (Savage *et al.*, 1991; Phang *et al.*, 1994). In particular, strain-induced interaction between islands in different layers leads to a self-organization of islands, with progressively improved uniformity in island size, shape, and spacing, as the multilayer thickness increases (Tersoff *et al.*, 1996; Teichert *et al.*, 1996; Liu *et al.*, 1999), which provides an attractive route to fabricating arrays and superlattices of quantum dots (Liu and Lagally, 1997a).

GLOSSARY

Adatom: A single atom adsorbed on a surface (terrace) bonded only to the surface atoms.

Basis: A group of atoms, molecules, or ions that constitute a single lattice point of the Bravais lattice.

Bravais Lattice: An infinite array of discrete points with an arrangement and orientation that appears the same from whichever of the points the array is viewed.

Crystal Systems: Point groups of the Bravais lattice (the symmetry of the basis is excluded).

Crystallographic Point Groups: Point groups of the general crystal structure including the symmetry of the basis.

Crystallographic Space Groups: Space groups of the general crystal structure.

Dangling Bond: A hybridized electron orbital on an atom without bonding on account of missing neighbors.

Dislocation: An extended line defect resulting from missing half of an atomic plane in the solid or from missing half of an atomic row in the surface.

Dividing Surface: A hypothetical geometrical surface separating two homogeneous phases to replace the real physical transition region between the two phases.

Epitaxial Growth: Growth of a film on a substrate with lattice structure of the film confined to that of the substrate.

Excess Surface Properties: The difference between the extensive thermodynamic property of a real system and that of an ideal system consisting of two homogeneous phases separated by the dividing surface.

Heteroepitaxy: Epitaxial growth of a material A on a substrate of different material B.

Homoepitaxy: Epitaxial growth of a material A on a substrate of the same material A.

Jahn–Teller Distortion: Lattice distortion that lowers the symmetry, removing the degeneracy in electronic levels and hence lowering the total energy.

Kink: A jog of any length in a step terminated by one inside corner and one outside corner.

Miller Indices: The indices of an atomic (lattice) plane that are coordinates of the shortest reciprocal lattice vector normal to the plane.

Misfit Strain: Strain induced by different equilibrium lattice constants (misfit) between the film and the substrate during heteroepitaxial growth.

Primitive Unit Cell: A volume (area) of space that, when translated through all the vectors in a Bravais lattice, just fills all the space without either overlapping itself or leaving a void.

Reciprocal Lattice: The set of all wave vectors $\boldsymbol{K}$ that yield plane waves with the periodicity of a given Bravais lattice.

Stacking Fault: A single plane where the stacking differs from the usual stacking sequence of atomic planes inside the crystal.

Stacking-Fault Region: A finite region of a stacking sequence of atomic planes different from that in the surrounding lattice.

Surface Reconstruction: Atomic displacements at a surface that change the surface periodicity relative to that of an ideally bulk-terminated surface.

Surface Relaxation: Atomic displacements at a surface that do not change the surface periodicity relative to that of an ideally bulk-terminated surface.

Surface Stress: (I) The reversible work per unit area required to deform a surface. (II) The force per unit length on a line cut in the surface.

Surface Tension: The reversible work required to create a new unit area of surface.

Vicinal Surface: A crystal surface formed by a miscut a few degrees away from a low-index atomic plane.

Works Cited

Abernathy, D. L., Mochrie, S. G., Zehner, D. M., Grübel, G., Gibbs, D. (1992), *Phys. Rev. B* **45**, 9272–9291.

Alerhand, O. L., Vanderbilt, D., Meade, R. D., Joannopoulos, J. D. (1988), *Phys. Rev. Lett.* **61**, 1973–1976.

Alerhand, O. L., Berker, N. A., Joannopoulos, J. D., Vanderbilt, D., Hamers, R. J., Demuth J. E. (1990), *Phys. Rev. Lett.* **64**, 2406–2409.

Altmann, M. S., Estrup, P. J., Robinson, I. K. (1988), *Phys. Rev. B* **38**, 5211–5214.

Arthur, J. R. (1974), *Surf. Sci.* **43**, 449–461.

Ashcroft, N. W., Mermin, N. D. (1976), *Solid State Physics*, London: Holt, Rinehart and Winston.

Badziag, P., Verwoerd, W. S. (1988), *Surf. Sci.* **201**, 87–96.

Barth, J. V., Brune, H., Ertl, G., Behm, R. J. (1990), *Phys. Rev. B* **42**, 9307–9318.

Bauer, E. (1958), *Z. Kristallogr.* **110**, 372–394.

Becker, R. S., Golovchenko, J. A., McRae, E. G., Swartzentruber, B. S. (1985a), *Phys. Rev. Lett.* **55**, 2028–2031.

Becker, R. S., Golovchenko, J. A., Swartzentruber, B. S. (1985b), *Phys. Rev. Lett.* **54**, 2678–2680.

Becker, R. S., Golovchenko, J. A., Higashi, G. S., Swartzentruber, B. S. (1986), *Phys. Rev. Lett.* **57**, 1020–1023.

Becker, R. S., Swartzentruber, B. S., Vickers, J. S., Klitsner, T. (1989), *Phys. Rev. B* **39**, 1633–1647.

Bennett, P. A., Feldman, L. C., Kuk, Y., McRae, E. G., Rowe, J. E. (1983), *Phys. Rev. B* **28**, 3656–3659.

Biegelsen, D. K., Swartz, L.-E., Bringans, R. D. (1990a), *J. Vac. Sci. Technol.* **A8**, 280–283.

Biegelsen, D. K., Bringans, R. D., Northrup, J. E., Swartz, L.-E. (1990b), *Phys. Rev. B* **41**, 5701–5706.

Biegelsen, D. K., Bringans, R. D., Northrup, J. E., Swartz, L.-E. (1990c), *Phys. Rev. Lett.* **65**, 452–455.

Binnig, G., Rohrer, H., Gerber, Ch., Weibel, E. (1983a), *Phys. Rev. Lett.* **50**, 120–123.

Binnig, G., Rohrer, H., Gerber, Ch., Weibel, E. (1983b), *Surf. Sci.* **131**, L379–L384.

Bohr, J., Feidenhans'l, R., Nielsen, M., Toney, M., Johnson, R. L., Robinson, I. K. (1985), *Phys. Rev. Lett.* **54**, 1275–1278.

Bott, M., Hohage, M., Morgenstern, M., Michely, Th., Comsa, G. (1996), *Phys. Rev. Lett.* **76**, 1304–1307.

Brennan, S., Stöhr, J., Jaeger, R., Rowe, J. E. (1980), *Phys. Rev. Lett.* **45**, 1414–1418.

Breuer, U., Prince, K. C., Bonzel, H. P., Oed, W., Heinz, K., Schmidt, G., Müller, K. (1990), *Surf. Sci.* **239**, L493–L497.

Brommer, K. D., Needels, M., Larson, B. E., Joannopoulos, J. D. (1992), *Phys. Rev. Lett.* **68**, 1355–1358.

Burton, W. K., Cabrera, N., Frank, F. C. (1951), *Philos. Trans. R. Soc. London Ser. A* **243**, 299–358.

Chadi, D. J. (1979), *Phys. Rev. Lett.*, **43**, 43–46.

Chadi, D. J. (1984), *Phys. Rev. Letters* **52**, 1911–1914.

Chadi, D. J. (1987a), *J. Vac. Sci. Technol.* **A5**, 834–837.

Chadi, D. J. (1987b), *Phys. Rev. Lett.* **59**, 1691–1694.

Chen, X., Wu, F., Zhang, Z. Y., Lagally, M. G. (1994), *Phys. Rev. Lett.* **73**, 850–853.

Cho, A. Y. (1971), *J. Appl. Phys.* **42**, 2074–2081.

Cornelissen, H. J., Cammack, D. A., Dalby, R. J. (1988), *J. Vac. Sci. Technol.* **B6**, 769–772.

Dabrowski, J., Pehlke, E., Scheffler, M. (1994), *Phys. Rev. B* **49**, 4790–4793.

Daley, R. S., Felter, T. E., Hildner, M. L., Estrup, P. J. (1993), *Phys. Rev. Lett.* **70**, 1295–1298.

de Miguel, J. J., Aumann, C. E., Kariotis, R., Lagally, M. G. (1991), *Phys. Rev. Lett.* **67**, 2830–2833.

de Miguel, J. J., Aumann, C. E., Jaloviar, S. G., Kariotis, R., Lagally, M. G. (1992), *Phys. Rev. B* **46**, 10257–10261.

Drathen, P., Ranke, W., Jacobi, K. (1978), *Surf. Sci.* **77**, L162–L166.

Duport, C., Nozières, P., Villain, J. (1995), *Phys. Rev. Lett.* **74**, 134–137.

Eaglesham, D. J., Cerullo, M. (1990), *Phys. Rev. Lett.* **64**, 1943–1946.

Ebina, A., Takahashi, T. (1982), *J. Cryst. Growth* **59**, 51–64.

Ehrlich, G., Hudda, F. G. (1966), *J. Chem. Phys.* **66**, 1039–1049.

Farrell, H. H., Palmstrøm, C. J. (1990), *J. Vac. Sci. Technol.* **B8**, 903–907.

Feibelman, P. J. (1995), *Phys. Rev. B* **51**, 17867–17875.

Feidenhans'l, R., Nielsen, M., Grey, F., Johnson, R. L., Robinson, I. K. (1987), *Surf. Sci.* **186**, 499–510.

Fery, P., Moritz, W., Wolf, D. (1988), *Phys. Rev. B* **38**, 7275–7286.

Foiles, S. M. (1987), *Surf. Sci.* **191**, L779–L786.

Foxon, C. T., Joyce B. A. (1977), *Surf. Sci.* **64**, 293–304.

Garcia, A., Northrup, J. E. (1993), *Phys. Rev. B* **48**, 17350–17353.

Gibbs, J. W. (1928), in *Collected Works*, Vol. 1, New York: Longmans, Green, p. 55.

Gossmann, H.-J., Bean, J. C., Feldman, L. C., Gibson, W. M. (1984), *Surf. Sci.* **138**, L175–L180.

Gossmann, H.-J., Bean, J. C., Feldman, L. C., McRae, E. G., Robinson, I. K. (1985), *Phys. Rev. Lett.* **55**, 1106–1109.

Gritsch, T., Coulman, D., Behm, R. J., Ertl, G. (1991), *Surf. Sci.* **257**, 297–306.

Haberern, K. W., Pashley, M. D. (1990), *Phys. Rev. B* **41**, 3226–3229.

Hamers, R. J., Tromp, R. M., Demuth, J. E. (1986a), *Phys. Rev. B* **34**, 5343–5357.

Hamers, R. J., Tromp, R. M., Demuth, J. E. (1986b), *Phys. Rev. Lett.* **56**, 1972–1975.

Hamers, R. J., Köhler, U. K., Demuth, J. E. (1989), *J. Vac. Sci. Technol.* **A8**, 195–200.

Haneman, D. (1961), *Phys. Rev.* **121**, 1093–1100.

Harrison, W. A. (1976), *Surf. Sci.* **55**, 1–19.

Harrison, W. A. (1981), *Phys. Rev. B* **24**, 5838–5843.

Himpsel, F. J. (1983), *Phys. Rev. B* **27**, 7782–7785.

Himpsel, F. J., Heimann, P., Chiang, T.-C., Eastman, D. E. (1980), *Phys. Rev. Lett.* **45**, 1112–1115.

Hirschorn, E. S., Lin, D. S., Leibsle, F. M., Samsavar, A., Chiang, T. C. (1991), *Phys. Rev. B* **44**, 1403–1406.

Ho, K. M., Bohnen, K. P. (1987), *Phys. Rev. Lett.* **59**, 1833–1836.

Hohage, M., Michely, Th., Comsa, G. (1995), *Surf. Sci.* **337**, 249–267.

Hohage, M., Bott, M., Morgenstern, M., Zhang, Z., Michely, Th., Comsa, G. (1996), *Phys. Rev. Lett.* **76**, 2366–2369.

Ichikawa, T., Sueyosi, T., Sato, T., Iwatsuki, M., Udagawa, F., Sumita, I. (1995), *Solid State Commun.* **93**, 541–545.

Ihm, J., Lee, D. H., Joannopoulos, J. D., Xiong, J. J. (1983), *Phys. Rev. Lett.* **51**, 1872–1875.

Jacobson, K. W., Nørskov, J. K. (1988), in: J. K. van der Veen, M. A. Van Hove (Eds.) *The Structure of Solid Surfaces II*, Berlin: Springer.

Kandel, D., Weeks, J. D. (1995), *Phys. Rev. Lett.* **74**, 3632–3635.

Kaxiras, E., Bar-Yam, Y., Joannopoulos, J. D., Pandey, K. C. (1986), *Phys. Rev. Lett.* **57**, 106–109.

Keating, P. N. (1966), *Phys. Rev.* **145**, 637–645.

Kevan, S. D., Stoffel, N. G. (1984), *Phys. Rev. Lett.* **53**, 702–705.

Kitamura, N., Swartzentruber, B. S., Lagally, M. G., Webb, M. B. (1993a), *Phys. Rev.* **B48**, 5704–5707.

Kitamura, N., Lagally, M. G., Webb, M. B. (1993b), *Phys. Rev. Letters* **71**, 2082–2085.

Kubby, J. A., Griffith, J. E., Becker, R. S., Vickers, J. S. (1987), *Phys. Rev. B* **36**, 6079–6093.

Lagally, M. G. (1993), *Jpn. J. Appl. Phys.* **32**, 1493–1501.

Lander, J. J., Gobeli, G. W., Morrison, J. (1963), *J. Appl. Phys.* **34**, 2298–2306.

Lander, J. J., Morrison, J. (1966), *Surface Sci.* **4**, 241–247.

Lang, E., Müller, K., Heinz, K., Van Hove, M. A., Koestner, R. J., Somorjai, G. A. (1983), *Surf. Sci.* **127**, 347–365.

LeGoues, F. K., Kesan, V. P., Iyer, S. S., Tersoff, J., Tromp, R. (1990), *Phys. Rev. Lett.* **64**, 2038–2041.

LeGoues, F. K., Mooney, P. M., Tersoff, J. (1993), *Phys. Rev. Lett.* **71**, 396–399.

Levine, J. D., Freeman, S. (1970), *Phys. Rev. B* **2**, 3255–3272.

Liu Feng, Davenport S. E., Erans, H. M., Lagally M. G. (1999), *Phys. Rev. Lett.* **82**, 2528–2531.

Liu Feng, Lagally, M. G. (1996), *Phys. Rev. Lett.* **76**, 3156–3159.

Liu Feng, Lagally, M. G. (1997a), *Surf. Sci.* **386**, 169–181.

Liu Feng, Lagally, M. G. (1997b) in: D. A. King, D. P. Woodruff (Eds.), *The Chemical Physics of Solid Surfaces*, Vol. 8, Amsterdam: Elsevier, p. 258–296.

Liu Feng, Wu Fang, Lagally, M. G. (1997), *Chem. Rev.* **97**, 1045–1061.

Lucas, C. A., Dower, C. S., McMorrow, D. F., Wong, G. C. L., Lamelas, F. J., Fuoss, P. H. (1993), *Phys. Rev. B* **47**, 10375–10382.

MacRae, A. U., Gobeli, G. W. (1966), in: R. K. Willardson, A. C. Beer (Eds.), *Semiconductors and Semimetals*, Vol. 2, New York: Academic, p. 115.

Marsh, J. B., Farnsworth, H. E. (1964), *Surface Sci.* **1**, 3–21.

Massies, J., Etienne, P., Dezaly, F., Linh, N. T. (1980), *Surf. Sci.* **99**, 121–131.

Materer, N., Starke, U., Barbieri, A., Döll, R., Heinz, K., Van Hove, M. A., Somorjai, G. A. (1995), *Surf. Sci.* **325**, 207–222.

Marchenko, V. I. (1981), *Pis'ma Zh. Eksp. Teor. Fiz.* **33**, 397–399 [*JETP Lett.* **33**, 397–399].

Matthews, J. W., Blakeslee, A. E. (1975), *J. Cryst. Growth* **29**, 273–280.

Matthews, J. W., Blakeslee, A. E. (1976), *J. Cryst. Growth* **32**, 265–273.

McRae, E. G. (1983), *Surf. Sci.* **124**, 106–128.

Meade, R. D., Vanderbilt, D. (1989), *Phys. Rev. B* **40**, 3905–3913.

Men, F. K., Packard, W. F., Webb, M. B. (1988), *Phys. Rev. Lett.* **61**, 2469–2471.

Mo, Y. W., Lagally, M. G. (1991), *J. Cryst. Growth* **111**, 876–881.

Mo, Y. W., Savage, D. E., Swartzentruber, B. S., Lagally, M. G. (1990), *Phys. Rev. Lett.* **65**, 1020–1023.

Mo, Y. W., Kleiner, J., Webb, M. B., Lagally, M. G. (1991), *Phys. Rev. Lett.* **66**, 1998–2001.

Mönch, W. (1995), *Semiconductor Surfaces and Interfaces*, 2nd ed., Berlin, New York: Springer-Verlag, Chap 7.

Moritz, W., Wolf, D. (1985), *Surf. Sci.* **163**, L655–L665.

Mullins, W. W., Hirth, J. P. (1963), *J. Phys. Chem. Solids* **24**, 1391–1404.

Nakada, T., Osaka, T. (1991), *Phys. Rev. Lett.* **67**, 2834–2837.

Naumovets, A. G., Fedorus, A. G. (1975), *Zh. Eksp. Teor. Fiz.* **68**, 1183 (1975) [*Sov. Phys. JETP* **41**, 587–589].

Neave, J. H., Joyce, B. A. (1978), *J. Cryst. Growth* **44**, 387–397.

Needels, M., Payne, M. C., Joannopoulos, J. D. (1987), *Phys. Rev. Lett.* **58**, 1765–1768.

Needs, R. J., Godfrey, M. J., Mansfield, M. (1991), *Surf. Sci.* **242**, 215–221.

Noonan, J. R., Davis, H. L. (1984), *Phys. Rev. B* **29**, 4349–4355.

Northrup, J. E. (1986), *Phys. Rev. Lett.* **57**, 154–157.

Northrup, J. E., Cohen, M. L. (1982), *Phys. Rev. Lett.* **49**, 1349–1352.

Northrup, J. E., Cohen, M. L. (1984), *Phys. Rev. B* **29**, 1966–1969.

Ourmazd, A., Taylor, D. W., Bevk, J., Davidson, B. A., Feldman, L. C., Mannaerts, J. P. (1986), *Phys. Rev. Lett.* **57**, 1332–1335.

Pandey, K. C. (1982), *Phys. Rev. Lett.* **49**, 223–226.

Park, R. L., Madden, H. H. (1968), *J. Surf. Sci.* **11**, 188–202.

Pashley, M. D. (1989), *Phys. Rev. B* **40**, 10481–10487.

Pehlke, E., Tersoff, J. (1991), *Phys. Rev. Lett.* **67**, 465–468, 1290.

Pendry, J. B., Heinz, K., Oed, W., Landskron, H., Müller, K., Schmidtlein, G. (1988), *Surf. Sci.* **193**, L1–L6.

Perderau, J., Biberian, J. P., Rhead, G. E. (1974), *J. Phys. F* **4**, 798–806.

Phang, Y. H., Teichert, C., Lagally, M. G., Peticolas, T. J., Bean, J. C., Kasper, E. (1994), *Phys. Rev. B* **50**, 14435–14445.

Poon, T. W., Yip, S., Ho, P. S., Abraham (1990), *Phys. Rev. Lett.* **65**, 2161–2164.

Qian G.-X., Chadi, D. J. (1987), *Phys. Rev. B* **35**, 1288–1293.

Ramstad, A., Brocks, G., Kelly, P. J. (1995), *Phys. Rev. B* **51**, 14504–14523.

Reichardt, D. (1991), *Prog. Surf. Sci.* **35**, 63–66.

Rous, P. J. (1995) in: F. R. de Boer, D. G. Pettifor (Eds.), *Cohesion and Structure of Surfaces*, Amsterdam: Elsevier, vol. 4, 1.

Sandy, A. R., Mochrie, S. G. J., Zehner, D. M., Grübel, G., Hwang, K. G., Gibbs, D. (1992), *Phys. Rev. Lett.* **68**, 2192–2195.

Savage, D. E., Schimke, N., Phang, Y. H., Lagally, M. G. (1991), *Phys. Rev. B* **71**, 3283–3293.

Saxena, A., Gawlinski, E. T., Gunton, J. D. (1985), *Surf. Sci.* **160**, 618–640.

Schlier, R. E., Farnsworth, H. E. (1959), *J. Chem. Phys.* **30**, 917–926.

Schlüter, M. (1988) in: King D. A. and Woodruff, D. P. (Eds.), *The Chemical Physics of Solid Surfaces*, Vol. 5, Amsterdam: Elsevier, p. 64.

Schottke, M., Behm, R. J., Ertl, G., Penka, V., Moritz, W. (1987), *J. Chem. Phys.*, **87**, 6191–6198.

Schwoebel, R. L., Shipsey, E. J. (1966), *J. Appl. Phys.* **37**, 3682–3686.

Singh, D., Krakauer, H. (1988), *Phys. Rev. B* **38**, 3999–4006.

Sinha, D. T., Sirota, E. B., Garoff, S., Stanley, H. B. (1988), *Phys. Rev. B* **38**, 2297–2311.

Štich, I., Payne, M. C., King-Smith, R. D., Lin, J.-S., Clarke, L. J. (1992), *Phys. Rev. Lett.* **68**, 1351–1354.

Swartzentruber, B. S., Mo, Y.-W., Kariotis, R., Lagally, M. G., Webb, M. B. (1990), *Phys. Rev. Lett.* **65**, 1913–1916.

Swartzentruber, B. S., Kitamura, N., Lagally, M. G., Webb, M. B. (1992), *Phys. Rev. B* **47**, 13432–13441.

Tabata, T., Aruga, T., Murata, Y. (1987), *Surf. Sci.* **179**, L63–L70.

Takayanagi, K., Tanishiro, Y., Takahashi, M., Takahashi, S. (1985a), *J. Vac. Sci. Technol.* **A3**, 1502–1510.

Takayanagi, K., Tanishiro, Y., Takahashi, M., Takahashi, S. (1985b), *Surf. Sci.* **164**, 367–392.

Teichert C., Lagally M. G., Peticolas L. J., Bean, J. C., Tersoff J. (1996), *Phys. Rev. B* **53**, 16334–16337.

Tersoff, J. (1991), *Phys. Rev. B* **43**, 9377–9380.

Tersoff, J., Pehlke, E. (1992), *Phys. Rev. Lett.* **68**, 816–819.

Tersoff, J., Phang, Y. H., Zhang, Z., Lagally, M. G. (1995), *Phys. Rev. Lett.* **75**, 2730–2733.

Tersoff, J., Teichert, C., Lagally, M. G. (1996), *Phys. Rev. Lett.* **76**, 1675–1678.

Tong, S. Y., Xu, G., Mei, W. N. (1984) *Phys. Rev. Lett.* **52**, 1693–1696.

Tong, X., Bennett, P. A. (1991), *Phys. Rev. Lett.* **67**, 101–104.

Tromp, R. M., Hamers, R. J., Demuth, J. E. (1985), *Phys. Rev. Lett.* **55**, 1303–1306.

Tromp, R. M., Hamers, R. J., Demuth, J. E. (1986), *Science* **234**, 304–309.

Tromp, R. M., Reuter, M. C. (1992), *Phys. Rev. Lett.* **68**, 820–822.

van Bommel, A. J., Crombeen, J. E., Dirschot, T. G. J. (1978), *Surf. Sci.* **72**, 95–108.

van der Merwe, J. H. (1963), *J. Appl. Phys.* **34**, 117; 123–127.

van Silfhout, R. G., van der Veen, J. F., Norris, C., MacDonald, J. E. (1990), *Faraday Discuss. R. Soc. Chem.* **89**, 169–180, 204–205.

Vanderbilt, D., Alerhand, O. L., Meade, R. D., Joannopoulos, J. D. (1989). *J. Vac. Sci. Technol.* **B7**, 1013–1016.

Venables, J. A., Spiller, G. D. T., Hanbücken, M. (1984), *Rep. Prog. Phys.* **47**, 399–459.

Wang, X.-S., Goldberg, J. L., Bartelt, N. C., Einstein, T. L., Williams, E. D. (1990), *Phys. Rev. Lett.* **65**, 2430–2433.

Webb, M. B., Men, F. K., Swartzentruber, B. S., Kariotis, R., Lagally, M. G. (1990), in: M. G. Lagally (Ed.), *Kinetics of Ordering and Growth at Surfaces*, New York: Plenum.

Webb, M. B., Men, F. K., Swartzentruber, B. S., Kariotis, R., Lagally, M. G. (1991), *Surf. Sci.* **242**, 23–31.

Wierenga, P. E., Kubby, J. A., Griffith, J. E. (1987), *Phys. Rev. Lett.* **59**, 2169–2172.

Wolkow, R. A. (1992), *Phys. Rev. Lett.* **68**, 2636–2639.

Wood, E. A. (1964), *J. Appl. Phys.* **35**, 1306–1312.

Woodruff, D. P. (1981), in: D. A. King, D. P. Woodruff (Eds.), *The Chemical Physics of Solid Surfaces and Heterogeneous Catalysis*, Vol. 1, Amsterdam: Elsevier, p. 175.

Woodruff, D. P. (1994), in: D. A. King, D. P. Woodruff (Eds.), *The Chemical Physics of Solid Surfaces*, Vol. 7, Amsterdam: Elsevier, p. 513.

Wu, F., Lagally, M. G. (1995), *Phys. Rev. Lett.* **75**, 2534–2537.

Wu, F., Chen, X., Zhang, Z. Y., Lagally, M. G. (1995), *Phys. Rev. Lett.* **74**, 574–577.

Xu, G., Hu, W. Y., Puga, M. W., Tong, S. Y., Yeh, J. L., Wang, S. R., Lee, B. W. (1985), *Phys. Rev. B* **32**, 8473–8476.

Yamamoto Y. (1994), *Surf. Sci.* **313**, 155–167.

Zandvliet, H. J., Elswijk, H. B., van Leonen, E. J. (1992), *Surf. Sci.* **272**, 264–268.

Zhang, Z. Y., Chen, H., Bolding, B. C., Lagally, M. G. (1993), *Phys. Rev. Lett.* **71**, 3677–3680.

Zhang, Z. Y., Chen, X., Lagally, M. G. (1994), *Phys. Rev. Lett.* **73**, 1829–1832.

Zhang, Z., Lagally, M. G. (1997), *Science*, **276**, 377–383.

Zubkus, V. E., Kundrotas, P. J., Molotkov, S. N., Zatarsi, V. V., Tornau, E. E. (1991), *Surf. Sci.* **243**, 295–302.

Further Reading

Bauer, E. (1958), *Z. Kristallogr.* **110**, 372–395.

Burton, W. K., Cabrera, N., Frank, F. C. (1951), *Philos. Trans. R. Soc. London Ser. A* **243**, 299–358.

Ertl, G., Küppers, J. (1974), *Low Energy Electrons and Surface Chemestry*, Weinheim: Verlag Chemie.

Herring, C. (1953), in: R. Gomer, C. S. Smith, (Eds.), *Structure and Properties of Solid Surfaces*, Chicago: The University of Chicago Press, p. 5.

King, D. A., Woodruff, D. P. (Eds.) (1981), *The Chemical Physics of Solid Surfaces and Heterogeneous Catalysis*, Vol. 7, Amsterdam, Elsevier.

King, D. A., Woodruff, D. P. (Eds.) (1998), *The Chemical Physics of Solid Surfaces*, Vol. 8, Amsterdam: Elsevier.

Liu, Feng, Wu, Fang, Lagally, M. G. (1997), *Chem. Rev.* **97**, 1045.

Mönch, W. (1995), *Semiconductor Surfaces and Interfaces*, 2nd ed., Berlin, New York: Springer-Verlag.

Yang, H. N., Wang, G. C., Lu, T. M. (1993), *Diffraction from Rough Surfaces and Dynamic Gorwth Fronts*, Singapore: World Scientific.

Zhang, Z., Lagally M. G. (1997), *Science* **276**, 377–383.

ATMOSPHERIC STRUCTURE

(An Addendum to *Encyclopedia of Applied Physics*, Volume 2, pages 201–224.)

RAYMOND G. ROBLE, *High Altitude Observatory, National Center for Atmospheric Research, Boulder, Colorado, U.S.A.*

Introduction 355
1. **Thermosphere and Ionosphere** ... 355
2. **Lower Thermosphere and Ionosphere** 356
3. **Stratosphere** 358
4. **Troposphere Weather and Climate** 358
Works Cited 359
Further Reading 359

The original article ATMOSPHERIC STRUCTURE describes the basic vertical structure of the Earth's atmosphere, with a brief description of latitudinal and seasonal variation and the response of the atmosphere to solar variability. This Addendum describes new insights into the atmospheric structure in regions that were poorly known when the original article was written but have been measured recently by new satellite- and ground-based instruments. A brief summary of the nature of the variability found in each atmospheric layer is also given, starting from the upper atmosphere downward to the Earth's surface.

1. THERMOSPHERE AND IONOSPHERE

Within the thermosphere and ionosphere most of the variability about the basic state is controlled by the solar radiative and plasma outputs. Since the original article was written there have been only a few new measurements of the sun's spectral irradiance in the extreme ultraviolet portion of the spectrum (10 to 120 nm), primarily from rockets. However, there have been extensive measurements of the solar UV spectrum (120 to 400 nm) and total solar irradiance that have been made from various satellites. These measurements have defined the range of variability of the solar radiative output over a solar cycle and also have shown variability associated with the 27-d solar rotation period, and shorter-term variability, such as during solar flares. In general, the solar radiative output over a solar cycle varies by about 10% near 200 nm, 1% near 300 nm, and 0.1% for the total solar irradiance. The older solar EUV measurements show considerable day-to-day variability that influences the basic temperature and compositional structure of the thermosphere and ionosphere as discussed in the original article.

There have also been new measurements from satellites and new analysis techniques that better define the energy and momentum inputs into the thermosphere and ionosphere from the aurora. The aurora is driven by the interaction of the solar-wind plasma with the Earth's magnetic field, and images of the aurora made by the NASA POLAR satellite in the UV and visible range have observed a tremendous source of variability introduced into the Earth's atmosphere especially during geomagnetic storms. During these storms the energy and momentum inputs into the Earth's upper atmosphere can vary by over two orders of magnitude. As a result, the high-latitude thermosphere and ionosphere are subjected to a wide range of dynamic phenomena such as large horizontal scale waves that travel globally, circulation changes to the basic wind structure, temperature and compositional perturbations that occur at high latitudes and propagate equatorward, and many others. Schematics showing the response to the aurora have been presented in the original article, and the new data have helped to better define the forcing that drive the variability and the small-scale structure.

In addition to the variability introduced into the thermosphere and ionosphere by solar and auroral variability, numerical models that couple the lower and upper atmosphere show

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

that another important source of variability in the thermosphere and ionosphere consists of gravity waves, tides, and other large-scale disturbances generated by weather systems in the lower atmosphere. These waves and disturbances propagate upward and deposit their energy and momentum in the thermosphere and are a major source of variability, especially during quiet auroral conditions and steady solar EUV output. During these conditions waves from the lower atmosphere are the major source of day-to-day variability that has been observed over a given station or by a satellite in orbit.

2. LOWER THERMOSPHERE AND MESOSPHERE

For many years the atmospheric structure and dynamics in the upper mesosphere and lower thermosphere between about 60 and 200 km altitude has not been measured on a global basis. Most data came from rockets and remote sensing from ground stations. The Upper Atmosphere Research Satellite (UARS) that was launched in 1991 had instruments on board that could measure the global temperature and winds between 30 and 500 km altitude. These measurements not only defined the basic structure of the region, but also illustrated the range of variability that exists in the upper atmosphere. Analysis of the data indicated that the equatorial upper atmosphere below about 110 km is strongly influenced by the diurnal tide that is forced by solar radiation being absorbed by water vapor in the troposphere and propagating upward to the 100-km altitude region before dissipating by molecular viscosity. At mid and high latitudes the tidal structure is dominated by a semidiurnal tide that is excited not only in the troposphere but also by *in situ* absorption of solar radiation primarily by ozone in the middle atmosphere near 50 km. During the summer months a strong signature of an intermittent two-day wave was found to dominate the winds in the upper mesosphere and lower thermosphere between 70 and 100 km altitude. The UARS satellite also found equatorially trapped fast Kelvin waves with periods of 3 to 5 d and regions of strong atmospheric gravity-wave activity that influenced the basic wind, temperature, and compositional structure of the upper mesosphere and lower thermosphere profoundly. In fact, the gravity-wave activity in this region of the upper atmosphere is so strong that it primarily controls the temperature and dynamic structure and drives it away from a radiative equilibrium structure that would result if only absorption of solar radiation acted upon the region. An example of the complex wind structure observed by the UARS satellite is shown in Fig. 1, which illustrates the zonal (positive eastward) and meridional (positive northward) winds for March equinox and December solstice at 1200 local time. These are the first global observations of the wind structure between 60 and 200 km, and they shown that the diurnal tide propagates to about 110 km at low latitudes and that there is a major transition in the dynamic structure above 110 km. This structural change occurs because the atmosphere is so thin above 100 km that the waves in the lower atmosphere are damped and strong solar EUV forcing in the thermosphere gives rise to a different circulation pattern. These figures show the changing wind structure for both equinox and solstice conditions near 100 km altitude. These data filled a critical gap in our knowledge of the wind structure in the region between 60 and 200 km altitude. These data, with the additional new data from ground stations such as LIDARs, radars, and airglow instruments, are greatly improving our understanding of the upper-atmosphere structure and dynamics.

In addition to the wind and temperature measurements, there have been composition measurements of the upper atmosphere. Briefly, the new measurements indicate that current aeronomical theory predicts too much hydroxyl, OH, that catalytically destroys ozone in the mesosphere. The ozone presently calculated by upper-atmosphere models is about 20% to 30% lower than measurements indicate. It is important to resolve this discrepancy because ozone is such an important constituent for absorbing damaging solar UV radiation before it reaches the Earth's surface. The current problem suggests that the laboratory rate coefficients for certain chemical reactions need to be re-evaluated because they may be different for upper-atmosphere environmental conditions. There are also problems

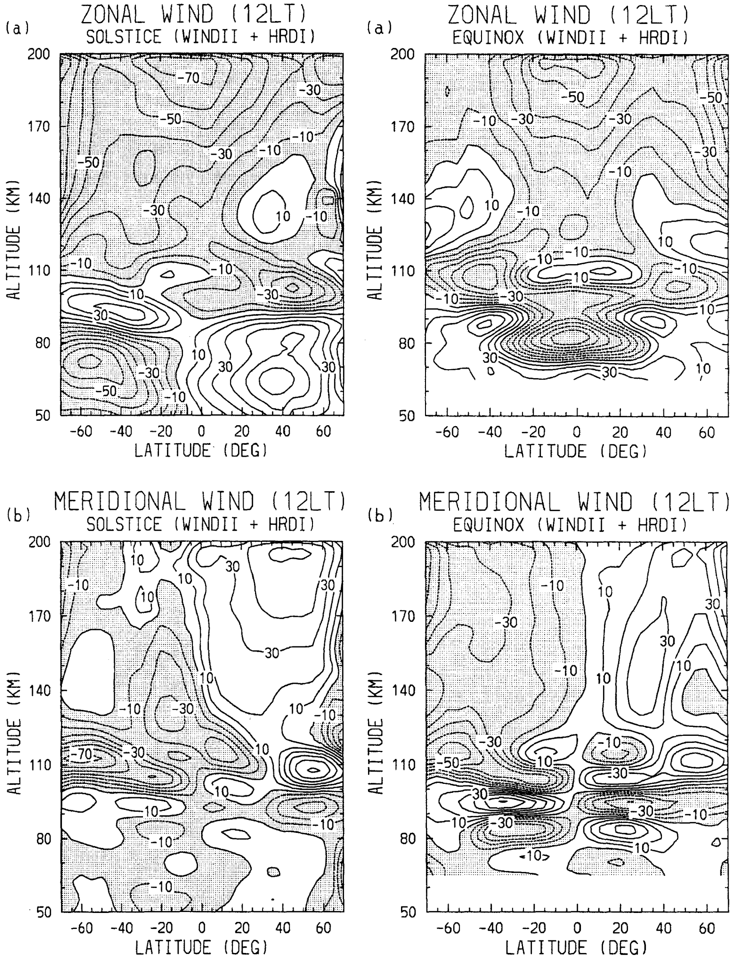

FIG. 1. Combined zonally averaged mesospheric and thermospheric winds for December solstice period (left panels) and March equinox period (right panels) at 1200 h local time measured by the Upper Atmosphere Research satellite. The upper frames are the zonal and the lower frames are the meridional components of the wind. A 10-m/s contour interval is employed. Shading is used for westward winds in the upper panels and for southward winds in the lower panels (McLandress *et al.*, 1996). The acronym WINDII represents the WIND Imaging Interferometer, and HRDI, the High Resolution Doppler Imager, instruments used to measure the winds from the Upper Atmosphere Research Satellite (UARS).

between theory and observations for other chemical species such as nitric oxide and carbon monoxide, indicating that there is still considerable work needed before we are able to understand the current compositional structure of the upper atmosphere.

A very important problem that governs the temperature structure of the upper mesosphere and lower thermosphere between 80 and 150 km altitude is the magnitude of the O–CO_2 vibrational exchange coefficient. This coefficient governs the infrared radiative cooling of the upper atmosphere. Current theory, laboratory measurements, and observational analysis of satellite infrared emissions indicate a factor of 6 uncertainty in the magnitude of this important rate coefficient, which governs the temperature structure of the upper atmosphere. Not only does this problem need to be resolved for understanding current-day atmospheric structure, but it is important for determining the magnitude of possible future global change.

There have been observations that suggest that the upper atmosphere above about 50 km altitude has been cooling during the last decade. This secular trend has been identified in satellite, LIDAR, and rocket data, but the cause of the cooling is not known. Studies of the upper-atmosphere response to global change and a CO_2 doubling from present-day levels indicate that the upper atmosphere will cool because of increased infrared radiation to space. But the major cooling is expected to occur slowly during the next century, and it is generally believed to be too small at the present time to account for the observed cooling. Another possible cause may be an association with the observed ozone decrease caused by chlorofluorocarbons (CFC's) affecting the ozone layer, or it may be a natural variability of the ocean–atmosphere system that is not well understood.

3. STRATOSPHERE

Ozone is one of the greenhouse gases that modify the radiative structure of our atmosphere. A change in ozone amount and distribution can influence climate. In addition, ozone protects the biosphere from harmful solar ultraviolet radiation. Ozone is continually being produced and destroyed by solar ultraviolet radiation. Additionally ozone is destroyed by catalytic reactions of nitrogen, hydrogen, chlorine, and bromine oxides. In the early 1970s it was predicted that human activity will lead to a diminishing ozone layer; and since the detection of the Antarctic ozone hole, the downward trends in the column total ozone have been well documented for the regions both inside and outside Antarctica by satellite, ground-based, and aircraft observations. Decadal trends have been observed, with the greatest ozone loss occurring in the Antarctic and the smallest in the equatorial region. Rates as large as −3%/yr have been observed in spring over Antarctica and −0.8%/yr in northern-hemisphere spring at mid-latitudes (Stolarski *et al.*, 1991). The ozone loss has also been linked to an observed cooling of the lower stratosphere. These observations all indicate that the basic structure described in the original article is subject to considerable variability and that these changes may be caused by a combination of natural variability, man's activities, changes in the solar radiative and plasma output, and changes associated with atmosphere–ocean and atmosphere–biosphere interactions. These complex interactions all contribute to a possible change in the vertical structure of the atmosphere.

Some of the major sources of atmospheric variability in the stratosphere are global oscillations associated with the quasi-biennial oscillation (QBO), an oscillation with an irregular period of about 26 months where the equatorial winds in the stratosphere oscillate in an easterly/westerly direction. A seasonal variation occurs with interannual variability superimposed and there are solar-cycle variations that have been detected within the stratosphere.

4. TROPOSPHERE WEATHER AND CLIMATE

It is well known that the Earth's climate has undergone considerable change over geologic history. In its early history the atmosphere was much warmer because of higher CO_2 levels. There were also ice ages and interglacial periods, all indicating that the basic structure of the atmosphere has undergone considerable change with time. Not only have the surface temperatures changed with time but most likely the entire atmospheric vertical structure has changed.

Variability in the lower atmosphere is governed by many factors. The most prominent that has emerged in recent years consists of the major changes to the climatological structure caused by El Niño. The period 1991 to 1995 was one of the longest El Niño events in recent times, and the major event that is presently underway in 1997 is expected to cause major changes in weather patterns. It is well documented that El Niño can affect the weather near the surface, but it is not clear whether there are associated changes that occur in the upper atmosphere. Interannual variability in the lower atmosphere is large and the major causes of this variability are uncertain but are subject to intensive study. Much of the variability is associated with atmosphere–ocean interactions, but there is also considerable variability introduced by nonlinear interactions of the large-scale dynamical processes.

Atmospheric chemistry is also having a strong influence on regional climate. The atmosphere is not only subjected to increasing amounts of important atmospheric trace gases such as CO_2, CH_4, and NO_x, but aerosol-forming gases give rise to cloud seeding and aerosol layers that affect the solar radiative balance of the lower atmosphere. All of these chemical processes exert some influence on the climatological balance within the lower atmosphere that in turn influences the structure of the entire atmosphere.

The limits of variability in the lower atmosphere have been quantified by use of historical records, but each year the record highs or lows for an individual station may be broken, indicating that the true range of atmospheric variability has not been determined. With possible global change occurring in the next century, computer models are suggesting that the range of variability may increase, with periods of prolonged drought and intense storms occurring in future climate scenarios. There is considerable effort being expended to identify and understand possible global change caused by increasing CO_2 concentrations and other greenhouse gases.

Works Cited

McLandress, C. M., Shepherd, G. G., Solheim, B. H., Burrage, M. D., Hays, P. B., Skinner, W. R. (1996), "Combined mesosphere/thermosphere winds using WINDII and HRDI data from the Upper Atmosphere Research Satellite," *J. Geophys. Res.* **101**, 10441–10455.

Stolarski, R. S., Bloomfield, P., McPeters, R. D., Herman, J. R. (1991), "Total ozone trends deduced from NIMBUS 7 TOMS data," *Geophys. Res. Lett.* **18**, 1015–1018.

Further Reading

Brekke, A. (1997), *Physics of the Upper Polar Atmosphere*, New York: Wiley.

Brasseur, G., Solomon, S. (1986), *Aeronomy of the Middle Atmosphere*, Dordrecht: D. Reidel Publishing.

Salby, M. L. (1996), *Fundamental of Atmospheric Physics*, San Diego: Academic Press.

BIOLOGICAL EFFECTS OF ELECTROMAGNETIC AND PARTICLE RADIATION

(An Addendum to *Encyclopedia of Applied Physics*, Volume 2, pages 365–401.)

JÜRGEN KIEFER, *Strahlenzentrum der Justus-Liebig-Universität, Giessen, Germany*

Introduction 361
1. **Deposition of Energy** 361
2. **Cellular Effects** 362
2.1 Delayed Reproductive Death, Apoptosis, Genomic Instability, and Low-Dose Responses 362
2.2 Repair Processes 362
2.3 Radiofrequency Electromagnetic Radiation 363
3. **Organ and Whole-Body Effects: Cancerogenesis** 363
4. **Applied Radiobiology** 364
4.1 New Developments in Biodosimetry 364
4.1.1 Fluorescence *in Situ* Hybridization for Chromosomal Translocations (FISH) 364
4.1.2 ESR Dosimetry................ 364
4.2 Developments in Radiation Therapy........................ 364
5. **Radioecology, Radiation Protection, and Guidelines** 365
5.1 Effects of the Chernobyl Accident 365
5.2 Ultraviolet Radiation 365
5.3 Radiofrequency Electromagnetic Radiation..... 366
5.4 Principles and Fundamentals of Radiation Protection........... 367
Glossary 368
Works Cited and Further Reading 368

INTRODUCTION

It is impossible to review here all new developments that took place after the publication of the original article, and so an admittedly subjective choice was made. Quite a few larger publications appeared, the most important being the 1990 recommendations of the International Commission for Radiological Protection (ICRP) (ICRP 60, 1991), a new report of the UN Scientific Committee on the Effects of Atomic Radiations (UNSCEAR) (UNSCEAR, 1994) and the proceedings of the 10th International Congress of Radiation Research (Hagen *et al.*, 1995). Particularly the last volume is a valuable source to be consulted for an overview of the present state of the art in the field.

1. DEPOSITION OF ENERGY

Energy deposition plays an important role in the understanding of the action of ionizing radiations; with non-ionizing radiations the actual spatial pattern depends more on the structure of the target than on the property of the radiation. These questions have so far received less attention and will not be treated here. Classical microdosimetry is limited to micrometer dimensions. For even smaller sizes one has to recur to Monte-Carlo calculations. They have now been performed for protons and alpha particles. Presently it is attempted to determine energy deposition in molecules of biological interest, particularly DNA, to understand the initial processes of lesion formation and link them to the biological effective-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ness of different radiations. It has been found that the yield of DNA double-strand breaks as measured by conventional techniques does not depend to a great extent on radiation type. Chromosomal aberrations, cell killing, and mutation induction, on the other hand, show a clear dependence. This is interpreted by assuming that the interaction between primary lesions and their proximity plays an important role. Densely ionizing particles produce strand breaks close to each other forming "clusters" not resolvable by the usual experimental approaches. They presumably give rise to errors in repair. An understanding of this phenomenon is essential to assessment of the risk of, e.g., neutrons and alpha particles (Rossi and Zaider, 1996).

2. CELLULAR EFFECTS

2.1 Delayed Reproductive Death, Apoptosis, Genomic Instability, and Low-Dose Responses

If surviving cells as defined by the colony-forming assay are further cultivated, they display a decrease in plating efficiency that depends on the dose originally given. It appears that they retain some kind of "cryptic" damage expressed at later stages. In many cases these cells die in a characteristic manner that has been termed *apoptosis* and constitutes some kind of a cellular "suicide program". The cells do not just disintegrate as in necrosis but follow a given scheme of interphase death: At first the cells shrink and their mean density increases; this is followed by changes in their morphological appearance. Most characteristically the chromatin is fragmented into nucleosome sizes, i.e., about 200 nucleotide pairs. Apoptosis is an active cellular process requiring active protein synthesis, mediated by enzymes and under genetical control. It can also be directly triggered by ionizing radiation and seems to play an important role particularly at low doses (Hendry *et al.*, 1995).

Cells surviving the exposure to ionizing radiation display also a phenomenon called *genomic instability*, which was first demonstrated after alpha-particle irradiation: As discussed in the original article, dicentric chromosomal aberrations can be scored after one or two replication rounds. Their number decreases with time because the affected cells die. After several generations there is another increase that is due to newly formed aberrations, indicating fragility of the genome caused by the previous irradiation. At the same time, mutations may arise that were not present in the population before. All these findings have an important bearing on the understanding of radiocarcinogenesis: it is now well established that it proceeds in a multistage fashion, and that more than one (presumably around 5) mutational events are required. It is extremely unlikely that they are all caused directly by radiation; more probably genomic instability is responsible for the later steps (Little, 1994).

The response of cells to low doses of ionizing radiation has received increased attention. It is now clear that gene activity and its regulation can be influenced by x rays as well as by UV. Studies using single-cell observations suggest that mammalian cells may be more sensitive to killing by low doses of x rays than anticipated from an extrapolation from the overall survival curve. At the same time it has been reported by several groups that the yield of chromosomal aberration after high doses is reduced if the cell received "priming doses" in the order of a few hundred milligrays several hours before. This "adaptive response" depends on an active cellular metabolism and may be related to inducible repair processes, but there is presently no proof for this supposition. Studies related to these questions can be found in a special issue of *Radiation Research* (Skov and Marples, 1994) and is also summarized in UNSCEAR (1994).

2.2 Repair Processes

The last years have brought impressive progress in the understanding of cellular repair of radiation-induced damage (Friedberg *et al.*, 1995). Most of the relevant genes have been cloned, both in microorganisms and in mammalian cells. Excision repair was shown to work preferentially in active genes via a coupling with transcription. Apart from this, "general" excision can also be found. A number of genes responsible for the repair of radiation-induced double-strand breaks (DSB) in mammalian cells could be identified. It is now clear that DSB repair in mammalian cells operates through an "end joining" mechanism, in contrast to the case in yeast where homologous recombination plays the most important role. A major breakthrough was the cloning of the *Ataxia telangiectasia* (AT) gene, although its role remains still elusive.

2.3 Radiofrequency Electromagnetic Radiation

Because of the wide-spread public interest in the possible adverse effects of power lines, e.g., the induction or promotion of cancer (see below), effects on the cell and tissue level have to be analyzed. There are a number of mechanisms that could be invoked: The quantum energy of rf (radio-frequency) electromagnetic radiation is much too low to cause any direct damage at the molecular level and also well below the "thermal noise": it is 2.5×10^{-13} eV at 60 Hz while the mean molecular kinetic energy ($\frac{3}{2}kT$) is 0.04 eV. Forces on charged particles by a field of an electric field of 1000 V/m or a magnetic field of 0.1 mT, ranging between 10^{-7} and 10^{-10} pN, are even smaller than those generated when a single hair cell in the inner cochlea of the ear is activated (about 1 pN), the most sensitive interaction known in biological systems. An interaction with isolated components can thus be ruled out and some kind of cooperative mechanism has to be invoked. Electromagnetic fields may alter the lifetime of radicals, which could cause biological effects, but the levels required are much larger than created from external electricity. Resonance phenomena have frequently been suggested, but a careful analysis reveals that they are not plausible in real biological systems. The main argument is that the medium causes rapid damping so that with the external fields applied resonance cannot build up. The sensitivity to the pickup of even very weak signals might be increased if a number of sensors cooperate in a coherent way. This has been found, e.g., in sharks, who are able to sense even 10^{-5} V/m, but no such arrangement has ever been found or even suggested in higher mammals; and it would also be extremely difficult to link it with carcinogenesis. The authors of a very careful and rigorous study (Valberg *et al.*, 1997) come to the conclusion that there is no plausible mechanism to invoke a harmful action of rf electromagnetic radiation on biological systems; and, particularly, no clue to support either tumour induction or tumour promotion can be provided (see also Malyapa, 1997a, 1997b).

3. ORGAN AND WHOLE-BODY EFFECTS: CANCEROGENESIS

UNSCEAR (1994) has reviewed the available epidemiological data concerning cancer induction by ionizing radiations by exposure both from external and from internal sources. The most important information still stems from the survivors of the atomic bombs in Hiroshima and Nagasaki, which is also the core of the new ICRP recommendations (see below). They are supported by quite a number of studies with medical radiation applications and occupational exposures. On the basis of the available data risk estimates for various cancer sites were computed; these are summarized in Table 1.

A new dosimetry system for the atomic bomb survivors was introduced in 1986 (Roesch, 1987), which resulted in a significant reduction of the neutron component compared to the older estimates. In the meantime new figures based on measurements of isotopes produced by neutron reactions became available, casting some doubt on the correct assessment of thermal neutrons. A compilation of the data (Straume *et al.*, 1992) suggests that particularly at larger distances from the epi-

Table 1. Risk of exposure-induced deaths for various cancer sites, estimated on the basis of data for the bomb survivors and cancer incidences in the Japanese population (UNSCEAR 1994). Figures are given in %/Sv and relate to a single acute exposure.

Cancer site	Male	Female	Both
Leukemia	1.3	0.9	1.1
Esophagus	0.3	0.7	0.5
Stomach	0.9	2.0	1.4
Colon	0.5	0.6	0.6
Liver	2.2	0.3	1.2
Bladder	0.4	0.2	0.3
Lung	1.8	3.1.	2.5
Breast	—	2.0	2.0
Ovary	—	0.5	0.3
Others	4.3	2.0	3.1
Total	**11.7**	**12.3**	**12.0**

centre the dose by thermal neutrons is underestimated by factors between at least 2 to 10. This does not drastically change the overall risk factor but may have implications for neutron weighting factors (see below).

4. APPLIED RADIOBIOLOGY

4.1 New Developments in Biodosimetry

4.1.1 Fluorescence *in Situ* Hybridization for Chromosomal Translocations (FISH). Chromosomal aberrations were discussed in the original article. Apart from their fundamental relevance they are an important tool of biodosimetry. If lymphocytes taken from exposed individuals are cultivated *in vitro*, aberrations can be scored and used as an indicator of dose. Classically, dicentric chromosomes ("dics") are used for this purpose since they are comparatively easy to recognize. The lowest dose limit resolvable is about 0.1 Gy with x or gamma rays. Dics interfere with further cellular proliferation; 50% of the cells carrying them will die at every division. This means that these "unstable" aberrations will rapidly fade out so that reliable dose estimates are impossible after months or years. Translocations, on the other hand, are stably transmitted to the progeny, allowing in principle biodosimetry at any time, even after decades. Technical problems of their scoring prevented the use until recently. The introduction of the so-called FISH technique has now led to an improvement: Monoclonal antibodies coupled to fluorescent dyes can be raised against human chromosomes so that every chromosome can be selectively "painted". Translocations are then visualized by colour changes along the chromosomal structures. Studies have shown that they persist indeed much longer than dics. Their dose dependence is similar to that of dics, and their detection limit lies in the same range (about 100 mGy for gamma rays). Developments for automatic scoring are already well under way (Lucas *et al.*, 1992; Yang, 1997).

4.1.2 ESR Dosimetry. Ionizing radiation induces the formation of free radicals in the exposed tissue, but their lifetime is usually very short. This is not the case if they are created in a solid matrix where their movability is severely restricted. In this case they may persist for very long periods and may thus be detected by electron spin resonance (ESR), e.g., in teeth or bone. The method has practical applications both in radiation protections and in radiotherapy (Kudynski *et al.*, 1993).

4.2 Developments in Radiation Therapy

The "classical" radiation therapy of tumours uses quantum or electron radiation provided commonly by linear accelerators. The energy is adjusted in such a way that the dose maximum lies at the tumour site. By careful treatment planning using multiple beams and beam shaping by custom-made absorbers, the tumour volume can be delineated and the healthy tissue spared to a great extent. This technique has now reached a high degree of sophistication. More recently particle therapy has also gained some importance. The most important is based on protons. Since they deliver most of their energy just before the end of their range, a very good dose discrimination can be achieved. There is typically a dose factor of 5 between the so-called "plateau" and the "peak" regions. As the dose maximum is very sharp, small tumours can be very effectively treated with an optimal sparing of the surrounding areas. This technique has been applied especially to ocular melanomas; several thousand patients have so far been treated with excellent results. Not only could the tumour be locally cured, but also eyesight could be saved to a large extent in most cases. While in the beginning treatments were performed at accelerators built for research in physics (the forerunner being Boston, Massachusetts, U.S.A.), more recently "dedicated" machines have been established in specialized hospitals. The largest is situated in Loma Linda, California, U.S.A., commissioned in 1990; others are located in Boston, Massachusetts, U.S.A., Nice, France, and Villigen, near Zürich, Switzerland. Also more deep-seated tumours are now treated, which requires, of course, higher proton energies, which now go up to about 250 MeV (Miller, 1995).

Protons possess a very favourable dose distribution, but their biological effectiveness is not much different from that of gamma or x rays in consequence of their comparably low linear energy transfer (LET). In order to deliver an even higher "biological" dose and

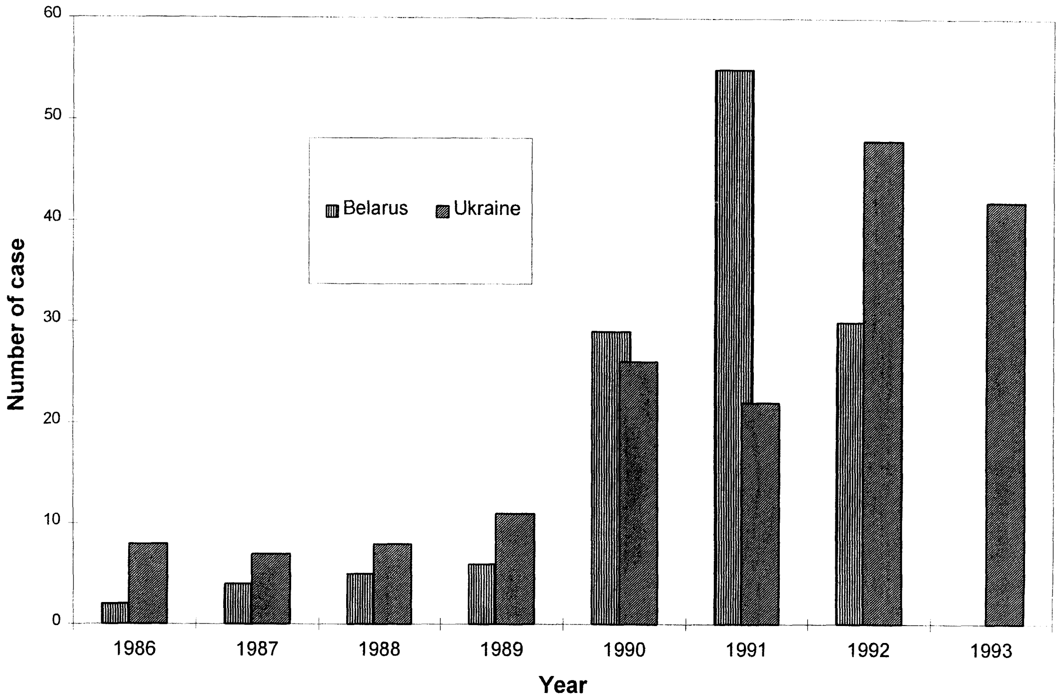

FIG. 1. Thyroid cancers in children up to age 14 in Belarus (Kazakov *et al.*, 1992) and Ukraine (Likhtarev *et al.*, 1995) as a consequence of the Chernobyl disaster. The 1992 value for Belarus is only for half a year.

overcome complications like the "oxygen effect" (see the original article), heavier ions are required. Studies along this line started many years ago in Berkeley, California, U.S.A., but had to be terminated because of the shutdown of the accelerator. Meanwhile two new studies were started, one in Chiba, Japan, where a dedicated heavy-ion accelerator was built, and in Darmstadt, Germany, which makes use of the existing heavy-ion accelerator and is still in the preclinical study phase. A new method for treatment planning has here been developed where the beam energy is modulated in such a way that the treatment volume matches as closely as possible the tumour ("raster scan") (Amaldi and Larsson, 1994).

5. RADIOECOLOGY, RADIATION PROTECTION, AND GUIDELINES

5.1 Effects of the Chernobyl Accident

The 1986 accident in the Soviet reactor in Chernobyl created a widespread radioactive contamination affecting more than 100 000 people. Because of uncertain dose assessment, firm epidemiological data are not readily available, but a few conclusions can now be drawn:

1. Contrary to common perceptions, there is **no** increase in childhood leukaemia, both in the Ukraine and in Belarus (Ivanov *et al.*, 1996; Drozdova *et al.*, 1996).
2. A dramatic increase in thyroid cancers was found in children up to age 14, as shown in Fig. 1.

5.2 Ultraviolet Radiation

All types of skin cancer show a world-wide dramatic increase. This is not only the case for the comparatively harmless basal- and squamous-cell carcinoma but also for the very aggressive melanoma. Figure 2 summarizes data obtained in British Columbia illustrating this alarming development. It is certainly not due to the reduction in stratospheric ozone but

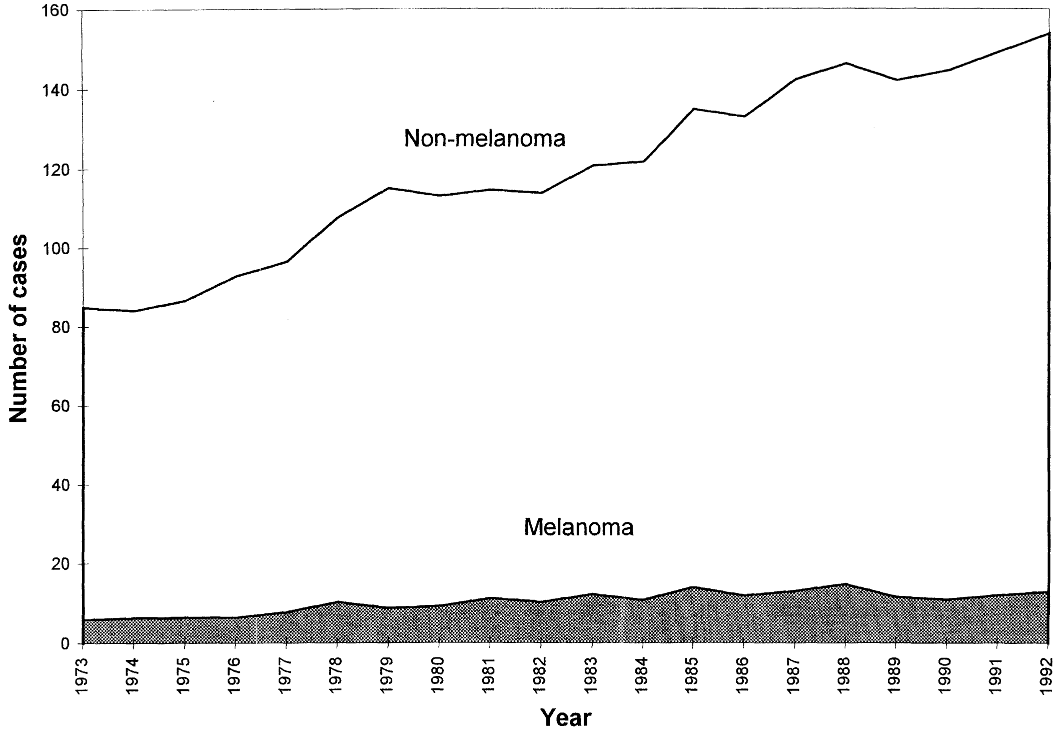

FIG. 2. Increase in non-melanoma and melanoma skin cancer in British Columbia from 1973 to 1992 (Philips *et al.*, 1996).

presumably to changed leisure habits and possibly also the increased use of sunbeds and other tanning devices.

There is a still on-going debate whether near UV (UV-A) has a carcinogenic action. Unfortunately there is not yet an action spectrum available for melanomas in mammals, but action spectra for other types of skin cancer obtained in hairless mice strongly indicate that also long-wavelength UV has clearly a carcinogenic activity.

Ozone depletion will add to this situation. Evidence is mounting that this is not only a theoretical possibility. It can now be seen world-wide, although the most dramatic effect has been registered in Antarctica. Future trends are difficult to predict and depend essentially on the use of chlorofluorocarbon compounds. It has to be kept in mind, however, that atmospheric chemistry operates at long time scales. A comprehensive review of all problems involved can be found in SSK (1996).

5.3 Radiofrequency Electromagnetic Radiation

The US National Research Council (NRC, 1997) has reviewed in a careful and comprehensive study all reports dealing with the question of a possible relationship between power lines and cancer. The result is best summarized by a sentence taken from the Executive Summary:

> *"Specifically, no conclusive and consistent evidence shows that exposures to residential electric and magnetic fields produce cancer, adverse neurobehavourial effects, or reproductive and developmental effects."*

The stress is on "conclusive and consistent," which does not rule out single, as-yet-unexplained findings. Taken together with biophysical considerations reviewed above it may be safely stated that scientific judgement does not provide any possibility for a link between normal environmental levels of electric and magnetic fields and human diseases.

Table 2. Radiation weighting factors (ICRP 60, 1991).

Type and energy range	Radiation weighting factor w_R
Photons, all energies	1
Electrons and muons, all energies	1
Neutrons, energy < 10 keV	5
10–100 keV	10
100 keV–2 MeV	20
2–20 MeV	10
>20 MeV	5
Protons, other than recoil, energy >2 MeV	5
Alpha particles, fission fragments, heavy nuclei	20

5.4 Principles and Fundamentals of Radiation Protection

The International Commission on Radiological Protection (ICRP) issued in 1990 a new set of fundamental recommendations (ICRP 60, 1991) that are likely to have also considerable impact on national legislation. More formally, the term "dose equivalent" has been changed to "equivalent dose" but also to some extent redefined. It is given by the product of dose and radiation weighting factors that replace the "quality factor". Radiation weighting factors are listed in Table 2.

A functional relationship with LET is no longer recommended, to avoid the impression of an accuracy that is inherently not possible. The step function for neutrons is somewhat artificial and commonly replaced by a smooth approximation.

Equivalent dose is measured over a tissue or an organ. To estimate a whole-body equivalent it has to be multiplied by "tissue weighting factors" w_T. This concept had already been used before but the actual values and organs to be considered have been altered, taking into account new medical results from the atomic-bomb survivors.

It has to be emphasized that the system described is only devised for application in occupational radiation protection. Its value in assessing the risk of medical radiation usage is obviously limited.

The ICRP has again reviewed all available evidence for the quantification of radiation carcinogenesis and models for risk projection. As a result new estimates for the expected occurrence of fatal cancer were produced, which must not be taken as firm evidence but rather as upper limits in order to limit possible risks. The numbers have increased considerably over the years, essentially as a result of the application of altered projection models. Consequently also new upper limits are recommended (Table 3). The data from the survivors of Hiroshima and Nagasaki are still the main source of information. Since they received acute doses in a short time, and as there is evidence that lowering the dose rate reduces the effect level, the ICRP introduced a "Dose and Dose Rate Effectiveness Factor" (DDREF) of 2 that is applicable in the common occupational situation. Extended studies with nuclear workers that became recently available are not at variance with the ICPR estimates.

Table 3. Recommended dose limits (ICRP 60, 1991).

	Occupational	Public
Effective dose (Whole body)	20 mSv/year averaged over 5 years	1 mSv/year
Eye lens (Equivalent tissue dose ETD)	150 mSv/year	15 mSv/year
Skin (ETD)	500 mSv/year	50 mSv/year
Hands or feet (ETD)	500 mSv/year	

ETD: Equivalent tissue dose

GLOSSARY

Apoptosis: "Programmed cell death": disintegration of a dying cell characterized by a sequence of distinct metabolic processes.

***Ataxia telangiectasia*:** Recessive hereditary disease in humans characterized by extreme sensitivity to ionizing radiations.

Chromosomal Aberrations: Change in the number of chromosomes (numerical aberration) or in their structure (structural aberrations). Radiation produces mainly the latter.

Dicentric Chromosomes: Chromosomal aberration where two chromosomes are joined in such a way that there are two centromeres (attachment sites for the spindle during mitosis).

DNA: Abbreviation for *deoxyribonucleic acid*, the macromolecule that carries the genetic information in cells.

Dose Equivalent: Formerly, the product of the (physical) dose and the quality factor for different radiation types. It is now replaced by "equivalent dose," which is the product of dose and the newly introduced "radiation weighting factor". If dose is measured in grays (Gy), the equivalent dose is given in sieverts (Sv). The old unit was "rem" with 100 rem = 1 Sv.

Double Strand Breaks (DSB): Rupture of the phosphate–sugar backbone of DNA in both strands opposite to each other. DSB are assumed to be the most important lesions caused by ionizing radiations.

ICRP: International Commission on Radiological Protection, an international non-government organisation that reviews the scientific literature and issues recommendations for radiation protection.

Linear Energy Transfer (LET): Energy imparted per unit path length by a charged particle.

Necrosis: Disintegration of a dying cell.

Recombination: Exchange of parts of genetic material between DNA molecules or chromosomes.

Translocations: Exchange of parts between different chromosomes without loss of genetic material.

Works Cited and Further Reading

Amaldi, U., Larsson, B. (Eds.) (1994), *Hadrontherapy in oncology*, Amsterdam: Elsevier, 755 pp.

Drozdova, V. P., Moroz, G. I., Kiriyewa, S. S., Kuzmina, S. G., Ivanova, I. A. (1996), "Descriptive epidemiology of leukemias in children in the regions of Ukraine caused by the Chernobyl APS accident," *Exp. Oncol.* **18**, 128–131.

Friedberg, E. C., Walker, G. C., Siede, W. (1995), *DNA repair and mutagenesis*, Washington D.C.: ASM Press, 698 pp.

Hagen, U., Harder, D., Jung, H., Streffer, C. (Eds.) (1995), *Radiation Research 1895–1995, 10th ICRR Congress*, Würzburg: 10th ICRR Society, 1236 pp.

Hendry, J. H., Potten, C. S., Merritt, A. (1995), "Apoptosis induced by high- and low-LET radiations," *Radiat. Env. Biophys.* **34**, 59–62.

ICRP 60 (1991), "1990 recommendations of the International Commission on Radiological Protection," *Ann. ICRP* **21**, 1–3.

Ivanov, E. P., Tolochko, G. V., Shavaeva, L. P., Becker, S., Nekolla, E., Kellerer, A. M. (1996), "Childhood leukemia in Belarus before and after the Chernobyl accident," *Radiat. Env. Biophys.* **35**, 75–88.

Kazakov, V. S., Demidchik, E. P. Astakhova, I. N. (1992), "Thyroid cancer after Chernobyl," *Nature* **359**, 21.

Kudynski, R., Kudynski, J., Buckmaster, H. A. (1993), "The application of ESR dosimetry for radiotherapy and radiation protection," *Appl. Radiat. Isotopes* **44**, 903–906.

Likhtarev, I. A., Sobolev, B. G., Kairo, I. A., Tronko, N. D., Bogdanova, T. I., Oleinic, V. A., Epshtein, E. V., Beral, V. (1995), "Thyroid cancer in the Ukraine," *Nature* **375**, 365.

Little, J. B. (1994), "Changing views of cellular radiosensitivity," *Radiat. Res.* **140**, 299–311.

Lucas, J. N., Awa, A., Straume, T., Poggensee, M., Kodama, Y., Nakano, M., Ohtaki, K., Weier, U., Pinkel, D., Gray, J., Littlefield, G. (1992), "Rapid translocation frequency analysis in humans decades after exposure to ionising radiation," *Int. J. Radiat. Biol.* **62**, 53–63.

Malyapa, R. S., Ahern, E. W., Straube, W. L., Moros, E. G., Pickard, W. F., Roti Roti, J. L. (1997a), "Measurement of DNA damage after exposure to 2450 MHz electromagnetic radiation," *Radiat. Res.* **148**, 608–617.

Malyapa, R. S., Ahern, E. W., Straube, W. L., Moros, E. G., Pickard, W. F., Roti Roti, J. L. (1997b), "Measurement of DNA damage after exposure to electromagnetic radiation in the cellular phone communication frequency band (835.62 and 847.74 MHz)," *Radiat. Res.* **148**, 618–627.

Miller, D. W. (1995), "A review of proton beam radiation therapy," *Med. Phys.* **22**, 1943–1953.

NRC (1997), *Possible health effects of exposure to residential electric and magnetic fields*, Washington, DC: National Research Council National Academy Press, 356 pp.

Philips, N. A., Gallagher, R. P., Coldman, A. J., Band, P. R. (1996), "Skin cancer incidence: what has happened and what can we expect," in: *Environmental UV-radiation, risk of skin cancer and primary prevention*, Veröffentlichungen der Strahlenschutzkommission, Band 34, Stuttgart: G. Fischer-Verlag, pp. 323–332.

Roesch, W. C. (Ed.) (1987), *U.S.–Japan Reassessment of Atomic Bomb Radiation Dosimetry in Hiroshima and Nagasaki—Final Report*, Vols. 1 and 2, Hiroshima: Radiation Effects Research Foundation.

Rossi, H. H., Zaider, M. (1996), *Microdosimetry and its applications*, Berlin: Springer, 333 pp.

Skov, K. A., Marples, B. (Eds.) (1994), "Molecular, cellular and genetic basis of radiosensitivity at low doses: a case of inducible repair?," *Radiat. Res. Suppl.* **138**, S1–S131.

SSK (1996), *Environmental UV-radiation, risk of skin cancer and primary prevention*, Veröffentlichungen der Strahlenschutzkommission Band, 34, Stuttgart: G. Fischer-Verlag, 443 pp.

Straume, T., Egbert, S. D., Woolson, W. A., Finkel, R. C., Kubik, P. W., Grove, H. E., Sharma, P., Hoshi, M. (1992), "Neutron discrepancies in the DS86 Hiroshima dosimetry system," *Health Phys.* **63**, 421–426.

UNSCEAR (1994), *Sources and effects of ionizing radiation*, Geneva: United Nations, 272 pp.

Valberg, P. A., Kavet, R., Rafferty, C. N. (1997), "Can low-level 50/60 electric and magnetic fields cause biological effects?," *Radiat. Res.* **148**, 2–21.

Yang, T. C (Ed.) (1997), "Space radiation damage and biodosimetry," *Radiat. Res. Suppl.* **148**, S1–S114.

BIOLOGICAL EFFECTS OF SOUND AND ULTRASOUND

(An Addendum to *Encyclopedia of Applied Physics*, Volume 2, pages 403–420.)

WESLEY L. NYBORG, *Department of Physics, University of Vermont, Burlington, Vermont, U.S.A.*

Introduction 371
1. **Biological Effects of Thermal Origin** 371
1.1 Safety Concerns 371
1.2 Thermal Therapy Using Ultrasound 372
2. **Biological Effects Mediated by Acoustic Cavitation** 373
2.1 Sonoluminescence 373
2.2 Extracorporeal Shock-Wave Lithotripsy 375
2.3 Pulsed Ultrasound 375
2.4 Contrast Agents 376
2.5 Safety Considerations 376
3. **Conclusions** 376
Glossary 376
Works Cited 377
Further Reading 377

INTRODUCTION

Since the original article (hereafter cited as A) was prepared, there has been considerable research activity into biological effects of sound, especially ultrasound, and its applications. Much of this has been motivated by medical applications, both to diagnosis, for which safety guidelines are desired, and to therapy, where the goal is to optimize treatment conditions. For either medical purpose, and for other applications, it is recognized as advantageous to know the mechanisms for the biological effects, whether they are adverse or beneficial. This brief review of selected findings and developments in the past six years is organized according to the mechanisms, both thermal and nonthermal, that are believed responsible for the various kinds of effects under consideration.

Theory for these mechanisms was discussed in A, and is not repeated here. In this Addendum the emphasis is on advances in experimental findings and applications. References are made to the literature for recent developments in theory.

1. BIOLOGICAL EFFECTS OF THERMAL ORIGIN

1.1 Safety Concerns

At frequencies in the megahertz range used in most medical ultrasound, acoustic absorption coefficients for mammalian tissues are high enough to allow temperature elevations produced in medical applications to be significant, even in some applications of diagnostic ultrasound. In tissues where cells divide rapidly, as in the mammalian embryo or fetus, a temperature rise of 2 °C or less can produce damage if prolonged sufficiently. For a wide range of teratogenic effects, the time t in minutes required for exposure to temperature T (in degrees Celsius) to produce the effect has been equated, approximately, to

$$t = 4^{43-T}, \qquad (1)$$

when T is not greater than 43 °C (Miller and Ziskin, 1989; NCRP, 1992); Eq. (1) was

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

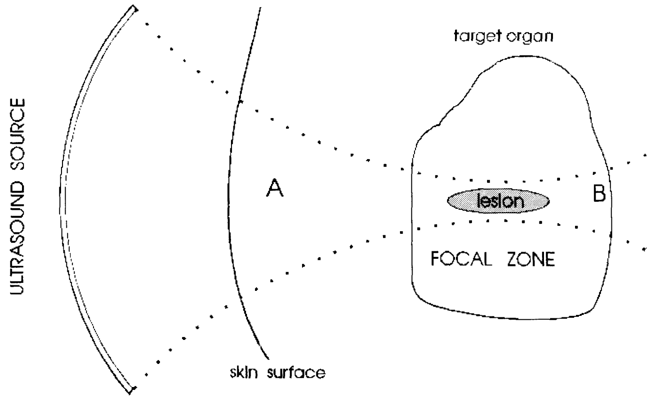

FIG. 1. Schematic diagram of arrangement used in surgical applications of focused ultrasound. A lesion (a region of severely damaged tissue that the body will remove by natural physiological processes) is produced in the focal region, while microscopic examination reveals no damage to tissue lying between the source and the target volume (A) or beyond it (B). Adapted from ter Haar (1995).

adapted from a more general equation given by Sapareto and Dewey (1984). According to Eq. (1), damage may occur after 4 min of exposure to a temperature of 42 °C, and after 16 min of exposure to 41 °C.

Among fetal tissues, the absorption coefficient is highest in bone, and the potential for heating increases with gestational age. Safety criteria derived from algorithms for calculating estimates of the maximum temperature rise that would be produced by ultrasonic exposure of bone and other tissues have been formulated during the past few years (NCRP, 1992; AIUM/NEMA, 1992). These criteria are expressed in terms of *thermal indices*, which are becoming familiar to users of diagnostic ultrasound equipment; many manufacturers of such equipment (in satisfying U.S. regulatory requirements) now display on the imaging screen a "real-time thermal index", i.e., an indicator of the maximum temperature rise being produced in the body of a patient during an ultrasound examination with current instrument settings.

Investigations of the accuracy of algorithms for temperature estimation, while not extensive, have included experimental measurements using special phantoms, as well as measurements made in living animals. The general conclusion is that the thermal indices, while based on estimates of temperature elevation that are subject to error, provide the user with information that is much more relevant to patient safety than was available before the indices were available; errors are usually in the direction of overestimating the temperature rise. It is hoped that in the future the accuracy of the indices can be improved, especially for the more critical situations.

1.2 Thermal Therapy Using Ultrasound

First introduced in the 1940's and 1950's, focused ultrasound has received increasing attention in recent years as a mean of modifying tissue in selected parts of the body; special emphasis has been given to its capability for elevating tissue temperature in a controlled manner. In *hyperthermia* applications, tissues are exposed to ultrasound at moderate intensity levels so that the temperature is maintained at a temperature of 43 to 45 °C for 10 to 30 minutes. In *noninvasive surgery* much higher intensities are used and the temperature raised to the range 60 to 90 °C for 0.1 to 10 seconds. The subject was reviewed recently by ter Haar (1995) and Sanghvi *et al.* (1996), and is featured in a special journal issue (Ebbini, 1996). Ultrasound hyperthermia and surgery are being considered in many fields of medicine, their significant applications at this time being in the fields of ophthalmology and urology. In Fig. 1 is a simplified diagram showing the basic features of an arrangement used in surgery for generating a beam of focused sound and using it to create a *lesion*, i.e., an ellipsoidally shaped volume of damaged tissue in soft tissue. Here the source is a spherical shell of piezoelectric ceramic a few centimeters in diameter, the frequency lies in the range 1 to 8 MHz, and the maximum intensity in the focal region is 300 to 2000 W/cm^2.

In other implementations the single source is replaced by an array of smaller ones, each with its own amplitude and phase; such an array can readily be steered, focused and programmed, so that multiple lesions can be produced in desired configurations. This facilitates the treatment of tumors, for example, by providing the capability of damaging the tissue in a controllable manner throughout regions of various sizes and shapes. Treatment design is greatly assisted by calculations of the expected temperature distribution, with computational or analytical solutions of the bio-heat transfer equation (A; NCRP, 1992; Ebbini, 1996; Sanghvi *et al.*, 1966) or, for very short exposures where the influence of blood flow can be neglected, the equation for thermal diffusion. Ultrasound and magnetic-resonance imaging methods are used to define the region to be treated and to visualize changes that the treatment has produced in the tissue.

2. BIOLOGICAL EFFECTS MEDIATED BY ACOUSTIC CAVITATION

2.1 Sonoluminescence

A major advance in experimental studies of cavitation came from the discovery (Gaitan and Crum, 1990; Gaitan *et al.*, 1992) that a continuous sound field of frequency 22 kHz and pressure amplitude 0.12 MPa can cause a single air bubble in liquid to undergo repeated "collapse events," i.e., implosions, at a regular rate synchronized with the sonic frequency for an indefinitely long period of time, and to generate a short flash of light during each im-

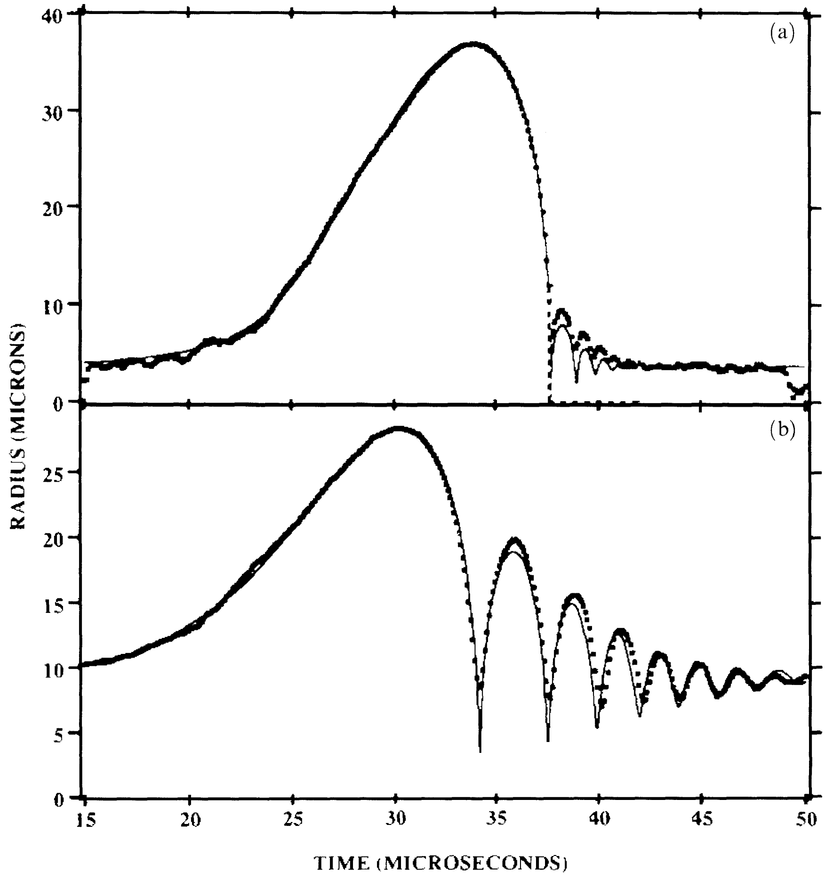

FIG. 2. Bubble radius as a function of time, repeated during each cycle, for a spherical air bubble in water driven by a sound field of frequency 26.5 kHz. The points are experimental; the solid curve is from theory. Sonoluminescence occurs under the conditions of (a), but not under those of (b), the only difference being that the pressure amplitude in the driving field was 0.133 MPa for (a) and 0.108 MPa for (b). From Barber and Putterman (1992).

plosion. These findings caused considerable interest because of the short duration of the flash (often less than 50 ps), the high regularity of the phenomenon, and the nature of the light spectrum. Optical techniques were used to determine the bubble radius R as a function of time t. A typical R *vs* t curve for one cycle of the motion for a frequency of 26.5 kHz is shown in Fig. 2; the points are experimental and the solid curves are from theory. The latter is a solution of the nonlinear equation of motion [similar to Eq. (13) of A]; the excellent agreement gives strong support for the theory. Flynn (1964) has pointed out that the violent collapse events come about because of the inertia of the liquid that rushes inward as the bubble radius decreases. The activity in and near a bubble undergoing the kind of motion illustrated in Fig. 2 has become known as *inertial cavitation* (known earlier as "transient cavitation"). The light production, known as *sonoluminescence* (*q.v.*), is critically dependent on the pressure amplitude in the driving sound field. For example, there was copious light production under conditions of Fig. 2(a) and none under those of Fig. 2(b), yet the conditions were the same except that the pressure amplitude was 0.133 MPa for the former and 0.108 MPa, about 20% less, for the latter. Interesting features of single-bubble sonoluminescence have been discussed by Crum (1994) and Putterman (1995).

Sonoluminescence produced by ultrasound in water or other liquids has been known for many years, and apparently emanates from activity of small bubbles that, in turn, owe their existence to pre-existing microscopic bodies of gas, stabilized against diffusion, called *cavitation nuclei* (see A). Such nuclei are ubiquitous in most liquids unless the latter have been very carefully prepared. An ultrasound field of sufficiently high pressure amplitude may cause free bubbles to emerge from these nuclei and, further, may produce inertial cavitation by causing the bubbles to expand quickly, then contract violently. The optical spectrum of "multibubble sonoluminescence" that appears when ultrasound is applied to water in a test tube or other vessel contains discrete lines characteristic of free radicals, especially O and OH; these are produced by dissociation of water molecules during brief periods of high temperatures (typically, several thousands of kelvins) following the violent contractions. Discrete lines are not seen in the spectra of single-bubble sonoluminescence; instead, a plot of spectral density *vs* wavelength published by Hiller *et al.* (1992) for the light generated by a single air bubble in water is a smooth curve which is fitted well over the 300 to 700 nm wavelength range by a 25000-K blackbody spectrum. Crum (1994) suggests that the conditions of single-bubble inertial cavitation, involving symmetry of the driving pressure field about the bubble center, can lead to production of much higher temperatures in the bubble (which is compressed with little loss of heat to the liquid) than are produced in the more typical experimental conditions where many bubbles are active. In the latter situation, because of high-speed bubble motions and interactions between bubbles, the latter probably do not maintain their spherical shapes as they contract. Instead, available information suggests that the motion during collapse is complicated by surface deformation accompanied by the formation of liquid jets and microbubbles (see A; SONOLUMINESCENCE). It can be expected that, because of the energy used in forming new surface area and because of increased heat transfer from the gas, the pressures and temperatures reached during collapse will be reduced.

There have been many studies in which ultrasound is applied to suspensions of biological cells *in vitro* with the consequence that inertial cavitation occurs and the cells are lysed or otherwise damaged. Sonoluminescence may have occurred, but biological effects have not been attributed directly to the light (even though its spectrum extends into the ultraviolet); instead, the effects have usually been attributed to mechanical or chemical aspects of inertial cavitation (Miller *et al.*, 1996). For example, Miller *et al.* (1995) showed that inertial cavitation generated by ultrasound in suspensions of Chinese hamster ovary (CHO) cells caused H_2O_2 to be produced, which entered surviving cells and caused single-strand DNA breaks. Kim *et al.* (1996) have shown that cavitation produced by continuous 1-MHz ultrasound with pressure amplitude about 400 kPa can make mammalian cells temporarily permeable to large molecules, and appears to provide a promising method of gene transfer. (Note also related findings for lithotriptor-produced shock waves, reported below.)

2.2 Extracorporeal Shock-Wave Lithotripsy

Much has been learned from experiments in which the source is an extracorporeal lithotriptor, a device used medically for breaking up kidney stones by means of focused ultrasound pulses of short duration and high pressure amplitude that are generated outside the patient's body and propagated into it. These high-amplitude pulses are often called *shock waves* and the procedure for applying them is *extracorporeal shock wave lithotripsy* (ESWL). Conditions are favorable for the production of inertial cavitation in water, in the field of a typical lithotriptor, because of the high acoustic pressures (up to 120 MPa) that exist in the focal region during a shock and because of the frequency spectrum, which includes considerable energy in the range 200 kHz and below.

Delius (1995) has reviewed a wide variety of effects produced by ESWL fields on cell suspensions, concluding that these are caused by shear forces resulting from cavitation. A novel finding (Gambihler *et al.*, 1994) is that mouse tumor cells can be made temporarily permeable, allowing large molecules to enter them, after which the cells remain viable. On the basis of such findings, Delius *et al.* (1994) have proposed that ESWL techniques be developed for application to patients as a mean of gene therapy. (Note related findings for ultrasound, reported above.)

Evidence has been obtained that inertial cavitation occurs in the body of patients during ESWL procedures. Coleman *et al.* (1995, 1996) used B-scan ultrasound imaging to examine a region exposed to an ESWL field while the latter was focused on stones in the kidney. In a region (possibly a fat layer) just outside the kidney, through which the shock wave is transmitted into the kidney, increases in image brightness, i.e., echogenicity, appeared when the output setting of the lithotriptor was raised to values above threshold. This increased echogenicity was attributed to gas-filled cavities that had been produced by the field. From hydrophone measurements of the acoustic pressure in water, it was determined that threshold values for the cavity formation were about 15 MPa for the positive peak pressure and 3.5 MPa for the negative peak during passage of a shock.

2.3 Pulsed Ultrasound

Acoustic cavitation, both inertial and non-inertial, capable of causing biological effects, is readily produced by ultrasound in water and in cell suspensions, even with pulsed ultrasound such as is used in medical diagnosis, where frequencies are higher (usually 2 MHz or above) than for a lithotriptor and the focal dimensions are smaller (Deng *et al.*, 1996; Miller *et al.*, 1996).

Even in some mammalian tissues, pulsed ultrasound, similar to that used in diagnostic applications, produces cavitation activity of some kind, though its nature is still under investigation; these are tissues where there are known gas-filled spaces, as in the mammalian lung and gastrointestinal system. The biological endpoint is usually structural damage, made evident by the leakage of erythrocytes from small blood vessels and capillaries on the inner surface of the lung and gastrointestinal tract. Systematic studies of lung damage produced by pulsed focused ultrasound have yielded information on thresholds for damage, in terms of the peak negative pressure p. All threshold values lie within an order of magnitude of 1 Mpa, but there is variation with age and species of the animal, and with frequency and pulse duration of the ultrasound. See, for example, Carstensen *et al.* (1992), Baggs *et al.* (1996), and Holland *et al.* (1996). Surprisingly, the thresholds for damage to mammalian lung are in the same range as thresholds for damage to fruit-fly larvae and to leaves of the plant *Elodea* (AIUM, 1993). The common element to all of these biological systems is the presence of stabilized gas-filled cavities; i.e., lung in the mammal, respiratory channels in the larvae, and intracellular gas-filled channels in the plant. For the last, the primary mechanism for damage appears to be viscous stress from acoustic microstreaming produced near vibrating intracellular channels (Miller, 1987; A). The detailed mechanisms for damage caused by ultrasound to mammalian lung or to insect larvae are, as yet, unknown.

2.4 Contrast Agents

A recent development in diagnostic ultrasound procedures is the increasing use of contrast agents (de Jong, 1996). These agents take various forms, but those being considered here are particulates that, when injected into the blood stream of a patient, cause small bodies of gas to exist there. These gas-filled cavities are surrounded by fluid, or are encapsulated by elastic shells or by viscoelastic films; they are very effective as scatterers of the incident ultrasound (at moderate pressure amplitudes) or as emitters of cavitation-generated harmonics (at higher pressure amplitudes), and thus add information to an ultrasound scan.

Contrast agents also can contribute to biological effects of ultrasound. It was found in experiments with the commercial product Albunex® that these gas-containing particles can serve as nuclei for cavitation and, when added to a cell suspension, can greatly increase the destruction of cells when the latter are exposed to pulsed ultrasound (Brayman *et al.*, 1996) or when they are exposed to lithotriptor shock waves (Miller and Thomas, 1996). Studies are in progress on the extent to which damage occurs to living mammals upon exposure to ultrasound when contrast agents are present in the circulatory system. The mechanisms for the action of contrast agents are not known in detail, but evidently include various aspects of inertial and noninertial cavitation.

2.5 Safety Considerations

Theoretical and experimental results have led to safety criteria, analogous to those developed for thermal mechanisms (NCRP, 1992), which define conditions under which tissue damage might occur as a result of cavitation. A criterion adopted by the Output Display Standard (AIUM/NEMA, 1992) is expressed in terms of a *Mechanical Index* (MI), the latter being defined as the ratio of the peak acoustic pressure, defined in a specified way and expressed in megapascals, to the square root of the center frequency, the latter expressed in megahertz. A standard formula is used for calculating the peak negative acoustic pressure in the patient's body, given the results of measurements in water. It is assumed that damage does not occur if the MI is less than 1.0 (or other assigned number). Many manufacturers provide for display of the "real-time MI", i.e., the MI appropriate for current instrument settings, on the screen of ultrasound imaging equipment, in satisfying U.S. regulatory requirements.

3. CONCLUSIONS

Biological effects of ultrasound are receiving increased attention, partly because of their relevance to safety of medical procedures, and partly because of the diverse therapeutic applications which are possible because of the ability of ultrasound to change the structure and function of cells and tissues. For continued progress, it will be necessary to achieve much better understanding of the mechanisms involved in the applications described above, and in others which it has not been possible to discuss in this brief review.

GLOSSARY

Acoustic Cavitation: The mechanical response of one or more gas-filled cavities to a sound field; the cavitation may be inertial or noninertial.

Cavitation Nuclei: Small bodies, usually gas-filled cavities, that can serve as sites for acoustic cavitation.

Extracorporeal Shock-Wave Lithotripsy (ESWL): A procedure used for destruction of kidney stones, involving production of short ultrasound pulses (with very high acoustic pressures) outside the body of a patient and focusing them on the kidney.

Hyperthermia: A procedure for killing tumor cells by raising their temperature to a critical value.

Inertial Cavitation: A class of acoustic cavitation involving growth and violent collapse of one or more cavities.

Lesion: A region of damaged tissue, such as can be produced by focused ultrasound under specific conditions.

Mechanical Index: A number calculated according to standards accepted in the U.S.A., used as a safety index for alerting the user of diagnostic ultrasound equipment to conditions under which damaging acoustic cavitation might occur in the body of a patient.

Noninvasive Surgery: A procedure, such as an application of focused ultrasound, for creating damaged tissue, especially in a tumor, which will subsequently be removed by natural processes.

Sonoluminescence: Production of light by sound in water or other liquid, usually as a consequence of inertial cavitation.

Thermal Index: A number calculated according to standard algorithms, used as a safety index for alerting the user of diagnostic ultrasound equipment to conditions under which damaging temperature elevation might occur in the body of a patient.

List of Works Cited

AIUM (1993), *Bioeffects and Safety of Diagnostic Ultrasound*, Laurel MD: American Institute of Ultrasound in Medicine.

AIUM/NEMA (1992), *Standard for Real-Time Display of Thermal and Mechanical Indices on Diagnostic Ultrasound Equipment*, Laurel, MD: American Institute of Ultrasound in Medicine, and Washington, DC: National Electrical Manufacturers Association.

Baggs, R., Penney, D. P., Cox, C., Child, S. Z., Raeman, C. H., Dalecki, D., Carstensen, E. L. (1996), *Ultrasound Med. Biol.* **22**, 119–128.

Barber, B. P., Putterman, S. J. (1992), *Phys. Rev. Lett.* **69**, 3839–3842.

Brayman, A. A., Azadniv, M., Cox, C., Miller, M. (1996), *Ultrasound Med. Biol.* **22**, 927–938.

Carstensen, E. L., Duck, F. A., Meltzer, R. S., Schwartz, K. Q., Keller, B. (1992), *Echocardiography* **6**, 605–623.

Coleman, A. J., Kodama, T., Choi, M. J., Adams, T., Saunders, J. E. (1995), *Ultrasound Med. Biol.* **21**, 405–417.

Coleman, A. J., Choi, M. J., Saunders, J. E. (1996), *Ultrasound Med. Biol.* **22**, 1079–1087.

Crum, L. A. (1994), *Phys. Today* **47** (9), 22–29.

de Jong, N. (1996), *IEEE J. Eng. Med. Biol.* **15**, 72–82.

Deng, C. X., Xu, Q., Apfel, R. E., Holland, C. K. (1996), *Ultrasound Med. Biol.* **22**, 939–948.

Delius, M. (1994), *Shock Waves* **4**, 55–72.

Delius, M., Hofschneider, P.-H., Lauer, U., Messmer, K. (1995), *Lancet* **345**, 1377.

Ebbini, E. S. (1996), *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **43**, 989–1129.

Flynn, H. G. (1964), "Physics of acoustic cavitation in liquids", in: W. P. Mason (Ed.), *Physical Acoustics*, Vol. 1B, New York: Academic Press, pp. 57–172.

Gaitan, D. F., Crum, L. A. (1990), in: M. F. Hamilton, D. T. Blackstock (Eds.), *Frontiers of Nonlinear Acoustics: Proceedings of 12th ISNA*, London: Elsevier Science Publishers pp. 459–463.

Gaitan, D. F., Crum, L. A., Church, C. C., Roy, R. A. (1992), *J. Acoust. Soc. Am.* **91**, 3166–3183.

Gambihler, S., Delius, M., Ellwart, J. W. (1994), *J. Membrane Biol.* **141**, 267–275.

Hiller, R., Putterman, S. J., Barber, B. P. (1992), *Phys. Rev. Lett.* **69**, 1182–1184.

Holland, C. K., Deng, C. X., Apfel, R. E., Alderman, J. L., Fernandez, L. A., Taylor, K. J. W. (1996), *Ultrasound Med. Biol.* **22**, 917–925.

Kim, H. J., Greenleaf, J. F., Kinnick, R. R., Bronk, J. T., Bolander, M. E. (1996), *Human Gene Therapy* **7**, 1339–1346.

Miller, D. L. (1987), *Ultrasound Med. Biol.* **13**, 443–470.

Miller, D. L., Thomas, R. M. (1996), *Ultrasound Med. Biol.* **22**, 1089–1095.

Miller, D. L., Thomas, R. M., Buschbom, R. L. (1995), *Ultrasound Med. Biol.* **21**, 841–848.

Miller, M. W., Ziskin, M. C. (1989), *Ultrasound Med. Biol.* **15**, 707–722.

Miller, M. W., Miller, D. L., Brayman, A. A. (1996), *Ultrasound Med. Biol.* **22**, 1131–1154.

NCRP (1992), *Exposure Criteria for Medical Diagnostic Ultrasound*, Report No. 113, National Council on Radiation Protection and Measurements, Bethesda, MD: NCRP Publications.

Putterman, S. J. (1995), *Sci. Am.* **272** (2) 46–51.

Sanghvi, N. T, Hynynen, K., Lizzi, F. L. (1996), *IEEE J. Engin. Med. Biol.* **15**, 83–92.

Sapareto, S. A., Dewey, W. C. (1984), *Int. J. Radiat. Oncol. Biol. Phys.* **10**, 787–800.

ter Haar, G. (1995), *Ultrasound Med. Biol.* **21**, 1089–1100.

Further Reading

AIUM (1993), *Bioeffects and Safety of Diagnostic Ultrasound*, Laurel, MD: American Institute of Ultrasound in Medicine.

Crum, L. A. (1994), *Phys. Today*, **47** (9) 22–29.

Ebbini, E. S. (Ed.) (1996), Special Issue on Therapeutic Ultrasound, *IEEE Trans. Ultrason. Ferroelec. Freq. Control* **43**, 989–1129.

Leighton, T. G. (1994), *The Acoustic Bubble*, London: Academic Press.

NCRP (1992), *Exposure Criteria for Medical Diagnostic Ultrasound*, Report No. 113. National Council on Radiation Protection and Measurements, Bethesda, MD: NCRP Publications.

Nyborg, W. L. (1991), "Biological effects of sound and ultrasound," in: G. L. Trigg (Ed.), *Encyclopedia of Applied Physics*, Vol. 2, New York: VCH, pp. 403–420.

CHARACTERIZATION AND ANALYSIS OF MATERIALS

(An Addendum to *Encyclopedia of Applied Physics*, Volume 3, pages 215–240.)

JOHN P. SIBILIA, *Sibilia Associates Inc., Livingston, New Jersey, U.S.A.*

1. **Optical Microscopy and Image Analysis** 379
2. **Molecular Spectroscopy** 379
2.1 Infrared Spectroscopy 379
2.2 Raman Spectroscopy 381
2.3 Nuclear Magnetic Resonance Spectroscopy 381
2.4 Mass Spectrometry 382
3. **Chromatography** 382
3.1 Supercritical Fluid Chromatography 382
3.2 Capillary Electrophoresis 382
4. **Scanning Probe Microscopy (Scanning Tunneling and Atomic force Microscopy)** 382
Works Cited 384

1. OPTICAL MICROSCOPY AND IMAGE ANALYSIS

Considerable strides have been made in recent years in developing quantitative image analysis (QIA) (Russ, 1990). Faster and greater processing-capacity computers and improvements in software have made the acquisition, digitization, storage, and image display processes much more efficient. For example, one of the most widely used applications is in the analysis of particle size distributions and particle shapes (Li *et al.*, 1995). The improved computational developments have resulted in improved resolution by enhancing the contrast between particles and displaying a variety of particles in a mixture by discrete colors. These techniques have been applied to the analysis of polymer composites, metal alloys, ceramic mixtures, and various combinations of polymer, metal, and ceramic composites (Sibilia, 1996). In addition to information on the particle sizes and shapes, QIA also provides information on the degree of mixing of the components, their ability to "wet", and their packing efficiency. Contiguity, a parameter which measures the degree of interaction between neighboring particles, is also determinable by QIA.

QIA has also been applied in the medical and biological sciences. The various features of cell structures have been much more clearly elucidated and quantified by image analysis. Furthermore, the technique has been applied to the analysis of virus-containing tissues and the micromolecular diffusion in skin.

2. MOLECULAR SPECTROSCOPY

2.1 Infrared Spectroscopy

The infrared dispersive spectrometers have essentially been replaced by Fourier-transform infrared (FTIR) spectrometers. These instruments utilize an interferometer in place of a monochromator to achieve the separation of absorbed frequencies when infrared light is passed through a sample (Ferraro and Krishnan, 1990). The advantages of FTIR are several. They have much higher source-radiation throughput, increased signal/noise ratio, reduced measurement times, and higher wavelength accuracy. These advantages have allowed infrared spectroscopy to be used in a number of new applications, as well as the conducting of more standardized experiments in a more efficient manner (Coleman, 1993).

One of the areas where FTIR has widened the scope of applications in infrared spectroscopy is in surface analysis. Infrared spectroscopy in general can be used to provide information about the molecular structure of the surfaces of materials. There are three available

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

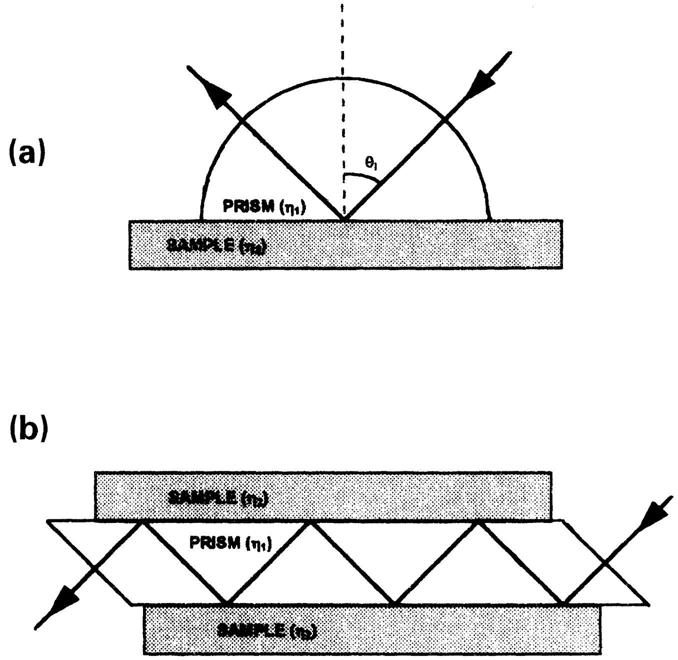

FIG. 1. Schematic of sampling apparatus for ATR. (a) Single-reflection hemispherical element; (b) multiple-reflection parallelogram element. From Sibilia (1996), p. 20, Fig. 2.1.1.

techniques, which provide information at three different depth levels in materials and for different kinds of materials. The most common of these FTIR surface analytical techniques is attenuated total reflectance (ATR), which probes the depth in the range of about 0.5 to 3 μm. In this technique the sample is held in contact with a prism as shown in Fig. 1. In Fig. 1(a), a single-reflection hemispherical prism is shown while in 1(b), a multiple-reflection prism is used with the sample pressed against both prism surfaces. Light entering the prism (usually KRS-5, or some other high-refractive-index, infrared-transparent inorganic crystal) undergoes internal reflection at the prism surfaces. At the interface, the light penetrates a small distance into the sample. The depth of penetration depends on the angle of incidence of the light beam and the difference between the index of refraction of the prism (n_1) and that of the sample (n_2) (Fig. 1). The light exiting the prism is attenuated at specific frequencies because of absorption by the sample. The resulting spectrum is similar to that obtained by transmission infrared spectroscopy. The sensitivity of the ATR technique is enhanced if the multiple-pass configuration [Fig. 1(b)] is used rather than single-pass mode shown in 1(a). Applications of the ATR technique include the surface analysis of thin films, adhesives, surface contaminants, surface compositions of materials, and migration of substances to the surface of materials.

Another infrared spectroscopic surface technique is photoacoustic spectroscopy (PAS) (*q.v.*). The sampling depth of PAS is about 1 to 10 μm. In this technique the incident infrared energy that is absorbed by the sample is converted into heat. Gas, which is in contact with the sample surface, is heated by the sample and expands. The expansion of the gas produces an acoustic signal when the incident beam intensity is modulated at a frequency in the acoustic range. Microphonic detection of this signal can be made to produce an IR spectrum when it is processed by the detector amplification electronics of an FTIR spectrometer. PAS is particularly useful for examining samples as received without any further preparation. For example, powders, pellets, films, opaque materials, and rough-surfaced materials can all have their surfaces analyzed by PAS.

The third infrared surface analytical technique, which is capable of examining ultrathin (monolayer) films, is grazing incidence reflectance (GIR) spectroscopy. In this methodology the IR beam intersects the surface at a high angle of incidence. This technique is highly sensitive but has been limited to the examination of only metallic surfaces.

Another infrared technology which has been expanded through the use of FTIR is infrared microspectroscopy. Infrared spectra, with a spatial resolution of about 10 μm, can be achieved by coupling a high-quality optical microscope with an FTIR spectrometer. Spectra in both transmission and reflectance modes can be obtained. It is possible to perform localized surface analyses *in situ* using specialized ATR and GIR objectives at lower spatial resolution (about 25 μm). Mapping of individual components can also be achieved by using an x-y translation stage. Infrared microspectroscopy has been applied in analysis of contaminants or defects on surfaces of ceramic, metallic, and polymer substrates; gels and defects in pharmaceutical tablets and packages; and electronic failures; and in forensic analysis of fibers, pigments, dirt etc.

Another 'combination' technique that has been extensively applied is chromatography-FTIR. Components in the microgram and subnanogram range have been identified by this combination technique. In the GC-IR mode, compounds can be identified in the gas phase in real time. Techniques such as liquid chromatography/IR or gel permeation chromatography/IR require that the sample first be separated and collected on a disc and then analyzed.

The study of kinetic systems and transient species has also been enhanced through the application of FTIR techniques. Spectra can be obtained in times as short as 1/20 of a second with the use of conventional FTIR. If the spectrometer is operated in a step-scan mode, microsecond and even nanosecond time domains can be achieved. This allows for real-time studies of chemical reactions to detect unstable intermediates, establish reaction mechanisms, and determine kinetic parameters.

2.2 Raman Spectroscopy

Both dispersive and Fourier-transform spectrometers are used routinely for Raman measurements. A choice of incident laser frequencies, ranging from the ultraviolet to visible regions of the electromagnetic spectrum, is usually available. FT-Raman spectrometers use the same principles as FTIR but operate in the near-IR region of the spectrum and use a near-IR laser at 1.06 μm. The advantage of the FT-Raman spectrometers is that sample fluorescence can be reduced. Sample fluorescence is a common problem in Raman spectroscopy; it results in masking of the spectrum and sometimes creates totally unusable spectra.

One recent application of Raman spectroscopy is in the analysis of carbon/graphite-containing materials. Infrared spectroscopy is of little use in the analysis of such materials since they absorb too strongly throughout the IR region and provide little structural information. Raman spectroscopy, on the other hand, can distinguish between the different forms of graphite (determined by their different degrees of graphitization) and diamond.

2.3 Nuclear Magnetic Resonance Spectroscopy

Significant advances have been made in multi-dimensional high-resolution NMR. Two-dimensional NMR (2D NMR) is a methodology that provides molecular structural information in a simple visual display. High-resolution one-dimensional NMR spectra of even relatively low–molecular-weight compounds can be complex on account of the overlapping array of signals arising from a multitude of different chemical shifts and spin–spin interactions. When a second frequency dimension (2D NMR) or in some cases even 3D or 4D is utilized, structural information becomes easier to extract because the interactions between specific nuclei in the molecule become more evident.

Advances in the application of cross-polarization (CP) and high-power decoupling (DD) have also made solid-state NMR a more valuable technique. These and other advances have accelerated the development of multidimensional solid-state NMR to applications in polymers (Mathias, 1991) as well as small-molecule organic and inorganic compounds. Magnetic resonance imaging of materials is another field that is also rapidly developing in solid-state NMR.

2.4 Mass Spectrometry

Two relatively new techniques in the field of structural mass spectrometry have been used to analyze the structures of high-molecular weight polymers. They are plasma desorption mass spectrometry (PDMS) and matrix assisted laser desorption mass spectrometry (MALDI). Both techniques utilize a pulse of energy to create a thermal spike in the sample, which in turn results in the ejection of both positively and negatively charged ions into the gas phase. Time-of-flight mass analyzers are now generally used for both PDMS and MALDI. Time-of-flight mass spectrometers have only moderate mass-resolving power but have no constraints on mass range. They are also ideally suited for this application because they have the ability to detect all the ions ejected by the ion source.

3. CHROMATOGRAPHY

3.1 Supercritical Fluid Chromatography

Supercritical fluid chromatography (SFC) is yet another type of chromatography in a growing field that is more appropriately known as separation sciences. The technique requires that the sample to be analyzed must be soluble in a solvent for dilution and injection purposes and must also be soluble in the mobile phase used in the SFC method. The mobile phase is a supercritical fluid, i.e., a "gas" above its critical pressure and above its critical temperature. Most analyses are carried out at room temperature or slightly higher but rarely above 100 °C. Sensitivities by this methodology are in the parts-per-million range or better depending on the method of detection used and the nature of the sample. The principle of the method is similar to that of other chromatographies where components are separated by their ability to partition between the mobile phase and the stationary phase, which is a component of the column. Some of the supercritical fluids used in SFC are carbon dioxide, ammonia, monochlorodifluoromethane, and dichlorodifluoromethane. SFC has advantages over GC and HPLC in certain analytical applications, but it does not share the wide range of applicability of these two well established techniques.

3.2 Capillary Electrophoresis

High-performance capillary electrophoresis (HPCE) (see Capillary Electrophoresis in this Supplement) is a technique that combines the strength of high-performance liquid chromatography (HPLC) and conventional electrophoresis to yield rapid, precise, automated, and highly efficient analysis of complex mixtures (Wenclawiak, 1992). It is particularly useful for analyzing both organic and inorganic compounds in aqueous media. Since HPCE utilizes such small sample sizes (nanoliters and less), it is most effective as an analytical rather than a preparative technique.

The basic principles of HPCE are rather simple. A small plug of sample solution is introduced into one end of a fused-silica capillary tube that contains an appropriate electrolyte. Upon application of a large electrical field to the solution flowing through the capillary column, the molecules in the sample migrate at different rates through the column. A number of different detectors such as ultraviolet/visible, fluorescence, conductivity, or mass spectrometry are used to detect the effluent molecules. The technique has found wide use in a broad range of applications, especially in the pharmaceutical industry.

4. SCANNING PROBE MICROSCOPY (SCANNING TUNNELING AND ATOMIC FORCE MICROSCOPY)

Scanning probe microscopy (SPM) (see Tunneling Microscopy and Spectrometry) is the term associated with a variety of techniques which measure the topography of surfaces at the atomic level. Scanning tunneling microscopy (STM) and atomic force microscopy (AFM) are two of the earlier SPM methodologies that were described in the original article. In Fig. 2 are schematic drawings showing the basic operations of an STM and an AFM. In STM the piezoelectric scanner probe passes over the surface of a relatively flat conducting sample. If a potential differ-

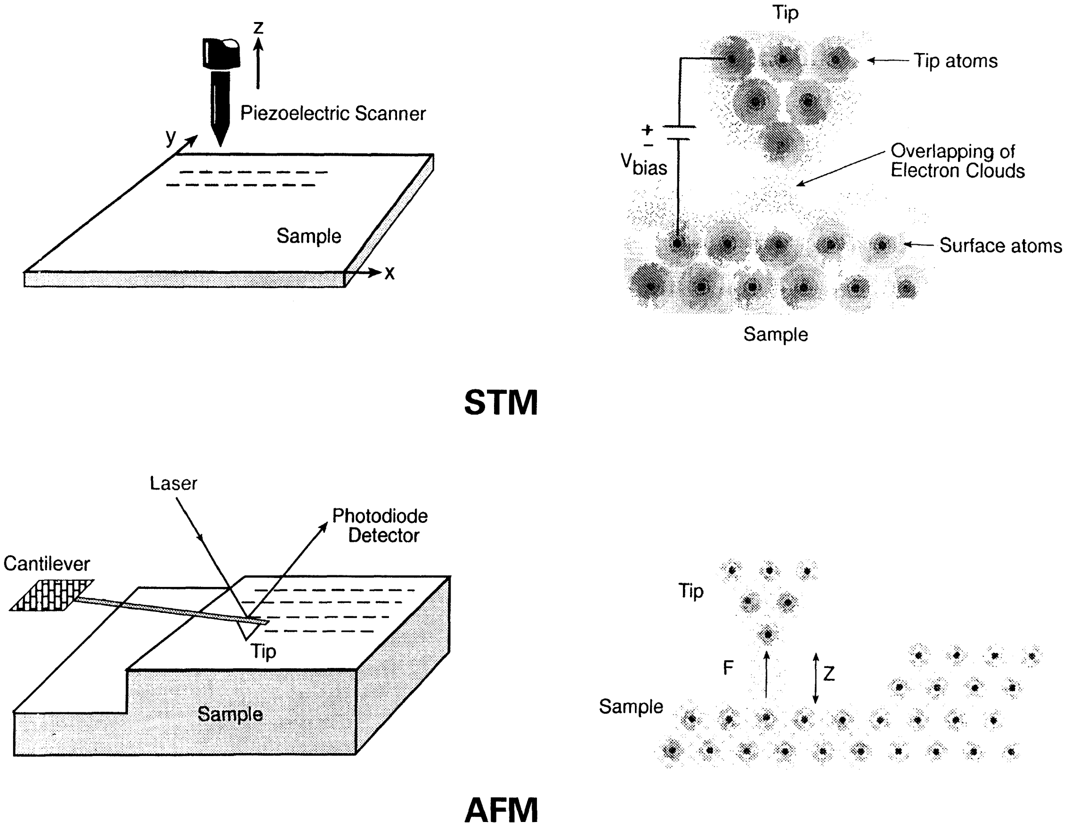

STM

AFM

FIG. 2. Schematic of the STM and AFM setups. From Sibilia (1996). p. 190, Fig. 8.18.

ence is maintained between the tip and sample surface as illustrated in Fig. 2, a current will be generated if the tip passes very close to the sample surface. The current is generated by electrons "tunneling" their way through the surface of the sample and through the air gap to the probe tip. As the tip is traversed over the surface it will experience an up-and-down motion corresponding to the topography of the surface at the atomic level. The motion of the probe can be sensed by a transducer and in turn can be converted to a contour map of the surface. This is only one mode of operation, and it involves feedback to the (vertical) positioning mechanism based on the magnitude of the tunneling current. In the AFM technique, the electronic forces between the probe tip and the sample surface cause a deflection in the cantilever arm. The deflections can be detected by means of a laser beam reflecting off the cantilever and into a photodiode detector. A mapping of the cantilever deflections as the probe is traversed over the sample surface results in a map of the atomic level surface topography. Basically, all the SPM techniques consist of a probe which scans the surface of a material and measures the surface topography and the specific surface interaction, which is dependent on the nature of the probe and sample type of surface being studied. Three more recent techniques (in addition to STM and AFM) to join this expanding field of SPM are magnetic force microscopy (MFM), scanning thermal microscopy (SThM), and scanning near-field optical microscopy (SNOM).

MFM is a variation of AFM that uses a ferromagnetic probe to detect the forces exerted on the tip by the stray magnetic fields of the sample. While AFM operates in a contact mode, the tip does not contact the surface in the MFM mode. Both surface topography signals, which dominate at close distances to the surface, and magnetic signals, which are more prevalent at greater distances from the surface, can be detected by moving the probe closer to or further away from the surface. This is important to establish the relationship between the topographic and magnetic data. Resolution down to 50 nm and the noncontact/nondestructive nature of MFM make this a useful addition to the SPM arsenal.

SThM uses the sample's thermal conductivity characteristics as a contrast mechanism in imaging microscopic features (Dinwiddie *et al.*, 1994). There are several variations of the thermal probe. It can be made with two wires to form a thermocouple junction at the scanning tip or it can comprise an integrated resistive thermal probe with the AFM cantilever. The probe is operated in a constant-force mode as in an AFM. SThM can collect and display topographic and thermal (conductivity or temperature) data simultaneously. Like the MFM technique, which provides both topographic and surface magnetic information, SThM provides both topographic and basic information about the conductivity/temperature characteristics of the sample surface.

SNOM is a scanning probe technique that provides optical microscopic information at a resolution better than 100 nm. The optical image is generated by means of a subwavelength optical probe that is held in close proximity to the sample (i.e., in the near field). The near-field interaction, which is based on fiber-optic waveguide theory, is monitored while the probe is scanned across the sample surface. The SNOM technique, which does not use lenses, has improved resolution over conventional optical microscopy since the diffraction limitations due to the use of lenses are eliminated. The technique has been finding application in the semiconductor, bioscience and magneto-optical device fields.

List of Works Cited

Coleman, P. B. (1993), *Practical Sampling Techniques for Infrared Analysis*, Boca Raton: CRC Press.

Dinwiddie, R. B., Pylkki, R. J., West, P. E. (1994), *Thermal Conductivity* **22**, 668–670.

Ferraro, J. R., Krishnan, K. (1990), *Practical Fourier Transform Infrared Spectroscopy*, New York: Academic Press.

Li, C.-W., Lui, S.-C., Goldacker, J. (1995), *J. Am. Ceram. Soc.* **78** (2), 449–450.

Mathias, L. J. (1991), *Solid State NMR of Polymers*, New York: Plenum Press.

Russ, J. C. (1990), *Computer-Assisted Microscopy, The Measurement and Analysis of Images*, New York: Plenum Press.

Sibilia, J. P. (Ed.) (1996), *A Guide to Materials Characterization and Chemical Analysis*, 2nd ed., New York: Wiley-VCH, Chapters 8 and 9.

Wenclawiak, B. (1992), *Analysis with Supercritical Fluids: Extraction and Chromatography*, New York: Springer.

CHEMICAL ANALYSIS

(An Addendum to *Encyclopedia of Applied Physics*, Volume 3, pages 307–343.)

KENNETH A. RUBINSON, *The Five Oaks Research Institute, Cincinnati, Ohio, U.S.A.*

Introduction 385

1. **Parallel Collection of Analytical Information** 385

2. **Electrophoresis** 386

2.1 Gel Electrophoresis 387

2.2 Capillary Electrophoresis 387

3. **Mass Spectrometry** 388

4. **Sample Preparation** 388

Further Reading 389

INTRODUCTION

Steady improvement in practically every area of chemical analysis has occurred since the original article appeared in this Encyclopedia. However, some areas have progressed so fast that little is similar over that time span. This update focuses on four major changes that have occurred:

1. in spectrometry, where a larger fraction of the light available for analyses can be collected through the use of multichannel monitoring;
2. in electrophoresis, where running the separations in capillary tubes has resurrected this old method of separations;
3. in mass spectrometry, where molecules with molecular weight up to 100 000 can be brought into the gas phase; and
4. in sample preparation, where microwave heating has helped to speed sample preparation to match the ever greater throughput of advanced instruments.

1. PARALLEL COLLECTION OF ANALYTICAL INFORMATION

A quiet revolution in detectors has occurred with the advent of silicon solid-state charge-transfer devices (CTDs), which include charge-coupled devices (*q.v.*) (CCDs). Contemporary CCDs now outperform photomultiplier tubes in nearly every application. Each photon that hits the CCD produces one (in the near-IR to visible range) or more (in the UV and x-ray ranges) electron-hole pairs, which are stored inside the volume under the location where the photon hit the detector. These charges are stored until the packets of charge from each pixel are read out individually and amplified. Figure 1 shows the difference between using a single detector such as a photomultiplier in conjunction with a monochromator in a scanning instrument and a CCD array in a dispersive, nonscanning instrument (a polychromator).

The scanning instrument gets it name because over time it scans the colors past a slit that selects a narrow range of the whole to produce a spectrum (a plot of light power *vs.* wavelength). However, the scanning instrument makes inefficient use of the light entering it. Most of the light—that not at the specific wavelength falling on the detector—is wasted. However, when the entire spectrum can be recorded simultaneously, far more signal is recorded per unit time. This inherently better performance is designated the multichannel advantage. Closely related is the multiplex advantage of infrared spectrometry.

Polychromators are used to great benefit in atomic emission spectrometry. Early polychromators had up to ten photomultiplier tubes set to record the lines from ten elements simultaneously. In contrast, contemporary instruments can quantitate essentially all the common elements in a sample in about a minute through a combination of the CCD-based polychromator and contemporary small computers. Improvement in the signal-to-noise ratio over the older methods can be up to a factor of 100.

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

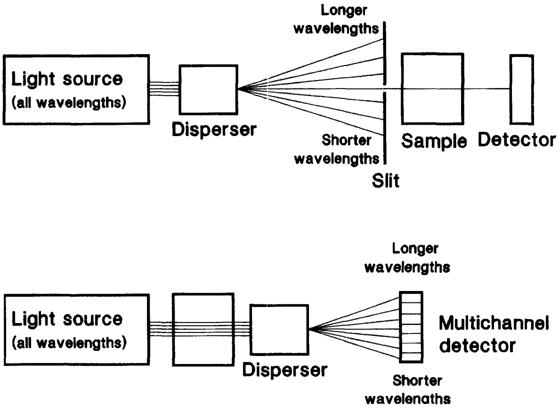

FIG. 1. Diagram of the general differences between a scanning spectrometer (top) and a polychromator system (bottom). The disperser consists of one or more gratings. The scanning spectrometer scans sequentially through the wavelengths. Here, it is shown in the middle of the range. The multichannel spectrometer has optics that are fixed in position.

Separations provide another application where polychromators are rapidly being applied. Compounds are separated in time (and volume) by column liquid chromatography (see CHROMATOGRAPHY) and then pass sequentially through a detector. The volumes containing the individual compounds are called bands. If the band holds a single compound, the spectrum is the same throughout. However, if a band is not pure, but composed of two compounds, it is likely that one is enriched at the leading edge and the other is enriched at the trailing edge. Two different spectra are seen at the edges and a combination of the two in the middle. The benefits of multiplexing are so great that multidimensional detectors have become the method of choice, since no more time is necessary for the chromatographic separation to be run. After all, why should all that extra data not be collected if it can be used to answer any questions that arise about the purity, identity, and properties of the compounds in each band?

2. ELECTROPHORESIS

Separations of different species based on their dissimilar behavior under the influence of an applied electric field are classified generally as electroseparations. There are six phenomenological rules that describe the factors that underlie the various types of electroseparations.

1. Negative ions tend to migrate toward the higher (positive) voltage and positive ions tend to migrate toward the lower (negative) voltage.
2. The average rate of migration of an analyte species is proportional to the average charge on the ion.
3. The average rate of migration of an analyte species is proportional to the average voltage applied.
4. The average rate of motion of an ion decreases as its average cross section increases. (The cross section is the area of the ion that lies perpendicular to the direction of the motion.)

The ability to separate slightly different molecules, such as deoxyribonucleic acids (DNAs) that differ by one base unit, is severely compromised by convection, however. Convection is the macroscopic mixing of liquids due to differences in density, and heating causes much convection in solutions. So,

5. the quality of electroseparations in liquids improves as the convection can be suppressed.

Heating of a liquid in electroseparations *cannot* be avoided since the electric current passing through the solution heats it. Finally,

6. the quality of electroseparations in liquids improves when the temperature can be held constant over the time of the experiment.

There are two ways to suppress convection in a liquid when gravity is present:

A. running the electroseparation in a gel, where the molecular strands of the gel stop macroscopic convection by trapping the liquid between the strands, and

B. running the electroseparation inside a small ($\sim 100\,\mu\text{m}$) capillary, where the walls of the capillary have the same beneficial effect as the gel molecules. In addition, the small size of the capillary lowers the temperature differences that can arise since the heat can leave relatively easily across the short distances to the surface of the tube.

The analytical methods that fall under the classification of electroseparations in solution are used to separate and quantitate analytes ranging from small molecules (formula weight ~ 100) to the largest macromolecules (formula weight up to 10^8) such as DNA chains or the largest proteins.

2.1 Gel Electrophoresis

The most commonly used methods of electrophoresis are done by carrying out the separation in a polymeric gel (as in *gelatin*). The gels are composed of a three-dimensional web of polymer strands having the spaces between the strands filled with liquid. Two different separating mechanisms exist: electrophoresis, that separates by the charge-to-size ratio (that is, the total electric charge on the molecule divided by its effective cross section), and sieving, that separates mostly by size. The sieving effect is caused by the macromolecules' migration being slowed by colliding with the gel molecules; larger molecules are slowed more than smaller ones. Especially prominent among the applications of gel electrophoresis is the analysis of DNA and RNA fragments, which forms the analytical basis of contemporary molecule biology of the gene.

Conditions for electrophoresis can be divided into two classes—denaturing and nondenaturing. To *denature* means to make the molecules biologically inactive. The molecules are still polymers with the monomers covalently linked in their original order, but the polymer is no longer folded into the form that it has when it is active. Detergents such as sodium dodecyl sulfate [SDS, $CH_3(CH_2)_{10}CH_2OSO_3^-Na^+$] and other nondetergent agents (*e.g.*, urea) are known to cause such denaturing. The names of the methods, such as *SDS polyacrylamide gel electrophoresis*, relate the type of gel and that it is run under denaturing conditions.

2.2 Capillary Electrophoresis

In an electrophoretic separation, both the speed of the separation and the resolution of the components improve as the applied electric field is increased. The reason that the voltage cannot be increased at will is that the heating increases to the point where the convection ruins the quality of the separation. The heating effects can be reduced drastically by running the electrophoresis inside a capillary tube having an inside diameter less than 0.1 mm and a length about 50 cm to 1 meter. Naturally, the technique is called *capillary electrophoresis* (CE) (see CAPILLARY ELECTROPHORESIS in this update volume). Another name is capillary zone electrophoresis (CZE). One reason why running in a capillary is effective follows from its electrical properties. The electrical resistance is so high that the current remains low (less than about $10\,\mu\text{A}$). As a result, the heating is reduced.

The apparatus for capillary electrophoresis is illustrated in Fig. 2. The technique has great resolving power, and methods have been developed so that electrophoresis can even be done for neutral, nonionic molecules. (This latter method is called micellar electrokinetic chromatography, abbreviated MECC.)

The analyte zones separated by capillary electrophoresis can be detected in their path through the capillary as they pass a detector, since the solution contained inside the tube move while the separation is occurring. This solution movement is called electroosmosis or electroendoosmosis.

Using a narrow capillary does involve a tradeoff: the samples must be about a nanoliter or less, and, after separation, the analytes are each in a similar volume. With such small liquid samples, only a few methods of detection can be used effectively. For example, light absorption is not routinely used to detect concentrations below $1\,\mu\text{M}$ in such small volumes. As a result, analytes are required to be at least $10\,\mu\text{M}$ to 0.1 mM. Nevertheless, nanogram to picogram quantities can be detected and quantitated, and it is this progress in detection that

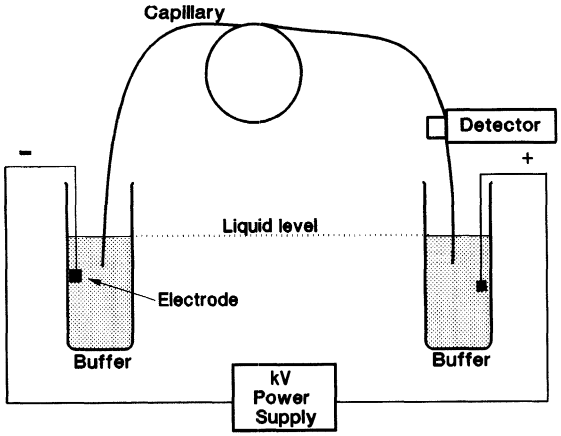

FIG. 2. Diagram of the basic parts of a capillary electrophoresis apparatus with spectroscopic detection. The sample is put in at the left side by placing the end of the capillary in the sample solution and raising its level so that the solution flows in. To run, the end is placed in a buffer set at the same height as the right end. While running, the levels of the two buffer solutions must be the same (as indicated by the dotted line) to avoid further bulk flow. The entire unit shown is held at a constant temperature.

has led to the revival of electrophoresis in this newer, capillary, form.

In running the separation, when the anode (+) is on the injection side of the capillary, this is called running in the normal mode. When the cathode (−) end is on the injection side, it is the reversed polarity mode. Voltages up to 1000 V cm^{-1} may be applied.

3. MASS SPECTROMETRY

Mass spectrometry (MS) is one of the most versatile and powerful tools of chemical analysis. Its broad usefulness is due mostly to the wide variety of capabilities possible for each of the three sections of a mass spectrometer: the mass source, the mass discriminator, and the transducer/detector. These were illustrated in the original article. The gas-phase ions are separated according to their differences in mass-to-charge ratio, abbreviated m/z.

Since its inception, mass spectrometry has been used effectively to analyze small molecules and atoms. However, a variety of sources have been devised to vaporize and ionize materials that are not initially gases, which has allowed mass spectrometry to analyze virtually all types of samples. Because of these newer vaporization/ionization methods, mass spectrometry is now firmly established in biochemical, polymer, and inorganic analysis. For molecules greater than 10 000 daltons [1 dalton (Da) = 1 unified atomic mass unit (u)], electrospray ionization (ESI), matrix-assisted laser desorption ionization (MALDI), and plasma desorption (PD) can be used. ESI has an advantage over the others in that multicharged ions (up to the range of 20 charges) are produced, which brings these heavy molecules into the m/z range of 500 to 2000. As a result, electrospray ionization permits mass spectrometry up to the 100 000-Da range on commonly available instruments. The limitation, then, is only whether such large molecules can be brought intact into the gas phase, and, apparently, many can.

Mass spectrometry also has become one of the two techniques of choice for the simultaneous analysis of many elements. The source of the ions of the elements is the inductively coupled plasma torch. This torch is a plasma formed from argon that is heated to about 3000 K with high-power radio waves. The sample in solution passes into the torch, where ions of the elements form and are passed into a mass spectrometer for analysis.

4. SAMPLE PREPARATION

Contemporary instrumentation has increased our ability to measure the contents of samples more quickly, and the bottleneck in productive analyses in many cases has become the relatively slow preparation of samples. For example, one might require a liquid solution for analysis, but the sample is soil from a field.

The conversion from a solid to the solution involves dissolution of soluble components and bringing otherwise insoluble components into solution. This latter process is called digestion.

A relatively new addition to accelerate sample dissolution and digestion is the microwave oven. A number of such ovens are specifically designed for chemical sample digestions with heating compartments lined with relatively inert materials and a means to remove corrosive vapors. Their power levels are far more precisely controlled than those of household microwave ovens, and monitors of the sample temperatures and pressures are provided. Gram-size samples can be heated to a given temperature from 10 to 100 times faster than by conventional thermal heating and with far better control of the speed of heating and the temperature. Microwave assistance also facilitates digestions in sealed containers. These containers are most often made of a type of Teflon^{TM} called (perfluoroalkoxy)ethylene (or Teflon PFA). Digestion in closed containers obviously is a superior way to ensure that analyses will be accurate and precise, since the loss of analyte is limited and introduction of interferents is easier to avoid. Both characteristics are especially important in trace and ultratrace analyses. In addition, the closed containers allow digestion to be accelerated since the boiling point can rise above the normal boiling point of the acid at atmospheric pressure. As an example, concentrated (constant boiling) nitric acid boils in an open container (at 1 atmosphere pressure) at 120.5 °C, while in a microwave pressure vessel, the temperature can be held at about 185 °C. A typical digestion reaction could be accelerated about 800-fold with this temperature difference. The limits of microwave digestion arise from the pressure and temperature limits of the containers and the electrical and chemical properties of the samples.

Further Reading

Bilhorn, R. B., Epperson, P. M., Sweedler, J. V., Denton, M. B. (1987), "Spectrochemical Measurements with Multichannel Integrating Detectors," *App. Spectrosc.*, **41**, 1127–1136.

Baker, D. R. (1995), *Capillary Electrophoresis*, New York: Wiley.

Johnstone, R. A. W., Rose, M. E. (1996), *Mass Spectrometry for Chemists and Biochemists*. Cambridge, U.K.: Cambridge Univ. Press.

Kingston, H. M., Jassie, L. B. (1988), *Introduction to Microwave Sample Preparation*, Washington, D.C.: American Chemical Society.

14

Design and Application of Cyclotrons

Richard A. Baartman, Robert E. Laxdal, George H. Mackenzie, and Yi-Nong Rao

14.1 **Introduction 453**
14.2 **Applications of Cyclotrons 453**
14.2.1 Cyclotrons as Sources of Neutrons 453
14.2.1.1 Radioisotope Production by Neutrons 454
14.2.1.2 Production of Thermal or Electrical Power 454
14.2.2 Production of Radioactive Ion Beams (RIBs) 455
14.2.3 Ion Beams in Industrial and Materials Research 456
14.2.3.1 Thin-Layer Activation 456
14.2.3.2 Constructive Processes in Materials 457
14.2.3.3 RIB Complementing Muon Spin Relaxation (μSR) Techniques 457
14.2.4 Medical Applications 457
14.2.4.1 Radioisotope Production by Ions 457
14.2.4.2 Proton Therapy 459
14.2.4.3 Neutron Therapy 460
14.3 **Advances in Cyclotron Design 460**
14.3.1 The Separated-Orbit Cyclotron 460
14.3.2 Injection from Radiofrequency Quadrupole Linacs 461
14.3.3 Altering the Beam Microstructure and Macrostructure 461
14.3.4 Limitations to Beam Intensity from Cyclotrons 463
14.3.4.1 Space-Charge Limits 463
14.3.4.2 Lifetime of Foils Used for Charge Exchange 465
References 466
Further Reading 466

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

14.1
Introduction

The cyclotron was invented in 1931, and from its inception to the 1960s installations were designed chiefly for research in nuclear physics and radiochemistry. Several laboratories developed strong affiliated programs in nuclear medicine, radiation oncology, and materials science, and these became the primary functions of some laboratories during the 1960s. Since then, more cyclotrons have been built for research purpose in nonnuclear fields and for industrial and medical applications. This scenario has been described in Chapter 13. That chapter mainly describes the history of cyclotrons and their design features in some detail, while this chapter introduces new fields of application, some speculative, and gives illustrations of recent activity in some older fields. It describes some new cyclotron technology and discusses the factors that may limit the intensity of the beam from cyclotrons.

An overview of the field can be found in the Proceedings of the 14th Conference on Cyclotrons and their Applications (Cornell, 1996). In the year 1998, there were more than 130 cyclotrons operating commercially and about 75 more in research institutions. Approximately 30 projects involving new machines or major expansions of existing facilities were under way.

14.2
Applications of Cyclotrons

14.2.1
Cyclotrons as Sources of Neutrons

It is becoming harder to convince the public to commission new reactors or to replace older ones. Accelerators are being perceived as a safer source of neutrons for the production of power by nuclear fission and for other traditional reactor applications. They function by directing an ion beam onto a target with a high atomic number. Spallation reactions, and sometimes fission reactions, produce a copious number of secondary particles including fast neutrons. These final products (neutrons) may then be slowed down in a moderator to give the velocity spectrum suited for the application. The optical source of the neutrons is less diffusible than in a reactor and much less total power is dissipated.

Examples of spallation neutron sources are ISIS at the Rutherford-Appleton Laboratory, near Oxford, United Kingdom, and LANSCE at the Los Alamos National Laboratory, New Mexico. ISIS is based on a 200-μA, 800-MeV proton synchrotron and

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

LANCSE on a linac and storage ring handling 100-μA, 800-MeV protons. The nature of operation of these types of accelerator is to produce neutrons with a pulsed time structure that can be useful for time-of-flight spectroscopy. The most powerful cyclotron proton beam at present is that at the Paul Scherrer Institute (PSI), Villigen, Switzerland, where 850 μA of the 1.5 mA of 590-MeV protons is extracted from the continuous wave PSI cyclotron is directed toward the SINQ cold neutron source to produce $10^{14}\, n\, \mathrm{s}^{-1}\, \mathrm{cm}^{-2}$; this is comparable to a medium-flux reactor. Liquid-deuterium moderators can be placed near the maximum neutron flux, and the continuous beam of very cold, 25-K or 2-meV, neutrons will match that of the current best source, the reactor at the Institut Laue-Langevin (ILL), Grenoble, France.

Physicists, chemists, and biologists use neutron beams to elucidate complicated molecular structures. Neutron therapy and neutron radiography are the other established applications that use beam from cyclotrons with $k = 40$–70. Two newly proposed applications are described below. Neutron sources based on cyclotrons have the advantage that they are smaller and more economical to operate than those using linacs or synchrotrons, although linacs may eventually produce more beam power.

14.2.1.1 Radioisotope Production by Neutrons

The most widely used medical radioisotope is $^{99\mathrm{m}}$Tc with millions of diagnoses done per year using this isotope. Because it has a 6-hour half-life, it is supplied in the form of a generator using a ^{99}Mo parent having 66-hour half-life. This isotope is obtained from reactors as the fission by-product of ^{235}U. There have been proposals to make $^{99\mathrm{m}}$Tc or ^{99}Mo by the proton bombardment of an enriched ^{100}Mo target; however, technetium produced directly would have to be used within few hours of production, and there may be difficulties in separating ^{99}Mo from the ^{100}Mo target. Ion Beam Applications (IBA), Louvain, Belgium, has proposed a facility to make ^{99}Mo by fission, as at present, but to create the neutrons by bombarding a lead target with a 1.5-mA beam of 150-MeV protons. Approximately one fast neutron would be produced per each proton. The target would be surrounded by a water moderator that would contain the secondary production targets of ^{235}U and would thermalize the neutrons. The system would operate safely below the critical level and the calculated flux of $2 \times 10^{14}\, n\, \mathrm{cm}^{-2}\, \mathrm{s}^{-1}$ should produce 2×10^{14} Bq per week of ^{99}Mo, which if efficiently extracted and distributed would match the world's demand.

14.2.1.2 Production of Thermal or Electrical Power

In the so-called energy amplifier (Rubbia, Mandrillon and Fietier, 1994), a beam from an accelerator would initiate a fission cascade in a target placed inside a calorimeter. The fission products would be slowed down in the calorimeter and their energy would be transformed into heat. Electrical energy transformed in turn from heat would exceed that required to operate the facility and the excess, as electricity or heat, could be delivered to external users. Thorium is proposed as the target; it is more abundant than uranium, fissile isotopes can be ‘‘bred’’ from it through nuclear interactions, and its use would result in a smaller amount of radioactive waste. Such ideas have been proposed before but the required technology seems closer now.

A schematic energy-generating complex is illustrated in Figure 14.1. In this proposal (Fiétier, Mandrillon and Rubbia, 1996),

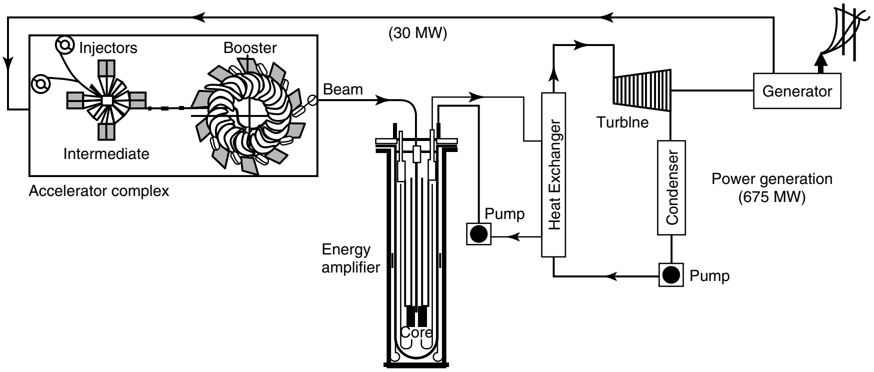

Fig. 14.1 The major components of an energy-producing complex. The cyclotron chain consumes 30 MW to produce 12 MW of 1-GeV protons. These initiate fission reactions in the core of the energy amplifier. About 1500 MW of thermal energy would be produced, which could be transformed to 675 MW of electrical energy. (Courtesy of P. Mandrillon, Cyclotron Laboratory, Nice, France).

two injector cyclotrons deliver 6 mA of 10-MeV H^- ions, and two separated-sector cyclotrons accelerate a combined 12 mA of protons first to 120 MeV and then to 1 GeV, for injection into an amplifier module. Interestingly, one 10-MeV beam would be extracted by stripping and the other using a deflector channel; the two would be overlapped in the injection line to reduce space-charge forces. Each amplifier module would produce 1500 MW_{TH}, of which 30 MW would be returned to operate the accelerators. The neutron flux in the target would be in the range of $10^{14}\ cm^{-2}\ s^{-1}$, well below criticality so that energy production will stop when the incident beam of protons is shut off, thus offering an extra measure of safety. A plant might consist of several such generators in parallel.

14.2.2
Production of Radioactive Ion Beams (RIBs)

Nuclear physicists have used separated isotopes in their targets and ion sources for several years; the isotopes could be unstable provided half-lives were reasonably long. The current interest is in the production on line of isotopes with short half-lives (minutes or less) and energies varying from $\sim$1 keV amu^{-1} to many MeV amu^{-1}. Nuclear physicists are interested in the structure of nuclei far from the line of stability in the mass (A)–nuclear charge (Z) diagram. Astrophysicists as well as nuclear physicists are interested in the reaction mechanisms of these nuclei, because they can be part of the chain of reactions that breed heavy elements in stars. Some experiments with the less easily produced isotopes can be performed by confining them in neutral-atom traps. Ions are neutralized by first implanting the beam in a surface and then re-evaporating them as atoms. These are cooled and confined by interchanging their momentum with crossed laser beams. The final ensemble may have the atomic shell structure elucidated by probing with lasers of different frequencies, and nuclear properties can be inferred from their effects on the atomic shells. Six thousand $^{38\mathrm{m}}$K atoms have been constrained to a volume of 2 mm^3 in such a trap; the half-life of the metastable isotope is about 1 second.

There are two principal methods to generate beams of short-lived isotopes. In the first method, an energetic beam is directed onto a thin target; the unstable ions are made by direct nuclear reactions and emerge with significant energy according to the laws of kinematics. They are captured in a large-acceptance beam-transport system and led to an experimental area. This method was used at the Bevalac high-energy synchrotron (now shut down) at Lawrence Berkeley Laboratory (LBL), California, and is presently employed using cyclotrons at Michigan State University (MSU), East Lansing, Michigan, and GANIL, Caen, France. Most new radioactive ion beam (RIB) facilities, including an upgrade at GANIL, will use a second method in which a high-power primary beam from an accelerator hits a thick target to produce both stable and unstable isotopes of many different species by spallation reactions. These continually diffuse out of the target, which is usually hot, into an ion source. Ions emerge at the source potential, and a mass separator selects the species of interest, which is then transported to an experimental area or to a second accelerator. Such an on-line isotope separator, ISOLDE, at CERN, Switzerland, has been producing unstable-ion beams of ~1 keV amu^{-1} for 20 years using first a proton beam from a synchrocyclotron and more recently a 1-μA, 1.2-GeV proton beam from a synchrotron. Isochronous cyclotrons offer a 10- to 100-fold increase in primary beam power and are the foundation of facilities under construction at TRIUMF at Vancouver, Canada, GANIL, INFN-LNS at Catania, Italy, and RIKEN at Saitama, Japan. One quite different proposal, PIAFE, would place the target and ion source, biased at 30 kV, within the pile of the reactor at ILL. Ions of the neutron-rich isotopes produced would be transported 400 m to the cyclotrons at the nearby Institut des Sciences Nucléaires for acceleration.

14.2.3
Ion Beams in Industrial and Materials Research

There are now two institutes dedicated to research on condensed matter that use beams from cyclotrons. These are the Hahn–Meitner Institute (HMI), Berlin, Germany, and the Institute for Advanced Materials (IAM), Ispra, Italy.

14.2.3.1 Thin-Layer Activation

The process of thin layer activation and its applications were discussed in Chapter 13. A recent example of its utility is the investigation of the spallation of the oxide scale that protects alloys from corrosion (Stroosnijder, 1996). This scale often has a different coefficient of thermal expansion from the underlying layer, and thermal cycles can cause cracks to appear and material to spall off. The conventional procedure for measuring resistance to corrosion involves subjecting a sample to several cycles of the industrial process and then measuring the change in mass. The difference in weight of the sample gives the amount of material loss, which may be only a small fraction of the sample mass. The amount of material spalled away can be measured much more accurately by first producing radioactivity in the surface layer by bombarding a specimen with a proton or deuteron beam, then subjecting the sample to cycling, and finally measuring the activity in the spalled material only.

IAM achieves a precision of nanograms per square centimeter. One may choose to irradiate only a part of a sample; for example, one need not irradiate regions of high curvature in a scaled-down model because these will not exist in the full-scale item.

The radioisotope to be produced by the beam for thin-layer studies must have an activity that can be easily measured, usually γ rays. It must have a half-life matched to the duration of the tests and may have to interact chemically with the surface, for example, in the formation of an oxide scale. In some cases, no suitable isotope can be made, or the material may be damaged by the beams that produce them. A solution may be to implant radioactive ions directly. However, not every element has an isotope with a suitable half-life or γ activity, and the range of the heavier ions may be too short. ^{7}Be has been used at MSU and at the Kurchatov Institute, Russia, to irradiate ceramic, polyamid, or viton machine parts and to measure their rate of wear during the operation of the machines. There is a proposal at TRIUMF to measure the wear of hip joint replacements by implanting ^{7}Be, or possibly ^{22}Na.

14.2.3.2 **Constructive Processes in Materials**

Emerging evidences suggest that ion beams can cause constructive processes, that is, those that create order. Ion beams passing through matter cause electronic excitation, ionization, and displacement of atoms. These processes usually destroy molecules, but it is well known that electron beams can cause cross linking of molecules in plastics. Constructive processes involving ions appear as the energy density transferred to excitation and ionization increases (Fink *et al.*, 1994), and at sufficiently high energy densities, the constructive yield can surpass the destructive yield. The passage of a heavy ion produces a dense cloud of positive charges within a few nanometers of the track, and this cloud in turn is surrounded by a penumbra of knocked-on electrons that excite nearby atoms. The ion thus produces a transient pulse with a high energy in a small volume that quickly quenches once the ion has passed. Such a sequence is suitable for the synthesis of buckminsterfullerene. C_{60} has been found within the tracks of 5-MeV $^{6}Li^{+}$ in polyimide at HMI and of 130 MeV amu^{-1} $^{161}Dy^{22+}$ on pyrolytic graphite at the UNILAC Laboratory, Darmstadt, Germany. Other examples are given in Apel (1996).

14.2.3.3 **RIB Complementing Muon Spin Relaxation (μSR) Techniques**

Research laboratories have used muon beams to investigate solid state structure, magnetism, chemical reaction rates, and so on. Unstable heavy ions also can be used in these fields. Polarized beams of β emitters can be stopped in a layer of material of less than 10 nm thickness, the isotope precesses in a magnetic field, and the observation of the modulation of the decay rate by a β detector will yield information analogous to muon spin relaxation (μSR). The precision may be better than μSR for some of these cases, because the sample is thinner and because more precession cycles may be observed since β decay lifetimes are much longer than muon lifetimes.

14.2.4
Medical Applications

14.2.4.1 **Radioisotope Production by Ions**

The most popular accelerator for isotope production remains the variable-energy H-cyclotron with 0.1–1-mA beam and a

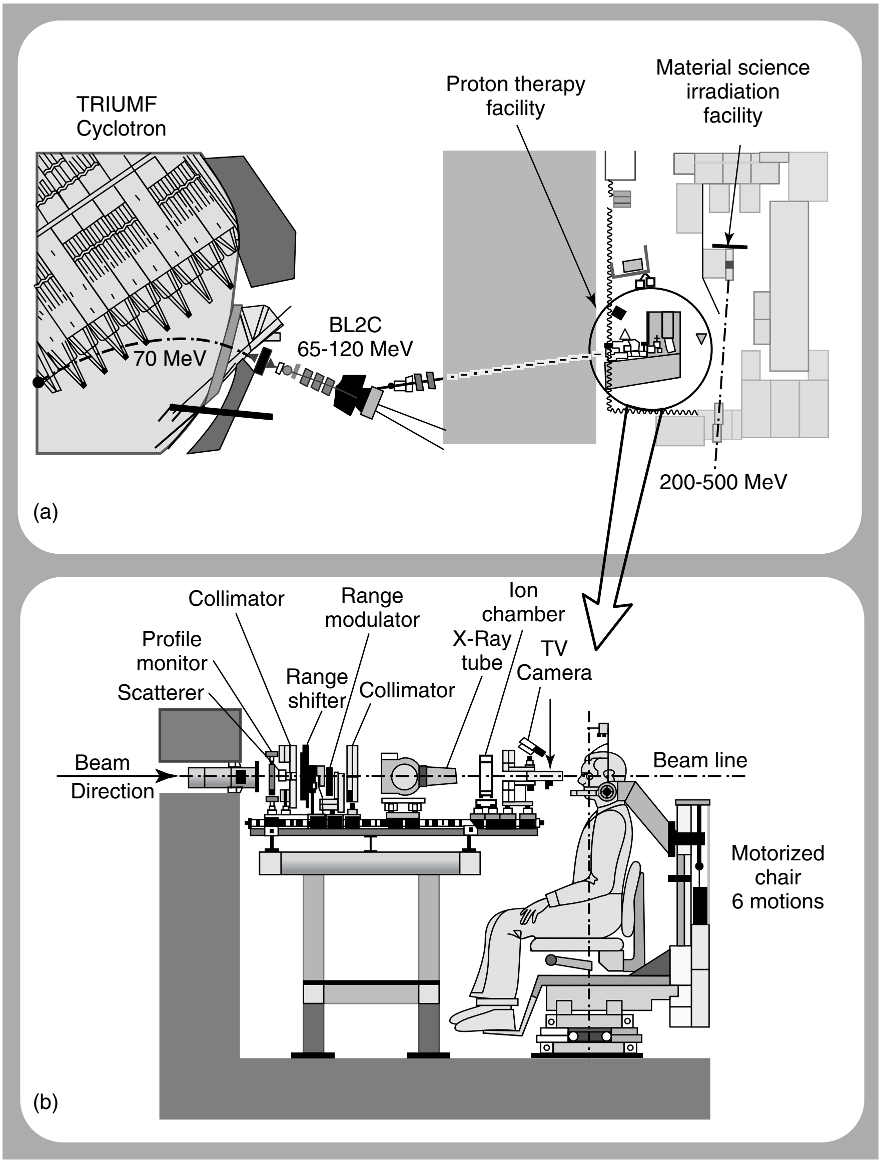
TRIUMF
Cyclotron
Proton therapy
facility
Material science
irradiation
facility
BL2C
65-120 MeV
70 MeV
200-500 MeV
(a)
Collimator
Range
modulator
Ion
chamber
Profile
monitor
X-Ray
tube
TV
Camera
Range
shifter
Collimator
Scatterer
Beam
Direction
Beam line
Motorized
chair
6 motions
(b)

◀ **Fig. 14.2** (a) A portion of the H^- beam circulating in the TRIUMF cyclotron is stripped at 70 MeV, at approximately half of the maximum orbit radius. The protons curl out of the cyclotron into a beam line, which focuses and directs them into a shielded patient treatment area. (b) The equipment used to treat tumors of the eye at TRIUMF. The scatterer introduces a broad Gaussian distribution in the angular divergence of the 70-MeV proton beam extracted from the cyclotron. The collimators trim the sloping sides of this distribution and reduce the flux from the ~1 nA extracted to a treatment level of about 10^7 protons $\text{cm}^{-2}\,\text{sec}^{-1}$. The shape of the final collimator is matched to the tumor being treated. The material in the range shifter reduces the proton energy to match the distance to the rear of the tumor; the variable-thickness range modulator spreads the dose over the tumor depth. The ion chamber measures the dose of ionizing radiation delivered. The X-ray tube is inserted to confirm the tumor position with respect to the proton beam line. (Courtesy of E. Blackmore, TRIUMF, Canada).

maximum energy in the range 18–30 MeV; over 100 of these have been installed in the last decade. They are economic and reliable in operation and have a high beam power for relatively low capital cost. The TRIUMF laboratory and Nordion Inc., Canada, have jointly demonstrated extracted beams of 30 kW of 30-MeV protons. Also their small size requires a smaller volume of shielding than some competing accelerators. Nine very small 12-MeV H-superconducting OSCAR cyclotrons have been sold by Oxford Instruments Corp., United Kingdom.

Other continuing developments include the production of generators to supply short-lived isotopes to institutions without production facilities. These generators are isotopes that decay into the short-lived species required for imaging and have a lifetime long compared with transportation times, for example, $Sr^{82}(25\ d) \rightarrow Rb^{82}$ (6 hours and 1 minute). There are also proposals to use focused beams of RIBs to accumulate high activities in a small volume.

14.2.4.2 **Proton Therapy**

The vast majority of medical treatments with particle beams are intended to destroy cancerous tissues while attempting to spare neighboring healthy tissues. Ion beams are well suited to this because they have a well-defined range (see Figure 13.15) and can be focused to spots in tissue within a diameter of a few millimeters. Proton beams have a lower magnetic rigidity than heavy-ion beams with the same range and thus the accelerator and beam-transport system will be smaller and cheaper. There are 19 proton irradiation facilities at present, and the number continues to increase.

The equipment for a 70-MeV beam to treat ocular melanoma is shown in Figure 14.2. Deeper tumors require higher particle energies. The 230-MeV protons would have sufficient range for most people but an accelerator may have to provide 250 MeV or even 270 MeV to compensate for energy lost in the beam-preparation system. Tumors with large area can be treated either by incorporating a second scatterer into the standard passive system illustrated in Figure 14.2(b) or by steering, usually termed *scanning*, a small-diameter beam over the tumor area. The second scatterer is designed to broaden the center of the Gaussian distribution produced by the first and to remove the penumbra. The scatterers absorb some of the beam energy and may produce a small background of secondary particles. Scanning techniques avoid these effects and are considered to match tumors of irregular shape better, but they are more complicated. Each scan treats a certain depth of tumor, and the energy of the beam entering the patient is changed to treat ''slices'' at different depths. The beam current must not vary by more than a few percent for scanning and must accommodate refinements such as synchronization with breathing. The beam from a cyclotron is well suited as it shows little modulation on a millisecond and longer timescale.

It can take an hour to change the energy of a beam from a variable-energy proton cyclotron. Consequently, the treatment facilities often operate with a fixed extraction energy and use energy-absorbing material followed by a magnetic momentum-selection system to obtain adequate energy for treatment. The radiation background and residual activity in the degrader and momentum collimator are acceptable. Both are well shielded, the beam currents used are $\leq 300\,\mathrm{nA}$, and the beam is usually switched on for less than 10% of the working day because much time is consumed for setting up the patients for treatment. Cyclotrons for therapy occupy a lesser space compared with most treatment rooms and are less expensive and easier than synchrotrons to build and operate. IBA have designed a compact 235-MeV proton cyclotron, beam-preparation system, and gantry specifically for therapy. The magnet of this cyclotron is not superconducting but it does operate with highly saturated iron; the hill field is 2.3 T and the outer diameter of the yoke is about 4.5 m. This results in a cyclotron of relatively small size without the necessity to provide a supply of liquid helium in a hospital. Two have been sold, the first to Massachusetts General Hospital, Boston, United States and the second to the National Cancer Center, Kishiwa, Japan.

14.2.4.3 Neutron Therapy

Salivary-gland and advanced prostate tumors are examples of the sites that respond best to direct irradiation by neutron beams, and there are perhaps 10 cyclotrons attached to hospitals performing such treatments. Boron-capture therapy is a new modality that incorporates ^{10}B into molecules that enter tumors easily; this is followed by irradiating the patient with neutrons. The reaction $^{10}\mathrm{B} + \mathrm{n} \rightarrow {}^{7}\mathrm{Li} + \alpha$ has a cross section of hundreds of barns for thermal neutrons, and the reaction products are highly ionizing and destructive over a short range. At present, trials are being carried out at reactors and hospitals such as The Department of Radiation Oncology, University of Washington, Seattle, that have cyclotrons that can produce 10^8 slow neutrons $\mathrm{s}^{-1}\,\mathrm{cm}^{-2}$ thermalized in the patient. Ideally about $10^9\,\mathrm{s}^{-1}\,\mathrm{cm}^{-2}$ thermal neutrons are required at the tumor, and if the method proves successful, an optimized cyclotron and neutron preparation system will be required for hospital use.

14.3
Advances in Cyclotron Design

14.3.1
The Separated-Orbit Cyclotron

A most innovative, but technically challenging, machine is the superconducting separated-orbit cyclotron (SOC) Tritron at Munich, Germany. It combines some of the features of a cyclotron and synchrotron. The typical isochronous fixed-frequency cyclotron with $\beta \approx r\omega/c$ accelerates a continuous stream of beam bunches but with a turn separation that unfortunately decreases as $(\beta\gamma^3)^{-1}$. Tritron is also isochronous, with a fixed field and accelerating frequency, but in addition, the turn separation is constant. For turn separation to be constant, the accelerating voltage should increase proportionately with radius. Ions in Tritron follow a fixed spiral path in the field of superconducting magnets, each turn occupying its own channel in the iron (Figure 14.3). Efficient superconducting radiofrequency (rf) accelerating cavities produce a radial turn separation of 40 mm, much larger than the beam width. Particles leading or lagging in phase with respect to the reference particle will deviate in energy and eventually in radius. The field gradient dB/dR in each separate channel can be chosen to provide a longitudinal (phase) restoring behavior as well as transverse focusing provided that the beam bunches lie at negative phases where $dT/d\phi > 0$ (Trinks, 1991). This characteristic resembles that of a synchrotron; it takes about four turns in Tritron to complete one synchrotron, or phase, oscillation. The existence of longitudinal focusing allows operation at 170 MHz with harmonic numbers $h = 14–55$, larger than the 1–10 typical of more conventional cyclotrons. All components are now in place and a beam of $^{32}S^{14}$ ions has been accelerated over several turns.

14.3.2
Injection from Radiofrequency Quadrupole Linacs

Linear accelerators usually operate with a fixed rf and a fixed particle-velocity profile, which results in a fixed output energy per nucleon (see Chapter 16). This is a limitation to their use as injectors to variable-frequency cyclotrons. Some years ago, a variable-frequency drift-tube linac, RILAC, was built at the RIKEN Laboratory, Saitama, Japan. The drift tubes were attached to the inner end of coaxial $\lambda/4$ lines, and the frequency was altered by varying the electrical length of these lines by means of a sliding short. These Wideröe-type linacs still require an injected beam of several hundredkilo electronvolts. Radiofrequency quadrupole (RFQ) linacs accept beams at a few keV amu^{-1}, bunch it, and accelerate it with high efficiency, but again they usually operate at a fixed frequency. Moreover their emittance is not well matched to a cyclotron acceptance. This situation is changing. A variable-energy and -frequency four-rod RFQ injector linac is being built by a group at Frankfurt University to match the dee frequencies, 80–110 MHz, of the $K = 134$ MeV separated-sector cyclotron at HMI. It will replace the present 8-MV tandem injector. Another four-rod RFQ has successfully operated at 18–40 MHz at RIKEN; it will be used to inject from a high-power ion source into RILAC. The efficiency of an RFQ in turning electrical power into particle energy is greater at high frequencies. RFQs could operate at harmonics of the cyclotron dee frequency, but the ion beam itself must be bunched to the lower, cyclotron, frequency to capture most of the ions leaving the ion source.

14.3.3
Altering the Beam Microstructure and Macrostructure

The typical cyclotron operates in continuous wave (CW) mode, producing a train of beam bunches, each a few nanoseconds long at the frequency of the accelerating field. These frequencies range from ~1 MHz for heavier ions to ~100 MHz for protons. There has been no macrostructure, that is, modulation at

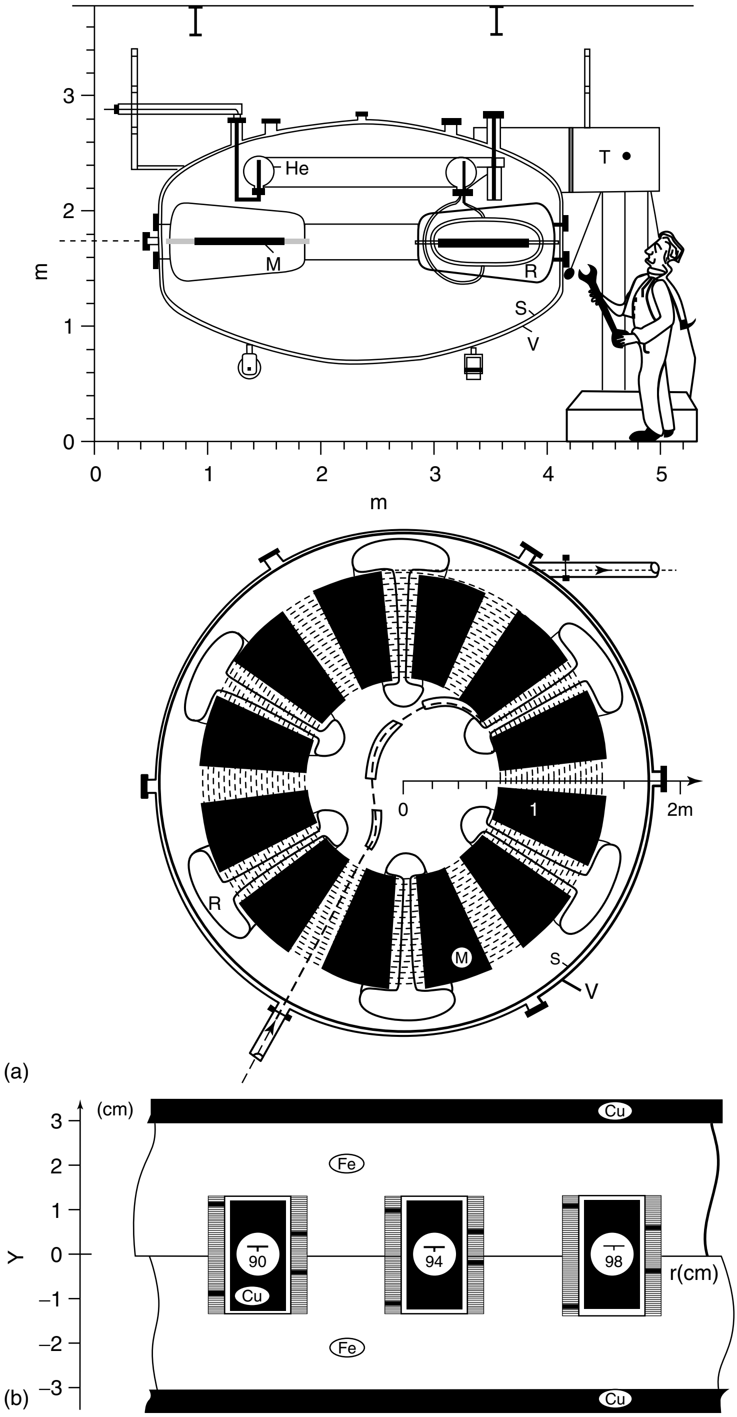
3
2
1
0
m
He
M
R
S
V
T
0
1
2
3
4
5
m
0
1
2m
R
M
S
V
(a)
(cm)
3
2
1
0
−1
−2
−3
Y
Cu
Fe
90
94
98
Cu
r(cm)
Fe
Cu
(b)

◀ **Fig. 14.3** (a) Cross sections of the Tritron cryostat. The rf cavity, R, and magnet sectors, M, are cooled by liquid helium from the reservoir He. S is a shield at 80 K and V is the vacuum vessel. (b) A cross section of three beam channels in a magnet sector. The superconducting coils exciting the field have two separate windings in each channel to produce the field gradient that provides focusing. (Courtesy of U. Trinks).

lower frequencies, unless it was imposed by chopping the beam with a pulsed deflector or by pulsing either the ion source or the rf voltage, and these methods reduce the total charge delivered. Recently, cavities operating at some harmonic h of the main rf have demonstrated modulation of the time structure without loss of beam.

Consider that such a cavity to be installed approaches the outer radius of a cyclotron and provides an additional energy gain per turn ΔT_h that can be phased to either augment or oppose the energy gain ΔT_1 from the main rf. The phase Φ_r of some particle at radius r is obtained from

$$\Delta T_1(r_0)\sin\Phi_0 = \Delta T_1(r_0)\sin\Phi r + [k\Delta T_h(r)/h]\sin(h\Phi_r) \tag{1}$$

where Φ_0 is the starting phase at radius r_0. The parameter $k = 1$ when the additional cavity is phased to assist the main cavity in accelerating the beam. In this case, the width of the phase band decreases during acceleration from r_0 to r and phase compression is said to occur. The phase width is expanded when the additional cavity opposes acceleration – k is set equal to -1 in Eq. (1). In the latter case, the phase band can be split into two when $\Delta T_h > \Delta T_1$. The results of an experiment to confirm this are shown in Figure 14.4(a). The beam extracted from TRIUMF was a single bunch 3 ns wide at half maximum. This split into two bunches 8.5 ns apart when an additional cavity was turned on and its phase was adjusted to oppose acceleration. Figure 14.4(b) shows how further splitting is possible with the addition of extra cavities. This effect may possibly be exploited to match a cyclotron with a postaccelerator of higher frequency.

Large betatron oscillations can be excited by a kick from a pulsed deflector or by a field oscillating at the betatron frequency. If an extraction system – stripping foil or dc deflector – is placed at the point of maximum excursion, the beam extracted will have the time structure of the driving term. For example, calculations have shown that a stripping foil placed at a radius where Q_z is 0.25 and positioned above a deflector pulsed to impart a vertical kick every fourth turn to drive this resonance can provide a beam with three unpopulated rf buckets followed by a bucket four times as bright as normal.

14.3.4
Limitations to Beam Intensity from Cyclotrons

The most important advantage of cyclotrons for many ion-beam applications is the high beam current that they can deliver. It is useful to consider the factors that may limit beam intensity.

14.3.4.1 Space-Charge Limits

Coulomb forces between the constituent ions of the beam being accelerated oppose the focusing forces provided by the magnetic and electric fields. The most common limitation to beam current arises at low energy when this transverse space-charge force increases the height of the beam to the limit of the acceptance. This may occur in an axial injection system, in an inflector, or over the first few turns in

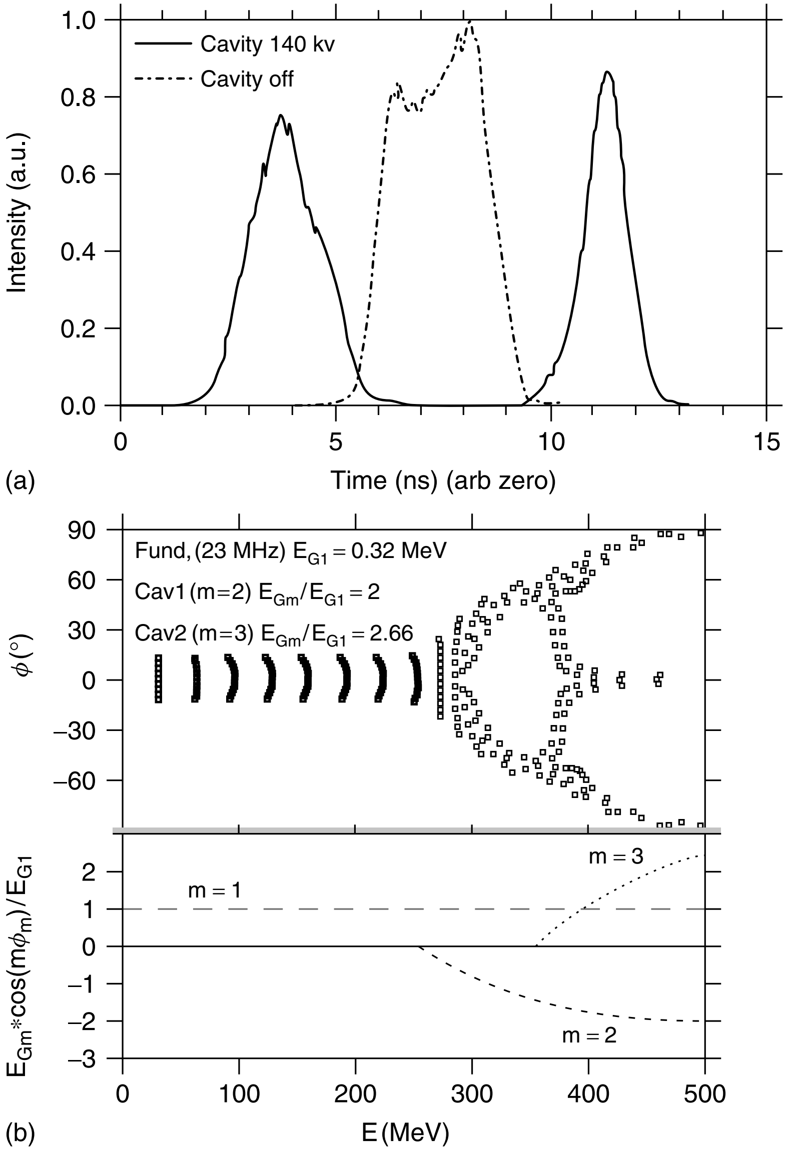

Fig. 14.4 (a) An experimental confirmation of beam splitting using one 300 keV/turn, $h = 4$ cavity at TRIUMF. One nanosecond is equivalent to 8.3° of phase for the main 23 MHz rf. (b) A computer simulation illustrating the splitting of a single phase band into three by means of two cavities operating at the second and third harmonics of the main accelerating rf.

the cyclotron where the limiting proton current, in amperes, is given by

$$I_{\max} = \beta \mathrm{B_f} \left(\frac{Q_z b \omega_{\mathrm{ion}}}{c} \right)^2 [7.8 \times 10^6] \qquad (2)$$

In Eq. (2), B_f is the bunching factor, defined as the ratio of average to local current, and b is the half-height of the accepting aperture. The other terms are defined in Chapter 13. A 1-MeV test-bed cyclotron with a solid pole, an intense external cusp H^- source, a spiral inflector, and an injection energy of only 25 keV has produced 2.5 mA of H^- ions at TRIUMF.

The ring cyclotron at PSI, which holds the record for beam power, has delivered 1.5 mA, or 0.9 MW, of 590-MeV protons. Beam is injected into the ring at 72 MeV, and low-energy space-charge problems are handled by the specially designed injector cyclotron. Within the ring itself, 10^{-4} of a 1 mA current is lost, and this loss scales as $1/N^3$, where N is the number of turns from injection to extraction. This is consistent with the limiting factor being longitudinal space charge, which tends to accelerate the front of the beam bunch and decelerate the rear. The bunch twists in the horizontal plane, increasing the overlap of turns and reducing the extraction efficiency. An extra rf cavity, oscillating at a harmonic of the main accelerating cavities, may ameliorate this situation by providing accelerating forces opposing those from space charge.

Cyclotrons normally operate in a steady state, and dynamic feedback between a beam signal and a subsystem is unnecessary. The rf cavities in the PSI ring cyclotron are illustrated in Figure 13.5. Four 50-MHz cavities accelerate the beam, and one 150-MHz cavity opposes their action to flatten the 50-MHz waveform over the length of a bunch. This flattening results in all protons having about the same energy in any given turn and thus increases the separation between turns and extraction efficiency. The 0.9 MW absorbed by the 1.5-mA beam is comparable to the power dissipated in the walls of the 50-MHz cavities. The 150-MHz cavity decelerates the beam and thus absorbs power from it, and at 1.5 mA, this absorbed power is also comparable with the cavity losses but has the opposite phase. The net phasor (voltage and phase) of the power supplied to this cavity depends strongly, therefore, on the beam current being delivered at that instant. PSI had to develop a fast feed-forward control system for the 150-MHz power supply in order to operate stably at these higher currents. The system uses the signal from a beam-current sensor to adjust the drive signal to the amplifier. In addition, an external load was added to the cavity to absorb both forward and reflected power at the cost of increasing the total power required (Sigg *et al.*, 1996).

14.3.4.2 Lifetime of Foils Used for Charge Exchange

Foils are used to extract a beam from a cyclotron by stripping (Cyclotrons, Section 14.3.5), or to increase the charge state of an ion to make more effective use of an accelerating voltage. The useful life of a foil is affected by the energy dissipated within it. Power dissipated by the passage of beam ions and stripped electrons can raise the temperature to a point where evaporation is important. Also the structure of the foil material may be altered at high temperatures, resulting in dimensional changes and curling or tearing of the foil; for example, pyrolytic graphite may be transformed into polycrystalline carbon. The carbon foils and monofilaments used to extract proton beams by stripping circulating H^- ions take 13 mAh mm^{-2}, or

$(2\text{ to }7) \times 10^{19}$ protons mm^{-2}, before curling. Lattice changes can also occur at normal temperatures; in fact this is the basis of some applications in the modification of materials. Heavy ions are most effective in displacing foil atoms because of their mass and short range. It was found at GANIL that carbon foils 150–400 μg cm^{-2} thick take only 3.6×10^{15} mm^{-2} of 13.6 MeV amu^{-1} argon ions before foil thickening renders the emittance of the outgoing beam unacceptable. For very thin foils Livingston, Berry and Thomas (1978) obtained

$$_{T}(p\mu\text{A- min -min}^{-2}) = \text{A} \times \text{T}^{1.13}(\text{MeV amu}^{-1}) \quad (3)$$

where the beam intensity is expressed in particle microamps (pμA) and the empirical constant A is 20 for argon. The rupture lifetime–intensity product τ may be extended >10-fold by more sophisticated manufacturing methods or by mounting the foils in an oscillating frame. Pure metal foils should be more resistant to lattice deformations than carbon; however, they must be thin to reduce scattering and to be able to handle the thermal power from the beam. The ISIS synchrotron uses 200 μg cm^{-2} Al_2O_3 foils successfully.

References

Apel, P.Yu. (**1996**) *Proceedings of the 14th International Conference on Cyclotrons and their Applications*, (ed. J.C. Cornell), World Scientific Publishing, Singapore, pp. 136–143.

Cornell, J.C. (ed.) (**1996**) *Proceedings of the 14th International Conference on Cyclotrons and their Applications*, World Scientific Publishing, Singapore.

Fiétier, N., Mandrillon, P., and Rubbia, C. (**1996**) in *Proceedings of the 14th International Conference on Cyclotrons and their Applications*, (ed. J.C. Cornell), World Scientific Publishing, Singapore, pp. 598–605.

Fink, D., Chadderton, L.D., Hosoi, F., Omichi, H., Sasuga, T., Schmoldt, A., Wang, L., Klett, R., and Hillenbrand, J. (**1994**) *Nucl. Instrum. Methods*, **B91**, 146–150.

Livingston, A.E., Berry, H.G., and Thomas, G.E. (**1978**) *Nucl. Instrum. Methods*, **148**, 125–127.

Rubbia, C., Mandrillon, P., and Fietier, N. (**1994**) in *Proceedings of the 4th European Particle Accelerator Conference*, (eds V. Suller and Ch. Petit-Jean-Genaz), Vol. **1**, World Scientific Publishing, Singapore, pp. 270–272.

Sigg, P., Cherix, J., Fetze, H.R., Frei, H., Märki, M., and Schryber, U. (**1996**) in *Proceedings of the 14th International Conference on Cyclotrons and their Applications*, (ed. J.C. Cornell), World Scientific Publishing, Singapore, pp. 161–168.

Stroosnijder, M.F. (**1996**) in *Proceedings of the 14th International Conference on Cyclotrons and their Applications*, (ed. J.C. Cornell), World Scientific Publishing, Singapore, pp. 144–147.

Trinks, U. (**1991**) *Nucl. Instrum. Methods*, **A306**, 27–35.

Further Reading

Many of the larger laboratories maintain Web pages which describe the laboratory facilities and give links to the experimental programs. Some examples are the following:

European Centre for Nuclear Research, (**2009**) Geneva, Switzerland. *http://www.cern.ch*.

Grand Accélérateur National d'Ions Lourds, (**2009**) Caen, France. *http://ganinfo.in2p3.fr*.

Indiana University Cyclotron Facility, (**2009**) Bloomington, IN. *http://www.iucf.indiana.edu*.

Laboratorio Nazionale del Sud, (**2009**) Catania, Italy. *http://www.lns.infn.it*.

National Superconducting Cyclotron Laboratory, (**2009**) Lansing, MI. *http://www.nscl.msu.edu*.

Paul Scherrer Institute, (**2009**) Villigen, Switzerland. *http://www.psi.ch*.

Institute of Physical and Chemical Research, (**2009**) Saitama, Japan. *http://www.riken.go.jp*.

Tri-University Meson Facility (TRIUMF), (**2009**) Vancouver, Canada. *http://www.triumf.ca*.

DISPLAY TECHNOLOGY

(An Addendum to *Encyclopedia of Applied Physics*, Volume 5, pages 101–126.)

DIETMAR THEIS, *Corporate Technology, Siemens AG, Munich, Germany*

Introduction 403

1. **Liquid-Crystal Displays** 403

1.1 Tailored Nematic Liquid-Crystal Materials 403

1.2 Novel Addressing Schemes for Direct-Addressed Supertwist LCDs 404

1.3 Reflective LCDs for Low-Power Systems 405

1.4 Chiral Nematic and Chiral Smectic Liquid Crystals 405

1.5 Active-Matrix–Addressed Liquid-Crystal Displays (AMLCDs) 407

1.6 Plasma-Addressed Liquid-Crystal Displays 409

2. **Active Displays** 409

2.1 Field-Emission Displays 410

2.2 Plasma Displays (PDPs) 411

2.3 Electroluminescent Displays (ELDs) 412

2.3.1 Electroluminescence in Inorganic Materials 412

2.3.2 Electroluminescence in Organic Materials 413

2.4 Visible-Light–Emitting Diodes (LEDs) 413

3. **Projection Displays** 414

Works Cited 415

INTRODUCTION

Since the beginning of the last decade of this century the technical and economic importance of electronic displays has dramatically increased. It is predicted that the ever-growing application of displays will continue over many years to come. We are now only at the threshold of the "information age," which inevitably will stimulate new applications and new technical developments in the field of visual man-machine interface. In recent years we have already witnessed the appearance of flat-panel displays as daily commodity items. The market penetration of display-based products and systems occurs in an environment where digitized hardware and networks, advances in display technologies, and user-friendly software are rapidly developing. The information age provides new modalities like multimedia, high-definition television, and virtual reality on an internetworked global scale. Advanced storage and data-compression technologies allow for the provision of huge image data bases at one's fingertip. All these developments pave the way for an increased demand for user-friendly light and flat displays.

The intense market penetration of flat-panel displays is nearly completely to be credited to the products of liquid-crystal display (LCD) technology. In 1995/96 the production volume of high-information LCDs was two orders of magnitude larger than the sum of all other high-information content flat-panel technologies.

This Addendum will cover some of the new developments in the mainstream LCD technology; in the active light-emitting display technologies like field emission, plasma, electroluminescence, visible-light–emitting diodes; and in projection display technologies.

1. LIQUID-CRYSTAL DISPLAYS

1.1 Tailored Nematic Liquid-Crystal Materials

Liquid crystals appear in several "mesophases," which exist between the isotropic liquid phase and the highly ordered crystalline structure. Technically, the most important mesophases are the "nematic" (string-like) and the "smectic" (soapy) mesophases. Nematic liquid-crystal molecules consist of a flexible component such as an alkyl group and rigid components such as a biphenyl group and a cyano group. The cyano group serves to pro-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

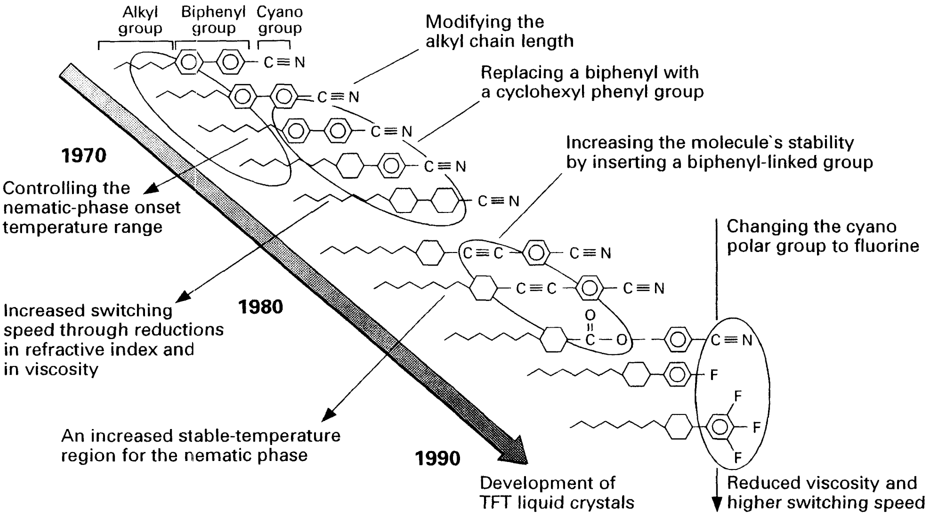
Alkyl group
Biphenyl group
Cyano group
C≡N
Modifying the alkyl chain length
Replacing a biphenyl with a cyclohexyl phenyl group
Increasing the molecule's stability by inserting a biphenyl-linked group
Changing the cyano polar group to fluorine
1970
Controlling the nematic-phase onset temperature range
Increased switching speed through reductions in refractive index and in viscosity
1980
An increased stable-temperature region for the nematic phase
1990
Development of TFT liquid crystals
Reduced viscosity and higher switching speed
C≡C
O
C−O
F

Fig. 1. Molecular design for liquid-crystal molecules.

duce positive dielectric anisotropy (the difference between the dielectric constant in the direction of the long axis of the liquid crystal molecule and the dielectric constant at a right angle to the long axis). Molecular design efforts in recent years have involved replacing or modifying the three parts of a liquid crystal in order to synthesize new liquid crystals. This is illustrated in Fig. 1, which shows the main results of chemical modifications on the device performance (Yamazaki *et al.*, 1996). Properties that cannot be produced by molecular design are achieved by mixing various liquid crystal materials, as is being done to increase the practical temperature range.

When the liquid-crystal molecules are oriented by a surface alignment layer or by an electric field, the bulk sample has all the optical properties of an anisotropic crystal. LCDs operate by electronically modulating this anisotropy between two distinct optical states. The anisotropy is made visible with the aid of crossed polarizers.

1.2 Novel Addressing Schemes for Direct-Addressed Supertwist LCDs

In LCD matrix displays the driving circuitry sequentially selects the address lines, while the data information is sent simultaneously to all pixels in the selected line. Liquid crystals have relatively slow turn-on times; they respond to the time-average voltage (rms) instead of the instantaneous voltage. A selected pixel, which spends most of its time with its driving voltage below threshold voltage, has only a slightly higher rms voltage than a nonselect pixel. All contrast needs to be produced with only this small difference in voltage. This is why a highly nonlinear response is required as is provided by supertwist (STN) LCDs.

A drawback of STN is the inherent comparatively long response time (300 to 100 ms) and the chromaticity of the cell. To get rid of the chromaticity, double supertwisted nematic cells (D-STN) have been developed. Here a second LC panel containing a helix of the opposite orientation is placed in front of the first device. This technique later gave way to film-compensated STN (F-STN) in which the undesired coloration was compensated for by a phase-difference film.

The inherent response time of the display must be many times longer than the refresh time to satisfy the rms response condition. This generally rules out displays with response times short enough to show moving images at video rates. Indeed, if the display uses a "slow" LC mixture, the difference between OFF and ON states never gets large enough to generate good contrast. If the display uses a "fast" responding LC, then the contrast of an individual row begins to fade before the next refresh cycle begins, again deteriorating contrast. This phenomenon is called frame response—instead of responding to the average voltage applied over several frames, the LCD reacts to the relatively high-voltage row pulse within a frame.

The problem can be avoided by a completely different approach (Scheffer and Clifton, 1992: "Active Addressing^{TM}"; Ihara *et al.*, 1992; "Multiple Line Selection"). Instead of sending the high-voltage pulses sequentially down the rows, a set of predetermined signals are applied to multiple rows simultaneously and a unique calculated analog voltage is applied to the columns, depending on the state of the pixels in that column. The key is to determine which rows need which signal and to calculate the proper column voltage within the time frame available. One still has to provide an LCD cell that is capable of switching at video rates, i.e., decrease the cell thickness and use fast material.

The ultimate success of these novel addressing schemes will depend on the added cost of making these enhancements and the competition with active-matrix LCDs in terms of cost and performance.

1.3 Reflective LCDs for Low-Power Systems

Reflective color LCD without back light would be the enabling technology for multifunctional portable information terminals by which one could receive and send any information any time at any place. Such displays do not need heavy batteries and represent the most promising approach to "electronic paper" today.

A wide variety of reflective LC devices have been developed so far. A comprehensive description of the basic LC modes and design concepts has been given by Uchida (1996).

1.4 Chiral Nematic and Chiral Smectic Liquid Crystals

Both nematic and smectic liquid crystals forming a helical (chiral) structure are optically active. Optically active (chiral) nematic liquid crystals are also called cholesteric liquid crystals. They have the property of selectively reflecting or passing the light of a specific wavelengh if the pitch of their helical structure is equal to the wavelengh λ. This is the condition for constructive interference of light waves occurring if identical regions in the sample are separated by a distance of $\lambda/2$. Cholesteric liquid crystals can form one of three major textures depending on boundary conditions and applied fields: planar, focal conic, or homeotropic (Fig. 2). With increasing field strength the planar structure first transforms into a focal conic structure, then into a homeotropic aligned nematic state—which is highly transparent. The planar texture is highly reflective at a preselected wavelength and band-

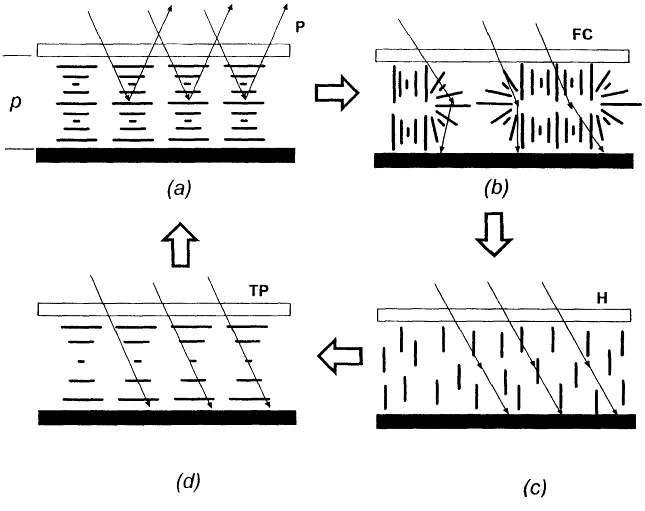

Fig. 2. Schematic textures in cholesteric liquid-crystal material: (a) light-reflecting planar texture; (b) nearly transparent focal-conic texture; (c) transparent homeotropic texture; (d) transient planar texture.

width. With the introduction of a polymer network or with suitable surface treatment of the cell substrates, a domain structure is introduced. This broadens the viewing angle, stabilizes the focal conic texture to create bistable switching, and introduces a gray-scale regime. The ability of the display to maintain pixels in both the ON and OFF states once they are written (without refresh) allows for unlimited resolution on a low-cost passive matrix without sacrificing contrast. Since a recent publication on a full-page size prototype (Pfeiffer *et al.*, 1995), renewed interest in this material is growing.

In the first engineering samples a total reflection around 40% was observed, substantially more than an ordinary reflective twisted nematic LCD. The achievable contrast is determined by transparency of the focal conic structure, which can exhibit a weak scattering. Different driving schemes are still under development. Real-time television has already been viewed on a cholesteric display cell at Kent State University (Doane and Stefanov, 1996). The technology has unique features like high resolution, low power due to bistability, and low cost, and seems to be very attractive for portable and handheld devices. The absence of polarization makes it possible to use low-cost plastic substrates—their inherent birefringence does not interfere with device performance. Principally polymer-stabilized cholesteric-texture devices could be one of the premier enabling technologies for electronic books, newspapers, or document viewers.

Optically active (chiral) smectic liquid crystals have a permanent electric dipole moment in the absence of an electric field, and for this reason they are called also ferroelectric liquid crystals (FLCs). The intrinsic polarization of these molecules, which is perpendicular to their long molecular axis (director), has the consequence that when subjected to an external field the molecules respond to the direction as well as the magnitude of the field. The helical structure would average out the direction of the spontaneous dipole, thus suppressing the ferroelectric behavior. Rubbed polymer surface alignment layers, similar to those in supertwist LCDs, remove the helical structure. The spacing between the two glass substrates should be about 2 μm. This spacing is also required for the ferroelectric liquid crystal cell to act as a half-wave plate. The operation of a FLC display depends on the switching of the FLC's dipole, whose orientation depends on the polarity of the addressing pulse (Fig. 3).

The FLC molecule cannot simply rotate

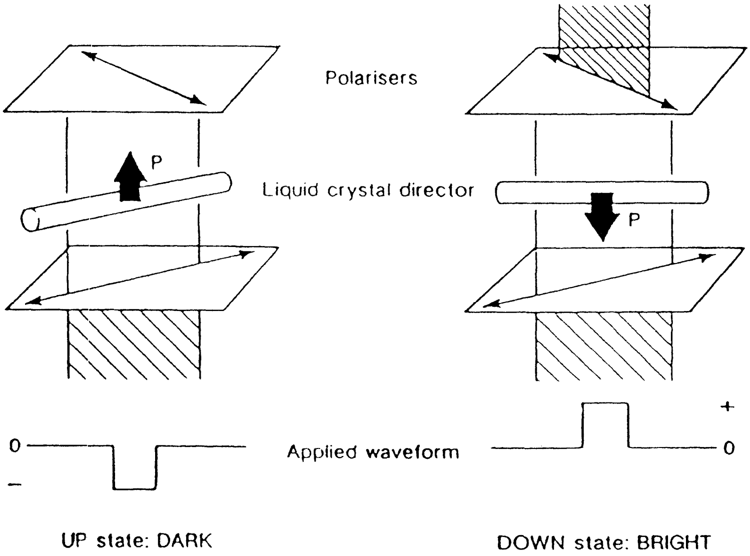

Fig. 3. An FLC display operates by changing the orientation of the LC molecule in the plane of the display.

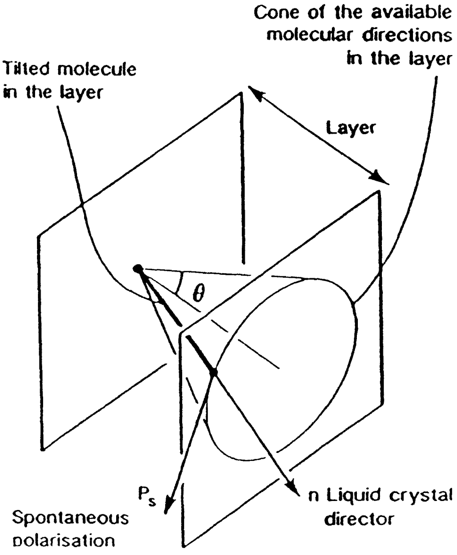
Cone of the available molecular directions in the layer
Tilted molecule in the layer
Layer
θ
P_s
Spontaneous polarisation
n Liquid crystal director

Fig. 4. The FLC molecule is constrained to rotate about an imaginary cone.

about its own axis; instead it is constrained to rotate about an imaginary cone (Fig. 4). If the cone half-angle θ, which is equal to the tilt angle of the molecules in the smectic layer, is approximately equal to 22.5°, then the change in the orientation of the FLC molecules, 2θ, will be 45°. This is the angular change that produces the optimum values of transmission and contrast ratio (Mosley, 1997).

The two main attributes of FLC displays are very fast switching—typically 100 μs per row—and bistability. This property can be made use of to produce a nonvolatile display where the image is maintained at zero volts. The molecules switch in the plane of the display rather than perpendicular to it (as in supertwist and active-matrix LCDs). This leads to a very good viewing angle. The main obstacles in the commercialization of ferroelectric LCDs have been production issues (very small distance between substrate plates), the absence of gray scale, and susceptibility to mechanical shock, although it is said that there are solutions to these problems. In addition to ferroelectric LCDs there are also antiferroelectric LCDs. Yamada *et al.* (1995) have developed a display with gray scale and have demonstrated high-quality video images on a $125 \times 80\text{-mm}^2$ display area.

1.5 Active-Matrix-Addressed Liquid-Crystal Displays (AMLCDs)

Active-matrix addressing has been the most broadly pursued technology to enhance liquid-crystal display performance. In 1996 the world market for AMLCDs was estimated to be about US\$$6.1 \times 10^9$; a global turn-over in the order of US\$$10 \times 10^9$ is expected around the year 2000.

Active-matrix addressing leaves the fundamental twisted nematic operation unchanged and uses pixel-external nonlinear devices to enhance the contrast of large multiplexed displays. The devices do not respond to the rms time-average voltages but rather to the instantaneous peak voltage. Two- or three-terminal thin-film devices act as switches, electrically isolating a pixel until its input voltage exceeds a certain specified threshold. Pixels in active-matrix devices are driven at duty cycles close to 100%. The voltage at the pixel remains practically constant during each frame, as if the pixel were directly driven. The pixel acts as a capacitor storing the charge and maintaining the voltage during the frame time. Thin-film transistors (TFTs) formed by sophisticated deposition and patterning technologies are the most frequently used switches in AMLCDs commercially available. The semiconductor material used for the TFT in most cases is amorphous silicon because it can be fabricated at relatively low temperatures and generates high image quality. The use of polysilicon material would allow the integration of driver circuits on the same glass plate because of the much higher mobility achievable. On account of high process temperatures polysilicon is not compatible with normal glass substrates. Laser or furnace annealing are key processes for the advancement of polysilicon AMLCD technology. For details of the fabrication the reader is referred to Morozumi (1996).

AMLCD manufacturers now are moving into the third level of development with significant improvement issues. The coming wave of AMLCDs is intended to at least partially replace CRT monitors. This means that competitive size, cost, and performance of the flat displays will be of paramount importance. In 1995 the AMLCD mass product was the VGA 10.4-in. (26.4-cm) panel. Now the trend has shifted to 11.3-in. (28.7-cm) and 12.1-in. (30.7-

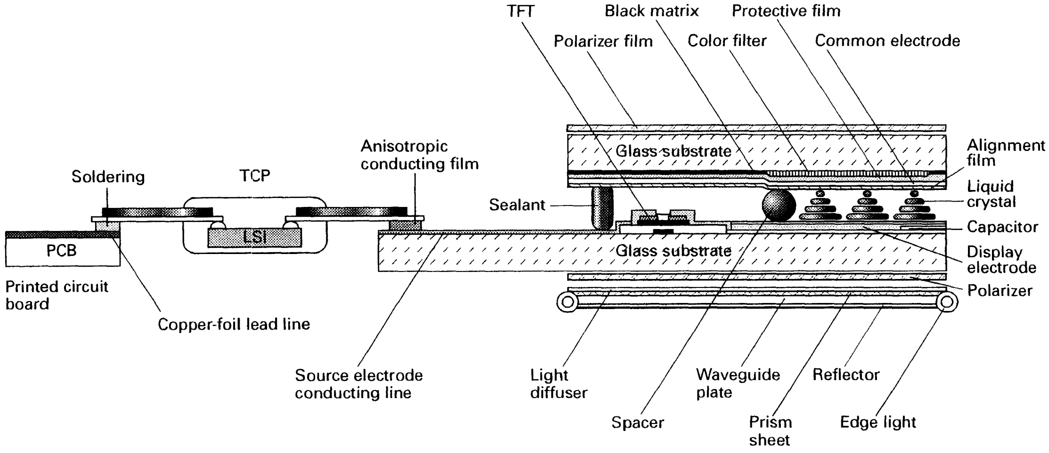
TFT
Black matrix
Protective film
Polarizer film
Color filter
Common electrode
Glass substrate
Alignment film
Liquid crystal
Capacitor
Display electrode
Polarizer
Anisotropic conducting film
Soldering
TCP
Sealant
LSI
PCB
Glass substrate
Printed circuit board
Copper-foil lead line
Source electrode conducting line
Light diffuser
Waveguide plate
Reflector
Spacer
Prism sheet
Edge light

Fig. 5. Cross section of a color AMLCD and connection to the addressing circuitry (after Yamazaki *et al.*, 1996).

cm) SVGA panels. The size enlargement will continue to 13.3 in. (33.8-cm) in XGA resolution with future options in the 15- to 17-in. (38.1- to 43.2-cm) range. Manufacturers are shifting their 390 × 470-mm^2 or 400 × 500-mm^2 substrate sizes to 550 × 650 mm^2 and will ultimately be targeting even 1 m diagonal size. They will simplify their array processes by lowering the number of photo masks and eliminate rubbing by photo alignment technologies. They will have to develop means to lower the resistance of bus lines. At the same time they are striving to widen the viewing angle and reduce power consumption of the panels. Figure 5 shows a cross section of a color AMLCD module and indicates how the driving circuitry is connected to the glass substrate carrying the TFTs. Mostly anisotropic conducting films are used to connect the ITO pads on the glass to the driver LSIs. Another connection method would be chip-on-glass.

Several technologies have been proposed to improve the originally rather narrow and nonuniform viewing characteristic of single-domain twisted nematic cells. Two of the major technologies are the formation of multi-domain cells with reversed LC orientations in each domain (by multiple rubbing and photolithographic processes) as suggested, e.g., by Kaneko *et al.* (1993) and the in-plane switching as first devised by Kiefer *et al.* (1992) and implemented by Kondo *et al.* (1996).

Standard AMLCD have five electrodes: three for gate, source, and drain of the transistor, one for the pixel electrode, and one common electrode. In most cases the common electrode is formed on the counter-substrate. As shown in Fig. 6, in-plane switching uses an interdigital electrode structure. The five electrodes are arranged in two layers on the TFT substrate side. The lower layer carries the gate and common electrodes, the upper one the source and drain electrodes, which are co-used as pixel electrodes. This approach avoids the indium-tin-oxide transparent electrode on the counter-substrate. The electric field is no longer vertically but horizontally oriented as indicated by the arrows in Fig. 6. Under the influence of this field the optical axis of the nematic LC molecules deviates from the polarization axis and the transmittance of incident light through crossed polarizers is gradually increased. Horizontally and vertically 70° viewing angle in both directions has been achieved at a contrast ration of 10.

The overall transmissivity of AMLCDs is determined by the transmissivities of the glass, 95%, the polarizers, 35%, the color filters, 25%, and the cell aperture (that portion of the cell which is transparent to light), which depending on design can vary between 40% and 65%. As a result the overall transmissivity of the cell is not more than 3.3% to 5.0%! One of the major research issues for AMLCDs is to improve transmissivity to reduce power consumption. The two other areas to work on for further

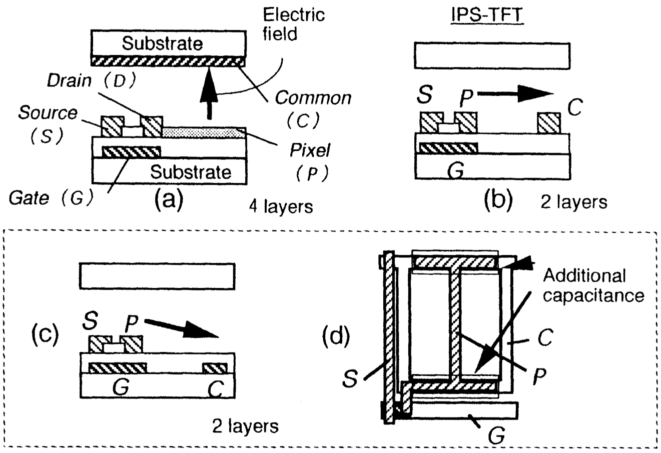

Fig. 6. (a) Cross section of conventional AMLCD and (b) in-plane-switching. (IPS-AMLCD). (c) In the IPS version the common electrode is located in the gate electrode layer. (d) The planar arrangement of (c) (after Kondo *et al.*, 1996).

power reduction are improvement of the efficiency of backlights and reduction of the power consumption of the drive-circuit unit. As a result of a comprehensive optimization the power consumption of a 10.4-in. (26.4-cm) SVGA AMLCD with a luminance of 70 cd/m^2 could be reduced from 3.5 W to 1.5 W (57% reduction) or with the same power consumption the luminance could be raised to 120 cd/m^2 (+72%) (Ohshita, 1996).

1.6 Plasma-Addressed Liquid-Crystal Displays

Despite their immense success in the portable display area and the anticipated large market share they might conquer in the monitor market in the near future, it is not predicted that AMLCDs will be available as meter-sized panels for consumer HDTV in the next ten years. The scaling difficulties of photolithography and metallization to sizes larger than 25 in. (63.5 cm) can elegantly be bypassed by a new approach called plasma-addressed liquid crystal (PALC) (Buzak, 1990). PALC uses the electrical switching properties of an ionized gas to address the liquid-crystal pixels actively. In the integrated LCD/plasma structure of the PALC, each scan line is defined by a plasma channel. The cathodes in the channels are sequentially scanned by applying a plasma discharge voltage that produces a full plasma within several microseconds. The analog video signals are fed to each column electrode. Except for the one row where the gas is ionized, the analog voltage has no effect on the light transmissivity of the liquid crystal since there is no electrical connection or conductor on the bottom side of the liquid crystal. Along the line above the ionized channel, however, the pixel capacitance can charge to the value determined by the data voltage. In effect, the plasma completes the electrical circuit between the data electrodes and the grounded channel electrode. Kakizaki *et al.* (1996) have presented a 25-in. (63.5-cm) PALC display for wall-hanging TV with 1.8 lm/W light efficiency, 250 cd/m^2, and high contrast even in bright ambient.

2. ACTIVE DISPLAYS

Active, self-emitting displays do not need any backlight. Every pixel emits light (and uses power) only as long as this is required for the image buildup. In the backlit LCDs every pixel including the OFF pixels receives full backlight power as long as the display is operating. In terms of energy management active displays have a conceptual advantage. On the other hand the separation of the functionalities "pixel addressing" and "light emission" allows a separate optimization of both. In the case of AMLCD this freedom in design has led to an overall energy consumption which is lower than in any active light-emitting device.

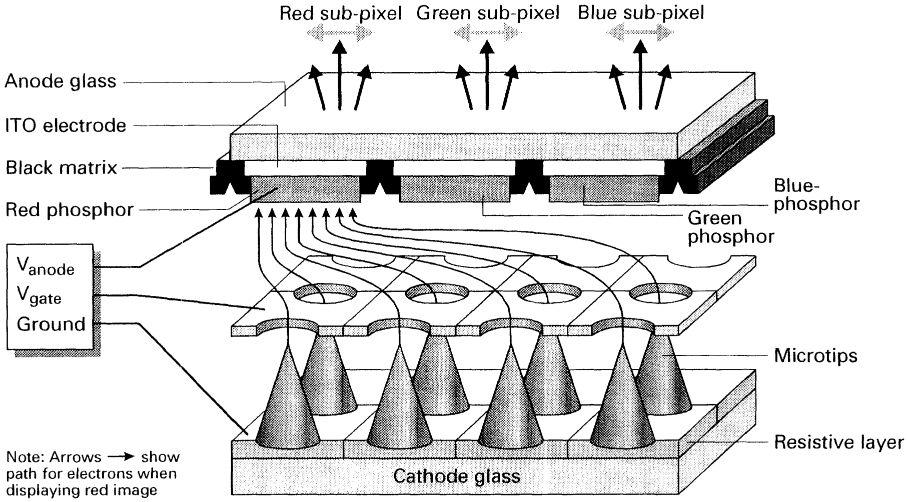

Fig. 7. Basic structure of field-emission displays (courtesy of PixTech).

2.1 Field-Emission Displays (FEDs)

The operating principle of these newly emerged vacuum electron devices is comparable to vacuum-fluorescent displays with the important difference that the electrons are not emitted by thermal activation from hot filaments but rather by field emission from a cold cathode. Electrons escaping from the cathode are accelerated towards a phosphor screen, where their energy is partly transformed into light (Fig. 7). In FEDs electrons are generated by a voltage difference between the cathode and closely positioned gates according to the Fowler–Nordheim tunneling equation for field emission.

Field enhancement induced by tips is the most widely investigated method to obtain useful currents at moderate macroscopic fields. One way of fabricating microtips is to make use of the wet etching properties of *n*-silicon (Jiang and White, 1993). This approach limits the display size to the available diameters of Si wafers. Meyer (1990) and his group have developed a fabrication process for molybdenum microtip cathodes on glass. The process also provides a gate electrode with a 1.4-μm self-aligned hole around each cathode tip. This original design was refined through the addition of a ballast resistive layer to average out random fluctuations of current yield between tips. The negative electron affinity on diamond-like carbon surfaces has fueled research in diamond emitters as a low-cost manufacturing alternative (Kumar *et al.*, 1994). The lwo-cost cathode architecture allows only the design of diode-type FEDs with the inherent disadvantage that gray scale is difficult and expensive to obtain (Courreges, 1996).

A big challenge in FED construction is the anode technology. The anode acceleration voltage determines the penetration depth of electrons into the phosphor grains. At low acceleration voltages, below 1000 V, the RGB phosphors are deposited on a conductive transparent electrode on the anode glass. If the phosphors are deposited on electrically isolated comb-like ITO structures, switched-anode FEDs become possible. Here the cathode is time multiplexed with a sequential addressing of color anodes as shown in Fig. 7. As a result of small electron penetration depth in the low-voltage approach, the grain surface largely determines the efficiency and the ultimate brightness. At higher acceleration voltages, above 5 kV, the phosphors can be covered by a thin layer of aluminum to provide surface conduction and avoid charging up. Here efficiency and brightness come close to those of CRTs. Low-voltage architecture avoids the need for focusing and allows for a low-cost spacer technology in the 200-μm cell gap. For high voltage the cell gap has to be increased at least ten times, which leads to

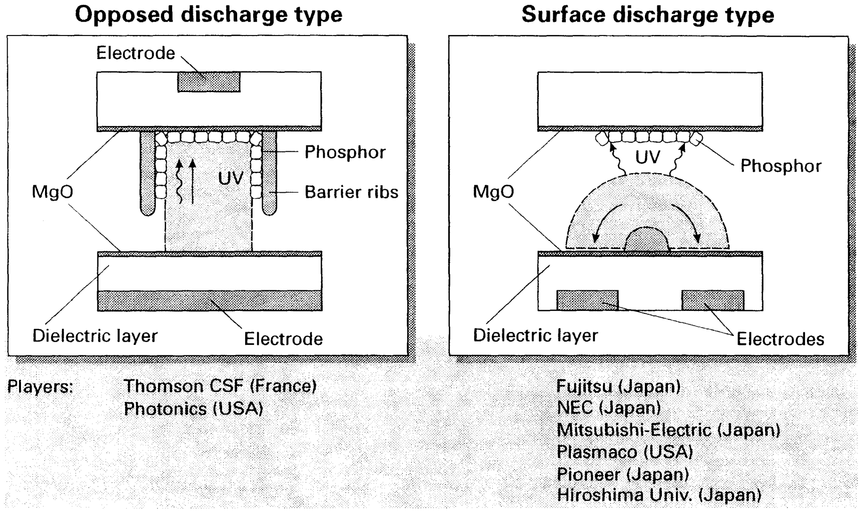
Opposed discharge type
Electrode
Phosphor
UV
MgO
Barrier ribs
Dielectric layer
Electrode
Surface discharge type
UV
Phosphor
MgO
Dielectric layer
Electrodes
Players:
Thomson CSF (France)
Photonics (USA)
Fujitsu (Japan)
NEC (Japan)
Mitsubishi-Electric (Japan)
Plasmaco (USA)
Pioneer (Japan)
Hiroshima Univ. (Japan)

Fig. 8. Two designs for color ac plasma displays.

complex solutions for the focusing and spacer technologies to suppress arcing and sputtering effects. The future will show whether the attractive properties of FEDs will prevail in terms of cost/performance ratio.

2.2 Plasma Displays (PDPs)

Plasma-display technology has gained a strong momentum in recent years. Demonstrations of exceptional large-area image quality have generated new interest in this 25-year-old technology. Plasma is now regarded as the leading direct-view flat-panel technology for large-area (meter-sized) displays. After the successful prototype demonstrations of 42-in. (107-cm) diagonal plasma displays, the major challenge now is to achieve high-volume manufacturing at affordable cost.

Like LCD panels, PDPs consist of two glass plates with helium or neon gas with an addition of 3 to 10% of xenon gas sealed between them. Electrodes are formed at the inner surface of the glass plates. To avoid catastrophic arc the current in the gas discharge must be limited: dc plasma displays use a resistor, and have the electrodes in intimate contact with the gas discharge; ac plasma displays limit the current with an internal glass dielectric covering the electrodes, which are capacitively coupled to the gas discharge.

In color PDPs the Xe component of the gas mixture emits ultraviolet light in the 147-nm range. In a line-at-a-time scanned-matrix display the overall brightness decreases with increasing number of scanned lines. It is very helpful in this respect that for both ac and dc plasma displays an inherent memory function is available. RGB light emission is achieved via the photoluminescence of phosphors excited by the UV light.

Two ac plasma-display electrode configurations are shown in Fig. 8.

The single-substrate structure separates the discharge cathode areas from the phosphor by applying the sustain voltage only the lower electrodes while the phosphor is on top. This approach avoids direct bombardment of the phosphor by energetic ions. The basic structure of single-substrate ac plasma displays is shown in Fig. 9.

The front glass carries the two parallel transparent sustain electrodes for each pixel. On the rear glass substrate the RGB phosphors are deposited on top of the address electrodes, separated by barrier ribs to avoid color cross-talk. The two glass plates are aligned orthogonally, forming pixels automatically at the intersection of the orthogonal electrodes. In PDP production the fabrication of the barrier ribs with their high aspect ratio (50 μm wide and 130 μm high) is a formidable

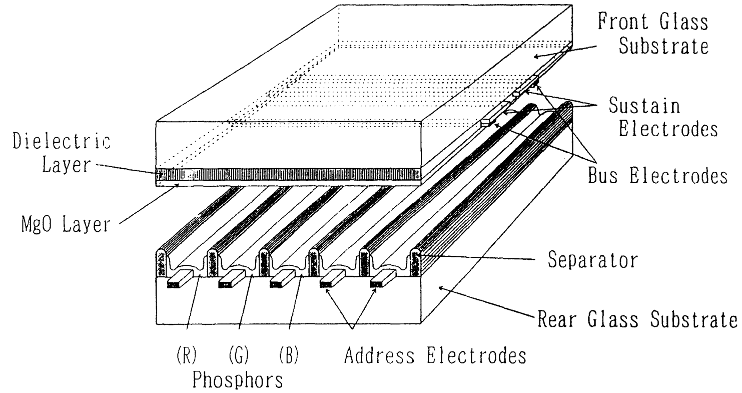

Fig. 9. Structural details of Fujitsu's ac plasma television screen.

problem. For low and medium resolution, multiple-screen printing can be used. For high-resolution displays this process is not suitable. For high-density barrier-rib formation, sand blasting and photosensitive molding of glass paste are being investigated.

In current generations of PDPs lamp phosphors like $(Y,Gd)BO_3:Eu$ are used for red, $BaMgAl_{14}O_{23}:Eu^{2+}$ for blue, and $Zn_2SiO_4:Mn$ for green emission. Phosphor aging is an important issue but seems to be solved. Probably the greatest challenge for plasma-display engineers is to increase the luminous efficiency of the devices. 0.7 lm/W is the highest efficiency reported so far for a commercial full-color plasma display (Shinoda, 1993). To increase this value improvements in the discharge efficiency, better optical coupling, and higher phosphor efficiencies have to be achieved.

To obtain gray scale in the memory mode of ac and dc plasma displays, the time that the pixel is ON in a given frame rate is modulated. This means that the pixels must be addressed multiple times per frame. In a binary chop technique, e.g., a 256-intensity gray scale would require 8 addresses per pixel per frame (Weber, 1996).

2.3 Electroluminescent Displays (ELDs)

2.3.1 Electroluminescence in Inorganic Materials. The performance of inorganic color thin-film electroluminescent (TFEL) displays has made significant progress in recent years due to the improvements of color phosphor materials with blue and broadband emissions. The reader is referred to the comprehensive article ELECTROLUMINESCENCE, together with its Addendum.

The classical EL phosphor is the yellow-emitting $ZnS:Mn$. This large–band-gap semiconductor is sandwiched between two insulating thin films and operated with ac voltage. The power dissipated in generating light is $8\,mW/cm^2$. Accordingly the power required to generate light from a 10-in. (25.4-cm) VGA display is just over 1 W; the luminous efficiency is 6 lm/W.

Two parallel roads have been taken to achieve color EL:

1. the development of an efficient white (broadband) phosphor that can be filtered and
2. the development of efficient red, green, and blue primary-color EL phosphors.

Tuenge and Karne (1991) have shown that an efficient red emission (1 lm/W) can be achieved with the combination of the standard yellow $ZnS:Mn$ phosphor and a red filter. The green $ZnS:Tb$ phosphor also has an efficiency near 1 lm/W (Härkönen *et al.*, 1990). In the blue part of the spectrum for some time the efficient $SrS:Ce$ phosphor was available. Actually this phosphor has a blue-green chromaticity and requires filtering to achieve a true blue color. Barrow *et al.* (1993) achieved an important breakthrough by showing that a group of ternary sulfides (thiogallates:

MGa_2S_4 where M stands for Ca, Sr, or Ba) activated with Ce were efficient blue-emitting phosphors.

White EL phosphors were first based on the broad emission of the system SrS : Ce,Eu, but the achieved brightness did not satisfy the needs of high-information displays. White emission from stacked layers of ZnS : Mn and SrS : Ce phosphors has achieved the luminance level required for filtered color ELDs. A white luminance of $340\,cd/m^2$, and red, green, and blue filtered luminance values of 39, 90, and $18\,cd/m^2$ at 60 Hz have been obtained by Velthaus *et al.* (1994) using a multilayer structure with 9 layers. The peak luminance could be further increased to $470\,cd/cm^2$ by dual layers grown by atomic-layer epitaxy (Soininen *et al.* 1994).

2.3.2 Electroluminescence in Organic Materials. Electroluminescence can also be obtained from numerous organic materials. Following the pioneering work of Tang and Van Slyke (1987) and Burroughes *et al.* (1990), research efforts have been directed at developing efficient organic EL devices based on thin films, driven typically below 10 V. Substantial improvements in efficiency, color, and stability of the devices have been achieved.

Organic EL devices generally consist of a hole-transport layer (HTL) and an electron-transport layer (ETL) sandwiched between current-injecting electrodes. Potential barriers at the interface of the two layers localize electrons and holes and thus enhance the probability of their recombination in this area. Also, the channel for radiative recombinations can be further optimized by introducing fluorescent centers near this interface. The two-layer structure is thus extended to include a thin emitting layer (EML) between HTL and ETL. The three-layer structure is particularly useful in tuning the EL colors (Tang, 1996).

The current in the organic layers in the EL device is principally space-charge or injection limited. Space-charge limitation is avoided by using organic films only a few hundred nanometers thick. As ETL organic materials, often a complex of 8-hydroxyquinoline with a *tris* ligand configuration around the central Al ion (called Alq) is used, while the EML contains a small amount of fluorescent molecules and the HTL is made of an aromatic amine material. Another successful approach is the use of light-emitting polymer material for the fabrication of flexible full-color displays. Low-barrier contacts are designed by matching the organic layers with electrode materials of proper work functions like Mg : Ag or Al : Li as cathode material and ITO for the anode. Red, green, and blue light-emitting devices have been obtained with efficiencies above 1 lm/W. Depending on the luminance output, operational half-life on the order of several thousand hours is achievable from orgainc EL emitters.

2.4 Visible-Light–Emitting Diodes (LEDs)

Visible LEDs have been on the market since the 1960's. Only in the last few years, however, have high-performance LEDs been available in all colors, allowing the fabrication of full-color flat-panel displays viewable in direct sunlight. Before 1985, when LED performance was generally in the range of less than 2 lm/W, LEDs were mostly limited to indicator-type applications. Since 1985 the performance has increased to as high as 50 lm/W for the highest performance type of red-emitting AlInGaP LEDs, With the development of GaInN devices, the range of colors has expanded to include the violet, blue, and pure green such that high-performance LEDs are available throughout the color spectrum. With an efficiency that exceeds many types of conventional lighting sources, LEDs can now be used in a variety of new application areas such as vehicle lighting, traffic signals, road signs, and large-area displays for advertisement. Red-light–emitting high-performance devices in AlGaAs technology were first introduced to the market around 1983. In 1991 red, orange, yellow, and green light-emitting diodes based on AlInGaP came up (Kuo *et al.*, 1990; Sugawara *et al.*, 1991). In 1993 this was followed by the successful commercialization of a new materials system for the green and blue part of the visible spectrum: the InGaN system (Nakamura *et al.*, 1995). The luminous efficiency vs. emission wavelength for commercial LEDs is shown in Fig. 10.

The AlInGaP system shows the highest performance in the red and amber region; the efficiency degrades sharply towards the green region. Nevertheless green AlInGaP LEDs can

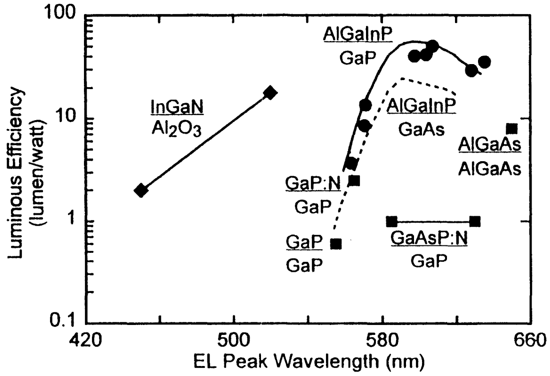

Fig. 10. Luminous efficiency of different materials systems vs. peak wavelength (after Craford, 1996).

be produced with an efficiency still doubling that of the conventional GaP : N green devices. The wavelength, however, being 560 nm or greater, is perceived as a yellowish green and therefore is not suitable for many full-color applications. The efficiency degradation towards the green is due to the fundamental transition in the band structure of the semiconductor alloy, which changes from "direct type" to "indirect type" with much less light-generation efficiency. The direct-to-indirect change is also the reason why AlGaAs LEDs work well in the infrared and red but not in the green part of the spectrum. The InGaN material system is a direct-type semiconductor all the way from the red to the ultraviolet. The best devices have been made in the blue and green regions. In the yellow there is a sharp falloff due to problems with materials growth and a bad match in atomic spacing between the different layers of the device structure.

3. PROJECTION DISPLAYS

In 1995 the total market for projection displays was in the order of US $ 4.6 billion. 58% of this market was covered by CRT-based products, and 42% by LCD-based products. The range of these products includes projectors for video home theater applications, projectors for overhead presentations, head-up displays for automotive systems, and many more.

At the heart of the projector is the light engine, which converts the electrical signal to an optical image (Stupp, 1996). The light engine can be either light emissive or light modulating. In the light-emissive group we find the CRT projectors and systems under development like field-emission or laser-based projection systems. Light-modulator systems employ either light-valve (LV) or light-amplifier (LA) systems to modulate the output light spatially.

The typical CRT projector uses three CRTs and three lenses—one for each of the primary colors. Consumer projectors are mostly CRT-based rear-projection systems with the light passing through a viewing screen to the viewer. The screen consists of a Fresnel lens and a double lenticular lens to minimize color separation in the image resulting from the angular separation of the sources.

LV or LA projector systems can be based on one or more (usually three) LVs. The systems based on multiple light valves often use dichroic mirrors and filters to separate the white light into color bands and to recombine the spatially modulated bands into a single converged image. Color information in single-LV systems can be obtained either by color-field sequence, by microfilters on the panel, or by angular color separation. The typical light source for LV systems is a high-intensity discharge lamp or occasionally a tungsten halogen lamp. The efficiency of light collection and utilization is determined by the size of the light arc, the area of the LV, and the usable angles of illumination. Since the peak white intensity is continuously generated, the maximum contrast achievable is determined by the attenuation capability of the LV. Active-matrix LCDs (frequently in the polysilicon version) are the most commonly used LVs in LCD-based projection systems. Polymer-dispersed LCDs are of specific interest, because they do not need polarizers with their inherent optical losses.

Another LV technology entering commercial application is based on micromechanics (Sampesell, 1993). The Digital Mirror Device^{TM} (DMD^{TM}) consists of an array of square mirrors ($17 \times 17\,\mu m^2$) torsionally suspended over a surface of monocrystalline silicon containing SRAM and other CMOS circuitry. The mirrors are tilted to $\pm 10\,^\circ$ under the influence of voltages from the underlying address circuitry. When the mirror is in the $+10^\circ$ position, the light from the lamp enters the projection lens and is transmitted to the screen. In the -10° position, the light misses the lens and is internally absorbed. Gray scale is achieved by time-multiplexing the position of the mirror. The switching speed is sufficient to achieve 8 bits per color in a 60-Hz frame time.

In LA devices the signal "writing" is done optically rather than electrically with an active matrix. The devices are multilayered. The input side consists of a photoconductor such as amorphous Si coupled to a reflective liquid-crystal cell that forms the output stage. A CRT or other light source is imaged onto the input photoconductor, generating an electrical image, which is transferred to the LC cell. An intense illumination onto this output stage is modulated by the cell, as with an LV. The input and output sides of the LA are separated by an efficient dielectric mirror that is part of the reflective LC cell and by a light-blocking layer to keep the output illumination from degrading the response of the input photoconductor.

Another emerging technology is AMLCDs fabricated on crystalline silicon wafers. These reflective-mode devices offer the potential of small pitch and high aperture ratio. In transmissive LVs, the aperture ratio is limited by the address lines and the TFTs. In reflective cells these can be fully hidden under the reflective back electrode. As with the DMD^{TM}, this produces small interpixel spacings with high aperture ratios.

Works Cited

Barrow, W. A., Coovert, R. C., Dickey, E., King, C. N., Laakso, C., Sun, S. S., Tuenge, R. T., Wentross, R. (1993) , *SID 93 Digest*, 761–764.

Burroughes, J. H., Bradley, D. D. C., Brown, A. R., Marks, R. N., MacKay, K., Friend, R. H., Burn, P. C., Holmes, A. B. (1990), *Nature* **347**, 539–543.

Buzak, T. S. (1990), *SID 90 Digest*, 420–423.

Courreges, F. (1996), *SID 96 Digest*, 45–48.

Craford, M. G. (1996), in: *HBF International Forum "Blue Light Emission for Future Imaging Technology,"* Tokyo: Hoso-Bunka Foundation, pp. 147–154.

Doane, J. W., Stefanov, M. E. pp. (1996), *Inf. Displ.* **12** (12), 18–21.

Härkönen, G., Härkönen, K., Törnqvist, R. (1990), *SID 90 Digest*, 232–235.

Ihara, S., Sugimoto, Y., Nakagawa, Y., Kuweta, T., Koh, H., Hasebe, H., Ruckmongathan, T. N. (1992), *SID 92 Digest*, 232–235.

Jiang, J. C., White, R. C. (1993), *SID 93 Digest*, 596–598.

Kakizaki, T., Tanamachi, S., Hayashi, M. (1996), *SID 96 Digest*, 915–918.

Kaneko, S., Hirai, Y., Sumiyoshi, K. (1993), *SID 93 Digest*, 265–268.

Kiefer, R., Weber, B., Windscheid, F., Baur, G. (1992), *J. Displ.* **92**, 547–550.

Kondo, K., Kinugawa, K., Konishi, N., Kawakami, H. (1996), *SID 96 Digest*, 82–84.

Kumar, N., Schmidt, H. K., Clark, M. H., *et al.* (1994), *SID 94 Digest*, 43–46.

Kuo, C. P., Fletcher, R. M., Osentowski, T. D., Lardizable, M. C., Craford, M. G. (1990), *Appl. Phys. Lett.* **57**, 2937–2939.

Meyer, R. (1990), *Eurodisplay* **90**, 374.

Morozumi S. (1996), *SID 96 Seminar Lecture Notes*, M8/1–35.

Mosley, A. (1997), "The Structure, Performance and Future of Passive LCDs," in: L. W. MacDonald, A. L. Lowe (Eds.), *Display Systems*, Chichester: John Wiley, pp. 173–189.

Nakamura, S., Senoh, M., Iwasa, N., Nagahama, S., Yamaka, T., Mukai, T. (1995), *J. Appl. Phys.* **34**, L1332–L1335.

Ohshita, S. (1996), *Displ. Dev.* **13**, 28–30.

Pfeiffer, M., Yang, D. K., Doane, J. W. (1995), *SID 95 Digest*, 706–709.

Sampsell, J. B. (1993), *SID 93 Digest*, 1012–1015.

Scheffer, T. J., Clifton, B. (1992), *SID 92 Digest*, 228–231.

Shinoda, T., Wakitami, M., Nanto, T., Yoshikawa, K., Ohtsuka, A., Hirose, T. (1993), *SID 93 Digest*, 161–164.

Soininen, E., Leppänen, M., Törnqvist, R., Viljanen, J. (1994), *7th International Workshop on EL, Beijing 1994*, pp. 97–100.

Stupp, E. (1996), *Inf. Displ.* **12** (10), 20–24.

Sugawara, H., Ishikawa, M., Hatakoshi, G. (1991), *Appl. Phys. Lett.* **58**, 1010–1012.

Tang, C. W. (1996), *SID 96 Digest*, 181–184.

Tang, C. W., Van Slyke, S. A. (1987), *Appl. Phys. Lett.* **151**, 913–915.

Tuenge, R. T., Kane, J. (1991), *SID 91 Digest*, 279–281.

Uchida, T. (1996), *SID 96 Digest*, 31–34.

Velthaus, K. O., Troppenz, U., Hüttl, B., Herrmann, R., Mauch, R. H. (1994), *International Display Research Conference 1994*, pp. 346–348.

Weber, L. F. (1996), *SID 96 Seminar Lecture Notes*, M6/1–41.

Yamada, Y., Yamamoto, N., Nakamura, K., Koshobu N. *et al.* (1995), *SID 95 Digest*, 789–792.

Yamazaki, T., Kawakami, H., Hori, H. (1996), in: SEMI Standard FDP Technology Group (Ed.), *Color TFT Liquid Crystal Displays*, Mountain View, CA: Semiconductor Equipment and Materials International.

EARTH, INTERIOR STRUCTURE OF THE

(An Addendum to *Encyclopedia of Applied Physics*, Volume 5, pages 127–148.)

SHUN-ICHIRO KARATO, *Department of Geology and Geophysics, University of Minnesota, Minneapolis, Minnesota, U.S.A.*

EIJI OHTANI, *Institute of Mineralogy, Petrology and Economic Geology, Tohoku University, Sendai, Japan*

1. **Earth Structure and Composition** 417
2. **Mantle Dynamics** 419
3. **Core Dynamics and Generation of the Geomagnetic Field** 420
Works Cited.................... 420
Further Reading 420

1. EARTH STRUCTURE AND COMPOSITION

Significant progress has occurred in seismic tomography. The resolution in detecting velocity heterogeneity has been improved, and it is now well established, on the basis of both P-wave and S-wave tomography, that some subducted slabs penetrate down to the bottom of the lower mantle (North and South America), although others appear to be deflected near the 660-km discontinuity (e.g., in the western Pacific) (e.g., van der Hilst *et al.*, 1997). These latter slabs, however, also are likely to reach to the bottom of the lower mantle after a certain time, judging from the correlation of high-velocity anomalies at the bottom of the mantle and the history of subduction. Variation of the behavior of subducted slabs appears to correlate with the velocity of trench migration and the temperature of slabs: slab penetration deep into the lower mantle occurs where the velocity of trench migration is small and/or the temperature of the slabs is high (young and/or fast subducting slabs). Thus, the style of mantle convection appears to be neither a simple whole-mantle convection nor strictly layered convection at the present time and during the recent past.

Seismic tomography has also shown large-scale low-velocity anomalies at the bottom of the mantle. They are located beneath the south Pacific and Africa. The presence of partial melting is suggested (Williams and Garnero, 1996). These regions are the likely source for plumes that result in ocean island volcanoes. However, plumes with shallower sources (transition zone) are also inferred (e.g., in Europe).

The presence of anisotropic structures in the inner core has been established both by the analysis of body-wave travel-time anomalies and by the analysis of anomalous splitting of free oscillation peaks. The anisotropy is rather large ($\sim 3\%$) and has nearly axial symmetry, with the symmetry axis slightly ($\sim 10°$) tilted with respect to the current rotation axis (e.g., Song, 1997). The anisotropy is best explained by the preferred orientation of anisotropic iron crystals in the inner core, although neither the physical mechanisms of alignment nor the crystal structure of iron in the inner core is well understood. By analyzing old seismic records, two groups of seismologists have detected the migration of the pole of anisotropy, suggesting rotation of the inner core with respect to the mantle (Song and Richards, 1996; Su *et al.*, 1996). It is remarkable that this "super-rotation" of the inner core was predicted by theoretical modeling of the geodynamo as caused by magnetic coupling between the inner and outer cores (Glatzmaier and Roberts, 1995).

In contrast to the inner core, the bulk of the Earth's mantle is nearly isotropic except for the near-surface layer (lithosphere and asthenosphere), the D$''$ layer (the bottommost layer of the mantle), and possibly the layer near the 660-km discontinuity. Anisotropy in the mantle is likely caused by lattice-preferred orientation of elastically anisotropic minerals such as olivine or perovskite or by laminated structures. Boundary layers of convection are the likely place to cause seismic anisotropy be-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

cause of highly concentrated deformation at high stresses and the accumulation of chemically distinct materials that tend to form laminated structures. Thus the presence of strong anisotropy in the top and the bottom (D″) layers indicates that they are the boundary layers of convection, and also that the possible presence of an anisotropic layer around the 660-km discontinuity suggests the presence of a mid-mantle boundary layer, implying that convection is somewhat layered. It is surprising that the bulk of the lower mantle is nearly isotropic despite the fact that its constituent minerals are highly anisotropic. The most likely explanation is to assume that the deformation mechanism in the lower mantle is grain-size–sensitive diffusion creep. In this case, rheology is linear (rate of deformation increases linearly with stress) and the grain-size reduction due to phase transformations in downgoing slabs will cause significant rheological weakening.

New seismological observations on gross Earth structure include the evidence of the 520-km discontinuity and a discontinuity around ~1000 km in some areas (e.g., Kawakatsu and Niu, 1994). These discontinuities are rather subtle as compared to those at ~410 km and at ~660 km, which are more distinct and are global features. The 520-km and ~1000-km discontinuities may be regional as opposed to global. The 520-km discontinuity may be attributed to a phase transformation in Ca-rich silicate to $CaSiO_3$ perovskite, but the origin of a discontinuity around 1000 km is uncertain. No phase transformations in major minerals are known at conditions equivalent to this depth. A chemical layering caused by subduction of oceanic lithosphere is a possible mechanism.

The chemical composition of the Earth's mantle has continuously been a subject of extensive studies. Advances in high-pressure technology combined with high-energy x-ray sources from synchrotron radiation now allow us to measure directly the molar volume of minerals under conditions equivalent to the shallow lower mantle (to ~1000 km depth). Experimental techniques to measure seismic-wave velocities under deep-Earth conditions (to ~500 km deep so far) have also been developed. With these techniques, the density, compressibility, and thermal expansion of $MgSiO_3$ perovskite have been determined under shallow lower-mantle conditions (Funamori *et al.*, 1996). Also, seismic-wave velocities of the α (olivine) and β (modified spinel) phases of $(Mg,Fe)_2SiO_4$ have been measured near conditions equivalent to the 410-km depth. The results of these studies have been compared with seismological observations to place constraints on the chemical composition of the Earth's mantle. The main issue here is if the Earth's mantle is chemically layered or not. Despite much progress in high-pressure mineral physics, the issue is far from resolved. The major difficulty is the trade-offs among various parameters (e.g., temperature, iron/silicon content). For example, seismological data can be fitted to either a chemically homogeneous model or a chemically layered model with some iron and/or silicon enrichment in the lower mantle, when a significant temperature increase around the 660-km discontinuity is assumed. However, in view of the strong evidence of penetration of some downgoing slabs into the lower mantle, a strictly chemically layered mantle structure appears to be ruled out.

One important discovery in mineral physics is a dramatic effect of Al on the crystal structure of $(Mg,Fe)SiO_3$ perovskite. It was found, from high-pressure experiments using a multi-anvil apparatus, that a small amount of Al changes the valence state and the crystallographic site of Fe in perovskite (Wood and Rubie, 1996). Fe appears to occur mostly as ferric iron and occupies six-coordinated sites (similar to Si) when a small amount of Al is present. This behavior affects the partitioning of Fe between perovskite and magnesiowüstite and also is likely to change other physical properties including electrical conductivity and elastic properties. Discussions on the chemical composition of the lower mantle will have to take this point into account.

Another potentially important source of information to place constraints on the chemical composition of the deep interior of the Earth is the chemical composition of rocks that are carried from the deep interior of the Earth. During the last few years, several rocks have been discovered that are considered to have originated in the transition zone or the lower mantle of the Earth on the basis of their mineralogy. They include peridotites from South Africa containing unusually silica-rich garnets, a peridotite from the Swiss Alps containing olivine grains with exotic Ti-rich inclusions, and diamonds from Brazil that contain a pair of minerals (magnesiowüstite and pyroxene) that could have been in chemical equilibrium in the lower mantle. These rock specimens provide direct information as to the chemical composition of the deep interior of the Earth. However, the relevance of their compositions to the composition of average Earth's mantle is not always clear, because these samples may have been carried from some unusual regions.

The sharpness and the regional variation in depth of seismic discontinuities have been studied in detail. Many of the seismic discontinuities appear to be rather sharp (major changes in physical properties occur within a few kilometers interval in most cases). Regional variation in the depths of these discontinuities has also been detected. The most marked one is the variation in the depth of "660-km" discontinuity. The depth of the "660-km discontinuity" (boundary between upper and lower mantles) appears to vary on a regional scale (Shearer and Masters, 1992). It is significantly deeper than average in the regions where the temperature is considered to be lower than average as a result of the subduction of oceanic lithosphere. This is consistent with the results of mineral physics, which indicate that the phase transition from upper-mantle minerals to lower-mantle minerals occurs at higher pressures at lower temperatures. The depth of the 410-km discontinuity does not vary as much as that of the 660-km discontinuity. Because the temperature sensitivity of the depth of the phase transformation corresponding to the 410-km discontinuity is as large as that of the 660-km discontinuity (but with a different sign), this observation suggests that lateral variation in temperature around the 410-km discontinuity is not as large as that around the 660-km discontinuity.

2. MANTLE DYNAMICS

The study of mantle convection through computer simulation has made major progress during the last few years. The increasing capability of computation allows incorporation of more and more "realistic" physical properties in numerical modeling. The most important result in this area is the demonstration that the density contrast associated with the 660-km discontinuity has an important effect in modifying the convection pattern. The lateral density contrast near the 660-km boundary as inferred from the undulation of the 660-km discontinuity tends to prevent vertical mass transfer across this boundary. The results of numerical modeling during the early 1990s showed that this effect is large enough to change convection pattern to intermittently layered convection when the contrast in viscosity between descending currents and the surroundings is small (e.g., Tackley, 1995).

In contrast to the effects of density variation, the effects of viscosity structures on the style of convection are less well understood. The viscosity profile in the Earth's mantle is much less constrained than the density profile, but viscosity could change very dramatically both radially and laterally (e.g., Karato, 1997). Recent analysis of vertical crustal movement associated with the post-glacial rebound combined with the analysis of Earth's rotation and geoid observations yields a unified model characterized by a significantly higher viscosity in the deep lower mantle than in the upper mantle (Forte and Mitrovica, 1996). Such a viscosity profile provides a natural explanation for the fixed position of hotspot volcanoes, which may be caused by plumes originating in the deep portions of the lower mantle. Lateral viscosity variation, particularly the viscosity of subducting lithosphere, has been shown to significantly affect convection patterns. The significant role of the density contrast associated with the 660-km discontinuity in changing the convection pattern is important only when the descending current does not have much higher viscosity than the surrounding mantle. The observed significant deflection of subducted oceanic lithosphere in the western Pacific therefore suggests that subducted lithosphere there is rather weak. Likely causes of weakness include grain-size reduction associated with phase transformation(s) (Riedel and Karato, 1997).

Deformation in deep subducted oceanic lithosphere often occurs through seismic faulting. The origin of these deep earthquakes has been an enigma, because brittle fracture that is responsible for shallow earthquakes is impossible under high confining pressures. The role of phase transformations in causing instabilities in deformation has been demonstrated in laboratory studies, which provides a potential mechanism for deep earthquakes (Green, 1994). However, the new data obtained from the Bolivia and the Fiji deep earthquakes during 1994 showed much wider fault widths than predicted from simple models. Active research is ongoing to understand better the role of phase transformations in deep earthquakes.

3. CORE DYNAMICS AND GENERATION OF THE GEOMAGNETIC FIELD

The last few years witnessed significant progress in the understanding of the origin of magnetic field generation in the Earth's core. The first successful modeling was made in U.S.A. and in Japan in which both the equation of motion of a rotating viscous fluid and the electromagnetic induction equation are simultaneously solved to show the generation of a nearly dipole magnetic field (Glatzmaier and Roberts, 1995, 1997; Kageyama and Sato, 1997). The driving forces in these models include both thermal and chemical buoyancy forces, the latter being due to the solidification of nearly pure iron from iron-rich alloy. The important effects of the inner core are also noted in these models. These models were able to show a stable nearly dipole magnetic field, but some instabilities causing reversals of magnetic field were also seen. One of the important outcomes of this type of model is the prediction of a "superrotation" of the inner core as a result of magnetic coupling between outer and inner cores due to a strong toroidal field (Glatzmaier and Roberts, 1995). "Superrotation" of the inner core may also be attributed to viscous coupling.

Observationally, one of the most fascinating observations is the suggestion that the virtual geomagnetic pole (VGP: the pole of magnetic dipole field projected to the Earth's surface) follows specific paths during the reversal. This notion was inferred from the analysis of magnetic records in sediments, and the results show that the VGP follows a band in the Earth along which seismic-wave velocities are anomalously high at the bottom of the mantle (Laj *et al.*, 1991). This suggests that the temperature anomalies at the core–mantle boundary caused by mantle convection may influence the convection pattern in the outer core that in turn causes the variation in magnetic field (Kuang and Bloxham, 1997).

List of Works Cited

Forte, A. M., Mitrovica, J. X. (1996), *Geophys. Res. Lett.* **33**, 1147–1150.

Funamori, N., Yagi, T., Utsumi, W., Kondo, T., Uchida, T. (1996), *J. Geophys. Res.* **101**, 8257–8269.

Glatzmaier, G. A., Roberts, P. H. (1995), *Phys. Earth Planet. Inter.* **91**, 63–75.

Glatzmaier, G. A., Roberts, P. H. (1997), *Nature* **377**, 203–209.

Green, H. W., II. (1994), *Sci. Am.* **271** (3), 64–71.

Kageyama, A., Sato, T. (1997), *Phys. Rev.* **E55**, 4617–4626.

Karato, S. (1997), in: D. J. Crossley (Ed.), *Earth's Deep Interior*, London: Gordon and Breach, pp. 223–272.

Kawakatsu, H., Niu, F. (1994), *Nature* **371**, 301–305.

Kuang, W., Bloxham, J. (1997), *Nature* **389**, 371–374.

Laj, C., Mazaud, A., Weeks, R., Fuller, M., Herro-Brevera, E. (1991), *Nature* **351**, 447.

Riedel, M. R., Karato, S. (1997), *Earth Planet. Sci. Lett.* **148**, 27–44.

Shearer, P. M., Masters, T. G. (1992), *Nature* **355**, 791–796.

Song, X., Richards, P. G. (1996), *Nature* **382**, 221–224.

Song, X. (1997), *Rev. Geophys.* **35**, 297–313.

Su, W. J., Dziewonski, A. M., Jeanloz, R. (1996), *Science* **274**, 1883–1887.

Tackley, P. J. (1995), *Rev. Geophys. Suppl.* (U.S. Natl. Rep. Int. Union Geo. Geophys., 1991–1994), 275–282.

van der Hilst, R. D., Widiyantoro, S., Engdhal, E. R. (1997), *Nature* **386**, 578–584.

Williams, Q., Garnero, E. J. (1996), *Science* **273**, 1528–1530.

Wood, B. J., Rubie, D. C. (1996), *Science* **273**, 1522–1524.

Further Reading

Lay, T., Wallace, T. C. (1995), *Modern Global Seismology*, San Diego: Academic Press. Introduction to global seismology with basic concepts and some recent results.

Crossley, D. J. (Ed.) (1997), *Earth's Deep Interior*, London: Gordon and Breach. A collection of review papers related to Earth's interior. Topics discussed include seismic tomography, mineral physics, mantle convection, and the geodynamo.

FUSION, INERTIAL CONFINEMENT

(An Addendum to *Encyclopedia of Applied Physics*, Volume 6, pages 545–573.)

BARRETT H. RIPIN, *The American Physical Society College Park, Maryland, U.S.A.*

Introduction 421
1. **Overview of Inertial Confinement Fusion (ICF)** 421
2. **Principles of X-Ray Drive** 423
3. **Current Status of ICF Research** 425
3.1 Laser Facilities 425
3.2 Target Experiments 425
3.3 Predictive Capability 426
3.4 NIF Ignition Target Point Design 426
3.5 Target Fabrication 427
4. **National Ignition Facility** 427
4.1 Science-Based Stockpile Stewardship 428
5. **Heavy-Ion Driver Development** 429
6. ***Z*-Pinch Driver** 429
7. **Alternate Concept: Fast Ignitor** 430
8. **Inertial Fusion Energy** 431
9. **Prognosis** 431
Works Cited 431
Further Reading 432

INTRODUCTION

Developments in the inertial-controlled fusion (ICF) program occurring in the mid to late 1990's have raised expectations that ignition of an ICF pellet in the laboratory may be demonstrated as early as 2005.

1. OVERVIEW OF INERTIAL CONFINEMENT FUSION (ICF)

The basic physics principles of the inertial-confinement fusion concept are described in detail in the original article. As a brief reminder of the concept, recall that the objective is to implode a small pellet containing deuterium (D) and tritium (T), heavy isotopes of hydrogen, to high enough density and temperatures that a substantial fraction of the fuel nuclei undergo fusion reactions and yield bursts of energy. Calculations predict that the fusion energy release in this process can be of the order of 10^2 times the energy driving the implosion. The ratio of released energy to driver energy is called gain. If high-gain ICF can be demonstrated, then it raises hopes that it might be further developed for use in civilian electric-power generation.

To achieve the conditions for high-gain ICF, the outside surface of a pellet of about 1 mm diameter that encloses a layer of cryogenic DT fuel on the shell's inner surface is irradiated by an intense (10^{14} to 10^{15} W/cm^2) energy pulse from a directed energy source, such as a laser. The energy deposition on the pellet surface causes it to ablate and heat, become a hot plasma, and blow away from the surface at high velocity. Pressures of tens of megabars created by the hot ablating plasma surface material drives the remaining shell and interior fuel inward, compressing the DT fuel to over 10^3 times solid density. The driver pulse characteristics, and hence the ablation pressure, are staged such that the bulk of the compressed fuel remains on a low isentrope (dense but cold material) except for a small fuel region in its center. The drive pulse is shaped in time so that the central region, called the ignitor, is heated to temperatures of about 10 keV. This configuration is shown schematically in Fig. 1. If the hot-spot density-radius product is high enough (>0.3 g/cm) to stop the fusion-produced alpha particles at this temperature, then self-heating from the nuclear fusion reactions will initiate a fusion "burn" front that propagates outward throughout the main fuel volume. If all this is done properly, the energy yield of the explosion will greatly exceed the incident energy, yielding net energy. Although most ICF research to date has used a large laser driver, in principle, particle beams from accelerators could also be used.

Figure 2 illustrates the two general ap-

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

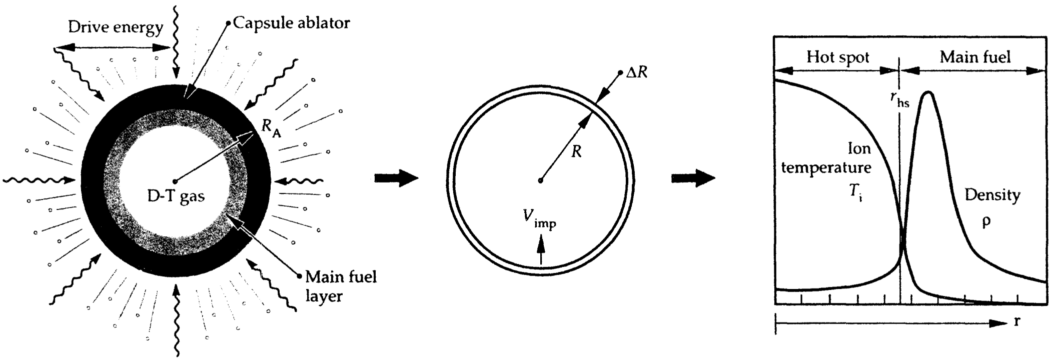
Drive energy
Capsule ablator
R_A
D-T gas
Main fuel layer
ΔR
R
V_{imp}
Hot spot
Main fuel
r_{hs}
Ion temperature
T_i
Density
ρ
r

Driver–target coupling
$\Rightarrow I_r \lesssim 10^{15}$ W/cm^2 or $\lesssim 300$ eV

To control:
- Absorption/preheat
- X-ray conversion
- Transport/drive

Symmetry:	$\frac{R_A}{r_{hs}}$ =	Convergence ratio	≈ 20–35 ⇒	Coupling $\eta \approx 10–15\%$
Stability:	$\frac{R}{\Delta R}$ =	In-flight aspect ratio	≈ 25–35 ⇒	$I_R \gtrsim 4 \times 10^{14}$ W/cm^2 or 250 eV Surface <1000 Å
Ignition:	• $T_{i,hs} = 10$ keV • $\rho r_{hs} \approx 0.3$ g/cm		⇒	V_{imp} 3–4 × 10^7 cm/s for E_{driver} = 1–2 MJ

FIG. 1. Elements of inertial confinement fusion. (Left) A schematic of an irradiated pellet, (middle) aspect ratio (R/ΔR) of imploding pellet, and (right) density–temperature profile of hot spot (ignitor) and main compressed DT fuel at point of ignition. Tables summarize constraints on drive symmetry, hydrodynamic stability requirements, and ignition parameters. [From LLNL (1995).]

proaches to ICF. When the driver energy beam directly irradiates the pellet, the ICF approach is called "direct-drive." If the driver energy first strikes high-Z material inside of an enclosure, a hohlraum, that converts the beam energy into intense x rays that subsequently strike, ablate, and implode a pellet, the approach is called "indirect drive," "x-ray drive," or the "classified approach." There was a major declassification and disclosure of the x-ray drive approach to ICF in 1993. Most of what follows is a description of the

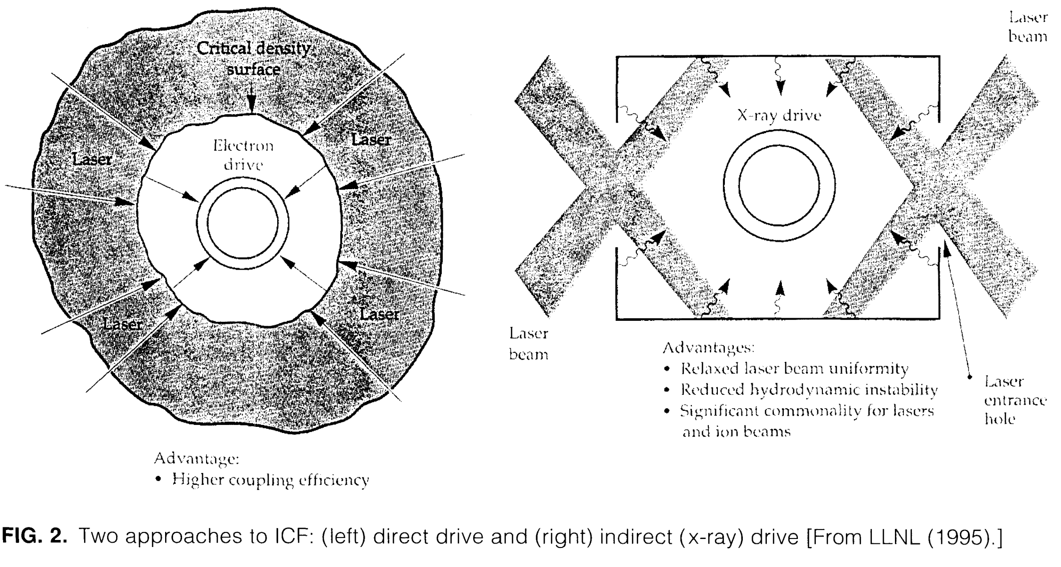
(a) Direct (electron) drive

(b) Indirect (x-ray) drive

Critical density surface
Laser
Laser
Electron drive
Laser
Laser
Advantage:
• Higher coupling efficiency
Laser beam
X-ray drive
Laser beam
Advantages:
• Relaxed laser beam uniformity
• Reduced hydrodynamic instability
• Significant commonality for lasers and ion beams
Laser entrance hole

FIG. 2. Two approaches to ICF: (left) direct drive and (right) indirect (x-ray) drive [From LLNL (1995).]

x-ray drive approach that was omitted in the original article because of classification restrictions, along with more recent advances.

There has also been a significant shift in the programmatic rationale for ICF since 1993. Most of the funding for ICF research in the United States comes from the Department of Energy's defense programs, ostensibly as a laboratory means to gain general information related to nuclear weapons design, operation, and effects. With the breakup of the Soviet Union and the advent of a comprehensive nuclear test ban, the ICF program has become a key component in the U.S. stockpile stewardship maintenance program (SSMP) in the mid-1990's. The purpose of SSMP is to maintain a nuclear deterrent capability in the absence of underground nuclear testing.

2. PRINCIPLES OF X-RAY DRIVE

The x-ray drive approach has actually been the field's principal effort since 1976, although little was known about it in the general scientific community because of U.S. Department of Energy (DOE) secrecy restrictions. The formerly classified information release, combined with recent advances in the scientific basis of the ICF research program and in driver and target fabrication technologies, has raised expectations that an x-ray–driven ICF target could achieve ignition in the laboratory in the not too distant future. A comprehensive review of the development of x-ray drive ICF approach can be found in Lindl (1995).

The distinctive element of x-ray drive is the hohlraum, which is a radiation-filled enclosure made of high-Z material (typically gold) that surrounds a pellet (also known as a 'capsule'), such as depicted in Fig. 2. Laser beams enter the hohlraum by focusing through small entrance holes. Copious x rays are produced where the beam spots strike the interior hohlraum walls. Most x rays are absorbed by and heat the remainder of the hohlraum walls, which in turn, reradiate x rays forming a bath of x rays with an equilibrium distribution. Typical beam–to–x-ray energy conversion efficiencies are reasonably high (70% to 80%). Laser beams are usually arranged in two clusters. Hohlraums are usually cylindrically shaped with an entrance hole in each end cap to accept the beam clusters. The optimal hohlraum design depends upon details of the driver type and pellet design. The beams are arranged to strike the walls such that the x-ray flux falling on the pellet surface is a uniform mixture of x-rays emitted directly from the laser spots and the somewhat lower intensity x-ray albedo from other portions of the hohlraum. The exact beam placement required to achieve the required uniformity and symmetry for high-convergence ratio implosions is an area of active research.

The pressure P_{a} applied at the ablation surface of the pellet is what determines the implosion dynamics of an ICF pellet, whether that pressure be generated by direct-drive or indirect-drive process. The ablation pressure accelerates the remaining shell material inward at a velocity given by the spherical version of the rocket equation, $V_{\mathrm{imp}} = V_{\mathrm{pl}} \ln(m_0/m)$. Here m_0 and m are respectively the original and final mass/surface areas of the pellet shell. The velocity of the plasma ablating away from the pellet, V_{pl}, is given by $V_{\mathrm{pl}} = P_a/[dm/dt]$).

X rays are well suited to generation of high ablation pressures because they deposit their energy directly into the high-density regions of the pellet shell. Dense material deposition results in a higher mass ablation rate (dm/dt) and an ablation velocity that is lower and better matched to the peak efficiency point (where $V_{\mathrm{imp}} \approx V_{\mathrm{pl}}$) of the rocket equation. X rays inside a hohlraum can be characterized by a blackbody temperature, at least within the relevant wavelength regime containing most of the energy. The x-ray intensity impinging on the ablation surface is given approximately by the relation $I_{\mathrm{a}} \approx T_{\mathrm{r}}^4$. By way of calibration, a hohlraum temperature of $T_{\mathrm{r}} = 300\,\mathrm{eV}$ has the equivalent x-ray intensity of about $I_{\mathrm{a}} = 10^{15}\,\mathrm{W/cm^2}$.

For typical ignition target designs, the effective radiation temperatures are in the range 225 to 300 eV. The high-temperature limit is set primarily by the need to keep incoming laser-beam irradiances low enough to avoid stimulating parametric instabilities in the hohlraum plasma. The low-temperature limit comes from the need to achieve the minimum shell velocity for ignition, which is about $3 \times 10^7\,\mathrm{cm/s}$. Here the consensus is that the pellet shell in-flight aspect ratio ($\mathrm{IFAR} = R/\Delta R$) is limited to 30 or less (i.e., this is the thinnest feasible shell) because of excessive hydrodynamic instability.

Capsules need to converge in radius by a factor of 25 to 35 in order to form the small hot-spot region needed to ignite the pellet at peak compression. For a pellet to converge this far and remain spherical requires x-ray fluxes uniform to within 1 to 2%. x-ray uniformity improves with larger hohlraums for a given pellet size, but with a loss of efficiency. Alternatively, optimal placement of beam deposition patterns and careful balance between them can reduce x-ray asymmetries inside a smaller hohlraum.

In addition to good overall x-ray symmetry impinging on the pellet surface, the pellet shell must resist excessive hydrodynamic instability. One of the most critical regions for hydrodynamic instability in indirect-drive pellets is at the inner surface of the main fuel. This is because hydrodynamic instability will tend to mix cold dense fuel into the hot-spot region, cooling it and quenching ignition. The most important factor affecting inner-surface mix is the initial inner-surface roughness of the fuel layer. The upper limit for the inner-surface roughness of the cryogenic DT in typical NIF (National Ignition Facility; see Sec. 4) ignition target designs is about 1 μm when a CH ablator shell is employed, and a factor of two or three higher if a beryllium ablator shell can be used. Hydrodynamic instability causing mix is fed by areal mass nonuniformities. Beryllium is a more tolerant material than CH because its density is lower so that the effect of a given roughness is reduced. Also, a Be ablator shell would be about twice as thick as a plastic (CH) shell, which further reduces the growth rate of the hydrodynamic instability that leads to mix.

Although there is a small loss of efficiency in the process of converting the driver energy into x rays, x-ray drive has a number of potential advantages over the direct drive approach.

1. The pellet finds itself immersed in an x-ray bath that tends to be naturally uniform on account of the broad pattern of the x-ray emission process plus absorption and reradiation by the hohlraum walls. The x-ray emission is relatively insensitive to nonuniformities in the driver beams and, to some degree, to the geometric pattern of incoming beam clusters.
2. X rays penetrate and directly deposit their energy in the vicinity of the pellet ablation surface ($\sim 10^{24}$ electrons/cm^3), which is a few times solid density. This is a more efficient process than the case of direct laser irradiation in which initial laser-light absorption occurs at much lower densities (near or below the critical density, where the laser-light frequency equals the plasma frequency, $\sim 10^{22}$ electrons/cm^3 for 0.35-μm light), requiring the energy then to diffuse up a density ramp to the ablation surface.
3. x-ray deposition creates a softer density gradient in the vicinity of the ablation region that reduces the shell's susceptibility to hydrodynamic instability over that of the direct-drive case.
4. The principal proponents of the x-ray approach, DOE weapons laboratories, have considerable experience, infrastructure, and data base upon which to draw in this area.

This is not to claim that the x-ray ICF approach is ideal or a sure thing. For example, only about 5% of the driver energy actually ends up as useful hydrodynamic implosion energy. This is a relatively inefficient use of high-quality energy from the driver beam. The losses take the form of direct conversion loss of beam energy into x-rays, loss of x-ray energy out the holes in the hohlraum and by reabsorption in the hohlraum walls, and waste heat carried away by the ablating plasma.

Scale sizes of plasmas generated inside a hohlraum can be large, creating a favorable environment for laser parametric instabilities, such as Brillouin or Raman backscatter instabilities. Sizable levels of backscatter would reduce beam energy available to generate x rays and thereby lower efficiency. Avoidance of parametric instabilities tends to cap the maximum laser-beam irradiances that can be utilized.

Finally, there always exists a small residual asymmetry in the x-ray albedo falling on the pellet surface due to the beam deposition geometry. Although the x-ray radiation pattern falling on the pellet is made nearly spherically symmetric by multiple absorption and reradiation of x rays by the hohlraum walls, there can still be enough of a residual asymmetry to affect the implosion symmetry adversely. Excessive asymmetry can seed hydrodynamic instabilities during pellet implosion that prevent formation of a small ignitor region, which, in turn, would increase the fuel needed to be brought to 10 keV and, consequently, the driver energy required for ignition.

For energy applications, the presence of extraneous mass, particularly high-Z materials typically used in hohlraums, is a practical disadvantage. This is because some of the accumulated material from the millions of capsules exploded in an energy reactor will become activated by the fusion neutron flux and must be extracted and disposed of.

Experiments on Nova and elsewhere, as well as computer simulation projections, have demonstrated that most of these potential problems can be averted or made tolerable, and that current x-ray–driven pellets designed for use in the NIF are likely to reach ignition yields.

In spite of the advantages of the indirect-drive approach, direct-drive research is also being pursued vigorously. If direct-drive ICF can be demonstrated, then it might permit ignition and high gain with lower-energy (smaller) drivers, since direct drive does not have energy losses associated with the conversion to x rays and other hohlraum losses. Direct-drive pellets would be simpler in design and cheaper, again since they do not require a carefully crafted hohlraum. Therefore, direct-drive targets are more desirable for energy reactor applications. Since either approach may yet encounter formidable, but presently unknown, problems as they scale up into new parameter regimes, the current ICF program strategy is to pursue both vigorously until it is clear which is superior.

3. CURRENT STATUS OF ICF RESEARCH

3.1 Laser Facilities

Two new ICF-laser facilities designed to pursue the direct-drive ICF approach, under construction in 1993, have been completed and put into operation. The Omega glass laser, at the University of Rochester's Laboratory for Laser Energetics, underwent a complete upgrade to deliver over 40 kJ of UV (0.35 μm wavelength) in each nanosecond-duration pulse in sixty uniform, balanced, and flexible beams. Figure 3 shows a photo of the Omega Upgrade target chamber. Although Omega Upgrade was designed primarily to do integrated direct-drive pellet implosion research, it is also being used to investigate aspects of x-ray–drive physics in preparation for ignition experiments on the NIF. An overview of the Omega Upgrade experimental program may be found in Soures *et al.* (1995). The other new laser facility, Nike, is a relatively small (4 kJ) krypton-fluoride gas laser built at the Naval Research Laboratory. Nike was designed with excellent beam-smoothing features to allow precision planar-geometry hydrodynamic-instability experiments that may help determine the upper yield limit of direct-drive pellet implosions.

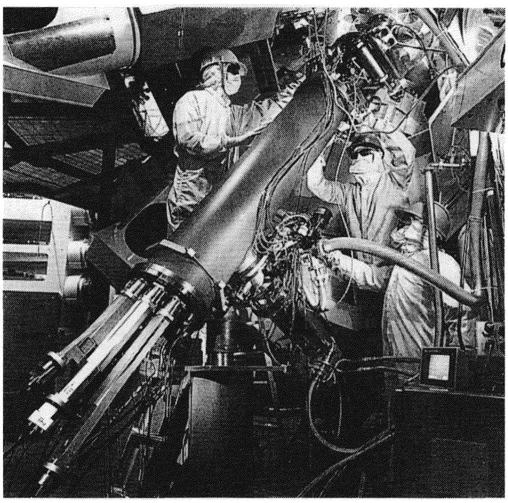

FIG. 3. Omega Upgrade target facility at the Laboratory for Laser Energetics, University of Rochester [From University of Rochester (1996).]

3.2 Target Experiments

The U.S. ICF program has made significant progress in reducing uncertainties in most technical areas required for ignition. Experiments on the Nova laser at Lawrence Livermore National Laboratory (LLNL) have provided a better physical understanding of ignition physics, especially those that may affect target performance with the NIF. The status of some of those investigations is provided in the following.

Currently, gas-filled hohlraums are the baseline targets for use on NIF. The purpose of the gas is to impede the flow of plasma from the hohlraum wall into its volume, thereby reducing the environment for parametric instabilities of the incoming beams. Early Nova experiments indicated levels of laser light backscattered from instabilities and dynamics shifts in beam pointing inside hohlraums large enough to affect adversely the x-ray drive effectiveness. It was conjectured that these effects resulted from high-intensity peaks in the fairly nonuniform Nova beams. Beam-smoothing techniques, such as use of phase plates, and an active beam-fluctuation averaging technique, such as smoothing by spectral dispersion (SSD), were applied to Nova's beams in 1996. Experiments and modeling then confirmed that use of smooth beams largely eliminated these potential problems and brought about a corresponding rise in x-ray radiation temperature and symmetry. These favorable results scale to expected NIF conditions according to current theory and models.

The Omega laser at the University of Rochester has also been used to do precision hohlraum x-ray–drive experiments. Excellent agreement was obtained on Omega with equivalent Nova experiments and LASNEX calculations (Decker *et al.*, 1997). Omega has also been used to test beam configurations planned for the NIF facility. About half of Omega's 60 beams have been arranged into a similar geometry to the NIF point-target design, i.e., two outer and two inner rings of beam irradiation spots inside the hohlraum. Experiments to exercise precise temporal and spatial control over the symmetry of x-ray emission by use of beam delays and pulse shapes were performed. Again good agreement between simulation and experiments was found, showing that similar beam phasing on NIF should provide adequate drive symmetry for ignition implosions.

3.3 Predictive Capability

A key component of successful extrapolation from Nova to NIF, and from NIF to a high-yield facility, is possession of reliable predictability through an accurate computer simulation capability. Most of the experimental results are used either to validate predictive codes, or to help add or correct the physics models required to do so. Increasingly, it has been found that two-dimensional (2D) codes are not sufficient to model implosion dynamics and that three-dimensional (3D) simulations are needed. In the mid-1990s, a radiation-hydrodynamics code, HYDRA, which includes full fusion burn physics, was developed. HYDRA is now routinely used for modeling ICF capsules. In particular, the surface-finish smoothness requirements for the inner surface of the deuterium-tritium ice layer in NIF ignition targets have been studied. HYDRA showed that a DT-containing shell made with a plastic ablator can tolerate up to 1-μm rms DT layer roughness, whereas the corresponding shell made of beryllium ablator can tolerate up to three times more roughness.

3.4 NIF Ignition Target Point Design

The baseline ignition capsule (pellet) for use on the NIF has a cryogenic DT-ice shell surrounding a central DT-gas region that resides inside a gas-filled gold hohlraum, as depicted in Fig. 4. The DT ice is surrounded by either a plastic or a beryllium shell that both forms a vessel to hold the DT and acts as the ablator material. With an expected NIF input of 1.8 MJ in a 500-TW pulse shaped as shown in Fig. 4, yields of order 15 MJ are predicted. This leaves about a safety margin of close to 10 above the ignition level for nonideal performance. There are over a dozen target designs that reach ignition levels, most of which have moderate robustness for nonideal behavior. As mentioned previously, the most sensitive parameter that can spoil performance is the surface finish of the inner DT-ice surface, which affects the level of hydrodynamic instability causing mix in the hot-spot region. Since significantly higher DT-ice roughness is tolerated with beryllium ablators than with plastic ablators, considerable research is being

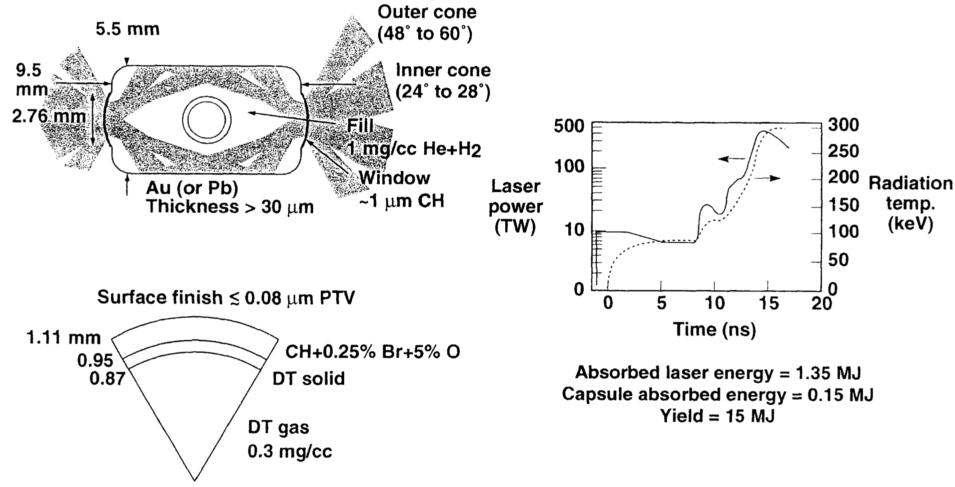
5.5 mm
Outer cone (48° to 60°)
9.5 mm
Inner cone (24° to 28°)
2.76 mm
Fill 1 mg/cc He+H2
Window ~1 μm CH
Au (or Pb) Thickness > 30 μm
Surface finish ≤ 0.08 μm PTV
1.11 mm
0.95
0.87
CH+0.25% Br+5% O
DT solid
DT gas 0.3 mg/cc
Laser power (TW)
500
100
10
0
0
5
10
15
20
Time (ns)
300
250
200
150
100
50
0
Radiation temp. (keV)
Absorbed laser energy = 1.35 MJ
Capsule absorbed energy = 0.15 MJ
Yield = 15 MJ

FIG. 4. Baseline ignition target design for the NIF (Courtesy of LLNL).

performed on fabricating suitable Be shells. A more complete description of ignition target design can be found in Haan *et al.* (1995).

3.5 Target Fabrication

Plastic shells are the mainline target for NIF ignition because their fabrication and characterization has been demonstrated. The relaxed specifications on surface finishes of cryogenic DT layers in Be targets have motivated efforts to develop the means to fabricate Be shells with the proper uniformity. Both machining and sputter-coating methods are being evaluated for this purpose. Although good progress has been achieved, the fabrication of beryllium shells with both sufficient uniformity and a means to load them with DT fuel needed further development as of 1998. It is expected that beryllium will indeed become the *de facto* ablator material, with plastic kept as the fallback option.

4. NATIONAL IGNITION FACILITY

Among the facility proposals under consideration in the early 1990s was a "Nova Upgrade," the purpose of which was to increase the existing 10-beam Nova laser's energy to a level that could approach pellet ignition. Subsequently, a very large glass laser facility of this type, called the National Ignition Facility (NIF), was designed and approved for construction in the United States. Figure 5 shows a sketch of the NIF as it is expected to look when completed. The new name reflects both the fact that the NIF is an entirely new facility and that it is being viewed as a national resource for use by all research participants in the U.S., and perhaps worldwide. The NIF is viewed as a key element of the U.S. stockpile stewardship management program. An ignition facility very similar to the NIF is also planned for construction in France.

29 May 1997 was the groundbreaking for the National Ignition Facility at Lawrence Livermore National Laboratory. When operational in the 2005 time frame, NIF will be the world's largest laser, about twenty times more powerful than Nova, the LLNL laser it is replacing. The heart of the system is an Nd laser with 192 beams that will produce in excess of 1.8×10^6 J of ultraviolet (0.35 μm) laser light energy in 3 ns (about 500 TW).

The NIF was designed to be very flexible, incorporating beams with excellent uniformity, interbeam balance, and pulse shaping. It was also constructed to allow relatively easy reconfiguration into direct-drive geometry facility. The size (energy) of the NIF was chosen to

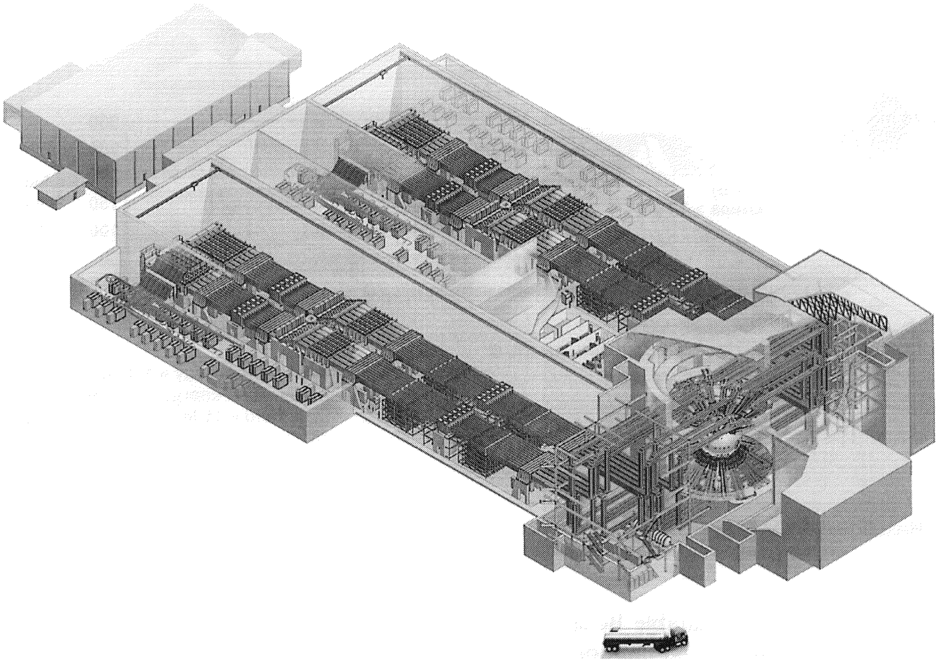

FIG. 5. The National Ignition Facility (artist's sketch courtesy of LLNL).

have a likelihood to achieve ignition level yields, yet small enough to contain costs. The completed facility is expected to cost about 1.2×10^9 (1997) dollars, including the laser, associated experimental and target fabrication facilities, and infrastructure. The NIF will consist of about 8000 large (over 1 ft^2) optical components and more than 30 000 smaller ones. The operational lifetime of the NIF is expected to be about 15 years.

Although located at LLNL, the NIF was designed, built, and operated as a national facility through a partnership with Los Alamos National Laboratory, Sandia National Laboratories, and the University of Rochester Laboratory for Laser Energetics, in addition to LLNL. General Atomics, Inc., is a partner in developing the target fabrication technology to be used in NIF experiments.

The NIF is scheduled to be completed in several phases, with the first set of eight beam lines ready for testing in about 2001. Beamlet, a prototype of one NIF beam, began operating at LLNL in 1994. The Beamlet prototype was necessary to validate the laser architecture used in NIF, which is much different from previous-generation lasers.

The light from NIF's beams will be directed into a 10-m-diam spherical target chamber and focused onto the relatively small (millimeter) targets. The target chamber will be loaded with sophisticated diagnostic instruments to measure the performance of the pellet performance. The entire NIF facility is being designed to accept either indirect-drive geometry, in which the beams enter the chamber primarily in clusters on the top and bottom of the target chamber, or direct-drive geometry, in which the beams enter the chamber more symmetrically. Although the indirect-drive configuration will be implemented first, the system will be reconfigured should direct-drive experiments at Omega Upgrade prove more promising.

4.1 Science-Based Stockpile Stewardship

Establishment of the science-based stockpile stewardship (SBSS) program, a subset of the comprehensive stockpile stewardship maintenance program (SSMP), was a consequence of the cessation of the cold war and President Clinton's order for a moratorium on U.S.

nuclear weapon testing in 1993. (Subsequently, in 1996, the U.S. and over 90 other nations signed a Comprehensive Test Ban Treaty.) The SBSS represents DOE's strategy to address the challenge of retaining nuclear capability in a test-free environment. The 1994 National Defense Authorization Act called for "a stewardship program to ensure the preservation of the core intellectual and technical competencies of the United States in nuclear weapons." The SSMP is directed to do this in a manner consistent with the broad arms-control and nonproliferation goals of the United States and most other nations.

The SBSS is structured to maintain a continuing capability to anticipate, detect, and correct problems related to aging of the nuclear stockpile. In addition to accumulating detailed knowledge about nuclear technologies, understanding of high–energy-density physics, reliable computer simulation capability, and the like, it requires that there continue to be a number of excellent scientists and engineers trained and knowledgeable in weapons related areas.

The role that ICF, and NIF in particular, play in the SBSS program is to explore the physics of matter and radiation at high energy density and the physics of thermonuclear fusion in a laboratory setting. Ignition of an ICF pellet in NIF is a major goal. The goals of SBSS and NIF are also fully consistent with the goals of both the defense and energy ICF applications. Many people believe that it was only the confluence of favorable events, such as of the end of the cold war, test bans, and nonproliferation policy, concurrent with readiness of glass-laser technology and advanced and positive ICF research results, that allowed the NIF to be authorized in the tight federal fiscal environment of the mid-1990s.

5. HEAVY-ION DRIVER DEVELOPMENT

Although the most advanced and utilized drivers for ICF pellet implosion research to date are lasers, there has been some progress in developing particle-beam accelerators for ICF driver use. Particle-beam driver concepts currently assume use of the x-ray drive approach (although direct-drive schemes are also possible in principle). Particle-beam energy is first converted in a hohlraum into x rays that, in turn, irradiate and drive pellet implosions. The technical feasibility of the development of a heavy-ion induction linear accelerator continues at Lawrence Berkeley Laboratory, but at a slow pace due to low funding levels ($\$7 \times 10^6$ level in FY97 out of a total ICF funding of $\$250 \times 10^6$). The reader is directed to the original article for further details and references that describe several heavy-ion driver approaches.

6. *Z*-PINCH DRIVER

There has been a program, from the early days of the ICF program, to determine if electric-pulse–power devices could be used as a driver. Pulse-power devices store large amounts of electrical energy in capacitor banks or inductive units and use switching elements to discharge the electrical energy rapidly into a load. Discharge-pulse parameters are in the tens of nanoseconds time scale with voltages in the multimegavolt range and with multimegampere currents.

In the 1970s, the approach used was to generate an electron beam in the load, frequently called a diode, to drive the pellet implosion. The electron-beam approach was dropped when the multi-MeV electrons from the driver beam were found to penetrate and preheat the pellet interiors to unacceptable levels. In the 1980s and early 1990s, the diodes were configured to have reverse polarity, and efforts were shifted to accelerating light ions (hydrogen or lithium) for pellet driver application. After encouraging initial results, light-ion driver ICF development stagnated for a number of years because no way was found to focus the beam to sufficiently high energy densities. This resulted in hohlraum temperatures much too low to be useful. The light-ion–beam accelerator approach was essentially abandoned in the mid-1990s.

In 1996, Sandia National Laboratory made a major advance in the generation of x rays with a different pulse-power approach. They rejuvenated and refined an old pulse-power device, called a *Z*-pinch, to generate x rays directly, without use of a particle beam inter-

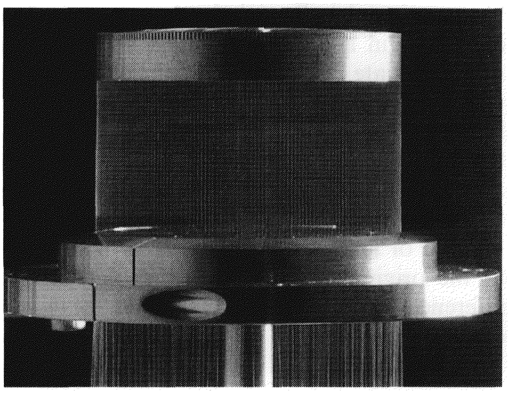

FIG. 6. Wire array load for Sandia *Z*-pinch x-ray source. The diameter of the cylindrical array of 300 tungsten wires is about 2 cm. The pulse-power voltage is applied across the top and bottom support rings. (Courtesy of Sandia National Laboratory.)

mediary. This *Z*-pinch consists of a wire array load that gets vaporized, is made into a plasma, and is pinched to high density and temperature by the magnetic field from the huge currents flowing through it. In the process, the pinched plasma emits copious x rays. Sandia produced a record 1.9 MJ of x-ray energy in a 200-TW pulse with the 300-wire array load pictured in Fig. 6. Using this x-ray source, they produced a 130-eV black body temperature inside of a 2.5-cm-diam hohlraum in 1997. While the hohlraum temperatures demonstrated are still at least a factor 2 to 3 too low to drive ignition-level pellets, the results are viewed as encouraging. Sandia proposed to build a larger version of this device, called X-1, that might be used for ignition-level experiments in the late 1990's. A *Z*-pinch ICF driver is best suited for defense applications that require single-shot capability (e.g., one shot per day or less), since the *Z*-pinch load region is destroyed with each firing. It is difficult to imagine a *Z*-pinch being compatible with the high repetition rate needed for a civilian power plant.

7. ALTERNATE CONCEPT: FAST IGNITOR

If inefficiencies and waste heat could be eliminated, high-gain fusion yield could be achieved with as little as 100 kJ energy input. This is approximately the energy required to compress low-temperature main pellet fuel to the high densities required to support propagating fusion burn. What is then needed, in principle, is only a few tens of kilojoules delivered to the central small ignitor region that raises its temperature enough to initiate propagating fusion burn. In conventional approaches (direct and x-ray drive) ignition is achieved by carefully programming in a set of shocks converging though the cold, dense main fuel region that coalesce at the right instant to heat and ignite the central region. This is a rather inefficient process, as only a small fraction of the shock energy transfers to the ignitor, even if done optimally.

The fast-ignitor concept would employ an ultrahigh-intensity laser to burrow through the corona surrounding the compressed fuel and deliver energy directly to an ignitor region in the form of energetic electrons, as depicted in Fig. 7. By way of an example, to heat a DT ignitor region of density 300 g/cm^3 to about 10-kV temperature using energetic electrons requires focal intensities of at least 10^{20} W/cm^2 with 1-μm light in a pulse duration of 1 ps (i.e., 10 kJ). Any laser–matter inefficiencies will increase the associated laser requirements. A laser prepulse of order 10^{19} W/cm^3 is needed over a few tens of picoseconds to create the immense pressures required to push the cold fuel aside and form a low-density channel through which the much more intense ignition pulse can be delivered.

The fast-ignitor concept is considered to be very speculative approach, since it requires specific (favorable) physics behavior through large extrapolations in unexplored, highly nonlinear portions of parameter space. Nonetheless, if it can be demonstrated to work, the costs to achieve high-yield ICF and the engineering requirements for the reactor chamber plummet and the variety of driver options decrease.

A petawatt laser (10^{15} W) has been developed at LLNL to test some of the physics basis of the fast-ignitor concept. The petawatt laser produces about 1 kJ of 1.053-μm light at irradiances above 10^{21} W/cm^2 with a pulse length in the 1-ps regime. The laser starts with short chirped pulses (broad bandwidth) that are stretched to about 1-ns duration with a dif-

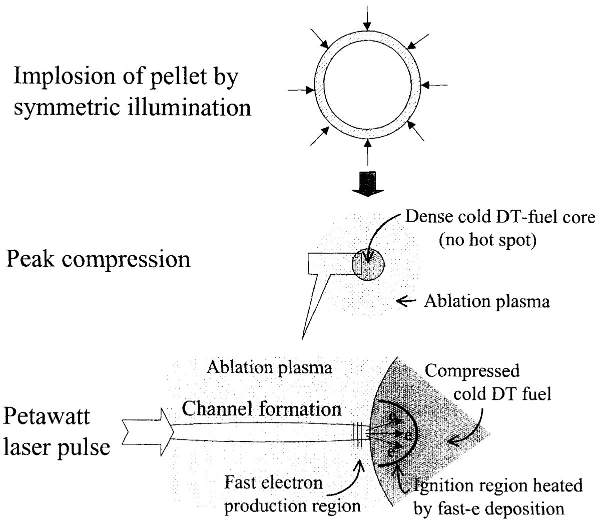
Implosion of pellet by symmetric illumination
Peak compression
Dense cold DT-fuel core (no hot spot)
Ablation plasma
Ablation plasma
Compressed cold DT fuel
Petawatt laser pulse
Channel formation
e
e
e
Fast electron production region
Ignition region heated by fast-e deposition

FIG. 7. Fast-ignitor concept. The pellet containing DT fuel is compressed without forming a hot spot. A picosecond laser pulse bores a hole through the dense plasma leading up to the compressed core and creates an intense burst of energetic (fast) electrons near the core that heat a small region to ignition.

fraction grating. The stretched pulse length is well matched for efficient amplification using a slightly modified Nova amplifier chain, followed by recompression of the pulse duration down to the picosecond regime using large recompression diffraction gratings. Proof-of-principle experiments of the critical fast-ignitor physics issues began in 1997 using the petawatt laser. A more complete description of the fast ignitor concept may be found in Tabak (1994).

8. INERTIAL FUSION ENERGY

Efforts directed specifically to produce a civilian energy source based on ICF, called inertial fusion energy (IFE), continue, but constitute a minor portion of the total ICF program. However, the essential stage of demonstrating ICF ignition in the laboratory is common to both energy and defense applications; hence the ICF program resources have been concentrated on this goal. A critical requirement for civilian energy use of ICF is a driver that is capable of being cycled with a repetition rate of about 1 shot/second with good efficiency and reliability, long lifetime, and compatibility with a reactor environment. Heavy-ion accelerators appear to be capable of meeting these requirements. Two laser technologies that might potentially satisfy the energy driver requirements, high–repetition-rate KrF-gas laser and laser-diode–pumped glass lasers, are also being investigated at low levels.

9. PROGNOSIS

The inertial-confinement fusion program has progressed to the point where the scientific basis and technology for the next major step is in place, that of demonstrating ICF ignition in the laboratory. When complete in about 2005, the National Ignition Facility will have this capability. Although the primary impetus for the program is defense related, this phase is also necessary for the development of inertial fusion as an electrical power source. Obviously, the next decade will be an exciting and telling period.

List of Works Cited

Decker, C. *et al.* (1997), "Hohlraum radiation drive measurements on the Omega laser," *Phys. Rev. Lett.* **79**, 1491–1494.

Haan, S. *et al.* (1995), "Design and modeling of ignition targets for the National Ignition Facility," *Phys. Plasmas* **2**, 2480–2087.

Lindl, John (1995), "Development of the indirect-drive approach to inertial confinement fusion and the target physics basis for ignition and gain," *Phys. Plasmas* **2**, 3933–4024.

LLNL (1995), *ICF Quart. Rep.* **5** (July–September).

Soures, J. *et al.* (1995), "Direct-drive laser fusion experiments with the OMEGA, 60-beam, >40 kJ, ultraviolet laser," *Phys. Plasmas* **3**, 2108–2112.

Tabak, Max *et al.* (1994), "Ignition and high gain with ultrapowerful lasers," *Phys. Plasmas* **1**, 1626–1634.

University of Rochester (1996), *LLE Rev.* **68** (July–September).

Further Reading

Berk, Herbert (1998), "Fusion, Magnetic Confinement," in: G. L. Trigg (Ed.), *Encyclopedia of Applied Physics*, Vol. 6, New York: VCH, pp. 575–607. An up-to-date review of magnetic confinement fusion research, the principle alternative approach to ICF.

Kruer, William (1988), *The Physics of Laser Plasma Interactions*, Reading, MA: Addison-Wesley. An excellent and physically intuitive description of the interaction physics between plasma and intense laser light.

Lindl, John (1995), "Development of the indirect-drive approach to inertial confinement fusion and the target physics basis for ignition and gain," *Phys. Plasmas* **2**, 3933–4024. A comprehensive technical review of the indirect-drive ICF approach.

Ripin, Barrett H. (1993), "Fusion, Inertial Confinement," in: G. L. Trigg (Ed.), *Encyclopedia of Applied Physics*, Vol. 6, New York: VCH, pp. 545–573. A summary of the basic physical principles of inertial confinement fusion.

Soures, J. *et al.* (1995), "Direct-drive laser fusion experiments with the OMEGA, 60-beam, >40 kJ, ultraviolet laser," *Phys. Plasmas* **3**, 2108–2112. A description of the current status of the direct-drive ICF research approach.

Tabak, Max *et al.* (1994), "Ignition and high gain with ultrapowerful lasers," *Phys. Plasmas* **1**, 1626–1634. A good general description of the fast-ignitor concept.

FUSION, MAGNETIC CONFINEMENT

(An Addendum to *Encyclopedia of Applied Physics*, Volume 6, pages 575–607.)

H. L. Berk, *Institute for Fusion Studies, The University of Texas at Austin, Austin, Texas, U.S.A.*

Introduction 433
1. **Power-Producing Tokamak Experiment** 433
1.1 Magnetic Field 433
1.2 Central Solenoid 433
1.3 Fusion Burn 434
1.4 Divertor 434
2. **Operation of a Burning Plasma** 434
2.1 Lawson Condition 434
2.2 Magnetic Field Limits 435
2.3 Empirical Scaling 435
2.4 Heat Load 436
2.5 Pumping 436
2.6 Ignition Physics 436
3. **Plasma Confinement Issues** 437
3.1 *H*-mode 437
3.2 Transport Barriers 437
3.3 Transport Theory 438
3.4 MHD Considerations 438
3.5 Reliability 439
4. **Alternative Approaches** 439
4.1 Alternative Tokamak Approaches 439
4.2 Other Concepts 440
4.3 Final Statement 441
Glossary 441
Works Cited 442
Further Reading 442

INTRODUCTION

In the 1990s there has been significant progress in the development of the tokamak concept and serious study by an international research group [the International Thermonuclear Experimental Reactor (ITER) team] on the feasibility and design of an experiment to demonstrate fusion power production. In this Addendum we present an overview of the status of tokamak research in relation to achieving this goal.

1. POWER-PRODUCING TOKAMAK EXPERIMENT

1.1 Magnetic Field

Recall that a tokamak is a toroidally shaped plasma in a magnetic field as shown in the schematic diagram of Fig. 1. Its cross section is shown in Fig. 2, with dimensions envisioned for a 1500-MW power-producing experiment. The dominant magnetic field is in the toroidal direction and is established by toroidal field (TF) coils that poloidally loop the plasma. A smaller, poloidally directed magnetic field arises from toroidally directed current flowing in the plasma itself and toroidal currents flowing in poloidal field coils that are outside and parallel to the plasma (currents external to the plasma control the shape and position of the plasma). All these currents produce magnetic field lines that circulate in the toroidal and poloidal directions and form magnetic flux surfaces. The q value (a parameter that will be frequently referred to) of a surface is the number of times a field line circulates in the toroidal direction in one poloidal circuit. The surfaces are represented by closed surfaces in Fig. 2.

1.2 Central Solenoid

The central solenoid is used to supply the toroidal voltage that sustains the plasma current. The voltage arises from the time rate of change of the entrained magnetic flux in the coil of the central solenoid. However, without other current drive methods being imposed, the establishment of the plasma equilibrium and the sustainment time of the discharge are limited by the time integral of the emf that is obtained by reversing this magnetic flux. This time is expected to be about 1000 s in a fusion-producing plasma.

Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

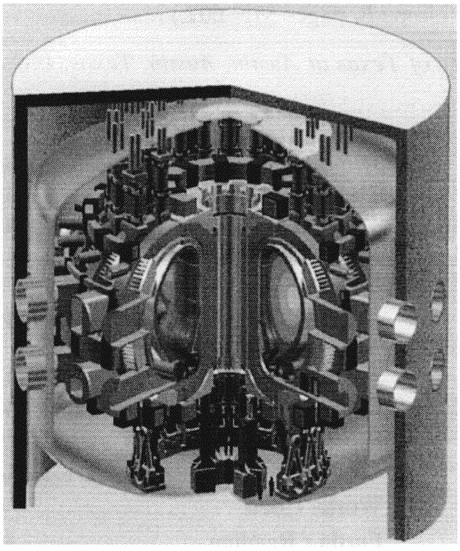

FIG. 1. An artist's conception of a tokamak reactor, with cutout showing plasma discharge.

1.3 Fusion Burn

Charged particles are trapped in continuous spirals that closely follow the entrapped field lines. This containment allows the particles to be heated to temperatures where fusion reactions occur. For a deuterium-tritium mixture, this temperature is in the range 10 to 25 keV where we use temperature in units of energy (where $1\,\text{eV} = 1.16 \times 10^4\,\text{K}$). When deuterium and tritium fuse ($d + t \Rightarrow {}^4\text{He} + n + 17.6\,\text{MeV}$), the 14-MeV neutrons leave the plasma and are absorbed by the walls. Their energy is to be used to produce steam that drives an electrical power plant. In addition the neutrons are needed to breed tritium (tritium is radioactive with a half-life of 12 years and is available only in limited amounts from fission reactors). Hence it will be necessary to surround a fusion reactor with a lithium blanket (see FUSION TECHNOLOGIES) to breed tritium ($n + {}^6\text{Li} \Rightarrow t + {}^4\text{He} + 4.8\,\text{MeV}$; $n +$

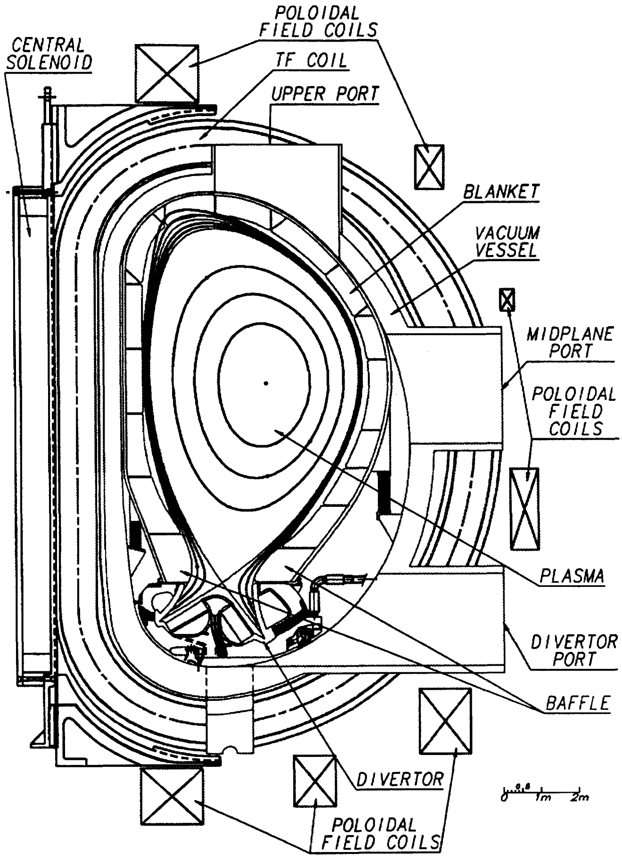
CENTRAL SOLENOID
POLOIDAL FIELD COILS
TF COIL
UPPER PORT
BLANKET
VACUUM VESSEL
MIDPLANE PORT
POLOIDAL FIELD COILS
PLASMA
DIVERTOR PORT
BAFFLE
DIVERTOR
POLOIDAL FIELD COILS
0
1m
2m

FIG. 2. Cross-section of a typical design of a tokamak for a fusion-producing experiment which has been produced by the ITER team.

$^7\text{Li} \Rightarrow t + {}^4\text{He} + n' - 2.5\,\text{MeV}$). The 3.5-MeV charged alpha particles (^{4}He) should be retained in the tokamak and their energy transferred to the background plasma by collisions. Then, even without external heating, a "burn" is established, whereby the plasma fuel remains hot while losing energy by bremsstrahlung, atomic radiation, and convection of heat to the "cold" solid walls. Ignition occurs when the fusion-produced alpha particles supply the required heating instead of an external source.

1.4 Divertor

Figure 2 also illustrates a single divertor configuration where there is a special magnetic flux surface called the separatrix (the one where the poloidal field lines appear to cross). Field lines beyond the separatrix (the scrape-off layer) impinge on structure called divertor plates designed to handle large escaping heat loads (charged particles that diffuse outside the separatrix strike here). Gas formed from the recombined plasma is to be pumped away from the surrounding divertor chamber into a closed tritium-recovery system.

2. OPERATION OF A BURNING PLASMA

2.1 Lawson Condition

In order for fusion reactions to be thus sustained, the confinement properties at fusion temperatures must satisfy the Lawson criterion, $n\tau_E \approx 2 \times 10^{20}\,\text{m}^3\,\text{s}$ (here n is the d-t ion density and τ_E the overall energy lifetime). The energy lifetime is determined by transport arising from particle collisions and plasma-induced short-wavelength turbulence. At fusion temperatures, the density is limited by magnetohydrodydamic (MHD) stability. This sets a limit on the parameter β, the ratio of the average pressure (the particle density times the temperature) to magnetic field pressure. In standard tokamak operation, where the current density is largest in the plasma's center, there is a further MHD limitation that prevents the q value (which is inversely proportional to the current density) from being too much less than unity. In the standard case the β is limited to about 5%. In addition there exists a less well understood empirical density limitation, perhaps associated with charged-particle recombination, known as the Greenwald limit, which seems to limit the maximum density to a level proportional to the mean current density of the discharge.

2.2 Magnetic Field Limits

Another limitation is set by the magnitude of the magnetic field that can be imposed. With pulsed operation ($\sim 10\,\text{s}$), TF coils with about a 20-T peak field (producing a 13-T field at the plasma center) have been designed. However, superconducting magnets are generally considered to be needed for a fusion experiment designed to operate in quasisteady conditions of order 1000 s. Then peak fields are limited to 12 T, which allows a 6-T field at the plasma center.

2.3 Empirical Scaling

The energy lifetime of a typical tokamak discharge is uncertain as it is primarily governed by incompletely understood turbulent processes. Thus the extrapolation of tokamak parameters to larger-size reactor regimes is typically based on empirical scaling to the experimental database, obtained from many different installations. A reasonable fit to many experiments is achieved by assuming that the energy lifetime is proportional to algebraic powers of the physical parameters defining the experiment (such as plasma current, magnetic field, plasma heating rate, etc.), with the value of each power the same for all the experiments. A general conclusion from this procedure is that confinement improves with increasing plasma current but degrades with increasing heating power. When this law is applied to an ignition experiment, the extrapolated lifetime is compatible with achieving ignition, but with some uncertainty due to the variance in the data.

2.4 Heat Load

The dimensions shown in Fig. 2 arise as a result of conforming to the constraints mentioned above. They lead to a total fusion-power production in the gigawatt regime (Post and Uckan, 1992). One fifth of this power (from alpha particles) directly heats the plasma. The heat exhausts on plasma-facing material with a rather modest average power load of about 0.2 MW m^{-2}. However, should most of this heat impinge on the divertor plate, an unacceptable 40-MW m^{-2} heat load would arise.

Operation with a divertor is viewed as essential. It has been found to be extremely important to control the impurities that enter the plasma, to design efficient pumping methods, and to establish the conditions for optimum plasma confinement in the so-called *H*-mode. The *H*-mode of operation improves the energy confinement time a factor of 2 or more compared to an alternative mode of operation, the so-called *L*-mode (see below for further discussion).

A fix to this heat exhaust problem is to radiate away most of the exhaust energy. For example, if 1500 GW of fusion power is being produced, then the plasma is directly heated by the 300 MW in alpha particle production. Roughly 100 MW radiates as bremsstrahlung. Impurity seeding at the plasma edge and diverter throats is required to radiate away about 150 MW so that the remaining 50 MW of heat flux in the escaping plasma allows a tolerable power-density load of less than 10 MW m^{-2} at the divertor plates. Favorable results are now available in several tokamak experiments that demonstrate that the heat load on the divertor plates is indeed reduced by such impurity radiation shielding, in a so-called semidetached mode of operation, while maintaining relatively good purity of the plasma. An important associated issue is whether optimal *H*-mode confinement properties can be achieved in semidetached operation. Most experiments observe confinement degradation, but there have been results that demonstrate, in semidetached operation, confinement properties close to that in typical *H*-mode operation.

2.5 Pumping

Pumping in the divertor region is extremely important in reactor operation. The helium accumulating in the core must be pumped away. Otherwise the burn quenches after ignition is achieved, as helium ash is replacing plasma fuel. Experiments have shown that the ratio of helium to energy lifetime is about 5, and with pumping and *d-t* gas puffing the helium accumulation can be limited to 10% of the plasma density, an acceptable level. The pumping in the divertor system is a principal control of the plasma impurities as well. The density of seeded impurities needs to be peaked near the scrape-off layer but diffuse in the plasma bulk. Experiments have demonstrated that this can be achieved.

2.6 Ignition Physics

There are several intrinsic physics issues associated with achieving ignition and burn. One is the control of the parameters of the plasma. There can be an ignition-induced instability, where once the fusion burn dominates, the plasma heating rapidly increases the plasma pressure. The increased pressure can cause the system to enter unstable MHD regimes where confinement deteriorates or is even lost. The conditions for such excursions are believed to be understood and in principle can be avoided. Ultimately the appropriate operation must be experimentally demonstrated.

Another burn issue is the physics associated with alpha-particle confinement. To achieve burn the alpha kinetic energy must be absorbed by the plasma. It is possible for energetic alpha particles to diffuse to the plasma edge and then embed themselves into the plasma-facing material. It is estimated that a limit of only 5% of the generated energetic alpha particles can be lost in this way if serious wall deterioration is to be averted.

Energetic alpha particles near the edge are primarily lost as a result of complicated motion arising from the presence of toroidal magnetic field ripple resulting from the spacing of a finite number of TF coils in the machine. It has been shown that if only the energetic alpha particles born near the edge are lost, there is a tolerable loss with a reasonable number of TF coils. However, there is still a concern that alpha particles can be fed into the edge region from plasma instabilities. One possibility is that Alfvén-wave instabilities can be spontaneously generated with the alpha particles themselves the instability source. Another possibility is that alpha particles are lost through MHD activity of the plasma, which either directly drives the alpha particles into the loss region or indirectly triggers the Alfvén instability by causing a rearrangement of the alpha-particle pressure profile. Such scenarios, with loss, have been observed in experiments that use energetic particles arising from neutral-beam injection or ion-cyclotron rf heating. In the TFTR tokamak, where up to 10 MW of fusion power has been produced, the alpha-particle–driven Alfvén wave has been observed, but in these cases the observed saturation level is too low to cause significant energetic particle loss. Theory predictions indicate that an alpha-particle–driven Alfvén instability is unlikely to arise for many operating parameters, but it may prevent some otherwise desirable set of plasma parameters from being achieved.

3. PLASMA CONFINEMENT ISSUES

3.1 *H*-mode

Clearly energy-confinement time is a critical parameter in determining the feasibility and design of fusion-power production. The plasma properties can be significantly altered if turbulence conditions are changed. This appears to be the case when there is a transition from *L*-mode to *H*-mode operation. In *H*-mode a pedestal in the temperature appears at the edge of the plasma with a radial thickness typically of an ion orbit width. In this layer there is a flowing plasma with radial shear. A widely accepted explanation of why *H*-mode is established (Itoh and Itoh, 1996), is that the escaping plasma causes a torque to be generated that drives a shearing flow. The shear in turn suppresses fluctuations in the pedestal region by either "ripping" apart the coherence of the fluctuations so that they saturate at a low level, or preventing unstable fluctuations from appearing. It is experimentally observed that a threshold power level is needed to establish an *H*-mode, and this observation is consistent with the hypothesis for the cause of shear flow. The determination of this level is still uncertain; there is no generally accepted quantitative theory, and empirical studies have too much scatter to predict whether, in the reactor operation described above, there is enough power flowing through the edge to establish an *H*-mode. This is an important issue that is to be resolved in ongoing studies.

3.2 Transport Barriers

The studies of the *H*-mode have been crucial in the discovery of new ways to improve energy confinement (Burrell, 1997). It has been found in many tokamaks that by applying torques in either narrow or wide regions of the plasma, by use of either neutral beams or rf sources, transport barriers can be established in the core of the plasma. High temperatures have been established in the central region, with a large drop in the temperature taking place in the transport barrier. It has been further found that establishing this barrier is aided by having a hollow current profile, which leads to a reversed q profile that has a local maximum at the plasma center and a minimum q value, typically ~ 2, further out radially. The thermal barrier is generally established near the minimum-q position. Stability theory indicates that in the region where q is radially decreasing, the strength of the diamagnetic-drift– and magnetic-curvature–driven instabilities (see FUSION, MAGNETIC CONFINEMENT) are reduced. Then the flow shear that is established is enough to stabilize the modes completely. Some experiments report that the residual transport can be explained by transport due only to collisional effects.

The major issue with modes that have hollow current profiles is that they are transient, unless a tailored emf can be applied from auxiliary-current drive methods involving rf or neutral beam sources. It needs to be demonstrated that power requirements for such a drive are compatible with reactor operation.

Another benefit from the hollow current profiles is that they lead to plasmas where a large fraction of the current is in the form of bootstrap current (see FUSION, MAGNETIC CONNEMENT), which does not decay in absence of voltage. If this can indeed be established, then true steady-state operation in a reactor with non-Ohmic current-drive power is feasible as it leads to acceptable power requirements for driving the nonbootstrap current component. Steady-state operation is extremely desirable. It avoids what may be delicate plasma control issues associated with startup and shutdown, as well as thermal stresses associated with repetitive operation that lead to material deterioration.

3.3 Transport Theory

There has been considerable development in theory to achieve transport predictions based on plasma-physics principles. The theories used are still in a post-processing phase, i.e., they explain the temperatures achieved in a large number of experiments but have not predicted the temperature of any experiment before it was performed. One reason is that a theory for energy transport is easier to develop than a theory for particle transport. Hence to determine temperature, a density profile has to be taken from experiment. In addition the transport processes in the core of the plasma are thought to differ from the processes at the edge, and typically the various predictive methods make allowances for less understood edge transport processes either by using experimentally observed temperatures at the edge or by adjusting edge transport coefficients to match empirical observation. As a result many theory-based transport calculations produce good correlation of the predicted and observed energy lifetimes for a variety of experiments (Connor *et al.* 1997). However, when the theories are extrapolated to reactor-type plasmas, significant deviations are predicted. One theory predicts that gigawatt fusion power operation can be expected only if a high edge temperature, 4 keV at the inner edge of the pedestal, can be achieved, while another theory indicates that such fusion power levels can be achieved even with a modest edge temperature. In such a case the former theory would predict a fusion power output that is about a factor of 2 larger than the external heating input, which would give a few hundred megawatts of fusion power when about 100 MW input power is available.

The issue of attaining a predictive theory is of high priority, and tests to resolve existing discrepancies are still in progress. Nonetheless, it is clear that there has been a great deal of basic progress in the understanding of the turbulent transport problem, and we can expect the theoretical reliability of this approach to improve rapidly.

3.4 MHD Considerations

If the plasma's MHD stability characteristics can be improved it can lead to better reactor properties. At a given temperature, a greater density can be achieved, making it easier to satisfy the Lawson criterion. One of the important parameters controlling MHD stability is the plasma shape. Note the vertical elongation and the triangular shape of the outer flux surfaces in Fig. 2. The decreasing pressure profile at the pedestal is a destabilizing effect on the outer part of the surface and is a stabilizing effect on the inner part of the surface. The triangularity allows stronger weighting from the stabilizing inside than would arise without this type of shaping. Then the temperature drop at the pedestal where the H-mode formed can be increased to a limit apparently determined by MHD oscillations known as elms. As one example we note that in 1996 the best bulk confinement parameter $n\tau_E T = 1.5 \times 10^{21}\ \mathrm{m}^{-3}\ \mathrm{s\ keV}$ was achieved in the JT-60U experiment, in part as a result of shape optimization.

There are many experiments that achieve higher β values, but they frequently have MHD fluctuations that reduce the energy-confinement time or cause a disruption where the plasma discharge rapidly disintegrates. One route to a disruption is through a so-called locked mode, where a perturbed wave "attaches" to an imperfection of the stationary magnetic field, which can cause deterioration of the discharge. Locked modes are avoided with a sufficiently large plasma rotation, and hence enough rotation needs to be generated in a fusion power-producing experiment.

There is a great deal of optimism that discharges with reversed q profiles in the center can attain significantly increased β values, particularly when a close-fitting conducting wall is used. However, in this case slowly growing modes still arise on account of the finite wall resistivity. Such modes can in principle be controlled by feedback with low-power sensing circuitry. More experiments on this topic are needed, and appropriate reactor designs still need to be developed.

3.5 Reliability

The demonstration of reliability is another challenge for tokamak operation. In conventional operation, pulsations, such as sawtooth oscillations and elms, are common. Typically the sawteeth arise when the q value of the discharge is somewhat below unity and the pressure profile peaks too strongly within the $q = 1$ surface. The time interval between sawtooth pulses is typically a fraction of the overall current diffusion time (the time interval is of order 100 ms in the larger present-day machines and is expected to be about 1 s in a tokamak like that described in Fig. 2). The pulsation flattens the pressure profile within the $q = 1$ surface, but it also has a beneficial effect of preventing further steepening of the central current profile, a destabilizing tendency. These pulsations generally cause rather small perturbations at the edge so that ordinary sawteeth by themselves are not likely to induce heat-flow problems at the divertor plates. However, in some operations energetic particles induce a change of plasma behavior that causes the sawtooth oscillations to be suppressed for long time intervals. Unfortunately, at times, a giant sawtooth pulse appears that disturbs edge conditions and can even disrupt the entire plasma discharge. Such behavior must be controlled in reactor operation.

The elms, mentioned earlier, are a plasma edge perturbation that cause enhanced heat pulses to outflow at regular time intervals. There is an acceptable type of elm operation, so-called grassy elms, where the pulses are frequent but low level, which leads to heat loads that are readily handled, good global confinement characteristics, and an observed advantageous property of keeping impurities from accumulating in the main part of the plasma. However, elms arise that give larger heat pulses and less favorable plasma confinement. A further difficulty is that a large elm pulse can cause termination of H-mode operation, which will then lead to a significant reduction of the fusion output unless active counter measures to restore H-mode operation are found.

Large disruptions are known to arise in tokamaks, leading to the loss of the plasma discharge. Frequently these disruptions are uncontrolled and are more likely to arise for tokamaks that operate at the limits of acceptable operation. Of concern is that large disruptions can cause damage to plasma-facing wall components due to localized heat deposition, the formation of so-called halo currents where the plasma current penetrates the surrounding conducting wall, and the formation and subsequent wall bombardment of relativistic "runaway" electrons accelerated from the inductive electric field.

Studies are in progress to develop reliable ways that disruptions can be guided to a "soft" landing, where the heat is distributed benignly over the entire surface of plasma-facing walls (for example, from injection of so-called killer pellets, that cause the heat to radiate), and the current in the remaining cold plasma, together with its associated magnetic field energy, is allowed to dissipate on the time scale of a second, a time interval long enough so that the inductively produced electric field is too low to cause runaway electrons.

4. ALTERNATIVE APPROACHES

4.1 Alternative Tokamak Approaches

On the basis of the knowledge known today there is a high confidence level that a fusion power-producing experiment can be designed to operate, at a minimum, in a fusion-power amplification mode where several hundred megawatts of power are produced. A more challenging goal is to achieve, reliably, quasi-steady burn conditions or a high enough energy amplification level to produce gigawatt power levels. Such success requires maintaining optimal tokamak performance (such as H-mode operation, steady alpha-particle heating, limited parameter excursions from pulsations, etc.) in the standard-type discharge discussed above.

Alternative directions of research will proceed either in conjunction with, or as an alternative to, the direct implementation of a fusion power experiment. Perhaps the most exciting development in the mid-1990s is the demonstration that turbulence levels can be dramatically diminished, leading to enhanced confinement, as a result of the combination of plasma flow shear and establishing q profiles that decrease radially over much of the discharge. The future challenge in the operation of this type of discharge is to demonstrate that such conditions can be steadily maintained with modest external power control that extrapolates properly to reactor-size plasmas. The intrinsic bootstrap current in such discharges is expected to be high enough that the power needed to maintain true steady-state reactor operation can be achieved. Significantly, non-inductive current-drive experiments in the Tore-Supra and in other tokamaks have already proven that plasma discharges can be sustained indefinitely, but the needed efficiency is still to be demonstrated.

Future research will also focus on achieving higher plasma pressure operation by implementing conducting-wall stabilization with feedback to control slowly growing resistive wall instabilities. The new methods of obtaining enhanced confinement may also lead to alternative designs for plasma shaping. Other studies are examining whether plasma confinement properties can be improved by increasing a tokamak's minor-to-major radius or major-to-minor radius ratio. Preliminary arguments and data give supportive reasons why such approaches can be beneficial.

There is extensive interest in building an experiment that demonstrates ignition and a short burn, but that does not address long-time reliability issues (Coppi *et al.*, 1994). By using the highest possible magnetic fields, plasmas can be established that are well within intrinsic parameter limits. Hence in such an experiment the likelihood of achieving a proof-of-principle controlled burn is high. Future improvements in the development of large superconducting magnets to the 20-T range would have a dramatic impact on the progress of controlled magnetic fusion. Then long-time burn experiments can be performed at high fields, allowing engineering issues associated with power production to be tested.

Another alternative is to attempt to obtain net fusion-power operation when the *d-t* temperature is greater than the electron temperature. It is significant that such operation characterizes experiments that have achieved some of the highest Lawson parameters in JT-60U, JET, and TFTR and the highest fusion power (16 MW in JET). If the $n\tau_E$ parameter can be increased another factor of 3 to 4 in the former two machines, the condition of a relatively compact reactor operation would nearly be established. There is a problem with maintaining this mode, because as confinement improves the electron and ion temperatures tend to approach each other, and in the process degrade the favorable features of this type of confinement. A novel mechanism known as energy channeling has been suggested, where through the controlled stimulation of rf waves, the alpha-particle energy directly heats the *d-t* fuel rather than the alpha particles primarily heating the electrons (as generally occurs). The technical feasibility of this idea is still under study.

4.2 Other Concepts

The most developed alternative to tokamak research is the stellarator concept (see the original article), where charged particles are confined by the three-dimensional magnetic field established by currents in external coils. This concept is intrinsically steady state; some plasma confinement regimes have been shown to be comparable to similar tokamak confinement regimes, and experiments have been free of disruptions. Of principal concern is that the achieved β values in stellarators are generally lower than in tokamaks, and in most designs a large class of energetic particles are not contained in the magnetic fields. A positive recent development has been the design of so-called quasisymmetric fields, which improves confinement of energetic particles. However, this design still needs to be made compatible with the achievement of reasonably large β limits. Another advantage of a stellarator is its typical property of a radially increasing q profile. This may allow easier access into regimes demonstrating enhanced confinement. A new large stellarator in Japan, the LHD, began operation in 1998, and it will further test the prospects for achieving high-grade plasma confinement suitable to reactor operation.

Research with different types of pinched plasmas has made significant progress during the 1990s. The simplest in conception is the *z*-pinch (see the original article), where a current is directed in the *z* direction, producing a magnetic field that winds poloidally around this direction. This concept was one of the first for which the controlled-fusion problem was investigated in the 1950s. The conclusion was that the configuration is unstable. Nonetheless, in recent years there has been appreciable progress in the study of such pinches. When they are produced from thin fibers, extremely dense and hot plasmas are produced that are useful as x-ray sources and for understanding the interaction of plasmas with solids. By using many parallel strands of fiber, it is found that for short time intervals (of the order of microseconds) uniformity in the pinch is maintained. Because of the high densities and temperatures that have been achieved, proposals have now been made for demonstrating fusion power production with this concept.

Another pinch concept, the field-reversed pinch (FRC; see the original article) is also a candidate for demonstrating fusion power production. In the latter case, it is proposed to compress an FRC to high density and temperature using an imploding metal liner that produces magagauss magnetic fields. However, both types of pinch experiments are intrinsically pulsed, and it is now too premature to tackle the difficult problem of obtaining a duty cycle that would allow the consideration of commercial fusion power production.

The reversed-field pinch (RFP; see the original article) is a pinch more like a tokamak, in that it forms a toroidal plasma with toroidal and poloidal magnetic fields. In the 1990s, significant improvement of its parameters of operation and in the understanding of its physical mechanisms of turbulence have been made. This concept has also contributed to the understanding of the plasma dynamo problem of interest in astrophysics and to the understanding of how to control locked modes, which is crucial to understand in tokamaks as well as the RFP.

4.3 Final Statement

As a result of a great deal of progress, a plan for the demonstration of fusion-power production in a tokamak has been made. A detailed design for a fusion power-producing experiment has been completed, but cost considerations are likely to cause postponement of its implementation or a downscaled design. Many scientific issues remain to be clarified and other alternative paths are likely to develop. The physics issues that need to be resolved and the goal of a long-term energy source remain challenges to the technical creativity of mankind.

GLOSSARY

Alfvén Wave: A plasma wave that may be excited by alpha particles in a fusion plasma.

Burn: The self-sustained heating of a plasma arising from the fusion reaction.

Divertor: The edge region of a tokamak where field lines divert escaping plasma away from the immediate vicinity of the main discharge.

Energy Channeling: A proposed process to allow charged fusion products, with assistance from radio-frequency excitation, to heat ions rather than directly heating electrons.

Halo Current: Currents that are partially carried by the plasma but pass through the surrounding walls.

Ignition: The process whereby plasma heating transfers from being externally supplied to being internally supplied from the fusion reactions.

Magnetic-Field Ripple: The spatial modulation in the magnitude of the magnetic field due to the discrete number of toroidal field coils.

Pedestal: A region of plasma adjacent to the edge, where the plasma temperature abruptly rises.

Semidetached: A divertor plasma condition where a large fraction of the escaping plasma has recombined into neutral atoms.

Shear Flow: Flow of plasma that varies with minor radius.

Thermal Barrier: A region in a plasma where local transport properties are suppressed.

List of Works Cited

Burrell, K. (1997), *Phys. Plasmas* **4**, 1499.

Connor, J. W. *et al.* (1997), in: *Plasma Physics and Controlled Nuclear Fusion Research, 1996*, Proceedings of the 16th Conference, Montreal, Vol. 2, Vienna: International Atomic Energy Agency, pp. 935–944.

Coppi, B., the Ignitor Project Group, (1994), *J. Fusion Energy* **13**, 111.

Itoh, K., Itoh, S., (1996), *Plasma Phys. Control. Fusion* **38**, 1.

Post, D. E., Uckan, N. A., (1992), *Fusion Technol.* **21**, 1427.

Further Reading

Aymar, R., the ITER Team (1997), in: *Plasma Physics and Controlled Nuclear Fusion Research, 1996*, Proceedings of the 16th Conference, Montreal, Vol. 2, Vienna: International Atomic Energy Agency, pp. 3–17 and 737–1002.

MODULATORS AND DEMODULATORS, ELECTRICAL

(An Addendum to *Encyclopedia of Applied Physics*, Volume 10, pages 379–391.)

Kohji Hohkawa, *Faculty of Engineering, Kanagawa Institute of Technology, Atsugi Kanagawa, Japan*

Introduction 443
1. **Pulse Modulation Methods for Data-Communication Systems** ... 443
1.1 Advanced Modulation Method in Quadrature Phase-Shift Keying Scheme 443
1.1.1 Offset Quadrature Phase-Shift Keying Modulation (OQPSK) .. 444
1.1.2 Quadrature Quadrature Phase-Shift Keying Modulation (Q^2PSK) 445
1.2 Digital Pulse Amplitude Modulation (PAM-DSB, PAM-SSB, and PAM-VSB) 445
1.2.1 Modulator and Demodulator ... 445
1.2.2 Timing Extraction 446
1.2.3 Equalization Process 446
1.3 Quadrature Amplitude Modulation (QAM) 447
1.3.1 Symbol Constellation and Mapping 447
1.3.2 Modulator and Demodulator ... 449
2. **Modulator and Demodulator for Spread-Spectrum Communication Systems** 449
2.1 Direct-Sequence Spread-Spectrum Communication 450
2.2 Frequency Hopping 452
3. **Offset Frequency-Division Multiplex** 454
4. **Conclusions** 457
Abbreviations 457
Further Reading 457

INTRODUCTION

With the advent of the Internet, communication systems are making an evolutional impact on people's lives all over the world. A large quantity of information in various categories, such as telephone, telegraph, computer data, and broadcast, and various services, including bi-directional moving pictures, are required to be transmitted through communication networks. Extremely high-capacity and flexible communication networks are required in order to serve these requirements. In these networks, pulse modulation schemes capable of transmitting a large quantity of data through a restricted communication band play an important role, as does the digital signal processing. This Addendum introduces modulators and demodulators widely used in up-to-date digital communication systems.

1. PULSE MODULATION METHODS FOR DATA-COMMUNICATION SYSTEMS

This section introduces advanced digital pulse modulation schemes and modulators and demodulators. Device and circuit technology are in common for various kinds of modulation schemes. The functions relating modulation and demodulation process, such as line filtering and timing extraction, represented in Sec. 1.2 are also used for other modulation schemes.

1.1 Advanced Modulation Method in Quadrature Phase-Shift Keying Scheme

Binary phase-shift keying (BPSK) and quadrature phase-shift keying (QPSK) modulation schemes have the highest performance and, therefore, are extensively used for various kinds of communication systems, such as sat-

digital Encyclopedia of Applied Physics
Copyright © WILEY-VCH Verlag GmbH & Co KGaA. All rights reserved.

ellite communication, mobile communication, and personal handy telephone systems. This section discusses modulators and demodulators for two types of advanced schemes based on QPSK:

1. offset quadrature phase-shift keying (OQPSK) modulation, which reduces signal distortion caused by nonlinear effects, and
2. quadrature quadrature phase-shift keying (Q^2PSK) modulation, which improves signal throughput.

1.1.1 Offset Quadrature Phase-Shift Keying Modulation (OQPSK). In the modulated signal using BPSK and QPSK schemes, there are abrupt phase transitions with maximum values of π at the boundaries of data. The large phase transition tends to increase signal distortion and to increase the outer band spectrum of the modulated signals when such signals pass through a relatively high-efficiency C-class power amplifier with saturation characteristics. There are several kinds of phase-shift keying modulation schemes to overcome the degradation by reducing the phase transition of the carrier at the data boundary. Offset quadrature phase-shift keying modulation (OQPSK) is the basic method among many kinds of modulation schemes with the small phase transition.

As shown in Fig. 1, the OPQSK modulator has about the same circuit structure as that for BPSK except that data in the quadrature channel are delayed by $T/2$, where T is the duration of the BPSK-modulated signal of in-phase and quadrature channels. Similarly, demodulation of OQPSK and MSK signals is done using a QPSK demodulator. In an OPQSK-modulated signal, the phase transition occurs in every $T/2$, but the signal spectrum does not increase because the operation is a simple summation. The sum signal at the phase transition point is produced by a vector sum of a continuous signal and a signal with $\pi/2$ (or $-\pi/2$) difference in phase at the transition point. Therefore, the maximum value of the phase transition reduces to $\pi/2$.

The degradation due to the phase transition at the boundary is minimized if we can realize a continuous phase condition at the data boundary. In the frequency-shift keying scheme, it is simply realized by selecting frequencies f_i to meet the orthogonal condition under which the total phase shift on each datum ($2\pi f_i T$) is equal to a multiple of 2π (this scheme is called CPFSK and extensively used). In the quadrature phase-shift keying scheme, it is realized by shaping each signal bit in both the in-phase channel (I channel) and the quadrature channel (Q channel) by weighting as shown in Fig. 1(c). The shaping is carried out by any of several weighting functions such as sinusoidal or Gaussian. The modulation scheme using a sinusoidal weighting function is called minimum-shift keying (MSK).

Reduction of the abrupt phase transition of OQPSK makes it possible to use the high-efficiency C-class power amplifier. This feature is effective for downsizing and light-weighting battery-assisted receivers.

1.1.2 Quadrature Quadrature Phase-Shift Keying Modulation (Q^2PSK). This modulation scheme is a focus of attention because of its effective utilization of the frequency resources and is now being applied in various communication systems. Figure 2 illustrates the structure of a Q^2PSK modulator. The scheme uses quadrature phase and an orthogonal shaping function as is used for the MSK. However, it encodes four bits per symbol, compared to two for conventional QPSK, and twice the data per modulation interval is accommodated. On account of the existence of an abrupt change of phase in the modulating signal as shown in waveform (*d*), this scheme requires an increase of 1.6 dB transmission power in order to provide an identical bit error rate for the same bandwidth. However, this scheme can obtain twice the throughput of QPSK.

A Q^2PSK modulated signal has both a phase transition and an amplitude variation in the overlap wave shape. In order to reduce the degradation due to nonlinear effects due to the power amplifier, a linearized saturation power amplifier, which has a saturation characteristic to the peak value without degradation on linearity against a wide range of input signal level, has been developed. Basically, this characteristic is attained by an A-class or AB-class amplifier whose gain is controlled according to the input signal level by some means, such as feed-forward technology, predistortion, and postdistortion. The linearized amplifier tends to be inefficient from the point of view of power consumption, but is now essential for a communication system having an abrupt

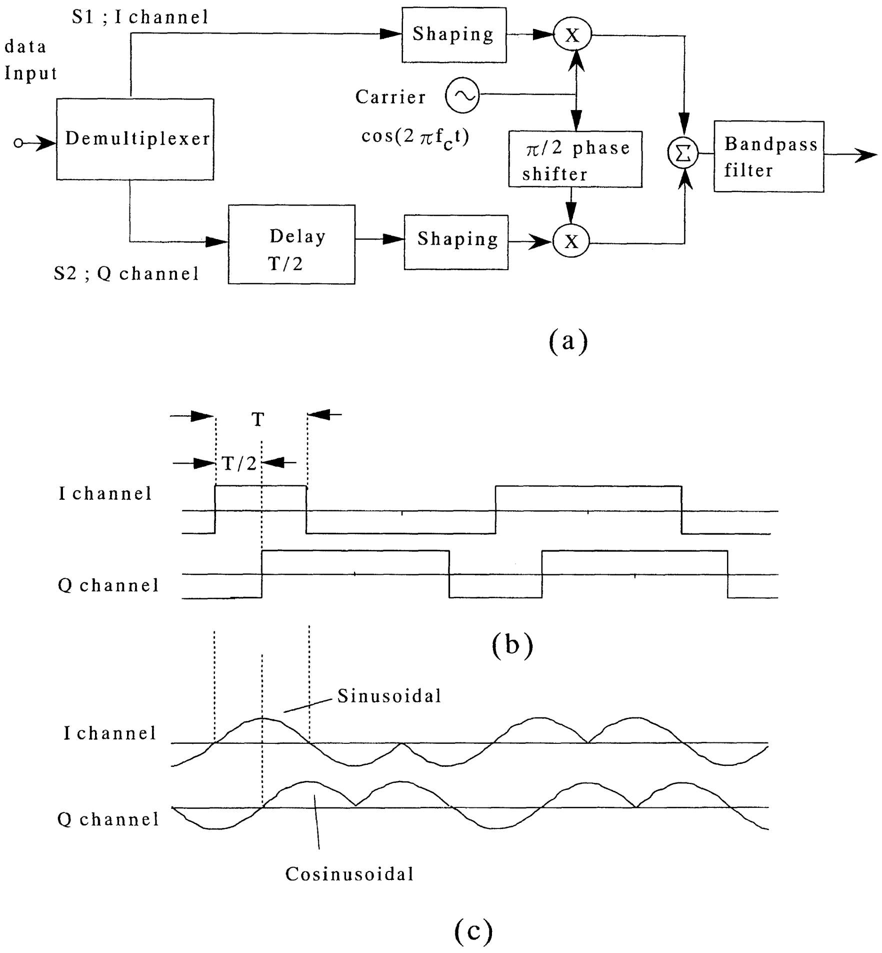
S1 ; I channel
data
Input
Demultiplexer
Shaping
X
Carrier
cos(2 πf_c t)
π/2 phase shifter
Σ
Bandpass filter
S2 ; Q channel
Delay T/2
Shaping
X
(a)
T
T/2
I channel
Q channel
(b)
Sinusoidal
I channel
Q channel
Cosinusoidal
(c)

FIG. 1. OQPSK modulation scheme: (a) modulator, (b) OQPSK signal wave forms, (c) MSK signal wave forms. In (a), ordinary OQPSK needs no shaping function. OQPSK with shaping using a half-cycle sinusoidal weighting function is called MSK. (b), (c) Wave forms for input data with the duration time of $T/2 = T_b$ and sequence "11010011110". The I-channel signal is modulated by a cosinusoidal carrier and the Q-channel by a sinusoidal one.

phase shift or a nonuniform amplitude envelope in its modulated signals. This technology is also applied on systems with modulation schemes presented in the following sections.

1.2 Digital Pulse Amplitude Modulation (PAM-DSB, PAM-SSB, and PAM-VSB)

This section introduces efficient digital modulation schemes for transmitting binary data through a restricted communication channel by using digital pulse amplitude modulation (PAM).

1.2.1 Modulator and Demodulator. Figure 3 illustrates a basic circuit configuration of a PAM modulator and demodulator. At the sending end, the input data sequence is subdivided and mapped onto a corresponding amplitude level of a multiple-level (M-ary) digital signal. The modulated signal is fed to the filter and transmitted through a baseband communication channel or modulated to form a bandpass data signal with a carrier using the conventional analog amplitude modulator: double sideband (DSB), single sideband (SSB), or vestigial sideband (VSB).

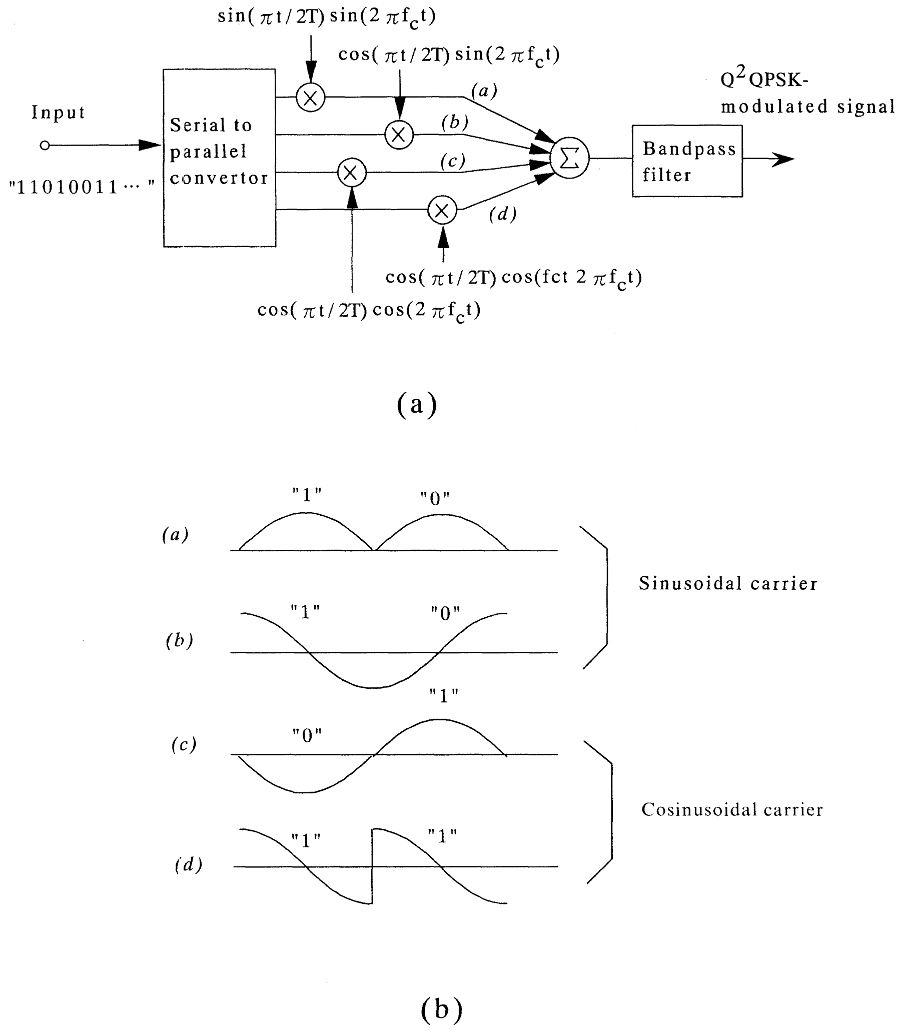
sin(πt/2T) sin(2πf_ct)
cos(πt/2T) sin(2πf_ct)
Input
"11010011…"
Serial to parallel convertor
(a)
(b)
(c)
(d)
Σ
Bandpass filter
Q²QPSK-modulated signal
cos(πt/2T) cos(fct 2πf_ct)
cos(πt/2T) cos(2πf_ct)
(a)
"1"
"0"
Sinusoidal carrier
Cosinusoidal carrier
(b)

FIG. 2. Modulator for quadrature quadrature phase-shift keying.

1.2.2 Timing Extraction. At the receiving end, the received signal must be sampled periodically at the symbol rate and the signal level must be detected by a decision circuit. We need a clock signal to perform the periodic sampling. One method for extracting the timing signal is for the transmitter to transmit simultaneously the clock frequency $1/T$ or a multiple of $1/T$ along with the information signal. The receiver may simply extract the clock signal by employing a narrow bandpass filter tuned to the transmitted clock frequency. This method has the advantage of being simple to implement but has several disadvantages, one of which is that the transmitter must allocate some of its available power and bandwidth to the transmission of the clock signal. There are a number of methods for extracting the clock from the received signal. Figure 4 shows typical circuits: In (a), the clock signal is detected directly using harmonics caused by nonlinear elements. In (b) the clock signal is generated by a VCO whose phase is controlled by the position of the input signal.

1.2.3 Equalization Process. The received signal is distorted by the communication channel by several factors, such as nonlinear amplitude and phase characteristics of the channel, or echo signal reflected at the discon-

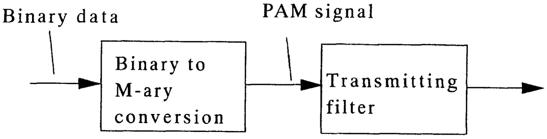

(a)

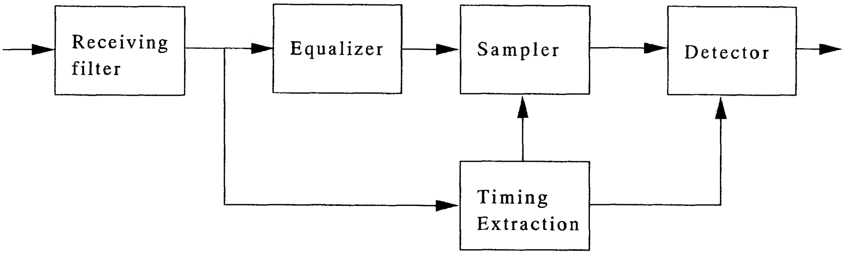

(b)

FIG. 3. (a) PAM modulator, (b) PAM demodulator.

tinuity. This results in intersymbol interference at the output of the demodulator and leads to an increase in the probability of error at the detector. A channel equalizer is used for correcting or undoing the distortion. If the channel is precisely set and its characteristics do not change with time, we can design the equalizer easily. One example is a linear equalizer, which uses a finite–impulse-response filter or a transversal filter, as shown in Fig. 5. We often encounter channels whose characteristics are unknown or change with time. In these circumstances, the channel equalizer tracks the variations of the channel and adopts its coefficients to reduce the intersignal interference dynamically.

1.3 Quadrature Amplitude Modulation (QAM)

This modulation scheme is also called quadrature amplitude-shift keying. It has high efficiency and is extensively used in the modem in computer networks and various kinds of data communication systems including mobile radio, satellite communication, and the information superhighway. In this method, a single pulse in modulated signals contains multiple bits of information in the form of phase and amplitude. This scheme may be viewed as a form of combined digital amplitude and digital phase modulation.

The QAM signal employs two quadrature carriers, $\cos 2\pi f_c t$ and $\sin 2\pi f_c t$, each of which is modulated by an independent sequence of information bits to form a digital PAM signal. Adapting N-level PAM, we can transmit $M = N \times 4$ signal points with one PAM symbol. Therefore, the bandwidth required can be reduced to $1/(2 + \log_2 N)$ times that required for the ordinary BPSK system. A QAM system with up to 256 data points M is being used in an existing system.

1.3.1 Symbol Constellation and Mapping. In this scheme, information bits are mapped onto a two-dimensional complex signal space. Figure 6 illustrates examples of allocation of data points in signal space for an $M = 16$ QAM. These figures are called *symbol constellations*, and *mapping* is the procedure for assigning data symbols onto data points of the

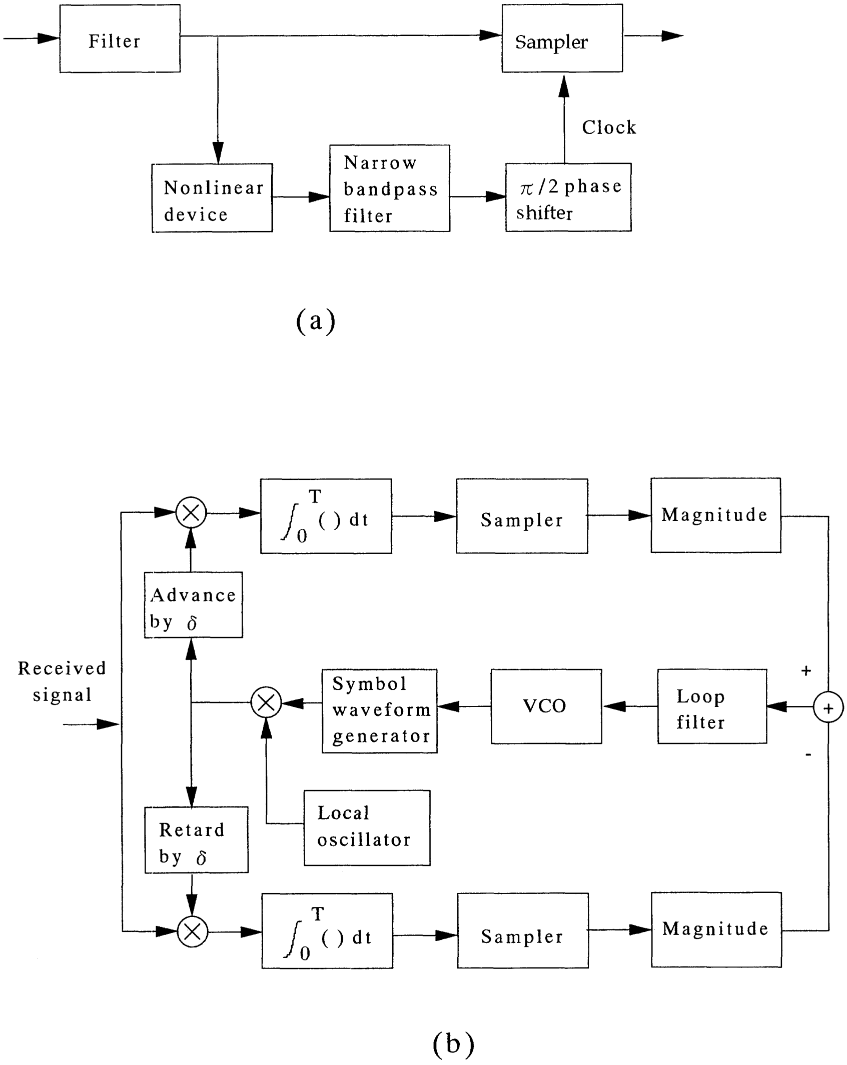
Filter
Sampler
Clock
Nonlinear device
Narrow bandpass filter
π/2 phase shifter
(a)
$\int_0^T () \, dt$
Sampler
Magnitude
Advance by δ
Received signal
Symbol waveform generator
VCO
Loop filter
+
-
Local oscillator
Retard by δ
$\int_0^T () \, dt$
Sampler
Magnitude
(b)

FIG. 4. Timing extraction for PAM. (a) Spectral-line method. Since the received signal has no power spectrum at the clock frequency, it is fed to the nonlinear circuit to generate second-harmonic components of $1/T$ (Hz) and higher harmonics. The clock signal is extracted by a narrow-bandpass filter. (b) Early-late gate synchronizer. In this method, the input signal is correlated by two correlators with a reference signal over the symbol interval $1/T$; one of the correlators starts integration δT early relative to the estimated optimum sampling time and the other starts it δT late relative to the estimated optimum sampling time. The optimal sample point is found by adjusting the phase of the voltage-controlled oscillator that generates the estimated optimum timing of the reference signal so as to make the peak values of the two correlated wave forms coincide with each other.

constellation. These procedures strongly affect system performance.

On designing a constellation, one must consider the following factors:

1. The Euclidian distance between phasors (coordinates on the complex plane) should be minimized in order to realize the characteristic noise immunity of the scheme.
2. The minimum phase rotation between phasors, which determines phase jitter immunity, should be sufficient to stand up against clock recovery imperfection and channel phase rotation.

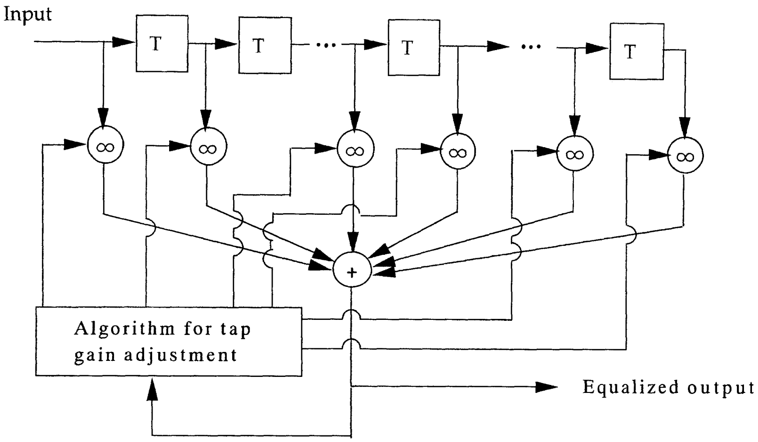

FIG. 5. Equalizer based on a transversal filter. In this circuit, tap coefficients are adjusted so as to realize the inverse frequency response of the channel.

3. The ratio of peak to average phasor power should be a minimum to ensure robustness against nonlinear distortion.

The rectangular constellation with Gray code mapping procedure as shown in Fig. 6(a) has merits with respect to power and signal-to-noise ratio. Systems using this scheme with the numbers of points up to 256 have been developed and are widely used as the standard modulation schemes for communication systems, such as modems. Figure 6(b) shows a circular (or star) constellation. This type of constellation does not have the equidistance condition between phasors. However, it has the advantage of being capable of employing a noncoherent demodulation process, which does not require complex carrier regeneration and AGC circuits. This scheme is extensively used for radio communication systems with a Layleigh fading channel, in which undesired multiple-echo signals with time-varying characteristics are superposed onto the desired receiving signal.

1.3.2 Modulator and Demodulator. Figure 7 shows a block diagram of a modulator and demodulator for a QAM signal with a rectangular constellation. In the modulator, the input data sequence is subdivided and mapped into two digital signal sequences whose amplitudes are modulated according to the input binary sequence, using the same circuit as that used for the PAM modulator. The two PAM signals produced are then modulated by sinusoidal carrier signals whose phases differ by $\pi/2$, using the same circuit as that used in a quadrature phase-shift keying (QPSK) modulator. In the demodulator, the received signal is divided into two PAM signals by multiplying the quadrature carriers in the same way as in the QPSK demodulation circuit. In the QAM demodulator, the same circuits or circuits based on the same principles as those of QAM and QPSK demodulators are employed to perform basic processes, such as synchronization and equalization.

In the circular constellation for 16-point QAM, the modulation process is performed through the encoding process shown in Fig. 8. The demodulation process is a differential decoding process of the modulated signal: The first signal is detected by comparing the signal amplitude at the data timing and that at the previous data timing. If there is a large change in amplitude then the first signal b1 is defined to be “1”, and otherwise “0”. This process is done by two adaptive threshold comparators. The remaining three signals are decoded by detecting the phase difference between the data and the previous data.

2. MODULATOR AND DEMODULATOR FOR SPREAD-SPECTRUM COMMUNICATION SYSTEMS

Spread-spectrum (SS) signals for digital communication were originally developed and

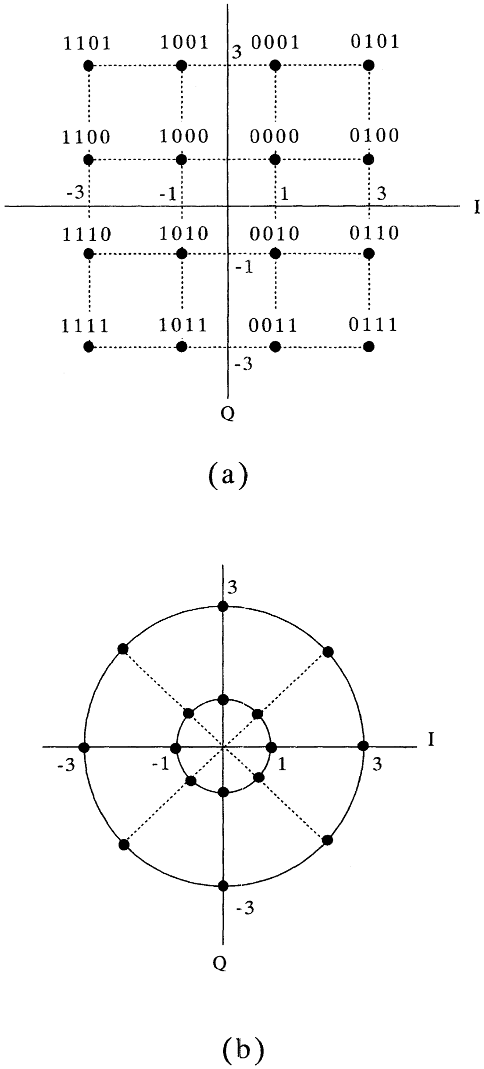

FIG. 6. Symbol constellation for QAM signal. (a) Rectangular constellation, so called from its shape in the signal space. This constellation has equidistant constellation points arranged in such a way that the average energy of the phasors is maximized. This scheme provides the higher signal-to-noise ratio to achieve the same bit error rate (BER). From the point of view of mapping, this example employ so-called Gray code mapping, in which the Hamming distance between a phasor and its closest neighbor is always 1. This procedure ensures that whenever a transmitted phasor is corrupted by noise that is incorrectly identified as a neighboring constellation point, the demodulator chooses a phasor with a single bit error. This minimizes the error probability. (b) Circular constellation. An example of the mapping procedure is illustrated in Fig. 8.

used for military communication to provide resistance to jamming, or to hide the signal by transmitting it at low power and thus making it difficult for the unintended listeners to detect its presence in noise. They are also useful for code-division multiple-access communications (CDMA) when two or more transmitters use the same common channel to transmit information, and they allow reliable communication through extremely noisy communication channels in which systems with conventional modulation schemes are useless. In addition to the robustness against interference, this modulation scheme has the excellent property of efficient usage of the limited bandwidth. Because of its features, this technology is expected to be useful for, and is now beginning to be applied for, a variety of civilian communication systems including satellite, mobile vehicle, and inter- and intra-office communication.

To overcome the problem of intentional or unintentional interference, the bandwidth of a transmitted signal is increased in this modulation scheme so that the bandwidth expansion factor W/R is much greater than unity (where W is the transmission bandwidth and R is the baseband data bandwidth). In addition, the information signal at the modulator is spread in bandwidth by means of a code (channel code) that is independent of the information sequence. The code has a pseudorandom property, so that it appears to be random to receivers other than the intended receiver, which uses the code to demodulate the signal. The usage of the channel codes allows great freedom in the choice of low interuser-interference codes. Proper choices of channel codes make it possible for a large number of users to communicate simultaneously through a limited bandwidth.

There are many kinds of modulation schemes in which the direct sequence scheme and the frequency-hopping scheme are generally used.

2.1 Direct-Sequence Spread-Spectrum Communication

A modulator for accomplishing the spreading in the direct-sequence (DS) scheme is shown in Fig. 9. Here, we assume that the binary information signal is at a rate $R\,(= 1/T_b$,

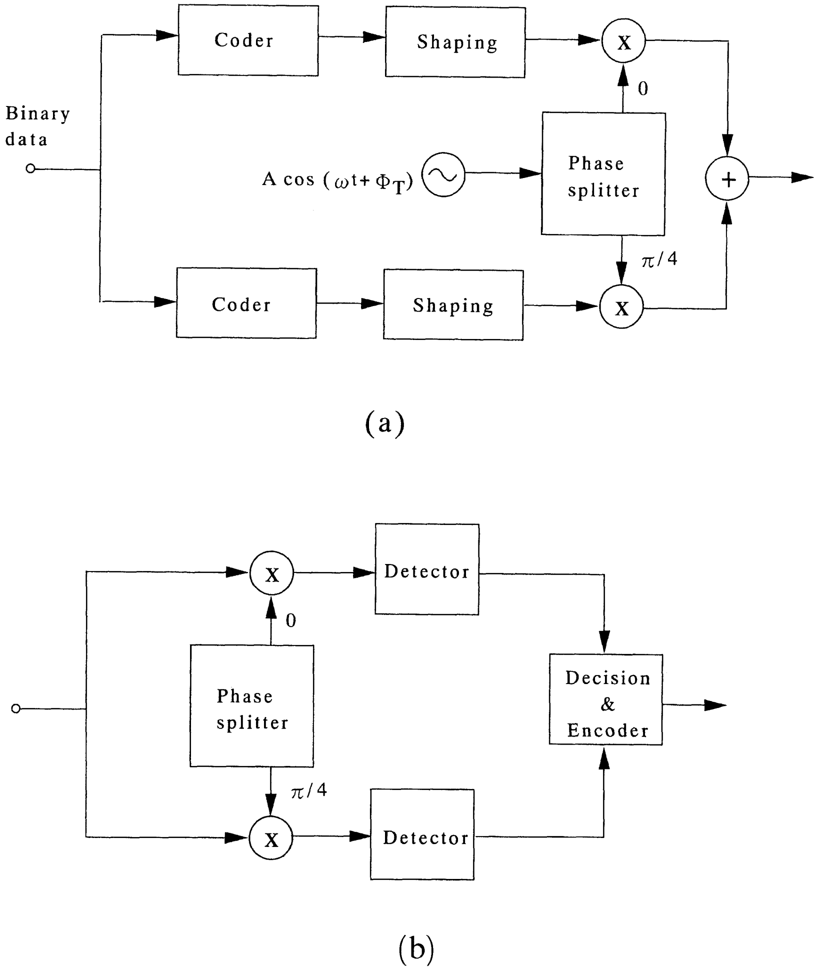

FIG. 7. (a) Modulator and (b) demodulator for QAM with rectangular constellation.

where T_b is the bit interval) bits per second and the available channel bandwidth is B_c Hz, where $B_c \gg R$. At the modulator, the information signal (a) is modulated with a bipolar type of pseudonoise [or pseudorandom noise, PN; this term is commonly used in this field as the same category of pseudorandom binary sequence (PRBS)] pulse sequence with its bit interval of $T_c(= 1/B_c)$ using the product modulator. The bandwidth of the information signal is expanded to $W = B_c$ Hz by shifting the phase of the carrier pseudorandomly at a rate of W times per second. The DC-SS signal is produced by modulating the expanded signal as shown in (c) with a sinusoidal carrier to form the BPSK signal and is transmitted through a communication channel.

We usually call a pulse in the PN code sequence the *chip*, T_c the *chip interval*, and $R_c = B_c$ the *chip rate*. The ratio of the bit interval to the chip interval $L_c = T_b/T_c$ is usually selected to be equal to the number of chips in the PN code sequence per information bit. The PN code sequence mentioned here is a code sequence of 1s and 0s whose autocorrelation has properties similar to those of white noise; the wave form of the autocorrelated signal has a steep peak at only one time point and at any other time position has a side lobe with a small noise amplitude. In the DS-SS

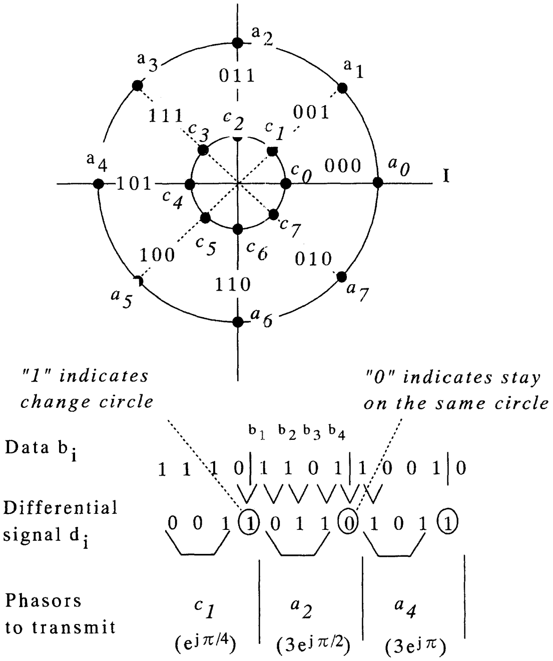

FIG. 8. Modulation process for 16-point QAM with a circular constellation. Of 4 bits in each symbol, b_1, b_2, b_3, and b_4, the first signal b_1 is differentially encoded onto the QAM phasor amplitude so that "1" causes a change to the amplitude circle that was not used in the previous signal and "0" causes the current symbol to be transmitted at the same amplitude. The remaining three signals are differentially Gray encoded on the circle with the same amplitude as that of the first signal.

system, the binary sequence with element $\{0, 1\}$ is mapped into a corresponding binary sequence with element $\{-1, 1\}$. The most widely known PN code sequence is the maximum-length sequence, which has a length $2^m - 1$ bits and is generated by an m-stage shift register as shown in Fig. 9(b).

The demodulation is the process of *despreading* of the spread signal. This is done by the circuits shown in Fig. 10. That in Fig. 10(a) uses direct cross correlation using a matched filter, such as a SAW correlator or convolver. This method has the advantage that demodulation can be performed with an extremely simple circuit and does not require a complicated synchronization process.

In the coherent demodulation process shown in Fig. 10(b), signal synchronization is critical to maintaining stable communication. The synchronization step can be separated into two distinct phases. The usual procedure for establishing initial synchronization is for the transmitter to send a known pseudorandom sequence to the receiver. The receiver is continuously in a search mode looking for this sequence to establish initial synchronization. In principle, matched filtering or cross correlation [as used in the circuit of Fig. 10(a)] are optimum methods for establishing initial synchronization in the presence of additive Gaussian noise. The matched filter establishes the initial synchronization in the same way as that of the modulator using a matched filter, and the demodulator enters the "data receive" mode. Examples of circuits for initial synchronization and tracking are shown in Fig. 11. Figure 11(a) shows an example of the initial synchronization method called a sliding correlator. The tracking is usually performed by means of a tracking loop called a delay-locked loop (DLL) as shown in Fig. 11(b). This implementation is equivalent to the early-late gate bit tracking synchronization described in Sec. 1.2.

2.2 Frequency Hopping

An example of the signal pattern of FH is shown in Fig. 12. In this scheme, the available channel bandwidth W is subdivided into a large number of nonoverlapping frequency slots. In any signaling interval the transmitted signal occupies one or more of the available

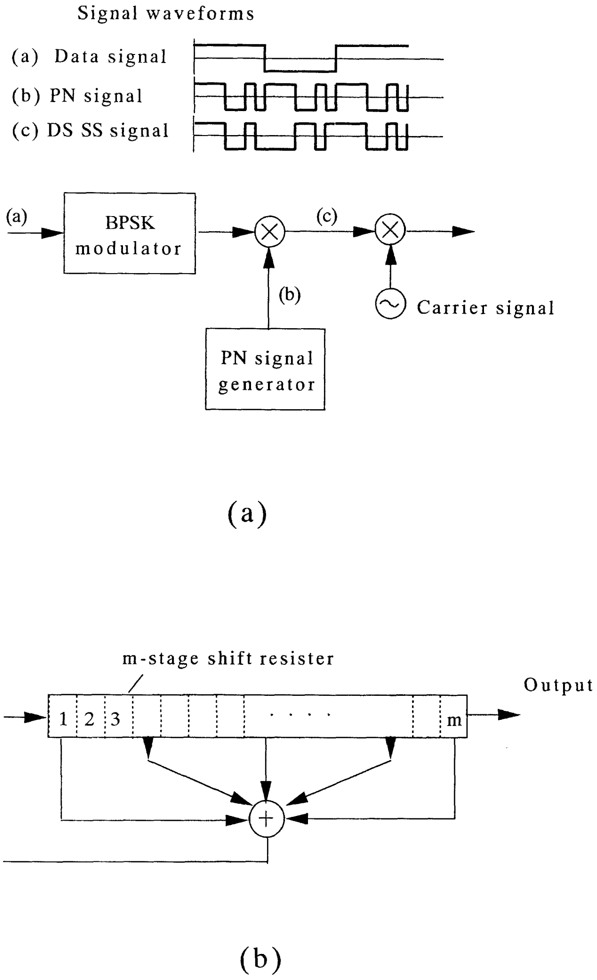

FIG. 9. Modulator for a direct sequence SS signal. (a) Block diagram, (b) maximum-sequence PN code generator.

frequency slots. The selection of the frequency slot in each signal interval is made pseudorandomly according to the output from a PN generator.

A block diagram of the modulator and demodulator for a FH-SS system is shown in Fig. 13. The modulation is either binary or *M*-ary FSK (MFSK). For example, if binary FSK is used, the modulator selects one of two frequencies, say f_0 or f_1, corresponding to the transmission of a 0 or a 1. The resulting binary FSK is translated in frequency by an amount that is determined by the output sequence from the PN generator, which used to select the frequency f_c that is synthesized by the frequency synthesizer. This frequency is mixed with the output of the output FSK modulator and the resultant frequency-translated signal is transmitted over the channel. For example, by taking m bits from the PN generator, we may specify $2^m - 1$ possible carrier frequencies.

At the receiver, there is an identical PN sequence generator, synchronized with the received signal, which is used to control the output of the frequency synthesizer. Thus, the pseudorandom frequency translation introduced at the transmitter is removed at the demodulator by mixing the synthesizer output with the received signal. The resultant signal is then demodulated by means of an FSK demodulator.

A signal for maintaining synchronization of the PN sequence generator with the FH received signal is usually extracted from the re-

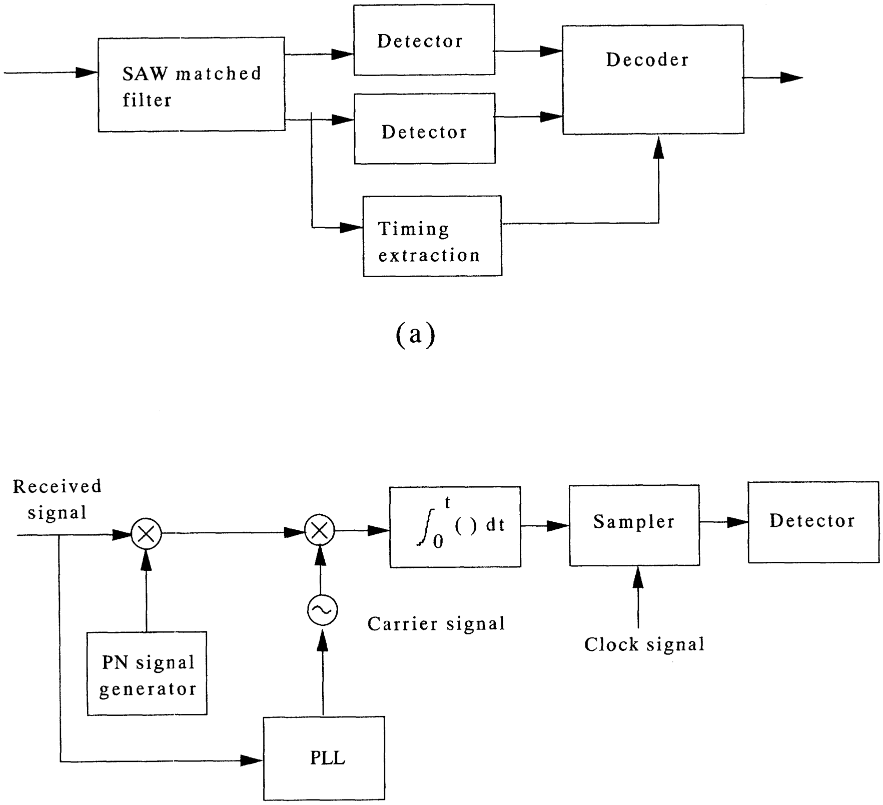

(b)

FIG. 10. Demodulator for a direct-sequence SS signal. (a) DS demodulation using SAW matched filter, (b) coherent demodulation. In (a), the received signal is autocorrelated by a matched filter and a compressed signal is generated directly by detecting the peak values from matched filters; thus the data will detected. In (b), the input signal is multiplied by the same PN sequence and is cross-correlated with a rectangular pulse of duration T_b. The data decision is performed by detecting whether the level of the output signal from the sampler exceeds thresholds.

ceived signal. Nonsynchronous detection is usually used for FH SS systems because it is difficult to maintain phase coherence both in the synthesis of the frequencies used in the hopping patterns and also in the propagation of the signal over the channel as the signal is hopped from one frequency to another over a wide bandwidth.

In the FH method, the initial acquisition process is to synchronize the known PN code sequence generated at the receiver that controls the hopping frequency pattern. The same procedures used in the DS signal synchronization procedure, such as serial search and sliding correlator, are useful for FH signals. However, it requires a long searching time. By employing multiple correlators in parallel, the searching time can be reduced at the expense of a more complex and costly implementation. For the tracking, the same technology as that used in DS-SS is applied.

3. OFFSET FREQUENCY-DIVISION MULTIPLEX

There exist some other modulation schemes with high efficiency in both transmission power and frequency resources. Among them,

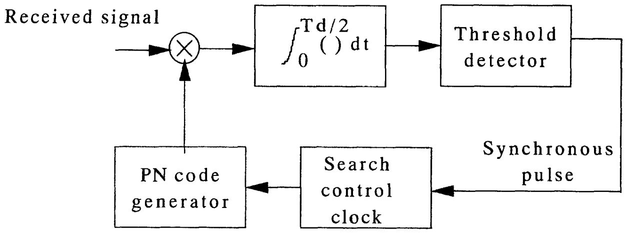
Received signal
$\int_0^{Td/2} (\ ) dt$
Threshold detector
Synchronous pulse
Search control clock
PN code generator

(a)

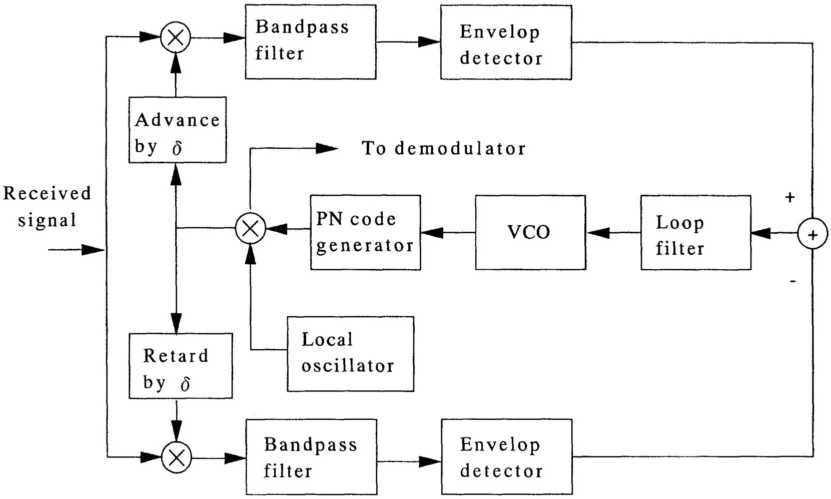
Bandpass filter
Envelop detector
Advance by δ
To demodulator
Received signal
PN code generator
VCO
Loop filter
+
+
-
Local oscillator
Retard by δ
Bandpass filter
Envelop detector

(b)

FIG. 11. Methods for synchronization. (a) Sliding correlator for DS signal acquisition. The received signal is cross correlated with a rectangular pulse of $T_C/2$ seconds or less. Then integration is performed over the time interval NT_C, and the peak value of the output is compared with a threshold to determine if the known signal sequence is present. If the known threshold is not exceeded, the known reference signal is advanced by $T_C/2$ seconds and the correlation process is repeated. These operations are performed until a signal is detected or until the search has been performed over the uncertain interval. (b) Tracking using delay-locked loop (DLL). In this loop, the received signal is applied to two multipliers, where it is multiplied by two outputs from the local PN code generator, that are delayed relative to each other by an amount $2\delta < T_C$. The product signals are bandpassed and detected, and then subtracted. They are then applied to the loop filter that drives the clock VCO. The VCO output serves as the clock for the PN code-signal generator. In this process if the synchronization is not exact, the filtered outputs from two correlators will be different from each other and the phase of the VCO will be appropriately advanced or delayed. At the equilibrium point, the two filtered outputs will be equally displaced from the peak value, and the PN code output fed to the demodulator will be exactly synchronized to the received signal.

an excellent scheme is the offset frequency-division multiplex (OFDM) modulation. The operation principle of this system is that the original bandwidth is divided into a large number of narrow sub-bands in which the mobile channel can be considered to be non-

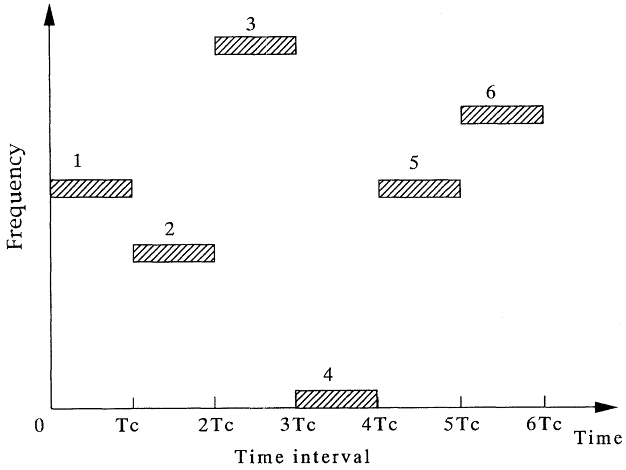

FIG. 12. An example of an FH signal.

dispersive. In other words, by setting the duration time of the modulated signal larger than the delay time of echo signals, it makes an undesirable echo no disturbance on the desired signal.

The modulation is performed by transforming serial time-sequence pulse data to parallel long-duration pulse signals with orthogonal carrier frequencies. This process is

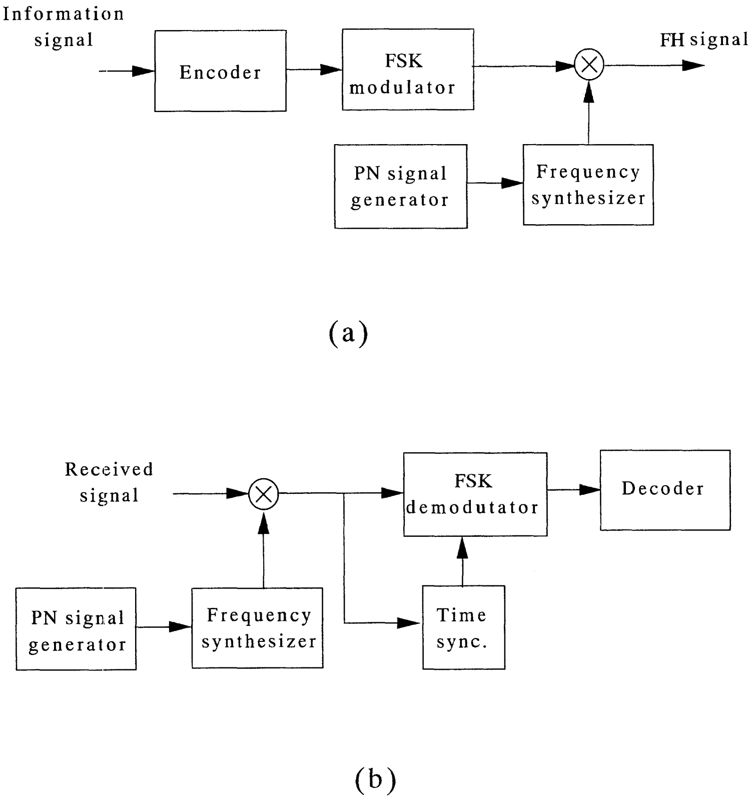
Information signal
Encoder
FSK modulator
FH signal
PN signal generator
Frequency synthesizer
(a)
Received signal
FSK demodutator
Decoder
PN signal generator
Frequency synthesizer
Time sync.
(b)

FIG. 13. (a) Modulator and (b) demodulator for FH-SS signal.

simply performed by a fast Fourier transformer (FFT) [or a discrete Fourier transformer (DFT)]. At the receiving end, the demodulation is also done by performing inverse discrete Fourier transform using an FFT. This scheme has the advantages of being implemented without a complex channel equalizer. Since the echo signal within a data interval contributes to an increase in the received signal power, this scheme has an advantage also in power efficiency, in addition to the robustness against multipath fading. Because of these features, this scheme is adopted for various digital wireless communication systems such as audio and digital TV broadcasting systems.

4. CONCLUSIONS

The basic concepts for modulation schemes stated here were proposed in relatively old days (several tens of years ago, or more). However, recent progress in device technology and global needs for communication systems as an infrastructure of social life accelerated the research and development of systems with these schemes. Technologies to improve system performances have also developed—for example, in DAM schemes, the development of linearized power amplifiers, adoption of error correction, and pilot-signal–assisted modulation—with the results that these systems are now in practical use in a wide area. As is represented by the information-superhighway concept, the importance of communication technologies is increasing still more, so that further development of technology of modulators and demodulators can be expected.

ABBREVIATIONS

DFT: Discrete Fourier transform.
FFT: Fast Fourier transform.
DLL: Delay-locked loop.
DS: Direct sequence.
DSB: Double sideband.
FH: Frequency hopping.
I-channel: In-phase channel.
Q-channel: Quadrature-phase channel.
MSK: Minimum-shift keying.
MFSK: M-ary frequency-shift keying.
OQPSK: Offset quadrature phase-shift keying.
PN: Pseudoranom noise.
PRBS: Pseudorandom binary sequence.
QAM: Quadrature amplitude modulation.
Q^2PSK: Quadrature quadrature phase-shift keying.
SAW: Surface acoustic wave.
SS: Spread spectrum.
SSB: Single sideband.
VCO: Voltage-controlled oscillator.
VSB: Vestigial sideband.

Further Reading

Alisouskas, V. F. A., Tamsi, W. (1985), *Digital and Data Communications*, Englewood Cliffs, NJ: Prentice-Hall.

Campbell, C. (1989), *Surface Acoustic Wave Devices and Their Signal Processing Applications*, London: Academic Press.

Dixon, R. C. (1976), *SpredSpectrum Systems*, New York: Wiley.

Gilcic, S. G., Leppanen, P. A. (Eds.) (1995), *Code Division Multiple Access Communications*, London: Kluwer Academic Publishers.

Proakis, J. G., Salhi, M. (1994), *Communication Systems Engineering*, Englewood Cliffs, NJ: Prentice-Hall.

Webb, W. T., Hanzo, L. (1995), *Modern Quadrature Amplitude Modulation—Principles and Applications for Fixed and Wireless Channels*, London: Pentech Press Publishers, and New York: IEEE PRESS.

Deforming Nanocrystalline Metals: New Insights, New Puzzles

Jörg Weissmüller and Jürgen Markmann

1
Introduction

Grain boundaries are obstacles for motion, as evidenced by the increase of the yield stress of polycrystals with decreasing grain size embodied in Hall-Petch type equations. Thus, it is expected that when the grain size of a metal is reduced to extremely small values, as in nanocrystalline (nc) materials, dislocation activity should become comparatively less important in favor of alternative mechanisms which involve transport at grain boundaries, most notably diffusion creep and grain boundary sliding [1, 2]. One of the consequences of having a mechanism which circumvents bulk dislocation plasticity would be that materials which, in the coarse-grained state, are brittle due to a lack of active slip systems, might become ductile when in the nanocrystalline form.

The recent past has seen a revived interest in studies of the mechanical properties of nanocrystalline metals. Among the driving forces for this trend are the immense increase in performance of modern computers, which has enabled more and more realistic simulations of the deformation of these materials, leading to predictions on the underlying microscopic processes which have stimulated new experiments. The synergy between experiment and simulation is critically assessed in the comprehensive review, Ref. 3 Equally important has been the advent of materials of improved quality. Notably, ultrafine-grained materials [4] (with grain size D of roughly 100–500 nm) can be produced in considerable quantity, and nanocrystalline sheets with D of the order of 50 nm can be obtained by electrodeposition. At the lower extreme of the size scale are materials produced by assembling inert-gas condensed clusters in situ to form bulk materials; grain sizes down to 5 nm can be obtained by this technique. As compared to early studies, from which the first suggestions for modified deformation mechanism and modified stress-strain relations at nanometer grain size emerged fifteen years ago [1, 2], we now have at hand samples of considerably reduced porosity and enhanced purity. As an example, Figure 1 displays an inert-gas condensed (igc) nanocrystalline Pd sample.

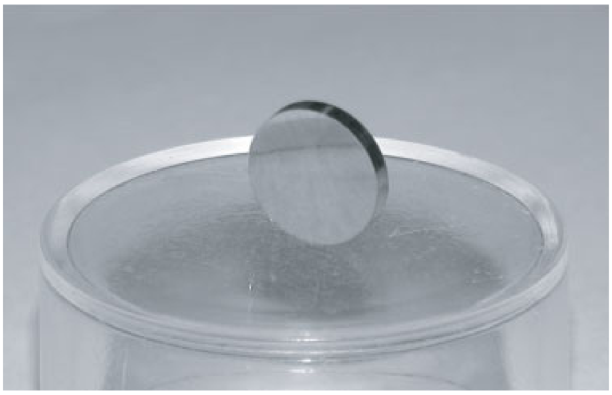

Fig. 1 Bulk nanocrystalline metal samples with grain sizes down to about 10 nm can be produced by in-situ consolidation of inert-gas condensed nanoparticles in an UHV environment. The Pd sample in the image has the diameter 8 mm and thickness of about 0.5 mm. The initial density, typically 91–93% of bulk Pd, is increased to 96–97% after deformation by rolling.

Previous studies of the deformation of igc materials, using micro tensile tests, found brittle behavior or strain to failure of not more than few percent [5]. By contrast, more recent studies showed that high density, high purity igc samples can be deformed to a true strain in excess of 0.5 by rolling. For the first time, this has now made it possible to study systematically the evolution of the defect structure and of the microstructure during the deformation of truly nanocrystalline bulk metals [6, 7]. Here, we highlight some of the issues which have arisen from these experimental studies.

2
Diffusion Creep

It has been suggested early-on that the rate of Coble creep, which is based on transport via grain boundary diffusion and which varies as D^{-3}, could become the dominant deformation mechanism at small grain size, even at room temperature. Recent studies of the deformation of nc Pd at small stress find the creep rate quantitatively compatible with the estimated rate of Coble creep [6]. There is a considerable uncertainty in such measurements, *i*) due to the extrapolation of the grain boundary diffusion coefficients – which were measured at high temperature – to room temperature, and *ii*) because no steady state creep was established. Still, the findings can be discussed with respect to the proposition of dominantly diffusive deformation: Markmann et al. measured strain rates of 10^{-10}/s at the stress of 10 MPa [6]. Since the stress for diffusive creep varies linearly with the strain rate, this means that an unphysically high stress, of the order of 10^{10} MPa, would be required to achieve the strain rates of the order of 10^{-1}/s required in industrial forming processes. Even when a large margin of error is admitted, this would seem to rule out Coble creep as a practicable deformation pathway at room temperature.

It is an open question how this finding can be reconciled with the observation that nanocrystalline metals, if their grain size is not stabilized – intentionally by alloying or unintentionally by voids or impurities – undergo significant grain growth even at room temperature [8]. This testifies to an extraordinarily high atomic mobility in the grain boundary cores. When the grain size doubles, typically within a few days for Pd at $D = 8\,\text{nm}$ and $T = 300\,\text{K}$, the number of grains has been reduced by the factor of eight, so at least 7/8 of the atoms have been displaced from their original lattice site. Since grain growth does not intrinsically entail a plastic deformation, it is conceivable that grain boundary migration simply does not couple to the stress and so does not contribute to the creep rate.

3
Grain Boundary Sliding and Grain Rotation

It is well established that grain boundary sliding plays an important role in the superplastic deformation of conventional coarse grained materials. Since the number of grain boundaries per volume increases as the grain size is reduced, it can be expected that sliding is a relatively more important process at smaller grain size. Intricately linked to this issue is that of grain rotation. Dislocation glide on preferred slip systems gives rise to rotation and to crystallographic texture. By contrast, grain rotation via grain boundary sliding along may randomize the grain orientation distribution. It is therefore remarkable that experiments on nc Pd deformed to a large true strain, 0.6, by rolling at room temperature show that the random grain orientation distribution of the starting material is maintained, whereas coarse grained Pd deformed under identical conditions develops a pronounced rolling texture [6]. This provides strong evidence for a significant role of grain rotation in the deformation of the nanocrystalline metal, similar to the grain rotation during superplastic deformation of coarse grained materials at elevated temperature. In fact, individual rotation events (apparently coupled, in the experiment, with grain growth) have recently been observed by in-situ transmission electron microscopy (TEM) [9].

Grain boundary sliding over finite distances requires changes in the grain shape. Therefore, an accommodation process must be active in parallel. Hahn and Padmanabhan [10] have proposed a mechanism that circumvents accommodation when the grain size of nanocrystalline ceramics becomes comparable to the grain boundary thickness. For metals, where the grain boundary thickness is of the order of one interplanar spacing [11], the required extremely small grain sizes have not been reached so far. However, the formation of mesoscopic shear planes – stretches of several coplanar grain boundaries in a row which facilitate fast deformation via sliding with little accommodation – during the superplastic deformation of metals with grain sizes in the micrometer-region is well supported by experiments [12]. These features have also been observed in nanocrystalline materials, both by computer simulation [13] and by experiment [6]. However, more systematic investigations are required to determine how frequent these features occur and how significantly they contribute to the deformation. Most observations of nanomaterials so far show serrated grain boundary networks between polygonal grains, in which sliding can only be understood if it is accompanied by accommodation. With diffusive transport ruled out, dislocation plasticity is the obvious candidate for the accommodation process.

4
Dislocation-Mediated Plasticity

4.1
Are Nanomaterials Dislocation-Free?

Isolated nanocrystals are often free of dislocations, which would be attracted to the surface by large image forces. By analogy, one may assume that undeformed nanocrystalline materials are free of dislocations as well. Experiment does not support this view: dislocations have been repeatedly observed, even in the immediate vicinity of grain boundaries

[7]. This is not too surprising when one considers the fact that grain boundaries can transmit both normal and shear components of the stress to the neighboring grains, so that the magnitude and even the sign of the image forces are less obvious than at free surfaces. Furthermore, microstrain is ubiquitous, and has even been proposed to be intrinsic to the small grain size [14]. The local strain may balance image forces, stabilizing dislocations in the interior of nanograins.

It is worth mentioning that a typical RMS microstrain value in nanocrystalline fcc metals is $\langle \mu^2 \rangle^{1/2} = 0.005$ at the grain size $D = 10\,\mathrm{nm}$; [14] for Pd this translates into an RMS stress of roughly 0.6 GPa, of the same order as typical yield stresses of nanocrystalline metals [5, 15]. This large variation in local stress values complicates the analysis of experiments in terms of microscopic mechanisms relying on a well-defined stress state.

4.2
Do Dislocations Contribute to the Plasticity?

At room temperature and not too small strain rates, conventional coarse grained metals deform by dislocation glide. The Frank-Read source provides a conceptual frame for the yield stress as well as a mechanism for dislocation multiplication. As the deformation proceeds, the dislocation density increases and the interaction between dislocations on different slips systems in coarse-grained materials leads to work hardening. Even at this most simple level of understanding, it is subject to doubt if the conventional concepts of dislocation plasticity apply to nanomaterials.

In spite of the observation of dislocations in nanograins, there is so far no reported observation of a Frank-Read source in a nanocrystalline solid. Once again this is not surprising, since the stresses involved would be prohibitive: by inspection of Figure 2 it is readily seen that the largest possible source would require a distance between pinning points of $L > D/3$. The Orowan stress is at least (in the case of a pinned edge dislocation) $\lambda = Gb/L$. For a full dislocation and a Shockley partial in Pd this evaluates to 2.4 GPa and 4.8 GPa at $D = 10\,\mathrm{nm}$, respectively, when $L = D/4$ is used. These stresses are above the highest yield stress reported so far for nanocrystalline pure metals.

In spite of the fact that the high critical stress makes the action of standard dislocation sources unlikely, experiments have recently confirmed that dislocations are active during the deformation of nanocrystalline fcc metals. By HRTEM, Chen et al. [16] found twins in deformed thin films of nc Al. Rösner and coworkers confirmed the existence of twins by investigation of deformed nc Pd [7]. Markmann et al. investigated bulk nc Pd deformed by rolling; by means of x-ray powder diffraction they found that a large density of stacking faults develops as the deformation proceeds [6]. Stacking faults have also been found in HRTEM work on deformed nc Al [17]. Faults and twins are created by the motion of partial dislocations, they are therefore indicative of deformation via the glide of lattice dislocations. Thus, there is now compelling experimental evidence of dislocation activity contributing to the deformation of bulk nanocrystalline materials.

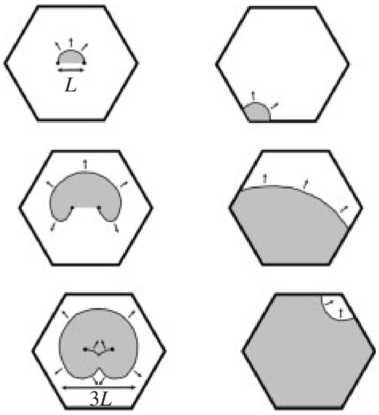

Fig. 2 Dislocation multiplication by activation of a Frank-Read source with distance L between pinning points in a conventional coarse-grained metal (left) and dislocation nucleation and adsorption in a nanograin (right). The hexagons represent the respective grains, and the shaded area represents the sheared region of the slip plane.

4.3
Dislocation Nucleation versus Multiplication

Molecular dynamics simulations explore the regime of very high stress, since high strain rates are imposed by the short computational timescale. Still, even at these high stresses no observation of dislocation emission from Frank-Read sources has been reported. Instead the simulations indicate that a fundamentally different mechanism is active [18, 19]: New dislocations, often partials, nucleate at grain boundaries, transverse the grain, and are then absorbed by the grain boundary on the opposite side of the grain, as sketched in Figure 2. Thus, the dislocation density does not increase as the deformation progresses. In agreement with this notion, nanocrystalline fcc metals exhibit unusually low work hardening [15, 20].

If one was to accept this mechanism of dislocations popping into a grain at grain boundaries, and popping back out at the opposing boundary as the elementary event of dislocation plasticity at small grain size, a number of emerging fundamental issues would require to be addressed:

i. What is the microscopic mechanism of dislocation nucleation at grain boundaries?
ii. Which type of dislocation will be emitted?
iii. What is the required (local?) stress?
iv. How is the local strain concentration at boundaries, associated with the nucleation and absorption accommodated?
v. Is there an energy barrier for nucleation, and what is the role of thermal activation in overcoming it?

So far, few arguments endowed with predictive power have been brought forward in this field. A possible exception will be critically discussed in the next section.

4.4
Faulting versus Twinning

Chen et al. [16] have made a suggestion on question *ii.* of the previous section: by analyzing the dislocation emission from a Frank-Read source they showed that, when the anchor spacing becomes very narrow, the emission of partials is favored over that of full dislocations. They argued that this explains why nc Al forms twins upon deformation, in spite of the materials high stacking fault energy. The observation of twinning in Pd [7], another high stacking fault metal, supports these findings.

The argument by Chen et al. exploits the fact that the area of the fault decreases faster, with diminishing dimension, than the line-length of the dislocation. Thus, at small size the low energy partial dislocation is preferred over the full dislocation, in spite of the extra energy stored in the faulted area of the glide plane. These considerations are convincing, but their combination with the Frank-Read source entails the complications addressed above: no experimental confirmation, unrealistically high activation stress. In fact, Rösner et al. [7] have reported an observation which is not readily explained in terms of the Chen model or of a recent refinement [21]: in Pd deformed by rolling at strain rates near $0.08\,\mathrm{s}^{-1}$ they found abundant faulting, but no twinning. By contrast, when the strain rate was raised to $0.3\,\mathrm{s}^{-1}$, abundant twinning was observed. In other words, the mere fact that partial dislocations can be generated does not automatically imply twinning. Instead, there appears to be a barrier to twinning which is only overcome at the higher strain rate. The nature of this barrier is not understood so far.

It is generally acknowledged that deformation twinning results form the cooperative passage of partial dislocations on successive glide planes. The details of this process are the subject of debate even for conventional materials, and it is not obvious in how far concepts such as the pole mechanism or double cross-slip are transferable to situations where – as it appears to be the case in nanomaterials – each individual partial dislocation has to nucleate separately from a high-angle grain boundary. Even more puzzling is the formation of arrays of very finely spaced twin lamella, where the process acquires an extra characteristic length-scale of periodicity, the lamellae spacing [7].

4.5
Slip System Selection and Alignment by Grain Rotation

A significant observation is that, within each individual grain, all deformation twin boundaries are coplanar, as is shown in the high resolution TEM image of Figure 3. Apparently, the twin networks – involving twin boundaries on more than one crystallographic set of planes – which were found in MD simulations [22] require the vastly higher strain rate (of the order of $10^7\,\mathrm{s}^{-1}$) of the numerical studies, and the associated high stresses. This indicates that large stresses would be required for transmitting dislocations through twin boundaries. An interesting consequence is that exceptionally high yield stresses and significant strain hardening can be reached with samples of micrometer grain size and nanoscale twin spacing, as was recently demonstrated for Cu [23]. Remarkably, however, similar strain hardening has not so far been reported for truly nanocrystalline samples in spite of the tendency to form deformation twins, nor are the products of dislocation-twin interactions observed. This indicates that an

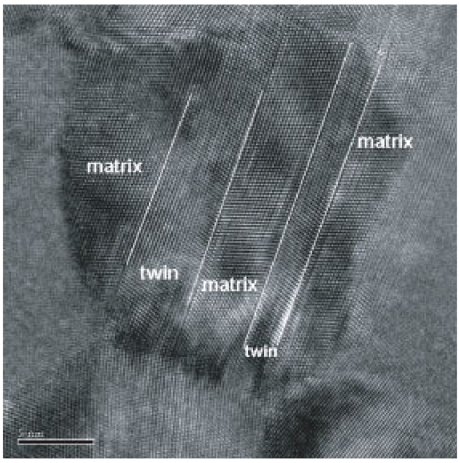

Fig. 3 High-resolution transmission electron micrograph of nc Pd deformed to a true strain of 0.32 by rolling at the strain rate μ = 0.3/s. After Ref. [7].

alternative deformation pathway may be active in nanocrystalline materials, which avoids the activation of slip systems that intersect the deformation twin boundaries.

Rösner et al. [7] have highlighted this issue from a different angle. They pointed out that the finding that twinning selects exclusively parallel planes in deformed nc Pd raises a puzzling question: it implies that only one glide system is active within each grain, significantly short of the five independent glide systems that are required for a general deformation in a polycrystal (after von Mises). In view of the large true strain value in their experiment, and of the presence of grain rotation (see above), it appears likely that the local direction of principal shear stress experienced by any given grain varied as the deformation proceeded. As the material apparently did not activate a second slip system once it had decided on the direction of twinning in the early stages of deformation, other processes than dislocation glide must then have become active for further deformation to be possible.

Rösner et al. suggested that grain rotation may provide the requested degree of freedom: it allows each nanograin to adjust its orientation in such a way that, throughout all stages of the deformation, the active glide plane remains aligned in parallel with the local shear direction. If this behaviour, which is sketched in Figure 4, could be confirmed in materials other than fcc metals, it could provide a – highly desirable – mechanism for achieving plasticity even in materials with less than five active slip systems if the grain size is small enough.

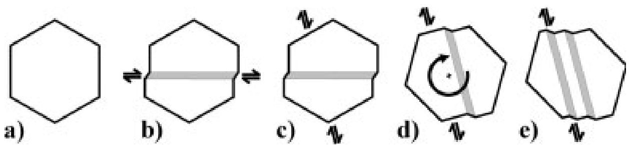

Fig. 4 Twinning and slip system alignment by rotation as suggested in Ref. [7]. The undeformed nanograin (a) is sheared along the slip system with maximum projected shear stress λ_{max} (arrows) by twinning (b), thereby selecting the single active slip system. At a later stage of deformation, the local direction of λ_{max} has changed (c). Since the deformation along the corresponding slip systems is blocked, grain rotation (d) is required to align the active slip system with λ_{max}, leading to further deformation (e).

5
Deformation Mechanism Maps

The recent experimental and simulation results on the deformation of nanocrystalline fcc metals have arguably raised as many questions as they have provided answers. Yet, the field has clearly moved beyond the early stage where the question 'what is the active deformation mechanism in nanocrystalline metals' summarized the direction of research. The body of available results indicates that there is not 'the' deformation mechanism. Instead, depending on the grain size, temperature, and stress or strain rate quite different mechanisms may be dominant. Quite analogously to conventional materials one may therefore attempt to summarise the insights on microscopic processes in the form of deformation mechanism maps for nanocrystalline materials.

Deformation maps can be drawn in various representations. We find it useful to plot the mechanism in the plane spanned by the grain size as the abscissa and the strain rate as the ordinate. A tentative map plotted in this way is shown in Figure 5. As discussed above, the map shows Coble creep at very small strain rate, followed by the combination of partial dislocation activity with grain rotation and grain boundary sliding as discussed above, including the transition from faulting to twinning. At the highest strain rate, the map lists the formation of twin networks as indicated by MD simulation [22]. Also a results of MD work is the transition to a dominantly sliding-controlled deformation at the very

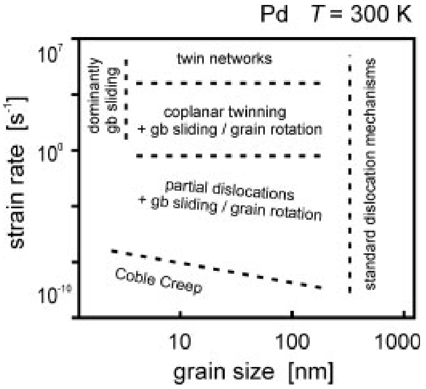

Fig. 5 Tentative deformation mechanism map for Pd at small grain size. See text for details.

smallest grain size and high strain rate [24], represented by the top left region of the map. It is emphasized that the regime of strain rates covered by the combined experimental and simulation studies extends over 17 orders of magnitude; in view of the few available results it appears highly likely that processes not so far identified are active in regions of the deformation map which have not yet been explored. Therefore, a representation such as Figure 5 cannot aspire to represent a complete picture of deformation mechanisms in nanomaterials; it should be viewed as an attempt to provide a snapshot of the available knowledge in a field which is evolving quickly.

As a further word of caution we emphasize that the map of Figure 5 combines data collected in the different stress states associated with uniaxial tension (MD simulations and creep) as compared to rolling. The hydrostatic stress component, compressive in rolling versus tensile in uniaxial tension, may affect the dislocation processes via their activation volumes. On the other hand, in view of the highly nonuniform local state of stress of a polycrystalline material during deformation, and in view of the very high stresses of both signs associated with the microstrain in nanocrystalline materials, attempts at rationalizing their behaviour in terms of processes related to a well-defined state of stress, with a uniform hydrostatic stress component, may not necessarily succeed in capturing the full phenomenology. Thus, in a qualitative way, the map of Figure 5 may still provide a useful picture of what is going on.

The entries in the map referring to MD simulation results highlight one of the mayor obstacles in comparing the simulation results to experiments: while the system sizes in MD simulation have been increasing over the years, to the extent where constraints on the size may have little effect on the predictions for atomic-scale processes, the time-scales remain exceedingly short, and strain rates many orders of magnitude faster than those in standard experiments as well as in industrial forming processes are required for significant deformation to occur within the time of the simulation. Inasmuch as the dominant deformation mechanism is strain-rate dependent, caution is required when concluding on realworld behaviour based on simulation data. This statement does nothing, of course, to diminish the extreme usefulness of MD simulations in stimulating both targeted experimental studies and new theoretical approaches for analysing and understanding the microscopic processes underlying the deformation of nanocrystalline materials. With the advent of improved samples and more systematic studies of deformation, the experimental data base on which this understanding rests is rapidly expanding.

References

1. J. Karch, R. Birringer, H. Gleiter, *Nature* 1987, **330**, 556.
2. A. H. Chokshi, A. Rosen, J. Karch, H. Gleiter, *Scripta Metal.* 1989, **23**, 1679.
3. D. Wolf, V. Yamakov, S. R. Phillpot, A. Mukherjee, H. Gleiter, *Acta mater*. 2005, **53**, 1.
4. R. Z. Valiev, *Nature mater*. 2004, **3**, 511.
5. M. Legros, B. R. Elliott, M. N. Rittner, J. R. Weertman, K. J. Hemker, *Phil. Mag. A* 2000, **80**, 1017.
6. J. Markmann, P. Bunzel, H. Rösner, K. W. Liu, K. A. Padmanabhan, R. Birringer, H. Gleiter, J. Weissmüller, *Scripta Mater*. 2003, **49**, 637.
7. H. Rösner, J. Markmann, J. Weissmüller, *Philos. Mag. Lett.* 2004, **84**, 321.
8. V. Y. Gertsman, R. Birringer, *Scripta Metal. et Mater*. 1994, **30**, 577.
9. Z. Shan, E. A. Stach, J. M. K. Wiezorek, J. A. Knapp, D. M. Follstaedt, S. X. Mao, *Science* 2004, **305**, 654.
10. H. Hahn K. A. Padmanabhan, *Philos. Mag. B* 1997, **76**, 559.
11. J. Löffler J. Weissmüller, *Phys. Rev. B* 1995, **52**, 7076.
12. M. G. Zelin A. K. Mukherjee, *Acta Metall. Mater*. 1995, **43**, 2359.
13. A. Hasnaoui, H. van Swygenhoven, P. M. Derlet, *Phys. Rev. B* 2002, **66**, 184112.
14. J. Weissmüller in: *Proc. 22nd Risø Int. Symp. Mater. Sci.: Science of Metastable and Nanocrystalline Alloys Structure, Properties and Modelling*, Roskilde, Denmark, 2001, p. 155.

15. N. Wang, Z. Wang, K. T. Aust, U. Erb, *Mater. Sci, Eng. A* 1997, **237**, 150.
16. M. Chen, E. Ma, K. J. Hemker, H. Sheng, Y. Wang, X. Cheng, *Science* 2003, **300**, 1275.
17. X. Z. Liao, S. G. Srinivasan, Y. H. Zhao, M. I. Baskes, Y. T. Zhu, F. Zhou, E. J. Lavernia, H. F. Xu, *Appl. Phys. Lett.* 2004, **84**, 3564.
18. V. Yamakov, D. Wolf, S. R. Phillpot, A. K. Mukherjee, H. Gleiter, *Nature Materials* 2002, **1**, 1.
19. P. M. Derlet, A. Hasnaoui, H. van Swygenhoven, *Scripta Mater.* 2003, **49**, 629.
20. Y. M. Wang, E. Ma, *Mater. Sci, Eng. A* 2004, 375–377, **46**.
21. Y. T. Zhu, X. Z. Liao, S. G. Srinivasan, Y. H. Zhao, M. I. Baskes, F. Zhou, E. J. Lavernia, *Appl. Phys. Lett.* 2004, **85**, 5049.
22. V. Yamakov, D. Wolf, S. R. Phillpot, H. Gleiter, *Acta Mater.* 2003, **51**, 4135.
23. E. Ma, Y. M. Wang, Q. H. Lu, M. L. Sui, L. Lu, K. Lu, *Appl. Phys. Lett.* 2004, **85**, 4932.
24. J. Schiotz, K. W. Jacobsen, *Science* 2003, **301**, 1357.

Nanotechnological Aspects in Materials for Hydrogen Storage

Maximilian Fichtner

Introduction

Hydrogen storage is regarded as one of the key issues which have to be solved in the developing field of hydrogen technology [1, 2]. Of the various ways of storing hydrogen for fuel cell-driven applications, however, conventional storage systems based on liquefied and pressurized hydrogen exhibit principal drawbacks.

Pressurized hydrogen can be stored in containers made of composite materials that have to withstand pressures of up to 700 bar in order to carry enough fuel for a reasonable driving cycle of 400–500 km. In this case, however, there are safety concerns related to a tank rupture in an accident. Technical problems arise e.g. from adiabatic effects when expanding and compressing the hydrogen during refuelling. Furthermore, current containers do not allow operation at temperatures below freezing point due to the risk of failure of the composite structure.

To increase the stored mass per container volume, liquefied hydrogen with a density of 70.8 kg/m^3 is particularly attractive. For this purpose, the hydrogen has to be cooled down to 21 K, which, however, needs about one third of the energy content of the stored hydrogen. Overall efficiency is further reduced by a finite heat transfer through the insulated vessel, leading to a loss of 2–3% of vaporized hydrogen per day. In order to prevent a high pressure build-up (the critical temperature of hydrogen is 32 K), the overpressure must be released from the tank, e.g. via a catalytic converter.

A practical solution to circumvent these drawbacks would be storage in a material that can readily take up and release large amounts of hydrogen. According to car manufacturers and the technical targets of the US Department of Energy for 2010 and 2015, more than 6 wt.% of hydrogen should be contained in such a system, including tank and valves, and the filling time should not exceed 5 min. Furthermore, the thermal properties of the material have to match the operation conditions of the fuel cell, which means that the temperature necessary to release the hydrogen from the storage should not exceed the temperature of the exhaust gas of the fuel cell.

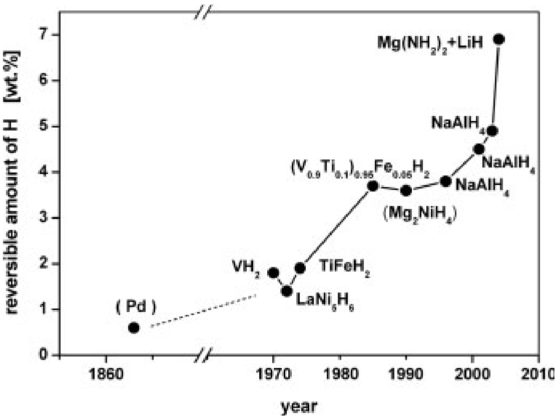

Fig. 1 Development of the gravimetric storage capacity of low- and medium-temperature hydride systems.

Development of such a material is a challenge and up to now no system has been found that would be able to fulfil all requirements. Although considerable progress has been made concerning the storage capacity in the last 30 years (see Fig. 1), the advancements resulting from the investigation of single compounds have not yet led to a solution. Hence, new concepts as well as a multidisciplinary approach are being pursued in order to both increase the reversible storage capacity and the kinetics of hydrogen exchange of systems with suitable thermodynamic properties.

Recent investigations (e.g. [3–6]) have shown that nanoscale materials may offer advantages if certain physical and chemical effects related to the nanoscale can be used efficiently. Associated phenomena, such as surface interactions, material transport, defects, vacancies, phase transitions, grain boundary phenomena, and the formation of new and metastable phases may also play an important role in the development of reversibly working hydrogen storage materials with a high cycling stability.

In practice, two basic mechanisms may be considered for hydrogen storage in nanoscale materials:

- **Physisorption** of molecular hydrogen in a nanomaterial.
- **Chemisorption**, including dissociation of hydrogen molecules and chemical bonding of the hydrogen atoms by integration in the lattice of a metal, an alloy, or by formation of a new chemical compound.

Each principle has its own prospects and constraints. Chemisorption materials may provide high volumetric and gravimetric storage capacities, but there is always the need to split or recombine the hydrogen molecule and chemically bond or release the atoms while forming the compounds. This is a process that inherently involves big enthalpies which have to be handled in a technical application.

Storing hydrogen by physisorption is not subject to this constraint, because the hydrogen stays in its molecular form. The problem rather is to provide light materials with a sufficient amount of bonding sites for the hydrogen per volume. Moreover, physisorption interaction between the H_2 molecule and the surface is in the lower kJ/mol range, as a result of which it may be necessary to go to very low working temperatures.

The present article shall give an overview of the current state of the art of nanomaterials developed for hydrogen storage on the basis of these principles. Material-relevant nanotechnological aspects of preparation and properties shall be outlined and strategies shall be presented for the further development of the materials.

1
Hydrogen Storage by Physisorption

Physisorption is a principle where the forces involved are weak intermolecular forces (van der Waals forces) which do not cause any significant change in the electronic orbital patterns of the relevant species. The challenge associated with the physisorption of hydrogen results from the fact that H_2 is the smallest molecule and only has two electrons. Hence, it is hard to polarize, and dispersion forces created by temporarily induced dipoles are comparably weak.

Despite these peculiarities, nanomaterials may offer advantages for molecular hydrogen storage when they exhibit high specific surface areas and when the chemical and spatial environment of the hydrogen can be designed on a molecular level.

An alternative for storing hydrogen by physisorption on large surfaces is hydrogen storage by encapsulation or trapping in microporous media. The working principle is that the guest molecules are forced into the cavities of the nanostructured host. Upon cooling, hydrogen is trapped inside the cavities. It can be released again by raising the temperature. Thus, trapping may be an instrument to overcome the low binding energies between the molecular hydrogen and the support.

For bonding hydrogen molecules in nanostructures, several approaches have been pursued:

1. Physisorption in materials with large surface areas.
2. Trapping in stable nanostructures.
3. Trapping in temporary nanostructures.

Of the currently investigated physisorption materials, nanostructured carbons, zeolites, metal-organic frameworks, and clathrates shall be discussed below.

1.1
Nanostructured Carbon

Nanostructured carbon materials, such as graphitic nanofibers (GNF), multiwalled carbon nanotubes (MWNT), single-walled carbon nanotubes (SWNT), and carbon nanohorns, exhibit novel properties and yield unusual scientific phenomena, as was shown in a number of studies. The chemistry of the non-planar carbon structures differs from that of the planar graphite, because the non-planar backbone reduces the orbital overlap of the π electrons. This causes localization of the double bonds, which affects the overall chemistry.

A particular interaction of hydrogen molecules with carbon nanomaterials was expected to be possible and in 1998, a storage capacity of 67 wt.% H was reported for graphitic nanofibers [9]. Since then, a number of experimental studies have been conducted, but such high values have not yet been reproduced by other groups. Moreover, the hypothesis of unusually high storage capacities was not supported by theoreticians. According to Wang [8], no graphite-hydrogen potential can account for the values reported.

At the same time, first publications reporting high storage capacities in carbon nanotubes resulted in extensive research activities yielding many experimental data and a similar, still ongoing controversy [10]. Figure 2 gives an overview of data published in this field. Nijkamp et al. [14] and Zuettel et al. [15] found that the amount of adsorbed hydrogen was proportional to the specific surface area (BET) of the carbon material and limited to 2 wt.% hydrogen at 1 bar H_2 and 77 K. Recent studies of hydrogen uptake at 77 K in various carbon nanostructures indicated that the storage capacity depends on the surface area only and not on the long-range

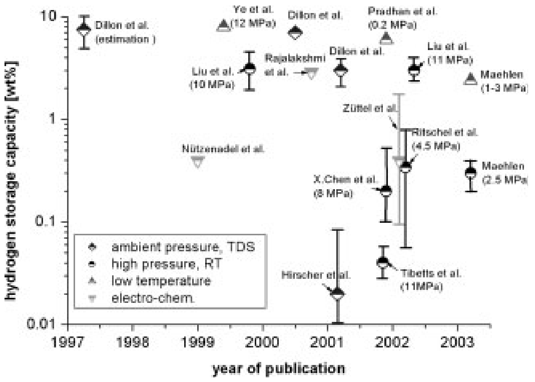

Fig. 2 Experimental data regarding hydrogen storage capacities in SWNTs versus publication year for different methods, pressures, and temperature regimes [11].

order or curvature of the graphene sheets [16]. When cathodically charging GNF, SWNT, and MWNT with hydrogen under ambient conditions, the reversible hydrogen uptake was 1.5 wt.% H per 1000 m^2/g [15].

Several attempts were made to modify carbon materials with the aim of producing designated binding sites for molecular hydrogen at introduced defects and stacking faults. Orimo et al. [12] showed that nanostructured graphite can be made to hold up to 7.4 wt.% hydrogen if the graphite is prepared by 80 h reactive ball milling under a 1.0 MPa hydrogen atmosphere. However, hydrogen desorption starts at about 600 K with maxima around 800 K and 1050K [17, 18]. Hirscher et al. [13, 18] concluded that the uptake of hydrogen was mainly due to the formation of covalent C-H bonds during ball milling rather than to bonding by physisorption at designated binding sites.

Use of a supported catalyst as a hydrogen source may considerably enhance the overall hydrogen uptake of a carbon material, as was reported by Lueking and Yang [7]. In this work, 1–3 $cm^3 H_2/g$ (STP) were adsorbed on various undoped carbon materials (SWNT, MWNT, GNF, and activated carbon) at 300 K and 0.1 MPa H_2. Mixing of carbon nanotubes with palladium nanoparticles (dispersion ~50%) increased the uptake of the carbons by a factor of three. This effect was explained by hydrogen spillover from the catalyst particles to carbon receptors at the surface of the nanotubes. However, baseline carbon adsorption was below 0.1 wt.% and the predominant factor in the magnitude of the overall hydrogen uptake.

A new type of nanoporous carbon with a controlled microporous or mesoporous structure, the so-called carbide-derived carbons (CDCs), was presented recently [105]. The material is produced by a high temperature process from carbides (e.g. SiC or Ti_3SiC_2), by removing metals and metalloids as volatile chlorides, leaving a noncrystalline carbon with up to 80% open pore volume. A regular porous structure is obtained with a narrow pore size distribution depending on the type of carbide and the conditions of chlorination. The specific surface area is up to 2000 m^2/g, the pore size distribution is narrow (0.5–1.5 nm), the pore volume is up to 1 cm^3/g of the material. Up to 3 wt.% hydrogen can be adsorbed at 77 K and 1 bar H_2 pressure. It was concluded in [105] that a uniform and open porosity with pore diameters smaller than 1 nm is the key to storing large amounts of hydrogen by physisorption.

1.2
Zeolites

Reversible occlusion of gases in zeolites is a well-known phenomenon, although little attention has been paid to their use as media for hydrogen storage. Weitkamp et al. [19] found a relationship between the amount of encapsulated hydrogen and the size of the exchanged cation in zeolite A. The number of zeolite cages per gram material was between $3.58 \cdot 10^{20}$ for NaA and NaX and $1.41 \cdot 10^{21}$ for sodalite, and the number of molecules H_2 per cage was between 0.1 and 0.25 for the different materials. Storage capacity was generally higher for zeolites having a high number of small cavities in their structure, and sodalite showed the highest hydrogen uptake of 9.2 cm^3/g (0.082 wt.%) when loaded at 573 K and 10 MPa hydrogen pressure. This implies that only one out of four to five sodalite cages was occupied by a hydrogen molecule. The reason why not every cage takes up at least one hydrogen molecule is unknown. If every cage in sodalite would contain a single hydrogen molecule, a maximum storage capacity of 54 cm^3/g or 0.48 wt.% could be expected.

The storage capacity can be increased substantially when loading zeolite samples at liquid nitrogen temperatures. Kazansky et al. [20] studied hydrogen adsorption on sodium forms of faujasite and found a hydrogen storage capacity of 1.2 wt.% H at 77 K. Moreover, a maximum hydrogen storage capacity of 1.81 wt.% (15 bar) was obtained for NaY zeolite in a recent study [21]. The authors noticed a close correlation between the BET surface area of the zeolite measured with nitrogen and the storage capacity for hydrogen at low temperatures. For all zeolites investigated, a type I isotherm was observed at 77 K, with desorption closely following the same path as adsorption. This indicates the process of microporous hydrogen physisorption in the zeolite structure at low temperatures.

A factor limiting the storage capacity of such systems certainly is the relatively large mass of the zeolite framework itself, which contains Si, Al, O, and heavy cations. Furthermore, many materials seem to have unnecessarily large diameters of their voids. A bottom-up design strategy to improve the storage capacity in microporous systems should therefore include

- a framework consisting of light elements.
- avoidance of nonessential free space in the structure.
- improving the interaction energy between hydrogen and the framework.

1.3
Metal-Organic Frameworks

Metal-organic frameworks (MOFs) are suitable systems for such a bottom-up approach. In their recent work on MOFs, Eddaoudi et al. [22] presented a synthesis strategy based on reticulating metal ions and organic carboxylate links into extended networks. The method uses a self-organizational process and allows for the design of porous structures, in which pore size and functionality can be varied systematically. MOFs are the lightest known crystalline materials with densities down to 0.21 g/cm^3 and have gained considerable interest because of their capability of storing in their framework hydrogen and light hydrocarbons, such as methane [23].

Figure 3(a) shows a cubic MOF structure of $Zn_4O(L)_3$ with Zn_4O clusters forming the corners and linked by linear carboxylates L. The building principle can be extended if other metalions, such as copper, or differently structured carboxylates are used. An example is shown in Figure 3(b), where the structure is formed by a trigonal carboxylate linking Zn_4O clusters.

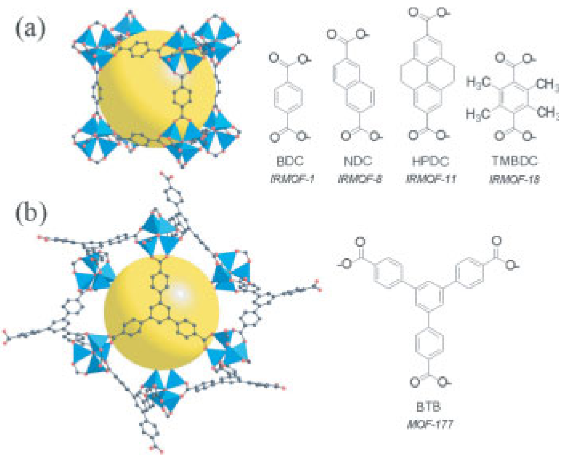

Fig. 3 (a) Metal organic frameworks, $Zn_4O(L)_3$, constructed by linking zinc oxide clusters with linear carboxylates L, such as those shown. (b) The structure of MOF-177, $Zn_4O(BTB)_2$, is formed by linking the same clusters with a trigonal carboxylate. The large void regions are illustrated by yellow spheres with diameters equal to the distance of separation between the frameworks' van der Waals surfaces (reproduced with permission from [24], Copyright 2004 Am. Chem. Soc.).

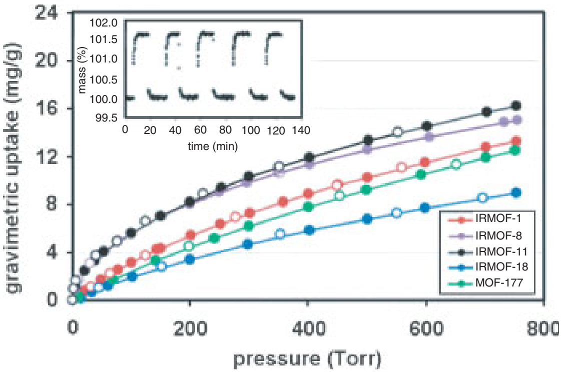

Fig. 4 Volumetrically measured hydrogen isotherms of the activated materials measured gravimetrically at 77 K (adsorption, ●; desorption, ○). The inset shows the time-dependent cycling of IRMOF-11 between 0 and 0.1 MPa of H_2 (reproduced with permission from [24], Copyright 2004 Am. Chem. Soc.).

Sorption isotherms for heavier gases in MOFs are of type I as is typical of microporous materials [24]. Their specific surface areas were calculated from nitrogen sorption isotherms with values ranging between 1466 m^2/g for IRMOF-8 (see Fig. 3a) and 4500 m^2/g for MOF-177 (see Fig. 3b). Isotherms for hydrogen adsorption in different MOF materials are shown in Figure 4. At 77 K and 0.1 MPa H_2 pressure, the maximum uptake of hydrogen was 1.6 wt.% for IRMOF-11. It is evident from this figure that the surfaces

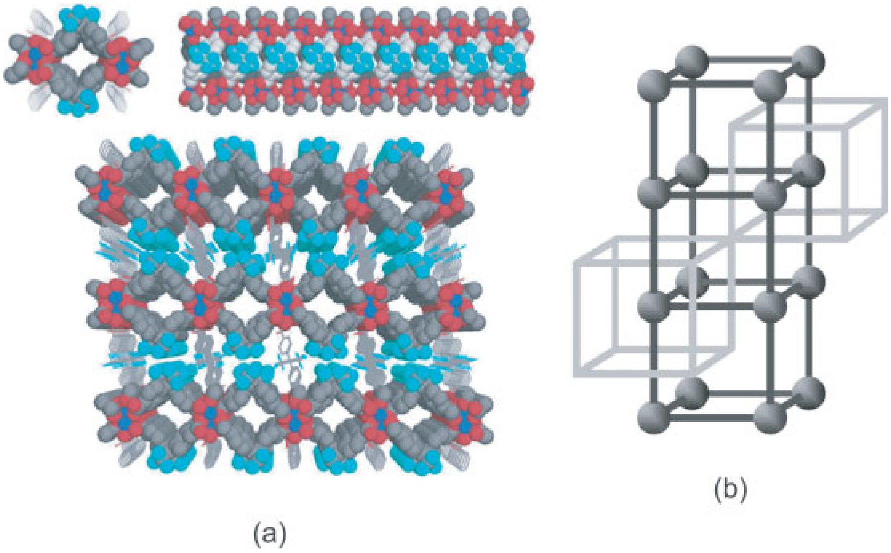

Fig. 5 (a) Structure of Cu-(hfipbb)(*H2hfipbb*)$_{0.5}$[*H2hfipbb* = 4, 4′-(hexafluoroisopropyl idene)-bis(benzoic acid)]. Top left: Cross section of a single microtube. Blue balls, Cu; red balls, O; grey balls, C; light blue balls, F. Top right: Side view of a single microtube. Bottom: Three-dimensional packing of microtubes (reproduced with permission from [26], Copyright 2004, American Chemical Society). (b) Schematic representation of an interpenetrating metal organic framework.

are not saturated by hydrogen at this pressure. Hence, the maximum uptake of each compound was not determined in the experiments. At room temperature and 2 MPa hydrogen pressure, the hydrogen uptake is about 1 wt.% [23].

Interestingly, there is a lack of correlation between the measured surface area and the hydrogen uptake. Apart from preparation effects, the nature of the organic linker was assumed to have a considerable influence on the number of hydrogen molecules attached to a linking unit [24]. This assumption is supported by the results of a theoretical study conducted by Hübner et al. [25] where the interaction energies of molecular hydrogen with different aromatic molecules were studied in ab-initio calculations. It was found that the binding energies between H_2 and differently substituted benzene rings vary systematically depending on the nature of the functional group. While electron-drawing groups lower the interaction energy, electron-pushing groups increase the values. Moreover, larger aromatic systems are superior to single benzene rings.

In the current state of the art, not much is known about the behavior of hydrogen in MOFs. Although the organic linkers seem to play an important role as binding sites, the role of the metal oxide clusters has not yet been elucidated. As the dynamics of hydrogen is another important factor, a critical aspect of pore quality is the relative size of the pores with regard to the hydrogen and its interaction energy with the pore's surface. As a design strategy, the pore diameters could be minimized to both increase the number of pores and the number of binding sites per volume, which would also enhance the volumetric storage density of the material.

For the practical implementation of the concept, other metal-organic frameworks with higher densities can be synthesized, an example was shown by Pan et al. [26] see Figure 5(a). Another approach is the synthesis of interpenetrating frameworks, as is shown in Figure 5(b). In this case, there would not be any limitation to using small linker molecules which might have less favorable properties concerning their interaction energy with hydrogen.

In order to provide for a spatial obstruction of the pores, Zhao et al. [29] synthesized nickel-based metal-organic frameworks with smaller channels and found that some of them exhibited a remarkable hysteresis in the adsorption and desorption of hydrogen at 77 K. 57 to 71% of the total pore volume of 0.149 and 0.181 cm^3/g H_2 were filled in two different structures with molecular-size windows. Of three different materials, the compound with the most distinctive isotherm hysteresis had the smallest window dimensions (smaller than the kinetic diameter of 0.29 nm of H_2) and cavities that were much larger than H_2.

Hysteresis decreases at higher temperatures and was therefore explained by kinetic trapping of the gas. This is supposed to result from the flexibility of the organic ligands (4, 4′-bipyridine) in the nickel-based MOF. It was assumed that desorption is kinetically limited at lower temperatures, so that hydrogen can be loaded at high pressures, but then stored at lower pressures.

1.4
Clathrates

Mao et al. [27] reported about a new type II hydrogen clathrate structure, where H_2 and H_2O mixtures crystallize at a pressure of 220 MPa and 250 K with an approximate H_2/H_2O molar ratio of 1:2. At higher temperatures, extremely high pressures are required to stabilize the material, e.g. 300 MPa at 280 K. The clathrate cages formed are multiply occupied, with a cluster of two H_2 molecules in the small cage and four in the large cage, see Figure 6. The quenched clathrate was found to be stable up to 145 K at ambient pressure. Raman spectroscopy revealed that the H_2 molecules in the cages still were in free rotational states. However, softening of the H–H stretching modes indicated increased intermolecular interactions between hydrogen and water molecules in the cage.

In a theoretical study, Patchovskii and Tse [30] essentially confirmed the experimental results and suggested that the stability of the clathrate was mainly caused by dispersive interactions between H_2 molecules and the water forming the cage walls. According to their calculations, the clathrate is thermodynamically stable at 2.5 MPa and 150 K.

Due to the high pressures needed for the formation of H_2/H_2O clathrates and their stability at low temperatures only, it would be favorable to further stabilize the structures and facilitate their formation. As a first promising result, Florusse et al. [28] showed that the presence of tetrahydrofuran as a guest component may indeed stabilize the clathrate at 280 K and 5 MPa pressure. PXRD measurements indicated that the same sII binary clathrate hydrate had formed as reported by Mao et al. [27] Obviously, the system is strongly sensitive to promoters, which gives room for further development.

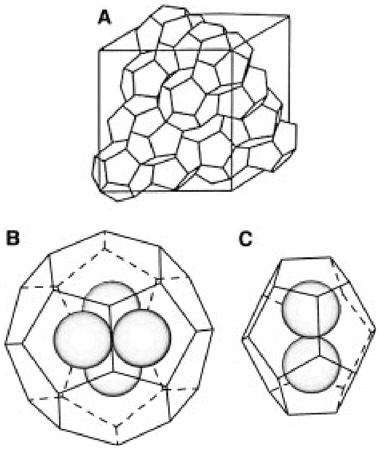

Fig. 6 The sII clathrate crystal structure consisting of $5^{12}6^4$ and 5^{12} building blocks. (B) A tetrahedral cluster of four hydrogen molecules in the $5^{12}6^4$ cage. (C) A cluster of two hydrogen molecules oriented towards opposite pentagonal faces in the $5^{12}6^4$ cage. The H_2O frameworks define the edges of the polyhedrons, with an oxygen atom at each corner and a hydrogen atom on each edge. The $5^{12}6^4$ and 5^{12} cages are drawn to scale relative to the enclathrated hydrogen molecule (sphere) of 2.72 Å diameter (reproduced with permission from [27], Copyright 2002, American Association for the Advancement of Science).

2
Hydrogen Storage by Chemisorption

The common way of storing hydrogen in solids is the chemisorption of dissociated hydrogen in noble metals [31] and alloys. This method is suited for obtaining high volumetric capacities due to the short hydrogen–hydrogen distances that can be realized in a hydride material. The minimum H–H separation in ordered binary metal hydrides is 0.21 nm, this limit being due to the repulsive interaction generated by the partially charged hydrogen atoms [32].

While the volumetric hydrogen density in transition metals and alloys is high, gravimetric density is inherently low due to the relatively high mass of the metal atoms (see Fig. 1). Consequently, systems consisting of light elements have moved into the focus of research and a new field has been opened up by investigating the reversible properties of complex hydrides $M^{+x}(BH_y)_x$, with M = Li, Na, K, Be, Mg, Ca,. . ., and B = B, N, Al. These so-called boranates, amides, and alanates are salt-like materials and consist of a cation

M^{+x} and a complex anion $BH_y{}^{-1}$, where the hydrogen atoms are covalently bound to the central atom B.

Nanoscale effects have been found for a number chemisorption materials. They are mostly related to:

1. The reversible hydrogen content,
2. The thermodynamic properties which define the theoretical working parameters of the material, i.e. the pressure/temperature combinations where hydrogen can be absorbed and desorbed, and
3. The kinetic properties that determine the rate at which hydrogen can be exchanged at a given temperature and pressure.

The following selected examples refer to these topics and were chosen in order to discuss recent findings and developments in this field. For further information, several excellent reviews on hydrogen in metals [33], intermetallic compounds [34, 35], and complex hydrides [36] can be found in literature.

2.1
Metal Clusters

Nanocrystalline metals and metal-hydrogen alloys with typical grain sizes of less than 20 nm possess many properties that differ from those of conventional coarse-grained materials. As an example, the thermodynamics of Pd-H clusters have been studied by several groups both theoretically and experimentally. Their hydrogenation behaviour was found to differ from that of coarse-grained Pd. Figure 7 shows isotherms of differently prepared Pd-H clusters of similar diameter, compared to bulk data [38]. The cluster data exhibit a remarkable increase in the hydrogen solubility in the solute solution phase (α-) and decrease in the hydride (α'-) phase, e.g. a considerable narrowing of the two-phase region.

Eastman et al. [39] reported that the entire volume of nanocrystalline samples was transformed to α'-PdH when exposed to hydrogen and suggested that the observed narrowing of the miscibility gap was due to different entropies and/or enthalpies of mixing in nanocrystalline and coarse-grained samples.

Pundt et al. [37, 38] proposed that an outer shell of the grains or particles would not participate in the phase transition, whereas the phase diagram of the core remains unchanged relative to that of the coarse-grained alloy. An enrichment of hydrogen at the surface is also supported by the results of Monte-Carlo simulations for hydrogen-loaded, multi-facetted Pd clusters of about 2 nm in size [40].

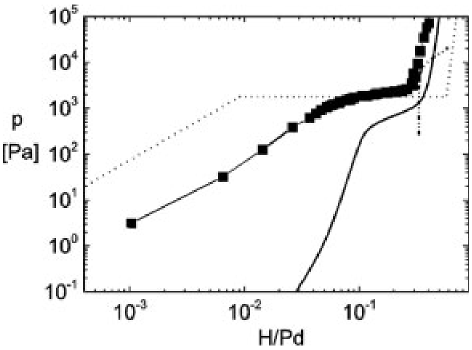

Fig. 7 Isotherms for Pd-H clusters of similar diameter (– compacted clusters d = 3.9 nm, 293 K, ■ quasi-free clusters d = 3.5 nm, 310 K, and bulk data (. . .) (reproduced with permission from [37], Copyright 2004, WILEY-VCH Weinheim).

Weissmüller et al. [41] showed that the observed lowering of the critical temperature was a function of the grain size [42] in Pd-H and the changes in the miscibility gap in the core could be explained solely by thermodynamic effects. It was found that hydrogen enrichment and depletion at grain boundaries of compacted nc $Pd_{81}Au_{19}$ result in a strongly size-dependent interface stress that is compensated by a pressure in the bulk [43]. Hence, long-range elastic interactions of grain boundaries and crystallites may shift the thermodynamic equilibrium in the crystallites and influence both the solubility and the total hydrogen content of hydrogen storage materials. As grain size decreases, interface stress must be taken into account for an appropriate determination of the hydrogen content from the lattice parameter and the prediction of the decrease in the critical temperature [44].

Thermodynamic effects were also observed in other materials, e.g. nanocrystalline $LaNi_5$ [46] and FeTi [47]. Here, an enhanced solubility of hydrogen at low pressures was found [48], similar to Pd. Moreover, the two-phase region of nc Fe-Ti is narrowed compared to coarse-grained FeTi.

2.2
Nanocomposites

For the preparation of nanocomposites, mechanical alloying has the advantage of simplicity as preparing technique which offers a free choice of various kinds of alloys [45, 48]. In addition, the technique allows for the mass production of non-equilibrium phases of alloys over a wide range of composition. It was also found that a number of difficulties in synthesis and processing of metal hydrides could be solved by applying this method.

In the first instance, improvements were made with respect to the so-called ''activation'' of hydride materials. This is a process, by means of which possibly present oxidic barriers at interfaces are broken by a mechanical treatment and high isotropic and anisotropic lattice strains are induced in the material so that a high density of dislocations in the order of $10^{12}\,\mathrm{cm}^{-2}$ is produced [49]. This promotes the transformation of the alloy, as a forming hydride phase may not coherently precipitate in a perfect lattice due to the large elastic energy associated with precipitation and the small formation enthalpy for hydrides. When a virgin sample of e.g. $LaNi_5$ is hydrogenated, hydride precipitation is considered to occur at free surfaces and at lattice defects that thread the surface where the elastic strain is largely relieved [50]. The formation of dislocations may therefore play a key role in the phase transformation during hydrogenation and dehydrogenation

A number of hydride materials that could be synthesized under severe conditions or in a complicated process only have been made accessible by mechanical alloying of the constituting elements. For example, the composition of raw materials is important when synthesizing $LaNi_5$, because La and Ni possess different melting vapor pressures. Moreover, the cooling process has to be controlled carefully to avoid a peritectic reaction and to obtain the product with the right composition. At the same time, the product can be produced easily by ball milling [51, 52]. A material synthesized in such a manner exhibits improved hydrogenation and dehydrogenation kinetics [53], see Figure 8. It was also shown that a nanocomposite produced by ball milling of the immiscible metals Fe and Mg exhibits better hydrogenation kinetics and reversible capacities for hydrogen than the same material produced by sintering [54]. Thus, the activation temperature for Mg_2FeH_6 could be lowered by 100–150 K.

A further developed method for the synthesis of ternary hydrides is reactive ball milling, where the primary materials are milled under a hydrogen atmosphere at elevated pressures. A hydride phase forms during the process, which is more brittle than the raw materials. It may facilitate the milling process and speed up synthesis. A number of complex and interstitial hydrides have been produced in this way, for example Mg_2FeH_6 [55], Mg_2NiH_4 [56], $TiFeH_x$ [57], and $Zr_{1-\delta}NiH_x$ [58].

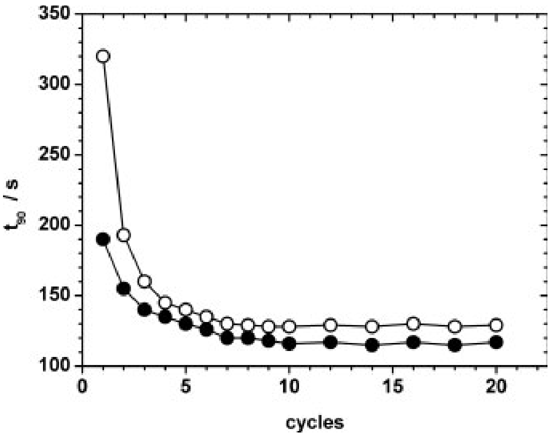

Fig. 8 Time of 90% of the hydrogenation reaction at room temperature versus the cycle for (●) the 5 h milled $LaNi_5$ powder, (○) the unmilled $LaNi_5$ powder (reproduced with permission from [53], Copyright 1999, Elsevier Science S.A.).

In a further attempt, Huot et al. showed that Mg_2FeH_6 can also be produced directly by mechanical alloying of the preformed hydride MgH_2 and Fe [59]. Synthesis of alkali metal and strontium alanates is feasible using the same principle according to Dymova et al. [60, 61] while adding a Ti-based dopant such as $Ti(OBu)_4$ [62] enhances the hydrogen exchange kinetics of the material produced. In addition, a number of solid-state reactions may take place in a ball mill to yield hydrides that can hardly be produced by other methods, such as: [63–65]

$$2\,\mathrm{NaH} + \mathrm{NaAlH_4} \longrightarrow \mathrm{Na_3AlH_6} \tag{1}$$

$$2\,\mathrm{NaH} + \mathrm{LiAlH_4} \longrightarrow \mathrm{Na_2LiAlH_6} \tag{2}$$

$$2\,\mathrm{NaAlH_4} + \mathrm{CaCl_2} \longrightarrow \mathrm{Ca(AlH_4)_2} + 2\,\mathrm{NaCl} \tag{3}$$

$$2\,\mathrm{AlH_3} + \mathrm{MgH_2} \longrightarrow \mathrm{Mg(AlH_4)_2} \tag{4}$$

A particular advantage of mechanical alloying is the opportunity to blend dopants with the hydrogen carrier materials, thus forming a nanocomposite and reducing kinetic barriers associated with hydrogen chemisorption and/or phase transformation. This is advantageous for practical use, as in some cases, e.g. in automotive applications, short refuelling times are needed, which can only be achieved when using such a nanocomposite. The following chapters address this emerging field in more detail, the focus being on a light metal hydride and alanates.

2.2.1 Magnesium Hydride

Magnesium is of particular interest to hydrogen storage applications because of its low costs and high content of 7.6 wt.% hydrogen. However, hydrogenation of untreated Mg is very slow and numerous attempts have been made to both increase the hydrogenation rate and decrease the desorption temperature of 573 K. This high temperature is mostly due to the high enthalpy of formation of −76 kJ/mol. On the other hand, these properties make magnesium hydride a favorable material e.g. for thermal and solar-thermal applications.

The absorption and desorption kinetics may be enhanced considerably by simple ball milling of pure MgH_2, and a correlation between the transformation rate and the grain size of Mg was reported by Zaluska et al. [66] Figure 9 shows that a freshly ball-milled Mg sample with a grain size of 1 μm is virtually inert to hydrogen, whereas material with grain sizes smaller than 50 nm can be hydrogenated to 80% of the theoretical capacity within 120 min only. Dehydrogenation curves of both unmilled and milled samples exhibit a sigmoidal shape, which suggests that a nucleation and growth process governs the transformation rate. The effect of milling was explained qualitatively by an increasing influence of grain boundaries and enhanced diffusion of hydrogen.

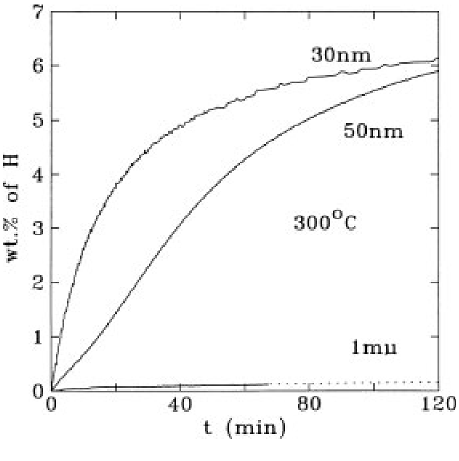

Fig. 9 Effect of grain size on hydrogen absorption of ball-milled magnesium powder: Absorption at 573 K, first cycle, no activation (reproduced with permission from [66], Copyright 2001, Springer-Verlag GmbH).

To further improve the H exchange kinetics, composite materials were prepared by ball milling of Mg with a series of transition metals either absorbing hydrogen (e.g. V, Nb, Pd, Pt[68, 69]) or not (e.g. Ni, Fe, Cu, Co [71, 70]). V, Nb, and Pd do not alloy with Mg, and small hydrogenated particles in the composite are supposed to act as gateways for the hydrogen to the magnesium matrix. For V- and Nb-doped nanocomposites, the dehydrogenation rate at temperatures below 573 K was again found to be limited by nucleation and two-dimensional growth. At higher temperatures, the reaction is interface-controlled, with a two-dimensional growth of the forming Mg phase [67].

Metal oxides may have a strong catalytic effect, even in small amounts, as was found for the oxides of Ti, V, Cr, Mn, Fe, and Mo [72, 73]. The role of the transition metals has not been settled yet, but the formation of hydrides was observed in the case of TiH_x and VH_x. They may act as a hydrogen donor [72]. Other oxides are known to incorporate hydrogen and form a bronze, such as H_xWO_3. Hydrogen bronzes provide for a high hydrogen mobility that might lead to a spillover effect in contact with Mg/MgH_2. However, the detailed mechanism is not known and subject of research at the moment.

2.2.2 Alanates

Due to their high storage capacity, aluminum hydrides or alanates have been considered promising materials for fuel cell applications. Alanates belong to the so-called complex hydrides and are composed of a "complex anion" $[AlH_4]^-$ or $[AlH_6]^{3-}$ and the respective cation. They are salt-like, insulating materials. The bond between the hydrogen atoms and aluminum can be regarded as covalent with a strong ionic character [75], the bond to the cation is ionic. Here, they differ from classical metal hydrides, where the H atoms are incorporated in tetrahedral and octahedral holes of a metallic host lattice. Nevertheless, similarities of both types of materials are observed in principle, for example when they are ball-milled and/or when dopants are added.

Several alanate compounds containing light cations have been investigated for hydrogen storage purposes 74, 76–78. Of these candidates, sodium alanate ($NaAlH_4$) currently is the most investigated and promising hydrogen carrier due to its high hydrogen content and auspicious thermodynamic properties. The pure macrocrystalline compound melts at 456 K and the melt starts to decompose and release H_2 at around 513 K, forming the hexahydride Na_3AlH_6 and elemental Al as solid decomposition products. Na_3AlH_6 decomposes at temperatures above 573 K and releases H_2 so that a mixture of two solid phases, NaH and Al, is obtained; see Eqs. (5) and (6). The decomposition of NaH occurs at temperatures above 723 K, which is considered too high for practical applications, the maximum reversible storage capacity for hydrogen therefore is 5.5 wt.% H [79]:

$$3\mathrm{NaAlH_4} \leftrightarrows \mathrm{Na_3AlH_6} + 2\mathrm{Al} + 3\mathrm{H_2}$$

$$(+\Delta H = 3 \times 37\ \mathrm{kJ/mol};\ 3.7\ \mathrm{wt.\%H}) \tag{5}$$

$$\mathrm{Na_3AlH_6} \leftrightarrows 3\mathrm{NaH} + \mathrm{Al} + 3/2\mathrm{H_2}$$

$$(+\Delta H = 47\ \mathrm{kJ/mol};\ 1.8\ \mathrm{wt.\%H}) \tag{6}$$

When the material is ball-milled, desorption already starts below the melting point of $NaAlH_4$, which is probably due to the lowering of kinetic barriers [80]. Although milling leads to such a pronounced effect, however, no thorough study has been performed so far to explain the reasons of this behavior. In particular, it is not yet clear whether the effect is caused by an increasing influence of interfaces and grain boundaries, by introduced defects, or by a combination of reasons.

A strong influence of doping agents was also found for the alanates. Bogdanovic and Schwickardi [74] discovered that the addition of small amounts of Ti compounds may considerably accelerate H exchange in $NaAlH_4$. Thus, working temperatures around 373 K can be achieved, which fits to the working conditions of a PEM fuel cell. A variety of metals have been tested as dopants in the meantime [74, 81] and Ti was always found to have the best catalytic activity for $NaAlH_4$. In the first work, Ti alkoxides were used in a wet impregnation process to form a finely dispersed layer of small Ti particles on macrocrystalline $NaAlH_4$. Later, it was shown that the simple addition of Ti alkoxides or halides to the alanate and ball milling of the mixture under inert conditions lead to improved kinetics and a simplification of the process 82–84. Another considerable increase in the transformation rate was achieved recently by using a solvent-stabilized Ti_{13} cluster as dopant [85, 86].

Figure 10 shows the isothermal hydrogen desorption behavior of different composite materials, as obtained from kinetic studies by Kircher et al. [87] The samples were prepared by ball milling of sodium alanate for 30 min under an argon atmosphere with different precursors and decomposed isothermally at 423 K. For comparison, the data of a ball-milled undoped sample are shown. The diagram reveals the slow decomposition of the undoped alanate despite the higher decomposition temperature of 443 K and indicates the advantage of using nanoscale Ti as a precursor in the doping process.

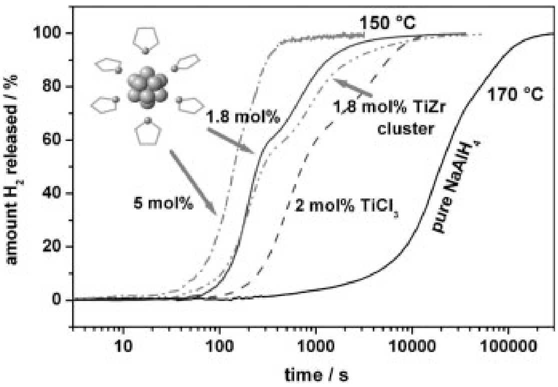

Fig. 10 Isothermal desorption behavior of different composite materials, all based on $NaAlH_4$ as hydrogen carrier, milling time 30 min. The doped samples were decomposed at 423 K in a vacuum.

A similar distinct improvement was observed for the rehydrogenation rates of dehydrogenated composites. Figure 11 shows the isothermal hydrogenation behavior after the third hydrogenation/ dehydrogenation cycle. As can be seen from the diagram, it was possible to refuel the material to 80% of the total absorption of about 4.5 wt.% H within only one minute, which would meet the refuelling time target of the US-DoE for 2015.

An interesting feature of both reversible transformation Reactions (5) and (6) is that a massive mass transport is involved, which covers distances in the range of up to 100 nm due to the grain sizes of the different phases. For illustration, Figure 12 shows a STEM of decomposed material after the 8th cycle, where the solid phases involved in the process have been identified by electron microscopy [89]. If such a decomposed sample is rehydrogenated, the separated solid phases NaH and Al have to establish a reactive interface to form Na_3AlH_6 in the presence of hydrogen. In the second step,

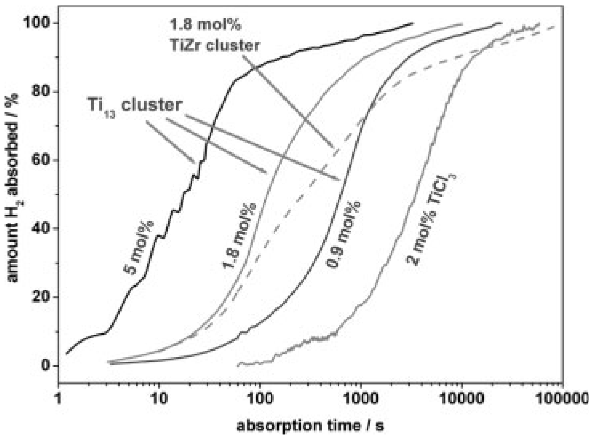

Fig. 11 Isothermal absorption behavior of differently doped alanate samples (milling time 30 min, absorption temperature 373 K, H_2 pressure 10 MPa).

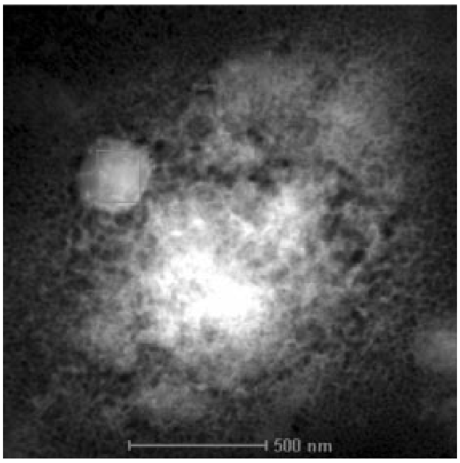

Fig. 12 STEM of $NaAlH_4$ doped with 5 mol% Ti clusters and desorbed after the 8th dehydrogenation/hydrogenation cycle. The encircled area contains a region of pure Al (white).

Na_3AlH_6 and Al react with H_2 and form $NaAlH_4$. Both reactions occur within minutes at temperatures around 370 K.

In order to examine the mechanisms underlying this process, the kinetics of the phase transformations were analyzed by a separate investigation of the various hydrogenation and dehydrogenation steps of Eqs. (5) and (6). For this purpose, a reactor and an experimental setup were used to minimize the effects of the reaction heat and thermal conductivity of the material [85, 87, 88]. As a result, the kinetic data of both absorption and decomposition curves could be fitted well over a wide range by applying a Johnson-Mehl-Avrami equation that describes the reacted fraction F when the rate-limiting process for the formation of the new phase is nucleation and growth (Eq. 7).

$$F = 1 - \exp\{-(kt)^{\eta}\} \tag{7}$$

The Avrami exponents η obtained were small and reflected diffusion-limited nucleation and growth of the new phase as a rate-determining step in the kinetics. In samples doped with higher amounts of Ti, the Avrami exponent changed during absorption, probably due to overlapping diffusion profiles of neighboring grains.

Several suggestions were made to explain the accelerating influence of Ti dopants on the transformation of the composites. It was proposed for example that Ti might facilitate the formation or decomposition of potentially mobile AlH_3 at grain boundaries and AlH_3 could transport the aluminum over larger distances [90]. As this species has not been detected up to now, more work is needed to prove or disprove the existence and potential role of AlH_3.

It was also proposed that Ti metal, a solid solution of Ti in Al, or a Ti-Al alloy is catalytically active at the surface of doped, ball-milled alanate. There, it could act as chemisorption site for the dissociation or recombination of molecular hydrogen. Indeed, chemisorption of H_2 on Ti-doped $NaAlH_4$ was found in isotope scrambling experiments [36] and Ti-Al phases were detected by EXAFS and XRD measurements 91–93. However, it is not clear at the moment whether chemisorption at the Ti-containing phases plays a decisive role in the kinetics. An addition of $TiAl_3$ as dopant and ball milling led to a slight improvement of the reaction rates only [91], but a certain influence cannot be excluded and further work is necessary to elucidate the role of these phases.

Another hypothesis proposed by Fichtner et al. [94] is based on the kinetics of the phase transformations and on data obtained with different dopants. As shown above, the rate-limiting steps of the phase transformations of sodium alanate and the associated hydrogen absorption or desorption are governed by nucleation and growth processes which are limited by the diffusion of the phases involved. Therefore, it was concluded that Ti-based dopants must have an accelerating influence on exactly these steps. A potential role of Ti in the acceleration of the transport processes could be the reduction of activation barriers for the mobility of atoms, e.g. by introducing defects. In fact, recent studies indicate that Ti may indeed exchange with Al in the alanate lattice, thus creating non-thermal defects and vacancies due to the differences in size and charge of the atoms. According to density functional calculations by Lovvik and Opalka [95], for example, Ti may replace Al close to the surface, thus leading to a metastable state of the system.

Apart from this potential function of the dopant, NMR experiments revealed that the mobility also of H atoms was enhanced in the nanocomposite after doping [96]. Provided that all these steps are necessary for the overall transformation reaction, it is likely that the process is a consecutive reaction of fast chemisorption and hydrogen diffusion steps and slower, rate-limiting diffusion processes in the solid. This would also explain why noble metal catalysts that have much better chemisorption capabilities than Ti were found to have no advantage for alanate materials – because they cannot exchange with Al in the lattice and the chemisorption step already is faster than the material transport. Recent studies of isotope effects in the rate limiting steps of the materials transformation suggest a species heavier than H as being the mobile species [106]. The data rather supports species such as Al_xH_y which could be responsible for the Al transport.

It should be mentioned that quantitative information about the different transformation steps in the nanocomposites is still lacking. Consequently, the conclusions drawn above are of preliminary character only. Again, more mechanistic work is needed in order to learn more about the details of hydrogen exchange and the role of the dopants in these materials. The outcome will be essential for the further development of the alanates as hydrogen storage materials.

2.2.3 Other Complex Hydrides

Other complex hydrides containing light elements, such as the boranates and amides, have a very high gravimetric hydrogen content of up to 18.4 wt.% ($LiBH_4$). First studies of their reversible hydrogen storage properties have been published in literature. Although promising because of their high hydrogen content, however, the boranates exhibit a high thermodynamic stability due to the strong B–H bond, and the working temperatures and pressures achieved in dehydrogenation/hydrogenation experiments are too high for current PEM fuel cell applications [97]. Furthermore, it is known from early works on boranates that volatile boranes may develop upon the decomposition of the solid. The gases are toxic and may poison the catalytic membranes of the fuel cell. Future work in this field will therefore be aimed at finding ways of destabilizing the complex hydride and avoiding the production of boranes. $Mg(BH_4)_2$ is one of the few complex hydrides which have the potential to meet the requirements for hydrogen storage materials, because it contains 14.9 mass% H and has suitable thermodynamic properties. It can be synthesized by various methods and is currently investigated for hydrogen storage applications ([107] and references therein).

In search of other systems for hydrogen storage, Chen et al. [98] presented a work on the reversible hydrogenation behavior of lithium nitride which transforms stepwise, similar to the alanates:

$$Li_3N + 2H_2 \leftrightarrows Li_2NH + LiH + H_2 \leftrightarrows LiNH_2 + 2LiH \quad (8)$$

The amount of H_2 absorbed at Li_3N is 10.4 wt.% in this reaction, which is a very high value. However, the overall heat of reaction is −161 kJ/mol and operation temperatures above 520 K were reported, which is a daunting value for low-temperature fuel cell applications. Nonetheless, successive activities have led to an interesting and promising new approach, as will be shown in the next chapter.

2.2.4 Combined Systems

Shortly after the work of Chen et al. [98] other groups presented their work on amide systems [101, 102]. Ichikawa et al. [99] pointed out that the enthalpy change of the second reaction of Eq. (7) is −44.5 kJ/mol only, which is almost ideal from the thermodynamic point of view, because the free enthalpy of reaction is close to zero at room temperature. To perform the reaction, they produced composite materials containing $LiNH_2$ and LiH by ball milling and investigated the hydrogenation/dehydrogenation behavior according to Eq. (9), with a theoretical amount of 6.5 wt.% H being involved.

$$Li_2NH + LiH + H_2 \leftrightarrows LiNH_2 + 2LiH \quad (9)$$

Small amounts of $TiCl_3$ as dopant were found to have a strong accelerating effect on the reaction and it was suggested that it might act as a transfer agent for intermediate NH_3 to LiH. In a subsequent work, it was stated that the uniform

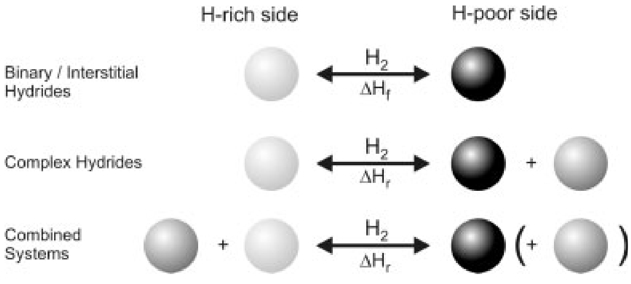

Fig. 13 Solid phases involved in the hydrogen exchange of different classes of storage materials.

distribution of nanometer-sized Ti metal between LiH and $LiNH_2$ plays an important role for the catalytic effect [100].

Compared to the binary and interstitial hydrides and the alanates, reactions of the type of Eq. (9) open up a new field in the development of hydrogen storage materials. For illustration, a simplified classification scheme of hydride systems shows a distinction of three classes with increasing degree of complexity, see Figure 13. The first class of systems includes the binary and interstitial metal hydrides, where the reaction enthalpy is the same as the enthalpy of formation and basically one H-rich and one H-poor solid phase are involved in the transition. The next, more complicated situation is found for metal hydrides, including the alanates, where one H-rich and two H-poor phases are involved. Finally, the novel, combined systems consist of two H-rich and one or two H-poor phases. The systems may be considered solid-state reactions of binary (or interstitial) metal hydrides with complex metal hydrides, under the exchange of hydrogen.

Even though this is an additional degree of complexity, the approach provides for a number of opportunities to design a storage system. In principle, many combinations of binary and complex hydrides are conceivable, so that an optimization can be made with respect to the exchangeable amount of hydrogen and the thermodynamics, i.e. reaction enthalpy and reaction entropy. Even compounds with originally unfavorable thermodynamic properties for low-temperature fuel cell applications, such as MgH_2, may be considered again when combined with an appropriate reaction partner. An example was published recently by Luo [103] where a nanocomposite was prepared by high-energy ball milling of a MgH_2:$LiNH_2$ mixture with a molar ratio of 1.1 : 2. The reversible behavior of the material was tested at temperatures between 473 K and 573 K and 6–6.5 wt.% H were gained upon absorption and desorption. At 473 K, hydrogen absorption took place at 3.2 MPa hydrogen pressure. Leng et al. [104] investigated a composite material produced by ball milling of a 3 : 8 molar mixture of $Mg(NH_2)_2$ and LiH under a 1 MPa H_2 atmosphere. The results showed that a large amount of hydrogen (approx. 7 wt.%) started to be desorbed at 413 K, and desorption peaked at 463 K. Hydrogenation was achieved, e.g. at 473 K and 3 MPa H_2 pressure. Based on X-ray diffraction measurements of the material in different stages, the following reversible reaction was proposed (Eq. 10).

$$3Mg(NH_2)_2 + 8LiH \leftrightarrows Mg_3N_2 + 4Li_2NH + 8H_2 \qquad (10)$$

The challenges associated with this new type of systems are the number of solid phases that are involved in the transformations. At lower working temperatures, severe kinetic barriers may interfere with the phase transformations and, therefore, with the H exchange. In any case, a transport in the solid and an interface reaction are necessary and means like

- shortening the diffusion path,
- enhancing the mobility, e.g. by introducing defects, and
- an intimate contact along the reactive phase boundaries

may be considered prerequisites for a properly and fast working system. All these requirements are influenced by the particle size and structure of the grain boundaries.

Accordingly, a nanotechnological approach is deemed advantageous for the development of this promising new type of systems.

3
Conclusion

Hydrogen storage in solids is, in the medium and long term, considered to be one of the most promising and challenging methods for the realization of hydrogen storage systems in fuel cell-driven mobile applications. Recent advances have given rise to the hope that the targets defined for 2010 or even 2015 could be reached, provided that the progress made in the different areas can be continued. A particular challenge is the combination of targeted storage capacities, thermodynamic properties, and hydrogen exchange kinetics of storage materials, thus opening up a multi-faceted research field that makes the use of a multidisciplinary and interdisciplinary approach mandatory. Recent developments have shown that a nanotechnological approach may be very effective in this field, because it combines physics, chemistry, and engineering science and closely looks at the microstructure of materials, thus giving clues for a tailor-made development by optimizing the bonding environment for molecular or atomic hydrogen.

References

1. L. Schlapbach, A. Zuettel, *Nature* 2001, **414**, 353.
2. J. A. Ritter, A. D. Ebner, J. Wang, R. Zidan, *Materials Today* 2003, **6**, 9, 18.
3. Several authors in *Nanoclusters and Nanocrystals,* (ed. H. S. Nalwa). American Scientific Publishers, Stevenson Ranch, Calif. 2002.
4. X. He, D. M. Antonelli, *Angewandte Chem. Int. Ed.* 2002, **41**, 214–229.
5. H. Gleiter, *Scripta Mater.* 2001, **44**, 1161.
6. J. D. Aiken, R. G. Finke, *J. Mol. Cat.* 1999, **1**.
7. A. D. Lueking, R. T. Yang, *Appl. Cat. A* 2004, **256**, 259.
8. Q. Wang, J. K. Johnson *J. Phys. Chem. B.* 1999 **103**, 277.
9. A. Chambers, C. Park, R. T. K. Baker, N. M. Rodriguez, *J. Phys. Chem. B* 1998, **102**, 4253.
10. M. Hirscher, M. Becher, *J. Nanosci. Nanotech.* 2003, **1/2**, 1.
11. M. Hirscher, private communication.
12. S. Orimo, G. Majer, T. Fukunaga, A. Züttel, L. Schlapbach, H. Fujii, *Appl. Phys. Lett.* 1999, **75**, 3093.
13. M. Hirscher, M. Becher, M. Haluska, A. Quintel, V. Skakalova, Y.-M. Choi, U. Dettlaff-Weglikowska, S. Roth, I. Stepanek, P. Bernier, A. Leonhardt, J. Fink, *J. Alloys Comp.* 2002, **330–332**, 654.
14. M. G. Nijkamp, J. E. Raaymakers, A. J. van Dillen, K. P. de Jong, *Appl. Phys. A* 2001, **72**, 619.
15. A. Zuettel, P. Sudan, Ph. Mauron, T. Kiyobayashi, Ch. Emmenegger, L. Schlapbach, *Int. J. Hydrogen Energy* 2002, **27**, 203.
16. H. G. Schimmel, G. J. Kearley, M. G. Nijkamp, C. Visser, K. de Jong, F. M. Mulder, *Chem. Eur. J.* 2003, **9**, 4764.
17. S. Orimo, T. Matsushima, H. Fujii, T. Fukunaga, G. Majer, *J. Appl. Phys.* 2001, **90**, 1545.
18. M. Hirscher, M. Becher, M. Haluska, F. von Zeppelin, X. Chen, U. Dettlaff-Weglikowska, S. Roth, *J. Alloys Comp.* 2003, **356–357**, 433.
19. J. Weitkamp, M. Fritz, S. Ernst, *Int. J. Hydrogen Energy* 1995, **20**, 967.
20. V. B. Kazansky, V. Y. Borovkov, A. Serich, H. G. Karge. *Microporous Mesoporous Mater.* 1998, **22**, 251.
21. H. W. Langmi, A. Walton, M. M. Al-Mamouri, S. R. Johnson, D. Book, J. D. Speight, P. P. Edwards, I. Gameson, P. A. Anderson, I. R. Harris, *J. Alloys Comp.* 2003, **356–357**, 710.
22. M. Eddaoudi, J. Kim, N. Rosi, D. Vodak, J. Wachter, M. O'Keeffe, O. M. Yaghi, *Science* 2002, **295**, 469.
23. N. L. Rosi, J. Eckert, M. Eddaoudi, D. T. Vodak, J. Kim, M. O'Keeffe, O. M. Yaghi, *Science* 2003, **300**, 1127.
24. J. Rowsell, A. R. Millward, K. S. Park, O. M. Yaghi, *J. Am. Chem. Soc.* 2004, **126**, 5666.
25. O. Huebner, A. Gloess, M. Fichtner, W. Klopper, *J. Phys. Chem. A* 2004, **108**, 3019.
26. L. Pan, M. B. Sander, X. Huang, J. Li, M. Smith, E. Bittner, B. Bockrath, J. K. Johnson, *J. Am. Chem. Soc.* 2004, **126**, 1308.
27. W. L. Mao, H.-K. Mao, A. F. Goncharov, V. V. Struzhkin, Q. Guo, J. Hu, J. Shu, R. J. Hemley, M. Somayazulu, Y. Zhao, *Science* 2002, **297**, 2247.
28. L. J. Florusse, C. J. Peters, J. Schoonman, K. C. Hester, C. A. Koh, S. F. Dec, K. N. Marsh, E. D. Sloan, *Science* 2004, **306**, 469.
29. X. Zhao, B. Xiao, A. J. Fletcher, K. M. Thomas, D. Bradshaw, M. J. Rosseinsky, *Science* 2004, **306**, 5698.
30. S. Patchovskii, J. S. Tse, *Proc. Nat. Acad. Soc.* 2003, **100**, 14649.
31. T. Graham, *Philos. Trans. R. Soc.* 1866, **156**, 415.
32. A. C. Switendick, *Z. Phys. Chem.* 1979, **117**, 89.
33. "Hydrogen in Metals", *Topics in Applied Physics*, Vols. **28 & 29** (Eds.: G. Alefeld, J. Völkl) Springer-Verlag, Berlin 1978.
34. "Hydrogen in Intermetallic Compounds III" *Topics in Applied Physics*, Vol. **73** (Ed. H. Wipf,) Springer-Verlag, Berlin 1997.
35. J. Huot in *Nanoclusters and Nanocrystals* (Ed. H. S. Nalwa) American Scientific Publishers, California 2003.
36. F. Schüth, B. Bogdanovic, M. Felderhoff, *Chem. Comm.* 2004, **20**, 2249.
37. A. Pundt, C. Sachs, M. Winter, M. T. Reetz, D. Fritsch, R. Kirchheim, *J. Alloys Comp.* 1999, **293–295**, 480.
38. A. Pundt, *Adv. Eng. Mater.* 2004, **1–2**, 6.
39. J. A. Eastman, L. J. Thompson, B. J. Kestel, *Phys. Rev. B* 1993, **48**, 84.
40. R. J. Wolf, M. W. Lee, R. C. Davis, P. J. Ray, *Phys. Rev. Lett.* 1994, **73**, 1971.
41. J. Weissmüller, Ch. Lemier, *Phys, Rev. Lett.* 1999, **82**, 213.
42. E. Salomons, R. Griessen, D. G. De Groot, A. Magerl, *Europhys. Lett.* 1988, **5**, 449.
43. J. Weissmüller, Ch. Lang, and Ch. Lemier, *Scripta mater.* 2001, **44**, 1899.
44. J. Weissmüller, C. Lemier, *Phil. Mag. Lett.* 2000, **80**, 411.
45. A. W. Weebar, H. Bakker, *Physica B* 1988, **153**, 93.
46. G. Liang, R. Schulz, J. Huot, *J. Alloys Comp.* 2001, **320**, 133.
47. P. Tessier, J. O. Ström-Olsen, R. Schulz, *J. Mater. Res.* 1998, **13**, 1538.
48. L. Zaluski, S. Hosatte, P. Tessier, D. H. Ryan, J. O. Ström-Olsen, M. L. Trudeau, R. Schulz, *Z. Phys. Chem.* 1994, **183**, 45.
49. T. Yamamoto, H. Inui, M. Yamaguchi, *Intermetallics* 2001, **9**, 987.

50. K. Tanaka, S. Okazaki, T. Ichitsubo, T. Yamamoto, H. Inui, M. Yamaguchi, M. Koiwa, *Intermetallics* 2000, **8**, 613.
51. C. C. Koch, J. D. Whittenberger, *Intermetallics* 1996, **4**, 339.
52. H. Sakaguchi, T. Sugioka, G. Adachi, *Chem.: Lett.* 1995, **7**, 561.
53. S. Corré, M. Bououdina, N. Kuriyama, D. Fruchart, G. Adachi, *J. Alloys Comp.* 1999, **292**, 166.
54. I. Konstantchuk, E. Yvanov, M. Pezat, B. Darriet, V. Boldyrev, P. Hagenmuller, *J. Less-Common Met.* 1987, **131**, 181.
55. J. Huot, H. Hayakawa, E. Akiba, *J. Alloys Comp.* 1997, **248**, 164.
56. S. Orimo, H. Fuji, *J. Alloys Comp.* 1996, **232**, L16.
57. C. Chiang, Z. Chin, T. Perng, *J. Alloys Comp.* 2000, **397**, 259.
58. S. Orimo, H. Fuji, T. Yoshino. *J. Alloys Comp.* 1995, **217**, 287.
59. J. Huot, S. Boily, E. Akiba, R. Schulz, *J. Alloys Comp.* 1998, **280**, 306.
60. T. N. Dymova, D. P. Aleksandrov, V. N. Konoplev, *Koord. Khim.* 1994, **20**, 4, 279.
61. T. N. Dymova, V. N. Konoplev, A. S. Sizareva, *Dokl. Ross. Akad. Nauk.* 1998, **359**, 2, 200.
62. B. Bogdanovic, M. Schwickardi, *Appl. Phys. A* 2001, **72**, 221.
63. J. Huot, S. Boily, V. Guther, R. Schulz, *J. Alloys Comp.* 1999, **283**, 304.
64. M. Fichtner, Ch. Frommen, O. Fuhr, *Inorg. Chem.* 2005, **44**, 3479.
65. T. N. Dymova, V. N. Konoplev, A. S. Sizareva, D. P. Aleksandrov, *Russ. J. Coord. Chem.* 1999, **25**, 5, 334.
66. A. Zaluska, L. Zaluski, J. O. Ström-Olsen, *Appl. Phys. A* 2001, **72**, 157.
67. G. Liang, J. Huot, S. Boily, R. Schulz, *J. Alloys Comp.* 2000, **305**, 239.
68. G. Liang, J. Huot, S. Boily, A. Van Neste, R. Schulz, *J. Alloys Comp.* 1999, **291**, 295.
69. O. Gutfleisch, N. Schlorke-de Boer, N. Ismail, M. Herrich, A. Walton, J. Speight, I. R. Harris, A. S. Pratt, A. Züttel, *J. Alloys Comp.* 2003, **356–357**, 598.
70. J.-L. Bobet, S. Pechev, B. Chevalier, B. Darriet, *J. Mater. Chem.* 1999, **9**, 315.
71. K. Zeng, T. Klassen, W. Oelerich, R. Bormann, *J. Alloys Comp.* 1999, **283**, 213.
72. M. Khrussanova, M. Terzieva, P. Pesdhev, I. Konstanchuk, E. Y. Ivanov, *Z. Phys. Chem. Neue Folge* 1989, **164**, 1261.
73. W. Oelerich, T. Klassen, R. Bormann, *J. Alloys Comp.* 2001, **315**, 237.
74. B. Bogdanovic, M. Schwickardi, *J. Alloys Comp.* 1997, **253**, 1.
75. A. Peles, J. A. Alford, Zhu Ma, Li Yang, M. Y. Chou, *Phys. Rev. B* 2004, **70**, 165105.
76. J. Chen, N. Kuriyama, Q. Xu, H. T. Takeshita, T. Sakai, *J. Phys. Chem. B* 2001, **105**, 11214.
77. M. Fichtner, O. Fuhr, *J. Alloys Comp.* 2002, **345**, 286.
78. M. Fichtner, J. Engel, O. Fuhr, O. Kircher, O. Rubner, *Mat. Sci. Eng. B* 2004, **108**, 42.
79. B. Bogdanovic, R. A. Brand, A. Marjanovic, M. Schwickardi, J. Tölle, *J. Alloys Comp.* 2000, **302**, 36.
80. A. Zaluska, L. Zaluski, J. O. Ström-Olsen, *J. Alloys Comp.* 2000, **298**, 125.
81. D. L. Anton, *J. Alloys Comp.* 2003, **356**, 400.
82. C. M. Jensen, R. A. Zidan, N. Mariels, A. G. Hee, C. Hagen, *Int. J. Hydrogen Energy* 1999, **24**, 461.
83. G. Sandrock, K. Gross, G. Thomas, *J. Alloys Comp.* 2002, **339**, 299.
84. K. J. Gross, G. J. Thomas, C. M. Jensen, *J. Alloys Comp.* 2002, **330**, 683.
85. M. Fichtner, O. Fuhr, O. Kircher, J. Rothe, *Nanotechnology* 2003, **14**, 778.
86. B. Bogdanovic, M. Felderhoff, S. Kaskel, A. Pommerin, K. Schlichte, F. Schüth, *Adv. Mater.* 2003, **15**, 1012.
87. O. Kircher, M. Fichtner, *J. Appl. Phys.* 2004, **95**, 7748.
88. O. Kircher, A. Léon, *Appl. Phys. A* (submitted).
89. A. Léon, H. Roesner, private communication.
90. K. J. Gross, S. Guthrie, S. Takara, G. Thomas, *J. Alloys Comp.* 2000, **297**, 270.
91. E. H. Majzoub, K. J. Gross, *J. Alloys Comp.* 2003, **356–357**, 363.
92. M. Felderhoff, K. Klementiev, W. Grünert, B. Spliethoff, B. Tesche, J. M. Bellosta von Colbe, B. Bogdanovic, M. Härtel, A. Pommerin, Ferdi Schüth, C. Weidenthaler, *J. Phys. Chem. Chem. Phys.* 2004, **6**, 4369.
93. H. W. Brinks, C. M. Jensen, S. S. Srinivasan, B. C. Hauback, D. Blanchard, K. Murphy, *J. Alloys Comp.* 2004, **376**, 215.
94. M. Fichtner, P. Canton, O. Kircher, A. Léon, *J. Alloys Comp.* (in press).
95. O. M. Lovvik, S. M. Opalka, *Phys. Rev. B* (in press).
96. E. Stanik, L. E. Valiente, G. Majer, O. Kircher, M. Fichtner, *J. Alloys Comp.* (in press).
97. A. Züttel, S. Rentsch, P. Fischer, P. Wenger, P. Sudan, Ph. Mauron, Ch. Emmenegger, *J. Alloys Comp.* 2003, **356–357**, 515.
98. P. Chen, Z. Xiong, J. Luo, J. Lin, K. L. Tan, *Nature* 2002, **420**, 302.
99. T. Ichikawa, N. Hanada, S. Isobe, H. Leng, H. Fujii, *J. Phys. Chem. B* 2004, **108**, 7887.
100. S. Isobe, T. Ichikawa, N. Hanada, H. Y. Leng, M. Fichtner, O. Fuhr, H. Fujii, *J. Alloys Comp.* (in press).
101. Y. H. Hu, E. Ruckenstein, *Ind. Eng. Chem. Res.* 2003, **42**, 5135.
102. Y. Nakamori, S. Orimo, *J. Alloys Comp.* 2004, **370**, 271.
103. W. Luo, *J. Alloys Comp.* 2004, **381**, 284.
104. H. Leng, T. Ichikawa, S. Hino, N. Hanada, S. Isobe, H. Fujii, *J. Phys. Chem. B* 2004, **108**, 8763.
105. Y. Gogotsi, R.K. Dash, G. Yushin, T. Yildirim, G. Laudisio, J.E. Fischer, *J. Am. Chem. Soc.* 2005, **127**, 16006–16007.
106. W. Lohstroh, M. Fichtner, *Phys. Rev. B* 2007(in press)
107. Ch. Chlopek, Ch. Frommen, A. Léon, O. Zabara, M. Fichtner, "Synthesis and Properties of Mg(BH4)2", *J. Mater. Chem.* (DOI: 10.1039/b702723k)

Nanotechnology: An Approach to Mimic Natural Architectures and Concepts

Pierangelo Gröning

1
Introduction

Most of the revolutions in the industrial evolution of the past 250 years were to a good extent accompanied by the abilities to produce smaller and smaller devices with higher and higher precision. In a historical perspective one will observe that the rate at which this ability to miniaturise structures has rapidly increased in the past 40 years. From the early ages of mankind on to the beginning of the antique cultures around 3000 B.C. the relevant scale of material processing was the meter down to the centimeter. Then, skilled craftsmen developed routes for fineries, e.g. by manufacturing metals with millimeter features. Despite the continuous decreasing tolerances in the machining, the millimeter remained the relevant scale of the smallest features in material processing till the second half of the last century. The step to the micrometer was taken in the 1970ies with the development of the semiconductor technology. From then on it took less than 30 years to extend materials processing to the nanometer range. Thus, mechanical engineering, electromechanics, micromechanics, electronics, microelectronics developed successively. Today, it is possible to fabricate micro-electromechanical systems (MEMS) of several hundred nanometers in size with few nanometers tolerance. Recent microprocessors exhibit gate structure widths of 90 nm only. Manufacturing these very small structures demands tremendous precision and stability on the lithography process and, consequently, on all the optical and mechanical elements and the construction of the production tools. For example, the thermal expansion of the mask material for photolithography (ceramics) must be below 0.1 nm or one atomic diameter per centimeter and degree Celsius – a requirement which lies at the physical limits. This example shows clearly that industrial production techniques have penetrated deeply into the nanometer scale.

We are certainly aware that chemistry has been working with nano-objects for several hundred years. Here, we want to highlight the advances in processing individual objects though, which would be called the top-down approach in contrast to the bottom-up approach delivered by chemistry. It might seem that nanotechnology is just the natural consequent step in the continuous process of modern industrialisation. However, nanotechnology is much more than only the next stage of the miniaturisation process following microtechnology. An additional and central aspect of nanotechnology is that by shrinking objects to nanometer dimensions their physical properties start to differ markedly from macroscopic, in fact even from the microscopic ones. Actually, the nano-objects reveal their discrete, atomistic structure and the description of phenomena on this scale in terms of classical physics is not valid any more, that quantum effects not only become important but dominant.

The invention of the Scanning Tunnelling Microscope (STM) in 1981 by *G. Binnig* and *H. Rohrer* [1, 2] at IBM Rüschlikon, opened the gate to the nanoworld. With the STM and the later invented Atomic Force Microscope (AFM) [3] it became possible to visualise and handle individual atoms and molecules, in short STM and AFM may be addressed as a "*Feynman* Machine" – *"The principles of physics, as far as I can see, do not speak against the possibility of manoeuvring things atom by atom"; R. P. Feynman*, 29. Dec. 1959 [4]. The prospect to control matter on molecular and atomic level fascinates many physicists, chemists, biologists and engineers worldwide to work on nanoscience and nanotechnology. In the approach to control matter on the ultimate level of molecules and atoms, many disciplines of classical science such as physics, chemistry, materials engineering but also biology and medicine meet on the nanoscopic scale. As a matter of fact, the biological and medical aspects of nanotechnology should especially be highlighted, since nature does nanotechnology at its best with proteins as nanomachines performing highly complex tasks with great efficiency.

2
Physical Benefits of Nano-particles

Breaking a macroscopic piece of matter into ever smaller parts, results first in an increase of the specific surface area, which is inverse proportional to the average diameter of the particles. Due to this, at first glance inconspicuous effect, nano-powders reach enormous specific surface areas in the order of 100–1000 m^2/g (Fig. 1). This makes nano-powders very interesting for applications involving surface reactions,

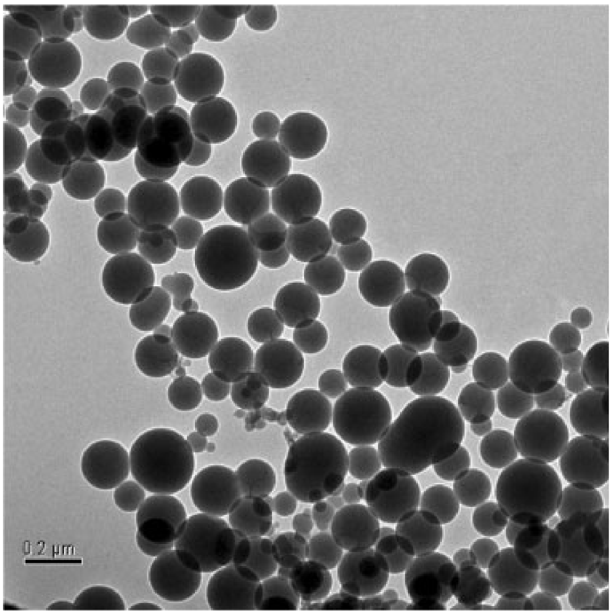

Fig. 1 TEM picture of non-aggregated spherical SiO_2 nano-particles produced by flame synthesis. (A. Vital, EMPA, Laboratory for High Performance Ceramics)

e.g. catalysts, membranes, filters or electrochemical double layer capacitors, often called super- or ultra-capacitor, will profit from a much extended efficiency by high surface area materials. The electrochemical double layer capacitor (EDLC) utilizes the double or *Helmholtz* layer formed at the electrode/electrolyte interface (Fig. 2) to store electrostatic energy. The electric charges accumulated on the electrode surface and the ions of opposite charge arranged on the electrolyte side of the interface form an electric dipole of an extremely small size of few nanometers. In first approximation, the capacity of the EDLC is given by the simple formula of the plate capacitor:

$$C_{Helmholtz} = \frac{\varepsilon_0 \cdot \varepsilon_r \cdot A}{d} \qquad (1)$$

where A is the apparent electrode surface, d the thickness of the dipole layer and ε relates to the dielectric constants. The specific capacity (C/A) of an EDLC is in the range of 5–50 $\mu F/cm^2$. In other words, an EDLC consisting of 100 g active electrode material with a specific surface area of 1000 m^2/g has the extremely high capacity of about 1 Farad. In comparison, a plate air capacitor with 1 mm electrode distance would have electrodes of 10,000 m × 10,000 m in size in order to produce the same capacity.

Reducing the particle-size does not only lead to an increase of the specific surface area, which would be a purely geometrical effect. Particles with sizes below 100 nm show a much reduced scattering of visible light, due to suppressed refraction. Therefore, nano-particles are practically invisible in dilute emulsions, which make them very interesting for applications as transparent, protective layers, e.g. scratch-proof lacquer for automotives, UV-light protection lacquer, sun-lotion or on ink-jet paper for photo quality printing. Since 2003, ''Mercedes-Benz'' cars can be ordered with a new type of clear lacquer containing ceramic nano-particles (Fig. 3). ''DaimlerChrysler'' promises that the nano-particles provide a three-fold improvement in the scratch resistance of the paintwork and ensures visibly enhanced gloss over an extended period of time [6].

When the particle size becomes extremely small, one has to realize that the number of surface atoms becomes comparable or even larger than the number of bulk atoms. In such situations the physical properties of the particle may be dominated by the properties of the surface atoms, which often differ from those in the bulk. As a classic example, the melting point of Au-clusters decreases from 1300 K to 300 K (room temperature!) if the cluster diameter decreases from 25 nm to 2.4 nm. The decrease in the average coordination of the gold atoms leads to a decrease in cohesive energy. For semiconductor particles with sizes below 10 nm the *de Broglie* wavelength of the electrons becomes comparable to the particle size. In this situation quantum confinement leads to a discretization of the valence electron state comparable to the textbook problem of an electron in a spherical box. The energy of the electron states depends on the particle diameter d as $E \sim 1/d^2$. A further result of the small particle size is that the momentum conservation in electronic transitions is lifted; in combination with the discrete level spectrum of the electron states, semiconductor particles may become fluorescent with colors depending only on the particle size (Fig. 4).

Silicon, macroscopically an indirect bandgap semiconductor is changing to a direct bandgap semiconductor as nano-crystal. Pavesi et al. [7] sandwiched silicon nano-crystals (ϕ = 3 nm) between layers of oxidized silicon. They tested them for optical gain by pumping with green laser light

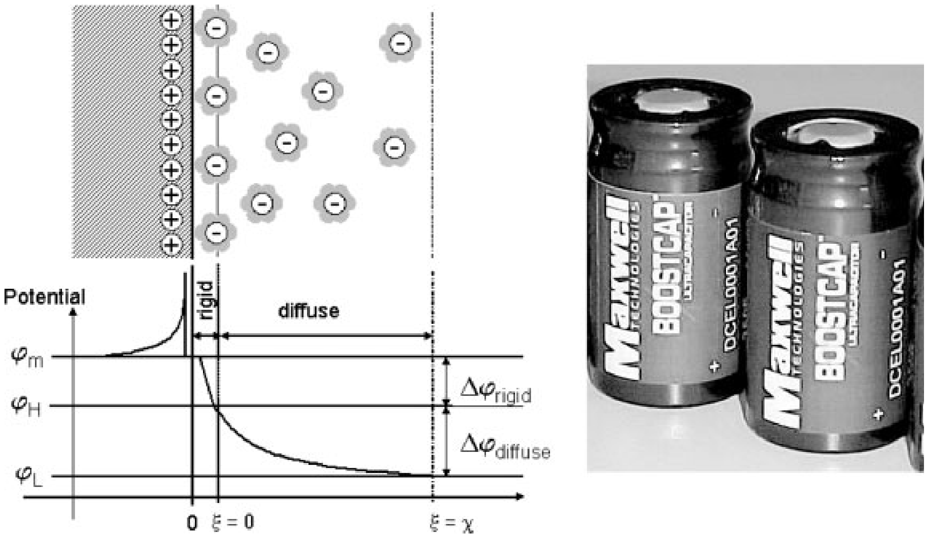

Potential
rigid
diffuse
φ_m
φ_H
φ_L
$\Delta\varphi_{rigid}$
$\Delta\varphi_{diffuse}$
0
$\xi = 0$
$\xi = \chi$
Maxwell TECHNOLOGIES
BOOSTCAP ULTRACAPACITOR
DCEL0001A01

Fig. 2 (a) Schematic illustration of the electrochemical double layer according to the model of Stern,[5] (b) 2.5 Volt Supercaps with 350 F capacity each. Size: 30 mm × 61.5 mm

(λ = 390 nm), while a second test laser shone through the sample. They observed amplification of the test beam in the red-infrared range with a peak at ~ 775 nm of the same order of magnitude as seen in gallium arsenide solid-state lasers.

Carbon nanotubes (CNTs) are considered as maybe the most famous objects in nanotechnology. The reason for this fame might be seen in a combination of several favorable circumstances. Firstly, CNTs represent a materials class with extraordinary properties in virtually every field of technological interest, be it mechanical, electronic, thermal, structural or optical. All these extraordinary properties arise truly from their nanoscopic dimensions and are markedly different from bulk graphite. Secondly, although being true nano-objects the CNTs are nanoscaled only in two dimensions. In the third

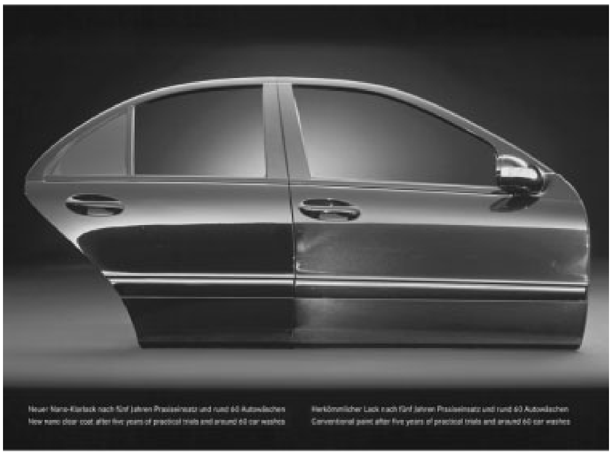

Neuer Nano-Klarlack nach fünf Jahren Praxiseinsatz und rund 60 Autowäschen
New nano clear coat after five years of practical trials and around 60 car washes
Herkömmlicher Lack nach fünf Jahren Praxiseinsatz und rund 60 Autowäschen
Conventional paint after five years of practical trials and around 60 car washes

Fig. 3 Comparison between the new "nano clear coat" (left door) and the conventional paint (right door) after five years of practical trials and around 60 car washes. (http://www.500sec.com/mercedesclearcoat.html)

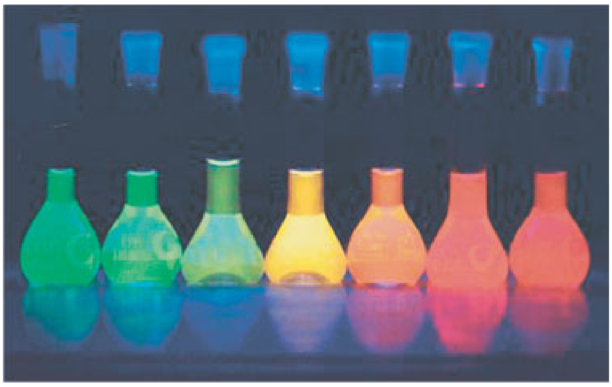

Fig. 4 Cadmium-Telluride (CdTe) particles fluoresce, the color depending only on particle size, from green (ϕ = 2 nm) to red (ϕ = 5 nm). (http://www.chemie.unihamburg.de/brochuere/pc/pc10.html)

dimension they remain micro-objects that can be handled with our conventional microtechnology. Thirdly, CNTs can be produced in relevant quantities using relatively simple processes and, therefore, they have a good availability. CNTs are macromolecules composed only of carbon atoms, discovered in 1991 by S. Iijima [8]. They can be viewed as a graphene sheet rolled into a cylinder and seamlessly welded together. Tubes consisting of a single graphene layer are called single-walled nanotubes (SWNT) and tubes consisting of multiple, concentric layers are called multi-walled nanotubes (MWNT). Typical dimensions of the CNTs are one to several nanometers in diameter and length up to several micrometers. Their applications arise from the remarkable properties of SWNT such as ballistic electron transport as phonon scattering is efficiently suppressed, highest thermal conductivity (3000 W/Km), highest Young's modulus, and the high aspect ratio structure. The high electric conductivity in combination with the aspect ratio makes CNTs very interesting as fillers for conductive ceramic or polymer composites. From percolation theory it is known that a mixture of conducting and insulating particles becomes conductive when the volume fraction of the conducting particles exceeds 16%. This value is size- but not shape-independent. For CNT with their high aspect ratio, the chance of contact increases and conduction can occur at lower volume fraction of the conductive particles. Sandler et al. [9] achieved in MWNT-epoxy composites percolation at loadings smaller than 0.01%. A very interesting application of CNT-polymer composites is electromagnetic shielding or antistatic coating. Because SWNTs have the highest *Young*'s modulus of about 1 TPa if normalized to their diameter, all types of CNTs have attracted much interest for low weight structural composites. The incorporation of SWNTs has been demonstrated to significantly improve the mechanical properties of composites; SWNT reinforced composites with extremely high *Young* moduli and tensile strengths have been reported [10–13]. S. Kumar et al. [14] have shown that 10 wt% SWNTs added to poly(p-) phenylene benzobisoxazole increases the tensile strength by about 50% in comparison to the pure polymer and reduces the shrinkage as well as high-temperature creep. It is worth noting that the mechanical properties of the CNT-polymer composites strongly depend on the ability to homogeneously disperse them throughout the matrix without destroying their integrity. The main difficulties to disperse an adequate level of CNT into a polymer matrix are: the insolubility of the material and thus the relatively weak chemical bond between the CNT and the polymer. Functionalisation of the CNT surface to improve the according with the polymer is a challenge.

Several methods have been developed to synthesise CNTs. All of them, namely, arc evaporation [15], laser ablation [16], chemical vapour deposition (CVD) [17], and plasma enhanced chemical vapour deposition (PECVD) [18, 19] need metal particles (e.g. Fe, Co, Ni) for the catalytic growth reaction and elevated temperatures ($T > 600\,^\circ C$). The catalytic synthesis of CNTs has its origin in the corresponding early works on carbon fibers [20, 21]. Backer and Harris [20] proposed a model for the catalytic carbon filament growth in which the hydrocarbon precursor is decomposed on the surface of the catalytic metal particle. Dissociated carbon dissolves into the particle and diffuses through the particle to be precipitated on the trailing face, forming the carbon filament. The dissociated reactive hydrogen keeps the metal particle surface catalytic active by reducing oxides. The application of this model to CNT has been demonstrated recently by imaging the CNT growth process by in-situ transmission electron microscopy [22].

The world-wide production capacity for CNTs is estimated to around 50 tons per year and will be doubled in two years. Thomas Swan & Co. Ltd. launched in April 2004 at Consett (UK) the first commercial CNT plant in Europe [23].

3
Micro- and Nano-Structured Surfaces

In the previous section we have looked at particles and the consequences of reducing their size to nanoscopic dimensions. Now we will focus our attention on the question what can be gained by producing nanoscopic structures on surfaces.

Nature makes use of submicron structured surfaces in manifold manners, e.g. to keep leaf surfaces clean, or to achieve unique optical effects. In Asian culture, the lotus flower is regarded as the symbol of cleanliness. This, because the lotus leaf exhibits self-cleaning, nicely illustrated by the series of environmental scanning electron microscope (ESEM) pictures shown in Figure 5. The Pictures 5a and 5b show the clean lotus leaf's surface, which is covered with cones of $\sim 5\,\mu m$ in diameter and roughly $5\,\mu m$ in height.

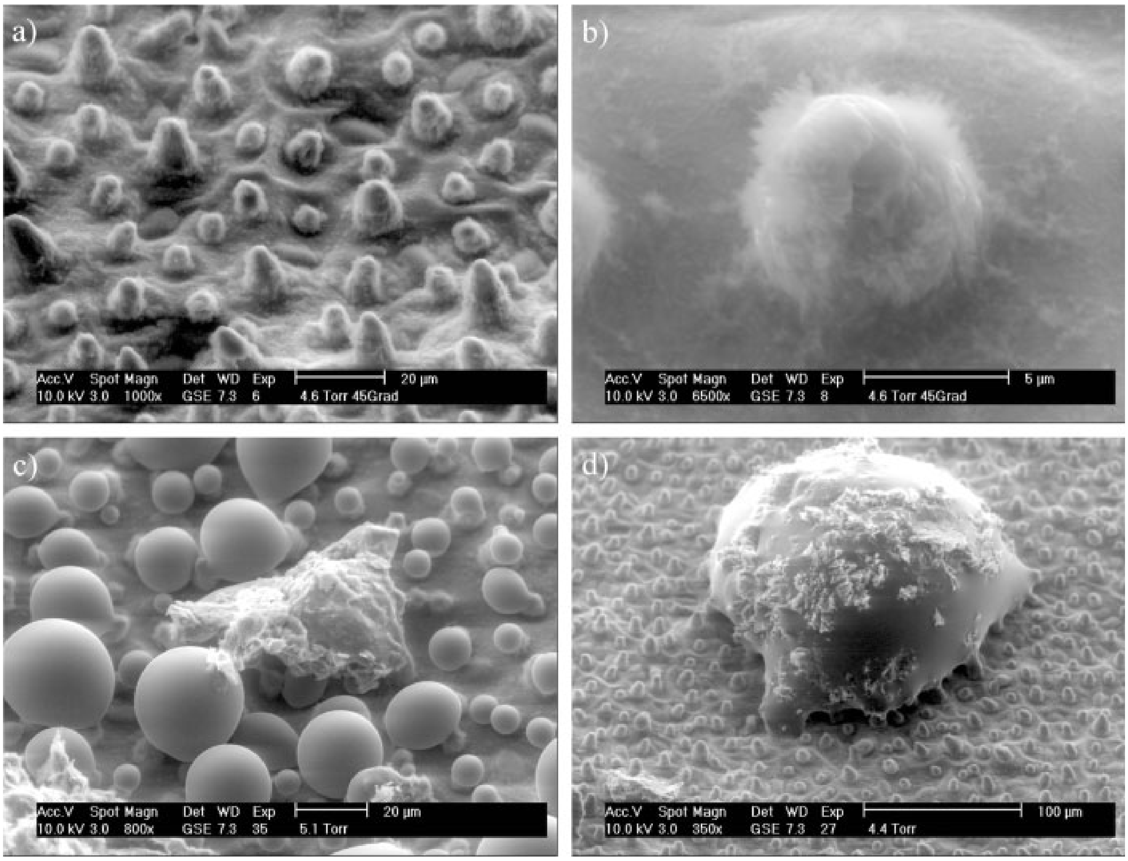
a)
Acc.V Spot Magn Det WD Exp 20 µm
10.0 kV 3.0 1000x GSE 7.3 6 4.6 Torr 45Grad
b)
Acc.V Spot Magn Det WD Exp 5 µm
10.0 kV 3.0 6500x GSE 7.3 8 4.6 Torr 45Grad
c)
Acc.V Spot Magn Det WD Exp 20 µm
10.0 kV 3.0 800x GSE 7.3 35 5.1 Torr
d)
Acc.V Spot Magn Det WD Exp 100 µm
10.0 kV 3.0 350x GSE 7.3 27 4.4 Torr

Fig. 5 (a) & (b) Environmental SEM (ESEM) pictures showing the micro- & nano-structure of a lotus leaf surface, c) & d) ESEM pictures showing water droplets condensing on the leaf surface and lifting up graphite particles. (L. Holzer, EMPA, Laboratory for Concrete/Construction Chemistry)

These cones are again covered with fine hairs less than 100 nm in diameter and 0.5 μm in length. By increasing the relative humidity in the ESEM one can observe that water condenses first at each cone on the leaf surface forming small micron sized droplets. These droplets stick rather strongly to the cones, which enables them to lift up dust particles present on the surface (Fig. 5c). Further condensation increases the water droplet size till they merge forming a macroscopic droplet heavy enough to run off from the slightly vaulted leaf, taking with them every dust particles (Fig. 5d). This so-called lotus effect has already been used in a range of products, e.g. as "anti-graffiti" spray or self-cleaning textiles.

South-American butterflies from the Morpho family appear from specific angles as vivid blue and green metallic lustrous but rather dull brown under others angles (Fig. 6a). This vibrant play of colours displays attractive behaviour of the butterfly, resp. simultaneously serves as a camouflage. Whereas the majority of insects uses pigments (absorption) or occasionally thin film multilayer structures (interference) to create colours, the wings of the butterflies from the Morpho family have a complex 3-dimensional arrangement of submicron grating structures. The physical principle behind this unique visual effect is zero-order diffraction [24, 25]. Light incident perpendicular on a fine grating structure is typically diffracted into a number of grating orders at angles θ_n given by the *Bragg* relation $\sin\theta_n = n\lambda/d$, where λ is the wavelength of the light, $n = 0 \pm 1\ldots$ and d the grating period. At zero diffraction order ($n = 0$) light is transmitted

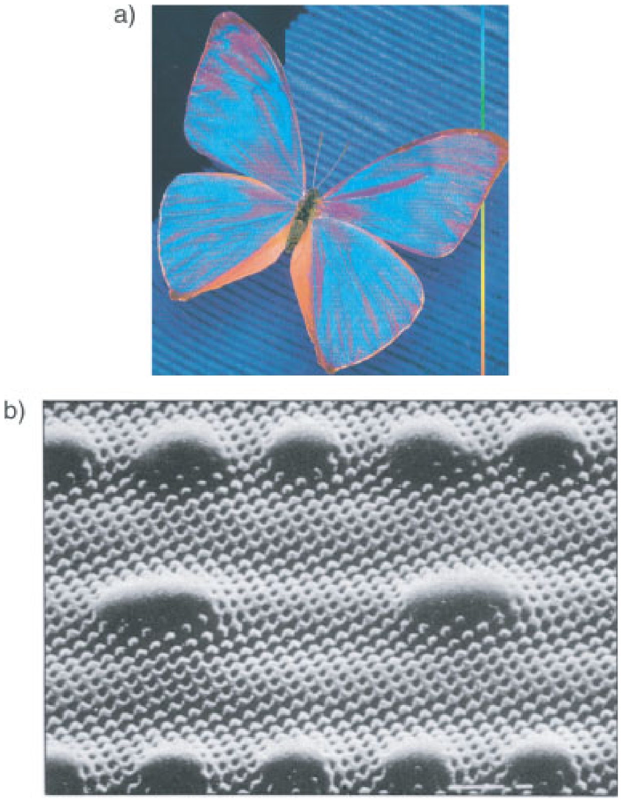

Fig. 6 (a) Butterfly from the Morpho family. Underneath, a SEM picture showing the submicron relief structure of its wings period (from[27]). (b) SEM picture of data written on the optical disk using moth eye antireflection structure (λ = 250 nm) (“Plasmon Data System”).

or reflected undeviated (independent of wavelength). If the grating period is finer than the wavelength no first order exists for illumination normal to the grating plane. Submicron gratings exhibit a strong polarisation dependence of reflected and transmitted light of wavelengths greater than the grating period, making zero-order diffractive optical elements an obvious candidate for polarisation applications. In contrast to holograms, zero-order optical structures can be viewed even under diffuse illumination, which makes them very interesting as security features. Zero-order optical structures based on embossed relief gratings, coated with a dielectric medium and fully integrated into a plastic sheet, are inherently more secure than holograms and, nowadays, widely used security feature on credit cards, banknotes, etc.

The night moth uses zero-order optical structures in a different way. The structure of its eyes is covered with a crossed grating microrelief with a period of about 250 nm. This structure acts as an antireflection layer at visible light by introducing a smooth, gradual change in the effective refractive index from the cornea to air, the eyes become less visible, and thus the whole moth to predators at night [26]. This moth eye antireflection structure has been investigated as an efficient absorber for thermal solar panels. The English Company “Plasmon Data System” developed an optical data storage disk based on the moth eye antireflection structure (Fig. 6b). The polycarbonate disk is covered by a uniform two-dimensional grating relief of about 250 nm period with a thin platinum layer on top. The structure is designed to have a low reflectance in the near infra-red. Data is written by heating the surface with the focused beam from a semiconductor laser, causing the formation of micron-sized bubbles with higher reflectance. The main technological advantage to create antireflection surfaces by optical microstructures instead of thin multilayer coatings is that the first can be produced very cheaply on a large scale by roll embossing into polymer foils or injection moulding.

Advanced technologies do not only allow copying natural structures, they are also able to produce nanostructures with applications not known in nature. Examples are the carbon nanotubes (CNTs) and their applications in vacuum micro-electronics. In opposition to the well established solid state microelectronics, which relies on the scattering dominated drift of electrons (or holes) in the conduction band (valence band) of a semiconductor such as silicon, vacuum microelectronics makes use of the collision free, ballistic motion of electrons in vacuum. The advantage of using a semiconductor as transport medium for electrons resides in the fact that the energies required to promote an electron from the valence band into the conduction band is in the order of 1 eV and, therefore, much lower than the work function ϕ, the energy required to put an electron from solid into vacuum which is usually in the order of 4–5 eV. Using doped semiconductors, the activation energy to produce carriers in the conduction band can be reduced further to a few meV, which makes it possible to have a saturated carrier concentration even at room temperature. Yet, the motion of electrons in a solid involves also some important disadvantages. On account of the collision dominated motion of the electrons in the conduction band the drift

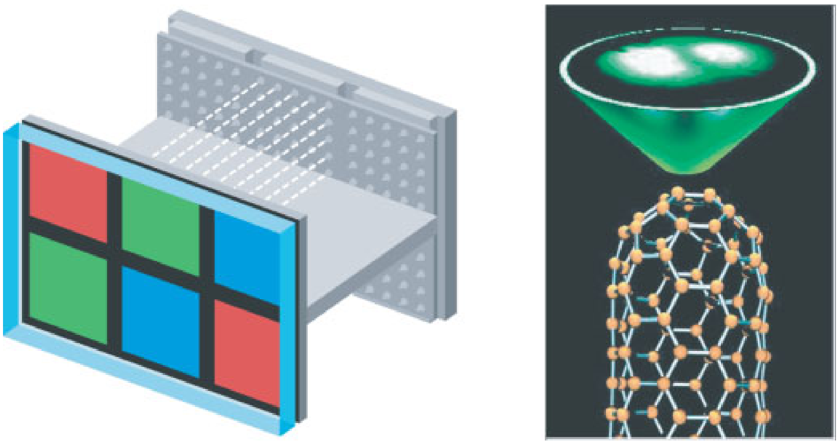

Fig. 7 (a) Sketch of 6 pixels of a Field Emission Display (FED). Each pixel on the screen is illuminated by about several hundreds electron emitters. (b) Schematic illustration of single CNT field emission pattern of a SWNT.

velocities are considerably smaller than for ballistic motion in vacuum at comparable electric fields. Furthermore, power is dissipated due to the interaction of the electrons with the lattice. Especially for high power and high frequency applications this power dissipation becomes a problem, because in the field of high power radio frequency and microwave amplifiers, vacuum electronic devices such as electron tubes, magnetrons or klystrons are still being used. In these devices the free electrons are usually generated by thermionic electron emitters. Albeit rather inexpensive and delivering high current densities, thermionic electron sources have a number of significant drawbacks: The large heat dissipation causes rather large energy consumption and, most important, the thermionic electron sources can neither be made very small nor be turned on and off at high frequencies. The term ‘‘small’’ here refers to length scales in the order of micrometers. Field electron emission from solids may open the door for the miniaturisation of electron sources. However, it was not until two decades ago that improvements in micro fabrication techniques triggered progress in manufacturing devices with miniaturised electron sources. These are commonly called vacuum microelectronic devices (VMED) and make use of a large number of micronsized field emitting electron sources. A device design based on a large number of parallel operating micron-sized electron sources can offer interesting technological applications, e.g., high power radio frequency amplifiers, electron guns for klystrons or travelling wave tubes, parallel e-beam lithography, or data storage where field emitted electron beams are used to read and write bits of data into a storing medium [28–31]. And perhaps in just a few years, we will find flat-screen televisions hanging on our walls, the working principle of which is based on the use of CNT as field electron emitter [32–34].

How is it that CNTs have found their way into the multibillion dollar application of flat panel displays? This stroke of luck is due to the fact that the first observation of their unique properties as electron emitter coincided with the development of a new field emission display (FED) technology in the second half of the 1990ies (Fig. 7). FEDs work like the conventional cathode ray tubes, except that each single pixel is illuminated by its own addressable, miniaturized electron source, which is directly located behind it. This arrangement allows displays to be made of only a few millimeter thickness still having the same outstanding image quality as cathode ray tubes. Classical cathode ray tubes use thermionic electron sources operating at temperatures between 950 and 2000 °C, making miniaturization impossible. In contrast, field electron emitters operating at room temperature hold great potential for miniaturization.

In 1928 R.H. Fowler and L.W. Nordheim [35] delivered the first generally accepted explanation for the electron field emission in terms of the newly developed theory of quantum mechanics. According to the Fowler-Nordheim theory, the field emission current density j is a function of the electric field E and the emitter work function ϕ according to following equation:

$$j = \frac{e^3}{4(2\pi)^2\hbar\phi} E^2 \exp\left(-\frac{4\sqrt{2m_e}}{3\hbar e E}\phi^{1.5}\right)$$

Fowler–Nordheim Relation (2)

The large electric fields of ~ 3000 $\mathrm{V\mu m^{-1}}$ required for electron emission from a flat surface can hardly be achieved, but can be generated by the field enhancing properties of tip-like structures. The field enhancement at the apex of a needle-shaped object of height h and radius r is in first approximation equal to the aspect ratio h/r [36]. A CNT with a diameter of 10 nm and a length of 2 μm exhibits, therefore, a field enhancement factor β up to 400. Therefore the required electric field E of 3000 $\mathrm{V\mu m^{-1}}$ for field emission can be generated at the CNT apexby a relatively low applied electric field $E_0 = E{\cdot}\beta^{-1}$ of 7.5 $\mathrm{V\mu m^{-1}}$.

The development of field emission electron sources has been hampered for half a century by the fact that tip-like field-enhancing structures are needed to locally create the high electric fields. Usually the higher the field enhancement factor β is, the smaller the effective emission area of the tip becomes, as the radius of curvature decreases. Although the field emission current density can easily exceed 100,000 $\mathrm{Acm^{-2}}$, the total emission current per single tip remains small as the emitting area, being the apex of the enhancing tip, is small. The current per single tip in field emission rarely surmounts 0.1 mA. As a result, field emitters are only used where high brightness rather than high currents are required, as it is in the case of high resolution scanning electron microscopes.

The key for reaching field emitter electron sources of high current density approaching 1 $\mathrm{Acm^{-2}}$ is in the integration of a large number of field emitters. Depending on the required emission current density, the density of the emitting tips has to be up to 10^8 $\mathrm{cm^{-2}}$. This limits their size to micron or even submicron dimensions. Today there is only one mature technology to produce field emitter arrays. This technology is the so-called *Spindt* type metal microtip process [37, 38]. The drawbacks of this technology are the expensive production, the critical lifetime in technical vacuum conditions, and the high extraction voltage. CNTs can be regarded as the potential second-generation technology to the *Spindt* type metal microtips (Fig. 8)[39, 42]. The advantages of CNTs are, above all, a potentially larger lifetime thanks to the chemical inertness, the highest current density, $10^6 - 10^{10}$ $\mathrm{Acm^{-2}}$ (over 1000 times that for copper), lower operating voltage due to higher local field enhancement and, last but not least, they can be very simply and economically manufactured [43–46]. In concrete terms, this means that the use of CNTs as electron emitter has made the production of FEDs economically feasible. More or less all big companies active in display

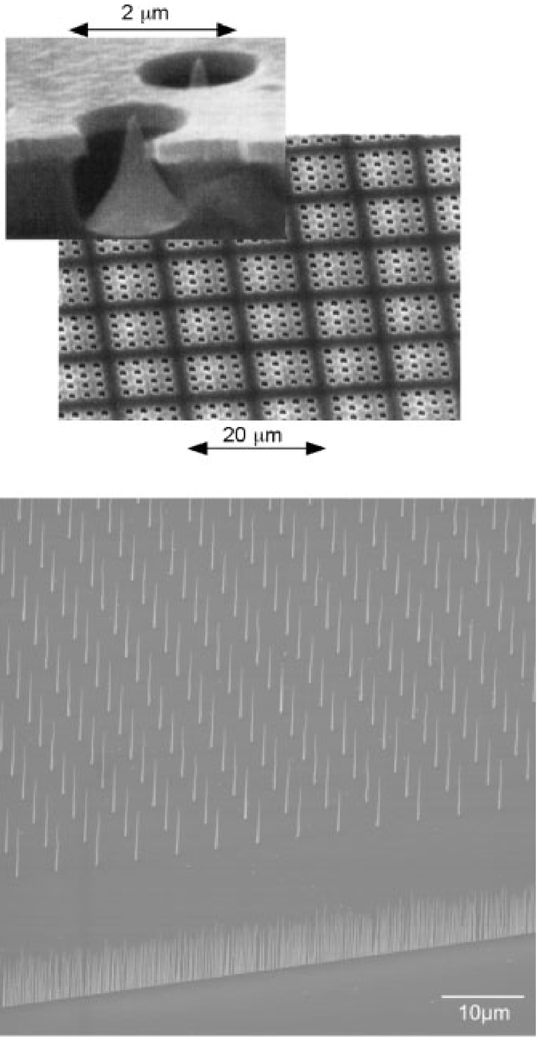

Fig. 8 (a) Field emitter array of evaporated metal tips, so called Spindt-tips[37]. (b) Array of vertical aligned CNT electron emitters grown by plasma enhanced chemical vapour deposition (PECVD (from.[39])

technology (Sony, Samsung, Toshiba, etc.) are intensively working on the development of CNT based FEDs. Prototypes have already been presented by the companies.

Quantum wells (QW) and quantum dots (QD) are further surface or interface nanostructures unknown in nature. These are nanostructures with two, respectively three dimensional carrier confinement, exhibiting strong quantised properties [47, 48]. The density of states (DOS) as well as the phonon modes vary sharply with the reduced degrees of freedom and dimensionality of the quantum system. Consequently, QWs and QDs can be designed with particular geometrical confinement and material combinations to achieve specific electronic and phonon properties. Today QWs are produced by growth on high-quality epitaxial layers of semiconductors by molecular beam epitaxy or metalorganic chemical vapour desposition. QDs, on the other hand, are mainly produced by two alternative ways: Firstly, by epitaxial growth with nano-scale patterning or surface self-assembly [47, 49, 50] and secondly, by the formation of nano-sized semiconductor particles through colloidal chemistry [51, 52]. The ability to tailor the electronic properties of materials to achieve specific optical and electronic functions by geometrical confinement was first recognised by R.D. Burnham et al. [53] and L. Esaki and R. Tsu [54] in the early 1970ies. With their pioneering works on lasing actions in hetero-structures and resonant tunnelling semiconductor superlattices to produce negative differential resistance, respectively, they opened up new horizons for a wide class of high functionality and high performance electronic and optoelectronic devices. Nowadays, through the use of advanced molecular beam epitaxy (MBE) technology, superlattice structures with almost any electronic energy levels, wave-functions and band structures can be deposited, making quantum engineering feasible. The quantum cascade laser (QCL) is an excellent example of how quantum engineering can be used to design materials with novel optoelectronic properties.

The QCL was invented in 1994 by J. Faist et al. [55] at "Bell Labs" in Murray Hill, New Jersey. The QCL is based on a fundamentally different principle to normal semiconductor lasers. It uses only one type of charge carrier, electrons, and is, therefore, called a unipolar laser. In essence the QCL operates like an electronic waterfall (Fig. 9). Passing through injectors and active transition regions repeatedly, a single electron can emit up to hundreds of photons as it travels through the corresponding number of active regions [56]. This is unlike double hetero-structure diode laser in which only one photon is emitted by electron hole recombination. Among the much higher quantum yield, the revolutionary aspect of the QCL is that it can be designed to emit any wavelength over an extremely wide range (4–24 μm), using the same combination of materials in the active region. In QCLs, the wavelength is essentially determined by quantum confinement, i.e. the layer thickness of the active region rather than by the bandgap of the material. QCLs can operate at high powers (up to 1 W in pulsed mode at room temperature) and are relative cheap to produce. All these benefits make the QCL an ideal light source in applications such as infrared spectroscopy, optical wireless communications and medical diagnostics [57].

4
Molecular Electronics and Molecular Self-assembly

Silicon-based microelectronics has advanced at exponential rates in performance and productivity throughout the past four decades. The general trend is to make devices smaller and faster and, therefore, to provide always cheaper electronic devices with greater performance for use in everyday life. It has been scaled so continuously that it is beginning to encounter fundamental limits associated with the laws of quantum mechanics, limitations of fabrication techniques, and fundamental materials properties, e.g. the breakdown voltage of ultrathin ($d < 3$ nm) gate oxides. In order to continue the miniaturization of electronic devices, entirely new nano-electronic device structures and computational paradigms are required. Although silicon will undoubtedly remain of central importance as microelectronics moves towards nanoelectronics, many would argue that carbon becomes very important in nanoelectronics too. What makes carbon so interesting for nanoelectronics? Carbon is surly the most flexible and versatile element of the periodic system. Due to the ability of its 2 s and 2 p valence electron states to form different hybrid orbitals even in the carbon-carbon bond, it can form 1-dimensional chains (sp^1 hybrid), 2-dimensional sheets with the well known graphite structure (sp^2 hybrid) and 3-dimensional networks like diamond or tetrahedral amorphous carbon (sp^3 hybrid). The properties of the different bond configurations are related to the varying number of delocalized π-electrons giving rise to conduction for the sp^1 and sp^2 hybrid case, and the very

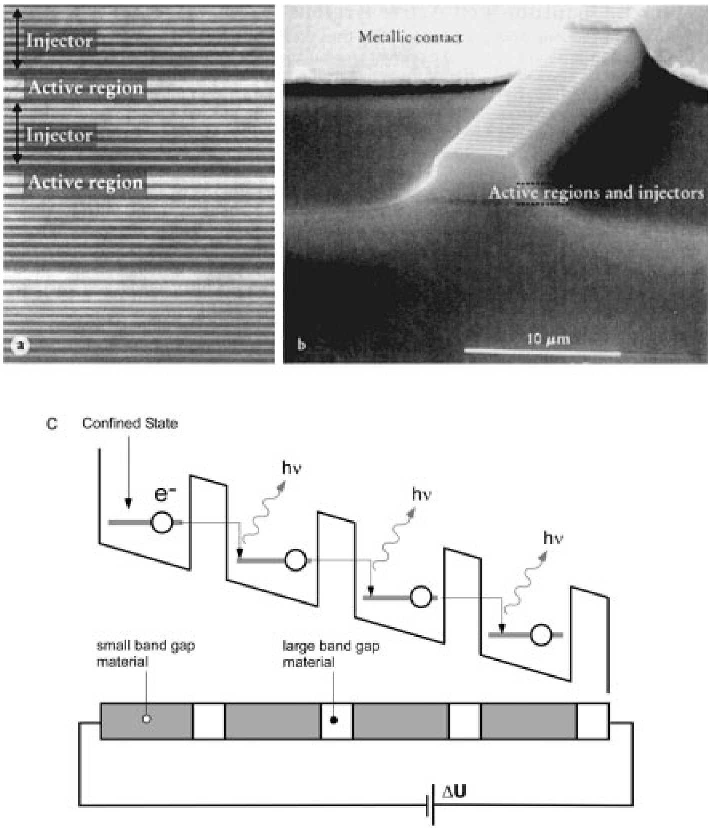

Fig. 9 (a) High resolution transmission electron microscope image showing several of the identical stages of the QCL structure, each consisting of an active region and an electron injector. The white layers in the active region are the quantum wells, made of GaInAs, separated by barrier layers (dark) made of AlInAs. The wide quantum wells in the active regions are 4.4 nm and 4.8 nm wide. (b) Scanning electron microscope image of a portion of the same QCL. The whole QCL is 8 μm wide and 2.5 mm long. (S.N. George Chu of "Agere Systems") c) Schematic energy diagram of a quantum cascade laser.

strong σ-bonds common to all hybrids leading to binding energies round 7 eV per carbon atom. Therefore, carbon materials span an enormous range of mechanical, thermodynamic and electronic properties that is larger than of any other elements. In its hexagonal sp^2-bonded graphite from, carbon is soft, semi-metallic and very anisotropic. In its four-fold-coordinated sp^3-bonded diamond form, carbon is the hardest material, the best thermal conductor and a wide band gap semiconductor with intrinsic good insulator properties. Alone on account of these two allotropic forms, carbon is a very promising material for electronic applications. But again, it was the discovery of CNTs which triggered key developments in carbon electronics.

SWNTs exhibit interesting electronic properties [58, 59]. The electronic band-structure of SWNT can be explained by the dispersion relation calculated for graphene. However, one important restriction on the allowed electronic states originates from the periodic boundary conditions in the circumferential direction of the tube. This reduces drastically the number of allowed wave-vectors in this direction and leads to the one-dimensionality of the bands, which expresses itself in the appearance of van Hove singularities in the density of states. Depending on the diameter and the chirality (Fig. 10), SWNTs can exhibit either metallic or semi-conducting electronic properties [60]. The ability to yield both metallic and semi-conducting SWNTs without

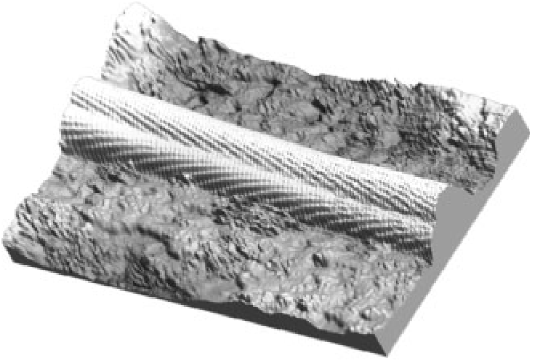

Fig. 10 Scanning Tunnelling Microscope (STM) picture of two SWNT with different chiralities. The larger tube has a diameter of $\phi = 1.2\,\text{nm}$.

doping is unique among solid-state materials and has led to the anticipation that SWNT might thus serve as a key building block for carbon based electronics.

Already in this early stage of research SWNT based field effect, transistors (FET) have shown to outperform state of the art p-Si-CMOS devices with regard to current drive and transconductance [61–64]. This clearly makes CNTs very promising candidates for electronic applications. Since CNTs are the mechanically and chemically most stable nanostructures, opportunities exist for integrating CNT electronics in other chemical, mechanical or biological systems. Examples include electrical sensing of individual biomolecules in physiological environment as well as electromechanical devices [65, 66]. Undoubtedly, CNTs are promising candidates for nanoelectronic building blocks, but the main problem to use CNT in nanoelectronics is the quality and purity of the material available. Although a great deal of work has been done, the progress to date in the controlled synthesis of CNTs has been modest. Until now there exists no synthesis route which allows the control of the chirality of the tubes. As a result, a batch of grown CNTs always contains metallic and semi-conducting tubes. Processes which allow the separation of metallic and semi-conducting tubes are slowly being developed [67, 68]. The separation of the CNTs with respect to their chirality will be an even greater challenge, and it is not at all clear whether this task can be realized at all.

A way out of this dilemma may be the use of organic molecules as electronic building blocks. Organic molecules can be relatively easily and reproducibly synthesized with extremely high purity. In addition, organic molecules can be functionalized in a great variety. This allows chemists to tailor molecules to building blocks in molecular architecture. Molecular architecture as self-assembly process is a very promising approach to create supramolecular structures with desired electronic or optical properties. As beneficial the use of molecules is with respect to tailoring properties and control purity, as enormous is the challenge in integrating specific molecules in a functional architecture e.g. adding electric contacts to molecules. Nevertheless, molecular architecture based on self-assembly processes seems a very promising approach to create supramolecular structures with desired electronic or optical properties.

A further aspect of nanotechnology is that it is not only associated with small, nanoscale objects, but also with adequate technologies to handle large numbers. As already shown, an A4 field emission display (FED) requires an array of about 10^{10} CNT electron emitters. The acceptable tolerance in length, diameter and interspacing of the CNTs for a high quality FED is in the range of 5%. This example shows clearly that nanotechnology demands drastically different approaches to manufacturing than previous technologies have taken. Instead of scaling materials down to create devices, it would be much more efficient to produce nano-sized objects by building them up piece by piece on a molecular level. The key-word of this bottom-up approach again is molecular self-assembly.

Nature is the master of self-assembly and self-organization since millions of years. Self-assembly is omnipresent in nature at both macroscopic and microscopic scales, from the assembly of flocks of birds in the sky, where the flock seems to develop behavior of its own, down to the molecular assembly of viruses or lipids. Nucleic acids, ribosomes, enzymes, proteins, etc. are molecular building blocks and molecular assemblers forming perfect biological systems by hierarchical self-assembly. The key feature here is the spontaneous organization of molecules under thermodynamic equilibrium conditions into structurally complex but well-defined and rather stable arrangements through a number of non-covalent interactions such as hydrogen bonds, ionic bonds, and van der Waals interactions (G.M. Withesides (1991)[69], J.-M. Lehn (1993)[70].) Further elements are chemical complementarity and structural compatibility. Many scientists believe that molecular self-assembly techniques will set the basis for manufacturing a wide range of objects, from electronic circuits to new materials, e.g.

advanced ceramics grown by biomineralization [72, 73]. Biomineralization is the process by which microorganisms mediate and catalyze inorganic reactions so as to form a composite assembled by minerals and organic material. An example is the mother of pearl or nacre, a remarkable robust nanocomposite despite the brittle nature of its main component aragonite (∼ 95%), the mineral of calciumcarbonate ($CaCO_3$). It is made up of about 10 µm in diameter and 500 nm thick regularly arranged aragonite platelets, held together by a 40 nm thick biopolymer glue. Although the content of the biopolymer glue is only about 5% it is dominantly responsible for the mechanical properties of nacre which has a toughness 1000 times greater than its constituents.

Since molecules are the smallest units that will ever have completely engineered electronic and optical properties, scientists have pursued the idea to build electronic circuits by molecules and molecular structures, called molecular electronics. It has been quite common in recent years to describe isolated molecular devices and their envisaged advantages [73–77], but the functional interconnection of such devices to higher levels has only rarely been achieved. One of the big challenges for molecular devices that needs to be overcome in the near future is developing strategies to build integrated circuits by bottom-up integration from the molecular scale. These molecule-based devices with advanced functions require the development of new and efficient approaches for integrating molecular building blocks into desired functional structures, ideally with these structures supported on suitable substrates. Recently, Y. Lou et al. [78] presented an evolutionary molecular design process to build up a two dimensional molecular crossbar memory device. Conventional crossbar random access memories (RAM) are made by ferroelectric and magneto-resistive materials called FeRAM and MagRAM, respectively. The electrical resistance of the crosspoints exhibits a hysteretic response to an applied field thus serving as the basis for non-volatile information storage. However, these systems also exhibit significant problems with respect to point addressability, known as "half-select" problem. The "half-select" problem arises, because in a crossbar circuits all of the devices are electrically interconnected, therefore, every junction (crosspoint) in the circuit subjected to at least some electrical fields, occasionally sufficient to alter the state of the junction since the poling process itself is based on a field driven nucleation event. This problem has seriously limited the development of crossbar RAM's. Using amphiphilic bi-stable[2]rotaxane molecules Y. Lou et al. build up a 64-bit addressable molecular switch RAM working with voltage pulses of +2 V for opening and −2 V for closing the switch. The ON/OFF current ratio was in the order of 300% which is at least one order of magnitude larger than for ferroelectric or magneto-resistive analogues. For the device fabrication, the molecules were self-assembled as Langmuir-Blodgett monolayers before being transferred to a substrate patterned with n-type polysilicon electrodes.

Supramolecular aggregation – or self-assembly – occurs spontaneously and can lead to controlled structures if selective and directional non-covalent interactions are exploited. It occurs in liquids as well as at surfaces and interfaces. In view of future molecular electronics self-assembly on solid surfaces is particularly important. A beautiful demonstration of this approach has recently been given by T. Yokoyame and co-workers [79] who have shown the formation of surface-supported supramolecular structures whose size and aggregation pattern are rationally controlled by tuning the non-covalent interactions between the individual adsorbed molecules. They demonstrated that structures formed by substituted porphyrins adsorbed on a gold surface correspond in a predictable fashion to the geometrical and chemical nature of the substituents that mediate the interactions between the adsorbed molecules. These findings suggest that careful positioning of functional groups that are able to participate in directed non-covalent interactions, will allow the rational design and construction of a wide range of molecular architectures adsorbed on surfaces. A first proof of principle of this approach has recently been given by. J. A. Theobald et al. [80] with the self-assembled growth of perfectly regular supramolecular hexagonal networks from perylene tetra-carboxylic di-imide (PTCDI) and melamine molecules on a Ag/Si(111) surface (Fig. 11).

Recently, R. Fasel et al. [81, 82] have demonstrated the transfer of chirality from single helical molecules into self-assembled monolayers. Although the self-organisation of chiral molecules into helical architectures is of fundamental importance in nature and has found important applications such as liquid crystal technologies [83], the mechanism of chirality induction is still poorly understood. By means of STM Fasel et al. have shown that helical heptahelicane molecules ([7]H, $C_{30}H_{18}$) absorbed into the close-packed monomolecular layer forms on Cu(111) handed supramolecular structures (Fig. 12). They demonstrated that

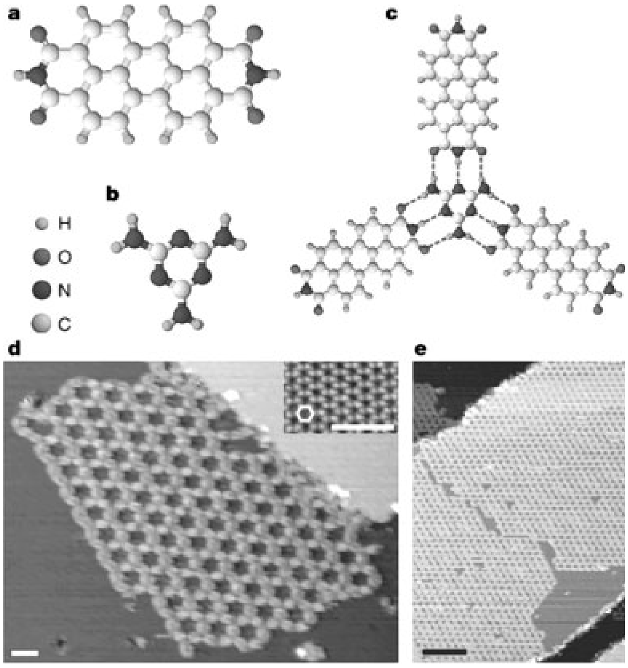

Fig. 11 Self-assembly of a PTCDI-melamine supramolecular network. Chemical structure of PTCDI (a) and melamine (b). c), Schematic diagram of a PTCDI-melamine junction. d) STM image of a PTCDI-melamine network. Scale bars, 3 nm. e) STM image of large-area network, with domains extending across terraces on the Ag/Si surface.[80] Scale bar, 20 nm. (from [80])

the self-assembly of molecules on surfaces is not only governed by the lateral molecule-molecule, but is also influenced by the molecule-substrate interaction, which, in turn, determines the mobility of the molecular species at the surface at a given temperature. They found, that the chirality transfer from the molecule to the self-assembled monolayer occurs through steric repulsion and does not require molecular flexibility or strong dipole interactions. The appearance of handed molecular clusters is due to a correlation of azimuthal molecular orientations which are driven by repulsive intermolecular forces in the close-packed 2D layer.

Molecular self-assembly is of course the most promising bottom-up concept in nanotechnology. Precisely controlled molecular self-assembly, as shown by the few examples above, has the potential to be a whole new engineering paradigm, much like engineering semiconductors.

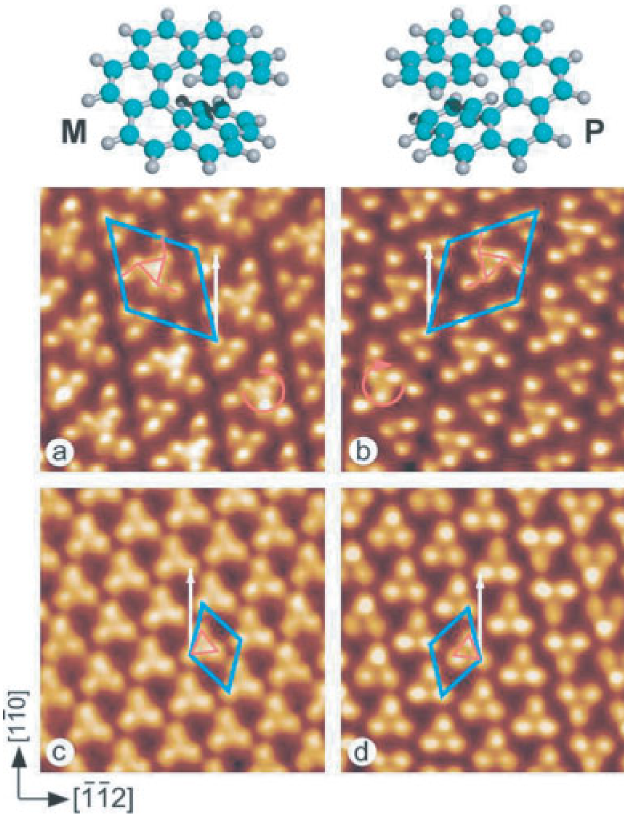

Fig. 12 Scanning Tunnelling Microscope (STM) pictures (10 nm × 10 nm) of (M)- and P-heptahelicene on Cu(111). (a) (M)-[7]H at 95% (θ = 0.95) of the saturated monolayer coverage. (b) (P)-[7]H at θ = 0.95. c) (M)-[7]H at θ = 1.0. d) (P)-[7]H at θ = 1.0. The (M)- and (P)-[7]H structures are mirror images of each other.(from [81])

5
Outlook

Whereas science aims at the advancement of our knowledge in the physical world and its phenomena, technology aims at the application of knowledge to improve our environment and life quality. It is natural that science and technology are linked by knowledge and that science usually precedes technology. We see today that in industry the awareness of processes and engineering at the nanometer level rapidly increases and products using materials engineered at the nanometer level emerge on the markets. In that sense we believe that it is absolutely justified to talk of nanotechnology as opposed to pure nanoscience. It may also be true that in many cases nanotechnology is not "visible" in the final product any more, as e.g. in an anti-scratch lacquer or an anti-reflection coating on viewing glasses, where nanotechnology might be much less related to radically new "nanoproducts" – but enabling new functionalities to well established products and new ways of producing them. In the long term, it is hardly possible to estimate, from the present point, where the ability to engineer materials on the nanoscale will lead us, whether the dream of a molecular nanotechnology with autonomous working machines built with atomic precision will ever become reality. Again nature might give us some hints, as it has realized such machines in the form of proteins working in our bodies all the time.

History teaches us, that in long term forecast future revolutionary technologies were always underestimated. Thus, experts believed, having invented the transistor in 1948, that one day, it might be possible to build a computer by semiconductors with 5000 operations per second weighing 1500 kg only, consuming 10 kilowatt. Today's PCs are in all respect 100'000 more powerful. Similarly it was with the electricity at the end of the nineteenth century or with telecommunication at the end of the twentieth century. Future will show whether nanotechnology will be placed in this line of revolutionary technologies.

References

1. G. Binnig, H. Rohrer, Ch. Gerber, E. Weibel, *Physica B* 1982, **109/110**, 2075
2. G. Binnig, H. Rohrer, Ch. Gerber, E. Weibel, *Phys. Rev. Lett.* 1982, **49**, 57.
3. G. Binnig, C. F. Quate, Ch. Gerber, *Phys. Rev. Let.* 1986, **56**, 930.
4. R. P. Feynman, *Engr. and Sci.* February 1960; http://www.zyvex.com/nanotech/feynman.html.
5. O. Z. Stern, *Elektrochem.* 1924, **30**, 508.
6. http://www.500sec.com/mercedesclearcoat.html.
7. L. Pavesi, L. Dal Negro, C. Mazzoleni, G. Franzo, F. Priolo, *Nature* 2000, **408**, 440.
8. S. Iijima, *Nature* 1991, **354**, 56.
9. J. K. W. Sandler, J. E. Kirk, I. A. Kinloch, M. S. P. Shaffer, A. H. Windle, *Polymer* 2003, **44**, 5893.
10. B. Vigolo, A. Penicaud, C. Coulon, C. Sauder, R. Pailler, C. Journet, P. Bernier, P. Poulin, *Science* 2000, **290**, 1331.
11. A. A. Mamedov, N. A. Kotov, M. Prato, D. M. Guldi, J. P. Wicksted, A. Hirsch, *Nature Materials* 2002, **1**, 190.
12. A. B. Dalton, S. Collins, E. Munoz, J. M. Razal, V. H. Ebron, J. P. Ferraris, J. N. Coleman, B. G. Kim, R. H. Baughman, *Nature* 2003, **423**, 703.
13. R. Sen, B. Zhao, D. Perea, M. E. Itkis, H. Hu, J. Love, E. Bekyarova, R. C. Haddon, *Nano Lettres* 2004, **4**, 459.
14. S. Kumar, T. D. Dang, F. E. Arnold, A. R. Bhattacharyya, B. G. Min, X. Zhang, R. A. Vaia, C. Park, W. W. Adams, R. H. Hauge, R. E. Smalley, S. Ramesh, P. A. Willis, *Macromolecules* 2002, **35**, 9039.
15. T. W. Ebbesen, *Annu. Rev. Mater. Sci.* 1994, **24**, 235.
16. T. Guo, P. Nikolaev, A. Thess, D. T. Colbert, R. E. Smalley, *J. Phys. Chem.* 1995, **55**, 10694.
17. M. Endo, K. Takeuchi, S. Igarashi, K. Kobori, M. Shiraishi, H. W. Kroto, *J. Phys. Chem. Solids* 1993, **54**, 1841.
18. S. Iijima, T. W. Ebbesen, P. M. Ajayan, *Nature* 1992, **358**, 229.
19. M. S. Dresselhaus, G. Dresselhaus, I. L. Spain, H. A. Goldberg, "*Graphite Fibers and Filaments*", Springer Verlag, New York 1988.
20. R. T. K. Baker, P. S. Harris, *Chem. Phys. Carbon* 1978, **14**, 8.
21. S. H. Tsai, C. W. Chao, C. L. Lee, H. C. Shih, *Appl. Phys. Lett.* 1999, **74**, 3462.
22. S. Helveg, C. López-Cartes, J. Sehested, P. L. Hansen, B. S. Clausen, J. R. Rostrup-Nielsen, F. Abild-Pedersen, J. K. Nørskov, *Nature* 2004, **427**, 426.
23. H. Swan, *European Coatings Journal* 2004, **11**, 18.
24. M. Gale, *Physics World* 1989, **2**, 24.
25. H. P. Herzig, "*Micro-Optics, Elements, Systems and Applications*", Taylor & Francis 1999.
26. S. J. Wilson, M. C. Hutley, *Optica Acta* 1982, **29**, 993.
27. www.cardag.ch.
28. Hewlett Packard, *Sci. American*, May 2000, 50.
29. G. Z. Yue, Q. Qiu, Bo Gao, Y. Cheng, J. Zhang, H. Shimoda, S. Chang, J. P. Lu, O. Zhou, *Appl. Phys. Lett.* 2002, **81**, 355.
30. W. B. Choi, D. S. Chung, J. H. Kang, H. Y. Kim, Y. W. Jin, I. T. Han, Y. H. Lee, J. E. Jung, N. S. Lee, G. S. Park, J. M. Kim. *Appl. Phys. Lett.* 1999, **75**, 3129.
31. Y. Saito, S. Uemura, *Carbon* 2000, **38**, 169.
32. R. G. Forbes, *Solid-State Electron.* 2001, **45**, 779.
33. O. Gröning, "Field Emission Properties of Carbon Thin Films and Carbon Nanotubes", Ph.D. Thesis, No. 1258, University of Fribourg, Switzerland 1999.
34. P. Gröning, L. Nilsson, P. Ruffieux, R. Clergereaux, O. Gröning, Carbon Nanostructures for Cold Electron Sources, in *Encyclopedia of Nanoscience and Nanotechnology*, 2004, **1**, 547, edited by H.S. Nalwa, American Scientific Publishers.
35. L. W. Nordheim, *Proc. Roy. Soc. London, Ser. A* 1928, **121**, 626.
36. F. Rohrbach, CERN Report 71-5/TC-L, 1971.
37. I. Brodie, C. Spindt, *Adv. Electron. Electron Phys.* 1992, **83**, 6.
38. P. R. Schwoebel, I. Brodie, *J. Vac. Sci. Technol. B* 1995, **13**, 1391.
39. K. B. K. Teo, S.-B. Lee, M. Chhowalla, V. Semet, Vu Thien Binh, O. Gröning, M. Castignolles, A. Loiseau, G. Pirio, P. Legagneux, D. Pribat, D. G. Hasko, H. Ahmed, G. A. J. Amaratunga, W. I. Milne, *Nanotechnology* 2003, **14**, 204.

40. E. Minoux, O. Gröning, K.B.K. Kenneth, S.H. Dalai, L. Gangloff, J.-P. Schnell, L. Hudanski, I.Y.Y. Bu, P. Vincent, P. Legagneux, G.A.J. Amaratunga, and W.I. Milne, *Nano Letters* 2005 **5**, 2135
41. W.I. Milne, K.B.K. Teo, E. Minoux, O. Groening, L. Gangloff, L. Hundanski, J.-P. Schnell, D. Dieumegard, F. Peauger, I.Y.Y. Bu, M.S. Bell, P. Legagneux, G. Hasko, G.A.J. Amaratunga *J. Vac.Sci. Technol. B* 2006 **24**, 345
42. S.M.C. Vieira, K.B.K. Teo, W.I. Milne, O. Gröning, L. Gangloff, E. Minoux, and P. Legagneux *Appl. Phys. Lett.* 2006 **89**, 022111
43. H. Kind, J.-M. Bonard, Ch. Emmenegger, L.-O. Nilsson, K. Hernadi, E. Maillard-Schaller, L. Schlapbach, L. Forro, K. Kern, *Adv. Mater*. 1999, **15**, 1285.
44. B. Q. Wie, R. Vajtai, Y. Jung, J. Ward, R. Zhang, G. Ramanath, P. M. Ajayan, *Nature* 2002, **416**, 495.
45. W. I. Milne, K. B. K. Teo, G. A. J. Amaratunga, P. Legagneux, L. Gangloff, J.-P. Schnell, V. Semet, Vu Thien Binh, O. Groening, *Journal of Materials Chemistry*, 2004, **14**, 933.
46. W. I. Milne, K. B. K. Teo, M. Chhowalla, G. A. J. Amaratunga, S. B. Lee, D. G. Hasko, H. Ahmed, O. Groening, P. Legagneux, L. Gangloff, J. P. Schnell, G. Pirio, D. Pribat, M. Castignolles, A. Loiseau, V. Semet, Vu Thien Bin, *Diamond and Rel. Mat*. 2003, **12**, 422.
47. "Semiconductor Quantum Dots", *MRS Bulletin*, 1998, **23**(2).
48. "Photonic Materials for Optical Communications", *MRS Bulletin* 2003, **28**(5).
49. W. G. Jeong, P. D. Dapkus, U. H. Lee, J. S. Yim, D. Lee, B. T. Lee. *Appl. Phys. Lett*. 2001, **78**, 1171.
50. P. M. Petroff, A. Lorke, A. Imamoglu, *Physics Today* 2001, **54**, 46.
51. G. Schedelbeck, W. Wegscheider, M. Bichler, G. Abstreiter, *Science* 1997, **278**, 1792.
52. B. Murray, S. Sun, W. Gaschler, H. Doyle, T. A. Betley, L. R. Kugan, *IBM J. Res. & Dev*. 2001, **45**, 47.
53. R. D. Burnham, N. Holonyak, D. R. Scifres, *Appl. Phys. Lett*. 1970, **17**, 455.
54. L. Esaki, R. Tsu, *J. Res. Develop*. 1970, **14**, 61.
55. J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, A. Y. Cho, *Science* 1994, **264**, 553.
56. F. Capasso, C. Gmachl, D. L. Sivco, A. Y. Cho, *Physics Today*, May 2002, p. 34.
57. R. Stevenson, *Compound Semiconductor*, September 2004, http://www.compoundsemiconductor.net/articles/magazine/10/9/4/1.
58. M. S. Dresselhaus, G. Dresselhaus, P. C. Eklund, *Science of Fullerenes and Carbon Nanotubes*, Academic Press 1996, ISB 0-12-22-221820-5.
59. S. Reich, C. Thomsen, J. Maultzsch, *Carbon Nanotubes*, WILEY-VCH Verlag, Weinheim 2004, ISBN 3-527-40386-8.
60. J. W. G. Wilder, L. C. Venema, A. R. Rinzler, R. E. Smalley, C. Dekker, *Nature* 1998, **391**, 59.
61. P. L. McEuen, *Nature* 1998, **393**, 15.
62. J. T. Sander, A. R. M. Verschueren, C. Dekker, *Nature* 1998, **393**, 49.
63. P. L. McEuen, M. S. Fuhrer, H. Park, *IEEE Transactions on Nanotechnology* 2002, **1**, 78.
64. R. Martel, T. Schmidt, H. R. Sea, T. Hertel, P. Avouris, *Appl. Phys. Lett*. 1998, **73**, 2447.
65. T. Rueckes, K. Kim, E. Joselevich, G. Y. Tseng, C. L. McEuen, *Science* 2000, **289**, 494.
66. T. W. Tombler, Z. Chongwu, L. Alxseyev, K. Jing, D. Hongjie, L. Lei, C. S. Jayanthi, T. Meijie, W. Shi-Yu, *Nature* 2000, **405**, 769.
67. D. Chattopadhyay D. I. Galeska I. F. Papadimitrakopoulos, *J. Am. Chem. Soc*. 2003, **125**, 3370.
68. R. B. Weisman, *Nature Materials* 2003, **2**, 569.
69. G. M. Whitesides, J. P. Mathias, C. T. Seto, *Science* 1991, **254**, 1312.
70. J. M. Lehn, *Science* 1993, **260**, 1762.
71. E. Baeuerlein (ed.), *Biomineralization, Process in Biology, Molecular Biology and Application*, WILEY-VCH, Weinheim, 2004, ISBN 3-527-31605-7.
72. M. E. Davis, *Materials Science* 2004, **305**, 480.
73. C. Joachim, J. K. Gimzewski, A. Aviram, *Nature* 2000, **408**, 41.
74. A. Aviram, M. A. Ratner (eds.), *Molecular Electronics II*, Vol. **960** Annals of the New York Academy of Sciences 2002.
75. M. A. Ratner, *Materials Today* 2002, **5**, 20.
76. K. S. Kwok, J. C. Ellenbogen, *Materials Today* 2002, **5**, 28.
77. M. A. Reed, J. Chen, A. M. Rawlett, D. W. Price, J. M. Tour, *Appl. Phys. Lett*. 2001, **78**, 3735.
78. Y. Luo, L. P. Collier, J. O. Jeppesen, K. A. Nielson, E. Delonno, G. Ho, J. Perkins, H.-R. Tseng, T. Yamamoto, J. F. Soddart, J. R. Heath, *ChemPhysChem* 2002, **3**, 519.
79. T. Yokoyama, S. Yokoyama, T. Kamikado, Y. Okuno, S. Mashiko, *Nature* 2001, **413**, 619.
80. J. A. Theobald, N. S. Oxtoby, M. A. Phillips, N. R. Champness, P. H. Beton, *Nature* 2003, **424**, 1029.
81. R. Fasel, M. Parschau, K.-H. Ernst, *Angew. Chem. Int. Ed*. 2003, **42**, 5178.
82. R. Fasel, M. Parschau, and K.-H. Ernst, *Nature* 2006 439, 449
83. H.-G. Kuball, T. Höfer in *Chirality in Liquid Crystals* (Eds.: H.-S. Kitzerow, C. Bahr), Springer, New York, 2001 p. 67.

Further Reading

84. J.D. Meindl, Q. Chen, J. A. Davis "Limits on Silicon Nanoelectronics for Terascale Integration" *Science* 2001 **293**, 2045
85. G. Schmid, M. Decker, H. Ernst, H. Fuchs, W. Grünwald, A. Grunwald, H. Hofmann, M. Mayor, W. Rathgeber, U. Simon, D. Wyrwa *Small Dimensions and Materials Properties: A Definition of Nanotechnology* Editor: Europäische Akademie, ISSN 1435-487 X, 2003 Print: Warlich Druck Ahrweiler GmBH, Bad Neuenahr-Ahrweiler

86. Vision 2020: Nanoelectronics at the centre of Change European Commission, 2004 ISBN 92-894-7804-7
87. Richard P. Feynman, 1959 http://www.its.caltech.edu/~feynman/plenty.html
88. Nanotechnology: Innovation for tomorrow's world European Commission, 2004 ISBN 92-894-7498 X
89. *Nanoscience and Nanotechnologies: Opportunities and Uncertainties* The Royal Society & The Royal Academy of Engineering, July 2004
90. Key Technologies for the 21st Century Scientific American, 150th Anniversary Issue, Sept. 1995

Glossary

STM	Scanning Tunnelling Microscope
AFM	Atomic Force Microscope
MEMS	Micro-Electromechanical Systems
NEMS	Nano-Electromechanical Systems
EDLC	Electrochemical Double Layer Capacitor
CNT	Carbon Nanotubes
SWNT	Single Walled Carbon Nanotubes
MWNT	Multi Walled Carbon Nanotubes
CVD	Chemical Vapour Deposition
PECVD	Plasma Enhanced Chemical Vapour Deposition
SEM	Scanning Electron Microscope
ESEM	Environmental Scanning Electron Microscope
TEM	Transmission Electron Microscope
HRTEM	High Resolution Transmission Electron Microscope
FED	Field Emission Display
VMED	Vacuum Microelectronic Devices
QD	Quantum Dot
QW	Quantum Wires
DOS	Density of States
QCL	Quantum Cascade Laser
MBE	Molecular Beam Epitaxy
CMOS	Complementary Metal Oxide Semiconductor
RAM	Random Access Memory
FeRAM	Ferroelectric Random Access Memory
MagRAM	Magneto-resistive Random Access Memory

Separation Techniques for Carbon Nanotubes

Ralph Krupke and Frank Hennrich

1
Introduction

The microelectronic revolution has been driven by the exponential enhancement of computer power over several decades, a development that had been predicted in 1965 and is known as Moore's law [1]. The progress has been enabled by lithography, a top to bottom process which allows to produce a deterministic network of transistors and wires. Advancements in lithography towards smaller feature size are ongoing, however within the next 10 to 15 years lithography might reach physical limits preventing further miniaturisation [2]. Moreover the exponentially growing lithography cost (second Moore's law) probably leads to a deviation from the actual rate of feature size shrinkage before the year 2015 [3].

A perspective towards smaller or cheaper circuitry might arise from molecular nanostructures. The developing field of nanoelectronics or molecular electronics explores the potential of organic molecules or inorganic clusters to be used as building blocks. In particular single-walled carbon nanotubes as quasi 1-dimensional molecular nanostructures have been identified as promising candidates for nanoscale transistors or nanoscale wires [4]. Yet, for applying carbon nanotubes in a future nanocircuitry, there exist several problems to be solved such as the mixed growth of metallic and semiconducting carbon nanotubes, the assembling of carbon nanotubes into functional structures or controlling the interface between a nanotube and the microscopic world. The purpose of this article is to report on recent progress on one of the grand challenges, the separation of metallic from semiconducting single-walled carbon nanotubes and to give an overview over this new emerging field.

A single-walled carbon nanotube (SWNT) can be thought of a graphene sheet seamlessly rolled up to a hollow tube with a one atomic layer wall of hexagonally arranged carbon atoms (Fig. 1). By varying the dimensions of the sheet and changing the pitch of the rolled up sheet, tubes with variations in diameter, length and chiral angle can be imagined. Apart from its length, a SWNT is defined by its roll-up or chiral vector (n,m) based on the unit vectors of the carbon atom lattice. It is remarkable that SWNTs, although composed purely out of carbon, may behave either as metals or as semiconductors just due to minute differences in the geometry. Quantum size effects determine that tubes with $n = m$ are metallic or quasi-metallic (with a tiny band-gap) if $n - m$ is divisible by 3. All others tubes are semiconducting with band gaps of the order of 0.5 eV.

Concerning nanoelectronic applications, metallic SWNTs (M-SWNT) are exceptional nanoscale leads with ballistic electron transport over several μm and current densities up to $10^9\,\mathrm{A/m^2}$. Semiconducting SWNTs (S-SWNT) can be wired up to function as fast nanoscale transistors or sensitive nanoscale sensors. For most electronic applications a preferential growth of only metallic or only semiconducting SWNTs would be highly desirable. Indeed Li and co-workers have concluded from their data that the ratio of metallic to semiconducting tubes depends on the production method, which might indicate that chirality controlled growth could be achieved in the future [5]. However in practice SWNTs are grown as mixtures of the two tube types, which makes separation schemes inevitable.

2
Separation Techniques

A pre-condition for efficient separation is that tubes are individualized within mixtures of metallic and semiconducting tubes, meaning that they do not form bundles as generally observed. This can be achieved by suspending SWNTs in aqueous surfactant solution under sonication, where many bundles split up and the individualized tubes then being hindered by the surfactant to rebundle. Remaining bundles can be removed by ultracentrifugation [6]. The scheme of O'Connell and co-workers is now used by other groups as well achieving with various surfactants such as sodium benzene sulfonate similar results. The fingerprints of a surfactant stabilized individual SWNT suspension are sharp absorption peaks in the optical absorption spectra in contrast to broad bands for suspended bundles. Furthermore surfactant stabilized SWNT suspensions show fluorescence in the near infrared region, emitted by semiconducting tubes which otherwise is quenched by the nearby metallic tubes in bundles. It should be noted that the observation of sharp absorption peaks and fluorescence indicates a rather large

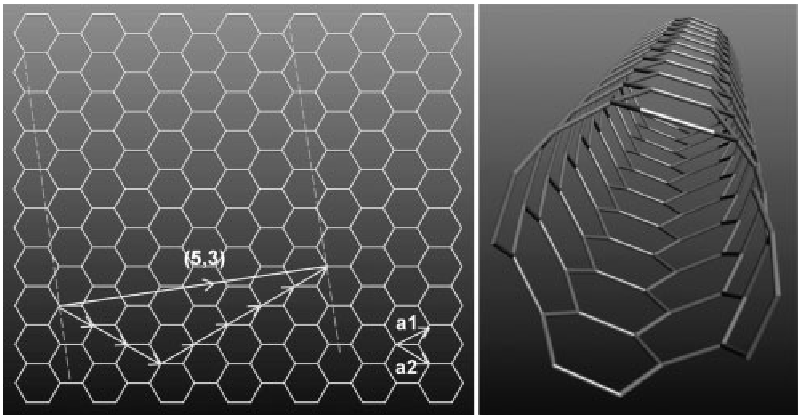

Fig. 1 A single-walled carbon nanotube is determined by its chiral vector (*n,m*) based on the unit vectors a1 and a2 of the graphene lattice, where one carbon atom is sitting at each corner of the lattice. As an example, a (5,3) SWNT (right panel) is formed by cutting a strip of a graphene sheet along the dashed lines (left panel) and rolling it up, such that the starting and the end point of the chiral vector fall on each other.

fraction of individualized SWNTs in the suspension but it is not sensitive to exclude the presence of a small fraction of bundles, an information which can be given by atomic force microscopy or transmission electron microscopy after transferring tubes onto a surface.

Before describing the separation techniques we emphasize that the analysis of supposedly separated material is not trivial. Electron transport measurements can be used on the level of a single tube to determine whether a tube is a M-SWNT or a S-SWNT and by doing so on many tubes, a statistical access to the ratio of M-SWNT to S-SWNT is given. In principle one could also determine for a single tube the (n,m) directly from STM and TEM measurements with atomic resolution. But although these measurements would give exact numbers of the M-SWNT and S-SWNT ratio, they are too time consuming as to be used for routine measurements. For those reasons most of the reviewed work is analysed by either optical absorption or resonant Raman spectroscopy. Since optical absorption cross-sections or Raman cross-sections are likely to be (n,m) dependent and are not known or under debate, both measurements do not provide absolute values for the M-SWNT and S-SWNT ratio at the moment. However the relative enrichment in M-SWNT or S-SWNT can be determined.

Optical absorption spectra can be taken either on suspended tubes or on thin films. In general absorption peaks or bands are observed in three regions of the UV-VIS-NIR spectra. Spectral features within two regions at low energy are associated with excitations of S-SWNTs (S1, S2) and an additional region at high energy coming from excitations of M-SWNTs (M1) (Fig. 2). The energetic positions of the regions depend on the mean diameter of the SWNTs. A ratio S1/M1 or S2/M2 can be obtained by dividing the integrated intensities of the peaks or the band of the regions respectively. When compared with a reference sample it allows to derive the degree of enrichment or even to calculate the sample composition, if the composition of the original reference sample is assumed to have the theoretically expected value of 2/1.

Raman spectroscopy on SWNTs has turned out to be an analysis tool which allows to characterize rather small quantities of material down to a single SWNT. In contrast to optical absorption not all tubes contribute to the spectra but only those which are resonantly excited (Fig. 2). Important features for analysing the spectra are the radial breathing modes (RBM) and the tangential modes (G-modes) of carbon atoms oscillating out of or in parallel to the nanotube surface, respectively. Although details of the spectra and their origin are still under discussion it is now widely accepted that the shape of the G-mode is significantly different for M-SWNT and S-SWNT. Furthermore, depending on the diameter distribution of the sample and the laser wavelength used, M-SWNT and S-SWNT with different diameter are excited, which splits the RBM range into regions associated with M-SWNT and S-SWNT. Estimate of the enrichment of

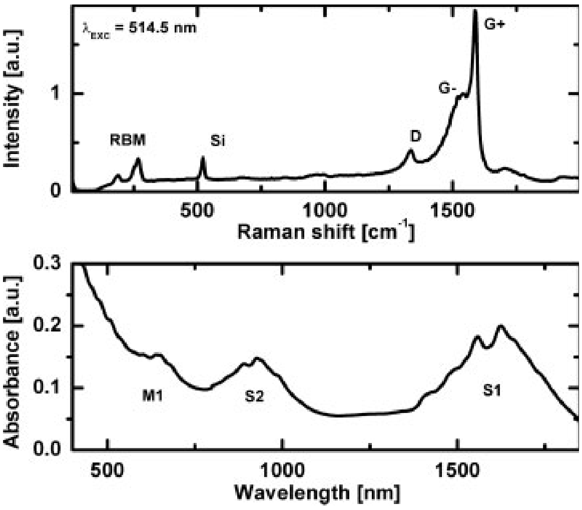

Fig. 2 Typical Raman- (top) and UV-VIS-NIR spectra (bottom) used for analysing the degree of separation or enrichment in metallic or semiconducting SWNTs.
Bottom: S1 and S2 denote the first and second excitation energies of semiconducting SWNTs, respectively. Due to inverse scaling of the excitation energy with the diameter, the position and the width of S1 and S2 depends on the mean diameter and diameter distribution of SWNTs in the sample. M1 denotes the first excitation energies of metallic SWNTs. The steep increase of absorbance below 500 nm is due to higher order excitations and the pi-plasmon. The sharp wiggles are only observed for samples with a large content of unbundled tubes.
Top: SWNTs which are resonantly excited contribute to Raman spectra comprising of radial breathing modes (RBM), the disorder mode (D) and the tangential modes (G+ and G−). In this particular example, two regions of RBMs are observed, coming from the simultaneous excitation of a few metallic and semiconducting SWNTs with different diameter. Also observed is a sharp and intense G^{+}-mode, typical for semiconducting SWNTs, and a broad G^{+}-mode, typical for metallic SWNTs (see also Fig. 4).

a sample is then done in a similar fashion as for the optical absorption measurement by calculating the ratio of the integrated peak intensities of RBM bands originating from M-SWNT and S-SWNT and comparing it with a reference sample. Furthermore variations in the shape of the G-mode for the two tube types can be used for analysis. Of course for optical absorption and Raman spectroscopy it is important to compare samples of similar morphology. For instance spectral changes can be observed upon bundling which are not related to compositional changes.

In the following separation techniques are promoted which, as a starting material, use individual SWNT suspensions with the above discussed optical properties. Then techniques are described which are using suspensions of bundled SWNTs and finally alternative concepts are presented.

2.1
Separation via Dielectrophoresis

Separating metallic from semiconducting SWNTs, starting from individual tube suspensions, was first obtained via alternating current dielectrophoresis [7]. The physical method makes use of the different polarisabilities of M-SWNT and S-SWNT induced by an electrical field. Theoretically it was expected that the dielectric constant of S-SWNT is of the order of $5{\cdot}\varepsilon_0$ compared to $80{\cdot}\varepsilon_0$ for aqueous surfactant solution. For M-SWNT the dielectric constant should be at least of the order of $1000{\cdot}\varepsilon_0$. The dielectric constants of S-SWNT and M-SWNT being smaller and larger than that of water gives access to separate the two tube types by exposing them to a strong and inhomogeneous electrical field. The interaction of the induced dipole moment with the inhomogeneous external field leads to a movement of the M-SWNT towards the high field region whereas the S-SWNT move into the opposite direction towards the low field region (Fig. 3).

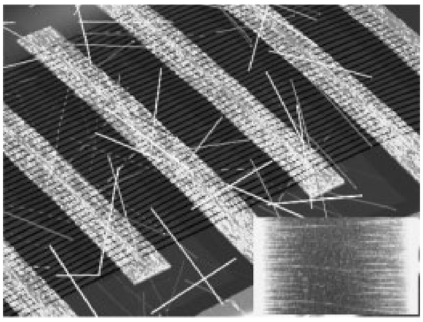

Fig. 3 Illustrated dielectrophoretic separation of metallic from semiconducting SWNTs. Planar microelectrodes, powered by an ac-voltage generator, impose a strong dielectrophoretic force on SWNTs suspended in aqueous solution. The highly polarizable metallic SWNTs (black) are dragged to the electrodes whereas the less polarizable semiconducting SWNTs (white) remain in suspension. The separation is observed for a nominal electrical field of 10^6 V/m and electric field frequencies above 3 MHz[17]. The inset shows an optical microscopy image of aligned metallic SWNTs forming bundles and ropes, eventually bridgeing the electrode gap.

This dielectrophoretic response is competing with the electrophoretic response which is due to the interaction of the field with charges sitting on the tubes when suspended in aqueous solution. This static charge accumulation does not only depend on the intrinsic electronic properties but also on the diameter and length and therefore is not an appropriate property to be used for separating M-SWNT from S-SWNT. To avoid an electrophoretic response an alternating electrical field of a few MHz has been used, where the oscillatory electrophoretic movement in the ac field results in a zero displacement of the tubes. In contrast the dielectrophoretic response leads to a net opposite movement of the two tube types since their induced dipole moment is always in phase with the ac field even at the highest frequency used in the experiment.

The experimental setup comprises of a planar interdigitated microelectrode array with electrode gaps of typically 20 μm which is powered by an ac power generator with a sinoidal waveform and a peak to peak voltage of typically 20 V. The setup produces a strong and inhomogeneous ac electrical field with a nominal field strength of the order of 10^6 V/m. Applying a drop of individual SWNT suspension onto the array leads within minutes to the deposition of metallic SWNTs forming bundles at the surface, which are growing along the field lines and eventually connect the opposing electrodes. The deposited material has been analyzed by resonant raman spectroscopy and showed that primarily M-SWNT had been deposited. In the original work it had been estimated that 80% of the deposited material are M-SWNT. By further reducing the residual bundle content in the starting suspension, a complete separation has been achieved in the sense that by Raman spectroscopy no S-SWNT can be detected in the dielectrophoretic samples (Fig. 4).

It should be noted, that in a similar experiment, albeit with bundled material, Lee and co-workers have found that S-SWNTs are strongly enriched in the leftover suspension, as expected [8].

In the proof-of-principle experiment about 1 ng of tubes can be separated from a 10 μl drop of tube suspension per run. For the technique to become an industrial technology at least 6 orders of magnitude in separation capacity have to be gained. To achieve this goal the technique has to be scaled up. A problem is that the dielectrophoretic force reduces within the first 10 μm above the electrodes by two orders of magnitude leaving most of the tubes in the drop unexposed to sufficiently strong dielectrophoretic forces. However, a microfluidic setup with the electrode array being embedded in a shallow channel would allow to expose the suspension to sufficiently strong dielectrophoretic forces.

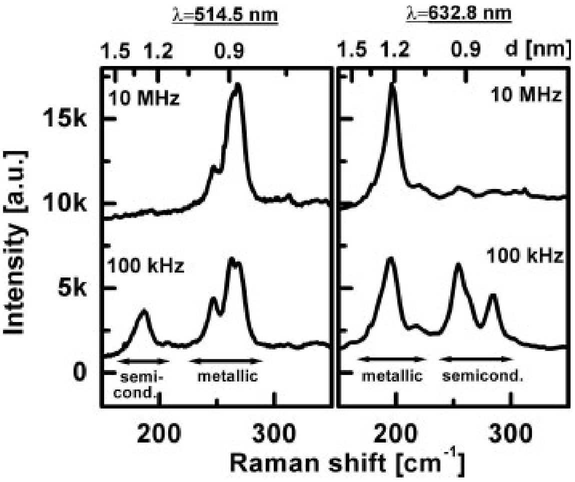

Fig. 4 Comparison of SWNTs deposited with ac-dielectrophoresis at different field frequencies (10 MHz and 100 kHz) by analyzing the Raman signal in the Radial Breathing Mode region with two laser excitation wavelengths λ = 514.5 nm and λ = 632.8 nm (left and right panel, respectively). At 514.5 nm small metallic SWNT with d $\approx$ 0.9 nm and large semiconducting SWNTs with d $\approx$ 1.2 nm are excited[7]. At 632.8 nm large metallic SWNTs with d $\approx$ 1.2 nm and small semiconducting SWNTs with d $\approx$ 0.9 nm are probed. The data demonstrates, that separation of metallic from semiconducting SWNTs is obtained for high frequency ac-dielectrophoresis at 10 MHz, independent from the tube diameter.

2.2
Chemical Selectivity to Metallic Tubes

More easily scalable would be a chemical access to separate M-SWNT from S-SWNT. Indeed M. Strano and co-workers reported that diazonium reagents, under carefully controlled conditions, primarily reacts with M-SWNT [9]. Also here the starting materials were surfactant stabilized individual SWNT suspensions. Confirmation of the selective reaction is given by optical absorption spectroscopy where mainly the absorption peaks associated with M-SWNT initially decay under addition of the diazonium reagents. Further confirmation is given by the evolution of the Raman spectra with increasing diazonium agent concentration, which shows that small M-SWNT are the most reactive species.

The model used to explain the selective reactivity describes the diazonium agent extracting electrons, thereby evolving N_2 gas and leaving a stable C-C covalent bond with the nanotube surface. It is assumed that the availability of electrons near the Fermi level stabilizes a charge-transfer state that is preceding the bond formation, which finally leads to a higher initial reactivity towards M-SWNT. However the reactivity is also anticipated to depend on the chiral angle, chemical doping and charged adsorbates.

Having found a chemistry which appears to be selective towards primarily M-SWNT it is now the challenge to extract the M-SWNT from the still mixed suspension. In that sense the M-SWNT are not yet separated from the S-MWNT but have been modified chemically such, that they are no more detectable by optical spectroscopy.

A selective electrochemical modifications of those M-SWNTs, which have been wired to electrodes on a surface, has been achieved by Balasubramanian and co-workers [10]. The authors used similar to Strano et al. a diazonium agent to functionalize M-SWNT. However in this experiment the reaction was driven by an electrical potential applied between electrodes in touch with tubes and a counter electrode. The authors found conditions in which primarily M-SWNTs did react, once the S-SWNTs were driven into the non-conducting state by an appropriate gate voltage. Experimental verification of the selective modification is given by transport measurements in which the signatures of M-SWNTs disappear after the treatment.

2.3
Selection via Ion Exchange Chromatography

Another chemical approach is to separate SWNTs by anion exchange chromatography. M. Zheng and co-workers have found out that single-stranded DNA interacts strongly with SWNT to form a stable DNA-SWNT hybrid that effectively disperses SWNTs in aqueous solution under sonication [11]. Such DNA-SWNT suspensions do show similar sharp optical absorption peaks and fluorescence as surfactant stabilized individual tube suspensions. Since the phosphate groups on a DNA-SWNT hybrid provide a negative charge density on the surface of a nanotube, the authors anticipated that a DNA-(M-SWNT) has less surface charge than DNA-(S-SWNT) due to screening of the charge by the M-SWNT. To take advantage of the difference for SWNT separation, an ion-exchange liquid chromatography experiment was performed, where the difference in surface charge should modify the interaction of the hybrid with the resin and hence the propagation speed through the column. Experimentally, the chromatogram (the evolution of the absorbance with time) shows two bands due to free DNA (first band) and DNA-SWNT (second band). In the initial experiment two fractions of the DNA-SWNTs bands were analysed, indicating by optical absorption measurements an enrichment of M-SWNT in the early fraction. In addition length variations of the two fractions were analyzed by atomic force microscopy, indicating that SWNTs in the earlier fraction are shorter. Hence the conclusion was that M-SWNT and S-SWNT separation is convoluted by the tube length variations.

More recently the authors conducted a systematic study whether the formation of the DNA-SWNT hybrid is sequence dependent [12]. The result was that DNA sequences rich in guanine and thymidine amino acids are particularly useful for the formation of stable aqueous suspensions. These suspensions were then split up into various fractions by the anion exchange liquid chromatography. Optical absorption spectra again revealed an enrichment in M-SWNT in the early fractions but now convoluted with a variation in diameter. SWNTs with a smaller diameter pass the column faster than SWNTs with a larger diameter. Recently M. Strano with M. Zheng and co-workers performed an extensive analysis on the very same samples by combined fluorescence and raman spectroscopy [13]. The result is that the separation of the suspensions by anion exchange liquid chromatography is primarily according to the diameter of the SWNTs, independent of the metallic or semiconducting nature of the SWNTs. The results are certainly complex, but definitely show that ion exchange chromatography allows to separate DNA-SWNT hybrids at least according to their diameter.

2.4
Selective Destabilisation of Tube Suspensions

We now come to separation techniques which use suspensions of bundled SWNT as starting material. It is obvious that separation can not be achieved if bundles are of heterogeneous composition, however for small bundles some degree of enrichment is possible. For instance Chattopadhyay and co-workers reported an enhanced solubility of S-SWNT when dispersed with octadecylamine (ODA) in tetrahydrofuran (THF) [14]. The authors assume that ODA physisorbs along the SWNT sidewalls, with the amino group having a stronger affinity to S-SWNT than to M-SWNT, and thereby providing additional stability to the physisorbed ODA. In contrast M-SWNTs were more likely to precipitate.

In the experiment ODA stabilized SWNT suspensions were prepared in THF with bundles containing between 2-6 SWNTs. Over a period of a few weeks sedimentation of parts of the ODA-dispersed SWNTs takes place, which can be accelerated to minutes by solvent evaporation. Analysis of the sediment by Raman spectroscopy indicates an enrichment in M-SWNT as compared to the supernatant.

Recent quantitative analysis of the Raman spectra on such samples by G. Samsonidze and co-workers revealed that the separation efficiency is convoluted by variations in the tube diameter [15]. In particular poor enrichment is observed for larger-diameter tubes, which becomes more efficient towards smaller tube diameter. The origin of this diameter dependence on the enrichment is not yet understood but the authors point onto electronic amine-SWNT interactions or curvature-induced stabilization.

In another work, Z. Chen and co-workers have added dropwise diluted bromine to Triton X-100 suspended SWNT and SWNT bundles, and observe a sedimentation of tube material enriched in M-SWNT [16]. No enrichment is observed for experiments using sodium dodecyl sulfonate as the surfactant. The authors take this as evidence that the enrichment mechanism involves bromine induced destabilization of the Triton X-100 surfactant shell, with greater specificity for the M-SWNT. The enrichment is shown by comparison of optical absorption spectra of films from bromine treated supernatant and sediment after thermal removal of the bromine.

3
Summary

It remains a challenge for separation methods to produce substantial quantities of suspensions containing only metallic SWNTs or only semiconducting SWNTs of pristine quality for applications. Certainly a chemical access to separation appears to be promising, since a chemical recipe should be easily scalable. However up to now a spatial separation of metallic SWNTs from semiconducting SWNTs by chemical means has not been obtained. Either metallic SWNTs have been selectively functionalized but still being in the treated sample and hence not spatially separated from the semiconducting SWNTs. Or fractions enriched with metallic or semiconducting SWNTs have been obtained, however with the degree of enrichment being diameter dependent. Also the dielectrophoretic separation, a physical method, exhibits a diameter dependence if performed at low electric field strength, and only enriches M-SWNTs if processed at low frequency [17]. However for sufficient high field strength and field frequency a spatial separation of metallic SWNTs from semiconducting SWNTs is observed, independent from the diameter (Fig. 4). Under these separation conditions only the different intrinsic polarisabilities of the two tube types matter.

It might well be that upon further improvement not one particular method will prevail but rather a smart combination of methods, adjusted to specific applications. Certainly it remains exciting to see what new methods will be discovered in this emerging field of nanotube separation. In this context we refer to recent progress in electronic type selective separation of nanotubes by density gradient centrifugation [18], chemical treatment [19, 20], and dielectrophoresis [21].

References

1. G. E. Moore, *Electronics* 1965, **38**, 114.
2. International Technology Roadmap For Semiconductors 2003, http://public.itrs.net.
3. G. E. Moore, *Proceedings of the SPIE* 1995, **2440**, 2–17.
4. P. McEuen, *Physics World* 2000, **13**, 31, and references therein.
5. Y. Li, D. Mann, M. Rolandi, W. Kim, A. Ural, S. Hung, A. Javey, J. Cao, D. Wang, E. Yenilmez, Q. Wang, J. F. Gibbons, Y. Nishi, H. Dai, *Nano Lett.* 2004, **4**, 317.
6. M. J. O'Connell, S. M. Bachilo, C. B. Huffman, V. C. Moore, M. S. Strano, E. H. Haroz, K. L. Rialon, P. J. Boul, W. H. Noon, C. Kittrell, J. Ma, R. H. Hauge, R. B. Weisman, R. E. Smalley, *Science* 2002, **297**, 593.
7. R. Krupke, F. Hennrich, H. v. Löhneysen, M. M. Kappes, *Science* 2003, **301**, 344.
8. D. S. Lee, D. W. Kim, H. S. Kim, S. W. Lee, S. H. Khang, Y. W. Park, E. E. B. Campbell, *Appl. Phys. A* 2005, **78**, 283, DOI: 10.1007/s00339-004-2992-4.
9. M. S. Strano, C. A. Dyke, M. L. Usrey, P. W. Barone, M. J. Allen, H. W. Shan, C. Kittrell, R. H. Hauge, J. M. Tour, R. E. Smalley, *Science* 2003, **301**, 1519.

10. K. Balasubramanian, R. Sordan, M. Burghard, K. Kern, *Nano Lett.* 2004, **4**, 827.
11. M. Zheng, A. Jagota, E. D. Semke, B. A. Diner, R. S. McLean, S. R. Lustig, R. E. Richardson, N. G. Tassi, *Nat. Mat.* 2003, **2**, 338–342.
12. M. Zheng, A. Jagota, M. S. Strano, A. P. Santos, P. Barone, S. G. Chou, B. A. Diner, M. S. Dresselhaus, R. S. McLean, G. B. Onoa, G. G. Samsonidze, E. D. Semke, M. Usrey, D. J. Walls, *Science* 2003, **302**, 1545.
13. M. S. Strano, M. Zheng, A. Jagota, G. B. Onoa, D. A. Heller, P. W. Barone, M. L. Usrey, *Nano Lett.* 2004, **4**, 543.
14. D. Chattophadhyay, I. Galeska, F. Papadimitrakopoulos, *J. Amer. Chem. Soc.* 2003, **125**, 3370.
15. G. G. Samsonidze, S. G. Chou, A. P. Santos, V. W. Brar, G. Dresselhaus, M. S. Dresselhaus, A. Selbst, A. K. Swan, M. S. Unlu, B. B. Goldberg, D. Chattopadhyay, S. N. Kim, F. Papadimitrakopoulos, *Appl. Phys. Lett.* 2004, **85**, 1006.
16. Z. Chen, X. Du, M.-H. Du, C. D. Rancken, H.-P. Cheng, A. G. Rinzler, *Nano Lett.* 2003, **3**, 1245.
17. R. Krupke, F. Hennrich, M. M. Kappes, H. v. Löhneysen, *Nano Lett.* 2004, **4**, 1395.
18. M. S. Arnold, A. A. Green, J. F. Hulvat, S. I. Stupp, M. C. Hersam, *Nature Nano.* 2006, **1**, 60
19. Y. Maeda, S. Kimura, M. Kanda, Y. Hirashima, T Hasegawa, T. Wakahara, Y. Lian, T. Nakahodo, T. Tsuchiya, T. Akasaka, J. Lu, X. Zhang, Z. Gao, Y. Yu, S. Nagase, S. Kazaoui, N. Minami, T. Shimizu, H. Tokumoto, R. Saito, *J. Am. Chem. Soc.* 2005, **127**, 10287
20. Y. Miyata, Y. Maniwa, H. Kataura, *J. Phys. Chem. B* 2006, **110**, 7316
21. R. Krupke, S. Linden, M. Rapp, F. Hennrich, *Adv. Mat.* 2006, **18**, 1468

Glossary

Single-walled carbon nanotube A single-walled carbon nanotube is a single sheet of graphite (graphene) rolled-up to a seamless cylinder. Its electronic property depends on the diameter and chirality due to quantum confinement of the electron wave function around the nanotube circumference. So far the electronic properties of nanotubes have not been controlled during growth.

Raman spectroscopy Raman spectroscopy is a spectroscopic technique used to study vibrational, rotational, and other low-frequency modes in a system. It relies on inelastic scattering, or Raman scattering of monochromatic light, usually from a laser in the visible, near infrared, or near ultraviolet range. The laser light interacts with phonons or other excitations in the system, resulting in the energy of the laser photons being shifted up or down. The shift in energy gives information about the phonon modes in the system. In single-walled carbon nanotubes the Raman spectra are resonantly enhanced due to singularities in the electronic density of states.

Dielectrophoresis Dielectrophoresis is the translational motion of a polarisable object in an inhomogenous electric field. Dielectrophoretic forces are usually much weaker than electrophoretic forces, which depend on the net charge of the object interacting with the electric field. Electrophoretic motion is suppressed in alternating current electric fields where dielectrophoresis can still be observed.

Ion exchange chromatography Ion-exchange chromatography is a process that allows the separation of dissolved ions and polar molecules based on their charge properties. Dissolved molecules (analyte) are driven with a mobile phase (solvent) through a stationary phase (resin) where retention times depend on the interaction between the analyte and the stationary phase.

Density gradient centrifugation A method of separating macromolecules by their differential rate of sedimentation in a centrifugal gradient and by their differential bouyancy in a density gradient. At equilibrium each particle settles in the gradient at a point equal to its density.

Ceramic Materials for Advanced Solid Oxide Fuel Cells

Peter Holtappels, Ulrich Vogt and Thomas Graule

1
Introduction

Although the SOFC has been known since 1911 and a first concept for large scale stationary SOFC systems was presented in 1937 by Baur and Preis [1], the first significant worldwide development wave did not start until the 1960 s when the Batelle Institutes in Switzerland (Geneva) and Germany (Frankfurt) began developing the technology. A detailed historical overview is given by Möbius [2].

The status of particular SOFC materials and cell development up to 1993 has been described in detail by Minh [3] and Hammou [4]. Of these early developments, only the tubular concept developed at Westinghouse (now Siemens-Westinghouse) has been pursued up to the present day. During the 1980s, advances in ceramics processing led to flat plate (planar) cell concepts, which compared to the Siemens-Westinghouse cells offer higher volumetric power densities.

SOFC development today covers cells and systems for stationary combined heat and power (CHP) production from 1 $\mathrm{kW_{electric}}$ for single-household applications to 200 $\mathrm{kW_{electric}}$ for decentralised and industrial CHP applications. Sulzer HEXIS [5] and Siemens-Westinghouse are two of the leading developers of stationary systems in Europe and the US, respectively. Japan also has considerable programmes devoted to stationary SOFCs [6]. Towards the end of the 1990s, BMW announced the development of SOFC systems to replace the alternator in cars, and this promoted increasing interest in SOFC systems for automotive applications worldwide. For detailed information the reader is referred to the most recent reviews [7, 8].

Current developments in SOFCs concentrate mainly on increasing the durability and lowering the cost of the system and the materials therein. A way to reach these goals is to decrease the operating temperature without incurring performance losses rather than further increasing the performance of the cell.

The scope of this contribution is to provide a general overview and selected examples of the most recent trends in materials development for advanced solid oxide fuel cells operating at temperatures around 800 °C.

digital Encyclopedia of Applied Physics.
Copyright © Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim. All rights reserved.

2
Fundamentals of Solid Oxide Fuel Cells

The operating principle of an SOFC is shown in Figure 1. The conversion of chemical into electrical energy is achieved by separating a chemical reaction into two electrochemical reactions taking place inside the two porous electrodes. At the fuel electrode (anode), hydrogen and carbon monoxide are oxidised with oxygen ions from the electrolyte to form water and carbon dioxide, respectively. The resulting electrons are transported through an external circuit to the air electrode (cathode) where they are used to reduce oxygen to oxygen ions, which are subsequently transported through the electrolyte to the fuel electrode.

The fuel and air atmospheres have different electrochemical potentials related to the difference in the oxygen partial pressures in the fuel and air compartments. The resulting potential difference across the cell is the so-called Nernst potential. A typical fuel gas mixture corresponding to partially-oxidised methane (natural gas) contains 0.23 bar hydrogen, 0.2 bar carbon monoxide, 0.03 bar water and 0.06 bar carbon dioxide, with nitrogen forming the balance to 1 bar. The resulting oxygen partial pressure is 10^{-20} bar at 850 °C. The difference against the air compartment (0.21 bar oxygen) leads to a voltage across an unloaded cell of 1.02 V.

During operation of the fuel cell (external electrical circuit closed), a current can be generated which is proportional to the electrochemical conversion rate of hydrogen and carbon monoxide to water and carbon dioxide, respectively. However, the electrode reactions and the oxygen ion transport through the electrolyte lead to losses that reduce the cell voltage with increasing current flow. In order to achieve high electrical power ($P = \mathcal{U} \cdot i$), the losses in both the electrodes (the so-called anodic ($\ddot{\eta}_\mathrm{a}$) and cathodic ($\ddot{\eta}_\mathrm{c}$) overpotentials, see Fig. 1) and the ohmic losses in the electrolyte layer ($i\,R_{\ddot{\alpha}1}$x Fig. 1) should be minimised. Therefore, highly electrochemically-active electrode materials and fast oxide ion-conducting materials are required for use as electrodes and the electrolyte, respectively.

The cell components, along with their function, specific requirements and the typical materials used to manufacture them are summarised in Table 1.

The solid oxide fuel cell can be regarded as a laminated structure of three different functional materials. The cell, therefore, is an all-solid-state system for which different designs may be realised. An overview of the various cell designs can be found in earlier reviews [4, 7, 8].

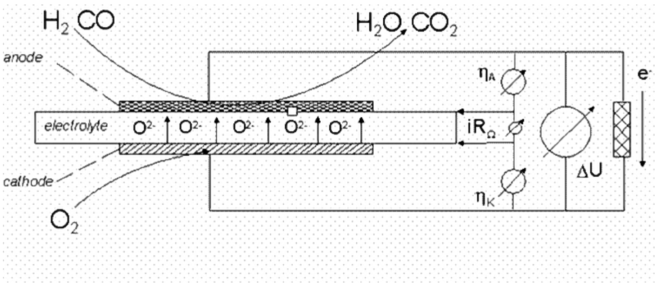

Fig. 1 Principle of a SOFC (including the partial electrode reactions).

Tab. 1 Overview of materials and requirements for SOFC cell components

Component	***Function***	***Requirements***	***Materials***
Cathode	$O_{2,g} + 4e^- + 2V_o^{\cdot\cdot} = 2O_o^x$	Long triple-phase boundary	Perovskites
	gas transport	Activity for oxygen surface exchange	$La_{1-x}Sr_xMnO_3$ (LSM)
	current pick-up	Porosity	$La_{1-x}Sr_xCo_{1-y}Fe_yO_3$ (LSCF)
		Electronic conductivity	$La_{1-x}Sr_xFeO_3$ (LSF)
Anode	Electrocatalytic activity	Long triple-phase boundary	Cermets
	$Red_g + nO_o^x = OxO_n + nV_o + 2ne^-$	Activity for oxygen surface exchange	Ni-$Y_{0.15}Zr_{0.85}O_{1.93}$
	gas transport	Porosity	(Ni-YSZ)
	current pick-up	Electronic conductivity	
		Redox stability	
		Sulphur tolerance	
Electrolyte	oxygen ion transport	Oxygen ion conductivity	$Y_{0.15}Zr_{0.85}O_{1.93}$
	electronic insulator	Gas tightness	(8-YSZ)
		Mechanical stability	

Both tubular and flat plate cells with either the electrolyte, or an electrode, forming the mechanically-supporting part of the cell have been developed up to a high level of performance. Reducing the electrolyte thickness by using supported electrolyte layers (e.g. by having the anode as the electrolyte-bearing component, as shown in Fig. 2) permits the ohmic losses in the electrolyte layer to be reduced and makes lower operating temperatures around 700–850 °C feasible. At operating temperatures below 850 °C, ferritic steels can be used as interconnect materials, and this significantly reduces the cell material costs.

Several concepts and designs for supported thin electrolyte cells have been pursued, with the anode-supported electrolyte cells having attracted particular interest during recent years.

After the Forschungszentrum Jülich demonstrated the feasibility of the anode-supported electrolyte design [9], most SOFC developers worldwide started their own development line of anode-supported electrolyte cells. Great progress has been made to improve the performance of this cell type, with state-of-the-art cells reaching up to 1 W/cm^2 at 750–850 °C (0.5 W/cm^2 at 0.7 V) in hydrogen fuels [10]. Stacks up to 10 kW have been successfully tested with hydrogen and with pre-reformed natural gas.

Today, most R&D activities worldwide are concentrating on a flat plate planar cell design with the fuel electrode (anode) as the cell-supporting component. This concept combines a high volumetric power density with the capability to operate at reduced temperatures around 800 °C. In this contribution we will focus on the main

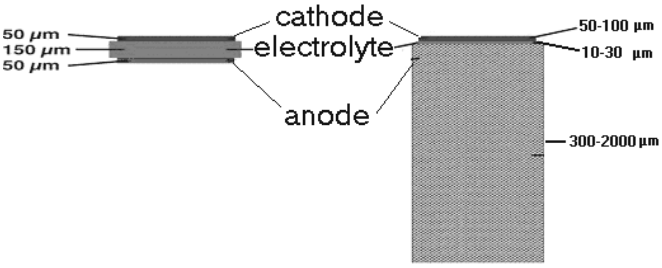

Fig. 2 Comparison of self-supported electrolyte cells (left) and anode-supported electrolyte cells (right) in the flat plate design.

development paths for materials and cell fabrication of advanced supported electrolyte cells which feature such reduced operating temperatures.

3
Materials for Advanced Solid Oxide Fuel Cells

3.1
Anode

At the anode, the fuel is oxidised in the presence of oxygen ions from the electrolyte according to Eq. 1.

$$\mathrm{Re}d_g + n\ O_O^x = Ox + n\ V_O^{\cdot\cdot} + 2n\ e^- \quad (1)$$

Equation 1 is written in the so-called Kröger-Vinck notation, where the subscript indicates the site in the ionic lattice and the superscript the electrical defect compared to the regular lattice. Accordingly, $O^x{}_O$ denotes an oxygen ion (oxidation state-II) on an oxygen site in the oxide electrolyte lattice, and $V_{O^{\cdot\cdot}}$ indicates a vacant oxygen site in the electrolyte lattice (representing a double positive charge in comparison to an occupied site). *Red* and *Ox* refer to the reduced and oxidised fuel species, respectively, and the subscript *g* refers to the gas phase (see Table 2). 2 *n* denotes the number of electrons involved in the electrochemical reaction.

Under reducing atmospheres in the fuel electrode, metals are stable over a wide range of operating conditions. For SOFCs operating at temperatures between 900 °C and 1000 °C, several materials and alloys have been investigated, and Ni has become the material of choice due to its high electrochemical activity for the hydrogen oxidation reaction, acceptable cost, and suitable chemical compatibility. In order to adjust the coefficient of thermal expansion (CTE) of a Ni-containing anode to that of the electrolyte (yttrium stabilised zirconia, YSZ), Ni powder is mixed with the electrolyte material to form a ceramic-metal composite, the so-called Ni-YSZ cermet. In the Ni-YSZ cermet, the ceramic forms a matrix which prevents the Ni powder from agglomerating and thus maintains the porous and highly dispersed microstructure of the anode. At Ni contents of around 40 vol%, the cermets are electronic conductors (500–1800 S/cm) and exhibit a thermal expansion coefficient close to that of the pure electrolyte material YSZ.

Up to now, Ni-YSZ cermets have been the state-of-the art anode material providing high performance hydrogen oxidation even at temperatures of 800 °C [11]. A measure for the electrochemical activity is the

Tab. 2 Reacting species in SOFC anodes

Red	*Ox*	*No. of electrons transferred in the overall electrochemical reaction*
H_2	H_2O	2
CO	CO_2	2
CH_4	$CO_2 + 2H_2O$	8
CH_4	$CO + 2H_2O$	6

area specific resistance, R_{as}, which reflects the energetic losses arising from the electrochemical processes in the electrode. The area specific resistance is usually expressed in $\ddot{\alpha}$ cm^2. The electrochemical activity of Ni-YSZ electrodes is such that the area specific resistance reaches values of ≈ 0.15 $\ddot{\alpha}$ cm^2 at 850 °C and <0.1 $\ddot{\alpha}$ cm^2 in the range 950–1000 °C.

The hydrogen oxidation reaction in Ni-YSZ cermet electrodes has been investigated for both technical electrodes and model systems. The results reveal that the reaction is limited by the length of the so called triple-phase boundary (TPB) region [12, 13], where the electronic conductor Ni, the oxygen ion conductor YSZ and the gas phase are in contact with each other. Possible steps in the hydrogen oxidation reaction in the vicinity of the TPB-line are adsorption, surface diffusion, and charge transfer on both the Ni and the YSZ electrolyte surface, as shown in Figure 3. Even though the detailed anode reaction mechanism is not yet fully understood, optimisation of the length of the triple-phase boundary line by reducing the particle and pore size in the electrochemically-active anode layer has led to high performance SOFC cells.

While a fine-grained microstructure is necessary for the Ni-YSZ anode to achieve maximum performance, the resulting high interfacial surface area is susceptible to degradation and is therefore a serious obstacle to achieving an anode operating life of several thousands of hours under varying conditions (e.g. gas phase, electrical loading, temperature cycles). In particular the high susceptibility of Ni to coking, the re-oxidation in case of fuel loss and the sensitivity to poisoning of the electrode by sulphur are the main concerns regarding the durability and degradation of Ni-YSZ anode-supported SOFC cells.

As an alternative to Ni-YSZ cermet anodes, the focus is on electronic and mixed-ionic electronically-conducting ceramics. Developments are concentrating on mixed conducting fluorites and perovskites, and also on more complex structures such as pyrochlores, layered perovskites and tungsten bronzes [14]. The ceramics in general exhibit stability with respect to redox-cycles (fuel-air) as well as a very low, sometimes even negligible, coking activity.

Of these alternative materials, only ceria-based anodes, such as $Ce_{0.6}Gd_{0.4}O_{1.8}$ [15], and $La_{0.8}Sr_{0.2}Cr_{0.5}Mn_{0.5}O_3$ (LSCM) [14] have been developed up to the point where their performance is competitive with the Ni-YSZ cermet. By dispersing a Ni catalyst throughout the porous ceria structure, area specific resistances less than 0.06 $\ddot{\alpha}$ cm^2 at 1000 °C in hydrogen have been achieved. This is similar to the values observed in Ni-YSZ anodes. No carbon deposition has been observed at high methane content, and redox as well as long-term stability are proven. The ceria electrodes

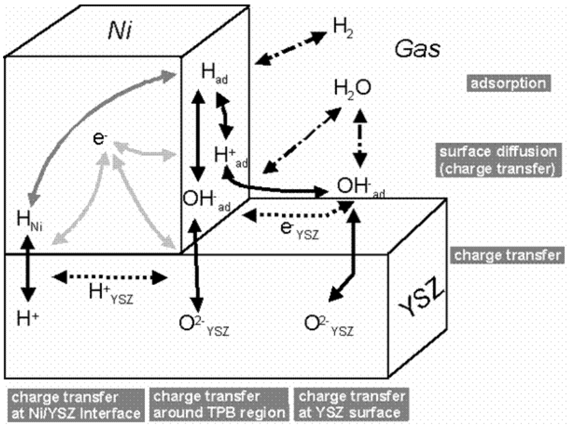

Fig. 3 Possible reaction steps in the hydrogen water reaction around the Ni/YSZ/gas exposed triple-phase boundary line.

do not fully match the CTE of the YSZ electrolyte and the electrode expands upon reduction. The problems resulting from the thermo-mechanical stresses at the ceria/YSZ interface have been overcome by introducing an anchoring construction at the interface to the YSZ electrolyte.

A disadvantage of the mixed conducting ceramics such as ceria or LSCM is the relatively low electronic conductivity which largely prevents the use of these materials at SOFC operating temperatures around 800 °C. This problem may be overcome by using cermets, in which a metallic phase acts as the current collector, instead of pure ceramic materials. Ceriabased composite anodes Cu/YSZ/ceria [16] and Ni/ceria [17] are being developed for direct application of hydrocarbon fuels. Direct feeding of methane, propane and higher hydrocarbons up to decane without carbon deposition has been reported.

With respect to compatibility with the YSZ solid electrolyte, yttrium-titania modified zirconia (YZT) is also a promising mixed-conducting fluorite, which has recently been investigated as a potential anode material. [18] The development of these advanced cermets is still at a rather fundamental stage and the materials remain to be demonstrated in cells of technologically-relevant size.

3.2
Anode Support

Due to the high electronic conductivity of Ni-YSZ cermets, relatively thick layers can easily be used as the supporting component in the SOFC cell. Besides the ability to pick up the current from the anode, the support has to be sufficiently mechanically stable and porous in order to allow the unlimited transport of the gaseous educts and products to and away from the anode. The thickness of the anode supports varies with the fabrication technique. Anode support layers of 300 ά m up to 2000 ά m thickness have been prepared by tape casting and die pressing, respectively.

The pressing technique seems to be an easy and upscalable fabrication procedure for the anode supports. Both warm and cold die pressing have been investigated. Ceramic compositions produced by the Coat Mix process have been successfully developed for warm die pressing at the Forschungszentrum Jülich and form the basis for their anode-supported SOFC development [9].

An alternative process is the cold die pressing of polymer-bonded Ni-YSZ granules. Cold die pressing has been suggested as an upscalable and cost effective technique for the fabrication of various Ni-YSZ cermet anode supports [19, 20]. The method is environmentally friendly due to the comparatively low amount of harmless organic binders used, and is upscalable for mass production. No complex debinding or drying procedure is needed prior to the firing step. The porosity of these anode supports can be tailored by controlling the particle sizes of the NiO and YSZ powders and by including additional pore-forming agents. Graphite in the form of plates and spheres with particle sizes (d_{50}) ranging from 5–50 ä m has been shown to be a suitable pore-former for water-based dispersions [20].

An important step in the fabrication of anode-supported cells and also multilayer anode supports is the cofiring of at least two layers (the electrolyte and the anode), and higher orders of multilayers. In order to avoid bending and warping of the samples, the shrinkage of the individual layers during sintering should be matched. Figure 4 shows the variation of the free shrinkage for various NiO-YSZ compositions with and without carbon as a pore former.

By controlling the shrinkage and the porosity of individual Ni-YSZ cermet compositions, multilayer structures with a gradually changing pore size have been prepared as shown in Figure 5. The concept of a gradually changing porosity in the anode support allows individual tailoring of the supporting anode layer and the electrochemically-active layer. While a coarse and open porous structure is essential for gas transport inside the anode support, much smaller pore and particle sizes leading to a greater TPB length are required towards the electrochemically-active layer. First results [20] have validated the feasibility of the concept and of the fabrication process. However, the microstructure in Figure 5 exhibits rather isolated

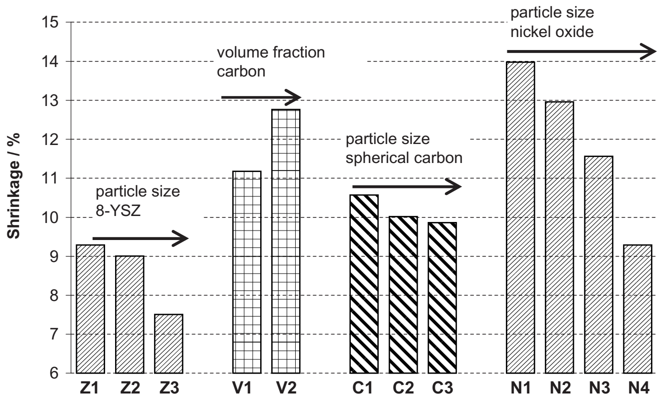

Fig. 4 Influence of the particle sizes and the amount of pore formers on the shrinkage of NiO-YSZ anode supports. D_{50}8-YSZ: Z1, 12.9 ä m; Z2, 16.7 ä m; Z3, 35.6 ä m; Vol% C: V1, 10%; V2, 30% ä m; D_{50} carbon: C1, 5.6 ä m; C2, 23.2 ä m; C3, 51.9 ä m; D_{50} NiON1, 0.5 ä m; N2, 1.0 ä m; N3, 2.4 ä m; N4, 6.2 ä m (See also Ref. [20]).

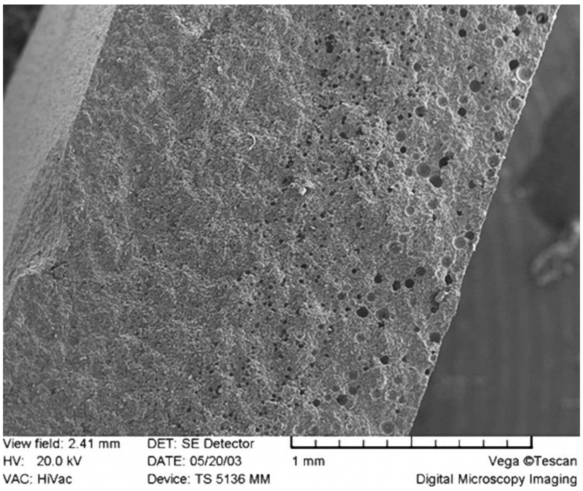

Fig. 5 Anode support consisting of four layers with graded porosity.

large pores which are not considered advantageous for gas transport through the porous anode support, thus further optimisation of the process is required.

3.3
Electrolyte

The solid oxide electrolyte must isolate the two gas atmospheres (fuel and air) from one another and transport the oxygen ions without significant losses from the cathode to the anode. The electrolyte, therefore, needs to be a sufficiently fast oxygen ion conductor and also has to be chemically stable in the large oxygen partial pressure gradient which exposes it simultaneously to highly reducing and highly oxidising conditions.

The transport of oxygen ions in the electrolyte occurs via oxygen vacancies in the oxygen sublattice. The concentration of vacancies and their mobility determines the electrolyte ion-conductivity. Among a number of oxide ion conductors, which includes ceria and La-gallate [3, 21], zirconia stabilised in the conductive cubic phase with up to 10 mol% of either yttria (YSZ) or scandia (SSZ) is the electrolyte material of choice. Fully-stabilized zirconia offers best fast oxygen ion conduction and matching secondary requirements with the other SOFC components such as chemical stability and durability. The conductivity of $Y_{0.15}Zr_{0.85}O_{1.92}$ is >0.1 S/cm at 950 °C–1000 °C and 0.03–0.003 S/cm at temperatures of 600–800 °C (temperature dependence corresponds to an activation enthalpy of 0.8 to 1 eV).

In the absence of suitable alternative materials, minimising the thickness of the YSZ layer to a few micrometers is essential to lower the SOFC operating temperature to around 800 °C. The use of nanopowders seems to be a promising approach for obtaining thin electrolyte films and also for reducing the sintering temperature. For industrial applications, organic solvents are of high interest since they are regarded as more stable and less sensitive to ambient conditions than waterbased solvent systems [22, 23, 24]. The dispersion of nanosized YSZ in organic media was recently investigated [25] in order to achieve high solids loadings in screen-printing slurries. Different long-chained organic acids and amines (dispersants) and various solvents with high boiling points were compared. The effectiveness of the dispersants was evaluated on the basis of rheological measurements, i.e. the lower the viscosity of the slurries, the better the dispersant. Organic acids show a higher dispersive efficiency compared to the amines, and the best results were obtained by using an oligomeric pentastearic acid (Solsperse 3000 (Avecia Additives)). This can be explained by considering the dispersant chain length: Pure steric stabilisation, which is needed in non-polar solvents, is achieved by adsorbed layers thicker than 10 nm, and this condition is fulfilled by the pentamer. The other dispersants, which are semi-steric dispersants, provide a layer of only about 4 nm thickness, which is not sufficient to overcome the van der Waals attraction between the particles. Regarding the solvents, slurries prepared with either kerosene or dodecane showed the lowest viscosities, but the best long-term stability was achieved by using terpineol.

Using these results, dispersions of micron- and submicronsized powders were prepared. Solids loadings as high as 57 vol% and 52 vol%, respectively, were reached (Fig. 6). With these dispersions, electrolyte layers were screen printed on anode substrates and sintered to dense layers on circular disks 12 cm in diameter. An example is shown in Figure 7. All the dispersions showed good compatibility with the substrate. The electrolyte layer could already be sintered at a temperature as low as 1300 °C (Fig. 7a). The sintered layers are between 10 and 20 α̈ m thick and contain no pinholes or cracks (Fig. 7b).

3.4
Cathodes

At the air electrode (cathode), gaseous oxygen is reduced and the resulting oxygen ions are incorporated into the electrolyte (Eq. 2).

$$O_{2,g} + 4e^- + 2V_O^{\cdot\cdot} = 2O_O^x \qquad (2)$$

The cathode environment is highly oxidising, and this fact precludes the

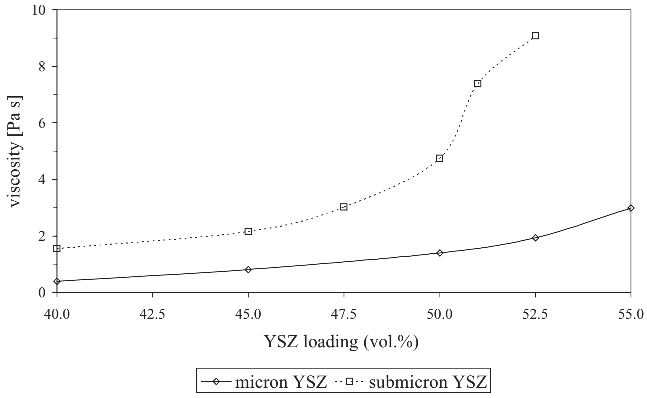

Fig. 6 Viscosity of the terpineol-Solsperse3000-zirconia dispersions as a function of powder size and solids loading.

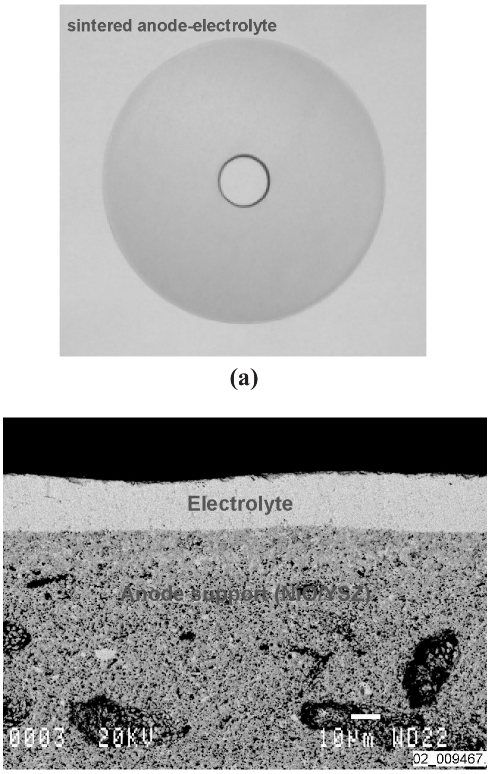

(a)

(b)

Fig. 7 Sintered electrolyte layer on the anode substrate: (a) top view, (b) cross section.

use of metals as electrode materials. The standard SOFC cathode material is an electronically-conductive oxide ceramic based on the perovskite lanthanum manganite in which La is partially replaced by Sr (i.e. $La_{1-x}Sr_xMnO_3$, or LSM). These perovskites are electronic p-type conductors for which the electrical properties are determined by the La/Sr ratio. The formation of $La_2Zr_2O_7$ and SrO is inhibited by incorporating an excess of Mn (1–10%) in the composition and keeping the sintering temperature below 1300 °C [26]. These two compounds impair cathode performance due to solid state reactions.

Significant attention has been paid to the understanding of the electrochemical process in SOFC LSM cathodes [27]. Similar to the Ni-YSZ anode, the electrochemically-active area is assumed to be around the triple-phase boundary where the electronic conducting perovskite, the YSZ electrolyte and the gas phase are in contact with each other. It is generally assumed that oxygen reduction is a multi-step reaction involving adsorption and surface diffusion in a region around the TPB (see Fig. 8).

Early attempts to optimise the LSM cathode activity were concerned with the incorporation of noble metal catalyst particles in order to facilitate −OO bond breaking [28]. A more effective development was the introduction of the composite LSM/YSZ (similar to the Ni-YSZ). Composite cathodes have been fabricated as bi-layers and functionally-graded structures using not

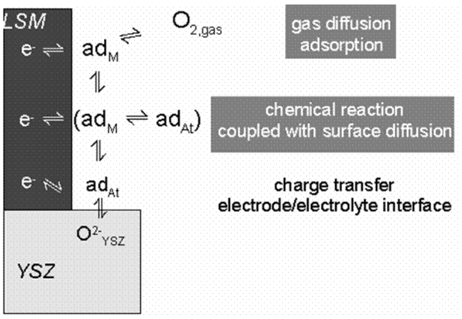
LSM
e⁻ ⇌ ad_M ⇌ O_2,gas
e⁻ ⇌ (ad_M ⇌ ad_At)
e⁻ ⇌ ad_At
O²⁻_YSZ
YSZ
gas diffusion adsorption
chemical reaction coupled with surface diffusion
charge transfer electrode/electrolyte interface

Fig. 8 Possible reaction steps at the air electrode in solid oxide fuel cells.

only conventional wet ceramics processes e.g. slurry deposition [29], but also using advanced deposition techniques such as plasma and thermal spraying [30, 31]. An electrochemical performance corresponding to 0.07–0.12 $\ddot{\alpha}\,\mathrm{cm}^2$ at 850 °C and 0.2–0.4 $\ddot{\alpha}\,\mathrm{cm}^2$ at 750 °C was achieved. Furthermore, a significant decrease in the temperature dependence of the cathodes was realised, with 140 kJ/mol being recorded for air sprayed cathodes and 70 kJ/mol for plasma sprayed cathodes (compared to 200 kJ/mol for pure LSM cathodes).

Despite these advances, there is still a need to further increase the cathodic performance for SOFC operation around 800 °C. Regarding optimisation of the cathode material, there is a fair agreement that a fast oxygen exchange reaction between the lattice oxygen and the gas phase, as well as ionic conductivity in the perovskite, would be advantageous. Thus, further mixed ionic electronic conducting perovskites have been the focus of recent R&D activities.

In the perovskite structure the chemical composition and also the oxygen stoichiometry can be varied to a large extent, and this permits tailoring of the material's electrical properties, especially the ionic conductivity and the oxygen exchange rate. Alternative perovskites based on $La_{1-x}Sr_xFeO_3$ (LSF) and $La_{1-x}Sr_xFe_{1-y}Co_yO_3$ (LSCF) have already been investigated, while the search for new or advanced materials is still going on.

Synthesis techniques that allow a variation of the chemical composition and also the powder morphology in a simple, but controlled, manner are, therefore, of high interest in the area of cathode development. Complex ternary and quaternary perovskite powders can meanwhile be produced in multiple kilogramme quantities by wet chemical processes (co-precipitation), spray processes (flame spraying, spray pyrolysis), and by solid state reaction (mixed oxides or carbonates).

A promising route suitable for such a medium size batch fabrication of complex oxides which has attracted increasing attention over the past 10 years is the spray pyrolysis method [32]. The principle is shown in Figure 9. By spraying a saturated solution of nitrate salts with the desired cation stoichiometry into a hot chamber, the desired perovskites can be prepared either directly, or by a subsequent thermal treatment of the resulting powders. As shown in Figure 10, various powder morphologies ranging from nanocrystallites to sintered microgranules can be obtained depending on the materials, the spraying conditions and the subsequent temperature treatment [33].

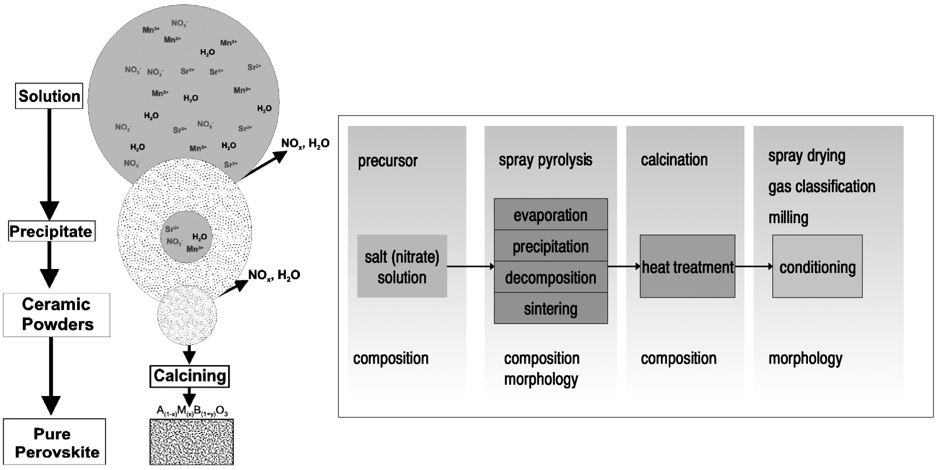
Solution
Precipitate
Ceramic Powders
Pure Perovskite
Calcining
NOx, H2O
NOx, H2O
A(1-x)M(x)B(1+y)O3
precursor
salt (nitrate) solution
composition
spray pyrolysis
evaporation
precipitation
decomposition
sintering
composition morphology
calcination
heat treatment
composition
spray drying
gas classification
milling
conditioning
morphology

Fig. 9 Principle of the spray pyrolysis procedure (left) and possible subsequent processing steps (right).

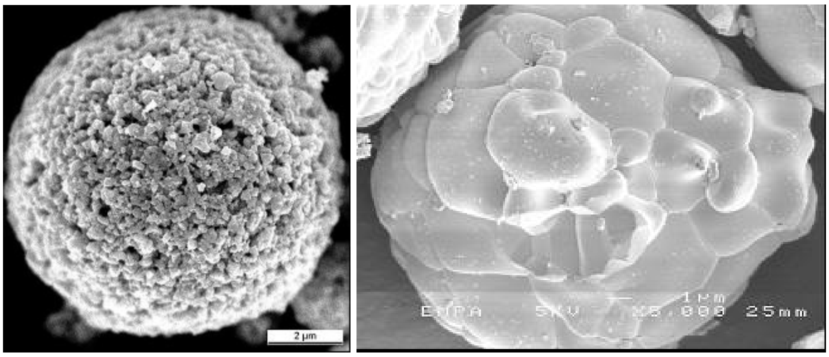
2 µm
EMPA 5kV X5,000 25mm
1µm

Fig. 10 Examples of perovskite granulates prepared by spray pyrolysis: (a) pyrolysed and spray dried particle; b) spray dried and sintered granule.

Tailoring the powder morphology in order to obtain a long triple-phase boundary line is an important focus for further development in the fabrication of high performance SOFC cathodes. For wet chemical processing techniques like screen printing, the use of submicron powders can serve to impart a fine, porous electrode structure. Figure 11 shows the micro-structure of a LSF-cathode prepared from spray-pyrolysed submicron powders. The performance test under technically relevant conditions indicated a significant improvement of the cell performance that was attributed to the use of nanostructured perovskite powders for the cathode [34, 35].

Although these results are encouraging, the fabrication of nanoparticles using rather simple fabrication routes carries with it the risk that the individual nano-sized grains will be of inhomogeneous chemical composition. While the average composition can be well controlled, the stoichiometry may vary on a microscopic

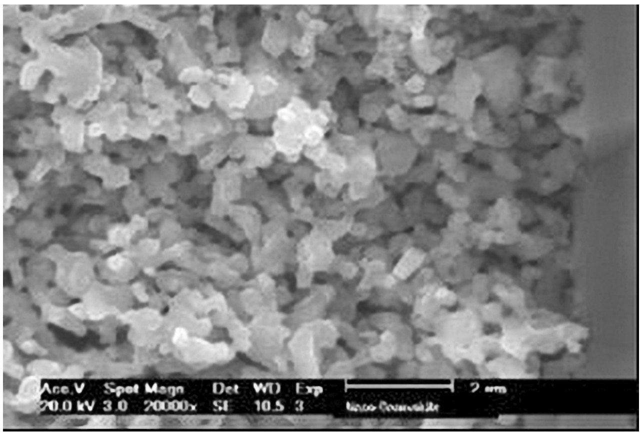

Fig. 11 SEM micrograph of an nanostructured cathode, as applied in HT ceramix SOFC cells.

level. Such inhomogeneities have indeed been observed by TEM-EDX in both spray pyrolysed and co-precipitated LSF and LSM powders [35]. It is likely that insufficient control of the seed crystal formation process is the cause. Subsequent grain growth during calcination and sintering steps are expected to homogenise the stoichiometry, however, the trend to lower heat treatment temperatures mentioned earlier means that the precise control of the nanopowder composition is becoming even more important. Possible alternative routes to produce powders with homogenous chemical compositions are the flame spray processes and combustion synthesis, where the latter is currently limited to laboratory-scale amounts of powders.

Although LSF shows promise as a SOFC cathode material, the chemical compatibility with the YSZ electrolyte has not been fully explored. Initial results have provided evidence for cation interdiffusion between YSZ and LSF, which needs to be further analysed and evaluated [36].

A well-known and highly-active cathode material is $La_{1-x}Sr_xCo_{1-y}Fe_yO_3$ (LSCF). This material is chemically incompatible with the YSZ electrolyte, with solid state reactions ocurring which can seriously deteriorate cathode performance. In order to use this material in a solid oxide fuel cell, a more complex cell arrangement and modifications to the cell fabrication process are required. An advanced cell concept for using LSCF as cathode is presented below.

4
Advanced SOFC Cell Concepts

The multilayer structure of the SOFC cell and the operating conditions impose strict requirements on the materials and the compatibility between the individual cell layers. Although there is significant research and development going on to find better and more suitable materials, not all the requirements for a given cell layer can currently be fulfilled by one single material. Engineering solutions under investigation to date include the use of composite materials such as cermets, or the use of additional layers (for instance in a bi-layer electrolyte) in order to

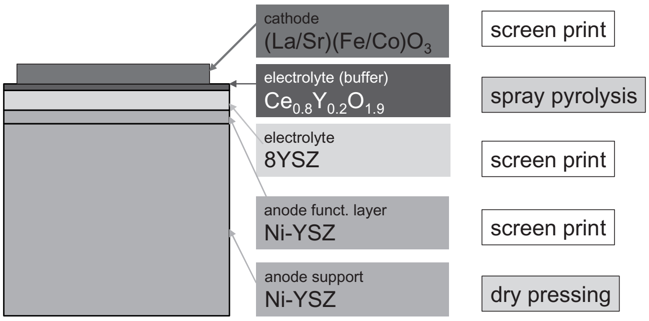

Fig. 12 Concept for advanced anode-supported cells for operation in steam-reformed natural gas at around 800 °C.

avoid direct contact between incompatible materials while maintaining the functionality of the respective layers.

An advanced concept using LSCF as a cathode material in anode supported SOFC cells by using low cost and upscalable fabrication processes has recently been presented [37]. The cell concept, presented here in Figure 12, features a Ni-YSZ anode on a anode support, and a bi-layer electrolyte layer consisting of a 20–30α̈ m thick 8-YSZ (8 mol% yttria stabilised zirconia) layer followed by a submicron yttria-modified ceria layer. The ceria acts as a so-called ‘‘barrier layer’’ to avoid contact, and thus chemical reaction, between the 8-YSZ electrolyte and the LSCF cathode.

Two alternative low-cost fabrication techniques were used to fabricate circular cells with an active cell area of 100 cm^2 of this advanced cell concept. Die pressing (see Sec. 3.2) was used to fabricate the Ni-YSZ cermet support, while ceria barrier layer was deposited using the spray pyrolysis technique (see Sec. 3.4). The technique involves spraying nitrate solutions onto a heated substrate surface where the solutions dry to salts and decompose directly to the desired ceramic phase without any subsequent firing or sintering step [38]. Dense layers with a thickness under one micron can be prepared [39]. In the advanced cell concept shown here, spray pyrolysis deposition was used to deposit thin (250 nm) barrier layers of yttria-modified ceria (CYO) on top of the screen-printed 8-YSZ electrolytes on ring-shaped cells with an inner and outer diameter of 2 and 12 cm, respectively (100 cm^2 active cell area, see also Fig. 7). The resulting microstructure of the cell is shown in Figure 13.

Such advanced cells have been assembled in a three-cell short stack in the Sulzer HEXIS design and tested in steam-reformed methane (a mixture of mainly hydrogen and carbon monoxide). The stack performance, and the performance of each individual cell, is shown in Figure 14. The open-circuit voltage is around 980 mV. Due to the absence of outer rim sealing in this stack design and the steam content in the fuel mixture, the unloaded cell voltage value confirms adequate electrolyte specifications. At 800 °C with 200 g/h cell air flow and 4 g/h cell fuel flow, a power output of 32 W/cell (320 mW/cm^2) with an electrical efficiency of 61% and a fuel utilization of 90% was achieved. The highest power output (43 W/cell, 430 mW/cm^2) resulted in an electrical efficiency of 32% and a fuel utilization of 45%. This correlates

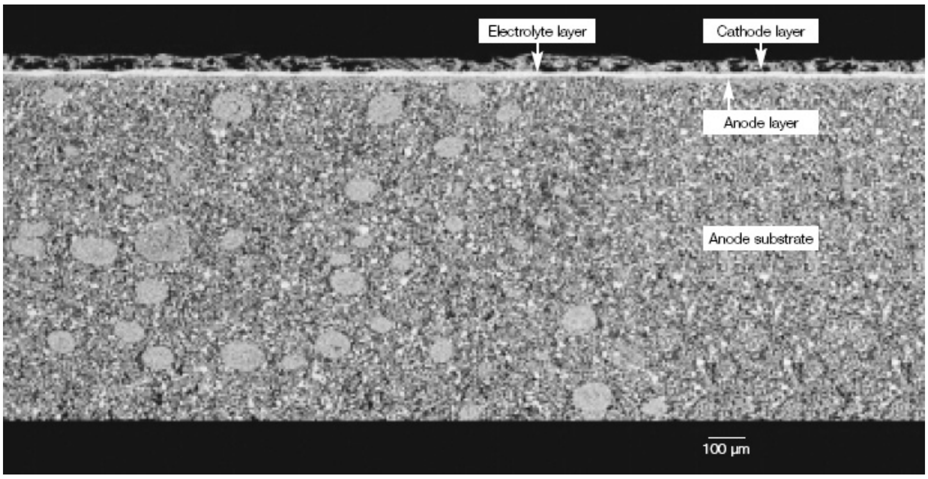

Fig. 13 Cross-section of an anode-supported cell (Ni-bidisperse YSZ)//YSZ/CYO//LSCF) after 330 h testing in steam-reformed natural gas (steam-to-carbon ratio = 2).

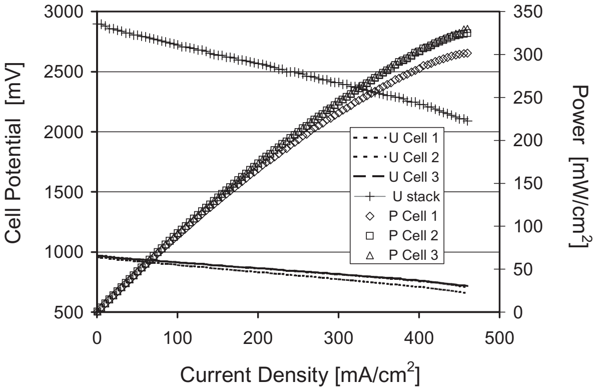

Fig. 14 Performance of a 3-cell short stack and the individual cells (Ni-bidisperse YSZ)//YSZ/CYO//LSCF) at 800 °C, natural gas flow per cell 4 g/h, air flow per cell 200 g/h.

well with the performance of the standard, electrolyte-supported Sulzer Hexis cells, which, however, are usually operated at much higher temperatures (920 °C). Therefore, a significant decrease in the operating temperature to 800 °C without losses in the electrical power output compared to the standard product has been demonstrated in these advanced anode-supported cells. A remaining challenge for this cell concept, and the techniques used for its fabrication, is the reproducibility of the cell fabrication process on a technical scale. Especially the processing of the Ni-YSZ supports needs to be further optimised.

As indicated in Section 3.1., the major functions of the anode support are current pick-up and gas transport to the active electrode layer. While the Ni-YSZ cermet was originally developed for operating temperatures between 900–1000 °C, the reduction of the operating temperature to 800 °C now allows the use of other materials such as ferritic steels, which are already used for the interconnect plate. However, conventional wet ceramic processes, especially the associated sintering steps in air at temperatures of 1300–1500 °C, still rule out the use of such steels in the cell assembly. In order to efficiently make use of metallic supports, alternative fabrication processes need to be applied.

Possible alternative cell fabrication processes for supported-solid oxide fuel cells are plasma spraying and thermal spraying [30, 31]. The plasma spray process for SOFC cells is currently being developed further by the German Aerospace Center (DLR) for both mobile (automotive) and stationary applications [40, 41]. In the plasma-spray process, the anode, cathode and electrolyte raw materials (NiO-YSZ, (La,Sr)MnO_3, and YSZ or SSZ, respectively) in the form of powders with a particle size of approximately 10–25α̈ m are injected into a high temperature plasma jet and deposited in order as the desired ceramic layers. In contrast to wet powder processing techniques, plasma spraying does not involve additional thermal treatments (e.g. a sintering step). Further advantages of the process are the high mechanical strength of the cells and the resulting thermal cycling capability, as well as the high porosity of the anode substrate, which ensures a uniform fuel supply to the electrode. For plasma sprayed cells on steel supports, power densities of 0.6 W/cm^2 in hydrogen at 900 °C have been reported [42]. At reduced temperatures of 800 °C, a maximum power of 0.23–0.35 W/cm^2 has been achieved using hydrogen as fuel [43].

5
Outlook

Even though a set of materials seems to be established for solid oxide fuel cells, the materials and processing development can by no means be considered as complete. Durability and cost reduction are major problems which still need to be solved before successful commercialisation of SOFC technology can be realised.

Presently there is a continuous trend to lower the operating temperature of the SOFC cells, and this requires that new materials and/or cell concepts need to be developed. The situation remains that no single material is known that can fulfil all requirements for a given individual cell component. Therefore, materials engineering solutions and the development of new fabrication techniques are expected to play a key role in future developments. While lower operating temperatures may allow the application of even more complex cell concepts, highly integrated processes such as the spray pyrolysis method for powder production/layer deposition, as well as thermal (plasma) spraying, also have a significant potential to further reduce fabrication costs. The various application-specific demands for small, larger-scale stationary, and especially automotive systems can be regarded as a strong and helpful "application pull" for future progress in SOFC technology.

References

[1] A. Baur, B. Preis, *Ztschr. Elektrochem.* **1937**, *43*, 727.
[2] H. H. Möbius, *J. Solid State Electrochem.* **1997**, *1*, 2.
[3] N. Q. Minh, *J. Am Ceram. Soc.* **1993**, *76*, 563.

[4] A. Hammou, J. Guindet, in: *The CRC Handbook of Solid State Electrochemistry*, CRC Press Inc., Boca Raton, FL 1997

[5] B. Pietzak, H. Raak, *Proc. The Fuel Cell World (2004)*, European Fuel Cell Forum, Oberrohrbach, CH, 2004, p. 227.

[6] H. Fujii, *Solid Oxide Fuel Cells, (SOFC VIII)*, The Electrochemical Soc. Proc Volume PV2003-07, The Electrochemical Soc. Inc, Pennington, NJ, 2003, 9.

[7] P. Holtappels, U. Stimming in: *Handbook of Fuel Cells – Fundamentals, Technology and Applications*, Vol. 1, J. Wiley&Sons, Ltd. Chichester 2003, p. 335.

[8] *High Temperature Solid Oxide Fuel Cells Fundamentals, Design and Applications*, (Eds.: S. C. Singhal, K. Kendall) Elsevier Advanced Technology, Oxford, UK 2003.

[9] H. P. Buchkremer, U. Diekmann, L. G. J. de Haart, H. Kabs, U. Stimming, D. Stöver, *Solid Oxide Fuel Cells, (SOFC V)*, The Electrochemical Soc. Proc Volume PV97-40, The Electrochemical Soc. Inc, Pennington, NJ, 1997, 160.

[10] L. G. J. de Haart, I. C. Vinke, A. Janke, H. Ringel, F. Tietz, *Solid Oxide Fuel Cells (SOFC VII)*, The Electrochemical Soc. Proc Volume 2001-16, The Electrochemical Soc. Inc, Pennington, NJ, 2001, 111.

[11] S. Linderoth, *Proc.* 4th *European Solid Oxide Fuel Cell Forum*, Luzern 2000, European Fuel Cell Forum, Oberrohrbach, CH, 2000, p. 19.

[12] P. Holtappels, I. C. Vinke, L. G. J. de Haart, U. Stimming, *J. Electrochem. Soc.* **1999**, *146*, 2976.

[13] M. Brown, S. Primdahl, M. Mogensen, *J. Electrochem. Soc.* **1999**, 146.

[14] S. W. Tao, J. T. S. Irvine, *Chemical Record* **2004**, *4*, 83.

[15] M. Mogensen, *J. Electroceram.* **2000**, *5*, 141.

[16] H. Kim, S. Park, J. M. Vohs, R. Gorte, *J. Electrochem. Soc.* **2001**, *147*, A693.

[17] E. P. Murray, T. Tsai, S. A. Barnet, *Nature* **1999**, 400.

[18] M. C. Verbraeken, B. A. Boukamp, D. H. A Blank, P. Holtappels, U. Vogt, *Proc* 9th *International Symposium on Solid Oxide Fuel Cells*, submitted.

[19] B. Gut, *SOFC-VI*, The Electrochemical Society Proc. Vol. PV 99-19, The Electrochemical Society Inc. Pennington, NJ, 1999, p. 840.

[20] P. Holtappels, C. Sorof, M. C. Verbraeken, S. Rambert, U. Vogt, *Proc. 6th European SOFC Forum Lucerne, 2004*, Vol. 1, The European Fuel Cell Forum Oberrohrdorf, CH. 2004, p. 201.

[21] A. I. Ishihara, M. Honda, T. Shibayama, H. Minami, H. Nishiguchi, Y. Takita, *J. Electrochem. Soc.* **1998**, *145*, 3177. B) A. Matrazek, D. Kobertz, L. Singheiser, K. Hilpert, *SOFC VII*. The Electrochemical Soc. Proc. Vol. 2001-16, 2001, p. 319.

[22] T. J. Graule, L. J. Gauckler, *Third Euro-Ceramics 1*, **1993**, 491.

[23] P. Hidber, T. J. Graule, L. J. Gauckler, *J. Eur. Ceram. Soc.* **1997**, *17*, 239.

[24] P. Hidber, T. J. Graule, L. J. Gauckler, *J. Am. Ceram. Soc.* **1996**, *79*, 1857.

[25] S. Zürcher, T. Graule, *J. European Ceram. Soc.* **2005**, *25*, 863.

[26] G. Stochniol, E. Syskakis, A. Naoumidis, *J. Am. Ceram. Soc.* **1995**, *78*, 929.

[27] M. J. Joergensen, M. Mogensen, *J. Electrochem. Soc.* **2001**, *148*, A433.

[28] J. W. Erning, Th. Hauber, U. Stimming, K. Wippermann, *J. Power Sources* **1996**, *61*, 205.

[29] P. Holtappels, C. Bagger, *J. Eur. Ceram. Soc.*, **2002**, *22*, 41.

[30] K. Barthel, S. Rambert, S, Siegmann, *J. Therm. Spray Techn.*, **2000**, *9*, 343.

[31] S. Rambert, A. J. McEvoy, K. Barthel, *J. Europ. Ceram. Soc.*, **1999**, *19*, 921.

[32] G. L. Messing, S.-C. Zhang, G. V. Jayanthi, *J. Am. Ceram. Soc.* **1993**, *76*, 2707.

[33] P. Holtappels, U. Vogt, H. Schindler, B. Gut, *Proc. 5th European Solid Oxide Fuel Cell Forum*, Vol. 1, Ed.: J. Huijsmans, Editor, European Fuel Cell Forum, Oberrohrdorf, Switzerland, 2002, p. 103.

[34] M. Molinelli, D. Larrain, R. Ihringer, L. Constantin, N. Audissier, O. Bucheli, D. Favrat, J. Van herle, *SOFC VIII*, The Electrochemical Society Proc. Vol. PV 2003-07, The Electrochemical Society Inc. Pennington, NJ 2003, p. 905.

[35] J. Sfeir,* S. Vaucher, P. Holtappels, U. Vogt, H.-J. Schindler, J. Van herle, E. Suvorova, P. Buffat, D. Perret, N. Xanthopoulos, O. Bucheli, *J. Electroceram.* **2005**, In Press.

[36] S. P. Simner, J. P. Shelton, M. D. Anderson, J. W Stevenson, *Solid State Ionics* **2003**, *161*, 11. a). M. D. Anderson, J. W Stevenson, S. P. Simner, *J. Power Sources* **2004**, *129*, 188.

[37] P. Holtappels, T. Graule, B. Gut, et al. SOFC VIII, The Electrochemical Society

Proc. Vol. PV 2003-07, The Electrochemical Society Inc. Pennington, NJ 2003, p. 1003.

[38] D. Perednis, K. Honegger, L. J. Gauckler, in *Proc. 4th European Solid Oxide Fuel Cell Forum*, Vol. 2, Ed.: A. J. McEvoy, European Fuel Cell Forum Oberrohrdorf, Switzerland, 2000, p. 819.

[39] D. Perednis, L. J. Gauckler, in *Proc. 5th European Solid Oxide Fuel Cell Forum*, Vol. 1, Ed.: J. Huijsmans, European Fuel Cell Forum, Oberrohrdorf, Switzerland 2002. p. 72.

[40] J. Zizelman, C. DeMinco, S. Mukerjee, J. Tachtler, J. Kammerer, P. Lamp, *Proc. 5th European Solid Oxide Fuel Cell Forum*, Vol. 2, Ed.: J. Huijsmans, Editor, European Fuel Cell Forum, Oberrohrdorf, Switzerland 2002, p. 1153.

[41] M. Lang, R. Henne, S. Schaper, G. Schiller, *J. of Thermal Spray Techn.* **2001**, Vol. 10 No. 4 618.

[42] G. Schiller, T. Franco, R. Henne, M. Lang, R. Ruckdäschl, *Solid Oxide Fuel Cells (SOFC VII)*, The Electrochemical Soc. Proc Volume 2001-16, The Electrochemical Soc. Inc, Pennington, NJ, 2001, 885.

[43] M. Lang, T. Franco, R. Henne, G. Schiller, S. Ziehm, *Extended Abstracts 2003 Fuel Cell Seminar, Miami Beach, Florida, USA, 3.-7.11.2003*, 2003, p. 794.

Temperature Measurement

Joachim Fischer

1
Definition of Temperature – An Introduction

The definition of temperature was first suggested by William Thomson, later named Lord Kelvin, and it is based on the amounts of heat entering and leaving an ideal heat engine:

$$\frac{Q_h}{T_h} = \frac{Q_c}{T_c} \tag{1}$$

where Q_h is the heat flowing into the engine from a hot reservoir at a temperature T_h, and Q_c is the heat flowing out of the engine to a cold reservoir at a temperature T_c. Thomson summarized that heat provides the basis of almost all power generation, in which the efficiency increases with increase in the temperature of the hot reservoir. Clausius observed that one consequence of Thomson's temperature definition is that the sum of all of the heats flowing out of a heat engine multiplied by $1/T$ is identically zero for reversible processes and greater than zero for irreversible processes, that is,

$$\sum_i \frac{Q_i}{T_i} \geq 0 \tag{2}$$

Clausius called the new quantity, $S = Q/T$, *entropy*, being a sort of transformed energy (from the Greek word *trope*, which means transformation).

The temperature defined by Thomson was based on the conceptual device of the ideal heat engine; however, such a device is not necessary. The mathematician Carathéodory showed that in any system there exists a unique state variable that characterizes the reversibility of a process, and this variable is proportional to the integral of the heats associated with the process divided by a unique "integrating factor" (Chandrasekhar, 1967). Comparison of Carathéodory's result with Eq. (2) shows that the state variable is the entropy defined by Clausius and the integrating factor is the temperature defined by Thomson. Carathéodory's result leads to the definition of temperature:

$$\frac{1}{T} = \frac{dS}{dU} \tag{3}$$

where U is the internal energy of the system and S is the conventional entropy.

Boltzmann subsequently showed that the entropy of a system is related to the number of ways the constituent atoms and molecules can be arranged into the

digital Encyclopedia of Applied Physics.
Copyright © Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim All rights reserved.

observed macroscopic state:

$$S = k \ln P \tag{4}$$

where k is a constant and P is the probability of the system being in the observed state. In his original work, Boltzmann deduced the relation (4) without the factor k. Naturally, this leads to a temperature measured in energy units (joule); hence, in principle we do not need a separate base unit for temperature, the kelvin. However, such a temperature scale would have impracticably small ($\sim 10^{-20}$) and unfamiliar values. For this reason, Planck introduced a constant k, later named after Boltzmann, to provide the link to conventional definitions of entropy and temperature (Planck, 1921).

Equation (3) applied to various idealized systems yields thermodynamic relations that can be used to measure temperature. For example, for an ideal gas, we can derive the equation of state:

$$pV_m = N_A kT \tag{5}$$

where p is the pressure, V_m is the molar volume, and N_A is the Avogadro constant, the number of particles per mole.

The quantity kT, which occurs in the equations of state, is a characteristic energy determining the energy distribution among the particles of the system when it is in thermal equilibrium. Thus for unbound atoms, temperature is proportional to the mean translational kinetic energy. Temperature is linear and rational: equal intervals or ratios of temperature correspond to equal differences or ratios of mean kinetic energy, and a single definition is required to fix the magnitude of the temperature unit. All other temperature values must then be determined experimentlly, using a suitable thermal system and equation of state.

Today, the kelvin is defined in terms of the temperature of the triple point of water (TPW), and the Boltzmann constant k is a measured quantity. However, there is a current proposal of the Comité International des Poids et Mesures (CIPM) to define a numerical value for k, from which it follows that all temperatures, including the TPW, must be measured. Of course, the adopted value for k will be such that the temperature values will, as far as possible, remain unchanged. The determination of the Boltzmann constant applying primary thermometers is discussed in Section 13.

The ideal gas as introduced with Eq. (5) is as fictitious as the ideal heat engine. However, using a noble gas at a very low pressure, a thermometer can be constructed that really measures the so-called thermodynamic temperatures. In fact, the gas thermometer was the first thermometer to derive thermodynamic temperatures. Such kinds of thermometers, whose basic relation between the measurand and T can be written down explicitly without unknown, temperature-dependent constants being introduced, are generally called *primary thermometers* (see Sections 2–6 for further examples). Any accurate realization of the thermodynamic temperature is very time consuming and requires extreme metrological effort. A gas thermometer is appropriate only as a primary standard in fundamental laboratory measurements. On this ground and to harmonize the differing national temperature scales for the day-to-day practical use, the so-called International Temperature Scales have been developed by the

Consultative Committee of Thermometry (CCT) and adopted by the General Conference of Weights and Measures (CGPM). The International Temperature Scales reflect the most recent state of metrological accuracy and therefore they are replaced by new versions from time to time (see Sections 7 and 8). Instruments and methods to disseminate the scales to the users in science and industry are discussed in Sections 9–12.

With a few exceptions, thermodynamic temperatures are determined only to establish internationally agreed temperature scales. Only at very low temperatures, thermometers are calibrated with thermodynamic methods of measurement. The following primary thermometric methods are described in some detail: As an introduction, constant-volume gas thermometry (CVGT) is discussed in Section 2. More recent gas thermometric work is performed at Physikalisch-Technische Bundesanstalt (PTB), Germany, employing dielectric-constant gas thermometry (DCGT, Section 2.1) and at the French and Italian metrology institutes LNE-INM/CNAM, INRIM, the National Institute of Standards and Technology (NIST), USA, and the University College London (UCL), UK, using acoustic gas thermometry (AGT, Section 2.2). Noise thermometry work and Doppler-broadening techniques are discussed in Sections 3 and 4. The total and spectral radiation thermometry are described in detail in Sections 5 and 6.

2
Primary Gas Thermometry

The classical gas thermometer is based on the equation of state for an ideal gas (5) and is reviewed here to introduce the principle of a primary thermometer. Gases have been used for thermodynamic thermometry since its inception, and we shall see that they are still preeminent for much of the range, although the techniques have changed. In classical gas thermometry, the virial equation of state for a gas is used to relate pressure p and molar density $\rho (= n/V)$, of the gas to T (Figure 1).

Although many gases exhibit a nearly ideal behavior at low temperatures, in view of the desired level of accuracy the small departures from the ideal behavior must be carefully considered.

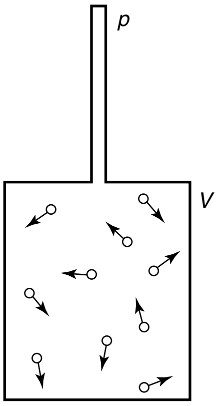

Fig. 1 Principle of constant-volume gas thermometry; pressure p and density n/V are the measurands.

This is done by measuring the relevant property that is dependent on the density. Then, the ideal behavior is deduced by an extrapolation to zero density applying an appropriate virial expansion. At sufficiently low pressures and densities, the behavior of a real gas can be described by the virial expansion, where $B(T)$ and $C(T)$ are the second and third density virial coefficients, and R is the molar gas constant:

$$pV = nRT\left(1 + \frac{B(T)}{V_m} + \frac{C(T)}{V_m^2} + \cdots\right) \tag{6}$$

For practical reasons, in primary gas thermometry, the volume is kept constant and therefore the method is called *constant volume gas thermometry* (*CVGT*). In the method of absolute pV-isotherm CVGT, the gas bulb at a constant but unknown temperature T is filled with a series of increasing amounts of gas to obtain a series of pressures p. pV/nR may be plotted as a function of $1/V_m$ according to Eq. (6). The intercept of the resulting isotherm is the temperature T. For this method, it is not necessary to know the virial coefficients because the extrapolation to zero pressure is made by fitting a virial expansion to the experimental data. There are different ways of measuring the amount of gas. One method is to weigh a bulb of known volume with and without the gas sample and deduce the amount from the weight difference. This method has not been used in recent high-precision CVGT owing to the practical difficulties inherent in the weighing of low-density gases. It was used in historical measurements at the TPW to determine the molar gas constant R. In view of the careful analysis of these measurements in Colclough (1984) and also considering systematic effects related to gas sorption, to the measurement of the gas-bulb volume, and to the pressure dilatation of the bulb, at present it does not seem possible to reach a level of relative uncertainty below about 10 parts in 10^6 (10 ppm).

Relative pV-isotherm CVGT is performed by measuring the pressures p_r and p of the same gas sample in the thermometer bulb at a known thermodynamic reference temperature T_r and the unknown temperature T. If the bulb volume V is assumed to be independent of temperature and pressure, in the case of an ideal gas, T is obtained from the following equation:

$$T = \frac{T_r p}{p_r} \tag{7}$$

However, for a real gas, the density virial coefficients have to be taken into account again. In addition, the volume of the gas bulb changes with the temperature due to the thermal expansion of the bulb material. The thermal expansion must therefore be known as accurately as possible, as it is one of the main sources of error in gas thermometry. A change in the gas pressure also causes the bulb to dilate, and this change in volume must either be calculated or the gas bulb must be surrounded by a second bulb filled with gas at the same pressure as the gas in the first bulb. In this case, the dilation of the bulb due to pressure is negligible. CVGT has been employed to establish the thermodynamic basis of the International Temperature Scale of 1990 (ITS-90) from very low temperatures up to 660 °C and is used to interpolate the scale between 3 and 25 K; for details see Fischer and Fellmuth (2005a).

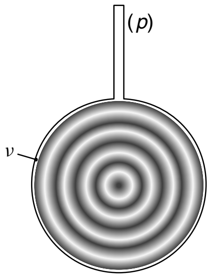

Fig. 2 Principle of acoustic gas thermometry: from the resonance frequency ν and the dimension of the resonator the speed of sound u is deduced.

2.1
Acoustic Gas Thermometers

The utility of measuring the speed of sound in gases as a primary means of measuring temperature has long been appreciated. The parameters are all intensive; hence in principle, quantities such as volumes and the amounts of gas contained in them need not be measured. Moreover, pressure need not be measured to the first order of accuracy, being only needed for evaluating the nonidealities.

For an ideal gas, the relation between the speed of sound u_0 and the thermodynamic temperature T is given by the following equation ($\gamma = C_p/C_V$ is the adiabatic exponent and M is the molar mass of the gas):

$$u_0 = \left(\gamma \frac{RT}{M}\right)^{1/2} \tag{8}$$

The pressure dependence of the speed of sound in a real gas can be expressed by a virial expansion:

$$u^2 = u_0{}^2\left(1 + \alpha p + \beta p^2 + \cdots\right) \tag{9}$$

that is, the influence of the pressure is of second order. The acoustic virial coefficients α and β can be expressed in terms of the density virial coefficients $B(T)$ and $C(T)$. As can be seen from Eq. (8), the gas constant R and the molar mass M directly influence the result. Thus, the uncertainty of T depends on their uncertainties. The influence of R and M may be eliminated if the speed of sound is determined at the temperature of the TPW with the same thermometer and the same gas, that is, if relative isotherms are measured. Applying Eq. (9), u_0 is determined by extrapolation to zero density and T is calculated using Eq. (8).

The main challenge of the method is the measurement of u with the necessary uncertainty, and in spite of significant advances this is still the limitation. In earlier works, a fixed-frequency, variable-path, cylindrical acoustic interferometer was used (Quinn, 1990; Pavese and Molinar, 1992). Such a system was used for the measurements of T in the temperature range from 4.2 to 20 K, with an uncertainty of 1–4 mK, and with a similar instrument the gas constant R was determined (Colclough, Quinn and Chandler, 1979).

Nowadays, variable-frequency, fixed-path spherical resonators are preferred (Figure 2). Their figure of merit is about an order of magnitude higher than that of cylindrical resonators. Furthermore, boundary layer effects are smaller and the problems due to the excitation of different modes are fewer. Moldover *et al.* (1988) assembled a spherical resonator and used it in 1986 to redetermine the gas constant R with an estimated relative uncertainty of 1.7 ppm, a factor of 5 smaller than the uncertainty of the best previous measurement. They measured the acoustic resonance frequencies of the argon-filled resonator and the microwave resonance frequencies of the same cavity when evacuated. The microwave data were used to deduce the thermal expansion of the cavity. More recent measurements lead to new values of $T - T_{90}$ (T_{90} temperature according to the ITS-90) between 217 and 303 K (Moldover *et al.*, 1999). Ewing and Trusler (2000) performed thermodynamic measurements in the range from 90 to 300 K. The estimated standard uncertainty ranges from 0.9 to 1.3 mK.

Later, a new spherical resonator was developed at NIST, including new acoustic transducers to cover the temperature range from 273 to 800 K. Results have been obtained for the melting point of gallium and the freezing points of indium and tin (Strouse *et al.*, 2003) and in the second experimental run between 271 and 552 K (Ripple, Strouse and Moldover, 2007). Independent acoustic measurements of the thermodynamic temperature have also been performed from 234 to 380 K by Benedetto *et al.* (2004) and from 7 to 24.5 K and from 90 to 273 K by Pitre, Moldover and Tew (2006) using spherical resonators. All results agree within the remarkably small combined uncertainty.

2.2
Dielectric-Constant Gas Thermometers

DCGT is based upon the variation with the temperature of an intensive property of the gas, the relative dielectric constant (permittivity) ε_r. The gas particles act in a capacitor as induced dipoles (Figure 3) with the static electric dipole polarizability α_0 according to the Clausius–Mosotti equation:

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{N}{V}\frac{\alpha_0}{3\varepsilon_0} \qquad (10)$$

where ε_0 is the exactly known electric constant and N/V is the number density. The basic idea of DCGT is to replace the density in the state equation of a gas by the dielectric constant. Thus, the state equation of an ideal gas (5) can be expressed in the following form:

$$p = \frac{kT(\varepsilon - \varepsilon_0)}{\alpha_0} \qquad (11)$$

with $\varepsilon_0\varepsilon_r = \varepsilon$ and the approximation $\varepsilon_r + 2 \approx 3$ valid for ideal gases. Therefore, DCGT avoids the density determination of the conventional gas thermometry that is complicated by the gas contained in dead spaces not at the

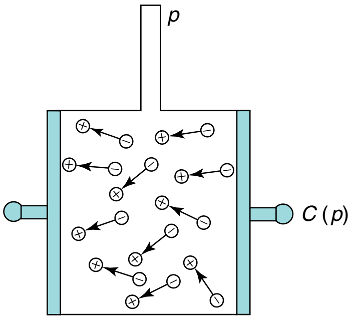

Fig. 3 Principle of dielectric-constant gas thermometry: the relative dielectric constant ε_r is determined by the change of the capacitance $C(p)$.

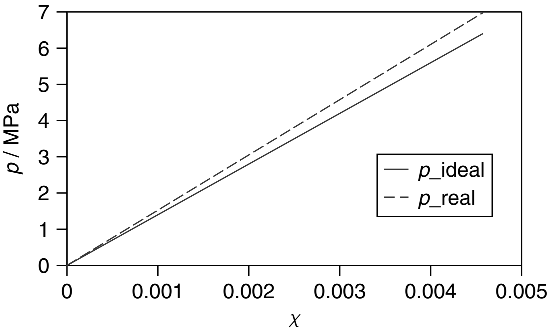

Fig. 4 Pressure versus dielectric susceptibility isotherms of helium at the TPW with (dashed curve *p*_real) and without (straight, full curve *p*_ideal) consideration of the interaction between the atoms.

measuring temperature T and by gas adsorption in the system. Other advantages are that the pressure-sensing tubes can be of any convenient size and that the thermometric gas can be moved in or out of the thermometer cell without the need to measure the amount of the gas involved.

DCGT can be used for primary thermometry in two ways. Absolute measurements require that the polarizability α_0 be known, with the required accuracy. Nowadays, this condition is fulfilled for helium, which became a model substance for evaluating the accuracy of ab initio calculations of thermophysical properties. Recent progress has decreased the uncertainty of the ab initio value of α_0 well below one part in 10^6 (Lach, Jeziorski and Szalewicz, 2004). However, one disadvantage of helium is its relatively small polarizability, which is, for instance, smaller than that of argon by a factor of 8. On the other hand, primary thermometry does not require a value of α_0 if measurements are made both at the triple point of water (T_{TPW}) and at the measuring temperature T. The ratio of the two temperatures is given by

$$\frac{T}{T_{\mathrm{TPW}}} = \left(\frac{p}{p_{\mathrm{TPW}}}\right) \frac{(\varepsilon_{\mathrm{TPW}} - \varepsilon_0)}{(\varepsilon(T) - \varepsilon_0)} \tag{12}$$

where the quantities measured at the TPW have the corresponding index.

By measuring the relative change in capacitance $(C(p) - C(0))/C(0) = \chi + (\varepsilon/\varepsilon_0)\kappa_{\mathrm{eff}}p$ of the capacitor as a function of gas pressure p, the slope $\varepsilon_0 kT/\alpha_0$ of the isotherms (11) is determined. κ_{eff} takes into account the effective compressibility of a suitable capacitor used to measure the susceptibility $\chi = \varepsilon_r - 1$. A polynomial fit to the resulting p versus χ data points (Figure 4), together with the knowledge of the dependence of the dimensions of the capacitor on $p(\kappa_{\mathrm{eff}})$, yields the temperature. Because of very small susceptibility (for instance, 0.003 at 4 K and 0.05 MPa for helium), this technique is extremely demanding for measurement of capacitance changes. For a real gas, the interaction between the particles has to be considered by combining the virial expansions of the state equation (5) and the Clausius–Mosotti equation (10) (see the dashed curve in Figure 4).

Recently, with this method, thermodynamic temperatures have been determined between 4 and 26 K (Gaiser, Fellmuth and Haft, 2007). Considering the progress in the measurement of pressure and capacitance changes, it is seen that DCGT has potential for both decreased uncertainty and increased application range. Measurements up to the TPW would be desirable because the density is measured in situ and hence the thermal expansion of the thermometer bulb does not cause problems as in CVGT. Furthermore, as an inverse application of DCGT, measurements at the TPW allow the determination of the Boltzmann constant k. The system described in Luther, Grohmann and Fellmuth (1996) would allow to achieve a standard uncertainty of k of about 15 parts in 10^6 and presently at PTB a system is under construction, which should yield an uncertainty that is an order of magnitude smaller (Gaiser, Fellmuth and Haft, 2007). Another inverse application of DCGT is the realization of a pressure standard near 1 MPa applying a toroidal cross capacitor filled with helium (Moldover, 1998). Since the interaction between the helium atoms at the realized pressure has to be considered, this project requires values calculated ab initio for both the polarizability α_0 and the second density virial coefficient.

3
Noise Thermometry

The noise thermometer is based on the temperature dependence of the mean square noise voltage, $\langle U^2 \rangle$, developed in a resistor. Nyquist derived Eq. (13) from thermodynamic calculations (Nyquist, 1928)

$$\langle U^2 \rangle = 4kTR\Delta f \tag{13}$$

valid for frequencies $f \ll kT/h$, where R is a frequency-independent resistance, Δf the bandwidth, and h Planck's constant. From the statistical nature of the measured quantity, long measuring times arise, which may be estimated from Eq. (14):

$$\frac{\Delta T}{T} \approx \frac{2.5}{\sqrt{t\Delta f}} \tag{14}$$

with t being the measuring time. One of the main problems is the accurate measurement of the very small voltages developed avoiding extraneous sources of noise and maintaining constant bandwidth and gain of the amplifiers. For details see Quinn (1990); White, Mason and Saunders (2002). In the past, in the high-temperature range, the uncertainties of noise thermometry have not been comparable to those of gas-based techniques owing to limitations from the nonideal performance of electronic detection systems. The noise generated by the wires that connect the sensor to the amplifiers has to be eliminated from the measurement. This is conveniently done by making a four-wire connection to the sensor and using two amplifier chains. The most successful technique to date is the switched-input digital correlator pioneered by Brixy *et al.* (1992) and Edler, Kühne and Tegeler (2004) (Figure 5a). The correlator is implemented by digitizing the signals from the two channels and carrying out the multiplication and averaging function by software. This eliminates the amplifier and transmission line noise superimposed on the thermal noise signal. In practial applications, the thermometer switches between a reference noise source at a reference temperature T_0 and the noise source at the unknown temperature T. The switching removes the effects of drift in the gain

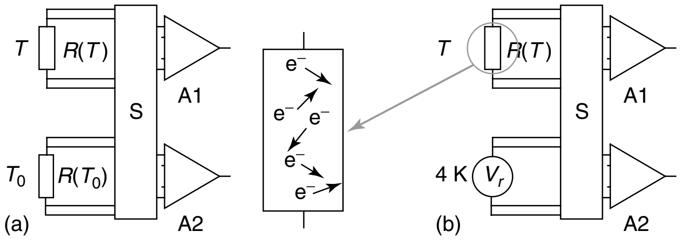

Fig. 5 (a) Block diagram for the conventional relative method with switched-input noise correlator; S, switching and A1, A2, amplification and digitization. (b) Block diagram for the new absolute method.

and bandwidth of the amplifiers and filters. A relative standard uncertainty of 2×10^{-5} at the zinc fixed point has been estimated.

The conventional switching approach does not contain sufficient free parameters to resolve the contradictory requirements for matching both the sensing resistances and the noise powers. Currently, a collaboration between NIST and MSL (Benz *et al.*, 2002) explores a new approach using the perfect quantization of voltages from the Josephson effect. This approach keeps the proven elements of the switched correlator, but separates the roles of the temperature reference and the voltage reference. The sensing resistor in the reference arm of the comparator is replaced by an AC Josephson voltage standard. A block diagram is shown in Figure 5(b). The long-term goal of the project is to build a noise thermometer with an uncertainty of 1×10^{-5} over the temperature range of 83–430 K. First steps were a proof of concept at the fixed points of gallium and water (Nam *et al.*, 2003), and a relative measurement of the zinc freezing point (Labenski *et al.*, 2007). Noise thermometry has been successfully used in the low-temperature range and the interested reader is referred to the review in Fischer and Fellmuth (2005a).

4
Doppler-Broadening Thermometry

Doppler-broadening thermometry is a standard means of diagnostics for high-temperature plasmas (Griem, 1964) and is based on the Doppler shift of the frequency of an electromagnetic wave in a moving frame of reference as compared to a frame at rest. Consider the propagation of a laser beam through an absorption cell containing an ideal gas of atoms or molecules with uniform temperature T. The Gaussian–Maxwell probability density for the velocity of the molecules v is proportional to $\exp[-(v/v_0)^2]$ with $v_0{}^2 = 2kT/m$ for atomic mass m. Around the absorption frequency ν_0, this translates into the corresponding Doppler-broadened absorption line profile with the Doppler width

$$\Delta\nu_\mathrm{D} = \left[\frac{2kT}{(mc_0{}^2)}\right]^{1/2} \nu_0 \qquad (15)$$

This relation allows the determination of the temperature by spectroscopic measurement of a Doppler-broadened absorption line profile and determination of its width. In principle, the measurement can be done using standard laser-spectroscopic techniques. As a main advantage compared to other

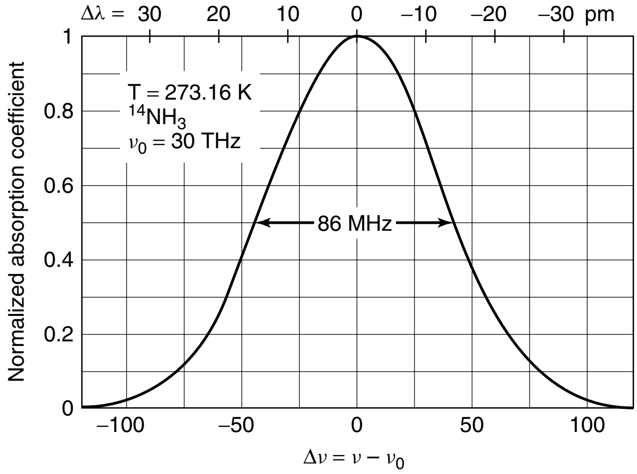

Fig. 6 Doppler-broadened absorption line profile for ammonia $^{14}NH_3$ at T_{TPW} assuming a central frequency $\nu_0 = 30$ THz.

optical methods such as absolute radiation thermometry, the Doppler profile can be determined by relative radiation measurements since it is only its width that is of interest here. Moreover, laser frequencies can be controlled with extremely small uncertainties. However, at the 10^{-6} or 10^{-7} uncertainty level, various other sources of uncertainty will have to be investigated in detail. Apart from the quadratic Doppler effect, these include, among many others, the effects brought about by interatomic interactions, notably the additional line broadening (collisional, transit time, and saturation broadening) and the reduction in Doppler broadening caused by a finite mean free path length (Dicke narrowing). Because of these, measurements may have to be performed at a series of pressure values and extrapolated to pressure zero. It should be noted that the extrapolation to zero pressure is problematic for reasons connected with the speed distribution of particles in monolayers on the cell's surface. Similarly, heating by the absorbed laser power, if it cannot be neglected at all, may require an extrapolation to vanishing laser power.

The possibility of obtaining accurate temperature values has been demonstrated (Daussy, 2007) in an experiment using an ammonia line probed by a CO_2 laser spectrometer close to 30 THz (Figure 6). The absorption signal was recorded by splitting the laser beam in two and then propagating one of the two beams through the ammonia cell for spectroscopy while the other was used as a reference beam. The two beams were amplitude-modulated by two acousto-optic modulators at two different frequencies and then recombined and focused on a single photodetector. On the basis of these preliminary results, Daussy *et al.* (2007) expect that the Doppler-broadening measurements can contribute to the determination of the Boltzmann constant with an uncertainty close to 10^{-6}.

5
Total Radiation Thermometry

It is only with the development of the cryogenic radiometer that accurate measurements of the total radiation emitted from a black body have become possible. The total radiant exitance $M(T)$ of a black body at a temperature T is given by

$$M(T) = \frac{2\pi^5 k^4}{15\, h^3 c^2} T^4 = \sigma T^4 \qquad (16)$$

where σ is the Stefan–Boltzmann constant and c the speed of light. To determine the total radiant exitance, it is necessary, for practical reasons, to make measurements over only a restricted solid angle rather than over a complete hemisphere. An aperture system must be interposed between the black body and the detector so that $M'(T) = gM(T)$, where g is the throughput of the optical system. Provided that g is independent of temperature, we may express

$$\frac{M(T)}{M(T_{\mathrm{TPW}})} = \frac{M'(T)}{M'(T_{\mathrm{TPW}})} = \left(\frac{T}{T_{\mathrm{TPW}}}\right)^4 \qquad (17)$$

where $M'(T)$ and $M'(T_{\mathrm{TPW}})$ are the quantities to be determined. Near room temperature, a measurement of the ratio $M'(T)/M'(T_{\mathrm{TPW}})$ to 1 part in 10^5 is sufficient to determine T to 1 mK or better. Quinn and Martin (1985) and Martin, Quinn and Chu (1988) have demonstrated that such measurements are possible and have obtained results between −130 and 100 °C. The principle is illustrated in Figure 7. The black body radiator at a temperature T irradiates an aperture system at liquid helium temperatures, which allows a beam of thermal radiation to enter the second black body held initially at a temperature of 2 K. The absorbing black body acts as a heat flow calorimeter or cryogenic radiometer. The radiant power absorbed in the calorimeter leads to a rise in its temperature until the radiant power absorbed is balanced to a close approximation by the heat flow along a poorly conducting heat link to a heat sink maintained at a very stable temperature near 2 K. The rise in temperature of the calorimeter is monitored. When equilibrium has been reached, a shutter at liquid helium temperature is closed, cutting off radiation from the black body radiator. At the same time, sufficient electrical power is supplied to a heater on the calorimeter to keep it at the same temperature. Provided a number of conditions are met, this easily measurable electrical power is a very precise equivalent of the thermal radiative power.

For measuring the ratio $M'(T)/M'(T_{\mathrm{TPW}})$, the following parameters of the system had to be evaluated: the emissivity of the radiator, its effective temperatures and the absorptivity of the calorimeter, the diffraction effects at the apertures, the effects of scattering of thermal radiation from surfaces between the apertures, the absorption of thermal radiation at the aperture edges, the departures from ideal geometry of the apertures, the equivalence of radiant and electrical heating of the calorimeter, the uncertainty in the measurements of the electrical power applied to the calorimeter, and the energy transfer from radiator to calorimeter by residual gas. Although some of these parameters had to be known absolutely, others needed to be known only to the extent that their dependence on the wavelength or temperature of the radiator was required. An absolute measurement of $M'(T_{\mathrm{TPW}})$ for the determination of the Stefan–Boltzmann

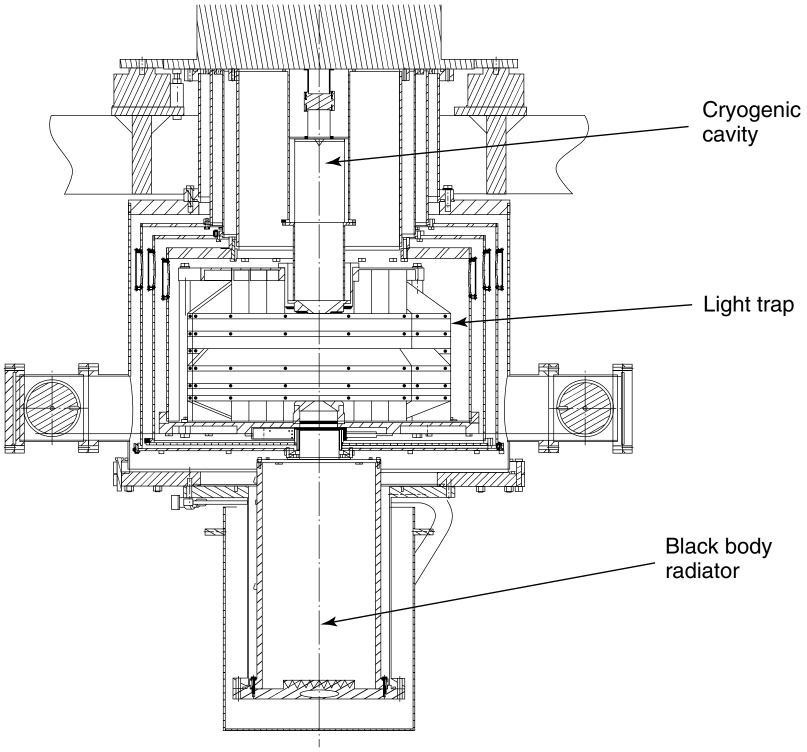

Fig. 7 Principle of cryogenic radiometer.

constant, however, would require an absolute knowledge of all of them (Quinn and Martin, 1985). The result of the determination of the Stefan–Boltzmann constant had an uncertainty of about 1.3 parts in 10^4.

Recently, NPL has built a new version of a total radiation thermometer (Figure 7), which is expected to measure σ with an uncertainty of 0.001% and thermodynamic temperatures between the Hg and Sn fixed points (Martin and Haycocks, 1998) with uncertainties of about 0.5 mK. It should be noted that the determination of σ with an uncertainty of 0.001% corresponds to the determination of the Boltzmann constant with an uncertainty of 2.5 parts in 10^6.

6
Spectral Radiation Thermometry

6.1
Relative Method Referenced to Known Temperature

As reviewed in Section 7, Eq. (18) defines the ITS-90 in the temperature range above the silver point, where $L_\lambda(\lambda, T)$ is the spectral radiance of a black body at a temperature T and wavelength λ, and c_2 is the second radiation constant. The measurement of radiance ratios can also be used to measure thermodynamic temperature if the temperature of one of the black bodies is known in terms of thermodynamic temperature as a reference. This method has been applied

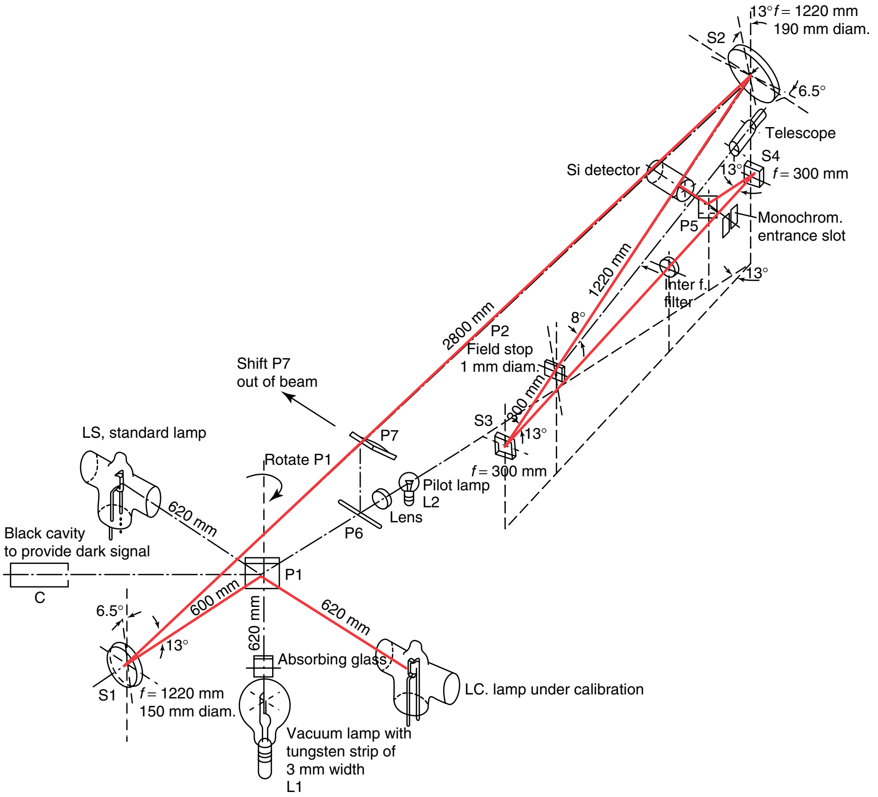
13°
f = 1220 mm
190 mm diam.
S2
6.5°
Telescope
S4
13°
f = 300 mm
Si detector
P5
Monochrom.
entrance slot
1220 mm
13°
Inter f.
filter
8°
2800 mm
P2
Field stop
1 mm diam.
Shift P7
out of beam
300 mm
S3
13°
f = 300 mm
LS, standard lamp
P7
Rotate P1
Pilot lamp
L2
Lens
620 mm
P6
Black cavity
to provide dark signal
P1
C
600 mm
620 mm
6.5°
13°
620 mm
Absorbing glass
S1
f = 1220 mm
150 mm diam.
LC. lamp under calibration
Vacuum lamp with
tungsten strip of
3 mm width
L1

Fig. 8 Optical layout of the spectral radiation thermometer of PTB. Instead of the lamps LC, LS black bodies B1 and B2 can be mounted.

to determine the temperatures of the Al, Ag, and Au fixed points of the ITS-90 and is described in the following:

$$\frac{L_\lambda(\lambda, T_{90})}{L_\lambda(\lambda, T_{90,\mathrm{ref}})} = \frac{\exp(c_2/[\lambda T_{90,\mathrm{ref}}]) - 1}{\exp(c_2/[\lambda T_{90}]) - 1} \tag{18}$$

The spectral radiation thermometer used by Jung (1984, 1986) and Fischer and Jung (1989) in the range 410–962 °C employed two simultaneously running black bodies B1 and B2. They are observed alternatively by means of a rotatable plane mirror P1 (Figure 8). One of the black bodies has the unknown temperature, and the other has the reference temperature. The radiation thermometer alternatively focused the apertures of both black bodies onto a linear detector. An interference filter defined the wavelength. The unknown temperature is calculated from the photocurrent ratio and the reference temperature using Planck's radiation law and taking into account the spectral responsivity of the radiation thermometer. During the first run of measurements, black body B1 had a reference temperature close to $T_{\mathrm{ref}} = 729$ K, a thermodynamic temperature as derived from gas thermometry (Guildner and Edsinger, 1976; Edsinger and Schooley, 1989). Black body B2, immersed in freezing aluminum, had the unknown temperature.

During the second run of measurements, black body B2, constantly running aluminum freezes, served as the reference radiator. The thermodynamic temperature T that corresponded to the temperature of black body B1 was unknown. The latter was set to values between 410 and 630 °C using standard platinum resistance thermometers (SPRTs) while the corresponding thermodynamic temperatures were obtained from the measured photocurrent ratios and the redetermined aluminum point. The results gave rise to the SPRT reference function used for the ITS-90 in the high-temperature range. Finally, during the third run, the freezing points of silver and gold have been determined (Fischer and Jung, 1989). The reference black body contained aluminum and the other one contained freezes of silver and gold, respectively.

6.2
Absolute Method

The measurement of radiance ratios does not require one to know the spectral responsivity of the radiation thermometer absolutely. However, the determination of the spectral radiance without referencing to a source of known temperature requires an instrument that has a known absolute spectral response with a well-defined geometric viewing system. This system is called, in its simplest version, a *filter radiometer* and has a set of two view-defining apertures with accurately known dimensions (Figure 9). Only recently has it become possible to measure the absolute spectral responsivity of a filter radiometer with sufficient accuracy to compete with the method used in the measurement of radiance ratios. This became possible through the use of cryogenic radiometers as primary reference standard defining a scale of spectral responsivity. The first cryogenic radiometer was constructed at NPL (Quinn and Martin, 1985) to measure the total radiation as described earlier. To establish scales of spectral responsivity, cryogenic radiometers use the spectral power of a monochromatic source, such as a laser, and subsequently calibrate the response of a transfer detector (Fox, 2002). The additional uncertainty of the calibration of the filter transmission has to be taken into account, which cannot be lower than the uncertainty of the cryogenic radiometer used for calibration. Altogether, this results in best relative uncertainties at the 10^{-4} level for this

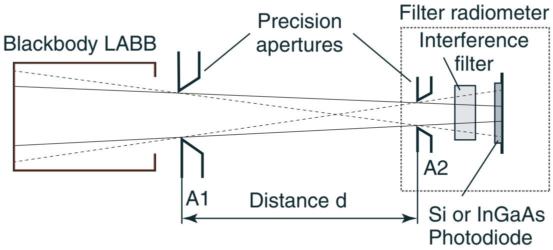

Fig. 9 Optical layout of the PTB primary radiation thermometer with the large-area black body LABB, two precision apertures A1 and A2, and the filter radiometer.

type of radiation thermometry and measurements at temperatures about 500 °C and above (Fox, 2002; Taubert *et al.*, 2003; Yoon *et al.*, 2005).

At the time of establishment of the ITS-90, there were no direct measurements of the thermodynamic temperature T above 729 K of sufficiently low uncertainty that could be used to anchor the scale. Instead, the fixed points of Al, Ag, Au, and Cu were determined through ratios of radiance using radiation thermometers as described in Section 6.1. The temperature value assigned to the reference point at 729 K (456 °C) was the mean of two separate NIST gas thermometry experiments (Rusby *et al.*, 1991). Since these two experiments differed by about 30 mK, this became the dominant uncertainty of ITS-90 realizations for all higher temperatures. This relatively large thermodynamic uncertainty of T_{90} propagates as $T^2{}_{90}$ and is thus around 50 mK at the gold point. To reduce these uncertainties, work is in progress applying acoustic thermometry (see Section 2.1) and spectral radiation thermometry.

Using absolute filter radiometry, measurements of $T - T_{90}$ in the temperature range from 660 to 962 °C (Stock *et al.*, 1996) and subsequently down to zinc-point temperatures (Hartmann, Taubert and Fischer, 2002; Taubert *et al.*, 2003; Noulkhow *et al.*, 2007) have been performed at PTB. They employed a large-area black body formed by two concentric sodium heat pipes so that the filter radiometer does not need imaging optics whose transmittance is to be calibrated (Figure 9).

A different approach is applied at NIST (Yoon *et al.*, 2005) and at NPL (Fox, 2002), where imaging filter radiometers were developed. With these systems, the thermodynamic temperatures of fixed-point black bodies with their small cavity apertures can be measured directly.

7
The International Temperature Scale of 1990

The International Temperature Scales reflect the most recent state of metrological accuracy and therefore they are replaced by new versions from time to time. The International Temperature Scale of 1927 (ITS-27) was the first to overcome the practical difficulties of the direct realization of thermodynamic temperatures by gas thermometry and the first universally acceptable replacement for the differing existing national temperature scales. Finally on 1 January 1990, the ITS-90 (Preston-Thomas, 1990a) came into force. We note that the thermodynamic temperature T (K) may also be expressed in terms of Celsius temperature t according to

$$t(^\circ\mathrm{C}) = T(\mathrm{K}) - 273.15 \qquad (19)$$

The ITS-90 accordingly defines both international Kelvin temperatures T_{90} and international Celsius temperatures t_{90} by the corresponding relation:

$$t_{90}(^\circ\mathrm{C}) = T_{90}(\mathrm{K}) - 273.15 \qquad (20)$$

Both the thermodynamic and the International Temperature Scale have the same units – the kelvin and the degree Celsius. Users sometimes prefer kelvin in the range below 273.15 K and degree Celsius above this point.

The thermodynamic basis of the ITS-90 is described in Rusby *et al.* (1991) and recommendations for its realization are given in Preston-Thomas, Bloembergen and Quinn (1990b). The ITS-90 extends upward from 0.65 K to the highest temperature practically measurable in terms of the Planck radiation law using monochromatic radiation. It is based on 17 well-reproducible thermodynamic states of equilibrium, the defining fixed points: boiling points (3–5 K with helium, 17 and 20.3 K with hydrogen), triple points (equilibrium hydrogen, neon, oxygen, argon, mercury, and water), melting point of gallium, and freezing points (indium, tin, zinc, aluminum, silver, gold, and copper). To these states, numerical values of the temperature T_{90} are assigned. These are the values that have been determined by measurements of thermodynamic temperatures T in several national metrology institutes (Rusby *et al.*, 1991). They are considered to be the best estimates at the time the scale was adopted. The defining fixed points are listed in Table 1.

Tab. 1 Defining fixed points of the ITS-90 with uncertainties $u(T_{90})$ of the best practical realization in terms of ITS-90 and uncertainties $u(T)$ of the thermodynamic temperature.

Fixed Point	T_{90} (K)	$u(T_{90})$ (mK)	$u(T)$ (mK)
Cu	1357.77	15	60
Au	1337.33	10	50
Ag	1234.93	1	40
Al	933.473	0.3	25
Zn	692.677	0.1	13
Sn	505.078	0.1	5
In	429.7485	0.1	3
Ga	302.9146	0.05	1
H_2O	273.16	0.02	0
Hg	234.3156	0.05	1.5
Ar	83.8058	0.1	1.5
O_2	54.3584	0.1	1
Ne	24.5561	0.2	0.5
e-H_2	≈20.3	0.2	0.5
e-H_2	≈17.0	0.2	0.5
e-H_2	13.8033	0.1	0.5
^{4}He	4.2221	0.1	0.3

All values are quoted as standard uncertainties. Values have been taken from Table 1.2 of the Supplementary Information for the ITS-90 (Preston-Thomas, Bloembergen and Quinn, 1990b).

7.1
Interpolating the Scale

For the first time, the ITS-90 comprises a number of ranges and subranges, throughout each of which temperatures T_{90} are defined differently. Several of these ranges or subranges overlap. In the range from 0.65 to 5 K, relations of helium vapor pressure to temperature are used to interpolate. From 3 up to 25 K, an interpolating version of the CVGT is applied to define the scale. From the triple point of equilibrium hydrogen (13.8033 K) to the freezing point of silver (1234.93 K), SPRTs are used to interpolate between the fixed points. The temperature dependence of the resistivity of high-purity platinum is too complicated for a description based only on the calibration at the available fixed points of sufficient quality (Nicholas, 1999; Quinn, 1990), but it is highly reproducible from sample to sample. Therefore, the ITS-90 prescribes two general reference functions for the ranges below and above the TPW representing a "typical" SPRT (Preston-Thomas, 1990a). Values and derivatives of the two functions are continuous at the TPW. This allows to describe the characteristic of an SPRT by the sum of a reference function and an individual deviation function. The coefficients of the deviation function are deduced from the results of the calibration at the defining fixed points of the ITS-90. The calibration can be performed for 11 subranges, which overlap more or less, using different deviation functions. To avoid the influences of the dimensions of the platinum wire, both the reference and deviation functions are specified for the resistance ratio $W(T_{90}) = R(T_{90})/R(273.16\text{ K})$, where $R(T_{90})$ is the resistance at a temperature T_{90} and $R(273.16\text{ K})$ is the resistance at the TPW. A detailed analysis of the uncertainty in the realization of the SPRT subranges of the ITS-90 is given in White *et al.* (2007).

An acceptable SPRT must be made from pure, strain-free platinum and satisfy one of the relations to be checked at the triple point of mercury ($W(-38.8344\,^\circ\text{C}) \leq 0.844235$) or the melting point of gallium ($W(29.7646\,^\circ\text{C}) \geq 1.11807$). An acceptable SPRT to be used up to the freezing point of silver (961.78 °C) must also satisfy the relation: $W(961.78\,^\circ\text{C}) \geq 4.2844$. The first two conditions guarantee a minimum purity of the platinum and the third aims at avoiding SPRTs with excessive leakage currents at high temperatures. Three types of different designs cover the SPRT range from 13.8 K to 962 °C, where the first two types have similar sensor elements:

1. The capsule-type SPRT from 13.8 to 273 K (sometimes 430 K). Characteristic data are as follows: $R(0\,^\circ\text{C}) = 25\ \Omega$, 5 mm diameter, 60 mm length, filled with 30 kPa helium at room temperature, the four platinum leads are taken out through a glass seal.
2. The conventional long-stem SPRT from −189 to 420 °C (sometimes 630 or 660 °C). Characteristic data are as follows: $R(0\,^\circ\text{C}) = 25\ \Omega$, 7 mm diameter, 600 mm length, filled with dry air. The sensors of 0.07-mm-diameter platinum wire are supported by insulators of mica, alumina, or silica.
3. The high-temperature long-stem SPRT from 0 to 962 °C (suited to −189 °C). Characteristic data are as follows: $R(0\,^\circ\text{C})$ between 0.25 and 2.5 Ω, 7 mm diameter, 600–800 mm length, filled with 90% argon + 10% oxygen at 20 kPa near room temperature. The sensors and leads are designed such as to minimize the strain on heating and cooling and the leakage through the insulation resistance. Platinum wire up to 0.4 mm diameter and insulators and sheaths of quartz are used to achieve this.

From 961.78 °C, the freezing point of silver, up to the highest practicably measurable temperatures T_{90}, is defined in terms of the ratio of the spectral radiances $L_\lambda(T_{90})$ and $L_\lambda(T_{90,\text{ref}})$ of two black bodies according to Eq. (18). One of them has the temperature T_{90} to be determined. The other has the reference temperature $T_{90,\text{ref}}$ that stands for one of the freezing points of silver, gold, or copper. The ratio is measured by means of a spectral radiation thermometer. The temperature T_{90} is calculated from the measured ratio using Planck's law for monochromatic radiation. Appropriate designs of the apparatus and good current practice of their application are extensively described in "Supplementary Information for the ITS-90" (Preston-Thomas, Bloembergen and Quinn, 1990b).

8
The Provisional Low-Temperature Scale PLTS-2000

In October 2000, the Comité International des Poids et Mesures adopted the

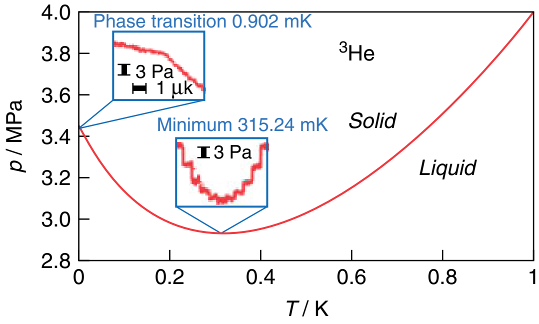

Fig. 10 Phase diagram of ^{3}He to define the Provisional Low-Temperature Scale PLTS-2000.

PLTS-2000 (CIPM, 2001). It is based on noise and magnetic thermometry performed at three institutes (Rusby *et al.*, 2002; Fellmuth, Hechtfischer and Hoffmann, 2003). Considering the uncertainty estimates for the thermometers used and the spread of the results obtained, the relative standard uncertainty of the PLTS-2000 in thermodynamic terms has been estimated to range from 2% at 1 mK to 0.05% at 1 K.

The PLTS-2000 is defined from 0.9 mK to 1 K by a polynomial with 13 terms that describes the temperature dependence of the melting pressure of ^{3}He, which is shown in Figure 10. Furthermore, four natural features on the ^{3}He melting curve can be used as intrinsic fixed points of temperature and pressure because their temperature and pressure values are also defined in the text of the PLTS-2000: the minimum pressure (315.24 mK, 2.93113 MPa), the transition to the superfluid "A" phase (2.444 mK, 3.43407 MPa), the "A-to-B" transition in the superfluid (1.896 mK, 3.43609 MPa), and the Néel transition in the solid (0.902 mK, 3.43934 MPa). The melting pressure of ^{3}He has been chosen as scale carrier for different reasons. First of all, it is a thermodynamic property of a pure substance, that is, no nonuniqueness due to a different interpolation behavior of thermometers occurs. The melting pressure can be reproduced much better than the readings of all other thermometers and a temperature range of about three decades is covered. Furthermore, apart from a narrow range near the minimum of the melting curve, a high resolution down to 0.1 µK can be achieved. The minimum is caused by the unusual fact that the entropy of the liquid phase becomes smaller than the entropy of the solid phase.

The occurrence of the intrinsic fixed points of temperature and pressure is of great importance for using the melting pressure of ^{3}He as scale carrier because melting-pressure thermometry at the highest level of accuracy requires the measurement of the pressure with an uncertainty, which is already close to that of the best national pressure standards. The use of the intrinsic fixed points decreases the uncertainty significantly since the dominating influence of the components caused by the calibration of the pressure balance applied as standard instrument of the highest quality and the hydrostatic-head correction is removed.

For the realization of the PLTS-2000, the melting pressure must be measured in situ because for temperatures below the minimum of the ^{3}He melting curve, the pressure-sensing capillary becomes blocked by solid ^{3}He. However, also above the minimum, the usual way of applying a ^{3}He melting-pressure sensor (MPS) is the "blocked capillary method," where the MPS is isolated during operation by a solid ^{3}He plug in the pressure-sensing capillary and thus the conditions are most stable. Up to now, exclusively homemade MPSs were used that are all capacitive diaphragm gauges, but differ in the detailed design (Rusby *et al.*, 2007). Recommendations concerning the facilities and procedures to be used for realizing the PLTS-2000 are given in Schuster, Hoffmann and Hechtfischer (2001). The state-of-the-art level of accuracy can be deduced from the uncertainty budgets presented in Engert, Fellmuth and Hoffmann (2004); Peruzzi and de Groot (2003).

The dissemination of the PLTS-2000 is difficult because the parameters of the available secondary thermometers cause a large uncertainty of their reading. For instance, resistance and capacitance thermometers suffer from a high instability and relatively large thermal resistances especially at temperatures below about 0.6 K. Therefore, the smallest uncertainty of the scale dissemination, being comparable with that of the scale realization, is obtained by calibrating superconductive reference samples as transfer standards for the calibration of MPSs in situ in the users' cryostats. These superconductive transfer standards can also be used for checking the parameters of other secondary thermometers together with the measuring equipment. They are of great importance for approximating the PLTS-2000 in dilution refrigerators without a demagnetization stage that do not allow to realize the three intrinsic fixed points at the lower end of the definition range (below 3 mK). The additional uncertainty component caused by the approximation may be very small if an appropriate combination of reference samples is used. For instance, it is possible to decrease this component below 20 μK in the whole range by calibrating an MPS at the minimum of the melting curve as well as at the super-to-normal-conducting transitions of tungsten (15 mK) and molybdenum (9.2 K) samples (Schuster, Hoffmann and Hechtfischer, 2001).

9
Metal-Carbon Eutectic Phase Transitions

At extremely high temperatures, most materials become so reactive that the choice of a crucible material, which does not affect the fixed-point material, becomes severely limiting. There are many types of metals with melting temperature higher than copper, the highest fixed point of the ITS-90, but none had been used successfully as practical fixed point because contamination by the graphite crucible would cause depression of the melting and freezing temperatures.

Yamada *et al.* (1999) showed that the use of metal–carbon eutectics as the fixed-point material offers a solution to the combined problem of the graphite crucible and the fixed-point materials. Graphite, the crucible material, is already a component of the fixed-point alloy and cannot cause contamination. As shown in the binary phase diagram for Ni–C in Figure 11 (Massalski, 1990), the liquid phase takes its lowest temperature at the eutectic point, and there cannot be any further depression of the freezing point by carbon contamination. The

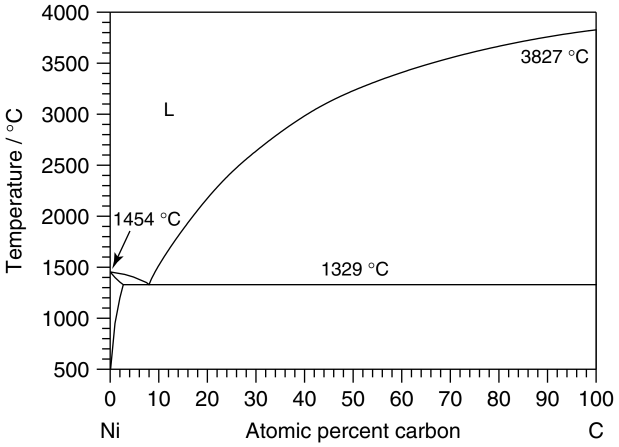

Fig. 11 Binary phase diagram of the eutectic system Ni–C, where L denotes the liquid phase.

molten metal, at slightly higher temperature than the eutectic point, would always be slightly richer in C content than the eutectic. However, solidification of graphite during the cooling would reduce the carbon content and the molten metal would reach eutectic composition when the freezing commences. Thus, reproducible plateaus are observed. There is no way to adjust the composition ratio because carbon could endlessly be supplied from the crucible. Furthermore, graphite crucibles form black bodies of high emissivity suitable for radiation thermometer calibration, and also there is no concern about reaction with the surrounding furnace material, which is also graphite.

Possible high-temperature fixed points using metal–carbon eutectics are listed in Figure 12. The melting temperatures are conveniently spaced out of the temperature range from the copper point up to 2500 °C. Metal–carbon eutectics, with no special precautions or procedures and with a wide variation in the cooling rates for the preceding freezes, have shown melting plateau repeatability of 20 mK (Woolliams *et al.*, 2006) in some cases. Evidently, material purity plays the major role as the metal–carbon eutectics appear to be more susceptible to impurities than pure-metal fixed points.

The eight metals in the box in Figure 12 are the ones that show the

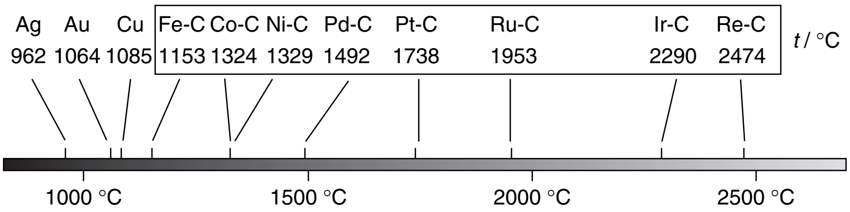

Fig. 12 Approximate melting temperatures of metal–carbon eutectic compositions. The three highest defining fixed points of the ITS-90 are shown on the left, outside the box.

simplest binary phase diagrams with carbon and are under consideration to become supplementary fixed points of the ITS-90. For most of the others, carbides are formed, which make the phase diagram complex. For instance, Ti gives TiC, with a melting temperature of 3067 °C and forms an eutectic with Ti and with graphite. The possible material combinations are Ti–TiC eutectic in a Ti (or TiC) crucible and TiC–C eutectic in a C (or TiC) crucible. The former is not easy to realize as reaction between the crucible material and the furnace material is difficult to prevent. On the other hand, TiC–C eutectic in a graphite crucible requires no additional technical improvements and is of more interest because the eutectic point has a temperature above the 3000 K mark. Other candidates are ZrC–C eutectic at 2927 °C and HfC–C eutectic at 3180 °C (Woolliams *et al.*, 2006).

In the ITS-90, the defining fixed point with the highest temperature is the copper point. Above this, no fixed point is available that can be widely used as a practical calibration device. High-temperature fixed points using metal–carbon eutectics may change this situation. The eutectics can be held in graphite crucibles, which makes the cells practical devices for use at very high temperatures. These fixed points can benefit the future high-temperature standards in various ways. They may play the role of transfer standards replacing the standard tungsten strip lamps in radiation thermometry or standard incandescent lamps in source-based radiometry. The temperature scale may be realized with much smaller uncertainty by interpolation at a selected number of these fixed points (Bloembergen *et al.*, 2003). Extension of this technique to above 3000 K is possible with metal carbide–carbon eutectics. However, the temperature values to be assigned to these new fixed points have to be determined beforehand with sufficiently low uncertainty by primary thermometry (Machin *et al.*, 2007).

10
Industrial Resistance Thermometers

Although it is not possible to develop precise formulae for the temperature dependence of the resistance of metals (Nicholas, 1999), the dependence exhibits an extremely stable and well-characterized form. Consequently, resistance thermometers are capable of quite reliable and economical industrial-quality temperature measurement. Besides nickel and copper, pure platinum proves to be the best material for most purposes; hence, the platinum resistance thermometer is the most common form of this instrument and most national standard specification codes, as well as the IEC-60751 (International Electrotechnical Commission, 1995a), are limited to platinum.

IEC-60751 defines the temperature–resistance relationship for industrial platinum resistance thermometers (IPRTs) in the temperature range of $0\,^\circ\mathrm{C} \le t \le 850\,^\circ\mathrm{C}$ by a simple quadratic equation:

$$R_t = R_0(1 + At + Bt^2) \tag{21}$$

For the range from −200 to 0 °C, an additional term is used:

$$R_t = R_0(1 + At + Bt^2 + C(t - 100)t^3) \tag{22}$$

R_0 is the resistance at the ice point, usually 100 Ω (Pt100 type). Values for the constants in these two equations are given for ITS-90, for example, in Bernhard (2004). Using these formulae, the IPRTs can be interchanged and will give temperature readings within the allowable tolerances without any calibration. For two classes of IPRTs, the temperature tolerances δt defined in IEC-60751 at a temperature t are as follows:

$$\text{class A}: \ \delta t = \pm(0.15 + 0.002|t(^\circ\mathrm{C})|)^\circ\mathrm{C} \quad (23)$$

$$\text{class B}: \ \delta t = \pm(0.30 + 0.005|t(^\circ\mathrm{C})|)^\circ\mathrm{C} \quad (24)$$

The tolerances of class A must not be exceeded in the temperature range $-200\,^\circ\mathrm{C} \leq t \leq 650\,^\circ\mathrm{C}$, and the class B values are valid in the whole usable temperature range specified by the manufacturer. In the new version of IEC-60751, which is currently in the final international adoption process, different classes with more temperature subranges, taking into account also the sensor construction, are proposed; for details see Bernhard (2004). In general, IPRTs should be considered as an alternative or replacement for liquid-in-glass thermometers when an uncertainty of better than 0.5 °C is required. It is not possible to batch IPRTs with tolerances close enough to work at a level of uncertainty of about 0.1% or better; hence, they have to be calibrated individually. Then, they may have the capability to be used as laboratory standards with uncertainties approaching 10 mK at temperatures near to room temperature.

Compared to thermocouples, IPRTs offer some advantages in sensitivity, although they are more limited in temperature range. An IPRT of the Pt100 type has an effective sensitivity of about 400 μV K^{-1} for the usually used measuring current of 1 mA. This is a factor of about 40 greater than the sensitivity of a type R or S thermocouple. The main disadvantage of an IPRT is the need of lead compensation, which usually requires three or four leads running to the sensor rather than two wires for thermocouples.

As described above, the platinum wire in the sensor of an SPRT is wound in a configuration that is as strain free as possible. This makes it rather susceptible to vibration and shock and, therefore, IPRT sensors have to be designed differently. For industrial use, in addition the sheathing must be more robust than for SPRTs. IPRT sensor constructions use often encapsulated Pt-wires wound around glass or ceramic materials whose thermal expansion coefficients match that of platinum. Cheaper sensors can be manufactured applying lithographic techniques. IPRTs of this type are available as thick-film or thin-film sensors, the latter also as an SMD. A comprehensive overview of the different constructions is given in Bernhard (2004). The direct contact between the platinum sensor element and the supporting material may cause strains or even an exceeding of the yield strength. The yield strength determines the maximum instability and corresponds to a temperature equivalent of the order of 0.5 K. Thus, the strains may cause a hysteresis of several tenths of a kelvin depending on the application range (Quinn, 1990).

So far we have described only resistance thermometers based on platinum sensor elements. For low temperatures down to about 0.5 K, that is, below the operation range of PRTs, especially rhodium–iron resistance thermometers (RIRTs) have been developed (Bedford *et al.*, 1990). Since their resistance versus temperature characteristic is complicated and nonunique, they have to be calibrated individually at 10–30 temperatures in the range from 0.5 to 30 K depending on the desired uncertainty level. RIRTs are available as precision thermometers designed similar to capsule-type SPRTs, the long-term instability of which may be of the order of only a few tenths of a millikelvin, as well as industrial thermometers with wire sensors wound on a ceramic substrate or thin-film sensors. The room-temperature resistance is between 25 and 100 Ω.

In contrast to metal-based resistance thermometers, for instance, thermistors manufactured from semiconducting metal oxides and resistance thermometers made from germanium, carbon, or some other materials show high-negative temperature coefficients of resistance. Thermistors have relative temperature coefficients of about $-0.04\,\mathrm{K}^{-1}$ near room temperature, compared with a 10-fold smaller and positive value of $+0.004\,\mathrm{K}^{-1}$ for metals. They typically can be obtained commercially with room-temperature resistances ranging from 100 Ω to 1 MΩ and can be useful for temperature measurement from −200 to 300 °C. Their high sensitivity permits use with inexpensive, direct reading instruments and, as the lead resistance is small compared to the sensor resistance, lead compensation is usually unnecessary. The characteristic of thermistors is very nonlinear and is described to a good approximation by the empirical relation based on the formula given in Steinhart and Hart (1968):

$$\frac{1}{T} = A + B\,\ln\left(\frac{R}{R_0}\right) + C\left[\ln\left(\frac{R}{R_0}\right)\right]^2 + D\left[\ln\left(\frac{R}{R_0}\right)\right]^3 \qquad (25)$$

where R_0 is a convenient reference resistance. For limited temperature ranges, measurement uncertainties below 0.1 °C can be easily obtained.

11
Thermocouples

In industry, thermocouples are the most widely used of all temperature sensors. Their basic simplicity and reliability have an obvious appeal for many applications. However, when accuracies greater than normal industrial requirements are called for, their simplicity in use is lost and their reliability cannot be assumed a priori.

The thermocouple is a device in which the difference between the electromotive forces (emfs) in its two arms gives an indication of the temperature difference between the hot and cold ends of the two wires (thermoelements). This difference is known as the *Seebeck effect*. The point of connection of the two dissimilar arms A and B, at a temperature T_1, is called the *measuring junction*, and the free ends are referred to as the *reference junction*. To permit measurements to be made of the thermoelectric potential difference between the free ends, an additional pair of identical conductors is attached, which leads to a detector. It is evident that the Seebeck effect is in no way a junction phenomenon but is, instead, a temperature-gradient phenomenon. For a proper understanding of the

Tab. 2 Thermocouple types in common use with their letter designations, typical compositions in percent weight, and temperature ranges for which tolerances apply.

Type	Thermoelements		Temperature range (°C)
	Positive	Negative	
B	Pt 30Rh	Pt 6Rh	600 to 1700
R	Pt 13Rh	Pt	0 to 1600
S	Pt 10Rh	Pt	0 to 1600
K	Chromel (Ni 9.5Cr 0.5Si)	Alumel (Ni 5(Si,Mn,Al))	−40 to 1200
N	Nicrosil (Ni 14.2Cr 1.4Si)	Nisil (Ni 4.4Si 0.1Mg)	−40 to 1200
E	Chromel (Ni 9.5Cr 0.5Si)	Constantan (Cu 44Ni)	−200 to 900
J	Fe	Constantan (Cu 44Ni)	−40 to 750
T	Cu	Constantan (Cu 44Ni)	−200 to 350

behavior of thermocouples, this cannot be overemphasized. The thermoelectric potential difference between the free ends, both at temperature T_0, is

$$E_{AB}(T_0 \to T_1) = \int_{T_0}^{T_1} [S_A(T) - S_B(T)]\,dT \qquad (26)$$

where $S_A(T)$ and $S_B(T)$ are the thermopowers of the respective conductors. $E_{AB}(T_0 \to T_1)$ is a unique function of T_0 and T_1 only if the thermoelements are homogeneous throughout the temperature gradient. The effect of the presence of an inhomogeneity is to add a small additional thermopower at the place in the thermoelement where the inhomogeneity occurs. The thermoelectric potential difference of an inhomogeneous thermocouple will be a function of its position as well as the temperature difference between the hot and cold ends. In using a thermocouple to explore the temperature gradient in a furnace, there is thus always the problem of deciding whether or not the change in potential difference, as the thermocouple is moved, indicates the presence of a temperature gradient in the furnace or an inhomogeneity in the thermocouple. Fortunately, it is usually possible to place an upper limit on the size of the effects of inhomogeneities.

Over 20 different types of thermocouples are extensively used and, of these, 8 have been standardized, with their emf–temperature relationships represented by internationally recognized standard reference functions (International Electrotechnical Commission, 1995b). The metals used in the standardized thermocouple types are given in Table 2, with the letters that identify them and the temperature ranges for which tolerances apply (International Electrotechnical Commission, 1989). In general, the tolerances are higher than those for IPRTs and are in the order of 1 K.

The thermocouples mentioned in Table 2 fall into three groups. The first is the rare-metal group, comprising types B, R, and S, based on platinum and its alloys with rhodium. They are the most accurate of the tabulated thermocouples and may be used at higher temperatures, but they are more expensive and are particularly sensitive to contamination.

The second group consists of the two nickel-based thermocouples, types

N and K. They are preferred for most applications not requiring the higher temperature limit or accuracy of the rare-metal thermocouples. The choice between types N and K depends especially on whether they are in the bare-wire or the mineral-insulated, metal-sheathed (MIMS) form. In the MIMS configuration, the thermoelements, their insulation, and a sealed sheath are integrated into a flexible cable. In contrast, bare-wire thermocouples, insulated in loose-fitting beads, are exposed to the local atmosphere and are vulnerable to the effects of oxygen, carbon, sulphur, and so on.

The third group, comprising the types E, J, and T, is based on the use of constantan as the negative arm. Constantan has the most negative value of thermopower; hence, the net thermopowers for these thermocouples are high.

Besides the standardized thermocouples, the elemental thermocouples – those whose thermoelements are pure elements – need to be mentioned. A detailed review of the recent developments of elemental thermocouples is given in Bentley (1998). A pure element is more homogeneous and thermoelectrically more stable than an alloy, because it is free of effects that arise from lattice ordering, selective volatilization, and oxidation. The most stable thermocouple is Au/Pt, although its use is limited to temperatures below about 1000 °C. The instabilities in Pt/Pd are greater, mainly due to inhomogeneity effects in the Pd wire. For example, the drift of a Pt/Pd thermocouple may be as low as 0.02 °C for 200 h at 963 °C (Burns and Ripple, 1996). Even so, these changes are about 10 times smaller than that would occur in a type R or S thermocouple under the same conditions.

Above 1750 °C, the most useful thermocouples are those based on W and Re. There are several thermocouple combinations in this category, such as W and various alloys from W 3Re to W 26Re. Their main difficulty is that on heating above recrystallization temperatures, they become brittle. The recrystallization temperature of W is 1200 °C and it has the greatest embrittlement problem, whereas W 26Re contains enough Re so that there is no problem of this kind. The thermocouple combination W 5Re versus W 26Re is possibly the best choice. It should not be used in air, but in a vacuum, in an inert atmosphere, or in dry hydrogen. Its upper limit for reliable operation is considered to be 2760 °C.

12
Radiation Thermometers

The use of noncontact temperature measurement by radiation thermometers has increased considerably in the last few years. Accordingly, a wide variety of different designs of radiation thermometers for laboratory and industrial use is available. In contrast to the procedure prescribed in ITS-90, for most practical radiation thermometers, there is no direct comparison against a reference black body of known temperature. Rather, the thermometers make use of the stability of their built-in detectors and after a calibration of their sensitivity against a temperature standard of the ITS-90 they read directly in temperature.

Practical total radiation thermometers, operating at room temperature, are the simplest radiation thermometers consisting generally of a very basic imaging system restricting the field of view with an aperture. The detector is usually a

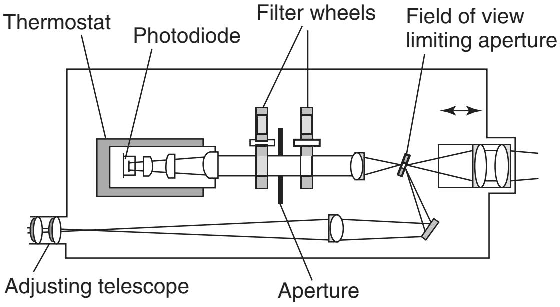

Fig. 13 Optical layout of the high-precision radiation thermometer LP3.

thermopile bolometer, which measures the rise in temperature of a blackened disk and allows the detection of the whole radiated spectrum. Because of their comparatively low sensitivity, they are, on the other hand, the most sensitive of the radiation thermometers to emissivity variations of the target. For example, a 10% overestimation of the emissivity will result in a 2.5% underestimation of the absolute temperature.

Spectral-band radiation thermometers use a filter to selectively measure only thermal radiation in a narrow band as described in Section 6. A relative bandwidth ($\Delta\lambda/\lambda$) less than a few percent is typical. Usually a silicon-semiconductor detector is used. Spectral-band radiation thermometers are used to overcome surface emissivity problems. In this case, the center wavelength λ is chosen to be as short as possible while still giving enough signal. From Eq. (18) it is derived that typically for this kind of radiation thermometers $c_2/\lambda T \cong (10-20)$, so that a 10% overestimation of the emissivity will lead to only a (0.5–1)% underestimation of the temperature.

The most common bands are around 0.65 μm for temperatures above 700 °C and 0.85–0.9 μm for temperatures above 500 °C. The design of the high-precision radiation thermometer LP3, widely used in national metrology institutes, is shown in Figure 13 (Fischer *et al.*, 2002). For lower temperatures, longer wavelengths and different detectors are applied. For example, at 3.43 μm many plastics have high absorptance. A radiation thermometer operating at this wavelength will thus see even a very thin plastic film as opaque and having high emissivity, allowing accurate temperature measurements during the manufacturing process at 200–300 °C.

The thermal radiation emitted by objects around room temperature is comparatively low and exclusively in the infrared (IR). To obtain sufficient signal, most radiation thermometers will use a wider band of wavelengths. Atmospheric absorption of IR radiation limits the choice of bands to mainly two windows, the (3–5) μm band and the (8–14) μm band. Pyroelectric or thermoelectric detectors are used to detect the radiation, and the wavelength range is often set by the transmission of the optics materials. Since the filter defining the spectral sensitivity, the optics, and the apertures will also radiate energy in this wavelength region, there must be some form of background temperature compensation within the instrument. Some instruments solve this problem by using a rotating sector disk providing a reference signal of known temperature. Recently, there has been considerable progress in the instrumentation for the calibration of low-temperature radiation thermometers (Hollandt *et al.*, 2003/4).

13
New Definition of the Kelvin

Experimentally testing of the fundamental laws of physics involves, in practice, the precise determination of the fundamental constants appearing in the laws. Since they are ultimately related to the physical units, the precise experimental realization of the latter is an unavoidable prerequisite for the progress in the development of our physical understanding of nature. Precision experiments relating the (space- and time-independent) fundamental constants to the system of units guarantee stability and reproducibility. The essence of current activities is that prototypes, which may vary uncontrollably with time and location, are replaced by abstract experimental prescriptions that relate the units to the constants. This ensures that the requirement of invariance with space and time is fulfilled. This approach is shown here for the definition of the kelvin and the Boltzmann constant. The unit of temperature T – the kelvin – is presently defined by the temperature of the TPW. Thus, the kelvin is linked to a material property. Instead, it would be advantageous to proceed in the same way as with other units: to relate the unit to a fundamental constant and fix its value. By this no temperature value and no measurement method would be favored. For the kelvin, the corresponding constant is the Boltzmann constant k, because temperature always appears as thermal energy kT in fundamental laws of physics.

As shown by the recently published CODATA values for fundamental physical constants (Mohr, Taylor and Newell, 2008), the 2006 recommended value of the Boltzmann constant k with $u_r(k) = 1.7 \times 10^{-6}$ is to a very large extent determined by the NIST result (Moldover *et al.*, 1988) and, therefore, is not yet regarded as sufficiently corroborated to replace the present definition of the kelvin. The important point to note here is that the measurement uncertainty of any value of k would be transferred to the value of T_{TPW} if that k value were taken to be the exact value of the Boltzmann constant and used to define the kelvin. Hence, if the 2006 CODATA recommended value is be fixed as the exact value of k in future, the best estimate of T_{TPW} would still be 273.16 K. However, this value would no longer be exact (as it is now as a result of the current definition of the kelvin) but would become uncertain by $u_r(T_{\mathrm{TPW}}) = 1.7 \times 10^{-6}$, which corresponds to 0.46 mK.

Considering the uncertainty estimates in the sections mentioned earlier, the two most promising methods for the reduction in the uncertainty of k currently are DCGT (Section 2.2) and AGT (Section 2.1). Doppler-broadening thermometry (Section 4) utilizing radiation measurements has only recently been proposed for the purpose of determining the Boltzmann constant (Daussy *et al.*, 2007) and is presently under investigation at the University Paris North, France, with respect to the uncertainty that can possibly be achieved. In a joint project of the universities of Milan and Naples (Fellmuth and Fischer, 2007), Italy, a similar experiment is under development. They will determine the Doppler broadening of an absorption line of water vapor probed by a diode-laser-based spectrometer system in the near IR. It seems to be possible that the DCGT method at PTB or the AGT work of different groups will have been advanced so far by the end of 2010 that they can contribute to an improved value of k or R with similar relative uncertainty as that obtained by Moldover *et al.* (1988) with the AGT in 1988. Thus, an improved value of the Boltzmann constant proposed for the definition of the kelvin would ideally have been determined by at least these two fundamentally different methods and be corroborated by other – preferably optical – measurements with larger uncertainty. We shall assume here that the experiments currently underway to measure R or k (Fellmuth, Gaiser and Fischer, 2006) will achieve a relative standard uncertainty by the end of 2010, which is a factor of about two smaller than the current u_r of approximately 2×10^{-6}, so that $u_r(T_{\mathrm{TPW}})$ will be reduced to about 1×10^{-6}, corresponding to about 0.25 mK, and that this will be small enough for the redefinition of the kelvin to be adopted by the 24th CGPM in 2011.

In the discussion about the new definition of the kelvin, it should also be recognized that the "practical" ITS-90 is a defined temperature scale that assigns an exact temperature value T_{90} to each defining fixed point. Hence, the ITS-90 value of the TPW temperature will remain 273.16 K, that is, $T_{\mathrm{TPW-90}} = 273.16$ K exactly. The value and uncertainty of T_{TPW} would only need to be taken into account if for some critical reason one has to know how well the thermodynamic temperature scale is represented by the ITS-90 at a particular temperature or in a particular temperature range. In fact, although the consistency of T_{TPW} as realized by different TPW reference cells can be as low as 50 μK and even less if the isotopic composition of the water used is taken into account (Strouse and Zhao, 2007), the uncertainties of the thermodynamic temperatures of all other defining fixed points, which are the basis for all practical thermometry, are significantly larger. In contrast to other units, the uncertainty of the realization of the kelvin varies greatly with temperature: at 1300 K, for instance, the uncertainty is roughly 100 times greater than at T_{TPW}. Hence, the fact that T_{TPW} will not be exactly known but have a standard uncertainty of 0.25 mK will have negligible practical consequences (Fischer *et al.*, 2007).

To put the new definition of the kelvin into practice, a *mise-en-pratique* has already been recommended to the CIPM by the CCT Recomm T3. The *mise-en-pratique* will allow direct determination of thermodynamic temperatures particularly at temperatures far away from the TPW in parallel to the realization described in the International Temperature Scale. In the high-temperature range, this will considerably reduce the uncertainty of the realization of the kelvin for many purposes for which the need to refer back to the TPW is anomalous, such as radiation thermometry.

In Mills *et al.* (2006), new definitions are suggested for each of the units kilogram, ampere, kelvin, and mole to be chosen by the CGPM in 2011. The following definition for the kelvin was proposed for the first time in Fischer *et al.* (2005b) and could be as simple as "The kelvin is the change of thermodynamic temperature that results in a change of the thermal energy kT by exactly $1.3806504 \cdot 10^{-23}$ joule."

Glossary

Black Body: Radiation source whose emission spectrum depends besides wavelength only on temperature and not on material properties.

Boltzmann Constant: Conversion factor between thermal and mechanical energy.

Calibration: Comparing the reading of an instrument with a standard.

Cryogenic Radiometer: Electrical substitution radiometer operated at cryogenic temperatures.

Emissivity: Ratio of radiance emitted by an object to the radiance emitted by a black body of the same temperature.

Gas Thermometer: Measures temperature by the variation in volume or pressure of a gas. The common apparatus is a constant-volume thermometer CVGT. Consists of a bulb connected by a capillary tube to a manometer.

Noise Thermometer: Measures temperature by the electronic noise generated by the thermal agitation of electrons inside an electrical conductor.

Primary Thermometer: Thermometer whose basic relation between the measurand and temperature can be written down explicitly without having to introduce unknown, temperature-dependent constants.

Radiation Thermometer: Measures temperature by detecting heat radiation.

Resistance Thermometer: Measures temperature by using the dependency of its own resistance on temperature developed across the metal.

Secondary Thermometer: Thermometer that needs calibration against a primary thermometer.

Thermocouple: A device for measuring temperature based on the junction of two dissimilar metals that have a voltage output proportional to the difference between the hot junction and the cold junction (lead wires).

Thermodynamic Temperature: Temperature of a state of equilibrium measured by a primary thermometer.

References

Bedford, R.E., Bonnier, G., Maas, H., and Pavese, F. (1990) *Techniques for Approximating the International Temperature Scale of 1990*, BIPM, Sèvres, Pavillon de Breteuil.

Benedetto, G., Gavioso, R.M., Spagnolo, R., Marcarino, P., and Merlone, A. (2004) *Metrologia*, **41**, 74–98.

Bentley, R.E. (ed.) (1998) *Handbook of Temperature Measurement*, Springer, Singapore.

Benz, S.P., Martinis, J.M., Nam, S.W., Tew, W.L., and White, D.R. (2002), *International Symposium on Temperature and Thermal Measurements in Industry and Science (TEMPMEKO)*, Vol. 8 (eds B. Fellmuth, J. Seidel, and G. Scholz), VDE Verlag GmbH, Berlin, pp. 37–44, isbn 3-8007-2676-9.

Bernhard, F. (ed.) (2004) *Technische Temperaturmessung*, Springer, Berlin, Heidelberg, New York.

Bloembergen, P., Yamada, Y., Yamamoto, N., and Hartmann, J. (2003) *Temperature Its Measurement and Control in Science and Industry*, Vol. 7 (editor-in-chief D.C. Ripple), American Institute of Physics, Melville, pp. 291–296.

Brixy, H., Hecker, R., Oehmen, J., Rittinghaus, K.F., Setiawan, W., and Zimmermann, E. (1992) *Temperature Its Measurement and Control in Science and Industry*, Vol. 6 (ed. J.F. Schooley), American Institute of Physics, New York, pp. 993–996.

Burns, G.W. and Ripple, D.C. (1996) *International Symposium on Temperature and Thermal Measurements in Industry and Science (TEMPMEKO)*, Vol. 6 (ed. P. Marcarino), Levrotto & Bella, Torino, pp. 171–176.

Chandrasekhar, S. (1967), *An Introduction to the Study of Stellar Structure*, Dover, New York, pp. 11–37.

Colclough, A.R. (1984) Gas constant, X-ray interferometry, nuclidic masses, other constants, and uncertainty assignments, in *Precision Measurement and Fundamental Constants II*, NBS Special Publication 617 (eds B.N. Taylor and W.D. Phillips), National Bureau of Standards, Washington, DC, pp. 263–275.

Colclough, A.R., Quinn, T.J., and Chandler T.R.D. (1979) *Proc. R. Soc. London, Ser. A*, **368**, 125–139.

Daussy, C., Guinet, M., Amy-Klein, A., Djerroud, K., Hermier, Y., Briaudeau, S., Bordé, Ch.J., and Chardonnet, C. (2007) *Phys. Rev. Lett.*, **98**, 250801.

Edler, F., Kühne, M., and Tegeler, E. (2004) *Metrologia*, **41**, 47–55.

Edsinger, R.E. and Schooley, J.F. (1989) *Metrologia*, **26**, 95–106.

Engert, J., Fellmuth, B., and Hoffmann, A. (2004) *J. Low Temp. Phys.*, **134**(1–2), 425–430.

Ewing, M.R. and Trusler, J.P.M. (2000) *J. Chem. Thermodyn.*, **32**, 1229–1255.

Fellmuth, B. and Fischer, J. (eds) (2007) *Workshop on Progress in Determining the Boltzmann Constant*, Report PTB-Th-3, PTB, Braunschweig. isbn 978-3-86509-684-5.

Fellmuth, B., Gaiser, Ch., and Fischer, J. (2006) *Meas. Sci. Technol.*, **17**, R145–R159.

Fellmuth, B., Hechtfischer, D., and Hoffmann, A. (2003) *Temperature Its Measurement and Control in Science and Industry*, Vol. 7 (editor-in-chief D.C. Ripple), American Institute of Physics, Melville, pp. 71–76.

Fischer, J. and Fellmuth, B. (2005a) *Rep. Prog. Phys.*, **68**, 1043–1094.

Fischer, J., Fellmuth, B., Seidel, J., and Buck, W. (2005b) *International Symposium on Temperature and Thermal Measurements in Industry and Science (TEMPMEKO)*, Vol. 9, (ed. D. Zvizdic), Laboratory for Process Measurement, Faculty of Mechanical Engineering and Naval Architecture, Zagreb, pp. 12–22.

Fischer, J., Gerasimov, S., Hill, K.D., Machin, G., Moldover, M.R., Pitre, L., Steur, P., Stock, M., Tamura, O., Ugur, H., White, D.R., Yang, I., and Zhang, J. (2007) *Int. J. Thermophys.*, **28**, 1753–1765.

Fischer, J. and Jung, H.J. (1989) *Metrologia*, **26**, 245–252.

Fischer, J., Neuer, G., Schreiber, E., and Thomas, R. (2002), *International Symposium on Temperature and Thermal Measurements in Industry and Science (TEMPMEKO)*, Vol. 8. (eds B. Fellmuth, J. Seidel, and G. Scholz), VDE Verlag GmbH, Berlin, isbn 3-8007-2676-9, pp. 801–806.

Fox, N.P. (2002) *International Symposium on Temperature and Thermal Measurements in Industry and Science (TEMPMEKO)*, Vol. 8, (ed B. Fellmuth, J. Seidel, and G. Scholz), VDE Verlag GmbH, Berlin, pp. 27–35, isbn 3-8007-2676-9.

Gaiser, C., Fellmuth, B., and Haft, N. (2008) *Int. J. Thermophys.*, **29**, 18–30.

Griem, H.R. (1964) *Plasma Spectroscopy*, McGraw-Hill, New York.

Guildner, L.A. and Edsinger, R.E. (1976) *J. Res. Natl. Bur. Stand. Sect., A*, **80**, 703–738.

Hartmann, J., Taubert, D.R., and Fischer, J. (2002) *International Symposium on Temperature and Thermal Measurements in Industry and Science (TEMPMEKO)*, Vol. 8, (eds B. Fellmuth, J. Seidel, and G. Scholz), VDE Verlag GmbH, Berlin, isbn 3-8007-2676-9, pp. 377–382.

Hollandt, J., Friedrich, R., Gutschwager, B., Taubert, D.R., and Hartmann, J. (2003/2004) *High Temp. – High Press.*, **35/36**, 379–415.

International Electrotechnical Commission (1989) Thermocouples – Part 2 Tolerances IEC 60584-2, Genève.

International Electrotechnical Commission (1995a) Industrial Platinum Resistance Thermometer Sensors IEC 60751, Genève.

International Electrotechnical Commission (1995b) Thermocouples – Part 1 Reference tables IEC 60584-1, Genève.

Jung, H.J. (1984) *Metrologia*, **20**, 67–69.

Jung, H.J. (1986) *Metrologia*, **23**, 19–31.

Labenski, J.R., Tew, W.L., Benz, S.P., Nam, S.W., and Dresselhaus, P. (2008) *Int. J. Thermophys.*, **29**, 1–17.

Łach, G., Jeziorski, B., and Szalewicz, K. (2004) *Phys. Rev. Lett.*, **92**, 233001–1–4.

Luther, H., Grohmann, K., and Fellmuth, B. (1996) *Metrologia*, **33**, 341–352.

Machin, G., Bloembergen, P., Hartmann, J., Sadli, M., and Yamada, Y. (2007) *Int. J. Thermophys.*, **28**, 1976–1982.

Martin, J.E. and Haycocks, P.R. (1998) *Metrologia*, **35**, 229–233.

Martin, J.E., Quinn, T.J., and Chu, B. (1988) *Metrologia*, **25**, 107–112.

Massalski, B.T. (ed.) (1990) *Binary Alloy Phase Diagrams*, Vol. 1, ASM Int, Materials Park, Ohio.

Mills, I.M., Mohr, P.J., Quinn, T.J., Taylor, B.N., and Williams, E.R. (2006) *Metrologia*, **43**, 227–246.

Mohr, P.J., Taylor, B.N., and Newell, D.B. (2008) *Rev. Mod. Phys.*, **80**, 633–730.

Moldover, M.R. (1998) *J. Res. NIST*, **103**, 167–175.

Moldover, M.R., Boyes, S.J., Meyer, C.W., and Goodwin, A.R.H. (1999) *J. Res. Natl. Inst. Stand. Technol.*, **104**, 11–46.

Moldover, M.R., Trusler, J.P.M., Edwards, T.J., Mehl, T.J., and Davis, R.S. (1988) *J. Res. Natl. Bur. Stand.*, **93**, 85–114.

Nam, S.W., Benz, S.P., Martinis, J.M., Dresselhaus, P., Tew, W.L., and White, D.R. (2003) *Temperature Its Measurement and Control in Science and Industry*, Vol. 7, (editor-in-chief D.C. Ripple), American Institute of Physics, Melville, NY, pp. 37–42.

Nicholas, J.V. (1999), *International Symposium on Temperature and Thermal Measurements in Industry and Science (TEMPMEKO)*, Vol. 7, (eds J.F. Dubbeldam and M.J. deGroot), IMEKO / NMi Van Swinden Laboratorium, Delft, pp. 100–105.

Noulkhow, N., Taubert, D., Meindl, P., and Hollandt, J. (2009) *Int. J. Thermophys.*, **30**, 134–143.

Nyquist, H. (1928) *Phys. Rev.*, **32**, 110–113.

Pavese, F. and Molinar, G. (1992) *Modern Gas-Based Temperature and Pressure Measurements*, Plenum Press, New York and London.

Peruzzi, A. and de Groot, M. (2003) *Proceedings of the 2nd International Seminar and Workshop on Low Temperature Thermometry* (eds A. Szmyrka-Grzebyk and A. Kowal), Institute of Low Temperature and Structure Research Polish Academy of Sciences, Wrocław, isbn 83-906218-5-1, pp. 35–40.

Pitre, L., Moldover, M.R., and Tew, W.L. (2006) *Metrologia*, **43**, 142–162.

Planck, M. (1921) *Theorie derWärmestrahlung*, 4th ed, Leipzig, Barth.

Preston-Thomas, H. (1990a) *Metrologia*, **27**, 3–10, 107.

Preston-Thomas, H., Bloembergen, P., and Quinn, T.J. (1990b) *Supplementary Information for the International Temperature Scale of 1990*, BIPM, Sèvres, Pavillon de Breteuil.

Quinn, T.J. (1990) *Temperature 2nd edition Monographs in Physical Measurement*, Academic Press, London, San Diego, New York, Boston, Toronto, Sydney, Tokyo.

Quinn, T.J. and Martin, J.E. (1985) *Philos. Trans. R. Soc. London, A*, **316**, 85–189.

CIPM (2005) Recommendation T 3 to the CIPM: creation of a *mise en pratique* of the definition of the kelvin. *BIPM Com. Cons. Thermométrie*, **23**, (Document CCT/05-32).

Ripple, D.C., Strouse, G.F., and Moldover, M.R. (2007) *Int. J. Thermophys.*, **28**, 1789–1799.

Rusby, R.L., Hudson, R.P., Durieux, M., Schooley, J.F., Steur, P.P.M., and Swenson, C.A. (1991) *Metrologia*, **28**, 9–18.

Rusby, R.L., Durieux, M., Reesink, A.L., Hudson, R.P., Schuster, G., Kühne, M., Fogle, W.E., Soulen, R.J., and Adams, E.D. (2002) *J. Low Temp. Phys.*, **126**, 633–642.

Rusby, R.L., Fellmuth, B., Engert, J., Fogle, W.E., Adams, E.D., Pitre, L., and Durieux, M. (2007) *J. Low Temp. Phys.*, **149**, 156–175.

Schuster, G., Hoffmann, A., and Hechtfischer, D. (2001) *Realisation of the Temperature Scale PLTS-2000 at PTB*, PTB, Braunschweig. PTB-ThEx-21, isbn 3-89701-742-3.

Steinhart, J.S. and Hart, S.R. (1968) *Deep Sea Res.*, **15**, 497–503.

Stock, M., Fischer, J., Friedrich, R., Jung, H.J., Werner, L., and Wende, B. (1996), *International Symposium on Temperature and Thermal Measurements in Industry and Science (TEMPMEKO)*, Vol., 6 (ed. P. Marcarino), Levrotto & Bella, Torino, pp. 19–24.

Strouse, G.F., Defibaugh, D.R., Moldover, M.R., and Ripple, D.C. (2003) *Temperature Its Measurement and Control in Science and Industry*, Vol. 7, (editor-in-chief D.C. Ripple), American Institute of Physics, Melville, NY, pp. 31–36.

Strouse, G.F. and Zhao, M. (2007) *Int. J. Thermophys.*, **28**, 1913–1922.

Taubert, D.R., Hartmann, J., Hollandt, J., and Fischer J. (2003) *Temperature Its Measurement and Control in Science and*

Industry, Vol. 7, (editor-in-chief D.C. Ripple), American Institute of Physics, Melville, NY, pp. 7–12.

White, D.R., Mason, R.S., and Saunders, P. (2002) *International Symposium on Temperature and Thermal Measurements in Industry and Science (TEMPMEKO)*, Vol. 8, (eds B. Fellmuth, J. Seidel, and G. Scholz), VDE Verlag GmbH, Berlin, isbn 3-8007-2676-9, pp. 129–134.

White, D.R., Ballico, M., del Campo, D., Duris, S., Filipe, E., Ivanova, A., Kartal Dogan, A., Mendez-Lango, E., Meyer, C.W., Pavese, F., Peruzzi, A., Renaot, E., Rudtsch, S., and Yamazawa, K. (2007) *Int. J. Thermophys.*, **28**, 1868–1881.

Woolliams, E., Machin, G., Lowe, D., and Winkler, R. (2006) *Metrologia*, **43**, R11–R25.

Yamada, Y., Sakate, H., Sakuma, F., and Ono, A. (1999) *Metrologia*, **36**, 207–209.

Yoon, H.W., Gibson, C.E., Allen, D.W., Saunders, R.D., Litorja, M., Brown, S.W., Eppeldauer G.P., and Lykke K.R. (2005), *International Symposium on Temperature and Thermal Measurements in Industry and Science (TEMPMEKO)*, Vol. 9, (ed. D. Zvizdic), Laboratory for Process Measurement, Faculty of Mechanical Engineering and Naval Architecture, Zagreb, pp. 59–70.

Further Reading

Bentley, R.E. (ed.) (1998) *Handbook of Temperature Measurement*, Springer, Singapore.

Bernhard, F. (ed.) (2004) *Technische Temperaturmessung*, Springer, Berlin, Heidelberg, New York.

Michalski, L., Eckersdorf, K., Kucharski, J., and McGhee, J. (2001) *Temperature Measurement*, 2nd edn, John Wiley & Sons Ltd, Chichester, New York, Weinheim, Brisbane, Singapore, Toronto.

Nicholas, J.V. and White, D.R. (2001) *Traceable Temperatures. An Introduction to Temperature Measurement and Calibration*, 2nd edn, John Wiley & Sons Ltd, Chichester, New York, Weinheim, Brisbane, Singapore, Toronto.

Preston-Thomas, H., Bloembergen, P., and Quinn, T.J. (1990b) *Supplementary Information for the International Temperature Scale of 1990*, BIPM, Sèvres, Pavillon de Breteuil.

Quinn, T.J. (1990) *Temperature, Monographs in Physical Measurement*, 2nd edn, Academic Press, London, San Diego, New York, Boston, Toronto, Sydney, Tokyo.

Introduction to Radiation Detectors

Claus Grupen

1
Introduction

The development of particle detectors practically starts with the discovery of radioactivity by Henri Becquerel in the year 1896. He noticed that the radiation emanating from uranium salts could blacken photosensitive paper. Almost at the same time (1895), X rays, which originated from materials after the bombardment by energetic electrons, were discovered by Wilhelm Conrad Röntgen.

The first nuclear-particle detectors were thus extremely simple. Also, the zinc-sulfide scintillators in use at the beginning of the twentieth century were very primitive. They were used, for example, for the study of scattering of α particles from nuclei. This technique required tedious and tiresome optical registration of scintillation light with the human eye.

Scintillations in the form of ‘‘northern lights’’ (aurora borealis) had already been observed since long. This fluorescence is due to charged solar particles entering the Earth’s atmosphere along the magnetic field lines at the geomagnetic poles causing excitations of nitrogen and oxygen atoms. Also, already about 50 years before the discovery of Cherenkov radiation, Heaviside (1892) showed that charged particles, moving faster than light in a transparent medium with refractive index n, emit an electromagnetic radiation at a certain angle with respect to the particle direction. Madame Curie, too, noticed a faint light emitted from concentrated solutions of radium in water, thereby unknowingly operating the first Cherenkov detector. The human eye can also act as Cherenkov detector. This has been shown by light flashes experienced by astronauts with eyes closed during their space missions. These light emissions are caused by energetic primary cosmic rays passing through the vitreous body of the eye.

In the course of time, the measurement methods have been greatly refined. Nowadays, it is generally insufficient to only detect particles and radiation. One wants to identify their nature, that is, one would like to know whether one is dealing, for example, with electrons, muons, pions, or energetic γ rays. In addition, accurate energy and momentum measurements are often required. For the majority of applications, an exact knowledge of the spatial coordinates of particle trajectories is of interest.

digital Encyclopedia of Applied Physics.
Copyright © Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim All rights reserved.

The trend of particle detection has shifted in the course of time from optical measurement to purely electronic means. In this development ever higher resolutions, for example, of time, spatial reconstruction, and energy resolutions, have been achieved. Early optical detectors, like cloud chambers, only allowed rates of one event per minute, while modern devices, like fast organic scintillators, can process data rates in the gigahertz regime. With gigahertz rates, new problems arise and questions of radiation hardness and aging of detectors become an issue.

Originally, particle detectors were used in cosmic rays and nuclear and particle physics. Meanwhile, these devices have found applications in medicine, biology, environmental science, metrology, radiation protection, oil exploration, civil engineering, archaeology, and arts, to name a few. While the most sophisticated detectors are still developed for particle physics and astroparticles, practical applications often require robust devices, which also function in harsh environments.

Radiation detectors have contributed significantly to the advancement of science. New detection techniques, like, for example, cloud chambers, bubble chambers, multiwire proportional and drift chambers, and micropattern detectors allowed essential discoveries. The development of new techniques in this field was also recognized by the award of a number of Nobel prizes (Charles T.R. Wilson, cloud chamber, 1927; Pavel A. Cherenkov, Ilja M. Frank, Igor J. Tamm, Cherenkov effect, 1958; Donald A. Glaser, bubble chamber, 1960; Luis W. Alvarez, bubble-chamber analysis, 1968; Georges Charpak, multiwire proportional chamber, 1992; Raymond Davis Jr., Masatoshi Koshiba, Riccardo Giacconi, novel detection techniques for neutrinos and X rays, 2002).

In this chapter, only a brief overview over the main radiation detectors is given. Within the scope of this chapter, the basic interaction mechanisms, which form the basis of radiation detectors, are only sketched.

2
Historical Detectors

Historical radiation detectors are mainly optical devices that have been used in the early days of cosmic rays and particle physics (Galison, 1997). Even though some of these detectors have been "recycled" for recent elementary particle physics experiments, like nuclear emulsions for the discovery of the τ neutrino (2000) or bubble chambers with holographic readout for the measurement of short-lived hadrons, these optical devices are nowadays mainly integrated into demonstration experiments in exhibitions or employed as eye-catchers in lobbies of physics institutes.

2.1
Cloud Chambers

The expansion cloud chamber ("Wilson chamber") is one of the oldest detectors for track and ionization measurement. The positron (1932) and the muon (1937) have been discovered in a cloud-chamber experiment. A cloud chamber is a container filled with a gas–vapor mixture (e.g., air–water vapor or argon–alcohol) at the vapor saturation pressure. If a charged particle traverses the cloud chamber, it produces

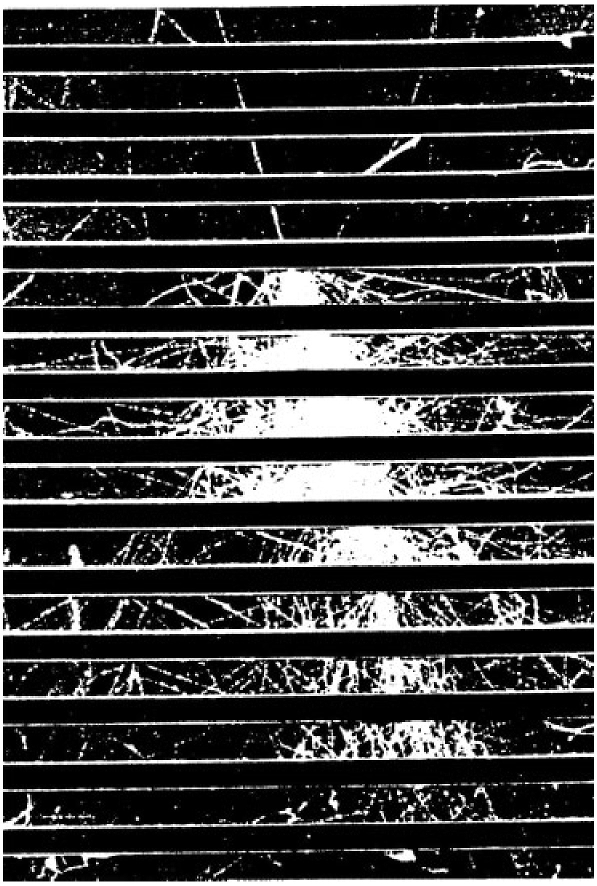

Fig. 1 Example of an electromagnetic cascade in a multiplate cloud chamber; Rochester and Turver, 1981. Private communication by G.D. Rochester; photo credit George D. Rochester.

an ionization trail. The lifetime of positive ions produced in the ionization process in the chamber gas is relatively long (approximately a few milliseconds). Therefore, after the passage of the particle a trigger signal, for example, can be derived from a coincidence of scintillation counters, which initiates a fast expansion of the chamber. By means of adiabatic expansion, the temperature of the gas mixture is lowered and the vapor gets supersaturated. It condenses on nuclei, which are represented by the positive ions yielding droplets marking the particle trajectory. The track consisting of droplets is illuminated and photographed (Figure 1).

In contrast to the expansion cloud chamber, a diffusion cloud chamber is permanently sensitive. The chamber is, like the expansion cloud chamber, filled with a gas–vapor mixture. A constant temperature gradient provides a region where the vapor is in a permanently supersaturated state. Charged particles entering this region produce a trail automatically without any additional trigger requirement. Zone widths (i.e., regions in which trails can form) with supersaturated vapor of 5–10 cm can be obtained. A clearing field removes the positive ions from the chamber.

2.2
Bubble Chambers

The bubble chamber allows the recording and reconstruction of events of high complexity with high spatial resolution. Therefore, it is perfectly suited to study rare events (e.g., neutrino interactions). In a bubble chamber, the liquid gas (hydrogen, deuterium, neon, or Freon, etc.) is held in a pressure container close to the boiling point. Before the expected event, the chamber volume is expanded by retracting a piston. The expansion of the chamber leads to a reduction in pressure, thereby exceeding the boiling temperature of the bubble-chamber liquid. If in this superheated liquid state a charged particle enters the chamber, bubble formation sets in along the particle track. The positive ions produced by the incident particles act as nuclei for bubble formation. The lifetime of these nuclei is too short to trigger the expansion of the chamber by the incoming particles. For this reason, the superheated state has to be reached before the arrival time of the particles. Therefore bubble chambers can only be used at accelerators, where the arrival time of particles in the detector is known and, therefore, the chamber can be expanded in time.

In the superheated state, the bubbles grow until the growth is stopped by a termination of the expansion. At this moment, the bubbles are illuminated by light flashes and photographed (Figure 2). The inner walls of the container have to be extremely smooth so that the liquid ''boils'' only in those places where bubble formation should occur, namely, along the particle trajectory, and not on the chamber walls. In order to measure short lifetimes in bubble chambers, the size of the bubbles must be limited. This means that

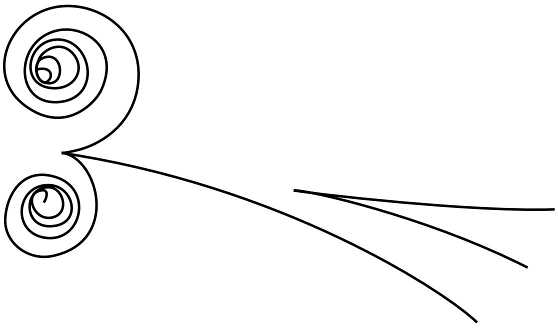

Fig. 2 This event shows the creation of an e^+e^- pair by a photon in the Coulomb field of an electron and an e^+e^- pair created in the field of a nucleus. In the first case, the target electron gets a substantial recoil and leaves a high-momentum track; Close and Marten, 1987. Photo credit Frank Close 2007.

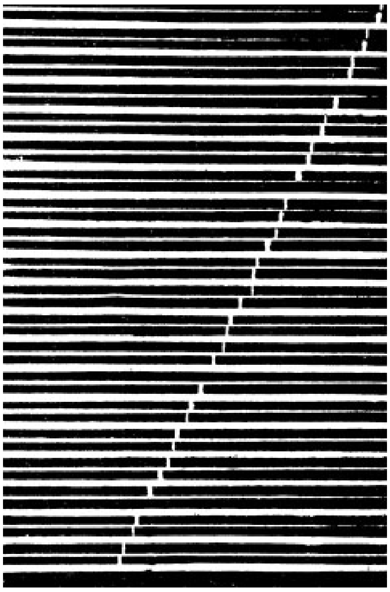

Fig. 3 Track of a cosmic-ray muon in a multiplate spark chamber; Kaftanov and Liubimov, 1963. Private communication by V.S. Kaftanov; photo credit Vitali S. Kaftanov.

the event under investigation must be photographed relatively soon after the onset of bubble formation when the bubble size is relatively small, thereby guaranteeing a good spatial resolution and, as a consequence, also good lifetime resolution for unstable particles. In any case, the bubble size must be small compared to the decay length of the particle.

2.3
Spark Chambers

In a spark chamber, a number of parallel plates are mounted in a gas-filled volume. Typically, a mixture of helium and neon is used as counting gas. Alternately, the plates are either grounded or connected to a high-voltage supply. The high-voltage pulse is normally triggered to every second electrode by a coincidence between two scintillation counters placed above and below the spark chamber. The gas amplification is chosen in such a way that a spark discharge occurs at the point of the passage of the particle (Figure 3). This is obtained for gas amplifications between 10^8 and 10^9. For lower gas amplifications sparks will not develop, while for larger gas amplifications sparking at unwanted positions (e.g., at spacers that separate the plates) can occur. The discharge channel essentially follows the electric field. Up to an angle of 30°, the conducting plasma channel can, however, follow the particle trajectory (track spark chamber). Between two discharges, the ions produced are removed from the detector volume by means of a clearing field. If the time delay between the passage of the particle and the high-voltage signal is less than the memory time of about 10 microseconds, the efficiency of the spark chamber is close to 100%. A clearing field, of course, removes also the primary ionization from the detector volume. For this reason, the time delay between the passage of the particle and the application of the high-voltage signal has to be chosen as short as possible to reach full efficiency. Also, the rise time of the high-voltage pulse must be short because otherwise the leading edge acts as a clearing field before the critical field strength for spark formation is reached.

2.4
Nuclear Emulsions

Tracks of charged particles in nuclear emulsions can be recorded by the photographic method. Nuclear emulsions consist of fine-grained silver-halide crystals (AgBr and AgCl), which are embedded in a gelatin substrate. A charged particle produces a latent image in the emulsion. Owing to the free charge carriers liberated in the ionization process, some halide molecules are reduced to metallic silver in the emulsion. In the subsequent development process, the silver-halide crystals are chemically reduced. This preferentially affects those microcrystals (nuclei) that are already disturbed and partly reduced. These are transformed into elemental silver. The process of fixation dissolves the remaining silver halide and removes it. Therefore the charge image, which has been transformed into elemental silver particles, remains stable.

The evaluation of the emulsion is usually done under a microscope, but it can also be performed by using a charge-coupled-device (CCD) camera and a semiautomatic pattern-recognition device. Fully automated emulsion-analysis systems have also been developed. The sensitivity of the nuclear emulsion must be high enough so that the energy loss of minimum-ionizing particles is sufficient to produce individual silver-halide microcrystals along the track of a particle. Commercially available photoemulsions do not have this property. Furthermore, the silver grains that form the track and also the silver-halide microcrystals must be sufficiently small to enable a high spatial resolution. The requirements of high sensitivity and low grain size are in conflict and, therefore, demand a compromise. In most nuclear emulsions, the silver grains have a size of 0.1–0.2 μm and so are much smaller than in commercial films (1–10 μm). The mass fraction of the silver halide (mostly AgBr) in the emulsion amounts to approximately 80%. Nuclear emulsions allow high spatial resolution of complex events with large multiplicities. They are permanently sensitive and cannot be triggered.

2.5
Track-Etch Detectors

Particles of high electric charge destroy the local structure in a solid along their tracks. This local destruction can be intensified by etching and thereby made visible. Solids such as inorganic crystals, glasses, plastics, minerals, or even metals can be used for this purpose. The damaged parts of the material react with the etching agent more intensively than the undamaged material and characteristic etch cones will be formed. If the etching process is not interrupted, the etch cones starting from the surface will merge and form a hole at the point of the particle track. The etching procedure will also remove some part of the surface material.

Track-etch detectors show a threshold effect: the minimum radiation damage caused by protons and α particles is frequently insufficient to produce etchable tracks. The detection and measurement of heavy ions, for example, in primary cosmic rays ($Z \geq 3$), will consequently not be disturbed by a high background of protons and α particles. The size of the etch cones (for a fixed etching time) is a measure of the energy loss of the particles. Therefore, if the velocity of the particles is known, it allows a determination of the charge of the nuclei. A stack of track-etch detectors, flown in a balloon at a residual atmosphere of several grams per square centimeter, thus permits a determination of the elemental abundance in primary cosmic rays. Specially prepared plastic materials can also be made sensitive to α particles and neutrons. Such detectors are used in the field of radiation protection for radon monitoring and neutron measurement.

3
Gas Detectors

3.1
Bethe–Bloch Energy-Loss Formula

Particles and radiation cannot be detected directly, but rather only through their interactions with matter. There are specific interactions for charged particles that are different from those of neutral particles, for example, of photons. The variety of interactions is quite rich and, as a consequence, a large number of detection devices for particles and radiation exist. The most important energy-loss mechanism for charged particles in gas detectors is the ionization loss, which can be described by the Bethe–Bloch formula, which represents the energy loss dE per length dx. It is given by Yao, 2006

$$-\frac{dE}{dx} = 4\pi N_A r_e^2 m_e c^2 z^2 \frac{Z}{A} \frac{1}{\beta^2} \left(\ln \frac{2 m_e c^2 \gamma^2 \beta^2}{I} - \beta^2 - \frac{\delta}{2} \right) \tag{1}$$

where z is the charge of the incident particle in units of the elementary charge;
Z, A are the atomic number and atomic weight of the absorber;
m_e is the electron mass;
r_e is the classical electron radius;
N_A is the Avogadro number (= number of atoms per mole);
I is the ionization constant, characteristic of the absorber material;
β is the velocity of the particle v normalized to the velocity of light c, $\beta = v/c$;
γ is the Lorentz factor, $\gamma = 1/\sqrt{1-\beta^2}$;
δ is a parameter that describes how much the extended transverse electric field of incident relativistic particles is screened by the charge density of the atomic electrons. In this way, the energy loss is reduced (“density effect,” “Fermi plateau” of the energy loss).

A useful constant appearing in the Bethe–Bloch equation is

$$4\pi N_A r_e^2 m_e c^2 = 0.3071 \, \frac{\text{MeV}}{\text{g/cm}^2} \tag{2}$$

In the low-energy domain, the energy loss decreases like $1/\beta^2$ and reaches a broad minimum of ionization near $\beta\gamma \approx 4$. Relativistic particles ($\beta \approx 1$), which have an energy loss corresponding to this minimum, are called *minimum-ionizing particles*. In light absorber materials, where the ratio $Z/A \approx 0.5$, the energy loss of minimum-ionizing particles can be roughly represented by

$$-\left.\frac{dE}{dx}\right|_{\text{min}} \approx 2 \frac{\text{MeV}}{\text{g/cm}^2} \tag{3}$$

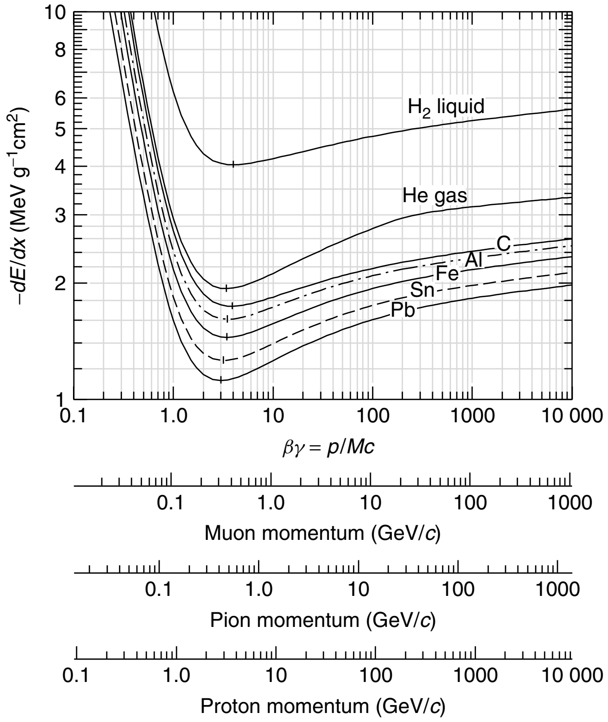

Fig. 4 Mean energy loss in liquid hydrogen, gaseous helium, carbon, aluminum, iron, tin, and lead. p is the momentum of the particle, M its mass, and c is the velocity of light; Yao, 2006.

The energy loss increases again for $\beta\gamma > 4$ ("logarithmic rise" or "relativistic rise"). The increase approximately follows a dependence like $\ln \gamma$ (see Figure 4). For thin absorbers, the ionization energy loss is subject to asymmetrical Landau fluctuations.

3.2
Ionization Yield

An interesting number for gas detectors is the ionization yield. One must distinguish between primary ionization, that is, the number of primary electron–ion pairs produced, and the total ionization. A sufficiently large amount of energy can be transferred to some primary electrons so that they also can ionize (knock-on electrons or δ rays). This secondary ionization together with the primary ionization forms the total ionization. The average energy required to form an electron–ion pair (W value) exceeds the ionization potential of the gas because inner shells of the gas atoms can also be involved in the ionization process, and a fraction of the energy of the incident particle can be dissipated by excitation processes, which do not lead to free electrons. The W value of a material is constant for relativistic particles and increases only slightly for low velocities of incident particles.

For gases, the W values are around 30 eV. They can, however, strongly depend on impurities in the gas. This number has to be compared to the average energy required for the production of an electron–hole pair in solid-state radiation detectors, which is only around 3.5 eV. Therefore, the statistical fluctuations in the number of charge carriers produced for a given energy loss is much smaller in solid-state detectors than in gaseous detectors.

3.3
Ionization Chambers

An ionization chamber is a gaseous detector that measures the amount of ionization produced by a charged particle passing through the gas volume. Neutral particles can also be detected by this device via secondary charged particles resulting from the interaction of the primary ones with electrons or nuclei. Charged particles are measured by separating the charge-carrier pairs produced by their ionization in an electric field and guiding the ionization products to the anode or cathode, respectively, where corresponding signals can be recorded. If a particle is totally absorbed in an ionization chamber, such a detector type measures its energy.

In the simplest case, an ionization chamber consists of a pair of parallel electrodes mounted in a gastight container that is filled with a gas mixture, which allows electron and ion drift. Charged particles produce electron–ion pairs along their track. The number of created charge carriers depends on the particle type and energy and on the gas.

The parallel electrodes of the ionization chamber producing a homogeneous electric field are acting as a capacitor, which is initially charged to a certain voltage. The drifting charge carriers induce an electric charge on the electrodes, which leads to certain change of the voltage, ΔU. The signal amplitude ΔU has contributions from fast-moving electrons and the slowly drifting ions. The disadvantage is that the output signal depends on where the particles enter the ionization chamber. This can be solved by introducing a Frisch grid, permitting the particles to enter the chamber only in a well-defined region.

Ionization chambers are mainly used in the spectroscopy of particles of higher charge, because, in this case, the deposited energies are, in general, larger (typically 5 MeV for α particles from radioactive sources) compared to those of singly charged minimum-ionizing particles. A minimum-ionizing particle passing a layer of 4 cm of argon will deposit only about 11 keV, which provides about 400 pairs. To detect such a small signal is a very difficult task.

Apart from planar ionization counters, cylindrical ionization counters are also in use. Because of the cylindrical arrangement of the electrodes the electric field in this case is no longer constant, but increases inversely proportional to the distance to the anode, which is usually a thin wire. The important point with cylindrical ionization chambers is that no multiplication of charge carriers occurs in the gas. Only the originally produced number of electrons and ions is collected.

3.4
Proportional Counters

In contrast to ionization chambers, the high voltage in proportional chambers is increased so that the initially produced charge carriers gain enough energy between two collisions to produce secondary electrons. In this way, an avalanche of secondary and tertiary charge carriers is started.

If N_0 primary electrons are produced, the number of particles $N(x)$ at the point x is calculated to be

$$N(x) = N_0 e^{\alpha x} \tag{4}$$

The first Townsend coefficient α depends on the field strength $\vec{E}$ and the pressure in the gas counter. The proportional range of a counter is characterized by the fact that the gas-amplification factor takes a constant value. As a consequence, the measured signal is proportional to the ionization produced. Gas-amplification factors of up to 10^6 are possible in the proportional mode. Typical gas amplifications are rather in the range between 10^4 and 10^5. At high field, collisions of electrons with atoms or molecules can cause not only ionization but also excitation. De-excitation is often followed by photon emission. However, photons produced in the course of the avalanche development are of no importance in the proportional regime.

Since the output signal in a proportional counter is proportional to the energy loss of the particle (or to the energy, if the particle stops in the counter), these radiation detectors can be used for spectroscopy, for example, of X rays or α particles.

If many anode wires are stretched in a planar chamber in parallel, one arrives at a multiwire proportional chamber (MWPC). The working principle of an MWPC is the same as that for a proportional counter. In addition to energy-proportional signals, MWPCs provide accurate track information if the wires are stretched with relatively small separations. As a contamination monitor, they can also be used in the field of radiation protection to allow the measurement of possible contaminations over larger areas.

3.5
Geiger–Müller Counters

The increase of the field strength in a proportional counter leads to a copious production of photons during the avalanche formation. As a consequence, the probability to produce further new electrons by the photoelectric effect increases. This photoelectric effect can also occur at points distant from the production of the primary avalanche. These electrons liberated by the photoelectric effect will initiate new avalanches, whereby the discharge will propagate along the anode wire, in contrast to the proportional chamber, where the discharge is localized to the point of particle passage.

The discharge is normally stopped by "quenching." In self-quenching counters, a quench gas is admixed to the counting gas, which is a noble gas in most cases. Hydrocarbons like methane (CH_4), ethane (C_2H_6), isobutane (iC_4H_{10}), alcohols like ethyl alcohol (C_2H_5OH) or methylal ($CH_2(OCH_3)_2$), or halides, like ethylbromide, are suitable as quenchers. These additions will absorb photons in the ultraviolet (UV) range (wavelength 100–200 nm), thereby reducing their range to a few wire radii (≈100 µm). The transverse propagation of the discharge proceeds only along and in the vicinity of the anode wire because of the short range of the photons. The photons have no chance to liberate electrons from the cathode by the photoelectric effect because they will be absorbed before they can reach the cathode. After a flux tube of positive ions has been formed along the anode wire, the external field is reduced by this space charge by such an amount that the avalanche development comes to an end. The positive ions drifting in the direction of the cathode will collide on their way with quench-gas molecules, thereby becoming neutralized. The molecule ions, however, have insufficient energy to liberate electrons from the cathode upon impact. Consequently, the discharge stops by itself.

3.6
Streamer Tubes

In Geiger counters, the fraction of counting gas to quenching gas is typically 90 : 10. The anode wires have diameters of 30 μm and the anode voltage is around 1 kV. If the fraction of the quenching gas is considerably increased, the lateral propagation of the discharge along the anode wire can be completely suppressed. One again obtains, as in the proportional counter, a localized discharge with the advantage of large signals (gas amplification $\geq 10^{10}$ for sufficiently high anode voltages), which can be processed without any additional preamplifiers. These streamer tubes, sometimes also called *Iarocci tubes*, are operated with "thick" anode wires between 50 μm and 100 μm. Gas mixtures with $\leq$60% argon and $\geq$40% isobutane can be used. Streamer tubes operated with pure isobutane also proved to function well.

3.7
Drift Chambers

MWPCs allow only modest spatial resolutions. For a wire spacing of only 2 mm resolutions around 500 μm are achievable. The number of anode wires required to cover large areas with MWPCs becomes prohibitive. The use of the drift chamber provides a way out of this problem. The time Δt between the moment of the particle passage through such a chamber and the arrival time of the charge cloud at the anode wire depends on the point of passage of the particle through the chamber. If v^- is the constant drift velocity of the electrons, a linear relation between the drift time and the distance to the anode wire holds. The precise measurement of this drift time allows to determine the spatial coordinate with high accuracy. The resolution is only limited by longitudinal diffusion and by primary ion statistics. Another advantage of drift chambers is that the anode-wire spacing can be much larger than in MWPCs, typically several centimeters. With an appropriate time readout, spatial resolutions under 100 μm are achievable. The drift velocity depends on the gas mixture, with typical values around several centimeters per microsecond. The consequence of this is that this detector does not permit high repetition rates. The drift chamber cannot distinguish between particles passing to the right or to the left of the readout wire. This left–right ambiguity can easily be resolved by using a double layer of chambers with staggered anode wires. In order to produce a suitable drift field, potential wires are introduced between neighboring anode wires.

Drift chambers can be made planar to cover large areas, such as muon chambers for fixed-target and collider experiments, but as vertex detectors in e^+e^- colliders often cylindrical arrangements are in use.

3.8
Time-Projection Chambers (TPCs)

The *crème de la crème* of track recording in cylindrical and planar gas detectors at the moment is realized with the time-projection chamber (TPC). Apart from the counting gas, this detector contains no other constructional elements and thereby represents the optimum as far as minimizing multiple scattering and photon conversions are concerned. The TPC is usually divided into two halves by means of a central electrode. A typical counting gas is a mixture of argon and methane (90 : 10).

The ionization produced by charged particles drifts in the electric field, in the direction of the endfaces of the chamber, which consist of MWPCs in most cases. For cylindrical chambers, the electric field is usually supplemented by a parallel magnetic field. The magnetic field suppresses the diffusion perpendicular to the field. This is achieved by the action of the magnetic forces on the drifting electrons, which, as a consequence, spiral around the direction of the magnetic field. For typical values of electric and magnetic field strengths, Larmor radii below 1 μm are obtained. The MWPCs at the endfaces determine the projected coordinates x and y. The arrival time of the ionization electrons at the endplates supplies the z coordinate along the cylinder axis. Therefore, the TPC is a truly three-dimensional imaging device (see Figure 5).

The gas amplification of the drifting ionization takes place at the anode wires of the MWPC, which are mostly stretched in azimuthal direction. To obtain accurate three-dimensional coordinates, the cathodes of the endcap-MWPC segments are usually structured as pads. The numbers of the fired wire and pad provide the coordinates x and y.

In addition, the analog signals on the anode wires give information on the specific energy loss and can consequently be used for particle identification. Typical values of the magnetic field are around 1.5 T and around 20 kV m^{-1} for the electric field. Since in most constructions

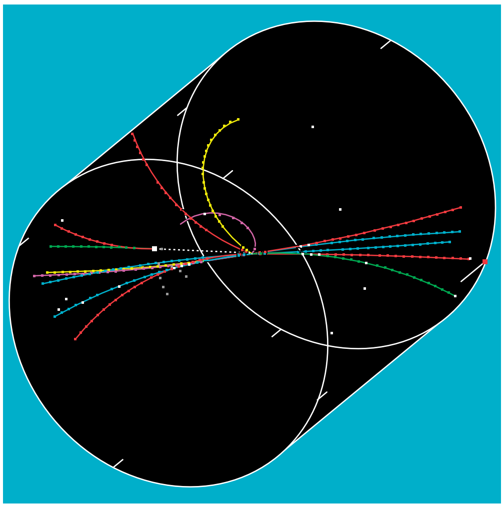

Fig. 5 Two-jet event from an electron–positron annihilation into hadrons as seen in the ALEPH time-projection chamber at the large electron–positron storage ring (LEP) at CERN; ALEPH Collaboration, 2000. Photo Credit ALEPH Collaboration; Roberto Tenchini.

electric and magnetic field are parallel, the Lorentz angle is zero and the electrons drift parallel to $\vec{E}$ and $\vec{B}$ (there is no "$\vec{E} \times \vec{B}$ effect").

However, a problem is caused by the large number of positive ions, which are produced in the gas-amplification process at the endplates and which have to drift a long way back to the central electrode. The strong space charge of the drifting positive ions would cause the field quality to deteriorate. This can be overcome by introducing an additional grid ("gate") between the drift volume and the endcap MWPC. The gate is normally closed. It is only opened for a short period of time if an external trigger signals an interesting event. In the closed state, the gate prevents ions from drifting back into the drift volume. Therefore, the quality of the electric field in the sensitive detector volume remains unchanged. This means that the gate serves a dual purpose. On the one hand, electrons from the drift volume can be prevented from entering the gas-amplification region of the endcap MWPC, if there is no trigger that would signal an interesting event. On the other hand – for interesting gas-amplified events – the positive ions are prevented from drifting back into the detector volume. TPCs can be made very large (diameter ≥ 3 m, length ≥ 5 m). They contain a large number of analog readout channels (number of anode wires $\approx$5000 and cathode pads $\approx$50 000). Several hundred energy-loss samples can be obtained per track, which ensures an excellent determination of the radius of curvature and allows an accurate measurement of the energy loss, which is essential for particle identification. The drawback of the TPC is that high particle rates cannot be handled, because the drift time of the electrons in the detector volume amounts to 40 microseconds (for a drift path of 2 m) and the readout of the analog information also requires several microseconds.

3.9
Micropattern Gaseous Detectors

The construction of MWPCs would be simplified and their stability and flexibility would be greatly enhanced if anodes were made in the form of strips or dots on insulating or semiconducting surfaces instead of stretching anode wires in the counter volume. The rate capability would improve by more than one order of magnitude for these devices because of the short drift times. At present, the class of micropattern gaseous detectors is already rather wide and many new promising devices are under study.

These microstrip gas chamber (MSGC) detectors are miniaturized MWPCs, in which the dimensions are reduced by about a factor of 10 in comparison to conventional chambers. The typical pitch is 100–200 μm and the gas gap ranges from 2 to 10 mm. This has been made possible because the electrode structures can be reduced with the help of electron lithography. The wires are replaced by strips that are evaporated onto a thin substrate. Cathode strips arranged between the anode strips allow for an improved field quality and a fast removal of positive ions. The segmentation of the otherwise planar cathodes in the form of strips or pixels also permits two-dimensional readout. Instead of mounting the electrode structures on ceramic substrates, they can also be arranged on thin plastic foils. In this way, even light, flexible detectors can be constructed, which exhibit a high spatial resolution. Possible disadvantages lie in the electrostatic charging-up of the insulating plastic structures, which can lead to time-dependent amplification properties because of the modified electric fields. The obvious advantages of these microstrip detectors – apart from their excellent spatial resolution – are the low dead time (the positive ions being produced in the avalanche will drift the very short distance to the cathode strips in the vicinity of the anodes), the reduced radiation damage (because of the smaller sensitive area per readout element), the high-rate capability, and the reduced occupancy.

Another structure providing charge multiplication is the gas electron multiplier (GEM). This is a thin ($\approx$50 μm) insulating kapton foil coated with a metal film on both sides. It contains chemically produced holes of 50–100 μm in diameter with 100–200 μm pitch. The metal films have different potentials to allow gas multiplication in the holes. A GEM-based detector contains a drift cathode separated from one or several GEM layers and an anode readout structure to read out the ionization information. The readout is then done in a similar way as in most other micropattern detectors.

3.10
Transition-Radiation Detectors (TRD)

Transition radiation is emitted when a charged particle traverses a medium with discontinuous dielectric constant. A charged particle moving toward a boundary, where the dielectric constant changes, can be considered to form together with its mirror charge an electric dipole whose field strength varies in time. The time-dependent dipole field causes the emission of electromagnetic radiation. This emission can be understood in such a way that although the dielectric displacement $\vec{D} = \varepsilon\varepsilon_0\vec{E}$ varies continuously in passing through a boundary, the electric field does not. The energy radiated from a single boundary (transition from vacuum to a medium with dielectric constant ε) is proportional to the Lorentz factor of the incident charged particle. This provides an excellent possibility to identify the nature of this particle.

Since the number of transition-radiation photons produced per boundary is rather low, usually a stack of transparent foils is employed. The generated photons – preferentially in the X-ray domain – are detected in a standard MWPC or drift chamber. It is important that the generator foils are made of low-Z material, so that they do not absorb the X-ray transition-radiation photons, and the counting gas in the MWPC is of high Z (e.g., xenon) to efficiently detect them. Transition-radiation detectors (TRDs) are mostly used for particle-identification purposes. This is related to the fact that the yield of transition-radiation photons is proportional to the Lorentz factor (i.e., energy) of the particle rather than the velocity, which saturates at high energies.

3.11
Neutron Counters

Neutrons, just as photons, can only be detected indirectly. Depending on the neutron energy, various reactions can be considered, which produce charged particles that are then detected via their ionization or scintillation in standard radiation detectors.

1. Low kinetic energies (<20 MeV)

$$
\begin{aligned}
n + {}^{6}\mathrm{Li} &\rightarrow \alpha + {}^{3}\mathrm{H} \\
n + {}^{10}\mathrm{B} &\rightarrow \alpha + {}^{7}\mathrm{Li} \\
n + {}^{3}\mathrm{He} &\rightarrow p + {}^{3}\mathrm{H} \\
n + p &\rightarrow n + p
\end{aligned}
\tag{5}
$$

   The conversion can be performed in a proportional counter with BF_3 or ^{3}He filling or a mixture with an addition of hydrocarbons, like CH_4. Also a scintillator consisting of suitable conversion material like LiI(Tl) can be used.
2. Medium energies
   ($20\ \mathrm{MeV} \leq E_{\mathrm{kin}} \leq 1\ \mathrm{GeV}$)
   The (n, p) recoil reaction can be used for neutron measurements in detectors, which contain many quasi-free protons in their sensitive volume (e.g., hydrocarbons, like, e.g., CH_4).
3. High energies ($E_{\mathrm{kin}} > 1\ \mathrm{GeV}$)
   Neutrons of high energy initiate hadron cascades in inelastic interactions, which are easy to identify in hadron calorimeters.

Neutrons are detected with relatively high efficiency at very low energies. Therefore, it is often useful to slow down neutrons with substances containing many protons (“moderation”), because neutrons can transfer a large amount of energy to collision partners of the same mass. In the field of radiation protection, it is also important to measure the neutron energy (e.g., by threshold counters), because the relative biological effectiveness depends on it.

4
Scintillators

Scintillator materials can be inorganic crystals, organic liquids or plastics, and gases. The scintillation mechanism in inorganic crystals is an effect of the lattice. Incident particles can transfer energy to the lattice by creating electron–hole pairs or taking electrons to higher energy levels below the conduction band. Recombination of electron–hole pairs may lead to the emission of light. Also, electron–hole bound states (excitons) moving through the lattice can emit light when hitting an activator center and transferring their binding energy to activator levels, which subsequently de-excite. Important parameters for inorganic scintillators are the light yield, the emission spectrum, and the decay time of the light signal. In thallium-doped NaI crystals, about 25 eV are required to produce one scintillation photon. The decay time in inorganic scintillators can be quite long (e.g., 1 microsecond in CsI(Tl)), but there are also fast scintillators like BaF_2 with a decay component of 0.7 nanoseconds. BaF_2 has its emission maximum in the UV and therefore requires quartz windows for the readout with photomultipliers. If scintillators are used as sampling elements in electromagnetic calorimeters, materials with a short radiation length X_0, like bismuth germanate (BGO: $Bi_4Ge_3O_{12}$), cadmium or lead tungstate ($CdWO_4$, $PbWO_4$), or cerium-activated lutetium oxyorthosilicate (LSO: $Lu_2SiO_5(Ce)$) are required.

In organic substances, the scintillation mechanism is different. Certain types of molecules will release a small fraction (≈3%) of the absorbed energy as optical photons. This process is especially marked in organic substances that contain aromatic rings, such as polystyrene, polyvinyltoluene, and naphthalene. Liquids that scintillate include toluene or xylene. The primary scintillation light is preferentially emitted in the UV range. The absorption length for UV photons in the scintillation material is rather short: the scintillator is not transparent for its own scintillation light. Therefore, this light is transferred to a wavelength shifter, which absorbs the UV light and reemits it at longer wavelengths (e.g., in the green). Owing to the lower concentration of the wavelength-shifter material, the reemitted light can get out of the scintillator and be detected by a photosensitive device (see Figure 6). The technique of wavelength shifting is also used to match the emitted light to the spectral sensitivity of the photomultiplier. For plastic scintillators, the primary scintillator and wavelength shifter are mixed with an organic material to form a polymerizing structure. In liquid scintillators, the two active components are mixed with an organic base.

About 100 eV are required to produce one photon in an organic scintillator. The decay time of the light signal in plastic scintillators is, in most cases, substantially shorter compared to inorganic substances (e.g., 30 nanoseconds in naphthalene and around 2–3 nanoseconds in polystyrene). Also polystyrene-based scintillators containing soluted organic compounds of heavy elements with ultrashort decay times (0.5 nanoseconds) for fast-timing purposes are available.

Gas scintillators are used in cosmic rays for extensive air-shower experiments, where the atmospheric nitrogen is used as fluorescence agent. Because of the low light absorption in gases, there is no need for wavelength shifting.

Plastic scintillators do not respond linearly to the energy-loss density. The number of photons produced by charged particles is described by Birk's semiempirical formula

$$N = N_0 \frac{dE/dx}{1 + k_B dE/dx} \tag{6}$$

where N_0 is the photon yield at low specific ionization density and k_B is

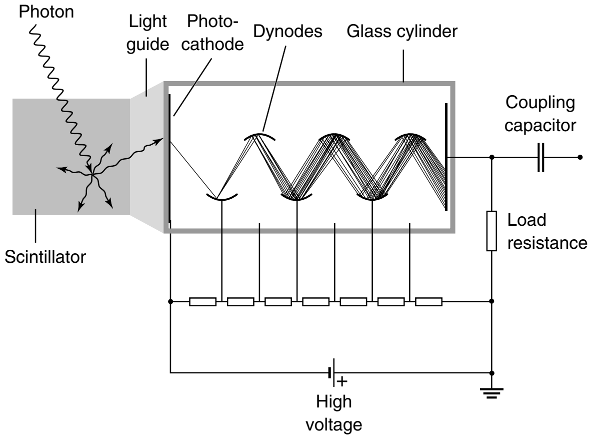

Fig. 6 Scintillation counter with photomultiplier readout; Grupen, 2008.

Birk's density parameter. For low energy losses, Equation (6) leads to a linear dependence

$$N = N_0 \cdot \frac{dE}{dx} \tag{7}$$

while for very high dE/dx saturation occurs at

$$N = \frac{N_0}{k_B} \tag{8}$$

5
Solid-State Detectors

Solid-state detectors are essentially ionization chambers with solids as a counting medium. Because of their high density compared to gaseous detectors, they can absorb particles of correspondingly higher energy. Charged particles or photons produce electron–hole pairs in a crystal. An electric field applied across the crystal allows the charge carriers produced to be collected.

The operating principle of solid-state detectors can be understood from the band model of solids. According to this theory, the electron energy levels of individual atoms or ions within a crystal are unified forming energy bands. According to the Pauli principle, each band can contain a finite number of electrons. So, some low-energy bands are fully filled with electrons while the high-energy bands are empty, at least at low temperature. The lowest partially filled or empty band is called *conduction band*, while the highest fully filled band is referred to as *valence band*. The gap between these bands is called *forbidden band* or band gap, E_g.

When the "conduction band" is partially filled, electrons can move easily under the influence of an electric field; hence, this solid is a conductor. Such a material cannot be used as an ionization counter. The solids that have basically empty conduction bands are divided conventionally into insulators (specific resistivity $10^{14}–10^{22}$ Ω cm at room temperature) and semiconductors ($10^9–10^{-2}$ Ω cm). The electric charge in these materials is carried by electrons, which have been excited to the conduction band from the valence band. The corresponding vacancies in the valence band are called *holes* and are able to drift in the electric field as well.

The common solid-state ionization counters are based on semiconductors. The specific resistivity of the material is determined as

$$\varrho = \frac{1}{e(n\mu_e + p\mu_p)} \tag{9}$$

where n and p are electron and hole concentrations, respectively, while μ_e, μ_p are their mobilities and e is the elementary charge. In a pure semiconductor the electron concentration, n, is equal to the hole concentration, p. Electron and hole concentrations in solid-state detectors are intentionally changed by special doping. A germanium or silicon crystal becomes n-conducting (i.e., conducting for negative charge carriers, electrons), when electron-donor impurities are introduced into the lattice. Germanium and silicon have four electrons in the outer shell. Atoms with five electrons in the outer shell act as electron-donor impurities. Similarly, germanium and silicon become p-conducting (i.e., conducting for holes) if trivalent atoms, which act as electron-acceptor impurities, are introduced into the crystal lattice. Phosphor and arsenic act as electron donors. The neighboring silicon (germanium) atoms can only bind four electrons. The fifth electron of the electron-donor impurity is only weakly bound and can easily reach the conduction band. The donor levels are situated approximately 0.05 eV below the edge of the conduction band. If trivalent electron-acceptor impurities like boron or indium are added to the lattice, one of the silicon bonds remains incomplete. This acceptor level, which is about 0.05 eV above the edge of the valence band, tries to attract one electron from a neighboring silicon atom. As a consequence, the state of the missing electron (the hole) migrates through the crystal. The addition of small amounts of energy can cause the holes to be transferred to the valence band, thereby causing a hole current. Lithium acts as an electron donor because it has only one weakly bound electron in the outer shell.

If a charged particle traverses an *n*- or *p*-conducting crystal, it will produce electron–hole pairs along its track. The primary electrons can produce further secondary electron pairs or excite lattice vibrations. A plasma channel is produced along the particle track with charge-carrier concentrations from 10^{15} up to $10^{17}/\mathrm{cm}^3$. The working principle of a solid-state detector now consists of collecting the free charge carriers in an external drift field before they can recombine with the holes. If this is successful, the measured charge signal is proportional to the energy loss of the particle or, if the particle deposits its total energy in the sensitive volume of the detector, it is proportional to the particle energy.

Solid-state counters can be used in nuclear physics for α, β, and γ spectroscopy. High-purity germanium detectors provide excellent energy resolution at the level of $\Delta E/E \approx 10^{-3}$ in the mega-electronvolt range (see Figure 7 and Figure 8). If the readout planes of silicon detectors are segmented into strips

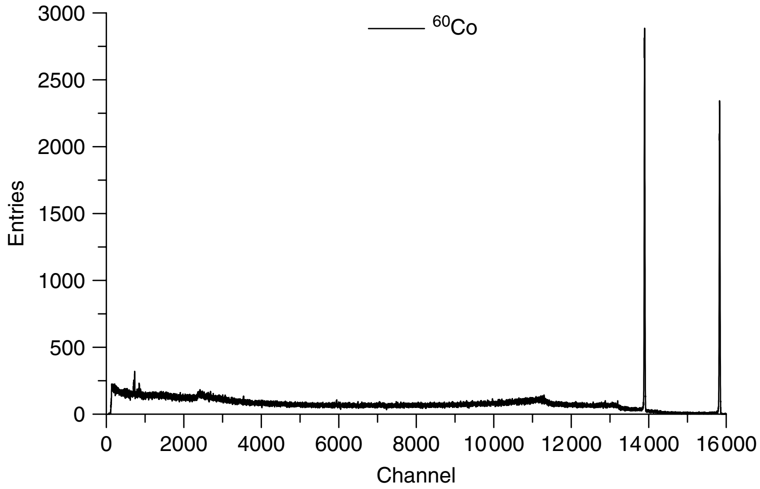

Fig. 7 γ-ray spectrum of ^{60}Co measured in a high-purity Ge detector. The energy resolution of the 1.17 MeV and 1.32 MeV γ-ray lines is rather impressive; Hausmann, 2002; private communication 2002. Photo credit Joachim Hausmann.

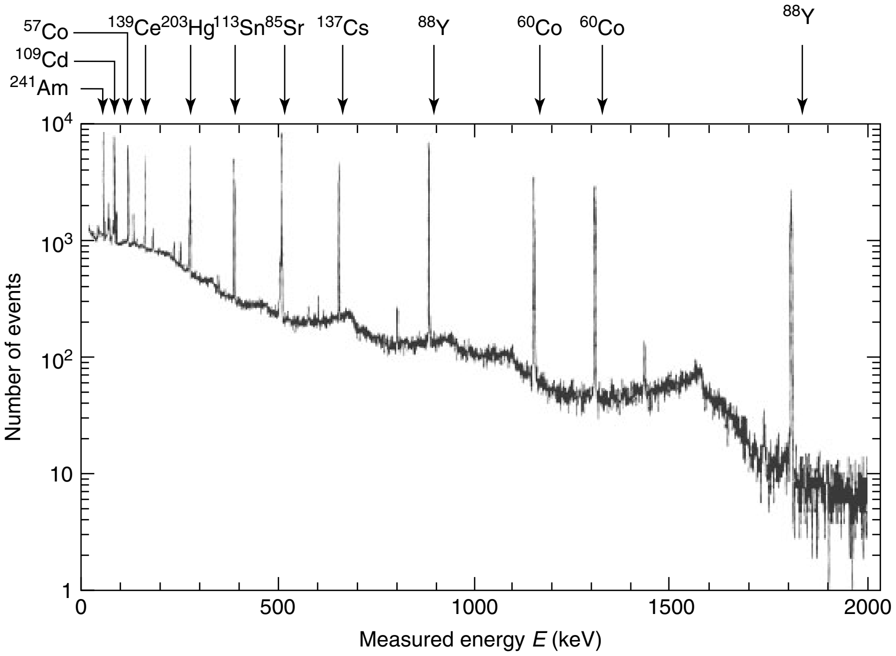

Fig. 8 Photopeak identification in a radioisotope sample using a high-purity germanium detector; Grupen, 2008.

or pixels, also outstanding spatial resolutions of about 10 μm can be achieved. This, however, requires a large number of readout channels equipped with low-noise charge-sensitive amplifiers. The number of readout elements can be reduced if, in addition, the drift time from the moment of charge generation to the arrival time of the drifting charge at the readout element is measured (silicon drift chamber).

For the operation of silicon pixel counters in harsh radiation environments, aging effects also become an issue.

Even better energy resolutions can be achieved if quantum transitions, which require less energy than ionization or electron–hole pair creation, are used. For example, phonons in solid-state materials have energies around 10^{-5} eV for temperatures of 100 mK. Other types of quasi-particles at low temperature are Cooper pairs in a superconductor, which are bound states of two electrons with opposite spin that behave like bosons. Cooper pairs in superconductors have binding energies in the range between $4 \cdot 10^{-5}$ eV (Ir) and $3 \cdot 10^{-3}$ eV (Nb). Thus, even extremely low energy depositions would produce or break up a large number of phonons or Cooper pairs. To avoid thermal excitations of these quantum processes, such radiation detectors, however, would have to be operated at extremely low temperatures, typically in the millikelvin range. For this reason, such devices are called cryogenic detectors. Cryogenic detectors can be subdivided into two main categories: first, detectors for quasi-particles in superconducting crystals, and secondly, phonon detectors in insulators. One detection method is based on the fact that the superconductivity of a substance is destroyed by the energy deposition if the detector element is sufficiently small. This is the working principle of superheated superconducting granules. The transition from the superconducting to the normal-conducting state can be detected by pick-up coils coupled to very sensitive preamplifiers or by superconducting quantum interference devices (SQUIDs). These quantum interferometers are extremely sensitive radiation detectors.

In contrast to Cooper pairs, phonons, which can be excited by energy deposition in insulators, can be detected with the methods of classical calorimetry. If ΔE is the absorbed energy, this results in a temperature rise of

$$\Delta T = \frac{\Delta E}{mc} \tag{10}$$

where c is the specific heat capacity and m is the mass of the detector. If these measurements are performed at very low temperatures, where c can be very small (the lattice contribution to the specific heat is proportional to T^3 at low temperatures), this method can be used to detect individual particles of extremely low energy.

6
Cherenkov Detectors

A charged particle, traversing a medium with refractive index n with a velocity v exceeding the velocity of light c/n in that medium, emits a characteristic electromagnetic radiation, called *Cherenkov radiation*. Cherenkov radiation is emitted because the charged particle polarizes atoms along its track so that they become electric dipoles. The time variation of the dipole field leads to the emission of electromagnetic radiation. As long as $v < c/n$, the dipoles are symmetrically arranged around the particle path, so that the dipole field integrated over all dipoles vanishes and no radiation occurs. If, however, the particle moves with $v > c/n$, the symmetry is broken resulting in a nonvanishing dipole moment, which leads to the radiation. The contribution of Cherenkov radiation to the energy loss is small compared to that from ionization and excitation even for minimum-ionizing particles. For gases with $Z \geq 7$, the energy loss by Cherenkov radiation amounts to less than 1% of the ionization loss of minimum-ionizing particles.

The angle between the emitted Cherenkov photons and the track of the charged particle can be worked out to be

$$\cos\theta_{\mathrm{C}} = \frac{1}{n\beta} \tag{11}$$

For the emission of Cherenkov radiation there is a threshold effect. Cherenkov radiation is only emitted, if $\beta > \beta_{\mathrm{C}} = \frac{1}{n}$. At threshold, Cherenkov radiation is emitted in the forward direction. The Cherenkov angle increases until it reaches a maximum for $\beta = 1$, namely,

$$\theta_{\mathrm{C}}^{\max} = \arccos\frac{1}{n} \tag{12}$$

Consequently, Cherenkov radiation is emitted only if the medium and the frequencies ν are such that $n(\nu) > 1$. The maximum emission angle, $\theta_{\mathrm{C}}^{\max}$, is small for gases ($\theta_{\mathrm{C}}^{\max} \approx 1.5^\circ$ for air) and becomes large for condensed media (about 45° for usual glass).

For fixed energy, the threshold Lorentz factor depends on the mass of the particle. Therefore, the measurement of Cherenkov radiation is suited for particle-identification purposes.

The number of Cherenkov photons emitted per path length with wavelengths between λ_1 and λ_2 is given by

$$\frac{dN}{dx} = 2\pi\alpha z^2 \int_{\lambda_1}^{\lambda_2} \left(1 - \frac{1}{n^2\beta^2}\right) \frac{d\lambda}{\lambda^2} \tag{13}$$

for $n(\lambda) > 1$, where z is the electric charge of the particle producing Cherenkov radiation and α is the fine-structure constant.

Neglecting the dispersion of the medium (i.e., n independent of λ) leads to

$$\frac{dN}{dx} = 2\pi\alpha z^2 \cdot \sin^2\theta_C \cdot \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right) \tag{14}$$

For the optical range ($\lambda_1 = 400$ nm and $\lambda_2 = 700$ nm) one obtains for singly charged particles ($z = 1$)

$$\frac{dN}{dx} = 490 \sin^2\theta_C \,\mathrm{cm}^{-1} \tag{15}$$

All transparent materials are candidates for Cherenkov radiators. In particular, Cherenkov radiation is emitted in all scintillators and in the light guides, which are used for the readout. The scintillation light, however, is approximately 100 times more intense than the Cherenkov light. A large range of indices of refraction can be covered by the use of solid, liquid, or gaseous radiators. Even indices of refraction intermediate between solids and gas can be achieved by using transparent silica aerogels.

Cherenkov detectors are mainly used for particle identification in beams of particles of fixed momentum. Lighter particles of fixed momentum are faster than heavier ones. Therefore threshold Cherenkov counters, being sensitive to the velocity, can be adjusted to distinguish between these particles. Special imaging devices, like the ring imaging Cherenkov detector (RICH) or other differential Cherenkov counters, can also provide a direct measurement of the particle velocity.

7
Calorimeters

7.1
Electron–Photon Calorimeters

The development of cascades induced by electrons, positrons, or photons is governed by bremsstrahlung of electrons and pair production of photons. Secondary particle production continues until photons fall below the pair-production threshold, and energy losses of electrons other than bremsstrahlung start to dominate: the number of shower particles decays exponentially.

Already a very simple model can describe the main features of particle multiplication in electromagnetic cascades: a photon of energy E_0 starts the cascade by producing an e^+e^- pair after one radiation length. Assuming that the energy is shared symmetrically between the particles at each multiplication step, one gets at the depth t

$$N(t) = 2^t \tag{16}$$

particles with energy

$$E(t) = E_0 \cdot 2^{-t} \tag{17}$$

The multiplication continues until the electrons fall below the critical energy E_c,

$$E_c = E_0 \cdot 2^{-t_{\max}} \tag{18}$$

From then on, ($t > t_{\max}$) the shower particles are only absorbed. The position of the shower maximum is obtained from Equation (18),

$$t_{\max} = \frac{\ln(E_0/E_c)}{\ln 2} \approx \ln E_0 \tag{19}$$

The total number of shower particles is

$$\begin{aligned} S &= \sum_{t=0}^{t_{\max}} N(t) \\ &= \sum 2^t = 2^{t_{\max}+1} - 1 \approx 2^{t_{\max}+1} \end{aligned}$$

$$S = 2 \cdot 2^{t_{\max}} = 2 \cdot \frac{E_0}{E_c} \propto E_0 \tag{20}$$

If the shower particles are sampled in steps t measured in units of the radiation length X_0, the total number of track segments is obtained as

$$S^* = \frac{S}{t} = 2\frac{E_0}{E_c} \cdot \frac{1}{t} \tag{21}$$

which leads to an energy resolution of

$$\frac{\sigma}{E_0} = \frac{\sqrt{S^*}}{S^*} = \frac{\sqrt{t}}{\sqrt{2E_0/E_c}} \propto \frac{\sqrt{t}}{\sqrt{E_0}} \tag{22}$$

In a more realistic description, the longitudinal development of the electron shower can be approximated by

$$\frac{dE}{dt} = \text{const} \cdot t^a \cdot e^{-bt} \tag{23}$$

where a, b are fit parameters.

With homogeneous electron–photon calorimeters, energy resolutions of

$$\frac{\sigma_E}{E} = \frac{3\%}{\sqrt{E/\text{GeV}}} \tag{24}$$

can be achieved.

The lateral spread of an electromagnetic shower is mainly caused by multiple scattering. It is described by the Molière radius

$$R_M = \frac{21\ \text{MeV}}{E_c} X_0 (\text{g/cm}^2) \tag{25}$$

In a homogeneous calorimeter, 95% of the shower energy is contained in a cylinder of radius $2R_M$ around the shower axis.

Since the position of the shower maximum depends only logarithmically on the energy, electron–photon calorimeters can cover quite a large range of energies from some hundred mega electronvolts up to several hundred giga electronvolts even at moderate detector sizes. Electron–photon calorimeters are usually total-absorption devices. Such calorimeters for electrons and photons are mostly made of scintillating crystals. However, also Cherenkov media, like lead glass, can be used as converter materials.

7.2
Hadron Calorimeters

The longitudinal development of electromagnetic cascades is characterized by the radiation length X_0 and their lateral width is determined by multiple scattering. In contrast to this, hadron showers are governed in their longitudinal structure by the nuclear interaction length λ and by transverse momenta of secondary particles as far as lateral width is concerned. Since for most materials $\lambda \gg X_0$ and $\langle p_T^{\text{interaction}} \rangle \gg \langle p_T^{\text{multiple scattering}} \rangle$, hadron showers are longer and wider (see Figure 9).

Part of the energy of the incident hadron is spent on breaking up nuclear bonds. This fraction of the energy is

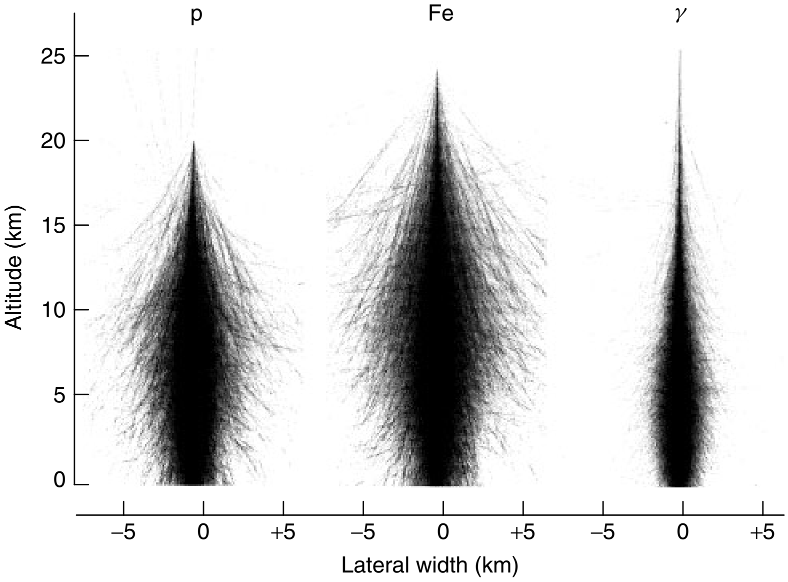

Fig. 9 Comparison between proton-, iron-, and photon-induced cascades in the atmosphere. The primary energy in each case is 10^{14} eV; Knapp and Heck, 1998. Photo credit Johannes Knapp 2007.

invisible in hadron calorimeters. Further energy is lost by escaping particles like neutrinos and muons as a result of hadron decays. Since the fraction of lost binding energy and escaping particles fluctuates considerably, the energy resolution of hadron calorimeters is systematically inferior to electron calorimeters.

Hadron calorimeters are usually sampling devices where active detector layers interspersed with passive targets collect the energy of shower particles. The active detector parts can be scintillators or liquid-argon or gaseous detectors. As passive targets mostly lead, uranium, copper, or iron is used. With a high-performance uranium–liquid-argon calorimeter energy resolutions of $\sigma_E/E = 35\%/\sqrt{E/\mathrm{GeV}}$ can be achieved.

Acknowledgments

It is a pleasure to thank Dr Tilo Stroh for a very careful reading of the chapter and his efficient help in laying out this chapter in a professional way in LaTeX.

Glossary

Acceptor: substance added as impurity to a semiconductor to create hole conductivity.

Activator: dopant used in scintillators to de-exite excitons, which then provide photons.

Aging: the performance of detectors may deteriorate in harsh radiation environments.

Avalanche Formation: if drifting electrons gain enough energy in an electric field on their mean free path, they themselves can ionize other atoms and initiate avalanche formation.

Bethe–Bloch Formula: describes the energy loss of charged particles by ionization and excitation.

Charge-Coupled Device (CCD): pixelized silicon detector, which digitizes the incident photon or particle pattern, which is then electronically read out.

Cherenkov Effect: particles with velocities v faster than light in media ($v \geq c/n$; n – index of refraction) create a characteristic electromagnetic radiation.

Critical Energy: energy at which the energy loss of electrons due to ionization and bremsstrahlung are the same.

Decay Length: track length of a short-lived particle before it decays.

Donor: substance added as impurity to a semiconductor, which can donate electrons to the conduction band.

Exciton: electron–hole bound state.

Fermi Plateau: saturation of the ionization loss at high energies due to polarization effects.

First Townsend Coefficient: characterizes the gas-amplification factor in its electric field and pressure dependence.

Frisch Grid: grid electrode in an ionization chamber, which provides signals that do not depend on the position of impact of the particles into the chamber.

Gating: by a grid in a TPC interesting events can be selected by applying suitable voltages to it.

Ionization Loss: energy loss of charged particles by transferring part of their energy to atomic electrons.

Ionization Yield: number of electron–ion pairs created by charged particles.

Knock-on Electron: atomic electron, which receives a rather large energy transfer in a collision with a charged particle (also called δ ray).

Landau Fluctuations: the energy-loss fluctuations due to the stochastic nature of the ionization process lead to asymmetric energy-loss distributions especially for thin targets.

Larmor Radius: bending radius of a charged particle gyrating in a magnetic field.

Lattice Vibrations: quantized oscillations of a lattice: phonons.

Light Yield: amount of light created in scintillators by charged particles or γ rays.

Lorentz Angle: in a combined electric and magnetic field electrons drift in the direction of $\vec{E} \times \vec{B}$.

Lorentz Factor: $\gamma = 1/\sqrt{1-\beta^2} = E/mc^2$, where $\beta = v/c$ and c is the velocity of light.

Minimum-Ionizing Particles (mips): relativistic charged particles with $\beta \cdot \gamma \approx 4$ exhibit the lowest energy loss by ionization and excitation.

Moderation: slowing down of neutrons by materials containing many quasi-free protons.

Molière Radius: characteristic size of an electromagnetic cascade describing its lateral spread.

Nuclear Interaction Length: characteristic distance between two collisions of a hadron.

Occupancy: probability that a readout element of a detector is occupied by a particle so that no further particles can be recorded on this element. High occupancy leads to dead times and inefficiency.

Pad: cathode-structure element, which enables the determination of the position of a charged particle along the wire in MWPCs.

Pitch: distance between anodes or readout structures.

Pixel: picture element as readout structure, mostly in silicon detectors.

Quenching: vapor additions to noble gases in Geiger–Müller counters limit the range of avalanche photons and cause the gas discharge to stop by itself.

Radiation Damage: degradation of the performance of detectors in a harsh radiation environment. Detectors that can be operated under such conditions are called *radiation hard* or *radiation tolerant*.

Radiation Length X_0**:** characteristic length describing the energy loss of electrons due to bremsstrahlung, $dE/dx = E/X_0$.

Relativistic Rise: rising energy loss at intermediate energies due to the relativistic deformation of the transverse electric field of fast charged particles.

RICH: ring imaging Cherenkov counter for the measurement of the velocity of charged particles.

Silica Aerogel: artificially produced Cherenkov medium of very low index of refraction consisting of a porous structure of silica and sub-micrometer air bubbles.

SQUID: superconducting quantum interference device used for the measurement of small magnetic flux changes.

Wavelength Shifter: agent that absorbs UV light and reemits visible light at lower frequency.

W Value: energy required to create an electron–ion pair in a gas or an electron–hole pair in a semiconductor.

References

ALEPH Collaboration (2000) http://aleph.web.cern.ch/aleph/.

Close, F. and Marten, M. (1987) *The Particle Explosion*, Oxford University Press, Oxford.

Galison, P. (1997) *Image and Logic: A Material Culture of Microphysics*, The University of Chicago Press, Chicago.

(a) Grupen, C. (2008) *Grundkurs Strahlenschutz, Springer, Heidelberg;* and (2009) *Introduction to Radiation Protection*, Springer, Heidelberg; (b) Kalthoff, O. (1996) *Berechnung der Photopeakeffizienz für koaxiale Reinstgermaniumdetektoren*, Diploma Thesis, Siegen.

Hausmann, J. (2002) *Development of a low noise integrated readout electronic for pixel detectors in CMOS technology for a Compton camera*, PhD Thesis, Siegen.

Kaftanov, V.S. and Liubimov, V.A. (1963) Spark chamber use in high energy physics. *Nucl. Instr. Meth.*, **20**, 195.

Knapp, J. and Heck, D. (1998) Luftschauersimulationsrechnungen mit dem CORSIKA programm. *Forschungszentrum Karlsruhe Nachrichten*, **30**, 27; and www.ast.leeds.ac.uk/~fs/.

Rochester, G.D. and Turver, K.E. (1981) Cosmic rays of ultra-high energy. *Contemp. Phys.*, **22**, 425.

Particle Data Group, Yao, W.-M. *et al.* (2006) Review of particle physics. *J. Phys. G*, **33**, 1–1232, http://pdg.lbl.gov/.

Further Reading

There are many conferences specialized in radiation detectors, like the annual "Nuclear Science Symposium" or the "Instrumentation and Measurement Technology Conference" organized by the IEEE, and the "Vienna conference on Instrumentation." Results on radiation detectors are mostly published in the journal "Nuclear Instruments and

Methods in Physics Research Section A." There are also a number of excellent monographs on radiation detectors, like K. Kleinknecht, *Detectors for Particle Radiation*, Cambridge University Press (1986); F. Sauli (ed.), *Instrumentation in High Energy Physics*, World Scientific, Singapore (1992); G.F. Knoll, *Radiation Detection and Measurement*, 3rd edition, John Wiley & Sons Inc., New York (1999/2000); D. Green, *The Physics of Particle Detectors*, Cambridge University Press (2000); C. Grupen & B. Shwartz, *Particle Detectors*, Cambridge University Press (2008).

1
Quantum Molecular Biological Methods Using Density Functional Theory

Karl James Jalkanen, Sándor Suhai, and Henrik G. Bohr

1.1 Introduction 9
1.2 Introduction to Density Functional Theory methods 22
1.3 Calculating Vibrational Absorption and Vibrational Circular Dichroism Spectra 28
1.3.1 Vibrational Absorption (VA) or Infrared (IR) Absorption and Vibrational Circular Dichroism 28
1.4 Calculating Raman Scattering and Raman Optical Activity Spectra 30
1.4.1 Raman Scattering 30
1.4.2 Raman Optical Activity 32
1.5 Classifying Conformational States and Structures based on VA, VCD, Raman, and ROA Spectra 35
1.5.1 Experimental Vibrational Frequencies for *N*-Methylacetamide, NALANMA, Peptides, and Proteins 35
1.5.2 Results for the DFT Calculations 35
1.5.2.1 L-Alanine 35
1.5.2.2 *N*-Acetyl L-Alanine *N*′-Methylamide 39
1.5.2.3 L-Alanyl L-Alanine 43
1.5.2.4 L-Histidine 49
1.5.2.5 *N*-Acetyl L-Histidine *N*′-Methylamide 51
1.5.2.6 *N*-Acetyl L-Alanine L-Proline L-Tyrosine *N*′-Methylamide 53
1.5.3 Future Perspectives and Applications for DFT-based Simulations 55
1.5.4 Future Perspectives of the Multidisciplinary Fields of Molecular Biophysics, Molecular Biology, Systems Biology, and Molecular Medicine and Molecular Psychiatry 56
Acknowledgments 60
References 60

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

1.1
Introduction

Toward the end of last century, when classical molecular dynamics (MD) became a very sought after technique for studying structure and conformations of large biomolecules, it was not an appropriate quest to employ and utilize quantum mechanical methods for studying dynamics of biomolecules (Warshel and Levitt, 1976; van Gunsteren *et al.*, 1983; Kollman, 1992). Initially, the potential energy surfaces were determined by fitting to experimental observed spectra (Colwell and Percival, 1983a,b), but later, also by parameterization/fitting to a combination of both experimental and theoretical data (Brooks *et al.*, 1983; Weiner *et al.*, 1984, 1986; Floris *et al.*, 1992; Brooks and Case, 1993; Cornell *et al.*, 1995; Dudek and Ponder, 1995; Maple *et al.*, 1998; MacKerell *et al.*, 1998). Owing to a limited amount of experimental data many assumptions and approximations needed to be made. These assumptions and approximations gained credence and support by being able to reproduce the (limited amount of) data. This has also created subsequent problems, as many of the assumptions and approximations involved throwing away/neglecting fundamental physics, which has subsequently been shown to be important and is essential to predict some experimental observables. In many cases, the acceptance and widespread use of simplified models and theories have prevented researchers from further developing and extending the models and theories, as it has been hard to get funding to further develop and extend methods when the general consensus in the "application" field is that the methods are already of a sufficient level to warrant general use in applications. This has subsequently immensely slowed down the development of methods, as the funding has focused on applications to the detriment of further developing the methods. The black box approach to theory becomes prevelant and then researchers are limited to what is available in the packages like Gaussian, Amsterdam Density Functional (ADF), Dalton, QChem, Turbomole, Cambridge Analytical Derivatives Package (CADPAC), Spanish Initiative for Electronic Simulations with Thousands of Atoms (SIESTA), Vienna Ab-initio Simulation Package (VASP), and so on. The development and refinement of the methods is then left to a small number of researchers who all have their so-called development versions of these codes/packages. More recently, there has

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

been a call for more open-source codes, such as ABINIT, and for the National Science Foundation USA (NSF), Australian Research Council (ARC), National Institutes of Health (NIH), European Union (EU), and the National Health and Medical Research Council Australia (NHMRC) to continue to fund the development of methods for use in quantum nanobiology and molecular biology and for a core of young researchers to be supported to learn these methods by summer schools, North Atlantic Treaty Organization (NATO) workshops, and so on. This is clearly a high priority if the field is to continue to thrive and flourish and for the methods to continually be extended and developed, as problems are identified by the so-called users of the academic and commercial codes. The time lag between identifying the problems and implementing the changes/new developments needs to be reduced.

Basically, the overall goal of utilizing wave function and density functional theory based quantum mechanical methods for biomolecules is to derive from first principles functional and spectroscopic data. Here one calculates the structures, electric and magnetic properties, and vibrational, nuclear magnetic resonance (NMR), electron spin resonance (ESR), and electronic spectra of isolated molecules, aggregates, and molecular complexes in various environments where measurements have been made (nonpolar solvents, embedded within membranes, in solvents of varying polarity and ionic strengths, and finally in mixed environments, the most complex being proteins and glycoproteins that have part outside the membrane, part within the membrane, and part inside the membrane). Here one realizes that the structure, function, and properties of these complex machines are lost when the machine (molecular complex) is broken down into its various parts (individual species). Here the combination of theory and experiment has a lot to offer, as either on its own is less useful.

Within the current set of assumptions and approximations, many physically measurable properties are predicted to be "zero," which clearly shows that there are occasions and phenomena which cannot be treated by making the simplifying assumptions and approximations. The most commonly used ones being the so-called Born–Oppenheimer (BO) approximation and the harmonic approximation (both mechanical (for vibrational frequencies) and electronic (for infrared/vibrational intensities)). Actually, the concept of the potential energy surface which one parameterizes to generate a molecular mechanics force field is based on the validity of the BO approximation, which as we shall see later breaks down for some very interesting phenomena, the so-called non-BO properties. Many of these properties involve electric currents, magnetic fields, and angular momenta. The development/parameterization of classical mechanical force fields (and other molecular property surfaces which depend only on the nuclear positions of the atoms) is a field within itself, and the interested are encouraged to read speciality papers on this (Maple *et al.*, 1998; MacKerell *et al.*, 1998). The point here is that even though these molecular mechanics force fields are commonly used in molecular biology, they should be used only in specific cases where the assumptions and approximations made are valid. Clearly, for the case of quantum molecular biology, where a biological process is initiated by the absorption of a photon of light, followed by subsequent electron and energy transfer, one must go beyond the use of molecular mechanics methods. Here, one can turn to either semiempirical-based methods or first principles/*ab initio*-based methods.

The field of developing and parameterizing molecular mechanical force fields (and property surfaces) has itself benefited immensely from both wave function quantum mechanics (WFQM)- and density functional theory (DFT)-based methods, as has the field of developing semiempirical-based methods, which are based on WFQM (Klopman, 1964a,b; Dewar *et al.*, 1985; Stewart, 1989, 2007) and DFT (Elstner *et al.*, 1998; Bohr *et al.*, 1999; Frimand and Jalkanen, 2002). In addition, as more accurate and complete experimental data becomes available, these data can be used both in the parameterization and testing/verifying stages. This is especially important for more exotic molecules and molecular complexes and assemblies (nanoparticles and protein–ligand complexes to name two), which have not been included in the data sets used for parameterizing and/or testing. In many cases, for example in non *ab initio* DFT methods, this new data provides a wealth of data to test the newly developed force fields, semiempirical methods, and new exchange and correlation (XC) functionals.

Since the beginning of this century, it has seemed much more appropriate to study, mainly, the electronic structure of large molecules with the help of quantum mechanical methods, with WFQM, DFT, and Green's functions being the three most commonly used. This is because the electronic structures and properties are necessary for deriving the spectroscopic data and for understanding experimental observables (electric and magnetic properties) for the structures of molecules and molecular complexes/aggregates (Farid, 2002; Onida, Reining and Rubio, 2002; Amos, Jalkanen and Stephens, 1988; Amos *et al.*, 1987; Amos, 1982, 1986, 1984, 1987; Buckingham, 1967). Additionally, the electronic structures and properties are important for understanding the reactivity and functionality of biomolecules and biomolecular complexes (Deslongchamps, 1983; Hanessian, 1983). Hence one must extend the methods of classical MD simulations, where only one calculates the energy and gradients, to include the calculation of the various electronic and magnetic properties at each step of the MD simulation. Here, one has various ways to proceed. One can parameterize molecular property surfaces, similar to potential energy surfaces (force fields). For some properties, the property may depend on the momenta of the particles in addition to the positions (Buckingham, 1967; Amos, 1982, 1987; Buckingham, Fowler and Galwas, 1987; Lee, Colwell and Handy, 1994; Ioannnou, Colwell and Amos, 1997; Hornicek, Kapralova and Bour, 2008). The properties (observables) that depend on the momenta, in addition to the positions, are the so-called non-BO property surfaces. In some cases, there are equivalent expressions in terms of positions, momenta, and torque, which in theory should give the same results, but in practice for a variety of reasons do not (Stephens *et al.*, 1990; Amos, Jalkanen and Stephens, 1988).

Hence, the methodology and techniques which have been developed to parameterize potential energy functions can and have been extended to parameterize molecular property surfaces

(Amos, 1987). By calculating the properties at each step of the MD simulation, in addition to the energy, one can then simulate various measured properties and other experimental observables that depend on these molecular properties and how they change as a function of both positions/displacements and momenta. In addition, one would like to perturb the system and observe the system "evolve" back to equilibrium, if indeed this occurs. In some cases, the perturbation may induce conformational changes and even chemical reactions. For example, the absorption of ultraviolet (UV) radiation by DNA or RNA often involves the formation of either cyclobutane or oxetane rings, damaging the DNA or RNA. One can determine the correlated changes that occur between various quantities as the molecules and molecular complexes not only change their positions but also change their linear and angular momenta. Many of these quantities also change and are correlated with each other (this occurs due to solvation, hydration, desolvation and ligand binding, and the presence of spectator molecules like the chaperon proteins). Here, one has the ability to do correct averaging (in principle), which is harder to do when one calculates only a finite number of local minima.

Depending on the barriers between the local minima, experimental conditions such as temperature, pressure, and presence of chaperon proteins, and/or other species, which may reduce the barriers, the species and conformers present may vary (Ertl, 2008). One assumes in many cases that the system is in thermal equilibrium and that small fluctuations do not perturb the system. In some cases, the complete phase space available to the molecule or molecular complex is sampled during the experiment (assuming the barriers are low enough). Here, one assumes that the system is in thermal equilibrium and that small fluctuations do not perturb the system, and so we do not need to take them, that is, the fluctuations in temperature (pressure), into account. In other cases, one may be trapped in a local minimum as the other local minima and even the global minimum are not accessible due to larger barriers. Hence one can measure the spectra and properties of a metastable state and think that one is in the global minimum due to the stability of the spectra and hence structure which give the unique and high resolution spectra. When one then perturbs the system and gives it enough energy to surmount the large barrier, the properties can then change drastically, which is thought to be the case for instance of the prion proteins. In this case, the global minimum of the individual molecules and the complexes (dimers, trimers, etc.) are not the same. For example, on dimerization fibrils form. The vibrational circular dichroism (VCD) and Raman optical activity (ROA) spectra of fibrils have been shown to reflect a large change from the nonfibril state. These spectroscopic techniques can hence be used to monitor such processes.

In some cases, one measures a superposition of the spectra of species present weighted by their Boltzmann populations, and in other cases, one actually measures an average spectrum, which may not even represent a real species/structure, but a sort of average structure. Hence, it is important to know both what one is measuring and one is simulating, and how to connect the two experiments, one in solution and one *in silico*. One would like to determine the structure of the species present and also conformers in the case of flexible molecules from just vibrational spectral measurements, similar to what one now does with the X-ray and NMR methods. Many proteins and other molecules, molecular complexes, and molecular assemblies have escaped X-ray and NMR structure determination and it would be remarkable if vibrational structure determination were developed. To date, one can determine the percentage of secondary structure elements present and the side chain torsion angles for specific residues, but not the complete three-dimensional structure of a protein. What one would really like is to determine the side chain and backbone torsion angles for each residue. One promising extension of normal vibrational spectroscopy includes isotopic substitution at each residue in the protein to shift the modes and hence make them separate from those for the other $N-1$ residues in the protein. This allows one in many cases to assign the backbone and side chain angles for only this specific residue, but not for the other $N-1$ residues in the protein, which makes this impractical as a general method.

Finally, in addition to the ground state potential energy and property surfaces, one is interested in various excited state energy and property surfaces, and of course, the transition rates and probabilities to jump from one surface to another and back, not only in the absence of radiation but also in the presence of coherent and incoherent electromagnetic radiation (Onida, Reining and Rubio, 2002; Farid, 2002; Farid, 1999). There are many cases in biological systems where a photon (quanta) of light is necessary to initiate a/the process, hence the term *quantum nanobiology*. One example is the repair of damaged DNA (due to UV radiation) in the presence of the photolyase protein and blue light (Jalkanen *et al.*, 2006a; Bohr, Jalkanen and Malik, 2005).

In this review, we present some of our contributions to the field of quantum nanobiology and also some main contributions from other researchers to this newly developing field and the outstanding problems that are yet to be addressed to make these simulations transparent and routine. This is done modestly by complete *ab initio* calculations of small biomolecules in various environments so as to derive structures, electronic and magnetic properties, and the various spectra of these molecules, aggregates, and molecular complexes. Hence, this area of research and study is a work in progress and each year more and more work is contributing to the field's development and evolution. The development of multidimensional infrared (IR) and Raman experiments, the analogs of two- and three-dimensional NMR, have recently been proposed (Cho, 2008; Scott, 2003; Jung, 2000; Mukamel, 1995), both of which in theory give more structural information than the simple one-dimensional spectra. What has limited the development and subsequent use of these techniques has been the development of a rigorous theory. The theories used to date involve many assumptions that have not been adequately tested, verified, and substantiated (Cho, 2008). Hence there is a need for the development of a rigorous high-level first principles theory and implementation, which can be used to test and benchmark the approximate classical, semiempirical, DFT, and *ab initio* methods. This is similar to the development of commercial VCD and ROA instruments. Until rigorous theories were developed, the commercial companies were unwilling to invest in commercializing instrumentation for which the analysis of the spectra was not considered rigorous and transparent. In addition, pattern recognition methods should be developed which can be used to go directly from the measured spectra to structure or possible structures and not to have to do a complete MD, BOMD (Born Oppenheimer Molecular Dynamics) or even non-BOMD simulation for each system studied. Hence the term *vibrational structure determination* like X-ray and NMR structure determination.

In the following, we describe and summarize several applications of quantum mechanics to small biomolecules. We review our own and other groups' restricted Hartree–Fock (RHF) and DFT studies for hydrated L-alanine (LA) (Jalkanen, Bohr and Suhai, 1997; Tajkhorshid, Jalkanen and Suhai, 1998; Frimand *et al.*, 2000; Jalkanen *et al.*, 2001, 2008a; Thar, Zhan and Kirchner, 2008), L-alanyl-L-alanine (LALA) (Knapp-Mohammady *et al.*, 1999; Jalkanen *et al.*, 2003), *N*-acetyl L-alanine *N*′-methylamide (NALANMA) (Jalkanen and Suhai, 1996; Deng *et al.*, 1996; Han *et al.*, 1998; Bohr *et al.*, 1999; Pohl *et al.*, 2007; Jalkanen *et al.*, 2008a), L-histidine (LH) (Bohr and Jalkanen, 2005; Jalkanen *et al.*, 2006a; Deplazes *et al.*, 2008), *N*-acetyl L-histidine *N*′-methylamide (NALHNMA) (Bohr and Jalkanen, 2005; Jalkanen *et al.*, 2006a), L-tryptophan (LW) (Jalkanen *et al.*, 2006a; Bombasaro, Rodriquez and Enriz, 2005), and capped L-alanine L-proline L-tyrosine (APY) [*N*-acetyl L-alanine L-proline L-tyrosine *N*′-methylamide (NALALPLYNMA) (Nardi, 1999; Nardi *et al.*, 2000; Jalkanen *et al.*, 2009c)] and examine the effects of water hydration on the structures, the vibrational frequencies, vibrational absorption (VA),VCD, Raman scattering, and ROA intensities. The large changes that hydration has on the structures, relative stability of conformers, and on the VA, VCD, Raman, and ROA spectra observed experimentally are qualitatively reproduced by our DFT calculations when one adequately treats the environment, which to date has been neglected in the literature where one has mostly tried to correlate features in the vibrational spectra of proteins and peptides with known X-ray or NMR structures (Zhu *et al.*, 2005; Barron *et al.*, 2000). The goal has been to decompose the VA, VCD, Raman, and ROA spectra into their principle components, so that these principle components are the characteristic spectra of either local minima or parts of the structures of local minima, which together make up the measured spectra. This is the premise in the attempts to deconvolute or partition the measured spectra into their principle components (Kubelka, Pancoska and Keiderling, 1999; Baello, Pancoska and Keiderling, 1997; Baumruk, Pancoska and Keiderling, 1996; Keiderling, 1996).

In the past, theoreticians have neglected to properly take into account the effects of the environment on stabilizing structures that were thought to be stable without environmental influences (Jalkanen, Bohr and Suhai, 1997; Tajkhorshid, Jalkanen and Suhai, 1998; Han *et al.*, 1998; Han *et al.*, 2000; Elstner *et al.*, 2000, 2001). Many of the routinely used molecular mechanics force fields have parameterized in local minima in the absence of solvent, where these local minima are only stable in the presence of the solvent (Jalkanen *et al.*, 2008a; Han *et al.*, 1998). Hence one must be careful in using molecular mechanics force fields; cases still exist where the various force fields give different results. To be fair though, this is also the case with many of the routinely used generalized gradient approximation (GGA) and hybrid exchange correlation functionals. The methods, which have been shown to be systematically correct, are the very expensive and rigorous *ab initio* DFT and coupled-cluster methods of the Bartlett group. The methods have been used to benchmark the more approximate and computationally faster methods: GGAs, hybrid XC functionals and, of course, the SCC-DFTB (Self Consistent Charge-Density Functional Tight Binding) and PM6 (Parametric Method 6) semiempirical DFT and WFQM methods.

In addition to the work on hydration of amino acids, peptides, and proteins via vibrational spectroscopy, there is a large body of related work on gas hydrates, both experimental and theoretical (van Klaveren *et al.*, 2001). The formation of gas hydrates in natural gas pipelines initiated a lot of fundamental work on characterizing and understanding the process of hydrate formation (Hammerschmidt, 1934). Furthermore, calorimetric methods have been used to determine the compositions, heat capacities, and enthalpies of dissociation for a number of hydrates (Handa, 1986a,b). MD simulation of N_2 and O_2 hydrates have been used to interpret the double maximum in large clathrate cages and single maxima in small clathrate cages for N_2 and double maximum in both cages for O_2 (Horikawa *et al.*, 1997). Here, one is able to use MD simulations to interpret the experimentally observed spectra of gases isolated in various cages of water molecules. Similar types of experiments and MD simulations can be used to understand the hydration of amino acids, peptides, and proteins.

Neutron and X-ray diffraction experiments have shown that the water molecules encapsulating peptides and proteins are different from bulk water (Sauter *et al.*, 2001). The question of the encapsulation of polar and nonpolar groups in amino acids, peptides, and proteins and the effects of hydration on the vibrational spectra, both frequencies and VA, VCD, Raman, and ROA intensities are very much related. To date, very little work has been done on analyzing the changes of VA, VCD, Raman, and ROA intensities due to the environment. This is due to the fact that large changes are observed because of conformational changes induced by the changing environment. Here, the two effects may not be able to be distinguished. Indeed, the work of Han *et al.* (1998) on NALANMA has shown that the species of interest may not be the isolated molecule, as it is in nonpolar solvents, in the gas phase or at low temperature in inert gas matrices, but the molecule and its bound water/solvent molecules, a molecular complex. The topology of the potential energy surface of the molecule appears to be very solvent/environment dependent, a fact that has not been appreciated in the past. Much of the complexity that appears in the gas phase appears to go away in strongly interacting polar media like water. But then the question is how does one determine the species of interest/study? This is especially important and relevant when one wishes to treat the systems in layers using various levels of theory, that is, different models for each layers.

The assumption that the properties of the molecular complex are a simple superposition or sum of the properties of the individual molecules (parts) may be in gross error. Hence the continued development of fields of complex systems and systems biology. Chaos theory is another example of a theory being developed which describes phenomena which cannot be explained using the standard model.

Finally, the treatment of solvent effects on crystallization and predicting the morphology of the crystals formed has recently been addressed by Gale and Rohl (2007). They have extended the COSMO (Conductor-Like Screening Model) model to treat solute molecules in solution to deal with solute molecules in a mother liquor at the crystal surface. Here they calculate both the surface energy and attachment energy in the presence of implicit solvent (Gay and Rohl, 1995). Previous work calculated these energies from the gas phase, which, though important, are not consistent with the media from which many crystals are grown. The molecules in the mother liquor can either adsorb on various crystal faces or desorb and go back into solution. MD simulations of this process are very time consuming and Gale and Rohl have developed a method that is highly efficient and uses classical potentials. The method can be extended to *ab initio* and DFT methods, but to date this has not been done.

In addition to changes in the vibrational spectra, the thermal conductivity of gas hydrates changes; it is much lower than that of ice and has been assigned due to coupling between the guest molecules and the water framework (Tse *et al.*, 2001; Inoue, Tanaka and Nakanishi, 1996). Changes in the vibrational spectra were also simulated and are further evidence for the coupling of the modes of the crystalline water cages and the embedded guest molecules. The low frequency modes have been observed in the incoherent inelastic neutron scattering (IINS) spectra, at frequencies which are very hard to observe in the IR and Raman spectra (Gutt *et al.*, 2002). Hence it is very useful to measure the IINS spectra for molecular hydrates and complexes, in addition to the VA, VCD, Raman, and ROA spectra (Oyama *et al.*, 2003; Subramanian and Sloan, 2002).

The IINS spectra for amides have also recently been analyzed with MD simulations (Itoh *et al.*, 2003). Early simulations of the IINS spectra for *N*-methylacetamide (NMA) using simple models have been shown to be in error (Fillaux *et al.*, 1993), even though they were consistent with the experimental data (Fillaux, 1999, 2004; Kearley *et al.*, 1994, 2001; Fillaux *et al.*, 1998; Kearley, 1986, 1995; Kearley, Tomkinson and Penfold, 1987). The models for the IINS reassigned the so-called amide A N–H stretch mode in NMA to be approximately half the value found in the isolated state in nonpolar solvents and in low temperature matrices. In the solid state, the NH hydrogen is hydrogen bonded with another C=O oxygen of an adjacent NMA molecule. The model used was a double-well potential and the model fit consistently all of the IINS data, so the work was accepted and caused quite a controversy in the conventional vibrational spectroscopy field, both experimental and theoretical. Subsequent work using a more rigorous *ab initio* model has shown that the initial assignment was right and the proposed new model, although able to reproduce the experimental data, was in error (Kearley *et al.*, 1994). This was due to the approximations and assumptions being made that are not strictly valid.

This emphasizes the fact that just because a simple model is able to reproduce a limited amount of experimental data does not mean it is correct. Hence the use of Ockham's razor does not always apply (it states that essentially one should use and accept the simplest possible explanation and/or model and/or theory). In this case, the more sophisticated model was necessary to get the physics correct . In some cases, one is happy to use a nonphysical model, as is quite often used in classical molecular mechanics and MD simulations that use harmonic potentials for bond stretches. But if one wants to make predictions and understand the physics of the phenomena, then one should strive to use models and theories which include the basic physics of the problem. Hence the term *ab initio* or *first principles theories*. But in many cases even the so-called first principles or *ab initio* theories involve approximations, the one most commonly made, but not explicitly stated being the BO approximation. There are many phenomena that cannot be described within the BO approximation, for example, the electronic contribution to the atomic axial tensors (AATs), which are needed to predict/simulate the VCD intensities.

Hence it is very important that the model is used to simulate as much of the available experimental data (Serdyuk, Zaccai and Zaccai, 2007). MD simulations have recently been extended to grow methyl hydrate sII from a supersaturated aqueous methane solution (Vatamanu and Kusalik, 2008). The next application is to develop inhibitors, as hydrate formation in natural gas pipelines is very problematic and much research has been undertaken to understand and to prevent it from occurring, including the use of proteins (Zeng *et al.*, 2003; Moon, Taylor and Rodger, 2003). Here the two fields merge. The multidisciplinary nature of many problems is finally being appreciated. Here researchers can really gain from working together and getting fresh new ideas and perspectives on their problems from researchers in other fields. In many cases, having to rigorously defend one's methods and approaches, and inherent assumptions, results in one's finding some of the weaknesses that do not get noticed when one just turns the crank. It is good every once in a while to analyze the gears and mechanism that go along with the crank.

Matrix isolation (Whittle, Dows and Pimentel, 1955; Pimentel, 1975), X-ray diffraction, and solid-state vibrational studies have been undertaken to try to determine relationships between structural parameters (both intermolecular, intramolecular, and intermolecular hydrogen bonding) and vibrational frequencies (VA, VCD, Raman, and ROA intensities) (Charles and Lee, 1965; Pimentel and Charles, 1963; Pimentel and Sederholm, 1956). Early Raman studies on lysozyme, ribonuclease, and α-chymotrypsin documented that vibrational studies could also contribute information to help understand the complexities of biomolecules in aqueous solution (Lord and Yu, 1970a,b).

In addition, one has also constructed software systems such as artificial neural networks to solve the inverse scattering problem of retrieving structural information of the biomolecule from spectroscopic data, that is, vibrational frequencies as well as VA and VCD intensities of isolated NALANMA (Bohr *et al.*, 2001; Ramnarayan, Bohr and Jalkanen, 2008).

Two possible routes have been considered to predict the structure of our test molecule from the spectroscopic data. One route is to use DFT to calculate all the possible structures/conformers and their corresponding frequencies, VA, VCD, Raman and ROA intensities and then compare them to the experimental data. The other is to train expert systems, for example, neural networks on a large combination of calculated and experimental spectra with known structures, either determined experimentally from X-ray or neutron diffraction or NMR. Then correlations and/or component spectra are used to infer or extrapolate new results, that is, to predict the structures, fold classes, and percentage of secondary structural elements present from only the measured VA, VCD, Raman, and ROA spectra.

In case of large biomolecules, one can determine whether there are correlations between the best-predicted structural details from spectroscopic data and the data connected to secondary structure stability and the structures predicted from neural network programs, which are based on only sequence and structural correlations. The main problem that prevented a complete protein structure to be determined from the combined VA, VCD, Raman, and ROA spectra has been to determine which spectroscopic data are the most important determiners of secondary structures, so that such information can be used to predict secondary structural elements, and then subsequently how these secondary elements fold up to form their tertiary and finally even quaternary structure. One assumes that the information is contained in the combined VA, VCD, Raman, and ROA spectra: frequencies, line widths, and integrated intensities; polarization properties of the intensities; and finally the frequency dependence of the Raman and ROA intensities. To date the methods are only able to predict percentages of various secondary structural elements, but not to assign the secondary structural elements to each and every residue in the protein.

Finally, in addition to the secondary structural element (ϕ_i and ψ_i angles) for each residue, one needs to determine the side chain torsion angles $\{\tau_{ij}\}$ and any disulfide or salt bridges connecting two nonadjacent residues (giving extra stability to this structure). In many cases, the protein folds up to its ‘‘native state’’ and is then posttranscriptionally modified/cleaved by an enzyme to achieve its functional state. Insulin is formed from the larger proinsulin by such a process. Hence, in many cases, the functionally active state of a protein does not correspond to any one gene or protein, but to a set of proteins/genes working in a concerted manner. Hence the term *systems biology*. One cannot really understand biological function by studying individual molecules or even individual molecules in their native environment. Once must know how the molecules interact with other molecules to do so. In the case of insulin, the ligand binds to the transmembrane insulin receptor extracellularly, which triggers dimerization and then a subsequent activation of intracellular tyrosine kinase activity, the so-called signal transduction process. This is as complex as its get, and one would like to be able to use spectroscopy to follow each and every step of this very complex process.

What has made this problem more difficult for biomolecules is the absorption of radiation by the environment, water being a strong absorber in the IR, and the strong interaction between the water molecules and the biomolecules via hydrogen bonding. Hence one must consider in their native environment, the biomolecules that interact with other biomolecules, counterions, cations, anions, and finally prosthetic groups and bound ligands, which in many cases are necessary for the biomolecules to function. On ligand binding, for so-called induced fit binding, both the protein and ligand can undergo conformational changes (Keskin *et al.*, 2008). Many of the current docking algorithms assume a lock- and key-type interaction, and hence fail to predict the binding of ligands via induced fit, and hence attempts have been made to improve and extend the current docking models (Böhm and Klebe, 1996; Davis and Teague, 1999; Cavasotto and Abagyan, 2004; Oliveira *et al.*, 2006).

The use of VA, VCD, Raman, and ROA spectroscopy are ideal tools to monitor such conformational changes in both the ligand and the protein on ligand binding. In addition, VCD and optical rotation have been shown to be sensitive enough to be able to observe the induced chirality in achiral solvent molecules and ligands on binding or even interacting with a chiral environment (Losada and Xu, 2007; Jalkanen *et al.*, 2008a; Mukhopadhyay *et al.*, 2007; Losada, Tran and Xu, 2008). The sesquiterpene molecules have multiple chiral centres and in many cases flexible side chains, which though achiral, have induced chirality due to the interactions with chiral cyclic system (Jalkanen *et al.*, 2008c; Kulcitki *et al.*, 2005; Weyerstahl *et al.*, 1995; Ayafor *et al.*, 1994; Newman, 1972).

The larger goal is to find/develop/derive ways to determine structural minima, not only for the individual molecules but also when they interact to form ligand–protein complexes. In addition, in many cases there are other participants in the stable and functional biological unit, a metal center, bound solvent molecules, prosthetic groups like the porphyrin ring in myoglobin and hemoglobin, and finally, cations or anions (Sauter *et al.*, 2001). For these minimum energy stable complexes, VA, VCD, Raman, and ROA intensities are calculated by DFT in order to produce training data, that is, sets of spectroscopic data correlated with $\phi - \psi$ angles and other structural determiners for the network. Other methods such as X-ray crystallography and NMR have only been utilized to determine the native states of proteins. The VA, VCD, Raman, and ROA spectra for proteins and peptides with known structures should also be added and be part of the training set. By using DFT methods, the VA, VCD, Raman, and ROA spectra of the so-called denatured states of proteins and peptides can also be added to the training set. Many of the so-called random coil and denatured states have been shown to have a lot of secondary structure in the CD, VCD, and ROA spectra. All that one can say for sure is that these structures are different from the so-called minimum energy structures under physiological conditions. In some cases, these structures may even be global minima and the structures which are biologically active may actually be in metastable states. For example, the prion proteins may be lower in energy in the so-called denatured state, thereby making attempts to refold the prion proteins very difficult. In addition, it has recently been proposed that the functional states of some proteins are only present when the ligand is bound and the so-called native isolated protein is not one conformer (global deep minimum energy state), but an ensemble of structures (Keskin *et al.*, 2008; Sauter *et al.*, 2001). Only on ligand binding or in complex with another protein, does a single unique functional conformer arise. This could account for the failure of X-ray and NMR techniques in determining the unbound state(s)and structure(s) of this class of proteins.

One of the main problems is to determine the structures from the VA, VCD, Raman, and ROA spectra of the denatured states of proteins. Keiderling and coworkers have utilized neural network methodology to find correlations of electronic absorption (EA), electronic circular dichroism (ECD), VA, and VCD spectra with the native states of proteins by utilizing the known NMR and X-ray crystallographic structures (Pancoska, Yasui and Keiderling, 1989, 1991; Pancoska, Blazek and Keiderling, 1992; Pancoska *et al.*, 1994, 1995; Pancoska, Janota and Keiderling, 1996; Baumruk, Pancoska and Keiderling, 1996; Baello, Pancoska and Keiderling, 1997; Pancoska, Kubelka and Keiderling, 1999; Kubelka, Pancoska and Keiderling, 1999). Our work compliments their work in providing correlations of VA, VCD, Raman, and ROA spectra and the higher energy denatured states of peptides and proteins. These denatured states can be produced under various experimental conditions, that is, in aqueous solution, for example, under various pHs and salt conditions and by the presence of urea and other denaturing conditions.

In that sense one should be able to predict higher level intermediate energy states during folding processes of biomolecules with the help of expert software systems such as artificial neural networks and principle component analyses that have to be trained on known data in order to predict structural data novel to the networks. Once they are trained on known sets of intermediate energy states, that is, the set of structural data we present to the network they should be able to extrapolate to new relations between spectrum and structure. The advantage of utilizing software systems such as neural network techniques and principle component analyses for the inverse scattering problem of deriving structural information from scattering data is that they go hand in hand with experiments and DFT calculations in the sense that one of the tools can support the other when it fails or is impractical. The big endeavor is to produce detailed structural data of proteins and peptides, both in the isolated state and in complex with each other and other proteins, and either measure the experimental VA, VCD, Raman, and ROA spectra and/or simulate them with DFT calculations. In many cases the spectra and structures of proteins can be assembled from the principle components, that is, small peptides that constitute the whole protein.

In many cases, the diverse conformational spectrum of small peptides is greatly reduced when they form part of a protein. Hence one may not need to determine the structures and spectra for all possible conformers of small peptides, which itself is a very challenging problem. But like the protein not sampling all possible conformational states, peptides may and probably do not sample all possible states either. This has also been confirmed via MD simulations. Hence the many groups and researchers working on methods to sample the complete conformational space for proteins and peptides, though interesting from an academic point of view, may not really be necessary. This may be necessary to do once for each amino acid and then for sequence of dipeptide, tripeptide, and so on, but is really a waste of time and resources to do again and again for each protein and/or for each new study/simulation. What is required is to assemble/collect the results from various laboratories and annotate and verify the data before making it available to the research community. This is similar to what has been done for both experimental X-ray and NMR structures. What is needed is a similar database for theoretically determined structures. In addition, what is important for the molecular biophysicists, molecular biologists, and quantum nanobiologists to understand are the biophysical forces and interactions that are responsible for not only stabilizing the conformers, structures, and assemblies but also the biophysical processes and events which are responsible for function, and dynamics. In the opinion of these authors, too much emphasis has been put on static structures and not enough on dynamics and function. In many cases, the dynamical fluctuations and function go hand in hand. Vibrational spectroscopies are ideal techniques to study and investigate both dynamics and function.

If biomolecules and biomolecular complexes only sample a small part of the total available phase and conformational space possible, then this is what one would like to simulate, and not large portions of phase space which are not sampled and/or relevant for the experiments being studied. Part of this question can be addressed with, for example, neural network and principle component analyses tools determining the dihedral angles in the protein that are actually populated from the VA, VCD, Raman, and ROA data for each subunit. If one is able to achieve this goal, then the combination of VA, VCD, Raman, and ROA spectroscopy will take a seat along with X-ray and neutron diffraction and NMR spectroscopies as standard tools for structure determination of proteins and peptides.

A library of various conformers for each amino acid along with its corresponding properties as a function of various environments (aqueous, nonpolar membrane, or mixed) would be very beneficial, similar to the current protein data base (PDB) for X-ray and NMR-determined structures and the Cambridge crystallographic database for small molecules and ligands. Of course, along with amino acids, its corresponding capped amino acid, and sequences of dipeptides, tripeptides, and so on, data of dipeptides, tripeptides, and so on should also be collected/assembled and most certainly annotated. Many experiments and theoretical simulations have not been able to be reproduced (a fundamental part of good science) due to the lack of being adequately annotated/documented. A simple case is of whether one used distilled water or distilled *and* deionized water *and* whether one carefully controlled the pH with a buffer solution and then, which buffer, HEPES (N-2 -Hydroxyethyl-piperazine-N^1-2-Ethane-sulfonic Acid), phosphate, and so on. Some effects due to different buffering solutions have even been observed. The regrettable fact is that, in many experiments, everything seems to matter, and hence needs to be properly documented and controlled.

This large (and ever increasing) amount of data could then be used either in the parameterization and/or testing of new molecular mechanics and semiempirical methods, and even newly developed XC functionals. It is important to continually test and retest and, if necessary, reparameterize and further develop and extend methods as problems and/or disagreements arise and are found. A classical example has been the so-called standard hybrid XC functional, the B3LYP, now being replaced by a variety of more accurate and general XC functionals, one such being the B3PW91.

In the next sections, we present our theoretical approach to simulate the VA, VCD, Raman, and ROA spectra of proteins and peptides. First, we present an introduction to DFT methods. Then a presentation is given on how one calculates/simulates VA and VCD spectra followed by the same for Raman and ROA spectra. The last section covers the use of VA/VCD/Raman/ROA experimental and theoretical spectroscopy in molecular biophysics, molecular biology, and quantum nanobiology.

1.2
Introduction to Density Functional Theory methods

DFT is in principle an *ab initio* method that allows one to calculate the structures and properties of single isolated molecules and various aggregated complexes (Hohenberg and Kohn, 1964; Mermin, 1965; Kohn and Sham, 1965; Pople, 1999; Kohn, 1999; Bartlett *et al.*, 2005; Bartlett, Lotrich and Schwiegert, 2005). It is more complete than Hartree–Fock methodology in that it includes electron correlation and is better than other correlated methods because it requires less computational resources (disk, memory, and CPU time) than otherwise, for the same computational complexity. DFT formulates the energy of the system in terms of the electron density and electron current density instead of the wave function. The trade-off is that one must determine the various terms in the Hamiltonian for the total energy of the system, the kinetic energy (nuclear and electron) and Coulomb energy (nuclear–nuclear, nuclear–electron, and electron–electron)in terms of the electron density, current density, and perhaps the gradients and higher order spatial derivatives of these so-called observables. Since these quantities are themselves functions of functions, they are called functionals. Various local and nonlocal electron density functionals and local electron current density functionals have been determined and tested in their ability to determine structures, relative energies, binding energies, dipole moments, polarizabilities, hyperpolarizabilities, dipole moment derivatives, NMR shielding tensors, and various other molecular and aggregate properties (Ma and Brueckner, 1968; Rajagopal and Callaway, 1973; Perdew and Levy, 1983; Sham and Schlüter, 1985; Lannoo, Schlüter and Sham, 1985; Vignale and Rasolt, 1987, 1988; Godby, Schlüter and Sham, 1988; Becke, 1988a, 1993a; Lee, Yang and Parr, 1988; Levine and Allan, 1989; Papai *et al.*, 1990; Perdew and Wang, 1992; Malkin, Malkina and Salahub, 1993a,b; Komornicki and Fitzgerald, 1993; Colwell *et al.*, 1993; Lee and Sosa, 1994; Johnson and Frisch, 1994; Schreckenback and Ziegler, 1995; Malkin *et al.*, 1995; Wilson, Amos and Handy, 1999; Wolff, 2005; Kongsted and Ruud, 2008).

The first Hohenberg–Kohn theorem showed that the electron density determines the energy and hence reformulated the basic equation to solve as one in which one has to determine the electron density rather than the wave function (Hohenberg and Kohn, 1964). The energy function can be written as

$$E_v[\rho] = T[\rho] + V_{\mathrm{ne}}[\rho] + V_{\mathrm{ee}}[\rho] \qquad (1)$$

$$= \int \rho(\mathbf{r})v(\mathbf{r})d\mathbf{r} + F_{\mathrm{HK}}[\rho] \qquad (2)$$

where

$$F_{\mathrm{HK}}[\rho] = T[\rho] + V_{\mathrm{ee}}[\rho] \qquad (3)$$

and $v(\mathbf{r})$ is the external potential.

The second Hohenberg–Kohn theorem provides the energy variational principle, which enables one to find the density that minimizes this energy functional (Kohn and Sham, 1965). The problem is we do not know the functional $F_{\mathrm{HK}}[\rho]$ and many functionals have been developed that try to address this problem. By introducing the Kohn–Sham orbitals, Kohn and Sham were able to address the problem of not being able to determine the electron kinetic energy functional in terms of only the density. At this point, they could have also used the Kohn–Sham orbitals to determine the exchange energy functional by simply substituting now the Kohn–Sham orbitals within a Slater determinant. But by doing this, they would have introduced a nonlocal exchange term and the simplicity and robustness of DFT would not have been apparent. So instead of taking this route, which would have been to search for the correlation energy functional also in terms of the Kohn–Sham orbitals, Kohn and Sham explored using the local density approximation (LDA) for both the exchange and correlation energy functionals. For problems in solid-state physics, LDA works well and has been used for decades now. Owing to problems with scaling in WFQM, quantum chemists turned to DFT as an alternative to the WFQM semiempirical based methods. For some problems LDA did fine, but for complex organic and bioorganic molecules, especially those involving hydrogen bonding, one requires higher accuracy than LDA can give. This led to the development of the so-called GGA functionals, where in addition to the density, the exchange and correlation energy functionals are functions of the gradient of the density, in addition to the density (Perdew, Burke and Ernyerhoh, 1996, 1997). Still higher accuracy was required, and to get this higher accuracy one went to the nonlocal Hartree–Fock form for the exchange energy functional, substituting the Kohn–Sham orbitals into the expressions instead of the Hartree–Fock orbitals. This led to the so-called hybrid XC functionals.

Commonly used hybrid functionals, the Becke 3LYP (B3LYP) (Becke, 1988a, 1993a; Lee, Yang and Parr, 1988; Perdew and Wang, 1992),

$$E[\rho(x, y, z), \nabla(\rho(x, y, z))] = E_{\mathrm{RB}} + E_{\mathrm{HF}} - E_{\mathrm{LYP}} \qquad (4)$$

and Becke 3PW91 (B3PW91)

$$E[\rho(x, y, z), \nabla(\rho(x, y, z))] = E_{\mathrm{RB}} + E_{\mathrm{HF}} - E_{\mathrm{PW91}} \qquad (5)$$

hybrid exchange correlation functionals have been implemented in CADPAC, Dalton, Turbomole, and Gaussian 94, 98 and 2003 (Stephens *et al.*, 1994; Lee and Sosa, 1994; Becke, 1993a; Lee, Yang and Parr, 1988). B3LYP and B3PW91 level analytical Hessian and atomic polar tensors (APT) calculations have also been implemented in Gaussian 94, 98, 03, CADPAC, and Dalton. More recently, the AAT have been implemented in the ADF suite of programs utilizing Slater-type orbitals (STOs) (Nicu *et al.*, 2008). Previous implementations had all been with the Gaussian-type orbitals (GTOs), which have been used by quantum chemists, since these allow one to analytically evaluate many integrals, in contrast to the use of numerical integrations to evaluate many of the integrals when using STOs. But with the use of numerical integration to evaluate many of the exchange and correlation integrals required for various XC functionals, the necessity and advantage of using strictly GTOs no longer exists, and hence the return to the use of STOs.

The B3LYP and B3PW91 level force fields have been shown to be more accurate than RHF level Hessians, which must be scaled to get good agreement with both experimental frequencies and VA and VCD intensities (Jalkanen and Stephens, 1991; Stephens *et al.*, 1994; Jalkanen and Suhai, 1996; Jalkanen *et al.*, 2008a,b; Deplazes *et al.*, 2008). The nature of the normal modes have been shown to depend very much on the scaling scheme one chooses to scale the Hessian (Jalkanen *et al.*, 1987, 2008b; Jalkanen and Stephens, 1991). The advantage of the B3LYP and B3PW91 hybrid exchange correlation functionals is that the Hessians appear to be accurate enough to predict the VA and VCD intensities when coupled with accurate APT and distributed origin (DO) gauge and gauge invariant atomic orbital (GIAO) AAT without scaling. The number of molecules for which the B3LYP and B3PW91 Hessians, APT and AAT, have been calculated and used to predict and help interpret the VA and VCD spectra has been quite extensive now. This level of field has became the standard where either B3LYP or B3PW91 are able to predict the correct structure (conformer) and relative energies. The good agreement shown to date has included only a small number of functional groups and the comparison has been with measurements of the VA and VCD spectra of molecules in nonpolar solvents. Recently, systematic errors with B3LYP have been found (Matsuda, Wilson and Xiong, 2006), and hence the need for a more systematic approach to develop XC functionals. The *ab initio* DFT method of Bartlett, which uses optimized effective potentials (OEPs) for both exchange and correlation (XC), appears to be the most promising (Bartlett *et al.*, 2005; Bartlett, Lotrich and Schwiegert, 2005). The work of Bartlett built on previous works with OEPS is covered briefly and cited in the classical quantum chemistry text by Levine (2000).

Even though much more expensive than the GGAs like PBE, OEPs give the correct answer for the correct reasons (correct physics and chemistry), and hence the additional effort and cost is justified. Therefore, the OEP XC functionals of the Bartlett group have become the new gold standard and, all the more, simple, approximate, and robust XC functionals should be tested and benchmarked against the Bartlett OEPs. With the work of Bartlett and coworkers, DFT has now been put on as a rigorous level as WFQM, with a systematic way to improve both the XC functionals achievable, though at much higher cost and complexity than originally formulated first by Thomas, then Thomas and Fermi, and finally by Hohenberg, Kohn, and Sham. The merging and integration of the works by Kohn and Pople and many others in DFT and WFQM appears to have finally been achieved. Now one can use DFT and WFQM and semiempirical methods based on each almost interchangeably, depending on the focus of one's work. Each method, as well as Greens functions-based methods, has its advantages and disadvantages, especially with respect to the treatment of excited electronic states and excited state potential energy surfaces, which are required for photon-induced photochemical reactions. Many reactions in molecular biology and molecular biophysics are photon initiated, so there is plenty of work to be done to further develop and refine DFT-based methods for electronic excited state MD. In this realm, WFQM and Greens functions-based methods are more accurate and should be used to test and benchmark the time-dependent density functional theory (TD-DFT)- based methods (Onida, Reining and Rubio, 2002; Reining *et al.*, 2002; Ismail-Beigi and Louie, 2003; Zhukov, Chulkov and Echenique, 2004).

In this review, we summarize previous works and present some new results on some small peptides (LALA, NALANMA, and NALHNMA) and two amino acids (LA and LH). LA and LH are two clear examples of a biological molecules whose properties and structures in water and gas phase are very different. We have found that without the water being explicitly or implicitly accounted for, the gas phase or isolated molecule calculations are of questionable use in analyzing the measurements on these molecules in aqueous solution, though they have found recent application in mass spectroscopic studies (Ipolyi *et al.*, 2006). The zwitterionic structure of LA, the predominant species in aqueous solution, is not stable in the gas phase according to our 6-31G* B3LYP *ab initio* calculations (Jalkanen, Bohr and Suhai, 1997; Tajkhorshid, Jalkanen and Suhai, 1998). Therefore we have explicitly included solvent, in this case water, to stabilize this structure and also to be able to compare to the measurements of the VA, VCD, Raman, and ROA spectra of LA in aqueous solution (Tajkhorshid, Jalkanen and Suhai, 1998; Frimand *et al.*, 2000; Jalkanen *et al.*, 2001, 2008a). The VA and VCD spectra of LA in aqueous solution were initially measured by the groups of M. Diem at Hunter College in NY, NY (Diem, 1988) and L. Nafie at Syracuse University in Syracuse, NY (Freedman *et al.*, 1988). Raman and ROA measurements and calculations on LA have also been reported by Barron and coworkers (Barron *et al.*, 1991a). At our *ab initio* 6-31G* B3LYP DFT-optimized geometries we have calculated the vibrational frequencies, VA and VCD intensities. The VA and VCD spectra were calculated with the 6-31G* B3LYP force field (Hessian) and APT and the 6-31G** RHF DO gauge AAT.

For LALA, three zwitterionic structures were able to be stabilized with four, six, and seven water molecules, respectively (Jalkanen, Bohr and Suhai, 1997; Bohr and Jalkanen, 1998). The zwitterionic structure was not stable relative to proton transfer of one of the protons of the NH_3^+ group to the C=O group of the adjacent amide group, whereas the proton of the N−H group was concurrently transferred to the closest oxygen of the CO_2^- group. Subsequently completely solvated (hydrated) structures of the LALA zwitterion were extracted from MD simulations and subsequently optimized at the B3LYP level of theory (Knapp-Mohammady *et al.*, 1999). Here only the water molecules directly hydrogen bonded with the polar groups were kept and the number of waters varied from 10 to 14, depending on the conformation of the LALA zwitterion. In a subsequent study, we used one of the structures and predicted the VA, VCD, Raman, and ROA spectra (Jalkanen *et al.*, 2003). Here we present our new results for the LALA plus 11 water clusters embedded within the polarizable continuum model (PCM). We present the VA, VCD, Raman, and ROA spectra for the LALA zwitterion with 11 water molecules. As one can see the water absorption masks many of the LALA modes. Note that in addition to masking the LALA modes, the water modes couple with the LALA modes and therefore the water molecules in the first solvation shell surrounding the LALA zwitterion are different than those in bulk water. This has been documented by X-ray and neutron diffraction studies as well as from VCD signals due to chiral water molecules (the chirality is due to the chiral environment of the LALA molecule) (Sauter *et al.*, 2001; Losada and Xu, 2007; Losada, Tran and Xu, 2008).

The goal there was and here is to determine the structure (a conformer) of the zwitterionic form of LALA in aqueous solution from the experimental vibrational spectra, VA, VCD, Raman, and ROA, if possible. As shown by our calculations on the neutral species, there are many possible conformers for this molecule and the structure(s) of this molecule in water has(have) not yet been determined, though the VA, VCD, Raman, and ROA spectra have been measured. The goal of our work is to help answer the question of the conformation and species of LALA in aqueous solution and to develop and document a methodology which will work for larger peptides and even proteins in both native and nonnative conditions.

We also present here optimized structures of NALANMA with four waters starting from our 6-31G* B3LYP optimized structures. The relative energies of these complexes are compared with the isolated molecule values. In addition we explore the effect of changing the hybrid exchange correlation functional, as recently the B3LYP hybrid exchange correlation functional has been called into question. Hence we have replaced the LYP correlation functional with the PW91 correlation functional and have optimized the structure of NALANMA with the four bound waters embedded within the PCM and conductor-like polarizable continnum model (CPCM) continuum solvent models with the B3PW91 hybrid exchange correlation functional (Jalkanen *et al.*, 2008a).

The goal here has been to model biomolecules by explicitly adding water molecules to provide calculations which can be used to critically evaluate solvent models and specific models developed for water (liquid simulations) (Jorgensen, Madura and Swenson, 1984; Jorgensen and Swenson, 1985a,b; Jorgensen and Tirado-Rives, 1988; Jorgensen, Maxwell and Tirado-Rives, 1996). The H-bonding properties as exemplified by some of the simple water models are clearly incorrect, and we feel that conclusions based on these models can be critically evaluated utilizing the better models of water (Berendsen *et al.*, 1981). In addition, the properties for the water potentials that are used in parameterization may not necessarily be those one is interested in. For models which do not include effects due to polarization, there is no change in the electronic charge density on hydrogen bonding, which is known to occur. Hence whenever one uses a classical potential, it should be benchmarked for the properties of interest. In many cases, a more refined model must be used for more sophisticated and elaborate experiments, for example, the vibrational and EA and CD spectra. In addition, since many models using parameters reduce the error between measured and simulated results, care must be taken to ensure that the parameters do not appear to correct the physics and chemistry that is *lacking* in the model. For example, to use parameters to "make" harmonic frequencies anharmonic. By doing so, other real anharmonic effects are lost.

Various models have been developed for implicitly and explicitly taking into account water at various levels (Onsager, 1936; Richmond, 1984; Wesson and Eisenberg, 1992; Schiffer *et al.*, 1992; Fortunelli and Tomasi, 1994; Schmidt and Fine, 1994; Tomasi and Persico, 1994; Fortunelli, 1996; Cossi *et al.*, 1996; Tawa *et al.*, 1996; Ösapay *et al.*, 1996; Craw *et al.*, 1996; Rauhut *et al.*, 1996; Mok, Nwumann and Handy, 1996; Jalkanen and Suhai, 1996, 1998; Jorgensen, Maxwell and Tirado-Rives, 1996; Tajkhorshid, Han *et al.*, 1998; Knapp-Mohammady *et al.*, 1999; Frimand *et al.*, 2000; Jalkanen *et al.*, 2001, 2003, 2006a; Tomasi, Mennucci and Cammi, 2005; Mantz *et al.*, 2006). At the molecular mechanics level, the force field can be parameterized against experimental data measured on the molecule in the aqueous solution. The force field is then an effective force field in that it is not useful for doing calculations on the molecule in the gas phase or for other solvents, the optimized potentials for liquid simulation (OPLS) force field being the most specific of this type (Jorgensen, Madura and Swenson, 1984; Jorgensen and Swenson, 1985a,b; Jorgensen and Tirado-Rives, 1988; Jorgensen, Maxwell and Tirado-Rives, 1996). Another approach to the solvent and/or hydration problem is to use a distance-dependent dielectric or some other perturbation to the gas phase potential energy surface and/or interactions to take into account the effect of water without actually adding explicit waters. The goal here is to save computational time and resources because the addition of explicit water molecules adds to the length of the calculation and the multiple minimum questions become exponentially worse when one tries to find the "global minimum" of the molecule solvated by water molecules. Finally the question of temperature, and hence entropic effects are very important when modelling hydrated systems, as there are many possible orientations of the water molecules encapsulating and/or hydrating the molecule.

The hydrated structures presented here for LA, NALANMA, LALA, LH, and NALHNMA can be used to test the various water models before one uses them in expensive molecular dynamic simulations on proteins and nucleic acids. The work is a part of our collaborative work involving groups at the German Cancer Research Center, the Technical University of Denmark, Helsinki University of Technology, Curtin University of Technology, and the University of Bremen to model proteins and nucleic acids along with various ligands in the presence of water and denaturing environments. In the next sections, we present the methodology to simulate the vibrational spectrum, not only the frequencies but also the intensities, followed by the use of the spectroscopic methods in molecular biology, molecular biophysics, and quantum nanobiology.

1.3
Calculating Vibrational Absorption and Vibrational Circular Dichroism Spectra

1.3.1
Vibrational Absorption (VA) or Infrared (IR) Absorption and Vibrational Circular Dichroism

Vibrational absorption and CD spectra are related to molecular dipole and rotational strengths via,

$$\varepsilon(\bar{\nu})=\frac{8\pi^3 N_A}{3000hc(2.303)}\sum_i \bar{\nu} D_i f_i(\bar{\nu}_i,\bar{\nu}) \tag{6}$$

$$\Delta\varepsilon(\bar{\nu})=\frac{32\pi^3 N}{3000hc(2.303)}\sum_i \bar{\nu} R_i f_i(\bar{\nu}_i,\bar{\nu})$$

where ε and $\Delta\varepsilon = \varepsilon_L - \varepsilon_R$ are molar extinction and differential extinction coefficients, respectively, D_i and R_i are the dipole and rotational strengths of the ith transition of wavenumbers $\bar{\nu}_i$ in cm^{-1}, and $f(\bar{\nu}_i,\bar{\nu})$ is a normalized line-shape function and N_A is Avogadro's number. For a fundamental (0 → 1) transition involving the ith normal mode within the harmonic approximation

$$D_i=\left(\frac{\hbar}{2\omega_i}\right)\sum_\beta\left\{\sum_{\lambda\alpha}S_{\lambda\alpha,i}P^\lambda_{\alpha\beta}\right\}\times\left\{\sum_{\lambda'\alpha'}S_{\lambda'\alpha',i}P^{\lambda'}_{\alpha'\beta}\right\} \tag{7}$$

$$R_i=\hbar^2\,\mathrm{Im}\sum_\beta\left\{\sum_{\lambda\alpha}S_{\lambda\alpha,i}P^\lambda_{\alpha\beta}\right\}\times\left\{\sum_{\lambda'\alpha'}S_{\lambda'\alpha',i}M^{\lambda'}_{\alpha'\beta}\right\}$$

where $\hbar\omega_i$ is the energy of the ith normal mode, the $S_{\lambda\alpha,i}$ matrix interrelates normal coordinates Q_i to Cartesian displacement coordinates $X_{\lambda\alpha}$, where λ specifies a nucleus and $\alpha = x$, y, or z:

$$X_{\lambda\alpha}=\sum_i S_{\lambda\alpha,i}Q_i \tag{8}$$

$P^\lambda_{\alpha\beta}$ and $M^\lambda_{\alpha\beta}$ (α, $\beta = x, y, z$) are the APT and AAT of nucleus λ. $P^\lambda_{\alpha\beta}$ is defined by

$$P^\lambda_{\alpha\beta}=\left\{\frac{\partial}{\partial X_{\lambda\alpha}}\langle\psi_G(\mathbf{R})|(\mu_{\text{el}})_\beta|\psi_G(\mathbf{R})\rangle\right\}_{\mathbf{R}_o} \tag{9}$$

$$=2\left\langle\left(\frac{\partial\psi_G(\mathbf{R})}{\partial X_{\lambda\alpha}}\right)_{\mathbf{R}_o}\middle|\left(\mu^e_{\text{el}}\right)_\beta\middle|\psi_G(\mathbf{R}_o)\right\rangle + Z_\lambda e\delta_{\alpha\beta}$$

where $\psi_G(\mathbf{R})$ is the electronic wavefunction of the ground state G, $\mathbf{R}$ specifies nuclear coordinates, $\mathbf{R}_o$ specifies the equilibrium geometry, $\boldsymbol{\mu}_{\text{el}}$ is the electric dipole moment operator, $\bar{\mu}^e_{\text{el}} = -e\sum_i \mathbf{r}_i$ is the electronic contribution to $\bar{\mu}_{\text{el}}$, and $Z_\lambda e$ is the charge on nucleus λ and $M^\lambda_{\alpha\beta}$ is given by

$$M^\lambda_{\alpha\beta}=I^\lambda_{\alpha\beta}+\frac{\mathrm{i}}{4\hbar c}\sum_\gamma\varepsilon_{\alpha\beta\gamma}R^o_{\lambda\gamma}(Z_\lambda e)$$

$$I^\lambda_{\alpha\beta}=\left\langle\left(\frac{\partial\psi_G(\mathbf{R})}{\partial X_{\lambda\alpha}}\right)_{\mathbf{R}_e}\middle|\left(\frac{\partial\psi_G(\mathbf{R}_e,B_\beta)}{\partial B_\beta}\right)_{B_\beta=0}\right\rangle \tag{10}$$

where $\psi_G(\mathbf{R}_o, B_\beta)$ is the ground state electronic wavefunction in the equilibrium structure $\mathbf{R}_e$ in the presence of the perturbation $-(\mu^e_{\mathrm{mag}})_\beta B_\beta$, where $\boldsymbol{\mu}^e_{\mathrm{mag}}$ is the electronic contribution to the magnetic dipole moment operator. $M^\lambda_{\alpha\beta}$ is origin dependent. Its origin dependence is given by

$$(M^\lambda_{\alpha\beta})^0 = (M^\lambda_{\alpha\beta})^{0'} + \frac{\mathrm{i}}{4\hbar c}\sum_{\gamma\delta} \varepsilon_{\beta\gamma\delta} Y^\lambda_\gamma P^\lambda_{\delta\alpha} \tag{11}$$

where $\mathbf{Y}^\lambda$ is the vector from O to O′ for the tensor of nucleus λ. Equation (11) permits alternative gauges in the calculation of the set of $(M^\lambda_{\alpha\beta})^0$ tensors. If $\mathbf{Y}^\lambda = 0$, and hence $\mathrm{O} = \mathrm{O}'$, for all λ the gauge is termed the *common origin (CO) gauge*. If $\mathbf{Y}^\lambda = \mathbf{R}^o_\lambda$, so that in the calculation of $(M^\lambda_{\alpha\beta})^0$ O′ is placed at the equilibrium position of nucleus λ, the gauge is termed the *DO gauge* (Stephens, 1985, 1987; Buckingham, Fowler and Galwas, 1987; Amos *et al.*, 1987; Amos, Jalkanen and Stephens, 1988).

The problem with the expressions for the APT and AAT presented initially is that they involved the wavefunction and wavefunction derivatives (usually calculated with Coupled Hartree–Fock theory). These expressions were implemented in CADPAC to calculate the APT at the self-consistent field (SCF) and MP2 levels and the AAT at the SCF level. The first rigorous VA and VCD spectral simulations were performed using the Gaussian and CADPAC programs (Stephens and Lowe, 1985; Amos *et al.*, 1987; Amos, Jalkanen and Stephens, 1988, 1991). The scaled quantum mechanical (SQM) force fields (Lowe *et al.*, 1986; Jalkanen *et al.*, 1987, 1988; Kawiecki *et al.*, 1991) were initially scaled using the scaling methods of Blom and Altona (1976) and Pulay *et al.* (1979). Later more accurate MP2 force fields appeared to be accurate enough, when combined with MP2 APT and RHF/SCF AAT tensors (Stephens *et al.*, 1993). This gave qualitative accuracy in general, and in some cases even quantitative accuracy, though there were cases which were problematic, one being *trans*-1,2-dicyanocyclopropane, which is now still considered problematic and a good test case for any new method (Jalkanen *et al.*, 2008b). Subsequently, the AAT have been implemented with GIAOs at both the SCF and multiconfiguration self-consistent field (MCSCF) levels of theory within the Dalton suite of programs (Bak *et al.*, 1993, 1994, 1995).

The advantages of DFT are numerous, but the main one is to extend rigorous methodology to the calculation of properties of large biological molecules. DFT seems to provide a way to do that. But the expressions must be reformulated as expressions amenable for DFT, where one has the density expressed in terms of the so-called Kohn–Sham orbitals and not in terms of rigorous wave functions. The Kohn–Sham orbitals are only used to generate the density, and one must be very careful in applying and using the orbitals and eigenvalues outside the scope for which they were intended.

The calculation of the DFT AATs has been implemented within the coupled perturbed Kohn–Sham formalism using GTOs in CADPAC, Dalton, and Gaussian (Cheesman *et al.*, 1996) and using STOs in the ADF program (Nicu *et al.*, 2008; Wolff, 2005; te Velde *et al.*, 2001). For many cases, SCF AAT appear to be of sufficient accuracy, but with the prevalence of implementations of DFT, DFT methodology appears to be taking over. A comparison of rotational strengths for molecules with both RHF and DFT AAT is still important, as in many cases one can use a different method for the tensors required for the VA and VCD (and also Raman and ROA) intensities, than one requires for the geometry optimization and subsequent Hessian and normal mode calculation. The implementation of VA and VCD spectral simulation in Gaussian, Dalton, CADPAC, and ADF all allow one to accurately predict the VA and VCD spectra for molecules in nonpolar solvents, in inert low temperature nobel gas and N_2 matrices, and in membranes where there is no strong interaction between the solute molecule and the solvent/environment.

Recently, various continuum models have been implemented in Gaussian, Dalton, and ADF that allow not only geometry optimizations but also analytical derivative methods for the calculation of the Hessian, APT and AAT, required to simulate the VA and VCD spectra. The continuum solvent models have required continuous refinements and extensions as one has started to use them to treat biolological systems, where many of the assumptions inherent in the development of the models break down. One of these is that there are no specific directional effects which must be treated. Hydrogen bonding effects have been shown to be inadequately treated, and hybrid models have been developed to treat these systems (Jalkanen and Suhai, 1996, 1998; Jalkanen, Bohr and Suhai, 1997; Han *et al.*, 1998; Tajkhorshid, Jalkanen and Suhai, 1998; Knapp-Mohammady *et al.*, 1999; Jalkanen *et al.*, 2008a). Especially when one attempts to treat hydrogen-bonded systems with implicit solvent models one must be very careful. One cannot treat an explicit directional and specific interaction between the charged NH_3^+ and CO_2^- groups in an amino acid, peptide and protein with solvent/water molecules, cations, or anions with an implicit continuum solvent model. Here we have introduced our embedded molecular (hydrogen-bonded) complex model (Jalkanen and Suhai, 1996; Jalkanen, Bohr and Suhai, 1997; Han *et al.*, 1998; Tajkhorshid, Jalkanen and Suhai, 1998; Knapp-Mohammady *et al.*, 1999; Jalkanen *et al.*, 2006a, 2008a; Jürgensen and Jalkanen, 2006). Here one needs to identify the important solvent molecules, cations, and anions which are (thought to be) important in stabilizing the structure/conformation/functional state of the molecule/molecular complex/assembly/aggregate. This complex/assembly/aggregate can be embedded within a solvent cavity and/or protein active site and/or micelle, where the cavity surface is treated implicitly by using a lower level of theory (perhaps using a molecular mechanics force field). A complete review of these methods is beyond the scope of this work on the use of vibrational spectroscopic methods in quantum nanobiology, molecular biology, and molecular biophysics.

1.4
Calculating Raman Scattering and Raman Optical Activity Spectra

1.4.1
Raman Scattering

If one wishes to simulate the Raman scattering intensities, one is required to calculate the electric dipole–electric dipole polarizability derivatives. Here one can use either the static values, as was done in most early works, or one can take into account the frequency response, as is now routine. In most conventional Raman scattering experiments one uses either 488, 532, or 1064 nm source, the later for molecules for which fluorescence is a problem.

The Raman intensity is proportional to the Raman scattering activity, which is given for the *jth* normal mode Q_j by

$$I_j^{\text{Ram}} = g_j(45\bar{\alpha}_j^2 + 7\bar{\beta}_j^2) \qquad (12)$$

where g_j is the degeneracy of the jth transition, and the electric dipole–electric dipole polarizability tensor derivative invariants $\bar{\alpha}_j^2$ and $\bar{\beta}_j^2$ are given by

$$\bar{\alpha}_j^2 = \frac{1}{9}\left(S_{\lambda\alpha,j}\alpha_{xx}^{\lambda\alpha} + S_{\lambda\alpha,j}\alpha_{yy}^{\lambda\alpha} + S_{\lambda\alpha,j}\alpha_{zz}^{\lambda\alpha}\right)^2 \qquad (13)$$

and

$$\begin{aligned}\bar{\beta}_j^2 = \frac{1}{9}\Big\{&(S_{\lambda\alpha,j}\alpha_{xx}^{\lambda\alpha} - S_{\lambda\alpha,j}\alpha_{yy}^{\lambda\alpha})^2 \\ &+ (S_{\lambda\alpha,j}\alpha_{xx}^{\lambda\alpha} - S_{\lambda\alpha,j}\alpha_{zz}^{\lambda\alpha})^2 \\ &+ (S_{\lambda\alpha,j}\alpha_{yy}^{\lambda\alpha} - S_{\lambda\alpha,j}\alpha_{zz}^{\lambda\alpha})^2 \\ &+ 6[(S_{\lambda\alpha,j}\alpha_{xy}^{\lambda\alpha})^2 + (S_{\lambda\alpha,j}\alpha_{yz}^{\lambda\alpha})^2 \\ &+ (S_{\lambda\alpha,j}\alpha_{xz}^{\lambda\alpha})^2]\Big\}\end{aligned} \qquad (14)$$

respectively. The $S_{\lambda\alpha,j}$ matrix relates the normal coordinates Q_j to the Cartesian displacement coordinates $X_{\lambda\alpha}$, where λ specifies a nucleus and $\alpha = x, y, or\ z$ (Diem, 1993): as previously defined. In addition to normal and Cartesian coordinates there are symmetry coordinates (Cotton, 1990). The advantage of using symmetry coordinates is that the Hessian will block diagonalize, that is, the Hessian will be reduced to independent sum matrices. Before the advent of supercomputers, symmetry coordinates were used for all symmetrical molecules. They are still used as in the IR and Raman modes for molecules with a center of inversion are mutually exclusive. That is, modes which are active in the Raman and inactive in the IR and vice versa. Here one can make use of character tables (Cotton, 1990). In addition, some modes are not seen in either the Raman or IR,and here the use of IINS can help. The selection rules for IR, Raman and IINS are all different, so the three experiments are complementary and all necessary if one wishes to get complete data for all of the vibrational modes.

The normal frequencies are the eigenvalues Λ of the Hessian matrix H (the second derivative of the energy with respect to nuclear displacements, evaluated at the equilibrium geometry):

$$H_{\lambda\alpha,\lambda'\alpha'} = \left(\frac{\partial^2 E(\mathbf{R})}{\partial X_{\lambda\alpha}\partial X_{\lambda'\alpha'}}\right)_{\mathbf{R}=\mathbf{R}_o} \qquad (15)$$

and

$$C^{-1}HC = \Lambda \qquad (16)$$

C is the eigenvector matrix, which defines the transformation matrix S which is given by

$$S = M^{-1/2}C \qquad (17)$$

where M is the mass matrix.

Therefore, through the matrix S (obtained by diagonalizing the Hessian matrix) and the polarizability derivatives (Electric Dipole-Electric Dipole Polarizability Derivatives EDEDPD), $\alpha_{\beta\gamma}^{\lambda\alpha}$, one can obtain the Raman scattering activity $\mathrm{I}_j^{\text{Ram}}$ of any normal mode Q_j. The first efficient method for determining the EDEDPD was implemented utilizing finite field perturbation theory by Komornicki and McIver (1979). Subsequently the analytical EDEDPD were derived and implemented in CADPAC (Amos, 1986,

1987) and Gaussian (Frish *et al.*, 1986) at the Hartree–Fock level of theory, and subsequently at the DFT level. The ED-EDPD have also been implemented at the DFT level and used to simulate the Raman spectra without any solvent treatment (Johnson and Florian, 1995; Stirling, 1996; Collier, Magdo and Klots, 1999; Halls and Schlegel, 1999a; van Caillie and Amos, 2000; Jalkanen *et al.*, 2006b) and subsequently by treating the solvent with explicit solvent molecules, with continuum solvent models, and finally using hybrid models which combine the two (Jürgensen and Jalkanen, 2006; Jalkanen *et al.*, 2006a, 2008a; Deplazes *et al.*, 2008). In contrast to solvent subtraction which is common in IR/VA spectroscopy, this is not always done with Raman measurements. Investigations of the performance of local, gradient-corrected, and hybrid density functional models in predicting Raman intensities have shown that the hybrid functionals are very accurate when the solvent effects are not important (Halls and Schlegel, 1999b; Jalkanen *et al.*, 2006b).

Hence the question of whether and how to perform a solvent subtraction for Raman measurements is a question that we and others have started to address (Jalkanen *et al.*, 2001, 2003, 2008a; Han *et al.*, 1998; Tajkhorshid, Jalkanen and Suhai, 1998).

1.4.2
Raman Optical Activity

Finally if one wishes to simulate the ROA intensities, one is required to calculate the electric dipole–magnetic dipole polarizability derivatives (ED-MDPD) and the electric dipole–electric quadrupole polarizability derivative (ED-EQPD) (Buckingham, 1967; Atkins and Barron, 1969; Barron and Buckingham, 1971, 2001; Fischer and Hache, 2005). The calculation of the EDMDP, G′, have been implemented in CADPAC (Amos, 1982). The EDMDP can be evaluated as the second derivative of the energy with respect to a static electric field and a time varying magnetic field. Hence the ED-MDPD is a third derivative with respect to the nuclear displacement.

$$G'^{\lambda\alpha}_{\beta\gamma} = \left(\frac{\partial^3 W_G(\mathbf{R}, F_\beta, \dot{B}_\gamma)}{\partial X_{\lambda\alpha}\partial F_\beta \partial \dot{B}_\gamma}\right)_{\substack{\mathbf{R}=\mathbf{R}_o, F_\beta=0,\\ \dot{B}_\gamma=0}} = \left(\frac{\partial G'_{\beta\gamma}(\mathbf{R})}{\partial X_{\lambda\alpha}}\right)_{\mathbf{R}=\mathbf{R}_o} \tag{18}$$

The electric dipole–electric dipole, electric dipole–magnetic dipole, and electric dipole–electric quadrupole polarizability tensors can also be calculated at the frequency of the incident light using SCF linear response theory. London atomic orbitals have been employed, hence the results are gauge invariant (Helgaker *et al.*, 1994). In this work, the electric dipole–magnetic dipole orbitals are gauge independent when using a finite basis set due to the use of gauge-invariant atomic orbitals as implemented in Gaussian 03. Previously we used conventional basis sets which were not gauge invariant, similar to the work of Polavarapu and coworkers (Polavarapu, 1990; Barron *et al.*, 1991a,b; Deng *et al.*, 1996) and our previous work on NALANMA (Han *et al.*, 1998), where that level of theory had been shown to give predicted ROA spectra in good agreement with the experimentally observed spectra. We are thus confident that the new level of theory can be used to answer the structural questions posed here, realizing that the agreement in absolute intensities could be improved by using much larger basis sets, as we shall do for the NALANMA4WC, but not the LA20WC.

The EDEQP, A, can be evaluated as the second derivative of the energy with respect to a static electric field and a static electric field gradient. Hence the EDEQPD is a third derivative, the third derivative being with respect to the nuclear displacement.

$$A^{\lambda\alpha}_{\beta,\gamma\delta} = \left(\frac{\partial^3 W_G(\mathbf{R}, F_\beta, F'_{\gamma\delta})}{\partial X_{\lambda\alpha}\partial F_\beta \partial F'_{\gamma\delta}}\right)_{\substack{\mathbf{R}=\mathbf{R}_0, F_\beta=0,\\ F'_{\gamma\delta}=0}} = \left(\frac{\partial A_{\beta,\gamma\delta}(\mathbf{R})}{\partial X_{\lambda\alpha}}\right)_{\mathbf{R}=\mathbf{R}_0} \qquad (19)$$

The Cartesian polarizability derivatives are required to calculate the ROA spectra. The quantities required have been derived by Barron and Buckingham (1971) and are $\bar{\alpha}_j\bar{G}'_j$, γ_j^2, and δ_j^2 (Barron, 1982, 2004; Buckingham, 1967; Barron, Bogaard and Buckingham, 1973; Hecht and Barron, 1990). The quantities are calculated by combining the various polarizability derivatives with the normal mode vectors in the following equations.

$\bar{\alpha}_j\bar{G}'_j$ is given by

$$\bar{\alpha}_j\bar{G}'_j = \frac{1}{9}\left(S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{xx} + S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{yy} + S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{zz}\right) \times \left(S_{\lambda\alpha,j}G'^{\lambda\alpha}_{xx} + S_{\lambda\alpha,j}G'^{\lambda\alpha}_{yy} + S_{\lambda\alpha,j}G'^{\lambda\alpha}_{zz}\right) \qquad (20)$$

γ_j^2 is given by

$$\begin{aligned}\gamma_j^2 = \frac{1}{2}\Big\{&(S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{xx} - S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{yy})(S_{\lambda\alpha,j}G'^{\lambda\alpha}_{xx} - S_{\lambda\alpha,j}G'^{\lambda\alpha}_{yy}) + (S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{xx} - S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{zz})\\ &\times(S_{\lambda\alpha,j}G'^{\lambda\alpha}_{xx} - S_{\lambda\alpha,j}G'^{\lambda\alpha}_{zz}) + (S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{yy} - S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{zz})(S_{\lambda\alpha,j}G'^{\lambda\alpha}_{yy} - S_{\lambda\alpha,j}G'^{\lambda\alpha}_{zz})\\ &+ 3\Big[(S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{xy})(S_{\lambda\alpha,j}G'^{\lambda\alpha}_{xy} + S_{\lambda\alpha,j}G'^{\lambda\alpha}_{yx}) + (S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{yz})(S_{\lambda\alpha,j}G'^{\lambda\alpha}_{yz} + S_{\lambda\alpha,j}G'^{\lambda\alpha}_{zy})\\ &+(S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{xz})(S_{\lambda\alpha,j}G'^{\lambda\alpha}_{xz} + S_{\lambda\alpha,j}G'^{\lambda\alpha}_{zx})\Big]\Big\}\end{aligned} \qquad (21)$$

and δ_j^2 is given by

$$\begin{aligned}\delta_j^2 = \frac{\omega}{2}\Big\{&(S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{yy} - S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{xx})S_{\lambda\alpha,j}A^{\lambda\alpha}_{z,xy}\\ &+ (S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{xx} - S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{zz})S_{\lambda\alpha,j}A^{\lambda\alpha}_{y,zx}\\ &+ (S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{zz} - S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{yy})S_{\lambda\alpha,j}A^{\lambda\alpha}_{x,yz}\\ &+ S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{xy}(S_{\lambda\alpha,j}A^{\lambda\alpha}_{y,yz} - S_{\lambda\alpha,j}A^{\lambda\alpha}_{z,yy}\\ &+ S_{\lambda\alpha,j}A^{\lambda\alpha}_{z,xx} - S_{\lambda\alpha,j}A^{\lambda\alpha}_{x,xz})\\ &+ S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{xz}(S_{\lambda\alpha,j}A^{\lambda\alpha}_{y,zz} - S_{\lambda\alpha,j}A^{\lambda\alpha}_{z,zy}\\ &+ S_{\lambda\alpha,j}A^{\lambda\alpha}_{x,xy} - S_{\lambda\alpha,j}A^{\lambda\alpha}_{y,xx})\\ &+ S_{\lambda\alpha,j}\alpha^{\lambda\alpha}_{yz}(S_{\lambda\alpha,j}A^{\lambda\alpha}_{z,zx} - S_{\lambda\alpha,j}A^{\lambda\alpha}_{x,zz}\\ &+ S_{\lambda\alpha,j}A^{\lambda\alpha}_{x,yy} - S_{\lambda\alpha,j}A^{\lambda\alpha}_{y,yx})\Big\}\end{aligned} \qquad (22)$$

The equations relating these tensor derivatives to the ROA spectra are given by Barron (1982, 2004); Hecht, Barron and Hug (1989), and Barron *et al.* (1994). These references provide a good introduction to the theory and application of ROA spectroscopy. Here we shall give the relevant equations for completeness. We follow very closely the notation of Barron from his Faraday Discussions review article (Barron *et al.*, 1994).

The commonly reported quantity from ROA measurements is the dimensionless circular intensity differential (CID) defined as

$$\Delta_\alpha = \left(\frac{I_\alpha^R - I_\alpha^L}{I_\alpha^R + I_\alpha^L}\right) \tag{23}$$

where I_α^R and I_α^L are the scattered intensities with linear α polarization in right- and left-circularly polarized incident light. In terms of the quantities, EDEDPD, EDMDPD and EDEQPD, the CIDs for forward, backward, and polarized and depolarized right-angle scattering from an isotropic sample for incident laser light are (Barron *et al.*, 1994):

$$\Delta(\text{forward}) = \left(\frac{8[45\alpha G' + \beta(G')^2 - \beta(A)^2]}{2c[45\alpha^2 + 7\beta(\alpha)^2]}\right) \tag{24}$$

$$\Delta(\text{backward})_{\text{ICP}} = \left(\frac{48[\beta(G')^2 + (1/3)\beta(A)^2]}{2c[45\alpha^2 + 7\beta(\alpha)^2]}\right) \tag{25}$$

$$\Delta(90^\circ, \text{polarized}) = \left(\frac{2[45\alpha G' + 7\beta(G')^2 + \beta(A)^2]}{c[45\alpha^2 + 7\beta(\alpha)^2]}\right) \tag{26}$$

$$\Delta(90^\circ, \text{depolarized}) = \left(\frac{12[\beta(G')^2 - (1/3)\beta(A)^2]}{6c\beta(\alpha)^2}\right) \tag{27}$$

$$\Delta(\text{backward})_{\text{DCP}_I} = \left(\frac{48[\beta(G')^2 + (1/3)\beta(A)^2]}{2c[6\beta(\alpha)^2]}\right) \tag{28}$$

where c is the speed of light and the isotropic invariants are defined as

$$\alpha = (1/3)\alpha_{\alpha\alpha} \tag{29}$$

and

$$G' = (1/3)G'_{\alpha\alpha} \tag{30}$$

and the anisotropic invariants as

$$\beta(\alpha)^2 = (1/2)(3\alpha_{\alpha\beta}\alpha_{\alpha\beta} - \alpha_{\alpha\alpha}\alpha_{\beta\beta}) \tag{31}$$

$$\beta(G')^2 = (1/2)(3\alpha_{\alpha\beta}G'_{\alpha\beta} - \alpha_{\alpha\alpha}G'_{\beta\beta}) \tag{32}$$

$$\beta(G')^2 = (1/2)(3\alpha_{\alpha\beta}G'_{\alpha\beta} - \alpha_{\alpha\alpha}G'_{\beta\beta}) \tag{33}$$

$$\beta(A)^2 = (1/2)\omega\alpha_{\alpha\beta}\varepsilon_{\alpha\gamma\delta}A_{\gamma\delta\beta} \tag{34}$$

Using these expressions, we can calculate the ROA spectra within the harmonic approximation for the vibrational frequencies and with the static field limit for the EDEDPD, EDMDPD, and EDEQPD. In addition, one can take into account the frequency dependence of the Raman and ROA intensities. A new formalism for the Raman and ROA intensities in the near resonance regime has recently been presented by Nafie, though not yet implemented (Nafie, 2008). Finally the analytical derivatives of the G' and A tensors have recently been reported by Cheeseman and Frisch (2008); Ruud (2008); Liegeois and Champagne (2008); Liegeois (2008). Hence the efficiency of calculation of Raman and ROA spectra is now the same as that for VA and VCD spectra. The availability of commercial Raman and ROA and VCD and VA spectrometers and programs to simulate the spectra has been a major breakthrough in the use of these spectroscopies in quantum nanobiology, molecular biology, and molecular biophysics. In the last section, we now present some interesting and pioneering work on the use of VA, VCD, Raman, and ROA spectroscopy in quantum nanobiology, molecular biology, and molecular biophysics.

1.5
Classifying Conformational States and Structures based on VA, VCD, Raman, and ROA Spectra

1.5.1
Experimental Vibrational Frequencies for *N*-Methylacetamide, NALANMA, Peptides, and Proteins

The amide modes have been determined experimentally from the simple amides formamide, acetamide, and NMA. NMA has been used as a model since its two methyl groups can be assumed to represent the two C_α carbons in proteins. The amide A mode is mainly a N−H stretch, amide I is mainly a C=O stretch, amide II a N−H bend coupled with C−N stretching, amide III N-bending and C−N stretching (only in simple amides as it has been shown to couple with the CH deformations on the adjacent C_α carbons with C−HR groups [nonglycine] (Diem, 1993; Jalkanen *et al.*, 2003), amide IV OCN bending, amide V in-plane N−H bending, amide VI out-of-plane N−H bending, and finally amide VII skeletal torsion (Jung, 2000; Jalkanen and Suhai, 1996). In addition to the amide modes for NMA being identified from both experimental and theoretical studies, these have also been identified in a number of peptides and proteins and the amide frequencies have been correlated with secondary structural elements (Jung, 2000). The mode most often correlated with secondary structure is the amide I mode (called *amide I'* when measured in D_2O). The amide I frequencies need to be measured in D_2O because the amide I mode overlaps with the H_2O bending modes. Hence running the spectra in D_2O when one wants to determine the C=O stretching frequency. Alternatively, one can make up a very concentrated solution and measure the spectra for very thin pathlengths.

In addition, by performing ^{13}C isotopic substitution for the C′ carbon in specific residues, one can shift the frequency of the amide I band for the specific residue. To date, the IR spectra of proteins and their corresponding amide I spectrum have been used to only determine the percent of a specific secondary structural element present, and one has not been able to determine the specific backbone angles for the specific residue and the corresponding side chain angles. This information for each residue would uniquely identify the structure of the protein and is the goal of the combined use of VA, VCD, Raman, and ROA spectroscopy, as has been achieved for both X-ray and neutron diffraction and NMR spectroscopies. In the next section, we review our and other groups theoretical vibrational spectra simulations for amino acids and peptides and present our new work on LALA and NALHNMA. The complete work on the later two molecules will be published in a special issue of Theoretical Chemical Accounts which focuses on molecular biophysics and bioinformatics, the Suhai Festschrift, which will appear in 2009. Hence for more details on the work on LALA and NALHNMA refer to the manuscripts in this special issue (Jalkanen *et al.*, 2009a,b).

1.5.2
Results for the DFT Calculations

1.5.2.1 L-Alanine

In Table 1.1, we present results for the LA zwitterionic species, which is the dominant species found in aqueous solution at neutral pH (Jalkanen, Bohr

Tab. 1.1 L-Alanine zwitterion structures.

Coordinate	I[a] 6-31G*	I[a] 6-31G**	I[a] ++G(2d,2p)	II[a] 6-31G*	III[b] 6-31G*	IV[c] 6-31G*	V[d] 6-31G*
r(N1−C2)	1.5115	1.5105	1.5129	1.5182	1.5142	1.5123	1.481
r(C2−C3)	1.5639	1.5638	1.5593	1.5673	1.5545	1.5516	1.534
r(C3−O4)	1.2893	1.2899	1.2854	1.2748	1.2615	1.2647	1.277
r(C3−O5)	1.2299	1.2297	1.2211	1.2421	1.2608	1.2601	1.247
r(C2−C6)	1.5267	1.5263	1.5229	1.5202	1.5227	1.5220	1.526
r(C2−H7)	1.0947	1.0945	1.0889	1.0940	1.0923	1.0925	1.096
r(N1−H8)	1.0466	1.0460	1.0344	1.0565	1.0536	1.0484	1.054
r(N1−H9)	1.0477	1.0472	1.0346	1.0559	1.0454	1.0464	1.057
r(N1−H10)	1.0340	1.0322	1.0250	1.0334	1.0529	1.0583	1.057
r(C6−H11)	1.0960	1.0950	1.0907	1.0940	1.0933	1.0938	1.094
r(C6−H12)	1.0916	1.0905	1.0866	1.0932	1.0942	1.0940	1.091
r(C6−H13)	1.0966	1.0954	1.0913	1.0937	1.0960	1.0955	1.093
θ(N1−C2−C3)	111.19	111.28	110.94	105.04	106.82	106.41	110.55
θ(C2−N3−O4)	116.14	116.18	115.80	117.31	115.31	115.83	119.06
θ(C2−N3−O5)	115.09	115.11	115.72	114.41	117.93	118.10	115.68
θ(N1−C2−C6)	109.96	110.05	109.95	110.18	109.95	109.85	111.05
θ(C3−C2−C6)	111.60	111.52	112.10	114.17	114.76	114.57	110.71
θ(N1−C2−H7)	106.19	106.32	106.13	106.06	105.79	105.78	106.84
θ(C3−C2−H7)	107.47	107.25	107.21	109.48	107.73	108.67	107.94
θ(C2−N1−H8)	109.05	108.83	109.82	108.64	112.36	111.69	111.54
θ(C2−N1−H9)	109.83	109.65	110.22	109.31	108.93	108.99	108.38
θ(H8−N1−H9)	106.25	106.27	105.95	106.22	109.40	109.14	110.87
θ(C2−N1−H10)	110.63	110.44	111.30	111.72	110.97	110.98	109.87
θ(H8−N1−H10)	110.27	110.54	109.88	110.57	109.60	110.36	107.78
θ(C2−C6−H11)	110.70	110.62	110.89	110.52	110.93	110.77	111.21
θ(C2−C6−H12)	108.27	108.21	108.42	109.44	109.85	109.84	109.80
θ(H11−C6−H12)	108.28	108.28	108.21	107.20	108.10	108.01	108.29
θ(C2−C6−H13)	111.52	111.52	111.56	111.36	110.91	110.87	110.88
θ(H11−C6−H13)	108.81	108.84	108.91	109.69	108.91	109.22	108.46
τ(N1−C2−C3−O4)	−0.53	0.08	−3.03	−101.08	94.86	86.84	−14.21
τ(N1−C2−C3−O5)	−179.57	−179.83	176.59	73.20	−81.69	−89.60	166.29
τ(C3−C2−N1−H8)	59.20	58.85	60.96	48.12	80.56	72.44	64.06
τ(C3−C2−N1−H9)	−56.85	−57.00	−55.40	−67.39	−40.83	−38.23	−58.27
τ(C3−C2−N1−H10)	−179.36	−179.64	−177.12	170.36	−156.34	−163.96	−176.50
τ(H11−C6−C2−H7)	177.17	177.27	177.02	−176.93	178.19	177.51	−175.98
τ(H12−C6−C2−H7)	−64.29	−64.26	−64.32	−59.09	−62.36	−63.27	−56.14
τ(H13−C6−C2−H7)	55.87	56.00	55.44	60.89	57.04	56.08	63.25

[a] Zwitterion stabilized by four water molecules.
[b] Zwitterion stabilized by nine water molecules.
[c] Zwitterion stabilized by nine water molecules + solvent model.
[d] Zwitterion stabilized by 20 water molecules + CPCM model, B3PW91.

and Suhai, 1997; Tajkhorshid, Jalkanen and Suhai, 1998; Frimand *et al.*, 2000; Jalkanen *et al.*, 2001). At the RHF level of theory the zwitterion was found to be stable (Barron *et al.*, 1991a), but one of the N−H bond lengths was found to be much longer than the other two. This resulted in the vibrational stretch frequency associated with N−H to be significantly lower in energy than that for the other two N–H bonds. If the three NH bonds were all the same, the three modes would be thought to be approximately the same and vibrational frequencies would be determined by both the local symmetry (thought to be C_{3v}) and the perturbation due to the local environment. Herein lies the additional information which can in theory be extracted from the experimental VA, VCD, Raman, and ROA spectra. Subsequently Yu and coworkers tried to improve the model for the LA zwitterion by embedding the zwitterion with a continuum solvent treatment, using the Onsager continuum model (Onsager, 1936), optimizing the structure and subsequently simulating the Raman and ROA spectra (Yu *et al.*, 1995). At the B3LYP DFT level of theory, we found that the zwitterionic structure was not stable in the isolated state (Jalkanen, Bohr and Suhai, 1997; Tajkhorshid, Jalkanen and Suhai, 1998).

We present three 6-31G* B3LYP zwitterionic structures for LA, two in the presence of four water molecules (Tajkhorshid, Jalkanen and Suhai, 1998) and one in the presence of nine water molecules (Frimand *et al.*, 2000). We also present the 6-31G** and 6-311++G(2d,2p) optimized geometries for one of the zwitterionic structures. As one can see from the basis set effect, as one adds more polarization functions and diffuse functions, the bond length changes are usually less than 0.004 Å. The valence angle changes are usually less than 0.5°. Finally we also present the LA zwitterion completely solvated by 20 water molecules and embedded with the PCM continuum solvent model (Jalkanen *et al.*, 2008).

Initially we solvated the zwitterion by adding water molecules to hydrogen bond with each of the polar groups, and subsequently extracted the positions from classical MD simulations (Tajkhorshid, Jalkanen and Suhai, 1998; Frimand *et al.*, 2000; Jalkanen *et al.*, 2001). In addition to the normal isotopomers of LA, we have additionally looked at the various deuteriated derivatives theoretically and compared to the experimental work of Diem *et al.* (1982); Abdali *et al.* (2002). We have recently determined the positions of the hydrogens and completely solvated the LA zwitterion with BO MD simulations (Degtyarenko *et al.*, 2007) and subsequently have simulated the VA, VCD, Raman, and ROA spectra for the completely solvated zwitterion and embedded within the COSMO and PCM continuum models (Jalkanen *et al.*, 2008a). Finally we have compared the zwitterionic structure derived from molecular dynamic simulations with that found in the crystal (Degtyarenko *et al.*, 2008). Hence the zwitterionic structure of LA has been thoroughly studied. The experimental VA, VCD, Raman, and ROA spectra have been able to be reproduced by treating the first solvation shell of water molecules explicitly and embedding the solvated zwitterionic structure within the continuum solvent models so as to treat the effects due to bulk water molecules. In Figure 1.1, we present the zwitterionic structure of the LAZ hydrated by 20 water molecules and this hydrated complex embedded within the continuum solvent model. As one can see in Figure 1.1 the 4 water molecules in the original LAZ plus 4 water molecules for the F' structure are very similar to 4 of the water molecules in the LAZ plus 20 water molecules. This can explain why the LAZ plus four water molecule plus Onsager model did quite well in reproducing the VA, VCD, Raman, and ROA spectra of the LAZ

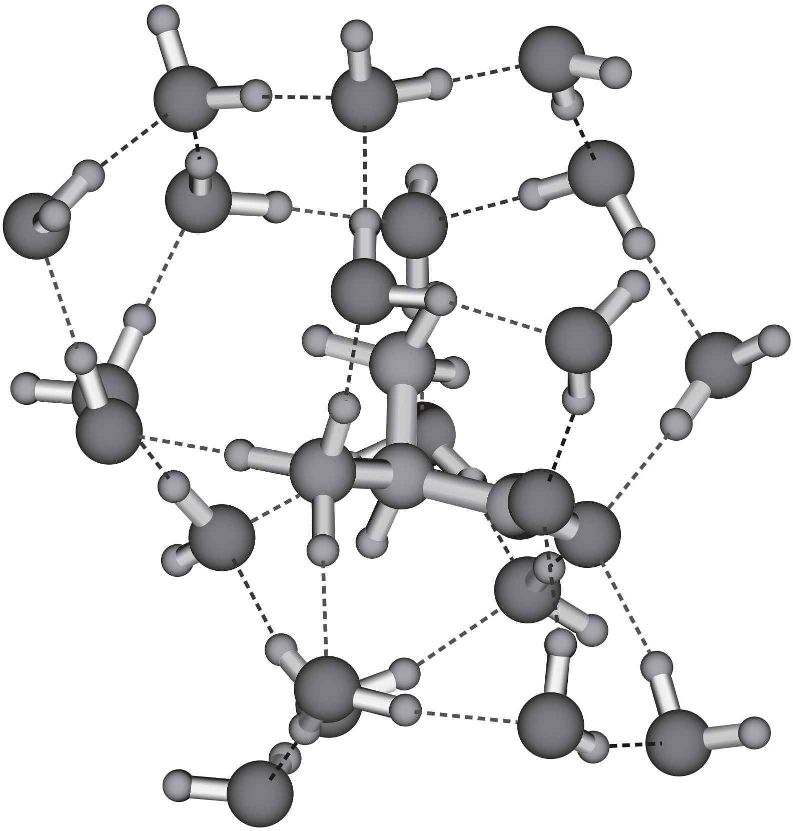

Fig. 1.1 L-Alanine zwitterion stabilized by 20 water molecules. B3PW91/6-31G* plus CPCM continuum solvent model optimized geometry. (Please find a color version of this figure on the color plates.)

(Tajkhorshid, Jalkanen and Suhai, 1998; Jalkanen *et al.*, 2001). Previously we have presented the experimental VA, VCD, Raman, and ROA spectra and the corresponding simulated VA, VCD, Raman, and ROA spectra (Tajkhorshid, Jalkanen and Suhai, 1998; Frimand *et al.*, 2000; Jalkanen *et al.*, 2001, 2008a). We showed that the LAZ plus 20 water molecule model gives better agreement with the experimental spectra, but that our minimal model is a good starting point. Our minimal model being the LAZ plus four water molecule complex embedded in one of the continuum solvent models (Onsager, PCM, CPCM, or Cosmo) (Tomasi, Mennucci and Cammi, 2005).

When one measures the VA, VCD, Raman, and/or ROA spectra, one has the problem of determining the spectra and spectra of the solvent. In many cases one assumes that one can subtract the spectra of the bulk solvent to get the spectra due to the isolated solute. This only applies when the solvent is inert and does not strongly interact with the solute. This is clearly not the case for amino acids and water. As stated above, the water molecules are necessary to stabilize the zwitterionic species. The relative energy of the neutral nonionic species is lower than that of the zwitterionic species, which is not even a stable minima. So one must rethink about how one could in principle subtract out the spectra of the so-called solvent, when the solute and solvent strongly interact. Of course, not all of the solvent molecules strongly interact via hydrogen bonding with the LA zwitterion. Only the first solvation shell waters are thought to be important. In our preliminary works, we included only four and then nine water molecules, but in our last work we included a complete solvation shell, which included 20 water molecules. In addition to the structure of the LA molecule being perturbed by the aqueous environment, the individual water molecules which are hydrogen bonded with the LA zwitterion are perturbed from both isolated water molecules and water molecules in bulk water. In the case of methyl lactate, water molecule(s) bound to polar groups have been shown to possess a VCD signature (Losada and Xu, 2007). Hence on the timescale of the VCD measurement, the water molecules strongly bound to the polar chiral molecule show chirality. Hence chirality can be due to weak hydrogen bonding, in addition to strong covalent bonding. With an increase in temperature the VCD spectra will go away, whereas a drop in temperature will make other water molecules less mobile and hence increase the VCD and also the ROA spectra. Hence by cooling down the aqueous solution of biomolecules, one should be able to learn more about the character of the hydration shells of biomolecules. From these temperature-dependent spectra one should be able to identify various water molecules in the first solvation shell, the strength of their binding, and their exchange rates and mobilities. This should also be the case for ligands bound in the binding pockets (active sites) of proteins.

In Tables 1.2a and 1.2b we present the calculated frequencies and VA and VCD intensities for the zwitterionic structures of L-alanine-d_0 We also give the experimental results reported by Diem and coworkers. In Table 1.2a we also give the calculated vibrational frequencies and VA intensities recently reported by Barron and coworkers. For a complete review of our works on LA, see our latest publication in the Stephens Festschrift in 2008 (Jalkanen *et al.*, 2008a).

1.5.2.2 ***N*-Acetyl L-Alanine *N′*-Methylamide**

In addition to the individual amino acids, one is interested in the amino acid residue as they appear in peptides and proteins. To model the amino acid in a protein, we have capped the LA amino acid with an *N*-acetyl group on the N-terminus and a *N′*-methylamide on the C-terminus. Some researchers have capped the N-terminus with the *N*-formyl group and the C-terminus with the simple amide group. In our work the methyl groups approximate the C_α carbon atoms in the adjacent residues, whereas in the simpler model, hydrogen atoms approximate the C_α carbons. With the speed of computers there is no longer a need to use the simpler model and hence we model our amino acid residues, LA/LH, with the NALANMA and NALHNMA molecules.

In Table 1.3, we present the relative energies of isolated NALANMA and with four bound water molecules. The values of ϕ and ψ (measures of secondary structure in proteins) are also given. The starting structures for the bound water optimizations were the 6-31G* B3LYP

Tab. 1.2a L-Alanine-d_0 zwitterion vibrational frequencies and VA intensities.

ν(cm^{-1})^{a}	ν(cm^{-1})^{b} I	A_i^{c}	ν(cm^{-1})^{b} II	A_i^{c}	ν(cm^{-1})^{b} III	A_i^{c}	ν(cm^{-1})^{d}	A_i^{d}
3080	3317	529.5	3327	497.5	3153	14.8	3809	65.3
3060	3170	3.6	3155	12.0	3132	32.2	3727	53.9
3020	3120	15.9	3132	19.4	3124	11.6	3333	3.7
3003	3105	336.6	3099	10.8	3114	96.1	3295	20.7
2993	3088	327.4	3057	29.0	3063	36.6	3258	32.2
2962	3077	502.0	2956	683.3	3009	63.9	3204	31.9
2949	3053	33.9	2940	691.6	2991	72.3	2813	581.0
1645	1770	324.3	1777	77.9	1775	60.3	1996	442.0
1625	1745	82.9	1745	181.4	1749	37.8	1822	42.1
1607	1741	56.8	1719	221.6	1696	544.0	1790	24.6
1498	1634	91.1	1647	78.2	1662	185.0	1641	11.9
1459	1534	2.2	1525	4.7	1526	21.5	1639	3.9
1459	1525	5.2	1513	6.5	1522	3.1	1568	47.0
1410	1433	41.3	1444	44.1	1454	92.1	1523	388.0
1375	1407	7.0	1415	5.8	1438	98.7	1501	69.9
1351	1377	34.6	1405	63.9	1428	8.9	1447	143.0
1301	1333	186.1	1342	157.5	1370	111.5	1395	238.0
1220	1281	37.5	1297	14.2	1298	22.9	1314	18.0
1145	1200	56.0	1199	20.2	1238	36.6	1205	41.8
1110	1124	16.9	1115	13.1	1121	16.9	1164	38.7
1001	1058	7.9	1075	7.0	1076	5.2	1080	72.9
995	1026	7.6	1023	1.5	1038	10.1	1060	22.6
922	901	15.3	896	17.9	921	17.7	942	99.2
850	833	43.4	840	28.9	850	34.5	881	34.1
775	771	9.8	799	13.4	804	17.9	838	16.2
640	620	4.2	658	6.8	672	8.9	681	7.9
527	602	3.7	639	1.9	658	0.7	562	34.9
477	521	16.7	533	2.3	549	9.8	413	3.0
399	432	15.7	403	28.0	400	11.3	352	111.0
296	368	39.4	352	45.4	335	34.9	303	30.4
283	287	8.1	279	3.7	283	14.1	275	17.2
219	262	4.2	231	4.2	232	3.6	255	3.3
184	178	6.9	170	3.4	171	2.4	54	3.5

[a] M. Diem and coworkers, *J. Am. Chem. Soc.* **104** (1982) 3329.

[b] 6-31G* B3LYP frequencies.

[c] 6-31G* B3LYP APT.

[d] Barron *et al.* (1991a).

optimized geometries. To each of these structures, four waters molecules were added by the Insight program (Biosym Technologies, San Diego, CA USA). The details of these calculations and the VA, VCD, Raman, and ROA spectra for this molecule have been presented (Jalkanen and Suhai, 1996; Han *et al.*, 1998; Jalkanen *et al.*, 2008a). Here we review the most important results. Note that recent developments by Cao *et al.* have reduced the linear birefringence in VCD measurements, which has increased significantly the sensitivity of modern

Tab. 1.2b L-Alanine-d_0 zwitterion vibrational frequencies and VCD intensities.

$\nu(\mathrm{cm}^{-1})^a$	$10^5(\Delta\varepsilon)^a$	$\nu(\mathrm{cm}^{-1})^b$ I	R_i^c	$\nu(\mathrm{cm}^{-1})^b$ II	R_i^c	$\nu(\mathrm{cm}^{-1})^b$ III	R_i^c
3080	–	3317	17.1	3327	28.0	3153	−0.1
3060	–	3170	−1.4	3155	−2.3	3132	−11.5
3020	–	3120	1.7	3132	−4.1	3123	9.2
3003	–	3105	−76.2	3099	−17.6	3063	0.1
2993	–	3088	29.5	3057	−1.7	2293	−1.3
2962	–	3078	15.7	2957	−55.9	2230	0.5
2949	–	3053	−11.6	2940	21.6	2162	1.8
1623	−1.1	1770	−53.8	1777	−77.3	1698	42.9
1625	–	1745	−0.6	1746	162.9	1525	−5.4
1607	–	1741	26.4	1719	−65.2	1523	6.4
1498	–	1635	−6.4	1647	−5.5	1453	21.5
1459	–	1534	−4.6	1525	−2.5	1432	−56.5
1459	–	1525	−5.1	1513	14.9	1406	1.3
1410	–	1433	32.3	1444	26.6	1350	32.8
1375	–	1407	19.3	1415	−35.5	1275	8.6
1351	−27	1377	−71.3	1405	28.6	1252	9.6
1301	16	1333	−13.5	1341	−122.0	1246	−55.8
1220	5.3	1281	60.1	1298	18.9	1188	−4.1
1145	−4.2	1200	−24.8	1198	37.8	1122	−9.8
1110	−4.7	1124	−0.7	1115	12.3	1077	10.9
1001	–	1058	0.4	1074	15.9	941	−1.4
995	–	1026	6.6	1022	−7.1	904	−45.6
922	–	901	−23.4	896	−20.8	900	67.0
850	–	833	21.1	839	16.6	835	−25.2
775	–	771	23.6	799	−5.2	786	−8.3
640	–	620	−1.0	658	−18.2	657	0.7
527	–	602	10.1	639	7.7	540	−10.9
477	–	521	−8.5	533	−7.7	471	1.5
399	–	432	14.9	403	43.7	383	16.5
296	–	368	−26.3	352	51.4	321	45.4
283	–	287	39.9	279	−23.7	281	−34.2
219	–	262	−22.9	231	−25.9	231	3.5
184	–	179	−9.3	170	−35.5	169	−14.8

[a] M. Diem, *J. Am. Chem. Soc.* **118** (1988) 6967–6970.
[b] 6-31G* B3LYP frequencies.
[c] 6-31G* B3LYP APT and 6-31G** RHF DO gauge AAT.

VCD instrumentation (Cao, Dukor and Nafie, 2008). This increased sensitivity is very important for amino acid, peptide, and protein VCD measurements in aqueous solution.

In Figure 1.2, we present the structure of the NALANMA plus four water molecular complex (Jalkanen *et al.*, 2008a). As one can see the structures and energetics of this molecule are greatly affected by the solvent, consistent with large changes in the VCD spectra of this molecule when one changes solvent from carbon tetrachloride to water. Note that the C_7^{eq} and C_5^{ext} conformers both converge to the same structure,

Tab. 1.3 *N*-Acetyl-L-alanine *N*′-methylamide conformational energies.

Conformer	ϕ^a	ψ^a	Energy[a] (kcal mol^{-1})	ϕ^b	ψ^b	Energy[b] (kcal mol^{-1})
C_7^{eq}	−82	72	0.000	−94	128	0.000
C_5^{ext}	−157	165	1.433	−94	128	0.000
C_7^{ax}	74	−60	2.612	59	−122	4.132
β_2	−136	23	3.181	−151	116	1.886
α_L	68	25	5.817	61	52	2.754
α_R	−60	−40	5.652	−82	−44	2.465
α_D	57	−133	6.467	67	−111	3.715
α'	−169	−38	6.853	−153	−92	15.140
Cryr	−84	155	–	−98	112	5.864

[a] Isolated NALANMA, 6-31G* B3LYP relative energies.
[b] NALANMA with 4 bound waters, 6-31G* B3LYP relative energies.

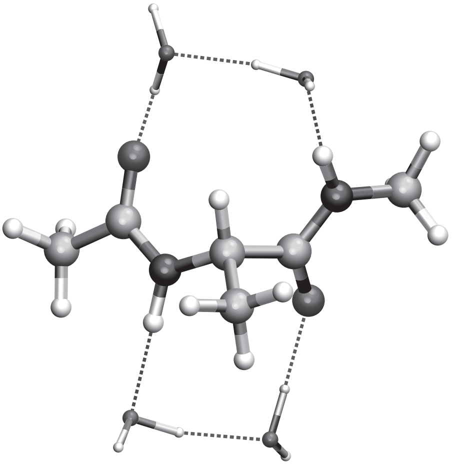

Fig. 1.2 *N*-Acetyl L-alanine *N*′-methylamide stabilized by four water molecules. B3PW91/6-31G* plus CPCM continuum solvent model optimized geometry. (Please find a color version of this figure on the color plates.)

which is the lowest energy structure of NALANMA with four bound waters found till date. This structure has also been used to reanalyze the NMR data for this molecule (Poon *et al.*, 2000; Weise and Weisshaar, 2003). The conclusions from both the NMR and VA, VCD, Raman, and ROA works is that NALANMA plus four water molecule complex is the entity (conformer) of interest, and not the isolated NALANMA molecule (Jalkanen *et al.*, 2008a). The structure of this molecule in aqueous solution had been studied by a variety of both experimental and theoretical methods, but had alluded all previous researchers. The problem had been that most of the previous researchers had failed to take into account the effects of the first solvation water molecules, which are responsible for stabilizing a structure which is not stable on the gas phase isolated-state potential energy surface. This work has been called the gold standard for vibrational spectroscopic structure determination for small peptides. The work on this molecule also documented the importance of measuring not only the VA and VCD or Raman and ROA but also all four spectra in addition to NMR. By measuring all these spectra and getting a structure or set of structures which are consistent with all of the measured spectra, one is able to get a self-consistent picture. This has also been advocated by Serdyuk, Zaccai and Zaccai (2007).

1.5.2.3 L-Alanyl L-Alanine

NALANMA is called the *alanine dipeptide* due to the formation of two peptide bonds by the capping. For biochemists dipeptides have only one peptide bond and keep the N-terminus and C-terminus groups. Hence to model dipeptides, one must now once again treat the N-terminus and C-terminus groups. In aqueous solution at pH 7 once again one has the neutral ionic species, the zwitterionic form of L-alanyl L-alanine.

In Tables 1.4a and 1.4b, we present the structures of zwitterionic structures stabilized by 4, 6, 7, 11, and 17 waters determined, respectively. Here one can see the effect of adding progressively more water molecules. As one can see the more water molecules one adds, the closer the LALA molecule comes to what it would be in aqueous solution. The more water molecules one adds, the more one is able to complete the first solvation shell. The polar groups are enough to hold a small number of water molecule in place, but when one starts to solvate the nonpolar groups one is faced with a problem that during the geometry optimization the water molecules tend to clump together, and this is even without any kinetic energy (i.e. zero temperature). Hence one is faced with the problem of completing the first solvation shell completely, as we have done for LA where we found we needed 20 water molecules to fully encapsulate the LA zwitterion. In a future, study we will completely solvate the LALA zwitterion, but in this work we have only presented and refer to our structures stabilized by 4, 6, 7, 11 and 17 water molecules. In Figure 1.3, we present the LALA zwitterion stabilized by 11 water molecules and the PCM continuum solvent model with the B3PW91 hybrid XC functional and the 6-311G(2d,p) basis set. The 11 water molecules are directly hydrogen bonded to the polar groups, whereas the PCM continuum model is added to treat the effects due to the bulk water. This is similar to our previous works and also the methods recently advocated by Thar, Zhan and Kirchner (2008). In Figure 1.4, we present the simulate VA and VCD spectra for the LALA zwitterion stabilized by 11 waters at the B3PW91/6-311G(2d,p) level of theory including the PCM continuum model, and in Figure 1.5 the corresponding Raman and ROA spectra. In Figure 1.6, we present the LALA zwitterion stabilized by 17 waters and PCM continuum model with the B3PW91 hybrid XC functional and the 6-31G* basis set. In a future work, we will present all of our work on the LALA zwitterion (Jalkanen *et al.*, 2009b).

We also present two zwitterionic structures stabilized by the Onsager model implemented in Gaussian (94, 98, and 03) with and without explicit waters. In addition to the Onsager continuum model, we have optimized the structures of a collection of LALA plus N water molecule complexes, similar to our work for NALANMA with both the PCM and

Tab. 1.4a L-Alanyl-L-alanine zwitterion structures: bond lengths and dihedral angles.

Coordinate	I[a]	I[b]	I[c]	II[d]	III[e]	IV[f]	V[g]
R(C2N1)	1.5232	1.5090	1.5065	1.5060	1.5086	1.4774	1.4854
R(C3C2)	1.5638	1.5532	1.5531	1.5516	1.5460	1.5297	1.5332
R(O4C3)	1.2384	1.2348	1.2383	1.2224	1.2311	1.2372	1.2375
R(N5C3)	1.3411	1.3437	1.3384	1.3651	1.3524	1.3297	1.3404
R(C6C2)	1.5299	1.5216	1.5219	1.5224	1.5288	1.5237	1.5256
R(H7C2)	1.0930	1.0955	1.0957	1.0967	1.0917	1.0946	1.1004
R(H8C6)	1.0954	1.0929	1.0935	1.0953	1.0944	1.0924	1.0955
R(H9C6)	1.0938	1.0928	1.0925	1.0908	1.0943	1.0894	1.0921
R(H10C6)	1.0948	1.0950	1.0953	1.0952	1.0964	1.0921	1.0949
R(H11N1)	1.0289	1.0359	1.0340	1.0340	1.1524	1.0493	1.0515
R(H12N1)	1.0487	1.0566	1.0547	1.0386	1.0227	1.0417	1.0532
R(H23N1)	1.0276	1.0381	1.0457	1.0554	1.0462	1.0494	1.0473
R(C13N5)	1.4726	1.4754	1.4771	1.4862	1.4731	1.4491	1.4550
R(H22N5)	1.0216	1.0191	1.0197	1.0187	1.0168	1.0234	1.0240
R(C14C13)	1.5777	1.5590	1.5607	1.5568	1.5590	1.5392	1.5443
R(O15C14)	1.2572	1.2773	1.2757	1.2750	1.2814	1.2500	1.2680
R(O16C14)	1.2617	1.2465	1.2449	1.2476	1.2457	1.2591	1.2614
R(C17C13)	1.5340	1.5265	1.5262	1.5251	1.5284	1.5288	1.5288
R(H18C13)	1.0974	1.0951	1.0948	1.0921	1.0898	1.0883	1.0951
R(H19C17)	1.0972	1.0936	1.0938	1.0941	1.0955	1.0932	1.0938
R(H20C17)	1.0933	1.0947	1.0950	1.0967	1.0950	1.0923	1.0948
R(H21C17)	1.0946	1.0944	1.0942	1.0953	1.0947	1.0901	1.9841
τ(O4C3C2N1)	−7.35	−123.42	−126.36	−146.63	109.83	−19.01	12.38
ψ_1(N5C3C2N1)	172.37	57.50	54.50	34.38	−67.77	162.77	−168.59
ω(C13N5C3C2)	178.91	177.48	179.06	162.30	155.23	175.15	173.89
ϕ_2(C14C13N5C3)	−169.65	−162.49	−164.92	−135.88	−84.87	−134.60	−72.35
ψ_2'(O15C14C13N5)	−2.30	67.31	67.56	85.50	123.81	−16.03	−30.38
ψ_2''(O16C14C13N5)	178.40	−108.61	−107.74	−88.70	−55.18	164.41	150.76
τ(H8C6C2H7)	178.23	179.03	177.67	175.82	177.19	−177.97	−179.94
τ(H9C6C2H7)	−62.17	−61.73	−63.13	−65.66	−62.35	−57.55	−58.89
τ(H10C6C2H7)	56.84	57.11	56.08	54.65	56.58	61.74	59.44
τ(C3C2N1H11)	127.44	−173.41	−174.97	−174.94	−47.16	178.40	171.26
τ(C3C2N1H12)	8.00	59.80	57.88	57.92	−167.63	60.66	49.88
τ(C3C2N1H23)	−110.70	−62.10	−63.68	−63.68	73.09	−61.85	−69.26
τ(H19C17C13H18)	177.98	−177.54	−176.62	−177.83	180.00	−176.93	178.19
τ(H20C17C13H18)	57.35	60.61	61.26	59.99	58.94	62.89	57.34
τ(H21C17C13H18)	−63.59	−59.17	−58.66	−59.33	−61.10	−56.27	−61.40

[a] Zwitterion stabilized by Onsager solvent model at B3LYP/6-31G* level of theory.

[b] Zwitterion stabilized by Onsager solvent model and seven water molecules at B3LYP/6-31G* level of theory.

[c] Zwitterion stabilized by seven water molecules at B3LYP/6-31G* level of theory.

[d] Zwitterion stabilized by six water molecules at B3LYP/6-31G* level of theory.

[e] Zwitterion stabilized by four water molecules at B3LYP/6-31G* level of theory.

[f] Zwitterion stabilized by PCM continuum model and 11 water molecules at B3PW91/6-311G(2d,p) level of theory.

[g] Zwitterion stabilized by PCM continuum model and 17 water molecules at B3PW91/6-31G* level of theory.

Tab. 1.4b L-Alanyl-L-alanine zwitterion structures: valence angles.

Coordinate	I[a]	I[b]	I[c]	II[d]	III[e]	IV[f]	V[g]
A(N1C2C3)	104.60	108.46	108.62	111.34	107.58	108.47	106.70
A(C2C3O4)	117.62	122.01	121.46	119.87	119.28	119.48	118.36
A(C2C3N5)	118.06	113.69	113.97	115.82	115.75	115.10	118.56
A(O4C3N5)	124.32	124.30	124.56	124.30	124.92	125.40	123.08
A(N1C2C6)	110.20	111.07	111.01	110.29	111.26	110.36	109.94
A(C3C2C6)	113.96	113.57	113.14	111.90	115.40	111.57	115.45
A(N1C2H7)	106.97	105.17	105.37	105.40	106.62	107.11	107.13
A(C3C2H7)	110.19	107.92	108.05	106.97	104.87	109.27	107.83
A(C6C2H7)	110.56	110.27	110.32	110.69	110.57	109.95	109.46
A(C2C6H8)	111.12	111.80	111.89	110.92	111.81	110.79	112.81
A(C2C6H9)	109.73	109.25	109.33	109.21	109.84	109.76	110.50
A(H8C6H9)	108.18	107.80	107.68	107.68	108.46	109.00	108.03
A(C2C6H10)	111.08	110.38	110.35	110.86	110.70	110.09	109.60
A(H8H6H10)	108.89	109.32	109.03	108.95	108.11	108.72	108.19
A(H9H6H10)	107.73	108.18	108.47	109.16	107.80	108.45	107.53
A(C2N1H11)	113.36	110.71	110.89	110.80	112.75	110.30	111.06
A(C2N1H12)	104.73	111.06	110.52	110.18	110.16	111.04	111.78
A(C2N1N23)	113.73	108.64	108.71	109.57	110.58	110.68	114.04
A(11N1H12)	109.61	113.31	113.79	114.03	107.80	106.44	108.53
A(11N1H23)	106.50	102.08	101.94	101.56	107.40	108.24	105.84
A(12N1H23)	108.83	110.64	110.59	110.36	107.98	110.03	105.18
A(C3N5C13)	123.93	125.32	125.71	122.12	122.25	122.52	118.95
A(N5C13C14)	109.30	104.48	103.88	103.54	108.17	111.98	111.66
A(C13C14O15)	116.15	114.42	114.25	115.12	118.40	118.14	118.58
A(C13C14O16)	115.02	118.54	118.09	116.93	116.18	114.70	116.49
A(O15C14O16)	128.83	126.90	127.47	127.65	125.41	127.15	124.93
A(N5C13C17)	111.91	111.91	111.66	111.44	110.49	112.01	111.66
A(C14C13C17)	110.77	113.70	113.67	113.73	111.73	110.52	110.35
A(N5C13H18)	108.21	107.24	107.43	106.98	107.05	106.77	107.82
A(C14C13H18)	108.11	108.86	109.19	109.85	108.40	106.94	106.71
A(C17C13H18)	108.42	110.31	110.63	110.87	110.84	108.35	108.43
A(C13C17H19)	110.36	110.28	110.08	110.12	110.64	111.10	112.77
A(C13C17H20)	111.33	110.40	110.61	110.57	110.92	110.13	109.96
A(H19C17H20)	108.56	110.07	110.27	110.32	108.97	108.45	108.20
A(C13C17H21)	108.19	109.90	109.85	110.28	109.84	109.35	109.26
A(H19C17H21)	108.54	107.52	107.31	107.52	107.81	109.24	108.24
A(H20C17H21)	108.56	108.61	108.64	107.95	108.58	108.52	108.31
A(C3N5H22)	124.46	118.18	118.21	117.63	117.85	119.37	121.34
A(C13N5H22)	111.31	116.38	116.00	114.00	114.98	118.07	118.49

[a] Zwitterion stabilized by Onsager solvent model at B3LYP/6-31G* level of theory.

[b] Zwitterion stabilized by Onsager solvent model and seven water molecules at B3LYP/6-31G* level of theory.

[c] Zwitterion stabilized by seven water molecules at B3LYP/6-31G* level of theory.

[d] Zwitterion stabilized by six water molecules at B3LYP/6-31G* level of theory.

[e] Zwitterion stabilized by four water molecules at B3LYP/6-31G* level of theory.

[f] Zwitterion stabilized by PCM continuum model and 11 water molecules at B3PW91/6-311G(2d,p) level of theory.

[g] Zwitterion stabilized by PCM continuum model and 17 water molecules at B3PW91/6-31G* level of theory.

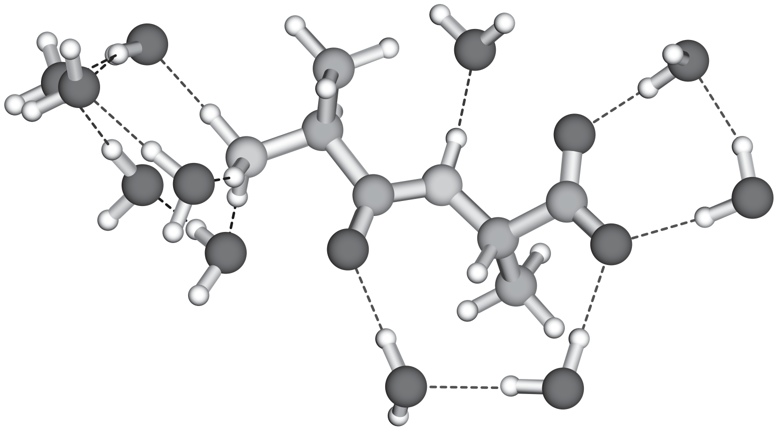

Fig. 1.3 L-alanyl-L-alanine zwitterion stabilized by 11 water molecules. B3PW91/6-311G2dp plus PCM continuum solvent model optimized geometry. (Please find a color version of this figure on the color plates.)

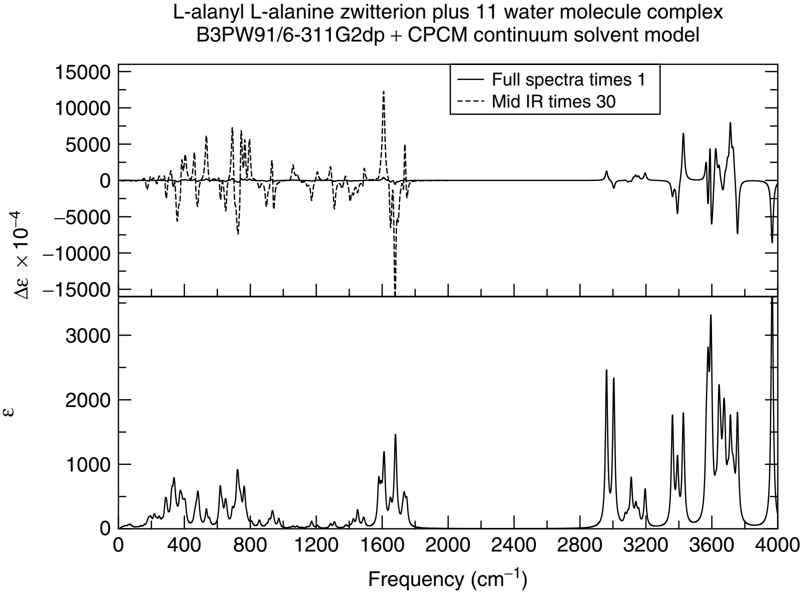

Fig. 1.4 L-Alanyl-L-alanine zwitterion stabilized by 11 water molecules. B3PW91/6-311G2dp plus PCM continuum solvent model for VA and VCD spectra.

CPCM continuum solvent models as implemented in Gaussian. The structures were extracted from classical MD simulations of the LALA zwitterion in a box of water molecules (Knapp-Mohammady *et al.*, 1999; Jalkanen *et al.*, 2003, 2009b). Only the waters which were directly hydrogen bonded with the polar groups were kept initially for the full DFT level calculations. This was the protocol which

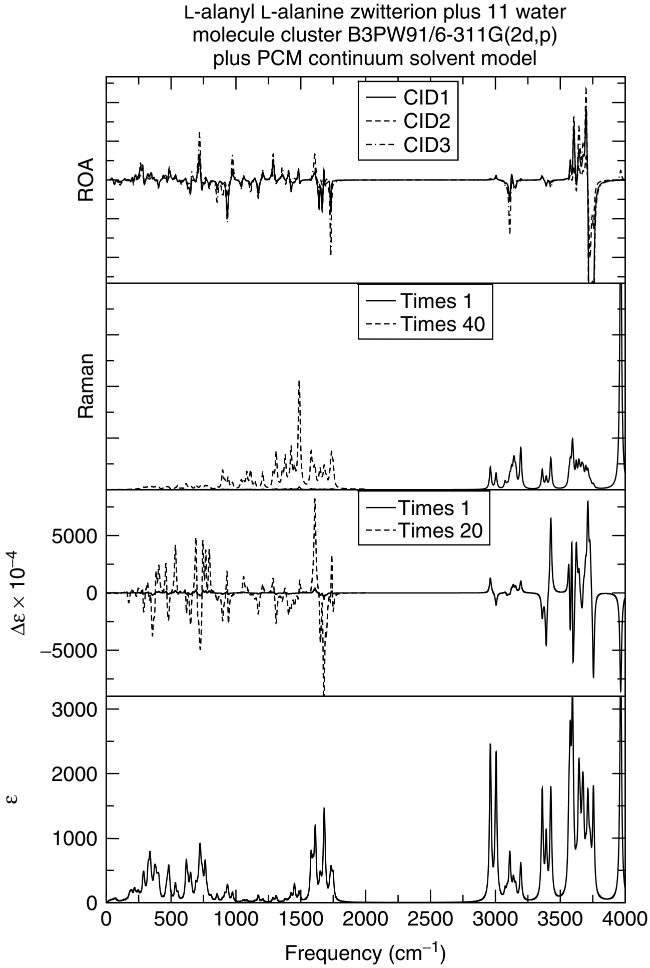

Fig. 1.5 L-Alanyl-L-alanine zwitterion stabilized by 11 water molecules. B3PW91/6-311G2dp plus PCM continuum solvent model for VA, VCD, Raman, and ROA spectra.

worked for LA and NALANMA, but it failed for the LALA zwitterion initially, as mentioned above. Since there is a competition for the hydrogen-bonded partners, some of the water molecules which were initially hydrogen bonded with the polar groups of the LALA zwitterion were pulled away, and hence polar groups of the LALA zwitterion ended up with no hydrogen bonding partners. Our simple solution to this problem was to explicitly add another water molecule to the polar group which now lacked a hydrogen-bonded water partner. The more general problem, which we have used for LA, is to optimize the structure of the solute plus a complete solvation shell. This is of course a much more stable entity than just the solute with the water molecules directly hydrogen bonded to the polar groups of

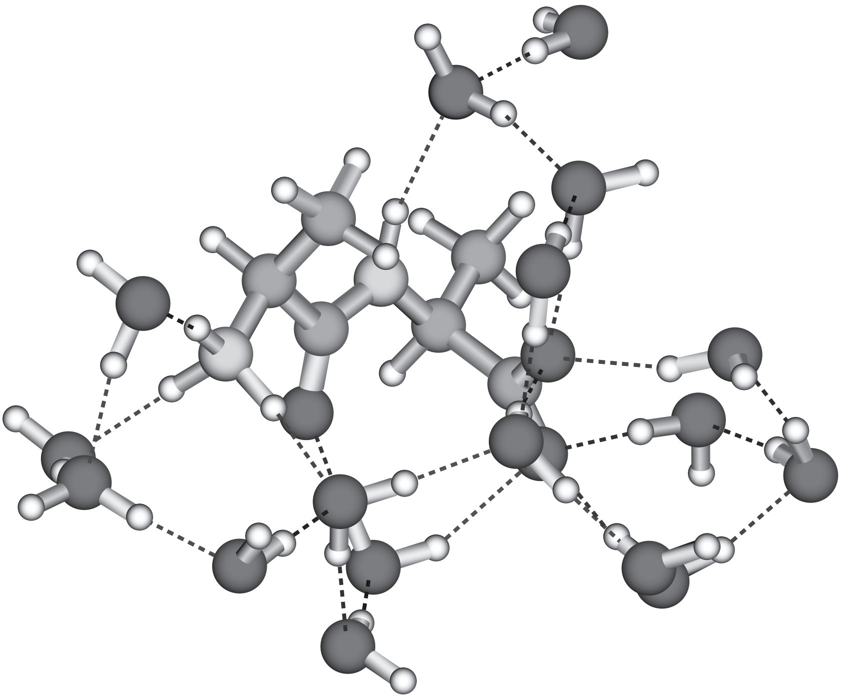

Fig. 1.6 L-Alanyl-L-alanine zwitterion stabilized by 17 water molecules. B3PW91/6-31G* plus PCM continuum solvent model optimized geometry. (Please find a color version of this figure on the color plates.)

the solute. In many cases the hydrogen bonding partners are necessary to orient and stabilize the water molecules that are hydrogen bonded to the polar groups (Degtyarenko *et al.*, 2007, 2008; Jalkanen *et al.*, 2008a).

Our goal is to try to interpret the measured VA, VCD, Raman, and ROA spectra and to be able to determine the structure or structures present under the condition of the experiment. If one cannot determine the conformer and species for simple peptides, one could not be expected to do so for polypeptides like Leu- and Met-enkephalin or proteins. If one wishes to get more than simple secondary structural elements present, which is the backbone and side chain torsion angles for all residues and the protonation state of all polar residues, including the N- and C-termini, then one must extend the methods of vibrational spectroscopy to many dimensions, as was done for NMR, which finally allowed one to get a complete 3D structure. In many cases, in addition to the amino acid, peptide, and protein, one must also include and determine the position of certain water molecules, cations, anions, ligands, and prosthetic groups which are responsible for stabilizing the structure of the molecular complex.

To determine the structure of the LALA zwitterion in aqueous solution requires a conformational search for the lowest energy zwitterionic structure of LALA. The results here only document the feasibility of explicitly adding water to the

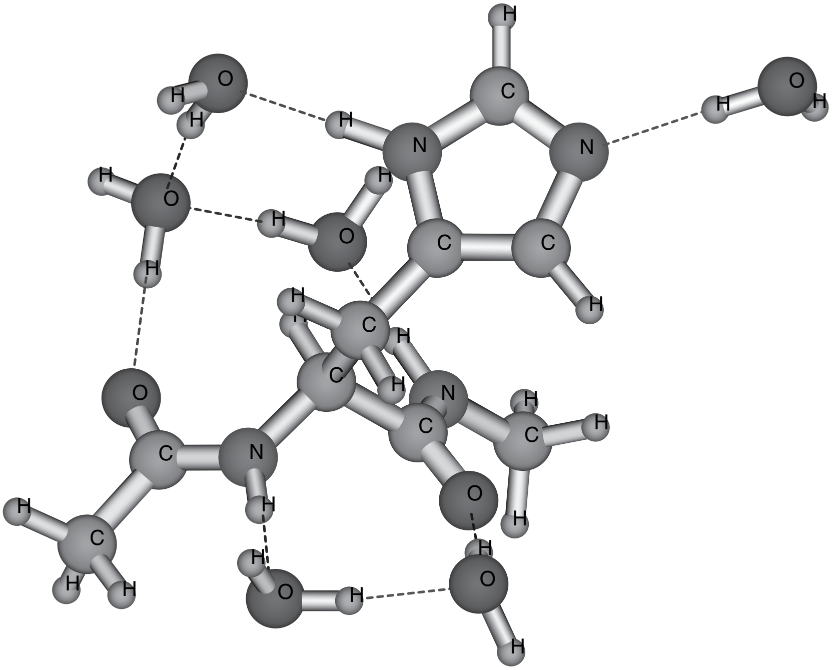

Fig. 1.7 *N*-Acetyl L-histidine *N*′-methylamide stabilized by six water molecules. B3PW91/6-31G* plus PCM continuum solvent model optimized geometry. (Please find a color version of this figure on the color plates.)

calculation of the zwitterionic structures of dipeptides and the calculation of the VA and VCD spectra. Our goal here is be able to determine the structure(s) of the zwitterionic state of LALA in aqueous solution. Our preliminary results show that the VCD spectra is very sensitive to the structure of the zwitterion of LALA and one should be able to distinguish which conformer(s) is(are) present if one knew what the predicted VCD spectra for each conformer was. We are pursuing the determination of the other possible structures of the zwitterion of LALA and will report the complete results in a future publication (Jalkanen *et al.*, 2009b).

In Table 1.5, we present our calculated vibrational frequencies, VA and VCD intensities, for three of the structures presented in Tables 1.4a and 1.4b.

1.5.2.4 **L-Histidine**

In addition to LA, NALANMA, and LALA, one is of course interested in other amino acids. One of the most interesting and functionally important amino acids is LH. Recently we have reported a combined experimental and theoretical study of LH in aqueous solution (Deplazes *et al.*, 2008). For complete details on this work one is referred to that paper. Here we only summarize the results. To determine the structure of LH in aqueous solution one must first stabilize the zwitterionic part of the amino acid. Here we can use the structure of the LA zwitterion, which we determined in our earlier works (Jalkanen, Bohr and Suhai, 1997; Tajkhorshid, Jalkanen and Suhai, 1998; Frimand *et al.*, 2000; Jalkanen *et al.*,

Tab. 1.5 L-Alanyl-L-alanine zwitterion VA and VCD spectra.

ν(cm^{-1})^a	$A_i{}^b$	$R_i{}^c$	ν(cm^{-1})^a	$A_i{}^b$	$R_i{}^c$	ν(cm^{-1})^a	$A_i{}^b$	$R_i{}^c$
	III			II			I	
1745	95.12	−249.30	1793	172.50	−106.04	1764	37.59	−100.23
1711	269.18	182.81	1769	198.20	116.73	1742	441.64	8.05
1701	284.78	−418.71	1711	641.68	157.82	1718	413.77	304.51
1678	187.10	193.42	1680	67.76	−105.94	1696	73.09	−64.26
1625	265.53	388.67	1615	148.81	10.31	1635	27.49	30.81
1588	190.59	94.32	1586	200.83	70.66	1625	162.49	−67.03
1532	2.92	3.80	1526	19.72	2.27	1530	3.88	12.57
1527	23.00	−25.74	1524	13.84	−15.51	1523	10.48	13.03
1522	2.33	−6.22	1518	7.16	−1.29	1523	5.72	−4.90
1520	2.67	3.40	1516	7.73	6.35	1516	15.82	3.13
1447	9.43	15.76	1440	24.90	2.70	1453	25.63	1.29
1439	40.84	11.22	1432	117.64	24.21	1431	102.61	11.16
1416	24.09	−12.95	1418	113.12	52.53	1419	45.70	51.86
1395	232.67	−161.32	1414	88.42	−132.80	1409	67.20	97.46
1372	9.22	−14.27	1393	11.50	73.31	1404	88.05	−201.64
1359	2.81	0.36	1384	44.91	−109.25	1372	12.27	21.19
1329	43.83	−129.21	1303	107.24	−175.64	1334	83.33	−69.71
1309	27.94	74.51	1275	28.14	−21.82	1292	68.84	22.65
1240	15.70	−23.91	1266	188.86	201.30	1287	15.75	18.11
1177	114.51	4.39	1203	42.40	−69.49	1210	21.24	11.89
1162	15.55	43.66	1150	4.67	−34.19	1146	13.12	1.91
1124	8.32	−40.32	1137	24.69	49.74	1141	8.25	14.75
1110	23.95	−6.83	1112	26.83	−17.96	1119	15.40	−9.43
1088	11.17	11.86	1070	3.69	−3.71	1072	1.98	−9.97
1088	11.17	11.86	1070	3.69	−3.71	1072	1.98	−9.97
1061	10.17	6.54	1049	11.13	17.96	1067	27.22	3.84
1018	7.54	−9.04	1031	10.09	−6.41	1035	2.63	−11.54
942	8.10	−10.99	949	14.38	−45.67	953	14.15	−28.20
925	8.47	−29.98	930	7.62	−11.18	924	1.78	−8.94
883	8.91	−10.74	883	13.21	11.28	883	8.71	10.84
863	52.91	−33.16	854	30.93	−5.04	853	20.83	16.04
800	49.65	34.95	804	15.15	30.85	791	12.67	37.58
797	14.39	−3.63	771	6.69	22.72	771	6.78	−28.23
750	62.17	−10.17	708	79.82	21.54	726	39.60	9.79
693	14.89	9.25	676	20.33	0.12	665	18.49	4.71
629	1.56	−21.74	581	12.38	−42.27	601	8.46	−24.13
586	3.27	−9.87	550	3.88	−49.09	543	17.07	3.19
535	9.27	−9.73	548	11.19	19.08	531	7.81	−37.80
505	11.96	32.14	468	11.67	25.32	486	6.19	7.61
394	34.00	16.21	447	9.32	36.02	443	32.06	46.97
385	13.66	23.12	368	4.97	5.29	366	8.57	21.29
342	6.62	−49.92	361	4.56	32.73	361	11.85	14.20
303	3.18	13.31	325	10.25	−37.13	311	4.21	−27.76
276	27.88	37.18	291	19.00	61.69	292	13.35	41.52
253	4.52	15.10	251	0.32	−6.70	254	2.37	−9.71
245	1.13	6.52	236	14.52	25.11	246	6.49	−11.81
222	10.18	−51.44	230	14.44	−59.15	235	5.07	7.62

Tab. 1.5 (*Continued*)

$\nu(\text{cm}^{-1})^a$	A_i^b III	R_i^c	$\nu(\text{cm}^{-1})^a$	A_i^b II	R_i^c	$\nu(\text{cm}^{-1})^a$	A_i^b I	R_i^c
193	27.45	−46.99	189	10.04	1.79	197	26.41	0.31
172	9.89	−15.30	174	12.99	−8.18	170	26.67	−53.71
165	2.71	26.63	134	4.57	−25.58	157	6.95	33.04
142	6.99	18.80	105	5.51	15.96	124	5.04	−44.78
125	2.36	−9.23	83	6.30	−13.56	109	0.44	2.06

[a] 6-31G* B3LYP frequencies.
[b] 6-31G* B3LYP APT.
[c] 6-31G* B3LYP APT and 6-31G** RHF DO gauge AAT.

2001) and mutate the methyl group to an imidazole group. Here we have two polar groups, the ring N–H group and the N atom. Following our previous work, we have added two water molecules to hydrogen bond with each group. This is our minimal explicit solvent model. The question we wished to address was what are the two side chain angles, χ_1 and χ_2. Hence we have optimized the structure for the various possible conformers. For the C_α–C_β bond, we have a sp^3–sp^3 bond, and hence one would expect three possibilities, gauche plus, trans, and gauche minus. For the C_β–C_γ bond, we have a sp^3–sp^2 bond, and hence one would expect various possibilities depending on the interactions between the various substituents. Here one could expect cis and trans, in addition to gauche plus and gauche minus. There is the possibility of the N–H group hydrogen bonding with the C=O group of the backbone or with one of the water oxygen atoms. In the presence of explicit water molecules, the backbone C=O oxygen atom would be screened from the ring NH hydrogen atom. Hence one would expect different results for the gas phase isolated state, for implicit solvent models (Onsager, PCM, CPCM, and/or COSMO), for explicit solvent models, and finally for our hybrid model, where we embed the LH + N water molecule complexes within the continuum solvent models. From our comparison with the VA, VCD, Raman, and ROA spectra measurements we concluded that it is important to include explicit water molecules and additionally that the gauche minus gauche minus conformer is the most likely one to be present, though we were not able to definitely rule out the presence of the gauche minus gauche plus.

We have previously presented the structures for the LH zwitterion calculated at the B3LYP/6-31G* plus PCM level of theory and the interested readers can refer to those works (Jalkanen *et al.*, 2006a; Deplazes *et al.*, 2008). In the next subsection, we present our new work on capped LH.

1.5.2.5 *N*-Acetyl L-Histidine *N*′-Methylamide

Similar to capping LA to produce capped LA, one can cap LH to produce capped LH. Here one has the backbone ϕ and ψ angles in addition to the χ_1 and χ_2 side chain angles. In Tables 1.6a,

Tab. 1.6a *N*-Acetyl-L-histidine *N*′-methylamide conformational energies.

Conformer	ϕ	ψ	χ_1	χ_2	Energy Hartrees	
gm gm	−162.26	173.64	−127.55	61.85	−720.613603718	Isolated[a]
gm gm	−124.91	132.35	−58.49	−78.23	−720.641205272	PCM[b]
gm gm	−119.14	128.43	−61.29	−87.41	−720.641751542	CPCM[c]
gm gp	−106.07	149.36	−68.19	57.63	−720.603271434	Isolated[a]
gm gp	−115.85	130.42	−61.20	85.91	−720.641232605	PCM[b]
gm gp	−116.36	131.30	−62.09	85.19	−720.641609495	CPCM[c]
gp gm	−151.60	177.47	67.74	−59.58	−720.605035221	Isolated[a]
gp gm	−142.19	161.99	62.14	−89.03	−720.640392649	PCM[b]
gp gm(ts)	−144.49	163.55	64.23	−86.38	−720.640865462	CPCM[c]
gp gm	−143.48	163.70	63.45	−86.64	−720.640832768	CPCM[c]
gp gp	−82.88	67.07	49.25	69.20	−720.618655982	Isolated[a]
gp gp	−135.38	126.74	43.41	52.26	−720.639519222	PCM[b]
gp gp	−141.30	129.55	79.86	49.23	−720.640240001	CPCM[c]
t gm	−82.84	67.74	−168.94	−70.77	−720.613246215	Isolated[a]
t gm	−132.57	118.03	179.93	−87.97	−720.641942617	PCM[b]
t gm	−135.20	119.68	179.83	−87.82	−720.641957463	CPCM[c]
t gp	−162.26	173.64	−127.54	61.85	−720.613603651	Isolated[a]
t gp (ts)	−126.44	125.05	−178.17	73.90	−720.641944678	PCM[b]
t gp	−130.52	125.55	−178.28	73.65	−720.641923448	PCM[b]
t gp	−142.74	130.04	−179.65	72.59	−720.642396472	CPCM[c]
t gp(6-31g**)	−134.57	126.85	−177.99	73.36	−720.667446267	CPCM[c]

[a] NALHNMA 6-31G* B3PW91 relative energies.
[b] NALHNMA 6-31G* B3PW91 + PCM relative energies.
[c] NALHNMA 6-31G* B3PW91 + CPCM relative energies.

1.6b, and 1.6c, we present the structures for NALHNMA, NALHNMA plus four water molecule complex, and NALHNMA plus six water molecule complex at the B3PW91/6-31G*, B3PW91/6-31G* plus PCM continuum solvent model and B3PW91/6-31G* plus CPCM continuum solvent model level of theory (Jalkanen *et al.*, 2009a). As one can see in Table 1.6a for the gauche minus gauche minus conformer, structures which are stable within the continuum solvent model are not stable without it. Here the continuum solvent model has effectively taken into account the screening due to to the solvent. In the absence of both explicit solvent molecules or the continuum solvent model, the histidine side chain rotates to form a strong hydrogen bond between the ring N–H group and the *N*-acetyl group C=O group. Hence the importance of taking into account solvent effects. The VA, VCD, Raman, and ROA spectra for all species will be presented in a future work (Jalkanen *et al.*, 2009a). In Figure 1.7, we present one of the conformers of NALHNMA stabilized by the presence of six explicit water molecules (NALHNMA hydrated complex) which has been embedded with the PCM continuum model that treats the effects due to the bulk water molecules.

Tab. 1.6b *N*-Acetyl-L-histidine *N*′-methylamide plus four water molecule complex conformational energies.

Conformer	ϕ	ψ	χ_1	χ_2	Energy Hartrees	
gm gm	−111.17	123.21	−52.37	−50.94	−1026.22240779	Isolated[a]
gm gm	−100.52	126.20	−64.60	−88.19	−1026.25855737	PCM[b]
gm gm	−100.43	126.23	−63.78	−87.51	−1026.25912256	CPCM[c]
gm gp	−106.58	120.16	−61.32	145.72	−1026.21352819	Isolated[a]
gm gp	−97.22	124.77	−65.36	93.20	−1026.25611672	PCM[b]
gm gp	−99.94	126.74	−74.53	93.88	−1026.25703059	CPCM[c]
gp gm	−130.40	155.34	83.26	−58.20	−1026.21719542	Isolated[a]
gp gm	−108.08	153.79	86.22	−67.84	−1026.25795559	PCM[b]
gp gm	−108.42	153.92	85.79	−67.44	−1026.25846269	CPCM[c]
gp gp	−151.34	115.42	42.30	54.74	−1026.21296490	Isolated[a]
gp gp	−118.97	135.83	59.70	93.17	−1026.25224287	PCM[b]
gp gp	−144.22	144.79	58.21	94.61	−1026.25444179	CPCM[c]
gp gp	−146.21	143.91	58.24	95.18	−1026.25454373	CPCM[c]
t gm	−104.15	114.33	176.24	−77.11	−1026.21907233	Isolated[a]
t gm	−98.87	127.28	−174.74	−92.38	−1026.25819396	PCM[b]
t gm	−98.65	126.85	−174.47	−92.48	−1026.25917833	CPCM[c]
t gp	−130.97	119.11	−163.46	66.23	−1026.22157806	Isolated[a]
t gp	−96.83	123.70	−179.36	82.00	−1026.25900659	PCM[b]
t gp	−96.82	123.67	−179.38	82.02	−1026.25896453	PCM[b]
t gp	−96.96	123.60	−179.40	82.00	−1026.25948994	CPCM[c]
t gp	−96.97	123.58	−179.41	81.01	−1026.25944361	CPCM[c]

[a] NALHNMA plus four water molecule complex 6-31G* B3PW91 relative energies.
[b] NALHNMA plus four water molecule complex 6-31G* B3PW91 + PCM relative energies.
[c] NALHNMA plus four water molecule complex 6-31G* B3PW91 + CPCM relative energies.

1.5.2.6 *N*-Acetyl L-Alanine L-Proline L-Tyrosine *N*′-Methylamide

In Figure 1.8, we present one of the 14 structures we have found to date for capped L-alanine L-proline L-tyrosine, *N*-acetyl L-alanine L-proline L-tyrosine *N*′-methylamide at the B3PW91 DFT level of theory. The structures have been extracted from classical molecular dynamics simulations (Nardi *et al.*, 2000; Nardi, 1999; Nardi, Worth and Wade, 1997) and the waters have been removed and the structures have been optimized with the COSMO/PCM hybrid continuum solvation model as implemented in Gaussian 03 (Jalkanen *et al.*, 2009c). Here we have not added explicit water molecules but have only used the continuum solvent treatment to see how well one can do with these larger peptides without actually including the explicit water molecules. In the case of the Raman and ROA spectral simulations, the inclusion of the explicit water molecules increased the cost of the Raman and ROA simulations, but with the new analytical methods developed by Ruud, the tensors required for the Raman and ROA spectral simulations are not longer rate limiting. Here one can calculate the

Tab. 1.6c *N*-Acetyl-L-histidine *N*′-methylamide plus six water molecule complex conformational energies.

Conformer	ϕ	ψ	χ_1	χ_2	Energy Hartrees	
gm gm	−103.93	115.93	−33.39	−65.51	−1179.02483018	Isolated[a]
gm gm	−107.63	115.49	−51.33	−66.53	−1179.06684481	PCM[b]
gm gm	−110.35	114.40	−51.87	−70.85	−1179.06740734	CPCM[c]
gm gm	−110.41	114.37	−51.89	−70.88	−1179.06740482	CPCM[c]
gm gp	−118.80	117.55	−52.05	111.03	−1179.01745486	Isolated[a]
gm gp					−1179.06371029	PCM[b]
gm gp	−118.55	116.37	−49.90	122.50	−1179.06482351	CPCM[c]
gp gm	−154.56	155.55	38.22	−101.72	−1179.01236234	Isolated[a]
gp gm	−144.70	148.25	25.69	−94.96	−1179.01307645	isolated[a]
gp gm	−116.83	149.29	61.84	−100.02	−1179.05500717	PCM[b]
gp gm	−116.44	152.83	63.15	−98.10	−1179.05524477	CPCM[c]
gp gp	−104.27	142.30	42.07	76.01	−1179.01765151	Isolated[a]
gp gp	−112.30	141.75	49.26	76.04	−1179.05968412	PCM[b]
gp gp	−112.28	141.68	49.23	76.14	−1179.06008150	CPCM[c]
t gm	−107.25	124.15	−174.10	−97.90	−1179.02992116	isolated[a]
t gm	−106.07	123.41	−179.19	−104.51	−1179.06675368	PCM[b]
t gm	−106.57	123.11	−179.52	−103.79	−1179.06717558	CPCM[c]
t gp	−113.35	122.95	−172.82	77.92	−1179.01858560	Isolated[a]
t gp	−110.05	116.25	−177.35	78.33	−1179.06840828	PCM[b]
t gp	−110.07	116.27	−177.35	78.36	−1179.06888670	CPCM[c]

[a] NALHNMA plus six water molecule complex 6-31G* B3PW91 relative energies.
[b] NALHNMA plus six water molecule complex 6-31G* B3PW91 + PCM relative energies.
[c] NALHNMA plus six water molecule complex 6-31G* B3PW91 + CPCM relative energies.

Hessian, VA, and VCD spectra in one calculation and the tensors required for Raman and ROA in a separate calculation, and then combine the results, that is, the Hessian, normal modes (eigenvalues and eigenvectors) from the first calculation with the EDEDPD, EDMDPD and the EDEQPD from the second, to get the most accurate Raman and ROA intensities. This was advocated earlier for the Raman intensities by Amos (1986), as the basis sets and level of theory to get accurate geometries and Hessians appear to be different than that required for accurate EDEDP, EDMDP, EDEQP and the EDEDPD, EDMDPD and the EDEQPD. But what one must at least be able to get is the structure/conformer which is present in aqueous solution. In our geometry optimizations with the waters removed, the structures stabilized by the waters were not stable (Jalkanen *et al.*, 2009c). This is consistent with our previous work on the alanine dipeptide, showing that this phenomenon for alanine peptide may not be an exceptional case, but the general case. Hence we advocate the use of our hybrid model where one includes the first solvation shell water molecules to treat the effects due to these waters and then embed this hydrated molecule within a continuum

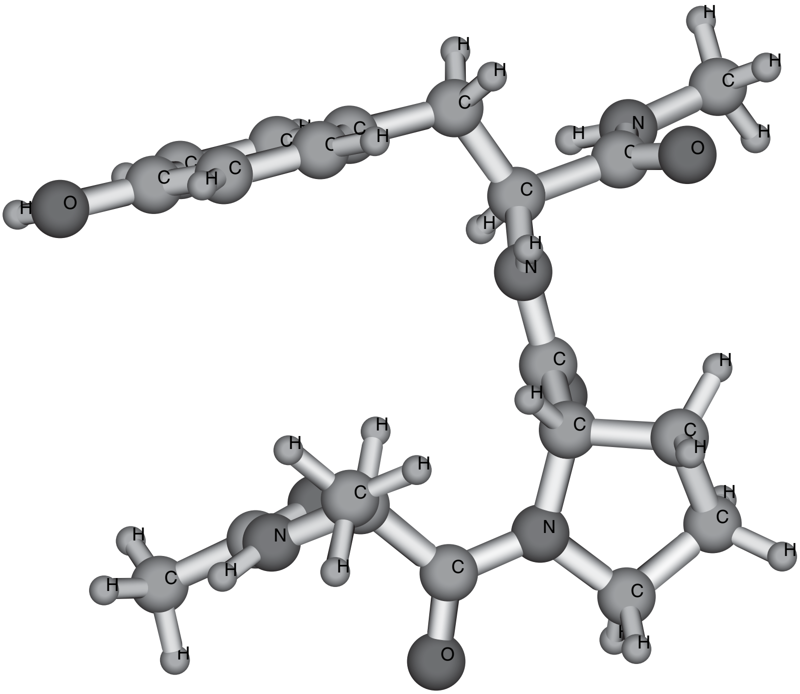

Fig. 1.8 *N*-Acetyl L-alanine L-proline L-tyrosine *N*′-methylamide molecule. B3PW91/6-31G** plus PCM continuum solvent model optimized geometry. (Please find a color version of this figure on the color plates.)

model to treat the effects due to bulk water (Jalkanen *et al.*, 2008a; Deplazes *et al.*, 2008; Frimand *et al.*, 2000; Han *et al.*, 1998; Tajkhorshid, Jalkanen and Suhai, 1998; Jalkanen, Bohr and Suhai, 1997; Jalkanen and Suhai, 1996). In some cases one may even have to include a second layer of water molecule. But the largest and most important effects appear to be due to only the first solvation/hydration layer, so this should be included minimally.

1.5.3
Future Perspectives and Applications for DFT-based Simulations

In this work, we have presented and discussed on the use of DFT simulations and VA, VCD, Raman, and ROA spectroscopy to investigate the conformers of the ligands and proteins and the changes which occur to both on binding. There are of course relevant other approaches to the same problem (Bruno *et al.*, 1997). In addition, the binding of insulin, insulin mimetics, and other ligands to the insulin receptor is an important problem. The insulin receptor is transmembrane receptor with insulin binding to the extracellular part. Recent work has tried to characterize a second ligand binding site of the insulin receptor (Hao, Whittaker and Whittaker, 2006). In addition to the binding of insulin to the extracellular domain, the insulin receptor has an intracellular (cytoplasmic) tyrosine kinase domain, the role of which has also been heavily investigated (Chou, 1990). Here we feel that VA, VCD, Raman, and ROA studies can contribute a lot to a better understanding of the structural changes which occur and to the mechanism/mode of action. Finally, in the last subsection of this work, we extend the multidisciplinary approach to more general problems in molecular medicine and in society in general.

1.5.4

Future Perspectives of the Multidisciplinary Fields of Molecular Biophysics, Molecular Biology, Systems Biology, and Molecular Medicine and Molecular Psychiatry

The future of chemistry, physics and mathematics appears to be secure and healthy through the application and use of the fundamental knowledge gained in these fields in the new fields of molecular biophysics, bionanotechnology and quantum nanobiology (Goodsell, 2002). In addition, the new field of molecular psychiatry has developed to the point where some behavioral disorders have been shown to be correlated with specificity genotypes (Kim-Cohen *et al.*, 2006). Monoamine oxidase A (MAOA) plays a role in regulating behavior by selectively degrading serotonin, norephinephrine, and dopamine following reuptake in the synaptic cleft. If MAOA is underexpressed, children who have been maltreated have been shown to develop conduct disorder, violent behavior and antisocial personalities in adulthood. On the other hand, children who overexpress MAOA do not appear to be affected. These results are very controversial and have not been fully accepted by everyone. Clearly more work needs to be done in this area. The possibility of preventing and/or hindering the antisocial behaviours, conduct disorders, and violent reactions and behaviours by either gene therapy, drugs, and counselling, or a combination of the above would go a long way to solve this problem, which formerly was thought to be only due to poor upbringing and socialization. With the ever increasing prison populations in the United States and China, clearly, any alternatives are well worth exploring. In addition, many of those currently in our prisons and criminal justice system need to be tested for genetic problems that may be partly responsible for behaviours and actions inherent in the genetic makeup.

Hence, in addition to determining tendencies for diabetes, high blood pressure, heart disease, and cancer, there now appear to be genetic markers that are able to predict tendencies for antisocial and combative behavioral patterns when the individuals are psychologically, emotionally, and physically stressed (maltreated, sexually molested, physically abused, and mobbed) as children or young adults. The increasing incidences of violent behavior like the Columbine and Virginia Tech shootings may be able to be prevented by genetically testing young students who have been and continue to be mobbed and socially ostracized. Rather than demonize these young people and their behavior, more research needs to be undertaken to understand the root causes of such behavior and the triggering mechanisms. Many veterans and survivors of the Korean, Vietnam, Gulf 1 and 2, Afghanistan, and Balkan wars have shown similar behavior after being exposed to inhumane treatment and/or seeing atrocities. The increase of extremism appears to be an ever increasing problem and there does not appear to be a lack of researchers trying to shed some light on this international problem (Gambetta, 2005; Abbas, 2005). In addition to psychological, part of the problem may indeed be biochemical. No one is saying that this is the full answer, but this field of research is well worth pursuing. Just like biomolecule interacting in various manners in various environments, it appears that many behavior patterns are the results of a complex set of physical, psychological, societal, economic, religious, and biochemical interactions. Ulcers were thought to be caused by stress, but subsequently were found to be caused by microorganisms, which when the immune response was weakened by other factors, were able to have adverse side effects that did not appear without the physical, mental, and psychological stress.

But these multidisciplinary and multifaceted fields require the researchers and scientists to be on top of many areas of endeavor. The pyramid principle of knowledge never made itself more apparent than in these problems. The idea of social workers, police and military officers, and our criminal justice worker, judges, and politicians are competent to understand these problems is really a fallacy. The systems are broken and this is shown by the record numbers of people being incarcerated in both the United States and China. When having to deal with complex problems, the simple black and white, with us or against us model is clearly not working. The world is not those who are with us and those who are against us. We have learned a lot from complex biological systems to see that the human body is a very complex machine and that when one try to fix one system with a drug, in many cases this causes adverse side effects, which in many cases are even worse than the symptoms one was/is trying to treat. The real root cause of the problem and/or disease is not even known or dealt with. This is the same in many societies, where one tries to treat and/or deal with the symptoms of some social, religious, economic, societal, and/or political problem, and one does not deal with the root cause or causes. We must be smarter and use the scientific method, logic, reasoning, and philosophy, to get to the root causes of the problems we are facing in society, be it behavioral, environmental, economic, and/or with the public and individual health.

Tackling the multifaceted and multidisciplinary problems has changed the paradigm for many departments, funding agencies, government agencies, educational and training institutions, and publishers. The cross-fertilization of knowledge and techniques has never been more fruitful. What has prevented this interaction in the past has been the artificial barriers between what was mathematics, physics, chemistry, biology, medicine, political science, psychology, sociology, anthropology, religious studies, criminal justice, economics, and environmental sciences. We now realize that these barriers were/are really artificial, and, in many cases, these are responsible for some our global problems, like air and water pollution, chemical poisoning of our environment, high increase of allergies, global warming, world hunger, epidemics, terrorism, and the economic crises being experienced by many countries. By introducing new technologies, governmental policies, economic systems, chemicals, and even alternative energy sources, one has bettered mankind and the quality of life in one area, but has had long-term detrimental effects in other areas. The idea that free trade and market economies would be the panacea for increasing the standard of living for everyone on the planet has been a big myth. The trade unions (organized labor), environmentalists, and religious leaders have not been included at the World Trade Organization (WTO) and G8 summit meeting. There has not been a consensus and trust from large segments of society. This has been due to the lack of experts and ideas from many of the major players: organized labor, environmentalists, religious leaders, and native indigenous nations which have been occupied, but who have almost no political, economic, and military strength (Tibetans, Tamil Tigers, Native Americans, Mexican Americans displaced by the land grab by the United States from Mexico: now the states of California, Nevada and Utah, and parts of the states of Wyoming, Texas, Arizona, Colorado, and New Mexico). Many of the 10 million so-called illegal aliens in the United States are a direct result of the failed policies and treaties resulting from the American–Mexican war. A settlement by politicians in Washington DC and Mexico City did not deal with the border problem. This was similar to the problem during the cold war between East and West, but in the Cold War the United States and the West welcomed Eastern Europe with open arms. Many so-called Americans are not so receptive now to accept the children and grandchildren of the Mexican Americans who had their land and resources stolen from them during the American–Mexican war. An open door policy similar to the expanded EU is the only fair, moral, ethical, and social solution of the American-Mexican border. Rather than build another Berlin Wall to keep the Mexican Americans out, the President Barack Obama should tear down the Bush/Republican Berlin Wall. Ronald Regan had it right when he told the Russians to tear down the Berlin wall and now many Mexican Americans and Americans are totally right when they told Bush and the Republicans and now tell Obama and the Democrats to open up the Mexican–American border. It will be a win win for both countries. Of course the elite in both countries who are responsible for the depressed economic conditions in both countries want a status quo. America can be like Finland (bilingual Swedish and Finnish), Switzerland (trilingual German, French, and Italian), and Belgium (bilingual French and Dutch). Note that the people of these three super economies and socially just countries also learn and are fluent in English. It is high time that Mexico and the United States (and also Canada) open their borders and expand like what EU has. If the EU can add 12 new countries and hopefully in the near future Turkey, then the United States, Canada, and Mexico can open their borders and have not only free trade, but free movement of people, thoughts, ideas, religions, resources, and economic wealth and capital. North Americans should all be Americans and not Mexicans, Canadians, Cubans and US-Americans, just like Europeans are no longer Germans, French, Spanish, Portuguese, Finnish, Swedish, and so on. We need to unite and be one, and not continue to separate and deny our brothers and sisters their birthrights and rights as fellow inhabitants of this planet.

The aim of this chapter has been a multidisciplinary and generalist approach to the problems in quantum nanobiology and how they can and have been tackled and approached by various researchers in different fields of endeavor: mathematicians, physicists, chemists, and biologists. Clearly in the future, the next generation of scientists, researchers, and even nontechnology working citizens must be better educated with respect to nature and how we and our machines, devices, vehicles, and energy sources interact with the environment and the world we live in, plant and animal life, as well as our most valuable, but taken for granted resources, pure and clean drinking water. The fate of future generations depends on the next generation being made better aware of the known and possible consequences of newly developed and being developed technologies, so that we can not only maintain our high standard of living in the Western societies but also expand it to the largest two developing nations India and China without them both having the same adverse side effects on their native environments and political systems as we have had. The goal and aim to learn from one's mistakes and errors and from those of others means that we must first admit when we have made errors, fully document them in the open literature, and not take draconian measures against honest mistakes that were made due to lack of knowledge. This will only keep the errors hidden and under the surface and prevent the root causes from being determined. In many cases, the errors and/or failures were due to the systems/approaches which were being used and the inherent assumptions. The more open the discussions and more accessible the data is made, the more honest, caring, and open researchers will tend to be. The faster the problems can be recognized, the quicker the solutions can be made. The withdrawal of the wonder drug Vioxx is a classical example of the new and open system in the field of pharmaceuticals. As soon as problems were recognized the drug was pulled which has allowed top researchers in the whole world to start working on determining and understanding the reasons which caused the adverse side effects, with the hope and goal that the next generation of drug should be even safer than the current generation. Merck is clearly a world leader in not only developing new and better drugs but also in the way to proceed as soon as one recognizes and/or is informed of possible adverse side effects due to one's new drug or technology (small nanoparticles, genetically modified organisms causing allergic reactions for some people).

Not just corporate houses, but each and every scientist, manager, member of the board of directors, and also the Wall Street investment houses need to be more concerned about the long-term sustainability of the world's environment and global standard of living and development instead of focusing only on the short-term bottom line, even if stock options and year end bonuses might suffer and/or be reduced. We owe this to the next generation which will follow us and have to deal with the consequences of our actions, or in many cases, lack of actions, or responses to known and/or possible adverse side effects. We can in many cases no longer claim ignorance. The goal is to rid the world of ignorance and educate not only scientists but also industrialists, politicians, investment bankers, venture capitalists, and of course the citizens and noncitizens globally. A very challenging task, but working with top universities and publishers like John Wiley & Sons, Inc. surely doable, if we put our hearts, souls, and minds to the great task at hand!

Acknowledgments

K.J. Jalkanen (KJJ) would like to thank the Deutscher Akadmemisher Austausch Dienst (DAAD) [German Academic Exchange Service], the Deutsches Krebsforschungzentrum (DKFZ) [German Cancer Research Center] in Heidelberg and the EU for making his five year stay at the German Cancer Research Center and the Steinbeiss Transfer Centrum possible. In addition, he would like to thank the Finnish Academy of Science for sponsoring many visits to Professor Risto Nieminen's Laboratory of Physics Laboratory in Helsinki University of Technology in Espoo, Finland. We would like to thank the Danish Foundation for Fundamental Research (www.dg.dk) for funding the establishment of the Quantum Protein (QuP) Centre in the Department of Physics at the Technical University of Denmark where KJJ has been an associate professor of biophysics before moving to the Nanochemistry Research Institute at Curtin University of Technology where he is currently on unpaid sabbatical leave from QuP and being funded by the Western Australian Premier's Research Fellowship program (grant awarded to Professor J.D. Gale, Western Australian Premier's Research Fellow). KJJ would also like to acknowledge and thank iVEC (The hub of advanced computing in Western Australia) in Perth, APAC (Australian Partnership for Advanced Computing) in Canberra, and the ARC for providing computational, people and financial resources which made parts of this work possible during KJJ's three-year sabbatical stay in Perth at Curtin University of Technology in the Nanochemistry Research Institute (NRI) in the Department of Applied Chemistry. Finally KJJ would like to thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) and the DAAD for financial support for his five months stay in the laboratory of Professor Thomas Frauenheim at the Bremen Center for Computational Material Science at the University of Bremen.

References

Abbas, H. (2005) *Pakistan's Drift into Extremism: Allah, the Army, and America's War on Terror*, M.E. Sharpe.

Abdali, S., Jalkanen, K.J., Bohr, H., Suhai, S., and Nieminen, R.M. (2002) *Chem. Phys.*, **282**, 219.

Amos, R.D. (1982) *Chem. Phys. Lett.*, **87**, 23.

Amos, R.D. (1984) *Chem. Phys. Lett.*, **108**, 185.

Amos, R.D. (1986) *Chem. Phys. Lett.*, **124**, 376.

Amos, R.D. (1987) Molecular property derivatives, in *Ab Initio Methods in Quantum Chemistry*, John Wiley and Sons Ltd, pp. 99–153.

Amos, R.D., Handy, N.C., Jalkanen, K.J., and Stephens, P.J. (1987) *Chem. Phys. Lett.*, **133**, 21.

Amos, R.D., Jalkanen, K.J., and Stephens, P.J. (1988) *J. Phys. Chem.*, **92**, 5571.

Atkins, P.W., and Barron, L.D. (1969) *Mol. Phys.*, **16**, 453.

Ayafor, J.F., Tchuedem, M.H.K., Nyasse, B., Tillequin, F., and Anke, H. (1994) *Pure Appl. Chem.*, **66**, 2327.

Baello, B.I., Pancoska, P., and Keiderling, T.A. (1997) *Anal. Biochem.*, **250**, 212.

Bak, K.L., Devlin, F.J., Ashvar, C.S., Taylor, P.R., Frisch, M.J., and Stephens, P.J. (1995) *J. Phys. Chem.*, **99**, 14918.

Bak, K.L., Jørgensen, P., Helgaker, T., Ruud, K., and Jensen, H.J.Aa. (1993) *J. Chem. Phys.*, **98**, 8873.

Bak, K.L., Jørgensen, P., Helgaker, T., Ruud, K., and Jensen, H.J.Aa. (1994) *J. Chem. Phys.*, **100**, 6621.

Barron, L.D. (1982) *Molecular Light Scattering and Optical Activity*, 1st edn, Cambridge University Press.

Barron, L.D. (2004) *Molecular Light Scattering and Optical Activity*, 2nd edn, Cambridge University Press.

Barron, L.D., Bogaard, M.P., and Buckingham, A.D. (1973) *J. Am. Chem. Soc.*, **95**, 603.

Barron, L.D., and Buckingham, A.D. (1971) *Mol. Phys.*, **20**, 1111.

Barron, L.D., and Buckingham, A.D. (2001) *Acc. Chem. Res.*, **34**, 781.

Barron, L.D., Ford, S.J., Bell, A.F., Wilson, G., Hecht, L., and Cooper, A. (1994) *Faraday Discuss.*, **99**, 217.

Barron, L.D., Gargaro, A.R., Hecht, L., and Polavarapu, P.L. (1991a) *Spectrochim. Acta, Part A*, **47**, 1001.

Barron, L.D., Gargaro, A.R., Hecht, L., and Polavarapu, P.L. (1991b) *Spectrochim. Acta, Part A*, **48A**, 261.

Barron, L.D., Hecht, L., Blanch, E.W., and Bell, A.F. (2000) *Prog. Biophys. Mol. Biol.*, **73**, 1.

Bartlett, R.J., Grabowski, I., Hirata, S., and Ivanov, S. (2005) *J. Chem. Phys.*, **122**, 034104.

Bartlett, R.J., Lotrich, V.F., and Schwiegert, I.V. (2005) *J. Chem. Phys.*, **123**, 062205.

Baumruk, V., Pancoska, P., and Keiderling, T.A. (1996) *J. Mol. Biol.*, **259**, 774.

Becke, A.D. (1988a) *Phys. Rev., A*, **38**, 3098.

Becke, A.D. (1988b) *J. Chem. Phys.*, **88**, 1053.

Becke, A.D. (1993a) *J. Chem. Phys.*, **98**, 1372.

Becke, A.D. (1993b) *J. Chem. Phys.*, **98**, 5648.

Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., and Hermans, J. (1981) Interaction models for water in relation to protein hydration, in *Intermolecular Forces*, Reidel, Dordrecht, Holland, pp. 331–42.

Blom, C.E., and Altona, C. (1976) *Mol. Phys.*, **31**, 1377.

Böhm, H.-J., and Klebe, G. (1996) *Angew. Chem. Int.*, **35**, 2588.

Bohr, H., Frimand, K., Jalkanen, K.J., Nieminen, R.M., and Suhai, S. (2001) *Phys. Rev. E*, **64**, 021905.

Bohr, H., and Jalkanen, K. (1998) *Condens. Matter Theor.*, **13**, 95.

Bohr, H.G., and Jalkanen, K.J. (2005) *Condens. Matter Theor.*, **20**, 375.

Bohr, H.G., Jalkanen, K.J., Frimand, K., Elstner, M., and Suhai, S. (1999) *Chem. Phys.*, **246**, 13.

Bohr, H., Jalkanen, K.J., and Malik, F.B. (2005) *Mod. Phys. Lett.*, **19**, 473.

Bombasaro, J.A., Rodriquez, A.M., and Enriz, R.D. (2005) *J. Mol. Struct.: Theochem.*, **724**, 173.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. (1983) *J. Comp. Chem.*, **4**, 187.

Brooks III, C.L., and Case, D.A. (1993) *Chem. Rev.*, **93**, 2487.

Bruno, I.J., Cole, J.C., Lommerse, J.P.M., Rowland, R.S., Taylor, R., and Verdonk, M.L. (1997) *J. Comput. Aided Mol. Des.*, **11**, 525.

Buckingham, A.D. (1967) *Adv. Chem. Phys.*, **12**, 107.

Buckingham, A.D., Fowler, P.W., and Galwas, P.A. (1987) *J. Chem. Phys.*, **112**, 1.

Cao, X., Dukor, R.K., and Nafie, L.A. (2008) *Theor. Chem. Acc.*, **119**, 69.

Cavasotto, C.N., and Abagyan, R.A. (2004) *J. Mol. Biol.*, **337**, 209.

Charles, S.W., and Lee, K.O. (1965) *Trans. Faraday Soc.*, **61**, 2081.

Cheeseman, J.R., and Frisch, M.J. (2008) in *First International Conference on Vibrational Optical Activity, 24*, (eds E. Blanch, and S. Abdali), Manchester Interdisciplinary Biocentre, University of Manchester Press, Manchester.

Cheesman, J.R., Frisch, M.J., Devlin, F.J., and Stephens, P.J. (1996) *Chem. Phys. Lett.*, **252**, 211.

Cho, M. (2008) *Chem. Rev.*, **108**, 1331.

Chou, C.K. (1990) *Clin. Biochem.*, **23**, 37.

Collier, W.B., Magdo, I., and Klots, T.D. (1999) *J. Chem. Phys.*, **110**, 5710.

Colwell, S.M., Murray, C.W., Handy, N.C., and Amos, R.D. (1993) *Chem. Phys. Lett.*, **210**, 261.

Colwell, S.M., and Percival, I.C. (1983a) *Chem. Phys.*, **75**, 215.

Colwell, S.M., and Percival, I.C. (1983b) *Chem. Phys.*, **75**, 225.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz Jr, K.M., Fergusson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. (1995) *J. Am. Chem. Soc.*, **117**, 5179.

Cossi, M., Barone, V., Cammi, R., and Tomasi, J. (1996) *Chem. Phys. Lett.*, **255**, 327.

Cotton, F.A. (1990) *Chemical Applications of Group Theory*, 3rd edn, John Wiley & Sons, New York.

Craw, J.S., Guest, J.M., Cooper, M.D., Burton, N.A., and Hillier, I.H. (1996) *J. Phys. Chem.*, **100**, 6304.

Davis, A.M., and Teague, S.J. (1999) *Angew. Chem. Int. Ed.*, **38**, 736.

Degtyarenko, I.M., Jalkanen, K.J., Gurtovenko, A.A., and Nieminen, R.M. (2007) *J. Phys. Chem. B*, **111**, 4227.

Degtyarenko, I.M., Jalkanen, K.J., Gurtovenko, A.A., and Nieminen, R.M. (2008) *J. Comp. Theor. Nanosci.*, **5**, 277.

Deng, Z., Polavarapu, P.L., Ford, S.J., Hecht, L., Barron, L.D., Ewig, C.S., and Jalkanen, K.J. (1996) *J. Phys. Chem.*, **100**, 2025.

Deplazes, E., van Bronswijk, B., Zhu, F., Barron, L.D., Ma, S., Nafie, L.A., and Jalkanen, K.J. (2008) *Theor. Chem. Acc.*, **119**, 155.

Deslongchamps, P. (1983) *Stereoelectronic Effects in Organic Chemistry*, Volume 1 of Organic Chemistry Series, Pergamon Press, Potts Point.

Dewar, J.S., Zoebisch, E., Healy, E.F., and Stewart, J.J.P. (1985) *J. Am. Chem. Soc.*, **107**, 3902.

Diem, M. (1988) *J. Am. Chem. Soc.*, **110**, 6967.

Diem, M. (1993) *Introduction to Modern Vibrational Spectroscopy*, John Wiley and Sons Ltd, New York.

Diem, M., Polavarapu, P.L., Oboodi, M., and Nafie, L.A. (1982) *J. Am. Chem. Soc.*, **104**, 3329.

Dudek, M.J., and Ponder, J.W. (1995) *J. Comp. Chem.*, **16**, 791.

Elstner, M., Jalkanen, K.J., Knapp-Mohammady, M., Frauenheim, Th., and Suhai, S. (2000) *Chem. Phys.*, **256**, 15.

Elstner, M., Jalkanen, K.J., Knapp-Mohammady, M., Frauenheim, Th., and Suhai, S. (2001) *Chem. Phys.*, **263**, 203.

Elstner, M., Porezag, D., Jungnickel, G., Elsner, J., Haugk, M., Frauenheim, Th., Suhai, S., and Seifert, G. (1998) *Phys. Rev., B*, **58**, 7260.

Ertl, G. (2008) *Angew. Chem. Int. Ed.*, **47**, 2.

Farid, B. (1999) *Electron Correlation in the Solid State, Chapter Ground and Low-lying Excited States of Interacting Electron Systems; A Survey and Some Critical Analyses*, World Scientific, Imperial College, pp. 103–262.

Farid, B. (2002) *Philos. Mag. B*, **82**, 1413.

Fillaux, F. (1999) *J. Mol. Struct.*, **511–512**, 35.

Fillaux, F. (2004) Vibrational spectroscopy and quantum localization, in *Energy Localisation and Transfer*, Volume 22 of Advanced Series in Nonlinear Dynamics, World Scientific, pp. 73–148.

Fillaux, F., Fontaine, J.P., Baron, M.-H., Kearley, G.J., and Tomkinson, J. (1993) *Chem. Phys.*, **176**, 249.

Fillaux, F., Nicolai, B., Baron, M.H., Lautié, A., Tomkinson, J., and Kearley, G.J. (1998) *Ber. Bunsenges. Phys Chem*, **102**, 384.

Fischer, P., and Hache, F. (2005) *Chirality*, **17**, 421.

Floris, F., Persico, M., Tani, A., and Tomasi, J. (1992) *Chem. Phys. Lett.*, **199**, 518.

Fortunelli, A. (1996) *Chem. Phys. Lett.*, **248**, 50.

Fortunelli, A., and Tomasi, J. (1994) *Chem. Phys. Lett.*, **231**, 34.

Freedman, T.B., Chernovitz, A.C., Zuk, W.M., Paterlini, M.G., and Nafie, L.A. (1988) *J. Am. Chem. Soc.*, **110**, 6970.

Frimand, K., and Jalkanen, K.J. (2002) *Chem. Phys.*, **279**, 161.

Frimand, K., Jalkanen, K.J., Bohr, H.G., and Suhai, S. (2000) *Chem. Phys.*, **255**, 165.

Frish, M.J., Yamaguchi, Y., Gaw, J.F., Schaefer, H.F., and Binkley III, J.S. (1986) *J. Chem. Phys.*, **84**, 531.

Gale, J.D., and Rohl, A.L. (2007) *Mol. Sim.*, **33**, 1237.

Gambetta, D. (ed.) (2005) *Making Sense of Suicide Missions*, Oxford University Press.

Gay, D.H., and Rohl, A.L. (1995) *J. Chem. Soc., Faraday Trans.*, **91**, 925.

Godby, R.W., Schlüter, M., and Sham, L.J. (1988) *Phys. Rev., B*, **37**, 10159.

Goodsell, D.S. (2002) The future of bionanotechnology, in *Biotechnology: Lessons from Nature*, John Wiley & Sons, Inc., Berlin, pp. 295–311.

Gutt, C., Baumert, J., Press, W., Tse, J.S., and Janssen, J.S. (2002) *J. Chem. Phys.*, **116**, 3795.

Halls, M.D., and Schlegel, H.B. (1999a) *J. Chem. Phys.*, **111**, 8819.

Halls, M.D., and Schlegel, H.B. (1999b) *J. Chem. Phys.*, **109**, 10587.

Hammerschmidt, E.G. (1934) *Ind. Eng. Chem.*, **26**, 851.

Han, W.-G., Jalkanen, K.J., Elstner, M., and Suhai, S. (1998) *J. Phys. Chem. B*, **102**, 2587.

Han, W.-G., Jalkanen, K.J., Elstner, M., Frauenheim, Th., and Suhai, S. (2000) *Int. J. Quantum Chem.*, **78**, 459.

Handa, Y.P. (1986a) *J. Chem. Thermodyn.*, **18**, 891.

Handa, Y.P. (1986b) *J. Chem. Thermodyn.*, **18**, 915.

Hanessian, S. (1983) *Total Synthesis of Natural Products: The 'Chiron' Approach*, Volume 3 of Organic Chemistry Series, Pergamon Press, Potts Point.

Hao, C., Whittaker, L., and Whittaker, J. (2006) *Biochem. Biophys. Res. Commun.*, **347**, 334.

Hecht, L., and Barron, L.D. (1990) *Appl. Spectrosc.*, **44**, 483.

Hecht, L., Barron, L.D., and Hug, W. (1989) *Chem. Phys. Lett.*, **158**, 341.

Helgaker, T., Ruud, K., Bak, K.L., Jørgensen, P., and Olsen, J. (1994) *Faraday Discuss.*, **99**, 165.

Hohenberg, P., and Kohn, W. (1964) *Phys. Rev.*, **136**, 864.

Horikawa, S., Itoh, H., Tabata, J., Kawamura, K., and Hondoh, T. (1997) *J. Phys. Chem. B*, **101**, 6290.

Hornicek, J., Kapralova, P., and Bour, P. (2008) *J. Chem. Phys.*, **127**, 084502.

Inoue, R., Tanaka, H., and Nakanishi, K. (1996) *J. Chem. Phys.*, **104**, 9569.

Ioannnou, A.G., Colwell, S.M., and Amos, R.D. (1997) *Chem. Phys. Lett.*, **278**, 278.

Ipolyi, I., Cicman, P., Denifl, S., Mtejcik, V., Mach, P., Urban, J., Scheier, P., Mark, T.D., and Matejcik, S. (2006) *Int. J. Mass Spectrom.*, **252**, 228.

Ismail-Beigi, S., and Louie, S.G. (2003) *Phys. Rev. Lett.*, **90**, 076401.

Itoh, H., Chazallon, B., Schober, H., Kawamura, K., and Kuhs, W.F. (2003) *Can. J. Phys.*, **81**, 493.

Jalkanen, K.J., Bohr, H.G., Suhai, S. (1997) in *Proceedings of the International Symposium on Theoretical and Computational Methods in Genome Research*, (S. Suhai), Plenum Press, New York, pp. 255–77.

Jalkanen, K.J., Degtyarenko, I.M., Nieminen, R.M., Cao, X., Nafie, L.A., Zhu, F., and Barron, L.D. (2008a) *Theor. Chem. Acc.*, **119**, 191.

Jalkanen, K.J., Gale, J.D., Jalkanen, G.J., McIntosh, D., El Azhary, A., and Jensen, G.M. (2008b) *Theor. Chem. Acc.*, **119**, 211.

Jalkanen, K.J., Gale, J.D., Lassen, P.R., Hemmingsen, L., Rodarte, A., Degtyarenko, I.M., Nieminen, R.M., Christensen, S.B., Knapp-Mohammady, M., and Suhai, S. (2008c) *Theor. Chem. Acc.*, **119**, 177.

Jalkanen, K.J., Gale, J.D., Rohl, A.L., Zhu, F., Barron, L.D., Ma, S., and Nafie, L.A. (2009a) *Theor. Chem. Acc.* (Suhai Festschrift) manuscript in preparation.

Jalkanen, K.J., Jürgensen, V.W., Claussen, A., Jensen, G.M., Rahim, A., Wade, R.C., Nardi, F., Jung, C., Nieminen, R.M., Degtyarenko, I.M., Herrmann, F., Knapp-Mohammady, M., Niehaus, T., Frimand, K., and Suhai, S. (2006a) *Int. J. Quantum Chem.*, **106**, 1160.

Jalkanen, K.J., Jürgensen, V.W., and Degtyarenko, I.M. (2006b) *Adv. Quantum Chem.*, **50**, 91.

Jalkanen, K.J., Knapp-Mohammady, M., Suhai, S., Nardi, F., Wade, R.C., Bohr, J., Bohr, H.G., Rohl, A.L., and Gale, J.D. (2009b) *Theor. Chem. Acc.* (Suhai Festschrift) manuscript in preparation.

Jalkanen, K.J., Nardi, F., Wade, R.C., Suhai, S., Rohl, A.L., Gale, J.D., and Jung, C. (2009c) *Theor. Chem. Acc.* (Suhai Festschrift) manuscript in preparation.

Jalkanen, K.J., Nieminen, R.M., Frimand, K., Bohr, J., Bohr, H., Wade, R.C., Tajkhorshid, E., and Suhai, S. (2001) *Chem. Phys.*, **265**, 125.

Jalkanen, K.J., Nieminen, R.M., Knapp-Mohammady, M., and Suhai, S. (2003) *Int. J. Quantum Chem.*, **92**, 239.

Jalkanen, K.J., and Stephens, P.J. (1991) *J. Phys. Chem.*, **95**, 5446.

Jalkanen, K.J., Stephens, P.J., Amos, R.D., and Handy, N.C. (1987) *J. Am. Chem. Soc.*, **109**, 7193.

Jalkanen, K., Stephens, P.J., Amos, R., and Handy, N. (1988) *J. Am. Chem. Soc.*, **110**, 2012.

Jalkanen, K.J., and Suhai, S. (1996) *Chem. Phys.*, **208**, 81.

Johnson, B.G., and Florian, J. (1995) *Chem. Phys. Lett.*, **247**, 120.

Johnson, B.G., and Frisch, M. (1994) *J. Chem. Phys.*, **100**, 7429.

Jorgensen, W.L., Madura, J.D., and Swenson, C.J. (1984) *J. Am. Chem. Soc.*, **106**, 6638.

Jorgensen, W.L., Maxwell, D.S., and Tirado-Rives, J. (1996) *J. Am. Chem. Soc.*, **118**, 11225.

Jorgensen, W.L., and Swenson, C.J. (1985a) *J. Am. Chem. Soc.*, **107**, 569.

Jorgensen, W.L., and Swenson, C.J. (1985b) *J. Am. Chem. Soc.*, **107**, 1489.

Jorgensen, W.L., and Tirado-Rives, J. (1988) *J. Am. Chem. Soc.*, **110**, 1657.

Jung, C. (2000) *J. Mol. Recognit.*, **13**, 325.

Jürgensen, V.W., and Jalkanen, K.J. (2006) *Phys. Biol.*, **3**, S63.

Kawiecki, R.W., Devlin, F.J., Stephens, P.J., and Amos, R.D. (1991) *J. Phys. Chem.*, **95**, 9817.

Kearley, G.J. (1986) *J. Chem. Soc., Faraday Trans. 2*, **82**, 41.

Kearley, G.J. (1995) *Nucl. Instrum. Methods Phys. Res. A*, **354**, 53.

Kearley, G.J., Fillaux, F., Baron, M.H., Bennington, S., and Tomkinson, J. (1994) *Science*, **264**, 1285.

Kearley, G.J., Johnson, M.R., Plazanet, M., and Suard, E. (2001) *J. Chem. Phys.*, **115**, 2614.

Kearley, G.J., Tomkinson, J., and Penfold, J. (1987) *Z. Phys. B*, **69**, 63.

Keiderling, T.A. (1996) Protein structural studies using vibrational circular dichroism spectroscopy, in *Spectroscopic Methods For Determining Protein Structure in Solution*, Chapter 8,VCH Publishers, Inc., Weinheim, pp. 163–89.

Keskin, O., Gursoy, A., Ma, B., and Nussinov, R. (2008) *Chem. Rev.*, **108**, 1225.

Kim-Cohen, J., Caspi, A., Taylor, A., Williams, B., Newcombe, R., Craig, I.W., and Moffitt, T.E. (2006) *Mol. Psychiatry*, **11**, 903.

Klopman, G. (1964a) *J. Am. Chem. Soc.*, **86**, 1463.

Klopman, G. (1964b) *J. Am. Chem. Soc.*, **86**, 4550.

Knapp-Mohammady, M., Jalkanen, K.J., Nardi, F., Wade, R.C., and Suhai, S. (1999) *Chem. Phys.*, **240**, 63.

Kohn, W. (1999) *Rev. Mod. Phys.*, **71**, 1253.

Kohn, W., and Sham, L.J. (1965) *Phys. Rev.*, **140**, 1133.

Kollman, P.A. (1992) *Curr. Opin. Struct. Biol.*, **2**, 765.

Komornicki, A., and Fitzgerald, G. (1993) *J. Chem. Phys.*, **98**, 1398.

Komornicki, A., and McIver, J.W. (1979) *J. Chem. Phys.*, **70**, 2014.

Kongsted, J., and Ruud, K. (2008) *Chem. Phys. Lett.*, **451**, 226.

Kubelka, J., Pancoska, P., and Keiderling, T.A. (1999) *Appl. Spectrosc.*, **53**, 666.

Kulcitki, V., Ungur, N., Gavagnin, M., Carbone, M., and Cimino, G. (2005) *Eur. J. Org. Chem.*, **2005** (9), 1816–22.

Lannoo, M., Schlüter, M., and Sham, L.J. (1985) *Phys. Rev., B*, **32**, 3890.

Lee, A.M., Colwell, S.M., and Handy, N.C. (1994) *Chem. Phys. Lett.*, **229**, 225.

Lee, C., and Sosa, C. (1994) *J. Chem. Phys.*, **100**, 9018.

Lee, C., Yang, W., and Parr, R.G. (1988) *Phys. Rev., B*, **37**, 785.

Levine, I.N. (2000) *Quantum Chemistry*, 5th edn, Prentice Hall, Upper Saddle River.

Levine, Z.H., and Allan, D.C. (1989) *Phys. Rev. Lett.*, **63**, 1719.

Liegeois, V. (2008) Ph.D. thesis Elaboration of Quantum Chemistry Methods for Predicting and Interpreting Vibrational Raman Optical Activity : Applications to Helical Structures, Facultes Universitaires Notre-Dame de la Paix, Namur.

Liegeois, V., and Champagne, B. (2008) in *First International Conference on Vibrational Optical Activity, 25*, (eds E. Blanch, and S. Abdali), Manchester Interdisciplinary Biocentre, University of Manchester Press, Manchester.

Lord, R.C., and Yu, N.-T. (1970a) *J. Mol. Biol.*, **50**, 509.

Lord, R.C., and Yu, N.-T. (1970b) *J. Mol. Biol.*, **51**, 203.

Losada, M., Tran, H., and Xu, Y. (2008) *J. Chem. Phys.*, **128**, 014508.

Losada, M., and Xu, Y. (2007) *Phys. Chem. Chem. Phys.*, **9**, 3127.

Lowe, M.A., Alper, J.S., Kawiecki, R.W., and Stephens, P.J. (1986) *J. Phys. Chem.*, **90**, 41.

Ma, S.-K., and Brueckner, K.A. (1968) *Phys. Rev.*, **165**, 18.

MacKerell, A.D.J., Bashford, D., Bellott, M., Dunbrack, J.R.L., Evanseck, J.D., Field, M.J., Fisher, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kucyera, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher III, W.E., Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998) *J. Phys. Chem. B*, **102**, 3586.

Malkin, V.G., Malkina, O.L., Eriksson, L.A., and Salahub, D.R. (1995) The calculation of NMR and ESR spectroscopy parameters using density functional theory, *Modern Density Functional Theory: A Tool for Chemistry*, Volume 2 of Theoretical and Computational Chemistry, Elsevier Science B.V., pp. 273–347.

Malkin, V.G., Malkina, O.L., and Salahub, D.R. (1993a) *Chem. Phys. Lett.*, **204**, 80.

Malkin, V.G., Malkina, O.L., and Salahub, D.R. (1993b) *Chem. Phys. Lett.*, **204**, 87.

Mantz, Y.A., Gerard, H., Iftimie, R., and Martyna, G.J. (2006) *J. Phys. Chem. B*, **110**, 13523.

Maple, J.R., Hwang, M.-J., Jalkanen, K.J., Stockfisch, T.P., and Hagler, A.T. (1998) *J. Comp. Chem.*, **19**, 430.

Matsuda, S.P.T., Wilson, W.K., and Xiong, Q. (2006) *Org. Biomol. Chem.*, **4**, 530.

Mermin, N. (1965) *Phys. Rev.*, **137**, A1441.

Mok, D.K., Nwumann, R., and Handy, N.C. (1996) *J. Phys. Chem.*, **100**, 6225.

Moon, C., Taylor, P.C., and Rodger, P.M. (2003) *Can. J. Chem.*, **81**, 451.

Mukamel, S. (1995) *Principles of Nonlinear Optical Spectroscopy*, Oxford University Press, Oxford.

Mukhopadhyay, P., Zuber, G., Wipf, P., and Beratan, D.N. (2007) *Angew. Chem. Int. Ed.*, **46**, 6450.

Nafie, L.A. (2008) *Theor. Chem. Acc.*, **119**, 39.

Nardi, F. (1999) Ph.D. thesis Molecular Dynamics and NMR Spectroscopic Studies of Local Interactions of Small Peptides in Aqueous Solution Involving Aromatic Residues, Universite Joseph Fourier, Grenoble.

Nardi, F., Kemmink, J., Sattler, M., and Wade, R.C. (2000) *J. Biomol. NMR*, **17**, 63.

Nardi, F., Worth, G.A., and Wade, R.C. (1997) *Fold. Des.*, **2**, S62.

Newman, A.A. (ed.) (1972) *Chemistry of Terpenes and Terpenoids*, Academic Press Inc., London.

Nicu, V.P., Neugebauer, J., Wolff, S.K., and Baerends, E.J. (2008) *Theor. Chem. Acc.*, **119**, 245.

Oliveira, F.G., Sant'Anna, C.M.R., Caffarena, E.R., Dardenne, L.E., and Barreiro, E.J. (2006) *Bioorg. Med. Chem.*, **14**, 6001.

Onida, G., Reining, L., and Rubio, A. (2002) *Rev. Mod. Phys.*, **74**, 601.

Onsager, L. (1936) *J. Am. Chem. Soc.*, **58**, 1486.

Ösapay, K., Young, W.S., Bradford, D., Brooks, C.L., and Case, D.A. (1996) *J. Phys. Chem.*, **100**, 2698.

Oyama, H., Ebinuma, T., Shimada, W., Takeya, S., Nagao, J., Uchida, T., and Narita, H. (2003) *Can. J. Phys.*, **81**, 485.

Pancoska, P., Bitto, E., Janota, V., and Keiderling, T.A. (1994) *Faraday Discuss. Chem. Soc.*, **99**, 287.

Pancoska, P., Bitto, E., Janota, V., Urbanova, M., Gupta, V.P., and Keiderling, T.A. (1995) *Protein Sci.*, **4**, 1384.

Pancoska, P., Blazek, M., and Keiderling, T.A. (1992) *Biochemistry*, **31**, 10250.

Pancoska, P., Janota, V., and Keiderling, T.A. (1996) *Appl. Spectrosc.*, **50**, 658.

Pancoska, P., Kubelka, J., and Keiderling, T.A. (1999) *Appl. Spectrosc.*, **53**, 655.

Pancoska, P., Yasui, S.C., and Keiderling, T.A. (1989) *Biochemistry*, **28**, 5917.

Pancoska, P., Yasui, S.C., and Keiderling, T.A. (1991) *Biochemistry*, **30**, 5089.

Papai, I., St-Alant, A., Ushio, J., and Salahub, D. (1990) *Int. J. Quantum Chem. Quantum Chem. Symp.*, **24**, 29.

Perdew, J.P., Burke, K., and Ernyerhoh, M. (1996) *Phys. Rev. Lett.*, **77**, 3865.

Perdew, J.P., Burke, K., and Ernyerhoh, M. (1997) *Phys. Rev. Lett.*, **78**, 1396.

Perdew, J.P., and Levy, M. (1983) *Phys. Rev. Lett.*, **51**, 1884.

Perdew, J.P., and Wang, Y. (1992) *Phys. Rev., B*, **45**, 13244.

Pimentel, G.C. (1975) *Angew. Chem. Int.*, **14**, 199.

Pimentel, G.C., and Charles, S.W. (1963) *Pure Appl. Chem.*, **7**, 111.

Pimentel, G.C., and Sederholm, C.H. (1956) *J. Chem. Phys.*, **24**, 639.

Pohl, G., Perczel, A., Vass, E., Magyarfalvi, G., and Tarczay, G. (2007) *Phys. Chem. Chem. Phys.*, **9**, 4698.

Polavarapu, P.L. (1990) *J. Phys. Chem.*, **94**, 8106.

Poon, C.-D., Samulski, E.T., Weise, C.F., and Weisshaar, J.C. (2000) *J. Am. Chem. Soc.*, **122**, 5642.

Pople, J.A. (1999) *Rev. Mod. Phys.*, **71**, 1267.

Pulay, P., Fogarasi, G., Pang, F., and Boggs, J.E. (1979) *J. Am. Chem. Soc.*, **101**, 2550.

Rajagopal, A.K., and Callaway, J. (1973) *Phys. Rev., B*, **7**, 1912.

Ramnarayan, K., Bohr, H.G., and Jalkanen, K.J. (2008) *Theor. Chem. Acc.*, **119**, 265.

Rauhut, G., Puyear, S., Wolinski, K., and Pulay, P. (1996) *J. Phys. Chem.*, **100**, 6310.

Reining, L., Olevano, V., Rubio, A., and Onida, G. (2002) *Phys. Rev. Lett.*, **88**, 066404.

Richmond, T.J. (1984) *J. Mol. Biol.*, **178**, 63.

Ruud, K. (2008) in *First International Conference on Vibrational Optical Activity, 28*, (eds E. Blanch, and S. Abdali), Manchester Interdisciplinary Biocentre, University of Manchester Press, Manchester.

Sauter, C., Otalora, F., Gavira, J.-A., Vidal, O., Giege, R., and Garcia-Ruiz, J.M. (2001) *Acta Crystallogr. D*, **57**, 1119.

Schiffer, C.A., Caldwell, J.W., Stroud, R.M., and Kollman, P.A. (2992) *Protein Sci.*, **1**, 396.

Schmidt, A.B., and Fine, R.M. (1994) *Mol. Sim.*, **13**, 347.

Schreckenback, G., and Ziegler, T. (1995) *J. Phys. Chem.*, **99**, 606.

Scott, A. (2003) *Nonlinear Science*, 2nd edn, Oxford University Press, Oxford.

Serdyuk, I.N., Zaccai, N.R., and Zaccai, J. (2007) *Methods in Molecular Biophysics: Structure, Dynamics, Function*, Cambridge University Press, Cambridge.

Sham, L.J., and Schlüter, M. (1985) *Phys. Rev., B*, **32**, 3883.

Stephens, P.J. (1985) *J. Phys. Chem.*, **89**, 748.

Stephens, P.J. (1987) *J. Phys. Chem.*, **91**, 1712.

Stephens, P.J., Chabalowski, C., Devlin, F., and Jalkanen, K.J. (1993) *Chem. Phys. Lett.*, **225**, 247.

Stephens, P.J., Devlin, F.J., Ashvar, C.S., Chabalowski, C.F., and Frisch, M.J. (1994) *Faraday Discuss.*, **99**, 103.

Stephens, P.J., Devlin, F.J., Chabalowski, C., and Frisch, M.J. (1994) *J. Phys. Chem.*, **98**, 11623.

Stephens, P.J., Jalkanen, K.J., Amos, R.D., Lazzeretti, P., and Zanasi, R. (1990) *J. Phys. Chem.*, **94**, 1811.

Stephens, P.J., and Lowe, M.A. (1985) *Ann. Rev. Phys. Chem.*, **36**, 213.

Stewart, J.J.P. (1989) *J. Comp. Chem.*, **10**, 209.

Stewart, J.J.P. (2007) *J. Mol. Model.*, **13**, 1173.

Stirling, A. (1996) *J. Chem. Phys.*, **104**, 1254.

Subramanian, S., and Sloan Jr, E.D. (2002) *J. Phys. Chem. B*, **106**, 4348.

Tajkhorshid, E., Jalkanen, K.J., and Suhai, S. (1998) *J. Phys. Chem. B*, **102**, 5899.

Tawa, G.J., Martin, R.L., Pratt, L.R., and Russo, T.V. (1996) *J. Phys. Chem.*, **100**, 1515.

te Velde, G., Bickelhaupt, F.M., Baerends, E.J., Guerra, C.F., van Gisbergen, S.J.A., Snijders, J.G., and Ziegler, T. (2001) *J. Comp. Chem.*, **22**, 931.

Thar, J., Zhan, S., and Kirchner, B. (2008) *J. Phys. Chem. B*, **112**, 1456.

Tomasi, J., Mennucci, B., and Cammi, R. (2005) *Chem. Rev.*, **105**, 2999.

Tomasi, J., and Persico, M. (1994) *Chem. Rev.*, **94**, 2027.

Tse, J.S., Shpakov, V.P., Belosludov, V.R., Trouw, F., Handa, Y.P., and Press, W. (2001) *Europhys. Lett.*, **54**, 354.

van Caillie, C., and Amos, R.D. (2000) *Phys. Chem. Chem. Phys.*, **2**, 2123.

van Gunsteren, W.F., Berendsen, H.J.C., Hermans, J., Hol, W.G.J., and Postma, J.P.M. (1983) *Proc. Natl. Acad. Sci.*, **80**, 4315.

van Klaveren, E.P., Michels, J.P.J., Schouten, J.A., Klug, D.D., and Tse, J.S. (2001) *J. Chem. Phys.*, **114**, 5745.

Vatamanu, J., and Kusalik, P.G. (2008) *J. Phys. Chem. B*, **112**, 2399.

Vignale, G., and Rasolt, M. (1987) *Phys. Rev. Lett.*, **59**, 2360.

Vignale, G., and Rasolt, M. (1988) *Phys. Rev., B*, **37**, 10685.

Vignale, G., Rasolt, M., and Geldart, D.J.W. (1988) *Phys. Rev., B*, **37**, 2502.

Warshel, A., and Levitt, M. (1976) *J. Mol. Biol.*, **103**, 227.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S., and Weiner Jr, P. (1984) *J. Am. Chem. Soc.*, **106**, 765.

Weiner, S.J., Kollman, P.A., Nguyen, D.T., and Case, D.A. (1986) *J. Comp. Chem.*, **7**, 230.

Weise, C.F., and Weisshaar, J.C. (2003) *J. Phys. Chem. B*, **107**, 3265.

Wesson, L., and Eisenberg, D. (1992) *Protein Sci.*, **1**, 227.

Weyerstahl, P., Schwieger, R., Schwope, I., and Hashem, Md.A. (1995) *Liebigs Ann.*, (7), 1389–92.

Whittle, E., Dows, D.A., and Pimentel, G.C. (1955) *J. Chem. Phys.*, **22**, 1943.

Wilson, P.J., Amos, R.D., and Handy, N.C. (1999) *Chem. Phys. Lett.*, **312**, 475.

Wolff, S.K. (2005) *Int. J. Quantum Chem.*, **104**, 645.

Yu, G.-S., Freedman, T.B., Nafie, L.A., Deng, T., and Polavarapu, P.L. (1995) *J. Phys. Chem.*, **99**, 835.

Zeng, H., Wilson, L.D., Walker, V.K., and Ripmeester, J.A. (2003) *Can. J. Phys.*, **81**, 17.

Zhu, F., Isaacs, N.W., Hecht, L., and Barron, L.D. (2005) *Structure*, **13**, 1409.

Zhukov, V.P., Chulkov, E.V., and Echenique, P.M. (2004) *Phys. Rev. Lett.*, **93**, 096401.

Color Plates

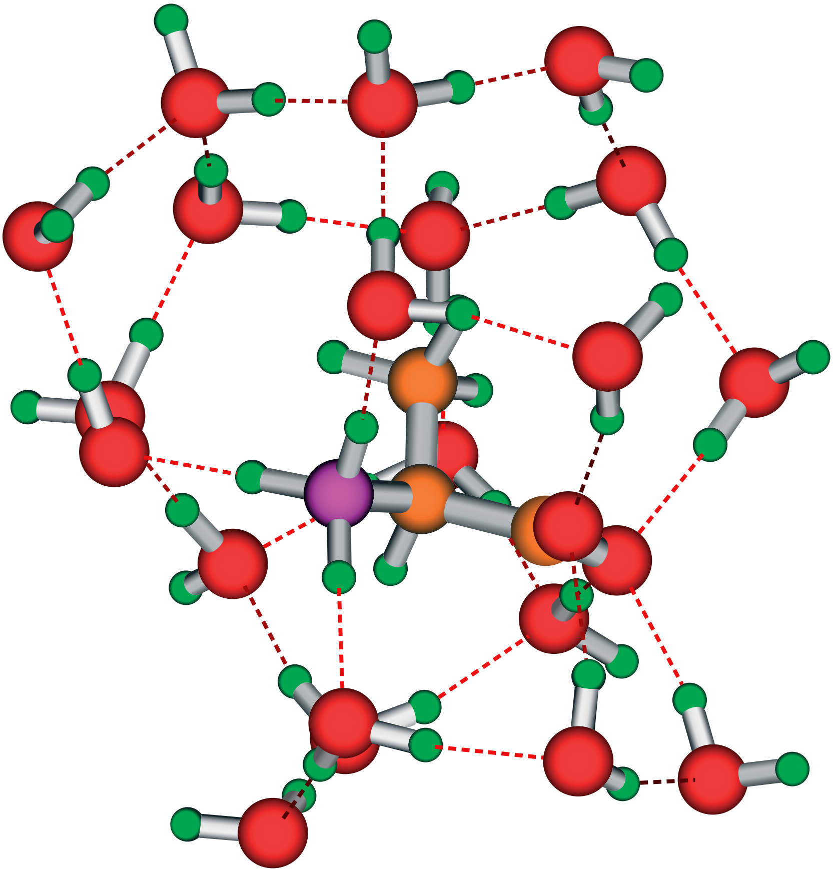

Fig. 1.1 L-Alanine zwitterion stabilized by 20 water molecules. B3PW91/6-31G* plus CPCM continuum solvent model optimized geometry. (This figure also appears on page 38.)

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

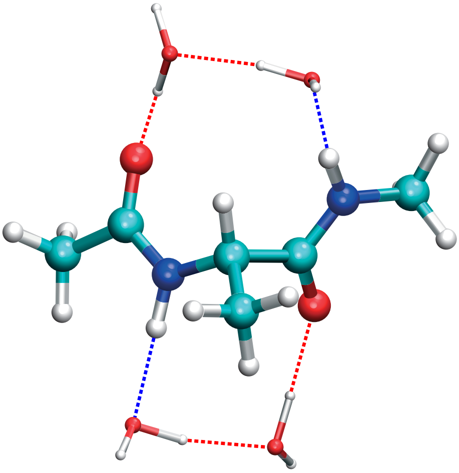

Fig. 1.2 *N*-Acetyl L-alanine *N*′-methylamide stabilized by four water molecules. B3PW91/6-31G* plus CPCM continuum solvent model optimized geometry. (This figure also appears on page 42.)

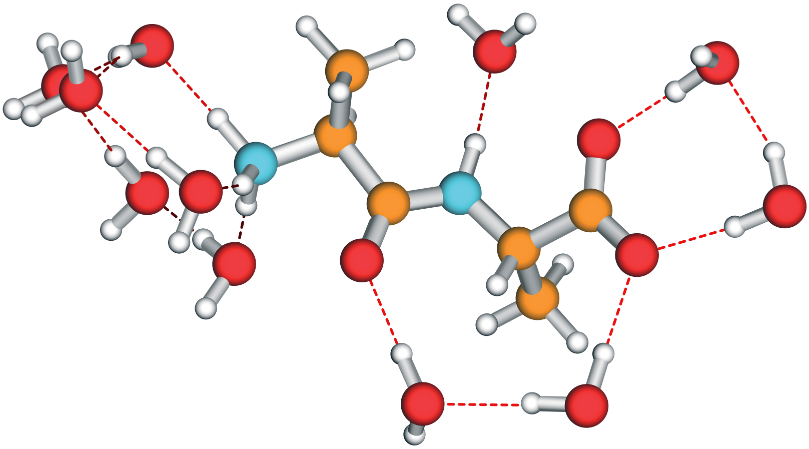

Fig. 1.3 L-alanyl-L-alanine zwitterion stabilized by 11 water molecules. B3PW91/6-311G2dp plus PCM continuum solvent model optimized geometry. (This figure also appears on page 46.)

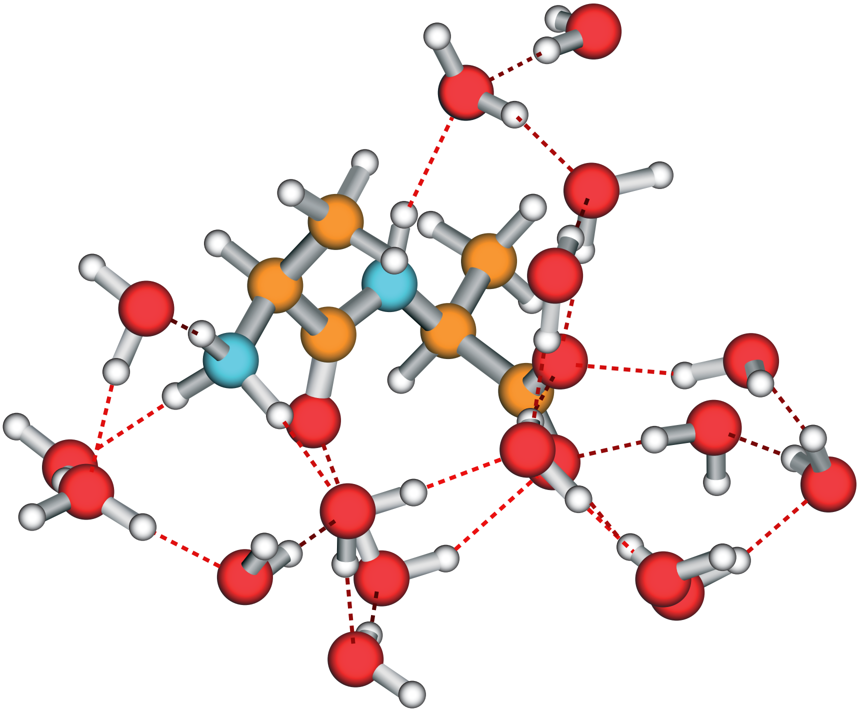

Fig. 1.6 L-Alanyl-L-alanine zwitterion stabilized by 17 water molecules. B3PW91/6-31G* plus PCM continuum solvent model optimized geometry. (This figure also appears on page 48.)

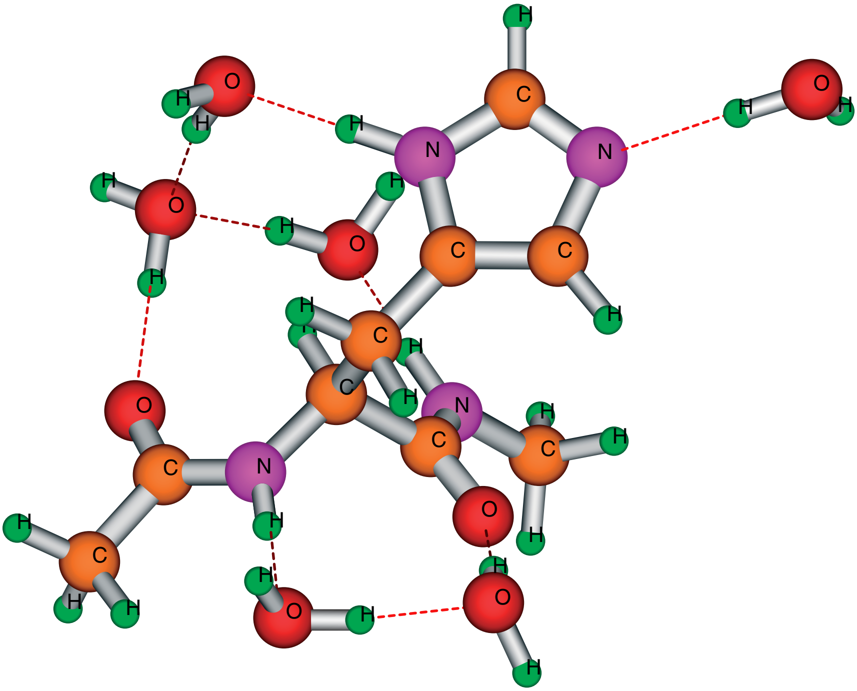
H
C
O
O
H
H
H
H
H
N
N
H
H
O
H
O
C
C
H
H
H
C
H
C
H
N
O
H
H
C
C
C
N
O
H
H
H
C
H
H
O
H
H
O
H
H

Fig. 1.7 *N*-Acetyl L-histidine *N*′-methylamide stabilized by six water molecules. B3PW91/6-31G* plus PCM continuum solvent model optimized geometry. (This figure also appears on page 49.)

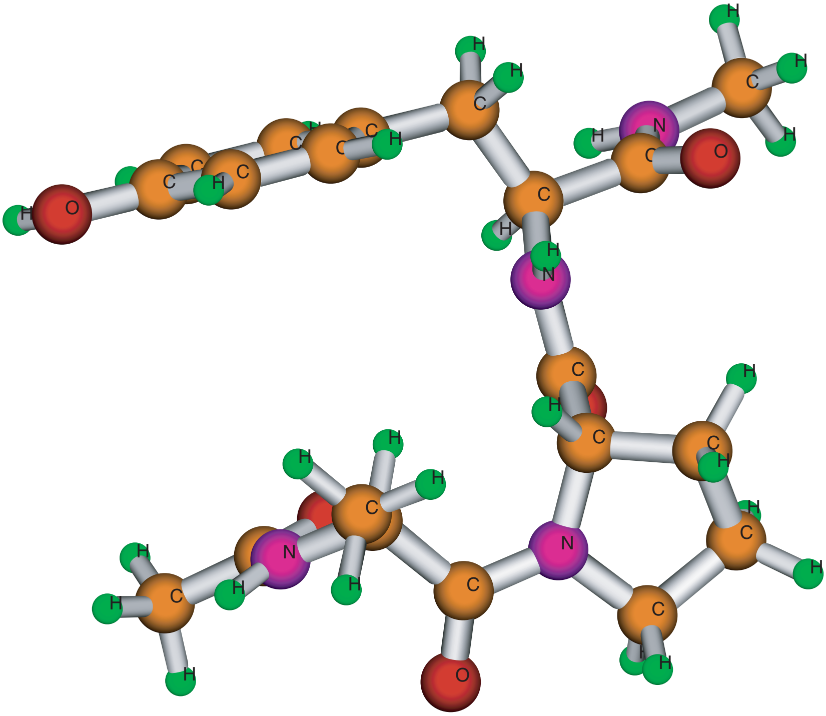
H
O
C
C
H
C
C
C
C
H
C
H
H
C
H
H
C
N
C
O
H
H
N
H
H
C
H
H
C
H
C
H
H
C
H
H
N
C
H
H
N
C
H
H
C
O
C
H
H

Fig. 1.8 *N*-Acetyl L-alanine L-proline L-tyrosine *N*′-methylamide molecule. B3PW91/6-31G** plus PCM continuum solvent model optimized geometry. (This figure also appears on page 55.)

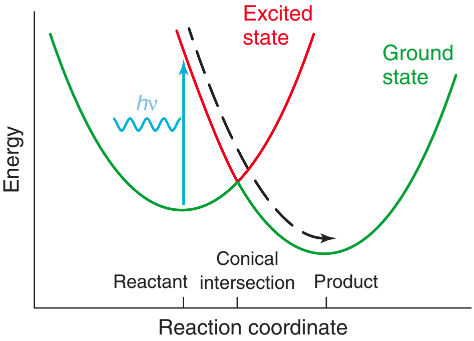
Excited state
Ground state
hν
Energy
Reactant
Conical intersection
Product
Reaction coordinate

Fig. 3.1 One-dimensional scheme of a photochemical reaction. After photon absorption (cyan arrow), evolution takes place on the excited state potential energy surface (red), until the system hits the conical intersection. At the intersection, a radiationless transition to the ground state occurs (green). After the decay, the system continues evolving in the ground state. (This figure also appears on page 96.)

2

Radiationless Auger-Like Electronic Transitions in Biomolecules: Bio-Auger Process

Henrik G. Bohr and F. Bary Malik

2.1 Introduction 69
2.2 Auger Process in Atoms 72
2.2.1 Nonrelativistic Expression for Auger Processes in Atoms and Ions 73
2.2.2 Expression for Radiative Transition 77
2.2.3 Bethe–Heitler Approximation for Auger Transition 77
2.2.4 Key Distinctions between Radiative and Auger Transitions 78
2.3 Distinctive Features of the Auger Processes in Atoms and those in Biomolecules 78
2.3.1 Distinctive Features of the Auger and the Bio-Auger Processes 79
2.3.2 Bio-Auger Process and Internal Conversion 81
2.4 Single Orbital Theory of Bio-Auger Process 82
2.4.1 Wentzel's Analogous Theory in Biomolecules 82
2.4.2 Theory of Nonradiative Decay in Mean Field Approximation in Biomolecules 83
2.5 Bio-Auger Process and Charge or Electron Transfer 85
2.6 Experimental Evidence 85
2.6.1 Radiationless De-Excitation from Singlet and Triplet States 86
2.6.2 Radiationless Transitions and Electron Transfer in a Protein Complex in Photosynthesis 87
2.6.3 Dark Structure 89
2.6.4 Exposures to Ultraviolet Radiation and Relaxation 89
2.6.5 Bio-Auger Process Involving Intermolecular Electronic States 89
Acknowledgments 90
References 91
Further Reading 92

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

2.1
Introduction

Living organisms on the earth are constantly being exposed to UV and other kinds of cosmic and georadiations that, in many instances, are capable of inducing electronic transitions in the atomic and molecular constituents of biomolecules, amino acids, or proteins. In most cases, these biomolecules return to their original states after being excited. The study of the mechanism of restoration of bio-organisms to their original states after being exposed to radiation capable of causing electronic transitions is still in its infancy, and in this chapter we present the current status of the nonradiative process caused by the interaction between two electrons, which is associated with the de-excitation. This process is likely to play a very important role in restoring them to their original states after being exposed to radiation.

It is now well recognized that biomolecules being exposed to electromagnetic radiation (EMR) or to collision by other projectiles such as electrons or protons of sufficient energy can undergo electronic transition from ground states to some excited states. For incident EMR, the wavelength, λ, required for this electronic excitation, is related to its energy, E, by the following relation:

$$E(\text{in eV}) = \frac{(1.237)10^3\ \text{eV nm}}{\lambda(\text{in nm})} \quad (1)$$

The most predominant excitation is caused by the dipole component of EMR and hence, the excited state must have the same electronic spin (but not total angular momentum) as the ground state, which in most of the biomolecules is a singlet state denoted in Figure 2.1 as S_0. Following the resonant absorption, the electron from S_0 is transferred to an excited singlet state, S_{ex}, commensurate with the energy given by Eq. (1), even though there are quite often triplet states at lower or nearby energy. Once the system is in the S_{ex} state, it may not primarily decay back immediately to S_0 because of the opening of many other decay channels. Frequently, intraband vibrational decay probabilities emitting infrared radiation are much faster than the electronic dipole radiation $S_{\text{ex}} \rightarrow S_0$, and hence, the system undergoes some vibronic decays, but then it could move from its original potential well to another one either by internal conversion (IC), caused by electronic–vibrational–rotational coupling, or to another potential well based on the triplet electronic state, T_{ex}, known as

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

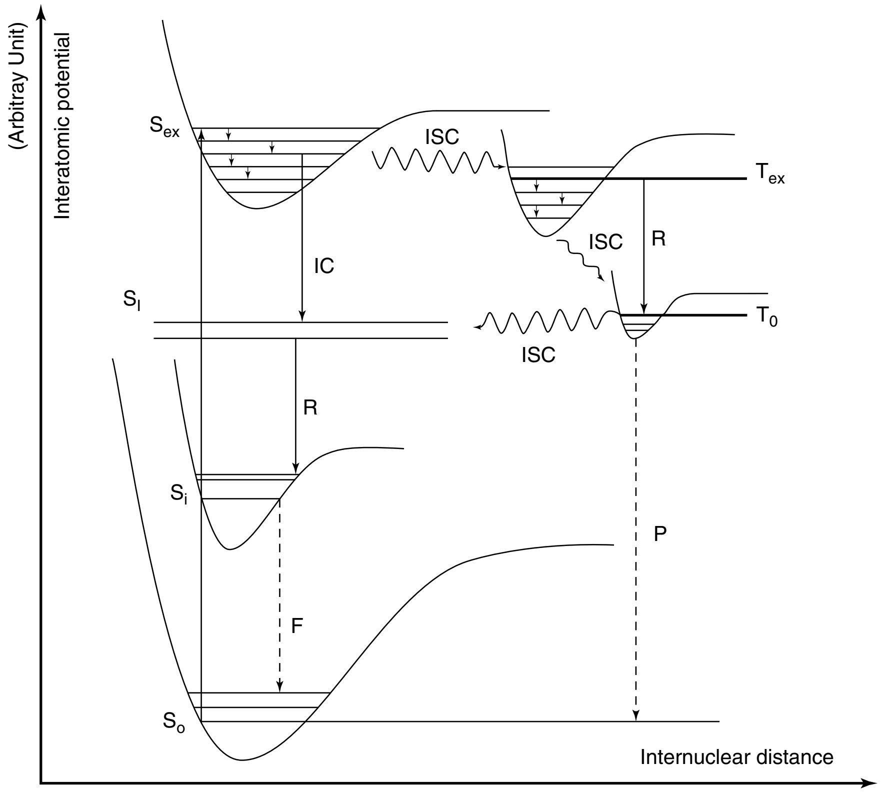

Fig. 2.1 Schematic pathway of radiative excitation and radiative and nonradiative relaxation in a biomolecule. S_0, S_l, S_ℓ, and S_{ex} refer to singlet states, and T_{ex} and T_0 to triplet states. ISC, R, IC, F, and P refer, respectively, to intersystem crossing, radiative transition, internal conversion, fluorescence and phosphorescence. The latter two are also radiative processes. The pathway for the Bio-Auger process is not shown.

intersystem crossing (ISC). The relaxation mechanism to the ground state, S_0, may occur in a variety of ways. For example, following IC, a system could decay to S_i, from which it could, via radiative dipole transition, return to S_0. This delayed radiative transition is termed *fluorescence*. Any one of the triplet states could bring an ISC back to a low-lying singlet state, S_l, which then decays to the ground state by fluorescence again. The relaxation pathway via triplet states to ground state could, however, happen directly by radiative transition of multipole of an order higher than dipole, known in atomic physics as forbidden transitions. In particular, the multipoles must be those capable of causing a triplet–singlet transition. These have long lifetimes and are termed *phosphorescence*. All these relaxation pathways to ground states have been experimentally observed in many polyatomic, aromatic, and nonaromatic organic molecules (McGlynn, Azumi, and Kinoshita, 1969; Lin, 1980; Schwoerer and Wolf, 2005; DiBartolo, 1980; Karplus and Porter, 1970). However, in many cases the total fluorescence and phosphorescence yields, collectively termed as *luminescence yields*, do not account for the magnitude of initial resonant absorption, pointing toward the existence of other decay modes, termed as *nonradiative decay mode*, or radiationless transition.

The study to understand the underlying causes of radiationless transitions started almost half a century ago (Henry and Kasha, 1968). Their occurrence in the relaxation process of biomolecules, particularly for many amino acids, were well established soon thereafter (Maki and Zuclich, 1975). It is now well recognized that many amino acids and proteins, following exposure to laser, (i) populate the triplet states substantially after undergoing ISC, (ii) the luminescence yields are extremely low at room temperature, implying that the transition to the ground state from the triplet state is primarily nonradiative (Pappalardo, 1980), and (iii) the luminescence lifetimes are very long.

In fact, the study of the emission process in radiatively exposed tryptophan, tyrosine, and phenylalanine revealed that, only at very low temperature, the nonradiative decay of the triplet state becomes negligible – in this particular case below 77 K (Maki and Zuclich, 1975). The decay properties of these three amino acids at 77 K are noted (taken from Maki and Zuclich (1975)) in Table 2.1. At higher temperature, the luminescence yields are well below one.

Bohr and his collaborators (Bohr, Jalkanen, and Malik, 2005; Bohr and Malik, 2007) proposed for the first time that the underlying physical process responsible for the radiationless transition in the relaxation of an electronically activated biosystem is akin to the Auger process in the atoms. At about the same time, Srinivasan *et al.* (2005) established "in aromatics such as benzenes and pyridines, upon absorption of light, a substantial portion of the energy is directed in nonradiative pathway" and the nonradiative decay process is a fast one and takes place in a few tenths of a picosecond. In a series of earlier studies, Zewail and his collaborators attributed the radiationless process to electron transfer (Wan *et al.*, 1999, 2000).

In Section 2.2, we present the basic theory of nonrelativistic Auger process in light atoms and atomic ions, which clearly quantifies that the process is a fast one and dominates over the radiative decay in light elements. This is followed by a section where we discuss the features distinguishing Auger

Tab. 2.1 Ultraviolet (UV) absorption and subsequent luminescence data of three aromatic amino acids (Maki and Zuclich, 1975) at 77 K.

	Tryptophan	**Tyrosine**	**Phenylalanine**
Absorption maxima (nm)	280, 218	276, 222	258, 205
Fluorescence maxima (nm)	330	300	290
Fluorescence quantum yield, Φ_f	0.72	0.47	0.41
Phosphorescence maxima (nm)	435	395	385
Phosphorescence quantum yield, Φ_{ph}	0.17	0.53	0.59
Phosphorescence lifetime (s)	5.8, 6.65	2.7, 2.9	5.5, 7.7
Luminescence yield $[\Phi_f + \Phi_{\mathrm{ph}}]$	0.89	1.0	1.0

processes in atomic systems from similar ones in biomolecules. The refined theory to deal with the Auger-like processes in biosystems, called *Bio-Auger processes*, developed in Bohr, Greissen, and Malik (2008), is presented in Section 2.4. Section 2.5 deals with the relation of the Bio-Auger process and electron or charge transfer. In the last section, we discuss the theoretical interpretation of the Bio-Auger process for biosystems exposed to solar radiation and laser beams for a number of cases including the restoration to the original state of UV activated damaged thymine dimer as it comes into contact with the optically activated photolyase protein and cofactor.

2.2
Auger Process in Atoms

It was discovered in 1926, that a helium-like atomic ion, upon exposure to EMR of enough energy capable of causing an electronic excitation, returns to its ground state primarily by two processes, of which one is radiative transition and the other, the dominant one in light atoms, is the Auger transition named after (Auger, 1923, 1925a, 1925b, 1926), who first demonstrated that an atomic ion with a vacancy in its K shell emits an electron in the process of filling the vacant K shell preponderantly over radiative transition in light elements. Although Klein and Rosseland (1921) in their study of inelastic scattering of electrons by atoms raised the possibility of such a process because of its being a reversible phenomenon to inelastic excitation with electrons, Auger was the first scientist to demonstrate it experimentally, and hence, it is now universally known as the *Auger Process* or the *Auger Effect* in atomic physics and the emitted electron is termed as the *Auger electron*. Auger's experiment was immediately confirmed by de Broglie and Thibaud (1925); Robinson (1926), and Robinson and Cassie (1926). The first comprehensive theoretical understanding of the Auger process was provided by Wentzel (1927), which still forms the basis of all modern theories in atomic physics.

Since then, the Auger process has been studied extensively for many atoms and atomic ions and extensive literature on the subject is available. Some of the monographs and review articles are noted in Burhop (1952); Burhop *et al.* (1972); Chatterji (1976); Åberg and Howat (1982); Craseman (1985); Bambynek *et al.* (1972), and Wuilleemier (1976). The key feature associated with the de-excitation of an atom or atomic ion with a vacancy in one of the inner shells in light elements up to about atomic number, $Z = 30$, is that it is filled primarily by radiationless transition where one of the electrons in occupied upper shells makes a transition to the empty lower shell and another electron from the occupied upper shell takes up the transition energy and is ejected in the continuum. The time to fill a vacancy in a K shell is essentially Z-independent or weakly dependent and of the order of $10^{-14}-10^{-17}$ seconds. The competing fluorescence process involving radiative dipole transition to a vacancy in K shell is strongly Z-dependent and for He-like ions, is slower by a million seconds, that is, the fluorescence lifetime is about 10^{-9} seconds.

For atoms heavier than about atomic number 30, the fluorescence process is faster than the Auger process in filling a K vacancy, but since the biomolecules are primarily composed of lighter elements, it is expected that the Auger-like process would be the dominant one in filling a vacancy.

Wentzel (1927), for helium-like ions, with a nucleus of charge Z, treating the interaction between two electrons as perturbation, showed, using Heisenberg's theory (Heisenberg, 1926), that the emission rate of an electron, $T(\mathrm{el})$, as the other electron makes a transition from L to K shell is

$$T(\mathrm{el}) \cong 10^{15}\ \mathrm{s}^{-1} \tag{2}$$

whereas the rate of emission by an electric dipole $T(E1)$ is about $10^9\ Z^4\ \mathrm{s}^{-1}$. (3)

$$\text{Thus,}\quad T(\mathrm{el})/T(E1) = 10^6\ Z^{-4} \tag{4}$$

Hence, the empty K shell of a helium-like atom is predominately filled by the Auger process up to atomic number of about 36. This important conclusion of Wentzel has essentially proved to be correct both experimentally and theoretically.

In Figure 2.1, taken from Walters and Bhalla (1971), the experimental as well as calculated ratio of K shell fluorescence yield, $\omega_K(R)$, due radiative dipole transition probabilities, $T(R)$, and the total transition probabilities being the sum of $T(A)$, the Auger transition rate, and $T(R)$, is plotted as a function of atomic number. $\omega_K(R)$ in the plot is defined as

$$\omega_K(R) = T(R)/(T(R) + T(A)) \tag{5}$$

The calculated and observed $\omega_K(R)$ are shown, respectively, by dots and a solid line. Clearly for low atomic numbers, the vacancies in K shell are primarily filled by the Auger transition and not radiative transitions. This point is further illustrated in Figure 2.2 taken from Walters and Bhalla (1971), where calculated Auger and dipole X-ray rates are individually plotted as a function of atomic number in filling a vacancy in K shell by electrons from L shell and once again in light atoms, vacant K shell is filled by the Auger transition predominantly.

The calculations illustrated in Figures 2.2 and 2.3 are based on a refined version of Wentzel's theory. The rate of the Auger Process is not independent of atomic number, but is weakly dependent and is about 10^{-17} seconds in light elements. An appropriate theory for Auger-like transitions in biomolecules at much lower energies, noted in a subsequent section, is developed in analogy with the nonrelativistic version of the theory of the Auger transition in atomic systems in Section 2.4.

2.2.1
Nonrelativistic Expression for Auger Processes in Atoms and Ions

We first present the basic theory of nonradiative transition involving relaxation of atoms and ions having a vacancy in one of the inner shells that is valid for light atoms, that is, the basic theory of the Auger process. For light atoms, the relativistic effects can be neglected and hence, we shall present the nonrelativistic approach. The transition probabilities, $T(i \rightarrow f)$ of a system from an initial state $|i\rangle$ to a final state $|f\rangle$ due to a perturbing potential, $H(p)$, can be obtained from a first-order time-dependent perturbation theory in the limit of time, $t \rightarrow \infty$ (Walters and Bhalla, 1971; Asaad and Burhop, 1958) and is given by

$$T(i \rightarrow f) = \frac{2\pi}{\hbar} |\langle f|H(p)|i\rangle|^2 \rho_f \tag{6}$$

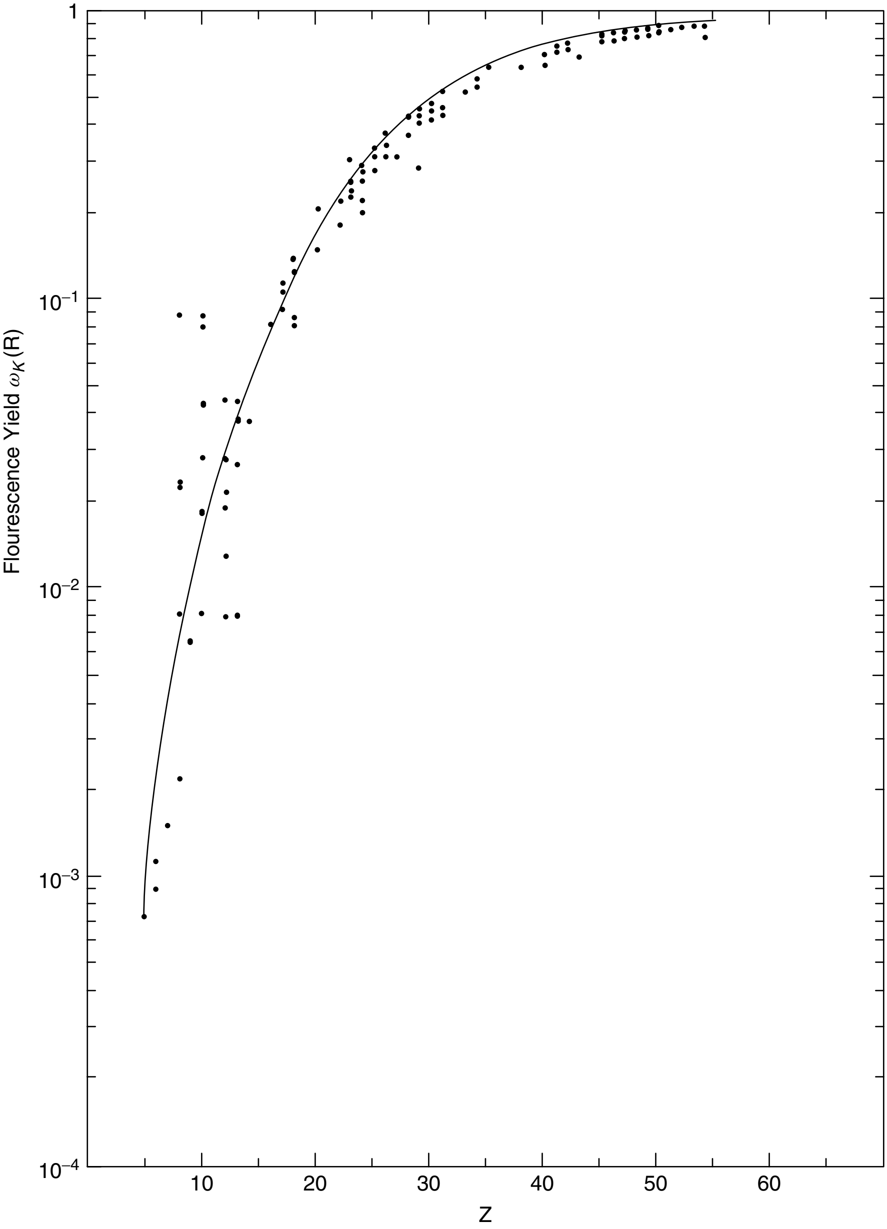

Fig. 2.2 Relative fluorescence yield defined by Eq. (5) is plotted as a function of atomic number *Z* for filling of the K shells in atoms taken from Walters and Bhalla (1971). Solid lines and dots are, respectively, theoretical calculations of Walters and Bhalla (1971) and experimental results noted in Walters and Bhalla (1971).

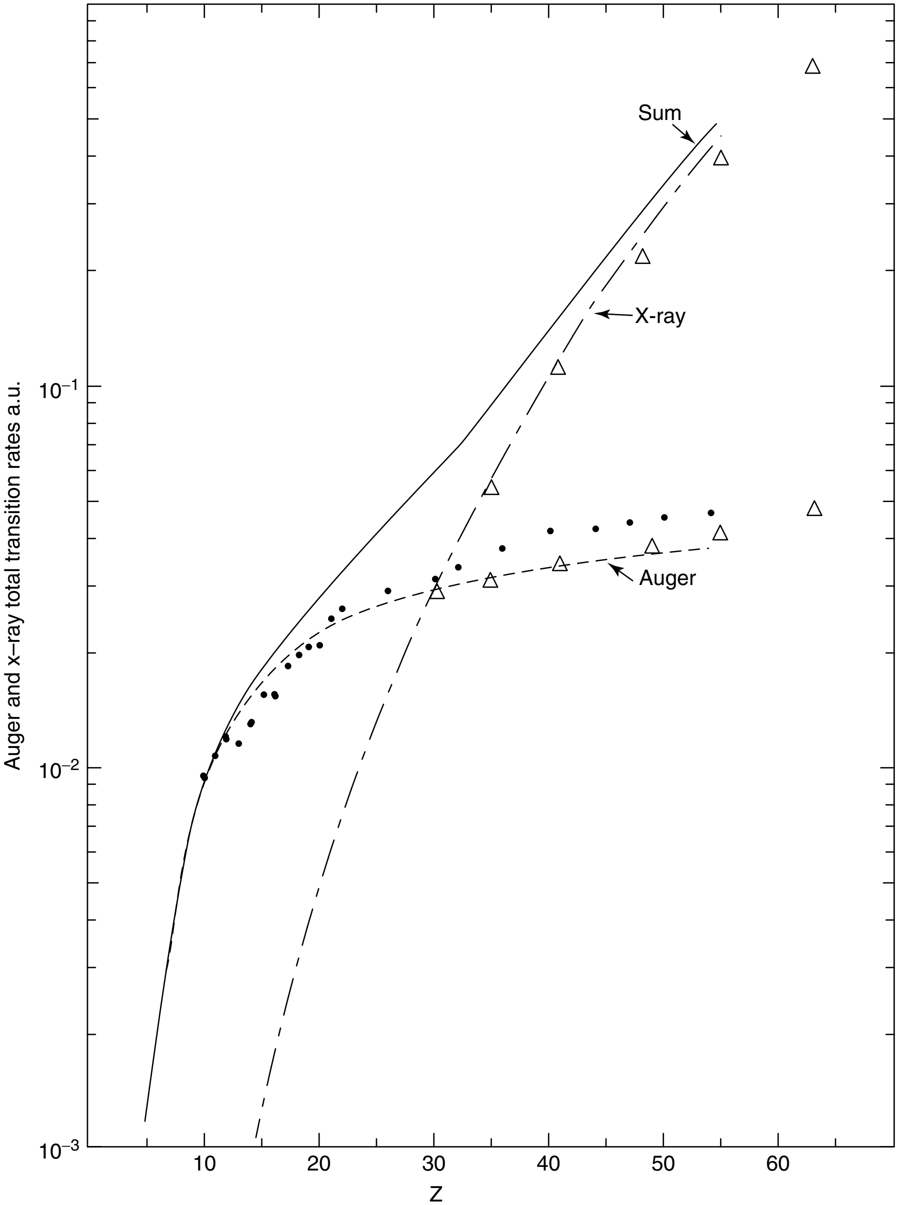

Fig. 2.3 Theoretical calculations of the Auger and X-ray rates done in Walters and Bhalla (1971), noted as Auger and X-ray, respectively, are plotted as a function of the atomic number *Z*. Experimental data for X-ray and Auger processes are shown as open triangles. The sum of theoretically calculated processes is shown as a solid line. au stands for atomic unit and 1 au for transition rate = 2.43 × 10^{-17} seconds. Taken from Walters and Bhalla (1971).

where ρ_f is the density of the final state. This is quite often referred to as the *golden rule of Fermi*. Since the process involves the transition of two electronic states, the initial and final states are, in the simplest approximation, the antisymmetric product of two electronic wave functions. Thus,

$$|i\rangle = [\varphi_\alpha(i)\varphi_\beta(j) - \varphi_\alpha(j)\varphi_\beta(i)] \tag{7}$$

and

$$|f\rangle = [\varphi_\gamma(i)\varphi_\delta(j) - \varphi_\gamma(j)\varphi_\delta(i)] \tag{8}$$

In Eqs (6)–(8), (i) and (j) stand for all coordinates of two electrons involved in the process, and (α, β) and (γ, δ) represent the complete set of quantum numbers of two electrons in the initial and the final states, respectively.

In the usual calculation of the Auger process in atoms and ions, the perturbing interaction, $H(p)$, is taken to be the unscreened Coulomb potential between two electrons involved in the transition, that is,

$$H(p) = e^2/|r_i - r_j| \equiv e^2/r_{ij} \tag{9}$$

where r_{ij} is the distance between two electrons, which can be expanded as follows:

$$\frac{1}{r_{ij}} = 4\pi \sum_{k,m} \frac{(k-|m|)!}{(k+|m|)!}\ \frac{r_<^k}{r_>^{k+1}} \times Y^*_{k,m}(\theta_i\varphi_i)Y_{k,m}(\theta_j\varphi_j) \tag{10}$$

In Eq. (10), the $Y_{k,m}$'s are normalized spherical harmonics and $r_<$ and $r_>$ are, respectively, the smaller and the larger of r_i and r_j.

Expanding φ's in spherical polar coordinates, for example,

$$\varphi_\alpha(i) = [P_{n_\alpha \ell_\alpha}(r_i)/r_i]Y_{\ell_\alpha m_\alpha}(\theta_i\varphi_i) \tag{11}$$

and similarly for φ_β, φ_γ, and φ_δ, one can evaluate the matrix element of Eq. (10) between final and initial states in the L–S coupling leading to (Burhop *et al.*, 1972; Walters and Bhalla, 1971)

$$\begin{aligned}\langle f|i/r_{ij}|i\rangle &= \tau(-1)^{L+\ell_\alpha+\ell_\delta} \\ &\times \sum_k \Bigg[R_k(\alpha,\beta,\gamma,\delta)\langle \ell_\gamma \| C^k \| \ell_\alpha\rangle \\ &\times \langle \ell_\delta \| C^k \| \ell_\beta \rangle \begin{Bmatrix} \ell_\gamma \ell_\delta L \\ \ell_\beta \ell_\alpha k \end{Bmatrix} \\ &+ (-1)^{L+S} R_k(\alpha,\beta,\gamma,\delta)\langle \ell_\gamma \| C^k \| \ell_\beta \rangle \\ &\times \langle \ell_\delta \| C^k \| \ell_\alpha \rangle \begin{Bmatrix} \ell_\gamma \ell_\delta L \\ \ell_\alpha \ell_\beta k \end{Bmatrix} \Bigg]\end{aligned} \tag{12}$$

where

$$\langle \ell_i \| C^k \| \ell_j \rangle = [(2\ell_i+1)(2\ell_j+1)]^{1/2} \times \begin{pmatrix} \ell_j & k & \ell_i \\ 0 & 0 & 0 \end{pmatrix} (-1)^{\ell_i}, \quad \begin{matrix} i = \gamma, \delta \\ j = \alpha, \beta \end{matrix} \tag{13}$$

In Eqs (12) and (13), the brackets represent $(6-j)$ and $(3-j)$ symbols of Wigner, respectively (Brink and Satchler, 1962). In evaluating Eq. (12), the antisymmetric nature of the wave functions in the initial as well as in the final state is taken fully into account and $\tau = 1$ if $(n_\alpha \ell_\alpha) \neq (n_\beta \ell_\beta)$ and $\tau = 1/\sqrt{2}$ if $(n_\alpha \ell_\alpha) = (n_\beta \ell_\beta)$. The symbol $R_k(\alpha, \beta, \gamma, \delta)$ represents the integration over the radial part,

$$\begin{aligned}R_k(\alpha,\beta,\gamma,\delta) &= \int_0^\infty dr_i P_{n_\alpha \ell_\alpha}(r_i) \\ &\times \left[\frac{1}{r_i^{k+1}} \int_0^{r_i} dr_j r_j^k + r_i^k \int_{r_i}^\infty \frac{dr_j}{r_j^{k+1}} \right] \\ &\times P_{n_\beta \ell_\beta}(r_j) P_{n_\gamma \ell_\gamma}(r_j) P_{n_\delta \ell_\delta}(r_i)\end{aligned} \tag{14}$$

The expression (14) is strictly true when $|i\rangle$ is a closed shell, but similar expressions are also valid for unfilled shells. In the atomic Auger process, $(n_\delta \ell_\delta)$ represents an ejected electron in a continuum state.

The calculated Auger transition shown in Figures 2.2 and 2.3 used these expressions with radial solutions of the self-consistent Hartree–Fock equations.

2.2.2
Expression for Radiative Transition

It is of interest to compare this expression with the one for radiative transition probability, $T(\text{rad}, i \rightarrow f)$, which is given by (Blatt and Weisskopf, 1952)

$$T(\text{rad}, i \rightarrow f) = 8\pi \left[\sum_{k=0} \frac{(k+1)}{k[(2k+1)!!]^2} \times (1/\hbar)(E_T/\hbar c)^{2k+1} B(k, I_i \rightarrow I_f) \right] \quad (15)$$

In Eq. (15), k is the polarity of the transition, E_T is the transition energy that is equal to the energy difference between the initial and final states, and $B(k, I_i \rightarrow I_f)$ is the reduced transition probability given by

$$B(k, I_i \rightarrow I_f) = (2I_i + 1)^{-1} \sum_{M_i M_f} \times |\langle f|O(k,m)|i\rangle|^2 \quad (16)$$

In Eq. (16), $O(k, m)$ is the electrostatic multipole operator of order k and z-projection m. I_i and I_f are the total spin of the initial and final states with respective z projection of M_i and M_f. The monopole term in Eq. (16) is zero because of the orthonormality of the initial and final wave functions. Noting that in atoms and molecules E_T is a few electron volts and $\hbar c = (197.327) \times 10^1$ eV Å, the dominant term in Eq. (16) is the dipole term, $k = 1$ in the summation. For dipole transition,

$$T(\text{rad}, i \rightarrow f) = 8\pi(2/3)(1/\hbar) \times (E_T/\hbar c)^3 B(1, I_i \rightarrow I_f) \quad (17)$$

A comparison of the integration over the radial part Eq. (14) involved in an Auger transition with that involved in $B(1, I_i \rightarrow I_f)$ in a radiative transition reveals that the dipole term in the two cases is identical if one omits the first integration in Eq. (14) and extends the lower limit in Eq. (14) to zero because of the normalization of $P(r_j)/r_j$ to one.

For a quick estimate, it is quite useful to use the classical expression for dipole transition, $T(1)$ given by (Heitler, 1954)

$$T(1) = (e^2/\hbar c^3)\nu^3 a^2 = \alpha(\nu a/c)^2 \nu \quad (18)$$

where α, ν, and a, respectively, are the fine structure constant (1/137.06), the transition frequency, and dipole length.

2.2.3
Bethe–Heitler Approximation for Auger Transition

In Eq. (9), one of the wave functions in the final state is outgoing continuum wave function, for example, $(rxP_{n_\delta \ell_\delta}(r_i))$, which is quite often approximated by a plane wave, $e^{ik_i \bullet r_i}$, which can be expanded as

$$e^{ik_i \bullet r_i} = \sum_{\ell=0}^{\infty} (2\ell+1) i^\ell j_\ell(k_i r_i) P_\ell(\cos\theta_i) \quad (19)$$

where j_i is the spherical Bessel function and θ_i, the angle between k_i and r_i.

Sometimes, in the matrix element Eq. (12), the initial state $P_{n_\alpha \ell_\alpha}(r_i)$ is also taken to be the radial part of a plane wave expansion Eq. (19). The situation is then akin to electron-impact ionization, first formulated by Bethe (1930). This approximate form of Eq. (12), known as the *Bethe approximation to the Auger Process*, has been applied to negative meson induced Auger transition in atoms (Lewis and Bethe, 1962).

Furthermore, if the integration over (dr_j) in the term in the square bracket in Eq. (14) is simply taken to be a single integral with limits from zero to infinity, the expression, as noted earlier, reduces to the matrix elements of electrostatic multipole transition, the dominant term being the dipole term. This is also termed sometimes as the *Bethe approximation*. For a quick estimate of the Auger transition rates, sometimes the classical expression of Heitler Eq. (18) for the dipole term is used, as was done in Bohr and Malik (2007). This is known as the *Bethe–Heitler approximation*.

2.2.4
Key Distinctions between Radiative and Auger Transitions

Whereas the operator acting on the particles involved in EMR is a one-body operator, that operating on particles responsible for the Auger transition is a two-body operator.

In case, particles undergoing radiative transition move in a well-defined mean field with very little correlation, radiative transitions involve transitions between two quantum states of a single particle. In the presence of correlations, radiative transitions can cause a change of quantum states of multiparticles, for example in the photo-ionization of the He-like ions; in a few percentage of cases, two electrons are emitted, but this is a relatively rare occurrence. But in all cases involving EMR, photons are emitted or absorbed.

The Auger process, on the other hand, is initiated by direct interaction between at least two particles. In case of electrons, quite often, the Auger transition is caused by electrostatic interaction between at least two electrons in a nonrelativistic case and the Møller interaction between two electrons in the relativistic situation. However, it may not be bare electromagnetic interaction but could be a modified version of it, because of the shielding of the interaction due to the presence of other electrons. Hence, the Auger transition is a nonradiative process. The Auger transition in atoms or ions is characterized by the transition of one electron from an excited state to an empty lower state, whereas another electron is emitted to a continuum or higher excited state by taking up the transition energy.

In case radiative transition is a dipole one, only transitions among the singlet state, for example $S_{\mathrm{ex}} \rightarrow S_0$ in Figure 2.1, or among the triplet states, for example $T_{\mathrm{ex}} \rightarrow T_1$ in Figure 2.1, are possible. On the other hand, Auger transitions could occur between triplet and singlet, singlet and singlet, and triplet and triplet states.

2.3
Distinctive Features of the Auger Processes in Atoms and those in Biomolecules

Bohr and his collaborators Bohr, Jalkanen, and Malik (2005); Bohr and Malik (2007) proposed for the first time that radiationless transitions akin to the Auger process in atoms could occur in

UV-activated biosystems. In particular, this process is responsible for repairing UV-induced damaged thymine dimer coming in contact with optically exposed photolyase, as observed in the experiment conducted by Mees *et al.* (2004). At about the same time, the experiment of Srinivasan *et al.* (2005) seems to establish that biomolecules in particular, being exposed to laser-induced electronic transition, relax substantially via a nonradiative process. Such nonradiative transitions are characteristic of the general nonradiative processes in biomolecules described by Bohr and his colleagues, and must be a quite common method of relaxation for electronically excited biomolecules composed of light atoms upto an atomic number of about 30. The reason for this is that the radiationless transition rate in the decay of the excited electronic state is likely to be more prevalent over radiative transition by a few orders of magnitude, if one extends the theory of atomic Auger transition to biomolecules, with vacancies in low-lying electronic states, involving light atoms. However, before extending this analogy, it is important to recognize the different physical situations between Auger processes in atoms and those in biomolecules as noted later.

The distinctive features of the Auger transition in atomic systems and that in biomolecules have some common and some differentiating characteristics, although both involve interaction between two electrons and the transition of one electron from an upper energy level to an empty level whereby this transition energy is taken up by the other participating electron.

Apart from this, the Bio-Auger process is distinct from the IC as discussed in Section 2.3.2.

2.3.1
Distinctive Features of the Auger and the Bio-Auger Processes

Auger transitions in atoms or atomic ions occur usually following an inner shell vacancy caused either by a photo process involving high-energy photons of X- or γ-rays or by collisions between atoms with incident particles such as electrons, protons, or other charged particles or in atom–atomic ion collisions. The incident energies of these projectiles are therefore usually greater than the ionization potentials of ejected electrons of atoms or atomic ions in consideration. In contrast, in biomolecules, the vacancies in electronic states are usually produced by UV radiations having wavelengths of about 200–400 nm (or optical radiation up to 700 nm, if the energy is sufficient to cause a transition between two electron states). The corresponding energies following Eq. (1) lie approximately between 2 and 6 eV in case of the biomolecules. Thus, the energy available for a radiationless Auger-like transition in biomolecules is noticeably lower than that for the usual Auger transitions in atoms and atomic ions.

In contrast, in standard Auger processes involving vacancies in the K shell of atoms or atomic ions, the transition energy, that is, the energy difference between an electron in an atom or in atomic ions in an initial state of the L, M, or higher shells prior to making the transition to the K shell and the electronic energy of the K shell, ranges from a few decades of electron volts to a few decades of kiloelectron volts. This requires a relativistic treatment for atoms heavier than those in the first transition metals in Mendeleev's periodic table.

In the case of the Auger process in atoms, atoms with a hole in a lower electronic state can relax to the ground state by Auger as well as by fluorescence. As shown in Figures 2.2 and 2.3, in light elements, the Auger process dominates, but still there are only two pathways for returning to the original ground state.

In the case of biomolecules exposed to UV or optical radiation, the electronic pathway following resonant absorption, if permitted by the various conservation laws, is likely to be the one depicted in Figure 2.1. An electron from the ground state singlet, S_0, upon absorbing a photon makes a transition to an excited state, S_{ex}, which is also a singlet state. The absorption is usually a dipole one, so there cannot be any change of electronic spin between the two states. The state, S_{ex}, may undergo some vibrational de-excitation emitting appropriate infrared radiation followed by a transition to a triplet state, T_{ex}, by ISC. This triplet state may or may not further decay to another electronic triplet state, T_0, following the emission of infrared vibronic radiation. The biomolecule may, then, return to its original S_0 ground state by radiative transition, which must be forbidden in character, since it involves a triplet to a singlet transition. Such forbidden transitions associated with phosphorescence have long lifetimes, for example, they could be a few seconds in some amino acids as noted in Table 2.1. On the other hand, the excited S_{ex} state can radiatively decay to a lower singlet state, which ultimately returns to the ground state by radiative transition akin to fluorescence in the atomic case.

Thus, the radiative relaxation or de-excitation to the ground state in biomolecules could take multiple paths of fluorescence and phosphorescence, whereas in atoms, the radiative relaxation involves only fluorescence.

The lifetimes associated with fluorescence in atoms and biomolecules could be very different. In light elements, the dominant relaxation pathway to the ground state is the Auger process. In biomolecules, it is now experimentally well established that the dominant relaxation modes to ground state are radiationless transition from an excited singlet (Crespo-Hernandez *et al.*, 2004) as well as triplet to ground states, which, as discussed in the next section, is akin to the Auger process and is due to electrostatic interaction, bare or shielded, between two or multiple electrons.

Thus, the pathways of de-excitation to ground state by the Bio-Auger process are, in principle, also multiple, although one or more particular ones may dominate.

Topological difference between the electronic configurations of biomolecules consisting usually of H, C, N, and O atoms building up the polypeptide protein backbone with aromatic or nonaromatic side-chains with occasionally ligand bonded metallic ions in some cases such as in metalloproteins, and those involved in the Auger process in the atomic case occurring in an atom or atomic ions is another important consideration in any theoretical formulation of the process. Whereas in the latter case, the transition is caused by the interaction between two electrons in the same atom or atomic ion, in biomolecules, it could be the interaction between electrons in different components of the molecules as well as in the same atom. In addition, the interaction involving electrons between two biomolecules in close proximity could lead to an Auger-like transition as is the case discussed in Bohr and Malik (2007).

In the Auger process in atoms, while an electron from a filled upper quantum state makes a transition to an unfilled energetically lower state, the other electron takes up the transition energy and is usually ejected to a continuum state because of the high energies involved. The electron undergoing Auger-like transition in the biomolecule, on the other hand, can impart its transition energy to many degrees of motion. Bio-Auger processes may also occur between states involving the same atomic or molecular constituent of the biomolecule. However, they could occur between electronic states of two different constituents. They could also take place between two biomolecules in close proximity and hence, all these multiple pathways are open and may contribute. In addition, associated with each electronic state, there are rotational and vibrational degrees of freedom, which could play a role in determining the transition rates. The rotational-vibronic states associated with an excited electronic state may have a lower energy than their counterpart associated with the electronic ground state because, the internuclear distance in the excited state is expected to be slightly larger compared to its value in the ground state due to the increased Coulomb repulsion among nuclei. As a result, the moment of inertia associated with the rotational degree of motion in the excited state increases, thus lowering rotational energy. The spring constant associated with the vibronic motion in the excited state is weaker than that in the ground state, resulting in a lower frequency. A part of the energy will have to be spent in replenishing the rotational-vibronic energy difference as an electron makes a transition from an upper to a lower or ground state. This is also the case for electromagnetic decay. Hence, the maxima of fluorescence spectra are expected to shift to longer wavelengths compared to those of the absorption spectra. This seems to be the case in Table 2.1.

Apart from this, the bulk of the transition energy of a few electron volts in Bio-Auger case may be taken up by electronic excitation or bond stretching or, if enough energy is available, in bond breaking or reorientation of biomolecules. (Note, in diatomic molecules, the electronic de-excitation could lead to dissociation, since sufficient energy of the order of a few electron volts is available.) In some cases, electrons may be emitted to the medium in which the biomolecules are embedded. The energy could also be transmitted to change the motion of the protein to which these biomolecules are attached as ligands.

To emphasize the differences in these two cases, these electronic transitions in biomolecules are termed as *Bio-Auger Processes.* The terminology coined as continuum of the Auger Process in Bohr, Jalkanen, and Malik (2005) is a special case of the general phenomenon of the Bio-Auger process.

2.3.2
Bio-Auger Process and Internal Conversion

An excited electronic state could also decay in radiationless fashion by IC. The underlying interaction for IC is the electron-vibronic coupling and hence, it involves the transition of one electron from an excited state to a lower or ground state. This coupling is usually spin independent and hence, IC occurs between either two singlet or two triplet states. In addition, no electron or charge transfer between two or more components of atoms of biomolecules is involved. It has, however, been difficult to quantitatively ascribe a picosecond timescale to this process (Crespo-Hernandez *et al.*, 2004). These characteristics of nonradiative transitions involving IC differ from those associated with Bio-Auger processes.

2.4
Single Orbital Theory of Bio-Auger Process

One may develop a theory of Bio-Auger process in an analogous way as Wentzel's theory of the Auger process, which assumes that both the initial and final wave functions of two electrons involved in transitions can be suitably expressed as an appropriate antisymmetric product of two electronic wave functions to take into account the Pauli principle, requiring that the electrons will return in the right spin state. On the other hand, wave functions of electronic states in biomolecules are likely to be highly correlated. Nevertheless, in the lowest approximations, for example, in tight binding or density functional theory, they are described by a single orbital wave function in the average or mean field generated by electron–electron, electron–nucleus, and nucleus–nucleus interactions. In the following, we discuss the theory of the Bio-Auger process in this approximation.

2.4.1
Wentzel's Analogous Theory in Biomolecules

A biomolecule, being activated by photons of sufficient energy has a hole electronic state in the ground state configuration and an electron in an excited state. The transition probabilities from this state, $\Psi_{\text{ex}} = |i\rangle$, to the original ground state, $\Psi_{\text{g}} = |f\rangle$, are again given by the expression Eq. (6), and in the lowest approximation $H(p)$ may be taken to be the Coulomb interaction between the electronic states.

Within the context of a density functional description of biomolecules, the electronic orbitals associated with these two states are Kohn–Sham's effective single particle orbitals (Kohn and Shams, 1965) that describe electronic motion in an average field generated by electrostatic attractions between a particular electron and all atomic nuclei and electrostatic repulsion generated by the interaction among all electrons and among all nuclei. Taking a particular nucleus as the reference point, electronic orbitals in the lowest approximation can be described by hydrogen-like orbitals with an effective charge, Z(eff). In that situation, a nonradiative transition from the state Ψ_{ex} to the ground state Ψ_{g} can be triggered by the electrostatic interaction between two electronic states associated with these two configurations. Unlike the Auger transition in atoms, the complete wave function in each of these states is not simply an electronic one, but in the Born–Oppenheimer approximation is the following:

$$\begin{aligned}\Psi_{\text{ex}}&[\varphi_{\gamma}(\text{el})\Phi_{\gamma}(\text{vib-rot})\Phi_{\gamma}(\text{sym})]\\ &\times[\varphi_{\delta}(\text{el})\Phi_{\delta}(\text{vib-rot})\Phi_{\delta}(\text{sym})]\\ &\equiv \varphi_{\gamma}(\text{el})\varphi_{\delta}(\text{el})C_{\gamma}D_{\delta} \qquad (20)\end{aligned}$$

$$\Psi_g[\varphi_\alpha(\text{el})\Phi_\alpha(\text{vib-rot})\Phi_\alpha(\text{sym})] \times [\varphi_\beta(\text{el})\Phi_\beta(\text{vib-rot})\Phi_\beta(\text{sym})] \equiv \varphi_\alpha(\text{el})\varphi_\beta(\text{el})A_\alpha B_\beta \quad (21)$$

$\varphi(\text{el})$, $\Phi(\text{vib-rot})$, and $\Phi(\text{sym})$ in each of the four states refer, respectively, to the electronic, vibrational-rotational, and any special symmetry function of these states. The vibronic-rotational and symmetry wave functions and their specific characteristic parameters, for example, vibrational frequency (or spring constants) and moment of inertial parameter characteristic of internuclear distance associated with each of the electronic states are, in principle, different in the initial and final states.

The electrostatic interaction acting between the two electronic states initiates a nonradiative transition with a half-life of $\tau(\text{non-R})$, which, following Wentzel, is expected to be nearly independent of $Z(\text{eff})$ (actually it should be weakly dependent) and is about 10^{-15} second. However, it is to be multiplied by an overlap matrix. Thus,

$$\tau(\text{non-R}) \cong (10^{-15}\,\text{s})|(A_\alpha B_\beta, C_\gamma D_\delta)|^2 \quad (22)$$

On the other hand, radiative half-lives from the excited to the ground state depend on the energy difference, E, multipolarity, k, and effective charges of the two states; but in case these effective charges are nearly equal, the half-life of the radioactive process $\tau(R)$ is

$$\tau(R) = f(E,k)Z(\text{eff})^4|(A_\alpha B_\beta, C_\gamma D_\delta)|^2 \quad (23)$$

$f(E,k)$ is noted in Section 2.2.2. As noted by Wentzel, for dipole-like transition, and for excitation of a few electron volts

$$\tau(R) \cong (10^{-9}\,\text{s})Z(\text{eff})^4|(A_\alpha B_\beta, C_\gamma D_\delta)|^2 \quad (24)$$

In the strictest sense, for radiative transitions the overlap integral involves only two vibronic states, for example, $|(A_\alpha, C_\gamma)|^2$ since $(B_\beta, D_\delta) = 1$. However, overlooking this difference, one gets the following relation:

$$\tau(\text{non-R})/\tau(R) \cong 10^6 Z(\text{eff})^{-4} \quad (25)$$

The maximum value of effective charge could be as large as ‘6’ for an electron in valence shell of 0; in that case, the nonradiative process is a million times faster. Even for an effective charge of (0.1), the nonradiative decay process is a hundred times faster. Since a typical radiative transition probability is about 10^{-9} second, the nonradiative half-lives are expected to be in the range of pico- to femtoseconds. Thus, the half-life associated with the nonradiation process in biomolecules is expected to be from pico- to femtoseconds, which is much faster than the corresponding half-lives for radiative transitions.

2.4.2
Theory of Nonradiative Decay in Mean Field Approximation in Biomolecules

Following the nonrelativistic theory of the Auger process in atoms, one can develop a theory of nonradiative relaxation in mean field approximation as done in Bohr, Greissen and Malik (2008). A mean field theory of radiationless transitions in biomolecules is to incorporate the following situations:

1. radiationless transitions within a particular atomic constituent of a biomolecule;
2. radiationless transitions caused by a pair of electrons from two different atomic or ionic constituents of a

biomolecule, or two biomolecules in close proximity;
3. the interaction causing the transition may not be a bare electrostatic one but a dressed or shielded one
4. the overall transition rate is to be multiplied by the overlap integral as in Eq. (22).

In addition, unlike the Auger process, an electron may not necessarily be ejected out of an atom or ion, but may be promoted to a higher unoccupied state permitted by the energy conservation; cause the change in bond structure or disruption of a bond or even dissociation; or be transmitted to protein motion in which the biomolecule is embedded. The cases where an electron is either ejected or promoted to a different electronic orbit can be dealt with within the basic theory of the Auger process noted under Section 2.2.1 with appropriate modifications.

The most important modification would be to identify single particle orbitals, φ_α, φ_β, φ_γ, and φ_δ with Kohn–Sham orbitals (Kohn and Shams, 1965) in the density functional approach. In addition, the origin of the coordinate system is to be carefully delineated. The interaction between two interacting electrons may not be a bare Coulomb interaction, but could be modified. Notwithstanding that, $H(p)$ is to be a function of relative distance, r_{ij}, due to the Galilean invariance. Thus, the general form of H_p is

$$H_p = \nu(r_{ij}) \tag{26}$$

One can expand $\nu(r_{ij})$ in Legendre polynomial of the angle between r_i and r_j, θ_{ij}. Any such $\nu(r_{ij})$ can be expanded in spherical harmonics $v(r_{ij}) = \sum_k v_k(r_i, r_j) P_k(\cos\theta_{ij})$, where $v_k(r_i, r_j) = (2k + 1)/2 \int v(r_{ij}) \ P(\cos\theta_{ij}) \ d(\cos\theta_{ij})$. Using the addition theorem one gets

$$v(r_{ij}) = \sum_{k,m} \left(\frac{4\pi}{2k+1}\right) v_k(r_i, r_j) \times Y^*_{km}(\theta_i\varphi_i) Y_{km}(\theta_j\varphi_j) \tag{27}$$

The origin of the coordinate system is to be chosen to be the same as the one used to determine the orbitals.

The matrix element of the transition in the L–S coupling is still given by Eq. (12), except the radial integration is

$$R_k(\alpha, \beta, \gamma, \delta) = \int_0^\infty dr_i \int_0^\infty dr_j P_{n_\alpha \ell_\alpha} \times (r_i) P_{n_\beta \ell_\beta}(r_j) v_k(r_i, r_j) \times P_{n_\gamma \ell_\gamma}(r_j) P_{n_\delta \ell_\delta}(r_i) \tag{28}$$

Thus, the transition rate is given by

$$T(i \to f) = (2\pi/\hbar)|(f|H_p|i)|^2 \rho_f \times |(A_\alpha B_\beta, C_\gamma D_\delta)|^2 \tag{29}$$

As noted earlier, in analyzing the overlap integral, one may take some simple clues from diatomic molecules. In these molecules the vibrational frequency is $f \approx \sqrt{k/\mu}$, where k and μ are, respectively, the spring constant and reduced mass. The potential surfaces of the harmonic oscillator idealizing the vibration in the ground and excited states could be different and in excited states it is likely to be much less steeper. Thus, the spring constant in the excited state is softer than that in the ground state – hence, the vibrational energy quanta in excited states could have less energy compared to that in ground state. Similarly, if the internuclear distance in the excited state becomes larger than that in the ground state, the moment of inertia of the excited state could become larger than that in the ground state, resulting in a lower rotational frequency. In that case, a part of the relaxation energy of an electron as it makes a transition to the ground state is to be imparted to increase the vibrational and rotational frequencies.

A possible way to improve on the Born–Oppenheimer approximation is to expand Ψ_{ex} in Eq. (20) and Ψ_g in Eq. (21) in a complete set of the electronic and vibrational-rotational wave function and also to introduce an electronic-vibrational-rotational coupling.

2.5
Bio-Auger Process and Charge or Electron Transfer

Quite frequently, a Bio-Auger process could result in charge or electron transfer in biomolecules, among its constituent atoms or molecules. To illustrate that, let us consider a biomolecule, $(A(S_0), B)$, with two constituents A and B. Upon absorbing a photon, it could go to an excited state, $(A(S_i), B)$, where S_i is an excited singlet state in A. Bio-Auger relaxation of this molecule could lead to $(A(S_0, S_j), B)$ where an electron is promoted to an excited orbital S_j as the other electron makes a transition to the ground state. In case S_j is a continuum state, one is left behind with $(A^+(S_0), B)$.

An alternative pathway is for $(A(S_i), B)$ to transfer an electron to an excited singlet or triplet orbital, $B(S_j)$ or $B(T_j)$ in B and the partner electron to ground or a lower excited state $A(S_0)$ or $A(S_n)$ of A, that is,

$$\begin{aligned}(A(S_i), B) \xrightarrow{BA} & (A^+(S_0)B^-(S_j)) \\ & \text{or } (A^+(S_0)B^-(T_j)) \\ & \text{or } (A^+(S_n)B^-(S_j)) \\ & \text{or } (A^+(S_n)B^-(T_j))\end{aligned} \tag{30}$$

where BA stands for Bio-Auger process, similar to ISC involving two constituents. Any one of the molecules could return to ground state $(A(S_0), B)$ by the Bio-Auger process by either emitting an electron or transmitting it to other components.

In case enough energy is available, a chain of charge exchange processes of the following type could occur

$$\begin{aligned}h\upsilon + (ABCD\ldots) &\rightarrow (A^*BCD\ldots) \\ &\rightarrow (A^{*+}B^{*-}CD\ldots) \\ &\rightarrow (A^{*+}B^*C^{*-}D\ldots) \\ &\rightarrow (A^*B^{*+}C^{*-}D\ldots) \\ &\rightarrow (A^*B^{*+}C^*D^-\ldots)\end{aligned} \tag{31}$$

Here $(ABCD\ldots)$ are constituent atoms or molecules, and * refers to an excited singlet or triplet state.

In particular any one of the charge transferred excited states could emit an electron in the medium, for example,

$$(A^{*+}B^{*-}CD\ldots) \rightarrow (A^{*+}B^*CD\ldots) + e \tag{32}$$

or transfer the energy to other degrees of freedom. It is to be noted that these states created by charge transfer could act as dipoles, which, by vibration, are capable of emitting EMR.

2.6
Experimental Evidence

In contrast to ascribing radiationless phenomena to vibronic processes (Lim, 1997; Avouris, Gilbert, and El-Sayed, 1977; Callomon, Perkin, and Lopez-Dalgado, 1972; Englman and Jortner, 1970; Jortner, Rice, and Hochstrasser, 1969; Bixon and Jortner, 1968, 1969; Siebrand, 1967) and/or proximity effect (Lim, 1997, 1986), a new aspect, namely the Bio-Auger process, leading to radiationless relaxation in excited biomolecule, is discussed herein. In reality, according to the principle of quantum physics, all these processes compete with each other and depending on various conservation laws such as energy, spin, and symmetry, one or the other may be dominant. In a de-excitation process, all of them compete with radiative transitions if permitted by the conservation laws. The characteristics of a Bio-Auger process are that (i) it can occur between singlet and singlet; singlet and triplet; and triplet and triplet states, (ii) in most cases, it resembles a charge transfer process, (iii) it is a fast process, the transition time being of the order of pico- to femtoseconds, and (iv) the transition energy can be distributed to various degrees of motion, such as electronic excitation, change in bond structures, for example, stretching or breaking, change in molecular orientation, or change in kinetic of the protein in which it is embedded, and so on.

2.6.1
Radiationless De-Excitation from Singlet and Triplet States

As noted in the review article DiBartolo (1980), the evidence pointing to radiationless transition is inferred from the fact that the yields from radiative transitions observed in photoprocesses are quite often less than unity, pointing toward the existence of other nonradiative decay modes. It is now well established that many simple aromatic and nonaromatic (McGlynn, Azumi, and Kinoshita, 1969; Lin, 1980; Schwoerer and Wolf, 2005; DiBartolo, 1980; Karplus and Porter, 1970; Henry and Kasha, 1968; Maki and Zuclich, 1975; Pappalardo, 1980) biomolecules such as amino acids, after absorbing radiation decay via fluorescence and phosphorescence, and the total yield of them are quite often less than the amount absorbed. Thus, the radiationless process is an important relaxation path in biomolecules and since some of these fluorescence rates are in the micro- to nanosecond region, the radiationless decay rates are faster than nanoseconds.

Radiative resonant absorption in biomolecules and in carbonyl compounds occurs in the 200–400 nm range, which corresponds to about 2.0–6.0 eV in excitation energy. Calculated excited energy levels of thymine using time-dependent density function theory with the hybrid B3LIYP taken from Bohr, Jalkanen, and Malik (2005) are noted in Table 2.2.

These are obtained by calculating the ground state configuration and then the appropriate energy difference between the highest occupied and the lowest unoccupied molecular orbitals. Clearly, there are a number of electronic states that could be photoexcited by laser emitting 250–300 nm radiation. It is interesting to note that the triplet excited states lie below the corresponding singlet state, thereby conforming to Hund's rule.

Theoretical calculations using multiconfigurational self-consistent field approach done by Park, Feenstra, and Zeweil (2006), and noted in Table 2.3, confirm the existence of singlet and triplet excited states in benzaldehyde and acetophene in a few electron volts excited region.

The fact that all three amino acids, tryptophan, tyrosine, and phenylalanine exhibit phosphorescence is indicative of the fact that following resonant absorption that can only excite the singlet

Tab. 2.2 Energies in electron volts of the first three electronic excited singlet and triplet states for thymine calculated at the B3LYP level of density functional theory (Bohr, Jalkanen, and Malik, 2005).

Electronic states in thymine	Energy (eV)	Transition wavelength (nm)
First excited singlet	3.8458	322.39
Second excited singlet	4.3019	288.21
Third excited singlet	4.4112	281.06
First excited triplet	3.3046	375.19
Second excited triplet	3.3477	370.36
Third excited triplet	3.8458	322.39

Tab. 2.3 Experimental, Expt, and theoretical calculations in multiconfiguration self-consistent field approximation, Thy, of first few singlet S_0, S_1, and S_2 and triplet states T_1, T_2, and T_3 in benzaldehyde and acetophene noted in (cm^{-1}) units, taken from Park, Feenstra, and Zeweil (2006) (1 cm^{-1} is equivalent to 1.234×10^{-4} eV)

	Benzaldehyde		Acetophene	
	Expt	Thy	Expt	Thy
S_0	0	0	0	0
S_1	26 921	26 541	27 279	–
S_2	35 191	35 978	35 402	36 074
T_1	25 183	24 844	25 791	25 556
T_2	–	24 821	–	25 102
T_3	–	–	–	36 361

ground states of these biomolecules to an excited singlet state, these singlet states make transitions to triplet states by ISC. These states must then partially decay to ground state by nonradiative processes, because at room temperature the total fluorescence and phosphorescence yields are much less than 100%.

The existence of nonradiative decay processes between two singlet states has now been established in many cases. For example, in bacteriorhodopsin, an excited S_n state decays to the first excited S_1 state by a nonradiative process and it is fast (Logunov *et al.*, 2001).

2.6.2
Radiationless Transitions and Electron Transfer in a Protein Complex in Photosynthesis

Zigmantas *et al.* (2002) have also reported a very fast 50–300 fs radiationless decay of the second excited S state to the first excited S state, which they ascribe to IC caused by electronic-vibronic coupling. However, as noted in Crespo-Hernandez *et al.* (2004), such interpretation for these fast transitions could be problematic. These fast rates are, however, characteristic of the Auger process in atoms.

Zigmantas *et al.* (2002) exposed a PCP complex to blue–green light, exciting the ground state, S_0, of peridinin to its second excited S_2 state, which then decays to the first excited S_1 state as well as transfers charge to an excited state, Q_x, in chlorophyll-*a* (chl-*a*). Direct excitation to S_1 state is not allowed by symmetry mismatch. This S_1 state not only relaxes to S_0 by radiationless transition but also causes a charge transfer to Q_y state in (chl-*a*). To accommodate this, they invoked the excitance of an admixture of special state termed *intramolecular charge transfer* (ICT) *state* with S_0.

In the context of the Bio-Auger process, the ICT and nonradiative decay go hand-in-hand. The electronically excited S_2 state, interacting with an adjacent electron, under the theory of the Bio-Auger process, could undergo nonradiatively to both S_1 and S_0 and transfer an electron to all states of (chl-*a*) if allowed by energy conservation laws (Bohr and Malik, 2008). Thus, the decay process for the $S_2 \rightarrow S_1$ transition produces simultaneously an electron transfer to both Q_x and Q_y states of (chl-*a*) and a nonradiative de-excitation to S_1, as shown in Figure 2.4. It is noteworthy that the decay rates for both processes are noted to be about 200 fs for the former case and 50–300 fs for the latter case in Zigmantas *et al.* (2002), that is, they could be equal, as it should be if they occur simultaneously. S_2 can also decay to S_1 by EMR. Thus, the total decay width $\Gamma_T(S_2)$ is a sum of the decay widths $\Gamma(BA : S_2 \rightarrow Q_x)$, $\Gamma(BA : S_2 \rightarrow Q_y)$, and $\Gamma(R : S_2 \rightarrow S_1)$

$$\Gamma(S_2) = \Gamma(BA : S_2 \rightarrow Q_x) + \Gamma(BA : S_2 \rightarrow Q_y) + \Gamma(R : S_2 \rightarrow S_1) \quad (33)$$

In Eqs (33) and (34), BA and R refer, respectively, to Bio-Auger and radiative transitions. The relation among corresponding half-lives, $T_T(S_2)$, $T(BA : S_2 \rightarrow Q_x)$, $T(BA : S_2 \rightarrow Q_y)$, and $T(R : S_2 \rightarrow S_1)$ is the following

$$\frac{1}{T_T(S_2)} = \frac{1}{T(BA : S_2 \rightarrow Q_x)} + \frac{1}{T(BA : S_2 \rightarrow Q_y)} + \frac{1}{T(R : S_2 \rightarrow S_1)} \quad (34)$$

Since $T(R : S_2 \rightarrow S_1)$ is expected to be much larger than the other two, the total half-life is essentially determined by the two half-lives associated with the Bio-Auger process transferring electrons to chl-*a*.

Once populated, S_1 could also relax by the Bio-Auger process transferring an electron to (chl-*a*) and simultaneously de-exciting nonradiatively to S_0. In that case, it would be unnecessary to invoke the idea of a charge transfer state mixed with S_1. One may also note that a component of Bio-Auger decay could retransfer an electron to S_2 if allowed energetically, which might appear as a misleading admixture of S_2 in S_1. The process is summarized schematically in Figure 2.4.

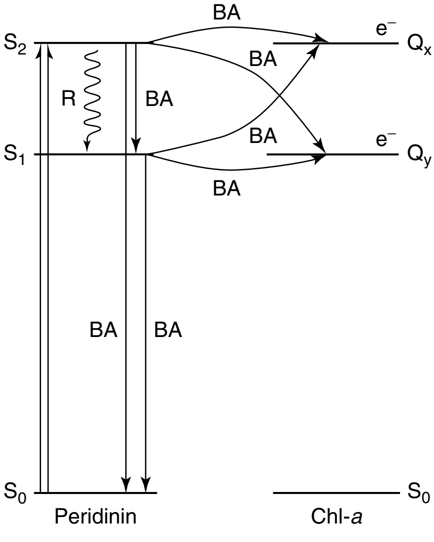

Fig. 2.4 Pathway of charge transfer from peridinin to chl-*a* in the PCP complex by a Bio-Auger process. Because of symmetry forbiddenness, direct transition of S_0 to S_1 by photo absorption cannot occur. Photoabsorption excites S_0 to S_2. S_2 state can undergo a Bio-Auger (BA) transition transferring an electron to Q_x and Q_y, if allowed by selection rules, and at the same time decay to S_1 by radiative IC processes (R). S_1 cannot decay to S_0 by a radiative process because of symmetry forbiddenness (it could probably decay weakly), but S_1 can undergo a BA transition transferring an electron to Q_x and Q_y and at the same time decay to S_0. The net result is a transfer of two electrons from peridinin to chl-*a*.

In short, following blue–green resonant absorption in PCP, exciting it directly to the second excited S_2 state, two electrons from peridinin could be transferred to (chl-*a*) component as S_2 relaxes to S_0 via S_1 and one has a (PCP) compound with a two-electron hole (i.e. the equivalent of two-positive charge) in peridinin and two-negatively charged (chl-*a*). This is a dipole, which, if it vibrates, could, of course, transfer its energy to its surroundings by emitting EMR or transfer two electrons to other molecules by transportation.

2.6.3
Dark Structure

The above analysis indicates that a biosystem undergoing electronic excitation could relax both electromagnetically and by a Bio-Auger process. In the latter case, it could in many instances be accompanied by intra- and intermolecular electron transfer. Such changes in the electronic configurations could lead to the change in the bond structure including rupture. The process is expected to take place in the pico- to femtosecond timescale. Such electron transfers and changes in molecular structure are consistent with the investigations in Wan *et al.* (1999, 2000) and Fiebig *et al.* (1999). The dark structure reported in Srinivasan *et al.* (2005) that accompanies radiationless transitions is the manifestation of these electronic transfers.

2.6.4
Exposures to Ultraviolet Radiation and Relaxation

Human beings and many other animals are being constantly exposed to the UV and blue optical radiation of sunlight. The energy of a few electron volts associated with these radiations, calculated using Eq. (1), is capable of electronic excitation in the biomolecules of exposed organic tissues. Experimentally, almost all of these tissues do not relax to ground state by emitting light or UV radiation, but they do relax and relax fast, since after being exposed to the sun, most animals including human beings do not glow and their skin remains intact. This relaxation is a radiationless process but cannot be thermal because 1-eV excitation is equivalent to about 10 500 K. The relaxation via a Bio-Auger process is a distinct possibility because it is fast and the energy of one electron could easily be transferred to the environment or the motion of the protein in which these molecules are embedded or attached to. The failure of these molecules to relax back to their original state could then inhibit the normal function of the host proteins leading to abnormal behaviors or cancer.

2.6.5
Bio-Auger Process Involving Intermolecular Electronic States

A Bio-Auger process is not restricted to electrons belonging to the same molecules or different molecular constituents of a complex biomolecule. It can occur between two electronic states in different molecular complexes in close proximity. In fact, the relaxation of the thymine dimer to two monomers in ground states, as it comes in close proximity of the negative ion cofactor, riboflavin, affected by the exposure of the photolyase protein to visible light, has been attributed to the Bio-Auger process

(Bohr, Jalkanen, and Malik, 2005; Bohr and Malik, 2007).

DNA nucleotides, for example thymine, being exposed to UV radiation of 200–300 nm lead to dimerization by causing a lesion between two adjacent nucleotides, in most cases forming a cyclobutane pyrimidine dimer (CPD) or duplex. This perturbic structure creates a kink in the helical axis of DNA leading to errors in genetic codes, which ultimately could lead to carcinogenetic behaviors. Very interestingly, this defect is repaired when the dimer comes in the proximity of photoactivated photolyse protein by optical radiation of 350–500 nm. The dimer coming in contact with this photoactivated photolyse is detached into two separate monomers in their ground states.

Density functional theoretical calculation indicates a number of excited triplet and singlet states of pyrimidine in the 3.5- and 5.0-eV excitation energy intervals as noted in Table 2.2. Resonant absorption of UV radiation would allow thymine to undergo a Frank–Condon transition to an excited S state, which after some vibronic de-excitation can cross over to a triplet state by ISC, also forming a duplex. Such a state is usually long lived and can be characterized as an isomeric state. Analogous states have also been identified in the photolysis of relatively simple polyatomic molecules such as formaldehyde (Moule and Walsh, 1975) and glyoxol (Tayer and Yardley, 1974). The energy of the dimer is less than the energy of the corresponding excited states of two monomers because of the radiative emission from the vibronic de-excitation.

Photolyase protein being exposed to optical radiation excites 7,8-didemethyl-8-hydroxy-5-deazariboflavin ($(FADH^-)$) cofactor to $(FADH^-)^*$, as discussed in Sancar (2003) and Sancar (1994). This excited negatively charged $(FADH^-)^*$ upon coming in proximity with the thymine dimer (which is actually in a hole excited state) interacts electrostatically by the Bio-Auger process leading an electron to make a transition to the ground state of thymine, thereby, detaching the dimer and in the process, de-exciting the cofactor to ground state, that is, to (FADH). The experiment of Mees *et al.* (2004) does indicate that (i) the thymine dimer is in a metastable state, (ii) a structural change in the stretch of damaged DNA containing the dimer exists, implying a change of electronic structure of the dimer compared to that of two monomers, (iii) a direct involvement of electrons in the detachment process of the dimer, and (iv) the detachment occurs only when $(FADH^-)^*$ formed in the exposure of photolyse protein to visible light, appears in the vicinity of the dimer.

Thus, the experimental analysis supports the theoretical picture. Furthermore, Mees *et al.* (2004) found that the repair process occurs in a time scale of about 10^{-10} second. The theoretical calculation of Bio-Auger transition rate done in the Bethe–Heitler approximation Eq. (13) agrees with this rate (Bohr and Malik, 2007).

Acknowledgments

The authors are thankful for the assistance of Mr Benedict Schurwanz and Mr M. Suhel Sabra in the preparation of the manuscript and figures, respectively. HGB is deeply appreciative of the grant from Danish Foundation for Quantum Protein Centre.

References

Åberg, T. and Howat, G. (1982) in *Theory of Auger effect, Handbuch d. Physik XXXI*, (ed. W. Mehlborn), Springer Verlag.

Asaad, W.N. and Burhop, E.H.S. (1958) *Proc. Phys. Soc., London*, **71**, 369.

Auger, P. (1923) *Comp. Rend.*, **177**, 169.

Auger, P. (1925a) *Comp. Rend.*, **180**, 65.

Auger, P. (1925b) *J. Phys.*, **6**, 206.

Auger, P. (1926) *Ann. Phys.*, **6**, 183.

Avouris, P., Gilbert, W.M., and El-Sayed, M.A. (1977) *Chem. Rev.*, **77**, 793.

Bambynek, W., Craseman, B., Fink, R.W., Freund, H.U., Mark, H., Swift, C.D., Price, R.E., and Venugopala Rao, P. (1972) *Rev. Mod. Phys.*, **44**, 716.

Bethe, H.A. (1930) *Ann. Phys. Leipzig*, **5**, 325.

Bixon, M. and Jortner, J. (1968) *J. Chem. Phys.*, **48**, 715.

Bixon, M. and Jortner, J. (1969) *J. Chem. Phys.*, **50**, 4061.

Blatt, J.M. and Weisskopf, V.F. (1952) *Theoretical Nuclear Physics*, John Wiley & Sons, Inc., New York.

Bohr, H.G., Jalkanen, K., and Malik, F.B. (2005) *Mod. Phys. Lett. B*, **19**, 473.

Bohr, H.G. and Malik, F.B. (2007) *Phys. Lett. A*, **362**, 460.

Bohr, H.G., Greissen, P., and Malik, F.B. (2008) *Int. Jour. Mod. Phys. B*, **22**, 4617.

Brink, D.M. and Satchler, G.R. (1962) *Angular Momentum*, Oxford University Press.

de Broglie, M. and Thibaud, J. (1925) *Comp. Rend.*, **180**, 179.

Burhop, E.H.S. (1952) *The Auger Effect and Other Radiationless Transitions*, Cambridge University Press.

Burhop, E.H.S., Asaad, W.N., Bates, D.R., and Esterman, I. (eds) (1972) *Advances in Atomic and Molecular Physics*, Vol. 8, Academic Press, p. 164.

Callomon, J.H., Perkin, J.E., and Lopez-Dalgado, R. (1972) *Chem. Phys. Lett.*, **13**, 125.

Chatterji, D. (1976) *The Theory of Auger Transitions*, Academic Press.

Craseman, B. (ed.) (1985) *Atomic Inner-Shell Physics*, Plenum Press.

Crespo-Hernandez, C.E., Cohen, B., Hare, P.M., and Kohler, B. (2004) *Chem. Rev.*, **104**, 1977.

DiBartolo, B. (ed.) (1980) *Radiationless Processes*, Plenum Press.

Englman, R. and Jortner, J. (1970) *J. Mod. Phys.*, **18**, 145.

Fiebig, T., Wan, C., Kelley, S.O., Barton, J.K., and Zewail, A.H. (1999) *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 1187.

Heisenberg, W. (1926) *Z.. Phys.*, **38**, 411.

Heitler, W. (1954) *The Quantum Theory of Radiation*, Oxford University Press.

Henry, B.R. and Kasha, M. (1968) *Annu. Rev. Phys. Chem.*, **19**, 161.

Jortner, J., Rice, S.A., and Hochstrasser, R.M. (1969) *Adv. Photochem.*, **7**, 149.

Karplus, M. and Porter, R.N. (1970) *Atoms and Molecules*, Chapter 7, Benjamin/Cumming Publication.

Klein, O. and Rosseland, S. (1921) *Z. Phys.*, **4**, 46.

Kohn, W. and Shams, L.J. (1965) *Phys. Rev., A*, **140**, 11.

Lewis, M. and Bethe, H.A. (1962) *Phys. Rev.*, **127**, 636.

Lim, E.C. (1986) *J. Phys. Chem.*, **90**, 6770.

Lim, E.C. (1997) *Adv. Photochem.*, **23**, 165.

Lin, S.H. (ed.) (1980) *Radiationless Transitions*, Academic Press.

Logunov, S.L., Volkov, V.V., Braun, M., and El-Sayed, M.A. (2001) *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 8475.

Maki, A.H. and Zuclich, J.A. (1975) *Topics in Current Chemistry*, Vol. 54, Springer Verlag, p. 115.

McGlynn, S.P., Azumi, T., and Kinoshita, M. (1969) *Molecular Spectroscopy of the Triplet States*, Prentice Hall.

Mees, A., Klar, T., Gnau, P., Hennecke, U., Eker, A.P.M., Carell, T., Essen, L.-O. (2004) *Science*, **306**, 2203.

Moule, D. and Walsh, A.D. (1975) *Chem. Rev.*, **75**, 67.

Pappalardo, R.G. (1980) *Radiationless Processes*, Plenum Press.

Park, A., Feenstra, J.S., and Zeweil, A. (2006) *J. Chem. Phys.*, **124**, 174707.

Robinson, H.R. (1926) *Nature*, **118**, 224.

Robinson, H.R. and Cassie, A.M. (1926) *Proc. R. Soc. London, Ser. A*, **113**, 282.

Sancar, A. (1994) *Biochemistry*, **33**, 2.

Sancar, A. (2003) *Chem. Rev.*, **103**, 2203.

Schwoerer, M. and Wolf, H.C. (2005) *Organic Molecular Solids*, Wiley-VCH Verlag GmbH.

Siebrand, W. (1967) *J. Chem. Phys.*, **46**, 440.

Srinivasan, R., Fenstra, J.S., Park, S.T., Xu, S., and Zewail, A.H. (2005) *Science*, **307**, 558.

Tayer, C.A. and Yardley, J.T. (1974) *J. Chem. Phys.*, **57**, 3992.

Walters, D.L. and Bhalla, C.P. (1971) *Phys. Rev., A*, **3**, 1919.

Wan, C., Fiebig, T., Kelley, S.O., Treadway, C.R., Barton, J.K., and Zewail, A.H. (1999) *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 6014.

Wan, C., Fiebig, T., Schiemann, O., Barton, J.K., and Zewail, A.H. (2000) *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 14052.

Wentzel, G. (1927) *Z. Phys.*, **43**, 524.

Wuilleemier, F.J. (ed.) (1976) *Photoionization and Other Probes of Many-Electron Interaction*, Plenum Press.

Zigmantas, D., Hiller, R.G., Sundström, V., and Polivka, T. (2002) *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 16760.

Further Reading

Bransden, B.H. and McDowell, M.R.C. (1992) *Charge Exchange and the Theory of Ion-Atom Collisions*, Clarendon Press, Oxford.

Harm, W. (1980) *Biological Effects of Ultraviolet Radiation*, Cambridge University Press.

Sinha, R.P. and Hader, D.-P. (2002) *Photochem. Photobiol. Sci.*, **1**, 225.

Hussain, I., Griffith, J., and Sancar, A. (1988) *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 2558.

Lin, S.H. (1966) *J. Chem. Phys.*, **44**, 3759.

Zigmantas, D., Polivka, T., Hiller, R.G., Yartsev, A., and Sundström, V. (2001) *J. Phys. Chem. A*, **105**, 10296.

3

Excited State Dynamics in Biomolecules

Gerrit Groenhof, Lars V. Schäfer, Martial Boggio-Pasqua, and Michael A. Robb

3.1 **Introduction** **95**
3.2 **Observing Excited State Dynamics on a Computer** **95**
3.2.1 Born–Oppenheimer Approximation 96
3.2.2 Conical Intersections 98
3.2.3 Excited State Molecular Dynamics 104
3.2.3.1 Surface Hopping 105
3.2.3.2 Excited State Quantum Chemistry 111
3.2.3.3 Mixed Quantum Classical Molecular Dynamics 113
3.3 **Applications** **116**
3.3.1 Photoreceptor Proteins 116
3.3.1.1 Retinal Proteins 116
3.3.1.2 Photoactive Yellow Protein 119
3.3.2 Fluoroproteins 121
3.3.2.1 Green Fluorescent Protein 121
3.3.2.2 asFP595 123
3.3.3 Photochemistry of DNA 126
3.3.4 Summary and Outlook 127
References **128**

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

3.1
Introduction

Organisms have evolved a wide variety of mechanisms to use and respond to sunlight. In many cases, the biological response is mediated by structural changes that follow photon absorption in the biomolecule. Exposure to high-energy photons, however, can cause irreversible damage in biological matter or even alter the genetic information. Despite the relevance of these photochemical processes in biology, little is known about the atomistic mechanisms by which irradiation can bring about changes in the chemical structure. Understanding these photobiological reactions requires insight into the underlying ultrafast molecular dynamics, which typically occur at femto- to picosecond timescales ($10^{-15}-10^{-12}$ s). As the relevant time and spatial resolutions are notoriously hard to access experimentally, computer simulations have become an essential tool to enhance our understanding of photobiological processes. Furthermore, the insights acquired by computation may be used directly to design artificial devices that mimic biological function. In this chapter, we first review the general theoretical framework that underlies our current understanding of photochemical reactions. Then we present how these theoretical concepts can be employed in practical simulation protocols. Finally, we discuss a number of representative applications, where such protocols have been used to elucidate the mechanisms of important photobiological processes.

3.2
Observing Excited State Dynamics on a Computer

The size and complexity of a typical biomolecular system, together with the timescales that must be reached, necessitates the use of classical molecular dynamics for the nuclear degrees of freedom in conjunction with a Molecular Mechanics force field. In molecular dynamics simulations, Newton's equations of motion are solved to obtain a trajectory of the dynamics of a molecule over a period of time, starting from an X-ray or nuclear magnetic resonance (NMR)-derived structure. To model the electronic rearrangement upon photon absorption, a quantum mechanical description is required for those parts of the system that are involved in the reaction. For the remainder, a

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

simple molecular mechanics force field model suffices. Both models are combined in a hybrid quantum mechanical/molecular mechanics (QM/MM) scheme. As schematically shown in Figure 3.1, a photochemical reaction starts in the excited state but ends in the ground state after radiationless decay via a conical intersection. To model the deactivation process, surface-hopping schemes are employed in the molecular dynamics (MD) simulations. We now present the theoretical foundations that provide the basis for computer simulations of excited state dynamics in complex biological molecules.

3.2.1
Born–Oppenheimer Approximation

All matter consists of nuclei and electrons that have a positive and negative charge, respectively. Electrostatic forces keep these particles together to form molecules of any size and complexity. The Schrödinger equation provides the mathematical description of the behavior of such many-body systems:

$$H\Psi = E\Psi \tag{1}$$

In this eigenvalue equation, H is the system's Hamilton operator, or Hamiltonian, which returns the total energy E of the system when operating on the many-body wave function Ψ. As in classical mechanics, the Hamiltonian is defined as the sum of the kinetic T and potential energy V:

$$H = T + V \tag{2}$$

From the wave function Ψ, all static properties of the system of interest can be derived. Dynamic information is obtained by integrating the time-dependent Schrödinger equation:

$$i\hbar\frac{\partial}{\partial t}\Psi = H\Psi, \tag{3}$$

with $\hbar$ the Planck constant divided by 2π. Unfortunately, an exact solution exists only if there are at most two interacting particles in the system, such as for instance the hydrogen atom. Solving the equations for any larger system requires approximations.

The first approximation is to ignore relativistic effects. Only for very heavy elements, which have very high nuclear charges, the inner shell electrons can move so fast that relativistic effects become important. Since biological molecules are mostly composed of first and second row elements, the electronic velocities are sufficiently low for

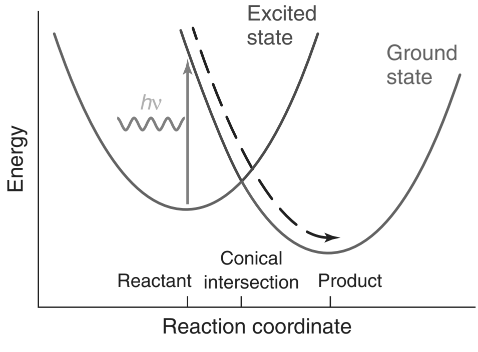

Fig. 3.1 One-dimensional scheme of a photochemical reaction. After photon absorption (cyan arrow), evolution takes place on the excited state potential energy surface (red), until the system hits the conical intersection. At the intersection, a radiationless transition to the ground state occurs (green). After the decay, the system continues evolving in the ground state. (Please find a color version of this figure on the color plates)

relativistic effects to be ignored. Within this approximation, the nonrelativistic Hamiltonian is given by

$$H = T_N + T_e + U(\mathbf{r}, \mathbf{R}) \tag{4}$$

where T_e and T_N are the operators of the kinetic energy of the electrons and the nuclei, respectively, and $U(\mathbf{r}, \mathbf{R})$ is the total potential energy of the electrons and nuclei together. Since relativistic effects are ignored, no spin interactions are included in the Hamiltonian. The vector **r** denotes the set of electronic coordinates and the vector **R** stands for the nuclear coordinates.

The next approximation is to separate the motion of the electrons from that of the nuclei. This so-called Born–Oppenheimer approximation provides the basis for modern quantum chemistry. The approximation rests on the mass difference between electrons and nuclei. Nuclei are much heavier than electrons, so that the latter adopt instantaneously to displacements of the former. Electrons "see" nuclei standing still, whereas the nuclei move on potential energy landscapes created instantly by the faster electrons. Within the Born–Oppenheimer approximation, the electronic and nuclear wave functions can be treated independently.

First, the Schrödinger equation is solved for electrons moving in a framework of fixed nuclei. Thus, the nuclear kinetic energy (T_N) is dropped from the full Hamiltonian (Eq. (4)) to yield

$$H^e = T_e + U(\mathbf{r}, \mathbf{R}) \tag{5}$$

where the superscript e indicates the *electronic* Hamiltonian. The electronic wave functions are the eigenfunctions of this Hamiltonian

$$H^e \psi_i(\mathbf{r}; \mathbf{R}) = V_i(\mathbf{R}) \psi_i(\mathbf{r}; \mathbf{R}) \tag{6}$$

with $\psi_i(\mathbf{r}; \mathbf{R})$ and $V_i(\mathbf{R})$ the electronic wave functions and their electronic energies, respectively, both of which depend *parametrically* on the nuclear coordinates. The wave functions are the adiabatic electronic states, representing the electronic ground state ($i = 0$), excited state ($i = 1$), second excited state ($i = 2$), and so on. How these wave functions are obtained in practical computations is the subject of modern quantum chemistry (Szabo and Ostlund, 1989) and is briefly discussed in Section 3.2.3.2.

The adiabatic wave functions can be made *orthonormal*, that is,

$$\langle \psi_i | \psi_j \rangle = \int_{-\infty}^{\infty} \psi_i(\mathbf{r}; \mathbf{R})^* \psi_j(\mathbf{r}; \mathbf{R}) d\mathbf{r} = \delta_{ij} \tag{7}$$

where the Dirac bra-ket notation ($\langle | \rangle$) has been introduced that is used throughout the text and where δ_{ij} is the Kronecker delta-function, which is 1 if $i = j$ and 0 otherwise.

Within the Born–Oppenheimer approximation, the adiabatic electronic states provide a complete basis to expand the molecular wave function in

$$\Psi(\mathbf{r}, \mathbf{R}) = \sum_j \chi_j(\mathbf{R}) \psi_j(\mathbf{r}; \mathbf{R}) \tag{8}$$

where the expansion coefficients $\chi_j(\mathbf{R})$ are the nuclear wave functions. Substituting the expression for the molecular wave function in the time-independent Schrödinger equation (Eq. (1)) and multiplying by the adiabatic electronic wave function $\psi_i^*(\mathbf{r}; \mathbf{R})$ from the left, followed by integration over the electronic coordinates **r** leads to the following set of coupled equations:

$$\sum_j H_{ij}(\mathbf{R}) \chi_j(\mathbf{R}) = E(\mathbf{R}) \chi_i(\mathbf{R}) \tag{9}$$

with

$$H_{ij}(\mathbf{R}) = \langle \psi_i(\mathbf{r};\mathbf{R})|H|\psi_j(\mathbf{r};\mathbf{R})\rangle = \langle \psi_i(\mathbf{r};\mathbf{R})|T_N|\psi_j(\mathbf{r};\mathbf{R})\rangle + V_i(\mathbf{R})\delta_{ij} \quad (10)$$

The nuclear kinetic energy operator is defined as

$$T_N = -\sum_k \frac{\hbar^2}{2M_k}\nabla^2_{\mathbf{R}_k} \quad (11)$$

where M_k is the mass of nucleus k, and the sum runs over all nuclei. Using this relation, Eq. (10) can be rewritten as (Köppel, Domcke and Cederbaum, 1984)

$$H_{ij}(\mathbf{R}) = [T_N + V_i(\mathbf{R})]\,\delta_{ij} - \Lambda_{ij}(\mathbf{R}) \quad (12)$$

The *nonadiabatic* operator elements $\Lambda_{ij}(\mathbf{R})$ are defined as

$$\Lambda_{ij}(\mathbf{R}) = \sum_k F^k_{ij}(\mathbf{R})\nabla_{\mathbf{R}_k} + G_{ij}(\mathbf{R}) \quad (13)$$

$$F^k_{ij}(\mathbf{R}) = \frac{\hbar^2}{M_k}\langle \psi_i(\mathbf{r};\mathbf{R})|\nabla_{\mathbf{R}_k}|\psi_j(\mathbf{r};\mathbf{R})\rangle \quad (14)$$

$$G_{ij}(\mathbf{R}) = \sum_k \frac{\hbar^2}{2M_k}\langle \psi_i(\mathbf{r};\mathbf{R})|\nabla^2_{\mathbf{R}_k}|\psi_j(\mathbf{r};\mathbf{R})\rangle \quad (15)$$

In contrast to H^e (Eq. (5)), where T_N was omitted, H_{ij} is *not* diagonal in the basis of the adiabatic electronic wave functions (ψ_i). The individual electronic states are coupled via nuclear motions. Nonadiabatic coupling is the key player in a photochemical reaction, as shown below.

In the limit that the electronic wave functions vary slowly with the nuclear dynamics, the nonadiabatic operators $F^k_{ij}(\mathbf{R})$ and $G_{ij}(\mathbf{R})$ are vanishingly small and can be safely neglected. Thus, the crux of the Born–Oppenheimer approximation is that H_{ij} is assumed to be diagonal everywhere:

$$H_{ij}(\mathbf{R}) = [T_N + V_i(\mathbf{R})]\,\delta_{ij} \quad (16)$$

With this assumption, the total molecular wave function becomes product of a *single* nuclear and a *single* electronic wave function:

$$\Psi^{\text{tot}}_i(\mathbf{R},\mathbf{r}) = \chi_i(\mathbf{R})\psi_i(\mathbf{r};\mathbf{R}) \quad (17)$$

which implies that nuclear motion takes place on a *single* electronic potential energy surface $V_i(\mathbf{R})$ of a given electronic state i. The Born–Oppenheimer approximation not only reduces the computational complexity of the equations that have to be solved but also provides a conceptually more intuitive picture of molecular structure and dynamics.

3.2.2
Conical Intersections

In regions of configuration space where the energy gaps between electronic states become comparable to the energy gaps between the vibrational states of the nuclei, resonance will occur between nuclear vibrations and electronic transitions. The populations of the adiabatic wave functions become strongly dependent on the nuclear dynamics and the nonadiabatic coupling operator ($\mathbf{\Lambda}$, Eq. (13)) can no longer be neglected. In regions of strong nonadiabatic coupling, the Born–Oppenheimer approximation breaks down, and nuclear dynamics induces population transfer between different electronic states, which is the origin of a radiationless transition. Furthermore, as we will demonstrate in this section, adiabatic potential energy surfaces can even intersect due to the coupling. These crossings provide efficient funnels for radiationless deactivation of the excited state and therefore play a crucial role in photochemistry.

In principle *all* electronic states $\psi_i(\mathbf{r};\mathbf{R})$ are involved in the nonadiabatic coupling. However, in practice, only states that are close in energy couple significantly. Therefore, the set of states included in the coupling can be truncated, thereby reducing the dimension of the nonadiabatic coupling matrix $\mathbf{\Lambda}$.

The nonadiabatic coupling operators Λ_{ij} are nonlocal derivative operators that depend inversely on the energy gap between the coupled adiabatic states (Worth and Cederbaum, 2004):

$$F_{ij}^{k}(\mathbf{R}) = \frac{\hbar^2}{M_k}\langle\psi_i(\mathbf{r};\mathbf{R})|\nabla_{\mathbf{R}_k}|\psi_j(\mathbf{r};\mathbf{R})\rangle = \frac{\hbar^2}{M_k}\frac{\langle\psi_i(\mathbf{r};\mathbf{R})|\nabla_{\mathbf{R}_k}H^e|\psi_j(\mathbf{r};\mathbf{R})\rangle}{V_j - V_i} \tag{18}$$

As the gap $(V_j - V_i)$ narrows, coupling increases and the nuclear wave function that is initially on one surface will spread to another without losing energy, and a radiationless transition occurs.

If the surfaces become degenerate, that is, $V_j - V_i = 0$, the coupling becomes infinite. Owing to the nonlocal nature of the coupling matrix elements in the adiabatic representation, it is convenient to switch to a diabatic representation, in which the nonadiabatic coupling operator becomes a local, potential-like operator. The switch to the diabatic wave functions φ^{dia} is achieved by a unitary transformation of the adiabatic wave functions ψ^{adia} at each point in space (Worth and Cederbaum, 2004):

$$\boldsymbol{\varphi}^{\text{dia}} = \mathbf{S}(\mathbf{R})\boldsymbol{\psi}^{\text{adia}} \tag{19}$$

From now on, we no longer use the superscripts dia and adia to indicate the diabatic φ and adiabatic ψ wave functions. In the diabatic representation, the complete Hamiltonian (Eq. (12)) becomes

$$H_{ij}(\mathbf{R}) = T_N\delta_{ij} + W_{ij}(\mathbf{R}) \tag{20}$$

and the molecular Schrödinger equation (Eq. (12)) can be written in matrix notation as

$$\mathbf{H}\boldsymbol{\chi} = [T_N\mathbf{1} + \mathbf{W}(\mathbf{R})]\,\boldsymbol{\chi} = \mathbf{E}\boldsymbol{\chi} \tag{21}$$

where $\mathbf{1}$ is the identity matrix, $\boldsymbol{\chi}$ the vector of nuclear wave functions (χ_i, Eq. (8)), and $\mathbf{W}(\mathbf{R})$ is the diabatic potential energy matrix, which in contrast to the adiabatic potential matrix $\mathbf{V}(\mathbf{R})$ contains only *local* terms.

To illustrate the concept of the surface crossing, we consider a molecule in which there is coupling between two diabatic states A and B only. We can expand the potential matrix elements in a Taylor series around an arbitrary point $\mathbf{R}_0$:

$$\mathbf{W}(\mathbf{R} - \mathbf{R}_0) = \mathbf{W}^{(0)} + \mathbf{W}^{(1)} + \mathbf{W}^{(2)} + \cdots \tag{22}$$

At $\mathbf{R}_0$, we can choose the diabatic and adiabatic states to be equal. Then, the zeroth-order matrix, $\mathbf{W}^{(0)}$, is a diagonal matrix in which the elements correspond to the energies E_A, E_B of the diabatic states φ_A and φ_B at $\mathbf{R}_0$, which, by our choice of origin, are identical to the *adiabatic* energies V_1 and V_2:

$$\mathbf{W}^{(0)} = \frac{E_A + E_B}{2}\mathbf{1} + \begin{pmatrix} -\frac{E_B - E_A}{2} & 0 \\ 0 & \frac{E_B - E_A}{2} \end{pmatrix} = \mathbf{V}(\mathbf{R}_0) \tag{23}$$

For small displacements $\Delta\mathbf{R}$ around $\mathbf{R}_0$, the Taylor expansion in Eq. (22) can be truncated after the first-order term

$$\mathbf{W}^{(1)} = \frac{\boldsymbol{\lambda} \cdot \Delta\mathbf{R}}{2}\mathbf{1} + \begin{pmatrix} -\frac{1}{2}\delta\boldsymbol{\kappa} \cdot \Delta\mathbf{R} & \boldsymbol{\kappa}^{AB} \cdot \Delta\mathbf{R} \\ \boldsymbol{\kappa}^{AB} \cdot \Delta\mathbf{R} & \frac{1}{2}\delta\boldsymbol{\kappa} \cdot \Delta\mathbf{R} \end{pmatrix} \tag{24}$$

in which we have introduced linear potential constants that are defined as

$$\begin{aligned} \delta\boldsymbol{\kappa} &\equiv \nabla_{\mathbf{R}} \left(E_B - E_A\right)|_{\mathbf{R}_0} \\ \boldsymbol{\kappa}^{AB} &\equiv \nabla_{\mathbf{R}} \left\langle\varphi_A|H^e|\varphi_B\right\rangle|_{\mathbf{R}_0} \\ \boldsymbol{\lambda} &\equiv \nabla_{\mathbf{R}} \left(E_A + E_B\right)|_{\mathbf{R}_0} \end{aligned} \tag{25}$$

We now choose $\mathbf{R}_0$ to be the point of degeneracy, where E_A and E_B are equal. Under this condition, $\mathbf{W}^{(0)}$ becomes a constant that we conveniently set to zero. The adiabatic potential energy surfaces V_1 and V_2 are obtained by diagonalizing the diabatic potential matrix $\mathbf{W}$. Thus for the two-state system considered here,

$$V_{1,2} = \frac{1}{2}\boldsymbol{\lambda} \cdot \Delta\mathbf{R} \pm \frac{1}{2} \times \sqrt{\left[\delta\boldsymbol{\kappa} \cdot \Delta\mathbf{R}\right]^2 + 4\left[\boldsymbol{\kappa}^{AB} \cdot \Delta\mathbf{R}\right]} \tag{26}$$

The necessary condition for a crossing between the two potential energy surfaces at $\mathbf{R}_0$ is that the two electronic energies are identical. Therefore, two conditions need to be fulfilled simultaneously:

$$\begin{aligned} \delta\boldsymbol{\kappa} \cdot \Delta\mathbf{R} &= 0 \\ \boldsymbol{\kappa}^{AB} \cdot \Delta\mathbf{R} &= 0 \end{aligned} \tag{27}$$

Thus, to first order, the degeneracy is lifted in the two-dimensional space spanned by the vectors $\delta\boldsymbol{\kappa}$ and $\boldsymbol{\kappa}^{AB}$, known as the *gradient difference vector* and *derivative coupling vector*, respectively. The space spanned by these two vectors is often referred to as the *branching space*, or *g–h plane*. Furthermore, as is evident from Eq. (26) and illustrated in Figure 3.2, the topology of the surfaces is that of a double cone, with the point of degeneracy at the apex. Orthogonal to the two-dimensional branching space exists the so-called intersection space (or seam space), in which the energies of the two states remain degenerate to first

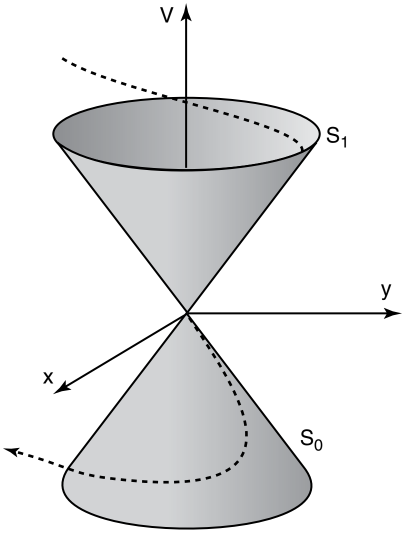

Fig. 3.2 A conical intersection plotted in the two-dimensional branching space that is spanned by the gradient difference vector (*g*) and the derivative coupling vector (*h*). Motion away from the intersection in the *g*–*h* plane lifts the degeneracy between the two electronic states S$_1$ and S$_0$. The dotted line shows a path of a nuclear trajectory passing from one electronic state (S$_1$) to another (S$_0$) through the intersection funnel.

order. In a molecule with N internal degrees of freedom, the intersection space thus forms an $N-2$ dimensional seam, each point of which is a conical intersection. If a molecule has less than two degrees of freedom, the two conditions for degeneracy (Eq. (27)) cannot be simultaneously fulfilled. Therefore in diatomic molecules states of the same symmetry do not cross, as was formulated by von Neumann and Wigner (1929).

The concept of the intersection seam is illustrated in Figure 3.3 for a hypothetical triatomic molecule. In our model system, there are three internal degrees of freedom: two bond lengths (x_1 and x_2) and one angle (α), which are sufficient to define the configuration of the molecule. Since *two* degrees of freedom are required to span the branching space, there is only *one* degree of freedom available to form the intersection seam. In our simple model system we let the branching space coordinates (i.e. the gradient difference and derivative coupling vectors) coincide with the two bond-length variables x_1 and x_2. In reality, the branching space coordinates are usually more complex linear combinations of the three internal degrees of freedom.

The first plot in Figure 3.3 shows the two adiabatic potential energy surfaces (denoted by S_0 and S_1) in the branching space. When projected onto this space, the surfaces adopt a double cone shape with a single intersection point at the apex. Any displacement away from the apex in either the x_1 or x_2 coordinate lifts the degeneracy.

The second plot shows the surfaces in the subspace spanned by the derivative coupling vector x_2 and the third independent degree of freedom, the bond angle α. The gradient difference vector x_1 is kept fixed at the conical intersection. Therefore, the only coordinate that can lift the degeneracy is x_2. If the latter is at the conical intersection as well, the surfaces intersect irrespective of the value of the angle variable (α). The intersection seam is thus a one-dimensional line that is parallel to α.

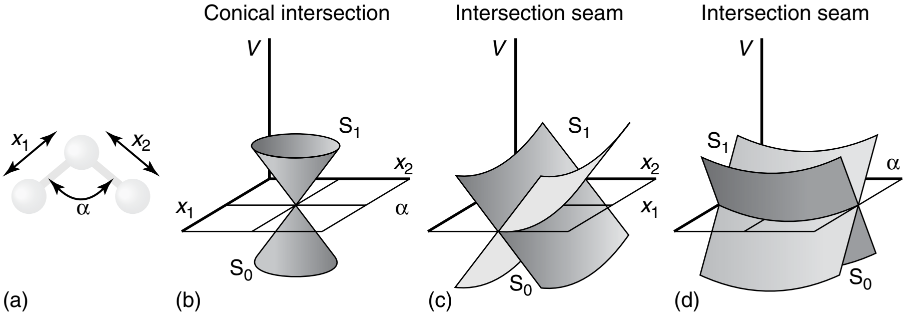

Fig. 3.3 Surface crossing between two potential energy surfaces S_1 and S_0 in a hypothetical triatomic molecule (a). We let x_1 and x_2 be parallel to the gradient difference vector and the derivative coupling vector, respectively, and α be the remaining degree of freedom. When projected onto the branching space spanned by x_1 and x_2, the surfaces of S_0 and S_1 intersect in a single point, the conical intersection (b). In the two other subspaces, spanned by either x_2 and α (c), or by x_1 and α (d), there is an intersection line between the surfaces. In the second plot x_1 and in the third x_2 are assumed to be at their conical intersection coordinates.

The third plot shows the projection of the two surfaces onto the subspace spanned by the gradient difference vector x_1 and the bond angle variable α. The derivative coupling vector x_2 is constrained to be at the conical intersection. Under this condition, only x_1 can lift the degeneracy and the surfaces intersect along a seam parallel to α.

In this idealized example, symmetry arguments could have also been used to illustrate the concept of the conical intersection seam. If x_1 or x_2 become identical, the point group of the molecule changes from C_s to C_{2v}. In the first point group, the electronic wave functions belong to the same irreducible representation, and their surfaces cannot intersect due to symmetry reasons. In the C_{2v} point group, in contrast, the wave functions belong to different irreducible representations, and a crossing is not symmetry forbidden.

The conical intersection has a number of peculiar characteristics that can be exploited in excited state molecular dynamics simulations. If we consider a small displacement $\mathbf{q}$ away from the cone tip in the branching plane, the diabatic energy matrix $U_{A,B}$ becomes

$$\begin{aligned} U_{A,B}(\mathbf{q}) = \begin{pmatrix} H_{AA}(\mathbf{q}) & H_{AB}(\mathbf{q}) \\ H_{AB}(\mathbf{q}) & H_{BB}(\mathbf{q}) \end{pmatrix} &= S(\mathbf{q})\mathbf{1} \\ &+ \begin{pmatrix} -\Delta H(\mathbf{q}) & H_{AB}(\mathbf{q}) \\ H_{AB}(\mathbf{q}) & \Delta H(\mathbf{q}) \end{pmatrix} \end{aligned} \tag{28}$$

with (see Eq. (24))

$$\begin{aligned} \Delta H(\mathbf{q}) &= \frac{H_{BB}(\mathbf{q}) - H_{AA}(\mathbf{q})}{2} = \boldsymbol{\delta\kappa} \cdot \mathbf{q} \\ H_{AB}(\mathbf{q}) &= \boldsymbol{\kappa}^{AB} \cdot \mathbf{q} \\ S(\mathbf{q}) &= \frac{H_{BB}(\mathbf{q}) + H_{AA}(\mathbf{q})}{2} \end{aligned} \tag{29}$$

The matrix $U_{A,B}(\mathbf{q})$ is the two-state Hamiltonian matrix defined in the basis of the eigenvectors at the reference point, $\mathbf{R}_0$, at which the diabatic and adiabatic bases are identical. The diabatic potential matrix $U_{A,B}(\mathbf{q})$ can be diagonalized by the rotation matrix $\mathbf{T}(\mathbf{q})$ (Atchity, Xantheas and Ruedenberg, 1991), to obtain the adiabatic energies $V_{1,2}$

$$\mathbf{T}(\mathbf{q}) = \begin{pmatrix} \cos[\theta(\mathbf{q})] & -\sin[\theta(\mathbf{q})] \\ \sin[\theta(\mathbf{q})] & \cos[\theta(\mathbf{q})] \end{pmatrix} \tag{30}$$

From Eq. (29), the rotation angle $\theta(\mathbf{q})$ is defined as (Atchity, Xantheas and Ruedenberg, 1991; Atchity and Ruedenberg, 1999; Yarkony, 2000; Koppel, Gronki and Mahapatra, 2001; Muller, Koppel and Cederbaum, 1993; Matsika and Yarkony, 2002):

$$\begin{aligned} \theta(\mathbf{q}) &= \frac{1}{2}\arctan\left[\frac{2H_{AB}(\mathbf{q})}{\Delta H(\mathbf{q})}\right] \\ &= \frac{1}{2}\arctan\left[\frac{2\boldsymbol{\kappa}^{AB}\cdot\mathbf{q}}{\boldsymbol{\delta\kappa}\cdot\mathbf{q}}\right] \\ &= \frac{1}{2}\arctan\left[\frac{y}{x}\right] \end{aligned} \tag{31}$$

where we have introduced the *scaled* coordinates x and y. Replacing these scaled coordinates by polar coordinates r and ϕ

$$\begin{aligned} r &= \sqrt{x^2 + y^2} \\ \phi &= \arctan\left[\frac{2y}{x}\right] \end{aligned} \tag{32}$$

yields the following expression for the rotation angle

$$\theta\left(H_{AB}, \Delta H\right) = \frac{\phi}{2} \tag{33}$$

Thus, a simple relationship exists between the polar angle (ϕ), which defines a rotation around the apex of the cone in the branching space on one hand, and the *mixing* angle (θ) for the diabatic states (φ_A, φ_B) on the other. Since the mixing depends *only* on the polar angle ϕ and *not* on the radius r, it is constant along any straight line with the origin at the apex of the double cone.

Because of this relationship, the adiabatic wave function changes sign upon a complete rotation around the apex of the cone in the branching space. This is demonstrated by comparing the wave functions at $\phi = \phi_0$ and $\phi = \phi_0 + 2\pi$. Substituting Eq. (33) into the rotation matrix (Eq. (30)) gives the following expression for the adiabatic states:

$$\psi_1 = \cos\left[\frac{\phi_0}{2}\right]\varphi_A - \sin\left[\frac{\phi_0}{2}\right]\varphi_B \quad (34)$$

and

$$\psi_2 = \sin\left[\frac{\phi_0}{2}\right]\varphi_A + \cos\left[\frac{\phi_0}{2}\right]\varphi_B \quad (35)$$

Next, we substitute $\phi = \phi_0 + 2\pi$ to show that the adiabatic wave function has indeed changed sign after the 360° rotation:

$$\begin{aligned}\psi_1 &= \cos\left[\frac{\phi_0 + 2\pi}{2}\right]\varphi_A \\ &\quad - \sin\left[\frac{\phi_0 + 2\pi}{2}\right]\varphi_B \\ &= \sin\left[\frac{\phi_0}{2}\right]\varphi_A - \cos\left[\frac{\phi_0}{2}\right]\varphi_B \\ &= -\psi_1 \end{aligned} \quad (36)$$

Because single-valuedness of the wave function is one of the basic postulates of quantum mechanics, this result implies that the conical intersection is a singularity. This singularity is a consequence of the separation between electronic and nuclear degrees of freedom, which is not valid near surface crossings. Thus, the singularity only exists for the adiabatic electronic wave function and is remedied by the nuclear wave function, so that the total molecular wave function is single-valued function. Furthermore, as illustrated schematically in Figure 3.4, rotation by 90° permutes the order of the electronic states, as is demonstrated by substituting $\theta + \pi$ into Eq. (33):

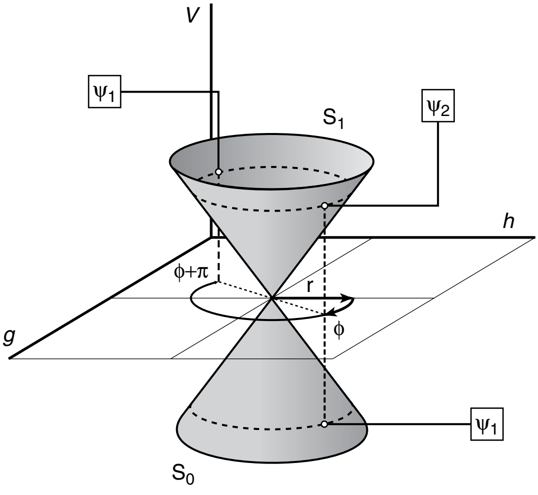

Fig. 3.4 In the branching space (x,y), points lying on a circle with a small radius r and centered at the apex of the double cone are *diabatically* related. If we select a point on one of the surfaces and move it to the opposite side of the circle ($\phi \to \phi + \pi$), its wave function becomes equal to the wave function of the other surface at the original position. If we complete the circle, the point is back at its original position, but its wave function has changed sign (not shown, see text).

$$\begin{aligned}\psi_1 &= \cos\left[\frac{\phi_0+\pi}{2}\right]\varphi_A \\ &\quad - \sin\left[\frac{\phi_0+\pi}{2}\right]\varphi_B \\ &= \sin\left[\frac{\phi_0}{2}\right]\varphi_A \\ &\quad - \cos\left[\frac{\phi_0}{2}\right]\varphi_B = -\psi_2 \qquad (37)\end{aligned}$$

As we will show in Section 3.2.3, this characteristic of the conical intersection can be used to detect when the seam is crossed in a classical molecular dynamics simulation of a photochemical reaction.

As we will discuss in the applications section, the conical intersection seam is the central mechanistic feature in a photochemical reaction. The conical intersection provides a funnel for efficient radiationless decay between electronic states (Figure 3.2). To illustrate the relationship between the surface crossing and photochemical reactivity, we can draw a parallel with the transition state in ground state chemistry. The transition state forms the bottleneck through which the reaction must pass on its way from reactants to products. A transition state separates the reactant and product energy minima along the reaction path. A conical intersection also forms a bottleneck that separates the excited state branch of the reaction path from the ground state branch. The crucial difference between conical intersections and transition states is that, although the transition state must connect the reactant minimum to a single product minimum via a single reaction path, an intersection is a spike on the ground state energy surface and thus connects the excited state reactant to two or more products on the ground state via several reaction paths.

3.2.3
Excited State Molecular Dynamics

To model the dynamics of a photochemical reaction, the ground- and excited state potential energy surfaces must be described accurately. As illustrated in Figure 3.1, the reaction starts in the excited state (S_1), but ends in the ground state (S_0). Therefore it is essential to model the radiationless transition between the excited and ground state surfaces in a manner that is consistent with a QM treatment.

If a sufficiently accurate description of the adiabatic Born–Oppenheimer potential energy surfaces is available, nuclear dynamics can be computed by numerically integrating either the time-dependent Schrödinger equation (Eq. (3)), or Newton's equations of motion. In the first case, quantum mechanics is used to follow the dynamics of nuclear wave packets ($\chi_j(\mathbf{R})$, Eq. (9)), evolving on electronic potential energy surfaces. Within this approach transitions between surfaces near or at conical intersections can be described correctly. A wave packet traveling through or near an intersection spawns new packets on both surfaces. The transfer probability is controlled by the nonadiabatic coupling elements. At the conical intersection, the coupling has its maximal strength, resulting in an efficient transfer of the complete wave packet to the lower surface.

A requirement for wave packet dynamics is that the relevant potential energy surfaces have been computed beforehand. As the computational costs associated with precomputing these surfaces increases rapidly with the number of coordinates in the system, wave packet dynamics is restricted to small isolated molecules. Alternatively, the multidimensional configuration space can be reduced to a lower dimensional subspace in which quantum dynamics is feasible (Burghardt, 2001). However, choosing suitable coordinates is not straightforward and requires averaging over the remaining degrees of freedom. Furthermore, the choice of the coordinates can strongly bias the outcome of a simulation. Therefore, for photobiological problems, the size and complexity of the systems involved severely limit the applicability of wave packet simulations.

Alternatively, Newton's equations of motion can be used to compute molecular dynamics trajectories. The underlying assumption in this case is to ignore the QM character of the nuclei. The nuclei are thus assumed to be classical particles moving on a single adiabatic potential energy surface. In terms of computational effort, classical molecular dynamics is orders of magnitude more efficient than wave packet dynamics, and is therefore routinely used for computing the time evolution of large biomolecular systems. Classical mechanics has the additional advantage that the potential energy surface can be computed on the fly. Forces are evaluated for the geometry at timestep t and used to compute the geometry at the next time step $t + \Delta t$. Thus, only at the configurations sampled by the classical trajectory, electronic structure calculations are required. For high-dimensional systems, for which computing potential energy surfaces beforehand is no longer possible, the on-the-fly strategy is the only way to perform a molecular dynamics simulation.

Classical mechanics provides a computationally cheap alternative to wave packet dynamics. However, because quantum effects are ignored, population transfer cannot occur and classical trajectories are restricted to a single potential energy surface. Therefore, in contrast to wave packet dynamics, radiationless transitions do not take place spontaneously. Instead, a binary decision to jump to a different electronic surface must be made at every timestep in a single trajectory. The criterion for switching between electronic states must result in a distribution of state populations over many trajectories that reflects the populations of a full quantum mechanical treatment. Furthermore, in contrast to wave packet dynamics, classical trajectories do not capture coherence effects. When a nuclear wave packet originally traveling on a single adiabatic potential energy surface encounters a crossing, it typically produces two subpackets, or offsprings, one on each surface (Thachuk, Ivanov and Wardlaw, 1998). After such spawning event, the new wave packets will still interfere with each other. However, as these subpackets travel on different surfaces and often have different initial velocities on their respective surfaces, they are well separated in phase space and it is highly unlikely that they will encounter each other again, especially in high-dimensional systems, such as solvated biomolecules. Therefore, the spawned wave packets are effectively incoherent (Thachuk, Ivanov and Wardlaw, 1998; Onuchic and Wolynes, 1998), and the classical trajectory approximation is justified in most situations.

3.2.3.1 **Surface Hopping**

The interest in understanding photochemical processes has prompted the development of methods for the treatment of nonadiabatic effects in classical molecular dynamics. Most, if not all of these methods are based on surface hopping: nuclei move on a single potential energy surface and nonadiabatic transitions are included by allowing the trajectory to hop from one surface to another. The size and complexity of biomolecular systems necessitates the use of computationally cheap surface-hopping algorithms. In this section we review two of the most widely used hopping procedures for excited state dynamics in biomolecules, diabatic hopping and Tully's fewest-switches surface hopping.

As formulated independently by Landau and Zener, in one dimension the probability of a transition between two electronic states, ψ_2 and ψ_1, is given by

$$P_{2\to 1} = \exp\left[-\frac{1}{4}\pi\xi\right] \tag{38}$$

The Massey parameter ξ is defined as (Desouter-Lecomte and Lorquet, 1979)

$$\xi = \frac{\Delta E}{\hbar\frac{\partial \mathbf{Q}}{\partial t}\cdot \mathbf{g}(\mathbf{Q})} \tag{39}$$

with $\mathbf{Q}$ representing a one-dimensional nuclear reaction coordinate, ΔE the energy gap between the adiabatic states, and

$$\mathbf{g}(\mathbf{Q}) = \langle\psi_1|\nabla_{\mathbf{Q}}\psi_2\rangle \tag{40}$$

After differentiation of ψ_2 with respect to t (via $\frac{\partial}{\partial t}\mathbf{Q}$), ξ can be rewritten as

$$\xi = \frac{\Delta E}{\hbar\left\langle\psi_1|\frac{\partial\psi_2}{\partial t}\right\rangle} \tag{41}$$

To decide when to undergo a transition to a different potential energy surface, $\langle\psi_1|\frac{\partial}{\partial t}\psi_2\rangle$ needs to be computed at every timestep of the simulation. For a small timestep Δt to integrate the classical equations of motion for the nuclei, we have at t:

$$\begin{aligned}\psi_1(t) &= \varphi_A\\ \psi_2(t) &= \varphi_B\end{aligned} \tag{42}$$

where the $\varphi_{A,B}$ are the diabatic electronic states. At $t+\Delta t$, the states will have become mixed due to the nonadiabatic coupling. In the limit of an infinitesimally small timestep, the change in the wave functions becomes

$$\begin{aligned}\psi_1(t+\Delta t) &= \varphi_A + \beta\varphi_B\\ \psi_2(t+\Delta t) &= -\beta\varphi_A + \varphi_B\end{aligned} \tag{43}$$

with β being a mixing coefficient. Numerical differentiation (finite differences) of the wave function gives

$$\frac{\partial}{\partial t}\psi_2 = -\beta\frac{\varphi_A}{\Delta t} \tag{44}$$

and yields the following expression:

$$\left\langle\psi_1|\frac{\partial}{\partial t}\psi_2\right\rangle \approx -\frac{\beta}{\Delta t} \tag{45}$$

Since

$$\langle\psi_1(t)|\psi_2(t+\Delta t)\rangle = -\beta \tag{46}$$

we can compute $\langle\psi_1(t)|\psi_2(t+\Delta t)\rangle$ as a numerical approximation for $\langle\psi_1|\frac{\partial}{\partial t}\psi_2\rangle$ in the Massey parameter (Eq. (41)).

With the energy gap ΔE and $\langle\psi_1(t)|\psi_2(t+\Delta t)\rangle$, we can use the Landau–Zener formula (Eq. (38)) to calculate at every timestep the probability of a transition to the other surface. In principle, the transition probabilities could be used to spawn off multiple trajectories. However, such spawning procedure leads to an enormous increase of computational effort. Instead, hops are usually restricted to situations where the transition probability approaches unity. This happens at the conical intersection seam, where $\Delta E \approx 0$ and $\| \langle \psi_1(t)|\psi_2(t+\Delta t)\rangle \| \approx 1$. The former is trivially true by the definition of an intersection, and the latter follows from Eq. (37): As shown in Figure 3.4, passing through the intersection during a single timestep (Δt) leads to geometries at t and $t+\Delta t$ that are opposite to each other with respect to the apex in the branching space. Therefore, crossing the seam is equivalent to rotating the wave function by half a circle (Figure 3.4):

$$\begin{aligned}
&\langle \psi_1(t)|\psi_2(t+\Delta t)\rangle \\
&\quad = \langle \psi_1(\phi)|\psi_2(\phi+\pi)\rangle \\
&\quad = -\langle \psi_1(\phi)|\psi_1(\phi)\rangle \\
&\quad = -\langle \psi_1(t)|\psi_1(t)\rangle \\
&\quad = -1
\end{aligned} \tag{47}$$

If hopping is allowed only at the conical intersection seam, a classical trajectory never leaves the diabatic surface, and energy and momentum are obviously conserved. In principle, this strict diabatic hopping criterium could lead to an underestimation of the population transfer probability, because a surface hop in regions with strong nonadiabatic coupling far from the intersection is prohibited. However, because of the high dimensionality of the seam, most trajectories can usually encounter such regions of high transfer probability. The Landau–Zener model is clearly an approximation, but helps one to keep a proper physical insight, which is crucial in understanding complex systems.

Tully and coworkers have proposed an alternative hopping scheme in which switches between adiabatic states are described by a stochastic process (Tully and Preston, 1972; Tully, 1990, 1991). The stochastic nature of the hops implies that only the average over many trajectories can be compared to observable quantities. At each timestep of the simulation, there is a probability that a trajectory propagating on one potential energy surface will hop to another surface. The hopping probability is derived from a *primary* electronic wave function $\Psi^e(t, \mathbf{r}, \mathbf{R})$. This wave function is expanded on the basis of the adiabatic states $\psi_j(\mathbf{r};\mathbf{R})$:

$$\Psi^e(t) = \sum_j c_j(t)\psi_j \tag{48}$$

where $c_i(t)$ are time-dependent expansion coefficients. Substituting the primary wave function into the time-dependent Schrödinger equation (Eq. (3)), multiplication on the left by $\psi_k(\mathbf{r};\mathbf{R})$ and integration over r yields

$$i\hbar\frac{\partial c_k(t)}{\partial t} = \sum_j c_j(t)\left\{H^e_{kj}(\mathbf{R}) - i\hbar\dot{\mathbf{R}}\cdot\mathbf{d}_{kj}\right\} \tag{49}$$

where the matrix elements are defined as

$$H^e_{kj}(\mathbf{R}) = \langle \psi_k(\mathbf{r};\mathbf{R})|H^e(\mathbf{R})|\psi_j(\mathbf{r};\mathbf{R})\rangle \tag{50}$$

The nonadiabatic coupling vectors

$$\mathbf{d}_{kj} = \langle \psi_k(\mathbf{r};\mathbf{R})|\nabla_{\mathbf{R}}\psi_j(\mathbf{r};\mathbf{R})\rangle \tag{51}$$

are obtained by applying the chain rule:

$$\begin{aligned}
&\langle \psi_k(\mathbf{r};\mathbf{R})|\frac{\partial}{\partial t}\psi_j(\mathbf{r};\mathbf{R})\rangle \\
&\quad = \langle \psi_k(\mathbf{r};\mathbf{R})|\nabla_{\mathbf{R}}\psi_j(\mathbf{r};\mathbf{R})\rangle\cdot\dot{\mathbf{R}}
\end{aligned} \tag{52}$$

Because the Hamiltonian H^e_{jk} is, by definition, diagonal in the adiabatic representation, population can only be transferred between the adiabatic states $\psi_j(\mathbf{r};\mathbf{R})$ and $\psi_k(\mathbf{r};\mathbf{R})$ by the nonadiabatic coupling between nuclei and electrons ($\dot{\mathbf{R}} \cdot \mathbf{d}_{kj}$).

For what follows, it is convenient to rewrite Eq. (49) in the equivalent density matrix notation. The elements of the density matrix $\boldsymbol{\rho}(t)$ are defined as

$$\rho_{kj}(t) = c_k(t)c_j^*(t) \tag{53}$$

In this notation Eq. (49) becomes

$$\begin{aligned} \mathrm{i}\hbar\dot{\rho}_{kj}(t) = \sum_l \rho_{lj}(t) \left[H^e_{kl} - \mathrm{i}\hbar\dot{\mathbf{R}} \cdot \mathbf{d}_{kl}\right] \\ - \rho_{kl}(t) \left[H^e_{lj} - \mathrm{i}\hbar\dot{\mathbf{R}} \cdot \mathbf{d}_{lj}\right] \end{aligned} \tag{54}$$

where we used the following properties on an orthonormal basis:

$$\mathbf{d}^*_{lj} = -\mathbf{d}_{jl} \tag{55}$$

and

$$\mathbf{d}_{jj} = 0 \tag{56}$$

The diagonal elements ρ_{jj} are the electronic state populations, and the off-diagonal elements ρ_{jk} describe coherence.

In Tully's approach, there is a distinction between the primary electronic wave function $\Psi^e(t)$ and the auxiliary adiabatic wave function ψ_n. The primary wave function defines all quantum observables, such as the populations of the adiabatic states, and is propagated in time along with the classical nuclear trajectory. The auxiliary adiabatic wave function defines the potential energy surface on which the nuclear trajectory evolves and is not propagated in time, but can make stochastic switches to other adiabatic states.

In order to maintain the correct quantum statistical distribution of states in an ensemble of surface-hopping trajectories, the probabilities of switching between different adiabatic states must reflect the populations of those adiabatic states in the primary electronic wave function. These populations or occupation probabilities are the diagonal elements of the density matrix $\boldsymbol{\rho}(t)$ (Eq. (53)). The rate at which the populations change along a nuclear trajectory is given by

$$\dot{\rho}_{kk} = \sum_{l \neq k} b_{kl} \tag{57}$$

where, from Eq. (54),

$$b_{kl} = 2\hbar^{-1}\mathrm{Im}\left[\rho^*_{kl}H^*_{kl}\right] - 2\mathrm{Re}\left[\rho^*_{kl}\dot{\mathbf{R}} \cdot \mathbf{d}_{kl}\right] \tag{58}$$

Since adiabatic basis states are used in the expansion of the primary electronic wave function $\Psi(t)$, the imaginary term in this expression is always zero.

During a very small time interval δt, the change in the population of state ψ_k is approximately $\dot{\rho}_{kk}\delta t$. If this state is only coupled to one other state, for example ψ_j, the change in population of state ψ_k is $\delta\rho_{kk} \approx \dot{\rho}_{kk}\delta t = b_{kj}\delta t$. Since no electrons are created or destroyed, the change in the population of state ψ_j, due to its coupling with ψ_k is $\delta\rho_{jj} = -\delta\rho_{kk} \approx -b_{kj}\delta t$. In what follows we assume that $b_{kj} > 0$, that is, the population of ψ_k increases due to the coupling, whereas that of ψ_j decreases. To maintain the correct quantum statistical distribution of adiabatic states in an ensemble of N_{ens} trajectories, the number of switches from ψ_j to ψ_k in the time interval δt must exceed the number of switches from ψ_k to ψ_j by $N_{\mathrm{ens}}b_{kj}\delta t$. Tully argued that in a simulation the total number of switches should be minimal (Tully, 1990). A minimum number of state switches is realized if $N_{\mathrm{ens}}b_{kj}\delta t$ switches from ψ_j to ψ_k occur, and none from ψ_k

to ψ_j. Since there are initially $N_{\text{ens}}\rho_{jj}$ trajectories in state ψ_j, the probability that one of these will switch to ψ_k during the time interval δt is

$$g_{jk} = \frac{N_{\text{ens}} b_{kj} \delta t}{N_{\text{ens}} \rho_{jj}} = \frac{b_{kj}}{\rho_{jj}} \delta t \tag{59}$$

and the probability of a switch from ψ_k to ψ_j should be zero. Therefore the fewest-switches algorithm computes the transition probabilities as

$$g_{jk} = \max\left[0, \frac{b_{kj}\Delta t}{\rho_{jj}}\right] \tag{60}$$

where Δt is the integration time step in a simulation. Being based on the rates of change of the populations, the fewest-switches algorithm ensures that surface hops only occur if the trajectory is in regions where the nonadiabatic coupling vectors $\mathbf{d}_{jk}$ are nonzero. Outside those regions, the nonadiabatic coupling vectors are negligible and the adiabatic state populations no longer change, that is, $\dot{\rho}_{kk} = 0$.

In order to conserve energy, the velocities of the nuclei are adjusted, whenever a state switch is made to an adiabatic state that has a different energy at the current nuclear configuration. Since nuclei can only induce transitions between electronic states if their velocity vectors overlap with the nonadiabatic coupling vectors ($\dot{\mathbf{R}} \cdot \mathbf{d}_{kj}$), Tully proposed to adjust the components of the nuclear velocities parallel to these coupling vectors (Tully, 1990). Thus the velocities of the nuclei are adjusted, as if the nuclei were subjected to a force in the direction of the nonadiabatic coupling vector, and the work done by that force was used to promote the system to another adiabatic state.

In a surface-hopping simulation, the new velocity of each nucleus $\dot{\mathbf{R}}^{i,\text{new}}$, after a state switch from ψ_k to ψ_j, is calculated as Fang and Hammes-Schiffer, 1999)

$$\dot{\mathbf{R}}^{i,\text{new}} = \dot{\mathbf{R}}^i - \frac{1}{M^i}\gamma_{kj}\mathbf{d}_{kj} \tag{61}$$

where M^i is the mass of nucleus i and $\mathbf{d}_{kj}$ the nonadiabatic coupling vector (Eq. (51)). The scaling factor γ_{kj} is defined as

$$\gamma_{kj} = \begin{cases} \dfrac{\beta_{kj} + \sqrt{\beta_{kj}^2 + 4\alpha_{kj}\left[V_k(\mathbf{R}) - V_j(\mathbf{R})\right]}}{2\alpha_{kj}} \\ \text{for} \quad \beta_{kj} < 0 \\ \dfrac{\beta_{kj} - \sqrt{\beta_{kj}^2 + 4\alpha_{kj}\left[V_k(\mathbf{R}) - V_j(\mathbf{R})\right]}}{2\alpha_{kj}} \\ \text{for} \quad \beta_{kj} \geq 0 \end{cases} \tag{62}$$

where $V_n(\mathbf{R})$ is the potential energy of state ψ_n at the current nuclear configuration, and we have introduced

$$\alpha_{kj} = \frac{1}{2M^i}\left|\mathbf{d}_{kj}\right|^2 \tag{63}$$

and

$$\beta_{kj} = \dot{\mathbf{R}}^i \cdot \mathbf{d}_{kj} \tag{64}$$

If there is not enough velocity (kinetic energy) in the direction of the nonadiabatic coupling vector to maintain energy conservation during a hop, the hop is classically forbidden. There are two generally accepted approaches for dealing with classically forbidden state switches. In the first approach, the hops are rejected, and the velocities parallel to the nonadiabatic coupling vector are reversed. If the system attempts to make the hop when there is not sufficient energy, it falls back as if it has encountered a too high barrier. In the other approach, the switch is also rejected, but now the velocities are not changed. The rationale for the second approach is that, if there is not enough energy for making a hop, nothing will happen and the system remains in the current state (Müller and Stock, 1997). Classically forbidden hops lead to discrepancies between the QM occupation probabilities of the adiabatic states and the fraction of trajectories found in those adiabatic states in the ensemble (Fang and Hammes-Schiffer, 1999).

Another source of error is the coherent propagation of the primary electronic wave function (Fang and Hammes-Schiffer, 1999; Thachuk, Ivanov and Wardlaw, 1998), which can lead to inconsistencies between quantum probabilities and ensemble averages. When a trajectory initially moving on state i passes a region of nonadiabatic coupling, the other adiabatic states become populated and the primary electronic wave function no longer exists as a single adiabatic state, but as a mixture of different states. After leaving the nonadiabatic region, the primary electronic wave function remains a superposition of adiabatic states, and is propagated along the trajectory as a mixed state in the rest of the simulation. This essentially means that, together with the real trajectory in state i, a number of virtual trajectories are evolved (Thachuk, Ivanov and Wardlaw, 1998). These virtual trajectories evolve on the adiabatic surfaces associated with the other adiabatic states that are populated in the primary wave function. Although they are on different surfaces, the virtual trajectories travel at the same velocity and in the same direction as the real trajectory and can interfere with it. As a consequence, coherence between the electronic wave functions of these virtual trajectories and the real trajectory is maintained. At a second passage through a region of nonadiabatic coupling, the interference of virtual trajectories with the real one influences the transition probabilities.

In reality, a wave packet quickly looses coherence with the wave packets on the other electronic surfaces after leaving a crossing region and will not interfere with the other wave packets when it encounters a next crossing (Thachuk, Ivanov and Wardlaw, 1998; Onuchic and Wolynes, 1998). This coherence loss is attributed to the rapid divergence in phase space of wave packets on different electronic surfaces. In an exact QM calculation of the system, a fully coherent propagation of the electronic wave function will not cause problems, since the nuclear wave functions quickly destroy any coherence between wave packets on different adiabatic surfaces. In surface hopping, on the other hand, nuclear wave functions are replaced by classical trajectories. Since a classical trajectory has no wave character, it can not exist on different adiabatic surfaces simultaneously. Propagating the primary electronic wave function along a single trajectory therefore introduces artificial coherence between this trajectory and virtual trajectories on other surfaces.

An *ad hoc* solution to the coherence problem in surface hopping is to reset the primary electronic wave function $\Psi^e(t)$ to the currently occupied auxiliary adiabatic state ψ_n (i.e. $c_j(t) = \delta_{jn}$) each time a trajectory leaves a hopping region (Fang and Hammes-Schiffer, 1999). The resetting introduces artificial decoherence in the coherent propagation of the primary electronic wave function, and thus washes out any interference with virtual trajectories. As a result, the ensemble averages are in agreement with the quantum mechanical probabilities.

In contrast to the diabatic hopping approach, hops are not restricted to occur at the conical intersection seam, but can occur wherever the nonadiabatic coupling is strong enough, even at nonzero energy gaps. In principle, population transfer is not underestimated by this method. Energy conservation, however, which is fulfilled in the diabatic hopping approach, is more difficult, and an *ad hoc* scaling of nuclear velocities is introduced. In terms of computational effort, the fewest-switches approach is more demanding as it requires calculation of the nonadiabatic coupling vector. Following Yarkony and coworkers, this vector can be related to the interstate coupling vector (Lengsfield, Saxe and Yarkony, 1984; Lengsfield and Yarkony, 1992)

$$\mathbf{h}_{ij}(\mathbf{R}) = \mathbf{C}_i^{\dagger} \frac{\partial \mathbf{H}}{\partial \mathbf{R}} \mathbf{C}_j \tag{65}$$

where $\mathbf{C}_i$ are the configuration interaction coefficients and $\mathbf{H}$ is the configuration interaction Hamiltonian matrix. With the interstate coupling, the nonadiabatic coupling vector can be computed as

$$\mathbf{d}_{ij}(\mathbf{R}) = \frac{\mathbf{h}_{ij}(\mathbf{R})}{V_i - V_j} + \sum_{a,b} \left(\Gamma_{ij}\right)_{ab} \left\langle \Phi_a \right| \frac{\partial \Phi_b}{\partial \mathbf{R}} \right\rangle \tag{66}$$

where Γ_{ij} is the one-electron transition density matrix and Φ denotes molecular orbitals. The computational cost associated with evaluating these gradients increases steeply with the number of nuclear degrees of freedom. These costly computations are avoided in the diabatic hopping method. Both approaches can yield mechanistic insights and, in the limit of a sufficiently high number of trajectories, can be used to estimate observable quantities, such as excited state lifetimes, and quantum yields.

3.2.3.2 **Excited State Quantum Chemistry**

In this subsection, we will highlight the challenges associated with excited state electronic structure calculations that are required for on-the-fly molecular dynamics simulations of photochemical reactions. A number of theoretical methods for the treatment of excited electronic states have been developed. Although highly accurate methods have become available over the past years, they are usually too time consuming for systems larger than a few atoms. Therefore, most of these methods are not yet applicable for molecular dynamics simulations of large biomolecular systems. Alternatively, simple force fields or existing *semiempirical* methods that are computationally efficient may be used, but their applicability is limited, unless properly reparameterized (Warshel, 1976; Toniolo *et al.*, 2004). Therefore, for on-the-fly molecular dynamics, a compromise between cost and accuracy has to be made.

A computationally feasible approach to describe excited state electronic structure is the equation of motion coupled cluster (EOM-CCSD) method and *ab initio* dynamics simulations of small isolated molecules have been performed at the EOM-CCSD level (Baeck and Martinez, 2003). However, EOM-CCSD can only work well if the underlying CCSD method provides a good description of the ground electronic state. This can cause problems when bonds are being broken or formed and the ground electronic state has a significant multiconfigurational character. The description of electronic states with a strong double excitation character also causes problems (Bartlett and Musial, 2007).

Time-dependent density functional theory (TD-DFT) offers a computationally very efficient approach to describe excited states and consequently has been used in excited state molecular dynamics simulations (Tavernelli, Rohrig and Rothlisberger, 2005). Similar to EOM-CCSD, TD-DFT suffers from the deficiencies of the underlying monoconfigurational DFT description of the ground state in regions of bond breaking and bond formation. Moreover, TD-DFT is known to encounter severe problems in describing valence states of molecules exhibiting extended π-systems, doubly excited states, and charge transfer excited states (Dreuw and Head-Gordeon, 2005).

The problems associated with methods that are based on a single reference configuration, such as EOM-CCSD and TD-DFT, demonstrate that for computing excited states, multiconfigurational methods are required to provide wave functions that are sufficiently flexible to describe bond rearrangements, electronic state mixing, and electronic reorganizations. In addition, to calculate molecular dynamics trajectories analytical energy gradients are necessary. Since the complete active-space self-consistent field (CASSCF) method fulfills these requirements, it has often been used in the framework of excited state dynamics simulations (Coe and Martinez, 2006, 2005; Weingart *et al.*, 2004; Boggio-Pasqua *et al.*, 2002; Sanchez-Galvez *et al.*, 2000). In CASSCF, a judicious set of occupied and virtual orbitals is chosen, the so-called active-space orbitals. In this active space, a full configuration interaction calculation is performed, while the other orbitals are kept doubly occupied or empty in all configurations. The active orbitals are optimized such that the electronic energy of the state considered is minimal. Alternatively, the average energy of the states under study is minimized (state-averaged approach), if state bias or root-flipping problems (Werner and Meyer, 1981) that occur near surface crossing regions (Worth and Robb, 2002), have to be avoided.

The CASSCF method captures to a large extent the so-called *static* electron correlation. However, owing to the necessary truncation of the active space, it does not recover *dynamic* electron correlation completely. Dynamic correlation is known to play a key role in the quantitative description of barrier heights and excitation energies. Thus, a higher-level treatment that includes dynamic electron correlation effects is desirable. Unfortunately, methods that resolve both static and dynamic correlations tend to be computationally too demanding and furthermore often lack the required analytical energy gradients. Because in most cases CASSCF describes the topology of the potential energy surfaces of the involved states sufficiently accurate, it is widely used for mechanistic studies of photochemical reactions (Lipkowitz and Boyd, 2000).

Dynamic correlation is accounted for in multireference perturbation theory approaches, such as Complete Active Space Perturbation Theory to Second Order (CASPT2) (Roos, 1999). CASPT2 provides a means to include dynamics correlation, while simultaneously describing static correlation. Recently, analytical CASPT2 energy gradients (Celani and Werner, 2003) have become available, which have opened the way for molecular dynamics simulations. Coe *et al.* have already used these gradients to perform an *ab initio* MD simulation on the excited state proton transfer reaction in methyl salicylate (Coe, Levine and Martinez, 2007). However, the computational cost of this method still prevents its use for larger biomolecular systems, in which the number of correlated electrons is too large.

Alternatively, dynamic electron correlation is included to some extent in the restricted active-space self-consistent field (RASSCF) method (Roos, 1999). RASSCF allows larger active spaces and thus provides more flexibility in the choice of the active orbitals than CASSCF (Boggio-Pasqua *et al.*, 2004). The larger active space is subdivided into three classes: orbitals with a limited number of valences, a fully active orbital set, and orbitals with a limited number of electrons. In addition to including dynamic electron correlation, RASSCF can be used as an approximation to CASSCF. By eliminating the unimportant configurations, the size of the configuration interaction problem can be greatly reduced in RASSCF compared to CASSCF without compromising accuracy. Because analytical gradients are available, the RASSCF method can be expected to be used in excited state dynamics simulations in the future. Unfortunately, however, it remains a difficult task even in RASSCF to select the correct orbital set for a given photochemical problem.

A promising alternative that circumvents the choice of active orbitals lies in the use of semiempirical configuration interaction methods. These methods offer a lower cost alternative, while still taking into account the correlation effects necessary to describe excited states. For example, the recently developed semiempirical OM2 method has been shown to describe accurately the well-known conical intersections of small molecules (Keal, Koslowski and Thiel, 2007). Thus, the use of new semiempirical methods, such as OM2, could hold great promise for nonadiabatic molecular dynamics simulations of very large molecular systems in the near future.

3.2.3.3 Mixed Quantum Classical Molecular Dynamics

MD computer simulations of biological systems have come of age. Since the first application of MD on a small protein in vacuum more than three decades ago (McCammon *et al.*, 1976), advances in computer power, algorithmic developments, and improvements in the accuracy of the used interaction functions have established MD as an important and predictive technique to study dynamic processes at atomic resolution (Berendsen, 2001). In the interaction functions, the so-called MM force field, simple chemical concepts are used to describe the potential energy of the system (Jensen, 2001):

$$
\begin{aligned}
V_{\mathrm{MM}} = & \sum_{i}^{N_{\mathrm{bonds}}} V_i^{\mathrm{bond}} + \sum_{j}^{N_{\mathrm{angles}}} V_j^{\mathrm{angle}} \\
& + \sum_{l}^{N_{\mathrm{torsions}}} V_l^{\mathrm{torsion}} \sum_{i}^{N_{\mathrm{MM}}} \sum_{j>i}^{N_{\mathrm{MM}}} V_{ij}^{\mathrm{Coul}} \\
& + \sum_{i}^{N_{\mathrm{MM}}} \sum_{j>i}^{N_{\mathrm{MM}}} V_{ij}^{\mathrm{LJ}} \qquad (67)
\end{aligned}
$$

where N_{MM} is the number of atoms in the system. Bonds and angles (V^{bond}, V^{angle}) are normally modeled by harmonic functions, and torsions by periodic functions (V^{torsion}). The pairwise electrostatic interaction between atoms with a partial charge (Q_i) is given by Coulomb's law:

$$V_{ij}^{\mathrm{Coul}} = \frac{e^2 Q_i Q_j}{4\pi\varepsilon_0 R_{ij}} \qquad (68)$$

where R_{ij} denotes the interatomic distance, e the unit charge, and ε_0 the dielectric constant. Short-range Pauli repulsion and long-range dispersion attraction are most often described by a single Lennard-Jones potential:

$$V_{ij}^{\text{LJ}} = \left(\frac{C_{12}^{ij}}{R_{ij}}\right)^{12} - \left(\frac{C_6^{ij}}{R_{ij}}\right)^{6} \qquad (69)$$

with C_{12}^{ij} and C_6^{ij} the repulsion and attraction parameters, respectively, which depend on the atom types involved.

Electrons are thus ignored completely in molecular mechanics force fields. Their influence is expressed by empirical parameters that are valid for the ground state of a given covalent structure. Therefore, processes that involve electronic rearrangements, such as photochemical reactions, cannot be described at the MM level. Instead, these processes require a quantum mechanics description. As we have discussed above, the computational effort associated with computing excited state electronic structure puts severe constraints on the size of the system that can be studied. To overcome this limitation for biological systems, which are typically orders of magnitude too large for a complete quantum chemical treatment, methods have been developed that treat a small part of the system at an appropriate QM level, while retaining the computationally cheaper force field (MM) for the remainder. This hybrid QM/MM strategy was originally introduced by Warshel and Levitt (1976) and is illustrated in Figure 3.5.

The justification for dividing a system into regions, which are described at different levels of theory, is the local character of chemical reactions in condensed phases. A distinction can usually be made between a "reaction center" with atoms that are directly involved in the reaction and a "spectator" region, in which the atoms do not directly participate in the reaction. For example, most reactions in solution involve the reactants and the first few solvation shells. The bulk solvent is hardly affected by the reaction, but can influence the reaction via long-range interactions. The same is true for most enzymes, in which the catalytic process is restricted to an active site. The rest of the protein provides an electrostatic background that can facilitate the reaction.

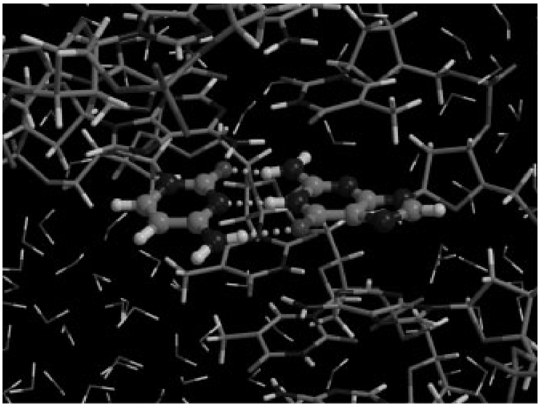

Fig. 3.5 The QM/MM partitioning scheme used in recent simulations of excited state proton transfer in a Watson–Crick Cytosine–Guanine base pair in DNA. The atoms of the QM subsystem are shown in ball-and-stick representation and MM atoms are shown as thin sticks. The CASSCF method was employed to model the electronic structure, while the Amber (Wang, Cieplak and Kollman, 2000) force field was used for the rest of the system.

The hybrid QM/MM Hamiltonian contains three classes of interactions: between atoms in the QM region, between atoms in the MM region, and between atoms in the QM and MM region

$$H^{\text{hybrid}} = H_{\text{QM}} + H_{\text{MM}} + H_{\text{QM/MM}} \tag{70}$$

The interactions within the QM and MM regions are relatively straightforward to describe, that is, at the QM and MM levels, respectively. The interaction between the two subsystems is more difficult to describe and several approaches have been proposed.

In the most simple approach, the QM subsystem is mechanically embedded in the MM system and kept in place by force field interactions, that is, bonds, angles, torsions and Lennard-Jones. With the exception of these interactions, the two systems are treated independently. Thus a quantum chemistry calculation is performed on an isolated QM subsystem, while a force field calculation is performed on the MM region. An improvement of the model is to use the isolated electronic wave function to derive partial atomic charges for the QM atoms and use these charges to compute the electrostatic QM/MM interactions with the MM atoms.

In the more popular electronic embedding scheme of Field and coworkers (Field, Bash and Karplus, 1990), the MM atoms enter the electronic Hamiltonian, as if they were QM nuclei:

$$\begin{aligned} H^{e}_{\text{QM/MM}} &= H^{e} + \sum_{i}^{n_e}\sum_{K}^{N_{\text{MM}}} \frac{Q_i}{4\pi\varepsilon_0 r_{iK}} \\ &= -\frac{\hbar^2}{2m_e}\sum_{i}^{n_e}\nabla_i^2 + \sum_{i}^{n_e}\sum_{j>i}^{n_e}\frac{e^2}{4\pi\varepsilon_0 r_{ij}} \\ &\quad - \sum_{i}^{n_e}\sum_{A}^{N_{\text{QM}}}\frac{e^2 Z_A}{4\pi\varepsilon_0 r_{iA}} + \sum_{i}^{n_e}\sum_{K}^{N_{\text{MM}}}\frac{e^2 Q_K}{4\pi\varepsilon_0 r_{iK}} \end{aligned} \tag{71}$$

where H^e is the original electronic Hamiltonian for the isolated QM system, defined in Eq. (5); n_e is the number of electrons, N_{QM} the number of QM nuclei, and N_{MM} the number of MM atoms; Z_A and Q_K are the nuclear and partial charges of QM nucleus A and MM atom K, respectively; and m_e denotes the electron mass. Because the MM atoms enter the Hamiltonian, the electronic wave function is polarized by the environment. Simultaneously, the electrons are exerting electrostatic forces on both QM nuclei and MM atoms. Problems may arise if the MM atoms near the QM region have high partial charges. In this case, the electrons are strongly attracted by such MM atoms, and the wave function can become overpolarized. Penetration of electron density in to the MM region is an artefact of ignoring the electrons of the MM atoms. A remedy for this spill-out effect is to use gaussian-shaped charge densities rather than point charges to represent partially charged MM atoms (Amara and Field, 2003).

Interactions between the nuclei in the QM region, and between QM nuclei and MM atoms are described by Coulomb's law:

$$\begin{aligned} H^{\text{nuc}}_{\text{QM/MM}} &= \sum_{A}^{N_{\text{QM}}}\sum_{B>A}^{N_{\text{QM}}}\frac{Z_A Z_B}{4\pi\varepsilon_0 R_{AB}} \\ &\quad + \sum_{A}^{N_{\text{QM}}}\sum_{K}^{N_{\text{MM}}}\frac{Z_A Q_K}{4\pi\varepsilon_0 R_{AK}} \end{aligned} \tag{72}$$

In addition to electrostatics, there are also Van der Waals interactions between the subsystems that are handled at the force field level (Eq. (69)), as if the QM nuclei were MM atoms. Similar to the previous model, bonded interactions, such as bonds, angles, and torsions involving both QM and MM atoms, are described by the respective force field functions.

If the QM and MM subsystems are connected by chemical bonds, care has to be taken when evaluating the QM wave function. Cutting the QM/MM bond creates one or more unpaired electrons in the QM subsystem. In reality, these electrons are paired with electrons belonging to the atoms on the MM side. A number of approaches to remedy the artefact of such open valences have been proposed. The most easy solution is to introduce a monovalent link atom at an appropriate position along the bond vector between the QM and MM atoms. Hydrogen is most often used, but there is no restriction on the type of the link atom and even complete fragments, such as methyl groups, can be used to cap the QM subsystem. The link atoms are present only in the QM calculation, and are invisible for the MM atoms. Alternative approaches for capping the QM region are the frozen orbital approach (Philipp and Friesner, 1999) and the generalized hybrid orbital method (Gao *et al.*, 1998).

The QM/MM scheme of Field and coworkers provides a conceptually intuitive way of including the effect of an environment on a chemical reaction. However, unless polarization is treated explicitly at the force field level, this QM/MM model is not internally consistent. In most force fields polarization is not explicitly included, but is implicitly accounted for via the parameters in the other terms, most notably, in the nonbonded interactions. Thus, although the MM region can induce polarization of the QM subsystem, the latter cannot back-polarize the MM region. A related problem arises from the use of standard Lennard-Jones and charge parameters, which implicitly contain polarization. When using these parameters without modification, there is both explicit and implicit polarization of the QM region. Therefore, the total polarization can be overestimated. To avoid such possible artefact, the nonbonded parameters in principle should be reparametrized for use in QM/MM simulations. However, this procedure requires the optimization of many parameters that also depend on the level of theory employed for the QM subsystem, and is most often skipped.

3.3
Applications

In this section, we review a number of representative applications of excited state dynamics simulations in the context of photobiology.

3.3.1
Photoreceptor Proteins

3.3.1.1 Retinal Proteins

Response to visible light in higher eukaryotes is mediated by the visual pigment Rhodopsin, which is a seven-helix transmembrane protein located in the photoreceptor rodd cells in the retina. Rhodopsin contains a 11-*cis* retinal molecule that is covalently linked to a lysine residue (Lys296) (Okada *et al.*, 2004), as shown in Figure 3.6). The initial step in the vision process is photoisomerization of this highly conjugated chromophore, which induces strain in the chromophore pocket. Relaxation of

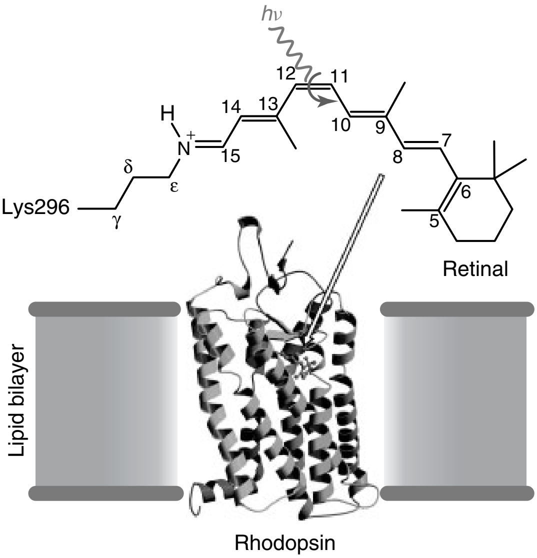

Fig. 3.6 Retinal (a) is covalently linked to a lysine residue in the membrane protein Bovine rhodopsin (b). Photon absorption by retinal induces a *cis*-to-*trans* isomerization of the conjugated $C_{11}{=}C_{12}$ bond that leads to an all-*trans* configuration. For clarity, the membrane is shown only schematically.

this strain is accompanied by conformational changes that ultimately trigger a signal transduction cascade in the cell.

In archaea, all-*trans* retinal isomerization in bacteriorhodopsin is the initial step in converting solar energy into chemical energy. For instance, in *Halobacterium Salinarium*, the conformational strain induced by photoisomerization is used to pump protons across the cell membrane against a concentration (pH) gradient. By controlling the proton flow back into the cell through other membrane proteins, energy is released that is subsequently used to synthesize high-energy compounds. As these examples show, photoisomerization of conjugated chromophores in proteins is a commonly employed theme in nature to allow organisms to respond to and use visible light.

Isolated retinal can undergo photoisomerization around any of its conjugated bonds (Koyama *et al.*, 1991). In contrast, when retinal is embedded in rhodopsin, photoisomerization exclusively takes place around the $C_{11}{=}C_{12}$ bond to form the all-*trans* photoproduct (Figure 3.6) (Mathies and Lugtenburg, 2000). In bacteriorhodopsin, which has a different protein cavity and retinal configuration, photon absorption induces an isomerization from an all-*trans* to 13-*cis* conformation (Tittor and Oesterhelt, 1990). These observations suggest that the protein environment is crucial for photoactivation. However, despite a wealth of structural and spectroscopic data, there is no consensus on how the protein matrix modulates the photochemical properties of retinal to ensure both high efficiency and bond selectivity of the isomerization process.

More than three decades ago, Warshel made the first serious attempt to address this issue. Using the empirical Quantum mechanical Consistent Force Field Method for π electron systems (QCFF/PI) to describe the excited state and a simple restraint on the retinal atoms to model the steric interaction with the protein matrix, he demonstrated that a space-saving photoisomerization reaction occurs, requiring minor rearrangements in the retinal cavity (Warshel, 1976). In a so-called bicycle-pedal mechanism, two double bonds rotate concertedly in opposite directions, such that the overall conformational change is minimized. Since then, Warshel and his coworkers have successively improved the model, employing better structural data and new computational developments (Warshel and Barboy, 1982; Warshel, Chu and Hwang, 1991; Warshel and Chu, 2001). In their most recent work, Warshel and Chu have used an advanced QM/MM strategy, with improved QCFF/PI parameters for the excited state of the chromophore, and polarization of the MM environment, to perform multiple excited state simulations of the photoisomerization event based on a high-resolution X-ray structure of bacteriorhodopsin (Luecke *et al.*, 1999).

The first *ab initio* QM/MM simulation of retinal photoisomerization in the protein environment of bacteriorhodopsin has been reported by Schulten and coworkers in 2003 (Hayashi, Tajkhorshid and Schulten, 2003). A small part of the all-*trans* retinal consisting of eight heavy atoms (C_{11} to C_ε, Figure 3.6), was described at the CASSCF level of theory. For this truncated chromophore, a full active space was used that comprised six electrons in six orbitals. The rest of the chromophore and the protein were modeled with the Amber94 force field. With this model, multiple simulations were carried out that have yielded a consistent picture of the dynamic events that follow photo absorption in the protein.

Very recently, Olivucci and coworkers, in a tremendous computational effort, have performed a QM/MM simulation of the excited state dynamics of retinal in bovine rhodopsin (Frutos *et al.*, 2007). Retinal was modeled at the CASSCF/6-31G* level, with an active space that included a complete π-system of 12 electrons in 12 orbitals. To compensate for the lack of dynamic electron correlation in the CASSCF formalism, the forces were scaled to match the forces evaluated at the CASPT2 level of theory. Surprisingly, a single linear scaling factor sufficed along the complete minimum-energy isomerization pathway. The rest of the protein was described by a modified version of the Amber force field, in which charges on the atoms in the vicinity of the retinal were reparameterized to reproduce the measured excitation wavelength. With this QM/MM model of rhodopsin, a 140 fs trajectory was computed that provided an atomistic description of the electronic and geometrical changes that prepare the system for decay to the ground state.

The simulation studies by Warshel, Schulten, and Olivucci have provided detailed atomistic insights into the sequence of events initiated by photon absorption and have established the role of the protein environment on the isomerization process. Immediately after excitation to S_1, bond-length relaxation takes place that occurs on a time scale of about 10 fs. This initial relaxation phase is followed by a space-saving isomerization process that includes the bicycle-pedal motion. Selectivity is achieved by steric interactions that hinder rotations around other bonds. Furthermore, in the protein environment the chromophore is rapidly channeled toward the seam region. The momentum acquired in the excited state ensures a high isomerization quantum yield after decay (Hayashi, Tajkhorshid and Schulten, 2003; Warshel and Chu, 2001).

Interestingly, recent work by Weingart on an isolated retinal chromophore model has shown that the isomerization quantum yield depends critically on the initial retinal configuration (Weingart, 2007). In the resting state of bovine rhodopsin (Okada *et al.*, 2004) the chromophore is distorted from planarity. By taking this nonplanar retinal structure as a starting point for a series of excited state CASSCF dynamics simulations, Weingart found that in 70% of the trajectories, the all-*trans* photoproduct was formed, in close agreement with the experimentally observed quantum yield. This result suggests that selectivity might originate predominantly from strain imposed by steric interactions between the chromophore and the protein *before* photon absorption.

3.3.1.2 **Photoactive Yellow Protein**

Photoactive yellow protein (PYP) is believed to be the primary photoreceptor for the photo avoidance response of the salt-tolerant bacterium *Halorhodospira halophila* (Hellingwerf, Hendriks and Gensch, 2003). PYP contains a deprotonated 4-hydroxy-cinnamic acid (or *p*-coumaric acid, pca) chromophore linked covalently to the γ-sulphur of Cys69 via a thioester bond (green, Figure 3.7). Upon absorbing a blue-light photon, PYP enters a fully reversible photocycle involving several intermediates on timescales ranging from a few hundred femtoseconds to seconds (Hellingwerf, Hendriks and Gensch, 2003).

By means of QM/MM MD simulations at the CASSCF(6,6)/3-21G level in combination with the Gromos96 force field, the detailed sequence of structural changes that follow photon absorption has been revealed (Groenhof *et al.*, 2004). The crucial step is a light-induced trans-to-cis isomerization of the covalently

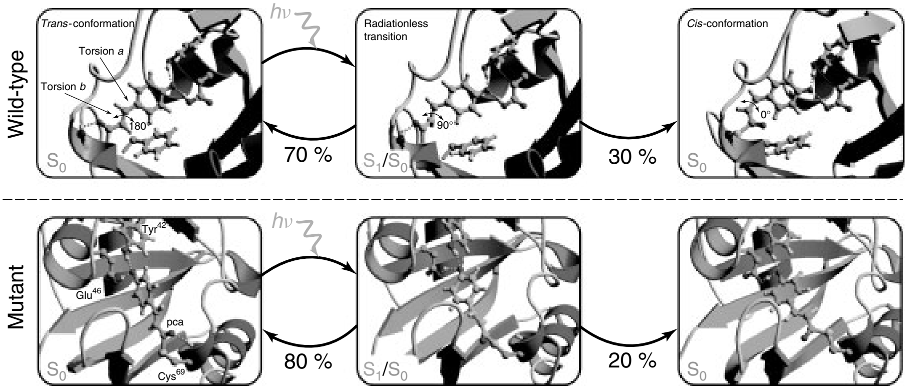

Fig. 3.7 Photochemistry of photoactive yellow protein. Snapshots from excited state trajectories of the (a) wild type , and (b) mutant, showing the chromophore (pca) in the active site pocket. The first snapshot is at the excitation. The second shows the configuration at the radiationless transition from S_1 to S_0. The third snapshot shows the photoproduct. In the wild type, isomerization takes place around the double bond (torsion *b*), whereas in the mutant, a single bond (torsion *a*) rotates. In both systems, the carbonyl oxygen of the thioester linkage flips, causing the break of the hydrogen bond to the backbone of Cys69.

bound chromophore (Figure 3.7). In simulations of an isolated chromophore, no double bond photoisomerization was observed, demonstrating that the protein environment is essential for the photoreaction. Furthermore, mutating one of the key amino acids in the chromophore cavity altered the photochemical reaction and led to a different photoproduct in the simulations (Groenhof *et al.*, 2008). These observations, which are difficult to predict by other means, demonstrate that the photoreactivity of the chromophore is extremely sensitive to the surrounding protein environment.

Figure 3.7 shows the primary events after photoexcitation of the wild-type protein. The chromophore rapidly decays to the ground state via a 90° rotation of the double bond (torsion *b*, Figure 3.7). Upon returning to the ground state the chromophore relaxes back either to the original *trans* conformation (180°), or it continues isomerizing to a *cis* conformation (0°). In the latter case, the relaxation also involves a flip of the thioester linkage, in which the carbonyl group rotates 180°. During this rotation, the hydrogen bond between the carbonyl oxygen and the Cys69 backbone amino group is broken (Figure 3.7). In total, 14 MD simulations were carried out, initiated from different snapshots from a classical ground state trajectory. The fluorescence lifetime (200 fs) and isomerization quantum yield (30%) in the simulations agree well with experiments (400 fs (Mataga *et al.*, 2000) and 35% (Hellingwerf, Hendriks and Gensch, 2003), respectively). Furthermore, it was found that in the protein environment the isomerization is enhanced by a preferential electrostatic stabilization of the chromophore's excited state by the guanidinium group of Arg52, located just above the negatively charged chromophore ring.

PYP mutants in which the arginine has been replaced by an uncharged amino acid can still enter the photocycle, albeit with a lower rate and quantum yield (Changenet-Barret *et al.*, 2007; Takeshita *et al.*, 2002). These findings indicate that Arg52 is important but not essential for photoactivation. To elucidate the role of this arginine in the activation process in more detail, excited state dynamics simulations have been performed on the X-ray structure of the Arg52Gln mutant of PYP (Shimizu *et al.*, 2006). In the mutant, the primary events after photoexcitation are different from those in the wild type (Groenhof *et al.*, 2008). First, the predominant excited state reaction in the mutant involves isomerization of a single bond in the chromophore (torsion *a*, Figure 3.7), rather than the double bond (torsion *b*). Although single bond isomerization does not result in the formation of the *cis* chromophore, a 180° flip of the thioester group and a rupture of the hydrogen bond to Cys69 was observed with a 20% quantum yield (Figure 3.7). Together with the experimental observation that the mutant has a photoactivation quantum yield of about 21% (Takeshita *et al.*, 2002), this suggests that the key step to enter the photocycle is the oxygen flip rather than the double bond isomerization.

To summarize, the simulations are consistent with experimental observations and have provide detailed structural and dynamic information at a resolution well beyond that achievable by other means. From the simulations, key amino acids have been identified and the mechanism by which they control the primary events in the photocycle of PYP. These are (i) photoisomerization around a double or a single bond and (ii) the break of a hydrogen bond between the chromophore and the protein backbone. These events trigger a proton transfer from the protein to the chromophore, which ultimately leads to the signaling state of PYP.

3.3.2
Fluoroproteins

3.3.2.1 Green Fluorescent Protein

The discovery of the green fluorescent protein (GFP) in the early 1960s had a pronounced impact on fluorescence microscopy and ultimately heralded a new era in the bio- and nanosciences. This success is based on the applicability of GFP and its variants as genetically encodable fusion tags that enable to monitor protein localizations and dynamics in live cells (Lippincott-Schwartz, Altan-Bonnet and Patterson, 2003; Miyawaki, Sawano and Kogure, 2003; Tsien, 1998). Recently, a new class of green fluorescent protein (GFP)-like proteins has been discovered, which can be reversibly photoswitched between a fluorescent (*on*) and a nonfluorescent (*off*) state (Ando, Mizuno and Miyawaki, 2004; Chudakov *et al.*, 2003a; Lukyanov *et al.*, 2000; Habuchi *et al.*, 2005; Wilmann *et al.*, 2006; Dedecker *et al.*, 2006). As the reversible photoswitching of photochromic organic molecules such as fulgides or diarylethenes is usually not accompanied by fluorescence (Crano and Guglielmetti, 1999), this switching reversibility is a very remarkable feature that may allow new applications. For example, the reversible photoswitching, also known as *kindling*, may provide nanoscale resolution in far-field fluorescence optical microscopy much below the diffraction limit (Hell, Jakobs and Kastrup, 2003; Hell, 2003a; Hell, Dyba and Jakobs, 2004; Hofmann *et al.*, 2005; Geisler *et al.*, 2007). Since fluorescence can be sensitively read out from a bulky crystal, the prospect of erasable three-dimensional data storage is equally intriguing (Sauer, 2005). Below, we will detail the excited state molecular dynamics of the GFP-like protein asFP595 as a prototype for a reversibly switchable fluorescent protein.

The GFP chromophore has been studied extensively, both by means of experimental and theoretical methods (Remington, 2006; Helms, 2002; Zimmer, 2002). Figure 3.8a shows the GFP protein with its β-barrel fold (Yang, Moss and Phillips, 1996; Ormö, Cubitt *et al.*, 1996). The chromophoric moiety is buried in the center of the protein structure. The *p*-hydroxybenzylidene-imidazolinone chromophore, shown in Figure 3.8b, is formed upon maturation of the protein in an autocatalytic cyclization–oxidation reaction of a tripeptide (Ser-Tyr-Gly in the *Aequorea victoria* wild type). The chromophore is embedded in the protein matrix and is surrounded by ionizable residues. Notably, this polar environment does not significantly change the absorbance of the chromophore as compared to the solution or the gas phase. The wild-type GFP exhibits two main absorption bands at 398 and 477 nm, which were assigned to the neutral and anionic chromophore species, respectively (Schellenberg *et al.*, 2001; He, Bell and Tonge, 2002; Bell *et al.*, 2003). However, the former absorption could also be due to the cationic chromophore (Voityuk, Michel-Beyerle and Rosch, 1998a; Dong, Solntsev and Tolbert, 2006). The GFP protein fluoresces with a high quantum yield of about 80% and a fluorescence

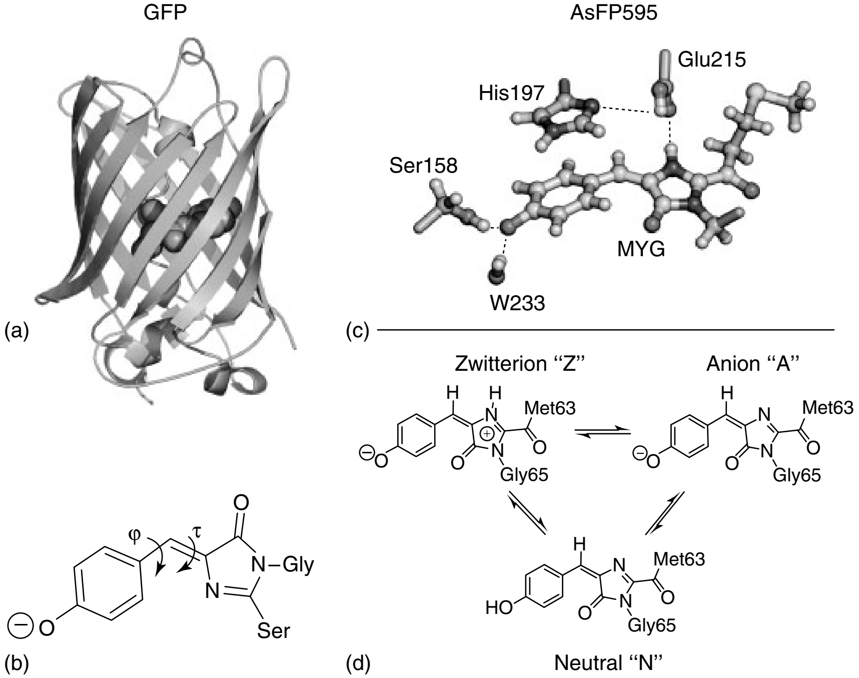

Fig. 3.8 Fluorescent proteins GFP and asFP595. (a) X-ray structure of GFP (PDB entry 1EMB), Ref. Brejc *et al.* (1997), with the β-barrel fold shown in green and the chromophore (in the center of the protein) in orange spheres. (b) Schematic drawing of the anionic form of the *cis* *p*-hydroxybenzylidene-imidazolinone chromophore. (c) The asFP595 chromophore (MYG) in the *trans* conformation and adjacent amino acid side chains. Dashed lines represent hydrogen bonds between MYG and Ser158, Glu215, and a crystallographic water molecule (W233), respectively. His197 is π-stacked to the MYG phenoxy-moiety and forms a hydrogen bond to Glu215. The carbon skeleton of the quantum mechanical (QM) subsystem is shown in cyan, and the carbon atoms modeled by molecular mechanics (MM) are shown in orange. (d) Schematic drawings of the important MYG protonation states. (Please find a color version of this figure on the color plates.)

lifetime of about 3 ns (Tsien, 1998); Striker *et al.*, 1999). Recent time-resolved Fourier transform infrared (tr-FTIR) measurements (Stoner-Ma *et al.*, 2005, 2006) together with classical molecular dynamics simulations (Lill and Helms, 2002) and static *ab initio* calculations (Vendrell *et al.*, 2006) suggest that in the protein, the fluorescent anionic chromophore species is formed out of the neutral species in a three-step excited state proton transfer (ESPT) reaction that involves a serine, a water molecule, and a glutamic acid residue that form a proton wire in the chromophore pocket (Agmon, 2005; Leiderman, Huppert and Agmon, 2006).

In contrast to the protein, model chromophores in solution do not feature fluorescence. Instead, after excitation, fast ground state recovery on the picoseconds timescale is observed (Webber, Litvinenko and Meech, 2001; Litvinenko, Webber and Meech, 2001).

Static calculations of vertical excitation energies and minimum-energy conical intersections and reaction paths have been carried out using *ab initio* (Martin, Negri and Olivucci, 2004; Altoe *et al.*, 2005; Sinicropi *et al.*, 2005), density functional (Marques *et al.*, 2003; Laino, Nifosi and Tozzini, 2004; Patnaik *et al.*, 2007), and *semiempirical* methods (Voityuk, Michel-Beyerle and Rosch, 1998b; Weber *et al.*, 1999; Voityuk *et al.*, 2001; Toniolo, Granucci and Martinez, 2003). The computational efficiency of *semiempirical* methods, in conjunction with a careful parameterization against *ab initio* data, enabled Toniolo *et al.* to carry out multiple-spawning nonadiabatic molecular dynamics simulations of the neutral chromophore, both in the gas phase and microsolvated in a small water droplet (Toniolo *et al.*, 2004; Martinez, 2006; Levine and Martinez, 2007). The simulations showed that the aqueous environment alters the topology of the involved electronic states such that the minimum energy S_0/S_1 conical intersection becomes the global minimum on S_1, thus accounting for the observed ultrafast ground state recovery. In addition, the authors predict a space-saving hula-twist *cis-trans* isomerization of the chromophore, that is, a simultaneous rotation around the ring-bridging torsions ψ and τ in Figure 3.8b. MD simulations based on a simple MM forcefield description of the excited state potential energy surface suggest a similar chromophore isomerization mechanism in the E2-GFP mutant (Nifosi and Tozzini, 2006).

3.3.2.2 **asFP595**

The fluoroprotein asFP595, a GFP-like protein that is expressed in the sea anemone *Anemonia sulcata*, converts absorbed green light of $\lambda_{\text{abs,max}} =$ 568 nm into a red fluorescence emission at $\lambda_{\text{fl,max}} = 595$ nm (Chudakov *et al.*, 2003a,b; Lukyanov *et al.*, 2000; Andresen *et al.*, 2005). The fluorescence quantum yield of asFP595 is very low (Lukyanov *et al.*, 2000, 2005), and subpicosecond ground state recovery without any structural damage is usually observed (Schüttrigkeit *et al.*, 2006). These photochemical properties suggest that the fluoroprotein might function as an efficient and stable sun blocker that protects the anemone against photodamage (Salih *et al.*, 2000). Remarkably, and in contrast to most other fluorescent proteins known so far, fluorescence of asFP595 can be reversibly and selectively switched *on* and *off* in response to irradiation of a particular wavelength (Andresen *et al.*, 2005). Analysis of the X-ray crystal structures (Andresen *et al.*, 2005) showed that this *on-off* switching of the fluorescence is accompanied by a *trans-cis* isomerization of the protein's chromophore, a 2-acetyl-5-(*p*-hydroxybenzylidene)-imidazolinone moiety formed autocatalytically from a Met-Tyr-Gly (MYG) tripeptide (Figure 3.8c).

The reversible photoswitching of asFP595 may allow fundamentally new applications. For example, as recently demonstrated by Hofmann and coworkers, asFP595 can provide nanoscale resolution in optical microscopy much below the diffraction limit (Hofmann *et al.*, 2005). Today, however, the photochromic properties of asFP595 still leave much room for improvement. To systematically exploit the potential of such switchable proteins and to enable a rational improvement of their properties, a detailed molecular understanding of the switching process is mandatory. Theoretical methods, in particular excited state molecular dynamics simulations, can yield the required atomistic insights into the photoswitching mechanism. Static quantum chemical calculations were carried out to predict vertical excitation energies of asFP595 model chromophores in the gas phase and in a continuum dielectric model (Amat *et al.*, 2006; Nemukhin, Topol and Burt, 2006) as well as in a minimal protein environment, the latter by means of QM/MM calculations (Grigorenko *et al.*, 2006).

In a two-step approach, QM/MM simulations were used to study the photoswitching mechanism of asFP595. First, because the positions of the protons were not resolved in the available X-ray structures, the aim was to locate the protons in the chromophore binding pocket. To this end, simulated UV/Vis spectra were compared to the available experimental data. This comparison was backed up by continuum electrostatics calculations, and enabled to unambiguously determine the protonation of the *on* and *off* states of asFP595 (Schäfer *et al.*, 2007). These calculations ascertained that the neutral chromophore form is dominant in the *cis* conformation (*on*-state), whereas the zwitterionic and the anionic chromophores predominate in the *trans* conformation (*off*-state) (Figure 3.8d). These results reveal that the photoinduced *trans–cis* isomerization of the chromophore is accompanied by proton transfer events. As shown in Figure 3.8c, these proton transfers are mediated by the ionizable residues Ser158, His197, and Glu215 in the chromophore pocket. His197 is connected through an extended hydrogen bonding network to the exterior interface between the protein and the surrounding solution (not shown in Figure 3.8).

In the second step of the study, nonadiabatic QM/MM molecular dynamics simulations were carried out to elucidate the photochemistry of each of the previously identified protonation states of the chromophore (Schäfer *et al.*, 2008). In these simulations, energies and gradients on the ground- and excited state were calculated on the fly at the CASSCF(6,6)/3-21G level of theory. Prior to the dynamics simulations, minima and conical intersections were characterized in vacuo at the RASSCF(18,7+4+5)[2,2]/6-31G* level of theory, that is, including all π-orbitals of the chromophore in the active space. The final six electron, six orbital active space used in the QM/MM dynamics simulations was selected from these RASSCF calculations such as to enable the simultaneous description of the ground state and the first excited state.

The quantum yields and excited state lifetimes from the QM/MM dynamics simulations agree well with experimental measurements (Schüttrigkeit *et al.*, 2006). In addition, the simulations enable to predict the structures of the hitherto unknown photochemical intermediates and the irreversibly fluorescent state. Furthermore, the simulations revealed how the proton distribution in the active site of the asFP595 controls the photochemical conversion pathways of the chromophore in the protein matrix. The suggested mechanism in Figure 3.9 shows that changes in the protonation state of the chromophore and some proximal amino acids lead to different photochemical states, which all turn out to be essential for the photoswitching mechanism. These photochemical

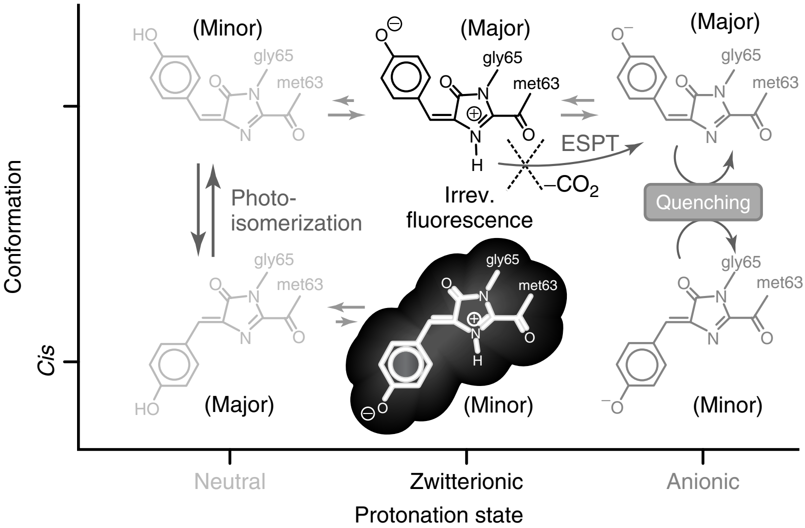

Fig. 3.9 Scheme of the reversible photoswitching mechanism of asFP595. The fluorescent state Z_{cis} is highlighted. The green arrows indicate ground state equilibria, whereas the red arrows indicate excited state processes. The major protonation states are the zwitterionic and the anionic chromophores in the *trans* conformation, and the neutral chromophore in the *cis* conformation, as indicated in the square brackets. (Please find a colour version of this figure on the color plates.)

states are (i) a neutral chromophore, which can photoisomerize back and forth between the *trans* and *cis* configuration on a subpicosecond timescale; (ii) an anionic chromophore, which rapidly undergoes radiationless decay after excitation in both the *cis* and *trans* configuration; and (iii) a long-lived and therefore putatively fluorescent zwitterionic chromophore. The *trans* zwitterion can rapidly return to the ground state through proton transfer to the neighboring Glu215 (Figure 3.8c). This alternative de-excitation pathway is not accessible for the *cis* isomer, thus explaining why only the *cis* form fluoresces. The overall stability (and thus the relative population) of the different protonation states is controlled by the isomeric state of the chromophore.

On the basis of the simulations, it was proposed that radiation-induced decarboxylation of the glutamic acid Glu215 could block the proton transfer pathways that enable the deactivation of both the *cis* and *trans* zwitterions and thus leads to irreversible fluorescence (Figure 3.9). Recent experiments on the structurally similar protein Dronpa (Habuchi *et al.*, 2006; Fron *et al.*, 2007) also provide strong support for the proposed protonation/deprotonation mechanism. The similarity between the chromophores in a variety of fluoroproteins suggests that during molecular evolution, the (*p*-hydroxybenzylidene)imidazolinone chromophoric moiety served as a template and that the photochromic properties – and thus the function – was fine-tuned by the protein environment.

3.3.3
Photochemistry of DNA

Deoxyribonucleic acid (DNA) carries the genetic information of all cellular forms of life. DNA forms a double helix, in which the nucleoside bases of the single strands are stacked upon each other, forming strong hydrogen bonds with the bases in the complementary strand (Watson–Crick configuration). Owing to the high absorbance of the bases in the harmful ultraviolet (UV) region of the spectrum, DNA is vulnerable to photochemical damage that can ultimately lead to cancer. To study the effects of UV radiation on DNA, molecular dynamics simulations have been performed of the ultrafast processes that follow photon absorption.

The fastest photochemical reaction involving base pairs in opposite strands was found to be proton transfer (Abo-Riziq *et al.*, 2005; Schwalb and Temps, 2007). Figure 3.10 shows the sequence of events that follows excitation of a cytosine–guanine base pair in DNA. According to the QM/MM simulations at the CASSCF(8,8)/3-21G level of theory (see Figure 3.5, for details on the QM/MM partitioning), photon absorption induces transfer of a proton from the guanine to the cytosine (Groenhof *et al.*, 2007). This proton transfer enhances ultrafast decay of the excited state. After the radiationless transition to the ground state, the original Watson–Crick configuration is quickly restored. The entire photochemical reaction is completed within 200 fs. The absorbed energy is converted into heat and therefore does not lead to structural damage. These findings have been recently corroborated by experimental observations (Schwalb and Temps, 2007; Miannay *et al.*, 2007). The existence of an ultrafast deactivation channel involving the observed proton transfer had been proposed previously by Sobolewski and Domcke on the basis of static high level *ab initio* computations (Sobolewski and

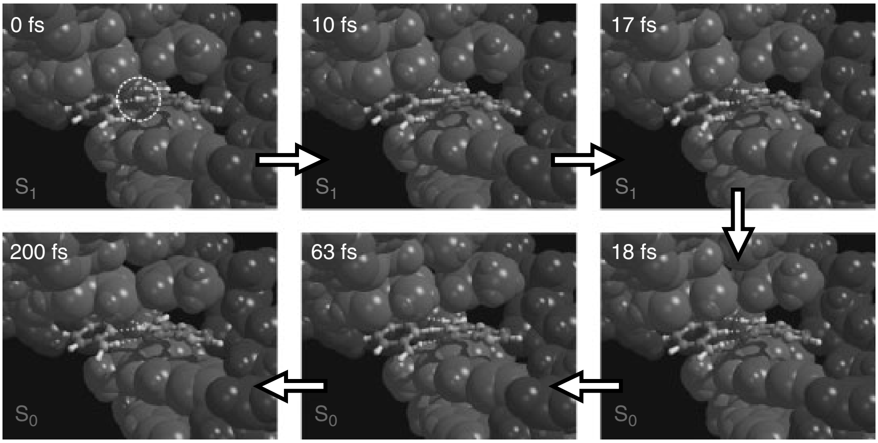

Fig. 3.10 Excited state dynamics of a Watson–Crick cytosine–guanine base pair after light absorption. The frames show the light-induced transfer of the central proton from a guanine to a cytosine. After this transfer, a radiationless transition to the ground state takes place (18 fs). The proton then rapidly returns to the guanine, restoring the Watson–Crick base pair (200 fs).

Domcke, 2004; Sobolewski, Domcke and Hättig, 2005).

Radiationless decay occurs along an extended hyperline that lies parallel to the proton transfer coordinate in the vicinity of the excited state minimum, as shown in Figure 3.11. The gradient difference vector and the derivative coupling vector (Eq. (27)) are colinear and thus span a one-dimensional branching space. The seam is displaced from the proton transfer pathway along a skeletal deformation of the bases. Decay can thus occur anywhere along the single proton transfer coordinate, accounting for the remarkably short lifetime of the excited base pair (Abo-Riziq *et al.*, 2005; Schwalb and Temps, 2007). In vacuum, decay occurs after a complete proton transfer, whereas in DNA decay can also occur much earlier. The origin of this effect lies in the temporal electrostatic stabilization of the dipole in the excited state by the solvated DNA environment. The seam topology also accounts for the recrossings (partial diabatic trapping) that were observed in the MD simulations. As schematically shown in Figure 3.11, vibrational motions perpendicular to the reaction coordinate repeatedly drive the system in and out of the conical intersection, until the base pair eventually returns to its stable ground state minimum.

3.3.4
Summary and Outlook

Understanding light-driven processes is one of the major goal of the bio- and nanosciences. The underlying molecular mechanisms are typically governed by subpicosecond atomic motions. Mechanisms on such ultrafast timescales are very challenging to probe by experiment. Here, molecular dynamics simulations have become an invaluable tool to understand such processes in atomic detail.

In this article, we have reviewed the state of the art in excited state molecular dynamics simulations. In the

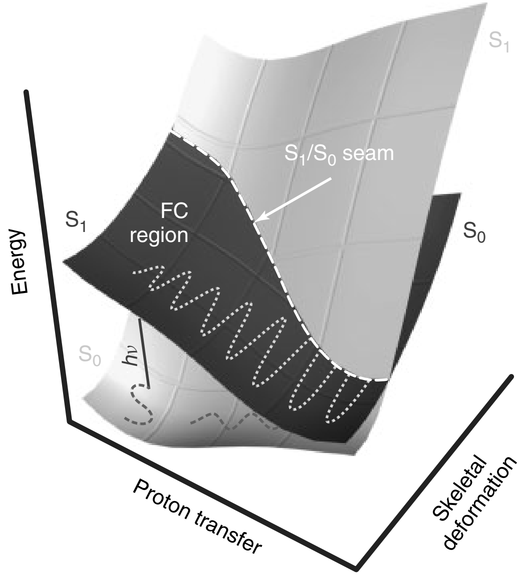

Fig. 3.11 Artist impression of the potential energy surfaces of the excited (S_1) and ground (S_0) states of the cytosine–guanine base pair versus protein transfer and skeletal deformation of the bonds. The dashed yellow and green lines represent the path sampled in a typical trajectory. Motion along the proton transfer coordinate connects the Franck-Condon geometry to the minimum of the S_1 state. Nonradiative decay occurs along the intersection seam between the surfaces. Note that the trajectory encounters the seam more than once and recrosses between the excited and ground states repeatedly. (Please find a color version of this figure on the color plates.)

applications that we have selected here, the simulations could provide detailed structural and dynamical information of the photobiological process at a resolution well beyond what is achievable experimentally. The applications also demonstrate what is feasible today with on-the-fly molecular dynamics simulations, and where the limits are. These limits are predominantly imposed by the current state of computer technology, which restricts both system size and timescale of the processes studied. However, the expected increase of computer power, complemented by the development of more efficient electronic structure methods and new algorithms, will enable the study of larger systems and longer timescales in the near future. Therefore, excited state molecular dynamics simulation has the potential to ultimately lead to a better understanding of photobiological reactions. Furthermore, the simulations will enable the prediction of photochemical properties and aid the rational design of new photoactive systems.

References

Abo-Riziq, A., Grace, L., Nir, E., Kabelac, M., Hobza, P., and deVries, M.S. (2005) Photochemical selectivity in guanine-cytosine base-pair structures. *Proc. Natl. Acad. Sci. U.S.A.*, **1021**, 20–23.

Agmon, N. (2005) Proton pathways in green fluorescence protein. *Biophys. J.*, **884**, 2452–61.

Altoe, P., Bernardi, F., Garavelli, M., Orlandi, G., and Negri, F. (2005) Solvent effects on the vibrational activity and photodynamics of the green fluorescent protein chromophore: a quantum-chemical study. *J. Am. Chem. Soc.*, **127**, 11534–35.

Amara, P. and Field, M.J. (2003) Evaluation of an ab initio quantum mechanical/molecular mechanical hybrid-potential link-atom method. *Theor. Chem. Acc.*, **109**, 43–52.

Amat, P., Granucci, G., Buda, F., Persico, M., and Tozzini, V. (2006) The chromophore of asFP595: a theoretical study. *J. Phys. Chem., B*, **110**, 9348–53.

Ando, R., Mizuno, H., and Miyawaki, A. (2004) Regulated fast nucleocytoplasmic shuttling observed by reversible protein highlighting. *Science*, **306**(5700), 1370–73.

Andresen, M., Wahl, M.C., Stiel, A.C., Gräter, F., Schäfer, L.V., Trowitzsch, S., Weber, G., Eggeling, C., Grubmüller, H., Hell, S.W., and Jakobs, S. (2005) Structure and mechanism of the reversible photoswitch of a fluorescent protein. *Proc. Natl. Acad. Sci. U.S.A.*, **10237**, 13070–74.

Atchity, G.J. and Ruedenberg, K. (1999) Orbital transformations and configurational transformations of electronic wavefunctions. *J. Chem. Phys.*, **111**, 2910–20.

Atchity, G.J., Xantheas, S.S., and Ruedenberg, K. (1991) Potential-energy surfaces near intersections. *J. Chem. Phys.*, **95**, 1862–76.

Baeck, K.K. and Martinez, T.J. (2003) Ab initio molecular dynamics with equation-of-motion coupled-cluster theory: electronic absorption spectrum of ethylene. *Chem. Phys. Lett.*, **375**, 299–308.

Bartlett, R.J. and Musial, M. (2007) Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.*, **79**, 291–352.

Bell, A.F., Stoner-Ma, D., Wachter, R.M., and Tonge, P.J. (2003) Light-driven decarboxylation of wild-type green fluorescent protein. *J. Am. Chem. Soc.*, **12523**, 6919–26.

Berendsen, H.J.C. (2001) Bio-molecular dynamics comes of age. *Science*, **271**, 954–55.

Boggio-Pasqua, M., Bearpark, M.J., Hunt, P.A., and Robb, M.A. (2002) Dihydroazulene/vinylheptafulvene photochromism: a model for one-way photochemistry via a conical intersection. *J. Am. Chem. Soc.*, **124**, 1456–70.

Boggio-Pasqua, M., Bearpark, M.J., Klene, M., and Robb, M.A. (2004) A computational strategy for geometry optimization of ionic and covalent excited states, applied to butadiene and hexatriene. *J. Chem. Phys.*, **120**, 7849–60.

Brejc, K., Sixma, T.K., Kitts, P.A., Kain, S.R., Tsien, R.Y., Ormo, M., and Remington, S.J. (1997) Structural basis for dual excitation and photoisomerization of the aequorea victoria green fluorescent protein. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 2306–11.

Burghardt, I. (2001) Reduced dynamics with initial conditions: multiconfigurational approach. *J. Chem. Phys.*, **114**, 89–101.

Celani, P. and Werner, H.J. (2003) Analytical energy gradients for internally contracted second-order multireference perturbation theory. *J. Chem. Phys.*, **119**, 5044–57.

Changenet-Barret, P., Plaza, P., Martin, M.M., Chosrowjan, H., Taniguchi, S., Mataga, N., Imamoto, Y., and Kataoka, M. (2007) Role of arginine 52 on the primary photoinduced events in the pyp photocycle. *Chem. Phys. Lett.*, **434**, 320–25.

Chudakov, D.M., Belousov, V.V., Zarisky, A.G., Novoselov, V.V., Staroverov, D.B., Zorov, D.B., Lukyanov, S., and Lukyanov, K.A. (2003a) Kindling fluorescent proteins for precise in vivo photolabeling. *Nat. Biotechnol.*, **21**, 191–94.

Chudakov, D.M., Feofanov, A.V., Mudrik, N.N., Lukyanov, S., and Lukyanov, K.A. (2003b) Chromophore environment provides clue to "kindling fluorescent protein" riddle. *J. Biol. Chem.*, **2789**, 7215–19.

Coe, J.D., Levine, B.G., and Martinez, T.J. (2007) Ab initio molecular dynamics of excited-state intramolecular proton transfer using multireference perturbation theory. *J. Phys. Chem., A*, **111**, 11302–10.

Coe, J.D. and Martinez, T.J. (2005) Competitive decay at two- and three-state conical intersections in excited-state intramolecular proton transfer. *J. Am. Chem. Soc.*, **12713**, 4560–61.

Coe, J.D. and Martinez, T.J. (2006) Ab initio molecular dynamics of excited-state intramolecular proton transfer around a three-state conical intersection in malonaldehyde. *J. Phys. Chem., A*, **110**, 618–30.

Crano, J.C. and Guglielmetti, J. (eds) (1999) *Organic Photochromic and Thermochromic Compounds: Main Photochromic Families*, Vol. 1, Plenum, New York.

Dedecker, P., Hotta, J., Ando, R., Miyawaki, A., Engelborghs, Y., and Hofkens, J. (2006) Fast and reversible photoswitching of the fluorescent protein Dronpa as evidenced by fluorescence correlation spectroscopy. *Biophys. J.*, **915**, L45–L47.

Desouter-Lecomte, M. and Lorquet, J. (1979) Nonadiabatic interactions in unimolecular decay. IV. Transition probability as a function of the massey parameter. *J. Chem. Phys.*, **71**, 4391–403.

Dong, J., Solntsev, K.M., and Tolbert, L.M. (2006) Solvatochromism of the green fluorescence protein chromophore and its derivatives. *J. Am. Chem. Soc.*, **128**, 12038–39.

Dreuw, A. and Head-Gordeon, M. (2005) Single-reference ab initio methods for the calculation of excited states of large molecules. *Chem. Rev.*, **105**, 4009–37.

Fang, J.Y. and Hammes-Schiffer, S. (1999) Improvement of the internal consistency in trajectory surface hopping. *J. Phys. Chem.*, **103**, 9399–407.

Field, M.J., Bash, P.A., and Karplus, M. (1990) A combined quantum-mechanical and molecular mechanical potential for molecular-dynamics simulations. *J. Comput. Chem.*, **11**, 700–33.

Fron, E., Flors, C., Schweitzer, G., Habuchi, S., Ando, R., Mizuno, H., Schryver, F.C.D., Miyawaki, A., and Hofkens, J. (2007) Ultrafast excited-state dynamics of the photoswitchable protein Dronpa. *J. Am. Chem. Soc.*, **12916**, 4870–71.

Frutos, L.M., Andruniow, T., Santoro, F., Ferre, N., and Olivucci, M. (2007) Tracking the excited-state time evolution of the visual pigment with multiconfigurational quantum chemistry. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 7764–69.

Gao, J., Amara, P., Alhambra, C., and Field, M.J. (1998) A generalized hybrid orbital (GHO) method for the treatment of boundary atoms in combined QM/MM calculations. *J. Phys. Chem., A*, **102**, 4714–21.

Geisler, C., Schonle, A., von Middendorff, C., Bock, H., Eggeling, C., Egner, A., and Hell, S.W. (2007) Resolution of lambda/10 in fluorescence microscopy using fast single molecule photo-switching. *Appl. Phys. A: Mater. Sci. Process.*, **882**, 223–26.

Grigorenko, B., Savitsky, A., Topol, I., Burt, S., and Nemukhin, A. (2006) Ground-state structures and vertical excitations for the kindling fluorescent protein asfp595. *J. Phys. Chem., B*, **11037**, 18635–40.

Groenhof, G., Bouxin-Cademartory, M., Hess, B., de Visser, S.P., Berendsen, H.J.C., Olivucci, M., Mark, A.E., and Robb, M.A.

(2004) Photoactivation of the photoactive yellow protein: why photon absorption triggers a trans-to-cis isomerization of the chromophore in the protein. *J. Am. Chem. Soc.*, **126**, 4228–33.

Groenhof, G., Schäfer, L.V., Boggio-Pasqua, M., Goette, M., Grubmüller, H. and Robb, M.A. (2007) Ultrafast deactivation of on excited cytosine-guanine base pair in DNA. *J. Am. Chem. Soc.*, **12921**, 6812–19.

Groenhof, G., Schäfer, L.V., Boggio-Pasqua, M., Grubmüller, H., and Robb, M.A. (2008) Arginine52 controls photoisomerization in photoactive yellow protein. *J. Am. Chem. Soc.*, **130**, 3250–51.

Habuchi, S., Ando, R., Dedecker, P., Verheijen, W., Mizuno, H., Miyawaki, A., and Hofkens, J. (2005) Reversible single-molecule photoswitching in the GFP-like fluorescent protein Dronpa. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(27), 9511–16.

Habuchi, S., Dedecker, P., Hotta, J.I., Flors, C., Ando, R., Mizuno, H., Miyawaki, A., and Hofkens, J. (2006) Photo-induced protonation/deprotonation in the GFP-like fluorescent protein Dronpa: mechanism responsible for the reversible photoswitching. *Photochem. Photobiol. Sci.*, **56**, 567–76.

Hayashi, S., Tajkhorshid, E., and Schulten, K. (2003) Molecular dynamics simulation of bacteriorhodopsin's photoisomerization using ab initio forces for the excited chromophore. *Biophys. J.*, **853**, 1440–49.

He, X., Bell, A.F., and Tonge, P.J. (2002) Synthesis and spectroscopic studies of model red fluorescent protein chromophores. *J. Phys. Chem., B*, **106**(23), 6056–66.

Hell, S.W. (2003) Toward fluorescence nanoscopy. *Nat. Biotechnol.*, **2111**, 1347–55.

Hell, S.W., Dyba, M., and Jakobs, S. (2004) Concepts for nanoscale resolution in fluorescence microscopy. *Curr. Opin. Neurobiol.*, **145**, 599–609.

Hell, S.W., Jakobs, S., and Kastrup, L. (2003) Imaging and writing at the nanoscale with focused visible light through saturable optical transitions. *Appl. Phys. A: Mater. Sci. Process.*, **777**, 859–60.

Hellingwerf, K.J., Hendriks, J., and Gensch, T. (2003) Photoactive yellow protein, a new type of photoreceptor protein: will this "yellow lab" bring us where we want to go? *J. Phys. Chem., A*, **107**, 1082–94.

Helms, V. (2002) Electronic excitations of biomolecules studied by quantum chemistry. *Curr. Opin. Struct. Biol.*, **122**, 169–75.

Hofmann, M., Eggeling, C., Jakobs, S., and Hell, S.W. (2005) Breaking the diffraction barrier in fluorescence microscopy at low light intensities by using reversibly photoswitchable proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(49), 17565–69.

Jensen, F. (2001) *Introduction to Computational Chemistry*, John Wiley & Sons, New York.

Keal, T.W., Koslowski, A., and Thiel, W. (2007) Comparison of algorithms for conical intersection optimisation using semiempirical methods. *Theor. Chem. Acc.*, **118**, 837–844.

Köppel, H., Domcke, W., and Cederbaum, L.S. (1984) Multimode molecular dynamics beyond the born-oppenheimer approximation. *Adv. Chem. Phys.*, **57**, 59–246.

Koppel, H., Gronki, J., and Mahapatra, S. (2001) Construction scheme for regularized diabatic states. *J. Chem. Phys.*, **115**, 2377–88.

Koyama, Y., Kubo, K., Komori, M., Yasuda, H., and Mukai, Y. (1991) Effect of protonation on the isomerization properties of normal-butylamine schiff-base of isomeric retinal as revealed by direct hplc analyses – selection of isomerization pathways by retinal proteins. *Photochem. Photobiol.*, **54**, 433–43.

Laino, T., Nifosi, R., and Tozzini, V. (2004) Relationship between structure and optical properties in green fluorescent proteins: a quantum mechanical study of the chromophore environment. *Chem. Phys.*, **298**(1–3), 17–28.

Leiderman, P., Huppert, D., and Agmon, N. (2006) Transition in the temperature-dependence of GFP fluorescence: from proton wires to proton exit. *Biophys. J.*, **903**, 1009–18.

Lengsfield, B.H., Saxe, P., and Yarkony, D.R. (1984) On the evaluation of nonadiabatic coupling matrix-elements using sa-mcscf/cl wave-functions and analytic gradient methods. 1. *J. Chem. Phys.*, **81**, 4549–53.

Lengsfield, B.H. and Yarkony, D.R. (1992) Nonadiabatic interactions between potential-energy surfaces – theory and applications. *Adv. Chem. Phys.*, **82**, 1–71.

Levine, B.G. and Martinez, T.J. (2007) Isomerization through conical intersections. *Annu. Rev. Phys. Chem.*, **58**, 613–34.

Lill, M.A. and Helms, V. (2002) Proton shuttle in green fluorescent protein studied by dynamic simulations. *Proc. Natl. Acad. Sci. U.S.A.*, **995**, 2778–81.

Lipkowitz, K.B. and Boyd, D.B. (eds) (2000) *Reviews in Computational Chemistry*, Wiley-VCH, New York.

Lippincott-Schwartz, J., Altan-Bonnet, N., and Patterson, G.H. (2003) Photobleaching and photoactivation: following protein dynamics in living cells. *Nat. Cell Biol.*, S7–S14.

Litvinenko, K.L., Webber, N.M., and Meech, S.R. (2001) An ultrafast polarisation spectroscopy study of internal conversion and orientational relaxation of the chromophore of the green fluorescent protein. *Chem. Phys. Lett.*, **346**(1–2), 47–53.

Luecke, H., Scholbert, B., Richter, H.T., Cartailler, J.P., and Lanyi, J.K. (1999) Structure of bacteriorhodopsin at 1.55 angstrom resolution. *J. Mol. Biol.*, **291**, 899–911.

Lukyanov, K.A., Chudakov, D.M., Lukyanov, S., and Verkhusha, V.V. (2005) Photoactivatable fluorescent proteins. *Nat. Rev. Mol. Cell Biol.*, **611**, 885–91.

Lukyanov, K.A., Fradkov, A.F., Gurskaya, N.G., Matz, M.V., Labas, Y.A., Savitsky, A.P., Markelov, M.L., Zaraisky, A.G., Zhao, X.N., Fang, Y., Tan, W.Y., and Lukyanov, S.A. (2000) Natural animal coloration can be determined by a nonfluorescent green fluorescent protein homolog. *J. Biol. Chem.*, **275**(34), 25879–82.

Marques, M.A.L., Lopez, X., Varsano, D., Castro, A., and Rubio, A. (2003) Time-dependent density-functional approach for biological chromophores: the case of the green fluorescent protein. *Phys. Rev. Lett.*, 9025.

Martin, M.E., Negri, F., and Olivucci, M. (2004) Origin, nature, and fate of the fluorescent state of the green fluorescent protein chromophore at the CASPT2//CASSCF resolution. *J. Am. Chem. Soc.*, **126**, 5452–64.

Martinez, T.J. (2006) Insights for light-driven molecular devices from ab initio multiple spawning excited-state dynamics of organic and biological chromophores. *Acc. Chem. Res.*, **392**, 119–26.

Mataga, N., Chosrowjan, H., Shibata, Y., Imamoto, Y., and Tokunaga, F. (2000) Effects of modification of protein nanospace structure and change of temperature on the femtosecond to picosecond fluorescence dynamics of photoactive yellow protein. *J. Phys. Chem., B*, **104**, 5191–99.

Matsika, S. and Yarkony, D.R. (2002) Intersecting conical intersection seams: their location, representation, and effect on local topography. *J. Phys. Chem., A*, **106**, 2580–91.

McCammon, J.A., Gelin, B.R., Karplus, M., and Wolynes, P.G. (1976) Hinge-bending mode in lysozyme. *Nature*, **262**, 325–26.

Miannay, F.A., Banyasz, A., Gustavsson, T., and Markovitsi, D. (2007) Ultrafast excited-state deactivation and energy transfer in guanine-cytosine DNA double helices. *J. Am. Chem. Soc.*, **129**(47), 14574–75.

Miyawaki, A., Sawano, A., and Kogure, T. (2003) Lighting up cells: labelling proteins with fluorophores. *Nat. Cell Biol.*, S1–S7.

Muller, H., Koppel, H., and Cederbaum, L.S. (1993) Topology and dynamics of vibronically coupled potential-energy surfaces – general-aspects and application to the ozone cation. *New J. Chem*, **17**, 7–29.

Müller, U. and Stock, G. (1997) Surface-hopping modeling of photoinduced relaxation dynamics on coupled potential-energy surfaces. *J. Chem. Phys.*, **107**, 6230–44.

Nemukhin, A.V., Topol, I.A., and Burt, S.K. (2006) Electronic excitations of the chromophore from the fluorescent protein asFP595 in solutions. *J. Chem. Theory Comput.*, **2**, 292–99.

Nifosi, R. and Tozzini, V. (2006) Cis-trans photoIsomerization of the chromophore in the green fluorescent protein variant e(2)gfp: a molecular dynamics study. *Chem. Phys.*, **323**(2–3), 358–68.

Okada, T., Sugihara, M., Bondar, A.N., Elstner, M., Entel, P., and Buss, V. (2004) The retinal conformation and its environment in rhodopsin in light of a new 2.2 angstrom crystal structure. *J. Mol. Biol.*, **342**, 571–83.

Onuchic, J. and Wolynes, P. (1998) Classical and quantum pictures of reaction dynamics in condensed matter: resonances, dephasing and all that. *J. Phys. Chem.*, **92**, 6495–503.

Ormö, E.M., Cubitt, A.B., Kallio, K., Gross, L.A., Tsien, R.Y., and Remington, S.J. (1996) Crystal structure of the *Aequorea* *victoria* green fluorescent protein. *Science,* **273**(5280), 1392–95.

Patnaik, S.S., Trohalaki, S., Naik, R.R., Stone, M.O. and Pachter, R. (2007) Computational study of the absorption spectra of green fluorescent protein mutants. *Biopolymers,* **853**, 253–63.

Philipp, D.M. and Friesner, R.A. (1999) Mixed ab initio QM/MM modeling using frozen orbitals and tests with alanine dipeptide and tetrapeptide. *J. Comput. Chem.*, **20**(14), 1468–94.

Mathies, R.A. and Lugtenburg, J. (2000) *Handbook of Biological Physics,* Elsevier, Amsterdam.

Remington, S.J. (2006) Fluorescent proteins: maturation, photochemistry and photophysics. *Curr. Opin. Struct. Biol.*, **166**, 714–21.

Roos, B.O. (1999) Theoretical studies of electronically excited states of molecular systems using multiconfigurational perturbation theory. *Acc. Chem. Res.*, **32**, 137–44.

Salih, A., Larkum, A., Cox, G., Kühl, M., and Hoegh-Guldberg, O. (2000) Fluorescent pigments in corals are photoprotective. *Nature,* **408**, 850–53.

Sanchez-Galvez, A., Hunt, P., Robb, M.A., Olivucci, M., Vreven, T., and Schlegel, H.B. (2000) Ultrafast radiationless deactivation of organic dyes: evidence for a two-state two-mode pathway in polymethine cyanines. *J. Am. Chem. Soc.*, **122**, 2911–24.

Sauer, M. (2005) Reversible molecular photoswitches: a key technology for nanoscience and fluorescence imaging. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(27), 9433–34.

Schäfer, L.V., Groenhof, G., Boggio-Pasqua, M., Robb, M.A., and Grubmüller, H. (2008) Chromophore protonation state controls photoswitching of the fluoroprotein asfp595. *PLOS Comput. Biol.*, **4**, e1000034.

Schäfer, L.V., Groenhof, G., Klingen, A.R., Ullmann, G.M., Boggio-Pasqua, M., Robb, M.A., and Grubmüller, H. (2007) Photoswitching of the fluorescent protein asFP595: mechanism, proton pathways, and absorption spectra. *Angew. Chem. Int., Ed.*, **46**, 530–36.

Schellenberg, P., Johnson, E., Esposito, A.P., Reid, P.J., and Parson, W.W. (2001) Resonance raman scattering by the green fluorescent protein and an analogue of its chromophore. *J. Phys. Chem., B,* **105**(22), 5316–22.

Schüttrigkeit, T.A., von Feilitzsch, T., Kompa, C.K., Lukyanov, K.A., Savitsky, A.P., Voityuk, A.A., and Michel-Beyerle, M.E. (2006) Femtosecond study of light-induced fluorescence increase of the dark chromoprotein asFP595. *Chem. Phys.*, **323**(2–3), 149–60.

Schwalb, N.K. and Temps, F. (2007) Ultrafast electronic relaxation in guanosine is promoted by hydrogen bonding with cytidine. *J. Am. Chem. Soc.*, **129**, 9272–73.

Shimizu, N., Kamikubo, H., Yamazaki, Y., Imamoto, Y., and kataoka, M. (2006) The crystal structure of the r52q mutant demonstrates a role for r52 in chromophore pk(a) regulation in photoactive yellow protein. *Biochemistry,* **45**, 3542–47.

Sinicropi, A., Andruniow, T., Ferre, N., Basosi, R., and Olivucci, M. (2005) Properties of the emitting state of the green fluorescent protein resolved at the CASPT2//CASSCF/CHARMM level. *J. Am. Chem. Soc.*, **127**(33), 11534–35.

Sobolewski, A.L. and Domcke, W. (2004) Ab initio studies on the photophysics of the guanine-cytosine base pair. *Phys. Chem. Chem. Phys.*, **610**, 2763–71.

Sobolewski, A.L., Domcke, W., and Hättig, C. (2005) Tautomeric selectivity of the excited-state lifetime of guanine/cytosine base pairs: the role of electron-driven proton-transfer processes. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(50), 17903–6.

Stoner-Ma, D., Jaye, A.A., Matousek, P., Towrie, M., Meech, S.R., and Tonge, P.J. (2005) Observation of excited-state proton transfer in green fluorescent protein using ultrafast vibrational spectroscopy. *J. Am. Chem. Soc.*, **127**(9), 2864–65.

Stoner-Ma, D., Melief, E.H., Nappa, J., Ronayne, K.L., Tonge, P.J., and Meech, S.R. (2006) Proton relay reaction in green fluorescent protein (GFP): polarization-resolved ultrafast vibrational spectroscopy of isotopically edited GFP. *J. Phys. Chem., B,* **110**(43), 22009–18.

Striker, G., Subramaniam, V., Seidel, C.A.M., and Volkmer, A. (1999) Photochromicity and fluorescence lifetimes of green fluorescent protein. *J. Phys. Chem., B,* **103**(40), 8612–17.

Szabo, A. and Ostlund, N.S. (1989) *Modern Quantum Chemistry*, Dover Publications, New York.

Takeshita, K., Imamoto, Y., Kataoka, M., Mihara, K., Tokunaga, F., and Terazima, M. (2002) Structural change of site-directed mutants of pyp: new dynamics during pr state. *Biophys. J.*, **83**, 1567–77.

Tavernelli, I., Rohrig, U.F., and Rothlisberger, U. (2005) Molecular dynamics in electronically excited states using time-dependent density functional theory. *Mol. Phys.*, **103**, 963–81.

Thachuk, M., Ivanov, M.Y., and Wardlaw, D.M. (1998) A semiclassical approach to intense-field above-threshold dissociation in the long wavelength limit. II. Conservation principles and coherence in surface hopping. *J. Chem. Phys.*, **109**, 5747–60.

Tittor, J. and Oesterhelt, D. (1990) The quantum yield of bacteriorhodopsin. *FEBS Lett.*, **263**, 269–73.

Toniolo, A., Granucci, G., and Martinez, T.J. (2003) Conical intersections in solution: a QM/MM study using floating occupation semiempirical configuration interaction wave functions. *J. Phys. Chem., A*, **107**(19), 3822–30.

Toniolo, A., Olsen, S., Manohar, L., and Martinez, T.J. (2004) Conical intersection dynamics in solution: the chromophore of green fluorescent protein. *Faraday Discuss.*, **127**, 149–63.

Tsien, R.Y. (1998) The green fluorescent protein. *Annu. Rev. Biochem.*, **67**, 509–44.

Tully, J.C. (1990) Molecular dynamics with electronic transitions. *J. Chem. Phys.*, **93**, 1061–71.

Tully, J.C. (1991) Nonadiabatic molecular dynamics. *Int. J. Quantum Chem.*, **25**, 299–309.

Tully, J.C. and Preston, R. (1972) Trajectory surface hopping approach to nonadiabatic molecular collisions: the reaction of h^+ with d_2. *J. Chem. Phys.*, **55**, 562–72.

Vendrell, O., Gelabert, R., Moreno, M., and Lluch, J.M. (2006) Potential energy landscape of the photoinduced multiple proton-transfer process in the green fluorescent protein: classical molecular dynamics and multiconfigurational electronic structure calculations. *J. Am. Chem. Soc.*, **128**, 3564–74.

Voityuk, A.A., Kummer, A.D., Michel-Beyerle, M.E., and Rosch, N. (2001) Absorption spectra of the GFP chromophore in solution: comparison of theoretical and experimental results. *Chem. Phys.*, **269**(1–3), 83–91.

Voityuk, A.A., Michel-Beyerle, M.E., and Rosch, N. (1998a) Quantum chemical modeling of structure and absorption spectra of the chromophore in green fluorescent proteins. *Chem. Phys.*, **231**(1), 13–25.

Voityuk, A.A., Michel-Beyerle, M.E., and Rosch, N. (1998b) Structure and rotation barriers for ground and excited states of the isolated chromophore of the green fluorescent protein. *Chem. Phys. Lett.*, **296**(3–4), 269–76.

von Neumann, J. and Wigner, E. (1929) Concerning the behaviour of eigenvalues in adiabatic processes. *Z. Phys.*, **30**, 467–70.

Wang, J.M., Cieplak, P., and Kollman, P.A. (2000) How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, **21**, 1049–74.

Warshel, A. (1976) Bicycle-pedal model for 1st step in vision process. *Nature*, **260**(5553), 679–83.

Warshel, A. and Barboy, N. (1982) Energy-storage and reaction pathways in the 1st step of the vision process. *J. Am. Chem. Soc.*, **104**, 1469–76.

Warshel, A. and Chu, Z.T. (2001) Nature of the surface crossing process in bacteriorhodopsin: Computer simulations of the quantum dynamics of the primary photochemical event. *J. Phys. Chem., B*, **105**, 9857–71.

Warshel, A., Chu, Z.T., and Hwang, J.K. (1991) The dynamics of the primary event in rhodopsins revisited. *Chem. Phys.*, **158**, 303–14.

Warshel, A. and Levitt, M. (1976) Theoretical studies of enzymatic reactions: dielectric, electrostatic and steric stabilization of carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, **103**, 227–49.

Webber, N.M., Litvinenko, K.L., and Meech, S.R. (2001) Radiationless relaxation in a synthetic analogue of the green fluorescent protein chromophore. *J. Phys. Chem., B*, **105**(33), 8036–39.

Weber, W., Helms, V., McCammon, J.A., and Langhoff, P.W. (1999) Shedding light on the dark and weakly fluorescent states of green fluorescent proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 6177–82.

Weingart, O. (2007) The twisted c11=c12 bond of the rhodopsin chromophore-a photochemical hot spot. *J. Am. Chem. Soc.*, **129**, 10618–19.

Weingart, O., Migani, A., Olivucci, M., Robb, M.A., Buss, V., and Hunt, P. (2004) Probing the photochemical funnel of a retinal chromophore model via zero-point energy sampling semiclassical dynamics. *J. Phys. Chem., A*, **108**, 4685–93.

Werner, H.J. and Meyer, W. (1981) A quadratically convergent mcscf method for the simultaneous-optimization of several states. *J. Chem. Phys.*, **74**, 5794–801.

Wilmann, P.G., Turcic, K., Battad, J.M., Wilce, M.C.J., Devenish, R.J., Prescott, M., and Rossjohn, J. (2006) The 1.7Å crystal structure of Dronpa: a photoswitchable green fluorescent protein. *J. Mol. Biol.*, **364**(2), 213–24.

Worth, G.A. and Cederbaum, L.S. (2004) Beyond Born-Oppenheimer: molecular dynamics through a conical intersection. *Annu. Rev. Phys. Chem.*, **55**, 127–58.

Worth, G.A. and Robb, M.A. (2002) Applying direct molecular dynamics to nonadiabatic systems. *Adv. Chem. Phys.*, **124**, 355–431.

Yang, F., Moss, L.G., and Phillips, G.N. (1996) The molecular structure of green fluorescent protein. *Nat. Biotechnol.*, **14**(10), 1246–51.

Yarkony, D.R. (2000) On the adiabatic to diabatic states transformation near intersections of conical intersections. *J. Chem. Phys.*, **112**, 2111–20.

Zimmer, M. (2002) Green fluorescent protein (GFP): applications, structure and related photophysical behavior. *Chem. Rev.*, **102**, 759–81.

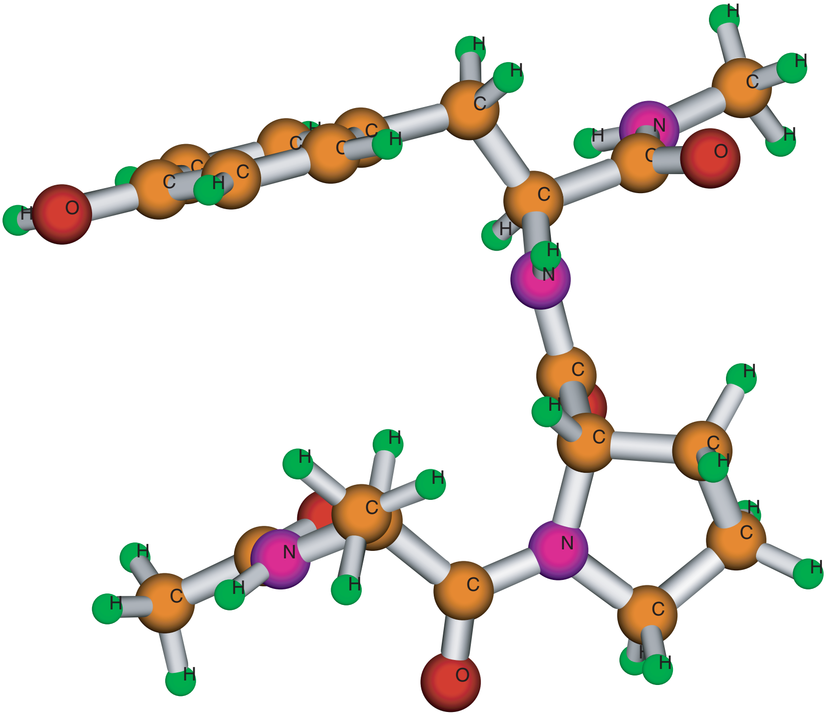
H
O
C
C
H
C
C
C
C
H
C
H
H
C
H
N
C
O
H
C
H
H
N
H
C
H
C
C
H
H
C
H
H
H
C
H
H
N
H
C
C
N
C
H
C
H
H
O
C
H
H

Fig. 1.8 *N*-Acetyl L-alanine L-proline L-tyrosine *N*′-methylamide molecule. B3PW91/6-31G** plus PCM continuum solvent model optimized geometry. (This figure also appears on page 55.)

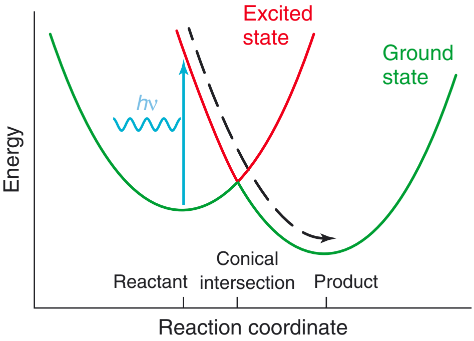
Excited state
Ground state
hν
Energy
Conical intersection
Reactant
Product
Reaction coordinate

Fig. 3.1 One-dimensional scheme of a photochemical reaction. After photon absorption (cyan arrow), evolution takes place on the excited state potential energy surface (red), until the system hits the conical intersection. At the intersection, a radiationless transition to the ground state occurs (green). After the decay, the system continues evolving in the ground state. (This figure also appears on page 96.)

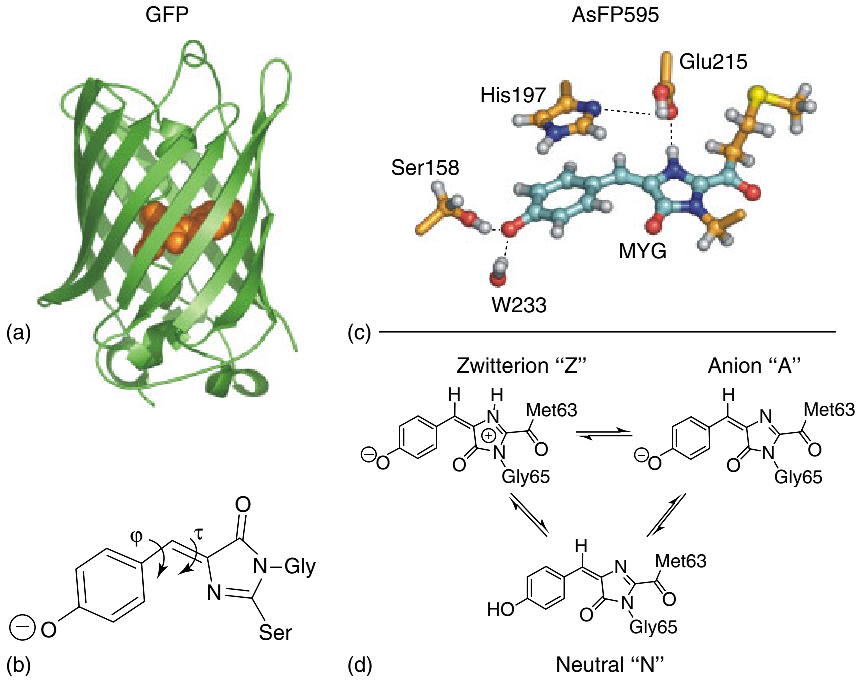
GFP
AsFP595
Glu215
His197
Ser158
MYG
W233
(a)
(c)
Zwitterion "Z"
Anion "A"
H
H
N
Met63
O
N
O
Gly65
H
N
Met63
O
N
O
Gly65
φ
τ
O
N−Gly
N
Ser
H
N
Met63
HO
O
N
O
Gly65
(b)
(d)
Neutral "N"

Fig. 3.8 Fluorescent proteins GFP and asFP595. (a) X-ray structure of GFP (PDB entry 1EMB), Ref. Brejc *et al.* (1997), with the β-barrel fold shown in green and the chromophore (in the center of the protein) in orange spheres. (b) Schematic drawing of the anionic form of the *cis* *p*-hydroxybenzylidene-imidazolinone chromophore. (c) The asFP595 chromophore (MYG) in the *trans* conformation and adjacent amino acid side chains. Dashed lines represent hydrogen bonds between MYG and Ser158, Glu215, and a crystallographic water molecule (W233), respectively. His197 is π-stacked to the MYG phenoxy-moiety and forms a hydrogen bond to Glu215. The carbon skeleton of the quantum mechanical (QM) subsystem is shown in cyan, and the carbon atoms modeled by molecular mechanics (MM) are shown in orange. (d) Schematic drawings of the important MYG protonation states. (This figure also appears on page 122.)

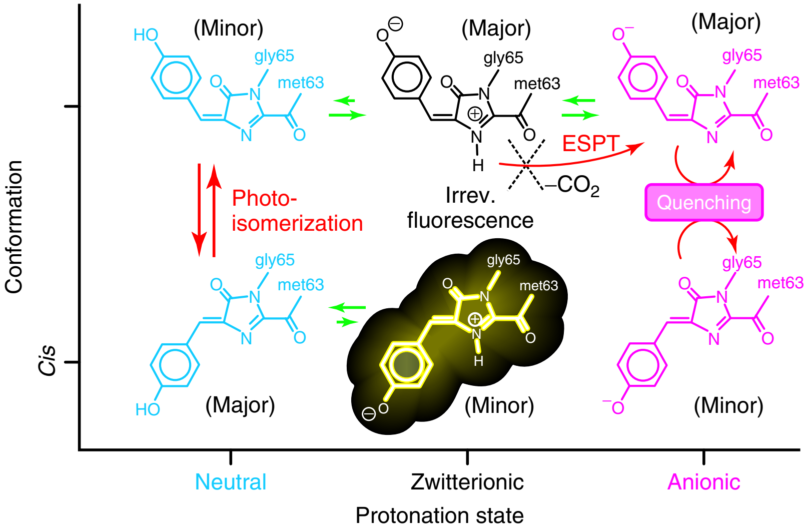
(Minor)
HO
gly65
met63
(Major)
gly65
met63
(Major)
gly65
met63
ESPT
Photo-
isomerization
Irrev.
fluorescence
–CO2
Quenching
gly65
met63
gly65
met63
gly65
met63
(Major)
(Minor)
(Minor)
Conformation
Cis
Neutral
Zwitterionic
Anionic
Protonation state

Fig. 3.9 Scheme of the reversible photoswitching mechanism of asFP595. The fluorescent state Z_{cis} is highlighted. The green arrows indicate ground state equilibria, whereas the red arrows indicate excited state processes. The major protonation states are the zwitterionic and the anionic chromophores in the *trans* conformation, and the neutral chromophore in the *cis* conformation, as indicated in the square brackets. (This figure also appears on page 125.)

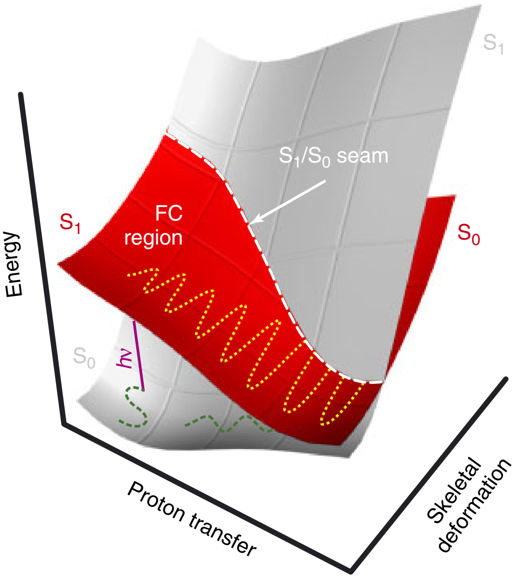
S$_1$
S$_1$/S$_0$ seam
S$_1$
FC region
S$_0$
Energy
S$_0$
hν
Proton transfer
Skeletal deformation

Fig. 3.11 Artist impression of the potential energy surfaces of the excited (S_1) and ground (S_0) states of the cytosine–guanine base pair versus protein transfer and skeletal deformation of the bonds. The dashed yellow and green lines represent the path sampled in a typical trajectory. Motion along the proton transfer coordinate connects the Franck-Condon geometry to the minimum of the S_1 state. Nonradiative decay occurs along the intersection seam between the surfaces. Note that the trajectory encounters the seam more than once and recrosses between the excited and ground states repeatedly. (This figure also appears on page 127.)

4

Electron Interactions and Currents in CFT, a Current Functional Theory Applied to Molecules

Jens N. Bang, Henrik G. Bohr, and Joao da Providencia

4.1 **Time-Dependent DFT 137**
4.1.1 Introduction 137
4.1.2 The Standard DFT Formalism 137
4.1.3 An Expression for the Exchange Correlation Energy 139
4.2 **A Framework for Describing Time Development of Electronic Structures 139**
4.2.1 Normal Modes 142
4.2.2 Single Electron Description 143
4.2.3 Electron–Nuclei Description 146
4.3 **Relativistic Electrons 148**
4.4 **The General Description for the Electronic Evolution Equations 150**
4.5 **Multielectronic Interaction and Collective Phenomena 153**
4.5.1 Collective Coordinates and Plasmons 153
4.5.2 Landau Damping 155
4.6 **Molecular Calculations 156**
4.6.1 Gaussian Integration 157
4.6.2 The Case of the ZnO Molecule 158
4.6.3 The Case of a Water Molecule 159
4.7 **Numerical Treatments 161**
4.7.1 Comparison of the Standard DFT's and Our Method's Numerics 161
Appendix A: **Gaussian Expansions and Numerical Calculations 162**
Appendix B: **Transition Charge Density 163**
References 164

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

4.1
Time-Dependent DFT

In the following, a time-dependent, quantum mechanical treatment of many-body systems, for example large molecules and atomic cluster, is proposed and formulated in terms of evolution equations (Bang and Bohr, 2008). For that purpose, the density functional theory (DFT) formalism is reformulated into a framework of currents so as to give the energy a parameter-dependent behavior, of, for example, time, which allows for following trajectories of the electrons. Furthermore, a more rigorous form of the exchange potentials is incorporated. The current formalism is describing the transition of electrons from one orbital to another. The formalism is also extended to the relativistic region.

4.1.1
Introduction

At the present time, electronic structure calculations of larger molecules are the focus of intense scientific activity. Here, DFT plays a prominent role of being the most used framework for performing quantum calculations of large molecular systems such as biomolecules. However, electronic calculations of giant biomolecules, for example, protein complexes and membrane systems, which are of interest for researchers in biotechnology, are difficult to be carried out in a quantum mechanical setting with the present- day techniques (Bohr, Jalkanen, *et al.*, 1999). They comprise, for example, the standard DFT computer programs that have been very successful in dealing with smaller molecules. This is also due to the fact that the many-body Schrödinger equation is very complicated to solve since it involves many coupled partial differential equations. We shall, as being explained in the following, instead be using the equations for electronic currents and the corresponding continuity equations that, with the help of vector algebra, lead to simple evolution equations for several electrons (Davidov, 1965). In the following, we shall first briefly mention some of the celebrated methods used today before explaining our own method.

4.1.2
The Standard DFT Formalism

The most prominent aim of the methods of quantum molecular calculations is to solve and obtain electron orbitals in molecules. We shall in the later chapters

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

show various approaches to that. In this section, we shall, however, review some of the standard features of DFT with the usual equations (Atkins and Friedman, 1997) before we develop new tools.

Ordinarily, the DFT formalism for large molecules is formulated as follows:

One starts with the energy, $E(\rho)$, as a function of the density of charged particles (electrons):

$$E(\rho) = -\frac{\hbar^2}{2me}\sum_{i=1}^{n}\int \Psi_i^*(r_1)\nabla^2\Psi_i(r_1)dr_1 - \sum_{I=1}^{N}\int \rho(r_1)\frac{Z_I e^2}{4\pi\varepsilon_o r_{I1}}dr_1 + \int \frac{\rho(r_1)\rho(r_2)e^2}{4\pi\varepsilon_o r_{I2}}dr_1 dr_2 + E_{XC}[\rho] \tag{1}$$

where the one-electron orbitals Ψ_i are the Kohn–Sham (KS) orbitals that satisfy the equation of motion:

$$\left\{-\frac{\hbar}{2m_e}\nabla_1^2 - \sum_{I=1}^{N}\frac{Z_I e^2}{4\pi\varepsilon_o r_{I1}} + \int dr_2 \times \frac{\rho(r_2)e^2}{4\pi\varepsilon_o r_{I2}} + V_{XC}\right\}\Psi_i(\mathbf{r}_1) = \varepsilon_i\Psi_i(\mathbf{r}_1) \tag{2}$$

The energy function is given as a function of the charge density at the exact ground state at the location $\mathbf{r}$ and this charge density is given by

$$\rho(\mathbf{r}) = \sum_{i=1}^{n}|\Psi_i(\mathbf{r})|^2 \tag{3}$$

In the energy expression above, the first term represents the kinetic energy of the electrons, the second term represents the interaction term between the nuclei (with index I and atomic number Z_I), and the third term being the Coulomb interaction between the electron charges of the one-electron orbitals with indices i ($i = 1, 2, \ldots$). The term ε_i is the KS orbital energies, and V_{XC} is the exchange correlation (xc) potential, which is a functional derivative of the xc energy $V_{XC}(\rho) = \frac{\delta E_{XC}}{\delta\rho}$, which means that if E_{XC} is known then V_{XC} can be obtained. The KS orbitals in the DFT formalism have the advantage that they allow the density, ρ, to be computed from the equations above. All the equations should be satisfied to ensure self-consistency.

The usual procedure for running a DFT procedure is first to guess the charge density of a certain molecular system where we use a superposition of our atomic densities. Secondly, we use some approximate form for the functional dependency of E_{XC} on the density, which then allows us to calculate the potential $V_{XC}(\rho)$ as a function of r. Next, one is in the position to get the first iterative estimate of the KS orbitals by solving the equations of motion. From these preliminary KS orbital values, one can compute an improved density from the expression for the charge density in terms of the orbitals. Such a procedure is repeated until the density and the xc energy have converged to within a certain tolerance. Finally, the total electronic energy $E(\rho)$ can be computed from Eq. (1). In each step of the procedure, the KS orbitals are numerically calculated and can be expressed in terms of an expansion in a set of basis functions whose coefficients are basically determined once, solving the equation of motion above. The computational time for DFT scales as the third power of the number of basis functions, the latter, thus, being the crucial number in the quantum calculations. An important step in this procedure is to come up with a good choice for the density and thus for the xc energy. We shall come back to that in the next sections. Let us first make a mathematical proposal for a useful xc energy in the following section.

4.1.3
An Expression for the Exchange Correlation Energy

Let us briefly mention an earlier derivation of an xc expressed as a difference, $W = V - V_H$, between the total potential V and the Hartree potential V_H, and with the help of a new central potential term, also used by Talman and Shadwick, see Bang and Hansteen (2002). We may then write

$$W = \sum_i \sum_{j\neq i} \int d\mathbf{r}'^3 \int d\mathbf{r}''^3 \frac{\psi_i(\mathbf{r})^* \psi_j(\mathbf{r})}{E_j - E_i} \times 2 \sum_k \frac{\psi_j(\mathbf{r}')^* \psi_k(\mathbf{r}') \psi_k(\mathbf{r}'')^* \psi_i(\mathbf{r}'')}{\mid \mathbf{r}' - \mathbf{r}'' \mid} \tag{4}$$

which can be simplified using the same restriction on the indices of Coulomb matrix elements, as in Bang and Hansteen (2002):

$$W = C \sum_i \sum_{j\neq i} \sum_k \frac{\psi_i(\mathbf{r})^* \psi_j(\mathbf{r})}{E_i - E_j} \times \int dr' \frac{(\mathbf{j}^L_{jk}(\mathbf{r}') \cdot \mathbf{j}^L_{ki}(\mathbf{r}'))}{(E_k - E_i)(E_j - E_k)} \tag{5}$$

where C is a constant that contains Planck's constant, $C = \hbar 8\pi$. It is quite remarkable that one can obtain such relatively simple expression for the xc. This is a contribution between different orbitals or quantum states i, j. In the case of $E_i = E_j$, this term vanishes and we are left with the zeroth-order Born term. For details, see Bang and Hansteen (2002).

4.2
A Framework for Describing Time Development of Electronic Structures

We shall discuss an alternative to the standard DFT for the description of *electron* motion of states of large, mainly organic, molecules. DFT, developed particularly by Kohn, Hohenberg, and Sham has proved efficient in many cases, and its use needs no further justification. Among the many formulations of this, we shall particularly keep near to that of Gross *et al.* (1996). As mentioned earlier, our foremost desire is to construct the time development of the electronic structural configurations in large molecules leading to an understanding of the electron trajectories.

First, we shall briefly mention a proof for the time dependence of the KS equation, which makes use of the currents as time derivatives of the densities, as explained in Gross *et al.* (1996). It is these currents that later in the text constitute our evolution equations for the electrons. However, in the proof of the KS Density Functional theorem, Gross *et al.* (1996), rather start with the currents and their continuity equation and then rederive the densities in time and space that satisfy the KS equation.

We start, as in Gross *et al.* (1996), with the usual equation of motion for the expectation value of a time-dependent operator, $\hat{O}(t)$:

$$\frac{\partial}{\partial t}\left\langle \Psi_1 \mid \hat{O}(t) \mid \Psi_0 \right\rangle = \left\langle \Psi_1 \left| \frac{\partial \hat{O}}{\partial t} - \mathrm{i}\left[\hat{O}(t), \hat{H}(t)\right] \right| \Psi_0 \right\rangle \tag{6}$$

where we, instead of considering a fixed orbital $\Psi(t)$, consider transitions from orbital 0 to 1. We can now consider a current density $\hat{\mathbf{j}}(r)$ as the operator, $\hat{O}$, and hence obtain

$$\begin{aligned}\frac{\partial}{\partial t}\mathbf{j}(r,t) &= \frac{\partial}{\partial t}\left\langle \Psi_1 \mid \hat{\mathbf{j}}(r) \mid \Psi_0 \right\rangle \\ &= -\mathrm{i}\left\langle \Psi_1 \mid [\hat{\mathbf{j}}(r), \hat{H}(t)] \mid \Psi_0 \right\rangle\end{aligned} \tag{7}$$

and taking two different KS potentials $V_0(t, r)$ and $V_1(t, r)$ with corresponding currents $\mathbf{j}_0(r, t)$ and $\mathbf{j}_1(r, t)$ we can, as in Gross *et al.* (1996), obtain the corresponding densities $n_0(r)$ and $n_1(r)$ from the following equation:

$$\begin{aligned}&\frac{\partial}{\partial t}(\mathbf{j}_a(r,t) - \mathbf{j}_b(r,t)) \\ &= -\mathrm{i}\left\langle \Psi_1 \mid \left[\hat{\mathbf{j}}(r), \hat{H}_a(t) - \hat{H}_b(t)\right] \mid \Psi_0 \right\rangle \\ &= -n_{10}(r)\mathbf{\nabla}(V_a(r,t) - V_b(r,t))\end{aligned} \tag{8}$$

The usual current densities are obtained with $\Psi_0 = \Psi_1$. An important relation between the currents and the densities is manifested in the continuity equation for different densities n, n':

$$\begin{aligned}&\frac{\partial}{\partial t}(n(r,t) - n'(r,t)) \\ &\quad = -\mathbf{\nabla}\cdot(\mathbf{j}(r,t) - \mathbf{j}'(r,t))\end{aligned} \tag{9}$$

The potentials, V, are generally written as an effective potential V_{ef}, containing a time- dependent external potential, a Hartree term as a function of densities $n(r, t)$ and an xc part, V_{XC}:

$$V_{\text{ef}} = V(r,t) + \int d^3r' \frac{n(r',t)}{r-r'} + V_{XC} \tag{10}$$

We will finally make use of a time-dependent KS equation formulated for currents instead of the usual KS orbitals:

$$\begin{aligned}&\frac{\partial}{\partial t}\mathbf{j}(r,t) = \dot{n}\nabla\varphi - n\nabla\frac{\delta E}{\delta n} \approx -n\nabla\frac{\delta E}{\delta n}, \\ &\quad \mathbf{j}(r,t) = n\nabla\varphi, \qquad (11) \\ &E = \int d^3r\left(\frac{1}{2}n(\nabla\varphi)\cdot(\nabla\varphi)\right. \\ &\qquad \left. + (\nabla\sqrt{n})\cdot(\nabla\sqrt{n}) + Vn\right)\end{aligned}$$

where the current density $\mathbf{j}(r, t)$ is connected to the KS orbitals $\Phi(r, t)$ by the equation

$$\begin{aligned}\mathbf{j}(r,t) &= \frac{1}{2i}(\Phi^*(r,t)\mathbf{\nabla}\Phi(r,t) \\ &\quad - \Phi(r,t)\mathbf{\nabla}\Phi^*(r,t)) \\ &= \Phi(r,t)\Phi^*(r,t)\nabla\varphi(r,t)\end{aligned} \tag{12}$$

The previous discussion applies to single electron motion, $\varphi(r, t)$, being the argument of $\Phi(r, t)$. Thus, we obtain transition charge densities that can be compared to electromagnetic transition amplitudes as in Gross *et al.* (1996).

However, the presented expression for the current applies only to single electron processes. The time development of a many-electron system requires the following analysis.

According to KS ideas, we consider an arbitrary density $n(x)$ and the best wave function in a variational sense, such that

$$\left\langle \Psi \middle| \hat{n}(x) \middle| \Psi \right\rangle = n(x) \tag{13}$$

where $\hat{n}(x)$ is the density operator,

$$\hat{n}(x) = \sum_{j=1}^{N} \delta(x - x_j) \tag{14}$$

Since H and $\hat{n}(x)$ are time even, $|\Psi\rangle$ is also time even. The equilibrium density $n_0(x)$ and the associated wave function

$|\Psi_0\rangle$ are obtained by minimizing the energy functional $E(n) = \langle\Psi|H|\Psi\rangle$, that is, $n_0(x)$ is the solution of $\delta E[n]/\delta n = 0$. The ground state wave function $|\Psi_0\rangle$ is stationary, but, for $n(x) \neq n_0(x)$, the wave function $|\Psi\rangle$ is not stationary, so that, in time, it will acquire time odd components. Given a time even wave function $|\Psi\rangle$, the wave function $\exp(\mathrm{i}P)|\Psi\rangle$, where P is an arbitrary *time odd* operator, is also time even. Time odd components are introduced into the time even wave function $|\Psi\rangle$ through a *time even* Hermitian operator Q, that is $|\Psi_Q\rangle = \exp(-\mathrm{i}Q)|\Psi\rangle$ contains time odd components. Of course, the time evolution of $|\Psi_Q\rangle$ is determined by the Schrödinger equation

$$\mathrm{i}\frac{\partial}{\partial t}|\Psi_Q\rangle = H|\Psi_Q\rangle \tag{15}$$

or by the quantum mechanical Lagrangian

$$\mathcal{L} = \frac{\mathrm{i}}{2}\left(\langle\Psi_Q|\dot{\Psi}_Q\rangle - \langle\dot{\Psi}_Q|\Psi_Q\rangle\right) - \langle\Psi_Q|H|\Psi_Q\rangle \tag{16}$$

Let us consider small deviations from equilibrium, so that $n(x) \neq n_0(x)$, but $\delta n(x) = n(x) - n_0(x)$ is small and $|\Psi\rangle$ is close to $|\Psi_0\rangle$. Moreover, Q is small. Expanding in powers of Q, we have, to leading order,

$$\frac{\mathrm{i}}{2}\left(\langle\Psi_Q|\dot{\Psi}_Q\rangle - \langle\dot{\Psi}_Q|\Psi_Q\rangle\right) \approx \frac{\mathrm{i}}{2}\left(\langle\Psi|\dot{\Psi}\rangle - \langle\dot{\Psi}|\Psi\rangle\right) + \langle\Psi|\dot{Q}|\Psi\rangle \tag{17}$$

$$\langle\Psi_Q|H|\Psi_Q\rangle \approx \frac{1}{2}\langle\Psi_0|[[Q, H], Q]|\Psi_0\rangle + \langle\Psi|H|\Psi\rangle \tag{18}$$

Notice the replacement of $|\Psi\rangle$ by the ground state wave function $|\Psi_0\rangle$ in $\langle\Psi|[[Q, H], Q]|\Psi\rangle$, which is consistent with the assumption of small deviations from equilibrium and of the smallness of Q. Since $|\Psi\rangle$ is time even, it follows that $\langle\Psi|\dot{\Psi}\rangle$ is real and $\langle\Psi|\dot{\Psi}\rangle - \langle\dot{\Psi}|\Psi\rangle = 0$ Thus, the Lagrangian for small fluctuations reads

$$\mathcal{L} = \langle\Psi|\dot{Q}|\Psi\rangle - \frac{1}{2}\langle\Psi_0|[[Q, H], Q]|\Psi_0\rangle - \langle\Psi|H|\Psi\rangle \tag{19}$$

As a natural and simple approximation, we set

$$Q = -\sum_{i=1}^{N}\phi(x_i, t) \tag{20}$$

It follows that

$$\begin{aligned}\langle\Psi|\dot{Q}|\Psi\rangle &= -\int dx\,\dot{\phi}\,n,\\ \langle\Psi_0|[[Q, H], Q]|\Psi_0\rangle &= \frac{1}{2}\int dx\,n_0\,(\nabla\phi)(\nabla\phi),\\ \langle\Psi|H|\Psi\rangle &= E[n]\end{aligned} \tag{21}$$

Finally,

$$\mathcal{L} = -\int dx\dot{\phi}n - \frac{1}{2}\int dx(\nabla\phi) \times (\nabla\phi)n_0 - E[n] \tag{22}$$

The equations of motion are written as

$$\dot{\phi} = -\frac{\delta E[n]}{\delta n}, \quad \dot{n} = -\nabla(n_0\nabla\phi) \tag{23}$$

The current operator reads

$$\hat{\mathbf{j}}(x) = -\frac{\mathrm{i}}{2}\sum_{j=1}^{N} \times\left(\delta(x - x_j)\nabla_j + \nabla_j\delta(x - x_j)\right) \tag{24}$$

Thus,

$$\begin{aligned}\mathbf{j}(x) &= \left\langle \Psi_Q | \hat{\mathbf{j}}(x) | \Psi_Q \right\rangle \\ &= i \left\langle \Psi | [\hat{\mathbf{j}}(x), Q] | \Psi \right\rangle \\ &= n_0(x) \nabla \phi(x) \end{aligned} \tag{25}$$

so that

$$\frac{\partial}{\partial t} n = -\nabla \cdot \mathbf{j}, \qquad \frac{\partial}{\partial t} \mathbf{j} = \nabla n_0 \frac{\delta E[n]}{\delta n} \tag{26}$$

We have recovered, for processes involving the participation of many electrons, the same equations that had been obtained for single electron processes.

4.2.1
Normal Modes

Let us come back to the equations of motion

$$\dot{\phi} = \frac{\delta E[n]}{\delta n}, \qquad \dot{n} = -\nabla \cdot (n_0 \nabla \phi) \tag{27}$$

For small fluctuations, we write

$$\frac{\delta E[n]}{\delta n}(x) = \int dy \kappa(x, y) \delta n(y), \quad \text{with}$$

$$\kappa(x, y) = \frac{\delta^2 E[n]}{\delta n(x) \delta n(y)}\Big|_{n=n_0} \tag{28}$$

Looking for oscillating solutions, the ansatz

$$\begin{aligned}\phi(x, t) &= \cos(\omega_\mu t) \phi_\mu(x), \\ \delta n(x, t) &= \sin(\omega_\mu t) \delta n_\mu(x) \end{aligned} \tag{29}$$

leads to

$$\omega_\mu \delta n_\mu = -\nabla(n_0 \nabla \phi_\mu) \tag{30}$$

$$\omega_\mu \phi_\mu = \kappa \delta n_\mu \tag{31}$$

Observe that $\omega_{\bar{\mu}} = -\omega_\mu$ is also an eigenvalue

$$\omega_{\bar{\mu}} \delta n_{\bar{\mu}} = -\nabla(n_0 \nabla \phi_{\bar{\mu}}) \tag{32}$$

$$\omega_{\bar{\mu}} \phi_{\bar{\mu}} = \kappa \delta n_{\bar{\mu}} \tag{33}$$

being $\phi_{\bar{\mu}} = -\phi_\mu$, $\delta n_{\bar{\mu}} = \delta n_\mu$. The normalization condition

$$\int dx \left(\phi_\alpha(x) \quad \delta n_\alpha(x) \right) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \times \begin{pmatrix} \phi_\beta(x) \\ \delta n_\beta(x) \end{pmatrix} = \frac{\omega_\alpha}{|\omega_\alpha|} \delta_{\alpha\beta} \tag{34}$$

may be imposed. An arbitrary perturbation may be expanded in the normal modes, that is,

$$\begin{pmatrix} \phi \\ \delta n \end{pmatrix} = \sum_\alpha c_\alpha \begin{pmatrix} \phi_\alpha \\ \delta n_\alpha \end{pmatrix} \tag{35}$$

where

$$c_\alpha = \frac{\omega_\alpha}{|\omega_\alpha|} \int dx \left(\phi_\alpha(x) \quad \delta n_\alpha(x) \right) \times \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \phi(x) \\ \delta n(x) \end{pmatrix} \tag{36}$$

It may be easily seen that the familiar energy weighted sum rule

$$\sum_{\alpha(\omega_\alpha > 0)} \omega_\alpha c_\alpha^2 = \frac{1}{2} \int dx \, n_0(x) (\nabla \phi)^2 + \frac{1}{2} \int dx \, dy \, \kappa(x, y) \times \delta n(x) \, \delta n(y) \tag{37}$$

is satisfied, which means that the expansion coefficients represent transition amplitudes from the ground state $|\Psi_0\rangle$ to the excited state $|\Psi_\alpha\rangle$,

$$c_\alpha = \left\langle \Psi_\alpha | (Q + P) | \Psi_0 \right\rangle \tag{38}$$

and the absolute values of the frequencies represent excitation energies,

$$|\omega_\alpha| = E_\alpha - E_0 \tag{39}$$

Indeed, for Q and P such that

$$Q = \sum_{j=1}^{N} \phi(x_j),$$

$$\delta n(x) = \mathrm{i}\langle \Psi_0 | [\hat{n}(x), P] | \Psi_0 \rangle \tag{40}$$

we have

$$\begin{aligned}\sum_{\alpha(\omega_\alpha>0)} \omega_\alpha c_\alpha^2 &= \frac{1}{2}\langle \Psi_0 | [Q, [H, Q]] | \Psi_0 \rangle \\ &+ \frac{1}{2}\langle \Psi_0 | [P, [H, P]] | \Psi_0 \rangle \\ &= \frac{1}{2}\langle \Psi_0 | [(P+Q), [H, (P+Q)]] | \Psi_0 \rangle\end{aligned} \tag{41}$$

which is in agreement with

$$\begin{aligned}\sum_\alpha (E_\alpha - E_0) | \langle \Psi_0 | (P+Q) | \Psi_\alpha \rangle |^2 \\ = \frac{1}{2}\langle \Psi_0 | [(P+Q), [H, (P+Q)]] | \Psi_0 \rangle\end{aligned} \tag{42}$$

The main purpose of this exercise has been to state the time-dependent equation of KS with the new interpretation of the currents as transition currents between different states $|\Psi_0\rangle$ and $|\Psi_1\rangle$.

4.2.2
Single Electron Description

In the following, the motion of the nuclei is mainly described by classical trajectories, whereas the density functionals are used to describe the electrons. We shall use atomic units everywhere. We shall denote the densities of electrons, as usual, by n and the corresponding Schrödinger many-particle wave function by $|\Psi(t)\rangle$. As in the standard literature about electronic calculations, we look at the $n(\mathbf{r}, t)$ as evolving from an initial state

$$|\Psi_0\rangle = |\Psi(t_0)\rangle \tag{43}$$

determined by the Schrödinger equation

$$\mathrm{i}\partial\Psi(t)1/\partial t = \hat{H}(t)\Psi(t) \tag{44}$$

$$\hat{H}(t) = \hat{T} + \hat{V}(t) + \hat{U} \tag{45}$$

$$\hat{V}(t) = \sum_{i=1}^{N} v(\mathbf{r}_i, t) \tag{46}$$

$$\hat{T} = \sum_{i=1}^{N} (-1/2)\nabla_i^2 \tag{47}$$

$$\hat{U} = 1/2 \sum_{i,j,i\neq j}^{N} 1/|\mathbf{r}_i - \mathbf{r}_j| \tag{48}$$

so V is an external potential (from nuclei) on electrons and U their mutual repulsion. The last term is by far the most difficult to handle.

It is obviously not small compared to the other terms. The method suggested in Bang and Hansteen (2002), however, facilitates the treatment of just that term.

We shall assume that a local single-particle potential, $v_s[n](\mathbf{r}, t)$, exists, which reproduces the time-dependent density

$$n(\mathbf{r}, t) = \sum_{j=1}^{N} |\phi_j(\mathbf{r}, t)|^2 \tag{49}$$

with orbitals satisfying the KS (or Schrödinger) equation

$$(\mathrm{i}\partial/\partial t + \nabla^2/2 - v_s[n](\mathbf{r}, t))\phi_j(\mathbf{r}, t) = 0 \tag{50}$$

Generally,

$$\begin{aligned}v_s[n](\mathbf{r}, t) &= v(\mathbf{r}, t) \\ &+ \int d^3r' n(\mathbf{r}', t)/|\mathbf{r} - \mathbf{r}'| + v_{\mathrm{xc}}[n](\mathbf{r}, t)\end{aligned} \tag{51}$$

where Eq. (49) *defines* the time-dependent xc potential. A general matrix element of $v_s[n]$, corresponding to a transition from single-particle state i to single-particle state j

$$\langle \phi_j | v_s[n] | \phi_i \rangle \tag{52}$$

can be viewed as a sum over the single-particle constituents (j') of n in the *second* term of Eq. (10)

$$\begin{aligned} v_s[n] &= \int d^3r' n(\mathbf{r}', t)/|\mathbf{r}-\mathbf{r}'| \\ &= \int d^3r' \sum_{j'} |\phi_j'|^2/|\mathbf{r}-\mathbf{r}'| \end{aligned} \tag{53}$$

which can again be considered as the trace of the more general matrix

$$\mathrm{trace}[v_s[n]_{i',j'}] = \mathrm{trace} \int d^3r' \phi_{i'}^* \phi_{j'}/|\mathbf{r}-\mathbf{r}'| \tag{54}$$

Now, as in Bang and Hansteen (2002), the idea is to use the Schrödinger equation, satisfied by all ϕ_m, and then derive from that the continuity equation for single-particle electron densities

$$\partial n(\mathbf{r}, t)/\partial t = -\nabla \mathbf{j}(\mathbf{r}, t) \tag{55}$$

to introduce, instead of the electron transition densities $n_{i,j}$, the corresponding *transition current densities*. In Gross *et al.* (1996) currents are also introduced, but in a more formal way, which does not demonstrate the full calculational advantage of this particular step.

The main point is now to use the time-dependent Schrödinger equation and, as the consequence of it, the continuity equation to eliminate the single-particle densities in favor of similar *longitudinal* single-particle currents

$$\frac{\partial \rho_{ki}}{\partial t} = \mathrm{i}\frac{div_r \mathbf{j}_{ki}}{E_k - E_i}\hbar \tag{56}$$

Since Eq. (55) can be separated in a time- and a space-dependent partial differential equation, a similar elimination can be made here. The single-particle energies E_i, and so on, are replaced here by the separation constants e_i, and so on, so also the denominators $E_{ki}/\hbar^2$ should be replaced by $e_{ki} = e_k - e_i$. Remember $\mathbf{r}_- = |\mathbf{r} - \mathbf{r}'|$.

We may, as explained in Appendix 4.B and Bang and Hansteen (2002), introduce the *longitudinal* currents

$$\mathbf{j}^L(\mathbf{r}) = -\nabla_r \int dr'^3 div_{r'} \mathbf{j}(\mathbf{r}')/|4\pi \mathbf{r}_-| \tag{57}$$

and by a partial integration, we get the general Coulomb matrix element, connecting the transition densities $n_{ki}(\mathbf{r})$, $n_{jh}(\mathbf{r}')$:

$$I_{ki;jh} = \int d^3r (\mathbf{j}_{ki}^L \cdot \mathbf{j}_{jh}^L) 4\pi / e_{ki} e_{jh} \tag{58}$$

and the similar equation for the transverse component $\mathbf{j}_{ki}^T$. Here, again the longitudinal and transverse currents in a spherical system take simple forms, which can be expressed by means of vector spherical harmonics. A longitudinal field can be written as

$$\begin{aligned} \mathbf{L}_{J,M} &= \nabla([\Phi^J]^{\mathrm{long}} Y_{J,M}) \\ &= -\sqrt{(J+1)/(2J+1)} \\ &\quad \times (d/dr - (J+0)/r)[\Phi^J]^{\mathrm{long}} Y_{J,J+1,M} \\ &\quad + \sqrt{(J+0)/(2J+1)} \\ &\quad \times (d/dr + (J+1)/r)[\Phi^J]^{\mathrm{long}} Y_{J,J-1,M} \end{aligned} \tag{59}$$

Analogously, a transverse field can be written as

$$\begin{aligned}\mathbf{T}_{J,M} &= \sqrt{(J+0)/(2J+1)}\\ &\quad\times \big(d/dr - (J+0)/r\big)\\ &\quad\times \big[\Phi^J\big]^{\mathrm{tran}}\, Y_{J,J+1,M}\\ &\quad+ \sqrt{(J+1)/(2J+1)}\\ &\quad\times \big(d/dr - (J+1)/r\big)\\ &\quad\times \big[\Phi^J\big]^{\mathrm{tran}}\, Y_{J,J-1,M} \end{aligned} \tag{60}$$

This is not the only transverse field; any field where the vector spherical harmonics are of the type

$$Y_{J,J,M} \tag{61}$$

is also transverse. But the total Fermion vector density with quantum numbers $[J, l, M]$, where l is different from J, may be written as $\mathbf{A}_{J,l,M} = \mathbf{L}_{J,l,M} + \mathbf{T}_{J,l,M}$. The occurrence of e_{ki}, and so on, in the denominators seems somewhat inconvenient since whenever $e_k = e_i$, these denominators become zero. The corresponding numerators also become zero, however. The construction of a general transition current, containing both longitudinal and transverse components, from the single-particle electron wave functions in a spherical field, was demonstrated in Bang and Hansteen (2002), Appendix 4.A. The vector spherical harmonics, which are eigenfunctions of an angular momentum **J** with magnetic components **M**, given by addition of an orbital angular momentum **L** and a unit angular momentum **S**, eigenvectors **e**, are expressed as

$$\mathbf{Y}_{J,L,M} = \sum_{m,q} Y_{L,m}(\Omega)\, \mathbf{e}_q(Lm1q|JM) \tag{62}$$

A gradient of a single-particle wave function can be written as

$$\begin{aligned}\nabla(\phi_n Y_{\nu,m_n}) &= \phi_{\nu+} Y_{\nu,\nu+1,m_n}\\ &\quad+ \phi_{\nu-} Y_{\nu,\nu-1,m_n}\end{aligned} \tag{63}$$

where we for the moment introduce the convention that a state with quantum number n has orbital angular momentum quantum number named by the corresponding Greek letter ν

$$\begin{aligned}\phi_{\nu+} &= -\sqrt{((\nu+1)/(2\nu+1))}\\ &\quad\times (d/dr - (\nu+0)/r)\phi_n\end{aligned} \tag{64}$$

$$\begin{aligned}\phi_{\nu-} &= +\sqrt{((\nu+0)/(2\nu+1))}\\ &\quad\times (d/dr + (\nu+1)/r)\phi_n\end{aligned} \tag{65}$$

The terms in the current contain the factors

$$Y^*_{\kappa,m_k}\mathbf{Y}_{\iota,\lambda,m_i} = \sum_{JlM} \mathbf{Y}_{J,l,M} X_{JlM} \tag{66}$$

where the factor X is constructed from 6J symbols, which are (for typographical reasons) replaced here and in Bang and Hansteen (2002) by the equivalent U symbols of Jahn.

$$\begin{aligned}&X_{JlM} = \sqrt{(2\kappa+1)(2\lambda+1)/4\pi(2l+1)}(-)^{m_k}\\ &U\big(1,\lambda,J,\kappa;\iota,l\big)\big(\kappa,-m_k,\iota,m_i|J,M\big)\\ &\quad\times\big(\kappa,0,\lambda,0|l,0\big)\end{aligned} \tag{67}$$

The spherical harmonics together with radial functions, $\phi_n(r)$, constitute the single-particle wave functions. To construct the single-particle transition currents, we need a general expression for the gradient of a single-particle wave function (Eqs (63)–(65)) together with the above combination formulas (Eq. (66)). From Eqs (66) and (67) above, we get the general expressions for longitudinal and transverse transition current densities

$$F^L_{J+}\mathbf{Y}_{J,J+1,M} + F^L_{J-}\mathbf{Y}_{J,J-1,M} \tag{68}$$

$$F^T_{J+}\mathbf{Y}_{J,J+1,M} + F^T_{J-}\mathbf{Y}_{J,J-1,M} \tag{69}$$

The sum of all the four last terms should be equivalent to the transition current densities calculated from Eqs (67) and (68), and since the vector spherical harmonics are orthogonal for different values of all three quantum numbers, the same equivalence must be true for the coefficients F. So, from these equations, we get the desired longitudinal current components.

4.2.3
Electron–Nuclei Description

In this chapter as well as in Gross *et al.* (1996), a possible motion of nuclei is also included, and hence one gets a set of coupled equations for electrons and nuclei:

$$\mathrm{i}\frac{\partial\phi_j}{\partial t} = \left(-\frac{1}{2}\nabla_r^2 + v_s[n, n_B](\mathbf{r}, t)\right) \times \phi_j(\mathbf{r}, t), \quad [j = 1, \ldots, N] \tag{70}$$

$$\mathrm{i}\frac{\partial\psi_{A\alpha}}{\partial t} = \left(-\frac{1}{2M_A}\nabla_R^2 + V_s^A[n, n_B](\mathbf{R}, t)\right) \times \psi_{A\alpha}(\mathbf{R}, t) \tag{71}$$

where nucleon α is associated with nucleus A, so that the nuclear densities read

$$n_A(\mathbf{R}, t) = \sum_{\alpha=1}^{N_A} n_{A,\alpha}(\mathbf{R}, t) = \sum |\psi_{A\alpha}(\mathbf{R}, t)|^2 \tag{72}$$

and electron densities

$$n(\mathbf{r}, t) = \sum_{j=1}^{N} |\phi_j(\mathbf{r}, t)|^2 \tag{73}$$

Correspondingly, we have the KS potentials

$$v_s[n, n_B](\mathbf{r}, t) = v_{\text{ext}}(\mathbf{r}, t) + \int d^3r'\frac{n(0\mathbf{r}', t)}{|\mathbf{r} - \mathbf{r}'|} - \sum_{B=1}^{K}\int d^3R\frac{Z_B n_B(\mathbf{R}, t)}{|\mathbf{r} - \mathbf{R}|} + v_{\text{xc}}[n, n_B](\mathbf{r}, t) \tag{74}$$

$$V_s^A[n, n_B](\mathbf{R}, t) = V_{\text{ext}}^A(\mathbf{R}, t) - Z_A\int d^3r' \times \frac{n(\mathbf{r}', t)}{|\mathbf{R} - \mathbf{r}'|} + Z_A\sum_{B=1}^{K}\int d^3R'\frac{Z_B n_B(\mathbf{r}', t)}{|\mathbf{R} - \mathbf{R}'|} + V_{\text{xc}}^A[n, n_B](\mathbf{R}, t) \tag{75}$$

corresponding to Eq. (54).

The important case of $e_i = e_k$ deserves special attention. Here, $div\mathbf{j}_{ki}/e_{ki}$ can be considered a limiting case of $e_k = e_i + d$ for $d \to 0$:

$$(\nabla(\psi_k^*\nabla\psi_i - \psi_i\nabla\psi_k^*))/e_{ki} \tag{76}$$

$$= ((\Delta\psi_i)\psi_k^* - (\Delta\psi_k^*)\psi_i)/e_{ki} \tag{77}$$

$$= bd\psi_i\psi_k^*/e_{ki} = b|\psi_i(r)|^2 \tag{78}$$

The last formulation assumes that we further have $\psi_i = \psi_k$, as in the nondegenerate case of single-particle n, again looking apart from spin. Here, b in atomic units is 1.

Gross *et al.* (1996) proved that different currents

$$\mathbf{j} = \langle\psi(t)|\hat{j}|\psi(t)\rangle \tag{79}$$

and

$$\mathbf{j}' = \left\langle\psi(t)'|\hat{j}|\psi(t)'\right\rangle \tag{80}$$

correspond to different external potentials v, v', and therefore to different solutions of KS equations.

Once the external potentials are fixed, all other matrix elements of one- and two-body operators between different states of the system determined by $\hat{H}$ are also given.

Although it is obvious that a completely quantum mechanical treatment of the electron motion is superior to a classical one, the motion of the nuclei seems fit for use of classical or semiclassical methods, as also realized in Gross *et al.* (1996). From Ehrenfests theorem applied to the nuclear KS equation, the classical trajectory is given as

$$\mathbf{R}^{\text{class}}_{A\alpha} = \left\langle \psi_{A\alpha}(t) | \hat{\mathbf{R}} | \psi_{A\alpha}(t) \right\rangle \tag{81}$$

$$= \int d^3R\mathbf{R}n_{A\alpha}(\mathbf{R}, t) \tag{82}$$

We see that

$$M_A d^2\mathbf{R}^{\text{class}}_{A\alpha}/dt^2(t) = \mathbf{F}_{A\alpha}(t) \tag{83}$$

with

$$\mathbf{F}_{A\alpha}(t) = -\left\langle \psi_{A\alpha}(t) | \nabla_R V^A_s | \psi_{A\alpha}(t) \right\rangle \tag{84}$$

In some cases, the nuclear densities may be rather narrowly peaked distributions around a classical trajectory. Such distributions may, for example, be forced upon a system by introducing appropriate screens in an experiment. The practical difficulties of such a procedure should not prevent us from using it in a theoretical description, since the spreading of wave packets is in many cases shown to be *relatively* small (as exemplified by scattering of slow nuclei in Coulomb fields). Under such circumstances, the equations of motions for the nuclei reduce to

$$M_A \frac{d^2}{dt^2}\mathbf{R}^{\text{class}}_{A\alpha}(t) = -\nabla\left[V^A_{\text{ext}}(\mathbf{R}^{\text{class}}_{A\alpha}, t) - \int d^3r \frac{Z_A n(\mathbf{r}, t)}{|\mathbf{R}^{\text{class}}_{A\alpha} - \mathbf{r}|} + \sum_{B=1}^{K}\sum_{\beta=1}^{N_B} \frac{Z_A Z_B}{|\mathbf{R}^{\text{class}}_{A\alpha} - \mathbf{R}^{\text{class}}_{B\beta}|}\right] \tag{85}$$

and the electronic KS equations simplify to

$$\mathrm{i}\frac{\partial \phi_j(\mathbf{r}, t)}{\partial t} = \left[-\nabla^2/2 + v_{\text{ext}}(\mathbf{r}, t) + \int d^3r' \frac{n(\mathbf{r}', t)}{|\mathbf{r} - \mathbf{r}'|} - \sum_{B=1}^{K}\sum_{\beta=1}^{N_B} \times \frac{Z_B}{|\mathbf{R}^{\text{class}}_{\mathrm{B}\beta}(t) - \mathbf{r}|} + v_{\text{xc}}\left[n, \mathbf{R}^{\text{class}}_{\mathrm{B}\beta}(t)\right](\mathbf{r}, \mathbf{t})\right]\phi_j(\mathbf{r}, t) \tag{86}$$

Equations (81) and (82) are evidently coupled, and they must be simultaneously solved, thus achieving the trajectory.

Such unique trajectories are, however, not the only possibility. In potentials that give rise to classical crossing points or branching points, the quantum mechanical wave packets may be split into two or more, centered around different trajectories, even giving rise to interference effects. In such cases, the classical treatment becomes insufficient, and a solution of the original Eqs 70) and 71 cannot be avoided, at least in the neighborhood of classical crossing points or branching points. Let us, however, first return to the simple case of one classical trajectory.

The interaction between nuclei and electrons is given here by a matrix of four indices (see Appendix 4.B and Bang and Hansteen (2002) Eqs (7)−(11) $[I_C]_{ki,jh}$ of which the first two (ki) correspond to the initial and final (external, center of mass (CM)) states of the nuclei, the last two (jh) to the initial and final states of electrons.

$$[I_C]_{ki,jh} C \int d^3r \left(\mathbf{j}(\mathbf{r})^{\mathrm{long}}_{ki} / e_{ki} \mathbf{j}(\mathbf{r})^{\mathrm{long}}_{jh} / e_{jh} \right) \tag{87}$$

The factor, C, in front of the integral is a scaling factor due to the transition from the nuclear coordinates to electronic ones used in the partial integration of (see Appendix 4.B and Bang and Hansteen (2002) Eq. (7))−(11). Each nucleus obviously contributes a current component that has a direction antiparallel to the momentum transfer in the transition current $\mathbf{j}_{jh}$ of the electron considered. The nuclear current component is contained in a sphere with radius approximately $1/M$ in atomic units. The total contribution of nuclei of one sort to the force on one electron undergoing a transition between states j, h is just

$$\sum_{B=1}^{K} \sum_{\beta=1}^{N_B} I_C \tag{88}$$

We now return to the case of trajectories with crossing points or branching points. We shall try to solve the coupled equations (Eqs (70) and (71)) with the boundary condition that at the time $t_1 < t_c$ the solution for the electrons is ϕ_a. Here, t_c is the time of the collision. For $t_2 > t_c$, the electronic solution is $a\phi_a + b\phi_b$ with $|a|^2 + |b|^2{=}1$. We see that $|a/b| = |\psi_{A\alpha}|/|\psi_{B\beta}|$, when $t_1 \ll t_2$. We have sketched here how an electron transfer may be described. It can either be a transfer between two atoms in one molecule or between atoms in different molecules. In the first case, we may again directly introduce the classical approximation for the motion of the center of mass of the nuclei. In the second case, the coupled equations must be used in the neighborhood of crossing points and branching points.

To get the complete description of wave functions and densities of *electrons*, we must further add the remaining terms of Eq. (81) to the Hamiltonian. It should be noted, though, that the terms denoted v_{xc} have been treated particularly in Bang and Hansteen (2002), Chapter 5. We see that the attraction of nuclei of one sort, lying on a single curve, can be simulated by a classical current contained in a thin (order of magnitude of width $1/M$) conductor on the same curve interacting with the electron transition current density with a contact interaction. The method can be extended to other nuclei, not lying on the curve, by using the methods of crossing and branching points, mentioned earlier. It seems that these methods are particularly fit for long, chain-like molecules. It is evident that the same longitudinal electron transition current densities are met in the nucleus–electron interaction and in the electron–electron interaction. These currents or rather their spherical components should therefore be stored, so much more as each transition current density can occur in several matrix elements.

4.3
Relativistic Electrons

Another interesting problem is connected with relativistic effects. The kinetic energies of electrons in atoms and molecules are only in the neighborhood of the very heaviest nuclei of such an order of magnitude that relativistic effects come into play in a direct way. However, magnetic interactions can always be considered as relativistic effects and a proper treatment as that may be, not only consistent but also the simplest way, as shown later.

The electric current density in the Dirac theory of relativistic electrons is given as the three-vector part of a four-vector:

$$s_{\nu=1,2,3,4} = \mathrm{i}ec\bar{\boldsymbol{\Psi}}\gamma_\nu\boldsymbol{\Psi} \tag{89}$$

$$s_{k=1,2,3} = \mathrm{j}_k; \ldots s_4 = \mathrm{i}c\rho \tag{90}$$

The 4-vector s_ν can be decomposed

$$s_\nu = s_\nu^C + s_\nu^P \tag{91}$$

so that the space part of s^C, the *conduction current* is of the same form as the nonrelativistic expression for j_k. The remaining part is the *polarization current* s_k^P. The Dirac State vector is here

$$\boldsymbol{\Psi} = \sum_\nu \psi_\nu \mathbf{e}_\nu \tag{92}$$

and the conjugate

$$\bar{\boldsymbol{\Psi}} = \boldsymbol{\Psi}^\dagger\alpha_0 = \sum_\nu \bar{\psi}_\nu \mathbf{e}_\nu^{\mathrm{T}} \tag{93}$$

where $\mathbf{e}_\nu^{\mathrm{T}}$ stands for the transposed of $\mathbf{e}_\nu$. Here, the ψ are functions of x, y, z, t and the $\mathbf{e}$ are orthogonal unit vectors in spin space. The Dirac equation is now

$$\begin{aligned}&\left[\boldsymbol{\alpha}_0 mc+\sum_k \boldsymbol{\alpha}_k\left(\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial x_k}-\frac{e}{c}A_k\right)+e\phi\right]\boldsymbol{\Psi}\\ &= \left[\boldsymbol{\alpha}_0 mc\boldsymbol{\Psi}+\boldsymbol{\alpha}\left(\frac{\hbar}{\mathrm{i}}\nabla\boldsymbol{\Psi}-\frac{e}{c}\mathbf{A}\boldsymbol{\Psi}\right)+e\phi\boldsymbol{\Psi}\right]\\ &= -\frac{\hbar}{\mathrm{i}c}\frac{\partial\boldsymbol{\Psi}}{\partial t}\end{aligned} \tag{94}$$

Here, $\boldsymbol{\alpha}$ is a vector with components α_k and $\alpha_0 = \rho_0$, the familiar Dirac matrices. The continuity equation in this relativistic case becomes

$$\frac{\partial(\boldsymbol{\Psi}^\dagger\boldsymbol{\Psi})}{\partial t} = -c\; div(\boldsymbol{\Psi}^\dagger\alpha\boldsymbol{\Psi}) \tag{95}$$

Any field in three dimensions may be split in a longitudinal and a transverse part

$$\mathbf{j} = \mathbf{j}^{\mathrm{long}} + \mathbf{j}^{\mathrm{tran}} \tag{96}$$

$$\begin{aligned}\mathbf{j}^{\mathrm{long}} &= -\nabla_r \int d^3r' \frac{div_{r'}\mathbf{j}(r')}{4\pi|\mathbf{r}-\mathbf{r}'|};\\ \mathbf{j}^{\mathrm{tran}} &= -\mathrm{curl}_r \int d^3r' \frac{\mathrm{curl}_{r'}\mathbf{j}(r')}{4\pi|\mathbf{r}-\mathbf{r}'|}\end{aligned} \tag{97}$$

Let the transition current field be given by its relativistic form in spherical coordinates

$$\mathbf{j}_{ki}^1 = iec\bar{\boldsymbol{\Psi}}_k\boldsymbol{\gamma}\boldsymbol{\Psi}_i \tag{98}$$

with the spinors

$$\boldsymbol{\Psi}_{i,J_i,l_i,M_i} = \mathcal{F}(r,t)_i\mathcal{Y}_{lm}\chi_\mu(lm1/2\mu|l1/2JM) \tag{99}$$

In the presence of a magnetic field, the one-electron current density takes the form

$$\begin{aligned}\mathbf{j}_{ki}^{\mathrm{A}} = \frac{\hbar}{2mi}\Big(&\boldsymbol{\Psi}_k^*\boldsymbol{\nabla}\boldsymbol{\Psi}_i - \boldsymbol{\Psi}_i\boldsymbol{\nabla}\boldsymbol{\Psi}_k^*\\ &-\frac{2ie}{c\hbar}\mathbf{A}\boldsymbol{\Psi}_k^*\boldsymbol{\Psi}_\mathrm{i}\Big)\end{aligned} \tag{100}$$

or in spherical form

$$\mathbf{j}^A = C_1\mathbf{j}_{JlM}^1 - C_2(\mathbf{A}\rho)_{JlM} \tag{101}$$

where the Cs are given in Eqs (71)–(73) and $(\mathbf{A}\rho)_{JlM}$ is given by the definition of vector spherical harmonics in the expression

$$\mathbf{A}\rho = \mathbf{A}\sum_{mq} Y_{lm}\mathbf{e}_q(lm1q|l1JM) \tag{102}$$

Remember that $\mathbf{A}$ is a vector and ρ a scalar. The *longitudinal* part of these vectors is now found in the same way as in Bang and Hansteen (2002). In order to treat the magnetic interactions between electrons and between nuclei as well as between nuclei and electrons, we shall look at the transverse transition current densities. At the same time we shall introduce a new form of these current densities, appropriate for the interactions with magnetic fields:

$$\mathbf{s}_{ki} = \mathbf{j}_{ki} - \frac{e^2}{c}\,\mathbf{A}\boldsymbol{\Psi}_k^*\boldsymbol{\Psi}_i \qquad (103)$$

$$\mathbf{s}^{\mathrm{T}} = \mathrm{curl}_r \int d^3\mathbf{r}'\,\frac{\mathrm{curl}_{r'}(\mathbf{s}(\mathbf{r}') - \frac{e^2}{c}\,\mathbf{A}\rho(\mathbf{r}'))}{4\pi\,|\mathbf{r}-\mathbf{r}'|} \qquad (104)$$

with

$$\rho(\mathbf{r}') = \Psi_k^*(\mathbf{r}')\Psi_i(\mathbf{r}') \qquad (105)$$

The new expression for the current density gives the same continuity expression as the previous one. The curl $\mathbf{A} = \mathbf{H}$ gives the connection to the magnetic field (which could originate from other electrons or nuclei, static or in transitions in general $(j{-}h)$). The magnetic interaction energy $[ki, jh]$ can again be written by means of the integral of a scalar product of two transition densities, $\mathbf{j}_{ki}^{T}$ and $\mathbf{j}_{jl}^{T}$. Note, however, that the two proportionality constants in the electric and magnetic energies, expressed by means of currents, are different. To show the effect of the method of giving the matrix elements by means of currents, we shall start from the final result and work backward

$$I_M = \int d^3r(\mathbf{j}(\mathbf{r})_{ki}^{\mathrm{T}}\,\mathbf{j}(\mathbf{r})_{jh}^{\mathrm{T}})C_{ki,jh} \qquad (106)$$

where

$$\mathbf{j}(\mathbf{r})^{\mathrm{T}} = \mathrm{curl}_r \int d^3\mathbf{r}'\,\frac{\mathrm{curl}_{\mathbf{r}'}\mathbf{j}(\mathbf{r}')}{4\pi\,|\mathbf{r}-\mathbf{r}'|} \qquad (107)$$

(Note, however, that the constant C in the magnetic matrix elements is different from the corresponding constant in the electric matrix element).

Note also that the current in the magnetic matrix element must include the electromagnetic potential $\mathbf{A}$, and that the transverse current densities have the same form as the source terms of *electric* multipole radiation (and vice versa with magnetic multipole radiation).

The occurrence of curl$\mathbf{j}$ in Eq. (107) and elsewhere may seem inconvenient, since this expression is not part of the common formulas, derived from Maxwell's or Schrödingers equations. Introducing

$$\mathbf{j}^{\mathrm{T}} = \mathbf{j} - \mathbf{j}^{\mathrm{L}} \qquad (108)$$

this formal problem is avoided.

4.4
The General Description for the Electronic Evolution Equations

In the earlier chapters, we introduced the transition currents for calculation of electric and magnetic matrix elements. The advantages of introducing that, as suggested here, are several:

First, this procedure, or these procedures, are known to lead to similar simplifications compared to Hartree–Fock methods as those of density methods Gross *et al.* (1996). Secondly, the currents give a simpler connection to dynamic problems, than do static densities. Thirdly, we have just shown that even magnetic interactions can, by means of the currents, be expressed in a particularly simple way. And fourthly, as already shown in Bang and Hansteen (2002), special methods for calculations of integrations in matrix elements permit us to avoid unwieldy, sometimes even singular, integrals.

As a fifth point, it should be mentioned that the classification used here, of dividing vector functions in longitudinal and transverse vector functions, is a very natural one, when thinking of local, as distinguished from global, classifications.

Further simplifications may be obtained by using a special representation for the wave functions, and consequently also for the currents. We shall here, neglecting spin, use the basis of wave functions of electrons around a point nucleus of infinite mass, and charge eZ, factorizing the wave function in the usual way:

$$\phi = 1/r\chi_l(r)Y_{lm}(\theta, \phi) \tag{109}$$

$$\chi_l'' + (-\gamma^2 + 2\gamma\kappa/r - l(l+1)/r^2)\chi_l = 0 \tag{110}$$

$$2mE/\hbar^2 = -\gamma^2 \tag{111}$$

$$\kappa = Ze^2m/(\hbar^2\gamma) \tag{112}$$

and introduce a dimensionless radial variable

$$z = r/\alpha_0 = r\gamma = rZ/n \tag{113}$$

Now according to Appendix 4.B and Bang and Hansteen (2002), the transition current density in this representation has matrix elements given by vector spherical harmonics, cf. Eqs (59)–(69).

$$\begin{aligned}[\mathbf{j}_{ki}]_{JlM} &= [\phi(r)_k Y^*_{l_k m_k} \nabla\phi(r)_i Y_{l_i m_i}]_{JlM} \\ &\quad - [\phi(r) - \mathrm{i} Y_{l_i m_i} \nabla\phi(r)_k Y^*_{l_k m_k}]_{JlM}\end{aligned} \tag{114}$$

The radial part of $\mathbf{j}_{ki}$ can be written as

$$F_{\alpha,J,l,M} = F^{\alpha} - F^{\alpha'} \tag{115}$$

Here, α' means that in the second term of $\mathbf{j}$ the roles of i, k are interchanged, although ϕ_k still occurs with complex conjugation cf. Bang and Hansteen (2002).

So equating the coefficients of $\mathbf{Y}_{J,l,M}$ in the transition current densities

$$\mathbf{j} = \sum_{J,l=J\pm1,M} F_{\alpha,J,l}\mathbf{Y}_{J,l,M} \tag{116}$$

and those in $\mathbf{L}_{JM} + \mathbf{T}_{JM}$, we get

$$(d/dr - J/r)u_J = F_{\alpha,J,J+1} \tag{117}$$

$$(d/dr + (J+1)/r)v_J = F_{\alpha,J,J-1} \tag{118}$$

where

$$\begin{aligned}u(r)_J &= \sqrt{J/(2J+1)}\Phi_J^T \\ &\quad - \sqrt{(J+1)/(2J+1)}\Phi_J^L\end{aligned} \tag{119}$$

$$\begin{aligned}v(r)_J &= \sqrt{J/(2J+1)}\Phi_J^L \\ &\quad + \sqrt{(J+1)/(2J+1)}\Phi_J^T\end{aligned} \tag{120}$$

The differential equations (Eqs 117 and 118) are to be solved to get $u(r)$, $v(r)$. Although this is not difficult for small molecules, in the case of larger molecules, it could be time consuming. An alternative method could be to expand the $F_{\alpha,J,l}$ on functions that lead to algebraic solutions of Eqs (117) and (118), as would be the case if the transition currents were currents in atoms with a fixed nucleus of point charge. Regardless of how Eqs (117) and (118) are solved, the next step should be to invert the equations to get the $\mathbf{\Phi}_J^L$, $\mathbf{\Phi}_J^T$ from which the respective electric, and magnetic electron, interactions are constructed:

$$\Phi_J^L = v_J\sqrt{J/(2J+1)} - u_J\sqrt{(J+1)/(2J+1)} \tag{121}$$

$$\Phi_J^T = u_J\sqrt{J/(2J+1)} + v_J\sqrt{(J+1)/(2J+1)} \tag{122}$$

These equations constitute the main framework for solving the evolution of the electronic structure in molecules. The equations are in time, meaning that the parameter, or the argument, of the two electron evolution equations can be considered as any variable since the potentials, and therefore also the Fs, are time dependent. The time dependence is basically through the nuclear dynamics, see Section 4.2.3. Ehrenfest theorem of the time evolution of the wave packet can also be applied in this case guaranteeing that the time evolution of $u(r, t)$, $v(r, t)$ will have an extra time parameter in the exponent. Such a parameter, however, disappears in the calculation of the density and the density of the currents. Overall, we build the formalism for the transition currents between different orbitals in time. Thus, we can write

$$u(r, t)_J = \sqrt{J/(2J+1)}\Phi(r, t)_J^T - \sqrt{(J+1)/(2J+1)}\Phi_J^L(r, t) \tag{123}$$

$$v(r, t)_J = \sqrt{J/(2J+1)}\Phi(r, t)_J^L + \sqrt{(J+1)/(2J+1)}\Phi_J^T(r, t) \tag{124}$$

and similarly

$$\Phi(r, t)_J^L = v_J\sqrt{J/(2J+1)} - u_J\sqrt{(J+1)/(2J+1)} \tag{125}$$

$$\Phi(r, t)_J^T = u_J\sqrt{J/(2J+1)} + v_J\sqrt{(J+1)/(2J+1)} \tag{126}$$

Equations (117) and (118), and the ones just mentioned, are of the linear type, which are described and, generally, solved in the comprehensive book by Kamke (1983). With the notation given in Kamke (1983), both the equations can be written as

$$dy/dx + f(x)y = g_y(x) \tag{127}$$

with the solution

$$y = \exp(-F)[a + \int_b^x g(x')\exp(F)dx'] \tag{128}$$

Here,

$$F(x)^u = F(r) = -\int_b^r (J+1)/r'dr' = r^{-J}/b^{-J} \tag{129}$$

whereas in the next equation

$$F(x)^v = \int_b^r (J)/r'dr' = r^{-J+1}/b^{-J+1} \tag{130}$$

The lower limit of the two integrals is close to zero. Next, we will write

$$g_u(r) = \exp(-yr)G_u(r) \tag{131}$$

$$g_v(r) = \exp(-yr)G_v(r) \tag{132}$$

and we shall suppose that the G-functions can be Taylor expanded around $r = 0$. Hence,

$$G_u = \sum_{n=0}^{N} A_n^u r^n; \quad G_v = \sum_{n=0}^{N} A_n^v r^n \tag{133}$$

The equations for $u_J(r)$, $v_J(r)$ have the solution

$$v_J = r^{-(J+1)}\int_0^r dr' r'^{(J+1)} F^v_{\alpha,J,J-1}(r') \tag{134}$$

and correspondingly for u_J, but with $A = -J$, which we may solve in a similar way, assuming that J is different from 1. (The case, $J = 1$, can be dealt with by a limiting procedure, $J \to 1$). Altogether, we get for the first term, for example, in the Taylor expansion of F

$$A_1^{\text{Taylor}} = F_0\big(r^J r^{-(J+1)} - b^J b^{-(J+1)}\big); \quad \text{where} \quad b \to 0 \tag{135}$$

We see that the electron–electron interaction strength corresponding to the two transitions

$$\alpha = [i, k] \tag{136}$$

and

$$\beta = [j, h] \tag{137}$$

is given directly by the scalar product of the corresponding transition current densities, integrated over all space. Antisymmetry of the wave functions simply means that the Pauli exclusion principle should be implemented on each pair of states. In Bang and Hansteen (2002), Eq. (7), another way of writing j^L_{ki} was suggested:

$$j(r)^L_{ki} = \int dr'^3 \nabla_{r'} (\rho(r)_{ki} \mathrm{i} E_{ki}/(4\pi \mid r_- \mid) \tag{138}$$

We see that with this formulation, inconvenient energy denominators E_{ki} are removed.

The transition matrix element from one orbital Ψ_0 to another Ψ_1 is described by

$$\left\langle \Psi_0 \mid \hat{\mathbf{j}}(r) \mid \Psi_1 \right\rangle \tag{139}$$

where the transition current $\mathbf{j}$ is given by Eq. (116). Note, however, that the definition (24), involving the explicit dependence on $\mathbf{r}$ is familiar in nuclear physics applications and may also lead to useful information. Let us consider a state vector $|\Psi(t)\rangle$ of some molecule, containing an admixture of excited states, $|\Psi(t)\rangle = e^{-\mathrm{i}E_0 t}|\Psi_0\rangle + c_1 e^{-\mathrm{i}E_1 t}|\Psi_1\rangle$, so that it is nonstationary. The current $\mathbf{j}(r, t) = \langle \Psi(t)|\hat{\mathbf{j}}(\mathbf{r})|\Psi(t)\rangle$ is clearly time dependent. Its Fourier components in time are given by the matrix elements of $\hat{\mathbf{j}}(\mathbf{r})$. Indeed, $\mathbf{j}(\mathbf{r}, t) = c_1\langle \Psi_0|\hat{\mathbf{j}}(\mathbf{r})|\Psi_1\rangle e^{-\mathrm{i}(E_1-E_0)t} + c_1^*\langle \Psi_1|\hat{\mathbf{j}}(\mathbf{r})|\Psi_0\rangle e^{\mathrm{i}(E_1-E_0)t}$, since $\langle \Psi_0|\hat{\mathbf{j}}(\mathbf{r})|\Psi_0\rangle = 0$. We have assumed $|c_1| \ll 1$, so that we have neglected quadratic terms in c_1. The excited states determine the current through the matrix elements $\langle \Psi_0|\hat{\mathbf{j}}(\mathbf{r})|\Psi_1\rangle$, which, therefore, are called transition currents.

4.5
Multielectronic Interaction and Collective Phenomena

Up to now, we have basically been treating the electronic media as single electronic equations of motion. If we should extend the single-particle description to a many-body framework, we should expect to have to deal with collective phenomena as far down as to collective coordinates.

4.5.1
Collective Coordinates and Plasmons

In this respect, classical dynamics may prove to be a useful tool for the description of collective phenomena involving a very large number of electrons, provided the Pauli Principle is properly taken into account, as explained in the sequel. As an example of the approach we have in mind, we consider a two-components system, namely, a large number of positive ions and a large number of electrons in interaction, in appropriate amounts to ensure charge neutrality. For each species, the Vlasov equation

$$\frac{\partial f}{\partial t} + \{h, f\} = 0 \tag{140}$$

describes the respective dynamics. Here, $f = f(\mathbf{r}, \mathbf{p}, t)$ denotes the familiar distribution function $h = h(\mathbf{r}, \mathbf{p}, t)$ – the single-particle hamiltonian – and $\{h, f\} = \sum_{i=1}^{3} \left(\frac{\partial h}{\partial x_i}\frac{\partial f}{\partial p_i} - \frac{\partial f}{\partial p_i}\frac{\partial h}{\partial x_i}\right)$ – the familiar Poisson bracket. Everything is classical, except that f is supposed to assume only two possible values, 0 and 1, according to the Pauli Principle requirements. It is easy to see that if this property is imposed at time $t = 0$, the dynamics will take care to preserve it forever.

We write $f = f_0 + \delta f$, $h = h_0 + \delta h$, where f_0 and h_0 are, respectively, the equilibrium distribution function and hamiltonian (so that $\{h_0, f_0\} = 0$) and δfandδh are the corresponding fluctuating parts. We denote by S the generator of the fluctuations, that is, $\delta f = \{S, f_0\}$. It has been shown that Nemes, Piza and da Providencia (1987); Brink and da Providencia (1989); Nielsen and da Providencia (1989) for small fluctuations, the Vlasov equation may be replaced by the equation for the time evolution of the generator S:

$$\frac{\partial S}{\partial t} + \{S, h_0\} + \delta h = 0 \tag{141}$$

Assuming an interaction of electrostatic origin (Coulomb), the dynamics of positive ions and electrons are described, respectively, by

$$\frac{\partial S_1}{\partial t} + \{S_1, h_0^{(1)}\} + \delta h^{(1)} = 0$$
$$\frac{\partial S_2}{\partial t} + \{S_2, h_0^{(2)}\} + \delta h^{(2)} = 0 \tag{142}$$

We may treat the positive ions nonrelativistically, because they are heavy, and the electrons relativistically, because they are light, so we write, for positive ions, $h_0^{(1)} = p^2/(2M)$, and for electrons $h_0^{(2)} = \sqrt{p^2 + m^2}$. For neutral matter, the density of positive ions must be the same as the density of electrons, so we assume $f_0^{(1)} = f_0^{(2)} = \theta(p^2 - p_F^2)$, where p_F is the common Fermi momentum. Due to translation invariance we assume that

$$S_1 = e^{\mathrm{i}(\omega t - \mathbf{k}\cdot\mathbf{x})}\hat{S}_1(p, \mathbf{k}\cdot\mathbf{p})$$
$$S_2 = e^{\mathrm{i}(\omega t - \mathbf{k}\cdot\mathbf{x})}\hat{S}_2(p, \mathbf{k}\cdot\mathbf{p}) \tag{143}$$

where $p = |\mathbf{p}|$, that is, a plane wave with wave vector $\mathbf{k}$. Then, we may write

$$\delta h^{(1)} = \frac{4\pi e^2}{k^2}(\rho_1 - \rho_2)$$
$$\delta h^{(2)} = \frac{4\pi}{k^2}e^2(\rho_2 - \rho_1) \tag{144}$$

$$\rho_1 = \int d^3p \left\{S_1, f_0^{(1)}\right\}$$
$$\rho_2 = \int d^3p \left\{S_2, f_0^{(2)}\right\} \tag{145}$$

where $k = |\mathbf{k}|$. Notice that

$$\begin{aligned}\int d^3p\{S_j, f_0\} &= -2\pi i \int p^2\, dp\, d\cos\theta\, 2kp \\ &\quad \times \cos\theta\, \delta(p^2 - p_F^2) \\ &\quad \times \hat{S}_j(p, \cos\theta) \\ &= -2\pi i \int d\cos\theta \cos\theta p_F^2 \\ &\quad \times \hat{S}_j(p_F, \cos\theta)\end{aligned} \tag{146}$$

Thus,

$$\left(\omega - \frac{kp\cos\theta}{M}\right)\hat{S}_1 = \frac{8\pi^2 e^2}{k} \times \int d\cos\theta\ \cos\theta(\hat{S}_1 - \hat{S}_2) \tag{147}$$

$$\left(\omega - \frac{kp\cos\theta}{h_{0F}}\right)\hat{S}_2 = \frac{8\pi^2e^2}{k} \times \int d\cos\theta\, \cos\theta(\hat{S}_2 - \hat{S}_1) \qquad (148)$$

Here, θ denotes the angle between the vectors $\mathbf{p}$ and $\mathbf{k}$, and $h_{0F} = h^{(2)}(p_F)$ is the electron Fermi energy. For $\omega > kp_F/h_{0F}$, a collective mode exists such that

$$S_2 = \frac{C_2}{\omega - \frac{kp_F\cos\theta}{h_{0F}}} \qquad S_1 = \frac{C_1}{\omega - \frac{kp_F\cos\theta}{M}} \qquad (149)$$

where C_1, C_2 are constants to be determined. Since, for $|\omega| > a$,

$$a\int_{-1}^{1} d\cos\theta\, \cos\theta\, \frac{1}{\omega - a\cos\theta} = -2 + \frac{\omega}{a}\log\frac{\omega + a}{\omega - a} \qquad (150)$$

the previous ansatz leads to

$$C_1 = \frac{8\pi^2e^2}{k^2p_F}\left(MC_1\left(-2 + \frac{\omega M}{kp_F}\log\frac{1 + \frac{kp_F}{\omega M}}{1 - \frac{kp_F}{\omega M}}\right) - h_{0F}C_2\left(-2 + \frac{\omega h_{0F}}{kp_F}\log\frac{1 + \frac{kp_F}{\omega h_{0F}}}{1 - \frac{kp_F}{\omega h_{0F}}}\right)\right) \qquad (151)$$

$$C_2 = -\frac{8\pi^2e^2}{k^2p_F} \times \left(MC_1\left(-2 + \frac{\omega M}{kp_F}\log\frac{1 + \frac{kp_F}{\omega M}}{1 - \frac{kp_F}{\omega M}}\right) - h_{0F}C_2\left(-2 + \frac{\omega h_{0F}}{kp_F}\log\frac{1 + \frac{kp_F}{\omega h_{0F}}}{1 - \frac{kp_F}{\omega h_{0F}}}\right)\right) \qquad (152)$$

The following dispersion relation is obtained:

$$1 = \frac{8\pi^2e^2}{k^2p_F}\left(M\left(-2 + \frac{\omega M}{kp_F}\log\frac{1 + \frac{kp_F}{\omega M}}{1 - \frac{kp_F}{\omega M}}\right) + h_{0F}\left(-2 + \frac{\omega h_{0F}}{kp_F}\log\frac{1 + \frac{kp_F}{\omega h_{0F}}}{1 - \frac{kp_F}{\omega h_{0F}}}\right)\right) \qquad (153)$$

Approximately,

$$1 \approx \frac{2}{3}\frac{8\pi^2e^2p_F}{\omega^2}\left(\frac{1}{M} + \frac{1}{h_{0F}}\right) \qquad (154)$$

Finally, the plasmon frequency reads

$$\omega_{pl}^2 = \frac{2}{3}8\pi^2e^2p_F\left(\frac{1}{M} + \frac{1}{h_{0F}}\right) \qquad (155)$$

4.5.2
Landau Damping

For $|\omega| < a$,

$$a\mathcal{P}\int_{-1}^{1} d\cos\theta\, \cos\theta\, \frac{1}{\omega - a\cos\theta} = -2 + \frac{\omega}{a}\log\frac{\omega + a}{|\omega - a|} \qquad (156)$$

where $\mathcal{P}$ denotes the principal value.

If $kp_F/M < \omega < kp_F/h_{0F}$, Landau damped modes of the form

$$S_2 = \delta\left(\omega - \frac{kp_F\cos\theta}{h_{0F}}\right) - \frac{C}{\omega - \frac{kp_F\cos\theta}{h_{0F}}} \qquad S_1 = \frac{C}{\omega - \frac{kp_F\cos\theta}{M}} \qquad (157)$$

are also easily obtained. These modes are associated with low-energy excitations.

In figure 4.1 the plasmon frequency is shown as a function of the wave vector. The full line represents the plasmon frequency versus the wave vector, Eq. (153), in arbitrary units. The

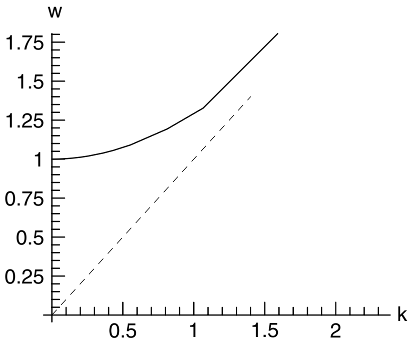

Fig. 4.1 The plasmon frequency, versus the wave vector, in arbitrary units.

dashed line represents the boundary of the region where Landau damping is present, i.e., for a given k, the frequency of the Landau damped mode, Eq. (157), extends continuously from zero to the dashed line.

The formalism described here in this chapter is interesting in the context of evolution equations for electrons and positive ions and as an extension to the single-particle treatment given in the former chapters. The extension has introduced distributions and collective coordinates as in a condensed matter framework. The relevance for such a description is found not only in many-body treatments of plasma systems but also in metallic clusters and could be applied to even metallic molecular clusters or even metalloproteins.

The systems having an amount of positive and negative charges such as positive ions and electrons within a large biomolecule are found in the repair protein, metallothionein, with many metal ions (up to 12 Zn-ions per protein) in a protein pocket containing many negative charges. This protein works as a buffer for Zn-ion signaling inside the cell with a diffusion equation that looks similar to the equations for S_1, S_2. Such a diffusion equation will also make use of distributions and write time derivative of concentrations $[\mathrm{Zn}^+]$ as $d[\mathrm{Zn}^+]/dt = DF([\mathrm{Zn}^+], \nabla[\mathrm{Zn}^+])$.

4.6
Molecular Calculations

In this chapter, we shall show how one can make use of Gaussian wave functions since they are the standard tools in the usual DFT programs that we wish to compare our theory with.

We shall first show some general tools for molecular calculations before going to analyze specific molecules. Basically, we are interested in a proof of the completeness of the current formalism of DFT, or what we call current functional theory (CFT).

When treating molecules, we shall first look at one electron interacting with several nuclei. Later, we shall look at the repulsion between electrons. Since the mass of nucleons is much larger than that of electrons, we shall describe the motion of nuclei classically and also neglect internal excitations corresponding to nuclear dynamics j, h given earlier, so we look at an electronic excitation from the level i with quantum numbers n, m, l and write the wave function

$$\chi_{nlm}(\mathbf{r}+\mathbf{R}) \tag{158}$$

so connected with a nucleus at **R**. It is by the nuclear Coulomb attraction moved to an orbit with quantum number n', m', l' round a nucleus at the origin. So the Coulomb transition matrix element is

$$\int d^3r\chi_{n'm'l'}(\mathbf{r})^* Z/r\chi_{nml}(\mathbf{r}+\mathbf{R}) \tag{159}$$

In the formulation of Shibuya and Wulfman (1965), this amplitude of exciting a state k from a transported χ_i is denoted as

$$\int d^3r\chi_i(\mathbf{r}+\mathbf{R}) = \sum_j \chi_i(\mathbf{r})S(\mathbf{R})^j_i \tag{160}$$

where, in agreement with the earlier,

$$S^j_i = \int d^3r\chi_i(\mathbf{r})^* Z/r\chi_j(\mathbf{r}+\mathbf{R})/k^2 \tag{161}$$

and where the last factor is because of normalization.

The interaction matrix between one electron with transition density $\rho(\mathbf{r})_{\tau'\tau}$ and several nuclei densities, $\rho(\mathbf{X}_a)_{T'T}$, is

$$S_{\tau'\tau} = \sum_a \int \rho_{\tau'\tau} d^3r 1/|r-X| \int \rho_{T'T} \tag{162}$$

Here, we follow (Avery and Shim, 2001) and let τ denote the center of the position of the electron density (of course $= X_a$) as well as its spherical quantum numbers. From the Schrödinger equation of the electron, as well as its consequence, the continuity of transition time-dependent transition densities as in Bang and Hansteen (2002) we arrive at the expression for $S_{\tau'\tau}$

$$S_{\tau'\tau} = \int_j L_{\tau'} J_{T'T}\hbar^2 4\pi d^3r/(E_{\tau'\tau}E_{T'T}) \tag{163}$$

where L stands for the longitudinal component of the transition probability. Since the nuclei are so much heavier than the electrons, we may approximate $J_{T'T}$ with a classical density of point charges, whereas $j^L_{\tau'\tau}$ is still the quantum density of longitudinal currents. Now, the $S_{\tau'\tau}$ obviously forms a representation of the group of translations based on Sturmian bound state densities, for example, as used by Avery and Shim (2001). The character of this group is just identical to the densities of currents as defined by Gross *et al.* (1996) . Thus, the completeness of the currents and densities is readily seen.

4.6.1
Gaussian Integration

Next we return to the Gaussian integration tools.

With a Gaussian expansion for the wave functions, it is natural to start from expression (136)

$$j(r)^L_{ki} = \text{grad}_r \int dr'^3 \rho(r')_{ki} \mathrm{i} E_{ki}/((r_-)4\pi\hbar) \tag{164}$$

$$Y_{J,J-1,M} \quad \text{for} \quad A = J+1 \tag{165}$$

$$Y_{J,J+1,M} \quad \text{for} \quad A = -J \tag{166}$$

each Gaussian term now get a radial form

$$w(r) = r^{-A}\int_0^r dr' r'^A \rho(r')C \tag{167}$$

where $A = J+1$ when the vector spherical harmonics is $Y_{J,J-1,M}$, but $A = -J$ for $Y_{J,J+1,M}$. The argument from (Bang and Hansteen, 2002), that Eq. (167) will lead to finite integrals, independently of the sign of A, is still valid. Let us look at one term in a Gaussian expansion of an ordinary electron density

$$j(r)^L_{kk}/E_{kk} = \int dr'^3 \mathrm{grad}_{r'} \times \rho(r')_{kk} \mathrm{i}/(4\pi\hbar|r_-|) \quad (168)$$

Considering one term in a Gaussian expansion of an electron density (an ordinary, nontransition density) $\rho(\mathbf{r})_{kk}$, we can use the above formula to get

$$\begin{aligned} \mathrm{grad}_r[\rho(r)] &= \mathrm{grad}_r(F(r)Y_{JM}) \\ &= -\sqrt{(J+1)/(2J+1)} \\ &\quad \times (d/dr - J/r)F(r)Y_{J,J+1,M} \\ &\quad + \sqrt{J/(2J+1)} \\ &\quad \times (d/dr + (J+1)/r)F(r)Y_{J,J-1,M} \end{aligned} \quad (169)$$

The two terms correspond to $A = -J$, respectively, $A = J + 1$.

$$F(r) = e^{-\beta r^2} \quad (170)$$

each radial differential equation has the solution

$$r^A \int dr'(r')^{-A}F(r') \quad (171)$$

When the F's are error functions, the integrals are given by

$$\int_{-\infty}^{0} (r')^A e^{-\beta(r')^2} = 1/2(-1)^A(\beta)^{(-1-A)/2}\Gamma[(1+A)/2] \quad (172)$$

$$\int_0^r (r')^A e^{-\beta(r')^2} = 1/2r^{1+A}(r^2\beta)^{1/2(-1-A)} \times \left(\Gamma[(1+A)/2] - \Gamma[(1+A)/2, r^2\beta]\right) \quad (173)$$

The Gaussian densities, with their short ranges are of course not very fit for calculation of the effects of the long-range Coulomb interactions, but since they are easy to calculate, they are widely used.

The vector spherical harmonics are orthogonal in the way that their scalar products, integrated over all the relevant space, are only different from zero if all three of their quantum numbers [J,l,M] are identical. So we get, for the energy term corresponding to the interaction between two electrons, where one goes between the states k and i, the other between the states j and h, the transitions called, respectively, α and β

$$I^L_{J,\alpha,\beta} = \int_0^\infty drr^2((F^L_{\alpha,J+}F^L_{\beta,J+}) + (F^L_{\alpha,J-}F^L_{\beta,J-}) \quad (174)$$

where

$$F^L_{\alpha,J+}(r) = -\sqrt{(J+0)/(2J+1)} \times (d/dr - (J+0)/r)F(r) \quad (175)$$

$$F^L_{\alpha,J-}(r) = +\sqrt{(J+1)/(2J+1)} \times (d/dr + (J+1)/r)F(r) \quad (176)$$

and so on.

4.6.2
The Case of the ZnO Molecule

As an example of the use of the suggested method, we shall look at the molecule of ZnO. Although this is not organic, it plays an important role in many organic compounds. We shall here, as suggested in Bang and Hansteen (2002), use elliptic coordinates, which separate the kinetic energy operator, and also the potential energy of one electron in the Coulomb potential from nuclei at rest, when it can be written in the form

$$V = (V_I(r_1 + r_2) + V_{II}(r_1 - r_2))/r_1 r_2(+V(\phi \cdots)) \tag{177}$$

Here, r_1 and r_2 are the distances of the electron to nuclei 1 and 2, respectively, and ϕ the angle of the position of the electron around an axis through both nuclei. The standard elliptical coordinates are now ϕ and

$$\xi = (r_1 + r_2)/R \tag{178}$$

$$\eta = (r_1 - r_2)/R \tag{179}$$

Here, R is the distance between the two nuclei. To get simple formulas, we chose units so that this distance is 2. The components of the gradient of the electron transition density are now

$$\left[\mathrm{grad}_\phi,\ \mathrm{grad}_\xi,\ \mathrm{grad}_\eta\right]\rho \tag{180}$$

which we shall list below in the same order:

$$\sqrt{(1 - cos^2\phi)/((\xi^2 - 1)(1 - \eta^2))}\,\partial\rho_{ki}/\partial\phi \tag{181}$$

$$\sqrt{(\xi^2 - 1)/(\xi^2 - \eta^2)}\,\partial(\rho_{ki})/\partial\xi \tag{182}$$

$$\sqrt{(1 - \eta^2)/(\xi^2 - \eta^2)}\,\partial(\rho_{ki})/\partial\eta \tag{183}$$

Omitting the spin again, the solution of the one-electron Schrödinger equation in elliptical coordinates has the form

$$\psi_{k,|m|}(\xi, \eta)e^{im\phi} = \sum_{n=0,l=|m|} c^k_{n,l,|m|} e^{im\phi} P_l^{|m|}(\eta) F_{k,n,|m|}(\xi) \tag{184}$$

Here, the solutions in terms of the η, ϕ coordinates are well known and simple, whereas the F functions in terms of ξ are only known in special cases. They are, nevertheless, easy to find from the separated Schrödinger equation in elliptical coordinates. As a rough approximation, we may take the number of relevant states as being the same as the number of Gaussian states in the single particle atomic orbital approximation, that is, with spin 54 basis functions. In the atomic orbital approximation, 22 of these are concentrated around O, the other 32 are concentrated around Zn. Overall, we see that it is possible to get an understanding of the occupation probability by taking the square of the equation given earlier, which mainly mounts to taking the square of $F_{k,n,|m|}$.

4.6.3
The Case of a Water Molecule

Although H_2O is again an inorganic molecule, it also plays a large role in many organic compounds and reactions. It is interesting as an example of a more complicated use of the suggested method

$$\chi(H_2O) = A\chi_1(H)\chi_2(H)\chi(O) \tag{185}$$

Here, A is an antisymmetrizing operator on the electron wave functions. It is not so important, because the electron wave functions for different atoms in a molecule have rather small overlaps. We treat the nuclei in a classical approximation. Now, the hydrogen electron wave functions are Shibuya–Wulfman integral factors (Shibuya and Wulfman, 1965; Avery and Shim, 2001), translated to the center of the oxygen atom

$$\chi_{i,\nu}(\mathbf{x} + \mathbf{R_i}) = \sum_{\nu'} \chi_{\nu'}(\mathbf{x}) S^{\nu}_{\nu'}(\mathbf{R_i}) \tag{186}$$

Hence the product of the two hydrogen functions in Eq. (185) becomes

$$A\chi_{1,\nu}(\mathbf{x})\chi_{2,\nu'}(\mathbf{x})\sum_{\nu''} S_{\nu'}^{\nu''}(\mathbf{R}_1)S_{\nu''}^{\nu}(\mathbf{R}_2) \tag{187}$$

Owing to the group properties of the Shibuya–Wulfman factors, the last sum of products of these can be written as

$$\sum_{\nu''} S_{\nu'}^{\nu''}(\mathbf{R}_1)S_{\nu''}^{\nu}(\mathbf{R}_2) = S_{\nu'}^{\nu}(\mathbf{R}_1+\mathbf{R}_2) \tag{188}$$

Although this expression looks a little complicated, it must be remembered that the Shibuya–Wulfman integral factors are an extremely useful tool in any molecular calculation. The square of the factor $S_{\nu'}^{\nu''}(\mathbf{R}_1)$ is a transition probability between ν and ν' states. These factors will hopefully be part of numerical projects such that programs for their calculation at different values of $\mathbf{R}_{\rm ij}$ should be available in the near future at centers of such calculations, thereby solving one main problem of these calculations: the interaction between electrons and nuclei for different configurations. The other main problem, the interaction between electrons, may then be solved by the methods suggested here or by some modifications of those.

In Bang and Hansteen (2002) one method of describing the Coulomb interaction between electrons in atoms and molecules was given. Here, we shall suggest a similar method for the interactions between electrons and nuclei. We start from Eq. (4) of Bang and Hansteen (2002). Let the quantum numbers of an electron orbital around a nucleus (charge Z) at $\mathbf{R}$ be nlm. In the following, we suppress lm and denote the orbital $\chi_n(\mathbf{r}+\mathbf{R})$ and an (excited) orbit $\chi_{n''}$. Similarly, we denote the states of the nucleus $\chi_{N'}$, $\chi_{N''}$. As in Bang and Hansteen (2002) the general Coulomb matrix element is

$$C\int dr^3 \int dr'^3 div j_{n''n} div j_{N''N'}/|r-r'| \tag{189}$$

but since r' is now the coordinate of a nucleus, we may assume, that its deviation from 0 is negligible, and just write a classical approximation (also neglecting its internal excitations)

$$j_{N''N''} = C_N\delta(r') \tag{190}$$

We then get the amplitude for exciting the electron from an orbital χ_n around $\mathbf{R}$ to another orbital $\chi_{n''}$ around 0 by means of the Coulomb field from a nucleus at 0:

$$C\int dr^3\chi_{n''}^*(\mathbf{r})Z/r\chi_n(\mathbf{r}+\mathbf{R}) \tag{191}$$

Note that the relations between transition charge densities and transition current densities are general and do not depend on whether the normalization is Schrödinger or Sturmian, or on the orthogonality of χ_n, $\chi_{n''}$; not even on whether the argument of the χ is $\mathbf{r}$ or $\mathbf{r}+\mathbf{R}$. Introducing C above as the reciprocal of the double kinetic energy of electrons (common to all Sturmian states) we get our amplitude equal to the $S_{n''}^{n}(\mathbf{R})$ introduced by Shibuya and Wulfman in their important work on molecular electronic wave functions. Using Sturmian states also for the complete molecular wave function, we come to the same secular equation as these authors:

$$\sum_{\tau}[\nu_{\tau'\tau} - k_\mu S_{\tau'\tau}]C_{\tau\mu} = 0 \tag{192}$$

Here, the index τ includes both the quantum numbers nlm and an index describing the position of the various nuclei of the molecule; the ν matrix represents the interaction between one electron and the totality of nuclei of the molecule. The solution of this secular equation thus gives the motion of one electron. To solve the total molecular problem, one should then include the effect of electron–electron interaction, which is the main theme of this paper, and add the classical description of the motion of nuclei, including their repulsion from each other, all three steps might be repeated to convergence.

In the Shibuya and Wulfman (1965) paper, much work is connected to the introduction of a momentum space representation of electron wave functions. Nevertheless, the final expression for their S is written in position space, like here.

4.7
Numerical Treatments

We shall finish the review with a proposal of how one could treat the electronic time development equations derived earlier, for larger molecular systems. This will make use of some of the modern computer software for deriving physical properties of molecules on the basis of their crystallographic databases. Many of the algorithms for these tasks are embedded in the molecular dynamics program packages.

Let us briefly describe here one of such procedures we have in mind, by giving the following example of a large molecular system. An example that illustrates the importance and usefulness of solving our equations of the time development of the electronic systems is zinc molecules or molecules where zinc atoms are incorporated. We have especially the proteins that contain copper ligands of which one medium-sized protein is alcohol dehydrogenase.

The first step is, as for any molecular dynamics simulations that do not start from *ab initio* configurations, to obtain the full set of atomic cartesian coordinates representing the initial configuration. This set of coordinates is obtained from the X-ray crystallographic databank. One might ask if that is necessary for a quantum mechanical calculation, but an initial molecular dynamics energy minimization is always required for positioning the surrounding solvent molecules such as water and other ligands and this classical molecular dynamics is also in our case necessary before solving for quantum time trajectories.

Once this initial molecular dynamics equilibration is performed, we can ask for electrostatic configuration, which is the next input to our electronic time-dependent equations. We wish to have the field strength tensor constructed and that is done either from a multipole expansion or directly from the force field where fields are obtained from the nablaient of the energy potentials built up from the atomic coordinates, the bond lengths and the bond angles. One of the basic effects of the charges in the molecule is to induce dipoles in the surroundings, given basically by the potential from the work the charges are performing: $V_o = 1/2\alpha E^2$.

4.7.1
Comparison of the Standard DFT's and Our Method's Numerics

In the following section, we shall describe how one can assess the various numerical treatments of the molecules

Tab. 4.1 A comparison of the different calculations of occupation probabilities (Mulliken) in atomic orbitals (AOs) of oxygen in various simple molecules. The two methods that are compared are the standard Density Functional Theory (DFT) calculation with Gaussian error functions and our method BB (Bang-Bohr) presented by spherical harmonics. The population is according to the Mulliken-prescription, that is, core contribution minus valence. The py and pz occupations are the same. Both spin states are listed as s_1/s_2. Therefore only px and py are significant.

Calculated occupation numbers						
Molecule type (MD)	**AO 2s (DFT)**	**AO px (DFT)**	**AO py (DFT)**	**AO 2s (BB)**	**AO px (BB)**	**AO py (BB)**
ZnO_2	1.02/0.95	0.65/0.43	1.03/0.74	1.05/0.98	0.62/0.41	1.01/0.72
$4H_2O$	0.90/0.91	1.00/0.71	0.97/0.69	0.94/0.95	0.97/0.69	0.94/0.67

mentioned earlier. We shall, within these examples, compare the numbers about electronic orbitals calculated with the help of usual theoretical DFT techniques and techniques based on our electronic wave equations.

We give the results from a DFT calculation of the small molecules using the Gaussian program from where we have listed all the occupation probabilities of the available quantum states. Thus, we can employ the formulas derived earlier to derive the occupation numbers from our method.

The first case we have studied numerically is the zinc-oxide molecule. In Table 4.1 we have listed the occupation probabilities of the electrons in the various ground states.

As is seen from the table there is reasonable agreement of the stationary population statistics in the lower atomic orbitals of the two methods. Our method is supposed to be superior with respect to excited state dynamics, the transition amplitudes, and other time-dependent quantities. However, it is hard to find other numbers to compare with than the stationary population dynamics.

Appendix A: Gaussian Expansions and Numerical Calculations

In the following, we want to compare results with standard DFT calculations using Gaussian functions. Hence, we shall, for the purpose of comparisons, introduce Gaussian expansions of the wave functions and similarly of the density ρ and its derivatives that are also taken as Gaussian like. It is thus natural to start from the current expression:

$$j^L_{ki} = \nabla_r \int dr'^3 \rho(r')^A_{ki} \mathrm{i} E_{ki}/((r-)4\pi\hbar) \times Y_{J,J-1,M} \tag{A1}$$

and each Gaussian term now gets a radial form:

$$w(r) = r^{-A} \int_0^r dr' r'^A \rho(r') C' \tag{A2}$$

The argument from Bang and Hansteen (2002) that concerns the finiteness of the integrals given earlier, independent of the sign of A, is still valid. Actually, extending the upper integration limit to infinity renders these Gaussian integrals manageable, such that

$$\int_0^\infty e^{-\lambda x^2} x^k = 1/2\lambda^{-(k+1)/m}/\,\Gamma(k+1/m) \tag{A3}$$

showing that no singularities are occurring in this formalism. The integrals, which we shall use, are of the form

$$Q(r) = -\sum_i \int dr' \int dr'' \psi_i(r)^* \times G_i(r, r') V_e(r', r'') \psi_i(r'') \tag{A4}$$

which can be simplified, using the procedures of Eqs (A2) and (A3), to

$$r^{-A} \int_0^r dr' (r')^A C \int_0^{r'} dr'' e^{[-r''^2]} \tag{A5}$$

where $A = J + 1$ corresponds to the vector spherical harmonics to be $Y_{J,J-1,M}$ but, in the case of $A = -J$, becomes $Y_{J,J+1,M}$. Considering just one term (diagonal) in a Gaussian expansion of an electron density (an ordinary, nontransition density) $\rho(r)_{kk}$, we can use the formula given earlier to get

$$\frac{j(r)^L_{kk}}{E_{kk}} = \int dr'^3 \nabla'_r \rho(r')_{kk} \mathrm{i}/84\pi\hbar \tag{A6}$$

where

$$\begin{aligned}\nabla_r[\rho(r)] &= \nabla_r(F(r)Y_{JM}) \\ &= -\sqrt{(J+1)/(2J+1)} \\ &\quad \times (d/dr - J/r)F(r)Y_{J,J+1,M} \\ &\quad + \sqrt{J/(2J+1)} \\ &\quad \times (d/dr - J/r)F(r)Y_{J,J-1,M}\end{aligned} \tag{A7}$$

The two terms correspond to $A = -J$ and $A = J + 1$, respectively. Terms with negative A can be eliminated in the radial integral, where the F functions now are of the Gaussian type. This can be seen by using a recursive procedure, repeated some times

$$\mathrm{i}^n \operatorname{erfc}(z) = -z/n\mathrm{i}^{n-1} \operatorname{erfc}(z) + 1/(2n)i^{n-2} \operatorname{erfc}(z) \tag{A8}$$

or

$$\begin{aligned}2(n+1)(n+2)\mathrm{i}^{n+2} \operatorname{erfc}(z) &= (2n+1+2z^2) \operatorname{erfc}(z) \\ &\quad - 1/2\mathrm{i}^{n-3} \operatorname{erfc}(z)\end{aligned} \tag{A9}$$

where erfc is the error function.

Appendix B: Transition Charge Density

For the sake of consistency of the chapter, this appendix brings a derivation of the expression, following the prescription of the Bang and Hansteen (2002) for the transition current densities that are used in Sections 4.2.1 and 42.2 of the present chapter. We start, as in Bang and Hansteen (2002), from an expression for the one-particle transition charge density:

$$\rho(r)_{ki} = e\psi(r)^*_k \psi(r)_i \tag{B1}$$

where $\psi(r)_k$, $\psi(r)_i$ are the one-particle energy eigen states and where the numbers i, k stand for the radial, angular momentum, and so on. The perturbative matrix element for the residual Coulomb interaction between two electrons situated at r and r' is then expressed by the integral

$$I_C = \int dr^3 \int dr'^3 \frac{\rho(r)_{ki}\rho(r')_{jl}}{|\,r - r'\,|} \tag{B2}$$

and applying the time-dependent Schrödinger equation for ψ the integral can be written as

$$I_C = -\int dr^3 \int dr'^3 \frac{1}{| r - r' |} \frac{\partial \rho(r)_{ki}}{\partial t} \times \frac{\partial \rho(r')_{jl}}{\partial t} (E_{ki} E_{jl})^{-1} (\hbar)^2 \qquad \text{(B3)}$$

where $E_{ij} = E_i - E_j$. If again the continuity equation of transition charge densities is applied, one can write the integral I_C as follows:

$$I_C = -\int dr^3 \int dr'^3 \frac{1}{| r - r' |} \times div_r \mathbf{j}(\mathbf{r})_{ki} div_{r'} \mathbf{j}(\mathbf{r}')_{jl} (E_{ki} E_{jl})^{-1} (\hbar)^2 \qquad \text{(B4)}$$

with j_{ij} being the transition current:

$$\mathbf{j}_{ij} = a(\psi_i^* \nabla \psi_j - \psi_j \nabla \psi_i^*) \equiv \mathbf{j}^\alpha \mathbf{j}^\beta \qquad \text{(B5)}$$

where $a = \frac{e\hbar}{2mi}$ and where α, β correspond to the quantum numbers of the interchange of the states i, j. One should remark that, fortunately, if the denominator E_{ij} is zero, then the numerator is also zero.

We are now in a position to derive the very important expressions for the current, which are used in Chapter 3. These expressions have been derived with the powerful tool of vector algebra where the vectors, in this case the transition current densities, are split into longitudinal parts, j^L, and transverse parts, j^T, as seen later:

$$\mathbf{j}(\mathbf{r})^L = -\mathrm{grad}_r \int d\mathbf{r}'^3 div_{r'} \times \mathbf{j}(\mathbf{r}') \frac{1}{(4\pi \mid \mathbf{r} - \mathbf{r}' \mid)} \qquad \text{(B6)}$$

which in terms of densities is:

$$\mathbf{j}(\mathbf{r})_{ij}^L = \int dr'^3 \mathrm{grad}_{r'} \boldsymbol{\rho}(\mathbf{r}')_{ij} \frac{iE_{ij}}{(4\pi \mid \mathbf{r} - \mathbf{r}' \mid \hbar)} \qquad \text{(B7)}$$

and

$$\mathbf{j}(\mathbf{r})^T = \mathrm{rot}_{r'} \int d\mathbf{r}'^3 rot_{r'} \mathbf{j}(\mathbf{r}') \frac{1}{(4\pi \mid \mathbf{r} - \mathbf{r}' \mid)} \qquad \text{(B8)}$$

which altogether can be written as $\mathbf{j} = \mathbf{j}^L + \mathbf{j}^T$.

With these expressions inserted in Eq. (A2) and subsequently performing a partial integrations, one finally obtains

$$I_C = \frac{\hbar^2 4\pi}{E_{ij} E_{kl}} \int d\mathbf{r}^3 \, \mathbf{j}(\mathbf{r})_{ki}^L \, \mathbf{j}(\mathbf{r})_{jl}^L \equiv \frac{\hbar^2 4\pi I^L}{E_{ij} E_{kl}} \qquad \text{(B9)}$$

which therefore transform the sixth dimensional integral at the start of the appendix to a much simpler three-dimensional integral of the scalar product of the two transition current densities. There are similar expressions valid for the transverse component. However, only the longitudinal ones are used for the calculation of the Coulomb matrix elements.

References

Atkins, P.W. and Friedman, R.S. (1997) *Molecular Quantum Mechanics*, 3rd edn, Oxford University Press.

Avery, J.M. and Shim, R. (2001) *Int. J. Quantum Chem.*, **83**, 1.

Bang, J. and Bohr, H.G. (2008) Time development for electrons in a current functional theory for atoms and molecules. *Phys. Rev. A. Submitted to J. Phys. A.* (2008).

Bang, J.M. and Hansteen, J.M. (2002) Coulomb Matrix Elements: outline of a description of electron interactions in purely Coulombic systems. *J. Phys. B: At., Mol. Phys.*, **35**, 3979.

Bohr, H.G., Jalkanen, K.J., *et al.* (1999)*Chem. Phys.*, **246**, 13.

Bohr, H.G., Jalkanen, K.J., *et al.* (1999) *Chem. Phys.*, **246**, 13.

Brink, D. and da Providencia, J. (1989) *Nucl. Phys. A*, **500**, 301.

Davidov, A.S. (1965) *Quantum Mechanics*, Pergamon Press, Oxford.

Gross, E.K.U., *et al.* (1996)Density-functional theory of time-dependant phenomena. *Top. Curr. Chem.*, **181**, 82.

Kamke, E. (1983) *Differentialgleichungen*, Teupner, Stuttgartt.

Nemes, M.C., Piza A.F.R.T., and da Providencia, J. (1987) *Physica A*, **146**, 289.

Nielsen, M. and da Providencia, J. (1989) *Phys. Rev. C*, **40**, 2377.

Shibuya, T. and Wulfman, C.E. (1965) *Proc. R. Soc. A*, **286**, 376.

5

Exploring Excited State Processes in Photosynthesis – Features of Photosynthetic Light Harvesting

Villy Sundstrom

5.1 Construction of Light Harvesting and Energy Converting Pigment Systems of Photosynthesis 169
5.2 Need for a Light-Harvesting Antenna 170
5.3 Spectral Coverage 170
5.4 Efficient Energy Flow Through the Light-Harvesting Antenna Systems 173
5.5 Intracomplex Energy Transfer 175
5.6 Delocalized Excitons in Photosynthetic Light Harvesting 177
5.7 Efficient Antenna–RC Coupling: Long Distance versus Short Distance 178
5.8 Carotenoid Molecules in Photosynthesis – Their Spectroscopy 179
5.9 Light Harvesting by Carotenoid Molecules 184
5.10 Protection of the Photosynthetic Machinery (NPQ) – Quenching of Chlorophyll Excited States by Carotenoids 185
5.11 Storing the Energy of Light – Photosynthetic Charge Separation 186
Acknowledgments 189
References 189

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

5.1
Construction of Light Harvesting and Energy Converting Pigment Systems of Photosynthesis

There is a large variety of photosynthetic organisms ranging from very old photosynthetic bacteria containing only one photosystem to more recently evolved cyanobacteria, algae, and green plants having two photosystems operating in series and capable of water oxidation and oxygen production. Despite large differences in composition and architecture of the pigment–protein systems of various organisms, the overall function of most photosynthetic light energy converting systems is surprisingly similar. Thus, it is generally seen that light is very efficiently collected by a light-harvesting antenna and within ~100 ps it is transferred to the photochemical reaction center (RC) where the light energy is converted to a stable charge-separated state with almost 100% quantum efficiency creating a transmembrane potential, which is the source of energy for all subsequent processes in the organisms (vanGrondelle, 1985; vanGrondelle *et al.*, 1994; Pullerits and Sundstrom, 1996; Sundstrom, Pullerits and vanGrondelle, 1999). This general pattern of overall time scale of the primary photosynthetic processes can easily be understood as a combined result of the high (~95%) overall quantum efficiency of light-to-charge conversion and the known nanosecond fluorescence lifetimes of many photosynthetic pigments (chlorophyll, bacteriochlorophyll, phycocyanin, etc.). To efficiently compete with fluorescence deactivation of the excited states, the functional energy and charge transfer processes have to be at least 10–100 times faster. These requirements for efficient conversion of light energy into chemical energy in pigment systems, which are robust against changes in environmental factors, variation in light intensity, presence of reactive species, and variation in pigmentation and structure, have through the eons of evolution led to a number of "design principles" that can be recognized in most photosynthetic systems. In what follows, we will identify the most important concepts and discuss how they contribute to achieve an efficient light harvesting and converting process. In the search for efficient man-made devices for solar energy conversion, the principles of nature are very interesting to pursue.

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

5.2
Need for a Light-Harvesting Antenna

Natural photosynthesis utilizes two coupled pigment systems, a light-harvesting antenna and a photochemical RC, for efficient conversion of light energy into stabilized charges, eventually used to synthesize chemically stable high-energy products. This is well illustrated by the photosynthetic unit (PSU) of purple bacteria (Figure 5.1a), where two different antenna systems, the LH2 peripheral antenna (smaller circles) and the core LH1 antenna (bigger circle), are surrounding the photosynthetic RC inside of LH1. Both pigment systems are absolutely essential for the operation of photosynthesis. The light-harvesting antenna is composed of a large number (~100) of pigment molecules, whose task is to collect sunlight over a broad spectral range and efficiently deliver it to the RC. The RC consists of a few highly specialized pigment molecules, whose organization and interactions are fine tuned to optimize unidirectional electron transfer across the photosynthetic membrane, achieving a long lived charge-separated state with high efficiency. In oxygenic photosynthesis, the oxidizing power of this charge-separated state is used to extract electrons from water, with the help of a catalytic manganese cluster.

The need for a coupled antenna–RC system is easily realized by considering the fact that the photochemical reaction in photosynthesis, splitting of water and formation of molecular oxygen, requires four photochemical cycles of the RC to occur in quite short time (typically milliseconds). On longer time scales, all charge-separated states undergo recombination reactions to the ground state, causing loss of the acquired photon energy. Owing to this constraint, without an antenna system coupled to the RC the conversion efficiency of a photosynthetic system would be very low at the ambient light fluxes delivered by the sun. Thus, an antenna coupled to a photochemical RC unit is a necessary feature for efficient photocatalytic operation at ambient solar light intensities. This is also an important lesson to learn from nature for constructing of an artificial solar energy conversion system. Figure 5.1a–d shows a few examples of photosynthetic antenna–RC pigment systems from different types of organisms. Their function is discussed below.

5.3
Spectral Coverage

Photosynthetic organisms live in habitats with greatly varying conditions of light intensity and spectral distribution. Land-based plants are exposed to the full solar spectrum, whereas organisms living in water or in the soil all are exposed to a more or less heavily filtered solar spectrum. This implies that species have developed their pigment composition to optimize light collection in each specific location. This fact is illustrated in Figure 5.2, where the solar spectrum is reproduced along with characteristic regions of light absorption of different organisms and pigments. Thus, we can see that the useful parts of the solar spectrum from the blue (~400 nm) to the near-infrared (~1100 nm) is completely covered by various pigments without any gaps; nature has devised pigment systems and organisms have evolved to efficiently harvest all colors of the sunlight impinging on the Earth. This large

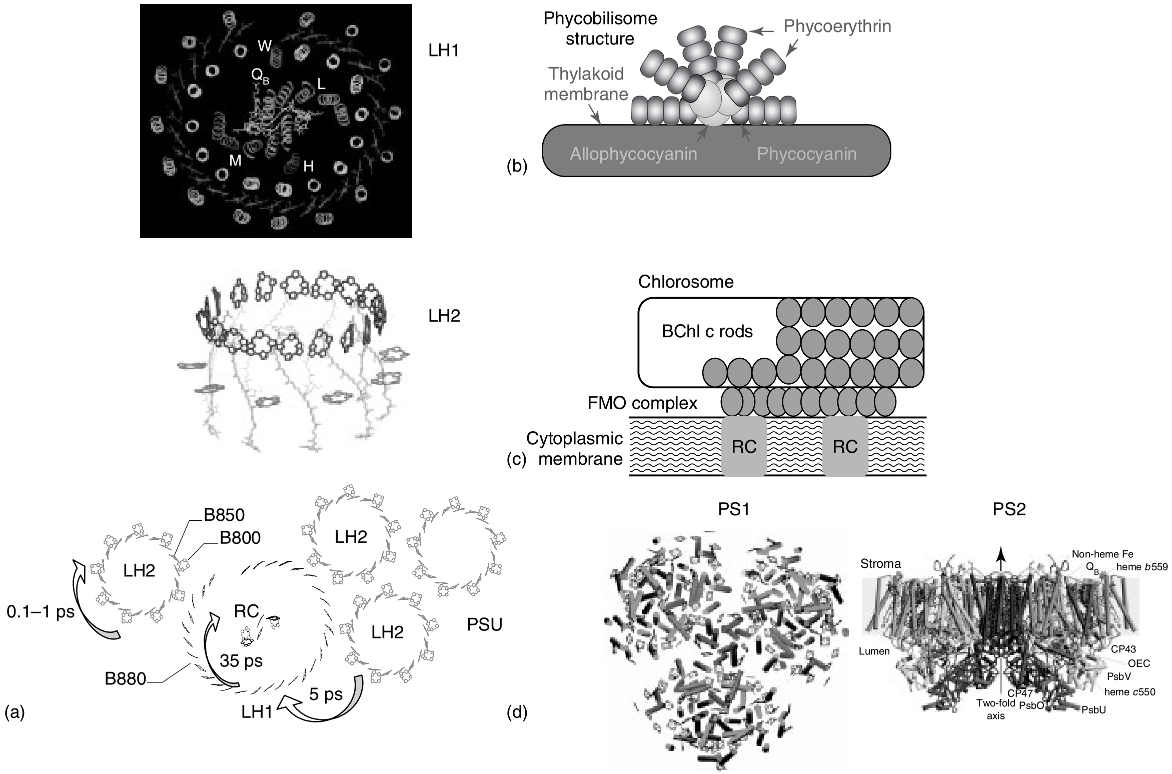
W
Q_B
L
M
H
LH1
LH2
B850
B800
LH2
0.1–1 ps
RC
35 ps
B880
LH2
LH2
PSU
5 ps
LH1
(a)
Phycobilisome structure
Phycoerythrin
Thylakoid membrane
Allophycocyanin
Phycocyanin
(b)
Chlorosome
BChl c rods
FMO complex
Cytoplasmic membrane
RC
RC
(c)
PS1
PS2
Stroma
Non-heme Fe
Q_B
heme *b*559
Lumen
CP43
OEC
PsbV
heme *c*550
CP47
Two-fold axis
PsbO
PsbU
(d)

Fig. 5.1 A few examples of photosynthetic pigment systems. (a) LH1, LH2, and full photosynthetic unit (PSU) of purple bacteria. (b) Phycobilisome antenna attached to the photosynthetic membrane of cyanobacteria. (c) Chlorosome and FMO antenna complexes attached to the photosynthetic membrane of green sulfur bacteria. (d) The PS1 (left) and PS2 (right) photosystems of green plants.

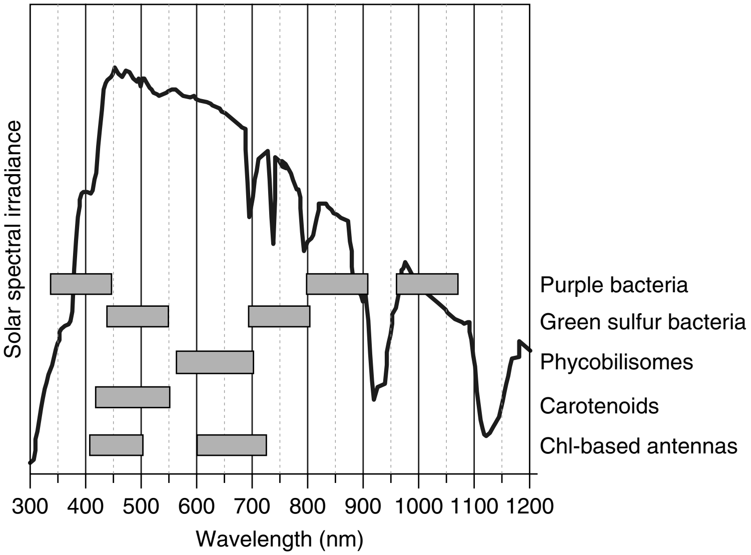

Fig. 5.2 Spectrum of the solar radiation that reaches the surface of the earth. Wavelength ranges of absorption for different pigments and a few types of photosynthetic organisms.

variety of pigmentation is integrated into a range of functional pigment–protein structures. In Figure 5.1 we show just a few examples to illustrate the great structural diversity of the photosynthetic pigment apparatus construction. In photosynthetic purple bacteria, both antenna and RCs are transmembrane proteins fully integrated into the photosynthetic membrane (Figure 5.1a). The antenna consists of a circular core antenna LH1 surrounding the RC and a peripheral LH2 complex also of circular shape. High-resolution crystal structures have been obtained for both LH1 (Roszak *et al.*, 2003) and LH2 (Mcdermott *et al.*, 1995). The pigments of these antennas have well-separated absorption and fluorescence spectra and the high symmetry of the complexes makes them amenable for quantum chemistry calculations and computer simulations of energy and electron transfer processes. It also turns out that the antennas of purple bacteria constitute a representation of various energy transfer mechanisms. Antennas and RCs of purple bacteria have therefore served as test grounds for studies of energy and electron transfer in photosynthesis; much of our knowledge today about these processes stems from studies of purple bacterial pigment–protein complexes. In cyanobacteria (Figure 5.1b) there is an extramembrane antenna, the phycobilisome (Gantt, 1981), attached to the thylakoid membrane holding the core antenna and the photochemical RC. The chromophores of phycobilisomes are open-chain tetrapyrolles (phycoerythrin, phycocyanin, and allophycocyanin) arranged in such a way that downhill energy transfer (from high- to low-energy pigments) occurs toward the membrane and chlorophyll containing core antenna and RC (vanGrondelle *et al.*, 1994). The overall construction of the photosynthetic pigment apparatus of green sulfur bacteria (Figure 5.1c) is reminiscent of that of cyanobacteria. Here, a large antenna containing thousands of bacteriochlorophyll molecules arranged in rod-like structures and encapsulated in cigar-shaped structures (chlorosomes) (Olson, 1980) are attached via a baseplate and a linking pigment–protein complex (the so-called Fenna-Mathews-Olsen (FMO) Fenna and Matthews, 1975, complex) to the membrane. One chlorosome can funnel energy to 10–15 RCs. To conclude this survey of different PSUs, we consider that of green plants consisting of the two photosystems PS1 (Jordan *et al.*, 2001) and PS2 (Zouni *et al.*, 2001; Ferreira *et al.*, 2004) (Figure 5.1d), operating in series. Here, chlorophyll *a* and chlorophyll *b* are the main chromophores of both antenna and RCs, and despite the presence of many different pigment–protein complexes, the spectral properties are quite similar. This difficulty to spectroscopically distinguish different parts of the photosynthetic apparatus has delayed a detailed understanding of energy and electron transfer pathways and mechanisms. However, using the most powerful spectroscopic methods combined with theoretical modeling, important progress has been made.

5.4
Efficient Energy Flow Through the Light-Harvesting Antenna Systems

Examination of different photosynthetic antennas reveals one of nature's remarkable achievements – despite often very large antennas of hundreds or sometimes even thousands of pigment molecules, all of them are coupled to each other such that energy flows through the system and is channeled into the RC with a total quantum conversion efficiency approaching unity. How remarkable this design is, may be best appreciated when one tries to build an artificial pigment system for light harvesting – it is very difficult to make a large system containing many pigment molecules that are all well-coupled and at the same time avoid excited state quenching. The distance dependence to the sixth power of Förster energy transfer (Forster, 1946), generally active in photosynthetic systems, implies that molecules have to be within ~1 to 3 nm for efficient energy transfer. If the distance is shorter, excited state quenching often occurs and efficiency decreases. Nature has solved this problem by using proteins that hold the pigment molecules in well-defined and fixed positions, which optimize energy transport and minimize quenching processes. In the PS1 and PS2 antennas (Figure 5.1d) more than half the mass is pigment and (Figure 5.1c) there is no protein in the interior of the chlorosome; the bacteriochlorophyll-*c* molecules are self-assembled into a very ordered aggregate. Nevertheless, the excited state lifetimes of the pigments in antennas are generally ~1 ns and there is plenty of time for energy transfer over quite long distances. As we will see below, typical energy transfer times between two pigment molecules are subpicosecond to few picoseconds (vanGrondelle *et al.*, 1994; Pullerits and Sundstrom, 1996; Sundstrom, Pullerits and vanGrondelle, 1999; Sundstrom *et al.*, 1986).

We will use the photosynthetic apparatus of purple bacteria (Figure 5.1a) to illustrate the concepts just discussed, since it contains spectrally and structurally distinct pigment systems, implying that using the tools of time-resolved spectroscopy we can, in detail, monitor the flow of energy through the antenna and relate observed dynamics to structural features.

Figure 5.3a illustrates the absorption spectrum of the purple bacterial PSU (Figure 5.1a), demonstrating the two infrared absorption bands of LH2 characteristic of the B800 and B850 pigments, as well as a red wing due to the LH1 core antenna. The RC absorption is overwhelmed by the absorption of many more LH1 antenna molecules. Energy flow through the pigment system can be monitored with time-resolved fluorescence or transient absorption spectroscopy. The well-separated spectral bands of the different pigments facilitate the resolution of the various transfer steps. This is illustrated by the time-resolved transient absorption spectrum of a purple bacterial PSU in Figure 5.3b. The peripheral LH2 antenna

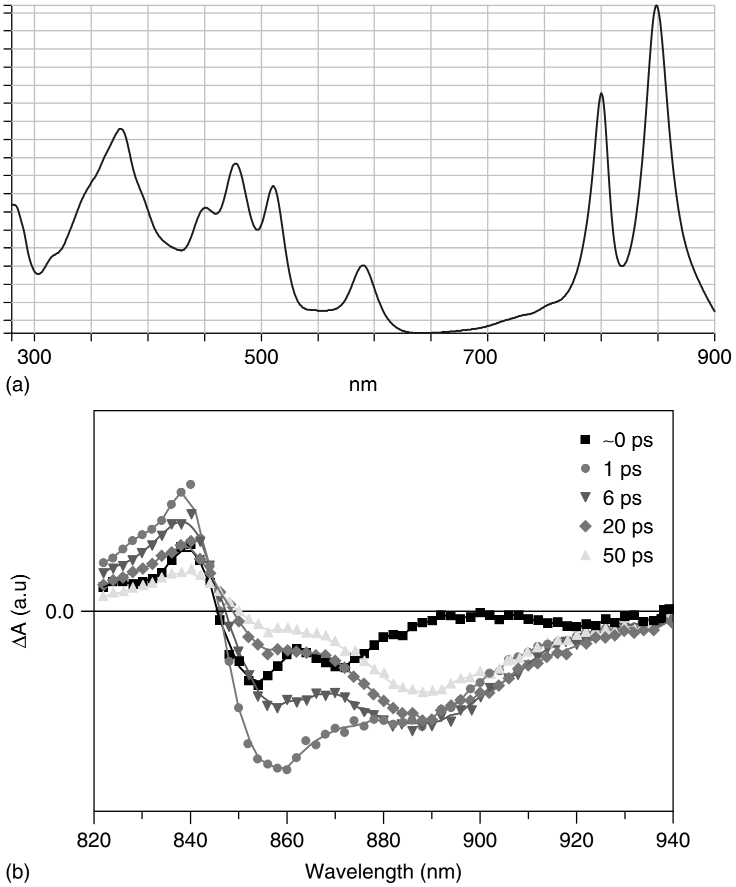

Fig. 5.3 (a) Absorption spectrum of the purple bacterial PSU. (b) Time evolution of the transient absorption spectrum of the purple bacterial photosynthetic unit, following excitation of the LH2 antenna with an ultrashort optical pulse.

was excited with an ultrashort femtosecond laser pulse. At the earliest times, after photoexcitation the B800 and B850 bands are bleached due to the direct optical excitation. On the tens of picosecond time scale the low-energy bands of LH1 and the RC grow in intensity at the expense of the higher energy LH2 bands, showing that the energy has reached the RC within this time (vanGrondelle *et al.*, 1994; Sundstrom, Pullerits and vanGrondelle, 1999; Sundstrom *et al.*, 1986).

Energy transfer from an LH2 to an LH1 ring was studied in a membrane preparation of *Rhodobacter sphaeroides* and concluded to be characterized by a ∼5 ps transfer time, consistent with the estimated inter-ring distance. Transfer between aggregated B850 rings was studied using time-resolved excitation annihilation combined with a Monte Carlo simulation, together with transmission electron microscopy (TEM) (Schubert *et al.*, 2004). The random-walk Monte Carlo calculation showed that the ring-to-ring hopping time is ∼30 ps and the number of B850 rings in an aggregate is ∼15, in good agreement with the spot size observed in the TEM images (Schubert *et al.*, 2004). The ring-to-ring hopping time of ∼30 ps implies that for a B850 ring surrounded by six neighboring rings the excitation residence time is ∼5 ps. This time is in good agreement with the 3–8 ps LH2–LH1 transfer time obtained for native membranes (Hess *et al.*, 1995; Nagarajan and Parson, 1997; Freiberg *et al.*, 1989).

Upon inspecting the structures of, for instance, the purple bacterial (Figure 5.1a) and PS1 (Figure 5.1d) PSUs the concept of multiple energy transfer pathways becomes obvious, that is from a given point in the antenna system there are many possible paths to reach the RC. This is a design principle that renders high efficiency and robustness to the antenna system. In the purple bacterial PSU, LH2 rings are connected to other LH2s as well as to LH1, implying that if an excitation in an LH2 ring for some reason is disconnected from the LH1 core antenna, it can always find an alternative path through other LH2s. Once in LH1, the excitation can enter the RC from any point on the LH1 circle. The multiple pathways of energy transfer in combination with the very fast (hundreds of femtosecond time scale) energy transfer within B850 of LH2 and LH1 result in a very fast and efficient percolation of excitation energy over the photosynthetic pigment system. The energy gradient of the light-harvesting pigments toward the core antenna and RC results in the observed very fast and efficient (∼50 ps and close to 100% quantum efficiency) overall transfer of excitations from antenna to RC.

5.5
Intracomplex Energy Transfer

Using biochemical methods, individual pigment–protein complexes can be isolated from the photosynthetic membrane. Spectroscopic studies of such preparations offer additional detail about the photosynthetic processes. We give one example, B800 to B850 energy transfer in the LH2 complex (Figure 5.1a). The BChls of B800 are well separated from each other and from the B850s and thereby have mainly monomeric spectroscopic properties. Excitation absorbed in the B800 ring is first transferred among B800 BChls (Sundstrom, Pullerits and vanGrondelle, 1999). Calculations based on the Förster theory (Forster, 1946) agree remarkably well with the measured

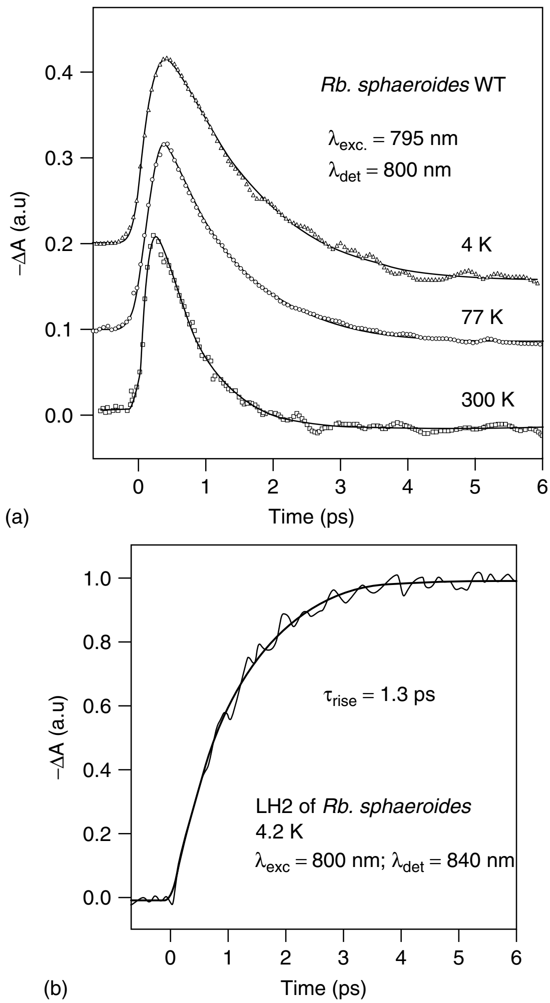

Fig. 5.4 Transient absorption kinetics illustrating energy transfer from B800 to B850 in LH2. B800 is excited by a femtosecond pulse at 800 nm and the energy transfer is monitored in (a) as the decay of B800 excited state population, and in (b) as the corresponding rise of B850 population. The transfer becomes somewhat slower at low temperature.

pairwise transfer time of 300 fs, suggesting that for the B800 ring the point-dipole approximation is applicable. Also low temperature transient absorption anisotropy kinetics measured at different wavelengths inside the B800 band were successfully simulated by a model of Förster hopping in a spectrally inhomogeneous ring of BChl molecules. B800 to B850 transfer occurs with a time-constant of 0.7 ps at room temperature (Sundstrom, Pullerits and vanGrondelle, 1999) and it slows down upon lowering the temperature to 1.2 ps at 77 K and to 1.5 ps at 4 K (Figure 5.4). Both B800 and B850 can be spectrally tuned by selective mutations of key amino acids and the B800 BChls can be exchanged to other similar BChl and Chl pigments (Sundstrom, Pullerits and van-Grondelle, 1999; Fraser *et al.*, 1999). A series of pigment–protein complexes, where either the B800 or B850 band was spectrally shifted, were studied and significant changes of the transfer time were observed (Sundstrom, Pullerits and vanGrondelle, 1999; Herek *et al.*, 2000). The complex with the most blue-shifted B800 pigment, at 670 nm, gave the transfer time of 8.3 ps, as compared to 0.7 ps for the wild-type complex. All the observed trends of the transfer time are in qualitative agreement with what one would expect based on Förster spectral overlap. However, quantitative Förster theory calculations of the transfer time based on the dipole–dipole interaction between B800 and nearby B850s failed to reproduce the observed lifetimes by almost an order of magnitude. Since the spectral overlap did describe the qualitative trend, it was suggested that the source of the discrepancy is the electronic coupling term (Herek *et al.*, 2000). It has been pointed out that the carotenoid molecule may contribute to the electronic coupling between B800 and B850 ((Sundstrom, Pullerits and vanGrondelle, 1999) and references therein). Alternatively, the nondiagonal and/or diagonal (Sumi, 1999) electron–phonon coupling may facilitate efficient excitation transfer to optically forbidden exciton levels. In the spirit of the same ideas, a modified Förster theory adapted for transfer to the collective exciton states with spectral inhomogeneity has been used for describing B800 to B850 transfer, apparently leading to quantitative agreement between theory and experiment (Scholes and Fleming, 2000).

5.6
Delocalized Excitons in Photosynthetic Light Harvesting

The B850 ring of LH2 (Figure 5.1a) forms a densely packed excitonically coupled aggregate. Shortly after the structure of the LH2 became available, the question of the extent of exciton delocalization in B850 became a hot research subject. Different experimental techniques and theoretical methods led to diverging and sometimes even contradicting conclusions. On the basis of nonlinear absorption experiments (Leupold *et al.*, 1996) it was concluded that the excitation is delocalized over almost the whole B850 ring whereas analyses of the transient absorption spectra suggested quite limited exciton delocalization on the picosecond timescale – about 4 BChl molecules (Pullerits, Chachisvilis and Sundstrom, 1996). Many of the following studies supported one or the other point of view (Meier, Chernyak and Mukamel, 1997; Monshouwer *et al.*, 1997; Chachisvilis *et al.*, 1997). The apparent controversy was resolved by considering that the exciton, and thereby also exciton delocalization, is time-dependent (Dahlbom *et al.*, 2001). Furthermore, different definitions of exciton delocalization length led to rather different numerical values for the quantity. We showed that at the moment of excitation certain measures of exciton delocalization indeed gave numerical values suggesting full-ring excitons. After a few hundred femtoseconds most of the measures gave the value ∼4 BChls (Dahlbom *et al.*, 2001). Figure 5.5 illustrates calculated time dependencies of exciton delocalization in B850 for several different definitions of the delocalization. The calculated results are correlated to various types of measurements of exciton delocalization.

At high light intensities it may happen that two or more excitations are simultaneously present in a set of molecules exchanging excitation, like B850. In this

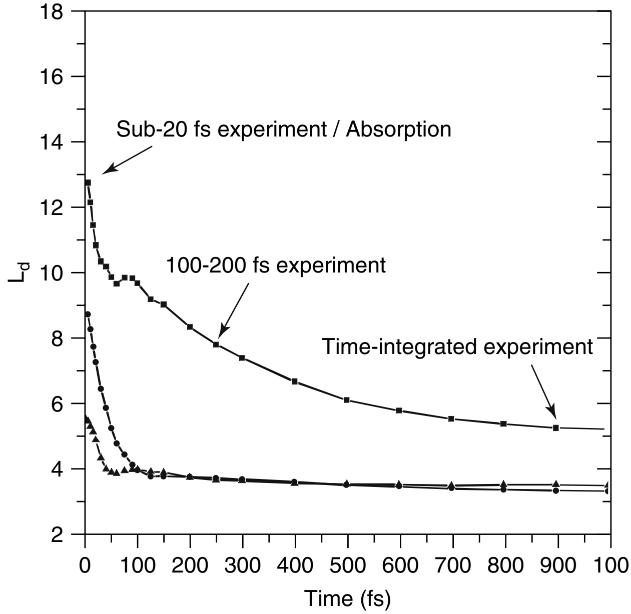

Fig. 5.5 Calculated time dependencies of exciton delocalization, L_d (number of BChl molecules over which exciton is delocalized), in B850 for several different definitions of the delocalization. The arrows indicate what would be the observed delocalization for different types of experiments.

case excitation annihilation can take place (Mauzerall, 1976; Denhollander, Bakker and vanGrondelle, 1983). The annihilation process is usually described as excitation transfer to an already excited molecule, where it produces a higher (doubly) excited electronic state. By very fast internal conversion the molecule relaxes to the lowest excited molecular state and one excitation is lost. Simultaneous analysis of excitation annihilation dynamics and transient absorption anisotropy decay in well-separated rings of LH2 was recently used for obtaining detailed information about excited state properties and dynamics in a B850 ring (Trinkunas *et al.*, 2001). It was confirmed that excitons in B850 are delocalized over 3–4 BChls.

5.7
Efficient Antenna–RC Coupling: Long Distance versus Short Distance

The last step in excitation transfer is transfer from the antenna to the special pair of the RC (energy trapping). Already early studies indicated that the trapping by the RC is slow (Visscher *et al.*, 1989). The rate of antenna to RC transfer was first determined by Timpmann *et al.* (1993) to be 30–40 ps, that is demonstrating that the final antenna–RC energy transfer step is the slowest and rate limiting step in the overall energy converting process. The trapping of excitation energy by the photosynthetic RC has often been discussed in terms of a trap-limited or diffusion-limited process, implying that the rate limiting process is charge separation in the RC, or the energy diffusion process through the antenna, respectively. The realization that the antenna to RC energy transfer step is rate limiting, was termed *transfer-to-trap-limited* (vanGrondelle *et al.*, 1994; Timpmann *et al.*, 1993). The structure of the purple bacterial LH1–RC complex (Figure 5.1a) explains the result, the distance between the RC and the LH1 BChls is the longest distance, which by virtue of the strong distance dependence (sixth power) of the energy transfer rate will be the slowest step. The functional importance of this large LH1–RC distance is that it prevents photo-oxidation of antenna BChls, and concomitant excited state quenching in the antenna, by the oxidized special pair of the RC.

An issue sometimes raised in the literature is the role of collective excited states, excitons, in photosynthetic light harvesting. Formation of collective excited states through interchromophore interaction leads to increased dipole strength of the collective state as compared to the monomeric chromophore. Energy transfer involving collective excited states will therefore be faster than energy transfer involving monomeric pigments under otherwise identical conditions. Since the last energy transfer step to the RC is the slowest, this step is most influenced by collective excitations. Thus, the overall trapping and thereby efficiency of photosynthesis is significantly improved by exciton formation in LH1.

Characteristic times of the different energy transfer steps discussed above are highlighted in the figure of the purple bacterial PSU in Figure 5.1a. Energy transfer between neighboring BChl molecules within the LH1 or LH2 pigment–protein complexes occurs on the ~0.1 to 1 ps time scale, whereas transfer between these complexes takes ~5 ps. The final energy transfer step from the LH1 ring to the special pair in the RC is characterized by the slowest energy transfer, ~35 ps. This results in efficient transfer of energy from any point in the network of antenna pigments to the RC, such that energy is stored in the RC within 50–100 ps after light absorption with an overall quantum efficiency approaching unity. Other pigment systems of other photosynthetic organisms, despite being composed of different pigments and having very different organization, show similar time scales of the energy transfer processes (vanGrondelle *et al.*, 1994; van Amerongen and van Grondelle, 2001; van Grondelle and Novoderezhkin, 2006; Croce *et al.*, 2000, 2003; Karapetyan *et al.*, 1999).

5.8
Carotenoid Molecules in Photosynthesis – Their Spectroscopy

Carotenoids are, along with chlorophylls, the most abundant pigments found in nature. They are present in most organisms including humans, but can be synthesized only by plants and microorganisms. While they are perhaps best known for their bright colors, they have well-documented multiple functions in nature: they serve as light-harvesting pigments in almost all photosynthetic organisms covering a region of the visible spectrum not accessible by (bacterio)chlorophylls and they protect against excessive light by quenching both singlet and triplet states of (B)Chls (van Amerongen and van Grondelle, 2001; Frank and Cogdell, 1993, 1996; Koyama *et al.*, 1996; Ritz *et al.*, 2000).

The diversity of carotenoid functions is unmatched by any other class of natural pigments. The functional variety is directly related to their unique spectroscopic properties resulting from the structure of the carotenoid molecule (Figure 5.6). The central pattern repeated in all carotenoids is a backbone consisting of alternating single and double carbon bonds that form a conjugated π-electron system responsible for most of the spectroscopic properties of carotenoids. An energy level scheme of a carotenoid molecule together with transitions, which will be used to describe the dynamics of excited states in the following sections, is depicted in Figure 5.7. The strong absorption of carotenoids (extinction coefficients of the order of 10^5) in the blue-green spectral region is caused by the strongly allowed S_0-S_2 transition. The S_0-S_2 transition of carotenoids usually exhibits a characteristic three-peak structure corresponding to the lowest three vibrational levels of the S_2 state. The energy of the S_0-S_2 transition decreases with the conjugation length (N) and can be

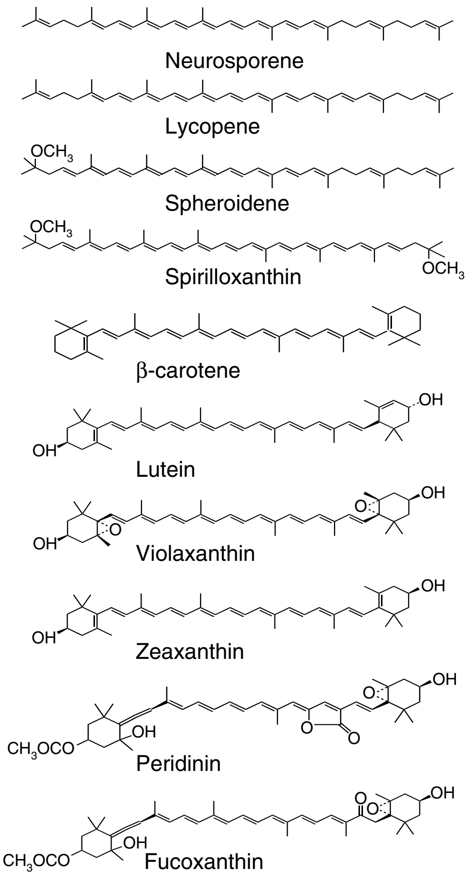
Neurosporene
Lycopene
OCH3
Spheroidene
OCH3
Spirilloxanthin
OCH3
β-carotene
OH
OH
Lutein
OH
O
O
OH
Violaxanthin
OH
OH
Zeaxanthin
OH
O
OH
O
O
CH3OCO
Peridinin
OH
O
O
OH
CH3OCO
Fucoxanthin

Fig. 5.6 A selection of carotenoid molecules found in photosynthetic organisms.

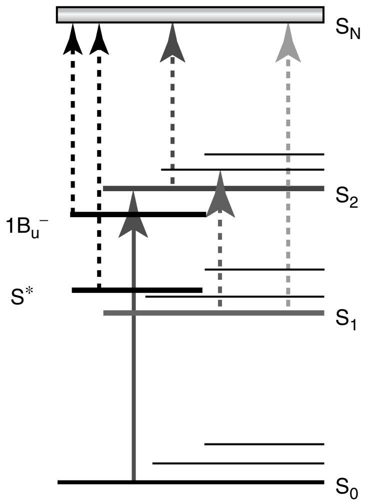

Fig. 5.7 Carotenoid energy level diagram.

in the first approximation expressed as $E = A + B/N$, where the parameter A is the energy of the 0–0 vibrational band of the S_0–S_2 transition for an infinitely long carotenoid (Polivka and Sundstrom, 2004 and references therein). Besides the number of conjugated double bonds, the 0–0 energy is also affected by the structure of the carotenoid. For example, the nonplanarity of the conjugated C=C bond located at the β-ionylidene ring leads to a decrease of the effective conjugation length. This results in higher S_2 energy of β-carotene and zeaxanthin than for their linear counterpart lycopene, which also possesses 11 conjugated double bonds.

Longer carotenoids violate Kasha's rule as they exhibit emission from the S_2 state. The critical conjugation length for which the S_1 emission that dominates for shorter conjugation changes to S_2 emission is around eight conjugated C=C bonds. Studies of a series of open-chain carotenoids demonstrated that those with a conjugation length of eight exhibit dual emission from both the S_1 and S_2 states (Frank *et al.*, 2002). The S_2 emission constitutes almost a mirror image of the absorption spectrum with a Stokes shift of 150–300 cm^{-1}, almost independent on conjugation length for carotenoids with conjugation lengths 9–13 (Christensen *et al.*, 1999; Fujii *et al.*, 2001).

While the radiative lifetime of the S_2 state as calculated by the Strickler–Berg equation (Strickler and Berg, 1962) is in the range of nanoseconds, measured quantum yields of the S_2 emission of the order of 10^{-5} showed that internal conversion on the subpicosecond time scale governs the S_2 dynamics. Thus, information about the dynamics of the S_2 state relies solely on ultrafast time-resolved techniques. Application of such techniques to a number of carotenoids employed in photosynthesis established that the S_2 lifetime is in the range of 100–300 fs (Ricci *et al.*, 1996; Akimoto *et al.*, 1999, 2002) and dependent on both conjugation length and solvent parameters. The very short excited state lifetime of the carotenoid S_2 state implies that the properties of the S_1 state are crucial for understanding photophysics of carotenoids, in other words, much of the functions of carotenoids involving excited states will originate from the S_1 state. Because of the forbidden nature of the S_0–S_1 transition, knowledge about energetics and dynamics of this key excited state was until recently very limited. Since the theoretical prediction in 1972 of the $2A_g^-$ state being the lowest excited state in longer polyenes and carotenoids, it took more than 20 years until the energy of this state could be determined experimentally for longer carotenoids.

Much effort has gone into determining the energy of the dark S_1 state of naturally occurring carotenoids having conjugation lengths in the range 7–13. An indirect approach that has been used is to determine the S_1 energy of short carotenoids from their S_1 fluorescence and extrapolate to longer carotenoids. Progress in very sensitive fluorescence detection during the 1990s enabled direct measurement of the very weak S_1 fluorescence even from photosynthetic carotenoids with conjugation lengths 9–13. Raman spectroscopy is another method that has been employed in carotenoid research. Under the conditions of resonance Raman when the excitation light is in resonance with an electronic transition of the studied molecule, the intensities of the Raman lines are greatly enhanced. Thus, tuning the detection to the frequency of either C=C or C–C stretching modes and scanning the excitation light over a broad spectral range, a resonance Raman profile can be obtained. The high sensitivity of Raman line intensities to the resonance conditions enables detection of even dark states with very low transition dipole moments. Using this technique, the S_1 state energy of several carotenoids has been determined.

A different approach employing femtosecond time-resolved spectroscopy to locate the S_1 energy of carotenoids, based on the fact that the S_1–S_2 transition is symmetry allowed, emerged in 1999 (Polivka *et al.*, 1999). A femtosecond excitation pulse populates the lowest vibrational band of the well-characterized S_2 state to avoid contribution from vibrational relaxation within this state. The excited molecule then relaxes to the S_1 state on the time scale of 50–300 fs as described above. By scanning the wavelength of the probe pulse within the time window dictated by the lifetime of the S_1 state, the S_1–S_2 resonance can be found. Given the known spectral profile of the S_2 state and lifetime of the S_1 state, the S_1–S_2 transition must (i) reflect the spacing between the S_2 vibrational bands and (ii) decay with the S_1 lifetime. If these two conditions are fulfilled, the energies of the vibrational bands of the S_1–S_2 transition can be determined. The precise location of the carotenoid S_1 state can then be obtained, since its 0–0 spectral origin can be calculated from the spectral origins of the S_0–S_2 and S_1–S_2 transitions. Although this approach is quite straightforward, its application requires tunable femtosecond pulses in the 1–2 μm spectral range where the S_1–S_2 transition is expected. Along with the availability of laser sources generating such pulses in the late 1990s, this method became a powerful tool to study the properties of the S_1 state of carotenoids. Typical S_1–S_2 profiles of two carotenoids with different conjugation lengths are shown in Figure 5.8. The first application of this method was used to locate the S_1

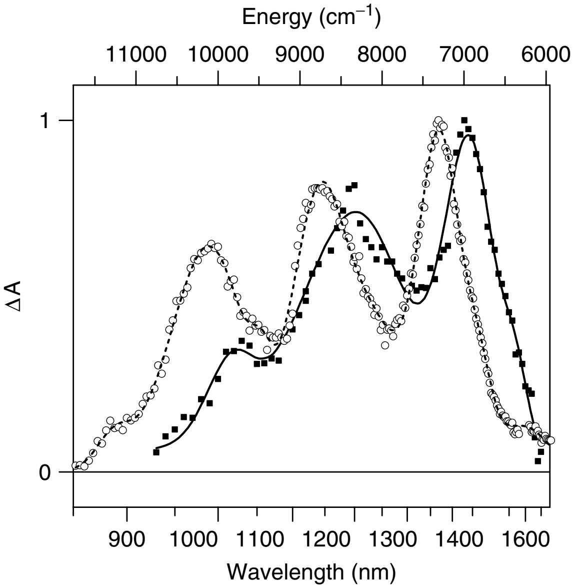

Fig. 5.8 S_1-S_2 transient absorption spectra of spheroidene in *n*-hexane (open circles) and violaxanthin in methanol (full squares) in the spectral region 850–1750 nm. From these spectra and the ground state absorption spectrum the energy of the forbidden lowest excited state (S_1) can be obtained (Polivka *et al.*, 1999).

energies of two carotenoids, violaxanthin and zeaxanthin, that participate in the xanthophyll cycle of higher plants, $14\,030 \pm 90\,\text{cm}^{-1}$ for zeaxanthin and $14\,470 \pm 90\,\text{cm}^{-1}$ for violaxanthin (Polivka *et al.*, 1999).

Comparing the various methods used to locate the S_1 energy of carotenoids, it is clear that all of these methods have their own advantages and disadvantages, and success of their application depends on the studied system. While fluorescence and resonance Raman methods are useful tools for measurements of the S_1 energy of shorter carotenoids in solution, their use is rather limited for carotenoids having more than 10 C=C bonds, since the signals become extremely weak, putting severe constraints on a reliable data analysis. On the other hand, the S_1-S_2 spectra are more reliable for longer carotenoids, since the longer carotenoids exhibit a larger S_1-S_2 energy gap, shifting the S_1-S_2 profile toward higher energies, thus making the detection easier. For carotenoids with $N \leq 9$, the 0–0 band of the S_1-S_2 transition can extend below $5000\,\text{cm}^{-1}$, where vibrational overtones originating from solvent may occur. Although higher vibrational bands of the S_1-S_2 transition can be used to obtain the S_1 energy, the precision is certainly lower in such a case. Despite these complications, the S_1-S_2 method is the only one providing information about both energy and lifetime of the S_1 state, which makes it very useful in establishing excited state properties.

5.9
Light Harvesting by Carotenoid Molecules

Chlorophyll and bacteriochlorophyll lack optical absorption in the blue-green part of the spectrum, while carotenoids have intense absorptions. Carotenoids therefore act as accessory light-harvesting pigments, but they have also many other important functions in photosynthesis. In addition to light harvesting, we will discuss their role in photoprotection. That carotenoids can transfer energy to Chls and BChls was known long before the X-ray structures were resolved. Initial suggestions regarding mechanisms and pathways of carotenoid–BChl energy transfer in LH2 and LH1 complexes proposed energy transfer via the S_1 state, because of extremely fast deactivation of the S_2 state. The forbidden nature of the S_1 state led to a suggestion that the Dexter electron exchange mechanism (Dexter, 1953) is active in this process (Cogdell and Frank, 1987). However, the first time-resolved experiments performed in the early 1990s showed that following the excitation of carotenoids the B850 Q_y state was populated in less than 200 fs, signaling that both the S_1 and S_2 states may be efficient donors in the carotenoid–BChl energy transfer process (Trautman *et al.*, 1990). Calculations of carotenoid–BChl energy transfer also led to the conclusion that both the S_1 and S_2 states can be active in energy transfer (Nagae *et al.*, 1993). A large number of time-resolved studies on various LH1 and LH2 complexes later showed that energy transfer occurs from both the S_1 and S_2 states of carotenoids to the Q_y and Q_x excited states of BChl (for more details see Polivka and Sundstrom, 2004). All the documented energy transfer pathways of an LH2 complex are summarized in Figure 5.9. Energy transfer from the optically excited carotenoid S_2 state was shown to occur in competition with the ultrafast S_2-S_1 internal conversion; transfer times varying between $\sim$50 and 150 fs, yielding energy transfer efficiencies for the S_2-Q_x channel of $\sim$50 to 60% have been reported. The mechanism of S_2 energy transfer was concluded to be of the Förster-type (Damjanovic, Ritz and Schulten, 1999), but contrary to the point dipole–point dipole interaction used in the Förster formula (Forster, 1946), full Coulombic interaction is usually used to calculate couplings between carotenoids and BChl. The reason for this is that the point-dipole description of interacting chromophores inherent in the traditional Förster mechanism fails for extended molecules at interaction distances comparable to or shorter than the size of

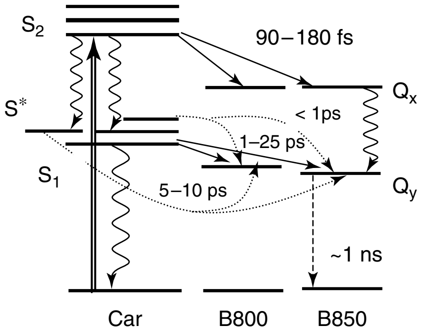

Fig. 5.9 Schematic energy level scheme for energy transfer from carotenoids to BChls in LH2. Major pathways are indicated with full arrows, while minor paths are dotted.

the involved molecules. The contribution of the electron-exchange (Dexter) energy transfer mechanism (Dexter, 1953) to S_2 energy transfer is negligible (Polivka and Sundstrom, 2004).

The energy level scheme of Figure 5.9 suggests that the carotenoid S_1 state must be higher in energy than the Q_y states of B800 and B850 BChls for efficient carotenoid–BChl transfer via the S_1 state. Before any carotenoid S_1 energy in a light-harvesting complex had been measured, estimates based on S_1 energies in solution indicated that carotenoids with $N \leq 10$ would have sufficiently high S_1 energies to efficiently transfer energy to both B800 and B850. By using several LH2 complexes having carotenoids of different conjugation lengths and comparing the S_1 lifetime of the carotenoids in solution and protein, this expectation was confirmed – carotenoids with $N \geq 11$ were shown to have very little S_1 energy transfer to BChl, while shorter carotenoids exhibited transfer efficiencies of 80–95%. The energetic requirement for efficient S_1-mediated energy transfer was finally demonstrated by direct measurement of carotenoid S_1 energies in light-harvesting complexes through recording of S_1–S_2 spectra (Polivka *et al.*, 2002a, 2002b). Both B800 and B850 were shown to accept energy from carotenoids, although B800 appears to be a somewhat more efficient acceptor. Calculations of carotenoid S_1–BChl energy transfer for several LH2 complexes showed that the observed S_1 transfer rates can be explained in terms of the same energy transfer mechanism as for the S_2 route. Although the very small transition dipole moment of the S_0–S_1 transition led initially to a suggestion that the Dexter mechanism had to be invoked, more detailed calculations later showed that the Dexter contribution is negligible and that higher-order Coulombic and polarization interactions dominate the S_1-mediated energy transfer (Scholes, Harcourt and Fleming, 1997).

5.10
Protection of the Photosynthetic Machinery (NPQ) – Quenching of Chlorophyll Excited States by Carotenoids

Quenching of various reactive excited states is probably the most important carotenoid function in photosynthesis. Carotenoids may quench both singlet and triplet (bacterio)chlorophyll excited states, as well as singlet oxygen. Here, we will only consider the quenching of Chl excited singlet states in the LHCII antenna complexes of green plants. Under certain conditions one of the carotenoid binding sites of LHCII can be occupied by zeaxanthin and this was suggested to trigger a structural change of the protein leading to quenching of Chl fluorescence (Crimi *et al.*, 2001), which is the origin of the photoprotective mechanism called *nonphotochemical quenching* (NPQ). This mechanism protects against damage caused by excessive photon flux by dissipating the energy thermally in PS2 (DemmigAdams and Adams, 1996; Horton, Ruban and Walters, 1996). It involves a cascade of processes triggered by high light conditions resulting in enzymatic conversion of violaxanthin to zeaxanthin by means of the xanthophyll cycle. Although a fundamental understanding of NPQ is still lacking, it is clear that one of the crucial components is a zeaxanthin-induced quenching of Chl-*a* excited states, as presence of this carotenoid reduces markedly the Chl-*a* fluorescence lifetime (Polivka *et al.*, 2002b; Richter *et al.*, 1999; Frank *et al.*, 2001). Several different mechanisms have been suggested for this quenching process. Frank *et al.* proposed the so-called "molecular gear shift mechanism" (Frank *et al.*, 1994) as a mechanism for NPQ. In the initial proposal based on S_1 energies estimated from the energy gap law, violaxanthin was supposed to have its S_1 energy above the Q_y transition of Chl-*a* serving as energy acceptor, while zeaxanthin had an S_1 energy sufficiently low to act as quencher of excited Chl-*a*. Then, exchange of violaxanthin to zeaxanthin by means of the xanthophyll cycle would enable increase of quencher concentration, explaining the mechanism of NPQ. Later, direct measurements of the S_1 energies in both solution and LHCII (Polivka *et al.*, 1999) challenged this mechanism, since the S_1 energies of both violaxanthin and zeaxanthin were shown to be below the Chl-*a* Q_y state. Nevertheless, although the initially proposed "molecular gear shift" would not work, the location of the zeaxanthin S_1 state below the Chl-*a* Q_y transition still allows direct quenching. Direct involvement of the zeaxanthin S_1 state in NPQ was also shown in a femtosecond transient absorption study of intact thylakoid membranes of *thaliana* (Ma *et al.*, 2003; Holt *et al.*, 2005). After excitation of either Chl-*b* or Chl-*a* by ~100 fs pulses, the zeaxanthin S_1 state was observed to be directly populated, signaling very fast energy transfer from Chl to zeaxanthin (and thus Chl excited state quenching). The formation of the zeaxanthin S_1 state was correlated to the presence of the PsbS protein that is known to be an essential part of NPQ (Li *et al.*, 2000), thus confirming the relation between population of the zeaxanthin S_1 state and NPQ. Further, time-resolved spectroscopy work (Holt *et al.*, 2005) and calculations (Ma *et al.*, 2003) indicated that the quenching mechanism could involve energy transfer from chlorophyll molecules to a chlorophyll–zeaxanthin heterodimer, which then undergoes charge separation and formation of a zeaxanthin$^+$–Chl$^-$ radical pair. Another hypothesis proposed that the essential function of zeaxanthin is to promote changes in structure and/or organization of the light-harvesting complexes that eventually lead to quenching conditions. One such mechanism involves a structural change facilitating a formation of Chl-*a* excitonic pairs, which than play a role of quenchers (Crofts and Yerkes, 1994). Clearly, additional work is required before the mechanism of NPQ is fully clarified.

5.11
Storing the Energy of Light – Photosynthetic Charge Separation

As already discussed above, excitation energy captured by the antenna is transferred to the pigments of the RC where a sequence of ultrafast electron transfer steps convert the excited state energy in the antenna to a transmembrane potential. The core of the RC is made up of two homologous membrane proteins, called L and M, while a third polypeptide, H, covers the cytosolic surface (Deisenhofer *et al.*, 1985). Both L and M possess five membrane-crossing helices forming a cage around the four BChls, two bacteriopheophytins and two quinones noncovalently bound to the RC (Figure 5.10). With respect to both the proteins and the cofactors, the RC shows

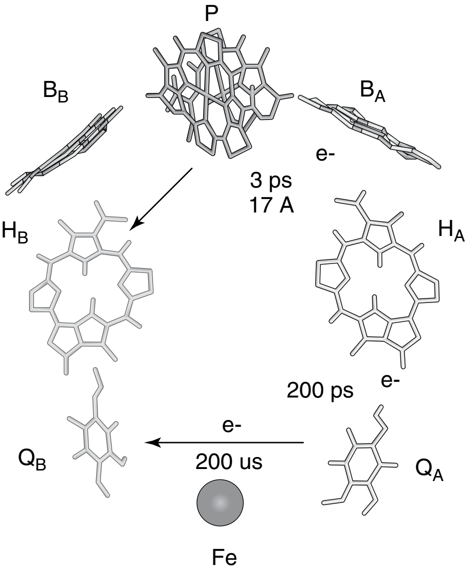

Fig. 5.10 Structure of the bacterial RC, showing the cofactors. P – special pair; B_A and B_B – accessory monomeric chlorophylls; H_A and H_B – pheophytins; Q_A and Q_B – quinines.

a remarkable C_2-symmetry with the symmetry axis running through the middle of the centrally located ‘‘special pair’’ (P) of BChl molecules and the nonheme iron. The pigments are arranged in two branches labeled A and B, but surprisingly, electron transfer occurs only along the A (or ‘‘active’’) branch. Following excitation of the special pair (transfer of excitation energy from the antenna), charge separation occurs within a few picoseconds to yield the oxidized special pair, P^+, and a reduced bacteriopheophytin, $H_A{}^-$. The electron then hops to a quinone molecule, Q_A, in about 200 ps, and on to a second quinone Q_B in 100 μs. P^+ is restored to neutrality by electron transfer from a cytochrome. This sequence of reactions is repeated after the absorption of a second photon, and then the doubly reduced Q_B molecule leaves the RC as Q_BH_2 and an electrochemical gradient has been created across the photosynthetic membrane in which the RC is positioned.

The understanding that primary charge separation occurs within ~3 ps with the formation of P^+ and $H_A{}^-$ was a result of subpicosecond absorption measurements in the mid-1980s. These results generated much discussion on the nature of the electron transfer between the special pair and the bacteriopheophytin. Taking into account that the edge-to-edge distance between P and H_A is 11 Å, a simple model predicted the time-constant for direct electron transfer to be approximately 1 ns. From the crystal structure of the RC it was apparent that the monomeric BChl (B_A) is ideally situated to be an intermediate in the electron transfer pathway $P \rightarrow B_A \rightarrow H_A$. However, in the time-resolved measurements there were no known spectral changes that could be unambigously interpreted as a transient population of a $B_A{}^-$ species. To solve these difficulties, a mechanism was proposed that included B_A as a virtual intermediate. In this so-called superexchange mechanism it is assumed that the coupling between P and H_A is provided by B_A, but that the radical pair $P^+B_A^-$ is never significantly populated since it is far above the excited state of P (P*) in free energy. Today, most workers in the field support the proposal originally made by Shuvalov and coworkers (Chekalin *et al.*, 1987) and later verified by Zinth and coworkers (Holzapfel *et al.*, 1990; Arlt *et al.*, 1993) that B_A is transiently observable as an electron transfer intermediate, with the transfer time of the electron away from P* (3–4 ps) almost three times slower than the transfer away to H_A (about 1 ps). An important confirmation that $P^+B_A^-$ is indeed a detectable intermediate in charge separation came from the discovery of the B_A^- anion absorption band around 1020 nm.

In spite of the obvious symmetry of the RC, the electron transfer essentially occurs only along the A-branch. Estimates of the coupling strengths along both paths do indeed favor the active branch, but not in the experimentally observed 200 : 1 ratio. One conspicuous feature in the RC structure is a tyrosine residue (TyrM208 in *Rhodopseudomonas viridis*, TyrM210 in *sphaeroides*) strategically positioned between the cofactors in the active branch, while the symmetry-related residue in the inactive branch is a phenylalanine (PheL181). A variety of groups studied RCs with mutations introduced at these positions and they observed dramatic changes in the rate of charge separation (e.g. Jia *et al.*, 1993; Visschers *et al.*, 1999). For most mutations the rate slowed down, sometimes dramatically. Surprisingly, for a mutation that restored the symmetry with tyrosines at positions L181 and M208 the rate increased slightly. These observations were explained within the framework of the Marcus theory (Marcus and Sutin, 1985) for electron transfer, and ascribed to changes in the free energy of the initial radical pair due to the mutation. The free energy dependence of the rate suggested that any change in the protein structure during electron transfer (the reorganization energy) is very small. Nevertheless, in none of the mutants was electron transfer observed along the inactive branch. Other types of mutants were also generated – mutants in which H-bonds of the protein were modified around P produced dramatic changes in its redox potential, again without affecting the asymmetry of electron transfer. In a double mutant with H_A replaced by a bacteriochlorophyll and $P^+B_A^-$ upshifted in free energy, Holten and coworkers detected a measurable amount of $P^+H_B^-$ (Kirmaier *et al.*, 1991). This suggests that the free energy of the charge-separated states controls the asymmetry of the electron transfer, rather than the differences in electronic coupling.

Following excitation of bacterial RCs with femtosecond excitation pulses, oscillatory patterns have been observed in the transient absorption changes (Vos *et al.*, 1993) and spontaneous emission (Stanley and Boxer, 1995). Typically these phenomena are explained by the generation of coherent wavepackets on the excited state surface due to the broad spectral bandwidth of the short excitation pulse. The Fourier transform spectrum of these oscillations indicated dominant modes at 15 and 77 cm^{-1} and a smaller contribution from a mode at 90–100 cm^{-1}. The modes are consistent with the ground state resonance Raman spectrum of P and possibly result from the protein or are weak low-frequency modes from the BChls that are enhanced in the special pair configuration. Observation of these oscillatory signals questions the conventional assumption that vibrational dephasing and relaxation occur on a time scale that is much shorter than the electron transfer, and therefore challenges conventional theory used for describing the primary electron transfer steps in RCs.

The charge transfer processes described above account for the main function of the RC converting the light energy delivered to the special pair from the light-harvesting antenna into a stable charge separation across the photosynthetic membrane. All of the pigments of the RC can of course directly absorb light as well; however, due to their relatively low concentration as compared to LH pigments, this contribution plays no practical role for the total light absorption of the PSU. Nonetheless, for characterizing the various pathways of energy and electron flow within the RC it is of interest to study these processes. It is generally assumed that excitation of all other pigments than P results in ultrafast energy transfer to P (Breton *et al.*, 1986), followed by the conventional charge separation to form the $P^{+}H_{A}^{-}$ radical pair. Indeed, a variety of ultrafast measurements in which B_A and B_B were selectively excited demonstrated that energy transfer to P occurs within a few hundred femtoseconds (Jonas *et al.*, 1996). It has also been demonstrated that direct excitation of the accessory B_A pigment leads to a charge separation $P^{+}H_{A}^{-}$ within ~200 fs, without the involvement of P* (van Brederode *et al.*, 1999). In the PS2 RC of green plants, the charge separation starting at the accessory chlorophyll was shown to be functionally important. From the first purification work on the PS2 RC, showing strong homology between the bacterial and PS2 RCs, it was speculated that the operation of the two RCs would be similar. However, the spectroscopic similarity of all chromophores in the PS2 RC led to a proposal that charge separation in the PS2 RC follows the minor pathway of charge transfer demonstrated for bacterial RCs and starts at the accessory chlorophyll. Data from spectroscopic and time-resolved measurements combined with theoretical modeling indeed now strongly suggest that the primary charge separation occurs between the accessory chlorophyll in the D_1 subunit and the pheophytin in the same subunit. Only by a secondary electron transfer step is the radical pair involving the special pair P680 and the pheophytin, $P680^{+}$-H^{-}, formed (Groot *et al.*, 2005; Novoderezhkin *et al.*, 2005).

Acknowledgments

I wish to thank all my coworkers at the Department of Chemical Physics, Lund University, for fruitful discussions. I thank prof. R. van Grondelle for the spectrum of Figure 5.3a.

References

Akimoto, S., Yamazaki, I., Sakawa, T., and Mimuro, M. (2002) *J. Phys. Chem. A*, **106**, 2237–43.

Akimoto, S., Yamazaki, I., Takaichi, S., and Mimuro, M. (1999) *Chem. Phys. Lett.*, **313**, 63–68.

Arlt, T., Schmidt, S., Kaiser, W., Lauterwasser, C., Meyer, M., Scheer, H., and Zinth, W. (1993) *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 11757–61.

Breton, J., Martin, J.L., Migus, A., Antonetti, A., and Orszag, A. (1986) *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 5121–25.

Chachisvilis, M., Kuhn, O., Pullerits, T., and Sundstrom, V. (1997) *J. Phys. Chem. B*, **101**, 7275–83.

Chekalin, S.V., Matveets, Y.A., Shkuropatova, A.Y., Shuvalov, V.A., and Yartsev, A.P. (1987) *FEBS Lett.*, **216**, 245–48.

Christensen, R.L., Goyette, M., Gallagher, L., Duncan, J., DeCoster, B., Lugtenburg, J., Jansen, F.J., and van der Hoef, I. (1999) *J. Phys. Chem. A*, **103**, 2399–407.

Cogdell, R.J. and Frank, H.A. (1987) *Biochim. Biophys. Acta*, **895**, 63–79.

Crimi, M., Dorra, D., Bosinger, C.S., Giuffra, E., Holzwarth, A.R., and Bassi, R. (2001) *Eur. J. Biochem.*, **268**, 260–67.

Croce, R., Dorra, D., Holzwarth, A.R., and Jennings, R.C. (2000) *Biochemistry*, **39**, 6341–48.

Croce, R., Muller, M.G., Bassi, R., and Holzwarth, A.R. (2003) *Biophys. J.*, **84**, 2508–16.

Crofts, A.R. and Yerkes, C.T. (1994) *FEBS Lett.*, **352**, 265–70.

Dahlbom, M., Pullerits, T., Mukamel, S., and Sandstrom, V. (2001) *J. Phys. Chem. B*, **105**, 5515–24.

Damjanovic, A., Ritz, T., and Schulten, K. (1999) *Phys. Rev. E*, **59**, 3293–311.

Deisenhofer, J., Epp, O., Miki, K., Huber, R., and Michel, H. (1985) *Nature*, **318**, 618–24.

DemmigAdams, B. and Adams, W.W. (1996) *Trends Plant Sci.*, **1**, 21–26.

Denhollander, W.T.F., Bakker, J.G.C., and vanGrondelle, R. (1983) *Biochim. Biophys. Acta*, **725**, 492–507.

Dexter, D.L. (1953) *J. Chem. Phys.*, **21**, 836–50.

Fenna, R.E. and Matthews, B.W. (1975) *Nature*, **258**, 573–77.

Ferreira, K.N., Iverson, T.M., Maghlaoui, K., Barber, J., and Iwata, S. (2004) *Science*, **303**, 1831–38.

Forster, T. (1946) *Naturwissenschaften*, **33**, 166–75.

Frank, H.A. and Cogdell, R.J. (1993) *Photochemistry of Carotenoids*, Chapman and Hall, 252–70.

Frank, H.A. and Cogdell, R.J. (1996) *Photochem. Photobiol.*, **63**, 257–64.

Frank, H.A., Cua, A., Chynwat, V., Young, A., Gosztola, D., and Wasielewski, M.R. (1994) *Photosynth. Res.*, **41**, 389–95.

Frank, H.A., Das, S.K., Bautista, J.A., Bruce, D., Vasil'ev, S., Crimi, M., Croce, R., and Bassi, R. (2001) *Biochemistry*, **40**, 1220–25.

Frank, H.A., Josue, J.S., Bautista, J.A., van der Hoef, I., Jansen, F.J., Lugtenburg, J., Wiederrecht, G., and Christensen, R.L. (2002) *J. Phys. Chem. B*, **106**, 2083–92.

Fraser, N.J., Dominy, P.J., Ucker, B., Simonin, I., Scheer, H., and Cogdell, R.J. (1999) *Biochemistry*, **38**, 9684–92.

Freiberg, A., Godik, V.I., Pullerits, T., and Timpman, K. (1989) *Biochim. Biophys. Acta*, **973**, 93–104.

Fujii, R., Ishikawa, T., Koyama, Y., Taguchi, M., Isobe, Y., Nagae, H., and Watanabe, Y. (2001) *J. Phys. Chem. A*, **105**, 5348–55.

Gantt, E. (1981) *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, **32**, 327–47.

Groot, M.L., Pawlowicz, N.P., van Wilderen, L.J.G.W., Breton, J., van Stokkum, I.H.M., and van Grondelle, R. (2005) *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 13087–92.

Herek, J.L., Fraser, N.J., Pullerits, T., Martinsson, P., Polivka, T., Scheer, H., Cogdell, R.J., and Sundstrom, V. (2000) *Biophys. J.*, **78**, 2590–96.

Hess, S., Chachisvilis, M., Timpmann, K., Jones, M.R., Fowler, G.J.S., Hunter, C.N., and Sundstrom, V. (1995) *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 12333–37.

Holt, N.E., Zigmantas, D., Valkunas, L., Li, X.P., Niyogi, K.K., and Fleming, G.R. (2005) *Science*, **307**, 433–36.

Holzapfel, W., Finkele, U., Kaiser, W., Oesterhelt, D., Scheer, H., Stilz, H.U., and Zinth, W. (1990) *Proc. Natl. Acad. Sci. U.S.A.*, **87**, 5168–72.

Horton, P., Ruban, A.V., and Walters, R.G. (1996) *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, **47**, 655–84.

Jia, Y.W., Dimagno, T.J., Chan, C.K., Wang, Z.Y., Du, M., Hanson, D.K., Schiffer, M., Norris, J.R., Fleming, G.R., and Popov, M.S. (1993) *J. Phys. Chem.*, **97**, 13180–91.

Jonas, D.M., Lang, M.J., Nagasawa, Y., Joo, T., and Fleming, G.R. (1996) *J. Phys. Chem.*, **100**, 12660–73.

Jordan, P., Fromme, P., Witt, H.T., Klukas, O., Saenger, W., and Krauss, N. (2001) *Nature*, **411**, 909–17.

Karapetyan, N.V., Holzwarth, A.R., and Rogner, M. (1999) *FEBS Lett.*, **460**, 395–400.

Kirmaier, C., Gaul, D., Debey, R., Holten, D., and Schenck, C.C. (1991) *Science*, **251**, 922–27.

Koyama, Y., Kuki, M., Andersson, P.O., and Gillbro, T. (1996) *Photochem. Photobiol.*, **63**, 243–56.

Leupold, D., Stiel, H., Teuchner, K., Nowak, F., Sandner, W., Ucker, B., and Scheer, H. (1996) *Phys. Rev. Lett.*, **77**, 4675–78.

Li, X.P., Bjorkman, O., Shih, C., Grossman, A.R., Rosenquist, M., Jansson, S., and Niyogi, K.K. (2000) *Nature*, **403**, 391–95.

Ma, Y.Z., Holt, N.E., Li, X.P., Niyogi, K.K., and Fleming, G.R. (2003) *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 4377–82.

Marcus, R.A. and Sutin, N. (1985) *Biochim. Biophys. Acta*, **811**, 265–322.

Mauzerall, D. (1976) *Biophys. J.*, **16**, 87–91.

Mcdermott, G., Prince, S.M., Freer, A.A., Hawthornthwaitelawless, A.M., Papiz, M.Z., Cogdell, R.J., and Isaacs, N.W. (1995) *Nature*, **374**, 517–21.

Meier, T., Chernyak, V., and Mukamel, S. (1997) *J. Phys. Chem. B*, **101**, 7332–42.

Monshouwer, R., Abrahamsson, M., van-Mourik, F., and vanGrondelle, R. (1997) *J. Phys. Chem. B*, **101**, 7241–48.

Nagae, H., Kakitani, T., Katoh, T., and Mimuro, M. (1993) *J. Chem. Phys.*, **98**, 8012–23.

Nagarajan, V. and Parson, W.W. (1997) *Biochemistry*, **36**, 2300–06.

Novoderezhkin, V.I., Andrizhiyevskaya, E.G., Dekker, J.P., and van Grondelle, R. (2005) *Biophys. J.*, **89**, 1464–81.

Olson, J.M. (1980) *Biochim. Biophys. Acta*, **594**, 33–51.

Polivka, T. and Sundstrom, V. (2004) *Chem. Rev.*, **104**, 2021–71.

Polivka, T., Herek, J.L., Zigmantas, D., Akerlund, H.E., and Sundstrom, V. (1999) *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 4914–17.

Polivka, T., Zigmantas, D., Herek, J.L., He, Z., Pascher, T., Pullerits, T., Cogdell, R.J., Frank, H.A., and Sundstrom, V. (2002a) *J. Phys. Chem. B*, **106**, 11016–25.

Polivka, T., Zigmantas, D., Sundstrom, V., Formaggio, E., Cinque, G., and Bassi, R. (2002b) *Biochemistry*, **41**, 439–50.

Pullerits, T., Chachisvilis, M., and Sundstrom, V. (1996) *J. Phys. Chem.*, **100**, 10787–92.

Pullerits, T. and Sundstrom, V. (1996) *Acc. Chem. Res.*, **29**, 381–89.

Ricci, M., Bradforth, S.E., Jimenez, R., and Fleming, G.R. (1996) *Chem. Phys. Lett.*, **259**, 381–90.

Richter, M., Goss, R., Wagner, B., and Holzwarth, A.R. (1999) *Biochemistry*, **38**, 12718–26.

Ritz, T., Damjanovic, A., Schulten, K., Zhang, J.P., and Koyama, Y. (2000) *Photosynth. Res.*, **66**, 125–44.

Roszak, A.W., Howard, T.D., Southall, J., Gardiner, A.T., Law, C.J., Isaacs, N.W., and Cogdell, R.J. (2003) *Science*, **302**, 1969–72.

Scholes, G.D. and Fleming, G.R. (2000) *J. Phys. Chem. B*, **104**, 1854–68.

Scholes, G.D., Harcourt, R.D., and Fleming, G.R. (1997) *J. Phys. Chem. B*, **101**, 7302–12.

Schubert, A., Stenstam, A., Beenken, W.J.D., Herek, J.L., Cogdell, R., Pullerits, T., and Sundstrom, V. (2004) *Biophys. J.*, **86**, 2363–73.

Stanley, R.J. and Boxer, S.G. (1995) *J. Phys. Chem.*, **99**, 859–63.

Strickler, S.J. and Berg, R.A. (1962) *J. Chem. Phys.*, **37**, 814–25.

Sumi, H. (1999) *J. Phys. Chem. B*, **103**, 252–60.

Sundstrom, V., Pullerits, T., and van Grondelle, R. (1999) *J. Phys. Chem. B*, **103**, 2327–46.

Sundstrom, V., vanGrondelle, R., Bergstrom, H., Akesson, E., and Gillbro, T. (1986) *Biochim. Biophys. Acta*, **851**, 431–46.

Timpmann, K., Zhang, F.G., Freiberg, A., and Sundstrom, V. (1993) *Biochim. Biophys. Acta*, **1183**, 185–93.

Trautman, J.K., Shreve, A.P., Violette, C.A., Frank, H.A., Owens, T.G., and Albrecht, A.C. (1990) *Proc. Natl. Acad. Sci. U.S.A.*, **87**, 215–19.

Trinkunas, G., Herek, J.L., Polivka, T., Sundstrom, V., and Pullerits, T. (2001) *Phys. Rev. Lett.*, **86**, 4167–70.

vanGrondelle, R. (1985) *Biochim. Biophys. Acta*, **811**, 147–95.

vanGrondelle, R., Dekker, J.P., Gillbro, T., and Sundstrom, V. (1994) *Biochim. Biophys. Acta*, **1187**, 1–65.

van Grondelle, R. and Novoderezhkin, V.I. (2006) *Phys. Chem. Chem. Phys.*, **8**, 793–807.

van Amerongen, H. and van Grondelle, R. (2001) *J. Phys. Chem. B*, **105**, 604–17.

van Brederode, M.E., van Stokkum, I.H.M., Katilius, E., van Mourik, F., Jones, M.R., and van Grondelle, R. (1999) *Biochemistry*, **38**, 7545–55.

Visscher, K.J., Bergstrom, H., Sundstrom, V., Hunter, C.N., and vanGrondelle, R. (1989) *Photosynth. Res.*, **22**, 211–17.

Visschers, R.W., Vulto, S.I.E., Jones, M.R., van Grondelle, R., and Kraayenhof, R. (1999) *Photosynth. Res.*, **59**, 95–104.

Vos, M.H., Rappaport, F., Lambry, J.C., Breton, J., and Martin, J.L. (1993) *Nature*, **363**, 320–25.

Zouni, A., Witt, H.T., Kern, J., Fromme, P., Krauss, N., Saenger, W., and Orth, P. (2001) *Nature*, **409**, 739–43.

6
Optical Tweezers Techniques

Lene B. Oddershede

6.1 **Introduction 197**
6.2 **Basic Principles 198**
6.2.1 Rayleigh Regime, $d \ll \lambda$ 198
6.2.2 Mie Regime, $d \gg \lambda$ 199
6.2.3 Intermediate Regime, $d \simeq \lambda$ 200
6.2.4 Finding the Forces 200
6.2.4.1 Stokes Drag Calibration 201
6.2.4.2 Brownian Motion Calibration 202
6.2.5 Proximity to Surfaces 203
6.3 **Equipment 204**
6.4 **Concerns 206**
6.4.1 Biological Damage 206
6.4.2 Heating of Specimen 207
6.5 **Technical Improvements 208**
6.5.1 Multiple Optical Traps 208
6.5.2 Compensating Spherical Aberrations 209
6.6 **What Can Be Trapped? 210**
6.7 **Applications 210**
6.7.1 Molecular Motors 211
6.7.1.1 Kinesin 211
6.7.1.2 RNA Polymerase 211
6.7.1.3 Viral Rotary Motor 211
6.7.2 Unfolding/Disassembly of Macromolecules 212
6.7.2.1 Release of DNA from Nucleosomes 212
6.7.2.2 RNA Structures 212
6.7.3 Cell Biophysics 213
6.8 **Future Directions 213**
References 213
Further Reading 214

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.

Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

ISBN: 978-3-527-40702-6

6.1
Introduction

Optical tweezers are a unique tool providing the ability to manipulate specimen with nanometer precision just by the use of light. With proper detection equipment, optical tweezers can also be used to measure the forces acting on, or by, a given micro- or nanoscale system. These are distances and forces typically present in biological systems at the macromolecular level. Therefore, optical tweezers have mediated new exciting discoveries within the field of bioscience. With proper chosen wavelength and intensity of the trapping laser light, optical tweezers can nearly nondestructively or noninvasively trap living microorganisms such as virus, bacteria, or eucaryotic cells. Besides, optical tweezers are the only nanotool that can measure forces on and manipulate naturally occurring objects, as, for example, organelles, inside living cells without destructively perturbing the cell wall.

Though the development of optical tweezers started only in the 1970s, the fact that light can exert force on particles has been known for centuries. Johannes Kepler proposed that radiation pressure from the sun is responsible for the tail of comets always pointing away from the sun. The basic physical principle behind this postulate was put into a theoretically consistent framework by J.C. Maxwell, proving that momentum is transferred when light interacts with an object.

With the invention of the laser, a highly collimated and very powerful source of light became available. This intense light source found use in the manipulation of atoms and molecules. In 1970, Arthur Ashkin made the first version of an optical trap; it was made by two counter propagating laser beams, and with this setup, dielectric particles could be optically trapped in liquids and air (Ashkin, 1970). Trapping was mediated by the counteracting radiation pressures from two opposing laser beams. In 1986, the development of optical tweezers took another major step forward – Arthur Ashkin demonstrated that using a green argon ion laser (Ashkin *et al.*, 1986) dielectric particles could be trapped with a single, highly focused, laser beam. Adding to the potential of the technique was also the fact that he was also able to trap living microorganisms such as yeast, bacteria, and virus (Ashkin, Dziedzic and Yamane, 1987). And the microorganisms did not appear too damaged by the laser light. In contrast to the previous trap based on radiation pressure, the trapping

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

capabilities of the tightly focused single laser beam are caused by the spatial gradient force, which always pulls dielectric objects toward the region of highest intensity. By tightly focusing a laser beam he found that the gradient force of a single beam could overcome the scattering force by the same beam, hence leading to a stable three-dimensional trapping condition.

Current technical developments have evolved the technique of optical tweezers to a stage where they can be used to measure subnanometer distances and forces as feeble as femto-Newton. In the following, there is first a description of the principles of optical trapping, of how to measure forces and distances, the basic equipment, and concerns. Then focus is on the applicability of the technique, on which objects can be trapped, and on some of the very exciting biological applications that have arisen, most noteworthy among them being the unraveling of the function of a single biological molecule.

6.2
Basic Principles

The functioning of an optical trap relies on the fact that photons carry momentum, $|\mathbf{p}| = h/\lambda$, λ being the wavelength of the light and h Planck's constant. Let $\mathbf{S}$ be the Poynting vector describing the light from the laser. Upon encounter with an object, the light interacts with the object and the object feels a force, $\mathbf{F}$, due to refraction and reflection of the light by the object:

$$\mathbf{F} = \frac{n}{c}\int (\mathbf{S}_{\text{in}} - \mathbf{S}_{\text{out}})\,\mathrm{d}A \qquad (1)$$

where $\mathrm{d}A$ is the projected area of the object, n the index of refraction, and c the speed of light. This means that light exerts force on all objects that reflect or refract light. The radiation force can be decomposed in components arising from refraction and reflection, respectively. The component arising from reflection pushes the object along the propagating direction of the laser light and is known as the *scattering force*. The component arising from the refraction is termed the *gradient force*, and this force pulls an object with refractive index higher than the surrounding media toward the most intense region of the laser beam.

In principle, one can use Eq. (1) to calculate the force acting on the particle. However, in practice, this is either very difficult or not possible. To simplify the treatment of how light can trap particles, two limiting cases can be considered, namely the Rayleigh regime, where the trapped object, with typical linear dimension d is considerably smaller than λ and the Mie regime, where the d is considerably bigger than λ and where the ordinary ray-optics picture can be applied. In any case, to achieve successful trapping, the gradient force must be larger than the scattering forces.

6.2.1
Rayleigh Regime, $d \ll \lambda$

As the size of the object is considerably smaller than λ, it can be described as a point dipole. Hence, the incoming electrical field, $\mathbf{E}$, induces a dipole $\mathbf{d}$, in the dielectric object,

$$\mathbf{d} = \alpha\mathbf{E} \qquad (2)$$

where α is the polarizability of the object. The potential, U, describing the interaction between the dipole and the electromagnetic field is given by

$$U = -\mathbf{d} \cdot \mathbf{E} \tag{3}$$

This potential gives rise to a force,

$$\mathbf{F}_{\text{grad}} = -\nabla U = \alpha \nabla E^2 \tag{4}$$

which is denoted the gradient force and is proportional to the gradient of the intensity profile.

6.2.2
Mie Regime, $d \gg \lambda$

This is the case where the trapped object is considerably larger than λ, a scenario which is depicted in Figures 6.1 and 6.2. In this regime, Eq. (1) is fairly easily applicable. Consider a situation as depicted in Figure 6.1 where a laser beam is incident from above with an intensity profile that is increasing toward the center of the beam (e.g. a Gaussian intensity profile). Let a bead having an index of refraction larger than the index of refraction of the surrounding media be slightly displaced to the left with respect to the center of the beam. The most intense part of the laser beam, denoted "1" in Figure 6.1, will be refracted according to Snells law upon interaction with the bead. Comparing the direction of part "1" of the laser beam before and after encounter with the bead, it is clear that it has changed direction and hence momentum. Momentum must be conserved. Hence, the bead experiences an equally sized but oppositely directed change of momentum, $\mathbf{P}_{1,\text{bead}}$. Likewise, the momentum of the dimmer part "2" of the laser beam undergoes a change upon encounter with the bead. This gives rise to an equal in size, but opposite in direction, change of momentum of the bead, $\mathbf{P}_{2,\text{bead}}$. The sum of these two contributions is

$$\mathbf{P}_{\text{tot,bead}} = \mathbf{P}_{1,\text{bead}} + \mathbf{P}_{2,\text{bead}} \tag{5}$$

The total effect is that the bead is pushed toward the center of the laser beam. Integrating this argument over the entire laser profile gives a similar result, and hence, explains trapping in a direction orthogonal to the propagational direction of the laser light.

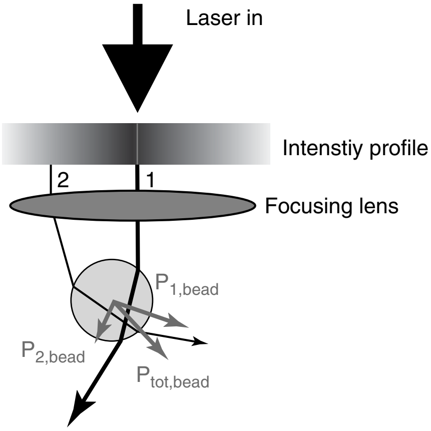

Fig. 6.1 Schematic drawing of the principle behind optical trapping in a direction orthogonal to the propagating laser light in the regime where ray optics apply.

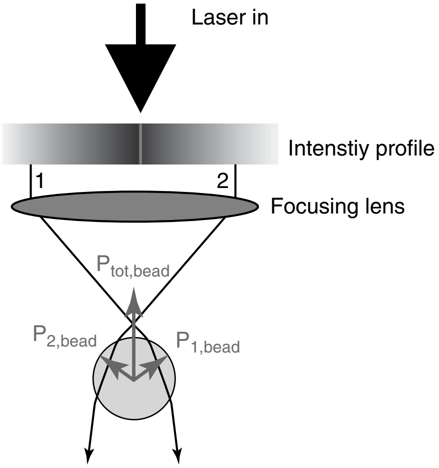

Fig. 6.2 Schematic drawing of the principle behind optical trapping in a direction parallel to the propagating laser light in the regime where ray optics apply.

To understand optical trapping in the parallel direction, consider a scenario as depicted in Figure 6.2. The laser beam has an intensity profile as in the case considered in Figure 6.1, but now it is important that the laser light passes through a focusing lens before encountering the bead. Let the bead be displaced slightly below the focus of the lens and consider two parts of the laser beam, parts "1" and "2," originating from opposite sides of the laser beam but carrying equal intensity. Again, parts "1" and "2" of the laser beam will experience a change in momentum upon passing the bead. Correspondingly, the bead experiences changes in momentum, $\mathbf{P}_{1,\text{bead}}$ and $\mathbf{P}_{2,\text{bead}}$, which are equal in size but opposite in direction to the changes of rays "1" and "2." The total contribution $\mathbf{P}_{\text{tot,bead}}$ will be directed upward the horizontal components canceling. Integrating this argument over the entire laser profile gives a similar result, and hence, explains trapping in a direction parallel to the propagational direction of the laser light.

6.2.3
Intermediate Regime, $d \simeq \lambda$

In this regime, the size of the trapped particle is on the same order as the wavelength of the trapping laser light. In fact, this is the situation most commonly encountered, and where an optical trap appears most efficient. Calculating the forces in this regime is difficult, mostly because how to apply Maxwell's equations is not trivial in a situation where the size and geometry of the object cannot be ignored. Recently, A. Rohrbach carried through a theoretical calculation based on electromagnetic force density, the so-called two-component approach (Jackson, 1975), to find the gradient and scattering forces by solving Eq. (1) (Rohrbach, 2005). The result was impressively close to experimental observations.

6.2.4
Finding the Forces

To find the forces exerted by an optical trap, consider a coordinate system where *x* and *y* are perpendicular to the direction of propagation of the laser, which is determined *z*. The three coordinates of the bead are uncorrelated, and the equation of motion for each of them is the same, but with different parameter values. Therefore, we can consider the motion in one dimension, for example in the *x* coordinate, only. The optical trap is assumed to exert a harmonic force, $\mathbf{F}_{\text{trap}}$, on a trapped object:

$$\mathbf{F}_{\text{trap}} = -\kappa \mathbf{x} \tag{6}$$

where κ denotes the spring constant of the optical trap, $\mathbf{x}$ is the deviation from the equilibrium position (center of trap), and the minus sign is due to the fact that the force always acts so as to push the object toward the center, opposite to the displacement, $\mathbf{x}$. As can be seen from the position the histogram shown in Figure 6.3, a harmonic potential is a reasonable approximation. The full line in Figure 6.3 is a Gaussian fit to the position data, and the inset shows the same but using a logarithmic ordinate axis.

As it is cumbersome, or maybe even impossible, in practice to theoretically calculate the forces exerted by optical tweezers, a far more reliable and easy method is to perform a force calibration. This can be done in several ways, of which the two most common procedures are mentioned here, namely the "Stokes drag calibration" and the "Brownian motion calibration."

6.2.4.1 Stokes Drag Calibration

Let a bead with a size well known in an aqueous chamber be held by an optical trap. By moving the trap with respect to

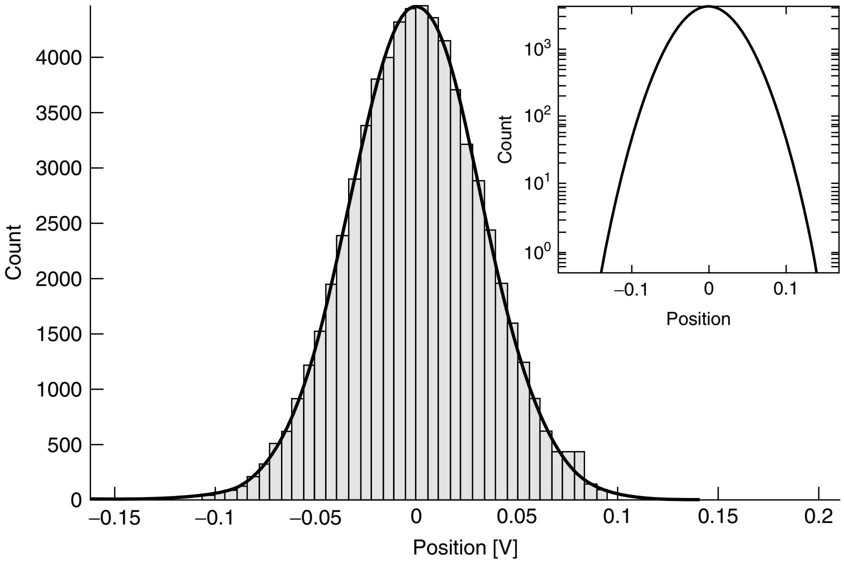

Fig. 6.3 Histogram of positions visited by a 2-µm bead held in an optical trap. The positions are measured by a photodiode in units of volts and can be converted to meters by multiplying by a conversion factor of 1.5×10^{-7} mV^{-1}. The standard deviation of the histogram is 0.03 V corresponding to ~5 nm. The full line is a fit by a Gaussian function to the data, and the inset shows the same but on semilogarithmic axes.

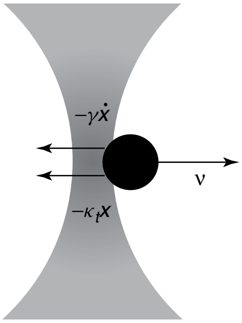

Fig. 6.4 Schematic drawing of velocity, v, of bead with respect to fluid and of the forces $\mathbf{F}_{\text{drag}} = -\gamma\dot{x}$ and $\mathbf{F}_{\text{trap}} = -\kappa_t x$ acting on an optically trapped bead.

the chamber with a constant velocity, one can exert a drag force, $\mathbf{F}_{\text{drag}}$, on the bead

$$\mathbf{F}_{\text{drag}} = -\gamma \boldsymbol{v}, \quad \gamma = 3\pi\eta d \tag{7}$$

where γ is the friction coefficient given by Stokes law, d the diameter of the bead, η the viscosity of the surrounding liquid, and $\boldsymbol{v} = \dot{x}$ the velocity of the surrounding media with respect to the bead. The minus sign shows that $\mathbf{F}_{\text{drag}}$ always points oppositely to the velocity vector, $\boldsymbol{v}$, as depicted in Figure 6.4. If $\boldsymbol{v}$ is kept constant, the bead will find a new equilibrium position, x_{new}. By measuring x_{new} and $\boldsymbol{v}$ and equating Eqs (7) and (6), the spring constant, κ characterizing the optical trap can be found.

By increasing $\boldsymbol{v}$ until the bead escapes the trap, a method known as *escape calibration*, the maximum force applicable by the optical trap can be found by using Eq. (7).

6.2.4.2 Brownian Motion Calibration

A precise and commonly used method to describe the forces exerted by optical tweezers utilizes the Brownian motion of the particle in the trap. This method requires a fast detection system (with respect to the relaxation time scales) that can determine the positions of the bead as a function of time.

To find the equation of motion of the particle, all forces acting on the particle must be considered. These are the harmonic force exerted by the optical trap, Eq. (6), the frictional forces exerted by the surrounding liquid, Eq. (7), and the bead is subject to a stochastic force $\mathcal{F}(\mathbf{T}, \mathbf{t})$ because of the Brownian motion of the surrounding liquid at temperature T and time t. $\mathcal{F}(\mathbf{T}, \mathbf{t})$ is assumed to be white noise with an average value of zero and a constant value of the power spectrum

$$|\tilde{F}(f)|^2 = 4\gamma k_B T \tag{8}$$

where k_B is the Boltzmann constant.

Since the mass of the bead is very small, and likewise the velocities, the Reynolds number describing the situation is very low, and the inertial term, $m\mathbf{a}$ (m is mass $\mathbf{a} = \ddot{x}$ is acceleration) is significantly smaller than the other forces mentioned. Hence, Newton's second law gives the equation of motion:

$$m\mathbf{a} = -\kappa\mathbf{x} - \gamma\boldsymbol{v} + \mathcal{F}(\mathbf{T}, \mathbf{t}) \simeq 0 \tag{9}$$

Upon Fourier transformation of Eq. (9), one obtains

$$2\pi\gamma(f_c - \mathrm{i}f)\tilde{x}(f) = \tilde{F}(f) \tag{10}$$

where the corner frequency f_c is given by

$$f_c = \frac{\kappa}{2\pi\gamma} \tag{11}$$

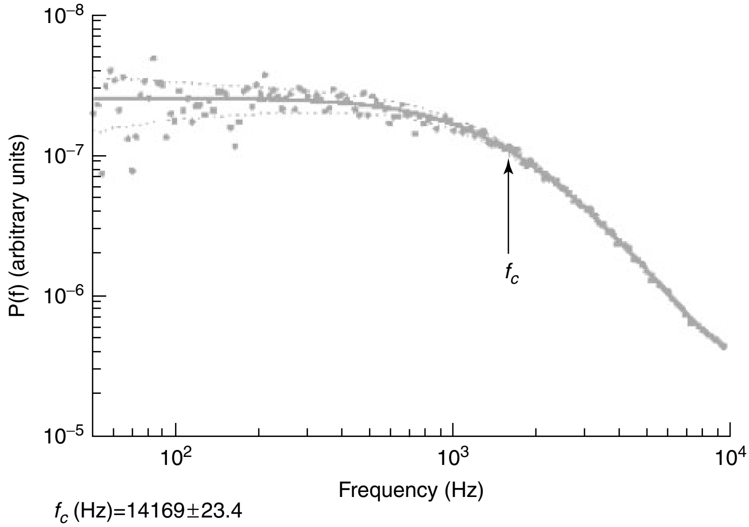

Fig. 6.5 Power spectrum. Fitting of data is performed by the procedures described in Hansen *et al.*, (2006).

Hence, the power spectrum, $S_x \equiv \langle |\tilde{x}(f)|^2 \rangle$, for the position x can be calculated as in Gittes and Schmidt (1998)

$$S_x = \frac{k_B T}{\pi^2 \gamma (f_c^2 + f^2)} \tag{12}$$

By plotting the positional power spectrum and fitting Eq. (11) to the data (as shown in Figure 6.5), one can obtain f_c for the data depicted in Figure 6.5, $f_c = 1417 \pm 23$ Hz. f_c separates two regimes from each other, namely the low-frequency regime, corresponding to large bead excursions limited in space by the trapping potential, thus resulting in a flat power spectrum, and the high-frequency regime, corresponding to small excursions, where the bead performs a simple Brownian motion inside the optical trap, this being characterized by a slop of -2 in the power spectrum. With γ known, for example, from Eq. (7), κ can be determined from Eq. (11). If x is measured and κ is known, then the force exerted by the optical trap is easily found, using Eq. (6).

In a harmonic potential, the equipartition theorem gives

$$\sigma^2 = \frac{k_B T}{\kappa} \tag{13}$$

With κ known from Eq. (11), σ (measured in meters), describing the standard deviation of the Gaussian position distribution as shown in Figure 6.3, can be found.

6.2.5
Proximity to Surfaces

The friction coefficient, γ is only well described by Stokes law (Eq. (7)) if surfaces are far away. However, often the geometry of a given experiment is such that the trapped object is close to a surface. In such cases, the friction coefficient is different, typically larger than that given by Stokes law. With proper detection, either using photodiodes or enhanced video microscopy, one can find the motion of a trapped particle in three dimensions. The corrections to γ depend on the direction of motion with respect to the surface. If one considers motion parallel to a surface (x or y directions with the coordinate system employed in previous sections), then the correctional factor, $k(h)$, to Stokes law given by Faxen's law conveniently formulated in Stout and Webb (1998):

$$k_{\text{parallel}}(h) = \left(1 - \frac{9}{16}\left(\frac{r}{h+r}\right) + \frac{1}{8}\left(\frac{r}{r+h}\right)^3 - \frac{45}{256}\left(\frac{r}{h+r}\right)^4 - \frac{1}{16}\left(\frac{r}{h+r}\right)^5\right)^{-1} \quad (14)$$

Here, h is the distance from the surface of the bead to the other surface, and r is the radius of the bead. Equation (14) is valid for $(h/r) > 0.02$.

The correction to diffusional motion in a direction normal to the surface is given by Brenner (1961):

$$k_{\text{normal}}(h) = \frac{4}{3}\sinh(\zeta) \times \sum_{n=1}^{\text{inf}} \frac{n(n+1)}{(2n-1)(2n+3)}\Omega(n,\zeta) \quad (15)$$

where

$$\Omega(n,\zeta) = \frac{2\sinh[(2n+1)\zeta] + (2n+1)\sinh(2\zeta)}{4\sinh^2[(n+\frac{1}{2})\zeta] - (2n+1)^2\sinh^2(\zeta)} - 1 \quad (16)$$

and

$$\zeta = \cosh^{-1}\left(1 + \frac{h}{r}\right) \quad (17)$$

The diffusion constant, D, is found as

$$D = \frac{k_B T}{3\pi\eta d k(h)} \quad (18)$$

The evolution of D in directions parallel to and normal to a surface is shown in Figure 6.6.

6.3
Equipment

To make optical tweezers on the basis of a single laser beam, in principle, one only needs laser light with a gradient in the intensity profile and a tightly focusing lens, that is, a lens with a high numerical aperture. To make an optical trap on the basis of the scattering forces from two counter propagating laser beams, the lenses directing the laser beams need not have a high numerical aperture.

For practical reasons and user friendliness, most labs find it useful to implement the tweezers in a microscope, typically an inverted microscope. Figure 6.7 shows a typical setup for optical tweezers.

The light emitted from the laser is expanded to a diameter, where it slightly overfills the back aperture of the objective. Then it passes through a variable attenuator, for example, consisting of a $\lambda/2$ plate and a polarizing beam splitter cube. The part of the laser light which is transmitted by the beam splitter cube is stirred into a microscope (the graey region on Figure 6.7) by means of a couple of mirrors. The orthogonally polarized part of the beam that is reflected by the beam splitter cube is often used as a second trap in the sample. Before entering the microscope, the beam passes through a telescope consisting of two lenses, one of these movable;

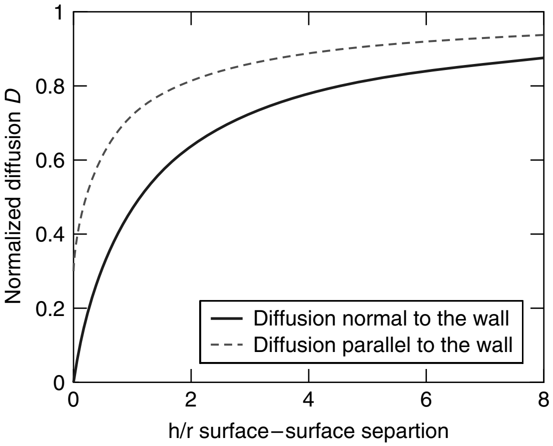

Fig. 6.6 The change of the diffusion constant, *D*, as a function of proximity to a surface. Dotted and full lines show the correction in a direction normal to and parallel to the surface, respectively. Figure is by courtesy of P.M. Bendix.

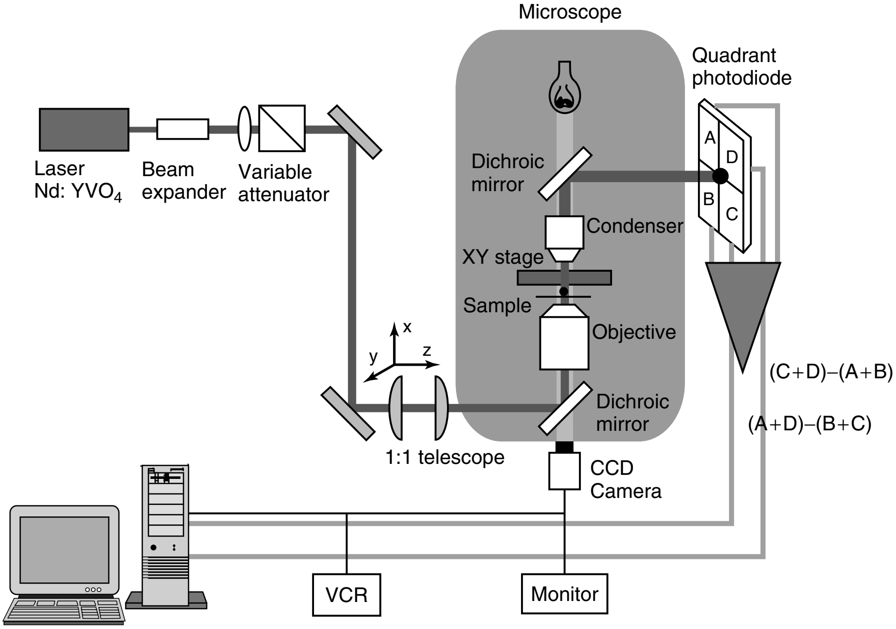

Microscope
Quadrant photodiode
A
D
B
C
Laser Nd: YVO4
Beam expander
Variable attenuator
Dichroic mirror
Condenser
XY stage
Sample
Objective
x
y
z
Dichroic mirror
1:1 telescope
CCD Camera
(C+D)–(A+B)
(A+D)–(B+C)
VCR
Monitor

Fig. 6.7 Typical optical tweezers setup.

this allows translation of the focus of the laser inside the sample. Using a dichroic mirror, light is passed on to a microscope objective, which performs a tight focusing of the laser light, hence creating the optical trap. The sample is typically mounted on a piezo translational stage.

The detection can be done in different ways, one of the two most commonly applied methods is by means of a CCD camera, the images of which are afterward analyzed, for example, by single particle-tracking routines. The other method is by using photodiodes, most often a quadrant photodiode or a position-sensitive diode. The diode can be placed such that the forward-scattered laser light collected by the condenser after having passed the sample is imaged onto the diode. Or the diode can be placed such that it uses the backward-scattered light from the sample. In a photodiode, the voltage across an active area is proportional to the number of photons hitting the diode. If the laser beam is centered with respect to a quadrant photo diode, then a trapped object will cast a shadow, create an interference pattern on the diode. The x or y positions of the trapped object is found by subtracting the signal of one half of the quadrant photodiode from the signal of the other half (as sketched in Figure 6.7). The z position of the trapped object is proportional to the total signal on the photodiode. Hence, by using a photodiode, one can obtain three-dimensional information about the positions visited by the trapped object with nanometer precision. Compared to the CCD camera, photodiodes have the advantage of a faster acquisition rate, furthermore, data analysis requires far less computational efforts than image processing.

The recorded signal, stemming from a photo diode, gives the positions x, y, and z of the bead in volts. However, it is desirable to know the position of the trapped object in spatial units. Therefore, it is necessary to find the conversion factor A, translating the recorded signal in volts, to a position in meters. From a histogram of the positions measured in volts, as shown in Figure 6.3, the variance σ_V^2 (subscript V to indicate measured in volts) can be determined

$$\sigma_V^2 = \langle x_V^2 \rangle - \langle x_V \rangle^2 \tag{19}$$

From the powerspectral analysis as outlined in Section 6.2.4.2, one can find σ measured in meters using Eq. (13). By comparing the value for σ measured in volts to the value measured in meters, the desired conversion factor can be found:

$$A = \frac{\sigma}{\sigma_V} \tag{20}$$

For the dataset depicted in Figures 6.3 and 6.5, $A = 1.5 \times 10^{-7}\,\mathrm{V\,m^{-1}}$. Knowledge of the conversion factor and κ makes the calibration procedure complete, and the forces exerted by the optical trap and the distances moved can easily be found.

6.4
Concerns

6.4.1
Biological Damage

Already, in the first trapping experiments using an argon ion laser with a wavelength of 514.5 nm, Ashkin had noticed a significant damage of the optically trapped latex particles (Ashkin *et al.*, 1986). The particles shrank in size until they finally disappeared. Possible damage on the trapped specimen induced by the laser light is a very appropriate concern to address.

The optical damage, the so-called opticution, depends on a number of factors, for example the wavelength of the trapping laser, the intensity of the laser light, and the presence of possible toxic radicals in the sample. In general, the longer the wavelength, the less is absorbed by biological specimen. In contrast, the longer the wavelength, the more pronounced the absorption of water, resulting in a heating of the sample. There seems to be a window of laser wavelengths around 800–1100 nm where the absorption by both biological specimen and water is so small that for many practical purposes it can be ignored. A wavelength of 1064 nm is very commonly used for optical trapping, because there exist YAG lasers with very high-quality Gaussian intensity profiles emitting at this wavelength.

The common notion that optical trapping essentially is not destructive, provided there is a correct choice of laser wavelength, first came about when A. Ashkin, in 1987, showed trapping of living microorganisms, which were even able to proliferate during the trapping (Ashkin, Dziedzic and Yamane, 1987). Since then, optical tweezers have been very extensively used to probe biological specimen at the nano level. Often, a handle in the form of, for example, a polystyrene bead is specifically attached to the system of interest. The optical trap then acts on the handle and hence, not on the biological system itself. The size of the trap is diffraction limited and hence does not exceed 1 μm^3. Therefore, it is fair to assume that the biological specimen is not significantly perturbed by the presence of the trap. However, the attachment of the handle might have an effect.

Recent experiments challenge the notion that optical trapping does not have a physiological effect, especially if the biological specimen is directly trapped. A study of how *Escherichia coli* bacteria were able to efficiently rotate their flagella while being optically trapped (Neuman *et al.*, 1999) showed that the tweezing indeed did have an effect. The higher the intensity of the trapping laser light, the lower was the rotational frequency. Also, they saw a pronounced effect with varying wavelengths, with 870 and 930 nm causing the most damage, and 830 and 970 nm, the least. They also proposed that the *E. coli* bacteria are increasingly sensitive in the presence of oxygen radicals, possibly produced by the trapping laser.

Other studies have shown that gene expression in *Caenorhabditis elegans* held in an optical trap could be induced more by a certain wavelength and less by other wavelengths. The stress response increased with laser power and irradiation time (Leitz *et al.*, 2002). Also, it has been shown that various bacterial types, for example *E. coli*, *Salmonella*, and *Bacilus subtilus*, are not capable of maintaining the physiologically necessary pH gradient across the outer cell wall while they are being optically trapped, this effect showing up even at very low laser intensities (Rasmussen *et al.*, 2008).

Hence, it seems a fact that living organisms do experience physiological stress upon optical trapping regardless of intensity and wavelength used. And, for each application, one has to judge whether this is considerable.

6.4.2
Heating of Specimen

Another issue concerning optical trapping is the possible heating of the specimen. The degree of heating is related to the amount of energy absorbed and the heat conductive properties of the surrounding solution. Peterman, Gittes and Schmitt (2003) have addressed this issue and found that under normal trapping conditions for example trapping a polystyrene particle in an aqueous solution with a 1064-nm laser and an intensity of 20 mW, the heating is significantly less than 1 °C. However, if trapping is performed, for example, in glycerol or other poorly conductive media, or if higher laser powers are used, one should be aware of a possible significant heating.

6.5
Technical Improvements

While operating in the nanoworld, it is always a challenge to minimize noise. Optical tweezer setups typically suffer from mechanical drift in various parts of the equipment. Therefore, efforts have been put into making very stable setups, which include minimizing air currents, mechanical drift of stages, and so on. One way of eliminating the effect of drift is to have two traps in the sample originating from the same laser beam and to use one of these traps to monitor the drift of a reference point, a fiducial mark; this drift can then be subtracted from the actual measurements thus significantly increasing the spatial resolution down to fractions of a nanometer (Nugent-Glandorf and Perkins, 2004).

Another very useful improvement is the so-called force clamp, which is a feed-back mechanism ensuring a constant force acting by the optical trap despite possible motion of the trapped specimen.

6.5.1
Multiple Optical Traps

In principle, the only necessary condition to trap a particle with an index of refraction different from the surroundings is to have an appropriate light intensity gradient. Hence, different trapping schemes with various intensity profiles can be imagined. One useful technique is the so-called multiple-beam optical tweezers, which are capable of trapping and manipulating up to hundreds of particles simultaneously in three dimensions. Multiple-beam optical tweezers can be constructed in fundamentally different ways. One of these is by means of a diffraction grating, a hologram, which is placed in the beam path typically before the telescope. When the beam hits the hologram it is split into a number of less intensive beams, each of which acts as an optical trap. This creates "holographic optical tweezers" (Curtis, Koss and Grier, 2002).

Other ways to construct multiple-beam optical tweezers include using a spatial light modulator (SLM) to impose a prescribed amount of phase shift at each pixel in an array by varying the local optical path length. Typically, this is done by controlling the orientation of molecules in a layer of liquid crystals. The generalized phase contrast technique is an elegant variation of this, which, by a simple computer interphase can be used to create arbitrary trapping schemes (Rodrigo, Daria and Gluckstad, 2004).

Such arrays of optical traps hold great promise, for example, for making 3D colloidal structures, for microassembly, sorting (MacDonald, Spalding and Dholakia, 2003), and manipulation. One drawback, however, is that, at present, no calibration methods are available for

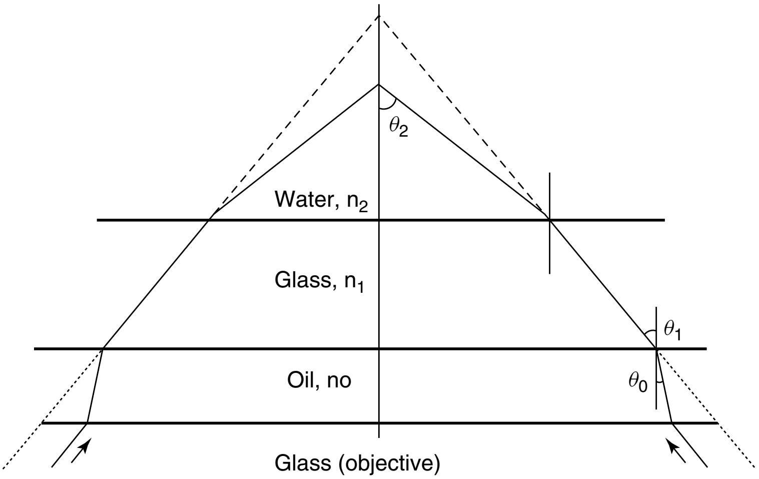

Fig. 6.8 Schematic drawing of marginal rays. At all interphases between media with different index of refraction, the rays are bend, hence giving rise to a defocusing of the optical trap.

multiple traps, hence, the forces cannot be either controlled or measured. Also, if many traps are made from the same laser beam, each of them is typically weak, sometimes only capable of two-dimensional trapping.

6.5.2
Compensating Spherical Aberrations

Because of possible opticution and heating of the sample, it is important to have as efficient an optical trap as possible, that is to use as little power as possible to create a trap of a given strength. When the laser beam passes from a media with one refractive index to a media with a different refractive index, for example, from the immersion media to the glass coverslip, and again from the glass coverslip to the sample environment, it undergoes refraction as sketched in Figure 6.8. The refractive angle depends on the distance to the center of the beam. Therefore, this refraction defocuses the laser light in the specimen, thus giving rise to a concept known as *spherical aberration*, which weakens the optical trap. Therefore, to have an efficient trap, it is desirable to avoid or minimize spherical aberrations.

Oil immersion objectives can have higher numerical aperture than water immersion objectives. Hence, in principle, they hold potential for producing stronger traps. However, because of the presence of immersion oil in the optical path, oil immersion objectives produce optical traps that suffer significantly more from spherical aberrations than traps produced by water immersion objectives. Hence, microscope manufacturers optimize the objectives to work with visible light so as to have as few distortions as possible for the corresponding wavelengths. However, most optical tweezers used to manipulate biological specimen are based on infrared lasers, which suffer significantly from spherical aberrations if focused by typical high numerical objectives.

One solution to the problem of induced spherical aberrations is to use another source to cancel the spherical aberrations induced along the beam path. An easy and very efficient method is simply to change immersion media to an oil with a higher refractive index (Reihani and Oddershede, 2007). This method increases the efficiency of a typical optical trap by a factor of 3. Another way of compensating spherical aberrations is by changing the length of the focusing tube (Reihani, Khalesifard and Golestanian, 2006). By combining these two methods, optical trapping of hundreds of microns can be performed very efficiently inside an aqueous sample, even by using an oil immersion objective.

6.6
What Can Be Trapped?

In the 1980s, Ashkin demonstrated single beam three-dimensional optical trapping of dielectric beads ranging in size from 26-nm silica beads to latex or polystyrene beads of several microns. Also, he demonstrated trapping of living bacteria, yeast, and virus. Optical trapping of metallic nanoparticles was first done by Svoboda and Block (1994), who trapped 36.2-nm gold nanoparticles. The forces acting on such small metallic particles are significantly larger than on, for example, latex particles of similar size. Hence, metallic nanoparticles hold great potential as handles for nanoscientific applications. Later, 3D optical trapping of individual gold nanoparticles with diameters ranging from 18 to 254 nm was achieved (Hansen, Bhatia and Oddershede, 2005), and even quantum dots could be optically manipulated.

Basically, all living cells can be individually trapped and manipulated, and even organelles inside living cells can be trapped (without perturbing the cell wall).

In principle, any object with dielectric properties that are different from the surroundings can be trapped with a proper detection scheme, provided that the trapping potential is deeper than the typical thermal energy of the particle. In Hansen, Bhatia and Oddershede (2005), the trapping potentials for gold nanoparticles have been calculated by integrating Eq. (4). For a metallic particle with diameter 20 nm the potential depth is $\sim 20\ k_BT$, whereas for a 12-nm particle it is $\sim 2\ k_BT$ (assuming a laser power of 800 mW). Hence, the limit of optical trapping of metallic nanoparticles probably is around 10–12 nm.

6.7
Applications

Optical tweezers are very versatile tools. This, along with the fact that they can be made nearly noninvasive, makes them perfect to study biological organisms and molecules at the micro and nano levels. One of the largest accomplishments mediated by optical tweezers is the unraveling of the functioning of individual biological molecules. Optical tweezers allow for the study of an individual molecule, thus with the potential of revealing spatial and temporal inhomogeneities that are masked in ensemble measurements. As a ‘‘macroscopic’’ example, let us imagine we are looking at a crowd of people from a distance. Let us suppose that half of the persons are blue, half are yellow, and that they are mixed together. From our point of view, the crowd appears green, but it would an erroneous conclusion to draw that the individuals are green. To really understand the nature of the crowd, the individuals need to be investigated.

6.7.1
Molecular Motors

A molecular motor is a single biological molecule that performs a biased motion and burns some fuel (typically in the form of triphosphates). Significant work has been performed using optical tweezers to pinpoint working mechanisms, such as typical speeds, stalling forces, and step sizes of molecular motors. In these experiments, a bead is typically and specifically attached to the molecule of interest, the bead being used as a handle for the optical techniques. In the following, a couple of the first molecular motors uncovered are described.

6.7.1.1 Kinesin

Kinesin is a two-headed molecule that moves along microtubules, thus mediating intracellular transport. It is a highly progressive molecule, typically taking hundreds of steps along the microtubules before falling off. (Schnitzer and Block, 1997) have done a careful investigation of this molecular motor. Using an attached silica bead as a handle for the optical trap, they studied the motion of the molecule. Without an optical trap present, the attached bead would perform a Brownian diffusion with an amplitude significantly larger than the step size of the motor. By holding the bead within the optical trap, the Brownian fluctuations are largely suppressed, hence, a step size of only 8 nm was resolved. Also, it was proved that a single ATP is hydrolyzed per step. Several mechanisms for kinesin movement can be envisioned; amongst the published suggestions are the hand-over-hand or the inch worm models. At present, the most favored model is a slightly asymmetric hand-over-hand mechanism (Asbury, Fehr and Block, 2003).

6.7.1.2 RNA Polymerase

RNA polymerase moves along DNA while performing transcription, that is copying the information stored in DNA into an RNA copy. The polymerase is also highly processive as it moves from the promotor site to the stop signal. By attaching the polymerase to a surface and attaching a bead, held by an optical trap to one end of a DNA string, the action of the RNA polymerase was investigated (Wang *et al.*, 1998). The polymerase was found to move with an average velocity of 16 nucleotides/s, corresponding to $\sim 5\,\mathrm{nm\,s^{-1}}$. By increasing the force applied with the optical trap on the motor the stalling force, causing the RNA polymerase to arrest, was found to be 30–35 pN.

6.7.1.3 Viral Rotary Motor

Viruses are tiny organisms, on the order of 50 nm in diameter, containing their heritage material very densely packed in a capsid. Upon infection, the virus squirts its heritage material through the cell wall of the host organism. The viral bacteriophage $\phi 29$ packages its 6.6-µm long DNA into its nanocapsid. This process is done by a rotary motor, the efficiency of which is remarkable because DNA must be packed to nearly crystalline density. The motor was studied by optical tweezers by Bustamante and coworkers (Smith *et al.*, 2001) where they investigated the action of the motor against an applied force. The motor is very strong, capable of working against a load of 57 pN, and it has an initial packaging rate of ~100 basepairs/s, corresponding to ~33 nm s^{-1} of DNA. Also, the motor is highly processive, working over lengths of 5 μm.

6.7.2
Unfolding/Disassembly of Macromolecules

Optical tweezers can exert forces up to some hundreds of pico-Newtons. They operate in a force regime lower than the atomic force microscope, which typically operates in the nano-Newton range. Macromolecules can be unfolded by any nanotool capable of applying forces in the correct regime, and whether tweezers or an atomic force microscope is preferable depends on the structure investigated. However, a problem common to any technique applied to unfold macromolecules is the fact that the force needed to unfold a given structure, and the time it takes before the structure unfolds, is highly dependent on the loading rate with which the force is applied. In principle, if one waits for a long enough time, any structure will unfold even in the absence of an external load. In the presence of an external load, the free energy as a function of the reaction coordinate is tilted (Wang *et al.*, 1998), and if the energy barrier is lowered, then the reaction will take place significantly faster. The interpretation of such nonequilibrium unfolding curves is not trivial, but can be done by the procedures proposed by Jarzynski (1997) or Crooks (1999).

6.7.2.1 Release of DNA from Nucleosomes

Nucleosomes organize eucaryotic DNA by wrapping it around the nucleosomes. The DNA on the nucleosomes is in a condensed state, also permitting regulated access to genetic information contained therein. Using optical tweezers, the dynamic structure of individual nucleosomal arrays was investigated by Wang and coworkers (Bower-Toland *et al.*, 2002). They find a reversible, multistage release of DNA with 76 basepairs being released at relatively low stretching force, at higher force 80 basepairs were released, and finally the histones were completely detached. These experimental observations form the basis of a ‘‘three-stage’’ model of nucleosomal opening.

6.7.2.2 RNA Structures

Optical tweezers have also been used to unfold secondary and tertiary mRNA structures. One example is the unfolding of RNA hairpins (Liphardt *et al.*, 2001). The RNA hairpin setup was also used as an example of how to apply the methods suggested by Jarzynski (1997) and Crooks (1999) for a correct treatment of an unfolding process through a nonequilibrium pathway.

mRNA pseudoknots are tertiary structures, which are often used by viral pathogens to induce frameshifting. Frameshifting is a mechanism by which the translating ribosome slips backward by one nucleotide and hence experiences a reading frame different from the original one. Optical tweezers were used to mechanically unfold structurally nearly identical mRNA pseudoknots, but which give rise to significantly different degrees of frameshifting (Hansen *et al.*, 2007). The conclusion was that the degree of frameshifting was correlated to the mechanical strength of the pseudoknot, thus indicating a physical mechanism for frameshifting.

6.7.3
Cell Biophysics

As optical tweezers are the only nanotool presently available that is capable of manipulating and measuring forces inside living cells in a nearly noninvasive fashion, they hold great potential use in cell biophysics. For instance, optical tweezers have found use in probing microrheological environments both of *in vitro* actin-based networks (Mizuno *et al.*, 2007) and inside living organisms (Tolic-Nørrelykke *et al.*, 2004). In comparison with the *in vitro* single-molecule experiments, probing a living organism has the advantage that the physiological conditions are biologically correct and relevant. However, interpretation of the results can be difficult because of the presence of a number of uncontrollable factors.

6.8
Future Directions

In the future, the limits of optical trapping will be pushed toward finer resolution and faster detection. Probably, efforts will be undertaken also to produce stronger traps, and to implement traps, for example, into lab-on-a-chip systems.

Combinations of optical tweezers with other techniques such as fluorescence, or even single-molecule fluorescence are beginning to appear (Lang *et al.*, 2004). Such combinations hold huge potential, for example, using the optical tweezers to apply a force to a particular structure, say, a molecule and to observe how the intrinsic components of the structure react to the applied force. Such experiments could shed light on the question of the sequence in which a protein folds and unfolds, or, for instance, how the use of energy is coupled to mechanical action of a molecular motor.

References

Asbury, C.L., Fehr, A.N., and Block, S.M. (2003) *Science*, **302**, 2130–34.

Ashkin, A. (1970) Acceleration and trapping of particles by radiation pressure. *Phys. Rev. Lett.*, **24**, 156–59.

Ashkin, A., Dziedzic, J.M., Bjorkholm, J.D., and Chu, S. (1986) Observation of a single-beam gradient force optical trap for dielectric particles. *Opt. Lett.*, **11**, 288–90.

Ashkin, A., Dziedzic, J.M., and Yamane, T. (1987) Optical trapping and manipulation of single cells using infrared laser beam. *Nature*, **330**, 769.

Bower-Toland, B.D., Smith, C.L., Yeh, R.C., Lis, J.T., Peterson, C.L., and Wang, M.D. (2002) *Proc. Natl. Acad. Sci.*, **99**, 1960–65.

Brenner, H. (1961) *Chem. Eng. Sci.*, **16**, 242.

Crooks, G.E. (1999) *Phys. Rev. E*, **60**, 2721–26.

Curtis, J.E., Koss, B.A., and Grier, D.G. (2002) Dynamic holographic optical tweezers. *Opt. Commun.*, **207**, 169–75.

Gittes, F. and Schmidt, C.F. (1998) *Methods Cell Biol.*, **55**, 129–56.

Hansen, P.M., Bhatia, V.K., and Oddershede, L. (2005) *Nano Lett.*, **5**, 1937–42.

Hansen, P.M., Tolic-Nørrelykke, I.M., Flyvberg, H., and Berg-Sørensen, K. (2006) *Comput. Phys. Commun.*, **17**, 518–20.

Hansen, T.M., Reihani, S.N.S., Oddershede, L.B., and Sørensen, M.A. (2007) *Proc. Natl. Acad. Sci.*, **104**, 5830–35.

Jackson, J.D. (1975) *Classical Electrodynamics*, Wiley, New York.

Jarzynski, C. (1997) *Phys. Rev. Lett.*, **78**, 2690–93.

Lang, M.J., Fordyce, P.M., Engh, A.M., Neuman, K.C., and Block, S.M. (2004) *Nat. Methods*, **1**, 1–7.

Leitz, G., Fallman, E., Tuck, S., and Axner, O. (2002) *Biophys. J.*, **82**, 2224–31.

Liphardt, J., Onoa, B., Smith, S.B., Tinoco, I. Jr., and Bustamante, C. (2001) *Science*, **292**, 733–37.

MacDonald, M.P., Spalding, G.C. and Dholakia, K. (2003) *Nature*, **426**, 421.

Mizuno, D., Tardin, C., Schmidt, C.F., and MacKintosh, F.C. (2007) *Science*, **315**, 370–73.

Neuman, K.C., Chadd, E.H., Liou, G.F., Bergman, K., and Block, S.M. (1999) *Biophys. J.*, **77**, 2856–63.

Nugent-Glandorf, L. and Perkins, T.T. (2004) *Opt. Lett.*, **29**, 2611–13.

Peterman, E.J.G., Gittes, F., and Schmidt, C.F. (2003) *Biophys. J.*, **84**, 1308–16.

Rasmussen, M.B., Oddershede, L.B., and Siegumfeldt, H. (2008) *Appl. Environ. Microbiol.*, **74**, 2441–46.

Reihani, S.N.S., Khalesifard, H.R., and Golestanian, R. (2006) *Opt. Commun.*, **259**, 204.

Reihani, S.N.S. and Oddershede, L.B. (2007) *Opt. Lett.*, **32**, 1998–2000.

Rodrigo, P.J., Daria, V.R., and Gluckstad, J. (2004) *Opt. Express*, **12**, 1417–25.

Rohrbach, A. (2005) *Phys. Rev. Lett.*, **95**, 168102.

Schnitzer, M.J. and Block, S.M. (1997) *Nature*, **388**, 386–90.

Smith, D.E., Tans, S.J., Smith, S.B., Grimes, S., Anderson, D.W., and Bustamante, C. (2001) *Nature*, **413**, 748–52.

Stout, A.L. and Webb, W.W. (1998) *Opt. Force Microsc.*, **55**, 99.

Svoboda, K. and Block, S.M. (1994) *Opt. Lett.*, **19**, 930–32.

Tolic-Nørrelykke, I.M., Munteanu, E.-L., Thon, G., Oddershede, L., and Berg-Sørensen, K. (2004) *Phys. Rev. Lett.*, **93**, 078102-1–078102-4.

Wang, M.D., Schnitzer, M.J., Yin, H., Landick, R., Gelles, J., and Block, S.M. (1998) *Science*, **282**, 902–7.

Further Reading

Grier, D.G. (2003) *Nature*, **424**, 810–16.

Methods in Cell Biology: Laser Tweezers in Cell Biology, Vol. 55, Academic Press (1998).

Svoboda, K. and Block, S.M. (1994) *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 247–85.

Mofitt, J.R., Chemla, Y.R., Smith, S.B., and Bustamante, C. (2008) *Recent Adv. Opt. Tweezers Ann. Rev. Biochem.*, **77**, 205–228.

7

Protein Dynamics – Concepts and Experimental Tools

Hans Frauenfelder

7.1 Introduction 217
7.2 Supercooled Liquids and Glasses 217
7.2.1 Glass Structure 217
7.2.2 The Energy Landscape 218
7.2.3 Viscosity, the Glass Transition, and Fragility 218
7.2.4 Fluctuation and Relaxation Processes 218
7.2.5 Primary (α) and Secondary (β) Relaxations 219
7.3 The Energy Landscape of Proteins 220
7.4 Vibrations and Normal Modes 222
7.5 Proteinquakes 222
7.6 The α Fluctuations and α Slaving 223
7.7 Internal Motions, Hydration, and β Slaving 224
7.8 Experimental 225
7.8.1 Single-Molecule Experiments 226
7.8.2 Slaving Experiments 226
Acknowledgments 226
References 226

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

7.1 Introduction

If proteins were in a unique conformation, without a hydration shell and a bulk solvent, they could not function. Proteins, however, are not in a unique conformation. They can assume a very large number of different conformational substates (CSs). At ambient temperature, they fluctuate continuously among the CSs. Working proteins consist of the folded amino acid chain, surrounded by a hydration shell, and embedded in a bulk solvent. All are involved in protein dynamics. The different CSs are organized in an energy landscape. A fundamental understanding of proteins and their functions must involve knowledge of the energy landscape, transitions among the CSs, and the influence of the hydration shell and the bulk solvent on the dynamics and the functions.

7.2 Supercooled Liquids and Glasses

Why discuss supercooled liquids and glasses? As will be shown below, the prominent motions inside proteins are slaved to the hydration shell and the bulk solvent. The hydration shell and the bulk solvent, however, are supercooled liquids or glasses, depending on the temperature. Information about their properties can therefore be extracted from the vast literature on such systems (Donth, 2001; Zalle, 1983; Brawer, 1985; Ediger, Angell and Nagel, 1996; Lubchenko and Wolynes, 2006). If a liquid is cooled, it can solidify discontinuously to a crystal. Some liquids resist forming a crystal even well below their freezing temperature; they become supercooled. Despite the technical importance of glasses and the challenging physics problems that they pose, standard condensed matter texts devote only few pages to them. The biological physicist has to resort to the few available books, reviews, and original papers to get the insight into the properties of amorphous systems that help understanding protein dynamics.

7.2.1 Glass Structure

A fundamental difference between a glass and a crystal is in the atomic-scale structure. Crystals are periodic and possess long-range order. Glasses are aperiodic and have no long-range order. In a crystal, every atom knows

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

where its place is. Glasses are frustrated (Toulouse, 1977; Fischer and Hertz, 1991; Lubchenko, 2008); atoms compete for the same position and a given atom may have two or more quasi-equilibrium positions. Crystals are in equilibrium at all temperatures; glasses are metastable.

7.2.2
The Energy Landscape

Supercooled liquids and glasses can be described by an energy landscape (EL) (Goldstein, 1969) that is organized in a number of tiers. The top tier contains glasses with different overall structures, called *polymorphs* (Poole *et al.*, 1997). Each polymorph can exist in a vast number of structures that are locally different. The organization continues to substates with smaller barriers. There are, however, substantial differences between the glass and the protein landscapes: glasses have some deep and very narrow wells, representing crystals. Proteins cannot form crystals. What is called a protein crystal is a crystal in which the individual components are proteins. Each individual protein can assume a large number of different structures.

7.2.3
Viscosity, the Glass Transition, and Fragility

The viscosity, $\eta(T)$, of a supercooled liquid increases very rapidly with decreasing temperature. The glass transition temperature, T_g, is often defined as the temperature where $\eta(T_g) = 10^{13}$ P. (Viscosity is measured in poise (P) or pascal second (Pa s), where 1 P = 0.1 Pa s.) The temperature dependence of the viscosity can be approximated by the Vogel–Fulcher–Tammann (VFT) expression:

$$\eta(T) = \eta_0 \exp\{DT_0/(T - T_0)\} \qquad (1)$$

Here, D is called the *fragility*. T_0 is the temperature where the viscosity would diverge if Eq. (1) is still valid at T_0.

7.2.4
Fluctuation and Relaxation Processes

Supercooled liquids and glasses undergo thermal fluctuations at all temperatures. If pushed out of thermal equilibrium, for instance by a temperature or pressure jump, they relax to the new equilibrium. Fluctuations and relaxations are connected by the celebrated fluctuation–dissipation theorem (Kubo, 1966; Chandler, 1987). The literature covering these processes in glasses and supercooled liquids is immense, but a number of reviews and papers summarize the essential results (Toulouse, 1977; Richert and Blumen, 1994; Lunkenheimer *et al.*, 2000; Angell *et al.*, 2000; Richert, 2002; Affouard, Descamps and Ngai, 2006). Two main approaches are used to study relaxations, time-domain and frequency-domain techniques. In the time-domain technique, a rapid perturbation forces the system out of equilibrium to a new state. The relaxation to the new state is observed, for instance, optically by following the position of a spectral line. In simple systems, the relaxation usually has the form

$$\phi(t) = \phi(0)e^{-kt} \qquad (2)$$

where $\phi(t)$ describes a quantity characterizing the measurement, and k is the relaxation rate coefficient. In most complex systems, however, Eq. (2) is not adequate, but a stretched exponential,

$$\phi(t) = \phi(0)\exp\{-(kt)^{\beta}\} \quad (3)$$

with $\beta < 1$, often approximates the observed $\phi(t)$ in some systems. Sometimes a powerlaw,

$$\phi(t) = \phi(0)\{1 + \mathrm{const}(kt)^{\beta}\}^{-1} \quad (4)$$

provides a better fit. A remark on data evaluation is in order here. The standard approach of the past, still in vogue in many laboratories, is to fit $\phi(t)$ with two exponentials if one does not fit. This technique is related to the fact that kinetic data in these laboratories are traditionally plotted as $\phi(t)$ versus t. In the same laboratories data would never be plotted versus H, where H is the hydrogen ion concentration. Instead pH is used, where p denotes $-\log$. Similarly kinetic data should *always* be plotted versus $-\mathrm{p}\,t$, or $\log t$. Of course, if data are only obtained over a small time range, the form of plotting is irrelevant, but such data are not of much use. A figure of $\log\phi(t)$ versus $\log t$ helps to distinguish between two or more exponential processes and a single nonexponential process. Time-domain techniques permit measurements from below picosecond to longer than kilosecond.

The frequency-domain approach, promoted by Debye (1929), permits the study of relaxation processes over more than 18 orders of magnitude in frequency (Kremer and Schoenhals, 2003; Kaatze and Feldman, 2006), from about 10^{-6} to more than $10^{12}\ \mathrm{s}^{-1}$. In the broadband dielectric spectrometry, the sample is placed in a capacitor and exposed to a monochromatic electric field $\mathbf{E}(\omega) = \mathbf{E}_0\exp(-\mathrm{i}\,\omega t)$. The resultant polarization $\mathbf{P}(\omega)$ is measured as a function of the radial frequency ω that is connected to the "technical" frequency υ by $\omega = 2\pi\upsilon$. The radial frequency ω used in the frequency domain is identical to the rate coefficient k in the time domain. The sample properties are expressed in terms of the complex electric permittivity:

$$\begin{aligned}\varepsilon(\omega) &= \varepsilon'(\omega) - \mathrm{i}\varepsilon''(\omega)\\ &= 1 + P(\omega)/\varepsilon_0 E(\omega) \quad (5)\end{aligned}$$

Here, ε_0 is the vacuum dielectric permittivity. The real part $\varepsilon'(\omega)$ represents the component of $\mathbf{P}(\omega)$ that is in phase with $\mathbf{E}(\omega)$ while $\varepsilon''(\omega)$ has a $\pi/2$ phase shift. The frequency dependence of the complex permittivity is caused by the fact that the polarization of a material does not respond instantaneously to the applied electric field. The dielectric permittivity spectrum shows a number of dispersion phenomena. The fluctuation–dissipation theorem then implies that the same phenomena should also occur in the equilibrium fluctuations. This fact is important for protein dynamics.

7.2.5
Primary (α) and Secondary (β) Relaxations

A very large number of experiments in many different supercooled liquids and glasses demonstrate that two main classes of relaxation (fluctuations) processes occur, called α and β. A typical example (Deegan and Nagel, 1995) is shown in Figure 7.1a. The two processes are characterized by the temperature dependence of their rate coefficients $k_{\alpha}(T)$ and $k_{\beta}(T)$, obtained from the peak positions of the respective peaks in $\varepsilon''(\omega)$ (Ngai and Paluch, 2004). The α relaxation generally follows a VFT law, Eq. (1).

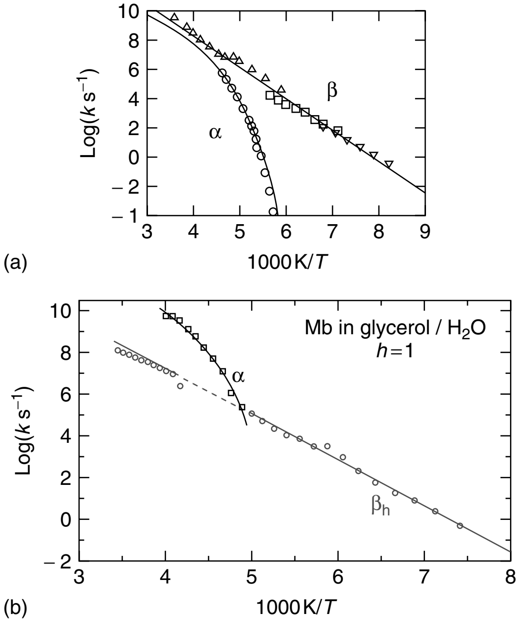

Fig. 7.1 (a) The temperature dependence of the α and β relaxations in a glass former (polybutadiene). (After Deegan and Nagel (1995).) The α relaxation can be approximated by the VFT relation, Eq. (1). The β relaxation follows the Arrhenius law. (b) The temperature dependence of the α and β_h relaxations in myoglobin embedded in a equal weight glycerol/water solvent, with hydration $h = 1$, where h is the weight ratio of water/protein.

The rate coefficient $k_\alpha(T)$ and the viscosity $\eta(T)$ are related by the mechanical Maxwell relation

$$k_\alpha(T) = G/\eta(T) \tag{6}$$

Here, G is a shear modulus that depends only weakly on temperature and on the material. At 300 K in glycerol–water mixtures G is approximately 4×10^{11} cP s^{-1}. The β relaxation usually follows an Arrhenius relation, $k_\beta(T) = A\exp\{H_\beta/k_BT\}$, and depends only weakly on viscosity. In some systems, more than one β relaxation occurs. Proteins surrounded by the hydration shell and embedded in a bulk solvent also show the α and β fluctuations (Figure 7.1b). The β relaxations in Figure 7.1a and b differ. In Figure 7.1a, the β relaxation is faster and appears to merge with the α relaxation at high temperature. In Figure 7.1b, the β relaxation crosses α and it continues to higher temperature, and is much slower than α above the crossing. It turns out that it originates in the hydration shell of the protein and we therefore denote it with a subscript "h".

7.3
The Energy Landscape of Proteins

Proteins, like glasses, can assume a large number of different CSs, distinguished by their different structures and dynamics. Evidence for the existence of CS in proteins comes from a wide variety of experiments, for instance nonexponential ligand binding (Austin *et al.*, 1975), the Debye–Waller factor in X-ray diffraction (Frauenfelder, Petsko and Tsernoglou, 1979; Garcia, Krumhansl and Frauenfelder, 1997; Chong *et al.*, 2001), and the width of spectral bands observed in various spectroscopies. If the number of different CSs were small, they could be individually described. The number is, however, extremely large and they can be characterized by introducing the EL, a construct in a high-dimensional space. At any instant of time, the structure of a protein, including its hydration shell and part of the solvent, is given by the coordinates $\mathbf{x_i} \equiv \{x_i, y_i, z_i\}$ of all N atoms in the protein, the hydration shell, and the adjacent solvent. The high-dimensional vector $\mathbf{x(t)} \equiv \{x_1, x_2, \ldots, x_N\}$ describes the structure at time *t*. To represent **x(t)**, the EL is introduced. It has about $3N$ coordinates, where N is about a thousand. The vector **x(t)** is a point in this space. These points, together with the barriers between them, form the EL. Changes of the protein structure in real space appear as the motion of **x(t)** in the hyperspace. Any attempt to draw the EL on a piece of paper always causes a headache. Nevertheless, such a figure can give some insight. Figure 7.2 is a projection of the EL on a "conformation plane". Each point in this figure describes one particular CS, given by the vector **x(t)**. The number of such points is extremely large, of the order of 10^{100}. Only a very small selection is shown. What has been gained by the introduction of the EL? Not much would be changed if the EL were featureless. However, in myoglobin (Mb) and most likely in all biomolecules the EL has energy valleys within energy values within energy valleys that are organized into a number of tiers. The top three tiers are relevant for function. Lower tiers do exist, which have been discovered in cryogenic experiments (Leeson *et al.*, 1997), but we do not know if they have functional importance. The top tier, tier 0, describes taxonomic substates; they are so called because they can be individually described and behave in some sense as if they are different proteins. They are the analog to polymorphs in glasses. In Mb, three taxonomic substates have been identified and two, called A_0 and A_1, have been explored in some detail. They have different functions (Frauenfelder *et al.*, 2001), catalysis and dioxygen storage, and can be described by their different spectra (Alben, 1982; Ansari, 1987) and different structures (Yang and Phillips,

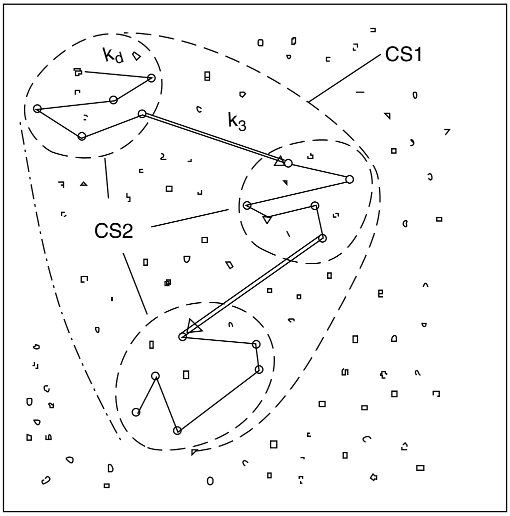

Fig. 7.2 The energy landscape. Each point represents a vector **x**(t) that describes the entire structure of the protein in a particular substate. Since 10^{100} of the order of CSs are there, only an extremely small selection is shown. A protein starting in a particular CS traces out a path in the EL that defines the substates in tiers 2 and 1.

1996). We now use the conformation plane to discuss tiers 1 and 2, and define these tiers by the average fluctuation rate coefficients at ambient temperature. Typical values are $k_\alpha \sim 10^{11}\ \mathrm{s}^{-1}$ and $k_\beta \sim 10^{8}\ \mathrm{s}^{-1}$ (Figure 7.1b). We therefore assign the α fluctuations to tier 2 and the β fluctuations to tier 1. Consider the fate of a protein starting out in a particular substate of tier 2. The protein will make a random walk, each step given by k_α. After an average of 10^3 steps, the protein uses a β fluctuation to jump to another substate of tier 1 where it will again remain approximately for the time $1/k_\beta$. These random walks trace out the CS2. A protein starting out in a different CS will also trace out CS2 and CS1. The EL is actually more complex than that that is shown in Figure 7.2. More work will be needed! (Frauenfelder, Sligar and Wolynes, 1991)

As discussed in Sections 7.6 and 7.7, both the α and β fluctuations are essential for motions in proteins. These fluctuations are actually not generated by the protein itself. The α fluctuations originate in the bulk solvent and drive these fluctuations in the protein. The β fluctuations originate in the hydration shell and generate internal protein motions. The beauty of work with proteins is that the two types can be separated. In a rigid solvent, the α fluctuations are so slow that they can be neglected. If the hydration shell is absent, there are no β fluctuations.

7.4
Vibrations and Normal Modes

In 1907, Einstein resolved a major difficulty in understanding the low-temperature behavior of the specific heat of crystals by quantizing the system, assuming that all atoms vibrate with the same frequency. Debye improved the model by considering the crystal as an elastic continuum. For proteins, these models are much too simple. Each protein atom can vibrate with its own frequency. What is of interest for dynamics are not these individual frequencies, but the normal modes (Cui and Bahar, 2006). These are time-independent constants of the motion and are described by the vibrational density-of-state spectrum, the number of normal modes per frequency interval (Hinsen, 2000; Melchers *et al.*, 1996). The normal modes are the solutions of the relevant eigenvalue equation. Modern computers can solve this equation even for proteins. Experimentally, information about the normal mode spectrum has been obtained, for instance, by using the phonon-assisted Mössbauer effect (Achterhold *et al.*, 2002) or by high-resolution X-ray measurements (Leu *et al.*, 2005). Normal mode analysis has limitations. Barriers between substates can be so high that vibrations cannot overcome them. Molecular dynamics calculations complement the normal mode approach by also investigating jumps between wells (Garcia, 1992; Kitao, Hayward and Go, 1998; Becker *et al.*, 2001).

7.5
Proteinquakes

An earthquake is the result of a sudden release of stored energy in the Earth's crust. Tectonic earthquakes caused by a sudden slip of plates result in two observables, the changes in "conformation" along the fault and seismic waves. Quakes occur on the moon, the planets, the sun, stars, and in proteins (Ansari *et al.*, 1985). Consider myoglobin in the bound-state structure. A laser flash breaks the CO–Fe bond and the protein relaxes to the deoxy structure. Part of the released energy is emitted as waves. When compared to the total binding energy of the system, proteinquakes are more powerful than earthquakes or starquakes, because the released bond energy can be of the same order of magnitude as the binding energy of the protein, about 100 kJ mol^{-1}. Proteinquakes show the two characteristic features of quakes, changes in the conformation and the emission of waves. Evidence for a structural change during a proteinquake comes from the fact that the X-ray structures of the MbCO and deoxyMb are different. Synchrotron radiation gives far more information; it permits the study of the time-dependence of the structure during the quake (Teng, Srajer and Moffat, 1994; Bourgeois *et al.*, 2006). Very fast volume changes, which are the equivalent of the waves in earthquakes, have also been observed (Miller, 1994; Nagy, Raicu and Miller, 2005). The primary features of proteinquakes are thus established. The properties of the structural changes in the quakes in Mb and also in other proteins, for instance in the photoactive yellow protein (PYP) (Xie *et al.*, 2001; Ito and Sasai, 2004), remain to be explored. The fluctuation–dissipation theorem implies that the motions seen in proteinquakes also appear in thermal fluctuations. Indeed, entrance of a ligand such as dioxygen into a protein and diffusion through the protein involve fluctuations. Experiments and computational and theoretical treatments (Onuchic and Wolynes, 1993; Dellerue *et al.*, 2001; Miyashita, Onuchic and Wolynes, 2003; Miyashita, Wolynes and Onuchic, 2005; Okazaki *et al.*, 2006) show the crucial roles of proteinquakes and the related fluctuations in protein function.

7.6
The α Fluctuations and α Slaving

The complexity of the EL implies that proteins can undergo a range of fluctuations and relaxations. Rather than studying the individual motions in many proteins as is often done in computational approaches, deeper insight comes from looking at general features. Experiments over a broad range of temperature and using Mb embedded in cryosolvents provide such information. The effect of solvent viscosity on the dynamics of myoglobin is crucial (Beece *et al.*, 1980; Doster, 1983; Sage, Shoemaker and Champion, 1995; Kleinert *et al.*, 1998; Shibata, Kurita and Kushida, 1999; Fenimore *et al.*, 2002). A problem arises in the interpretation of such experiments. Although the viscosity in the bulk solvent can be measured, the viscosity in the interior of a protein must be calculated (Ansari, 1999). It is therefore better to use the dielectric relaxation rate coefficient $k_\alpha(T)$, which is related to the viscosity by the Maxwell relation (Eq. (5)). The experiments show that large-scale motions depend strongly on viscosity (Shibata, Kurita and Kushida, 1999). Examples of such large-scale motions are the exit of ligands from myoglobin, the exchange between taxonomic substates, and fluctuations of the entire protein. The protein motions follow $k_\alpha(T)$ over many orders of magnitude in rate, but are slower by large factors. This observation leads to the concept of slaving: large-scale protein motions follow the large-scale α fluctuations in the bulk solvent, but can be slower. Large-scale protein motions are not like opening and closing a rigid gate; they take place in a large number of small steps. These motions can be understood in terms of the hierarchical EL. Assume that the initial state i represents a closed conformation, with CO inside the protein, and f the substate where a gate to the solvent is open. Assume further that each step occurs with a rate coefficient $k_\alpha(T)$ and that the protein makes $n(T)$ steps to reach the open conformation. The exit rate is then given by

$$k_{\text{exit}}(T) = k_\alpha(T)/n(T) \tag{7}$$

The fluctuations in the protein follow the fluctuations in the solvent over about eight orders of magnitude, justifying the concept of slaving: The protein follows the solvent and not the other way around (Fenimore *et al.*, 2005). Moreover, the protein follows the solvent up to the smallest solvent viscosity, about 1 cP. It is therefore the protein that adjusts to the solvent. The protein behaves more like a liquid than a solid (Doster, Cusack and Petry, 1990; Hinsen *et al.*, 2002). The internal protein viscosity is about 1 cP, or even smaller, in agreement with Ansari's calculation, which shows that the internal friction is similar to that of water (Ansari, 1999). The bulk solvent contributes the enthalpic barrier and the protein contributes the entropic barrier for the control of large-scale protein motions. Slaving is also important for protein folding (Frauenfelder *et al.*, 2006). Lubchenko and Wolynes have formulated a theory of slaving (Lubchenko, Wolynes and Frauenfelder, 2005; Lubchenko, 2006). The theory shows how the molecular motions of the glass-forming solvent distort the boundary of the protein, thereby slaving large-scale conformational motions of the protein.

7.7
Internal Motions, Hydration, and β Slaving

Most protein functions stop below 200 K and it is often claimed that below this temperature proteins undergo only harmonic motions. However, internal motions, for instance the transit of small molecules from one cavity to another or some enzymatic reactions, occur well below 200 K. The α fluctuations alone cannot be responsible for these internal motions because internal motions occur even in a solid bulk environment, which suggests different fluctuations must be responsible. Therefore, the β fluctuations, briefly introduced in Section 7.2, are candidates for these internal motions. Experiments support this suggestion: (i) the internal motions continue even if the protein is encapsulated in a solid environment, such as poly(vinyl) alcohol (Austin *et al.*, 1975); (ii) internal motions disappear if the protein is dehydrated; and (iii) the activation enthalpies of the internal ligand transit are of the order of 40 kJ mol^{-1} (Sage, Shoemaker and Champion, 1995). The activation enthalpies of the β fluctuations in thin water films on a wide range of substances cluster around 45 kJ mol^{-1} (Swenson, Jansson and Bergman, 2006). These three observations together suggest that the hydration shell (Rupley and Careri, 1991; Halle, 2004; Levy and Onuchic, 2006; Helms, 2007), about two layers of water surrounding the protein, is the source of the β fluctuations. Experiments to test this conjecture lead to a remarkable conclusion: The internal motions as seen by the heme iron in myoglobin and the internal hydrogen protons, mainly in the side chains of the backbone, follow the β fluctuations in the hydration shell. The fluctuations in the hydration shell are measured using dielectric spectroscopy, and the internal motions are observed using the Mössbauer effect (Goldanski and Herber, 1968; Parak, Knapp and Kucheida, 1982) and neutron scattering (Doster, Cusack and Petry, 1989). Comparison of the external fluctuation rates to the internal fluctuation rates shows that the fluctuations seen by the heme iron and the hydrogen atoms follow fluctuations in the hydration shell. The conclusion is surprising. Naively one would expect that different parts of the protein's interior fluctuate with different rates. The observation that the heme iron and interior protons fluctuate similar to the hydration shell suggests a different picture: major internal motions are governed by the β fluctuations in the hydration shell. In other words, internal motions are slaved to the motions in the hydration shell. A word of caution. The results discussed here are relatively new and they have mainly been obtained by using one protein, myoglobin. The experiments have to be repeated, extended, and applied to other proteins and also to nucleic acids. Moreover, the fact that major motions can be explained as being due to the α and β fluctuations does not mean that other fluctuations do not play a role (Fenimore *et al.*, 2004).

7.8
Experimental

Conformational fluctuations in biomolecules extend from below picosecond to above kilosecond, more than 15 orders of magnitude. They involve small-scale and large-scale motions and can depend on external factors such as temperature, pressure, hydration, pH, viscosity, and solvent composition. It is obvious that no single technique and no single experiment can provide all the information necessary to fully understand biomolecular dynamics. The fact evident from experiments is that data taken over a small range of variables, such as measurements at one temperature only, one viscosity only, a small range of time say 1–100 ps, or one hydration level, cannot give the insight into the dynamics that is needed to understand the connection between motions and functions of protein. Nearly every modern experimental technique provides part of the necessary information. Many of the important approaches are discussed in other chapters of this volume or in the various Annual Reviews. Connecting experimental studies to computational and theoretical work is also needed to reach a picture with predictive power. Here, we provide a potpourri of a few useful experimental techniques and give relevant references.

Two properties of proteins determine what types of experimental techniques are most useful. The first property is the inherent existence of CSs. Proteins are highly inhomogeneous and not two in a sample have exactly the same structure and the same dynamics; protein properties must be described by distributions. In many instances, such distributions can be extracted from bulk experiments. Single-molecule experiments, however, give direct information. The second property is the observation that protein motions are to a large extent slaved to the fluctuations in the hydration shell and the bulk solvent. Thus, motions in the interior, the hydration shell, and the bulk solvent should be measured separately.

7.8.1
Single-Molecule Experiments

Because a protein can assume an extremely large number of conformations, any experiment on a sample yields an average result. Single-molecule experiments provide complementary information to bulk data (Deniz, Mukhopadhyay and Lemke, 2008).

7.8.2
Slaving Experiments

The relaxation and fluctuation experiments have already been described in Section 7.3. Broadband dielectric spectrometry is an excellent tool to study the α and β fluctuations in the bulk solvent and the hydration shell over a very wide range of time and temperature. As already pointed out, the two can be separated. In a solid bulk, the α fluctuation is essentially absent. The β fluctuation is absent in a dehydrated protein. The internal fluctuations can be observed with many different techniques. The Mössbauer effect and neutron scattering (Bée, 2003) have already been mentioned. Other techniques such as nuclear magnetic resonance (NMR) (Palmer, 2001; Mittermaier and Kay, 2006; Henzler-Wildman and Kern, 2007) also can yield the insight. But note that measuring only at one temperature does not provide the critical information whether two fluctuation processes with different rates are related or are independent (Lu, Xun and Xie, 1998; Michalet, Weiss and Jäger, 2006; Myong, Stevens and Ha, 2006; Joo *et al.*, 2008).

Acknowledgments

The work was supported by the US Department of Energy contract W-7405-ENG-36. I have been greatly benefited from discussions and interactions with Joel Berendzen, Guo Chen, Paul Fenimore, Ben McMahon, Jan Swenson, and Robert Young.

References

Achterhold, K., Keppler, C., Ostermann, A., van Buerck, U., Sturhan, W., alp, E.E. and Parak, F.G. (2002) Vibrational dynamics of myoglobin determined by the phonon-assisted Moessbauer effect. *Phys. Rev. E*, **65**, 061916.

Affouard, F., Descamps, M. and Ngai, K.L. (eds) (2006) *Relaxations in Complex Systems*, Elsevier, Netherlands. The papers have also been published in *J. Non-Cryst. Solids*, **352**, 4371–5227, 2006.

Alben, J.O. *et al.* (1982) Infrared spectroscopy of photodissociated carbonmonoxymyoglobin at low temperatures. *Proc. Natl. Acad. Sci. U.S.A.*, **79**, 3744–48.

Angell, C.A., Ngai, K.L., McKenna, G.B., McMillan, P.F. and Martin, S.W. (2000) Relaxation in glassforming liquids and amorphous solids. *J. Appl. Phys.*, **88**, 3113–50.

Ansari, A. (1999) Langevin modes analysis of myoglobin. *J. Chem. Phys.*, **110**, 1774–80.

Ansari, A., Berendzen, J., Bowne, S.F., Frauenfelder, H., Iben, I.E.T., Sauke, T.B., Shyamsunder, E. and Young, R.D. (1985) Protein states and protein quakes. *Proc. Natl. Acad. Sci. USA*, **82**, 5000–4.

Ansari, A. *et al.* (1987) Rebinding and relaxation in the myoglobin pocket. *Biophys. Chem.*, **26**, 337–55.

Austin, R.H., Beeson, K.W., Eisenstein, L., Frauenfelder, H. and Gunsalus, I.C. (1975) Dynamics of ligand binding to myoglobin. *Biochemistry*, **14**, 5355–73.

Becker, O.M., MacKerell, A.D. Jr., Roux, B. and Watanabe, M. (eds) (2001) *Computational Biochemistry and Biophysics*, Marcel Decker, Inc., New York.

Bée, M. (2003) Localized and long-range diffusion in condensed matter: state of the art of QENS studies and future prospects. *Chemical. Physics.*, **292**, 121–41.

Beece, D., Eisenstein, L., Frauenfelder, H., Good, D., Marden, M.C., Reinisch, L., Reynolds, A.H., Sorenson, L.B. and Yue, K.T. (1980) Solvent viscosity and protein dynamics. *Biochemistry*, **19**, 5147.

Bourgeois, D., Vallone, B., Arcovito, A., Sciara, G., Schotte, F., Anfinrud, P.A. and Brunori, M. (2006) Extended subnanosecond structural dynamics of myoglobin revealed by Laue crystallography. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 4924–29.

Brawer, S. (1985) *Relaxation in Viscous Liquids and Glasses*, American Ceramic Society, Columbus Ohio.

Chandler, D. (1987) *Introduction to Modern Statistical Mechanics*, Oxford University Press, Oxford.

Chong, S.-H., Joti, Y., Kidera, A., Go, N., Ostermann, A., Gassmann, A. and Parak, F. (2001) Dynamical transition of myoglobin in a crystal: comparative studies of X-ray crystallography and Mössbauer spectroscopy. *Eur. Biophys. J.*, **30**, 319–29.

Cui, Q. and Bahar, I. (eds) (2006) *Normal Mode Analysis. Theory and Applications to Biological and Chemical Systems*, Capman & Hall/CRC.

Debye, P. (1929) *Polar Molecules*, Dover.

Deegan, R.D. and Nagel, S.R. (1995) *Phys. Rev. B*, **52**, 5653.

Dellerue, S., Petrescu, A.-J., Smith, J.C. and Bellisent-Funel, M.-C. (2001) Radially softening diffusive motions in a globular protein. *Biophys. J.*, **81**, 1666–76.

Deniz, A.A., Mukhopadhyay, S. and Lemke, E.A. (2008) Single-molecule biophysics: at the interface of biology, physics, and chemistry. *Interface*, **5**, 15–45.

Donth, E. (2001) *The Glass Transition*, Springer, Berlin.

Doster, W. (1983) Viscosity scaling and protein dynamics. *Biophys. Chem.*, **17**, 97–103.

Doster, W., Cusack, S. and Petry, W. (1989) Dynamical transition of myoglobin revealed by inelastic neutron scattering. *Nature*, **337**, 754–56.

Doster, W., Cusack, S. and Petry, W. (1990) Dynamic instability of liquidlike motions in a globular protein observed by inelastic neutron scattering. *Phys. Rev. Lett.*, **65**, 1080–83.

Ediger, M.D., Angell, C.A. and Nagel, S.R. (1996) Supercooled liquids and glasses. *J. Phys. Chem.*, **100**, 13200–12.

Fenimore, P.W., Frauenfelder, H., McMahon, B.H. and Parak, F.G. (2002) Slaving: solvent fluctuations dominate protein dynamics and functions. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 16047–51.

Fenimore, P.W., Frauenfelder, H., McMahon, B.H. and Young, R.D. (2004) Bulk-solvent and hydration-shell fluctuations, similar to α- and β-fluctuations in glasses, control protein motions and functions. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 14408–13.

Fenimore, P.W., Frauenfelder, H., McMahon, B.H. and Young, R.D. (2005) Proteins are paradigms of stochastic complexity. *Phys. A*, **351**, 1–13.

Fischer, K.H. and Hertz, J.A. (1991) *Spin Glasses*, Cambridge University Press, Cambridge.

Frauenfelder, H., Fenimore, P.W., Chen, G. and McMahon, B.H. (2006) Protein folding is slaved to solvent motions. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 15469–72.

Frauenfelder, H., McMahon, B.H., Austin, R.H., Chu, K. and Groves, J.T. (2001) The role of structure, energy landscape, dynamics and allostery in the enzymatic function of myoglobin. *Proc. Nat. Acad. Sci. U.S.A.*, **98**, 2370–74.

Frauenfelder, H., Petsko, G.A. and Tsernoglou, D. (1979) Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. *Nature*, **280**, 558–63.

Frauenfelder, H., Sligar, S.G. and Wolynes, P.G. (1991) The energy landscape and motions of proteins. *Science*, **254**, 1598–603.

Garcia, A.E. (1992) Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.*, **68**, 2696–99.

Garcia, A.E., Krumhansl, J.A. and Frauenfelder, H. (1997) Variations on a theme by Debye and Waller: from simple crystals to proteins. *Proteins*, **29**, 153–60.

Goldanski, V.I. and Herber, R.H. (eds) (1968) *Chemical Applications of Mössbauer Spectroscopy*, Academic Press, New York.

Goldstein, M. (1969) Viscous liquids and the glass transition: a potential energy barrier picture. *J. Chem. Phys.*, **51**, 3728–39.

Halle, B. (2004) Protein hydration dynamics in solution: a critical survey. *Phil. Trans. R. Soc. Lond. B*, **359**, 1207–24.

Helms, V. (2007) Protein dynamics tightly connected to the dynamics of surrounding and internal water molecules. *ChemPhysChem*, **8**, 23–33.

Henzler-Wildman, K.A. and Kern, D. (2007) A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, **450**, 913–16.

Hinsen, K. (2000) Normal mode theory and harmonic potential approximations. *Contemp. Phys.*, **41**, 15–36.

Hinsen, K., Petrescu, A.J., Dellerue, S., Bellissent-Funel, M.C. and Kneller, G.R. (2002) Liquid-like and solid-like motions in proteins. *J. Mol. Liquids*, **98–99**, 381–98.

Ito, K. and Sasai, M. (2004) Dynamical transition and protein quake in photoactive yellow protein. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 14736–41.

Joo, C., Balci, H., Ishitsuka, Y., Buranana-chai, C. and Ha, T. (2008) Advances in single-molecule fluorescence methods for molecular biology. *Annu. Rev. Biochem.*, **77**, 51–76.

Kaatze, U. and Feldman, Y. (2006) Broadband dielectric spectrometry of liquids and biosytems. *Meas. Sci. Technol.*, **17**, R17–R35.

Kitao, A., Hayward, S. and Go, N. (1998) Energy landscape of a native protein: jumping-among-minima model. *Proteins: Struct. Funct. Genet.*, **33**, 496–517.

Kleinert, T., Doster, W., Leyser, H., Petry, W., Schwarz, V. and Settles, M. (1998) Solvent composition and viscosity effects on the kinetics of CO binding to horse myoglobin. *Biochemistry*, **37**, 717–33.

Kremer, F. and Schoenhals, A. (2003) *Broadband Dielectric Spectroscopy*, Springer, Berlin.

Kubo, R. (1966) The fluctuation–dissipation theorem. *Rep. Prog. Phys.*, **29**, 255–84.

Leeson, D.T., Wiersma, D.A., Fritsch, K. and Friedrich, J. (1997) Energy landscape of myoglobin: an optical study. *J. Phys. Chem. B.*, **101**, 6331–40.

Leu, B.M. *et al.* (2005) Vibrational dynamics of biological molecules: multi-quantum contributions. *J. Phys. Chem. Solids*, **66**, 2250–56.

Levy, Y. and Onuchic, J.N. (2006) Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.*, **35**, 389–415.

Lu, H.P., Xun, L. and Xie, X.S. (1998) Single-molecule enzymatic dynamics. *Science*, **282**, 1877–82.

Lubchenko, V. (2006) Quantitative theory of structural relaxation in supercooled liquids and folded proteins. *J. Non-Cryst. Solids*, **352**, 4400–9.

Lubchenko, V. (2008) Competing interactions create functionally through frustration. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 10635–36.

Lubchenko, V. and Wolynes, P.G. (2006) Theory of structural glasses and supercooled liquids. *Annu. Rev. Phys. Chem.*, **58**, 235–66.

Lubchenko, V., Wolynes, P.G. and Frauenfelder, H. (2005) Mosaic energy landscapes of liquids and the control of protein conformational dynamics by glass-forming solvents. *J. Phys. Chem. B*, **109**, 7488–99.

Lunkenheimer, P., Schneider, U., Brand, R. and Loidl, A. (2000) Glassy dynamics. *Contemp. Phys.*, **41**, 15–36.

Melchers, B., Knapp, E.W., Parak, F., Cordone, L., Cupane, A. and Leone, M. (1996) Structural fluctuations of myoglobin from normal-modes, Mössbauer, Raman, and absorption spectroscopy. *Biophys. J.*, **70**, 2092–99.

Michalet, X., Weiss, S. and Jäger, M. (2006) Single-molecule fluorescence studies of protein folding and conformational dynamics. *Chem. Rev.*, **106**, 1785–813.

Miller, R.J. (1994) Energetics and dynamics of deterministic protein motion. *Acc. Chem. Res.*, **27**, 145–50.

Mittermaier, A. and Kay, L.E. (2006) New tools provide new insights in NMR studies of protein dynamics. *Science*, **312**, 224–28.

Miyashita, N., Onuchic, J.N. and Wolynes, P.G. (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 12570–75.

Miyashita, O., Wolynes, P.G. and Onuchic, J.N. (2005) Simple energy landscape model for the kinetics of functional transitions in proteins. *J. Phys. Chem. B*, **109**, 1959–69.

Myong, S., Stevens, B.C. and Ha, T. (2006) Bridging conformational dynamics and function using single-molecule spectroscopy. *Structure*, **14**, 633–43.

Nagy, A.M., Raicu, V. and Miller, R.J.D. (2005) Nonlinear optical studies of heme protein dynamics; implications for proteins as hybrid states of matter. *Biochim. Biophys. Acta*, **1749**, 148–72.

Ngai, K.L. and Paluch, M. (2004) Classification of secondary relaxation in glass-formers based on dynamic properties. *J. Chem. Phys.*, **120**, 857–73.

Okazaki, K., Koga, N., Takada, S., Onuchic, J.N. and Wolynes, P.G. (2006) Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: structure-based molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 11844–49.

Onuchic, J.N. and Wolynes, P.G. (1993) Energy landscapes, glass transitions, and chemical reaction dynamics in biomolecular or solvent environment. *J. Chem. Phys.*, **93**, 2218–24.

Palmer, A.G. III (2001) NMR probes of molecular dynamics: overview and comparison with other techniques. *Annu. Rev. Biophys. Biomol. Struct.*, **30**, 129–55.

Parak, F., Knapp, E.W. and Kucheida, D. (1982) Protein dynamics – Mössbauer spectroscopy on deoxymyoglobin crystals. *J. Mol. Biol.*, **161**, 177–94.

Poole, P.H., Grande, T., Angell, C.A. and McMillan, P.F. (1997) Polymorphism in liquids and glasses. *Science* **275**, 322–23.

Richert, R. (2002) Heterogeneous dynamics in liquids: fluctuations in space and time. *J. Phys. Condens. Matter*, **14**, R703–R738.

Richert, R. and Blumen, A. (eds) (1994) *Disorder Effects on Relaxational Processes*, Springer, Berlin.

Rupley, J.A. and Careri, G. (1991) *Adv. Protein Chem.*, **41**, 37–172.

Sage, J.T., Shoemaker, K.T. and Champion, P.M. (1995) Solvent-dependent structure and dynamics in myoglobin. *J. Phys. Chem.*, **99**, 3394–405.

Shibata, Y., Kurita, A. and Kushida, T. (1999) Solvent effects on conformational dynamics of Zn-substituted myoglobin observed by time-resolved hole-burning spectroscopy. *Biochemistry*, **38**, 1789–801.

Swenson, J., Jansson, H. and Bergman, R. (2006) Relaxation processes in supercooled confined water and implications for protein dynamics. *Phys. Rev. Lett.*, **96**, 247802.

Teng, T.-Y., Srajer, V. and Moffat, K. (1994) Photolysis-induced structural changes in single crystals of carbonmonoxy myoglobin at 40 K. *Nat. Struct. Biol.*, **1**, 701–5.

Toulouse, G. (1977) *Commun. Phys.*, **2**, 115.

Xie, A., Kelemen, L., Hendriks, J., White, B.J., Hellingwerf, K.J. and Hoff, W.D. (2001) *Biochemistry*, **40**, 1510–1517.

Yang, F. and Phillips, G.N. Jr. (1996) Crystal structures of CO- deoxy- and met-myoglobins at various pH values. *J. Mol. Biol.*, **256**, 762–74.

Zalle, R. (1983) *The Physics of Amorphous Solids*, John Wiley & Sons, New York.

8

Molecular Dynamics Computations for Proteins: A Case Study in Membrane Ion Permeation

Igor Vorobyov and Toby W. Allen

8.1 Introduction 233
8.2 Molecular Mechanical Force Fields 235
8.2.1 Potential Energy Function 236
8.2.2 Intramolecular Parameters 238
8.2.3 Nonbonded Parameters 239
8.2.4 Polarizable Force Fields 242
8.2.5 Water, Ion, and Lipid Models 245
8.2.6 Coarse-Grained Force Fields 249
8.3 Molecular Dynamics Method for Protein Systems 251
8.3.1 Approximations to Handle Finite Systems 252
8.3.2 Maintaining Ensembles 256
8.3.3 Approximating Long-Ranged Forces 259
8.4 Calculation of Observables for Protein Structure and Function 262
8.4.1 Time Series and Average Properties 262
8.4.1.1 Averages and Correlations 262
8.4.1.2 Protein and Solvent Fluctuations 264
8.4.2 Free Energy Calculation 266
8.4.2.1 Potential of Mean Force Calculation 266
8.4.2.2 Free Energy Perturbation 267
8.4.2.3 Energetic Decompositions 268
8.4.2.4 Multidimensional PMFs 270
8.4.2.5 Computing Equilibrium Constants: The Case of pK_a Calculation 271
8.4.2.6 Computing Equilibrium Constants: An Example of Binding Constants 272
8.4.2.7 Computing Flux along a Free Energy Path 273
8.4.3 Advanced Sampling and Nonequilibrium Approaches 273
8.5 Case Studies 276
8.5.1 Ion Channel Permeation 276
8.5.1.1 Simulation Details 277
8.5.1.2 Free Energy Profile Calculation 277
8.5.1.3 Connection with Observables 280

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

8.5.1.4 Two-dimensional Free Energy Surfaces Reveal Mechanisms 280
8.5.2 Models to Explore Voltage-Gated Ion Channel Activation 284
8.5.2.1 Simulation Details 285
8.5.2.2 Mechanism of Arg Translocation 286
8.5.2.3 Free Energy Profiles Across the Membrane 287
8.5.2.4 pK_a Profiles Across the Membrane 288
8.5.2.5 Thermodynamic Decomposition 288
8.5.2.6 Accuracy of Force Fields for Arg–Membrane Interactions 290
8.6 Conclusion and Prospects 292
References 293

8.1
Introduction

Experiments have provided a wealth of structural and functional data for protein systems, yet atomic-level descriptions and femtosecond time resolution remain out of the reach of experimental techniques (Karplus and Petsko, 1990). With improved computational resources available today, molecular dynamics (MD) simulation has become an essential tool for investigating a wide range of biological phenomena. Recent advances in structural determination of proteins with electron microscopy, nuclear magnetic resonance (NMR), and X-ray diffraction have provided a springboard from which computer simulation can be used to obtain a deeper understanding of function. The aim of this chapter is to overview present-day MD methods, molecular mechanical (MM) force fields, and quantitative analysis, and to explore the challenges facing studies of protein structure and function today.

The first MD simulations of condensed-phase systems used highly simplified potential functions for simple systems such as hard-sphere models of liquid argon or the simple three-point water model (Alder and Wainwright, 1957; Rahman, 1964; Rahman and Stillinger, 1971). Atomistic computer simulations of proteins date back to 1977, when McCammon and Karplus reported a 10 ps simulation of the bovine pancreatic trypsin inhibitor (BPTI), containing 58 amino acids or $\sim$ 600 atoms (McCammon *et al.*, 1977). The first simulations of myoglobin in 1979 (Case and Karplus, 1979) revealed a path for O_2 access to the heme pocket that was not evident in crystal structures (Perutz and Mathews, 1966; Takano, 1977), demonstrating the utility of dynamic simulations of proteins. However, it is only in the last decade that enough computer power has been available to allow simulation of large protein–solvent systems with accurate parameterizations and today a great diversity of systems have been studied. Protein sizes range from hundreds of atoms, e.g. gramicidin (Allen *et al.*, 2003b) to large protein clusters, viral capsids or cell organelles comprising millions of atoms with simulation times spanning up to microseconds (Arkhipov *et al.*, 2006a; Arkhipov *et al.*, 2006b; Gumbart and Schulten, 2006; Freddolino *et al.*, 2006; Kitao *et al.*, 2006; Sanbonmatsu and Tung, 2007; Freddolino *et al.*, 2008; Jayachandran *et al.*, 2006). We are now at an exciting point in time for computer

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

simulation of proteins, but one where we must ensure accurate descriptions of protein interactions and develop rigorous statistical mechanical methods for computational analysis.

MD simulations can be performed using different levels of theory depending on the size of the system and the timescales of the protein motions of interest. In addition, the protein environment can be the dominant factor in determining system size, with water usually constituting the majority of the simulation system. Furthermore, the timescales of protein fluctuations can range from multi-femtosecond bond vibrations to millisecond to second global protein motions and folding. The challenge, therefore, is to use the best level of theory possible to describe the system that is tractable, given its size and timescales of interest. Arguably, the highest level of theory available for protein simulation is quantum mechanics (QM), which includes *ab initio*, density functional theory (DFT) and semiempirical methods, all of which explicitly describe electronic structure. DFT-based MD simulations, based on the Car–Parrinello formalism (Car and Parrinello, 1985), are emerging for protein modeling (Andreoni *et al.*, 2001), but are limited to hundreds of atoms and tens of picoseconds; well below the timescales of protein function. Recently, some advances have been made in the development of linear-scaling semiempirical methods, which allow for systems of thousands of atoms (Yang and Lee, 1995; Lee *et al.*, 1996; Dixon and Merz, 1997; Anikin *et al.*, 2004). However, hybrid QM/MM methods, pioneered by Warshel and Karplus (1972), have evolved as a valuable tool for studying protein structure and dynamics (Gao *et al.*, 2008). In this approach, the system is divided into a small QM region and a large MM region that encompasses all other atoms. (Warshel and Levitt, 1976; Bash *et al.*, 1987) The challenge remains to optimally separate QM and MM regions and to extend to greater system sizes and timescales (Gao *et al.*, 2008).

At the other end of the spectrum, coarse-grained (CG) methods (Levitt and Warshel, 1975), where each amino acid residue is represented by one or a few particles, have been developed to permit extension to greater sizes and timescales (Nielsen *et al.*, 2004; Tozzini, 2005; Biggin and Bond, 2008; Monticelli *et al.*, 2008) These methods can be used for multi-microsecond to millisecond simulations for millions of atoms (Nielsen *et al.*, 2004). Therefore, they can, in principle, be used to study protein folding (Tozzini, 2005), membrane protein insertion into lipid bilayers (Biggin and Bond, 2008) and simulation of large molecular complexes such as the ribosome (Trylska *et al.*, 2005) and nucleosome (Voltz *et al.*, 2008). Multiscale approaches, which combine atomistic MM and CG methodologies for large biomolecular systems, can allow for extended system sizes while maintaining atomic level information for protein function, which is currently an active area of development (Ayton *et al.*, 2007; Ayton *et al.*, 2008; Praprotnik *et al.*, 2008).

However, fully atomistic MD simulation with classical force fields form the heart of present-day protein simulation because of their realistic descriptions of protein–protein, protein–ligand, and protein–nucleic acid interactions and their ability to cover much of the relevant length and timescales for protein function. MD simulation can complement experimental observations by not only depicting molecular motions and interactions, but also by providing quantitative measures of function (Becker *et al.*, 2001; Kukol, 2008; Karplus and McCammon, 2002). Here we describe some of the challenges facing quantitative descriptions of protein function with MD simulation due to timescales and MM force field descriptions. We illustrate the successes and challenges by investigating particularly challenging applications involving simple model membrane-protein systems. Membrane proteins make up one-third of the human genome and over 70% of drug targets. They carry out vital physiological functions in the body. While there has been much experimental progress in understanding the structure and function of these proteins, a molecular-detail description has not been fully resolved. In this chapter we concentrate on exploring computational methodologies to understand the microscopic mechanisms of membrane protein function. We particularly focus on ion channel proteins that facilitate the selective transport of charged molecules across cell membranes and which are associated with many neurological and cardiovascular disorders (Ashcroft, 2000).

Ion translocation is an electrodiffusive process, in which specific channels can distinguish between different ions on the basis of valence and size (Hille, 2001). After the pioneering studies of Hodgkin and Katz (1949), Hodgkin and Huxley (1952), nearly 60 years ago, many structural and functional studies have sought to understand the mechanisms of permeation and selectivity. Recent successes in structural determination of ion channel proteins (Doyle *et al.*, 1998; Jiang *et al.*, 2002; Jiang *et al.*, 2003a; Miyazawa *et al.*, 2003; Chang *et al.*, 1998; Dutzler *et al.*, 2002) have provided a solid basis for detailed mechanistic studies. Ion channels also profit from decades of studies examining sensitivity to physiological and pharmacological conditions (Seeman, 1997; Hille, 2001). Yet, to fully understand the mechanisms of ion permeation and selectivity, one needs to understand the atomic breakdown of ion channel energetics. We also report studies that explore the ability of charged proteins to move within membranes to understand the voltage gating of potassium channels. Both of these studies push the limits of classical force fields, beyond what was originally envisaged when the force fields were developed, and require extensive thermodynamic analysis to extract the mechanisms of protein function.

8.2
Molecular Mechanical Force Fields

An empirical force field is a combination of a potential energy function, $U(\mathbf{R})$, describing the molecular system as a function of coordinates of the constituent atoms, $\mathbf{R}$, along with a set of empirically derived parameters. Simple empirical energy functions for computer simulation have been used for more than three decades for simple organic molecules (Burkert and Allinger, 1982) and biological systems (McCammon *et al.*, 1977). Examples of empirical force fields commonly used for protein simulations include Amber (Cornell *et al.*, 1995), CHARMM (MacKerell, 1998b), GROMOS (Oostenbrink *et al.*, 2004), and OPLS-AA (Jorgensen *et al.*, 1996), as well as other more general force fields for organic and biological molecules (Rappe *et al.*, 1992; Allinger *et al.*, 1996; Halgren, 1996; Oldziej, 2005; Daggett, 2006). There is also a broad category of CG force fields, in which a group of atoms or molecules is treated as single particles (Nielsen *et al.*, 2004). Here the main focus is on the atomistic force fields for which, arguably, the most realistic models of protein systems exist today. The popular software packages, which can be used for protein MD simulations include AMBER (www.ambermd.org) (Weiner and Kollman, 1981; Pearlman *et al.*, 1995; Case *et al.*, 2005), CHARMM (www.charmm.org) (Brooks *et al.*, 1983; MacKerell *et al.*, 1998a), GROMOS (www.igc.ethz.ch/GROMOS) (Scott *et al.*, 1999; Christen *et al.*, 2005), GROMACS (www.gromacs.org) (Berendsen *et al.* 1995; Lindahl *et al.*, 2001; Van der Spoel *et al.*, 2005), LAMMPS (lammps.sandia.gov) (Plimpton, 1995), NAMD (www.ks.uiuc.edu/Research/namd) (Nelson *et al.*, 1996; Kale *et al.*, 1999; Phillips *et al.*, 2005), and TINKER (dasher.wustl.edu/tinker).

8.2.1
Potential Energy Function

The potential energy function used in MD simulation includes several terms describing both internal (bonded) and external (nonbonded) interactions, which ultimately determine both structure and dynamics of a molecular system. The bonded terms are associated with the interactions between covalently bonded atoms separated by one, two, or three chemical bonds, whereas the nonbonded terms are associated with through-space electrostatic and van der Waals (vdW) interactions. The potential energy function used in the all-atom CHARMM empirical protein force field (MacKerell, 1998b), for example, is

$$
\begin{aligned}
U(\mathbf{R}) &= \sum_{\text{bonds}} K_b(b-b_0)^2 \\
&+ \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 \\
&+ \sum_{\text{dihedrals}} \sum_{n} K_\chi[1+\cos(n\chi-\delta)] \\
&+ \sum_{\substack{\text{nonbonded}\\ \text{pairs} i,j}} \left\{ \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right. \\
&\left. + \varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]\right\} \qquad (1)
\end{aligned}
$$

where we describe each term below. Once the potential energy of the system is calculated using Eq. (1), the forces on all atoms can be obtained from the first derivate with respect to atomic positions $\{R_i\}$ from which trajectories of atomic positions can be generated.

The bonded energy terms (first three terms in Eq. (1)) include harmonic terms associated with changes in bond length, b, (stretching) and bond angle, θ, (bending) with respect to their reference values (b_0 and θ_0) with stiffness governed by force constants (K_b and K_θ), as well as a series of cosine terms for rotation around a bond associated with the barrier height (K_χ), multiplicity (n), and phase angle δ to describe changes in the dihedral angle χ. Nonbonded interactions (last terms of Eq. (1)) are described by a simple Coulomb potential for electrostatic interactions between atomic partial charges q_i and q_j at a distance r_{ij}, with permittivity of free space, ε_0, employed due to the explicit nature of the simulated dielectric medium, as well as a Lennard-Jones (LJ) 6–12 potential for vdW interactions between atomic sites. In the representation used here, ε_{ij} is the depth of the vdW energy minimum and σ_{ij} is the position of that minimum (as opposed to the position of the zero of the potential energy used in other representations). The parameters in Eq. (1) are fitted to reproduce experimental and/or quantum mechanical reference data for a number of model systems (MacKerell, 2004). The quality of these parameters and associated target data are key factors that determine the accuracy of the empirical force field.

Parameterization is typically an iterative multistep procedure, which requires substantial computational resources (Foloppe and MacKerell, 2000; MacKerell, 2004). Because it is nearly impossible to perform parameterization for all atoms in each protein system of interest, the concept of "atom type" is introduced, which allows an easy transfer of the parameters developed for a limited set of model compounds to a much wider ranger of molecular systems: referred to as the "transferability" of the force field. Atoms belonging to a particular type typically share the same bonded and non-bonded parameters, even though they can belong to different molecules or even functional groups. The assignment of atom types ("typing") relies not only on atomic numbers, but also on their chemical environment (e.g. charge and hybridization state, hydrogen bonding potential, etc.) (MacKerell, 2004; Guvench and MacKerell, 2008). The larger number of atom types, in principle, provides a more accurate description of a molecular system but decreases force field transferability. Therefore, some compromise is sought, where several dozens of different atom types are used in modern biomolecular force fields.

Even though parameters in biomolecular force fields such as CHARMM or Amber are developed for specific residues, they can be used for modeling of other compounds with similar functional groups, allowing for simulations of protein complexes with different substrates, cofactors, inhibitors, and so on (Guvench and MacKerell, 2008). As a rule, force fields for different classes of biomolecules (proteins, nucleic acids, lipids, carbohydrates) can be used together, which allows simulations of proteins in lipid bilayers or protein–nucleic acid complexes, for example (Wang *et al.*, 2001; MacKerell and Nilsson, 2008; Sapay and Tieleman, 2008). For instance, the CHARMM protein force field (MacKerell, 1998b) can be used together with CHARMM lipid (Feller and MacKerell, 2000) and/or nucleic acid (Foloppe and MacKerell, 2000) force fields. There is also a trend toward the development of the more general force fields and software packages that allow automatic assignment of atomic types and partial atomic charges (e.g. the general Amber force field, GAFF) (Wang *et al.*, 2004).

The choice of model compounds is an important factor for force field development. They should be relatively small, representative of protein functional groups, and have available reference data. Beside model compounds, some test molecules should be selected that are not included during parameter optimization, but which can be used to test the accuracy and transferability of parameters (MacKerell, 1998). For instance, for protein backbone parameter development, *N*-methylacetamide

(NMA) is often used as a model compound since it contains important functional moiety, amide linkage, capped by methyl groups (MacKerell, 1998b; MacKerell, 1998). In addition, a large number of structural, vibrational, condensed-phase thermodynamic data are available for this compound, which makes it a good parameterization target (MacKerell, 1998). NMA is used for parameterization of nonbonded and some internal parameters, whereas a larger molecule, alanine dipeptide (ALAD) is commonly used for the optimization of backbone conformational properties (MacKerell, 1998b). MD simulations using other di- and tetrapeptides have been used for parameter validations. Parameters for each protein side chain are also obtained using side chain analogs, that is molecules having the same atoms as amino acid side chain truncated at C_α (for instance, acetamide for Asn, methyl indole for Trp) with some exceptions (MacKerell, 1998). The experimental and/or QM data for these model compounds are then used for parameter optimization.

8.2.2
Intramolecular Parameters

The mathematical forms for the terms in the empirical potential energy function usually represent a compromise between simplicity and accuracy (MacKerell, 2001). For instance, harmonic terms for bond stretching and angle bending are sufficiently accurate at ambient conditions. In some force fields, commonly used for small molecules, for example MMFF (Halgren, 1996), MM3 (Lii and Allinger, 1991), MM4 (Allinger *et al.*, 1996), QMFF (Ewig *et al.*, 2001), the bonded terms are expanded to include higher-order terms to account for anharmonicities and/or crossterms to take into account coupling of different bonded interactions. These force fields are termed Class II force fields, whereas biomolecular force field missing such terms are referred to as Class I force fields (Leach, 2001) In some potential energy functions, the Morse potential is used for bonds and cosine-based term applied for angles (Mayo *et al.*, 1990; Rappe *et al.*, 1992). These terms increase the force field accuracy for molecular geometries far from equilibrium structures but are usually not required for typical protein simulations. Equilibrium bond lengths and angles are parameterized on the basis of spectroscopic measurements, if available, or QM geometries of model compounds, whereas corresponding force constants are fitted to best reproduce experimental or QM gas-phase vibrational spectra (MacKerell, 1998; MacKerell, 2004).

For the dihedral energy term, a Fourier cosine series (Eq. (1)), with different values of multiplicity n (ranging from 1 to 6) corresponding phase angles χ, and force constants K_χ, are fitted to reproduce a QM potential energy surface (PES) corresponding to a rotation around given bond. This is especially important for protein force fields, where an accurate treatment of both backbone and side chain PES is critical for protein structure and dynamics. For instance, it was recently shown that, to different extents, standard CHARMM22, Amber, and OPLS-AA force fields do not provide sufficiently good agreement for the distribution of protein backbone ϕ and ψ dihedral angles compared to crystal structure and QM calculations (Hu *et al.*, 2003; MacKerell *et al.*, 2004). This deficiency manifested itself in the presence of an unrealistically large fraction of π-helical conformations in simulated proteins (Shirley and Brooks, 1997; Feig *et al.*, 2003). This led to revision of corresponding dihedral parameters for Amber (Wang *et al.*, 2000; Okur *et al.*, 2003) and OPLS (Kaminski *et al.*, 2001). In CHARMM a different approach was used: dihedral crossterms of grid-based energy correction maps (CMAP) were utilized to accurately reproduce a high-level PES for alanine, glycine, and proline dipeptides with subsequent empirical adjustments (MacKerell *et al.*, 2004; MacKerell, 2004). By using this approach, the agreement between CHARMM and experimental (PDB survey data) or QM/MM ϕ and ψ distributions in solution was substantially improved (MacKerell, 2004).

Two additional terms are present in the CHARMM potential energy function. The first one is a Urey-Bradley angular term (Urey and Bradley, 1931), which treats the 1–3 interaction by a harmonic potential depending on the distance between terminal atoms and helps reproduce more accurately reference vibrational spectra of model compounds (Guvench and MacKerell, 2008). The second term (also present in some other bimolecular force fields) is a separate improper dihedral energy term, treated in CHARMM with a harmonic potential similar to bond and angle terms. Such a term is used to maintain planarity of some functional groups (e.g. amide linkage in proteins) and can be used to improve the agreement with reference vibrational spectra. Improper torsion terms are also used in united atom force fields to prevent chiral center inversion during MD simulations (Leach, 2001). In these force fields, hydrogen atoms are treated as a part of non-H atoms, to which they are bonded, and thus the latter are called *united or extended atoms*. Most often such treatment applies to nonpolar H atoms only, whereas polar hydrogen atoms are still treated explicitly to properly describe hydrogen bonding interactions. Some (mostly earlier) versions of popular biomolecular force fields CHARMM19 (Neria *et al.*, 1996), Amber ff84 (Weiner *et al.*, 1984) and ff06 (Yang *et al.*, 2006), GROMOS87 and GROMOS96 (van Gunsteren *et al.*, 1996; Daura *et al.*, 1996; Daura *et al.*, 1998), and OPLS-UA (Jorgensen and Tiradorives, 1988) belong to the united atom force field category, whereas CHARMM22 (MacKerell, 1998b), OPLS-AA (Jorgensen *et al.*, 1996), and Amber ff95 (Cornell *et al.*, 1995) are all-atom force fields.

8.2.3
Nonbonded Parameters

The accurate treatment of nonbonded interactions is usually a key component of any biomolecular force field development because of the strong influence of the environment on protein properties (MacKerell, 2001). These interactions are not only limited to those between different molecules but also within large molecules between nonbonded atoms and can have an influence on the conformational energetics and to less extent molecular geometries and vibrational spectra (Guvench and MacKerell, 2008). The functional form used for nonbonded interactions, shown in Eq. (1) includes a Coulomb potential for electrostatics and a LJ 6–12 potential for vdW interactions, representing the combination of attractive dispersion forces and exchange electron repulsion. The implementation of an LJ potential may vary between different biomolecular force fields with the use of different combination rules and special treatment of 1–4 interactions. Combination rules are used to simplify LJ parameter development and are based on the assumption that parameters for cross-interactions between atom types i and j, ε_{ij} and σ_{ij}, can be obtained by some combination of parameters for types i and j. In CHARMM and Amber, Lorentz–Berthelot combining rules are used, where LJ well-depths are combined using geometric mean (the Berthelot rule) $\varepsilon_{ij} = \sqrt{\varepsilon_{ii}\varepsilon_{jj}}$ and interatomic separations using arithmetic mean (the Lorentz rule): $\sigma_{ij} = 0.5(\sigma_{ii} + \sigma_{jj})$ (MacKerell, 1998b; Guvench and MacKerell, 2008). Geometric means for both ε_{ij} and σ_{ij} are used in OPLS-AA and GROMOS force fields. The validity of using Lorentz–Berthelot combining rules in molecular simulations has been recently questioned, and the use of the more advanced Kong combining rules (Kong and Chakrabarty, 1973) has been suggested (Delhommelle and Millie, 2001).

Typically nonbonded interactions between atoms separated by one or two chemical bonds (1–2 and 1–3 interactions) are not included in atomistic force fields, whereas those for atoms separated by three bonds (1–4 interactions) are typically included, but might be scaled to differentiate from the rest of non-bonded interactions. This is motivated by the need to tune conformational energetics while keeping the same set of parameters for intra- and intermolecular interactions (Guvench and MacKerell, 2008).

In most biomolecular force fields, LJ parameters are obtained by reproducing condensed-phase properties such as densities or molecular volumes and heats of vaporization of neat liquids at ambient conditions (MacKerell, 2004): an approach pioneered during the development of the original OPLS force field (Jorgensen and Tiradorives, 1988). Alternatively, crystal parameters and heats of sublimation can be used as target properties (Warshel and Lifson, 1970; MacKerell *et al.*, 1995). The common problem encountered with the above approach is the so-called parameter correlation, where different sets of LJ parameters can reproduce the same condensed-phase properties (MacKerell, 2001). On the other hand, the direct application of QM data for LJ parameter optimization is difficult owing to the need for a very high level of theory and basis sets to properly account for dispersion interactions. To alleviate these problems the combined QM–empirical approach has been used where high-level *ab initio* data on model compound interaction with rare gas atoms were used in the LJ parameter optimization along with experimental pure solvent properties (Yin and Mackerell, 1998; Chen *et al.*, 2002).

Different approaches have been used in biomolecular force field for the determination of partial atomic charges. In Amber, charges for model compounds are obtained from the restrained fitting to QM electrostatic potentials (RESP) (Bayly *et al.*, 1993; Cieplak *et al.*, 1995). Hyperbolic restraints help reduce charges on buried atoms and remove some dependence on the molecular conformation (Leach, 2001). In Gromacs and OPLS, charges are empirically adjusted along with LJ parameters to reproduce condensed-phase properties. In the CHARMM force field, partial atomic charges are optimized to reproduce QM interaction energies (IEs) and geometries of model compounds with water (MacKerell, 1998b). Polarization effects not present in the empirical CHARMM force field functional form are included implicitly by choosing partial charges that overestimate gas-phase dipole moments (MacKerell, 2004). Therefore, reference QM IE are scaled up (by a factor of 1.16 for neutral models) and CHARMM minimum-energy interaction distances are kept shorter than their QM counterparts. This method allows balancing protein–protein, protein–water, and water–water interactions (Guvench and MacKerell, 2008).

Some empirical force fields also include an explicit hydrogen bonding term, which can be modeled, for example by 10–12 or 4–6 function similar to an LJ potential, but often direction dependent through the use of cosine factors (Goodford, 1985; Vedani, 1988). However, it was shown that the combination of Coulomb and LJ potentials can accurately reproduce nonbonded interactions between molecules including hydrogen-bonded interactions without the need for such a term (Reiher, 1985). If the hydrogen bond directionality is not reproduced well, it can be mimicked through the use of off-site charge distributions reminiscent of lone pairs (Dixon and Merz, 1997; Harder *et al.*, 2006). Recently, as a part of OPLS-AA force field revision, the off-plane charge model for aromatic rings was proposed, which better mimics their electronic configuration and provides more accurate tryptophan side chain hydration free energy as well as cation–π interactions (Xu *et al.*, 2007). The idea to use off-site charge distribution was also used in both early and more recent four- and five-site water models (Bernal and Fowler, 1933; Stillinger and Rahman, 1974; Jorgensen *et al.*, 1983; Mahoney and Jorgensen, 2000) (see below).

Comparison of experimental and calculated properties such as solvation free energies, dielectric constants, isothermal compressibilities, NMR relaxation times can be also used to test the quality of the developed force field parameters. For instance, testing of these properties was, for instance, an integral part of CHARMM alkane and ether parameter development (Yin and Mackerell, 1998; Klauda *et al.*, 2007; Vorobyov *et al.*, 2007). In general, agreement with experimental properties is reasonably good. However, many biomolecular force fields (CHARMM, Amber, OPLS) tend to underestimate hydration free energies of neutral amino acid side chain analogs (Shirts *et al.*, 2003; Deng and Roux, 2004). Solvation free energies have also been used as target data in some force fields such as OPLS and recent GROMOS versions. In GROMOS, for example, solvation-free energies in both water and cyclohexane were used as target data (Oostenbrink *et al.*, 2004). In principle, using target partitioning data for parameterization is a logical step in ensuring correct protein structures, in particular, for membrane proteins, which require separation of hydrophobic and hydrophilic residues (discussed in greater detail below). We note, however, that this is limited in the case of ionizable side chains, for which direct experimental measurement of charged species in nonpolar phases is lacking. However, indirect measures of solvation free energies can be obtained. For instance, using CHARMM, the free energy of a charged arginine (Arg) analog, methyl guanidinium, is accurately reproduced in water, but is severely underestimated in cyclohexane (Vorobyov *et al.*, 2008). This deficiency stems from incorrect dielectric constant of nonpolar substances, such as liquid hydrocarbons, which appear vacuum-like (dielectric constant of ~1), instead of the experimental value of ~2 (Vorobyov *et al.*, 2005; Vorobyov *et al.*, 2007) owing to the lack of explicit electronic polarizability. We now discuss recent attempts to incorporate electronic polarizability into force fields and return to the effect on amino acid solvation later in this chapter.

8.2.4
Polarizable Force Fields

Electrostatic interactions in biomolecular force fields are normally treated using fixed partial atomic charges interacting through Coulomb law. The electronic polarization is not explicitly included but treated implicitly through assignment of partial atomic charges overestimating gas-phase molecular dipole moments and IE to reproduce condensed-phase properties. Such force fields are commonly referred to as additive or non polarizable with CHARMM22 (MacKerell, 1998b), AMBER ff95 (Cornell *et al.*, 1995), OPLS-AA (Jorgensen *et al.*, 1996), and GROMOS96 being representatives. They provide a fairly accurate description of aqueous condensed-phase environments for protein simulations but cannot simultaneously provide accuracy for systems with distinctly different polarity, for example, gas phase or liquid hydrocarbon (MacKerell, 1998b; MacKerell, 2004). Empirical force fields, which explicitly include an electronic polarizability contribution, can provide a remedy for such deficiencies. Polarizability can be included in Eq. (1) as a separate term and/or modification of the Coulomb electrostatic term. Electronic polarizability can be introduced via point dipoles, fluctuating charges, classical Drude oscillators (shell), or a combination of those models. Comprehensive reviews of the different models are given elsewhere (Halgren and Damm, 2001; Rick and Stuart, 2002; Ponder and Case, 2003; MacKerell, 2004; Yu and van Gunsteren, 2005; Patel and Brooks, 2006; Warshel *et al.*, 2007).

In the point dipole moment model, explicit polarization is included in the form of induced dipoles assigned to polarizable atoms. This was the first approach to be applied for modeling polarization in biomolecular systems using the Langevin dipole model by Warshel and Levitt (1976). The induced dipole moment, $\boldsymbol{\mu}_i$, is proportional to the electric field $\mathbf{E}_i$, where the constant of proportionality, α_i, is the atomic polarizability, which can be considered as a force field parameter (Rick and Stuart, 2002; MacKerell, 2004). The electrostatic energy of a polarizable system includes an induction term:

$$U_{\text{ind}} = -\frac{1}{2}\sum_i \boldsymbol{\mu}_i \cdot \mathbf{E}_i^0 \tag{2}$$

where $\mathbf{E}_i^0$ is the electric field due to permanent charges (and dipoles, quadrupoles in some models (Ren and Ponder, 2003)). The expression for μ_i can be written as

$$\boldsymbol{\mu}_i = \alpha_i \left[\mathbf{E}_i^0 - \sum_{j \neq i} \mathrm{T}_{ij} \boldsymbol{\mu}_j \right] \tag{3}$$

where T_{ij} is the dipole field tensor describing interactions between induced dipoles i and j. Since each induced dipole feels the presence of other induced dipoles, their values can be obtained iteratively or through matrix inversion

(Rick and Stuart, 2002). The interactions between induced dipoles can be screened at short distances as was first suggested by Thole (1981) to avoid the "polarization catastrophe," that is infinite polarizability component as dipoles come closer than $(4\alpha_i\alpha_j)^{1/6}$) (Rick and Stuart, 2002; Yu and van Gunsteren, 2005). Alternatives to solve this problem include increasing vdW repulsion, reducing polarizabilities or spreading them over several centers, using a non linear dependence of $\boldsymbol{\mu}$ on the electric field (Yu and van Gunsteren, 2005).

In the fluctuating charge (FC) models (Rick and Stuart, 2002; Patel and Brooks, 2006) the partial atomic charges are not fixed but can redistribute between atoms in a molecule following the principle of electronegativity neutralization (Patel and Brooks, 2006). Such charge transfer depends on the atomic electronegativities (abilities to attract electrons), χ_i^0, and hardnesses (resistance to change the number of electrons, defined as one-half of the energy required to transfer electron from one atom to another), J_{ii} (Rick and Stuart, 2002; Patel and Brooks, 2006). Both χ_i^0 and J_{ii} are typically considered force field parameters to be optimized (Rick and Stuart, 2002). The electrostatic energy of an isolated molecule in this model can be represented by

$$U_{\text{elec}} = \frac{1}{4\pi\varepsilon_0}\left(\sum_i \chi_i^0 q_i + \frac{1}{2}\sum_i\sum_j q_i q_j J_{ij}(r_{ij})\right) \tag{4}$$

J_{ij} can be assigned to be simply $1/r_{ij}$ or some screening factor can be used. In most FC models, charges are allowed to fluctuate only within each molecule, so there is no intermolecular charge transfer, and interactions between atoms in different molecules are calculated through Coulomb law (Rick and Stuart, 2002). The values of atomic charges q_i are obtained iteratively by seeking the stationary point $\partial U_{\text{elec}}/\partial q_i = 0$ for atoms in the system, $i = 1, \ldots N$ under constraint of molecular charge preservation, which is equivalent to electronegativity neutralization within each molecule (Rick and Stuart, 2002; Patel and Brooks, 2006).

In the classical Drude oscillator model, auxiliary charged particles are attached to polarizable atoms via harmonic springs with a force constants, k_{Di}. The partial atomic charges, q_i, are distributed between the atomic core (c) and Drude particle (D): $q_i = q_{\text{ci}} + q_{\text{Di}}$. The magnitudes of q_{Di} can be obtained from atomic polarizabilities α_i as $q_{\text{Di}} = \sqrt{\alpha_i k_{Di}}$. In turn, values of α_i along with partial atomic charges q_i are force field parameters determined from restrained fitting to the QM electrostatic potential maps perturbed by background charges placed around a molecule (Anisimov *et al.*, 2005). The electrostatic potential term in the potential energy function is modified to include Coulomb interactions between atomic cores and Drude particles. The Coulomb interactions between Drude particles on atoms separated by one or two bonds (1–2 and 1–3) are included but smeared by using the exponential screening function similar to that used for point dipole models as proposed by Thole (1981). A harmonic term describing the self-energy of a polarizable atom $0.5k_{Di}d^2$ (where d is the displacement of the Drude particle from the atomic core) is included. The Drude force constant k_{Di} can be treated as a tensor, allowing for anisotropic polarizability (Harder *et al.*, 2006). The positions of Drude particles can be determined self-consistently under the Born–Oppenheimer approximation (Anisimov *et al.*, 2005).

The iterative nature of solutions for polarizable potentials, which needs to be applied at every time step of MD to find equilibrium values of induced dipoles, charges or Drude particle positions, make these calculations very time consuming compared to simulations using nonpolarizable force fields. The application of polarizable models to MD simulations of biomolecular systems was facilitated through the development of extended Lagrangian simulation protocols, where polarizable degrees of freedom were assigned small masses and propagated with atomic degrees of freedom (Rick *et al.*, 1994; Vanbelle *et al.*, 1992; Tuckerman and Martyna, 2000; Lamoureux and Roux, 2003). Polarizable force field development to date has concentrated much on water models (Barnes *et al.*, 1979; Caldwell *et al.*, 1990; Dang *et al.*, 1991; Vanbelle *et al.*, 1992; Rick *et al.*, 1994; Lamoureux *et al.*, 2003; Ren and Ponder, 2003; Yu and van Gunsteren, 2004; Lamoureux *et al.*, 2006) owing to its importance in biomolecular systems and as a simple test-case for methodology testing. We discuss water models in the following section and concentrate here on protein force fields.

While many early studies on polarizable protein force fields used scaled charges and other parameters from nonpolarizable models (Caldwell and Kollman, 1995a; Meng *et al.*, 1996; Wang *et al.*, 2000), a lot of efforts were made for a careful parameter development for model compounds, building blocks of proteins and other biomolecules, such as hydrocarbons (Vorobyov *et al.*, 2005; Patel and Brooks, 2006a; Lopes *et al.*, 2007), ethers (Hermida-Ramon and Rios, 2000; Borodin and Smith, 2006; Vorobyov *et al.*, 2007), alcohols (Gao *et al.*, 1995; Patel and Brooks, 2005a; Patel and Brooks, 2005; Anisimov *et al.*, 2007; Geerke and Van Gunsteren, 2007), aldehydes (Hermida-Ramon and Rios, 1998; Mannfors *et al.*, 2000), as well as amines and amides (Ding *et al.*, 1995; Meng *et al.*, 1996; Rick and Berne, 1996; Mannfors *et al.*, 2001; Xie, *et al.*, 2007; Harder *et al.*, 2008). The parameterization for NMA as a model for the protein backbone can be considered as one of the cornerstones of the polarizable protein force field development and has proven to be challenging (Patel and Brooks, 2004; Xie, *et al.*, 2007; Harder *et al.*, 2008). Protein force fields based on parameterization of a number of model compounds have been reported as well (Banks *et al.*, 1999; Stern *et al.*, 1999; Kaminski *et al.*, 2002; Palmo *et al.*, 2003; Ponder and Case, 2003; Kaminski *et al.*, 2004; Patel and Brooks, 2004; Donchev *et al.*, 2008).

MD simulations of proteins using polarizable protein force fields are still scarce but are rapidly emerging (Patel *et al.*, 2004; Harder *et al.*, 2005; Kim *et al.*, 2005; Maple *et al.*, 2005; Friesner, 2006; Friesner, 2006; Khoruzhii *et al.*, 2008; Jiao *et al.*, 2008; Masella *et al.*, 2008). Some earlier studies were performed either in the gas phase or using implicit solvent models (Banks *et al.*, 1999; Kaminski *et al.*, 2002), whereas peptide and protein polarizable MD simulations using explicit solvent representation have become more widespread (Liang and Walsh, 2007; Patel *et al.*, 2004; Harder *et al.*, 2005; Kim *et al.*, 2005; Jiao *et al.*, 2008; Khoruzhii *et al.*, 2008; Masella *et al.*, 2008). Polarization was found to have a substantial effect on solvent structure and dynamics near charged residues (Kim *et al.*, 2005)

In other calculations of protein–ligand binding energies using polarizable protein and solvent models have been performed (Maple *et al.*, 2005; Friesner, 2006; Khoruzhii *et al.*, 2008; Jiao *et al.*, 2008) with good agreement with experimental data (Khoruzhii *et al.*, 2008; Jiao *et al.*, 2008). Even though polarizable force fields might be considered superior to additive models, they might be not available for all systems of interest and might be harder to parameterize owing to larger number of parameters and a wider range of reference data aimed to be reproduced. In many cases nonpolarizable force models might be sufficient as shown later in this chapter. Polarizable force fields also have limitations and might not be able to reproduce polarizability limits (Giese and York, 2004), in which case QM or QM/MM approach might be a method of choice.

8.2.5
Water, Ion, and Lipid Models

Water is the dominant component of any biological system: its interaction with biomolecules determining their structure and dynamics. Therefore, the use of an accurate water model is crucial. The most widely used nonpolarizable water models in protein MD simulations, TIP3P (three-point transferable intermolecular potential) (Jorgensen *et al.*, 1983) and SPC (simple point charge) (Berendsen *et al.*, 1981) or simple point charge–extended (SPC–E) (Berendsen *et al.*, 1987), are rigid three-site water models. This means that there are three Coulomb/LJ interaction sites centered on O and H nuclei with fixed bond lengths and angle. These models have some differences in geometry (e.g. HOH angle of 104.52° for TIP3P and tetrahedral angle of 109.47° for SPC and SPC/E) as well as nonbonded parameters. In another popular TIP4P model, the site of negative charge is moved away from oxygen along the HOH bisector, which improves structure and properties compared to TIP3P (Jorgensen *et al.*, 1983). These models reproduce condensed-phase properties such as density, heat of vaporization and water structure well at ambient conditions are widely used in conjunction with biomolecular force fields. However, all water models have some limitations, for example too high diffusion coefficient, lack of long-range structure, or overestimation of bulk potential energy (MacKerell, 2004). There also exist five-site water models, such as ST2 (Stillinger and Rahman, 1974) and TIP5P (Mahoney and Jorgensen, 2000), where oxygen charge is distributed between two lone pair sites. The TIP5P water model was found to better reproduce experimental temperature-dependent properties (such as the density vs. temperature profile) and liquid water structure (Jorgensen and Tirado-Rives, 2005), at the expense of increased computational cost. Flexible water models, on the basis of SPC or TIP4P models, have also been reported (Ferguson, 1995; Lawrence and Skinner, 2003; Wu *et al.*, 2006), with improved dielectric constant, diffusion coefficient, and liquid-phase infrared spectra, for example. However, the use of such models is more computationally expensive and requires the use of a smaller time step for MD simulations.

A fundamental limitation of water models, mentioned earlier, is the lack of explicit electronic polarizability, because gas-phase properties had to be sacrificed to get accurate condensed-phase properties. In particular, the water dipole is expected to be quite different in the bulk liquid, around ionic species or in some hydrophobic environment, which is not possible in a non-polarizable model. Polarizable models can provide good agreement for a wide range of condensed-properties without sacrificing the gas-phase properties. Many polarizable water models are capable of reproducing both gas-phase (dipole moment, water cluster structure and energetics) and condensed-phase properties (molecular density, heat of vaporization, and dielectric constant). One example is the four-site SWM4-DP (simple water model with 4 sites and Drude polarizability) Drude polarizable water on the basis of the TIP4P geometry (Lamoureux *et al.*, 2003; Lamoureux *et al.*, 2006). This model was optimized to reproduce a wide range of gas-phase, condensed-phase, and interfacial properties (Lamoureux *et al.*, 2006). However, atomic polarizabilities need to be scaled down from their gas-phase values to reproduce condensed-phase properties (Alfredsson *et al.*, 1998; Halgren and Damm, 2001; Lamoureux *et al.*, 2003). The need for such reduction was attributed to the energy cost of the overlapping electron clouds opposing induction (Halgren and Damm, 2001; Lamoureux *et al.*, 2003; Anisimov *et al.*, 2005) Such polarizability scaling was also extended to other model compounds (Anisimov *et al.*, 2005; Vorobyov *et al.*, 2007; Anisimov *et al.*, 2007), although the choice of the appropriate scaling factors is still an open question.

This does not mean, though, that previous MD simulations using nonpolarizable potentials produced incorrect results. For instance, it was recently reported that the nonpolarizable TIP3P water model describes two extreme cases of ionic solvation in the bulk phase and in strongly anisotropic single file of water inside a narrow pore very similar to the Drude polarizable SWM4-DP water model (Allen *et al.*, 2004b). Nevertheless, polarizable models of water, and other solvents might provide better descriptions of their structural reorganization and dynamics around charged and polar species. Polarizable models also offer the possibility of describing ionization. For example, the dissociable water model PM6 consists of O^{2-} and H^{+} moieties (Stillinger and David, 1978) and has been used to study the mechanism of proton transfer along hydrogen-bonded chains of water molecules (Pomes and Roux, 1998) and to elucidate the proton blockage mechanism in aquaporins (Chakrabarti *et al.*, 2004).

In this review we focus our application on ion channels. The study of ion channels relies on a correct parameterization of metal and halide ion interactions with water and protein. Accurate polarizable models for ion solvation (Caldwell *et al.*, 1990; Dang *et al.*, 1991; Meng *et al.*, 1994; Stuart and Berne, 1996; Bryce *et al.*, 1998; Grossfield *et al.*, 2003; Lamoureux and Roux, 2006; Whitfield *et al.*, 2007; Warshel *et al.*, 2007; Warren and Patel, 2007), partitioning (Dang, 1999) and ion pair interactions (Shelley *et al.*, 1993; Soetens *et al.*, 1997) are also an active area of development. Polarizability has been found to be important for ion solvation energetics, ion-solvent distributions, binding to interfaces and ion pair interaction potentials (Halgren and Damm, 2001; Warshel *et al.*, 2007). It was also found that cation–π interactions, which are commonplace in protein structures (Dougherty, 1996; Crowley and Golovin, 2005), are more accurately described by polarizable force fields (Caldwell and Kollman, 1995b).

The development of accurate ion force field model is challenged by the absence of reliable experimental solvation data for ions in water and even more so in other solvents. Only hydration free energies of ion pairs and $\Delta\Delta G_{\mathrm{sol}}$ between same-charge ions (e.g. Na^+ and K^+) are known with certainty. The reported absolute solvation free energies of ions typically rely on some extrathermodynamic assumption regarding the magnitude of ΔG_{sol} for H^+ or OH^-. The solvent interfacial potential should be also taken into account while comparing different ion solvation free energy estimates and can account for observed differences (Lamoureux and Roux, 2006). Neutron diffraction data might provide another source of reference information. However, a radial distribution function (RDF) from diffraction studies may have peak positions only accurate to around ±0.05 to 0.1 Å. Such a leeway in σ of the LJ parameters can lead to variations in solvation energetics of the order of 10 kcal mol^{-1} (Allen *et al.*, 2006a). This almost "free" parameter requires other accurate experimental benchmark data to constrain its value, such as gas-phase interactions of ions with water molecules. However, in the framework of nonpolarizable force fields it has been a challenge to reproduce both gas phase interactions while at the same time reproducing solution phase solvation energetics (Berneche and Roux, 2001; Allen *et al.*, 2006b). This deficiency can be circumvented by using polarizable ion and solvent potentials (Grossfield *et al.*, 2003; Lamoureux and Roux, 2006; Warren and Patel, 2007). The consistent procedure for the parameter development for a series of alkali metal and halide ions in the framework of the classical Drude oscillator model on the basis of a wide range of experimental data has been reported recently (Lamoureux and Roux, 2006). The developed parameters were found to be consistent with a unique hydration free energy scale corresponding to the absolute intrinsic hydration free energy of proton of −247 kcal mol^{-1}, thus allowing an estimate of an important quantity not available from experiment (Lamoureux and Roux, 2006).

For membrane protein and especially ion channel MD simulations, an accurate description of ion solvation not only in water, but also liquid amides is of great importance, because an ion may leave the aqueous environment to be solvated by protein groups (Allen *et al.*, 2006b; Allen *et al.*, 2006a). In nonpolarizable CHARMM, the LJ interactions of ions with amides have been adjusted to reproduce either bulk solvation data or gas-phase interactions (Allen *et al.*, 2006b). Reasonable agreement for both ion gas-phase and condensed-phase energetics in water and formamide was demonstrated by a polarizable AMOEBA potential (Grossfield *et al.*, 2003). In a recent combined experimental and computational study, the Drude polarizable force field was found to provide a good agreement with ion–NMA gas-phase reference data as well as experimental ion solvation in liquid NMA, water, and partitioning between these solvents thus ensuring a proper balance of different interactions for ion channel MD simulations.

Lipid models are also essential for membrane protein MD simulations (Merz, 1997; MacKerell, 2004; Sapay and Tieleman, 2008; Klauda *et al.*, 2008). The development of accurate atomistic lipid models is challenging since at physiological conditions they are in the fluid-like liquid-crystalline state, which leads to a lack of experimental high-resolution structural data necessary for their careful parameterization. Careful optimization of parameters of lipid components such as hydrocarbon, phosphate, choline, glycerol, ether moieties are performed using model compounds, and then the derived lipid parameters are tested against experimental low-resolution X-ray, solid-state deuterium NMR, neutron diffraction data (MacKerell, 2004). The most commonly used lipid models used in atomistic MD simulations are those associated with CHARMM and GROMOS force fields. The phospholipid model available in CHARMM has been developed over the years to include parameters for saturated (Schlenkrich *et al.*, 1996; Feller and MacKerell, 2000; Klauda *et al.*, 2005; Klauda *et al.*, 2008), monounsaturated (Feller *et al.*, 1997), and polyunsaturated (Feller *et al.*, 2002) phosphatidylcholines (PC). Parameters for other lipid classes such as phosphatidylethanolamine and sterols (Cournia *et al.*, 2005) are also available in CHARMM (Klauda *et al.*, 2008) CHARMM lipid force field development was on the basis of reproduction of a range of experimental and high-level QM data and has undergone several revisions related to improvement of dihedral and nonbonded parameters (Feller and MacKerell, 2000; Klauda *et al.*, 2005; Klauda *et al.*, 2008). The latest version of this force field (CHARMM27R) was found to provide excellent agreement for a number of experimental parameters for lipid bilayers and monolayers such as structure factors, electron density profiles, deuterium order parameters, NMR spin lattice relaxation times, and translational diffusion constants (Klauda *et al.*, 2008). There were also considerable efforts toward the development of the GROMOS lipid force field and several models have emerged over the years (Berger *et al.*, 1997; Daura *et al.*, 1998; Chiu *et al.*, 1999a; Chiu *et al.*, 1999b; Chandrasekhar *et al.*, 2003). Parameters for other lipid classes (for example cholesterol (Baran and Mazerski, 2002), sphingomyelin (Chiu *et al.*, 2003), and phosphatidylserine (Mukhopadhyay *et al.*, 2004)) have also been developed. GROMOS lipid force fields are united atom models with zero charges on carbon atoms, thus allowing substantial speedup of MD simulation using these parameters.

One of the controversial issues related to lipid force field parameters is the surface area per lipid and appropriate value of surface tension during MD simulations. The area per lipid is a good structural parameter to check force field quality. However, experimental values are available only for a few bilayers and their values are subject to uncertainties (Nagle and Tristram-Nagle, 2000). Nevertheless, MD simulations of lipid bilayers using constant experimentally determined surface area are commonplace. MD simulations of dipalmiotylphosphatidylcholine (DPPC) lipid bilayers using the latest CHARMM force field produce nonzero surface tension (20 dyn/cm per leaflet for systems consisting of 72 lipids) when simulated at experimental surface area (Klauda *et al.*, 2008). This result is in disagreement with experimental data on macroscopic lipid films and vesicles indicating nearly zero surface tension (Tien and Diana, 1968; Israelachvili *et al.*, 1977; Jahnig, 1996). The reason for such a discrepancy is still under contention (Klauda *et al.*, 2008). However, the nonzero surface tension may be related to the relatively small size of the membrane patch used in MD simulations (Feller and Pastor, 1999; Klauda *et al.*, 2008). Undulations, known to occur in macroscopic membranes, might balance the positive surface tension for small membrane patches, not sampled by atomistic MD simulations (Klauda *et al.*, 2008). Alternatively, MD simulations at constant surface area producing nonzero surface tension maybe considered (Feller and Pastor, 1999; Klauda *et al.*, 2008). However, in the popular GROMACS force field implementation by Berger *et al.* (1997), the positive surface tension was chosen to be corrected by reducing lipid head group charges. A similar approach was used in a recent attempt to modify CHARMM lipid force field by Sonne *et al.* (2007). Such modifications might, however, result in the loss of balance of different lipid–lipid, lipid–solvent, and lipid–protein interactions.

Lipid bilayers exhibit dipole potentials that arise from a nonuniform distribution of charge at their interfaces. These positive potentials create extra barriers for cation permeation. The dipole potential can be determined from MD simulation by double integration of charge density using Poisson's equation (Lamoureux *et al.*, 2003). Experimental estimates for lipid bilayers are available from freeze-fracture experiments (cryo-electron microscopy) (Wang *et al.*, 2006a), or the partitioning of large hydrophobic ions (Clarke, 2001). These estimates indicate that for saturated PC bilayers the dipole potential should be on the order of 200–500 mV (Clarke, 2001; Wang *et al.*, 2006a). However, MD simulations of these lipids in CHARMM, GROMOS, and other nonpolarizable force fields typically produce dipole potential values higher than experimental estimates (e.g. 850 mV for CHARMM27R) (Wang *et al.*, 2006a; Klauda *et al.*, 2008). This discrepancy can be corrected by using a polarizable force field. Our preliminary studies indicate that the use of Drude polarizable force field parameters for lipid hydrocarbon tails alone results in a dipole potential consistent with experiment while maintaining the correct lipid geometry and order (unpublished results).

8.2.6
Coarse-Grained Force Fields

CG force fields treat a group of atoms and/or molecules as a single particle or interaction site. The united atom approximation described earlier can be considered as a limiting type of CG approach. There are even methods where the whole protein is modeled as one or few CG particles (Tozzini, 2005). Typically, simplified (and "softer") potential energy functions are used in CG force fields allowing a substantial increase in MD time step. These methods can provide a speedup of several orders of magnitude and can be used to study phenomena occurring on much larger time and length scales. Protein CG force fields have been devised for the variety of applications (Tozzini, 2005) including refinement of low-resolution structures (Delarue and Dumas, 2004), normal mode analyses (Tama *et al.*, 2002; Bahar and Rader, 2005), studies of protein folding (Levitt and Warshel, 1975; Go and Taketomi, 1978; Miyazawa and Jernigan, 1985; Liwo *et al.*, 1997; Liwo *et al.*, 2005; Maupetit *et al.*, 2007), large-scale motions of proteins and macromolecular assemblies (Trylska *et al.*, 2005; Tozzini and McCammon, 2005; Arkhipov *et al.*, 2006a), and membrane-associated protein and peptide simulations (Venturoli *et al.*, 2005; Smeijers *et al.*, 2006).

CG lipid models are also needed for membrane protein simulations. A number of CG models of lipid membranes have been reported (see Goetz and Lipowsky, 1998; Shelley *et al.*, 2001; Stevens *et al.*, 2003; Marrink *et al.*, 2004; Murtola *et al.*, 2004; Orsi *et al.*, 2008 or Venturoli *et al.*, 2006; Muller *et al.*, 2006; Ayton *et al.*, 2008 for recent reviews). One of the popular lipid CG models by Marrink *et al.* (2004) has been recently adapted for simulations of membrane associated peptides and proteins (Shih *et al.*, 2006; Bond and Sansom, 2006; Bond *et al.*, 2007; Bond and Sansom, 2007; Scott *et al.*, 2008). This model is based on approximately 4-to-1 mapping of nonhydrogen atoms to one CG particle and a gain up to 3–4 orders of magnitude in the computational efficiency compared to atomistic simulations (Marrink *et al.*, 2004), while maintaining reasonable agreement with experimental structural, dynamical and thermodynamic data for model compounds (Marrink *et al.*, 2004; Baron *et al.*, 2007), and lipid assemblies (Marrink *et al.*, 2004).

This model has been applied to study lipid bilayer insertion for a number of membrane proteins providing good agreement with atomistic simulations and experimental data (Bond and Sansom, 2006; Bond *et al.*, 2007; Scott *et al.*, 2008; Sansom *et al.*, 2008). This CG model has been used to model the membrane insertion of the S4 helix and entire voltage sensor domain of the voltage-gated K^+ channel KvAP (Bond and Sansom, 2007). The S4 peptide contains multiple Arg side chains, but was shown to exhibit almost equal preference to adopt membrane-spanning or interfacially bound orientations (Bond and Sansom, 2007). This suggests little penalty for exposing charged Arg to lipid hydrocarbon in sharp contrast to atomistic simulations (Dorairaj and Allen, 2007; Li *et al.*, 2008a). Such discrepancy was found to originate from an incorrect physical representation of charged side chain interactions with solvent and lipids in this CG model, since interaction and partitioning free energies are an order of magnitude smaller than those from atomistic simulations. This leads to underestimation of free energy barriers for Arg translocations across lipid bilayers (Vorobyov *et al.*, 2008). A more detailed description of atomistic simulations for charged side chain translocation across lipid bilayers is provided in Section 8.5.2. The revised version of this force field, called MARTINI, has more CG particle types and improved parameters to approximately reproduce experimental partitioning free energies between polar and nonpolar media for a large number of model compounds (Marrink *et al.*, 2007) achieved good agreement with atomistic simulations, although some discrepancies for charged residues still persist (Monticelli *et al.*, 2008). This revised force field has been used, for example, in MD simulations of mechanosensitive (Yelimov *et al.*, 2008) and voltage-gated potassium channels (Treptow *et al.*, 2008).

Another approach is multiscale coarse-graining (MS-CG) method of Voth and co workers relying on force-matching from atomistic MD simulations (Izvekov and Voth, 2005; Noid *et al.*, 2008a; Noid *et al.*, 2008b). This methodology was successfully applied to simulations of the gramicidin A (gA) ion channel in a hydrated lipid bilayer, where the protein was modeled by atomistic force field whereas lipid and water molecules were described using CG representation (Shi *et al.*, 2006). The use of such methods, which allows resolution exchange between CG and atomistic representations (Ayton *et al.*, 2007), might extend applicability of CG simulations and increase their accuracy (Monticelli *et al.*, 2008).

8.3
Molecular Dynamics Method for Protein Systems

Here we discuss the basic methods for carrying out MD simulations of protein systems. More detailed reviews can be found elsewhere (Allen and Tildesley, 1987; Frenkel and Smit, 2002; Leach, 2001; Vangunsteren and Berendsen, 1990; Becker *et al.*, 2001; Rapaport, 2004; Adcock and McCammon, 2006). MD involves the numerical solution of Newton's equation of motion $F = ma$. First, positions and velocities are assigned to all atoms from experimental data (e.g. crystal or NMR structure), minimized geometry or a previous MD simulation. Initial velocities are usually assigned based on a Maxwell–Boltzmann distribution for a given temperature. Forces, F, acting on every particle, and thus accelerations, $a = F/m$, are calculated from gradients of the potential energy: $F = -\nabla U(\mathbf{r})$. Numerical integration algorithms are based on Taylor series expansion of atomic positions in time and can be derived using the Liouville operator formalism (Tuckerman *et al.*, 1992). The Verlet algorithm (Verlet, 1967) expresses position at the next time step $r(t + \Delta t)$ in terms of the previous and current positions ($r(t - \Delta t)$ and $r(t)$) and acceleration $a(t)$ through a difference of Taylor expansions forward and backward in time, eliminating the need to keep track of velocities which are then estimated via finite differences (FDs). The leap-frog algorithm (Hockney, 1970), instead calculates velocities at half-integer time steps (i.e. $v(t - \Delta t/2)$ and $v(t + \Delta t/2)$) and uses them to calculate new positions $r(t + \Delta t)$. In this algorithm, however, particle positions and their velocities are not calculated at the same time, therefore it is impossible to directly compute total energy at a point in time. The velocity Verlet integrator (Swope *et al.*, 1982), more commonly used in biomolecular simulations today, also explicitly incorporates velocity into equations of motion but calculates positions and velocities for a given time step from their values at previous time step as well as accelerations at current and previous time steps: $r(t + \Delta t) = r(t) + v(t)\Delta t + (1/2)a(t)\Delta t^2$ and $v(t + \Delta t) = v(t) + (1/2)(a(t) + a(t + \Delta t))\Delta t$. All these schemes should, in principle, lead to the same trajectories. There also exist algorithms in which higher-order derivatives of the particle coordinates are involved, such as the predictor–corrector algorithm (Gear, 1971). Such methods are more computationally expensive but can allow longer time steps (Leach, 2001). The most important criteria to be considered while choosing an integrator are the conservation of energy and momentum, time reversibility and ability to use longer time steps (Allen and Tildesley, 1987; Frenkel and Smit, 2002).

In all MD algorithms there are numerical errors, revealed by root mean square (RMS) energy fluctuations. Since the error depends on the time step, Δt, with smaller values not only providing better accuracy, but also limiting phase space sampling. Therefore, some compromise is sought to provide numerical stability as well as reasonable trajectory length. Δt is, however, limited by the fastest molecular motions. For instance, the vibration of a C–H bond is ~10 fs and thus requires the use of ~1-fs time step.To use longer time steps, the bond lengths can be fixed at equilibrium values. In CHARMM protein simulations, all bonds involving hydrogens are fixed with holonomic constraints using the SHAKE algorithm (Ryckaert *et al.*, 1977), in which a system of nonlinear constraint equations is solved using a method of Lagrange multipliers. A similar algorithm called Rattle (Andersen, 1983) is used in a new CHARMM velocity Verlet integrator (Lamoureux and Roux, 2003). The SETTLE algorithm (Miyamoto and Kollman, 1992), which solves for three constraints at a time, can be used to constrain the geometry of rigid water molecules. Another algorithm called LINCS (linear constraint solver) (Hess *et al.*, 1997) was reported to be several times faster than SHAKE and is used in GROMACS simulations to constrain all bond lengths (not just those involving hydrogen). An alternative approach for imposing rigidity is the use of the equations of motion in terms of the quaternions (Evans and Murad, 1977; Allen and Tildesley, 1987; Miller *et al.*, 2002): though this is seldom used in present-day simulations. All such constraint algorithms allow one to increase the time step during MD runs of biomolecular systems to 2 fs or more (Feenstra *et al.*, 1999). An alternative approach is to use multiple time steps for different atomic motions in the system (Streett *et al.*, 1978; Tuckerman *et al.*, 1992; Martyna *et al.*, 1996).

8.3.1
Approximations to Handle Finite Systems

The molecular system will have a finite size that is limited to 10^3–10^6 atoms with present-day computational resources. However, this creates artificial boundary effects. One approach to alleviate these effects is to impose periodic boundary conditions (PBC), in which a system is replicated in three dimensions to provide an infinite periodic lattice. In the MD simulation, once the particle leaves the primary cell in one direction, it would enter from the opposite direction (a process called *image centering*) (Allen and Tildesley, 1987; Leach, 2001). The distance between two atoms i and that in the closest image j is considered to be the minimum image distance (Darden, 2001). The use of PBC has drawbacks, however. For instance, it inhibits fluctuations of the particles in the systems with wavelengths greater than the size of the box (Allen and Tildesley, 1987).

For the simplest cubic or orthorhombic cell, 26 nearest neighbor image cells are present. Depending on the molecular system being modeled, different shapes of the primary cell can be used. For instance, for simulations of roughly spherical proteins it is desirable to limit the number of solvent molecules by using a truncated octahedron or rhombic dodecahedron cell (Allen and Tildesley, 1987; Leach, 2001). For simulations of oriented proteins or proteins elongated in one dimension (e.g. proteins embedded into lipid bilayers), it is sometimes desirable to use hexagonal periodic boundaries (Allen *et al.*, 2003a; Dorairaj and Allen, 2007). The simplest and most widely used space group used in MD simulation is P1, where unit cells are replicated only by translation. It is possible, however, to use different crystal symmetry operations, such as inversions and screw axes, to impose symmetry that may assist in maintaining desired system properties. For instance, $P2_1$ allows, for example, lipid molecules to switch leaflets in membrane simulations while maintaining chirality (Dolan *et al.*, 2002).

An alternative to PBC is to use an implicit representation of the protein environment (e.g. solvent and/or lipid membrane). The interactions with the environment can be modeled in an average way on the basis of some effective potential. In this case the dynamics of the system is typically described by Langevin equation of motion to take into account random collisions with the solvent molecules via a random force, $R(t)$, and a frictional drag, γ (Allen and Tildesley, 1987; Roux *et al.*, 2004):

$$m_i a_i(t) = F_i(t) - \gamma_i v_i(t) m_i + R_i(t) \tag{5}$$

The value of γ is typically in the range 5–100 ps^{-1} (Feig, 2008). If γ_i is large, the inertial term $m_i a_i(t)$ can be neglected resulting in Brownian dynamics (BD) (Roux *et al.*, 2004).

Electrostatic interactions between solute and solvent can be obtained by solving the Poisson–Boltzmann (PB) equation (Sharp and Honig, 1990; Fogolari *et al.*, 2002; Roux *et al.*, 2004), which is a modification of the Poisson equation in the presence of mobile ions. The Poisson equation is the second-order differential equation, relates the variation in the electrostatic potential ϕ with the charge density ρ in the medium with given spatially varying dielectric response $\varepsilon(\mathbf{r})$ (Jackson, 1999)

$$\nabla\varepsilon(\mathbf{r})\nabla\phi(\mathbf{r}) = -\rho(\mathbf{r})/\varepsilon_0 \tag{6}$$

If mobile ions are present, the term describing Boltzmann distribution of these ions is added giving rise to the PB equation, which in the linear form (first-order truncation of a Taylor series expansion of $\sinh[\phi(\mathbf{r})]$) can be written as:

$$\nabla\varepsilon(\mathbf{r})\nabla\phi(\mathbf{r}) - \kappa^2(\mathbf{r})\phi(\mathbf{r}) = -\rho(\mathbf{r})/\varepsilon_0 \tag{7}$$

where κ is the Debye–Hückel screening factor (Feig, 2008). The PB equation can be solved to obtain ϕ, from which the electrostatic contribution to solvation free energy, $\Delta G_{\text{solv, elec}}$, can be obtained (Roux *et al.*, 2004; Feig, 2008).

To perform MD simulations, gradients in $\Delta G_{\text{solv, elec}}$ must be calculated (Feig, 2008). Yet, analytical solutions to the PB equation are available only for a handful of simple boundary conditions (plane, sphere, infinite cylinder, toroid), leading to the need for numerical solutions using approaches such as the boundary element (Zauhar and Morgan, 1985; Rashin, 1990; Zhou, 1993; Vorobjev and Scheraga, 1997; Bordner and Huber, 2003), finite element method (Orttung, 1977; You and Harvey, 1993; Cortis and Friesner, 1997; Holst *et al.*, 2000; Baker *et al.*, 2000) or most commonly used FD method (Warwicker and Watson, 1982; Klapper *et al.*, 1986; Gilson and Honig, 1987; Davis and McCammon, 1989; Holst and Saied, 1995). In the finite difference approach, the solution is usually performed using a cubic lattice, which is superimposed onto the molecular system with each grid point assigned values of ρ, ε, ϕ and screening factor κ. The values of ρ are obtained from the distribution of partial atomic charges over the adjacent grid points, whereas the values of ε are assigned on the basis of the solute–solvent boundary (Darden, 2001). Atomic radii for determination of this solute-solvent boundary are typically parameterized to reproduce solvation free energies of model compounds (Sitkoff *et al.*, 1994; Nina *et al.*, 1997). The boundary can be defined as either molecular (vdW) surface or (more accurately) as solvent-accessible surface by rolling a molecular probe (hard sphere representing water or other solvent moiety, e.g. CH_2 for lipid tails) (Allen *et al.*, 2004a). Partial atomic charges can be also assigned using similar fitting (Sitkoff *et al.*, 1994) or can be taken from an atomistic force field (Nina *et al.*, 1997). Numerical approaches to solve the PB equation have been implemented in several software packages such as DelPhi (Klapper *et al.*, 1986; Gilson and Honig, 1987), UHBD (Davis *et al.*, 1991), APBS (Baker *et al.*, 2001), and CHARMM (Nina *et al.*, 1997; Im *et al.*, 1998).

PB-based methods have been applied not only to the modeling of aqueous protein solvation but also to modeling of lipids and proteins in a hydrated lipid bilayer environment using implicit membrane models (Murray *et al.*, 1997; BenTal *et al.*, 1996; Kessel *et al.*, 2000; Roux *et al.*, 2000; Im and Roux, 2002). In these models the interior of the membrane is modeled as a hydrocarbon-like slab with ε of 1 or 2 and the exterior as water-like with ε of 80. The contribution from the transmembrane (TM) potential originating from a small charge imbalance near the membrane–solution interfaces can be accounted for by a modified PB (PB-V) equation (Roux, 1999; Roux *et al.*, 2004).

In practice, combination of PB with MD is problematic as the PB equation would need to be solved iteratively every time step (Luo *et al.*, 2002; Lu *et al.*, 2002; Prabhu *et al.*, 2004), and may suffer from large fluctuations (Feig, 2008). However, numerical PB methods can be also useful in post-simulation analysis, for example to examine electrostatic potentials in the molecular system (Honig and Nicholls, 1995; Jo *et al.*, 2008) or estimate periodicity artifacts (Hunenberger and McCammon, 1999a; Hunenberger and McCammon, 1999b; Allen *et al.*, 2006b). The diffusional trajectories of ionic species can be also computed by PB methods using BD (Fogolari *et al.*, 2002; Roux *et al.*, 2004), which has been applied to study ion conduction through biological pores (Allen *et al.*, 2000; Im *et al.*, 2001; Im and Roux, 2002; Noskov *et al.*, 2004).

An alternative approach for the calculation of $\Delta G_{\mathrm{solv,\,elec}}$ using an implicit solvent representation is the generalized Born (GB) approximation (Still *et al.*, 1990). This phenomenological model is the generalization of the Coulomb interaction (Warshel and Papazyan, 1998) that incorporates Born solvation energetics (Born, 1920) for an atom with charge q and radius a in the medium with the dielectric constant ε, for which

$$\Delta G^{\mathrm{Born}}_{\mathrm{solv,\,elec}} = -\frac{1}{8\pi\varepsilon_0}\left(1-\frac{1}{\varepsilon}\right)\frac{q^2}{a} \tag{8}$$

In the GB model, $\Delta G_{\mathrm{solv,\,elec}}$, for collection of N atoms with charges q_i with Born radii a_i, can be written as:

$$\Delta G^{\mathrm{GB}}_{\mathrm{solv,\,elec}} = -\frac{1}{8\pi\varepsilon_0}\left(1-\frac{1}{\varepsilon}\right)\sum_{i,j=1}^{N}\frac{q_i q_j}{f_{ij}} \tag{9}$$

where f_{ij} depends on distance between atoms i and j, r_{ij} and Born radii (Still *et al.*, 1990): $f_{ij} = \sqrt{r_{ij}^2 + a_i a_j e^{-r_{ij}^2/(4a_i a_j)}}$.

The advantage of the GB method is that it is reasonably accurate and provides analytical derivatives (Feig *et al.*, 2004; Feig, 2008), thus allowing fast MD simulations. Various GB models (Schaefer and Karplus, 1996; Dominy and Brooks, 1999; Onufriev *et al.*, 2000; Lee *et al.*, 2003; Onufriev *et al.*, 2004) have been used to perform simulations of proteins in solution (Calimet *et al.*, 2001; Onufriev *et al.*, 2004; Fan *et al.*, 2005; Chocholousova and Feig, 2006; Geney *et al.*, 2006), as well as implicit membrane models (Spassov *et al.*, 2002; Im *et al.*, 2003; Tanizaki and Feig, 2005). They are typically two-dielectric membrane models with $\varepsilon = 1$ implicitly assumed inside the membrane slab and $\varepsilon = 80$ everywhere else, although multiple dielectric layer models exist (Feig, 2008). In addition to the electrostatic contribution to solvation free energy, non-polar terms are also important components of the implicit solvation or membrane models. They include such contributions as the cost of the cavity formation and solute–solvent vdW interaction term, which are typically combined into one equation: $\Delta G_{\text{solv, nonp}} = \gamma A + b$, where A is the total solvent accessible surface area (SASA) of the solute, γ and b are empirical constants with b often assigned to be 0. Values of γ and b are often determined from the fitting to some reference (experimental or calculated from all-atom simulations) solvation or partitioning free energies for nonpolar compounds (alkanes, molecular oxygen) (Leach, 2001; Feig, 2008). For implicit membrane simulations lateral forces owing to membrane deformation upon molecule insertion into a bilayer can be also incorporated through an anisotropic contribution to $\Delta G_{\text{solv, nonp}}$ (Feig, 2008).

There are other implicit solvation approaches, which can be used for protein simulations. For instance, in the Langevin dipole method (Warshel and Levitt, 1976), the solute beyond its SASA is surrounded by a three-dimensional grid of rotatable point solvent dipoles with the size and direction determined iteratively using the Langevin equation. This methodology has been also applied to membrane protein modeling (Alden *et al.*, 1995; Grossfield *et al.*, 2000).

An alternative is to include a limited number of solvent molecules around the solute and represent the rest of them with an effective solvent boundary potential. One of the first implementations of this formalism was stochastic boundary condition method (Berkowitz and Mc Cammon, 1982), in which the system is divided into three spherical regions, such as a central reaction, a buffer, and the outermost static reservoir regions. As the solute moves, the corresponding regions move through the simulation box as well. The atoms in the reservoir region are typically kept fixed but exert forces on other regions in the system. Other mixed implicit/explicit schemes have been devised over the years (see (Roux, 2001) for review). A different approach called spherical solvent boundary potential (SSBP) (Beglov and Roux, 1994) which is based on the separation of the multidimensional solute–solvent configurational integral in terms of "inner" (closest to solute) and "outer" solvent molecules. The solute and a finite number of "inner" explicit solvent molecules are confined inside the hard sphere of a variable radius (to include the most distant "inner" solvent molecules), which fluctuates during MD simulation at constant pressure (Beglov and Roux, 1994; Roux, 2001). It was shown that SSBP gives amino acid analog solvation free energies in good agreement with those obtained using more computationally expensive PBC calculations (Deng and Roux, 2004). Furthermore, ion solvation free energies are fairly insensitive to the number of explicit water molecules (Beglov and Roux, 1994) and agree well with PBC results when adjusted for the value of the interfacial potential (Lamoureux *et al.*, 2006). SSBP has been applied to simulations of various biomolecular systems (Im *et al.*, 2001). There might also some artifacts associated with artificial structure and dynamics of solvent near the boundary and the assumption that there is isotropic bulk liquid beyond the boundary sphere (Darden, 2001).

The extension of the SSBP methodology to macromolecular systems with arbitrary geometries is the generalized solvent boundary potential (GSBP) method (Im *et al.*, 2001). A small part of the biomolecular system is represented in atomic detail, whereas the influence of remaining distant atoms is modeled with an effective boundary potential. In the GSBP approach, contributions from distant atoms are represented as solvent-shielded static field and a solvent-induced reaction field (Im *et al.*, 2001; Banavali *et al.*, 2002). These terms are invariant to atom configurations in the inner region and thus are calculated only once using FD PB equation. It was found that GSBP can provide accurate estimates of electrostatic free energy contributions in biomolecular systems at relatively low computational cost (Im *et al.*, 2001; Banavali *et al.*, 2002). GSBP has been used for simulations of biological systems mainly for calculations of protein–ligand binding free energies (Banavali *et al.*, 2002; Wang *et al.*, 2006b; Deng and Roux, 2006).

8.3.2
Maintaining Ensembles

We seek to compute average properties of our system, as well as thermodynamics, from the partition function in a given ensemble. In the absence of any treatment of the interaction with the surroundings, the equations of motion used in MD will produce microcanonical (*NVE*) ensemble (i.e. maintain constant number of particles N, volume V, and energy E). Such constraints are not typically used in real-world experiments and the fixed energy represents a significant constraint on phase space. It is therefore desirable to have a computational approach to sample from the canonical, *NVT* (constant N, V, and T) or isothermal–isobaric, *NPT* (constant N, P, and T) ensembles (McQuarrie, 2000). In the canonical ensemble the configurational integral of the partition function has the form

$$Z = \int d^{3N}\mathbf{R}\, e^{-U(\mathbf{R})/kT} \tag{10}$$

This allows computation of any average property $\mathcal{X}(\mathbf{R})$ that is a function of molecular configuration $\mathbf{R}$,

$$\langle \mathcal{X} \rangle_{(NVT)} = \frac{\int d^{3N}\mathbf{R}\, \mathcal{X}(\mathbf{R})\, e^{-U(\mathbf{R})/kT}}{\int d^{3N}\mathbf{R}\, e^{-U(\mathbf{R})/kT}} \tag{11}$$

and the configurational contribution to the Helmholtz free energy (defined to within a constant)

$$A = -kTlnZ \tag{12}$$

In the *NPT* ensemble this takes the form

$$\Delta^{\text{config}} = \int dV \int d^{3N}\mathbf{R}\, e^{-U(\mathbf{R})/kT} e^{-pV/kT} \tag{13}$$

$$\langle \mathcal{X} \rangle_{\text{NPT}} = \frac{1}{\Delta^{\text{config}}} \int dV \int d^{3N}\mathbf{R}\, \mathcal{X}(\mathbf{R}) \times e^{-U(\mathbf{R})/kT} e^{-pV/kT} \tag{14}$$

(identical to $\langle \mathcal{X} \rangle_{(NVT)}$ in the limit of infinite system size) and the configurational contribution to the Gibbs free energy (defined to within a constant)

$$G = -kT \ln \Delta^{\text{config}} \tag{15}$$

A seemingly simple approach to maintain constant temperature is to perform velocity rescaling at every time step by a factor of $(T_{\text{required}}/T_{\text{current}})^{1/2}$, the so-called isothermal MD (Woodcock, 1971), excluding temperature fluctuations altogether. In a more general approach, proposed by Berendsen (Berendsen *et al.*, 1984), the system is coupled to an external heat reservoir maintained at a certain temperature. The heat flow from the reservoir is then used to maintain the temperature of the system around the required value. The rate of temperature change is determined by a coupling parameter τ. Typically, the value of τ used in atomistic MD simulations with 1- to 2-fs timestep is on the order of 0.1 ps (i.e. around every 100 time steps), which corresponds to weak coupling. If τ is equal to the time step, this method is equivalent to a simple velocity rescaling. A problem associated with these methods is that they do not generate thermodynamic averages corresponding to *NVT*, *NPT* or any other known ensemble (Leach, 2001; Frenkel and Smit, 2002). The application of the weak coupling scheme to multicomponent systems (e.g. protein in the solvated lipid bilayer) may also result in artifacts, where the temperatures of components differ substantially while preserving the desired temperature for the whole system (the so-called hot solvent–cold solute problem). One method commonly used to circumvent this problem is to use multiple thermostats for different parts of the system, but this can also introduce artifacts (Lingenheil *et al.*, 2008). In an alternative approach, the system coupling to the external heat bath is achieved through stochastic collisions which act on randomly selected particles (Andersen, 1980). Between collisions the system evolves in the *NVE* ensemble. The strength of coupling is determined by the frequency of collisions. Even though this method generates canonical distribution and provides good results for time-independent properties, dynamic properties depend on the collision frequency (Frenkel and Smit, 2002).

Instead, extended system methods, first introduced by Nosé (1984) and Hoover (1985) are more commonly used today. In this approach, a thermal reservoir is considered as an integral part of the system and treated as an additional degree of freedom with coordinate s, with conjugate momentum p_s, in the equations of motion. Therefore, these methods can be considered as an application of extended Lagrangian methodology to temperature control (Frenkel and Smit, 2002). The temperature fluctuations in the system are controlled by a coupling parameter Q, which can be considered as the fictitious mass of the heat reservoir. A small value of Q results in rapid temperature fluctuations and vice versa. The application of the Nosé–Hoover algorithm to MD simulations results in a correct canonical distribution if there is only one conserved quantity (i.e. total energy of the extended system) or the center of mass (COM) of the system remains fixed in the absence of external forces (Frenkel and Smit, 2002). For a more detailed description of different methods used for constant temperature simulations, see the textbooks (Allen and Tildesley, 1987; Leach, 2001; Frenkel and Smit, 2002 and reviews Becker and Watanabe, 2001; Hunenberger, 2005).

In many cases, simulations at constant pressure P are more representative of systems studied experimentally and can alleviate artifacts from a predefined boundary dimension. Therefore, the temperature control should be combined with a pressure control algorithm, in which case the volume becomes the fluctuating variable (instead of the pressure). In the extended system approach (Andersen, 1980), coupling of the system to a piston is introduced via a coordinate ε, with conjugate momentum p_ε, and some fictitious mass parameter W. The value of W for this barostat defines the decay time for volume fluctuations (Becker and Watanabe, 2001). This method was later extended to non-cubic simulation cells and to allow non-isotropic pressure coupling (Parrinello and Rahman, 1980; Parrinello and Rahman, 1981; Nose and Klein, 1983). The drawback of this method is that the dynamics of the system might depend on the piston mass (Feller *et al.*, 1995a). In an alternative approach, the system can be coupled to a "pressure bath," with coupling constant τ_p that is inversely proportional to the rate of the pressure change (Berendsen *et al.*, 1984). The simulation box volume is rescaled by a factor γ, which corresponds to atomic coordinate rescaling by $\gamma^{1/3}$ for isotropic coupling. As for the temperature control weak coupling scheme, this method does not produce trajectories corresponding to any known ensemble (Allen and Tildesley, 1987; Becker and Watanabe, 2001), but allows for the damping of large volume fluctuations and is well suited for the equilibration stage of MD runs (Anezo *et al.*, 2003). The Langevin piston method developed by Feller *et al.* (1995a) resolves difficulties associated with the extended system and weak coupling methods of pressure control. In this method, the motion of the piston is described by a Langevin equation (Eq. (5)) with a collision frequency γ, eliminating the unphysical volume fluctuations related to the piston mass (Feller *et al.*, 1995b).

Sometimes it is necessary to use anisotropic pressure coupling for protein MD simulations, especially for those involving interfacial systems (for example proteins associated with hydrated lipid bilayers), possibly involving a surface tension γ (not to be confused with the coupling constants described earlier) at the interface (Zhang *et al.*, 1995). γ for a system with interface oriented with its normal in the z direction can be calculated from the diagonal components of the pressure tensor as

$$\gamma = 0.5 < L_z[P_{zz} - 0.5(P_{xx} + P_{yy})] \tag{16}$$

where L_z is the height of the cell, and the first factor 0.5 takes into account the presence of two interfaces in a simulation cell (Zhang *et al.*, 1995; Klauda *et al.*, 2008). For interfacial systems MD runs using constant normal pressure P_N and fixed surface area A (NP_NAT ensemble) or constant surface tension γ ($NP_N\gamma T$) are commonly used (Zhang *et al.*, 1995; Feller *et al.*, 1995b; Feller *et al.*, 1997; Feller and Pastor, 1999; Klauda *et al.*, 2008). Membrane-protein simulations using $NP_N\gamma T$ ensemble are complicated by the uncertainty in the appropriate reference value of γ for lipid bilayers (Feller and Pastor, 1999; Klauda *et al.*, 2008), leading to the frequent use of NP_NAT simulations instead. The area is set on the basis of the experimental surface area per lipid and the estimated projected surface area of the membrane proteins or peptide (Woolf and Roux, 1994; Berneche *et al.*, 1998).

8.3.3
Approximating Long-Ranged Forces

The majority of computer time in MD simulations of biomolecular systems is spent on the calculation of nonbonded forces, with the required time scaling as the square of the number of atoms in the system, $O(N^2)$ for a pairwise-additive potential energy function. If PBC are used and interactions between atoms in the primary cell and all image cells are counted, the number of nonbonded terms, which needs to be calculated becomes infinite. Therefore the minimum image convention is used, where the nonbonded potential is truncated to include interactions with the nearest images only, thus including only $(1/2)N(N-1)$ terms (Allen and Tildesley, 1987). However, even this approximation can be prohibitively expensive for large biomolecular systems comprising of tens of thousands of atoms. To reduce this computational cost, spherical cutoffs can be used, where the interactions between pairs of atoms further than some cutoff distance r_c are set to 0. The cutoff distance should be large enough to ensure a small effect owing to truncation, but not larger than half of the box length so as to maintain the minimum image convention. Cutoffs on the order of 12 Å are typically used. To determine, which atoms are located within this cutoff, nonbonded neighbor lists (Verlet, 1967) are generated. These lists are updated at regular intervals (e.g. every 10–100 steps) or using some heuristic algorithm to test for the movement of atoms in the system (Frenkel and Smit, 2002). For very large systems (with $L \gg r_c$) the cell index method for the nonbonded list updates can be used, where the simulation box is divided into cells with length longer than r_c (Leach, 2001; Frenkel and Smit, 2002). To expedite the handling of nonbonded interactions it is sometimes advantageous to use group-based cutoffs, where a molecule is divided into one or several groups consisting of a few atoms. These groups are often made electroneutral if possible, which eliminates long-range charge–charge and charge–dipole interactions (with energy varying as $1/r$ and $1/r^2$, respectively) between these groups and only dipole–dipole (energy varying as $1/r^3$) and higher-order terms surviving. It has been shown, however, that the use of group-based cutoffs can also lead to artifacts such as free rotation of polar groups just outside the cutoff distance (Darden, 2001; Steinbach and Brooks, 1994).

A common problem with both atom- and group-based cutoffs is that they produce discontinuities in both energies and forces at the cutoff boundary, r_c (Leach, 2001). To eliminate this problem, the potential can be modified to gradually reduce energies and/or forces to 0 at r_c. One approach is to shift the potential by adding a term, such that $U(r_c) = 0$. Alternatively, the force maybe shifted by adding a term to the potential to ensure $dU(r_c)/dr = 0$. The shifted potential or force is therefore modified even for $r < r_c$, with artifacts. An alternative is to apply a switching function, $S(r)$ (varying from 1 to 0), that can be used to gradually move the potential from $U(r_s)$ at some switching distance, r_s ($r_s < r_c$, to 0 at r_c). The modified potential then has the form $U'(r) = U(r)S(r)$. Typically in protein MD simulations, the value of r_s is set to be ~ 2 Å smaller than r_c to avoid rapid variation in force over short distances (Steinbach and Brooks, 1994). In the force switching method, the force (instead of the potential) is switched and the potential energy is then obtained by integrating the switched force between r_s and r_c with the constant term added to potential at $r < r_s$ to ensure its continuity at r_s, and was shown to lead to stable MD simulations (Steinbach and Brooks, 1994).

The use of finite cutoffs necessarily leads to the omission of some nonbonded interactions that may affect the accuracy of simulation results. The inclusion of contributions beyond cutoffs can be achieved in several ways. One of the simplest methods is to use two nonbond cutoffs: a "short" cutoff (e.g. 10 Å) where interactions are computed every time step, and a "long" cutoff (e.g. 30 Å) where interactions are computed more seldomly (such as at nonbond list update) It has been demonstrated that such twin-range cutoffs perform well for membrane simulations (Anezo *et al.*, 2003). Long-range corrections to the energy and pressure can then be estimated assuming a uniform distribution of atoms for $r > r_c$ (Allen and Tildesley, 1987), applied during or after simulations (Lague *et al.*, 2004) These corrections are relatively small (but still nonnegligible) for LJ interactions since they are short-ranged ($\sim 1/r^6$). However, for electrostatic interactions, which decay as $1/r$, these corrections are more significant. With such long-ranged interactions the use of any cutoff-based scheme and/or minimum image convention can cause artifacts (Allen and Tildesley, 1987; Feller *et al.*, 1996), which have been observed in simulations of biomolecular systems leading to solvent ordering (Anezo *et al.*, 2003).

One approach to circumvent the artifacts of truncated long-ranged interactions is to employ lattice methods, such as Ewald summation (Ewald, 1921), widely used in MD simulations today. In this approach, electrostatic interactions of a charged atom with all of its periodic images are calculated. The sum of these interactions is conditionally convergent, yet when converted into a sum of two series, convergence can be achieved (Leach, 2001). Each charge is surrounded by the neutralizing (i.e. of the same magnitude and opposite sign) Gaussian distribution $\rho_i(r) = q_i\kappa^3 \exp(-\kappa^2 r^2)/\pi^{3/2}$ where r is the distance from the center of this distribution and the parameter κ determines the width of this distribution. The interactions between the charges, plus their neutralizing distributions, (called the *real-space sum*) are now short-ranged and converge rapidly with distance (Allen and Tildesley, 1987). To obtain only the potential owing to the original charges, a second, canceling distribution is then added, with contributions summed, for simplicity in reciprocal space. A correction term, which accounts for the interaction of the neutralizing distribution with itself, is also added (Allen and Tildesley, 1987). The Ewald method is computationally expensive and scales as poorly as N^2 (Darden, 2001). To speed up calculations of the computationally demanding reciprocal sum term, several methods have been proposed. In the particle-mesh Ewald (PME) (Darden *et al.*, 1993; Essmann *et al.*, 1995) and related methods (Hockney and Eastwood, 1988; Luty *et al.*, 1994; Deserno and Holm, 1998), the point charges are interpolated to grid points (e.g. using B-spline interpolation), which is followed by the calculation of reciprocal-space sum using a fast Fourier transform (FFT). These methods scale as $N \log N$ and subsequently are used in most simulations of protein systems today (MacKerell, 2004).

Ewald summation also has artifacts originating from the imposed PBC, which become more pronounced for systems with low-dielectric solvent, small system size, nonzero net charge or large dipole (Hummer *et al.*, 1996; Smith and Pettitt, 1996; Darden, 2001; Anezo *et al.*, 2003). One such artifact is the nearly zero net force between charges separated by half of the box length (Smith and Pettitt, 1996; Leach, 2001). For periodic systems with nonzero net charge, a uniform charge distribution with the same magnitude and opposite sign (“neutralizing plasma”) is added to avoid divergence (Hummer *et al.*, 1996; Bogusz *et al.*, 1998; Darden, 2001). However, spurious interaction of atomic charges with this neutralizing distribution leads to artifacts in both energies and forces, which need to be corrected (Bogusz *et al.*, 1998). For solvated systems with net charge, there is also an artifact associated with perturbation of solvent molecules by periodic copies of a charged solute (Hunenberger and McCammon, 1999b). Both of these artifacts might be quite large (several kilocalories per mole) but often substantially cancel each other for polar systems (Bogusz *et al.*, 1998; Hunenberger and McCammon, 1999b; Warren and Patel, 2007) and can be estimated using continuum electrostatics (Hunenberger and McCammon, 1999a; Hunenberger and McCammon, 1999b).

An alternative approach for the treatment of long-range electrostatic interactions is the reaction field (RF) method (Allen and Tildesley, 1987; Tironi *et al.*, 1995; Leach, 2001). In this approach, interactions between atoms within some cutoff are treated explicitly, whereas interactions with those outside are modeled approximately via a continuous dielectric continuum. (Barker and Watts, 1973). While this might be a reasonable approximation for bulk liquids, it may fail for many heterogeneous biomolecular systems. Another approach is to treat distant interactions approximately using a multipolar expansion of the electrostatic potential (Darden, 2001). In cell multipole methods the simulation space is divided into cubic cells, for which multipole moments (charge, dipole, and quadrupole) are calculated. Then the interactions of atoms in a given cell with all adjacent cells are calculated explicitly, but with atoms in nonadjacent cells approximated using a multipolar expansion (Leach, 2001; Darden, 2001). With more distant atoms grouped into larger cells this method can be fairly efficient without substantial loss in the accuracy

(Leach, 2001; Darden, 2001), which inspired the development of more efficient $O(N)$ tree-based methods such as fast multipole algorithm (FMA) (Greengard and Rokhlin, 1987).

8.4
Calculation of Observables for Protein Structure and Function

MD trajectories maybe utilized in different ways: (i) to study the time evolution of structure, dynamics, interactions and correlations of a thermally fluctuating system and visualize microscopic phenomena; (ii) to sample phase space for the calculation of ensemble averaged equilibrium observables and the thermodynamics governing microscopic processes; and (iii) to force the system to ensure sampling or observation of a process of interest that may otherwise not occur in a reasonable simulation timeframe. Here we summarize some of the basic approaches used for these three kinds of studies.

8.4.1
Time Series and Average Properties

MD trajectories contain snapshots of the molecular system, coordinates (and velocities), sampled with a predefined sampling interval, $\Delta\tau$, typically much larger than the time step Δt. The optimal value of $\Delta\tau$ is determined on the basis of the timescale of the phenomenon under investigation. In atomistic protein simulations sampling is typically every 100 –1000 steps (i.e. $\Delta\tau$ are 0.2–2 ps using Δt of 2 fs). Calculations of some time-dependent property $\mathcal{X}$ that is a function of system coordinates and/or velocities for frames of the trajectory produce a time series $\mathcal{X}(t)$. Energy components, volume, temperature, pressure, and so on, are typically reported by a MD simulation program during the run. Time series using instantaneous values of $\mathcal{X}(t)$ tend to be noisy but can be smoothed by using block or running averages. Time series of energy, volume, IE between system components are useful in monitoring convergence of MD simulations. The root mean square deviation or displacement (RMSD) from some reference or initial structure can also be used for such monitoring (Biggin and Bond, 2008; Lindahl and Sansom, 2008). The RMSD is defined as

$$\mathrm{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{r}_{i,(t)} - \mathbf{r}_{i,0}\right)^2} \tag{17}$$

where $\mathbf{r}_i(t)$ and $\mathbf{r}_{i,0}$ are coordinates of atom i for current and reference structure, preoriented by some prescription (Leach, 2001). The RMSD for different parts of the molecular system, for example C_α atoms of the protein (Biggin and Bond, 2008; Lei and Duan, 2008), can be calculated to monitor changes in protein structure and simulation convergence. In principle any scalar, vector, or tensor quantity maybe expressed as a time series from MD simulation. Once this time series is obtained it can be used for statistical analysis.

8.4.1.1 Averages and Correlations

The statistical analysis of MD simulation data is also an important part of data processing. It may involve calculation of averages, fluctuations, variances and covariances, higher order moments, and correlation functions. Some of these properties are briefly reviewed here, a more general and comprehensive description can be found elsewhere (Allen and Tildesley, 1987; Leach, 2001; Becker and Watanabe, 2001; Frenkel and Smit, 2002).

Time averages of $\mathcal{X}(t)$ can be obtained as

$$\bar{\mathcal{X}} = \frac{1}{\tau}\sum_{t=1}^{\tau}\mathcal{X}(t) \tag{18}$$

where τ is the simulation length or number of frames in the MD run (Allen and Tildesley, 1987; Frenkel and Smit, 2002). It is assumed that for sufficiently long MD run the time average of some property of the system is the same as its average over the statistical ensemble if the system is ergodic ($\langle\mathcal{X}\rangle = \bar{\mathcal{X}}$) (Frenkel and Smit, 2002). Therefore we can use all principles of ensembles in statistical mechanics on the basis of time averages.

An ensemble average of a quantity $\mathcal{X}$ which is function of atomic coordinates can be computed using Eq. (14) in the NPT ensemble. Such mean values can be structural (e.g. RDFs, correlations, NMR order parameters), dynamical (e.g. translational and rotational diffusion), or mechanical (e.g. elastic moduli) and each adds to our understanding of the origins of microscopic behavior. It is often necessary to look not at just single average values of some properties from MD simulations but at their distributions, binned into histograms. For example, the RDF, or $g_{ij}(r)$ is a pair distribution function, which describes probability of finding a particle i (atom, set of atoms, molecule, etc.) a distance r from another particle compared to the distribution in the homogeneous system with the same density (Allen and Tildesley, 1987).

$$g_{\alpha\beta}(r) = \frac{N_\beta(r, r+\delta r)}{4\pi r^2 \delta r \rho_\beta} \tag{19}$$

where $N_\beta(r, r+\delta r)$ is the number of atoms of species β between a distance r and $r+\delta r$ around an atom of species α with bulk reference density ρ_β. For neat liquids, $g(r)$ is a useful source of structural information and can be compared with experimental diffraction data (Leach, 2001). The integration of the unnormalized $g_{ij}(r)$ from 0 to r provides number of coordinating species inside a sphere with radius r. Coordination numbers can be calculated at the position of the first minimum in the $g(r)$ to obtain the number of species inside the first solvation shell (Li *et al.*, 2008a). Thus these numbers provide valuable structural information about protein surroundings. Moreover, by integration of the product of $g_{ij}(r)$ and another pair function $a_{ij}(r)$ the ensemble average $\langle a_{ij}\rangle$ is obtained, which can be used for calculating thermodynamic quantities such energy, pressure, and chemical potential (Allen and Tildesley, 1987).

To shed more light on function, it is usually essential to analyze interactions between different parts of the system not only in terms of distributions (e.g. RDFs), but also in terms of IEs (interaction energies) which offer a more quantitative description of interactions and their contributions to the governing thermodynamics of a process. An IE can be calculated simply as the difference in energies of a combined system (complex) and two interacting species, or simply as a sum of interactions between atomic paris. Statistical analysis of IEs can be used to identify the most relevant interactions to protein structure and function. If one is interested in a measure of "packing" of solvent around a solute or protein, then it maybe preferable to turn to measures other than energies that are less chemically specific. For instance, packing scores can be defined as (Grossfield *et al.*, 2006)

$$s = \sum_i \sum_j \frac{1}{r_{ij}^6} \tag{20}$$

(analogous to the attractive part of the LJ potential and related to the rate of magnetization transfer in NMR) and have been used to explore locations of specific protein–lipid interactions (Grossfield *et al.*, 2006).

Another important quantity that can reveal much about the interplay between components of a biological system is the covariance $\langle \delta A \delta B \rangle$ where δA, δB are fluctuations (e.g. $\delta A = A- < A >$). For instance, in a recent study of the signal-transduction mechanism in rhodopsin, the covariance between IEs of two pairs of protein residues has been calculated to determine how tightly they are coupled (Kong and Karplus, 2007). Time correlation functions between properties A and B are defined as (Allen and Tildesley, 1987; Leach, 2001; Becker and Watanabe, 2001)

$$C_{AB}(t) = \langle A(0)B(t) \rangle = \lim_{T\to\infty} \frac{1}{T-t} \times \int_0^{T-t} d\tau A(\tau)B(t+\tau) \tag{21}$$

If $A = B$ then quantity is called the *autocorrelation function* which reveals how a quantity is related to itself at some point later in time. The integral of the autocorrelation function $C_A(t)$ from 0 to ∞ is called *correlation* or relaxation time (Allen and Tildesley, 1987) and can be viewed as time taken to lose correlation (Leach, 2001). For instance, rotational correlation for water molecules can be determined by analyzing the decay of the dipole autocorrelation (Allen *et al.*, 1999) as a measure of rotational diffusion. One may also express correlations in space (as opposed to time) to determine a correlation length between two propertieis (e.g. the positions or orientations of molecules in space). The MD simulation time should be much longer than the correlation time of the slowest process of interest in the system, whereas the size of the system greater than the correlation length.

8.4.1.2 **Protein and Solvent Fluctuations**

Besides RMSD, root mean square fluctuations (RMSF) of protein residue coordinates with respect to their time-averaged values can also be calculated as

$$\mathrm{RMSF} = \sqrt{\frac{1}{\tau_{\mathrm{run}}} \sum_{t=1}^{\tau_{\mathrm{run}}} (\mathbf{r}(t) - \langle \mathbf{r}(t) \rangle_t)^2} \tag{22}$$

RMSF describe protein flexibility and can be related to 3D thermal (B) factors ($B = \frac{8\pi^2}{3} RMSF^2$) (Willis and Pryor, 1975) measured by X-ray crystallography.

The protein structure fluctuations are often hard to interpret owing to the complexity of motions in a large biomolecular system. However, it is possible to discern major collective modes of fluctuations using quasi-normal mode or principal component analysis (PCA) (Garcia, 1992; Amadei *et al.*, 1993; Kitao *et al.*, 1991). This method is based on the diagonalization of the covariance matrix of protein atom fluctuations. The principal modes with the largest eigenvalues corresponding to their contribution to the total fluctuation are identified, and the resultant eigenvectors for protein motions can be determined (Hess, 2000; Hayward and de Groot, 2008).The principal modes with the largest amplitude typically represent the slowest protein motions. In many cases they can be linked to function (Hayward and de Groot, 2008). Thus protein dynamics in the subspace of these modes is termed *essential dynamics* (Amadei *et al.*, 1993). The results of such analysis can be used for advanced sampling techniques such as conformational flooding (Grubmuller, 1995) and essential dynamics sampling (Amadei *et al.*, 1996). Thus, the PCA is a powerful tool for both analysis and simulations of large collective motions in proteins (Kitao and Go, 1999; Hayward and de Groot, 2008). It should be noted, however, that the timescale of modern MD simulations might be not long enough to sample these modes properly (Hess, 2000).

To examine the dynamics of solvent, the 3D diffusion coefficient, D, can be computed by mean square displacement (MSD) analysis via the relation

$$D = \lim_{t\to\infty} \frac{1}{6t} \left\langle |\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2 \right\rangle \qquad (23)$$

which can be derived from the Langevin equation of Eq. (5), in the absence of a systematic force, F, and provides a prescription for the friction coefficient, γ and thus diffusion, $D = kT/m\gamma$ (Allen and Tildesley, 1987). If PBC are used, the effect of image centering and changes in the box dimensions need to be removed from trajectories to calculate the diffusion coefficient using Eq. (23). Moreover, calculation of D from 23 does not normally come from analysis of a single trajectory, but of all trajectories with every possible origin of time, by sliding an analysis window along the MD trajectory (Rapaport, 2004). One may also extract the following relationship from the Langevin equation in terms of the velocity (McQuarrie, 2000)

$$D = \frac{1}{3} \int_0^\infty \langle \boldsymbol{v}(0) \cdot \boldsymbol{v}(t) \rangle dt \qquad (24)$$

(in simplified notation for a correlation integral such as in Eq. (21))
where the diffusion coefficient comes from the velocity autocorrelation. One may also extract diffusion from the fluctuation–dissipation theorem by analyzing the force autocorrelation

$$D = \frac{3(kT)^2}{\displaystyle\int_0^\infty \langle \mathbf{R}(0)\mathbf{R}(t) \rangle dt} \qquad (25)$$

One concern when computing diffusion is the need to separate random and systematic forces, as Eqs. (23) and (24) were derived in the absence of systematic forces and Eq. (25) reveals that it is only the random contribution to the force that determines D. In some cases, incorrect values of D may be obtained. For example, if a particle is trapped in a deep energy well it is still diffusing despite its RMSD being small. One way to overcome this is to "clamp" the particle and measure the random forces at that fixed position using Eq. (25) (Mamonov *et al.*, 2006). Another, more theoretically complicated approach, is based on analysis of the Laplace transform of the velocity autocorrelation function on the basis of the generalized Langevin equation for a harmonic oscillator (Crouzy *et al.*, 1994). This approach is very useful for analysis of umbrella sampling biased trajectories to be described below. The following function is evaluated

$$D(z) = \lim_{s\to 0} \frac{-\hat{C}(s;z)\langle\delta z^2\rangle_{(i)}\langle\dot{z}^2\rangle_{(i)}}{\hat{C}(s;z)\left[s\langle\delta z^2\rangle_{(i)} + \langle\dot{z}^2\rangle_{(i)}/s\right] - \langle\delta z^2\rangle_{(i)}\langle\dot{z}^2\rangle_{(i)}} \tag{26}$$

where $C(t;z)$ is the velocity autocorrelation function for position z, and $\hat{C}(s;z)$ is its Laplace transform. To estimate the value of the limit as $s \to 0$, linear extrapolation is needed owing to a divergence. However, the above analysis is in fact equivalent to a real-space analysis derived by Hummer and coworkers that circumvents such extrapolation (Yeh and Hummer, 2004). We now have methods to extract measure of diffusion that are needed for calculations of rates of translocation, such as through aqueous solution and ion channels (Section 8.5.1).

8.4.2
Free Energy Calculation

A quantitative calculation of free energies is very important problem in computational biology (Brooks *et al.*, 1988; Kollman, 1993; Simonson, 2001; Meirovitch, 2007). Here we describe two main approaches to free energy calculation, potential of mean force (PMF), and alchemical mutation, and then venture into some more detailed approaches to extract microscopic mechanisms of protein function from MD simulations.

8.4.2.1 Potential of Mean Force Calculation

Many studies simulate for long periods and hope to observe a process of interest. This hit and miss approach does not make the best use of computational resources, nor it is reliable as outcomes can depend on initial configurations. The obvious difficulty of all-atom MD is that the trajectory of a large number of atoms must be computed, even though only a few coordinates maybe truly relevant to function. Rather than trying to observe rare events with "brute force" simulation, a better approach is to calculate a free energy landscape, characterized by identified "relevant" coordinates. One may project out the uninteresting degrees of freedom (i.e. integrate over them) to study the most relevant coordinates (Zwanzig, 2001; Mori, 1965), under the action of a mean force that is an average over the thermal fluctuations of the other $3N - 1$ degrees of freedom, $\mathbf{R}$ (Roux *et al.*, 2004). The PMF, W, is the thermodynamic reversible work along a chosen coordinate z, relative to some reference value z':

$$W(z) = -\int_{z'}^{z} d\zeta\, \langle F(\zeta)\rangle \tag{27}$$

or, in differential form the mean force is

$$\langle F(z)\rangle = -\frac{dW(z)}{dz} \tag{28}$$

where $F(z)$ is the instantaneous force acting on the chosen coordinate, z, with integration variable, ζ. One can easily show that the derivative of the following expression, that relates the free energy to a configurational integral, leads to the same mean force

$$W(z) = -kT\ln \times \left(\int d^{3N-1}\mathbf{R}\, e^{-U(z,\mathbf{R})/kT}\right) + \mathrm{C} \tag{29}$$

which by setting the free energy at some reference point, z' to zero the constant C maybe evaluated such that

$$W(z) = -kT \ln \left(\frac{\int d^{3N-1}\mathbf{R}\, e^{-U(z,\mathbf{R})/kT}}{\int d^{3N-1}\mathbf{R}\, e^{-U(z',\mathbf{R})/kT}} \right)$$

$$= -kT \ln \frac{\langle \rho(z) \rangle}{\langle \rho(z') \rangle} \qquad (30)$$

This ratio of configurational integrals therefore also determines the PMF. The ratio of configurational integrals also equals the relative density at z and z' and thus the equilibrium constant for the process of changing that degree of freedom from a value z' to a value z. Thus, by computing average density along the reaction coordinate one knows the free energy change, and thus the reversible work involved in moving along the reaction coordinate. PMF calculations entail simulating long enough so that all interesting values of the relevant coordinates are encountered, and an equilibrium distribution of uninteresting coordinates are integrated over. However, in many cases it is difficult to sample all values of the relevant coordinates, simply because they correspond to the rate limiting steps and thus usually involve significant free energy barriers; not thermally accessible in a manageable simulation timeframe. To assist, one may apply some biasing potentials to ensure sampling of all interesting values of z. For example, in the method of Umbrella Sampling (Torrie and Valleau, 1977), a series of biasing potentials, as simple as harmonics, $w_i(z) = \frac{1}{2}K_i\,(z - z_i)^2$, maybe added to the system potential, $U(z, \mathbf{R})$, in a series of separate simulations/ "windows," i. Because it is known how the coordinate, z, was biased, it is possible to remove this bias postsimulation. Trajectories may be combined and rigorously unbiased with the weighted histogram analysis method (WHAM) (Kumar *et al.*, 1992). In this method one uses the biased histograms to solve for the optimal estimate for the unbiased density, ρ, and thus the PMF using Eq. (30) above.

$$\langle \rho(z) \rangle = \frac{\sum_{i=1}^{N} n_i \,\langle \rho(z) \rangle_{(i)}^{\text{bias}}}{\sum_{i=1}^{N} n_i e^{-[w_i(z)-F_i]/kT}} \qquad (31)$$

where N is the number of WHAM windows and n_i is the number of independent data points. Free energy constants F_i can be determined using the optimal estimate for the distribution function:

$$e^{-F_i/kT} = \int dz\, e^{-w_i(z)/kT} \,\langle \rho(z) \rangle \qquad (32)$$

Alternatively one may simply compute the mean force from these simulations directly, or simply use the magnitude of the constraint force (which is equal and opposite to the instantaneous force on the reaction coordinate) to obtain the PMF via integration.

8.4.2.2 **Free Energy Perturbation**

Free energy perturbation (FEP) calculations are popular in biomolecular simulation because of their ability to explore reaction coordinates other than spatial coordinates. In particular, alchemical transformations such as single site mutations of key residues associated with protein folding and function

(Funahashi *et al.*, 2003; Park and Lee, 2005; Pan and Daggett, 2001; Wang *et al.*, 1998), ion mutations to explore ion selectivity of ion channels (Allen *et al.*, 1999; Allen *et al.*, 2000; Luzhkov and Aqvist, 2001; Berneche and Roux, 2002; Noskov *et al.*, 2005), and protonation states of titratable groups in proteins (Merz, 1991; Simonson *et al.*, 2004; Yoo and Cui, 2008) can be sampled at equilibrium. The procedure involves introducing a thermodynamic coupling parameter λ, varying from 0 to 1, to transform between an unperturbed, $U(\lambda = 0)$, and some perturbed, $U(\lambda = 1)$, potential energy function of the system (Kollman, 1993)

$$U(\lambda) = \lambda U(\lambda = 1) + (1 - \lambda) U(\lambda = 0) \tag{33}$$

The free energy difference, ΔG, can be computed via free energy perturbation (Zwanzig, 1954; Kollman, 1993)

$$\Delta G = -kT \times \sum_{\lambda=0}^{1} \ln \left\langle \exp\left[-\frac{U(\lambda + \Delta\lambda) - U(\lambda)}{kT} \right] \right\rangle_{(\lambda)} \tag{34}$$

or, equivalently, via thermodynamic integration (TI) (Kirkwood, 1935)

$$\Delta G = \int_0^1 d\lambda \left\langle \frac{\partial U}{\partial \lambda} \right\rangle_{(\lambda)} \tag{35}$$

where the subscript (λ) signifies that ensemble averaging has been carried out for that particular value of the scaling parameter. In reality a finite number of scaling parameters are used. However, if the magnitude of the perturbation is great then the distribution of configurations at one value of λ and the next will be very different. To ensure good convergence of the free energy, one should typically ensure that entropic changes remain within a couple of kT units between windows (Lu *et al.*, 2003) One may also employ non-linear scaling to avoid rapid changes at a given λ value (Shirts *et al.*, 2003). Analysis of FEP simulations may also be improved through a WHAM analysis since the free energy contribution for a given value of λ is computed using all information from the simulation in an optimal way (Souaille and Roux, 2001).

A problem may occur in solvation free energy calculations that arises from the linear coupling scheme for the vdW energy term upon "particle annihilation." At the end point of the FEP simulation as, $\lambda \to 0$, there is a singularity owing to the repulsive portion of LJ potential (Simonson, 2001; Deng and Roux, 2004). This problem can be alleviated by using nonlinear scaling of vdW interactions (Simonson, 1993; Simonson, 2001). An efficient way to solve this problem was proposed in the staged protocol developed by Deng and Roux (2004) It is based on the Weeks, Chandler, and Andersen scheme to separate repulsive and attractive parts of the LJ potential (Weeks *et al.*, 1971). Then the linear coupling scheme is used for a slowly varying attractive LJ contribution, whereas the repulsive part is gradually transformed into a soft core potential controlled by a nonlinear staging parameter varying from 0 to 1 (Deng and Roux, 2004). This approach can be efficiently used for solvation and protein-binding free energy calculations (Deng and Roux, 2004; Deng and Roux, 2006).

8.4.2.3 **Energetic Decompositions**

It is often of interest to determine how much a particular subset of atoms in the system, or a particular energetic term, contributes to the free energy. The simplest contributions to be calculated are just energetic contributions. One may compute a contribution from a particular interaction between some groups of atoms, or perhaps a type of interaction (e.g. electrostatic or LJ) as a function of reaction coordinate.

$$\langle U_\alpha(\mathbf{R}, z)\rangle (z) = \frac{\displaystyle\int d^{3N-1}\mathbf{R}\; U_\alpha(\mathbf{R}, z) e^{-U(\mathbf{R},z)/kT}}{\displaystyle\int d^{3N-1}\mathbf{R}\; e^{-U(\mathbf{R},z)/kT}} \quad (36)$$

One may also wish to divide energies up into some meaningful contributions for a given problem. For example, if the reaction coordinate is the position of a solute moving through a system, one may wish to divide the energy into "direct" interactions with the solute and "indirect" interactions that change as the solute passes (e.g. strains on the system). For example,

$$\langle U_I(\mathbf{R})\rangle (z) = \frac{\displaystyle\int d^{3N-1}\mathbf{R}\; U_I(\mathbf{R}) e^{-[U_D(z,\mathbf{R})+U_I(\mathbf{R})]/kT}}{\displaystyle\int d^{3N-1}\mathbf{R}\; e^{-[U_D(z,\mathbf{R})+U_I(\mathbf{R})]/kT}} \quad (37)$$

which shows that all energy terms affect the mean contribution from any one, via the Boltzmann factor in the configurational integral. IEs maybe computed through an IE calculation between two subsets of atoms, possibly with some truncation of nonbonded terms, as explained previously. However, potential energy contributions do not tell the entire story.

One may also calculate entropic contributions to the PMF. The usual practice for extracting the entropy is to analyze temperature variations in free energy

$$S = -\frac{dG}{dT} \quad (38)$$

but which requires simulations at multiple temperatures (Maccallum and Tieleman, 2003), followed by FD calculation. A simple alternative is to estimate the entropic contributions at a single temperature via

$$-T\Delta S = \Delta G - \Delta H \quad (39)$$

where the enthalpic contribution includes the average internal energy of the system as well as the pressure–volume work. This approach is simple but does suffer the limitation that large fluctuations in the energy are transferred to the resulting entropic estimate.

In many instances, entropic contributions can be as significant as enthalpic contributions to a free energy change. One would really prefer to have a means of decomposing the free energy into contributions from particular interactions or energetic terms. The relationship between the PMF and the mean force (Eq. (28)) permits a decomposition from different interactions, α

$$W_\alpha(z) = W_\alpha(z') - \int_{z'}^{z} d\zeta \langle F_\alpha(\zeta)\rangle \quad (40)$$

where F_α is the instantaneous force of that component. This mean force decomposition is meaningful because forces acting on the protein are additive and related to free energy by simple integration.

8.4.2.4 Multidimensional PMFs

Often a process is governed by more than one coordinate, The objective then is to compute the function, $W(z, s)$ as a function of a secondary coordinate, s, also.

$$W(z, s) = -kT \ln\langle \rho(z, s)\rangle \tag{41}$$

A similar expression could be written for more than two coordinates to create a multidimensional free energy surface. In principle, the distribution of many coordinates could be obtained from unbiased simulation. However, when the time to equilibrate one or more of those degrees of freedom exceeds manageable simulation times one may wish to bias the sampling of more than one coordinate.

In the case where one coordinate has already been biased, calculation of the 2D free energy surface is simple. Once the WHAM equations (Eqs. (31) and (32)) have been iterated to convergence for the 1D PMF calculation, any distribution $\langle \rho(z, s)\rangle_{(i)}^{\text{bias}}$, biased by window function $w_i(z)$ acting on the z coordinate, but involving some secondary variable s, may be unbiased via an extension of the 1D WHAM equations (Allen *et al.*, 2006a),

$$\langle \rho(z, s)\rangle = \frac{\displaystyle\sum_{i=1}^{N} n_i \langle \rho(z, s)\rangle_{(i)}^{\text{bias}}}{\displaystyle\sum_{i=1}^{N} n_i e^{-[w_i(z)-F_i]/kT}} \tag{42}$$

where n_i is the number of time steps in window i, and F_i are the free energy constants obtained from solution to the 1D WHAM equations. However, this unbiasing procedure can be avoided. Defining the function $\rho(s|z)$ by

$$\langle \rho(z, s)\rangle_{(i)}^{\text{bias}} = \langle \rho(z)\rangle_{(i)}^{\text{bias}} \cdot \rho(s|z) \tag{43}$$

where the conditional probability $\rho(s|z)$ (probability of having a value of s, given a value of z) is normalized, then the unbiased distributions, related to the biased distributions via Eq. (42), possess a similar relation

$$\langle \rho(z, s)\rangle = \langle \rho(z)\rangle \cdot \rho(s|z) \tag{44}$$

The 2D PMF, $W(z, s)$, is then simply obtained from $W(z)$ by

$$\begin{aligned} W(z, s) &= -kT \ln \langle \rho(z, s)\rangle \\ &= W(z) - kT \ln \langle \rho(s|z)\rangle + C \end{aligned} \tag{45}$$

where C is some additive constant. However, if the second variable also requires biasing then a full 2D WHAM analysis of 2D umbrella sampling windows is required (Souaille and Roux, 2001).

Having obtained a multidimensional free energy surface it is often interesting to extract the free energy along any one coordinate; where the other coordinates are sampled in their equilibrium distributions. To do this, one may integrate over the secondary variable, s, to reveal the 1D PMF, $W(z)$,

$$e^{-W(z)/kT} = C \int ds\, e^{-W(z,s)/kT} \tag{46}$$

In some cases, the definition of the 1D PMF from a 2D PMF can be problematic, if the secondary coordinate becomes unbound for any range of z. Such a situation emerges from analysis of ion permeation through ion channels, for example, where an ion leaving the channel becomes unbounded and experiences an infinite entropic increase (Roux *et al.,* 2008). In such cases of binding free energy profiles one must place a limitation on the bounds of the secondary coordinate that maybe easily sampled during a simulation (i.e. the lateral displacements of the ion exiting the ion channel). This constraint must then be rigorously unbiased. Because the configurational integral of Eq. (46) is not defined as the ion becomes unbounded (outside the channel) we can restrict lateral displacements with a flat-bottom cylindrical constraint of radius R. We can represent this by a Heaviside step function $H(x, y)$ which is 1 inside the bounding cylinder and 0 outside, such that (Allen *et al.,* 2006a)

$$e^{-W(z)/k_\mathrm{B}T} = C \int dx\, dy\, H(x, y)\, e^{-W(x,y,z)/kT} \tag{47}$$

where C is some constant. By setting $W(z') = 0$ at the position $z = z'$ far from the protein, C becomes

$$C = \frac{1}{\displaystyle\int dx\, dy\, H(x, y)\, e^{-W(x,y,z')/kT}} \tag{48}$$

If W is independent of x and y at $z = z'$, then $W(x, y, z') = W(0, 0, z')$ and

$$C = \frac{e^{+W(0,0,z')/kT}}{\pi R^2} \tag{49}$$

such that (Allen *et al.,* 2004b)

$$e^{-W(z)/k_\mathrm{B}T} = \frac{1}{\pi R^2} \int dx\, dy\, H(x, y) \times e^{-[W(x,y,z)-W(0,0,z')]/kT} \tag{50}$$

The multidimensional PMF can also be used for extracting average properties $\mathcal{X}(z, s)$ or $\mathcal{X}(z)$ through similar projections to the PMF. A 2D histogram of a property can be formed on the same grid as the 2D PMF, $\langle \mathcal{X}(z, s) \rangle$, and then the second coordinate can be integrated over

$$\langle \mathcal{X}(z) \rangle = \frac{\displaystyle\int ds\, \langle \mathcal{X}(z, s) \rangle\, e^{-W(z,s)/kT}}{\displaystyle\int ds\, e^{-W(z,s)/kT}} \tag{51}$$

Similarly one may analyze mean force decompositions in this way (Dorairaj and Allen, 2007). Other manipulations of multidimensional PMFs can also be carried out, such as the remapping of free energy to obtain a PMF along any linear combination of PMF coordinates (see (Li *et al.,* 2008a) for an illustration).

8.4.2.5 **Computing Equilibrium Constants: The Case of pK_a Calculation**

The typical approach for pK_a computation is to simulate the deprotonation process via FEP. If one wishes to compute the pK_a of some residue, R, at a position, $\mathbf{r}$, in space, then one must compute the free energy change for the process

$$\mathrm{RH}^+(\mathbf{r}) \rightarrow \mathrm{R}^0(\mathbf{r}) + \mathrm{H}^+(\mathrm{aq}) \tag{52}$$

This free energy is equal to

$$\Delta G_\mathrm{deprot}(\mathbf{r}) = (G_{R^0}(\mathbf{r}) + G_{H^+}) - G_{\mathrm{RH}^+}(\mathbf{r}) \tag{53}$$

as well as at a position in the bulk $\mathbf{r}'$

$$\Delta G_\mathrm{deprot}(\mathbf{r}') = \left(G_{R^0}(\mathbf{r}') + G_{H^+}\right) - G_{\mathrm{RH}^+}(\mathbf{r}') \tag{54}$$

The difference between these two free energies leads to a cancellation of the free energy of the unconstrained aqueous proton G_{H^+} and is

$$\begin{aligned}\Delta\Delta G_{\text{deprot}} &= \left(G_{R^0}(\mathbf{r}) - G_{R^0}(\mathbf{r}')\right) \\ &\quad - \left(G_{\text{RH}^+}(\mathbf{r}) - G_{\text{RH}^+}(\mathbf{r}')\right) \\ &= (G_{R^0}(\mathbf{r}) - G_{\text{RH}^+}(\mathbf{r})) \\ &\quad - \left(G_{R^0}(\mathbf{r}') - G_{\text{RH}^+}(\mathbf{r}')\right) \\ &= \Delta G_{\text{deprot}}(\mathbf{r}) - \Delta G_{\text{deprot}}(\mathbf{r}') \end{aligned} \tag{55}$$

Thus, the deprotonation process corresponds simply to a mutational transformation from protonated RH^+ to neutral R at two positions. The free energy difference, ΔG, can be computed via thermodynamic perturbation. As we shall show later in this chapter, convergence of this procedure can be difficult to achieve if the protonation event involves significant conformational change of the protein or surrounding membrane or solvent. An alternative approach is to employ a thermodynamic cycle where $\Delta\Delta G_{\text{deprot}}$ can be seen to equal the difference between the PMFs for charged and neutral R from a reference position, z' to a position, z along the chosen reaction coordinate:

$$\Delta\Delta G_{\text{deprot}} = W_{R^0}(z) - W_{\text{RH}^+}(z) \tag{56}$$

(where the PMF has already been defined relative to the reference position, z', via Eq. (30)). Then the change in free energy of deprotonation is related to the pK_a shift via

$$\Delta \text{p}K_a = \Delta\Delta G_{\text{deprot}}/2.3\ kT \tag{57}$$

8.4.2.6 Computing Equilibrium Constants: An Example of Binding Constants

A single-particle dissociation constant K_D can be expressed in terms of the three-dimensional PMF, $W(\mathbf{r})$ (Roux, 1999)

$$K_D^{-1} = \frac{\int d\mathbf{r}\, \mathcal{H}_{\text{site}}(\mathbf{r})\, e^{-W(\mathbf{r})/kT}}{e^{-W(\mathbf{r}')/kT}} \tag{58}$$

where the binding site is defined by a Heaviside step function $\mathcal{H}_{\text{site}}(\mathbf{r})$ which is 1 inside the binding site and 0 outside. K_D may also be expressed in terms of the 1D PMF, $W(z)$ (Allen *et al.*, 2004b). We define the binding site to be within the range $z_{\min} \le z \le z_{\max}$

$$\mathcal{H}_{\text{site}}(\mathbf{r}) = h_{\text{site}}(z)\, H(x, y) \tag{59}$$

where

$$h_{\text{site}}(z) = \begin{cases} 1; & z_{\min} \le z \le z_{\max} \\ 0; & z < z_{\min} \\ 0; & z > z_{\max} \end{cases} \tag{60}$$

then, with $\mathbf{r}' = (0, 0, z')$, Eq. (58) becomes

$$\begin{aligned} K_D^{-1} &= \int dx\, dy\, dz\, \mathcal{H}_{\text{site}}(x, y, z) \\ &\quad \times e^{-[W(x,y,z) - W(0,0,z')]/kT} \\ &= \int dz\, h_{\text{site}}(z) \int dx\, dy\, H(x, y) \\ &\quad \times e^{-[W(x,y,z) - W(0,0,z')]/kT} \end{aligned} \tag{61}$$

Then, using Eq. (50), we find the relationship between the dissociation constant and the 1D PMF

$$\begin{aligned} K_D^{-1} &= \pi R^2 \int dz\, h_{\text{site}}(z) e^{-W(z)/k_B T} \\ &= \pi R^2 \int_{z_{\min}}^{z_{\max}} dz\, e^{-W(z)/k_B T} \end{aligned} \tag{62}$$

where R is the radius of the bounding cylinder used to restrain lateral/secondary coordinates (Allen *et al.*, 2004b).

8.4.2.7 Computing Flux along a Free Energy Path

Having mapped out the free energy surface for a process, one maybe interested in estimating the stationary flux along a given path. Such a calculation involves estimation of the spatially varying diffusion constant and applying a rate model. For instance, the Kramer's transition rate can be computed in the high friction, large activation regime (Allen *et al.*, 2003b; Crouzy *et al.*, 1994)

$$k_{tr} = \frac{D(z_{\text{barrier}})}{2\pi kT} \times \left[-W''(z_{\text{barrier}})W''(z_{\text{well}})\right]^{1/2} \times e^{-\Delta W_{\text{activation}}/kT} \qquad (63)$$

where $D(z_{\text{barrier}})$ is the diffusion constant measured near the top of the barrier in the z-direction, $W''(z)$ the second derivative of the PMF at the barrier and wells, and $\Delta W_{\text{activation}}$ is the activation barrier height for the transition. The stationary flux (J) across a 1D free energy profile can be determined from a 1D Nernst–Planck equation (Levitt, 1986)

$$J = -D(z)\frac{dP(z)}{dz} - P(z)\frac{D(z)}{kT}\frac{dW_{\text{total}}(z)}{dz} \qquad (64)$$

where $D(z)$ is the diffusion coefficient on that reaction coordinate and $P(z)$ is the spatial probability density. The PMF has been substituted for a "total" PMF in this expression to allow for the application of a nonequilibrium driving potential, such as an electrical membrane potential that would drive ions across ion channels in a cell membrane (Allen *et al.*, 2004b). In that case, W_{total} can be expressed as a sum of the equilibrium PMF W, and the interaction of atomic charges, q_i (at positions z_i), with the scalar potential arising from the asymmetrical charge distribution across the membrane, ϕ.

$$W_{\text{total}}(z) = W(z) - kT\ln\left\langle e^{-\sum_i q_i\phi(z_i)/kT}\right\rangle \qquad (65)$$

where z is the location of the ion and the brackets represent an equilibrium average. Different levels of approximation to Eq. (65) can be made, as discussed in (Allen *et al.*, 2006b). In particular, the charge distribution of the protein and its hydration waters can be represented by the moments of its dipole distribution and a cumulant expansion of the above expression obtained. However, at physiological voltages the impact of the voltage on the PMF is dominated by the movement of the ion where, to a first approximation, one can assume that the system charge distribution is independent of ion position and therefore that (Allen *et al.*, 2006a)

$$W_{\text{tot}}(z) \approx W(z) + q_{\text{ion}}\phi_{\text{mp}}(z) \qquad (66)$$

In the limit of zero applied voltage simple expressions for the stationary flux that have been obtained (Allen *et al.*, 2006a) will be solved later in this chapter.

8.4.3
Advanced Sampling and Nonequilibrium Approaches

PMF computation is complicated by the wide-ranging length and timescales of the thermal fluctuations in the biological system (Cooper *et al.*, 1985; Allen *et al.*, 2004a). The timescales of protein conformational changes and membrane dynamics add a further challenge. Improvements in sampling can be obtained through the Umbrella Sampling technique by using more sophisticated biasing potentials. In particular, adaptive biasing approaches (Darve and Pohorille, 2001; Darve *et al.*, 2002; Henin and Chipot, 2004; Rodriguez-Gomez *et al.*, 2004; Chipot and Henin, 2005) allow the simulation to "learn" from previous trajectory to better bias the coordinate and obtain a more uniform distribution. However, in some instances the very choice of order parameters for a PMF calculation becomes complicated, as several coordinates may equilibrate on very long timescales.

The Replica Exchange Method (Swendsen and Wang, 1986; Sugita and Okamoto, 1999) (REM), expedites configurational sampling on a rugged energy surface by introducing random walks by exchanging parallel simulations with different temperatures. This method lends itself to ordinary computer clusters because the replicas are run, for the most part, independently on separate CPUs (Fukunishi *et al.*, 2002). However, as the number of degrees of freedom, f, increases, the distribution of energies in each system becomes very narrow, reducing the transition probabilities between neighboring replicas. As a result, the number of systems required for sufficient exchanges grows as $O(f^{1/2})$. This seems an unnecessary burden when most of the degrees of freedom are associated with solvent. A proposed remedy for this is Hamiltonian REM, where instead of changing temperature between replicas (i.e. effectively scaling the entire Hamiltonian), just part of the Hamiltonian, associated with slow varying coordinates, is scaled leading to severalfold increase in efficiency (Fukunishi *et al.*, 2002). Replicas with weaker interactions will have a smoother energy surface and better configurational sampling. Replicas are mixed via Monte Carlo (MC) exchanges during the simulation using a Metropolis (Metropolis *et al.*, 1953) transition probability

$$W(X, U_m; Y, U_n) = \begin{cases} 1; & \Delta \leq 0 \\ e^{-\Delta} & \Delta > 0 \end{cases} \tag{67}$$

where replica m is in configuration X, with Hamiltonian $U_m(X)$, and replica n is in configuration Y, with Hamiltonian $U_n(Y)$, and

$$\Delta = [(U_m(Y) + U_n(X)) - (U_m(X) + U_n(Y))]/kT \tag{68}$$

Taking this method a step further, Hamiltonian REM could be combined with umbrella sampling (by exchanging Hamiltonians and window functions) to obtain enhanced sampling and PMFs as a function of desired coordinates (Sugita *et al.*, 2000).

A completely different approach to these methods is to use nonequilibrium sampling. By simulating with an applied force one can explore rare configurations in relatively short times and this has been used successfully to study processes such as ion permeation (Isralewitz *et al.*, 2001a; Gullingsrud and Schulten, 2003; Liu *et al.*, 2006; Zhang *et al.*, 2006; Bastug *et al.*, 2008), protein–ligand binding/unbinding (Isralewitz *et al.*, 1997; Isralewitz *et al.*, 2001b), protein translocation (Monticelli *et al.*, 2005) and protein unfolding (Krammer *et al.*, 1999; Paci and Karplus, 2000; Isralewitz *et al.*, 2001b; Gao *et al.*, 2003), for example. In steered molecular dynamics (SMD) simulations the time-dependent force is applied to pull the system along some particular degrees of freedom (Isralewitz *et al.*, 2001b). In the constant velocity pulling method an atom being pulled is attached via a harmonic spring to some restraint point (dummy atom) moving with certain velocity (Lu *et al.*, 1998). In another method the restraint point does not move but the spring stiffness increases over time (Isralewitz *et al.*, 2001a). SMD methodology is closely related to experimental techniques such as atomic force microscopy or optical tweezers (Isralewitz *et al.*, 2001b). The direction of the applied force plays an important role in SMD simulations and can be chosen to be random, guessed from structural information or obtained from the so-called targeted MD, which employs time-dependent holonomic constraints to move from one state of the system (e.g. protein conformation) to another (Schlitter *et al.*, 1993). The direction of movement from initial to final state can be then used in SMD (Isralewitz *et al.*, 2001a). The magnitude and pathway of the pulling can be also controlled manually using haptic device in a technique called interactive molecular dynamics (IMD) (Grayson *et al.*, 2003).

But such nonequilibrium approaches can also be used to learn about equilibrium properties of the system using the Jarzynski equality (JE) (Jarzynski, 1997)

$$\Delta G = -kT \ln\left\langle e^{-\tilde{W}/kT}\right\rangle \tag{69}$$

where $\tilde{W}$ is work done on the system. The JE was derived first on the basis of Hamiltonian equations of motions but generalized later by Crooks (1998, 2000) for other systems, for which dynamics is Markovian and microscopically reversible. JE demonstrates that the work required to follow each of an ensemble of nonequilibrium paths, $\tilde{W}(z)$, can be used to calculate the equilibrium PMF. This methodology combined with SMD thus can serve as a viable alternative to umbrella sampling techniques (Park and Schulten, 2004; Kosztin *et al.*, 2006). Moreover, SMD might be considered equivalent to umbrella sampling MD simulations in the limit of weak slowly changing in time external forces inducing only minor changes (Isralewitz *et al.*, 2001b).

JE has been used in both simulations using SMD or similar methods (Xiong *et al.*, 2006; Bastug and Kuyucak, 2007; Bastug *et al.*, 2008) as well as experiments (Hummer and Szabo, 2001); (Liphardt *et al.*, 2002). The application of very large forces not only allows one to sample very rare events in short simulations, but also can be unbiased accordingly (Wereszczynski and Andricioaei, 2006). This formalism can also be applied to calculation of other thermodynamic properties, for example entropic and enthalpic contributions without the necessity to run simulations at several temperatures (Nummela *et al.*, 2008). The major concern with JE-based simulation methodology is sampling, especially for large biomolecular systems (Frenkel and Smit, 2002; Xiong *et al.*, 2006).

A wide array of improved sampling methods, both equilibrium and nonequilibrium, have been developed in recent years (see Adcock and McCammon (2006) for a recent review), yet in this chapter we focus on standard techniques as illustrations.

8.5
Case Studies

8.5.1
Ion Channel Permeation

We consider an important class of membrane-proteins, known as *ion channels,* which allow selective permeation of charged molecules across the cell membrane. Ion channels are excellent probes for the state of the bilayer because they have a direct measure of function (electrical current) and benefit from decades of structural and functional studies. We consider the gA channel (Urry, 1971) as a prototypical example because of its simplicity and extensive experimental characterization. The gA channel forms by TM dimerization of single stranded, right-handed β-helical monomers (Urry, 1971). With known structure (Arseniev *et al.*, 1986; Townsley *et al.*, 2001; Ketchem *et al.*, 1997; Allen *et al.*, 2003b) and extensive functional studies (Hladky *et al.*, 1979; Eisenman and Horn, 1983; Andersen and Koeppe, 1992; Busath, 1993; Andersen *et al.*, 2005), gA is an ideal prototype for studies of ion permeation and selectivity in narrow pores. The dimer is shorter than typical lipid bilayers and can lead to a ‘‘hydrophobic mismatch’’ and corresponding free energy cost for the membrane to accommodate the channel (Lundbaek *et al.*, 1997). As a result, gA activity is sensitive to the composition of the surrounding membrane and has become a useful ‘‘probe’’ for membrane structure and composition. The gA channel poses a significant challenge for computational investigation because of the need to accurately represent both the interactions between ions and bulk water and ions and the protein in an almost dehydrated state.

For the sake of simplicity, ion channel proteins have often been pictured as rigid macromolecules, immersed in an implicit solvent continuum. Many permeation models have used a single rigid structure and continuum methods to represent interactions inside narrow ion channels, yet ion energetics can be very sensitive to the precise location of protein atoms and their ability to undergo a flexible response (Allen *et al.*, 2004a). It has long been known that proteins have a rather fluid-like nature (Cooper, 1976) that undergo thermal fluctuations, from subpicosecond vibrations, to multinanosecond global reorientations and side chain isomerizations and eventually μs to second conformational changes (Karplus and McCammon, 1981). In particular, with single permeation events requiring 10–100 ns, extensive protein and solvent fluctuations occur during ion translocation through a pore. The effects of fluctuations on protein function has been known for several decades (Northrup *et al.*, 1982; Perutz and Mathews, 1966; Takano, 1977) and numerous simulations have suggested significant atomic fluctuations in ion channels such as gA (Mackay *et al.*, 1984; Roux and Karplus, 1991) and the KcsA potassium channel (Allen *et al.*, 1999; Guidoni *et al.*, 1999; Allen *et al.*, 2000; Aqvist and Luzhkov, 2000; Berneche and Roux, 2000; Shrivastava and Sansom, 2000; Noskov *et al.*, 2004). Experimental results on thermal fluctuations in proteins (Frauenfelder *et al.*, 1979; Parak and Knapp, 1984; Ringe and Petsko, 1986; Gabel *et al.*, 2002), the ion-induced reorganization of the KcsA pore (Zhou *et al.*, 2001) as well as simulation-based calculations all show that proteins are flexible on subangstrom length scales, with significant response to perturbative forces, such as those arising from ions, which cannot be neglected in narrow ion channels (Allen *et al.*, 2004a). In addition to thermal fluctuations, discrete water molecules gives rise to structural effects on length scales under 10 Å (Israelachvili and Wennerström, 1996). It is therefore essential to study ion permeation using fully atomistic MD simulation.

8.5.1.1 **Simulation Details**

Simulations were carried out with the program CHARMM (Brooks *et al.*, 1983) using the PARAM27 (MacKerell, 1998b; Schlenkrich *et al.*, 1996) protein/lipid force field with TIP3P water (Jorgensen *et al.*, 1983) parameters and ion LJ parameters from Beglov and Roux (1994). Dimyristoylphosphatidylcholine (DMPC) bilayers of different sizes were built around the gA protein using extensions of existing membrane-building techniques (Woolf and Roux, 1996) (Figure 8.1a). The gA helical dimer (PDB:1JNO) (Townsley *et al.*, 2001), has been shown to provide agreement with a range of experimental solution and solid-state nuclear magnetic resonance observations (Allen *et al.*, 2003b) and was used for these studies. Membranes consisted of ~1 or 3 shells of phospholipids, containing 20 or 96 lipid molecules (and 1080 or 3996 water molecules), respectively (though for illustration we shall show results only for the smaller system here). Hexagonal periodic boundaries were chosen, because they provide more efficient use of lipid patch, with respective xy-translation lengths 32.1 and 61.9 Å. Pressure coupling was employed in the z direction (parallel to the membrane normal, with average height ~75 Å), while the x–y dimensions of the box remained fixed to ensure correct lipid structure for this CHARMM force field. Ionic solutions of 1 M KCl were used to ensure good electrolyte sampling. Particle-mesh Ewald, SHAKE on bonds involving H atoms and Nose–Hoover constant pressure and temperature algorithms were employed. The temperature of 330 K ensured the system remained well above the gel-phase transition temperature for the lipid bilayer and enhanced configurational sampling.

8.5.1.2 **Free Energy Profile Calculation**

At moderate concentrations the gA pore is singly occupied such that a one-ion PMF, $W(\mathbf{r})$, can be used to model the flux of ions through the pore. We choose a reference $W(\mathbf{r}') = 0$ at a position $\mathbf{r}\prime = 30$ Å in the bulk where the free energy surface has become flat. The one-ion pore region has been determined to be defined approximately by a sphere of radius 14 Å (Allen *et al.*, 2006a) around the COM of the gA dimer. This is the largest sphere for which two cations on one side of the channel is not seen, and for which anions are not observed. However, it is possible to have one cation on each side of the channel (Allen *et al.*, 2006a) and thus a spherical repulsive flat-bottom spherical harmonic restraint with force constant 5 kcal mol^{-1} was used to maintain a one-ion PMF.

Because conduction is driven by a transmembrane voltage difference we choose a reaction coordinate z, parallel to the fixed membrane normal, that is the distance of the ion to the COM of the gA dimer. Because the configurational integral of Eq. (30) will not be defined as the ion leaves the channel and becomes unbounded, we restrict lateral displacements with a flat-bottom

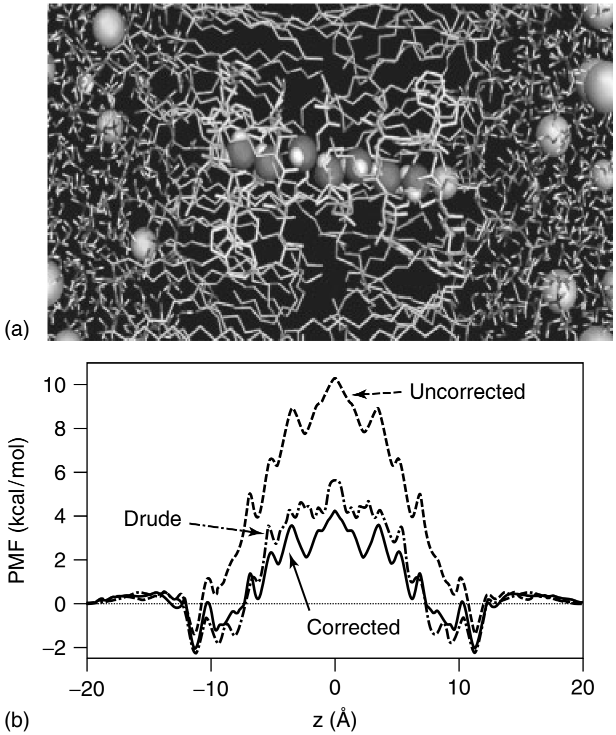

Fig. 8.1 Gramicidin A ion channel. (a) gA protein (yellow) embedded in a hydrated DMPC membrane (gray hydrocarbon tails, water molecules as red and white sticks, potassium ions as green balls and chloride ions as gray balls) with membrane normal oriented horizontally. Water molecules inside the narrow pore are shown as balls. (b) 1D PMF for permeation before any artifact corrections ("uncorrected", dashed), after postsimulation corrections ("corrected", solid) and after simulation with a Drude-polarizable membrane hydrocarbon core ("Drude", dash-dot) (adapted from (Allen *et al.*, 2004b) with "Drude" curve added). (Please find a color version of this figure on the color plates.)

cylindrical constraint of radius $R = 8$ Å and force constant 10 kcal mol^{-1} Å^{-2} (Allen *et al.*, 2004b). This radius is large enough such that it is not felt for an ion inside the channel but small enough such that it maybe thoroughly sampled in a short simulation. The use of alternative constraints on lateral displacements can cause poorly defined artifacts in PMF calculations (Roux *et al.*, 2008). We calculated this PMF with umbrella sampling using 101 simulations, biased by harmonic window functions with force constant 10 kcal mol^{-1} Å^{-2} that hold the ion at 0.5 Å increments in $z = (-20, +30)$ Å.

To obtain starting configurations it is important not to generate too large a perturbation owing to the sudden placement or pulling of an ion that may take the system too far from equilibrium. We avoid any "dragging" procedure that would pull ions into the narrow pore so as to maintain a correct protein and vicinal water geometry. We instead place ions at regular positions by replacing either existing ions, or if ions do not exist closeby, at water positions and then displace that water molecule. This procedure involves searching every frame of a 4 ns unbiased MD trajectory for the best possible starting configuration (Allen *et al.*, 2006a). Simulation times of 2 ns/window are needed in order to obtain reliable simulations, as shown in the analysis of Allen *et al.* (2006a). Shorter simulations can lead to asymmetries in the 1D PMF of several kilocalories per mole. One should simulate long enough that asymmetries drop to a desired level (1 kcal mol^{-1} or less, perhaps) and then may impose symmetry on the PMF. Such levels of asymmetry correspond to very small average force across the channel (orders of magnitude smaller than the average magnitude of the instantaneous force). Symmetrization should not be imposed after solving the WHAM equations but before, as this can assist in the solution of the iterative equations.

Figure 8.1b shows the symmetrized 1D PMF after 2 ns of simulation per window (dashed curve), over a reduced range $-20 \leq z \leq 20$, with average error of $\sim$0.3 kcal mol^{-1} (Allen *et al.*, 2006a). The PMF exhibits a central barrier of $\sim$11 kcal mol^{-1} with respect to the binding site. We refrain from calculating a barrier with respect to the unbound bulk reference because this is dependent on the choice of lateral constraining method. There is a deep outer binding site at $z = 11.3$ Å, and a less deep inner binding site at 9.7 Å. The positions of these minima are consistent with experiments, with the 2D PMF surfaces, to be described below, shedding more light on this agreement.

The magnitude of the barrier in Figure 8.1b has been shown to be too great for agreement with experimental flux measurements (Allen *et al.*, 2004b). However, the system setup suffers from artifacts owing to the finite size of the membrane patch and the inadequate description of the hydrocarbon core of the lipid membrane (Allen *et al.*, 2004b). The finite size of the patch, combined with periodic boundaries, leads to an artificial destabilization of an ion as it enters a lower dielectric membrane environment relative to the high dielectric aqueous phase. Poisson–Boltzmann solutions have shown that the finite size effect can overestimate the barrier by 1.6 kcal mol^{-1}, as confirmed by simulations with larger membrane patches (Allen *et al.*, 2006a). The lacking electronic polarizability of the MD force field for hydrocarbon results in a dielectric constant of $\sim$1 instead of the experimental value of $\sim$2 for hydrocarbon tails of lipids. The effect has been calculated by either Poisson–Boltzmann solutions using a thermodynamic cycle as described previously (Allen *et al.*, 2006a) with a correction of 2.1 kcal mol^{-1} needed. This correction has also been estimated by obtaining average shifts in energy when polarizability of the lipid tails is added *a posteriori*. Drude oscillators (Vorobyov *et al.*, 2005) have been added to all lipid carbon atoms to obtain a result consistent with the Poisson–Bolztmann solutions, but slightly larger corrections up to 3.6 kcal mol^{-1} at the center of the channel (see solid "corrected" curve). The lacking electronic polarizability of the lipid tails may also introduce other artifacts, such as an artificially high dipole potential of the membrane (Klauda *et al.*, 2008). However simulations of the membrane-protein system with the polarizable lipid core included during simulations (dash-dot "Drude" curve) have led to similar results to post-corrections, indicating that the changes in membrane dipole potential are not experienced by the ion inside the ion channel.

8.5.1.3 Connection with Observables

We calculated the flux through the ion channel in the limit of low membrane voltage. This maximum conductance, based on the corrected PMF (solid curve in Figure 8.1b), was found to be 9.3 pS, only a factor of 2 from the experimental value of 21 pS (in similar branched DPhPC (diphytanoylphosphatidylcholine) bilayers with 1 M KCl at 100 mV (Busath *et al.*, 2003)). This indicates that the PMF for K^+ conduction through the gA ion channel is likely within a kilocalories per mole of the true free energy profile.

We calculate the single-ion dissociation constant for the channel of 0.14 M using the same corrected PMF (and 0.07 M with the "Drude" PMF), in agreement with experimental values from NMR and conductance studies: 0.07 M (Andersen *et al.*, 2005) (in DPhPC/*n*-decane bilayers); 0.017 M (Jing *et al.*, 1995) (measured in dodecylphosphocholine micelles); 0.019–0.73 M (Hinton *et al.*, 1986) (in aqueous lysophosphatidylcholine dispersions); and 0.035 M (Thompson *et al.*, 2001) (in glycerylmonoolein bilayers). We have also determined that the ion preferentially binds to the outer binding site ($10.2 < z < 12.5$ Å) with $K_D = 0.83$ M as compared to the inner site ($6.9 < z < 10.2$ Å) with $K_D = 3.6$ M. This is in agreement with Tian and Cross (1999) who observed a larger chemical shift for the Leu12-Trp13 than the Leu10-Trp11 linkage for K^+, but is in contrast to X-ray scattering by Olah *et al.* (1991) who see preferred binding nearer the inner site. We again warn that the 1D PMF above establishes a connection to experimental observations of binding and conductance but cannot establish a connection between the channel and bulk environment in terms of relative free energies due to the inability to define the free energy in the unconstrained aqueous phase.

8.5.1.4 Two-dimensional Free Energy Surfaces Reveal Mechanisms

Figure. 8.2a reveals the 2D PMF as a function of membrane normal, z, and radial, r, coordinates (for a greater range see Allen *et al.*, 2006b). This (uncorrected) 2D PMF reveals the free energies of binding sites and the central barrier relative to bulk. The 2D PMF also shows the extent of lateral ion motion, which, in the frame of the membrane normal, includes the influence of channel tilting (12° on average (Allen *et al.*, 2004b)) on the lateral displacements. The outer binding site at $z = 11.3$ Å is the deepest point in the free energy surface, being -3.2 kcal mol^{-1} relative to the bulk. In the narrowest part of the channel, an ion experiences a barrier of 7.2 kcal mol^{-1} relative to the bulk. This barrier has been confirmed to be accurate to the order of $\sim$1 kcal mol^{-1}, with the chosen force field, with free energy perturbation calculations (Allen *et al.*, 2004b).

Although the membrane normal coordinate is pertinent to ion conduction down a membrane potential gradient, an instantaneous channel-axis frame PMF

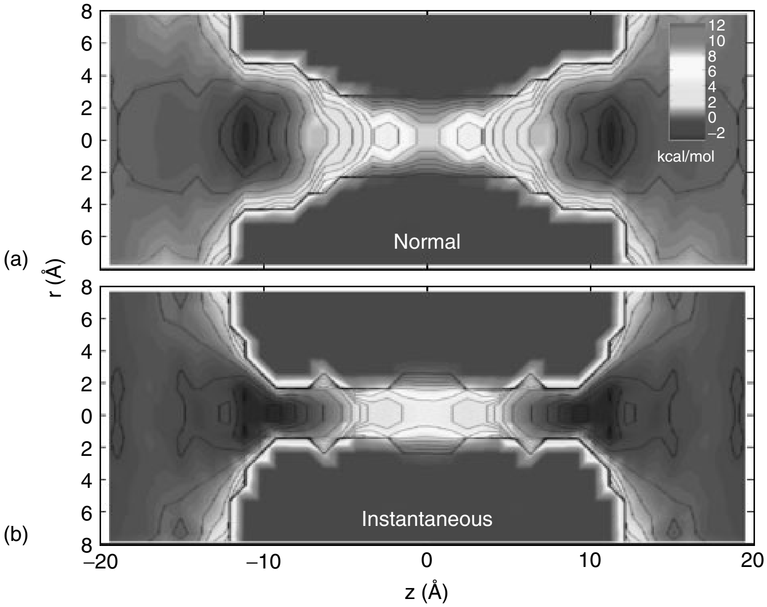

Fig. 8.2 Two-dimensional PMFs for permeation through gA. (a) obtained using a reaction coordinate that is parallel and perpendicular to the membrane normal, (b) obtained using an instantaneously fluctuating channel axis coordinate (adapted from (Allen *et al.*, 2006a)).

should reveal much more about the ion–protein–water interactions (Allen *et al.*, 2006a). We have defined this fluctuating axis coordinate as the position parallel to the vector connecting the centers of mass of the two monomers of gA. Figure 8.2b reveals no wide "free energy vestibules" as seen in Figure 8.2a. It also shows that the outer binding sites are in fact off-axis, as suggested experimentally (Finkelstein and Andersen, 1981), where the ion attains a hydration by three water molecules as opposed to two in the inner site. But what this PMF shows is that the inner binding sites are in fact deeper (by ∼0.5 kcal mol^{-1}) than the outer sites, whereas they were almost nonexistent in the membrane-normal-frame 2D PMF. It was owing to the entropic contribution from being off-axis (occupying a greater volume) that the outer sites were preferred in the original lab-frame PMF. This finding allows one to better reconcile simulation and experiment. In particular, off-axis binding implies greater translational freedom such that the protein backbone will require less distortion to stabilize the ion, and the effect of tilting also must play a role in interpretation of X-ray experiments.

The large stabilization that results from interactions with the single-file pore water (Allen *et al.*, 2004b; Allen *et al.*, 2006a) led us to question whether dehydration could be rate determining. For example, if the ion must wait at a particular position for dehydration to occur then this could suggest that an additional dehydration order parameter may be needed to describe the permeation process (Allen *et al.*, 2004b; Bolhuis *et al.*, 2000). $W(z, n)$ of Figure 8.3a shows the free energy of an ion at z, hydrated by n water molecules in the first shell, relative to $z = 30$ Å where $n \sim 6-7$. Dehydration seems to be a continuous process when entering the narrow pore, dropping to three in the outer site and two in the inner binding site.

However, water in this narrow pore plays a more complicated role. While the water column can be oriented with its dipole running one way or the other in the absence of an ion, ion movement requires an alignment of the water column with oxygen atom adjacent to the cation. Figure 8.3b shows the 2D PMF, $W(z, \mu)$, as a function of the ion position and the dipole moment, μ, of the single-file water molecules. With the ion outside the channel there are two possible dipole moments, ±15 D, evidenced by free energy minima. But when the ion occupies the outer binding site, the preferred state has the water aligned such that the last oxygen atom is coordinating the ion ($\mu = -15$ D). But the ion cannot enter the channel inner site until the column is oriented the correct way. There is a 2–3 kcal mol^{-1} barrier to flip the column while the ion is outside the channel and a smaller barrier of <1 kcal mol^{-1} to flip the

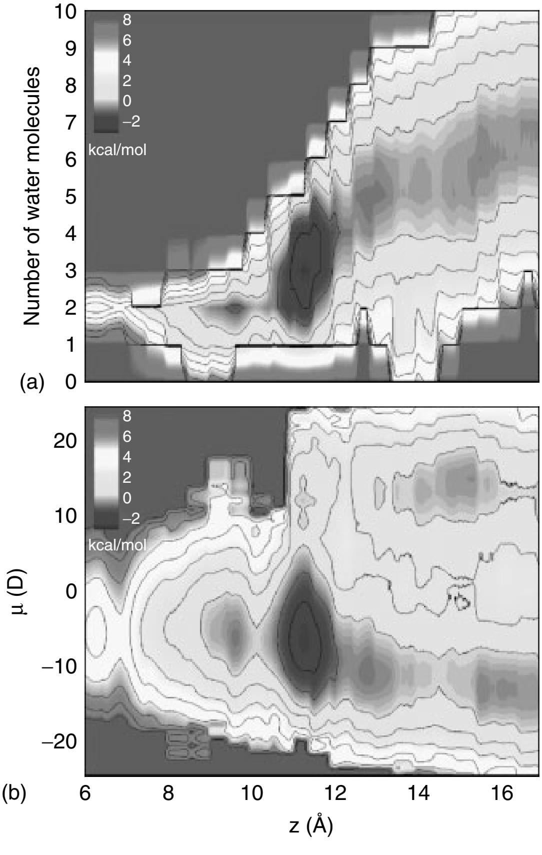

Fig. 8.3 PMFs as a function of position and solvation coordinates. (a) As function of coordination number and (b) as function of dipole moment of the single-file column of water. The restricted range of z ranges from just inside to just outside the narrow pore (adapted from (Allen *et al.*, 2004b)).

single-file column when the ion is in the outer site. Regardless, a sudden change in μ is required for entry with an associated free energy barrier opposing permeation that is not a function of the coordinate z. Any PMF that uses only a single positional order parameter as a reaction coordinate must ensure adequate sampling of slow coordinates such as this water column flipping. Otherwise, two order parameters should be used to describe ion permeation (z and μ) (Allen *et al.*, 2004b).

The dipole moment of the single-file water also complicates conduction calculations at large membrane potentials because it couples to the external potential with nontrivial contributions (Allen *et al.*, 2006a). Moreover, at higher concentrations, multiple ion mechanisms will become more significant. However, the dipole of the single file will contribute to the interaction between multiple ions over the length of the channel. For instance, as an ion approaches from the right in Figure 8.3b, it requires a dipole moment of -15 D to enter. Once it crosses, the chain of waters will have been ''zipped up,'' leaving the dipole moment at $+15$ D, making it difficult for another ion to enter from the right. This suggests a structural long-range coupling across the pore that would be absent in continuum solvent models (Allen *et al.*, 2004b). The mechanism is, however, not so simple because multiple ions will perturb this single-file arrangement.

This case study on gramicidin permeation is by no means complete and has been provided as an illustration of the power of MD simulation when combined with free energy calculation methods. A more complete survey should include permeation studies in other channels, such as potassium, calcium, sodium, and chloride channels. Furthermore, MD simulation has been used not just to study permeation but also to explain how ion channels precisely select between different ion types on the basis of valence (Allen *et al.*, 2006a; Corry *et al.*, 2002), as well as size (Allen *et al.*, 1999; Guidoni *et al.*, 1999; Allen *et al.*, 2000; Luzhkov and Aqvist, 2001; Berneche and Roux, 2001; Noskov *et al.*, 2005; Leung *et al.*, 2006; Bostick and Brooks, 2007). The mechanisms of size selectivity have been an issue of much controversy in recent years. Early explanations of size-based ionic selectivity of potassium channels in terms of the precise geometry of a pore perfectly adapted to snugly fit potassium, but not sodium (Doyle *et al.*, 1998; Yellen, 2002), once one takes into consideration the effects of thermal fluctuations of the protein, are unlikely to be correct (Allen *et al.*, 2004a). While the spatial arrangement of the pore-lining groups is important, it has long been recognized that channels use a combination of geometric, chemical, and energetic factors to select among different ions. Thus while the size of the ion and the channel pore must be similar to allow selective permeation (the so-called sieving mechanism (Mullins, 1960; Hille, 1973)), a complete description of selectivity must include the specific microscopic interactions and their influence on solvation and selectivity (Eisenman, 1961; Eisenman and Horn, 1983; Allen *et al.*, 2004a). More recently, computational models have attempted to elucidate the mechanisms of ion selectivity on the basis of coordination (Bostick and Brooks, 2007; Leung *et al.*, 2006) and specific ion–ligand and ligand–ligand energetics (Noskov *et al.*, 2005) that have brought us much closer to a microscopic understanding of ion channel function.

8.5.2
Models to Explore Voltage-Gated Ion Channel Activation

The second case study considered here is a series of all-atom MD simulations to explore the possibility of the movement of membrane-bound protein segments that contain multiple ionizable amino acid side chains. This study was motivated by recent controversy over the activation mechanisms of voltage-gated ion channel proteins. The voltage-gated potassium channel KvAP contains several Arg side chains in its voltage sensor domains that are designed to sense changes in the membrane voltage and initiate a conformational change in the protein to open up the lumen to allow potassium ion permeation. One proposed model, called the "paddle model," suggests that the Arg-rich S4 segments of the voltage-sensing domains move across the hydrophobic membrane core during activation (Jiang *et al.*, 2003b). Such a motion necessitates the movement of several Arg side chains, known to remain charged, exposed to lipid hydrocarbon. However, theoretically, the cost of moving charges from water across a slab of hydrocarbon is very large and the barrier for this activation has been estimated to be of the order of 200–300 kcal mol^{-1} (Grabe *et al.*, 2004).

Elevating this controversy, small apparent free energies of $\sim$2 kcal mol^{-1} have been measured for moving an Arg side chain along an inserted hydrophobic TM helix (Hessa *et al.*, 2005a). These measurements come from recent cell biological experiments conducted using the translocon/oligosaccharyl transferase machinery of protein synthesis: a very complex process that is currently poorly understood. The nearly zero free energy for membrane insertion of the KvAP S4 peptide alone was also suggested from these studies (Hessa *et al.*, 2005b) as well as recent CG MD simulations (Bond and Sansom, 2007), shown to be inadequate in their ability to reproduce water–membrane partitioning energetics (Vorobyov *et al.*, 2008). Each of these studies has challenged a long-held view that lipid membranes act as barriers for charged and polar species (Parsegian, 1969) and the corresponding theoretical estimates of the associated barriers for a charged Arg to cross a lipid bilayer on the basis of continuum membrane models (Honig and Hubbell, 1984; Kessel and Ben-Tal, 2002; Dorairaj and Allen, 2007; Ulmschneider *et al.*, 2007). However, as pointed out nearly 40 years ago by Parsegian in his pioneering work on membrane permeability, a fundamental limitation of continuum models is the simplified picture of the membrane as a low-dielectric slab segregated from bulk water. In fact, a membrane is fluid-like and will deform under the large electric fields around an ion (Parsegian, 1969). Thus, any model of ion–membrane interactions must allow for this flexible response. We therefore turn to atomistic MD studies that have shown that lipid membranes can undergo substantial deformations in the presence of charged species (Freites *et al.*, 2005; Johansson and Lindahl, 2006; Dorairaj and Allen, 2007; MacCallum *et al.*, 2007; Li *et al.*, 2008a), and elucidate the mechanisms and thermodynamics of such processes using simple model protein systems.

We have obtained spatially resolved free energy profiles of Arg side chain transfer across the lipid bilayer using the all-atom CHARMM force field (MacKerell, 1998b). To test whether low energies seen in translocon experiments

(Hessa *et al.*, 2005) can result from Arg deprotonation, we performed these simulations for both charged and neutral Arg side chain models (Dorairaj and Allen, 2007; Li *et al.*, 2008a), despite the large aqueous pK_a of Arg (12–13.7) (Nozaki and Tanford, 1967; Angyal and Warburton, 1951; Hall and Sprinkle, 1932). We have used simple side chain analog models, methyl guanidinium (MguanH$^+$) and methyl guanidine (Mguan^0), but will not describe those models here. Instead, we focus on simple TM helix models consisting of a charged (ArgH$^+$) or neutral (Arg^0) side chains in the middle of a long 80-residue polyleucine TM helix (Dorairaj and Allen, 2007; Li *et al.*, 2008b; Li *et al.*, 2008a). Other similar studies have been performed recently using different analog models and force fields (MacCallum *et al.*, 2007; MacCallum *et al.*, 2008; Yoo and Cui, 2008; Johansson and Lindahl, 2008).

8.5.2.1 **Simulation Details**

The simulation details of our calculations are provided in the corresponding papers (Dorairaj and Allen, 2007; Li *et al.*, 2008a; Vorobyov *et al.*, 2008) and will be only briefly mentioned here. All simulations were performed using program CHARMM with standard protein (MacKerell, 1998b) lipid (Feller and MacKerell, 2000; Yin and Mackerell, 1998), ion (Beglov and Roux, 1994), and recently developed neutral Arg side chain (Li *et al.*, 2008b) parameters as well as TIP3P water model (Jorgensen *et al.*, 1983). The TM peptide was built initially in the ideal α-helical conformation. The peptide was then placed at the position of interest with the new hydrated membrane built around it asdescribed previously (Woolf and Roux, 1996; Dorairaj and Allen, 2007; Li *et al.*, 2008a). Membrane patches consisted of $\sim$ 2 layers or 48 DPPC lipid molecules. The systems were then solvated with $\sim$ 0.5 M aqueous solution of KCl to provide good sampling of ionic bath and neutralize the positive charge of protonated Arg side chain. Hexagonal PBC with an xy-translation length of $\sim$ 44 Å and an average height of $\sim$ 190 Å were used. The lateral dimensions were based on the experimental area per lipid of 64 Å^2 (Nagle, 1993) and estimated protein cross-sectional area of $\sim$ 180 Å^2, whereas the height was chosen to avoid the TM helix passage through periodic boundaries and minimize its interactions with image peptides.

MD simulations were performed under constant temperature of 330 K (above the gel-phase transition for a pure DPPC bilayer) using a Nose–Hoover thermostat and constant normal pressure of 1 atm while keeping fixed lateral box dimensions using the Langevin piston method. Free energy profiles of Arg side chain translocation across the lipid bilayer were calculated with umbrella sampling using biasing potentials that hold the z component of the peptide COM relative to that of the lipid bilayer. Force constants of 2.5 kcal mol^{-1} Å^{-2} were used, corresponding to RMSF of 0.5 Å and providing good overlap of umbrella sampling windows (Li *et al.*, 2008a). These simulations ran for 5–10 ns after initial equilibration periods to converge solvation IEs. Electrostatic interactions were calculated using PME with a coupling parameter of 0.34 and sixth-order spline for mesh interpolation. Nonbond pair lists were maintained out to 16 Å, whereas vdW as well as real-space electrostatic energies and forces were computed up to 10 Å, with the LJ term truncated via an atom-based shift algorithm (Li *et al.*, 2008a).

8.5.2.2 **Mechanism of Arg Translocation**

Fully atomistic MD simulations have shown that the charged Arg side chain movement across the lipid bilayer is a not a simple dehydration as envisioned by continuum membrane models. When ArgH^+ moves into the membrane core, it disrupts the uniform planar geometry of the bilayer and remains coordinated by ≥ 5 water and ≥ 1 lipid phosphate moieties at all positions across the membrane (see the left-hand inset in Figure 8.4 for an example), as revealed by RDF analysis (Li *et al.*, 2008a). These membrane deformations are due to strong interactions between charged Arg side chain and lipid head groups and water molecules, which are as large as -150 and $-50\,\text{kcal mol}^{-1}$, respectively, and impose significant forces on the bilayer (Dorairaj and Allen, 2007; Li *et al.*, 2008a; Vorobyov *et al.*, 2008). This also can be viewed as the formation of an interfacial connection, which effectively results in local membrane thinning. Furthermore, the Arg side chain can "snorkel" toward a membrane interface, complicating sampling. The proper sampling of interfacial connections and side chain rotamers is necessary to get accurate free energy profiles because connections to both the closest and distant membrane interface must be sampled. This is challenging, however, due to large forces acting on the Arg side chain in the membrane core (up to $10\,\text{kcal mol}^{-1}\,\text{Å}^{-1}$). Once the interfacial connection is established, it does not break during multinanosecond MD simulations (and likely taking hundreds of nanoseconds (Dorairaj and

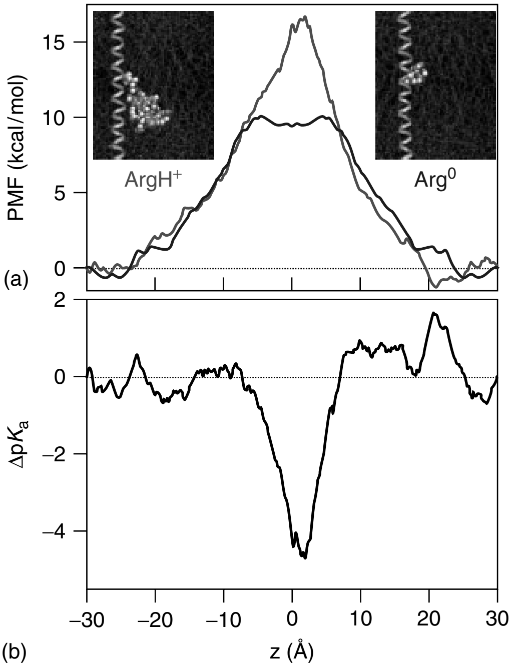

Fig. 8.4 Arginine side chain attached to a TM helix. (a) PMF across the membrane (adapted from (Li *et al.*, 2008a)). MD systems for the charged (ArgH^+) and neutral Arg side chain (Arg^0) with the side chain held near the center of the bilayer are shown as insets revealing membrane deformations in the charged case. Arg is drawn as blue/gray balls, the TM helix as a green ribbon and the hydrated DPPC bilayer with gray C atoms, orange P atoms and red/white water molecules. Lipid phosphate and water atoms that have been pulled into the bilayer core are drawn as balls. The PMFs for charged (red) and neutral (blue) Arg from umbrella sampling simulations are shown. (b) Arg $\text{p}K_\text{a}$ shift profile across the membrane calculated from free energy profiles. (Please find a color version of this figure on the color plates.)

Allen, 2007)). In the innermost ~10 Å the guanidinium group may contact either interface as a result of snorkeling yet in single window simulations will only contact one interface, leading to large artificial average forces and an erroneous PMF calculation. To overcome this challenge, 2D umbrella sampling simulations were carried out where both the peptide COM and relative Arg side chain guanidinium carbon position were biased to produce the 2D PMF, $W(z, s)$. The 1D PMF $W(z)$ was then obtained by integrating $W(z, s)$ over the second variable s using Eq. (46). In similar analog simulations the equilibrium distribution of interfacial connections was obtained through restricted phase space sampling approach (Li *et al.*, 2008a), not being complicated by side chain structural isomers.

In neutral Arg side chain simulations no membrane deformations were observed (see right-hand inset of Figure 8.4 for an example), which is a consequence of much weaker interactions (by almost an order of magnitude) with water and membrane components. Arg^0 does, however, experience dehydration from 10 water molecules in bulk water to just 1–2 in the middle of membrane. The wide rotamer distribution for Arg^0 in different regions of the membrane and the absence of interfacial connections suggests that there is no need for special sampling. However, owing to the weak interactions of neutral Arg side chain with the membrane, longer equilibration times of 4–5 ns per window were required (Li *et al.*, 2008a).

8.5.2.3 Free Energy Profiles Across the Membrane

The free energy profile for the charged Arg side chain is shown in Figure 8.4a as a red line. The barrier to cross the lipid bilayer is large, 17 kcal mol^{-1}, and climbs more steeply as the charge moves deeper into the membrane, which can be explained by greater local deformations of the membrane structure (i.e. the cost to bring in more water molecules and lipid head groups into the hydrocarbon core of the membrane). We note that the full extent of the barrier is not included in this profile because this is the PMF with a reaction coordinate that is the COM of the TM helix, allowing the guanidinium to snorkel by several angstroms. The guanidinium charge is never resting at the center of the membrane. We have shown that the barrier would climb by another 3–4 kcal mol^{-1} if the guanidinium were forced to pass through the center, as shown through a remapping and integration of the 2D PMF (not shown) (Li *et al.*, 2008a). Such a barrier was also shown to be consistent with that found for an Arg analog molecule (Li *et al.*, 2008a).

The large free energy barrier indicates that the transfer of the charged Arg side chain across the lipid bilayer is highly improbable. From the slope of these free energy profiles, it can be estimated that a conformational change in a protein that requires the movement of a single Arg side chain from across the hydrophobic core must overcome a force of over 1 kcal mol^{-1} Å^{-1}, requiring the equivalent of over 1 V across the membrane (Dorairaj and Allen, 2007). This is an order of magnitude greater than physiological values and argues against any model of voltage gating of ion channels that requires a charged Arg side chain to reside or move into the hydrocarbon core of the lipid membrane.

The focus on the charged state of Arg was a result of the very high aqueous pK_a of that amino acid, because it is known that Arg does carry charge during voltage-gated ion channel activation. Yet, having excluded the possibility of "paddle"-like motions, we now question the assumption that Arg would remain protonated. Perhaps deprotonantion is a possible cause for low experimental observations in recent translocon experiments? The calculation of pK_a requires the computation of the free energy profile for the neutral Arg side chain. The Arg^0 PMF is shown in Figure 8.4a as a blue curve and exhibits a 10 kcal mol^{-1} barrier. While this is smaller than for the charged species, it is still prohibitively large, owing to the energetic costs of losing hydrogen-bonds for the neutral guanidine (Li *et al.*, 2008a). The profile for the neutral species is also different in shape, displaying a plateau instead a sharp peak. This again is a result of the lack of membrane deformations and their energetic costs.

8.5.2.4 **pK_a Profiles Across the Membrane**

On the basis of PMFs for both the charged and neutral Arg side chain, its pK_a shift profile across a lipid membrane was calculated using a thermodynamic cycle and Eq. (56). The ΔpK_a profile is shown in Figure 8.4b and remains close to 0 throughout the membrane, dropping by just ~4.5 units within a few angstroms from the membrane centre. Taking into account its high aqueous pK_a (12–13.7), this indicates that Arg remains protonated throughout the membrane at pH 7. The relatively small pK_a shift for the Arg side chain in the membrane core is owing to membrane deformations preferentially stabilizing the charged state. In the absence of such deformations the pK_a shifts would be several times greater as shown by using implicit membrane calculations as well as by extraction of the contributions from displaced water molecules and lipid head groups (see below) (Vorobyov *et al.*, 2008).

Alternatively, the ΔpK_a profile could, in principle, be obtained using FEP for Arg deprotonation at different positions across the membrane. This approach was recently used by us and others (Yoo and Cui, 2008) to calculate pK_a values inside a lipid bilayer to obtain some consistent results. However, we concluded that the convergence of membrane deformations for partially charged species in the FEP simulations was uncertain, favoring the thermodynamic cycle approach above (Li *et al.*, 2008a; Li *et al.*, 2008b).

8.5.2.5 **Thermodynamic Decomposition**

To better understand the mechanism of Arg permeation across the membrane, various approaches were used to calculate contributions to free energy profiles. IE calculations between different components of the system, both direct (such as solvent-protein and lipid-protein) and indirect (such as solvent–lipid, lipid–lipid, solvent–solvent) revealed that the energetics of Arg side chain transfer across the bilayer is a balance of very large contributions that do not completely cancel and thus can explain some part of the free energy barrier. Direct energy contributions $U_D(z, \mathbf{R})$ lead to forces, acting on the reaction coordinate z, whereas indirect energy components $U_I(\mathbf{R})$ involve interactions between other atoms and are functions of their coordinates, $\mathbf{R}$, only. For instance, the reduction in the strength of direct interactions with solvating molecules near the membrane center was observed, which would contribute to the large free energy barrier. The indirect energy calculations revealed large strain energies across the bilayer core, which resist deformations extending to the membrane center. However, energy contributions account only for a fraction of the overall free energy profile. The other, entropic component was found to be equally important for shaping overall PMF profiles, owing to the cost of immobilizing water molecules and lipid head groups solvating Arg in the membrane (Li *et al.*, 2008a).

The large contribution from entropy tells us that one must compute free energy contributions and not just energetic values. Using the relationship with the mean force (Eq. (40)), we computed contributions from lipid head groups and water molecules that are located within the membrane core ($|z| \leq 13$ Å) as a function of z to understand how the deformed membrane is assisting in the stabilization of the Arg side chain in a process that is clearly distinct from previous theoretical models (Dorairaj and Allen, 2007). Figure 8.5 reveals that hydrocarbon core-located water molecules stabilize ArgH^+ by up to 35 kcal mol^{-1} near the membrane center (red curve) because water molecules are efficient in stabilizing charges away from the aqueous phase (Allen *et al.*, 2004a). Lipid head groups (blue curve) provide up to 5 kcal mol^{-1} stabilization near the outer core but contribute ~15 kcal mol^{-1} less near the membrane center resulting in net destabilization despite their large attractive interactions with the Arg side chain (Dorairaj and Allen, 2007), thus emphasizing the difference between energy and free energy components. This destabilization results from the significant strain energy owing to bilayer deformations necessary to allow Arg–lipid head group interactions (Dorairaj and Allen, 2007). This profile is interesting in that it suggests membrane proteins may leave the Arg side chains at the membrane interface and lipid phosphates will coordinate and stabilize the protein, but that such interactions when Arg is placed deeper in the membrane are clearly destabilizing and thus not favored. By subtracting these contributions from the total PMF, a "remainder" component was obtained, which exhibits a barrier of 36 kcal mol^{-1} and represents contributions from the bulk solvent and unperturbed lipid bilayer (Dorairaj and Allen, 2007; Li *et al.*, 2008a). This can be

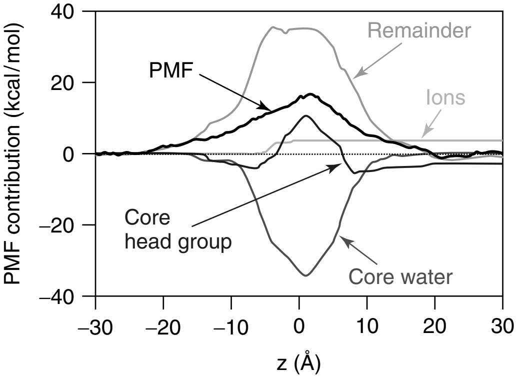

Fig. 8.5 Free energy decomposition for an Arg side chain across the membrane. The black curve is the PMF from Figure 8.4, the red curve is the contribution from water molecules inside the membrane core, the blue curve is from lipid head groups drawn into the core and the green curve is the remainder that makes up the total PMF. (adapted from (Li *et al.*, 2008a)). (Please find a color version of this figure on the color plates.)

viewed as the dielectric barrier of the unperturbed membrane some contribution from the bilayer dipole potential, as well as consistent with continuum estimates (Li *et al.*, 2008a; Vorobyov *et al.*, 2008). This calculation also demonstrates that without the membrane deformation the barrier for the charged species would be ~20 kcal mol^{-1} higher and would effectively eliminate the possibility of Arg carrying charge in a membrane.

8.5.2.6 **Accuracy of Force Fields for Arg–Membrane Interactions**

MD simulations using empirical all-atom CHARMM force field have been used to map out the thermodynamics of Arg side chain as a function of both position and protonation state across a membrane. However, since biomolecular force fields rely on empirical parameters, they are not expected to work in all cases, especially for situations not envisaged during original parameterization such as the modeling of a lipid-exposed ionizable Arg side chain. The accuracy of calculated free energy and $\mathrm{p}K_\mathrm{a}$ profiles was estimated by employing model systems to measure thermodynamics of Arg–membrane interactions using CHARMM force field and compare them to available experimental and QM data (Vorobyov *et al.*, 2008).

Solvation free energies of both charged and neutral Arg side chain analogs (propyl and methyl guanidinium and guanidine) in water and cyclohexane as well as partitioning free energies between these liquids were obtained by performing FEP simulations of these models in bulk solvents using *NPT* MD runs with PBC and in the gas phase using Langevin dynamics (Vorobyov *et al.*, 2008). The linear coupling scheme was used for electrostatic and dispersion free energy components, whereas the repulsive term was transformed into a soft-core potential and calculated in multiple stages using recently developed protocol (Deng and Roux, 2004). The calculated free energies were corrected for the dipole potentials of the corresponding interface (water–air, cyclohexane–air, and water–cyclohexane) and periodicity artifacts for the charged species (Vorobyov *et al.*, 2008). Calculated solvation and partitioning free energies were then compared to experimental values (if available) or estimates from continuum solvation models. CHARMM estimates of hydration free energies for both charged and neutral Arg side chain analogs were within ~1 kcal mol^{-1} of corresponding reference values (e.g. −62 kcal mol^{-1} for MGuanH^{+} and −11 kcal mol^{-1} for MGuan^{0}). Solvation free energies of neutral Arg side chain analogs in cyclohexane and their water–cyclohexane partitioning free energies are also in good agreement (within 1 kcal mol^{-1}) with experimental data (Vorobyov *et al.*, 2008). However, the lack of the electronic polarizability in this force field caused substantial underestimation of solvation free energies of the charged Arg side chain analogs in cyclohexane and thus overestimation of their unfavorable water–cyclohexane partitioning energetics by ~20 kcal mol^{-1} (Vorobyov *et al.*, 2008). FEP simulations of these ions using a recent Drude polarizable cyclohexane model (Vorobyov *et al.*, 2007) provided solvation and partitioning free energies close to values predicted by continuum solvation models, around −30 and 30 kcal mol^{-1}, respectively (Vorobyov *et al.*, 2008). However, by comparing relative energies of different system configurations with and without Drudes, this effect was found to be substantially smaller (on the order of 1 kcal mol^{-1}) for Arg side chain or analog transfer across the bilayer owing to their shielding from hydrocarbon lipid tails by water molecules and lipid head groups (Dorairaj and Allen, 2007; Vorobyov *et al.*, 2008).

To estimate the effect of polarizability on these PMF profiles directly we also carried out umbrella sampling MD simulations of MguanH^+ across the bilayer core with Drude particles on lipid hydrocarbon tails (Vorobyov *et al.*, 2008). These simulations were performed using 1 fs time step and the extended Lagrangian formalism by assigning small atomic masses of 0.4 amu to Drude particles and keeping them at low temperature of 1 K with a separate thermostat (Lamoureux and Roux, 2003). Two separate simulations for each window were run: starting with nonpolarizable CHARMM equilibrated configurations where bilayer deformations were already present or deformation-free initial structures. Interfacial connections with MguanH^+ in the membrane core (similar to those observed with a nonpolarizable membrane) were established and/or remained stable during the course of simulations. The resultant free energy profiles are very similar to that from the nonpolarizable CHARMM MguanH^+ simulation with the barrier lower by 1 kcal mol^{-1} (Vorobyov *et al.*, 2008). This confirms a surprisingly small effect of lipid polarizability on PMF despite much larger effect on water–hydrocarbon partitioning energetics and also substantially lower dipole potential for the polarizable lipid model (around 550 mV vs. 900 mV for the nonpolarizable lipid). The membrane deformations shield Arg side chain from interactions with lipid hydrocarbon and also prevents it from seeing the full extent of the dipole potential.

It was also necessary to determine how accurately CHARMM accounts for interactions of the Arg side chain with coordinating water molecules and lipid head groups, which drive the membrane deformations. This was done by comparing CHARMM and QM IEs for the corresponding clusters extracted from umbrella sampling simulations of ArgH^+ and Arg^0 models. Arg side chains were transformed into MGuanH^+ or MGuan^0, respectively, whereas lipid phosphate moiety was turned into dimethyl phosphate anion (DMPA). Water molecules within the first solvation shell of both MGuan and DMPA (within 3 Å) were kept. CHARMM IEs between Mguan, DMPA, and water were calculated using infinite cutoffs for nonbonded interactions. In QM calculations ten clusters for a pair of adjacent umbrella sampling windows were used with the most typical number of water molecules (six for charged and seven for neutral Arg) and CHARMM IEs closest to the average values. QM IEs used as reference values were calculated with Gaussian03 (Frisch, 2004) at the MP2/6-31+G(d) level and corrected for the basis set superposition error (BSSE) using the counterpoise technique (Boys and Bernardi, 1970). Both CHARMM and QM results were averaged for different regions of the membrane (interface, central and outer core) where Arg side chain was located. IEs between MguanH^+ and dehydrated DMPA are underestimated by CHARMM by ~7 kcal mol^{-1}, whereas those between MGuanH^+ or MGuan^0 and water clusters are overestimated by ~4 or ~5 kcal mol^{-1}, respectively. The interactions between MGuanH^+

or MGuan^0 and hydrated DMPA (i.e. DMPA/water clusters as one unit) are also made too attractive by CHARMM by ~6 kcal mol^{-1} (Vorobyov *et al.*, 2008). This is expected since nonpolarizable CHARMM force field was parameterized to reproduce condensed phase where ion–ion interactions are weakened and ion–solvent might get stronger compared to gas phase owing to cooperativity effects. Thus in nonpolarizable CHARMM, where these effect modeled implicitly, for example through enhanced dipole moments, the gas-phase behavior is inevitably sacrificed. The effect of these errors on PMF can be calculated using FEP formalism:

$$\Delta G = -kT \ln \left\langle e^{-\Delta U/kT} \right\rangle_{(\mathrm{CHARMM})} \quad (70)$$

where ΔU here is the difference between QM and CHARMM IEs for each configuration. For interactions between MguanH$^+$ or Mguan^0 and DMPA/water clusters, these ΔG and thus errors in free energy profiles are just 1.6 or 2.5 kcal mol^{-1}, respectively (i.e. CHARMM free energy barriers are too low), and are fairly constant across the membrane (Vorobyov *et al.*, 2008). The much smaller ΔG compared with $\langle \Delta U \rangle$ can be explained in terms of the stronger interactions giving rise to higher indirect energies and entropy costs. (Vorobyov *et al.*, 2008).

We also estimated that simulation artifacts related to the use of the finite bilayer patch (periodicity artifacts), fixed surface area, finite cutoffs for nonbonded interactions affect the accuracy of PMF profiles by around 0.5 kcal mol^{-1} or less (Vorobyov *et al.*, 2008). On the basis of all these tests it can be concluded that CHARMM free energy barriers for charged or neutral Arg side chain translocation across the membrane can be trusted to 1–2 kcal mol^{-1}, whereas pK_a shift profiles are accurate to within the order of 1 unit. (Vorobyov *et al.*, 2008).

8.6
Conclusion and Prospects

This chapter has outlined some of the basic principles of MD simulation of proteins. We have concentrated on modern force fields and algorithms for protein simulations as well as methods for quantitative analysis of protein activity. Experiment has provided a wealth of insight into the structure and function of both globular and membrane proteins. However, microscopic simulation, combined with rigorous statistical mechanical techniques, can assist in uncovering some microscopic mechanisms.

Though there exists an impressive range of protein simulations published today, we chose to illustrate just two related case studies involving membrane-bound ion channels, with the goal of elucidating the mechanisms by which proteins can control the passage of charged molecules across biological membranes. To better illustrate the utility of MD simulation we have made use of model peptides, which reduces the size and complexity of the problem while making the microscopic mechanisms of interest more general and transparent. We demonstrated how computer simulation can be used to generate configurational sampling of protein systems using biased sampling protocols to calculate equilibrium free energy surfaces governing protein function. Umbrella sampling and free energy perturbation are the basic tools of MD free energy simulation that permit a chosen spatial or alchemical reaction coordinate to be sampled in a reasonable simulation time. While there are several advanced simulation techniques, barely touched upon in this review, these methodologies all offer a direct route to understanding function and making contact with experimental observables. We have also seen how free energy surfaces along several choices of protein and solvent coordinates can reveal much about the mechanisms of function, as demonstrated here for ion and solvent contributions to a permeation process. In principle, however, any array of relevant order parameters can be examined in this manner.

One of the major focuses of this review has been the design and choice of MM force fields for proteins and membrane systems. Yet our illustrations concerned interactions for which the force fields were not specifically designed. That is, while protein and membrane force fields have undergone independent parameterizations, they can be combined successfully as a result of a common underlying paramterization strategy. We went on to show that for the most challenging of situations where a charged protein side chain enters the hydrocarbon core of a membrane, interactions and solvation free energies remained accurate to the order of kT. The key feature of the solvation environment of the amino acid side chain that led to this accuracy was the penetration of water molecules to maintain a nearly aqueous environment, because force fields were developed to reproduce such environments.

However, this does not mean there are no errors in present-day force fields. We have seen that the description of nonpolar molecules, such as the hydrocarbon interior of a lipid membrane, is fundamentally flawed owing to the lacking description of electronic polarizability. The description of electronic polarizability for nonpolar molecules as well as polar molecules, in particular around sites of high charge density, is currently an active area of research. The next generation of force fields will allow us to take MD simulations of proteins to even higher levels of accuracy and better reveal hidden mechanisms of function.

References

Adcock, S.A. and McCammon, J.A. (2006) *Chem. Rev.*, **106**, 1589–615.

Alden, R.G., Parson, W.W., Chu, Z.T., and Warshel, A. (1995) *J. Am. Chem. Soc.*, **117**, 12284–98.

Alder, B.J. and Wainwright, T.E. (1957) *J. Chem. Phys.*, **27**, 1208–9.

Alfredsson, M., Brodholt, J.P., Hermanson, K., and Vallauri, R. (1998) *Mol. Phys.*, **94**, 873–76.

Allen, T.W., Andersen, O.S., and Roux, B. (2003a) *J. Am. Chem. Soc.*, **125**, 9868–77.

Allen, T.W., Bastug, T., Kuyucak, S., and Chung, S.H. (2003b) *Biophys. J.*, **84**, 2159–68.

Allen, T.W., Andersen, O.S., and Roux, B. (2004a) *J. Gen. Physiol.*, **124**, 679–90.

Allen, T.W., Andersen, O.S., and Roux, B. (2004b) *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 117–22.

Allen, T.W., Andersen, O.S., and Roux, B. (2006a) *Biophys. Chem.*, **124**, 251–67.

Allen, T.W., Andersen, O.S., and Roux, B. (2006b) *Biophys. J.*, **90**, 3447–68.

Allen, T.W., Kuyucak, S., and Chung, S.H. (1999) *Biophys. J.*, **77**, 2502–16.

Allen, T.W., Kuyucak, S., and Chung, S.H. (2000) *Biophys. Chem.*, **86**, 1–14.

Allen, M. and Tildesley, D. (1987) *Computer Simulation of Liquids*, Clarendon Press, Oxford.

Allinger, N.L., Chen, K.S., and Lii, J.H. (1996) *J. Comput. Chem.*, **17**, 642–68.

Amadei, A., Linssen, A.B.M., and Berendsen, H.J.C. (1993) *Proteins: Struct. Funct. Genet.*, **17**, 412–25.

Amadei, A., Linssen, A.B.M., deGroot, B.L., vanAalten, D.M.F., and Berendsen, H.J.C. (1996) *J. Biomol. Struct. Dyn.*, **13**, 615–25.

Andersen, H.C. (1980) *J. Chem. Phys.*, **72**, 2384–93.

Andersen, H.C. (1983) *J. Comput. Phys.*, **52**, 24–34.

Andersen, O. and Koeppe, R. (1992) *Physiol. Rev.*, **72**, S89–S158.

Andersen, O.S., Koeppe II, R.E., and Roux, B. (2005) *IEEE Trans. Nanobiosci.*, **4**, 10–20.

Andreoni, W., Curioni, A., and Mordasini, T. (2001) *Ibm J. Res. Dev.*, **45**, 397–407.

Anezo, C., de Vries, A.H., Holtje, H.D., Tieleman, D.P., and Marrink, S.J. (2003) *J. Phys. Chem. B*, **107**, 9424–33.

Angyal, S.J. and Warburton, W.K. (1951) *J. Chem. Soc.*, 2492–94.

Anikin, N.A., Anisimov, V.M., Bugaenko, V.L., Bobrikov, V.V., and Andreyev, A.M. (2004) *J. Chem. Phys.*, **121**, 1266–70.

Anisimov, V.M., Lamoureux, G., Vorobyov, I.V., Huang, N., Roux, B., and MacKerell, A.D. (2005) *J. Chem. Theory Comput.*, **1**, 153–68.

Anisimov, V.M., Vorobyov, I.V., Roux, B., and MacKerell, A.D. (2007) *J. Chem. Theory Comput.*, **3**, 1927–46.

Åqvist, J. and Luzhkov, V. (2000) *Nature*, **404**, 881–84.

Arkhipov, A., Freddolino, P.L., Imada, K., Namba, K., and Schulten, K. (2006a) *Biophys. J.*, **91**, 4589–97.

Arkhipov, A., Freddolino, P.L., and Schulten, K. (2006b) *Structure*, **14**, 1767–77.

Arseniev, A.S., Lomize, A.L., Barsukov, I.L., and Bystrov, V.F. (1986) *Biol. Membr.*, **3**, 1077–104.

Ashcroft, F.M. (2000) *Ion Channels and Disease: Channelopathies*, Academic Press.

Ayton, G.S., Blood, P.D., and Voth, G.A. (2007) *Biophys. J.*, **92**, 3595–602.

Ayton, G.S., Izvekov, S., Noid, W.G., and Voth, G.A. (2008) Multiscale simulation of membranes and membrane proteins: Connecting molecular interactions to mesoscopic behavior, Computational Modeling of Membrane Bilayers, in *Current Topics in Membranes*, Vol. 60, Elsevier Academic Press, San Diego, pp. 111–30.

Bahar, I. and Rader, A.J. (2005) *Curr. Opin. Struct. Biol.*, **15**, 586–92.

Baker, N., Holst, M., and Wang, F. (2000) *J. Comput. Chem.*, **21**, 1343–52.

Baker, N.A., Sept, D., Joseph, S., Holst, M.J., and McCammon, J.A. (2001) *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 10037–41.

Banavali, N.K., Im, W., and Roux, B. (2002) *J. Chem. Phys.*, **117**, 7381–88.

Banks, J.L., Kaminski, G.A., Zhou, R.H., Mainz, D.T., Berne, B.J., and Friesner, R.A. (1999) *J. Chem. Phys.*, **110**, 741–54.

Baran, M. and Mazerski, J. (2002) *Chem. Phys. Lipids*, **120**, 21–31.

Barker, J.A. and Watts, R.O. (1973) *Mol. Phys.*, **26**, 789–92.

Barnes, P., Finney, J.L., Nicholas, J.D., and Quinn, J.E. (1979) *Nature*, **282**, 459–64.

Baron, R., Trzesniak, D., de Vries, A.H., Elsener, A., Marrink, S.J., and van Gunsteren, W.F. (2007) *Chemphyschem*, **8**, 452–61.

Bash, P.A., Field, M.J., and Karplus, M. (1987) *J. Am. Chem. Soc.*, **109**, 8092–94.

Bastug, T., Chen, P.C., Patra, S.M., and Kuyucak, S. (2008) *J. Chem. Phys.*, **128**, 155104.

Bastug, T. and Kuyucak, S. (2007) *Chem. Phys. Lett.*, **436**, 383–87.

Bayly, C.I., Cieplak, P., Cornell, W.D., and Kollman, P.A. (1993) *J. Phys. Chem.*, **97**, 10269–80.

Becker, O.M., MacKerell, A.D., Roux, B., and Watanabe, M. (2001) *Computational Biochemistry and Biophysics*, Marcel Dekker, New York.

Becker, O.M., Watanabe, M. (2001) Dynamics methods, in *Computational Biochemistry and Biophysics* (eds O.M. Becker, A.D. MacKerell, B. Roux, and M. Watanabe), Marcel Dekker, New York, p. xii, 512.

Beglov, D. and Roux, B. (1994) *J. Chem. Phys.*, **100**, 9050–63.

BenTal, N., BenShaul, A., Nicholls, A., and Honig, B. (1996) *Biophys. J.*, **70**, 1803–12.

Berendsen, H.J.C., Grigera, J.R., and Straatsma, T.P. (1987) *J. Phys. Chem.*, **91**, 6269–71.

Berendsen, H.J.C., Postma, J.P.M., Van Gunsteren, W.E., and Hermans, J. (1981) Interaction models for water in relation to protein hydration, in *Intermolecular Forces* (B. Pullman), Springer, Dordrecht, pp. 331–42.

Berendsen, H.J.C., Postma, J.P.M., Vangunsteren, W.F., Dinola, A., and Haak, J.R. (1984) *J. Chem. Phys.*, **81**, 3684–90.

Berendsen, H.J.C., Vanderspoel, D., and Vandrunen, R. (1995) *Comput. Phys. Commun.*, **91**, 43–56.

Berger, O., Edholm, O., and Jahnig, F. (1997) *Biophys. J.*, **72**, 2002–13.

Berkowitz, M. and Mc Cammon, J. (1982) *Chem. Phys. Lett.*, **90**, 215.

Bernal, J.D. and Fowler, R.H. (1933) *J. Chem. Phys.*, **1**, 515–48.

Berneche, S., Nina, M., and Roux, B. (1998) *Biophys. J.*, **75**, 1603–18.

Berneche, S. and Roux, B. (2000) *Biophys. J.*, **78**, 2900–17.

Berneche, S. and Roux, B. (2001) *Nature*, **414**, 73–77.

Berneche, S. and Roux, B. (2002) *Biophys. J.*, **82**, 772–80.

Biggin, P.C. and Bond, P.J. (2008) Membrane-associated proteins and peptides, in *Molecular Modeling of Proteins* (A. Kukol), Humana Press, Totowa, pp. 147–60.

Bogusz, S., Cheatham, T.E., and Brooks, B.R. (1998) *J. Chem. Phys.*, **108**, 7070–84.

Bolhuis, P.G., Dellago, C., and Chandler, D. (2000) *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 5877–82.

Bond, P.J., Holyoake, J., Ivetac, A., Khalid, S., and Sansom, M.S.P. (2007) *J. Struct. Biol.*, **157**, 593–605.

Bond, P.J. and Sansom, M.S. (2006) *J. Am. Chem. Soc.*, **128**, 2697–704.

Bond, P.J. and Sansom, M.S.P. (2007) *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 2631–36.

Bordner, A.J. and Huber, G.A. (2003) *J. Comput. Chem.*, **24**, 353–67.

Born, M. (1920) *Z. Phys.*, **1**, 45–48.

Borodin, O. and Smith, G.D. (2006) *J. Phys. Chem. B*, **110**, 6279–92.

Bostick, D.L. and Brooks III, C.L. (2007) *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 9260–65.

Boys, S.F. and Bernardi, F. (1970) *Mol. Phys.*, **19**, 553–66.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. (1983) *J. Comput. Chem.*, **4**, 187–217.

Brooks, C.L.I., Karplus, M., and Pettitt, B.M. (1988) *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*, Vol. 71, John Wiley & Sons, Inc., New York.

Bryce, R.A., Vincent, M.A., Malcolm, N.O.J., Hillier, I.H., and Burton, N.A. (1998) *J. Chem. Phys.*, **109**, 3077–85.

Burkert, U. and Allinger, N.L. (1982) *Molecular Mechanics*, American Chemical Society, Washington, DC.

Busath, D. (1993) *Annu. Rev. Physiol.*, **55**, 473–501.

Busath, D.D., Thulin, C.D., Hendershot, R.W., Phillips, L.R., Maughan, P., Cole, C.D., Bingham, N.C., Morrison, S., Baird, L.C., Hendershot, R.J., Cotten, M., and Cross, T.A. (2003) *Biophys. J.*, **75**, 2830–44.

Caldwell, J., Dang, L.X., and Kollman, P.A. (1990) *J. Am. Chem. Soc.*, **112**, 9144–47.

Caldwell, J.W. and Kollman, P.A. (1995a) *J. Phys. Chem.*, **99**, 6208–19.

Caldwell, J.W. and Kollman, P.A. (1995b) *J. Am. Chem. Soc.*, **117**, 4177–78.

Calimet, N., Schaefer, M., and Simonson, T. (2001) *Proteins: Struct. Funct. Genet.*, **45**, 144–58.

Car, R. and Parrinello, M. (1985) *Phys. Rev. Lett.*, **55**, 2471–74.

Case, D.A., Cheatham, T.E., Darden, T., Gohlke, H., Luo, R., Merz, K.M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R.J. (2005) *J. Comput. Chem.*, **26**, 1668–88.

Case, D.A. and Karplus, M. (1979) *J. Mol. Biol.*, **132**, 343–68.

Chakrabarti, N., Tajkhorshid, E., Roux, B., and Pomes, R. (2004) *Structure*, **12**, 65–74.

Chandrasekhar, I., Kastenholz, M., Lins, R.D., Oostenbrink, C., Schuler, L.D., Tieleman, D.P., and van Gunsteren, W.F. (2003) *Eur. Biophys. J. Biophys. Lett.*, **32**, 67–77.

Chang, G., Spencer, R., Lee, A., Barclay, M., and Rees, D. (1998) *Science*, **282**, 2220–26.

Chen, I.J., Yin, D.X., and MacKerell, A.D. (2002) *J. Comput. Chem.*, **23**, 199–213.

Chipot, C. and Henin, J. (2005) *J. Chem. Phys.*, **123**, 244906.

Chiu, S.W., Clark, M.M., Jakobsson, E., Subramaniam, S., and Scott, H.L. (1999a) *J. Phys. Chem. B*, **103**, 6323–27.

Chiu, S.W., Jakobsson, E., Subramaniam, S., and Scott, H.L. (1999b) *Biophys. J.*, **77**, 2462–69.

Chiu, S.W., Vasudevan, S., Jakobsson, E., Mashl, R.J., and Scott, H.L. (2003) *Biophys. J.*, **85**, 3624–35.

Christen, T., Hunenberger, P.H., Bakowies, D., Baron, R., Burgi, R., Geerke, D.P., Heinz, T.N., Kastenholz, M.A., Krautler, V., Oostenbrink, C., Peter, C., Trzesniak, D., and Van Gunsteren, W.F. (2005) *J. Comput. Chem.*, **26**, 1719–51.

Chocholousova, J. and Feig, M. (2006) *J. Phys. Chem. B*, **110**, 17240–51.

Cieplak, P., Cornell, W.D., Bayly, C., and Kollman, P.A. (1995) *J. Comput. Chem.*, **16**, 1357–77.

Clarke, R.J. (2001) *Adv. Colloid Interface Sci.*, **89–90**, 263–81.

Cooper, A. (1976) *Proc. Natl. Acad. Sci.*, **73**, 2740–41.

Cooper, K., Jakobsson, E., and Wolynes, P. (1985) *Prog. Biophys. Mol. Biol.*, **46**, 51–96.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. (1995) *J. Am. Chem. Soc.*, **117**, 5179–97.

Corry, B., Hoyles, M., Allen, T.W., Walker, M., Kuyucak, S., and Chung, S.-H. (2002) *Biophys. J.*, **82**, 1975–84.

Cortis, C.M. and Friesner, R.A. (1997) *J. Comput. Chem.*, **18**, 1591–608.

Cournia, Z., Smith, J.C., and Ullmann, G.M. (2005) *J. Comput. Chem.*, **26**, 1383–99.

Crooks, G.E. (1998) *J. Stat. Phys.*, **90**, 1481–87.

Crooks, G.E. (2000) *Phys. Rev. E*, **61**, 2361–66.

Crouzy, S., Woolf, T., and Roux, B. (1994) *Biophys. J.*, **67**, 1370–86.

Crowley, P.B. and Golovin, A. (2005) *Proteins*, **59**, 231–39.

Daggett, V. (2006) *Chem. Rev.*, **106**, 1898–916.

Dang, L.X. (1999) *J. Phys. Chem. B*, **103**, 8195–200.

Dang, L.X., Rice, J.E., Caldwell, J., and Kollman, P.A. (1991) *J. Am. Chem. Soc.*, **113**, 2481–86.

Darden, T. (2001) Treatment of long-range forces and potential, in *Computational Biochemistry and Biophysics* (eds O.M. Becker, A.D. MacKerell, B. Roux, M. Watanabe), Marcel Dekker, New York, p. xii, 512.

Darden, T., York, D., and Pedersen, L. (1993) *J. Chem. Phys.*, **98**, 10089–92.

Darve, E. and Pohorille, A. (2001) *J. Chem. Phys.*, **115**, 9169–83.

Darve, E., Wilson, M.A., and Pohorille, A. (2002) *Mol. Simul.*, **28**, 113–44.

Daura, X., Mark, A.E., and van Gunsteren, W.F. (1998) *J. Comput. Chem.*, **19**, 535–47.

Daura, X., Oliva, B., Querol, E., Aviles, F.X., and Tapia, O. (1996) *Proteins: Struct. Funct. Genet.*, **25**, 89–103.

Davis, M.E., Madura, J.D., Luty, B.A., and McCammon, J.A. (1991) *Comput. Phys. Commun.*, **62**, 187–97.

Davis, M.E. and McCammon, J.A. (1989) *J. Comput. Chem.*, **10**, 386–91.

Delarue, M. and Dumas, P. (2004) *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 6957–62.

Delhommelle, J. and Millie, P. (2001) *Mol. Phys.*, **99**, 619–25.

Deng, Y.Q. and Roux, B. (2004) *J. Phys. Chem. B*, **108**, 16567–76.

Deng, Y.Q. and Roux, B. (2006) *J. Chem. Theory Comput.*, **2**, 1255–73.

Deserno, M. and Holm, C. (1998) *J. Chem. Phys.*, **109**, 7678–93.

Ding, Y.B., Bernardo, D.N., Kroghjespersen, K., and Levy, R.M. (1995) *J. Phys. Chem.*, **99**, 11575–83.

Dixon, S.L. and Merz, K.M. (1997) *J. Chem. Phys.*, **107**, 879–93.

Dolan, E.A., Venable, R.M., Pastor, R.W., and Brooks, B.R. (2002) *Biophys. J.*, **82**, 2317–25.

Dominy, B.N. and Brooks, C.L. (1999) *J. Phys. Chem. B*, **103**, 3765–73.

Donchev, A.G., Galkin, N.G., Illarionov, A.A., Khoruzhii, O.V., Olevanov, M.A., Ozrin, V.D., Pereyaslavets, L.B., and Tarasov, V.I. (2008) *J. Comput. Chem.*, **29**, 1242–49.

Dorairaj, S. and Allen, T.W. (2007) *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 4943–48.

Dougherty, D.A. (1996) *Science*, **271**, 163–68.

Doyle, D., Cabral, J., Pfuetzner, R., Kuo, A., Gulbis, J., Cohen, S., Chait, B., and MacKinnon, R. (1998) *Science*, **280**, 69–77.

Dutzler, R., Campbell, E.B., Cadene, M., Chait, B.T., and MacKinnon, R. (2002) *Nature*, **415**, 287–94.

Eisenman, G. (1961) in *Symposium on Membrane Transport and Metabolism* (eds A. Kleinzeller and A. Kotyk), Academic Press, New York, pp. 163–79.

Eisenman, G. and Horn, R. (1983) *J. Membr. Biol.*, **76**, 197–225.

Essmann, U., Perera, L., Berkowitz, M.L., Darden, T., Lee, H., and Pedersen, L.G. (1995) *J. Chem. Phys.*, **103**, 8577–93.

Evans, D.J. and Murad, S. (1977) *Mol. Phys.*, **34**, 327–31.

Ewald, P. (1921) *Ann. Phys.*, **64**, 253–87.

Ewig, C.S., Berry, R., Dinur, U., Hill, J.R., Hwang, M.J., Li, H.Y., Liang, C., Maple, J., Peng, Z.W., Stockfisch, T.P., Thacher, T.S., Yan, L., Ni, X.S., and Hagler, A.T. (2001) *J. Comput. Chem.*, **22**, 1782–800.

Fan, H., Mark, A.E., Zhu, J., and Honig, B. (2005) *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 6760–64.

Feenstra, K.A., Hess, B., and Berendsen, H.J.C. (1999) *J. Comput. Chem.*, **20**, 786–98.

Feig, M. (2008) Implicit membrane models for membrane protein simulations, in *Molecular Modeling of Proteins* (A. Kukol), Humana Press, Totowa, p. xi, 390.

Feig, M., MacKerell, A.D., and Brooks, C.L. (2003) *J. Phys. Chem. B*, **107**, 2831–36.

Feig, M., Onufriev, A., Lee, M.S., Im, W., Case, D.A., and Brooks, C.L. (2004) *J. Comput. Chem.*, **25**, 265–84.

Feller, S.E., Gawrisch, K., and MacKerell, A.D. (2002) *J. Am. Chem. Soc.*, **124**, 318–26.

Feller, S.E. and MacKerell, A.D. (2000) *J. Phys. Chem. B*, **104**, 7510–15.

Feller, S.E. and Pastor, R.W. (1999) *J. Chem. Phys.*, **111**, 1281–87.

Feller, S.E., Pastor, R.W., Rojnuckarin, A., Bogusz, S., and Brooks, B.R. (1996) *J. Phys. Chem.*, **100**, 17011–20.

Feller, S.E., Yin, D.X., Pastor, R.W., and MacKerell, A.D. (1997) *Biophys. J.*, **73**, 2269–79.

Feller, S.E., Zhang, Y.H., Pastor, R.W., and Brooks, B.R. (1995a) *J. Chem. Phys.*, **103**, 4613–21.

Feller, S.E., Zhang, Y.H., and Pastor, R.W. (1995b) *J. Chem. Phys.*, **103**, 10267–76.

Ferguson, D.M. (1995) *J. Comput. Chem.*, **16**, 501–11.

Finkelstein, A. and Andersen, O.S. (1981) *J. Membr. Biol.*, **59**, 155–71.

Fogolari, F., Brigo, A., and Molinari, H. (2002) *J. Mol. Recognit.*, **15**, 377–92.

Foloppe, N. and MacKerell Jr, A.D. (2000) *J. Comput. Chem.*, **21**, 86–104.

Frauenfelder, H., Petsko, G.A., and Tsernoglou, D. (1979) *Nature*, **280**, 558–63.

Freddolino, P.L., Arkhipov, A.S., Larson, S.B., McPherson, A., and Schulten, K. (2006) *Structure*, **14**, 437–49.

Freddolino, P.L., Liu, F., Gruebele, M., and Schulten, K. (2008) *Biophys. J.*, **94**, L75–L77.

Freites, J.A., Tobias, D.J., von Heijne, G., and White, S.H. (2005) *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15059–64.

Frenkel, D. and Smit, B. (2002) *Understanding Molecular Simulation: from Algorithms to Applications*, 2nd edn, Academic Press, San Diego.

Friesner, R.A. (2006) Peptide modeling polarization in proteins and protein-ligand complexes: Methods and preliminary results, in *Advances in Protein Chemistry*, Vol. 72, Peptide Solvation and H-Bonds, Elsevier Academic Press, San Diego, pp. 79–104.

Frisch, M.J. (2004) *Gaussian 03, Revision C.02*, Gaussian, Wallingford.

Fukunishi, H., Watanabe, O., and Takada, S. (2002) *J. Chem. Phys.*, **116**, 9058–67.

Funahashi, J., Sugita, Y., Kitao, A., and Yutani, K. (2003) *Protein Eng.*, **16**, 665–71.

Gabel, F., Bicout, D., Lehnert, U., Tehei, M., Weik, M., and Zaccai, G. (2002) *Q. Rev. Biophys.*, **35**, 327–67.

Gao, M., Craig, D., Lequin, O., Campbell, I.D., Vogel, V., and Schulten, K. (2003) *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 14784–89.

Gao, J.L., Habibollazadeh, D., and Shao, L. (1995) *J. Phys. Chem.*, **99**, 16460–67.

Gao, J., Major, T.D., Fan, Y., Lin, Y.-I., Ma, S., and Wong, K.-Y. (2008) Hybrid quantum and classical methods for computing kinetic isotope effects of chemical reactions in solutions and in enzymes, in *Molecular Modeling of Proteins* (A. Kukol), Humana Press, Totowa, p. xi, 390.

Garcia, A.E. (1992) *Phys. Rev. Lett.*, **68**, 2696–99.

Gear, C.W. (1971) *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs.

Geerke, D.P. and Van Gunsteren, W.F. (2007) *Mol. Phys.*, **105**, 1861–81.

Geney, R., Layten, M., Gomperts, R., Hornak, V., and Simmerling, C. (2006) *J. Chem. Theory Comput.*, **2**, 115–27.

Giese, T.J. and York, D.M. (2004) *J. Chem. Phys.*, **120**, 9903–6.

Gilson, M.K. and Honig, B.H. (1987) *Nature*, **330**, 84–86.

Go, N. and Taketomi, H. (1978) *Proc. Natl. Acad. Sci. U.S.A.*, **75**, 559–63.

Goetz, R. and Lipowsky, R. (1998) *J. Chem. Phys.*, **108**, 7397–409.

Goodford, P.J. (1985) *J. Med. Chem.*, **28**, 849–57.

Grabe, M., Lecar, H., Jan, Y.N., and Jan, L.Y. (2004) *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 17640–45.

Grayson, P., Tajkhorshid, E., and Schulten, K. (2003) *Biophys. J.*, **85**, 36–48.

Greengard, L. and Rokhlin, V. (1987) *J. Comput. Phys.*, **73**, 325–48.

Grossfield, A., Feller, S.E., and Pitman, M.C. (2006) *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 4888–93.

Grossfield, A., Ren, P.Y., and Ponder, J.W. (2003) *J. Am. Chem. Soc.*, **125**, 15671–82.

Grossfield, A., Sachs, J., and Woolf, T.B. (2000) *Proteins: Struct. Funct. Genet.*, **41**, 211–23.

Grubmuller, H. (1995) *Phys. Rev. E*, **52**, 2893–906.

Guidoni, L., Torre, V., and Carloni, P. (1999) *Biochemistry*, **38**, 8599–604.

Gullingsrud, J. and Schulten, K. (2003) *Biophys. J.*, **85**, 2087–99.

Gumbart, J. and Schulten, K. (2006) *Biophys. J.*, **90**, 2356–67.

van Gunsteren, W.F., Billeter, S.R., Eising, A.A., Hünenberger, P.H., Krüger, P., Mark, A.E., Scott, W.R.P., and Tironi, I.G. (1996) *Biomolecular Simulation: The GROMOS96 Manual and User Guide*, Hochschuleverlag AG an der ETH Zürich, Zürich.

Guvench, O. and MacKerell Jr, A.D. (2008) Comparison of protein force fields for molecular dynamics simulations, in *Molecular Modeling of Proteins* (A. Kukol), Humana Press, Totowa, p. xi, 390.

Halgren, T.A. (1996) *J. Comput. Chem.*, **17**, 490–519.

Halgren, T.A. and Damm, W. (2001) *Curr. Opin. Struct. Biol.*, **11**, 236–42.

Hall, N.F. and Sprinkle, M.R. (1932) *J. Am. Chem. Soc.*, **54**, 3469–85.

Harder, E., Anisimov, V.M., Vorobyov, I.V., Lopes, P.E.M., Noskov, S.Y., MacKerell, A.D., and Roux, B. (2006) *J. Chem. Theory Comput.*, **2**, 1587–97.

Harder, E., Anisimov, V.M., Whitfield, T.W., MacKerell, A.D., and Roux, B. (2008) *J. Phys. Chem. B*, **112**, 3509–21.

Harder, E., Kim, B.C., Friesner, R.A. and Berne, B.J. (2005) *J. Chem. Theory Comput.*, **1**, 169–80.

Hayward, S., de Groot, B.L. (2008) Normal modes and essential dynamics, in *Molecular Modeling of Proteins* (A. Kukol), pp. 89–106.

Henin, J. and Chipot, C. (2004) *J. Chem. Phys.*, **121**, 2904–14.

Hermida-Ramon, J.M. and Rios, M.A. (1998) *J. Phys. Chem. A*, **102**, 10818–27.

Hermida-Ramon, J.M. and Rios, M.A. (2000) *Chem. Phys.*, **262**, 423–36.

Hess, B. (2000) *Phys. Rev. E*, **62**, 8438–48.

Hess, B., Bekker, H., Berendsen, H.J.C., and Fraaije, J. (1997) *J. Comput. Chem.*, **18**, 1463–72.

Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S.H., and von Heijne, G. (2005) *Nature*, **433**, 377–81.

Hessa, T., Kim, H., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S.H., and von Heijne, G. (2005a) *Nature*, **433**, 377–81.

Hessa, T., White, S.H., and von Heijne, G. (2005b) *Science*, **307**, 1427.

Hille, B. (1973) *J. Gen. Physiol.*, **61**, 669–86.

Hille, B. (2001) *Ionic Channels of Excitable Membranes*, Sinauer, Sunderland.

Hinton, J.F., Whaley, W.L., Shungu, D.C., Koeppe II, R.E., and Millett, F.S. (1986) *Biophys. J.*, **50**, 539–44.

Hladky, S.B., Urban, B., and Haydon, D. (1979) *Membrane Transport Processes*, Vol. 3, pp. 89–103.

Hockney, R.W. (1970) *Methods Comput. Phys.*, **9**, 136–211.

Hockney, R.W., Eastwood, J.W. (1988) *Computer Simulation using Particles*, Special student ed., Adam Hilger, Bristol, Philadelphia.

Hodgkin, A. and Huxley, A. (1952) *J. Physiol. (Lond.)*, **117**, 500–44.

Hodgkin, A. and Katz, B. (1949) *J. Physiol. (Lond.)*, **108**, 37–77.

Holst, M., Baker, N., and Wang, F. (2000) *J. Comput. Chem.*, **21**, 1319–42.

Holst, M.J. and Saied, F. (1995) *J. Comput. Chem.*, **16**, 337–64.

Honig, B.L. and Hubbell, W.L. (1984) *Proc. Natl. Acad. Sci. U.S.A.*, **81**, 5412–16.

Honig, B. and Nicholls, A. (1995) *Science*, **268**, 1144–49.

Hoover, W.G. (1985) *Phys. Rev. A*, **31**, 1695–97.

Hu, H., Elstner, M., and Hermans, J. (2003) *Proteins: Struct. Funct. Genet.*, **50**, 451–63.

Hummer, G., Pratt, L.R., and Garcia, A.E. (1996) *J. Phys. Chem.*, **100**, 1206–15.

Hummer, G. and Szabo, A. (2001) *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 3658–61.

Hunenberger, P. (2005) *Advanced Computer Simulation Approaches for Soft Matter Sciences I*, Vol. 173, pp. 105–47.

Hunenberger, P.H. and McCammon, J.A. (1999a) *J. Chem. Phys.*, **110**, 1856–72.

Hunenberger, P.H. and McCammon, J.A. (1999b) *Biophys. Chem.*, **78**, 69–88.

Im, W., Beglov, D., and Roux, B. (1998) *Comput. Phys. Commun.*, **111**, 59–75.

Im, W., Berneche, S., and Roux, B. (2001) *J. Chem. Phys.*, **114**, 2924–37.

Im, W., Feig, M., and Brooks, C.L. (2003) *Biophys. J.*, **85**, 2900–18.

Im, W. and Roux, B. (2002) *J. Mol. Biol.*, **322**, 851–69.

Israelachvili, J.N., Mitchell, D.J., and Ninham, B.W. (1977) *Biochim. Biophys. Acta*, **470**, 185–201.

Israelachvili, J. and Wennerström, H. (1996) *Nature*, **379**, 219–25.

Isralewitz, B., Baudry, J., Gullingsrud, J., Kosztin, D., and Schulten, K. (2001a) *J. Mol. Graph. Model.*, **19**, 13–25.

Isralewitz, B., Gao, M., and Schulten, K. (2001b) *Curr. Opin. Struct. Biol.*, **11**, 224–30.

Isralewitz, B., Izrailev, S., and Schulten, K. (1997) *Biophys. J.*, **73**, 2972–79.

Izvekov, S. and Voth, G.A. (2005) *J. Chem. Phys.*, **123**, 134105.

Jackson, J.D. (1999) *Classical Electrodynamics*, 3rd edn, John Wiley & Sons, Inc., New York.

Jahnig, F. (1996) *Biophys. J.*, **71**, 1348–49.

Jarzynski, C. (1997) *Phys. Rev. Lett.*, **78**, 2690–93.

Jayachandran, G., Vishal, V., and Pande, V.S. (2006) *J. Chem. Phys.*, **124**, 164902.

Jiang, Y., Lee, A., Chen, J., Cadene, M., Chait, B., and MacKinnon, R. (2002) *Nature*, **417**, 523–26.

Jiang, Y., Lee, A., Chen, J.Y., Ruta, V., Cadene, M., Chait, B.T., and MacKinnon, R. (2003a) *Nature*, **423**, 33–41.

Jiang, Y., Ruta, V., Chen, J., Lee, A., and MacKinnon, R. (2003b) *Nature*, **423**, 42–48.

Jiao, D., Golubkov, P.A., Darden, T.A., and Ren, P. (2008) *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 6290–95.

Jing, N., Prasad, K.U., and Urry, D.W. (1995) *Biochim. Biophys. Acta*, **1238**, 1–11.

Jo, S., Vargyas, M., Vasko-Szedlar, J., Roux, B., and Im, W. (2008) *Nucleic Acids Res.*, **36**, W270–W275.

Johansson, A.C.V. and Lindahl, E. (2006) *Biophys. J.*, **91**, 4450–63.

Johansson, A.C.V. and Lindahl, E. (2008) *Proteins-Struct. Funct. Bioinformatics*, **70**, 1332–44.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. (1983) *J. Chem. Phys.*, **79**, 926–35.

Jorgensen, W.L., Maxwell, D.S., and Tirado-Rives, J. (1996) *J. Am. Chem. Soc.*, **118**, 11225–36.

Jorgensen, W.L. and Tiradorives, J. (1988) *J. Am. Chem. Soc.*, **110**, 1657–66.

Jorgensen, W.L. and Tiradorives, J. (2005) *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 6665–70.

Kale, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K., and Schulten, K. (1999) *J. Comput. Phys.*, **151**, 283–312.

Kaminski, G.A., Friesner, R.A., Tirado-Rives, J., and Jorgensen, W.L. (2001) *J. Phys. Chem. B*, **105**, 6474–87.

Kaminski, G.A., Stern, H.A., Berne, B.J., and Friesner, R.A. (2004) *J. Phys. Chem. A*, **108**, 621–27.

Kaminski, G.A., Stern, H.A., Berne, B.J., Friesner, R.A., Cao, Y.X.X., Murphy, R.B., Zhou, R.H., and Halgren, T.A. (2002) *J. Comput. Chem.*, **23**, 1515–31.

Karplus, M. and McCammon, J. (1981) *CRC Crit. Rev. Biochem.*, **9**, 293–349.

Karplus, M. and McCammon, J.A. (2002) *Nat. Struct. Biol.*, **9**, 646–52.

Karplus, M. and Petsko, G.A. (1990) *Nature*, **347**, 631–39.

Kessel, A. and Ben-Tal, N. (2002) *Curr. Top. Membr.*, **52**, 205–53.

Kessel, A., Cafiso, D.S., and Ben-Tal, N. (2000) *Biophys. J.*, **78**, 571–83.

Ketchem, R.R., Roux, B., and Cross, T.A. (1997) *Structure*, **5**, 11655–69.

Khoruzhii, O., Donchev, A.G., Galkin, N., Illarionov, A., Olevanov, M., Ozrin, V., Queen, C., and Tarasov, V. (2008) *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 10378–83.

Kim, B.C., Young, T., Harder, E., Friesner, R.A., and Berne, B.J. (2005) *J. Phys. Chem. B*, **109**, 16529–38.

Kirkwood, J. (1935) *J. Chem. Phys.*, **3**, 300.

Kitao, A. and Go, N. (1999) *Curr. Opin. Struct. Biol.*, **9**, 164–69.

Kitao, A., Hirata, F., and Go, N. (1991) *Chem. Phys.*, **158**, 447–72.

Kitao, A., Yonekura, K., Maki-Yonekura, S., Samatey, F.A., Imada, K., Namba, K., and Go, N. (2006) *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 4894–99.

Klapper, I., Hagstrom, R., Fine, R., Sharp, K., and Honig, B. (1986) *Proteins: Struct. Funct. Genet.*, **1**, 47–59.

Klauda, J.B., Brooks, B.R., MacKerell Jr, A.D., Venable, R.M., and Pastor, R.W. (2005) *J. Phys. Chem. B*, **109**, 5300–11.

Klauda, J.B., Venable, R.M., MacKerell, A.D., and Pastor, R.W. (2008) Considerations for lipid force field development, in *Computational Modeling of Membrane Bilayers*, Vol. 60, Elsevier Academic Press, San Diego, pp. 1–48.

Klauda, J.B., Wu, X.W., Pastor, R.W., and Brooks, B.R. (2007) *J. Phys. Chem. B*, **111**, 4393–400.

Kollman, P. (1993) *Chem. Rev.*, **93**, 2395–417.

Kong, C.L. and Chakrabarty, M.R. (1973) *J. Phys. Chem.*, **77**, 2668–70.

Kong, Y. and Karplus, M. (2007) *Structure*, **15**, 611–23.

Kosztin, I., Barz, B., and Janosi, L. (2006) *J. Chem. Phys.*, **124**, 064106.

Krammer, A., Lu, H., Isralewitz, B., Schulten, K., and Vogel, V. (1999) *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 1351–56.

Kukol, A. (2008) *Molecular Modeling of Proteins*, Humana Press, Totowa.

Kumar, S., Bouzida, D., Swendsen, R.H., Kollman, P.A., and Rosenberg, J.M. (1992) *J. Comput. Chem.*, **13**, 1011–21.

Lague, P., Pastor, R., and Brooks, B. (2004) *J. Phys. Chem. B*, **108**, 363–68.

Lamoureux, G., Harder, E., Vorobyov, I.V., Roux, B., and MacKerell, A.D. (2006) *Chem. Phys. Lett.*, **418**, 245–49.

Lamoureux, G., MacKerell, A.D., and Roux, B. (2003) *J. Chem. Phys.*, **119**, 5185–97.

Lamoureux, G. and Roux, B. (2003) *J. Chem. Phys.*, **119**, 3025–39.

Lamoureux, G. and Roux, B. (2006) *J. Phys. Chem. B*, **110**, 3308–22.

Lawrence, C.P. and Skinner, J.L. (2003) *Chem. Phys. Lett.*, **372**, 842–47.

Leach, A.R. (2001) *Molecular Modelling: Principles and Applications*, Pearson Education Limited, Harlow.

Lee, M.S., Feig, M., Salsbury, F.R., and Brooks, C.L. (2003) *J. Comput. Chem.*, **24**, 1348–56.

Lee, T.S., York, D.M., and Yang, W.T. (1996) *J. Chem. Phys.*, **105**, 2744–50.

Lei, H. and Duan, Y. (2008) Protein foldng and unfolding by all-atom molecular dynamics simulations, in *Molecular Modeling of Proteins* (ed A. Kukol) Humana Press, Totowa, pp. 277–96.

Leung, K., Remper, S.B., and Lorenz, C.D. (2006) *Phys. Rev. Lett.*, **96**, 095504.

Levitt, D. (1986) *Annu. Rev. Biophys. Biophys. Chem.*, **15**, 29–57.

Levitt, M. and Warshel, A. (1975) *Nature*, **253**, 694–98.

Li, L., Vorobyov, I., and Allen, T.W. (2008a) *J. Phys. Chem. B*, **112**, 9574–87.

Li, L., Vorobyov, I., MacKerell, A.D., and Allen, T.W. (2008b) *Biophys. J.*, **94**, L11–L13.

Liang, T. and Walsh, T.R. (2007) *Mol. Simul.*, **33**, 337–42.

Lii, J.H. and Allinger, N.L. (1991) *J. Comput. Chem.*, **12**, 186–99.

Lindahl, E., Hess, B., and van der Spoel, D. (2001) *J. Mol. Model.*, **7**, 306–17.

Lindahl, E. and Sansom, M.S. (2008) *Curr. Opin. Struct. Biol.*, **18**, 425–31.

Lingenheil, M., Denschlag, R., Reichold, R., and Tavan, P. (2008) *J. Chem. Theory Comput.*, **4**, 1293–306.

Liphardt, J., Dumont, S., Smith, S.B., Tinoco, J.I., and Bustamante, C. (2002) *Science*, **296**, 1832–35.

Liu, Z., Xu, Y., and Tang, P. (2006) *J. Phys. Chem. B*, **110**, 12789–95.

Liwo, A., Khalili, M., and Scheraga, H.A. (2005) *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 2362–67.

Liwo, A., Oldziej, S., Pincus, M.R., Wawak, R.J., Rackovsky, S., and Scheraga, H.A. (1997) *J. Comput. Chem.*, **18**, 849–73.

Lopes, P.E.M., Lamoureux, G., Roux, B., and MacKerell, A.D. (2007) *J. Phys. Chem. B*, **111**, 2873–85.

Lu, B.Z., Chen, W.Z., Wang, C.X., and Xu, X.J. (2002) *Proteins: Struct. Funct. Genet.*, **48**, 497–504.

Lu, H., Isralewitz, B., Krammer, A., Vogel, V., and Schulten, K. (1998) *Biophys. J.*, **75**, 662–71.

Lu, N., Kofke, D.A., and Woolf, T.B. (2003) *J. Phys. Chem. B*, **107**, 5598–611.

Lundbaek, J.A., Maer, A.M., and Andersen, O.S. (1997) *Biochemistry*, **36**, 5695–701.

Luo, R., David, L., and Gilson, M.K. (2002) *J. Comput. Chem.*, **23**, 1244–53.

Luty, B.A., Davis, M.E., Tironi, I.G., and Vangunsteren, W.F. (1994) *Mol. Simul.*, **14**, 11–20.

Luzhkov, V. and Åqvist, J. (2001) *Biochim. Biophys. Acta*, **1548**, 194–202.

MacCallum, J.L., Bennett, W.F.D., and Tieleman, D.P. (2007) *J. Gen. Physiol.*, **129**, 371–77.

MacCallum, J.L., Bennett, W.F.D., and Tieleman, D.P. (2008) *Biophys. J.*, **94**, 3393–404.

Maccallum, J.L. and Tieleman, D.P. (2003) *J. Comput. Chem.*, **24**, 1930–35.

Mackay, D.H.J., Berens, P.H., Wilson, K.R., and Hagler, A.T. (1984) *Biophys. J.*, **46**, 229–48.

MacKerell, A.D. (1998) Protein force fields, in *Encyclopedia of Computational Chemistry* (P.v.R. Schleyer), John Wiley & Sons, Inc., New York, Chichester, p. 5, xxix, 3429.

MacKerell, A.D. (2001) Atomistic models and force fields, in *Computational Biochemistry and Biophysics* (eds O.M. Becker, A.D. MacKerell, B. Roux, and M. Watanabe), Marcel Dekker, New York, p. xii, 512.

MacKerell, A.D. (2004) *J. Comput. Chem.*, **25**, 1584–604.

MacKerell, J.A.D., Brooks, B., Brooks, I.C.L., Nilsson, L., Roux, B., Won, Y., and Karplus, M. (1998a) CHARMM: the energy function and its paramerization with an overview of the program, in *Encyclopedia of Computational Chemistry*, Vol. 1 (eds P. Schleyer, N. Allinger, T. Clark, J. Gasteiger, P. Kollman, H.F. Schaefer III, and P. Schreiner), John Wiley & Sons, Inc., Chichester, pp. 271–77.

MacKerell, A.D. (1998b) *J. Phys. Chem. B*, **102**, 3586–616.

MacKerell, A.D., Feig, M., and Brooks, C.L. (2004) *J. Comput. Chem.*, **25**, 1400–15.

MacKerell, A.D. and Nilsson, L. (2008) *Curr. Opin. Struct. Biol.*, **18**, 194–99.

MacKerell, A.D., Wiorkiewiczkuczera, J., and Karplus, M. (1995) *J. Am. Chem. Soc.*, **117**, 11946–75.

Mahoney, M.W. and Jorgensen, W.L. (2000) *J. Chem. Phys.*, **112**, 8910–22.

Mamanov, A.B, Kurnikova, M.G. and Coalson, R.D. (2006) *Biophys. Chem.*, **124**, 268–78.

Mannfors, B., Mirkin, N.G., Palmo, K., and Krimm, S. (2001) *J. Comput. Chem.*, **22**, 1933–43.

Mannfors, B., Palmo, K., and Krimm, S. (2000) *J. Mol. Struct.*, **556**, 1–21.

Maple, J.R., Cao, Y.X., Damm, W.G., Halgren, T.A., Kaminski, G.A., Zhang, L.Y., and Friesner, R.A. (2005) *J. Chem. Theory Comput.*, **1**, 694–715.

Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P., and de Vries, A.H. (2007) *J. Phys. Chem. B*, **111**, 7812–24.

Marrink, S.J., de Vries, A.H., and Mark, A.E. (2004) *J. Phys. Chem. B*, **108**, 750–60.

Martyna, G.J., Tuckerman, M.E., Tobias, D.J., and Klein, M.L. (1996) *Mol. Phys.*, **87**, 1117–57.

Masella, M., Borgis, D., and Cuniasse, P. (2008) *J. Comput. Chem.*, **29**, 1707–24.

Maupetit, J., Tuffery, P., and Derreumaux, P. (2007) *Proteins-Struct. Funct. Bioinformatics*, **69**, 394–408.

Mayo, S.L., Olafson, B.D., and Goddard, W.A. (1990) *J. Phys. Chem.*, **94**, 8897–909.

McCammon, J.A., Gelin, B.R., and Karplus, M. (1977) *Nature*, **267**, 585–90.

McQuarrie, D.A. (2000) *Statistical Mechanics*, University Science Books, Sausalito.

Meirovitch, H. (2007) *Curr. Opin. Struct. Biol.*, **17**, 181–86.

Meng, E.C., Caldwell, J.W., and Kollman, P.A. (1996) *J. Phys. Chem.*, **100**, 2367–71.

Meng, E.C., Cieplak, P., Caldwell, J.W., and Kollman, P.A. (1994) *J. Am. Chem. Soc.*, **116**, 12061–62.

Merz, K.M. (1991) *J. Am. Chem. Soc.*, **113**, 3572–75.

Merz Jr, K.M. (1997) *Curr. Opin. Struct. Biol.*, **7**, 511–17.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953) *J. Chem. Phys.*, **21**, 1087–92.

Miller, T.F., Eleftheriou, M., Pattnaik, P., Ndirango, A., Newns, D., and Martyna, G.J. (2002) *J. Chem. Phys.*, **116**, 8649–59.

Miyamoto, S. and Kollman, P.A. (1992) *J. Comput. Chem.*, **13**, 952–62.

Miyazawa, A., Fujiyoshi, Y., and Unwin, N. (2003) *Nature*, **423**, 949–55.

Miyazawa, S. and Jernigan, R.L. (1985) *Macromolecules*, **18**, 534–52.

Monticelli, L., Kandasamy, S.K., Periole, X., Larson, R.G., Tieleman, D.P., and Marrink, S.J. (2008) *J. Chem. Theory Comput.*, **4**, 819–34.

Monticelli, L., Tieleman, D.P., and Colombo, G. (2005) *J. Phys. Chem. B*, **109**, 20064–67.

Mori, H. (1965) *Prog. Theor. Phys.*, **33**, 423–55.

Mukhopadhyay, P., Monticelli, L., and Tieleman, D.P. (2004) *Biophys. J.*, **86**, 1601–9.

Muller, M., Katsov, K., and Schick, M. (2006) *Phys. Rep.-Rev. Sect. Phys. Lett.*, **434**, 113–76.

Mullins, L.J. (1960) *J. Gen. Physiol.*, **43**, 105–17.

Murray, D., BenTal, N., Honig, B., and McLaughlin, S. (1997) *Structure*, **5**, 985–89.

Murtola, T., Falck, E., Patra, M., Karttunen, M., and Vattulainen, I. (2004) *J. Chem. Phys.*, **121**, 9156–65.

Nagle, J. (1993) *Biophys. J.*, **64**, 1476–81.

Nagle, J.F. and Tristram-Nagle, S. (2000) *Biochim. Biophys. Acta, Rev. Biomembr.*, **1469**, 159–95.

Nelson, M.T., Humphrey, W., Gursoy, A., Dalke, A., Kale, L.V., Skeel, R.D., and Schulten, K. (1996) *Int. J. Supercomput. Appl. High Perform. Comput.*, **10**, 251–68.

Neria, E., Fischer, S., and Karplus, M. (1996) *J. Chem. Phys.*, **105**, 1902–21.

Nielsen, S.O., Lopez, C.F., Srinivas, G., and Klein, M.L. (2004) *J. Phys.: Condens. Matter*, **16**, R481–R512.

Nina, M., Beglov, D., and Roux, B. (1997) *J. Phys. Chem. B*, **101**, 5239–48.

Noid, W.G., Chu, J.W., Ayton, G.S., Krishna, V., Izvekov, S., Voth, G.A., Das, A., and Andersen, H.C. (2008a) *J. Chem. Phys.*, **128**, 244114.

Noid, W.G., Liu, P., Wang, Y., Chu, J.W., Ayton, G.S., Izvekov, S., Andersen, H.C., and Voth, G.A. (2008b) *J. Chem. Phys.*, **128**, 244115.

Northrup, S.H., Pear, M.R., Lee, C.Y., McCammon, J.A., and Karplus, M. (1982) *Proc. Natl. Acad. Sci. U.S.A.*, **79**, 4035–39.

Nose, S. (1984) *Mol. Phys.*, **52**, 255–68.

Nose, S. and Klein, M.L. (1983) *Mol. Phys.*, **50**, 1055–76.

Noskov, S.Y., Berneche, S., and Roux, B. (2004) *Nature*, **431**, 830–34.

Noskov, S.Y., Lamoureux, G., and Roux, B. (2005) *J. Phys. Chem. B*, **109**, 6705–13.

Nozaki, Y. and Tanford, C. (1967) *Methods Enzymol.*, **11**, 715–34.

Nummela, J., Yassin, F., and Andricioaei, I. (2008) *J. Chem. Phys.*, **128**, 024104.

Okur, A., Strockbine, B., Hornak, V., and Simmerling, C. (2003) *J. Comput. Chem.*, **24**, 21–31.

Olah, G.A., Huang, H.W., Liu, W., and Wu, Y. (1991) *J. Mol. Biol.*, **218**, 847.

Oldziej, S. (2005) *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 7547–52.

Onufriev, A., Bashford, D., and Case, D.A. (2000) *J. Phys. Chem. B*, **104**, 3712–20.

Onufriev, A., Bashford, D., and Case, D.A. (2004) *Proteins-Struct. Funct. Bioinformatics*, **55**, 383–94.

Oostenbrink, C., Villa, A., Mark, A.E., and Van Gunsteren, W.F. (2004) *J. Comput. Chem.*, **25**, 1656–76.

Orsi, M., Haubertin, D.Y., Sanderson, W.E., and Essex, J.W. (2008) *J. Phys. Chem. B*, **112**, 802–15.

Orttung, W.H. (1977) *Ann. N. Y. Acad. Sci.*, **303**, 22–37.

Paci, E. and Karplus, M. (2000) *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 6521–26.

Palmo, K., Mannfors, B., Mirkin, N.G., and Krimm, S. (2003) *Biopolymers*, **68**, 383–94.

Pan, Y. and Daggett, V. (2001) *Biochemistry*, **40**, 2723–31.

Parak, F. and Knapp, E.W. (1984) *Proc. Natl. Acad. Sci.*, **81**, 7088–92.

Park, H. and Lee, S. (2005) *Biophys. Chem.*, **114**, 191–97.

Park, S. and Schulten, K. (2004) *J. Chem. Phys.*, **120**, 5946–61.

Parrinello, M. and Rahman, A. (1980) *Phys. Rev. Lett.*, **45**, 1196–99.

Parrinello, M. and Rahman, A. (1981) *J. Appl. Phys.*, **52**, 7182–90.

Parsegian, A. (1969) *Nature*, **221**, 844–46.

Patel, S. and Brooks, C.L. (2004) *J. Comput. Chem.*, **25**, 1–15.

Patel, S. and Brooks, C.L. (2005) *J. Chem. Phys.*, **122**, 024508.

Patel, S. and Brooks, C.L. (2005a) *J. Chem. Phys.*, **123**, 164502.

Patel, S. and Brooks, C.L. (2006) *Mol. Simul.*, **32**, 231–49.

Patel, S.A. and Brooks, C.L. (2006a) *J.Chem. Phys.*, **124**, 204–706.

Patel, S., Mackerell, A.D., and Brooks, C.L. (2004) *J. Comput. Chem.*, **25**, 1504–14.

Pearlman, D.A., Case, D.A., Caldwell, J.W., Ross, W.S., Cheatham, T.E., Debolt, S., Ferguson, D., Seibel, G., and Kollman, P. (1995) *Comput. Phys. Commun.*, **91**, 1–41.

Perutz, M.F. and Mathews, F. (1966) *J. Mol. Biol.*, **21**, 199–202.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L., and Schulten, K. (2005) *J. Comput. Chem.*, **26**, 1781–802.

Plimpton, S. (1995) *J. Comput. Phys.*, **117**, 1–19.

Pomes, R. and Roux, B. (1998) *Biophys. J.*, **75**, 33–40.

Ponder, J.W. and Case, D.A. (2003) Force fields for protein simulations, *Protein Simulations*, Vol. 66, p. 27.

Prabhu, N.V., Zhu, P.J., and Sharp, K.A. (2004) *J. Comput. Chem.*, **25**, 2049–64.

Praprotnik, M., Delle Site, L., and Kremer, K. (2008) *Annu. Rev. Phys. Chem.*, **59**, 545–71.

Rahman, A. (1964) *Phys. Rev. A Gen. Phys.*, **136**, A405–A411.

Rahman, A. and Stillinger, F.H. (1971) *J. Chem. Phys.*, **55**, 3336–59.

Rapaport, D.C. (2004) *The Art of Molecular Dynamics Simulation*, 2nd edn, Cambridge University Press, Cambridge.

Rappe, A.K., Casewit, C.J., Colwell, K.S., Goddard, W.A., and Skiff, W.M. (1992) *J. Am. Chem. Soc.*, **114**, 10024–35.

Rashin, A.A. (1990) *J. Phys. Chem.*, **94**, 1725–33.

Reiher III, W.E. (1985) *Theoretical Studies of Hydrogen Bonding (Molecular Mechanics, ab initio Calculation, Potential Surfaces, Formamide, Water* Ph.D. Thesis, Harvard University.

Ren, P.Y. and Ponder, J.W. (2003) *J. Phys. Chem. B*, **107**, 5933–47.

Rick, S.W. and Berne, B.J. (1996) *J. Am. Chem. Soc.*, **118**, 672–79.

Rick, S.W. and Stuart, S.J. (2002) *Rev. Comput. Chem.*, **18**, 89–146.

Rick, S.W., Stuart, S.J., and Berne, B.J. (1994) *J. Chem. Phys.*, **101**, 6141–56.

Ringe, D. and Petsko, G.A. (1986) *Methods Enzymol.*, **131**, 389–433.

Rodriguez-Gomez, D., Darve, E., and Pohorille, A. (2004) *J. Chem. Phys.*, **120**, 3563–78.

Roux, B. (1999) *Biophys. J.*, **77**, 139–53.

Roux, B. (2001) Implicit solvent models, in *Computational Biochemistry and Biophysics* (eds O.M. Becker, A.D. MacKerell, B. Roux, and M. Watanabe), Marcel Dekker, New York, p. xii, 512.

Roux, B., Allen, T., Berneche, S., and Im, W. (2004) *Q. Rev. Biophys.*, **37**, 15–103.

Roux, B., Andersen, O.S., and Allen, T.W. (2008) *J. Chem. Phys.*, **128**, 227101.

Roux, B., Berneche, S., and Im, W. (2000) *Biochemistry*, **39**, 13295–306.

Roux, B. and Karplus, M. (1991) *J. Phys. Chem.*, **95**, 4856–68.

Ryckaert, J.P., Ciccotti, G., and Berendsen, H.J.C. (1977) *J. Comput. Phys.*, **23**, 327–41.

Sanbonmatsu, K.Y. and Tung, C.S. (2007) *J. Struct. Biol.*, **157**, 470–80.

Sansom, M.S.P., Scott, K.A., and Bond, P.J. (2008) *Biochem. Soc. Trans.*, **36**, 27–32.

Sapay, N. and Tieleman, D.P. (2008) *Computational Modeling of Membrane Bilayers*, Vol. 60, pp. 111–30.

Schaefer, M. and Karplus, M. (1996) *J. Phys. Chem.*, **100**, 1578–99.

Schlenkrich, M., Brickmann, J., Mackerell, J., Alexander, D., and Karplus, M. (1996) An empirical potential energy function for phospholipids: criteria for parameter optimization and applications, *Biological Membranes: A Molecular Perspective from Computation and Experiment*, pp. 31–81.

Schlitter, J., Engels, M., Kruger, P., Jacoby, E., and Wollmer, A. (1993) *Mol. Simul.*, **10**, 291.

Scott, K.A., Bond, P.J., Ivetac, A., Chetwynd, A.P., Khalid, S., and Sansom, M.S.P. (2008) *Structure*, **16**, 621–30.

Scott, W.R.P., Hunenberger, P.H., Tironi, I.G., Mark, A.E., Billeter, S.R., Fennen, J., Torda, A.E., Huber, T., Kruger, P., and van Gunsteren, W.F. (1999) *J. Phys. Chem. A*, **103**, 3596–607.

Seeman, P. (1997) *Pharmacol. Rev.*, **24**, 583–655.

Sharp, K.A. and Honig, B. (1990) *Annu. Rev. Biophys. Biophys. Chem.*, **19**, 301–32.

Shelley, J.C., Shelley, M.Y., Reeder, R.C., Bandyopadhyay, S., Moore, P.B., and Klein, M.L. (2001) *J. Phys. Chem. B*, **105**, 9785–92.

Shelley, J.C., Sprik, M., and Klein, M.L. (1993) *Langmuir*, **9**, 916–26.

Shi, Q., Izvekov, S., and Voth, G.A. (2006) *J. Phys. Chem. B*, **110**, 15045–48.

Shih, A.Y., Arkhipov, A., Freddolino, P.L., and Schulten, K. (2006) *J. Phys. Chem. B*, **110**, 3674–84.

Shirley, W.A. and Brooks, C.L. (1997) *Proteins: Struct. Funct. Genet.*, **28**, 59–71.

Shirts, M.R., Pitera, J.W., Swope, W.C., and Pande, V.S. (2003) *J. Chem. Phys.*, **119**, 5740–61.

Shrivastava, I. and Sansom, M. (2000) *Biophys. J.*, **78**, 557–70.

Simonson, T. (1993) *Mol. Phys.*, **80**, 441–47.

Simonson, T. (2001) Free energy calculations, in *Computational Biochemistry and Biophysics* (eds O.M. Becker, A.D. MacKerell Jr, B. Roux, and M. Watanabe), Marcel Dekker, New York, pp. 169–97.

Simonson, T., Carlsson, J., and Case, D.A. (2004) *J. Am. Chem. Soc.*, **126**, 4167–80.

Sitkoff, D., Sharp, K.A., and Honig, B. (1994) *J. Phys. Chem.*, **98**, 1978–88.

Smeijers, A.F., Pieterse, K., Markvoort, A.J., and Hilbers, P.A.J. (2006) *J. Phys. Chem. B*, **110**, 13614–23.

Smith, P.E. and Pettitt, B.M. (1996) *J. Chem. Phys.*, **105**, 4289–93.

Soetens, J.C., Millot, C., Chipot, C., Jansen, G., Angyan, J.G., and Maigret, B. (1997) *J. Phys. Chem. B*, **101**, 10910–17.

Sonne, J., Jensen, M.O., Hansen, F.Y., Hemmingsen, L., and Peters, G.H. (2007) *Biophys. J.*, **92**, 4157–67.

Souaille, M. and Roux, B. (2001) *Comput. Phys. Commun.*, **135**, 40–57.

Spassov, V.Z., Yan, L., and Szalma, S. (2002) *J. Phys. Chem. B*, **106**, 8726–38.

Steinbach, P.J. and Brooks, B.R. (1994) *J. Comput. Chem.*, **15**, 667–83.

Stern, H.A., Kaminski, G.A., Banks, J.L., Zhou, R.H., Berne, B.J., and Friesner, R.A. (1999) *J. Phys. Chem. B*, **103**, 4730–37.

Stevens, M.J., Hoh, J.H., and Woolf, T.B. (2003) *Phys. Rev. Lett.*, **91**, 188102.

Still, W.C., Tempczyk, A., Hawley, R.C., and Hendrickson, T. (1990) *J. Am. Chem. Soc.*, **112**, 6127–29.

Stillinger, F.H. and David, C.W. (1978) *J. Chem. Phys.*, **69**, 1473–84.

Stillinger, F.H. and Rahman, A. (1974) *J. Chem. Phys.*, **60**, 1545–57.

Streett, W.B., Tildesley, D.J., and Saville, G. (1978) *Mol. Phys.*, **35**, 639–48.

Stuart, S.J. and Berne, B.J. (1996) *J. Phys. Chem.*, **100**, 11934–43.

Sugita, Y., Kitao, A., and Okamoto, Y. (2000) *J. Chem. Phys.*, **113**, 6042–51.

Sugita, Y. and Okamoto, Y. (1999) *Chem. Phys. Lett.*, **314**, 141–51.

Swendsen, R.H. and Wang, J.S. (1986) *Phys. Rev. Lett.*, **57**, 2607–9.

Swope, W.C., Andersen, H.C., Berens, P.H., and Wilson, K.R. (1982) *J. Chem. Phys.*, **76**, 637–49.

Takano, T. (1977) *J. Mol. Biol.*, **110**, 537–68.

Tama, F., Wriggers, W., and Brooks III, C.L. (2002) *J. Mol. Biol.*, **321**, 297–305.

Tanizaki, S. and Feig, M. (2005) *J. Chem. Phys.*, **122**, 124706.

Thole, B.T. (1981) *Chem. Phys.*, **59**, 341–50.

Thompson, N., Thompson, G., Cole, C.D., Cotten, M., Cross, T.A., and Busath, D.D. (2001) *Biophys. J.*, **81**, 1245–54.

Tian, F. and Cross, T. (1999) *J. Mol. Biol.*, **285**, 1993–2003.

Tien, H.T. and Diana, A.L. (1968) *Chem. Phys. Lipids*, **2**, 55–101.

Tironi, I.G., Sperb, R., Smith, P.E., and Vangunsteren, W.F. (1995) *J. Chem. Phys.*, **102**, 5451–59.

Torrie, G.M. and Valleau, J.P. (1977) *J. Comput. Phys.*, **23**, 187–99.

Townsley, L.E., Tucker, W.A., Sham, S., and Hinton, J.F. (2001) *Biochemistry*, **40**, 11676–86.

Tozzini, V. (2005) *Curr. Opin. Struct. Biol.*, **15**, 144–50.

Tozzini, V. and McCammon, J.A. (2005) *Chem. Phys. Lett.*, **413**, 123–28.

Treptow, W., Marrink, S.J., and Tarek, M. (2008) *J. Phys. Chem. B*, **112**, 3277–82.

Tuckerman, M., Berne, B.J., and Martyna, G.J. (1992) *J. Chem. Phys.*, **97**, 1990–2001.

Tuckerman, M.E. and Martyna, G.J. (2000) *J. Phys. Chem. B*, **104**, 159–78.

Trylska, J., Tozzini, V., and McCammon, J.A. (2005) *Biophys. J.*, **89**, 1455–63.

Ulmschneider, M.B., Ulmschneider, J.P., Sansom, M.S.P., and Di Nola, A. (2007) *Biophys. J.*, **92**, 2338–49.

Urey, H.C. and Bradley, C.A. (1931) *Phys. Rev.*, **38**, 1969–78.

Urry, D.W. (1971) *Proc. Natl. Acad. Sci. U.S.A.*, **68**, 672–76.

Vanbelle, D., Froeyen, M., Lippens, G., and Wodak, S.J. (1992) *Mol. Phys.*, **77**, 239–55.

Van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., and Berendsen, H.J.C. (2005) *J. Comput. Chem.*, **26**, 1701–18.

Vangunsteren, W.F. and Berendsen, H.J.C. (1990) *Angew. Chem. Int. Ed. Engl.*, **29**, 992–1023.

Vedani, A. (1988) *J. Comput. Chem.*, **9**, 269–80.

Venturoli, M., Smit, B., and Sperotto, M.M. (2005) *Biophys. J.*, **88**, 1778–98.

Venturoli, M., Sperotto, M.M., Kranenburg, M., and Smit, B. (2006) *Phys. Rep.-Rev. Sect. Phys. Lett.*, **437**, 1–54.

Verlet, L. (1967) *Phys. Rev.*, **159**, 98.

Voltz, K., Trylska, J., Tozzini, V., Kurkal-Siebert, V., Langowski, J., and Smith, J. (2008) *J. Comput. Chem.*, **29**, 1429–39.

Vorobjev, Y.N. and Scheraga, H.A. (1997) *J. Comput. Chem.*, **18**, 569–83.

Vorobyov, I., Anisimov, V.M., Greene, S., Venable, R.M., Moser, A., Pastor, R.W., and MacKerell, A.D. (2007) *J. Chem. Theory Comput.*, **3**, 1120–33.

Vorobyov, I.V., Anisimov, V.M., and MacKerell, A.D. (2005) *J. Phys. Chem. B*, **109**, 18988–99.

Vorobyov, I., Li, L., and Allen, T.W. (2008) *J. Phys. Chem. B*, **112**, 9588–602.

Wang, L., Bose, P.S., and Sigworth, F.J. (2006a) *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 18528–33.

Wang, J.Y., Deng, Y.Q., and Roux, B. (2006b) *Biophys. J.*, **91**, 2798–814.

Wang, J.M., Cieplak, P., and Kollman, P.A. (2000) *J. Comput. Chem.*, **21**, 1049–74.

Wang, W., Donini, O., Reyes, C.M., and Kollman, P.A. (2001) *Annu. Rev. Biophys. Biomol. Struct.*, **30**, 211–43.

Wang, L., Veenstra, D.L., Radmer, R.J., and Kollman, P.A. (1998) *Proteins*, **32**, 438–58.

Wang, J.M., Wolf, R.M., Caldwell, J.W., Kollman, P.A., and Case, D.A. (2004) *J. Comput. Chem.*, **25**, 1157–74.

Warren, G.L. and Patel, S. (2007) *J. Chem. Phys.*, **127**, 064509.

Warshel, A. and Karplus, M. (1972) *J. Am. Chem. Soc.*, **94**, 5612–25.

Warshel, A., Kato, M., and Pisliakov, A.V. (2007) *J. Chem. Theory Comput.*, **3**, 2034–45.

Warshel, A. and Levitt, M. (1976) *J. Mol. Biol.*, **103**, 227–49.

Warshel, A. and Lifson, S. (1970) *J. Chem. Phys.*, **53**, 582–94.

Warshel, A. and Papazyan, A. (1998) *Curr. Opin. Struct. Biol.*, **8**, 211–17.

Warwicker, J. and Watson, H.C. (1982) *J. Mol. Biol.*, **157**, 671–79.

Weeks, J.D., Chandler, D., and Andersen, H.C. (1971) *J. Chem. Phys.*, **54**, 5237.

Weiner, P.K. and Kollman, P.A. (1981) *J. Comput. Chem.*, **2**, 287–303.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. (1984) *J. Am. Chem. Soc.*, **106**, 765–84.

Wereszczynski, J. and Andricioaei, I. (2006) *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 16200–5.

Whitfield, T.W., Varma, S., Harder, E., Lamoureux, G., Rempe, S.B., and Roux, B. (2007) *J. Chem. Theory Comput.*, **3**, 2068–82.

Willis, B.T.M. and Pryor, A.W. (1975) *Thermal Vibrations in Crystallography*, Cambridge University Press, London.

Woodcock, L.V. (1971) *Chem. Phys. Lett.*, **10**, 257.

Woolf, T.B. and Roux, B. (1994) *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 11631–35.

Woolf, T. and Roux, B. (1996) *Proteins: Struct. Funct. Genet.*, **24**, 92–114.

Wu, Y.J., Tepper, H.L., and Voth, G.A. (2006) *J. Chem. Phys.*, **124**, 024503.

Xie, W.S., Pu, J.Z., MacKerell, A.D., and Gao, J.L. (2007) *J. Chem. Theory Comput.*, **3**, 1878–89.

Xiong, H., Crespo, A., Marti, M., Estrin, D., and Roitberg, A.E. (2006) *Theor. Chem. Acc.*, **116**, 338–46.

Xu, Z.T., Luo, H.H., and Tieleman, D.P. (2007) *J. Comput. Chem.*, **28**, 689–97.

Yang, W.T. and Lee, T.S. (1995) *J. Chem. Phys.*, **103**, 5674–78.

Yang, L.J., Tan, C.H., Hsieh, M.J., Wang, J.M., Duan, Y., Cieplak, P., Caldwell, J., Kollman, P.A., and Luo, R. (2006) *J. Phys. Chem. B*, **110**, 13166–76.

Yeh, I.C. and Hummer, G. (2004) *J. Phys. Chem. B*, **108**, 15873–79.

Yelimov, S., van der Giessen, E., Onck, P.R., and Marrink, S.J. (2008) *Biophys. J.*, **94**, 2994–3002.

Yellen, G. (2002) *Nature*, **419**, 35–42.

Yin, D.X. and Mackerell, A.D. (1998) *J. Comput. Chem.*, **19**, 334–48.

Yoo, J. and Cui, Q. (2008) *Biophys. J.*, **94**, L61–L63.

You, T.J. and Harvey, S.C. (1993) *J. Comput. Chem.*, **14**, 484–501.

Yu, H. and van Gunsteren, W.F. (2004) *J. Chem. Phys.*, **121**, 9549–64.

Yu, H.B. and van Gunsteren, W.F. (2005) *Comput. Phys. Commun.*, **172**, 69–85.

Zauhar, R.J. and Morgan, R.S. (1985) *J. Mol. Biol.*, **186**, 815–20.

Zhang, Y.H., Feller, S.E., Brooks, B.R., and Pastor, R.W. (1995) *J. Chem. Phys.*, **103**, 10252–66.

Zhang, D., Gullingsrud, J., and McCammon, J.A. (2006) *J. Am. Chem. Soc.*, **128**, 3019–26.

Zhou, H.X. (1993) *Biophys. J.*, **65**, 955–63.

Zhou, R., Berne, B.J., and Germain, R. (2001) *Proc. Natl. Acad. Sci.*, **98**, 14931–36.

Zwanzig, R.W. (1954) *J. Chem. Phys.*, **22**, 1420–26.

Zwanzig, R.W. (2001) *Nonequilibrium Statistical Mechanics*, Oxford University Press, New York, Oxford.

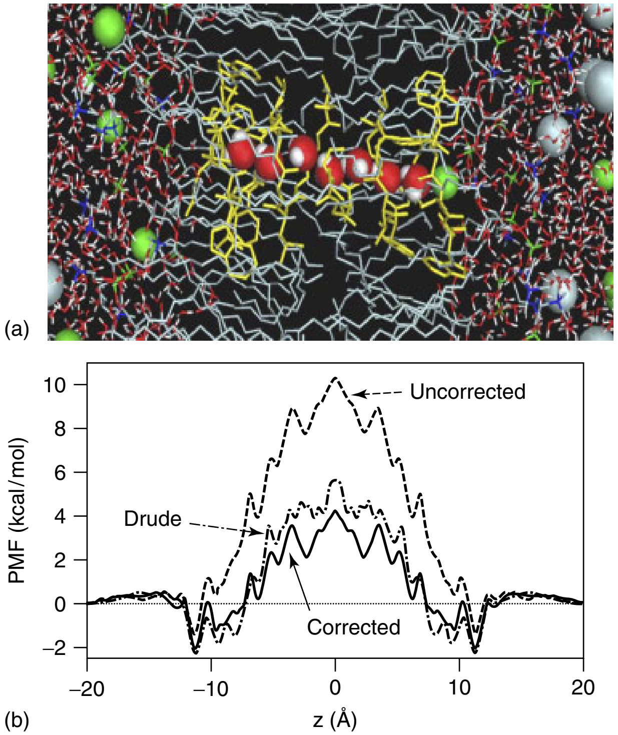

Fig. 8.1 Gramicidin A ion channel. (a) gA protein (yellow) embedded in a hydrated DMPC membrane (gray hydrocarbon tails, water molecules as red and white sticks, potassium ions as green balls and chloride ions as gray balls) with membrane normal oriented horizontally. Water molecules inside the narrow pore are shown as balls. (b) 1D PMF for permeation before any artifact corrections (''uncorrected'', dashed), after postsimulation corrections (''corrected'', solid) and after simulation with a Drude-polarizable membrane hydrocarbon core (''Drude'', dash-dot) (adapted from with ''Drude'' curve added). (This figure also appears on page 278.)

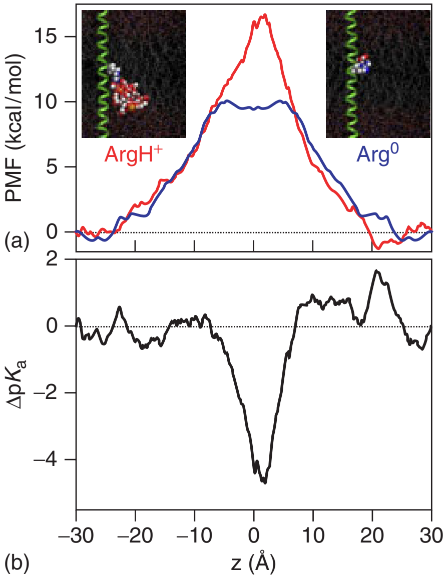

Fig. 8.4 Arginine side chain attached to a TM helix. (a) PMF across the membrane (adapted from). MD systems for the charged (ArgH^+) and neutral Arg side chain (Arg^0) with the side chain held near the center of the bilayer are shown as insets revealing membrane deformations in the charged case. Arg is drawn as blue/gray balls, the TM helix as a green ribbon and the hydrated DPPC bilayer with gray C atoms, orange P atoms and red/white water molecules. Lipid phosphate and water atoms that have been pulled into the bilayer core are drawn as balls. The PMFs for charged (red) and neutral (blue) Arg from umbrella sampling simulations are shown. (b) Arg pK_a shift profile across the membrane calculated from free energy profiles. (This figure also appears on page 286.)

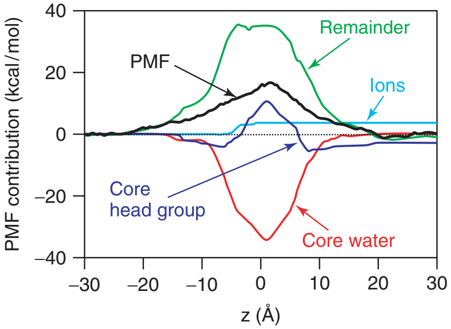

Fig. 8.5 Free energy decomposition for an Arg side chain across the membrane. The black curve is the PMF from Figure 8.4, the red curve is the contribution from water molecules inside the membrane core, the blue curve is from lipid head groups drawn into the core and the green curve is the remainder that makes up the total PMF. (adapted from). (This figure also appears on page 289.)

9

Dynamics of Individual Muscular Protein Molecules

Bjørn Gilbert Nielsen

9.1 Introduction 309
9.2 Organization of Muscle Cells 310
9.3 Sequence of Events Leading to Muscle Contraction 311
9.4 On the Forces Produced by an Array of Springs 314
9.5 Activation Dynamics 320
9.6 Conclusion 322
References 322

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

9.1
Introduction

> *"Muscle is the only known piece of machinery which can be cooked in many ways." Thomas A. McMahon in Muscles, Reflexes, and Locomotion*

One defining characteristic of life is the ability to use energy to actively and purposefully transport matter against opposing forces, thus bringing the organism far away from thermodynamic equilibrium, a defining property of so-called dissipative systems (Prigonine, 1978). Matter is transported from one part of a single cell to another, from the roots to the branches of a tree, or even from our dinner plates to our mouths. Such transport depends on a complex and highly idiosyncratic interplay between thousands of different molecules in a process collectively termed *cell motility*, justly considered "one of the crowning achievements of evolution." (Lodish *et al.*, 1995) This chapter focuses only on those molecules that are most evidently involved in the movement of whole organisms, with special attention to the molecular composition of muscles and the parameters governing their molecular interactions. This chapter is mainly focused on the forces and interactions existing between only a small subset of proteins that are directly involved in muscle activation and movement (these are listed in Table 9.1), and should not be seen as a full treatment of the literature, rather it is hoped that this work will provide the reader with some general knowledge from which to initiate further explorations into the different topics covered.

When trying to describe the dynamics of individual muscle proteins, one should keep in mind that the biological systems we have access to study today are the result of a selection process that has now lasted several billion years. In this arms-race of evolution, the ability to move has played a decisive role in determining which molecular interactions give the most viable solutions. During this selection process, most of these molecules have acquired a multiplicity of roles added step by step as adaptations providing new functionality into an already working system. This ultimately implies that a single protein might be involved in many very different cellular processes due to its interaction with many different molecules. Of these proteins, myosin is by weight the most abundant protein in mammals, and it is also the protein whose function is most directly evident: to produce tensile

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

Tab. 9.1 Proteins involved in muscle contraction.

Protein	Abbreviation	Function
Acetylcholine receptor	ACh-R	Activation and excitability
Voltage gated Na^+ channel	vNa	
Voltage gated K^+ channel	vK	
Ryanodine receptor	RyR	Ca^{2+} binding and transport
Dihydropyridine receptor	DHPC	
Parvalbumin	par	
Calsequestrin	cal	
Tropomyosin	Tro	
Troponin C	TrC	
Troponin T	TrT	
Troponin I	TrI	
Sarco(endo)plasmic reticulum calcium ATPase	SERCA	
Titin	Tt	Scaffolding and connection
Connexin	Cx	
Actin	Ac	
Myosin	Mys	Force production

force. However, this force would never be produced were it not for several dozen other proteins responsible for activation, modulation, regulation, scaffolding, strong reversible bonding, and so on (see Table 9.1).

9.2
Organization of Muscle Cells

Muscles are collections of single cells, each of which has the ability to contract. By acting in synchrony, the small contribution generated by each cell adds up to considerable forces in the whole muscle (>50 N cm^{-2}). On a first approximation, it is precisely the organization of such cells which determines whether a muscle belongs to the skeletal, the cardiac, or the smooth muscle type:

- Skeletal muscle: These are highly organized into sarcomeres with all muscle fibers arranged in parallel and pulling synchronously. Skeletal muscle fibers are activated by synapses using the neurotransmitter acetylcholine, and are under voluntary control.
- Cardiac muscle: The muscle fibers are organized into sarcomeres, with individual fibers arranged in parallel. Cardiac muscle cells are connected with gap junctions so that an electric pulse generated at one place (the sinoatrial node) will propagate throughout the cardiac muscle through such junctions. These pulses are myogenic (self-initiated) and do not require direct synaptic activity, but are modulated by sympathetic and parasympathetic nervous activity.
- Smooth muscle: The filamentous muscle proteins are not organized into sarcomeres, and may be activated either by cell–cell propagation through gap junctions, via synaptic action or via hormonal effects. The

spindle shaped smooth muscle cells respond to autonomic control and to hormonal stimulation.

In the following, we specifically consider the skeletal muscle type, keeping in mind that there are significant differences to the other muscle types. However, it can be argued that these differences are more of a structural nature, and thus a quantitative biophysical model of muscle should, by its nature, be adaptable to any of the possible muscle types insofar as it contains the right type of parameters.

A single cell in a muscle is known as a *muscle fiber*, and in addition to all the usual organelles present in animal cells (e.g. nucleus, mitochondria, and Golgi apparatus), it predominantly houses the sarcoendoplasmic reticulum (SR) enveloping several myofibrils (mfs). SR and mfs dominate the muscle fiber almost to the point of excluding other organelles. As will be clarified shortly, the division of labor is such that the SR transforms the electric signals from the nerves and T-tubule system into Ca^{2+} signals within the mfs, whereas the mf is responsible for the production of mechanical force and subsequent contraction.

9.3
Sequence of Events Leading to Muscle Contraction

Motor neurons in the spinal cord innervate muscle tissue are therefore the only major class of neurons that make synaptic contact with non-nervous tissue. A distinction is made between α, β, and γ type motorneurons, depending on the actual type of muscle tissue that they contact. By definition, α motor neurons in the spinal cord directly synapse with muscle fibers whereas γ type neurons activate contractile tissue within the muscle spindle (a sensory organ within muscles that monitors rate of change of length and tension in the muscle). Beta Motor neurons innervate both types of fibers.

When an action potential arrives at the neuromuscular junction, small packets of the neurotransmitter acetylcholine (ACh) are released from the presynaptic region of the neuromuscular junction. Each released packet (or vesicle) contains on the order of 10000 ACh molecules, which diffuse across the narrow synaptic gap separating the motor neuron from the muscle cell, and subsequently bind to postsynaptic acetylcholine receptors (ACh-R). Each molecule of ACh is capable of binding to a single ACh receptor only once before it becomes hydrolyzed into acetate and choline (by acetylcholinesterase, an enzyme in the synaptic cleft), but two molecules of ACh are required to effectively activate the receptor.

Thus, the ACh-R is the first protein that becomes activated upon binding of neurotransmitter initially released following the arrival of an action potential to the neuromuscular junction. When ACh binds to the receptor the associated nonspecific ohmic ion channel opens, creating a small conductance $g_{\mathrm{ACh}} \sim 30$ pS and the ensuing current, I_{ACh}, will drive the local membrane potential, V_m, toward the reversal potential for the ion channel, $E_{\mathrm{ACh}} = 0$ mV. Formally the current through the ion channel may be expressed as

$$I_{\mathrm{ACh}} = g_{\mathrm{ACh}}(V_m - E_{\mathrm{ACh}}) \qquad (1)$$

Given that muscle cells typically have a resting membrane potential of approximately $V_m = -90$ mV inside relative to outside, a depolarization toward the reversal potential of the acetylcholine receptor will effectively cause an influx of positively charged ions that are driven by the electrochemical gradient, primarily of Na^+ initially. The "cloud" of acetylcholine released from a single presynaptic spike is typically responsible for activating on the order of 10^5 individual ACh-receptors, resulting in a total current of approximately -500 nA (by convention currents are measured as outward positive). This current causes a buildup of charge across the muscle cell's membrane, which effectively functions as an insulator with a specific capacitance of $1\,\mu F\,cm^{-2}$ and a specific resistance of $20-60\,k\Omega\,cm^{-2}$ (roughly equivalent to an RC-circuit with a time constant of 20–60 ms). When the electric field across the membrane reaches a certain threshold (usually 10–20 mV above the resting membrane potential), another class of membrane proteins are poised for action to initiate an action potential in the muscle cell, namely the voltage-gated Na^+ and K^+ channels. Given the electrochemical gradients of Na^+ and K^+, the former type of channels will activate quickly to further depolarize the membrane by initiating a positive feedback loop in which depolarization leads to more Na^+ channels opening, Na^+ ions flow into the cell thus giving further depolarization, until the reversal potential for Na^+ is reached around $> +30$ mV. Conversely, the later (K^+) channel will be activated at a higher threshold and with a short time delay and will drive the membrane toward the reversal potential for K^+ (at around -90 mV) due to the resulting efflux of K^+, ultimately reaching the resting membrane potential.

This results in a wave of depolarization called a *muscle action potential*, an event that is very similar to the action potential observed in neurons and other excitable cell types. However, the muscle action potential will not result in the release of neurotransmitter, instead it will result in the contraction of the muscle fiber. The muscle action potential propagates deep into the muscle fiber via the T-tubule network, see Figure 9.1, where it activates dihydropyridine receptors (DHPRs). These are a specialized type of voltage dependent L-type Ca^{2+} channels that exist within the T-tubule membrane of all muscles (see reviews by Franzini-Armstrong (1999); Dulhunty *et al.* (2002)) and respond to the incoming wave of depolarization by opening a Ca^{2+} selective ion channel and/or by interacting directly with the ryanodine receptor (RyR), a protein embedded in the membrane of the SR, reviewed in Fill and Copello (2002). The amount of Ca^{2+} released from the extracellular fluid into the myoplasm via DHPR channels is not sufficient to activate muscles by itself, an additional release of Ca^{2+} is required (see reviews by Bottinelli and Reggiani (2000); Dulhunty *et al.* (2002). In cardiac muscle, the RyR channel becomes activated directly by the Ca^{2+} released through the DHPR channel (Nakai *et al.*, 1996; Gurrola *et al.*, 1999; Protasi *et al.*, 2002; Kugler *et al.*, 2004); however, in skeletal muscle an additional direct molecular interaction between DHPR and RyR is required to activate the RyR channel (Bottinelli and Reggiani, 2000). In either case, the release of additional Ca^{2+} into the myoplasm is mediated by the activation of the RyR channel, which in fact is a Ca^{2+} selective ion channel that permits the release of substantial amounts of Ca^{2+} from the internal buffers within the lumen of the SR, primarily from

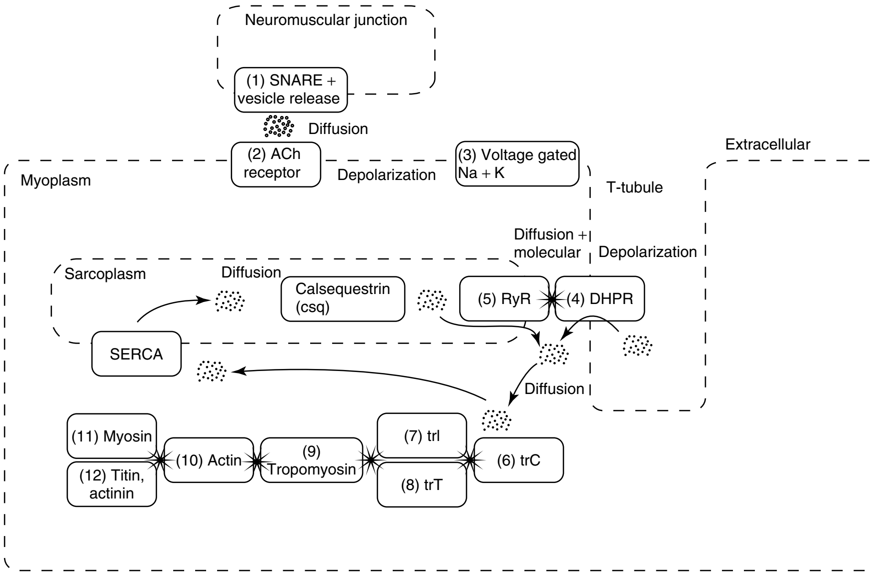
Neuromuscular junction
(1) SNARE + vesicle release
Diffusion
(2) ACh receptor
Depolarization
(3) Voltage gated Na + K
Extracellular
Myoplasm
T-tubule
Diffusion + molecular
Depolarization
Sarcoplasm
Diffusion
Calsequestrin (csq)
(5) RyR
(4) DHPR
SERCA
Diffusion
(11) Myosin
(12) Titin, actinin
(10) Actin
(9) Tropomyosin
(7) trI
(8) trT
(6) trC

Fig. 9.1 Activation scheme for a muscle, emphasizing the role of proteins in the process of muscle contraction. "Clouds" of dots indicate the presence of freely diffusing substances like neurotransmitters and Ca^{2+}, stars indicate direct contact based protein–protein interaction, arrows indicate transport or diffusion gradient, dashed lines indicate the membranes of different cellular compartments within muscle cells, and rounded squares are actual proteins. Numbering roughly corresponds to activation order in the muscle activation sequence. Abbreviations used are SNARE, soluble N-ethylmaleimide sensitive fusion protein attachment receptor; ACh, acetylcholine; DHPR, dyhydropyridine receptor; RyR, ryanodine receptor; trC, troponin C; trI, troponin I; trT, troponin T; SERCA, sarco(endo) plasmic reticulum calcium ATPase. See text for further details on the activation dynamics.

calsequestrin (csq) (Jorgensen, Kalnins and MacLennan, 1979; Jorgensen *et al.*, 1983; Beard, Laver and Dulhunty, 2004). Once Ca^{2+} is present in the myoplasm, it will either be sequestered by endogenous buffers or it will bind to Troponin C (trC) which by direct molecular interactions will cause conformational changes in troponin I and troponin T, and tropomyosin (see Collins (1991); Alberts *et al.* (2002) for review), ultimately leading to a small displacement of tropomyosin relative to the actin filament around which it is intertwined. The resting function of tropomyosin, that is, when no Ca^{2+} is present, is to block specialized binding pockets where the actin binding region of the myosin molecule will bind in the presence of ATP, so when Ca^{2+} is present, these binding sites will be available for myosin and contraction may begin. It should be noted, that in some types of muscle fiber (e.g. fast fatigable fibers) relatively high concentrations of buffer proteins are present in the myoplasm, especially parvalbumin (par) which assists in rapidly sequestering Ca^{2+} from the myoplasm, leading to contractions of shorter duration due to the shorter exposure time of Troponin to Ca^{2+} (Celio and Heizmann, 1982).

Once bound to actin, the neck region of myosin will undergo a conformational change in which it will cause the myosin head region to rotate approximately 90°, essentially shortening the end-to-end distance of the myosin molecule (Lymn and Taylor, 1971; Holmes, 1996). This conformation change thus produces a force that induces a translation of actin with respect to myosin, subsequently transmitted via connective proteins such as α actinin, capZ, nebulin (nebulette), and connectin (titin) to neighboring sarcomeres, ultimately causing a contraction of the whole muscle fiber.

9.4
On the Forces Produced by an Array of Springs

The smallest functional unit in muscle is the sarcomere, a structure which in mammals has a typical diameter of approximately 1 μm, and length between 1.5 and 3.5 μm, depending on the state of contraction (and type of muscle). Within the sarcomere, thick filaments composed of myosin and thin filaments composed of actin are arranged in parallel so that each thick filament is surrounded by exactly six actin filaments. A cross section of this system reveals a clearly hexagonal pattern. The cross-sectional area of such a hexagon is $\sim 1200\,\text{nm}^2$, so approximately 650 thick filaments contribute to force in a single sarcomere. Along the thick filament two pairs of myosin heads are found sticking out in opposite directions every 14.3 nm, and each such opposing group of myosin molecules is rotated 60 °C with respect to the preceding group. The myosin head covered region of the sarcomere can be calculated from the typical length of the A-band (anisotropic band) minus the typical length of the H-band (bright band), which gives approximately 1.4 μm. Note, that two halves of a sarcomere are separated by the H-band and pull in opposite directions, so effectively the working unit is the half sarcomere. The total number of working myosin head pairs spanning one half of the sarcomere on a single thick filament can thus be calculated to be $2 \times 700\,\text{nm}/14.3\,\text{nm} \sim 100$. When attached to actin, a single myosin head produces forces in the range of 1–10 pN, thus a quick calculation reveals that if one assumes all myosin heads are simultaneously active, the maximal holding force that a single thick filament can produce in one direction will be in the range of 0.1–1 nN, and the total force that may be produced by a single muscle may be estimated from its cross-sectional area, ranging approximately from 8 to $80\,\text{N cm}^{-2}$. Although the origin of this force is generally accepted to be in compliance within the actomyosin complex, the exact details pertaining to how the force production is initiated and maintained remain the subject of intense study (for a recent review and account see Karatzaferi, Chinn and Cooke (2004). Much attention has been given to determining how and when the energy released from the hydrolysis of a single ATP molecule ($\sim$0.06 aJ or $36\,\text{kJ mol}^{-1}$) brings about the appropriate conformational changes of the myosin neck and head regions. Some hints may be gained by examining the details of myosin–actin interactions. In fact, at any given time only a fraction of the myosin heads are attached to actin and therefore contribute to force. Myosin goes through an attachment–detachment cycle known as the *cross-bridge cycle*. Within this cycle, myosin has two preferred conformations, one when bound to ATP and one when bound to ADP, distinguishable by the angle between the myosin neck–head region with respect to the main trunk of the thick filament. When bound to ATP, myosin detaches from actin and the preferred conformation is one where the myosin molecule as a whole is more contracted, with an angle of about 45 °C between the myosin tail and the myosin neck–head region. Myosin is an ATPase and will therefore break down ATP into ADP + P, but this must occur in several steps. First a complex of ADP·P is formed during which the preferred conformation of myosin changes to be much more extended, with the neck–head region of myosin now forming an angle of 90 °C to the myosin tail. At this position, the actin binding site on the myosin head comes into perfect apposition to the myosin binding sites on the actin filament, and a weak bond is formed. Once bound, phosphate is released from the ADP·P complex, and the preferred conformation of myosin changes back to the 45 °C configuration; however, this change cannot occur immediately because at this point actin is bound to the myosin head, and must be pulled from its current position to a position corresponding to the new preferred conformation of myosin. This rotation of the myosin head causes a 5.3 nm displacement of the myosin filament with respect to the actin filament, and the amount of work produced by this rotation will be calculated in the next paragraph.

Detailed biophysical models of this system have been developed and carefully described in several articles and books, and the reader is encouraged to consult these for further details, especially (McMahon, 1984; Bagshaw, 1994). What follows here is a highly simplified model that captures only the bare essentials, to give a sense of the type of information such models require and provide. Given the highly organized almost crystalline structure of muscle, it is straightforward to produce a simple model of force production. It seems certain that all active crossbridges within a half sarcomere will act in parallel, so their individual contributions are additive. To keep things simple assume that each myosin–actin crossbridge acts like a spring with an extension dependent spring force, $s(x_i)$, where x_i is the extension of the particular crossbridge indexed by i. Assuming that there are N crossbridges formed, the total force, F_{tot}, produced by such an array of parallel springs amounts to

$$F_{\mathrm{tot}} = \sum_{i=1}^{N} s(x_i) \tag{2}$$

It is here assumed that all bonds form at an initial extension length of approximately $x = 5.3$ nm, and during muscle contractions these bonds shorten until bond release occurs at $x = 0$ nm, while during muscle stretching these bonds lengthen until bond breaking at approximately $x = 11$ nm (in agreement with experimental evidence reported in Brown and Loeb (1999)). Hence, during contraction the crossbridge will have contracted from $a = 5.3$ nm to $b = 0$ nm, producing the total work $W_{\mathrm{tot}} = \int_a^b s(x)dx$. For simplicity a linear spring model has most predominantly been used in the literature to represent the compliance of the myosin–actin cross-bridge, thus $s(x) = kx$, although the actual origin of this force has not to date been properly specified. In this case the total work of contraction becomes $W_{\mathrm{tot}} = \frac{1}{2}\mathrm{ka}^2$, which for a spring with $k = 2$ pN nm^{-1} amounts to 0.028 aJ of work (vs. 0.06 aJ for ATP hydrolysis). By a similar argument, the amount of work required to stretch the bond from initial state of 5.3 nm to bond breaking point at 11 nm is approximately 0.09 aJ. From the point of view that the force producing region is in fact a polypeptide that is folded in a particular way, it would be reasonable to assume the compliance to be nonlinear, perhaps following the force-extension profiles described for other polypeptide systems (as proposed by Howard and Spudich (1996); Nielsen (2002)). Actual stretching experiments on this system have proven difficult to perform, and therefore only a limited amount of experimental data is available. Experiments by Molloy *et al.* (1995); Lewalle *et al.* (2008) support the linear compliance view, even though in these experiments myosin exists under rather unnatural conditions as single filaments bound to very large spheres. In reality myosin exists within a rather constrained environment, and it is difficult to evaluate how this constrainment affects functionality. Stretching experiments performed *in silico* using steered molecular dynamics reveal that the form of the force-extension profile of a single scallop myosin molecule (in rigor) to some extent depends on the velocity of stretching, see Figure 9.2b. Although very approximative, this result hints to the fact that effects such as friction, damping, geometric constraints and hydrostatic forces might also have a role to play under nonisometric conditions.

Thus, to a first approximation, the myosin-actin crossbridge seems to have a linear force-extension relationship within the extension range relevant for muscle contractions, the remaining question is then: where does this linearity come from? Given that myosin is a complex polymer molecule that is being mechanically deformed one would expect that entropy and Brownian motions to a large extent dictate the forces present at the force producing neck region of myosin, similar to what has been described for many other proteins and polypeptides (Smith, Finzi and

Fig. 9.2 Stretching experiment performed *in silico* using steered molecular dynamics in order to apply controlled forces on different parts of a myosin molecule. (a) The myosin molecule was stretched out to a predetermined extension and then held at that position to sample the reactive force produced by the molecule. The colored blobs indicate force-extension samples resulting from five different stretch–hold experiments. A line fitting these samples is shown together with a family of dashed lines corresponding to best fits using the worm like chain model expressed by Eq. (3) and using persistence lengths from 2 to 5 Å, and contour lengths of 20–40 nm (explanation in text). (b) The molecule was continuously stretched at a high velocity and the reactive force was plotted. The best fit worm like chain model is also shown as a black line. In both cases coordinate data corresponding to scallop myosin in near rigor state was obtained from the Protein Data Bank (pdb-id: 1SR6), from which a sub region was selected to represent the neck region of the molecule, corresponding to residues 473–584, 693–728, and 738–781, and shown in the figure as the first molecular structures surrounding the plots (upper left structure). This subregion was equilibrated using molecular dynamics (NAMD2 developed by Kalé *et al.* (1999), and a number of α carbon atoms was fixed while the α carbon of residue 774 was pulled in different directions depending on the case, either as a stretch–hold experiment (a) or constant velocity pulling experiment (b), resulting in the other molecular structures shown. (Please find a color version of this figure on the color plates.)

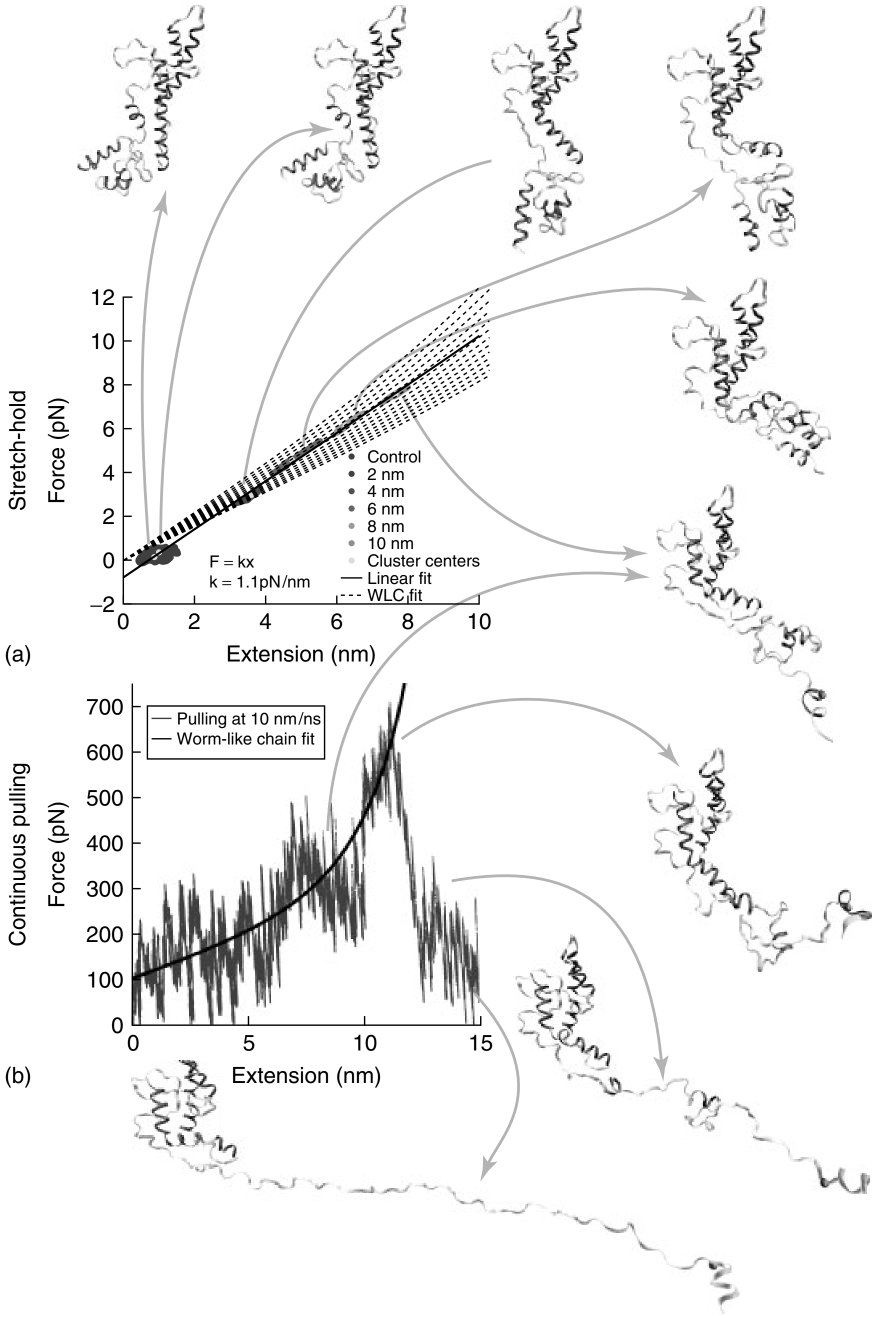
Stretch-hold
Force (pN)
12
10
8
6
4
2
0
–2
0
2
4
6
8
10
Extension (nm)
Control
2 nm
4 nm
6 nm
8 nm
10 nm
Cluster centers
Linear fit
WLC fit
F = kx
k = 1.1pN/nm
(a)
Continuous pulling
Force (pN)
700
600
500
400
300
200
100
0
0
5
10
15
Extension (nm)
Pulling at 10 nm/ns
Worm-like chain fit
(b)

Bustamante, 1992; Bustamante *et al.*, 1994; Kellermayer *et al.*, 1998; Smith and Radford, 2000; Marszalek *et al.*, 1999; Rief *et al.*, 1999; Oberhauser *et al.*, 1998). Hence, the apparent linearity of myosin must be derived from a first principles perspective and not just accepted without question. One possible scenario which would allow for linearity at short extension ranges while still maintaining an entropy driven folding process, makes use of the so-called Worm-Like Chain model of polypeptides (initially developed for DNA by Marko and Siggia (1995)). Within this model, statistical mechanics are used to describe the thermodynamic forces affecting a polymer by assuming that the polymer can be approximated by a flexible chain consisting of linked units (here amino acids). The model predicts that a linked flexible chain in a thermal bath will have a natural tendency to fold up to minimize free energy, thereby maximizing entropy. Work done on such a chain will reduce its entropy until the chain is fully stretched, and the force required to perform this work depends on the extension of the polymer, and although the actual derivation is rather involved (see derivation by Marko and Siggia (1995), the resulting expression is in fact quite simple:

$$s(x) = \frac{k_B T}{A}\left(\underbrace{\frac{x}{L} - \frac{1}{4}}_{\text{linear}} + \underbrace{\frac{1}{4\left(1 - \frac{x}{L}\right)^2}}_{\text{nonlinear}}\right) \tag{3}$$

where k_B is Boltzmann's constant, T is the temperature, A is the persistence length of the chain (equal to one half the Kuhn length of the chain, b, the length of the freely jointed links in the chain), x is the amount of extension away from the unperturbed state, and finally L is the contour length of the chain (i.e. the unfolded length of the polymer under consideration). The contour length is here related to the persistence length as $L = Nb = 2NA$, where N is the number of links making up the polymer. Substances obeying this type of force-extension relationship are said to possess *entropic elasticity*, and behave linearly at extensions much smaller than the length of the folded region of polymer, but nonlinearity at greater extensions. Figure 9.2a shows a family of force-extension curves (dashed lines) showing a range of persistence lengths from 2 to 5 Å, and contour lengths of 20–40 nm. Note here that within the extension range relevant for muscles (0–11 nm) all of these curves are close to linear. The implication here is, that the individual links in the chain are identical with the rigid peptide bond, and that steric hindrances imposed by side chain protruding from the C_α atoms lead to further rigidity resulting in a Kuhn length which is very close to 0.7 nm, and thus corresponding to the length of a dipeptide. The further implication is that the number of peptide bonds participating in the contour length should be in the range of 30–60, given that the contour length is about 20–40 nm. This compares well to the apparently deformable region observed during molecular dynamics simulations, and is not a bad estimate given the almost naive simplicity of this model. Further details about how these results have been obtained will be published elsewhere.

The next step is to define the form and appropriate values for the distribution of crossbridge extensions represented by x_i. In one approach (developed further in Nielsen (2002)), it is assumed that during contraction there is a uniform distribution of lengths between $a = 0$ and $b = 5.3$ nm, thus $x_i = x_1, x_2, \ldots, x_{n-1}, x_n = \{\delta, 2\delta, 3\delta, \ldots, (n-1)\delta, 5.3\}$, where $\delta = (b - a/N_{\text{cb}})$ thus the total force produced by the system can be found as the integral sum over the forces for all the different lengths:

$$F_{\text{tot}} = \frac{N_{\text{cb}}}{b-a} \int_a^b s(x)dx \tag{4}$$

which assuming $s(x)$ has the form given by Eq. (3) gives the total force:

$$F_{\text{tot}} = \frac{N_{\text{cb}}}{b-a} \int_a^b \frac{k_B T}{A} \times \left(\frac{x}{L} - \frac{1}{4} + \frac{1}{4(1-\frac{x}{L})^2}\right) dx \tag{5}$$

$$= \frac{N_{\text{cb}}}{b-a} \frac{k_B T}{A} \left(\frac{b^2 - a^2}{2L} - \frac{1}{4}(b-a) \right. \tag{6}$$

$$\left. + \frac{L}{4}\left(\frac{1}{1-b/L} - \frac{1}{1-a/L}\right)\right) \tag{7}$$

The remaining task is to determine the actual number of crossbridges that have formed and to this end it is useful to consider, that bond formation takes time, and that the amount of available time falls off linearly as a function of sliding velocity between filaments. At increasing velocity there must be a trade off between the increased frequency with which myosin heads encounter actin binding sites, and the reduced time available for interaction. A simple scheme to describe this considers the probability that a bond actually forms as a simple linearly decreasing function of velocity:

$$p_{\text{bond}} = 1 - \kappa \frac{|v|}{v_{\max}} \tag{8}$$

Where κ is a muscle dependent parameter, and $v_{\max}$ is the maximal sliding velocity of the muscle. One of the cornerstones in experimental muscle science was the observation by Hill (1938); Hill (1964), that the force of a muscle and its corresponding contraction velocity seem to follow a very simple hyperbolic relationship given by $F_{\text{hill}} = (1 - v)/[1 + v/k)]$, where k is a parameter that depends on the muscle type in question (typical values are in the range from 0.15 to 0.25). Any muscle model should as a minimum be able to account for this result, and the very simple approach adopted here does indeed lead to similar results (see e.g. Nielsen (2002); Nielsen (2003)). Furthermore, one should always consider the ATP utilization rate and the rate of energy liberation predicted by a model, and again the results here compare quite nicely with experiment (as reported in Nielsen (2002)). Thus we may assume that the main ingredients of muscle contraction have been captured by the model. Further refinements are of course necessary to truly claim that this model describes muscle accurately, for example, one should consider the duty ratios of myosin more carefully, and also it would be necessary to include many of the compliances acting in series with the muscle section considered here. Furthermore, the binding rates of myosin to actin may be included in much more detail (as has been done by other workers, see e.g. Zahalak (2000); Huxley (1957); Ford, Huxley and Simmons (1977)).

9.5
Activation Dynamics

Most proteins listed in Table 9.1 are involved in Ca^{2+} binding and Ca^{2+} transport within muscle tissue, and are crucial to muscle activation and dynamics but do not participate in the generation of force as such. Muscles are traditionally considered as part of the nervous system in lieu of their very close interaction with neurons and their intrinsic excitability. This excitability is mainly caused by the activation of voltage-gated Na^+ and K^+ channels, and is initiated by the opening of acetylcholine receptors. What really sets muscle cells apart from other excitable cells is the cascade of molecular events following excitation. As mentioned in a previous section, and hinted at in Figure 9.1, depolarization leads to activation of the DHPR, and subsequently to release of calcium from the sarcoplasm through RyR. The exact details depend on the muscle type in question, but regardless of type and mechanism, it is useful to consider the movement of Ca^{2+} through the system as such. If we assume that Ca^{2+} is a limited resource within the cell, then one may keep track of Ca^{2+} by determining which of the proteins in Table 9.1 act as sinks and/or sources of Ca^{2+}, and which of the compartments in the muscle cell are to be considered Ca^{2+} reservoirs. In brief, Ca^{2+} may be said to exist either as free Ca^{2+} or as buffer bound Ca^{2+}. Working through the detailed description given previously on Ca^{2+} movement and summarized in Figure 9.1, the following fractions of Ca^{2+} may be considered:

φ_{myo}	Free Ca^{2+} within the myoplasm
φ_{sar}	Free Ca^{2+} within the longitudinal regions of the sarcoplasm
φ_{csq}	Ca^{2+} associated with calsequestrin or diffusing freely in the sarcoplasmic terminal
φ_{trc}	Ca^{2+} in association with troponin C
φ_{par}	Ca^{2+} in association with parvalbumin (=0 in slow-twitch fibers)

Given that these are fractions of the total amount of calcium, their sum must be unity. To describe sudden movement of a certain fraction, ε of calcium moving from one reservoir, a to another, b, at time t_0 simple equations may be used of the form:

$$\frac{d\varphi_a}{dt} = \frac{\varphi_b}{\tau} - \varepsilon\varphi_b\delta(t-t_0) \tag{9}$$

$$\frac{d\varphi_b}{dt} = -\frac{\varphi_b}{\tau} + \varepsilon\varphi_b\delta(t-t_0) \tag{10}$$

where τ is a time constant controlling the rate of movement of Ca^{2+} from reservoir b to reservoir a. One such equation is needed for each suggested Ca^{2+} reservoir. Considering the interaction network presented in Figure 9.1, six time constants are considered rate limiting in the present model:

τ_{csq}	Time constant governing $\varphi_{\text{sar}} \rightarrow \varphi_{\text{csq}}$
τ_{serca}	Time constant governing $\varphi_{\text{myo}} \rightarrow \varphi_{\text{sar}}$
$\tau_{\text{trc,on}}$	Binding time governing $\varphi_{\text{myo}} \rightarrow \varphi_{\text{trc}}$
$\tau_{\text{trc,off}}$	Dissociation time constant governing $\varphi_{\text{trc}} \rightarrow \varphi_{\text{myo}}$

$\tau_{\mathrm{par,on}}$ Binding time constant governing $\varphi_{\mathrm{myo}} \rightarrow \varphi_{\mathrm{par}}$

$\tau_{\mathrm{par,off}}$ Dissociation time constant governing $\varphi_{\mathrm{par}} \rightarrow \varphi_{\mathrm{myo}}$

Several highly interesting and useful models of this so-called excitation–contraction coupling have been developed (see e.g. Hannaford (1990); Chou and Hannaford (1992); Otazu, Futami and Hoshimiya (2001); Wexler, Ding and Binder-Macleod (1997); Ding, Binder-MacLeod and Wexler (1998)), focusing on the different roles of the muscle proteins involved. The model presented here follows a recent approach by this author Nielsen (2006a); Nielsen (2006b) aimed at reducing the complexity and number of parameters of the previous models, in addition to provide a possible interpretation to the spike "doublets" often observed in spinal motor neurons (as defined in Simpson (1969), see also Maton and Bouisset (1975); Søgaard *et al.* (2001); Garland and Griffin (1999)). As described previously each muscle fiber is activated via an axon originating from a motor neuron in the spinal cord. The collection of all individual muscle fibers that a single motor neuron innervates is known as *motor unit*, a term coined by Sherrington in his early studies of the motor system (Sherrington, 1897). The ratio of innervation, that is the number of muscle fibers contacted by a single motor neuron, is highly correlated with the type of muscle fiber in question, so that strong fast fatigable fibers have very high innervation ratios, and the weaker and slow fatigue resistant fibers have very low innervation ratios (few fibers per motorneuron). This being so, one may wonder how precise control of muscle force is accomplished, and to this end there are several contributing factors. First, there are many more slow and weak motor units than fast and fatigable motor units. In fact, the actual number of motor units, n, producing a given force, ϕ, strongly depends on the force and roughly has the form $n(\phi) = \exp\{-\beta\phi\}$, where β is a scaling parameter that depends on the actual muscle in question, thus small and weak units are far more numerous than large and strong units (for more information regarding this distribution consult (Milner-Brown, Stein and Yemm, 1973)). Second, according to the Henneman size principle (Henneman, Somjen and Carpenter, 1965), motor units are recruited in order of increasing size, so the smallest units are recruited first, and the largest units last. This orderly recruitment leads to the observation that even when 50% of all motor units in the muscle have been recruited, only 20% of the total force is being produced (Walmsley, Hodgson and Burke, 1978), allowing for a very gradual increase in force.

Of course these facts have to be seen in proportion to the type of tasks that the skeletal musculature is usually requested to do. For example, one may calculate, that at 246.2 $\mathrm{km\,h^{-1}}$ it takes a tennis ball 346 ms to traverse the full length of a tennis court (record-breaking serve by Andy Roddick in 2004). Within this time, the opponent receiving the serve must judge the ball's trajectory, precisely move arm and racquet to a position where the serve may be countered and apply a controlled amount of force to return the ball and perhaps even score a point. Common values for reaction times to visual cues are in the order of 200 ms for simple behavioral responses like pressing buttons. More elaborate responses require more time. It is evident that precise voluntary movements require that muscles be rapidly brought from one predefined state of tension to another within precisely timed intervals. According to previously invoked size principle, slow and weak muscle fibers are recruited first, followed by strong, fast and fatigable fibers. Counting from the onset of a regularly firing spike train, it takes on the order of 20–100 ms for a relaxed fast muscle fiber to reach target tension. The corresponding figure for slow muscle fibers is about 300–1000 ms. This delay is not due to resiliency by the contractile machinery in the muscles, but is due to a relatively slow cumulative buildup of Ca^{2+} within the myoplasm following each spike. Thus, fast fibers might make it to the goal in time, but not so with slow fibers even if they are recruited first. In work by Nielsen (2006a); Nielsen (2006b) it is suggested, that in order to jump-start even slow (but strong) muscles, the nervous system might adopt a strategy where the insertion of spike "doublets" in an otherwise regular spike train will bring the muscle force up to a particular force level almost instantaneously (within milliseconds rather than the usually required hundreds of milliseconds), thus taking advantage of the so-called "catch-like" effect of skeletal muscle (first documented by Burke, Rudomin and Zajac (1970)).

9.6
Conclusion

It should be clear from the exposition above, that there is a very direct link between the action of specific proteins and the contraction of muscle. Simply put, proteins are required for every single step in the excitation–contraction coupling. However, it is only within the last decades where the tremendous advances in molecular biology and biophysics have allowed researchers to identify exactly which proteins are relevant and in which types of muscle, and to map out the intricate interaction networks between such proteins. Given the availability of such highly detailed interaction networks it is now feasible to routinely develop quite sophisticated network models accounting for different phenomena that one might observe in living muscle. Now comes the major task of putting in the right parameters, be they expressed as time constants, concentrations, coupling constants, rate constants, release fractions, force-extension profiles, and so on, that will make such models useful not only from a theoretical point of view, but also from the clinical perspective in order to start asking "what if" questions, and to develop novel approaches for improving or even repairing the body when it is broke. In the present exposition I have only tried to present some of the parameters that are usually considered important for understanding muscle function, focusing primarily on those parameters that have a direct link to protein action, and therefore no claims are made in favor of one or the other actual interaction model as this is considered outside the scope of this work. It is hoped that the reader will find inspiration in this work, perhaps as a stepping stone toward further exploration into the fascinating subject of muscle activation.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walther, P. (2002) *Molecular Biology of the Cell*, 4th edn, Garland Science, New York.

Bagshaw, C. (1994) *Muscle Contraction*, 2nd edn, Chapman & Hall, London.

Beard, N., Laver, D., and Dulhunty, A. (2004) Calsequestrin and the calcium release channel of skeletal and cardiac muscle. *Prog. Biophys. Mol. Biol.*, **85**, 33–69.

Bottinelli, R. and Reggiani, C. (2000) Human skeletal muscle fibres: molecular and functional diversity. *Prog. Biophys. Mol. Biol.*, **73**, 195–262.

Brown, I. and Loeb, G. (1999) Measured and modeled properties of mammalian skeletal muscle. i. The effects of post-activation potentiation on the time course and velocity dependencies of force production. *J. Muscle Res. Cell. Motil.*, **20**, 443–56.

Burke, R., Rudomin, P., and Zajac, F. (1970) Catch property in single mammalian motor units. *Science*, **168**(3927), 122–24.

Bustamante, C., Marko, J., Siggia, E., and Smith, S. (1994) Entropic elasticity of lambda-phage DNA. *Science*, **265**(5178), 1599–600.

Celio, M. and Heizmann, W. (1982) Calcium-binding protein parvalbumin is associated with fast contracting muscle fibres. *Nature*, **297**, 504–6.

Chou, C. and Hannaford, B. (1992) Dual stable point model of muscle activation and deactivation. *Biol. Cybern.*, **66**, 511–23.

Collins, J. (1991) Myosin light chains and troponin c: structural and evolutionary relationships revealed by amino acid sequence comparisons. *J. Muscle Res. Cell. Motil.*, **12**, 3–25.

Ding, J., Binder-MacLeod, S., and Wexler, A. (1998) Two-step, predictive, isometric force model tested on data from human and rat muscles. *J. Appl. Physiol.*, **85**(6), 2176–89.

Dulhunty, A., Haarmann, C., Green, D., Laver, D., Board, P., and Casarotto, M. (2002) Interactions between dihydropyridine receptors and ryanodine receptors in striated muscle. *Prog. Biophys. Mol. Biol.*, **79**, 45–75.

Fill, M. and Copello, J. (2002) Ryanodine receptor calcium release channels. *Physiol. Rev.*, **82**, 893–922.

Ford, L., Huxley, A., and Simmons, R. (1977) Tension responses to sudden length change in stimulated frog muscle fibres near slack length. *J. Physiol.*, **269**, 441–515.

Franzini-Armstrong, C. (1999) The sarcoplasmic reticulum and the control of muscle contraction. *FASEB J.*, **13**, S266–S270.

Garland, S. and Griffin, L. (1999) Motor unit double discharges: statistical anomaly or functional entity. *Can. J. Appl. Physiol.*, **24**(2), 113–30.

Gurrola, G., Arovalo, C., Sreekumar, R., Lokuta, A., Walker, J., and Valdivia, H. (1999) Activation of ryanodine receptors by imperatoxin A and a peptide segment of the II-III loop of the dihydropyridine receptor. *J. Biol. Chem.*, **274**(12), 7879–86.

Hannaford, B. (1990) A nonlinear model of the phasic dynamics of muscle activation. *IEEE Trans. Biomed. Eng.*, **37**(11), 1067–75.

Henneman, E., Somjen, G., and Carpenter, D. (1965) Excitability and inhibitibility of motoneurones of different sizes. *J. Neurophysiol.*, **28**, 599–620.

Hill, A. (1938) The heat of shortening and the dynamic constants of muscle. *Proc. R. Soc. B*, **126**, 136–95.

Hill, A. (1964) The effect of load on the heat of shortening of muscle. *Proc. R. Soc. Lond., B*, **159**, 297–318.

Holmes, K. (1996) Muscle proteins – their actions and interactions. *Curr. Opin. Struct. Biol.*, **6**, 781–89.

Howard, J. and Spudich, J. (1996) APPENDIX: is the lever arm of myosin a molecular elastic element? *Proc. Natl. Acad. Sci.*, **93**, 4462–64.

Huxley, A. (1957) Muscle structure and theories of contraction. *Prog. Biophys. Biophys. Chem.*, **7**, 255–318.

Jorgensen, A., Kalnins, V., and MacLennan, D. (1979) Localization of sarcoplasmic reticulum proteins in rat skeletal muscle by immunofluorescence. *J. Cell Biol.*, **80**, 372–84.

Jorgensen, A., Shen, A., Campbell, K., and MacLennan, D. (1983) Ultrastructural localization of calsequestrin in rat skeletal muscle by immunoferritin labeling of ultrathin frozen sections. *J. Cell Biol.*, **97**, 1573–81.

Kalé, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K., and Schulten, K. (1999) NAMD2: greater scalability for parallel molecular dynamics. *J. Comput. Phys.*, **151**, 283–312.

Karatzaferi, C., Chinn, M., and Cooke, R. (2004) The force exerted b a muscle cross-bridge depends directly on the strength of the actomyosin bond. *Biophys. J.*, **87**, 2532–44.

Kellermayer, M., Smith, S., Bustamante, C., and Granzier, H. (1998) Complete unfolding of the titin molecule under external force. *J. Struct. Biol.*, **122**, 197–205.

Kugler, G., Weiss, R., Flucher, B., and Grabner, M. (2004) Structural requirements of the dihydropyridine receptor α_{1S} II-III loop for skeletal-type excitation-contraction coupling. *J. Biol. Chem.*, **279**(6), 4721–28.

Lewalle, A., Steffen, W., Stevenson, O., Ouyang, Z., and Sleep, J. (2008) Single-molecule measurement of the stiffnes of the rigor myosin head. *Biophys. J.*, **94**, 2160–69.

Lodish, H., Baltimore, D., Berk, A., Zipursky, S., Matsudaira, P., and Darnell, J. (1995) *Molecular Cell Biology*, Scientific American Books, New York.

Lymn, R. and Taylor, E. (1971) Mechanism of adenosine tri-phosphate hydrolysis of actomyosin. *Biochemistry*, **10**, 4617–24.

Marko, J. and Siggia, E. (1995) Stretching DNA. *Macromolecules*, **28**, 8759–70.

Marszalek, P., Lu, H., Li, H., Carrion-Vazquez, M., Oberhauser, A., Schulten, K., and Fernandez, J. (1999) Mechanical unfolding intermediates in titin modules. *Nature*, **402**, 100–3. Authors present EMG data on reach to grab task.

Maton, B. and Bouisset, S. (1975) Motor unit activity and preprogramming of movement in man. *Electroencephalogr. Clin. Neurophysiol.*, **38**, 658–60.

McMahon, T. (1984) *Muscles, Reflexes, and Locomotion*, 1st edn, Princeton University Press, Princeton, NJ.

Milner-Brown, H., Stein, R., and Yemm, R. (1973) The orderly recruitment of human motor units during voluntary isometric contractions. *J. Physiol.*, **230**, 358–70.

Molloy, J., Burns, J., Kendrick-Jones, J., Tregear, R., and White, D. (1995) Movement and force produced by a single myosin head. *Nature*, **378**, 209–12.

Nakai, J., Dirksen, R., Nguyen, H., Pessah, I., Beam, K., and Allen, P. (1996) Enhanced dihydropyridine receptor channel activity in the presence of ryanodine receptor. *Nature*, **380**, 72–75.

Nielsen, B. (2002) The role of muscle spindles in constraining motor control. *Neurocomputing*, **44**, 943–49.

Nielsen, B. (2003) Unfolding transitions in myosin give rise to the double-hyperbolic force-velocity relation in muscle. *J. Phys. Condens. Matter*, **15**, S1759–S1765.

Nielsen, B. (2006a) Ca$ 2+$resource allocation model of the catch-like effect in skeletal muscle. *J. Muscle Res. Cell. Motil.*, **27**, 527.

Nielsen, B. (2006b) Frequency-dependent interactions between ion channels in the triad junction and the induction of post-tetanic potentiation. *Proc. Physiol. Soc.*, **4**, PC10.

Oberhauser, A., Marszalek, P., Erickson, H., and Fernandez, J. (1998) The molecular elasticity of the extracellular matrix protein tenascin. *Nature*, **393**, 181–85.

Otazu, G., Futami, R., and Hoshimiya, N. (2001) A muscle activation model of variable stimulation frequency response and stimulation history, based on positive feedback in calcium dynamics. *Biol. Cybern.*, **84**, 193–206.

Prigonine, I. (1978) Time, structure and fluctuations. *Science*, **201**, 777–85.

Protasi, F., Paolini, C., Nakai, J., Beam, K., Franzini-Armstrong, C., and Allen, P. (2002) Multiple regions of ryr1 mediate functional and structural interactions with α_{1S}-dihydropyridine receptors in skeletal muscle. *Biophys. J.*, **83**, 3230–44.

Rief, M., Pascual, J., Saraste, M., and Gaub, H. (1999) Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles. *J. Mol. Biol.*, **286**, 553–61.

Sherrington, C. (1897) The central nervous system, in *Textbook of Physiology*, (M. Foster), Macmillan, p. 60.

Simpson, J. (1969) Terminology of electromyography. *Electroencephalogr. Clin. Neurophysiol.*, **26**, 224–26.

Smith, S., Finzi, L., and Bustamante, C. (1992) Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science*, **258**(5085), 1122–26.

Smith, D. and Radford, S. (2000) Pulling back the frontiers. *Curr. Biol.*, **10**, R662–R664.

Søgaard, K., Sjøgaard, G., Finsen, L., Olsen, H., and Christensen, H. (2001) Motor unit activity during stereotyped finger tasks and computer mouse work. *J. Electromyogr. Kinesiol.*, **11**, 197–206.

Walmsley, B., Hodgson, J., and Burke, R. (1978) Forces produced by medial gastrocnemius and soleus muscles during locomotion in freely moving cats. *J. Neurophysiol.*, **41**(5), 1203–16.

Wexler, A., Ding, J., and Binder-Macleod, S. (1997) A mathematical model that predicts skeletal muscle force. *IEEE Trans. Biomed. Eng.*, **44**(5), 337–48.

Zahalak, G. (2000) The two-state cross-bridge model of muscle is an asymptotic limit of multi-state models. *J. Theor. Biol.*, **204**, 67–82.

Fig. 9.2 Stretching experiment performed *in silico* using steered molecular dynamics in order to apply controlled forces on different parts of a myosin molecule. (a) The myosin molecule was stretched out to a predetermined extension and then held at that position to sample the reactive force produced by the molecule. The colored blobs indicate force-extension samples resulting from five different stretch–hold experiments. A line fitting these samples is shown together with a family of dashed lines corresponding to best fits using the worm like chain model expressed by Eq. (3) and using persistence lengths from 2 to 5 Å, and contour lengths of 20–40 nm (explanation in text). (b) The molecule was continuously stretched at a high velocity and the reactive force was plotted. The best fit worm like chain model is also shown as a black line. In both cases coordinate data corresponding to scallop myosin in near rigor state was obtained from the Protein Data Bank (pdb-id: 1SR6), from which a sub region was selected to represent the neck region of the molecule, corresponding to residues 473–584, 693–728, and 738–781, and shown in the figure as the first molecular structures surrounding the plots (upper left structure). This subregion was equilibrated using molecular dynamics (NAMD2 developed by Kalé *et al* (1999), and a number of α carbon atoms was fixed while the α carbon of residue 774 was pulled in different directions depending on the case, either as a stretch–hold experiment (a) or constant velocity pulling experiment (b), resulting in the other molecular structures shown. (This figure also appears on page 316-317.)

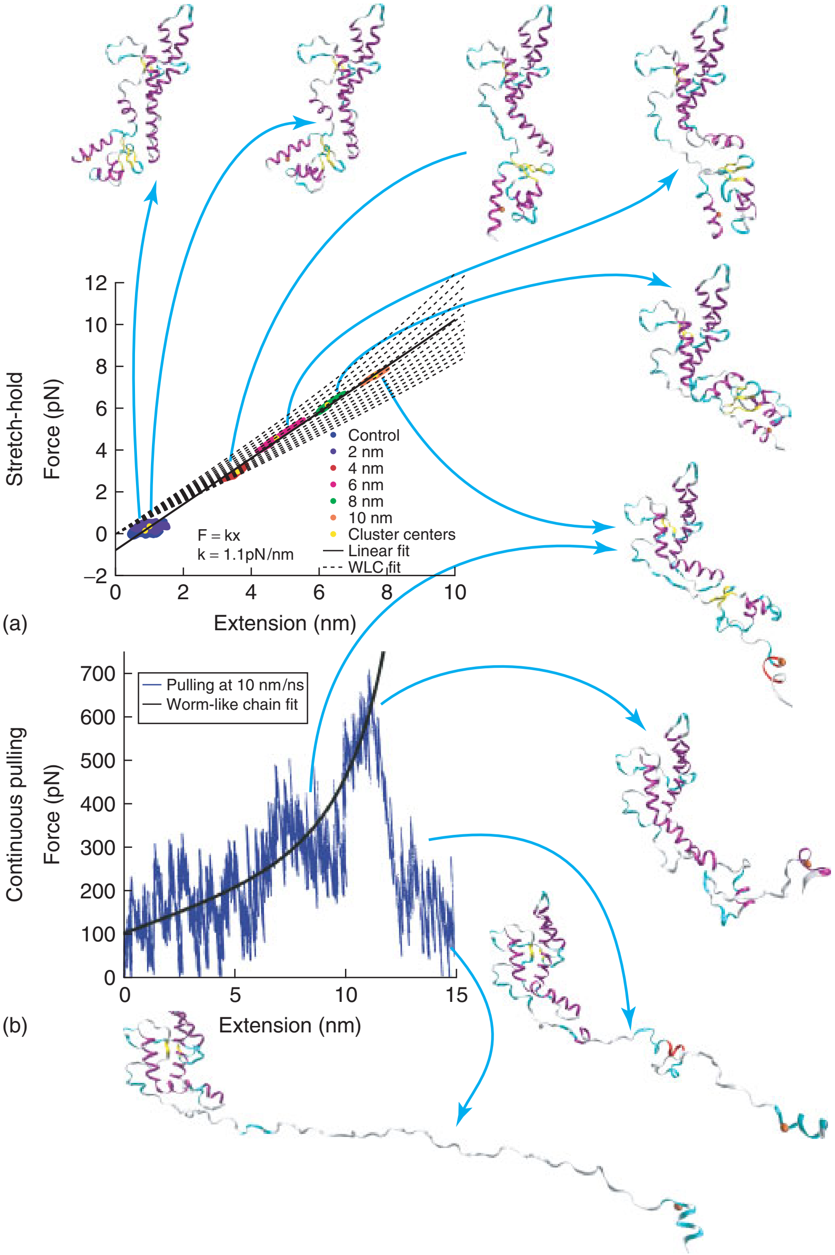
Stretch-hold
Force (pN)
12
10
8
6
4
2
0
−2
0
2
4
6
8
10
Extension (nm)
F = kx
k = 1.1pN/nm
● Control
● 2 nm
● 4 nm
● 6 nm
● 8 nm
● 10 nm
● Cluster centers
Linear fit
WLC fit
(a)
Continuous pulling
Force (pN)
700
600
500
400
300
200
100
0
0
5
10
15
Extension (nm)
Pulling at 10 nm/ns
Worm-like chain fit
(b)

10
Protein–Lipid Interactions in Biological Membranes

Claus Helix Nielsen

10.1 Introduction 331
10.2 Hydrophobic Matching Principle 333
10.3 The Lipid–Protein Matrix 333
10.3.1 Biological Membranes 334
10.3.1.1 Properties of Fluid Lipid Bilayers 334
10.3.1.2 Mechanical Properties of Lipid Bilayers 335
10.3.2 Transmembrane Spanning Proteins 338
10.3.2.1 General Energetic Considerations 339
10.3.2.2 Helical Bundles 339
10.3.2.3 β-Barrels 339
10.3.2.4 Interfacial Anchoring of Transmembrane Segments 340
10.3.2.5 Proteins as Embedded Inclusions 340
10.4 Determining Membrane-Spanning Protein Structure and Orientation in Fluid Lipid Bilayers 341
10.4.1 Some Insights from Nuclear Magnetic Resonance (NMR) Spectroscopy 341
10.4.2 Some Insights from Electron Paramagnetic Resonance (EPR) Spectroscopy 343
10.5 Elastic Membrane Deformations and Lipid–Protein Interactions 344
10.5.1 Quantifying Bilayer Deformation Energy 344
10.5.2 Bilayer Deformation Energy and Protein Function 348
10.5.2.1 Model Systems 348
10.5.2.2 Integral Membrane Proteins 349
10.6 Summary and Perspectives 350
References 351

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

10.1
Introduction

The canonical picture of a biological membrane is that of an amphiphatic barrier to polar solvents defining enclosed cellular and subcellular spaces. Its fundamental constituents are lipid-class molecules, specifically phospholipids arranged in a double layer – a lipid bilayer, and a variety of proteins involved in transport of material and material across the membrane. The cohabitation of proteins and lipids naturally leads to the question of how the two species mutually interact.

Historically, the picture of biological membranes has emerged over the last century and a half, starting in 1855, with the observation of Swiss botanist Nägeli and his student Cramers of the different rates of pigment penetration into damaged and undamaged plant cells (Nägeli and Cramer, 1855). Their conclusion was that plant cells contain an outer layer which was coined the *plasma membrane*. Later, in 1877, Pfeffer devised a permeable membrane for the study of osmosis, and the mathematical relationship to predict osmotic pressure was developed by Jacobus van't Hoff (Andrews, 1929). Pfeffer concluded that the membrane essentially was a permeability barrier to water and polar solutes. This view was modified by Overton who concluded that the entry of any molecule into a cell is the result of its "selective solubility" in the cell's boundary, and that the more soluble in lipids the molecule is, the greater is its permeability – now known as *Overtons Rule* (Overton, 1899).

Thanks to the work of the French chemist Chevreul in the early 1800s, the chemical and physical nature of lipids and fatty acids was understood reasonably well (Chevreul, 1889; McNamara, Warnick and Cooper, 2006). Overton further suggested that the cell membrane contained cholesterol and lecithins (generally denominating a mixture of glycolipids, triglycerides, and phospholipids, but biochemically used as synonym for phosphatidylcholine) (Overton, 1899). Although later studies, including Walther Nernst's formal introduction of the semi-permeable membrane concept in 1908 (Nernst, 1909), refined the conceptual framework, and confirmed Overton's findings, membrane structure remained highly controversial. An important observation, made in 1917 by Langmuir, was that lipid molecules spread out into a one-molecule thick layer at air–aqueous buffer interfaces (Langmuir, 1917). This finding was

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

complemented in 1925 by Gorter and Grendel leading to the bilayer model of the plasma membrane (Gorter and Grendel, 1925). About 10 years later Danielli, working with Harvey, observed that proteins could bind to mackerel oil, and in 1935, together with Davson, Danielli proposed a biological membrane model consisting of a lipid bilayer sandwich covered with proteins on both sides facing the aqueous phases (Danielli and Davson, 1935). With the advent of electron microscopy in the 1950s, Robertson was able to resolve a trilaminar structure and he interpreted this as reflecting two protein coats encapsulating the lipid bilayer and coined the term *unit membrane* for this arrangement (Robertson, 1966). Around the same time, the globular structure of soluble proteins was recognized and generally attributed to the spherical arrangement with the hydrophobic part shielded from the aqueous solvent. This prompted (Lenard and Singer, 1966), and (Wallach and Zahler, 1966) independently to propose a different model for the arrangement of proteins and lipids in biological membranes. Owing to the hydrophobic interior of the lipid bilayer, proteins could in fact be globular entities penetrating into and sometimes spanning the hydrophobic thickness of the lipid bilayer, provided that these proteins adopted conformations that would allow this arrangement. Singer later elaborated the thermodynamic basis for membrane-associated proteins having distinct hydrophobic regions that would be shielded from the aqueous phase by inserting themselves into the bilayer hydrophobic core (Singer, 1971).

In 1971, at a meeting at the New York Academy of Sciences (Green, 1972), a final consensus emerged that biological membranes are liquid-crystalline bilayers. This was summarized in the notorious *fluid-mosaic* Singer–Nicolson model (Singer and Nicolson, 1972). In this model, the outer surfaces are composed primarily of charged and polar lipid head groups and interact with the aqueous phases, whereas the inner portion of the bilayer is composed of the hydrocarbon chains of the lipids aligned parallel to each other, having some flexibility allowing for methylene group rotamatic states. The nonpolar acyl chains are in contact with each other at the terminal methyl group, and this arrangement creates a hydrophobic barrier impermeable to polar solvents, for example, ions, but allow small nonpolar molecules to pass, for example, ethanol and diatomic neutral gasses. Although the original model has been extended and refined to include nonhomogeneous distribution of lipid components, glycocalyx, and cytoskeletal elements (Sackmann, 1984; Mouritsen and Andersen, 1998), its prime conceptual feature is that the membrane is a dynamic (fluid) matrix in which the proteins and lipids can move and interact.

The permeability of pure lipid bilayers is consistent with the approximation that lipid bilayers essentially are 5-nm thick two-dimensional sheets of liquid hydrocarbon (Fettiplace and Haydon, 1980). This suggests that the Singer–Nicolson model implies an impermeable lipid matrix acting as a passive, albeit, fluid solvent for embedded proteins, but this view does not provide a viable basis for understanding the increasing evidence for specific and unspecific, and, to some extent, reciprocal interactions between the lipid bilayer and membrane proteins (Andersen, Sawyer and Koeppe, 1992).

In the following, we limit the scope to integral membrane proteins and focus on their interaction with the host lipid bilayer. First, the general structure, the lipid bilayer, and transmembrane-spanning protein structure is presented. Second, some examples of how spectroscopic techniques can be used to understand protein conformational structure in lipid bilayers are discussed. Finally, the concept of elastic membrane deformations is presented and implications of deformation free energies for protein function are discussed.

10.2
Hydrophobic Matching Principle

The permeability barrier properties of biological membranes entail that membrane proteins are inserted into the lipid bilayer without compromising these properties and/or compromising protein function and stability. The evolutionary consequence of this is that membrane and membrane proteins must have coevolved. The structural consequence is that proteins and lipids must be compatible in the sense that transmembrane-spanning proteins must contain regions that are sufficiently “lipid-like” to match the surrounding membrane constituents (see Figure 10.1). This is expressed in the notion of hydrophobic matching arising from the hydrophobic effect. For an acyl chain, each CH_2 group adds 0.8 kT of energy (Cevc and Marsh, 1987), and it would require >100 kT to expose a membrane edge with the size corresponding to an integral membrane protein to water (Sharp *et al.*, 1991). Thus understanding the mutual lipid–protein matching ability requires an understanding of both bilayer and protein structure.

10.3
The Lipid–Protein Matrix

Both lipids and membrane-spanning proteins have proven to be elusive entities in the sense that it is difficult to obtain structural information. Lipids are generally in the fluid state rendering it difficult to construct structural models. Membrane proteins are notoriously difficult to crystallize and obtain structural information from despite very impressive

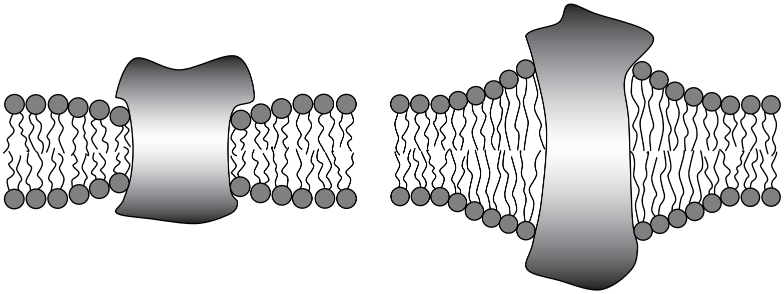

Fig. 10.1 The hydrophobic matching (Mouritsen and Bloom, 1984) implies that whenever the hydrophobic length of an integral membrane-spanning protein is different from the hydrophobic thickness of the unperturbed bilayer, there will be a bilayer deformation involving bending and compression (a) or expansion (b) of the bilayer.

progress made over the last decade, for example, Doyle *et al.* (1998); Jiang *et al.* (2003); Toyoshima *et al.* (2000); Morth *et al.* (2007). Moreover, the intricate interplay between a large number of various proteins, lipids, and cytoskeletal elements remains the major obstacle in a detailed functional understanding of all elements of the biological membrane.

In energetic terms, the understanding of lipid–protein interplay translates into quantifying the plethora of component energies for the various interactions between the elements of the membrane. One of these component energies constituting the total free energy associated with two different conformations of a bilayer embedded protein is based on the reciprocal coupling between lipid and protein arising from the hydrophobic matching principle. In this way, the hydrophobic matching leads to a reciprocal hydrophobic coupling between bilayer (structural and mechanical) properties and protein structure and function (Mouritsen and Bloom, 1984; Andersen and Koeppe, 2007). To assess the energetics of this hydrophobic coupling, structural information about both lipids and proteins is necessary.

10.3.1
Biological Membranes

Despite the apparent simplicity of the Singer–Nicolson model for lipid bilayers, progress in understanding the molecular structure of biological membranes has been hampered by the complex lipid biochemistry. A typical biological membrane contains hundreds of chemically distinct lipids with great variation in the structures of the acyl chains and lipid headgroups (Dowhan, 1997). In addition, there is substantial variability in composition across cell types. Evidently, the cellular requirements for lipid-based chemical signaling constitute a major reason for this complexity, but still it is not generally clear how the lipid composition for a membrane is maintained (Hazel, 1995) and what an acceptable/optimal composition is, for its embedded proteins to function.

10.3.1.1 Properties of Fluid Lipid Bilayers

Although diffraction studies of phospholipid crystals (i.e. below the main gel-to-liquid transition temperature T_m) at low hydration can provide detailed images in which the mean relative positions are well defined and measurable (Pascher *et al.*, 1992), this is not possible for fluid lipid bilayers (i.e. above T_m) where the acyl chains have considerable flexibility with varying torsion angles (the most stable being *trans* (180°) and *gauche* (60°)) along the acyl chain. Instead, one must consider the average spatial distributions of lipid molecule moieties. Thus, one can obtain a bilayer profile in which the various components (e.g. phosphate and choline groups and the methylene region) can be identified as peaks at various positions from the bilayer midplane. White and colleagues have combined X-ray and neutron-diffraction data using joint-refinement procedures (Wiener and White, 1991a, 1991b) and provided multi-Gaussian distribution quasi-molecular models of hydrated (66% relative humidity) dioleoylphosphatidylcholine ((diC18:1)PC) (Wiener, King and White, 1991; Wiener and White 1992a, 1992b). The interfacial regions appear fuzzy in these models. Nevertheless, the bilayer *equilibrium* thickness d_0 can be regarded as a well-defined entity.

Using deuterium quadrupole coupling nuclear magnetic resonance (NMR) an order parameter describing the disorder arising from thermal motion of the acyl chains can be obtained for fluid liquid-crystalline bilayers (Seelig and Seelig, 1974; Seelig, 1977). The order parameter displays a plateau for acyl chain carbons 2–8, whereas segments 10–16 (in the case of dipalmitoylphosphatidylcholine (DPPC)) show increased frequency of rotational isomerization toward the terminal methyl group. The order parameter value for the plateau is quite low with 60% disorder. Interestingly, the order parameter plateau length for DPPC is mirrored in the structure of biological membranes. Prokaryote lipids are typically monounsaturated with the double bond between carbon 9 and 10 (Kenyon, 1978). Eukaryote lipids contain sterols where the rigid polycyclic sterol ring system lies in the membrane between carbon 2 and 9 with its flexible side chain resting between carbon 10 and the terminal end of the acyl chains (Engelman and Rothman, 1972).

The T_m for a series of monounsaturated dioctadec-*cis*-enoyl phosphatidylcholines into which the single *cis*-double bond was inserted at positions 2–17 displays a minimum with the double bond at position 9 (Barton and Gunstone, 1975). Although this observation could lead to the conclusion that the double bond at position 9 serves to increase the "fluidity" of the bilayer, this is not the case. T_m gives information about the stability of the gel state, not the fluid state. Fluidity (i.e. the inverse of viscosity) above T_m, is not affected by the position of the double bond.

Even though the (fluid) liquid-crystalline organization of the lipid bilayer is necessary to allow for conformational changes involving the protein–lipid boundary, the focus on membrane fluidity as *the* determinant for membrane function as it has developed over the last three decades is problematic. First, the term itself is not a well-defined experimental entity. Fluidity is proportional to the rate of motion, but the behavior of many probes with which fluidity has been measured is sensitive to both rate and range of probe motion. Second, important in the context of lipid–protein interactions, it has been pointed out by Lee (1991) that changes in fluidity itself cannot affect the conformational equilibrium between interconverting membrane protein conformations – it only affect the kinetics. Thus the equilibrium distribution among protein conformational states is not affected by changes in fluidity. To understand how the membrane may affect protein function, one therefore has to focus on the energetic modulation of protein conformational equilibrium by the host membrane (Andersen and Koeppe, 2007).

10.3.1.2 **Mechanical Properties of Lipid Bilayers**

The energetics of bilayers is closely linked to their mechanical properties. In this respect, it is useful to treat the bilayer as a continuum entity with physical–chemical properties (Zimmerberg and Gawrisch, 2006). The theoretical description of elastic membrane deformations has evolved greatly since the seminal contribution from Helfrich (1973). In this description, the bilayer material properties are defined by the hydrophobic bilayer thickness d_0, the intrinsic lipid curvature (or monolayer equilibrium curvature) c_0, and the elastic moduli for bilayer compression – expansion K_a and bending K_c. The thickness d_0 of a fully hydrated saturated phosphatidylcholine in the liquid-crystalline state lipids is related to acyl chain length through the semi-empirical approximation:

$$d_0 = 1.75(n-1) \tag{1}$$

where n is the number of acyl chain carbon atoms (Lewis and Engelman, 1983; Sperotto and Mouritsen, 1988). Introducing two monounsaturated acyl chains lowers d by approximately 2.5 Å. As an example, according to Eq. (1), a distearoylphosphatidylcholine (di(C18:0)PC) bilayer has thickness of 29.7 Å, whereas ((diC18:1)PC) has a thickness of 27.2 Å, the latter value in agreement with 27.1 Å from refined diffraction measurements (Nagle and Tristram-Nagle, 2000). A lipid bilayer made from 1-stearoyl-2-oleoyl-phosphatidylcholine (C18:0–C18:1PC), a lipid that can be seen as a representative for a typical plasma membrane lipid, has a thickness of 30.7 Å (Rawicz *et al.*, 2000). Measurements of bilayer electrical capacitance can also provide information about d_0 as the specific capacitance $C_m = \varepsilon\varepsilon_0/d_0$ where ε is the dielectic constant corresponding to a long hydrocarbon ($\varepsilon = 2.1$ for hydrocarbons (Fettiplace, Andrews and Haydon, 1971) and increases by approximately 0.2 per double bond (Lide, 1996)), and $\varepsilon_0 = 8.85 \cdot \times 10^{-12}$ F m^{-1} is the permittivity of free space. This has been investigated in bilayers formed from natural and synthetic lipids dissolved in different n-alkanes (Andrews and Haydon, 1968; Läuger *et al.*, 1967; White and Thompson, 1973), a system often used in protein reconstitution experiments, for example Green, Weiss and Andersen, (1987). By measuring C_m the bilayer thickness was found to decrease with increasing alkane length. For (diC18:1)PC, $d_0 = 49.3$ Å with n-octane as solvent $d_0 = 29.8$ Å with n-hexadecane as solvent (Benz and Janko, 1976). For nominally solvent free membranes formed by the folding of two spread monolayers (Montal and Mueller, 1972; Bamberg and Benz, 1976), $d_0 = 25.8$ Å for(diC18:1)PC is in good agreement with the diffraction data mentioned above.

In general, literature values for lipid bilayer structure (e.g. thickness and area per lipid headgroup) show significant variation, partly due uncertainties in hydration level (Rand and Parsegian, 1989), but recent progress has reduced this problem producing values for fully hydrated liquid-crystalline lipid bilayers (Nagle and Tristram-Nagle, 2000; Katsaras, 1997, 1998).

As d_0 is orders of magnitude smaller than the lateral micrometer extent of biological lipid bilayers, membrane mechanical properties can be approximated to those of a thin deformable sheet. In general, a sheet can be deformed in three principal ways: stretching, bending, and shearing. While fluid lipid bilayers have no resistance to shear (Evans and Skalak, 1980), compression–expansion (CE) and bending or splay-distortion (SD) can be quantified in terms of energy per unit area as

$$E_{\mathrm{CE}} = \frac{1}{2} K_{\mathrm{A}} \left(\frac{u}{d_0}\right)^2 \tag{2}$$

where u is the thickness change from equilibrium thickness d_0, and

$$E_{\mathrm{CE}} = \frac{1}{2} K_{\mathrm{C}} \left[\left(\frac{1}{R_1} + \frac{1}{R_2}\right)^2 - \frac{1}{R_0^2}\right] \tag{3}$$

where R_1 and R_2 are the principal radii of curvatures and $R_0 = 1/c_0$ the equilibrium curvature radius). In a sheet (shell) description of the bilayer (Evans and Skalak, 1980), K_A and K_C are related to each other via d_0:

$$K_{\mathrm{C}} = d_0^2 K_{\mathrm{A}} \tag{4}$$

In addition, there is a surface energy term due to changes in bilayer surface area ΔA relative to the equilibrium area A_0 upon bending with a characteristic interfacial tension Γ

$$E_{\mathrm{ST}} = \Gamma \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] \tag{5}$$

Bending and curvature are related through the lateral stress profile as a function of distance normal to the bilayer surface for the monolayer $t(z)$. The lateral stress profile is a result of intermolecular interactions within the bilayer, which leads to positive and negative lateral pressures in a direction parallel to the bilayer surface (Cantor, 1997a,b, 1999). In the headgroup region, these interactions include steric, hydrational, and electrostatic repulsion (Seddon and Templer, 1993). In the two interfaces between polar and nonpolar regions, there is an attractive lateral tension as a result of the hydrophobic effect attempting to minimize water contact with the hydrophobic core region. Finally, in the core acyl chain region, interactions include van der Waaals attractive forces, as well as repulsive forces exerted by thermally activated *trans–gauche* rotameric states of the acyl chains. For a monolayer in equilibrium, the net tension is zero, which corresponds to

$$\int_{\text{Monolayer total thickness}} t(z)dz = 0 \tag{6}$$

The torque acting on the monolayer is given by the first moment of $t(z)$:

$$\int_{\text{Monolayer total thickness}} zt(z)dz = -K_c c_0 \tag{7}$$

Thus, whenever $c_0 \neq 0$ there will be a net torque on the monolayer acting relative to the neutral surface (the surface of inextension during pure bending), which will force the monolayer to curve away from or toward the aqueous phase. This corresponds to an imbalance of forces making up $t(z)$. For a bilayer made of two opposing monolayers each with $c_0 \neq 0$ this will result in a frustrated bilayer in which energy is stored.

The lateral stress profile is often described in simple geometric terms leading to the concept of effective lipid molecule shape (Luzzati and Husson, 1962; Israelachvili, Marcelja and Horn, 1980; Israelachvili, Mitchell and Ninham, 1976; Israelachvilli, 1992) (see Figure 10.2). Thus lipids can be described as having an effective cylindrical shape when the torque is zero ($c_0 = 0$), and these lipids form relaxed bilayers. When the effective cross-sectional area of the headgroup region is larger than that of the acyl chain region, the monolayer has a $c_0 > 0$. The result is effectively cone-shaped lipids forming a convex surface when viewed from the aqueous phase. Conversely, when the effective cross-sectional area of the headgroup region is smaller than that of the acyl chain region, the monolayer has a $c_0 < 0$. This results in effectively inverted cone-shaped lipids forming a concave surface when viewed from the aqueous phase. In a bilayer, the two monolayers are coupled in the sense that they must have complementary curvatures.

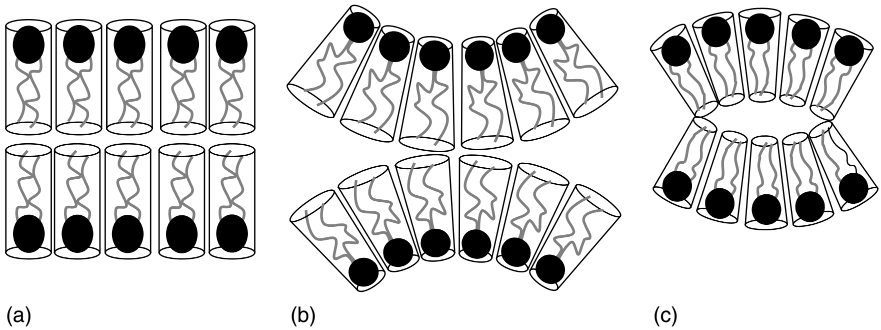

Fig. 10.2 The effective shape concept for lipid molecule geometry. Effectively, cylindrically shaped lipid molecules self-assemble into in stress-free bilayers (a) Whenever the effective shape is noncylindrical, the monolayers will have a propensity to form curved interfaces with $c_0 < 0$ (b) or $c_0 > 0$ (c). Forcing two opposed curving monolayers into a bilayer will result in a curvature-stressed or frustrated bilayer in which energy is stored.

Hence, the monolayer equilibrium curvature is determined by the effective molecular shape, whereas the coupling between monolayers determines the bilayer curvature.

The effective shape image of lipid molecules has been very useful in predicting pure lipid phases (Seddon, 1990; Cullis and deKruijff, 1979) and "curvature-stress" has been described as a major (if not the most important) regulating factor for membrane protein regulation (Gruner, 1985; Stubbs and Slater 1996; Ces and Mulet, 2006). It is, however, important to underline that curvature-stress (given by the "effective values" of K_c and c_0) constitutes only part of the description of mechanical properties of bilayers. Also, assigning shapes is, in fact, a retrospective operation, in the sense that shape and c_0 are based on structural data on spontaneous (macroscopic) phases. Although one may intuitively predict curved phases from say lysophospholipids (Lundbæk and Andersen, 1994), it is not straightforward to generalize the effects of, say, polyunsaturated fatty acids (PUFAs) on membrane properties based on their apparent molecular shape only (Bruno, Koeppe and Andersen, 2007) despite progress in predicting lipid effective structures (and phases) from microscopic models (Li and Schick, 2000).

10.3.2
Transmembrane Spanning Proteins

Integral (bitopic or multitopic) membrane proteins contain one or more regions that allow them to insert partially into, or fully traverse the hydrophobic core of the lipid bilayer. The association with the lipid bilayer implies that detergents or some other nonpolar solvents are required to dissociate them from the lipid matrix cf. (Ollivon *et al.*, 2000). Both classes can be post-translationally modified with added fatty acid chains (acylated), prenyl chains (prenylated), or glycosylphosphatidylinositol (GPI), where these additions serve as to anchor the protein in the lipid bilayer (Resh, 2006; Paulick and Bertozzi, 2008).

10.3.2.1 General Energetic Considerations

For all structures of membrane-spanning peptide chains known to date, the condition of complete satisfied hydrogen bonds is satisfied. The reason for this lies in the fact than an unsatisfied hydrogen bond in a hydrophobic environment is energetically unfavorable. The difference in free energy change for transferring of peptide bonds, with and without hydrogen bonds, amounts to 6.7 kT (Roseman, 1988). For a 20-amino acid segment, this corresponds to 133 kT, which prohibits insertion of nonstructured (nonhydrogen bonded) segments into the membrane. All known integral membrane protein structures traverse the membrane with either α-helices or β-strands producing helical bundle proteins or β-barrel proteins.

10.3.2.2 Helical Bundles

In the case of an α-helical structure with bacteriorhodopsin as the canonical example, the amino acids are arranged in a right-handed structure (favored by steric constraints given by the fact that as all amino acids except glycine have the l-configuration). As each amino acid corresponds to a 100° turn in the helix, the helix has 3.6 residues per turn, corresponding to a shift of 1.5 Å along the helical axis. The N—H group of an amino acid forms a hydrogen bond with the C=O group of the amino acid four residues earlier; this repeated hydrogen bonding pattern defines an α-helix (Pauling, Corey and Branson, 1951). The α-helix is packed with vanishingly little free space within the helix. The amino acid side chains are all on the outside of the helix and oriented toward the *N*-terminus.

How do helix–helix interactions affect the structural arrangement of transmembrane-helical proteins? The recent development of simple transmembrane helical bundle model systems opens possibilities for a quantitative understanding, and several groups have studied these systems. For example, the dimerization of GpA – a small erythrocyte transmembrane protein – has been extensively studied in bilayers, native membranes, and detergent micelles (for a review see Lemmon and Engelman, 1994). Site-directed mutagenesis has revealed an association motif LIXXGVXXGVXXT where X represents any amino acid (Lemmon *et al.*, 1992; Mingarro *et al.*, 1996). Using solid states NMR, the crossing angle between the two monomers of GpA in dodecyl phosphatidylcholine bilayers was found to be 25°. This is in agreement with modeling studies (Treutlein *et al.*, 1992) showing that the two glycines from each monomer line the dimer interface allowing the two helical polar backbones to bind tightly. The valine, leucine, and threonine residues in the motif further extend the interface, and when the helices are docked for maximal interaction they form an interhelical angle of 30–40°.

10.3.2.3 β-Barrels

For β-barrel proteins (e.g. bacterial porins), the secondary and tertiary structure is closely linked as the protein contains β-strands that twist to form a closed structure in which the first strand is hydrogen bonded to the last. The strands are generally arranged in an antiparallel arrangement so that the *N*-terminus of one strand is adjacent to the *C*-terminus of the next. This structure has a high stability because it allows the interstrand hydrogen bonds between carbonyls and amines to be in their preferred planar orientation.

β-Barrels can be classified in terms of two parameters: the number of strands in the β-sheet, n, and the so-called shear number, S, a measure of the stagger of the strands in the β-sheet (Liu, 1998; McLachlan, 1979). n and S are related to the inclination-angle of the β-strands relative to the axis of the barrel. The stagger (or strand tilt) originates in the chirality of the amino acids in the strand. This implies that even though only 7 amino acid residues are needed to span the membrane, typical transmembrane β-strands have 9–11 residues giving a strand tilt between 20° and 45°. The result is a circular pattern where none of the strands are left without a partner. For an eight-strand β-barrel protein, breaking all hydrogen bonds in the bilayer environment would cost 536 kT (8 strands with 10 bonds per strand times 6.7 kT per bond). With the number of strands varying from 8 to 22 implies that β-barrels are extremely stable in an bilayer environment.

10.3.2.4 Interfacial Anchoring of Transmembrane Segments

A ubiquitous feature of both α-helical and β-barrel membrane-spanning proteins seems to be a preference of tryptophan (W) and tyrosine (Y) residues at the interfacial lipid headgroup region (Schiffer, Chang and Stevens, 1992). Statistical analysis of membrane protein sequences indicates that this preference is a general feature (Landolt-Marticorena *et al.*, 1993; von Heijne, 1994). The preference has been discussed in terms of so-called anchoring by flanking residues in the sense that the aromatic or charged residues serve as to symmetrically anchor membrane-spanning proteins at the interfacial regions in the membrane (Schiffer, Chang and Stevens, 1992; Killian and Heijne, 2000; de Planque and Killian, 2003; Reithmeier, 1995). Studies with small membrane-adsorbed model peptides and model transmembrane helical proteins suggest that the preference for aromatic residues at the interface is related to distinct interfacial interactions (Jacobs and White, 1989; Wimley and White, 1996; Kachel, Asuncion-Punzalan and London, 1995; Morein *et al.*, 1997; Yau *et al.*, 1998) including electrostatic and cation–π interactions (Gaede, Yau and Gawrisch, 2005; Petersen, Jensen and Nielsen, 2005).

10.3.2.5 Proteins as Embedded Inclusions

Since the volumetric compressibility moduli for globular proteins in aqueous solution is about $10^{10} - 10^{11}$ N m^{-2} (Gekko and Noguchi, 1979) compared to 10^9 N m^{-2} for liquid-crystalline bilayers (Liu and Kay, 1977) and 10^8 N m^{-1} for bilayer thickness compressibility (Evans and Hochmuth, 1978), both α-helical and β-barrel proteins essentially are rigid structures embedded in a soft fluid membrane. Owing to the architecture of β-barrels, it is not likely that their function involve extensive conformational changes. On the other hand, the function of α-helical proteins typically depends on reorganization of the structure, that is they can convert (discretely or gradually) between various conformational states where the free energy for conformational transitions not only comes from the protein (e.g. helix–helix interactions) but also from "external" sources. One of the sources contributing to the free energy associated with conformational transitions or helical tilting is the membrane environment (Andersen and Koeppe, 2007) because proteins are hydrophobically coupled to the membrane and also in several cases anchored by interfacial aromatic residues. In this regard, single transmembrane-spanning α-helical proteins/peptides constitute a special group of inclusions in the sense that they have the smallest possible hydrophobic inclusion volume and are therefore likely to respond to changes in bilayer properties in ways not possible for larger (helical bundle/barrel) proteins by, for example, tilting the entire protein relative to the bilayer normal or by allowing formation of kinks in the helix. This raises the question as how to assess protein structure (i.e. conformation and helical tilt) *in situ* in the fluid bilayer.

10.4
Determining Membrane-Spanning Protein Structure and Orientation in Fluid Lipid Bilayers

Many techniques including X-ray crystallography, fluorescence resonance energy transfer (FRET), and quenching can be used to study lipid–protein interactions. Fluorescence techniques are especially relevant in studies of partition of peptides into membranes, protein/peptide aggregation, and lipid selectivity (Reyes Mateo *et al.*, 2006). X-ray is a powerful structural technique for membrane proteins *per se* but in the context of lipid–protein interactions it literally provides a frozen picture at low temperature and thus providing information about a membrane environment very different from that of a fluid liquid-crystalline bilayer (Lee, 2005). In the study of integral membrane protein structure and orientation in fluid bilayers, two classes of techniques are especially valuable: NMR-based techniques and electron paramagnetic resonance (EPR) (also referred to as electron spin resonance (ESR)) techniques.

10.4.1
Some Insights from Nuclear Magnetic Resonance (NMR) Spectroscopy

Structural determination using NMR has previously been limited to small proteins. However, the development of sophisticated methods such as transverse relaxation optimized spectroscopy (TROSY) and solution NMR has provided complete structures of bacterial β-barrel proteins such as OmpX, OmpA, and PagP reconstituted in micelles (Fernandez *et al.*, 2004; Arora *et al.*, 2001; Hwang *et al.*, 2002).

Recent developments in aligned-sample solid-state NMR techniques have allowed for detailed information about transmembrane protein conformation. By combining chemical shift and spin–spin coupling, it is possible to obtain detailed information about the orientation of transmembrane α-helices and β-barrels in fluid bilayers (Marassi and Opella, 2000; Gong, Choi and Marassi, 2006; Wang *et al.*, 2000). In order to appreciate this technique, it is to be noted that the angular dependence of the dipolar coupling D is given as $D(\theta) \propto 3\cos^2\theta - 1$, which implies that $D = 0$ at the so-called magic angle: $\theta = \arccos 1/\sqrt{3}$ where the effect of magnetic susceptibility is also minimal. Spinning the samples at the magic angle relative to the $\mathbf{B}$ field has, for example, been a successful tool in determining lipid structure from ^{2}H spectra (Herzfeld *et al.*, 1980; Yeagle and Frye, 1987; Oldfield, Bowers and Forbes, 1987). The principle in magic angle spinning (MAS) has been further developed and is now used in characterization of lipid–protein interactions.

Using a phase and frequency-shifted pulse sequence for the **B** field, it is possible to lock all the ^{1}H spins at the magic angle so that their dipole coupling goes to zero. Simultaneously, another set of synchronous pulses in antiphase allows the ^{15}N nuclei to exchange spin with the ^{1}H nuclei. When this spin exchange has stopped, the ^{15}N magnetizations that are modulated by what happened during spin exchange can be recorded. Fourier transforming from the time to the frequency domain yields a two-dimensional spectrum. This is the so-called polarization inversion with spin exchange at the magic angle (PISEMA) (Wu, Ramamoorthy and Opella, 1994). A two-dimensional ^{1}H-^{15}H dipolar coupling/^{15}N chemical shift PISEMA resonance spectrum displays a characteristic structure referred to as a *polar index slant angle* (*PISA*) diagram (Marassi and Opella, 2000). For an α-helical protein, each point in the wheel-like diagram corresponds to the NH half of each peptide bond (H-bonded to the CO of a peptide four bonds away in the chain) whose ^{15}N chemical shift and ^{1}H $-^{15}$H dipolar coupling both depend on the orientation of the bond with respect to the **B** field. As a result, the shape and position of the PISA wheel is a direct measure of helical orientation (or tilt) in the membrane.

Spectral calculations based on an ideal immobile α-helix or β-strand with uniform dihedral angles and with the helix/strand long axis making a fixed angle with the external **B** field yield precise theoretical PISA diagrams (Gong, Choi and Marassi, 2006), but predicting them from crystallographic data is not straightforward due to the relatively low structural resolution and the high sensitivity of the wheels to local structural variations (shielding variation). This, combined with the fact that membrane proteins are dynamic entities, has led to doubts whether PISA spectra would be generally observed in membrane proteins (Straus, Scott and Watts, 2003). There are now 10 structures in the Protein Data Bank characterized by aligned-sample solid-state NMR techniques (Li *et al.*, 2007), and recent advances in probe technology (Gor'kov *et al.*, 2007) and sample preparation have made it possible to obtain spectra of fully aligned full-length proteins confirming the existence of measurable PISA spectra in membrane proteins.

The use of model peptides in NMR studies has also revealed how isolated membrane-spanning segments orient themselves in lipid bilayer and, in particular, how they respond to changes in the hydrophobic thickness of the bilayer. Specifically, studies of model peptides based on a Acetyl-GW$_2$(LA)$_n$W$_2$A-amide (WALP) and Acetyl-GK$_2$(LA)$_n$K$_2$A-amide (KALP) peptides have provided valuable information (for recent reviews see (de Planque and Killian 2003; Killian and Nyholm, 2006)). For example, insight into the possible role of anchoring residues was obtained by de Planque *et al.* using ^{31}P and ^{2}H NMR (de Planque *et al.*, 1999) who found that W-residues anchor at the interface, whereas K-residues preferentially positions its positively charged amino group close to the lipid phosphate groups. The length of K-flanked peptides can be effectively modulated by "snorkeling," that is stretching of the charged group toward the aqueous phase (de Planque *et al.*, 1999; Strandberg *et al.*, 2002), whereas W-residues resist displacement of their preferred interfacial localization upon a hydrophobic mismatch challenge (de Planque *et al.*, 2003) pointing to the role of W-residues in protein anchoring. ^{15}N NMR studies on lysine (K)-flanked model peptides suggest a considerable tilt (Harzer and Bechinger, 2000) when inserted in a bilayer with a smaller hydrophobic thickness compared to the hydrophobic length of the peptide. Also the *N*-terminal channel forming domain of the membrane protein VpU from HIV-1 showed tilt angles sufficient to compensate for mismatch. In contrast, ^{2}H NMR studies of WALP and KALP only showed a small response to a similar hydrophobic mismatch (Strandberg *et al.*, 2004; Ozdirekcan *et al.*, 2005), although very long WALPs and KALPs adopt helical structures with a central kink in thin lipid bilayers (Daily *et al.*, 2008). Whether this discrepancy is due to differences in experimental parameters or intrinsic peptide properties (e.g. side chain chirality and/or the unnatural poly(AL) sequence in WALP/KALP) still remains an open question.

10.4.2
Some Insights from Electron Paramagnetic Resonance (EPR) Spectroscopy

For more than 30 years, EPR has been used in the studies of lipid organization (Lee, 2005). In these studies, nitroxide spin labels are attached to selected positions in the acyl chains. One of the major results is the detection of so-called boundary or annular lipids, a subpopulation of highly immobilized lipids seen in lipid–protein preparation but not in pure lipid systems (Jost *et al.*, 1973; Jost and Griffith, 1978; Brotherus *et al.*, 1981). The number of annular lipids, that is, the number of lipids that bind to the hydrophobic exterior of a membrane protein, can be quantified (Jost and Griffith, 1978; Brotherus *et al.*, 1981), and the number fits well from estimates based on average area per lipid and protein circumference (Marsh and Horvath, 1998; Marsh and Pali, 2004). However, due to the ESR timescale of 10^{-8} seconds, identification and quantification of annular lipids must be performed at low temperatures (Lee, 2005). At higher temperatures rates for lipid molecule movement in and out of the annular region (described by on and off rate constants) can be detected. The on rate is diffusion controlled, whereas the off rate reflects any specificity in binding (Marsh and Horvath, 1998; East, Melville and Lee, 1985; Davoust and Devaux, 1982). At 30 °C, off rates are about $1–2{\cdot}10^{7}\ s^{-1}$ and this is somewhat lower than the diffusional exchange of $8 \cdot 10^{7}\ s^{-1}$ in the bulk membrane. However, combined with results from fluorescence quenching (East and Lee, 1982), it appears that apart from specific hotspots with selectivity for anionic lipids the lipid–protein interaction, nevertheless, is rather nonspecific (Lee, 2005).

Recent developments in EPR techniques have opened up for investigations of helical peptides incorporated in aligned bicelles (discoid bilayers) (Inbaraj *et al.*, 2006), nanotube arrays (Karp *et al.*, 2006), and bilayers (Inbaraj, Laryukhin and Lorigan, 2007; Mayo *et al.*, 2008). Using the tetra-substituted unnatural amino acid, 2,2,6,6-tetramethyl piperidine 1-oxyl-4-amino-4-carboxylic acid (TOAC) that exhibits highly restricted motion (Inbaraj *et al.*, 2006) attached to the backbone of a transmembrane, helix information about helix tilt angle can be obtained quite accurately (Inbaraj *et al.*, 2006). This method can be applied for peptides and small proteins (<60 amino acids) and appears superior to approaches using standard cystein-bound 1-oxyl-2,2,5,5-tetramethyl-d3-pyrroline-3-methylme-thane-thiosulfonate spin label (MTSSL) in which there are several flexible bonds between the helical peptide backbone and the nitroxide group. Also, introducing the TOAC label in membrane proteins does not alter secondary structure (Karim *et al.*, 2004).

Using the 23-amino acid, long M2δ domain of the nicotinic acetylcholine receptor (AchR) ion channel as a model for a transmembrane helical segment, the tilting value could be compared with known helical tilt information based on the PISEMA approach mentioned in Section 4.1. The M2δ segment was found to tilt 14° with respect to the bilayer normal both in bicelles and bilayers, consistent with NMR data (Inbaraj *et al.*, 2006). The advantage of the TOAC EPR approach is that it is orders of magnitude more sensitive than ^{15}N NMR solid-state spectroscopy (Inbaraj *et al.*, 2006). Further, the TOAC label be placed at several positions along the helix and tilt angles can be determined more accurately. Future work could thus potentially resolve issues related to WALP/KALP helical tilt as mentioned in Section 4.1.

10.5
Elastic Membrane Deformations and Lipid–Protein Interactions

The hydrophobic matching between lipids and proteins, which entails a hydrophobic coupling between the bilayer mechanical properties of protein function, implies that the hydrophobic bilayer thickness d close to the protein inclusion should closely match the hydrophobic length l of the protein. This matching serves to minimize the energetic penalty associated with exposing hydrophobic lipid or protein moieties to the aqueous phase (see Figure 10.1). If $l \neq d_0$ (the bilayer equilibrium thickness), then $d \neq d_0$, that is, the "soft" bilayer responds to the presence of the "rigid" inclusion. Thus the bilayer deformation generally will involve both CE and SD of the lipid acyl chains in the two opposing monolayers.

Even though the bilayer can be considered as soft, the mechanical properties of lipid bilayers, nevertheless, entail that the bilayer deformation energy can regulate protein function by affecting the free energy associated with protein conformational changes ΔG_{int}, which has contributions from both the protein and the deformation energy of the bilayer ΔG_{def} (Andersen and Koeppe, 2007; Andersen *et al.*, 1999):

$$\Delta G = \Delta G_{\text{int}} + \Delta G_{\text{def}} \tag{8}$$

This regulation is bidirectional: the lipid bilayer can respond to a protein, whereas the protein may in turn respond to the bilayer by adjusting its conformation (effective hydrophobic dimensions) or in the case of single spanning α-helices, its helical tilt relative to the bilayer normal.

10.5.1
Quantifying Bilayer Deformation Energy

The bilayer deformation energy can be estimated using the continuum theory of elastic bilayer deformations (Huang,

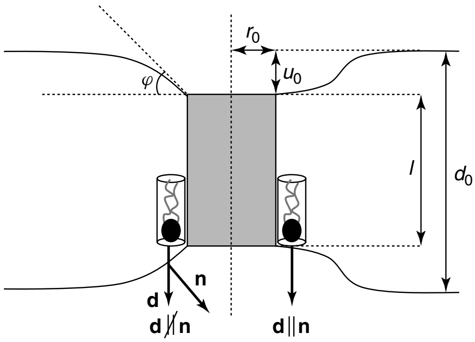

Fig. 10.3 Inclusion-induced bilayer deformation. The total deformation free energy is determined by the bilayer deformation profile associated with a hydrophobic mismatch between the bilayer and the inclusion. The inclusion is treated as a rotationally symmetrical deformation of the bilayer. Assuming midplane symmetry (a symmetrical bilayer), the problem can be reduced to a radially varying deformation of a monolayer with unperturbed thickness $d_0/2$, where $u(r)$ denotes the local perturbation in monolayer thickness at the distance r from the luminal axis. At the channel–bilayer contact surface, r_0, the deformation is u_0. The first derivative at the contact surface (the contact slope) du/dr is $s = \tan(\varphi)$. The deformation profile to the left of the inclusion corresponds to $\mathbf{d} \neq \mathbf{n}$ and $s = s_{\rm min}$, the profile to the right of the protein correspond to $\mathbf{d} = \mathbf{n}$ and $s = 0$. The deformation profiles obtained using the elastic membrane model define the average hydrocarbon/headgroup boundary. The actual (instantaneous) profile will be less well-defined due to thermal motion of the bilayer lipids perpendicular to the bilayer/solution interface (Wiener and White, 1992).

1986; Helfrich and Jakobsson, 1990). The energy density has contributions from CE (Eq. (2)) and bending (Eq. (3)). Combining these contributions, the total free energy of deformation for a cylindrical protein with radius r_0 can be written as

$$\begin{aligned}\Delta G_{\rm def} = {}& \frac{1}{2}\int_{r_0}^{\infty}\left[K_a\left(\frac{2u}{d_0}\right)^2 + K_c(c_1\right.\\ &\left. + c_2 - c_0)^2 + \Gamma\left(\frac{du}{dr}\right)^2\right]2\pi r dr\\ &- \frac{1}{2}\int_{r_0}^{\infty} K_c c_0^2 2\pi r dr \pi \int_{r_0}^{\infty}\\ &\times\left[K_a\left(\frac{2u}{d_0}\right)^2 + K_c(c_1 + c_2)^2\right.\\ &\left. - K_c(c_1 + c_2)c_0 + \Gamma\left(\frac{du}{dr}\right)^2\right] r dr\end{aligned} \tag{9}$$

where $c_1 = 1/R_1$ and $c_2 = 1/R_2$. To determine the minimal free energy, one must minimize Eq. (9) with respect to variations in $u(r)$. Four boundary conditions are needed to solve Eq. (9).

Three of these are self-evident. Far away from the protein $u = 0$ and $du/dr = 0$ for $r \to \infty$. Close to the protein, hydrophobic matching implies that $u(r_0) = (d_0 - 1)/2$. The fourth boundary condition $du/dr(r = r_0) = s = \tan(\varphi)$ can be formulated in several ways (Huang, 1986; Helfrich and Jakobsson, 1990). Two extreme cases can be considered (Nielsen, Goulian and Andersen, 1998). A “free slope” boundary condition, is one where the slope is free to assume a value that will minimize s. This $s = s_{\rm min}$ condition implies that the lipids are allowed to tilt, that is, have their director **d** deviating from the bilayer surface normal vector **n**. The other extreme case is a "constrained slope" boundary condition, which generally implies that $\mathbf{d}|\mathbf{n}$ and specifically for cylindrical inclusion, $s = 0$ (see Figure 10.3).

It turns out that the c_1c_2-dependent Gaussian term can be neglected (Nielsen and Andersen, 2000) and that the c_0-dependent term in Eq. (9) only depends on s (Ring, 1996) corresponding to an energy term

$$\Delta G_{\mathrm{MEC}} = 2\pi K_c r_0 s c_0 \tag{10}$$

Thus, we may minimize ΔG_{def} by first solving the Lagrange equation corresponding Eq. (9) with $c_0 = 0$, and subsequently adding ΔG_{MEC} to the total minimized ΔG_{def}. The Lagrange equation is

$$K_c\nabla^4 u - \Gamma\nabla^2 u - (K_a/d_0^2)u = 0 \tag{11}$$

From Eq. (11), a solution can be obtained analytically (Nielsen, Goulian and Andersen, 1998) as

$$\Delta G_{\mathrm{def}} = \pi r_0 K_c \left[s\left[\frac{A_+K_0\left(\frac{\sqrt{i}r_0}{\lambda}\right) - A_-K_0\left(\frac{\sqrt{-i}r_0}{\lambda}\right)}{i\lambda^2}\right] + u_0\left[\frac{A_+K_1\left(\frac{\sqrt{i}r_0}{\lambda}\right) + iA_-K_1\left(\frac{\sqrt{-i}r_0}{\lambda}\right)}{\sqrt{i}\lambda^3}\right] + \Gamma u_0 s \right] \tag{12}$$

where K_0 and K_1 are the modified first-order Bessel functions of the second kind and

$$\lambda = \sqrt[4]{\frac{d_0^2 K_c}{K_a}} \tag{13a}$$

and

$$A_+ = \frac{u_0K_1\left(\frac{\sqrt{-i}r_0}{\lambda}\right) - sK_0\left(\frac{\sqrt{-i}r_0}{\lambda}\right)\frac{\lambda}{\sqrt{-i}}}{K_0\left(\frac{\sqrt{i}r_0}{\lambda}\right)K_1\left(\frac{\sqrt{-i}r_0}{\lambda}\right) - iK_0\left(\frac{\sqrt{-i}r_0}{\lambda}\right)K_1\left(\frac{\sqrt{i}r_0}{\lambda}\right)} \tag{13b}$$

and

$$A_- = \frac{-iu_0K_1\left(\frac{\sqrt{i}r_0}{\lambda}\right) - sK_0\left(\frac{\sqrt{i}r_0}{\lambda}\right)\frac{\lambda}{\sqrt{-i}}}{K_0\left(\frac{\sqrt{i}r_0}{\lambda}\right)K_1\left(\frac{\sqrt{-i}r_0}{\lambda}\right) - iK_0\left(\frac{\sqrt{-i}r_0}{\lambda}\right)K_1\left(\frac{\sqrt{i}r_0}{\lambda}\right)} \tag{13c}$$

Rewriting, we get

$$\begin{aligned}\Delta G_{\mathrm{def}} &= -\pi r_0 K_{\mathrm{c}} \\ &\left[\frac{u_0^2[K_1(r_0/(\sqrt{-i}\lambda))K_1(r_0/\sqrt{i}\lambda) - K_1(r_0/(\sqrt{i}\lambda))K_1(r_0/\sqrt{-i}\lambda)]/\lambda^4}{K_0(r_0/\sqrt{i}\lambda)K_1(r_0/(\sqrt{-i}\lambda))/(\sqrt{-i}\lambda) - K_0(r_0/\sqrt{-i}\lambda)K_1(r_0/(\sqrt{i}\lambda))/(\sqrt{i}\lambda)}\right. \\ &+ \frac{su_0[K_0(r_0/(\sqrt{-i}\lambda))K_1(r_0/\sqrt{i}\lambda) - K_1(r_0/\sqrt{i}\lambda)K_0(r_0/\sqrt{-i}\lambda)]/(\sqrt{i}\lambda^3)}{K_0(r_0/\sqrt{i}\lambda)K_1(r_0/(\sqrt{-i}\lambda))/(\sqrt{-i}\lambda) - K_0(r_0/\sqrt{-i}\lambda)K_1(r_0/(\sqrt{i}\lambda))/(\sqrt{i}\lambda)} \\ &+ \frac{s^2[K_0(r_0/\sqrt{-i}\lambda)K_0(r_0/\sqrt{i}\lambda) - K_0(r_0/\sqrt{i}\lambda)]K_0(r_0/\sqrt{-i}\lambda)]/\lambda^2}{K_0(r_0/\sqrt{i}\lambda)K_1(r_0/(\sqrt{-i}\lambda))/(\sqrt{-i}\lambda) - K_0(r_0/\sqrt{-i}\lambda)K_1(r_0/(\sqrt{i}\lambda))/(\sqrt{i}\lambda)} \\ &\left. + \Gamma u_0 s\right] \qquad (14)\end{aligned}$$

For realistic values of Γ, it turns out that the relative energetic contribution from surface tension can be neglected (Nielsen, Goulian and Andersen, 1998). Consequently Eq. (14) is effectively biquadratic in u_0 and s.

$$\Delta G_{\mathrm{def}}(u_0, s) = a_1 u_0^2 + a_2 u_0 s + a_3 s^2 \qquad (15)$$

For both $s = s_{\min}$ and $s = 0$, Eq. 15 reduces to a linear spring formalism with a phenomenological coefficient H

$$\Delta G_{\mathrm{def}} = H(2u_0)^2 \qquad (16a)$$

With $$H_{s=s_{\min}} = a_1 - a_2^2/4a_3 \qquad (16b)$$

and $$H_{s=0} = a_1/4, \qquad (16c)$$

respectively. In general, with the added ΔG_{MEC} contribution we get

$$\begin{aligned}\Delta G_{\mathrm{def}}(u_0, s) &= a_1 u_0^2 + a_2 u_0 s + a_3 s^2 \\ &\quad + 2\pi K_c r_0 s c_0 \qquad (17)\end{aligned}$$

where the a coefficients are functions of the material moduli, bilayer thickness, and inclusion (protein) radius. Equation (17) can be re-written as

$$\begin{aligned}\Delta G_{\mathrm{def}}(u_0, s) &= a_1 u_0^2 + (a_2 u_0 \\ &\quad + 2\pi K_c r_0 c_0)s + a_3 s^2 \qquad (18)\end{aligned}$$

For both free and constrained slope boundary conditions we can write ΔG_{def} as

$$\begin{aligned}\Delta G_{\mathrm{def}} &= H_B(d_0 - l)^2 \\ &\quad + H_X(d_0 - l)^2 c_0 + H_C c_0^2 \\ \Delta G_{\mathrm{def}} &= H_b(d_0 - l)^2 \\ &\quad + H_X(d_0 - l)^2 c_0 + H_C c_0^2 \qquad (19)\end{aligned}$$

where H_B, H_X and H_C are functions of the a coefficients, r_0, and K_c (Andersen and Koeppe, 2007; Nielsen and Andersen, 2000).

Formally, we can express the coefficients in Eq. (18) as

$$a_{\mathrm{i}} = K_{\mathrm{c}} r_0 f_{\mathrm{i}}(r_0/\lambda) \qquad (20)$$

where

$$f_1(r_0/\lambda) = \frac{\pi[K_1(r_0/(\sqrt{-i}\lambda))K_1(r_0/\sqrt{i}\lambda) - K_1(r_0/(\sqrt{i}\lambda))K_1(r_0/\sqrt{-i}\lambda)]/\lambda^4}{K_0(r_0/\sqrt{i}\lambda)K_1(r_0/(\sqrt{-i}\lambda))/(\sqrt{-i}\lambda) - K_0(r_0/\sqrt{-i}\lambda)K_1(r_0/(\sqrt{i}\lambda))/(\sqrt{i}\lambda)} \tag{21a}$$

$$f_2(r_0/\lambda) = \frac{\pi[K_0(r_0/(\sqrt{-i}\lambda))K_1(r_0/\sqrt{i}\lambda) - K_1(r_0/\sqrt{i}\lambda)K_0(r_0/\sqrt{-i}\lambda)]/(\sqrt{i}\lambda^3)}{K_0(r_0/\sqrt{i}\lambda)K_1(r_0/(\sqrt{-i}\lambda))/(\sqrt{-i}\lambda) - K_0(r_0/\sqrt{-i}\lambda)K_1(r_0/(\sqrt{i}\lambda))/(\sqrt{i}\lambda)} \tag{21b}$$

$$f_3(r_0/\lambda) = \frac{\pi[K_0(r_0/\sqrt{-i}\lambda)K_0(r_0/\sqrt{i}\lambda) - K_0(r_0/\sqrt{i}\lambda)]K_0(r_0/\sqrt{-i}\lambda)]/\lambda^2}{K_0(r_0/\sqrt{i}\lambda)K_1(r_0/(\sqrt{-i}\lambda))/(\sqrt{-i}\lambda) - K_0(r_0/\sqrt{-i}\lambda)K_1(r_0/(\sqrt{i}\lambda))/(\sqrt{i}\lambda)} \tag{21c}$$

Equations (15)–(21) constitute a general formalism in which bilayer deformation energy can be estimated using estimates for K_a, K_c, d_0, l, r_0, c_0, and s. It is important to note that ΔG_{def} arises as a minimization of Eq. (9) for both the constrained and relaxed boundary condition (Nielsen, Goulian and Andersen, 1998). This implies that the energy components (CE and bending) are interrelated and one cannot interpret ΔG_{def} (and the resulting bilayer deformation profile) as arising from only one energy component or descriptor (e.g. thickness or curvature) of bilayer properties.

10.5.2
Bilayer Deformation Energy and Protein Function

Reliable quantification of bilayer deformation energies is based on macroscopic bilayer parameters K_a, K_c, and c_0. This raises two concerns: are macroscopic parameters valid in a description of bilayer deformations on length-scales corresponding to the radial extent of annular lipids (see Section 4.2) or the bilayer thickness? If they are, how will experimental uncertainties in these parameters affect the estimate for ΔG_{def}? and thus any predictions of how the bilayer may regulate protein function? Furthermore, are there experimental model systems in which the elastic theory and ΔG_{def} estimates can be evaluated?

10.5.2.1 Model Systems

A useful experimental tool for investigating elastic bilayer formation takes advantage of the unique formation of gramicidin ion channels in lipid bilayers (Andersen *et al.*, 1999; Huang 1986; Goulian *et al.*, 1998; Lundbaek and Andersen, 1999). Gramicidin channels are formed by transmembrane dimerization of two nonconducting $\beta^{6.3}$-helical subunits (O'Connell, Koeppe and Andersen, 1990) (see Figure 10.4). They are permeable to monovalent cations and channel formation (channel average lifetimes) is detectable using standard voltage clamp techniques (Andersen, 1983). As the hydrophobic length of the gramicidin dimer (22 Å) is shorter than the

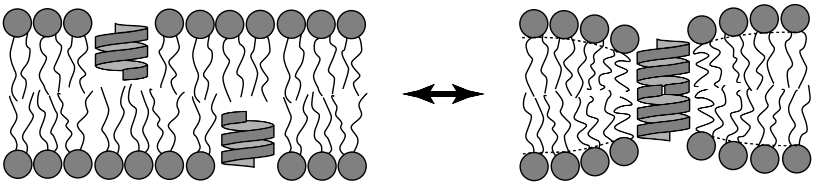

Fig. 10.4 Gramicidins as molecular force transducers. Two nonconducting dimers (left) join and form a channel (right). Because channel formation involves a well-defined deformation of the bilayer, the gramicidin channel average lifetime, which is detectable using standard voltage-clamp techniques, will vary as a function of the hydrophobic mismatch through Eq. (22) (Andersen and Koeppe, 2007; Andersen, 1999).

hydrophobic thickness of the bilayer (which in turn is dependent on, e.g., acyl chain length and solvent properties), channel formation leads to a well-defined deformation of the bilayer. Thus gramicidin can be used as molecular force transducers (Andersen *et al.*, 1999) detecting the disjoining force on the dimer exerted by the bilayer:

$$\begin{aligned} F &= -\left(-\frac{\partial \Delta G}{\partial (d_0 - l)}\right) \\ &= 2H_B(d_0 - l) + H_X c_0 \qquad (22) \end{aligned}$$

The elastic description has been tested using gramicidins as a function of changes in d_0 implemented as changes in acyl chain length and alkane solvent length (Huang 1986; Lundbaek and Andersen, 1999; Elliott *et al.*, 1983) or tension (effectively thinning the bilayer) (Goulian *et al.*, 1998).

Using independently determined values for the mechanical moduli, quantitative agreement could be obtained (Goulian *et al.*, 1998; Lundbaek and Andersen, 1999) assuming that the energetic cost of tilt ($\mathbf{d} \not\parallel \mathbf{n}$) is high, thereby effectively constraining the lipid packing close to the protein and increasing ΔG_{def} (Huang, 1986; Lundbaek and Andersen, 1999). Although the "conformational change" involved in gramicidin channel dimerization is unique and the dimerization always is associated with a dimpling of the bilayer, gramicdins, nevertheless, can be used to quantify the deformation energy (i.e. the phenomenological spring constants).

10.5.2.2 **Integral Membrane Proteins**

As described in Section 10.4, the structures and conformational changes involved in function of integral proteins are generally difficult to detect in fluid lipid bilayers. While β-barrel proteins appear to be rather "inert" in the sense that their function is unlikely to involve major conformational changes, the function of α-helical bundle proteins is likely to involve sliding and tilting of their transmembrane segments (Unwin, 1989; Kaback and Wu, 1997; Sakmar, 1998). Part of the free energy for these conformational changes depends on intrinsic protein factors (Eq. (8)) such as unbinding of the helix–helix association motif (see Section 10.3.2.2). The open–close transition in gap junction channels, for example, involves a tilt of the domains by about 8° (Unwin and Ennis, 1984) corresponding to a hydrophobic length change of 0.3 Å. Both the open and closed states are likely to perturb the bilayer with deformation energies $\Delta\Delta G_{\text{def,o}}$ and $\Delta\Delta G_{\text{def,c}}$

$$\Delta\Delta G_{\text{def}} = \Delta G_{\text{def,o}} - \Delta G_{\text{def,c}} \qquad (23)$$

The closed → open transition probability is

$$P = \frac{1}{1 + K^*_{c\to 0}\exp(\Delta\Delta G_{\text{def}}/kT)} \qquad (24)$$

where $K^*_{c\to 0}$ is the intrinsic equilibrium constant of the transition. Equations (23) and (24) constitute the link between bilayer deformation energy (as determined by the bilayer material/elastic properties) and integral membrane protein function. The analysis shows that the energies can be considerable: for a protein embedded in a bilayer with an initial mismatch of 1 Å, an increase to 1.3 Å increases P by 10-fold (Nielsen, Goulian and Andersen, 1998). Also recent NMR studies on rhodophsin show that the protein can adjust its helical content to match changes in bilayer hydrophobic thickness (Soubias *et al.*, 2008) showing that deformation energies are sufficient to change H-bonding patterns and that the hydrophobic matching is reciprocal: both protein and bilayer can respond to it.

The link between ΔG_{def} and protein function (Eqs (23) and (24)) reflects not only hydrophobic matching but also the hydrophobic coupling between bilayer material properties and membrane protein function. Using this framework as a basis, it is possible to rationalize the effect on protein function of *changing* the bilayer material moduli (e.g. by incorporating pharmacological amphiphatic compound), for a comprehensive overview see Lundbaek (2006).

10.6
Summary and Perspectives

Support for a physical (mechanical) view of the membrane comes from reconstitution experiments where membrane proteins can be inserted into lipid bilayers consisting of only one or a few lipid species, for example, see Criado, Eibl and Barrantes (1984); Johannsson, Smith and Metcalfe (1981). This suggests that membrane protein function is not necessarily crucially dependent on specific lipid–protein interactions. Nevertheless, the function of many different types of bilayer embedded proteins can be modulated by maneuvers that changes bilayer mechanical properties, see Andersen and Koeppe (2007) and references therein.

The elastic description of protein-induced bilayer deformation provides a strong basis for understanding how the bilayer can regulate protein function. This understanding is both qualitative and quantitative. With calculated deformation energies, it is possible to predict how the bilayer properties and, in particular, how induced changes in these properties will affect protein function. However, it is important to stress the energetic consequences of lipid packing around the protein. The effective packing translates into an effective slope at the protein-bilayer boundary, which again determines ΔG_{def}. In fact, $\Delta\Delta G_{\text{def}}(c_0)$ (Eq. (23)) and thus $P(c_0)$ (Eq. (24)) can be either increasing or decreasing functions of c_0 dependent on the choice of boundary condition (Nielsen and Andersen, 2000). Among the future challenges is the need for a better understanding of the energetic consequences of local lipid packing and more specific interactions. Binding of lipids in crystal structures of membrane proteins as seen, for example, for bacteriorhopsin (Corcelli *et al.*, 2002), photosynthetic reaction centers (Nogi *et al.*, 2000; Fyfe *et al.*, 2004), and aquaporin (Gonen *et al.*, 2005) could be taken as evidence for specific lipid–protein interactions, but for lipids binding so tightly, they should, in fact, be regarded as structural cofactors (Lee, 2005).

The fact that the membrane deformation energy is an unspecific regulator of protein function depending on local lipid packing around the protein also resolves the apparent problem that a stored elastic energy (curvature-stress) is a thermodynamic property of the entire bilayer. Thus, any change in the bilayer deformation energy would change the conformational equilibrium of all proteins in a cell membrane, which would compromise the cell's control of the membrane and its function. This would indeed be the case if all proteins were having similar hydrophobic shapes and thus "felt" the deformation energy in the same manner through similar boundary conditions and the membrane has a uniform structure. The increasing number of high-resolution membrane protein structures shows that despite obvious similarities (e.g. the occurrence of interfacial aromatic residues) transmembrane protein structures display a substantial structural variety on the Å to nanometer scale. Also, despite considerable controversies, there is evidence that the lateral heterogeneity of the membrane leads to various domains with different lipid–protein composition and thus mechanical properties (Simons and Ikonen, 1997; Brown and London, 1998; Jacobson, Mouritsen and Anderson, 2007; Edidin, 2003). Thus, although a general change in the bilayer deformation energy is global, the "sensing" of this energy is different for different protein structures located in different domains (Lundbaek *et al.*, 2003). Nevertheless the general mechanism is the same: hydrophobic matching leads to hydrophobic coupling between protein function and bilayer properties and the energies involved are of sufficient magnitude to regulate protein structure and function.

References

Andersen, O.S. and Koeppe II, R.E. (2007) Bilayer thickness and membrane protein function: an energetic perspective. *Annu. Rev. Biophys. Biomol. Struct.*, **36**, 107–30.

Andersen, O.S., Sawyer, D.B., and Koeppe II, R.E. (1992) Modulation of channel function by the host bilayer. *Biomembr. Struct. Funct.*, 227–44.

Andersen, O.S. (1983) Ion movement through gramicidin a channels. Single-channel measurements at very high potentials. *Biophys. J.*, **41**, 119–33.

Andersen, O.S. *et al.* (1999) Ion channels as tools to monitor lipid bilayer-membrane protein interactions: gramicidin channels as molecular force transducers. *Methods Enzymol.*, **294**, 208–24.

Andrews, F.M. and Pfeffer, W. (1929) *Plant Physiol.*, **4**, 284–88.

Andrews, D.M. and Haydon, D.A. (1968) Electron microscope studies of lipid bilayer membranes. *J. Mol. Biol.*, **32**(1), 149–50.

Arora, A. *et al.* (2001) Structure of outer membrane protein a transmembrane domain by NMR spectroscopy. *Nat. Struct. Biol.*, **8**(4), 334–38.

Bamberg, E. and Benz, R. (1976) Voltage-induced thickness changes of lipid bilayer membranes and the effect of an electric field on gramicidin a channel formation. *Biochim. Biophys. Acta*, **426**, 570–80.

Barton, P.G. and Gunstone, F.D. (1975) Hydrocarbon chain packing and molecular motion in phospholipid bilayers formed from unsaturated lecithins. Synthesis and properties of sixteen positional isomers of 1,2-dioctadecenoyl-sn-glycero-3-phosphorylcholine. *J. Biol. Chem.*, **250**(12), 4470–76.

Benz, R. and Janko, K. (1976) Voltage-induced capacitance relaxation of lipid bilayer membranes. Effects of membrane composition. *Biophys. Biochim. Acta*, **455**, 721–38.

Brotherus, J.R. *et al.* (1981) Lipid–protein multiple binding equilibria in membranes. *Biochemistry*, **20**(18), 5261–67.

Brown, D.A. and London, E. (1998) Functions of lipid rafts in biological membranes. *Annu. Rev. Cell Dev. Biol.*, **14**, 111–36.

Bruno, M.J., Koeppe, R.E., and Andersen II, O.S. (2007) Docosahexaenoic acid alters bilayer elastic properties. *Proc. Natl. Acad. Sci. U.S.A.*, **104**(23), 9638–43.

Cantor, R. (1997a) The lateral pressure profile in membranes: a physical mechanism of general anesthesia. *Biochemistry*, **36**, 2339–44.

Cantor, R. (1997b) Lateral pressures in cell membranes: a mechanism for modulation of protein function. *J. Phys. Chem. B*, **101**, 1723–25.

Cantor, R. (1999) Lipid composition and the lateral pressure profile in bilayers. *Biophys. J.*, **76**, 2625–39.

Ces, O. and Mulet, X. (2006) Physical coupling between lipids and proteins: a paradigm for cellular control. *Signal Transduct.*, **6**, 112–32.

Cevc, G. and Marsh, D. (1987) in *Phospholipid bilayers*, (ed G. Cevc), John Wiley & Sons, New York, pp. 34–36.

Corcelli, A. *et al.* (2002) Lipid–protein stoichiometries in a crystalline biological membrane: NMR quantitative analysis of the lipid extract of the purple membrane. *J. Lipid Res.*, **43**(1), 132–40.

Criado, M., Eibl, H., and Barrantes, F.J. (1984) Functional properties of the acetylcholine receptor incorporated in model lipid membranes. Differential effects of chain length and head group of phospholipids on receptor affinity states and receptor-mediated ion translocation. *J. Biol. Chem.*, **259**, 9188–98.

Cullis, P.R. and deKruijff, B. (1979) Lipid polymorphism and the functional roles of lipids in biological membranes. *Biochim. Biophys. Acta*, **559**, 399–420.

Daily, A.E. *et al.* (2008) Helical distortion in tryptophan- and lysine-anchored membrane-spanning alpha-helices as a function of hydrophobic mismatch: a solid-state deuterium NMR investigation using the geometric analysis of labeled alanines method. *Biophys. J.*, **94**(2), 480–91.

Danielli, J.F. and Davson, H. (1935) A contribution to the theory of permeability of thin films. *J. Cell. Comp. Physiol.*, **5**, 495–508.

Davoust, J. and Devaux, P.F. (1982) *J. Magn. Reson.*, **48**, 475–94.

de Planque, M.R. and Killian, J.A. (2003) Protein–lipid interactions studied with designed transmembrane peptides: role of hydrophobic matching and interfacial anchoring. *Mol. Membr. Biol.*, **20**(4), 271–84.

de Planque, M.R. *et al.* (1999) Different membrane anchoring positions of tryptophan and lysine in synthetic transmembrane alpha-helical peptides. *J. Biol. Chem.*, **274**(30), 20839–46.

de Planque, M.R. *et al.* (2003) Interfacial anchor properties of tryptophan residues in transmembrane peptides can dominate over hydrophobic matching effects in peptide–lipid interactions. *Biochemistry*, **42**(18), 5341–48.

Dowhan, W. (1997) Molecular basis for membrane phospholipid diversity: why are there so many lipids? *Annu. Rev. Biochem.*, **66**, 199–232.

Doyle, D.A. *et al.* (1998) The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science*, **280**, 69–77.

East, J.M. and Lee, A.G. (1982) Lipid selectivity of the calcium and magnesium ion dependent adenosinetriphosphatase, studied with fluorescence quenching by a brominated phospholipid. *Biochemistry*, **21**(17), 4144–51.

East, J.M., Melville, D., and Lee, A.G. (1985) Exchange rates and numbers of annular lipids for the calcium and magnesium ion dependent adenosinetriphosphatase. *Biochemistry*, **24**(11), 2615–23.

Edidin, M. (2003) The state of lipid rafts: from model membranes to cells. *Annu. Rev. Biophys. Biomol. Struct.*, **32**, 257–83.

Elliott, J.R. *et al.* (1983) The effects of bilayer thickness and tension on gramicidin single-channel lifetime. *Biochim. Biophys. Acta*, **735**, 95–103.

Engelman, D.M. and Rothman, J.E. (1972) The planar organization of lecithin-cholesterol bilayers. *J. Biol. Chem.*, **247**(11), 3694–97.

Evans, E.A. and Hochmuth, R.M. (1978) Mechanochemical properties of membranes. *Curr. Top. Membr. Transp.*, **10**, 1–64.

Evans, E. and Skalak, R. (1980) *Mechanics and thermodynamics of biomembranes*, CRC Press Boca Raton.

Fernandez, C. *et al.* (2004) NMR structure of the integral membrane protein OmpX. *J. Mol. Biol.*, **336**(5), 1211–21.

Fettiplace, R. and Haydon, D.A. (1980) Water permeability of lipid membranes. *Physiol. Rev.*, **60**(2), 510–50.

Fettiplace, R., Andrews, D.M., and Haydon, D.A. (1971) *J. Membr. Biol.*, **5**, 277–96.

Fyfe, P.K. *et al.* (2004) Disruption of a specific molecular interaction with a bound lipid affects the thermal stability of the purple bacterial reaction centre. *Biochim. Biophys. Acta*, **1608**(1), 11–22.

Gaede, H.C., Yau, W.-M., and Gawrisch, K. (2005) Electrostatic contributions to indole–lipid interactions. *J. Phys. Chem. B*, **109**, 13014–23.

Gekko, K. and Noguchi, H. (1979) Compressibility of globular proteins in water at 25 °C. *J. Phys. Chem.*, **83**, 2706–14.

Gonen, T. *et al.* (2005) Lipid–protein interactions in double-layered two-dimensional AQP0 crystals. *Nature*, **438**(7068), 633–38.

Gong, X.-M., Choi, J., and Marassi, F.M. (2006) NMR of membrane proteins in lipid environments: the Bcl-2 family of apoptosis regulators, in *Protein–lipid interactions* (eds C. Reyes Mateo, J. Gómez, J. Villalain, and J.M. Gonzales Ros) Springer, Berlin Heidelberg, pp. 35–61.

Gor'kov, P.L. *et al.* (2007) Using low-E resonators to reduce RF heating in biological samples for static solid-state NMR up to 900 MHz. *J. Magn. Reson.*, **185**(1), 77–93.

Gorter, E. and Grendel, F. (1925) On bimolecular layers of lipids on chromatocytes of blood. *J. Exp. Med.*, **41**, 439–43.

Goulian, M. *et al.* (1998) Gramicidin Channels Kinetics under Tension. *Biophys. J.*, **74**, 328–37.

Green, D.E., ed. *Membrane structure and its biological applications*. Ann. N.Y. Acad. Sci. Vol. 195. 1972: The New York Academy of Science, New York.

Green, W.N., Weiss, L.B., and Andersen, O.S. (1987) Batrachotoxin-modified sodium channels in planar lipid bilayers. Ion permeation and block. *J. Gen. Physiol.*, **89**, 841–72.

Gruner, S.M. (1985) Intrinsic curvature hypothesis for biomembrane lipid composition: a role for nonbilayer lipids. *Proc. Natl. Acad. Sci. U.S.A*, **82**, 3665–69.

Harzer, U. and Bechinger, B. (2000) Alignment of lysine-anchored membrane peptides under conditions of hydrophobic mismatch: a CD, ^{15}N and ^{31}P solid-state NMR spectroscopy investigation. *Biochemistry*, **39**(43), 13106–14.

Hazel, J.R. (1995) Thermal adaption in biological membranes: is homeoviscous adaption the explanation? *Annu. Rev. Physiol.*, **57**, 19–42.

Helfrich, W. (1973) Elastic properties of lipid bilayers: theory and possible experiments. *Z. Naturforsch.*, **28C**, 693–703.

Helfrich, P. and Jakobsson, E. (1990) Calculation of deformation energies and conformations in lipid membranes containing gramicidin channels. *Biophys. J.*, **57**, 1075–84.

Herzfeld, J. *et al.* (1980) Magic angle sample spinning in inhomogeneously broadened biological systems. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **289**(1037), 459–69.

Huang, H.W. (1986) Deformation free energy of bilayer membrane and its effect on gramicidin channel lifetime. *Biophys. J.*, **50**, 1061–70.

Hwang, P.M. *et al.* (2002) Solution structure and dynamics of the outer membrane enzyme PagP by NMR. *Proc. Natl. Acad. Sci. U.S.A.*, **99**(21), 13560–65.

Inbaraj, J.J., Laryukhin, M., and Lorigan, G.A. (2007) Determining the helical tilt angle of a transmembrane helix in mechanically aligned lipid bilayers using EPR spectroscopy. *J. Am. Chem. Soc.*, **129**(25), 7710–11.

Inbaraj, J.J. *et al.* (2006) Determining the topology of integral membrane peptides using EPR spectroscopy. *J. Am. Chem. Soc.*, **128**(29), 9549–54.

Israelachvili, J., ed. *Intermolecular & Surface Forces*. 1992, Academic Press: London.

Israelachvili, J.N., Marcelja, S., and Horn, R.G. (1980) Physical principles of membrane organization. *Q. Rev. Biophys.*, **13**(2), 121–200.

Israelachvili, J.N., Mitchell, D.J., and Ninham, B.W. (1976) Theory of self-assembly of hydrocarbon amphiphiles into micelles and bilayers. *J. Chem. Soc., Faraday Trans. 2*, **72**, 1525–68.

Jacobs, R.E. and White, S.H. (1989) The nature of the hydrophobic binding of small peptides at the bilayer interface: implications for the insertion of transbilayer helices. *Biochemistry*, **28**, 3421–37.

Jacobson, K., Mouritsen, O.G., and Anderson, R.G. (2007) Lipid rafts: at a crossroad between cell biology and physics. *Nat. Cell Biol.*, **9**(1), 7–14.

Jiang, Y. *et al.* (2003) X-ray structure of a voltage-dependent K^+ channel. *Nature*, **423**(6935), 33–41.

Johannsson, A., Smith, G.A., and Metcalfe, J.C. (1981) The effect of bilayer thickness on the activity of (Na^{++} K^+) -ATPase. *Biochim. Biophys. Acta*, **641**, 416–21.

Jost, P.C. and Griffith, O.H. (1978) The spin-labeling technique. *Methods Enzymol.*, **49**, 369–418.

Jost, P.C. *et al.* (1973) Evidence for boundary lipid in membranes. *Proc. Natl. Acad. Sci. U.S.A.*, **70**(2), 480–84.

Kaback, H.R. and Wu, J. (1997) From membrane to molecule to the third amino acid from the left with a membrane transport protein. *Q. Rev. Biophys.*, **30**, 333–64.

Kachel, K., Asuncion-Punzalan, E., and London, E. (1995) Anchoring of tryptophan and tyrosine analogs at the hydrocarbon-polar boundary in model membrane vesicles: parallax analysis of fluorescence quenching induced by nitroxide-labeled phospholipids. *Biochemistry*, **34**(47), 15475–79.

Karim, C.B. *et al.* (2004) Phospholamban structural dynamics in lipid bilayers probed by a spin label rigidly coupled to the peptide backbone. *Proc. Natl. Acad. Sci. U.S.A.*, **101**(40), 14437–42.

Karp, E.S. *et al.* (2006) Electron paramagnetic resonance studies of an integral membrane peptide inserted into aligned phospholipid bilayer nanotube arrays. *J. Am. Chem. Soc.*, **128**(37), 12070–71.

Katsaras, J. (1997) Highly aligned lipid membrane systems in the physiologically relevant "excess water" condition. *Biophys. J.*, **73**(6), 2924–29.

Katsaras, J. (1998) Adsorbed to a rigid substrate, dimyristoylphosphatidylcholine multibilayers attain full hydration in all mesophases. *Biophys. J.*, **75**(5), 2157–62.

Kenyon, C.N. (1978) Complex lipids and fatty acids in photosynthetic bacteria, in *The Photosynthetic Bacteria* (eds R.K. Clayton and W.R. Sistrom) Plenum Press, New York, pp. 281–313.

Killian, J.A. and Heijne, G. (2000) How proteins adapt to a membrane–water interface. *Trends Biochem. Sci.*, **25**, 429–34.

Killian, J.A. and Nyholm, T.K. (2006) Peptides in lipid bilayers: the power of simple models. *Curr. Opin. Struct. Biol.*, **16**(4), 473–79.

Landolt-Marticorena, C. *et al.* (1993) Non-random distribution of amino acids in the transmembrane segments of human type I single span membrane proteins. *J. Mol. Biol.*, **229**(3), 602–8.

Langmuir, I. (1917) The shapes of group molecules forming the surfaces of liquids. *Proc. Natl. Acad. Sci. U.S.A.*, **3**(4), 251–57.

Läuger, P., *et al.* (1967) Electrical properties of bimolecular phospholipid membranes. *Biochim. Biophys. Acta*, **135**, 20–32.

Lee, A.G. (1991) Lipids and their effects on membrane proteins: Evidence against a role for fluidity. *Prog. Lipid Res.*, **30**, 323–48.

Lee, A.G. (2005) Cell biology: a greasy grip. *Nature*, **438**(7068), 569–70.

Lee, A.G. (2005) How lipids and proteins interact in a membrane: a molecular approach. *Mol. Biosyst.*, **1**(3), 203–12.

Lemmon, M.A. and Engelman, D.M. (1994) Specificity and promiscuity in membrane helix interactions. *Q. Rev. Biophys.*, **27**(2), 157–218.

Lemmon, M.A. *et al.* (1992) Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry*, **31**(51), 12719–25.

Lenard, J. and Singer, S.J. (1966) Protein conformation in cell membrane preparations as studied by optical rotatory dispersion and circular dichroism. *Proc. Natl. Acad. Sci. U.S.A.*, **56**(6), 1828–35.

Lewis, B.A. and Engelman, D.M. (1983) Lipid bilayer thickness varies linearly with acyl chain length in fluid phosphatidylcholine vesicles. *J. Mol. Biol.*, **166**, 211–17.

Li, X. and Schick, M. (2000) Theory of lipid polymorphism: application to phosphatidylethanolamine and phosphatidylserine. *Biophys. J.*, **78**(1), 34–46.

Li, C. *et al.* (2007) Uniformly aligned full-length membrane proteins in liquid crystalline bilayers for structural characterization. *J. Am. Chem. Soc.*, **129**(17), 5304–5.

Lide, D.R. (1996) *Handbook of chemistry and physics*, CRC, Boca Raton. Vol. 77.

Liu, W.M. (1998) Shear numbers of protein beta-barrels: definition refinements and statistics. *J. Mol. Biol.*, **275**(4), 541–45.

Liu, N.I. and Kay, R.L. (1977) Redetermination of the pressure dependence of the lipid bilayer phase transition. *Biochemistry*, **16**(15), 3484–86.

Lundbaek, J.A. (2006) Regulation of membrane protein function by lipid bilayer elasticity – a single molecule technology to measure the bilayer properties experienced by an embedded protein. *J. Phys. Condens. Matter*, **18**, S1305–S1344.

Lundbæk, J.A. and Andersen, O.S. (1994) Lysophospholipids modulate channel function by altering the mechanical properties of lipid bilayers. *J. Gen. Physiol.*, **104**, 645–73.

Lundbaek, J. and Andersen, O.S. (1999) Spring constants for channel-induced lipid bilayer deformations – estimates using gramicidin channels. *Biophys. J.*, **76**, 889–95.

Lundbaek, J.A. *et al.* (2003) Cholesterol-induced protein sorting: an analysis of energetic feasibility. *Biophys. J.*, **84**(3), 2080–89.

Luzzati, V. and Husson, F. (1962) The structure of the liquid-crystalline phases of lipid–water systems *J. Cell Biol.*, **12**, 207–19.

Marassi, F.M. and Opella, S.J. (2000) A solid-state NMR index of helical membrane protein structure and topology. *J. Magn. Reson.*, **144**(1), 150–55.

Marsh, D. and Horvath, L.I. (1998) Structure, dynamics and composition of the lipid–protein interface. Perspectives from spin-labelling. *Biochim. Biophys. Acta*, **1376**(3), 267–96.

Marsh, D. and Pali, T. (2004) The protein–lipid interface: perspectives from magnetic resonance and crystal structures. *Biochim. Biophys. Acta*, **1666**(1–2), 118–41.

Mayo, D.J. *et al.* (2008) Comparing the structural topology of integral and peripheral membrane proteins utilizing electron paramagnetic resonance spectroscopy. *J. Am. Chem. Soc.*, **130**(30), 9656–57.

McLachlan, A.D. (1979) Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.*, **128**(1), 49–79.

McNamara, J.R., Warnick, G.R., and Cooper, G.R. (2006) A brief history of lipid and lipoprotein measurements and their contribution to clinical chemistry. *Clin. Chim. Acta*, **369**(2), 158–67.

Chevreul, M.E. (1889) *Science*, **13**(323), 271.

Mingarro, I. *et al.* (1996) Ala-insertion scanning mutagenesis of the glycophorin A transmembrane helix: a rapid way to map helix–helix interactions in integral membrane proteins. *Protein Sci.*, **5**(7), 1339–41.

Montal, M. and Mueller, P. (1972) Formation of biomolecular membranes from lipid monolayers and a study of their electrical properties. *Proc. Natl. Acad. Sci. U.S.A.*, **69**, 3561–66.

Morein, S. *et al.* (1997) Influence of membrane-spanning alpha-helical peptides on the phase behavior of the dioleoylphosphatidylcholine/water system. *Biophys. J.*, **73**(6), 3078–88.

Morth, J.P. *et al.* (2007) Crystal structure of the sodium-potassium pump. *Nature*, **450**(7172), 1043–49.

Mouritsen, O.G. and Andersen, O.S. (1998) *In search of a new biomembrane model.* Biol. Skr. Dan. Vid. Selsk.. Vol. 49.: Munksgaard, Copenhagen.

Mouritsen, O.G. and Bloom, M. (1984) Mattress model of lipid–protein interactions in membranes. *Biophys. J.*, **46**, 141–53.

Nägeli, C. and Cramer C.E. (1855) *Pflanzenphysiologische Untersuchungen, 1–4 Heft, Zurich Schulthess.*

Nagle, J.F. and Tristram-Nagle, S. (2000) Structure of lipid bilayers. *Biochim. Biophys. Acta*, **1469**(3), 159–95.

Nernst, W. (1909) Zur Theorie des electrischen Reizes. *Pfluegers Arch. Physiol.*, **122**, 275–314.

Nielsen, C. and Andersen, O.S. (2000) Inclusion-induced bilayer deformations: effects of monolayer equilibrium curvature. *Biophys. J.*, **79**(5), 2583–604.

Nielsen, C., Goulian, M., and Andersen, O.S. (1998) Energetics of Inclusion-Induced Bilayer Deformations. *Biophys. J.*, **74**, 1966–83.

Nogi, T. *et al.* (2000) Crystal structures of photosynthetic reaction center and high-potential iron–sulfur protein from Thermochromatium tepidum: thermostability and electron transfer. *Proc. Natl. Acad. Sci. U.S.A.*, **97**(25), 13561–66.

O'Connell, A.M., Koeppe, R.E., and Andersen II, O.S. (1990) Kinetics of gramicidin channel formation in lipid bilayers: transmembrane monomer association. *Science*, **250**, 1256–59.

Oldfield, E., Bowers, J.L., and Forbes, J. (1987) High-resolution proton and carbon-13 NMR of membranes: why sonicate?. *Biochemistry*, **26**(22), 6919–23.

Ollivon, M. *et al.* (2000) Vesicle reconstitution from lipid-detergent mixed micelles. *Biochim. Biophys. Acta*, **1508**(1–2), 34–50.

Overton, E. (1899) Über die osmotischen Eigenschaften derZelle und ihre Bedeutung für die Toxikologie und Pharmakologie. *Vierteljahrsschr. Naturforsch. Ges. Zürich*, **44**, 88–114.

Ozdirekcan, S. *et al.* (2005) Influence of flanking residues on tilt and rotation angles of transmembrane peptides in lipid bilayers. A solid-state ^{2}H NMR study. *Biochemistry*, **44**(3), 1004–12.

Pascher, I. *et al.* (1992) Crystal structures of membrane lipids. *Biochim. Biophys. Acta*, **1113**, 339–73.

Paulick, M.G. and Bertozzi, C.R. (2008) The glycosylphosphatidylinositol anchor: a complex membrane-anchoring structure for proteins. *Biochemistry*, **47**(27), 6991–7000.

Pauling, L., Corey, R.B., and Branson, H.R. (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A*, **37**, 205–11.

Petersen, F.N., Jensen, M.O., and Nielsen, C.H. (2005) Interfacial tryptophan residues: a role for the cation-pi effect? *Biophys. J.*, **89**(6), 3985–96.

Rand, R.P. and Parsegian, V.A. (1989) Hydration forces between phospholipid bilayers. *Biochim. Biophys. Acta*, **988**, 351–76.

Rawicz, W. *et al.* (2000) Effect of chain length and unsaturation on elasticity of lipid bilayers. *Biophys. J.*, **79**(1), 328–39.

Reithmeier, R.A. (1995) Characterization and modeling of membrane proteins using sequence analysis. *Curr. Opin. Struct. Biol.*, **5**(4), 491–500.

Resh, M.D. (2006) Trafficking and signaling by fatty-acylated and prenylated proteins. *Nat. Chem. Biol.*, **2**(11), 584–90.

Reyes Mateo, C., Rodrigo, F.M., Loura, L.M.S., and Prieto, M. (2006) From lipid phases to membrane protein organization: fluorescence methodologies in the study of lipid–protein interactions, in *Protein–Lipid Interactions* (eds C. Reyes Mateo, J. Gómez, J. Villalain, and J.M. Gonzales-Ros) Springer, Berlin Heidelberg, pp. 1–33.

Ring, A. (1996) Deformation free energy of bilayer membrane and its effects on gramicidin channel lifetime. *Biochim. Biophys. Acta*, **1278**, 147–59.

Robertson, J.D. (1966) Granulo-fibrillar and globular sub-structure in unit membranes. *Ann. N.Y. Acad. Sci.*, **137**, 421–40.

Roseman, M.A. (1988) Hydrophobicity of the peptide C=O···H–N hydrogen-bonded group. *J. Mol. Biol.*, **201**, 621–23.

Sackmann, E. (1984) Physical basis of trigger processes and membrane structures, in *Biological Membranes* (ed D. Chapman) Academic Press Inc., Ltd., London, pp. 105–43.

Sakmar, T.P. (1998) Rhodopsin: a prototypical G protein-coupled receptor. *Prog. Nucleic Acid Res. Mol. Biol.*, **59**, 1–34.

Schiffer, M., Chang, C.-H., and Stevens, F.J. (1992) The functions of tryptophan residues in membrane proteins. *Protein Eng.*, **5**, 213–14.

Seddon, J.M. (1990) Structure of the inverted hexagonal (H^{II}) phase, and non-lamellar phase transitions of lipids. *Biochim. Biophys. Acta*, **1031**, 1–69.

Seddon, J.M. and Templer, R.H. (1993) Cubic phases of self-assembled amphiphilic aggregates. *Philos. Trans. R. Soc. London, Ser. A*, **344**, 377–401.

Seelig, J. (1977) Deuterium magnetic resonance: theory and application to lipid membranes. *Q. Rev. Biophys.*, **10**, 353–418.

Seelig, A. and Seelig, J. (1974) The dynamic structure of fatty acyl chains in a phospholipid bilayer measured by deuterium magnetic resonance. *Biochemistry*, **13**(23), 4839–45.

Sharp, K.A. *et al.* (1991) Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science*, **252**, 106–9.

Simons, K. and Ikonen, E. (1997) Functional rafts in cell membranes. *Nature*, **387**(6633), 569–72.

Singer, S.J. (1971) The molecular organization of membranes, in *The Structure and Function of Biological Membranes* (ed L.I. Rothfield), Academic Press, New York, pp. 145–222.

Singer, S.J. and Nicolson, G.L. (1972) The fluid mosaic model of the structure of cell membranes. *Science*, **175**, 720–31.

Soubias, O. *et al.* (2008) Rhodopsin adjust its helical content to match membrane hydrophobic thickness. *Biophys. J. Abstr.*, Suppl. **182a**, 894-plat.

Sperotto, M.M. and Mouritsen, O.G. (1988) Dependence of lipid membrane phase temperature on the mismatch of protein and lipid hydrophobic thickness. *Eur. Biophys. J.*, **16**, 1–10.

Strandberg, E. *et al.* (2002) Lipid dependence of membrane anchoring properties and snorkeling behavior of aromatic and charged residues in transmembrane peptides. *Biochemistry*, **41**(23), 7190–98.

Strandberg, E. *et al.* (2004) Tilt angles of transmembrane model peptides in oriented and non-oriented lipid bilayers as determined by ^{2}H solid-state NMR. *Biophys. J.*, **86**(6), 3709–21.

Straus, S.K., Scott, W.R., and Watts, A. (2003) Assessing the effects of time and spatial averaging in ^{15}N chemical shift/^{15}N –^{1}H dipolar correlation solid state NMR experiments. *J. Biomol. NMR*, **26**(4), 283–95.

Stubbs, C.D. and Slater, S.J. (1996) The effects of non-lamellar forming lipids on membrane protein–lipid interactions. *Chem. Phys. Lipids*, **81**, 185–95.

Toyoshima, C. *et al.* (2000) Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 A resolution. *Nature*, **405**(6787), 647–55.

Treutlein, H.R. *et al.* (1992) The glycophorin A transmembrane domain dimer: sequence-specific propensity for a right-handed supercoil of helices. *Biochemistry*, **31**(51), 12726–732.

Unwin, N. (1989) The structure of ion channels in membranes of excitable cells. *Neuron*, **3**, 665–76.

Unwin, P.N.T. and Ennis, P.D. (1984) Two configurations of a channel-forming membrane protein. *Nature*, **307**, 609–13.

von Heijne, G. (1994) Membrane proteins: from sequence to structure. *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 167–92.

Wallach, D.F. and Zahler, P.H. (1966) Protein conformations in cellular membranes. *Proc. Natl. Acad. Sci. U.S.A.*, **56**(5), 1552–59.

Wang, J. *et al.* (2000) Imaging membrane protein helical wheels. *J. Magn. Reson.*, **144**(1), 162–67.

White, S.H. and Thompson, T.E. (1973) Capacitance, area, and thickness variations in thin lipid films. *Biochim. Biophys. Acta*, **323**(1), 7–22.

Wiener, M.C. and White, S.H. (1991a) Fluid bilayer structure determination by the combined use of X-ray and neutron diffraction. I. Fluid bilayer models and the limits of resolution. *Biophys. J.*, **59**(1), 162–73.

Wiener, M.C. and White, S.H. (1991b) Fluid bilayer structure determination by the combined use of X-ray and neutron diffraction. II. "Composition-space" refinement method. *Biophys. J.*, **59**(1), 174–85.

Wiener, M.C. and White, S.H. (1992a) Structure of a fluid dioleoylphosphatidylcholine bilayer determined by joint refinement of X-ray and neutron diffraction data. III. Complete structure. *Biophys. J.*, **61**(2), 434–47.

Wiener, M.C. and White, S.H. (1992b) Structure of a fluid dioleoylphosphatidylcholine bilayer determined by joint refinement of X-ray and neutron diffraction data. II. Distribution and packing of terminal methyl groups. *Biophys. J.*, **61**(2), 428–33.

Wiener, M.C., King, G.I., and White, S.H. (1991) Structure of a fluid dioleoylphosphatidylcholine bilayer determined by joint refinement of X-ray and neutron diffraction data. I. Scaling of neutron data and the distributions of double bonds and water. *Biophys. J.*, **60**(3), 568–76.

Wimley, W.C. and White, S.H. (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Mol. Biol.*, **3**, 842–48.

Wu, C.H., Ramamoorthy, A. and Opella, S.J. (1994) High resolution heteronuclear dipolar solid-state NMR spectroscopy. *J. Magn. Reson. A*, **109**, 270–72.

Yau, W. *et al.* (1998) The preference of tryptophan for membrane interfaces. *Biochemistry*, **37**, 14713–18.

Yeagle, P.L. and Frye, J. (1987) Effects of unsaturation on ^{2}H-NMR quadrupole splittings and ^{13}C-NMR relaxation in phospholipid bilayers. *Biochim. Biophys. Acta*, **899**(2), 137–42.

Zimmerberg, J. and Gawrisch, K. (2006) The physical chemistry of biological membranes. *Nat. Chem. Biol.*, **2**(11) 564–67.

11
DNA Biomechanics

Wilma K. Olson, Annankoil R. Srinivasan, Andrew V. Colasanti, Guohui Zheng, and David Swigon

11.1 **Introduction 361**
11.2 **DNA Elastic Rod 361**
11.3 **Physical–Chemical Perspectives of DNA 363**
11.3.1 DNA Polyanion 363
11.3.2 DNA Fiber-Diffraction Models 365
11.3.3 Base Sequence and Helical Structure 365
11.3.4 Helical Fragments and Bending 366
11.4 **Structural Features of DNA 367**
11.4.1 High-Resolution Structures 367
11.4.2 DNA Base-Pair Parameters 367
11.4.3 Bending and Translational Anisotropy 369
11.4.4 Coupling of Parameters 369
11.4.5 Sequence-Dependent Twist 370
11.5 **Sequence-Dependent Potentials 372**
11.5.1 Statistical Perspective 372
11.5.2 Knowledge-Based Potentials 372
11.5.3 Pyrimidine–Purine Deformability 374
11.6 **Sequence-Dependent DNA Elasticity 374**
11.6.1 Naturally Discrete Treatment of Polymeric DNA 374
11.6.2 Intrinsic Motions of "Real" DNA Sequences 375
11.6.3 Sequence-Dependent Effects in Monte-Carlo Simulations of Short DNA 376
Acknowledgments 377
Glossary 377
References 378

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

11.1
Introduction

Encoded in the strings of DNA bases that make up the genomes of living species are codes that regulate, control, and describe all sorts of biological processes. The underpinnings of these codes lie in the base sequence-dependent micromechanical properties of DNA, which determine the degree to which the long, threadlike molecule fluctuates and how it responds to the proteins that control its processing and govern its packaging.

Insight into these properties of DNA comes from a variety of sources, ranging from classical physical measurements of polymeric behavior to new biophysical techniques that allow for the manipulation of single DNA molecules and new computational approaches that incorporate the sequence-dependent structural and energetic information encoded in high-resolution DNA structures. This chapter first introduces the reader to the classic representation of DNA as an elastic rod and the breakdown of this model in the interpretation of recent experimental data. Next follows an overview of the classical physical–chemical perspective of DNA mechanics in terms of its polyanionic character and fiber-diffraction properties and a summary of early ideas on the connection of base sequence to helical structure and DNA bending. The narrative then turns to the unique, sequence-dependent spatial information that has been gleaned from the many solved crystal structures of DNA and its complexes with other molecules. Following a brief discussion of the knowledge-based elastic functions that incorporate the observed sequence-dependent structural properties of DNA, the chapter concludes with examples of some of the new insights into DNA polymeric behavior gained with new computational approaches that incorporate the sequence-dependent features of neighboring base-pair steps.

11.2
DNA Elastic Rod

Although the biological literature tends to focus on the double-helical structure of short pieces of DNA, the classic physics literature dwells on the macroscopic properties of DNA and the thin, inextensible, naturally straight, homogeneous elastic-rod model (Kirchhoff, 1859, 1876) used to interpret the features of

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

very long molecules. The elastic energy of the rod is taken as the integral of a bending energy that depends on the curvature κ of the smoothly deforming structure and a twisting contribution that depends on the difference, $\Delta\omega$, between the twist of residues in the stress-free state and the current state,

$$E = \frac{A}{2}\int_0^L \kappa(s)^2 ds + \frac{C}{2}\int_0^L \Delta\omega(s)^2 ds \tag{1}$$

Here s is the arc length, L the contour length, A the bending modulus, and C the torsional modulus of the rod.

The bending constant A is directly proportional to the persistence length a, a measure of DNA extension found from both classic (Hagerman, 1988; Crothers *et al.*, 1992; Bednar *et al.*, 1995) and single-molecule (Smith, Finzi and Bustamante, 1992; Baumann *et al.*, 1997) experiments to be in the range of 500–600 Å for mixed sequence DNA under physiological salt conditions, that is $A = ak_BT$, where k_B is the Boltzmann constant and T the temperature in Kelvin. The magnitude of the torsional modulus, however, is a matter of debate. Values of C obtained with recent rotor-bead tracking measurements (Bryant *et al.*, 2003) ($4.1 \pm 0.3 \times 10^{-19}$ erg cm) substantially exceed the values deduced from many solution studies (Barkley and Zimm, 1979; Thomas *et al.*, 1980; Hurley *et al.*, 1982; Millar, Robbins and Zewali, 1982; Horowitz and Wang, 1984; Heath *et al.*, 1996), but roughly match the values extracted from the observed populations (Shore and Baldwin, 1983) of short, covalently closed DNA duplexes with different linking numbers. Proponents of the solution data suggest that bending strain restricts the twisting of small, circular molecules (Heath *et al.*, 1996) and also note the closer correspondence of the observed ratio of the bending and twisting moduli, A/C, deduced from the solution data with the predicted value of 1.5 for an incompressible rod. That is, under conditions where the persistence length is 500 Å and the bending constant A is 2×10^{-19} erg-cm, the ratio A/C is roughly unity when the torsional modulus is based on the solution data, but ~0.5 when the higher estimates of C are used. Moreover, the negative Poisson ratio σ deduced from the latter values, $\sigma = A/C - 1$, is indicative of a thickening of the rod upon extension, a behavior unknown in homogeneous, isotropic elastic materials and also inconsistent with the structural properties of double-helical DNA (Kosikov *et al.*, 1999).

Measurements of the response of individual DNA molecules to imposed forces have stimulated extension of the elastic theory to account for the extensibility of DNA in terms of a stretching modulus S that measures the resistance of the polymer to longitudinal deformation (Odijk, 1995; Marko and Siggia, 1995). Application of the theory to these experiments yields S values of ~1000 pN when the persistence length is 500 Å. The derived variation of S with changes in ionic strength, however, does not match theoretical expectations. Although the theory predicts S to be directly proportional to the persistence length, the apparent stretching modulus decreases under conditions that increase a and vice versa (Baumann *et al.*, 2000).

Other indication of the breakdown of the DNA elastic-rod model comes from the unexpected, spontaneous cyclization of DNA molecules less than 100 bp in length (Cloutier and Widom, 2004, 2005). The observed levels of closed circular molecules, as measured by the Jacobson–Stockmayer *J* factor, the well-known ratio of the equilibrium constants for cyclization versus bimolecular association of a linear molecule (Jacobson and Stockmayer, 1950), are up to five orders of magnitude greater than those expected for a short DNA chain modeled as an ideal elastic rod (Shimada and Yamakawa, 1984). Both the apparent resistance of DNA to bending and the contributions of twist that bring the ends of the duplex in register for ring closure are much smaller than the elastic moduli cited above. Importantly, the sequences found to close most easily into tight minicircles contain a well-known nucleosome-positioning sequence (Lowary and Widom, 1998), that is a sequence that binds more tightly than others on the surface of the nucleosome core particle, with periodically repeating chemical features. This sequence includes two notable patterns: (i) TA base-pair steps, which recur at increments of ~10 bp, that is roughly every complete double-helical turn, and (ii) AT-containing dimers (AA, TA, or TT), which alternate at half-helical turns with GC base-pair steps. The classic theory takes no account of the special features of these steps that may give rise to the aberrant ring-closure properties.

11.3
Physical–Chemical Perspectives of DNA

11.3.1
DNA Polyanion

Of course, there is no reason other than mathematical convenience to assume that double-stranded DNA, with complementary antiparallel strands wrapped around a core of Watson–Crick base pairs (Figure 11.1), should behave like an inextensible elastic rod. The sugar–phosphate linkage acts like a spring in the sense that it constrains the movements of neighboring base pairs, with the

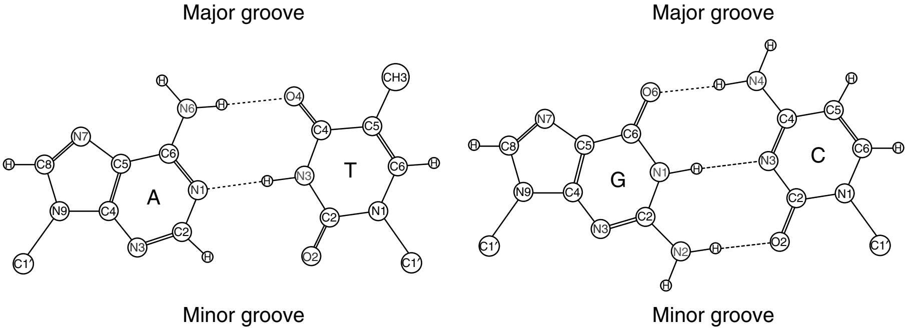

Fig. 11.1 Comparison of hydrogen-bonding interactions and chemical structures (including double bonds) of the purine and pyrimidine bases comprising normal A.T and G.C Watson–Crick pairs. Hydrogen bonds designated by dashed lines. Proton donor and acceptor atoms denoted in red and blue, respectively. Note the pseudosymmetric positioning of the purine N3 and pyrimidine O2 atoms on the minor-groove edges of the base pairs and the unique sequence-dependent signatures (Seeman, Rosenberg and Rich, 1976) of the unfulfilled donor and acceptor atoms on the major-groove edges of A.T versus G.C. (Please find a color version of this figure on the color plates.)

repulsive interactions between the negatively charged phosphates helping to maintain the straight helical structure (Manning, 2006). Neutralization of the charge, such as by deposition of DNA on positively coated surfaces (Podestà *et al.*, 2005), has a dramatic effect on polymer extension, decreasing the persistence length by fivefold. Asymmetric reduction of charge, by site-specific neutralization of phosphates on one face of the DNA (Strauss and Maher, 1994), similarly bends the duplex toward the chemically modified sites. The nonuniform build-up of counterions in the grooves of DNA (Williams and Maher, 2000) presumably bends the double helix along the same lines, with the preferential localization of positively charged species in the major groove of G.C-rich sequences (Denisov and Halle, 2000) and the minor groove of A.T-rich sequences (Hud and Feigon, 1997) governing the intrinsic structure and deformability (see Figures 11.1 and 11.2 and discussion below of the DNA grooves).

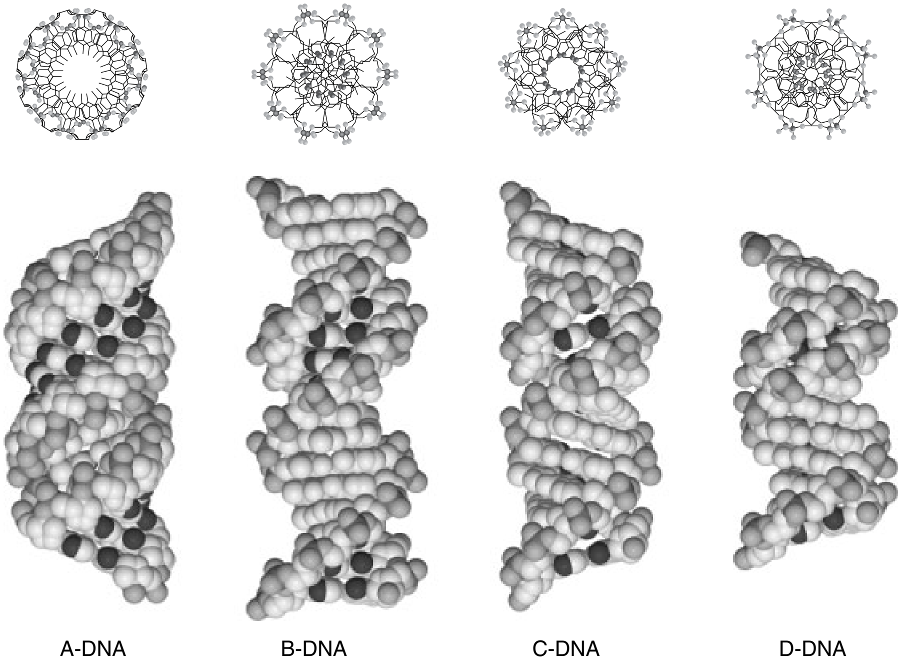

Fig. 11.2 Top and side views illustrating the global features of selected right-handed double-helical models deduced from X-ray fiber-diffraction measurements: A-DNA (Franklin and Gosling, 1953), 11 bp/helical turn; B-DNA (Watson and Crick, 1953), 10 bp/helical turn; C-DNA (Marvin *et al.*, 1958), 9 bp/helical turn; D-DNA (Davies and Baldwin, 1963), 8 bp/helical turn. The dark balls, located on the purine N3 and pyrimidine O2 atoms, reveal the relative exposure of the minor-groove edges of the Watson–Crick base pairs in the different helical forms. The lighter balls, located on the phosphate groups, help to visualize the accompanying changes in groove widths. DNA is represented by stick models and smaller balls on the aforementioned atoms in the top views, and by regular space-filling models in the side views. Note, in addition, the differences in base-pair inclination, width of the central "channel," and overall DNA extension. Images rendered with Raster3D (Merritt and Bacon, 1997).

11.3.2
DNA Fiber-Diffraction Models

The classic models of stretched DNA fibers provide hints of the ways in which the double-helical structure responds to changes in chemical environment. For example, variation in the ion composition or water content of DNA fibers introduces a large-scale conformational rearrangement of the classic right-handed B-DNA double helix with 10 bp per turn (Watson and Crick, 1953) to under- and over-wound A-DNA, C-DNA, and D-DNA structures of the same helical sense with 11, 9, and 8 bp per turn, respectively (Franklin and Gosling, 1953; Marvin *et al.*, 1958; Davies and Baldwin, 1963) (Figure 11.2). These changes in helical repeat are coupled to changes in the inclination and displacement of base pairs, which are linked, in turn, to the helical pitch and groove structure. Although the planes of the Watson–Crick base pairs stack roughly perpendicular to the global helical axis of B-DNA, they deviate by 10–20° from this alignment in the A, C, and D forms. Inclination of this magnitude, when coupled to the movement of the base pairs toward the outside of the helix, reduces the pitch, that is the projected displacement of successive base pairs along the global helical axis, and brings different atoms to the surface of the helix. Although the edge of the Watson–Crick base pair containing the pyrimidine O2 and the purine N3 atoms, also called the *minor-groove edge,* is exposed, the edge on the opposite side of the base pair, called the *major-groove edge,* is hidden in A-DNA compared to B-DNA (see Figures 11.1 and 11.2). By contrast, the minor groove is less accessible and the major groove is more accessible in C- and D-DNA compared to B-DNA. In other words, the B → A transition displaces and reorients base pairs in the opposite sense from that seen in the B → C and B → D transitions (Calladine and Drew, 1984; van Dam and Levitt, 2000; Vargason, Henderson and Ho, 2001; Lu and Olson, 2003). The displacement of base pairs also opens a channel through the center of the duplex during the course of these transitions. The atoms that line the inner passages provide potential binding sites for water, ions, and other small molecules from the surrounding environment.

11.3.3
Base Sequence and Helical Structure

The models of DNA deduced from the diffraction patterns of regularly repeating polynucleotide chains offered the first suggestions of a connection between base-pair sequence and three-dimensional double-helical structure (Chandrasekaran and Arnott, 1989). For example, the synthetic poly(dG)·poly(dC) homopolymeric complex is “A-philic” (Ivanov *et al.*, 1974) in the sense that it adopts an A-like structure under conditions where the poly(dA)·poly(dT) duplex is B-like (Arnott and Selsing, 1974a, 1974b). Moreover, poly(dA)·poly(dT) differs from canonical B-DNA in taking up a slightly modified B′ helical structure with highly distorted base pairs (Arnott *et al.*, 1983; Alexeev, Lipanov and Skuratovskii, 1987a, 1987b). Other sequences favor the C and D forms. For example, the repetition of A.T or I.C base pairs seems to accommodate the deep, tightly closed minor groove typical of D-DNA (Chandrasekaran and Arnott, 1989; Arnott, 1999; Premilat and Albiser, 1986). The purine rings of these two base pairs, adenine and inosine (also called *hypoxanthine*), do not contain the bulky amino group attached to the minor-groove edge of guanine (see Figure 11.1) and thus fit more easily in the close confines of a narrowed minor groove.

11.3.4
Helical Fragments and Bending

The alternation of different helical forms along the DNA sequence, such as the insertion of an A-type stretch of GG·CC steps within a B-like duplex, not only brings selected base atoms to the surface for easy recognition by other molecules but also introduces an appreciable bend in the molecule (Selsing *et al.*, 1979) without disruption of base stacking, that is the characteristic van der Waals' association of neighboring base-pair planes one above the other. Bending of this sort is maximized when the deformed segment spans roughly a half-helical turn and forms a distinct wedge (Figure 11.3). The abutment of the inclined base pairs of a 5-bp fragment of A-DNA or C-DNA next to the base pairs of a B-type model introduces a bend of more than 20°, that is a change of this magnitude in the direction of the DNA axis. The bending would be even more pronounced if ~5-bp segments of A- and C- or D-like structures were repeated in tandem, forming bends of ~45° per 10-bp fragments and superhelical structures with ~80 bp per turn (Figure 11.3). Models of this type were initially used to account for the "action-at-a-distance" effects,

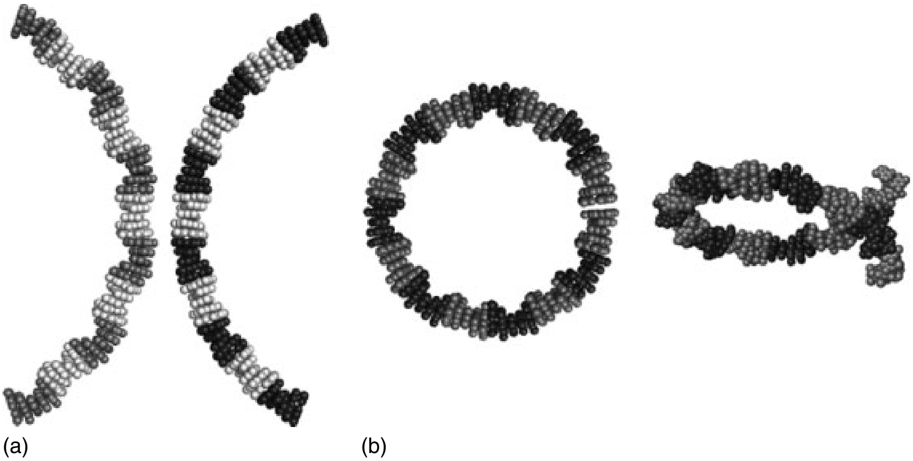

Fig. 11.3 Schematic illustration of curved DNA structures generated by the repeated abutment of 5-bp stretches of A-, B-, and C-like helices: (a) views down the global axes of semicircular structures constructed of (left) A/B and (right) B/C wedges; (b) views down and perpendicular to the axis of the tight superhelical pathway, with nearly 80 bp/superhelical turn, constructed of A/C wedges. The alignment of the structures in (a), along the central B-DNA stretch of each fragment, emphasizes their similar 150-bp circular curvature and opposing directions of global bending. Structures generated with the 3DNA software package, with A-, B-, and C-like base pair steps assigned (roll and slide) values of (7.4°, −1.2 Å), (0°, 0 Å), and (−7.4°, 1.2 Å), respectively. Tilt, twist, shift, and rise are fixed at 0°, 36°, 0 Å, and 3.4 Å throughout. Images prepared using PyMOL (DeLano, 2002).

for example melting enhancement and binding affinities, observed spectroscopically (Burd *et al.*, 1975) in abutting helical segments of synthetic $C_nA_n \cdot T_nG_n$ block copolymers and subsequently confirmed by the coexistence of A and B helical forms in $d(G_nT_n) \cdot d(C_nA_n)$ oligodeoxynucleotides in solution (Early *et al.*, 1977; Wang *et al.*, 1987, 1989) and protein-bound DNA crystals (Lu, Shakked and Olson, 2000). The GC half of the latter molecules exhibits A-like spectra, whereas the AT fraction retains B-like properties in concentrated salt solution. The coexistence of A- and B-type helices is also relevant to the structure of hybrid DNA–RNA complexes in which the double-helical DNA segments adopt B-form helices and the complementary RNA and DNA residues associate as A-type structures (Saenger, 1984). The superhelical arrangement of abutting stretches of A- and C-DNA, shown here for the first time, resembles the pathway taken by DNA around the core of histone proteins in the nucleosome core particle (Luger *et al.*, 1997)

11.4
Structural Features of DNA

11.4.1
High-Resolution Structures

The variation of DNA helical form and global bending found in early fiber models is being revealed today at atomic resolution in an ever-growing number of high-resolution crystal structures of “pure” DNA oligomers and DNA complexes with proteins and/or drugs. As anticipated, these structures are base-sequence dependent, with individual residues assuming characteristic spatial forms. This section summarizes the simple picture of base-sequence-dependent DNA structure and deformability that has emerged from these data, with emphasis on the subtle changes in local conformation and flexibility associated with the sequential arrangement of the four common bases in the observed right-handed helices. The more dramatic conformational changes that give rise to the left-handed Z-DNA helical structure (Wang *et al.*, 1979; Gessner *et al.*, 1989) are not addressed here. Z-DNA belongs to a totally different class of double-helical structures, which is widely separated from A-, B-, C-, and D-DNA by conformational barriers substantially higher than those that partition the major right-handed helical forms (Harvey, 1983; Olson *et al.*, 1983). Indeed, many of the solution properties attributed to Z-DNA can be accounted for by left-handed forms of the A–D helices that differ only subtly, in terms of the arrangement of the sugar–phosphate backbone and glycosyl linkage between the sugar and base, from conventional right-handed DNA; see Olson (1982) for further details.

11.4.2
DNA Base-Pair Parameters

Among the most interesting aspects of the known high-resolution DNA structures are the subtle irregularities of the double helix dictated by the locations of purine and pyrimidine residues along the chain backbone. In order to examine these perturbations, Dickerson and Drew (1981) reintroduced several parameters originated by Arnott, Dover, and Wonacott (1969) to define the

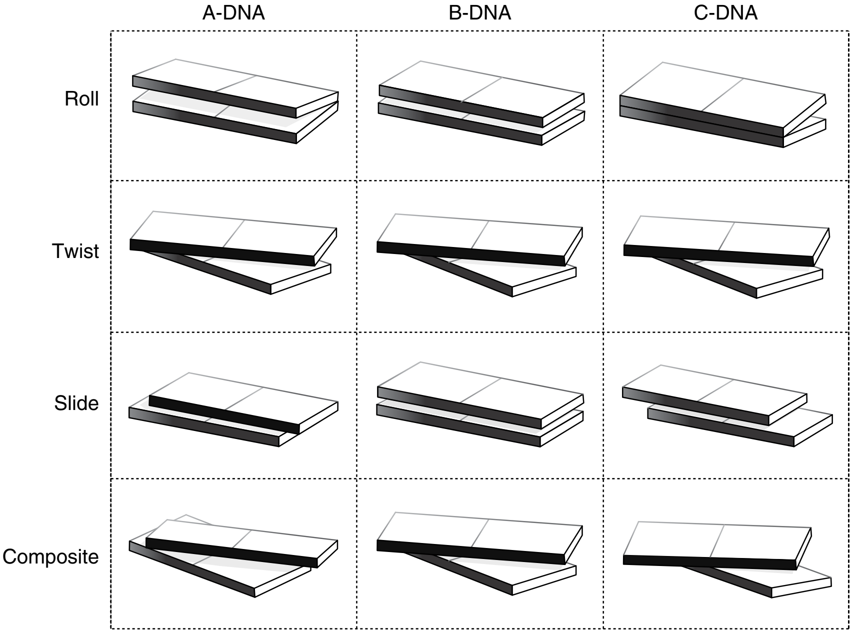

Fig. 11.4 Pictorial definitions of the three key rigid-body parameters – twist, roll, and slide – describing the spatial arrangements of sequential base pairs in DNA structures in A-, B-, and C-like steps. Base and base-pair ''bricks'' drawn such that the minor-groove edges are lightly shaded. Images generated with 3DNA (Lu and Olson, 2008), rendered with Raster3D (Merritt and Bacon, 1997), and plotted with PyMOL (DeLano, 2002).

orientation and displacement of adjacent base pairs in DNA fiber-diffraction models. The most important of these parameters for understanding the global helical transformations and sequence-dependent properties of DNA – twist, roll, and slide (Dickerson *et al.*, 1989) – are illustrated schematically in Figure 11.4. Twist is the rotation of successive base pairs about the normal to their planes, roll the rotation of the residues about the long axis of the dimer step, and slide the displacement of the base-pair centers parallel to the long axis. Variation in roll has a direct effect on the inclination of the DNA base pairs and the width of the major and minor grooves, while changes in slide alter the pitch and the depth of the grooves (Lu and Olson, 2003). Complete specification of DNA structure, however, requires knowledge, at each base-pair step, of (i) the six rigid-body parameters (buckle, propeller, opening, shear, stretch, and stagger) that describe the deviations of Watson–Crick base-pair geometry from the ideal planar form (Dickerson *et al.*, 1989), (ii) the three base-pair-step parameters (tilt, shift, and rise) that play a secondary role in most DNA structures (Dickerson *et al.*, 1989), and (iii) the set of torsion angles that describe the intervening sugar–phosphate linkages (IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN), 1983). Tilt is the component of bending about the short axis of each dimer (an accordion-like deformation that concomitantly compresses and stretches the strands of an intact duplex), shift the displacement of successive base pairs along their short axis, and rise the vertical, that is van der Waals', spacing between the mean planes of the base pairs. Specification of the base positions places limits on the states of the backbone (Srinivasan and Olson, 1987) (also see below).

The values of the step parameters reported herein are based on the 3DNA suite of programs (Lu and Olson, 2003, 2008). The values are defined by a matrix-based scheme (Zhurkin, Lysov and Ivanov, 1979; Bolshoy *et al.*, 1991; El Hassan and Calladine, 1995) that is rigorous and reversible in that it allows for the exact reconstruction of a three-dimensional structure from a complete set of derived parameters. The right-handed coordinate frame attached to each base follows established guidelines (Olson *et al.*, 2001), with (i) the x-axis pointing in the direction of the major groove along what would be the pseudodyad axis of an ideal Watson–Crick base pair, that is the perpendicular bisector of the $C1' \cdots C1'$ vector spanning the base pair, (ii) the y-axis running along the long axis of the idealized base pair in the direction of the sequence strand, parallel to the $C1' \cdots C1'$ vector, and displaced so as to pass through the intersection on the (pseudodyad) x-axis of the vector connecting the pyrimidine Y(C6) and purine R(C8) atoms, and (iii) the z-axis is defined by the right-handed rule.

11.4.3
Bending and Translational Anisotropy

As implied above, twist, roll, and slide are more susceptible to variation than the three other base-pair-step parameters. The bending and translational anisotropy is evident in scatter plots of the bending angles and shear displacements found in 6606 dimer steps from a nonredundant set of 239 known protein–DNA crystal structures of 2.5 Å or better resolution taken from the Nucleic Acid Database (Berman *et al.*, 1992) (Figure 11.5). The range of observed roll values is many times that of tilt as evident from the relative lengths of the major and minor axes of the ellipses fitted to the data. The local bending flexibility suggested by the scatter of data, however, is somewhat deceptive since individual dimer sequences cluster in sequence-specific sectors of the roll–tilt ellipse (Olson *et al.*, 2008). The shear deformations plotted in Figure 11.5b similarly show greater variation of slide than shift, that is, preferential displacement along the long axes of base-pair steps, albeit with less striking directional differences than the angular deformations. The slide and shift values also group in sequence-specific zones.

11.4.4
Coupling of Parameters

Significantly, variations in roll and slide accompany changes in twist, with an increase in roll and a decrease in slide generally attended by a decrease in twist and vice versa. This conformational interdependence, which also occurs in the X-ray structures of "pure" DNA oligomers (Olson *et al.*, 1998), reflects the chiral nature of the dinucleotide linkage. That is, the constraints of chemical structure limit the range of sterically acceptable conformations and the routes

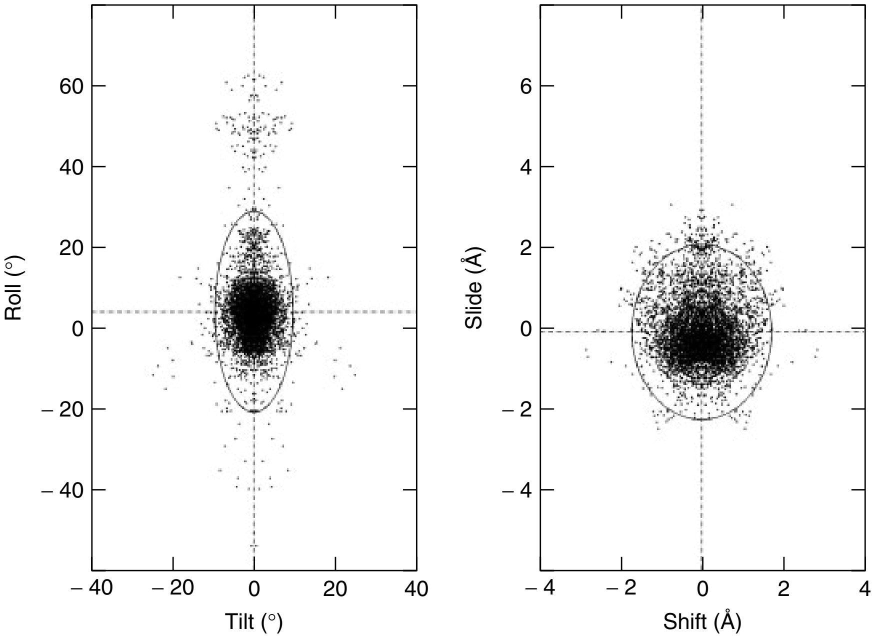

Fig. 11.5 Collective scatter plots of the tilt and roll bending angles (left) and the shift and slide shear displacements (right) in 6606 dimer steps from 239 protein–DNA crystal structures taken from the Nucleic Acid Database (Berman *et al.*, 1992).

by which they can be achieved. For example, the positive roll and negative slide values characteristic of A-DNA structures occur in concert with changes in the puckering of the sugar ring and the relative orientation of the sugar with respect to the base (Schneider, Neidle and Berman, 1997). The occurrence of high twist is often coupled to a conformational change of the phosphodiester backbone (Dickerson *et al.*, 1987), implicating the chemical pathway of duplex overwinding to a so-called $B_I \rightarrow B_{II}$ interconversion of the torsion angles about the C3′–O3′ and O3′–P bonds. Such B_{II} steps repeat at ~10-bp intervals in the nucleosome core particle structure (Richmond and Davey, 2003). Overtwisted states like these, with negative roll and positive slide, are identified with the C form of DNA (van Dam and Levitt, 2000).

11.4.5
Sequence-Dependent Twist

The sequence-dependent twisting observed in B-DNA and protein–DNA crystal structures (Gorin, Zhurkin and Olson, 1995; Olson *et al.*, 1998) mirrors simple geometrical features of the constituent base pairs; dimers with exocyclic groups involved in severe steric contacts in the major and minor grooves exhibit a tendency to unwind (Gorin, Zhurkin and Olson, 1995). These trends match the structural variability of dimer steps predicted over 25 years ago by Kabsch, Sander, and Trifonov (1982) from a combination of solution and gel

Tab. 11.1 Comparison of the sequence-dependent twist angles observed in high-resolution crystal structures with early predictions based on the gel and solution properties of DNA.[a]

Base-pair step	B-DNA[b]	P + DNA[c]	Prediction
	Purine–purine steps		
AG·CT	$28.5_{(5.3)}$	$32.2_{(5.1)}$	27.7
GG·CC	$35.3_{(5.4)}$	$33.2_{(4.4)}$	33.7
AA·TT	$36.1_{(3.1)}$	$32.6_{(6.8)}$	35.6
GA·TC	$40.4_{(2.2)}$	$35.8_{(4.7)}$	36.9
	Purine–pyrimidine steps		
AT·AT	$32.1_{(2.5)}$	$28.8_{(5.9)}$	31.5
AC·GT	$33.8_{(4.1)}$	$31.2_{(3.8)}$	34.4
GC·GC	$40.7_{(2.8)}$	$33.7_{(5.0)}$	40.0
	Pyrimidine–purine steps		
CG·CG	$31.7_{(5.7)}$	$33.5_{(4.6)}$	29.8
TA·TA	$41.9_{(6.7)}$	$35.6_{(8.9)}$	36.0
CA·TG	$42.4_{(8.7)}$	$36.1_{(6.4)}$	34.5

[a] Twist values, in degrees, computed with 3DNA (Lu and Olson, 2003), for "internal" dimer steps extracted from 145 B-DNA and 239 nonredundant protein–DNA (P + DNA) structures of 2.5 Å or better resolution from the Nucleic Acid Database (Berman *et al.*, 1992). Standard deviations are noted by the subscripted values in parentheses immediately following the corresponding averages. Values are identical for complementary steps, that is Tw(AA) = Tw(TT), and so on. Predicted data taken from Kabsch, Sander, and Trifonov (1982).

[b] B-DNA values based on 51 AG, 29 GG, 113 AA, 72 GA, 124 AT, 38 AC, 208 GC, 234 CG, 62 TA, 68 CA steps.

[c] P + DNA values based on 364 AG, 317 GG, 421 AA, 338 GA, 498 AT, 418 AC, 344 GC, 326 CG, 574 TA, 341 CA steps.

measurements (Table 11.1). Although some dimers, such as GG·CC steps, which are known to be "A-philic," are prone to undertwisting, other base-pair steps show a tendency to overtwist. Specifically, TA·TA and CA·TG dimer steps adopt very high values of Twist, coupled to a decrease in roll and an increase in slide, in many B-DNA and protein–DNA crystal structures. These preferences mimic the overwinding characteristic of synthetic polynucleotides, such as the sodium salt of the alternating poly(dAT)·poly(dAT) fiber, which forms an eightfold D-like helix rather than the A-type structures adopted under the same low-humidity conditions by mixed sequence DNA and its poly(dGC)·poly(dGC) counterpart (Davies and Baldwin, 1963; Arnott *et al.*, 1974; Leslie *et al.*, 1980). On average, the binding of proteins tends to untwist DNA (note the lower average twist values for B-DNA compared to P + DNA entries in Table 11.1).

Although the occurrences of high twist in B-DNA crystal structures have been attributed to intermolecular packing forces in the different crystal forms in which they occur and/or to the sequence context in which the dimer steps are found (Larsen, Kopka and Dickerson, 1991; Subirana and Faria, 1997), the persistence of the same extreme YR steps in protein-bound DNA structures (Suzuki and Yagi, 1995; Dickerson, 1998; Tolstorukov *et al.*, 2007) indicates that these dimers are also capable of locally overwinding the double helix and narrowing the minor groove in response to specific outside forces. The fact that DNA–DNA (Larsen, Kopka and Dickerson, 1991; Subirana and Faria, 1997) or DNA–protein contacts (Li and Olson, unpublished data) usually accompany such deformations suggests that the overtwisted C-like state is of high energy compared to standard B-DNA.

11.5
Sequence-Dependent Potentials

11.5.1
Statistical Perspective

The observed spread of DNA base-pair step parameters in high-resolution crystal structures provides one of the best available estimates of the natural, sequence-dependent deformability of the double-helical molecule (Olson *et al.*, 1998). The parameters of the 10 unique dimer steps cluster in distinctive, quasi-normal distributions consistent with harmonic behavior (Figure 11.6). The average values of the step parameters are suggestive of the rest state preferred by each dimer, and the pairwise covariance of variables, that is the differences between the mean squares and the squares of the means of all pairs of step parameters, is a measure of the intrinsic stiffness. This statistical perspective yields the smooth energy landscape that best connects the many distinct crystallographic examples and offers information about the likely overall motions of neighboring base pairs. By contrast, the thermal ellipsoids that define the range of individual atomic motions in some crystal structures (Willis and Pryor, 1975) characterize the more restricted motions of DNA within the various low-energy substates on the real "dimpled" potential landscape (Matsumoto and Olson, 2002).

The many diverse configurations of DNA seen in its complexes with different proteins constitute one of the best datasets for this sort of analysis. Since the various proteins impose different types of forces on DNA, the resulting distortions of double-helical structure are expected to cancel one another (major vs. minor groove bending, etc.), and the natural conformational response of DNA is expected to surface after averaging. By contrast, the number of "pure" DNA crystal structures, particularly those of B-DNA, is quite small and highly biased in terms of the types of available sequences and lattices in which they have been crystallized. Such constraints severely restrict the range of DNA conformational states compared to that seen in the presence of proteins.

11.5.2
Knowledge-Based Potentials

Although the equipotential contours in Figure 11.5 reflect the overall structural features of protein-bound DNA without regard to sequence, the contours superimposed on the roll–slide scatter plots in Figure 11.6 depict the distinctive

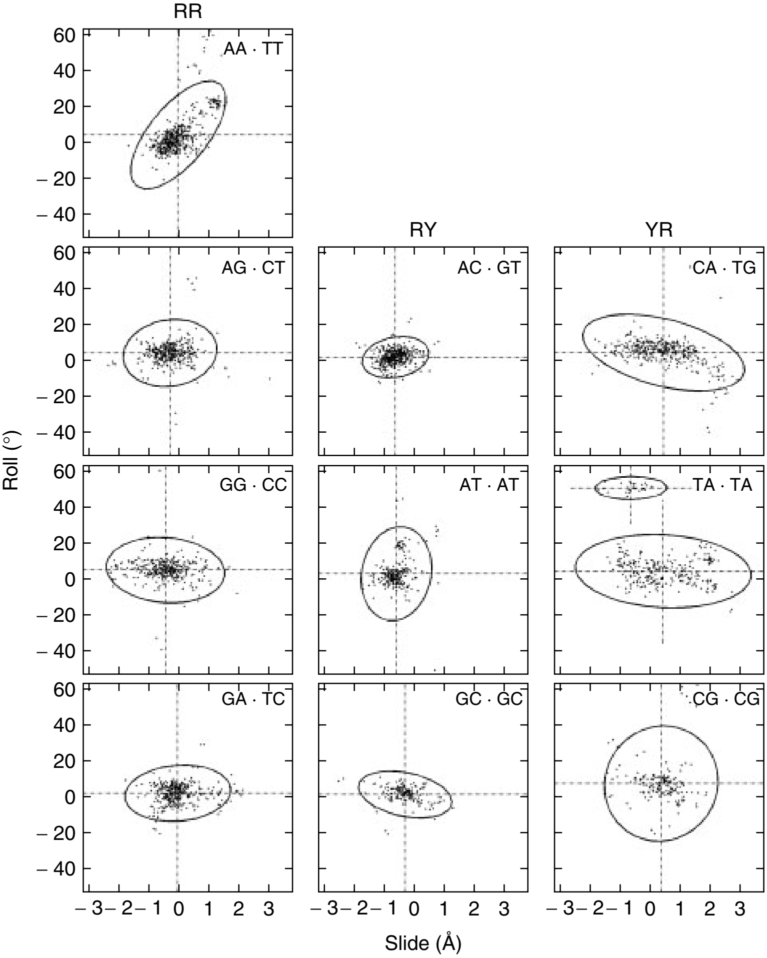

Fig. 11.6 Collective scatter plots in the roll–slide plane of base-pair step parameters found in high-resolution protein–DNA crystal complexes and derived sequence-dependent potentials of the 10 unique dimer steps. Dots correspond to the points used to derive the potentials (solid contours). Elliptical contours are projections of the multidimensional potentials on the roll–slide plane obtained from the 2 × 2 covariance matrix of observed roll–slide values. Contours correspond to deviations of parameters equal to two times the combined root-mean-square deviations of parameters. Average values of roll and slide are highlighted by dashed lines. The three columns show the respective deformational profiles of individual purine–purine (RR), purine–pyrimidine (RY), and pyrimidine–purine (YR) steps. Note the bimodal distribution of observed TA states and derived potentials in the two regimes.

deformational behavior of the 10 bp steps. As noted above, the direction of most probable change (along the long axes of the ellipses) often involves some combination of roll and slide. The extent and direction of parametric coupling depend upon sequence. For example, CA·TG dimers show a propensity to combine the negative roll that bends DNA into the minor groove with an increase in slide, whereas the AC·TG dimers tend to shear in the opposite sense with a decrease in roll.

11.5.3
Pyrimidine–Purine Deformability

The pyrimidine–purine (YR) steps stand out as the most easily deformed dimers. The intrinsic variability of the YR dimers is presumably related to their relatively weak stacking interactions (Ornstein and Fresco, 1983). Remarkably, the sharp bends in DNA observed in protein–DNA crystal complexes occur almost exclusively at YR steps (Suzuki and Yagi, 1995; Dickerson, 1998; Tolstorukov *et al.*, 2007). Moreover, the complementary base pairs remain intact in these steps. Note, in particular, the distinct bimodal character of bending (via roll) and shearing (via slide) of TA dimers. The YR steps apparently act as flexible hinges fitting the duplex to the protein surface. By contrast, the RY dimer steps are resistant to large-scale conformational distortions, serving as an indirect hindrance to DNA conformational change.

11.6
Sequence-Dependent DNA Elasticity

11.6.1
Naturally Discrete Treatment of Polymeric DNA

A coarse-grained model of DNA that makes use of the sequence-dependent dimeric potentials described herein brings new understanding of the properties of long polymers not possible with conventional elastic treatments. Incorporation of this type of model in a new theory of sequence-dependent DNA elasticity (Coleman, Olson and Swigon *et al.*, 2003; Olson, Swigon and Coleman, 2004) illustrates how the properties of "real" DNA can be misinterpreted in the context of an approximate model. For example, minicircles composed of naturally straight dimers, some of which are subject to roll–twist coupling, mimic the global response of an ideal elastic rod to imposed twisting, showing a sudden jump in the equilibrium configuration from a circle to figure of eight at a critical value of the imposed linking number. The elastic energies of such chains are smaller than the corresponding values for the ideal rod and, in some cases, show departures from ideal quadratic behavior that can be interpreted as changes in the effective elastic moduli. For example, if the coupled steps are distributed along the DNA in phase with the helical repeat, the elastic energy is nearly independent of imposed twisting, suggestive of a near-zero effective twist modulus.

Application of the sequence-dependent treatment of DNA to the loops formed by the binding of the *Escherichia coli* Lac repressor (LacR) assembly allows one to relate the fine structure of the LacR–DNA complex to a wide range of heretofore-unconnected experimental observations (Swigon, Coleman and Olson, 2006; Swigon and Olson, 2008). Examination of the 92-bp wild-type O3–O1 loop (Figure 11.7) finds antiparallel orientations of the DNA sequence on the V-shaped LacR assembly (labeled A1, A2 in the figure) to be more favorable than parallel arrangements (P1,P2). The

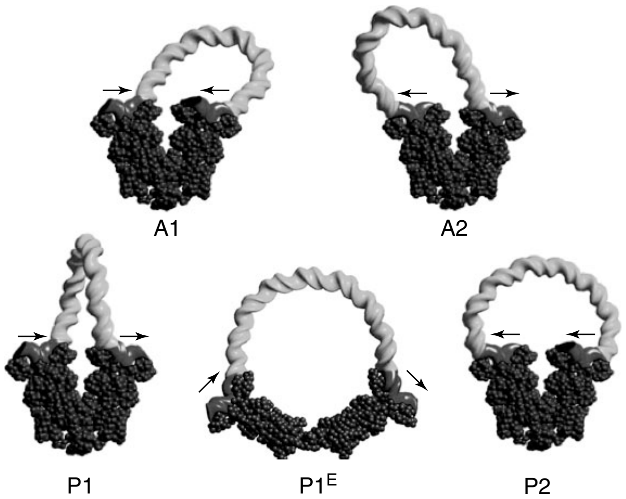

Fig. 11.7 Minimum-energy configurations of the 92-bp DNA loops formed upon binding of the Lac repressor (LacR) protein assembly to the so-called O3 and O1 operator sites on the *Escherichia coli lac* operon. DNA shown in aqua, LacR in red, and operator sites in blue. Arrows above operator sites (O3 left, O1 right) denote the orientations of DNA – antiparallel (A1, A2) and parallel (P1, P1^E, P2) – on the protein assembly. (Please find a color version of this figure on the color plates.)

free energy of one of the parallel loops (P1) is lowered substantially when the V-shaped structure is allowed to open ~80° (P1^E), falling close to or below the values computed for the antiparallel loops, depending upon the estimated cost of protein deformation. One of the antiparallel loops (A1) has the DNA fine structure consistent with the $KMnO_4$ sensitivity of the native loop (Borowiec *et al.*, 1987), that is localized overtwisting and exposure of the major-groove atoms in the hypersensitive TTTAT base-pair stretch ~37 bp upstream of (preceding) the O1 site (highlighted in blue on the right side of the protein assemblies illustrated in Figure 11.7). The capability to draw such connections with the chemical literature highlights one of the clear advantages of a sequence-dependent approach over more simplified models currently used in the interpretation of DNA looping

11.6.2
Intrinsic Motions of "Real" DNA Sequences

A new computational approach for studying the dynamic properties of relatively long DNA molecules also keeps track of the local sequence-dependent features of the double helix (Matsumoto and Olson, 2002; Matsumoto, Tobias and Olson, 2005a, 2005b). The description of the collective motions of DNA in terms of base-pair steps builds upon general formulations developed by others to express the normal modes of proteins in terms of internal chemical coordinates (Noguti and Go, 1983;

Braun, Yoshioki and Go, 1984; Higo, Seno and Go, 1985). Although normal-mode analysis has been previously used to study DNA dynamics, it is only now possible, with a base-pair-level perspective, to examine the vibrations of longer pieces of DNA. Apparent mechanical constants, which describe the overall elastic behavior of linear DNA, can be obtained by substituting the computed normal-mode frequencies in classical expressions for the normal modes of an ideal elastic rod (Strutt (Baron Rayleigh), 1945). The calculations show how the natural coupling of local conformational variables affects the global motions of DNA. For example, successful correspondence of the computed stretching modulus with experimental data requires that the DNA base pairs be inclined with respect to the direction of stretching or reoriented upon forced extension (Matsumoto and Olson, 2002). Chain extension is thereby effected by low-energy transverse motions that vary the overlap but preserve the strong van der Waals' stacking of neighboring base-pair planes. The computed stretching constant significantly exceeds observed values if the DNA is pulled in a direction perpendicular to the base-pair planes.

End-to-end constraints placed on a naturally straight DNA molecule, in combination with natural conformational features of the double helix, also account for the substantially larger torsional moduli determined with state-of-the-art, single-molecule experiments (Bryant *et al.*, 2003; Gore *et al.*, 2006) compared to values extracted from solution measurements. The computed normal-mode frequencies and torsional moduli increase substantially if base pairs are inclined with respect to the double-helical axis and the deformations of selected conformational variables follow known interdependent patterns (Matsumoto and Olson, 2006). The changes are greatest if the fluctuations in dimeric twisting are coupled to parameters such as roll, slide, and rise, which directly alter the end-to-end displacement. Imposed restraints that mimic the end-to-end conditions of single-molecule experiments then impede the twisting of base pairs and increase the torsional modulus. Thus, the twisting motions of DNA are predicted to be strongly coupled to other motions, such as global bending and stretching, in which the end-to-end distance of the DNA changes. Indeed, new single-molecule studies (Lionnet *et al.*, 2006; Gore *et al.*, 2006) show that DNA twisting and stretching are coupled at the macromolecular level in the same sense as the high-resolution structural information used in the knowledge-based potentials. That is, when stretched, the double helix overwinds and when compressed, it unwinds.

11.6.3
Sequence-Dependent Effects in Monte-Carlo Simulations of Short DNA

A new, computationally efficient Monte-Carlo technique (Czapla, Swigon and Olson, 2006) provides estimates of the ring-closure properties of short, realistically modeled DNA chains. The method has been used to investigate the cyclization properties of arbitrary-length DNA with greatly enhanced sampling sizes, that is $O(10^{14})$ configurations, and to estimate J factors lower than 0.1 pM with high accuracy. The methodology was checked against classic theoretical predictions of the cyclization properties of an ideal, inextensible, naturally straight, DNA elastic rod (Shimada and

Yamakawa, 1984) and then applied to investigate the extent to which one can account for the unexpectedly large *J* factors of short DNA chains without the need to invoke significant distortions of double-helical structure. Several well-known structural features of DNA – including the presence of intrinsic curvature, roll–twist coupling, or enhanced pyrimidine–purine deformability – bring the computed *J* factors in line with the observed data. Moreover, periodically distributed roll–twist coupling reduces the magnitude of oscillations in *J*, seen in plots of *J* versus chain length, to the extent found experimentally. Thus, in contrast to earlier suggestions (Yan and Marko, 2004; Cloutier and Widom, 2005; Du *et al.*, 2005; Wiggins, Phillips and Nelson, 2005), it is not necessary to melt, kink, or freely twist the DNA double helix to bring the computed *J* factors in line with the enhanced cyclization propensities of molecules much shorter than the persistence length (Cloutier and Widom, 2004, 2005).

Acknowledgments

The US Public Health Service under research grants GM20861 and GM34809 and instrumentation grant RR022375 has generously supported this work. AVC gratefully acknowledges traineeship support from the US Public Health Service (training grant GM08319) and fellowship support from the Program in Mathematics and Molecular Biology (Burroughs–Wellcome Fund Interfaces Program) and DS research support from an A.P. Sloan Fellowship and NSF grant DMS-05-16646. We also thank Mr Mauricio Esguerra Neira for assistance with the preparation of figures.

Glossary

A-philicity: the tendency of a DNA sequence to adopt A-like helical structures (Ivanov and Minchenkova, 1995).

Linking number: a topological measure of the number of times the two strands of a cyclic DNA molecule are interlinked (White, 1989).

Nucleosome core particle: the basic packaging unit of DNA in eukaryotic cells, that is cells with nuclei, containing ~147 bp wrapped ~1.6 left-handed superhelical turns around a core of eight histone proteins.

Poisson ratio: the negative of the ratio of radial compression to axial elongation of an elastic rod (Landau and Lifshitz, 1986).

Purine: a heterocyclic organic molecule constituted of fused six- and five-membered rings that makes up one of the two groups of nitrogeneous bases in DNA. The two purines, adenine (A) and guanine (G), form hydrogen-bonded complexes with the complementary pyrimidines, thymine (T) and cytosine (C).

Pyrimidine: a six-membered heterocyclic organic molecule with nitrogen atoms at the 1 and 3 positions of the ring that makes up one of the two groups of nitrogeneous bases in DNA. See *Purine*.

Superhelix: a higher-order molecular structure formed by regularly spaced distortions of a helix, such as the wrapping of double-helical DNA around the histone proteins in the nucleosome core particle.

Unique dimers: the 10 DNA base-pair steps with unique sequence and structural identities, four of which are self-complementary, with the same sequence in each strand (AT·AT, CG·CG, GC·GC, TA·TA), and six of which are non-self-complementary (AA·TT, AG·CT, GG·CC, GA·TG, AC·GT, CA·TG).

References

Alexeev, D.G., Lipanov, A.A., and Skuratovskii, I.Y. (1987a) Poly(dA)·poly(dT) is a B-type double helix with a distinctively narrow minor groove. *Nature*, **325**, 821–23.

Alexeev, D.G., Lipanov, A.A., and Skuratovskii, I.Y. (1987b) The structure of poly(dA)·poly(dT) as revealed by an X-ray fibre diffraction. *J. Biomol. Struct. Dyn.*, **4**, 989–1011.

Arnott, S. (1999) Polynucleotide secondary structures: an historical perspective, in *Oxford Handbook of Nucleic Acid Structure*, (ed S. Neidle) Oxford University Press, Oxford, UK, pp. 1–38.

Arnott, S., Chandrasekaran, R., Hall, I.H., and Puigjaner, L.C. (1983) Heteronomous DNA. *Nucleic Acids Res.*, **11**, 4141–51.

Arnott, S., Chandrasekaran, R., Hukins, D.W.L., Smith, P.J.C., and Watts, L. (1974) Structural details of a double-helix observed for DNAs containing alternating purine and pyrimidine sequences. *J. Mol. Biol.*, **88**, 523–33.

Arnott, S., Dover, S.D., and Wonacott, A.J. (1969) Least-squares refinement of the crystal and molecular structures of DNA and RNA from X-ray data and standard bond lengths and angles. *Acta Crystallogr.*, **B25**, 2192–206.

Arnott, S. and Selsing, E. (1974a) Structures for the polynucletoide complexes poly(dA). poly(dT) and poly(dT). poly(dA). poly(dT). *J. Mol. Biol.*, **88**, 509–21.

Arnott, S. and Selsing, E. (1974b) The structure of polydeoxyguanylic acid. polydeoxycytidylic acid. *J. Mol. Biol.*, **88**, 551–52.

Barkley, M.D. and Zimm, B.H. (1979) Theory of twisting and bending of chain macromolecules; analysis of the fluorescence depolarization of DNA. *J. Chem. Phys.*, **70**, 2991–3007.

Baumann, C.G., Smith, S.B., Bloomfield, V.A., and Bustamante, C. (1997) Ionic effects on the elasticity of single DNA molecules. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 6185–90.

Baumann C.G., Bloomfield V.A., Smith, S.B., Bustamante, C., Wang, M.D., and Block, S.M. (2000) Stretching of single collapsed DNA Molecules. *Biophys. J.* **78**, 1965–78.

Bednar, J., Furrer, P., Katritch, V., Stasiak, A.Z., Dubochet, J., and Stasiak, A. (1995) Determination of DNA persistence length by cryo-electron microscopy. Separation of the static and dynamic contributions to the apparent persistence length of DNA. *J. Mol. Biol.*, **254**, 579–94.

Berman, H.M., Olson, W.K., Beveridge, D.L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A.R., and Schneider, B. (1992) The Nucleic Acid Database: a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–59.

Bolshoy, A., McNamara, P., Harrington, R.E., and Trifonov, E.N. (1991) Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc. Natl. Acad. Sci. U.S.A.*, **88**, 2312–16.

Borowiec, J.A., Zhang, L., Sasse-Dwight, S., and Gralla, J.D. (1987) DNA supercoiling promotes formation of a bent repression loop in *lac* DNA. *J. Mol. Biol.*, **196**, 101–11.

Braun, W., Yoshioki, S., and Go, N. (1984) Formulation of static and dynamic conformational energy analysis of biopolymer systems consisting of two or more molecules. *J. Phys. Soc. Jpn.*, **53**, 3269–75.

Bryant, Z., Stone, M.D., Gore, J., Smith, S.B., Cozzarelli, N.R., and Bustamante, C. (2003) Structural transitions and elasticity from torque measurements on DNA. *Nature*, **424**, 338–41.

Burd, J.F., Wartell, R.M., Dodgson, J.B., and Wells, R.D. (1975) Transmission of stability (telestability) in deoxyribonucleic acid. *J. Biol. Chem.*, **250**, 5109–13.

Calladine, C.R. and Drew, H.R. (1984) A base-centered explanation of the B-to-A transition in DNA. *J. Mol. Biol.*, **178**, 773–82.

Chandrasekaran, R. and Arnott, S. (1989) The structures of DNA and RNA helices in oriented fibers, in *Landolt-Börnstein Numerical Data and Functional Relationships* *in Science and Technology, Group VII/1b, Nucleic Acids* (ed W. Saenger) Springer-Verlag, Berlin, pp. 31–170.

Cloutier, T.E. and Widom, J. (2004) Spontaneous sharp bending of double-stranded DNA. *Mol. Cell*, **14**, 355–62.

Cloutier, T.E. and Widom, J. (2005) DNA twisting flexibility and the formation of sharply looped protein–DNA complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 3645–50.

Coleman, B.D., Olson, W.K., and Swigon, D. (2003) Theory of sequence-dependent DNA elasticity. *J. Chem. Phys.*, **118**, 7127–40.

Crothers, D.M., Drak, J., Kahn, J.D., and Levene, S.D. (1992) DNA bending, flexibility, and helical repeat by cyclization kinetics. *Methods Enzymol.*, **212**, 3–29.

Czapla, L., Swigon, D., and Olson, W.K. (2006) Sequence-dependent effects in the cyclization of short DNA. *J. Chem. Theory Comput.*, **2**, 685–95.

Davies, D.R. and Baldwin, R.L. (1963) X-ray studies of two synthetic DNA copolymers. *J. Mol. Biol.*, **6**, 251–55.

DeLano W.L. (2002) The PyMOL molecular graphics system. http://www.pymol.org.

Denisov, V.P. and Halle, B. (2000) Sequence-specific binding of counterions to B-DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 629–33.

Dickerson, R.E. (1998) DNA bending: the prevalence of kinkiness and the virtures of normality. *Nucleic Acids Res.*, **26**, 1906–26.

Dickerson, R.E., Bansal, M., Calladine, C.R., Diekmann, S., Hunter, W.N., Kennard, O., von Kitzing, E., Lavery, R., Nelson, H.C.M., Olson, W.K., Saenger, W., Shakked, Z., Sklenar, H., Soumpasis, D.M., Tung, C.-S., Wang, A.H.-J., and Zhurkin, V.B. (1989) Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.*, **208**, 787–91.

Dickerson, R.E. and Drew, H.R. (1981) Structure of a B-DNA dodecamer II. Influence of base sequence on helix structure. *J. Mol. Biol.*, **149**, 761–86.

Dickerson, R.E., Goodsell, D.S., Kopka, M.L., and Pjura, P.E. (1987) The effect of crystal packing on oligonucleotide double helix structure. *J. Biomol. Struct. Dyn.*, **5**, 557–79.

Du, Q., Smith, C., Shiffeldrim, N., Vologodskaia, M., and Vologodskii, A. (2005) Cyclization of short DNA fragments and bending fluctuations of the double helix. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 5397–402.

Early, T.A., Kearns, D.R., Burd, J.F., Larson, J.E., and Wells, R.D. (1977) High resolution proton nuclear magnetic resonance investigation of the structural and dynamic properties of $d(C_{15}A_{15}){\cdot}d(T_{15}\ G_{15})$. *Biochemistry*, **16**, 541–51.

El Hassan, M.A. and Calladine, C.R. (1995) The assessment of the geometry of dinucleotide steps in double-helical DNA: a new local calculation scheme. *J. Mol. Biol.*, **251**, 648–64.

Franklin, R.E. and Gosling, R.G. (1953) Molecular configuration in sodium thymonucleate. *Nature*, **171**, 740–41.

Gessner, R.V., Frederick, C.A., Quigley, G.J., Rich, A., and Wang, A.H.-J. (1989) The molecular structure of the left-handed Z-DNA double helix at 1.0 Ångstrom atomic resolution. Geometry, conformation, and ionic interactions of d(CGCGCG). *J. Biol. Chem.*, **264**, 7921–35.

Gore, J., Bryant, Z., Nöllmann, M., Le, M.U., Cozzarelli, N., and Bustamante, C. (2006) DNA overwinds when stretched. *Nature*, **442**, 836–39.

Gorin, A.A., Zhurkin, V.B., and Olson, W.K. (1995) B-DNA twisting correlates with base pair morphology. *J. Mol. Biol.*, **247**, 34–48.

Hagerman, P.J. (1988) Flexibility of DNA. *Annu. Rev. Biophys. Biophys. Chem.*, **17**, 265–86.

Harvey, S.C. (1983) DNA structural dynamics: longitudinal breathing as a possible mechanism for the B?Z transition. *Nucleic Acids Res.*, **11**, 4867–78.

Heath, P.J., Clendenning, J.B., Fujimoto, B.S., and Schurr, J.M. (1996) Effect of bending strain on the torsion elastic constant of DNA. *J. Mol. Biol.*, **260**, 718–30.

Higo, J., Seno, Y., and Go, N. (1985) Formulation of static and dynamic conformational energy analysis of biopolymer systems consisting of two or more molecules–avoiding a singularity in the previous method. *J. Phys. Soc. Jpn.*, **54**, 4053–58.

Horowitz, D.S. and Wang, J.C. (1984) Torsional rigidity of DNA and length dependence of the free energy of DNA supercoiling. *J. Mol. Biol.*, **173**, 75–91.

Hud, N.V., and Feigon, J. (1997) Localization of divalent metal ions in the minor groove of DNA A-tracts. *J. Am. Chem. Soc.*, **119**, 5756–57.

Hurley, I., Osei-Gyimah, P., Archer, S., Scholes, C.P., and Lerman, L.S. (1982) Torsional motion and elasticity of the deoxyribonucleic acid double helix and its nucleosomal complexes. *Biochemistry*, **21**, 4999–5009.

IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN) (1983) Abbreviations and symbols for the description of conformations of polynucleotide chains. Recommendations 1982. *Eur. J. Biochem.* **131**, 9–15.

Ivanov, V.I. and Minchenkova, L.E. (1995) The A-form of DNA: in search of biological role (a review). *Mol. Biol.*, **28**, 780–88.

Ivanov, V.I., Minchenkova, L.E., Minyat, E.E., Frank-Kamenetskii, M.D., and Schyolkina, A.K. (1974) The B to A transition of DNA in solution. *J. Mol. Biol.*, **87**, 817–33.

Jacobson, H. and Stockmayer, W.H. (1950) Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys.*, **18**, 1600–06.

Kabsch, W., Sander, C., and Trifonov, E.N. (1982) The ten helical twist angles of B-DNA. *Nucleic Acids Res.*, **10**, 1097–104.

Kirchhoff, G.J. (1859) Über das Gleichgewicht und die Bewegung eines unendlich dunnen elastischen Stabes. *J. f. Reine. Angew. Math. (Crelle)*, **56**, 285–313.

Kirchhoff, G. (1876) *Vorlesungen über mathematische Physik, Mechanik*, B. G. Tübner, Leipzig.

Kosikov, K.M., Gorin, A.A., Zhurkin, V.B., and Olson, W.K. (1999) DNA stretching and compression: large-scale simulations of double helical structures. *J. Mol. Biol.*, **289**, 1301–26.

Landau, L.D. and Lifshitz, L.M. (1986) *Theory of Elasticity*, Chapter 2, Pergamon Press, Oxford.

Larsen, T.A., Kopka, M.L., and Dickerson, R.E. (1991) Crystal structure analysis of the B-DNA dodecamer CGTGAATTCACG. *Biochemistry*, **30**, 4443–49.

Leslie, A.G.W., Arnott, S., Chandrasekaran, R., and Ratliff, R.L. (1980) Polymorphism of DNA double helices. *J. Mol. Biol.*, **143**, 49–72.

Lionnet T., Joubaud, S., Lavery, R., Bensimon, D., and Croquette, V. (2006) Wringing out DNA. http://lanl.arxiv.org/abs/q-bio.BM/0602008.

Lowary, P.T. and Widom, J. (1998) New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.*, **276**, 19–42.

Lu, X.-J. and Olson, W.K. (2003) 3DNA: a software package for the analysis, rebuilding, and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–21.

Lu, X.-J. and Olson, W.K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding, and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.*, **3**, 1213–27.

Lu, X.-J., Shakked, Z., and Olson, W.K. (2000) A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.*, **300**, 819–40.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, **389**, 251–60.

Manning, G.S. (2006) The contribution of transient counterion imbalances to DNA bending fluctuations. *Biophys. J.*, **90**, 3208–15.

Marko, J.F. and Siggia, E.D. (1995) Stretching DNA. *Macromolecules*, **28**, 8759–70.

Marvin, D.A., Spencer, M., Wilkins, M.H.F., and Hamilton, L.D. (1958) A new configuration of deoxyribonucleic acid. *Nature*, **182**, 387–88.

Matsumoto, A. and Olson, W.K. (2002) Sequence-dependent motions of DNA: a normal mode analysis at the base-pair level. *Biophys. J.*, **83**, 22–41.

Matsumoto, A. and Olson, W.K. (2006) Predicted effects of local conformational coupling and external restraints on the torsional properties of single DNA molecules. *Multiscale Model. Simul.*, **5**, 1227–47.

Matsumoto, A., Tobias, I., and Olson, W.K. (2005a) Normal mode analysis of circular DNA at the base-pair level. I. Comparison of computed motions with the predicted behavior of an ideal elastic rod. *J. Chem. Theory Comput.*, **1**, 117–29.

Matsumoto, A., Tobias, I., and Olson, W.K. (2005b) Normal mode analysis of circular DNA at the base-pair level. II. Large-scale configurational transformation of a naturally curved molecule. *J. Chem. Theory Comput.*, **1**, 130–42.

Merritt, E.A. and Bacon, D.J. (1997) Raster3D: photorealistic molecular graphics. *Methods Enzymol.*, **277**, 505–24.

Millar, D.P., Robbins, R.J., and Zewail, A.H. (1982) Torsion and bending of nucleic acids studied by subnanosecond time-resolved fluorescence depolarization of intercalated dyes. *J. Chem. Phys.*, **76**, 2080–94.

Noguti, T. and Go, N. (1983) Dynamics of native globular proteins in terms of dihedral angles. *J. Phys. Soc. Jpn.*, **52**, 3283–88.

Odijk, T. (1995) Stiff chains and filaments under tension. *Macromolecules*, **28**, 7016–18.

Olson, W.K. (1982) Theoretical studies of nucleic acid conformation: potential energies, chain statistics, and model building, in *Topics in Nucleic Acid Structure, Vol. Part II* (ed S.E. Neidle) Macmillan Press, London, pp. 1–79.

Olson, W.K., Bansal, M., Burley, S.K., Dickerson, R.E., Gerstein, M., Harvey, S.C., Heinemann, U., Lu, X.-J., Neidle, S., Shakked, Z., Sklenar, H., Suzuki, M., Tung, C.-S., Westhof, E., Wolberger, C., and Berman, H.M. (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**, 229–37.

Olson, W.K., Colasanti, A.V., Czapla, L., and Zheng, G. (2008) Insights into the sequence-dependent macromolecular properties of DNA from base-pair level modeling, in *Coarse-Graining of Condensed Phase and Biomolecular Systems* (ed G.A. Voth), Taylor and Francis Group, LLC, Boca Raton, FL, pp. 205–23.

Olson, W.K., Gorin, A.A., Lu, X.-J., Hock, L.M., and Zhurkin, V.B. (1998) DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 11163–68.

Olson, W.K., Srinivasan, A.R., Marky, N.L., and Balaji, V.N. (1983) Theoretical probes of DNA conformation examining the B leads to Z conformational transition. *Cold Spring Harb. Symp. Quant. Biol.*, **47**, 229–41.

Olson, W.K., Swigon, D., and Coleman, B.D. (2004) Implications of the dependence of the elastic properties of DNA on nucleotide sequence. *Philos. Trans. R. Soc. London, Ser. A*, **362**, 1403–22.

Ornstein, R.L. and Fresco, J.R. (1983) Correlation of T_m and sequence of DNA duplexes with ΔH computed by an improved empirical potential method. *Biopolymers*, **22**, 1979–2000.

Podestà, A., Indrieri, M., Brogioli, D., Manning, G.S., Milani, P., Guerra, R., Finzi, L., and Dunlap, D. (2005) Positively charged surfaces increase the flexibility of DNA. *Biophys. J.*, **89**, 2258–563.

Premilat, S. and Albiser, G. (1986) DNA models for A, B, C and D conformations related to fiber X-ray, infrared and NMR measurements. *J. Biomol. Struct. Dyn.*, **3**, 1033–43.

Richmond, T.J. and Davey, C.A. (2003) The structure of DNA in the nucleosome core. *Nature*, **423**, 145–50.

Saenger, W. (1984) *Principles of Nucleic Acid Structures*, Chapter 14, Springer-Verlag, New York.

Schneider, B., Neidle, S., and Berman, H.M. (1997) Conformations of the sugar–phosphate backbone in helical DNA crystal structures. *Biopolymers*, **42**, 113–24.

Seeman, N.C., Rosenberg, J.M., and Rich, A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **73**, 804–8.

Selsing, E., Wells, R.D., Alden, C.J., and Arnott, S. (1979) Bent DNA: visualization of a base-paired and stacked A-B conformational junction. *J. Biol. Chem.*, **254**, 5417–22.

Shimada, J. and Yamakawa, H. (1984) Ring-closure probabilities for twisted wormlike chains. Application to DNA. *Macromol.*, **17**, 689.

Shore, D. and Baldwin, R.L. (1983) Energetics of DNA twisting. II. Topoisomer analysis. *J. Mol. Biol.*, **170**, 983–1007.

Smith, S.B., Finzi, L., and Bustamante, C. (1992) Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science*, **258**, 1122–26.

Srinivasan, A.R. and Olson, W.K. (1987) Nucleic acid model building: the multiple backbone solutions associated with a given base morphology. *J. Biomol. Struct. Dyn.*, **4**, 895–938.

Strauss, J.K. and Maher III, L.J. (1994) DNA bending by asymmetric phosphate neutralization. *Science*, **266**, 1829–34.

Strutt (Baron Rayleigh), J.W. (1945) *Theory of Sound*, Chapter 7, Dover Publications, London.

Subirana, J.A. and Faria, T. (1997) Influence of sequence on the conformation of the B-DNA helix. *Biophys. J.*, **73**, 333–38.

Suzuki, M. and Yagi, N. (1995) Stereochemical basis of DNA bending by transcription factors. *Nucleic Acids Res.*, **23**, 2083–91.

Swigon, D., Coleman, B.D., and Olson, W.K. (2006) Modeling the Lac repressor-operator assembly: the influence of DNA looping on Lac repressor conformation. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 9879–84.

Swigon, D. and Olson, W.K. (2008) Mesoscale modeling of multi-protein–DNA assemblies: the role of the catabolic activator protein in Lac-repressor-mediated looping. *Int. J. Nonlinear Mechanics* **43**, 1082–93.

Thomas, J.C., Allison, S.A., Apellof, C.J., and Schurr, J.M. (1980) Torsion dynamics and depolarization of fluorescence of linear macromolecules. II. Fluorescence polarization anisotropy measurements on a clean viral 29 DNA. *Biophys. Chem.*, **12**, 177–88.

Tolstorukov M.Y., Colasanti, A.V., McCandlish, D., Olson, W.K., and Zhurkin, V.B. (2007) *A novel "roll-and-slide" mechanism of DNA folding in chromatin.* Implications for nucleosome positioning. submitted.

van Dam, L. and Levitt, M.H. (2000) BII nucleotides in the B and C forms of natural-sequence polymeric DNA: a new model for the C forms of DNA. *J. Mol. Biol.*, **304**, 541–61.

Vargason, J.M., Henderson, K., and Ho, P.S. (2001) A crystallographic map of the transition from B-DNA to A-DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 7265–70.

Wang, A.H.-J., Quigley, G.J., Kolpak, F.J., Crawford, J.L., van Boom, J.H., van der Marel, G., and Rich, A. (1979) Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*, **282**, 680–86.

Wang, Y., Thomas, G.A., and Peticolas, W.L. (1987) Sequence dependent conformations of oligomeric DNA's in aqueous solutions and in crystals. *J. Biomol. Struct. Dyn.*, **5**, 249–74.

Wang, Y., Thomas, G.A., and Peticolas, W.L. (1989) A duplex of the oligonucleotides d(GGGGGTTTTT) and d(AAAAACCCCC) forms an A to B conformational junction in concentrated salt solutions. *J. Biomol. Struct. Dyn.*, **6**, 1177–87.

Watson, J.D. and Crick, F.H.C. (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature*, **171**, 964–67.

White, J.H. (1989) An introduction to the geometry and topology of DNA structure, in *Mathematical Methods for DNA Sequences* (ed M.S. Waterman), CRC Press, Boca Raton, FL, pp. 225–53.

Wiggins, P.A., Phillips, R., and Nelson, P.C. (2005) Exact theory of kinkable elastic polymers. *Phys. Rev. E*, **71**, 021909.

Williams, L.D. and Maher III, L.J. (2000) Electrostatic mechanisms of DNA deformation. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 497–521.

Willis, B.T.M. and Pryor, A.W. (1975) *Thermal Vibrations in Crystallography*, Cambridge University Press, London.

Yan, J. and Marko, J.F. (2004) Localized single-stranded bubble mechanism for cyclization of short double helix DNA. *Phys. Rev. Lett.*, **93**, 108108.

Zhurkin, V.B., Lysov, Y.P., and Ivanov, V.I. (1979) Anisotropic flexibility of DNA and the nucleosomal structure. *Nucleic Acids Res.*, **6**, 1081–96.

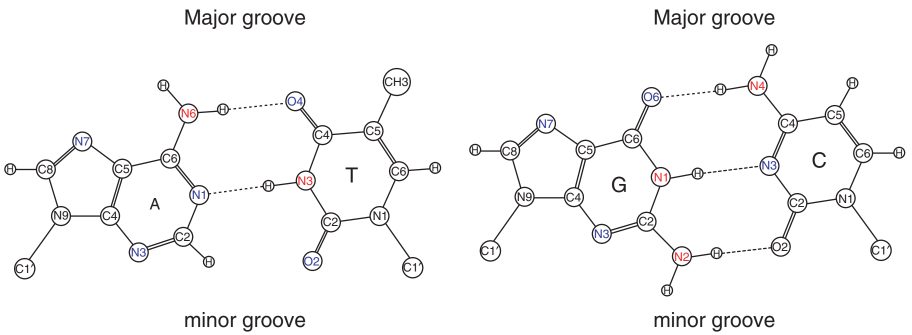
Major groove
Major groove
A
T
G
C
N6
H
O4
CH3
C4
C5
C6
N7
C5
C8
N1
N3
N9
C4
N3
C2
C2
N1
O2
C1'
O6
N4
N1
N2
minor groove
minor groove

Fig. 11.1 Comparison of hydrogen-bonding interactions and chemical structures (including double bonds) of the purine and pyrimidine bases comprising normal A.T and G.C Watson–Crick pairs. Hydrogen bonds designated by dashed lines. Proton donor and acceptor atoms denoted in red and blue, respectively. Note the pseudosymmetric positioning of the purine N3 and pyrimidine O2 atoms on the minor-groove edges of the base pairs and the unique sequence-dependent signatures (Seeman, Rosenberg and Rich, 1976) of the unfulfilled donor and acceptor atoms on the major-groove edges of A.T versus G.C. (This figure also appears on page 363.)

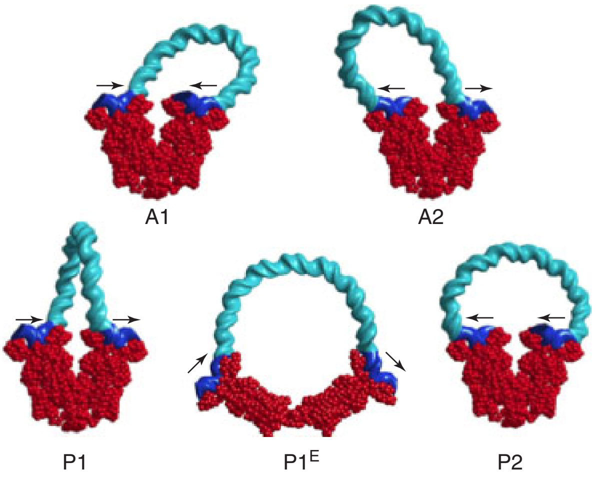
A1
A2
P1
P1^E
P2

Fig. 11.7 Minimum-energy configurations of the 92-bp DNA loops formed upon binding of the Lac repressor (LacR) protein assembly to the so-called O3 and O1 operator sites on the *Escherichia coli lac* operon. DNA shown in aqua, LacR in red, and operator sites in blue. Arrows above operator sites (O3 left, O1 right) denote the orientations of DNA – antiparallel (A1, A2) and parallel (P1, P1^E, P2) – on the protein assembly. (This figure also appears on page 375.)

12
Models of Membrane Fusion

Sylvio May

12.1 Introduction 385
12.2 Bilayer Fusion 387
12.2.1 Lamellar-to-Inverted Phase Transitions 387
12.2.1.1 Lipidic Intramembraneous Particles 388
12.2.1.2 Extended Lipid Conformations 388
12.2.1.3 Direct Formation of Inverted Tubular Micelles 389
12.2.1.4 Stalks 389
12.2.1.5 Transmonolayer Contact and Interlamellar Attachment 390
12.2.2 Divalent Cations 390
12.2.3 Polymer-Induced Fusion 391
12.2.4 Lipids as Fusion Modulators 391
12.2.5 Peptides as Fusion Modulators 393
12.2.6 Oppositely Charged Membranes 393
12.3 Viral Fusion 395
12.3.1 Internalization through Endocytosis 396
12.3.2 Viral Fusion Proteins 396
12.3.3 Hemagglutinin 397
12.4 Cellular Fusion 398
12.4.1 SNAREs 398
12.4.2 Cell–Cell Fusion 401
12.4.3 Fusion in Mitochondria 402
12.5 Modeling Fusion 403
12.5.1 Forces between Membranes 403
12.5.1.1 Electrostatic Forces 403
12.5.1.2 Van der Waals Forces 403
12.5.1.3 Undulation Forces 404
12.5.1.4 Hydration Forces 404
12.5.1.5 Polymer-Induced Forces 404
12.5.2 Membrane Elasticity 404
12.5.3 Membrane Pores 406
12.5.4 Modeling Fusion Intermediates 406

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

12.6 Simulating Fusion 407
12.6.1 Molecular Dynamics Simulations 408
12.6.2 Dissipative Particle Dynamics 409
12.6.3 Solvent-Free Simulations 410
12.7 Conclusions 411
Acknowledgments 411
Glossary 411
References 412
Further Reading 414

12.1
Introduction

Biomembranes fulfill multiple functions in living cells, such as compartimentalization, formation of permeability barriers, regulation of transport, and service as platform for numerous biochemical activities. The molecules that enable this variability in functions are lipids. Driven by the hydrophobic effect, lipids spontaneously self-assemble into a variety of ordered structures in which the lipid headgroups cooperatively shield the hydrocarbon tails from the unfavorable exposure to the aqueous environment. The planar lipid bilayer is just one such structure; among a plethora of other aggregation geometries are spherical- and rod-like micelles, as well as inverse and bicontinous phases.

Fluid lipid bilayers are soft materials with remarkable physical properties that include the ability to change their shape and topological structure. Contained in the latter is membrane rupture (which leads to the formation of a pore), membrane fusion–loosely defined as the merger of two lipid bilayers into a single one–and fission (the inverse of fusion). Hence, membrane fusion is already integrated into the design of the biomembrane's basic building blocks. It is therefore plausible that the fusion of *bare* bilayers has been studied intensively in the past (Section 12.2). Initially, some focus in these studies was put on understanding the fusion between highly charged vesicles induced by divalent cations, especially calcium (Ca^{2+}). Here, a complex set of interactions leads to the attraction of like-charged vesicles followed by a large-scale aggregation and condensation of the apposed membranes (Arnold, 1995). This contrasts other more localized fusion pathways in bare membranes and also in biological systems where the two membranes are brought together and become finally merged by a specialized fusion machinery.

Two closely apposed lipid vesicles (and more generally, any two membrane-enclosed aqueous compartments) need not necessarily fuse with each other; fusion is just one out of several possible outcomes. Depending on a variety of different factors the vesicles may remain aggregated, undergo only part of the fusion reaction, or rupture. The latter case may eventually lead to the formation of a single vesicle but it is accompanied by leakage. A definition of membrane fusion that excludes this alternative scenario has been stated by Lentz (2007): Membrane fusion is characterized by

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

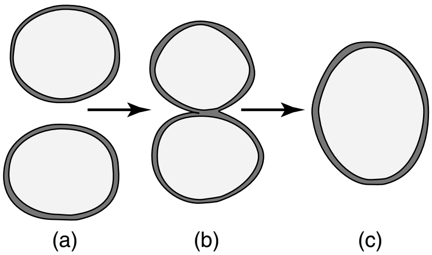

Fig. 12.1 Fusion of two lipid vesicles: (a) Prior to fusion, the two vesicles are separated; (b) The fusion process is generally thought to proceed through intermediate states such as the merger of only the proximal monolayers followed by the formation of a fusion pore; (c) Finally, after fusion is completed, only one single vesicle remains. Both lipids and contents of the initially two vesicles are mixed in the final state, and little leakage has occurred.

"mixing of entrapped content between two membrane-enclosed aqueous compartments accompanied by mixing of membrane components (lipids and likely proteins) but with little accompanying leakage of entrapped contents." The fusion of two vesicles according to this definition is illustrated in Figure 12.1.

Topological changes that are reminiscent of membrane fusion also occur in structural membrane transitions of lipid bilayers, particularly in the transition from the planar bilayer to the inverse hexagonal phase (the $L_\alpha \rightarrow H_{II}$ transition) and to inverse cubic phases (the $L_\alpha \rightarrow Q_{II}$ transition). Both the planar phase (L_α) and the inverse hexagonal phase (H_{II}) are illustrated in Figure 12.4 (Section 12.2.4). The forces that drive the transition result from a change of the preferred aggregation geometry of the involved lipids (here too see Section 12.2.4). At least in some cases, the transition involves the formation of an intermediate structure–a so-called stalk–that very likely also appears during fusion of model and perhaps even biological membranes (Chernomordik, Zimmerberg and Kozlov, 2006). Studying structural phase transitions (Section 12.2.1) therefore also contributes to the understanding of membrane fusion.

That fusion proceeds through the formation of a stalk intermediate has been suggested early (Kozlov and Markin, 1983). The stalk hypothesis has since then become a central paradigm of membrane fusion. A large number of experimental results can be explained based on a stalk-involving fusion pathway. This includes the way in which nonlamellar phase-forming lipids promote or inhibit fusion (Section 12.2.4). Other fusion promoters, such as fusogenic peptides, are less easily understood. Indeed, the ability of certain peptides to promote fusion appears to depend sensitivity on their amino acid structure, often with no simple correlation between structure and mechanism of action. Generally, stalk formation and subsequent opening of the fusion pore have attracted considerable modeling efforts, based on elasticity theory (Section 12.5.4) and computer simulations (Section 12.6).

Biological fusion is vital for a multitude of functions–fertilization, formation of multinuclear cells in development, intracellular membrane trafficking and recycling, exocytosis, and infections caused by enveloped viruses–all require the merger of two biomembranes. Biological fusion is catalyzed by special proteins. The best-studied cases are viral fusion proteins (and here especially the fusion protein hemagglutinin (HA) of the influenza virus; see Section 12.3.3) and the synaptic vesicle fusion machinery (centered around SNARE complexes;

Section 12.4.1). Yet many other systems are being studied intensively as well, among them class II viral fusion proteins (Section 12.3.2) and cell–cell fusion (Section 12.4.2). A central question is whether–despite large differences in the structure of different fusion proteins–there are common fusion mechanisms, such as the formation of a stalk intermediate. Related is to what extent biological fusion is lipid or protein based (Mayer, 2001). In other words, do fusion proteins merely lower the energy barriers to the fused state or do they create an entirely new fusion path that does not exist for bare membranes?

12.2
Bilayer Fusion

This section is concerned with the fusion of membranes in the absence of fusion-proteins. When vesicles in aqueous solution approach each other by means of thermal motion or by an external field, a repulsive force between the respective bilayers acts to prevent close contact. The net interbilayer force can be decomposed into electrostatic forces, undulation forces, and hydration forces (Section 12.5.1). Particularly the latter ones, which represent structural forces due to the ordering of solvent molecules close to the membrane interfaces, become very strong at short distances. Vesicle aggregation can nevertheless be induced through divalent cations for like-charged membranes (Section 12.2.2), through polymers (Section 12.2.3), or by using oppositely charged membranes (Section 12.2.6). These aggregation methods prime for fusion but do not intrinsically induce it. That is, the membranes may reside in an aggregated state for a long time period. For example, large concentrations of sodium (of about 0.5M) induce aggregation but not fusion of sonicated phosphatidylserine (PS) vesicles (Day *et al.*, 1980). Needed for fusion is an additional destabilization (i.e. stress) of the planar lipid bilayer. In bare lipid bilayers, stress occurs as *curvature stress* or *lateral tension.* They express deviations of the curvature and lateral area, respectively, of a lipid layer from the preferred values.

Fusion is generally believed to proceed through intermediate states. The presence of intermediates reflects local minima in the energy landscape of the fusion path. One of the most frequently discussed fusion pathways–based on the stalk-pore model (Kozlov and Markin, 1983)–is shown in Figure 12.2. It starts from two closely apposed membranes (a), evokes the formation of a *stalk* (b) which then widens into a *transmembrane contact* (c). Further, radial expansion gives rise to an extended *hemifusion diaphragm* which ruptures (e) into a *fusion pore* (f).

12.2.1
Lamellar-to-Inverted Phase Transitions

Inverted lipid phases include the inverse hexagonal phase (H_{II} phase) and various bicontinous cubic phases (Q_{II} phase) (Seddon *et al.*, 2000). The inverse hexagonal phase (H_{II} phase) consists of long rod-like inverted micelles that are parallel and closely packed into an hexagonal array. Cubic phases lack the translational symmetry of the inverse hexagonal phase; they form a network of inverted micellar tubes that has cubic symmetry. The packing geometry in the Q_{II} phases is generally more *inverted cone*-like (Section 12.2.4) than

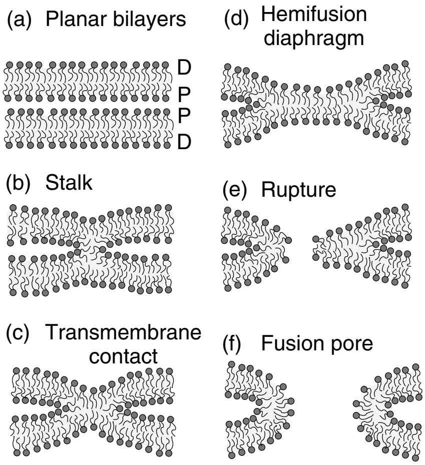

Fig. 12.2 The fusion pathway according to the stalk-pore model. Two closely apposed planar membranes (a) connect their proximal monolayers through a *stalk* (b) which widens into a *transmembrane contact* (c). A fusion pore (f) may form already at this point or after radial expansion into an extended *hemifusion diaphragm* (d) and subsequent rupture (e). All diagrams show two-dimensional schematic illustrations of three-dimensional systems. The lipid headgroups are represented by dark circles, lipid tails comprise the hydrocarbon region as indicated. The proximal and distal monolayers of the planar bilayers in diagram (a) are marked by the letters "P" and "D", respectively.

in the H_{II} phase. Lipids with small, weakly hydrated headgroups or with a bulky hydrocarbon core are expected to form such structures. Lamellar-to-inverted phase transitions can be triggered by increasing the temperature or by changing the hydration level (Gawrisch *et al.*, 1992). A typical example is dioleoylphosphatidylethanolamine (DOPE) with a transition temperature of 3 °C in excess water (Toombes *et al.*, 2002). Increase in temperature disorders the hydrocarbon chain region to a larger extent than the headgroup region, thus shifting the preferred packing geometry toward more inverted cone-like.

Transitions between lamellar and inverted phases, especially involving the inverse hexagonal phase, have been studied intensively in the past, partly because of their possible relevance for membrane fusion. A central objective is the identification of intermediate structures between the L_α and H_{II} phases, making them candidates to appear as fusion intermediates as well. Experimental observations, based on techniques such as X-ray diffraction, spectroscopy, freeze-fracture and cryo-electron microscopy, support a number of distinct intermediates.

12.2.1.1 Lipidic Intramembraneous Particles

Early observations based mainly on freeze-fracture microscopy revealed so-called lipidic intramembraneous particles which are spatially localized, typically string-like, defects, interpreted as either (a) inverted micelles sandwiched between the two leafs of a single bilayer (Verkleij *et al.*, 1979) (illustrated in Figure 12.3a) or two bilayers (Figure 12.3b) or (b) intermembrane attachment sites (Hui and Stewart, 1981). Note that the latter interpretation corresponds closely to the modern terminology of stalk-like intermediates. Theoretical analysis suggests that the structure in Figure 12.3a has generally larger energy than that in Figure 12.3b Siegel (1984).

12.2.1.2 Extended Lipid Conformations

The so-called extended lipid conformation refers to lipids that have their

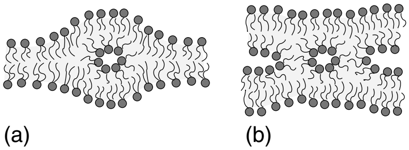

Fig. 12.3 "Lipidic intramembraneous particles" were among the first structures suggested as intermediates in the $L_\alpha \rightarrow H_{II}$ transition. The illustration shows a single inverted micelle (which can be spherical or tubular) sandwiched between a single bilayer (a) or between two bilayers (b). Note that a structure similar to (b) was recently suggested to be involved in the growth mechanism of the tubular rods in the H_{II} phase (Rappolt *et al.*, 2003).

two hydrocarbon chains reside in the proximal monolayers of two adjacent membranes. Instead of an essentially parallel chain alignment, these lipids reside in an extended conformation with the two hydrocarbon tails pointing in opposite directions. Such an arrangement is prohibited in a single bilayer by the hydrophobic effect but has been suggested (Kinnunen and Holopainen, 2000) to relieve packing stress for two apposed membranes that are about to undergo the transition to an inverted phase. The actual transition is triggered by the extended lipids in a "zipper"-like mechanism. Support for this hypothesis derives mainly from nuclear magnetic resonance (NMR) and fluorescence spectroscopy.

12.2.1.3 **Direct Formation of Inverted Tubular Micelles**

X-ray scattering-based structure reconstitution of multilamellar palmitoylphosphatidylethanolamine (POPE) vesicles within the region of L_α/H_{II} phase coexistence suggests a model of cooperative formation of inverted tubular micelles sandwiched within the multilamellar bilayer stack (Rappolt *et al.*, 2003). Since the thickness of the free water layer between adjacent membranes was found to be very small (less than 1 nm) the inverted tubular micelles are suggested to form directly from two proximal monolayers without having to pass through an intermediate. The model is structurally related to that of inverted micelles sandwiched between two bilayers (Figure 12.3b).

12.2.1.4 **Stalks**

A stalk is a lipid connection between the proximal monolayers of two fusing bilayers (Figure 12.2b). Note that the distal monolayers have not established direct contact between their hydrocarbon chain regions. The formation of a stalk occurring as an intermediate during either a structural phase transition between the lamellar and an inverted phase or during a fusion reaction was first suggested as a hypothesis (Kozlov and Markin, 1983), but has subsequently gained experimental support. Evidence derives mainly from cryo-electron microscopy (Siegel and Epand, 1997) and X-ray scattering (Yang and Huang, 2002). The latter showed that stalks can form as an equilibrium structure for several different lipid systems at room temperature and low hydration. Structure reconstitution through X-ray scattering on a membrane stack was possible because the stalks between any two membranes arranged into an hexagonal lattice. A similar structure was also observed to form in a coarse-grained molecular dynamics (MD) simulation (Marrink and Mark, 2004).

12.2.1.5 Transmonolayer Contact and Interlamellar Attachment

Pathways of structural transitions involving stalks must branch at some point depending on whether the transition proceeds to H_{II}- or Q_{II} phase. Models for both scenarios have been developed (Siegel, 1999; Siegel, 1993). Stalks are suggested to evolve into transmonolayer contacts (TMCs)–illustrated in Figure 12.2c–in which the hydrocarbon chains of the distal monolayers start contacting each other. The H_{II} phase could then form upon elongation of TMCs followed by an aggregation. On the other hand, immediate opening of a TMC creates a pore, the so-called interlamellar attachment (ILA). Lattice formation of ILAs would then lead to the Q_{II} phase.

12.2.2
Divalent Cations

A number of biologically important lipids such as PS, phosphatidyglycerol (PG), and phosphatidic acid (PA) are anionic. Vesicles that contain large fractions of these charged lipids can be fused by adding divalent cations such as Ca^{2+} or Mg^{2+}. The threshold concentration required to induce fusion is in the millimolar range (Arnold, 1995) and is generally lower for Ca^{2+} than for Mg^{2+}. For example, Ca^{2+} is able to fuse both large and small unilamellar PS vesicles, whereas Mg^{2+} is unable to fuse large unilamellar PS vesicles but mediates aggregation. Monovalent cations such as Na^+ and K^+ (with the exception of H^+) are able to induce vesicle aggregation but usually not fusion. Generally, the ability of a given cation to induce aggregation is distinct and well separated from its ability to induce fusion. Both processes are separate and superimposable. For example, Mg^{2+} enhances the rate of Ca^{2+}-mediated fusion in large unilamellar PS vesicles through its ability to preaggregate the vesicles (Wilschut, Düzgüne and Papahadjopoulos, 1981)

No nonlamellar phase-forming lipids are needed for divalent ion-mediated fusion. Hence, the fusion mechanism does not likely involve the formation of a stalk intermediate. Instead, the two apposed membranes form a highly condensed, dehydrated complex in which the cation bridges and rigidifies the two adhering membranes. The condensed complex would tend to maximize its contact area thus exerting a large bending moment and tension onto the remaining (not aggregated) vesicle parts, particularly at the rim region. It is plausible that this leads to either membrane rupture or fusion, but the detailed molecular mechanism is unclear at present. A conceptual model that captures this scenario is the adhesion–condensation mechanism (Kozlov and Markin, 1984).

Cytosolic calcium concentration is usually about 100 nM, whereas external calcium is present in the millimolar range. Intracellular calcium concentrations are tightly regulated as this ion is involved in numerous signaling processes. For example, in secretory cells, calcium enters into the cytosol with micromolar concentrations from both the extracellular space and internal stores located in the endoplasmatic reticulum (ER). Due to the generally low concentrations, and also due to the fairly low charge densities of biological membranes, it may not seem likely that calcium is involved directly in the fusion of sectretory vesicles with the negatively charged plasma membrane.

12.2.3
Polymer-Induced Fusion

Water-soluble polymers are able to promote the fusion of lipid vesicles. A convenient polymer is polyethylene glycol (PEG) which has been studied most extensively. The role of the polymer is to induce a depletion attraction between vesicles (Section 12.5.1). The method is often used as an auxiliary tool to enhance fusion rates of membranes that are already primed for fusion. For example, the presence of PEG reduces the concentration of calcium needed to trigger the fusion between anionic vesicles (Ohki and Arnold, 1990). Another example is small unilamellar vesicles (SUVs) whose curvature stress can be relieved by fusing them into larger vesicles. As has been pointed out, the use of SUVs could affect the fusion mechanism as compared to large vesicles because they limit the size of possible fusion intermediates and require trans-monolayer redistribution of lipids (Siegel, 2004).

PEG-induced fusion has been employed to investigate the dynamics of the fusion process. One of the key studies (Lee and Lentz, 1997) lead to the identification of a sequence of steps for the particular system of 45 nm large vesicles consisting of a mixture of dioleoylphosphatidylcholine (DOPC) and dilinolenoylphosphatidylcholine of composition 85/15 mol/mol. The steps involve two intermediates that are consistent with stalk formation (Figure 12.2b) followed by transmonolayer contact (TMC, Figure 12.2c). Kinetic modeling of this system predicted activation energies of about 60 k_BT to reach the first intermediate from an initially aggregated state, 50 k_BT for the second intermediate, and 36 k_BT for the final fusion pore. (Here k_BT is the thermal energy with k_B being Boltzmann's constant and T the absolute temperature.) The subsequent analysis for other systems lead to a generalized model (Weinreb and Lentz, 2007) for PEG-induced fusion as a two-state sequential process with the two intermediates accounting in a general sense for an ensemble of structures related to stalks and TMCs. PEG has been used extensively in the past to study the fusion of bare membranes but, of course, it is equally valuable to study how proteins promote fusion (Lentz *et al.*, 2000).

12.2.4
Lipids as Fusion Modulators

Bilayer stability can be characterized on a qualitative level in terms of the *effective shape* of lipid molecules (Israelachvili, 1992). This shape is dynamic, it results from an averaging over many ($\approx 10^{10}$ s^{-1}) different conformations of the flexible hydrocarbon tails.

The geometry of planar bilayers indicates a preferential *cylinder-like* shape of the lipid molecules. Canonical example of a bilayer-forming lipid is phosphatidylcholine (PC), but others such as the anionic lipids phosphatidylglycerol (PG) and PS form as well stable bilayers at physiological conditions. Lipids with preferred *conical* shape such as the single-tailed lyso-phosphatidylcholine tend to aggregate into spherical or cylindrical micelles. Correspondingly, the preferred shape of an *inverted cone* implies formation of inverted phases. For example, phosphatidylethanolamine (PE) typically forms the inverse hexagonal

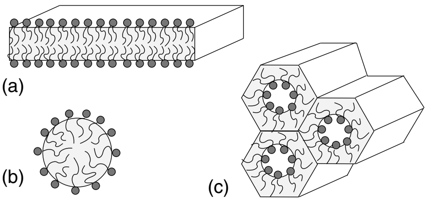

Fig. 12.4 Schematic illustration of three basic aggregation geometries: (a) planar bilayer (L_α); (b) spherical or cylindrical micelle; (c) inverse hexagonal phase (H_{II}). The corresponding effective shape of the lipids in these aggregates is cylindrical, conical, and inverse conical respectively.

phase. Figure 12.4 schematically illustrates three basic aggregation geometries. Note that the effective shape of a lipid species depends also on external conditions (such as temperature, level of hydration, pH, and salt content). For example, lowering the temperature induces the $H_{II} \rightarrow L_\alpha$ transition in DOPE. Upon also lowering the level of hydration, a transition back to the L_α phase may occur (Gawrisch *et al.*, 1992). Note also that a quantitative measure of the propensity to adopt curved aggregation geometries is provided by elasticity theory. As outlined in Section 12.5.2, monolayers composed of cone-like (inverted cone-like) lipids imply a positive (negative) *spontaneous curvature*; see Eq. (1).

Generally, lipids that favor inverse phases also facilitate fusion. PE is a well-studied example of such "fusogenic" lipids (Cullis and de Kruijff, 1978). The PE headgroup is sterically small; compared to PC it lacks three methyl groups. Beyond that, it forms an intermolecular hydrogen bond network which is accompanied by a low level of hydration, thus reducing the hydration repulsion (Section 12.5.1) between apposed bilayers.

If the fusion path involves highly curved intermediates, it can be expected that small concentrations of lipids that form nonlamellar phases affect the propensity to undergo fusion. Thus, if the preferred molecular shape of the added nonlamellar lipids fits that of the intermediate, fusion (or, at least, the rate of adopting the intermediate state) will be facilitated. Based on this notion, experiments can be designed to probe putative intermediate structures. For example, the stalk-pore model (Figure 12.2) predicts that stalk formation is promoted (inhibited) by inserting inverted cone-like (cone-like) lipids into the proximal monolayers. In other words, a negative spontaneous curvature of the proximal monolayers supports stalk formation. Conversely, opening of the fusion pore is predicted to be facilitated by a positive spontaneous curvature of the distal monolayers. This opposite influence of nonlamellar lipids on different stages of fusion, depending on whether they are present in the proximal or distal monolayers, is indeed observed experimentally (Chernomordik *et al.*, 1995).

Diacylglycerol (DAG) is a lipid of some interest because it modulates the physical properties of the host membrane and also acts as a second messenger-signaling lipid. More specifically, upon activation, the membrane-bound enzyme phospholipase C splits the lipid phosphatidyl inositol-bisphosphate (PIP_2) into its water- soluble headgroup inositol triphosphate (IP3) and DAG. The headgroup-lacking lipid remains in the membrane where it could subsequently act as a fusion promotor through its strongly inverted cone-like shape. Support comes from model membranes where the production of DAG by phospholipase C triggers fusion in a manner that correlates with the formation of cubic phase intermediates (Goñi and Alonso, 1999).

12.2.5
Peptides as Fusion Modulators

Naturally occurring and synthetic peptides as well as protein fragments are widely studied in terms of their ability to destabilize lipid bilayers and to induce structural changes. For example, antimicrobial peptides, used by many animals as host defense systems, trigger pore formation or solubilize their target membrane (Bechinger and Lohner, 2006). Sometimes, peptides leave the bilayer structure intact but tubulate the shape of the membrane. In other cases, transmembrane peptides, both natural (such as gramicidin A) and synthetic, are able to induce structural transitions from a planar lipid bilayer to inverse nonlamellar phases. Among the most remarkable examples is the synthetic peptide WALP (acetyl-GWW(LA)$_n$ WWA-ethanolamide with $n = 5-10$) which induces the H_{II} phase in a PC membrane, given the so-called hydrophobic mismatch between membrane and peptide is sufficiently large (Killian, 1998). Intermediate structures for transmembrane peptide–lipid systems during the $L_\alpha \rightarrow H_{II}$ phase transition are not known.

Viral fusion peptides have evolved to trigger fusion. Not surprisingly, some of these peptides are also able to induce inverse nonlamellar phases (Epand, 1997). Yet, the mechanism by which viral fusion peptides destabilize their target membrane during viral fusion is not well understood. One possibility is softening of the host membrane in order to decrease the energy to form fusion intermediates (Tristram-Nagle and Nagle, 2007). On the basis of molecular modeling it has been suggested that oblique orientations of fusion peptides in the membrane are crucial for their function (Peuvot *et al.*, 1999). Most detailed experimental evidence is available for the fusion peptides of the influenza and immunodeficiency viruses. The structure of both peptides as determined by NMR in micelles (Tamm, Lai and Li, 2007) is shown in Figure 12.5. In the latter case, the helix is straight and very likely oblique in the membrane. Yet, the fusion peptide of HA (the fusion protein of influenza virus) consists of two α-helical segments linked by a rigid kink region. The peptide thus forms a boomerang-like shape with both its arms being tilted with respect to the membrane interface (Tamm, Lai and Li, 2007). Boomerang-like insertion has also been found for the fusogenic peptide sequence B18 from the sea urchin fertilization protein bindin (Afonin *et al.*, 2004) which could indicate a general structural motif for fusogenic peptides.

12.2.6
Oppositely Charged Membranes

The repulsion between two apposed bilayers can also be overcome electrostatically by using oppositely charged membranes. Indeed, oppositely charged giant unilamellar vesicles (GUVs) spontaneously adhere onto each other and subsequently may undergo hemifusion or full fusion as evidenced by fluorescence video microscopy (Pantazatos, Pantazatos and MacDonald, 2003). Owing to the mutual neutralization of oppositely charged lipids, the two apposed membranes condense

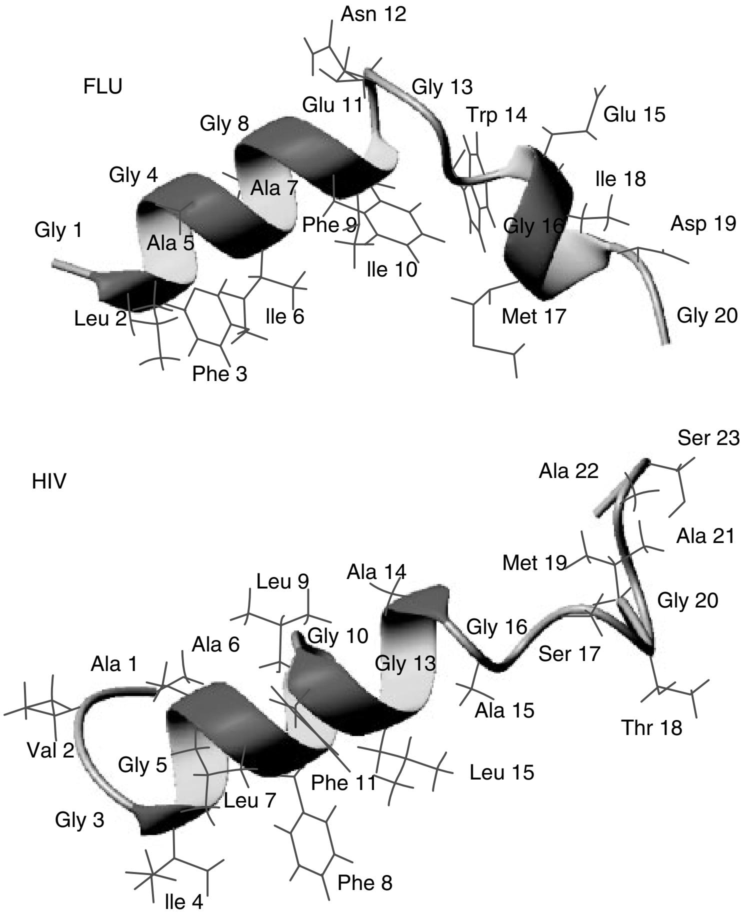
FLU
Asn 12
Gly 13
Glu 11
Gly 8
Trp 14
Glu 15
Gly 4
Ala 7
Ile 18
Gly 1
Ala 5
Phe 9
Gly 16
Asp 19
Ile 10
Leu 2
Ile 6
Met 17
Gly 20
Phe 3
HIV
Ser 23
Ala 22
Ala 21
Met 19
Leu 9
Ala 14
Gly 20
Gly 10
Gly 16
Ala 6
Ala 1
Gly 13
Ser 17
Ala 15
Thr 18
Val 2
Gly 5
Phe 11
Leu 15
Leu 7
Gly 3
Phe 8
Ile 4

Fig. 12.5 The structures of the influenza (FLU) and human immunodeficiency viruses (HIV) fusion peptides as determined by solution NMR in single-chain lipid micelles (dodecylphosphocholine), with permission from Tamm *et al.* (2007).

onto each other which makes the adhesion–condensation mechanism (Kozlov and Markin, 1984) likely to describe the fusion process, similarly to the fusion induced by divalent cations. For highly charged GUVs, membrane rupture and engulfment of smaller by larger vesicles are other possible outcomes. Engulfment becomes especially frequent if GUVs contain some lipids with covalently attached PEG (Lei and Macdonald, 2008). On the other hand, cholesterol stabilizes the hemifused state (García *et al.*, 2001).

Fusion of oppositely charged membranes is unlikely to be directly relevant for living cells because of the absence of cationic lipids. However, cationic lipids are being used in gene therapy as nonviral delivery systems for DNA. They spontaneously form complexes (so-called lipoplexes) with negatively charged DNA that are able to enter their target cells and release the DNA into the cytoplasm. The advantages over viral vectors have stimulated intensive research to improve transfection efficiency. Many different cationic lipids have been synthesized and tested so far but conclusive correlations between biophysical properties of lipoplexes and transfection efficiency is still lacking. However, it has been realized early that the use of fusogenic lipids such as PE as the so-called helper lipid often enhances transfection efficiency (Hope *et al.*, 1998). Two of the crucial steps in the delivery pathway are the entry of the lipoplex into the cytoplasm and the release of the DNA. Some evidence exists of lipoplexes being internalized into cells via the endocytotic pathway. To what extent the escape from the endosome and possibly also the release of DNA from the lipoplexes involve fusion remains to be established (Zuhorn, Engberts, and Hoekstra, 2007).

12.3
Viral Fusion

Viral fusion is the best studied case of exoplasmatic fusion. Relevant with respect to the process of membrane fusion are enveloped viruses. These viruses have their nucleocapsid enclosed in a lipid bilayer membrane that the virus acquires from a host's cell membrane during virus budding. Delivery of the nucleocapsid cargo into the cytoplasm of a new host cell requires fusion of the viral membrane with that of the host cell. An exception is the virus-triggered formation of syncytia (multinuclear cells) that allows the direct passage of viral material from cell to cell.

Nonenveloped viruses do not possess a lipid membrane and thus have to use alternative strategies to translocate their genome into the host cell. These strategies are often not well understood but they include breaking the hydrophobic membrane barrier through the formation of membrane pores–this is the entry path of polioviruses–or directly through membrane rupture. For example, adenoviruses, which have potential applications as viral transport vehicles of genetic material in gene therapy, become internalized into endosomes where they initiate membrane rupture.

Enveloped viruses bind to the target membrane through specific interactions of viral proteins with individual or multiple cell membrane receptors. Nonspecific interactions, especially electrostatic attraction between membrane-associated glycosaminoglycans and ionizable groups of certain viruses, may contribute to the initial recognition of a potential target cell (Sieczkarski and Whittaker, 2004). Membrane receptors include members of the immunoglobulin family, the chemokine receptor family, the low-density lipoprotein receptor family, integrins, or specific carbohydrate groups such as sialic acid. Enveloped viruses have their receptor binding sites located on spike-like glycoproteins. A single virus contains many (often hundreds) of these proteins that can make a multitude of contacts with their corresponding receptors. Thus, even if the binding energy to an individual receptor is weak, the sum of all binding contacts amounts to a large binding energy, much larger than the thermal energy k_BT. It is no surprise that some viruses such as influenza virus and paramyxovirus are in need of special receptor-destroying enzymes (so-called neuroaminidases) that release the virus in case of unproductive binding.

The fusion process of enveloped viruses can take place either directly at the plasma membrane or, for the majority of enveloped viruses, after internalization via the endocytotic pathway (Section 12.3.1). In the former case, direct fusion with the plasma membrane, the viruses need not be exposed to a mildly acidic environment. Instead, fusion proceeds at neutral pH. Examples include most retroviruses, paramyxoviruses, herpesviruses, poxviruses, and coronaviruses. Upon direct entry into the cell through the plasma membrane, the nucleocapsid is released into the cytoplasm from where it is transported to the appropriate transcription or translation sites via the cytoskeletal transport machinery.

12.3.1
Internalization through Endocytosis

Most enveloped viruses become internalized into the cell through endocytosis. These include the orthomyxoviruses, togaviruseses, rhabdoviruses, flaviviruses, and bunyaviruses. Of the two endocytotic trafficking pathways, receptor-mediated endocytosis and bulk-phase endocytosis, the former one appears to be most relevant for virus internalization. To what extent bulk-phase endocytosis–the unspecific mechanism of internalization (pinocytosis)–contributes to viral infections is not well known. Receptor-mediated endocytosis is based on the specific, high-affinity interaction of the virus with cell membrane receptors that trigger endocytosis. The major route of receptor-mediated endocytosis is through clathrin-coated pits. Clathrin is the building block of coated vesicles. It has a tendency to self-assemble on the plasma membrane into a curved cage that invaginates the bilayer and eventually pinches off forming a coated vesicle. A significant protein in this process is dynamin (a guanosine triphosphate (GTP)-binding protein) which may act as a force-generating molecule driving the coated vesicles to pinch-off. Members from the nonenveloped virus families of picornaviruses and polyomaviruses appear not to have to rely on the presence of clathrin. Instead, they are able to interact with lipid rafts or caveolae to become internalized (Sieczkarski and Whittaker, 2004).

Using an endocytotic vesicle as a transport vehicle allows the virus to pass beyond the cytoskeletal cortex (which is typically most dense beneath the cell membrane) and to reach a location closer to the nucleus. Endocytic vesicles are transported to and then fuse with early endosomes. Being located in the endosome, the virus is still separated by the endosomal membrane from the cytosol. Yet, in the endosome, the virus encounters a mildly acidic environment which initiates the fusion reaction between the viral and endosomal membranes.

12.3.2
Viral Fusion Proteins

Enveloped viruses generally possess specialized proteins–type I membrane glycoproteins–that mediate the fusion process and are therefore referred to as *fusion proteins* (Teissier and Pecheur, 2007). Fusion proteins have a large ectodomain in order to interact with the target membrane. They become proteolytically cleaved by proteases of their initial host cell, but the subunits remain associated. Fusion proteins often oligomerize. Despite significant differences between fusion proteins, they all contain a transmembrane domain and a fusion peptide, the latter being a highly conserved short hydrophobic sequence of about 10–12 amino acid residues that is thought to play a central role in the destabilization of the target membrane. Prior to the fusion reaction, the fusion peptides are buried within the aggregates of the fusion proteins. Fusion is initiated by receptor binding (for viruses that enter directly through the plasma membrane) or by the low pH encountered in the endosome (for viruses that are internalized through endocytosis). In both cases, the fusion proteins undergo drastic conformational changes that ultimately lead to the cooperative insertion of several fusion peptides into the target cell.

On a structural level, fusion proteins can be divided into at least two major classes: class I fusion proteins and class II fusion proteins. Class I fusion proteins have large α-helical content and form spike-like trimers that point perpendicularly away from the viral membrane. Retroviruses, paramyxoviruses, filoviruses, and orthomyxoviruses all contain class I fusion proteins. Upon activation, the fusion peptide, located at or close to the N-terminal end of each protein, is released. This process is accompanied by a large conformational change into an extended coiled-coil that contains three α-helices. The following refolding toward a (low-energy) hairpin-like post-fusion conformation of the coiled-coils is thought to lead to the approach and eventually to the merger of the viral and cellular lipid membranes.

Class II fusion proteins–contained in togaviruses, hepadnaviruses, and flaviviruses (Stiasny and Heinz, 2006)–are mostly composed of beta-strands, and the fusion peptide is located in an internal loop rather than at the N terminus. Among the best-characterized examples are the fusion glycoprotein E of the tick-borne encephalitis (TBE) virus and the protein E1 of the Semliki Forest virus (SFV). In the former case, the proteins form heterodimers (together with a viral accessor protein) and in the latter case, homodimers. These prefusion dimers have their long axis oriented roughly parallel to the viral membrane where they contribute to the formation of an ordered icosahedral shell that increases the structural stability of the virus envelope. Fusion is activated by low pH, which induces a major oligomeric rearrangement (with only minor conformational changes) leading to a homotrimeric state for both E and E1. This step is accompanied by the insertion of the fusion peptide into the target membrane and a subsequent folding-back process which leads to the final merger of the two membranes.

Not all known viral fusion proteins belong to either class I or class II (Kielian and Rey, 2006). Fusion proteins of rhabdoviruses and herpesviruses display properties that possibly constitute another class. Structurally, they contain a mixture of α-helical (in particular three-stranded-coiled coils) and beta-sheet secondary structural elements. Their conformational changes induced by low pH in the absence of a target membrane are (in contrast to class I and II) reversible.

12.3.3
Hemagglutinin

The fusion glycoprotein from influenza virus (member of the family of orthomyxoviruses) is HA. Being extensively studied (Bentz and Mittal, 2003), it forms the canonical member of class I fusion proteins. HA is rendered fusion competent after cleavage by cellular proteases into the two subunits HA1 and HA2. HA1 is a globular subunit that contains the sialic acid binding domain and HA2 includes the hydrophobic fusion peptide and the transmembrane domain. In the prefusion state, three HAs form a mushroom-like trimer with the HA1 subunits topping a triple-stranded coiled-coil formed by the HA2 subunits. The fusion peptide is buried within the HA2 trimer. After endocytosis the virus is exposed to a mildly acidic pH in the endosome. This induces a major conformational change. HA1 separates and through a loop-to-helix transition HA2 expels into an extended conformation, terminated by the fusion peptide which inserts into the target membrane (Figure 12.5 for the boomerang-like membrane structure of the fusion peptide). Host and target membrane are now connected through the fully extended HA2 subunits of the fusion proteins. To initiate membrane deformations, a number of HA trimers are thought to arrange into a rosette-like cluster. Then, a folding back of the HA2 subunits through a helix-to-loop transition would plausibly generate a merger of initially the outer monolayers of host and target membrane (possibly through a stalk) which further evolves into a fusion pore. The sequence of these events is illustrated in Figure 12.6 (Acheson, 2007).

12.4
Cellular Fusion

Various biological processes involve cellular fusion, among them vesicular trafficking (intracellular fusion), as well as sexual reproduction and formation of multinucleated cells (cell–cell fusion). Fusion of biological membranes is tightly regulated by a fusion machinery that involves a multitude of proteins. How they act together is the subject of intensive research, especially for the specific case of synaptic vesicle fusion. Principal questions concern the diversity of the fusion pathway: Is there a universal fusion scheme with common intermediates? If so, do the intermediates resemble those of the fusion between bare membranes? And finally, how actively is the lipid membrane involved?

Two principal views on the nature of the fusion process are expressed in *protein-* and *lipid-based* models of fusion pores (Mayer, 2001). The former consists of transmembrane proteins that line up a fusion pore. This is envisioned to be similar to a gap junction where two hemichannels (connexons) combine across the two membranes into a pore (Lindau and Almers, 1995). Protein-based fusion pores are invoked, for example, to explain ion channel-like dynamics of pore opening and closing. In the other (lipid-based) model, the fusion proteins mainly reduce the large energy barriers that exist for fusion of bare membranes. This is accomplished by closely apposing the membranes, affecting their elastic properties, and locally perturbing the bilayer structure. Here, the energy landscape is structurally conserved compared to bare membranes and thus similar non-bilayer intermediates (such as stalks and hemifusion diaphragms) exist (Chernomordik and Kozlov, 2003).

12.4.1
SNAREs

Intracellular compartments such as the ER, lysosome, and Golgi apparatus participate in vesicular trafficking

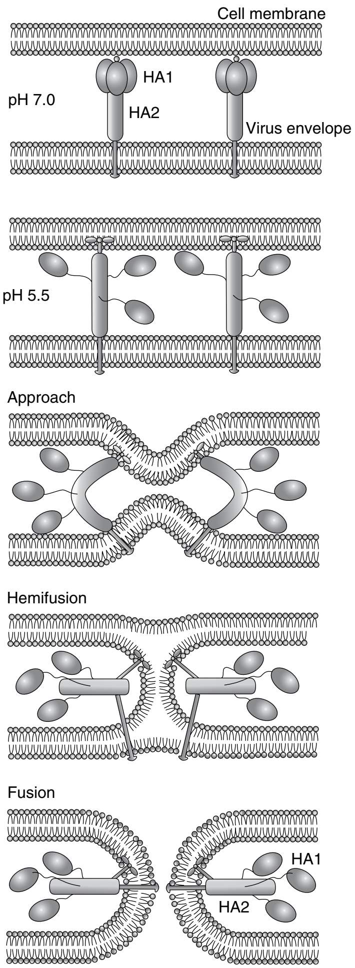

Fig. 12.6 Schematic illustration of fusion between influenza virus and endosomal membrane mediated by hemagglutinin, with permission from Acheson (2007).

and thus engage in membrane fusion events. Fusion also occurs in exocytosis, which is the final step in the secretory pathway and refers to the fusion of trafficking vesicles with the plasma membrane. We distinguish between constitutive (not triggered by Ca^{2+}) and Ca^{2+}-regulated exocytosis (Jackson and Chapman, 2006). The latter includes fusion of neurotransmitter-filled presynaptic vesicles with the plasma membrane where "full-collapse" and "kiss-and-run" fusion constitute two different modes of a vesicle recycling process (Harata, Aravanis and Tsien, 2006). Both secretory pathways involve a similar fusion machinery centered about SNARE (soluble *N*-ethylmaleimide-sensitive factor attachment protein receptor) complexes, which are widely considered to be crucial for the fusion between transport vesicle and plasma membrane. Specific sets of SNAREs exist for each fusion pathway and specificity is ensured by the recruitment of cognate SNAREs to the corresponding fusion sites (Bethani *et al.*, 2007).

Synaptic SNARE complexes contain a highly conserved core that consists of a parallel four-helix bundle originating from three different SNAREs: syntaxin (which is associated with the plasma membrane), two helices from SNAP-25 (synaptosome-associated protein of 25 kDa), and synaptobrevin (also referred to as *vesicle-associated membrane protein* or *VAMP*). The coiled-coil four-helix bundle contains conserved heptad repeats and is highly stable. A model of a SNARE complex (and synaptotagmin) is shown in Figure 12.7. Interestingly, each helix contributes a highly conserved ionic amino acid side chain to the hydrophobic center of the helix bundle; one arginine (R) and three glutamine (Q). The structural conservation of a 1 : 3 ratio between R and Q has lead to the new nomenclature of Q-SNAREs and R-SNAREs, replacing the original t-SNAREs (t-target) and v-SNAREs (v-vesicular).

The emphasis on SNARE complexes as a "minimal fusion machinery" originates to some extent from their ability to induce fusion *in vitro* when the corresponding SNAREs are reconstituted into separate liposomes (Weber *et al.*, 1998). Yet, fusion under these conditions is infrequent and slow–on a timescale of minutes as compared to milliseconds for synaptic fusion. A suggested mechanism for the docking of vesicles involves the "zippering" of SNARE complexes. Beginning at the membrane-distal N-terminal ends and proceeding toward the membrane-proximal C-termini, the membranes are pulled together and thus brought to close apposition. It is likely that the energy released during the formation of these so-called SNAREpins is used to produce fusion intermediates. Interestingly, the propensity to fuse is completely abolished if the two transmembrane domains of the SNARE complex (from syntaxin and synaptobrevin) are replaced by lipid anchors (McNew *et al.*, 2000). Vesicles are still being docked onto each other but fusion appears to require more than merely close contact. Transmembrane segments of fusion proteins appear to generally participate in the fusion reaction, perhaps by perturbing

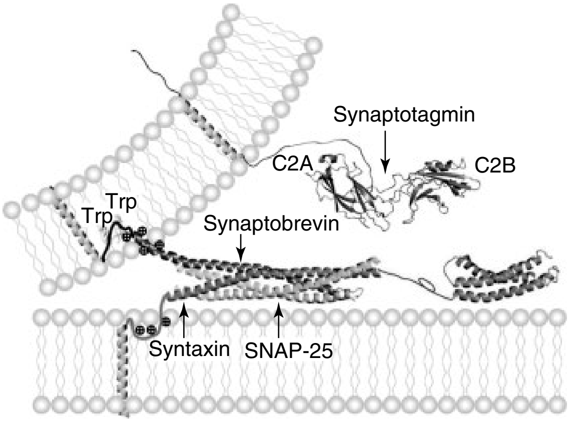

Fig. 12.7 A Model of a SNARE complex and synaptotagmin; with permission from Jackson and Chapman, 2006.

the bilayer, inducing membrane tension or fencing off the hemifusion structure (Langosch, Hofmann and Ungermann, 2007).

Most studies of reconstituted synaptic SNARE-mediated fusion systems confirm the need to add a calcium sensor in order to accelerate fusion (Stein *et al.*, 2007). Present evidence points at an ability of the calcium sensor synaptotagmin to promote curvature in the target membrane (Martens, Kozlov and McMahon, 2007). Indeed, synaptotagmin interacts in the presence of Ca^{2+} with anionic phospholipids through its cytoplasmic C2 tandem domain (C2A and C2B; Figure 12.7) thereby inserting four hydrophobic loops partially into the lipid bilayer. Such insertions generally affect the spontaneous curvature and thus possibly reduce the energy barrier of the fusion reaction (Zimmerberg and Kozlov, 2006). Despite recent progress, to date no fusion system has been reconstituted that fully reproduces the dynamics of synaptic vesicle fusion. How exactly SNAREs and interacting proteins fuse membranes remains to be established.

12.4.2
Cell–Cell Fusion

Cell–cell fusion (Chen *et al.*, 2007) begins with sperm–egg fusion during mammalian fertilization and continues during development. Here, it occurs during skeletal muscle differentiation when many mononucleate myoblasts fuse into multinucleate muscle fibers. Multinucleated cells can also form upon fusion of macrophages into osteoclasts–bone-resorbing cells–and giant cells. Finally, cell–cell fusion contributes to the differentiation of the placenta where multinucleated cells (syncytiotrophoblasts) form a thick layer that serves as barrier between maternal and fetal blood vessels. Cells fuse not only during development but also as a result of disease. This is evidenced by the fusion of somatic cells induced through viruses. Generally, the molecular events that underlie cell–cell fusion remain largely unknown.

A prototype for studying developmental fusion is the nematode Caenorhabditis *elegans*. Syncytial cells in this roundworm form at all developmental stages. Most are binucleate but the epidermal giant cells grow to a total of 138 nuclei. So far, two single-pass transmembrane fusion proteins have been identified, *epithelial fusion failure 1* (EFF-1) and *anchor-cell fusion failure 1* (AFF-1), which independently regulate the fusion of different cell types (Oren-Suissa and Podbilewicz, 2007). The gene encoding for EFF-1 also encodes for a hydrophobic peptide that seems to localize EFF-1 to sites of cell–cell contact prior to fusion (del Campo *et al.*, 2005). The two fusion proteins constitute the founding members of the family of developmental fusogens. The proteins contain large ectodomains with 16 conserved cysteins which perhaps suggest the presence of a ‘‘zippering’’ mechanism, although different from that leading to SNARE complexes. As opposed to the cognate pairing of SNAREs, EFF-1 and AFF-1 require a homotypic arrangement (i.e. the same protein needs to be present on both membranes). Even though the molecular mechanism of EFF-1 and AFF-1-mediated fusion is unknown, it likely involves a hemifusion intermediate.

Sperm–egg fusion in mammals is preceded by a number of prefusion priming events. One is the fusion of the acrosome’s outer membrane with the sperm’s plasma membrane. This allows the enzyme-rich content of the acrosome–a cap-like organelle between the sperm's plasma membrane and nucleus–to secrete. This Ca^{2+}-regulated exocytosis involves a variant of the SNARE-based fusion machinery (Mayorga, Tomes and Belmonte, 2007). Only after subsequent penetration of the sperm into the zona pellucida–a glycoprotein membrane surrounding the plasma membrane of an oocyte–can fusion occur between the plasma membranes of sperm and egg. To date, some of the players of sperm–egg fusion have been identified; others previously hypothesized were shown to be dispensable (Rubinstein *et al.*, 2006).

Myoblast fusion leads to skeletal muscle fibers; cells in these fibers have hundreds of nuclei. Fusion proceeds from "founder cells" that fuse with fusion-competent cells. A number of proteins have been identified that are involved in myoblast fusion. Yet, none of them contains familiar fusion motifs such as fusion peptides or coiled-coil helix bundles, suggesting that (as for EFF-1 and AFF-1) the fusion mechanism is different.

Placental fusion (i.e. the fusion of placenta trophoblasts into syncytiotrophoblasts) is interesting from an evolutionary perspective: for higher primates and some rodents, it requires the protein syncytin which is related to the envelope proteins of endogenous retroviruses (ERVs). This suggests that a captured (class I) fusion protein catalyzes this fusion reaction. In fact, there is evidence that the retroviral genes were integrated and adapted for placenta trophoblasts fusion twice and independently in higher primates and rodents (Oren-Suissa and Podbilewicz, 2007).

12.4.3
Fusion in Mitochondria

Mitochondria are double-membrane intracellular organelles that besides providing the cell with adenosine triphosphate fulfill a large number of additional functions. Originating from a bacterial ancestor, they have evolved into semiautonomous entities with variable shape and morphology. Even more, mitochondria constantly redefine their identity by undergoing fusion and fission. Suppression of fusion leads to defragmentation, suppression of fission to elongation. The balance of these two antagonistic processes provides a dynamic equilibrium that appears to be important in a yet not well-understood manner. Fusion and fission allow mitochondria to communicate with each other, and both are needed to maintain proper respiratory function. A possible advantage of the continuous remodeling is greater efficiency of inheritance and intracellular distribution of mitochondrial DNA (Hoppins, Lackner and Nunnari, 2007). In addition, mitochondrial fusion and fission appear to correlate with the protection from and promotion of apaptosis.

There are distinct machineries for fusion and fission in mitochondria. Fusion of the mitochondia's outer membranes in mammalian cells depends on the GTPases mitofusins, Mfn1 and Mfn2. (Similar proteins exist in other organisms.) They are thought to mediate the tethering of the outer membranes based on a homotypic arrangement (i.e. their presence on both outer membranes) and interaction through a coiled-coil of their heptad repeat regions (Chan, 2006). The subsequent events that induce fusion have not yet been identified. Note that fusion of outer and inner membranes is correlated, but the inner membranes have a separate fusion machinery.

12.5
Modeling Fusion

Experimental data inspire models that feed back into the design of experiments. Of course, this general scheme applies to the fusion of lipid membranes as well. Modeling is therefore not only a supplement but also an integral part in our process of understanding membrane fusion. This chapter treats *modeling* and *simulating* fusion as separate sections because computer simulations are sometimes considered as experiments *in silico*. Yet, the separation between the two is not well defined because computer simulations often employ a set of assumptions comparable to phenomenological models and, conversely, phenomenological models can often only be evaluated using powerful computers.

Lipid membranes are self-assembled materials that interact with each other and resist mechanical forces to some degree. If subject to stress, membranes may respond in numerous ways, depending on the energy landscape and the ability to explore this landscape through fluctuations. Phenomenological models of membrane fusion drastically reduce the dimensionality of this energy landscape by considering only a small number of degrees of freedom, the so-called order parameters. This kind of modeling can be very valuable because it potentially identifies the most relevant degrees of freedom and underlying mechanisms. For example, fusion intermediates were first suggested and analyzed based on phenomenological modeling. Some of these models are certainly too simple but their conceptual value is highly significant.

12.5.1
Forces between Membranes

Before membrane fusion can be initiated, the two membranes must approach each other. Only uncharged (or very weakly charged) membranes have a small equilibrium distance of a few (2–3) nanometers (Rand and Parsegian, 1989). Charged membranes–including biological membranes–usually repel each other even at larger distances. A short account of the different forces that act between two membranes includes the following.

12.5.1.1 Electrostatic Forces

Upon approach of two charged membranes in aqueous solution, their respective diffuse layers of counter- and co-ions are perturbed. A force arises that is usually repulsive if both membranes are anionic (for cationic membranes, see Section 12.2.6). A useful model to describe this force is provided by Poisson–Boltzmann theory which assumes point-like ions and neglects all correlations (Andelman, 1995). Poisson–Boltzmann theory always predicts repulsion between two negatively charged membranes. However, if one of its assumptions is not fulfilled, Poisson–Boltzmann theory fails and attractive interactions can occur. This is the case if the membranes are highly charged or if the ions in solution are either multivalent or have an internal structure. Predicting the force is then not a simple matter.

12.5.1.2 Van der Waals Forces

Induced dipole–dipole interactions lead to an (usually) attractive potential $\sim r^{-6}$

which can be integrated over two membranes. This gives rise to an attractive force per unit area $-H/(6\pi r^3)$ for membrane-to-membrane distances r much smaller than the membrane thickness d and $-2Hd^2/(\pi r^5)$ for $r \gg d$. Here, $H \approx k_BT$ is the Hamaker constant. In colloid science, van der Waals and electrostatic forces together constitute DLVO (Derjaguin–Landau–Verwey–Overbeek) theory.

12.5.1.3 Undulation Forces

Bending fluctuations of two fluid membranes (characterized by the bending stiffness κ; see Eq. (1)) are suppressed when they approach each other, leading to a repulsive force per unit area of magnitude $3\pi^2(k_BT)^2/(64\kappa r^3)$ as a function of the membrane-to-membrane distance r.

12.5.1.4 Hydration Forces

Structural ordering of water molecules occurs in lipid membranes close to the apolar/polar interface. For two identical membranes, interference of the ordered layers leads to a short-range repulsive force per unit area of magnitude $C\exp(r/\lambda)$ with $C = 10^6–10^8\text{N m}^{-2}$ and $\lambda = (0.15–0.3)$ nm. This force dominates at membrane-to-membrane distances r smaller than about 1–3 nm. Note that the hydration force can also be attractive for surfaces that induce different directionality of the water ordering (Rand *et al.*, 1988).

12.5.1.5 Polymer-Induced Forces

The presence of water-soluble polymers can modify the interaction forces between two membranes in several ways (Netz and Andelman, 2003). Membrane-anchored polymers will contribute to the steric repulsion force between two membranes due to their conformational entropy loss. If free polymers are present in the aqueous solution, there will be an attractive depletion force. Finally, membrane-adsorbing polymers may give rise to a bridging interaction that can be attractive or favor an equilibrium distance.

The repulsion between bare (protein-free) membranes can be overcome by using divalent cations (Section 12.2.2), polymers such as PEG (Section 12.2.3), or oppositely charged membranes (Section 12.2.6). Initiating contact between biological membranes is accomplished by proteins. These proteins tether membranes together; as, for example, SNAREs accomplish through formation of four-helix bundles (Section 12.4.1), viral fusion proteins achieve via receptor-specific binding (Section 12.3.2), or mitofusins are thought to do for mitochondrial membranes (Section 12.4.3).

12.5.2
Membrane Elasticity

Most phenomenological modeling efforts of fusion intermediates are based on the elastic behavior of lipid layers. A number of deformation modes are commonly discussed; they include stretching, curvature, and tilt; as illustrated in Figure 12.8. In a lateral stretching deformation the average cross-sectional area a per lipid deviates from its equilibrium value $a_0 \approx 0.6\ \text{nm}^2$. The corresponding free energy, measured per unit area of the lipid layer, can be written as $(K/2)(a/a_0 - 1)^2$ where $K \approx 20\ k_BT/\text{nm}^2$ is the lateral compressibility modulus. A second mode corresponds to a bending deformation which can be described in terms of the two principal curvatures c_1

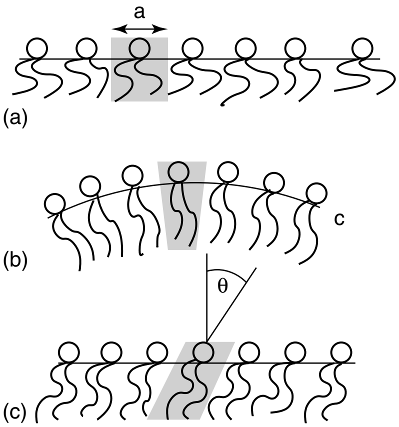

Fig. 12.8 Schematic illustration of three elastic deformation modes of a fluid lipid layer: (a) stretching, (b) bending, and (c) tilt. Here, *a* is the lateral cross-sectional area per lipid, *c* is the curvature of the lipid layer, and θ is the tilt angle of the lipid director.

and c_2 at any given location within the lipid layer. The corresponding bending energy F, measured per unit area A, can be written as (Helfrich, 1973)

$$\frac{F}{A} = \frac{\kappa}{2}\left(c_1 + c_2 - c_0\right)^2 + \bar{\kappa} c_1 c_2 \qquad (1)$$

Here, $\kappa \approx 10\, k_B T$ and $\bar{\kappa}$ are the bending stiffness and Gaussian modulus, respectively. The bending stiffness relates to changes in the mean curvature $(c_1 + c_2)/2$ or, equivalently, to a *splay* deformation at the level of individual lipids. The Gaussian modulus $\bar{\kappa}$ accounts for the resistance with respect to changing the Gaussian curvature $c_1 c_2$ or, at the level of individual lipids, for a *saddle–splay* deformation. Consequently, κ and $\bar{\kappa}$ can alternatively be referred to as *splay modulus* and *saddle-splay modulus*. The intrinsic tendency of a lipid layer to adopt a preferred curvature is expressed by the spontaneous curvature c_0 in Eq. (1). Note that F in Eq. (1) is invariant with respect to $c_1 \rightarrow c_2$ and $c_2 \rightarrow c_1$ which accounts for lateral fluidity. Moreover, the curvatures c_1 and c_2 must be measured at the so-called neutral surface where stretching and curvature deformations of the lipid layer decouple.

Another elastic mode is associated with a tilt deformation. Here, the average direction of the lipid molecules deviates from the normal direction of the lipid layer. Denoting the angle between the two by θ, the free energy of a tilt deformation, measured per unit area, can be written as $(\kappa_t/2)\theta^2$ where $\kappa_t \approx 20\, k_B T/\text{nm}^2$ is the tilt modulus (May *et al.*, 2004).

The different deformation modes can be expressed together in terms of a vector field $h\mathbf{n}$ where h is the local effective length of the lipids and $\mathbf{n}$ is a field of unit vectors, describing the directions into which the lipid tails point on average. Note that $\mathbf{n}$ is referred to as a *director* field in analogy to the theory of liquid crystals. The following part of the free energy per unit area

$$\frac{\kappa}{2}(\nabla \cdot \mathbf{n} - c_0)^2 + \frac{\kappa_t}{2}(\mathbf{n} - \nabla h)^2 \qquad (2)$$

(where ∇ denotes the nabla operator) then shows that splay can be produced from both curvature $\nabla^2 h$ at constant tilt and changing tilt at constant curvature (Hamm and Kozlov, 2000; Fournier, 1998). General free energy expressions that are based on Eq. (2) can be used to analyze elastic membrane deformations, including fusion intermediates.

12.5.3
Membrane Pores

The formation of pores often appears as alternative outcome in membrane fusion assays, signified by content leakage of vesicles. Pore formation in a bare membrane can be induced by lateral tension and, similar to fusion, requires an activation energy. This is seen most simply in a model inspired by classical nucleation theory (Litster, 1975). Here, a circular pore of radius r and line tension of the pore rim λ in a membrane that is subject to lateral tension Λ has the free energy

$$\lambda 2\pi r - \Lambda \pi r^2 \tag{3}$$

which adopts a minimum at $r = 0$ and a maximum at $r = \lambda/\Lambda$. Unbounded pore expansion thus occurs for an activation energy larger than $\pi\lambda^2/\Lambda$. The lateral tension Λ can generally have electrical or mechanical origin. The former is the basis for electroporation (Chang *et al.*, 1992), whereas the latter is thought to play an important role in the rupture of the hemifusion diaphragm (Figure 12.2d and e). The line tension λ–the energy needed to create a membrane rim of unit length–depends on the molecular structure of the involved lipids and is typically found to be in the range of $\lambda = (2.5–7.5)\ k_B T/\text{nm}$. The exposure of hydrocarbon chains to the aqueous environment would predict a much larger value, $\lambda = \gamma d = 36\ k_B T/\text{nm}$, based on a hydrocarbon water surface tension $\gamma = 12\ k_B T/\text{nm}$ and a hydrophobic membrane thickness of $d = 3\text{nm}$. The much lower value of λ found experimentally indicates a structural rearrangement of the rim-forming lipids. That is, some lipids reorient themselves in order to use their polar headgroups to protect the hydrocarbon chains from exposure to water as illustrated in Figure 12.9. The assumptions of cylindrical pores and constant line tension λ (Eq. (3)) need not be fulfilled in real systems. Small line tension will lead to significant fluctuations in pore shape or even in the spontaneous formation of transient pores. Also, the line tension λ is generally not a constant, especially not for small pore sizes where rim–rim interactions become relevant.

12.5.4
Modeling Fusion Intermediates

A stalk is a lipid connection between the two proximal monolayers of two lipid membranes, illustrated in Figure 12.2b. Its existence during the transition from the lamellar-to-inverted nonlamellar phases has been shown experimentally (discussed in Section 12.2.1), and some indirect evidence suggests its presence in the fusion of bare and biological membranes as well.

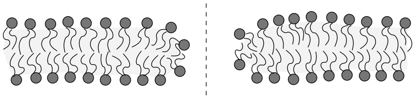

Fig. 12.9 Schematic illustration of a cross-section through a membrane pore. The broken line indicates the axis of cylindrical symmetry.

Since its postulation (Kozlov and Markin, 1983), the stalk model has been revised several times. The different generations of stalk models have been discussed recently (Siegel, 2004). The original stalk model was based entirely on the bending energy of the individual monolayers in each of the two fusing membranes; see Eq. (1). It made a number of predictions that agree with experimental observations, including the dependence of hemifusion rates on membrane composition. Specifically, the addition of lipids with $c_0 < 0$ to the proximal monolayers of two apposed membranes is predicted to render stalk formation more favorable. This is indeed observed experimentally as discussed in Section 12.2.4.

A different possibility to increase fusion rates is by adding alkanes (Walter, Yeagle and Siegel, 1994). The mechanism is likely to be the same as for the well-known stabilization of the inverse hexagonal phase by alkanes. In both cases, the packing energy of the hydrocarbon tails is reduced by allowing alkanes to partition into the interstitial region–the contact region of three monolayers. The original stalk model was not able to account for this observation which motivated Siegel to add to the bending energy a chain packing contribution (Siegel, 1993; Siegel, 1999). A problem that emerged in calculating the optimal energy of the modified stalk model was its high energy, generally exceeding 100 k_BT.

The problem of the large stalk energy was solved by allowing for edges in the shape of the stalk (Kozlovsky and Kozlov, 2002). The existence of edges is intimately related to nonvanishing tilt of the lipid chains. Tilt was introduced in Section 12.5.2 as an elastic deformation mode of a lipid layer. Appropriate tilt deformations completely eliminate the interstitial energy. As a result, the total stalk energy is contained in its elastic deformation energy. Yet, even when accounting for tilt, the energy remains large for a smooth stalk. What reduces the stalk energy to values below 40 k_BT is the existence of edges both in the waist of the stalk and at the dimples of the distal monolayers. An example of a calculation that compares a stalk with and without edges is shown in Figure 12.10.

According to the stalk-pore model, an initial stalk can radially expand into a hemifusion diaphragm; see Figure 12.2. This process is predicted to occur spontaneously if the lipid layer has a sufficiently large negative spontaneous curvature. In this case, the rim of the hemifusion diaphragm creates a lateral tension in the enclosed bilayer that may be sufficiently large to lead to membrane rupture and thus formation of a fusion pore (Kozlovsky, Chernomordik and Kozlov, 2002). Yet, this is not the only possible scenario. Different pathways have been suggested based not only on computer simulations (discussed in Section 12.6.3) but also on membrane elasticity. For example, the formation of multiple stalks (somewhat similar to Figure 12.3b) could lead to pore formation through stalk coalescence (Lukatsky and Frenkel, 2004).

12.6
Simulating Fusion

Computer simulations generally constitute a rapidly growing, diverse set of methods, ranging from atomistic to solvent-free approaches. Among those,

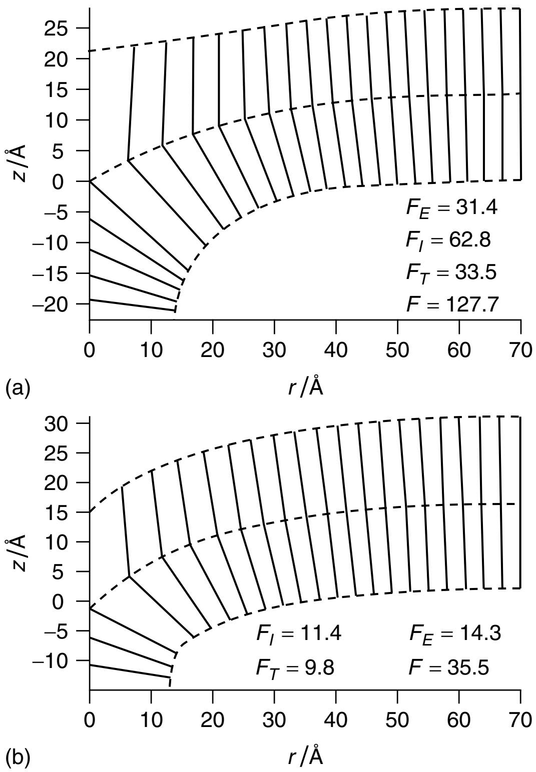

Fig. 12.10 Stalk structures on the basis of membrane elasticity theory, with permission from May (2002). Both stalks were calculated for the same system but the upper one was constrained to be smooth everywhere. The broken lines mark the polar–apolar interfaces and the bilayer midplane. The solid lines denote local lipid directors. For each structure, the energies of the distal monolayers (F_E), proximal monolayers (F_I), and the remaining lipid tether (F_T) are indicated. The total stalk energy is $F = F_E + F_I + F_T$.

atomistic MD simulations represent a method based on *first principles*. This most closely corresponds to actually producing the events of an experiment in silico. Accessible time and length scales are still restricted; atomistic simulations of fusion events are just starting to become feasible. Coarse-grained simulations allow to simulate larger systems. The term *coarse-graining* generally refers to ignoring certain degrees of freedom. This can range from joining groups of atoms (as in most coarse-grained MD simulations) to neglecting the entire solvent (as in solvent-free simulations). Which degrees of freedom can be ignored is often nontrivial, especially for biological systems that have an intrinsic multiscale character. This includes biological membrane fusion where, on the one hand, hundreds of lipids are involved and, on the other hand, a single point mutation of the fusion proteins may completely abolish fusion competence. Multiscale simulation methods may eventually provide means to self-adjust the coarse-grainedness of a simulation in space and time.

12.6.1
Molecular Dynamics Simulations

Atomistic MD simulations provide the most detailed information of structure and motion of lipid membranes. All atoms in a given system are accounted for explicitly as classical particles, and all interactions–bonded and nonbonded–are represented by suitable and sufficiently simple potentials. At each time, the forces acting on all atoms can be computed which allows to numerically solve Newton's equations of motion. The small time step needed for this integration ($\approx$ 2 fs) limits the total simulated time of a system with typical linear extension of 20 nm to the order of 10 ns. This is much less than the sub-millisecond timescale of the fastest known biological fusion (synaptic vesicle fusion). Therefore, almost all attempts to model membrane fusion on the basis of MD employ coarse-grained models. Still, the feasibility of atomistic MD simulations to produce fusion on the nanosecond timescale from a pre-formed stalk was recently demonstrated (Knecht and Marrink, 2007).

In coarse-grained MD simulations a group of atoms are represented by a single interaction site. Coarse-grained lipids contain only a few different segment types (such as polar, nonpolar, apolar, and charged (Marrink *et al.*, 2007)) which greatly reduces the number of required interaction parameters. The design of coarse-grained simulations is guided by matching their predictions with data from atomistic simulations and experiments. Depending on the model, coarse-grained MD simulations sample systems with larger spatial ($\approx$50 nm) and temporal extension (μs). This renders entire fusion paths accessible. By and large, available simulation results are consistent with the stalk-pore hypothesis. That is, a stalk initially connects the proximal leafs of two apposed membranes and then transforms into a transmembrane contact. Upon further radial expansion, a metastable hemifusion diaphragm appears that subsequently ruptures and thus gives rise to a fusion pore. Also in agreement with predictions from continuum stalk models, lipids rearrange through tilt and splay to prevent formation of a large void space in the interstitial regions of the stalk, and nonlammellar lipids strongly affect fusion activation energies (Marrink and Mark, 2003; Stevens, Hoh and Woolf, 2003; Smeijers *et al.*, 2006; Kasson and Pande, 2007). Figure 12.11 displays an example of a specific coarse-grained MD simulation where a small vesicle fuses with a planar bilayer (Smeijers *et al.*, 2006). Shown is a sequence of events characteristic for the stalk-pore model. An interesting aspect of this particular simulation is the transient formation of a hydrophobic void (Figure 12.11b).

Extensive sampling of many MD trajectories for the fusion of small vesicles suggests the existence of two distinct pathways: opening of the fusion pore directly from the "transmembrane contact"-intermediate or only after the formation of an extended hemifusion diaphragm (Kasson *et al.*, 2007).

12.6.2
Dissipative Particle Dynamics

Dissipative particle dynamics (DPD) is a particle-based simulation method that preserves correct hydrodynamic behavior at mesoscopic length and time scales. Conservative, random, and dissipative forces act on beads which represent fluid droplets–clusters of molecules–with all internal degrees of freedom integrated out. The conservative forces can be used to construct connectivities that represent complex molecules or molecular aggregates, such as polymers or lipids. Soft interaction potentials allow for large timesteps in integrating Newton's equations of motion which adopt

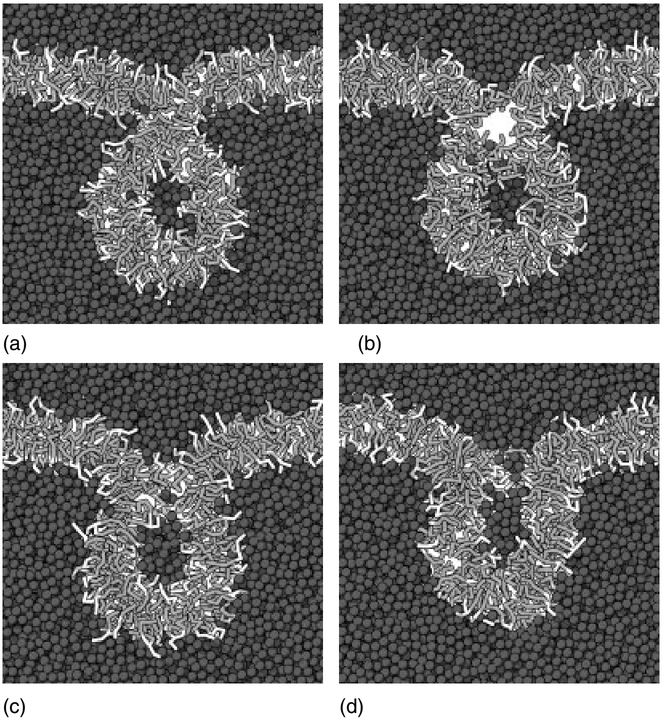

Fig. 12.11 Example of a coarse-grained MD simulation result from Smeijers *et al.* (2006), with permission. Shown is the fusion of a small vesicle with a planar bilayer. The snapshots refer to (a) stalk at 15.4 ns, (b) radial expansion of stalk and formation of a hydrophobic void 15.6 ns, (c) a hemifusion diaphragm at 15.7 ns, (d) fusion pore at 16.5 ns.

the form of stochastic differential equations. Application of the method allows to simulate lipid vesicles that are considerably larger in diameter ($\approx$30 nm) compared to those in coarse-grained MD simulations ($\approx$15 nm). Recent studies have focused on how lateral tension in a planar lipid membrane promotes the fusion with a vesicle in close proximity (Shillcock and Lipowsky, 2005). Fusion can be favorable in this case as it relieves the tension without forming an energetically costly pore. Yet, reasonable increase of fusion probability requires large lateral tension, making membrane rupture a frequent event.

12.6.3
Solvent-Free Simulations

The first solvent-free computer simulation of vesicle fusion employed Brownian dynamics. This method too is based on a stochastic differential equation–the Langevin equation–but does not reproduce hydrodynamic behavior. Fusion of two vesicles with a total of 1000 rod-shaped lipids predicts a fusion pathway quite distinct from the stalk-pore model. Here, a stalk forms and expands asymmetrically into a bent structure which leads to pore opening (Noguchi and Takasu, 2001). A related scenario is predicted on the basis of Monte Carlo simulations using a lattice description for single chain amphiphiles that are not far from the $L_\alpha \rightarrow H_{II}$ transition. Two closely apposed planar bilayers under tension initially develop a stalk that induces multiple pore formation close to the stalk in both bilayers. The synchronous appearance of two pores in direct apposition induces a zipper-like propagation of the stalk along the hole's edges, resulting in a fusion pore (Müller, Katsov and Schick, 2002). These two asymmetric stalk propagation models receive some support from coarse-grained MD simulations (Smeijers *et al.*, 2006).

12.7
Conclusions

Studying membrane fusion is an interdisciplinary process in which physics, chemistry, and biology meet each other. None of these fields alone provides sufficient tools to fully understand how fluid-like, highly dynamic bilayers are brought together and then merge by a specialized protein machinery. One of the difficulties in probing the fusion process is its spatial and temporal confinement: only a small fraction of all available lipids and proteins carry out the fusion act. This requires special techniques to selectively probe these molecules. Another challenge is the intrinsic disorder of lipid membranes which renders entropy a major contribution to the free energy. Indeed, measurements often suggest an ensemble of possible outcomes and computer simulations are just beginning to capture this variability.

Despite the large body of experimental data, there is as yet no rigorous molecular understanding of membrane fusion, neither for bare nor biological membranes. Experiments support a number of models for membrane fusion; most of these are discussed in this chapter. One of them, the stalk-pore model, is currently considered the most universal fusion pathway with relevance to certain fusion types of bare as well as biological membranes. The latter includes viral and SNARE-based fusion. Here, fusion appears to follow a lipid-based pathway with similar intermediates as for bare membranes. Little is known on a molecular level about other fusion types, and it remains to be seen if these follow a similar general fusion path.

Acknowledgments

I would like to thank Drs Gene Berry and Dave Siegel for their helpful comments. The author acknowledges support during the preparation of this chapter from NIH through R15-GM077184 and NSF through DMR-0605883.

Glossary

Coiled-coil: structural motif in proteins where two or more α-helical subunits are wrapped around each other forming a long extended rod-like segment.

Ectodomain: the part of a membrane protein that extends into the extracellular space.

Endoplasmatic fusion: fusion events that occur within a cell. Examples include fusion events during vesicular membrane transport between organelles, in mitochondria, in the Golgi apparatus, and in other inner organelles.

Exocytosis: fusion of the plasma membrane with secretory vesicles, which constitutes the final step in the secretory pathway.

Exoplasmatic fusion: fusion that involves the merger of plasma membranes.

Hemagglutinin: a glycoprotein from influenza (and other) viruses, which mediates binding and fusion with a target cell; see Section 12.3.3

Hydration force: a structural force between two surfaces immersed in aqueous solution due to the ordering of water molecules.

Lipids: large class of hydrophobic, fat-soluble molecules including waxes, oils, phospholipids, and steroids. Membrane-forming lipids are amphiphilic; they combine hydrophilic and hydrophobic properties.

Lipid polymorphism: amphiphilic molecules such as lipids are able to self-assemble into various aggregation geometries

Stalk: lipidic connection between the proximal leafs of two closely apposed lipid bilayers; see Section 12.2.1.

Syncytia: multinuclear cells that form through cell–cell fusion.

SNARE: soluble *N*-ethylmaleimide-sensitive factor attachment protein receptor. A protein family involved in the fusion of intracellular transport vesicles.

Spontaneous curvature c_0**:** denotes the propensity of a lipid layer to adopt curved geometries; see Eq. (1)

Undulation force: repulsive force between two apposed membranes, caused by steric repulsion due to bending fluctuations.

Vesicle: aqueous compartment enclosed by one (unilamellar) or several (multilamellar) lipid bilayers.

References

Afonin, S., Dürr, U.H.N., Glaser, R.W., and Ulrich, A.S. (2004) *Magn. Reson. Chem.*, **42**(2), 195–203.

Andelman, D. (1995) *Structure and Dynamics of Membranes*, (eds R. Lipowsky and E. Sackmann), Vol. 1, Chapter 12, Elsevier, pp. 603–42.

Arnold, K. (1995) *Structure and Dynamics of Membranes*, (eds. R. Lipowsky and E. Sackmann), Vol. 1, Chapter 19, Elsevier, pp. 903–57.

Bechinger, B. and Lohner, K. (2006) *Biochim. Biophys. Acta*, **1758**(9), 1529–39.

Bentz, J. and Mittal, A. (2003) *Biochim. Biophys. Acta*, **1614**(1), 24–35.

Bethani, I., Lang, T., Geumann, U., Sieber, J.J., Jahn, R., and Rizzoli, S.O. (2007) *EMBO J.*, **26**(17), 3981–92.

Chan, D.C. (2006) *Annu. Rev. Cell Dev. Biol.*, **22**, 79–99.

Chang, D.C., Chassy, B.M., Sounders, J.A., and Sowers, A.E. (eds) (1992) *Guide to Electroporation and Electrofusion*, Academic Press, New York.

Chen, E.H., Grote, E., Mohler, W., and Vignery, A. (2007) *FEBS Lett.*, **581**(11), 2181–93.

Chernomordik, L., Chanturiya, A., Green, J., and Zimmerberg, J. (1995) *Biophys. J.*, **69**, 922–29.

Chernomordik, L.V. and Kozlov, M.M. (2003) *Annu. Rev. Biochem.*, **72**, 175–207.

Chernomordik, L.V., Zimmerberg, J., and Kozlov, M.M. (2006) *J. Cell Biol.*, **175**(2), 201–7.

Cullis, P.R. and de Kruijff, B. (1978) *Biochim. Biophys. Acta*, **513**(1), 31–42.

Day, E.P., Kwok, A.Y., Hark, S.K., Ho, J.T., Vail, W.J., Bentz, J., and Nir, S. (1980) *Proc. Natl. Acad. Sci. U. S. A.*, **77**(7), 4026–29.

del Campo, J.J., Opoku-Serebuoh, E., Isaacson, A.B., Scranton, V.L., Tucker, M., Han, M., and Mohler, W.A. (2005) *Curr. Biol.*, **15**(5), 413–23.

Epand, R.M. (1997) *Curr. Top. Membr.*, **44**, 237–52.

Fournier, J.B. (1998) *Europhys. Lett.*, **43**, 725–30.

García, R.A., Pantazatos, S.P., Pantazatos, D.P., and MacDonald, R.C. (2001) *Biochim. Biophys. Acta*, **1511**(2), 264–70.

Gawrisch, K., Parsegian, V.A., Hajduk, D.A., Tate, M.W., Gruner, S.M., Fuller, N.L., and Rand, R.P. (1992) *Biochemistry*, **31**, 2856–64.

Goñi, F.M. and Alonso, A. (1999) *Prog. Lipid Res.*, **38**, 1–48.

Hamm, M. and Kozlov, M.M. (2000) *Eur. Phys. J. E.*, **3**, 323–35.

Harata, N.C., Aravanis, A.M., and Tsien, R.W. (2006) *J. Neurochem.*, **97**(6), 1546–70.

Helfrich, W. (1973) *Z. Naturforsch.*, **28**, 693–703.

Hope, M.J., Mui, B., Ansell, S., and Ahkong, Q.F. (1998) *Mol. Membr. Biol.*, **15**, 1–14.

Hoppins, S., Lackner, L., and Nunnari, J. (2007) *Annu. Rev. Biochem.*, **76**, 751–80.

Hui, S.W. and Stewart, T.P. (1981) *Nature*, **290**, 427–28.

Israelachvili, J.N. (1992) *Intermolecular and Surface Forces*, 2nd edn, Academic Press.

Jackson, M.B. and Chapman, E.R. (2006) *Annu. Rev. Biophys. Biomol. Struct.*, **35**, 135–60.

Kasson, P.M. and Pande, V.S. (2007) *PLoS Comput. Biol.*, **3**(11), 2228–38.

Kasson, P.M., Zomorodian, A., Park, S., Singhal, N., Guibas, L.J., and Pande, V.S. (2007) *Bioinformatics*, **23**(14), 1753–59.

Kielian, M. and Rey, F.A. (2006) *Nat. Rev. Microbiol.*, **4**(1), 67–76.

Killian, J.A. (1998) *Biophys. Biochim. Acta*, **1376**, 401–16.

Kinnunen, P.K.J. and Holopainen, J.M. (2000) *Biosci. Rep.*, **20**(6), 465–82.

Knecht, V. and Marrink, S.J. (2007) *Biophys. J.*, **92**(12), 4254–61.

Kozlov, M.M. and Markin, V.S. (1983) *Biofizika*, **28**, 255–61.

Kozlov, M.M. and Markin, V.S. (1984) *Gen. Physiol. Biophys.*, **5**, 379–402.

Kozlovsky, Y., Chernomordik, L.V., and Kozlov, M.M. (2002) *Biophys. J.*, **83**(5), 2634–51.

Kozlovsky, Y. and Kozlov, M.M. (2002) *Biophys. J.*, **82**, 882–95.

Langosch, D., Hofmann, M., and Ungermann, C. (2007) *Cell. Mol. Life Sci.*, **64**(7–8), 850–64.

Lee, J. and Lentz, B.R. (1997) *Biochemistry*, **36**(21), 6251–59.

Lei, G. and Macdonald, R.C. (2008) *J. Membr. Biol.*, **221**, 97–106.

Lentz, B.R. (2007) *Eur. Biophys. J.*, **36**(4–5), 315–26.

Lentz, B.R., Malinin, V., Haque, M.E., and Evans, K. (2000) *Curr. Opin. Struct. Biol.*, **10**(5), 607–15.

Lindau, M. and Almers, W. (1995) *Curr. Opin. Cell Biol.*, **7**(4), 509–17.

Litster, J.D. (1975) *Phys. Lett.*, **53**(A), 193–94.

Lukatsky, D.B. and Frenkel, D. (2004) *Eur. Phys. J. E.*, **14**(1), 3–6.

Marrink, S.J. and Mark, A.E. (2003) *J. Am. Chem. Soc.*, **125**(37), 11144–45.

Marrink, S.J. and Mark, A.E. (2004) *Biophys. J.*, **87**(6), 3894–900.

Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P., and de Vries, A.H. (2007) *J. Phys. Chem. B*, **111**(27), 7812–24.

Martens, S., Kozlov, M.M., and McMahon, H.T. (2007) *Science*, **316**(5828), 1205–8.

May, S. (2002) *Biophys. J.*, **83**, 2969–80.

May, S., Kozlovsky, Y., Ben-Shaul, A., and Kozlov, M.M. (2004) *Eur. Phys. J. E.*, **14**, 299–308.

Mayer, A. (2001) *Trends Biochem. Sci.*, **26**(12), 717–23.

Mayorga, L.S., Tomes, C.N., and Belmonte, S.A. (2007) *Iubmb Life*, **59**(4–5), 286–92.

McNew, J.A., Weber, T., Parlati, F., Johnston, R.J., Melia, T.J., Sollner, T.H., and Rothman, J.E. (2000) *J. Cell Biol.*, **150**(1), 105–17.

Müller, M., Katsov, K., and Schick, M. (2002) *J. Chem. Phys.*, **116**, 2342–45.

Netz, R.R. and Andelman, D. (2003) *Phys. Rep.*, **380**, 1–95.

Noguchi, H. and Takasu, M. (2001) *J. Chem. Phys.*, **115**, 9547–51.

Ohki, S. and Arnold, K. (1990) *J. Membr. Biol.*, **114**, 195–203.

Oren-Suissa, M. and Podbilewicz, B. (2007) *Trends Cell Biol.*, **17**(11), 537–46.

Pantazatos, D.P., Pantazatos, S.P., and MacDonald, R.C. (2003) *J. Membr. Biol.*, **194**, 129–39.

Peuvot, J., Schanck, A., Lins, L., and Brasseur, R. (1999) *J. Theor. Biol.*, **198**(2), 173–81.

Rand, R.P., Fuller, N., Parsegian, V.A., and Rau, D.C. (1988) *Biochemistry*, **27**, 7711–22.

Rand, R.P. and Parsegian, V.A. (1989) *Biophys. Biochim. Acta., Biomem. Rev.*, **988**, 351–76.

Rappolt, M., Hickel, A., Bringezu, F., and Lohner, K. (2003) *Biophys. J.*, **84**(5), 3111–22.

Rubinstein, E., Ziyyat, A., Wolf, J.P., Naour, F.Le., and Boucheix, C. (2006) *Semin. Cell Dev. Biol.*, **17**(2), 254–63.

Seddon, J.M., Robins, J., Gulik-Krzywicki, T., and Delacroix, H. (2000) *Phys. Chem. Chem. Phys.*, **2**(20), 4485–93.

Shillcock, J.C. and Lipowsky, R. (2005) *Nat. Mater.*, **4**(3), 225–28.

Sieczkarski, S.B., Whittaker, G.R. (2004) *Membrane Trafficking in Viral Replication, Current Topics in Microbiology and Immunology*, Vol. 285, Springer, pp. 1–23.

Siegel, D.P. (1984) *Biophys. J.*, **45**, 399–420.

Siegel, D.P. (1993) *Biophys. J.*, **65**, 2124–40.

Siegel, D.P. (1999) *Biophys. J.*, **76**, 291–313.

Siegel, D.P. (2004) *The Structure of Biological Membranes*, (ed P. L. Yeagle), Chapter 8, CRC Press, pp. 255–308.

Siegel, D.P. and Epand, R.M. (1997) *Biophys. J.*, **73**, 3089–111.

Smeijers, A.F., Markvoort, A.J., Pieterse, K., and Hilbers, P.A. (2006) *J. Phys. Chem. B*, **110**(26), 13212–19.

Stein, A., Radhakrishnan, A., Riedel, D., Fasshauer, D., and Jahn, R. (2007) *Nat. Struct. Mol. Biol.*, **14**(10), 904–11.

Stevens, M.J., Hoh, J.H., and Woolf, T.B. (2003). *Phys. Rev. Lett.*, **91**(18), 188102.

Stiasny, K. and Heinz, F.X. (2006) *J. Gen. Virol.*, **87**, 2755–66.

Tamm, L.K., Lai, A.L., and Li, Y. (2007) *Biochim. Biophys. Acta*, **1768**(12), 3052–60.

Teissier, E. and Pecheur, E.I. (2007) *Eur. Biophys. J.*, **36**, 887–99.

Toombes, G.E.S., Finnefrock, A.C., Tate, M.W., and Gruner, S.M. (2002) *Biophys. J.*, **82**(5), 2504–10.

Tristram-Nagle, S. and Nagle, J.F. (2007) *Biophys. J.*, **93**(6), 2048–55.

Verkleij, A.J., Momvers, C., Leunissen-Bijvelt, J., and Ververgaert, P.H. (1979) *Nature*, **279**, 162–63.

Walter, A., Yeagle, P.L., and Siegel, D.P. (1994) *Biophys. J.*, **66**, 366–76.

Weber, T., Zemelman, B.V., McNew, J.A., Westermann, B., Gmachl, M., Parlati, F., Sollner, T.H., and Rothman, J.E. (1998) *Cell*, **92**(6), 759–72.

Weinreb, G. and Lentz, B.R. (2007) *Biophys. J.*, **92**(11), 4012–29.

Wilschut, J., Düzgüne, N., and Papahadjopoulos, D. (1981) *Biochemistry*, **20**(11), 3126–33.

Yang, L. and Huang, H.W. (2002) *Science*, **297**(5588), 1877–79.

Zimmerberg, J. and Kozlov, M.M. (2006) *Nat. Rev. Mol. Cell Biol.*, **7**(1), 9–19.

Zuhorn, I.S., Engberts, J.B.F.N., and Hoekstra, D. (2007) *Eur. Biophys. J.*, **36**(4–5), 349–62.

Further Reading

Acheson, N.H. (2007) *Fundamentals of Molecular Virology*, Chapter I.4, John Wiley & Sons.

Blumenthal, R., Clague, M.J., Durell, S.R., and Epand, R.M. (2003) Membrane fusion.*Chem. Rev.*, **103**(1), 53–69.

Chen, E.H. and Olson, E.N. (2005) Unveiling the mechanisms of cell-cell fusion. *Science*, **308**(5720), 369–73.

Earp, L.J., Delos, S.E., Park, H.E., and White, J.M. (2004) The many mechanisms of viral membrane fusion proteins *Membrane Trafficking in Viral Replication, Current Topics in Microbiology and Immunology*, Vol. 285, Springer, pp. 25–66.

Hafez, I.M. and Cullis, P.R. (2001) Roles of lipid polymorphism in intracellular delivery. *Adv. Drug Deliv. Rev.*, **47**(2–3), 139–48.

Jahn, R. and Grubmüller, H. (2002) Membrane fusion. *Curr. Opin. Cell Biol.*, **14**, 488–95.

Sapir, A., Avinoam, O., Podbilewicz, B., and Chernomordik, L.V. (2008) Viral and developmental cell fusion mechanisms: Conservation and divergence. *Dev. Cell*, **14**(1), 11–21.

Shillcock, J.C. and Lipowsky, R. (2006) The computational route from bilayer membranes to vesicle fusion. *J. Phys. Cond. Mat.*, **18**(28), S1191–219.

Siegel, D.P. (2004) Lipid membrane fusion, in *The Structure of Biological Membranes*, (ed P. L. Yeagle), Chapter 8, CRC Press, pp. 255–308.

13

Raman Spectroscopy in Virus Structure Analysis

Daniel Němeček and George J. Thomas, Jr

13.1 Introduction 419
13.2 Assembly Pathways of Representative Viruses 420
13.3 Advantages of Raman Spectroscopy in Virus Research 421
13.4 Experimental Methods for Raman Spectroscopy 423
13.5 Analysis of Raman Spectroscopic Data 424
13.5.1 General Considerations 424
13.5.2 Singular value decomposition 425
13.5.3 Interpretation of Polarized Raman Spectra 426
13.6 Characteristics of Raman Spectra of Viral Proteins and Nucleic Acids 428
13.6.1 The Protein Raman Spectrum 430
13.6.1.1 Raman Markers of Main Chain Conformation: Amide I and Amide III Modes 430
13.6.1.2 Raman Markers of Side Chains 431
13.6.2 The Nucleic Acid Raman Spectrum 433
13.7 Applications 435
13.7.1 Assembly and Maturation of the Bacteriophage P22 Capsid 435
13.7.2 Assembly and Maturation of the Bacteriophage HK97 Head 439
13.7.2.1 Structural Considerations 439
13.7.2.2 Thermodynamic Properties of the Δ-Domain 440
13.7.3 Structural Studies of Filamentous Bacteriophages 443
13.7.3.1 Conformations of Packaged ssDNA 444
13.7.3.2 Structural Details of the Capsid Subunits 446
13.8 Concluding Summary 450
Acknowledgments 452
References 452

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

13.1
Introduction

The vibrational Raman effect, which was predicted on the basis of theoretical considerations by Smekal (1923) and confirmed experimentally by Raman and Krishnan (1928), was first applied to model peptides in the 1950s by Edsall and collaborators (Edsall, Otvos and Rich, 1950; Garfinkel, 1958a, 1958b, 1958c, 1958). Raman spectroscopy was subsequently demonstrated as an effective experimental probe of nucleic acids and native proteins by R. C. Lord and collaborators in the late 1960s and early 1970s (Lord and Thomas, 1967b, 1968; Lord and Yu, 1970a, 1970b). With the advent of the continuous wave laser as an effective excitation source for biomolecular applications, the Raman method soon gained use as a structural probe of supramolecular biological assemblies, including viruses (Hartman, Clayton and Thomas, 1973; Miura and Thomas, 1995b; Nemecek *et al.*, 2007; Rodriguez-Casado *et al.*, 2001; Thomas and Murphy, 1975; Thomas *et al.*, 1976; Thomas, Prescott and Olins, 1977; Tsuboi *et al.*, 2003b; Tsuboi, Benevides and Thomas, 2007a; Tuma and Thomas, 1997). Biological applications of Raman spectroscopy are greatly advantaged over infrared spectroscopy by both the relatively low interference of liquid water in the Raman effect and the diversity of sample states amenable to Raman analysis (solutions, suspensions, gels, precipitates, fibers, single crystals, amorphous solids, etc.). These considerations facilitate the correlation of biomolecular solution structures with corresponding high-resolution structures determined by X-ray crystallography (Helfand *et al.*, 2007; Li *et al.*, 1992b; Miura and Thomas, 1994; Thamann *et al.*, 1981; Totir *et al.*, 2007). More recently, the development of improved statistical approaches has allowed exploitation of Raman spectroscopy for detailed quantitative analyses of thermodynamic and kinetic properties of complex biomolecular systems (Hanus *et al.*, 1999, 2001; Nemecek *et al.*, 2008; Prochazka *et al.*, 2002). Concurrent developments in Raman instrumentation, including the advent of charge-coupled-device (CCD) detectors, efficient solid-state lasers, and efficient Raman microscopes, have similarly advanced biochemical, biophysical, pharmaceutical, and commercial applications of the related techniques of polarized Raman spectroscopy, Raman optical activity (ROA), ultraviolet resonance Raman (UVRR) spectroscopy, and

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

surface enhanced Raman spectroscopy (SERS) (Blanch *et al.*, 2002; Lewis and Edwards, 2001; Mikhonin and Asher, 2005; Mikhonin *et al.*, 2006; Pappas, Smith and Winefordner, 2000; Premasiri *et al.*, 2005; Tsuboi and Thomas, 2007b).

The objective of the present chapter is to consider representative applications of Raman spectroscopy, including UVRR and polarized Raman spectroscopy, that address fundamental questions of virus structure and assembly. The focus is on bacterial viruses (bacteriophages) that comprise only protein and DNA constituents. Often, the structural, mechanistic, and thermodynamic information obtainable from Raman spectra can be augmented by time-resolved data, such as those collected in the course of hydrogen/deuterium (H/D) exchanges of labile protein and DNA sites. UVRR approaches offer advantages of selectivity and sensitivity over conventional Raman spectroscopy, whereas polarized Raman spectroscopy is particularly well suited to determine orientations of specific moieties in viral assemblies.

13.2
Assembly Pathways of Representative Viruses

The maturation pathways of highly organized biomolecular assemblies such as viruses usually proceed in several discrete steps. Figure 13.1 shows this stepwise process for a typical DNA virus, the bacteriophage P22, which infects strains of the bacterium *Salmonella typhimurium* and is considered a paradigm for the assembly of many viruses that package a double-stranded (ds) DNA genome within an icosahedral protein shell (Casjens and Weigele, 2005). As shown in Figure 13.1, the pathway begins with the coassembly of 420 copies of the coat protein subunit (product of viral gene 5, or simply gp5) plus 12 copies of the portal protein subunit (gp1) and approximately 300 copies of the scaffolding protein subunit (gp8) to form a *procapsid*. The procapsid exterior consists of the gp5 shell incorporating the portal channel at a unique vertex, while the gp8 molecules are enclosed within. The procapsid also contains a few copies of the minor (ejection) proteins gp7, gp16, and gp20. At the time of DNA packaging, the procapsid shell matures into an expanded *capsid* shell and expels its scaffolding component. The genome packaging event (DNA translocation), which proceeds through the portal channel, is fueled by the hydrolysis of adenosine triphosphat (ATP) through the action of a virally encoded hetero-oligomeric enzyme called *terminase* (proteins gp2 and gp3). The last step in the assembly/maturation pathway involves the closure of the portal channel by accessory proteins (gp4, gp10, and gp26) and the attachment of six trimeric tailspike proteins (gp9), which serve as the host cell adhesion.

The filamentous bacterial viruses exhibit assembly pathways distinct from those of icosahedral viruses. The best-studied example is that of the filamentous virion Ff (strains fd, f1, and M13) (Figure 13.2), which infects *Escherichia coli*. The viral particle recognizes the conjugative pilus of the host cell and exploits the host transmembrane network of proteins (Tol–Pal system) for penetration. When the viral adhesion protein (pIII) binds to the Tol–Pal complex, the viral capsid is compromised and releases its single-stranded (ss) DNA genome. The major (pVIII) and minor (pIII, pVI, pVII, pIX) proteins of the

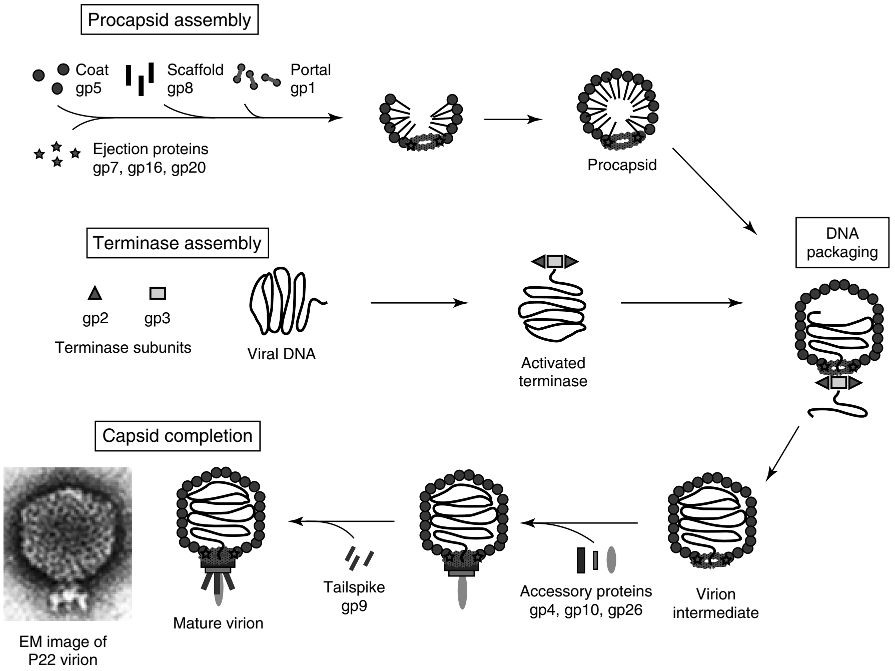
Procapsid assembly
Coat gp5
Scaffold gp8
Portal gp1
Ejection proteins gp7, gp16, gp20
Procapsid
DNA packaging
Terminase assembly
gp2
gp3
Terminase subunits
Viral DNA
Activated terminase
Capsid completion
EM image of P22 virion
Mature virion
Tailspike gp9
Accessory proteins gp4, gp10, gp26
Virion intermediate

Fig. 13.1 The assembly pathway of the *Salmonella* bacteriophage P22, which serves as a model for the assembly of many icosahedral double-stranded DNA viruses (Casjens and Weigele, 2005). The initial step (upper left) is the coassembly of 420 copies of the coat protein (gp5), 12 copies of the portal protein (gp1), approximately 300 copies of the scaffolding protein (gp8), and a few copies of minor ejection proteins (gp7, gp16, and gp20) to form a precursor shell (procapsid) incorporating a unique portal vertex. Subsequently, DNA packaging through the portal is fueled by ATP hydrolysis under the control of the phage terminase (subunits gp2 and gp3) and is accompanied by release of scaffolding and shell maturation. The final steps involve closure of the portal channel by accessory proteins (gp4, gp10, and gp26) and attachment of six trimeric tailspike proteins (gp9). The inset at the lower left shows an electron micrograph of the mature P22 virion.

capsid are disassembled and retained in the host inner membrane for subsequent use in the assembly of progeny virions. The ssDNA genome is converted into a replicative form for expression of viral structural and nonstructural proteins. Ultimately, the progeny virions exit the bacterial cell through a complex transmembrane channel formed from multiple nonstructural viral proteins (Overman and Thomas, 2008). The mature viral particle consists of the ssDNA genome in the form of a covalently closed circle extending from one end of the filament to the other and sheathed by thousands of copies of the major coat protein plus a few copies of each of two minor proteins at each of the filament ends.

13.3
Advantages of Raman Spectroscopy in Virus Research

Raman spectroscopy, like infrared spectroscopy, is a method for probing

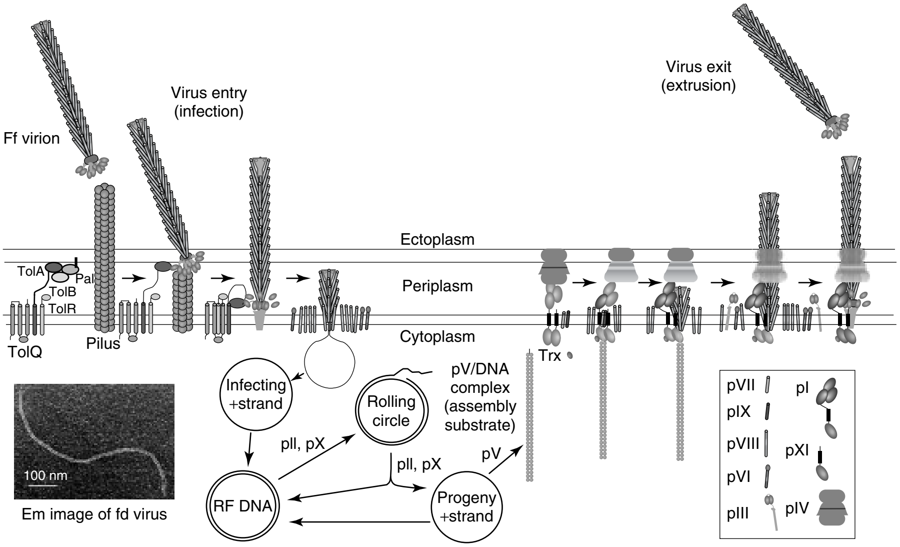
Ff virion
Virus entry
(infection)
Virus exit
(extrusion)
Ectoplasm
Periplasm
Cytoplasm
TolA
Pal
TolB
TolR
TolQ
Pilus
Trx
Infecting
+strand
Rolling
circle
pV/DNA
complex
(assembly
substrate)
pII, pX
pII, pX
pV
RF DNA
Progeny
+strand
100 nm
Em image of fd virus
pVII
pIX
pVIII
pVI
pIII
pI
pXI
pIV

Fig. 13.2 The life cycle and assembly pathway of the filamentous bacteriophage Ff, which infects *Escherichia coli* (Overman and Thomas, 2008). Essential features of membrane recognition and penetration are discussed in the text. The mature viral particle consists of a single-stranded DNA genome in the form of a covalently closed circle that extends from one end of the capsid sheath to the other. The sheath comprises thousands of copies of the major coat protein (pVIII) plus a few copies of minor proteins at each filament end. The inset at the lower left shows an electron micrograph of the mature Ff virion.

structures and interactions of molecules by measuring the frequencies (wavenumbers) of molecular vibrations. Although the two methods differ fundamentally in the mechanism of interaction between radiation and matter, one obtains in both cases a vibrational spectrum consisting of a number of discrete bands, the wavenumbers and intensities of which are determined by the nuclear masses in motion, the equilibrium molecular geometry, and the molecular force field. An important advantage of Raman over IR for applications to viral proteins, nucleic acids, and their complexes is the virtual transparency of liquid water (both H_2O and D_2O) in the Raman effect. This greatly simplifies the analysis of solution spectra and also facilitates the investigation of hydrogen-isotope-exchange phenomena. Other notable advantages of Raman spectroscopy are ease of implementation for samples in different physical states (solutions, suspensions, gels, precipitates, fibers, single crystals, amorphous solids, etc.), applicability to large supramolecular assemblies, a spectroscopic time scale ($\sim 10^{-14}$ seconds) that is fast in comparison to both biomolecular structure transformations and protium/deuterium (H/D) exchanges, nondestructiveness of data collection protocols, relatively low mass (<1 mg) and volume ($\sim 1 \mu l$) requirements for most sampling protocols, and freedom from the need for chemical labels or probes. Additionally, a large database of Raman spectra of model compounds has been established for applicability to viral proteins and nucleic acids.

The technique of Raman difference spectroscopy provides a complementary and powerful experimental approach for applications to viruses. For example, digital difference methods can be exploited to: (i) facilitate definitive Raman band assignments by detection of small wavenumber shifts and/or intensity changes accompanying residue-specific isotopic substitutions or mutations (Overman *et al.*, 1994; Overman and Thomas, 1999; Raso *et al.*, 2001), (ii) assess qualitatively and quantitatively the structural changes induced in proteins and nucleic acids when solution conditions, such as temperature, pH, ionic strength, or ligand concentration are altered (Benevides, Overman and Thomas, 2005b; Thomas, 1999; Tuma and Thomas, 2002), (iii) detect localized interactions and conformational changes in viruses (Aubrey, Casjens and Thomas, 1992; Overman *et al.*, 1998a; Wen, Overman and Thomas, 1997; Wen, Armstrong and Thomas, 1999), and (iv) probe changes in solvent accessible surfaces along the viral assembly pathway (Reilly and Thomas, 1994; Tuma, Prevelige and Thomas, 1998b).

Methods of polarized Raman spectroscopy have also been successfully applied to viruses. The data obtained from polarized Raman measurements are particularly informative when the virus specimen can be precisely oriented with respect to the directions of polarization of incident and scattered laser light. This is readily achieved for the examination of oriented filamentous viruses in the Raman microscope (Tsuboi *et al.*, 2003a). A comprehensive review of applications has been given recently (Tsuboi and Thomas, 2007b).

13.4
Experimental Methods for Raman Spectroscopy

Instrumentation for Raman spectroscopy of viruses is generally similar to that employed for investigating other macromolecules in condensed states (solutions, crystals, fibers, etc.) (Pappas, Smith and Winefordner, 2000). Diode-pumped solid-state lasers are available to provide stable excitations throughout the visible and near-infrared (NIR) wavelength regions. Excitation in the 750–800 nm interval, such as those provided at 785 nm by the Invictus laser (Kaiser Optical Systems, Inc., Ann Arbor, MI), offers the additional advantage of freedom from potentially interfering fluorescence transitions. Spectrometers for biomolecular applications ordinarily incorporate the holographic notch filter for rejection of elastically scattered light, a single axial transmissive grating to maximize photon throughput and a liquid-nitrogen-cooled CCD detector to optimize spectrophotometric sensitivity and selectivity. Recent sampling innovations, which are particularly useful for studying time-dependent processes in viruses, include solution flow dialysis (Caille, Pigeon-Gosselin and Pezolet, 1983; Tuma and Thomas, 1996b) and crystal diffusion (Carey, 2006; Helfand *et al.*, 2007). In principle, time-resolved Raman spectroscopy can span a wide temporal domain, from tens of femtoseconds (using pulsed lasers) to indefinitely long periods. The slower structural transformations that characterize viral assembly and maturation processes, including genome packaging enzyme activities, are relatively easily addressed (Jenkins *et al.*, 1999; Li, Johnson and Thomas, 1993b; Tuma and Thomas, 1996b).

Solutions of viruses are generally sealed in thin glass capillaries and thermostated at the appropriate temperature (Thomas and Barylski, 1970) for collection of Raman spectra. The exciting laser beam is focused on the sample volume (1–10 μl) and the light scattered at 90° is directed to the entrance slit of the spectrograph, where it is analyzed and directed to the CCD detector (Tuma *et al.*, 1993).

Solid samples, including oriented fibers and crystals that typically have dimensions of the order ∼100 μm, are most effectively probed by means of a Raman microspectrophotometer, which consists of a confocal microscope coupled to an efficient single grating spectrophotometer. The microscope functions as the sample illumination chamber and typically is coupled to a video camera and monitor to simplify optical alignment and focusing of the laser beam with respect to the sample. The incident laser beam is directed onto the oriented specimen through the microscope objective; the Raman scattering, preferably at 180°, is collected with the same objective and then directed to a beam splitter, polarization analyzer, optical scrambler, monochromator, and CCD detector (Benevides *et al.*, 1993; Tsuboi and Thomas, 2007b). A thermoelectrically cooled sample stage is also advantageous. The success of the method depends ultimately upon achieving a known, uniform, and stable orientation of the sample molecules with respect to the directions of the electric vectors of the exciting and scattered radiation. Experimental techniques appropriate to Raman microspectroscopy of oriented viruses and their protein and nucleic acid constituents have been surveyed recently (Tsuboi and Thomas, 2007b).

We note that NIR lasers and NIR-optimized detectors are finding increasing use in commercial Raman spectrometers and microscopes. Often, it is desirable to compare the spectra collected using such instrumentation with spectra reported in the literature that were collected using excitation wavelengths in the visible region. It should be kept in mind that the relative intensities of bands in NIR-excited Raman spectra of proteins or nucleic acids may differ significantly from those observed in the corresponding spectra obtained using visible excitation wavelengths (Benevides, Overman and Thomas, 2005a).

13.5
Analysis of Raman Spectroscopic Data

13.5.1
General Considerations

Raman spectra of viral proteins and nucleic acids are usually collected repetitively, and the individual data collections are averaged to achieve an acceptable signal-to-noise ratio (∼10: 1 or greater). The length of each accumulation (exposure time) is dictated by the stability of the sample at the experimental conditions employed and the objective of the experiment. The raw data typically require one or more of the following refinement steps: (i) precise wavenumber calibration versus known standards (e.g. inert gas emission spectra and/or Raman spectra of organic liquids), (ii) removal of transient drift in the spectral baseline by polynomial fitting, (iii) elimination of scattering contributions from the silica glass capillary by spectral subtraction, (iv) elimination of detector artifacts caused by cosmic rays or spurious light reaching the CCD, and (v) normalization of the intensities of all exposures in the accumulation to the same ordinate scale. Artificial refinements of the data, such as the use of computational algorithms for smoothing of natural spectral noise, narrowing of natural spectral bandwidths, or reduction of spectral band overlap, should be scrupulously avoided. Whenever a series of spectra is processed, particular care must be taken to avoid introduction of any systematic errors (Benevides *et al.*, 2002; Movileanu, Benevides and Thomas, 2002a).

Quantitative analysis of the ordered secondary structures present in native proteins – specifically α-helix and β-strand conformations of the polypeptide chain – can be accomplished by decomposition of the prominent amide I (1640–1680 cm^{-1}) or amide III (1230–1310 cm^{-1}) bands of the Raman spectrum. Several approaches have been described (Asher *et al.*, 2001; Bandekar, 1992; Sane, Cramer and Przybycien, 1999; Williams, 1983) including decomposition of the amide I profile in terms of a relatively simple basis set (Berjot, Marx and Alix, 1987). Although the method of Berjot *et al.* is constrained by the limited reference data set upon which it is based (obtained mainly from spectra of small globular proteins), it is particularly useful for viral proteins and their assemblies, especially when quantifying a change in the structure. Fourier deconvolution of complex bands can also prove helpful in certain cases, although such methods should be used with caution and the results interpreted accordingly (Thomas and Agard, 1984).

Analysis of Raman spectral changes that reflect structural transformations or interactions of viral proteins and nucleic acids can be exploited to elucidate kinetic and thermodynamic parameters. Relevant applications have been described recently (Movileanu, Benevides and Thomas, 2002b; Tuma, Preveliage and Thomas, 1998b). More powerful approaches – based upon factor analysis (Malinowski, 2002) – offer greater sensitivity and selectivity, and improved insight into the structural basis of biomolecular interaction. Several equivalent algorithms are available to statistically treat large sets of Raman spectra, including principal component analysis (PCA) and singular value decomposition (SVD). The latter enables testing of specific theoretical models and provides quantitative parameters (Hanus *et al.*, 1999, 2004; Prochazka *et al.*, 2002). Physicochemical properties and structural information are derived simultaneously by SVD and can provide a comprehensive characterization of the system under investigation.

13.5.2
Singular value decomposition

Using the SVD algorithm, a series of N Raman spectra can be represented as $Y_i(\nu)$, where $i = 1, \ldots, N$ and ν is the number of data points in each spectrum. The set $Y_i(\nu)$ is transformed into a set of N orthonormal components (singular vectors) given as $S_j(\nu)$, where $j = 1, \ldots, N$. The vectors S_j decrease in statistical weight with increasing j. We express the statistical weights as W_j (singular values); the relative contribution of the jth component to the i-th Raman spectrum is expressed in terms of the unitary matrix V_{ij} by Eq. (1):

$$Y_i(\nu) = \sum_{j=1}^{N} W_j V_{ij} S_j(\nu) \qquad (1)$$

The series of N Raman spectra $[Y_i(\nu)]$ can be approximated within the experimental error(s) by a subset of $S_j (j = 1, 2, \ldots, M)$ components, where $M < N$. This value of M also represents the number of independent spectral species in the original series and constitutes the dimension of the factor analysis (Malinowski, 2002).

The SVD approach is illustrated in Figure 13.3 for a simple series of spectra, namely, the 28 spectra ($N = 28$) of Na_2SO_4 at two different concentrations (4 and 20 mM) in H_2O and as a function of temperature over the interval $11 < t < 70\,^{\circ}C$. Here, we have monitored the spectral interval 800–1200 cm^{-1} ($\nu = 400$), which includes the prominent 980 cm^{-1} band of the sulfate ion (symmetrical S–O stretching mode). The SVD analysis of Figure 13.3 shows that all of the spectral information is expressed by three independent spectral components $S_1(\nu)$, $S_2(\nu)$, $S_3(\nu)$. Thus, in this simple example, the number of independent species is identified as three ($M = 3$). In biomolecular systems the number of independent species is usually not known and the determination of factor dimension by SVD can be crucial. In such cases, the corresponding SVD analysis can also provide information of direct physicochemical significance, as will be illustrated by an example below (see Section 13.7.2.2).

An important and convenient criterion for determination of the factor dimension in the SVD analysis is the residual error between the parent spectra and those generated from the M subspectra. As long as an increase of M results in a better than linear decrease of the residual error, the additional subspectra describe significant and relevant spectral changes (Malinowski, 2002). Auxiliary criteria are: (i) randomness of the relative contributions to the original Raman spectra (rows of the matrix V_{ij}), (ii) signal-to-noise ratio in the subspectrum $S_j(\nu)$, and (iii) singular value W_j. The factor dimension can, however, exceed the number of genuine spectral components owing to the presence of various artifacts in the Raman spectra. Typical artifacts may be (i) nonuniform background correction within the series, (ii) wavenumber scale drift, or (iii) traces of parasitic signals, such as the contributions from the glass capillary cell. Exceptional care must therefore be taken to eliminate all artifacts during the processing of the spectral series.

Overall, SVD accomplishes a substantial reduction in the size of the experimental data set, generally from many experimentally obtained spectra (N) to a few computed subspectra (M). Fitting of a theoretical model leads to additional data reduction (elimination of the spectral components). In effect, the model is applied only to the relevant columns of the matrix V_{ij} multiplied by their corresponding singular values W_j (Hanus *et al.*, 1999, 2004; Prochazka *et al.*, 2002).

13.5.3
Interpretation of Polarized Raman Spectra

Up to the present, polarized Raman spectroscopy has not been widely used to study protein structures. This is due in part to the difficulty of determining the orientations of protein molecules and their subgroups in macroscopic samples suited for polarized Raman analysis, and in part to the complexity of interpreting the polarized Raman scattering. Raman

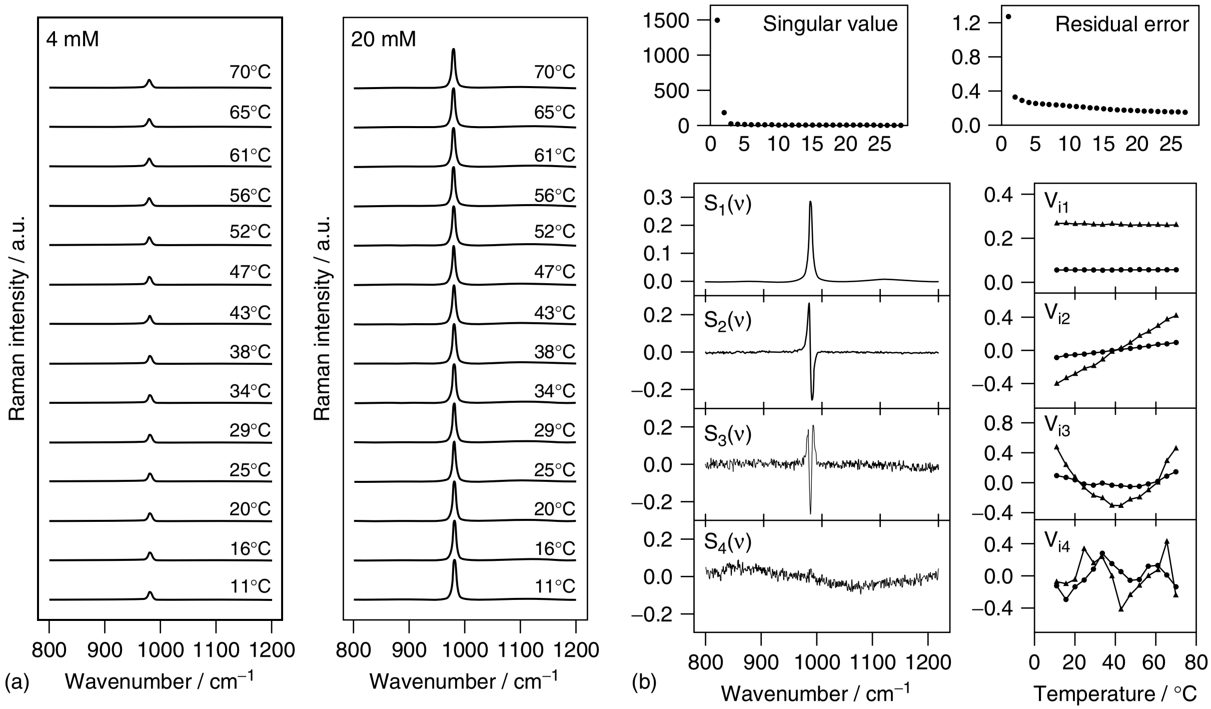

Fig. 13.3 (a) A series of Raman spectra showing the symmetrical stretching band of the sulfate ion (SO_4^{-2}) near 980 cm^{-1}. Data were obtained from solutions of Na_2SO_4 (4 and 20 mM) over the temperature range $11 < t < 70\,°C$, as labeled. (b) Factor analysis treatment of the spectral series using the SVD algorithm and showing singular value (upper left), singular vectors with the highest singular values $S_1(\nu)$, $S_2(\nu)$, $S_3(\nu)$, and $S_4(\nu)$ (lower left panel), parameter-dependent distribution of the singular vectors V_{i1}, V_{i2}, V_{i3}, and V_{i4} (lower right), and residual error (upper right). The vector $S_1(\nu)$ represents a weighted average spectrum of the series, where contributions V_{i1} reflect the higher intensity of the 980 cm^{-1} band in the more concentrated solution (▲). The singular vectors $S_2(\nu)$, $S_3(\nu)$, and $S_4(\nu)$ represent linearly independent spectral changes of, respectively, decreasing statistical importance. $S_2(\nu)$ shows the shift of the 980 cm^{-1} band to lower wavenumber with increasing t; $S_3(\nu)$ shows the corresponding broadening; and S_4 (and other vectors not shown) represent random noise in the data. Relative contributions of the significant singular vectors to the original Raman spectra are expressed in the V_{i1}, V_{i2}, and V_{i3} scores, respectively.

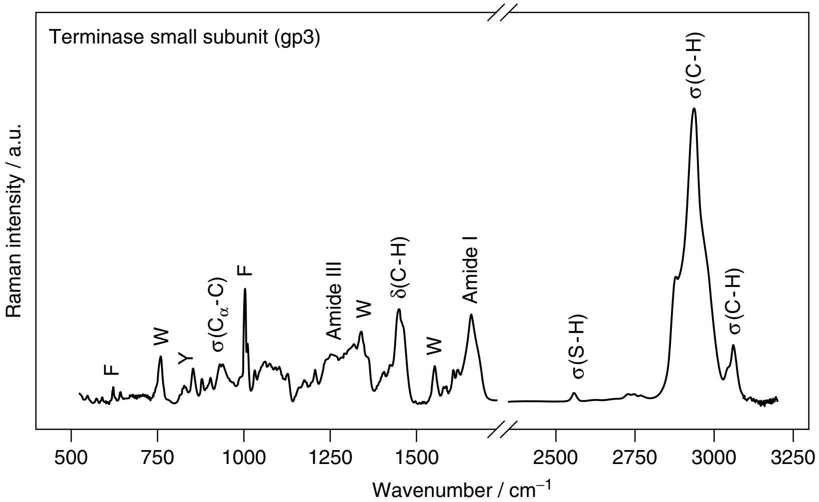

Fig. 13.4 Raman signature of the terminase small subunit (gp3) of bacteriophage P22. The spectrum is divided into: (i) the region 520–1750 cm^{-1}, where the often utilized amide modes of the polypeptide main chain and skeletal stretching and bending modes of the side chains predominate, and (ii) the often neglected region 2300–3200 cm^{-1}, where hydrogenic stretching modes of cysteinyl SH and aliphatic and aromatic CH groups occur. Assignments for many prominent peaks are shown. Labels indicate stretching (σ) and bending (δ) modes of specific bonds or assigned side chain markers (1-letter amino acid abbreviations). See also Table 13.1.

anisotropy is defined by a set of relative spectral intensities, I_{jk}, for which the indices j and k refer to a rectangular coordinate system (*abc*) that is fixed to the protein sample, such that j and k represent, respectively, the directions of the incident and Raman scattered electric vectors. Measurement and interpretation of the complete set of Raman intensities I_{aa}, I_{bb}, I_{cc}, I_{ab}, I_{bc} and I_{ca}, can represent a formidable challenge.

13.6
Characteristics of Raman Spectra of Viral Proteins and Nucleic Acids

The structural interpretation of Raman and UVRR spectra of viral proteins and nucleic acids requires definitive spectral band assignments. The considerable progress achieved in this area has been reviewed elsewhere (Austin, Jordan and Spiro, 1993; Benevides, Overman and Thomas, 2005a, 2005b; Miura and Thomas, 1995a; Tsuboi and Thomas, 2007b; Tuma and Thomas, 2002; Tuma, 2005). In summary, the required definitive assignments can be achieved by diverse approaches, including: (i) comparison of Raman wavenumbers and intensities with corresponding infrared spectral data; (ii) determination of spectral wavenumber shifts accompanying stable isotope substitutions, usually $^{1}H \rightarrow {}^{2}H$ (D), $^{12}C \rightarrow {}^{13}C$, $^{16}O \rightarrow {}^{18}O$, or $^{14}N \rightarrow {}^{15}N$; (iii) effects of pH, temperature, and/or environmental parameters on the Raman and infrared spectra; (iv) detailed and well analyzed Raman and infrared spectra of simpler molecules, which are structurally related

Tab. 1 Raman wavenumbers and assignments for the prominent bands in spectra of H_2O and D_2O solutions of proteins (Benevides *et al.*, 2004; Overman and Thomas, 1999; Tuma *et al.*, 1998a)

Wavenumber (cm^{-1})[a] H_2O	D_2O	Assignment[b]
621	621	Phe
643	642	Tyr
699	–	amide V
722	722	Met, Tyr, δ(C–C–C)
758	754	Trp, Ala
825	825	Tyr (Fermi doublet)
850	850	Tyr (Fermi doublet)
877	–	Trp
897	887	Lys, Arg, Ala
–	910	amide III′, side-chains
–	927	amide III′ (β-strand)
934	–	ν(C^α-C^β) (α-helix)
–	950	amide III′ (α-helix)
957	–	Val. Leu, δ(CH_3)
–	963	amide III′ (coil)
–	992	amide III′ (extended conformation)
1003	1003	Phe
1012	–	Trp
1032	1032	Phe
1060	–	Lys, Arg, Ala, ν(C–C, C–O)
1081	–	Lys, ν(C–N)
1101	–	Pro, Ala, ν(C^α–N), ν(C–C)
1127	–	Trp, ν(C–C), ν(C–N)
1156	–	δ(CH_3)
1174	1174	Tyr, δ(CH_3)
1206	1206	Tyr, Phe, Trp
1230–1240	–	amide III (β-strand)
1245–1255	–	amide III (coils, turns)
1264–1272	–	amide III (α-helix)
1264	1260	Tyr
1300	–	amide III (α-helix)
1318	1318	δ(CH_2)
1340	–	Trp (Fermi doublet), δ(C–H)
1358	–	Trp (Fermi doublet), δ(C–H)
1406	1404	ν(CO_2)
1450	1450	δ(CH_2), δ(CH_3)
–	1458	amide II′ (α-helix)
–	1470	amide II′ (extended conformation)
1552	1552	Trp
1605	1608	Phe, Tyr
1618	1614	Tyr, Trp
1645–1654	1633–1640	amide I/I′ (α-helix)
1665–1674	1655	amide I/I′ (β-strand)
1660–1665	–	amide I (coil)
1684	1669	amide I/I′ (extended conformation)

[a] Wavenumber values are believed accurate to ±1 cm^{-1}.

[b] Abbreviations: ν, stretching vibration; δ deformation vibration.

to the viral protein or nucleic acid; (v) Raman analyses of single crystals of related molecules, the structures of which have been authenticated by X-ray diffraction analysis, or analysis of corresponding solution data obtained by high-resolution NMR methods; (vi) *ab initio* calculations of intramolecular force fields and normal coordinate calculations based upon them; (vii) depolarization ratios of the Raman bands for randomly oriented molecules; (viii) polarized Raman measurements on oriented fibers or single crystals, and (ix) Raman excitation profiles, which correlate Raman intensities with laser excitation wavelengths.

Generally, Raman spectra of viral proteins are dominated by localized vibrations of multiply bonded atoms, such as the amide vibrational modes (amide I and amide III) associated with the peptide group ($C^{\alpha}-NH-C{=}O$) and aromatic ring modes associated with the amino acid side chains of tyrosine (Tyr), tryptophan (Trp), and phenylalanine (Phe). The sulfhydryl (S–H) stretching mode of cysteine (Cys) is also a significant contributor. Conversely, contributions from vibrational modes of nonaromatic side chains are generally weak. However, when the numbers of such side chains are large, their cumulative contributions can be significant. Additionally, chemical groups that are common to many side chains, such as the $C^{\beta}H_2$ moiety, can contribute importantly to the protein Raman spectrum (Overman and Thomas, 1999). Prominent bands represented in Raman spectra of H_2O and D_2O solutions of most viral proteins and their corresponding group assignments are summarized in Table 13.1.

Raman spectra of viral nucleic acids (DNA and RNA) are similarly dominated by spectral bands attributable to ring vibrations involving multiply bonded atoms of the purine and pyrimidine nucleotide bases (A, T or U, G, C). Symmetrical stretching modes of the phosphodiester (C–O–P–O–C) and phosphodioxy (O–P=O) linkages of the sugar–phosphate backbone are also major contributors (Thomas and Tsuboi, 1993). By comparison, the vibrational modes of the furanosyl methylenes in the deoxyribosyl (DNA) or ribosyl (RNA) sugars lead to feeble Raman scattering. However, as noted above for the aliphatic groups of nonaromatic amino acid side chains in proteins, the intrinsically weak Raman scattering of the furanosyl groups can achieve prominence in the spectra of DNA or RNA because of the relatively large numbers of such groups represented in the nucleic acid chain. Prominent bands in the Raman spectrum of DNA have been listed in previous surveys (Benevides, Overman and Thomas, 2005b; Thomas, 1999; Tuma and Thomas, 2002).

13.6.1
The Protein Raman Spectrum

13.6.1.1 Raman Markers of Main Chain Conformation: Amide I and Amide III Modes

The Raman spectrum of a specific viral protein, the terminase small subunit (gp3) of bacteriophage P22, is shown in Figure 13.4. Assignments for the most prominent spectral bands of gp3 correspond to the listing given in Table 13.1. The dominating amide I and amide III markers, which occur respectively in the intervals 1640–1680 cm^{-1} and 1225–1320 cm^{-1} (Bandekar, 1992; Chen and Lord, 1974a; Lippert, 1976; Miura and Thomas, 1995a; Thomas, Prescott and Urry, 1987; Yu, Lippert and Peticolas, 1973), are particularly useful as quantitative indicators of protein secondary structure (Berjot, Marx and Alix, 1987; Williams, 1983). Deuteration effects on these amide markers, which shifts them respectively to circa 1630–1670 cm^{-1} (designated amide I′) and circa 900–1000 cm^{-1} (amide III′) have also been profitably exploited (Thomas and Day, 1981; Tuma and Thomas, 1997; Williams *et al.*, 1980).

In addition to the amide I/I′ and amide III/III′ markers, other conformation-sensitive bands have been identified in Raman spectra of peptide model compounds and proteins. For example, vibrations localized in the peptide main chain near 950 and 1330 cm^{-1}, although not considered amide modes, are sensitive quantifiers of α-helix content (Benevides *et al.*, 2004; Tsuboi *et al.*, 2000). The amide II′ mode (circa 1400–1500 cm^{-1}), which involves peptide C–N stretching and achieves prominence only in spectra of deuterated proteins, can also be exploited for conformational analyses (Tuma *et al.*, 1996a; Tuma, Prevelige and Thomas, 1998b).

Because all Raman amide modes typically undergo significant and well-characterized wavenumber shifts with peptide group deuteration (NH → ND), they are conveniently exploited as probes of main chain dynamics as well as conformation (Tuma and Thomas, 1997; Tuma, Prevelige and Thomas, 1998b).

13.6.1.2 **Raman Markers of Side Chains**

Raman bands of tryptophan have been investigated extensively and many useful structural correlations have been developed for tryptophan side chains in proteins (Combs *et al.*, 2005; Harada, Miura and Takeuchi, 1986; Miura, Takeuchi and Harada, 1988, 1989; Takeuchi and Harada, 1986; Takeuchi, 2003). The intense and sharp tryptophan marker in the interval 1540–1560 cm^{-1} is indicative of the magnitude of the indolyl side chain torsion angle $\chi^{2,1}$ ($C^{\delta1}-C^{\gamma}-C^{\beta}-C^{\alpha}$). The tryptophan Fermi doublet with components at 1340 and 1360 cm^{-1} reflects the local environment of the indolyl ring, such that the intensity ratio I_{1360}/I_{1340} increases with increasing hydrophobicity. The tryptophan marker near 875 cm^{-1} is somewhat sensitive to indolyl NH hydrogen bonding. Strong N–H···X interactions lead to a band center near 871 cm^{-1}, while a hydrophobic environment shifts the band center to 883 cm^{-1}. The breathing vibration of the indolyl ring generates a tryptophan marker near 755 cm^{-1}, the intensity of which has been considered inversely related to the hydrophobicity of the local ring environment. Typically, the band intensity increases with decreasing hydrophobicity of the indolyl ring environment. Assignments, deuteration effects, and structural interpretations for tryptophan markers have been discussed in greater detail by Takeuchi (2003). Recently, the Raman markers associated with a nonhydrogen-bonding environment for the tryptophan side chain were inferred from a comprehensive vibrational analysis of the indolyl moiety in the gas phase (Combs *et al.*, 2005).

It has long been recognized that the *para*-substituted phenyl ring of tyrosine can generate a pair of relatively intense Raman bands near 830 and 850 cm^{-1} that corresponds to a Fermi interaction doublet (Lord and Yu, 1970a; Siamwiza *et al.*, 1975). The correlation between the ratio of intensities of the doublet components (I_{850}/I_{830}) and the hydrogen-bonding state of the phenolic OH group, which was initially quantified by Siamwiza *et al.* (1975) and later extended to the nonhydrogen-bonded phenoxyl state by Arp *et al.* (2001), is quite useful in viral protein applications (Overman *et al.*, 1994; Overman and Thomas, 1995; Tsuboi *et al.*, 2001, 2003a; Wen, Overman and Thomas, 1997). The doublet intensity distribution is governed by Fermi coupling between the phenolic ring-stretching fundamental expected near 840 cm^{-1} and the first overtone of the phenolic ring deformation mode occurring near 420 cm^{-1}, such that the following intensity distribution applies: (i) If the tyrosine phenoxyl group functions as a strong hydrogen-bond donor only, then $I_{850}/I_{830} = 0.30$, (ii) if the phenoxyl group functions as both a donor and an acceptor, then $I_{850}/I_{830} = 1.25$, (iii) if the phenoxyl group functions as a strong hydrogen-bond acceptor only, then $I_{850}/I_{830} = 2.5$, and (iv) if the phenoxyl group is not hydrogen bonded, then $I_{850}/I_{830} = 6.7$. For a protein containing a single tyrosine, the I_{850}/I_{830} ratio is unambiguously interpretable. Conversely, if the protein contains multiple tyrosines, then the observed I_{850}/I_{830} reflects the average hydrogen-bonding state of all tyrosine phenoxyls. In the latter eventuality, site-directed isotopic substitutions or mutagenesis can be exploited to deduce the hydrogen-bonding states of individual tyrosyl residues (Overman *et al.*, 1994). The nonhydrogen-bonded phenoxyl state, although unlikely to occur for tyrosines either within folded globular proteins or at solvent-exposed protein surfaces, is often encountered in the tightly packed α-helical subunits of filamentous virus capsids (Arp *et al.*, 2001; Overman *et al.*, 1994; Overman and Thomas, 1995).

The Raman band resulting from the cysteine sulfhydryl bond (S–H) stretching vibration is a unique probe of the local S–H structure and dynamics of the cysteine side chain (Li and Thomas, 1991; Li, Wurrey and Thomas, 1992a; Tuma *et al.*, 1993). The sulfhydryl Raman marker occurs in a region of the spectrum (2500–2600 cm^{-1}) that is virtually devoid of interference from any other Raman band of a protein or its aqueous solvent. The relatively large Raman cross-section associated with the S–H stretching mode and the sensitivity of the bond stretching vibration to hydrogen-bonding interactions and side chain conformation are well documented. Specific correlations have been established between the strength of sulfhydryl hydrogen bonding to an appropriate acceptor atom ($\text{S–H}\cdots\text{X}$ bonding) and the wavenumber of the S–H bond stretching vibration (Li and Thomas, 1991; Li, Wurrey and Thomas, 1992a). The stoichiometric dependence of the S–H Raman band on thiol concentration also confers a unique spectrophotometric value on the S–H marker for assessing the pK_a values of thiol/thiolate equilibria (Li *et al.*, 1993a; Vohnik *et al.*, 1998). Because the S–H marker requires the reduced state of cysteine, it can also prove useful for distinguishing this state from its oxidized counterpart, cystine, when either is possible in a protein.

Raman bands originating from bond stretching vibrations of the disulfide network of cystine ($\text{C}^{\alpha}\text{–C}^{\beta}\text{–S–S–C}^{\beta}\text{–C}^{\alpha}$) have been considered in detail (Lord and Yu, 1970a, 1970b; Sugeta, 1975; Van Wart *et al.*, 1973; Van Wart and Scheraga, 1977, 1986b). Most useful is the disulfide (S–S) stretching marker that ordinarily occurs within the 450–550 cm^{-1} interval and is sensitive in wavenumber to the

$C^\beta-S-S-C^\beta$ dihedral angle (Van Wart and Scheraga, 1986a).

Raman bands of histidine and the histidinium ion are relatively weak compared to those of the aromatic amino acids (tyrosine, tryptophan, and phenylalanine). However, when the protein does not contain many aromatic residues, some histidine bands can be identified. At neutral pH, the Raman spectrum of the imidazolium ring of histidine exhibits several weak bands, whereas at acidic pH, their Raman intensities are redistributed to fewer bands of intermediate intensities. In its deuterio form, the imidizolium ring exhibits a relatively strong band at 1408 cm^{-1}, which has been assigned to the $N_1-C_2-N_3$ symmetric stretch (Tasumi *et al.*, 1982). The band is particularly useful for monitoring histidine H/D exchange dynamics as well as histidine–histidinium equilibria in D_2O (Harada *et al.*, 1982; Okada, Miura and Takeuchi, 2003). A linear correlation between the wavenumber of the 1408 cm^{-1} Raman band and the histidine side chain conformation has been proposed (Takeuchi *et al.*, 1991; Takeuchi, 2003). The histidine markers are also sensitive to the binding of transition metals at imidazole sites and the Raman spectrum has been extensively exploited for monitoring such interactions (Miura *et al.*, 1998; Takeuchi, 2003).

13.6.2
The Nucleic Acid Raman Spectrum

The concept of localized group frequencies, as referred to above, may break down in biological macromolecules as a consequence of mechanical coupling between vibrational modes of covalently linked groups. Although this phenomenon is rarely encountered in Raman spectra of proteins, it occurs often in Raman spectra of nucleic acids – in the form of coupling between modes of the nucleotide purine or pyrimidine bases [adenine (A), cytosine (C), guanine (G), and thymine (T)] and those of the sugar–phosphate moiety. The resulting delocalized vibrational modes may then generate Raman bands that combine the relatively high intensity typical of localized purine and pyrimidine vibrational modes with the conformational dependency expected of modes associated with the sugar–phosphate backbone. Several such Raman bands have been identified for both DNA and RNA (Erfurth, Kiser and Peticolas, 1972; Thomas, Medeiros and Hartman, 1971). These Raman bands manifest dependence of wavenumber and/or intensity upon the nucleic acid macromolecular conformation (secondary structure) and are often referred to as Raman conformation markers.

Two categories of Raman conformation markers have been delineated (Thomas and Wang, 1988b), those relating to the overall conformation of the sugar–phosphate backbone and those relating to the conformation of the constituent nucleosides. The former are sensitive primarily to changes in the torsions of the phosphodiester bond network of the backbone ($-P-O5'-C5'-C4'-C3'-O3'-P-$), but do not depend strongly upon the identity of the attached base. These are listed in Table 13.2. Thus, the occurrence of unique backbone conformation markers for the three known secondary structures of DNA, that is *Z* DNA (745 ± 3 cm^{-1}), *A* DNA (807 ± 3 cm^{-1}) and *B* DNA (835 ± 7 cm^{-1}), facilitates the use of Raman spectroscopy to distinguish the

Tab. 2 Raman bands diagnostic of the *A*, *B* and *Z* backbone conformations of double-stranded DNA[a].

Mode type	*A* DNA	*B* DNA	*Z* DNA
σ(C–O)	706 ± 5		
σ_{sym}(O–P–O)	807 ± 3[b]	790 ± 5	745 ± 3
σ_{sym}(O–P–O)		835 ± 7[c]	
σ_{sym}(PO_2^-)	1099 ± 1	1092 ± 1	1095 ± 2
δ(C2′H_2)	1418 ± 2	1420 ± 2	1425 ± 2

[a] Wavenumbers (cm^{-1} units) are determined from Raman spectra of DNA crystals and fibers of known structure. Abbreviations: σ, stretching; sym, symmetric; δ, deformation.

[b] The corresponding band in *A* RNA is observed at 813 ± 2 cm^{-1}.

[c] In *B* DNA containing both GC and AT pairs, the band is broad and centered close to 835 cm^{-1}. In *B* DNA containing only GC or only AT pairs, the band is sharp and centered, respectively, close to 830 or 840 cm^{-1}.

A, *B*, and *Z* conformations from one another. The relatively large range quoted in Table 13.2 for the *B* marker (±7 cm^{-1}) reflects the structural complexity of *B* DNA in comparison to *A* and *Z* DNAs. In fact, the diversity of Raman markers observed for *B* DNAs of different base composition suggests a family of structures, the precise conformations of which are dependent upon the A–T and G–C base-pair content and sequence. Thus, regions of *B* DNA containing predominantly G–C and C–G base pairs exhibit the *B* marker near 830 cm^{-1}, while those containing predominantly A–T and T–A pairs exhibit the marker near 840 cm^{-1}. This distinction is believed due to the slightly different phosphodiester torsions associated with the groove dimensions of G–C/C–G rich (narrow minor groove) and A–T/T–A rich (wide minor groove) sequences. Consistent with this finding is a resultant broad band centered near 835 cm^{-1} in DNAs containing comparable percentages of G–C/C–G and A–T/T–A pairs.

Nucleoside conformation markers, on the other hand, originate primarily from certain base ring-stretching vibrations that are weakly coupled via the glycosyl bond to motions of the furanose ring atoms. These are listed in Table 13.3. For each nucleoside listed the markers are sensitive to either furanose pucker, as measured by torsion angle δ (defined by atoms C5′–C4′–C3′–O3′ of the sugar), or glycosyl orientation, as measured by the torsion angle χ (defined by atoms O4′–C1′–N1–C2 or O4′–C1′–N9–C4 for pyrimidine or purine nucleosides, respectively), or both.

These considerations apply also to RNA, with the important distinctions that the base uracil (U) replaces thymine (T) and the sugar ribose replaces deoxyribose. Importantly, only one type of secondary structure occurs for double-stranded RNA, namely *A* RNA, which exhibits many architectural and spectroscopic similarities to *A* DNA.

The following additional generalizations can be made regarding relative Raman band intensities of the nucleic acid constituents: (i) In-plane stretching vibrations of conjugated CC, CN, and CO double bonds of the bases yield

Tab. 3 Raman bands diagnostic of the conformations of deoxynucleosides and nucleosides of nucleic acids[a].

Residue	C3′ endo/anti	C2′ endo/anti	C3′ endo/syn	C2′ endo/syn
G	664 ± 2[b]	682 ± 2	625 ± 3	671 ± 2
	1318 ± 2	1333 ± 3[c]	1316 ± 2	1324 ± 2
A	644 ± 4	663 ± 2	624 ± 3	
	1335 ± 2	1339 ± 2	1310 ± 5	
C	780 ± 2	782 ± 2	784 ± 2	
	1252 ± 2	1255 ± 5	1265 ± 2	
T	642 ± 2	665 ± 2		
	745 ± 2	748 ± 2		
	777 ± 2	790 ± 3		
	1239 ± 2	1208 ± 2		

[a] Wavenumbers (cm^{-1} units) are determined from Raman spectra of DNA crystals and fibers of known structure.
[b] The corresponding band in structures containing rG is usually observed at 668 ± 1 cm^{-1}.
[c] A weak companion band near 1316 cm^{-1} is also observed in *B* DNA.

Raman bands of the highest intensity. Conversely, out-of-plane ring vibrations yield relatively weak bands (Lord and Thomas, 1967a). (ii) Symmetrical stretching vibrations of the diesterified orthophosphate moiety, namely phosphodioxy (PO_2^-) and phosphodiester (C–O–P–O–C) group vibrations, generate intense Raman bands. On the other hand, antisymmetrical and asymmetrical modes of the orthophosphate moiety are usually too weak to be observed (Shimanouchi, Tsuboi and Kyogoku, 1964). (iii) Vibrations localized in the sugar moieties of DNA (deoxyribose) or RNA (ribose) exhibit very low Raman intensities, unless as noted above, they are coupled to modes of the bases (Erfurth, Kiser and Peticolas, 1972; Lord and Thomas, 1967a; Shimanouchi, Tsuboi and Kyogoku, 1964; Thomas and Hartman, 1973). (iv) On the basis of Raman excitation profiles reported by several groups (Austin, Jordan and Spiro, 1993; Blazej and Peticolas, 1977; Kubasek, Hudson and Peticolas, 1985; Tsuboi, Nishimura and Hirakawa, 1987; Ziegler, 1984), UVRR intensities of the nucleic acid bases are generally greatest when excitation wavelengths in the vicinity of 260 nm are employed. Sugar and phosphate group vibrations are absent from UVRR spectra excited at wavelengths in the range 200–300 nm (Wen and Thomas, 1998).

13.7
Applications

13.7.1
Assembly and Maturation of the Bacteriophage P22 Capsid

The shell of bacteriophage P22 is constructed from 420 copies of a 47-kDa coat protein subunit (gp5) (see also Figure 13.1). Raman spectroscopy has provided details of the self-assembly of gp5 into the viral shell precursor (procapsid) and the subsequent rearrangement of the subunits of the procapsid shell to form the architecture of the mature capsid lattice (Prevelige *et al.*, 1993; Tuma and Thomas, 1997; Tuma, Prevelige and Thomas, 1998b; Tuma *et al.*, 2001). Structural characterization of the initial assembly step, that is, the formation of the procapsid shell from gp5 monomers was made possible by improvements in the sensitivity of Raman instrumentation that allow acquisition of spectra of the monomer at very low concentration (Tuma *et al.*, 2001). Although Raman signatures of gp5 in the monomeric and procapsid shell states share many features, it is clear that procapsid assembly significantly perturbs the structure of the unassembled monomer (Figure 13.5). The computed difference spectrum, as shown in the bottom trace of Figure 13.5, highlights the spectral changes accompanying monomer-to-procapsid assembly. Importantly, the center of the Raman amide I marker is shifted from 1667 cm^{-1} in the monomer to 1663 cm^{-1} in the procapsid shell. This shift gives rise to the Raman difference peaks observed at 1655 cm^{-1} (assigned to α-helix) and 1673 cm^{-1} (β-strand) and corresponding troughs at lower and higher wavenumber values. The difference band pattern reflects a more ordered secondary structure for the subunit in the viral shell than in the monomer at the expense of disordered chain segments. The more highly ordered secondary structure of the shell subunit is confirmed in the difference spectrum in Figure 13.5, by the diminished

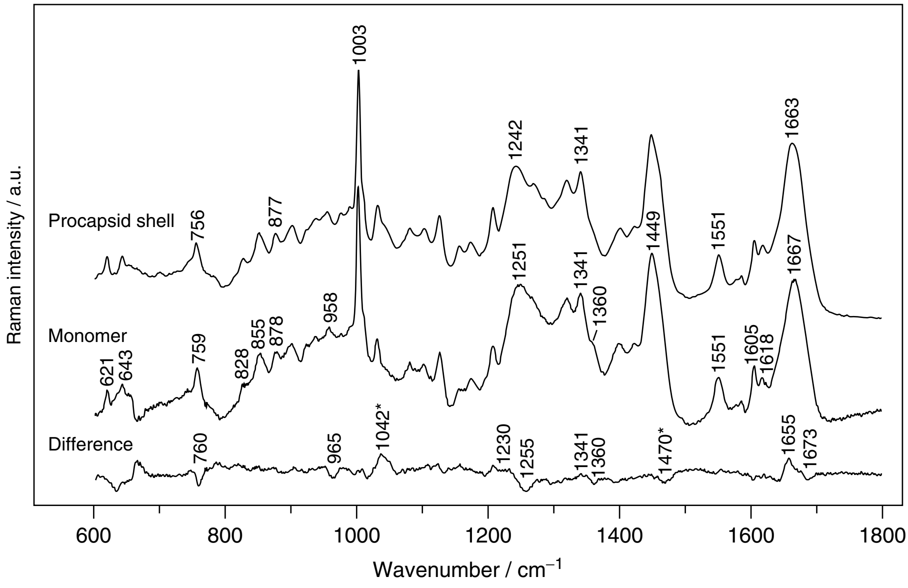

Fig. 13.5 Raman characterization of *in vitro* assembly of the bacteriophage P22 procapsid. Data were obtained with 514 nm excitation at low gp5 concentration (~15 μg/μL) and temperature (~10 °C) (Tuma *et al.*, 2001). The spectral difference profile (bottom trace) between gp5 monomer (middle trace) and shell (top trace) is modest. The amide band peaks and troughs reflect only minor reordering of the polypeptide backbone. Other different features are relatively small but significant, consistent with coupling between subunit folding and assembly. The monomer fold is, nevertheless, largely conserved with shell assembly.

intensity of the Raman amide III band of the monomer (1251 cm^{-1} band center), which leads to a difference trough at 1255 cm^{-1} (diminished irregular chain conformation) and difference peak at 1230 cm^{-1} (enhanced β-strand). These results show that the polypeptide backbone of the gp5 monomer undergoes refolding during procapsid assembly. Overall, about $4 \pm 1\%$ of the backbone, or 17 of 428 peptide bonds, are estimated to undergo refolding. Thus, assembly of the P22 procapsid leads to local reordering of a small but significant percentage of the peptide residues and demonstrates limited coupling of subunit folding with the shell assembly process. The native fold of the gp5 subunit is thus largely conserved with procapsid assembly.

In addition to the changes in subunit secondary structure noted above, the Raman spectrum also provides information on the effect of procapsid assembly on subunit tertiary structure. For example, the shift of the tryptophan (Trp) marker band at 760 cm^{-1} in the monomer to 756 cm^{-1} in the procapsid shell with a concomitant large decrease in intensity signifies a change in the amphipathic environment of the average Trp side chain. Broadening of the 1360 cm^{-1} component of the Trp Fermi doublet indicates that this change is due to a more hydrophilic indolyl environment in the procapsid shell subunit. Overall, the change in average Trp environment is small. Additional Raman markers show that the cysteine (Cys) sulfhydryl and tyrosine (Tyr) phenoxyl hydrogen-bonding environments are essentially invariant to the assembly process. Thus, despite measurable refolding of the coat protein backbone with procapsid assembly, the tertiary structure of the subunit is not greatly affected (Tuma *et al.*, 2001).

The folding state of gp5 in the monomeric and procapsid shell states has also been characterized by deuterium exchange of peptide NH groups (Tuma *et al.*, 2001). Figure 13.6a compares Raman profiles of the exchanged ND groups (amide III′ region, 900–1000 cm^{-1}) of gp5 in monomeric and procapsid shell states, following exposure of each to D_2O (pD 7.8) under native conditions (2 °C for monomer and 10 °C for procapsid shell). Also shown is the amide III′ profile of a gp5 control in which all NH groups were exchanged by deuterium under denaturing conditions prior to refolding of the fully exchanged protein in native D_2O buffer. These data show that NH → ND exchange is more extensive in the monomer than in the procapsid shell subunits. In particular, greater amide III′ intensity is observed at 930–945 cm^{-1} assignable to deuterated α-helix (Overman and Thomas, 1998b; Yu, Lippert and Peticolas, 1973), at 960 cm^{-1} assignable to deuterated disordered peptides (Chen, Lord and Mendelsohn, 1974b; Chen and Lord, 1974a; Tuma, Prevelige and Thomas, 1998b), and at 980–990 cm^{-1} assignable to deuterated β-strand (Chen, Lord and Mendelsohn, 1974b; Chen and Lord, 1974a; Tuma, Prevelige and Thomas, 1998b; Yu, Lippert and Peticolas, 1973). The amide III′ exchange profile of the monomer indicates that the native fold of the monomer comprises both unprotected and protected residues in regions of α-helix, β-strand, and disordered secondary structures. Figure 13.6a also shows extensive exchange protection. Neither the monomer nor procapsid shell subunit undergoes complete NH → ND exchange of α-helical, β-stranded, or disordered regions. The data suggest

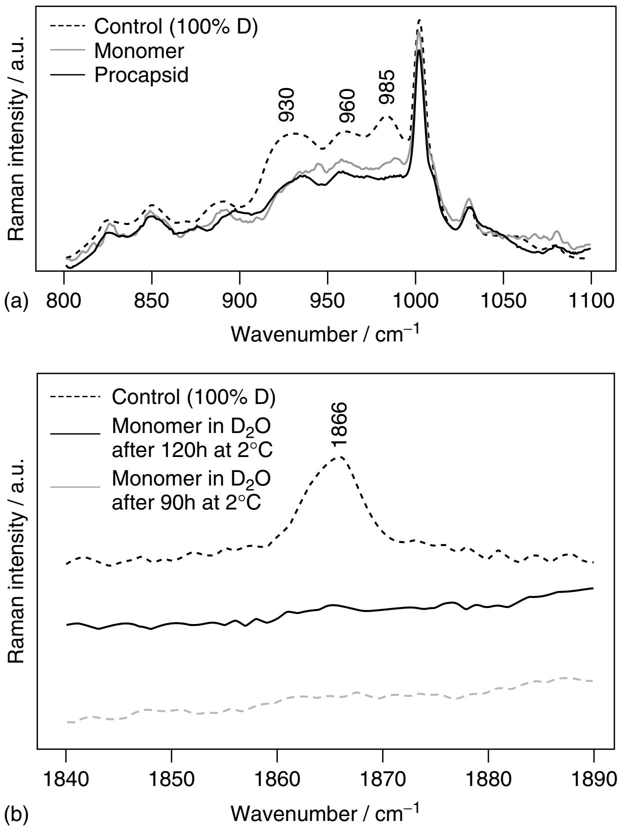

Fig. 13.6 Normalized Raman signatures (514 nm excitation) depicting native-state peptide NH $\rightarrow$ ND exchanges of P22 gp5 in the monomeric state (2 °C) and procapsid shell state (10 °C), each following 12 h exposure to D_2O buffer. (a) Amide III′ profiles of the P22 procapsid, native gp5 monomer and fully exchanged gp5 control. (b) Deuteriocysteinyl S–D stretching bands of the native gp5 monomer, following different conditions of exposure to D_2O (Tuma *et al.*, 2001).

that the most protected residues of the monomer are those located in regions of β-strand structure.

Raman-monitored H/D exchange of selected side chains can be used to identify residues of the protein folding core (Thomas, 1999). For example, the degree of protection of the S–H group of the single cysteine (Cys 405) in the coat protein is a measure of the accessibility of the *C*-terminal region to H/D exchange. Figure 13.6b depicts Raman spectra in the region 1840–1890 cm^{-1}, where the stretching vibration of the deuterated sulfhydryl group (S–D) is expected (Tuma *et al.*, 2001). The absence of an S–D band in the spectrum of the native gp5 monomer shows that the Cys 405 sulfhydryl is fully protected against H/D exchange. Protection persists for samples maintained at native conditions (pD 7.8, 2 °C) for up to 6 days. The S–H group can be exchanged only at conditions that favor global unfolding (40 °C). Thus, the Cys 405 sulfhydryl is buried within the hydrophobic folding core of the gp5 monomer. A similar degree of sulfhydryl exchange protection is observed when gp5 is assembled into the procapsid shell, but not upon maturation to the capsid shell (Tuma, Prevelige and Thomas, 1998b). This shows that the procapsid-to-capsid lattice transformation is accompanied by sequestering of the subunit *C*-terminal region from H/D exchange.

In summary, during P22 shell assembly about 20 peptide linkages of the gp5 monomer undergo local refolding without greatly affecting the global structure of the monomer. Compared to monomeric gp5, the procapsid subunits achieve enhanced protection against peptide H/D exchange. On the other hand, the monomer possesses a robust exchange-protected folding core that encompasses the Cys 405 sulfhydryl. Both the monomeric and procapsid states of the subunit constitute a distinct and compact late-folding intermediate, which ultimately matures to the native state of the capsid. Details of the additional changes in gp5 subunit structure and exchange protection that occur with the final procapsid → capsid lattice transformation (Figure 13.1) have also been quantified by Raman methods and are discussed in detail elsewhere (Tuma, Prevelige and Thomas, 1998b).

13.7.2
Assembly and Maturation of the Bacteriophage HK97 Head

13.7.2.1 Structural Considerations

The assembly pathway of the *E. coli* bacteriophage HK97, which differs fundamentally from that of P22, is shown schematically in Figure 13.7. Icosahedral structures of precursor (*prohead*) and mature (*head*) states of the bacteriophage HK97 shell have been determined by application of electron cryomicroscopy and X-ray crystallography to shells constructed *in vitro* from recombinantly expressed subunits (gp5, 42 kDa) (Wikoff *et al.*, 2000; Wikoff *et al.*, 2006). Mechanistic steps involving the initial assembly of the prohead (Prohead I), its transformation to Prohead II, and eventually its maturation to the Head are not well understood. Of particular interest is the

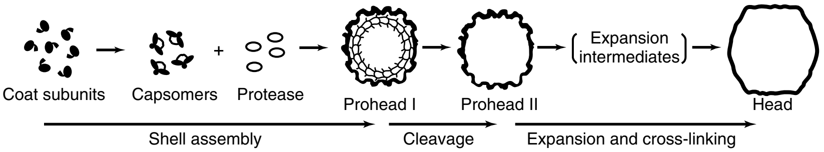

Fig. 13.7 Schematic representation of the *in vitro* assembly pathway of the icosahedral capsid of *E. coli* bacteriophage HK97. This pathway differs fundamentally from that of P22 (Figure 13.1) in that the precursor shell (Prohead I) is assembled from coat the protein monomer (gp5) via an oligomeric intermediate (capsomer) and without the participation of scaffolding protein subunits. Prohead I subsequently submits to a virally encoded protease that cleaves from gp5 the *N*-terminal domain (Δ-domain, residues 1–102) yielding the Prohead II shell. Subsequent shell expansion and subunit cross-linking generate the mature Head architecture (Benevides *et al.*, 2004; Hendrix and Duda, 1998).

structure and role of the gp5 *N*-terminal segment (Δ-domain, residues 1–102), which is cleaved from the Prohead I subunits by a virally encoded protease. This cleavage results in a reorganization of the shell lattice and the Prohead II architecture (Hendrix and Duda, 1998).

Raman spectra of Prohead I and Prohead II particles identify a conserved subunit domain with an α/β-fold, as well as a highly α-helical (74%) Δ-domain specific only to the Prohead I subunit (Figure 13.8a). The diagnostic tryptophan (1339 and 1551 cm^{-1}) and tyrosine (825 and 850 cm^{-1}) Raman markers are identical in subunits of both shells and indicate minimal changes in the tertiary structure of the *C*-terminal segment of gp5 (residues 104–385). However, a shift of the sulfhydryl Raman marker of Cys 362 from 2579 to 2573 cm^{-1} indicates a strengthening of the cysteinyl S–H···X bond with the prohead shell transformation (Benevides *et al.*, 2004).

The Raman signature of the Δ-domain was calculated by subtracting the Prohead II spectrum from the Prohead I spectrum (Figure 13.8b) (Benevides *et al.*, 2004). Interestingly, this difference spectrum exhibits a close correspondence to the Raman spectrum of the P22 scaffolding protein (Tuma and Thomas, 1997) and supports the notion that the Δ-domain fulfills a role in HK97 assembly similar to that of the scaffolding protein in P22 assembly. In contrast, the maturation of Prohead II to the Head involves a significant reordering of the secondary structure of gp5 (residues 104–385), as well as extensive changes in the subunit tertiary structure. The difference spectrum between the Prohead II and Head of HK97 (Figure 13.8c) demonstrates: (i) an increase in ordered secondary structure, (ii) altered interactions of tyrosine and tryptophan side chains, and (iii) additional strengthening of the S–H···X bond of the Cys 362 side chain.

13.7.2.2 Thermodynamic Properties of the Δ-Domain

Recently, Raman spectra of exceptional signal-to-noise quality were obtained from highly purified samples of Prohead I and Prohead II as a function of solution temperature in the interval 10–92 °C (Figure 13.9a and b). Quantitative analysis of the data allows elucidation of the thermostability of the Δ-domain within the framework of the prohead shell assembly pathway (Nemecek and Thomas, unpublished results). For both Prohead I and Prohead II, major spectral changes are observed between 85 and 92 °C, which can be attributed to thermal

Fig. 13.8 (a) Raman spectra of the HK97 Prohead I (top) and Prohead II (bottom) particles. The two spectra have been scaled to match the intensities of environment invariant tyrosine Raman markers at 642 cm^{-1}, as both particles contain the same number of tyrosine residues. (b) The Raman spectrum of the Δ-domain (bottom) obtained by subtracting the Raman spectrum of Prohead II from that of Prohead I. The intensity scale is expanded threefold over that shown in (a). The Raman signature of the P22 scaffolding protein (top), which is shown for comparison with that of the Δ-domain, reveals the fundamental similarities between the two protein structures. (c) Raman spectra of the HK97 mature Head (top) and Prohead II precursor (second from top), each normalized to the intensity of the 1449 cm^{-1} peak. The computed difference spectrum (bottom) is also shown (Benevides *et al.*, 2004).

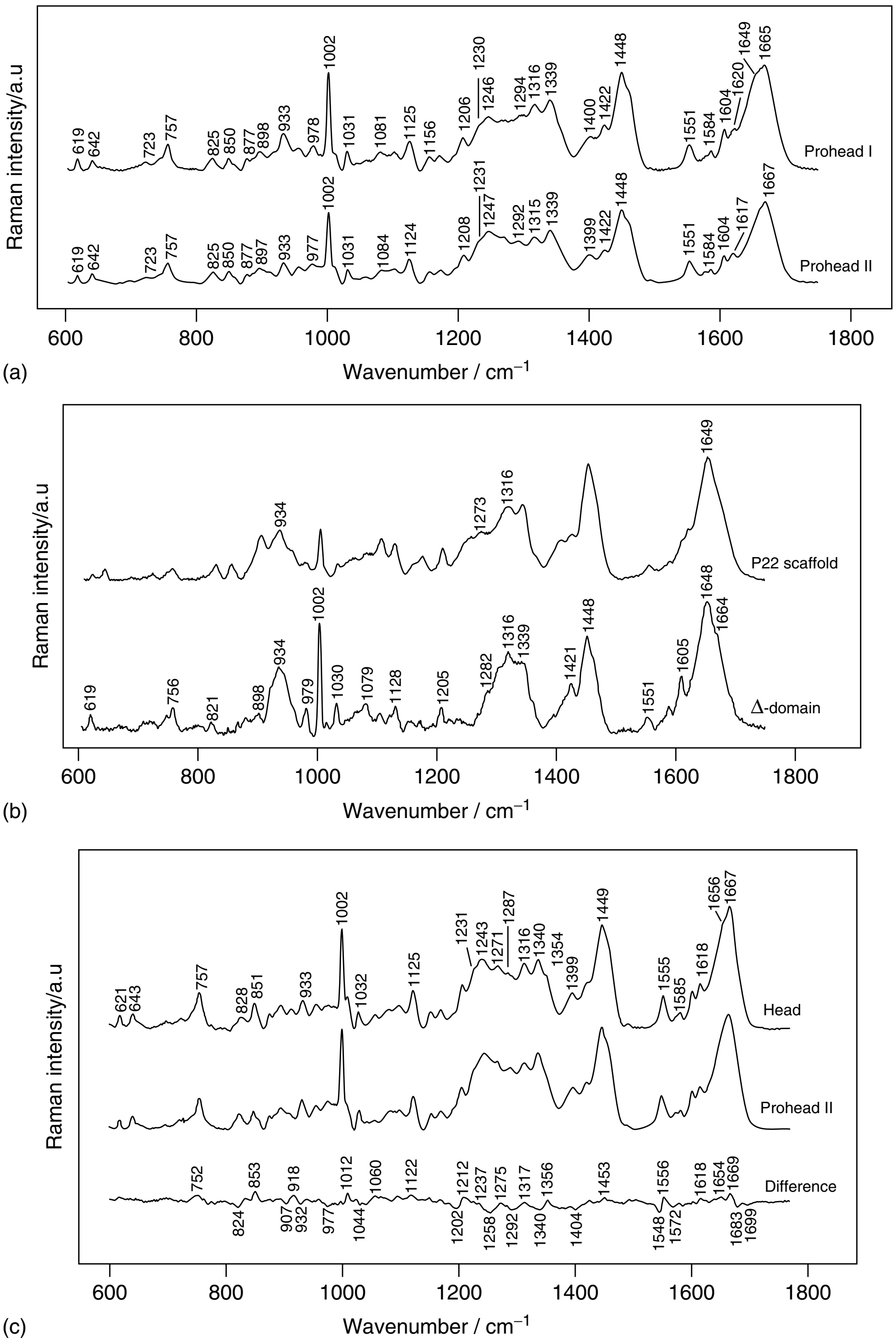
Raman intensity/a.u
619
642
723
757
825
850
877
898
933
978
1002
1031
1081
1125
1156
1206
1230
1246
1294
1316
1339
1400
1422
1448
1551
1584
1604
1620
1649
1665
Prohead I
619
642
723
757
825
850
877
897
933
977
1002
1031
1084
1124
1208
1231
1247
1292
1315
1339
1399
1422
1448
1551
1584
1604
1617
1667
Prohead II
600
800
1000
1200
1400
1600
1800
(a)
Wavenumber / cm^{-1}
Raman intensity/a.u
934
1273
1316
1649
P22 scaffold
619
756
821
898
934
979
1002
1030
1079
1128
1205
1282
1316
1339
1421
1448
1551
1605
1648
1664
Δ-domain
600
800
1000
1200
1400
1600
1800
(b)
Wavenumber / cm^{-1}
Raman intensity/a.u
621
643
757
828
851
933
1002
1032
1125
1231
1243
1271
1287
1316
1340
1354
1399
1449
1555
1585
1618
1656
1667
Head
Prohead II
752
853
918
1012
1060
1122
1212
1237
1275
1317
1356
1453
1556
1618
1654
1669
Difference
824
907
932
977
1044
1202
1258
1292
1340
1404
1548
1572
1683
1699
600
800
1000
1200
1400
1600
1800
(c)
Wavenumber / cm^{-1}

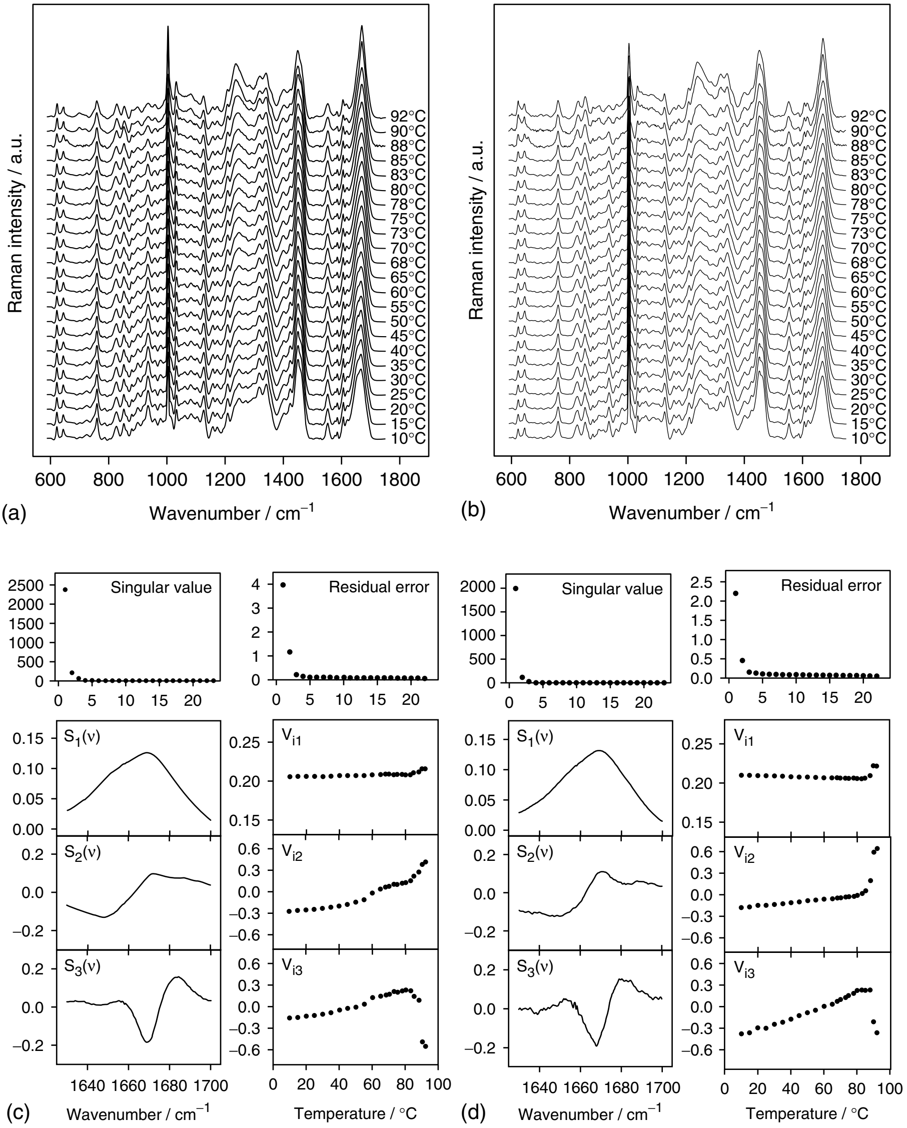

Fig. 13.9 Raman spectra of the bacteriophage HK97 Prohead I (a) and Prohead II (b) shells were collected over the temperature interval $10 < t < 92\,^\circ$C. Changes of secondary structure were analyzed by factor analysis (SVD algorithm) of the amide I conformational marker (1630–1700 cm^{-1}) of Prohead I (c) and Prohead II (d). Each spectral set can be described within the experimental noise level by three independent components. The scores V_{i2} of the second component manifest two distinct structural transitions in Prohead I, while Prohead II exhibits only one transition. For both Prohead I and Prohead II the high-temperature transition (~89 °C) corresponds to melting (thermal denaturation of the subunit). The lower temperature transition in Prohead I (~60 °C) represents unfolding of the Δ-domain.(Němeček and Thomas, manuscript in preparation).

denaturation of the subunits that are released upon shell disassembly. Spectral changes between 10 and 85 °C, on the other hand, reflect initial partial unfolding events in the shell subunits followed by the onset of shell disassembly. In the case of Prohead I, the observed changes include the effects of unfolding of the Δ-domain. The latter were deciphered and quantified by factor analysis of the Prohead I and Prohead II spectral series using the SVD algorithm. Figure 13.9c and d shows the factor analysis results for the amide I bands of Prohead I and Prohead II, inclusive of the third SVD component (i.e. V_{i3} terms). The contribution of the second component (V_{i2} term) to the respective parent spectrum reveals two sigmoidal transitions for Prohead I, but only a single transition for Prohead II. The first transition for Prohead I can be attributed to unfolding of the Δ-domain, while the second likely corresponds to Prohead I disassembly. Only the second transition is observed for Prohead II, consistent with the absence of the Δ-domain from the Prohead II subunit. Both parent spectral series can be approximated within the precision of the observed noise level by three spectral components $[S_1(\nu), S_2(\nu), S_3(\nu)]$. Consequently, each spectral series is said to contain three independent spectral species. For Prohead I, these are (i) the natively folded, (ii) Δ-domain unfolded, and (iii) globally unfolded shell subunit (residues 1–385). For Prohead II, these are (i) the natively folded, (ii) locally unfolded, and (iii) globally unfolded shell subunit (residues 104–385). Species (ii) of the Prohead II spectral series represents only a small structural change in comparison to that of species (ii) of Prohead I, which is reflected by a 50% smaller singular value of the second SVD component. Thermodynamic parameters that describe the stability of the Δ-domain in Prohead I were obtained by fitting V_{ij} scores of the significant components. A specific thermodynamic model for Δ-domain unfolding has been developed from these results (Nemecek *et al.*, 2008).

Raman spectra of the independent spectral species considered in the SVD analysis were also derived from the fitted parameters. It can be concluded from these studies that the Prohead I assembly confers structural thermostability on the Δ-domain and enhances the cooperativity of Δ-domain unfolding. This is accomplished by a secondary structural fold – specifically α-helix – that is converted to a β-strand conformation during Δ-domain unfolding. In a forthcoming publication these results will be further discussed (Nemecek *et al.*, 2008).

13.7.3

Structural Studies of Filamentous Bacteriophages

All filamentous bacteriophages are cylindrically shaped, flexible particles (~6 nm diameter and up to 2000 nm in length) that comprise an ss and covalently closed DNA genome packaged within a thin protein shell (capsid). The ssDNA molecule typically constitutes a small percentage (5–15%) of the total virion mass, while the capsid consists of a highly symmetrical arrangement of several thousand copies of a small subunit. The capsid of the filamentous bacteriophage Ff consists of approximately 2700 copies of the major coat protein (pVIII) and a few copies of minor proteins at the filament ends. Assembly of Ff takes place within the membrane of the *E. coli* host cell (Figure 13.2). The overall assembly architecture of Ff (strain fd) is depicted schematically in Figure 13.10 (bottom left panel). Also shown in Figure 13.10 is an electron micrograph of a single negatively stained fd particle (upper left panel) and a three-dimensional reconstruction of a small segment of the capsid surface, which was generated from images obtained by electron cryomicroscopy (right panel). Fiber X-ray diffraction and solid-state NMR studies of highly oriented specimens of fd provide useful information about the symmetry of the capsid subunit arrangement and are consistent with a predominantly α-helical secondary structure for the pVIII capsid subunit (Marvin *et al.*, 2006; Thiriot *et al.*, 2004; Zeri *et al.*, 2003).

Raman spectroscopy has been used extensively over the past several decades to probe further details of the molecular architectures of fd and other filamentous bacteriophages of the same (class I) and different (class II) capsid symmetries. Methods of off-resonance Raman, UVRR, and polarized Raman have all been exploited effectively for this purpose. Details of pVIII side chain orientations and hydrogen-bonding interactions, as well as information about the conformation and organization of the packaged ssDNA within the viral capsids have been obtained. Comprehensive surveys of the Raman studies of fd and other filamentous bacteriophages have been given elsewhere (Benevides, Overman and Thomas, 2005a; Tsuboi and Thomas, 2007b). Here, we consider a few representative applications.

13.7.3.1 Conformations of Packaged ssDNA

High resolution structural methods, such as X-ray crystallography and NMR solution spectroscopy, have not been applied successfully to the highly flexible filamentous virus particles. Additionally, the low DNA content of filamentous

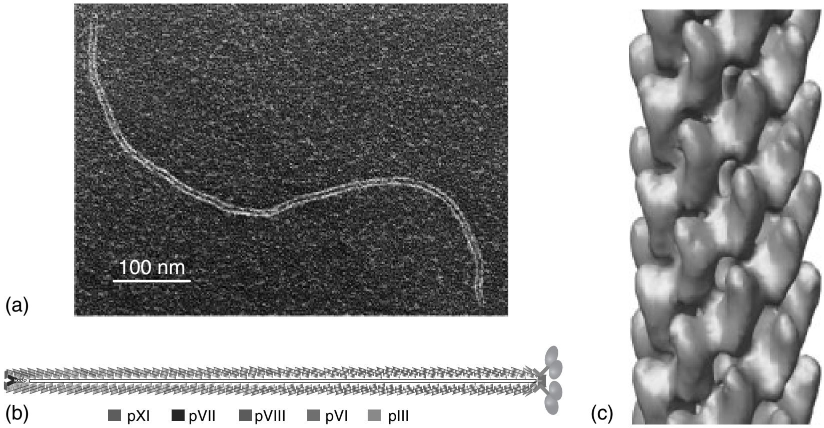

Fig. 13.10 The filamentous ssDNA bacteriophage fd. (a) A negatively stained electron micrograph showing one complete virion. The flexibility of the virion, which is apparent, is consistent with the experimentally determined equilibrium persistence length of 1.0 µm. (b) Cartoon of fd (not to scale) showing locations of the minor coat proteins. (c) Reconstruction of a 24 nm segment of fd at ~8 Å resolution from images obtained by electron cryomicroscopy (courtesy of E. H. Egelman, University of Virginia). (Please find a color version of this figure on the color plates.)

virions precludes definitive determination of conformational properties of the packaged ssDNA from available fiber X-ray diffraction and solid-state NMR data. Raman spectroscopy, on the other hand, provides a convenient approach for probing the structure and organization of the packaged ssDNA in filamentous viruses (Thomas and Murphy, 1975; Thomas *et al.*, 1988a). In the well-studied phages fd, Pf1, Pf3, and PH75, Raman markers of the deoxynucleosides (Tables 13.2 and 13.3) indicate, respectively, C3′-*endo*/*anti* (665 cm^{-1}) (Thomas *et al.*, 1988a; Wen, Overman and Thomas, 1997), C2′-*endo*/*anti* (682 cm^{-1}) (Thomas *et al.*, 1988a; Wen, Armstrong and Thomas, 1999), C3′-*endo*/*anti* (663 cm^{-1}) (Thomas *et al.*, 1988a; Wen *et al.*, 2001), and C2′-*endo*/*anti* (679 cm^{-1}) (Overman *et al.*, 2005) conformations. In a related application, interactions of packaged ssDNA genomes of fd and Pf1 with metal ions have also been probed by Raman methods (Day *et al.*, 1988).

In combination with the Raman tensors determined for local DNA vibrations (Thomas *et al.*, 1995), polarized Raman microspectroscopy has been applied to oriented fibers of the filamentous bacteriophages fd, Pf1, Pf3, and PH75 to reach conclusions about the orientations of ssDNA subgroups (Tsuboi *et al.*, 2003b, 2003a, 2005). The polarized Raman measurements show that while the bases of the packaged fd and Pf1 ssDNA genomes are nonrandomly oriented within their respective capsids, those of PH75 ssDNA lack uniformity of orientation (Benevides, Overman and Thomas, 2005b). Polarized Raman results on Pf1 DNA also show that the average DNA phosphodioxy group (PO_2^-) is oriented so that the line connecting the oxygen atoms ($O \cdots O$) forms an angle of 71° ± 5° with the virion axis (Tsuboi *et al.*, 2003a).

In the off-resonance Raman spectra of filamentous bacteriophages, Raman bands of the ssDNA genome are relatively weak in comparison to those of the capsid protein, owing to the low DNA content of the virion. Further, the ssDNA bands are extensively overlapped by those of the protein, especially in the 1200–1700 cm^{-1} region. The intensities of the DNA Raman bands can be selectively enhanced through the UVRR mechanism by use of 257-nm laser excitation. This approach, which has been pursued for filamentous bacteriophages fd, Pf1, Pf3, and PH75, (Overman *et al.*, 2005; Wen, Overman and Thomas, 1997; Wen, Armstrong and Thomas, 1999; Wen *et al.*, 2001) provides an effective means of deciphering spectral contributions and structural properties of the packaged ssDNA. The Raman wavenumbers and intensities (hypochromicities) of UVRR bands of the bases in the 1200–1700 cm^{-1} interval are uniquely informative of the base environments and their interactions within the viral capsid (Wen and Thomas, 1998).

The following procedure has been employed to evaluate and interpret UVRR hypochromicities of selected base vibrations of the packaged ssDNAs of filamentous phages. The scattering cross-section (σ_n) of a Raman band of wavenumber ν_n is obtained by comparison of its peak height (I_n) with the peak height (I_s) of an internal standard (ν_s) of known scattering cross-section (σ_s) by use of Eq. (2):

$$\sigma_n = \sigma_s(I_n/I_s)(C_s/C_n) \times [(\nu_0 - \nu_s)/(\nu_0 - \nu_n)]^4 \qquad (2)$$

where C_s and C_n are the molar concentrations of the internal standard and target species, respectively, and ν_0 is the wavenumber of the exciting line, expressed in cm^{-1} (The prominent UVRR band of sodium sulfate at 981 cm^{-1}, for which $\sigma_s = 0.16$ mb at 257 nm, is employed as the internal standard (Fodor *et al.*, 1989).). In applying Eq. (2) to a deoxynucleoside constituent of ssDNA, C_n is replaced by C_b ($\equiv X_b C_n$, where X_b is the base mole fraction). The fractional hypochromicity of an ssDNA Raman band at wavenumber ν_n is defined as $1 - \gamma_n$, where the hypochromic ratio, γ_n, is the quotient of intensities in Raman spectra of the nucleic acid and constituent deoxynucleosides. In terms of the corresponding Raman scattering cross-sections,

$$\gamma_n^{\mathrm{DNA}} \equiv I_n^{\mathrm{DNA}} / I_n^{\mathrm{nuc}} = \sigma_n^{\mathrm{DNA}} / \sigma_n^{\mathrm{nuc}} \quad (3)$$

where σ_n^{DNA} and σ_n^{nuc} are Raman scattering cross-sections for the band in DNA and in a solution of the free deoxynucleosides, respectively.

The calculated UVRR cross-sections and measured hypochromic ratios for selected UVRR bands of the packaged ssDNAs of fd, Pf1, Pf3, and PH75 and comparison with dsDNA are given in Table 13.4. Representative data (for Pf3) are shown in Figure 13.11. These results are based on the assumption that for each DNA a mixture of its free deoxynucleosides would be the equivalent of a completely unstacked arrangement of the bases ($\gamma_n = 1$). In brief, the results show that the bases of Pf1 exhibit very large hypochromic ratios, indicating minimal stacking of the bases in the packaged Pf1 genome. The bases of Pf3 exhibit slightly more stacking than those of Pf1. In contrast, the hypochromicities observed for the packaged ssDNAs of fd and PH75 indicate that these genomes have very highly stacked arrangements of the bases. Surprisingly, the hypochromic effects for fd and PH75 exceed that of dsDNA, implying even closer base stacking than in *B* DNA.

13.7.3.2 **Structural Details of the Capsid Subunits**

Knowledge of the fd viral capsid symmetry and Raman tensors for appropriate vibrational modes of the pVIII subunits, many of which have been independently determined from polarized Raman studies of model compounds of known structure, permit assessment of subunit orientations with respect to the filament axis. For example, knowledge of the amide I Raman tensor (Tsuboi, Ikeda and Ueda, 1991) enables the polarized Raman spectrum of phage fd to be interpreted in terms of the angle of orientation of the subunit α-helix axis with respect to the virion axis ($16 \pm 4^\circ$) (Overman, Tsuboi and Thomas, 1996). Similar protocols were applied to the polarized Raman spectra of the class II bacteriophages Pf1, Pf3, and PH75 (Tsuboi *et al.*, 2000, 2003b, 2003a, 2005). Similarly, polarized Raman methods can be exploited to determine orientations of key tryptophan and tyrosine side chains in the subunits of both class I and class II filamentous viruses (Tsuboi, Overman and Thomas, 1996a; Tsuboi *et al.*, 1996b, 2001, 2003b, 2005). Representative polarized Raman data are shown in Figure 13.12. Details of the experimental protocols and data interpretation have been reviewed elsewhere (Tsuboi and Thomas, 2007b).

Analysis of the polarized Raman amide I and amide III signatures of the filamentous virus PH75 reveals that

Tab. 4 UVRR cross-sections (σ) and hypochromic ratios (γ) for DNA bands of fd, Pf1, Pf3 and PH75.

Band	**fd virus**			**Pf1 virus**			**Pf3 virus**			**PH75 virus**			**ssDNA**			**dsDNA**		
	σ^{DNA}	σ^{nuc}	γ	σ^{DNA}	σ^{nuc}	γ	σ^{DNA}	σ^{nuc}	γ	σ^{DNA}	σ^{nuc}	γ	σ^{DNA}	σ^{nuc}	γ	σ^{DNA}	σ^{nuc}	γ
1335	83	395	0.21	222	381	0.58	47	75	0.63	135	469	0.29	138	395	0.35	97	363	0.27
1485	60	244	0.25	198	230	0.86	67	92	0.73	85	277	0.31	124	244	0.51	84	220	0.38
1575	17	70	0.24	80	97	0.82	31	42	0.75	93	302	0.31	38	70	0.54	22	74	0.44
1650	32	110	0.29	200	120	1.70	19	28	0.67	36	116	0.31	39	110	0.36	26	110	0.23

The UVRR bands (left column) are in cm^{-1} units. Raman cross-sections for residues of the packaged viral ssDNA (σ^{DNA}) and corresponding free nucleoside (σ^{nuc}) are given in units of mbarn and hypochromic ratios are as defined in Eq. (3) of the text (Overman *et al.*, 2004; Wen, Overman, and Thomas, 1997; Wen, Armstrong, and Thomas, 1999; Wen *et al.*, 2001).

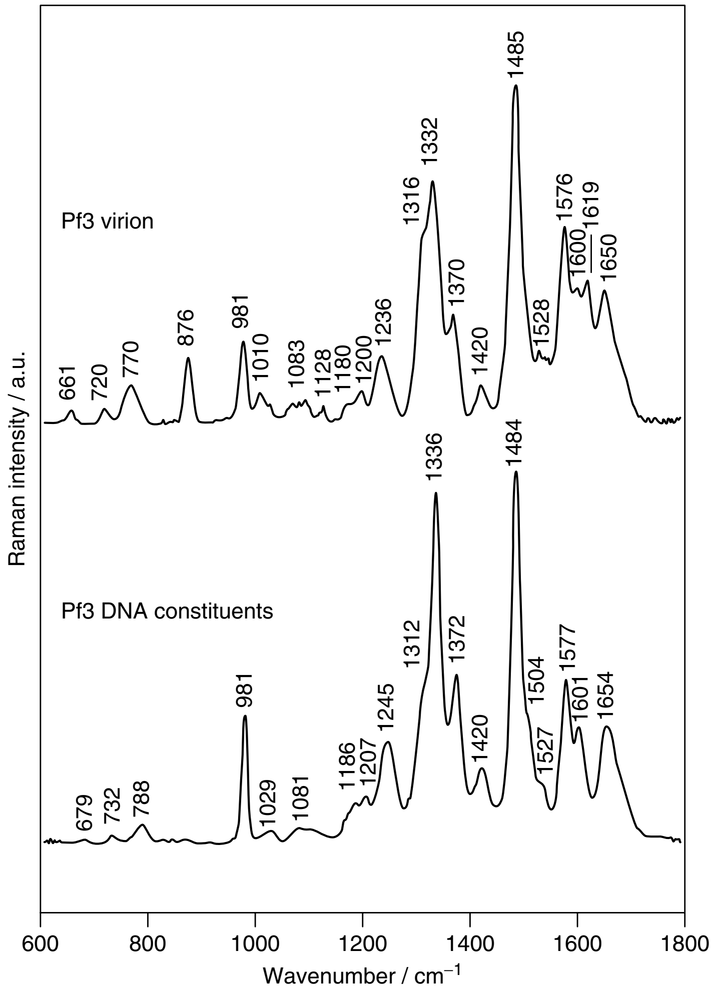

Fig. 13.11 UVRR spectra (600–1800 cm^{-1}, 257 nm excitation) of Pf3 (top) and a deoxynucleoside mixture having the same base composition (20% dA, 35% dT, 24% dG, 21% dC) as Pf3 DNA (bottom). Concentrations of the solutions of Pf3 and deoxynucleoside mixture in terms of moles of DNA phosphate are 1.04 and 0.80 mM, respectively. The concentrations of the Na_2SO_4 internal standard in these solutions are 91 and 200 mM, respectively (Wen *et al.*, 2001).

while the capsid subunit is predominantly α-helical (87%), it also contains a significant number of residues (6 ± 1 or 13 ± 3 %) differing from the canonical α-helix. This minor structural component is not apparent in the capsid subunits of the other filamentous phages (fd, Pf1, and Pf3) examined by polarized Raman spectroscopy. The Raman spectrum of PH75 also manifests local environments

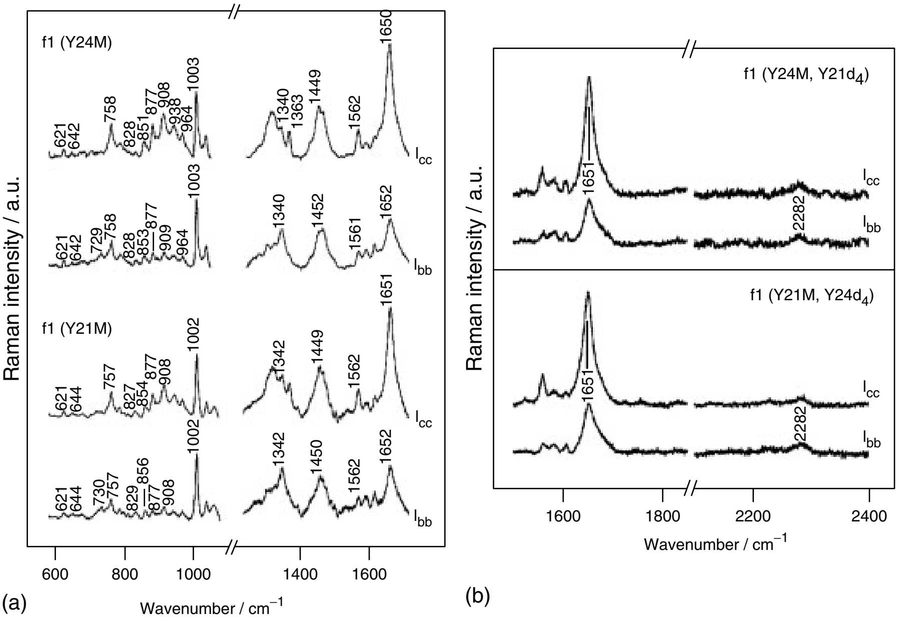

Fig. 13.12 (a) Polarized Raman spectra (I_{cc} and I_{bb}) in the spectral regions 600–1100 and 1200–1700 cm^{-1} of oriented fibers of two variants of the Ff virus (strain fl), incorporating the single-site mutations Tyr 24 → Met [fl (Y24M), top two traces] and Tyr 21 → Met [fl (Y21M), bottom two traces] in their respective capsid subunits. Raman intensities are normalized to the intensity of the isotropic marker band at 1002 cm^{-1} (Phe). (b) Polarized Raman spectra (I_{cc} and I_{bb}) in the spectral regions 1600–1800 and 2200–2400 cm^{-1} of oriented fibers of the same Ff variants in (a), which additionally carry tetradeuterio isotopomers of Tyr 21 [fl (Y24M,Y21d$_4$)] and Tyr 24 [fl (Y21M,Y24d$_4$)]in their respective capsid subunits. Accordingly, the Raman bands in the C–D stretching region are markers of the individual tyrosines (Tsuboi, 2001).

for alanine, tryptophan, and tyrosine side chains that are distinct from those observed for fd, Pf1, or Pf3. These structural anomalies were proposed to contribute to the high thermostability of the PH75 virion (Tsuboi *et al.*, 2005). Interestingly, the polarized Raman spectra of PH75 also reveal that the average peptide group of the PH75 capsid subunit is oriented with greater angular displacement from the virion axis (25°) than peptide groups of fd, Pf1, or Pf3 subunits (16°), which the indolyl ring of Trp 37 in the PH75 capsid subunit projects nearly equatorially from the α-helix axis in contrast to the more axial orientations that occur for Trp 26 of fd and Trp 38 of Pf3, and that the phenolic rings of Tyr 15 and Tyr 39 of PH75 project along the subunit α-helix axis such that one phenoxyl engages in hydrogen-bonding interaction that has no counterpart in the tyrosines of either fd or Pf1.

Analysis of the contributions of subunit tyrosine side chains (Tyr 21 and Tyr 24) to the polarized Raman spectrum of fd provides a good example of the value of the method for structural investigations of viruses. Here, single-site mutations of Tyr 21 and Tyr 24 (to methionine, Met) in the capsid subunits of fd were exploited to determine independently the orientation of each tyrosine. The mutants were also subjected to deuterium isotope-editing of the nonmutated tyrosyl ring (Overman and Thomas, 1995) in order to introduce novel tyrosyl Raman markers (C−D stretching bands in the 2200–2400 cm^{-1} region) amenable to polarized Raman analysis. The required Raman tensors for the C−D stretching bands were determined by analysis of the polarized Raman spectra obtained from an oriented single crystal of the model compound L-tyrosine-2,3,4,5-d_4 (Tsuboi *et al.*, 1998). The rationale for use of the C−D modes is based upon the fact that the corresponding Raman bands occur in a spectral region free of interference from any other Raman bands; also, the highly localized nature of the C−D stretching vibrations is considered favorable to transferability of the Raman tensors from the tyrosyl model compound to the viral subunit tyrosines. The procedure, results, and interpretation are summarized in Figure 13.13, which illustrates the coordinate systems defined for the virion and local tyrosyl Raman tensors [panel (a)], the nature of the experimental data [panel (b)], and the molecular model derived for two neighboring capsid subunits from the polarized Raman analysis (Tsuboi *et al.*, 2001) and complementary structural studies [panel (c)] (Marvin *et al.*, 1994; Overman and Thomas, 1995; Overman, Tsuboi and Thomas, 1996; Symmons *et al.*, 1995)].

13.8
Concluding Summary

Raman, UV-resonance Raman and polarized Raman spectroscopy are valuable methods for investigating structures of viral proteins and nucleic acids. The methods are also well suited to probe

Fig. 13.13 (a) The drawing at the left shows the coordinate system for the uniaxially oriented fd fiber (*abc* coordinate system) in relation to the α-helical subunit and a tyrosine side chain for which a representative Raman tensor coordinate system (*xyz*) is indicated. The drawing (not to scale) represents a small segment of the ~880 nm fd virion and a few α-helical turns of one subunit of the capsid. The drawing at the right indicates the Eulerian angles θ and χ, which define the orientation of the *xyz* coordinate system with respect to the *abc* system. Here, θ is the tilt angle (c−O−z) between the Raman tensor principal axis z and the fiber axis c, while χ is the angle (y−O−N) formed by the Raman tensor principal axis y and the line of intersection between the plane (*ab*) normal to the fiber axis and the plane (*xy*) of the phenolic ring of tyrosine. (b) Polarized Raman spectra (I_{cc}, I_{bb}) in the region 1500–2400 cm^{-1} of oriented fibers of the deuterium-labeled mutant fd viruses [designated here as f1 (Y21M;Y24d_4) and f1 (Y24M;Y21d_4) to represent virions in which, respectively, Tyr 21 is mutated to Met while Tyr 24 contains the deuterium labeling, and Tyr 24 is mutated to Met while Tyr 21 contains the deuterium labeling]. (c) Stereo diagram of the energy-minimized structure of neighboring subunits in the assembly model proposed for fd on the basis of the atomic coordinates developed for Tyr 21 and Tyr 24 from the polarized Raman spectra of panel (b) and coordinates for other residues as previously described (Marvin *et al.*, 1994; Overman, Tsuboi, and Thomas, 1996; Symmons *et al.*, 1995; Tsuboi, 1996a, 2001). The virion axis runs vertically and the subunit *N*-termini are at the top. The subunits shown correspond to index numbers $k = 0$ and $k = 5$, as defined in the Protein Databank Accession Code (Entry) 1IFJ. The molecular graphics were obtained using MOLSCRIPT (Kraulis, 1991).

the assembly pathways and architectures of native viruses. The sensitivity and selectivity shared by these spectroscopic approaches are well exemplified by the applications reviewed here to icosahedral and filamentous viruses and their constituent molecules. Significant developments in Raman instrumentation and data processing, coupled with the diversity of available sample-handling procedures and molecular biological techniques, have vastly improved the scope of Raman methods for studies not only of viruses but also of other biological macromolecular assemblies.

In the applications considered above, it must be stressed that accurate Raman vibrational band assignments are

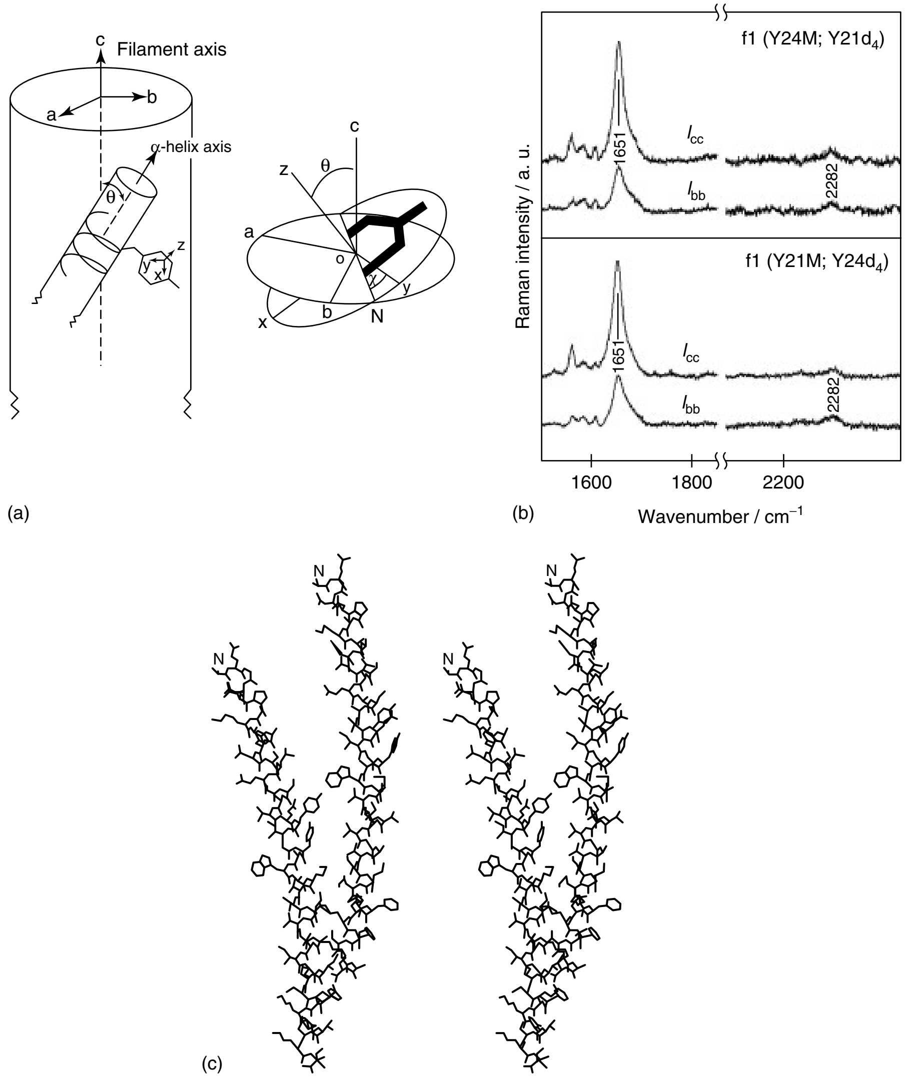
c
Filament axis
a
b
α-helix axis
θ
z
y
x
c
z
θ
a
o
χ
y
x
b
N
(a)
f1 (Y24M; Y21d$_4$)
I_{cc}
I_{bb}
1651
2282
f1 (Y21M; Y24d$_4$)
I_{cc}
I_{bb}
1651
2282
Raman intensity / a. u.
1600
1800
2200
Wavenumber / cm^{-1}
(b)
N
N
N
N
(c)

essential. Only with reliable assignments can the Raman spectrum be effectively exploited to investigate molecular structural details, including sites of intermolecular recognition and local and global changes in structure attendant with biomolecular recognition events. We have shown how the use of isotope-editing procedures and recombinantly engineered mutational substitutions can be used to advance Raman vibrational assignments that lead to definitive interpretations of the spectroscopic data. Statistical treatment of the spectra is also shown to facilitate quantitative analysis of physicochemical properties and provide improved understanding of the biological significance of the results.

Acknowledgments

We thank Dr Stacy A. Overman for assistance with the preparation of illustrations and proofreading. Research support from the National Institutes of Health (Grant GM50776 to GJT) is gratefully acknowledged.

References

Arp, Z., Autrey, D., Laane, J., Overman, S.A., and Thomas Jr, G.J. (2001) *Biochemistry*, **40**, 2522–29.

Asher, S.A., Ianoul, A., Mix, G., Boyden, M.N., Karnoup, A., Diem, M., and Schweitzer-Stenner, R. (2001) *J. Am. Chem. Soc.*, **123**, 11775–81.

Aubrey, K.L., Casjens, S.R., and Thomas Jr, G.J. (1992) *Biochemistry*, **31**, 11835–42.

Austin, J.C., Jordan, T., and Spiro, T.G. (1993) Ultraviolet Resonance Raman Studies of Proteins and Related Compounds in *Biomolecular Spectroscopy, Part A* (eds R.J.H. Clark and R.E. Hester) John Wiley and Sons, New York, pp. 55–127.

Bandekar, J. (1992) *Biochim. Biophys. Acta*, **1120**, 123–43.

Benevides, J.M., Bondre, P., Duda, R.L., Hendrix, R.W., and Thomas Jr, G.J. (2004) *Biochemistry*, **43**, 5428–36.

Benevides, J.M., Juuti, J.T., Tuma, R., Bamford, D.H., and Thomas Jr, G.J. (2002) *Biochemistry*, **41**, 11946–53.

Benevides, J.M., Overman, S.A., and Thomas Jr, G.J. (2005a), Raman spectroscopy of proteins, in *Current Protocols in Protein Science* (eds J.E. Coligan, B.M. Dunn, H.L. Ploegh, D.W. Speicher, and P.T. Wingfield), pp. 17.8.1–17.8.35, John Wiley & Sons, New York.

Benevides, J.M., Overman, S.A., and Thomas, G.J. (2005b) *J. Raman Spectrosc.*, **36**, 279–99.

Benevides, J.M., Tsuboi, M., Wang, A.H.J., and Thomas Jr, G.J. (1993) *J. Am. Chem. Soc.*, **115**, 5351–59.

Berjot, M., Marx, J., and Alix, A.J.P. (1987) *J. Raman Spectrosc.*, **18**, 289–300.

Blanch, E.W., Hecht, L., Syme, C.D., Volpetti, V., Lomonossoff, G.P., Nielsen, K., and Barron, L.D. (2002) *J. Gen. Virol.*, **83**, 2593–600.

Blazej, D.C. and Peticolas, W.L. (1977) *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 2639–43.

Caille, J.P., Pigeon-Gosselin, M., and Pezolet, M. (1983) *Biochim. Biophys. Acta*, **758**, 121–27.

Carey, P.R. (2006) *Annu. Rev. Phys. Chem.*, **57**, 527–54.

Casjens, S. and Weigele, P. (2005) DNA packaging by bacteriophage P22 in *Viral Genome Packaging Machines: Genetics, Structure, and Mechanism* (ed C.E. Catalano) Landes Bioscience/Eurekah.com and Kluwer Academic/Plenum Publishers, Georgetown, TX and New York, pp. 80–88.

Chen, M.C. and Lord, R.C. (1974a) *J. Am. Chem. Soc.*, **96**, 4750–52.

Chen, M.C., Lord, R.C., and Mendelsohn, R. (1974b) *J. Am. Chem. Soc.*, **96**, 3038–42.

Combs, A., McCann, K., Autrey, D., Laane, J., Overman, S.A., and Thomas Jr, G.J. (2005) *J. Mol. Struct.*, **735**(36), 271–78.

Day, L.A., Casadevall, A., Prescott, B., and Thomas Jr, G.J. (1988) *Biochemistry*, **27**, 706–11.

Edsall, J.T., Otvos, J.W., and Rich, A. (1950) *J. Am. Chem. Soc.*, **72**, 474–77.

Erfurth, S.C., Kiser, E.J., and Peticolas, W.L. (1972) *Proc. Natl. Acad. Sci. U.S.A.*, **69**, 938–41.

Fodor, S.P.A., Copeland, R.A., Grygon, C.A., and Spiro, T.G. (1989) *J. Am. Chem. Soc.*, **111**, 5509–18.

Garfinkel, D. (1958) *J. Am. Chem. Soc.*, **80**, 3827–31.

Garfinkel, D. and Edsall, J.T. (1958a) *J. Am. Chem. Soc.*, **80**, 3807–12.

Garfinkel, D. and Edsall, J.T. (1958b) *J. Am. Chem. Soc.*, **80**, 3818–23.

Garfinkel, D. and Edsall, J.T. (1958c) *J. Am. Chem. Soc.*, **80**, 3823–26.

Hanus, J., Barvik, I., Ruszova-Chmelova, K., Stepanek, J., Turpin, P.Y., Bok, J., Rosenberg, I., and Petrova-Endova, M. (2001) *Nucleic Acids Res.*, **29**, 5182–94.

Hanus, J., Nemecek, D., Stepanek, J., and Turpin, P.-Y. (2004) *J. Raman Spectrosc.*, **35**, 418–25.

Hanus, J., Stepanek, J., Turpin, P.-Y., and Bok, J. (1999) *J. Mol. Struct.*, **480**(81), 437–42.

Harada, I., Miura, T., and Takeuchi, H. (1986) *Spectrochim. Acta*, **42A**, 307–12.

Harada, I., Takamatsu, T., Tasumi, M., and Lord, R.C. (1982) *Biochemistry*, **21**, 3674–77.

Hartman, K.A., Clayton, N., and Thomas Jr, G.J. (1973) *Biochem. Biophys. Res. Commun.*, **50**, 942–49.

Helfand, M.S., Taracila, M.A., Totir, M.A., Bonomo, R.A., Buynak, J.D., van den, A.F., and Carey, P.R. (2007) *Biochemistry*, **46**, 8689–99.

Hendrix, R.W. and Duda, R.L. (1998) *Adv. Virus Res.*, **50**, 235–88.

Jenkins, R.H., Tuma, R., Juuti, J.T., Bamford, D.H., and Thomas Jr, G.J. (1999) *Biospectroscopy*, **5**, 3–8.

Kraulis, P.J. (1991) *J. Appl. Crystallogr.*, **24**, 946–50.

Kubasek, W.L., Hudson, B., and Peticolas, W.L. (1985) *Proc. Natl. Acad. Sci. U.S.A.*, **82**, 2369–73.

Lewis, I.R. and Edwards, H.G.M. (2001) *Handbook of Raman Spectroscopy: From the Research Laboratory to the Process Line*, Marcel Dekker, New York, USA.

Li, T., Chen, Z., Johnson, J.E., and Thomas Jr, G.J. (1992b) *Biochemistry*, **31**, 6673–82.

Li, H., Hanson, C., Fuchs, J.A., Woodward, C., and Thomas Jr, G.J. (1993a) *Biochemistry*, **32**, 5800–8.

Li, T., Johnson, J.E., and Thomas Jr, G.J. (1993b) *Biophys. J.*, **65**, 1963–72.

Li, H. and Thomas, G.J. (1991) *J. Am. Chem. Soc.*, **113**, 456–62.

Li, H., Wurrey, C.J., and Thomas Jr, G.J. (1992a) *J. Am. Chem. Soc.*, **114**, 7463–69.

Lippert, J.L., Tyminski, D., and Desmeules, P.J. (1976) *J. Am. Chem. Soc.*, **98**, 7075–80.

Lord, R.C. and Thomas Jr, G.J. (1967a) *Spectrochim. Acta*, **23A**, 2551–91.

Lord, R.C. and Thomas Jr, G.J. (1967b) *Biochim. Biophys. Acta*, **142**, 1–11.

Lord, R.C. and Thomas Jr, G.J. (1968) *Dev. Appl. Spectrosc.*, **6**, 179–99.

Lord, R.C. and Yu, N.T. (1970a) *J. Mol. Biol.*, **50**, 509–24.

Lord, R.C. and Yu, N.T. (1970b) *J. Mol. Biol.*, **51**, 203–13.

Malinowski, E.R. (2002) *Factor Analysis in Chemistry*, John Wiley & Sons, Inc., New York, USA.

Marvin, D.A., Hale, R.D., Nave, C., and Helmer-Citterich, M. (1994) *J. Mol. Biol.*, **235**, 260–86.

Marvin, D.A., Welsh, L.C., Symmons, M.F., Scott, W.R., and Straus, S.K. (2006) *J. Mol. Biol.*, **355**, 294–309.

Mikhonin, A.V. and Asher, S.A. (2005) *J. Phys. Chem. B*, **109**, 3047–52.

Mikhonin, A.V., Bykov, S.V., Myshakina, N.S., and Asher, S.A. (2006) *J.Phys. Chem. B*, **110**, 1928–43.

Miura, T., Satoh, T., Hori-i, A., and Takeuchi, H. (1998) *J. Raman Spectrosc.*, **29**, 41–47.

Miura, T., Takeuchi, H., and Harada, I. (1988) *Biochemistry*, **27**, 88–94.

Miura, T., Takeuchi, H., and Harada, I. (1989) *J. Raman Spectrosc.*, **20**, 667–71.

Miura, T. and Thomas Jr, G.J. (1994) *Biochemistry*, **33**, 7848–56.

Miura, T. and Thomas Jr, G.J. (1995a) *Subcell. Biochem.*, **24**, 55–99.

Miura, T. and Thomas Jr, G.J. (1995b) *Biochemistry*, **34**, 9645–54.

Movileanu, L., Benevides, J.M., and Thomas Jr, G.J. (2002a) *Biopolymers*, **63**, 181–94.

Movileanu, L., Benevides, J.M., and Thomas Jr, J.G. (2002b) *Nucleic Acids Res.*, **30**, 3767–77.

Nemecek, D., Gilcrease, E.B., Kang, S., Prevelige Jr, P.E., Casjens, S. and Thomas Jr, G.J. (2007), *J. Mol. Biol.*, **374**, 817–36.

Nemecek, D., Overman, S.A., Hendrix, R.W., Thomas Jr, G.J. (2008), *Biochemistry*, (submitted).

Okada, A., Miura, T., and Takeuchi, H. (2003) *Biochemistry*, **42**, 1978–84.

Overman, S.A., Aubrey, K.L., Reilly, K.E., Osman, O., Hayes, S.J., Serwer, P., and Thomas Jr, G.J. (1998a) *Biospectroscopy*, **4**, S47–S56.

Overman, S.A., Aubrey, K.L., Vispo, N.S., Cesareni, G., and Thomas Jr, G.J. (1994) *Biochemistry*, **33**, 1037–42.

Overman, S.A., Bondre, P., Maiti, N.C., and Thomas Jr, G.J. (2005) *Biochemistry*, **44**, 3091–100.

Overman, S.A., Kristensen, D.M., Bondre, P., Hewitt, B., and Thomas Jr, G.J. (2004) *Biochemistry*, **43**, 13129–36.

Overman, S.A. and Thomas Jr, G.J. (1995) *Biochemistry*, **34**, 5440–51.

Overman, S.A. and Thomas Jr, G.J. (1998b) *Biochemistry*, **37**, 5654–65.

Overman, S.A. and Thomas Jr, G.J. (1999) *Biochemistry*, **38**, 4018–27.

Overman, S.A., Thomas Jr, G.J. (2008), Filamentous ssDNA bacterial viruses, in *Encyclopedia of Virology* (eds. B.W.J. Mahy, M.H.V. van Regenmortel) Third Ed., Vol.2, Elsevier, Oxford, UK, pp. 190–198.

Overman, S.A., Tsuboi, M., and Thomas Jr, G.J. (1996) *J. Mol. Biol.*, **259**, 331–36.

Pappas, D., Smith, B.W., and Winefordner, J.D. (2000) *Talanta*, **51**, 131–44.

Premasiri, W.R., Moir, D.T., Klempner, M.S., Krieger, N., Jones, G., and Ziegler, L.D. (2005) *J.Phys. Chem. B*, **109**, 312–20.

Prevelige Jr, P.E., Thomas, D., King, J., Towse, S.A., and Thomas Jr, G.J. (1993), *Biochemistry*, **32**, 537–43.

Prochazka, M., Stepanek, J., Turpin, P.Y., and Bok, J. (2002) *J.Phys. Chem. B*, **106**, 1543–49.

Raman, C.V. and Krishnan, K.S. (1928) *Indian J. Phys.*, **2**, 419.

Raso, S.W., Clark, P.L., Haase-Pettingell, C., King, J., and Thomas Jr, G.J. (2001) *J. Mol. Biol.*, **307**, 899–911.

Reilly, K.E. and Thomas Jr, G.J. (1994) *J. Mol. Biol.*, **241**, 68–82.

Rodriguez-Casado, A., Moore, S.D., Prevelige Jr, P.E., Thomas Jr, G.J. (2001), *Biochemistry*, **40**, 13583–91.

Sane, S.U., Cramer, S.M., and Przybycien, T.M. (1999) *Anal. Biochem.*, **269**, 255–72.

Shimanouchi, T., Tsuboi, M., and Kyogoku, Y. (1964) *Adv. Chem. Phys.*, **7**, 435–98.

Siamwiza, M.N., Lord, R.C., Chen, M.C., Takamatsu, T., Harada, I., Matsuura, H. and Shimanouchi, T. (1975) *Biochemistry*, **14**, 4870–76.

Smekal, A.G. (1923) *Die Naturwissenschaften*, **43**, 873–75.

Sugeta, H. (1975) *Spectrochim. Acta A Mol. Spectrosc.*, **31**, 1729–37.

Symmons, M.F., Welsh, L.C., Nave, C., Marvin, D.A., and Perham, R.N. (1995) *J. Mol. Biol.*, **245**, 86–91.

Takeuchi, H. (2003) *Biopolymers*, **72**, 305–17.

Takeuchi, H. and Harada, H. (1986) *Spectrochim. Acta*, **42A**, 1069–78.

Takeuchi, H., Kimura, Y., Koitabashi, I., and Harada, I. (1991) *J. Raman Spectrosc.*, **22**, 233–36.

Tasumi, M., Jarada, I., Takamatsu, T., and Takahashi, S. (1982) *J. Raman Spectrosc.*, **12**, 149–51.

Thamann, T.J., Lord, R.C., Wang, A.H., and Rich, A. (1981) *Nucleic Acids Res.*, **9**, 5443–57.

Thiriot, D.S., Nevzorov, A.A., Zagyanskiy, L., Wu, C.H., and Opella, S.J. (2004) *J. Mol. Biol.*, **341**, 869–79.

Thomas Jr, G.J. (1999) *Annu. Rev. Biophys. Biomol. Struct.*, **28**, 1–27.

Thomas Jr, G.J. and Agard, D.A. (1984) *Biophys. J.*, **46**, 763–68.

Thomas Jr, G.J. and Barylski, J. (1970) *Appl. Spectrosc.*, **24**, 463–64.

Thomas Jr, G.J., Benevides, J.M., Overman, S.A., Ueda, T., Ushizawa, K., Saitoh, M., and Tsuboi, M. (1995) *Biophys. J.*, **68**, 1073–88.

Thomas Jr, G.J. and Day, L.A. (1981) *Proc. Natl. Acad. Sci. U.S.A.*, **78**, 2962–66.

Thomas Jr, G.J. and Hartman, K.A. (1973) *Biochim. Biophys. Acta*, **312**, 311–22.

Thomas, G.J., Medeiros, G.C., and Hartman, K.A. (1971) *Biochem. Biophys. Res. Commun.*, **44**, 587–92.

Thomas Jr, G.J. and Murphy, P. (1975) *Science*, **188**, 1205–7.

Thomas Jr, G.J., Prescott, B., McDonald-Ordzie, P.E., and Hartman, K.A. (1976) *J. Mol. Biol.*, **102**, 103–24.

Thomas Jr, G.J., Prescott, B., and Olins, D.E. (1977) *Science*, **197**, 385–88.

Thomas Jr, G.J., Prescott, B., Opella, S.J., and Day, L.A. (1988a) *Biochemistry*, **27**, 4350–57.

Thomas Jr, G.J., Prescott, B., and Urry, D.W. (1987) *Biopolymers*, **26**, 921–34.

Thomas Jr, G.J. and Tsuboi, M. (1993), *Raman Spectroscopy of Nucleic Acids and Their Complexes* (ed C.A. Bush), JAI Press Inc, Greenwich, CN, pp. 1–70.

Thomas Jr, G.J. and Wang, A.H.J. (1988b), Laser raman spectroscopy of nucleic acids, in *Nucleic Acids & Molecular Biology*, Volume 2 (eds F. Eckstein and D.J.M. Lilley), Springer-Verlag, Berlin, pp. 1–30.

Totir, M.A., Helfand, M.S., Carey, M.P., Sheri, A., Buynak, J.D., Bonomo, R.A. and Carey, P.R. (2007) *Biochemistry*, **46**, 8980–87.

Tsuboi, M., Benevides, J.M., Bondre, P., and Thomas Jr, G.J. (2005) *Biochemistry*, **44**, 4861–69.

Tsuboi, M., Benevides, J.M., and Thomas Jr, G.J. (2007a) *Biophys. J.*, **92**, 928–34.

Tsuboi, M., Ezaki, Y., Aida, M., Suzuki, M., Yimit, A., Ushizawa, K., and Ueda, T. (1998) *Biospectroscopy*, **4**, 61–71.

Tsuboi, M., Ikeda, T., and Ueda, T. (1991) *J. Raman Spectrosc.*, **22**, 619–26.

Tsuboi, M., Kubo, Y., Ikeda, T., Overman, S.A., Osman, O., and Thomas Jr, G.J. (2003a) *Biochemistry*, **42**, 940–50.

Tsuboi, M., Nishimura Y., Hirakawa, A.Y., and Petikolas W.L., (1987) Resonance Raman Spectroscopy and Normal Modes of the Nucleic Acid Bases, in *Biological Applications of Raman Spectroscopy* (ed. T.G. Spiro), Wiley Interscience, Vol. 2, New York, pp. 109–179.

Tsuboi, M., Overman, S.A., Nakamura, K., Rodriguez-Casado, A., and Thomas Jr, G.J. (2003b) *Biophys. J.*, **84**, 1969–76.

Tsuboi, M., Overman, S.A., and Thomas Jr, G.J. (1996a) *Biochemistry*, **35**, 10403–10.

Tsuboi, M., Suzuki, M., Overman, S.A., and Thomas Jr, G.J. (2000) *Biochemistry*, **39**, 2677–84.

Tsuboi, M., Thomas Jr, G.J. (2007b) Polarized Raman and polarized infrared spectroscopy of proteins and protein assemblies, in *Protein Structures: Methods in Protein Structure and Stability Analysis* (eds V.N. Uversky and Eu.A. Permyakov), Nova Science Publishers, Hauppage, New York, pp. 153–94.

Tsuboi, M., Ueda, T., Ushizawa, K., Ezaki, Y., Overman, S.A., and Thomas Jr, G.J. (1996b) *J. Mol. Struct.*, **379**, 43–50.

Tsuboi, M., Ushizawa, K., Nakamura, K., Benevides, J.M., Overman, S.A., and Thomas Jr, G.J. (2001) *Biochemistry*, **40**, 1238–47.

Tuma, R. (2005) *J. Raman Spectrosc.*, **36**, 307–19.

Tuma, R., Bamford, J.K.H., Bamford, D.H., Russell, M.P., and Thomas Jr, G.J. (1996a) *J. Mol. Biol.*, **257**, 87–101.

Tuma, R., Parker, M.H., Weigele, P., Sampson, L., Sun, Y., Krishna, N.R., Casjens, S., Thomas Jr, G.J., and Prevelige Jr, P.E. (1998a), *J. Mol. Biol.*, **281**, 81–94.

Tuma, R., Prevelige Jr, P.E., Thomas Jr, G.J. (1998b), *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 9885–90.

Tuma, R. and Thomas Jr, G.J. (1996b) *Biophys. J.*, **71**, 3454–66.

Tuma, R. and Thomas Jr, G.J. (1997) *Biophys. Chem.*, **68**, 17–31.

Tuma, R., Thomas Jr, G.J. (2002), Raman spectroscopy of viruses, in *Handbook of Vibrational Spectroscopy, Volume 5, Applications in Life, Pharmaceutical and Natural Sciences* (eds J.M. Chalmers and P.R. Griffiths), John Wiley & Sons, Chichester, U.K., pp. 3519–35.

Tuma, R., Tsuruta, H., Benevides, J.M., Prevelige Jr, P.E., Thomas Jr, G.J. (2001), *Biochemistry*, **40**, 665–74.

Tuma, R., Vohnik, S., Li, H.M., and Thomas, G.J. (1993) *Biophys. J.*, **65**, 1066–72.

Van Wart, H.E., Lewis, A., Scheraga, H.A., and Saeva, F.D. (1973) *Proc. Natl. Acad. Sci. U.S.A.*, **70**, 2619–23.

Van Wart, H.E. and Scheraga, H.A. (1977) *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 13–17.

Van Wart, H.E. and Scheraga, H.A. (1986a) *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 3064–67.

Van Wart, H.E. and Scheraga, H.A. (1986b) *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 3064–67.

Vohnik, S., Hanson, C., Tuma, R., Fuchs, J.A., Woodward, C., and Thomas Jr, G.J. (1998) *Protein Sci.*, **7**, 193–200.

Wen, Z.Q., Armstrong, A., and Thomas Jr, G.J. (1999) *Biochemistry*, **38**, 3148–56.

Wen, Z.Q., Overman, S.A., Bondre, P., and Thomas Jr, G.J. (2001) *Biochemistry*, **40**, 449–58.

Wen, Z.Q., Overman, S.A., and Thomas Jr, G.J. (1997) *Biochemistry*, **36**, 7810–20.

Wen, Z.Q. and Thomas Jr, G.J. (1998) *Biopolymers*, **45**, 247–56.

Wikoff, W.R., Conway, J.F., Tang, J., Lee, K.K., Gan, L., Cheng, N., Duda, R.L., Hendrix, R.W., Steven, A.C., and Johnson, J.E. (2006) *J. Struct. Biol.*, **153**, 300–306.

Wikoff, W.R., Liljas, L., Duda, R.L., Tsuruta, H., Hendrix, R.W., and Johnson, J.E. (2000) *Science*, **289**, 2129–33.

Williams, R.W. (1983) *J. Mol. Biol.*, **166**, 581–603.

Williams, R.W., Cutrera, T., Dunker, A.K., and Peticolas, W.L. (1980) *FEBS Lett.*, **115**, 306–8.

Yu, T.J., Lippert, J.L., and Peticolas, W.L. (1973) *Biopolymers*, **12**, 2161–75.

Zeri, A.C., Mesleh, M.F., Nevzorov, A.A., and Opella, S.J. (2003) *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 6458–63.

Ziegler, L.D., Hudson, B., Strommen, D.P., and Peticolas, W.L. (1984) *Biopolymers*, **23**, 2067–81.

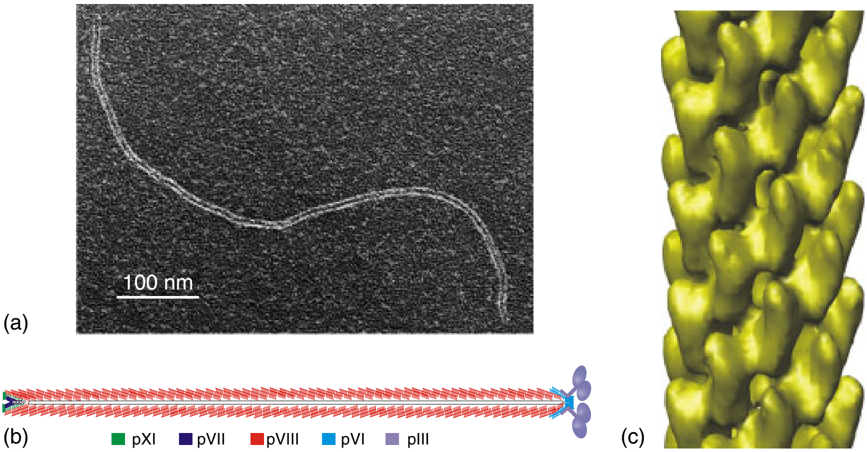

Fig. 13.10 The filamentous ssDNA bacteriophage fd. (a) A negatively stained electron micrograph showing one complete virion. The flexibility of the virion, which is apparent, is consistent with the experimentally determined equilibrium persistence length of 1.0 μm. (b) Cartoon of fd (not to scale) showing locations of the minor coat proteins. (c) Reconstruction of a 24 nm segment of fd at ~8 Å resolution from images obtained by electron cryomicroscopy (courtesy of E. H. Egelman, University of Virginia). (This figure also appears on page 444.)

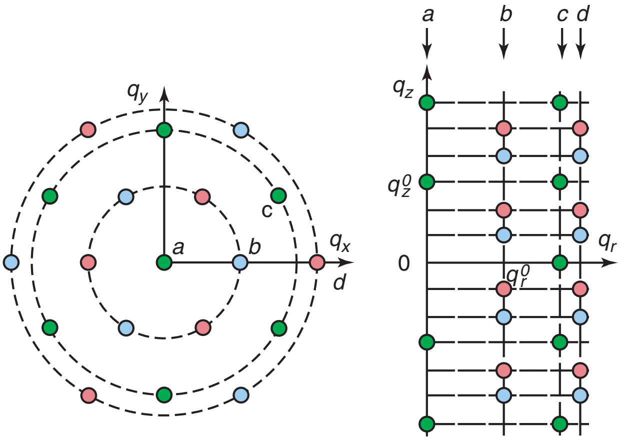

Fig. 14.17 The reciprocal lattice of rhombohedral symmetry on the $q_x - q_y$ plane (left) and the diffraction peaks detected on the $q_z - q_r$ plane (right). The reciprocal lattice of rhombohedral symmetry is composed of a stack of hexagonal layers stacked up in the ABCABC... sequence (A green, B blue, C red). In the $q_z - q_r$ plane are the four columns of the observed diffraction peaks. In Figure 14.18, the swelling method was applied to each set of the peaks having the same (H, K) (same color): column $a(H = 0, K = 0)$, $b(1, 0)$, $c(1, 1)$, and $d(2,0)$. The set of red peaks and the set of blue peaks in each of column b and d are equivalent by centrosymmetry (Yang and Huang, 2003). (This figure also appears on page 485.)

14

X-Ray Methods for Investigation of Structures of Lipid Assemblies

Huey W. Huang and Lin Yang

14.1 Introduction 459
14.1.1 Basic Structure of Lipid Bilayers 459
14.1.2 Nonlamellar Structures of Lipid Assemblies 460
14.2 Sample Preparation and Instrumentation 462
14.2.1 Sample Preparation 463
14.2.1.1 Powder Sample 463
14.2.1.2 Oriented Lipid Multilayers 464
14.2.1.3 Other Sample Configurations 465
14.2.2 Instrumentation 465
14.3 Structural Problems in Lamellar Phases 468
14.3.1 Ion-Binding Sites of the Gramicidin Channel 469
14.3.2 Hydrophobic Matching between Lipids and Proteins 474
14.3.3 Membrane Thinning Effect of Antimicrobial Peptides 475
14.4 In-Plane Scattering 475
14.4.1 Scattering by Proteins Embedded in Fluid Multilayers 477
14.4.2 Radial Distribution Function of Protein in Membrane 478
14.5 Periodically Ordered Lattices of Lipid Structures 481
14.5.1 Method of Measurement and Data Reduction 481
14.5.2 Methods of Phase Determination 483
14.5.2.1 Swelling Method and Pattern Recognition Method (With an Example in the Rhombohedral Phase) 484
14.5.2.2 Multiwavelength Anomalous Dispersion Method (With Examples in Hexagonal, Distorted Hexagonal, and Rhombohedral Phases) 486
References 498

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

14.1
Introduction

X-ray scattering measures instantaneous spatial correlations of electron density and more recently through the development of inelastic X-ray scattering, space–time correlations of electron density. By various modes of applications, it can provide structural as well as dynamical information. Compared with many other experimental tools that have been applied to lipid assemblies, X-ray has a clear advantage on structure determination, as evidenced by its definitive proof for the lipid bilayer structure in cell membranes. This article describes X-ray methods for structural determination of lipid assemblies and emphasizes applications to biophysical studies. So far X-ray methods have seldom been applied to native biological membranes. This is because for biophysical applications the purpose is to find the cause–effect relationships underlying biological phenomena. Diffraction or scattering from cell membranes that typically contain thousands of different lipid components and thousands of different proteins can never have an unambiguous interpretation. Thus the methods we describe here apply to artificial membranes composed of selected lipid components to which selected membrane-active molecules, such as proteins or drugs, are added.

14.1.1
Basic Structure of Lipid Bilayers

Until the early 1970s, the hypothesis that the basic structural element of biological membranes is a bimolecular layer of phospholipid molecules was seriously doubted due to the lack of direct evidence. The direct evidence was finally provided by X-ray diffraction of membranes and lipid bilayers (Levine and Wilkins, 1971; Wilkins *et al.*, 1971). Diffraction from dispersions of *Mycoplasma laidlawii*, erythrocyte, and other membranes showed the same basic features of diffraction from phospholipid bilayers, proving that a phospholipid bilayer is the major structural component of native membranes. These classic experiments established the salient features of diffraction from lipid bilayers, both from dispersions of bilayer vesicles and from oriented stacks of bilayers.

Diffraction from oriented stacks of lipid bilayers consists of low-angle lamellar reflections and a paraffin-like 0.46-nm diffuse band oriented at right angles to the lamellar diffraction. The

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

complete absence of any other low-angle diffraction indicates that the lipid chains are uniformly packed in the bilayers. This is contrary to many polymers that give strong low-angle diffraction due to irregularities in their chain packing. The 0.46-nm diffraction band is informative about the configuration and packing of the lipid chains. For fully saturated chains, the band shifts to a sharp 0.415-nm peak at low temperatures corresponding to the hexagonal packing of extended (all-*trans*) *n*-alkanes. The distance between two parallel, partially disordered chains is roughly 0.46 nm, as in liquid paraffins. The angular variation of the radially integrated intensity of the band provides a measure for the orientation of the chains in a bilayer. The radially integrated intensity is maximum at the equator and generally drops to less than 20% of the maximum within 40° as the angle increases toward the meridian. Conceptually one may picture the hydrocarbon region of the bilayer as an assembly of domains of locally correlated chain segments. The direction of chain segment varies continuously throughout the medium, but within a small domain the chain segments are nearly parallel. The domains diffract independently in a plane perpendicular to the axes of the chains. Thus the majority of the domains have the chain orientations more or less perpendicular to the plane of the bilayer. Such chain configurations put the terminal methyl groups near the middle of the bilayer and the glycerol ends of the chains to the surface of the hydrocarbon layer. This basic structure of lipid bilayers was reconfirmed by the electron density profiles constructed from the low-angle lamellar diffraction.

In many biophysical applications, the goal is to measure the structural changes of the host lipid bilayer caused by the addition of membrane-active components, or determine the distribution of the added components in the host bilayer. Therefore, it is important to remember the basic feature of the diffraction pattern from a lipid bilayer as described above. The simplest application is the measurement of the changes of the electron density profiles across the bilayer. This is very useful for studying the interaction of membrane-active molecules with membranes. We will illustrate three different examples: location of ion-binding sites, hydrophobic matching, and membrane thinning by antimicrobial peptides. The distribution of added molecules in a lipid bilayer can be measured by in-plane scattering, making use of the clean background outside of the lamellar diffraction and the diffuse band due to the chains. We will illustrate the method of in-plane scattering with gramicidin channels embedded in two different lipid bilayers.

14.1.2
Nonlamellar Structures of Lipid Assemblies

Besides bilayers, lipids and water have the potential to form wide varieties of nonlamellar assemblies, such as micellar, hexagonal, and cubic phases (Luzzati, 1968; Luzzati *et al.*, 1968; Tardieu *et al.*, 1973). It has long been recognized that many functions of membranes, particularly their abilities of shape changes (such as fusion, fission, and pore formation) are related to the intrinsic properties of lipid components that are capable of forming nonlamellar phases (Luzzati, 1968; Gruner, 1989). It has also been speculated that the lipid components that have the tendency of forming nonlamellar phases might create stresses in the lipid bilayer and consequently play a role of regulating the functions of membrane proteins (Gruner, 1989).

The majority of known lipid–water phases were discovered and studied by Luzzati and collaborators in the 1960s (Luzzati, 1968; Luzzati *et al.*, 1968; Tardieu *et al.*, 1973) (the earlier work referenced in Luzzati, 1968), but discoveries of new phases still continue as investigators broadened the range of chemical variables. New phases, even within lamellar phases (e.g. Yang and Fukuto, 2005), will undoubtedly continue to be discovered as the number of lipid components increases and greater varieties of lipid are investigated. There is no doubt in our mind that the field of structures of lipid assemblies is still in its infancy and will continue to grow as its impact on biology and technology increases.

As pointed out by Luzzati (1968) and Tardieu *et al.* (1973), one of the most remarkable characteristics of lipid–water mixtures is their ability to combine in one phase a periodically ordered long-range organization in one, two or three dimensions and a highly disordered short-range conformation in the unit cells. It is this unique property that makes the structural problems of lipid assemblies amenable to X-ray analysis. As in protein crystallography, the phase problem is the main obstacle for the reconstruction of electron density. But for two basic reasons, the well-developed phasing methods for protein crystallography are in most cases not applicable to lipidic structures. First, due to the short-range disorder a lipidic structure is described by an electron density distribution, not by atomic positions. Secondly, while most proteins are not centrosymmetric, most lipid assemblies prepared from lipid/protein/water mixtures are. The phasing methods of protein crystallography are based on atomicity (the electron density is zero between atoms) and are often specifically for noncentrosymmetric structures (Grossmann and Arnold, 2001). Thus a routine technique for phasing the diffraction amplitudes from lipid assemblies is yet to be developed, except for the simple case of lamellar phases. The phase problem for lamellar phases has been routinely solved by the swelling method (Perutz, 1954; Blaurock, 1971). But the phase problem for every nonlamellar assembly is a challenge.

In this article we describe methods and applications by examples. Besides applications in lamellar systems, we show successful solutions to structural problems in several nonlamellar phases, including inverted hexagonal phase, distorted hexagonal phase, and rhombohedral phase. Missing from the list are the structural problems of cubic phases. Since Scriven (1976) suggested that the bicontinuous cubic phases are based on the infinite periodic minimal surfaces, there have been a great deal of theoretical and experimental interests on these structures (Longley and McIntosh, 1983; Mariani *et al.*, 1988; Luzzati *et al.*, 1992; Seddon and Templer, 1993; Harper and Gruner, 2000; Harper *et al.*, 2000). Cubic phases are different from all other lipid–water phases in having a large water spacing (swollen), hence free of bilayer–bilayer interactions. It is considered the ideal system to study the curvature and the chain-packing energies. Structural problems of cubic phases are on-going research projects (Harper and Gruner, 2000; Harper *et al.*, 2000; Conn *et al.*, 2007). It is premature to be included in encyclopedia.

As for the fundamental theory of X-ray diffraction, we recommend the book by Warren (1969). Since lipid–water systems diffract easily even for poorly prepared samples, it is worth pointing out that a rigorous analysis of diffraction data is possible only for well-defined samples with a well-defined diffraction geometry (Warren, 1969). For this reason, we start with sample preparation and the diffraction technique for each type of sample.

14.2
Sample Preparation and Instrumentation

In this chapter, we discuss the types of samples that are used to study molecular assemblies of lipids and the X-ray scattering method suitable for measuring them. We will only focus on elastic scattering (including diffraction), in which the scattered X-rays have the same energy as the incident X-ray. Inelastic scattering (e.g. Chen *et al.*, 2001) and X-ray spectroscopy (e.g. Biou *et al.*, 2005) of lipid assemblies are possible but will not be discussed here, since their applications are still very limited. Coherent X-ray scattering may also be useful for studying the dynamics of lipid assemblies. However, to date its application to membrane structures remains a fascinating possibility rather than a reality.

An elastic X-ray scattering measurement is equivalent to taking the Fourier transform of the electron density distribution in the sample. The recording on the detector is in the reciprocal space. The structures at different length scales in the sample manifest as diffraction or scattering features at different q values in the reciprocal space. Here $q = 4\pi \sin\theta/\lambda$ is the scattering vector, or the momentum transfer of the scattering, with 2θ being the scattering angle between the incident and scattered rays and λ the X-ray wavelength. The magnitude of q is inversely proportional to the length scale probed by the X-rays. Sharp diffraction peaks are characteristic of periodically ordered structures. Bragg's law correlates the peak position in the reciprocal space to the spacing of the periodic order, while the direction of q is the orientation of the periodicity. Structures without long-ranged order (e.g. liquid-like in-plane structures, fluctuation of membranes) do not produce Bragg peaks, so their analyses are more complex. However, the reciprocal relationship between q and real space spacing still holds true. A diffused peak often signifies a characteristic length or an average spacing.

Oriented stacks of lipid bilayers produce three diffraction or scattering features, schematically shown in Figure 14.1. The most prominent are the Bragg peaks associated with the periodic layers, with the first order typically at $q_z \sim 1\ \text{nm}^{-1}$ (lamellar spacing $D \sim 6$ nm). The second feature is due to the chain–chain correlations in the bilayers, which manifest as a broad peak at $q_r \sim 14\ \text{nm}^{-1} (q_r = \sqrt{{q_x}^2 + {q_y}^2})$, corresponding to $\sim$0.45 nm spacing between parallel chains. Finally, if there are supramolecular structures (e.g. membrane pores or protein assemblies) within the membrane, additional scattering peaks may appear. It is important to remind ourselves that all scattering features from a system are related. Therefore, a good experimental procedure should record the complete scattering pattern first, even if the purpose is to analyze only one of the many features,for example the lamellar diffraction.

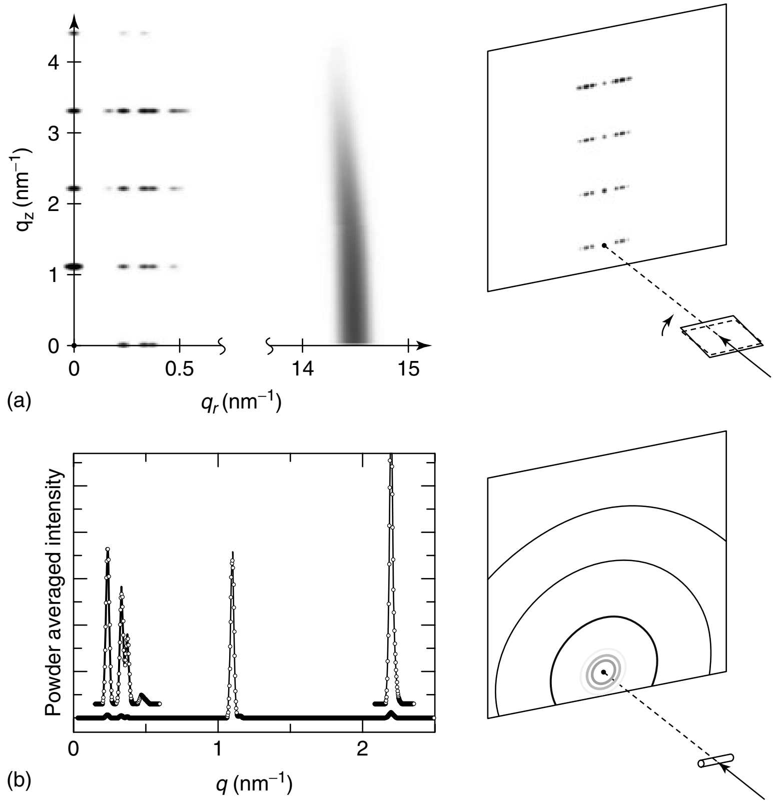

Fig. 14.1 (a) Simulated diffraction pattern on the $q_z - q_r$ coordinates from oriented stacks of bilayers in the square phase (Yang and Fukuto, 2005). The pattern includes lamellar peaks on q_z, a paraffin-like broad band, and satellite peaks next to the lamellar peaks. The last feature represents a supramolecular ordering within the bilayers. All three features are seen clearly in this diffraction geometry. (b) The same diffraction pattern measured by powder diffraction.

14.2.1
Sample Preparation

14.2.1.1 **Powder Sample**

Powder samples that contain randomly oriented structures are fairly easy to prepare and measure. These samples can be prepared by simply mixing lipids with water. Repeated freeze–thaw cycles are used to improve sample homogeneity. The sample is often housed in a thin-walled capillary tube and measured in transmission geometry.

Since the structures within a powder sample are randomly oriented, the scattering pattern contains only rings concentric to the direct beam, as shown in Figure 14.1b. Scattering data can be recorded using either an area detector or a linear detector. In order to ensure good powder averaging, the capillary tube is often rotated during measurements.

Powder samples are convenient if the purpose is to determine the lattice constant or to determine the lattice symmetry, but not ideal for the studies of structures. Diffraction peaks that are clearly resolved in the $q_z - q_r$ plane (Figure 14.1a) may have a similar magnitude of $|q|$ and therefore not distinguishable in powder diffraction. This is especially problematic for lipid assemblies because their diffraction peaks are not sharp due to their soft nature. The situation is worse if the lattice constant is large. For example, the unit cell in the square phase (Yang and Fukuto, 2005; Figure 14.1a) is ~10 times the layer spacing. In such cases even the identification of the lattice symmetry becomes difficult by power diffraction.

14.2.1.2 **Oriented Lipid Multilayers**

To record the complete diffraction pattern of membranes one must use oriented multilayers. All examples presented in the following chapters are based on this type of samples. The most important advantage of aligned samples is the separation of the q_r and q_z components of the scattering vector. The three diffraction features of membranes are clearly separated (Figure 14.1a). And the identification of lattice symmetry is greatly simplified.

There are several methods for preparing aligned lipid multilayers (Huang and Olah, 1987; Ludtke *et al.*, 1995). The most commonly used method is solvent casting. Lipids and other components, such as membrane-active proteins, are first codissolved in organic solvents. The solution is directly deposited on a clean, flat substrate (glass or silicon). Lipid multilayers (even nonlamellar phases have layered structures) form spontaneously as the solvents evaporate and the lipids hydrate from water vapor. The choice of the solvent is crucial for forming a uniform film (Ludtke *et al.*, 1995). The primary requirement is that the solvent system must have an appropriate surface tension and evaporation rate for the size of the substrate being used. For example, pure chloroform causes more sample to aggregate at the center of the substrate and tends to leave "rings" behind as it evaporates. Pure trifluoroethanol, on the other hand, causes a very smooth deposition in the center of the substrate, but most of the material aggregates along the edges of the substrate. Of course, chloroform evaporates much more rapidly than TFE, and what really matters is the surface tension near the end of the evaporation process. A mixture of three parts chloroform and one part TFE produces good results for most pure lipids. The mixing ratio may need to be adjusted depending on the materials used and other solvent may be added to improve the solubility of nonlipid components. Trial and error is required when working with a new lipid–protein system, but once the method has been refined, new samples can be made quite rapidly and reproducibly.

Thick aligned multilayers can be prepared by (i) lyophilizing the lipid–protein mixture, (ii) hydrating the mixture by saturated water vapor, and (iii) sandwiching the mixture between two flat plates (Huang and Olah, 1987). Although beryllium has been used for this purpose, for safety reason, thin silicon plates might be a better choice. Thin silicon plates are also partially transparent to long wavelength visible light (red) as silicon transmission starts to rise toward IR spectrum. Optical observation of the sample (for the inspection of smectic defects in the sample (Huang and Olah, 1987; He *et al.*, 1996)) is therefore possible.

Freestanding films are another form of aligned multilayers that have been widely used in liquid crystal research community. They are prepared by spreading a powder sample across an aperture. Although successful experiment has been carried out with freestanding films (e.g. Smith *et al.*, 1988) and in some cases even the number of bilayer in the film can be precisely controlled (Chiang *et al.*, 1995), in general these samples are not easy to prepare and the film tends to rupture when the lipid undergoes a phase transition.

14.2.1.3 **Other Sample Configurations**

Lipid vesicles are attractive membrane samples because they resemble the morphology of cell membranes. Vesicle samples can be prepared by extrusion or sonification of a lipid suspension (e.g. Brzustowicz and Brunger, 2005). The advantage of vesicles is that not only additional molecules could be incorporated into the lipid bilayer, but the chemical environment of the membrane could also be varied. However, the vesicle concentration must be low to avoid vesicle–vesicle correlations. Therefore, one must work with data of low intensity. Like small angle scattering, the data analysis is typically by model fitting.

Another type of lipid membrane sample that has been widely studied is the lipid monolayer, which can be prepared by spreading a lipid solution in organic solvent onto the surface of water. A Langmuir trough provides the means of controlling lipid area density or the surface pressure of the monolayer by moving a barrier that changes the surface area of the monolayer. Owing to the very small sample volume, X-ray scattering from a monolayer is weak. The measurements are often conducted with the X-ray incident angle below the critical angle so as to take advantage of surface-enhancement of scattering intensity. For small incident angles, this can be achieved by using a deflecting mirror. More flexible scattering geometry requires a dedicated liquid diffractometer. From the scientific point of view, monolayers may not be good representatives of biological membranes for the obvious reason that the latter are bilayers. There are examples that have shown that transmembrane proteins (e.g. gramicidin) in monolayers behave very differently from those in bilayers.

Recently, X-ray scattering measurements from a single bilayer freely floating between two boundaries (Daillant *et al.*, 2005) or supported on a solid substrate have been reported (Miller *et al.*, 2005). The presence of the surrounding bulk water necessitates the use of relatively high-energy X-rays (e.g. 1 cm of water transmits ~28% of 16 keV photons but only 0.004% of 8 keV photons). In either case, the diffraction geometry was restricted by the sample arrangement and the scattering from a single membrane was weak. We expect further progress from these methods.

14.2.2
Instrumentation

Figure 14.2 illustrates the typical experimental setup for measuring X-ray scattering from lipid samples. The X-rays generated by a source are trimmed by multiple sets of collimators (slits or apertures). Focusing optics (mirror or bent multilayer) are often utilized to increase the flux of X-rays that pass through the collimators or to improve the angular resolution. The sample is usually kept

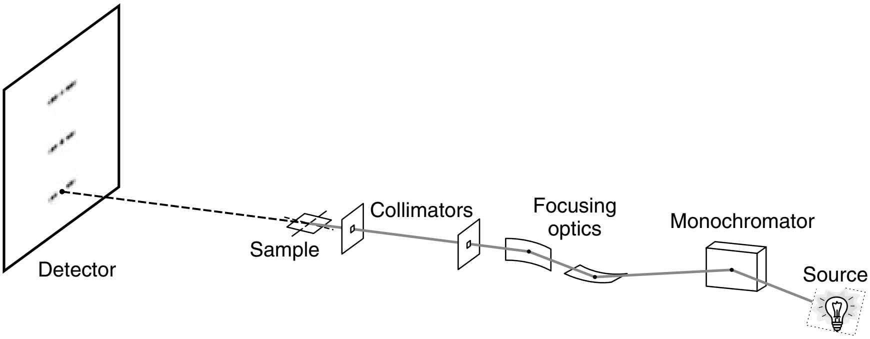

Fig. 14.2 Typical experimental setup for X-ray scattering measurements of lipid assemblies. A side-bounce monochromator, an oriented sample and an area detector are shown. The components can be implemented in many different ways.

in an enclosure in which the sample environment (temperature and humidity) is under experimental control. The sample chamber is mounted on a manipulator that can translate and re-orient the sample with respect to the X-ray beam. The detector at the opposite end collects the scattering data. Whenever possible, helium-filled or evacuated beam paths are used to reduce air absorption and scattering.

Most measurements on powder or aligned multilayer samples can be carried out using a sealed tube X-ray generator. This is especially true as more and more lab-based diffractometers are now equipped with focusing optics. However, for measurements that require extremely high angular resolution, very short data collection time or energy tunability (e.g. anomalous diffraction), synchrotron-based instruments are necessary.

One important aspect of the X-ray experiment is the manipulation of the sample's position and orientation during the measurement. For example, the powder sample of a lamellar phase is a collection of multilayered vesicles, which are usually not an ideal powder (giving nonuniform powder rings). The powder averaging can be improved by rotating the sample during the measurement. An aligned sample needs to be oriented correctly with respect to the incident X-ray and then rotated during the measurement to give the complete scattering pattern. The diffraction geometry must be well defined relative to the sample position for the scattering intensity to be correctly normalized between data points.

Not often appreciated is the fact that aligned multilayers are two-dimensional powder samples. Its powder nature (orientational randomness) is impossible to control during the sample preparation. In fact, a well-aligned sample may have larger domain sizes that tend to make the powder averaging less complete. An example is given in Figure 14.3, which lacks the left–right symmetry in intensity due to the incomplete powder averaging. An in-plane rotation of the sample during the measurement can alleviate this problem. An alternative is to translate the sample to increase the averaging sample volume. Sample translation has the added benefit of reducing the radiation dose that would decrease the possibility of radiation damage. Indeed,

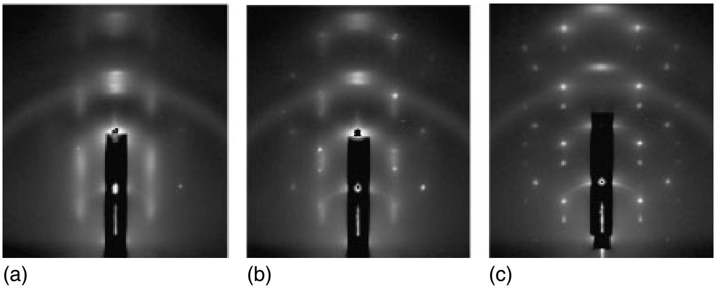

Fig. 14.3 An example of imperfect powder averaging. Aligned multilayers of a peptide–lipid mixture (magainin in 3 : 1 DMPC/PG) measured at 25 °C. (a) At 98% RH magainin formed pores in fluid bilayers. The satellite diffuse bands are due to the interference between the diffusing pores as in a 2D liquid. (b) and (c) at 96% RH and 93% RH, respectively, the pores were correlated to form a periodic lattice. The pattern in (b) shows left–right asymmetry indicating imperfect in-plane powder averaging.

radiation damage is of great concern for diffraction experiment of lipid assemblies, particularly when synchrotron radiation is used. Careful experimenters examine the samples by thin layer chromatography (Skipski *et al.*, 1962) after the measurement.

Another aspect of sample manipulation is the control of the environment surrounding the sample. The lipid structure can be very sensitive to the temperature and the water content of the sample (Figure 14.3). Accurate control of the sample temperature is usually not a problem using commercial devices. For powder samples, the water content can be adjusted by using the gravimetric method (by weight) or by mixing the sample with a polyethylene glycol (PEG) solution of known osmolality. For aligned multilayers, the water content in the sample is controlled by the ambient humidity. Unfortunately, good humidity controls are not commercially available. A practical humidity control can be achieved by heating/cooling a water reservoir based on the feedback from a humidity sensor (Yang and Huang, 2003).

The detector is selected based on the consideration of the desired angular resolution and the speed of measurement. A point detector with collimators (Figure 14.4a) that limit the direction of X-rays entering the detector usually gives the best angular resolution. The collimators also keep background scattering minimal by allowing only the X-rays scattered from the sample to be counted. However, the point-by-point data collection necessarily prolongs the sample exposure to X-rays. A linear detector collects scattering data simultaneously in a linear array of pixels. High angular resolution is achievable in the direction perpendicular to the pixel array by utilizing Soller slits (Figure 14.4b).

For the sake of speed, an area detector is much preferred. However, one must be aware of its potential problems. First, the background noise could be strong and uneven because the entire path of the incident beam is the potential source for unwanted scattering. Secondly, the diffraction peak may appear to be smeared, especially at wide scattering angles. This is because the detector at

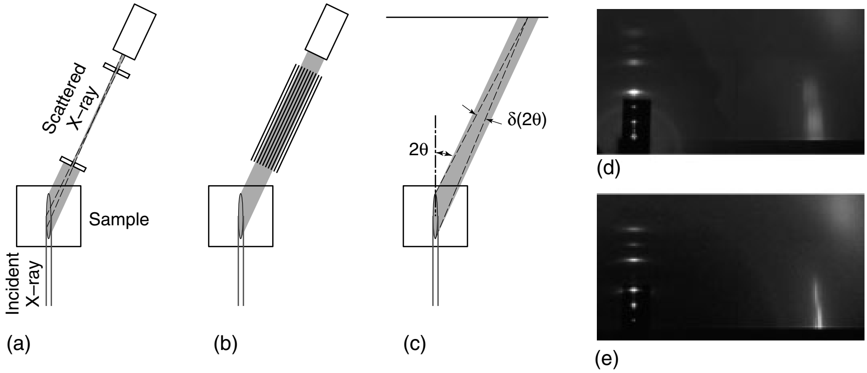

Fig. 14.4 Angular resolution (top view) in a grazing incidence reflection measurement. (a) A point detector with slit collimation gives the best resolution. (b) A linear detector with Soller slits improves the resolution in one direction. (c) An area detector may have problems with resolution, particularly at wide angles due to the long incident footprint. (d) and (e) are scattering patterns from the same sample under identical conditions, except that in (d) the full length (20 mm) of the sample was illuminated by X-rays, whereas in (e) the X-ray beam was focused to ~20 μm at the sample in the direction along sample normal and therefore resulted in a footprint of ~3 mm at 0.4° incident angle.

a wide scattering angle "sees" a large incident beam footprint (Figure 14.4c). The improvement in resolution by a focused beam (Yang, 2005) is illustrated in Figure 14.4d and e. If the available beam is not a focused beam, reducing the sample size could also help.

14.3
Structural Problems in Lamellar Phases

The peaks of lamellar diffraction are confined to the *c*-axis normal to the periodic layers (designated as the z-axis). Therefore, the constructed electron density represents the structure averaged over the $x - y$ plane and has only the z dependence. In a fluid-phase phospholipid bilayer, the only structural parameter that can be measured to a high accuracy is the phosphate-peak to phosphate-peak distance (PtP). PtP is a measure of the membrane thickness. Since a lipid bilayer is an elastic body, a change in the membrane thickness from its unperturbed value would induce an elastic stress that could act directly on an embedded protein. Thus, a protein-induced change in the PtP is a quantity that directly measures the degree of protein–lipid interaction. Two such examples will be illustrated below. In another example, we show that it is possible to find the location of heavy atoms in a fluid lipid bilayer.

Oriented stacks of lipid bilayers are preferred for lamellar diffraction. Owing to the difference in the Lorentz factor, a powder sample (that has a Lorentz factor $\sim 1/h^2$, h being the Bragg order) rarely produces more than four Bragg orders, whereas aligned multilayer samples (Lorentz factor $\sim 1/h$) typically produce 6–8 orders. Since the diffraction pattern is limited to one axis, a point detector rather than an area detector is preferred for limiting the background signal to minimum. In particular, if a diffraction beam monochromator is used, the Compton scattering and the fluorescence from the sample can be screened, and it can also be set up to bias against higher harmonics. Keeping the background to minimum allows the measurement of weak high-order peaks. The relative contribution of high-order peaks to the electron density is magnified by the Lorentz factor; therefore, small high-order peaks are not negligible. This is important if one wishes to measure the PtP as accurately as possible.

Multilayer samples are measured in a temperature–humidity chamber. In most cases one wishes to measure the membrane properties in the fluid phase of the lipid chains and as close to full hydration as possible. However, it is well known that membranes in full hydration undergo undulation fluctuations that damp the high-order diffraction peaks and broaden the remaining ones (Caille, 1972), making structural analysis difficult. It has been shown that although the lamellar spacing D tends to diverge as the hydration level reaches 100% humidity, the bilayer thickness PtP does not. Indeed, the PtP value at full hydration can be extrapolated from a series of measurement at less than full hydration levels (Chen *et al.*, 1997). For practical purpose, one can measure the bilayer thickness of full hydration approximately at the highest humidity level (~98% RH) without the damping effect of undulation fluctuations.

We illustrate the technical details of the method in the first example below.

14.3.1
Ion-Binding Sites of the Gramicidin Channel

Gramicidin is by far the most extensively studied membrane-active peptide that forms a transmembrane ion channel. A complete characterization of this channel is desirable as a case study of ion transport across a membrane. The channel selectively facilitates the diffusion of monovalent cations across the membrane, but does not transmit anions or divalent cations. Experiments have indicated that there are two monovalent binding sites in the channel, but their locations cannot be determined from the molecular structure of the channel, nor does gramicidin form crystals in the channel form. The problem then was to determine the ion-binding sites in the gramicidin channels embedded in fluid lipid bilayers.

Gramicidin channels were incorporated into dilauroyl-phosphatidylcholine (DLPC) bilayers to a peptide–lipid ratio of 1 : 10. Identical multilamellar samples were prepared with K^+, Tl^+, Ba^{++}, Mg^{++}, or without salt (Huang and Olah, 1987). The sample was kept in a temperature– humidity chamber and Bragg diffraction was measured with a point detector using the $\omega - 2\theta$ geometry.

One unique feature of membrane diffraction is that the repeat spacing is more than 30 times the X-ray wavelength. The scattering vector $\boldsymbol{q}$ satisfying the Bragg law lies very close to the surface of the Ewald sphere. As a consequence, it is very easy to obtain incorrect diffraction amplitudes with the sample misaligned in χ without realizing it (χ is the rotational angle around the line of intersection between the plane of scattering and the sample plane). A procedure for positioning and orienting lamellar samples was described in Wu *et al.*, (1995). A two-dimensional $\omega - \theta$ scan around the second or third Bragg order can be used to check the alignment of the ω-angle. A correctly aligned sample has the peak

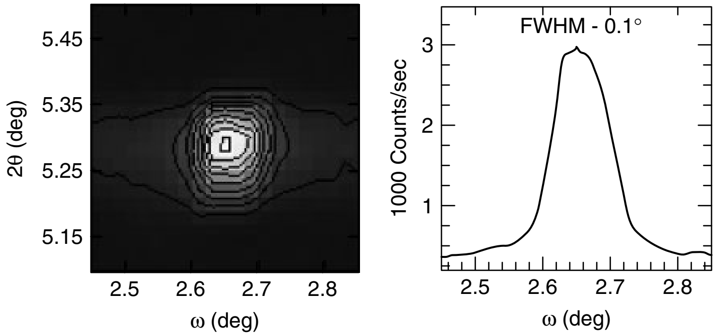

Fig. 14.5 (Left) An example of two-dimensional $\omega - \theta$ scan around the third Bragg peak of an aligned multilayer sample of peptide–lipid mixtures. The sample was well aligned and showed a single peak on a flat background. (Right) A cut through the center of the peak at constant θ corresponds to the ω rocking curve. The FWHM of the rocking curve is 0.18, the same as the limiting resolution of the apparatus. The background below the Gaussian fit to the rocking curve is negligible, implying very little powder diffraction (Weiss *et al.*, 2003).

position exactly at $2\omega = 2\theta$ in the $\omega - \theta$ plane. Also, the quality of the sample could be assessed from this scan: a well-aligned sample exhibited a single Bragg peak on a flat even background, whereas poor alignment showed up as a broad shoulder along ω. An example of well-aligned sample is shown in Figure 14.5.

The $\omega - 2\theta$ scan was recorded with a step size of $\Delta\theta = 0.01°$ (Figure 14.6). To make use of the swelling method, the scan was repeated at several humidity

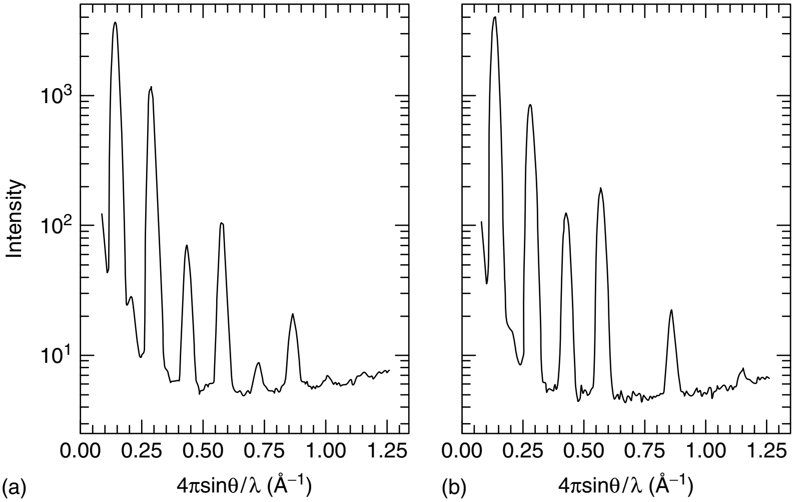

Fig. 14.6 Lamellar diffraction patterns of gramicidin–DLPC multilayers: (a) with Tl^+ ions and (b) without salt. Eight Bragg peaks were recorded. The lamellar spacings were 4.34 nm for both patterns (Olah *et al.*, 1991).

levels. Only consecutive scans that overlapped within the experimental errors indicating that the sample condition did not change during the scan were averaged and subsequently analyzed. Data reduction consisted of the following steps on the basis of the correction factors expressed as (Wu *et al.*, 1995).

$$E = I_0|F|^2 \, C_P C_L C_A C_V \qquad (1)$$

E is the measured intensity. The form factor square $|F|^2$ is the quantity we wish to obtain. I_0 is proportional to the incident intensity, and C's are the correction factors. (i) Background subtraction. This usually presents no problem except for the small high-order peaks. If that were the case, the high-angle region was scanned over a longer time until the background was smooth. (ii) Corrections for C_A, the factor of intensity decrease due to absorption, and for C_V, the diffraction volume. These corrections (particularly the first few Bragg orders) are sensitive to the beam geometry, so they are best calculated by constructing a ray diagram based on the sample and beam geometry (Wu *et al.*, 1995). The rest of the data reduction was not sensitive to the small variation of θ, so after the corrections (i) and (ii) the intensity was integrated for each Bragg peak to obtain the integrated peak intensities. (iii) The polarization factor C_P depends on the beam polarization and the diffraction geometry (Warren, 1969). For an unpolarized beam and with a diffraction beam monochromator is $C_P = (\cos^2 2\theta \cos^2 2\theta_M + 1)/2$, where θ_M is the monochromator angle. $C_P = (\cos^2 2\theta + 1)/2$ without the monochromator. (iv) The Lorentz factor $C_L = 1/\sin 2\theta$ (Warren, 1969).

We note that there is one more possible correction to be considered. If the incident X-ray beam has an angular divergence and the sample is mosaic, the vertical size of the Bragg reflections will increase with order. In the vertical direction, the slit width may be smaller than the width of the Bragg reflection; therefore, a correction factor has to be introduced that accounts for the increasing loss in intensity with order (Saxena and Schoenborn, 1977). This correction factor is usually important only for neutron diffraction.

After the corrections, we obtained the relative magnitudes of the diffraction amplitude $|F(q)| = |F(2\pi h/D)|$ for the diffraction pattern of lamellar spacing D and Bragg order h. The amplitude $F(q)$ is a real quantity since the bilayer made from a mixture is centrosymmetric. The phase problem is to determine its sign. This is usually done by the swelling method (Perutz, 1954; Blaurock, 1971) based on the idea that in most cases the structure of a lipid bilayer does not change much with hydration changes. Let us assume that the structure of the bilayer is invariant under a change of hydration level, which merely changes the thickness of the water layer. Then one can show that the diffraction amplitudes at lamellar spacing D satisfy Blaurock's scaling relation (Blaurock, 1971)

$$\sum_h \left| F\left(\frac{2\pi h}{D}\right)\right|^2 = \frac{D}{D_0} \qquad (2)$$

where D_0 is an arbitrary constant. Therefore, the diffraction amplitudes obtained from the raw data at each hydration level are first scaled to satisfy the relation (2). The scaled amplitudes are then plotted in the q space, and the phases are chosen so that a smooth curve would connect all the data points: we call this a phasing diagram. Examples are given in Figure 14.7. This is because if the bilayer is invariant, the form factor of the lipid bilayer (minus the water background) should be the same independent of the lamellar spacing. Indeed, one can construct the form factor $F(q)$ from one set of data of lamellar spacing D, called *the Shannon curve,*

$$F(q) = \sum_h F\left(\frac{2\pi h}{D}\right) \frac{\sin(qD/2 - h\pi)}{qD/2 - h\pi} \tag{3}$$

and the curve should go through the data points of all other lamellar spacings, provided the phases have been chosen correctly. As pointed out by Torbet and Wilkins (1976) and Olah *et al.* (1991), in real systems the bilayer is most likely slowly changing with the hydration. In that case the phases should still be chosen so that the data points of different D are smoothly connected as in Figure 14.7. This method apparently works for the majority of lamellar diffraction. With the phases determined, the diffraction amplitudes can be used to construct the unnormalized electron density $\rho_{\mathrm{ex}}(z)$

$$\rho_{\mathrm{ex}}(z) = \sum_h F(2\pi h/D)\cos(2\pi hz/D) \tag{4}$$

Because the diffraction intensity was not normalized (to the amount of sample) and the zeroth order was not measurable, the experimental density is related to the true density $\rho(z)$ by two unknown constant parameter a and b

$$\rho(z) = a\rho_{\mathrm{ex}}(z) + b \tag{5}$$

We need two conditions to determine a and b. For example, if the molecular composition of the bilayer is known (one might need to estimate the number of water molecules per lipid), one condition is $\int \rho(z)dz =$ number of electrons from the molecular composition. Another condition is that the density at the center of bilayer is that of the methyl group, $1.7 \times 10^2\ \mathrm{e\,nm^{-3}}$. The details of the normalization procedure for the

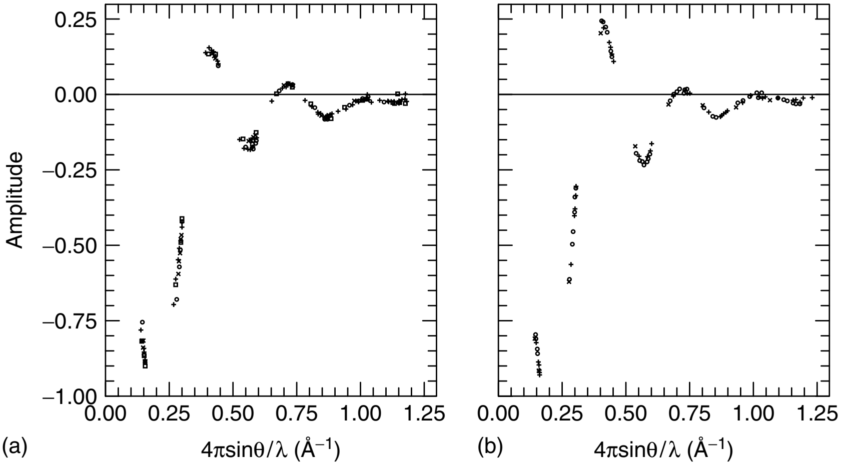

Fig. 14.7 Phasing diagrams. The signs of the diffraction amplitudes were chosen so that the data points fall on a single smooth curve. (a) the Tl^+ sample and (b) the salt-free sample were measured at several different hydration levels (Olah *et al.*, 1991).

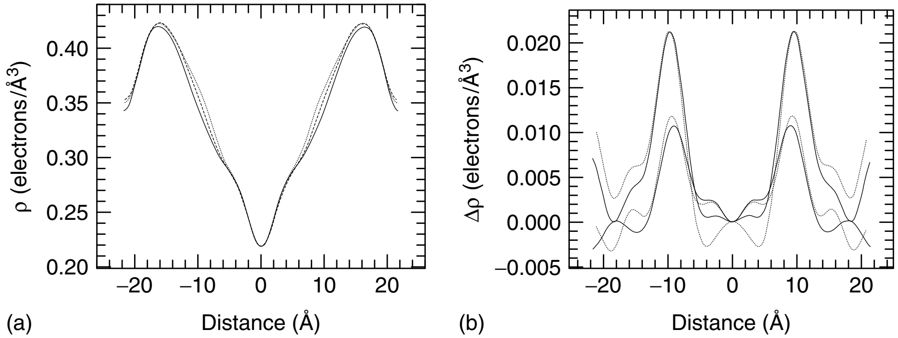

Fig. 14.8 (a) Normalized electron density profiles of gramicidin–DLPC bilayers with Tl^+ (dotted line), with K^+ (broken line), and without salt (solid line), obtained from the diffraction amplitudes at lamellar spacing 4.34 nm. (b) Difference in elelctron density profiles. The top 2 are $\rho(Tl^+)-\rho$(salt-free); the bottom 2 $\rho(Tl^+) - \rho(K^+)$. The solid and dotted lines are of lamellar spacing at 4.34 and 4.24 nm, respectively (Olah *et al.*, 1991).

gramicidin samples were given in (Olah *et al.*, 1991). If one is only interested in measuring the peak to peak distance PtP, normalization is unnecessary (Chen *et al.*, 1997).

The normalized electron density profiles of the DLPC bilayer containing gramicidin channels are shown in Figure 14.8a with K^+ or with Tl^+. The difference in electron density profile (Figure 14.8b) revealed a pair of symmetrically located ion-binding sites for monovalent cations at 0.96 (±0.03) nm from the midpoint of the gramicidin channel. Similarly, Figure 14.9 shows a pair of binding sites for divalent cations at 1.30 (±0.02) nm from the midpoint of the gramicidin channel. To date, the X-ray method is the only experimental technique capable of directly measuring the locations of the ion-binding sites in the gramicidin channel. These ion-binding sites have been crucial for the molecular dynamics studies of the gramicidin channel (Roux, Prod'hom and Karplus, 1995).

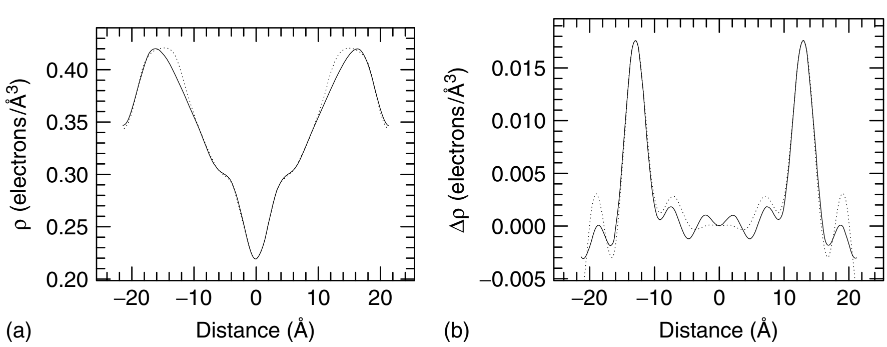

Fig. 14.9 (a) Normalized electron density profiles of gramicidin–DLPC bilayers with Ba^{2+} (dotted line) and with Mg^{2+} (solid line), obtained from the diffraction amplitudes of lamellar spacing 4.28 nm. (b) Difference in electron density profiles $\rho(Ba^{2+}) - \rho(Mg^{2+})$ at lamellar spacings 4.28 and 4.44 nm (Olah *et al.*, 1991).

14.3.2
Hydrophobic Matching between Lipids and Proteins

The gramicidin channel is a head-to-head hydrogen bonded dimmer. In lipid bilayers, the channel undergoes association–dissociation (dimer–monomers) reactions. As a result, a channel has a well-defined lifetime (its survival probability exponentially decreases in time) that can be measured accurately by single-channel conductance experiment (Hladky and Haydon, 1970, 1984). It has been shown that the gramicidin channel lifetime decreases with increasing thickness of lipid bilayer (Elliott *et al.*, 1983). This membrane thickness effect was explained as a consequence of hydrophobic matching between the gramicidin channel and the lipid bilayer. When there is a hydrophobic mismatch between the channel and the unperturbed state of the lipid bilayer, the effect of hydrophobic matching would create a local deformation in the lipid bilayer according to the elasticity theory. This membrane deformation energy and its effect on the gramicidin channel lifetime have been calculated, and the result successfully explained the membrane thickness effect (Huang, 1986).

Hydrophobic matching, in which a transmembrane protein causes the surrounding lipid bilayer to adjust its hydrocarbon thickness to match the hydrophobic surface of the protein, is a commonly evoked hypothesis. However, direct experimental evidence is scarce. To test the hypothesis, gramicidin was incorporated, at 1 : 10 peptide–lipid ratio, in two lipid bilayers of different thickness: DLPC and dimyristoylphosphatidylcholine (DMPC). The bilayer thickness, PtP, with or without gramicidin, was measured by X-ray diffraction (Harroun *et al.*, 1999). As we can see from the result shown in Figure 14.10, the discussion of membrane thickness is complicated by its temperature and hydration dependence. In each temperature, PtP was measured as a function of the lamellar spacing *D* by varying the relative humidity (RH) surrounding the sample. (*D* can be measured accurately to ±0.01 nm, whereas the accuracy of commercial RH sensors is typically ±1.5%.) In the fluid phase of DMPC bilayers, PtP approaches a constant value at high *D* values (or as RH approaches full hydration). It is well known that each phospholipid, particularly those with saturated chains like DMPC, has a fluid–gel main transition

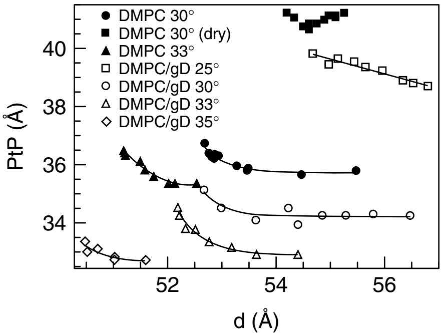

Fig. 14.10 Phosphate peak-to-phosphate peak (PtP) distance of the electron density profile as a function of lamellar spacing *D* for pure DMPC bilayers and DMPC bilayers containing gramicidin at the peptide-to-lipid ratio 1 : 10 (Harroun *et al.*, 1999).

temperature (~24 °C for DMPC). However, it would be a mistake to assume that DMPC containing gramicidin has the same transition temperature. In general, the transition temperature increases and broadens with inclusion of foreign molecules, but it is impossible to predict. Thus, part of the experiment was in finding that the transition temperature of gramicidin–DMPC 1 : 10 mixture extended over a range from 27 to 32 °C. In the fluid phase (≥32 °C) and near full hydration, the PtP was 3.53 nm for pure DMPC but 3.27 nm for the DMPC–gramicidin mixture (Harroun *et al.*, 1999). On the other hand, the PtP was 3.08 nm for pure DLPC but 3.21 nm for the DLPC–gramicidin mixture (Olah *et al.*, 1991; Chen *et al.*, 1997). Gramicidin apparently thickened DLPC and thinned DMPC toward a common thickness as expected by hydrophobic matching. The result is consistent with the effect of membrane thickness on the gramicidin channel lifetime (Elliott *et al.*, 1983; Huang, 1986).

14.3.3
Membrane Thinning Effect of Antimicrobial Peptides

In another example of measuring the membrane thickness, we investigated the molecular mechanism of antimicrobial peptides. Gene-encoded antimicrobial peptides act on the lipid matrix of cell membranes. Their bactericidal activities exhibit a so-called all-or-none concentration dependence, that is, little activity below a threshold concentration and quickly reaching the maximum activity as the concentration exceeds the threshold. To understand this phenomenon, we have performed experiments on the peptide–lipid systems with X-ray, neutron, oriented circular dichroism, and giant unilamellar vesicles (Huang, 2006). We found that above a concentration threshold the peptides form pores in the membranes. But the crucial data for understanding the mechanism came from the X-ray measurement of membrane thickness.

Representative peptides (alamethicin, magainin, protegrin, and melittin) were incorporated into various lipid bilayers (complete references in Huang (2006)). In each case the bilayer thickness was measured as a function of the peptide concentration (peptide–lipid molar ratio P/L), and in all cases the PtP decreased linearly with P/L below a threshold P/L^* (Figure 14.11). Above the threshold, PtP was constant. This indicates that the peptides initially bind to the lipid interface causing an area expansion, hence thinning the hydrocarbon region. Membrane thinning incurs an elastic energy proportional to the square of P/L. Hence, the energy level of inter-racial binding increases with P/L. As the peptide concentration increases over the threshold concentration P/L^*, the interfacial binding energy level surpasses the energy of pore formation. So, above the threshold all excessive peptides form pores, which if occuring in cell membranes cause cytolysis.

14.4
In-Plane Scattering

We now describe the method of measuring the distributions of proteins or protein assemblies in membranes. As an example, we use the X-ray diffraction taken from oriented stacks of lipid bilayers containing gramicidin channels (Yang *et al.*, 1999).

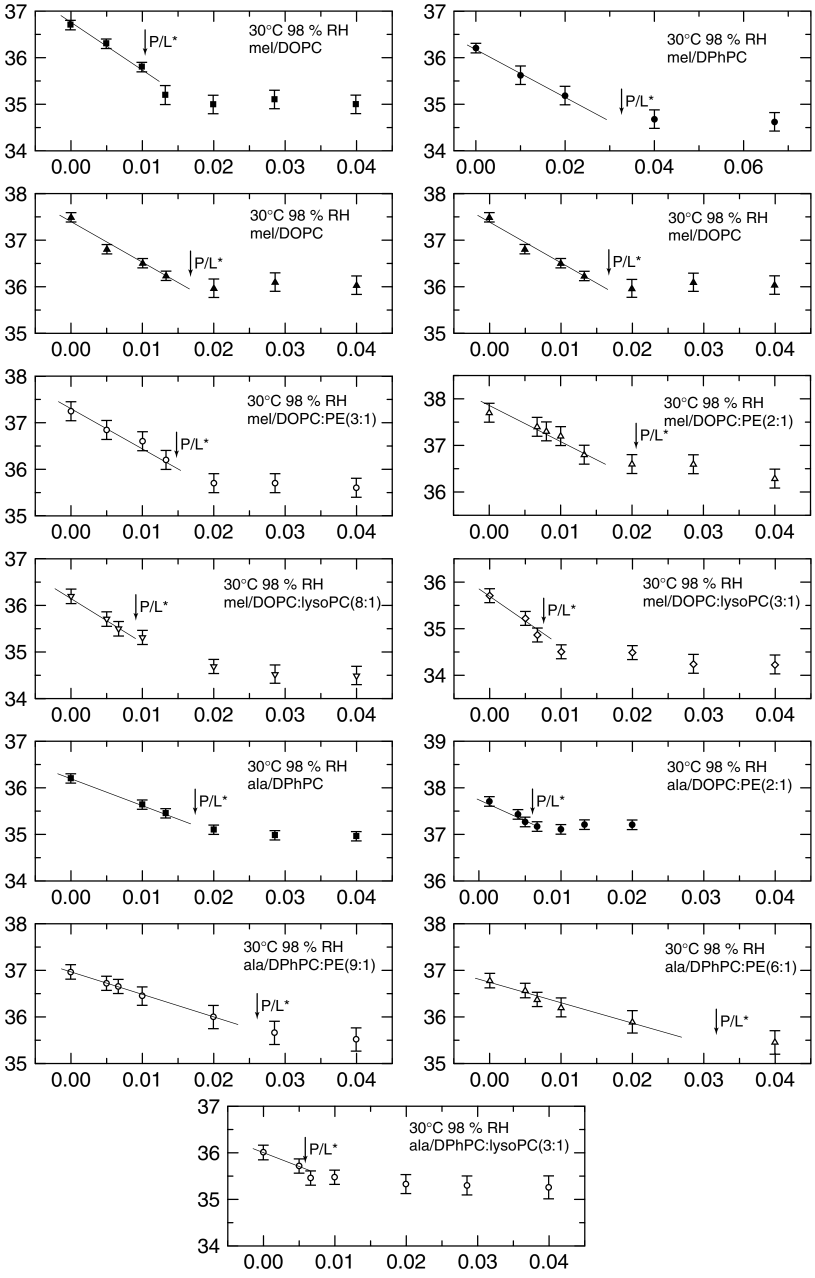

Fig. 14.11 Phosphate peak-to-phosphate peak (PtP) of lipid bilayers as a function of the peptide concentration P/L. Each panel shows a different peptide–lipid combination. P/L^* in each panel is the threshold concentration measured by oriented circular dichroism. At concentrations above P/L^* the peptides form pores in the membranes (Huang, 2006).

14.4.1
Scattering by Proteins Embedded in Fluid Multilayers

We start with X-ray scattering intensity $I(\boldsymbol{q})$ given by the following equation:

$$I(\boldsymbol{q}) = \left\langle \left|A(q)\right|^2 \right\rangle = \left\langle \left| \int \rho(\boldsymbol{r}) \exp(\mathrm{i}\boldsymbol{q} \cdot \boldsymbol{R}) d\boldsymbol{R} \right|^2 \right\rangle \quad (6)$$

where $A(\boldsymbol{q})$ is the scattering amplitude and $\rho(\boldsymbol{R})$ is the electron density at position $\boldsymbol{R}$ in the sample. In Eq. (6) a pre-factor representing the scattering intensity from an electron has been omitted for simplicity (Warren, 1969). We write $\rho(\boldsymbol{R}) = \rho_0(\boldsymbol{R}) + \phi(\boldsymbol{R})$ where $\rho_0(\boldsymbol{R})$ represents the density of pure lipid multilayers prepared without gramicidin, and $\phi(\boldsymbol{R})$ is the density contrast of gramicidin channels against the lipid background. Defining

$$A_0(q) = \int \rho_0(r) \exp(\mathrm{i}\boldsymbol{q} \cdot \boldsymbol{R}) d\boldsymbol{R},$$
$$A_\phi(q) = \int \phi(r) \exp(\mathrm{i}\boldsymbol{q} \cdot \boldsymbol{R}) d\boldsymbol{R} \quad (7),$$

we have,

$$I = \langle |A_0 + A_\phi|^2 \rangle = \langle |A_0|^2 \rangle + \langle |A_\phi|^2 \rangle + 2\langle \mathrm{Re}(A_0 A_\phi^*) \rangle \quad (8)$$

The contribution of A_0 has been described in Section 14.1.1. As illustrated in Section 14.2, the scattering amplitude of proteins is in the low-angle region normal to the bilayers' lamellar diffraction and does not overlap with the diffraction from lipid chains around ~14 nm^{-1}. Therefore, we may consider the scattering by $\phi(\boldsymbol{R})$ separately. We denote the positions of the centers of the gramicidin channels by $\boldsymbol{R}_n$, $n = 1, 2, \ldots, N$, so that $\phi(\boldsymbol{R}) = \sum_n \phi_c(\boldsymbol{R} - \boldsymbol{R}_n)$ where $\phi_c(\boldsymbol{R} - \boldsymbol{R}_n)$ is the density contrast for a gramicidin channel. Then the scattering intensity can be written in the form

$$I(q_z, q_r) = N|F_\phi(q_z, q_r)|^2 S(q_z, q_r) \quad (9)$$

with the structure factor $S(q_z, q_r)$ and the form factor $F_\phi(q_z, q_r)$ given by

$$S(q_z, q_r) = \frac{1}{N} \left\langle \left| \sum_n \exp(\mathrm{i}\boldsymbol{q} \cdot \boldsymbol{R}_n) \right|^2 \right\rangle \quad (10)$$

$$F_\phi(q_z, q_r) = \int \phi_c(\boldsymbol{R} - \boldsymbol{R}_n) \times \exp[\mathrm{i}\boldsymbol{q} \cdot (\boldsymbol{R} - \boldsymbol{R}_n)] d\boldsymbol{R} = F_{\mathrm{G}}(q_z, q_r) - F_{\mathrm{lb}}(q_z, q_r) \quad (11)$$

$F_{\mathrm{G}}(q_z, q_r)$ and $F_{\mathrm{lb}}(q_z, q_r)$ are, respectively, the form factor of the gramicidin channel and the form factor of a patch of lipid bilayer that would have filled the space occupied by the gramicidin channel. Note that the z-coordinate of $\boldsymbol{R}_n$ is discrete, denoted by mD, $m = 0, \pm 1, \pm 2, \ldots$ In this case the structure factor can be written as (Yang *et al.*, 1999)

$$S(q_z, \boldsymbol{q}_r) = 1 + \sum_m \exp(\mathrm{i}q_z mD) \int n_{0m}(\boldsymbol{r}) \times \exp(i\boldsymbol{q}_r \cdot \boldsymbol{r})\, d\boldsymbol{r} \quad (12)$$

where $\boldsymbol{r}$ is the two-dimensional coordinate in the plane parallel to the membranes. The correlation function $n_{0m}(\boldsymbol{r})d\boldsymbol{r}$ is defined as the average number of gramicidin channels within the area element d$\boldsymbol{r}$ at the position $\boldsymbol{r}$ in the mth bilayers relative to an arbitrarily chosen gramicidin channel in the zeroth bilayer.

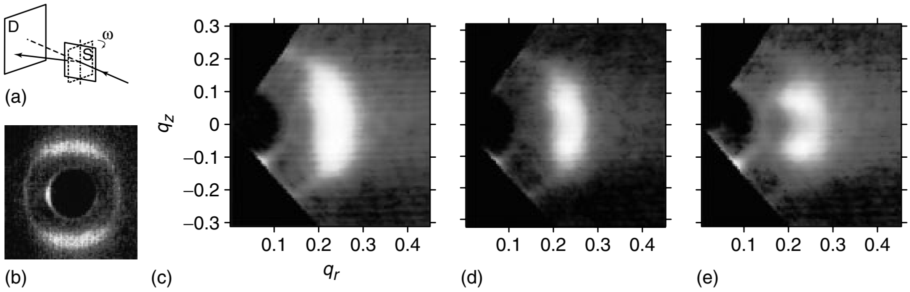

Fig. 14.12 (a) The diffraction geometry. The multilayers were 60° with respect to the detector. (b) The detector image of DLPC bilayers containing gramicidin at the peptide–lipid ratio 1 : 10 at 25 °C and >95% RH. (c) The detector image (a) was translated to the $q_z - q_r$ coordinates. (d) and (e) the diffraction pattern of the same sample at 90 and 80% RH, respectively (Yang *et al.*, 1999).

By symmetry $n_{0m} = n_{0-m}$. Expanding the sum in Eq. (12) and making use of the fact that the distribution of channels in a fluid bilayer is isotropic in the plane, we have

$$S(q_z, \mathbf{q}_r) = S(q_z, q_r) = S_{00}(q_r) + 2\cos \times (q_z D) S_{01}(q_r) + \cdots \qquad (13)$$

$$S_{0m}(q_r) = \delta_{0m} + \int [n_{0m}(r) - \overline{n}] \times J_0(q_r r) 2\pi r \, dr \qquad (14)$$

Note that in Eq. (13) all $S_{0m}(q_r)$ factors are multiplied by a q_z-dependent factor, except for the $S_{00}(q_r)$ term.

The distribution of gramicidin channels in a membrane is described by the radial distribution function $2\pi r n_{00}(r)\, dr$, which is the average number of channels within the ring of radius r and width dr, centered at an arbitrarily chosen channel. This function can be obtained from $S_{00}(q_r)$ by Bessel transform

$$2\pi r n_{00}(r) = 2\pi r \overline{n} + r \int [S_{00}(q_r) - 1] \times J_0(q_r r) q_r \, dq_r \qquad (15)$$

14.4.2
Radial Distribution Function of Protein in Membrane

Mixtures of lipids and gramicidin at the molar ratio of 10 : 1 were lyophilizied and then hydrated by water vapor (Huang and Olah, 1987). Approximately 5 mg of hydrated sample was sandwiched between two polished, SiO_2 coated beryllium plates (0.5 mm thick, 15 mm in diameter). The sandwiched samples were manipulated to rid of smectic defects (the procedures were described in Huang and Olah, 1987). The scattering pattern was recorded with the multilayers oriented at 60° with respect to the detector (Figure 14.12a). The detector image (Figure 14.12b) was translated to the (q_z, q_r) plane (Figure 14.12c). The scattering pattern of gramicidin is in a region free of the diffraction from the lipid multilayers. The data shown in Figure 14.12c was taken at RH >95%. The same sample was also measured at ~90% RH (Figure 14.12d) and at ~80% RH (Figure 14.12e).

Inter-bilayer correlations are usually absent when the multilayers are fully hydrated since the bilayers are well separated. However, this needs to be

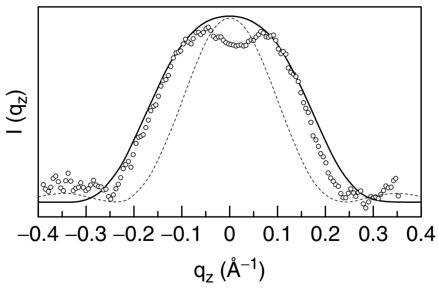

Fig. 14.13 The q_z-dependence of the diffraction pattern shown in Figure 14.12c. The data (o) were taken along the q_z coordinate at the value of q_r fixed at the maximum intensity. The solid line is the form factor of the gramicidin channel contrasted against the lipid background $|F_\iota(q_z, q_r)|^2 = |F_G(q_z, q_r) - F_{\mathrm{lb}}(q_z, q_r)|^2$. In contrast the dotted line is the form factor of the gramicidin channel $|F_G(q_z, q_r)|^2$. It shows that the contribution of the lipid background is important. The data show a small dip at the top indicating a small sign of inter-bilayer correlations. It implies that the sample was not quite fully hydrated (Yang *et al.*, 1999).

demonstrated before one uses the data to analyze the distribution of proteins in the membranes. This can be seen in the case of Figure 14.12c. Figure 14.13 shows that the q_z dependence of the scattering intensity $I(q_z, q_r)$ (Figure 14.12c) is the same as the q_z dependence of the form factor $|F_{phi}(q_z, q_r)|^2$, where the value of q_r is fixed at the peak value along the q_r axis. That means that according to Eq. (9) the structure factor $S(q_z, q_r)$ has no significant q_z dependence. Therefore, according to Eqs (14.12) and (14.13), $S(q_z, q_r)$ is essentially $S_{00}(q_r)$. All $S_{0\,m}(q_r)$, $m \neq 0$ terms in the expansion of $S(q_z, q_r)$ (Eq. (13)) vanish. Only under this condition one may obtain $S_{00}(q_r)$ by $S_{00}(q_r) = I(q_z, q_r)/(N|F_\phi(q_z, q_r)|^2)$. One can then use Eq. (15) to obtain the radial distribution function of proteins.

We now compare the distribution of gramicidin channels in DLPC and DMPC bilayers, both prepared at the peptide-to-lipid molar ratio of 1 : 10. Under the condition of no inter-bilayer correlations, the scattering intensities $I(q_z = 0, q_r)$ are shown in Figure 14.14: DMPC/gD at 20 °C (gel phase) and at 32 °C (fluid phase) and DLPC/gD at 25 °C (fluid phase). Their respective radial distribution functions are shown in Figure 14.15 (Harroun *et al.*, 1999). The distribution function rises from near zero at small r to a first maximum at the distance where there is a high probability for finding neighboring channels. This distance is the most probable nearest-neighbor separation between channels. Beyond this distance, $2\pi r n_{00}(r)$ oscillates below and above the average density curve $2\pi r\bar{n}$. The most interesting result is in the difference between DMPC/gD and DLPC/gD in the fluid phase. Apparently the channel–channel nearest-neighbor distance is smaller in DMPC (2.23 nm) than in DLPC (2.68 nm). Individual gramicidin channels diffuse randomly in the fluid membranes, but on an average the channels are 13% closer to each other in DMPC than in DLPC. In Section 14.3, we have seen that gramicidin channels caused a greater thickness change in DMPC than in DLPC because DMPC has a greater hydrophobic mismatch with the gramicidin channel than does DLPC. The thickness change around each channel creates a strain field in the lipid bilayer. This strain field creates an attractive membrane-mediated potential between gramicidin channels (Harroun *et al.*, 1999). As expected, the X-ray measurement showed that this attractive potential

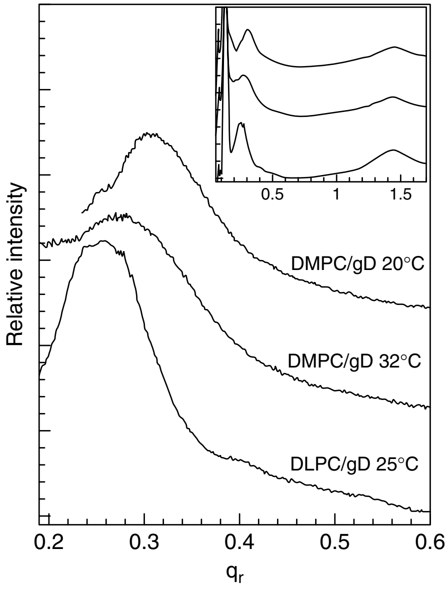

Fig. 14.14 X-ray in-plane scattering curves. The inset show (from top to bottom) the raw data of DMPC/gD (10 : 1) at 20 °C, DMPC/gD (10 : 1) at 32 °C, and DLPC/gD (10 : 1) at 25 °C. The main frame shows the in-plane scattering curves after removing the pure lipid background. These are the in-plane scattering curves of gramicidin (Harroun *et al.*, 1999).

is greater in DMPC than in DLPC. There have long been speculations that protein–lipid hydrophobic mismatch would cause membrane-mediated attractive interactions between proteins, and such interactions might be responsible for inhomogeneous lateral distributions of intrinsic membrane proteins and consequently the functional specialization of regions in cell membranes (Marcelja, 1976; Schroeder, 1977; Owicki and McConnell, 1979). As far as we know, the result shown in Figure 14.15 provided the first and only experimental evidence.

Finally a short note on inter-bilayer correlations. At lower hydration levels (Figure 14.12d and e), the q_z dependence of $I(q_z, q_r)$ shows undulations that are apparently different from the q_z dependence of $|F_\phi(q_z, q_r)|^2$ in Figure 14.7. This indicates existence of inter-bilayer correlations. Analyses of inter-bilayer correlations can be found in Yang *et al.*, 1999.

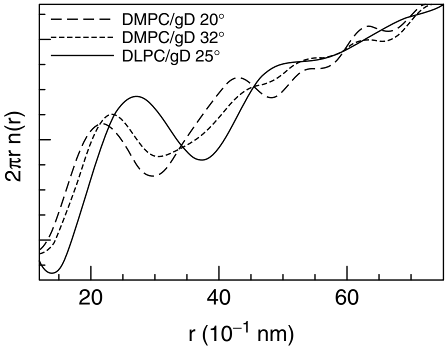

Fig. 14.15 Radial distribution functions obtained from the data shown in Figure 14.14. The position of the first peak gives the most probable nearest-neighbor separation between gramicidin channels in the bilayer (Harroun *et al.*, 1999).

14.5 Periodically Ordered Lattices of Lipid Structures

14.5.1 Method of Measurement and Data Reduction

As mentioned in Section 14.2, periodically ordered lattices of lipid assemblies (excluding the cubic phases) can be conveniently prepared on the surface of a solid substrate (a silicon wafer or a polished glass plate). The lattice constants of lipid assemblies are typically ∼5 nm, and their diffraction peaks tend to be limited within $<10\ \mathrm{nm}^{-1}$. The lattice domains in such samples are randomly oriented in the plane of the substrate. Therefore, with the exception of the reciprocal vectors normal to the substrate (z-axis), all Bragg peaks form rings parallel to the plane of substrate (the x–y plane). With X-rays of ∼8–15 keV incident at a grazing angle, the Ewald sphere intercepts the majority of the Bragg rings (reflection image in Figure 14.16). Each Bragg ring will be registered as a diffraction peak on the detector. However, there are two types of Bragg peaks not recorded by the grazing angle reflection. First, at a fixed incident angle, the Ewald sphere intersects the q_z axis twice: one at the origin of reciprocal space and another at the specular angle. Therefore, the diffraction pattern at one incident angle does not include the Bragg peaks on the z-axis. These peaks are measured by rotating the sample continuously to vary the incident angle, so that the intercept of the Ewald sphere with the z-axis moves along this axis at constant speed (ω-scan in Figure 14.16). Secondly, the Bragg rings with $q_z = 0$ are inaccessible by

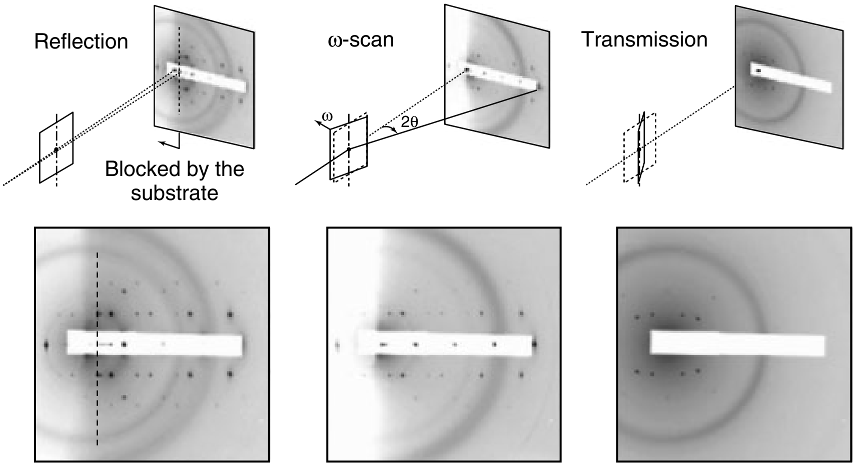

Fig. 14.16 A complete diffraction pattern for a nonlamellar phase is obtained by a combination of three recordings. (a) Reflection: X-ray beam was incident at a grazing angle (0.7°) to the substrate. On the detector, the region to the left of the dashed line was blocked by the substrate. The spot between the dashed line and (001) was the specular reflection. (b) ω-scan: this was the conventional $\omega - 2\theta$ scan along the q_z axis. The intensities accumulated during the ω-scan at the peak positions off the q_z axis were not suitable for analysis. (c) Transmission: the incident angle was −30°. Each detector image was the average of two 120-seconds exposures (Yang and Huang, 2003).

reflection, because, in practice, reflection always imparted a finite q_z. These Bragg peaks on the x–y plane can be recorded by transmission diffraction (transmission image in Figure 14.16). These three recordings combine to give a complete diffraction pattern for a periodically ordered lattice on a substrate. However, if one takes into account the symmetry of the lattice, the recording can be simplified for some lattices. For example, one grazing angle reflection is sufficient to include the complete diffraction pattern for an inverted hexagonal phase.

Removing the background from the diffraction peaks needs to be done with care. Usually the intensities of the diffraction peaks are integrated directly using the detector image. One may fit each peak with a two-dimensional Gaussian function plus a background and then integrate the function without the background. Or, one may integrate the intensities first in one direction (e.g. parallel to the substrate) over a width slightly wider than the apparent peak widths. The result is plotted along the substrate normal. On this one-dimensional profile the background is obtained by using the intensities between peaks and extrapolated into the peak regions. After the background subtraction, each peak is integrated. It would be reassuring if the two methods give the same result.

The procedure for data reduction starts with the translation of a vector (P_x, P_y, P_z) in the detector frame to a vector $(P^s{}_x, P^s{}_y, P^s{}_z)$ in the sample frame. (i) Use a standard sample (e.g. silver behenate, $D = 5.84$ nm) to determine the beam center O on the detector frame and the sample-to-detector distance $\overline{\mathrm{SD}}$. (ii) Determine the physical coordinates of each peak on the detector (P_x, P_z) (P_x is parallel to the plane of substrate). (iii) The diffracted beam vector **S** is defined by the direction $(P_x, \overline{\mathrm{SD}}, P_z)$ in the detector frame. (iv) Determine the incident angle α from the reflection point. Then the direction of **S** in the sample frame is given by

$$\begin{aligned}\begin{pmatrix} P^s_x \\ P^s_y \\ P^s_z \end{pmatrix} &= |P^s| \begin{pmatrix} S_x \\ S_y \\ S_z \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix} \\ &\quad \times \begin{pmatrix} P_x \\ \overline{SD} \\ P_z \end{pmatrix} \qquad (16)\end{aligned}$$

We also define the following angles: $(S_x, S_y, S_z) = (\cos\nu \ \sin\gamma,\ \cos\nu \ \cos\gamma,\ \sin\nu)$. The scattering vector in the sample frame is $q = (q_x, q_y, q_z) = (2\pi/\lambda)$ $[-(S_x, S_y, S_z)](0, \cos\alpha, -\sin\alpha)$.

The integrated intensity for a Bragg peak E is related to unnormalized form factor squared $|F|^2$ by $E = I_0|F|^2 C_\mathrm{P} C_\mathrm{L} C_\mathrm{abs} C_\mathrm{geo} \Delta T$ (Table 14.1). The reduction includes the polarization factor C_P, the Lorentz factor C_L, the absorption factor C_abs, and a factor for diffraction volume C_geo. The principles for all these factors can be found in Warren's book (Warren, 1969). The values of these factors are given in Table 14.1. The last factor ΔT is for relative normalization between the reflection recording and the ω-scan recording. Given the same incident beam intensity, if the reflection is recorded for a time ΔT in seconds, the equivalent factor for the ω-scan is $2\pi/\Omega$ where Ω is the angular speed of ω-rotation in rad s^{-1} (Yang and Huang, 2003; Ding *et al.*, 2005). This allows us to obtain the relative magnitude of the diffraction amplitude $|F|$ for all Bragg peaks.

Tab. 14.1 Formula for data reduction $E = I_0 |F|^2 C_P C_L C_{\mathrm{abs}} C_{\mathrm{geo}} \Delta T$.

Reflection	**ω-scan**
$C_p = \sin^2 \psi$	$C_p = \cos^2 2\theta$
$C_L = \dfrac{1}{\cos\alpha \cos\nu \sin\gamma}$	$C_L = 1/\sin 2\theta$
$C_{\mathrm{abs}} = \dfrac{1 - \exp\left[-\mu a \left(\dfrac{1}{\sin\nu} + \dfrac{1}{\sin\alpha}\right)\right]}{\mu a \left(\dfrac{1}{\sin\nu} + \dfrac{1}{\sin\alpha}\right)}$	$C_{\mathrm{abs}} = \dfrac{[1 - \exp(-2\mu a/\sin\theta)]\sin\theta}{2\ \mu a}$
$C_{\mathrm{geo}} = \dfrac{b/\sin\alpha}{w}$ or 1 if greater than 1	$C_{\mathrm{geo}} = \dfrac{b/\sin\alpha}{w}$ or 1 if greater than 1
ΔT	$2\pi/\Omega$

Notations (also applicable to neutron diffraction except for noted): 2θ, the angle between the incident beam and the diffracted beam; ψ, the angle between the incident polarization vector and the scattered beam; $\alpha = \omega$, the angle between the incident beam and the substrate surface (the $x - y$ plane); ν, the angle between the diffracted beam and the $x - y$ plane; γ, the angle between the in-plane $(x - y)$ projection of incident beam and the in-plane projection of the diffracted beam; μ, the linear absorption coefficient of sample; μ_s, the linear absorption coefficient of substrate; a, the sample thickness; and a_s, the substrate thickness.

The polarization factor C_P is for X-ray diffraction only.
C_{abs} in the table is for reflection geometry. For transmission geometry the absorption correction is

$$C_{\mathrm{abs}} = \exp\left[-\frac{\mu_s a_s + \mu a}{\sin\alpha}\right] \frac{1 - \exp\left[-\mu a \left(\dfrac{1}{\sin\nu} - \dfrac{1}{\sin\alpha}\right)\right]}{\mu a \left(\dfrac{1}{\sin\nu} - \dfrac{1}{\sin\alpha}\right)}$$

C_{geo} is the ratio of the length of the beam footprint $b/\sin\alpha$ over the length of the sample w (b is the beam width). ΔT is the scan time. Ω is the angular velocity of the ω-scan.

14.5.2
Methods of Phase Determination

The unit cell structure of a lipid assembly prepared from a homogeneous mixture can be reasonably assumed to be centrosymmetric $\rho(\boldsymbol{r}) = \rho(-\boldsymbol{r})$. Then from the definition of diffraction amplitude

$$F(\boldsymbol{q}) = \int_{\mathrm{cell}} \rho(\boldsymbol{r}) \exp(\mathrm{i}\boldsymbol{q} \cdot \boldsymbol{r})\, d\boldsymbol{r} \qquad (17)$$

one can easily show that $F(\boldsymbol{q}) = F^*(\boldsymbol{q})$. In other words, $F(\boldsymbol{q})$ is real and its phase is either 0 or π, or $F(\boldsymbol{q})$ is either positive or negative. At first glance, the phase problem seems simple. Given a finite number n of diffraction amplitudes, there are only 2^n possible phase combinations and one of them must be right. This might be a comforting thought, but it is sometimes difficult to pick the correct one by inspecting all possible combinations even in the case of lamellar diffraction (Wang *et al.*, 2006).

The two methods described below have recently been used with some success.

14.5.2.1 **Swelling Method and Pattern Recognition Method (With an Example in the Rhombohedral Phase)**

One of the most interesting discoveries concerning lipidic structures is the rhombohedral phase (Yang *et al.*, 2000). In Section 14.4 we described gramicidin channels embedded in oriented stacks of lipid bilayers and saw that the channels developed inter-bilayer correlations when the hydration level decreased. The same phenomenon occurred when the peptide-induced transmembrane pores were present in membranes (Yang *et al.*, 2000), and in these cases the inter-bilayer correlations eventually developed into long-range correlations. The pores formed a lattice of rhombohedral symmetry (space group $R\bar{3}$). Subsequently it was found that not only peptide-induced pores formed this lattice, pure lipid bilayers with negative spontaneous curvatures also formed this lattice in dehydrated condition (Yang and Huang, 2002; Yang *et al.*, 2003). Thus the rhombohedral phases seem to provide a common lattice for many different lipidic structures.

A rhombohedral lattice consists of hexagonal layers, stacked up in the ABCABC... fashion, similar to face-centered cubic (As expected, the hexagonal layers are always parallel to the surface of the substrate.). Between the hexagonal layers are intercalating layers of water. In some cases, the rhombohedral lattices allow a variation of the vertical lattice constant in response to a change of hydration level. This is called *swelling*, and the swelling method of phase determination could be used, similar to the lamellar phases described in Section 14.3. Here we describe an example of solving the phase problem for a rhombohedral lattice using the swelling method.

The rhombodedral diffraction pattern in Figure 14.16 is on a lattice described by the set of reciprocal vectors $\boldsymbol{b}_1 = (1/a, 1/\sqrt{3}a, -2/3c)$, $\boldsymbol{b}_2 = (0, 2/\sqrt{3}a, -1/3c)$, and $\boldsymbol{b}_3 = (0, 0, 1/c)$ indexed by (h, k, ℓ), which correspond to the crystal axes $\boldsymbol{a}_1 = (a, 0, 0)$, $\boldsymbol{a}_2 = (-a/2, \sqrt{3}a/2, 0)$, and $\boldsymbol{a}_3 = (a/2, a/2\sqrt{3}, c)$. $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ define the primitive unit cell. a and c are the hexagonal and vertical lattice constants, respectively. Equivalently, the lattice is described by the reciprocal vectors $\boldsymbol{B}_1 = (1/a, 1/\sqrt{3}a, 0)$, $\boldsymbol{B}_2 = (0, 2/\sqrt{3}a, 0)$, and $\boldsymbol{B}_3 = (0, 0, 1/3c)$ indexed by (H, K, L) and the crystal axes $\boldsymbol{A}_1 = (a, 0, 0)$, $\boldsymbol{A}_2 = (-a/2, \sqrt{3}a/2, 0)$, and $\boldsymbol{A}_3 = (0, 0, 3c)$. The cell defined by $\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3$ contains three primitive unit cells positioned at $(0, 0, 0)$, $(a/2, a/2\sqrt{3}, c)$, and $(0, a/\sqrt{3}, 2c)$. (H, K, L) indices are used in this article.

The detected Bragg peaks for the rhombohedral phase of the lipid diphytanoyl-phophatidylcholine (DPhPC) are shown in Figure 14.17. By the same principle described in Section 14.3.1, the relative phases of the diffraction amplitudes along each vertical column a, b, c, and d in Figure 14.17 can be phased by the swelling method (Yang and Huang, 2003).

After the relative phases within each column are determined, there are 2^3 (8) possible choices for the relative phases between the four columns. We can narrow the choices by using a pattern recognition method proposed by Luzzati and his collaborators (Mariani *et al.*, 1988). This method recognizes the fact that the basic structural element of lipid in water is the lipid monolayer. A change of the structure of a lipid–water system is

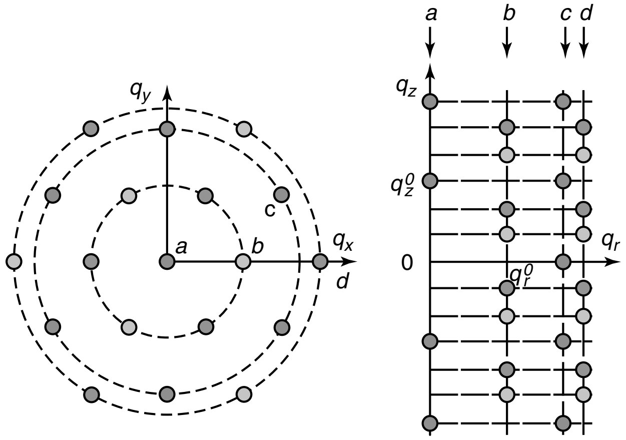

Fig. 14.17 The reciprocal lattice of rhombohedral symmetry on the $q_x - q_y$ plane (left) and the diffraction peaks detected on the $q_z - q_r$ plane (right). The reciprocal lattice of rhombohedral symmetry is composed of a stack of hexagonal layers stacked up in the ABCABC... sequence (A green, B blue, C red). In the $q_z - q_r$ plane are the four columns of the observed diffraction peaks. In Figure 14.18, the swelling method was applied to each set of the peaks having the same (H, K) (same color): column $a(H = 0, K = 0)$, $b(1, 0)$, $c(1, 1)$, and $d(2,0)$. The set of red peaks and the set of blue peaks in each of column b and d are equivalent by centrosymmetry (Yang and Huang, 2003). (Please find a color version of this figure on the color plates.)

a rearrangement of the lipid monolayers. Therefore, if a lipid–water system has more than one structures, the different structures should have comparable electron density contrast histograms $H(\Delta\rho)$. For example, the structure of DPhPC in the L_α phase was known (Hung *et al.*, 2000) and so its $H(\Delta\rho)$ could be used as a reference. For the rhombohedral phase, we computed $H(\Delta\rho)$ for each of the eight possible choices of relative phases and compared the results with that of the L_α phase. For practical purposes, the method was carried out by computing the fourth moment $\langle(\Delta\rho)^4\rangle$ following the procedure of Mariani *et al.*, 1988

$$\langle(\Delta\rho)^4\rangle = \frac{\displaystyle\sum_{H,K,L}\left|\sum_{H',K',L'} F(\boldsymbol{q}_{H'-H,K'-K,L'-L})F(\boldsymbol{q}_{H',K',L'})\right|^2}{\left(\displaystyle\sum_{H,K,L} F(\boldsymbol{q}_{H,K,L})\right)^2} \tag{18}$$

Among the eight possible choices (Figure 14.19), two with $\langle(\Delta\rho)^4\rangle = 2.610$ and 2.640 are closest to the value of the L_α

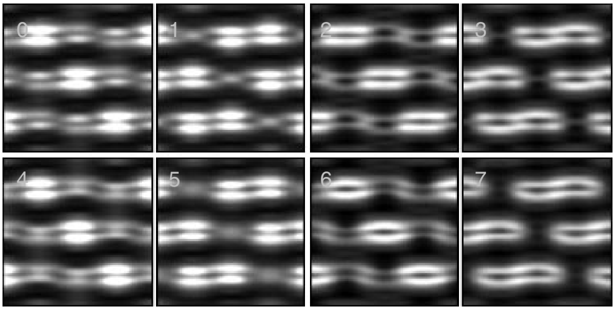

Set #	Relative phase	$<(\Delta p)^4>$
0	++++	4.163
1	+−++	3.625
2	++−+	3.013
3	+−−+	2.610
4	+++−	4.416
5	+−+−	3.550
6	++−−	3.063
7	+−−−	2.640

Fig. 14.18 $\langle(\Delta\rho)^4\rangle$ for the eight possible choices of relative phases are shown in the table on the right. The corresponding electron density maps are shown on the left. The choice 3 and 7 are closest to the value of the L_α phase, which has $\langle(\Delta\rho)^4\rangle = 1.750$. The electron density maps of choice 3 and 7 are very similar, both showing a stalk structure. However, upon close examination, the electron density map of the choice 3 shows somewhat uneven densities for the lipid headgroup along the lipid monolayer (need a close inspection of the density map). Therefore, the choice 7 is considered the most reasonable (Yang and Huang, 2003).

phase, which has $\langle(\Delta\rho)^4\rangle = 1.750$. The electron density distributions of these two choices (3 and 7 in Figure 14.18) are very similar, both show a stalk structure. However, upon close inspection, we found that the choice 3 produced somewhat uneven electron densities in the lipid headgroup region along the monolayers. The most reasonable choice of phases (7 of Figure 14.18) gave the electron density distribution

$$\rho(\boldsymbol{r}) = \frac{1}{V} \sum_{H,K,L} F(q_{H,K,L}) \cos(\boldsymbol{q}_{H,K,L} \cdot \boldsymbol{r}) \tag{19}$$

as shown in Figure 14.19.

The unit cell structure shown in Figure 14.19 suggests that two bilayers under the osmotic pressure merged into an hourglass-shaped structure called a *stalk* (Yang and Huang, 2002). The discovery of this lipid structure validated the stalk-pore hypothesis for membrane fusion (Chernomordik and Kozlov, 2005). This ended the long-time dispute about the mechanism of biomembrane fusion, whether it is lipidic in nature (stalk-pore hypothesis) or it is the proteins that play the dominant role and only make way for lipids at some point late in the process (protein pore hypothesis).

14.5.2.2 Multiwavelength Anomalous Dispersion Method (With Examples in Hexagonal, Distorted Hexagonal, and Rhombohedral Phases)

In protein crystallography the most reliable phase-determining method is the method of multiwavelength anomalous dispersion (MAD) (Grossmann and Arnold, 2001). In this method, one uses an equation relating three quantities, that is, the magnitude of the normal diffraction amplitude of the host molecules, the magnitude of the normal diffraction amplitude of the anomalous atoms, and the relative phase angle between the two amplitudes, with coefficients that are wavelength dependent. A set of diffraction measurements are then recorded at wavelengths below and above the absorption edge of the anomalous atom so as to vary these coefficients as much as possible in order to allow the three unknowns

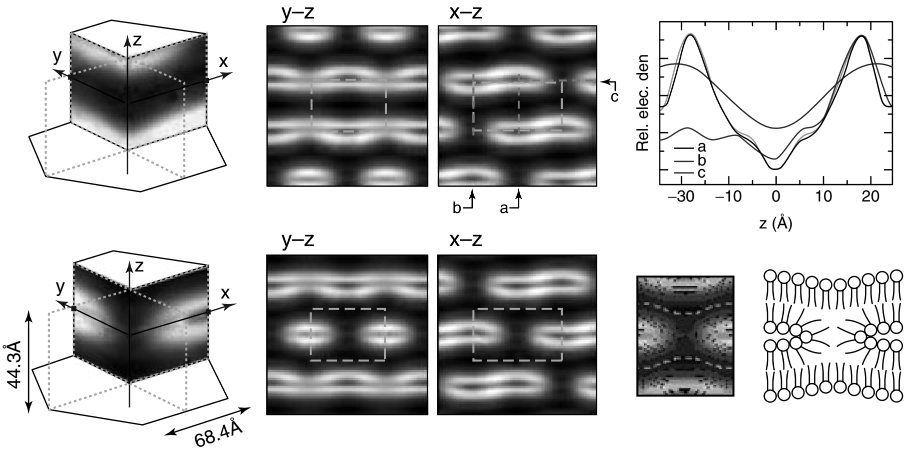

Fig. 14.19 Electron density distribution in the unit cell of the rhombohedral lattice (High density is white, low density is black.) (Top row): The unit cell was based on the phases determined by the swelling and pattern recognition methods. The stalk structure is not obvious in this unit cell. To its right are the density maps in the $y - z$ and $x - z$ planes. The dashed green rectangles are the boundaries of the unit cell in each map. On the $x - z$ map, the electron density profiles were taken along three lines a, b, and c, and shown in the top right panel. Also shown in the panel is a trans-bilayer density profile from the L_α phase. The bilayer portion in the rhombohedral phase remained very close to the L_α phase. (Bottom row): The unit cell was obtained by a translation from the previous one. A stalk structure appeared at the center of the cell. The maps on the $y - z$ and $x - z$ planes are shown to its right. The central part of the $y - z$ map is re-plotted with density contours on the right hand side, along with a cartoon of a talk where two lipid monolayers merged to form a hourglass structure (Yang and Huang, 2003). (Please find a color version of this figure on the color plates.)

to be determined. This standard method can be difficult for lipid samples since their sample thicknesses are usually not precisely known. The correction for X-ray absorption can be significantly uncertain if the absorption coefficient varies a great deal, such as below and above the absorption edge.

To alleviate the difficulty of absorption correction, we suggested a method of using only X-ray energies below the absorption edge, where the X-ray absorption coefficient is relatively small and almost independent of wavelength. It is possible to measure anomalous diffraction at 8–10 different X-ray energies below the edge so as to solve the unknowns by a straight-line fitting instead of solving nonlinear equations. We will demonstrate this method (Wang *et al.*, 2006) with brominated lipid 1,2-distearoyl(9,10-dibromo)-sn-glycero-3-phosphocholine, abbreviated as di18:0(9,10Br)PC, using the K absorption edge of Br to perform the MAD analysis.

It is well known that the absorption edge and the scattering factor near the edge are influenced by the chemical environment of the atom. Therefore, one needs to measure the anomalous scattering factor of the label atom in the actual sample. The absorption curve of di18:0(9,10Br)PC was measured over both sides of the Br K-edge 13.474 keV (Pan *et al.*, 2006). The absorption curve was then converted to the imaginary scattering factor f'' by using the theoretical values calculated by (Cromer and Libermann, 1970). The real part f' was calculated by the dispersion relation using the CHOOCH program by (Evans and Pettifer, 2001). Below the absorption edge, 10 energies were chosen with a step size $\Delta f' = 0.5$ (in the unit of electron) for each successive energy as shown in Table 14.2, where the values of f' and f'' of bromine atom are listed.

The diffraction amplitude from a system containing atoms of an element with anomalous scattering factor $f = f^n(q) + f'(\lambda) + \mathrm{i}f''(\lambda)$ is written as

$$\begin{aligned} F_\lambda &= \sum_j f_j^n \exp(\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}_j) \\ &\quad + \sum_k (f^n + f' + \mathrm{i}f'') \exp(\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}_k) \\ &= F_0 + \frac{f' + \mathrm{i}f''}{f^n} F_2 \end{aligned} \tag{20}$$

Tab. 14.2 Anomalous scattering factor at 10 sub-K edge X-ray energies (K edge at 13.474 keV).

n	En (eV)	f′	f″
1	13 470.7	−7.53	0.87
2	13 468.1	−6.98	0.68
3	13 464.8	−6.51	0.59
4	13 459.0	−5.99	0.54
5	13 449.8	−5.50	0.51
6	13 434.8	−5.00	0.50
7	13 409.1	−4.50	0.51
8	13 366.3	−4.00	0.52
9	13 294.1	−3.50	0.51
10	13 156.7	−3.00	0.52

where $\boldsymbol{q}$ is the X-ray scattering vector. f_j^n is the normal scattering factor of atom j at position r_j. The index j includes all atoms except for the anomalous atoms. The index k includes all the anomalous atoms. F_0 is the normal diffraction amplitude of the whole system. F_2 is the normal diffraction amplitude of the anomalous atoms alone. We will assume that the unit cell of the lipid structure is centrosymmetric, so that both F_0 and F_2 are real (rather than complex) quantities and their phases are the signs of the amplitudes. The subsequent analysis will test whether this assumption is valid. From Eq. (20), we have

$$|F_\lambda|^2 = \left(F_0 + \frac{f'}{f^n}F_2\right)^2 + \left(\frac{f''}{f^n}\right)^2 F_2^2 \tag{21}$$

However, the second term is about 1% of the first term, because at energies below the absorption edge the value of f'' is ~0.5, which is about 10% of $|f'|$ (Table 14.2). Therefore, we have the approximate relation

$$|F_\lambda| \approx \pm\left(F_0 - \frac{|f'|}{f^n}F_2\right) \tag{22}$$

where we have replaced f' by $-|f'|$, since f' is negative below the absorption edge.

The experimental value $|F_\lambda|$ is then plotted against $|f'|/f^n$ for each peak. If the data for every peak satisfy a linear relation, we may conclude that the method is validated. Since Eq. (22) is the combined consequence of (i) both F_0 and F_2 are real quantities and (ii) Eq. (21) is linearized to Eq. (22), the agreement with the data implies that both the assumption of centrosymmetry and the assumption of linear approximation are valid. From the linear plot, one finds the magnitude

$|F_0|$ from the intercept of the fitted line; the magnitude $|F_2|$ from the magnitude of the slope; and the sign of $-F_0/F_2$ from the sign of the slope.

This MAD analysis reduces the phase problem of the whole system (F_0) to the phase problem of the label atoms alone (F_2). The latter problem is much simpler than the original phase problem. And once the phases of F_2's are determined, so are those of F_0's.

Example 1: Hexagonal Phase The inverted hexagonal phase is the simplest nonlamellar phase, yet the phase problem of its diffraction is still very difficult (Turner and Gruner, 1992). This system was the first example investigated by the MAD method described above (Pan *et al.*, 2006). The mixture of di18:0(9,10Br)PC and cholesterol (in 2 : 1 ratio) is in the inverted hexagonal (H_{II}) phase below 44% RH at 25 °C. Figure 14.20a

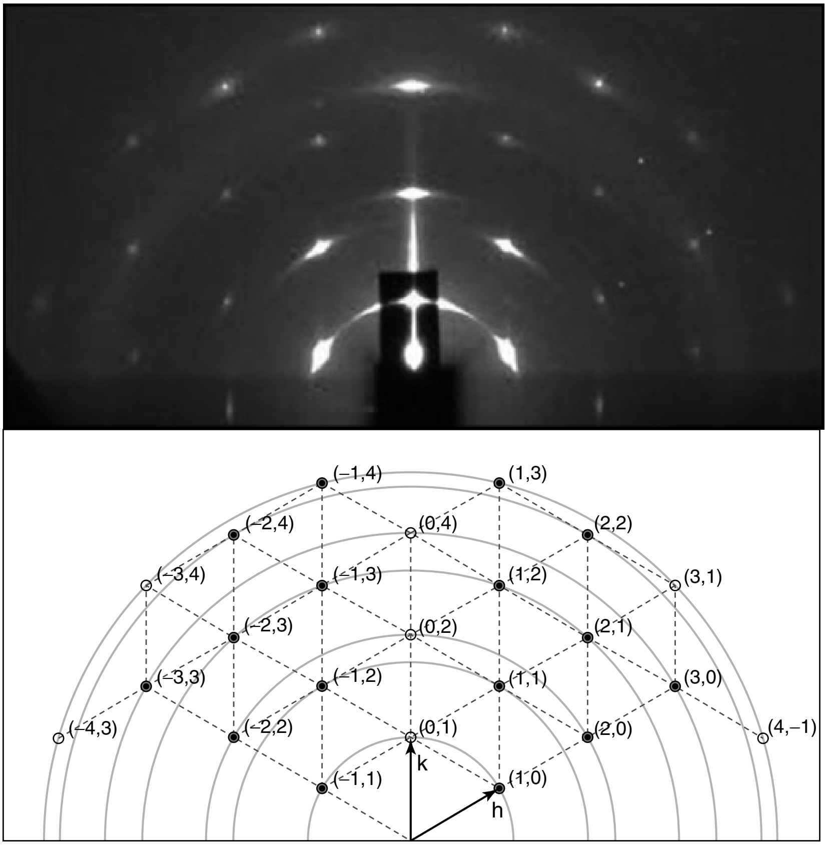

Fig. 14.20 (top) Diffraction pattern of the inverted hexagonal phase measured by reflection. The mixture of di18:0(9,10Br)PC and cholesterol (in 2 : 1 ratio) was prepared in an aligned lamellar phase in high humidity on a solid substrate, and then dehydrated to below 44% RH. (bottom) The detected Bragg peaks on a hexagonal reciprocal lattice. The peaks shown as solid circles were measured and analyzed. The lattice points at open circles are related to solid circles by symmetry (Pan *et al.*, 2006).

shows its diffraction pattern. Anomalous diffraction was measured at eight energies below the Br K-edge. Taking into account of the hexagonal symmetry, the diffraction pattern (Figure 14.20b) gave seven independent peaks. For each independent peak, the measured value $|F_\lambda|$ was plotted against $|f'|/f^n$ (Figure 14.21), which showed linearity within the experimental errors. From the linear fits, we obtained $|F_0|$, $|F_2|$, and the sign of the $-F_0/F_2$. The advantage of the MAD method is that we can work with the simpler distribution of bromine atoms alone (i.e. F_2) rather than with the distribution of the whole lipid (F_0).

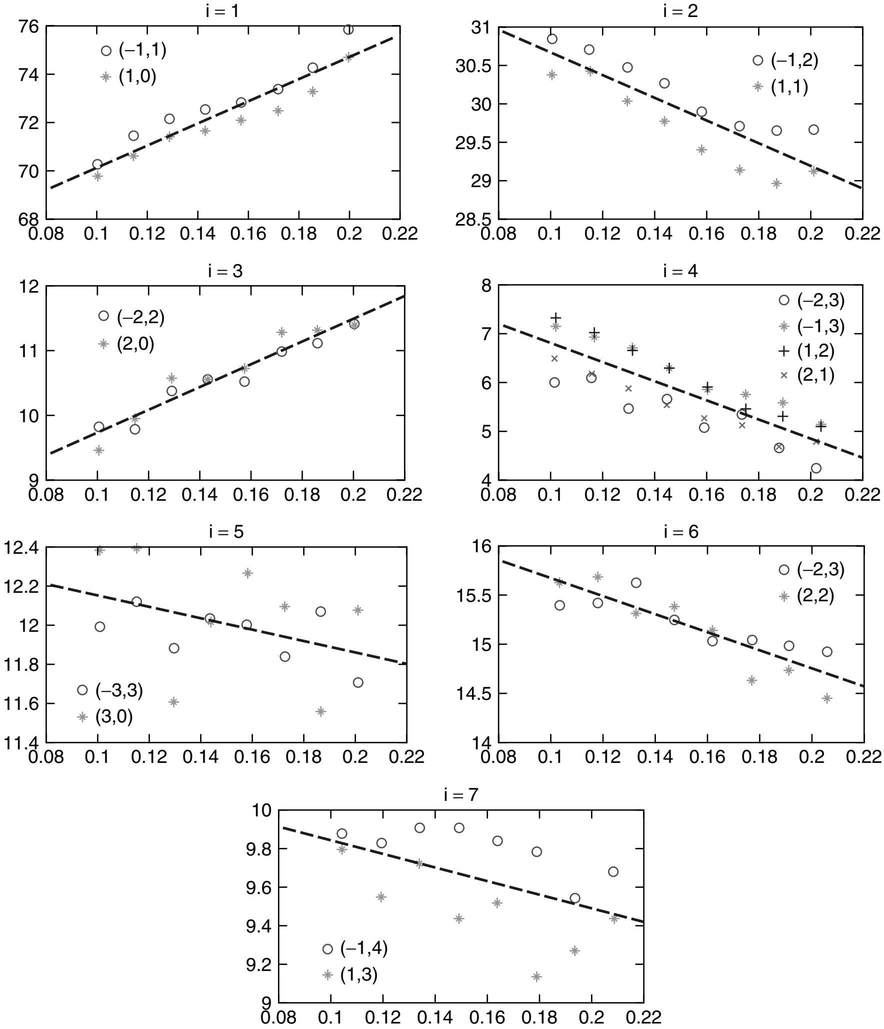

Fig. 14.21 The MAD analysis. For each panel, the ordinate is $|F_\lambda|$ and the abscissa is $|f'_\lambda|/f^n$. From the slope and intercept of the linear fit, we obtain $|F_0|$ and $|F_2|$ and the ratio F_0/F_2 (Pan *et al.*, 2006).

We solved the phase problem by modeling the distribution of bromine atoms. First, the Patterson map was constructed

$$P(r,\theta) = \sum_{(h,k)} |F_2(h,k)|^2 \times \cos\left\{\frac{2\pi}{D} r\left[\left(\frac{\sin\theta}{\sqrt{3}} + \cos\theta\right) \times h + \left(\frac{2}{\sqrt{3}}\sin\theta\right) k\right]\right\} \tag{23}$$

as shown in Figure 14.22. The fact that the Patterson function contains only one single peak implies that the Br distribution is localized at one radius and the width of the peak indicates the width of the Br distribution around that radius. Therefore, we built a model as a circular ring of Br distribution around the center of the unit cell. In a reference unit cell, the model is defined by

$$G[x_c, y_c] = \exp\left[\frac{-\left(\sqrt{(r\cos\theta - x_c)^2 + (r\sin\theta - y_c)^2} - r_0\right)^2}{2\sigma^2}\right] \tag{24}$$

where (r,θ) are the polar coordinates with the origin at the center of the reference unit center. $[x_c, y_c]$, referred to the same origin, are the coordinates of the center of a unit cell, which is either the reference unit cell or one of its adjacent unit cells. Within each unit cell, Br atoms are radially Gaussian-distributed, peaked at the radius r_0 with a width σ. Equation (23) describes the Br distribution in an isolated unit cell. To obtain the Br distribution in a lattice, we need to include the density contribution of the six adjacent unit cells, because the distribution defined by Eq. (23) spreads beyond the boundary of the unit cell. Thus the electron density measured by X-ray diffraction in the [0,0] unit cell is calculated from

$$\begin{aligned}\rho_{\text{mod}}^{\text{Br}}(r,\theta) = {} & G[0,0] + G[D,0] \\ & + G\left[\frac{D}{2}, \frac{\sqrt{3}D}{2}\right] \\ & + G\left[-\frac{D}{2}, \frac{\sqrt{3}D}{2}\right] \\ & + G[-D,0] \\ & + G\left[-\frac{D}{2}, -\frac{\sqrt{3}D}{2}\right] \\ & + G\left[\frac{D}{2}, -\frac{\sqrt{3}D}{2}\right]\end{aligned} \tag{25}$$

The model amplitudes are then calculated by

$$F_{\text{mod}}^{\text{Br}}(h,k) = \int\!\!\int_{r,\theta} \rho_{\text{mod}}^{\text{Br}}(r,\theta) \times \cos\left\{\frac{2\pi}{D} r\left[\left(\frac{\sin\theta}{\sqrt{3}} + \cos\theta\right) h + \left(\frac{2}{\sqrt{3}}\sin\theta\right) k\right]\right\} r\,dr\,d\theta \tag{26}$$

where the integration is over one unit cell. We adjusted the model parameters r_0 and σ until we obtained the best fit between the model amplitudes $|F_{\text{mod}}^{\text{Br}}(h,k)|$ and the seven experimental amplitudes $|F_2(h,k)|$ by maximizing the T function

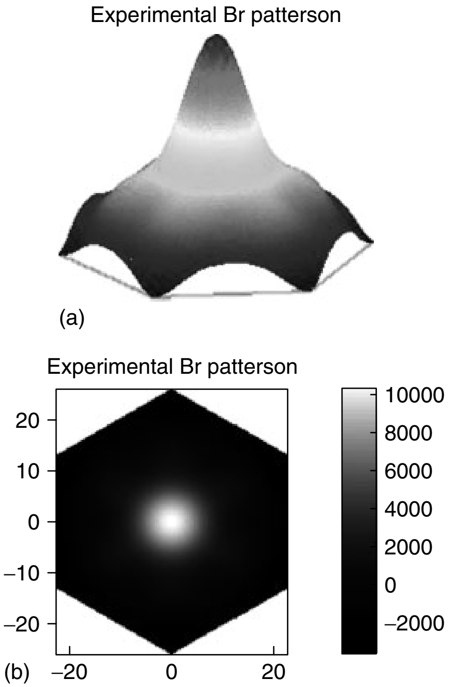

Fig. 14.22 The Patterson function for the bromine distribution: (a) 3D view and (b) 2D top view. The horizontal x, y coordinates are in the unit of Å. The vertical scale for the value of the Patterson function is arbitrary (Pan *et al.*, 2006).

(Pannu and Read, 1996; Saldin *et al.*, 2001)

$$T = \frac{\left[\sum_i (1/\varepsilon_i^2)|F_i^{\text{exp}}| \cdot |F_i^{\text{mod}}|\right]^2}{\sum_i (1/\varepsilon_i^2)(F_i^{\text{exp}})^2 \cdot \sum_i (1/\varepsilon_i^2)(F_i^{\text{mod}})^2} \tag{27}$$

where ε_i is the standard deviation for extracting F_i^{exp} from the raw data during the MAD analysis. The value of T is ≤ 1, with 1 being the perfect agreement between model and experiment. Our best fit gave a $T = 0.96622$. It is possible to refine the model to improve the T function to 0.97229 (Pan *et al.*, 2006) but it did not change the phases determined from the initial fit. These phases were then used to construct the experimental electron density distribution for the Br atoms as well as for the whole lipids (Figure 14.23) by the equation,

$$\rho_{\text{exp}}(r, \theta) = \sum_{(h,k)} F_{\text{exp}}(h, k) \times \cos\left\{\frac{2\pi}{D} r\left[\left(\frac{\sin\theta}{\sqrt{3}} + \cos\theta\right) h + \left(\frac{2}{\sqrt{3}}\sin\theta\right) k\right]\right\} \tag{28}$$

The undulation of Br distribution around the ring (Figure 14.23) was observed for the first time, and it contained useful information. This was due to confining a circular lipid tube in a hexagonal pipe. The inverted hexagonal phase has been used to study the chain packing in a hydrophobic interstice where three monolayers meet. Hydrophobic interstices are of great interest, because they

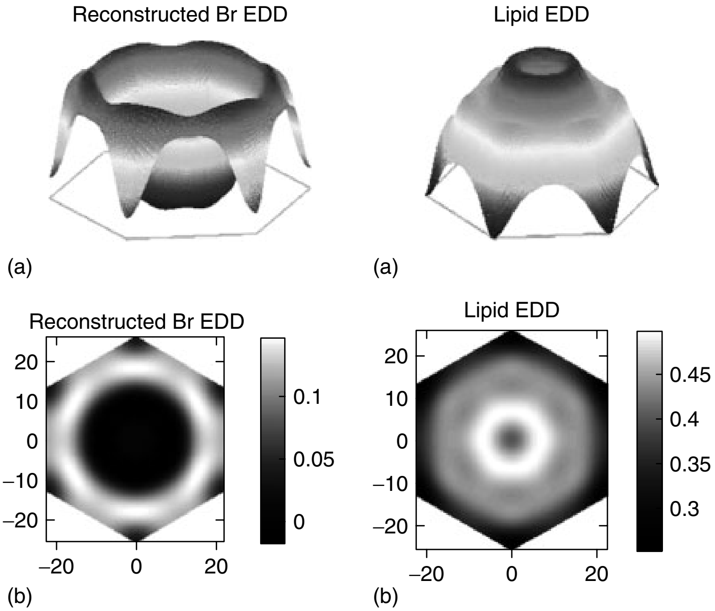

Fig. 14.23 (Left) Electron density distribution of bromine atoms reconstructed from the diffraction amplitudes: (a) 3D view and (b) 2D top view. (Right) Electron density distribution of the 2 : 1 mixture of di18:0(9-10dibromo)PC and cholesterol in a unit cell of the inverted hexagonal phase: (a) 3D view and (b) 2D top view. The horizontal x, y coordinates are in the unit of Å. The vertical scale for the electron density is e Å^{-3} (Pan *et al.*, 2006).

occur in the intermediate states of membrane fusion. It is generally believed that chain packing in such a region is energy costly. Consequently it has been speculated that the inverted lipid tube is likely to deviate from a circular shape, and the chain density distribution might be nonuniform. The bromine distribution obtained from the MAD analysis provided the information for the chain packing in the hexagonal unit cell. The intensity of the bromine distribution is undulated around the unit cell. The analysis shows that the lipid chains pack the hexagonal unit cell at constant volume per chain, with no detectable effect from a high-energy interstitial region (Pan *et al.*, 2006).

Example 2: Distorted Hexagonal Phase The mixture of di18:0(9,10Br)PC and cholesterol (in 2 : 1 ratio) studied above also exhibited a distorted hexagonal ($H_{II\delta}$) phase between 44% and 60% RH at 25 °C. We solved the structure of its unit cell to examine the distributions of two lipid components in a monolayer of nonuniform curvature (Wang *et al.*, 2007). Recent experiments suggested that cholesterol and other lipid components of high negative spontaneous curvature facilitate membrane fusion. This is taken as evidence supporting the stalk-pore model of membrane fusion in which the lipid bilayers go through intermediate structures of high curvature. How do the high-curvature lipid

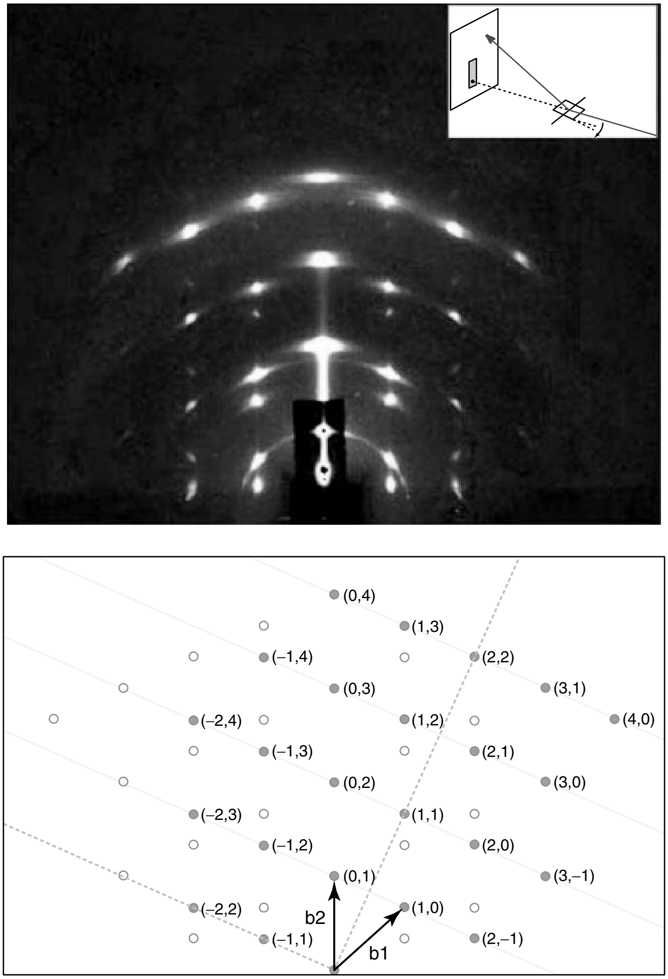

Fig. 14.24 (Top) The detector image of X-ray reflection from a 2 : 1 mixture of di18:0(9,10Br)PC and cholesterol in the $H_{II\delta}$ phase. The inset shows a schematic of the diffraction geometry. (Bottom) The diffraction pattern was translated to the reciprocal space according to the diffraction geometry. The two reciprocal vectors are $\mathbf{b}_1 = (0.1314\text{Å}^{-1} \sin 48.2^\circ, 0.1314\ \text{Å}^{-1} \cos 48.2^\circ)$ and $\mathbf{b}_2 = (0, 0.1314\ \text{Å}^{-1})$, expressed in (q_x, q_z). The angle between the two basis reciprocal vectors is 48.2°, which substantially deviates from the hexagonal angle of 60°. The solid green dots are the lattice defined by $\mathbf{b}_1$ and $\mathbf{b}_2$. The open circles are the mirror image of the lattice reflected by the mirror plane along the q_z and perpendicular to q_x – this left–right symmetry is due to the sample being a 2D powder on the plane of the substrate. Owing to the equality $|\mathbf{b}_1| = |\mathbf{b}_2|$, there are two mirror planes (purple lines) in the crystal lattice, one bisecting $\mathbf{b}_1$ and $\mathbf{b}_2$, another bisecting $-\mathbf{b}_1$ and $\mathbf{b}_2$ (Wang *et al.*, 2007). (Please find a color version of this figure on the color plates.)

components lower the free energy of the curved structure? Do the high-curvature lipid components modify the average spontaneous curvature of the relevant monolayer, thereby facilitate its bending, or do the lipid components redistribute in the curved structure so as to lower the free energy? This question is fundamental to the curvature elastic energy for lipid mixtures.

The diffraction pattern of the distorted hexagonal phase is shown in Figure 14.24, and the MAD analysis is shown in Figure 14.25. Again we used a model fitting to the Br distribution to solve the phase problem. The experimental electron density distributions of the Br atoms and the whole lipids are shown in Figure 14.26. Clearly the brominated lipids are distributed mainly in the low curvature region, leaving the high curvature region to cholesterol. Note that the lipid distribution of the same mixture is homogenous and uniform in the lamellar and hexagonal phases. But in the distorted hexagonal phase the lipid monolayer has nonuniform curvature and cholesterol almost entirely concentrates in the high curvature region. This finding demonstrates that

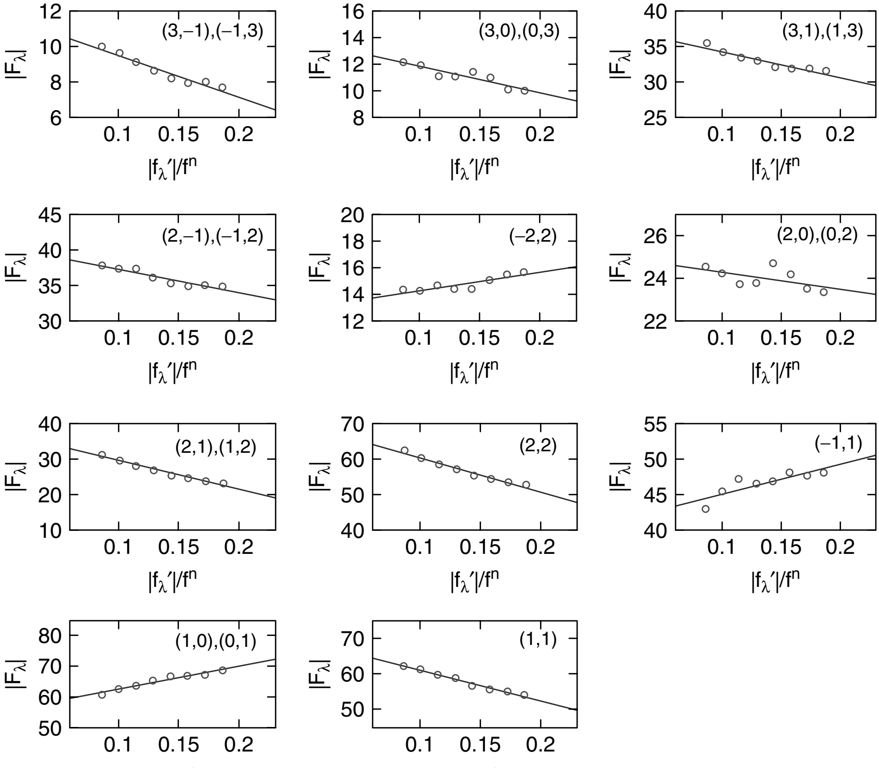

Fig. 14.25 MAD analyses for the detected peaks. For each independent peak, the square-root of the integrated intensity $|F_\lambda|$ is plotted as a function of $|f'_\lambda|/f^n$. The data are fit with a straight line, from which $|F_0|^2$, $|F_2|^2$, and the ratio F_0/F_2 are obtained (Wang *et al.*, 2007).

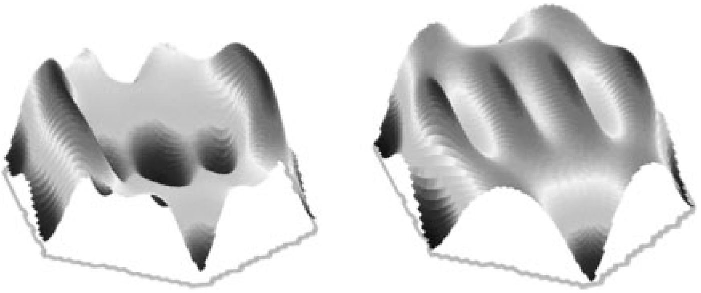

Fig. 14.26 (Left) The electron density distribution of bromine atoms in the unit cell. (Right) The electron density distribution of the di18:0(9,10Br)PC and cholesterol mixture. The high density regions are phosphate headgroups and bromine atoms (Wang *et al.*, 2007).

the association energies between lipid molecules vary with the curvature of membrane. Thus lipid components in a mixture may redistribute under conditions of nonuniform curvature, for example during membrane fusion. In such cases, the spontaneous curvature depends on the local lipid composition and the free energy minimum is determined by lipid distribution as well as curvature.

Example 3: Rhombohedral Phase The MAD method was also applied to a rhombohedral phase. Alamethicin is a well- known antimicrobial peptide. The peptide is known to form pores in membranes by ion conduction (Hall *et al.*, 1984), neutron scattering (He *et al.*, 1995; He *et al.*, 1996), and oriented circular dichroism (Huang and Wu, 1991) experiments, but its pore structure has never been observed. Thus our goal was to use the MAD method to directly observe the pore structure in membrane (Qian *et al.*, 2008).

We first examined alamethicin and di18:0(9,10Br)PC mixtures in full hydration. Using the method of oriented circular dichroism (Wu *et al.*, 1990; He

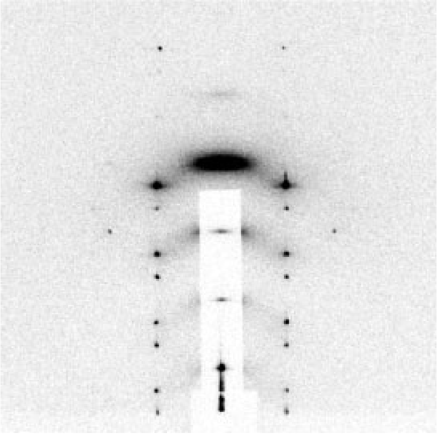

Fig. 14.27 Diffraction patterns of alamethicin/di18:0(9.10Br)PC $P/L = 1:20$ at 40% < RH < 55%, 25 °C. The patterns are of rhombohedral symmetry (space group $R\bar{3}$). The dimensions of the unit cells are $c = 4.95$ nm, $a = 5.94$ nm (the same notation as in Section 14.5.2.1) (Qian *et al.*, 2008).

et al., 1995, 1996), we found that alamethicin formed pores in the lipid bilayers if its concentration exceeded 1 : 30 peptide–lipid molar ratio. When such mixtures were dehydrated below 55% RH at 25 °C, the system transformed from the lamellar phase to the rhombohedral (R) phase.

The diffraction pattern of the R phase is shown in Figure 14.27 and the MAD analysis is shown in Figure 14.28. The phase problem was solved by fitting a model to the Br distribution. The experimental electron density of the lipid–alamethicin mixture shows a barrel-stave type of pore (Figure 14.29).

For the first time a peptide-induced pore in membrane is directly imaged by reconstructing its electron density based on X-ray diffraction. Because the

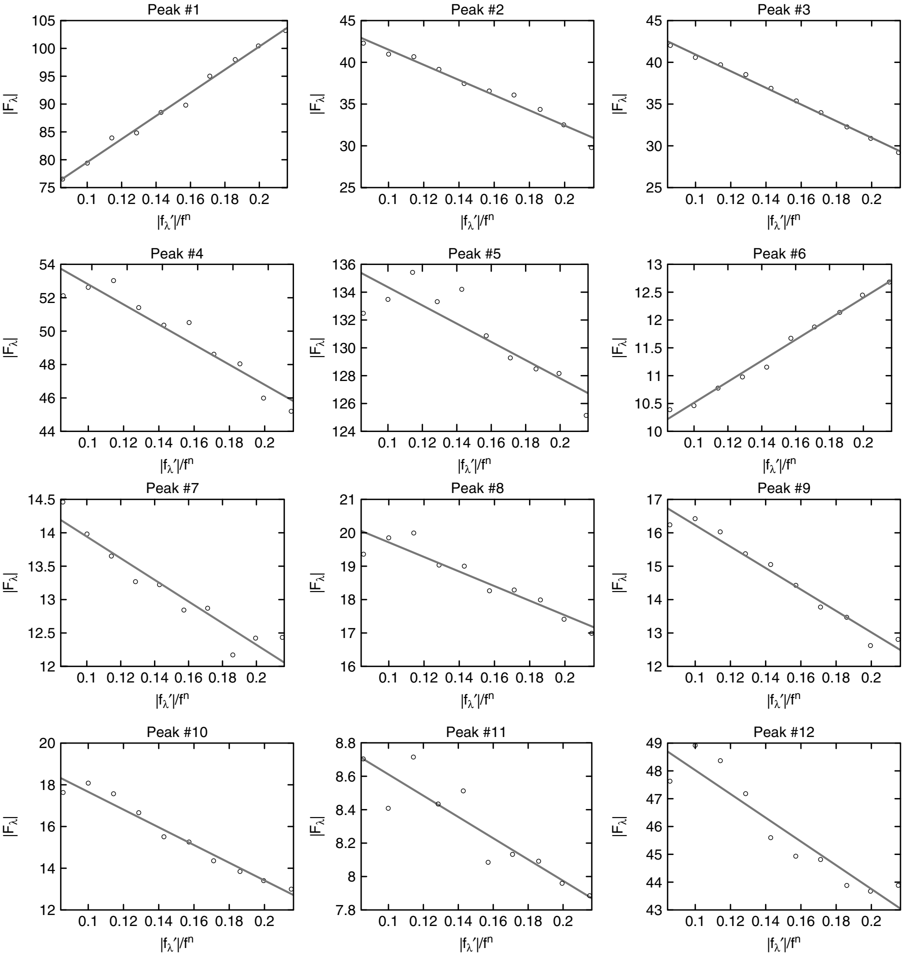

Fig. 14.28 MAD analyses for the detected peaks in Figure 14.27. For each independent peak, the square-root of the integrated intensity, $|F_\lambda|$, is plotted as a function of $|f'_\lambda|/f^n$. The data are fit with a straight line, from which $|F_0|$, $|F_2|$, and the ratio F_0/F_2 are obtained (Qian *et al.*, 2008).

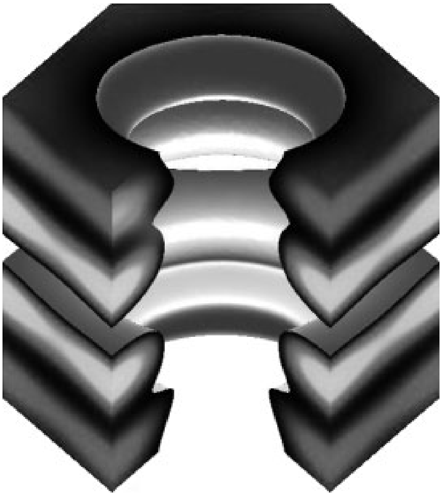

Fig. 14.29 Three dimensional view of a unit cell. One-sixth of the unit cell is removed to show the interior. The surface of electron density distribution is chosen at a density value such that the central hole has a radius 2.0 nm, according to the model. The low-resolution diffraction essentially reveals the layers of phosphate groups (top and bottom layers) and the layers of bromine atoms (the two interior layers). An alamethicin channel of eight monomers arranged in the barrel-stave fashion fits the hole perfectly (Qian *et al.*, 2008).

molecular assembly of the pore is defined by the atomic distribution rather than by atomic positions, and also because soft-matter structures have a high intrinsic disorder, its diffraction pattern is limited to relatively small q values. Only the high density parts are clearly revealed by the low-resolution diffraction. The clearly visible regions include the distribution of phosphate groups of the lipids and the distribution of the heavy atoms bromines that are bound to carbon 9 and 10 of each hydrocarbon chain. Unfortunately the peptide alamethicin does not possess high electron density; therefore their positions are not visible. Nevertheless, the lipid assembly is unambiguously defined by the distributions of the phosphate group and the bromines bound to carbons 9 and 10. The lipid assembly for the alamethicin pore is unmistakably that of a barrel-stave construction. This lipid assembly accommodates a channel made of n alamethicin helixes. From the crystallographic data (Fox and Richards, 1982), we know that an alamethicin monomer is a cylinder of 3.2 nm long and 1.1 nm in diameter. A channel of eight alamethicin helices gives an outside radius of 1.95 nm that fits the lipid assembly perfectly.

References

Biou, V., Bösecke, P., Bois, J.M., Brandolin, G., Kahn, R., Mas, C. *et al.* (2005) X-ray spectroscopy and X-ray diffraction at wavelengths near the K-absorption edge of phosphorus. *J. Synchrotron Radiat.*, **12**(4), 402–9.

Blaurock, A.E. (1971) Structure of the nerve myelin membrane: proof of the low-resolution profile. *J. Mol. Biol.*, **56**(1), 35–52.

Brzustowicz, M.R. and Brunger, A.T. (2005) X-ray scattering from unilamellar lipid vesicles. *J. Appl. Crystallogr.*, **38**, 126–31.

Caille, A. (1972) Physique cristalline: remarques sur la diffusion des rayons X dans les smectiques. *C. R. Acad. Sci. Paris Ser. B*, **274**, 891–93.

Chen, F.Y., Hung, W.C. *et al.* (1997) Critical swelling of phospholipid bilayers. *Phys. Rev. Lett.*, **79**, 4026–29.

Chen, S.H., Liao, C.Y., Huang, H.W., Weiss, T.M., Bellisent-Funel, M.C., Sette, F. *et al.* (2001) Collective dynamics in fully hydrated phospholipid bilayers studied by inelastic X-ray scattering. *Phys. Rev. Lett.*, **86**(4), 740.

Chernomordik, L.V. and Kozlov, M.M. (2005) Membrane hemifusion: crossing a chasm in two leaps. *Cell*, **123**(3), 375–82.

Chiang, H.T., Chenwhite, V.S., Pindak, R., and Seul, M. (1995) Ordering transitions in thin, freely-suspended phospholipid films. *J. Phys. II*, **5**(6), 835–57.

Conn, C., Ces, O., *et al.* (2008). Pressure-jump time-resolved X-ray diffraction study of cubic-cubic transition kinetics in monoolein *Langmuir*, **24**(6), 2331–40.

Cromer, D.T. and Libermann, D. (1970) Relativistic calculation of anomalous scattering factors for X rays. *J. Chem. Phys.*, **53**, 1891–98.

Daillant, J., Bellet-Amalric, E. *et al.* (2005) Structure and fluctuations of a single floating lipid bilayer. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(33), 11639–44.

Ding, L., Weiss, T.M. *et al.* (2005) Distorted hexagonal phase studied by neutron diffraction: lipid components demixed in a bent monolayer. *Langmuir*, **21**(1), 203–10.

Elliott, J.R., Needham, D. *et al.* (1983) The effects of bilayer thickness and tension on gramicidin single-channel lifetime. *Biochim. Biophys. Acta*, **735**, 95–103.

Evans, G. and Pettifer, R.F. (2001) CHOOCH: a program for deriving anomalous-scattering factors from X-ray fluorescence spectra. *J. Appl. Crystallogr.*, **34**, 82–86.

Fox Jr, R.O. and Richards, F.M. (1982) A voltage-gated ion channel model inferred from the crystal structure of alamethicin at 1.5-A resolution. *Nature*, **300**(5890), 325–30.

Grossmann, M.G. and Arnold, E. (eds) (2001) *International Tables for Crystallography, Vol. F. Crystallography of Biological Macromolecules*, Kluwer Academic Publishers, Dordrecht.

Gruner, S.M. (1989) Stability of lyotropic phases with curved interfaces. *J. Phys. Chem.*, **93**, 7562–70.

Hall, J.E., Vodyanoy, I. *et al.* (1984) Alamethicin. A rich model for channel behavior. *Biophys. J.*, **45**(1), 233–47.

Harper, P.E. and Gruner, S.M. (2000) Electron density modeling and reconstruction of infinite periodic minimal surfaces (IPMS) based phases in lipid–water systems. I. Modeling IPMS-based phases. *Eur. Phys. J. E*, **2**, 217–28.

Harper, P.E., Gruner, S.M. *et al.* (2000) Electron density modeling and reconstruction of infinite periodic minimal surfaces (IPMS) based phases in lipid-water systems. II. Reconstruction of D surface based phases. *Eur. Phys. J. E*, **2**, 229–45.

Harroun, T.A., Heller, W.T. *et al.* (1999) Experimental evidence for hydrophobic matching and membrane-mediated interactions in lipid bilayers containing gramicidin. *Biophys. J.*, **76**(2), 937–45.

He, K., Ludtke, S.J. *et al.* (1995) Antimicrobial peptide pores in membranes detected by neutron in-plane scattering. *Biochemistry*, **34**(48), 15614–18.

He, K., Ludtke, S.J. *et al.* (1996) Neutron scattering in the plane of membranes: structure of alamethicin pores. *Biophys. J.*, **70**(6), 2659–66.

Hladky, S.B. and Haydon, D.A. (1970) Discreteness of conductance change in biomolecular lipid membranes in the presence of certain antibiotics. *Nature*, **225**, 451–53.

Hladky, S.B. and Haydon, D.A. (1984). *Ion Movements in Gramicidin Channels*, (W.D. Stein), Academic Press, New York **21**: 327–72.

Huang, H.W. (1986) Deformation free energy of bilayer membrane and its effect on gramicidin channel lifetime. *Biophys. J.*, **50**, 1061–71.

Huang, H.W. (2006) Molecular mechanism of antimicrobial peptides: the origin of cooperativity. *Biochim. Biophys. Acta*, **1758**(9), 1292–302.

Huang, H.W. and Olah, G.A. (1987) Uniformly oriented gramicidin channels embedded in thick monodomain lecithin multilayers. *Biophys. J.*, **51**(6), 989–92.

Huang, H.W. and Wu, Y. (1991) Lipid–alamethicin interactions influence alamethicin orientation. *Biophys. J.*, **60**, 1079–87.

Hung, W.C., Chen, F.Y. *et al.* (2000) Order–disorder transition in bilayers of diphytanoyl phosphatidylcholine. *Biochim. Biophys. Acta*, **1467**(1), 198–206.

Hunt, J.F., McCrea, P.D., Zaccai, G., and Engelman, D.M. (1997) Assessment of the aggregation state of integral membrane proteins in reconstituted phospholipid vesicles using small angle neutron scattering. *J. Mol. Biol.*, **273**(5), 1004–19.

Levine, Y.K. and Wilkins, M.H. (1971) Structures of oriented lipid bilayers. *Nat. New Biol.*, **230**, 69–72.

Longley, W. and McIntosh, T.J. (1983) A bicontinuous tetrahedral structure in a liquid-crystalline lipid. *Nature*, **303**, 612–14.

Ludtke, S., He, K. *et al.* (1995) Membrane thinning caused by magainin?. *Biochemistry*, **34**(51), 16764–9.

Luzzati, V. (ed) (1968) X-ray diffraction studies of lipid-water systems, in *Biological Membranes*, Academic Press, New York.

Luzzati, V., Gulik-Krzywicki, T. *et al.* (1968). X-ray study of model systems: structure of the lipid-water phases in correlation with the chemical composition of the lipids. *J. Gen. Physiol.* **51**(5), Suppl: 37 S+.

Luzzati, V., Vargas, R. *et al.* (1992) Lipid polymorphism: a correction. The structure of the cubic phase of extinction symbol Fd–consists of two types of disjointed reverse micelles embedded in a three-dimensional hydrocarbon matrix. *Biochemistry*, **31**(1), 279–85.

Marcelja, S. (1976) Lipid-mediated protein interaction in membranes. *Biochim. Biophys. Acta*, **455**(1), 1–7.

Mariani, P., Luzzati, V. *et al.* (1988) Cubic phases of lipid-containing systems. Structure analysis and biological implications. *J. Mol. Biol.*, **204**(1), 165–89.

Miller, C.E., Majewski, J., Gog, T., and Kuhl, T.L. (2005) Characterization of biological thin films at the solid–liquid interface by X-ray reflectivity. *Phys. Rev. Lett.*, **94**(23), 238104.

Olah, G.A., Huang, H.W. *et al.* (1991) Location of ion-binding sites in the gramicidin channel by X-ray diffraction. *J. Mol. Biol.*, **218**(4), 847–58.

Owicki, J.C. and McConnell, H.M. (1979) Theory of protein–lipid and protein–protein interactions in bilayer membranes. *Proc. Natl. Acad. Sci. U.S.A.*, **76**(10), 4750–54.

Pabst, G., Rappolt, M., Amenitsch, H., and Laggner, P. (2000) Structural information from multilamellar liposomes at full hydration: Full q-range fitting with high quality X-ray data. *Phys. Rev. E*, **62**(3), 4000.

Pan, D., Wangchen, W. *et al.* (2006) Chain packing in the inverted hexagonal phase of phospholipids: a study by X-ray anomalous diffraction on bromine-labeled chains. *J. Am. Chem. Soc.*, **128**, 3800–7.

Pannu, N.S. and Read, R.J. (1996) Improved structure refinement through maximum likelihood. *Acta Crystallogr.*, **A52**, 659–68.

Perutz, M.F. (1954) The structure of haemoglobin. III. Direct determination of the molecular transform. *Proc. R. Soc. London, Ser. A*, **225**, 264–86.

Qian, S., Wang, W. *et al.* (2008). Structure of the alamethicin pore reconstructed by X-ray diffraction analysis *Biophys. J.* **94**(9), 3512–22.

Roux, B., Prod'hom, B., and Karplus, M. (1995) Ion transport in the gramidin channel: molecular dynamics study of single and double occupancy. *Biophys. J.*, **68**, 876–92.

Saldin, D.K., Harder, R. *et al.* (2001) Solving the structure completion problem in surface crystallography. *Comput. Phys. Commun.*, **137**, 12–24.

Saxena, A.M. and Schoenborn, B.P. (1977) Correction factors for neutron diffraction from lamellar structures. *Acta Crystallogr.*, **33**, 813–18.

Schroeder, H. (1977) Aggregation of proteins in membranes: an example of fluctuation-induced interactions in liquid crystals. *J. Chem. Phys.*, **67**, 1617–19.

Scriven, L.E. (1976) Equilibrium bicontinuous structure. *Nature*, **263**, 123–25.

Seddon, J.M. and Templer, R.H. (1993) Cubic phases of self-assembled amphiphilic aggregates. *Philos. Trans. R. Soc. London, Ser. A*, **344**, 377–401.

Smith, G.S., Sirota, E.B., Safinya, C.R., and Clark, N.A. (1988) Structure of the L_β phases in a hydrated phosphatidylcholine multimembrane. *Phys. Rev. Lett.*, **60**(9), 813.

Skipski, V.P., Peterson, R.F. *et al.* (1962) Separation of phosphatidyl ethanolamine, phosphatidyl serine, and other phospholipids by thin layer chromatography. *J. Lipid Res.*, **3**, 467–70.

Tanaka, M. and Sackmann, E. (2005) Polymer-supported membranes as models of the cell surface. *Nature*, **437**, 656–63.

Tardieu, A., Luzzati, V. *et al.* (1973) Structure and polymorphism of the hydrocarbon chains of lipids: a study of lecithin–water phases. *J. Mol. Biol.*, **75**(4), 711–33.

Torbet, J. and Wilkins, M.H. (1976) X-ray diffraction studies of lecithin bilayers. *J. Theor. Biol.*, **62**(2), 447–58.

Turner, D.C. and Gruner, S.M. (1992) X-ray diffraction reconstruction of the inverted hexagonal (HII) phase in lipid–water systems. *Biochemistry*, **31**(5), 1340–55.

Wang, W., Pan, D. *et al.* (2006) Method of X-ray anomalous diffraction for lipid structures. *Biophys. J.*, **91**, 736–43.

Wang, W., Yang, L. *et al.* (2007) Evidence of cholesterol accumulated in high curvature regions: implication to the curvature elastic energy for lipid mixtures. *Biophys. J.*, **92**(8), 2819–30.

Warren, B.E. (1969). *X-ray Diffraction*, Dover Publications, Inc., New York, pp. 41–45, 112–14.

Weiss, T.M. van der Wel, P.C.A. *et al.* (2003) Hydrophobic mismatch between helices and lipid bilayers. *Biophys. J.*, **84**, 379–85.

Wilkins, M.H., Blaurock, A.E. *et al.* (1971) Bilayer structure in membranes. *Nat. New Biol.*, **230**(11), 72–76.

Wu, Y., He, K. *et al.* (1995) X-ray diffraction study of lipid bilayer membrane interacting with amphiphilic helical peptides: diphytanoyl phosphatidylcholine with alamethicin at low concentrations. *Biophys. J.*, **68**, 2361–69.

Wu, Y., Huang, H.W. *et al.* (1990) Method of oriented circular dichroism. *Biophys. J.*, **57**, 797–806.

Yang, L., Ding, L. *et al.* (2003) New phases of phospholipids and implications to the membrane fusion problem. *Biochemistry*, **42**(22), 6631–35.

Yang, L. and Huang, H.W. (2002) Observation of a membrane fusion intermediate structure. *Science*, **297**, 1877–79.

Yang, L. and Huang, H.W. (2003) A rhombohedral phase of lipid containing a membrane fusion intermediate structure. *Biophys. J.*, **84**(3), 1808–17.

Yang, L., Weiss, T.M. *et al.* (1999) Supramolecular structures of peptide assemblies in membranes by neutron off-plane scattering: method of analysis. *Biophys. J.*, **77**(5), 2648–56.

Yang, L., Weiss, T.M. *et al.* (2000) Crystallization of antimicrobial pores in membranes: Magainin and protegrin. *Biophys. J.*, **79**, 2002–9.

Yang, L. and Fukuto, M. (2005) Modulated phase of phospholipids with a two-dimensional square lattice. *Phys. Rev. E*, **72**(1), 010901.

Yang, L. (2005) The X21 SAXS instrument at NSLS for studying macromolecular systems. *Macromol. Res.*, **13**, 538–41.

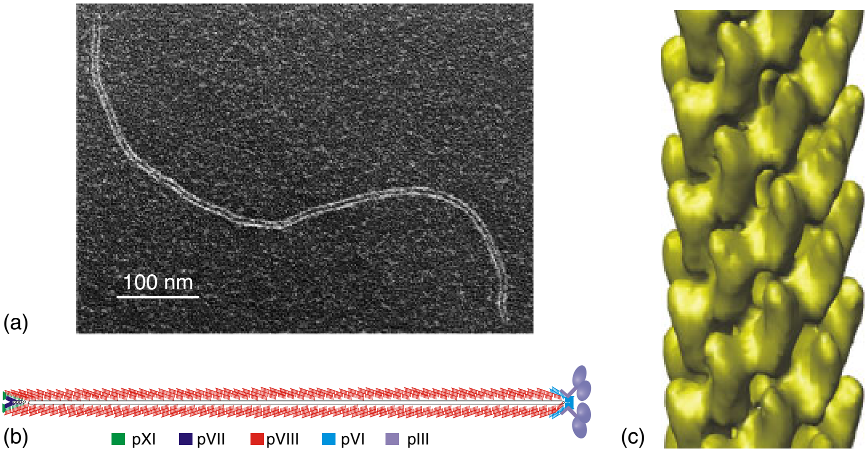

Fig. 13.10 The filamentous ssDNA bacteriophage fd. (a) A negatively stained electron micrograph showing one complete virion. The flexibility of the virion, which is apparent, is consistent with the experimentally determined equilibrium persistence length of 1.0 µm. (b) Cartoon of fd (not to scale) showing locations of the minor coat proteins. (c) Reconstruction of a 24 nm segment of fd at ~8 Å resolution from images obtained by electron cryomicroscopy (courtesy of E. H. Egelman, University of Virginia). (This figure also appears on page 444.)

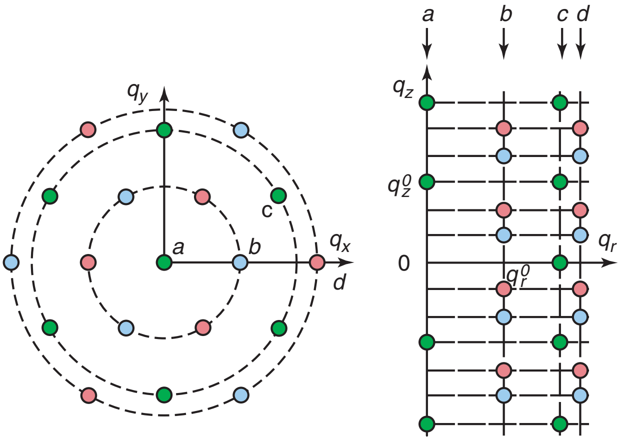

Fig. 14.17 The reciprocal lattice of rhombohedral symmetry on the $q_x - q_y$ plane (left) and the diffraction peaks detected on the $q_z - q_r$ plane (right). The reciprocal lattice of rhombohedral symmetry is composed of a stack of hexagonal layers stacked up in the ABCABC... sequence (A green, B blue, C red). In the $q_z - q_r$ plane are the four columns of the observed diffraction peaks. In Figure 14.18, the swelling method was applied to each set of the peaks having the same (H, K) (same color): column $a(H = 0, K = 0)$, $b(1, 0)$, $c(1, 1)$, and $d(2,0)$. The set of red peaks and the set of blue peaks in each of column b and d are equivalent by centrosymmetry (Yang and Huang, 2003). (This figure also appears on page 485.)

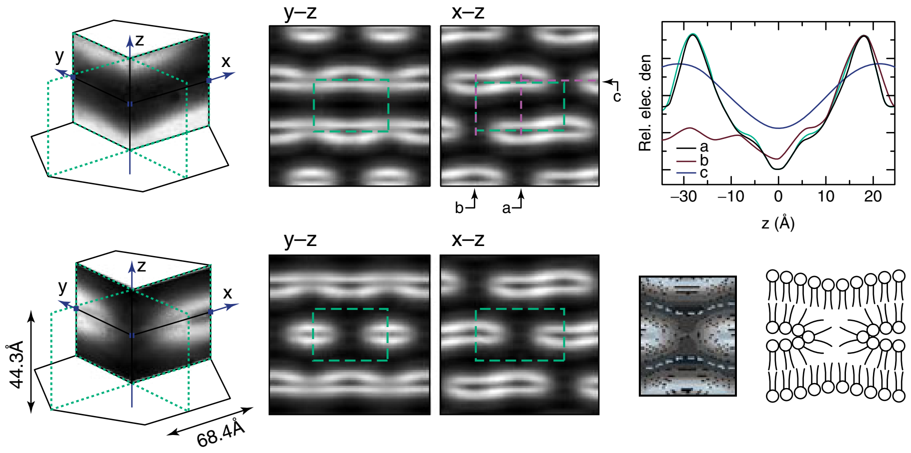

Fig. 14.19 Electron density distribution in the unit cell of the rhombohedral lattice (High density is white, low density is black.) (Top row): The unit cell was based on the phases determined by the swelling and pattern recognition methods. The stalk structure is not obvious in this unit cell. To its right are the density maps in the $y - z$ and $x - z$ planes. The dashed green rectangles are the boundaries of the unit cell in each map. On the $x - z$ map, the electron density profiles were taken along three lines a, b, and c, and shown in the top right panel. Also shown in the panel is a trans-bilayer density profile from the L_α phase. The bilayer portion in the rhombohedral phase remained very close to the L_α phase. (Bottom row): The unit cell was obtained by a translation from the previous one. A stalk structure appeared at the center of the cell. The maps on the $y - z$ and $x - z$ planes are shown to its right. The central part of the $y - z$ map is re-plotted with density contours on the right hand side, along with a cartoon of a talk where two lipid monolayers merged to form a hourglass structure (Yang and Huang, 2003). (This figure also appears on page 487.)

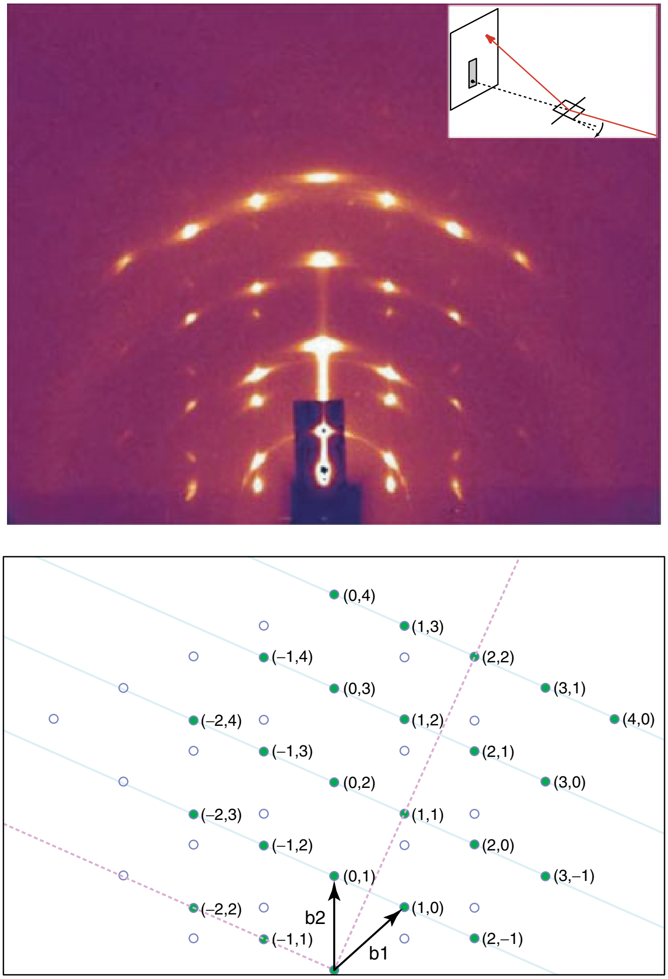

Fig. 14.24 (Top) The detector image of X-ray reflection from a 2 : 1 mixture of di18:0(9,10Br)PC and cholesterol in the $H_{II\delta}$ phase. The inset shows a schematic of the diffraction geometry. (Bottom) The diffraction pattern was translated to the reciprocal space according to the diffraction geometry. The two reciprocal vectors are $\mathbf{b}_1 = (0.1314\text{Å}^{-1} \sin 48.2^\circ, 0.1314\ \text{Å}^{-1} \cos 48.2^\circ)$ and $\mathbf{b}_2 = (0, 0.1314\ \text{Å}^{-1})$, expressed in (q_x, q_z). The angle between the two basis reciprocal vectors is 48.2°, which substantially deviates from the hexagonal angle of 60°. The solid green dots are the lattice defined by $\mathbf{b}_1$ and $\mathbf{b}_2$. The open circles are the mirror image of the lattice reflected by the mirror plane along the q_z and perpendicular to q_x – this left–right symmetry is due to the sample being a 2D powder on the plane of the substrate. Owing to the equality $|\mathbf{b}_1| = |\mathbf{b}_2|$, there are two mirror planes (purple lines) in the crystal lattice, one bisecting $\mathbf{b}_1$ and $\mathbf{b}_2$, another bisecting $-\mathbf{b}_1$ and $\mathbf{b}_2$ (Wang *et al.*, 2007). (This figure also appears on page 494.)

15

The Use of Microwaves in Molecular Biophysics

Henrik G. Bohr, Mogens H. Jacobsen, and I.B. Laursen

15.1 **Introduction** **505**
15.2 **Introduction to Microwave Technology** **505**
15.3 **Rotational Spectroscopy of Molecules** **506**
15.3.1 Linear Molecules 507
15.3.2 Symmetric Tops 507
15.3.3 Spherical Tops 508
15.3.4 Asymmetric Tops 508
15.3.5 Rotational Spectra of Small Molecules from Microwave Experiments 508
15.4 **Microwave Absorption by Protein Molecules** **509**
15.5 **The EM Absorption Spectra of Proteins** **510**
15.5.1 Theory 511
15.5.2 Side Chain Rotors Initiating Wringons 513
15.5.3 Hydrolysis of Proteins and DNA Damages 513
15.5.4 Previous Results 514
15.6 **Experiments on Absorption of Microwaves in Proteins** **515**
15.7 **Results and Discussion** **517**
References **518**

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.

Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

ISBN: 978-3-527-40702-6

15.1
Introduction

A review is given on the usage of microwave technology in biomolecular physics. This mainly concerns the rotational spectroscopy of small organic molecules, since relevant frequencies are in the microwave region, that is, 0.5–1000 GHz. Many applications can be found on microwave spectroscopy, such as air-pollution measurements, molecular investigations in astrophysics, conformational assessments of small molecules, and investigations of quantum-dot molecules (Townes and Schawlow, 1975; Kroto, 1992; Balle and Flygare, 1981; Arunan and Gutowsky, 1993; Emilsson, 1993; Rotational Micro-Wave Spectra, 2008; Larhed and Hallberg, 2001; Dutton, 2002).

Later an analysis is presented on the topological structure of protein backbones with a focus on the differential geometrical aspect of winding and torsion of the peptide chain. By a careful study of the three-dimensional crystallographic coordinates of the backbone atoms for known proteins, the geometrical analysis provides profiles of local winding as well as global twists throughout the protein backbones. From such studies, physical, material science implications are derived for the early stages of protein folding in which ribosome-driven processes and processes where cyclosylation and signal peptides take part, are important. The physical phenomena of such processes include long range excitations of twistons and thermally driven transitions that are part of a topological phase diagram description. Interesting applications for possible protein folding experiments are discussed. The analysis is carried out for specific proteins where clear conclusions are being derived (Bohr and Bohr, 2000a,b; Lax and Button, 1962).

15.2
Introduction to Microwave Technology

In the second half of last century, after the Second World War, electromagnetic radiation in the microwave region of 0.5–1000 GHz has not only been used as a tool for, primarily, detection of molecular structures in chemical research (Arunan and Gutowsky, 1993; Emilsson *et al.*, 2000) but also as a tool for enhancement of chemical reaction kinetics (Bohr and Bohr, 2000a; Larhed and Hallberg, 2001; Dutton, 2002). The reason was first of all the availability

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

of microwave generators due to the invention of the magnetron that provided a cheap and effective way of generating microwaves of a few gigahertz corresponding to a wavelength of around 10 cm. $\nu = 1\ \text{GHz} \sim \lambda = 30$ cm in air. Thus, with the magnetron more precise radar detection was possible. Later the efficient magnetron found its way into the kitchen household as a cooking device for food. This was due to the fact that the magnetron, generating microwaves at 2.24 GHz, could quickly cook food with a large percentage of water and with a penetration depth of about 3 cm. The reason for this rapid heating was due to rotation of water dielectricum or (alternatively explained) a lining of dipoles according to the E-field oscillation with a peak at around 22 GHz. Micro-wave radiation generated at about 2 GHz would still be in the large tail of the water absorption peak and with a reasonable penetration depth that otherwise would be small right at the peak of absorption. The fact that rotational modes of small molecules could be stimulated by microwaves later became a possible tool for investigating atomic structure of molecules as described below. However, the drawback for MW as means of detection required ultra high vacuum for molecules in the gas phase.

15.3
Rotational Spectroscopy of Molecules

It is well known that electronic dynamics is quantized when electrons are bound to nuclei in the atoms. The quantization of the dynamics of the electron binding results in electrons being confined to distinct orbitals around the nucleus.

Similarly, rotational dynamics is quantized – a fact that is relevant especially to spherically asymmetric molecules and also actually to a spherical nuclei. Hence, the rotational energy and the angular momentum can only take specific, distinct values. This chapter reviews the rotational quantization and the corresponding spectroscopy of molecules. The fact that spherical asymmetry plays a big role in the quantization of rotational degrees of freedom can be seen in the dependance of moment of inertia in energy of rotation. Thus, electromagnetic radiation can change the quantum states, the rotational energy states of a molecule and certainly also give information of the molecular structure, for example, in terms of moments of inertia, by observing the absorption in rotational spectroscopy.

Below in Figure 15.1 is shown a typical experimental rotational spectrum of the HCl molecule in the gas phase (Rotational Micro-Wave Spectra, 2008). The spectrum exhibits a slight asymmetry corresponding to the asymmetry of the molecular structure due to the difference in mass of H and Cl. Along the vertical axis is intensity of absorption and horizontal line represents the wavelength.

The structure of these rotating molecules is importantly given, as mentioned above, by moments of inertia. Any molecule in three dimensions will in general have three moments of inertia: I_A, I_B, I_C about three mutually orthogonal axes: A, B, C and with an origin being the center of mass of the molecular system.

Some cases of symmetry are topically described by these three moments. For example, linear molecules with cylindrical symmetry will have negligible moments of inertia along the rotational axis

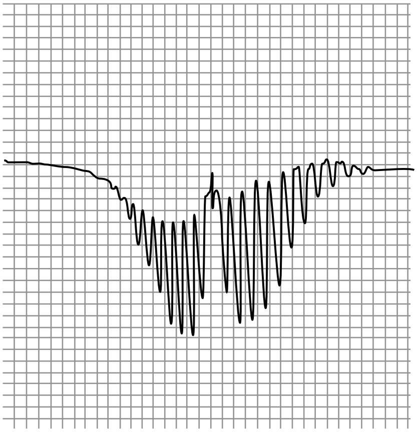

Fig. 15.1 A rotational spectrum of an HCl molecule in the gas phase (Rotational Micro-Wave Spectra, 2008).

A and therefore have ($I_A \ll I_B = I_C$). Thus the molecules' rotational behavior can be used to classify molecular structure and conveniently one can use four structural classes (Townes and Schawlow, 1975; Kroto, 1992) described below in subsections:

15.3.1
Linear Molecules

Here, one of the moments of inertia is negligible, for example, $I_A \ll I_B = I_C$. Often one sets $I_A = 0$. Therefore, the line separation in the rotational spectra can be uniquely related to the linear extension of the molecule. One of the most prominent cases of linear molecules are diatomic molecules such as, for example, H_2, O_2, $O \equiv C$ and also acetylene $HC \equiv CH$.

In order to interpret the rotational spectra, we have to specify the rotational energy levels. The energy, $E(J)$, for the linear molecules (rigid rotors) are quantized with a rotational quantum number, J, as:

$$E(J) = K_r J(J+1), \quad J = 0, 1, 2, \ldots \tag{1}$$

where K_r is a rotational constant of the molecule and related to the moments of inertia ($I_B = I_C$) by

$$K_r = \frac{h}{8\pi c I_B} \tag{2}$$

and with selection rules $\Delta J = J^1 - J^2 = \pm 1$ during absorption or emission. The transition lines in the rotational spectrum are spaced equally in wave numbers. Figure 15.1 shows transitions between the energy levels in a schematic spectrum of a linear molecule (Kroto, 1992).

15.3.2
Symmetric Tops

These molecular structures are generalization of the first case in the sense that they have two equal moments of inertia but a third one not negligible, for example, $I_A = I_B < I_C$. The symmetric tops are again divided into two classes: oblate symmetric tops (disc shaped) like above with $I_A = I_B < I_C$ and prolate symmetric tops (cigar shaped), with $I_A < I_B = I_C$. Here again, the spectra are very recognizable and bonding geometry can be deduced from those. Examples are benzene (oblate) C_6H_6 and chloroform (prolate) $CHCl_3$.

In the case of the symmetric top, one needs two independent rotational quantum numbers J, K to describe the energy, $E(J)$. The quantum numbers J, K correspond to rotations around the two independent axes B, A (prolate) or C (oblate). Actually, the number J is here associated with the total angular momentum and K is the number corresponding to the axis with different moment of inertia either in the oblate or the prolate case. The rotational energy is then in quantized form:

$$E(J, K) = K_r^1 J(J+1) + (K_r^2 - K_r^1)K^2 \tag{3}$$

with $J = 0, 1, 2, \ldots$ and $K = -J, -J+1, \ldots, -1, 0, 1, \ldots, J-1, J$ and where $K_r^1 = (h/8\pi c I_B)$ and $K_r^2 = (h/8\pi c I_C)$ (oblate) or $K_r^2 = (h/8\pi c I_A)$ (prolate) are the rotational constants of the molecule related to the moments of inertia I_B, I_C, or I_A.

The selection rules are in this case $\Delta J = \pm 1$ and $\Delta K = 0$ during absorption or emission. The reason for $\Delta K = 0$ is because the molecule is symmetric around the corresponding axis and therefore has no resulting dipole moment along that axis. In the rotational spectrum, one observes the same lines as in the case of linear molecules but with an extra "dimension" in the rotational spectrum and again with lines spaced equally in wave numbers (Rotational Micro-Wave Spectra, 2008).

15.3.3
Spherical Tops

This case, being a special case of 2, has all moments of inertia equal $I_A = I_B = I_C$. Thus, there are no specific spectral characteristics. Examples of that molecules are carbon tetrachloride CCl_4 and phosphorus tetramer P_4.

15.3.4
Asymmetric Tops

In this case, all moments of inertia are different. Most molecules will have such an asymmetrical structure and spectroscopically there will be no clear patterns in the spectra. It is, however, possible that the molecule can be divided into substructures that each satisfies one of the above symmetries and in that case one can still read out structural properties of the spectra. On the other hand, smaller molecules can form larger aggregates thereby making a structural deduction from the spectra very difficult. This is the case with, for instance, water. Examples of asymmetric tops are nitrogen dioxide NO_2, water H_2O and nucleic acids or amino acids.

15.3.5
Rotational Spectra of Small Molecules from Microwave Experiments

Very elaborate experimental constructions have been made for determining the rotational spectra and hence the accurate structures of smaller molecules in the gas phase. This involves large apparatus, often cylindrical containers maintaining the sample molecule in ultra high vacuum irradiated with microwaves at different frequencies around a few gigahertz (Balle and Flygare, 1981). The fact that rotational spectra can be recorded for, for example, organic molecules makes it possible to derive parts of the structural data about the molecule. For example, it is possible to distinguish between linearly extended molecules and symmetric tops according to the paragraphs mentioned above. For the main class of asymmetric tops a large number of structural data can be derived with the help of computer algorithms. Therefore, a large amount of research has been produced in the area of structural chemistry (Arunan and Gutowsky, 1993; Emilsson *et al.*, 2000). Such methods are, however, hard to carry out for large biomolecules.

One apparatus for detecting the rotational states of smaller molecules in the gas phase has been the pulsed nozzle Fourier transform microwave spectrometer developed 30 years ago (Balle and Flygare, 1981) but improved by Gutowsky and coworkers (Arunan and Gutowsky, 1993; Emilsson *et al.*, 2000). The apparatus has been especially useful for locating the hydrogen atoms in molecules and complexes due to the sensitive detection of small hyperfine interactions via the quadrupole couplings and spin–spin interactions. Figure 15.2 shows the rotational spectra of *benzene* – $C_6H_6 + H_2O$.

15.4 Microwave Absorption by Protein Molecules

The complex process of protein folding has many interesting facets (Anfinsen and Scheraga, 1975; Dobson and Fersht, 1995; Creighton, 1993; Levitt, 1983; Wolynes, 1990; Larhed and Hallberg, 2001; Dutton, 2002). One interesting aspect that has not been much addressed is to what extent external radiation can cause conformational effects. It has been proposed (Bohr and Bohr, 2000a; Bohr, Bohr and Brunak, 1996) and demonstrated (Bohr and Bohr, 2000b) earlier that microwaves can cause such effects through a stimulation of the coherent intrinsic dynamics in biomolecular chain molecules, for example, DNA, RNA, and proteins. Such behavior is consistent with the view that protein folding can be considered as the result of a competition between torsion and bending forces, and not only driven by entropy (Bohr and Bohr, 2000a; Bohr, Bohr and Brunak, 1996). Folded structures that are

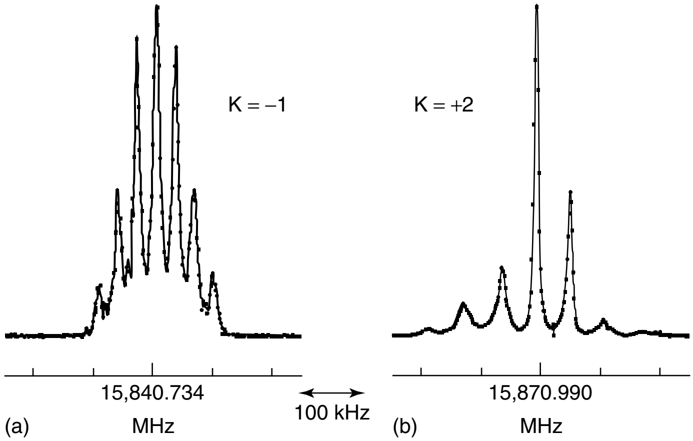

Fig. 15.2 A constructed picture of a spectrum in a large region containing rotational, vibrational, and electronic excitation, all appearing as absorption spectroscopy (Arunan and Gutowsky, 1993).

stabilized by hydrogen bonds, disulfide bridges, and so on, can maintain a significant fraction of folding in balance with unfolded chains, and the biological evolution of living organisms may be considered partly as a selection of proteins with the ability to acquire stable conformations.

An experimental proof for absorption of microwave radiation in protein molecules is given in Chapter 6, see also (Bohr, Jacobsen and Laursen, 2007). For being able to eliminate well-known absorbance of microwave radiation in water, the protein sample was carefully freeze dried and filtered to contain only a few coater molecules per protein molecule and then dissolved in carbon tetrachloride and with a sensitive electron spin resonance (ESR) microwave cavity was shown to exhibit a nonzero absorbance coefficient ε of the relative dielectric constant. The maximum of this absorbance was close to 2.3 GHz in frequency.

Previously, microwave effects have been reported in studies of bond breaking in plasmid DNA (Sagripanti, Swicord and Davis, 1987), in a method for selective peptide-bond cleavage of proteins (Wu *et al.*, 1992; Ramos, Winter and Kielland-Brandt), and in immunostaining of cytopathology specimens (Reynolds *et al.*, 1994). It is expected that these effects certainly include a component that is nonthermal (Bohr and Bohr, 2000b).

15.5
The EM Absorption Spectra of Proteins

Microwave spectra of proteins give unexpected new information about the various modes one can find in axial-symmetric chain molecules.

The much studied part of the electromagnetic spectrum is that of vibrational spectra of biomolecules in the infrared region and, to some extent, the rotational spectra in the higher radio frequency region of 1–10 MHz. The infrared spectra of biomolecules contain the bond vibrations typical for the carbon geometry such as the amide band that is a signal for secondary structures in protein molecules.

The region of 1–10 GHz, being in the lower end of microwave frequencies is not usually expected to entail phenomena concerning biomolecular structures and thus not expected to give rise to any absorption when the biomolecules are being irradiated with microwaves. Absorption is caused only if the solvent contains water as the most common constituent of biological solvents. Water has a strong ability to absorb microwaves due to the rotational energy of the water dielectric giving rise to a large dominant peak around 22 ± 20 GHz, the latter number being due to dispersion. As mentioned above, amino acids have, as can be calculated, rotational spectra in the gigahertz region. Experimentally, the amino acid, glycine, has been found to have a line at around 1 GHz.

In order to distinguish this large factor of absorption by water, which (often) override any phenomena that might be due to other phenomena of microwave irradiation on biomolecules, one must carefully select a solvent that had no absorption of microwave radiation, that is, transparent to microwaves. At the same time, it has to be a solvent that proteins can be dissolved in, without destroying the basic structure of the polypeptide backbone of proteins, although the proteins might undergo a certain degree of unfolding. Thus, one should choose carbon tetrachloride as a solvent for this purpose.

15.5.1
Theory

A possible reason for why microwave can cause structural changes of proteins when being irradiated can be the excitation of collective intrinsic modes in the protein. One can estimate the frequency of intrinsic torsional modes and show that they are well matched to the technical microwave sources with frequencies in the region of 1–3 GHz.

In physical and biological systems, interactions of biological chain molecules will take place with the aqueous environment, with other chain molecules, with membranes, and so on. For closed chain molecules, such as circular DNA, the topological conservation of White (1969) implies that the sum of the torsion and the writhing is constraint to a constant (the linking number). Owing to the finite viscosity of the aqueous medium, proteins are topological restricted when considered on a sufficient short time scale, a fraction of a microsecond and less. The torsion term can be considered separately as there is not sufficient time for the writhing term in the White equation (White, 1969; Fuller, 1971,1978; Crick, 1976). Coherent intrinsic motion of higher frequencies can therefore take place.

The eigenfrequency, ν, can be estimated from a simple calculation of the energy having a rotational and torsion contribution as follows:

$$E = E_{\mathrm{rot}} + E_{\mathrm{tor}} = \frac{1}{2}\left(\frac{\partial\phi(\sigma)}{\partial t}\right)^2 + \frac{1}{2}\alpha\left(\frac{\partial\phi(\sigma)}{\partial\sigma}\right)^2 \tag{4}$$

where σ is a spatial variable along the backbone chain and $\phi(\sigma)$ is a vector field orthogonal to the chain axis measuring the displacement. I is the moment of inertia and α is the torsion constant. The equation of motion is as follows:

$$\left(\left|I\frac{\partial^2}{\partial t^2} - \alpha\frac{\partial^2}{\partial\sigma^2}\right|\right)\phi(t,\sigma) = 0 \tag{5}$$

One might actually, in more analytical terms, write the Lagrangian and derive the full set of the Euler–Lagrange equations:

$$\begin{aligned} L = {} & \frac{1}{2}\frac{M}{L}\int_0^L \left(\frac{\partial \mathbf{r}(\sigma,t)}{\partial t}\right)^2 d\sigma \\ & + \beta\int_0^L \left(\left|\frac{\partial \mathbf{r}(\sigma,t)}{\partial\sigma}\right| - 1\right)^2 \\ & + \frac{1}{2}k_{\kappa,j(\sigma)}\int_0^L (\kappa(\sigma,t) - \kappa_{j(\sigma)})^2 d\sigma \\ & + \frac{1}{2}\alpha\int_0^L \left(\frac{\partial\phi}{\partial\sigma}\right)^2 d\sigma \\ & + \frac{1}{2}\frac{I}{L}\int_0^L \left(\frac{\partial\phi}{\partial t}\right)^2 d\sigma \end{aligned} \tag{6}$$

where k_κ is the bending rigidity constant and α the torsional constant. The bending is represented by the curvature $\kappa = |(\partial^2\mathbf{r}/\partial\sigma^2)|_n$ and the torsion by $\tau = (\partial\phi/\partial\sigma)$. L is the length of the string and M is the total mass. The bending process in the folding of the protein involves several energy minima $j(\sigma)$ parametrized along the string.

The equation of motion (the Lagrange–Euler equation) for the string then becomes:

$$\frac{M}{L}\frac{\partial^2 \mathbf{r}(\sigma)}{\partial t^2}+\frac{I}{L}\frac{\partial^2 \phi(\sigma)}{\partial t^2}=c_1\frac{\partial^2 \mathbf{r}(\sigma)}{\partial \sigma^2}+c_2\frac{\partial^2 \phi(\sigma)}{\partial \sigma^2}+c_3 \qquad (7)$$

We shall study two simple solutions to the corresponding equations of motion by proposing two easy ansatzes. More complicated solutions can be obtained by perturbation around any of the static solutions.

1. Straight rod. The protein as a straight rod is described by all $\kappa_{j(\sigma)}$ being equal to κ_o. In this situation, the elastic restoring force is around the direction of the rod:

$$\mathbf{r}(\sigma)=\mathbf{r}_o+\sigma(\mathbf{r}_L-\mathbf{r}_o)/L \qquad (8)$$

$\mathbf{r}_L$ and $\mathbf{r}_o$ are the start and end positions, respectively. We will now assume that it is physically along the x axis from 0 to L. We obtain the wave solution with the ansatz $\mathbf{r}=\mathbf{r}(\sigma-vt)$ and $\phi=\phi(\sigma-vt)$ that leads to the $\mathbf{r}(\sigma,t)=(L\sigma,0,0)$ solution of longitudinal and transverse oscillations:

Longitudinal oscillations:

$$\mathbf{r}(t\pm v\sigma)=\left(L\sigma+\sum_a a_n \times \cos n(t/v\pm\sigma)/l\pi, 0, 0\right) \qquad (9)$$

Transverse oscillations:

$$\mathbf{r}(t\pm v\sigma)=\left(L\sigma, \sum_b a_n\cos n(t/v\pm\sigma)/l\pi, \sum_c a_n\cos n(t/v\pm\sigma)/l\pi\right) \qquad (10)$$

and torsional oscillations

$$\phi(t\pm v_t\sigma)=\sum b_n\cos n(t/v_t\pm\sigma)/l\pi \qquad (11)$$

We have

a. Pure torsional modes: the rod is fixed while the angular variable can vary in time and space.

b. Pure bending modes: the bending can vary in time and space while the torsion is constantly zero.

2. Helical structure. Here we assume that $\kappa_j(\sigma)$ are equal to κ_o, which means that the restoring force is around a favored curvature. We represent the helical structure by

$$\mathbf{r}=(R\cos(b\sigma), R\sin(b\sigma), a\sigma) \qquad (12)$$

where $R=\sqrt{(1-a^2/b^2)}$ is the radius and a, b are constants. We have

a. Pure torsional modes: the torsion, $\tau=ab$, is changing while the curvature, $\kappa=b\sqrt{b^2-a^2}$, is constant. Stretching the helix leads to a larger radius R.

b. Pure bending modes: the curvature is changing while the torsion is constant. Stretching the helix generally leads to solutions with both smaller and larger radius depending on the exact value of the constants.

The simple version of the equations above is solved with the well-known wave ansatz $\phi(\sigma-vt)$, where the velocity v is given by

$$v=\sqrt{\frac{\alpha}{I}} \qquad (13)$$

The eigenfrequency, ν, can then be derived from the velocity v to be

$$\nu=\frac{1}{L}\sqrt{\frac{\alpha_l}{I}} \qquad (14)$$

where L is the length of the chain molecule that participate in one period of the wring mode, α_l is the torsion constant per inverse unit length and i is the moment of inertia of the backbone per unit length. Typical values are not easy to obtain. For proteins, we may take a rotation of $\pi/2$ to correspond to about 1 eV Å^{-1}, in agreement with common bond energies, although, the energy involved in strained chemical bonds are often smaller. Hence, the torsion constant per inverse unit length, α_l, is about 0.4 eV/. The moment of inertia of the backbone (per unit length), i, is about 100 J s^2, depending on to which degree the side chains are involved in the wringing. A typical value for L is 475 Å, and hence, the frequency ν is in the range of 1–5 GHz. This estimate is based on the assumption of the presence of a linear torsion term. However, as the torsion term is not linear the estimate is only an upper limit. We think that possibly it is a factor of 10 or too large. It should further be noticed that the width of the resonance will be very large because of dispersion. As can be seen, the frequency of such coherent modes in biomolecular chain molecules turn out to be of the same order of magnitude than those typical of radio- and micro-wave applications. Later, we shall give experimental "hints" about the resonance frequency for proteins derived from microwave absorption spectra.

15.5.2
Side Chain Rotors Initiating Wringons

The rotational energy spectrum of the amino acids contains lines that can be related to the wringon-induced twist excitations. Little experimental information is available on rotational spectra of amino acids side chains since they are hard to detect. Often they are referred to as *flip transitions* in the literature. The rotational frequency ω of a side chain group can be calculated from the quantized energy levels (as in Chapter 3):

$$\frac{\hbar^2}{2I}J(J+1), \quad J = 0.1, 2, \ldots \tag{15}$$

where I is the moment of inertia. One can calculate the rotational spectra for various amino acids with a standard program (Emilsson, 1993).

Both tyrosine and phenylalanine have low-level transitions in the interested region and especially the tyrosine line at 1.6 GHz and the phenylalanine line at 2.1 GHz could be the origin of the twist modes. It is appealing to assume that side chain rotors, for example, peptides are important and operating as oscillators for initiating topological resonance phenomena on the polypeptide backbone. This is because nature can exploit the fact that the square of the amplitudes of separate oscillators add up in a constructive interference phenomenon.

15.5.3
Hydrolysis of Proteins and DNA Damages

As mentioned in Chapter 4, bond breaking effects have been detected in a study on microwave radiation applied to plasmid DNA (Sagripanti, Swicord and Davis, 1987). In another study, a method has been developed for rapid and selective peptide-bond cleavage of proteins in a weak acidic solution (Wu *et al.*, 1992).

The energy of a single quanta of microwave is negligible when compared with chemical bonding energies and consequently microwaves is not characterized as ionizing radiation. Therefore, at first it may seem surprising that the electromagnetic energy deposited in DNA can be concentrated and focused in such a way that the energy subsequently can be released for the breaking of a single/double strand. The topological wring modes described above can explain this. Before the breaking of the strand, the energy of the wring mode is distributed along the chain molecule. However, locally the angular strain may lead to conformational changes of the DNA and eventually to breaking. As the DNA breaks, it can no longer support the wring mode that induced the breaking, and the energy is released in the breaking process by a topological process. In Bohr and Bohr (2000a), an illustrative comparison is made between chain molecules (e.g. proteins/DNA) and that of chains of springs, which thereby provide a possible mechanism and explanation for the bond breaking or hydrolysis by microwaves.

15.5.4
Previous Results

In the following is given indirect evidence for microwaves having effect on protein folding. Experiments were carried out on β-lactoglobulin, which was chosen as a test protein because of the accessibility of the cold denaturation, folded, and hot denaturation states. In a 4 M urea solution, the protein has a cold denaturation to folded state transition around 20 °C. The protein is derived from cow milk and has a barrel-like structure. The pronounced content of secondary structure elements is sensitive to polarization measurements as obtained from the Optical Rotational Dispersion (ORD) experiment. Hence, working with β-lactoglobulin provides the advantage of working with the easily accessible cold denaturation phase so that all refolding experiments can exclusively deal with cold denaturation and avoid hot denaturation which often involve protein aggregation. The basic refolding experiment for β-lactoglobulin consists of bringing the sample from the cold denaturation state over to the folded state by raising the temperature from 5 to 44 °C and then back to the cold denaturation state at 5 °C, where the system is allowed to completely equilibrate over about 3 hours. When this refolding experiment is done during a matter of a few hours the heating–cooling cycle will be monitored as a kinetic loop (Figure 15.3). If the measurements at each temperature step are performed with the sample at perfect equilibrium

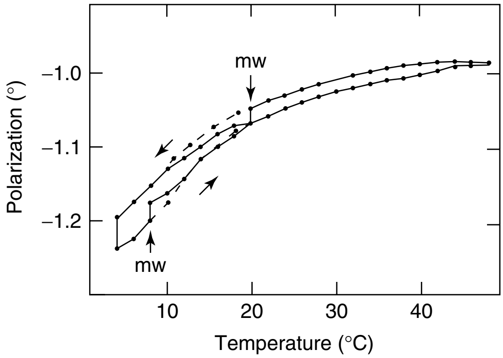

Fig. 15.3 A sketch of the heating–cooling cycle together with applied microwave radiation on a protein (β-lactoglobulin) sample in an aqueous solution. Bohr and Bohr (2000b).

(equilibrium between the concentration of protein in the denatured and folded state) the loop will collapse to a line of polarization values as a function of temperature. The appearance of a kinetic hysteresis loop is a result of the slow-reaction kinetics, and the measurements of the loop therefore allows to study the abrupt changes in the reaction kinetics. The temperature can be monitored inside the sample holder of the polarimeter by a thermocouple wire.

Then, in the next series of experiments the previously described heating-cooling cycles are studied when the microwave irradiation is applied to a protein sample at certain temperatures during the refolding and denaturation process.

The microwave radiation was provided from a standard magnetron of a microwave oven operating at 2.45 GHz and with an effect of 800 W. The microwave oven was remodeled so that it could operate on the sample holder of the polarimeter while it was constantly in contact withthe heath-bath with a circulating silicone liquid. Inside the oven, several containers, containing in all 2 l of water, were placed in order to absorb some of the microwave effect released during the operation. Then microwaves were applied, the sample holder was inserted into the microwave oven without disconnecting it from the heath-bath. The temperature was monitored before and after microwave irradiation. The microwave irradiation was applied for 5 seconds at 800 W causing a temperature increase of the protein sample of about 0.3 °C. The power absorbed by the 3 ml sample was about 0.75 W. Quickly after the microwave irradiation, the sample holder is put back into the polarimeter and the polarization is remeasured.

The previous experiment is repeated starting with the denatured β-lactoglobulin solution equilibrated at 4 °C. First, the temperature is increased to 48 °C without the application of microwaves. Then, the temperature is lowered back to 4 °C while at 20 °C microwaves is applied. After reequilibration at 4 °C, the temperature is raised to 48 °C again. At 8 °C microwaves are applied. The small difference in the polarization values when compared with Figure 15.2a is due to a slight difference in protein concentration in the sample.

Beside the old experiments already published by Bohr and Bohr (1998, 2000b), one has been able to continue with the experiments, trying to change the frequencies of the applied microwave irradiation in the experiment above and hence try to establish the optimal frequency with the largest microwave effects for protein folding/denaturation. Thus, a microwave generator has been applied to a changeable microwave cavity being able to operate approximately in the range of 2–3 GHz but at much lower effect, approximately 10 W, constantly applied in the whole range. The result of the experiment carried out, at the same protein sample and with the same solution as that mentioned above, is shown in Figure 15.3 and shows a clear optimum around 2.4, which corresponds to roughly the same parameters as mentioned or proposed in Bohr and Bohr (2000a).

15.6
Experiments on Absorption of Microwaves in Proteins

The experimental setup is a modified S-band spectrometer for ESR, consisting

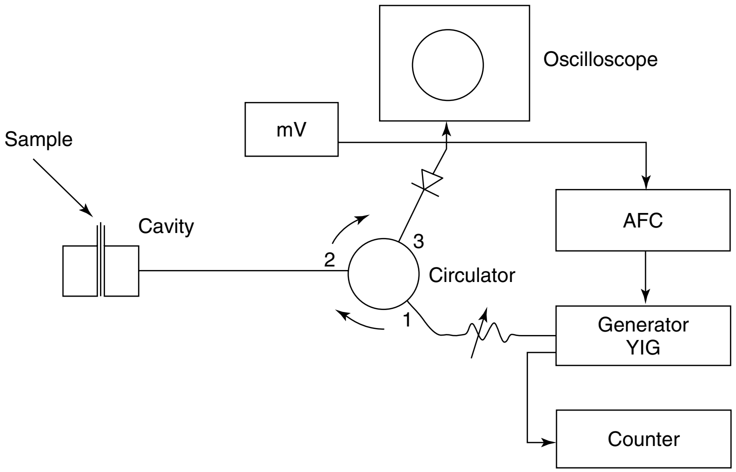

Fig. 15.4 A picture of the experimental setup for measuring microwave absorption of large molecules.

mainly of a YIG-oscillator, a circulator, a high-Q cavity, a counter, and a diode and a digital millivoltmeter (Figure 15.4). Power from the oscillator is lead through the circulators port 1 and 2 to the cavity, which is coupled near critical. The power reflected from the cavity passes the circulators port 2 and 3 to be detected by the diode. The resulting DC/AC voltage V is measured by the DC-millivoltmeter and a scope.

By means of a control circuit connected to the oscillator (AFC), frequency setting and frequency modulation is adjusted. The control unit may be operated in two different modes: a tune mode produces a rather broad frequency modulation and is used for introductory adjustment of the oscillator frequency and cavity coupling as seen on the oscilloscope. A lock mode produces a very small frequency modulation and uses feedback from the diode to lock the oscillator frequency to the resonance frequency of the cavity during measurements.

The rectangular cavity is operated in the TE-101 mode, corresponding to a resonance frequency at around 2350 MHz and a vertical E-field, with maximum at the center of the cavity, where the sample tube is situated. Coupling from the coax line to the cavity is achieved by means of a loop at the end of the line, adjustable to obtain a slightly under critical coupling. In this situation, a larger microwave absorption in the sample will produce a larger reflected power from the cavity and a higher numerical value of diode output. Assuming only small perturbations of the total absorption in the cavity, the measured changes of diode voltage V may be considered linearly dependent on sample absorption, $\Delta V = C\varepsilon''$, where the constant C depends on microwave power, cavity coupling and Q, sample filling factor and diode sensitivity. However, instead of calculating C, we compare $\Delta V(p)$ for protein with $\Delta V(w)$ measured after addition of 10 μl of water.

Therefore, the procedure was to measure first the diode voltage $V(1)$ for a pure CCl_4 sample and next $V(2)$ for CCl_4 + protein. The difference $\Delta V(p) =$

$V(1) - V(2)$ is then taken as a measure of absorption due to protein in the sample. Finally, $\Delta V(w)$ corresponding to the difference between (i) pure CCl_4 and (ii) $CCl_4 + 10\,\mu l$ water was measured for comparison.

15.7
Results and Discussion

In Table 15.1 is presented the measured diode voltages V corresponding to power reflected from the cavity at resonance for the cases (i) no sample material, (ii) pure solution, and (iii) solution and protein. The average value at 2.30 GHz is $V(p) = 1.7$ mVDC. For the difference between (i) pure CCl_4 and (ii) $CCl_4 + 10\,\mu l$ water, we found $\Delta V(w) = 2.5$ mVDC, that is, the same order of magnitude as $\Delta V(p)$ for the protein. According to results from optical measurements of attached water in the protein solution, the water content in the protein sample is far less than 10 µl. Therefore, we conclude that the absorption measured for the protein solution must be almost entirely due to the protein itself and not to water impurities in the solution or to water attached to the protein molecules. The increase of absorption $\delta\varepsilon''$ due to the protein itself may be calculated using the known absorption of water for calibration. From (Mansfield, 1994) we find $\varepsilon''(w) = 12.0$ at 3 GHz and 25 °C. When weighing these data with the values of δV measured and the volumes of protein solution and water, respectively, we found $\delta\varepsilon'' = 0.20$.

The eigenfrequency, ν of the cavity is strongly dependent on the exact geometry of the cavity and can be altered by adding metallic pieces on strategic positions of the cavity and sample box. This metallic box has closable holes in the top in which biochemical samples in plastic tubes can be submerged into the inner vacuum of the cavity.

In Table 15.2 is presented the average values of $\Delta V(p)$ at three

Tab. 15.1 Diode voltage corresponding to reflected power from cavity at resonance, and difference $\Delta V(p)$ corresponding to addition of protein to CCl_4.

Time sequence at 2.30 GHz	V(0) mV for Vacuum	V(1) mV for pure CCl_4	V(2) mV for CCl_4 + protein	Difference ΔV(p)
Frequency $\nu = 2.30$ GHz. Average $\Delta V(p) = 1.7$ mV				
T1	89.0	82.5	83.8	1.3
T2	89.0	81.2	84.9	3.7
T3	89.0	83.1	85.1	2.0
T4	87.5	82.6	83.7	1.1
T5	88.0	82.2	82.9	0.7
T6	88.0	82.1	83.6	1.5
Frequency $\nu = 2.36$ GHz. Average $\Delta V(p) = 1.9$ mV				
T1	80.0	68.4	70.5	2.1
T2	80.5	68.6	70.6	2.0
T3	82.5	69.8	71.6	1.8
T4	81.9	69.7	71.4	1.7
T5	81.7	69.3	71.1	1.8
T6	81.2	68.8	70.6	1.8

Tab. 15.2 Change of diode voltage and corresponding protein absorbance at three different frequencies, ν.

Frequency ν(GHz)	V(1) mV for pure CCl_4	V(2) mV for pure CCl_4 + protein	ΔV(p) mV	$\Delta\varepsilon''$
2.167	37.1	37.5	0.4	0.06
2.345	69.7	71.4	1.7	0.20
3.050	51.5	53.4	1.9	0.10

different frequencies. Taking ε'' at 2.345 GHz as reference, one may find the corresponding values for the other two frequencies, utilizing that $\varepsilon''(p)$ is proportional to $\Delta V(p)$ and of the cavity volume, which is again proportional to $\propto \nu^{-3}$ (Lax and Button, 1962). Thus, one finds the values of ε'' indicated in Table 15.2, showing that frequencies of around 2.3 GHz exhibit maximum in absorbance.

The main conclusion of this study is that proteins, that have been carefully dehydrated, do absorb electromagnetic radiation in the microwave range with a maximum absorption at the frequencies around 2.4 GHz. Correspondingly, the ORD measurements under the influence of microwave irradiation, which have been reported in earlier studies (Bohr and Bohr, 2000a,b), is briefly shown in Figure 15.3. A novel observation on ORD with varying microwave frequency shows a maximum effect around 2.35 GHz. The absorption data in Tables 15.1 and 15.2 are in agreement with the ORD data, with respect to the optimal frequency of irradiation being somewhat around 2.4 GHz. The ESR absorption experiments were performed at 10 W generated by AC circuits, amplifiers, and using adjustable MW cavities.

References

Anfinsen, C.B. and Scheraga, H.A. (1975) Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.*, **29**, 205–300.

Arunan, E. and Gutowsky, H.S. (1993) *J. Chem. Phys.*, **98**, 4294.

Baker, D., Shiau, K.S., and Aggard, D.A. (1993) The role of pro regions in protein folding. *Curr. Biol.*, **5**, 966.

Balle, T.J. and Flygare, W.H. (1981) *Rev. Sci. Instrum.*, **52**, 1.

Berman, A.L., Kolker, E., and Trifonov, E.N. (1994) *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 4044.

Bohr, J. and Bohr, H. (1998) *Z. Phys. D*, **40**, 186–89.

Bohr, H. and Bohr, J. (2000a) *The Topology of Protein Folding*, Ch. 4 in Chemical Topology, Gordon and Breach Science Publishers, Malaysia, p. 239–94.

Bohr, H. and Bohr, J. (2000b) Microwave-enhanced folding and denaturation of globular proteins. *Phys. Rev. E.*, **61**, 4310.

Bohr, H. and Wolynes, P.G. (1992) *Phys. Rev. A*, **46**, 5242.

Bohr, J., Bohr, H., and Brunak, S. (1995) Resonator driven protein folding, in *Protein Folds, A Distance Based Approach* (eds H. Bohr and S. Brunak), CRC Press, New York, p. 243.

Bohr, J., Bohr, H., and Brunak, S. (1996) The formation of protein structure. *Europhys. News*, **27**, 50.

Bohr, J., Bohr, H., and Brunak, S. (1997) Protein folding and wring resonances. *Biophys. Chem.*, **63**, 97–106.

Bohr, H., Jacobsen, M.H., and Laursen, I. (2008) Microwave absorption in biomolecules. *Phys. Rev. E.*, Submitted.

Creighton, T.E. (1993) *Proteins,* 2nd edn, Freeman & Company, New York.

Crick, F.H. (1976) Linking numbers and nucleosomes. *Proc. Natl. Acad. Sci. U.S.A.,* **73**, 2639–43.

Dobson C.M. and Fersht A.R. (eds) (1995) *Protein Folding*, Cambridge University Press, Cambridge.

Dutton, G. (2002) *Genet. Eng. News,* **22**, 12.

Emilsson, T. (1993) *Computer Program for Calculation of Rotational Spectra of Molecules,* Noyes Laboratory University of Illinois, Urbana.

Emilsson, T., Gutowsky, H.S., de Oliveira, G., and Dykstra, C.E. (2000) *J. Chem. Phys.*, **112**, 1287.

Fuller, F.B. (1971) The writhing number of a space curve. *Proc. Natl. Acad. Sci. U.S.A.*, **68**, 815–19.

Fuller, F.B. (1978) Decomposition of the linking number of a closed ribbon: a problem from molecular biology. *Proc. Natl. Acad. Sci. U.S.A.*, **75**, 3557–61.

Kroto, H.W. (1992) *Molecular Rotation Spectra,* Dover Publication, New York.

Larhed, M. and Hallberg, A. (2001) *Drug Discov. Today,* **6**, 406.

Lax, B. and Button, K.J. (1962) *Microwave Ferrites and Ferromagnetics*, McGraw-Hill.

Levitt, M. (1983) Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.*, **170**, 723–64.

Mansfield, M.L. (1994) *Nat. Struct. Biol.*, **1**, 4.

Ramos, C., Winter, J.R., and Kielland-Brandt, M.C. (1994) Requirement of the propeptide in *in vivo* formation of active yeast carboxypeptidase Y. *J. Biol. Chem.,* **269**(9), 7006–7012.

Reynolds, G.M., Young, F.I., Young, J.A., Williams, A., and Rowlands, D.C. (1994) Microwave-oven antigen retrieval applied to the immunostaining of cytopathology specimens. *Cytopathology* **5**, 345–58.

Rotational Micro-Wave Spectra (2008) *Wikipedia Under "Rotational Spectroscopy",* copy-right free.

Sagripanti, J.L., Swicord, M.L., and Davis, C.C. (1987) *Radiat. Res.,* **110**, 219–31.

Shi S.R., Gu J., Kalra K.L., Chen T., Cote R.J., and Taylor C.R. (1995) *Antigen Ret. Tech., Cell Vision*, **2**, 7–22.

Townes, C.H. and Schawlow, A.L. (1975) *Microwave Spectroscopy*, Dower Publication.

White, J.H. (1969) Self-linking and the Gauss-integral in higher dimensions. *Am. J. Math.*, **91**, 693–728.

Wolynes, P.G. (1990) Spin glass ideas in the protein folding problem, in *Spin Glasses and Biology* (D.L. Stein), World Scientific Press, New York.

Wu, C.Y., Chen, S. T., Chiou, S.H., and Wang, K.T. (1992) *J. Protein Chem.*, **11**, 45–50.

17
Ion Channels[1]

Olaf S. Andersen, Helgi I. Ingólfsson, and Jens A. Lundbæk

17.1 Introduction 559
17.1.1 Ion Channels Are Not Just Water-Filled Pores 559
17.1.2 Ion Channels Are Efficient Catalysts of Transmembrane Ion Movement 560
17.1.3 Catalytic Rate Enhancement 561
17.2 Channels and Diseases 562
17.3 Diversity of Ion Channels 562
17.3.1 Classification of Ion Channels 564
17.3.1.1 Pore-Loop Channels 564
17.3.1.2 Cys-Loop Channels 569
17.3.1.3 Other Channel Types 569
17.4 Ion Permeation and Membrane Potentials 570
17.4.1 Membrane Potentials and Single-Channel Currents 570
17.4.2 Ion Selectivity 572
17.4.3 Membrane Potential Changes and Transmembrane Charge Movements 572
17.4.4 Bulk Ion Movement 574
17.5 Channel Gating 575
17.5.1 Gating Transitions 575
17.5.2 Channel Block 576
17.5.3 Ligand Activation 577
17.5.4 Voltage-Dependent Gating 578
17.5.4.1 Voltage- and Ligand-Activated Channels 579
17.5.5 Mechanosensitive Channels 579
17.5.6 Other Mechanisms 580
17.6 Regulation of Channel Gating 580
17.6.1 Ion Channels Are Special 580

---

[1] This chapter is reprinted with permission from the Wiley Encyclopedia of Chemical Biology, John Wiley & Sons, Inc., 2009. All rights reserved.

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

17.6.2 Hydrophobic Matching and Hydrophobic Coupling 580
17.7 Mechanisms of Drug-Induced Ion Channel Modulation 582
17.8 Conclusion 584
Acknowledgments 584
References 584
Further Reading 592

17.1
Introduction

Bilayer-spanning ion channels are ubiquitous constituents of prokaryotic and eukaryotic plasma (and organellar) membranes (Hille, 2001); channel-encoding genes are found even in viral genomes (Pinto, Holsinger and Lamb, 1992; Kang *et al.*, 2004). Ion channels serve many biological functions by being conduits for rapid, regulated ion movement across cellular membranes. Notably, they allow for the changes in membrane potential that underlie the function of the nervous system (Fain, 1999) and other excitable tissues.

Ion conducting channels catalyze transmembrane ion movement by forming continuous paths through which ions can cross the barrier imposed by the membranes' lipid bilayer moieties (Hille, 2001; Andersen and Koeppe, 1992; Chung, Andersen and Krishnamurthy, 2007). In biological membranes, integral membrane proteins provide the structural basis for ion channels by forming the walls of water-filled, high-dielectric pores that allow for electrodiffusive ion movement from one aqueous phase to the other. Ion channels thus provide an alternate permeation (or reaction) path for transmembrane ion movement, in the sense that the permeating ions do not cross the membrane by moving through the bilayer hydrophobic core *per se* (Figure 17.1).

17.1.1
Ion Channels Are Not Just Water-Filled Pores

Most channels are not just "simple" water-filled pores; they possess substrate specificity (ion selectivity) and so can discriminate between chemically closely related ion species such as K^+ and Na^+ (Hille, 2001). That is, ion channels are highly specific protein catalysts or enzymes (Andersen, 1989; Eisenberg, 1990). Ion channels constitute a special class of enzymes, however, because no covalent bonds are made or broken during the ions' passage through a channel pore.

Although channel pores are generally water-filled, some pores are so narrow that the water chain is interrupted by the permeating ions, as shown for the selectivity filter in potassium channels (Doyle *et al.*, 1998), in which the water molecules are separated by K^+. It also has been proposed that H^+ movement through voltage-dependent proton channels proceeds

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

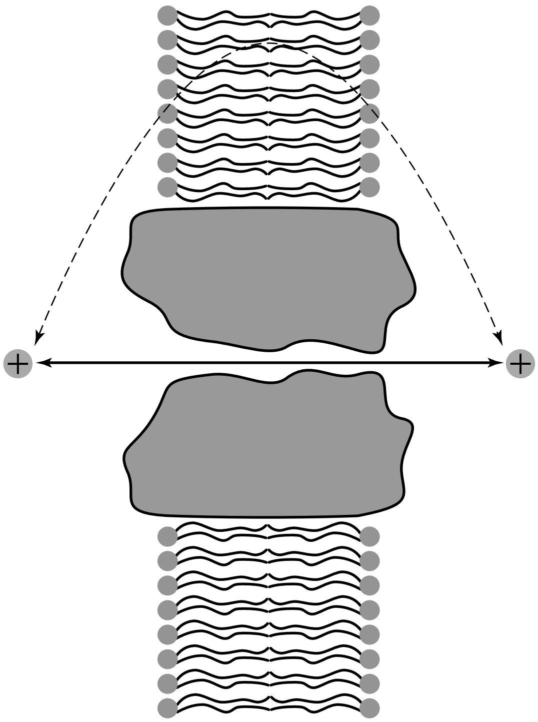

Fig. 17.1 Channel-mediated ion movement. The channel protein forms the walls of an aqueous pore through which ions, water, and other small solutes can move across the membrane (indicated by the solid line). Ions can also traverse the membrane by moving through the bilayer itself (indicated by the interrupted curve). Because the solubility of small inorganic ions in the membrane interior is very low, this uncatalyzed ion movement through the lipid bilayer core usually can be ignored

along a hydrogen-bonded network that may not contain a continuous chain of water molecules (DeCoursey and Cherny, 1998).

17.1.2
Ion Channels Are Efficient Catalysts of Transmembrane Ion Movement

Similar to other enzymes, the important determinants of ion channel function are the channels' substrate specificity (ion selectivity), their turnover number and catalytic power, and their regulation. Because no covalent bonds are formed or broken when ions pass through a channel, the energy barrier for channel-catalyzed ion movement may be quite low. A distinguishing feature of ion channels, therefore, is their turnover number, which can be $> 10^8\ \mathrm{s}^{-1}$ (Latorre and Miller, 1983) – as compared with $< 10^6\ \mathrm{s}^{-1}$ for most regular enzymes (see Hille (2001), Table 11.2A) – and a turnover number greater than $10^5\ \mathrm{s}^{-1}$ is commonly considered to be a defining feature of ion channels. Though there is no lower bound on the ion flux through membrane-spanning channels, the typically very high turnover allows for electrophysiological measurements of ion movement (the single-channel current) through individual channels (Figure 17.2).

Indeed, the first single-molecule measurements were single-channel recordings (Bean *et al.*, 1969; Hladky and Haydon, 1970), which provide the ability to resolve single-channel current transitions (single protein conformational transitions) at the microsecond timescale (Maconochie, Fletcher and Steinbach,

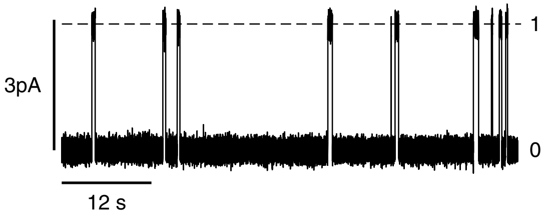

Fig. 17.2 Single-channel current trace recorded in planar lipid bilayers doped with gramicidin A (gA). The gA channels do not open and close but form and disappear. The numbers to the right in the figure indicate the number of conducting channels; the stippled line indicates the current level for the conducting channel. Experimental conditions as described in Greathouse *et al.* (1999) in 121.0 M NaCl, 200 mV applied potential, and dioleoylphosphatidylcholine/*n*-decane bilayer

1995; Shapovalov and Lester, 2004). The high temporal resolution provided by electrophysiological measurements makes ion channels important not only because of their intrinsic physiological functions but also because they can be used as tools to study general features of (membrane) protein function.

Ion movement through ion channels occurs as an electrodiffusive barrier crossing, in which the ion movement is largely uncoupled from protein conformational transitions. To appreciate ion channels' catalytic power, it is instructive to compare the measured conductances with predictions based on a simple water-filled pores with solvent properties similar to bulk water, while neglecting the diffusional access resistance to the channel entrance (Läuger, 1976; Andersen, 1983). In this case, the predicted single-channel conductance for monovalent ions, g_{pred}, becomes (Andersen, Koeppe and Roux, 2007, p. 51):

$$g_{\mathrm{pred}} \approx \lambda^{\circ} C \frac{\pi (r_p - r_i)^2}{l_p} \qquad (1)$$

where λ° is the limiting molar conductivity (a measure of the ion mobility in bulk aqueous solutions, units $(\mathrm{S \cdot cm^2})\ \mathrm{mol^{-1}}$), C the permeant ion concentration in the bulk aqueous phase, r_p the pore radius, r_i the ionic radius, and l_p the pore length. [The appropriate radius in Eq (1) is the difference between r_p and r_i because the ion centers are constrained to move within a narrow cylinder of radius $r_p - r_i$ (Ferry, 1936).] For $\mathrm{Na^+}$ permeation through the cation selective gA channels [$\lambda^{\circ}_{\mathrm{Na}} = 50.1\ (\mathrm{S \cdot cm^2})\ \mathrm{mol^{-1}}$ (Robinson and Stokes, 1959), $r_{\mathrm{Na}} \approx 0.95$ Å (Hille, 1975a), $r_p = 2$ Å, and $l_p = 25$ Å (Finkelstein and Andersen, 1981)], g_{pred} is predicted to be about ≈ 70 pS in 1.0 M NaCl and 7 pS in 0.1 M NaCl – similar to the experimental values of 12.5 pS (in 1.0 M NaCl) and 5.3 pS (in 0.1 M NaCl) (Finkelstein and Andersen, 1981). This similarity does not mean that ion channels are just aqueous pores. Ions may bind with high affinity into the pore of ion channels, such that the ion mole fraction in the pore may be several orders of magnitude higher than in the aqueous solution (see next section).

17.1.3
Catalytic Rate Enhancement

It is possible to extend this argument by estimating the catalytic rate enhancement provided by an ion channel (Andersen, Koeppe and Roux, 2007). By analogy with conventional enzymes (Wolfenden and Snider, 2001), the catalytic rate enhancement can be defined as the rate of channel-mediated ion movement (k_{cat}) relative to the rate of noncatalyzed movement through the bilayer (k_{non}). To a first approximation, $k_{\mathrm{cat}}/k_{\mathrm{non}}$ can be equated with $K_{w\to p}/K_{w\to m}$, where $K_{w\to p}$ and $K_{w\to m}$ denote the ion partition coefficients from bulk water into the pore and the bilayer hydrophobic core, respectively. $K_{w\to p}$ can be estimated from apparent dissociation constants for the permeant ions, which are in the micromolar range for calcium (Almers and McCleskey, 1984) and potassium (Neyton and Miller, 1988) channels. That is, $K_{w\to p}$ can be approximated as $n_i{}^p/n_i{}^b$, where $n_i{}^p$ and $n_i{}^b$ denote the ion mole-fractions in the pore and the bulk solution, which can be as high as 10^6. $K_{w\to m}$ can be estimated from the "leak" conductance ($G_0 \approx 10^{-9}$ S cm^{-2}) of unmodified bilayers in 1.0 M salt (Hanai, Haydon and Taylor, 1965) using the relation (Andersen, 1989; Hodgkin and Katz, 1949)

$$G_0 = N_A \frac{(ze)^2}{k_B T} \frac{D_m}{d_0} K_{w\to m} C \qquad (2)$$

where N_A is Avogadro's number, k_B Boltzmanns constant, T the temperature in Kelvin, z the ion's valence, e the elementary charge, and D_m the ion's diffusion coefficient in the bilayer hydrophobic core [about 10^{-5} cm$^2\cdot$s^{-1} (Schatzberg, 1965)]. On the basis of Eq. (2), $K_{w\to m}$ is predicted to be 10^{-14} (or less), such that the catalytic rate enhancement may be as high as 10^{20} – higher than the rate enhancements observed for conventional enzymes (Wolfenden and Snider, 2001).

17.2
Channels and Diseases

Because ion channels are such effective catalysts for transmembrane ion movement, disruptions in channel function can lead to disease, and an increasing number of human diseases have been shown to be caused by abnormal channel function caused by mutations in channel genes (Ashcroft, 2006; Cannon, 2006). These so-called channelopathies can develop from defective channel synthesis that is caused by missense mutations or splice defects (Cannon, 2006), inappropriate trafficking of the mutant channels to their target membrane [as is the case for the cystic fibrosis transmembrane conductance regulator or CFTR channel (Du, Sharma and Lukacs, 2005)], or channel retrieval [as is the case for the epithelial sodium channel (ENaC) in Liddle's syndrome (Rossier *et al.*, 2002)]. They also can develop from altered gating, which may increase or decrease the number of conducting channels (Bennett *et al.*, 1995); altered ion selectivity, which may disrupt the cellular electrolyte and volume balance (Kofuji *et al.*, 1996); or the appearance of new channel functions such as novel pathways for ion movement (Sokolov, Scheuer and Catterall, 2007), which again may compromise the cellular electrolyte and volume balance (Struyk and Cannon, 2007).

17.3
Diversity of Ion Channels

Ion channels can be classified based on their structure and function. Structurally, ion channels are usually hetero- or homo-oligomers of pore-forming subunits, sometimes in association with accessory/modulatory subunits. Channels in plasma membranes have predominantly α-helical bilayer-spanning transmembrane (or TM) domains, whereas channels in bacterial outer membranes (and the outer membrane of mitochondria) have a predominantly β-barrel structure (Figure 17.3). Although the focus is usually on the TM domains, for many

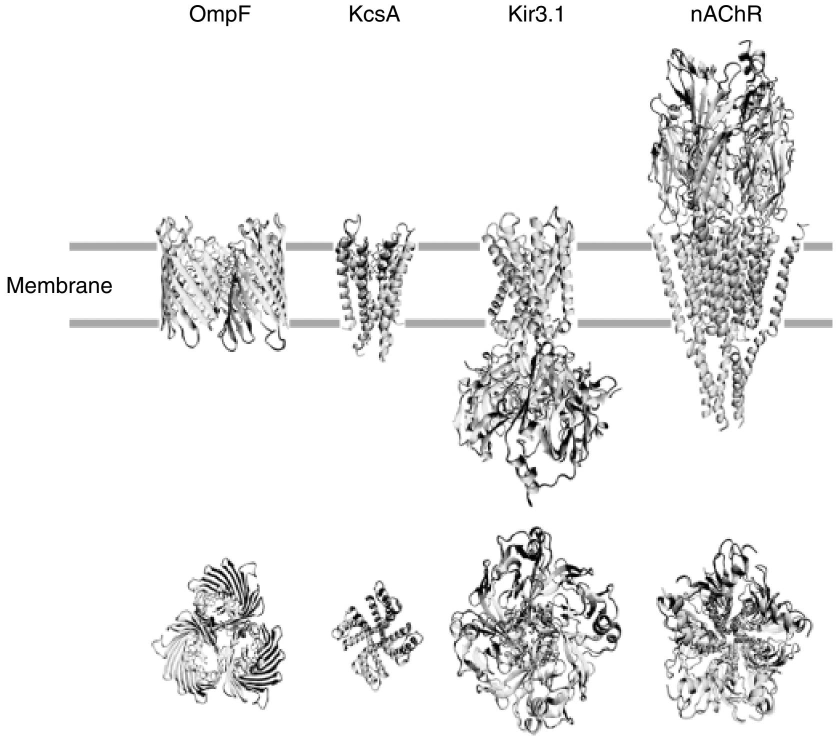

Fig. 17.3 Structural representation of different transmembrane ion channels (from left to right): the bacterial outer membrane protein OmpF [PDBid: 1OPF (Cowan *et al.*, 1995)]; the bacterial potassium channel KcsA [PDBid: 1 K4C (Zhou *et al.*, 2001)]; a chimera between the mammalian inward rectifier potassium channel Kir3.1 and the bacterial inward rectifier KirBAC1.3 [PDBid: 2QKS (Nishida *et al.*, 2007); in this structure a large part of the transmembrane pore domain of Kir3.1 has been replaced by the corresponding fragment of the pore domain of the prokaryotic KirBac1.3]; and the mammalian nicotinic acetylcholine-gated channel nAChR [PDBid: 2BG9 (Unwin, 2005)]. α-Helices are indicated as coiled ribbons, β-sheets as straight ribbons, and the loops as lines. Top row: side views of the channels relative to the outer membrane (OmpF, with the periplasmic surface down) or the plasma membrane (KcsA, Kir3.1, and nAChR, with the intracellular surface down). Bottom row: the channels viewed along the pore: OmpF is viewed from the periplasmic solution; KcsA and nAChR are viewed from the extracellular solution; and Kir3.1 is viewed from the intracellular solution. The biological unit of the OmpF consists of three functional channels, and each pore is formed by a single subunit. The KcsA and Kir3.1 channels are formed by four subunits that line a single pore; the nAChR channel is formed by five subunits that line a single pore

channels, majority of the molecular mass is outside the membrane, illustrated in Figure 17.3 for a Kir3.1 chimera and nicotinic acetylcholine receptor (nAChR).

Functionally, ion channels are distinguished by their permeability and gating properties and by their cellular localization. Most channels are valence selective, and many channels exhibit exquisite selectivity among ions of the same valence [as indicated by their names – calcium channels, chloride channels, potassium channels, and sodium channels, e.g., Hille (2001)]. The major exceptions are the large-diameter channels, such as the porins in the outer membranes of gram-negative bacteria and the connexins (or gap junctions) that couple adjacent cells together – both electrically and by mediating cell-to-cell movement of low-molecular weight signaling molecules (Harris, 2001). These channels allow for the relatively nonselective transmembrane movement of both anion and cations (Harris, 2007), yet such channels may exhibit selectivity among (larger) organic solutes (Harris, 2007; Kullman, Winterhalter and Bezrukov, 2002; Kanaporis *et al.*, 2008). Indeed, although channels with large-diameter pores allow for the passage of large solutes, they do not necessarily have high single-channel conductances; no clear relation is found between a channel's pore dimensions and its conductance (Finkelstein, 1975).

Ion channel function may be altered by changes in the surface density of channels in the membrane as well as by rapid transitions between nonconducting and conducting channel states – a process denoted as *channel gating*. The two classic examples of channel gating are the *voltage-dependent gating* (described in Section 17.5.4) of sodium and potassium channels that underlie the action potentials in nerve and muscle cells, in which the channel state is controlled by the transmembrane potential, and the *ligand-dependent gating* (described in Section 17.5.3) of the channels involved in synaptic (chemical) transmission of electrical signals from cell to cell, such as the nAChRs in skeletal muscle, in which the channel state is controlled by ligand (acetylcholine) binding (Hille, 2001). The principles underlying channel gating are discussed in Section 17.5.

17.3.1
Classification of Ion Channels

Tables 17.1 and 17.2 list ion channels according to their gene superfamilies, number of TM (α-helical) segments, and function. The emphasis is on plasma membrane channels because most is known about them.

17.3.1.1 Pore-Loop Channels

The largest family of plasma membrane ion channels is the family of pore-loop channels, listed in Table 17.1. The defining feature of these channels is their central core, which forms the ion permeable pore with the pore lining being formed by two TM α-helices plus the intervening sequence that loops into the membrane to form the selectivity filter – the pore region that determines which ions can permeate the channel. The archetypical example of this so-called 2TM,1P (P for pore) motif is found in the KcsA channel (Figure 17.3), the first ion-selective plasma membrane channel whose high-resolution structure was determined (Doyle *et al.*, 1998). The channel is a tetramer of four pore-forming 2TM,1P subunits, in which the

Tab. 17.1 Ion channel classes: plasma membrane pore-loop and cys-loop channels.

Channel class	Organization	Type	Properties	Reference
Pore-loop channels	2TM,1P (tetramers)	KcsA	K^+ selective (proton gated)	Doyle *et al.* (1998) and Chakrapani, Cordero-Morales and Perozo (2007a, 2007b)
		Kir	K^+ selective (inward rectifying)	Lu (2004) and Kubo *et al.* (2005)
	4TM,2P (dimers)	K_{2P}	K^+ selective (nonrectifying; stretch- and amphiphile gated)	Goldstein *et al.* (2005) and Honore (2007)
		K_V	K^+ selective (voltage dependent)	Gutman *et al.* (2005) and Korn *et al.* (2007)
		M channels	K^+ selective (voltage dependent; G protein inhibited)	Delmas and Brown (2005)
		hERG	K^+ selective (slow voltage dependent)	Sanguinetti and Tristani-Firouzi (2006)
	6TM,1P (tetramers)	K_{Ca} (SK)	K^+ selective (Ca^{2+})/calmodulin activated	Wei *et al.* (2005) and Bond, Maylie and Adelman (2005)
		CNG	Cation selective (cyclic nucleotide dependent)	Hofmann, Biel and Kaupp (2005) and Craven and Zagotta (2006)
		HCN	Cation selective (voltage- and cyclic nucleotide dependent)	Craven and Zagotta (2006)
		TRP	Cation selective (sensory, mechano-, and amphiphile activated)	Clapham *et al.* (2005) and Voets *et al.* (2007)
	7TM,1P (tetramers)	K_{Ca} (BK)	K^+ selective (Ca^{2+}- and voltage activated)	Wei *et al.* (2005) and Cox *et al.* (2007)

Continued

Tab. 17.1 *Continued*

Channel class	Organization	Type	Properties	Reference
	24TM,4P (monomers)	Na_V	Na^+ selective (voltage activated)	Catterall, Goldin and Waxman (2005) and Hanck *et al.* (2007)
		Ca_V	Ca^{2+} selective (voltage activated)	Catterall, Goldin and Waxman (2005) and Corry *et al.* (2007)
	3TM,1P (monomers)	GluR	Cation selective (glutamate activated)	Mayer (2006) and Oswald *et al.* (2007)
Cys-loop channels	4TM (pentamers)	AChR, $5HT_3R$	Cation selective (acetylcholine- or serotonin activated)	Unwin (2005) and Colquhoun (2006)
		$GABA_AR$, GlyR	Anion selective (γ-aminobutyric acid- or glycine activated)	Sine and Engel (2006) and Lynch *et al.* (2007)

Tab. 17.2 Ion channel classes: channels other than pore-loop and Cys-loop channels.

Localization	Organization	Type	Properties	Reference
Plasma membrane	4TM	CRAC	Ca^{2+} selective(Ca^{2+} store depletion activated)	Lewis (2007)
	4TM	H_v	H^+ selective (voltage dependent)	Decoursey (2003) and Ramsey *et al.* (2006)
		ENaC	Na^+ selective	Garty and Palmer (1997) and Snyder (2005)
	2TM (trimers)	Deg	Na^+ selective (mechanosensitive)	Goodman MBSchwarz (2003)
		ASIC	Na^+ selective (H^+ activated)	Wemmie, Price and Welsh (2006) and Jasti *et al.* (2007)

Tab. 17.2 *Continued*

Localization	Organization	Type	Properties	Reference
	2TM (trimers)	P_2X	Ca^{2+} (cation) selective (ATP dependent)	Egan, Samways and Li (2006) and Roberts *et al.* (2006)
	10TM (dimers)	ClC	Cl^- (anion) selective (some are Cl^-/H^+ antiporters) (Cl^-- and voltage dependent)	Miller (2006) and Pusch *et al.* (2007)
	12TM (monomers)	CFTR	Cl^- (anion) selective (pK_a dependent phosphorylation)	Gadsby, Vergani and Csanady (2006)
	4TM (hexamers)	Connexins	Selectivity among cytoplasmic signaling molecules (H^+ gated)	Harris (2001) and Harris (2007)
	4TM	Pannexins	Selectivity among cytoplasmic signaling molecules	Dahl and Locovei (2006)
		Innexins		Yen and Saier (2007)
	4TM	Bestrophins	Cl^--selective (some are Ca^{2+} activated)	Hartzell *et al.* (2008)
	2TM (pentamers)	MscL	Nonselective (mechanosensitive, amphiphile activated)	Sukharev and Corey (2004) and Martinac *et al.* (2007)
	3TM (heptamers)	MscS	Nonselective (mechanosensitive)	Martinac *et al.* (2007)
Organellar membrane	6TM (tetramers)	RyR	Ca^{2+} (cation) selective (Ca^{2+} activated)	Zalk, Lehnart and Marks (2007)
		IP_3R	Ca^{2+} (cation) selective (IP_3- and Ca^{2+} activated)	Bosanac *et al.* (2004)
Outer membrane	β-barrels	Porins VDAC	Nonselective (except for some nutrients)	Delcour (2002) and Rostovtseva, Tan and Colombini (2005)
Viral channels	1TM (tetramers)	M2	H^+ selective	Pinto and Lamb (2006)

selectivity filter is lined by four extended peptide strands (one from each subunit) in a loop between the TM helices. This bilayer-spanning motif is found not only in KcsA but also in the Kir family of inward rectifier potassium channels (Kubo *et al.*, 2005); the latter class of channels also has a large cytoplasmic domain (cf. Figure 17.3).

The basic 2TM,1P motif has evolved in three different directions (Figure 17.4). The first direction evolved by a process of internal gene duplication that led to the so-called 2-pore (K_{2P}) channels with a 4TM,2P organization. The second direction evolved by a more complex process that involved an inversion of the TM topology, the accretion of an additional TM segment, plus an extracellular domain with sequence similarity to bacterial periplasmic amino acid binding proteins (Wo and Oswald, 1995), which led to the glutamate receptor family of ligand-activated channels with a 3TM,1P organization. And the third direction evolved by acquiring a 4TM voltage-sensing domain [that by itself can form a voltage-dependent proton channel (H_V) (Musset *et al.*, 2008)] leading to the 6TM,1P group of channels, which are largely potassium selective and include the voltage-dependent potassium channels (K_V). This 6TM,1P motif has evolved into the 7TM,1P family of high-conductance Ca^{2+}-activated potassium channels (K_{Ca} or BK, for Big K^+ channel), which also have large C-terminal Ca^{2+} binding domains. Finally, the 6TM,1P motif also evolved by internal gene duplication to form the so-called two-pore channels that have two domains and the voltage-dependent sodium (Na_V) and calcium (Ca_V) channels that have four domains. In both cases, each domain corresponds to a duplicated 6TM,1P pore-loop motif. [The emphasis in Tables 17.1 and 17.2 and Figure 17.4 is on the pore-forming subunits, called *α subunits*; functional channels are usually coassemblies of four pore-forming subunits or

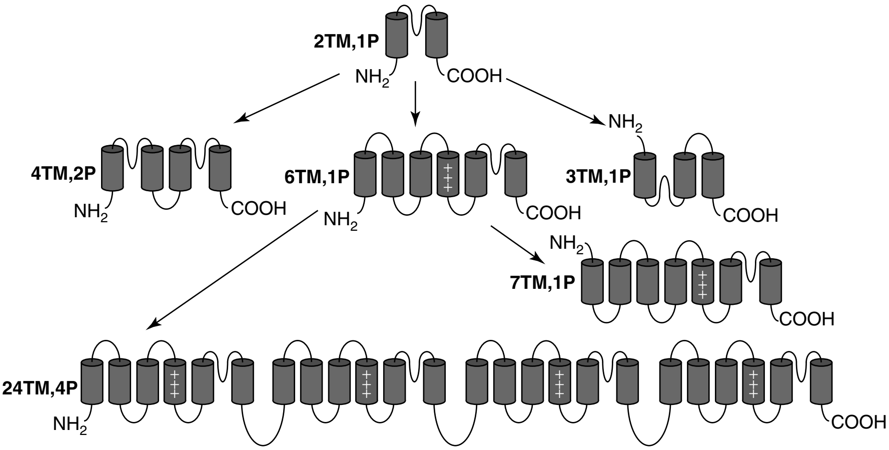

Fig. 17.4 Schematic diagram that shows the TM segment organization for different pore-loop plasma membrane ion channels (listed in Table 17.1). The ancestral 2TM,1P motif (exemplified by KcsA; Figure 17.3) has evolved by gene duplication and the accretion of additional TM domains – and an inversion of transmembrane topology in the case of the 3TM,1P glutamate receptors

domains plus additional cytoplasmic or TM regulatory subunits (Hille, 2001).]

17.3.1.2 Cys-Loop Channels

Another major family of plasma membrane ion channels is the family of ligand-gated Cys-loop receptors, also listed in Table 17.1. These channels can be subdivided based on their activating ligand: acetylcholine (ACh) for the nAChR (Figure 17.3); serotonin (5-HT) for the 5-HT_3R; glycine (Gly) for the GlyR; and γ-aminobutyric acid (GABA) for the $GABA_AR$. ACh, 5-HT, and GABA also bind to other receptors, for example, muscarinic acetylcholine receptor (AChR) and $GABA_BR$, which are G-protein-coupled receptors. These channels also can be categorized based on their ion selectivity and function: nAChR and 5-HT_3R are cation selective channels involved in excitatory synaptic transmission and $GABA_AR$ and GlyR are anion-selective channels involved in inhibitory synaptic transmission. [In addition to the listed channels, a H^+-gated Cys-loop channel has been reported (Beg *et al.*, 2008).]

17.3.1.3 Other Channel Types

The remaining channel types, listed in Table 17.2, are more difficult to categorize, and many of them have no obvious evolutionary relationship.

The major exception is the family of ENaCs, the mechanosensitive channels (Deg, for degeneration because Deg activation triggers cell death) first identified in *Caenorhabditis elegans*, and the H^+-gated sodium channels (ASIC, acid sensitive ion channel) found in various cell types. These channels are also related to a family of invertebrate peptide-gated channels [the FMRFamide-gated channels (Lingueglia, Deval and Lazdunski, 2006)].

There may also be an evolutionary relationship between the intracellular Ca^{2+} release channels, the ryanodine receptors (RyRs) that mediate the release of Ca^{2+} from the sarcoplasmic reticulum, and the IP_3 receptors (IP_3Rs) that mediate the release of Ca^{2+} from many intracellular organelles.

In the family of ClC proteins, some members are expressed in the plasma membrane where they form anion-selective channels. These channels occur as dimers, with each subunit forming a TM pore. Other members of this family are expressed in organellar membranes, where many of them form H^+/Cl^- countertransporters (Jentsch, 2008), as first demonstrated for a bacterial plasma membrane ClC protein (Accardi and Miller, 2004) – thereby validating a conjecture by Läuger (1980) that there would be a continuum of molecular mechanisms catalyzing TM solute movement, ranging from conformational transporters to TM channels.

The porins in the outer bacterial and mitochondrial membranes serve as conduits to the periplasmic space. Some porins occur as oligomers, each forming a TM pore (Figure 17.3). Many porins and the voltage-dependent anion-selective channel (VDAC) appear to catalyze the relatively nonspecific transfer of solutes below a certain size; other porins are involved in more specific solute transfer (Duy, Soll and Philippar, 2007). The H^+ permeable M2 channel, which is encoded by the influenza virus, is important for viral replication and is the target for the antiviral drug amantadine (Pinto and Lamb, 2006).

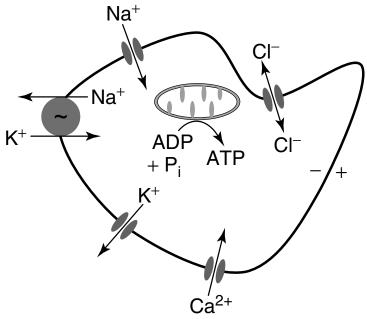

Fig. 17.5 Schematic representation of a cell, with channels that are selective for Cl$^-$, Ca^{2+}, K$^+$, and Na$^+$ plus the Na$^+$-, K$^+$-ATPase that maintains the ion distribution between the extracellular and intracellular compartments via an ATP-driven extrusion of Na$^+$ and uptake of K$^+$. The ATP is generated in the mitochondria by oxidative phosphorylation from ADP and inorganic phosphate (P$_i$). In resting cells, when no net charge movement is across the plasma membrane, the intracellular compartment is electrically negative relative to the extracellular compartment

17.4
Ion Permeation and Membrane Potentials

Most cell membranes are endowed with different types of ion channels (Figure 17.5) that differ in their ion selectivity and gating properties (cf. Tables 17.1 and 17.2).

Because the extracellular and intracellular ion concentrations differ (Table 17.3), the membrane potential ($V_m = V_i - V_e$, where V_i and V_e denote the electrical potential of the intracellular and extracellular compartment, respectively) will vary as a function of the number and type of conducting channels in the membrane (Section 17.4.3).

Although channel-catalyzed ion movement occurs as an electrodiffusive barrier crossing (Andersen, 1989; Cooper, Jakobsson and Wolynes, 1985; Levitt, 1986; Gates, Cooper and Eisenberg, 1990; Partenskii and Jordan, 1992; Kuyucak, Andersen and Chung, 2001; Roux *et al.*, 2004), the permeating ions are usually not distributed uniformly along the pore; they tend to be localized in discrete regions, or energy minima (e.g. see Doyle *et al.* (1998)), such that the ion movement can be approximated as a series of discrete transitions: entry, translocation through the pore, and exit (Figure 17.6).

Although the kinetics of channel-mediated ion movement may be described using discrete-state kinetics, the individual transitions depicted in Figure 17.6 extend over distances that are much longer than the ions' mean free path, which means that one cannot use Eyring's transition state theory with a prefactor of k_BT/h, where his Planck's constant, to relate the measured/deduced rate constants to an underlying energy profile (Andersen, 1999).

17.4.1
Membrane Potentials and Single-Channel Currents

The rate of ion movement through a channel, the single-channel current (i), varies as a function of V_m and the channel's reversal potential, V_{rev}, defined as the membrane potential where $i = 0$:

$$i = g(V_m - V_{\text{rev}}) \tag{3a}$$

where g is the single-channel conductance, which varies as a function of channel type, the permeant ion concentration(s), and V_m – with

Tab. 17.3 Extracellular and intracellular ion concentrations and equilibrium potentials.

Ion	Extracellular concentration (mM)	Intracellular concentration (mM)	Equilibrium potential (mV)
Na^+	145	~12	+67
K^+	4.5	~150	−94
Ca^{2+}	~1.5	~0.0001	+129
H^+	0.00004	~0.0001	−24
Cl^-	115	~10	−65
$HCO_3{}^-$	25	~10	−24

Ion concentrations in millimoles per liter of water. Values for mammals are modified from Ruch, Patton and Scher (1974), Tables 26-5 and 26-6. The $[Ca^{2+}]$ are the free ion concentrations; intracellular $[Cl^-]$ varies considerably among cell types, ranging from ~ 4 mM in skeletal muscle to ~ 80 mM in red blood cells. The equilibrium potentials were calculated for $T = 37\,^\circ C$ using the listed concentrations.

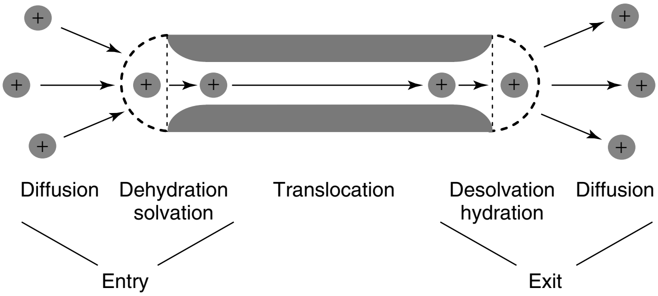

Fig. 17.6 Schematic representation of ion movement through a bilayer-spanning channel. Ions diffuse into the channel entrance, where they undergo dehydration and solvation by polar groups that line the pore wall, translocate through the pore in a diffusive barrier crossing for them to desolvate and rehydrate, and finally diffuse into the bulk solution

most channels having conductances that range between 5 and 50 pS (see Hille (2001), Figure 12.8). The total membrane current (I) that is carried across by all channels of a given type will be

$$\begin{aligned} I &= Ni = Ng(V_m - V_{\text{rev}}) \\ &= G(V_m - V_{\text{rev}}) \end{aligned} \tag{3b}$$

where N denotes the number of *conducting* channels in the membrane and G denotes the total membrane conductance contributed by the channels in question. In the case of highly selective channels that catalyze the TM movement of only a single ion type, V_{rev} becomes equal to the ion's equilibrium (or Nernst) potential E:

$$E = \left(\frac{-k_B T}{ze}\right) \ln\left\{\frac{C_i}{C_e}\right\} \tag{4}$$

where C_i and C_e denote the intracellular and extracellular ion concentrations, respectively. Table 17.3 lists values of E for the four major ions involved in cellular electrical activity. (Strictly, E

should be expressed in terms of the ion activities, but activity coefficients are neglected in Eq. (4) because the ionic strengths of the extracellular and intracellular solutions are similar.)

If only a single type of highly selective ion channels is found in the membrane, then the resting membrane potential – the time-invariant potential of a cell "at rest" when the net current across the membrane is 0 – will be equal to the equilibrium potential for the ion in question. If the membrane is endowed with different types of highly selective ion channels, then V_m becomes a weighted average of the equilibrium potentials for the different ions, for example:

$$V_m = \left(\frac{G_{\text{Na}} E_{\text{Na}} + G_{\text{K}} E_{\text{K}} + G_{\text{Ca}} E_{\text{Ca}} + G_{\text{Cl}} E_{\text{Cl}}}{G_{\text{Na}} + G_{\text{K}} + G_{\text{Ca}} + G_{\text{Cl}}} \right) \quad (5)$$

where the subscripts denote different ions. In the resting cell, G_{K} usually is much larger than the membrane conductances for other ions, so V_m will be close to E_{K}. In general, V_m and the cell functions that are regulated by V_m will vary as a function of the relative contributions of the different ions to the membrane conductance, $G_m = G_{\text{Na}} + G_{\text{K}} + G_{\text{Ca}} + G_{\text{Cl}} + \ldots$

17.4.2
Ion Selectivity

Some ion channels, notably potassium and calcium channels, have remarkable ion selectivity, to the extent that one in many cases can neglect the movement of other ions. Most ion channels, however, are measurably permeable to several different ions, in which case their V_{rev} will vary as function of the permeant ion concentrations and the ions' relative permeabilities, as expressed by the so-called Goldman–Hodgkin–Katz equation (Eq. (1)), which for monovalent cations becomes

$$V_{\text{rev}} = \frac{-k_B T}{e} \ln \left\{ \frac{P_{\text{Na}}[\text{Na}^+]_i + P_{\text{K}}[\text{K}^+]_i}{P_{\text{Na}}[\text{Na}^+]_e + P_{\text{K}}[\text{K}^+]_e} \right\} \quad (6)$$

where P_{Na} and P_{K} denote the permeability coefficients for Na^+ and K^+, respectively, and the ratio $P_{\text{Na}}/P_{\text{K}}$ is a measure of the channel's ability to discriminate between Na^+ and K^+. [Eq. (6) is widely used to quantify ion channels' selectivity, but it is an approximation that is valid only in special cases (Hille, 1975b; Hille and Schwarz, 1978); the permeability ratios, e.g. $P_{\text{Na}}/P_{\text{K}}$, therefore should be regarded as empirical descriptors of a channel's ion selectivity.]

Table 17.4 lists representative values for the permeability coefficients of different channels types (both selective and nonselective cation permeable channels as well as anion permeable channels) and V_{rev} calculated using the ion concentrations in Table 17.3 and Eq. (6). (See Hille (2001), Chapter 14, for a more extensive compilation and discussion.)

Comparing the ion equilibrium potentials in Table 17.3 with the channel reversal potentials in Table 17.4 shows that a channel's reversal potential may differ substantially from the name-giving ion's equilibrium potential.

17.4.3
Membrane Potential Changes and Transmembrane Charge Movements

Ion channels provide the structural underpinnings for the ion movements that underlie the rapid changes in membrane potential that are involved in the electrical signaling (the action

Tab. 17.4 Ion selectivities and reversal potentials for selected ion channels.[a]

Channel type	Permeability ratio	V_{rev} (mV)
Voltage-dependent sodium channel (Na_v)	$P_K/P_{Na} < 0.1$	$> +45$
Epithelial sodium channel (ENaC)	$P_K/P_{Na} \approx 0.001$	$+66$
Voltage-dependent potassium channels (K_v)	$P_{Na}/P_K < 0.01$	< -86
Voltage-dependent calcium channels (Ca_v)	$P_{Na}/P_{Ca} \approx 0.0008$ $P_K/P_{Ca} \approx 0.0003$	$+63$
Cyclic nucleotide-activated channels (CNG)	$P_K/P_{Na} \approx 1.1$ $P_{Ca}/P_{Na} \approx 90$	$+17$
Nicotinic acetylcholine receptor (nAChR)	$P_K/P_{Na} \approx 1.1$ $P_{Ca}/P_{Na} \approx 0.2$	-4
$GABA_AR$	$P_{HCO_3}/P_{Cl} \approx 0.4$	-58

[a] Modified from Hille (2001) (Tables 14.2– 14.6), except for $GABA_AR$, where P_{HCO3}/P_{Cl} is, from Fatima-Shad and Barry (1993).

V_{rev} was calculated using the ion concentrations in Table 17.3 and the permeability ratios in Table 17.4 and Eq. (6) ($T = 37\,°C$) or, for channels permeable to both mono- and divalent cations, a generalized version of Eq. (6) (Lewis, 1979).

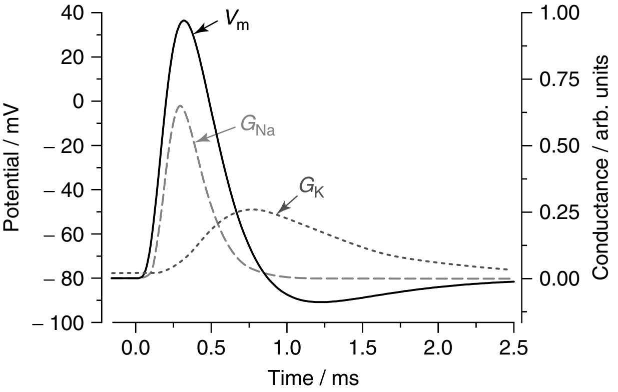

Fig. 17.7 Schematic time course of an action potential. The initial rising, or depolarizing, phase of the action potential is because of an increase in the conductance to Na^+ (G_{Na}). At the peak, the membrane potential approaches the Na^+ equilibrium potential. The later falling, or repolarizing, phase is due to an increase in the conductance to K^+(G_K) and a decrease in G_{Na}. Toward the end of the action potential, the membrane potential approaches E_K because G_K is much higher than at rest

potentials) in nerve and muscle cells (Fain, 1999) and in many other cell functions. Figure 17.7 shows the time course of the potential change during an action potential together with the time courses of the underlying changes in the membrane conductances to Na^+ and K^+. The initial depolarization is due to the rapid activation (opening) of Na_V channels; the later repolarization is because of the slower activation of K_V channels and deactivation and inactivation of Na_V channels, which eventually lead to a transient hyperpolarization (because G_K becomes so large that V_m approaches E_K).

Because cell membranes have a finite capacitance $C_m[C_m = AC_{sp}$, where C_{sp} is the specific membrane capacitance, $\sim 0.9\,\mu F\ cm^{-2}$ (Niles, Levis and Cohen, 1988; Gentet, Stuart and Clements, 2000) and A the membrane area], these changes in membrane potential require only a small TM charge movement ΔQ:

$$\Delta Q = C_m V_m \tag{7}$$

The net ion movements that underlie electrical signaling are minute: according to Eq. (7), a 100 mV change in V_m is associated with a net TM ion movement of $\sim 6 \times\ 10^3$ ions/μm^2. For a spherical cell of radius $10\,\mu$ m, the resulting change in the (volume-averaged) intracellular ion concentration would be $\sim 3\,\mu M$; for a nerve fiber of radius $0.5\,\mu$ m, the intracellular concentration change would be $\sim 40\,\mu M$. The actual ion movements are about threefold higher than the minimum value estimated using Eq. (7) (Hodgkin (1964), pp. 45–46); even then, the changes in V_m that underlie the action potential usually will not cause chemically measurable changes in the intracellular or extracellular ion concentrations (as can be seen by comparing these concentration changes with the concentrations listed in Table 17.3). The major exceptions to this rule are ions that are present at very low concentrations, such as cytoplasmic Ca^{2+} and H^+ [where the concentration changes, however, will be attenuated by cytoplasmic Ca^{2+} and H^+ buffers (Roos and Boron, 1981; Neher, 1995)]; cell processes with a very high surface-area-to-volume ratio (Helmchen, Borst and Sakmann, 1997); and tissues with small extracellular spaces, where K^+ and H^+ can accumulate (Beg *et al.*, 2008; Frankenhaeuser and Hodgkin, 1956; Filosa *et al.*, 2006).

Electrical signaling results from controlled changes in the cell membrane's permeability (conductance) to different ions, which result in changes in V_m (cf. Eq. (5)). The steady-state V_m is determined only by the relative conductances for the different ions (Eq. (5)), but the time course of changes in V_m (such as during an action potential; Figure 17.7) depends on the absolute membrane conductances (channel densities in the membrane). Approximating the change in V_m as an exponential transient with a time constant ($\tau = C_m/G_m$) of $10\,\mu$ s, the membrane conductance will need to be $\sim 10^{-9}$ S μm^{-2} – corresponding to a channel density of $\sim 100\,\mu m^{-2}$ (for channels with $g = 10$ pS) – meaning that a 100 mV change in V_m is associated with a net movement of only ~ 200 ions/channel. Although ion channels have very high turnover numbers, the actual number of ions that move through a channel during an action potential is quite small.

17.4.4
Bulk Ion Movement

In addition to the importance of ion channels for cellular electrical activity, where one to a first approximation can disregard the bulk movement of ions, ion channels serve two additional functions. First, as noted above, in the case of Ca^{2+}, the channel-mediated Ca^{2+} movement may cause a measurable increase in the cytoplasmic $[Ca^{2+}]$, which in turn regulates many different cell functions, including neurotransmitter release, enzyme and hormone secretion, muscle contraction, and gene activation (Berridge, Bootman and Roderick, 2003). Secondly, the bulk absorption of solutes across epithelial membranes depends not only on the presence of adenosine triphosphate (ATP)-dependent ion pumps, in particular the Na^+, K^+-ATPase, but also on the ion channels embedded in the apical and basolateral membranes (Palmer *et al.*, 2007).

17.5
Channel Gating

Ion channel function is tightly regulated to maintain normal function, such as the generation of action potentials, and to avoid compromising the cellular electrolyte, volume, and overall metabolic balance. Channel function is regulated by different stimuli, for example, membrane potential changes, usually positive-going (also called *depolarizing*) potential changes; changes in the concentration of extracellular (neuro)transmitters or intracellular messengers, which may be inorganic ions such as H^+, Ca^{2+}, and Zn^{2+}; light adsorption; covalent modification, such as phosphorylation or dephosphorylation; a mechanical perturbation; and so forth.

17.5.1
Gating Transitions

Channel gating usually involves transitions among different channel states (Figure 17.8): closed, open, and inactivated (in the case of voltage-dependent channels) or desensitized (in the case of ligand-activated channels).

Channel inactivation and desensitization describe transitions from conducting to nonconducting channel states at a sustained membrane depolarization or elevated ligand concentration (Hille, 2001). The inactivated or desensitized states are unresponsive to further depolarization or increases in the activating ligand concentration. The physiological effect of inactivation and desensitization is to reduce ion movement during excessive stimulation.

Whatever the gating mechanism or channel type, the total number of active, that is, conducting or open, channels (N) in a membrane will be equal to the total number of channels of that type in the membrane (N_{tot}) times the probability the channel is in the open state (W_O):

$$N = N_{\text{tot}} W_O \tag{8}$$

where W_O will vary as a function of the stimulus strength and time (after a given stimulus is applied). (Because the letter P is used to denote permeability, we use the letter W, for the German "Warscheinlichkeit," to denote probabilities.)

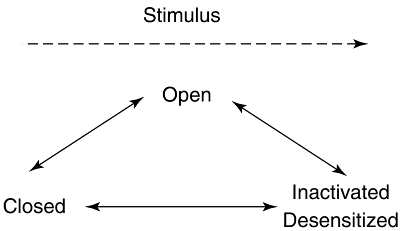

Fig. 17.8 The major transitions in channel gating. An activating stimulus leads to channel opening; the stimulus also causes the channel to enter one or more nonconducting states, which differ from the resting closed states and are denoted inactivated states (in the case of voltage-dependent channels) or desensitized states (in the case of ligand-activated channels)

Channel gating involves transitions among several closed (resting nonconducting) states, one or more open (conducting) states, and, maybe, one or more nonconducting inactivated or desensitized states (Figure 17.8). The principles underlying channel gating can be understood by considering just transitions between two states, closed (C) and open (O), where the stimulus shifts the distribution between the C and O states (Figure 17.9).

17.5.2
Channel Block

In the simplest case (Figure 17.9a), channel gating involves a (voltage-dependent)

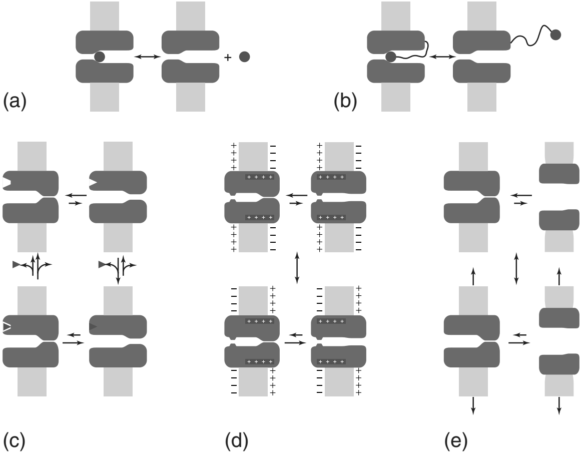

Fig. 17.9 Ion channel gating mechanisms. In each diagram, the extracellular solution is to the left and the intracellular solution is to the right of the channel and its host bilayer. The horizontal arrows in panels a and b denote the reversible transition between the two states; the vertical arrows in the middle of panels d and e denote transitions between the upper and lower equilibria. (a) Block by a small molecule that can enter the pore, where it binds to occlude the lumen and thereby block ion movement. (b) Block by a covalently attached channel domain that is able to enter the pore to occlude the lumen and block ion movement. (c) Ligand activation by a (neuro)transmitter, where the channel exists in two states, nonconducting and conducting, and the equilibrium distribution between the two states is shifted when the activating ligand binds. (d) Voltage activation, changes in the membrane potential (cf. the upper and lower panels) shifts the distribution between different positions of a voltage sensor (indicated by the bar with positive charges); the channel is in the conducting state when the voltage sensor is in the outward position. (e) Mechanical activation, where an increase in membrane tension (indicated by heavy arrows in the middle of each membrane) shifts the distribution between a closed state and an open state that has a greater cross-sectional area

block of the ion permeable pore (Woodhull, 1973). The blocking molecule (B) binds at a site in the pore lumen, and, because electrical potential varies along the pore, B's dissociation constant (K_B) will vary as a function of V_m. In the case of an impermeable cytoplasmic blocker that cannot pass through the pore, the blocker's reaction with the channel can be described as

$$\mathrm{O} + \mathrm{B} \rightleftharpoons \mathrm{OB};\ k_B = \frac{[O]\cdot[B]}{[\mathrm{OB}]};$$
$$k_B(V_m) = k_B(0)\exp\left\{-\frac{ze(1-\delta)V_m}{k_BT}\right\} \tag{9}$$

where O and OB denote the conducting (open) and nonconducting (obstructed) state, z the blocker's valence, and δ the fraction of V_m that falls between the extracellular solution and the site where the blocker binds. The probability that the channel is conducting thus becomes

$$W_O = \frac{k_B(0)\exp\left\{-\dfrac{ze(1-\delta)V_m}{k_BT}\right\}}{k_B(0)\exp\left\{-\dfrac{ze(1-\delta)V_m}{k_BT}\right\}} + [\mathrm{B}] \tag{10}$$

This mechanism describes many aspects of the rectification in inward rectifying potassium channels (Kir; cf. Table 17.1 and Figure 17.3), where cytoplasmic amines such as spermine and spermidine can enter the pore to block ion movement (Lu, 2004). It is also involved in the local anesthetic-induced block of Na_V by (positively) charged anesthetics, although local anesthetics also alter other channel properties (Yanagidate and Strichartz, 2007).

In a more elaborate version of this gating mechanism (Figure 17.9b), the blocking particle is covalently attached to the channel, but still can enter the pore to block ion movement. This mechanism is observed in the fast inactivation of voltage-dependent channels (Yellen, 2002).

17.5.3
Ligand Activation

In the case of ligand-activated channels (Figure 17.9c), such as the Cys-loop receptors (cf. Table 17.1 and Section 17.3.3), an agonist (A) binds with different affinities to the C and O states, and the difference in ligand binding energy to the two states shifts the O/C distribution (Karlin, 1967; Colquhoun, 1973; Jackson, 1989):

$$\frac{L^{CA\to OA}}{L^{C\to O}} = \frac{K_A^C}{K_A^O} = \exp\left\{-\frac{\Delta G_A^O - \Delta G_A^C}{k_BT}\right\} \tag{11}$$

where $L^{C\to O}$ and $L^{CA\to OA}$ denote the equilibrium constant for the interconversion between C and O and between their agonist-bound states CA and OA, respectively; K_A^C and K_A^O denote the dissociation constants and ΔG_A^C and ΔG_A^O denote the standard free energies, respectively, for agonist binding to C and O. The dose-response curve for channel activation by the agonist thus becomes

$$W_O([\mathrm{A}]) = \left(\frac{W_O(0)EC_{50} + W_O(\infty)[\mathrm{A}]}{EC_{50} + [\mathrm{A}]}\right) \tag{12}$$

where $W_O(0)$ and $W_O(\infty)$ denote the channel open probabilities in the limits $[\mathrm{A}] \to 0$ and $[\mathrm{A}] \to \infty$, respectively:

$$W_O(0) = \left(\frac{L^{C\to O}}{L^{C\to O}+1}\right);$$
$$W_O(\infty) = \left(\frac{L^{CA\to OA}}{L^{CA\to OA}+1}\right) \tag{13}$$

and

$$EC_{50} = \frac{L^{C\to O}+1}{L^{CA\to OA}+1}K_A^C = \frac{L^{C\to O}+1}{L^{C\to O}K_A^C + K_A^O}K_A^C K_A^O \qquad (14)$$

is the agonist concentration at which

$$W_O(EC_{50}) = \frac{W_O(0) + W_O(\infty)}{2} \qquad (15)$$

which differs from both K_A^C and K_A^O [cf. Colquhoun (1998)].

The gating of ligand-activated channels is usually more complicated than summarized here, see, for example, Mayer (2006) and Hansen, Yuan and Traynelis (2007). First, channel activation usually involves binding of two (or more) agonist molecules; second, sustained increases in the agonist concentration promote an additional gating transition – to a nonconducting, desensitized state (cf. Figure 17.8) – which is associated with additional conformational transitions (Mayer, 2006; Hansen, Yuan and Traynelis, 2007; Unwin, 1995).

17.5.4
Voltage-Dependent Gating

In the case of voltage-activated channels (Figure 17.9d), such as the 6TM,1P K_V channels or the 24TM,4P Na_V or Ca_V channels (cf. Table 17.1, Section 17.3.2, and Figure 17.4), changes in V_m are sensed by the translocation of charges or reorientation of dipoles by the electric field (Armstrong and Bezanilla, 1973; Bezanilla, 2008). In the case of the voltage-dependent K_V, Na_V, and Ca_V channels, the charge translocation develops from a reorientation of the positively charged S4 segment (Figure 17.4), which leads to an effective TM transfer of about three charges per subunit or domain (Schoppa *et al.*, 1992; Hirschberg *et al.*, 1995). This charge movement is coupled to conformational changes that cause the channel to open when the membrane is depolarized (cf. Hodgkin and Huxley (1952) and Zagotta, Hoshi and Aldrich (1994)). Atomic resolution structures are available for several K_V channels (Jiang *et al.*, 2003; Long, Campbell and Mackinnon, 2005; Long *et al.*, 2007), but the mechanism underlying the charge movement has not been fully clarified (Bezanilla, 2008).

Although the channels are activated (opened) by depolarizing the membrane, the C ↔ O equilibrium usually is strongly shifted toward O at $V_m = 0$ mV (Hodgkin and Huxley, 1952); that is, the depolarization removes a "stimulus" that normally keeps the channels closed.

The voltage-dependent shift in the C ↔ O equilibrium is given by

$$\frac{L^{C\to O}(V_m)}{L^{C\to O}(0)} = \exp\left\{\frac{z_a e V_m}{k_B T}\right\} \qquad (16)$$

where $L^{C\to O}(V_m)$ denotes the voltage-dependent C ↔ O equilibrium constant, and z_a denotes the gating valence (the charge movement associated with the C → O transition; z_a is positive for an outward movement of positive charge). The probability of the channel being open at a given voltage, V_m, thus becomes

$$W_O(V_m) = \frac{L^{C\to O}(0)}{L^{C\to O}(0) + \exp\{-z_a e V_m/k_B T\}} = \frac{1}{1+\exp\{-z_a e(V_m - V_{1/2})/k_B T\}} \qquad (17)$$

where $V_{1/2}$ is the potential at which $W_0(V_{1/2}) = 0.5$ (the midpoint potential for channel activation):

$$V_{1/2} = \frac{k_B T}{z_a e} \ln\{L^{C\to O}(0)\} \tag{18}$$

As is the case for the ligand-activated channels, the gating of voltage-activated channels is more complicated than summarized here: first, there are usually multiple closed states in the activation (Hodgkin and Huxley, 1952; Zagotta, Hoshi and Aldrich, 1994; Zheng and Sigworth, 1998); secondly, sustained depolarizations tend to promote transitions to one or more nonconducting, inactivated states (Figure 17.8) (cf. Yellen (2002)).

17.5.4.1 Voltage- and Ligand-Activated Channels

The distinction between voltage- and ligand activation is not always clear-cut. The function of many (if not all) voltage-dependent channels can be modulated by small molecules (including drugs and toxins) (Catterall *et al.*, 2007, see also Section 17.6) and the gating of ligand-activated channels can be modulated by changes in V_m (Magleby and Stevens, 1972). The most striking example of this dual regulation is the BK Ca^{2+}-activated potassium channels, which are both voltage- and ligand activated (Cox *et al.*, 2007; Magleby, 2001; Horrigan and Aldrich, 2002).

17.5.5
Mechanosensitive Channels

Mechanosensitive channels are activated by various mechanical stimuli (Sukharev and Corey, 2004; Markin and Hudspeth, 1995; Hamill and Martinac, 2001; Markin and Sachs, 2007), including changes in membrane tension (σ), which will shift the C ↔ O equilibrium when the cross-sectional channel area differ for the C and O states:

$$\frac{L^{C\to O}(\sigma)}{L^{C\to O}(0)} = \exp\left\{\frac{\Delta A\sigma}{k_B T}\right\} \tag{19}$$

where $L^{C\to O}(\sigma)$ denotes the tension-dependent, C ↔ O equilibrium constant, and $\Delta A = A_O - A_C$ where A_O and A_C denote the cross-sectional area of the open and closed state, respectively. The probability of the channel being open at a given tension, σ, thus becomes

$$\begin{aligned} W_O(\sigma) &= \frac{L^{C\to O}(0)}{L^{C\to O}(0) + \exp\{-\Delta A\sigma/k_B T\}} \\ &= \frac{1}{1 + \exp\{-\Delta A(\sigma - \sigma_{1/2})/k_B T\}} \end{aligned} \tag{20}$$

where $\sigma_{1/2}$ is the membrane tension at which $W_O(\sigma_{1/2}) = 0.5$:

$$\sigma_{1/2} = -\frac{k_B T}{\Delta A} \ln\{L^{C\to O}(0)\} \tag{21}$$

As is the case for ligand- and voltage-activated gating, the gating of mechanosensitive channels is more complicated than summarized here: changes in bilayer tension will cause changes in bilayer thickness (Evans and Needham, 1987), which in their own right may alter channel gating as has been demonstrated in the case of the gramicidin channel monomer ↔ dimer equilibrium (Goulian *et al.*, 1998); there are usually multiple closed, or intermediate conductance, states in the activation path (Anishkin, Chiang and Sukharev, 2005); and sustained stretch can promote transitions to one or more nonconducting, desensitized states [Figure 17.8; cf. Honore (2007)].

17.5.6
Other Mechanisms

Channel function also may be regulated by channel insertion into, or retrieval from, the target membrane – mediated by fusion, or budding, of vesicles in which the channels are incorporated in the vesicle membrane (Snyder, 2005). These regulatory mechanisms will not be discussed more here.

17.6
Regulation of Channel Gating

Ion channels, as other enzymes, are functionally regulated by covalent modification of the constituent protein [e.g. by phosphorylation (Hille, 2001) or oxidation (Barford, 2004)]; by small molecules that bind reversibly to the channel protein (Li *et al.*, 2005; Kaczorowski *et al.*, 2008; see also Ligand-Operated Ion Channels and Drug Design Strategies for Targeting G-Protein-Coupled Receptors); and by amphiphiles that bind the surrounding host bilayer (Andersen, 2008; Lundbæk, 2008). The mechanisms that underlie the regulation of globular proteins, therefore, also pertain to the regulation of ion channels (membrane proteins) (cf. Berg, Tymoczko and Stryer (2007), Chapters 7–10).

17.6.1
Ion Channels Are Special

Ion channels, however, are subject to additional regulatory mechanisms that do not pertain to globular proteins and which will be the focus here. These novel regulatory mechanisms arise from a key difference between ion channels (and other membrane proteins) and globular proteins, namely that the channels are inserted into and span the membranes' lipid bilayer. This is important because the conformational changes that underlie ion channel function involve the channels' TM domain and, therefore, perturb the packing of the adjacent lipids (Unwin, 1995; Perozo, Cortes and Cuello 1999; and Perozo *et al.*, 2002; cf. Figure 17.10], which incurs an energetic cost (Huang, 1986).

17.6.2
Hydrophobic Matching and Hydrophobic Coupling

The hydrophobic interactions between the channel's hydrophobic TM domain and the host lipid bilayer cause the bilayer to adapt to the channel's exterior, an adaptation called *hydrophobic matching* (Mouritsen and Bloom, 1984); channel conformational transitions that involve the channel/bilayer interface will alter this local bilayer perturbation, which incurs an energetic cost (Andersen and Koeppe, 2007, see also Chapter 10). For example, a transition between two different channel states, a closed (C) and an open (O) state, will be associated with a local bilayer deformation (Figure 17.10). The total free energy change for the protein conformational change $\Delta G_{\text{tot}}^{C\to O}$ therefore may be described as

$$\begin{aligned}\Delta G_{\text{tot}}^{C\to O} &= \Delta G_{\text{prot}}^{C\to O}\\ &\quad + \Delta\Delta G_{\text{bilayer, local}}^{C\to O}\\ &\quad + \Delta\Delta G_{\text{bilayer, bulk}}^{C\to O} \qquad (22)\end{aligned}$$

where $\Delta G_{\text{prot}}^{C\to O}$ is the energetic contribution from the protein conformational change *per se* (not related to the bilayer

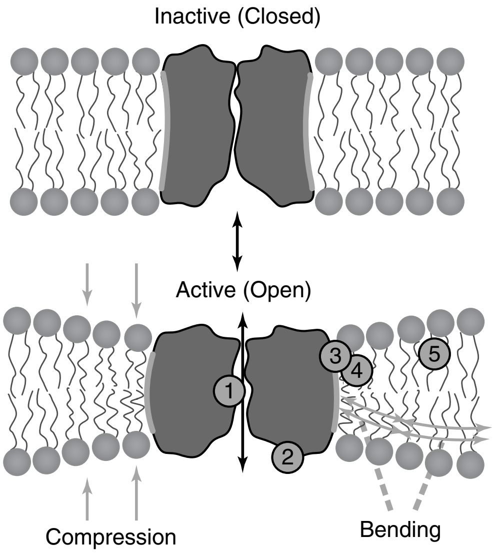

Fig. 17.10 Schematic representation of different, nonexclusive mechanisms by which one can regulate ion channel function. For simplicity, the conformational change underlying the C ↔ O is illustrated as a decrease in the length of the channel's hydrophobic TM domain, which is associated with a local deformation of the bilayer hydrophobic core (compression and bending of the two bilayer leaflets). (i) A drug may bind in, or very close to, the pore to block ion movement, (ii) a drug may bind to a site wholly formed by the protein to either inhibit or potentiate the channel function by altering the free energy difference between different channel states, (iii) an amphiphilic drug may bind specifically to a site composed of both the protein and the bilayer lipids, in which case its effects can involve also changes in the bilayer deformation energy associated with channel conformational changes, (iv) an amphiphilic drug may accumulate nonspecifically at the protein/bilayer interface to alter local lipid packing and thereby alter the bilayer deformation energy contribution to channel's conformational changes, and (v) an amphiphilic drug may adsorb at the lipid bilayer/solution interface to alter the lipid bilayer material properties and thereby the bilayer deformation energy associated with channel conformational changes

perturbation), $\Delta\Delta G^{C\to O}_{\text{bilayer, local}}$ is the energetic cost of the changes in local lipid packing ($\Delta\Delta G^{C\to O}_{\text{bilayer, local}} = \Delta G^{O}_{\text{bilayer, local}} - \Delta G^{C}_{\text{bilayer, local}}$, where $\Delta G^{C}_{\text{bilayer, local}}$ and $\Delta G^{O}_{\text{bilayer, local}}$ denote the energetic cost of local lipid packing at the channel/bilayer interface for C and O, respectively), and $\Delta\Delta G^{C\to O}_{\text{bilayer, bulk}}$ ($= \Delta G^{O}_{\text{bilayer, bulk}} - \Delta G^{C}_{\text{bilayer, bulk}}$) is the corresponding energetic cost of the bulk bilayer perturbation (Nielsen and Andersen, 2000). The equilibrium distribution between the two protein conformations is given by Andersen and Koeppe (2007) and Lundbæk (2006):

$$\frac{[\mathrm{O}]}{[\mathrm{C}]} = \exp\left\{-\frac{\Delta G^{C\to O}_{\text{tot}}}{k_B T}\right\} = \exp\left\{-\frac{\Delta G^{C\to O}_{\text{prot}} + \Delta\Delta G^{C\to O}_{\text{bilayer, local}} + \Delta\Delta G^{C\to O}_{\text{bilayer, bulk}}}{k_B T}\right\} \qquad (23)$$

The latter two energetic contributions vary as a function of the bilayer material properties [thickness, lipid intrinsic curvature, and the associated elastic moduli (Nielsen and Andersen, 2000)]; that is, there is *hydrophobic coupling* between the bilayer material properties and the function of bilayer-embedded proteins (ion channels) (Lundbæk, 2006).

17.7
Mechanisms of Drug-Induced Ion Channel Modulation

Small molecules (e.g. drugs) can regulate ion channel function by at least five different mechanisms (Andersen, 2008; cf. Figure 17.10). Specific examples of each of these mechanisms are listed in Table 17.5. The molecules may bind in – or near – the channel pore to physically occlude the lumen and block ion movement (see mechanism 1 in Figure 17.10 and Section 17.5.2). The molecules may regulate ion channel gating allosterically by binding to sites that are formed by the protein itself (as a conventional ligand binding site) to alter the $\Delta G_{\text{prot}}^{C\to O}$ contribution to $\Delta G_{\text{tot}}^{C\to O}$ (see mechanism 2 in Figure 17.10 and Section 17.5.3). These two mechanisms are general and apply also to globular proteins.

Because many drugs are amphiphiles [being hydrophobic and yet having some aqueous solubility (Lipinski *et al.*, 2001; Leeson and Springthorpe, 2007)], they will adsorb to the host lipid bilayer in addition to binding to the target channel (membrane protein). This means that they can regulate channel function by additional allosteric mechanisms. First, they can bind at the channel/bilayer interface to alter the $\Delta\Delta G_{\text{bilayer, local}}^{C\to O}$ (and also $\Delta G_{\text{prot}}^{C\to O}$) contribution to $\Delta\Delta G_{\text{tot}}^{C\to O}$ (mechanism 3 in Figure 17.10). Secondly, because the lipid bilayer material properties vary when amphiphiles adsorb to the bilayer/electrolyte interface (Ly and Longo, 2004), they can accumulate in the bilayer near the channel/bilayer interface and alter the $\Delta\Delta G_{\text{bilayer, local}}^{C\to O}$ contribution to $\Delta G_{\text{tot}}^{C\to O}$ (mechanism 4 in Figure 17.10), and they can adsorb to the bulk bilayer and alter the $\Delta\Delta G_{\text{bilayer, bulk}}^{C\to O}$ contribution to $\Delta G_{\text{tot}}^{C\to O}$ (mechanism 5 in Figure 17.10).

The common feature among mechanisms 2–5 in Figure 17.10 is that ligand binding, to the protein or the bilayer/solution interface, alters the contributions to $\Delta G_{\text{tot}}^{0}$ for the conformational transitions between different channel states (cf. Eq. (22)). At one extreme, mechanism 2 involves changes in $\Delta G_{\text{prot}}^{C\to O}$, a type of regulation that is shared by other (membrane-associated or water-soluble) enzymes and can be described by standard theories of ligand–receptor interactions (cf. Section 17.5.3). At the other extreme, mechanism 5 depends solely on changes in $\Delta\Delta G_{\text{bilayer, bulk}}^{C\to O}$, which are determined by the bilayer continuum material properties. Mechanisms 3 and 4 depend on changes in $\Delta\Delta G_{\text{bilayer, local}}^{C\to O}$ (and, in the case of mechanism 3, also $\Delta G_{\text{prot}}^{C\to O}$), which like $\Delta\Delta G_{\text{bilayer, bulk}}^{C\to O}$ are determined by the molecular composition of the bilayer immediately surrounding the channel. They also may involve local accumulation of molecules to diminish the energetic constraints on the local packing of lipid molecules adjacent to the channel – as deduced for polyunsaturated fatty acids (Bruno, Koeppe and Andersen, 2007). The key difference between these two mechanisms and mechanism 5 is that the magnitude

Tab. 17.5 Specific examples of mechanisms 1–5.

Mechanism	Channel	Molecule	Reference
1	Na_V	Tetrodotoxin	Hille (1975c)
1	Ca_V	ω -Conotoxins	Olivera *et al.* (1994)
1	K_V	Tetraethylammonium	Armstrong (1969)
2	Na_V	α- and β-Scorpion toxins	Catterall *et al.* (2007)
2	nAChR	ACh	Colquhoun and Sakmann (1985)
2	$GABA_A$	GABA	Kash, Trudell and Harrison (2004)
2’	$GABA_A$	Diazepam	Rudolph and Möhler (2004)
2’	nAChR	α-Bungarotoxin	Czajkowski, Kaufmann and Karlin (1993)
3	Kir	PIP_2	Hilgemann, Feng and Nasuhoglu (2001) and Suh and Hille (2005)
3	ENaC	PIP_2	Hilgemann, Feng and Nasuhoglu (2001) and Suh and Hille (2005)
3	TRPV	PIP_2	Hilgemann, Feng and Nasuhoglu (2001) and Suh and Hille (2005)
4 & 5	Na_V	Capsaicin	Lundbæk *et al.* (2005)
4 & 5	$GABA_A$	Docosahexaenoic acid*	Søgaard *et al.* (2006)
4 & 5	CFTR	Butanedione monoxime	Artigas *et al.* (2006)

Mechanism 2′ denotes the case in which the binding of one compound alters the affinity for the native ligand (e.g. GABA for $GABA_A$ and Ach for AChR).
PIP_2,phosphatidylinositol(4,5)bisphosphate.
* Docosahexanoic acid may exert its effect also through mechanism 3.

of $\Delta\Delta G^{C\to O}_{\text{bilayer, local}}$ also depends on the protein “shape.” The difference between mechanisms 3 and 4 is that the site of modulation in 3 is formed in part by the channel’s TM domain, which means that specific binding of the modulator to this site may occur (as well as a contribution from $\Delta G^{C\to O}_{\text{prot}}$).

A given molecule may affect ion channel function by any combination of these mechanisms. Mechanism 2 is the conventional allosteric mechanism. (Some drugs, such as diazepam, exert their action by binding to a site on the target and potentiating (or inhibiting) the action of another drug, such as GABA on the $GABA_A$ receptor; we denote this mechanism 2′, cf. Table 17.5.) To the extent that $\Delta\Delta G^{C\to O}_{\text{bilayer, local}}$ or $\Delta\Delta G^{C\to O}_{\text{bilayer, bulk}}$ are significant, on the order of $k_B T$, mechanisms 3–5 become important, and the lipid bilayer will be an allosteric regulator of ion channel (membrane protein) function (Andersen and Koeppe, 2007). Only mechanisms 1 and 2, and to a lesser extent 3, depend on specific, stoichiometric interactions with the channel; only mechanism 5 is altogether nonspecific. A challenge in future drug development becomes to identify the relative contribution of each of these mechanisms to a potential drug's overall effects – for then to optimize for the desired action. In this context, it is important that drugs with high affinity for the lipid bilayer/solution interface, relative to sites 1–3, may exert a significant part of their action through mechanisms 4 and 5, which make them promiscuous modulators of membrane protein function.

We finally note that molecules that reduce channel currents often are described as channel "blockers," whether or not the molecule in question actually enters the pore to occlude it. This is bad practice; the term blocker should be reserved only for mechanism 1, where there is a physical block or occlusion of the pore. Other molecules that reduce the channel-mediated ion movement are best described as channel "inhibitors." This terminology becomes important when a current reduction that is potentiated by repeated or frequent channel activation (as often is the case) is described as a "use-dependent block" – with the implication that mechanism 1 is involved. Any combination of mechanisms "1" to "5" may cause such an observation [cf. Lundbæk *et al.* (2005)], which is better described as "frequency dependent inhibition."

17.8
Conclusion

Ion channels constitute an important group of membrane proteins and are involved in many different biological functions. Because ion channels are generally efficient catalysts, they must be tightly regulated by mechanisms that control the channels' surface density and by mechanism that regulate the functional state of each individual channel. In addition to their physiological importance, the high turnover numbers allow for high-resolution measurements at the single-molecule level, which allow for detailed studies of how membrane protein function can be modulated/regulated. Finally, ion channels can be regulated – in addition to the mechanisms that have been established for globular proteins – by mechanisms that are unique for membrane proteins.

Acknowledgments

The preparation of this chapter is supported, in part, by National Institute of Health grants GM021342 and GM070971. We thank Ruchi Kapoor for helpful comments on the manuscript.

References

Accardi, A., and Miller, C. (2004) Secondary active transport mediated by a prokaryotic homologue of ClC Cl-channels. *Nature*, **427**, 803–7.

Almers, W., and McCleskey, E.W. (1984) Nonselective conductance in calcium channels of frog muscle: calcium selectivity in a single-file pore. *J. Physiol.*, **353**, 585–608.

Andersen, O.S. (1983) Ion movement through gramicidin A channels. Studies on the diffusion-controlled association step. *Biophys. J.*, **41**, 147–65.

Andersen, O.S. (1989) Kinetics of ion movement mediated by carriers and channels. *Meth. Enzymol.*, **171**, 62–112.

Andersen, O.S. (1999) Graphic representation of the results of kinetic analyses. *J. Gen. Physiol.*, **114**, 589–90.

Andersen, O.S. (2008) Perspectives on how to drug an ion channel. *J. Gen. Physiol.*, **131**, 395–97.

Andersen, O.S., and Koeppe, R.E. II (1992) Molecular determinants of channel function. *Physiol. Rev.*, **72**, S89–S158.

Andersen O.S., Koeppe II R.E., and Roux B.. (2007). Gramicidin channels. Versatile tools, in *Biological Membrane Ion Channels* (eds S.-H. Chung, O.S. Andersen, and V. Krishnamurthy), Springer Verlag, New York, pp. 33–80.

Andersen, O.S., and Koeppe, R.E. II (2007b) Bilayer thickness and membrane protein function: an energetic perspective. *Annu. Rev. Biophys. Biomol. Struct.*, **36**, 107–30.

Anishkin, A., Chiang, C.S., and Sukharev, S. (2005) Gain-of-function mutations reveal expanded intermediate states and a sequential action of two gates in MscL. *J. Gen. Physiol.*, **125**, 155–70.

Armstrong, C.M. (1969) Inactivation of the potassium conductance and related phenomena caused by quaternary ammonium ion injection in squid axons. *J. Gen. Physiol.*, **54**, 553–75.

Armstrong, C.M., and Bezanilla, F. (1973) Currents related to movement of the gating particles of the sodium channel. *Nature*, **242**, 459–61.

Artigas, P., Al'aref, S.J., Hobart, E.A., Diaz, L.F., Sakaguchi, M., Straw, S., and Andersen, O.S. (2006) 2,3-Butanedione monoxime affects cystic fibrosis transmembrane conductance regulator channel function through phosphorylation-dependent and phosphorylation-independent mechanisms: the role of bilayer material properties. *Mol. Pharmacol.*, **70**, 2015–26.

Ashcroft, F.M. (2006) From molecule to malady. *Nature*, **440**, 440–47.

Barford, D. (2004) The role of cysteine residues as redox-sensitive regulatory switches. *Curr. Opin. Struct. Biol.*, **14**, 679–86.

Bean, R.C., Shepherd, W.C., Chan, M., and Eichner, J. (1969) Discrete conductance fluctuations in lipid bilayer protein membranes. *J. Gen. Physiol.*, **53**, 741–57.

Beg, A.A., Ernstrom, G.G., Nix, P., Davis, M.W., and Jorgensen, E.M. (2008) Protons act as a transmitter for muscle contraction in *C. elegans*. *Cell*, **132**, 149–60.

Bennett, P.B., Yazawa, K., Makita, N., and George, A.L. Jr (1995) Molecular mechanism for an inherited cardiac arrhythmia. *Nature*, **376**, 683–85.

Berg, J.M., Tymoczko, J.L., and Stryer, L. (2007) *Biochemistry*, 6th edn, W. H. Freeman, New York.

Berridge, M.J., Bootman, M.D., and Roderick, H.L. (2003) Calcium signalling: dynamics, homeostasis and remodelling. *Nat. Rev. Mol. Cell Biol.*, **4**, 517–29.

Bezanilla, F. (2008) How membrane proteins sense voltage. *Nat. Rev. Mol. Cell Biol.*, **9**, 323–32.

Bond, C.T., Maylie, J., and Adelman, J.P. (2005) SK channels in excitability, pacemaking and synaptic integration. *Curr. Opin. Neurobiol.*, **15**, 305–11.

Bosanac, I., Michikawa, T., Mikoshiba, K., and Ikura, M. (2004) Structural insights into the regulatory mechanism of IP3 receptor. *Biochim. Biophys. Acta*, **1742**, 89–102.

Bruno, M.J., Koeppe, R.E. I, and Andersen, O.S. (2007) Docosahexaenoic acid alters bilayer elastic properties. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 9638–43.

Cannon, S.C. (2006) Pathomechanisms in channelopathies of skeletal muscle and brain. *Annu. Rev. Neurosci.*, **29**, 387–415.

Catterall, W.A., Cestele, S., Yarov-Yarovoy, V., Yu, F.H., Konoki, K., and Scheuer, T. (2007) Voltage-gated ion channels and gating modifier toxins. *Toxicon*, **49**, 124–41.

Catterall, W.A., Goldin, A.L., and Waxman, S.G. (2005) International Union of Pharmacology. XLVII. Nomenclature and structure-function relationships of voltage-gated sodium channels. *Pharmacol. Rev.*, **57**, 397–409.

Catterall, W.A., Perez-Reyes, E., Snutch, T.P., and Striessnig, J. (2005) International Union of Pharmacology. XLVIII. Nomenclature and structure-function relationships of voltage-gated calcium channels. *Pharmacol. Rev.*, **57**, 411–25.

Chakrapani, S., Cordero-Morales, J.F., and Perozo, E. (2007a) A quantitative description of KcsA gating I: macroscopic currents. *J. Gen. Physiol.*, **130**, 465–78.

Chakrapani, S., Cordero-Morales, J.F., and Perozo, E. (2007b) A quantitative description of KcsA gating II: single-channel currents. *J. Gen. Physiol.*, **130**, 479–96.

Chung, S.H., Andersen, O.S., and Krishnamurthy, V. (2007) *Biological Membrane Ion Channels: Dynamics, Structure and Applications*, Springer, New York.

Clapham, D.E., Julius, D., Montell, C., and Schultz, G. (2005) International Union of Pharmacology. XLIX. Nomenclature and structure-function relationships of transient receptor potential channels. *Pharmacol. Rev.*, **57**, 427–50.

Colquhoun, D. (1973) The relation between classical and cooperative models for drug action, in *Drug Receptors*, MacMillan Press, London, UK, pp. 149–82.

Colquhoun, D. (1998) Binding, gating, affinity and efficacy: the interpretation of structure-activity relationships for agonists and of the effects of mutating receptors. *Br. J. Pharmacol.*, **125**, 924–47.

Colquhoun, D. (2006) Agonist-activated ion channels. *Br. J. Pharmacol.*, **1471**, S17–S26.

Colquhoun, D. and Sakmann, B. (1985) Fast events in single-channel currents activated by acetylcholine and its analogues at the frog muscle end-plate. *J. Physiol.*, **369**, 501–57.

Cooper, K.E., Jakobsson, E., and Wolynes, E. (1985) The theory of ion transport through membrane channels. *Prog. Biophys. Mol. Biol.*, **46**, 51–96.

Corry, B. and Hool, L. (2007) *Calcium Channels* (eds S.H. Chung, O.S. Andersen., and V. Krishnamurthy), Springer, New York. pp. 241–99.

Cowan, S.W., Garavito, R.M., Jansonius, J.N., Jenkins, J.A., Karlsson, R., Konig, N., Pai, E.F., Pauptit, R.A., Rizkallah, P.J., Rosenbusch, J.P., Rummel, G., and Schirmer, T. (1995) The structure of OmpF porin in a tetragonal crystal form. *Structure*, **3**, 1041–50.

Cox, D.H. (2007) BKCa-channel structure and function, in *Biological Membrane Ion Channels. Dynamics, Structure, and Applications* (eds S.H. Chung, O.S. Andersen, and V. Krishnamurthy), Springer, New York, pp. 171–218.

Craven, K.B. and Zagotta, W.N. (2006) CNG and HCN channels: two peas, one pod. *Annu. Rev. Physiol.*, **68**, 375–401.

Czajkowski, C., Kaufmann, C., and Karlin, A. (1993) Negatively charged amino acid residues in the nicotinic receptor delta subunit that contribute to the binding of acetylcholine. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 6285–89.

Dahl, G. and Locovei, S. (2006) Pannexin: to gap or not to gap, is that a question?. *IUBMB Life*, **58**, 409–19.

DeCoursey, T.E. and Cherny, V.V. (1998) Temperature dependence of voltage-gated H+ currents in human neutrophils, rat alveolar epithelial cells, and mammalian phagocytes. *J. Gen. Physiol.*, **112**, 503–22.

Decoursey, T.E. (2003) Voltage-gated proton channels and other proton transfer pathways. *Physiol. Rev.*, **83**, 475–579.

Delcour, A.H. (2002) Structure and function of pore-forming beta-barrels from bacteria. *J. Mol. Microbiol. Biotechnol.*, **4**, 1–10.

Delmas, P. and Brown, D.A. (2005) Pathways modulating neural KCNQ/M (Kv7) potassium channels. *Nat. Rev. Neurosci.*, **6**, 850–62.

Doyle, D.A., Cabral, J.M., Pfuetzner, R.A., Kuo, A., Gulbis, J.M., Cohen, S.L., Chait, B.T., and MacKinnon, R. (1998) The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science*, **280**, 69–77.

Du, K., Sharma, M., and Lukacs, G.L. (2005) The Δ F508 cystic fibrosis mutation impairs domain-domain interactions and arrests post-translational folding of CFTR. *Nat. Struct. Mol. Biol.*, **12**, 17–25.

Duy, D., Soll, J., and Philippar, K. (2007) Solute channels of the outer membrane: from bacteria to chloroplasts. *Biol. Chem.*, **388**, 879–89.

Egan, T.M., Samways, D.S., and Li, Z. (2006) Biophysics of P2X receptors. *Pflugers Arch.*, **452**, 501–12.

Eisenberg, R.S. (1990) Channels as enzymes. *J. Membr. Biol.*, **115**, 1–12.

Evans, E. and Needham, D. (1987) Physical properties of surfactant bilayer membranes: thermal transitions, elasticity, rigidity, cohesion, and colloidal interactions. *J. Phys. Chem.*, **91**, 4219–28.

Fain, G.L. (1999) *Molecular and Cellular Physiology of Neurons*, Harvard University Press, Cambridge, MA.

Fatima-Shad, K. and Barry, P.H. (1993) Anion permeation in GABA- and glycine-gated channels of mammalian cultured hippocampal neurons. *Proc. R. Soc. Lond., B, Biol. Sci.*, **253**, 69–75.

Ferry, J.D. (1936) Statistical evaluation of sieve constants in ultrafiltration. *J. Gen. Physiol.*, **20**, 95–104.

Filosa, J.A., Bonev, A.D., Straub, S.V., Meredith, A.L., Wilkerson, M.K., Aldrich, R.W., and Nelson, M.T. (2006) Local potassium signaling couples neuronal activity to vasodilation in the brain. *Nat. Neurosci.*, **9**, 1397–403.

Finkelstein, A. (1975) Membrane channels and conductance. *Ann. N. Y. Acad. Sci.*, **264**, 244–46.

Finkelstein, A. and Andersen, O.S. (1981) The gramicidin A channel: a review of its permeability characteristics with special reference to the single-file aspect of transport. *J. Membr. Biol.*, **59**, 155–71.

Frankenhaeuser, B. and Hodgkin, A.L. (1956) The after-effects of impulses in the giant nerve fibres of Loligo. *J. Physiol.*, **131**, 341–76.

Gadsby, D.C., Vergani, P., and Csanady, L. (2006) The ABC protein turned chloride channel whose failure causes cystic fibrosis. *Nature*, **440**, 477–83.

Garty, H. and Palmer, L.G. (1997) Epithelial sodium channels: function, structure, and regulation. *Physiol. Rev.*, **77**, 359–96.

Gates, P.Y., Cooper, K.W., and Eisenberg, R.S. (1990) Analytical diffusion models for membrane channels. *Ion Channels*, **2**, 223–81.

Gentet, L.J., Stuart, G.J., and Clements, J.D. (2000) Direct measurement of specific membrane capacitance in neurons. *Biophys. J.*, **79**, 314–20.

Goldstein, S.A., Bayliss, D.A., Kim, D., Lesage, F., Plant, L.D., and Rajan, S. (2005) International Union of Pharmacology. LV. Nomenclature and molecular relationships of two-P potassium channels. *Pharmacol. Rev.*, **57**, 527–40.

Goodman MBSchwarz, E.M. (2003) Transducing touch in *Caenorhabditis elegans*. *Annu. Rev. Physiol.*, **65**, 429–52.

Goulian, M., Mesquita, O.N., Fygenson, D.K.K., Nielsen, C., Andersen, O.S., and Libchaber, A. (1998) Gramicidin channel kinetics under tension. *Biophys. J.*, **74**, 328–37.

Greathouse, D.V., Koeppe, R.E. II, Providence, L.L., Shobana, S., Andersen O.S. (1999) Design and characterization of gramicidin channels. *Meth. Enzymol.*, **294**, 525–50.

Gutman, G.A., Chandy, K.G., Grissmer, S., Lazdunski, M., McKinnon, D., Pardo, L.A., Robertson, G.A., Rudy, B., Sanguinetti, M.C., Stuhmer, W., and Wang, X. (2005) International Union of Pharmacology. LIII. Nomenclature and molecular relationships of voltage-gated potassium channels. *Pharmacol. Rev.*, **57**, 473–508.

Hamill, O.P. and Martinac, B. (2001) Molecular basis of mechanotransduction in living cells. *Physiol. Rev.*, **81**, 685–740.

Hanai, T., Haydon, D.A., and Taylor, J. (1965) The variation of capacitance and conductance of bimolecular lipid membranes with area. *J. Theor. Biol.*, **9**, 433–43.

Hanck, D.A. and Fozzard, H.A. (2007) Voltage-gated sodium channels, in *Biological Membrane Ion Channels. Dynamics, Structure, and Applications* (eds S.H. Chung, O.S. Andersen, and V. Krishnamurthy), Springer, New York, pp. 219–39.

Hansen, K.B., Yuan, H., and Traynelis, S.F. (2007) Structural aspects of AMPA receptor activation, desensitization and deactivation. *Curr. Opin. Neurobiol.*, **17**, 281–88.

Harris, A.L. (2001) Emerging issues of connexin channels: biophysics fills the gap. *Q. Rev. Biophys.*, **34**, 325–472.

Harris, A.L. (2007) Connexin channel permeability to cytoplasmic molecules. *Prog. Biophys. Mol. Biol.*, **94**, 120–43.

Hartzell, H.C., Qu, Z., Yu, K., Xiao, Q., and Chien, L.T. (2008) Molecular physiology of bestrophins: multifunctional membrane proteins linked to best disease and other retinopathies. *Physiol. Rev.*, **88**, 639–72.

Helmchen, F., Borst, J.G., and Sakmann, B. (1997) Calcium dynamics associated with a single action potential in a CNS presynaptic terminal. *Biophys. J.*, **72**, 1458–71.

Hilgemann, D.W., Feng, S., and Nasuhoglu, C. (2001) The complex and intriguing lives of PIP2 with ion channels and transporters. *Sci. STKE*, RE19.

Hille, B. (1975a) Ionic selectivity of Na and K channels in nerve membranes, in *Membranes. Lipid Bilayers and Biological Membranes: Dynamic Properties* (G. Eisenman) Marcel Dekker, Inc., New York.

Hille, B. (1975b) Ionic selectivity, saturation, and block in sodium channels. A four-barrier model. *J. Gen. Physiol.*, **66**, 535–60.

Hille, B. (1975c) The receptor for tetrodotoxin and saxitoxin. A structural hypothesis. *Biophys. J.*, **15**, 615–19.

Hille B. (2001) *Ionic Channels of Excitable Membranes*, 3rd edn, Sinauer, Sunderland, MA.

Hille, B. and Schwarz, W. (1978) Potassium channels as multi-ion single-file pores. *J. Gen. Physiol.*, **72**, 409–42.

Hirschberg, B., Rovner, A., Lieberman, M., and Patlak, J. (1995) Transfer of twelve charges is needed to open skeletal muscle Na+ channels. *J. Gen. Physiol.*, **106**, 1053–68.

Hladky, S.B. and Haydon, D.A. (1970) Discreteness of conductance change in bimolecular lipid membranes in the presence of certain antibiotics. *Nature*, **225**, 451–53.

Hodgkin, A.L. (1964) *The Conduction of the Nervous Impulse*, Liverpool University Press, Liverpool, UK.

Hodgkin, A.L. and Huxley, A.F. (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, **117**, 500–44.

Hodgkin, A.L. and Katz, B. (1949) The effect of sodium ions on the electrical activity of the giant axon of the squid. *J. Physiol.*, **108**, 37–77.

Hofmann, F., Biel, M., and Kaupp, U.B. (2005) International Union of Pharmacology. LI. Nomenclature and structure-function relationships of cyclic nucleotide-regulated channels. *Pharmacol. Rev.*, **57**, 455–62.

Honore, E. (2007) The neuronal background K2P channels: focus on TREK1. *Nat. Rev. Neurosci.*, **8**, 251–61.

Horrigan, F.T. and Aldrich, R.W. (2002) Coupling between voltage sensor activation, Ca2+ binding and channel opening in large conductance (BK) potassium channels. *J. Gen. Physiol.*, **120**, 267–305.

Huang, H.W. (1986) Deformation free energy of bilayer membrane and its effect on gramicidin channel lifetime. *Biophys. J.*, **50**, 1061–70.

Jackson, M.B. (1989) Perfection of a synaptic receptor: kinetics and energetics of the acetylcholine receptor. *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 2199–203.

Jasti, J., Furukawa, H., Gonzales, E.B., and Gouaux, E. (2007) Structure of acid-sensing ion channel 1 at 1.9 Å resolution and low pH. *Nature*, **449**, 316–23.

Jentsch, T.J. (2008) CLC chloride channels and transporters: from genes to protein structure, pathology and physiology. *Crit. Rev. Biochem. Mol. Biol.*, **43**, 3–36.

Jiang, Y., Lee, A., Chen, J., Ruta, V., Cadene, M., Chait, B.T., and MacKinnon, R. (2003) X-ray structure of a voltage-dependent K^+ channel. *Nature*, **423**, 33–41.

Kaczorowski, G.J., McManus, O.B., Priest, B.T., and Garcia, M.L. (2008) Ion channels as drug targets: the next GPCRs. *J. Gen. Physiol.*, **131**, 399–405.

Kanaporis, G., Mese, G., Valiuniene, L., White, T.W., Brink, P.R., and Valiunas, V. (2008) Gap junction channels exhibit connexin-specific permeability to cyclic nucleotides. *J. Gen. Physiol.*, **131**, 293–305.

Kang, M.A., Moroni, A., Gazzarrini, S., DiFrancesco, D., Thiel, G., Severino, M., and Van Etten, J.L. (2004) Small potassium ion channel proteins encoded by chlorella viruses. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 5318–24.

Karlin, A. (1967) On the application of "a plausible model" of allosteric proteins to the receptor for acetylcholine. *J. Theor. Biol.*, **16**, 306–20.

Kash, T.L., Trudell, J.R., and Harrison, N.L. (2004) Structural elements involved in activation of the γ-aminobutyric acid type A (GABAA) receptor. *Biochem. Soc. Trans.*, **32**, 540–46.

Kofuji, P., Hofer, M., Millen, K.J., Millonig, J.H., Davidson, N., Lester, H.A., and Hatten, M.E. (1996) Functional analysis of the weaver mutant GIRK2 K^+ channel and rescue of weaver granule cells. *Neuron*, **16**, 941–52.

Korn, S.J., and Trapani, J.G. (2007) Voltage-gated potassium channels, in *Biological Ion Channels: Dynamics, Structure, and Applications* (eds S.H. Chung, O.S. Andersen, and V. Krishnamurthy), Springer, New York, pp. 119–70.

Kubo, Y., Adelman, J.P., Clapham, D.E., Jan, L.Y., Karschin, A., Kurachi, Y., Lazdunski, M., Nichols, C.G., Seino, S., and Vandenberg, C.A. (2005) International Union of Pharmacology. LIV. Nomenclature and molecular relationships of inwardly rectifying potassium channels. *Pharmacol. Rev.*, **57**, 509–26.

Kullman, L., Winterhalter, M., and Bezrukov, S.M. (2002) Transport of maltodextrins through maltoporin: a single-channel study. *Biophys. J.*, **82**, 803–12.

Kuyucak, S., Andersen, O.S., and Chung, S. (2001) Models of permeation in ion channels. *Rep. Prog. Phys.*, **64**, 1427–72.

Latorre, R. and Miller, C. (1983) Conductance and selectivity of potassium channels. *J. Membr. Biol.*, **71**, 11–30.

Läuger, P. (1976) Diffusion-limited ion flow through pores. *Biochim. Biophys. Acta*, **455**, 493–509.

Läuger, P. (1980) Kinetic properties of ion carriers and channels. *J. Membr. Biol.*, **57**, 163–78.

Leeson, P.D. and Springthorpe, B. (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.*, **6**, 881–90.

Levitt, D.G. (1986) Interpretation of biological ion channel flux data–reaction-rate versus continuum theory. *Annu. Rev. Biophys. Biophys. Chem.*, **15**, 29–57.

Lewis, C.A. (1979) Ion-concentration dependence of the reversal potential and the single channel conductance of ion channels at the frog neuromuscular junction. *J. Physiol.*, **286**, 417–45.

Lewis, R.S. (2007) The molecular choreography of a store-operated calcium channel. *Nature*, **446**, 284–87.

Li, S., Gosling, M., Poll, C.T., Westwick, J., and Cox, B. (2005) Therapeutic scope of modulation of non-voltage-gated cation channels. *Drug Discov. Today*, **10**, 129–37.

Lingueglia, E., Deval, E., and Lazdunski, M. (2006) FMRFamide-gated sodium channel and ASIC channels: a new class of ionotropic receptors for FMRFamide and related peptides. *Peptides*, **27**, 1138–52.

Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, **46**, 3–26.

Long, S.B., Campbell, E.B., and Mackinnon, R. (2005) Crystal structure of a mammalian voltage-dependent Shaker family K^+ channel. *Science*, **309**, 897–903.

Long, S.B., Tao, X., Campbell, E.B., and MacKinnon, R. (2007) Atomic structure of a voltage-dependent K^+ channel in a lipid membrane-like environment. *Nature*, **450**, 376–82.

Lu, Z. (2004) Mechanism of rectification in inward-rectifier K^+ channels. *Annu. Rev. Physiol.*, **66**, 103–29.

Lundbæk, J.A. (2006) Regulation of membrane protein function by lipid bilayer elasticity: a single molecule technology to measure the bilayer properties experienced by an embedded protein. *J. Phys. Condens. Matter*, **18**, S1305–S44.

Lundbæk, J.A. (2008) Lipid bilayer-mediated regulation of ion channel function by amphiphilic drugs. *J. Gen. Physiol.*, **131**, 421–29.

Lundbæk, J.A., Birn, P., Tape, S.E., Toombes, G.E., Søgaard, R., Koeppe, R.E. II, Gruner, S.M., Hansen, A.J., and Andersen, O.S. (2005) Capsaicin regulates voltage-dependent sodium channels by altering lipid bilayer elasticity. *Mol. Pharmacol.*, **68**, 680–89.

Ly, H.V. and Longo, M.L. (2004) The influence of short-chain alcohols on interfacial tension, mechanical properties, area/molecule, and permeability of fluid lipid bilayers. *Biophys. J.*, **87**, 1013–33.

Lynch, J.W. and Barry, P.H. (2007) in *Ligand-Gated Ion Channels: Permeation and Activation* (eds S.H. Chung, O.S. Andersen, and V. Krishnamurthy), Springer, New York, pp. 335–67.

Maconochie, D.J., Fletcher, G.H., and Steinbach, J.H. (1995) The conductance of the muscle nicotinic receptor channel changes rapidly upon gating. *Biophys. J.*, **68**, 483–90.

Magleby, K.L. (2001) Kinetic gating mechanisms for BK channels: when complexity leads to simplicity. *J. Gen. Physiol.*, **118**, 583–87.

Magleby, K.L. and Stevens, C.F. (1972) A quantitative description of end-plate currents. *J. Physiol.*, **223**, 173–97.

Markin, V.S. and Hudspeth, A.J. (1995) Gating-spring models of mechanoelectrical transduction by hair cells of the internal ear. *Annu. Rev. Biophys. Biomol. Struct.*, **24**, 59–83.

Markin, V.S. and Sachs, F. (2007) Thermodynamics of mechanosensitivity. *Curr. Top. Membr.*, **58**, 87–119.

Martinac, B. (2007) Mechanosensitive channels, in *Biological Membrane Ion Channels. Dynamics, Structure, and Applications* (eds S.H. Chung, O.S. Andersen, and V. Krishnamurthy), Springer, New York, pp. 369–98.

Mayer, M.L. (2006) Glutamate receptors at atomic resolution. *Nature*, **440**, 456–62.

Miller, C. (2006) ClC chloride channels viewed through a transporter lens. *Nature*, **440**, 484–89.

Mouritsen, O.G. and Bloom, M. (1984) Mattress model of lipid-protein interactions in membranes. *Biophys. J.*, **46**, 141–53.

Musset, B., Cherny, V.V., Morgan, D., Okamura, Y., Ramsey, I.S., Clapham, D.E., and DeCoursey, T.E. (2008) Detailed comparison of expressed and native voltage-gated proton channel currents. *J. Physiol.*, **586**, 2477–86.

Neher, E. (1995) The use of fura-2 for estimating Ca buffers and Ca fluxes. *Neuropharmacology*, **34**, 1423–42.

Neyton, J. and Miller, C. (1988) Potassium blocks barium permeation through a calcium-activated potassium channel. *J. Gen. Physiol.*, **92**, 549–67.

Nielsen, C. and Andersen, O.S. (2000) Inclusion-induced bilayer deformations: effects of monolayer equilibrium curvature. *Biophys. J.*, **79**, 2583–604.

Niles, W.D., Levis, R.A., and Cohen, F.S. (1988) Planar bilayer membranes made from phospholipid monolayers form by a thinning process. *Biophys. J.*, **53**, 327–35.

Nishida, M., Cadene, M., Chait, B.T., and MacKinnon, R. (2007) Crystal structure of a Kir3.1-prokaryotic Kir channel chimera. *EMBO J.*, **26**, 4005–15.

Olivera, B.M., Miljanich, G.P., Ramachandran, J., and Adams, M.E. (1994) Calcium channel diversity and neurotransmitter release: the ω-conotoxins and ω-agatoxins. *Annu. Rev. Biochem.*, **63**, 823–67.

Oswald, R.E., Ahmed, A., Fenwick, M.K., and Loh, A.P. (2007) Structure of glutamate receptors. *Curr. Drug Targets*, **8**, 573–82.

Palmer, L.G. (2007) Ion channels in epithelial cells, in *Biological Membrane Ion Channels: Dynamics, Structure, and Applications* (eds S.H. Chung, O.S. Andersen, and V. Krishnamurthy), Springer, New York, pp. 425–45.

Partenskii, M.B. and Jordan, P.C. (1992) Theoretical perspectives on ion-channel electrostatics: continuum and microscopic approaches. *Q. Rev. Biophys.*, **25**, 477–510.

Perozo, E., Cortes, D.M., and Cuello, L.G. (1999) Structural rearrangements underlying K^+-channel activation gating. *Science*, **285**, 73–78.

Perozo, E., Cortes, D.M., Sompornpisut, P., Kloda, A., and Martinac, B. (2002) Open channel structure of MscL and the gating mechanism of mechanosensitive channels. *Nature*, **418**, 942–48.

Pinto, L.H., Holsinger, L.J., and Lamb, R.A. (1992) Influenza virus M2 protein has ion channel activity. *Cell*, **69**, 517–28.

Pinto, L.H. and Lamb, R.A. (2006) The M2 proton channels of influenza A and B viruses. *J. Biol. Chem.*, **281**, 8997–9000.

Pusch, M. (2007) Chloride transporting CLC proteins, in *Biological Membrane Ion Channels. Dynamics, Structure, and Applications* (eds S.H. Chung, O.S. Andersen, and V. Krishnamurthy), Springer, New York, pp. 301–33.

Ramsey, I.S., Moran, M.M., Chong, J.A., and Clapham, D.E. (2006) A voltage-gated proton-selective channel lacking the pore domain. *Nature*, **440**, 1213–16.

Rizkallah, P.J., Rosenbusch, J.P., Rummel, G., and Schirmer, T. (1995) The structure of OmpF porin in a tetragonal crystal form. *Structure*, **3**, 1041–50.

Roberts, J.A., Vial, C., Digby, H.R., Agboh, K.C., Wen, H., Atterbury-Thomas, A., and Evans, R.J. (2006) Molecular properties of P2X receptors. *Pflugers Arch.*, **452**, 486–500.

Robinson, R.A. and Stokes, R.H. (1959) *Electrolyte Solutions*, 2nd edn, Butterworth, London, UK.

Roos, A. and Boron, W.F. (1981) Intracellular pH. *Physiol. Rev.*, **61**, 296–434.

Rossier, B.C., Pradervand, S., Schild, L., and Hummler, E. (2002) Epithelial sodium channel and the control of sodium balance: interaction between genetic and environmental factors. *Annu. Rev. Physiol.*, **64**, 877–97.

Rostovtseva, T.K., Tan, W., and Colombini, M. (2005) On the role of VDAC in apoptosis: fact and fiction. *J. Bioenerg. Biomembr.*, **37**, 129–42.

Roux, B., Allen, T.W., Bernèche, S., and Im, W. (2004) Theoretical and computational models of biological ion channels. *Q. Rev. Biophys.*, **37**, 15–103.

Ruch, T.C., Patton, H.D., and Scher, A.M. Physiology and Biophysics, *Circulation, Respiration and Fluid Balance*, 20th edn, Vol. II, W. B. Saunders, Philadelphia, PA.

Rudolph, U. and Möhler, H. (2004) Analysis of GABAA receptor function and dissection of the pharmacology of benzodiazepines and general anesthetics through mouse genetics. *Annu. Rev. Pharmacol. Toxicol.*, **44**, 475–98.

Sanguinetti, M.C. and Tristani-Firouzi, M. (2006) hERG potassium channels and cardiac arrhythmia. *Nature*, **440**, 463–69.

Schatzberg, P. (1965) Diffusion of water through hydrocarbon liquids. *J. Polym. Sci., Part C*, **10**, 87–92.

Schoppa, N.E., McCormack, K., Tanouye, M.A., and Sigworth, F.J. (1992) The size of gating charge in wild-type and mutant Shaker potassium channels. *Science*, **255**, 1712–15.

Shapovalov, G. and Lester, H.A. (2004) Gating transitions in bacterial ion channels measured at 3 μs resolution. *J. Gen. Physiol.*, **124**, 151–61.

Sine, S.M. and Engel, A.G. (2006) Recent advances in Cys-loop receptor structure and function. *Nature*, **440**, 448–55.

Snyder, P.M. (2005) Minireview: regulation of epithelial Na+ channel trafficking. *Endocrinology*, **146**, 5079–85.

Søgaard, R., Werge TMBertelsen, C., Lundbye, C., Madsen, K.L., Nielsen, C.H., and Lundbæk, J.A. (2006) GABAA receptor function is regulated by lipid bilayer elasticity. *Biochemistry*, **45**, 13118–29.

Sokolov, S., Scheuer, T., and Catterall, W.A. (2007) Gating pore current in an inherited ion channelopathy. *Nature*, **446**, 76–78.

Struyk, A.F. and Cannon, S.C. (2007) A Na+ channel mutation linked to hypokalemic periodic paralysis exposes a proton-selective gating pore. *J. Gen. Physiol.*, **130**, 11–20.

Suh, B.C. and Hille, B. (2005) Regulation of ion channels by phosphatidylinositol 4,5-bisphosphate. *Curr. Opin. Neurobiol.*, **15**, 370–78.

Sukharev, S. and Corey, D.P. (2004) Mechanosensitive channels: multiplicity of families and gating paradigms. *Sci. STKE*, re4.

Unwin, N. (1995) Acetylcholine receptor channel imaged in the open state. *Nature*, **373**, 37–43.

Unwin, N. (2005) Refined structure of the nicotinic acetylcholine receptor at 4 Å resolution. *J. Mol. Biol.*, **346**, 967–89.

Voets, T., Owsianik, G., and Nilius, B. (2007) TRP channels, in *Biological Membrane Ion Channels: Dynamics, Structure, and Applications* (eds S.H. Chung, O.S. Andersen, and V. Krishnamurthy), Springer, New York, pp. 399–423.

Wei, A.D., Gutman, G.A., Aldrich, R., Chandy, K.G., Grissmer, S., and Wulff, H. (2005) International Union of Pharmacology. LII. Nomenclature and molecular relationships of calcium-activated potassium channels. *Pharmacol. Rev.*, **57**, 463–72.

Wemmie, J.A., Price, M.P., and Welsh, M.J. (2006) Acid-sensing ion channels: advances, questions and therapeutic opportunities. *Trends Neurosci.*, **29**, 578–86.

Wo, Z.G. and Oswald, R.E. (1995) Unraveling the modular design of glutamate-gated ion channels. *Trends Neurosci.*, **18**, 161–68.

Wolfenden, R. and Snider, M.J. (2001) The depth of chemical time and the power of enzymes as catalysts. *Acc. Chem. Res.*, **12**, 938–45.

Woodhull, A.M. (1973) Ionic blockage of sodium channels in nerve. *J. Gen. Physiol.*, **61**, 687–708.

Yanagidate, F. and Strichartz, G.R. (2007) Local anesthetics. *Handb. Exp. Pharmacol.*, **177**, 95–127.

Yellen, G. (2002) The voltage-gated potassium channels and their relatives. *Nature*, **419**, 35–42.

Yen, M.R. and Saier, M.H. Jr (2007) Gap junctional proteins of animals: the innexin/pannexin superfamily. *Prog. Biophys. Mol. Biol.*, **94**, 5–14.

Zagotta, W.N., Hoshi, T., and Aldrich, R.W. (1994) Shaker potassium channel gating. III: Evaluation of kinetic models for activation. *J. Gen. Physiol.*, **103**, 321–62.

Zalk, R., Lehnart, S.E., and Marks, A.R. (2007) Modulation of the ryanodine receptor and intracellular calcium. *Annu. Rev. Biochem.*, **76**, 367–85.

Zheng, J. and Sigworth, F.J. (1998) Intermediate conductances during deactivation of heteromultimeric Shaker potassium channels. *J. Gen. Physiol.*, **112**, 457–74.

Zhou, Y., Morais-Cabral, J.H., Kaufman, A., and MacKinnon, R. (2001) Chemistry of ion coordination and hydration revealed by a K^+ channel-Fab complex at 2.0 Å resolution. *Nature*, **414**, 43–48.

Further Reading

Ashcroft, F.M. (2000) *Ion Channels and Disease*, Academic Press, Sand Diego, CA.

Sten-Knudsen, O. (2002) *Biological Membranes: Theory of Transport, Potentials and Electric Impulses*, Cambridge University Press, Cambridge, UK.

18
Physical Properties of Biological Membranes

Thomas Heimburg

18.1 Introduction 595
18.2 Composition of Membranes 596
18.2.1 Lipid Composition 596
18.2.1.1 Hydrocarbon Chains and Chain Length 597
18.2.1.2 Head Groups 597
18.2.2 Lipid Composition of Biomembranes 598
18.2.3 Dependence on Thermodynamic Variables 598
18.2.3.1 Growth Temperature 598
18.2.3.2 Growth Pressure 599
18.2.3.3 Solvents and Anesthetics 599
18.2.4 Membrane Proteins 599
18.3 Membrane Melting 600
18.3.1 Chain Melting 600
18.3.1.1 Temperature Dependence and Heat Capacity 600
18.3.1.2 Pressure Dependence 601
18.3.1.3 Solvent Dependence 601
18.3.1.4 Voltage Dependence 601
18.3.1.5 Dependence on Membrane Proteins 601
18.3.2 Melting in Biomembranes 602
18.4 Phases and Domains 602
18.4.1 Gibbs Phase Rule 603
18.4.2 Phase Diagrams 604
18.4.3 Domain Formation 605
18.4.3.1 Domain Formation in Artificial Membranes and Monolayers 605
18.4.3.2 Domain Formation in Biological Systems 605
18.5 Curvature and Elastic Constants 606
18.5.1 Principle Curvatures 606
18.5.2 Curved Phases 607
18.5.3 Elastic Constants 608
18.5.4 Temperature and Pressure Dependence of Elastic Constants 608
18.6 Electrostatic Potential 609

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

18.6.1 Ionic Strength Dependence 609
18.6.2 Dependence on Charged Lipid Concentration 610
18.6.3 Dependence on Bound Proteins 610
18.7 Permeability 610
18.7.1 Protein Ion Channels 611
18.7.2 Lipid Ion Channels 611
18.8 Pulse Propagation 612
18.8.1 Nerve Pulses 612
18.8.2 Soliton Propagation 612
18.8.3 Reversible Heat 612
18.8.4 Mechanical Changes 613
18.9 Anesthesia 613
References 613

18.1
Introduction

Biological membranes are thin layers surrounding cells and their organelles, which define inside and outside of compartments. Their major components are lipids and proteins. Lipids are small amphiphilic molecules that are hydrophilic on one side and hydrophobic on the other side. When brought into contact with water, most lipids spontaneously form bilayer membranes. Our present picture of biomembranes stems from Gorter and Grendel (1925), Danielli and Davson (1935) and Singer and Nicolson (1972). The latter paper described the membrane as a two-dimensional bilayer of fluid lipids with proteins adsorbed or embedded. The Singer–Nicolson paper still dominates our present understanding of membranes, but it experienced some modifications. (Mouritsen and Bloom, 1984) proposed that the membrane is laterally heterogeneous as schematically shown in Figure 18.1. Here one sees domain formation and proteins that associate with different domains with different likelihoods.

The typical thickness of a lipid bilayer is about 5 nm. Thus, these are extremely thin layers as compared to the typical dimensions of cells (several 10 μm). Such membranes are flexible, heterogenous, and permeable. However, these properties depend not only on the conditions, for example on temperature, pressure, electrical field, pH, and salt concentration, but also on the presence of proteins and protein conformation. This basically means that the state of a biological membrane depends on all thermodynamic variables. These dependencies are discussed throughout this chapter. Thus, in this chapter, we focus rather on the properties of membranes as a whole than on the properties and functions of the individual molecules.

The composition of membranes is complex (see Section 18.2 for details). There are hundreds or even thousands of different lipid species and further thousands of different membrane proteins. The composition of membranes is different not only between different species, but also between different cell types of the same organism, and even between the membranes of different organelles within the same cell. Proteins act, for example, as catalysts and the importance of their role seems intuitively clear, while the role of the lipid membrane and the range of different lipid compositions is less obvious, and there is no agreement in the biophysical community yet on

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

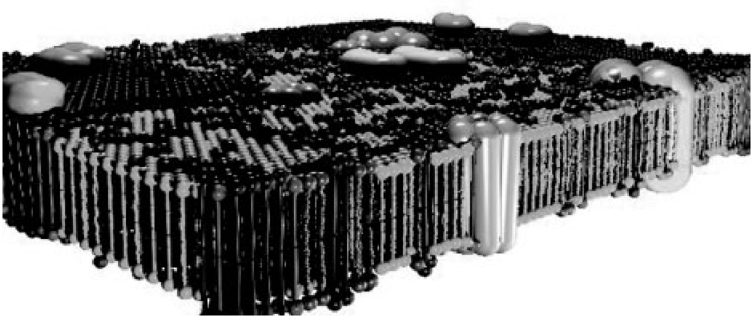

Fig. 18.1 Schematic representation of a biological membrane with various lipids in different physical states and embedded or adsorbed proteins. Courtesy Heiko Seeger.

what the purpose of the heterogeneity of the lipids precisely is.

Lipid membranes can undergo transitions from a rigid state into a liquid state (Section 18.3), where both the lateral order of lipid molecules and the order of the lipid hydrocarbon chains changes. In biological membranes, this occurs at temperatures that typically are slightly below body or growth temperature. Both states form two-dimensional membranes but their physical properties may be quite different. The lipid composition has an influence on membrane melting. The melting events in membranes lead to changes in the distribution of lipids within the membrane plane. In experiments, one finds regions of different composition and state. If they are of macroscopic dimensions, they are called *phases*; if they are on microscopic or nanoscopic scales, they are called *domains* or sometimes *rafts* in cell biology. The investigation of domain formation received much interest because of its putative influence on reaction cascades between proteins, and on diffusion processes within membranes. The melting transition in membranes is also interesting because of its influence on the elastic constants of membranes. In particular, at their melting temperatures, the membranes are very soft and easy to bend (Section 18.5). Related to this, the membranes also become very permeable for water, ions, and larger molecules (Section 18.7). Under some conditions, they may even propagate mechanical perturbations or pulses that resemble nerve pulses (Section 18.8).

18.2
Composition of Membranes

When discussing the composition of biological membranes, one has to distinguish proteins that are encoded in the genome, and the lipid composition that is not encoded in the genome. The latter rather adapts to the environmental conditions, partially by the control of membrane active proteins that display an activity depending on the physical state of the membrane.

18.2.1
Lipid Composition

There are hundreds of different lipids. What precisely counts as a lipid is only

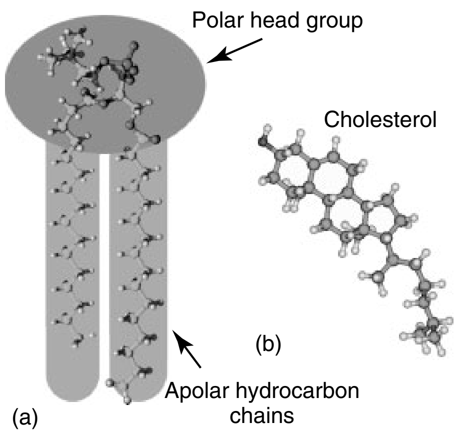

Fig. 18.2 (a) Dipalmitoyl phosphatidylcholine (DPPC) is a phospholipid molecule with zwitterionic polar head group and apolar fatty acid chains. (b) The steroid lipid cholesterol displays a quite different and much more rigid structure.

loosely defined. What lipids have in common is that they are small amphiphilic molecules that dissolve in membranes. By far the majority of these lipids are phospholipids with two hydrocarbon chains and a phosphate containing head group as shown in Figure 18.2a. However, one also finds minor fractions of lipids that only have one chain, or molecules of completely different structure as cholesterol (Figure 18.2a). This is a quite rigid molecule not forming membranes on its own. It is nevertheless a major component in some biological membranes, for example, in erythrocytes (Sackmann, 1995). One of the most obvious features being controlled by the chemical nature of the lipids is the melting point of the membranes, which is discussed in more detail in Section 18.3. Another property is the electrostatic potential of the membranes if the lipid head groups are charged.

18.2.1.1 Hydrocarbon Chains and Chain Length

Most lipids possess two hydrocarbon chains as shown in Figure 18.2. These chains vary in length from 12 to about 24 carbons. A second feature is the degree of unsaturation in the chain, that is, the number of double bonds in the chain. The chemical nature of the chains has a major influence on the melting temperatures of membranes as described in Section 18.3.

18.2.1.2 Head Groups

Lipids may also possess a variety of different polar head groups. Most lipids are phospholipids containing a phosphate group. About 40 mol% of the lipids in eucariotic cells are phosphatidylcholines (= lecithins) that are zwitterionic. They carry one negative and one positive charge in the physiological pH range. One example is shown in Figure 18.2. Another abundant heat group is the zwitterionic phosphatidylethanolamine that also contributes about 40 mol%. The rest are various other lipids that occur in smaller fractions, for example, the negatively charged phosphatidylserines, phosphatidylglycerols, phosphatidylinositols, and cardiolipins. Other lipids are sphingolipids and ceramides.

Tab. 18.1 Head group composition of the membranes of some mammalian liver cells, erythrocytes, and nerve cells in weight percent. Adapted from Jamieson and Robinson (1977).

Membrane	PC	PE	PS	PI	SM	CL	Glycolipid	Cholesterol	Others
Erythrocyte (human)	20	18	7	3	18	–	3	20	11
Plasma (rat liver)	18	12	7	3	12	–	8	19	21
ER	48	19	4	8	5	–	tr	6	10
Golgi	25	9	3	5	7	–	0	8	43
Lysosome	23	13	–	6	23	≈5	–	14	16
Nuclear membrane	44	17	4	6	3	1	tr	10	15
Mitochondria	38	29	0	3	0	14	tr	3	13
Neurons	48	21	5	7	4	–	3	11	1
Myelin	11	17	9	1	8	–	20	28	6

PC, phosphatidylcholines; PE, phosphatidylethanolamines; PS, phosphatidylserines; PI, phosphatidylinositols; SM, sphingomyelin; CL, cardiolipin.

18.2.2
Lipid Composition of Biomembranes

The lipid composition of membranes from different sources can be quite different. Examples are shown in Table 18.1. Human erythrocytes and rat liver plasma membranes, for instance, contain about 20 wt% cholesterol while rat liver mitochondria contain only 3 wt%. The phosphatidylcholine content of the endoplasmic reticulum is 48 wt% while the content in nerve myelin is only 11 wt%. Other examples can be found in Heimburg (2007) or Sackmann (1995). The differences in lipid composition between different cells, and even the organelles of the same cell are remarkable, but the reason is not completely understood.

18.2.3
Dependence on Thermodynamic Variables

It seems likely that the control of the thermodynamic parameters associated with the melting points of membranes (described in more detail in Section 18.3) belong to the most important features that are influenced by the lipid composition. This is a result of studying the lipid compositions of various organisms that live at conditions of variable temperature and pressure.

18.2.3.1 Growth Temperature

It has been shown that the lipid composition of cold-blooded organisms depends strongly on the ambient growth temperature. *Bacillus subtilis* bacteria show a largely altered lipid composition when grown at different temperatures (van de Vossenberg *et al.*, 1999). The composition of trout livers is very different in trouts caught in winter (water temperature ≈ 5 °C) and trouts in summer (water temperature ≈ 20 °C) (Hazel, 1979). In particular, the amount of unsaturated lipids is much higher in the livers of trouts in winter while the amount of saturated lipid is higher in summer. Unsaturated lipids have much lower melting temperatures than saturated lipids. It seems as if organisms adapt to their lipid compositions according to temperature in order to maintain a particular physical state of their membranes.

18.2.3.2 Growth Pressure

(DeLong and Yayanos, 1985) investigated barophilic marine bacteria. They found that the lipid composition of their membranes is highly dependent on growth pressure. Like temperature, pressure also has an influence on membrane melting. At high pressure, the melting point increases. (DeLong and Yayanos, 1985) found that the ratio of saturated to unsaturated chains linearly increases with ambient pressure.

18.2.3.3 Solvents and Anesthetics

Similar effects have been seen when *Escherichia coli* bacteria are grown in media containing solvents such as acetone or aniline (Ingram, 1977). Depending on the concentration of the solvents, one observes drastic changes in the composition of the membranes. In Section 18.3.1.3 it is shown that the adaptation of the lipid composition is accompanied by a change of the melting temperature of membranes.

18.2.4
Membrane Proteins

In contrast to the lipid structures, proteins are genetically encoded. The human genome contains about 30 000 genes that encode proteins, many of them being membrane proteins. Membrane proteins are much more difficult to crystallize than soluble proteins. Therefore, the number of known membrane protein structures is still small. Membrane proteins have various putative functions, but many are thought to be involved in transport processes across membranes. This includes the active transport of protons by the light driven rhodopsin, of sodium and potassium by the adenosine triphosphate (ATP)-consuming Na^+/K^+-ATPase, and passive transport by ion channel proteins, for example, Na^+-channel and K^+-channel (Figure18.3).

In this chapter, we focus on the physical properties of membranes as a whole rather than on the function of their components. The interested reader

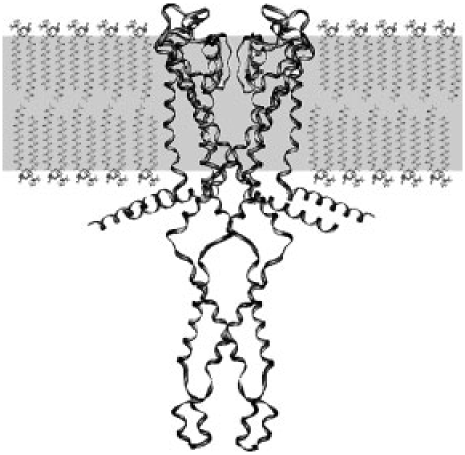

Fig. 18.3 The potassium channel protein spans through the lipid bilayer. Crystal structure of the protein from Cortes, Cuello and Perozo (2001).

may therefore, refer to other excellent publications on membrane proteins.

18.3
Membrane Melting

One of the important features of membranes is that they can melt. Melting processes are order transitions between two planar membrane phases or states that display different physical properties.

18.3.1
Chain Melting

The molecular origin of chain melting is shown in Figure 18.4. At low temperatures, lipid chains are ordered into an all-trans configuration (Figure 18.4, top left). At high temperatures these chains are disordered due to rotations around the C–C bonds within the lipid chains (Figure 18.4, top right). The membranes are in an ordered "gel" phase at low temperature while they are in a disordered "fluid phase" at high temperatures (Figure 18.4, center). The detailed nature of the phases involved in the melting process is described in more detail in Section 18.4.

18.3.1.1 Temperature Dependence and Heat Capacity

In the chain melting process, the molecules absorb heat (enthalpy) and the entropy increases due to the increase

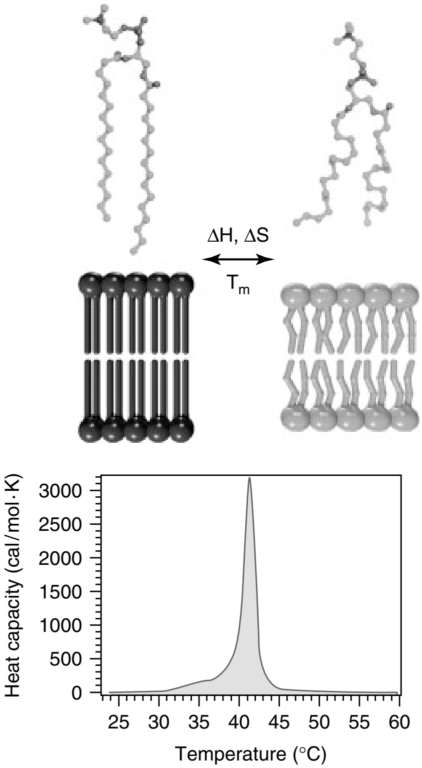

Fig. 18.4 Upon temperature increase, lipids undergo a melting transition where chains become unordered. This melting can be observed as a peak maximum of the heat capacity as measured in calorimetry.

in the number of possible chain configurations. The melting processes is cooperative, meaning that the lipids do not melt independently of each other. The absorption of heat is typically monitored by measuring the heat capacity, that is, the amount of heat absorbed per degree temperature increase. The heat capacity at constant pressure is defined by $c_p = (dH/dT)_p$. It is shown in Figure 18.4 (bottom). At the melting temperature, T_m, it displays a pronounced maximum.

18.3.1.2 **Pressure Dependence**

During the melting process, membranes also change their volume and area (Heimburg, 1998). As a rule of thumb the volume changes by about 4% and the area by about 25%. Owing to these changes, the melting temperature is pressure dependent. If one increases the hydrostatic pressure by about 40 bars, the melting temperature increases by 1°. Similarly, if one increases the lateral tension within the membrane, the melting temperature also changes. The latter effect can be studied on a Langmuir trough, where lipid monolayers are compressed by changing the film area while simultaneously monitoring the lateral pressure (Möhwald, 1990).

18.3.1.3 **Solvent Dependence**

There are other factors that change the melting temperature, for example, the partitioning of organic solvents in membranes. Typically, solvents do not dissolve well in the ordered low-temperature phase but dissolve quite well in the fluid phase. This gives rise to a lowering of the melting temperatures due to a well known effect called *freezing point depression*. As discussed in Section 18.9 this is likely to be related to anesthesia since most anesthetics resemble organic solvents.

18.3.1.4 **Voltage Dependence**

Membranes contain zwitterionic and charged lipids. The zwitterionic lipids have a head group that is an electric dipole. Thus, a monolayer of lipids possesses a strong dipole potential across that layer, which can be of the order of 500 mV (e.g. Pickard, Sehgal and Jackson (1979)). Upon area changes in a monolayer experiment, one finds that the lateral pressure changes in close relationship to this dipole potential. This implies that similar to the change in pressure, voltage changes can induce a transition in the lipid monolayers. Since a membrane consists of two monolayers, these potentials partially compensate. In the presence of charged lipids, the membranes additionally possess a net electrostatic potential that is discussed in Section 18.6.

18.3.1.5 **Dependence on Membrane Proteins**

As already mentioned in Section 18.2.4, many proteins are embedded into or absorbed on membranes. This also has an influence on melting processes (Zhang *et al.*, 1995). Some proteins shift transitions to higher temperatures, for example, the band 3 protein of erythrocytes (Morrow *et al.*, 1986) while others shift it rather to lower temperatures, for example, cytochrome b5 (Freire *et al.*, 1983). For integral proteins, the shift is related to the hydrophobic matching between proteins and lipid bilayer thickness (Mouritsen and Bloom, 1984).

18.3.2
Melting in Biomembranes

Most of the above has been investigated in model membranes or monolayers made of synthetic lipids. However, all of it gains biological relevance due to the fact that biological membranes also possess such melting transitions of their membranes, typically slightly below body temperature. This can be seen in Figure 18.5. Here, heat capacity profiles of native *E. coli* membranes are shown. The two traces are from bacteria grown at two different temperatures, one at 37 °C and the other one at 15 °C, respectively. Numerous peaks can be seen in these traces. The membrane melting peaks are shaded in gray. The other peaks represent the unfolding of proteins at higher temperatures. The membrane melting peak occurs at lower temperature for bacteria grown at 15 °C than for those grown at 37 °C. The bacteria adapt their membrane lipid compositions such that the melting point of the membranes is always below growth temperature. That the lipid composition responds to the changes in the environment has already be described in Section 18.2.

18.4
Phases and Domains

The thermodynamics of ensembles of molecules allows for the formation of quite different phases. If the conditions are right several phases may coexist. This happens in temperature, pressure, or concentration intervals well defined by the thermodynamic parameters of the system. The graphical representation of the coexistence of such phases is called a it phase diagram (cf. Figure 18.7).

The most common phases found in artificial and biological membranes are the solid ordered (SO) phase, the liquid ordered (LO) phase, and the liquid

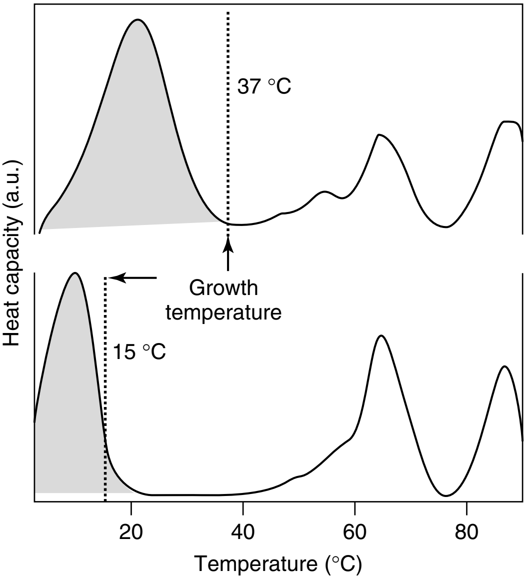

Fig. 18.5 The native cell membranes of *E. coli* bacteria also display chain melting of their membranes at temperature slightly below growth temperature (shaded gray). The lipid membrane melting peaks adapt to the growth temperature (dotted line). The peaks above growth temperature are protein unfolding peaks. Their position does not depend on growth temperature.

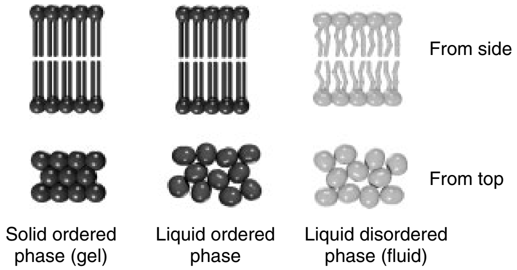

Fig. 18.6 The three most common membrane phases seen as cross sections and from top: solid ordered (SO), liquid ordered (LO) and liquid disordered (LD) phases. The first two phases (SO and LO) are low enthalpy phases, while the LD phase is a high enthalpy phase mostly existing at high temperatures.

disordered (LD) phase (Figure 18.6). "Solid" means that the lipids arrange on a (typically triangular) lattice in the membrane plane while "liquid" indicates the loss of lateral packing (Figure 18.6, bottom). The attributes "ordered" and "disordered" indicate the internal order of the lipid chains (Figure 18.6, top). The transition between ordered and disordered states is linked to the uptake of significant amounts of enthalpy (i.e. heat), as already described in Section 18.3 (Figure 18.4). The SO phase is also often called the *gel phase* and the trivial name of the LD phase is the "fluid phase." In the literature one often finds those trivial names of the phases.

Typically, the different phases differ in their physical properties. For instance, they may display very different elasticities, auto-diffusion constants or electrostatic potential. Under most circumstances, biological membranes seem to exist in the LD phase. Owing to the proximity to melting transitions, however, this can be fine tuned by the organism.

18.4.1
Gibbs Phase Rule

In a lipid mixture like the biological membrane, one may find more than the three phases mentioned above. For instance, one could have two LD phases coexisting.

Let us assume we have K lipid (or protein) components, and P phases. Of course, one has always at least one phase, that is, $P \geq 1$. Further, we introduce the number of degrees of freedom, F, that indicates how many variables can be changed independently of each other. Such variables are the concentrations of the different membrane component and the temperature. The number of degrees of freedom must always be positive, that is, $F \geq 0$.

There is a famous thermodynamic rule about how many different phases can coexist for a given number of components and degrees of freedom that is called *Gibbs phase rule*. At constant pressure it is given by:

$$F = K - P + 1 \quad \text{Gibbs' phase rule} \qquad (1)$$

Let us give a simple example for this. In a binary lipid mixture as described in Figure 18.7 one has two components, that is, $K = 2$. Let us now check all possibilities:

$$\begin{aligned} P = 1 &\rightarrow F = 2 \\ P = 2 &\rightarrow F = 1 \\ P = 3 &\rightarrow F = 0 \end{aligned} \qquad (2)$$

The first case indicates that one has only one phase, and one can vary both the concentration of the second component and

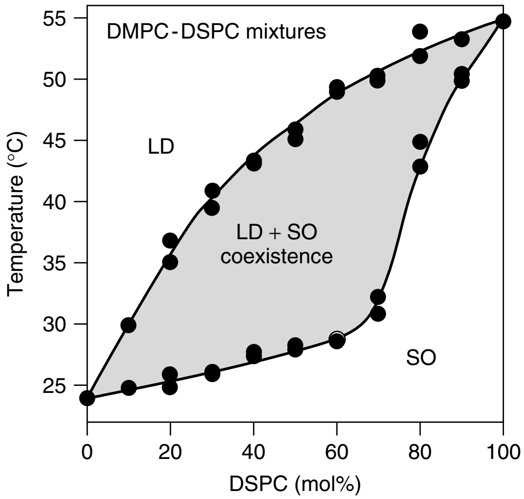

Fig. 18.7 Phase diagram of a simple binary lipid mixture. The gray-shaded region indicates the regime of temperature and concentration where a LD (gel) and a LD (fluid) phase coexist. Adapted from Hac *et al.* (2005).

the temperature freely without leaving the one-phase region. The second case indicates that one may have two different phases coexisting. However, when one of the variables (e.g. temperature) is varied, the second variable (concentration of the components in the two phases) is automatically determined and can not be varied freely. The last case in Eq. (2) indicates that there may be points (with both defined temperature and composition) where three phases may coexist.

18.4.2
Phase Diagrams

An example for a simple phase diagram is shown in Figure18.7. It represents the mixture of two synthetic lipids dimyristoyl phosphatidylcholine (DMPC), and distearoyl phosphatidylcholine (DSPC) in aqueous environment. The two variables in this system are temperature and the fraction of DSPC. These two variables represent the two axes. In the gray shaded region SO and LD phases coexist. At a given temperature, the composition of the two phases are given by the lines in the phase diagram (for details see Lee (1977)). Above the upper line (liquidus line), the whole membrane is in the LD (fluid) phase, while it is in the SO (gel) phase below the lower line (solidus line).

Biological membranes, however, consist of hundreds of components, which are the different lipids, the membrane proteins, and the many small molecule components. What is immediately clear from the above is that in a biological membrane with several hundred components, one may also have many different phases coexisting. For $K = 100$ one may in principle find up to 101 coexisting phases ($P = 101$). Thus, the thermodynamic behavior of such membranes may be very complex, and it may be practically impossible to unravel the detailed behavior of such complex mixtures. Nevertheless, clearly such phase coexistence is present in biological membranes and must strongly influence their behavior. Therefore, to make general statements on the physical behavior of biomembranes, one may wish to find general relations that do not require the detailed knowledge of the phase diagram. Such relations exist and will be outlined in Section 18.5.

18.4.3
Domain Formation

The concept of a phase is a strictly theoretical one assuming, for instance, that the system under consideration is infinitely big, and that the phases are separated from each other (i.e. their interface can be neglected). Both assumptions are not quite true for biological membranes. Cells are relatively small (on the order of 10^9 lipids) and phases cannot be macroscopically separated. Instead, one typically obtains finite size domains that interact at their boundaries. What can be concluded from the above is that both artificial membranes are made of lipid mixtures, and biological membranes must display the coexistence of many domains of different physical properties.

18.4.3.1 Domain Formation in Artificial Membranes and Monolayers

Domain formation in artificial systems is actually quite easy to observe. Already in the early 1980s, domains were described in monolayer films (e.g. Weis and McConnell (1984); Möhwald (1990)). The first demonstration of domain formation on bilayer membranes was from Korlach *et al.* (1999). In Figure 18.8 (left) we show a confocal fluorescence microscopy image of a giant vesicle (Ø ≈30 μm) made of a lipid mixture of dilauroyl phosphatidylcholine (DLPC) and dipalmitoyl phosphatidylcholine (DPPC) measured at room temperature. According to the phase diagram of this lipid mixture it should display phase coexistence at room temperature. This can be seen as dark and light grey domains that have been labeled with two different fluorescence markers with affinity for the two different phases. Dark grey domains are in the SO phase, while the light grey regions are LD phases. Such domains are obviously rather big, that is, of order of 5–10 μm.

18.4.3.2 Domain Formation in Biological Systems

Domain formation in biological membranes is much more difficult to monitor. In the 1990s, several authors proposed the existence of domains called *rafts* (Simons and Ikonen, 1997; Brown and London, 1998; London, 2002; Edidin, 2003b; Edidin, 2003a). They are thought to be rich in cholesterol, sphingolipids, and some GPI-anchored proteins that are in the LO phase (Figure 18.9). The physical evidence for such domains in real biomembranes is relatively weak. If they exist, they seem to be smaller than the resolution of a light microscope (<200 nm). Further, owing to the many

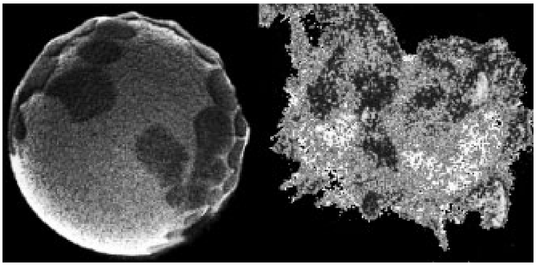

Fig. 18.8 Domain formation. Left: On a vesicle made of a lipid mixture of DLPC and DPPC (courtesy M. Fidorra, NBI). Right: On a living fibroblast cell (adapted from Gaus *et al.* (2003)). The dimension of both object is about 30 μm.

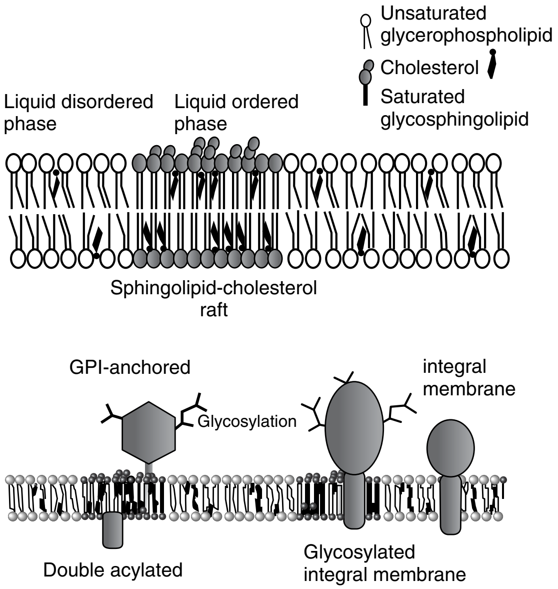
Unsaturated glycerophospholipid
Cholesterol
Saturated glycosphingolipid
Liquid disordered phase
Liquid ordered phase
Sphingolipid-cholesterol raft
GPI-anchored
Glycosylation
integral membrane
Double acylated
Glycosylated integral membrane

Fig. 18.9 The formation of cholesterol-rich rafts with GPI (glycosylphospat-idylinositol)-anchored proteins as envisioned by Bagnat and Simons (2002).

different components of biological membranes there is no need to constrain the concept of domain formation to particular domain compositions. There may be domains of all kinds of compositions, for example, protein clusters (already shown in the famous paper by Singer and Nicolson (1972)), but also SO and LD domains. One of the nicest demonstrations of domains in cells to date is from Gaus *et al.* (2003) in fibroblast cells shown in Figure 18.8 (right).

18.5
Curvature and Elastic Constants

Membranes are typically not flat rigid sheets but often display curvature when they form vesicles and organelles. They are soft and display both fluctuations in area and curvature. Further, when membranes are asymmetric they may curve spontaneously.

18.5.1
Principle Curvatures

Bending can be characterized by its radius of curvature, R, or its curvature, $c = 1/R$ (Figure 18.10). Since the surface has two dimensions one typically deals with two principle curvatures, c_1 and c_2, that may be different.

To curve the membrane one has to provide free energy. The free energy density of a membrane, g, is given by a famous formula by Helfrich (1973):

$$g = \frac{1}{2} K_B \left(c_1 + c_2 - c_0\right)^2 + K_G c_1 c_2 \quad (3)$$

This equation contains three material constants that have to be determined

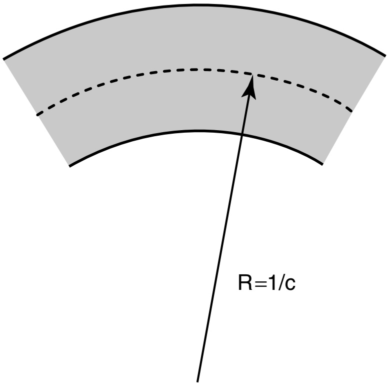

Fig. 18.10 Curvature and radius of curvature for a bend membrane.

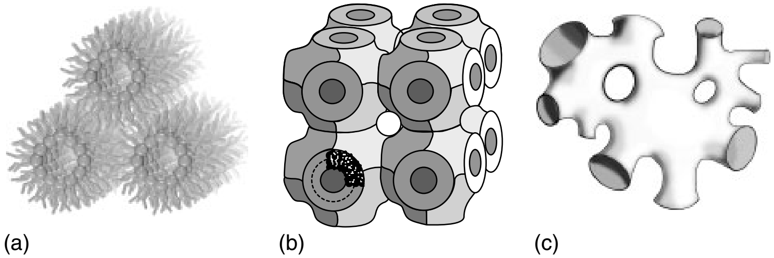

Fig. 18.11 Lipid phases with different curvatures. (a) Inverse hexagonal phases (from Heimburg (2007)), (b) Cubic phase (adapted from Lindblom and Rilfors (1989)). (c) Sponge phase (from Heimburg (2007)).

experimentally: K_B is the bending modulus, K_G is the Gaussian curvature modulus, and c_0 is the spontaneous curvature. The spontaneous curvature indicates the equilibrium curvature of the membrane. It mainly depends on the asymmetry of the membrane, for example, different lipid and protein compositions in the two opposing monolayers of the membrane, but also on electrostatic potential differences, pH differences, and so on.

18.5.2
Curved Phases

There exist various curved lipid phases.

The inverse hexagonal phase (Figure 18.11a) is not a membrane phase. It is formed due to the spontaneous curvature of the monolayers. There has been quite some discussion on whether the potency of biological membranes to form such phases may be linked to biological function (Toombes *et al.*, 2002).

The cubic phase (Figure 18.11c) resembles a periodic three-dimensional lattice of vesicles that are connected. It is a bicontinuous bilayer structure with no inside and outside. It displays saddle-point structures. These are the geometries where the two principle curvatures, c_1 and c_2 have opposite sign.

The sponge phase (Figure 18.11c) has the same topology compared to the cubic phase, but it does not display any periodic regularity.

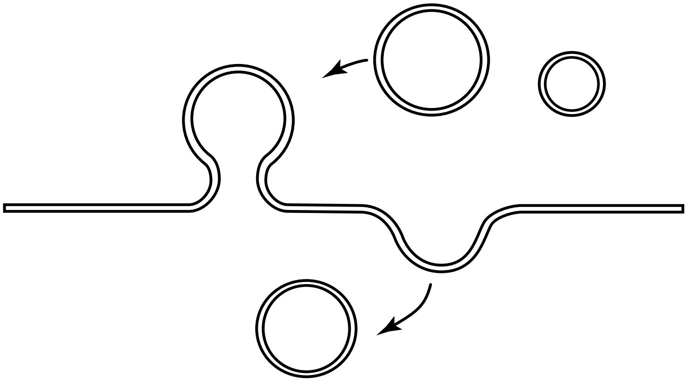

Fig. 18.12 Fusion and fission events as in endocytosis and exocytosis (adapted from Heimburg (2003)) that involve changes in local curvature.

Figure 18.12 shows the biological processes of endocytosis and exocytosis, that is, fusion and fission of vesicles and membranes. They obviously display a very similar saddle-point geometry at the site of fusion as the cubic and the sponge phases. This triggered the interest in such membrane phases (Lohner, 1996; Toombes *et al.*, 2002).

18.5.3
Elastic Constants

The volume compressibility is the function that indicates how much the volume changes when the pressure is increased by an incremental amount. Very compressible substances display a large change in volume when the pressure increases. The area compressibility is the respective function that indicates how much the area changes upon changes in lateral pressure or surface tension.

The curvature elasticity is the inverse of the curvature modulus as defined in Eq. (3). (Evans, 1974) has shown that the curvature elasticity is linearly related to the lateral compressibility.

18.5.4
Temperature and Pressure Dependence of Elastic Constants

The elastic constants are actually not real constants. They themselves depend in a sensitive manner on the thermodynamic variables, for example, on temperature, pressure, and so on. In particular, it has been shown by Heimburg (1998) and Ebel, Grabitz and Heimburg (2001) that all the elastic constants depend in a simple manner on the heat capacity of the membrane that assumes high values in the transitions of membranes (see Figures 18.4 and 18.5). This means that within transitions, membranes are very elastic. They are soft, easy to bend, and easy to compress. Since biological membranes display such transitions close to body or growth temperature, this is of quite high significance. Everything that is related to the elastic constants will respond very sensitively to changes in temperature and other variables like pH, calcium content, and so on. This includes membrane fusion events as well as the spontaneous formation of pores in the membrane as shown in Section 18.7. We further show in Section 18.8 that close to such transitions one obtains the possibility to propagate stable pulses along cylindrical membranes.

18.6
Electrostatic Potential

On average, about 10% of all lipids are negatively charged. As mentioned in Section 18.2, the relative amount of negative charge is very different depending on the particular organelle of a cell. Especially charged are the membranes of mitochondria that contain up to 20% cardiolipin that have two negative charges each. Further, the charges often distribute asymmetrically such that the inner membranes of cells and organelles are more charged (Rothman and Lenard, 1977; Rothman and Kennedy, 1977).

18.6.1
Ionic Strength Dependence

The electrostatic theory of surfaces originates from Gouy and Chapman (see, e.g. Jähnig (1976)). It describes how the electrostatic potential on surfaces depends on the charge density and on the ionic strength of the salt solution. For membranes one finds that the electrostatic potential decays fast as a function of the distance from the membrane. The characteristic decay length is called the *Debye length*. For a 100 mM solution of NaCl it is 0.97 nm. For 10 mM NaCl it is about 3.1 nm and for 1 mM it is 9.7 nm. Close to the membrane, the Na^+ concentration

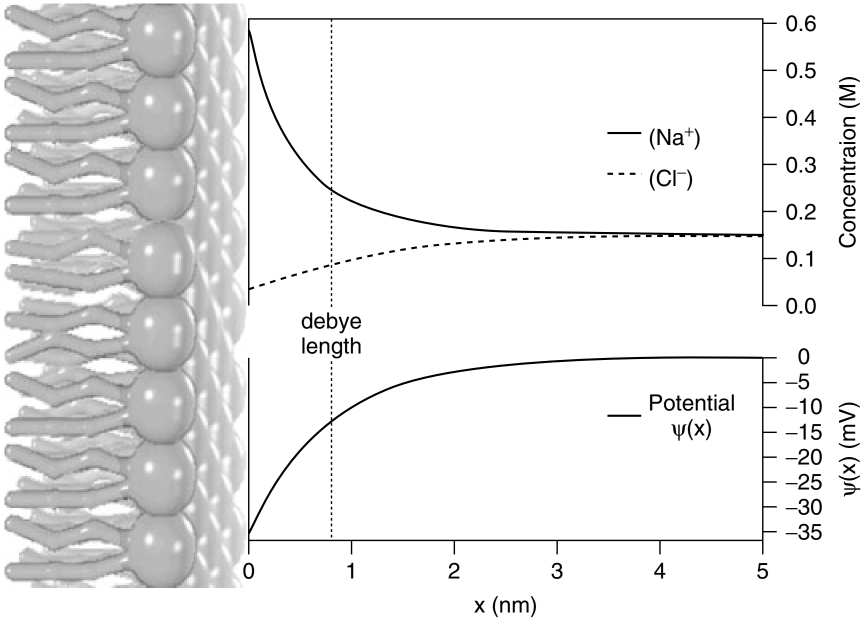

Fig. 18.13 Distribution of positive (Na^+) and negative ions (Cl^-), and magnitude of the electrostatic potential, Ψ, as a function of distance from a membrane containing acidic lipids, calculated for a NaCl concentration of 150 mM. From Heimburg (2007).

is increased, while the Cl^- concentration is reduced (Figure 18.13).

18.6.2
Dependence on Charged Lipid Concentration

Naturally, the electrostatic potential at a membrane surface also depends on the number of charged molecules the form the membrane. For the ionic strength conditions that are normally found in biology, the surface potential is just proportional to the charge density on the surface and inversely proportional to the square root of the ionic strength.

18.6.3
Dependence on Bound Proteins

Proteins also carry charges. Many soluble proteins are positively charged and therefore interact with negatively charged lipid membranes. One example is cytochrome c, a protein from the respiratory chain in mitochondria, with a net positive charge of about 4. It binds to the highly charged mitochondrial membranes and reduces the overall electrostatic potential.

18.7
Permeability

Many biology textbooks state that lipid membranes are impermeable to water, ions and small molecules. Although this seems intuitive because the inner core of a membrane is made of hydrocarbons, this is not quite correct. The lipid layers are very thin and due to fluctuations in molecular there is always some finite likelihood that spontaneous pores are formed. This likelihood is dramatically increased in the phase transition temperature regime of lipid membranes, that is, slightly below body temperature for biomembranes (Papahadjopoulos *et al.*, 1973); (Nagle and Scott, 1978). Figure 18.14 shows the permeability of artificial membranes for a

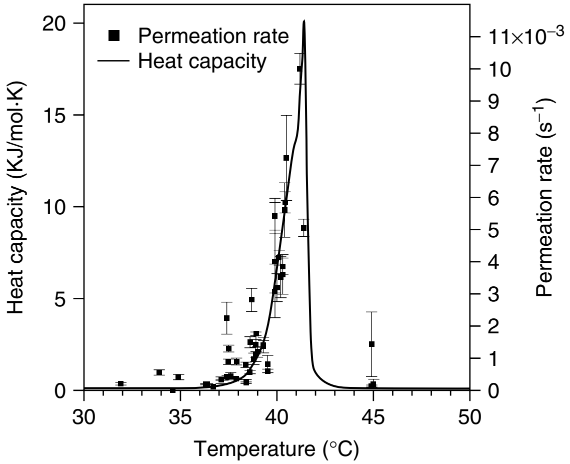

Fig. 18.14 Permeability of a lipid vesicles for a fluorescent dye as a function of temperature compared to the melting profile (Blicher *et al.*, 2009).

fluorescence dye as a function of temperature compared to the heat capacity. One can recognize that the permeability is closely related to the heat capacity.

18.7.1
Protein Ion Channels

Since the permeability of lipid membranes is thought to be small, most of the permeation events through membranes have been attributed to proteins that form pores or ion channels. In patch clamp or black lipid membrane (BLM) measurements on finds quantized ion currents through the membrane indicated localized events. Figure 18.15b shows the first of such traces recorded in the presence of the acetylcholine receptor protein published by Neher and Sakmann (1976). Both scientists received the Nobel prize in 1991. The currents occur in discrete steps that can be seen in the histogram at the right hand side. For a detailed description of ion channel proteins see Hille (1992).

18.7.2
Lipid Ion Channels

As already seen in Figure 18.14 the lipid membranes are also permeable to molecules if they are close to their phase

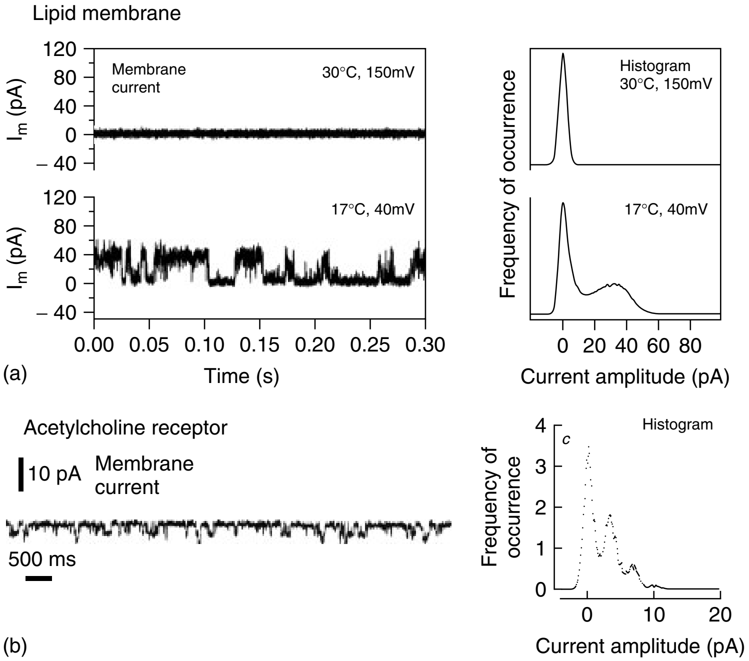

Fig. 18.15 *Microscopic permeability of membranes in the presence and absence of proteins. (a) Spontaneous ion currents through a synthetic black lipid membrane at its phase transition (17°C). Above its transition at 30°C the currents disappear (Fidorra, 2007).* (b) Spontaneous ion currents through a biological membrane in the presence of the acetylcholine receptor adapted from the original paper of Neher and Sakmann (1976).

transition temperature. In patch clamp or BLM measurements one finds quantized ion currents through the membrane. This can be seen in Figure 18.15a for a lipid mixture with transition maximum at 17 °C. The currents at the transition temperature resemble those seen in the biological membrane describe above. At 30 °C the currents disappear. This effect is not as well known as the protein conductance but has been described by a number of authors in the literature (e.g. Antonov *et al.*, 1980; Kaufmann and Silman, 1983b; Kaufmann and Silman, 1983a; Gögelein and Koepsell, 1984; Antonov *et al.*, 2005; Blicher *et al.*, 2009). Protein ion channels are always embedded in membranes. Biological membranes are close to their transition temperature where they can also show quantized channels due to events in the lipid membrane. Therefore, it does not seem straight-forward how to distinguish between the two events.

18.8
Pulse Propagation

Nerve cells form long membrane cylinders that are called *dendrites* or *axons*. Along these cables one finds the propagation of voltage pulses that are called *action potentials*. However, mechanical pulses can also travel along nerve membranes. It turns out that these two seemingly different events may well be caused by one single phenomenon.

18.8.1
Nerve Pulses

Action potentials have been described by Hodgkin and Huxley (1952) (Nobel prize 1963). Their theory is based on the existence of voltage-gated ion channels (in particular of sodium and potassium channels, see Figure 18.3). It is an electrical theory in which the lipid membrane serves as a capacitor and the ion channel proteins are electrical resistors. The resulting differential equation has the shape of a propagating voltage pulse very similar to that observe in real nerves. Since it is a purely electrical model, it does not contain and describe any changes in other variables, for example, heat, volume, thickness, and area.

18.8.2
Soliton Propagation

Biological membranes are slightly above their transition temperature where both the elastic constants and the permeability change drastically. (Heimburg and Jackson, 2005) showed that under these conditions stable mechanical pulses (called *solitons*) can travel along membranes. Since membranes possess both dipole potentials and electrostatic potentials, these mechanical pulses are accompanied by changes in voltage. The properties of such pulses are based on reversible physics. Area, thickness, and voltage of the membranes change during the pulses. Since the events are reversible, the process is isentropic, that is, no heat is dissipated during the pulse. During the pulse, all variables change that are associated to the phase transition in membranes. The free energy necessary to generate a pulse is given by its distance to the phase transitilidocaineon.

18.8.3
Reversible Heat

A problem for the Hodgkin–Huxley model is the experimental fact that no heat is dissipated during the nerve pulse (Abbott, Hill and Howarth, 1958; Howarth, Keynes and Ritchie, 1968; Ritchie and Keynes, 1985). Since currents flow through resistors (the ion channels), heat production would be expected during the action potential. Instead, one first finds a phase of heat liberation followed by a phase of heat reabsorption. This finding is rather in agreement with a reversible electromechanical pulse as described by the soliton theory.

18.8.4
Mechanical Changes

Nerves change their dimensions during the action potential. Tasaki and coworkers found that during the action potential the nerves change their thickness by about 1 nm (e.g. Iwasa and Tasaki (1980); Iwasa, Tasaki and Gibbons (1980); Tasaki and Byrne (1990) and that nerves also get shorter upon excitation (Tasaki, Kusano and Byrne, 1989)). Thus, in addition to the electrical phenomena mechanical changes also exist during nerve pulses.

18.9
Anesthesia

Anesthesia is a phenomenon caused by a number of small molecules that partition in the biological membrane. For more than 100 years, it is known that the effectiveness of anesthetics is proportional to their solubility in olive oil (that has the properties of the membrane interior). This rule is known as the *Meyer–Overton rule* (Overton, 1901). It holds over several orders of magnitude ranging from laughing gas, N_2O, and over halothane to lidocaine. Even the noble gas xenon is an anesthetic. This excludes any specific binding to macromolecules (e.g. proteins) if one is searching for a generic explanation of anesthesia. It has also been known for a long time that anesthetics cause a lowering of phase transition temperatures (Trudell *et al.*, 1975; Kharakoz, 2001; Heimburg and Jackson, 2007b). This has already been discussed in Section 18.3.1.3. There are strong indications that the effect of anesthetics is related to this finding. As shown in Section 18.3.1.2, phase transitions are pressure dependent. While pressure increases transitions, anesthetics lower them. It has in fact been found that pressure reverses the effect of anesthesia (Johnson and Flagler, 1950). In the context of the soliton theory (Section 18.8.2), anesthetics increase the free energy necessary to generate an electromechanical pulse (Heimburg and Jackson, 2007a; Heimburg and Jackson, 2007b). There is no obvious connection between the Hodgkin–Huxley model and anesthesia, but there are attempts to relate the action of anesthesia to specific binding to in channel proteins.

References

Abbott, B.C., Hill, A.V., and Howarth, J.V. (1958) The positive and negative heat production associated with a nerve impulse. *Proc. R. Soc. London, Ser. B*, **148**, 149–87.

Antonov, V.F., Anosov, A.A., Norik, V.P., and Smirnova, E.Y. (2005) Soft perforation of planar bilayer lipid membranes of dipalmitoylphosphatidylcholine at the temperature of the phase transition from the liquid crystalline to gel state. *Eur. Biophys. J.*, **34**, 155–62.

Antonov, V.F., Petrov, V.V., Molnar, A.A., Predvoditelev, D.A., and Ivanov, A.S. (1980) The appearance of single-ion channels in unmodified lipid bilayer membranes at the phase transition temperature. *Nature*, **283**, 585–86.

Bagnat, M. and Simons, K. (2002) Lipid rafts in protein sorting and cell polarity in budding yeast saccharomyces cerevisiae. *Biol. Chem.*, **383**, 1475–80.

Blicher, A. Wodzinska, K. Fidorra, M. Winterhalter, M. and Heimburg, T. (2009) The temperature dependence of lipid membrane Permeability, its quantized nature, and the influence of anesthetics. *Biophys. J.*, in print (arXiv:0807.4825v1 [physics.bio-ph]).

Brown, D.A. and London, E. (1998) Functions of lipid rafts in biological membranes. *Annu. Rev. Cell Dev. Biol.*, **14**, 111–36.

Cortes, D.M., Cuello, L.G., and Perozo, E. (2001) Molecular architecture of full-length KCSA: role of cytoplasmic domains in ion permeation and activation gating. *J. Gen. Physiol.*, **117**, 165.

Danielli, J.F. and Davson, H. (1935) A contribution to the theory of permeability of thin films. *J. Cell. Comp. Physiol.*, **5**, 495–508.

DeLong, E.F. and Yayanos, A.A. (1985) Adaptation of the membrane lipids of a deep-sea bacterium to changes in hydrostatic pressure. *Science*, **228**, 1101–03.

Ebel, H., Grabitz, P., and Heimburg, T. (2001) Enthalpy and volume changes in lipid membranes. i. the proportionality of heat and volume changes in the lipid melting transition and its implication for the elastic constants. *J. Phys. Chem. B*, **105**, 7353–60.

Edidin, M. (2003a) Lipids on the frontier: A century of cell-membrane bilayers. *Nat. Rev. Mol. Cell Biol.*, **4**, 414–18.

Edidin, M. (2003b) The state of lipid rafts: from model membranes to cells. *Annu. Rev. Biophys. Biomol. Struct.*, **32**, 257–83.

Evans, E.A. (1974) Bending resistance and chemically induced moments in membrane bilayers. *Biophys. J.*, **14**, 923–31.

Fidorra, M. (2007) Confocal Microscopy, Calorimetry and Permeability Studies on Giant Lipid Vesicles Containing Ceramides. Ph.D. thesis, University of Southern Denmark.

Freire, E., Markello, T., Rigell, C., and Holloway, P.W. (1983) Calorimetric and fluorescence characterization of interactions between cytochrome b_5 and phosphatidylcholine bilayers. *Biochemistry*, **28**, 5634–43.

Gaus, K., Gratton, E., Kable, E.P.W., Jones, A.S., Gelissen, I., Kritharides, L., and Jessup, W. (2003) Visualizing lipid structure and raft domains in living cells with two-photon microscopy. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 15554–59.

Gögelein, H. and Koepsell, H. (1984) Channels in planar bilayers made from commercially available lipids. *Pfluegers Arch.*, **401**, 433–34.

Gorter, E. and Grendel, F. (1925) On bimolecular layers of lipid on the chromacytes of the blood. *J. Exp. Med.*, **41**, 439–43.

Hac, A., Seeger, H., Fidorra, M., and Heimburg, T. (2005) Diffusion in two-component lipid membranes – a fluorescence correlation spectroscopy and monte carlo simulation study. *Biophys. J.*, **88**, 317–33.

Hazel, J.R. (1979) Influence of thermal acclimation on membrane lipid composition of rainbow trout liver. *Am. J. Physiol. Regul. Integr. Comp. Physiol.*, **287**, R633–41.

Heimburg, T. (1998) Mechanical aspects of membrane thermodynamics. Estimation of the mechanical properties of lipid membranes close to the chain melting transition from calorimetry. *Biochim. Biophys. Acta*, **1415**, 147–62.

Heimburg, T. (2003) Coupling of chain melting and bilayer structure: domains, rafts, elasticity and fusion, in *Planar Lipid Bilayers (BLMs) and their Applications*, (eds H.T. Tien and A. Ottova-Leitmannova), Elsevier, Amsterdam, 269–93.

Heimburg, T. (2007) *Thermal Biophysics of Membranes*, Wiley VCH, Berlin, Germany.

Heimburg, T. and Jackson, A.D. (2005) On soliton propagation in biomembranes and nerves. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 9790–95.

Heimburg, T. and Jackson, A.D. (2007a) On the action potential as a propagating density pulse and the role of anesthetics. *Biophys. Rev. Lett.*, **2**, 57–78.

Heimburg, T. and Jackson, A.D. (2007b) The thermodynamics of general anesthesia. *Biophys. J.*, **92**, 3159–65.

Helfrich, W. (1973) Elastic properties of lipid bilayers: theory and possible experiments. *Z. Naturforsch.*, **28c**, 693–703.

Hille, B. (1992) *Ionic Channels of Excitable Membranes*, Cambridge University Press, Cambridge.

Hodgkin, A.L. and Huxley, A.F. (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, **117**, 500–44.

Howarth, J.V., Keynes, R., and Ritchie, J.M. (1968) The origin of the initial heat associated with a single impulse in mammalian non-myelinated nerve fibres. *J. Physiol.*, **194**, 745–93.

Ingram, L.O. (1977) Changes in lipid composition of escherichia coli resulting from growth with organic solvents and with food additives. *Appl. Environ. Microbiol.*, **33**, 1233–36.

Iwasa, K. and Tasaki, I. (1980) Mechanical changes in squid giant-axons associated with production of action potentials. *Biochem. Biophys. Res. Commun.*, **95**, 1328–31.

Iwasa, K., Tasaki, I., and Gibbons, R.C. (1980) Swelling of nerve fibres associated with action potentials. *Science*, **210**, 338–39.

Jähnig, F. (1976) Electrostatic free energy and shift of the phase transition for charged lipid membranes. *Biophys. Chem.*, **4**, 309–18.

Jamieson, G.A. and Robinson, D.M. (1977) *Mammalian Cell Membranes*, Vol.2, Butterworth, London.

Johnson, F.H. and Flagler, E.A. (1950) Hydrostatic pressure reversal of narcosis in tadpoles. *Science*, **112**, 91–92.

Kaufmann, K. and Silman, I. (1983a) The induction by protons of ion channels through lipid bilayer membranes. *Biophys. Chem.*, **18**, 89–99.

Kaufmann, K. and Silman, I. (1983b) Proton-induced ion channels through lipid bilayer membranes. *Naturwissenschaften*, **70**, 147–49.

Kharakoz, D.P. (2001) Phase-transition-driven synaptic exocytosis: a hypothesis and its physiological and evolutionary implications. *Biosci. Rep.*, **210**, 801–30.

Korlach, J.J., Schwille, P., Webb, W.W., and Feigenson, G.W. (1999) Characterization of lipid bilayer phases by confocal microscopy and fluorescence correlation spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.*, **96**, 8461–66.

Lee, A.G. (1977) Local anesthesia: the interaction between phospholipids and chlorpromazine, propranolol, and practolol. *Mol. Pharmacol.*, **13**, 474–87.

Lindblom, G. and Rilfors, L. (1989) Cubic phases and isotropic structures formed by membrane lipids - possible biological relevance. *Biochim. Biophys. Acta*, **988**, 221–56.

Lohner, K. (1996) Is the high propensity of ethanolamine plasmalogens to form non-lamellar lipid structures manifested in the properties of biomembranes? *Chem. Phys. Lipids*, **81**, 167–84.

London, E. (2002) Insights into lipid raft structure and formation from experiments in model membranes. *Curr. Opin. Struct. Biol.*, **12**, 480–86.

Möhwald, H. (1990) Phospholipid and phospholipid-protein monolayers at the air/water interface. *Annu. Rev. Phys. Chem.*, **41**, 441–76.

Morrow, M.R., Davis, J.H., Sharom, F.J., and Lamb, M.P. (1986) Studies of the interaction of human erythroyte band 3 with membrane lipids using deuterium nuclear magnetic resonance and differential scanning calorimetry. *Biochim. Biophys. Acta*, **858**, 13–20.

Mouritsen, O.G. and Bloom, M. (1984) Mattress model of lipid-protein interactions in membranes. *Biophys. J.*, **46**, 141–53.

Nagle, J.F. and Scott, H.L. (1978) Lateral compressibility of lipid mono- and bilayers lateral compressibility of lipid mono- and bilayers. theory of membrane permeability. *Biochim. Biophys. Acta*, **513**, 236–43.

Neher, E. and Sakmann, B. (1976) Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature*, **260**, 779–802.

Overton, C.E. (1901) Jena, Germany. English translation: studies of Narcosis, Chapman and Hall, 1991, in *Studien über die Narkose*, (ed R. Lipnick), Verlag Gustav Fischer.

Papahadjopoulos, D., Jacobson, K., Nir, S., and Isac, T. (1973) Phase transitions in phospholipid vesicles. fluorescence polarization and permeability measurements concerning the effect of temperature and cholesterol. *Biochim. Biophys. Acta*, **311**, 330–40.

Pickard, W.F., Sehgal, K.C., and Jackson, C.M. (1979) Measurement of phospholipid monolayer surface-potentials at a hydrocarbon-electrolyte interface. *Biochim. Biophys. Acta*, **552**, 1–10.

Ritchie, J.M. and Keynes, R.D. (1985) The production and absorption of heat associated with electrical activity in nerve and electric organ. *Q. Rev. Biophys.*, **392**, 451–76.

Rothman, J.E. and Kennedy, E.P. (1977) Asymmetrical distribution of phospholipids in the membrane of bacillus megaterium. *J. Mol. Biol.*, **110**, 603–18.

Rothman, J.E. and Lenard, J. (1977) Membrane asymmetry. *Science*, **195**, 743–53.

Sackmann, E. (1995) Physical basis of self-organization and function of membranes: physics of vesicles, in *Structure and Dynamics of Membranes: From Cells to Vesicles*, (eds R. Lipowsky and E. Sackmann), Elsevier, Amsterdam, 213–304.

Simons, K. and Ikonen, E. (1997) Functional rafts in cell membranes. *Nature*, **387**, 569–72.

Singer, S.J. and Nicolson, G.L. (1972) The fluid mosaic model. *Science*, **175**, 720–31.

Tasaki, I. and Byrne, M. (1990) Volume expansion of nonmyelinated nerve fibers during impulse conduction. *Biophys. J.*, **57**, 633–35.

Tasaki, I., Kusano, K., and Byrne, M. (1989) Rapid mechanical and thermal changes in the garfish olfactory nerve associated with a propagated impulse. *Biophys. J.*, **55**, 1033–40.

Toombes, G.E.S., Finnefrock, A.C., Tate, M.W., and Gruner, S.M. (2002) Determination of l(alpha)-h(ii) phase transition temperature for 1,2-dioleoyl-sn-glycero-3-phosphatidylethanolamine. *Biophys. J.*, **82**, 2504–10.

Trudell, J.R., Payan, D.G., Chin, J.H., and Cohen, E.N. (1975) The antagonistic effect of an inhalation anesthetic and high pressure on the phase diagram pf mixed dipalmitoyl-dimyristoylphosphatidylcholine bilayers. *Proc. Natl. Acad. Sci. U. S. A.*, **72**, 210–13.

van de Vossenberg, J.L.C.M., Driessen, A.J.M., da Costa, M.S., and Konings, W.N. (1999) Homeostasis of the membrane proton permeability in bacillus subtilis grown at different temperatures. *Biochim. Biophys. Acta*, **1419**, 97–104.

Weis, R. and McConnell, H. (1984) Two-dimensional chiral crystals of phospholipids. *Nature*, **310**, 47–49.

Zhang, Y.-P., Lewis, R.N.A.H., Hodges, R.S., and McElhaney, R.N. (1995) Peptide models of helical hydrophobic transmembrane segments of membrane proteins. 2. Differential scanning calorimetry and FTIR spectroscopc studies of the interaction of Ac-Kmacrosrightcurlybrace2-(LA)12-K2-amide with phosphatidylcholine membranes. *Biochemistry*, **34**, 2362–71.

19
Chiral Raman and Single Molecule Spectroscopy

Salim Abdali and Ewan W. Blanch

19.1 **Introduction 621**
19.1.1 Overview of ROA 622
19.1.2 Overview of SERS 623
19.1.3 Overview of SEROA 624
19.2 **Theory of SEROA 624**
19.3 **ROA Instrumentation 628**
19.4 **SEROA, Measurement Results 629**
19.4.1 SEROA of Small Molecules 629
19.4.2. SEROA of Protein 632
19.5 **Future Perspectives 639**
Acknowledgments 639
References 639

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.

Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

ISBN: 978-3-527-40702-6

19.1
Introduction

Single molecule detection and chiral spectroscopy are two powerful tools within Raman spectroscopy (RS), and they have been combined into one tool in order to provide chiral information obtained at the single molecule level. This had first been predicted by (Efrima, 1983, 1985); however, the lack of an experimental evidence has until recently been dominant in the field. It is worth mentioning that the experimental results, which have been published and those which are discussed here, cannot yet be considered as results of single molecule measurements, in that they provide knowledge about the average for a number of molecules that contribute to the measured signal. A single molecule measurement should naturally concern a detection of 1.6×10^{-24} mol of an analyte. This requires a very small interaction cross section, an optical property that defines the ratio of the intensity of the incident to the scattered light. So far fluorescence emission has been the most utilized tool for single molecule detection (Morner, 2006), where a typical cross section of a chromophore can be 10^{-16} cm^2/molecule, a limit that is known to define single molecule detection. Raman scattering (RS), an alternative tool, which has the capability of providing more information than fluorescence, especially at room temperature, could not be used when information on a single molecule is desired. This is because the scattering cross section of RS is about 10^{-30} cm^2/molecule. However, during the last three decades and after the discovery of another Raman tool, known as *surface-enhanced Raman scattering* (*SERS*), this has become an alternative to fluorescence, especially when the resonance effect, known as *surface-enhanced resonance Raman scattering* (*SERRS*), takes place at the same time. In such conditions, the cross section can approach the limit for detecting a single molecule (Nie and Emory, 1997). In general, there are relatively few reports on single molecule detection, including fluorescence, and in these reports one finds that the dye rhodamine is the most common sample from which the results were reported, with only very few reports on biomolecules, for example, porphyrin (Bizzari and Cannistrato, 2002), enkephalin (Kneipp *et al.*, 2004) and adenine (Koo *et al.*, 2004).

Another Raman tool, which is sensitive to molecular chirality, is Raman optical activity (ROA) from which

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

knowledge about the molecule structure and stereochemistry can be obtained. However, the ROA signal intensity is very weak, and its combination with a single molecule technique is desirable. In the following sections, ROA, SERS, and the attempts at combining them, both theoretically and experimentally, have been discussed.

19.1.1
Overview of ROA

Chiral Raman spectroscopy, known as *Raman optical activity (ROA)*, is a technique that has mostly been employed for molecular studies owing to the fact that ROA spectra provide information about the studied molecule in solution, thus, it is extremely useful for bioanalysis in native environments. Similar to the Raman effect, ROA is a result of interaction between monochromatic electromagnetic (EM) radiation, usually in the visible region of the EM spectrum, and a biomolecule (Barron, 2004). However, it differs from Raman spectroscopy as circular polarized light is used, and the difference in intensity of scattering of the right- and left scattered light represents the ROA. Therefore, ROA can provide information on the molecular optical activity.

The optical activity effect has been known for a long time now, being first observed by Dominique Arago (1811); however, the discovery of stereoisomers and the formulation of chirality were introduced by Lord Kelvin (1904) as follows: *"I call any geometrical figure or group of points chiral, and say that it has chirality if its image in a plane mirror, ideally realized, cannot be brought into coincidence with itself."*

Using Raman spectroscopy the first ROA theory was formulized in 1969, in which a general theory of the polarization characteristics of Rayleigh, the elastic scattering, and Raman scattered light from chiral molecules based on a new interference mechanism between light waves scattered via the polarizability and optical activity property tensors was presented by Atkins and Barron (1969). In this work they were able to predict that the small difference in the scattered intensity in the right- and left-circular polarized incident light can be detected, and it hence represents a fingerprint of a chiral molecule. Barron and Buckingham succeeded in 1971 in incorporating the contribution of the electric dipole-magnetic dipole optical activity tensor to Raman scattering (Barron and Buckingham, 1971), and they defined in their work the experimental parameter of the dimensionless circular intensity difference (CID, or Δ) as follows:

$$\Delta = \frac{(I^R - I^L)}{(I^R + I^L)} \tag{1}$$

where I^R and I^L are the scattered intensities in right- and left-circular polarized incident light, respectively. As Eq. (1) indicates, the CID can be very small, and in fact this intensity is three to four orders of magnitude lower than the conventional Raman scattered intensity, equal to $(I^R + I^L)$, for the same molecule. Therefore, better tools, including stable monochromatic light sources, high efficiency filters to define the monochromatic wavelength, optical components for ensuring the circular polarization (CP) of the beam, and high sensitive detectors, were all required.

The technical requirements have made measurements difficult, at least until recently, and limited the use of ROA to a handful of laboratories worldwide, with the first reliable spectrum being published in 1973 (Barron, Bogaard and Buckingham, 1973) and confirmed two years later (Hug *et al.*, 1975). In spite of these requirements, ROA has proven to be an incisive probe of stereochemistry and molecular structure. Numerous studies by the Barron group at the University of Glasgow pioneered the use of ROA for conformational analysis of proteins, nucleic acids, viruses, and glycoproteins (Barron, 2004; Blanch, Hecht and Barron, 2003, more recent references). This body of work in conjunction with the introduction of the first commercial ROA spectrometer by BioTools Inc. in 2001 has generated the recent increase in the number of ROA researchers and the scope of applications. For a thorough description of ROA theory, results, and applications, the reader is directed to other reviews (Barron *et al.*, 2004).

19.1.2
Overview of SERS

At the same time as ROA became reliable, another phenomenon, known as *surface-enhanced Raman scattering (SERS)*, in Raman scattering was discovered (Fleischmann, Hendra and McQuillan, 1974; Jeanmaire and Van Duyne, 1977; Albrecht and Creighton, 1977). This effect takes place when a molecule is adsorbed on, or is in the vicinity of, a metallic nanoparticle, typically Ag, Au, or Cu. Consequently, an enhancement in the scattered field takes place. This effect was very soon adapted by most research laboratories within the Raman community because in general it provides, in spite of the difficulty of having stable nanoparticles, an enhancement of minimum 10^3 to 10^7 of the scattered Raman intensity (Kerker, 1990). The SERS effect has since its discovery been widely studied and the enhancement can be attributed to the EM mechanism as well as to the chemical (Ch) mechanism, and the signal can be defined as follows:

$$\phi^{\mathrm{SERS}} = \frac{I_I}{\hbar\omega_I}\sigma^R \sum_{i=1}^{N} M_i^{\mathrm{EM}} M_i^{\mathrm{Ch}} \qquad (2)$$

where I_{I} and ω_{I} are the intensity and the angular frequency of the incident beam, respectively, σ^{R} the Raman cross section (a parameter which is proportional to the square of the first polarizability derivative for a vibrational transition and the fourth power of the scattering frequency), and M_i^{EM} and M_i^{Ch} are the enhancements due to the EM and Ch contributions, respectively. This giant enhancement has been of great importance especially in biomolecular studies because within this effect the cross section is increased to the level of fluorescence, and on the other hand, and most importantly, as the intensity is proportional to the number of molecules taking part in the scattering, SERS enables studies of molecules at lower concentration compared to instant ordinary Raman scattering. That is, molecules at lower concentration under enhancement conditions can now provide the same intensity of signal, by keeping other experimental parameters constant. This can be expressed as follows:

$$P_{\mathrm{SERS}}(\nu) = NI(\nu)|A(\nu_L)|^2|A(\nu_s)|^2\sigma_{\mathrm{ads}} \qquad (3)$$

where N is the number of molecules, $I(\nu)$ the intensity of the excitation laser, $A(\nu_{\mathrm{L}})$ and $A(\nu_{\mathrm{S}})$ the field confinement of the optical coefficients related to the incident and the scattered light frequency, respectively, and σ_{ads} is the adsorption cross section. The field enhancement factor depends also on the distribution of the colloids in the measured sample.

19.1.3
Overview of SEROA

It seems natural, as it did at the time when SERS was developed, that the surface enhancement phenomenon could also be employed in ROA, and when it succeeds it will allow researchers to overcome the weak signal, high concentration, and longer time acquisition required for ROA. Nevertheless, many attempts were made and no successful result was reported, in spite of the fact that Efrima clearly expressed that "I propose that such measurements (i.e. surface-enhanced Raman optical activity (SEROA)) are feasible," and he predicted that these measurements could contain information on both the target molecule and "local electric fields, their gradients and, in general, local dielectric properties of the metal–air (or metal–solution) interface" (Efrima, 1983). Later, Efrima developed the first theoretical treatment of the metal surface enhancement of ROA bands in a more extensive paper (Efrima, 1985). In 1994, Hecht and Barron developed expressions for ROA scattering, as well as for the case of linearly polarized difference Raman scattering, from chiral molecules on isotropic surfaces, including contributions from antisymmetric contributions of polarizability tensor components for the first time (Hecht and Barron, 1994, 1995). The authors suggested that ROA could be a promising technique for probing the chirality of metal surfaces. Despite these developments, only two instances of SEROA were reported for two decades (Higuchi, 1990; Yu, 1994), but the signal-to-noise level in each case was low and so the results were equivocal. The results, mentioned below, can be considered as the first reported experimental SEROA (Abdali, 2006; Abdali *et al.*, 2007; Johannessen and Abdali, 2007; Abdali and Blanch, 2008).

19.2
Theory of SEROA

Efrima was the first to consider the effects of large electric field gradients on the electric dipole–magnetic dipole and electric dipole–electric quadrupole optical activity contributions to ROA scattering. In his first published report describing this phenomenon, he showed theoretically that additional molecular specific information could be obtained "in a feasible manner" in the vibrational spectra (Efrima, 1983). However, this work only discussed the effect of electric field gradients on ROA, and concluded that the CID (Δ) could be of the order of unity. Efrima also surmised that SEROA spectra should appear very different to the ROA spectra in bulk solution owing to the effects of the plasmon resonance on the contributions to ROA scattering and to the anisotropic environment of the analyte–colloid interface and the local dielectric properties. He also raised the first recommendation for caution in interpreting SEROA spectra. Though this was specifically in terms of the possible responsible mechanism, subsequent studies have indeed shown that the analysis of measured SEROA spectra requires care owing to the experimental artifacts that can arise. Changes in direction of the local field and the simple magnitude of the electric field were shown to be insufficient in themselves to generate large CID values, and that large field gradients are required for this to occur (Efrima, 1985). These field gradients were proposed to be far larger than those associated with the radiation field, which implied that even if the electric field associated with a plasmon resonance was too small in intensity to generate a large SERS signal it may still generate a measurable SEROA signal. Efrima derived the following relationship between CID values for the bulk for nonresonant ROA (Δ_b) and for the surface enhanced, that is SEROA, case (Δ_s) as follows:

$$\frac{\Delta_s}{\Delta_b} \approx \frac{[(E_{\mathrm{loc}}/d)/E_{\mathrm{loc}}]}{[(E_i/2\pi\lambda)/E_i]} \tag{4}$$

where E_i is the incident electric field, E_{loc} is the local contribution to the electric field, d is the range of the local field, and λ is the wavelength of the radiation used.

Reflection from metal surfaces is known to be able to change the polarization state of circularly polarized light to the elliptically polarized state, introducing a linearly polarized component. Therefore, the effect of reflection on the polarization state needs to be taken into account. Efrima concluded that for imperfect metal surfaces, such as metal colloids, a significant degree of CP is still likely to exist in the scattered radiation, and is derived from this work formalisms for CID intensity in both the parallel ($\Delta^{||}$) and perpendicular ($\Delta^{\perp}$) polarizations (Efrima, 1985).

Although most of the scattered radiation is expected to be circularly polarized, it is possible that any linear polarization introduced by reflection may lead to significant artifacts being introduced into measured SEROA spectra. These artifacts would have to be carefully controlled and minimized. Efrima found that the magnetic field component was not expected to make a large contribution to SEROA spectra, which is different to the situation for ROA measurements made in solution, and that near the metal surface only the electric polarizability–electric polarizability contribution was likely to be significant, and a change of the local electric fields direction would not influence the value of the CID. Efrima further suggested that another requirement for the generation of large SEROA signals was either a significant imaginary component to the electric polarizabilities of the analyte on the metal colloid surface or a phase difference between the electric field and its gradient, which is defined as

$$\nabla_\beta E_I = \delta_{\beta,z} \sum_i f_i \hat{i} \tag{5}$$

where $i = x, y, z$ and

$$\begin{aligned} f_x &= [\partial R_x(\theta_i)/\partial z] E^0_{||} \cos\theta_i \\ f_y &= [\partial R_y(\theta_i)/\partial z] E^0_{\perp} \\ f_z &= [\partial R_z(\theta_i)/\partial z] E^0_{||} \sin\theta_i \end{aligned} \tag{6}$$

where the reflection coefficient R depends on the scattering angle and the frequency of the scattered light. As for the CID derivation, Efrima defined the Cartesian component of the electric field emitted in the direction given by the wave vector k and detected at distance r as

$$\begin{aligned} E^\alpha_d = {} & (\omega^2 \mu_0 / 4\pi r) \exp[r/c - t)] \\ & \cdot [1 + R_\alpha(\theta_d)] \cdot \left[\left(\mu_\alpha - \sum_\beta k^\alpha_d k^\beta_d \mu_\beta \right) \right. \\ & - (1/c) \sum_{\beta,\gamma} \varepsilon_{\alpha\beta\gamma} k^\beta_d m_\gamma - (\mathrm{i}\omega/3c) \\ & \left. \times \sum_\beta \left(k^\beta_d \theta_{\alpha\beta} - \sum_\gamma k^\alpha_d k^\beta_d k^\gamma_d \theta_{\alpha\gamma} \right) \right] \end{aligned} \tag{7}$$

where μ and m are the electric- and magnetic-induced dipoles, respectively, and θ is the induced molecular electric quadrupole tensor. These molecular multipoles are defined as (Barron and Buckingham, 1971)

$$\begin{aligned}\mu_\alpha &= \sum_\beta \Big\{\alpha_{\alpha\beta} E_I^\beta + G_{\alpha\beta} B_I^\beta \\ &\qquad + (1/3) \sum_\gamma A_{\alpha,\beta\gamma} (\nabla_\beta E_I^\gamma) + \cdot \Big\} \\ m_\alpha &= \sum_\beta G^*_{\beta\alpha} E_I^\beta + \cdot \\ \theta_{\alpha\beta} &= \sum_\gamma A^*_{\gamma,\alpha\beta} E_I^\gamma + \cdot \end{aligned} \tag{8}$$

where α is the electric dipole–electric dipole polarizability tensor, G the electric dipole–magnetic dipole polarizability tensor, and A is the electric dipole–electric quadrupole polarizability. Evaluation of these CID expressions showed that no matter what the cause of the enhancement will be, whether resonance or chemical effect, there is always a fair chance that ROA could be measured for an adsorbed molecule, and the CID will always be sensitive to the orientation of the adsorbed molecule.

Hecht and Barron (1994, 1995) considered the case of a pure electric dipole surface ROA mechanism, in their attempt to develop the SEROA theory. In these two papers, the authors included antisymmetric polarizability components for the first time, and an incoherent scattering from a chiral molecule attached to an isotropic surface fixed with respect to the laboratory frame was also considered, with averaging of the ROA tensor products over two dimensions, and they showed that CID values much larger than those found for ROA measurements in solution may be possible. That is, the CID expressions for SEROA experiments may be significantly larger than those in conventional ROA experiments. Hecht and Barron found that these CIDs in the vicinity of metal surfaces originate from interference between adsorptive $\alpha_{\alpha\beta}(a)$ and dispersive $\alpha_{\alpha\beta}(d)$ electric dipole–electric dipole molecular polarizability tensor components:

$$\tilde{\alpha}_{\alpha\beta} = \alpha_{\alpha\beta}(d) + \mathrm{i}\alpha_{\alpha\beta}(a) \tag{9}$$

where $\tilde{\alpha}_{\alpha\beta}$ characterizes Cartesian components of the complex electric dipole–electric dipole polarizability tensor, and both the adsorptive and dispersive parts of the polarizability are purely real in the absence of a time-odd influence such as a static external magnetic field.

This situation will only be observed under resonant, or preresonant, conditions. It was pointed out by Hecht and Barron that isotropic surface ROA would vanish for achiral molecules and would give signals of equal magnitude, but opposite sign for two enantiomers, thereby leading to zero ROA. If a plane containing the normal to the metal surface and one of the axes parallel to the surface is considered, then a reflection across this plane changes the sign of off-diagonal components in the polarizability tensor. This reflection generates an enantiomeric isotropic surface for chiral molecules, but an identical surface for achiral molecules. Hecht and Barron also noted that although some achiral point groups may exhibit chiral projections on a surface, these examples would not be expected to give ROA from isotropic surfaces owing to the cancellation of enantiomeric projections, though surface ROA could still arise from an ordered surface.

A full orientational averaging and dressed molecular response tensors for scattered circular polarization (SCP) ROA was recently reported, though this formulasim is equally applicable to all ROA experimental configurations (Janesko and Scuseria, 2006). These researchers considered the fields produced by the multipoles that were induced in a molecule by an oscillating light wave, and the response they would generate in the metal substrate in order to specify the multipole moments of the molecule–substrate system. Janesko and Scuseria derived expansion coefficients for these multipole moments and thus determined general selection rules and scaling laws for SEROA experiments. For the simple model system of a dipolar sphere, they found that predicted SERS intensities agreed well with measured SERS intensities and that SERS enhancements for vibrational modes, whose polarizability derivatives are parallel to the surface normal, are 16 times larger than for those modes whose polarizability derivatives are parallel to the surface. Janesko and Scuseria also found that SEROA intensity scales with respect to an order parameter and that there were no SEROA contributions lower than first order of this parameter when full orientational averaging was taken into account. This finding is in contrast to the earlier model proposed by Efrima (1985) and leads to CID values smaller than those for ROA of bulk samples.

Janesko and Scuseria then considered quadrupolar spheres as substrates, and found that the CID values for these substrates could be much greater due to the existence of a gradient enhancement mechanism across the metal substrate particle rather than just across the molecule. They determined that the maximum CID using this model was of the order of 10^{-2}, and CID values could possibly be even higher for more realistic substrates. Dressed ROA tensors derived by Janesko and Scuseria contained contributions from the molecular polarizability, which gives rise to SEROA scattering that scales purely as the mean polarizability. This finding is similar to the chiral surface effect considered by Hecht and Barron (1994, 1995), and was again predicted to be nonzero only for chiral molecules. Janesko and Scuseria also predicted differences between incident circular polarization (ICP) and SCP ROA spectra, which are similar to the case of ROA measured under resonance or near-resonance conditions with an excited electronic state of the molecule, and this may be owing to differences in the molecular response to the incident and scattered fields. The introduction of quadrupolar responses to the model further suggested that CIDs could also be enhanced for dipolar nanorods used as substrates.

This multipole EM field expansion model was recently reworked by Bour into a matrix formalism (Bour, 2007). Using this new approach, Bour computed SEROA spectra of CHFClBr considering a single colloidal sphere and for the situations of an analyte between two colloidal spheres or between four chirally arranged colloidal spheres. Large differences were found between computed SEROA bands, both in sign and intensity, for these three situations, showing that SEROA spectral profiles should be highly sensitive to changes in both analyte–colloid and colloid–colloid interactions, than the corresponding SERS spectra. Bour also found that increasing the number of colloidal spheres around an analyte molecule increased

CID values, and confirmed the importance of quadrupolar contributions.

Recent work has also been published, modelling polarization effects in SERS, which is particularly relevant to SEROA (Etchegoin, Galloway and Le Ru, 2006). These authors reported that the local electric field of the plasmon resonance has a significant effect on the polarization state of both the incident and scattered waves. Such effects would need to be carefully considered both for the design of SEROA experiments and for the interpretation of measured SEROA spectra. The effect of relative orientation of a dimer hot spot with respect to the angle of incidence of the light wave suggested that large CID enhancement factors for SEROA measurements were not likely in typical situations. Etchegoin *et al.* further cautioned that measurement of SEROA may require differentiation between small signals originating from the analyte and large background, or artifactual, signals.

19.3
ROA Instrumentation

An ROA spectrometer has only recently become commercially available, as mentioned above, and it is mainly based on novel work by Hug and Hangartner (1999). The use of CP of light is fundamental to any ROA instrument, and different combinations exist depending on whether the CP takes place before the sample, that is, in the incident path (ICP), or after scattering from the sample, that is, in the scattered path (SCP), or their combination in the form of dual circular polarization (DCP), that is, in both the incident and scattered directions. An excellent discussion of these different forms of ROA has been presented by (Nafie and Che, 1994). The different mechanisms are schematically shown in Figure 19.1. It is worth mentioning that the existing commercial ROA instrument, Chiral*RAMAN*, employs the SCP mechanism, although half waveplates are also mounted, both in the incident and scattering path to average out any linear birefringence in the beam. This eliminates the so-called flicker noise, which usually exists (and influences the results) due to dust particles, density fluctuations, laser power fluctuations, and so on. This is important because the intensity difference measurements required to extract the circularly polarized components of the scattered beam are taken between two orthogonal components of the scattered light measured during the same acquisition period. Employing such a technique, that is, cancellation of the "flicker noise," has greatly improved signal-to-noise characteristics, by approximately a factor of 5 compared to the ICP ROA design previously used, allowing ROA spectra to be measured for lower sample concentrations and in far shorter timescales. Today, the Chiral*RAMAN* instruments in use utilize excitation wavelength of 532 nm and the ROA spectra recorded are virtually identical to ROA spectra measured using ICP instruments. Recently, a new Chiral*RAMAN* SCP ROA spectrometer has been reported, and this spectrometer utilizes near-infrared excitation at 780 nm (Nafie *et al.*, 2007). Such an excitation is intentionally chosen so that it will minimize the presence of fluorescence, and because of the properties of metals, it will allow investigation of the SEROA phenomenon using gold substrates rather than the silver colloids investigated so far. DCP ROA

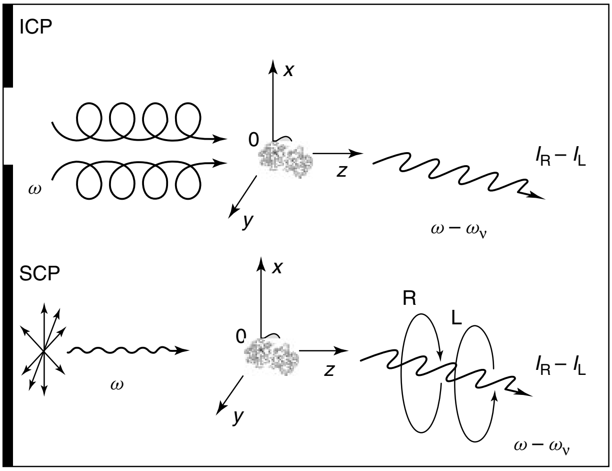

Fig. 19.1 Different mechanisms of ROA are schematically shown. The incident circular polarization (ICP) makes use of the circular polarization of the incident light (top), whereas scattered circular polarization (SCP) makes use of the circular polarization of the scattered light (bottom). I_R and I_L refer the intensity of the beam for the right and left circular polarized light, respectively.

spectroscopy has been less widely exploited and it relies upon either in-phase (DCP_I) or out-of-phase (DCP_{II}) measurement of scattering using both incident and SCP. All the results presented here were obtained using Chiral*RAMAN* instruments at 532 nm.

19.4
SEROA, Measurement Results

19.4.1
SEROA of Small Molecules

As the theoretical work discussed above has indicated, care must be exercised in the analysis of measured SEROA spectra. Figure 19.2 presents relevant SERS and SEROA spectra measured for proline. The top two panels show the conventional Raman and ROA spectra collected for a solution of proline (0.939 M), with 700 mW of laser excitation and 81 minutes of data collection. In the third panel are shown the SERS spectra measured for 8.8 mM proline in the presence of citrate- and hydroxylamine-reduced silver colloids, and with 20 mM K_2SO_4 used as an aggregating agent for both sols. The corresponding SEROA spectra for proline in these two silver sols are shown in the fourth panel with 100 mW laser excitation at 532 nm and 60 minutes of data acquisition. These spectra are unprocessed and show that there are a number of bands common to both measurements, in the range ~1300–1550 cm^{-1}, which indicates possible SEROA bands originating from the proline analyte. The CID values of these SEROA features appear to be of the order 10^{-3} to 10^{-4}, which is similar to CID values measured for the conventional ROA spectra of proline. There are clearly significant differences between the features displayed in the

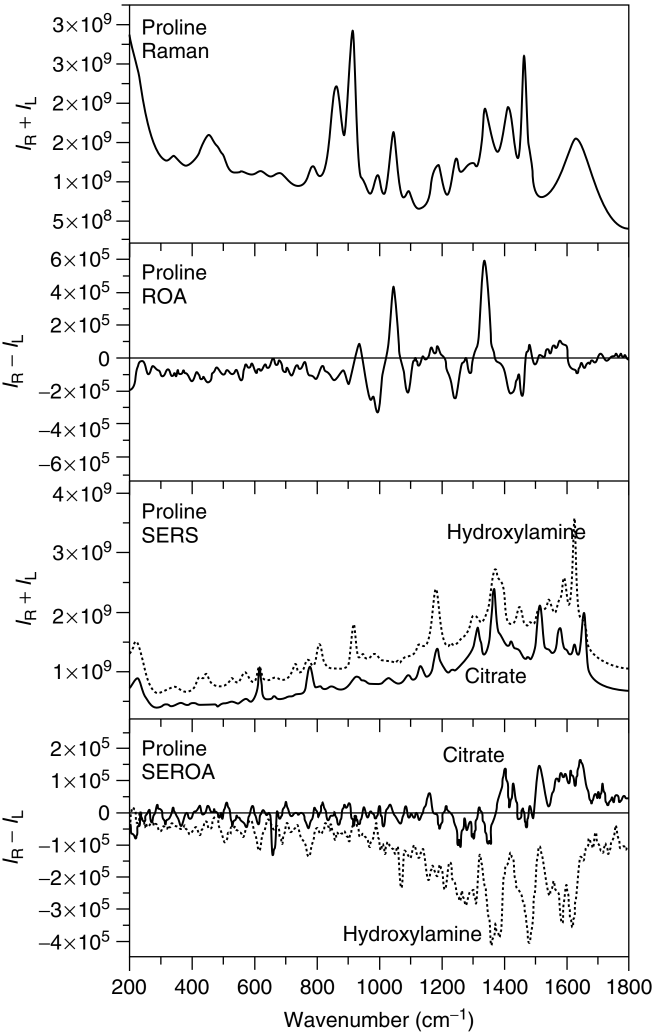

Fig. 19.2 SERS and SEROA spectra measured for proline. The top two panels show the conventional Raman and ROA spectra collected for a solution of proline (0.939 M) at 700 mW laser excitation and 80 minutes of data acquisition. The third panel shows the SERS spectra collected for a solution of proline (8.8 mM) in two Ag colloids; citrate and hydroxylamine. The corresponding SEROA spectra are shown in the fourth panel, at 100 mW laser excitation and 60 minutes of data acquisition.

ROA and SEROA spectra, as Efrima suggested would be observed (Efrima, 1985). There are also differences between the two SEROA spectra shown, with the SEROA spectrum measured for the hydroxylamine-reduced colloids being distorted in a negative direction compared to the SEROA spectrum measured for citrate-reduced colloids. Thus, several SEROA features have been distorted, particularly the two negatively signed bands found near 1600 cm^{-1}. Similar distortions are sometimes observed in conventional ROA spectra from polarized Raman bands. These polarization effects induced by plasmon resonances may be significant and so should be considered when evaluating SEROA spectra (Etchegoin, Galloway and Le Ru, 2006).

The distortion effects of these polarizable SERS bands on the features of SEROA spectra are also observed for the nucleoside cytidine in Figure 19.3. SERS spectra for cytidine (at 5.8 μM concentration, 75 mW laser excitation, 3.5 minutes data acquisition per spectrum, using citrate-reduced silver colloids and 3.8 mM NaCl as aggregating agent) are shown in this figure at time intervals of 20 minutes throughout the aggregation process. We can see that the SERS spectral profile is consistent for each of these spectra with the absolute intensity of each band changing in the same proportion to all others shown. In contrast to this, greater changes are apparent in the corresponding SEROA spectra, as shown in the lower panel in Figure 19.3. Several bands in this panel, originating from intense SERS bands, appear to decrease in relative intensity, indicating that these bands are not real SEROA bands but are probably artifacts and are being averaged out over time. After approximately 45 minutes several weak bands begin to appear; such as those at ∼400, 1270,

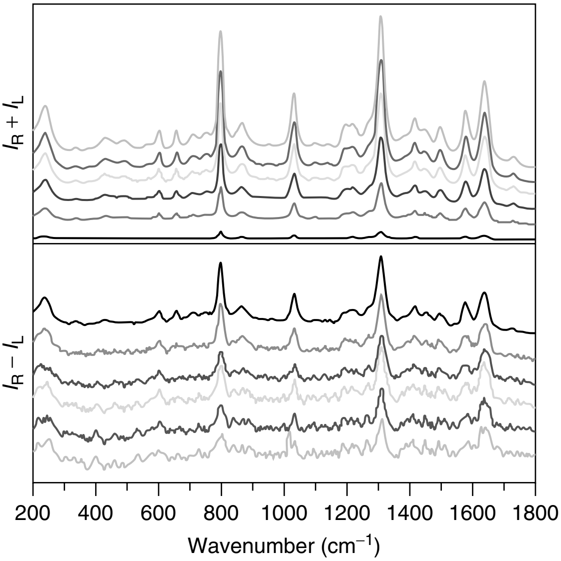

Fig. 19.3 SERS (top) and SEROA (bottom) spectra measured for cytidine with the following time interval of 20 minutes, starting with black (after 5 minutes) and ending with green (after 105 minutes). SERS intensities increase with time, whereas SEROA features associated with large SERS bands, and therefore prone to artifacts, become less intense, revealing weaker features in the SEROA spectra. (Please find a color version of this figure on the color plates.)

and 1460 cm^{-1}; suggesting that these weak features are real SEROA bands originating from the cytidine analyte. Although the SEROA spectra presented in Figures 19.2 and 19.3 are not yet definitive proof of the SEROA technique, they do show the importance of careful characterization of SEROA bands and how these can vary with respect to time and the aggregation behavior of the colloids in the sol.

19.4.2.
SEROA of Protein

A SEROA measurement of a chiral pentapeptide, enkephalin, was reported at the ICORS conference in 2004, and the results were later published (Abdali, 2006). This work illustrates the possibility of detecting SEROA, and the same group succeeded to employ SEROA in a series of recent publications. In these reports, it has been shown that recording a SEROA spectrum does not take place whenever a SERS spectrum is recorded, as it is the case for ROA and RS. Here, great care should be taken for optimizing the system, and SEROA is conditioned by optimization of the concentration of the analyte, colloid, and aggregating agent composing the sample (Abdali *et al.*, 2007). In Figure 19.4, the intensities of two resonance SERS peaks for myoglobin at 532 nm are plotted as a function of the concentration of myoglobin, Ag colloids, and NaCl, respectively. The two peaks, representing the symmetric (1374 cm^{-1}) and antisymmetric (1169 cm^{-1}) ν (pyr half ring) stretching modes were chosen to monitor SERS intensity owing to the fact that these peaks were found to be independent of the signal-to-noise ratio.

The concentration conditions are clearly seen to influence the results, in that any variation of the concentration of one of the sample components yields variation in the corresponding SERS. Accordingly, a series of different concentrations was investigated, from which the optimum conditions were found to be 10^{-7} M myoglobin, 70% Ag, and 0.1 M NaCl, in order to achieve the best SERS measurement of myoglobin, as shown in Figure 19.4.

It is known that the number of molecules per unit volume is important in any Raman measurement, whereas the number of molecules currently adsorbed at the SERS active sites influences the measured SERS intensity. This explains the behavior of the intensity observed in Figure 19.4, in which as the sample concentration initially increases the number of SERS active sites will also increase, and consequently higher SERS intensity will be obtained, a tendency that continues until a maximum value of the intensity is reached. Beyond this point, any increase in the sample concentration will yield a decrease in SERS intensity due to an increase in the background or/and increased fluorescence, which results from the aggregation of sol, leading to an increase in opacity, especially in the case of colored samples, such as myoglobin.

Resonance enhancement of Raman scattering (Resonance Raman Spectroscopy) occurs when an electronic transition is triggered by the incident light. The electronic excitation increases the Raman scattering probability of vibrational normal modes in the spatial vicinity of the chromophore, thereby enhancing these signals due to the scattering cross section enhancement by usually up to five orders of magnitude.

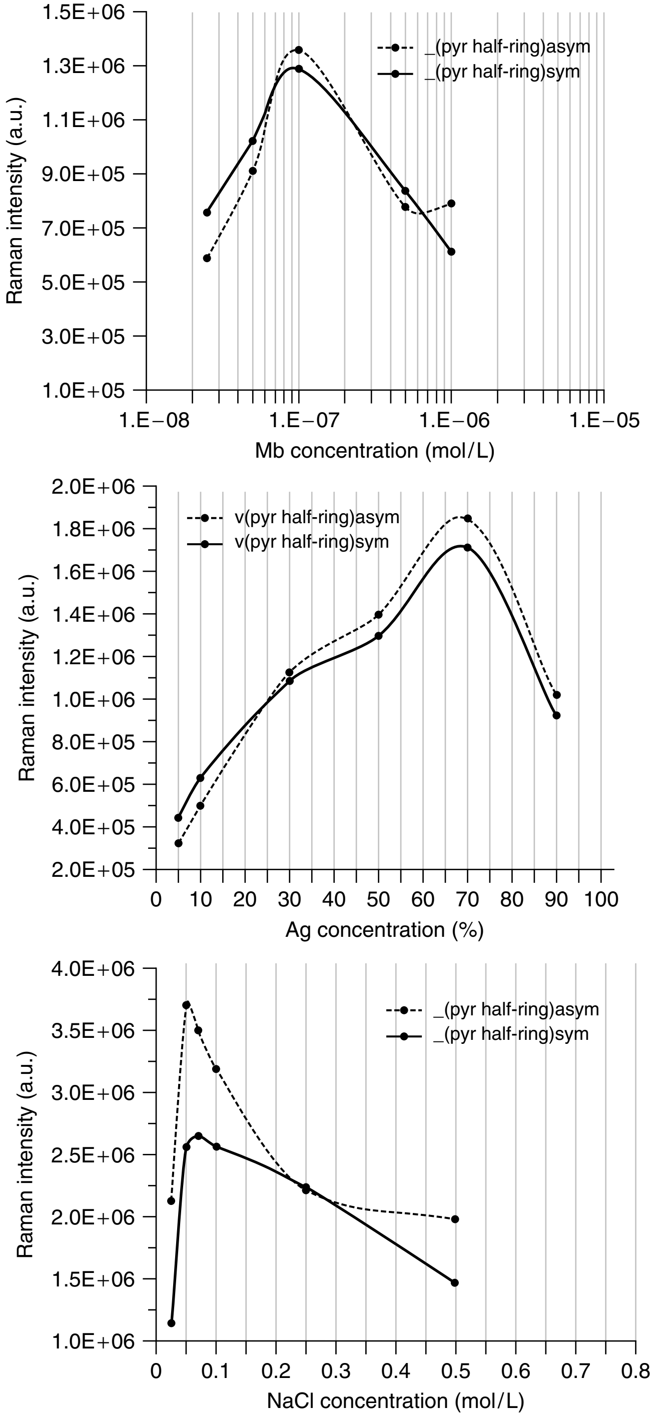

Fig. 19.4 The optimization of the Mb concentration (top), the colloids Ag (middle), and the aggregating agent NaCl (bottom).

When RRS enhancement couples with surface plasmon enhancement, a further enhancement of the resonant vibrations occurs, known as *surface-enhanced resonance Raman scattering (SERRS)*. Therefore, the SERS measurement of myoglobin using 532 nm excitation shown in Figure 19.4 is in fact a SERRS spectrum. Of course, SERRS measurements are only possible when there is sufficient overlap between the electronic transition frequency of the sample and the surface plasmon resonance of the nanoparticles. It is also known that measurement of ROA under electronic resonance conditions leads to resonance Raman optical activity (RROA), and the combination of electronic resonance with SEROA generates resonance surface-enhanced Raman optical activity (SERROA).

The RROA and SERROA spectra of myoglobin, using 532 nm line source, are also shown in Figure 19.5, with the SERROA measurement being carried out under the optimized conditions, as shown in Figure 19.4. Although chiral peaks in the spectrum were observed after only a few minutes, in order to achieve a better signal-to-noise ratio, the measurement was performed for 5 minutes. The SERRS signal was simultaneously monitored for erratic hot spots, and found to be stable during the experiment. In order to verify the authenticity of the surface-enhanced effect, an RROA spectrum of myoglobin of 100 μM concentration was also recorded, as shown in Figure 19.5, and was obtained during no less than 32 hours. This time was necessary to achieve a sufficient signal-to-noise ratio. Comparing the RROA and SERROA spectra in Figure 19.5, it is noticeable that the chiral signals of myoglobin, labeled with corresponding frequencies, are similar in both the spectra. In particular, the region 1300–1400 cm^{-1}, which also exhibits the largest signal enhancement in the SERRS spectrum, has a considerable signature overlap around 1395, 1353, 1321, and 1299 cm^{-1}. However, the SERROA peak around 1353 cm^{-1} appears to have bigger line width (full width at half maximum (FWHM)) than the corresponding RROA feature as shown in Figure 19.5, which may explain the absence of bands at ~1366 and 1344 cm^{-1}. The Raman bands in this region have previously been assigned to vibrations of the heme group in the protein, and

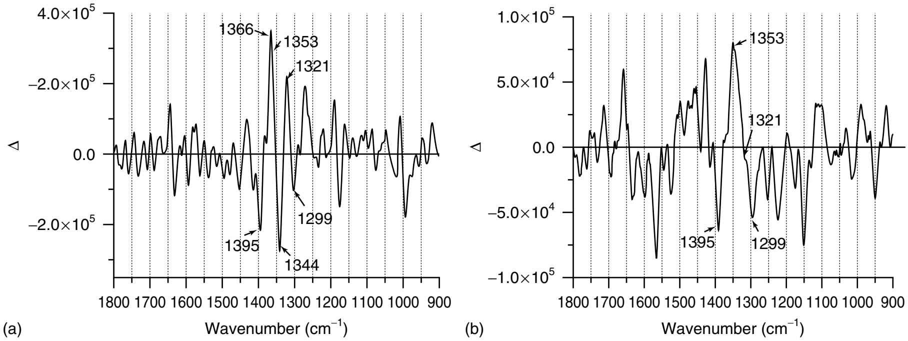

Fig. 19.5 Myoglobin spectra using (a) RROA for 100 μM and 32 hours acquisition and (b) SERROA for 100 nM and 30 minutes acquisition. The SERROA measurement carried out under the optimized conditions is shown in Figure 19.4.

have been shown to be enhanced by the resonance Raman effect, using an excitation source of 532 nm (Bizzari and Cannistrato, 2002). At 100 μM concentration of the sample, conventional ROA bands from polypeptide backbone vibrations would be too weak to measure, so the peaks found in the RROA spectrum in Figure 19.5 were assigned to the resonance enhancement of the porphyrin vibrations.

Employing the same optimization conditions, SERRS and SERROA spectra of another porphyrin-containing protein, cytochrome *c*, were also reported (Johannessen and Abdali, 2007). However, in this work hydroxylamine-reduced silver colloids were used (Leopold and Lindle, 2003). The RRS, RROA, SERRS, and SERROA spectra of cytochrome *c* are shown in Figure 19.6. It is worth mentioning that RROA features observed for cytochrome *c* are not similar to those usually observed for globular proteins. This clearly illustrates that the measured chiral signals originate due to vibrations of the porphyrin ring. In the range 1100–1700 cm^{-1}, most of the RRS bands observed for cytochrome *c* are assigned to strongly enhanced in-plane vibrations that also give rise to contributions for the major features in the RROA spectrum shown. Although the ν_{10} and ν_{11} transitions have positive RROA, the ν_{19} signal, located between these vibrations, appears to be a negative–positive couplet, indicating contributions from two transitions. Furthermore, it is clear that the observed signals originate from the porphyrin group in a chiral environment.

The bands ν_{20} and ν_4, both of which exhibit positive ROA, are due to the in-plane transition. In this region of the spectrum, the predominant feature is due to the ν_4 transition, whose resonance enhancement is highly sensitive to

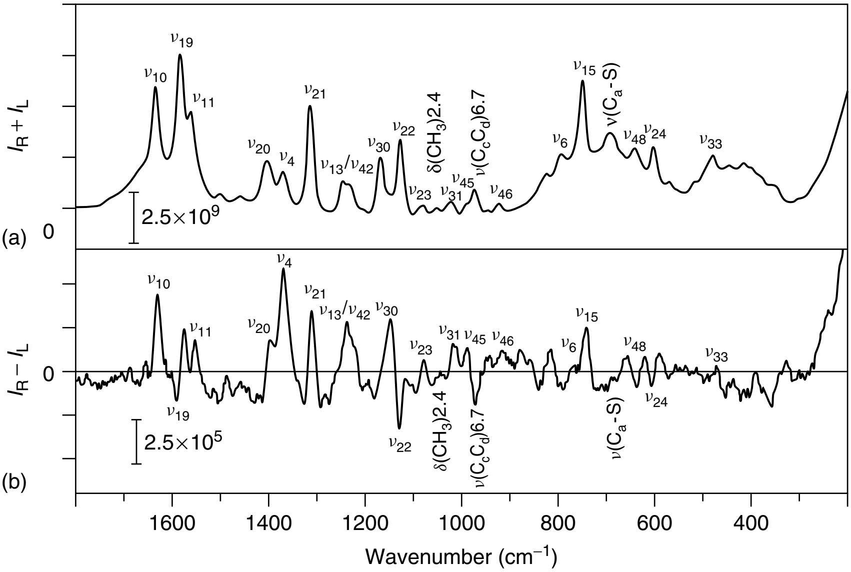

Fig. 19.6 Vibrational (a) RRS and (b) RROA spectra of cytochrome *c* with concentration of 100 μM and 16 hours acquisition time.

the excitation wavelength; however, it does not seem to be very strongly enhanced by the Q-band excitation caused by the 532 nm laser used for this measurement. This is apparently because the collective half ring stretching vibration, responsible for the ν_4 signal, is highly chiral and results in a strong signal.

Local in-plane deformations and side chain vibrations having weak ROA contributions can be assigned to the region between 800 and 1100 cm^{-1}, where less enhanced SEROA signals can be seen. Nevertheless, the two overlapping bands assigned to ν_{45} and the propionate stretching vibration at $\sim$974 cm^{-1} appear as a positive–negative couplet in the RROA spectrum. In the region below 800 cm^{-1}, ν_{15} is strongly enhanced and represents the main feature in this region, and it yields a positive RROA signal, whereas limited or no ROA signals can be found in this region.

It is noticeable that to a certain degree the enhancements of Raman active transitions in SERRS correlate with the corresponding SERROA signals. In the wavenumber region ($>$1100 cm^{-1}), the signals corresponding to the ν_{10} and ν_{11} transitions are observed as positive peaks in the RROA spectrum, whereas the signal of ν_{19} couplet appears to be within the noise level. This shows that although the first two transitions, ν_{10} and ν_{11}, are enhanced, the plasmon effect is not the same on the later transition ν_{19}. The absence of SERROA here may also be a consequence of an increased polarization of the transition (or transitions), which yields artifact prone data collection. Such comparisons can also show the considerable enhancement of ν_{20} SERROA signal due to the plasmon effect than that seen in the RROA spectrum. On the other hand, the adjacent ν_4 transition seems to undergo no further enhancement in the SERROA spectrum. It is therefore of importance to conclude that the RRS spectrum of cytochrome *c* shows a remarkable enhancement of the ν_4 transition compared to ν_{20}, whereas this is absent in the SERROA spectrum, an effect that is apparently the opposite for the ν_{20} transition.

Other significant differences between RROA and SERROA spectra of cytochrome *c* are also clearly observed in number of bands. The ν_{30} band in the SERROA spectrum is very weak compared to that in the RROA spectrum. This may be due to the changes of the dipole moment and polarization of the molecule, upon adsorption to the silver colloids, or as a result of the symmetry changes in the porphyrin ring. Such inversions taking place in the prosthetic group can explain the total inversion of the ν_{22} band in the SERROA spectrum. Another inversion can also be seen in the couplet around 1000 cm^{-1}, which is assigned to overlap the ν_{45} and the propionate stretching vibration at 974 cm^{-1} in the RROA spectrum. While this couplet appears as a positive-negative feature in the RROA spectrum, the SERROA spectrum reveals a weak negative-positive couplet, the negative portion of the couplet commencing immediately after ν_{31}. In addition to being an interesting contribution to the general understanding of the plasmon effect and the influence on adsorbed molecules to metal surfaces, such behaviors of these bands also necessitates more caution, when assignments are to be made using a comparison between ROA SERROA spectra. In this regard, Janesko and Scuseria (2006) found that quadrupolar responses of the substrate may significantly enhance ROA in certain cases, which could explain the apparent lack of SERRS signal from, for example, ν_{21} as the SERROA signal is enhanced.

Like ROA spectra, SEROA spectra have also been shown to have sensitivity to ligand binding (Johannessen, White and Abdali 2007). In this work, the authors extended the SERROA study of myoglobin discussed in Figure 19.5 to investigate the sensitivity of SEROA spectra to small changes in the molecule, specifically the effect of adding a ligand (azide) to myoglobin. The results showed that the SERRS spectrum of the myoglobin–azide complex was almost identical to the SERRS spectrum of myoglobin, as shown in Figure 19.4, whereas larger differences were observed between their corresponding SERROA spectra. This result shows that the sensitivity to perturbations of the molecule structure inherent to ROA spectroscopy can be retained in SERROA.

It is clear from Figure 19.7 that although spectral changes occur in the RRS spectra of myoglobin, above 1500 cm^{-1}, upon complexation of the azide ligand, only few changes are found between the SERROA bands. This implies that the localized porphyrin transitions, which are found in this region, are less susceptible to chiral analysis. However, it should be emphasized that the structural changes induced by complexation of azide, or other small ligands, in metalloproteins that contain porphyrin are very small, that is, <0.5 Å. Such changes cannot be determined directly even by X-ray diffraction. Therefore, the

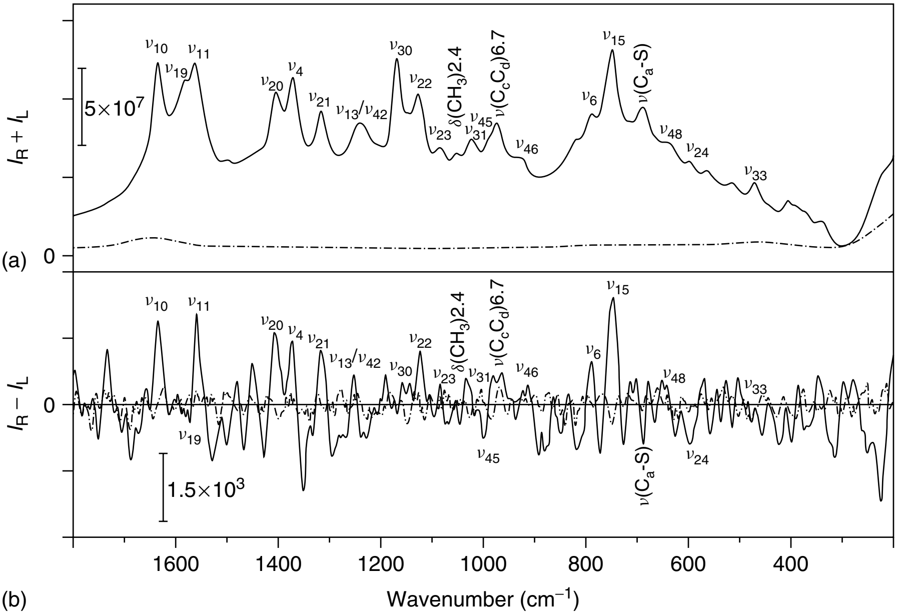

Fig. 19.7 Vibrational (a) SERRS and (b) SERROA spectra of the sample in Figure 19.6 mixed with hydroxylamine-reduced Ag colloids. The sample concentration here is 100 nM and with 30 minutes acquisition time. The control experiment is seen in dashed line.

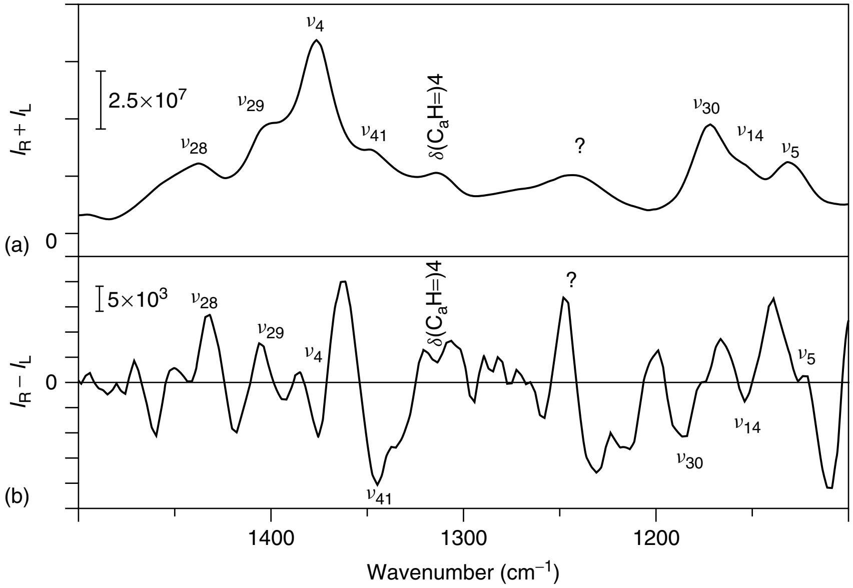

Fig. 19.8 SERRS and SERROA spectra of myoglobin of concentration of 100 nM.

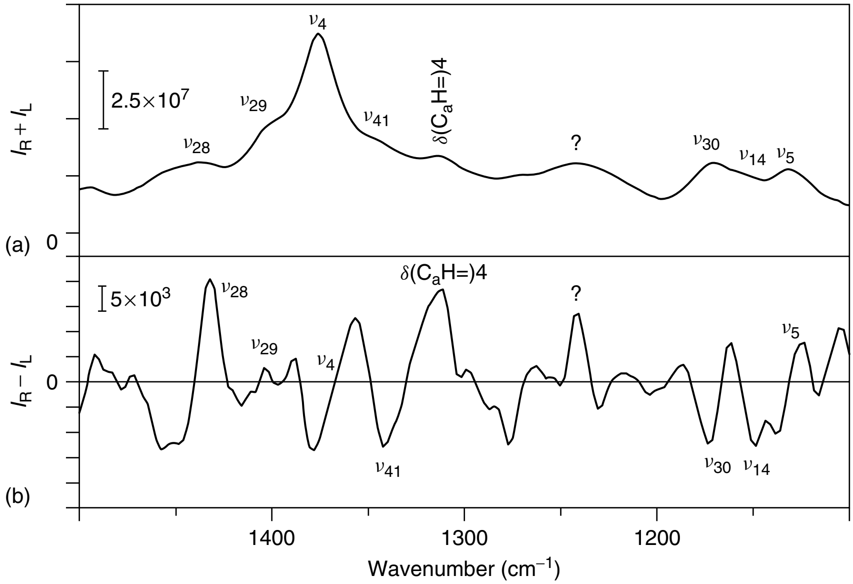

Fig. 19.9 SERRS and SERROA spectra of myoglobin–azide complex of concentration of 100 nM.

ability of SERROA to detect these structural changes is highly promising for future work of this nature.

19.5
Future Perspectives

The gold rush after methods and tools to detect single molecules has been ongoing since a long time, and there has certainly been great progress in this field. Raman variant techniques, that is, Raman spectroscopy, ROA, SERS, SEROA, have potentially important roles in biology due to the capability of these tools to operate *in vivo, in vitro*, and *in silico*, and without damaging the studied molecule. Today, these tools are applied in biology, chemistry, medicine, drugs industry, art and archaeology, forensic chemistry, and many other fields. Further developments in Raman and Raman-combined methods are expected to continue, particularly in instrumentation, which are expected to widely spread the applications of vibrational spectroscopy, and this can already be seen in combining Raman to many spectroscopic methods such as fluorescence, optical tweezers, atomic force microscopy (AFM), and the new tool, designated *tip-enhanced Raman scattering* (*TERS*) (Domke, Zhang and Pettinger, 2006). TERS seems to have wide potential, for instance, when combined with surface-enhanced method, for example tip enhanced Raman optical activity (TEROA); the effect will be studied in a more controlled manner. Theoretical models of SEROA are also expected to quickly develop in the next few years as polarized spectroscopies are becoming a more popular field of research and the sophistication of computational methods is now sufficient to calculate these spectra. Such models will continue to be informative about the mechanisms responsible for SEROA and they will provide a basis for interpretation of measured SEROA spectra, which can otherwise be difficult. A more detailed understanding of the theory of Raman techniques coupled with more advanced instrumentation, such as optical tweezers and stretchers, patch clamps, and so on, will undoubtedly take Raman single molecule detection to a much more sensitive level, and we may expect to see these tools being utilized in the near future for diagnostic analysis and control in different biomolecular fields.

Acknowledgments

Coauthors are greatly acknowledged for collaboration on providing the results of this work, and the Danish National Research Foundation is acknowledged for their financial support for QuP centre (2001–2009).

References

Abdali, S. (2006) *J. Raman Spectrosc.*, **37**, 1341.
Abdali, S. and Blanch, E.W. (2008) *Chem. Soc. Rev.*, **37**, 980.
Abdali, S., Johannessen, C., Nygaard, J. and Norbygaard, T. (2007) *J. Phys.: Condens. Matter.*, **19**, 285205.
Albrecht, M.G. and Creighton, J.A. (1977) *J. Am. Chem. Soc.*, **99**, 5215.
Atkins, P.W. and Barron, L.D. (1969) *Mol. Phys.*, **16**, 453.
Barron, L.D. (2004) *Molecular Light Scattering and Optical Activity*, 2nd edn, Cambridge University Press, Cambridge.
Barron, L.D., Bogaard, M.P. and Buckingham, A.D. (1973) *J. Am. Chem. Soc.*, **95**, 603.

Barron, L.D. and Buckingham, A.D. (1971) *Mol. Phys.*, **20**, 1111.

Barron, L.D., Hecht, L., McColl, I.H. and Blanch, E.W. (2004) *Mol. Phys.*, **102**, 731.

Bizzari, A.R. and Cannistrato, S. (2002) *Appl. Spectrosc.*, **56**, 1531.

Blanch, E.W., Hecht, L. and Barron, L.D. (2003) *Methods*, **29**, 196.

Bour, P. (2007) *J. Chem. Phys.*, **126**, 136101.

Domke, K.F., Zhang, D. and Pettinger, B. (2006) *J. Am. Chem. Soc.*, **128**, 14721.

Efrima, S. (1983) *Chem. Phys. Lett.*, **102**, 79.

Efrima, S. (1985) *J. Chem. Phys.*, **83**, 1356.

Etchegoin, P.G., Galloway, C. and Le Ru, E.C. (2006) *Phys. Chem. Chem. Phys.*, **8**, 2624.

Fleischmann, M., Hendra, P.J. and McQuillan, A.J. (1974) *Chem. Phys. Lett.*, **26**, 163.

Hecht, L. and Barron, L.D. (1994) *Chem. Phys. Lett.*, **225**, 525.

Hecht, L. and Barron, L.D. (1995) *J. Mol. Struct.*, **348**, 217.

Higuchi, S., Ikoma, N., Honma, T. and Gohshi, Y. (1990) in *Proceedings of the Twelfth International Conference on Raman Spectroscopy* (eds S.R. Durig and and J.F. Sullivan), John Wiley & Sons, Chichester.

Hug, W. and Hangartner, G.A. (1999) *J. Raman Spectrosc.*, **30**, 841.

Hug, W., Klint, S., Bailey, G.F. and Scherer, J.R. (1975) *J. Am. Chem. Soc.*, **97**, 5589.

Janesko, B.G. and Scuseria, G.E. (2006) *J. Chem. Phys.*, **125**, 124704.

Jeanmaire, D.L. and Van Duyne, R.P. (1977) *J. Electroanal. Chem.*, **84**, 1.

Johannessen, C. and Abdali, S. (2007) *Spectroscopy*, **21**, 143.

Johannessen, C., White, P. and Abdali, S. (2007) *J. Phys. Chem. A*, **111**, 7771.

Kelvin, W.T. (1904) *Baltimore Lectures on Molecular Dynamics and the Wave Theory of Light*, C.J. Clay and Sons, London.

Kerker, M. (ed.) (1990) *Selected Papers on Surface-enhanced Raman Scattering*, SPIE Optical Engineering Press, Bellingham, WA.

Kneipp, K., Kneipp, H., Abdali, S., Berg, R.W. and Bohr, H. (2004) *Spectrosc. Int. J.*, **18**, 433–40.

Koo, T.-W., Chan, S., Sun, L., Su, X., Zhang, J. and Berlin, A.A. (2004) *Appl. Spectrosc.*, **58**, 1401–7.

Leopold, N. and Lindle, B. (2003) *J. Phys. Chem. B*, **107**, 5723.

Morner, W.E. (2006) *J. Phys. Chem. B*, **106**, 901–27.

Nafie, L.A., Brinson, B.E., Cao, X., Rice, D.A., Rahim, O.M., Dukor, R.K. and Halas, N.J. (2007) *Appl. Spectrosc.*, **61**, 1103.

Nafie, L.A. and Che, D. (1994) *Adv. Chem. Phys.*, **85**, 105.

Nie, S. and Emory, S.R. (1997) *Science*, **275**, 1102.

Yu, G.-S. (1994) Ph.D. Thesis, Syracuse University, New York, USA.

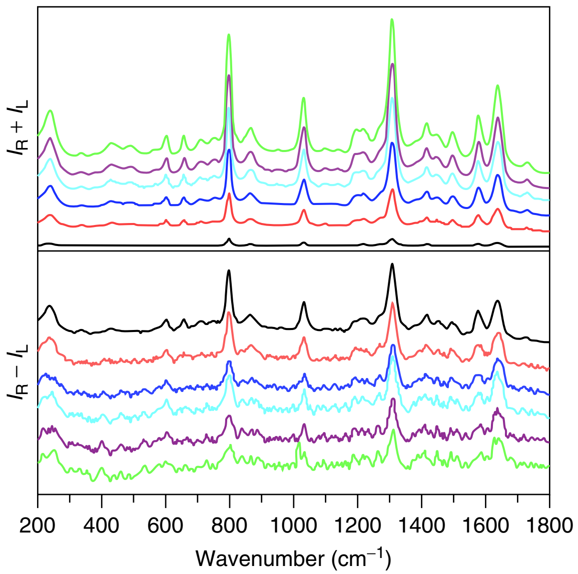

Fig. 19.3 SERS (top) and SEROA (bottom) spectra measured for cytidine with the following time interval of 20 minutes, starting with black (after 5 minutes) and ending with green (after 105 minutes). SERS intensities increase with time, whereas SEROA features associated with large SERS bands, and therefore prone to artifacts, become less intense, revealing weaker features in the SEROA spectra. (This figure also appears on page 631.)

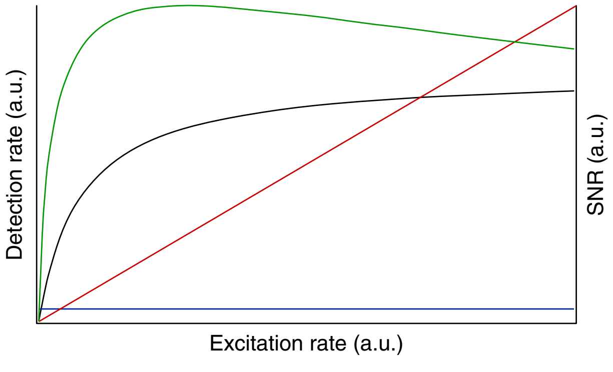

Fig. 20.5 Signal-to-noise ratio as a function of excitation laser intensity according to Eqs (16) and (6) (green). Also plotted are the fluorescence intensity of the single molecule (black), the dark signal (blue), and the laser-induced background (red). (This figure also appears on page 650.)

20
Single Molecule Biophysics with Fluorescence

Michael Prummer and Christian Huebner

20.1 **Introduction** **643**
20.2 **Photophysical Basics** **644**
20.2.1 Energy Levels and Transitions 644
20.2.2 Energy Transfer 646
20.2.3 The Dye Molecule as a Dipole 648
20.3 **Techniques** **649**
20.3.1 Signal-to-Noise Considerations 649
20.3.2 Wide-Field Fluorescence Microscopy Techniques 650
20.3.2.1 Epifluorescence Microscopy 651
20.3.2.2 TIRF Microscopy 651
20.3.3 Confocal Optical Microscopy 652
20.4 **Applications** **654**
20.4.1 Protein Folding 654
20.4.1.1 The Unfolded State 654
20.4.1.2 Single-Molecule Folding/Unfolding Transitions 656
20.4.2 Enzymes 657
20.4.2.1 Static and Dynamic Heterogeneity 657
20.4.2.2 Conformational Changes and Catalytic Activity 660
20.5 **Conclusions** **664**
References 665

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.

Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

ISBN: 978-3-527-40702-6

20.1
Introduction

The first detection of the fluorescence light of a single molecule dates back to 1976 (Hirschfeld, 1976). At that time, the molecule under investigation bore about 100 fluorescent chromophores. With the advent of extremely sensitive photon detectors and effective filters, the detection of the fluorescence of just a single fluorophore became possible in 1990 (Orrit and Bernard, 1990), however, in a crystalline host at liquid helium temperature. In the same year, single fluorophores were observed in solution at room temperature (Shera *et al.*, 1990), paving the way to the application of single-molecule fluorescence on biological systems in their native surrounding. The next milestones for the perception of single-molecule fluorescence as a useful biophysical technique were the demonstration of fluorescent resonance energy transfer on an individual basis (Ha *et al.*, 1996) and the direct observation of the diffusion of a dye-labeled protein in a lipid bilayer (Schmidt *et al.*, 1995). These two key experiments sparked a tremendous development of the technique over the following years, and already in 2001 the coming-of-age of single-molecule spectrosocopy was stated (Kelley, Michalet and Weiss, 2001). Today, the number of groups performing single-molecule fluorescence detection as well as the number of publications in the field is rapidly growing. The pioneers of single-molecule fluorescence have set out their children, and their grand children are starting their own groups. This chapter, therefore, cannot be comprehensive, but rather a selection of prototype experiments demonstrating the enormous potential of single- molecule fluorescence for molecular protein biophysics, with an emphasis on protein conformation in the context of protein folding and catalysis. Owing to space limitations, the fields of single-molecule observations of molecular motors and of membrane diffusion will be omitted.

Therefore, the aspects of the respective applications will be balanced by the methodological developments. The very aim of this approach is to provide the reader with a framework of knowledge that helps decide whether a single-molecule approach to a certain problem might be advantageous, and which method is to be preferred.

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

The basic photophysical principles underlying fluorescence detection, particularly at the high excitation rates necessary to observe the faint light of a single dye molecule, will be outlined in Section 20.2. Section 20.3 will briefly introduce the experimental layouts used, emphasizing their advantages and drawbacks. Having put those ingredients together, Section 20.4 can concentrate on selected applications.

20.2
Photophysical Basics

20.2.1
Energy Levels and Transitions

The fluorescence of single dye molecules is based on the same energy level scheme as ordinary molecular bulk fluorescence, consisting of, in the simplest case, an electronic ground state, an excited state, and a triplet state. In each electronic level, several vibronic states exist. The high repitition rates needed to obtain a measurable signal, however, demand a second look at the energy levels and transitions involved. Besides the levels found in every textbook, higher excited states are shown in the Jablonski diagram in Figure 20.1. Those higher levels exist in the singlet as well as in the triplet manifold. At high laser intensities, in particular in pulsed excitation, further excitations from the states S_1 and T_1 are possible.

Taking first only transitions between the electronic ground state S_0, the first excited singlet state S_1, and the first excited triplet state T_1 into account and neglecting stimulated emission, the following set of rate equations can be derived:

$$\begin{aligned}\dot{n}_1 &= -I\sigma n_1 + k_{\mathrm{rad}} n_2 + k_{\mathrm{ic}} n_2 + k_{\mathrm{risc}} n_3\\ \dot{n}_2 &= I\sigma n_1 - k_{\mathrm{rad}} n_2 - k_{\mathrm{ic}} n_2 - k_{\mathrm{isc}} n_2\\ \dot{n}_3 &= k_{\mathrm{isc}} n_2 - k_{\mathrm{risc}} n_3\end{aligned} \tag{1}$$

Here, n_1, n_2, and n_3 are the occupation probabilities of S_0, S_1, and T_1, respectively, k_{rad} the radiative rate, k_{ic} the rate of internal conversion, k_{isc} the intersystem crossing rate, k_{risc} the reverse intersystem crossing rate, I the excitation intensity, and σ the absorption cross section of the fluorophore.

The quantum yield of fluorescence Φ_f is the ratio of radiative transitions from

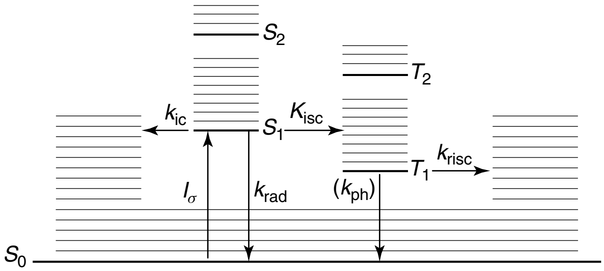

Fig. 20.1 Jablonski diagram showing the electronic ground state S_0, the first and second electronically excited singlet states S_1 and S_2, the first and the second electronically excited triplet states T_1 and T_2, and the most important transitions between the states as detailed in the text.

S_0 to all other transitions from that state:

$$\Phi_f = \frac{k_{\text{rad}}}{k_{\text{ic}} + k_{\text{isc}} + k_{\text{rad}}} = k_{\text{rad}}\tau_f \qquad (2)$$

Here, τ_f is the fluorescence lifetime, which is the experimentally accessible time constant of the decay of fluorescence after pulsed excitation. Therefore, the quantum yield of fluorescence can be determined by measuring the fluorescence lifetime and deriving k_{rad} from the absorption and emission spectra of the dye, making use of the relation between the Einstein coefficients for absorption and spontaneous emission (Strickler and Berg, 1962):

$$k_{\text{rad}} = 2.88 \times 10^{-9} n^2 \frac{\int F(\bar{\nu}) d\bar{\nu}}{\int F(\bar{\nu}) d\bar{\nu}/\bar{\nu}^3} \times \int \frac{\varepsilon(\bar{\nu})}{\bar{\nu}} d\bar{\nu} \qquad (3)$$

Here $F(\bar{\nu})$ is the fluorescence and $\varepsilon(\bar{\nu})$ the absorption spectrum of the dye, with $\bar{\nu}$ the wave number in cm^{-1}. Fluorophores for single-molecule spectroscopy should have quantum yields of fluorescence close to unity to ensure a strong enough fluorescence signal to be detected. Therefore, the intersystem crossing rate as well as the rate of internal conversion should be small.

Under stationary conditions (all time derivatives are zero), the emission rate R as a function of the excitation rate from the set of rate equations (1) is given by

$$R = n_2 k_{\text{rad}} = \frac{k_{\text{rad}}}{1 + \frac{k_{\text{isc}}}{k_{\text{risc}}} + \frac{k_{\text{rad}} + k_{\text{ic}} + k_{\text{ies}}}{I\sigma}} \qquad (4)$$

Owing to increasing population of the triplet state T_1 at high excitation rates, the maximum achievable emission rate with infinitely strong excitation is R_∞:

$$R_\infty = \frac{k_{\text{rad}}}{1 + \frac{k_{\text{isc}}}{k_{\text{risc}}}} \qquad (5)$$

The emission rate R can be rewritten as

$$R = R_\infty \frac{1}{\left(1 + \frac{I_{\text{sat}}}{I}\right)} \qquad (6)$$

with the saturation intensity I_{sat}, which is the excitation intensity where half of the maximum emission rate is reached, given by

$$I_{\text{sat}} = \frac{k_{\text{risc}}}{\sigma} \frac{k_{\text{rad}} + k_{\text{ic}} + k_{\text{isc}}}{k_{\text{risc}} + k_{\text{isc}}} \qquad (7)$$

The saturation of the average emission rate due to triplet excursions is known as the *triplet bottleneck*.

The solution of the set of rate equations (1) with the initial condition of $n_1 = 1$, $n_2 = n_3 = 0$ provides the time evolution of the occupation probability of the first electronically excited singlet state $n_2(t)$, which is related to the fluorescence intensity autocorrelation function or second-order autocorrelation function $g^{(2)}(t)$ defined as

$$g^{(2)}(t) = \frac{< n(\tau) n(t+\tau) >_\tau}{< n(\tau) >_\tau^2} = \frac{n_2(t)}{< n_2 >} \qquad (8)$$

and which reads

$$g^{(2)}(t) = 1 - \frac{k_{\text{isc}}^{\text{eff}} + k_{\text{risc}}}{k_{\text{risc}}} e^{-(k_{\text{rad}}+2I\sigma)t} + \frac{k_{\text{isc}}^{\text{eff}}}{k_{\text{risc}}} e^{-(k_{\text{isc}}^{\text{eff}}+k_{\text{risc}})t} \quad (9)$$

where the effective intersystem crossing rate $k_{\text{isc}}^{\text{eff}}$ is given by

$$k_{\text{isc}}^{\text{eff}} = < n_2 > k_{\text{isc}} \quad (10)$$

The theoretical autocorrelation function of a single dye molecule according to Eq. (9) is shown in Figure 20.2 on a logarithmic timescale. This function drops to zero at the temporal origin, which is intuitively understood taking into consideration that a single quantum system can never emit two photons simultaneously. This behavior is referred to as *photon antibunching*. Thereafter, the autocorrelation function takes a value above one, indicating that the occupation probability of S_1 at this time is higher than on average. This can be understood since the system will always be excited first to S_1, but on average the occupancy $< n_2 >$ is limited by the triplet bottleneck. Because of triplet excursions, a single molecule emits its photons in bunches (cycling in the singlet manifold under photon emission), separated by dark periods (when the molecule is shelved in the triplet state). Therefore, the second term in Eq. (9) is referred to as the *photon bunching* term. If dynamical processes are to be studied by single-molecule fluorescence, care has to be taken that photophysical effects do not lead to misinterpretations.

20.2.2
Energy Transfer

Energy transfer between two or more fluorescent entities is a major subject in single-molecule fluorescence. It plays an important role in light harvesting complexes, and, on the other hand, is the basis of one of the most popular single-molecule techniques, fluorescence resonance energy transfer (FRET), which is

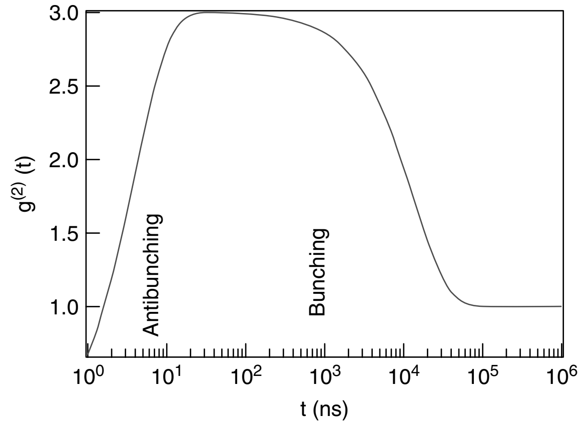

Fig. 20.2 Intensity autocorrelation function of the fluorescence of a single dye molecule under room temperature excitation.

referred to as a *molecular ruler* (Stryer and Haugland, 1967) because of its capability to measure distances between two fluorophores by spectral means.

In FRET, one fluorophore, called *donor*, is excited, and the exitation energy is transferred in a radiationless manner to a nearby fluorophore called *acceptor*. The two chromophores may be chemically identical (homotransfer) or distinct (heterotransfer). The transfer mechanism is based on the dipole–dipole interaction between the two fluorophores in the weak coupling regime, that is, the interaction energy is small compared to the transition energies. The transfer efficiency theoretically derived by Förster (1946) depends on the relative location and orientation of the transition dipoles of the fluorophores, their distance R, the refractive index of the medium n in the vicinity of both, and the overlap integral J. The latter is taken into account for the resonance condition and is the integral of the product of the normalized emission spectrum of the donor $F_D(\lambda)$ and the absorption spectrum (molar extinction coefficient in $\text{cm}^{-1}\text{M}^{-1}$) $\varepsilon_A(\lambda)$ of the acceptor:

$$J = \int_0^\infty F_D(\lambda)\varepsilon_A(\lambda)\lambda^4 d\lambda \tag{11}$$

The transfer rate, that is, the rate with which the excited donor transfers its energy to the acceptor, is given by

$$\begin{aligned} k_t &= 8.8 \times 10^{-23}\text{mol}\frac{1}{\tau_D}\frac{1}{R^6}\frac{\kappa^2\Phi_D}{n^4}J \\ &= \frac{1}{\tau_D}\left(\frac{R_0}{R}\right)^6 \end{aligned} \tag{12}$$

with the distance R taken in Angstrom and R_0 the Förster radius. Here, the relative position and orientation of the dyes is taken into account through the orientation factor κ^2. Values of κ^2 for some cases are shown in Figure 20.3. The orientation dependence of the energy transfer rate may jeoperdize distance measurements. Therefore, a free rotation of the fluorophores on a timescale faster than the fluorescence lifetime is aimed at, leading to an average orientation factor $<\kappa^2>=\frac{2}{3}$. Linkers with several carbon atoms are used to assure unhindered rotation of the dye. However, interactions between the dye and the protein may lead to blocking of the free rotation. Fluoresence anisotropy measurements provide information on the rotational flexibility of the dyes and should therefore always be conducted as a control. Fortunately, even a partial blocking of rotational freedom does not significantly alter the average orientation factor. If, for example, one of the dyes cannot rotate at all while the other one is free to rotate, then the average orientation factor

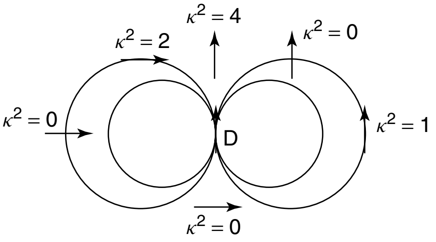

Fig. 20.3 Orientation factors κ^2 for different cases of relative Donor–acceptor placement.

takes the value $< \kappa^2 >= 0.63$ (Berberan-Santos and Prieto, 1988).

In the experiment, the transfer rate is not directly accessible but rather the transfer efficiency, which is the fraction of transferred excitation quanta, and which is related to the distance R as

$$E_t = \frac{1}{1 + \left(\frac{R}{R_0}\right)^6} \tag{13}$$

Whereas in ensemble experiments the transfer efficiency is measured via comparison of donor fluorescence in the presence and absence of the acceptor (e.g. by photo bleaching the acceptor), in single-molecule experiments the fluorescence signals of the donor F_A and of the acceptor F_D are utilized:

$$E_t = \frac{F_A}{\gamma F_D + F_A} \tag{14}$$

The correction factor γ is taking into account the quantum efficiencies of fluorescence of the donor and the acceptor and their respective detection efficiencies. If there is direct excitation of the donor (i.e. not FRET mediated), the acceptor signal needs to be corrected by substraction of that signal. A very robust method to determine the transfer efficiency is based on the fluorescence lifetime of the donor τ_{D0} and τ_{DA} in the absence and presence of the acceptor:

$$E_t = 1 - \frac{\tau_{DA}}{\tau_{D0}} \tag{15}$$

The fluorescence lifetime of the donor without the acceptor is determined in a control sample.

In single-molecule FRET experiments, the photophysics of the dyes has to be treated carefully. Although triplet excursions of the donor simply render the whole donor/acceptor pair dark, the situation is more complex if the acceptor is in the triplet state. If only one excited singlet and triplet state is considered, one would expect that no energy transfer occurs and that all excitation quanta lead to an emission of the donor in this case, resulting in an anticorrelated fluctuation of the donor and acceptor fluorescence. The existence of higher excited states for the acceptor, however, in the singlet as well as in the triplet system, gives rise to other pathways for the excitation energy. In particular, FRET-type energy transfer can result in an excitation of the acceptor to a higher triplet state T_n. Depending on the efficiency of this transfer process, the donor fluorescence may increase, stay constant, decrease, or even completely vanish when the acceptor is triplet shelved (Maus *et al.*, 2001). The latter process is known as *singlet-triplet annihilation (STA)*.

20.2.3
The Dye Molecule as a Dipole

Classically, a dye molecule can be treated as a dipole antenna. Therefore, its absorption and emission have the same anisotropy. Only the component of the electric field parallel to the absorption dipole axis can excite the fluorophore, and the probability curve of photon emission has the same shape as the emission intensity of a dipole antenna. This holds in an isotropic medium as well as in the vicinity of an interface between media with different refractive indices. If a fluorophore is, for example, located close to an interface to a medium with a higher refractive index (imagine a fluorophore in an aqeous surrounding close to a glass coverslip), most of its fluorescence is emitted toward the interface (Figure 20.4). Additionally, the light emitted by a fluorophore is linearly polarized in the dipole direction. The anisotropy as well as the polarization of the emission can be exploited for the determination of the orientation of the dipole in space or even for the relative orientation of the dipoles in a donor/acceptor pair (Hübner *et al.*, 2004).

20.3
Techniques

20.3.1
Signal-to-Noise Considerations

The key prerequisite for succesful single-molecule fluorescence experiments is a high signal-to-noise ratio. The fluorescence signal in single-molecule detection (SMD) is typically the photon count within a given time interval. Owing to the stochastic nature of photon emission, this photon number is Poisson distributed. The noise of the signal therefore is the shot noise of the total number of detected photons. The total number of detected photons is composed of the signal itself and background contributions. The latter can be further divided into background, induced by the excitation laser, and the background signal coming from the detectors in the absence of any source of light, the dark signal.

The laser-induced background originates from elastic Rayleigh and inelastic Raman scattering, as well as from fluorescence in the optical system and the solvent where the molecule under observation is embedded. Its contribution is linearly dependent on the laser intensity. The dark signal, on the other hand, is a constant. The total signal-to-noise ratio can be given by

$$\begin{aligned}\mathrm{SNR} &= \frac{\eta_{\mathrm{det}} \cdot R \cdot T_{\mathrm{int}}}{\sqrt{\eta_{\mathrm{det}} \cdot R \cdot T_{\mathrm{int}} + C_b \cdot I_L \cdot T_{\mathrm{int}} + N_d \cdot T_{\mathrm{int}}}} \\ &= \frac{\eta_{\mathrm{det}} \cdot R}{\sqrt{\eta_{\mathrm{det}} \cdot R + C_b \cdot I_L + N_d}} \cdot \sqrt{T_{\mathrm{int}}} \end{aligned} \tag{16}$$

where R is the emission rate of the fluorophore, T_{int} is the integration time, η_{det} is the detection efficiency, C_b is the constant of proportionality between laser intensity I_L and laser-induced

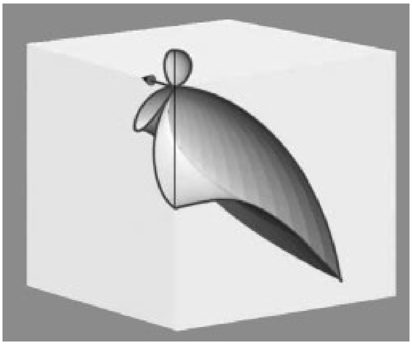

Fig. 20.4 Fluorescence emission characteristics of a dipole close to an interface to a medium with higher refractive index.

background photon rate, and N_d is the dark count rate of the detector.

Saturation of the fluorescence signal leads to an optimum exciation rate for a given integration time where the SNR reaches its maximum, as shown in Figure 20.5.

20.3.2
Wide-Field Fluorescence Microscopy Techniques

In wide-field fluorescence microscopy, the whole field of view of an optical microscope is illuminated to excite fluorescence, and the whole field of view is also imaged onto a camera. The great advantage of wide-field microscopy is that on the one hand, many molecules can be observed in parallel, and that on the other hand, the trace of a moving molecule can be followed. However, the time resolution is limited by the cameras in use.

For single-molecule fluorescence imaging, a very sensitive camera is necessary. If slow processes are to be followed, backthinned cooled charge-coupled device (CCD) cameras are the best choice. These CCDs convert incoming photons into electrons with an efficiency close to unity. However, the readout process of the charges introduces noise of typically a few electrons. Therefore, single photon detection is not possible, and longer exposure times are necessary.

If single photons are to be detected with a camera and a short integration time is wanted, then intensified cameras or electron-multiplying CCDs (EMCCDs) can be used. Intensified CCDs have an image intensifier coupled to the CCD camera. The image intensifier is a multichannel plate photomultiplier (MCP PM) with a phosphor behind the anode. Thus, an impinging photon is first converted into a photoelectron, which is multiplied in the PM, and the electron avalanche is converted back into light. The light from the phosphor is then detected by the CCD. Typical quantum efficiencies of the photo cathodes of image intensifiers are in the 10% to 20% range.

A great breakthrough in camera technology was the introduction of an electron-multiplying or gain register in the CCD. In this register, the single

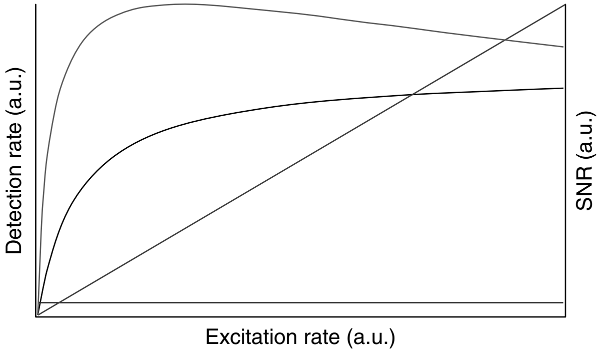

Fig. 20.5 Signal-to-noise ratio as a function of excitation laser intensity according to Eqs (16) and (6) (green). Also plotted are the fluorescence intensity of the single molecule (black), the dark signal (blue), and the laser-induced background (red). (Please find a color version of this figure on the color plates.)

charges are multiplied in an avalanche process principally similar to a PM. Thus, the single electron generated by a single photon is multiplied to a number that is higher than the readout noise and can therefore be detected. EMCCDs have the same high quantum efficiency of photon detection as CCDs close to unity.

The time resolution of a CCD camera is typically limited by the readout process. However, if only a short sequence of images is to be taken, the charges can be first shifted on the CCD chip and subsequently read out. Depending on the size of the image, exposure times in the submillisecond range are possible this way.

With respect to excitation of fluorescence, two different illumination modes can be applied, namely, conventional epifluorescence and total internal reflection (TIR) excitation as discussed in the following sections.

20.3.2.1 Epifluorescence Microscopy

The simplest way to observe the fluorescence of single fluorophores is via the use of a conventional epifluorescence microscope. Here, the whole field of view is illuminated by parallel laser light (Figure 20.6a). Therefore, the laser is focused into the back focal plane of the microscope objective. This technique, however, requires a sample that is thinner than the focal depth of the microscope, since otherwise out-of-focus molecules would give rise to a high background signal. Such thin samples are, for example, molecules tethered on the surface of a coverslip or embedded in a thin polymer film. A cell membrane as well is representing a thin sample, so that epifluorescence microscopy is the method of choice to study, for example, membrane diffusion (Schütz, Kada, Pastushenko and Schindler, 2000).

20.3.2.2 TIRF Microscopy

If the requirement of a thin sample for epifluorescence microscopy is not fulfilled, then illumination of out-of focus molecules can be circumvented by TIR excitation. The excitation laser is totally reflected typically at a glass–water interface and only fluorophores in the evanescent field at the interface are excited. Depending on the angle of incidence and the wavelength of the laser, the evanescent field penetrates approximately 10 nm into the aqueous medium. The small penetration depth is advantegeous with respect to lowering the background, but on the other hand, prevents observation of features further away from the interface.

Two different realizations for total internal reflection fluorescence (TIRF) excitation are used, objective-type TIRF and prism-type TIRF.

In objective-type TIRF, illumination is carried out similar to epifluorescence by

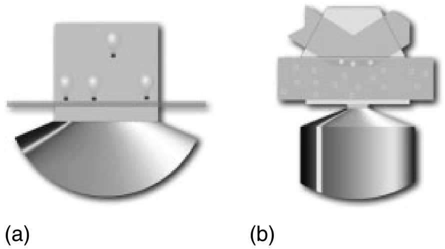

Fig. 20.6 (a) Sketch of epifluorescence and (b) TIRF illumination modes in wide-field microscopy.

focussing the laser in the back focal plane of an oil immersion microscope objective. In contrast to epifluorescence, this focus is offset from the optical axis so that the emerging parallel light is tilted by an angle larger than the critical angle for total reflection at the glass–water interface of 61°. The highest achievable tilt with a numerical aperture (N.A.) = 1.4 oil immersion objective is 67°. Dedicated TIRF objectives have N.A. = 1.45 corresponding to a maximum tilt angle of 73°. The advantage of objective-type TIRF is that the fluorescence is collected through the glass cover slip, in a direction where most of the fluorescence is emitted (cf. Figure 20.4). Thus, a high detection efficiency is assured.

In prism-type TIRF (Figure 20.6b), the laser is directed into a prism and the excitation occurs on the prism surface with TIR. In contrast to objective TIRF, the fluorescence is collected through the aqueous medium. Thus, the detection efficiency is considerably lower. The advantage of prism-type TIRF is the higher flexibility in the choice of the angle of incidence of the laser light.

20.3.3
Confocal Optical Microscopy

The main principle behind confocal optical microscopy invented by Marvin Minsky (1957) is that a point light source and a point detector are both being imaged in the same position in the sample. The point spread function (PSF) of the microscope, that is, the image that is obtained from a point in the sample, therefore is the product of the two PSFs of illumination and detection. Typically, the microscope objective is used for both focusing the laser and detecting the fluorescence in an epi-illumination scheme, and excitation and fluorescence light are seperated in a dichroic beam splitter as can be seen in Figure 20.7. The product of the excitation and detection PSFs defines the confocal volume of about 0.1 fl. The shape of the confocal volume depends on the illumination conditions, but can be approximated quite well by a three-dimensional Gaussian intensity distribution. The advantage of the smaller PSF is not primarily the somewhat higher resolution but rather a considerable reduced background. Laser-induced background signal is in a first approximation proportional to the detection volume. Compared to wide-field microscopy, this volume is reduced by a factor of about $2^{3/2}$.

To obtain an image, the confocal volume is raster scanned across the sample. In contrast to wide-field microscopy, where all points of the image are obtained in parallel in a single exposure, in confocal microscopy the image is built in a serial manner, point by point. This results in a drastically prolonged time to record a single image as compared to wide-field microscopy.

To build the image, either the sample or the confocal volume is moved. For sample scanning, piezo-driven scan stages are used, whereas beam or laser scanning is realized by galvanometer mirror scanners. For the latter, a telecentric lens system generates an image of the scanning mirrors in the back focal plane of the microscope objetive. Thus, a tilt of the scanning mirror results in a shift of the confocal volume in the sample.

Sample scanning has the clear advantage that the light propagates always on the optical axis through the microscope objective. Therefore, aberrations due to tilt of the beam with respect to the optical axis like koma or astigmatism are

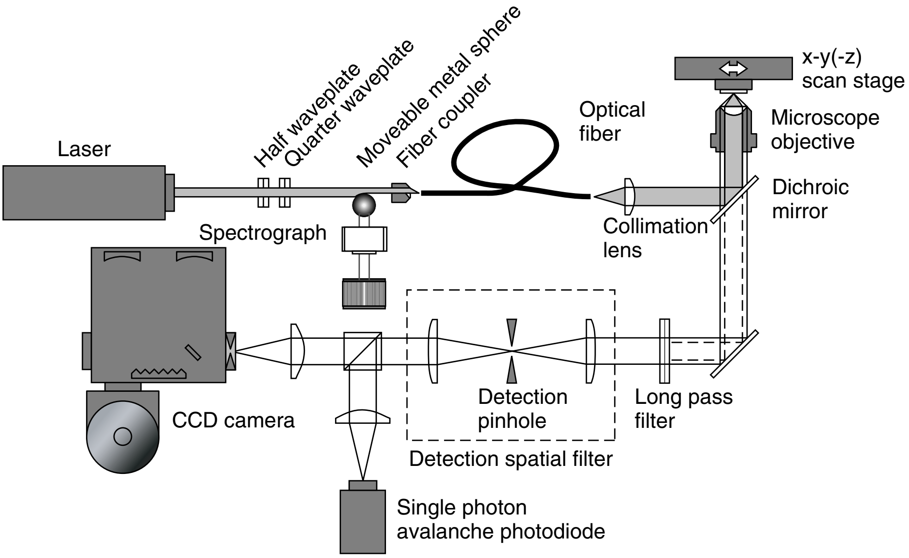
Laser
Half waveplate
Quarter waveplate
Moveable metal sphere
Fiber coupler
Optical fiber
x-y(-z) scan stage
Microscope objective
Dichroic mirror
Collimation lens
Spectrograph
CCD camera
Detection pinhole
Long pass filter
Detection spatial filter
Single photon avalanche photodiode

Fig. 20.7 Schematic drawing of a sample scanning confocal optical microscope used by the authors.

avoided. The allowed size of the image is only limited by the scan range. In contrast, in beam-scanning microscopes the size of the image is limited by the microscope objective. On the other hand, beam scanning has also several advantages: The achievable scan speed is considerably higher, the field of view can be changed by using microscope objectives with different magnifications, and the sample can be mounted in a cryostat.

The point-like light source required for confocal imaging is realized either by the use of an excitation pinhole or by feeding the laser through a single-mode optical fiber. A fiber typically guarantees a perfect Gaussian mode, whereas a pinhole is advantegeous with respect to polarization control.

The other element defining the confocal volume is the detection pinhole. For single-molecule fluorescence experiments, the size of the confocal pinhole is chosen with respect to minimal loss of fluorescence signal. A diameter corresponding to the diameter of the airy disk turned out to be a good choice.

Key elements of the confocal microscope are the dichroic or multichroic beam splitter and the emission filters. Although in the past, holographic notch filters were primarily used to block the backscattered laser light, the development of ever better dielectric filters has rendered them a cost-effective alternative. In particular, sputtered dielectric filters confirm this fact with steep edges, high optical density at the laser wavelength, high transimission in the pass band, and their ruggedness. Care should be taken that the filters are subpixel registered, that is, that they do not introduce beam deviations.

The confocal optical microscope is used in two different regimes for single-molecule experiments: In solution experiments, the confocal volume is placed stationary in the sample solution, and freely diffusing molecules lead to fluorescence bursts when they cross it. Thus, distributions of fluorescence properties of the molecules are obtained. If there is dynamics present on a timescale faster than the transit time, this dynamics can be studied. Dynamics on longer timescales, however, is not accessible with this method.

In immobilized experiments, the Brownian motion is prohibited by appropriate immobilization techniques like embedding in a gel or tethering to a surface. The confocal volume is placed on a particular molecule localized earlier by an image scan, and the (transient) fluorescence is recorded. This way, dynamical processes on all time scales can be studied.

20.4
Applications

20.4.1
Protein Folding

The mechanism by which the unstructured polypeptide chain of a protein finds its unique structure remains one of the great mysteries in protein biophysics. It is believed that there is no single pathway but rather a bunch of ways from the ensemble of unstructured chains to the final three-dimensional structure. This heterogeneity is an appealing target for single-molecule studies. Its ability to resolve subensembles allows for the study of the unfolded state and of folding kinetics and chain dynamics under equilibrium conditions, where either of the folded or unfolded state is weakly populated.

20.4.1.1 **The Unfolded State**

The best-studied subject in protein folding with single molecules is the unfolded state. The interest in the unfolded state arises from the idea that knowledge about its conformation and dynamics under nativelike conditions provides information about the initial steps on the folding pathway. In the first experimental single-molecule FRET study, the well-investigated protein chymotrypsin inhibitor 2 was specifically labeled with tetramethylrhodamine as the donor and Cy5 as the acceptor (Figure 20.8a) (Deniz *et al.*, 2000). The confocal volume was placed in the sample solution, and Brownian diffusion gave rise to a burst of fluorescence each time a molecule was passing by. Each burst was analyzed with respect to the transfer efficiency and added to a histogram. This way, a distribution of FRET efficiencies is obtained. The histograms shown in Figure 20.8b consist of two distinct, rather broad peaks, corresponding to the folded state (high transfer efficiency) and the unfolded state (low transfer efficiency). From the areas under the peaks, the concentrations of the folded and unfolded states can be directly obtained. Strikingly, the center position of the unfolded peak shifts to higher transfer efficiencies with decreasing denaturant concentration, as can be seen in Figure 20.8c. This shift is indicative of a compaction of the unfolded polypeptide chain under nativelike conditions.

Using the same methodology, Schuler *et al.* studied the folding transition of the cold shock protein from *Thermotoga maritima* (Schuler, Lipman and Eaton, 2002). Csp*Tm* was labeled with a donor and an acceptor dye (Figure 20.9a and b), and single-molecule bursts were analyzed for their FRET efficiency and compiled

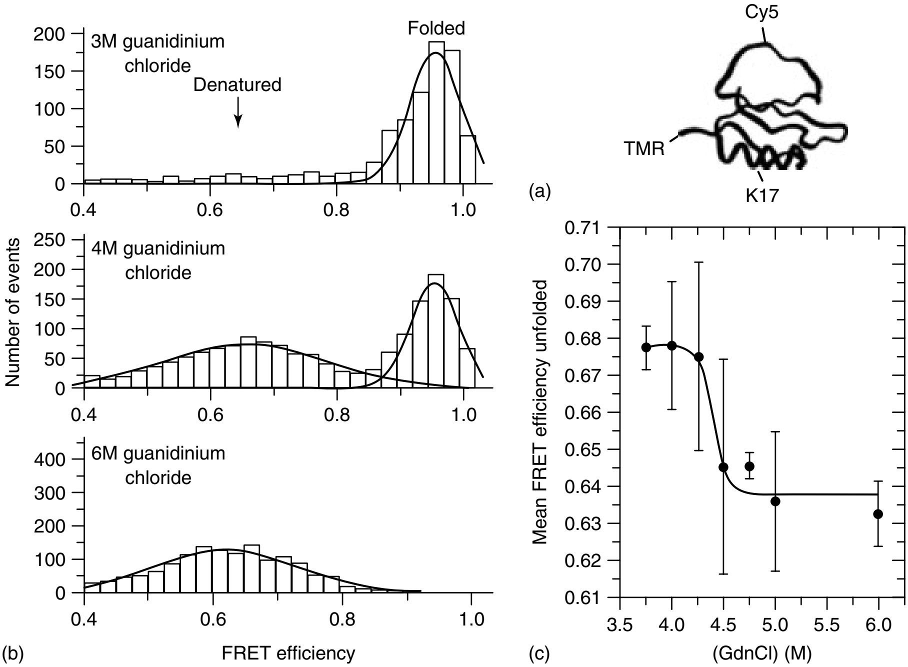

Fig. 20.8 (a) A structural sketch of chymotrypsin inhibitor 2 showing the labeling sites with tetramethylrhodamine as the donor and Cy5 as the acceptor. (b) Energy transfer histograms for different GdnCl concentrations. (c) Mean FRET efficiency as a function of denaturant concentration. Adapted from Deniz *et al.* (2000).

in a histogram shown in Figure 20.9c. Again, a compaction of the unfolded chain is observed (Figure 20.9d). Furthermore, from the finding that the width of the transfer efficiency distribution is shot noise limited it can be concluded that the reconfiguration time is considerably shorter than the transit time of the molecule through the focus. Thus, an upper limit for the reconfiguration time for the polypeptide chain is obtained, which allows for estimating a lower bound for the energy barrier height between the unfolded and folded states of the protein.

Recently, Hofmann et al. have studied the influence of different denaturants, namely, guanidinium chloride (GdnCl) and urea, on the collapse of the unfolded chain of the model protein barstar (Hofmann *et al.*, 2008). Using the same FRET pair as Schuler *et al.*, the subpopulation-resolved transfer efficiencies were measured as a function of denaturant concentration. It was found that for low denaturant concentrations, the chain is more compact in GdnCl than in urea. One possible explanation for this finding is that the charges in the polypeptide chain are shielded in GdnCl, thus reducing electrostatic repulsions. This interpretation was further evidenced by unfolding data in urea and the addition of salt, showing a clear compaction of the chain on addition of salt. This experiment shows that electrostatic interactions are important determinants of the conformation of the unfolded peptide chain.

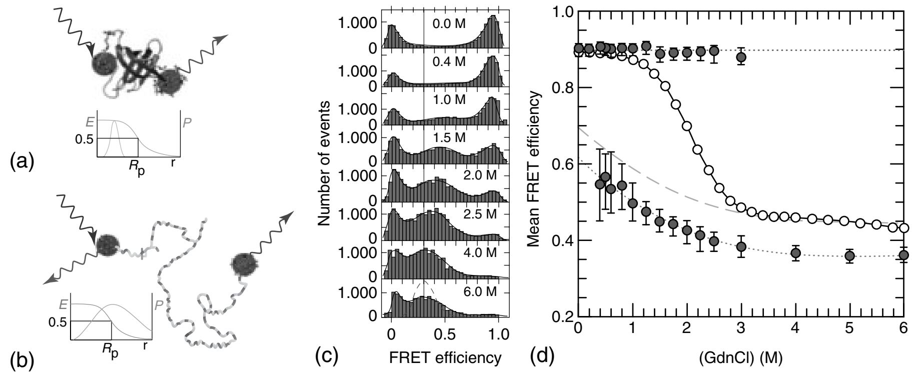

Fig. 20.9 The doubly labeled protein CsP*Tm* in the (a) folded and (b) unfolded states additionally showing the label-distance distribution and the distance-dependent FRET efficiency. The donor dye AlexaFluor 488 is shown in green and the acceptor AlexaFluor 594 in red. (c) The distributions of apparent FRET efficiencies of single CsP*Tm* molecules diffusing through the confocal volume for increasing GdnCl concentrations. The three peaks can be associated to (from left tp right) molecules missing an acceptor dye, unfolded, and folded molecules, respectively. (d) Mean transfer efficiencies derived from the histograms for the folded and unfolded population. The open circles are from ensemble measurements. Adapted from Schuler, Lipman and Eaton (2002). (Please find a color version of this figure on the color plates.)

The dynamics of the unfolded chain of donor/acceptor labeled CspTM was directly monitored by Nettels et al. with a sophisticated technique based on a Hanbury-Brown and Twiss detection scheme (Nettels *et al.*, 2007). With this method, the intensity autocorrelation function can be measured down to the nanosecond timescale by circumventing the dead time of the detectors using a pair of them. Chain dynamics associated with fluctuations of the donor–acceptor distance correspondingly results in intensity fluctuations of the donor and acceptor. Thus, the intensity autocorrelation shows a decay on the timescale of the fluctuations. Figure 20.10 shows the donor and acceptor interphoton time histograms for the unfolded population featuring photon antibunching on the nanosecond timescale typical for a single emitting quantum system. Strikingly, an additional decay of the correlation function is found. The time constant of this exponential decay is 44 ns, being the reconfiguration time of the chain, which is related to the frequency of attempts to cross the energy barrier separating the unfolded and folded states. Thus, single-molecule photon statistics is a valuable tool for the study of chain dynamics associated with the folding reaction.

20.4.1.2 **Single-Molecule Folding/Unfolding Transitions**

A direct observation of folding/unfolding transitions succeeded for the protein adenylate kinase from Escherichia coli, immobilized by encapsulation in surface-tethered vesicles (Rhoades, Gussakovsky and Haran, 2003). The protein is labeled with a donor/acceptor pair, and their fluorescence is transiently recorded as a function of time. To this end, the confocal volume is centered on a vesicle with an encapsulated protein. Surprisingly, rather slow transitions

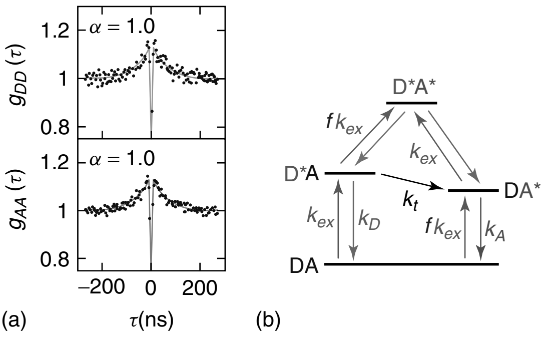

Fig. 20.10 (a) Interphoton time histogram of donor (top) and acceptor (bottom) fluorescence showing antibunching at short times and a correlation on the 100-nanosecond timescale. (b) Photophysical model with the ground states A and D and the respective excited states A* and B*. Fluctuations of k_t lead to the decay in the correlation functions shown in (a). Adapted from Nettels *et al.* (2007).

between the two states are occassionally observed (Figure 20.11a–c). These slow transitions might indicate the search for the folding pathway in the rugged energy landscape between the folded and unfolded state. Further evidence for the existence of local minima in this landscape is provided by the finding of small steps in the transfer efficiency, indicating partial folding/unfolding events (Figure 20.11). Furthermore, there is a broad distribution of step sizes for both folding and unfolding transitions (Fig. 20.11e). The stochastic nature of the walk on the energy landscape renders it impossible to observe those in ensemble experiments, clearly pinpointing the advantage of the single-molecule approach.

20.4.2

Enzymes

Single-molecule investigations of enzymatic activity are very appealing for a number of reasons. The most obvious one is common to all single-molecule experiments: the study of heterogeneity. Single-molecule fluorescence, on the other hand, also allows measurements of enzyme kinetics under equilibrium conditions. Furthermore, under nonequilibrium conditions, the system may remain stationary, that is, the concentration of educts and products stays almost constant, because of the low enzyme concentration in single-molecule studies. Eventually, the investigation of conformational changes associated with enzymatic activity may elucidate the mechanisms of catalysis.

20.4.2.1 **Static and Dynamic Heterogeneity**

The prototype experiment for the study of static and dynamic heterogeneity in enzyme activity was performed by Sunney Xie and coworkers with the enzyme cholesterol oxidase (COx) (Lu, Xun and Xie, 1998). Flavin adenine dinucleotide (FAD), the cofactor of this enzyme, is fluorescent in the oxidized and nonfluorescent in the reduced state. During enzymatic turnover, FAD switches between the oxidized and reduced states.

The COx molecules are embedded in an agarose gel (Figure 20.12a), thus freezing their translational mobility while preserving rotational freedom. The latter is very important to avoid fluctuations in the fluorescence signal owing to slow reorientation. The confocal volume was centered at one COx molecule, and FAD fluorescence recorded as a function of time.

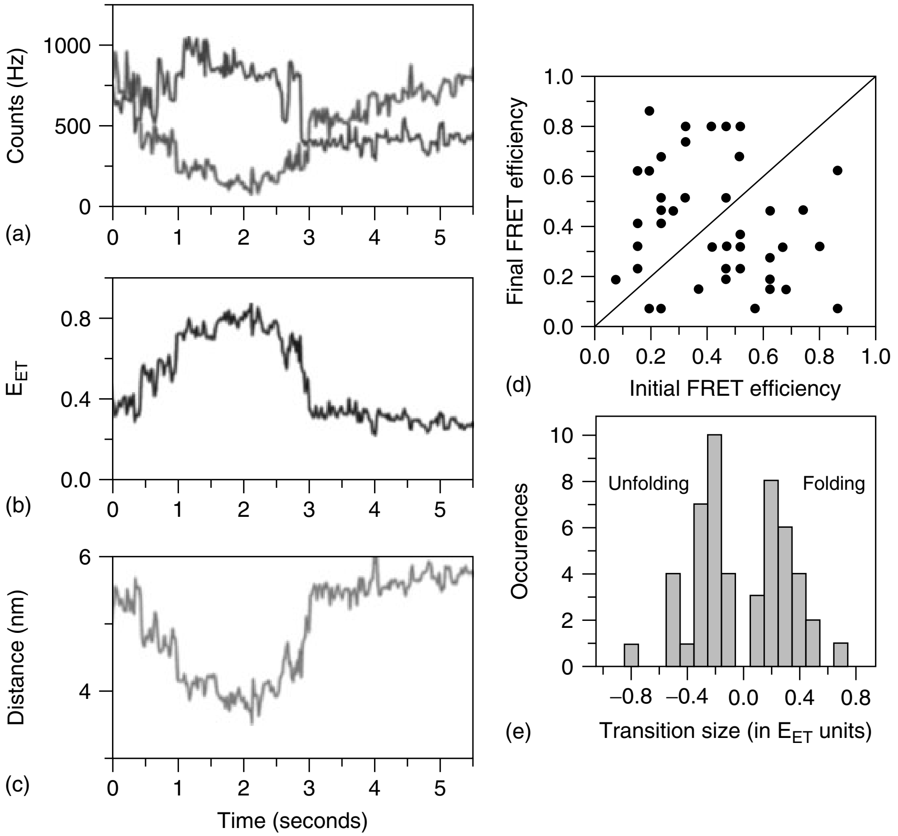

Fig. 20.11 (a) Transient donor and acceptor intensities for a single adenylate kinase molecule inside a vesicle. The anticorrelated behavior is indicative of a change in FRET efficiency shown in (b), which in turn can be related to a donor–acceptor distance change shown in (c). (d) Distribution of initial/final FRET efficiencies for folding (below the diagonal line) and unfolding (above the diagonal line) transitions. (e) Distribution of FRET efficiency step sizes of folding/unfolding. Adapted from Rhoades, Gussakovsky and Haran (2003).

Binary jumps in the transient fluorescence intensity in Figure 20.12b are indicative of cycling between the fluorescent and nonfluorescent states of FAD. Therefore, the catalytic activity is directly accessible, where one turnover is just one cycle of high and low fluorescence intensity. The steady-state kinetics can be obtained from the distributions of the on and off states of the fluorescence corresponding to the two modifications of FAD within the reaction cycle. Hence, both half-cycles of the reaction can be analyzed within one single-molecule experiment.

Strikingly, there is a rather broad distribution of the on times corresponding to the FAD reduction half-cycle for 33 CO*x* molecules, indicative of static disorder, that is, different enzymatic activities of the different enzyme molecules. Now the question arises if there is also dynamic disorder, which is a fluctuation in enzymatic activity over time. To adress this

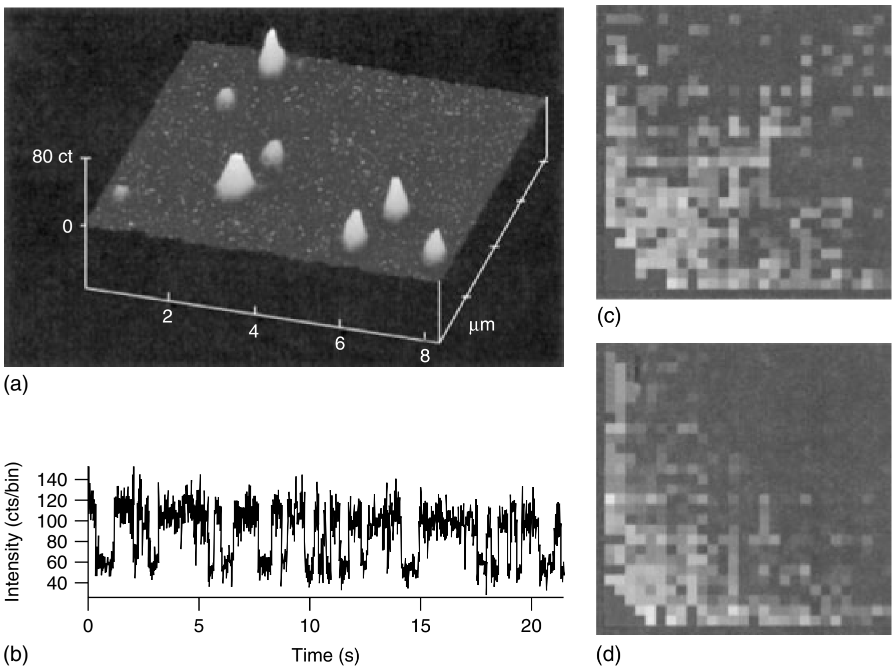

Fig. 20.12 (a) Fluorescence intensity image showing single COx molecules. (b) Transient fluorescence intensity of a single COx molecule showing binary on/off behavior . (c) and (d) 2D conditional probability distribution for a pair of two adjacent on times and for a pair of on times separated by 10 turnovers. Adapted from Lu, Xun and Xie (1998).

question, pairs of on times separated by a defined number of turnovers are considered and their conditional probabilities calculated. If an on time of a certain duration is with a high probability followed by an on time of the same length after this certain number of turnovers, the conditional probability is high. If there is no correlation between the length of the pairs of on times separated by the number of turnovers, the conditional probability is low. In a two-dimensional conditional probability distribution histogram a diagonal feature is indicative of a correlation of the on time durations of on states. Such a feature is visible for adjacent on times (Figure 20.12c), but absent for on times separated by 10 turnover cycles (Figure 20.12d). A covariance analysis even allows for determination of the characteristic timescale for the fluctuations in catalytic activity. Fluctuations of the spectral mean of FAD emission in the absence of cholesterol on the same time scale are interpreted in terms of conformational changes, with the simplest model of interchange between two conformational substates.

Edman and Rigler have used a similar approach for the study of catalytic activity of the enzyme horseradish peroxidase (Edman and Rigler, 2000). Using dihydrorhodamine 6G as a substrate that is nonfluorescent, the catalytic turnover can be followed by the generation of the product rhodamine 6G. The memory landscapes of different proteins are indicative of non-Markovian behavior. Furthermore, oscillations in these memory landscapes can be interpreted by a model where the molecule passes through a number of inactive (conformational) substates and one active substate.

Recently, static heterogeneity was also reported for single enzyme molecules using a fluorescence-based activity assay with much larger statistics (Rissin, Gorris and Walt, 2008). However, this experiment was based on the determination of product concentration and is therefore not capable of accessing the turnovers of a single enzyme molecule.

20.4.2.2 **Conformational Changes and Catalytic Activity**

The prime example for the study of conformational dynamics of single enzyme molecules is the rotation of F_0F_1 ATPase (Kagawa and Racker, 1966). This membrane protein synthesizes adenosine triphosphate (ATP) from adenosine diphosphate (ADP) and phosphate driven by a proton gradient across the membrane. The stepwise rotary motion was first directly observed by videomicroscopy of a fluorescent actin filament attached to the F_1 part (Noji *et al.*, 1997). Although ensemble methods are able to show that rotation occurs, only single-molecule experiments allow for resolution of a sequence of steps. After this key experiment, the rotation was investigated by different single-molecule fluorescence methods (Rondelez *et al.*, 2005; Yasuda *et al.*, 2003; Yasuda *et al.*, 2001; Hisabori, Kondoh and Yoshida, 1999; Iino *et al.*, 2005; Itoh *et al.*, 2004; Masaike *et al.*, 2000; Nishizaka *et al.*, 2004; Adachi *et al.*, 2007; Kaim *et al.*, 2002; Börsch *et al.*, 2002; Diez *et al.*, 2004; Zimmermann *et al.*, 2006; Steigmiller *et al.*, 2004). Kaim et al. show by polarized detection the stepwise rotation of complete F_0F_1 ATPase from *Propionigenium modestum* both in synthesis and hydrolysis (Kaim *et al.*, 2002). To this end, the c subunit was labeled by a fluorescent dye with rigid joining (Figure 20.13a). From the transient fluorescence detected in two channels corresponding to orthogonal polarizations, the polarization P can be calculated according to

$$P = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}} \tag{17}$$

The time-dependent polarization in the presence of ATP and Na^+ in Figure 20.13b shows stepwise changes between three levels, indicating rotational motion in three steps. A data analysis by means of the autocorrelation function of the transient polarization revealed that a model with three pairs of 90°/30° steps as proposed by Yasuda *et al.* (2001) was most consistent with the data.

Single-molecule FRET experiments were also able to show stepwise rotation of F_0F_1 ATPase. Incorporating doubly labeled ATPase (Figure 20.13) in liposomes freely diffusing in solution, Diez et al. recorded FRET efficiency transients during passage of the liposomes through the focal volume (Diez *et al.*, 2004). Those transients show stepwise changes both in hydrolysis and in synthesis between three levels (Figure 20.13d and e). Strikingly, the order of the levels is reversed in hydrolysis/synthesis, thus elegantly demonstrating the reverse of rotation. Recently, Yasuda et al. were able to identify the transiently populated ATP waiting conformation by single-molecule FRET experiments on immobilized ATPase molecules (Yasuda *et al.*, 2003).

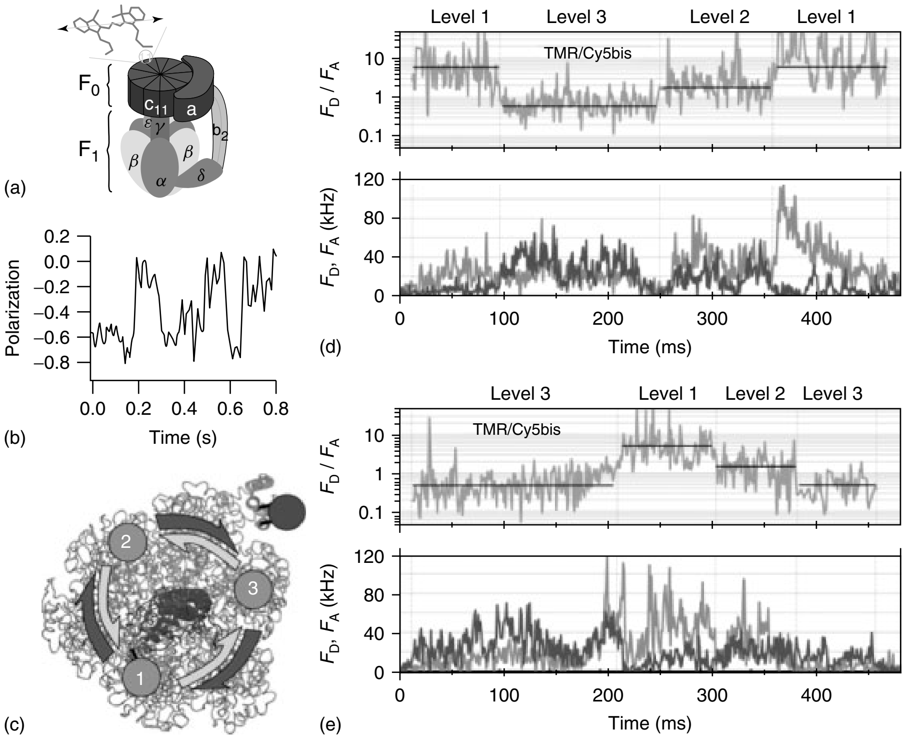

Fig. 20.13 F_0F_1 ATPase labeled at the c subunit (a) shows steps in the polarization transient (b). Donor/acceptor labeled ATPase (c) exhibits jumps between three levels in the ratio between the fluorescence of the donor and acceptor in hydrolysis (d) as well as in synthesis (e) with reversed order. Adapted from Kaim *et al.* (2002) and Diez *et al.* (2004).

The first single-molecule FRET demonstration of conformational dynamics associated with enzymatic activity succeeded with the enzyme *staphylococcal nuclease* (SNase) in the group of S. Weiss (1999). SNase labeled with a donor/acceptor pair showed fluctuations of the transfer efficiency that can be quantitatively analyzed in terms of the autocorrelation function of the transfer efficiency shown in Figure 20.14. The decay in the E_t autocorrelation function can be fit by an exponential with time constants ranging from approximately 10 to 1000 milliseconds, with an average of 41 milliseconds. In the presence of pTp, a known inhibitor of SNase activity, this mean is shifted to 133 milliseconds. Therefore, it can be concluded that the inhibitor affects the conformational flexibility that is essential for SNase function. Via labeling of the SNase with donor and the deoxyribonucleic acid (DNA) substrate with the acceptor, SNase activity can be directly monitored via the association time, which is the time period of high FRET efficiency. A clear difference in association times is observed if the DNA is labeled either at the 5' or 3' end. This difference is in agreement with SNase processivity in the 3' to 5' direction.

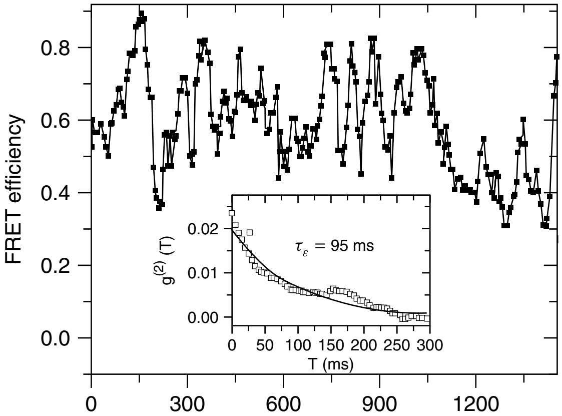

Fig. 20.14 FRET efficiency fluctuations of a single SNase molecule immobilized on a coverslip. The inset shows the autocorrelation function of the the FRET efficiency together with an exponential fit. Adapted from Ha *et al.* (1999).

Quenching of the dye AlexaFluor 488 was exploited for the investigation of conformational transitions of the enzyme dihydrofolate reductase (DHFR) (Zhang *et al.*, 2004). DHFR is a two-substrate enzyme catalyzing the reduction of dihydrofolate (DHF) to tetrahydrofolate (THF) by nicotinamide adenine dinucleotide phosphate (NADPH). To this end, AlexaFluor 488 was specifically attached to the structural loop that closes over the substrates after binding. Single DHFR molecules were tethered on biotynilated cover slips via a biotin–streptavidin–biotin sandwich. In a TIRF illumination wide-field setup, the transient intensities of many DHFR molecules could be recorded in parallel by an intensified CCD camera. On addition of DHF, conformational dynamics manifests itself in intensity fluctuations of the Alexa dye. In the simplest model, DHFR is switching between two conformational states with an opening and closing rate. Thus, using a threshold criterion, open and closed states can be assigned, delivering the dwell times in the respective states. The histograms of the dwell times provide the rates of opening and closing. From the concentration dependence of the closing rate in conjunction with transient ensemble kinetics measurements, a model can be derived where rapid binding of DHF is followed by a conformational change.

Henzler-Wildman *et al.* have investigated the conformational heterogeneity and dynamics of Adk from *aquifex aeolicus* by X-ray crystallography, molecular dynamics simulations, nuclear magnetic resonance (NMR) spectroscopy, and smFRET (Henzler-Wildman *et al.*, 2007). Adk is a two-substrate enzyme catalyzing phosphotransfer between AMP and ATP in the presence of Mg^{2+}. From X-ray crystallography in the presence of the substrates, it is known that a conformational change occurs when the substrates are bound. Adk was specifically labeled with AlexaFluor 488 as the donor and AlexaFluor 633 as the acceptor at the two substrate-binding lids. First, freely diffusing Adk was investigated, and burstwise histograms of the transfer efficiencies were derived. Surprisingly, even in the absence of any substrate Adk shows at least two subpopulations, where one corresponds to the conformation where both lids are opened and the other one corresponds to the closed conformation, which can be concluded comparing the distances from the FRET experiments with crystallographic data. On addition of the inhibitor diadenosin pentaphosphate (AP5A), the equilibrium between the two conformations shifts toward the closed conformation.

To study the dynamics of conformational exchange, labeled Adk was immobilized on a functionalized cover slip via a His_6 tag. Inspite of the considerable noise level that results from the short integration time dictated by the required time resolution transitions between the open and closed states can be observed. The overlap of the transfer efficiency histograms of two conformations renders a single threshold criterion for the discrimination of the two states unusable. Instead, a threshold zone was defined, where a change-of-state required a complete crossing of the zone. After assignment of the open and closed states, respectively, the dwell times in these states are compiled in histograms. From the histograms shown in Figure 20.15a and b the rate constants for opening and closing are derived. The values found for both cases, $k_{\text{open}} = 6500 \pm 500\,\text{s}^{-1}$ and $k_{\text{close}} = 2000 \pm 200\,\text{s}^{-1}$ for the substrate-free and $k_{\text{open}} = 40 \pm 20\,\text{s}^{-1}$ and $k_{\text{close}} = 350 \pm 200\,\text{s}^{-1}$ for the inhibitor-loaded enzyme, turn out to be very similar to the rates of chemical exchange found by NMR.

The assignment of open/closed states based solely on the apparent FRET efficiency is error prone owing to photophysical processes, which may also lead to fluctuations in the apparent FRET efficiency. It needs to be checked explicitly whether anticorrelated intensity changes of the donor and acceptor fluorescence signals occur. Therefore, a correlation analysis was performed. Clear evidence for anticorrelated behavior of donor/acceptor fluorescence intensity is provided by an anticorrelation in the donor/acceptor cross-correlation function as can be seen in Figure 15c. At shorter times, a correlated fluctuation of donor and acceptor fluorescence is found, which is attributed to triplet dynamics of the donor and/or triplet dynamics of the acceptor and efficient STA. A global fit of the three correlation functions, namely, the donor and acceptor autocorrelation and their respective cross-correlation functions, provides the overall rate of distance dynamics and triplet kinetics. The overall rate of distance dynamics of $k_{\text{tot}} = 7000 \pm 2000\,\text{s}^{-1}$ is in good agreement with the rate from NMR and from the states analysis. During catalysis, opening and closing rates of $k_{\text{open}} = 40 \pm 10\,\text{s}^{-1}$ and $k_{\text{close}} = 1000 \pm 500\,\text{s}^{-1}$ are found matching the NMR data on chemical exchange. Taking into account that NMR can only provide information about the occurance and timescale of dynamics, but no spatial information, these experiments nicely demonstrate that single-molecule FRET experiments perfectly complement other biophysical methods.

Single-molecule FRET experiments have also been performed on *E. coli* AdK, where opening and closing rates of $k_{\text{open}} = 120 \pm 40\,\text{s}^{-1}$ and $k_{\text{close}} = 220 \pm 70\,\text{s}^{-1}$ were found for substrate-free Adk (Hanson *et al.*, 2007).

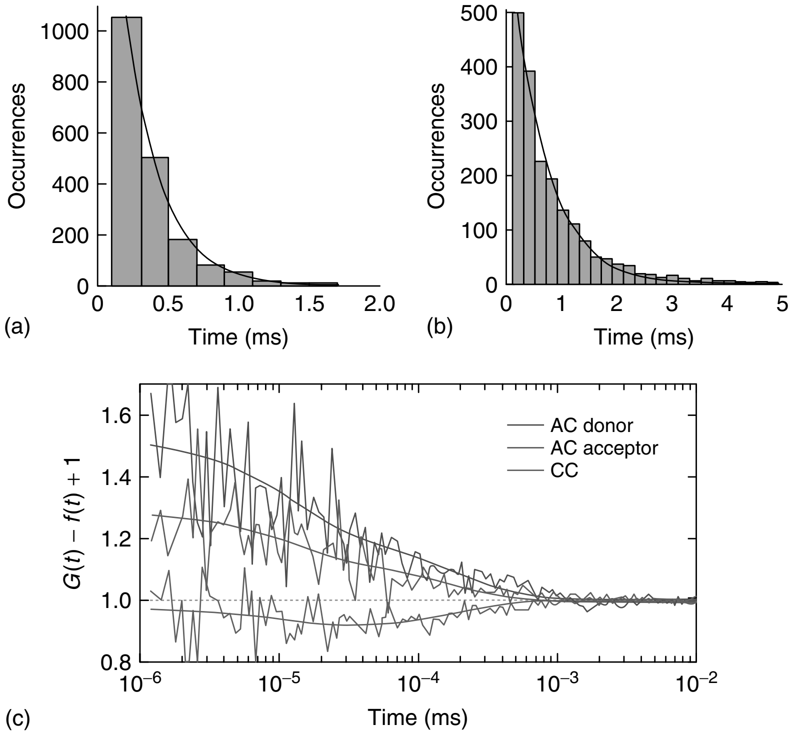

Fig. 20.15 Histograms of (a) low FRET efficiency and (b) high FRET efficiency states in the transient fluorescence of doubly labeled adenylate kinase from aquifex aeolicus attributed to an open and closed conformation. (c) Donor (green) and acceptor (red) autocorrelation and donor–acceptor cross-correlation function (blue) of the same transients. Adapted from Henzler-Wildman *et al.* (2007) (Please find a color version of this figure on the color plates.)

20.5
Conclusions

In recent years, single-molecule fluorescence has proven to be a valuable tool to obtain information on the conformational dynamics and heterogeneity of proteins that is not accessible by ensemble methods. However, the method is still far from being routinely applicable, mostly for problems with photophysics and photochemistry of the dye molecules. There have been several attempts made to improve the photostability of the dyes, but what works out for a particular dye may be even counterproductive for another. More effort has to be made on both synthesizing new dyes with better stability (Weil *et al.*, 2005) and improving the photostability of existing dyes (Rasnik, McKinney and Ha, 2006).

On the other hand, the noisy signal of a single molecule remains a challenge for data analysis, which might be confronted by methods like, for example, Hidden-Markov modeling Markov (2006).

References

Adachi, K., Oiwa, K., Nishizaka, T., Furuike, S., Noji, H., Itoh, H., Yoshida, M., and Kinosita, K. Jr. (2007) Coupling of rotation and catalysis in F-1-ATPase revealed by single-molecule imaging and manipulation. *Cell*, **130**(2), 309–21.

Berberan-Santos, M. and Prieto, M. (1988) Monte carlo simulation of orientational effects on direct energy transfer. *J. Chem. Phys.*, **88**(10), 6341–49.

Börsch, M., Diez, M., Zimmermann, B., Reuter, R., and Gräber, P. (2002) Stepwise rotation of the gamma-subunit of EF0F1-ATP synthase observed by intramolecular single-molecule fluorescence resonance energy transfer. *FEBS Lett.*, **527**(1–3), 147–52.

Deniz, A., Laurence, T., Beligere, G., Dahan, M., Martin, A., Chemla, D., Dawson, P., Schultz, P., and Weiss, S. (2000) Single-molecule protein folding: diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *PNAS*, **97**(10), 5179–84.

Diez, M., Zimmermann, B., Börsch, M., König, M., Schweinberger, E., Steigmiller, S., Reuter, R., Felekyan, S., Kudryavtsev, V., and Seidel, C.A.M., *et al.* (2004) Proton-powered subunit rotation in single membrane-bound F0F1-ATP synthase. *Nat. Struct. Mol. Biol.*, **11**(2), 135–41.

Edman, L. and Rigler, R. (2000) Memory landscapes of single-enzyme molecules. *Proc. Natl. Acad. Sci. U.S.A.*, **97**(15), 8266–71.

Förster, T. (1946) Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Phys.-Berlin*, **6**, 166–75.

Ha, T., Enderle, T., Ogletree, D., Chemla, D., Selvin, P., and Weiss, S. (1996) Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 6264–68.

Ha, T.J., Ting, A.Y., Liang, J., Caldwell, W.B., Deniz, A.A., Chemla, D.S., Schultz, P.G., and Weiss, S. (1999) Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism. *Proc. Natl. Acad. Sci. U.S.A.*, **96**(3), 893–98.

Hanson, J.A., Duderstadt, K., Watkins, L.P., Bhattacharyya, S., Brokaw, J., Chu, J.W., and Yang, H. (2007) Illuminating the mechanistic roles of enzyme conformational dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, **104**(46), 18055–60.

Henzler-Wildman, K.A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M.A., Petsko, G.A., and Karplus, M., *et al.* (2007) Intrinsic motions along an enzymatic reaction trajectory. *Nature*, **450**(7171), 838–844.

Hirschfeld, T. (1976) Optical microscopic observation of single small molecules. *Appl. Opt.*, **15**(12), 2965–66.

Hisabori, T., Kondoh, A., and Yoshida, M. (1999) The gamma subunit in chloroplast F-1-ATPase can rotate in a unidirectional and counter-clockwise manner. *FEBS Lett.*, **463**(1–2), 35–38.

Hofmann, H., Golbik, R.P., Ott, M., Hübner, C.G., and Ulbrich-Hofmann, R. (2008) Coulomb forces control the density of the collapsed unfolded state of barstar. *J. Mol. Biol.*, **376**(2), 597–605.

Hübner, C.G., Ksenofontov, V., Nolde, F., Müllen, K., and Basché, T. (2004) Three-dimensional orientational colocalization of individual donor-acceptor pairs. *J. Chem. Phys.*, **120**(23), 10867–70.

Iino, R., Rondelez, Y., Yoshida, M., and Noji, H. (2005) Chemomechanical coupling in single-molecule F-type ATP synthase. *J. Bioenerg. Biomembr.*, **37**(6), 451–54.

Itoh, H., Takahashi, A., Adachi, K., Noji, H., Yasuda, R., Yoshida, M., and Kinosita, K. (2004) Mechanically driven ATP synthesis by F-1-ATPase. *Nature*, **427**(6973), 465–68.

Kagawa, Y. and Racker, E. (1966) Partial resolution of enzymes catalyzing oxidative phosphorylation. 9. reconstruction of oligomycin-sensitive adenosine triphosphatase. *J. Biol. Chem.*, **241**(10), 2467–74.

Kaim, G., Prummer, M., Sick, B., Zumofen, G., Renn, A., Wild, U.P., and Dimroth, P. (2002) Coupled rotation within single F0F1 enzyme complexes during ATP synthesis or hydrolysis. *FEBS Lett.*, **525**(1–3), 156–63.

Kelley, A.M., Michalet, X., and Weiss, S. (2001) Chemical physics - Single-molecule spectroscopy comes of age. *Science*, **292**(5522), 1671–72.

McKinney, S.A., Joo, C., and Ha, T. (2006) Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys. J.*, **91**(5), 1941–51.

Lu, H.P., Xun, L.Y., and Xie, X.S. (1998) Single-molecule enzymatic dynamics. *Science*, **282**(5395), 1877–82.

Masaike, T., Mitome, N., Noji, H., Muneyuki, E., Yasuda, R., Kinosita, K., and Yoshida, M. (2000) Rotation of F-1-ATPase and the hinge residues of the beta subunit. *J. Exp. Biol.*, **203**(1), 1–8.

Maus, M., De, R., Lor, M., Weil, T., Mitra, S., Wiesler, U.M., Herrmann, A., Hofkens, J., Vosch, T., Müllen, K., *et al.* (2001) Intramolecular energy hopping and energy trapping in polyphenylene dendrimers with multiple peryleneimide donor chromophores and a terryleneimide acceptor trap chromophore. *J. Am. Chem. Soc.*, **123**(31), 7668–76.

Minsky, M. (1957) *Microscopy Apparatus*, US patent 3013467.

Nettels, D., Gopich, I.V., Hoffmann, A., and Schuler, B. (2007) Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc. Natl. Acad. Sci. U.S.A.*, **104**(8), 2655–60.

Nishizaka, T., Oiwa, K., Noji, H., Kimura, S., Muneyuki, E., Yoshida, M., and Kinosita, K. (2004) Chemomechanical coupling in F-1-ATPase revealed by simultaneous observation of nucleotide kinetics and rotation. *Nat. Struct. Mol. Biol.*, **11**(2), 142–48.

Noji, H., Yasuda, R., Yoshida, M., and Kinosita, J.K. (1997) Direct observation of the rotation of F/sub 1/-ATPase. *Nature*, **386**(6622), 299–302.

Orrit, M. and Bernard, J. (1990) Single pentacene molecules detected by fluorescence excitation in a p-terphenyl crystal. *Phys. Rev. Lett.*, **65**(21), 2716–19.

Rasnik, I., McKinney, S., and Ha, T. (2006) Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat. Methods*, **3**(11), 891–93.

Rhoades, E., Gussakovsky, E., and Haran, G. (2003) Watching proteins fold one molecule at a time. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 3197–202.

Rissin, D.M., Gorris, H.H., and Walt, D.R. (2008) Distinct and long-lived activity states of single enzyme molecules. *J. Am. Chem. Soc.*, **130**(15), 5349–53.

Rondelez, Y., Tresset, G., Nakashima, T., Kato-Yamada, Y., Fujita, H., Takeuchi, S., and Noji, H. (2005) Highly coupled ATP synthesis by F-1-ATPase single molecules. *Nature*, **433**(7027), 773–77.

Schmidt, T., Schütz, G.J., Baumgartner, W., Gruber, H.J., and Schindler, H. (1995) Characterization of photophysics and mobility of single molecules in a fluid lipid membrane. *J. Phys. Chem.*, **99**(49), 17662–68.

Schuler, B., Lipman, E.A., and Eaton, W.A. (2002) Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, **419**, 743–47.

Schütz, G.J., Kada, G., Pastushenko, V.P., and Schindler, H. (2000) Properties of lipid microdomains in a muscle cell membrane visualized by single molecule microscopy. *EMBO J.*, **19**(5), 892–901.

Shera, E.B., Seitzinger, N.K., Davis, L.M., Keller, R.A, and Soper, S.A. (1990) Detection of single fluorescent molecules. *Chem. Phys. Lett.*, **174**, 553–57.

Steigmiller, S., Zimmermann, B., Diez, M., Börsch, M., and Grüber, P. (2004) Binding of single nucleotides to H+-ATP synthases observed by fluorescence resonance energy transfer. *Bioelectrochemistry*, **63**(1–2), 79–85.

Strickler, S.J. and Berg, R.A. (1962) Relationship between absorption intensity and fluorescence lifetime of molecules. *J. Chem. Phys.*, **37**(4), 814–22.

Stryer, L. and Haugland, R. (1967) Energy transfer: a spectroscopic ruler. *PNAS*, **58**, 719–26.

Weil, T., Abdalla, M.A., Jatzke, C., Hengstler, J., and Müllen, K. (2005) Water-soluble rylene dyes as high-performance colorants for the staining of cells. *Biomacromolecules*, **6**(1), 68–79.

Yasuda, R., Masaike, T., Adachi, K., Noji, H., Itoh, H., and Kinosita, K. (2003) The ATP-waiting conformation of rotating F-1-ATPase revealed by single -pair fluorescence resonance energy transfer. *Proc. Natl. Acad. Sci. U.S.A.*, **100**(16), 9314–18.

Yasuda, R., Noji, H., Yoshida, M., Kinosita, K., and Itoh, H. (2001) Resolution of distinct rotational substeps by submillisecond kinetic analysis of F-1-ATPase. *Nature*, **410**(6831), 898–904.

Zhang, Z.Q., Rajagopalan, P.T.R., Selzer, T., Benkovic, S.J., and Hammes, G.G. (2004) Single-molecule and transient kinetics investigation of the interaction of dihydrofolate reductase with NADPH and dihydrofolate. *Proc. Natl. Acad. Sci. U.S.A.*, **101**(9), 2764–69.

Zimmermann, B., Diez, M., Börsch, M., and Gräber, P. (2006) Subunit movements in membrane-integrated EF0F1 during ATP synthesis detected by single-molecule spectroscopy. *Biochim. Biophys. Acta*, **1757**(5–6), 311–19.

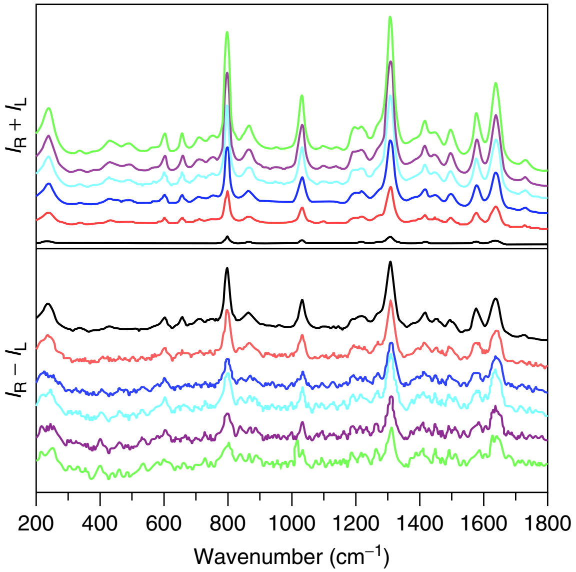

Fig. 19.3 SERS (top) and SEROA (bottom) spectra measured for cytidine with the following time interval of 20 minutes, starting with black (after 5 minutes) and ending with green (after 105 minutes). SERS intensities increase with time, whereas SEROA features associated with large SERS bands, and therefore prone to artifacts, become less intense, revealing weaker features in the SEROA spectra. (This figure also appears on page 631.)

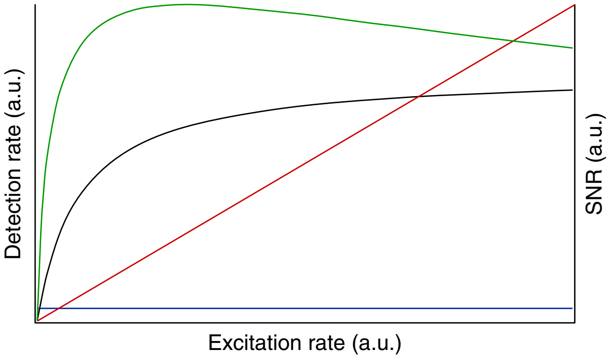

Fig. 20.5 Signal-to-noise ratio as a function of excitation laser intensity according to Eqs (16) and (6) (green). Also plotted are the fluorescence intensity of the single molecule (black), the dark signal (blue), and the laser-induced background (red). (This figure also appears on page 650.)

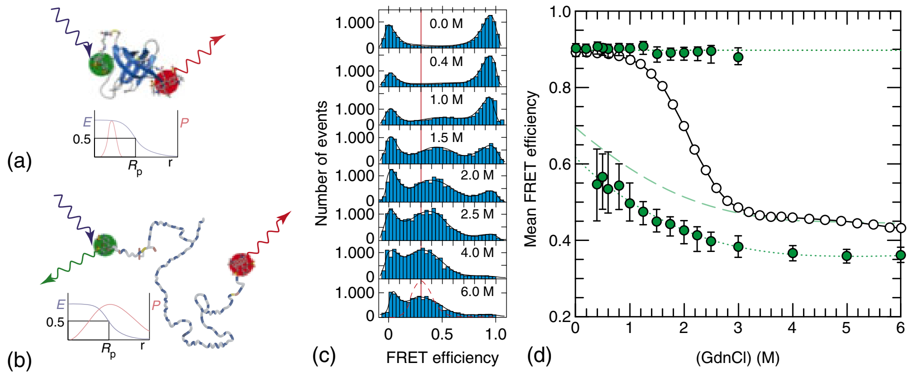

Fig. 20.9 The doubly labeled protein CsP*Tm* in the (a) folded and (b) unfolded states additionally showing the label-distance distribution and the distance-dependent FRET efficiency. The donor dye AlexaFluor 488 is shown in green and the acceptor AlexaFluor 594 in red. (c) The distributions of apparent FRET efficiencies of single CsP*Tm* molecules diffusing through the confocal volume for increasing GdnCl concentrations. The three peaks can be associated to (from left tp right) molecules missing an acceptor dye, unfolded, and folded molecules, respectively. (d) Mean transfer efficiencies derived from the histograms for the folded and unfolded population. The open circles are from ensemble measurements. Adapted from Schuler, Lipman and Eaton (2002). (This figure also appears on page 656.)

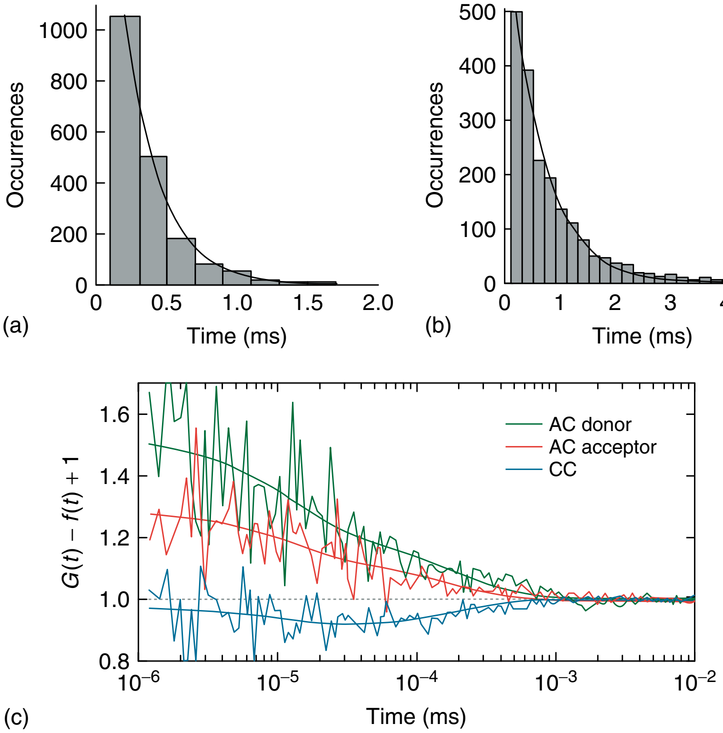

Fig. 20.15 Histograms of (a) low FRET efficiency and (b) high FRET efficiency states in the transient fluorescence of doubly labeled adenylate kinase from aquifex aeolicus attributed to an open and closed conformation. (c) Donor (green) and acceptor (red) autocorrelation and donor–acceptor cross-correlation function (blue) of the same transients. Adapted from Henzler-Wildman *et al* (2007) (This figure also appears on page 664.)

22
Lipid Membrane Technology

Luis A. Bagatolli

22.1 **Introduction** **713**
22.2 **Historical Perspective: Lipid Model Systems** **715**
22.3 **Model Membrane Systems: a Reductionist Approach** **718**
22.3.1 Lipid Monolayers 718
22.3.2 Liposomes 721
22.3.2.1 Small and Large Unilamellar and Multilamellar Vesicles 721
22.3.2.2 Membrane Protein Reconstitution in Lipid Vesicles: Proteoliposomes 723
22.3.2.3 Giant Unilamellar Vesicles 723
22.3.3 Planar Lipid Bilayers 725
22.3.3.1 Supported Lipid Bilayers 725
22.3.3.2 Black Lipid Membranes 726
22.4 **Lipids and their Interactions with Macromolecules** **726**
22.4.1 Lipid–Lipid Interaction 726
22.4.2 Lipid–Protein Interactions 730
22.4.3 Lipid–DNA Interactions 730
22.5 **Lateral Structure of Biological Membranes and Lessons from Model Membranes** **731**
22.6 **Liposomes in Action** **733**
22.6.1 Liposomes and Delivery of Substances 733
22.7 **Concluding Remarks** **734**
Acknowledgments **735**
References **735**

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

22.1
Introduction

As nicely described by Mouritsen in the prolog of his book "Life as a Matter of Fat," lipids have been traditionally considered to play a fairly nonspecific role in biological systems, being rather dull with respect to those described for proteins and genes (Mouritsen, 2005a). This probably arose from the role assigned to lipids in the well-known Fluid Mosaic model proposed by Singer and Nicholson in 1972 to represent the structure and dynamics of biological membranes (Singer and Nicolson, 1972). In this model, lipids were postulated to be merely structural builders, forming a fluid double layer (bilayer) in which few embedded membrane proteins perform particular functions. This perspective has been changed dramatically, particularly in the recent years, during which the importance of lipids in cell function and health has been increasingly recognized (Mouritsen, 2005a). For example, it is well known that there are no genes coding for lipids as such – only the enzymes for their synthesis. If this is the case, how does a particular biological membrane end up with a particular combination of hundreds of different lipid species assembled in particular molar ratios? Is there a coherent code for lipids (as for genes and proteins) still hidden and waiting to be discovered (Hilgemann, 2003)? Certainly, the answers are not yet known. These aforementioned and still open questions represent in my opinion one of the most interesting unsolved aspects of biological systems.

Lipids comprise a diverse group of compounds (e.g. phospholipids, sphingolipids, sterols, triglycerides, fatty acids) that participate in essential ways in diverse biological processes. Along with proteins, genes (DNA and RNA), and sugars, these molecules are fundamental building blocks of all living matter. For example, it is clear that lipids play a key role as structural components of cell membranes, but, *together* with membrane proteins and sugars. Lipids themselves also serve as energy storage sources, act as receptors (Velter *et al.*, 2007), and participate in diverse signaling pathways in biological systems (Hilgemann, 2003; Won, Singh, and Singh, 2007).

One very distinctive and peculiar property of lipid molecules is the self-aggregation of these molecules in water. This phenomenon is driven by an increase in the entropy of water in the

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

presence of lipids, a phenomenon known as the *hydrophobic effect* (King and Marsh, 1987). In consequence, lipids form complex multimolecular aggregates with characteristic physical properties (e.g. micelles, bilayers, hexagonal, and cubic phases) (Epand, 2007; Seddon and Cevc, 1993). These different structures play different functional and structural roles in biological systems.

Even though many different lipid structures exist in nature, this chapter will deal mainly with bilayer lipid structures. As mentioned above, bilayers are two-lipid-molecule-thick structures that together with particular proteins and sugars comprise biological membranes. Lipid bilayers themselves show different physical properties depending on their composition and the existing environmental conditions (e.g. temperature, pH, ionic strength). Correspondingly, the remarkable diversity of the molecular species present in biological membranes (lipids, proteins, and sugars) generate a broad range of physicochemical properties that are generally connected with membrane function. Connected to this, the compositional ratio between lipids and proteins is an important parameter to take into account (Marsh, 2008).

Despite the fact that the classical picture of a biological membrane includes a lipid bilayer with few diffusing proteins (Fluid Mosaic Model, see above), it is fair to also say that in many cases biological membranes more resemble proteins sheets with embedded lipid bilayers (the protein–lipid ratio is high). These two last pictures represent extreme cases of a broad range of different compositional possibilities existing in biological membranes and must be taken into account when reductionist approaches are considered.

Given the compositional complexity of biological membranes, reductionist approaches are normally used in order to explore physical (structure and dynamics) and chemical properties of membranes (i.e. the field of membrane biophysics). For example, several model membrane systems (presented in Section 22.3) composed of key lipid species, with or without particular membrane proteins, have been extensively studied for many decades. By using such model lipid systems one can tune, for instance, the membrane compositions going from a very simple lipid mixture to lipid mixtures extracted from biological membranes, or even generate these model systems with the full composition of the native biological membrane (Baumgart *et al.*, 2007; Bernardino de la Serna *et al.*, 2004; Montes *et al.*, 2007). The possibility of systematically increasing the membrane compositional complexity makes it possible to perform interesting comparative studies using an array of different experimental techniques, under very well-controlled and specific environmental conditions (Bagatolli, 2006; Baumgart *et al.*, 2007). Studies performed in theses model systems include lipid–lipid and also lipid–protein and lipid–DNA interactions (Section 22.4). The information obtained from these studies not only provides important foundations about the lateral structure of biological membranes (discussed in Section 22.5) (Silvius, 2005a, 2003; Simons and Vaz, 2004), but also some interesting applications of the lipid model systems over a very broad range of fields (discussed in Section 22.6) (Lasic, 1998). For example, reports regarding the delivery of substances using liposomes are extraordinarily wide-ranging with important applications in science and industry, including molecular biology, biophysics, immunology, cosmetic, pharmaceutical, and food sciences, as well as medical applications (Lasic, 1998). The breadth of liposome applications is such that these lipid entities are also used in marine fish larviculture to encapsulate lipids and amino acids in order to feed the live organisms used for fish larvae food (Barr and Helland, 2007)!

The aim of the following sections is to describe different model lipid systems and some of their applications, as well as a historical perspective of the lipid model systems, information about preparation methods, and some examples of their use in basic sciences and technology. This chapter represents a selective and personal viewpoint about the use of lipid model systems and is not intended to present an exhaustive review of results obtained in the many different and specialized topics involving lipids. The enormous amount of information already reported in the literature would make it impossible to summarize it all in one chapter, as reflected in the fact that the word lipid presently retrieves more that 800,000 citations in PubMed!

22.2
Historical Perspective: Lipid Model Systems

Lipid model systems help us to understand the basic structural and dynamic aspects of biological membranes. From a historical perspective the earliest reports of lipid model systems are, in modern terms, monomolecular lipid films spread at the air–water interface, that is, so-called Langmuir films. The earliest record (reported as oil films floating in water) comes from cuneiform-written Babylonian clay tablets from the Hammurabi period (eighteenth century BC) (Dynarowicz-Latka, Dhanabalan, and Oliveira, 2001). Babylonians used it as one of their forms of divination, based on pouring oil on water and observing the shapes formed with the spreading of the oil. This application was also used in Greek, Chinese, and ancient Japanese cultures (Dynarowicz-Latka, Dhanabalan, and Oliveira, 2001). The most cited report about applications of lipid films is perhaps the calming effect of ripples on the water surface caused by spreading oil, which was already observed by sailors in ancient times and described by Plutarch, Aristotle, and Plinius (Dynarowicz-Latka, Dhanabalan, and Oliveira, 2001). Pouring oil to calm stormy seas is believed to have come from the period 429–651 AD (Fulford, 1968). The scientific investigation of this phenomenon (i.e. using the scientific method) started in Franklin, Brownrigg, and Farish (1774), followed by contributions Pockels (1894), as reported by Rayleigh (1899) and Langmuir provided the first evidence of the monomolecular nature of the films in 1917 (Langmuir, 1917). Such monomolecular films are also known as *Langmuir–Blodgett* films due to the first detailed description of sequential monolayer transfer to a solid support by Blodgett (1935). The term *Langmuir film* is normally reserved for a floating monolayer at the air–water interface while Langmuir–Blodgett films refer to monomolecular films supported in a solid substrate. Incorporation of a transmembrane protein into a supported monolayer was first reported by McConnell in 1983 "... to study specific binding of cytotoxic T cells to well-defined model membranes" (Nakanishi, Brian, and McConnell, 1983). This work was published almost 50 years after Blodgett's contribution. Additionally, an extensive characterization of the lateral structure of lipids using this model system was reported by the McConnell laboratory (McConnell, Tamm, and Weis, 1984). Successive reports about formation of supported lipid *bilayers* on solid supports from Langmuir films were first reported between 1983 and 1985 (Tamm and McConnell, 1985; Tancrede, *et al.*, 1983). Supported planar bilayers are currently used to study a broad range of scientific and practical applications of membrane-related phenomena (Sackmann, 1996) and thus constitute an essential bilayer model system.

Langmuir films were used in a historically "key" experiment connected to our contemporary vision of a biological membrane. In 1925, two Dutch physicians, E. Gorter and F. Grendel, compared the area of "total" lipid extracts from erythrocyte membranes, measured in a Langmuir trough, with the average surface area of red blood cells (Gorter and Grendel, 1925). In their paper the authors report that "... cromocytes of different animals are covered by a layer of lipoids just two molecules thick." Disregarding the fact that lipid bilayers are not biological membranes and a pair of compensating experimental mistakes[1] are present in this article (Bar, Deamer, and Cornwell, 1966), the relevance of the authors' observation cannot be underestimated. The last was the first experimentally supported statement that cell lipids are arranged in a bilayer configuration (Przestalski *et al.*, 2000).

Liposomes (formed by most phospholipids including some sphingolipids when dispersed in aqueous solutions) can be defined as *spherical vesicles*, essentially an aqueous core surrounded by a lipid bilayer. The discovery of liposomes is attributed to (Bangham and Horne, 1964). In their classical paper they provide clear evidence, using electron microscopy, suggesting that dispersion of phospholipids in water formed such baglike structures. In the same paper they argue for first time that the integrity of a bimolecular leaflet of phospholipids depends as much on the presence of water as it does on the precise composition and structure of the participating lipid molecules (Bangham and Horne, 1964). The fundamental structure of this model lipid system agrees very well with the structure and some of the physical properties (e.g. the permeability) proposed previously in earlier models of biological membranes, for example, Gorter and Grendel, and especially that of Danielli and Davson (Danielli and Davson, 1935; Gorter and Grendel, 1925; Robertson, 1981). Subsequent reports from Bangham's laboratory show that these entities are indeed osmotically active, and much more permeable to water than to simple ions or glucose, also demonstrating that the sign and magnitude of the surface charge seemed to determine their relative permeability to cations and anions (Bangham, 1993). During this period, liposomes were adopted as "classic model systems," used to mimic processes occurring in biological membranes, and

---

1) In Gorker and Grendel experiments the area of the red blood cell was underestimated, and due to the roughness of the extraction techniques used at that time they did not extract all the lipids from the erythrocyte membrane. The interesting remark is that the area occupied by the proteins was not considered in their experiments, since there was no concept about integral membrane proteins at that time.

became key players in an enormous range of biotechnological process (Lasic, 1998). Although the credit for discovery of lipid vesicles is given to Bangham and Horne, reports about the colloidal behavior of lecithin and other phospholipids can be found much earlier in literature (Lasic, 2000; Segota and Tezak, 2006). In fact, hydration of dried extracts of lecithin from egg yolk or brain and the subsequent formation of myelin figures (as happens during liposome preparation) were documented in Neubauer (1867); Virchow (1854). Additionally, detailed description of lipid solutions as colloidal systems were obtained during the early 1920s (Lasic, 2000). However, it took more than 100 years and an electron microscope (Bangham and Horne, 1964) to realize that these lipid colloidal suspensions are indeed closed lipid bilayer structures encapsulating the same liquid medium in which they freely diffuse (Lasic, 2000).

Other colloidal suspensions composed of amphiphilic substances (e.g. fatty acids or their salts; lysolipids) are called *micelles*. Micelles are particular aggregates of surfactant molecules dispersed liquid media. A typical micelle in aqueous solution forms an aggregate with the hydrophilic "head" regions in contact with surrounding solvent, sequestering the hydrophobic tail regions in the micelle center. Upon addition of an organic solvent, these micelles can swell, and if the amount of organic solvent exceeds that of the water, the surfactant molecules can invert to form inverted micelles. Micelles are closely connected to soaps, since solutions of these compounds in water will likely contain these structures. Even though the earliest recorded evidence of the production of soap-like materials dates back to around 2800 BC in ancient Babylon (Willcox, 2000), structures and physical properties of micelles were not scientifically studied until the beginning of the twentieth century. Pioneering work with micelles was carried out by James William McBain at the University of Bristol (Rideal, 1953). As early as 1913 he postulated the existence of "colloidal ions" to explain the good electrolytic conductivity of sodium palmitate solutions (McBain, 1913). Micelle formation is essential for the absorption of fat-soluble vitamins and lipids in biological systems. Bile salts formed in the liver and secreted by the gall bladder allow micelles of fatty acids to form, allowing the absorption of lecithin and lipid soluble vitamins (A, D, E, and K) by the intestine. Besides the scientific interest in performing basic types of scientific studies, micelles have a major technological application as emulsifiers (e.g. detergents) with special applications, for example, in emulsion polymerization (Capek, 2001).

Related to freestanding bilayer model systems (e.g. liposomes), a beautiful technique for preparing credible model bilayers using lipid molecules, the so-called black lipid membrane (BLM), was introduced in Mueller *et al.* (1962). This system consists of a freestanding bilayer (generally composed of lipids and proteins) formed across a hole in an inert solid support (Teflon, for example) (Tien and Ottova-Leitmannova, 2003a; Winterhalter, 2000). The name of this model membrane system is connected to previous observations of Newton (Newton, 1704) of black spots in single soap films (the so-called Newton Black films) (Mueller *et al.*, 1962; Tien and Ottova-Leitmannova, 2003a). These black spots observed in film of soaps correspond to bimolecular soap leaflets that are the limiting unit of the smectic, mesomorphic soap phase (Mueller *et al.*, 1962). This conclusion was based on experimental observations of Johonnott, Perrin and Bragg during the 1900–1940 period (Bragg, 1939; Johonnott, 1899, 1906; Perrin, 1918). The same black spots are observed during formation of lipid bilayers under water (the BLMs) (Tien and Ottova-Leitmannova, 2003a). As the film starts to thin out, colored reflections are observed and finally a gray–black spot starts to spread over the film. This annealing is due to the interference of the reflected light from the water–lipid and phase-shifted light from the underlying lipid–water interface, which becomes destructive as the pathway through the membrane and back becomes negligible in comparison to the wavelength (Winterhalter, 2000). From a historical perspective and since the early 1960s, BLMs have become a key bilayer model system to measure transport-associated phenomena across membranes. For example, this model system has been extensively used to perform electrophysiological experiments, measuring ion transport across membranes under controlled compositional and environmental conditions (Morera *et al.*, 2007).

22.3
Model Membrane Systems: a Reductionist Approach

The reductionist approach helps us to understand the nature of complex systems by reducing them to the interactions of simplified descriptions of their parts. Since biological membranes are compositionally complex systems, this approach is extensively used to understand the basis of molecular and supramolecular interactions occurring in membranes, including their consequences on structure, dynamics and ultimately their function. Different model membrane systems, composed of representative molecules, are available (Figure 22.1). In view of the fact that the basic structure of biological membranes is strongly related to lipid bilayers, several bilayer model systems have been developed and are broadly used to study membrane-related phenomena.

22.3.1
Lipid Monolayers

Lipid monolayers have a broad application in the basic sciences (e.g. chemistry, biology, physical-chemistry, polymer science to mention a few). This section, however, will only discuss selected examples related to particular applications of lipid monolayers in biologically relevant problems. Even though this model represents ''half of a bilayer'' (one molecule thick), it is still extensively used to mimic basic molecular and supramolecular interactions in membranes. In fact, a particularly broad variety of different lipid molecules (single species or their mixtures) have been studied using this technique (Brown and Brockman, 2007).

Two main types of model systems are generally studied: (i) Langmuir films (at the air–water interface) and (ii) supported lipid monolayers (on mica or other solid supports). Both systems comprise monomolecular film of lipids. For example, studies reported in these systems include (to mention a few) the lateral structure of compositionally different lipid films (Brown and

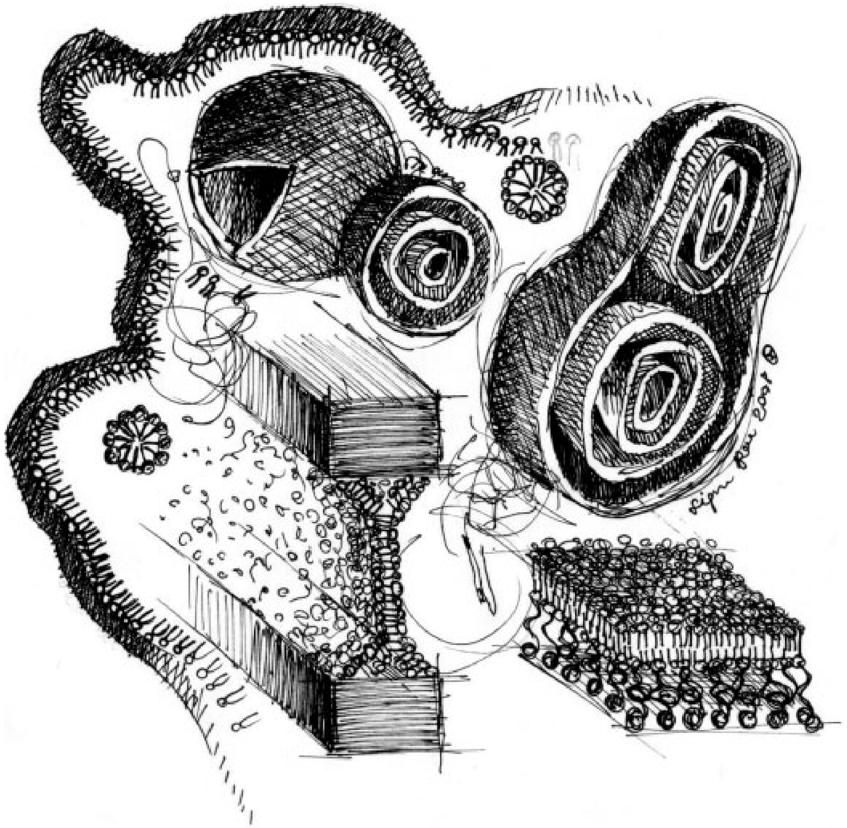

Fig. 22.1 "*Da jeg fandt lykken.*" Sketches with pencil (from Signe Boe private art collection, 2008). Representation of the different lipid models systems.

Brockman, 2007; Maggio, 1994; McConnell, Tam, and Weis, 1984; Norlen *et al.*, 2007), penetration studies of proteins, peptides or other drugs (anesthetics, for example) into a monolayer (Choi, Oh, and Lee, 2001; Fidelio *et al.*, 1986; Maget-Dana, 1999), stability of peptide or protein monolayers (Ambroggio *et al.*, 2004; Churbanova *et al.*, 2006), and enzymatic actions of lipases (Bianco *et al.*, 1991; Pieroni *et al.*, 1990). Studies including rheological, topological, electrical, and mechanical properties of monomolecular films are also accessible using these systems (Brown and Brockman, 2007; Leneweit *et al.*, 2008). For instance, studies of surface pressure and interfacial electrical potential as a function of average cross-sectional molecular area in Langmuir films provide insights into lipid hydrocarbon chain ordering, lateral compressibility/elasticity, and electric dipole effects under various conditions including those that well approximate one leaflet of a bilayer (Brown and Brockman, 2007; Leneweit, *et al.*, 2008).

Importantly, lipid monolayers have a connection with relevant structures existing in biological systems. For example, it is well known that a proteolipid surfactant material exists in our body, for example in the respiratory airways, having important biological functions (Notter, 2000), which is required to avoid lung collapse (by changing the surface tension on the surface of the alveoli) during breathing cycles (Notter, 2000). The organization of lung surfactant at the air–tissue interface has been modeled using monolayers (either Langmuir films or supported monolayers) composed of particular lipid mixtures (with or without proteins), surfactant lipid extracts or the native material (Nag *et al.*, 1998, 2002; Paul, Hassett, and Reiss, 1977; Piknova, Schram, and Hall, 2002; Wang *et al.*, 2007; Yan *et al.*, 2007). In these experiments the role of different components of the material is evaluated under controlled environmental conditions (Casals, 2001; Perez-Gil, 2001).

The preparation of Langmuir films is straightforward and very well described in the literature (Gaines, 1966). Generally, organic lipid solutions are directly spread at the air–water interface in an apparatus that has been traditionally referred to as a *Langmuir trough* (Gaines, 1966). This device consists of a trough, usually made of hydrophobic materials like Teflon to contain the water phase, that includes movable barriers that span the water surface (Dynarowicz-Latka, Dhanabalan, and Oliveira, 2001). Movement of the barriers allows the monomolecular film to be compressed laterally at the air–water interface. The change in the surface tension is generally monitored with a Langmuir balance or a Wilhelmy plate (Cruz and Perez-Gil, 2007; Dynarowicz-Latka, Dhanabalan, and Oliveira, 2001; Gaines, 1966). Plots of the change in surface pressure (π; a measure of the decrease in surface tension, i.e. a two-dimensional analogue of bulk pressure) as a function of the average area per molecule on the aqueous subphase are then obtained at fixed temperature, thus called *pressure-area isotherms* (Cruz and Perez-Gil, 2007; Dynarowicz-Latka, Dhanabalan, and Oliveira, 2001; Gaines, 1966). Under optimized conditions, the monolayer may be amenable to transfer onto hydrophilic/hydrophobic solid supports at different lateral pressures either by the vertical dipping method or horizontal touching methods (Cruz and Perez-Gil, 2007; Dynarowicz-Latka, Dhanabalan, and Oliveira, 2001).

There is an array of different experimental techniques that can be use to study these model systems. For Langmuir films, the most classical measurements perhaps are the pressure–area and surface potential–area isotherms, but also rheological measurements including also measurement based in spectroscopic methods such as Brewster angle, fluorescence microscopy, UV–vis, IR, Raman, and Second harmonic generation (Dynarowicz-Latka, Dhanabalan, and Oliveira, 2001). For supported monolayers, the most commonly applied techniques are atomic force microscopy, ellipsometry, quartz crystal microbalance, grazing angle X-ray diffraction, X-ray and neutron reflectivity, fluorescence microscopy, and near-field microscopy (Dynarowicz-Latka, Dhanabalan, and Oliveira, 2001; Hollars and Dunn, 1998). Interestingly, new applications using mass spectrometry were recently described to ascertain the spatial distribution of molecules on supported monolayers showing phase separated domains (Sostarecz *et al.*, 2004; Zheng *et al.*, 2007). This spatially resolved compositional information is extremely valuable, for example, to define complete phase diagrams of lipid mixtures or localize particular lipid components (or other molecules) into different lipid domains (Kraft *et al.*, 2006). Lastly, different techniques based on computer simulations have also been applied to understand molecular (molecular dynamics simulations approach) or supramolecular (e.g. coarse-grained molecular dynamics approach) details of these model systems (Giner-Casares *et al.*, 2008; Pink, 1984). Structural and dynamic features of particular model systems can be predicted using these computational approaches. In general, the obtained information strongly promotes further development of new experimental strategies in order to corroborate the predictions from the simulations.

22.3.2
Liposomes

As was discussed in Section 22.2, liposomes (lipid vesicles) are one of the most popular model systems containing lipid bilayers. They are classified with respect to their size and the number of lipid lamellaes (single bilayers) existing in their structure (Lasic and Papahadjopoulos, 1995). For example, liposomes containing a single bilayer, known as *unilamellar vesicles*, are classified with respect to size as small, large, or giant. When the liposome is formed from multiple bilayers, the vesicles are called *oligolamellar* or *multilamellar* depending on the number of bilayers present in the lipid aggregate. These lipid structures have been extensively used to study molecular and supramolecular features of lipid–lipid, lipid-protein, and lipid–DNA interactions (Section 22.4 below). In the following subsections, basic descriptions of the model systems and references to protocols for liposome preparation including the experimental techniques used to study liposomes will be addressed.

22.3.2.1 **Small and Large Unilamellar and Multilamellar Vesicles**

The lipid dispersions obtained by directly swelling and shaking dried phospholipid films in aqueous solvent are called *multilamellar vesicles* (MLVs; hand-shaken vesicles) (Lasch, Weissig, and Brandl, 2003). The number of lamellae (bilayers) obtained on these particular structures is high, resulting in a very broad size distribution of these preparations ranging from hundred of nanometers to several micrometers in diameter. If additional energy is supplied to the system it is possible to destabilize these multilamellar structures and form single bilayer structures (Winterhalter and Lasic, 1993). Small unilamellar vesicles (SUVs) have diameters ranging from 30 to 50 nm and thus represent membrane structures with high curvature (Winterhalter and Lasic, 1993). On the other hand, the large unilamellar vesicles (LUVs) have a diameter in the range of 100–200 nm. The curvature of these aggregates is thus much less pronounced with respect to that observed in SUVs (Winterhalter and Lasic, 1993).

A large amount of information is available in the literature regarding the different methods for vesicle preparation, including mechanisms of vesicle formation (Lasch, Weissig, and Brandl, 2003; Lasic, 1988; Lasic and Papahadjopoulos, 1995; Mozafari, 2005; Segota and Tezak, 2006; Szoka and Papahadjopoulos, 1980;

Walde and Ichikawa, 2001). The preparation methods can be categorized as those: (i) that start from dry lipid films, (ii) methods that use micelle-forming detergents (detergent depletion methods), (iii) that are based on mixing nonaqueous lipid solution with aqueous solution (solvent injection method), and (iv) that use emulsions containing water-immiscible solvents as starting systems (Lasch, Weissig and Brandl, 2003; Lasic, 1988; Mozafari, 2005; Ollivon *et al.*, 2000; Szoka and Papahadjopoulos, 1980; Walde and Ichikawa, 2001). Generally, MLVs are an intermediate step in the formation of unilamellar vesicles (e.g. LUVs or SUVs) from dried lipid films. In these cases solutions of MLVs are subsequently treated, for example, using sonication, freeze-thawing cycles, or extrusion through polycarbonate filters of well-defined pore size. The sketch shown in Figure 22.2 represents some of the aforementioned methods used to generate liposomes.

In preparing vesicles (particularly when the starting point involves dried lipid films) it is important to keep in mind that temperatures above the lipid gel–fluid phase transition temperature (T_m) must be used during vesicle formation. The lipid layers in the gel phase have particular mechanical properties (high rigidity, high lateral packing) that do not favor an optimum hydration (swelling) process. Furthermore, when MLVs are

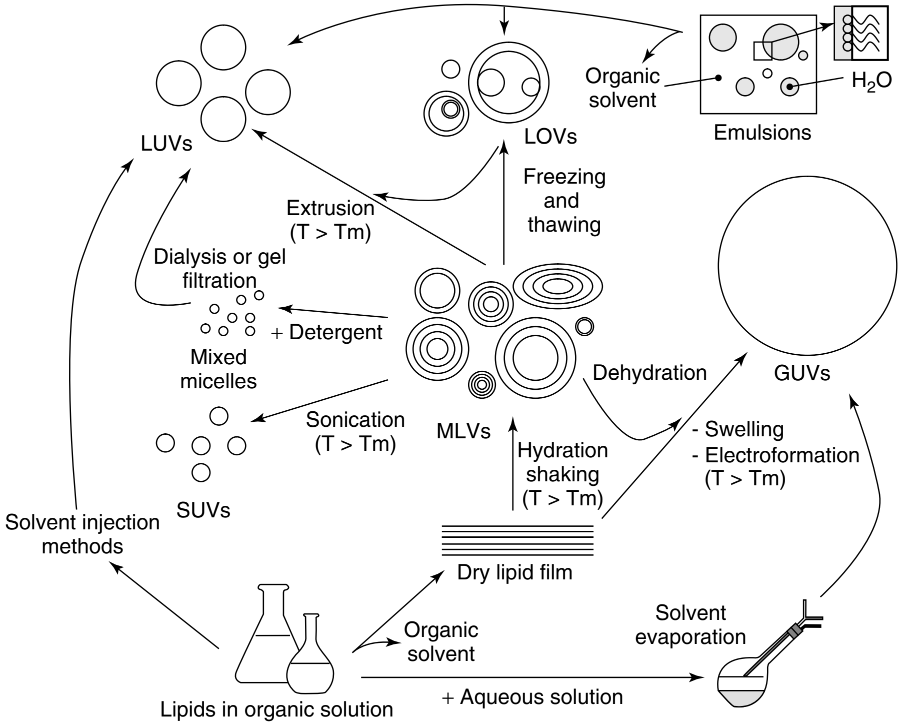
LUVs
LOVs
Organic solvent
H_2O
Emulsions
Extrusion (T > Tm)
Freezing and thawing
Dialysis or gel filtration
+ Detergent
Mixed micelles
Sonication (T > Tm)
MLVs
Dehydration
GUVs
- Swelling
- Electroformation (T > Tm)
Hydration shaking (T > Tm)
SUVs
Solvent injection methods
Dry lipid film
Organic solvent
Solvent evaporation
+ Aqueous solution
Lipids in organic solution

Fig. 22.2 Sketch representing the different protocols to generate liposomes. SUVs, small unilamellar vesicles; LUVs, large unilamellar vesicles; LOVs large oligolamelar vesicles; MLVs, multilamellar vesicles; GUVs giant unilamellar vesicles.

sonicated or extruded at temperatures below T_m, the efficiency of SUV or LUV generation is extremely low (Lasch, Weissig and Brandl, 2003; Lasic, 1988; Mozafari, 2005; Szoka and Papahadjopoulos, 1980; Walde and Ichikawa, 2001).

Another issue involving liposome preparation is related to trapping particular compounds inside the vesicles. As discussed in the Section 22.1, this procedure has several applications in basic sciences and industry, most popularly related to the delivery of substances (e.g. enzymes, drugs, vitamins, or other biological compounds, chemicals, etc.) into particular systems. The efficiency of trapping compounds into liposomes is a major issue in optimizing the different protocols (Mozafari, 2005; Taylor *et al.*, 2005; Walde and Ichikawa, 2001). Generally, this entails only slight modifications of the methods summarized in Figure 22.1. The reader can find additional information about these preparation methods in excellent reviews by Walde and Ichikawa (2001) and Taylor *et al.* (2005) and a book completely devoted to liposome preparation and characterization edited by Torchilin and Weissig (2003).

22.3.2.2 **Membrane Protein Reconstitution in Lipid Vesicles: Proteoliposomes**

Detergent removal methods (Figure 22.2) are of great utility in generating liposomes containing membrane proteins since membrane proteins are generally extracted and purified using detergent based-methods (Ollivon *et al.*, 2000; Seddon, Curnow, and Booth, 2004; Yamaguchi *et al.*, 1981; Yu and Branton, 1976). This particular model system allows protein activity to be evaluated in different lipid environments, permitting for example, correlation of the phase state of the membrane with protein structure/activity. The choice of detergents to perform the reconstitution is critical and is closely related to molecular and supramolecular mechanisms of the vesicle- to-micelle transition (Ollivon *et al.*, 2000). Generally, two common approaches have been used to generate the proteoliposomes: the dilution approach that consists of diluting the protein–detergent mixture with a liposome solution, and first introducing detergents to preformed liposomes before adding the protein–detergent mixture (Rigaud, Pitard, and Levy, 1995; Seddon, Curnow, and Booth, 2004). Also, there are a number of criteria that proteoliposomes should fulfill to be optimized for functional studies, particularly size distribution and uniform distribution of proteins into the liposomes, including the proper protein orientation in the bilayer (Seddon, Curnow, and Booth, 2004). For example, extrusion methods are sometimes applied after detergent depletion in order to achieve homogeneous size distribution (Seddon, Curnow, and Booth, 2004). Vesicles with dimensions corresponding to LUVs represent a common size for this model system although giant-sized proteoliposomes are also used (see next section). Last but not least, lipids and detergents are critical to the stabilization and formation of membrane protein crystals for X-ray diffraction analysis. Comprehensive reviews in this respect are available in the literature (Caffrey, 2003; Nollert, Navarro, and Landau, 2002).

22.3.2.3 **Giant Unilamellar Vesicles**

Many papers have appeared in recent years describing the use of giant unilamellar vesicles (GUVs) as model systems to study different physical aspects of membranes (lateral structure, mechanical properties), not only the effect of lipid–lipid, but also lipid–DNA, lipid–peptide and lipid–protein interactions; (Bagatolli, 2006) and references therein. Giant vesicles are suitable membrane model systems in part because of their size, of the order of a few tens of micrometers which is similar to the size of the plasma membrane of many cells. Single vesicles can thus be directly observed using techniques involving microscopy (such as phase contrast and fluorescence microscopy). Additionally, because experiments are performed at the level of single vesicles, undesired effects of heterogeneity in shape and size and the presence of multilamellar vesicles are eliminated. One significant aspect of using giant vesicles as model systems is the ability to control the molecular composition of the membrane as well as the environmental conditions. For instance, studies of the lateral structure of membranes using giant vesicles as model systems were normally restricted to giant vesicles composed of single lipids species or mixtures with few components (Bagatolli and Gratton, 2000a, 2000b, 1999; Baumgart, Hess, and Webb, 2003; Dietrich *et al.*, 2001; Feigenson and Buboltz, 2001; Kahya *et al.*, 2003; Kahya, Scherfeld, and Schwille, 2005b; Korlach *et al.*, 1999; Sot *et al.*, 2006b; Veatch and Keller, 2002, 2003, 2005b; Veatch *et al.*, 2004). However, as recently reported in the literature, it is also possible to form giant vesicles from natural lipid extracts (Bagatolli *et al.*, 2000; Bernardino de la Serna *et al.*, 2004; Dietrich *et al.*, 2001; Nag *et al.*, 2002) and native membranes (Baumgart *et al.*, 2007; Bernardino de la Serna *et al.*, 2004; Montes *et al.*, 2007; Ruan *et al.*, 2004). Additionally, GUVs containing membrane proteins can also be generated (Bacia *et al.*, 2004; Girard *et al.*, 2004; Kahya, 2006; Kahya, Brown, and Schwille, 2005a; Kahya *et al.*, 2001; Koster *et al.*, 2003) often involving an electroformation method (see below) using proteoliposomes as starting point.

There is no general agreement about "unique" conditions required to obtain giant vesicles mainly because the mechanism underlying giant vesicle formation is still obscure. As a consequence, there are many different methods described in the literature to obtain GUVs (Akashi *et al.*, 1996; Angelova and Dimitrov, 1986; Angelova *et al.*, 1992; Bernardino de la Serna *et al.*, 2004; Girard *et al.*, 2004; Kahya *et al.*, 2001; Moscho *et al.*, 1996; Reeves and Dowben, 1969). However, most of these methods are based on two main experimental protocols: the gentle hydration method, originally described by Reeves and Dowben (1969), and the electroformation method introduced by Angelova and Dimitrov (1986); Angelova *et al.* (1992), see Figure 22.2. Of these two experimental protocols, the electroformation method has the advantage that it provides the more homogeneous population of GUVs, with sizes between 10 and 60 μm in diameter. Additionally, the electroformation protocol requires less time compared to the gentle hydration method (~1–2 hours vs. 12–24 hours respectively) and provides a high yield of giant unilamellar vesicles (~95%) (Bagatolli, Parassasi, and Gratton, 2000; Düzgünes *et al.*, 2003).

Perhaps the main drawback of the use of GUVs as model systems for biological membranes is that low salt concentrations (generally below 10 mM NaCl) are normally required for their preparation using the above-mentioned protocols. This problem has recently been overcome by the introduction of a new electroformation protocol by Pott, Bouvrais, and Méléard (2008). This method, originally developed for single phospholipid species was successfully tested for preparing GUVs from compositionally complex lipid mixtures, including native biological membranes (Montes *et al.*, 2007). Additionally, GUVs containing asymmetric lipid distributions were recently reported by Pautot, Frisken, and Weitz (2003), providing a promising protocol to generate GUVs with different lipid compositions in the two leaflets of the membrane. Generally, bilayer compositional asymmetry is highly relevant since biological membranes are naturally asymmetric. In this method, GUVs are produced from size-controlled drops of water in oil–lipid emulsions and subsequent centrifugation against a lipid monolayer placed in an oil–water interface (Pautot, Frisken, and Weitz, 2003). However, the downside of this method is the impracticality of incorporating membrane proteins into the GUVs. The reader can find additional information about giant vesicles in an excellent review by Menger and Keiper (1998) and a book completely devoted to giant vesicles edited by Luisi and Walde (2000).

22.3.3
Planar Lipid Bilayers

As was stated above, planar bilayers (supported or free standing) are alternative model systems with which to explore the particular interactions among the constituents of biological membranes. One of the most interesting features of these model systems is that they permit generation of highly controlled asymmetric bilayers, that is, with different compositions in the two leaflets (Collins and Keller, 2008; Merkel, Sackman, and Evans, 1989). This aids not only to mimic the asymmetric nature of biological membranes but to improve the reductionist approach. From a compositional perspective these model systems represent an improvement in that include not only lipid but also membrane proteins or peptides. These systems are generally utilized to perform basic research investigation (lipid–lipid, lipid–protein interactions), and in the particular case of the supported planar membrane they are used in technological applications as biosensors (Sackmann, 1996; Trojanowicz, 2001).

22.3.3.1 Supported Lipid Bilayers

Different types of supported bilayers can be constructed (Sackmann, 1996), including (i) integrated bilayers with the inner monolayer fixed to the substrate either covalently or by ion bridges; (ii) freely supported protein bilayers separated from the substrate by ultrathin water layers (~10 Å); (iii) bilayers resting on ultrathin, soft hydrated polymer films; and (iv) freestanding bilayers resting on the aforementioned integrated bilayers called *double-supported bilayers* (Leidy *et al.*, 2004b). Consequently, a variety of methods are utilized to prepare these model membranes. Perhaps the most popular method is based on the successive transfer of monolayers using the Langmuir–Blodgett technique, but methods based in vesicle spreading or hydration of dried lipid films on different solid supports are also used (Crane and Tamm, 2007; Leidy *et al.*, 2004b; Rainey and Sykes, 2005; Simonsen and

Bagatolli, 2004; Steinem *et al.*, 1996). Membrane proteins can be incorporated into these models by spreading proteoliposomes (vesicles containing proteins see below) onto solid surfaces (Arora *et al.*, 2000; Ishiguro and Nakanishi, 1984; Slade *et al.*, 2002). Different solid supports are used to fabricate the different model systems, including mica (commonly used since it is atomically flat and very suitable to perform atomic force microscopy experiments), glass, quartz, and gold. Additionally these solid supports can be modified using polymers or other chemicals in order to generate surfaces of particular hydrophobicity or hydrophilicity (Leonenko, Carnini, and Cramb, 2000; Puu and Gustafson, 1997; Sackmann and Tanaka, 2000; Wagner and Tamm, 2000).

22.3.3.2 **Black Lipid Membranes**

The original method to experimentally generate BLMs consists of painting a solution of lipid in an organic solvent onto an aperture separating two aqueous solutions using a special chamber (Mueller *et al.*, 1962). This particular model thus makes it possible to perform experiments in a planar but freestanding bilayer. Several other methods however, have been described to generate BLM, such as brush, dipping, bubble, syringe, and monolayer techniques as detailed in (Tien, 1974; Tien and Ottova-Leitmannova, 2000). In particular, the monolayer technique allows asymmetric membranes to be formed, a feature that is almost unique to planar lipid systems (i.e. even though some methods are available to generate asymmetric liposomes most of the aforementioned liposomes model systems are symmetric). As was briefly mentioned in the historical perspective section, the major application of BLM is related to electrophysiological measurements. This is due in large part to the fact that the BLM experimental setup easily allows the concentration of ions on both sides of the bilayer to be changed, without causing osmotic stress (as in liposomes). In addition, this model allows for the easy incorporation of proteins (by the brush technique for example) to preformed lipid bilayers. In particular, incorporation of membrane proteins (e.g. channels) and peptides in the lipid bilayer and subsequent measurement of electrical signals through the membrane is commonly performed with BLMs. The reader can find additional information and a vast number of applications in the book “Planar lipid bilayers (BLMs) and their applications” (Chapters 11–26) edited by Tien and Ottova-Leitmannova (2003b).

22.4
Lipids and their Interactions with Macromolecules

This section provides a general summary describing some examples of studies of lipid–lipid, lipid–protein, and lipid–DNA interactions, including brief descriptions of the experimental techniques used in these experiments. Since the volume of the literature is very vast only a few specific (but representative) examples will be presented here.

22.4.1
Lipid–Lipid Interaction

Membranes can be classified as soft materials since they are basically structured liquids made of molecules with substantial conformational complexity (Mouritsen, 2005b). In his book "Life as a Matter of Fat" Mouritsen pointed out "... the physical properties of soft materials are largely determined by soft and fluctuating interfaces, the physics of which is dominated by entropy" (Mouritsen, 2005b). As mentioned above, the anisotropic nature of this material offers highly interconnected "observation windows" at several length and time scales. For example, understanding how a liposome can mimic the natural shape of red blood cells provides insight into which features of these soft materials are relevant to interpret the actual shape of a cell (Mouritsen, 2005c). The mechanical properties of these model membranes (e.g. bending modulus) are studied using different techniques such as micropipette manipulation or flickering analysis (Döbereiner, 2000; Mouritsen, 2005b; Nedham and Zhelev, 2000). Since light microscopic techniques are generally used in these experiments the obvious choice of a lipid model system is a giant vesicle.

A detailed understanding of temperature-dependent membrane phase equilibria in compositionally simple model systems (composed of one, two, and three-component lipid mixtures) constitutes another important step toward an understanding of the lateral organization of compositionally complex membranes. To this end, over the last 40 years there has been extensive research to elucidate the presence of particular lipid phases (including their coexistence) in compositionally diverse membranous systems using an array of experimental techniques such as fluorescence spectroscopy, differential scanning calorimetry, infrared spectroscopy, flourescence recovery after photobleaching (FRAP), electron spin resonance (ESR), nuclear magnetic resonance (NMR), quartz crystal microbalance, grazing angle X-ray diffraction, X-ray and neutron reflectivity, to mention a few, that provide important structural and dynamicl information about the organization of lipid membranes (Almeida, Vaz, and Thompson, 1992; Arnold, Losche, and Gawrisch, 1981; Bagatolli *et al.*, 1997; Blume *et al.*, 1982; Bultmann *et al.*, 1991; Caffrey and Hing, 1987; de Almeida, Fedorov, and Prieto, 2003; de Almeida *et al.*, 2005; Lee, 1975; Lentz, Barenholz, and Thompson, 1976a, 1976b; Mabrey and Sturtevant, 1976; Maggio *et al.*, 1986; Parasassi *et al.*, 1993; Schram, Lin, and Thompson, 1996; Shimshick and McConnell, 1973; van Dijck *et al.*, 1977; Vaz, Melo, and Thompson, 1989, 1990). It also includes theoretical treatments using computer simulations (Ipsen and Mouritsen, 1988; Jorgensen and Mouritsen, 1995). In several cases the results have been utilized to construct phase diagrams for different lipid mixtures, providing important thermodynamic understanding of membrane lateral organization phenomena under equilibrium conditions (Silvius, 2005b; Veatch and Keller, 2005a). It is important to note that the experimental techniques mentioned above only provide average values of parameters on the basis of data collected from bulk solutions of many liposomes (e.g. MLVs, SUVs, large oligolamellar vesicles (LOV)s, LUVs) or lipid hydrated multilayers, and hence lack visual spatial correlation of the lateral structural phenomena observed at the level of *single* membranes. Such information is particularly important in cases of compositionally complex membranes, where assigning lipid components as main actors for observing phenomena using the aforementioned experimental techniques (as generally happens in compositionally simpler situations) is complicated. Furthermore, the construction of phase diagrams is generally limited to a maximum of three components, (Veatch and Keller, 2005a) and is thus not applicable to biological membranes.

Visual spatial correlation of certain physical phenomena, reflected at the nano- and micrometer scales, is provided by microscopy techniques such as fluorescence microscopy or atomic force microscopy. The latter provides nanometer lateral resolution and Ångstrom resolution vertically, allowing visualization of phase separated lipid domains using supported planar membranes as model systems (Connell and Smith, 2006). On the other hand, the main advantage of using fluorescence microscopy techniques in membrane studies is clear: the sensitivity and flexibility of a microscope with the addition of fluorescence spectroscopy allows for the collection of spatially resolved information at a submicrometer size scale, that is ~250-nm resolution (Bagatolli, 2006). This latter information connects membrane morphology with the dynamical and structural information obtained at the molecular level using fluorescence spectroscopy (such as lipid mobility, hydration, etc.).

The additional "visual" data obtained in these microscopy experiments using GUVs (Figure 22.3) or planar membranes have generated a new and very valuable piece of information that complements very well the "bulk" information obtained from the more classical approaches (Bernardino de la

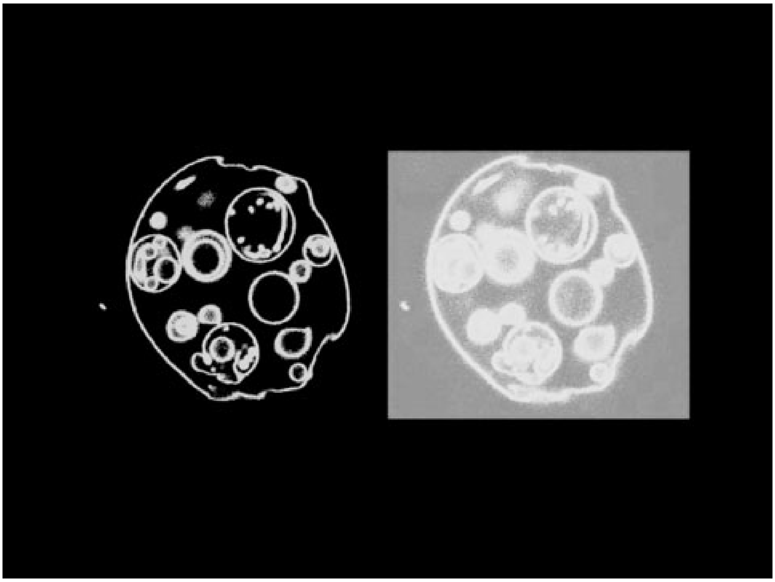

Fig. 22.3 " *Polifacético.*" Picture manipulation (giant lipid structures) (Bagatolli, 2006). Picture manipulations of lipid giant structures can be seen at the web page "science in your eyes" http://scienceinyoureyes.memphys.sdu.dk/index_en.php and in the author's personal page at http://www.memphys.sdu.dk/people/personal_pages/bagatolli/index1.php

Serna *et al.*, 2004; Dietrich *et al.*, 2001; Fidorra *et al.*, 2006; Nag *et al.*, 1998; Plasencia, Norlen and Bagatolli, 2007). This visual information can be used as an alternative to link the lateral structure of compositionally complex membranes (more than three components) with that observed in the compositionally simple cases (Bagatolli, 2006). Figure 22.4 represents an example of the overall information obtained combining differential scanning calorimetry and fluorescence microscopy techniques. In this particular case, complementary information is obtained from two different model systems (MLVs and GUVs) composed of pulmonary surfactant (Figure 22.4). This helps to establish correlations between this compositionally complex membrane and canonical (compositionally simpler) model lipid mixtures (DOPC:1, 2-Dioleoyl-sn-Glycero-3-phosphocholine /DPPC:1,2-Dipalmitoyl-sn-Glycero-3-phosphocholine/cholesterol in this case) (Bernardino de la Serna *et al.*, 2004), from which it is clear that lipid model systems have the potential to mimic not only molecular but also supramolecular levels of lipid membrane organization and dynamics.

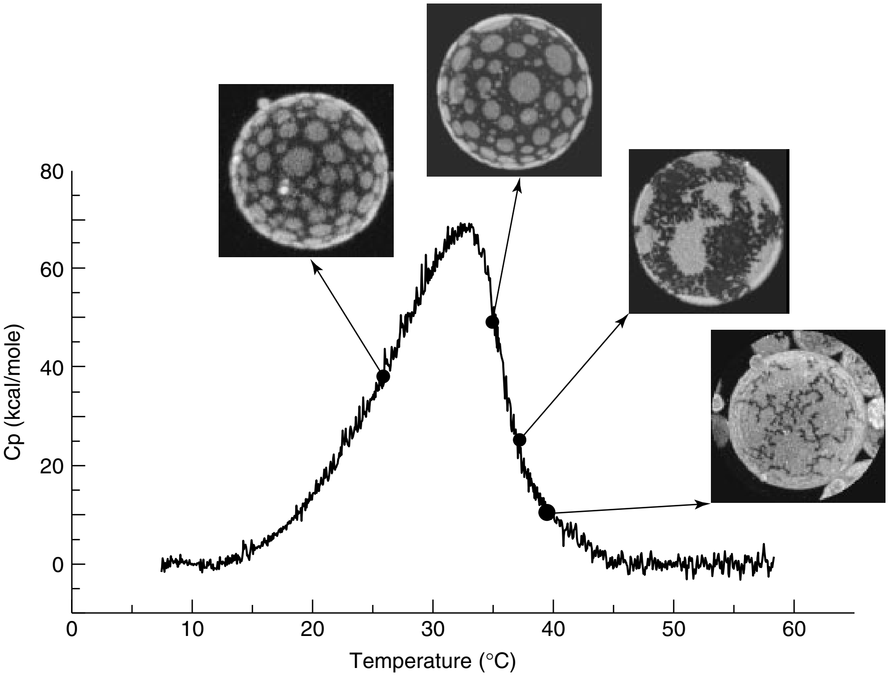

Fig. 22.4 Calorimetric profile of native pulmonary surfactant membranes (differential scanning calorimetry experiments) showing a low cooperative transition. The pictures show the actual lateral structure of the material at different temperatures (GUVs and fluorescence microscopy). The different areas in the microscopy pictures correspond to liquid ordered-(dark gray) and liquid disordered-like phases (light gray). Notice that dramatic changes in the structure of the material occur at temperatures slightly above physiological temperatures (37 °C). The figure shows the complementary information obtained from the different experimental techniques. For more details see (Bernardino de la Serna *et al.*, 2004).

22.4.2
Lipid–Protein Interactions

The vast bibliography connected with the study of lipid–protein interactions can be subdivided into studies that focus more on general effects of proteins in lipid membranes (and vice versa) and those related to a detailed description of specific cases involving particular proteins and lipids. Examples of the first category involve lipid-related enzymatic actions, membrane structural analyses (related for example to the influence of membrane proteins), hydrophobic matching, lateral pressure profile, membrane transport, membrane mechanics, and protein (peptide)-induced membrane permeability (Marsh, 2008; Mouritsen, 2005d).

Generally, the aforementioned experimental techniques (Section 22.4.1) are all used in lipid–protein interaction experiments (computer simulations, microscopy-related techniques, "bulk" techniques) providing important structurally and kinetically related information on different multimolecular systems. In these studies, the model systems are indeed based on proteoliposomes that are used directly or as the starting point to generate the model systems discussed above (i.e. planar membranes or GUVs). For example, the experiments performed with GUVs composed of pulmonary surfactant (presented in Figure 22.4) also explore the partition of surfactant hydrophobic proteins (SP-B and SP-C) when membrane heterogeneity is present in the native membrane (Bernardino de la Serna *et al.*, 2004). The partition or action of proteins (peptides) on model lipid mixtures were also explored visually, again providing a very good complement to "bulk" techniques (Ambroggio *et al.*, 2005a, 2005b; Kahya, 2006; Leidy *et al.*, 2004b; Sanchez *et al.*, 2002; Simonsen, Jensen, and Hansen, 2006; Thormann *et al.*, 2007) which also include mechanical measurements of peptide or protein interaction of membranes (Longo, Waring, and Hammer, 1997).

Additionally, the practical use of particular lipid–protein interactions in planar membrane systems as biosensors or for example, as "phantom cells" to study the interplay of specific "lock and key" relations between receptors and specific ligands (Sackmann, 1996) are a few of the interesting applications that emerge from basic studies related to protein–lipid interactions. Alternatively, the use of GUVs as bioreactors, encapsulating the cell-free expression system of bacteria and maintaining membrane protein synthesis for ~4 days (Noireaux and Libchaber, 2004) is one excellent model to study aspects of protein–lipid interactions. These studies try to mimic the process involved in first cellular life where the use of lipid models is a key element to develop proper experimental systems (Luisi, 2002; Monnard and Deamer, 2002).

22.4.3
Lipid–DNA Interactions

The interaction between nucleic acids and membranes has been extensively studied over the last decade due in large part to the great potential of these systems in gene delivery for gene therapy applications. One of the most important advantages for using liposomes in gene therapy is the lack of immunogenicity after *in vivo* administration, as compared to viral vectors (Smyth Templeton, 2002). One of the most important aspects of this liposome system is that is contains cationic lipids. Curiously, positively charged lipids are scarcely found as constitutive elements of natural membranes. The major problem with liposomes comprising natural lipids (neutral or negatively charged) is the low efficiency of encapsulating nucleic acids (e.g. plasmid DNA) (Sullivan, Gong, and Hughes, 2003). It is suggested that the presence of positive charges in the liposomes helps complexation with nucleic acids which are negatively charged (Sullivan, Gong, and Hughes, 2003).

This cationic lipid–DNA complex is called a *lipoplex* and was first reported during the late 1980s by Felgner *et al.* (1987). An effort has been directed toward designing different cationic lipids (Bajaj, Kondaiah, and Bhattacharya, 2008; Li and Szoka, 2007; Liu *et al.*, 2008) for generating cationic liposomes that can efficiently deliver nucleic acids. Generally, neutral lipids (called *lipid helpers*; e.g. 1,2-Dioleoyl-sn-Glycero-3-phosphoethanolamine (DOPE)) are also incorporated in these cationic liposomes to stabilize the liposome–DNA complex and modulate the particular physical properties of the liposome for optimum delivery (Li and Szoka, 2007; Sullivan, Gong and Hughes, 2003). This also includes the incorporation of polyethylene glycol (PEG)-lipids and lipid–polymer conjugates into these complexes to prolong liposome circulation (for *in vivo* delivery, called *stealth liposomes*), and to increase the efficiency of DNA release during delivery (Li and Szoka, 2007), respectively.

Several preparation methods have been proposed to generate nano-sized lipoplexes (Li and Szoka, 2007). These nanoparticles are designated according to the features of the generation methods: stabilized plasmid-lipid particle (SPLP), lipid-deoxyribonucleic acid particle (LDP), stable nucleic-acid-lipid particle (SNALP), nanolipoparticle (NLP), and pegylated immunoliposome (PIL). Methods to generate lipoplexes include many of the aforementioned liposome formation protocols such as detergent dialysis, ethanol injection and extrusion for example (Li and Szoka, 2007). *In vivo* delivery of these lipid–DNA complexes has extensively been tested and includes several routes such as direct injection (e.g. intratumoral), intravenous, intraperitoneal, mucosal (nasal, vaginal, rectal), transdermal, and subcutaneous to mention a few. The most promising delivery pathway seems to be the intravenous, and many investigators believe that this route can be efficient in the treatment of cancer, cardiovascular, and other inherited or acquired diseases (Smyth Templeton, 2002).

Finally, studies of lipid–DNA interaction have also been performed in GUVs. These studies focus on microinjection of giant DNA in the context of minimal life models (Nomura and Yoshikawa, 2000) and stability of cationic GUVs upon DNA injection (Angelova, Hristova, and Tsoneva, 1999).

22.5
Lateral Structure of Biological Membranes and Lessons from Model Membranes

The idea that lipids are simply randomly organized building blocks of membranes changed following the report of the coexistence of stable lipid domains in lipid bilayers three decades ago (Gebhardt, Gruler, and Sackmann, 1977; Shimshick and McConnell, 1973). However, the consequences of the nonrandom lateral organization of lipid membranes in the field of biological membranes, have not been acknowledged until recently, when the raft hypothesis was postulated (Brown and London, 1998; Edidin, 2003; Simons and Ikonen, 1997; Simons and Vaz, 2004). This hypothesis postulates that sphingomyelin/cholesterol rich areas are present in biological membranes as regions displaying a higher lateral order than the rest of the membrane (Simons and Vaz, 2004). Several membrane proteins are partitioned selectively into this area to promote signaling processes (Simons and Vaz, 2004). The main and serious drawback of this hypothesis is the way that these regions are identified, generally with the use of cold detergent extraction (Simons and Vaz, 2004). At present it is very clear (although several papers still are being published using this approach), that the addition of detergents at low temperatures *per se* generates membrane heterogeneity and this heterogeneity should thus not be associated with the presence of rafts (Lichtenberg, Goni, and Heerklotz, 2005; Sot *et al.*, 2006a). Additionally, insufficient attention is paid to the lipid composition and lipid–protein ratio of the studied membrane. In my opinion, the lateral heterogeneity of membranes is rather specific and related to compositional factors, rather than a general feature as is normally accepted. Still, and after 10 years of exhaustive investigation, the physical basis behind the lateral organization of biological membranes (including potential links with membrane function) remains obscure. For example, from the information obtained from membrane model lipid systems, the lateral heterogeneity of membranes can be rationalized simply as local (transient) compositional fluctuations or phase separation phenomena (Jacobson, Mouritsen, and Anderson, 2007). However, very little information is available about the occurrence of these two phenomena in real biological membranes. For instance, the presence of liquid ordered phases is repeatedly claimed in the literature but not always strictly demonstrated in many systems invoking rafts. This last observation points to the necessity of developing new experimental approaches in order to clarify the above-mentioned issues (Jacobson, Mouritsen, and Anderson, 2007). In particular, experimental information that allows correlation between the lateral structure of compositionally simple and complex membranes (as discussed above, Section 22.4.1) may help to identify the basis of membrane lateral heterogeneity in biologically relevant situations.

It is commonly accepted that raft size in biological membranes is of the order of tens to hundreds of nanometers, and thus not resolvable by conventional microscopy techniques. However, there are published examples of biological systems that display membrane domains with sizes of the order of micrometers; these data are not generally quoted in the raft literature, but are consistent with the lateral phase separation that is observed in some lipid model mixtures (Dietrich *et al.*, 2001; Kahya *et al.*, 2004; Veatch and Keller, 2005a). For example, transient micron-sized regions were reported Gaus *et al.* (2003) in living macrophages using the fluorescence probe 2-dimethylamino-6-lauroylnaphthalene (LAURDAN) and two-photon excitation fluorescence microscopy. This study demonstrated the presence of microscopically visible lateral phase separation in living cells, strongly supporting the notion that cholesterol had an effect on lateral organization in the membrane under study. Interestingly, this result is consistent with a report by Gousset *et al.* (2002) in which micron-sized domains were also observed in activated platelets. In the latter study, lipid mixtures with defined compositions (1-Palmitoyl-2-Oleoyl-sn-Glycero-3-phosphocholine (POPC), sphingomyelin, and cholesterol) based on the abundance of the major lipid components of the platelet membrane were used to mimic the lateral heterogeneity observed in this biological membrane (Gousset *et al.*, 2002; Leidy *et al.*, 2004a). This is similar to the case of pulmonary surfactant that contains high amounts of the saturated phospholipid DPPC, together with cholesterol, unsaturated phospholipids, and a low percentage of different membrane proteins, in which coexistence of liquid domains is observed (Figure 22.4) similar to that obtained in POPC/DPPC/Chol model systems (Bernardino de la Serna *et al.*, 2004). Perhaps the more exciting point of this work is that lateral phase separation is likely to be very relevant for the proper functioning of lung surfactant (Bernardino de la Serna *et al.*, 2004). The experimental approach described above can also be extended to other membranous systems, as was demonstrated by Ruan *et al.* (2004). In particular, the recent availability of GUVs containing the full composition of native biological membranes (Baumgart *et al.*, 2007; Montes *et al.*, 2007) is opening up exciting possibilities to proceed with this type of studies.

In the spite of the body of knowledge gained from model systems about biological membranes over the last decades, a call for new model systems is evident. Further developments into model membrane systems are expected in order to study the effect of other very important interactions, such as with the cytoskeleton (Miyata *et al.*, 2000) in order to study their influence on the actual lateral membrane organization.

22.6
Liposomes in Action

22.6.1
Liposomes and Delivery of Substances

Liposomes may satisfy very well the conditions that Paul Helrich, the father of modern medicinal chemistry, proposed for the magic bullet, that is, the perfect drug that targets diseased cells without affecting the healthy ones. Liposomes are water-containing capsules composed of biocompatible, nontoxic amphipatic materials (lipids) that may carry hydrophophilic substances in their interior and/or hydrophobic or amphipatic substances in the lipid bilayer. Moreover, they are soft and of variable size (the smallest can travel into the finest capillaries), and can be associated with specific chemical groups at the liposome surface that can act as homing devices targeting specific cells (Mouritsen, 2005e). In fact, the applications of liposomes for delivery of substances are beyond medicinal chemistry, although drug delivery for clinical applications is the most widely investigated area. These lipid entities are also utilized in the food industry and many cosmetic formulations (Lasic, 1998). Regarding the application of liposomes in medicine, there is an active effort trying to fight cancer by incorporating drugs into liposomes (Gabizon, Shmeeda, and Zalipsky, 2006). For instance, the abrupt change in permeability of liposomes during the gel–fluid phase transition was exploited by Needham *et al.* (2000) to release the anticancer drug doxorubicinin in combination with mild hyperthermia. The idea behind this experiment is that liposomes are sealed at physiological temperatures (while circulating) but upon heating (in the area where the tumor is situated) the liposome becomes leaky and releases the drug. The therapy was found to be significantly more effective than free drug or current liposome formulations (30 times or more) (Needham *et al.*, 2000). In the context of the Magic Bullet, the main disadvantage for this last application is that the tumor must be detected beforehand in order for the treatment to be efficient (in many cases this situation is not achievable). Another interesting strategy is the use of local tumor factors to induce local drug release from the liposome. For example, the fact that elevated amounts of phospholipases are present in tumors was used to degrade the liposomes containing drugs. It was further reported by Jorgensen and collaborators that lipidic nontoxic prodrugs can be circulated in stealth liposomes and be transformed into the active drug by the action of the abnormally high levels of phospholipases in the tumor, killing the malignant cells (Jorgensen, Davidson, and Mouritsen, 2002). These liposomes are called *LiPlasomes* and constitute a very promising approach to fight cancer (Mouritsen, 2005e). Several other clinical applications using liposomes include gene therapy, antibacterial and antifungal disorders, and vaccine entrapment (Banerjee, 2001; Goyal *et al.*, 2005; Gregoriadis *et al.*, 1999), showing the extraordinary versatility of these lipid model systems. Last but not least, liposomes are also utilized in the food science industry particularly to perform controlled delivery of proteins and enzymes, vitamins, and flavors as well as stabilization of food components against degradation (Taylor *et al.*, 2005).

22.7
Concluding Remarks

This chapter was intended to offer a broad overview of some of the many aspects where lipids are involved, particularly those related to membrane model systems and practical applications of lipid self-assembled structures. Lipids are fascinating molecules whose functions are not totally clear. Their capability to establish supramolecular (noncovalent) structures in aqueous media with very peculiar physical properties (related to soft materials) are exploited by nature to regulate many key functions of living systems. In the same context, practical use of lipid model systems have been realized with very promising applications that are surely still to come. In my opinion, rather than being dull molecules, lipids and their supramolecular structures are extraordinarily attractive and exciting biological elements. As can be concluded from this chapter, lipids and their structures are being examined by multiple disciplines with multiple approaches and objectives. One necessary point that can be drawn is that there is room for many more new ideas to help disentangle the many features of these important biological elements; some of them still hidden and waiting to be discovered.

Acknowledgments

This work was supported by funds from Forskningsrådet for Natur og Univers (FNU, Denmark) and the Danish National Research Foundation (which supports MEMPHYS-Center for Biomembrane Physics). The author would like to thank Robert Cantor for the critical reading of the manuscript and Signe Boe for her contribution of Figure 22.1.

References

Akashi, K., Miyata, H., Itoh, H., and Kinosita Jr, K. (1996) *Biophys. J.*, **71**, 3242–50.

de Almeida, R.F., Fedorov, A., and Prieto, M. (2003) *Biophys. J.*, **85**, 2406–16.

de Almeida, R.F., Loura, L.M., Fedorov, A., and Prieto, M. (2005) *J. Mol. Biol.*, **346**, 1109–20.

Almeida, P.F., Vaz, W.L., and Thompson, T.E. (1992) *Biochemistry*, **31**, 7198–10.

Ambroggio, E.E., Kim, D.H., Separovic, F., Barrow, C.J., Barnham, K.J., Bagatolli, L.A., and Fidelio, G.D. (2005a) *Biophys. J.*, **88**, 2706–13.

Ambroggio, E.E., Separovic, F., Bowie, J.H., Fidelio, G.D., and Bagatolli, L.A. (2005b) *Biophys. J.*, **89**, 1874–81.

Ambroggio, E.E., Separovic, F., Bowie, J., and Fidelio, G.D. (2004) *Biochim Biophys Acta.*, **1664**, 31–37.

Angelova, M.I. and Dimitrov, D.S. (1986) *Faraday Discuss. Chem. Soc.*, **81**, 303–11.

Angelova, M.I., Hristova, N., and Tsoneva, I. (1999) *Eur. Biophys. J.*, **28**, 142–50.

Angelova, M.I., Soléau, S., Meléard, P., Faucon, J.F., and Bothorel, P. (1992) *Prog. Colloid Polym. Sci.*, **89**, 127–31.

Arnold, K., Losche, A., and Gawrisch, K. (1981) *Biochim Biophys Acta.*, **645**, 143–48.

Arora, A., Rinehart, D., Szabo, G., and Tamm, L.K. (2000) *J. Biol. Chem.*, **275**, 1594–600.

Bacia, K., Schuette, C.G., Kahya, N., Jahn, R., and Schwille, P. (2004) *J. Biol. Chem.*, **279**, 37951–55.

Bagatolli, L.A. (2006) *Biochim Biophys Acta.*, **1758**, 1541–56.

Bagatolli, L.A. and Gratton, E. (1999) *Biophys. J.*, **77**, 2090–101.

Bagatolli, L.A. and Gratton, E. (2000a) *Biophys. J.*, **79**, 434–47.

Bagatolli, L.A. and Gratton, E. (2000b) *Biophys. J.*, **78**, 290–305.

Bagatolli, L.A., Gratton, E., Khan, T.K., and Chong, P.L. (2000) *Biophys. J.*, **79**, 416–25.

Bagatolli, L.A., Maggio, B., Aguilar, F., Sotomayor, C.P., and Fidelio, G.D. (1997) *Biochim Biophys Acta.*, **1325**, 80–90.

Bagatolli, L.A., Parasassi, T., and Gratton, E. (2000) *Chem. Phys. Lipids*, **105**, 135–47.

Bajaj, A., Kondaiah, P., and Bhattacharya, S. (2008) *Biomacromolecules*, **9**, 991–99.

Banerjee, R. (2001) *J. Biomater. Appl.*, **16**, 3–21.

Bangham, A.D. (1993) *Chem. Phys. Lipids*, **64**, 275–85.

Bangham, A.D. and Horne, R.W. (1964) *J. Mol. Biol.*, **8**, 660–68.

Bar, R.S., Deamer, D.W., and Cornwell, D.G. (1966) *Science*, **153**, 1010–12.

Barr, Y. and Helland, S. (2007) *J. Liposome Res.*, **17**, 79–88.

Baumgart, T., Hammond, A.T., Sengupta, P., Hess, S.T., Holowka, D.A., Baird, B.A., and Webb, W.W. (2007) *Proc. Nat. Acad. Sci. U.S.A.*, **104**, 3165–70.

Baumgart, T., Hess, S.T., and Webb, W.W. (2003) *Nature*, **425**, 821–24.

Bernardino de la Serna, J., Perez-Gil, J., Simonsen, A.C., and Bagatolli, L.A. (2004) *J. Biol. Chem.*, **279**, 40715–22.

Bianco, I.D., Fidelio, G.D., Yu, R.K., and Maggio, B. (1991) *Biochemistry*, **30**, 1709–14.

Blodgett, K.B. (1935) *J. Am. Chem. Soc.*, **57**, 1007–22.

Blume, A., Wittebort, R.J., Das Gupta, S.K., and Griffin, R.G. (1982) *Biochemistry*, **21**, 6243–53.

Bragg, W.H. (1939) *Proc. R. Inst. G. B.*, **30**, 687.

Brown, R.E. and Brockman, H.L. (2007) *Methods Mol. Biol.*, **398**, 41–58.

Brown, D.A. and London, E. (1998) *J. Membr. Biol.*, **164**, 103–14.

Bultmann, T., Vaz, W.L., Melo, E.C., Sisk, R.B., and Thompson, T.E. (1991) *Biochemistry*, **30**, 5573–79.

Caffrey, M. (2003) *J. Struct. Biol.*, **142**, 108–32.

Caffrey, M. and Hing, F.S. (1987) *Biophys. J.*, **51**, 37–46.

Capek, I. (2001) *Adv. Colloid Interface Sci.*, **91**, 295–334.

Casals, C. (2001) *Pediatr. Pathol. Mol. Med.*, **20**, 249–68.

Choi, S., Oh, S., and Lee, J. (2001) *Colloids Surf., B*, **20**, 239–44.

Churbanova, I.Y., Tronin, A., Strzalka, J., Gog, T., Kuzmenko, I., Johansson, J.S., and Blasie, J.K. (2006) *Biophys. J.*, **90**, 3255–66.

Collins, M.D. and Keller, S.L. (2008) *Proc. Nat. Acad. Sci. U.S.A.*, **105**, 124–28.

Connell, S.D. and Smith, D.A. (2006) *Mol. Membr. Biol.*, **23**, 17–28.

Crane, J.M. and Tamm, L.K. (2007) *Methods Mol. Biol.*, **400**, 481–88.

Cruz, A. and Perez-Gil, J. (2007) *Methods Mol. Biol.*, **400**, 439–57.

Danielli, J.F. and Davson, H. (1935) *J. Cell. Comp. Physiol.*, **5**, 495–508.

Dietrich, C., Bagatolli, L.A., Volovyk, Z.N., Thompson, N.L., Levi, M., Jacobson, K., and Gratton, E. (2001) *Biophys. J.*, **80**, 1417–28.

van Dijck, P.W., Kaper, A.J., Oonk, H.A., and de Gier, J. (1977) *Biochim Biophys Acta.*, **470**, 58–69.

Döbereiner, H. (2000) in *Giant Vesicles*, (eds P.L. Luisi and P. Walde), John Wiley & Sons, Ltd, New York, pp. 149–67.

Düzgünes, N., Bagatolli, L.A., Meers, P., Oh, Y.K., and Straubinger, R.M. (2003) in *Liposomes: A Practical Approach*, 2nd edn (eds V. Weissig and V. Torchilin), Oxford University Press, Oxford, pp. 105–47.

Dynarowicz-Latka, P., Dhanabalan, A., and Oliveira Jr, O.N. (2001) *Adv. Colloid Interface Sci.*, **91**, 221–93.

Edidin, M. (2003) *Annu. Rev. Biophys. Biomol. Struct.*, **32**, 257–83.

Epand, R.M. (2007) *Methods Mol. Biol.*, **400**, 15–26.

Feigenson, G.W. and Buboltz, J.T. (2001) *Biophys. J.*, **80**, 2775–88.

Felgner, P.L., Gadek, T.R., Holm, M., Roman, R., Chan, H.W., Wenz, M., Northrop, J.P., Ringold, G.M., and Danielsen, M. (1987) *Proc. Nat. Acad. Sci. U.S.A.*, **84**, 7413–17.

Fidelio, G.D., Austen, B.M., Chapman, D., and Lucy, J.A. (1986) *Biochem. J.*, **238**, 301–04.

Fidorra, M., Duelund, L., Leidy, C., Simonsen, A.C., and Bagatolli, L.A. (2006) *Biophys. J.*, **90**, 4437–51.

Franklin, B., Brownrigg, W., and Farish, M. (1774) *Philos. Trans. (1683–1775)*, **64**, 445–60.

Fulford, G.D. (1968) *Isis*, **59**, 198–99.

Gabizon, A.A., Shmeeda, H., and Zalipsky, S. (2006) *J. Liposome Res.*, **16**, 175–83.

Gaines, G.L.J. (1966) *Insoluble Monolayers at Liquid-Gas Interfaces*, Wiley InterScience, New York.

Gaus, K., Gratton, E., Kable, E.P., Jones, A.S., Gelissen, I., Kritharides, L., and Jessup, W. (2003) *Proc. Nat. Acad. Sci. U.S.A.*, **100**, 15554–59.

Gebhardt, C., Gruler, H., and Sackmann, E. (1977) *Z. Naturforsch., C: Biosci.*, **32**, 581–96.

Giner-Casares, J.J., Camacho, L., Martin-Romero, M.T., and Cascales, J.J. (2008) *Langmuir*, **24**, 1823–28.

Girard, P., Pecreaux, J., Lenoir, G., Falson, P., Rigaud, J.L. and Bassereau, P. (2004) *Biophys. J.*, **87**, 419–429.

Gorter, E. and Grendel, F. (1925) *J. Exp. Med.*, **41**, 439–443.

Gousset, K., Wolkers, W.F., Tsvetkova, N.M., Oliver, A.E., Field, C.L., Walker, N.J., Crowe, J.H. and Tablin, F. (2002) *J. Cell. Physiol.*, **190**, 117–128.

Goyal, P., Goyal, K., Vijaya Kumar, S.G., Singh, A., Katare, O.P. and Mishra, D.N. (2005) *Acta Pharm.*, **55**, 1–25.

Gregoriadis, G., McCormack, B., Obrenovic, M., Saffie, R., Zadi, B. and Perrie, Y. (1999) *Methods*, **19**, 156–162.

Hilgemann, D.W. (2003) *Annu. Rev. Physiol.*, **65**, 697–700.

Hollars, C.W. and Dunn, R.C. (1998) *Biophys. J.*, **75**, 342–353.

Ipsen, J.H. and Mouritsen, O.G. (1988) *Biochim Biophys Acta.*, **944**, 121–134.

Ishiguro, T. and Nakanishi, M. (1984) *J. Biochem.*, **95**, 581–583.

Jacobson, K., Mouritsen, O.G. and Anderson, R.G. (2007) *Nat. Cell Biol.*, **9**, 7–14.

Johonnott, E.S. (1899) *Philos. Mag.*, **47**, 501–522.

Johonnott, E.S. (1906) *Philos. Mag.*, **11**, 746–753.

Jorgensen, K., Davidsen, J. and Mouritsen, O.G. (2002) *FEBS Lett.*, **531**, 23–27.

Jorgensen, K. and Mouritsen, O.G. (1995) *Biophys. J.*, **69**, 942–954.

Kahya, N. (2006) *Chem. Phys. Lipids*, **141**, 158–168.

Kahya, N., Brown, D.A. and Schwille, P. (2005a) *Biochemistry*, **44**, 7479–7489.

Kahya, N., Scherfeld, D. and Schwille, P. (2005b) *Chem. Phys. Lipids*, **135**, 169–180.

Kahya, N., Pecheur, E.I., de Boeij, W.P., Wiersma, D.A. and Hoekstra, D. (2001) *Biophys. J.*, **81**, 1464–1474.

Kahya, N., Scherfeld, D., Bacia, K., Poolman, B. and Schwille, P. (2003) *J. Biol. Chem.*, **278**, 28109–28115.

Kahya, N., Scherfeld, D., Bacia, K. and Schwille, P. (2004) *J. Struct. Biol.*, **147**, 77–89.

King, M.D. and Marsh, D. (1987) *Biochemistry*, **26**, 1224–1231.

Korlach, J., Schwille, P., Webb, W.W. and Feigenson, G.W. (1999) *Proc. Nat. Acad. Sci. U.S.A.*, **96**, 8461–8466.

Koster, G., VanDuijn, M., Hofs, B. and Dogterom, M. (2003) *Proc. Nat. Acad. Sci. U.S.A.*, **100**, 15583–15588.

Kraft, M.L., Weber, P.K., Longo, M.L., Hutcheon, I.D. and Boxer, S.G. (2006) *Science*, **313**, 1948–1951.

Langmuir, I. (1917) *J. Am. Chem. Soc.*, **39**, 1848–1906.

Lasch, J., Weissig, V. and Brandl, M. (2003) in *Liposomes*, 2nd edn (eds V.P. Torchilin and V. Weissig), Oxford University Press, New York, pp. 3–29.

Lasic, D.D. (1988) *Biochem. J.*, **256**, 1–11.

Lasic, D.D. (1998) *Trends Biotechnol.*, **16**, 307–321.

Lasic, D.D. (2000) in *Giant Vesicles*, (eds P.L. Luisi and P. Walde), John Wiley & Sons, Ltd, pp. 11–24.

Lasic, D.D. and Papahadjopoulos, D. (1995) *Science*, **267**, 1275–1276.

Lee, A.G. (1975) *Biochim Biophys Acta.*, **413**, 11–23.

Leidy, C., Gousset, K., Ricker, J., Wolkers, W.F., Tsvetkova, N.M., Tablin, F. and Crowe, J.H. (2004a) *Cell Biochem. Biophys.*, **40**, 123–148.

Leidy, C., Mouritsen, O.G., Jorgensen, K., and Peters, G.H. (2004b) *Biophys. J.*, **87**, 408–18.

Leneweit, G., Vranceanu, M., Winkler, K., and Nirschl, H. (2008) *Biophys. J.*, **94**, 3924–34.

Lentz, B.R., Barenholz, Y., and Thompson, T.E. (1976a) *Biochemistry*, **15**, 4529–37.

Lentz, B.R., Barenholz, Y., and Thompson, T.E. (1976b) *Biochemistry*, **15**, 4521–28.

Leonenko, Z.V., Carnini, A., and Cramb, D.T. (2000) *Biochim Biophys Acta.*, **1509**, 131–47.

Li, W. and Szoka Jr, F.C. (2007) *Pharm. Res.*, **24**, 438–49.

Lichtenberg, D., Goni, F.M., and Heerklotz, H. (2005) *Trends Biochem. Sci.*, **30**, 430–36.

Liu, D., Qiao, W., Li, Z., Chen, Y., Cui, X., Li, K., Yu, L., Yan, K., Zhu, L., Guo, Y., and Cheng, L. (2008) *Chem. Biol. Drug Des.*, **71**, 336–44.

Longo, M.L., Waring, A.J., and Hammer, D.A. (1997) *Biophys. J.*, **73**, 1430–1439.

Luisi, P.L. (2002) *Anat. Rec.*, **268**, 208–14.

Luisi, P. and Walde, P. (2000) *Giant Vesicles*, John Wiley & Sons Ltd, England.

Mabrey, S. and Sturtevant, J.M. (1976) *Proc. Nat. Acad. Sci. U.S.A.*, **73**, 3862–66.

Maget-Dana, R. (1999) *Biochim Biophys Acta.*, **1462**, 109–40.

Maggio, B. (1994) *Prog. Biophys. Mol. Biol.*, **62**, 55–117.

Maggio, B., Fidelio, G.D., Cumar, F.A., and Yu, R.K. (1986) *Chem. Phys. Lipids*, **42**, 49–63.

Marsh, D. (2008) *Biochim. Biophys. Acta.*, **1778**, 1545–75.

McBain, J.W. (1913) *Trans. Faraday Soc.*, **9**, 99–101.

McConnell, H.M., Tamm, L.K., and Weis, R.M. (1984) *Proc. Nat. Acad. Sci. U.S.A.*, **81**, 3249–53.

Menger, F.M. and Keiper, J.S. (1998) *Curr. Opin. Chem. Biol.*, **2**, 726–32.

Merkel, R., Sackmann, E. and Evans, E. (1989) *J. Phys. (Paris)*, **50**, 1535–55.

Miyata, H., Ohki, K., Marriot, G., Nishiyama, S., Akashi, K., and Kinosita Jr, K. (2000) in *Giant Vesicles*, (eds P.L. Luisi and P. Walde), John Wiley & Sons, LTD, New York, pp. 319–34.

Monnard, P.A. and Deamer, D.W. (2002) *Anat. Rec.*, **268**, 196–207.

Montes, L.R., Alonso, A., Goni, F.M., and Bagatolli, L.A. (2007) *Biophys. J.*, **93**, 3548–54.

Morera, F.J., Vargas, G., Gonzalez, C., Rosenmann, E., and Latorre, R. (2007) *Methods Mol. Biol.*, **400**, 571–85.

Moscho, A., Orwar, O., Chiu, D.T., Modi, B.P., and Zare, R.N. (1996) *Proc. Nat. Acad. Sci. U.S.A.*, **93**, 11443–47.

Mouritsen, O.G. (ed.) (2005a) in *Life as a Matter of Fat: The Emerging Science* *of Lipidomics*, Springuer-Verlag, Berlin-Heidelberg-New York, pp. 1–5.

Mouritsen, O.G. (ed.) (2005b), in *Life as a Matter of Fat: The Emerging Science of Lipidomics*, Springuer-Verlag, Berlin-Heidelberg-New York, 53–62.

Mouritsen, O.G. (ed.) (2005c), in *Life as a Matter of Fat: The Emerging Science of Lipidomics*, Springuer-Verlag, Berlin-Heidelberg-New York, pp. 63–71.

Mouritsen, O.G. (ed.) (2005d), in *Life as a Matter of Fat: The Emerging Science of Lipidomics*, Springuer-Verlag, Berlin-Heidelberg-New York, pp. 159–72.

Mouritsen, O.G. (ed.) (2005e), in *Life as a Matter of Fat: The Emerging Science of Lipidomics*, Springuer-Verlag, Berlin-Heidelberg-New York, pp 209–23.

Mozafari, M.R. (2005) *Cell. Mol. Biol. Lett.*, **10**, 711–19.

Mueller, P., Rudin, D.O., Tien, H.T., and Wescott, W.C. (1962) *Nature*, **194**, 979–80.

Nag, K., Pao, J.S., Harbottle, R.R., Possmayer, F., Petersen, N.O., and Bagatolli, L.A. (2002) *Biophys. J.*, **82**, 2041–51.

Nag, K., Perez-Gil, J., Ruano, M.L., Worthman, L.A., Stewart, J., Casals, C., and Keough, K.M. (1998) *Biophys. J.*, **74**, 2983–95.

Nakanishi, M., Brian, A.A. and McConnell, H.M. (1983) *Mol Immunol.*, **20**, 1227–31.

Nedham, D. and Zhelev, D. (2000) in *Giant Vesicles*, (eds P.L. Luisi and P. Walde), John Wiley & Sons, Ltd, New York, pp. 103–47.

Needham, D., Anyarambhatla, G., Kong, G., and Dewhirst, M.W. (2000) *Cancer Res.*, **60**, 1197–201.

Neubauer, C. (1867) *Z. Anal. Chem.*, **6**, 189–95.

Newton, I. (1704) *Optiks*, Dover, New York, (reprinted 1952), pp. 215–32.

Noireaux, V. and Libchaber, A. (2004) *Proc. Nat. Acad. Sci. U.S.A.*, **101**, 17669–74.

Nollert, P., Navarro, J., and Landau, E.M. (2002) *Methods Enzymol.*, **343**, 183–99.

Nomura, S. and Yoshikawa, K. (2000) in *Giant Vesicles*, (eds P.L. Luisi and P. Walde), John Wiley & Sons, Ltd, Chichester, pp. 313–18.

Norlen, L., Gil, I.P., Simonsen, A., and Descouts, P. (2007) *J. Struct. Biol.*, **158**, 386–400.

Notter, R. (2000) *Lung Surfactants: Basic Science and Clinical Applications (Lung Biology in Health and Disease)*, Marcel Dekker Inc, New York.

Ollivon, M., Lesieur, S., Grabielle-Madelmont, C., and Paternostre, M. (2000) *Biochim Biophys Acta.*, **1508**, 34–50.

Parasassi, T., Loiero, M., Raimondi, M., Ravagnan, G., and Gratton, E. (1993) *Biochim Biophys Acta.*, **1153**, 143–54.

Paul, G.W., Hassett, R.J., and Reiss, O.K. (1977) *Proc. Nat. Acad. Sci. U.S.A.*, **74**, 3617–20.

Pautot, S., Frisken, B.J., and Weitz, D.A. (2003) *Proc. Nat. Acad. Sci. U.S.A.*, **100**, 10718–21.

Perez-Gil, J. (2001) *Pediatr. Pathol. Mol. Med.*, **20**, 445–69.

Perrin, J. (1918) *Ann. Phys.*, **10**, 160.

Pieroni, G., Gargouri, Y., Sarda, L., and Verger, R. (1990) *Adv. Colloid Interface Sci.*, **32**, 341–78.

Piknova, B., Schram, V., and Hall, S.B. (2002) *Curr. Opin. Struct. Biol.*, **12**, 487–94.

Pink, D.A. (1984) *Can. J. Biochem. Cell Biol.*, **62**, 760–77.

Plasencia, I., Norlen, L., and Bagatolli, L.A. (2007) *Biophys. J.*, **93**, 3142–55.

Pockels, A. (1894) *Nature*, **50**, 223–24.

Pott, T., Bouvrais, H. and Méléard, P. (2008) *Chem. Phys. Lipids*, **154**, 115–9.

Przestalski, S., Sarapuk, J., Kleszczynska, H., Gabrielska, J., Hladyszowski, J., Trela, Z., and Kuczera, J. (2000) *Acta Biochim. Pol.*, **47**, 627–38.

Puu, G. and Gustafson, I. (1997) *Biochim Biophys Acta.*, **1327**, 149–61.

Rainey, J.K. and Sykes, B.D. (2005) *Biophys. J.*, **89**, 2792–805.

Rayleigh, L. (1899) *Philos. Mag.*, **48**, 151.

Reeves, J.P. and Dowben, R.M. (1969) *J. Cell. Physiol.*, **73**, 49–60.

Rideal, E.K. (1953) *No Suggestion Available*, **8**, 529–47.

Rigaud, J.L., Pitard, B., and Levy, D. (1995) *Biochim Biophys Acta.*, **1231**, 223–46.

Robertson, J.D. (1981) *J. Cell Biol.*, **91**, 189s–204s.

Ruan, Q., Cheng, M.A., Levi, M., Gratton, E., and Mantulin, W.W. (2004) *Biophys. J.*, **87**, 1260–67.

Sackmann, E. (1996) *Science*, **271**, 43–48.

Sackmann, E. and Tanaka, M. (2000) *Trends Biotechnol.*, **18**, 58–64.

Sanchez, S.A., Bagatolli, L.A., Gratton, E., and Hazlett, T.L. (2002) *Biophys. J.*, **82**, 2232–43.

Schram, V., Lin, H.N., and Thompson, T.E. (1996) *Biophys. J.*, **71**, 1811–22.

Seddon, J.M. and Cevc, G. (1993), in *Phospholipid Handbook,* (eds G. Cevc), Marcel Dekker, Inc, New York, pp. 403–54.

Seddon, A.M., Curnow, P., and Booth, P.J. (2004) *Biochim Biophys Acta.*, **1666**, 105–17.

Segota, S. and Tezak, D. (2006) *Adv. Colloid Interface Sci.*, **121**, 51–75.

Shimshick, E.J. and McConnell, H.M. (1973) *Biochemistry*, **12**, 2351–60.

Silvius, J.R. (2003) *Biochim Biophys Acta.*, **1610**, 174–83.

Silvius, J. (2005a) *Q. Rev. Biophys.*, **38**, 373–83.

Silvius, J.R. (2005b) *Biochim Biophys Acta.*, **1746**, 193–202.

Simons, K. and Ikonen, E. (1997) *Nature*, **387**, 569–72.

Simons, K. and Vaz, W.L. (2004) *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 269–95.

Simonsen, A.C. and Bagatolli, L.A. (2004) *Langmuir*, **20**, 9720–28.

Simonsen, A.C., Jensen, U.B., and Hansen, P.L. (2006) *J. Colloid Interface Sci.*, **301**, 107–15.

Singer, S.J. and Nicolson, G.L. (1972) *Science*, **175**, 720–31.

Slade, A., Luh, J., Ho, S., and Yip, C.M. (2002) *J. Struct. Biol.*, **137**, 283–91.

Smyth Templeton, N. (2002) *DNA Cell Biol.*, **21**, 857–67.

Sostarecz, A.G., McQuaw, C.M., Ewing, A.G., and Winograd, N. (2004) *J. Am. Chem. Soc.*, **126**, 13882–83.

Sot, J., Bagatolli, L.A., Goni, F.M., and Alonso, A. (2006a) *Biophys. J.*, **90**, 903–14.

Sot, J., Bagatolli, L.A., Goñi, F.M., and Alonso, A. (2006b) *Biophys. J.*, **90**, 903–14.

Steinem, C., Janshoff, A., Ulrich, W.P., Sieber, M., and Galla, H.J. (1996) *Biochim Biophys Acta.*, **1279**, 169–80.

Sullivan, S., Gong, Y., and Hughes, J. (2003), in *Liposomes*, 2nd edn (eds V.P. Torchilin and V. Weissig), Oxford University Press, New York, pp. 289–301.

Szoka, F. and Papahadjopoulos Jr, D. (1980) *Annu. Rev. Biophys. Bioeng.*, **9**, 467–508.

Tamm, L.K. and McConnell, H.M. (1985) *Biophys. J.*, **47**, 105–13.

Tancrede, P., Paquin, P., Houle, A., and Leblanc, R.M. (1983) *J. Biochem. Biophys. Methods*, **7**, 299–310.

Taylor, T.M., Davidson, P.M., Bruce, B.D., and Weiss, J. (2005) *Crit. Rev. Food Sci. Nutr.*, **45**, 587–605.

Thormann, E., Simonsen, A.C., Nielsen, L.K., and Mouritsen, O.G. (2007) *J. Mol. Recognit.*, **20**, 554–60.

Tien, H.T. (1974) *Bilayer Lipid Membranes (BLM): Theory and Practice*, Marcel Dekker, Inc, New York.

Tien, H.T. and Ottova-Leitmannova, A. (eds) (2000) in *Membrane Biophysics as Viewed from Experimental Bilayer Lipid Membranes (Planar Lipid Bilayers and Spherical Liposomes)*, Elsevier Science B.V., Amsterdan, pp. 149–220.

Tien, H.T. and Ottova-Leitmannova, A. (eds) (2003a) in *Planar Lipid Bilayers (BLMs) and their Applications*, Elsevier Science B.V., Amsterdan, pp. 1–73.

Tien, H.T. and Ottova-Leitmannova, A. (2003b) in *Planar Lipid Bilayers (BLMS) and their Applications*, Chapters 11 to 26, Elsevier Science B.V., Amsterdam, pp. 347–723.

Torchilin, V.P. and Weissig, V. (2003) *Liposomes*, 2nd edn, Oxford University Press, New York.

Trojanowicz, M. (2001) *Fresenius J. Anal. Chem.*, **371**, 246–60.

Vaz, W.L., Melo, E.C., and Thompson, T.E. (1989) *Biophys. J.*, **56**, 869–76.

Vaz, W.L., Melo, E.C., and Thompson, T.E. (1990) *Biophys. J.*, **58**, 273–75.

Veatch, S.L. and Keller, S.L. (2002) *Phys. Rev. Lett.*, **89**, 268101.

Veatch, S.L. and Keller, S.L. (2003) *Biophys. J.*, **85**, 3074–83.

Veatch, S.L. and Keller, S.L. (2005a) *Biochim Biophys Acta.*, **1746**, 172–85.

Veatch, S.L. and Keller, S.L. (2005b) *Phys. Rev. Lett.*, **94**, 148101.

Veatch, S.L., Polozov, I.V., Gawrisch, K., and Keller, S.L. (2004) *Biophys. J.*, **86**, 2910–22.

Velter, I.A., Politi, M., Podlipnik, C., and Nicotra, F. (2007) *Mini Rev. Med. Chem.*, **7**, 159–70.

Virchow, R. (1854) *Virchows Arch.*, **6**, 562–72.

Wagner, M.L. and Tamm, L.K. (2000) *Biophys. J.*, **79**, 1400–14.

Walde, P. and Ichikawa, S. (2001) *Biomol. Eng.*, **18**, 143–77.

Wang, L., Cruz, A., Flach, C.R., Perez-Gil, J., and Mendelsohn, R. (2007) *Langmuir*, **23**, 4950–58.

Willcox, M. (2000) in *Poucher's Perfumes, Cosmetics and Soaps*, 10th edn (eds H. Butler), Kluwer Academic, pp. 453–466.

Winterhalter, M. (2000) *Curr. Opin. Colloid Interface Sci.*, **5**, 250–255.

Winterhalter, M. and Lasic, D.D. (1993) *Chem. Phys. Lipids*, **64**, 35–43.

Won, J.S., Singh, A.K., and Singh, I. (2007) *J. Neurochem.*, **103**((),Suppl 1), 180–91.

Yamaguchi, M., Sakai, M., Horigome, T., Omata, S., and Sugano, H. (1981) *Biochem. J.*, **194**, 907–13.

Yan, W., Biswas, S.C., Laderas, T.G., and Hall, S.B. (2007) *J. Appl. Physiol.*, **102**, 1739–45.

Yu, J. and Branton, D. (1976) *Proc. Nat. Acad. Sci. U.S.A.*, **73**, 3891–95.

Zheng, L., McQuaw, C.M., Ewing, A.G., and Winograd, N. (2007) *J. Am. Chem. Soc.*, **129**, 15730–31.

25
Biosimulation and Computations in Systems Biology

Olga Sosnovtseva, Jakob L. Laugesen, and Erik Mosekilde

25.1 **Introduction: Aiming to Explain 841**
25.2 **Approaches to a Systems Description 842**
25.3 **Biological Rhythms and the Role of Feedback Regulation 844**
25.4 **Instabilities and Far-From-Equilibrium Conditions 846**
25.5 **Luteinizing Hormone 847**
25.5.1 Model of the Pulsatile Release of Luteinizing Hormone 847
25.6 **Nephron Pressure and Flow Regulation 852**
25.6.1 Physiological Background 853
25.7 **Single Nephron Model 855**
25.8 **Simulation and Analysis 860**
25.9 **The Promise of Systems Biology 864**
References 865
Further Reading 866

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

25.1
Introduction: Aiming to Explain

Functioning of the human organism depends on regulatory processes that span an enormous range of spatial scales, from the nanometer scale of molecular conformation changes to the transmission of nerve signals over distances of the order of a meter. The fundamental elements of biochemical networks are nucleic acids (gene regulatory network), biomolecules (signaling networks), or small organic compounds (metabolic networks). The human genome may only have about 25 000 genes, but these genes code for more than 100 000 proteins, many of which serve more than one function. Although some are more likely than others, the number of conceivable interactions between all the human genes and their protein products is completely incomprehensive. At the cellular level, biomolecular signaling networks implement a new set of functions, including signal transduction, rhythm generation, and intercellular signaling. Tissues and organs, the next level of biological organization, again display completely new properties. Communication among the cells leads them to function in a coordinated manner that can be very different from the isolated behavior of the cells that constitute the tissue. Biological complexity cuts across time as well as space. At the molecular level, timescales of 10^{-9} seconds, characteristic of Brownian motion, are important, while aging processes and the development of diseases such as cancer and diabetes may involve timescales of the order of a human lifetime or 10^9 seconds.

Within this area of multidisciplinary research, systems biology and biosimulation are characterized by the following (Alon, 2007):

- Global rather than local analysis, or holistic rather than reductionistic perspective. This refers to the attempt of systems biology to capture and analyze many interconnected aspects of a biological system simultaneously. This is in contrast to the reductionistic approach where one or a few aspects are studied at a time.
- The simultaneous study of different levels of biological organization, often with the aim of describing how new properties and functions arise through the coordinated interaction of many units.
- The investigation of biological reaction cascades and feedback networks, including gene regulatory,

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

protein interaction as well as signaling and metabolic networks.
- The explicit incorporation of time in order to capture and analyze dynamical behaviors and stimuli-response patterns of biological systems and processes;
- The use of concepts and tools from engineering and control theory to examine the role of particular feedback regulations and to determine the significance and stability of such regulations.
- The direct involvement of concepts from nonlinear dynamics and nonequilibrium thermodynamics to explain and describe observed temporal patterns, synchronization phenomena, and nonlinear wave propagation.
- The use of computational means to capture, model, and simulate biological processes and systems.

Biosimulation, as we define it, is the application of systems biology to solve concrete problems in the pharmaceutical industry or the health sector at large (Bertau, Mosekilde and Westerhoff, 2008). Successful work in this area clearly requires close collaboration between pharmacists, medical doctors, and modelers with a good understanding of feedback control, stability criteria, and nonlinear dynamics.

25.2
Approaches to a Systems Description

Computationally oriented biology involves two distinct approaches: (i) pattern recognition or data-mining that aims at extracting the hidden information from huge quantities of experimental data, in order to discover new phenomena, quantify their significance, and formulate hypotheses about the involved processes (Baldi and Brunak, 1998) and (ii) simulation-based analysis (or hypothesis-driven research) that aims at establishing a dynamic model that captures the key features of a system and is consistent with all available experimental information (Bertau, Mosekilde and Westerhoff, 2008).

Knowledge discovery and pattern recognition are used extensively within bioinformatics for such tasks as the prediction of exon-intron position and protein structure from a given sequence or the inference of gene regulatory networks from expression profiles (Baldi and Brunak, 1998; Kitano, 2001). These methods typically use predictions based on statistical discriminators that often involve sophisticated approaches (such as hidden Markov models) or other algorithms. In contrast, the simulation approach attempts to predict the dynamics of a system so that the validity of the underlying causal structures can be tested (Figure 25.1).

Although bioinformatics has long been used for genome analysis, protein structure determination, and drug-likelyhood predictions, simulation-based approaches have so far received less attention. This now appears to change. Convincing results obtained with the systems approach have lead to a significant boost in interest, and combined with substantial advances in software and computational power this has enabled the creation and analysis of reasonably realistic, yet intricate biological models, such as the full-heart model developed

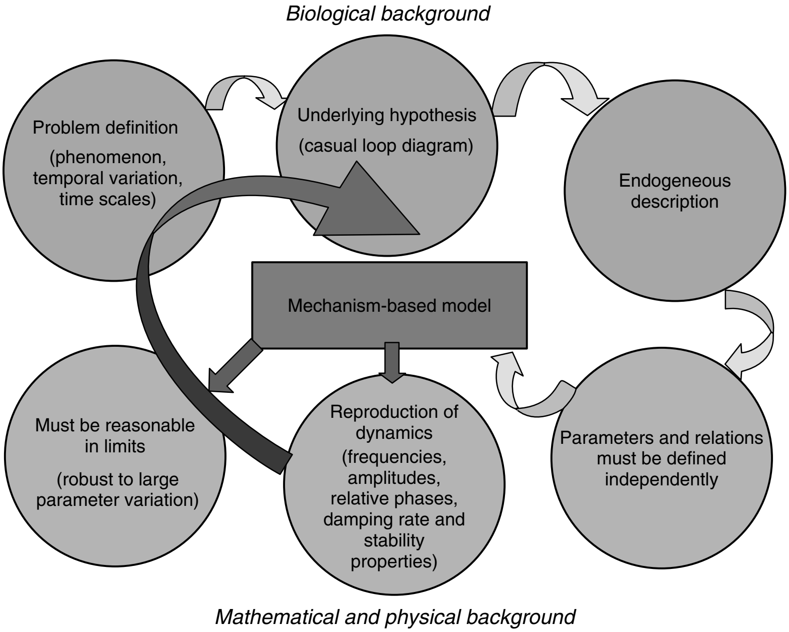
Biological background
Problem definition
(phenomenon, temporal variation, time scales)
Underlying hypothesis
(casual loop diagram)
Endogeneous description
Mechanism-based model
Must be reasonable in limits
(robust to large parameter variation)
Reproduction of dynamics
(frequencies, amplitudes, relative phases, damping rate and stability properties)
Parameters and relations must be defined independently
Mathematical and physical background

Fig. 25.1 Illustration of some of the main aspects of mechanism-based modeling as outlined in the main text. Note in particular the important role of problem definition. It is through this process that we can separate different timescales and decide what is important and what is not.

by the universities in Oxford and Auckland (Hunter and Nielsen, 2005; Noble, 2002).

Mechanism-based modeling stems from engineering and other fields of science that have strong theoretical traditions. Computer simulation allows engineers to test the function of a new construction and compare it with other possible constructions long before the first prototype is made. The results are enormous savings in time and money. Direct application of the same approach to biomedical problems is primarily restricted by the overwhelming complexity of living systems. From physics, we have inherited the idea that one can separate processes and phenomena based on their temporal and spatial scales (Figure 25.1). This is obviously much more difficult for biological systems with their many and strongly coupled processes at overlapping time and space scales. Moreover, some of the most interesting problems are precisely to explain the coordination of processes that occur over separate timescales or to explain the properties of a tissue or the function of an organ in terms of the processes and interactions of the constituting cells or functional units.

A mechanism-based model is validated through its ability to generate behaviors with amplitudes, periodicities, phase relationships, response dynamics, damping characteristics, and stability properties in accordance with the experimentally observed dynamics (Mosekilde, Sosnovtseva and Holstein-Rathlou, 2005). At the end of the day, the model is validated by its ability to predict phenomena not previously experienced.

Two different paradigms can be recognized in the approach to model construction. In a bottom-up approach, the individual base elements of the system are first described in great detail. These elements are then linked together to form larger subsystems, which in turn are linked, sometimes in many levels, until a complete model is developed at the top level. This strategy often resembles a "seed" model, where the beginning is small, but the model eventually grows in both complexity and coverage.

The theory of complex systems suggests that the rules by which parts of a system interact may be more important than the functioning of the individual parts. In a top-down approach, an overview of the system is first formulated, specifying but not detailing relevant first-level subsystems. Each subsystem is then refined in yet greater detail, sometimes in many additional levels of subsystems, until the entire specification is reduced to basic elements. The top-down approach is most appropriate in a professional simulation environment that aims at developing a large, lasting, and flawless model. The bottom-up approach is generally used in research environments where the goal is to test hypotheses and gain understanding. A judicious combination of the two, sometimes referred to as a "middle-out" approach, may turn out to be the most effective solution in some situations (Bertau, Mosekilde and Westerhoff, 2008).

The purpose of this chapter is to illustrate the use of mechanism-based modeling and Biosimulation. We will first demonstrate some of the basic aspects of the model-building process by means of a strongly simplified description of the processes involved in regulation of the male sex hormone. Testosterone is found to be released in pulses with approximately 2-hour intervals, and we want to illustrate the possible role of feedback regulation in generating such ultradian oscillations. The second example provides a much more detailed description of the mechanisms by which the individual nephron of the kidney regulates its afferent blood flow to compensate, for instance, for variations in the arterial blood pressure. Additionally, this regulation is found to lead to self-sustained oscillatory dynamics.

25.3
Biological Rhythms and the Role of Feedback Regulation

Investigations performed during the last decades have revealed a great variety of rhythms of significance for the regulation and function of normal physiological systems. Hormonal secretion provides several illustrative examples with the release of luteinizing hormone (LH) (Smith, 1980), growth hormone (Bassett and Gluckman, 1986), and insulin (Polonsky *et al.*, 1988) displaying pronounced ultradian oscillations with 2- to 3-hour periods. The release of insulin also displays a faster rhythm with a period of 9–15 minutes (Bergsten and Hellman, 1993). These rapid oscillations are likely to be of significance for the metabolic control processes in the liver, and there are a number of examples to support the idea that rhythmic administration of a hormone can be more effective than constant administration at the same average rate. It is also known that disruption of certain biological rhythms can lead to a state of disease

(Glass and Mackey, 1988) while, in other cases, synchronization of ultradian release processes can cause abnormal biological conditions such as, for instance, hot flashes (Kronenberg *et al.*, 1984).

Many hormonal secretion processes also exhibit strong 24 hours, or circadian, components. This is true, for instance, for cortisol, antidiuretic hormone, and growth hormone. The secretion of growth hormone is markedly increased during the early periods of sleep, and the secretion of antidiuretic hormone also reflects the sleep–wake cycle. The stress hormone cortisol, on the other hand, shows a higher release rate during the last stages of sleep and in the early morning. This serves as a sort of wakeup call, and insulin secretion is similarly increased around the wakening hours in order to activate the body's muscular processes. The mechanisms underlying these oscillations can often be traced back to cyclical variations in the activity of the central nervous system (Hastings, Reddy and Maywood, 2003; Levi, 2001). At the same time, the circadian rhythm modulates the above-mentioned ultradian oscillations.

Feedback regulation is an important feature of all biological systems (Murray, 2002). If a process has an effect on other processes and these processes again have an effect on the first process, then a "circular" relationship or "loop" exists. This generally helps a system to maintain stability. Regulation of the blood–glucose concentration may serve as an example. Increasing concentration of glucose in the blood after a meal stimulates the pancreas to produce more insulin. The insulin is carried with the blood around in the body and diffuses through the capillary walls into the interstitial space, causing the cells to increase their uptake of glucose. This in turn reduces the amount of glucose in the blood, thus bringing the system back toward its initial state.

The insulin–glucose feedback is said to be a negative feedback in that an initial increase in blood–glucose concentration is counteracted by the release of insulin. Negative feedback regulations are generally considered as stabilizing to the system's dynamics. At the same time they reduce the significance of parameter variations, that is, two individuals can have different values for the cellular sensitivity to insulin and yet function more or less in the same way. Positive feedback, on the other hand, is destabilizing and leads to runaway phenomena. A typical example is the growth of cancer cells, which often occurs in a nearly exponential fashion. Positive feedback regulation is not common in physiological systems. However, one can find examples where a temporary positive feedback helps to stimulate a fast activation. Peptides in the stomach, for instance, stimulate the release of chloric acid that breaks down proteins and thereby generate even more peptides. Similarly, during the release of an ovum from the ovary there is a phase in which some of the sex hormones enter into a positive feedback.

However, a negative feedback regulation is stabilizing only as long as the feedback is fast enough, that is, the variables in the system must not change much before the change is registered and corrections applied to the variables. If there is a delay in the feedback, the regulation may become unstable, and the system will start to oscillate. This is the situation in both of the two examples we are going to discuss in this chapter. The presence of a delay is also the mechanism through which many other rhythmic phenomena arise in biological systems. We say that the system has undergone a Hopf bifurcation in which the original equilibrium point has become unstable (an unstable focus), and a self-sustained oscillation (or limit cycle) has taken over the role as the stationary solution (Thompson and Stewart, 1986).

25.4
Instabilities and Far-From-Equilibrium Conditions

From a physical point of view, the rhythmic phenomena are related to the fact that biological systems are maintained under far-from-equilibrium conditions through a continuous dissipation of energy (Nicolis and Prigogine, 1977). However, nonequilibrium conditions can give rise to even more complicated behaviors. Chaotic dynamics, for instance, can arise either as a regular rhythmic process is destabilized and develops through a cascade of period-doubling bifurcations (Feigenbaum, 1980), by torus destruction in connection with the interaction of two or more rhythms with incommensurate periods (Afraimovich and Shilnikov, 1991), or via different types of intermittency (Bergé, Pomeau and Vidal, 1984).

Deterministic chaos is characterized by the facts that (i) the trajectory never repeats itself, but continues to find new ways to go and (ii) the dynamics is sensitive to small changes in the initial conditions (Bergé, Pomeau and Vidal, 1984). Deterministic chaos may resemble random noise, and it can be difficult to distinguish the two. However, while a noisy system always responds in the same way to an external perturbation, the response of a chaotic system can vary significantly over the phases of its trajectory. In relation, for instance, to the medical treatment of a patient, this implies that while in a linear system the effect of a drug will always be the same, the precise timing of the administration may be significant for a rhythmic (or chaotic system). During certain phases of the oscillation, even a major dose may have little effect while a smaller dose may be highly effective (or perhaps even toxic) during other phases. Attempts to take advantage of this phenomenon are made, for instance, in experiments with chronotherapy of cancer and other diseases (Mormont and Levi, 2003).

Synchronization is another universal phenomenon in nonlinear dynamical systems (Pikovsky, Rosenblum and Kurths, 2001), a phenomena by which two (or more) rhythmic processes tend to adjust their dynamics relative to one another so as to attain a state where they are completely entrained or, alternatively, a state in which there is a rational relation between their periods. The synchronization of our circadian rhythm to the local time and the aforementioned role of hormonal entrainment in the generation of hot flashes are typical examples. It is also well known that the beating of the heart may synchronize with the respiration, typically with 1:3 or 1:4 relations between the two periods. By forcing an oscillating system with an external periodic signal and detecting the various synchronization regions, one can obtain information about the nonlinearities at play in the system. An example of this type of work is the investigation by Sturis *et al.*(1995) of how the endogenous insulin secretion can synchronize with an externally forced periodic glucose infusion.

The significance of nonlinear dynamic phenomena seems to be even more pronounced at the cellular level. Coupling between the cells takes place via a variety of different mechanisms, including the short-range diffusive exchange of ions and small molecules through gap junction and the response of the individual cell to variations in the intercellular Ca^{2+}-concentration produced by the bursting activity of adjacent cells (Chay, 1997). Hence, one can observe synchronization of the bursting activity between neighboring cells as well as waves of cytoplasmic calcium propagating across larger groups of pancreatic cells (Gylfe, Grapengiesser and Hellman, 1991). The above-mentioned rapid oscillations in the secretion of insulin may be associated with a modulation of the bursting activity of the individual cells that arises from such intercellular interactions.

Synchronization of cellular activity is known from many other types of tissue, and transitions between different types of synchronization and between smaller and larger clusters of synchronized units may represent an important component in the overall regulation of a physiological system. In their inactive state, smooth muscle cells, for instance, are found to exhibit incoherent waves of cytoplasmic Ca^{2+} produced by an instability associated with calcium-induced calcium release. When activated, however, the cells synchronize in a regular oscillatory pattern (Peng *et al.*, 2001). On the other hand, several cases are known where a similar transition is related to the development of a state of disease. It has long been recognized, for instance, that the onset of Parkinsonian tremor is associated with a synchronization of the firing activity for a group of cells in the brain (Tass, 2001).

25.5
Luteinizing Hormone

In this section, we will use a simple example of hormonal regulation to illustrate the basic principles of systems biology and feedback control. By means of a system dynamics flow diagram (Morecroft and Sterman, 1994), we build a model of the control of the production of the most important male sex hormone, testosterone, and we use this model to discuss some of the aspects of the physiological mechanisms that cause the pulsatile behavior of the system.

System dynamics (Morecroft and Sterman, 1994) makes use of flow diagrams to illustrate the structure of interacting mechanisms that produces the dynamic phenomena we observe in managerial and economic systems. However, the same approach is applicable to biomedical systems. Flow diagrams can be considered as a language following a set of principles that must be used, in order to avoid ambiguous interpretations. In system dynamics, a system is described by a set of state variables, rates of change of these variables, auxiliary variables, exogenous variables and delays, each represented by a specific symbol in the flow diagrams.

25.5.1
Model of the Pulsatile Release of Luteinizing Hormone

In humans (and other mammals), the secretion of sex hormones, testosterone for males, and estradiol and progesterone for females, are controlled by the production of LH, which again is controlled by the production of a releasing hormone called gonadotropin-releasing hormone (GnRH). In turn, the production of GnRH appears to be controlled by the amount of circulating testosterone (Yen, Jaffe and Barbieri, 1999).

A particular interesting aspect of the gonadotropin-releasing hormone–luteinizing hormone–testosterone (GnRH–LH–T) system is that the hormone secretion takes the form of pulses. A GnRH-pulse is followed a few minutes later by an LH pulse, while it takes approximately 30 minutes for LH to stimulate the production of T to its maximum. A pulsatile release of GnRH is crucial for the system to function normally. It was shown by Knobil (1980) that constant replacement of GnRH production fails to establish normal LH concentrations. On the other hand, various diseases, such as hypothalamic amenorrhea and hypogonadotropic hypogonadism, have been treated successfully with pulses of GnRH (Leyendecker and Wildt, 1985).

System dynamics makes a clear distinction between amounts and concentrations with respect to their function in the models. The amounts of different hormones in their respective distribution volumes and their rates of change through production, degradation, transfer to other distribution volumes, and so on, are the elements of the material conservation equations for the system, and they constitute the backbone of the models. Concentrations do not satisfy similar conservation equations, and use of concentrations as basic variables can often cause confusion in models with different or time-varying distribution volumes. On the other hand, biological effects are controlled by concentrations, and the concentration of various hormones, therefore, enter the model's causal relations. Concentrations are, of course, calculated from the corresponding amounts by division by the relevant distribution volume. In this way, causal relations will always start at a level (or state) variable and end at a rate variable.

When the concentration of GnRH increases, it stimulates an increasing production of LH and, as long as the production of LH exceeds its rate of degradation in the organism, this leads to increasing amounts of LH. This first step in constructing a model is described by the flow diagram in Figure 25.2, where the valves indicate the rates

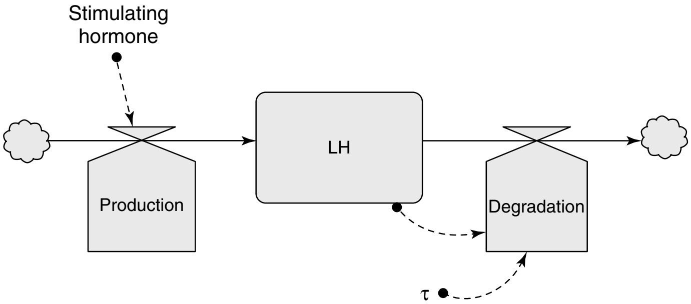

Fig. 25.2 Flow diagram of a production/degradation process. LH denotes the amount of hormone, and the arrows indicate flow directions with valves to illustrate that the flow is controlled by other variables or constants.

of production and decay. Rate-related parameters, for instance hormone lifetimes, are indicated by dashed curves pointing into the valve. The boxes represent material accumulations of a given hormone. The solid straight arrows indicate flow direction and dashed curves indicate causal dependences. The rate of degradation depends both on the lifetime and on the amount of hormone, since the more hormone is present the more hormone will degrade per unit time. The rate of hormone production depends on the current concentration of the stimulating hormone through a nonlinear relationship, which typically follows an S-shaped function. This implies that very little LH is produced at low concentrations of the stimulating hormone. Then follows a range of concentrations where the production of LH increases rapidly before it finally levels off at a maximum production rate.

In a similar way, stimulation of testosterone production in dependence of LH follows an S-shaped curve, such that increasing amount of LH stimulates an increasing production of T. An increase in T, on the other hand, is assumed to have an inhibitory effect on the production of GnRH. The system thus constitutes a negative feedback loop. A flow diagram of the described hormonal feedback system is shown in Figure 25.3. Note, that material flows are expressed by fully drawn straight arrows and causal dependences (i.e. physical and chemical laws or biological control mechanisms) are described by dashed curves. Dashed curves always emanate from a state variable and end at a rate.

These three interhormonal dependences have been experimentally studied (Veldhuis, 1991) and to a first approximation they follow the S-shaped curves shown in Figure 25.4.

From a biological point of view the most appropriate functional relationship for the S-shaped curves are Hill functions of the form

$$S(x) = \frac{Vx^n}{K^n + x^n} \tag{1}$$

where V is a saturation production rate, K is a constant that determines the position of the equilibrium point, and n a parameter that determines the slope of S at the equilibrium concentration x_0, which is also the turning point of the curve. Note, that the value of n is negative for $S_R(T)$.

The corresponding equations for the production of the three hormones are

$$\begin{aligned} \frac{dR}{dt} &= S_R(T) - R/\tau_R \\ \frac{dL}{dt} &= S_L(R) - L/\tau_L \\ \frac{dT}{dt} &= S_T(L) - T/\tau_T \end{aligned} \tag{2}$$

where the parameters τ_R, τ_L, and τ_T are hormonal lifetimes and $S(\cdot)$ are appropriately scaled S-shaped functions. Variables and indices R, L, and T refers to GnRH, LH, and T. The equilibrium point of this system is found by requiring the flows to vanish, whereby we find the equilibrium point

$$\begin{aligned} R_0 &= K_R \\ L_0 &= K_L \\ T_0 &= K_T \end{aligned} \tag{3}$$

The stability of the equilibrium is found by evaluating of the Jacobian matrix J at

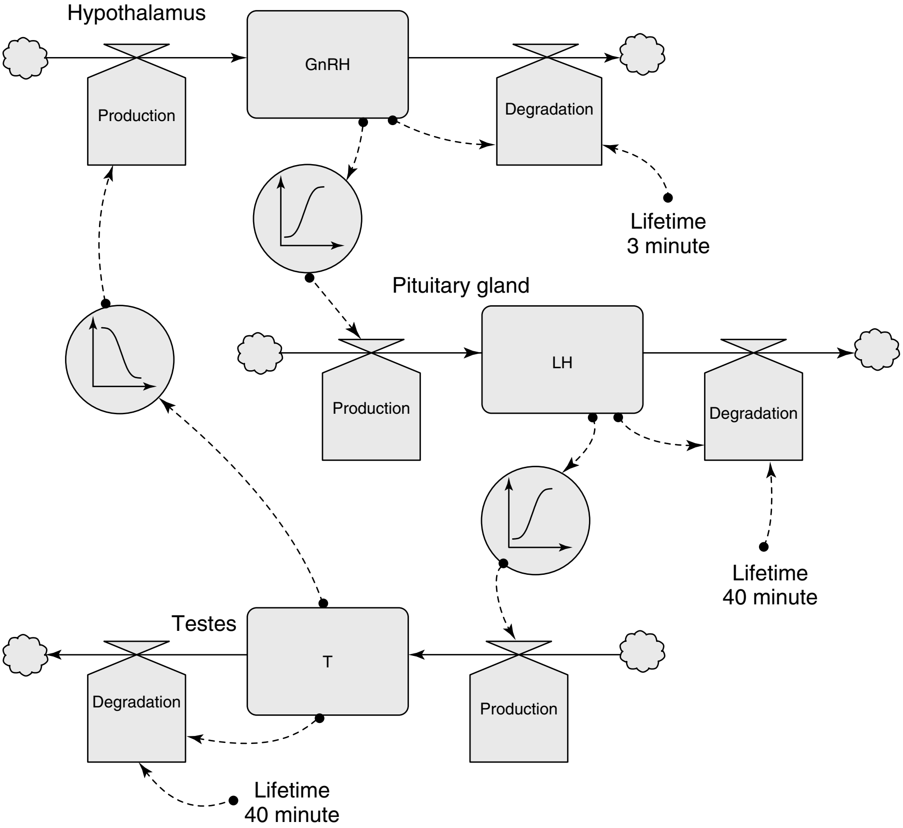
Hypothalamus
GnRH
Production
Degradation
Lifetime
3 minute
Pituitary gland
LH
Production
Degradation
Lifetime
40 minute
Testes
T
Degradation
Production
Lifetime
40 minute

Fig. 25.3 Flow diagram describing the production of the releasing hormone (GnRH), luteinizing hormone (LH), and testosterone (T) and their mutual stimulatory effects are modeled by S-functions. The effect of T on GnRH is suppressing, and this makes the overall feedback negative.

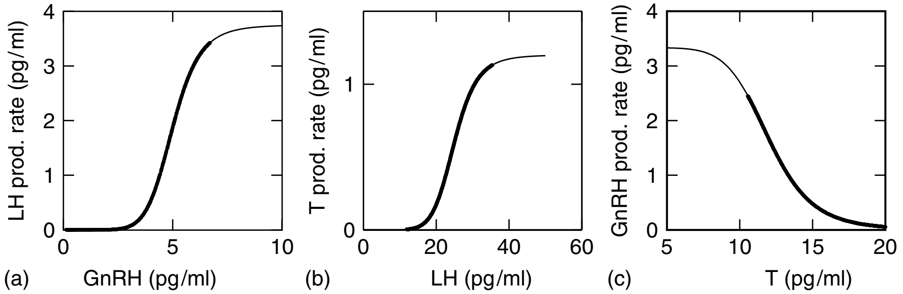

Fig. 25.4 S-shaped functions used to control the production of hormone in dependence on the concentration of the stimulating hormone. (a) and (b) are the functions controlling production of LH and T in dependence on GnRH and LH, respectively, while the negative sloped function in (c) controls the suppression of GnRH production at high concentrations of T.

the equilibrium point

$$J = \begin{pmatrix} -1/\tau_R & 0 & n_R V_R/(4K_R) \\ n_L V_L/(4K_L) & -1/\tau_L & 0 \\ 0 & n_T V_T/(4K_T) & -1/\tau_T \end{pmatrix} \tag{4}$$

and taking the determinant to obtain the characteristic polynomium. For a third-order system, this polynomium is also of third order $P(\lambda) = \lambda^3 + A\lambda^2 + B\lambda + C$, and for the present model the coefficients are

$$\begin{aligned} A &= \frac{1}{\tau_R} + \frac{1}{\tau_L} + \frac{1}{\tau_T} \\ B &= \frac{1}{\tau_R\tau_L} + \frac{1}{\tau_L\tau_T} + \frac{1}{\tau_T\tau_R} \\ C &= \frac{1}{\tau_R\tau_L\tau_T} - \frac{n_R n_L n_T}{64}\frac{V_L V_R V_T}{K_L K_R K_T} \end{aligned} \tag{5}$$

The Routh–Hurwitz stability criterion (Meirovitch, 1970) for a third-order system states that the coefficients of the characteristic polynomium must fulfill the conditions

$$A > 0,\ B > 0,\ C > 0 \text{ and } AB - C > 0 \tag{6}$$

The first three conditions are clearly fulfilled for all physical relevant values of the delays constants τ_R, τ_L, and τ_T. A more complete analysis of the system identifies the critical point, where $AB - C = 0$, as a Hopf bifurcation. The last equality can be explicitly written as

$$\frac{(\tau_T + \tau_L)(\tau_T + \tau_R)(\tau_R + \tau_L)}{(\tau_R\tau_L\tau_T)^2} > \frac{n_L n_R n_T}{64}\frac{V_L V_R V_T}{K_L K_R K_T}. \tag{7}$$

With experimentally measured lifetimes of $\tau_R = 3$ minutes, $\tau_L = 40$ minutes, and $\tau_T = 40$ minutes (Zolman, 1985) and equilibrium values of $R_0 = 5$ pg ml^{-1}, $L_0 = 25$ pg ml^{-1}, and $T_0 = 12$ pg ml^{-1} (Smith, 1980) we obtain the saturation values $V_R = 3.3$ pg (ml s) $^{-1}$, $V_L = 3.75$ pg (ml s) $^{-1}$, and $V_T = 1.2$ pg (ml s) $^{-1}$. The equilibrium point is stable for Hill exponents fulfilling the condition

$$|n_R n_L n_T| > 41.5 \tag{8}$$

By means of this inequality, we can decide the values of the slopes such that the hormonal release over time shows an oscillating behavior. It is assumed that the response of L to R is fast since the hypothalamus and pituitary gland is located relative closely to the hypothalamus in the brain, while the response of T to L is slow since L must be transported over a significant distance to the testes where the production of T takes place. Realistic values would be $n_R = -6$, $n_L = 6$, and $n_T = 6$. The time evolution of the simulated hormone concentration with these values is shown in Figure 25.5. Clearly, the behavior is oscillatory, however, the period is only approximately 50 minutes, which is more than a factor 2 too fast as compared with experimental findings. The parameters that most directly control the period are the lifetimes of the hormones. However, since the values have been measured experimentally there appears only to be one solution to the problem, namely to amend the production rate functions. The parameter V also influences the period, but this parameter has a strong influence on the amplitudes, and it is therefore not possible by varying this parameter to obtain both the amplitudes and the period correct at the same time. An alternative solution could be to examine the effect of delays that may

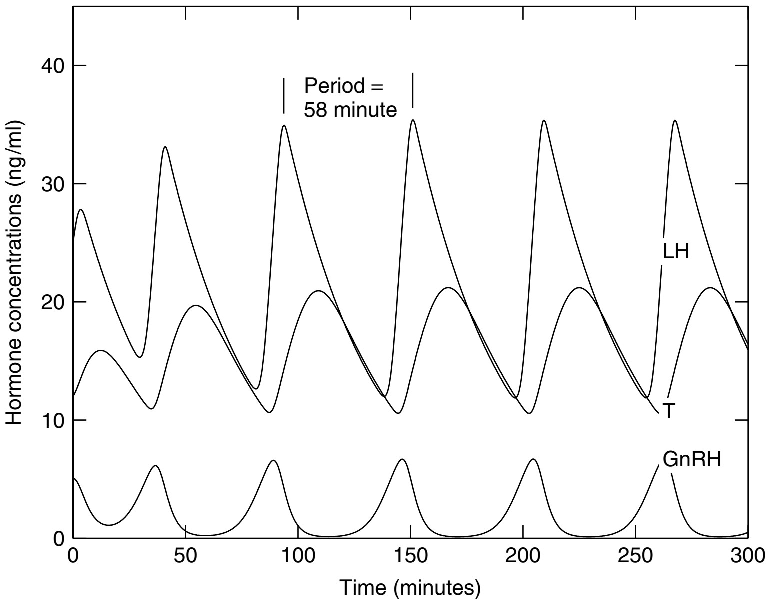

Fig. 25.5 Simulated time evolution for releasing hormone (GnRH), luteinizing hormone (LH), and testosterone (T). The parameters used in this simulation yield a period of 58 minutes between the bursts of LH secretion. In reality the period is closer to 2 hours.

be involved in the hormonal production. The presence of such delays could at one and the same time provide a longer period and a more pulsatile dynamics.

One possible way of increasing the period is to insert an extra delay in order to account for the time it takes to activate the cells to change their hormonal production.

A significant set of functions may fit the experimentally observed S-shaped hormone production response. However, few of them have a physiological interpretation in terms of biochemical processes. For instance, the logistic function, $V/(1+\exp\{K(x-x_0)\})$, could fit the relationship between hormone concentration and production rate quite well. But it is not possible to relate the exponential function to the mechanisms that control the receptor dynamics. Hill functions, on the other hand, are accepted as a mechanistic model of receptor dynamics, since there is some correspondence between the parameters of the function with the chemical constants of the biological process.

Finally, we note that there might also be an inhibitory interaction between LH and GnRH, such that high concentrations of LH may slow down the production of GnRH, whereby the full development of a cycle is retarded.

25.6
Nephron Pressure and Flow Regulation

As an example of a more detailed application of the mechanism-based approach we will consider a model that combines a description of nephron blood flow regulation with a description of the reabsorption of water and salts in the loop of Henle. A model of nephron autoregulation has been presented previously (Leyssac and Holstein-Rathlou, 1989). The purpose of that model was to examine the processes underlying the observed rhythmic and chaotic variations in pressures and flows observed for normotensive and hypertensive rats, respectively. Reabsorption of water and salts in the loop of Henle has been described in Sakai *et al.* (1986), but in a spatially extended model that is too complicated to allow detailed bifurcation analysis. The idea of the present study is to combine the two processes into a lumped model that retains sufficient simplicity to permit us to perform a more detailed bifurcation analysis.

25.6.1

Physiological Background

By regulating the excretion of salts, water, and metabolic end products, the kidneys play an important role in maintaining a suitable environment for the cells of the body. In particular, the kidneys control the plasma osmolality, the proportion of various blood solutes, the extracellular fluid volume, and the blood pressure. To protect their own function and secure a relatively constant blood flow, the kidneys also dispose of mechanisms that can compensate for variations in arterial blood pressure. It has long been recognized that this ability partly rests with controls in the individual nephron primarily the so-called tubuloglomerular feedback (TGF) (Moore, 1984). This is a negative feedback that regulates the diameter of the afferent arteriole, and hence the incoming blood flow, in dependence of the chloride or sodium concentration of pre-urine leaving the loop of Henle.

The nephron may be considered as the functional unit of the kidney. The human kidney contains about 1 million nephrons, while the rat kidney contains some 30 000 nephrons. However, the structural features, the main physiological processes, and the qualitative behavior appear to be similar to those of the human kidney. The differences are almost completely associated with the parameter values. The model we present here is a model of a single nephron initially formulated by Jensen *et al.* (1986) and later studied in greater detail by Barfred, Mosekilde and Holstein-Rathlou (1996). The intricacies of the nephron–nephron coupling, discussed in a number of recent studies (Postnov *et al.*, 2001), are neglected here.

Figure 25.6 shows the tubular structure of a nephron. The inset to the nephron from the body's vascular system is the afferent arteriole, originating from sequential branching of arteries. The blood enters the glomerulus, which extracts water and various substance, except blood cells and proteins. The extracted fluid flows further into the convoluted proximal tubulus. Approximately 70% of the glomerular filtrate is reabsorbed from the proximal tubule into the capillary network and collected by the outgoing venous system. This isosmotic process is controlled completely by a hydrostatic pressure gradient. The residual fraction becomes important for the feedback mechanism later in the filtration. The remaining 30% of the filtrate enters the thin descending loop of Henle where additionally 15% of the water is reabsorbed giving rise to an increased concentration of NaCl as the limb is impermeable to salts. On passing the

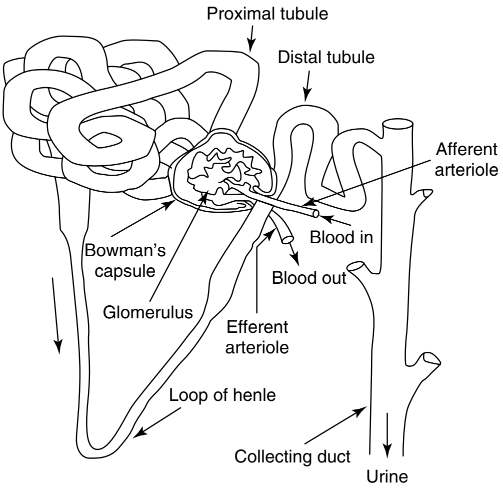

Fig. 25.6 Sketch of a nephron with its main components. The incoming blood from the afferent arteriole is filtered in the glomerulus. In the proximal tubulus a large part of the filtered fluid is reabsorbed and further reabsorption of water takes place in the descending limb of Henle, while salt reabsorption takes place in the ascending limb. At the point of contact between the distal tubulus and the glomerulus macula densa cells measure the salt concentration and send transmitters to the smooth muscle cells, which cause a constriction of the afferent arteriole at large salt concentrations.

ascending limb the filtrate is again diluted as Na^+ and Cl^- are reabsorbed, while the limb is largely impermeable to water. At the end of Henle's loop, the tubulus comes into cellular contact with the afferent arteriole of the same nephron. At these points special cells, the so-called macula densa (MD) cells, sense the Na^+ concentration. If the rate of glomerular filtration, due for instance to an elevated arterial blood pressure, is too high, the active reabsorption of NaCl is not complete, and the salt concentration at the MD is too high. The MD cells then produce transmitter agents that cause the flow resistance of the afferent arteriole to increase by radial contraction, hence lowering the glomerular pressure and thereby reducing the filtration rate. A second mechanism that contributes to the autoregulation is the myogenic response of the afferent (and possibly the efferent) arteriole. Basically it is an active response that counteracts increases in arteriolar pressure by increasing constriction/tension of the vessel caused by the smooth muscle cells in the arteriolar wall. The detailed chemical mechanics of the myogenic response are still not fully understood. In the present model a semiempirical macrophysiological description of this process is adopted.

Although most tissues in the organism dispose of a myogenic mechanism that allows them to control the incoming blood flow in relation to their metabolic and other needs, the TGF mechanism is specific to the nephrons of the kidney. This mechanism is seen as a way for the nephrons to protect their own function against variations in the arterial pressure. The causal loop diagram in Figure 25.7 illustrates the main elements of the TGF mechanism as described earlier. The feedback is negative in that increasing arterial pressures lead to higher filtration rates that result in higher NaCl concentrations at the MD leading to arteriolar contraction and reduced filtration rates. However, the feedback involves a delay of approximately 12 seconds associated with the time it takes the filtrate to flow through the loop of Henle.

Early experiments by Leyssac and Baumbach (1983) and by Holstein-Rathlou and Leyssac (1986) demonstrated

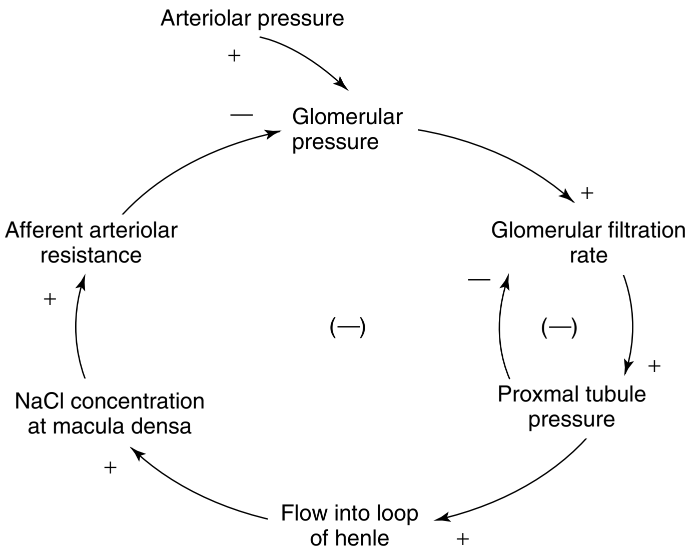

Fig. 25.7 Causal loops for TGF. Minus indicates a negative influence and plus indicates a positive influence in the arrow direction. Altogether the loop represents a negative feedback. However, because of the delay associated primarily with the fluid flow through the loop of Henle, and because of a relatively high gain factor the regulation tends to become unstable and produce self-sustained oscillatory dynamics.

that the feedback regulation could become unstable and generate self-sustained oscillations in the proximal intratubular pressure with characteristic periods of 30–40 seconds. Although for normotensive rats the oscillations had the typical appearance of a limit cycle, highly irregular oscillations were found for spontaneously hypertensive rats (SHRs). Oscillations were observed for 50–80% of the investigated nephrons. For nonoscillatory nephrons, self-sustained oscillations could be elicited by microperfusion with artificial tubular fluid, that is, by artificially increasing the rate of flow through the loop of Henle. Subsequently, it has been observed that similar irregular oscillations can develop for normal rats, if the arterial pressure is increased by ligating the blood supply to the other kidney.

25.7
Single Nephron Model

In this section we construct the model step by step and collect all the subsystems into one coherent description of the tubular and arteriolar system. We have already presented the fluid flow through the nephron. In the following the total function of the nephron is grouped into specific submodels representing glomerular filtration rate (GFR), flow through the loop of Henle, the arteriolar resistance, and a macroscopic model of the smooth muscle cells in the afferent arteriole. This description will follow the structure of the corresponding flow diagram as depicted in Figure 25.8.

In Figure 25.8 (left), we follow the flow of blood from the arterial system through the afferent arteriole, the glomerular capillaries, and the efferent arteriole to the venous system. The two arterioles are represented by their flow resistances of which the efferent resistance is assumed to be constant. The afferent flow resistance varies as part of the pressure and flow regulation. The blood flow through the glomerular capillaries is determined by the pressure difference between the arterial and venous sides and the two flow resistances. The glomerular capillaries are represented as a (small) compliant blood volume. The GFR depends on the difference between the glomerular pressure and the proximal tubular pressure accounting, of course, for the difference in protein

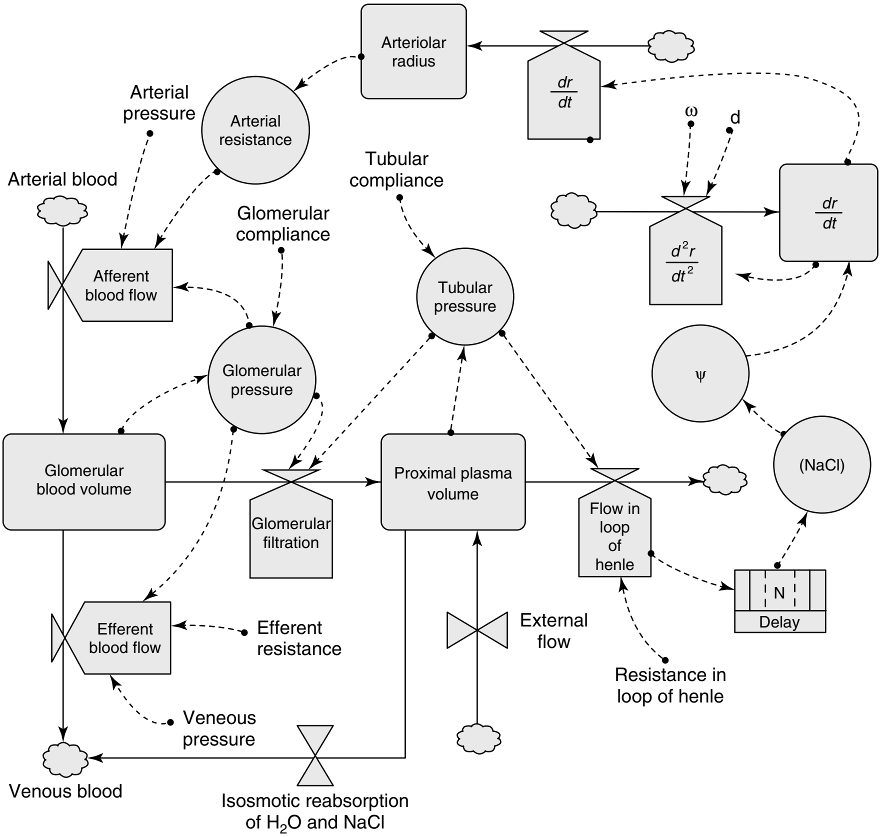
Arteriolar radius
$\frac{dr}{dt}$
Arterial pressure
Arterial resistance
ω
d
Arterial blood
Tubular compliance
$\frac{dr}{dt}$
Glomerular compliance
$\frac{d^2r}{dt^2}$
Afferent blood flow
Tubular pressure
Glomerular pressure
ψ
Glomerular blood volume
Proximal plasma volume
(NaCl)
Glomerular filtration
Flow in loop of henle
N
Delay
Efferent blood flow
Efferent resistance
External flow
Resistance in loop of henle
Veneous pressure
Venous blood
Isosmotic reabsorption of H_2O and NaCl

Fig. 25.8 Flow diagram for the single nephron model. There are four main flows in the diagram: (i) The blood flow through the glomerular capillaries, (ii) the glomerular filtration, (iii) the flow through the loop of Henle, and (iv) the rate of change of arterial radius and its second-order rate of change. The feedback on the arterial blood flow is realized by the dependence of the arterial resistance on the arterial response to pressure. Description of the processes that take place in the loop of Henle is here reduced to a single flow rate and a delay. A lumped representation is shown in greater detail in Figure 25.9.

osmotic pressure across the capillary wall. It is assumed that the blood reaches pressure equilibrium before it leaves the glomerular capillaries. The equation for the proximal pressure is derived from the in- and outflows to the tubule

$$\frac{dP_t}{dt} = (F_{filt} - F_{reab} - F_H)\,\frac{1}{C_t} \tag{9}$$

where F_{filt} denotes the flow of filtrated fluid, F_{reab} the reabsorption, F_H the flow from the proximal tubule in the loop of Henle, and C_t is the proximal tubular compliance. The large reabsorption of both water and salt in this early tubule is represented in the flow diagram by the arrow pointing from the tubular fluid volume to the venous system. The salt concentration remains approximately constant in the proximal tubule as the tubular wall is equally permeable to water and salts.

The loop of Henle is modeled as a tube divided into a number of consecutive compartments, each with their specific physiological properties. In order to account for the mechanistic differences between water and salt reabsorption, the amounts of water and salt must be considered separately as shown in Figure 25.9, where the upper part of the chain is the water volumes and flows between the compartments and the lower part represents the amounts of salt in the various compartments along the chain. Again, we model this system by the mass conservation of the filtrate for each compartment and apply our knowledge about the physical properties of the tube and fluid, which for instance are the compliance and resistance of the tube and the transit times between the compartments. The flow rate of water entering the first compartment of the thin descending limb from the proximal tubules is the inset to the flow diagram. This flow rate depends on the compliance and resistance of this compartment and on a delay time representing the time needed for the fluid to move from the proximal tubulus to the first compartment. Reabsorption of water into the capillary system takes place through the interstitium, however this process is considered to fall outside the boundaries of the model, since it only has little influence on the TGF mechanism. Reabsorption of water is osmotic and therefore depends on the concentration difference across the tubular wall and on the permeability of the wall.

In Figure 25.9 the lumped loop model is depicted with two descending compartments, each with reabsorption of water. The ascending limb is represented by three compartments. While there is no reabsorption of water in the ascending limb, reabsorption of salt takes place according to usual Michaelis–Menten kinetics. However, there is also an osmotic backleak of salt from the interstitium into the tubule. This backleak has only little influence on the dynamics of the model when operated in the normal working regime. However, if the flow through the loop becomes very small, the reabsorption of water will decrease, while the active salt reabsorption remains almost unchanged, giving rise to an increased salt gradient across the ascending limb. The backleak mechanism seeks to re-establish the normal salt balance by allowing paracellular diffusion.

At the end of the loop in Figure 25.9 the amounts of both water and salt yields a salt concentration which passes the MD cells at the intersect of the distal and juxtaglomerular apparatus, permitting

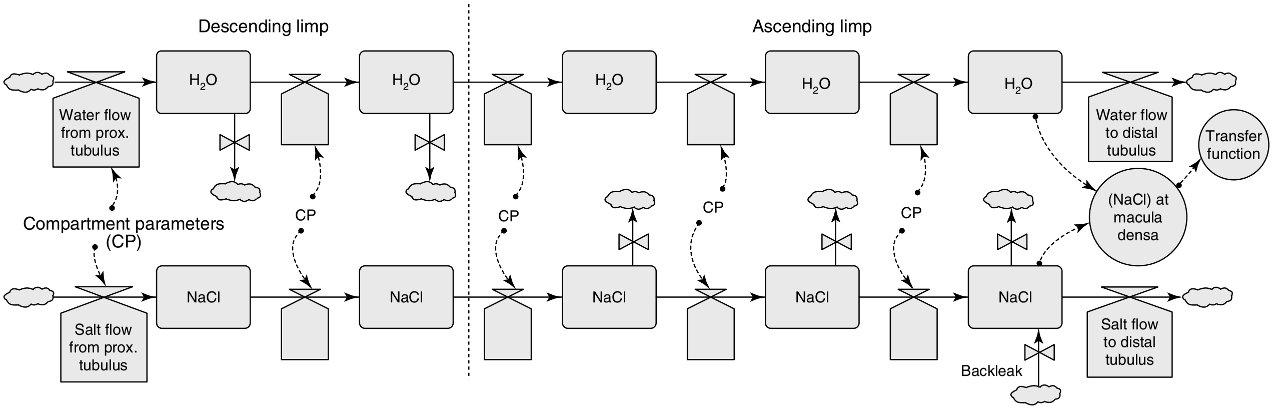
Descending limp
Ascending limp
H_2O
H_2O
H_2O
H_2O
H_2O
Water flow from prox. tubulus
Water flow to distal tubulus
Transfer function
Compartment parameters (CP)
CP
CP
CP
CP
(NaCl) at macula densa
NaCl
NaCl
NaCl
NaCl
NaCl
Salt flow from prox. tubulus
Backleak
Salt flow to distal tubulus

Fig. 25.9 Flow diagram with two descending and three ascending compartments in the loop of Henle. The incoming flows come from the proximal tubule and the outgoing flows enter the distal tubule close to the point where the macula densa cells are located. Reabsorption is symbolized by the rates between water/salt amounts and interstitial volume, which here is assumed to be an infinite sink.

the arteriolar model to respond to the NaCl concentration through the TGF transfer function.

Several experiments have shown that the radius of arterioles may perform spontaneous or self-sustained oscillations (Funk *et al.*, 1983; Aalkjær and Nilsson, 2005), that is, independent of the pressure variation due to TGF. According to Poiseuille's law, the resistance of the arteriole is inversely proportional to the fourth power its radius. Hence, constrictions of the arteriolar radius cause a reduction in blood flow into the glomerulus. This mechanism is modeled by a second-order differential equation of the form

$$\frac{d^2r}{dt^2} + d\frac{dr}{dt} = \omega^2 \frac{P_{\mathrm{trans}}}{P_0} \tag{10}$$

where r is the normalized radius, d is a damping coefficient, P_{trans} the transmural pressure, ω a characteristic angular frequency, and P_0 a normalization pressure. The transmural pressure is a quite complicated variable to describe for this system. It is expressed as the difference between the average pressure P_{av} of the fluid in the vessel and the equilibrium pressure P_{eq} for which the arteriole is in equilibrium with this pressure at its present radius and muscular tone.

$$P_{\mathrm{trans}} = P_{\mathrm{av}} - P_{\mathrm{eq}} \tag{11}$$

On top of the elastic reaction of the arteriolar wall P_{el}, activation of the smooth muscle cells in this wall gives rise to an additional contribution ψP_{ac} to the equilibrium pressure that the wall can sustain at a given radius, that is,

$$P_{\mathrm{eq}} = P_{\mathrm{el}} + \psi[\mathrm{NaCl}]P_{\mathrm{ac}} \tag{12}$$

Here, $\psi[\mathrm{NaCl}]$ is the strength by which the active constriction takes place. If the salt concentration at the MD, [NaCl] is low, then ψ is also low, thus reducing the contraction of the afferent arteriole and, hence, increases the glomerular filtration. The dynamics of the active contribution, better known as the myogenic response, is discussed in the following paragraphs. The corresponding equations are represented in the flow diagram of Figure 25.8 by two rates labeled d^2r/dt^2 and dr/dt with corresponding state variables.

The myogenic response is a process that is best understood on the cellular level. It is often observed that models on some space and time level need to include processes on other levels. We have already seen one example of this: the Michaelis–Menten transport of salt in the ascending loop of Henle. This process can also be described on the chemical level, but the thermodynamic average causes a macroscopic relationship to be a sufficient description of the process. By using a similar approach, Feldberg and Colding-Jorgensen (1995) have derived a macroscopic relationship between the constriction of arterioles and blood pressure, which includes the effect of the smooth muscle cells. The model is based on the assumption of nonlinear elastic stress–strain relationship and a linear relationship between the muscle tone and the active stress component.

The degree to which the afferent arterial resistance is changed due to an increased glomerular filtration is called the TGF transfer function (or loop gain). As mentioned earlier the values returned by the function is called the muscular tone. Actually, the TGF transfer function can be defined as a function of any variable along the loop of Henle; however, it is mechanistically most obvious to let it be a function of salt concentration at the site of the MD cells, since this is the terminal of the tubular system relevant for the autoregulation mechanism. The function is an S-shaped function with lower and upper limits, turning point and a slope at the turning point which can be determined by microperfusion experiments (Leyssac and Holstein-Rathlou, 1989). Especially the slope and upper limit of the function show characteristic differences between normotensive and hypertensive rats.

25.8
Simulation and Analysis

The model we have discussed in the preceding sections can be formulated mathematically as a set of ordinary differential equations. For the model with two descending and three ascending compartments of the loop of Henle 18 equations are required. Two of the equations describe the oscillations of the arterial wall and one describes the pressure in the proximal tubulus. Each compartment, whether descending or ascending, is described by three equations representing fluid volume, amount of salt, and the delay of the fluid transport between the compartments. Computer simulations of the equations show a great variety in the dynamical behavior.

The TGF transfer function has been experimentally recorded by inserting a plug of wax in the middle of the proximal tubulus of the rat, so that the flow in the tubular system is blocked. On the other side of the wax plug a micropipette is inserted and artificial tubular fluid is infused at a constant rate. With time, the system approaches a steady state and the pressure in the early proximal tubule is recorded.

With our mathematical model simulations of the microperfusion experiments can be made by setting the GFR to zero and setting the flow into the loop of Henle to a constant rate, representing the externally forced infusion.

Complete bifurcation analyses are hard to perform on systems of the size of the present model. However, biological systems are often infected by noise making it very difficult, if not impossible, to experimentally observe, for instance, a second period-doubling bifurcation after the first period doubling has occurred. For the present model observations of a steady state, a cycle of double period and chaos allows us to limit the analysis to bifurcations leading to these dynamical behaviors. Figure 25.10 shows a bifurcation analysis with the total delay time T and strength α of the TGF transfer function as the parameter. For small α values the system is in equilibrium with no time-dependent variations. As α is increased, the equilibrium point becomes unstable in a Hopf bifurcation, and a limit cycle is born. As it is clear from the figure, a further increase of α may give rise to a cycle of double period followed by a period-doubling cascade leading to chaotic dynamics. The existence of all four types of behaviors is important for the validity of the model. But additional validation is required in order to claim that the model and parameters used are correct. As mentioned earlier, the period of the oscillations has been measured to be in the range 30–40 seconds. The period of the simulated oscillations depends mainly on the delay time T. For delays in the range 9–13 seconds, the

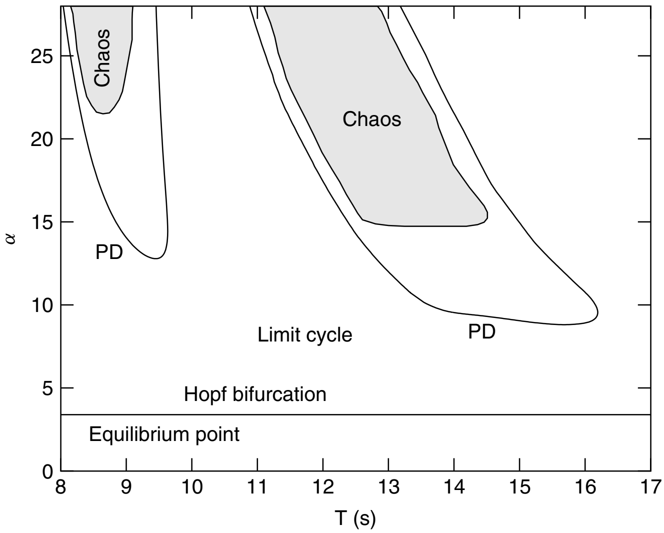

Fig. 25.10 Two-parameter bifurcation diagram for the extended nephron model showing the Hopf bifurcation at $\alpha \approx 4$, two tongues delineated by period-doubling bifurcations, and the regions dominated by chaotic dynamics.

period resembles those measured experimentally, a range that partly covers the region with all of the four types of dynamics (Figure 25.10). The scan through the diagram at constant $T = 12.5$ seconds is displayed. Figure 25.11 shows the dynamical behavior in more detail. The period doubling at $\alpha = 14$ is followed by a cascade of period-doubling transitions leading to deterministic chaos. At larger α a reverse period-doubling cascade terminates the chaotic region, a new cascade is initiated, and new regions with chaos appear. At $\alpha = 12.5$ and $\alpha = 17$, the periodic and chaotic attractors are shown (A and B) in Figure 25.11.

Fast oscillations of the proximal pressure and arterial wall arise as a consequence of the pulsatile contractions of the smooth muscle cells when the arterial pressure is high. For the period-1 cycle shown in Figure 25.11 the fast oscillations are clearly visible, as the orbit makes five fast cycles for each of the slow TGF-mediated oscillations. The number of fast oscillations per cycle of the slow oscillations depends on the parameters of the model, and as we increase the value of α (or T) we move from a region with 5:1 dynamics into a region with six oscillations for each slow oscillation.

The idea of creating a lumped model of the loop of Henle is to model the salt concentration along the loop. With the chosen number of compartments, we obtain average concentrations within each compartment, and parameters specific to the tubules wall anatomy are either integrated over the compartment length if originally specified as a value per length or divided by the number of compartments if originally given as an effective parameter for the whole loop.

The temporal evolution of the tubular pressures along the loop is shown in Figure 25.12 for a delay of $T = 12.5$

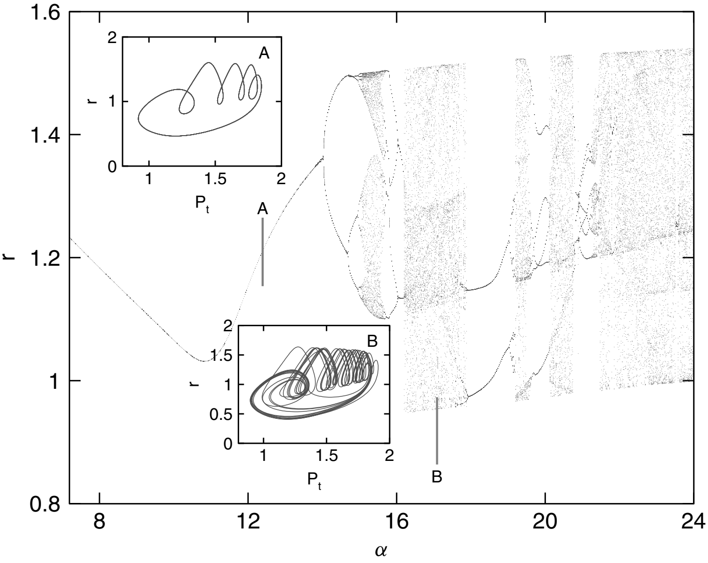

Fig. 25.11 Diagram showing the bifurcations and solutions as the TGF gain α is varied while the delay is kept constant at $T = 12.5$ seconds. At low α, the solution is oscillatory with a period of 38 seconds. As α increases, a period doubling takes place at $\alpha \approx 14$. Thereafter a cascade of period doublings leads to chaotic dynamics interrupted by periodic windows.

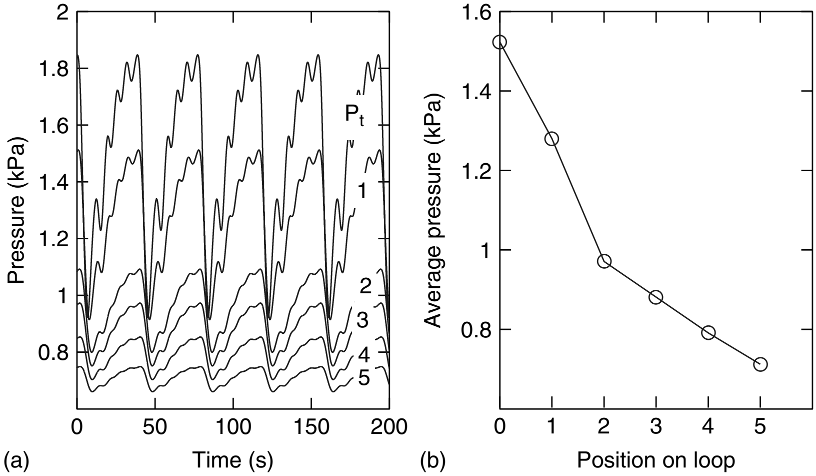

Fig. 25.12 Tubular pressures. (a) Time evolution of the pressures along the loop beginning with the proximal pressure (labeled P_t). Oscillations have a period of approximately 38 seconds. Note, how the amplitude of the oscillations is increasingly damped. (b) Time average of the pressures along the loop.

seconds and a TGF feedback gain of $\alpha = 10$. For this value of the total delay the period is 38 seconds, and we also see that there is a phase shift of 11.5 seconds between the proximal pressure and the distal NaCl concentration. The phase shift depends on the tubular compliances along the loop, since a highly compliant tubular segment can accommodate fluid at high pressure and release it as the pressure decreases. Also the decreasing amplitude of the pressure is accounted for by the relative large compliance in connection with the flow resistance.

Variation in time of the salt concentration along the loop is shown in Figure 25.13. In the proximal tubule, the concentration is defined constant equal to 150 mM in accordance with experimental knowledge. With appropriate absorption rates, the concentration increases in the descending limb to 350 mM in the second compartment corresponding to the bend of the loop. In the subsequent compartments that represent the ascending limb the Michaelis–Menten reabsorption of NaCl accounts for the decrease toward a concentration of 30 mM at the end of the loop.

Both this phase shift and the presence of period doubling and chaos depend on the compliance of the loop of Henle. For increasing compliance the phase shift increases and the period doubling and chaotic regions appear at larger values of α. Compliances measured in vitro (Sakai *et al.*, 1986) have indicated that the Henle compliance may be approximately 20 times larger than the proximal compliance. However, simulations with these larger values completely destroy the picture of the above findings. This seems to demonstrate that, when living tissue is taken out of its natural environment, the properties may change drastically. One should also not forget that the tubular system of the nephron is surrounded

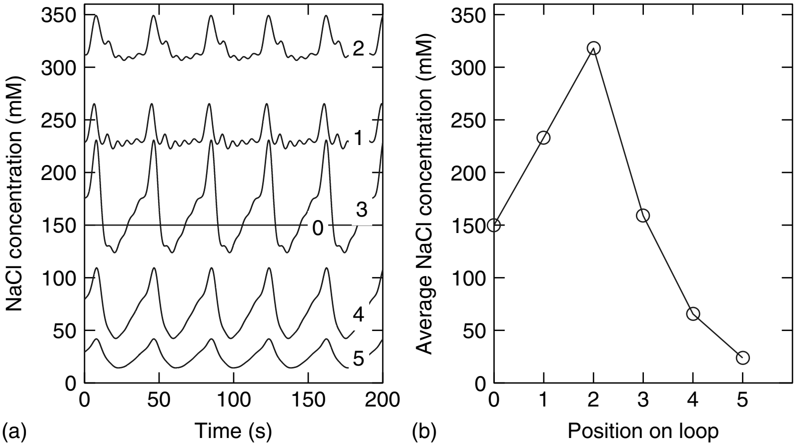

Fig. 25.13 NaCl concentrations. (a) The time evolution of the NaCl concentration along the loop with the proximal concentration defined constant equal to 150 mM. Along the loop a phase shift evolves and at the distal tubulus the phase shift between the proximal pressure and the distal salt concentration is 11.5 seconds. (b) Time average of the concentrations. Clearly, the concentration increases in the descending limb (0–2) and decreases in the ascending limb (3–5).

by the tissue of the interstitium, which may have a large effect of the dynamical properties.

In this chapter, we have described a lumped model of a single nephron in the kidney. This model combines a description of the interacting myogenic and TGF-mediated mechanisms of nephron autoregulation with a description of the reabsorption of water and salts in the loop of Henle. The following are the key features of nephron autoregulation we have aimed to reproduce:

- Four types of qualitative different dynamical behaviors: equilibrium state, oscillatory, oscillatory with double period, and chaotic.
- A phase shift in the oscillation between the proximal pressure and distal NaCl concentration of 10–15 seconds.
- TGF transfer functions obtained by microperfusion experiments on normo- and hypertensive rats.
- Increasing NaCl concentration along the descending limb and decreasing NaCl concentration along the ascending limb of Henle's loop.
- Slow and fast oscillations of periods 30–40 seconds and 5–6 seconds, respectively, the fast oscillations being caused by the model for the afferent arteriole.

Other aspects of the nephronic system are described in our recent publications (Sosnovtseva *et al.*, 2003; Sosnovtseva *et al.*, 2007).

25.9
The Promise of Systems Biology

Understanding the complex dynamic phenomena of the living world is one of the main scientific challenges of the coming decades. The way to acquire such understanding must necessarily involve a more direct and effective integration of experimental and computational research. Hypothesis-driven or mechanism-based modeling is an approach in which the physiological, pathological, and pharmacological processes of relevance to a given problem are represented as directly as possible. This approach allows us to (i) test whether assumed hypotheses are consistent with observed behavior, (ii) examine the sensitivity of a system to parameter variation, (iii) learn about processes not directly amenable to experimentation, and (iv) predict system behavior under conditions not previously experienced. By virtue of the direct physiological interpretation of the involved parameters, mechanism-based modeling provides a tool for translation between species and interpolation among groups. This implies, for instance, that the pharmaceutical industry to a certain extent can predict human responses to a new drug from animal experiments by replacing the physiological parameters with parameters relevant to man.

Systems biology is a scientific approach that seeks to understand how all the individual components of a biological system interact in time and space to determine the functioning of the system. This involves description of

- the structure of a biological system; that is, the significant physiological components and their structural relationships at the appropriate spatial scale;
- the dynamical behavior of the system under different conditions and external perturbations;

- the mechanisms that control the states and behavior of the system;
- the causes of system changes in connection with various diseases and the methods by which such changes can be counteracted or compensated.

A model of a biological system can only be expected to reproduce experimentally observed features, if the mechanisms responsible for these features are implemented in the model. Therefore, a model can never be expected to reproduce all experimental findings. However, a model is usually build with the intension to model specific phenomena and behaviors. If some of these key features fail to be produced, a revision of the model is required.

References

Aalkjær, C. and Nilsson, H. (2005) *Br. J. Pharmacol.*, **144**, 605–16.

Afraimovich, V.S. and Shilnikov, L.P. (1991) *Amer. Math. Soc. Transl.*, **149**, 201–12.

Alon, U. (2007) *An introduction to Systems Biology: Design Principles of Biological Circuits*, Chapman & Hall, London.

Baldi, P. and Brunak, S. (1998) *Bioinformatics: The Machine Learning Approach*, MIT Press, Massachusetts.

Barfred, M., Mosekilde, E., and Holstein-Rathlou, N.-H. (1996) *Chaos*, **6**, 280–87.

Bassett, N.S. and Gluckman, P.D. (1986) *J. Endocrinol.*, **109**, 307–12.

Bertau, M., Mosekilde, E., and Westerhoff, H. (eds) (2008) *Biosimulation in Drug Development*, Wiley-VCH, Weinheim.

Bergé, P., Pomeau, Y., and Vidal, C. (1984) *Order within Chaos*, John Wiley & Sons, Inc., New York.

Bergsten, P. and Hellman, B. (1993) *Biochem. Biophys. Res. Commun.*, **192**, 1182–88.

Chay, T.R. (1997) *Biophys. J.*, **73**, 1673–88.

Feigenbaum, M.J. (1980) *Comm. Math. Phys.*, **77**, 65–86.

Feldberg, R. and Colding-Jorgensen, M. (1995) *Am. J. Physiol.*, **269**, F581–F593.

Funk, W., Endrich, B., Messmer, K., and Intaglietta, M. (1983) *Int. J. Microcirc. Clin. Exp.*, **2**, 11–25.

Glass, L. and Mackey, M.C. (1988) *From Clocks to Chaos: The Rhythms of Life*, Princeton University Press, Princeton.

Gylfe, E., Grapengiesser, E., and Hellman, B. (1991) *Cell. Calcium.*, **12**, 229–40.

Hastings, M.H., Reddy, A.B., and Maywood, E.S. (2003) *Nat. Rev. Neurosci.*, **4**, 649–61.

Holstein-Rathlou, N.-H. and Leyssac, P.P. (1986) *Acta Physiol. Scand.*, **126**, 333–39.

Hunter, P.J. and Nielsen, P. (2005) *Physiology*, **20**, 316–25.

Jensen, K.S., Mosekilde, E., and Holstein-Rathlou, N.-H. (1986) Mondes en Developement, 54/55, 91–109.

Kitano, H. (2001) *Foundations of Systems Biology*, MIT Press, Cambridge.

Knobil, E. (1980) *Recent Prog. Horm. Res.*, **36**, 53–88.

Kronenberg, F., Cote, L.J., Linkie, D.M., Dyrenfurth, I., and Downey, J.A. (1984) *Maturitas*, **6**, 31–43.

Levi, F. (2001) *Lancet Oncol.*, **2**, 307–15.

Leyendecker, G. and Wildt, L. Proceedings of the 3rd Ferring Symposiom, Noordwijk, Netherlands, september 11–13(1985).

Leyssac, P.P. and Baumbach, L. (1983) *Acta Physiol. Scand.*, **117**, 415–19.

Leyssac, P.P. and Holstein-Rathlou, N.-H. (1986) *Pflügers Arch.*, **407**, 285–91.

Leyssac, P.P. and Holstein-Rathlou, N.-H. (1989) *Pflügers Arch.*, **413**, 267–72.

Meirovitch, L. (1970) *Methods of Analytical Dynamics*, McGraw-Hill, New York.

Moore, L.C. (1984) *Am. J. Physiol.*, **247**, F267–F276.

Morecroft, J.D.W. and Sterman, J.D. (1994) *Modeling for Learning Organizations*, Productivity Press, Portland.

Mormont, M.C. and Levi, F. (2003) *Cancer*, **97**, 155–69.

Mosekilde, E., Sosnovtseva, O.V., and Holstein-Rathlou N.H.(2005) *Basic Clin. Pharmacol. Toxicol.*, **96**, 212–34.

Murray, J.D. (2002) *Mathematical Biology*, Springer, New York.

Nicolis, G. and Prigogine, I. (1977) *Self-Organization in Nonequilibrium Systems*, John Wiley & Sons, Inc., New York.

Noble, D. (2002) *Science*, **295**, 1678–82.

Peng, H., Matchkov, V., Ivasen, A., Aalkjaer, C., and Nilsson, H. (2001) *Circ. Res.*, **88**, 810–15.

Pikovsky, A., Rosenblum, M., and Kurths, J. (2001) *Synchronization–A Universal Concept in Nonlinear Sciences*, Cambridge University Press, New York.

Polonsky, K.S., Given, B.D., Hirsch, L.J., Tillil, H., Shapiro, E.T., Beebe, C., Frank, B.H., Galloway, J.A., and Van Cauter, E. (1988) *N. Engl. J. Med.*, **318**, 1231–39.

Postnov, D.E., Sosnovtseva, O.V., Mosekilde, E., and Holstein-Rathlou, N.-H. (2001) *Int. J. Mod. Phys. B*, **15**, 3079–98.

Sakai, T., Craig, D.A., Wexler, A.S., and Marsh, D.J. (1986) *Biophys. J.*, **50**, 805–913.

Smith, W.R. (1980) *Bull. Math. Biol.*, **42**, 57–78.

Sosnovtseva, O.V., Postnov, D.E., Mosekilde, E., and Holstein-Rathlou, N.-H. (2003) *Chaos Solitons Fractals*, **15**, 343–69.

Sosnovtseva, O.V., Pavlov, A.N., Mosekilde, E., Yip, K.-P., Holstein-Rathlou, N.-H., and Marsh, D. (2007) *Am. J. Physiol.*, **293**, F1545–F1555.

Sturis, J., Knudsen, C., O'Meara, N.M., Thomsen, J.S., Mosekilde, E., Van Cauter, E., and Polonsky, K.S. (1995) *Chaos*, **5**, 193–99.

Tass, P.A. (2001) *Biol. Cybern.*, **85**, 343–54.

Thompson, J.M.T. and Stewart, H.B. (1986) *Nonlinear Dynamics and Chaos*, John Wiley & Sons, Inc., New York.

Veldhuis, J.D. (1991) The hypothalamic-pituitary-testicular axis in *Reproductive Endocrinology*, 3rd edn (eds S.S.C. Yen and R.B. Jaffe), Saunders, Philadelphia.

Yen, S.S.C., Jaffe, R.B., and Barbieri, R.L. (1999) *Reproductive Endocronology: Physiology, Pathophysiology, and Clinical Management*, Saunders, Philadelphia.

Zolman, J. (1985) *Am. J. Physiol.*, **248**, R312–R319.

Further Reading

British Academy of Medical Sciences and the Royal Academy of Engineering (2007) *Systems, Biology: a vision for engineering and medicine*. http://www.acmedsci.ac.uk/p99puid97.html.

European Science Foundation (2007) *Systems, Biology: A Grand Challenger for Europe.* http://www.esf.org/research-areas/medical-sciences/activities/science-policy/.

Special issue (2008) *Biomedical Applications of Systems, Biology and Biological Physics. Phil. Trans. Roy. Soc. A*, **366.**

26
Organic Superconductors

Claus Schelde Jacobsen

26.1 **Introduction** **869**
26.2 **What is Superconductivity?** **870**
26.2.1 Basic Properties 870
26.2.2 Thermodynamics, Condensation, and Order Parameter 870
26.2.3 Microscopic Understanding 872
26.2.4 Molecular Superconductivity and Little's Proposal 873
26.2.5 History of Superconducting Materials 874
26.3 **Organic Superconductors** **876**
26.3.1 Organic Conductors 876
26.3.2 One-Dimensional Physics and Instabilities 878
26.3.3 Quasi-One-Dimensional Organic Superconductors 881
26.3.4 Quasi-Two-Dimensional Organic Superconductors 885
26.4 **Conclusion and Outlook** **887**
Glossary 888
References **890**
Further Reading **891**

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

26.1
Introduction

A superconductor is crudely speaking a material that can conduct electric currents without losses, that is, without dissipation. Fritz London (1900–1954), one of the greatest contributors to the understanding of superconductivity, in his book on superfluids (London, 1950) speculates about the possible occurrence of the phenomenon in biological macromolecules. After the quantum mechanical foundation for understanding superconductivity had been developed (Bardeen, Cooper, and Schrieffer, 1957), William Little pursued the idea and argued that properly designed organic polymers might show superconductivity at ambient temperatures (Little, 1964). These ideas spurred an enormous interest in developing and studying electrically conducting organic materials, and beginning in 1980 (Jerome *et al.*, 1980) several families of organic superconductors were synthetized and investigated in great detail. Typically they superconduct at very low temperatures (most often below 10 K), and to date there is no convincing evidence for superconductivity in biological materials, except for reports on the so-called proximity induced superconductivity in DNA (Kasumov *et al.*, 2001). However, the activity in the area has continued partly because of lack of a deep understanding of the inorganic high-temperature superconductors, which work at moderate but still low temperatures (around 100 K). It is hoped that the high flexibility of organic chemistry would lead to the fulfillment of the old promise of a room-temperature superconductor.

It is the aim of this short review to describe the main advances in the field of organic superconductivity. We start out with a brief introduction to the phenomenon of superconductivity, on the macroscopic as well as the microscopic scale. For a more detailed treatment, the reader is referred to standard textbooks on the subject (see, for example, Tinkham, 1996). Next, an overview of the development of the various groups of superconducting materials is presented. In the main sections, we deal with the basic properties of the superconducting organics including their rich behavior in terms of ordered phases, which to a large extent is due to the reduced effective dimensionality encountered. We do not include materials with π-bound carbon sheets like C-60 compounds and carbon nanotubes, which are hardly more organic

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

than graphene and graphite. Finally, we conclude with an outlook for future directions in the field.

26.2
What is Superconductivity?

26.2.1
Basic Properties

Superconductivity is associated with an ordered phase of the conduction electrons in a material. The main signatures for an ideal superconductor are the absence of electrical resistance (up to a certain critical current) and the Meissner effect (Meissner and Ochsenfeld, 1933) – the expulsion of magnetic fields (up to a critical magnetic field strength, H_C). If the critical values are exceeded, the material reverts to the normal resistive state. As with other ordered phases, an increasing temperature favors a high entropy, that is, a disordered state. For a superconductor in zero field, this is reflected by the fact that there is a continuous phase transition at a well-defined transition temperature, T_C, above which superconductivity is lost. These basic properties are illustrated in Figure 26.1.

26.2.2
Thermodynamics, Condensation, and Order Parameter

Experimentally, it is found that the specific heat is exponentially activated below T_C. After a drop at T_C, it follows the behavior of a normal metal. This suggested a two-fluid model (Gorter and Casimir, 1934), where starting at T_C a macroscopic fraction of the conduction electrons condenses into a collective state of zero resistance and zero entropy. The fraction increases with decreasing temperature and approaches unity at $T << T_C$. At a given temperature, statistical equilibrium between condensate and normal (“evaporated”) electrons exists. This approach requires a condensation or binding energy for the condensate electrons. It also allows the establishment of a relation between the critical magnetic field and the difference in free energy between the normal fluid and the superconducting fluid. The phenomenological approach to superconductivity culminated with the

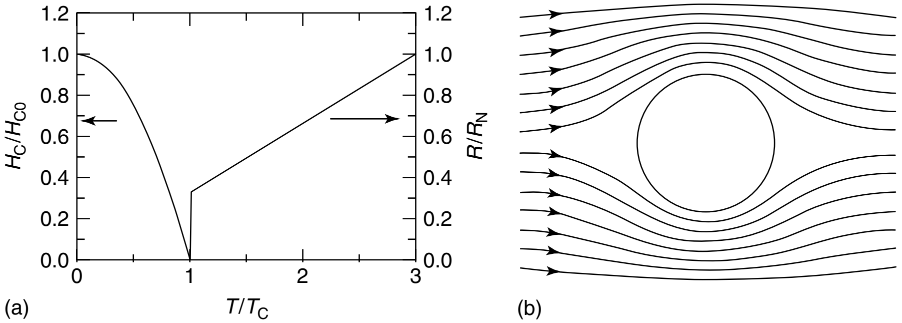

Fig. 26.1 (a) Behavior of resistance and critical magnetic field as function of temperature for ideal model superconductor. The values are normalized to the room-temperature and zero temperature values, respectively. (b) Schematic picture of the Meissner state. The magnetic field lines are expelled from the superconductor.

two equations of Fritz London and Heinz London, which account for the fascinating electrodynamics of superconductors (London and London, 1935). The first of the London equations simply states the absence of friction when the superfluid accelerates in an electric field, while the second is the integral with respect to time of Faraday's law of induction with the additional postulate that the constant of integration is zero. The London equations fully account for the Meissner effect, since they show that a superconductor in the ideal Meissner state carries a superconducting sheet current in the surface, which gives rise to a magnetic field contribution in the interior, which exactly cancels the applied field. The current density is not infinite; thus, the magnetic field has a penetration depth, λ, which from the London equations is

$$\lambda = \sqrt{\frac{m}{\mu_0 n_s e^2}} \tag{1}$$

In this expression, m and e are the electron mass and charge, respectively, $\mu_0 = 4\pi \times 10^{-7}$ Vs Am^{-1} is the vacuum permeability, while n_s is the density of electrons in the superfluid. Typical densities give numbers for the penetration depth in the range 50–200 nm well below T_C (i.e. corresponding to many interatomic distances). As the condensate density decreases with increasing temperature, the field penetrates deeper, and finally $\lambda \to \infty$ as the critical temperature is approached.

Systems undergoing continuous phase transitions are often described in terms of an order parameter. In a ferromagnet, for example, the order parameter is the magnetization vector, a three-dimensional quantity. For superconductors, it is found that a two-dimensional order parameter suffices. It is conveniently given in complex form as

$$\Psi(\mathbf{r}) = \sqrt{n_s(\mathbf{r})}\exp[i\theta(\mathbf{r})] \tag{2}$$

hence with an amplitude equal to the square root of the superfluid density, but with a variable phase, θ. For simplicity, we have assumed stationary conditions, but allowed for spatial variations. It turns out that the order parameter may be interpreted as a genuine macroscopic quantum mechanical wave function describing the superfluid and it is fully consistent with the London equations.

The order parameter is the basis of development of a much more detailed phenomenological theory for superconductivity (Ginzburg and Landau, 1950). The free energy may be expanded as a power series in the order parameter, and minimization leads to the two Ginzburg–Landau equations that integrate the thermodynamic, the quantum mechanical, as well as the electrodynamical properties of superconductors. Although in principle they are only valid near T_C, they work fine in a much wider range and constitute a standard tool for physicists working in this area. They still do not explain the physical origin of the condensation energy, which, however, as noted above is simply related to the critical magnetic field.

Before turning to the microscopic theory, let us introduce another fundamental length, the coherence length ξ (Pippard, 1953), which is the shortest distance over which the magnitude of the order parameter can vary significantly (e.g. dropping from a finite level to zero). A simple expression for ξ may be obtained from the uncertainty principle. Pippard assumed that only electrons with energies in a range kT_C around the chemical potential (the Fermi energy) participate in the ordering, which appears at T_C. The corresponding momentum range is kT_C/v_F, v_F being the velocity of the electrons at the Fermi level. Hence, we may estimate

$$\xi \cong \frac{\hbar}{\Delta p} = \frac{\hbar v_F}{kT_C} \tag{3}$$

The coherence length is quite material dependent. For example, in aluminum ($T_C = 1.2$ K) it is 1600 nm, while in the commonly used niobium ($T_C = 9$ K) it is 40 nm. The relation reflects an important issue, namely, that the higher the T_C, the shorter the coherence length will be. In the high-temperature superconductors, it is of order of a few nanometers. Thus, much of the robustness of superconductivity, which is due to the fact that the order parameter cannot be destroyed over short length scales, is lost as the transition temperature and the condensation energy increase. The coherence length may also be derived from the Ginzburg–Landau equations, showing that it diverges when approaching T_C. To some extent, the expression in Eq. (3) may be understood as the low-temperature extrapolation of the Ginzburg–Landau result.

26.2.3
Microscopic Understanding

After many unsuccessful attempts, the now universally accepted Bardeen–Cooper–Schrieffer (BCS) theory of superconductivity appeared in 1957 (Bardeen, Cooper, and Schrieffer, 1957). At that time it was realized that a weak indirect attraction between electrons in a metal existed. One electron traveling through a particular spot will start polarizing the surrounding ionic lattice. Because of the heavy ions, this is a fairly slow process; hence, when another electron passes by a little later, it will experience a lowering of its energy (Fröhlich, 1950). Because the two electrons in question need not be close in distance, the effective attraction may dominate over the direct Coulomb repulsion.

No matter what is the origin of the net attraction, it makes the Fermi-sea of normal electrons unstable to the formation of Cooper pairs (Cooper, 1956). In the simplest picture, Cooper pairs combine electrons of opposite momentum and spin. We will skip the details, which rest on a combination of a net attraction and on the restrictions on occupancy of the electronic states imposed by the Pauli principle. It is important to understand that a Cooper pair does not consist of two particular bound electrons. Rather there is a rapid dynamical exchange of electrons near the Fermi level. As might be expected, there is a close relation between the size of a Cooper pair and the superconducting coherence length (Eq. (3)).

The BCS theory proposes a ground state wave function where all electrons are combined into Cooper pairs, but with occupancies depending on the wave vector. The elementary excitations are combined electron- and holelike with a threshold energy (half the energy gap, $E_G = 2\Delta$) and statistical mechanics provided in the simplest weak coupling limit the following expressions for the energy gap and transition temperature:

$$E_G = 2\Delta = 3.5kT_C \cong 4\hbar\omega_0 \times \exp\left[-\frac{1}{N(0)V}\right] \tag{4}$$

The relation between the energy gap, which in a sense represents the binding energy of a Cooper pair, and the transition temperature is a general consequence of the BCS theory being a mean-field theory, where all fluctuations are neglected. The expression on the right-hand side contains the prefactor ω_0, which is the characteristic frequency of the mediator of the interaction (in the model discussed, it is the lattice vibration (phonon) frequency), and in the exponential function we find the density-of-states at the Fermi level for the metal, $N(0)$, and the effective attractive interaction at play near the Fermi surface, V. This already provides some guidelines for optimizing T_C: pick a mediator of high frequency, select a metal with a high density-of-states, and make sure that the effective interaction is as high as possible.

The BCS theory in many ways explains the behavior of all known superconductors. For example, the energy gap for excitations prevents scattering of the conduction electrons, and hence predicts the zero resistance state. It is possible to show that the Ginzburg–Landau equations follow from the BCS theory; so all the electrodynamics, and so on, is accounted for.

A particularly important experimental observation concerns the Cooper pair formation. A consequence of the electrodynamics of superconductors is that magnetic flux through a superconducting ring is quantized in units of h/q_s, where q_s is the charge of the "particle" making up the superfluid. The flux quantization is fairly easy to measure, and in all known cases it has been found that $q_s = 2e$, consistent with the existence of pairing.

Although the BCS theory seems to be highly successful in providing a general framework for understanding superconductivity, it is very nonspecific regarding the detailed microscopic mechanism responsible for the attraction between electrons. The mediators of the attraction could well be magnetic or purely electronic with quite different energy scales. The traditional test for the involvement of the ionic lattice has been studying the isotope effect. The vibration frequency, ω_0, in Eq. (4) scales with the inverse of the square root of the ionic mass. Thus, on substituting isotopes, the transition temperature should change in a systematic matter, which has been confirmed in many of the traditional low-temperature superconductors. However, in strong coupling, high-temperature superconductors the effect may be masked, because a simple expression such as Eq. (4) no longer applies, and several subtle side effects of substitution have also been observed.

Exotic mediators may also result in more complex Cooper pair formation. The traditional Cooper pair has a spherically symmetric *s*-wave orbit and a spin singlet (antiparallel spin directions). In interpreting results for several unconventional superconductors, *p*-wave and *d*-wave symmetries as well as triplet states (parallel spins) are encountered. They are all compatible with the BCS theory, but it is quite difficult to distinguish them experimentally. However, there is rather strong evidence for substantial *d*-wave pairing in the ceramic high-temperature superconductors.

26.2.4
Molecular Superconductivity and Little's Proposal

One signature of superconductivity is the occurrence of persistent currents in closed rings. They can be set up by cooling a ring in a magnetic field to below T_C and then switching off the field. The enclosed flux will be trapped. Now it is generated by a persistent current in the ring, which will continue to run until the ring is heated to above T_C. Likewise, a superconducting ring being exposed to a magnetic field will set up a persistent current to keep flux out of the ring. Sometimes it has been argued that the diamagnetic screening current in, for example, benzene molecules, which is also persistent in the sense that it continues to run as long as a magnetic field is present, may be explained by the fact that the π-electrons in the ring have superconductivity. However, this is a simple quantum mechanical consequence of having switched on a magnetic field combined with the unavailability of other non-current-carrying states in the same energy range, and is also known to occur in nanoscale normal metal ring structures. In a truly superconducting ring, the decay of a persistent current is due to the vanishing probability that all the Cooper pairs change their momentum state simultaneously, which is necessary to retain the BCS state. Thus the physics involved is quite different, and we will only by superconductivity understand superfluid formation involving a condensation energy.

There are several approaches to maximize T_C in designing a material. The basic idea of Little (1964) was to use long unsaturated polyene chains and attach, on the side of the chains, small and highly polarizable molecules, for example, dye molecules. The mediators of attraction between conduction electrons on the chains would then be essentially electronic transitions (excitons) on the dye molecules with a frequency, ω_0, in the optical range, which is obviously much higher than any lattice frequency. Also, if the chemistry allowed the dye molecules to be placed close enough to the chains, the interaction V would be appreciable. According to Little's sensible theoretical estimates, transition temperatures of order 1000 K should be achievable! Unfortunately, there is no firm evidence for this kind of excitonic superconductivity to date, but the proposal has had – and to some extent still has – a strong impact on research activities.

26.2.5
History of Superconducting Materials

For natural reasons there has been a lot of interest in the development of superconducting transition temperatures. Figure 26.2 shows a few of the more well known advances in terms of T_C versus year of discovery. It all started in 1911, when Kamerlingh-Onnes shortly after liquefying helium for the first time found mercury (Hg) to be superconducting below 4.2 K. Several other elemental metals (squares) were found to exhibit superconductivity, including niobium (Nb) with $T_C = 9.5$ K. These are the so-called type I superconductors, where theoretically $\xi > \sqrt{2}\lambda$. Type I superconductors have typically a rather low T_C and display the simple magnetic field behavior outlined above with a Meissner state below a single critical magnetic field, which is also rather low, typically 0.01–0.2 T. This ruins their applications in high-current, high-magnetic-field applications.

Indeed, in materials where ξ is smaller than λ, a different behavior is found. This is related to the fact that the interface energy at a boundary between the normal metal (N) and the superconductor

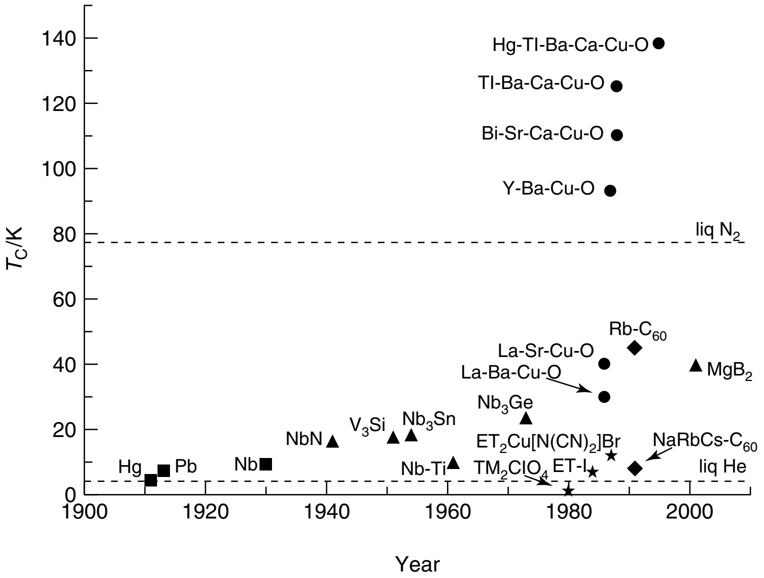

Fig. 26.2 Schematic picture of the historical development of superconductors: transition temperatures versus year of discovery. Squares: elemental superconductors. Black triangles: alloys and intermetallics. Stars: organic conductors. Circles: ceramic copper-oxide (high temperature) superconductors. Diamonds: doped C-60 (Bucky balls). The temperatures of liquid helium and liquid nitrogen are indicated with dashed lines [Sources: http://superconductors.org/ and Butch *et al.*, 2008.].

(S) goes from being positive to being negative. When ξ is small, the boundary layer is almost completely superconducting; hence, condensation energy is gained, while the magnetic field is allowed to penetrate to a greater depth. The fact is that it becomes advantageous for the material to have as much NS-interface as possible. For very low fields (below a lower critical field, H_{C1}), the Meissner state with complete flux expulsion is still found. Above H_{C1}, magnetic flux lines quantized in units of $\Phi_0 = h/2e$ start penetrating the bulk of the material. These entities are called *vortices*, since they have a normal core of approximate radius ξ, while the surrounding superconductor is shielded from the penetrating magnetic flux by a current vortex of approximate radius λ. For increasing magnetic field, the vortex density increases, and only when the normal cores start to overlap, superconductivity breaks down. Thus this happens at an upper critical field, which may be estimated as $H_{C2} = \Phi_0/\mu_0 \pi \xi^2$. In the so-called mixed state between the two critical fields, the vortices tend to arrange themselves in a two-dimensional triangular lattice (Abrikosov, 1957). In such type II superconductors, the upper critical field is often tens of tesla and in the high-temperature superconductors it may be hundreds of tesla.

In Figure 26.2 the triangles give examples of traditional type II superconductors, often alloys like Nb–Ti ($T_C = 9.5$ K) or intermetallic compounds like Nb_3Sn

($T_C = 18$ K). These two materials are the ones most often used for high-current, high-field applications (e.g. for magnetic field coils in magnetic resonance imaging MRI systems), but they still require expensive cooling most often in the form of liquid helium.

In 1973, Nb_3Ge with $T_C = 23$ K was found and for a number of years this remained the record. Then in the 1980s, the organic superconductors appeared on the scene (the stars in Figure 26.2), but the transition temperatures have so far remained modest. The discovery of the ceramic high-temperature superconductors in 1986 revitalized the field and the magic temperature of liquid nitrogen was soon exceeded. Currently a lot of research and development focuses on the technological development of these materials. Finally Figure 26.2 gives examples of doped fullerenes (C-60) as well as the recently discovered MgB_2 ($T_C = 39$ K).

26.3
Organic Superconductors

26.3.1
Organic Conductors

The flexibility of organic chemistry encouraged researchers to work with semiconducting organic materials as early as in the 1950s. For a review of this early work, see, for example, Gutmann and Lyons (1967). Excluding polymers, which so far have not shown bulk superconductivity, the overwhelming majority of organic conductors are based on charge transfer complexes. Usually organic molecules are closed shelled, so when incorporated in a crystal the electronic structure will consist of completely filled and completely empty energy bands. Thus insulators result. A charge transfer complex contains an electron donor (low ionization potential) and an electron acceptor (high electron affinity). The charge transfer provides unpaired charge carriers. Donor or acceptor or both may be organic, or one of them inorganic (for a thorough review, see Jerome and Schulz, 1982; see also Batail, 2004).

To achieve delocalization of charge carriers in the crystal, often π-bound rather planar molecules are used. A number of representative molecules are shown in Figure 26.3. The π-orbitals are oriented perpendicular to the molecular plane and tend to favor the formation of molecular stacks with significant intermolecular contacts within the stacks.

The first really excellent organic metal TTF–TCNQ (tetrathiafulvalene–tetracyanoquinodimethane (Ferraris *et al.*, 1973)) has a schematic crystal structure as shown in Figure 26.4.

The energy balance within the crystal results in an average transfer of 0.55–0.59 electron from TTF to TCNQ. Both donor and acceptor stacks form narrow one-electron bands of widths of order 0.6 eV (Jacobsen, 1988). Such a small bandwidth suggests that short-range electron–electron repulsion is important. The room-temperature conductivity along the stacks is of the order of 60,000 $(\Omega m)^{-1}$, about thousand times smaller than that of copper. Perpendicular to the stacks it is 100–1000 times lower. In an early work, the stacking axis conductivity was found to increase rapidly with decreasing temperature, reaching a maximum of order 500 times the room-temperature value near 60 K, but below this temperature the material becomes

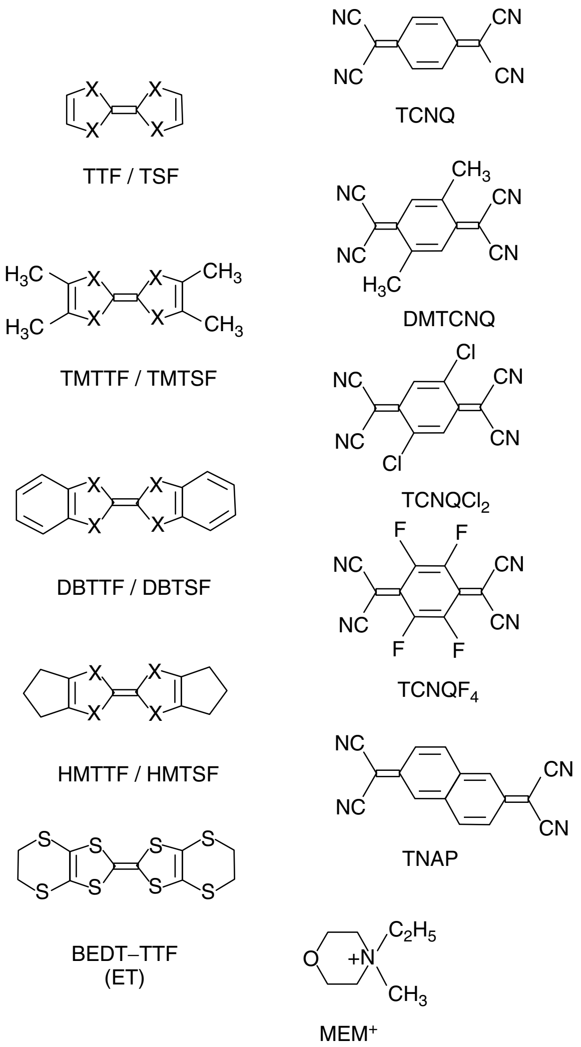
TTF / TSF
TMTTF / TMTSF
DBTTF / DBTSF
HMTTF / HMTSF
BEDT–TTF (ET)
TCNQ
DMTCNQ
TCNQCl2
TCNQF4
TNAP
MEM+
X
S
NC
CN
CH3
H3C
Cl
F
O
+N
C2H5

Fig. 26.3 Examples of donor (left) and acceptor (right) organic molecules used in building organic metals and semiconductors. X in the top four donor molecules can be sulfur or selenium. TT(S)F, tetrathia(selena)fulvalene; TMTT(S)F, tetramethyl-tetrathia(selena)fulvalene; DBTT(S)F, dibenzo-tetrathia(selena)fulvalene; HMTT(S)F, hexamethylene-tetrathia(selena)fulvalene; BEDT–TTF, bis(ethylenedithio)–tetrathiafulvalene; TCNQ, tetracyanoquinodimethane; DMTCNQ, dimethyltetracyanoquinodimethane; TCNQCl$_2$, dichlorotetracyanoquinodimethane; TCNQF$_4$, tetrafluorotetracyanoquinodimethane; TNAP, tetracyanonaphtoquinodimethane. MEM$^+$, methyl-ethyl-morpholinium (closed shell ion).

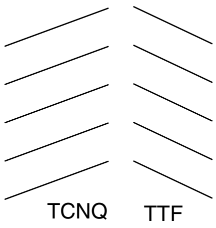

Fig. 26.4 Schematic stacking arrangement of TCNQ and TTF in the 1 : 1 charge transfer salt.

a semiconductor with a now quickly decreasing conductivity (Coleman *et al.*, 1973). The peak was interpreted as arising from superconducting fluctuations and the low-temperature semiconducting state resulting from a Peierls instability (see below). The reported unusually high-conductivity peak may have been a consequence of measurement errors easily encountered when dealing with highly anisotropic systems, but relative rise by a factor of 15–20 is usually found in pure crystals, and this is quite difficult to explain within a single electron scattering model.

26.3.2
One-Dimensional Physics and Instabilities

The Peierls instability is quite typical for a quasi-one-dimensional system like TTF–TCNQ and is seen in many similar materials. Hence, we will briefly describe the physics involved. The electronic structure of, for example, the TCNQ stack may be represented as shown in Figure 26.5.

As far as interstack interactions can be neglected, the one-electron band structure is truly one-dimensional. There is virtually no dispersion perpendicular to the stacks. Owing to the small electronic overlap along the stacks, a tight binding approximation is excellent and at low temperature the band is filled to the Fermi wave vector, k_F, which is at about 30% of the zone boundary. Peierls (1955) was the first to demonstrate that such a one-dimensional metal is unstable in the presence of a finite electron–lattice (–phonon) interaction. This is easily seen by considering the effect of a 2 k_F-potential on the electronic structure (see Figure 26.6).

The result is the opening of an energy gap, E_G, right at the Fermi level. From the density-of-states, it is evident that this new nonmetallic state has a lower electronic energy than the original metallic state shown in Figure 26.5. The system may introduce such a 2 k_F-potential through the electron–phonon coupling by creating a periodic lattice distortion. It can be shown that while the gain in energy will scale with $E_G{}^2 \ln(E_G)$, the cost in elastic energy only varies with $E_G{}^2$. Hence, the system is unstable for any magnitude of the electron–phonon coupling. It is important to emphasize that the instability depends on the ability for a potential at a single wave vector to open an energy gap over the entire Fermi surface. This may still be possible for a weak dispersion perpendicular to the chain direction but is not in two- or three-dimensional systems.

If 2 k_F is commensurate with the crystal lattice, the distortion will be locked to the underlying original lattice and a metal–insulator transition is predicted. However, if, as is the case

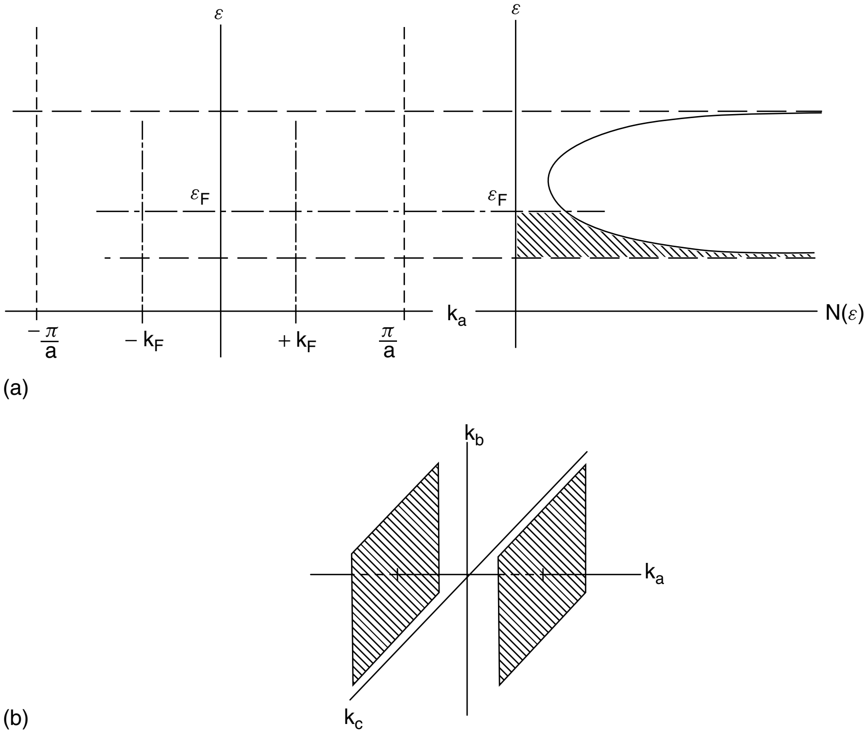

Fig. 26.5 (a) Quasi-one-dimensional electronic band structure in the TCNQ stack in TTF–TCNQ. The electronic wave number $k = 2\pi/\text{wavelength}$, while a is the lattice constant along the stack. The symbol ε is the one-electron energy. The right-hand drawing shows the density-of-states. The available states are filled to the Fermi level, ε_F, except for a thermal smearing of order a few kT. The corresponding wave number is k_F. (b) The Fermi surface (i.e. the transition between occupied and unoccupied states in **k**-space) simply consists of two planar and parallel sheets separated by 2 k_F.

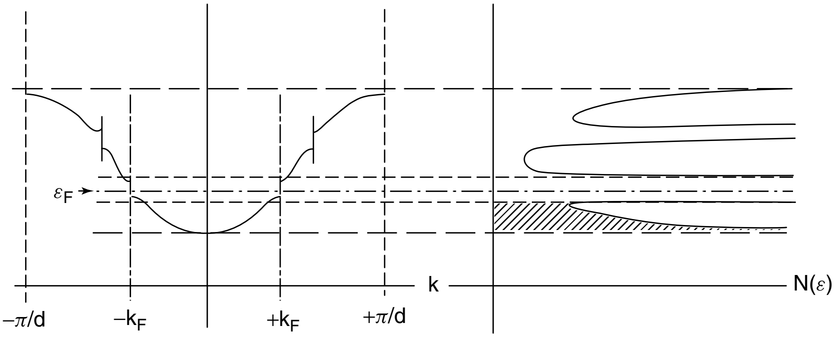

Fig. 26.6 Effect of a 2 k_F periodic lattice distortion along the stack on the quasi-one-dimensional band structure of the TCNQ stack in TTF–TCNQ. A gap opens over the entire Fermi surface and the electronic energy is lowered.

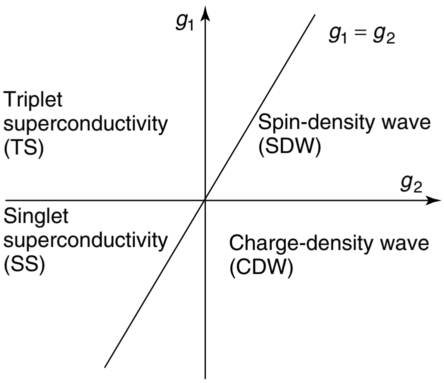
g_1
$g_1 = g_2$
Triplet superconductivity (TS)
Spin-density wave (SDW)
g_2
Singlet superconductivity (SS)
Charge-density wave (CDW)

Fig. 26.7 "G-ology" map of ground states in an idealized one-dimensional metal. The coupling constants g_1 and g_2 describe electron–hole interactions with momentum transfers near $2k_F$ and near 0, respectively.

with TTF–TCNQ, $2\,k_F$ is incommensurate with the underlying lattice, the energy of the lattice distortion is independent of the exact position along the stack, and the coupled electron–lattice distortion (called a charge-density wave (CDW)) may slide through the crystal without friction. Since the energy gap prevents scattering of the individual electrons, this is a mechanism for superconductivity, originally proposed by Fröhlich (1954). The sharp conductivity peak in TTF–TCNQ is sometimes attributed fully or partially to the Fröhlich mechanism. In any case, it simultaneously demonstrates that Fröhlich conductivity is susceptible to pinning both by defects and by ordering caused by three-dimensional interactions (e.g. Coulomb interactions). Any such effect will tend to lock the CDWs to the lattice, thus turning the material into an insulator. This is in contrast to the BCS mechanism for superconductivity, which is quite robust to nonmagnetic defects.

The CDW instability is just one example of possible instabilities in one-dimensional metals. Arbitrary electron–electron interactions may be assumed and characterized by two coupling constants g_1 and g_2. Owing to the one-dimensionality, low-energy electron– hole interactions are possible only with momentum transfers near $2\,k_F(g_1)$ and near 0 (g_2). Mean-field calculations then lead to a "phase diagram" for the ground state as schematically shown in Figure 26.7 (see, for example, Jerome and Schulz, 1982).

As is seen, spin-density waves (SDW) as well as CDWs, and both singlet and triplet superconductivity may be encountered. Again these ground states are genuine one-dimensional results, although the real systems when temperature goes to zero must always be influenced by three-dimensional effects.

Also it must be borne in mind that long range order in a one-dimensional system can only exist at zero temperature. Phase transitions will always involve higher dimensional effects, and hence the transitions tend to be suppressed significantly below the temperatures predicted by mean-field theories. In the temperature range between the mean-field scaling temperature and the real transition temperature, important one-dimensional fluctuations are likely to occur. This is presumably how the conductivity behavior of TTF–TCNQ should be understood. Between 60 and

150 K, the conductivity is enhanced by fluctuating short-ranged CDWs.

26.3.3
Quasi-One-Dimensional Organic Superconductors

Historically, superconductivity in organic crystals was first found in the Bechgaard salts $TMTSF_2X$, where X is an inorganic anion like PF_6^-, AsF_6^-, SbF_6^-, BF_4^-, NO_3^-, and so on (Bechgaard *et al.*, 1980). These are selenium-based sister compounds of the sulfur-based $TMTTF_2X$ family (Delhaes *et al.*, 1979). The characteristic crystal structure is shown in Figure 26.8. Apart from the electronic instabilities described below, those materials with less than octahedral symmetry of the anions are also prone to various anion ordering phenomena, which can have a drastic influence on the electronic properties (for an extended review see, for example, Jerome, 2004).

Again molecular stacks are formed, but because of the anions a zigzag pattern is seen. The overlap between stacks is small in the c'-direction (passing the anions), but may be considerable in the b'-direction. This is evidenced by

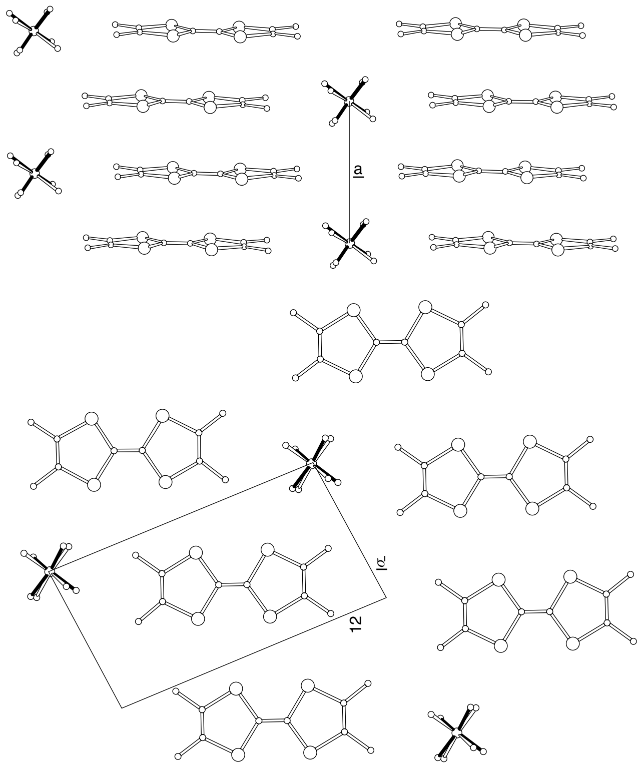

Fig. 26.8 Crystal structure of $TMTSF_2ClO_4$, the first ambient pressure organic superconductor. (a) Side view on the stacks and (b) view along the stacks. Two possible orientations of the perchlorate-ions are indicated. The ions are ordered in the superconducting state (Soling *et al.*, 1981).

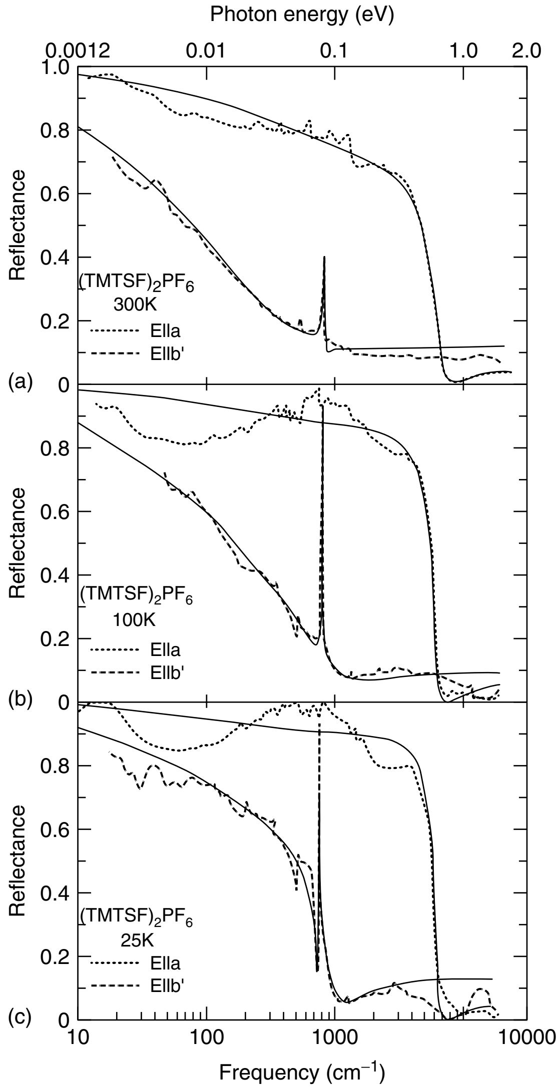

Fig. 26.9 Polarized infrared reflectance spectra of TMTSF$_2$ PF$_6$ at three temperatures: (a) 300 K, (b) 100 K, and (c) 25 K. The rise of a plasma edge along *b*′ is noted. The solid lines are Drude model fits (Reproduced with permission from Jacobsen, Tanner, and Bechgaard, 1981. Copyright American Physical Society.).

the optical reflectance studies shown in Figure 26.9, which in contrast to, for example, TTF–TCNQ and TMTTF$_2$X, show metallic reflectance both in the stacking direction and at low temperature in the *b*′-direction as well.

The data suggest a bandwidth anisotropy of order 10 and consequently a

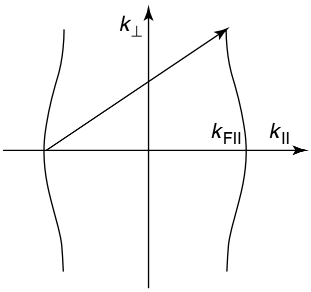

Fig. 26.10 Schematic drawing showing the warping of the Fermi surface (cross section shown) in the $TMTSF_2X$ conductors. The warping is due to significant interstack interactions.

warping of the Fermi surface such as that in Figure 26.10. Note that the Fermi surface is still open and may be almost completely nested by a single wave vector. Thus the basic electronic structure is still susceptible to the characteristic one-dimensional instabilities, and the materials are classified as being quasi-one-dimensional.

It is quite interesting to compare the isostructural $TMTSF_2X$ and the $TMTTF_2X$-families. Although the larger selenium atom gives a larger intermolecular overlap, the stacking axis bandwidth in the selenium-based compounds is just 20% bigger than it is in the sulfur-based compounds (Jacobsen, 1988). The main qualitative difference seems indeed to be in the interstack contacts, allowing coherent transport in the selenium-based family. The difference in transport properties is drastic. In Figure 26.11 we present the stacking axis resistivities as a function of temperature for $TMTSF_2$ PF_6 and $TMTTF_2$ PF_6.

The resistivity of $TMTTF_2$ PF_6 increases strongly on cooling, a behavior understood as being caused by a dominating short-range Coulomb interaction, which prevents delocalization of the charge carriers. It can also be described as a 4 k_F-instability and CDW, where the carriers correlate to stay apart (an equidistant arrangement of the charge carriers corresponds to a period $2\pi/4\ k_F$, forming what is named a *Mott insulator*). In contrast, $TMTSF_2$ PF_6 shows a metallic resistivity with an approximate T^2-dependence down to approximately 12 K, where a transition into a semiconducting 2 k_F-SDW state is found (for a review see Jerome, 2004).

Then in 1980 the finding of the first organic superconductor was reported. Jerome *et al.* (1980) discovered $TMTSF_2$ PF_6 to be superconducting at an elevated hydrostatic pressure of 900 MPa with a transition temperature of 1.0 K. Shortly after, the $TMTSF_2ClO_4$ compound, which due to a smaller anion has a slightly denser packing, was found to be superconducting at ambient pressure with a transition temperature of 1.2 K (Bechgaard *et al.*, 1981a, 1981b). Figure 26.12 shows measurements of the suppression of superconductivity by a small magnetic field perpendicular to the stacking axis.

The specific heat of the ClO_4 compound displays almost a textbook example of a superconducting transition (Garoche *et al.*, 1982). It may be analyzed to give a thermodynamic critical field of order 4 mT. However, this does not mean that the superconducting state in these materials is understood. Careful measurements have revealed that the critical magnetic fields, in particular, for fields

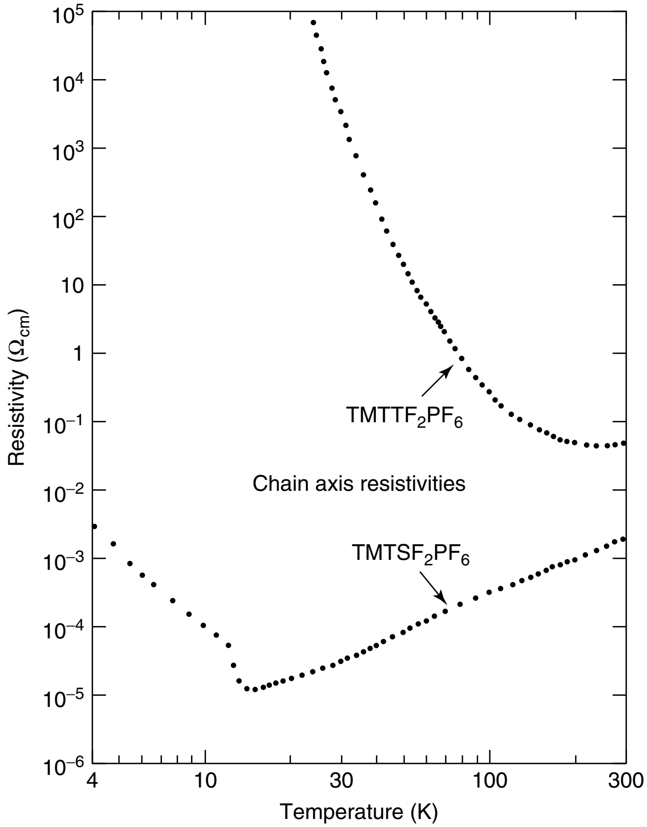

Fig. 26.11 Temperature-dependent resistivities of isostructural $TMTSF_2PF_6$ and $TMTTF_2PF_6$ at ambient pressure. Notice the logarithmic axes. The resistivity of $TMTSF_2PF_6$ shows an approximate T^2-dependence in the metallic state, and below about 12 K semiconducting behavior due to a frozen SDW state. The $TMTTF_2PF_6$-data bears evidence for the gradual Coulomb driven localization of the conduction electrons on cooling (Jacobsen, unpublished data).

along b' (where the interstack contacts are appreciable), are as large as 5–6 T (Lee *et al.*, 1997; Oh and Naughton, 2004). Thus the materials are extreme type II superconductors, and such values exceed the Pauli limit, where the pairing into singlet Cooper pairs should be destroyed by the field aligning the spins. It has therefore been taken as an argument for triplet superconductivity in these materials, where the spins are aligned in the BCS state. However, recent nuclear magnetic resonance (NMR) studies rather support the singlet-model (Shinagawa *et al.*, 2007). Another puzzling feature is a rather strong sensitivity of the superconductivity to nonmagnetic defects (see, for example, Jerome, 2004).

The nonconventional nature of superconductivity in TM_2X may very well be

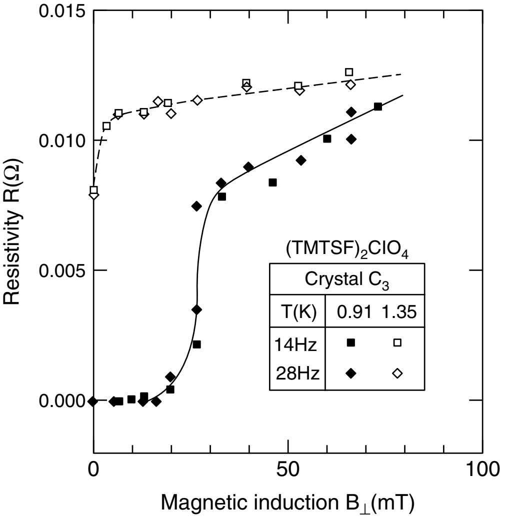

Fig. 26.12 Resistance of $TMTSF_2ClO_4$ as a function of a magnetic field perpendicular to the stacks for two temperatures, 0.91 K and 1.35 K, the latter being above T_C (Reproduced with permission from Bechgaard *et al.*, 1981b. Copyright American Physical Society.).

related to the occurrence of a whole range of interesting states in these materials. Jerome (1991) has suggested a generic pressure–temperature phase diagram for the family, which we reproduce in Figure 26.13. Ordered phases include superconductivity bordering to SDW formation, charge localization (described above), and the spin-Peierls state, where the spins in the charge localized state order with a 2 k_F-periodicity and quench the magnetic susceptibility. Also the whole picture is complicated by the occurrence of the so-called field-induced SDW states and quantum Hall effect states at fields in the 5–30 T range. In conclusion, these materials display very rich and challenging physics, which is yet to be understood.

26.3.4
Quasi-Two-Dimensional Organic Superconductors

An important modification of TMTTF results in the molecule bis(ethylenedithio) tetrathiafulvalene (BEDT–TTF or in brief ET, see Figure 26.3). It combines

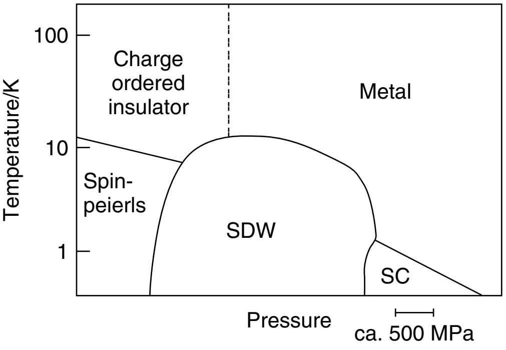

Fig. 26.13 Generic temperature–pressure phase diagram for the TMTT(S)F_2X conductors. SC, superconductivity and SDW, spin-density-wave state. $TMTTF_2PF_6$ is in the left part of the diagram, while $TMTSF_2ClO_4$ is in the right part (Adapted with permission from Jerome, 1991. Copyright American Association for the Advancement of Science.).

with various inorganic and organic anions to produce a variety of crystal structures, where those relevant for superconductivity have layered structures. The ET molecules order in stacks with appreciable contact between stacks. This is due to the high number of sulfur atoms along the periphery of the molecule (eight instead of four in TMTTF). This makes the electronic band structure effectively two-dimensional. This may be exemplified by the polarized reflectance spectra of β-ET_2I_3 shown in Figure 26.14. Analysis indicates a bandwidth anisotropy within the layer of $W_{\mathrm{stack}}/W_{\mathrm{perpendicular}} \approx 2$ (Jacobsen, Williams, and Wang, 1985; Jacobsen *et al.*, 1987), and suggests a closed Fermi surface in the layer plane, which cannot be nested by a single wave vector. Thus, such materials should not be susceptible to the typical quasi-one-dimensional instabilities (for an extended review see, for example, Singleton and Mielke, 2002).

The ET salts may be classified according to type of crystal structure, where those pertinent for superconductivity are named as α, β, β'', λ, and κ phases (Mori, 1998; Mori, Mori, and Tanaka, 1999; Mori, 1999). Only the β-phase has the simple two-dimensional Fermi surface described above. The others seem to possess both two-dimensional pockets and quasi-one-dimensional sections, thus again opening up for the possibility of for example, CDW formation (see, for example, Singleton, 2000, and Ishiguo, Yamaji, and Sauto, 1998).

Superconductivity at ambient pressure was first discovered in ET_2I_3 with $T_C = 1.4$–2.5 K (Yagubskii *et al.*, 1984). In the β-ET_2X family, it is observed that $T_C = 1.5$ K for $X = I_3^-$, $T_C = 2.7$ K for $X = IBr_2{}^-$, and $T_C = 4.9$ K for $X = AuI_2{}^-$ (see, for example, Ishiguo, Yamaji,

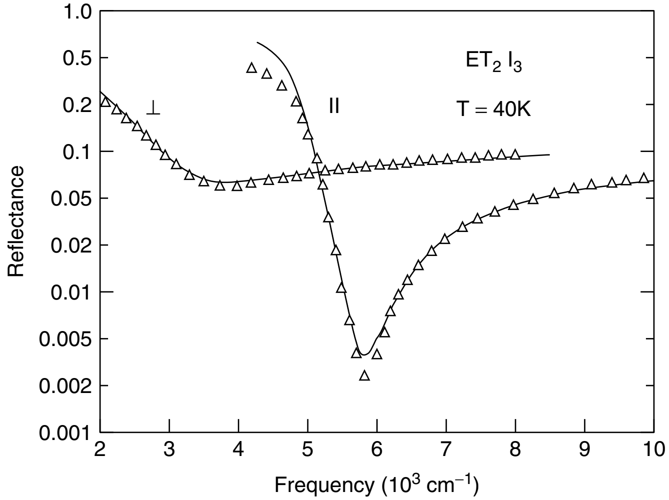

Fig. 26.14 Polarized infrared reflectance spectra of ET_2I_3 at $T = 40$ K. A distinct plasma edge is observed perpendicular to as well as along the molecular stacks. The solid lines are Drude model fits (Reproduced with permission from Jacobsen, Williams, and Wang, 1985. Copyright Elsevier.).

and Sauto, 1998). Hydrostatic pressure in general reduces T_C. Consistent with this, the highest transition temperatures in the ET family is found using rather large anions. Here, we just mention κ-ET$_2$Cu(NCS)$_2$ with $T_C = 10.4$ K (Urayama *et al.*, 1988; Carlson *et al.*, 1988) and κ-ET$_2$[N(CN)$_2$]Br with $T_C =$ 11.6 K (Kini *et al.*, 1990). Slightly higher values have been found using hydrostatic pressure. Altogether several tens of ET compounds have superconducting ground states and a variety of modified donor molecules have also been found to be superconductors (see, for example, Ishiguo, Yamaji, and Sauto, 1998).

The superconducting state has been investigated in detail, in particular, in the κ-family. Just as with the TMTSF-compounds, there is no consensus as to the detailed mechanism or the symmetry of the pairing. In the following, we just describe some typical characteristics.

Determination of the upper critical field in κ-ET$_2$Cu(NCS)$_2$ is hampered by motion of vortices in the mixed state but is in the range 3–8 T for fields perpendicular to the layers (for a review see Singleton and Mielke, 2002). For fields exactly in plane, the upper critical field is 35 T. Triangular vortex lattices have been seen.

Considerable discussion has centered around the symmetry of the order parameter (gap function, Δ). The standard *s*-wave order parameter is isotropic, and the gap is complete. Support for a completely gapped superconducting state comes from specific heat measurements on κ-ET$_2$[N(CN)$_2$]Br. They show a discontinuity and fit a theory for strong coupling superconductivity (Elsinger *et al.*, 2000). In contrast, *d*-wave symmetry is characterized by an anisotropic gap including nodes, where the gap goes to zero. Such *d*-wave symmetry is supported by NMR measurements. A detailed study on κ-ET$_2$[N(CN)$_2$]Cl is described by Lefebvre *et al.* (2000) as a function of temperature and hydrostatic pressure and apparently excludes *s*-wave symmetry. It also reveals an interesting phase diagram. Superconductivity exists above 30 MPa and below 13 K, while the ambient pressure ground state is antiferromagnetic. There is a first-order borderline between a magnetic phase and superconductivity with evidence for coexisting phases. While measurements of the penetration depth, which can also give indications about the gap, are rather inconclusive, tunneling experiments on κ-ET$_2$Cu(NCS)$_2$ have shown a convincing anisotropy (Arai *et al.*, 2001).

In conclusion, the superconducting members of the ET family are highly anisotropic type II materials. There is evidence for antiferromagnetic fluctuations playing a role in the pairing, and also for non-*s*-wave character of the order parameter. So while the situation is far from settled, these materials must also be placed in the category of unconventional superconductors.

26.4
Conclusion and Outlook

Almost 30 years after the discovery of the first organic superconductor, the microscopic mechanism is still not understood, and the area continues to attract many research groups throughout the world. The materials are similar to the ceramic high-temperature superconductors, which are also not understood, in the sense that superconductivity occurs next to magnetism. The richness of different ground states as well as the behavior in high magnetic fields also makes the materials group a favorite playground for physicists. The highest transition temperature in the charge transfer compounds is about 12 K, but without a detailed understanding of the phenomenon it is impossible to predict how high it can be.

To date no polymers has shown bulk superconductivity. Oxidized polypropylene films have been reported to show behavior that is interpreted as superconductivity at room temperature. However, the high conductivity is confined to narrow channels perpendicular to the films, and whether it has relation to BCS superconductivity is not so clear (see, for example, Grigorev, 1998).

Concerning biomolecules, speculations about superconductivity continue and the observation of signs of proximity induced superconductivity in micrometer DNA strings at ultralow temperatures suggests that it is a possibility (Kasumov *et al.*, 2001). The proximity effect exploits the fact that a superconducting order always to some extent leaks into a normal conductor, which is in close contact with the superconductor. In the present case, rhenium/carbon electrodes with $T_C = 1$ K were used. There is little doubt that efforts to map out the electronic quantum transport properties of macromolecules will continue.

Glossary

C-60: hollow spheres of 60 carbon atoms (“bucky balls”).

Carbon nanotubes: tubes formed as rolls of graphene. They can be single-walled as well as multiwalled and exist in both conducting and insulating forms.

Charge transfer complex: consists typically of pairs of donor and acceptor molecules, where a finite electron transfer takes place. It is one approach to create unpaired charge carriers in molecular crystals.

Conduction electrons: electrons that are not bound to the lattice sites but are free to move through the material.

Continuous phase transition: a phase transition without latent heat. Examples are ferromagnetism and superconductivity.

Density-of-states: number of allowed one-electron states per energy. It has a high value at the Fermi level in a metal and is zero within an energy gap.

Dissipation: loss of organized energy into heat. Example includes scattering of electrons caused by lattice vibrations leads to a loss of electric energy.

Energy band: description of one-electron quantum states in crystals. The energy band is the one-electron energy as function of wave number (or more generally wave vector). The states are so closely spaced that it is appropriate to talk about bands. There are a number of different bands making up the band structure.

Energy gap: range of energies, where no quantum states exist. It is typical for waves in periodic lattices. A material where the chemical potential is located in an energy gap is insulating or semiconducting.

Faraday’s law: the law of induction. A time-dependent magnetic flux generates an electric field circulating around it. The magnitude of the corresponding electromotive force equals the time derivative of the flux.

Fermi level: separates the occupied electronic states in a metal from those unoccupied. More precisely it is the low-temperature limit of the chemical potential.

Graphene: single carbon layer in graphite.

Mean-field theory: simple way of including ordering phenomena in statistical physics. The order is treated as an average field. Thus all fluctuations are neglected. It works rather well for superconductivity because of the finite coherence length.

Momentum: mass times velocity. In the presence of a magnetic field, it is necessary to distinguish between this kinematical momentum and the canonical momentum, which depends on the field and is present in the equations of motion.

Mott insulator: material with a finite one-electron density-of-states at the chemical potential, which, nevertheless, is semiconducting or insulating owing to important short-range Coulomb repulsion between the electrons.

Orbital symmetry: symmetry of wave function. For superconductivity, it often refers to the symmetry of the Cooper pairs (using the language of atomic physics: s-wave, p-wave, d-wave, etc.).

Organic molecule: molecule based on carbon chemistry. Pure forms of carbon are excluded.

Pauli principle: the principle states that a full one-electron quantum state (including spin) can house zero or one electron.

Persistent current: dissipation free current in a closed loop. In superconductors they have been observed running for years with negligible decay.

Phonon: quantized crystal lattice vibration.

π -bond: chemical bond formed from overlap of p-type orbitals perpendicular to the connection line between the nuclei. Example includes benzene.

Plasma edge: the sharp fall of reflectance of a metal at a frequency, where wave propagation sets in. It coincides with the plasmon frequency, that is, the frequency of electronic density oscillations. The plasma edge is often well described by a classical Drude model.

Polyene: unsaturated organic compound that contains one or more sequences of alternating double and single carbon–carbon bonds.

Proximity effect: leaking of superconductivity into a nearby normal metal in good contact with the superconductor. It is caused by second-order quantum processes.

Resistance: the ratio of the voltage drop across a conductor and the resulting electric current. According to the widely valid Ohm's law, the two quantities are proportional. A superconductor has zero resistance.

Semiconductor: material with a relatively small energy gap at the chemical potential. Hence, charge carriers can be generated easily thermally or by chemical or electrical control.

Superconductor: material that loses its resistivity completely below a critical temperature, T_C. In the superconducting state small magnetic fields are expelled.

Tight binding approximation: method of calculating the electronic band structure. It only includes the nearest neighbor interactions and works well for molecular crystals.

Wave number: one of the quantum number for free electrons and electrons moving in a periodic potential. The magnitude equals 2π divided by the wave length. In a crystal, it is generalized to a wave vector.

References

Abrikosov, A.A. (1957) *Zh. Eksperim. i Teor. Fiz.*, **32**, 1442.

Arai, T., Ichimura, K., Nomura, K., Takasaki, S., Yamada, J., Nakatsuji, S., and Anzai, H. (2001) *Phys. Rev. B*, **63**, 104518.

Batail, P. (ed.) (2004) *Chem. Rev.*, **104**(11) (special issue on Molecular Conductors).

Bardeen, J., Cooper, L.N., and Schrieffer, J.R. (1957) *Phys. Rev.*, **108**, 1175.

Bechgaard, K., Jacobsen, C.S., Mortensen, K., Pedersen, H.J., and Thorup, N. (1980) *Solid State Commun.*, **33**, 1119.

Bechgaard, K., Carneiro, K., Rasmussen, F.B., Olsen, M., Rindorf, G., Jacobsen, C.S., Pedersen, H.J., and Scott, J.C. (1981a) *J. Am. Chem. Soc.*, **103**, 2440.

Bechgaard, K., Carneiro, K., Olsen, M., Rasmussen, F.B., and Jacobsen, C.S. (1981b) *Phys. Rev. Lett.*, **46**, 852.

Butch, N.P., de Andrade, M.C., and Maple, M.B. (2008) *Am. J. Phys.*, **76**, 106.

Carlson, K.D., Geiser, U., Kini, A.M., Wang, H.H., Montgomery, L.K., Kwok, W.K., Beno, M.A., Williams, J.M., Cariss, C.S., Crabtree, G.W., Whangbo, M., and Evain, M. (1988) *Inorg. Chem.*, **27**, 967.

Coleman, L.B., Cohen, M.J., Sandman, D.J., Yamagishi, F.G., Garito, A.F., and Heeger, A.J. (1973) *Solid State Commun.*, **12**, 1125.

Cooper, L.N. (1956) *Phys. Rev.*, **104**, 1189.

Delhaes, P., Coulon, C., Amiell, J., Flandrois, S., Torreilles, E., Fabre, J.M., and Giral, L. (1979) *Mol. Cryst. Liq. Cryst.*, **50**, 43.

Elsinger, H., Wosnitza, J., Wanka, S., Hagel, J., Schweitzer, D., and Strunz, W. (2000) *Phys. Rev. Lett.*, **84**, 6098.

Ferraris, J., Cowan, D.O., Walatka Jr., V., and Perlstein, J.H. (1973) *J. Am. Chem. Soc.*, **95**, 948.

Fröhlich, H. (1950) *Phys. Rev.*, **79**, 845.

Fröhlich, H. (1954) *Proc. Roy. Soc. (London)*, **A223**, 296.

Garoche, P., Brusetti, R., and Bechgaard, K. (1982) *Phys. Rev. Lett.*, **49**, 1346.

Ginzburg, V.L. and Landau, L.D. (1950) *Zh. Eksperim. i Teor. Fiz.*, **20**, 1064.

Gorter, C.J. and Casimir, H.B.G. (1934) *Phys. Z.*, **35**, 963.

Grigorev, L.N. (1998) *Phil. Mag. B*, **78**, 353.

Gutmann, F. and Lyons, L.E. (1967) *Organic semiconductors*, Wiley, New York.

Ishiguo, T., Yamaji, K., and Sauto, G. (1998) *Organic Superconductors*, Springer-Verlag, Berlin.

Jacobsen, C.S. (1988) *Semiconductors and Semimetals* Vol. 27, Academic Press, New York, p. 293.

Jacobsen, C.S., Tanner, D.B., and Bechgaard, K. (1981) *Phys. Rev. Lett.*, **46**, 1142.

Jacobsen, C.S., Williams, J.M., and Wang, H.H. (1985) *Solid State Commun.*, **54**, 937.

Jacobsen, C.S., Tanner, D.B., Williams, J.M., Geiser, U., and Wang, H.H. (1987) *Phys. Rev. B*, **35**, 9605.

Jerome, D. (1991) *Science*, **252**, 1509.

Jerome, D. (2004) *Chem. Rev.*, **104**, 5565.

Jerome, D., Mazaud, A., Ribault, M., and Bechgaard, K. (1980) *J. Phys. Lett.*, **41**, L95.

Jerome, D. and Schulz, H.J. (1982) *Adv. Phys.*, **31**, 299.

Kasumov, A.Yu., Kociak, M., Guéron, S., Reulet, B., Volkov, V.T., Klinov, D.V., and Bouchiat, H. (2001) *Science*, **291**, 280.

Kini, A.M., Geiser, U., Wang, H.H., Carlson, K.D., Williams, J.M., Kwok, W.K., Vandervoort, K.G., Thompson, J.E., Stupka, D.L., Jung, D., and Whangbo, M. (1990) *Inorg. Chem.*, **29**, 2555.

Lee, I.J., Naughton, M.J., Danner, G.M., and Chaikin, P.M. (1997) *Phys. Rev. Lett.*, **78**, 3555.

Lefebvre, S., Wzietek, P., Brown, S., Bourbonnais, C., Jerome, D., Meziere, C., Fourmigue, M., and Batail, P. (2000) *Phys. Rev. Lett.*, **85**, 5420.

Little, W.A. (1964) *Phys. Rev.*, **134**, A1416.

London, F. (1950) *Superfluids*, Vol. 1, John Wiley & Sons, Inc., New York.

London, F. and London, H. (1935) *Proc. Roy. Soc. (London)*, **A149**, 71.

Meissner, W. and Ochsenfeld, R. (1933) *Naturwissenschaften*, **21**, 787.

Mori, T. (1998) *Bull. Chem. Soc. Jpn.*, **71**, 2509.

Mori, T., Mori, H., and Tanaka, S. (1999) *Bull. Chem. Soc. Jpn.*, **72**, 179.
Mori, T. (1999) *Bull. Chem. Soc. Jpn.*, **72**, 2011.
Oh, J.I. and Naughton, M.J. (2004) *Phys. Rev. Lett.*, **92**, 067001.
Peierls, R.E. (1955) *Quantum Theory of Solids*, Oxford University Press, Oxford, p. 108.
Pippard, A.B. (1953) *Proc. Roy. Soc. (London)*, **A216**, 547.
Shinagawa, J., Kurosaki, Y., Zhang, F., Parker, C., Brown, S.E., Jerome, D., Christensen, J.B., and Bechgaard, K. (2007) *Phys. Rev. Lett.*, **98**, 147002.
Singleton, J. (2000) *Rep. Prog. Phys.*, **63**, 1111.
Singleton, J. and Mielke, C. (2002) *Contemp. Phys.*, **43**, 63.
Soling H., Rindorf, G., and Thorup, N. (1981). Private communication.
Tinkham, M. (1996) *Introduction to Superconductivity*, 2nd edn, McGraw-Hill, New York.
Urayama, H., Yaochi, H., Saito, G., Nozaea, K., Sugano, T., Kinoshita, M., Sato, S., Oshima, K., Kawamoto, A., and Tanaka, J. (1988) *Chem. Lett.*, **1988**, 55.
Yagubskii, E.B., Shchegolev, I.F., Laukhin, V.N., Kononovich, P.A., Karatsovnik, M.V., Zvarykina, A.V., and Buravov, L.I. (1984) *Pis'ma Zh. Eksp. Teor. Fiz.*, **39**, 12. [transl.: *JETP Lett.* **39**,, 12].

Further Reading

Farges, J.-P. (ed) (1994) *Organic conductors: Fundamentals and applications*, Dekker, New York.
Williams, J.M., Ferrarao, J.R., Thorn, J.R., Carlson, K.D., Geiser, U., Wang, H.H., Kini, A.M., and Whangbo, M. (1992) *Organic Superconductors*, Prentice Hall, New Jersey.
Wosnitza, J. (2001) Recent aspects of organic metals and superconductors, *Curr. Opin. Solid State Mater. Sci.*, **5**, 131–141.

28

Raman Spectroscopy in Medicine

Christoph Krafft, Petra Rösch, and Jürgen Popp

28.1 Introduction 987
28.2 Raman Instrumentation for Medical Applications 988
28.2.1 Raman Spectra of Biomolecules 988
28.2.2 Near-Infrared Advantage 989
28.2.3 Raman Microscopy 990
28.2.4 Raman Imaging 991
28.2.5 Fiber-Optic Probes 991
28.2.6 Resonance Raman and Surface-Enhanced Raman Spectroscopy 992
28.3 Raman Spectroscopy in Pathology 992
28.3.1 Oncology 993
28.3.1.1 Neuro-Oncology 993
28.3.1.2 Tumors of Epithelial Tissue 998
28.3.2 Mineralized Tissue 1003
28.3.2.1 Bone 1003
28.3.2.2 Teeth 1004
28.3.2.3 Calculi 1004
28.3.3 Skin 1005
28.3.3.1 Corneocyte Biology and Pharmacology 1006
28.3.3.2 Skin Cancer 1006
28.3.3.3 Raman Instrumentation for Skin Studies 1007
28.3.4 Atherosclerosis 1008
28.3.5 Ophthalmology 1008
28.3.5.1 Macular Degeneration of the Retina 1009
28.3.5.2 Chemical Composition of the Vitreous and the Lens 1009
28.4 Raman Spectroscopy in Clinical Chemistry 1010
28.4.1 Determination of Glucose 1010
28.4.2 Serum Analysis 1011
28.4.3 Malaria Investigations 1011
28.4.4 Analysis of Hair and Nails 1012
28.5 Conclusions and Outlook 1012
References 1013

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

28.1
Introduction

After the inelastic scattering of light by molecules was predicted theoretically, the effect was observed experimentally in 1928 by the Indian physicist C.V. Raman. In honor of his discovery, the effect was named after him. In the past 80 years, the method was applied in various disciplines such as condensed matter physics and organic and inorganic chemistry. One field which gained significant recognition in the past 10 years was the application of Raman spectroscopy in medicine to assess cells and tissues. The principle of the Raman effect is that an excited photon interacts with a molecule, and an emitted photon possesses a different energy. The energy difference excites a molecular vibration. In this way, a set of vibrations, called Raman spectrum, can be obtained, which provides a highly specific fingerprint of the sample. In contrast to other photon emitting techniques such as fluorescence or phosphorescence, Raman spectroscopy offers the following advantages: (i) No labels are required, which suffer from limited stability and accessibility because vibrations are an inherent property of all molecules. (ii) The number of vibrations can be calculated from the number of atoms N by the formula $3N - 6$. Therefore, the number of Raman bands is much larger than the number of fluorescence bands, which depend on electronic transitions. (iii) As Raman bands are narrower, many bands can be resolved giving detailed information of the chemical composition and structure of samples. Molecular vibrations can also be excited by absorption of radiation in the mid infrared (IR) wavelength range between 2.5 and 25 μm. However, the prospects of Raman spectroscopy in medicine have been recognized to be superior compared with IR spectroscopy mainly due to easier sample preparation and the tolerance of spectral contributions from the most abundant biomolecule water. The principle of Raman spectroscopy to study cells and tissues is that diseases and other pathological anomalies cause chemical and structural changes at the molecular level. These changes also alter the Raman spectra that can be used as sensitive phenotypic marker. First, the current contribution describes the technical progress that was required before Raman spectroscopy could be applied to cells and tissues that are composed of thousands of biomolecules in a complex way. Then, medical applications are summarized, which complement pathology

Handbook of Molecular Biophysics. Methods and Applications. Edited by Henrik G. Bohr.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40702-6

to assess tissue *ex vivo* and *in vivo*, and clinical chemistry to determine the composition of body fluids.

28.2
Raman Instrumentation for Medical Applications

The basic setup of a dispersive Raman spectrometer consists of a laser as an intense and monochromatic light source, a device that separates the elastically (Rayleigh) scattered light of the sample from the Raman (inelastically) scattered light, a spectrograph and a detector. Multichannel detectors enable to register the whole Raman spectrum simultaneously within a fraction of seconds. Raman spectrometers utilizing the Fourier transform (FT) principle consist of an interferometer instead of a spectrograph. The interferometer converts the Raman scattered radiation into an interferogram, which is registered by a fast detector. After Fourier transformation of the interferogram, the whole Raman spectrum can be obtained.

28.2.1
Raman Spectra of Biomolecules

Starting point for medical applications of Raman spectroscopy is to study the Raman spectra of the main constituents of cells and tissues such as proteins, nucleic acids, and lipids. Representative spectra of purified samples are shown in Figure 28.1. In general, the spectra depend on the monomeric units

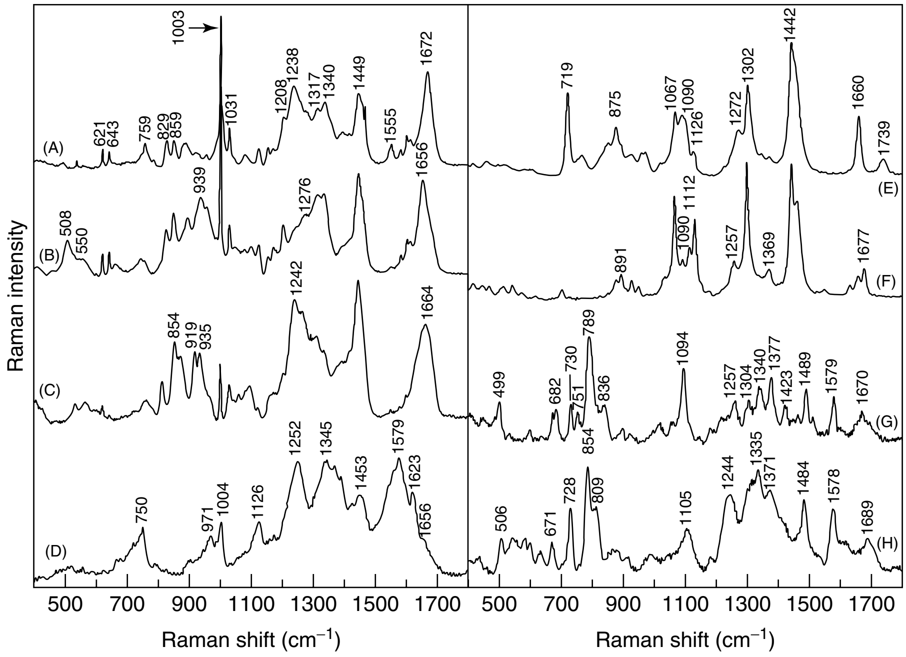

Fig. 28.1 Raman spectra from 400 to 1800 cm^{-1} of proteins (A, B, C, D), lipids (E, F), and nucleic acids (G, H): concananvalin A (A), serum albumin (B), collagen type 1 (C), hemoglobin (D), phosphatidylcholin (E), galactocerobroside (F), DNA (G), and RNA (H).

(amino acids in case of proteins and nucleotides in case of nucleic acids) and the geometry of the backbone. Bands in the Raman spectrum of the all β-protein concanavalin A (spectrum A) are assigned to phenylalanine (621, 1003, 1031, and 1208 cm^{-1}), tyrosine (643, 829, and 853 cm^{-1}), tryptophan (759, 1340, and 1555 cm^{-1}), $\mathrm{CH}_{2/3}$ of aliphatic amino acids (1317 and 1449 cm^{-1}), and amide I (1672 cm^{-1}) and amide III (1238 cm^{-1}) of the peptide backbone. Additional bands at 508 and 550 cm^{-1} in the Raman spectrum of the all α-protein serum albumin (spectrum B) are indicative of disulfide bonds that stabilize the protein structure. Amide I and III bands are shifted to 1656 and 1276 cm^{-1}, respectively. Together with the band at 939 cm^{-1}, they are typical for α-helices. Increased bands at 854, 919, 935, 1242, and 1664 cm^{-1} are observed in the Raman spectrum of collagen (spectrum C) due to the triple helix structure and the unusual amino acid composition, which contains high content of prolin and hydroxyprolin. As bands of the cofactor heme in the Raman spectrum of hemoglobin are enhanced at 750, 971, 1004, 1126, 1252, 1453, 1579, and 1623 cm^{-1}, they are more intense than the amide I band of the protein at 1656 cm^{-1} (spectrum D). The enhancement effect is described in Section 28.2.6. As the heme group is important for the function of the protein and Raman spectroscopy can sensitively probe the vibrations of the heme group, this method has frequently been applied to this group of proteins (Wood and McNaughton, 2006).

Lipids are composed of a backbone group, a hydrophobic moiety that is often a long fatty acid chain, and a hydrophilic moiety that is derived from a sugar or alcohol compound. The Raman spectrum of phosphatidylcholine (spectrum E) shows bands due to fatty acids (1067, 1126, 1302, and 1442 cm^{-1}), the choline group (719 and 875 cm^{-1}), the phosphate group (1090 cm^{-1}), and the ester group (1739 cm^{-1}). Bands at 1660 and 1272 cm^{-1} indicate unsaturated fatty acids chains. The Raman spectrum of galactocerobroside shows similar fatty acids bands (spectrum F). However, the ceramide backbone instead of the glyceride backbone is characterized by amide I and III bands near 1677 and 1257 cm^{-1}, respectively. The galactose ring instead of the phosphorylated choline group is characterized by bands at 891, 1090, 1112, and 1369 cm^{-1}. Raman spectra of other lipids have previously been reported (Krafft *et al.*, 2005a).

Bands in the Raman spectrum of DNA (spectrum G) are assigned to the nucleotides bases guanine, adenine, thymine, and cytosine (499, 682, 730, 751, 789, 1257, 1304, 1340, 1377, 1489, 1579, and 1670 cm^{-1}), the phosphodiester groups (836 cm^{-1}), and the phosphodioxy groups (1094 cm^{-1}) in B-form geometry. Changes in the Raman spectrum of RNA (spectrum H) are consistent with a different sugar pucker geometry, the nucleotide base thymine is replaced by uridine, and the phosphate backbone adopt a A-form geometry, which shift the bands to 809 and 1105 cm^{-1}. Details of the band assignment can be found elsewhere (Prescott, 1984).

28.2.2
Near-Infrared Advantage

Raman spectroscopy of natural, unprocessed biological material is more complicated because spectral contributions of all molecules overlap and intrinsic fluorescence often masks the weaker Raman signals when using visible lasers for excitation. Even trace amounts of fluorescent material can cause an intense signal due to several orders of magnitude larger efficiency. As most tissues and body fluids show minimum absorption in the wavelength interval from 700 to 900 nm, the excited autofluorescence is at minimum and the penetration of the exciting and scattered radiation is at maximum. Fluorescence in Raman spectra can also be avoided by laser excitation in the deep ultraviolet (UV) because a fluorescence-free window exists with excitation below 270 nm. However, deep UV excitation also harbors the risk of inducing photodegradation damage, which should be avoided in *in vivo* situations.

In the late 1980s, near-infrared (NIR) FT-Raman spectrometers with Nd-YAG laser excitation at 1064 nm have been developed to overcome the fluorescence problem. Because the scattering intensities decrease with the forth power of the frequency of the exciting radiation, the Raman signals were $2^4 = 16$ times less intense compared with 532 nm excitation. Together with the low sensitivity of the FT-Raman spectrometers and the low Raman cross-sections of most biomolecules, collection times were impractically long. Since the mid 1990s, high-throughput optical elements for Rayleigh scattering rejection (notch or edge filters) and for spectral dispersion (holographic transmissive gratings) have been introduced in Raman spectrometers. In combination with NIR lasers emitting at or near 785 nm and sensitivity-optimized, silicon-based charge-coupled device (CCD) detectors, these Raman systems enabled collecting Raman spectra of unprocessed biological material in the spectral range from 200 to 3600 cm^{-1} within seconds.

28.2.3
Raman Microscopy

The coupling of a Raman spectrometer with a microscope offers two main advantages. First, lateral resolutions can be achieved up to the Abbe's limit of diffraction below 1 μm. This property is important to resolve subcellular details in single cells. Even lower resolutions have been demonstrated using near field microscopic approaches. As a consequence of their complexity, these techniques have not been applied in medicine yet. High axial resolutions are obtained by confocal microscopes. Secondly, maximum sensitivity can be achieved because the photon flux (photon density per area) at the focused laser beam onto the sample is at maximum and the collection efficiency of scattered photons from the sample is at maximum. The lateral resolution, the diameter of the focus, and the collection efficiency depend on the numerical aperture (NA) of the microscope objective, which is the product of the sine of the aperture angle and the diffraction index. High NA objectives have small working distances below 1 mm in air. The working distance of water immersion objectives can be larger due to the larger diffraction index of water. Oil immersion objectives in combination with oil should be avoided because the oil can contaminate the sample and the spectral contributions of oil overlap with the Raman spectrum of the sample.

28.2.4
Raman Imaging

Raman imaging combines the spectral and the lateral information. Most Raman images are collected in the point-mapping mode. Here, the laser is focused onto the sample, the scattered light is registered, and subsequently the focus is moved by a scanning mirror or the sample is moved by a motorized stage to the next position. The total exposure time depends on the number of spectra and the exposure time per spectrum. Accumulating more than thousand spectra with typical times of a few seconds per spectrum usually takes hours. In the case of laser line illumination of the sample, the spatial data can be registered on the detector on a line parallel to the entrance slit of the spectrometer and the spectral information is dispersed perpendicularly. The second spatial dimension of an image is recorded by scanning in the direction perpendicular to that line. This so-called line-mapping registration mode is faster because only one dimension instead of two dimensions in the point-mapping mode has to be scanned. The parallel registration approaches called direct or wide-field Raman imaging employs intense, global sample illumination. The inelastically scattered light from the sample is projected onto a two-dimensional CCD detector. Most wide-field Raman imaging spectrometers use filters to select the wavelength such as dielectric, acousto-optic tunable, and liquid–crystal tunable filters. Although Raman spectrometers using this principle are commercially available (Falcon by Chemicon, USA), medical applications have not been reported yet.

28.2.5
Fiber-Optic Probes

In vivo diagnostic tools are much needed in many fields of medicine such as the guidance of surgical interventions to delineate lesions margins or to replace random biopsies of suspicious tissues by targeted biopsies that, in turn, would reduce unnecessary tissue excisions, pathology costs, and biopsy-associated risks. Several probe geometries have been designed for use with Raman spectrometers. A miniaturized probe that consisted of one central excitation fiber, six surrounding collection fibers, internal "in the tip" filters, and beveled fiber ends for optimized light collecting efficiency was designed (Shim *et al.*, 1999) and applied in various biomedical studies (Bakker Schut *et al.*, 2000; Boere *et al.*, 2003; Buschmann *et al.*, 2000; Crow *et al.*, 2005; Molckovsky *et al.*, 2003; Shim *et al.*, 2000). A micro-optical fiber probe for use in an intravascular Raman endoscope has been developed, which consists of one excitation fiber, eight collection fibers, integrated filters, and a lens (Komachi *et al.*, 2005a. Its total diameter is just 600 µm. Its use has been reported in two studies (Hattori *et al.*, 2007; Komachi, Sato and Tashiro, 2006). Other Raman probes that are, however, not commercially available use 15 fibers and a ball lens (Motz *et al.*, 2004), six fibers (Barbosa *et al.*, 2002) or one fiber for signal collection (Huang *et al.*, 2001; Mahadevan-Jansen *et al.*, 1998; Utzinger *et al.*, 2001). A probe with integrated filters, one excitation and one collection fiber each from Inphotonics Inc. (USA) was recently applied to record Raman images of murine brains (Krafft *et al.*, 2007). However, probe dimensions

(13 mm diameter, 100 mm length) are likely too large for endoscopic applications. A small, unfiltered probe was suggested as an alternative, because silica with a low content of hydroxyl groups as core material shows only low spectral contributions in the high wavenumber region from 2400 to 3800 cm^{-1} (Santos *et al.*, 2005). Other new developments include the PhAT probe from Kaiser Optical System (USA) with global illumination and an array of 50 collection fibers (Schulmerich *et al.*, 2006), hollow-core photonic crystal fiber-optic probes (Konorov *et al.*, 2006), a single flexible hollow waveguide coated with a silver film (Komachi *et al.*, 2005b), and concentric rings of optical fibers used in spatially offset Raman spectroscopy (SORS) (Matousek *et al.*, 2006). In SORS, Raman spectral contributions from subsurface layers can be obtained depending on the distance between the excitation and collection fibers.

28.2.6
Resonance Raman and Surface-Enhanced Raman Spectroscopy

Raman signals can be enhanced by various nonlinear techniques using ultrashort (femtoseconds), but high-intensity (megawatt) laser pulses. As they have not been applied in medicine so far, they are not within the scope of this chapter. However, resonance Raman spectroscopy (RRS) and surface-enhanced Raman spectroscopy (SERS) are variations of conventional Raman spectroscopy that can be used to enhance its specificity and sensitivity. In SERS, surface plasmons of rough metal surfaces (often gold or silver nanoparticles) are excited by the laser resulting in an increase in the electric fields surrounding the metal. Given that Raman intensities are proportional to the electric field, the measured signal of molecules that are adsorbed on these nanoparticles can be strongly increased (by up to 10^{15}) (Nie and Emory, 1997). Most plasmon frequencies are found in the visible range. In resonance Raman (RR), the excitation wavelength is matched to an electronic transition of the molecule, so that vibrational modes associated with the excited state are enhanced (by as much as a factor of 10^6). Electronic transitions of nucleic acids and proteins without prosthetic groups are found below 280 nm. The enhancement is restricted to vibrational modes of the chromophore, which reduces the complexity of large biomolecules. Both effects can be combined in surface-enhanced resonance Raman spectroscopy (SERRS) giving even higher enhancement factors.

28.3
Raman Spectroscopy in Pathology

Disease diagnosis generally involves gross and microscopic visual examination of tissues with special stains employed to visualize specific proteins and other substances. Rudolf Virchow (1821–1902) is generally recognized to be the father of microscopic pathology. Although the microscope has been invented earlier, Virchow was one of the first prominent physicians to emphasize the study of manifestations of disease, which were visible only at the cellular level. Virchow's student Julius Cohnheim (1839–1884) combined histology techniques with experimental manipulations making him one of the first experimental pathologists. Cohnheim also pioneered the use of the frozen section.

A version of this technique is still widely employed by modern pathologists to render diagnosis and provide other clinical information. New research techniques, such as electron microscopy, immunohistochemistry, and molecular biology, have expanded the means by which biomedical scientists can study disease. Raman spectroscopy has been combined with microscopy to complement microscopic pathology. As combination with fiber-optic probes enables applying the method under *in vivo* conditions, Raman spectroscopy is one of the optical methods currently under investigation for improvement of endoscopic applications, such as elastic light scattering, optical coherence tomography (OCT), and fluorescence spectroscopy.

28.3.1
Oncology

This section summarizes the prospects of Raman spectroscopy in oncology. As carcinoma, which is the collective term for tumors of epithelial tissue, constitutes the largest group of malignant tumors, Raman spectroscopy has frequently been applied for their detection. Before carcinomas of the gastrointestinal tract, breast, lung, urinary tract, and cervix uteri are discussed, the concepts of Raman spectroscopy in oncology is demonstrated in the context of neuro-oncology, which was previously described in detail (Krafft and Salzer, 2008).

28.3.1.1 Neuro-Oncology

Brain tumors are generally divided into primary and secondary brain tumors. According to a recommendation of the World Health Organization (WHO) (Kleihues and Cavenee, 2000), primary tumors that begin in brain tissue are classified by the cell types from which they originate, which include glial cells, neurons, the meninges, and cells of the pituitary gland. Astrocytic gliomas originate from astrocytes, a special class of glial cells. They account for approximately 50% of all primary brain tumors in adults and are classified by increasing malignancy into grade II (diffuse astrocytoma), grade III (anaplastic astrocytoma), and grade IV (glioblastoma multiforme, GBM). GBM either progresses from diffuse astrocytomas (secondary GBM) or arises *de novo* (primary GBM). Meningeomas are the second most frequent intracranial tumors after gliomas and account for approximately 20% of all primary brain tumors. They originate from the brain encapsulating meninges and form several subtypes, most of which are histologically benign (grade I): they exhibit slow growth pattern, and usually do not infiltrate brain tissue or cause metastases. Metastasis is the spread of cancer. Cancer that begins in other parts of the body may spread to the brain and cause secondary brain tumors, also called brain metastases. Brain metastases occur in about 15–40% of all cancer patients. This proportion has increased in recent years partly due to the increased sensitivity of tomographic techniques to detect these tumors, and also to the longer survival of these patients. The diagnosis of brain metastasis in a patient with a known systemic cancer is a simple matter. However, in almost 50% of patients undergoing surgery for brain metastases, the secondary brain tumor represents the first indication of cancer and the primary tumor is not known at the time of neurosurgery. Therefore, besides tumor resection, the scope of neurosurgical procedures is to collect tissue specimens for diagnosis. Nevertheless, in up to 15% of the patients with brain metastases the primary tumor will not be identified despite thorough investigation by standard screening techniques. Thus, brain metastases with unknown primary tumors are estimated to be almost equal in number to primary brain tumors (Polyzoidis, Miliaras and Pavlidis, 2005).

Histopathologic diagnosis of hematoxylin and eosin (H&E) stained tissue sections is the so-called golden standard to determine grade and type of the tumor, which decide its effective management and provide an indication of the prognosis. The location and the size of the brain tumor are determined by computed tomography (CT), magnetic resonance tomography (MRT), or positron emission tomography (PET), which usually require contrast enhancing agents. In spite of all the advances in understanding the pathomechanism, tomographic imaging, and availability of new therapeutic tools, treatment of cancer in the brain is one of the most challenging problems in oncology as well as neurosurgery. The medial survival rate of patients with GBM of 12 months has been prolonged only slightly and cure remains elusive. The typical median survival is more than five years for grade II gliomas and two to five years for grade III gliomas. None of the currently available surgical tools, including operative microscopy, lasers, and image-guided surgery, enable detection and removal of all tumor tissue. There are several success stories in treatment of other cancers, but similar success has not been achieved for GBM. Characteristics of GBM that hinder efforts to cure it include the following: (i) access of chemotherapeutic agents to the tumor is more restricted than to other organs due to the blood–brain and blood–tumor barriers; (ii) aggressive local spread and diffuse infiltrating growth of the tumor; (iii) location of the tumor within the brain obviates radical extirpation; (iv) the residual tumor grows rapidly and resists curative treatment. Management of gliomas encompasses surgery to establish the histological diagnosis and to reduce the tumor burden and the mass effect, with the maintenance of the patient's neurological function in mind. Tissue autofluorescence was suggested to distinguish brain tumors from normal brain tissue by their reduced fluorescence emission (Goujon *et al.*, 2003). False positive assignments were problematic as not all regions with reduced autofluorescence were tumorous. In order to improve the selectivity, clinical studies were performed using a combination of autofluorescence and diffuse reflectance (Toms *et al.*, 2005). A multicenter clinical study used a fluorescence marker to guide the tumor resection intraoperatively (Stummer *et al.*, 2006). It could be demonstrated that more radical tumor resection leads to a slower regrowth of the tumor. However, the medial survival rate did not change significantly, probably because recurrences were formed by tumor cells that were not labeled by the marker and thus could not be detected. Here, false negative assignments were problematic as not all regions without fluorescence were normal. Raman spectroscopy has been suggested to improve both the selectivity and sensitivity as it does not depend on exogenous markers to characterize tissue.

In a first set of Raman studies, consecutive thin tissue sections were prepared. One section was transferred onto a glass slide, stained by H&E and served as a reference to select regions of interest before data acquisition. Another section was transferred onto slides made of quartz, calciumfluoride, or bariumfluoride, which exhibit lower background signals with 785 nm excitation than glass. Alternatively, the unstained tissue section can be stained for histopathological analysis after data acquisition. It is important to prepare the tissue as cryosection without embedding the sample in paraffin. Paraffin penetrates the sample and causes intense overlapping spectral contributions. Removal of paraffin in tissue sections by organic solvents should be avoided as it also dissolves hydrophobic components such as lipids. Raman maps of GBM tissue sections from 20 patients were collected (Koljenovic *et al.*, 2002). Vital tumor was discriminated from necrotic tissue by lower spectral contributions of lipids. Raman maps of dura and meningeoma tissue sections from 20 patients were collected (Koljenovic *et al.*, 2005a). Raman spectra of dura and meningeoma were distinguished by high collagen content of dura and increased lipid content of tumors. Raman and Fourier transform infrared (FTIR) spectroscopy mapping of human primary brain tumors were compared for three tissue sections (Krafft *et al.*, 2006). Smaller details such as cell nuclei and microcrystals were better resolved in Raman maps than in FTIR maps. The formation of microcrystals was an artifact, which was induced in hydrophobic constituents such as cholesterol or cholesterol ester by the drying process. Raman maps were also obtained from tissue sections of rat brains with tumors (Andrade *et al.*, 2007). Here, healthy tissue was discriminated from tumor and necrotic tissue. The abovementioned studies used the algorithm cluster analysis to segment the spectra into groups of similar spectral response. Details of clustering algorithms to analyze vibrational spectroscopic images have been reported elsewhere (Lasch *et al.*, 2004) and will not be described in this chapter.

Figure 28.2 shows a photomicrograph and a 56 × 11 Raman map of a meningeoma section. The sample area was raster scanned with a step size of 100 μm. Before cluster analysis, low-quality and low-intensity spectra were removed from the data set. Their positions are represented by white pixels. The remaining spectra were segmented into five groups by k-means clustering. The gray scaled plot of the cluster membership is shown in Figure 28.2b. The average spectra of each cluster are displayed in the bottom part of Figure 28.2. Spectrum 2C representing the dark gray cluster contains spectral contributions of collagen at 858, 941, 1249, and 1666 cm^{-1} (compare with spectrum 1C of Figure 28.1). In spectrum 2D, representing the light gray cluster, the intensities of collagen bands are reduced and the bands of α-helical and β-sheet proteins dominate with an amide I band at 1660 cm^{-1}. In spectrum 2E, representing the light gray shaded cluster, spectral contributions of DNA are more intense than in spectrum 2D. The distinction of both the clusters demonstrates that even small differences can be detected if the signal-to-noise ratio is sufficiently high. As spectral contributions of collagen are even more intense in normal dura (Krafft *et al.*, 2005b), the dark gray cluster is assigned to meningeoma with higher collagen content and the light gray cluster is assigned to meningeoma with

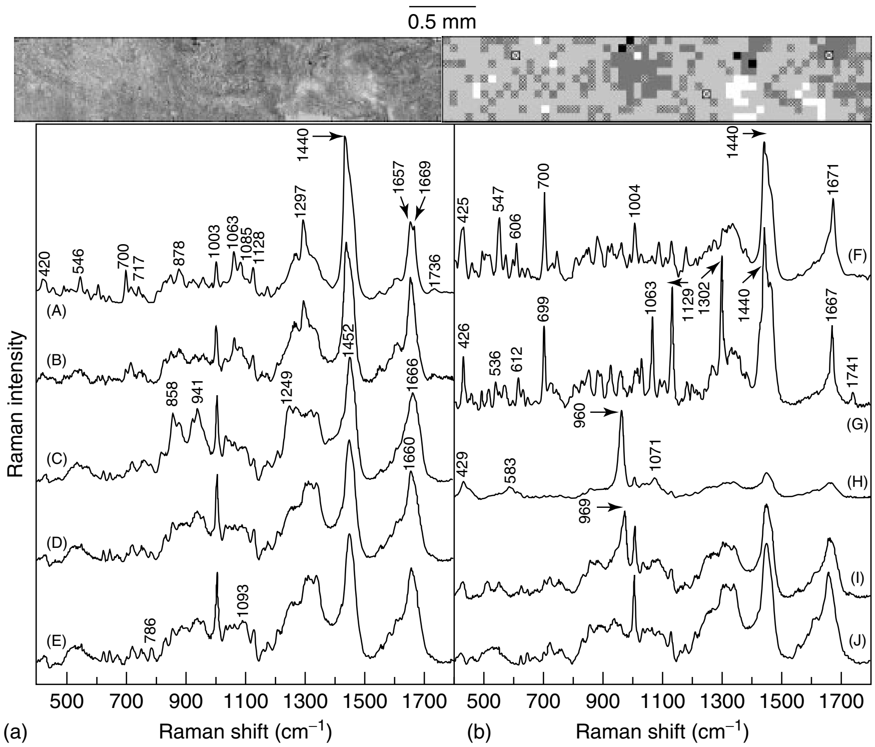

Fig. 28.2 (a) Photomicrograph and (b) cluster analysis of 56 × 11 Raman map of a meningeoma tissue section. Raman spectra from 400 to 1800 cm^{-1}: white (A) and gray (B) matter of normal brain tissue, meningeoma with higher collagen content (C), meningeoma with lower collagen content (D), meningeoma with higher DNA content (E), cholesterol microcrystal (F), cholesterol ester microcrystal (G), hydroxyapatite in brain tumor (H), calcium oxalate in brain tumor (I), glioblastoma multiforme (J).

low collagen content. Two additional spectral signatures are identified. Spectrum 2G representing the white crossed cluster resembles the Raman spectrum of almost pure cholesterol ester showing fatty acid bands (1063, 1129, 1302, and 1440 cm^{-1}), the ester band (1741 cm^{-1}), and bands near 425, 547, 606, 1671 cm^{-1}. This spectrum can clearly be distinguished from the Raman spectrum of cholesterol (spectrum 2F), which is often identified in normal brain tissue. Spectrum 2H representing the black cluster contains an intense band at 960 cm^{-1}, which is indicative of hydroxyapatite. Another stretching vibration band of hydroxyapatite is found at 1071 cm^{-1}, bending vibrations are located at 429 and 583 cm^{-1}. Calcification has also been observed in one of the first Raman studies of brain tumors (Mizuno *et al.*, 1994). The shift of the calcium phosphate band to 969 cm^{-1} in spectrum 2I is typical for calcium oxalate. The vibrational bands of calcium phosphate crystals were previously reviewed (Koutsopoulos, 2002). Calcification of tissue is not an unusual phenomenon in pathological tissues.

Calcium phosphates are formed as a result of functional irregularities and can be used as a marker to distinguish the pathological state and the type of tumors. For comparison, Raman spectrum of white matter (spectrum 2A) and gray matter (spectrum 2B) are included in Figure 28.2. These spectra represent the most abundant tissue types in normal brain. Spectral contributions of cholesterol are very intense in the spectrum of white matter, whereas they are less intense in the spectrum of gray matter. Both spectra are characterized by high lipid content, which offers the possibility to identify abnormal brain tissue including brain tumors. These results clearly demonstrate that the Raman spectrum provides a sensitive signature of tissue without additional markers, which can be used to analyze the chemical composition and determine the tissue type with supervised classification models.

In the second set of papers, nondried brain tissue has been studied by Raman spectroscopy. These experiments are important prerequisites to transfer the results from dried tissue sections toward *in vivo* applications. Seven Raman maps of native, human brain specimens were collected *ex vivo* using a Raman spectrometer coupled to a microscope and a motorized stage (Krafft *et al.*, 2005b). Normal white matter was confirmed to contain maximum lipid content, normal gray matter maximum water content, normal meninges maximum collagen content, and brain tumors lower lipid content. The methodology was modified to acquire Raman maps from whole murine brains (Krafft *et al.*, 2007). Instead of a microscope, the Raman spectrometer was coupled to a fiber-optic probe. The data coincided well with FTIR images from consecutive unstained tissue sections and with photomicrographs from consecutive H&E stained tissue sections. Brain metastases of malignant melanoma were detected in Raman maps.

The setup was applied to study brain tumors for the first time in a living mouse as shown in Figure 28.3a. The mouse was anesthetized and a cranial window made of CaF_2 was inserted onto the skull. The mouse was placed on a motorized table and a Raman map was collected from a predefined grid with 2 seconds exposure time per spectrum until the mouse awoke from anesthesia. The Raman spectra were segmented into the following groups by cluster analysis (Figure 28.3b): normal brain tissue (blue, cyan), bone (brown), the midline which contained blood (black), and brain metastases from malignant melanoma (red). The cluster distribution below the dashed line seems to be affected by motion artifacts of the animal. The cluster-averaged spectra representing each tissue type are shown in Figure 28.3c. The unprocessed spectra also contain the high wavenumber region from 1800 to 3550 cm^{-1} with the OH stretching vibrations of water near 3400 cm^{-1} and the CH stretching vibration at 2939 cm^{-1}. Owing to the high similarity of all spectra, the high wavenumber region is recognized to be less relevant for tissue recognition. The tumor detection in the low wavenumber region was facilitated by resonance enhancement of the Raman signals from melanin near 1600 cm^{-1} because the metastatic cells express this pigment. As a consequence of the inoculation of tumor cells in the carotid artery, the tumor was located in one brain hemisphere (here: left), whereas the other brain hemisphere (here: right) was tumor-free. The Raman band at 960 cm^{-1} is consistent

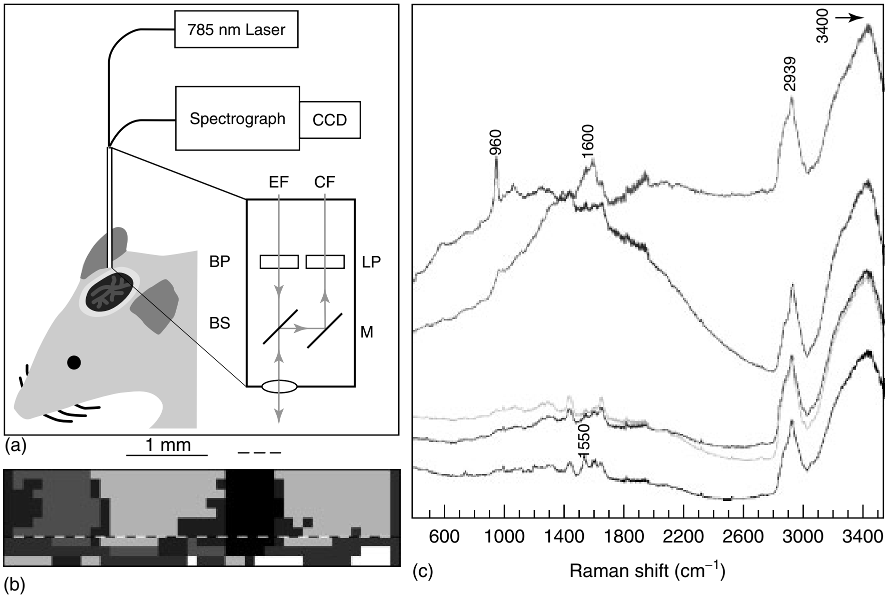

Fig. 28.3 (a) Schematic diagram of the Raman instrumentation: Radiation from a 785 nm laser is coupled into a fiber-optic probe. The scattered radiation is guided to the spectrograph and registered by a CCD detector. The filtered fiber-optic probe consists of an excitation fiber (EF), band pass (BP) filter, beam splitter (BS), lens (L), mirror (M), long pass (LP) filter, and collection fiber (CF). (b) Cluster analysis of 41×10 Raman map. (c) Unprocessed Raman spectra from 400 to 3550 cm^{-1} representing the average of each cluster. Color code: normal brain tissue (blue/cyan), oxygenated blood (black), bone (brown), and brain metastasis of malignant melanoma (red). (Please find a color version of this figure on the color plates.)

of hydroxyapatite of bone. The band at 1550 cm^{-1} belongs to a set of bands characteristic of oxygenized hemoglobin (Wood and McNaughton, 2007). Raman spectra of normal brain tissue are dominated by spectral contributions of lipids and proteins. Although the probe used in this study was too large for endoscopic applications, porcine brain tissue was recently characterized using a single, unfiltered fiber-optic probe (Koljenovic *et al.*, 2007). Instead of maps, single Raman spectra in the high wavenumber region were obtained of 11–19 anatomical structures in seven pig brains. These studies clearly illustrate the potential that Raman spectroscopy holds as a guidance tool during neurosurgical procedures.

28.3.1.2 Tumors of Epithelial Tissue

Epithelium is the collective term for cover and glandular tissue. It is composed of layers of cells that line the outer and inner surfaces of organs. The superficial layer is used to classify the stratification into simple, which consists of a single layer of cells, and stratified, which consists of more than one layer of cells. Stratified epithelium is usually composed of three layers. Cell division takes place in the basal cell layer (layer 1). Cells migrate and differentiate in the intermediate cell layer (layer 2) until they reach the superficial cell layer (layer 3). As the epithelium covers the surface of organs, it is exposed to the environment and is in contact to a broad range of potentially aggressive or harmful chemical and physical conditions that can induce a deregulation of cells' division in the basal cell layer. Therefore, tumors of the epithelium – usually called *carcinomas* – are the most common forms of cancer. Raman studies included carcinomas of the gastrointestinal tract, breast, lung, the urinary tract, cervix uteri, and skin (Section 28.3.4).

Gastrointestinal Tract: Oral Tissue, Esophagus, Intestinal Tissue The Raman and IR fingerprints of normal colon tissue and their distribution in control samples were previously compared in detail (Krafft *et al.*, 2008a). This sample served as a model for the analysis of soft tissue because it contains all four major tissue types such as muscle tissue, connective tissue, epithelium, and also nerve cells. The compilation of these data is important before the methods can be applied to pathological colon tissue such as colorectal adenocarcinoma, which is the third most common form of cancer and the second leading cause of death among cancers in the Western world. Colorectal adenocarcinomas originate from epithelial cells and are able to infiltrate the subjacent layer of colon and rectum. Figure 28.4a and b shows a photomicrograph and the cluster analysis of a 39 × 71 Raman map of an unstained mucosa section. Mucosa is the innermost, highly specialized layer of the gastrointestinal tract that is responsible for absorption and secretion. The Raman map was segmented into clusters of mucus at the top (pink, magenta, violet), submucosa (red), and muscularis

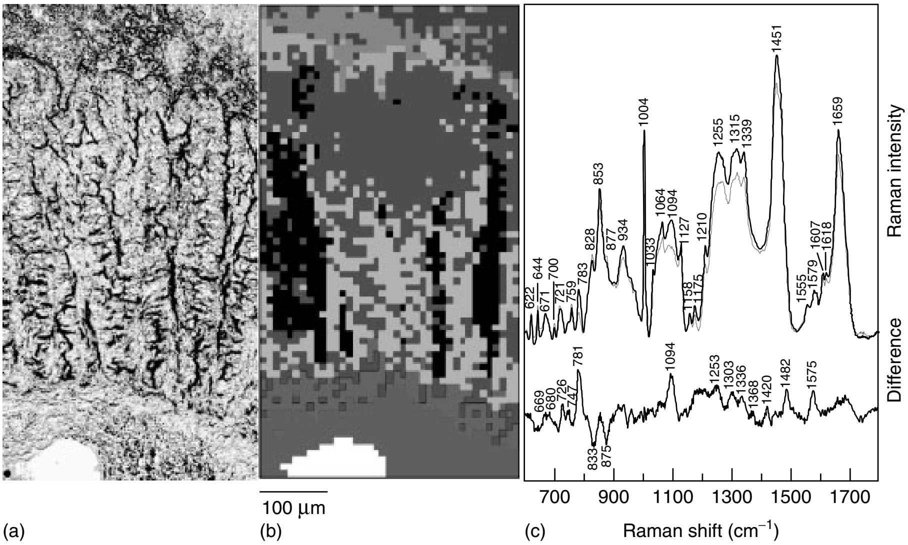

Fig. 28.4 (a) Photomicrograph and (b) cluster analysis of Raman map of mucosa in a colon tissue section. Raman spectra from 600 to 1800 cm^{-1} representing epithelial tissue (cyan cluster, black trace), glandular tissue (black cluster, gray trace), and difference spectrum = epithelial – glandular tissue. (Please find a color version of this figure on the color plates.)

mucusae (olive) at the bottom. The black cluster correlates with glandular features within epithelial tissue (blue/cyan). The segmentation seems to be affected by noise as evident by some heterogeneity in the cluster separation. The Raman spectra representing glands (black cluster) and epithelium (cyan cluster) are overlaid in Figure 28.4c. Although intensity differences in DNA bands at 783, 1094, and 1579 cm^{-1} are visible upon direct comparison, more subtle changes can be inspected in the difference spectrum. Before calculating the difference spectrum, both spectra were scaled to minimize differences in protein bands at 1004, 1451, and 1659 cm^{-1}. All positive difference bands coincide with DNA bands in spectrum 1G of Figure 28.1. The intensity of the DNA bands in the Raman spectra can be correlated with the grade of divisional activity. The mucosal epithelium of the gastrointestinal tract contains epithelial cells that undergo a continual turnover and renewal, and that has a higher proliferation rate than other colonic cells. This property makes these cells susceptible for tumor genesis. Negative difference bands at 833 and 875 cm^{-1} that are hardly visible in the overlaid spectra are associated with the production of mucus in glands.

The feasibility of fiber-optic coupled Raman spectroscopy for disease classification was demonstrated during *in vivo* gastrointestinal endoscopy (Shim *et al.*, 2000). The fiber-optic probe was passed through the endoscope instrument channel, placed in contact with the tissue surface, and spectra could be obtained in 5 seconds. The same instrument was also applied to differentiate adenomatous from hyperplastic polyps in the colon (Molckovsky *et al.*, 2003). A small endoscope system equipped with a micro-Raman probe was developed for *in vivo* Raman measurements of the esophagus and stomach in living rats (Hattori *et al.*, 2007).

Dysplasia is a term used in pathology to refer to an abnormality in maturation of cells within a tissue. It is often indicative for an early neoplastic process. The detection of dysplasia and early cancer is important because of the improved survival rates associated with an early treatment. Dysplasia was induced in the palate of rats by topical application of a carcinogen. Raman spectra were collected *in vivo* by a fiber-optic probe coupled to a Raman spectrometer with an NIR excitation laser (Bakker Schut *et al.*, 2000). Integration time was 100 seconds to record spectra for classification models. These models were trained to distinguish normal-, low-, and high-grade dysplastic tissue. Using signal collection times of 10 seconds, an independent set of spectra was obtained to test the classification models.

Long periods of exposure to gastroesophageal reflux can lead to a change in the lining of esophagus, whereby the normal squamous epithelial lining is replaced by protective columnar epithelial cells. This is termed *Barrett's esophagus* and is associated with an increased risk of developing esophageal adenocarcinoma. Patients with Barrett's esophagus undergo periodic endoscopic surveillance with multiple biopsy examinations in an attempt to differentiate inflammation from dysplastic and early cancerous lesions at a time when intervention can be beneficial. However, this surveillance approach is hindered by low sampling yield, random sampling error, pathology-associated costs, and delay in diagnosis. Fiber-optic Raman spectroscopy was recognized to enhance lesion detection in premalignant conditions. A Raman study of 89 esophageal samples from 44 patients was reported (Kendall *et al.*, 2003; Stone, Kendall and Barr, 2008; Stone *et al.*, 2002). Other normal and abnormal epithelial tissues from the larynx, tonsil, stomach, bladder, and prostate were included in these *ex vivo* investigations. Comparison of prediction of Raman spectral classification models with pathology gave sensitivities of 73–100% and specificities of 90–100%. Another *ex vivo* study collected Raman images to analyze tissue sections from 29 esophageal biopsies (Shetty *et al.*, 2006). Pathologists reviewed contiguous histopathologically stained tissue sections. Objective was to study the biochemical changes that distinguish the different stages of the disease Barrett's esophagus. Raman spectra of the rat esophagus were collected *ex vivo* with three different fiber-optic probes in order to mimic instrument calibration, probe-to-probe and day-to-day variations (Boere *et al.*, 2003). Barrett's and normal epithelium could be discriminated with accuracy higher than 93%.

Lung Tissue Lung cancer is the second most common cancer in humans and the most common cause of cancer deaths in the world. A Raman study of bronchial tissue investigated 28 specimens (Huang *et al.*, 2003) using a Raman spectrometer with 785 nm excitation coupled to a fiber-optic probe, which was developed earlier for skin studies (Huang *et al.*, 2001). Raman spectra were collected with 5 seconds exposure time and the spectral differences between normal and neoplastic tissues were analyzed. Another experimental approach using 1064 nm excitation, a fiber-optic probe, and a newly developed multichannel detector was presented to measure Raman spectra of seven lung tissue specimens (Kaminaka *et al.*, 2002). The data obtained with 400 seconds exposure time distinguished normal and cancerous tissue and also two different types of lung carcinomas. Characterization of the biochemical composition of normal bronchial tissue is important to understand the biochemical changes that accompany lung cancer development. Frozen sections of normal bronchial tissue were studied by Raman imaging (Koljenovic *et al.*, 2004). Subsequent comparison of Raman images with histologic evaluation of stained sections enabled to identify the spectral features of bronchial mucus, epithelium, fibrocollagenous stroma, smooth muscle, glandular tissue, and cartilage. Raman maps were collected from lung tissue and congenital cystic adenomatoid malformation (CCAM) of two patients and compared with FTIR images (Krafft *et al.*, 2008b). At low lateral resolution, a lower red blood cell content and higher lipid content were found in CCAM than in normal lung tissue. At higher lateral resolution, accumulations of lipids and glycogen were identified in CCAM. The lipid aggregates consist of phosphatidylcholine.

Breast Tissue Worldwide, breast cancer is the most common cancer amongst women, with an incidence rate more than twice that of colorectal cancer and cervical cancer and about three times that of lung cancer. Therefore, Raman spectroscopy was applied to complement the existing diagnostic tools. Raman mapping assessed the chemical composition and morphology of breast tissue (Shafer-Peltier *et al.*, 2002) and lymph nodes in tissue sections of breast cancer patients (Smith *et al.*, 2003). Shafer-Peltier (2002) presented a model, which fits macroscopic tissue spectra with a linear combination of basis spectra derived from spectra of the cell cytoplasm, cell nucleus, fat, β carotene, collagen, calcium hydroxyapatite, calcium oxalate dihydrate, cholesterol-like lipid deposits and water. Raman fiber-optical probes are under development to classify normal breast tissue, benign disease, and malignant breast cancer *in vivo* during breast lumpectomy surgeries (Motz *et al.*, 2005). The instrument is capable of collecting and processing Raman spectra in less than 2 seconds. Previous research showed the ability of Raman spectroscopy to accurately diagnose normal, benign, and malignant lesions of the breast *in vivo* in a laboratory setting with high sensitivities and specificities (Haka *et al.*, 2005). Raman spectra were also collected *in vivo* during partial mastectomy breast surgery (Haka *et al.*, 2006) using the previously developed fiber-optical instrument (Motz *et al.*, 2005). Kerr-gated Raman spectroscopy enables to non-destructively probe calcifications buried within breast tissue (Baker *et al.*, 2007). The method used a picosecond-pulsed laser combined with a fast temporal gating of Raman scattered light to collect spectra from a specific depth within scattering media. Spectra characteristic of both hydroxyapatite and calcium oxalate were obtained at depth up to 0.96 mm. As calcium oxalate is associated with benign lesions and hydroxyapatite is found mainly in proliferative lesions including carcinoma, the detection of calcification can indicate the disease state. Typical Raman spectra are already shown in Figure 28.2. Earlier, microcalcifications in benign and malignant breast lesions and their chemical composition were probed by conventional Raman microscopic spectroscopy (Haka *et al.*, 2002).

Urinary Tract: Prostate, Bladder Prostate tissue is structurally complex, consisting primarily of glandular ducts lined by epithelial cells and supported by heterogeneous stroma, and containing also blood vessels, blood, nerves, ganglion cells, lymphocytes, and stones. Benign prostatic hyperplasia is one of the most common diseases in elderly men. It describes an overgrowth of the epithelial and fibromuscular tissues in the human prostate, in which both epithelium and stroma continue to proliferate at different rates. Twenty-nine bladder samples and 38 prostate samples were studied *in vitro* by a Raman system coupled to a fiber-optic probe (Crow *et al.*, 2005). The spectra were correlated with the histologic features and used to develop classifications models. The bladder algorithm was able to differentiate benign samples (e.g. normal and cystitis) from malignant samples (e.g. transitional cell carcinoma). The prostate algorithm was able to differentiate benign samples (e.g. prostatic hyperplasia and prostatitis) from malignant samples (prostate cancer). Kerr-gated Raman spectroscopy uses a picosecond-pulsed laser as well as fast temporal gating of collected Raman scattered light. This technique suppresses signals from both surface and deep layers in bladder and prostate tissue and enables collecting Raman spectra from different depths (Prieto *et al.*, 2005). Another new approach used high wavenumber Raman microspectroscopy to image bladder tissue sections (Koljenovic *et al.*, 2005b). Raman spectroscopy was applied to discriminate between nontumor and tumor bladder tissue sections from 15 patients and to determine the biochemical differences therein (de Jong *et al.*, 2006).

Cervix Uteri Cervical cancer is malignant cancer of the cervix uteri. According to the WHO (fact sheet 297: cancer), cervical cancer is the fifth most deadly cancer in woman. While FTIR spectroscopy was mainly applied to screen cell specimens collected during Pap smear tests, focus of the Raman spectroscopic research on cervical tissue was *in vivo* assessment of squamous dysplasia by fiber-optic probes (Mahadevan-Jansen *et al.*, 1998; Utzinger *et al.*, 2001). The first study reported the design and testing of a fiber-optical Raman system, whereas the second study presented the results of a pilot clinical trial at three clinical sites. Raman spectra were measured from one colposcopically normal and one abnormal area of the cervix in 13 patients. Comparison to histologic analysis revealed that the spectral data were able to identify high-grade squamous dysplasia, misclassifying only one sample.

28.3.2
Mineralized Tissue

Raman spectroscopy has been used for several years to characterize mineralized structures in living organisms, notably bone, teeth, and calculus. Calculus is the collective term for a stone that forms in an organ or duct of the body. The application of vibrational spectroscopy to the study of mineralized tissue was reviewed (Carden and Morris, 2000). Mineralized biological samples are composed of an inorganic matrix (calciumphosphate) and an organic matrix. With the application of Raman imaging spectrometers to mineralized biological samples, spatial variations of the mineral content, mineral crystallinity, and the content/nature of the organic matrix consisting predominantly of collagen type 1 and cells are obtained. The spectral data can monitor the extent of mineral formation in tissue using a mineral to protein ratio. Other parameters monitor the size and perfection of the crystals and the ratio of carbonate ions and phosphate ions in the hydroxyapatite lattice. Furthermore, the interactions between prosthetic implants and tissue can be characterized.

28.3.2.1 Bone

The structural properties of bone were examined by Raman imaging at 785 nm excitation with micrometer range spatial resolution (Carden *et al.*, 2003; Timlin *et al.*, 1999). In contrast to conventional histological techniques, Raman spectroscopic methods do not require a special sample preparation, for example, homogenization, decalcification, extraction, or dilution. They enable to investigate hard tissue under native conditions. Recently, Raman images from bone were also collected at 532 nm excitation after the fluorescence background was reduced by photochemical bleaching (Golcuk *et al.*, 2006). Advantages of the shorter excitation wavelength are higher quantum efficiency of CCD detectors and more intense Raman signals due to the fourth power dependence of scattering intensity with excitation frequency. Both effects result in shorter acquisition times. Typically, Raman images of $300 \times 100\ \mu m^2$ areas with 2 μm resolution just require 10 minutes. The first transcutaneous Raman spectrum of human bone *in vivo* has been obtained at skin-safe laser illumination levels using SORS (Matousek *et al.*, 2006). This technique, which is based on collecting Raman spectra away from the point of laser illumination using concentric rings of optical fibers, provides chemically specific information on deep layers of human tissue well beyond the reach of existing comparative approaches. Raman imaging studies of bone-related diseases include craniosynostosis, which is an abnormal condition of bones making up the skull (Crane *et al.*, 2005; Tarnowski *et al.*, 2004). Microdamages in bone that were induced using a cyclic fatigue loading regime were investigated by Raman imaging (Timlin *et al.*, 2000). The bone chemical structure response to mechanical stress was studied by high-pressure Raman spectroscopy (de Carmejane *et al.*, 2005). The results enabled a deeper understanding of the response of tissue to mechanical stress and demonstrated that individual mineral and matrix constituents respond differently to pressure.

28.3.2.2 **Teeth**

In the context of dentistry, Raman imaging studies analyzed the chemical composition of the enamel/adhesive resin interface (Miyazaki *et al.*, 2003), the monomer/mineral distribution in intertubular region of adhesive/dentin interfaces (Wang, Spencer and Yao, 2006) and the adhesive/caries-affected dentin interface (Wang, Spencer and Walker, 2007). Dental resin composites cured with different light sources (Conti *et al.*, 2005), the post-polymerization shrinkage in dental resins (Truffier-Boutry *et al.*, 2006), and natural and deproteinated tooth enamel and dentin (Fattibene *et al.*, 2005) were studied. Another interesting application of Raman spectroscopy is the detection of dental caries in teeth. While OCT was used to screen carious sites and to determine lesion depth, fiber-optical Raman spectroscopy provided biochemical confirmation of caries (Ko *et al.*, 2005). OCT imaging of tooth samples demonstrated increased light backscattering intensity at sites of suspicious carious lesions as compared to the sound enamel. Furthermore, sound tooth enamel exhibited strong Raman polarization anisotropy, whereas early caries consistently showed a lower degree of Raman polarization anisotropy. Raman spectroscopy was also applied to assess how IR laser photobiomodulation improves the quality of healing bone around dental implants in rabbits (Lopes *et al.*, 2007).

28.3.2.3 **Calculi**

A calculus consists of a concretion of material, usually mineral salts. Stones cause various benign and malignant conditions that include nephrolithiasis (kidney stones), cholelithiasis (gallstones), and testicular microlithiasis. The Raman spectroscopic investigation is direct, fast, and nondestructive, and does not require tedious sample preparation. There are at least two Raman-based approaches to the quantitative analysis of the stone composition. Partial least squares (PLS) techniques will yield highly precise results if the composition of the unknown material is restricted to a reasonably well-defined range, with predictable components present. Another approach is based on supervised methods or on library searching. An unknown sample spectrum is then compared to a number of well-assigned library spectra and the best correlated spectrum is found. In particular, Raman spectroscopy is well suited to detect phosphate-type kidney stones because the main constituent hydroxyapatite has a distinctive line at 961 cm^{-1} which can be used as a "marker" band. The Raman and IR spectra of calculi have previously been summarized (Nguyen and Daudon, 1997).

Kidney Stones Approximately 5% of the population will suffer from kidney stones at some point in their life. About 70% of all kidney stones are composed of calcium oxalate, and small amounts of calcium hydroxyapatite, uric acid, and magnesium ammonium phosphate. Other types of stones consist of 50% hydroxyapatite or calcium monohydrogen phosphate or are composed of cystine. Raman spectroscopic studies on kidney stones demonstrated that vibrational spectroscopy deliver useful information on the stones' structure and composition. Raman spectra of standard phosphate compounds were obtained and applied to the analysis of phosphate-type kidney stones (Kodati *et al.*, 1991). Monohydrate and dihydrate forms of oxalate inclusions were differentiated in tissue section (Pestaner *et al.*, 1996). The detection of struvite in urinary stones was compared between Raman and IR spectroscopy (Takasaki, 1996).

Gallstones Ten to fifteen percent of the population carries gallstones that stay asymptomatic in most cases (ca. 75%). Gallstones are commonly classified into three classes: cholesterol (white), pigment (black), and mixed (brown) stones. Each type of stones has its own treatment regimens. Determination of composition is one fundamental part in establishing the cause and likelihood of recurrence of stones, as well as an appropriate treatment. Raman spectroscopy was used as an analytical technique to determine the structural composition of gallstones from 40 patients (Paluszkiewicz *et al.*, 1998). The utilization of Raman microscopic techniques permitted to map the distribution of the gallstone components leading to identification and characterization of the site of nucleation (Vitetta, Best and Sali, 2000). Polymorph modifications of calcium carbonate in gallstones were investigated using X-ray powder diffraction and spectroscopic techniques including Raman spectroscopy (Palchik and Moroz, 2005).

Testicular Microliths Testicular microlithiasis is associated with various benign and malignant conditions. However, the molecular constitution of these microliths is largely unknown. Microliths in six samples were investigated by Raman imaging in order to identify their molecular constitution (de Jong *et al.*, 2004). The results showed that microliths are composed of pure hydroxyapatite and that glycogen deposits surround all microliths in the samples associated with germ cell neoplasms but not in the benign case.

28.3.3
Skin

According to a review (Krafft and Sergo, 2006), Raman spectroscopic applications to skin had the highest frequency from 1996 to 2006 in the biomedical field, probably because it is the largest human organ and its surface location makes it easily accessible. Applications to corneocyte biology, pharmacology, and disease diagnosis have been reported. Human skin can be considered as a multilayered system, which is divided into the stratum corneum, the epidermis, and the dermis. The stratum corneum is theskin's outermost layer and the main protective barrier against water loss, microorganisms, and toxic agents. This thin (10–20 μm) superficial region has a biphasic structure consisting of anucleated, keratin-rich corneocytes embedded in a highly ordered, lamellar lipid network of ceramides, fatty acids, and cholesterol. A major function of the underlying epidermis is to generate the stratum corneum. The principle cell of the epidermis is the keratinocyte, which differentiates as it migrates toward the stratum corneum. Terminally differentiated keratinocytes generate corneocytes. During differentiation of keratinocytes to corneocytes, a large insoluble protein called *profilaggrin* is converted to filaggrin. Processing of filaggrin includes proteolysis of the protein into its constituents – amino acids, amino acids derivatives, and salts. The resulting mixture, known as *natural moisturizing factor* (*NMF*), provides sufficient hydration to help keep the skin flexible and to facilitate various enzymatic reactions in the pathway that terminates with corneocytes desquamation. The thickness of the epidermis ranges from 40 μm on the eyelid to >1 mm on the palms. The dermis (0.4–4 mm thick) is the underlying layer of the epidermis. The dermis consists of tough connective tissue along with specialized structures. Collagen comprises approximately 75% of the dry weight of the dermis.

28.3.3.1 Corneocyte Biology and Pharmacology

Combined approaches of Raman imaging with FTIR imaging studied the permeation of lipids into skin (Xiao *et al.*, 2005a, 2005b) and the structure during corneocyte maturation (Zhang *et al.*, 2006). Confocal Raman microspectroscopy was applied to the stratum corneum to identify variations in keratin-containing areas (Xiao *et al.*, 2004) and to obtain depth profiles of molecular concentration gradients and transdermal drug delivery up to 200 μm (Caspers *et al.*, 2001). The setup was applied as a noninvasive *in vivo* optical method to determine water concentration profiles of the stratum corneum and to determine changes therein as a result of hydration of the skin (Caspers *et al.*, 2000). Its combination with the imaging modality confocal scanning laser microscopy provides optical sections of the skin without physically dissecting the tissue (Caspers *et al.*, 2003). Natural variations and reproducibility of Raman spectra of normal human skin was studied by a fiber-optic FT-Raman spectrometer (Knudsen *et al.*, 2002). However, exposure times of 8 minutes were reported with this technique, which is much longer than the exposure times for dispersive NIR Raman systems of 30 seconds for the low wavenumber region and 3 seconds for the high wavenumber region (Caspers *et al.*, 2001).

28.3.3.2 Skin Cancer

Skin cancers that include squamous cell carcinoma, malignant melanoma, and basal cell carcinoma (BCC) are the cancers with the highest incidence worldwide. Understanding the molecular, cellular, and tissue changes that occur during skin carcinogenesis is central to cancer research in dermatology. As for many other tissues, vibrational spectroscopy has been used to evaluate these changes (review: (Eikje, Aizawa and Ozaki, 2005)). BCC is the most common cancer of the skin. Raman images were acquired from 15 sections of BCC and compared with histopathology (Nijssen *et al.*, 2002). In this sample set, 100% sensitivity and 93% selectivity were demonstrated. Melanoma is the most aggressive skin cancer and is invariably fatal if left untreated. Clinical diagnostic sensitivity is in the order of 80% for trained dermatologist and much worse for untrained medical personnel. Melanoma removal at an early stage is almost always curative leading to a good prognosis for the patient. Melanoma, pigmented nevi, BCC, seborrheic keratoses, and normal skin were studied by Raman spectroscopy (Gniadecka *et al.*, 2004). The sensitivity and specificity of an artificial neural network classification for diagnosis of melanoma were 85% and 99%, respectively. Recently, it was demonstrated that micro-Raman imaging is also well suitable to distinguish between normal and malignant skin cells at a subcellular level (Short *et al.*, 2006). The Raman scattering was excited at 633 and 515 nm. An excitation power of less than 2 mW ensured that cells were not damaged.

28.3.3.3 **Raman Instrumentation for Skin Studies**

The potential of Raman spectroscopy as a tool for skin studies is now being realized by two commercially available Raman systems. A third Raman system has recently been introduced (Lieber and Mahadevan-Jansen, 2007). A first Raman system was designed, which can be applied noninvasively to record depth concentration profiles of living skin with a resolution of $<5\ \mu m$ (River Diagnostics, Netherlands). Using laser excitation at 671 and 785 nm, a custom built detection system and software tools, various parameters of the skin can be analyzed. Owing to the limited penetration of typically 20–40 μm, most studies are restricted to the stratum corneum. Applications include the determination of (i) the hydration of the skin, (ii) the composition of the NMF, and (iii) the penetration of topically applied materials. This technique can be expected to experience growing use by skin scientists in the cosmetics, pharmaceutical, and dermatological research fields.

A second Raman system determines the carotenoid level in skin (Pharmanex, USA). Carotenoids are known to be important in human health. Most of the health benefits of carotenoids are associated with their action as antioxidants, that is, they protect cells and tissues from the effects of free radicals and reactive oxygen species. As carotenoids are not soluble in water, they are transported in blood by low-density lipoproteins (LDLs). An increase in blood carotenoids is reflected in an increase of carotenoid in all the organs in the body, which can take up lipoproteins, including skin. Thus, the direct measurement of carotenoids on skin provides a good indication of the carotenoid concentrations in blood and other tissues and of the body's overall antioxidant status. Owing to the lack of any accepted noninvasive technology for the detection of carotenoids in living human tissue, RRS was applied as a novel and rapid approach (Hata *et al.*, 2000). The Raman signals originating from single bond and double bond stretch vibrations of the π-conjugated carbon backbone of carotenoids at 1159 and 1525 cm^{-1} are enhanced as high as a factor of 100 000 under proper conditions (Koyama, 1995). The instrument uses light emitting diodes (LEDs) that are tuned to 471.3 and 473 nm in place of a laser. Although there is a large native fluorescence level in the skin upon Raman excitation with blue light, carotenoid levels over a large concentration range can be measured.

28.3.4
Atherosclerosis

Atherosclerosis is a disease affecting arterial blood vessels. The atherosclerotic plaque is divided into (i) the atheroma, which is composed of macrophages nearest the lumen of the artery; (ii) underlying areas of cholesterol crystals; and (iii) calcification at the outer base of older and more advanced lesions. Previous research has suggested that the microscopic morphology and chemical composition rather than the anatomy of an atherosclerotic plaque determine plaque stability and disease progression. Developments of compact clinical Raman systems, specially developed Raman catheters and future directions of Raman spectroscopy in cardiovascular medicine were summarized (van de Poll *et al.*, 2002c). The performance of a fiber-optic probe was tested *in vitro* with aorta tissue (Motz *et al.*, 2004). Raman spectra of normal aorta, noncalcified atherosclerotic plaque, and calcified atherosclerotic plaque that were collected in only 1 second with 100 mW excitation could be fitted by model spectra to assess the chemical composition. Another intravascular catheter for Raman spectroscopic detection and analysis of coronary atherosclerotic disease has been developed (Komachi, Sato and Tashiro 2006). The catheter having an outer diameter of 2 mm consisted of a side-view type micro-Raman probe, an imaging fiber bundle, a working channel, and a balloon. By inflating the balloon, the probe was brought close to the inner wall of a modeled blood flow system and detected a phantom target buried in the wall. Raman maps of human atherosclerotic plaques were generated to investigate the chemical composition of the pigment ceroid in cross-sections of the intimal surface (van de Poll *et al.*, 2002a) and to investigate plaque development in mice (van de Poll *et al.*, 2002b). Compact clinical Raman systems and dedicated, miniaturized fiber-optic Raman catheters with a side viewing geometry were used *ex vivo* in human coronary arteries (van de Poll *et al.*, 2003) and *in vivo* in lambs and sheep to illuminate the blood vessel wall and to collect Raman scattered light (Buschmann *et al.*, 2000). The *in vivo* intravascular Raman signal obtained from a blood vessel was found to be a simple summation of signal contributions of the blood vessel wall and of blood. Therefore, algorithms previously developed from single point Raman spectra (Römer *et al.*, 1998) could be adapted to extract information about the chemical composition of blood vessel walls from *in vivo* Raman spectra.

28.3.5
Ophthalmology

Ophthalmology is the branch of medicine, which deals with the diseases and surgery of the visual pathways. Because Raman signals are typically weak, intense lasers in combination with sophisticated light collection are usually used. Although intense laser radiation can potentially harm the delicate structures in the visual system, ocular tissue has been found to be a very suitable target for Raman spectroscopy for two reasons. First, the ocular media (cornea, lens, and vitreous) generally have good optical clarity, which enables high penetration of laser excitation and optical detection of scattered radiation. Secondly, the retina also contains the pigment carotenoid. Owing to its two orders of magnitude higher abundance compared with skin, the excitation intensities can be reduced to light levels that are within established safety ranges. No other biological molecules found in significant concentrations in human ocular tissues exhibit similar resonant enhancement with blue laser excitation so the *in vivo* carotenoid Raman spectra are remarkably free of confounding responses.

28.3.5.1 Macular Degeneration of the Retina

Age-related macular degeneration (AMD) is the leading cause of irreversible blindness in the elderly. Carotenoid macular pigments may provide protection against light-induced oxidative damage and lower the risk for this pathology. To explore the role of carotenoids in modifying the risk of AMD further, clinical studies to quantify their levels in the living human macula are impaired by the lack of appropriate instruments. Raman spectroscopy has been shown to assess the concentration and the spatial distribution of carotenoids in the retina through undilated eyes noninvasively within a fraction of a second (Bernstein *et al.*, 2004). Ocular Raman instruments were designed for this special application (Spectrotek LC, USA). However, the company is no longer in business yet. A new generation of ocular RR scanners will incorporate wide-field imaging of macular carotenoid distributions (Gellermann *et al.*, 2002).

28.3.5.2 Chemical Composition of the Vitreous and the Lens

Imaging vitreous is a quest to view what is, by design, invisible. The combined use of techniques including Raman spectroscopy could provide better imaging of vitreous for future investigational and clinical purposes. Raman spectroscopy was applied to image the vitreous in order to detect altered vitreous molecules such as glycated collagen and other proteins in diabetic vitreopathy (Sebag, 2004). Diabetes mellitus succeeded by diabetic cataract was induced to rats, and eye properties deterioration was monitored from clinical standpoint and by Raman spectroscopy (Furic, Mohacek-Grosev and Hadzija, 2005). The formation of nuclear cataract was induced in guinea pigs with hyperbaric oxygen, and changes in disulfides (S–S) and sulfhydryls (S–H) were measured in isolated lenses by Raman spectroscopy (Gosselin *et al.*, 2007). Using filipin as an external Raman label, which binds specifically to cholesterol, the distribution of cholesterol in a rat eye lens was determined by Raman spectroscopy (Sijtsema *et al.*, 1996). The protein distribution was obtained by using intrinsic Raman protein bands. Calcification of senile cataractous lens was determined by Raman microscopic spectroscopy (Chen *et al.*, 2005).

28.4
Raman Spectroscopy in Clinical Chemistry

28.4.1
Determination of Glucose

A number of common clinical chemistry tests have proven to be feasible using Raman spectroscopy. Among the most important applications is the fast and reliable determination of the blood glucose level of diabetic patients. Normally, glucose level is tightly regulated in human body. Failure to maintain blood glucose in the normal range between 70 and 150 mg dl^{-1} leads to conditions of persistently high (hyperglycemia) or low (hypoglycemia) blood sugar. Diabetes mellitus, which is characterized by persistent hyperglycemia of several causes, is the most prominent disease related to failure of blood sugar regulation. Raman spectroscopy was explored as a reagent-free tool for predicting the concentrations of different parameters in blood serum and serum ultrafiltrate (Rohleder, Kiefer and Petrich, 2004). In an investigation using samples from 247 blood donors, the concentrations of glucose, triglycerides, urea, total protein, cholesterol, high-density lipoprotein (HDL), LDL, and uric acid were determined with accuracy within the clinically interesting range. Furthermore, it was shown that ultrafiltration can efficiently reduce fluorescent light background to improve prediction accuracy such that the relative coefficient of variation decreased for glucose and urea in ultrafiltrate by more than a factor of 2 when compared to serum.

Another approach for the detection of blood glucose level is the application of SERS substrates. The normal Raman cross-section of glucose is rather small and additionally the adsorption of glucose molecules on the surface of SERS substrates made out of silver or gold is weak. Therefore, a direct identification of glucose using SERS is limited. Wu *et al.* (2006) applied a modified analysis system: in a first step glucose was modified by an enzymatic reaction to gluconolacton, which converts *o*-phenylenediamine in an oxidation–condensation to azoaniline. Azoaniline, on the other hand, can easily be detected by SERS spectroscopy. With this approach, a glucose concentration of minimal 0.5 mM could be correctly predicted.

A more direct approach is the use of metal films over nanospheres (FON) as SERS substrates. These substrates were modified with a partition layer in which glucose can penetrate. This enables then to monitor the glucose SERS signals in combination with the SERS signals of the partition layer. Yonzon *et al.* (2004) used Cu FON to monitor glucose concentration in the presence of bovine serum albumin to simulate blood environment.

Stuart *et al.* (2005) used both Au and Ag FON in order to monitor the time stability of the SERS substrate for the detection of glucose. Lyandres *et al.* (2005) used Ag FON in combination with a mixed decanethiol (DT)/mercaptohexanol (MH) partition layer to improve the detection level.

Beside the classical analysis of the glucose levels in blood, the correlation of blood with other body fluids can be used to measure noninvasively. One possibility is the aqueous humor in the anterior chamber of the eye. Since the aqueous humor is an ultrafiltrate of the blood, its glucose level correlates with the blood glucose level. Pelletier, Lambert, and Borchert (2005) investigated human aqueous humor samples collected from cataract surgeries as well as artificial aqueous humor samples as a model system. With this model, they could show that a model out of glucose and different other substances such as lactate or urea could be used to predict the glucose level of human aqueous humor *in vitro*. Lambert, Pelletier, and Borchert (2005) tested this assumption on the anterior chamber of rabbits and could predict the blood glucose level without damaging the retina. Sideroudi *et al.* (2007) measured the concentrations of antibiotics, for example, ciprofloxacine, in an artificial anterior chamber.

28.4.2
Serum Analysis

Beside the status of blood cells, also the composition of the serum reveals information about the health condition. Qi and Berger (2007) investigated serum samples with a liquid core optical fiber (LCOF). In this study 71 blood serum and 61 urine samples were investigated with NIR excitation in combination with PLS analyzing the concentration of 13 chemicals, for example bilirubine, cholesterol HDL and LDL at the same time. Rohleder *et al.* (2005) analyzed serum and serum ultrafiltrate samples from 247 blood donors to predict the concentration of, for example, glucose, triglycerides, HDL, and LDL with an accuracy within the clinical interesting range.

Harz *et al.* studied cryoprecipitated plasma samples by means of UV-resonance Raman spectroscopy. With this method, the increase of von Willebrand Factor multimers can be identified in samples from diseased persons enabling for early detection of microangiopathy (Harz *et al.*, 2006).

28.4.3
Malaria Investigations

Malaria is one of the most common infectious diseases and an enormous public health problem. The disease is caused by protozoan parasites of the genus Plasmodium. The parasites multiply within red blood cells, also known as *erythrocytes*, where they digest a major proportion of red cell hemoglobin. As free heme (ferriprotoporphyrin IX), a product of hemoglobin digestion, is toxic for the parasite it detoxifies free heme by conversion into an insoluble crystal called *hemozoin* or *malaria pigment*. The early detection of parasite infection is vital to the application of an early and appropriate therapy. Especially, malaria is a disease where the mode of action of the applied pharmaceuticals is not fully understood. Using resonance Raman spectroscopy, the characterization of hemozoin under different physiological conditions (Frosch *et al.*, 2007b) as well as the influences of quinoline additives (Solomonov *et al.*, 2007) or chloroquine (Frosch *et al.*, 2004) of β-hematin crystals can be investigated. The synthetic compound β-hematin is a structural analog of hemozoin.

Comparison of resonance Raman spectra of β-hematin, normal erythrocytes, parasite-infected erythrocytes, and lysed samples of the parasite-infected erythrocytes at different wavelengths enables for the clear distinction between normal and infected erythrocytes (Ong *et al.*, 2002). The characterization of hemozoin and hemoglobin inside the food vacuole of *Plasmodium falciparum* and β-hematin as well as the monomer hematin shows the identical Raman spectra of hemozoin and β-hematin with an exceptionally strong marker band at 1374 cm^{-1} (Frosch *et al.*, 2007a; Wood *et al.*, 2003). Using micro-Raman spectroscopy, the structural changes of hemozoin in *P. falciparum* during chloroquine exposure can be monitored (Webster *et al.*, 2008). Raman acoustic levitation spectroscopy (RALS) of lysed red blood cells allows for the faster detection of malaria infection inside 10 µl droplet (Puskar *et al.*, 2007).

28.4.4
Analysis of Hair and Nails

For forensic application, Widjaja and Seah Regina Kim (2006) were able to differentiate between Raman spectra of fingernails and toenails using NIR excitation. Beside nails, hair is a good reservoir for accumulated chemicals over time as well as a possibility to investigate physiological changes with minimal invasive effort. Trichothiodystrophy (TTD), for example, is a disease induced by a reduced content of cystine in the proteins, where the hair breaks very easily. Osada *et al.* studied the molecular change in the hair protein during TTD. Patients with TTD exhibit hair with altered disulfide bonds. Here, the Raman signals were shifted to higher wavenumber region indicating a change toward a *gauche–gauche–trans* conformation in the TTD hair (Osada *et al.*, 2003). Normal hair shafts and hair shafts from patients exhibiting TTD were compared by Raman line scanning over the hair (Schlücker *et al.*, 2006) or by Raman point measurement (Liang *et al.*, 2006). Both studies found conformational change of the disulfide bonds and correlated this to the hair brittleness.

28.5
Conclusions and Outlook

The increased interest in Raman spectroscopy for applications in medicine and life science is clearly demonstrated by recent book publications. A book entitled "Vibrational Spectroscopy in Life Science" described Raman and IR spectroscopy of specific biological systems, such as proteins and membranes (Siebert and Hildebrandt, 2007). Another book entitled "Vibrational Spectroscopy for Medical Diagnosis" summarized various biomedical applications including pathology, clinical chemistry, detection of cancer, and identification of pathogenic microorganisms (Diem, Griffiths and Chalmers, 2008). The forthcoming book "Infrared and Raman Spectroscopic Imaging" focuses on imaging applications in biology and medicine as well as in chemical, pharmaceutical, and food industries (Salzer and Siesler, 2009). According to a recent review (Krafft and Sergo, 2006), IR spectroscopy was applied two times more frequently from 1996 to 2006 than Raman spectroscopy to tissue diagnosis. Although single Raman spectra can be acquired in seconds, acquisition of Raman images that consist of thousands of spectra is usually more time consuming and the signal-to-noise ratio is inferior compared with FTIR imaging. Nevertheless, Raman imaging offers advantages in terms of (i) studying nondried specimens, (ii) the maximum, diffraction limited lateral resolution is higher, (iii) the spectral range can easily be extended to lower wavenumbers, (iv) the technique can be coupled to fiber-optic probes, and (v) enhancement effects can be utilized. For example, the RR effect was used for rapid and sensitive detection of carotenoids in skin

(Hata *et al.*, 2000) and retina (Bernstein *et al.*, 2004), the malaria pigment hemozoin in erythrocytes (Frosch *et al.*, 2007b), and melanin in brain metastases of malignant melanomas (Krafft *et al.*, 2007). Using nonlinear Raman effects such as coherent anti-Stokes Raman spectroscopy (CARS) signal intensities can also be enhanced by several orders of magnitude. CARS is a four-photon process, which uses two picosecond-pulsed laser for excitation under a defined geometry. Owing to the intense signals, CARS in combination with fast scanning mirrors offers an attractive perspective to reduce the acquisition time of Raman images. CARS imaging with video rate microscopy of skin tissue in a live mouse (Evans *et al.*, 2005) and brain structures (Evans *et al.*, 2007) at subcellular resolution has already successfully been performed. With the ongoing progress in light sources, optical components, detectors, and algorithms for data processing, the number of Raman spectroscopic applications in medicine will increase in the future. Although in its early stages, one can be optimistic that Raman spectroscopy has the potential to be fully accepted as a diagnostic tool with sufficient sensitivity and specificity for rapid and nondestructive *in vitro, ex vivo*, and *in vivo* analyses.

References

Andrade, P.O., Bitar, R.A., Yassoyama, K., Martinho, H., Santo, A.M.E., Bruno, P.M., and Martin, A.A. (2007) *Anal. Bioanal. Chem.*, **387**, 1643–48.

Baker, R., Matousek, P., Ronayne, K.L., Parker, A.W., Rogers, K., and Stone, N. (2007) *Analyst*, **132**, 48–53.

Bakker Schut, T.C., Witjes, M.J.H., Sterenborg, H.J.C.M., Speelman, O.C., Roodenburg, J.L.N., Marple, E.T., Bruining, H.A., and Puppels, G.J. (2000) *Anal. Chem.*, **72**, 6010–18.

Barbosa, C.J., Vaillancourt, F.H., Eltis, L.D., Blades, M.W., and Turner, R.F.B. (2002) *J. Raman Spectrosc.*, **33**, 503–10.

Bernstein, P.S., Zhao, D.Y., Sharifzadeh, M., Ermakov, I.V., and Gellermann, W. (2004) *Arch. Biochem. Biophys.*, **430**, 163–69.

Boere, I.A., Bakker Schut, T.C., van den Boogert, J., de Bruin, R.W.F., and Puppels, G.J. (2003) *Vib. Spectrosc.*, **32**, 47–55.

Buschmann, H.P., Marple, E.T., Wach, M.L., Bennet, B., Bakker Schut, T.C., Bruining, H.A., Bruschke, A.V., van der Laarse, A., and Puppels, G.J. (2000) *Anal. Chem.*, **72**, 3771–75.

Carden, A. and Morris, M.D. (2000) *J. Biomed. Opt.*, **5**, 259–68.

Carden, A., Rajachar, R.M., Morris, M.D., and Kohn, D.H. (2003) *Calcif. Tissue Int.*, **72**, 166–75.

Caspers, P.J., Lucassen, G.W., Bruining, H.A., and Puppels, G.J. (2000) *J. Raman Spectrosc.*, **31**, 813–18.

Caspers, P.J., Lucassen, G.W., Carter, E.A., Bruining, H.A., and Puppels, G.J. (2001) *J. Invest. Dermatol.*, **116**, 434–42.

Caspers, P.J., Lucassen, G.W., and Puppels, G.J. (2003) *Biophys. J.*, **85**, 572–80.

Chen, K.H., Cheng, W.T., Li, M.J., Yang, D.M., and Lin, S.Y. (2005) *J. Microsc.*, **219**, 36–41.

Conti, C., Giorgini, E., Landi, L., Putignano, A., and Tosi, G. (2005) *J. Mol. Struct.*, **744–747**, 641–46.

Crane, N.J., Morris, M.D., Ignelzi, M.A., and Yu, G. (2005) *J. Biomed. Opt.*, **10**, 031119.

Crow, P., Molckovsky, A., Stone, N., Uff, B., Wilson, B.C., and Wong Kee Song, L.M. (2005) *Urology*, **65**, 1126–30.

de Carmejane, O., Morris, M.D., Davis, M.K., Stixrude, L., Tecklenburg, M., Rajachar, R.M., and Kohan, D.H. (2005) *Calcif. Tissue Int.*, **76**, 207–13.

de Jong, B.W.D., Bakker Schut, T.C., Maquelin, K., van der Kwast, T., Bangma, C.H., Kok, D.-J., and Puppels, G.J. (2006) *Anal. Chem.*, **78**, 7761–69.

de Jong, B.W., de Gouveia Brazao, C.A., Stoop, H., Wolffenbuttel, K.P., Oosterhuis, J.W., Puppels, G.J., Weber, R.F., Looijenga, L.H., and Kok, D.J. (2004) *J. Urol.*, **171**, 92–96.

Diem, M., Griffiths, P.R., and Chalmers, J.M. (2008) *Vibrational Spectroscopy for Medical Diagnosis*, Wiley&Sons, Chichester.

Eikje, N.S., Aizawa, K., and Ozaki, Y. (2005) *Biotechnol. Annu. Rev.*, **11**, 191–225.

Evans, C.L., Potma, E.O., Pouris'haag, M., Cote, D., Lin, C.P., and Xie, X.S. (2005) *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 16807–12.

Evans, C.L., Xu, X., Kesari, S., Xie, X.S., Wong, S.T.C., and Young, G.S. (2007) *Opt. Express*, **15**, 12076–87.

Fattibene, P., Carosi, A., de Coste, V., Sacchetti, A., Nucara, A., Postorino, P., and Dore, P. (2005) *Phys. Med. Biol.*, **50**, 1095–1108.

Frosch, T., Koncarevic, S., Zedler, L., Schmitt, M., Schenzel, K., Becker, K., and Popp, J. (2007a) *J. Phys. Chem. B*, **111**, 11047–56.

Frosch, T., Schmitt, M., Bringmann, G., Kiefer, W., andPopp, J.(2007b) *J. Phys. Chem. B*, **111**, 1815–22.

Frosch, T., Kuestner, B., Schlücker, S., Szeghalmi, A., Schmitt, M., Kiefer, W., and Popp, J. (2004) *J. Raman Spectrosc.*, **35**, 819–21.

Furic, K., Mohacek-Grosev, V., and Hadzija, M. (2005) *J. Mol. Struct.*, **744–747**, 169–77.

Gellermann, W., Ermakov,I.V., McClane, R.W., and Bernstein, P.S. (2002) *Opt. Lett.*, **27**, 833–35.

Gniadecka, M., Philipsen, P.A., Sigurdsson, S., Wessel, S., Nielsen, O.F., Christensen, D.H., Hercogova, J., Rossen, K., Thomsen, H.K., Gniadecka, R., *et al.* (2004) *J. Invest. Dermatol.*, **122**, 443–49.

Golcuk, K., Mandair, G.S., Callender, A.F., Sahar, N., Kohn, D.H., and Morris, M.D. (2006) *Biochim. Biophys. Acta*, **1758**, 868–73.

Gosselin, M.-E., Kapustij,C.J., Venkateswaran, U.D., Leverenz, V.R., and Giblin, F.J. (2007) *Exp. Eye Res.*, **84**, 493–99.

Goujon, D., Zellweger, M., Radu, A., Grosjean, P., Weber, B.C., van den Bergh, H., Monnier, P., and Wagnieres, G. (2003) *J. Biomed. Opt.*, **8**, 17–25.

Haka, A.S., Shafer-Peltier, K.E., Fitzmaurice, M., Crowe, J., Dasari, R.R., and Feld, M.S. (2002) *Cancer Res.*, **62**, 5375–80.

Haka, A.S., Shafer-Peltier, K.E., Fitzmaurice, M., Crowe, J., Dasari, R.R., and Feld, M.S. (2005) *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 12371–76.

Haka, A.S., Volynskaya, Z., Gardecki, J.A., Nazemi, J., Lyons, J., Hicks, D., Fitzmaurice, M., Dasari, R.R., Crowe, J.P., and Feld, M.S. (2006) *Cancer Res.*, **66**, 3317–22.

Harz, M., Claus, R.A., Bockmeyer, C.L., Baum, M., Roesch, P., Kentouche, K., Deigner, H.P., and Popp, J. (2006) *Biopolymers*, **82**, 317–24.

Hata,T.R., Scholtz,T.A., Ermakov,I.V., McLane,R.W., Khachik, F., and Gellermann, W. (2000) *J. Invest. Dermatol.*, **115**, 441–48.

Hattori, Y., Komachi, Y., Asakura, T., Shimosegawa, T., Kanai, G.I., Tashiro, H., and Sato, H. (2007) *Appl. Spectrosc.*, **61**, 579–84.

Huang, Z., McWilliams, A., Lui, H., McLean, D.I., Lam, S., and Zeng, H. (2003) *Int. J. Cancer*, **107**, 1047–52.

Huang, Z., Zeng, H., Hamzavi, I., McLean, D., and Lui, H. (2001) *Opt. Lett.*, **26**, 1782–84.

Kaminaka, S., Ito, T., Yamazaki, H., Kohda, E., and Hamaguchi, H. (2002) *J. Raman Spectrosc.*, **33**, 498–502.

Kendall, C., Stone, N., Shepherd, N., Geboes, K., Warren, B., Bennett, R., and Barr, H. (2003) *J. Pathol.*, **200**, 602–09.

Kleihues, P. and Cavenee, W.K. (2000) *Pathology and Genetics of Tumors of the Nervous System*, IARC Press, Lyon.

Knudsen, L., Johansson, C.K., Philipsen, P.A., Gniadecka, M., and Wulf, H.C. (2002) *J. Raman Spectrosc.*, **33**, 574–79.

Ko, A.C., Choo-Smith, L.P., Hewko, M., Leonardi, L., Sowa,M.G., Dong,C.C., Williams, P., and Cleghorn, B. (2005) *J. Biomed. Opt.*, **10**, 031118.

Kodati, V.R., Tomasi, G.E., Turumin, J.L., and Tu, A.T. (1991) *Appl. Spectrosc.*, **45**, 581–83.

Koljenovic, S., Bakker Schut, T.C., van Meerbeek, J.P., Maat, A.P., Burgers, S.A., Zondervan, P.E., Kros, J.M., and Puppels, G.J. (2004) *J. Biomed. Opt.*, **9**, 1187–97.

Koljenovic, S., Bakker Schut, T.C., Vincent, A., Kros, J.M., and Puppels, G.J. (2005a) *Anal. Chem.*, **77**, 7958–65.

Koljenovic, S., Bakker Schut, T.C., Wolthuis, R., de Jong, B., Santos, L.F., Caspers, J.M., Kros, J.M., and Puppels, G.J. (2005b) *J. Biomed. Opt.*, **10**, 031116.

Koljenovic, S., Bakker Schut,T.C., Wolthuis, R., Vincent, A.J.P.E., Hendriks-Hagevi, G., Santos, L.F., Kros, J.M., and Puppels, G.J. (2007) *Anal. Chem.*, **79**, 557–64.

Koljenovic, S., Choo-Smith, L.P., Bakker Schut, T.C., Kros, J.M., van den Bergh, H., and Puppels, G.J. (2002) *Lab. Invest.*, **82**, 1265–77.

Komachi, Y., Sato, H., Aizawa, K., and Tashiro H. (2005a) *Appl. Opt.*, **44**, 4722–32.

Komachi, Y., Sato, H., Matsuura, Y., Miyagi, M., and Tashiro, H. (2005b) *Opt. Lett.*, **30**, 2942–44.

Komachi, Y., Sato, H., and Tashiro, H. (2006) *Appl. Opt.*, **45**, 7938–43.

Konorov, S.O., Addison, C.J., Schulze, R.F., Turner, R.F., and Blades, M.W. (2006) *Opt. Lett.*, **31**, 1911–13.

Koutsopoulos, S. (2002) *J. Biomed. Mater. Res.*, **62**, 600–12.

Koyama, Y. (1995) Resonance Raman spectroscopy, in *Carotenoids: Spectroscopy* (eds G. Britton, S. Liaaen-Jensen, and H. Pfander), Birkhäuser, Basel, pp. 135–46.

Krafft, C., Codrich, D., Pelizzo, G., and Sergo, V. (2008a) *J. Biophoton.*, **1**, 154–69.

Krafft, C., Codrich, D., Pelizzo, G., and Sergo, V. (2008b) *Analyst*, **133**, 361–71.

Krafft, C., Kirsch, M., Beleites, C., Schackert, G., and Salzer, R. (2007) *Anal. Bioanal. Chem.*, **389**, 1133–42.

Krafft, C., Neudert, L., Simat, T., and Salzer, R. (2005a) *Spectrochim. Acta A*, **61**, 1529–35.

Krafft, C., Sobottka, S.B., Schackert, G., and Salzer, R. (2005b) *Analyst*, **130**, 1070–77.

Krafft, C. and Salzer, R. (2008) Neurooncological applications if infrared and Raman spectroscopy, in *Vibrational Spectroscopy for Medical Diagnosis* (eds M. Diem, J.M. Chalmers, and P.R. Griffiths), Chichester, Jon Wiley & Sons Ltd., pp. 231–59.

Krafft, C. and Sergo, V. (2006) *Spectroscopy*, **20**, 195–218.

Krafft, C., Sobottka, S.B., Schackert, G., and Salzer, R. (2006) *J. Raman Spectrosc.*, **37**, 367–75.

Lambert, J.L., Pelletier, C.C., and Borchert, M. (2005) *J. Biomed. Opt.*, **10**, 031110–031111–031110–031118.

Lasch, P., Haensch, W., Naumann, D., and Diem, M. (2004) *Biochim. Biophys. Acta*, **1688**, 176–86.

Liang, C., Morris, A., Schluecker, S., Imoto, K., Price, V.H., Menefee, E., Wincovitch, S.M., Levin, I.W., Tamura, D., and Strehle, K.R., *et al.* (2006) *J. Invest. Dermatol.*, **126**, 2210–16.

Lieber, C.A. and Mahadevan-Jansen, A. (2007) *Opt. Express*, **15**, 11874–82.

Lopes, C.B., Pinheiro, A.L.B., Sathaiah, S., Da Silva, N.S., and Salgado, M.A.C. (2007) *Photomed. Laser Surg.*, **25**, 96–101.

Lyandres, O., Shah,N.C., Yonzon,C.R., Walsh, J.T., Glucksberg, M.R., and Van Duyne, R.P. Jr. (2005) *Anal. Chem.*, **77**, 6134–39.

Mahadevan-Jansen, A., Mitchell, M.F., Ramanujam, N., Utzinger, U., and Richards-Kortum, R. (1998) *Photochem. Photobiol.*, **68**, 427–31.

Matousek, P., Draper, E.R., Goodship, A.E., Clark, I.P., Ronayne, K.L., and Parker, A.W. (2006) *Appl. Spectrosc.*, **60**, 758–63.

Miyazaki, M., Sato, H., Onose, H., Moore,B.K., and Platt Jeffery, A. (2003) *Oper. Dent.*, **28**, 136–42.

Mizuno, A., Kitajima, H., Kawauchi, K., Muraishi, S., and Ozaki, Y. (1994) *J. Raman Spectrosc.*, **25**, 25–29.

Molckovsky, A., Wong Kee Song, L.M., Shim, M.G., Marcon, N.E., and Wilson, B.C. (2003) *Gastrointest. Endosc.*, **57**, 396–402.

Motz, J.T., Gandhi, S.J., Scepanovic, O.R., Haka, A.S., Kramer, J.R., Dasari, R.R., and Feld, M.S. (2005) *J. Biomed. Opt.*, **10**, 031113–031111–031113–031117.

Motz,J.T., Hunter, M., Galindo,L.H., Gardecki, J.A., Kramer, J.R., Dasari, R.R., and Feld, M.S. (2004) *Appl. Opt.*, **43**, 542–54.

Nguyen, Q.D. and Daudon, M. (1997) *Infrared and Raman Spectra of Calculi*, Elsevier, Paris.

Nie, S. and Emory, S.R. (1997) *Science*, **275**, 1102–06.

Nijssen, A., Bakker Schut, T.C., Heule, F., Caspers, P.J., Hayes, D.P., Neumann, M. H.A., and Puppels, G.J. (2002) *J. Invest. Dermatol.*, **119**, 443–49.

Ong, C.W., Shen, Z.X., Ang, K.K.H., Kara, U.A.K., and Tang, S.H. (2002) *Appl. Spectrosc.*, **56**, 1126–31.

Osada, M., Gniadecka, M., Brandrup, F., Hansen, L.K., and Wulf, H.C. (2003) *Br. J. Dermatol.*, **148**, 600–1.

Palchik, N.A. and Moroz, T.N. (2005) *J. Cryst. Growth*, **283**, 450–56.

Paluszczewicz, W.M., Kwiatek, W.M., Galka, M., Sobieraj, D., and Wentrup-Byrne, E. (1998) *Cell. Mol. Biol.*, **44**, 65–71.

Pelletier, C.C., Lambert, J.L., and Borchert, M. (2005) *Appl. Spectrosc.*, **59**, 1024–31.

Pestaner, J.P., Mullick, F.G., Johnson, F.B., and Centeno, J.A. (1996) *Arch. Pathol. Lab. Med.*, **120**, 537–40.

Prescott, B., Steinmetz, W., and Thomas, Jr. G.J. (1984) Characterization of DNA structures by laser Raman spectroscopy. *Biopolymers*, **23**, 235–56.

Polyzoidis, K.S., Miliaras, G., and Miliaras Pavlidis, N. (2005) *Cancer Treat. Rev.*, **31**, 247–55.

Prieto, M.C.H., Matousek, P., Towrie, M., Parker, A.W., Wright, M., Ritchie, A.W., and Stone, N. (2005) *J. Biomed. Opt.*, **10**, 044006–044001–044006–044006.

Puskar, L., Tuckermann, R., Frosch, T., Popp, J., Ly, V., McNaughton, D., and Wood, B.R. (2007) *Lab Chip*, **7**, 1125–31.

Qi, D. and Berger, A.J. (2007) *Appl. Opt.*, **46**, 1726–34.

Rohleder, D., Kiefer, W., and Petrich, W. (2004) *Analyst*, **129**, 906–11.

Rohleder, D., Kocherscheidt, G., Gerber, K., Kiefer, W., Kohler, W., Mocks, J., and Petrich, W. (2005) *J. Biomed. Opt.*, **10**, 031108.

Römer, T.J., Brennan, J.F., Schut, T.C., Wolthuis, R., van den Hoogen, R.C., Emeis, J.J., Van der Laarse, A., Bruschke, A.V., and Puppels, G.J. (1998) *Atherosclerosis*, **141**, 117–24.

Salzer, R. and Siesler, H.W. (2009) *Infrared and Raman Spectroscopic Imaging*, Wiley-VCH, Weinheim.

Santos, L.F., Wolthuis, R., Koljenovic, S., Almeida, R.M., and Puppels, G.J. (2005) *Anal. Chem.*, **77**, 6747–52.

Schlücker, S., Liang, C., Strehle, K.R., DiGiovanna, J.J., Kraemer, K.H., and Levin, I.W. (2006) *Biopolymers*, **82**, 615–22.

Schulmerich, M.V., Finney, W.F., Fredricks, R.A., and Morris, M.D. (2006) *Appl. Spectrosc.*, **60**, 109–14.

Sebag, J. (2004) *J. Biomed. Opt.*, **9**, 38–46.

Shafer-Peltier, K.E., Haka, A.S., Fitzmaurice, M., Crowe, J., Myles, J., Dasari, R.R., and Feld, M.S. (2002) *J. Raman Spectrosc.*, **33**, 552–63.

Shetty, G., Kendall, C., Shephard, N., Stone, N., and Barr, H. (2006) *Br. J. Cancer*, **94**, 1460–64.

Shim, M.G., Wilson, B.C., Marple, E., and Wach, M. (1999) *Appl. Spectrosc.*, **53**, 619–27.

Shim, M.G., Wong Kee Song, L.M., Marcon, N.E., and Wilson, B.C. (2000) *Photochem. Photobiol.*, **72**, 146–50.

Short, M.A., Lui, H., McLean, D.I., Zeng, H., and Chen, M.X. (2006) *Proc. SPIE*, **6093**, 60930E–60931–60930E–60936.

Sideroudi, T., Pharmakakis, N., Tyrovolas, A., Papatheodorou, G., Chryssikos, G.D., and Voyiatzis, G.A. (2007) *J. Biomed. Opt.*, **12**, 034005–034001–034005–034006.

Siebert, F. and Hildebrandt, P. (2007) *Vibrational Spectroscopy in Life Science (Tutorials in Biophysics)*, Wiley-VCH, Weinheim.

Sijtsema, N.M., Duindam, J.J., Puppels, G.J., Otto, C., and Greve, J. (1996) *Appl. Spectrosc.*, **50**, 545–51.

Smith, J., Kendall, C., Sammon, A., Christie-Brown, J., and Stone, N. (2003) *Technol. Cancer Res. Treat.*, **2**, 327–32.

Solomonov, I., Osipova, M., Feldman, Y., Baehtz, C., Kjaer, K., Robinson, I.K., Webster, G.T., McNaughton, D., Wood, B.R., Weissbuch, I., and Leiserowitz, L. (2007) *J. Am. Chem. Soc.*, **129**, 2615–27.

Stone, N., Kendall, C., and Barr, H. (2008) Raman spectroscopy as a potential tool for early diagnosis of malignancies in esophageal and bladder tissues, in *Vibrational Spectroscopy for Medical Diagnosis* (eds M. Diem, P.R. Griffiths, and J.M. Chalmers), John Wiley & Sons Ltd., Chichester, pp. 203–30.

Stone, N., Kendall, C., Shepherd, N., Crow, P., and Barr, H. (2002) *J. Raman Spectrosc.*, **33**, 564–73.

Stuart, D.A., Yonzon, C.R., Zhang, X., Lyandres, O., Shah, N.C., Glucksberg, M.R., Walsh, J.T., and Van Duyne, R.P. (2005) *Anal. Chem.*, **77**, 4013–19.

Stummer, W., Pichlmeier, U., Meinel, T., Wiestler, O.D., Zanella, F., and Reulen, H.J. (2006) *Lancet Oncol.*, **7**, 392–401.

Takasaki, E. (1996) *Int. J. Urol.*, **3**, 27–30.

Tarnowski, C.P., Ignelzi, M.A., Wang, W., Taboas, J.M., Goldstein, S.A., and Morris, M.D. (2004) *J. Bone Miner. Res.*, **19**, 64–71.

Timlin, J.A., Carden, A., Morris, M.D., Bonadio, J.F., Hoffler, C.E., Kozloff, K., and Goldstein, S.A. (1999) *J. Biomed. Opt.*, **4**, 28–34.

Timlin, J.A., Carden, A., Morris, M.D., Rajachar, R.M., and Kohn, D.H. (2000) *Anal. Chem.*, **72**, 2229–36.

Toms, S.A., Lin, W.C., Weil, R.J., Johnson, M.D., Jansen, E.D., and Mahadevan-Jansen, A. (2005) *Neurosurgery*, **57**, 382–91.

Truffier-Boutry, D., Demoustier-Champagne, S., Devaux, J., Biebuyck, J.-J., Mestdagh, M., Larbanois, P., and Leloup, G. (2006) *Dent. Mater.*, **22**, 405–12.

Utzinger, U., Heintzelmann, D.L., Mahadevan-Jansen, A., Malpica, A., Follen, M., and Richards-Kortum, R. (2001) *Appl. Spectrosc.*, **55**, 955–59.

van de Poll, S.W.E., Bakker Schut, T.C., Van der Laarse, A., and Puppels, G.J. (2002a) *J. Raman Spectrosc.*, **33**, 544–51.

van de Poll, S.W.E., Delsing, D.J.M., Jukema, J.W., Princen, H.M.G., Havekes, L.M., Puppels, G.J., and Van der Laarse, A. (2002b) *Atherosclerosis*, **164**, 65–71.

van de Poll, S.W.E., Romer, T.J., Puppels, G.J., and van der Laarse, A. (2002c) *J. Cardiovasc. Risk*, **9**, 255–61.

van de Poll, S.W.E., Kastelijn, K., Bakker Schut, T.C., Strijder, C., Pasterkamp, G., Puppels, G.J., and van der Laarse, A. (2003) *Heart*, **89**, 1078–82.

Vitetta, L., Best, S.P., and Sali, A. (2000) *Med. Hypotheses*, **55**, 502–6.

Wang, Y., Spencer, P., and Walker, M.P. (2007) *J. Biomed. Mater. Res. A*, **81A**, 279–86.

Wang, Y., Spencer, P., and Yao, X. (2006) *J. Biomed. Opt.*, **11**, 024005–024001 –024005–024007.

Webster, G.T., Tilley, L., Deed, S., McNaughton, D., and Wood, B.R. (2008) *FEBS Lett.*, **582**, 1087–92.

Widjaja, E. and Seah Regina Kim, H. (2006) *Appl. Spectrosc.*, **60**, 343–45.

Wood, B.R., Langford, S.J., Cooke, B.M., Glenister, F.K., Lim, J., and McNaughton, D. (2003) *FEBS Lett.*, **554**, 247–52.

Wood, B.R. and McNaughton, D. (2006) *Expert Rev. Proteomics*, **3**, 525–44.

Wood, B.R. and McNaughton, D. (2007) *Anal. Bioanal. Chem.*, **387**, 1691–1703.

Wu, Z.-S., Zhou, G.-Z., Jiang, J.-H., Shen, G.-L., and Yu, R.-Q. (2006) *Talanta*, **70**, 533–39.

Xiao, C., Flach, C.R., Marcott, C., and Mendelsohn, R. (2004) *Appl. Spectrosc.*, **58**, 382–89.

Xiao, C., Moore, D.J., Flach, C.R., and Mendelsohn, R. (2005a) *Vib. Spectrosc.*, **38**, 151–58.

Xiao, C., Moore, D.J., Rerek, M.E., Flach, C.R., and Mendelsohn, R. (2005b) *J. Invest. Dermatol.*, **124**, 622–32.

Yonzon, C.R., Haynes, C.L., Zhang, X., Walsh, J.T. Jr., and Van Duyne, R.P. (2004) *Anal. Chem.*, **76**, 78–85.

Zhang, G., Moore, D.J., Mendelsohn, R., and Flach, C.R. (2006) *J. Invest. Dermatol.*, **126**, 1088–94.

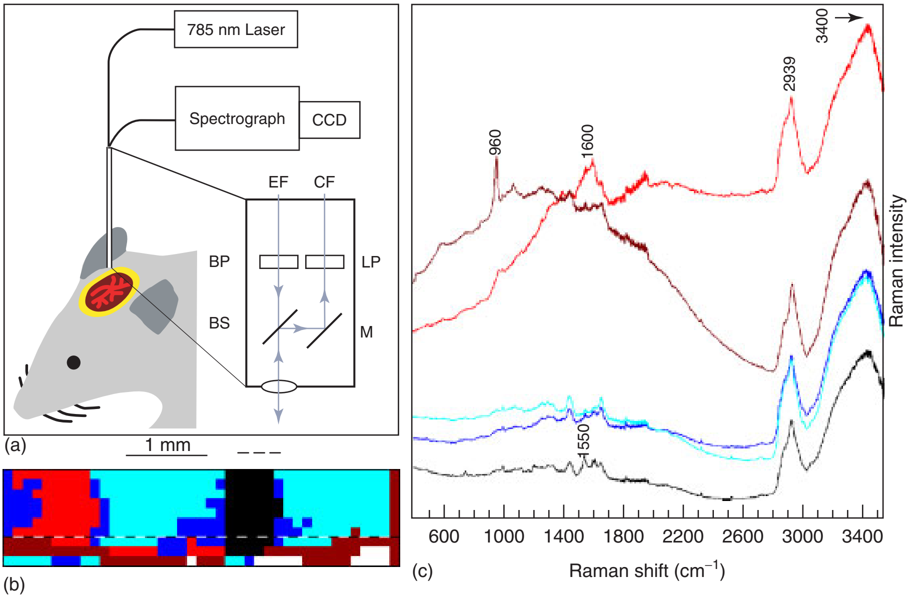

Fig. 28.3 (a) Schematic diagram of the Raman instrumentation: Radiation from a 785 nm laser is coupled into a fiber-optic probe. The scattered radiation is guided to the spectrograph and registered by a CCD detector. The filtered fiber-optic probe consists of an excitation fiber (EF), band pass (BP) filter, beam splitter (BS), lens (L), mirror (M), long pass (LP) filter, and collection fiber (CF). (b) Cluster analysis of 41 × 10 Raman map. (c) Unprocessed Raman spectra from 400 to 3550 cm^{-1} representing the average of each cluster. Color code: normal brain tissue (blue/cyan), oxygenated blood (black), bone (brown), and brain metastasis of malignant melanoma (red). (This figure also appears on page 998.)

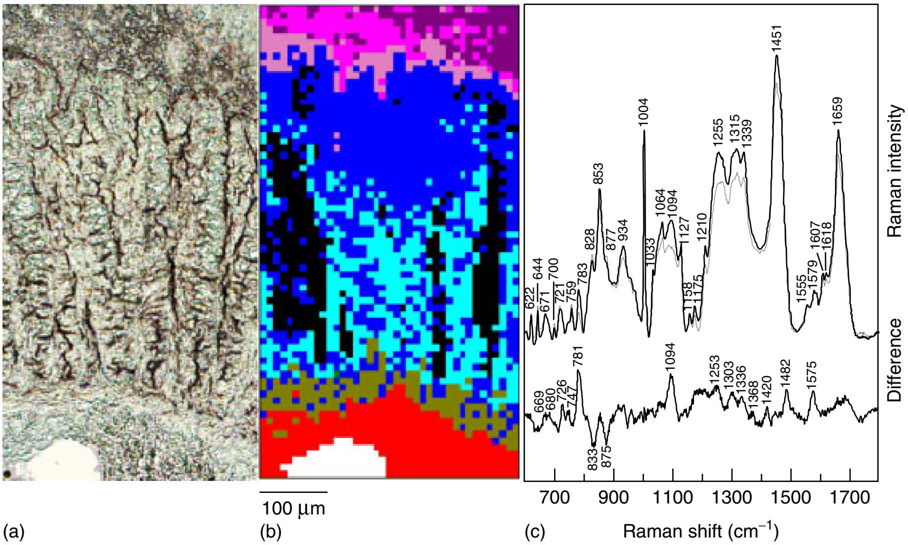

Fig. 28.4 (a) Photomicrograph and (b) cluster analysis of Raman map of mucosa in a colon tissue section. Raman spectra from 600 to 1800 cm^{-1} representing epithelial tissue (cyan cluster, black trace), glandular tissue (black cluster, gray trace), and difference spectrum = epithelial − glandular tissue. (This figure also appears on page 999.)

3
Mössbauer Spectroscopy

De-Ping Yang

3.1 **Introduction 53**
3.2 **The Mössbauer Effect 53**
3.3 **The Recoil-Free Fraction 56**
3.4 **Sources of Mössbauer Radiation 58**
3.5 **The Mössbauer Spectrometer 60**
3.6 **The Shape and Intensity of a Spectral Line 62**
3.7 **Hyperfine Interactions 64**
3.7.1 Isomer Shift 64
3.7.2 Magnetic Dipole Interaction 65
3.7.3 Electric Quadrupole Interaction 68
3.7.4 Combined Electric Quadrupole and Magnetic Dipole Interactions 71
3.7.5 Saturation Effect in the Presence of Hyperfine Splittings 72
3.8 **Second-Order Doppler Effect 73**
3.9 **Applications of Mössbauer spectroscopy 73**
3.10 **Mössbauer Spectroscopy Using Synchrotron Radiation 76**
3.10.1 Monochromatization of Synchrotron Radiation 77
3.10.2 Time-Domain Mössbauer Spectroscopy 78
3.10.3 Applications of Time-Domain Mössbauer Spectroscopy 81
3.11 **Synchrotron Methods versus Conventional Methods 82**
References 83
Further Reading 84

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

3.1
Introduction

Mössbauer spectroscopy is a unique technique based on the nuclear resonance via recoil-free emission, absorption, and scattering of γ-rays. Because of its extremely high energy resolution and its particular timescale, a Mössbauer spectrum can provide valuable information on structural, chemical, magnetic, and dynamic properties of a variety of solid materials. Mössbauer spectroscopy finds applications in various fields of scientific research, such as physics, metallurgy, materials science, chemistry, mineralogy, biology, medicine, environmental science, archeology, and art. In this chapter, the theoretical principles, experimental techniques, and various applications of Mössbauer spectroscopy are briefly described. For rigorous theoretical treatment and complete experimental details, the reader is referred to specialized books and review articles listed in the Further Reading section at the end of this chapter. There are more than 50 000 journal articles reporting research using Mössbauer spectroscopy in the literature. They have been systematically collected by the Mössbauer Effect Data Center at the University of North Carolina at Asheville and compiled in the Center's publications *The Mössbauer Effect Data Index* (1966–77) and *The Mössbauer Effect Reference and Data Journal* (since 1978).

3.2
The Mössbauer Effect

The principle of Mössbauer spectroscopy is based on the recoil-free nuclear γ-resonance, known as the *Mössbauer effect*, named after its discoverer Rudolf L. Mössbauer. γ-Rays are photons emitted from certain nuclei when they make a transition from an excited state to a lower state. These γ-rays may be absorbed by other nuclei of the same isotope in their lower state, making a transition to the same excited state, much like any other resonance absorption phenomenon. This lower state is usually the stable ground state, whereas the excited state has a typical lifetime of $\tau \approx 100$ nanoseconds. According to the uncertainty principle $(\Delta E)(\Delta t) \sim \hbar$, where the Planck constant $\hbar \approx 10^{-15}$ eV s, this lifetime corresponds to a natural linewidth of $\Gamma \approx 10^{-8}$ eV. Since the γ-photon energies are typically in the order of $E_\gamma \approx 10^4$ eV, this linewidth $(\Gamma/E_\gamma \approx 10^{-12})$ is very narrow and characterizes an extremely "monochromatic" nature of

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

γ-rays emitted from a particular nuclear isotope. We are fortunate to have access to nuclear γ-rays possessing such a unique quality, but this is a double-edged sword. On the one hand, an extraordinarily high energy resolution can be achieved if γ-rays are used in resonance spectroscopy. On the other hand, a precise resonance condition within 10^{-8} eV must be strictly satisfied between the emitter and the absorber.

During the first half of the twentieth century, considerable amount of research was devoted to finding ways of matching the energies of emitted γ-photons with the energy requirement for nuclear resonant absorption. The main obstacle was recoil of the nucleus, during emission as well as absorption of γ-rays. When a free nucleus emits a photon of energy E_γ (Figure 3.1), the photon carries with it a momentum of

$$p = \frac{E_\gamma}{c} \tag{3.1}$$

where c is the speed of light. According to momentum conservation, the nucleus recoils in the opposite direction with the same moment p, which corresponds to a recoil kinetic energy of

$$E_R = \frac{p^2}{2M} \tag{3.2}$$

where M is its mass. Because of this recoil energy, the actual energy of the emitted photon is reduced by E_R, and Eq. (3.2) can be rewritten as

$$E_R = \frac{(E_\gamma/c)^2}{2M} = \frac{E_\gamma^2}{2Mc^2} \tag{3.3}$$

The absorption of a γ-photon by a free nucleus involves exactly the opposite recoil process – the nucleus will gain a momentum of p and a recoil kinetic energy of E_R. Therefore, a photon of merely an energy of E_γ does not suffice to cause a nuclear transition, but an excess energy of E_R is needed to compensate the recoil energy imparting the motion to the nucleus.

Since the rest mass M_ec^2 of an electron is about 0.5 MeV and a typical nucleus is about 10^5 times more massive than the electron, the nuclear rest mass Mc^2 is of the order of 10^{11} eV. Using $E_\gamma \approx 10^4$ eV, we estimate the recoil energy

$$E_R \approx \frac{(10^4)^2}{10^{11}} = 10^{-3}\ \text{eV} \tag{3.4}$$

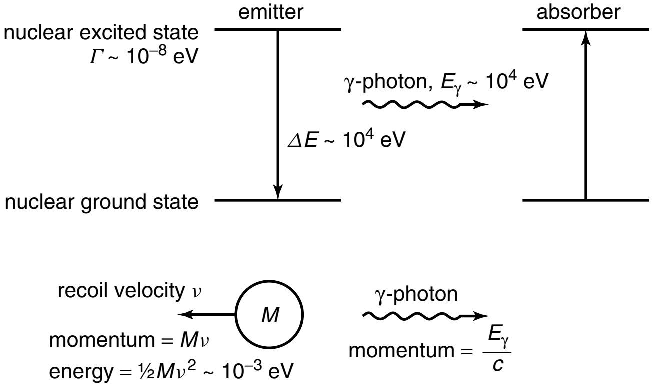

Fig. 3.1 Emission of a γ-photon and recoil of the nucleus.

This is about 100 000 times larger than the natural linewidth Γ of the photon energy; therefore, recoil of a free nucleus shifts the photon energy far beyond its "bandwidth" and completely destroys the resonance condition.

Various methods, mostly based on the Doppler effect, were used to compensate the recoil energy by raising the emitted photon energy so that the emission and absorption lines overlapped, at least partially. For example, it was well known that the Doppler broadening associated with thermal motion of the nuclei could add energy to photons emitted from those fast-moving nuclei that had the same direction as the γ-rays. Alternatively, macroscopic mechanical motion ($\sim 10^5$ cm s^{-1}) of the emitting nuclei (i.e., a source) toward the absorber could shift the photon energy to meet the resonance condition. In certain situations, a preceding nuclear reaction may result in a fast-moving nucleus so that the Doppler shift in the emitted photon could become large enough to cause resonance absorption. These techniques did allow the observation of nuclear resonance, but the recoil energy was merely compensated, often at the expense of the energy resolution and the loss of γ-ray intensity.

Mössbauer (1958) discovered a distinctively new phenomenon that, when the emitting nuclei are bound in a solid, a fraction of the emitted γ-rays have exactly the energy E_γ, corresponding to a recoil-free process. This process seemingly violates the conservation of momentum when we consider one individual emission. However, we should realize that the recoil momentum and energy can no longer be transferred to the emitting nucleus alone, but to a large number of tightly bound atoms within the entire solid, in the form of their simultaneous vibrations (i.e., creating phonons). While using γ-rays from ^{191}Ir and comparing their resonant scatterings by Ir and Pt, Mössbauer experimentally observed higher scattering intensities by Ir at lower temperatures, which contradicted the theory based on thermal broadening of the γ-rays. He then successfully observed higher resonant absorption of the 129 keV radiation from Ir at a temperature of 88 K and correctly attributed the phenomenon to the zero-phonon process (Mössbauer, 1958, 2000). Because the typical phonon frequency is in the order of 10^{13} Hz (corresponding to $\hbar\omega \sim 10^{-2}$ eV), a single recoil ($\sim 10^{-3}$ eV) does not have enough energy to create one phonon (Figure 3.2).

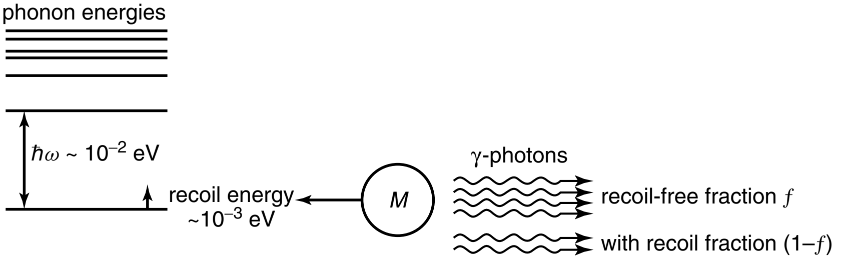

Fig. 3.2 Recoil energy compared with the phonon energy in a typical solid. Because the recoil energy due to the emission of one photon is not large enough to create one phonon, a fraction of photons are emitted in the recoil-free fashion.

Therefore, there is a sizeable probability of recoil-free emission. If all emissions (both recoil-free and with recoil) are considered collectively, momentum and energy conservation laws are completely obeyed.

When these γ-rays, emitted in the recoil-free fashion, are allowed to be absorbed, also in a similar recoil-free way, resonance experiments can be carried out without the need for energy compensation. Now we have essentially two categories of γ-photons; some are emitted, absorbed, or scattered at exactly the energy corresponding to a nuclear transition, and others carry higher or lower energies corresponding to the creation or annihilation of certain phonons. The first group (recoil-free) is wonderfully suitable for resonance experiments with extremely high energy resolution, whereas the second group is suitable for studying lattice dynamics because these γ-rays (with recoil) carry information about phonon distribution in the emitter, absorber, or scatterer. In recognition of his experimental discovery and theoretical interpretation of this recoil-free phenomenon, Mössbauer was awarded the Nobel Prize in physics in 1961 and this phenomenon has since been known as the *Mössbauer effect*.

3.3
The Recoil-Free Fraction

The recoil-free γ-photons constitute a fraction of the total γ-radiation emitted from a particular source. This parameter is known as the *recoil-free fraction* f. The value of f obviously depends on how the recoil energy E_R compares with the phonon energies $\hbar\omega$. Only when $E_R << \hbar\omega$ will f be reasonably large. According to Lipkin's sum rule (Lipkin, 1960), when a large number of absorption events are considered, the average energy transferred to the lattice must be exactly equal to E_R. Let a total of m γ-photons be emitted, among which n are recoil-free and, among the remaining photons, each excites a single phonon (neglecting double phonons). Then

$$mE_R = (m - n)\hbar\omega \tag{3.5}$$

On the basis of the Einstein model of lattice vibrations of a single frequency ω, we arrive at an approximate expression for the recoil-free fraction f:

$$f = \frac{n}{m} = 1 - \frac{E_R}{\hbar\omega} \tag{3.6}$$

A more precise expression for the recoilless fraction is

$$f = e^{-k^2\langle u^2\rangle} \tag{3.7}$$

where $\langle u^2\rangle$ is the mean-square displacement of a nucleus along the direction of the wave vector $\boldsymbol{k}$ of the emitted γ-ray. The wave vector $\boldsymbol{k}$ is related to the photon energy E_γ by $\hbar k = E_\gamma/c$. This expression in Eq. (3.7) points out that to have a relative large f, both the γ-photon energy E_γ and the atomic vibrations $\langle u^2\rangle$ should be small. The first condition of small $\boldsymbol{k}$ can be satisfied by selecting isotopes with low-energy γ-transitions. At present, Mössbauer effect has been observed from more than 100 nuclear isotopes (e.g., ^{57}Fe, ^{119}Sn, ^{191}Ir, etc.), many of which have transition energies less than 50 keV. The second condition of small $\langle u^2\rangle$ requires a rigidly bound solid (characterized by a high Einstein or Debye temperature), perhaps cooled to a low temperature (small vibration amplitudes). This implies that in a liquid or a gas, Mössbauer effect is extremely difficult to observe because of the large $\langle u^2\rangle$ values. For any γ-ray energy higher than 100 keV, the source and the absorber are usually kept at low temperatures to reduce their $\langle u^2\rangle$ values.

The mean-square displacement $\langle u^2\rangle$ is intimately related to lattice dynamics, that is, how the collective atomic vibrations take place at a given temperature. We adopt a particular model for lattice dynamics, for example, the Debye model with a phonon density of states (DOS) $g(\omega) = 3\omega^2/{\omega_D}^3$, where ω_D is related to the parameter known as the *Debye temperature* θ_D through the Planck and Boltzmann constants, $\hbar\omega_D = k_B\theta_D$. Applying the Bose–Einstein distribution of phonons among the available phonon states at temperature T, we arrive at a more specific expression for the recoil-free fraction,

$$f = \exp\left[-k^2\frac{3\hbar^2}{Mk_B\theta_D}\left(\frac{1}{4}+\frac{T^2}{\theta_D^2}\int_0^{\theta_D/T} \times \frac{xdx}{(e^x-1)}\right)\right] \tag{3.8}$$

where $x = (\hbar\omega)/(k_BT)$ is a dimensionless variable for the definite integral, which covers the range from 0 to θ_D/T. In the high-temperature limit ($\theta_D/T \ll 1$), x only takes small values, $e^x \approx 1 + x$, and the above expression simplifies to

$$\begin{aligned} f &\approx \exp\left[-k^2\frac{3\hbar^2}{Mk_B\theta_D}\left(\frac{1}{4}+\frac{T}{\theta_D}\right)\right] \\ &\approx \exp\left[-k^2\frac{3\hbar^2 T}{Mk_B\theta_D^2}\right] \end{aligned} \tag{3.9}$$

In the low-temperature limit ($\theta_D/T \gg 1$), the integral has a value of $\pi^2/6$, and the above express simplifies to

$$\begin{aligned} f &\approx \exp\left[-k^2\frac{3\hbar^2}{Mk_B\theta_D}\left(\frac{1}{4}+\frac{\pi^2T^2}{6\theta_D^2}\right)\right] \\ &= \exp\left[-k^2\frac{\hbar^2}{2Mk_B\theta_D}\left(\frac{3}{2}+\frac{\pi^2T^2}{\theta_D^2}\right)\right] \end{aligned} \tag{3.10}$$

At absolute zero ($T = 0$ K), it is simply

$$f = \exp\left[-k^2\frac{3\hbar^2}{4Mk_B\theta_D}\right] \tag{3.11}$$

The Debye model assumes isotropic, harmonic vibrations only. When the lattice has lower symmetry, $\langle u^2\rangle$ depends on the atomic vibration direction with respect to the lattice symmetry axes, resulting in an anisotropic behavior of the recoil-free fraction f. This affects the quantitative measurement of Mössbauer effect, which was first observed by Goldanskii in polycrystalline samples (Goldanskii *et al.*, 1962) and was first explained theoretically by Karyagin (Karyagin, 1963). This phenomenon is known as the *Goldanskii–Karyagin* (*G–K*) effect, which can be used to study anisotropy in lattice dynamics. External pressure applied to the solid may also cause a change in the mean-square displacement $\langle u^2\rangle$ and produce additional anisotropy, which will be reflected in a change in the recoil-free fraction. At high temperatures, if the Mössbauer effect still persists, anharmonicity in the atomic vibration may render the Debye model invalid, and the theoretical treatment of f may involve higher order terms and become a complex function of temperature. Even at low temperatures, anharmonic behavior has been detected in several solids.

It is important for resonant absorption experiments that a relatively large

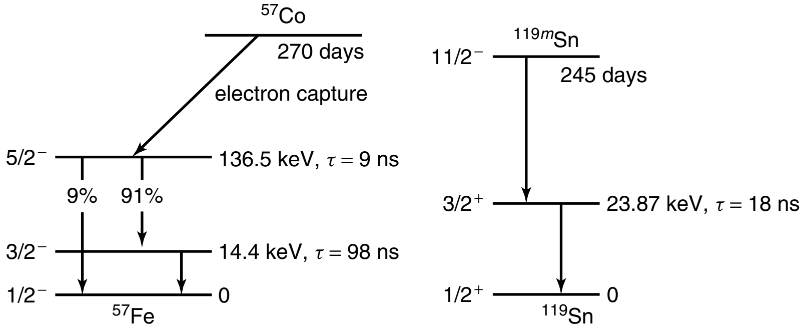

Fig. 3.3 Nuclear decay schemes of ^{57}Co and ^{119m}Sn.

fraction is available and its variation is precisely known (e.g., whether f remains constant when comparing results from a series of samples, or how f changes within the temperature range of the experiments). It is even more important for investigating lattice dynamics using Mössbauer effect because f carries all the information about phonon behavior. Special attention has always been paid to the precise measurement of the recoil-free fraction f. The precision has reached 1% when radioactive sources are used, and it can be better than 0.4% when synchrotron radiation is employed. Very often, we are more interested in how f changes with sample composition, temperature, or pressure than in the absolute value of f.

3.4
Sources of Mössbauer Radiation

Mössbauer effect has been observed in about 100 isotopes. ^{40}K is the lightest isotope, followed by ^{57}Fe. The Mössbauer isotopes are not distributed evenly in the periodic table – about 75% of them are concentrated in elements of atomic numbers between 50 and 80. There are only about 20 Mössbauer isotopes that are in practical use. ^{57}Fe and ^{119}Sn, whose decay schemes are given in Figure 3.3, are the most popular. ^{57}Fe is by far the most important Mössbauer isotope; more than 70% of the research work involves ^{57}Fe. Other frequently used isotopes are ^{151}Eu, ^{197}Au, ^{129}I, ^{121}Sb, and ^{125}Te. Increased attention has been paid to ^{237}Np, ^{155}Gd, ^{161}Dy, and especially to ^{67}Zn and ^{181}Ta, which are used for taking high-resolution spectra.

For ^{57}Fe Mössbauer spectroscopy, the source usually contains the radioactive ^{57}Co, which captures a K-electron and becomes ^{57}Fe at the 136-keV metastable state. It then makes transitions to the ground state, accompanied by γ-ray emissions, among which the 14.4-keV Mössbauer radiation is the most useful. The half-life of the ^{57}Co decay is $T_{1/2} =$ 270 days.

For Mössbauer spectroscopy, the radioactive isotope is embedded in a solid host material (matrix) to form a Mössbauer source. Mössbauer sources are specially prepared via nuclear reactions, and they are commercially available. Sources can also be made by "in-beam" implantation, allowing the use of isotopes with short life times, such as ^{57}Mn ($T_{1/2} = 1.45$ minutes), which decays directly to the 14.4-keV excited state ^{57}Fe (Yoshida *et al.*, 2002).

The quality of a Mössbauer source depends on the properties of the isotope and the selection of the host material. The Mössbauer isotope should satisfy the following criteria:

1. The γ-photon energy E_γ should be in the range 5–150 keV, preferably less than 50 keV. This is because both the recoil-free fraction f and the resonance cross section σ_0 decrease as E_γ increases, especially f, which decreases more rapidly.
2. The desired lifetime τ of the excited state is 1–100 nanoseconds. A long τ corresponds to a narrow natural width Γ_n, and a slight mechanical vibration may destroy the resonance condition. A short τ means that Γ_n may be too large to yield a high-resolution spectrum.
3. The internal conversion coefficient α should be small (<10), to ensure a relatively large probability of available γ-ray emission.
4. It is preferred that the parent nucleus has a long half-life $T_{1/2}$ and allows for easy production of a high activity source.
5. The Mössbauer isotope should not have a high nuclear spin, which would produce complicated spectra and make analysis more difficult.
6. The Mössbauer isotope should have a reasonably large natural abundance. Among the Mössbauer isotopes, ^{57}Fe fulfills the above requirements most satisfactorily, except for its low natural abundance of 2.14%. ^{119}Sn also possesses many good qualities.

A good Mössbauer source requires an appropriate host, a suitable fabrication process, and so on. The following are the main aspects that should be taken into consideration:

1. The radiation from the source should be monochromatic with an energy width as close to the natural width as possible, which requires that the host matrix be a nonmagnetic material of a cubic lattice and of high purity.
2. The host material should have a high Debye temperature to maximize the recoil-free fraction f; metals (and occasionally ionic crystals) of high melting temperatures are usually chosen for the matrix material.
3. The number of stable Mössbauer nuclei in the host material should be minimized to prevent resonant self-absorption.
4. The host material should be made very thin, to reduce the photoelectric effect and Compton scattering caused by the Mössbauer γ-rays.
5. The host should be chemically stable and properly sealed for protection against oxidation, leakage of radioactive material, and so on. For ^{57}Fe, rhodium (Rh) and palladium (Pd) are good hosts, giving an excellent f-value of about 0.78 at room temperature.

In the last two decades, researchers have also been attracted to synchrotron radiation (SR) as a new Mössbauer radiation source. SR provides polarized pulsed radiation of high intensity, high collimation, and narrow beams. The high intensity of the SR source has been utilized for scattering experiments where conventional radiation sources cannot provide adequate results. The pulsed nature of SR radiation is most suitable for measuring time spectra where the entire nuclear ensemble is excited simultaneously and its coherent decay can be observed at different time intervals. One more advantage associated with SR is that γ-rays are naturally polarized, either linearly or circularly, which allow selective absorption and therefore provide additional information. At present, SR Mössbauer radiation can only be produced in large facilities, such as the European Synchrotron Radiation Facility (ESRF) in Grenoble (France), the Advanced Photon Source (APS) in Argonne (USA), and the Super Photon ring (SPring-8) in Hyogo (Japan).

One of the drawbacks of SR radiation is its relatively wide energy distribution. Therefore, it must be collimated and monochromatized before being used in any absorption or scattering experiments. Going through a double-crystal premonochromator Si(111), followed by a high-resolution monochromator based on particular Bragg reflections, the bandwidth can be reduced to the order of milli electronvolts, or even micro electronvolts.

3.5
The Mössbauer Spectrometer

The most common technique for obtaining a Mössbauer spectrum is using a velocity-scanning spectrometer in the transmission geometry, a block diagram of which is shown in Figure 3.4. It consists of a radiation source, an absorber, a detector with its electronic recording system, a wave-form generator, a drive circuit, and a transducer. A beam of γ-rays from a source containing the Mössbauer isotope is directed toward a thin absorber containing the stable nuclei of the same isotope. The transmitted γ-rays are detected by a suitable detector and recorded in a computer system. To obtain a resonance spectrum (i.e., the amount of transmitted γ-rays versus photon energy), the energy of the incoming γ-rays must be modulated

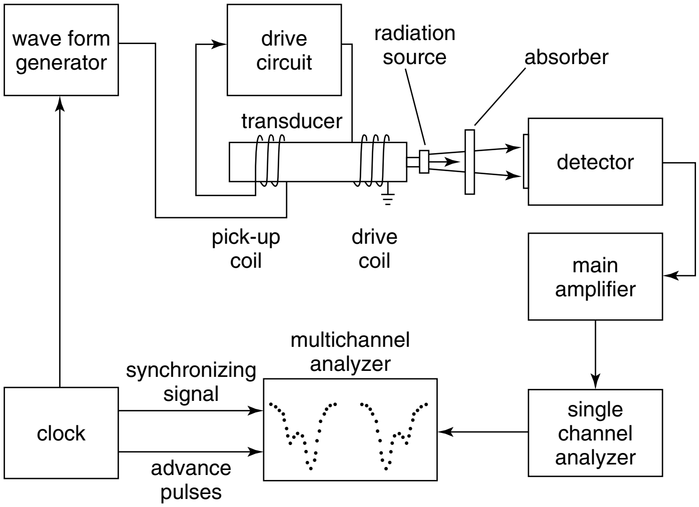
wave form generator
drive circuit
radiation source
absorber
transducer
detector
pick-up coil
drive coil
main amplifier
multichannel analyzer
synchronizing signal
clock
advance pulses
single channel analyzer

Fig. 3.4 Block diagram of a Mössbauer spectrometer in the transmission geometry.

using Doppler effect. In most cases, the source undergoes a mechanical motion with a velocity v toward or away from the absorber, whereas the absorber is at rest, so that it is easier to change its temperature or to apply an external magnetic field to the absorber. The energy of the radiation is therefore Doppler shifted to

$$E = E_0 \left(1 + \frac{v}{c}\right)\left(1 - \frac{v^2}{c^2}\right)^{-1/2} \tag{3.12}$$

where E_0 is the photon energy when the source is at rest. The source velocities that are required to cover the energy range of a typical Mössbauer spectrum are usually less than 1 m s^{-1}, thus $v/c \ll 1$ and a very good approximation of the above equation is

$$E = E_0 \left(1 + \frac{v}{c}\right) \quad \text{or} \quad \Delta E = E_0 \frac{v}{c} \tag{3.13}$$

Because of the energy modulation by source velocity, the energy scale in a Mössbauer spectrum is customarily presented in terms of the source velocity. For example, 1 mm s^{-1} corresponds to an energy shift of 4.80×10^{-8} eV for a ^{57}Fe Mössbauer spectrum ($E_0 =$ 14.4 keV).

The source is driven by a velocity transducer, generally operating in the velocity scan mode. The drive coil converts an applied current to the velocity of a shaft. A pickup coil provides a signal proportional to the actual velocity; a negative feedback circuit ensures accuracy and stability. The radiation source is attached to one end of the shaft, and a prism (or a mirror) for measuring velocity is mounted on the other end. The source thus scans periodically through the velocity range of interest, the simplest way of scanning velocity being constant acceleration.

The most widely employed detectors are gas proportional counters and NaI(Tl) scintillation counters, followed by semiconductor detectors. A detector should have a high efficiency and energy resolution, as well as quick response time. Typically, a gas proportional counter has a cylindrical metal tube (cathode) filled with a mixture of gases and a metal wire on the axis (anode). Operating under a high voltage of 1500–3000 V, the gas proportional counter has a high signal-to-noise ratio and a high count rate. The NaI(Tl) scintillation counters are made of sodium iodine crystals doped with thallium with detection efficiency as high as 100% and are easy to use. The semiconductor detector is essentially a p–n junction, with a reverse-biased high voltage, creating a region that is sensitive to γ-rays. They offer the best resolution over the entire range of energies of interest in Mössbauer experiments, and their efficiency is comparable to proportional or scintillation counters.

In Mössbauer spectroscopy, the absorber is usually the sample to be investigated. In the transmission geometry, the thickness of the absorber should be carefully chosen as it significantly affects the quality of the spectrum. When the sample is too thin, it would contain too few Mössbauer nuclei, resulting in weak spectral intensity, high background, and large statistical uncertainty. When the sample is too thick, significant atomic absorption would occur, causing a decrease in resonant absorption and distortion in the spectral shape. Obviously, there should be an optimal thickness between these two extremes. The effective thickness of the absorber is defined by

$$t_a = n_a f_a \sigma_0 d \tag{3.14}$$

where n_a is the number of Mössbauer nuclei in the absorber per milligram, f_a is the recoil-free fraction of the absorber, σ_0 is the maximum resonance cross section (in centimeter squared), and d is the thickness of the absorber (in milligrams per centimeter squared). For most of the absorption experiments, the effective thickness should be $t_a \approx 1$.

Sample preparation for the absorber is quite straightforward. Metallic or alloy samples are usually wrought or roll-milled into foils of the appropriate thickness. If only small pieces of sample foils are available, they may be arranged to cover the entire sample area with as few gaps and overlaps as possible. Most of the materials are in the powder form, which can be pressed to the desired thickness between two pieces of thin plastic sheets. For low-temperature measurements, the powder sample is usually mixed with a solid chemical of light atoms such as a sugar, and pressed into a "free-standing" sample. For high-temperature measurements, the powder sample is usually placed between two pieces of boron nitride sheets. Liquid samples are usually sealed in the sample holder and refrigerated until frozen. ^{57}Fe has a relatively low natural abundance; if the ^{57}Fe content in the sample is not sufficient to give a satisfactory spectrum, ^{57}Fe enrichment in the sample may be necessary.

The Mössbauer spectrometer may also include an external magnetic field (static or high frequency) to be applied to the absorber or the source. Sample temperature may be varied by using an oven or by using a cryogenic system. Windows for γ-rays to enter and exit the sample chamber should be made from a low-atomic-number element, such as beryllium.

Emission Mössbauer spectra may also be recorded from a transmission spectrometer, where the source is the material under investigation and the absorber is a well-characterized material.

In addition to the transmission geometry, Mössbauer effect can be utilized in the scattering geometry, which does not require a thin sample but probes into the surface of bulk materials. Experiments in the scattering geometry can detect not only the elastically scattered γ-rays but also the accompanying conversion electrons (conversion electron Mössbauer spectroscopy, CEMS), the X rays following the conversion events (conversion X-ray Mössbauer spectroscopy), or the inelastically scattered γ-rays (Mössbauer diffraction).

3.6
The Shape and Intensity of a Spectral Line

Because the nuclear excited state has a certain natural width Γ_s, the emitted γ-rays from a Mössbauer source are not completely monochromatic, but follow the Lorentzian distribution around E_γ (the Breit–Wigner formula),

$$\mathcal{L}(E)dE = \frac{\Gamma_s}{2\pi}\frac{1}{(E - E_\gamma)^2 + \Gamma_s^2/4}dE \tag{3.15}$$

After a resonant absorption, the nuclei in the absorber are in the excited state (with a natural width of Γ_a). They decay to the ground state through internal conversion or γ-ray emission in all directions. The cross section of resonant absorption of γ-rays (as a function of photon energy E)

is also described by the Breit–Wigner formula:

$$\sigma_a(E) = \sigma_0 \frac{\Gamma_a^2/4}{(E - E_\gamma)^2 + \Gamma_a^2/4} \tag{3.16}$$

where the maximum resonance cross section is

$$\sigma_0 = \frac{\lambda^2}{2\pi} \frac{1 + 2I_e}{1 + 2I_g} \frac{1}{1 + \alpha} \tag{3.17}$$

λ is the wavelength of the γ-ray, I_e and I_g are the nuclear spins of the excited and ground states, respectively, and α is the internal conversion coefficient.

A Mössbauer absorption spectrum is a record of transmitted γ-ray counts through the absorber as a function of γ-ray energy. Since the experiment involves both emission and absorption, we expect the line shape to be a convolution of the emission spectrum and the absorption cross section. Assuming a single-line emission and single-line absorption, we have

$$\begin{aligned}\sigma_a^{\exp}(E) &\propto \int_{-\infty}^{+\infty} \mathcal{L}(E - x)\sigma(x)dx \\ &= \frac{\sigma_0 \Gamma_a}{\Gamma_a + \Gamma_s} \frac{(\Gamma_s + \Gamma_a)^2/4}{(E - E_\gamma)^2 + (\Gamma_s + \Gamma_a)^2/4}\end{aligned} \tag{3.18}$$

which also has a Lorentzian shape, with a full width at half maximum equal to $\Gamma_s + \Gamma_a$ (the sum of the natural widths of the Mössbauer nuclei in the source and the absorber).

Experimentally, we usually measure resonance absorption by detecting the transmitted or scattered γ-rays. The vertical axis of a Mössbauer spectrum records the counts of detected γ-photons. In the transmission geometry, the original total intensity I_0 emitted from the source includes both recoil-free and recoiled γ-rays. After going through the absorber, the intensity is reduced mainly through two absorption processes: Mössbauer resonant absorption and nonresonant atomic absorption. When the source velocity does not meet the resonance condition, only the second process takes place and we observe an ideally flat baseline of the spectrum. When the source velocity is within the linewidth of absorption, resonant absorption significantly reduces the transmitted γ-rays and the spectrum shows an intensity minimum at the center of the resonance. The amount of maximum absorption may range from a few percent to 40% of the baseline count. The inverted peak can be characterized by parameters such as position of the resonance, full width at half-maximum absorption, and depth and area of the resonance line. When spectra are obtained in the scattering geometry, similar parameters are used, except that peaks representing the detected γ-rays, X rays, or conversion electrons are now above the baseline.

The horizontal axis of a Mössbauer spectrum is customarily labeled by the source velocity in millimeters per second. Velocity calibration is usually done by certain standard samples. For example, metallic α-iron is commonly used to determine the zero of source velocity and the scale of energy shift for ^{57}Fe spectra. The energy scale may also be independently calibrated by using a Michelson interferometer installed at the end of the oscillating shaft of the velocity transducer.

Since $\Gamma_s \approx \Gamma_a \sim 10^{-8}$ eV and E_γ is typically 10^4 eV, the energy resolution of a Mössbauer line can be as good as 10^{-12}. Because of this extremely high energy resolution, Mössbauer spectroscopy is ideally suited for studying hyperfine interactions.

3.7
Hyperfine Interactions

Because of the interactions between the nucleus and the surrounding electrons, nuclear energy levels are slightly altered and certain energy degeneracies are lifted. Some of the Mössbauer spectral lines will shift and others will split because of hyperfine interactions. These interactions are mainly electromagnetic but must be treated using quantum mechanics. The energies involved may be as small as 10^{-8} eV. Hyperfine interactions with nuclei at their ground states were initially observed in optical spectroscopy, followed by more accurate measurements from resonance experiments such as nuclear magnetic resonance (NMR) and electron paramagnetic resonance (EPR). Mössbauer spectroscopy offers an extra dimension of advantage. In addition to its high energy resolution, Mössbauer effect allows the observation of interactions between electrons and the nuclei in excited states, not just those in the ground state.

There are three main types of hyperfine interactions in Mössbauer effect:

1. electric monopole interaction (E0), which causes isomer shift δ, a shift of the entire resonance spectrum;
2. magnetic dipole interaction (M1), which causes magnetic hyperfine splittings of the spectral lines (Zeeman splittings);
3. electric quadrupole interaction (E2), which causes quadrupole splittings.

The electric dipole interaction (E1) is forbidden by parity requirement, and all interactions of higher orders (M2, E3, etc.) are much weaker to be detected.

3.7.1
Isomer Shift

We assume that a nucleus of atomic number Z is a uniformly charged sphere (monopole) when it is either in the ground state (with radius R_g) or in the excited state (with radius R_e). These radii have finite values and are different ($R_g \neq R_e$), although each is nearly zero compared with the size of an electron orbit. Let $\psi(0)$ represent the wave function of an s-electron within the volume of the nucleus ($r < R_g$, $r < R_e$, or practically, $r \approx 0$). The Coulomb energy between the nuclear charge and the s-electron's presence within the nuclear volume can be calculated to be

$$E_g = S(Z)Ze\frac{2\pi}{5}R_g^2 e|\psi(0)|^2 \quad \text{and}$$
$$E_e = S(Z)Ze\frac{2\pi}{5}R_e^2 e|\psi(0)|^2 \tag{3.19}$$

for the nuclear ground state and excited state, respectively. In these expressions, $S(Z)$ is the relativity factor due to relativistic effects in heavy elements (Shirley, 1964). Since the ground state is now raised by E_g and the excited state is raised by E_e, γ-transition energy is modified accordingly by

$$\Delta E = \frac{2\pi}{5}S(Z)Ze^2\left(R_e^2 - R_g^2\right)|\psi(0)|^2 \tag{3.20}$$

The same type of interaction occurs in the absorber (ΔE_a) as well as in the source (ΔE_s). Since the source and the absorber are usually not chemically identical, $\psi_a(0)$ differs from $\psi_s(0)$. As a consequence, the γ-photon energy change ΔE_a in the absorber is not identical to ΔE_s in the source. These energy changes are reflected in the Mössbauer spectrum as a shift of each line by an amount equal to

$$\delta = \frac{2\pi}{5} S(Z)\, Ze^2 \left(R_e^2 - R_g^2\right) \times \left(|\psi_a(0)|^2 - |\psi_s(0)|^2\right) \qquad (3.21)$$

This is due to two concomitant differences: one difference between the nuclear radius R_g in the ground state and its counterpart R_e in the excited state (a nuclear isomer), and the other difference between the s-electron density $\psi_a(0)$ in the absorber and its counterpart $\psi_s(0)$ in the source. This spectral shift δ is known as the *isomer shift* because of the volume difference in the nuclear isomers. It is also known as *chemical shift*, referring to its origin from the difference in chemical properties of the source and the absorber. The first difference is a fixed parameter once a nuclear isotope is chosen, whereas the second difference reflects the chemical contrast between the source and the absorber. In Mössbauer spectroscopy, we exploit the latter difference by using a series of absorbers of different chemical environments, for example, different oxidation states and coordination numbers.

Isomer shift is a relative quantity. If the nucleus increases in size when excited ($R_e > R_g$), the isomer shift increases linearly with the absorber's electron density $e|\psi_a(0)|^2$. For ^{57}Fe, the nucleus volume actually decreases when excited ($R_e < R_g$), and the isomer shift behavior is the opposite. The actual value of isomer shift depends on the choice of not only the source but also the calibration standard. Therefore, when reporting experimental results, it is important to state the type of source (e.g., ^{57}Co in rhodium matrix) and the reference point for isomer shift (e.g., with respect to the spectral center of α-Fe at room temperature). Typical ranges of isomer shifts for iron compounds in various oxidation states and spin states are available in the literature, for example, in Gibb's *Principles of Mössbauer Spectroscopy* (Gibb, 1976).

The isomer shift has been extensively investigated in many areas of chemistry and in materials science. Using isomer shift for studying the electronic structure in solids has been considered an extremely useful experimental method. The isomer shift δ can provide important information on the character of a chemical bond, as well as on oxidation state, spin state, electronegativity of a ligand, coordination number, and so on. A large amount of experimental isomer shift data has been accumulated to date; however, the interpretation of these results is not an easy task. The main difficulty is the lack of a unified model for the chemical bonds that can satisfactorily explain the isomer shift data. For practical applications, isomer shift measurements from a series of samples of similar properties are usually required. Through a comparative study, one can extract information on the electronic structure.

It should be pointed out that Mössbauer effect is so far the only method that can measure isomer shift, because the nuclear volume of the excited state is involved. The information on the electronic structure provided by isomer shift is not available from NMR, EPR, or any other methodologies.

3.7.2
Magnetic Dipole Interaction

The interaction between the nuclear magnetic dipole moment $\boldsymbol{\mu}$ and the magnetic field $\boldsymbol{B}$ produced at the site of nucleus by the surrounding electrons

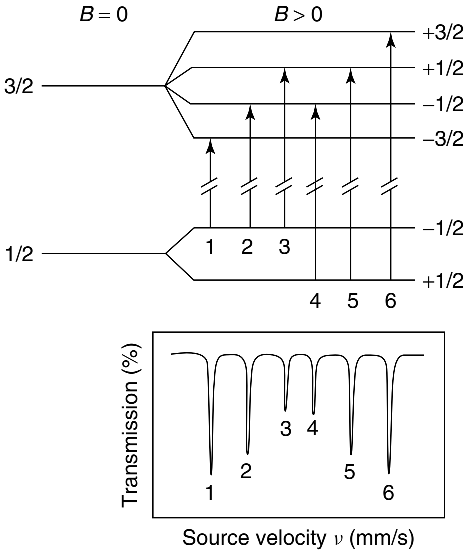

Fig. 3.5 Magnetic splittings of the ^{57}Fe nuclear energy levels and a sextet in the Mössbauer spectrum due to magnetic splittings.

or ions is called *the magnetic hyperfine interaction*. For a nucleus of spin $I \geq 1/2$, this interaction lifts the degeneracy of the energy level and splits it into $(2I + 1)$ sublevels, similar to the Zeeman effect in atomic spectroscopy. This type of splitting in the nuclear ground state had been observed in NMR and EPR. Using Mössbauer effect, nuclear Zeeman effect was first observed by Hanna *et al.* (1960).

The Hamiltonian of the interaction between a nuclear magnetic dipole moment $\boldsymbol{\mu}$ and a magnetic field $\boldsymbol{B}$ is

$$\mathcal{H}_M = -\hat{\mu} \cdot B = -g\mu_N \hat{I}_z B \tag{3.22}$$

where g is the dimensionless nuclear g-factor ($g = \mu/I\mu_N$), μ_N is the nuclear Bohr magneton (5.05×10^{-27} J T^{-1}), and the z-axis is along the direction of the magnetic field. The corresponding sublevel energies are

$$E_M = -gmB\mu_N \tag{3.23}$$

where $m = I, I - 1, \ldots, -I$.

The first excited state of ^{57}Fe has a spin of $I_e = 3/2$ and splits into four sublevels equally separated by $g_e B\mu_N$. The ground state has a spin of $I_g = 1/2$ and splits into two sublevels separated by $g_g B\mu_N$, as shown in Figure 3.5. The g_g-factor of the ground state, in general, is different from g_e of the excited state (for ^{57}Fe, $g_e = -0.1031$ and $g_g = 0.1808$). Therefore, the separation between the sublevels in the ground state is different from that in the excited state. Since the nuclear transition in ^{57}Fe is of the magnetic dipole type (M1), it can take place provided the selection rule ($\Delta m = \pm 1$ or 0) is obeyed. Thus there are only six allowed transitions, resulting in a sextet in the absorption spectrum as shown in Figure 3.5. The transitions with $\Delta m = \pm 2$ are forbidden.

The position of each line in the characteristic sextet can be easily calculated on the basis of values of $g_e B\mu_N$ and $g_g B\mu_N$. The separation between lines 3 and 4 equals $(g_g - |g_e|)B\mu_N$, while all other separations between adjacent lines are each equal to $g_g B\mu_N$.

The intensity of each absorption line is proportional to the transition probability, which can be calculated using the Clebsch–Gordan coefficients. Assuming a thin absorber with an isotropic recoil-free fraction f, we arrive at the following symmetric intensity ratio of the six absorption lines:

$$3:\frac{4\sin^2\theta}{(1+\cos^2\theta)}:1:1:\frac{4\sin^2\theta}{(1+\cos^2\theta)}:3 \tag{3.24}$$

where θ is the angle between the magnetic field $\boldsymbol{B}$ and the γ-ray direction. When the magnetic field is aligned with the γ-ray direction ($\theta = 0°$), the relative intensities of the sextet are in the ratio of 3 : 0 : 1 : 1 : 0 : 3. When the magnetic field is perpendicular to the γ-ray direction ($\theta = 90°$), the ratio is 3 : 4 : 1 : 1 : 4 : 3. If the magnetic field vectors at the nuclei are randomly oriented, then an integration over the entire solid angle gives an intensity ratio of 3 : 2 : 1 : 1 : 2 : 3.

A convenient way to describe the magnetic hyperfine interaction relies on the use of the effective magnetic field $\boldsymbol{B}_{\text{eff}}$, which has two major contributions: a local magnetic field $\boldsymbol{B}_{\text{loc}}$ at the Mössbauer nucleus by the lattice and a hyperfine magnetic field $\boldsymbol{B}_{\text{hf}}$ by the Mössbauer atom's own electrons,

$$\boldsymbol{B}_{\text{eff}} = \boldsymbol{B}_{\text{loc}} + \boldsymbol{B}_{\text{hf}} \tag{3.25}$$

The local field by the lattice may be due to the material's magnetic ordering, or may be applied externally, or both. It may have the following contributions:

$$\boldsymbol{B}_{\text{loc}} = \boldsymbol{B}_{\text{ext}} - D\boldsymbol{M} + \frac{4\pi}{3}\boldsymbol{M} \tag{3.26}$$

where $\boldsymbol{B}_{\text{ext}}$ is an external field, $\boldsymbol{M}$ is magnetization, $D\boldsymbol{M}$ represents the demagnetization field, and $4\pi\boldsymbol{M}/3$ represents the Lorentz field. In general, the local field is much smaller than the hyperfine field. The hyperfine field $\boldsymbol{B}_{\text{hf}}$ has three contributions:

$$\boldsymbol{B}_{\text{hf}} = \boldsymbol{B}_s + \boldsymbol{B}_L + \boldsymbol{B}_D \tag{3.27}$$

The first term $\boldsymbol{B}_s$ is called the *Fermi contact field* produced by spin polarization of *s*-electrons within the volume of the nucleus, and may be expressed as

$$\boldsymbol{B}_s = -\frac{4\mu_0}{3}\mu_B \sum_n [|\psi_{ns\uparrow}(0)|^2 - |\psi_{ns\downarrow}(0)|^2] \tag{3.28}$$

where μ_B is the Bohr magneton (9.27×10^{-4} J T^{-1}), $|\psi_{ns\uparrow}(0)|^2$ and $|\psi_{ns\downarrow}(0)|^2$ represent the *ns* spin-up and spin-down electron densities within the nucleus volume, respectively. The spin polarization of the *s*-electrons is caused by unpaired *d*-electrons because of an exchange interaction that slightly reduces the repulsion between electrons of the same spin. In iron compounds, $\boldsymbol{B}_s$ may range from 20 to 60 T and is the largest among the three terms in $\boldsymbol{B}_{\text{hf}}$, owing to a large number of unpaired *d*-electrons.

The second term $\boldsymbol{B}_L$ is called the *orbital field* owing to the orbital motions of the unpaired electrons around the nucleus. This motion constitutes a circular current, which in turn produces a magnetic field at the nucleus:

$$B_L = -\frac{\mu_0}{2\pi}\mu_B\langle r^{-3}\rangle\langle L\rangle \quad \text{or}$$
$$B_L = -\frac{\mu_0}{2\pi}\mu_B\langle r^{-3}\rangle(g-2)\langle S\rangle \tag{3.29}$$

where $\langle L\rangle$ and $\langle S\rangle$ are the expectation values of the orbital and spin angular momenta of the unpaired electrons.

For Fe^{3+}, $g \approx 2$, and consequently $\boldsymbol{B}_L \approx 0$. But for high-spin Fe^{2+}, $\boldsymbol{B}_L \approx 20$ T (opposite to $\boldsymbol{B}_s$) and, as expected, $\boldsymbol{B}_L$ values in rare-earth compounds are relatively large.

The third term $\boldsymbol{B}_D$ is the dipole field at the nucleus, produced by the total spin magnetic moment of the valence electrons. It can be written as

$$B_D = \frac{\mu_0}{8\pi}\mu_B \langle r^{-3}\rangle \langle 3\cos^2\theta - 1\rangle \langle S\rangle \tag{3.30}$$

$\boldsymbol{B}_D$ is obviously zero for a charge distribution with a cubic symmetry. For Fe group ions, $\boldsymbol{B}_D$ is small even in noncubic systems, ranging only from 0 to 8 T. However, in rare-earth compounds where the orbital momentum is not quenched, $\boldsymbol{B}_D$ can be quite large.

Since the separation between the lines of a sextet is proportional to the effective field $\boldsymbol{B}_{\text{eff}}$, in which the magnetic hyperfine field $\boldsymbol{B}_{\text{hf}}$ usually dominates, we can easily deduce $\boldsymbol{B}_{\text{hf}}$ from the spectrum. It is often very strong ($\boldsymbol{B}_{\text{hf}} \approx -33$ T in α-Fe) and it is local (within the vicinity of the nucleus). To measure the sign of the magnetic hyperfine field, we may apply an external magnetic field of 2 to 5 T to the sample and detect whether $\boldsymbol{B}_{\text{eff}}$ increases or decreases. When it increases, $\boldsymbol{B}_{\text{hf}}$ is positive; otherwise it is negative.

Mössbauer effect has found extensive applications in magnetism and in research of magnetic materials. Without requiring an applied external magnetic field, Mössbauer spectroscopy can be used to study the temperature dependence of spontaneous magnetization, the magnitude and orientation of hyperfine fields, and the magnetic structure of new materials. It can also be used to measure the ordering temperatures (T_C, T_N) and spin reorientation temperature, to determine phase transitions and phase compositions, to study magnetic lattice anisotropy and relaxation phenomena, and so on. Magnetism arises mainly from the atomic magnetic moments. Transition metals ($3d, 4d, 5d$), the lanthanides ($4f$), and the actinides ($5f$) all have unfilled valence electrons and have atomic magnetic moments. Fortunately, many isotopes of these elements are Mössbauer nuclei, for example, ^{57}Fe, ^{61}Ni, ^{99}Ru, ^{149}Sm, ^{151}Eu, ^{155}Gd, ^{159}Tb, ^{161}Dy, ^{165}Ho, ^{166}Er, ^{169}Tm, ^{170}Yb, ^{193}Ir, and ^{237}Np. Obviously, ^{57}Fe is most utilized because iron is the most important element in magnetism. The field of magnetism would not be as successful without Mössbauer spectroscopy.

3.7.3
Electric Quadrupole Interaction

If a nucleus (in either the ground state or an excited state) has a spin $I \geq 1$, it possesses an electric quadrupole moment eQ. For a nucleus surrounded by an asymmetric charge distribution, there is an electric field gradient (EFG), described by a tensor $\nabla \boldsymbol{E}$, at the site of the nucleus. In general, the EFG tensor contains nine second-order derivatives of electric potential ($V_{xy} = \partial^2 V/\partial x \partial y$, etc.). By choosing the principal z-axis along the highest gradient direction, we can diagonalize the EFG tensor so that $|V_{zz}| \geq |V_{xx}| \geq |V_{yy}|$. Since it is also traceless, $V_{xx} + V_{yy} + V_{zz} = 0$, only two independent parameters are needed to describe EFG. Customarily, V_{zz} is represented by eq, and the quantity $(V_{xx} - V_{yy})/V_{zz}$ is represented by η, known as the *asymmetry parameter* ($0 \leq \eta \leq 1$).

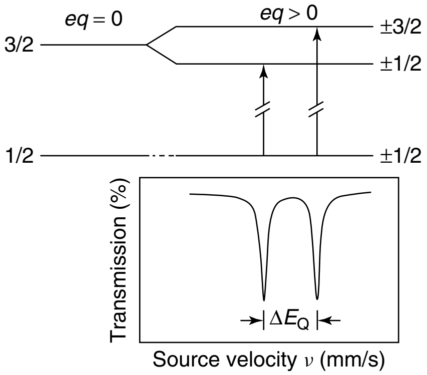

Fig. 3.6 Quadrupole splitting of the ^{57}Fe excited state and a doublet in the Mössbauer spectrum.

The Hamiltonian of this electric quadrupole interaction is

$$\mathcal{H}_Q = e\hat{Q}\nabla E = \frac{e^2qQ}{4I(2I-1)} \times \left[3\hat{I}_z^2 - \hat{I}^2 + \frac{1}{2}\eta(\hat{I}_+^2 + \hat{I}_-^2)\right] \tag{3.31}$$

where $\hat{I}_+ = \hat{I}_x + i\hat{I}_y$ and $\hat{I}_- = \hat{I}_x - i\hat{I}_y$ are the raising and lowering operators, respectively. The eigenvalues of this Hamiltonian are

$$E_Q(m) = \frac{e^2qQ}{4I(2I-1)}[3m^2 - I(I+1)] \times \left(1 + \frac{\eta^2}{3}\right)^{1/2} \tag{3.32}$$

where $m = I, I-1, \ldots -I$.

The ^{57}Fe excited state (14.4 keV) has a nuclear spin of $I = 3/2$, and this energy level splits into two sublevels ($m = \pm 3/2$ and $m = \pm 1/2$) due to quadrupole interaction. Because only m^2 appears in Eq. (3.32), each sublevel is still doubly degenerate. The energy eigenvalues of the two sublevels are

$$E_0 + E_Q\left(\pm\frac{3}{2}\right) = E_0 + \frac{e^2qQ}{4}\left(1 + \frac{\eta^2}{3}\right)^{1/2} \tag{3.33}$$

$$E_0 + E_Q\left(\pm\frac{1}{2}\right) = E_0 - \frac{e^2qQ}{4}\left(1 + \frac{\eta^2}{3}\right)^{1/2} \tag{3.34}$$

The ground state of ^{57}Fe has $I = 1/2$, so $Q = 0$, and the energy level does not split, as shown in Figure 3.6. Therefore, in the absence of Zeeman splitting, nuclear transition may take place from the ground state to either of the two excited states, resulting in a doublet in the Mössbauer spectrum, as shown in Figure 3.6. The separation between the two resonance lines, known as *quadrupole splitting*, is

$$\Delta E_Q = \frac{e^2qQ}{2}\left(1 + \frac{\eta^2}{3}\right)^{1/2} \tag{3.35}$$

Assuming an isotropic recoil-free fraction f, the relative intensities of the spectral lines of the quadrupole doublet in a single crystal can be calculated as

$$I_{\pm 3/2} = \frac{C}{\sqrt{1 + (\eta^2/3)}}\left(4\sqrt{1 + (\eta^2/3)} + 3\cos^2\theta - 1 + \eta\sin^2\theta\cos 2\phi\right) \tag{3.36}$$

$$I_{\pm 1/2} = \frac{C}{\sqrt{1+(\eta^2/3)}}\left(4\sqrt{1+(\eta^2/3)} - 3\cos^2\theta + 1 - \eta\sin^2\theta\cos 2\phi\right) \tag{3.37}$$

where C is a constant, and θ and ϕ are the polar and azimuthal angles of the incident γ-ray direction in the EFG principal-axis system. In the case of an axially symmetric EFG ($\eta = 0$), the intensity ratio of the two absorption lines is

$$\frac{I_{\pm 3/2}}{I_{\pm 1/2}} = \frac{1+\cos^2\theta}{\frac{5}{3}-\cos^2\theta} \tag{3.38}$$

For a polycrytalline sample or a powder sample, the crystal axes are randomly oriented. Integrating over the entire solid angle for each intensity, we obtain

$$\frac{\langle I_{\pm 3/2}\rangle}{\langle I_{\pm 1/2}\rangle} = \frac{\frac{1}{4\pi}\int I_{\pm 3/2}\sin\theta\, d\theta\, d\phi}{\frac{1}{4\pi}\int I_{\pm 1/2}\sin\theta\, d\theta\, d\phi} = 1 \tag{3.39}$$

If the recoil-free fraction is not isotropic but $f(\theta)$ is angular dependent, the above integrals must each be weighted by $f(\theta)$ and the ratio will not be 1 (a manifestation of the Goldanskii–Karyagin effect),

$$\frac{\langle I_{\pm 3/2}\rangle}{\langle I_{\pm 1/2}\rangle} = \frac{\frac{1}{4\pi}\int I_{\pm 3/2}f(\theta)\sin\theta\, d\theta\, d\phi}{\frac{1}{4\pi}\int I_{\pm 1/2}f(\theta)\sin\theta\, d\theta\, d\phi} \neq 1 \tag{3.40}$$

The quadrupole splitting $\Delta E_Q = 1/2e^2q\, Q(1+\eta^2/3)^{1/2}$ as measured from a Mössbauer spectrum only provides the product of eQ and eq, not their individual values or their signs. The nuclear quadrupole moment eQ is a fixed parameter for a particular isotope and may be independently measured (especially that of the ground state). The size and sign of the z-component of EFG eq ($= V_{zz}$) is determined by partially filled valence orbitals (valence) as well as by the charges on the neighboring ions and ligands surrounding the Mössbauer atom (lattice). Theoretical calculations of total V_{zz} at the site of the nucleus gave the following result:

$$V_{zz} = (1-R)(V_{zz})_{\text{val}} + (1-\gamma)(V_{zz})_{\text{lat}} \tag{3.41}$$

where $(V_{zz})_{\text{val}}$ and $(V_{zz})_{\text{lat}}$ are contributions from valence and lattice, respectively, each with an empirical parameter R (Sternheimer shielding factor) and a γ (Sternheimer antishielding factor) (Sternheimer, 1966). Therefore, ΔE_Q provides information on the symmetry of local chemical and electronic environment (oxidation state and spin state). Both $(V_{zz})_{\text{val}}$ and $(V_{zz})_{\text{lat}}$ are inversely proportional to r^3. Since the valence electrons are closer to the nucleus than the lattice charges, the former is much larger than the latter. For high-spin $^{57}\text{Fe}^{3+}$, the valence shell is $3d^5$ (exactly half filled), and $(V_{zz})_{\text{val}} = 0$ because of symmetry. For high-spin $^{57}\text{Fe}^{2+}$, the sixth electron in $3d^6$ is the main source of EFG. The quadrupole splitting for Fe^{2+} ($\Delta E_Q \approx 3.0\ \text{mm s}^{-1}$) has been found to be much larger than that for Fe^{3+} ($\Delta E_Q \approx 0.5\ \text{mm s}^{-1}$). The ΔE_Q values in low-spin $^{57}\text{Fe}^{2+}$ ($<0.8\ \text{mm s}^{-1}$) are smaller than the ΔE_Q values in low-spin $^{57}\text{Fe}^{3+}$ (0.7–$1.7\ \text{mm s}^{-1}$), again reflecting the one-electron difference in their d-orbitals.

$(V_{zz})_{\text{val}}$ may be strongly temperature dependent when a crystal field with lower-than-cubic symmetry splits the orbital energy levels. These suborbitals will be populated according to Boltzmann distribution, and hence the temperature dependence, which is reflected in the Mössbauer spectra taken at different temperatures (Chen *et al.*, 1992).

3.7.4
Combined Electric Quadrupole and Magnetic Dipole Interactions

When the above three hyperfine interactions are simultaneously present, the combined Hamiltonian is

$$\mathcal{H} = \mathcal{H}_{\mathrm{IS}} + \mathcal{H}_M + \mathcal{H}_Q \tag{3.42}$$

where the isomer shift Hamiltonian $\mathcal{H}_{\mathrm{IS}}$ is independent of the nuclear spin I, but both the magnetic dipole $\mathcal{H}_M$ and electric quadrupole $\mathcal{H}_Q$ Hamiltonians involve the same operator $\hat{I}$. When $\mathcal{H}_M$ and $\mathcal{H}_Q$ are considered separately, the quantization axis (the z-axis) for $\hat{I}$ in $\mathcal{H}_M$ is most conveniently chosen as the direction of the magnetic field $\boldsymbol{B}$; for $\hat{I}$ in $\mathcal{H}_Q$ it is along one principal axis of the EFG tensor. When both $\boldsymbol{B}$ and EFG are present, it is likely that the magnetic field is not aligned with the EFG principal axis. The combined Hamiltonian therefore takes a more complicated form:

$$\begin{aligned}\mathcal{H} = \mathcal{H}_{\mathrm{IS}} &+ \frac{e^2qQ}{4I(2I-1)} \\ &\times \left[3\hat{I}_z^2 - \hat{I}^2 + \frac{1}{2}\eta(\hat{I}_+^2 + \hat{I}_-^2)\right] \\ &- g\mu_N B\left\{\left[\frac{1}{2}(\hat{I}_+ + \hat{I}_-)\cos\phi \right.\right. \\ &\left. + \frac{1}{2}(\hat{I}_+ - \hat{I}_-)\sin\phi\right] \\ &\left. \times \sin\theta + \hat{I}_z \cos\theta \right\}\end{aligned} \tag{3.43}$$

where the z-axis is along the EFG principal axis, and θ and ϕ are the Euler angles describing the magnetic field direction within the EFG principal axes system. It is possible to obtain analytical solutions of eigenvalues and eigenvectors of this Hamiltonian, which involves the mixing of all original states and consequently allows the previously "forbidden" transitions (Häggström, 1974). To obtain complete analytical solutions is a tedious procedure and the solutions are very complicated for practical purposes.

In several special cases, the analysis is greatly simplified. When the EFG is axially symmetric ($\eta = 0$) and its z-axis happens to be aligned with the magnetic field $\boldsymbol{B}$ ($\theta = 0$), $\mathcal{H}_M$ and $\mathcal{H}_Q$ commute. The Zeeman states of $\mathcal{H}_M$ are also eigenvectors of $\mathcal{H}_Q$, and the new energy levels are simply combinations of the original E_M and E_Q:

$$\begin{aligned}E(m) &= E_{\mathrm{IS}} + E_M + E_Q \\ &= E_{\mathrm{IS}} - gmB\mu_N \\ &\quad + (-1)^{|m|+1/2}\frac{e^2qQ}{4}\end{aligned} \tag{3.44}$$

The Zeeman energy levels are no longer equally spaced; in the case of $I = 3/2$, the inner two energy levels are shifted in one direction and the outer two in the opposite direction.

It is often the case where either the magnetic hyperfine interaction or electric quadrupole interaction is dominant while the other occurs as a perturbation. For example, electric quadrupole interaction in a magnetically ordered solid is much weaker than the magnetic hyperfine interaction. On the other hand, an externally applied magnetic field to a paramagnetic material may result in a magnetic hyperfine interaction much weaker than the existing electric quadrupole interaction.

When the magnetic hyperfine interaction dominates, the energy levels are

$$E(m) = E_{\mathrm{IS}} - gmB\mu_N + (-1)^{|m|+1/2}\frac{e^2qQ}{4} \times \frac{3\cos^2\theta - 1 + \eta\sin^2\theta\cos 2\phi}{2} \tag{3.45}$$

which is similar to the previous case, except that the additional shift now depends on the relative angles (θ and ϕ) between the EFG and the magnetic field $\boldsymbol{B}$, as well as on the EFG's asymmetry parameter (η). When the electric quadrupole interaction dominates, the two states corresponding to $m = \pm 1/2$ will mix and give the following two different energy levels:

$$E(\pm ½) = E_{\mathrm{IS}} \pm \frac{1}{2}gB\mu_N \times \sqrt{4 - 3\cos^2\theta} - \frac{e^2qQ}{4} \tag{3.46}$$

and the two states corresponding to $m = \pm 3/2$ have the following energy levels:

$$E(\pm \tfrac{3}{2}) = E_{\mathrm{IS}} \pm \frac{3}{2}gB\mu_N\cos\theta + \frac{e^2qQ}{4} \tag{3.47}$$

Compared with the perfectly aligned and symmetric EFG cases, spectra from these more general cases are more difficult to analyze, but they allow us to extract more information, such as the sign of the EFG component V_{zz} and the relative direction of the hyperfine magnetic field.

3.7.5
Saturation Effect in the Presence of Hyperfine Splittings

It has been observed that in a sextet spectrum due to a magnetic hyperfine field the spectral intensities (absorption areas A) do not strictly follow the theoretical predictions of 3 : 2 : 1 : 1 : 2 : 3 for powder samples. This is because the area A is not a linear function of the effective thickness t_a of the absorber, but has a smaller slope for a larger t_a, manifesting a saturation effect as shown in Figure 3.7. Among the six absorption lines, each pair comes from participation of only one portion of the total population of Mössbauer nuclei: lines 1 and 6 are due to one-half of the total population, lines 2 and 5 are due to one-third of them, and lines 3 and 4 are due to only a sixth. Therefore, each pair corresponds to a different partial effective thickness. Because of the nonlinear saturation

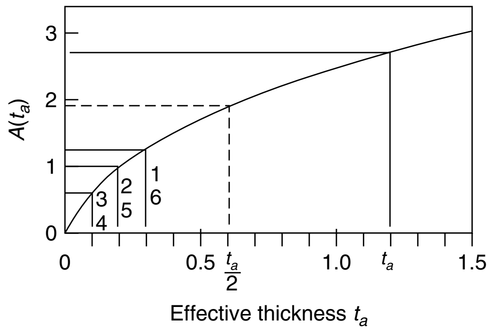

Fig. 3.7 Absorption area A as a function of effective thickness t_a. Because of the saturation effect, $2A\left(\frac{t_a}{2}\right) > A(t_a)$, and $12A\left(\frac{t_a}{12}\right) > 6A\left(\frac{t_a}{6}\right) > 4A\left(\frac{t_a}{4}\right)$.

effect, the absorption areas are not strictly in the ratio of 3 : 2 : 1, but $x : y : 1$, with $x < 3$ and $y < 2$.

Another important aspect is that the splitting of a single line will result in an increase of the total absorption area. Suppose a single line (effective thickness t_a) has an area of $A(t_a)$. When it splits into two sublines (each of partial effective thickness $\frac{t_a}{2}$), the total area is $2A(\frac{t_a}{2})$, which is clearly larger than $A(t_a)$, as illustrated in Figure 3.7.

When the absorption intensities are measured and used to infer any quantitative conclusions about the structural or chemical composition of the absorber, the saturation effect must be carefully taken into account.

3.8
Second-Order Doppler Effect

The second-order Doppler effect is a relativistic effect that results in the shift of the entire spectrum. Expanding Eq. (3.12) and retaining up to the second-order term, we obtain

$$\Delta E = E_0 \left(\frac{v}{c} - \frac{v^2}{2c^2} \right) \tag{3.48}$$

When atomic vibrations in the absorber are considered, their average velocity strongly depends on temperature. The corresponding Doppler shift is then

$$\langle \Delta E \rangle = -E_0 \frac{\langle v^2 \rangle}{2c^2} \tag{3.49}$$

This second-order Doppler effect modifies the energy of the γ-rays. In a Mössbauer spectrum, second-order Doppler shift δ_{SOD} is manifested as a relative quantity between the source and the absorber:

$$\delta_{\mathrm{SOD}} = \frac{\langle \Delta E \rangle}{E_0} = \frac{\langle v^2 \rangle_s - \langle v^2 \rangle_a}{2c^2} \tag{3.50}$$

where $\langle v^2 \rangle_s$ and $\langle v^2 \rangle_a$ are the mean-square velocities of the source and absorber, respectively.

Both the isomer shift and second-order Doppler shift have the effect of displacing the entire Mössbauer spectrum. They can be separated by utilizing the fact that the second-order Doppler shift depends strongly on temperature while the isomer shift is a measure of the s-electron density at the nucleus and approximately temperature-independent. Because the specific behavior of $\langle v^2 \rangle$ is determined by lattice vibration, δ_{SOD} is also useful in studying lattice dynamics.

3.9
Applications of Mössbauer spectroscopy

Mössbauer spectroscopy is mainly used to study the properties of materials through hyperfine interactions. Isomer shift, quadrupole splitting, magnetic hyperfine field, second-order Doppler effect, and the recoil-free fraction are all important parameters in Mössbauer spectroscopy. They are determined by the positions and the areas of spectral lines, as well as the spectral shapes and linewidths. These parameters also depend on external conditions such as temperature, pressure, and magnetic field. It is clear that Mössbauer spectra can provide an abundance of information.

Using hyperfine interactions, Mössbauer effect serves as a bridge linking the nucleus with its environmental details, including the crystal structures of materials, the perfectness of a crystal, lattice periodicity, magnetic ordering, the amorphous state, oxidation state in a compound, coordination number, effect of ligands, and the vibrations and diffusions of atoms in the solid. Because of its extremely diverse applications, Mössbauer spectroscopy has developed into a truly unique interdisciplinary field of science.

The isomer shift (chemical shift) δ provides information about the chemical environment, because of its sensitivity to changes in electron density at the site of the nucleus. Any changes in oxidation state, spin state, valence orbitals, type of chemical bond, ligand electronegativity, coordination number, substitutional or interstitial impurities, and so on, will be reflected in the isomer shift. Because it is often difficult to separate these causes, an investigation of a series of similar samples is usually required, and a comparison of their isomer shift data then provides unequivocal conclusions. In ^{57}Fe Mössbauer spectroscopy, isomer shift is very sensitive to the oxidation and spin states of iron. The δ-value (relative to α-Fe) is typically in the range from $+0.8$ to $+1.5$ mm s^{-1} for high-spin Fe^{2+} ($S = 2$), but in a distinctively different range of -0.2 to $+0.5$ mm s^{-1} for low-spin Fe^{+2} ($S = 0$). For Fe^{3+}, the isomer shift values for low spin ($S = 1/2$) and high spin ($S = 5/2$) are, respectively, $-0.2 - +0.4$ mm s^{-1} and $+0.2 - +0.6$ mm $^{-1}$s (Gibb, 1976). They overlap in the region from $+0.2 - +0.4$ mm s^{-1}; if a compound's isomer shift falls in this region, an independent determination of the spin state is required. Also, the two low-spin ions Fe^{2+} and Fe^{3+} have almost identical isomer shift values, but fortunately their quadrupole splittings are quite different, allowing unambiguous determinations of these two oxidation states.

The quadrupole splitting provides information about the lack of symmetry of the local electronic environment in which the nucleus resides. The asymmetry arises from nonspherical molecular orbitals or noncubic ligand coordination from the surrounding lattice. When combined with isomer shift data, quadrupole splitting values reflect the space group of local symmetry, the types of orbitals, relative orientations of local chemical bonds, and so on. Quadrupole splitting values may also be used to distinguish long-range and short-range orders in solids of less perfect lattices, such as amorphous materials and metallic glasses. Quadrupole splittings are also observed in magnetic materials, especially when they are in the paramagnetic phase, in the form of spin glasses, or in the superparamagnetic configuration. The behavior of quadrupole splitting as a function of temperature can be used to determine the respective phase transition temperatures.

The ubiquitous sextet ^{57}Fe Mössbauer spectrum, because of magnetic hyperfine interaction, contains a wealth of information about the magnetic properties of the material. Combined with result from bulk magnetic measurements, a spectrum with a sextet or several superimposed sextets indicates whether the material is ferromagnetic, antiferromagnetic, or ferrimagnetic. Fitting the experimental spectrum according to energy values of the Hamiltonian, we obtain parameters such as hyperfine magnetic field and isomer shift, which in turn allows us to calculate the magnetic moment of the atom. Experiments over a wide range of temperatures are beneficial for detecting magnetic phase transitions and for establishing a relationship between hyperfine magnetic field $\boldsymbol{B}_{\mathrm{hf}}$ (at the nuclear site) and the exchange field $\boldsymbol{B}_E$ (causing spontaneous magnetic ordering). For most magnetic ordered materials, $\boldsymbol{B}_{\mathrm{hf}}$ and $\boldsymbol{B}_E$ are proportional to each other, and they have identical temperature dependence. Therefore, information from Mössbauer spectroscopy can be directly correlated to macroscopic magnetic measurements (e.g., magnetization) as well as other microscopic measurements (e.g., NMR and muon resonance). For amorphous magnetic materials, the Mössbauer sextet is usually poorly resolved and broadened; it may be necessary to use a continuous distribution of hyperfine magnetic fields, instead of a few discrete values, to fit the spectrum.

Mössbauer spectroscopy has a particular timescale, characterized by the lifetime of the excited state from which the γ-rays are emitted. Usually of the order of 0.1 microseconds, this timescale τ is fast for many atomic processes, but it is slow relative to certain fluctuating mechanisms. Therefore, it is convenient to use Mössbauer spectroscopy to study materials that involve weak interactions between atoms, molecules, or clusters where relaxation times τ_0 may be comparable to τ. An example is superparamagnetism in which nanosized magnetic particles undergo fast thermal fluctuations, the magnetic hyperfine interaction is averaged out, and only quadrupole interaction is observed (Yang *et al.*, 2003). At lower temperatures when fluctuation slows down, hyperfine magnetic field begins to be observed. Near the so-called blocking temperature, a Mössbauer spectrum may reveal both coexisting phases, each with broadened spectral lines.

An external magnetic field may be applied to the absorber. A constant DC magnetic field applied to a magnetic material will allow us to determine the direction of the hyperfine field with respect to the atomic magnetic moment and, in the case of ferrimagnetic material, the direction and size of each component. When a constant external magnetic field is applied to a nonmagnetic material, one may observe Zeeman splittings as a perturbation to the quadrupole interaction, which will elucidate the direction of EFG relative to the crystallographic axes when the sample is a single crystal. An external high-frequency magnetic field may be applied to a soft magnetic material to induce a fast fluctuation of the hyperfine magnetic field, and therefore eliminate the Zeeman splittings in the Mössbauer spectrum, revealing only the quadrupole splittings and simplifying the analysis (Kopcewicz, 1989).

Other applications of Mössbauer spectroscopy include the following: measuring nuclear properties, testing general theory of relativity (Pound and Rebka, 1959; Kholmetskii, 2000), elucidating mechanism of catalysis (Jiang *et al.*, 2002), identifying minerals and ores (Stevens, 1998), studying corrosion in civil engineering (Cook, 2004), classifying chemical compositions of extraterrestrial samples and meteorites (Morris *et al.*, 2004), determining protein functions in biology (Bauminger and Harrison, 2003), analyzing iron compounds in atmospheric aerosols (Kopcewicz and Kopcewicz, 2000), exploring possibility of cancer therapy (Mills *et al.*, 1988), characterizing painting pigments in fine arts (Kuno *et al.*, 2004; Keisch, 1973), and archaeometry (Lazzarini *et al.*, 1980).

Mössbauer spectroscopy has enjoyed continued success because of the following extraordinary characteristics and advantages:

1. It has decisively the highest energy resolution of all spectroscopic methods.
2. Traditional Mössbauer spectroscopy requires only a relatively simple apparatus, in contrast to other nuclear physics research systems that are usually huge in size and exorbitantly expensive.
3. It is a nondestructive method for studying a solid and obtains microscopic statistical information on the atomic scale rather than a macroscopic average.

Mössbauer spectroscopy complements other nuclear methods such as neutron scattering, perturbed angular correlation, and nuclear magnetic resonance.

Like other experimental methods, Mössbauer spectroscopy has its own limitations. Although the total number of Mössbauer isotopes now exceeds 100, some of them have very short half-lives, and others require liquid-nitrogen or even liquid-helium temperatures for the observation of Mössbauer effect. By and large, for most isotopes, either the corresponding sources are very expensive or their Mössbauer effect is difficult to observe. ^{57}Fe remains to be the best Mössbauer isotope. Iron is found in a large array of materials. Many minerals have certain amount of Fe, magnetic materials are largely based on Fe, and there are numerous iron alloys and iron compounds. Because of this economic implication of iron and iron products, research using ^{57}Fe Mössbauer spectroscopy has always assumed the leading role.

Mössbauer spectroscopy has a history of 50 years since the first experiments of resonant absorption of nuclear γ-rays. New aspects of Mössbauer effect are still being investigated; especially, as SR became available, new theories and new techniques are being developed to make use of this artificial form of radiation to the fullest.

3.10
Mössbauer Spectroscopy Using Synchrotron Radiation

Electromagnetic radiation from a synchrotron accelerator may be used to observe Mössbauer effect. SR provides polarized pulsed radiation of high intensity, high collimation, and narrow beams. The only drawback is that SR is far from monochromatic. However, its energy can be adjusted and it can cover an energy range for a majority of Mössbauer transitions. The high intensity of SR sources has been utilized for scattering experiments where conventional radiation sources cannot provide adequate results. More importantly, the pulsed nature of SR radiation is most suitable for measuring time spectra – using a short SR pulse ($<10^{-10}$ seconds) to excite a nuclear ensemble, forming an "exciton" and allowing the observation of its coherent decay at different time intervals. The method of measuring time spectra is called *time-domain Mössbauer spectroscopy*, whereas the transmission method is referred to as *energy-domain Mössbauer spectroscopy*. In the last two decades, significant progress has been made in synchrotron Mössbauer spectroscopy, especially in the time-domain method, providing a direct and efficient approach to the study of Mössbauer effect and hyperfine fields. There emerged several new research areas that are not accessible with the conventional radiation sources. One example is the direct measurement of phonon DOS using SR sources. It was known that the phonon DOS could be measured, in principle, by Mössbauer effect, but because of various difficulties it was not achieved until SR became available in the 1990s. In 1999 and 2000, the journal *Hyperfine Interactions* devoted volumes 123, 124, and 125 to a comprehensive collection of 40 review articles written by the pioneers in the theory and experiments of nuclear resonant scattering of SR (Gerdau and de Waard, 1999, 2000).

The most distinct advantage of SR is its high brilliance. The brilliance of a third-generation SR source is about 9–10 orders of magnitude higher than that from a rotating target X-ray generator, and about 12 orders of magnitude higher than a ^{57}Co source of 3.7×10^8 Bq (10 mCi). SR can be tuned to the Mössbauer transition energies not only for the most common isotope ^{57}Fe but also for others such as ^{83}Kr, ^{119}Sn, ^{151}Eu, and ^{161}Dy. However, SR sources require a large and costly facility, and will not be available in ordinary Mössbauer laboratories. Therefore, the SR sources are unlikely to replace the conventional radiation sources.

3.10.1
Monochromatization of Synchrotron Radiation

The bandwidth of an SR beam is at least 10 orders of magnitude wider than the nuclear resonant width ($\sim 10^{-8}$ eV). After going through a double-crystal premonochromator Si(111), SR bandwidth can be reduced to the order of electronvolts. Further collimation and monochromatization are necessary. There are mainly two types of monochromators, one based on the scattering by electrons and the other on resonant scattering by nuclei. The first type can offer a bandwidth within a few milli electronvolts, and the second type can provide the desired bandwidth of a few micro electronvolts for nuclear resonant experiments.

Using monochromators based on electron scattering and specifically designed detectors, many lattice dynamics experiments as well as nuclear forward scattering experiments have been performed with acceptable signal-to-noise ratios, although a milli electronvolt bandwidth is still about 10^6 times wider than the resonance linewidth of ^{57}Fe. A typical scatterer contains a large number of electrons in addition to the resonant nuclei, and those electrons will nonresonantly scatter all the SR in the meV band. Therefore, only 10^{-6} of the detected photon count is due to nuclear resonant scattering. The electron scattering process is prompt in time, whereas the nuclear resonant scattering is a time-delayed process since the typical lifetime of the nuclear isomeric state is long ($\tau \sim 100$ nanoseconds) compared to the incident SR pulse ($\sim$0.1 nanoseconds). Using this time difference, photons from nonresonant electron scattering can be discriminated by a detector with nanosecond rise and fall times. For certain crystalline scatterers, we may choose specific scattering angles so that the electronic Bragg reflections are canceled or forbidden, for example, using the $(2n+1\ 2n+1\ 2n+1)$ plane in α-Fe_2O_3.

Through the use of nuclear Bragg scattering, monochromators such as single crystal ^{57}FeBO$_3$, multilayers 25× [^{57}Fe

(22 Å)/Sc(11 Å)/^{56}Fe(22 Å)/Sc(11 Å)], or the grazing incident antireflection (GIAR) film can further reduce the bandwidth to 10^{-6}–10^{-8} eV, approaching the natural width of the nuclear energy level. Because hyperfine interactions are involved in the above nuclear Bragg scattering, the resultant radiation may contain a few separate spectral lines. This multicomponent radiation may be directly used in experiment; but if a truly single-line monochromatic source is desired, an additional absorber may be added to filter out the unwanted spectral components.

The following description represents a specific example of monochromatization and polarization of SR (Smirnov, 2000). SR is first reflected off a double-crystal Si(1 1 1) monochromator, reducing the bandwidth to 2.8 eV at the nuclear resonance energy. A channel-cut Si(8 4 0) polarizer reduces the π-polarized component from 1% to less than 10^{-4} % and the bandwidth to milli electronvolts. It is then reflected off the (3 3 3) plane of a single-crystal ^{57}FeBO$_3$, placed in an oven (above its Neèl temperature to eliminate magnetic hyperfine interaction) with an external magnetic field in the (3 3 3) plane. The outcome is an extremely narrow, linearly polarized, completely recoil-free, and highly intense Mössbauer source for ^{57}Fe, equivalent to a ^{57}Co source of an enormous activity of $\sim 3.7 \times 10^{13}$ Bq (1000 Ci).

3.10.2
Time-Domain Mössbauer Spectroscopy

In time-domain Mössbauer spectroscopy (TDMS), coherent decays in resonant absorption, nuclear Bragg scattering, or forward scattering are observed at different times (~10–100 nanoseconds) after the nuclear system has been excited. Mössbauer parameters and hyperfine interactions can be studied by TDMS through the analysis of several new phenomena such as dynamical beats (DBs), quantum beats (QBs), and speed-up effect of initial decay.

In time-domain experiments, a strong pulse of SR simultaneously excites all Mössbauer nuclei in a sample and they decay in a coherent fashion. Each nucleus is no longer isolated, but interacts with others. It is possible for a nuclear excitation to propagate elastically throughout the entire ensemble of nuclei (a γ-photon emitted by one excited nucleus may be reabsorbed or scattered by other nuclei that are identical to the emitting nucleus). Without this interaction, individually excited nucleus would decay exponentially with its natural lifetime. However, an interacting ensemble of nuclei behaves differently. Such a collective nuclear excitation by an SR pulse is known as a *nuclear exciton*. Elastic decay of this exciton deviates from exponential behavior and contains intensity modulations characterized by beats.

In experiments using traditional Mössbauer radiation source, we may detect how the intensity of radiation, emitted from a source, evolves with time as it goes through an absorber. Typically, a delay coincidence circuit is used; a block diagram is shown in Figure 3.8 where an oscillating absorber allows absorption at a certain resonance velocities. In the case of traditional ^{57}Fe spectroscopy using a ^{57}Co source, it is convenient that 91% of the decay process follows a cascade through two excited states (136.5 keV → 14.4 keV → ground state). The time zero is set by the detection of a 122.1-keV photon (γ_1), which signifies that the nucleus

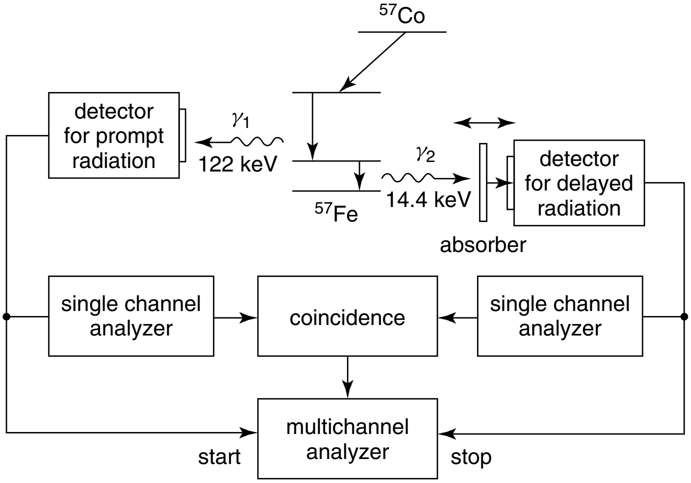

Fig. 3.8 Block diagram of the prompt and delay coincidence detection circuits.

is now in the 14.4-keV state (natural lifetime $\tau = 98$ nanoseconds). This signal starts the time measurement t until a 14.4-keV photon (γ_2) is detected. A time-domain spectrum is then constructed by plotting the number of events $N(t)$ against the time delay t. With no absorber, this curve is perfectly exponential $N(t) \propto e^{-t/\tau}$ and measures the lifetime τ. With the oscillating absorber in place, the detected γ_2 photons include those that have energies just above resonance and those just below resonance. Because these two groups of photons have very distinct dispersion characteristics (traveling through the absorber with very different speeds), $N(t)$ at the detector is an "interference pattern" instead of a purely exponential decay. It is characterized by aperiodic modulations, known as *dynamical beats* (DBs), and by the speed-up of initial decay.

Theoretically, this can be modeled as the propagation of a truncated harmonic wave through a dispersive medium, and the result is

$$N(t) \propto e^{-t/\tau} \left[J_0(\sqrt{t_a t/\tau}) \right]^2 \tag{3.51}$$

where the exponential behavior is now modulated by a Bessel function of the zeroth order J_0, oscillating aperiodically with the apparent periods increasing with time (Figure 3.9). In addition, the modulation depends on t_a, the effective thickness of the absorber. For a very thin absorber ($t_a \ll 1$) or very short times ($t \ll \tau$), we may expand the Bessel function and retain the first term. This portion of the time-domain spectrum becomes

$$N(t) \propto e^{-t/\tau} e^{-t_a t/2\tau} \tag{3.52}$$

where the second factor is also exponential, making the apparent decay rate to be $\tau/(1 + t_a/2)$, shorter than the natural lifetime τ. This phenomenon is known as *speed up* of initial decay.

DBs and speed up of initial decay can also be observed in experiments using SR, where a short pulse (~0.1

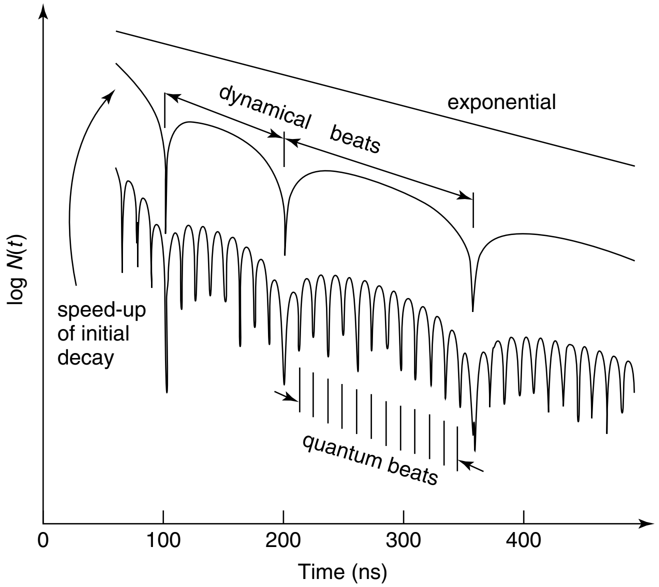

Fig. 3.9 Main features in time-domain Mössbauer spectra, showing speed up of initial decay, dynamical beats, and quantum beats.

nanosecond) of radiation of energies within a particular bandwidth is applied. The intensity of nuclear Bragg scattering or forward scattering is then observed as a function of time. (For readers familiar with pulsed NMR, this is similar to one single RF pulse followed by the observation of "free induction decay," or FID.) The theoretical treatment now requires the inclusion of all the energies within the bandwidth of the SR radiation. The resultant time-domain spectrum (DB) can be expressed in terms of a Bessel function of the first order J_1,

$$N(t) \propto e^{-t/\tau}\left[t_a \frac{J_1(\sqrt{t_a t/\tau})}{\sqrt{t_a t/\tau}}\right]^2 \qquad (3.53)$$

For $t_a \ll 1$ and $t \ll \tau$, we have again the speed up of initial decay:

$$N(t) \propto t_a^2 e^{-t/\tau} e^{-t_a t/4\tau} \qquad (3.54)$$

with an apparent rate of $\tau/(1 + t_a/4)$.

When the scatterer contains hyperfine interactions and therefore allows multiple resonant scatterings, another interference phenomenon takes place: periodic modulations (known as *quantum beats* (*QBs*)) are now superimposed onto the DBs, as shown in Figure 3.9. For example, a scatterer with two resonance energies E_1 and E_2 would give

$$N(t) \propto e^{-t/\tau}\left[t_a \frac{J_1(\sqrt{t_a t/\tau})}{\sqrt{t_a t/\tau}}\right]^2 \times \cos^2\left(\frac{E_2 - E_1}{2\hbar}t\right) \qquad (3.55)$$

In this case, the QBs are periodic and they contain information about the hyperfine splitting.

Comparing time-domain Mössbauer spectroscopy with energy-domain Mössbauer spectroscopy, we find similarities and differences. Both are phenomena of nuclear resonance. The energy-domain

Mössbauer spectra are mainly based on resonance absorption. The transmitted counts of photons are measured as functions of their energies (an energy spectrum), which represents an incoherent sum of the spectral components of the transmitted radiation. In other words, the transmitted spectrum reflects the incoherent process of nuclear resonance absorption by individual nuclei. By contrast, time-domain Mössbauer spectroscopy belongs to the scattering method. A scattering spectrum measured as a function of time is a coherent sum of the spectral components of the scattered radiation from nuclei collectively excited by an SR pulse. This leads to those important interference effects in TDMS.

A time-domain spectrum may be Fourier-transformed to an energy-domain spectrum. A purely exponential time-domain spectrum gives a familiar Lorentzian energy spectrum. However, time-domain spectra with QB and DB will give energy-domain spectra with interesting shapes, such as a double-hump structure.

3.10.3
Applications of Time-Domain Mössbauer Spectroscopy

Time-domain Mössbauer spectroscopy has found many applications, among which measurements of the recoil-free fraction f and the phonon DOS are perhaps the most interesting and the easiest to understand. (Strictly speaking, one does not measure the recoil-free fraction f directly in scattering experiments, but use the Lamb–Mössbauer factor f_{LM}. In an isotropic scatterer, f_{LM} and f are identical.)

In the time domain, we measure the recoil-free fraction f through determining t_a because they are linearly related according to Eq. (3.14), $t_a = n_a f \sigma_0 d$. Using nuclear forward scattering, t_a may be extracted from the speed up of initial decay or from the minimum positions of DBs. An excellent set of experiments has been carried out using, as the scatterer, a polycrystalline α-Fe foil of thickness of 11 μm and a 95% ^{57}Fe enrichment (Bergmann *et al.*, 1994). A magnetic field of 0.6 T was applied in the plane of the foil so that only the $\Delta m = 0$ transitions were allowed. As the scatterer temperature was varied between 9.7 and 1048 K, time-domain Mössbauer spectra were obtained using SR of bandwidth 10 meV. A large time window was open to observe as many periods of DB as possible so that t_a could be deduced more accurately. The time spectra contained both QBs and DBs, from which temperature dependence of the magnetic hyperfine field $\boldsymbol{B}_{\mathrm{hf}}$ were determined from the QB periods and that of the Lamb–Mössbauer factor f_{LM} from the DB "periods." The results were in excellent agreement with traditional energy-domain Mössbauer spectroscopy and bulk magnetic measurements.

Using SR of energy resolution 6 meV tuned precisely at the Mössbauer transition energy, inelastic nuclear resonant scattering has been applied to measure the phonon DOS directly (Sturhahn *et al.*, 1995). A very short pulse (<0.1 nanosecond) of SR excites essentially all electrons and nuclei in the scatterer. Nonresonantly scattered radiation (by nuclei other than the Mössbauer isotope and all electrons) appears promptly in time (within a few nanoseconds), whereas photons that are resonantly absorbed by a Mössbauer nuclei need a delayed time (comparable to the natural life time of the excited state) to be reemitted. The prompt nonresonantly scattered radiation was eliminated by only counting in a time window of 30 to 600 nanoseconds after the arrival of the SR flash. Photons with precisely the Mössbauer transition energy will naturally cause nuclear resonance and hence elastic scattering. Photons with less energy can also excite nuclear resonance by annihilation of a phonon (lattice vibration), and photons with more energy can excite nuclear resonance by creation of a phonon. The probability of annihilation or creation of phonons is proportional to the local DOS. By detecting the delayed *K*-fluorescence radiation due to conversion electrons caused by the deexcitation of the Mössbauer nuclei, the phonon DOS can be directly measured. This phonon spectrum can also be normalized and integrated to give the recoil-free fraction f or the Lamb–Mössbauer factor f_{LM}, which provides an alternative method independent of isotope abundance, sample thickness, resonance cross section, or hyperfine interactions.

3.11
Synchrotron Methods versus Conventional Methods

Synchrotron Mössbauer spectroscopy has attracted a significant amount of attention from researchers and become a well-established methodology in the last decade. SR provides extremely strong and narrow photon beams, which facilitates spectral measurement under special experimental conditions, such as high temperature, high pressure, high magnetic field, and working with small samples ($\sim$1 mm^2) or nanostructured thin films. In biological samples, the concentration of Mössbauer isotope is usually too low for conventional Mössbauer spectroscopy, but SR should be strong enough to produce detectable signals. Usually, results for the recoil-free fraction f from synchrotron Mössbauer spectroscopic have higher accuracy than results from the conventional methods, partly due to the absence of the saturation effect in synchrotron Mössbauer methods. Scattering experiments require only relatively short measurement time, ranging from a few minutes to a few hours. The conventional Mössbauer spectroscopy may require up to hundreds of hours. Hyperfine parameters measured from synchrotron experiments have accuracies comparable with those from conventional experiments. In QB experiments, sufficient data accumulation is required in the chosen time window to yield satisfactory measurements of the periods.

Perhaps the most important contribution from synchrotron Mössbauer spectroscopy is its ability to measure phonon DOS directly. So far, this has not been achieved by conventional Mössbauer spectroscopy. Phonon DOS may be deduced from inelastic neutron scattering by extracting force constants from the fitted dispersion curves. However, on using inelastic nuclear resonant scattering with SR where the Mössbauer nuclei serve as analyzers, phonon DOS can be measured without phonon dispersion relations; hence this is a model-independent and more accurate method. Furthermore, inelastic nuclear resonant scattering allows us to measure *partial* density of states (PDOS) related to a specific isotope or a particular impurity.

However, synchrotron Mössbauer spectroscopy suffers from several shortcomings. Because time-domain experiments are based on the interference phenomenon, the corresponding spectra are very complex, whereas conventional Mössbauer spectra provide certain direct visual information. If two or more hyperfine fields are involved, time spectra may be severely modulated. The second major drawback of synchrotron Mössbauer spectroscopy is obviously the high expenses in constructing and maintaining such a large centralized synchrotron facility. Synchrotron Mössbauer spectroscopy is an important supplement to the conventional Mössbauer spectroscopy. The former will never completely replace the latter, but will help solve problems that cannot be studied or solved satisfactorily by using conventional methods.

References

Bauminger, E.R. and Harrison, P.M. (2003) Ferritin, the path of iron into the core, as seen by Mössbauer spectroscopy. *Hyperfine Interact.*, **151/152**, 3–19.

Bergmann, U., Shastri, S.D., Siddons, D.P., Batterman, B.W., and Hastings, J.B. (1994) Temperature dependence of nuclear forward scattering of synchrotron radiation in α-^{57}Fe. *Phys. Rev. B*, **50**, 5957–5961.

Chen, Y.L., Xu, B.F., Chen, J.G., and Ge, Y.Y. (1992) Fe^{2+}–Fe^{3+} ordered distribution in chromite spinels. *Phys. Chem. Miner.*, **19**, 255–259.

Cook, D.C. (2004) Application of Mössbauer spectroscopy to the study of corrosion. *Hyperfine Interact.*, **153**, 61–82.

Gerdau, E. and de Waard, H. (1999) *Hyperfine Interact.*, **123/124**, reviews and references therein, 3–879.

Gerdau, E. and de Waard, H. (2000) *Hyperfine Interact.*, **125**, reviews and references therein, 3–221.

Gibb, T.C. (1976) *Principles of Mössbauer Spectroscopy*, Chapman & Hall, London.

Goldanskii, V.I., Gorodinskii, G.M., Karyagin, S.V., Korytko, L.A., Krizhanskii, L.M., Makarov, E.F., Suzdalev, I.P., and Khrapov, V.V. (1962) The Mössbauer effect in tin compounds. *Proc. Acad. Sci. USSR (Phys. Chem. Sect.)*, **147**, 766–768.

Häggström, L. (1974) *Determination of Hyperfine Parameters from 1/2 → 3/2 Transitions in Mössbauer Spectroscopy*, Uppsala University Institute of Physics Report UUIP-851, Uppsala, Sweden.

Hanna, S.S., Heberle, J., Littlejohn, C., Perlow, G.J., Preston, R.S., and Vincent, D.H. (1960) Polarized spectra and hyperfine structure in Fe^{57}. *Phys. Rev. Lett.*, **4**, 177–180.

Jiang, K.Y., Yang, X.L., Yuan, Y.T., Mao, L.S., and Yang, D.P. (2002) A Mössbauer effect study on the structural components in potassium-promoted iron oxide catalysts for dehydrogenation of ethylbenzene. *Hyperfine Interact.*, **139**, 97–105.

Karyagin, S.V. (1963) A possible cause for the doublet component asymmetry in the Mössbauer absorption spectrum of some powdered tin compounds. *Proc. Acad. Sci. USSR (Phys. Chem. Sect.)*, **148**, 110–112.

Keisch, B. (1973) Mössbauer effect studies in the fine arts. *Archaeometry*, **15**, 79–104.

Kholmetskii, A.L. (2000) Proposal for a test of the relativity principle by using Mössbauer synchrotron radiation. *Hyperfine Interact.*, **126**, 411–416.

Kopcewicz, M. (1989) in *Mössbauer Spectroscopy Applied to Inorganic Chemistry* (eds G.J. Long and F. Grandjean), Vol. III, Plenum, New York, pp. 243–287.

Kopcewicz, B. and Kopcewicz, M. (2000) Ecological aspects of Mössbauer study of iron-containing atmospheric aerosols. *Hyperfine Interact.*, **126**, 131–135.

Kuno, A., Matsuo, M., Pascual Soto, A., and Tsukamoto, K. (2004) Mössbauer spectroscopic study of a mural painting from Morgadal Grande, Mexico. *Hyperfine Interact.*, **156/157**, 431–437.

Lazzarini, L., Calogero, S., Burriesci, N., and Petrera, M. (1980) Chemical, mineralogical and Mössbauer studies of Venetian and

Paduan Renaissance sgraffito ceramics. *Archaeometry*, **22**, 57–68.

Lipkin, H.J. (1960) Some simple features of the Mössbauer effect. *Ann. Phys.*, **9**, 332–339.

Mills, R.L., Walter, C.W., Venkataraman, L., Pang, K., and Farrell, J.J. (1988) A novel cancer therapy using a Mössbauer-isotope compound. *Nature*, **336**, 787–789.

Morris, R.V. *et al.* (2004) Mineralogy at Gusev Crater from the Mössbauer spectrometer on the Spirit Rover. *Science*, **305**, 833–836.

Mössbauer, R.L. (1958) Kernresonanzfluoreszenz von Gammastrahlung in Ir^{191}. *Z. Phys.*, **151**, 124–143.

Mössbauer, R.L. (2000) The discovery of the Mössbauer effect. *Hyperfine Interact.*, **126**, 1–12.

Pound, R.V. and Rebka, G.A. (1959) Gravitational red-shift in nuclear resonance. *Phys. Rev. Lett.*, **3**, 439–441.

Shirley, D.A. (1964) Application and interpretation of isomer shifts. *Revs. Mod. Phys.*, **39**, 339–351.

Smirnov, G.V. (2000) Synchrotron Mössbauer source of ^{57}Fe radiation. *Hyperfine Interact.*, **125**, 91–112.

Sternheimer, R.M. (1966) Shielding and antishielding effects for various ions and atomic systems. *Phys. Rev.*, **146**, 140–160.

Stevens, J.G. (1998) Documentation and evaluation of Mössbauer data for minerals. *Hyperfine Interact.*, **117**, 71–81.

Sturhahn, W., Toellner, T.S., Alp, E.E., Zhang, X., Ando, M., Yoda, Y., Kikuta, S., Seto, M., Kimball, C.W., and Dabrowski, B. (1995) Phonon density of states measured by inelastic nuclear resonant scattering. *Phys. Rev. Lett.*, **74**, 3832–3835.

Yang, D.P., Lavoie, L.K., Zhang, Y.D., Zhang, Z.T., and Ge, S.H. (2003) Mössbauer spectroscopic and x-ray diffraction studies of structural and magnetic properties of heat-treated $(Ni_{0.5}Zn_{0.5})Fe_2O_4$ nanoparticles. *J. Appl. Phys.*, **93**, 7492–7494.

Yoshida, Y., Kobayashi, Y., Yoshida, A., Diao, X., Ogawa, S., Hayakawa, K., Yukihira, K., Shimura, F., and Ambe, F. (2002) In-beam Mössbauer spectroscopy after GeV-ion implantation at an on-line projectile-fragments separator. *Hyperfine Interact.*, **141/142**, 157–162.

Further Reading

Bancroft, G.M. (1973) *Mössbauer Spectroscopy: An Introduction for Inorganic Chemists and Geochemists*, John Wiley & Sons, Ltd, New York.

Chen, Y.L. and Yang, D.P. (2007) *Mössbauer Effect in Lattice Dynamics*, Wiley-VCH Verlag GmbH, Weinheim.

Dickson, D.P.E. and Berry, F.J. (eds) (1986) *Mössbauer Spectroscopy*, Cambridge University Press, Cambridge.

Frauenfelder, H. (ed.) (1962) *The Mössbauer Effect*, W.A. Benjamin Publishers, New York.

Gonser, U. (ed.) (1975) *Mössbauer Spectroscopy*, Springer-Verlag, New York.

Gonser, U. (ed.) (1981) *Mössbauer Spectroscopy II: The Exotic Side of the Method*, Springer-Verlag, Berlin.

Gruverman, I.J. (ed.) (1965–1975) *Mössbauer Effect Methodology*, Plenum Press, New York, Vols. 1–10.

Herber, R.H. (ed.) (1984) *Chemical Mössbauer Spectroscopy*, Plenum Press, New York.

Long, G.J. (ed.) (1984) *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, Vol. 1, Plenum Press, New York.

Long, G.J. (ed.) (1987) *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, Vol. 2, Plenum Press, New York.

Long, G.J. and Grandjean, F. (eds) (1989) *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, Vol. 3, Plenum Press, New York.

Long, G.J. and Grandjean, F. (eds) (1993) *Mössbauer Spectroscopy Applied to Magnetism and Materials Science*, Vol. 1, Plenum Press, New York.

Long, G.J. and Grandjean, F. (eds) (1996) *Mössbauer Spectroscopy Applied to Magnetism and Materials Science*, Vol. 2, Plenum Press, New York.

Kopcewicz, M. (1994) Mössbauer effect, in *Encyclopedia of Applied Physics* (ed G.L. Trigg), Vol. 11, VCH Publishers, New York, pp. 1–22.

Kolk, B. (1984) Studies of dynamical properties of solids with the Mössbauer effect, in *Dynamical Properties of Solids* (ed G.K. Horton and A.A. Maradudin), Vol. 5, North-Holland Physics Publishing, Amsterdam, pp. 1–328.

Maddock, A. (1997) *Mössbauer Spectroscopy: Principles and Applications*, Horwood Publishing, Chichester.

Stevens, J.G. (1981) Mössbauer spectroscopy, in *Handbook of Spectroscopy* (ed. J.W. Robinson), Vol. III, CRC Press, Boca Raton, pp. 403–528.

Stevens, J.G. and Shenoy, G.K. (eds) (1981) *Mössbauer Spectroscopy and its Chemical Applications*, American Chemical Society, Washington, DC.

11
Ultraviolet and Visible Absorption Spectroscopy

Robert H. Lipson

11.1 **Introduction** **355**
11.2 **Spectral Background** **355**
11.2.1 State Labels 355
11.2.1.1 Atoms 355
11.2.1.2 Molecules 356
11.2.2 Electronic Selection Rules 357
11.2.3 Born–Oppenheimer Approximation and the Franck–Condon Principle 358
11.2.4 Rotational Fine Structure 361
11.3 **Absorption Spectra** **362**
11.3.1 Beer–Lambert Law 362
11.3.2 Chemometrics 363
11.3.3 Derivative Spectroscopy 363
11.3.4 Circular Dichroism Spectroscopy 363
11.4 **The Origin of UV/Visible Transitions in Organic Molecules** **364**
11.4.1 Chromophores 365
11.5 **The Origin of UV/Visible Transitions in Inorganic Molecules** **368**
11.6 **The Origin of UV/Visible Transitions in Solids** **370**
11.7 **Experimental Techniques for UV/Visible Absorption Spectroscopy** **371**
11.7.1 Light Sources 371
11.7.1.1 Thermal Sources 371
11.7.1.2 Luminescent Sources 371
11.7.1.3 Light-Emitting Diodes 372
11.7.1.4 Synchrotrons 372
11.7.2 Wavelength Dispersive Instruments 372
11.7.2.1 Fastie–Ebert Configuration 373
11.7.2.2 Czerny–Turner Configuration 374
11.7.2.3 Single-Beam Versus Double-Beam Spectrometers 374
11.7.3 Detectors 374
11.7.3.1 Photodiodes 374
11.7.3.2 Photomultipliers (PMTs) 375

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

11.7.3.3 Microchannel Plates (MCP) 375
11.7.3.4 Charge Coupled Devices (CCD) 376
11.8 Laser-Based Techniques 376
11.8.1 Fluorescence Excitation Spectroscopy 376
11.8.2 Excitation Spectroscopy with Mass Detection 377
11.8.3 Cavity Ring-Down Spectroscopy 378
References 378
Further Reading 379

11.1
Introduction

An overview of spectroscopy fundamentals for atoms and molecules is provided as they pertain to the ultraviolet (UV) and visible spectral regions. Topics include state labels, electronic selection rules, the Born–Oppenheimer approximation, and the Franck–Condon principle. The Beer–Lambert law, chemometrics, derivative spectroscopy, and circular dichroism (CD) spectroscopy are explained as are the origins of UV/visible transitions in organic and inorganic compounds as well as solids. Relevant experimental methods and instrumentation such as light sources, monochromators, and detectors, which can be used to record UV/visible absorption spectra, are described. A brief discussion on laser-based techniques, including fluorescence excitation spectroscopy, excitation spectroscopy with mass detection, and cavity ring-down spectroscopy, is also given.

11.2
Spectral Background

Absorption occurs when electromagnetic radiation interacts with matter resulting in a loss of beam intensity after passing through the sample. Light absorption can take place in the gas phase, liquid phase, solid phase, or in a plasma, and can involve a single species or a complex mixture.

The UV spectral region lies between ~200 and 400 nm while the longer wavelength visible region lies between ~400 and 700 nm. The conversions to frequency (ν), cm^{-1} wave number ($\tilde{\nu}$), and energy (E) units are presented in Table 11.1.

UV/visible photons are usually sufficiently energetic to promote electrons from the atomic or molecular ground states to higher lying valence or Rydberg states. In the case of molecules, the energy can also excite excited-state vibrational and rotational motions.

11.2.1
State Labels

11.2.1.1 Atoms

The orbital angular momentum of an electron in an atom can only adopt integer values : $\ell = 0, 1, 2,$ and so on. The total orbital angular momentum of a many-electron system **L** can be calculated from the vector sum $\mathbf{L} = \overrightarrow{\sum_i \ell_i}$. Although **L** can only have integer

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

Tab. 11.1 Wavelength, λ, frequency, ν, wave number, $\tilde{\nu}$, and energy, E, ranges for the ultraviolet and visible spectral regions.

Range	λ (nm)	ν (Hz)	$\tilde{\nu}$ (cm^{-1})	E (J)
UV	200–400	1.5×10^{15}–7.5×10^{14}	50 000–25,000	9.94×10^{-19}–4.97×10^{-19}
Visible	400–780	7.5×10^{14}–3.8×10^{14}	25 000–12820.5	4.97×10^{-19}–$2.52 \times 10^{1-19}$

values = 0, 1, 2, and so on, it should be appreciated that a single multielectron configuration can often lead to many values of **L**. The terms are called $S, P, D, F \ldots$ when $L = 0, 1, 2, 3 \ldots$, respectively (Bernath, 1995).

Each electron has spin $s = 1/2$ and the total electron spin of the system can be calculated from the vector sum $\mathbf{S} = \sum_i \mathbf{s}_i$, where $\mathbf{s}_i$ is the electron spin of electron, i. A singlet state arises when the total electron spin is zero. This terminology comes from the definition of the spin multiplicity = $2S + 1$. If the total spin of a state is $S = 1(2S + 1 = 3)$, then the term is called a *triplet state*. One-electron systems or atoms with unpaired electrons can have terms with half-integer total electron spin; that is, $S = 1/2$, $3/2$, $5/2$, and so on, leading to doublet states, quartet state, sextet states, and so on. These arguments hold for molecules as well.

The total angular momentum, **J**, can be calculated in two ways. For light atoms with relatively few electrons (typically, atoms with atomic numbers $Z < 40$), **J** can be deduced from the vector sum of **L** and **S** (Russell–Saunders coupling scheme). Again one term with specific values of **L** and **S** can give rise to many J-levels. Russell–Saunders atomic terms are labeled by ${}^{2S+1}L_J$.

For heavier atoms, the individual electron orbital and spin angular momenta, $\vec{\ell}_i$ and $\mathbf{s}_i$ first vector-couple to form total angular momenta $\mathbf{j}_i$. The total angular momentum for the system is found from the vector sum $\mathbf{J} = \sum_i \mathbf{j}_i$ (j–j coupling scheme). The total angular momentum **J** can have integer or half-integer values depending on whether **S** is integer or half integer, and their numerical value serves as the label for the term.

The parity, I, of a one-electron state can be determined from $I = (-1)^{\ell}$ where ℓ is the orbital angular momentum of the electron involved. For a many-electron system, the parity is determined not by the total angular momentum L, but by a sum of the individual electron orbital angular momenta in the configuration; that is, by $I = (-1)^{\sum_i \ell_i}$. A state or term is considered even or odd depending on whether I is +1 or −1, respectively. The term symbols for an electron state would then be denoted by g (gerade; even) or u (ungerade, odd).

11.2.1.2 **Molecules**

The electronic states of diatomic or linear molecules are usually labeled by the component of orbital angular momentum along the bond axis Λ. When $\Lambda = 0, 1, 2, 3, \ldots$, the states are called Σ, Π, Δ, Φ, $\cdots$, respectively. Because linear systems have cylindrical symmetry, each state with $\Lambda > 0$ is twofold degenerate, corresponding physically to clockwise and counterclockwise electron orbital motion about the molecular axis. This degeneracy can be broken,

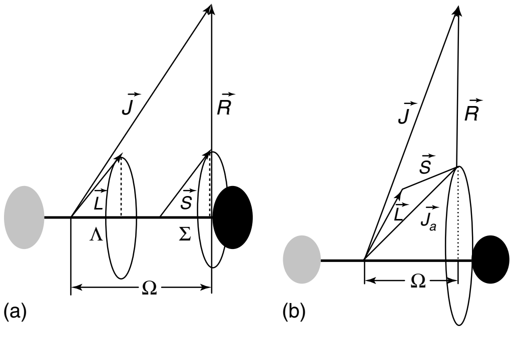

Fig. 11.1 (a) Hund's coupling case (*a*); (b) Hund's coupling case (*c*) for a diatomic molecule. **R**, **L**, **S**, and **J** refer to the rotational, orbital, spin, and total angular momenta, respectively. $\mathbf{J}_a$ is the vector sum of **L** and **S** in Hund's case (*c*). The precession of **J** about the internuclear axis in part a) is not indicated. In part (b) the precession of **L** and **S** about $\mathbf{J}_a$ or **J** about the internuclear axis is not shown. Λ, Σ, and Ω are the projections of **L**, **S** and **J** along the bond axis.

however, by molecular rotation in a process called Λ-*doubling* (Herzberg, 1950). The symmetry of Σ-states with respect to reflection in a plane containing the molecular axis σ_v must be considered. Σ-states that are symmetric with respect to σ_v are labeled Σ^+, while those that are antisymmetric with respect this symmetry operation are called Σ^-. These superscripts are not included for states with $\Lambda > 0$ since they already include +and − components. Molecular terms are denoted by ${}^{2S+1}\Lambda_\Omega$ where $\Omega = |\Lambda + \Sigma|$ is the component of total angular momentum along the molecular axis. Here Σ does not mean $\Lambda = 0$, but is instead the component of electron spin S along the bond axis, which ranges from $+S$ to $-S$ in integer steps.

The labels discussed above are appropriate for Hund's coupling case (*a*) (Figure 11.1a). There are five common coupling cases (*a*)–(*e*) (Herzberg, 1950). Hund's case (*a*) resembles the Russell–Saunders scheme presented above for light atoms. Hund's case (*c*) (Figure 11.1b) is the molecular analog of j–j coupling when spin–orbit coupling is important. Here Λ and Σ are no longer well defined, and the states are labeled only by $\Omega = 0, 1/2, 1, 3/2$, and so on, depending on the total electron spin. Again, the + and − labels must be added to the 0 states, and g and u subscripts included if needed.

The electronic states of nonlinear molecules are labeled by the Mulliken label for their irreducible representation, Γ, within their particular point group (Cotton, 1990), and again the g and u superscripts are used if appropriate.

11.2.2
Electronic Selection Rules

UV/visible transitions are induced when an incident electromagnetic field interacts with the different electric or magnetic multipole moments, $\tilde{\mathrm{M}}$, of the electron distribution of the molecule. Quantum-mechanical selection rules dictate transitions which will be observed. If ψ_e' and ψ_e'' are the upper and lower electronic states, respectively, a particular transition between these states will be allowed if the transition matrix element $\int \psi_e'^{*} \hat{\mathrm{M}} \psi_e'' d\tau_e$ is nonzero (Herzberg, 1966). Here $\hat{\mathrm{M}}$ is the operator corresponding to the particular charge distribution in the molecule, and the integration is over the electronic coordinates, τ_e. The intensity of the transition is proportional to the square of the transition matrix element although this can be very difficult to calculate. As a rule of thumb, the strongest transitions (sometimes referred to as *E1 transitions*) involve the electric dipole moment: $\hat{\mathrm{M}} = -\sum_i e_i \boldsymbol{r}_i$. E2 and $E3$ transitions involve the electric quadrupole and electric octupole moments, respectively, but are considerably weaker than E1 transitions. M1, M2, and M3 transitions involve the magnetic dipole, magnetic quadrupole, and the magnetic octupole moments, respectively, and are also relatively weak. Typically, M1 and E2 transitions are comparable in intensity. Both require high sensitivity to be detected, and usually will only be observed in frequency regions where E1 transitions are forbidden. The rest of this chapter will focus on single-photon electric dipole transitions.

There are additional symmetries that govern selection rules. Single-photon electric dipole transitions are allowed when there is a change in parity. The selection rule governing parity, the so-called Laporte selection rule, can be written for one-electron systems as $\Delta\ell = \pm 1$. For many-electron systems, the selection rule becomes $\Delta L = 0, \pm 1$.

Electron spin does not usually change during an electronic transition that leads to a $\Delta S = 0$ selection rule. The spin selection rule can be relaxed, however, when the spin–orbit interaction becomes important. Then the selection rules for **L** and **S** becomes less rigorous, and those for the total angular momentum, **J** must be considered. It can be shown that the selection rules are $\Delta J = 0, \pm 1$; $J' = 0 \leftrightarrow J'' = 0$ forbidden. The selection rule forbidding single-photon transitions between states with $J = 0$ can be understood as a consequence of conserving angular momentum in an absorption process involving photons that have their own intrinsic nonzero angular momentum.

The selection rules for linear molecules in Hund's case (*a*) are $\Delta\Lambda = 0, \pm 1$, $\Delta S = 0$, $\mathrm{g} \leftrightarrow \mathrm{u}$, $\left\{\begin{matrix}+\\-\end{matrix}\right\} \leftrightarrow \left\{\begin{matrix}+\\-\end{matrix}\right\}$. In Hund's case (*c*) the rules are $\Delta\Omega = 0, \pm 1$, $\mathrm{g} \leftrightarrow \mathrm{u}$, $\left\{\begin{matrix}+\\-\end{matrix}\right\} \leftrightarrow \left\{\begin{matrix}+\\-\end{matrix}\right\}$.

In general, a molecular transition will be allowed if the direct product of the irreducible representations of the ground-state wave function, excited-state wave function, and the multipole moment contains the totally symmetric representation of the particular point group, $\Gamma(\mathrm{A}_1)$, that is, $\Gamma(\psi_e') \otimes \Gamma(\hat{\mathrm{M}}) \otimes \Gamma(\psi_e'') \subset \Gamma(\mathrm{A}_1)$. Equivalently, a transition will be allowed if the direct product $\Gamma(\psi_e') \otimes \Gamma(\psi_e'') \subset \Gamma(\hat{\mathrm{M}})$.

11.2.3
Born–Oppenheimer Approximation and the Franck–Condon Principle

Vibrations play a significant role in determining the intensity of transitions between molecular electronic states (Hollas, 1996). An appropriate starting point is the Born–Oppenheimer approximation, which allows the electronic and nuclear motions to be considered separately because of the large mass differences between the nuclei of a molecule and an electron. Within this approximation, the total wave function of a state, Ψ, can be expressed as $\Psi(r, R) = \psi(r, R)\chi(R)$ where χ is a nuclear wave function that depends on nuclear coordinates $\{R\}$ and ψ is an electronic wave function that depends on the electronic coordinates $\{r\}$ and parametrically on the position of the nuclei. The electronic wave function is therefore found by solving the Schrodinger wave equation at fixed nuclear positions.

A potential energy surface can then be constructed by finding the electronic energy at different nuclear coordinates.

The Franck–Condon principle states that electronic transitions will occur on a much faster timescale than that it takes the nuclei to adjust to their new equilibrium positions. A single-photon electric dipole transition matrix element is defined as $\int \Psi' \hat{\mu} \Psi'' drdR$, where Ψ' and Ψ'' are the wave functions of the excited and ground state, respectively, and $\hat{\mu}$ is the electric dipole operator. As a consequence of the Born–Oppenheimer approximation, the matrix element can be written as a product of an electronic component $R_e = \int \psi' \hat{\mu} \psi'' dr$ and a vibrational overlap integral $S_{v',v''} = \int \chi'_{v'} \chi_{v''} dR$. The intensity of an electronic transition will only be appreciable if the electronic transition probability $|R_e|^2$ is nonzero. This requires that the transition obey the appropriate selection rules for the orbital and spin quantum numbers. The square of $S_{v',v''}$, called the *Franck–Condon factor,* however, modulates the intensity distribution of the vibronic bands observed for a specific electronic transition.

For the sake of simplicity, the Franck–Condon principle is illustrated here for diatomic molecules because they have only one vibrational motion (bond stretching), and therefore their potential energy surfaces reduce to potential energy curves. Quantum mechanically, except for the lowest vibrational level ($v = 0$), which has its maximum wave function amplitude at the equilibrium bond length of the molecule in that electronic state, the largest amplitudes of the vibrational wave functions for $v > 0$ are found at the classical turning points. These correspond to bond lengths where the total energy is potential energy only; that is, the turning points are bond lengths where the nuclear velocities are zero and classically the bond is either at its maximum extension (outer turning point) or its maximum compression (inner turning point). As shown in Figure 11.2(a) the ($v' = 0$, $v'' = 0$) vibronic transition will have the largest Franck–Condon factor (vibrational wave function overlap) when the ground-state equilibrium bond length r''_e, is $\approx$ the excited equilibrium bond length r'_e. The Franck–Condon principle predicts that this transition will be the most intense feature in the absorption spectrum. However when $r'_e > r''_e$ (Figure 11.2b) or $r'_e < r''_e$ the intensity maximum will no longer be the (0,0) transition. In both cases, the spectrum will exhibit a (v',0) progression with a Gaussian intensity distribution maximizing at a value of v', which depends on Δr_e.

In some molecular systems, the Franck–Condon factors are largest for vibronic transition where $v' \gg 0$. One interesting example is the transitions between the ground and ion-pair states of the diatomic halogens. There are 20 ion-pair states of a diatomic halogen XY (X and Y = Cl, Br, I) that dissociates to X^+ (^3P, ^1D, ^1S) + Y^- (^1S). The strongest transitions from their ground states terminate at the first tier $D0_u{}^+$ or $E0^+$ state depending on whether the halogen is homonuclear or heteronuclear, respectively. A molecular orbital analysis shows that these excited ion-pair states have antibonding character. However since the long-range binding is coulombic in nature, the resultant excited states have deeper potential energy curves than the ground states, longer bond lengths, and smaller vibrational frequencies. As a result, the most intense Franck–Condon transitions from

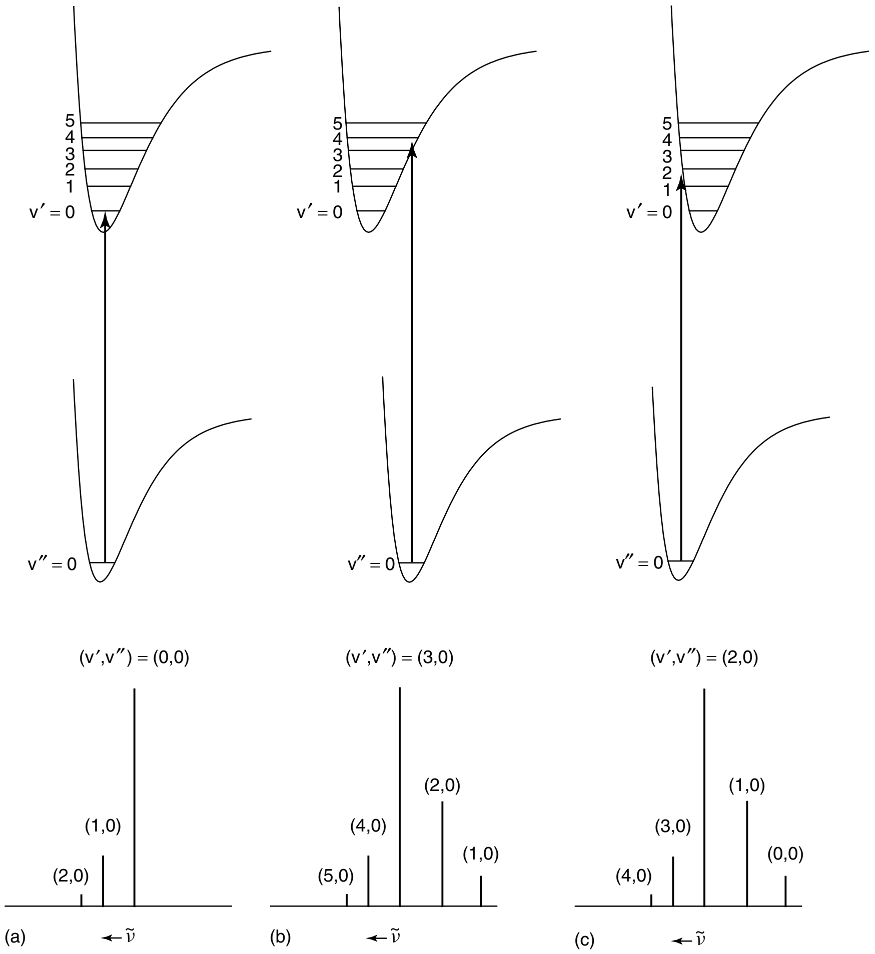

Fig. 11.2 Potential energy curves and resultant stick spectra for the case where (a) $r_e' \approx r_e''$; (b) $r_e' > r_e''$, and (c) $r_e' < r_e''$.

$v'' = 0$ of the IBr ground state, for example, probe the inner wing of the excited E0$^+$ state potential energy curve in the vacuum UV spectral region exciting high vibrational levels between $v' = 260$ and 383 (Lipson and Hoy, 1989).

In general, a vibronic transition will be allowed if $\int \psi_e'^{*} \chi_{v'} \hat{M} \psi_e'' \chi_{v''} d\tau_r d\tau_R$ is nonzero, where $\hat{M}$ is the electric dipole operator. This condition holds if $\Gamma(\psi_e') \otimes \Gamma(\chi_{v'}) \otimes \Gamma(\hat{M}) \otimes \Gamma(\psi_e'') \otimes \Gamma(\chi_{v''}) \subset \Gamma(A_1)$. Within the Born–Oppenheimer approximation, the vibronic transition matrix element will be a product of $3N - 6$ (nonlinear molecules) or $3N - 5$ (linear molecules) vibrational overlap integrals. At room temperature, the ground-state population resides primarily in $v'' = 0$ of the normal modes. These states transform as the total symmetric representation of the molecular point group. If the excited-state vibration is totally symmetric the selection rule on v is $\Delta v_i = 0, \pm 1, \pm 2, \ldots$. If the vibration is nontotally symmetric the Franck–Condon factor vanishes when $\Delta v_i = \pm 1, \pm 3, \pm 5, \ldots$ because $\Gamma(\chi_{v'}) \otimes \Gamma(\chi_{v''})$ does not contain the totally symmetric representation. The selection rule therefore becomes $\Delta v_i = \pm 2, \pm 4, \pm 6, \ldots$.

A series of vibronic transitions originating from a common lower vibrational state and probing a range of excited vibrational states is called a *progression*. For example, the progression originating from the lowest ground-state vibrational level would be ($v' = 0, 1, 2, \ldots, v'' = 0$), and the intensities of the observed bands are determined by their Franck–Condon factors. Vibronic transitions where $\Delta v = v' - v'' =$ constant (a positive or negative integer) form *sequences*. For example, a $\Delta v = 0$ sequence would be made up of the (0,0), (1,1), (2,2), ... transitions. Since the population in $v'' > 0$ is expected to be small for most normal modes, hot-band progressions (originating from $v'' > 0$) and sequence transitions involving $v'' > 0$ are not expected to be particularly intense.

11.2.4
Rotational Fine Structure

Usually, rotational fine structure can only be resolved for molecules in the gas phase, where, typically, the linewidths found in UV/visible spectra are Doppler-broadened. Rotationally resolved spectra are much more difficult to obtain for liquid- or solid-phase samples where other line broadening mechanisms are operative.

A full discussion regarding the rotational fine structure associated with electronic transitions of molecules is beyond the scope of this chapter. Instead, a short discussion pertaining to diatomic molecules is presented here. The rotational selection rules for diatomic molecules (Hund's case (*a*), when $\Delta\Lambda = 0, \pm 1$) are $\Delta J = J' - J'' = 0, \pm 1$ where J' and J'' are the upper and lower state rotational quantum numbers. This leads to three branches labeled P ($\Delta J = -1$), Q ($\Delta J = -1$), and R ($\Delta J = +1$). However, only P and R branches are found for electronic transitions where $\Delta\Lambda = 0$, $\Lambda' = \Lambda'' = 0$. This can be understood by conservation of angular momentum arguments. On the other hand, if $\Delta\Lambda = 0$, $\Lambda' = \Lambda'' \neq 0$, the three branches are allowed but the Q-branch will be weak, while the P and R branches will be strong.

The wave number ($\tilde{\nu}$) expressions for these branches ignoring centrifugal distortion effects, are given by

$$\begin{aligned}
\tilde{\nu}_P &= \tilde{\nu}_o - (B_{v'} + B_{v''})J'' \\
&\quad + (B_{v'} - B_{v''})J''^2 \\
\tilde{\nu}_Q &= \tilde{\nu}_o + (B_{v'} - B_{v''})J''(J'' + 1) \\
\tilde{\nu}_R &= \tilde{\nu}_o + 2B_{v'+} + (3B_{v'} - B_{v''})J'' \\
&\quad + (B_{v'} - B_{v''})J''^2
\end{aligned} \tag{11.1}$$

where $B_{v'}$ and $B_{v''}$ are the upper and lower state rotational constants in states v' and v'', respectively, and $\tilde{\nu}_o$ is the "pure" vibronic band origin. A B_v-value is inversely proportional to bond length, r_v, in a particular vibronic state, and hence a detailed rotational analysis can provide quantitative details about the upper and lower state vibronic geometries. These values can be used to find the equilibrium structure of the molecule.

The equations in Eq. (11.1) are quadratic in J. If $B_{v'} > B_{v''}$, that is, $r_{v'} < r_{v''}$, the rotational line spacings will decrease in the P-branch and increase in the R-branch. At some value of

J'' a band head will form, which is characteristic edge structure (in low resolution) composed of many rotational lines. The observed vibronic bands will be degraded to the higher energies (blue-shaded). Conversely, if $B_{v'} < B_{v''}$, that is, $r_{v'} > r_{v''}$, the spacings will decrease in the R-branch and increase in the P-branch. The resultant vibronic bands will then be red-shade.

11.3
Absorption Spectra

11.3.1
Beer–Lambert Law

In general, a UV/visible absorption spectrum is found by recording a plot of transmitted light versus wavelength. Experimentally, the amount of absorption or absorbance, A, is determined by measuring the transmitted light intensity, I, relative to the incident light intensity I_o using

$$A = -\log_{10}\left(\frac{I}{I_o}\right) \tag{11.2}$$

The absorbance, which is unitless and is sometimes called the *optical density*, is proportional to the number of molecules in the path of the beam, and can therefore in the solution phase be calculated using the Beer–Lambert law:

$$A = \varepsilon(\lambda)[C]\ell \tag{11.3}$$

where $\varepsilon(\lambda)$, which is the absorption coefficient of the sample at a specific wavelength, λ, $[C]$ is the sample concentration, and ℓ is the path length. ε is also called the *molar absorptivity* or *molar extinction coefficient* and has units of L mol^{-1} cm^{-1} when $[C]$ and ℓ are expressed in molarity and centimeters, respectively. An analogous expression can be written for gas-phase samples using number densities (in molecules per centimeter cubed) instead of concentrations. Then, the absorption coefficient is identified as an absorption cross section, σ (centimeter squared). The molar extinction coefficient is related to σ by

$$\varepsilon = \sigma(N_0 - N_1) \tag{11.4}$$

where N_0 and N_1 are the ground-state and excited-state number densities, respectively.

Typically, ε (in liters per mole per centimeter) can vary between 0 and 10^6 where values $\geq 10^4$ are related to high-intensity absorptions while values $<10^3$ are considered low-intensity absorptions. Forbidden transitions will have molar absorptivities ranging from 0 to $\sim 10^3$.

If more than one species exists in solution, the overall absorbance will be equal to the sum of the absorbances of the individual species. For an N-component mixture, A can be written as

$$A(\lambda) = \ell \sum_{i=1}^{N} \varepsilon(\lambda)[C_i] \tag{11.5}$$

The spectrum of a solution containing a pair of absorbing species can exhibit one or more isosbestic points, which correspond to wavelengths where the molar absorptivities of the species are identical. Even when a 1 : 1 reaction occurs the absorbance of a solution of the two reactants and product at an isosbestic point will remain invariant to the degree of reaction; that is, independent of their particular concentrations.

The Beer–Lambert law assumes that A will be a constant if the product of $[C]$ and ℓ is constant. However, this may not always be the case. For example, there could be molecular association at higher concentrations, or the species could undergo photochemistry upon irradiation. The Beer–Lambert law may also fail for acids, bases, and salts that exhibit concentration-dependent ionization, for samples that are strongly fluorescent, or if a thermal equilibrium exists between the ground state and a low-lying excited state. As a result, the linearity associated with the Beer–Lambert law should always be confirmed experimentally before an analysis.

11.3.2
Chemometrics

As noted above the Beer–Lambert law is additive. This allows the composition of a mixture of compounds to be determined if their individual absorption spectra are known. This technique lends itself very well to analysis by computer software. The field where mathematical or statistical methods are applied to data analysis is called *chemometrics*. The basic approach can be illustrated by considering a mixture of three compounds, a, b, and c with concentrations $[C_a]$, $[C_b]$, $[C_c]$ respectively (Rouessac and Rouessac, 2007). The absorbance of this mixture is measured at three wavelengths λ_1, λ_2, and λ_3 resulting in three measurements A_1, A_2, and A_3, respectively. By using a path length of $\ell = 1$ cm the absorbances at the three wavelengths can be written in matrix form as

$$\tilde{\mathrm{A}} = \tilde{\varepsilon}\tilde{\mathrm{C}} \tag{11.6}$$

where

$$\tilde{\mathrm{A}} = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix}, \quad \tilde{\varepsilon} = \begin{pmatrix} \varepsilon_a^1 & \varepsilon_b^1 & \varepsilon_c^1 \\ \varepsilon_a^2 & \varepsilon_b^2 & \varepsilon_b^2 \\ \varepsilon_a^3 & \varepsilon_b^3 & \varepsilon_c^3 \end{pmatrix}$$

$$\text{and} \quad \tilde{\mathrm{C}} = \begin{pmatrix} [C_a] \\ [C_b] \\ [C_c] \end{pmatrix}$$

If the nine absorbances in the $\tilde{\varepsilon}$ matrix are known, $\tilde{\mathrm{C}}$ can then be found by computing $\tilde{\mathrm{A}}\varepsilon^{-1}$. The precision of the composition determination can be improved for mixtures that have similar spectra by using many tens of data points, even though, mathematically, the system of equations in Eq. (11.6) would be overdetermined.

11.3.3
Derivative Spectroscopy

It can be helpful, instead of plotting A versus λ, to mathematically plot the absorbance derivatives $dA/d\lambda$ as a function of wavelength. This can be done for higher order derivatives as well. Such plots serve to amplify weak slope variations of the zeroth-order absorption spectrum. First-derivative plots will pass through zero at the location of each peak maximum, while second-derivative plots produce zero values at wavelengths corresponding to both the minima and maxima of the original absorption spectrum. This technique is useful when the zeroth-order spectrum has a uniformly increasing baseline due to a strong absorption at shorter wavelengths or because of light scattering. It is also helpful when there are weak absorbances that are overshadowed by stronger bands in the original spectrum.

11.3.4
Circular Dichroism Spectroscopy

CD spectroscopy measures the differential absorption of left-hand and right-hand circularly polarized light as a

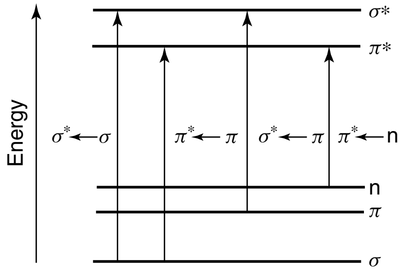

Fig. 11.3 An energy level diagram for a generic organic molecule indicating the possible types of electronic transitions, and their relative energy ordering.

function of exciting wavelength (Berova, Di bari and Pescitelli, 2007). The technique is sensitive to absolute configurations and therefore is able to discriminate different enantiomers (stereoisomers whose mirror images are not superimposable). Formally, a chiral molecule is one that does not possess an improper axis of rotation. CD is defined as

$$CD = A^L - A^R \tag{11.7}$$

where A^L and A^R are the absorptions of left and right-hand circularly polarized light, respectively. The difference in the molar absorption coefficients, $\Delta\varepsilon$, at a specific wavelength can be expressed in terms of CD by

$$\Delta\varepsilon = \varepsilon^L - \varepsilon^R = \frac{CD}{[C]\ell} \tag{11.8}$$

CD values for enantiomers are always equal and opposite, and therefore, when scanning through an absorption band $\Delta\varepsilon$ will change rapidly in one direction, pass through zero at the peak maximum, and then change rapidly in the opposite direction. This is known as the *Cotton effect* after its discoverer. Positive and negative Cotton effects refer to the cases where $\Delta\varepsilon$ is >0 and <0, respectively.

Although electric dipole moments, $\boldsymbol{\mu}_{ij}$, are larger than magnetic dipole moments, $\mathbf{m}_{ij}$, the rotational strength of a CD signal, R_{ij}, which is proportional to $\int \Delta\varepsilon d\nu$, depends on the product $\boldsymbol{\mu}_{ij} \cdot \mathbf{m}_{ij}$. R_{ij} is a signed quantity, and as such can be used to interpret CD spectra and to make configurational assignments. The assignments depend strongly on the chromophores (Section 11.4.1) and their arrangements in the molecule. In general, related complexes that have the same CD sign for a given absorption have the same absolute configuration.

11.4
The Origin of UV/Visible Transitions in Organic Molecules

A generic energy level diagram of an organic molecule is presented in Figure 11.3.

The filled molecular orbitals of an organic molecule are designated as σ-bonding, π-bonding, or nonbonding, n, depending on the nature of the atomic orbitals involved. The most probable electronic transitions will take place from the highest occupied molecular orbital (HOMO) to the lowest occupied molecular orbital (LUMO), which are typically antibonding in nature (σ^* or π^*). In general, the energies E of the possible transitions are ordered: $E(\pi^* \leftarrow n) < E(\pi^* \leftarrow \pi) < E(\sigma^* \leftarrow n) < E(\pi^* \leftarrow \sigma) < E(\sigma^* \leftarrow \sigma)$. Of these transitions, $\pi^* \leftarrow \pi$ and $\pi^* \leftarrow n$ dominate in the UV/visible region.

11.4.1
Chromophores

Although molecular orbitals involved in absorption extend over the entire molecule, it is often possible to identify a specific atom or group of atoms in the molecule as a chromophore; that is, as a functional group that acts as the dominant light-absorbing unit. A list of chromophore properties for the most common functional groups is presented in Table 11.2 (Pavia, Lampman and Kriz, 1996)

Most $\pi^* \leftarrow n$ transitions involving heteroatoms are weak because the spatial orthogonality of the n-bonding orbitals relative to the π^*-system make the transition's symmetry forbidden. Similarly, there is no overlap between σ and π orbitals, and therefore they can be considered separately. The $\pi^* \leftarrow \pi$ transitions listed in Table 11.2 are not particularly intense. However, their transition wavelengths can shift dramatically to longer wavelengths and grow in intensity (hyperchromism) with increasing degree of conjugation owing to a reduction in the HOMO–LUMO energy gap in much the same way that the energy level spacing of a one-dimensional particle in a box decreases with increasing box size. This trend can be examined in Table 11.3 for various polyenes although similar trends have been observed for many other unsaturated systems. The absorption intensity growth with conjugation reflects an increasing transition dipole that lies approximately along the direction of the π-system. The deviations evident in Table 11.3 arise because the polyene backbones are not completely delocalized but, instead, consist of alternating single

Tab. 11.2 Absorption wavelength maxima, λ_{max}, and molar absorption coefficient maxima, ε_{max}, for various transitions of different chromophores.

chromophore	λ_{max} (nm)	ε_{max} (L mol^{-1} cm^{-1})	Transition type
C−C	<180	1000	$\sigma^* \leftarrow \sigma$
C−H	<180	100	$\sigma^* \leftarrow \sigma$
R−OH	180	316	$\sigma^* \leftarrow n$
R−O−R	180	3162	$\sigma^* \leftarrow n$
$R-NH_2$	190	3162	$\sigma^* \leftarrow n$
R−SH	210	1000	$\sigma^* \leftarrow n$
$R_2C=CR_2$	175	1000	$\pi^* \leftarrow \pi$
$R-C\equiv C-R$	170	1000	$\pi^* \leftarrow \pi$
$R-C\equiv N$	160	<10	$\pi^* \leftarrow n$
R−N=N−R	340	<10	$\pi^* \leftarrow n$
$R-NO_2$	271	<10	$\pi^* \leftarrow n$
R−CHO	190	100	$\pi^* \leftarrow \pi$
	290	10	$\pi^* \leftarrow n$
R_2CO	180	1000	$\pi^* \leftarrow \pi$
	280	32	$\pi^* \leftarrow n$
RCOOH	205	32	$\pi^* \leftarrow n$
RCOOR′	205	32	$\pi^* \leftarrow n$
$RCONH_2$	210	32	$\pi^* \leftarrow n$

and double bonds. Furthermore, the absorption $\lambda_{\max}$ and $\varepsilon(\lambda_{\max})$ are sensitive to cis–trans isomerization (Klessinger and Michl, 1995).

Substituents that increase the intensity and often the $\lambda_{\max}$ of absorption peaks when attached to a chromophore are called *auxochromes*. Woodward (1941, 1942) and Fieser, Fieser and Rajagopalan (1948) established a set of empirical rules for dienes that allow the absorption maximum of a compound to be determined for different additional chromophores present. Similar rules have also been deduced for conjugated carbonyl compounds, and mono- and di-substituted benzenes. These rules can be useful for structure determination.

A proper solvent for solution-phase UV/visible spectroscopy should be nonabsorbing at wavelengths absorbed by the solute of interest. A list of common solvents and their short wavelength transmission cutoffs is presented in Table 11.4 (Gauglitz and Vo-Dinh, 2003). While water and ethanol are good solvents for many substances, hexane and other hydrocarbons are appropriate for less polar compounds. The latter solvents also interact more weakly with the solute, allowing a finer structure to be resolved.

Tab. 11.4 Ultraviolet cutoff wavelength for various commonly used solvents.

Solvent	**Cutoff wavelength (nm)**
Hexane, C_6H_{14}	200
Ethanol, C_2H_5OH	210
Water, H_2O	210
Methanol, CH_3OH	210
Acetonitrile, CH_3CN	215
Cyclohexane, C_6H_{12}	215
Chloroform, $CHCl_3$	250
Carbon tetrachloride, CCl_4	280
Benzene, C_6H_6	280
Pyridine, C_5H_5N	310

Tab. 11.3 Absorption wavelength maxima, $\lambda_{\max}$, and molar absorption coefficient maxima, $\varepsilon_{\max}$, for various polyenes.

$H(CH{=}CH)_n H$	$\lambda_{\max}$ (nm)	ε ($L\,mol^{-1}\,cm^{-1}$)
1	162	10 000
2	217	21 000
3	258	35 000
4	296	52 000
5	335	118 000
8	415	210 000
11	470	185 000
15	547	150 000

Solvent polarity can play a role in determining the position of peak maxima in an absorption spectrum. This effect, called *solvatochromism*, arises because the HOMO and LUMO orbitals involved in a UV/visible absorption have different polarities and therefore interact differently with the solvent. Bathochromic (red) shifts result when the LUMO is stabilized in the solvent relative to the ground state. This effect is common for $\pi^* \leftarrow \pi$ transitions. Conversely, hypsochromic (blue) shifts arise when the HOMO is preferentially stabilized by the solvent. This type of shift is commonly observed for $\pi^* \leftarrow n$ transitions.

Peptide bonds have intense $\pi^* \leftarrow \pi$ transitions between 190 and 210 nm while aromatic amino acids absorb strongly between 205 and 280 nm (Table 11.5, Gauglitz and Vo-Dinh, 2003). The $\lambda_{\max}$ values for the UV absorptions of tryptophan, tyrosine and phenylalanine come between 255 and 280 nm, and are often used to determine

Tab. 11.5 Absorption wavelength maxima, λ_{max}, for various amino acid residues.

Amino acid	**Structure**	**λ_{max} (nm)**
Histidine	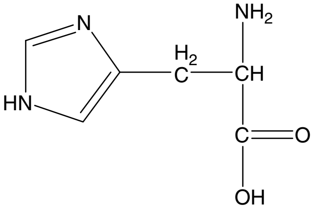	210
Tryptophan	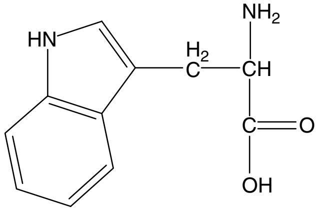	220, 280
Tyrosine	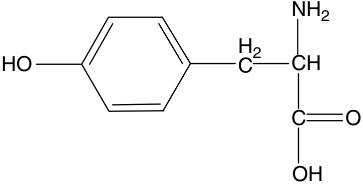	195, 222, 275
L-Cysteine	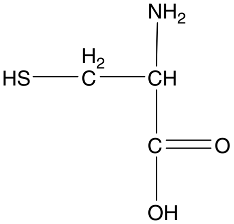	235
Phenylalanine	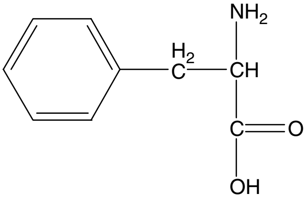	190, 205, 255

the total protein concentration in solution. Furthermore these wavelengths tend not to degrade the sample. The phenolic chromophore on tyrosine is sensitive to the pH of the solution in that its absorption maximum shifts to the red with increasing solution acidity. Similarly, increases in the nonpolar environment will also red-shift the absorption maxima of tyrosine and tryptophan. The UV/visible spectra of these residues can therefore be used to probe the conformational differences of different protein states.

11.5
The Origin of UV/Visible Transitions in Inorganic Molecules

Transition metal complexes are known to be important in biological systems, as catalysts, as components of solar cells, as light emitters, nonlinear media, and so on. One of their most striking features is the wide variety of colors they can exhibit both as solids and in solution. The energies of the five d-orbitals for a given principal quantum number $n \geq 3$ in a free atom or ion are essentially degenerate. Upon ligation, however, the degeneracy of the d-orbital set is lifted in a manner determined by the nature of the ligands and the symmetry of the complex. Although ligand-field theory (through a group theoretical analysis) is the appropriate formalism needed to understand the nature of the molecular orbitals that arise because of coordination, the simpler crystal-field theory also provides insight into the visible absorptions of the complexes (Figgis, 1966)

Consider a ML_6 octahedral complex where M and L are the metal and ligand, respectively. In crystal-field theory, the ligands are approximated as point-negative charges although similar results can be deduced even if the ligands are neutral but dipolar. If the metal ion is placed at the origin of a coordinate system, the d_{z^2} and $d_{x^2-y^2}$ orbitals form a twofold degenerate set (with e_g symmetry) that have lobes that point along the axes directly toward the ligands, which are raised in energy by Coulombic repulsion. The remaining orbitals (d_{xy}, d_{yz}, d_{zx}) form a threefold degenerate set with t_{2g} symmetry that lies lower in energy than $\{e_g\}$ because they are orientated between the axes and therefore, do not point directly at the ligands. Ligand-field theory also predicts this splitting pattern. However, ligand-field theory shows explicitly that the e_g orbitals unlike the t_{2g} set are not pure metal d-orbitals but have ligand character as well.

The d-orbital splitting patterns for complexes of lower symmetries can be deduced by removing specific ligands, and noting if this yields a net stabilization of some of the d-orbitals relative to the others. As shown in Figure 11.4, for example, forming a square planar ML_4 complex by removing the 2 ligands orientated along the z-axis lowers the energy of the d_{z^2} orbital in the e_g set relative to $d_{x^2-y^2}$, as well as the energy of the d_{yz} and d_{xz} orbitals relative

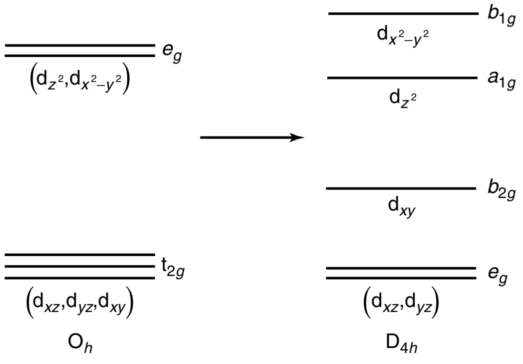

Fig. 11.4 Crystal-field splitting of the d-orbitals in an octahedral (Oh) and square planar (D_{4h}) geometry. The d-orbitals are labeled by their irreducible representations in the appropriate point group.

to d_{xy} in the t_{2g} set. Thus, the energy ordering of the d-orbitals in such a case might be $E(d_{xy}, d_{yz}) < E(d_{z^2}) < E(d_{xy}) < E(d_{x^2-y^2})$. Alternatively, the energy of d_{z^2} may even fall below that of the d_{xy}, d_{yz} pair depending on the metal ion and the nature of the ligands.

The splitting between the e_g and t_{2g} set in an octahedral complex is denoted by Δ_o and its magnitude is a strong function of the ligands making up the complex. This splitting is not trivial. For example, Δ_o for the $Ti(H_2O)_6{}^{3+}$ ion is 20 000 cm^{-1}. Ligands that yield a large Δ_o value are called *strong-field ligands* while those that result in small splittings are called *weak-field ligands*. The ligands can be empirically ordered in this regard for a given metal and geometry in a so-called spectrochemical series, which can be rationalized by considering ligand electronegativities, polarizabilities, permanent dipole moments, π-bonding characteristics, and so on. A partial list from strong field to weak field is given by

$$\begin{aligned} CO \approx CN^- &> NO_2{}^- > NCS^- > H_2O \\ &> OH^- > F^- > SCN^- > Cl^- \\ &> Br^- > I^- \end{aligned}$$

The magnitude of Δ_o also depends on the metal ion and its charge when the geometry and the ligands are held constant, although the rationale for the ordering is not as obvious. A partial spectrochemical series ranging from strong-field to weak-field ions is give by

$$\begin{aligned} Pt^{4+} &> Ir^{3+} > Rh^{3+} > Co^{3+} > Cr^{3+} \\ &> Fe^{3+} > Fe^{2+} > Co^{2+} > Ni^{2+} \\ &> Mn^{2+} \end{aligned}$$

The d-orbitals fill according to Hund's rules, which states that the terms of highest spin multiplicity lies lowest in energy. Electrons will spin-pair in the t_{2g} set before filling the e_g orbitals if the magnitude of Δ_o is large (strong-field ligands). Complexes of this sort are termed *low-spin systems*. Conversely, electrons will fill both the t_{2g} and e_g orbitals before spin-pairing if the ligands are weak field. These complexes are designated as high-spin systems.

UV visible transitions in a ML_6 complex involve promoting a d-electron from t_{2g} to e_g. Formally, however, d–d transitions are electric-dipole Laporte forbidden. However, this rule can be relaxed through vibronic coupling because the total wave function of a term includes not only an electronic component but also a vibrational normal mode. If that normal mode has the correct symmetry (in this case u-symmetry) the transition can become allowed. This is known as the *Herzberg–Teller effect*. Physically, the vibrational motion breaks the inversion symmetry of the molecule. Even so, octahedral complexes tend to have small molar absorptivities. On the other hand, the Laporte selection rule does not apply to noncentrosymmetric metal complexes geometries such as tetrahedral or trigonal bipyramidal, and so the d–d transitions tend to be stronger although the absorption coefficients are only in the range $\varepsilon \approx$ 100–200.

The strongest metal complex absorptions usually lie in the UV and can be assigned to metal–ligand charge-transfer ($g \leftrightarrow u$) transitions. If an electron is transferred from an orbital lying principally on the ligand to an orbital lying principally on the metal, or vice versa, the transitions are termed a *ligand-to-metal* or *metal-to-ligand* charge transfer, respectively. These can occur for d^0 and d^{10} systems as well as open d-shell complexes. Indeed, all polyatomic molecules exhibit such transitions, although they often come at much shorter wavelengths than the UV/visible. A charge-transfer transition, however, can occur at longer wavelengths if the metal is easily oxidized and the ligand easily reduced, or vice versa. These absorptions tend to be very intense and can often mask the crystal-field transitions. Notable in this regard are metal iodide complexes, because I^- is readily oxidized. Typical ε values for octahedral hexahalo anions range from ~600 to 60 000 (Lever, 1968). Charge-transfer transitions can be used in the analysis of ions such as Fe and Mg in biologically important systems such as hemoglobin and chlorophyll, respectively.

11.6
The Origin of UV/Visible Transitions in Solids

While transition metal crystal-field splittings are also responsible for many of the striking colors observed in mineral gemstones, the broad subject of UV/visible transitions in solids is limited here to two additional examples – semiconductors and color centers.

UV/visible radiation can excite many semiconductors from their valence band to their conduction band. These can be classified as direct or indirect (phonon-assisted) transitions. A partial list of direct band-gap transitions for some binary semiconductors, and their band gaps and absorption edges at room temperature are given in Table 11.6 (Berger, 1997).

Tab. 11.6 Electronic band gap and wavelength absorption edge for several common semiconductor solids.

Semiconductor	Band gap (eV)	λ_{edge} (nm)
ZnO	3.2	387
ZnS	2.8	443
ZnTe	2.4	517
CdS	2.4	517
HgS	2.1	590
CuCl	3.4	365
CuBr	2.9	428
CuI	2.9	428
AgI	2.9	428
BN	5.8	214
AlN	6.2	200
GaN	3.4	365
InN	2.0	605

Band gaps increase in energy (decrease in wavelength) at lower temperatures. Indirect band-gap materials tend to have shorter wavelength absorption edges. Furthermore, if defect levels are located within the band gap they can absorb light at longer wavelengths than the band edge.

Color centers are formed by first producing vacancies in a crystal lattice, usually by irradiating the material with X-rays. Electrons or holes are then trapped in the lattice defects. The number of color centers produced is a function of the irradiation time. Depending on the crystal, color centers can persist for a period of time ranging from days to years. The alkali halide crystals have been the subject of considerable scrutiny in this regard (Schulman and Compton, 1962), although defects can also be generated in minerals including diamond. In an F-center, one or more free electrons located in an anionic vacancy become trapped within the electrostatic field of the enclosure. They can therefore undergo energy level transitions by the absorption of UV/visible wavelengths, thus imparting color to the crystal. Defect electronic transitions in alkali halides have been the basis of several commercial infrared lasers (Silfvast, 1996).

11.7
Experimental Techniques for UV/Visible Absorption Spectroscopy

In addition to the sample, the basic UV/visible absorption spectrometer consists of three parts: a light source, a wavelength dispersive instrument such as a monochromater, and a detector. Each is considered separately.

11.7.1
Light Sources

The most common UV/visible light sources can be grouped into two categories: thermal sources and luminescent sources. The merits of light-emitting diodes (LEDs) and synchrotrons are also discussed (Schmidt, 2005).

11.7.1.1 Thermal Sources

Thermal sources convert electrical power into heat and emit light according to Planck's Law. The simple incandescent light bulb is the most common example. In these devices a tungsten filament is placed in an evacuated glass bulb and is resistively heated by an electric current. The resultant Planck Law temperature is typically ~2500 K. Over time, small amounts of the tungsten will vaporize from the filament, coating and blackening the bulb. This effect can be reduced by adding an inert gas to the bulb such as Ar, Kr, or Xe.

Halogens lamps essentially operate the same way as incandescent lamps except here a halogen gas (usually a bromide) is added to the bulb. This additive will react with the evaporated tungsten vapor to form a tungsten bromide compound that falls back onto the filament where it is converted to elemental tungsten. The sizes of these sources, which operate at T ~3450 K, are usually small, and therefore, make good point-light sources.

11.7.1.2 Luminescent Sources

The mechanism behind luminescent sources involves electronic excitation of atoms that emit light upon relaxation. Gas-discharge lamps are the most common devices in this group and, depending on the pressure of the device, they can operate as continuum sources (broad wavelength outputs) or as atomic line sources (specific wavelengths).

Hg lamps consist of a bulb containing an inert gas and a trace amount of mercury. When a high voltage is applied and the gas breaks down, electrons are accelerated to kinetic energies that cause the rare gas to be ionized. Ion–electron recombination excites the rare gas, which subsequently transfers its energy to the Hg atoms, which, in turn, emit light. At low pressures ($\sim 10^2$ Pa), most the light comes out at the Hg resonance line at 253.7 nm. At medium pressure ($\sim 10^5$ Pa) the output has many spectral lines, and the device usually requires higher currents to operate. At high pressures ($> 3 \times 10^6$ Pa) the spectral output consists of broad overlapping bands that extend from the UV to the near infrared. However, the flux in the red and blue–green regions of the visible spectrum is relatively small.

Xe arc lamps operate at a Planck's Law temperature of ~6000 K. They put out less radiant energy than an Hg lamp in the UV region, but more in the visible region. Their chief strength is that the spectral output is relatively smooth in the visible although there are some spectral lines in the blue and near infrared regions.

Among the best continuum UV light sources are the hydrogen lamps where emission results from the recombination reaction: $H_2 \leftarrow 2H$, at a surface. Heat conduction can be minimized in the device and the radiant energy increased by using deuterium instead of hydrogen. The emission from these sources maximizes at 290 nm. Often, the one strong atomic H or D Balmer series emission line at 486.12 or 485.99 nm, respectively, can be used for wavelength calibration.

In most spectrophotometers, two lamps are needed to cover the entire UV/visible spectral range. Typically, a D_2 lamp is used for the UV and a halogen lamp is used for the visible ranges. There is an obvious complexity to the instrumentation that will arise when one lamp is switched to the other.

11.7.1.3 Light-Emitting Diodes

LEDs are solid-state devices where the recombination of electrons and holes generates light. Depending on the semiconductor material, they can emit light between ~350 and 1550 nm. While their efficiencies and radiant energies are usually low, they have long lifetimes and are very inexpensive. White LEDs have now been developed, which operate on one of two mechanisms. The most dominant approach is to combine a 450–470-nm gallium nitride (GaN) LED with an yttrium aluminum garnet (YAG) phosphor. The blue output of the GaN LED causes the phosphor to glow white out to ~700 nm. The second method involves combining the outputs of red, green, and blue LEDs in the proper proportion to obtain white light.

11.7.1.4 Synchrotrons

Synchrotron emission from an electron storage ring provides wavelength coverage from the infrared to γ-rays. Each pulse of synchrotron light is highly polarized and short ($<10^{-10}$ seconds), and the repetition rate of the pulse train can reach 5×10^8 s^{-1}. Despite this wavelength versatility, synchrotrons are "big science" and as such, would probably not be the first choice for work in the UV/visible where cheaper tabletop sources abound. They are, however, indispensable for high-resolution spectral studies and crystallography in the X-ray region.

11.7.2
Wavelength Dispersive Instruments

Dispersive instruments spatially separate a light beam on the basis of wavelength. The simplest optic for performing this feat is a prism, usually made of quartz or glass, whose dispersive power is a function of its wavelength-dependent index of refraction $n(\lambda)$. A geometrical analysis shows that the wavelength-dependent beam deflection, δ through a prism is given by $d\delta/d\lambda = -B/H\, dn/d\lambda$, where $dn/d\lambda$ is the change in index of refraction with wavelength, B the dimension of the base of the prism, and H is the height of the incident beam.

Higher resolution dispersion can be obtained using diffraction gratings. Diffraction gratings are ruled or holographically produced substrates containing several hundred to $>10^3$ lines per millimeter. A grating gives rise to wavelength-dependent interference patterns by Bragg diffraction when illuminated. The directions of constructive interference, or intensity maxima, arise when the path differences between scattered rays are an integer multiple of the wavelength. These integers, m, are called *diffraction orders.*

Gratings can operate in transmission (slits) or in reflection (grooves), and their resolving power R is defined as

$$R = \frac{\lambda}{d\lambda} \tag{11.9}$$

R is a measure of how well it can separate two close-lying wavelengths, and is related to the number of grooves, N through

$$R = mN \tag{11.10}$$

Although the linear dispersion (spatial separation in a plane), $d\lambda/dx$, will increase with order m, at a fixed grating orientation light, for example, at 800 nm in first order, will overlap 400 nm light in second order, 266.7 nm light in third order, and so on. The different orders can be separated using filters, or, alternatively, by modifying (blazing) the shape of the grooves so that the scattered energy is concentrated primarily into one order.

Diffraction gratings are the main dispersion elements in monochromators. These instruments are used to measure the intensity of different wavelengths sequentially as a grating is rotated. The two common geometries used for UV/visible spectroscopy are the Fastie–Ebert configuration and a Czerny–Turner design.

11.7.2.1 **Fastie–Ebert Configuration**

The optical arrangement of a Fastie–Ebert spectrograph is shown in Figure 11.5. The main components are a spherical mirror and one plane diffraction grating. The entrance and exit slits of the instrument are both placed at the focal plane of the spherical mirror. Light entering the first slit and reflected off the mirror is dispersed by the grating. The

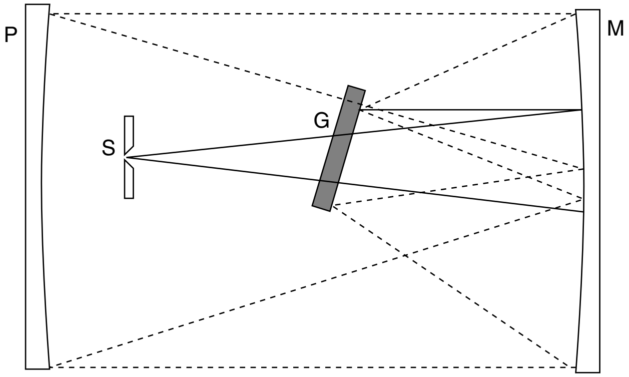

Fig. 11.5 A schematic of a Fastie–Ebert spectrograph. P, photographic plate; S, slit; G, grating; and M, concave mirror. The solid and dashed lines indicate the light path above and in line with the grating.

light is then rereflected and collimated by the spherical mirror to form an image at the exit slit. A UV/visible spectrum can be obtained by detecting the output light intensity photographically, or with a detector at the exit slit as a function of grating rotation.

11.7.2.2 **Czerny–Turner Configuration**

Czerny–Turner monochromators operate in much the same manner as the Fastie–Ebert design, in that the entrance and exit slits are fixed, and the monochromator is tuned by rotating a plane diffraction grating. The main difference is that the single spherical mirror in the former design is now replaced by two concave mirrors. This provides more flexibility in the geometry of the mirrors and can eliminate certain optical aberrations such as coma. This is also the preferred design when designing larger instruments where a single spherical mirror with good optical properties may be difficult to manufacture.

11.7.2.3 **Single-Beam Versus Double-Beam Spectrometers**

Single-beam spectrometers disperse one beam of light passed from a source; the beam is then passed through the sample and onto the detector. The measurement can be then repeated for a reference solvent. Double-beam spectrometers, on the other hand, use a split dispersed light beam that is passed through the sample and reference solvent at the same time (Skoog, 1985).

Single-beam instruments are technically simpler and often produce spectra with a better signal-to-noise ratio because the full radiant light energy of the lamp is used. However, the quality of the spectrum is also determined by the constancy of experimental conditions such as the output power of the lamp and electrical noise.

Conversely, dual-beam instrument operation compensates for lamp and electrical fluctuations, but the spectra often have poorer signal-to-noise ratios because only half the light beam is used for either the sample or the reference solvent. The instruments are also technically more sophisticated.

In general, single-beam instruments are adequate for absorbance measurements done at a fixed wavelength, while dual-beam instruments may be more appropriate for recording absorption spectra over a broad wavelength range.

11.7.3
Detectors

Many UV/visible detectors have been developed. A short survey of the most common devices is presented here.

11.7.3.1 **Photodiodes**

UV/visible light can be measured using an Si-based device known as a *reverse bias photodiode*. Here a voltage depletes the conductance at the junction of the p–n junction to nearly zero. When light falls on the depleted layer, holes and electrons are formed, which provide a current proportional to the light intensity. The equivalent quantum efficiency of a photodiode is considerably higher than that of a photomultiplier (PMT) tube (Section 5.3.2), but has a lower sensitivity because there is no signal amplification. The diode sensitivity can be increased using a p–i–n or PIN structure. Here an intrinsic (undoped) layer, i, is placed between the p- and n-doped regions, which, when illuminated, also creates hole and electrons that contribute to the photocurrent. The quantum efficiency of this arrangement is larger because the depletion layer is thicker. The output of a diode provides an indication of the impinging light energy over a time interval.

11.7.3.2 **Photomultipliers (PMTs)**

Many UV/visible detectors rely on the photoelectric effect to convert optical signals to an electrical output. Here, a material (termed a *photocathode*) with an appropriate work function will eject electrons upon illumination. Typical photocathodes are made of mixed alkalis (Table 11.7, Lytle, 1974). Under an applied voltage the elected electrons are measured as a photocurrent. The quantum efficiency of the photocathode is defined as the probability that a photon will eject an electron, and is usually less than unity.

In a PMT tube, the electrons ejected from the photocathode strike a chain of additional photoemissive electrodes, or dynodes, each at a successively higher positive potential. The first dynode will emit several electrons for each electron striking it. Those electrons cause the second dynode to emit several electrons, rapidly leading to a cascade multiplication of electrons at the end of the chain. The total gain in many devices can exceed 10^6. PMT-tube sensitivity is usually limited by their dark-current emission, but this can be almost eliminated by cooling the tube to low temperatures.

11.7.3.3 **Microchannel Plates (MCP)**

A microchannel is a small glass tube ~1-mm long, with typical inner diameters between 8 and 20 μm, whose inner surface is coated with a semiconductor material (Dereniak and Crowe, 1984). Like a dynode stage of a PMT, the inner surface will emit secondary electrons when struck by a primary electron. The gain in a channel under an applied voltage of $\approx 10^3$ V is on the order of 10^4. Microchannel plates (MCPs) are arrays of these tubes sandwiched between metal electrodes, and there can be between 10^4–10^7 microchannels per plate. The primary electrons come from a photocathode placed in front of the array. The gain can be increased to $\approx 10^7$ by placing two MCPs in series. Each channel can produce an intensified picture element

Tab. 11.7 Absorption maximum, λ_{max}, and quantum efficiency for several typical photocathode materials.

Nominal Composition of Photocathode material (response[a])	λ_{max} (nm)	Quantum efficiency (%)
Cs_8Sb (S-11)	440	16
Cs_8Sb (S-4)	400	12.4
K–Cs–Sb (117)	400	25
K–Cs–Sb (115)	400	25
(Cs)Na_2KSb (S-20)	420	20
(Cs)Na_2KSb (112)	420	18

[a] trade name; all "S" numbers are (Radio Corporation of America) RCA designations.

or pixel in an imaging system or the total current can be collected and plotted as a function of wavelength to yield a spectrum.

11.7.3.4 Charge Coupled Devices (CCD)

Although charge coupled devices (CCDs) were first invented as a computer memory circuit, they have become prevalent as light detectors in imaging applications. CCD cameras are either one-dimensional or two-dimensional pixel arrays of metal oxide semiconducting capacitors (based on Si) (Eastman Kodak, 2001). The number of electrons and holes produced within a pixel by an incident light beam is linearly proportional to the light level and exposure time at that location. There is also a nonlinear dependence on wavelength because of a changing penetration depth into the photocapacitor. When a voltage is applied to the pixel, a potential well is formed, which "collects," or integrates, the charge. The charge formed in each pixel is then transferred serially within the silicon substrate using a series of voltage gates, converted to a voltage signal, amplified, and then processed (by other equipment) to give an image. These devices can be employed in an absorption spectrometer by removing the exit slit and imaging the output off the diffraction grating onto a pixel array.

Overall, PMTs have excellent sensitivity, linearity, and dynamic range for measuring instantaneous light powers, and are reasonably priced. MCPs and CCDs are also excellent detectors with the added advantage that they can be used for spatial imaging. While these latter devices have recently become more affordable, they are usually more expensive than most PMTs.

11.8
Laser-Based Techniques

Although lasers have had a profound impact on the field of spectroscopy, they are surprisingly poorly suited for direct absorption measurements in the UV/visible range. The reason for this is that a direct absorption measurement entails measuring a small decrease in the output of an intense but usually fluctuating monochromatic source due to sample absorption. Nevertheless, several powerful laser-based methods have been developed, which circumvent these difficulties and allow UV/visible absorption spectra to be recorded particularly for gas-phase species.

11.8.1
Fluorescence Excitation Spectroscopy

Laser-induced fluorescence involves detecting optical emission from a sample that has been excited by the absorption of laser radiation. As shown in Figure 11.6, fluorescence excitation spectra can be recorded by measuring the total undispersed laser-induced emission as a function of the tunable exciting light wavelength using an optical detector such as a PMT. It can be shown that if the light collection efficiency and fluorescence quantum yield are independent of wavelength, the resultant spectrum is equivalent to an absorption spectrum (Lipson and Shi, 2002). This technique is considerably more sensitive than direct absorption because, in principle, optical detection is measured against a zero-light background. Nevertheless, care must be taken to eliminate scattered light, which can be orders of magnitude more intense than the fluorescence. Spatial pinhole filters before the PMT can

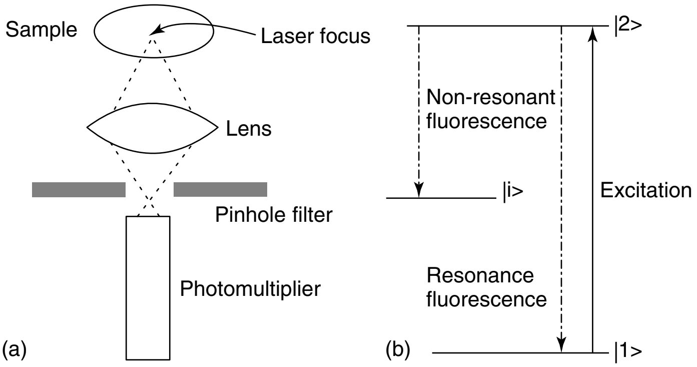

Fig. 11.6 (a) Schematic diagram of a typical experimental arrangement required to record laser-induced fluorescence excitation spectra. (b) A three-level diagram showing a single-photon transition between states $|1>$ and $|2>$, and possible resonance fluorescence and nonresonant fluorescence to an intermediate level $|i>$.

help in this regard. Sub-Doppler resolution gas-phase spectra can be obtained by sample delivery in a molecular beam or a supersonic jet expansion (Smalley, Wharton and Levy, 1977), and detecting fluorescence in a direction perpendicular to the gas flow.

11.8.2
Excitation Spectroscopy with Mass Detection

Not every molecule will fluoresce when excited. In such cases, one can use one or more laser beams to ionize the molecule of interest for detection in a time-of-flight mass spectrometer (Ashfold, 1998). The process of resonant two-photon ionization (R2PI) is described by

$$\mathrm{M}\xrightarrow{\mathrm{h}\nu}\mathrm{M}^*\xrightarrow{\mathrm{h}\nu'}\mathrm{M}^+ + \mathrm{e}^- \qquad (11.11)$$

where the first photon at frequency ν resonantly excites a molecule M while a second photon at frequency ν' ionizes the excited molecule. The ionization reaction is a one-colour process if the photons involved come from the same laser and have the same frequency. If photons are used from different lasers, and ν and ν' are different, the process is called a *two-colour ionization*. Mass selected, vibronically resolved excited-state spectra can be obtained by varying the laser excitation wavelength while monitoring the molecular mass peak. R2PI is the simplest variant of $(m + n')$ resonance enhanced multiphoton ionization (REMPI) where m and n corresponds to the number of photons required to excite the molecule and ionize the excited state, respectively, and the prime indicates that the ionizing photons come from a laser different from the excitation source.

If the molecule predissociates or undergoes some photochemical reaction, it is possible to monitor the mass of a daughter fragment ion produced to obtain an absorption spectrum of the parent. This general approach of detecting a consequence of absorption to obtain the spectrum of a parent molecule is called *action spectroscopy*.

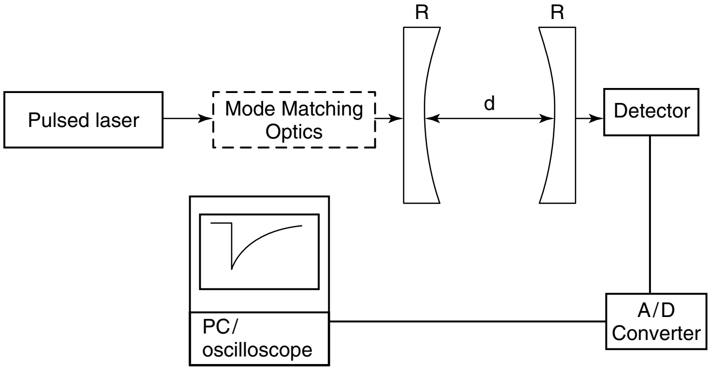

Fig. 11.7 Schematic of a typical cavity ring-down spectroscopy setup. *R* is the reflectivity of the mirrors making up the cavity of length *d*. The mode-matching optics may be optional.

11.8.3
Cavity Ring-Down Spectroscopy

Absorption spectra in cavity ring-down spectroscopy are obtained by measuring the rate of decay of light intensity inside a high-*Q* resonator (Figure 11.7) (Paldus and Kachanov, 2005). The intensity of light injected into the cavity will decay exponentially with a rate constant, $R(\lambda)$, that depends on the optical losses in the cavity.

The optical loss in an empty cavity, $L_{\text{cav}}(\lambda)$, is the sum of the round-trip scattering losses, $L_{\text{scat}}(\lambda)$ and the round-trip transmission losses through the high-reflectivity mirrors, $L_{\text{trans}}(\lambda)$. When a sample is added to the cavity, there is an additional loss due to absorption, $A(\lambda)$, which is governed by Beer's Law (Eq. 11.2). These factors are related to $R(\lambda)$ through

$$R(\lambda) = \frac{L_{\text{cav}}}{t_{\text{rt}}} + c\varepsilon(\lambda)[C] \qquad (11.12)$$

where $t_{\text{rt}} = \ell_{\text{rt}}/c$ is the photon round-trip transit time in the resonator, and ℓ_{rt} and c are the cavity round-trip length and the speed of light, respectively. $R(\lambda)^{-1}$ is called the *cavity ring-down time*. Measuring $R(\lambda)$ from the exponential intensity decay provides an accurate measurement of $\varepsilon(\lambda)$. An absorption spectrum is acquired by measuring $\varepsilon(\lambda)$ as a function of exciting laser wavelength, λ.

If high-reflectivity mirrors are used (>99.99%) that have low scattering losses (<0.001%), the effective total path length, which is determined by the average photon lifetime in the cavity, can be enhanced by factors of $>10^4$. This makes cavity ring-down spectroscopy very sensitive, even relative to setups that employ a multipass configuration. Cavity ring-down measurements are independent of the initial light intensity inside the cavity and the path length. Both pulsed and continuous wave sources can be used, and the technique is chemically specific because of the narrow linewidths available for many laser sources.

References

Ashfold, M.N.R. (1998) Multiphoton ionization (MPI) and Resonance enhanced multiphoton (REMPI) spectroscopy, in

Nonlinear Spectroscopy for Molecular Structure Determination (eds R.W. Field, E. Hirota, J.P. Maier, and S. Tsuchiya), Blackwell Science, Malden, pp. 127–145.

Berger, L.I. (1997) *Semiconductor Materials*, CRC Press, Boca Raton.

Bernath, P.F. (1995) *Spectra of Atoms and Molecules*, Oxford University Press, New York.

Berova, N., Di bari, L., and Pescitelli, G. (2007) *Chem. Soc. Rev.*, **36**, 914–931.

Cotton, F.A. (1990) *Chemical Applications of Group Theory*, 3rd edn, John Wiley & Sons, Inc., New York.

Dereniak, E.L. and Crowe, D.G. (1984) *Optical Radiation Detectors*, John Wiley & Sons, Inc., New York.

Eastman Kodak (2001) *Primer on CCDs*, http://www.kodak.com/ezpres/business/ccd/global/plugins/acrobat/en/support docs/chargeCoupleDevice.pdf.

Fieser, L.F., Fieser, M., and Rajagopalan, S. (1948) *J. Org. Chem.*, **13**, 800–806.

Figgis, B.N. (1966) *Introduction to Ligand Fields*, John Wiley & Sons, Inc., New York.

Gauglitz, G. and Vo-Dinh, T. (2003) *Hanbook of Spectroscopy*, Vol. 1, Wiley-VCH Verlag GmbH& Company kGaA, Weinheim.

Herzberg, G. (1950) *Molecular Spectra and Molecular Structure I. Spectra of Diatomic Molecules*, Van Nostrand Reinhold Company, New York.

Herzberg, G. (1966) *Molecular Spectra and Molecular Structure III. Electronic Spectra and Electronic Structure of Polyatomic Molecules*, Van Nostrand Reinhold Company, New York.

Hollas, J.M. (1996) *Modern Spectroscopy*, 3rd edn, John Wiley & Sons, Ltd, Chichester.

Klessinger, M. and Michl, J. (1995) *Excited States and Photochemistry of Organic Molecules*, VCH Publisher, New York.

Lever, A.B.P. (1968) *Inorganic Electronic Spectroscopy*, Elsevier Publishing Company, Amsterdam.

Lipson, R.H. and Hoy, A.R. (1989) *Mol. Phys.*, **68**, 1311–1319.

Lipson, R.H. and Shi, Y.J. (2002) Ultraviolet and vacuum ultraviolet laser spectroscopy using fluorescence and time-of-flight mass detection, in *Ultraviolet Spectroscopy and UV Lasers, Practical Spectroscopy Series, Vol. 30* (eds P. Misra and M.A. Dubinskii), Marcel Dekker, Inc., New York.

Lytle, F.E. (1974) *Anal. Chem.*, **46**, 545A–557A.

Paldus, B.A. and Kachanov, A.A. (2005) *Can. J. Phys.*, **83**, 975–999.

Pavia, D.L., Lampman, G.M., Kriz, G.S. (1996) *Introduction to Spectroscopy: A guide for Organic Chemistry, Saunders Golden Sunburst Series*, 2nd edn, Harcourt Brace College Publishers, Fort Worth.

Rouessac, F. and Rouessac, A. (2007) *Chemical Analysis*, 2rd edn, John Wiley & Sons, Ltd, Chichester.

Schmidt, W. (2005) *Optical Spectroscopy in Chemistry and Life Sciences*, Wiley-VCH Verlag GmbH & Company KGaA, Weinheim.

Schulman, J.H. and Compton, W.D. (1962) *Color Centers in Solids*, Pergamon Press, New York.

Silfvast, W.T. (1996) *Laser Fundamentals*, Cambridge University Press, Cambridge.

Skoog, D.A. (1985) *Principles of Instrumental Analysis*, 3rd edn, Saunders College Publishing, Philadelphia.

Smalley, R.E., Wharton, L., and Levy, D.H. (1977) *Acc. Chem. Res.*, **10**, 139–145.

Woodward, R.B. (1941) *J. Am. Chem. Soc.*, **63**, 1123–1126.

Woodward, R.B. (1942) *J. Am. Chem. Soc.*, **64**, 72–77.

Further Reading

General Spectroscopy Textbooks

Bernath, P.F. (2005) *Spectra of Atoms and Molecules*, 2rd edn, Oxford University Press, New York.

Herzberg, G. (1950) *Molecular Spectra and Molecular Structure*, Vols. I-III, VonNostrand Reinhold Company, New York (classic texts on molecular spectroscopy).

Hollas, J.M. (2004) *Modern Spectroscopy*, 4rd edn, John Wiley & Sons, Ltd, Chichester.

Karplus, M. and Porter, R.N. (1970) *Atoms and Molecules: An Introduction for Students of Physical Chemistry*, The Benjamin/Cummings Publishing Company, Menlo Park.

Group Theory Texts with Spectroscopy Applications

Bishop, D.M. (1993) *Group Theory and Chemistry*, Dover Publications, Inc., New York.

Carter, R.L. (1998) *Molecular Symmetry and Group Theory*, John Wiley & Sons, Inc., New York.

Cotton, F.A. (1990) *Chemical Applications of Group Theory*, 3rd edn, John Wiley & Sons, Inc., New York.

Jaffé, H.H. and Orchin, M. (2002) *Symmetry in Chemistry*, Dover Publications, Inc., Mineola, New York.

Organic Spectroscopy

Lambert, J.B., Shurvell, H.F., Lightner, D., and Cooks, R.G. (1987) *Introduction to Organic Spectroscopy*, Macmillan Publishing Company, New York.

12
Colorimetry

Robert T. Marcus

12.1 Introduction 383
12.2 Color Perception 383
12.2.1 Color Constancy and Metamerism 384
12.3 The CIE System of Colorimetry 385
12.3.1 Light and Light Sources 385
12.3.1.1 Visible Light 385
12.3.1.2 Standard Light Sources and Illuminants 386
12.3.1.3 Color-Matching Booths 388
12.3.2 Characterizing Objects 388
12.3.3 The Standard Observers 390
12.3.3.1 The Eye and Color Vision 390
12.3.3.2 Flare and Metamerism 391
12.3.3.3 Characterizing Observers 391
12.3.3.4 Calculating Tristimulus Values 393
12.3.3.5 Chromaticity Coordinates and Chromaticity Diagrams 394
12.4 Color Spaces 395
12.4.1 The Munsell Color Space 395
12.4.2 The 1976 CIE $L^*a^*b^*$ (CIELAB) Color Space 396
12.4.3 The 1976 CIE $L^*u^*v^*$ (CIELUV) Color Space 397
12.5 Color Differences and Tolerances 398
12.5.1 Single-Number Color Scales 400
12.5.1.1 Whiteness Scales 400
12.5.1.1 Yellowness Indices 401
12.5.1.2 Scales for Liquids 401
12.5.2 The CIELAB and CIELUV Color-Difference Equations 401
12.5.2.1 The CIELAB Color-Difference Equation 401
12.5.2.2 The CIELUV Color-Difference Equation 402
12.5.3 The CMC($l : c$) Color-Tolerance Equation 403
12.5.4 The CIEDE 2000 Color-Difference Equation 403
12.5.5 The DIN 99 Color-Tolerance Equation 405
12.5.6 Setting Instrumental Color Tolerances 406

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

12.6 Instrumental Color Measurement 406
12.6.1 Measuring Reflectance 406
12.6.1.1 Specular and Diffuse Reflectance 406
12.6.1.2 Illuminating and Viewing Geometries 407
12.6.1.3 Spectrophotometers, Colorimeters, Spectrocolorimeters, and LED-Based Instruments 410
12.6.1.4 Standardization 411
12.6.1.5 Repeatability, Reproducibility, and Accuracy 411
12.6.1.6 Measuring Fluorescent Materials 412
12.6.1.7 Measuring Materials Containing Metallic Flakes and Interference Pigments 412
12.6.2 Measuring Transmittance 415
12.6.2.1 Regular and Diffuse Transmittance 415
12.6.2.2 Illuminating and Viewing Geometries 415
12.6.2.3 Standardization 415
12.6.2.4 Measuring Transmittance 416
12.6.3 Radiometric Measurements 416
12.7 Summary 416
References 417

12.1
Introduction

The Earth, our home, is a blue planet. Color abounds from the blue sky, the striking sunsets, and the lush green of the rain forests to the bright reds and oranges of the fall foliage. Since prehistoric times man has used color to brighten life. Ancient civilizations created brightly dyed fabrics and decorated pottery. Michelangelo, Renoir, and Picasso are among the master artists who used color to enhance their art. Imagine how dull the world would be without color. Until recently, products were available in a limited variety of colors. You could buy a Model T automobile in any color you wanted as long as it was black. However, times have changed and today's consumers demand products in a wide variety of colors. Modern materials provide us with products in a multitude of colors that could only have been dreamt of in the past. Materials colored with metallic and interference pigments change colors as the illuminating and viewing conditions change. Materials colored with thermochromic pigments change color with temperature. The use of fluorescent pigments results in materials with eye-popping colors.

Scientist and technicians use colorimetry, the measurement of color, to ensure that consumers get the colors they expect consistently over time. In colorimetry, we try to simulate the human experience to quantify and control color.

In this chapter we consider color perception, the mathematics of simulating the perception of color, the color difference between two objects, and the instruments used to measure color. Several special measurement cases are also discussed.

12.2
Color Perception

In conversation we speak about a yellow school bus or a green traffic light as if the object's color was an intrinsic property of the object. Color, however, is a perception. Light falls upon an object. The object reflects and/or transmits that light depending upon the pigments used to color the object. The reflected and/or transmitted light enters the observer's eyes and the brain translates the resulting nerve impulses into the perception of color. Color is but one aspect of an object's appearance. Gloss and texture are two other important aspects of appearance. Color perception

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

is an extremely complex phenomenon. For example, the background on which an object is viewed can have a major effect on the perceived color of that object. The ambient light to which the eye becomes adapted also influences the color of objects. When we speak of color in the remainder of this chapter, we are speaking of the perception of a single, independent color by an individual with normal color vision (99.5% of women and 92% of men).

Color is three dimensional; hence, three terms are needed to describe a specific color. Typically, colors are described with reference to their lightness, chroma, and hue. Hue would be the general name for a color, for example, blue, green, yellow, red, and purple. Colorists think of hue as a circular dimension around a central axis of neutral (white to black) colors. Lightness is the dimension used to distinguish among white, gray, and black. Pure white would approach the highest possible lightness, whereas jet black would approach the lowest possible lightness. Grays range between white and black. Lightness can distinguish between colors with distinctive hues, for example, the difference between a light blue and a dark blue. Chroma is a more difficult concept. Chroma distinguishes a color's colorfulness or difference from a neutral gray. This would be the difference between a brick red and a red cherry. If the brick and the cherry have the same hue (the brick is neither yellower nor bluer than the cherry) and the same lightness, then the brick would be described as being a low-chroma red and the cherry would be described as being a high-chroma red. The terms dull (low chroma) and saturated (high chroma) are sometimes applied to chroma. Colors that differ only in lightness (white to gray to black) are called *achromatic colors*. Those colors with perceptible hue and chroma are called *chromatic colors*.

12.2.1
Color Constancy and Metamerism

The color of an object may change as the lighting conditions change. This change in color is called *flare*. For example, a red ball may look bluer outside in daylight than in a living room lighted with incandescent lamps. We would say that the ball flared red when illuminated by incandescent light. The magnitude of the change is dependent upon the pigments used to color the object. Colorists judge the effect of different illuminating conditions using industrial light booths (cabinets). Most light booths provide a simulated daylight lamp, an incandescent lamp, and a fluorescent lamp. Some light booths also have ultraviolet lamps for determining whether an object is fluorescent. The eye will try to adjust to different lamps so that they appear to be "white" (chromatic adaptation) when looked at directly.

In most industrial situations, it is preferred to keep an object's color the same as the lamp is changed. That effect is called *color constancy*. Changes in color resulting from changes in the light source can cause some very undesirable results. For example, two yellow objects are colored using different pigments but have the same color when illuminated by a simulated daylight lamp. When an incandescent lamp is used to illuminate the two objects, one flares green and the other flares red. The colors of the two objects no longer match. This effect is called *metamerism*, and we describe the two objects as being metameric to one another. Metamerism is a major problem for industrial color matchers, particularly when trying to match a target color obtained in one material (such as a dyed bath towel) to that made in a different material (such as a plastic toothbrush holder). Can you imagine buying a suit whose jacket only matches the pants when illuminated by incandescent light?

12.3
The CIE System of Colorimetry

The Commission Internationale de l'Éclairage (International Commission of Illumination), which is usually referred to by its French initials – CIE, is an organization of countries that is concerned with light and its measurement. The scientists and engineers whose countries were members of the CIE did the pioneering work in color measurements. The CIE, by the consent of all its member countries, specifies the basic standards for use in colorimetry. The first major recommendations regarding colorimetric standards were made by the CIE in 1931 and formed the basis of modern colorimetry. Since then many additional recommendations have been made by the CIE and adopted in the field of color measurement (CIE, 2004)

12.3.1
Light and Light Sources

12.3.1.1 Visible Light

The Earth is constantly bombarded with electromagnetic radiations. This includes cosmic rays, gamma rays, visible light, and radio lights. Visible light makes up only a small portion of the electromagnetic spectrum. Visible light is the electromagnetic energy that stimulates the cells in the human eye. In colorimetry, we characterize visible light by its wavelength in nanometers (nm). The CIE considers visible light to range from 360 to 830 nm. Not everybody can perceive light throughout that entire wavelength range. Most humans can perceive light between 400 and 700 nm and many color measuring instruments can only measure light within that more restrictive wavelength range.

Each of the wavelengths (or a narrow band of wavelengths) of light is perceived as a color. The wavelengths between 360 and 500 nm will appear to be shades of violet and blue; the wavelengths between 500 and 580 nm will appear to be shades of green; and the wavelengths between 580 and 830 nm will appear to move from shades of yellow to shades of orange to shades of red.

A light source emitting a broad band of wavelengths may appear to have color even though the eye adapts and tries to make it appear white. The sun appears red, yellow, or white depending upon the time of day and the scattering of the light by particles in the atmosphere. An incandescent bulb will be described as yellow, and fluorescent lamps will usually look white. If you take a walk around dusk, your eyes will adapt to the bluish color of daylight. If you pass a house with a window to an interior room lit by an incandescent lamp, you will immediately notice how yellow the incandescent lamp is compared to outside daylight.

If you heat a block of pure carbon, it will change color from black to yellow, to white, and then blue as the temperature increases. The carbon block can be heated until its color matches the color of the light source. The temperature of the carbon block (in kelvin, K) when its color matches the light source would be the color temperature of the light source. An incandescent light bulb that appears to be yellow has a color temperature of about 2850 K, whereas north sky daylight has a color temperature of about 7500 K. If a match between the heated carbon and the light source cannot be made, the temperature of the closest color to the light source would be used and is called the *correlated color temperature.* In conversation, and often in writing, many people do not distinguish between the color temperature and the correlated color temperature and use the terms interchangeably.

A light source can be characterized by its spectral power distribution (SPD), which is the relative power emitted by the light source at each wavelength of interest. Some light sources emit radiation outside of the region of visible light. Wavelengths of interest for those light sources may extend below the 360 nm visible range into the ultraviolet region and above the 830 nm visible range into the infrared region.

12.3.1.2 **Standard Light Sources and Illuminants**

For hundreds of years, north sky daylight has been the favorite light source for color matchers. Unfortunately, the use of natural daylight limits the time available for color matching. In our modern world, teams of color matchers may have to work round the clock to keep up with the demands for their products. Natural daylight is not constant and changes during the course of the day. In the morning, natural daylight will have a color temperature of about 3000 K and at noontime about 5000 K. Average daylight has a color temperature of 6500 K and north sky daylight has a color temperature of 7500 K. Average daylight and north sky daylight would appear bluish if it were not for the chromatic adaptation of the eye, which makes daylight appear to be white. The bluish cast of daylight can sometimes be seen in photographs.

Recognizing the problem with natural daylight, the CIE defined three reproducible light sources for use in colorimetry. CIE source A was a tungsten filament light bulb with a color temperature of 2854 K. CIE sources B and C were created by filtering the light from source A. Source B approximated noon sunlight and source C approximated average daylight. Source B is no longer recommended for colorimetry.

A distinction must be made between the terms *light source* and *illuminant.* A light source is a real entity that can be turned on and off to illuminate objects. An illuminant is an SPD representing a light source. CIE standard illuminants A, B, and C were created by measuring the SPDs of standard sources A, B, and C. Illuminants are used in colorimetric calculations.

Although CIE source C is an approximation to average daylight, there are some significant differences. Since source C is based on an incandescent lamp, it does not have much power in the blue and ultraviolet regions of the spectrum, whereas natural daylight does. When measuring fluorescent materials, it is important to simulate the amount of ultraviolet radiation in natural daylight.

In 1968, the CIE adopted the D series of standard illuminants duplicating the SPDs of various phases of natural daylight (CIE, 2004; ASTM 308). The name of the various D illuminants is based upon their correlated color temperature, such as D50 (5000 K) and D65 (6500 K).

D65 is commonly used in the textile, paint, and plastics industries, whereas D50 is typically used in the graphic arts. Very few real light sources simulate any of the D illuminants satisfactorily; hence, the CIE recommended procedures for assigning the quality of daylight simulators (CIE, 2004).

Fluorescent lamps are widely used in stores and offices. As a result, the CIE defined a series of fluorescent (F) illuminants to represent commonly used fluorescent lamps (CIE, 2004). F illuminants are grouped according to the types of phosphors used in the manufacture of lamps. F1–F6 represent the traditional fluorescent lamps, which use two semibroadband emissions of antimony and manganese activations in a calcium halo-phosphate phosphor. F2 represents a typical cool white fluorescent lamp and is the most used of the F illuminants. F7–F9 represent the "broad-band" group of fluorescent lamps, which are made using multiple phosphors in order to have better color-rendering properties than the standard group. Color rendering is the ability of the illuminant to keep a group of colors the same as the reference illuminant, typically D65. F10–F12 represent the three narrow-band group of fluorescent lamps, which were developed to save energy and still had excellent color-rendering properties. Each of these illuminants has a narrow-band emission in the red, green, and blue wavelength regions. Rare-earth phosphors are used to make lamps representing these illuminants. Illuminants D65, A, and F2 are illustrated in Figure 12.1 and illuminants F2 and F11 are illustrated in Figure 12.2.

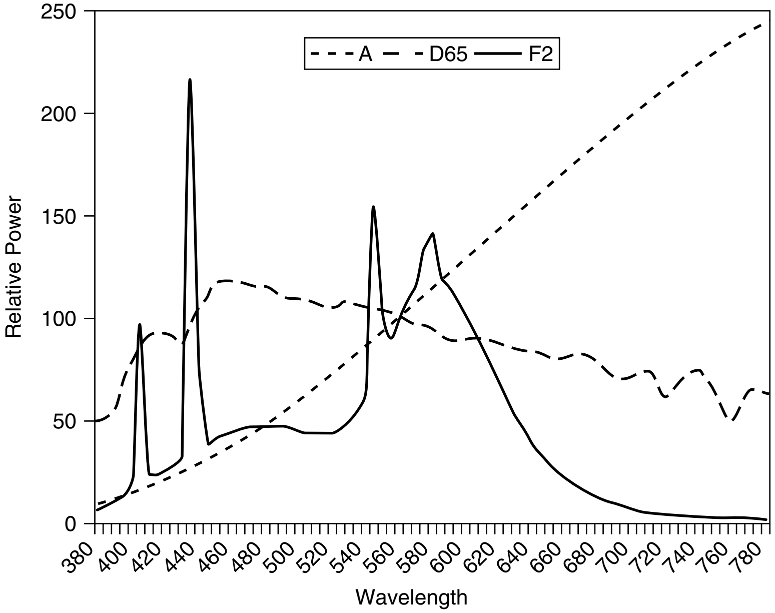

Fig. 12.1 The relative spectral power distribution of CIE standard illuminants A, D65, and F2. The curves are for shape comparison and normalized so that they have a value of 100.00 at 560 nm.

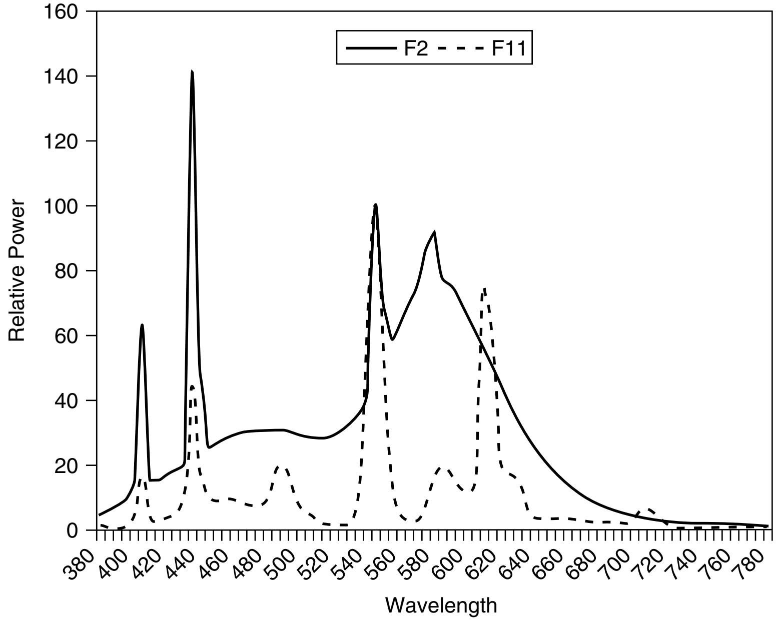

Fig. 12.2 The relative spectral power distribution of CIE standard illuminants F2 and F11. The curves are for shape comparison and normalized so that they have a value of 100 at 540 nm.

12.3.1.3 **Color-Matching Booths**

Commercial color matching takes place during the day and at night. Daylight is not available at night and other light sources are required to determine metamerism. It is essential that the visual evaluation of color be done under standardized illumination, such as that provided by a color-matching light booth. The light sources in the booth attempt to duplicate the CIE standard illuminants. Some manufacturers do better than others and it is extremely difficult to come close to achieving the SPD of real daylight with an artificial light source. However, a carefully manufactured and maintained light booth permits a colorist to make a visual evaluation of color with the confidence that the illumination stays constant over time.

12.3.2
Characterizing Objects

When light illuminates an object, the object will absorb, reflect, emit, and/or transmit that light. The events that take place depend upon the object and the pigments used to color it. If all of the light is reflected from an object, it will appear white. If all of the light is transmitted through an object, it will appear clear. If all of the light is absorbed by an object, it will appear black. Most objects will selectively absorb some wavelengths of light and will reflect or transmit the other wavelengths. The color of the object depends upon which wavelengths are absorbed and which are reflected or transmitted. Fluorescent objects absorb some wavelengths of light and then (in addition to reflection or transmission) emit light at longer wavelengths. Phosphorescent objects will emit light after the initial light no longer illuminates them and will "glow in the dark".

In colorimetry, we represent objects by their spectral reflectance or transmittance curves, the amount of light reflected or transmitted at each wavelength of the visible spectrum. The reflectance of an object is the ratio of reflected radiant flux to that of the incident flux. In common usage, the term *reflectance* is often used as an abbreviation for the term *reflectance factor*. The reflectance factor is the ratio of the intensity of reflected radiant energy to that reflected from the perfect reflecting diffuser under the same geometric and spectral conditions of measurement. The perfect reflecting diffuser is defined as an ideal reflecting surface that neither absorbs nor transmits light, but reflects diffusely, with the radiance of the reflecting surface being the same for all reflecting angles, regardless of the angular distribution of the incident light (ASTM E284). The perfect reflecting diffuser is the theoretical standard used for the instrumental measurement of reflectance. For the remainder of this chapter, we generally refer to the reflectance of an object since the calculations do not distinguish between the two. The transmittance of an object is the ratio of transmitted flux to that of the incident flux under specified geometric and spectral conditions (ASTM E284).

Figure 12.3 shows the spectral reflectance curves of black ink and of a white ceramic tile. Figure 12.4 shows the spectral reflectance of a blue ink, a green ink, and a red ink. Black (and gray) objects absorb all of the wavelengths of light. A perfect white would reflect all of

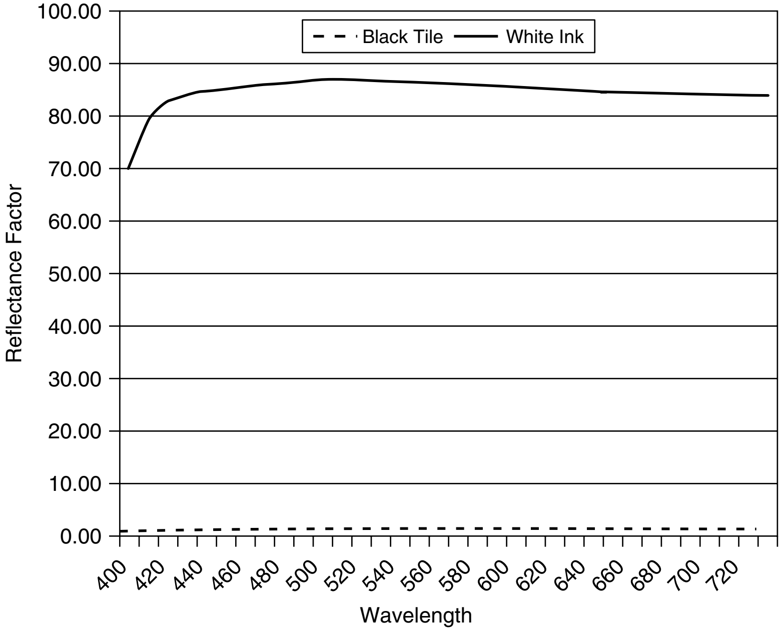

Fig. 12.3 The spectral reflectance factor curves of black ink and of a white tile.

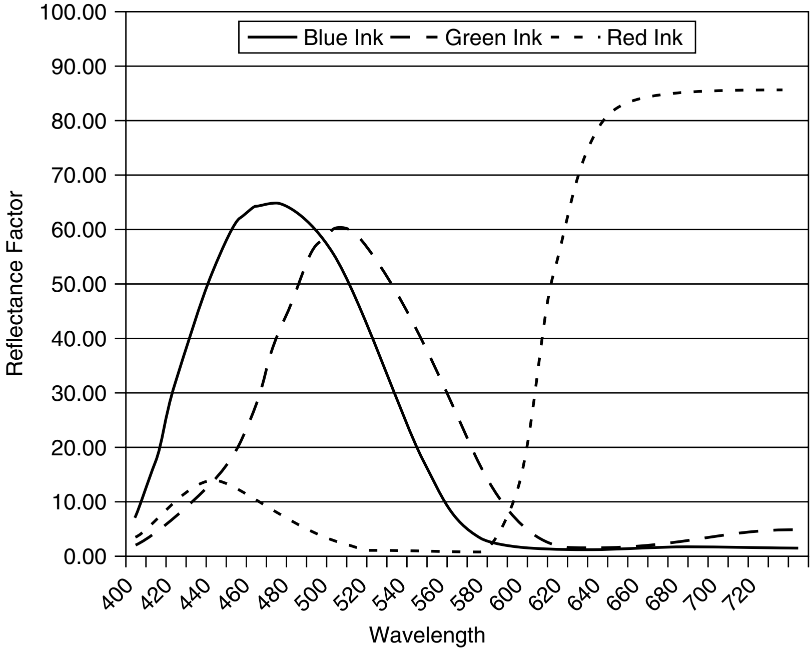

Fig. 12.4 The spectral reflectance factor curves of a blue ink, a green ink, and a red ink.

the wavelengths of light, but white objects in the real world are never perfect and will have some absorption. Many white objects have significant absorption below about 420 nm because they are colored with rutile titanium dioxide, a white pigment that absorbs strongly in that region of the spectrum. Blue objects absorb the green and red wavelengths of light, green objects absorb the blue and red wavelengths of light, and red objects absorb the blue and green wavelengths of light.

A spectrophotometer is used to measure the reflectance or transmittance factor of an object as a function of wavelength. The measured reflectance or transmittance factor of an object depends on the instrument used to make the measurements and the conditions under which the measurements are made. Several standard measuring geometries were developed for colorimetry by the CIE. These geometries are discussed in greater detail later in this chapter.

12.3.3
The Standard Observers

12.3.3.1 The Eye and Color Vision

The human eye is much like a camera. The lens images the scene on the light-sensitive retina. Two types of receptors, rods and cones, in the retina absorb a portion of the incident light and generate a signal, which is processed and interpreted by the brain. Rods detect small amounts of light and are responsible for night vision. They are not sensitive to color and are inactive at higher light levels. Cones have a much lower sensitivity to light but become active at daytime light levels. There are three types of cones called *S*, *M*, and *L*. S cones are sensitive to the short wavelengths of light, M cones are sensitive to the medium wavelengths of light, and L cones are sensitive to the long wavelengths of light. Although the spectral sensitivities of the cones slightly overlap, they are sensitive to wavelengths that roughly correspond to blue, green, and red. The cones reduce the entire spectrum of light into three signals. The distribution of rods and cones varies throughout the retina, with the cones distributed most densely at the center of the eye in an area called the *fovea*. Light triggers impulses from the cone cells that are either subtracted or added to create the black-white, red-green, or yellow-blue signals that are sent to the brain. These opponent signals travel to the brain via the optic nerve where they are interpreted as a colored image.

Most modern theories of color vision combine impulses from the rods and the three cone receptors of the eye into opponent color signals in the brain. So far, a theory of color vision that accounts for all the effects related to color perception has not been proposed.

12.3.3.2 Flare and Metamerism

Flare occurs when an object illuminated by one light source produces a different set of opponent signals than when that object is illuminated by a different light source.

The color of two objects matches when each object produces the same opponent signals. Since the cones in the eye reduce the spectrum into three signals, the two objects do not have to have the same spectral reflectance distribution to produce the same opponent signals. Metamerism occurs when the two objects have different spectral reflectance distributions but produce the same opponent signals when illuminated by one light source but produce different opponent signals when illuminated by a different light source.

12.3.3.3 Characterizing Observers

Since all color vision results from the opponent signals of the cones, shining combinations of red, green, and blue lights on the cones can produce all the colors. This was the principle used to characterize the two CIE Standard Observers.

If a single wavelength (or very narrow band of wavelengths) of light from the visible spectrum is projected upon one half of a screen, an observer should be able to match the perceived color by varying the intensities of a red, a green, and a blue light projected on the other half of the screen as shown in Figure 12.5. The amounts of each

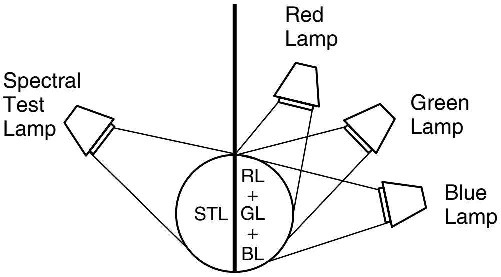

Fig. 12.5 Matching spectral colors with combinations of red, green, and blue lights.

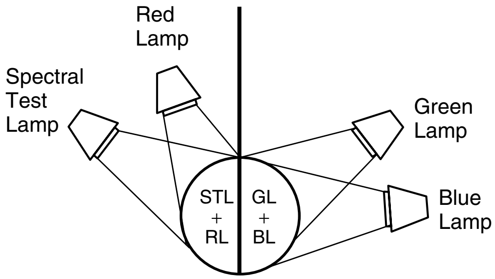

Fig. 12.6 Some spectral colors could not be matched using the combinations of red, green, and blue lights so that one of the lights had to be moved to make a match.

the three lights needed to match the spectral color are called *color-matching functions*. Observers are characterized by determining their color-matching functions for each wavelength of the visible spectrum.

Ideally, the red, green, and blue lights used in these experiments should match the responses of the three cones; unfortunately, they do not. As a result, some of the spectral colors could not be matched using combinations of real lights. When this occurred, one of the three matching lights was moved so that it illuminated the same side of the screen as the spectral color and could then be matched with the remaining two lights as shown in Figure 12.6. Moving a matching lamp so that it illuminates the same side of the observer's field of view as the test lamp results in a negative color-matching function. To avoid using negative numbers in the colorimetric calculations, the color-vision scientists created three "imaginary" lights, X, Y, and Z, to replace the real red, green, and blue lights used in the experiment. The three sets of color-matching functions (one for each light) were transformed mathematically in three sets of color-matching functions, $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$, that are always positive numbers. The symbol λ is used to indicate that the color-matching functions are wavelength dependent.

In the experiments that resulted in the 1931 CIE 2° Standard Observer, a group of observers viewed a circular field that covered 2° of their visual field, which is about the size of a dime viewed at 18 inches. Most of the cones are contained within this field of view. A 2° field of view is similar to that used to observe signal lights, flares, and other small objects.

Often a color is viewed using a much larger field of view. For example, industrial color matchers may often use samples that are at least 4 × 6 inches. Rods and the distribution of cones affect the perception of these colors. In 1964, the CIE recommended the CIE 1964 Supplemental Standard Observer. The experiments that resulted in the 1964 CIE 10° Standard Observer were based on observers viewing a circular field that covered 10° of their visual field, which is about the size of a 3-in. circle at 18 in. The three sets of color-matching functions for the CIE 10° Standard Observer are designated $\bar{x}_{10}(\lambda)$, $\bar{y}_{10}(\lambda)$, and $\bar{z}_{10}(\lambda)$ and are shown in Figure 12.7.

Since the two CIE Standard Observers were based on averaging the color-matching functions of a number of observers, it is unlikely that they will be the same as any single observer. The differences from the CIE Standard Observer usually do not present a

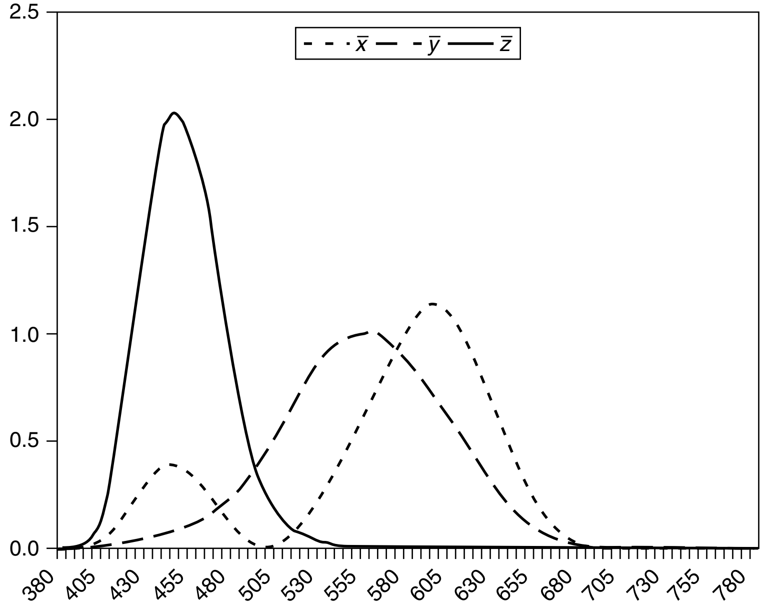

Fig. 12.7 The CIE color-matching functions for the 10° standard observer.

problem. However, they can affect the evaluation of highly metameric samples.

12.3.3.4 **Calculating Tristimulus Values**

Tristimulus values are the amounts of the X, Y, and Z lights needed to match a color. Each tristimulus value is calculated by finding the integral of the product of the spectral power distribution of the illuminant, $S(\lambda)$, the reflectance (or transmittance) of the object, $R(\lambda)$, and the appropriate color-matching function of the observer over the visible wavelength region as shown in Figure 12.8. The integral is approximated by taking the sum of those products at 1-nm intervals from 360 to 830 nm and then normalizing the sum as shown in Eqs (12.1)–(12.4). For transmitting objects substitute the transmittance factor, $T(\lambda)$, for the reflectance factor $R(\lambda)$. The factor k normalizes Y so that it is equal to 100.00 for the perfect reflecting diffuser or perfect transmitter, that is, $R(\lambda)$ or $T(\lambda)$ is equal to 1.00 for all wavelengths of the visible spectrum:

$$X = k\sum S(\lambda)R(\lambda)\bar{x}(\lambda) \tag{12.1}$$

$$Y = k\sum S(\lambda)R(\lambda)\bar{y}(\lambda) \tag{12.2}$$

$$Z = k\sum S(\lambda)R(\lambda)\bar{z}(\lambda) \tag{12.3}$$

$$k = \frac{100}{\sum S(\lambda)\bar{y}(\lambda)} \tag{12.4}$$

Reflectance and transmittance factors are often displayed as percentages ranging from 0.00 to 100.00. When calculating tristimulus values, the reflectance and transmittance factors should be expressed as a decimal ranging from 0.00 to 1.00. The above-mentioned summations are also valid for reflectance or transmittance measurements made at 5 nm intervals.

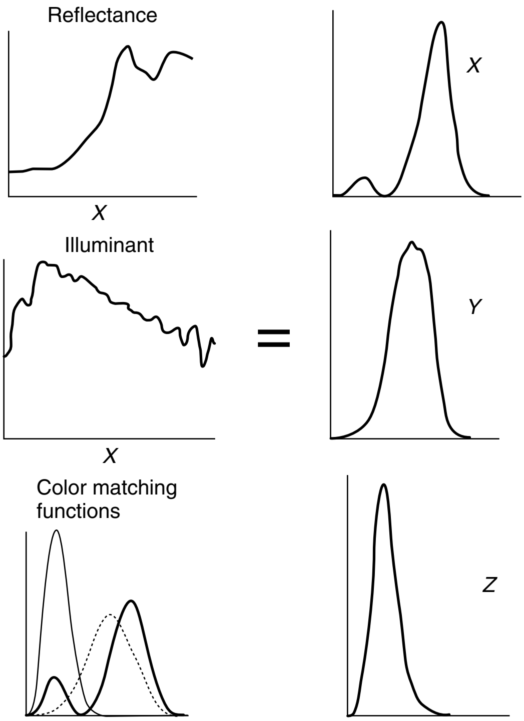

Fig. 12.8 Calculating the tristimulus values of a color.

Most commercial spectrophotometers designed for color measurement measure reflectance or transmittance factor at 10 nm intervals and some instruments measure at 20 nm intervals. Tristimulus values cannot be correctly calculated by only using the table values at 10 or 20 nm intervals. To facilitate the calculation of tristimulus factors from measurements made at 10 and 20 nm intervals, tables of normalized weighting factors, $W_x(\lambda)$, $W_y(\lambda)$, and $W_z(\lambda)$, have been published by the ASTM (ASTM 308). Weighting factors are derived from the combination of the illuminant and observer data along with other adjustments based on the wavelength interval and range selected. The tristimulus values are computed by summing the product of the weighting factor and the reflectance at the same wavelength as indicated in Eqs (12.5)–(12.7):

$$X = \sum W_X(\lambda)R(\lambda) \tag{12.5}$$

$$Y = \sum W_Y(\lambda)R(\lambda) \tag{12.6}$$

$$Z = \sum W_Z(\lambda)R(\lambda) \tag{12.7}$$

12.3.3.5 **Chromaticity Coordinates and Chromaticity Diagrams**

Most people cannot visualize a color based upon its tristimulus values. Thus, it is very useful to be able to make a graph to help visualize colors and their relationship with other colors, such as a standard and a batch. Graphing tristimulus values would require a three-dimensional plot that is very difficult to make. Separating a color into two components, its lightness and its chromaticity, allows for

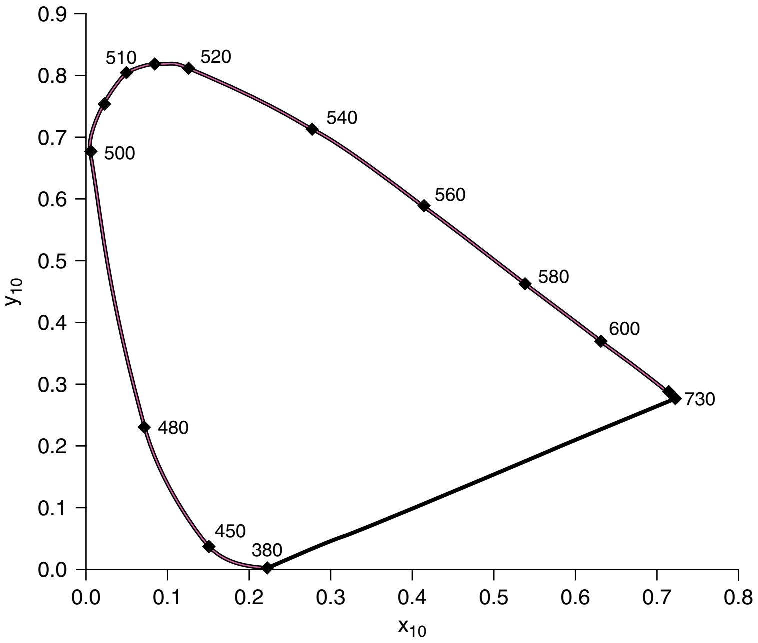

Fig. 12.9 The CIE 1964 chromaticity diagram.

a two-dimensional plot. A color's chromaticity is made up of its hue and chroma. The Y tristimulus value defines the color's lightness. The color's chromaticity coordinates, x and y, are calculated from the tristimulus values as shown in Eqs (12.8) and (12.9):

$$x = \frac{X}{X+Y+Z} \tag{12.8}$$

$$y = \frac{Y}{X+Y+Z} \tag{12.9}$$

A chromaticity diagram based on the 1964 Standard Observer is shown in Figure 12.9. The horseshoe shape contains the chromaticities of the wavelengths of the visible spectrum. The nonspectral purples are created by mixing blue light with a wavelength of 380 nm and red light with a wavelength of 780 nm. These colors fall on the straight line connecting the ends of the horseshoe.

Chromaticity diagrams are very useful for people working with colored lights and phosphors. Mixtures of two lights or phosphors will fall on the line connecting the chromaticities of each at a position predicted by the relative luminance of each light in the mixture. Plotting the chromaticities of accepted batches relative to the standard color helps define industrial color tolerances (ASTM 3134).

12.4
Color Spaces

12.4.1
The Munsell Color Space

Albert Munsell, an artist and educator, developed a visually uniform color space in the early 1900s. Munsell's system had three coordinates – hue, value, and chroma.

Value is an artist's term for lightness. A perfect black would have a value of 0, and a perfect white would have a value of 10. Each step (0–1, 1–2, 3–4,. . ., 9–10) appears to be equally different in lightness.

Munsell defined a hue circle consisting of 10 hues (red, R; yellow-red, YR; yellow, Y; green-yellow, GY; green, G; blue-green, BY; blue, B; purple-blue, PB; purple, P; and red-purple, RP). Each pair of adjacent hues in the hue circle will appear to have the same color difference as any other pair of adjacent hues in the circle when the samples have the same lightness and same chroma.

Chroma would designate a color's distance from the central neutral axis. When Munsell created his system, he used his brightest vermilion pigment to anchor the chroma at 10. Colors can have chromas greater than 10. Munsell created an atlas to exemplify his system. In his first atlas, he divided the chroma scale in visually equal steps of 2.0 chroma units.

On a perfect, visually uniform chromaticity diagram, the Munsell hue circle would plot as a perfect circle.

12.4.2
The 1976 CIE *L*a*b** (CIELAB) Color Space

Figure 12.10 illustrates a Munsell hue circle at a value of 5 and a chroma of 8 plotted on the 1931 CIE chromaticity diagram. The hue circle looks more like a pear than a circle. The members of the CIE wanted a more uniform color space than either the 1931 space or the 1964 color space. Work in the late 1960s and early 1970s resulted in the CIE 1976 $L^*a^*b^*$ (CIELAB) color space (CIE, 2004). CIELAB has gained wide acceptance in industry.

CIELAB is an opponent-type color space. Values for the lightness axis L^*,

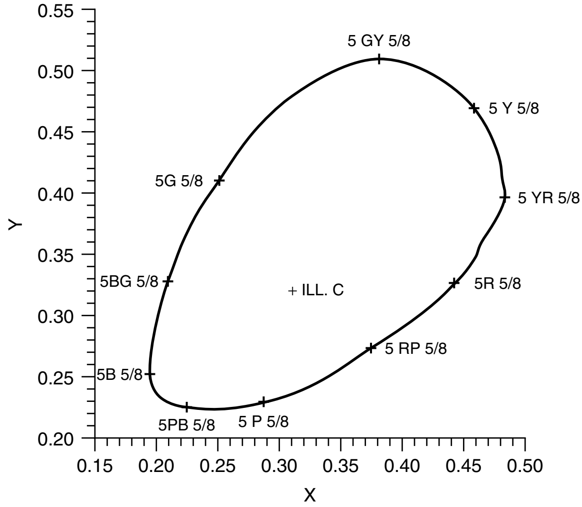

Fig. 12.10 The Munsell hue circle using the 1931 CIE chromaticity coordinates.

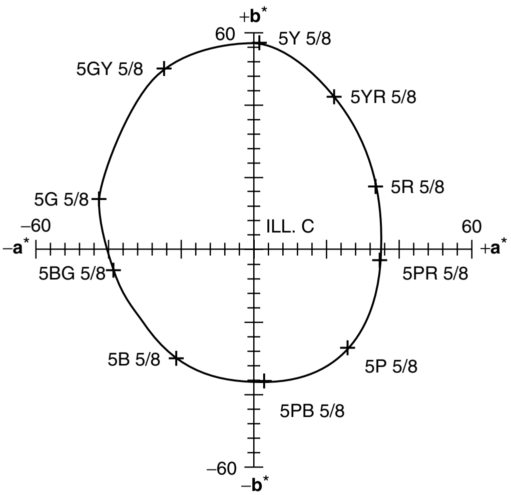

Fig. 12.11 The Munsell hue circle plotted in CIELAB color space.

the redness (positive values)-greenness (negative) axis a^*, and the yellowness (positive)-blueness (negative) axis b^* are calculated from Eqs (12.10)–(12.12):

$$L^* = 116f\left(\frac{Y}{Y_n}\right)^{1/3} - 16 \qquad (12.10)$$

$$a^* = 500\left[f\left(\frac{X}{X_n}\right)^{1/3} - f\left(\frac{Y}{Y_n}\right)^{1/3}\right] \qquad (12.11)$$

$$b^* = 200\left[f\left(\frac{Y}{Y_n}\right)^{1/3} - f\left(\frac{Z}{Z_n}\right)^{1/3}\right] \qquad (12.12)$$

When X/X_n, Y/Y_n, $Z/Z_n \geq 0.008856$, then $f(X/X_n) = (X/X_n)^{1/3}$, $f(Y/Y_n) = (Y/Y_n)^{1/3}$ and $f(Z/Z_n) = (Z/Z_n)^{1/3}$, where X_n, Y_n, and Z_n, are the tristimulus values of the illuminant or white point.

When X/X_n, Y/Y_n, $Z/Z_n < 0.008856$, then $f(X/X_n) = 7.787(X/X_n) + 16/116$, $f(Y/Y_n) = 7.787(Y/Y_n) + 16/116$, and $f(Z/Z_n) = 7.787(Z/Z_n) + 16/116$.

For workers who prefer working with lightness, chroma, and hue coordinates, the CIE defined chroma and hue angle as shown in Eqs (12.13) and (12.14). Note that the hue angle is measured in degrees.

$$C^* = \left(a^{*2} + b^{*2}\right)^{1/2} \qquad (12.13)$$

$$h = \tan^{-1}\left(\frac{b^*}{a^*}\right) \qquad (12.14)$$

These correlate well with visual judgments of lightness, chroma, and hue, respectively.

Figure 12.11 illustrates the Munsell hue circle plotted in CIELAB color space. CIELAB is much closer to a visually uniform color space than the 1931 or 1964 color spaces.

A chromaticity diagram was not defined for the CIELAB color space.

12.4.3
The 1976 CIE $L^*u^*v^*$ (CIELUV) Color Space

Since some industries, such as television and video displays, find chromaticity diagrams useful, the CIE also recommended the CIELUV color space that does have a chromaticity diagram. Values for the CIELUV coordinates are calculated from Eqs (12.15)–(12.22):

$$L^* = 116\left(\frac{Y}{Y_n}\right)^{1/3} - 16 \text{ when } \frac{Y}{Y_n} > 0.008856 \quad (12.15)$$

$$L^* = 903.3\left(\frac{Y}{Y_n}\right)^{1/3} \text{ when } \frac{Y}{Y_n} \leq 0.008856 \quad (12.16)$$

$$u^* = 13L^*(u' - u'_n) \quad (12.17)$$

$$v^* = 13L^*(v' - v'_n) \quad (12.18)$$

where the chromaticity coordinates of the color (u', v') are defined as

$$u' = \frac{4X}{X + 15Y + 3Z} \quad (12.19)$$

$$v' = \frac{9X}{X + 15Y + 3Z} \quad (12.20)$$

and the chromaticity coordinates of the illuminant or white point are defined as

$$u'_n = \frac{4X_n}{X_n + 15Y_n + 3Z_n} \quad (12.21)$$

$$v'_n = \frac{9X_n}{X_n + 15Y_n + 3Z_n} \quad (12.22)$$

The CIE also defined chroma, C^*_{uv}, and hue angle, h_{uv}, coordinates for CIELUV that are calculated using Eqs (12.23) and (12.24):

$$C^*_{uv} = \left(u^{*2} + v^{*2}\right)^{1/2} \quad (12.23)$$

$$h_{uv} = \arctan\left(\frac{v^*}{u^*}\right) \quad (12.24)$$

In addition, the CIE defined a correlate to perceived saturation, S^*_{uv}, that is calculated using either Eq. (12.25) or its equivalent Eq. (12.26):

$$S^*_{uv} = \frac{C^*_{uv}}{L^*} \quad (12.25)$$

$$S^*_{uv} = 13\left[\left(u' - u'_n\right)^2 + \left(v' - v'_n\right)^2\right]^{1/2} \quad (12.26)$$

The perceived saturation, S^*_{uv}, remains constant for a series of colors of constant chromaticity as the lightness increases or decreases (Hunt, 1991).

12.5
Color Differences and Tolerances

Most of the time, industrial color matching and production involve the evaluation and description of color differences. A colorist may be evaluating two lots of different pigments or dyes, comparing a target color with a laboratory sample or comparing a product standard with a production batch. Qualitatively describing the color differences is essential and being able to quantify the color difference is extremely useful. People are most sensitive to differences in hue, more sensitive to differences in chroma, and least sensitive to differences in lightness. This is the result of how colors change in nature. For example, a shadow falling across a fence results in the fence being darker in the shadow area. Although this is an actual color difference based on the difference in lightness, a person attributes the difference to the shadow and not to the color of the fence itself.

Distinguishing between the perceptibility of two colors and the acceptability of a color match are extremely important. A color difference between two samples is perceptible if you can see the color difference between them. Although you may be able to perceive the color difference between a product standard and a production batch, the difference in color may still be acceptable to the user or consumer. A color tolerance tells the colorist to what extent a color difference is acceptable between the standard and the batch. The size of the color tolerance depends on the product and how it is being used. For example, the color difference between two bolts of dyed fabric must be very small if you intend to make a suit using both bolts. However, the tolerance may be much larger if the bolts would not be used together and will be viewed as separate products.

Colorists and designers have used many terms for describing color differences. They describe the differences between two colors as being stronger, duller, slightly blue, and too gray. Designers have asked colorists to make a color "happier." Unfortunately, most of those terms do not help the colorist decide how to adjust the batch since one person's "slightly" may be another person's "strongly." And how do you make a color "happier"? For chromatic colors, the author recommends using the terms lightness, chroma, and hue to describe color differences. The batch is either equal in lightness to the standard or darker or lighter. The batch may have the same, a higher, or a lower chroma than the standard. Hue differences are usually described in terms of redness, yellowness, greenness, or blueness. A red or green sample may be yellower or bluer than the standard, and a yellow or blue sample may be redder or greener than the standard. For example, a red sample may be moderately yellower, slightly darker, and lower in chroma than the standard. For neutral- and low-chroma colors, it is often useful to describe color differences in terms of lightness, redness-greenness, and yellowness-blueness. The visual color difference between a low-chroma color standard and a batch can often be larger and still acceptable if both are in the same red-green and yellow-blue quadrants of color space.

Although many people rely upon an instrumental evaluation of color difference, the color of the final product is often evaluated visually. Regardless of whether or not an instrumental color difference is within a specified tolerance, there will be problems if the color does not look right. Therefore, before instrumental measurements are made, color and gloss differences between a standard and a sample should be evaluated visually. Visual evaluations should be made in light booths using standardized viewing conditions (ASTM 4449; ASTM 1729; ASTM 3928; ASTM 2616). Small-to-moderate color differences can be quantitatively evaluated visually by using a gray scale (ASTM 2616). Visually prepared color-tolerance sets consist of samples having acceptable color differences at the extremes of lightness–darkness, high- and low-chroma and the redness-greenness or yellowness-blueness differences in hue. A visual color-tolerance set can be measured to provide a set of instrumental color tolerances.

To obtain the most accurate instrumental evaluation of color difference, the standard and the batch should have the same texture and gloss. Some instruments will make measurements that minimize the affect of texture and gloss differences while others will accentuate the differences. The human eye–brain evaluation of color difference has a perceptibility tolerance as does instrumental measurements. An instrumental measurement, and the resulting color-difference calculation, may produce differences that are below those perceptible to a person. Care must be taken not to set an instrumental tolerance below which a person can perceive.

Some research has taken place on the optimum methods for calculating color differences. This has led to a number of published color-difference equations. Some of the equations correlate better with human perception than others. The goal of a color-difference equation is generally to determine a single number that will determine whether a batch is acceptable to a standard. Even though most industrial colorists tend to use a single-number color-difference tolerance, better results are usually obtained when tolerances are set for each component of the color difference.

12.5.1
Single-Number Color Scales

In some cases, the color of a sample varies along a single direction in color space, which can be adequately described by a single scale value. This occurs in the evaluation of whiteness and yellowness and in the evaluation of some liquids.

12.5.1.1 Whiteness Scales

Whiteness scales indicate the deviation from an ideal white, which is typically assigned a whiteness value of 100. Equation (12.27) represents a commonly used whiteness scale equation (ASTM 313):

$$WI = Y + (WI, x)(x_n - x) + (WI, y)(y_n - y) \qquad (12.27)$$

where Y, x, and y are the luminance factor and the chromaticity coordinates of the batch; x_n and y_n are the chromaticity coordinates for the CIE standard illuminant and source used; and WI, x and WI, y are numerical coefficients that are specified in Table 12.1. The whiteness index is valid only for specimens with $40 < WI < (5Y - 280)$.

A tint index often accompanies a whiteness index to indicate the direction in which the white varies. Equation (12.28) represents the tint index for the whiteness scale in Eq. (12.27):

$$T = (T, x)(x_n - x) + (T, y)(y_n - y) \qquad (12.28)$$

where x and y are the chromaticity coordinates of the batch; x_n and y_n are the chromaticity coordinates for the CIE standard illuminant and source used;

Tab. 12.1 Coefficients for the equation for CIE whiteness index and tint.

Value	CIE Standard Illuminant and Observer					
	C, 1931	D_{50}, 1931	D_{65}, 1931	C, 1964	D_{50}, 1964	D_{64}, 1964
x_n	0.3101	0.3457	0.3127	0.3104	0.3477	0.3138
y_n	0.3161	0.3585	0.3290	0.3191	0.3595	0.3310
WI, x	800	800	800	800	800	800
WI, y	1700	1700	1700	1700	1700	1700
T, x	1000	1000	1000	900	900	900
T, y	650	650	650	650	650	650

Tab. 12.2 Coefficients of the equation for yellowness index.

Value	CIE Standard Illuminant and observer			
	C, 1931	D_{65}, 1931	C, 1964	D_{64}, 1964
C_X	1.2769	1.2985	1.2871	1.3013
C_Z	1.0592	1.1335	1.0781	1.1498

and T, x and T, y are numerical coefficients that are specified in Table 12.1. The tint index is valid only for specimens with $-3 < T < +3$.

12.5.1.1 **Yellowness Indices**

Yellowness indices are often used to describe the yellowing of clear plastics and white paints. They show the departure of a batch from neutral toward yellow. Equation (12.29) represents a commonly used yellowness index equation (ASTM 313):

$$YI = \frac{100\,(C_X X - C_Z Z)}{Y} \qquad (12.29)$$

where X, Y, and Z are the tristimulus values of the batch and C_X and C_Z are numerical coefficients that are specified in Table 12.2.

12.5.1.2 **Scales for Liquids**

When only a limited range of color is involved, such as in the testing of the color of lacquer used in the paint industry, a simple comparison of the color of the batch with standard colored solutions or glasses ranging from colorless to highly colored can be made. A standardized series of these solutions or glasses is used to provide a specialized color scale. The color is often a measure of concentration of ingredient. One difficulty in using these special color scales is that the color of the specimen may not match with that of the standard; this can make rating on a single-number scale difficult. Nevertheless, their simplicity, low cost, and adaptability to special situations have resulted in wide use of single-number scales for certain applications (ASTM 156; ASTM 1209; ASTM 1500; ASTM 6045; ASTM 1544; ASTM 6166; ASTM 1686).

12.5.2
The CIELAB and CIELUV Color-Difference Equations

In addition to recommending the CIELAB and CIELUV color spaces, in 1976 the CIE also recommended color-difference equations for those spaces (CIE, 2004).

12.5.2.1 **The CIELAB Color-Difference Equation**

The popularity of this color space increased through the 1980s and became the most commonly used color space in industry. To calculate CIELAB color difference, the L^*, a^*, and b^* values of the batch or sample color are subtracted from those of the standard or target color (ASTM 2244):

$$\Delta L^* = L^*_{\mathrm{batch}} - L^*_{\mathrm{standard}} \qquad (12.30)$$

$$\Delta a^* = a^*_{\mathrm{batch}} - a^*_{\mathrm{standard}} \qquad (12.31)$$

$$\Delta b^* = b^*_{\mathrm{batch}} - b^*_{\mathrm{standard}} \qquad (12.32)$$

A negative value of ΔL^* indicates that the batch is darker than the standard, whereas a positive value indicates that the batch is lighter than the standard. A negative value of Δa^* indicates that the batch is greener than the standard, whereas a positive value of Δa^* indicates that the batch is redder than the standard. A negative value of Δb^* indicates that the batch is bluer than the standard, whereas a positive value indicates that the batch is yellower than the standard.

The total color difference between the batch and standard is given by Eq. (12.33):

$$\Delta E^*_{ab} = \left[\left(\Delta L^*\right)^2 + \left(\Delta a^*\right)^2 + \left(\Delta b^*\right)^2\right]^{1/2} \tag{12.33}$$

The chroma difference between the batch and standard is given by

$$\Delta C^*_{ab} = C^*_{\text{batch}} - C^*_{\text{standard}} \tag{12.34}$$

The hue difference is not the difference in the hue angle. It is that part of the color difference that is left after accounting for the lightness and chroma differences. The hue difference between the batch and standard, $\Delta H^*{}_{ab}$, is calculated using Eq. (12.35):

$$\Delta H^*_{ab} = \left[\left(\Delta E^*_{ab}\right)^2 + \left(\Delta L^*\right)^2 + \left(\Delta C^*_{ab}\right)^2\right]^{1/2} \tag{12.35}$$

The sign of $\Delta H^*{}_{ab}$ is taken as positive if the hue angle of the batch is greater than that of the standard and is taken as negative if the hue angle is less than that of the standard.

The notations DL^*, Da^*, Db^*, DC^*, DH^*, and DE^* are often substituted for ΔL^*, Δa^*, Δb^*, ΔC^*_{ab}, ΔH^*_{ab}, and ΔE^*_{ab}.

12.5.2.2 **The CIELUV Color-Difference Equation**

Calculating the CIELUV color differences is very similar to calculating the CIELAB color differences (ASTM 2244):

$$\Delta L^* = L^*{}_{\text{batch}} - L^*{}_{\text{standard}} \tag{12.36}$$

$$\Delta u^* = u^*{}_{\text{batch}} - u^*{}_{\text{standard}} \tag{12.37}$$

$$\Delta v^* = v^*{}_{\text{batch}} - v^*{}_{\text{standard}} \tag{12.38}$$

A negative value of ΔL^* indicates that the batch is darker than the standard, whereas a positive value indicates that the batch is lighter than the standard. Δu^* *and* Δv^* represent differences in the chromaticity coordinates but are not associated with color names as are Δa^* and Δb^*.

The total color difference between the batch and standard is given by Eq. (12.39):

$$\Delta E^*_{ab} = \left[\left(\Delta L^*\right)^2 + \left(\Delta u^*\right)^2 + \left(\Delta v^*\right)^2\right]^{1/2} \tag{12.39}$$

The chroma difference between the batch and standard is given by

$$\Delta C^*_{uv} = C^*_{uv,\text{batch}} - C^*_{uv,\text{standard}} \tag{12.40}$$

The hue difference is not the difference in the hue angle. It is that part of the color difference that is left after accounting for the lightness and chroma differences. The hue difference between the batch and the standard, ΔH^*_{uv}, is calculated using Eq. (12.41):

$$\Delta H^*_{uv} = \left[\left(\Delta E^*_{uv}\right)^2 + \left(\Delta L^*\right)^2 + \left(\Delta C^*_{uv}\right)^2\right]^{1/2} \tag{12.41}$$

The sign of $\Delta H^*{}_{uv}$ is taken as positive if the hue angle of the sample is greater than that of the standard and is taken as negative if the hue angle is less than that of the standard.

The notations DL^*, Da^*, Db^*, DC^*, DH^*, and DE^* are often substituted for ΔL^*, Δa^*, Δb^*, ΔC^*_{uv}, ΔH^*_{uv}, and ΔE^*_{uv}.

12.5.3
The CMC(*l* : *c*) Color-Tolerance Equation

A number of new color-difference and color-tolerance equations have been developed since the adoption of the CIELAB equations. One of these, known as the *CMC*($l:c$) *equation* (CIE, 2004; ASTM 2244; McDonald, 1980; AATCC, 173–1992), is a modification of CIELAB that has improved uniformity of visual perception of its color differences. The CMC($l:c$) equation has found widespread use in the industry and is now more popular than CIELAB. The CMC($l:c$) equation is more appropriately called a *color-tolerance equation* rather than a color-difference equation. A color-difference equation calculates the same numeric value regardless of which specimen is used – the batch or the standard. A color-tolerance equation will calculate different values depending upon the position of the standard in color space. The CMC($l:c$) equation modifies the ΔL^*, ΔC^*, and ΔH^* components of the CIELAB color difference before calculating the total color difference:

$$\Delta E_{\text{CMC}}\left(l:c\right) = cf* \sqrt{\left(\frac{\Delta L^*}{l*S_L}\right)^2 + \left(\frac{\Delta C^*}{c*S_C}\right)^2 + \left(\frac{\Delta H^*}{S_H}\right)^2} \tag{12.42}$$

$$S_L = \frac{0.040975*L^*}{1+(0.01765*L^*)} \quad \text{for } L^* \geq 16 \tag{12.43}$$

$$S_L = 0.511 \quad \text{for L}^* < 16 \tag{12.44}$$

$$S_C = \frac{0.0638*C^*}{1+(0.0131*C^*)} + 0.638 \tag{12.45}$$

$$S_H = S_C*((F*T)+1-F) \tag{12.46}$$

in which

$$F = \left\{\frac{(C)^4}{(C)^4+1900}\right\}^{1/2} \tag{12.47}$$

$$T = 0.56 + \text{abs}\left|0.2*\cos\left(h+168^\circ\right)\right| \quad \text{if } 164^\circ < h < 345^\circ \tag{12.48}$$

or

$$T = 0.36 + \text{abs}\left|0.4*\cos\left(h+35^\circ\right)\right| \quad \text{for other values of } h \tag{12.49}$$

where "abs" indicates the absolute, that is, positive value of the term inside the brackets.

The parameters l and c weight the relative contributions of lightness to chromaticity in the color difference and compensate for systematic bias or parametric effects such as texture and sample separation. The most common values for $l:c$ are 2 : 1. The parameter *cf* is a commercial factor (AATCC 173–1992) used to adjust the total volume of the tolerance region so that accept/reject decisions can be made on the basis of a unit value of the tolerance.

12.5.4
The CIEDE 2000 Color-Difference Equation

Research was continued to find a better color-difference equation than CMC ($l:c$). The CIE developed an equation that correlates better with human perception than the CMC($l:c$) equation. This equation was recommended by the CIE in 2001 and is called *CIEDE2000* (CIE, 2001). Equations (12.50)–(12.66) are required to calculate CIEDE2000 (ASTM 2244):

$$L' = L^* \tag{12.50}$$

$$a^* = (1 + G)^* a^* \tag{12.51}$$

$$b' = b^* \tag{12.52}$$

$$C' = \sqrt{(a')^2 + (b')^2} \tag{12.53}$$

$$h' = \arctan\left(\frac{b'}{a'}\right) \tag{12.54}$$

$$G = 0.5 * \left(1 - \sqrt{\frac{(\overline{C}^*)^7}{(\overline{C}^*)^7 + 25^7}}\right) \tag{12.55}$$

where $\overline{C}^*$ is the arithmetic mean of the CIELAB C* values for the pair of specimens (standard and batch).

$$\Delta L' = L'_B - L'_S \tag{12.56}$$

$$\Delta C' = C'_B - C'_S \tag{12.57}$$

$$\Delta H = \frac{(a'_S * b'_B) - (a'_B * b'_S)}{\sqrt{0.5 * [(C'_S * C'_B) + (a'_S * a'_B) + (b'_S * b'_B)]}} \tag{12.58}$$

$$\Delta E_{00}(K_L, K_C, K_H) = \sqrt{\left(\frac{\Delta L^*}{K_L * S_L}\right)^2 + \left(\frac{\Delta C^*}{K_C * S_C}\right)^2 + \left(\frac{\Delta H^*}{K_H * S_H}\right)^2 + R_T * \left(\frac{\Delta C' * \Delta H'}{K_C * S_C * K_L * S_H}\right)} \tag{12.59}$$

The factors K_L, K_C, and K_H are correction terms for variation in perceived color-difference due to the viewing conditions. To obtain color-differences similar to CMC(2 : 1), set $K_L = 2$ and $K_C = K_H = 1$:

$$S_L = 1 + \frac{0.015 * \left(\overline{L}' - 50\right)^2}{\sqrt{20 + \left(\overline{L}' - 50\right)^2}} \tag{12.60}$$

where $\overline{L}'$ is the arithmetic mean of the CIELAB L' values for the pair of specimens (standard and batch).

$$S_C = 1 + \left(0.045 * \overline{C}'\right) \tag{12.61}$$

where $\overline{C}'$ is the arithmetic mean of the CIELAB C' values for the pair of specimens (standard and batch).

$$S_H = 1 + \left(0.015 * \overline{C} * T\right) \tag{12.62}$$

$$\begin{aligned} T = 1 &- \left[0.17 * \cos\left(\overline{h}' - 30^\circ\right)\right] \\ &+ \left[0.24 * \cos\left(2\overline{h}'\right)\right] \\ &+ \left[0.32 * \cos\left(3\overline{h}' + 6^\circ\right)\right] \\ &- \left[0.20 * \cos\left(4\overline{h}' - 63^\circ\right)\right] \end{aligned} \tag{12.63}$$

where $\overline{h}'$ is the arithmetic mean of the CIELAB L' values for the pair of specimens (standard and batch). All angles are in degrees. Care should be taken while calculating the mean hue angle if the color-difference pair has samples in different quadrants. For example, a color-difference pair has hue angles of 30° and 300°. The mean hue angle for this example is 345°. To determine the mean correctly, calculate the absolute difference of the hue angles. If the absolute difference is larger than

180°, then add 360° to the smaller hue angle and divide that sum by 2:

$$R_T = -R_C * \sin(2 * \Delta\theta) \qquad (12.64)$$

$$R_C = 2 * \sqrt{\frac{\left(\overline{C}'\right)^7}{\left(\overline{C}'\right)^7 + 25^7}} \qquad (12.65)$$

$$\Delta\theta = 30 * \exp\left\{-\left[\frac{\overline{h}' - 275^\circ}{25}\right]^2\right\} \qquad (12.66)$$

Although the CIEDE2000 equation correlates better with the existing data sets, it is not as popular as the CMC($l : c$) equation.

12.5.5
The DIN 99 Color-Tolerance Equation

The concept behind the CMC($l : c$) and CIEDE2000 equations involved changing the color-difference components of the CIELAB color-difference equation to make the final equation agree more closely with human perception. A committee of the Deutsches Institut für Normung e.V. (DIN) in Berlin, Germany, took a different approach for developing an improved color-difference equation. The idea behind DIN99 is to retain a Euclidean color-difference formula by correcting the nonuniformity of CIELAB color space (DIN 6176). Equations (12.67)–(12.78) define the new color space and the steps for the calculation of a DIN99 color difference:

Red-green axis:

$$e = a^* \cos(16^\circ) + b^* \sin(16^\circ) \qquad (12.67)$$

Yellow-blue axis:

$$f = 0.7\left(-a^* \ \sin(16^\circ) + b^* \ \cos(16^\circ)\right) \qquad (12.68)$$

Chroma:

$$G = \sqrt{e^2 + f^2} \qquad (12.69)$$

Hue angle:

$$h_{ef} = \arctan\left(\frac{f}{e}\right) \qquad (12.70)$$

which are used to develop a new coordinate system defined as

$$L_{99} = 105.51 \frac{\ln(1 + 0.0158\ L^*)}{k_e} \qquad (12.71)$$

$$a_{99} = C_{99}\ \cos(h_{99}) \qquad (12.72)$$

$$b_{99} = C_{99}\ \sin(h_{99}) \qquad (12.73)$$

where

$$C_{99} = \frac{\ln(1 + 0.045\ G)}{0.045 k_{CH} k_e} \qquad (12.74)$$

$$h_{99} = h_{ef} \frac{180}{\pi} \qquad (12.75)$$

and k_{CH} and k_e are adjustable parameters. The color difference is given by the Euclidean distance:

$$\Delta E_{99} = \sqrt{(\Delta L_{99})^2 + (\Delta a_{99})^2 + (\Delta b_{99})^2} \qquad (12.76)$$

or

$$\Delta E_{99} = \sqrt{(\Delta L_{99})^2 + (\Delta C_{99})^2 + (\Delta H_{99})^2} \qquad (12.77)$$

where ΔH_{99} is defined as

$$\Delta H_{99} = \sqrt{(\Delta a_{99})^2 + (\Delta b_{99})^2 - (C_{99})^2} \qquad (12.78)$$

which is similar to the corresponding CIELAB definition.

It remains to be seen which approach to the color-difference problem will gain acceptance. The CIE continues to have groups working on the problem of a more uniform color space as well as improved color-difference equations.

12.5.6
Setting Instrumental Color Tolerances

As with all measurements, instrumental color measurement has variation. The sources of the variation include not only the instrument itself but also variation in the sample itself. Some samples are more homogeneous than others and some instruments have less variation than others. In addition, all manufacturing processes cause variation in color. A colorist must consider the variation of the whole manufacturing and measurement system when setting instrumental color tolerances. Unfortunately, there is a tendency among a number of colorists to expect the total color difference to be zero. Making multiple measurements on real samples demonstrate that this is not possible. Setting a tolerance below the perceptibility limit is a waste of time because the customers will not see that color difference. The closer the tolerance is to the perceptibility limit, the more expensive it is to produce the product. A compromise must be made in determining the limits of an acceptable, instrumental color-difference tolerance.

The first step in determining an instrumental tolerance is to know the variation in the manufacturing and measurement processes. Measuring multiple samples of multiple batches can provide this information. Unless the seller wants to select, or blend, batches of product, the tolerances cannot be set to limits less than those found in this step. Preliminary tolerances are usually set based on past experience where it is not practical, or too expensive, to determine actual production limits initially for a new color. Final tolerances are determined after the seller gets some experience with making the color. Measuring a number of batches that the customer has judged relative to the standard helps to more accurately determine the acceptability limits (ASTM 3134).

Another method of setting instrumental tolerances is from measurements of a visually prepared color-tolerance set. Designers and marketing personnel usually determine these color-difference limits of these sets before the color is actually manufactured. It is possible that the color-tolerance set might be tighter than the manufacturing variations. The colorist preparing the color-tolerance set must caution the buyer not to set the tolerances so tight that the product cannot be produced within the tolerances.

12.6
Instrumental Color Measurement

12.6.1
Measuring Reflectance

12.6.1.1 Specular and Diffuse Reflectance

If we shine a beam of light onto a perfect mirror from the left at 45° from its normal, the light will be reflected to the right at 45° from the normal. The light does not penetrate the surface of the mirror. The mirrorlike reflection from the surface of a sample is called *specular reflection.*

If we shine a beam of light onto a very good diffuser, such as a compressed pellet of barium sulfate, the light will enter the material and get scattered (redirected) many times by the barium sulfate particles. Eventually the light will find its way out of the interior of the pellet and the barium sulfate tablet will appear to reflect light in all directions as shown in Figure 12.12.

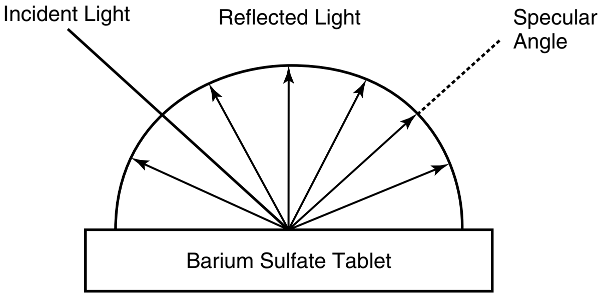

Fig. 12.12 Diffuse light reflected from a barium sulfate tablet.

Most materials have some specular reflection and some diffuse reflection as shown in Figure 12.13. Materials with very high gloss will have very narrow cones at the specular angle. As the gloss decreases, the specular cone widens. About 4% of the light reflected from glossy paint samples and glossy plastic samples will be specular reflection. The remaining 96% will be diffusely reflected.

12.6.1.2 **Illuminating and Viewing Geometries**

Color-measuring instruments consist of an illuminator, a specimen holder, and a receiver. The illuminator contains the light source and associated optics necessary to illuminate the specimen. An illuminator may also contain a diffraction grating or filters, diffusers, and various electronics. The receiver contains the components and optics necessary to gather and analyze the light reflected from the specimen. A receiver may also contain a diffraction grating or filters, diffusers, and various electronics. These components may be arranged in a variety of ways depending upon what the optical designer wants to measure. The CIE has recommended four illuminating and viewing geometries for color measurement (CIE, 2004; CIE, 1995). Two of them are bidirectional and two of them are diffuse. A number of factors should be considered when deciding upon which illuminating and viewing geometry is best for a given type of color

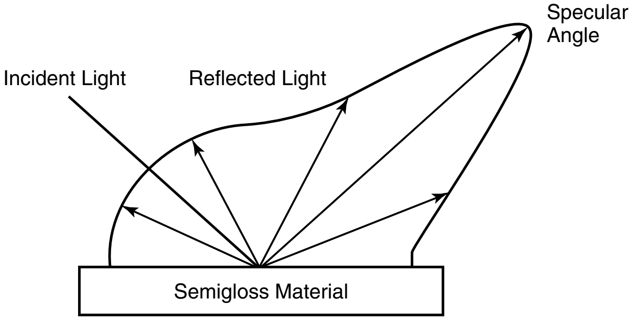

Fig. 12.13 Light reflected from a semigloss material with both a diffuse component and a specular component.

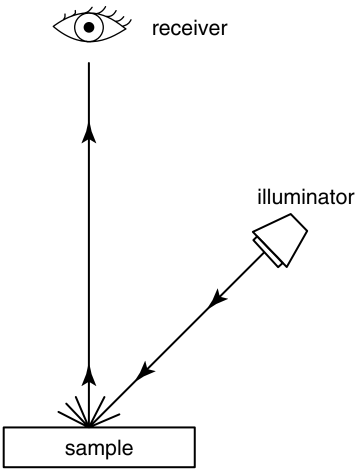

Fig. 12.14 The 45 : 0 illuminating and viewing geometry.

measurement (ASTM 179; Rich, 1988; Mabon, 1992).

Bidirectional geometries illuminate the specimen with a narrow beam of light and gather the reflected light with a receiver having a narrow field of view. One of the CIE bidirectional recommendations is to illuminate the specimen at 45° from its normal and view the reflected light along the specimen's normal. This illuminating and viewing geometry is called 45 : 0 and is shown in Figure 12.14. The CIE also recommends the reverse geometry (0 : 45) in which the specimen is illuminated along its normal and viewed at 45° from its normal. Both geometries produce equivalent results for nonfluorescent samples that do not contain metallic flakes or interference pigments (ASTM 179; Billmeyer and Marcus, 1969). The 45 : 0 and 0 : 45 geometries do not measure the specular reflection from a specimen and simulate the viewing conditions under which many specimens are evaluated visually. These measurements are sensitive to the texture of the specimen and any polarization affects that may occur when light enters the specimen. Multiple measurements should be made to average the differences due to the sample's texture. Using annular or circumferential illumination (or viewing) also minimizes texture and polarization affects. Circumferential illumination provides illumination (or the receiver will gather the reflected light) in many beams distributed at uniform intervals around the 45° cone about the sample's normal, *c* : 45 (or 45 : *c*). Annular illumination provides illumination (or the receiver will gather the reflected light) continuously and uniformly around the 45° cone about the specimen's normal, *a* : 45 (45 : *a*). Because of the surface texture influence, measurements made with bidirectional geometries are sometimes described as measuring appearance rather than color. Instruments with bidirectional geometries are very popular in the graphics arts community.

Figure 12.15 illustrates the most popular of the CIE diffuse geometries called *diffuse/normal* or *d* : 0. The illuminator shines light on the wall of an integrating sphere. An integrating sphere is a hollow metal sphere coated with an efficient white diffusing material. The most

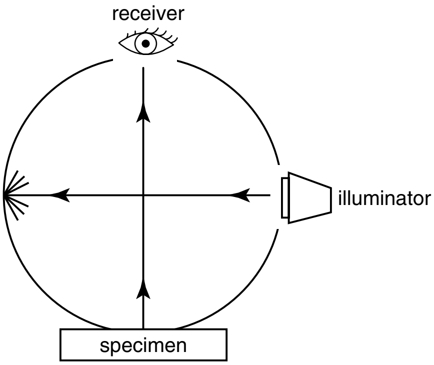

Fig. 12.15 The diffuse-normal (*d* : 0) illuminating and viewing geometry.

common integrating sphere coatings are barium sulfate ($BaSO_4$) and polytetrafluoroethylene (PTFE). Hollowing a block of PTFE has been used to create some spheres. Ports are cut into the wall of the integrating sphere, which are needed for the illuminator and the receiver. The light from the illuminator is scattered in all directions, which uniformly illuminates the integrating sphere that then acts as a diffuse hemispherical illuminator to illuminate the sample. The receiver views the specimen along its normal. Baffles are used to prevent light from the integrating sphere wall from entering the receiver directly. In the other CIE diffuse geometry (normal/diffuse or 0 : *d*), the specimen is illuminated directly along its normal and the diffuse light from the specimen uniformly illuminates the integrating sphere. The receiver then views the wall of the integrating sphere. Both geometries produce equivalent results.

Neither of the two recommended diffuse geometries measures the specular reflection from the sample. By making a slight modification to the diffuse geometries, the specular reflection from the sample can also be measured. The modification, shown in Figure 12.16, involves moving the receiver or illuminator slightly off the sample's normal (6–8°). When this modification is used, a third port (the specular port) is cut in the

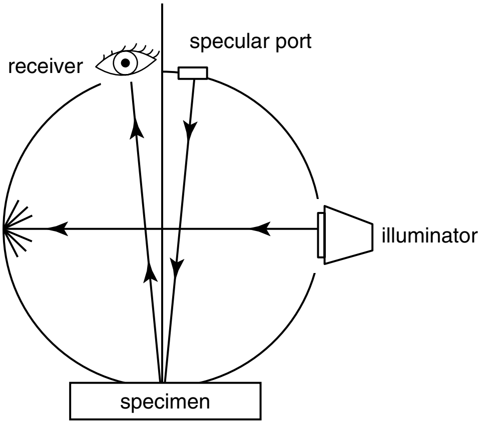

Fig. 12.16 The diffuse near-normal (*d* : 0 : *i*) illuminating and viewing geometry.

integrating sphere wall. To include the sample's specular reflection, a white cap is used to cover the specular port. The white cap should have the same reflectance as the wall of the integrating sphere. The two geometries including the specimen's specular reflection are called, respectively, *total/near-normal* ($d : 0 : i$) and *near-normal/total* ($0 : d : i$). To exclude the specimen's specular reflection, a light trap is used for the specular port. The two geometries excluding the specimen's specular reflection are called *diffuse/near-normal* ($d : 0 : e$) and *near-normal/diffuse* ($0 : d : e$). The gloss and texture of the specimen will have little effect on the color measurements when the specimen's specular reflection is included. Thus, many colorists prefer this geometry for computerized color matching. Measuring highly glossy specimens using one of the diffuse geometries with the specular reflection excluded will correlate with a bidirectional 45 : 0 instrument. This correlation will also occur for perfectly matte specimens for which there are not a noticeable specular reflection peak. Since the shape of the specular reflection varies by material and specimen and is poorly defined, designing the perfect universal light trap, without affecting the measurement of the diffuse reflectance, is impossible (Budde, 1980). The amount of the specular reflection trapped is also dependent upon the design of the instrument and of the integrating sphere.

12.6.1.3 **Spectrophotometers, Colorimeters, Spectrocolorimeters, and LED-Based Instruments**

Spectral reflectance values are required to calculate tristimulus values and color coordinates. Spectrophotometers measure the reflection (or transmission) characteristics of a specimen at different wavelengths of the visible spectrum. Spectrophotometers have been manufactured in all of the recommended illuminating and viewing geometries. Although some spectrophotometers can measure the reflectance continuously at all wavelengths of the spectrum, others (abridged spectrophotometers) can only measure selected wavelengths. Many commercial spectrophotometers measure the reflectance in 10 nm increments.

Reflectance is the ratio of the amount of light reflected by a specimen relative to the amount of light illuminating the specimen. Most spectrophotometers actually measure reflectance factor rather than reflectance. Reflectance factor is the ratio of the amount of light reflected by a specimen relative to the amount of light that would be reflected from a perfect reflecting diffuser under the same geometric and spectral conditions of measurement. Many colorists use the term reflectance when they mean reflectance factor. This distinction becomes very important when measuring fluorescent materials, materials with metallic flakes, and materials with interference pigments. Although the reflectance from a specimen can never be greater than 1.0 (100%), the reflectance factor may exceed 1.0 by a significant amount for some materials.

Colorimeters measure broad areas of the light reflected from a specimen using broad-band filters or some other spectral defining device. The amount of light reflected is converted directly into tristimulus values or color coordinates such as L^*, a^*, and b^*. Because colorimeters usually are customized for only one illuminant/observer combination, they cannot measure metamerism.

Colorimeters are generally not as accurate as spectrophotometers because of the difficulties in matching the light source and filters to the combination of CIE illuminant spectral power distributions and color-matching functions. Many of the current colorimeters are based on video camera technology. These instruments are coupled with advanced image technology to make color measurements on small areas of a complex scene. This technology is particularly useful for materials whose appearance is naturally or intentionally nonuniform such as woven and patterned fabrics. Video-based instruments may also be used for online applications such as controlling color on web printing application.

Spectrocolorimeters are spectrophotometers that output only tristimulus values or color coordinates.

Several instruments that illuminate a specimen with light emitting diodes (LEDs) have been developed. Spectral reflectance factors, tristimulus values, and color coordinates are calculated from the reflected light.

12.6.1.4 **Standardization**

The first step in making accurate color measurements is to standardize the instrument. This means the zero point of the instrument and the high end of the photometric scale must be set. A light trap is often used to set the zero point of the instrument. Most instrument manufacturers provide a white ceramic tile for setting the high end of the photometric scale. The values of the white are traceable to a national standardizing laboratory such as the US National Institute for Standards and Technology (NIST). A number of other materials have been used to set the high end of the photometric scale. They include optical grade barium sulfate (BsSO4), optical grade PTFE, and opal glasses (ASTM 259; Carter, Billmeyer and Rich, 2000). Instruments should be standardized at regular intervals. In most industrial settings, they are standardized once per 8-hour shift.

12.6.1.5 **Repeatability, Reproducibility, and Accuracy**

Top-of-the-line color measuring instruments claim a short-term repeatability of 0.01 CIELAB units. This is based upon multiple measurements of a white standard. The interinstrument agreement (reproducibility) is claimed to be 0.15 CIELAB units (maximum), with an average of 0.08 CIELAB units. This claim is based upon reading a set of 12 colored tiles. It should be noted that the very low repeatability and reproducibility numbers could only be achieved with stable standards at a tightly controlled temperature. Accuracy is determined by comparing instrumentally measured results with those of a known standard. There are no known standards available for color measurement with ‘‘absolute’’ values. Fortunately most industrial color measurement is concerned with the comparative measurement of a standard (or target) to a batch (or trial). As long as the instruments are using the same illuminating and viewing conditions and have comparable repeatability and reproducibility, these measurements are adequate. Long-term instrument stability should be monitored as part of a good quality assurance program. A set of stable ceramic standards was designed for this purpose (Clark, 1971), but any set of stable standards can be used. Some instrument manufacturers and private laboratories have programs to help the colorist monitor his/her instruments.

The repeatability and reproducibility claimed by instrument manufacturers are tested under tightly controlled conditions with highly uniform samples. It is extremely unlikely that a colorist would be able to duplicate these claims in the field. More variation in a color measurement comes from the sample rather than from the instrument. Careful sampling and sample preparation can help reduce the variability. The colorists should determine the repeatability and reproducibility using their own materials under their typical measurement conditions. Knowing how the instrument performs under field conditions helps determine practical color tolerances. The variability in measuring real specimens in the field can be reduced by making multiple measurements (ASTM 1345).

12.6.1.6 Measuring Fluorescent Materials

Fluorescent materials do more than just reflect light. They absorb radiation at some wavelengths and emit light at wavelengths longer than those that were absorbed. This is in addition to the light reflected in the same region of the spectrum. For example, many white papers will absorb ultraviolet radiation (below 400 nm) and emit blue light. The higher energy in the blue region of the spectrum makes the paper appear whiter than paper that does not fluoresce.

If an instrument with monochromatic light illuminates a fluorescent specimen, and a receiver sensitive to the entire visible spectrum is used to measure the reflectance, the receiver will not be able to distinguish the amount of light reflected at that wavelength from the light emitted at the longer wavelength so that the reflectance will appear abnormally high. Monochromatic light illuminating the sample at the wavelength where light is emitted will result in the measurement of the true reflectance at that wavelength. Spectrofluorimeters that both illuminate the sample with monochromatic light and view the sample at a single (narrow band) of wavelengths for the complete measurement of fluorescent samples were developed.

Most industrial fluorescent materials are measured with instruments that illuminate the sample with a broadband polychromatic illuminator simulating CIE Illuminant D65. Some of these instruments have special filters to control the amount of ultraviolet illumination from the source to better simulate D65 as the lamp in the illuminator ages. These instruments measure the combined reflectance and emittance, which is the total spectral reflectance factor. Because of the emittance, the total spectral reflectance factor can be greater than 1.0.

To properly measure fluorescent samples, only the 45 : 0 or 0 : 45 illuminating/viewing geometry should be used (ASTM 991). Emitted fluorescent light in an integrating sphere lowers the efficiency of the integrating sphere, resulting in lower total spectral reflectance factors than when using the other geometries (McKinnon, 1987).

Gundlach and Terstiege (1994) recommended using fluorescent calibration standards to standardize single polychromator instruments to ensure the proper measurement of fluorescent specimens.

12.6.1.7 Measuring Materials Containing Metallic Flakes and Interference Pigments

Metallic flakes and interference pigments are called *effect pigments*. When a

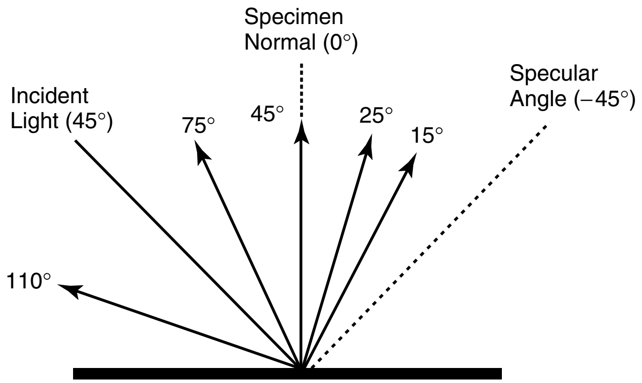

Fig. 12.17 The illuminating and viewing geometries used for the measurement of metal flake pigmented materials. The ASTM recommends using the aspecular angles of 15, 45, and 110°.

material is colored with these pigments, one or more of the attributes of color change as the illumination or viewing angle is changed. This is called *gonioapparance.*

Materials colored with only metallic flakes appear lighter when viewed near the specular angle than when viewed at angles farther away from the specular angle. This change in lightness is called *flop* in the coatings industry. The larger the metallic flakes, the greater the flop. Metallic flakes also add a sparkling appearance to the material. Multiangle instruments (sometimes called *goniospectrophotometers*) were designed to measure materials containing metallic flakes. These instruments are modifications of the common 45 : 0 bidirectional instruments. The ASTM specifies the angles of illumination and viewing by the illumination anormal angle, the viewing (detection) anormal angle, and the viewing aspecular angle enclosed in parenthesis (ASTM 2175). An anormal angle is an angle measured from the specimen normal in the illuminator plane unless otherwise specified. Positive anormal angles are usually diagrammed to the left of the specimen normal on the same side as the illuminator. An aspecular angle is a viewing angle measured from the specular direction in the illuminator plane unless otherwise specified (ASTM E284). Positive aspecular angles are in the direction toward the illuminator axis. Research instruments let the colorist specify the angles of illumination and view. Portable instruments illuminate the sample at 45° from the normal. Portable instruments view the specimen at three to five aspecular angles (see Figure 12.17). Most five-angle instruments would view the specimen at aspecular angles of 15, 25, 45, 75, and 110°. A three-angle instrument would use either 15° (preferred) or 25, 45° and either 75 or 110° (preferred) (ASTM 2194). The aspecular angle of 45° corresponds to an illuminating/viewing geometry of 45 : 0 or 45° : 0° (as 45) using the ASTM notation.

Measuring materials colored with interference pigments is a more difficult problem. The color of an interference

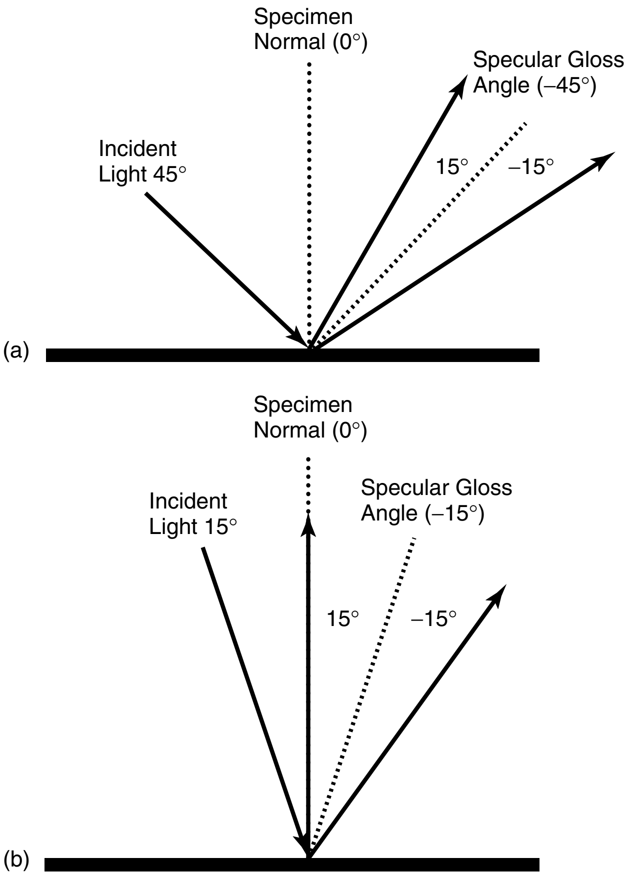

Fig. 12.18 The ASTM recommended viewing geometries used for the multiangle color measurement of interference pigments. (a) The specimen is illuminated by incident light at an anormal angle of 45°. (b) The specimen is illuminated by incident light at an anormal angle of 15°. The specimen normals and specular gloss angles are shown as anormal angles. The two viewing angles for each illumination angle are shown as aspecular angles.

pigment is seen only near the angle of the specular reflection. As the angle of the illuminator changes, the angle of specular reflection changes. The hue and chroma of the interference pigment change as the specular angle changes. Thus, an observer can see extreme changes in hue as the specular angle changes. To measure this effect, you need more than one angle of illumination in addition to multiple angles of view (see Figure 12.18a and b). The ASTM specifies that the minimum requirement for an instrument measuring interference pigments is to have two illuminators, one at an anormal angle of 45° and one at an anormal angle of 15°. Two aspecular viewing angles (−15 and +15°) are required for each illumination angle for a total of four angles (ASTM 2539). Some commercial

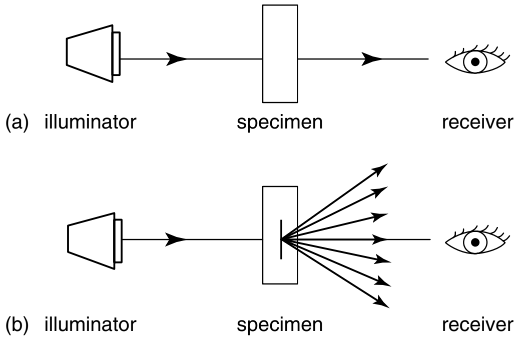

Fig. 12.19 The 0 : 0 illuminating and viewing geometry for measuring transmittance. This is only accurate for samples with no scattering (a). Hazy samples (b) cannot be measured using this illuminating and viewing geometry.

multiangle instruments have a third illuminator at an anormal angle of 65°.

Significant visual effects can be obtained by coloring a material using both metallic flakes and interference pigments. These specimens can be measured with multiangle instruments with multiple angles of illumination and multiple angles of view.

12.6.2
Measuring Transmittance

12.6.2.1 Regular and Diffuse Transmittance

Regular transmittance occurs in materials that can only absorb light. They may appear crystal clear or colored but totally transparent. Hazy materials not only absorb light but also scatter it as it passes through. They will appear to be almost, but not totally, transparent and may be clear or colored. Translucent materials reflect, transmit, and scatter light as it passes through. They will appear to be milky and may be white or colored.

12.6.2.2 Illuminating and Viewing Geometries

Although the CIE has recommended four geometries for transmittance, only three of them are commonly used for color measurement.

Spectrophotometers used for analytical chemistry measure transmittance using 0 : 0 illuminating/viewing geometry as shown in Figure 12.19(a) and (b). The illuminator illuminates the specimen along its normal and the receiver views the sample on the other side along it normal. This geometry measures transmittance accurately only for specimens that have regular transmission.

The diffuse-normal (*d* : 0) shown in Figure 12.15 and its reverse geometry normal-diffuse (0 : *d*) can be used to measure materials that transmit regularly or diffusely. To measure regular transmission, the specimen is placed away from the instrument's integrating sphere. To measure diffuse transmission, the specimen is placed in contact with the instrument's integrating sphere. When integrating sphere instruments designed for measuring reflectance are used, a white material with reflectance as close to that of the integrating sphere's wall is placed in the sample port.

12.6.2.3 Standardization

The instrument's zero point is usually set by blocking the light beam with an opaque material.

There are several ways of setting the high end of the photometric scale. The easiest technique is to make a measurement with nothing in the sample holder and set the transmittance factor to 1.0 at each wavelength. Transmittance measurements will then be relative to air. A totally colorless blank can be measured and the transmittance factor is set to 1.0 at each wavelength. It is important that this blank specimen be made of the same material and has the same thickness as the specimen to be measured. Transmittance measurements will then be relative to the clear specimen.

Liquid samples must be put into a sample holder before they are measured. An empty holder can be measured and the transmittance factor is set to 1.0 at each wavelength. Although the measurements would still be relative to air, the affect on transmission of the sample holder will be eliminated. Lastly, the liquid sample holder can be filled with water or a solvent of similar optical properties as that of the specimen. The transmittance measurements will then be relative to water or the solvent.

12.6.2.4 Measuring Transmittance

Specimens being measured should be flat with parallel sides. Curved specimens, such as lenses, are very difficult to measure because the curvature changes the path of the light. Liquid sample holders should be made of optical glass to minimize the effect of the sample holder on the measurement. Transmission changes with thickness and concentration. If two specimens are being compared, they must be prepared in the same way if the comparison is to be valid. This is particularly important if a numeric standard is used. A haze index can be calculated from multiple measurements of transmittance (ASTM 1003).

12.6.3
Radiometric Measurements

Spectral radiometers and radiometric colorimeters were designed to measure lamps, light sources, and visual display units. These instruments are closely related to spectrophotometers and colorimeters, but they need not supply an illuminator because the object being measured is the source of light.

Lamps and light sources can be measured either directly or indirectly by measuring the reflectance of a stable white reflecting surface being illuminated by the lamp or light source (ASTM 1341). Measuring televisions and visual display units usually requires that the emitted light be directly imaged on the optics of the instrument (ASTM 1336; ASTM 1455).

12.7
Summary

Color is an integral part of commerce and entertainment. Color perception depends upon the light source illuminating an object, the pigments and dyes used in manufacturing the object, and the eye of the observer whose impulses are translated by the brain as a color. Color is three dimensional and thus three terms (lightness, chroma, and hue) are needed to describe a color. Color is only one aspect of appearance. Members of the CIE developed a system to simulate the color perception. They standardized illuminants (mathematical light sources) and color-matching functions (simulating the eye–brain combination). Spectrophotometers are used to characterize the object completely. Tristimulus values and color coordinates can be calculated from the spectrophotometric measurements, the standardized illuminants, and the color-matching functions. Color differences between a target or standard color and a trial or batch color can be calculated from the measurements. For some situations, a single-number color scale is sufficient. For other situations, a three-component color difference is required. There are a number of equations available for calculating color differences.

An object may absorb light, scatter and reflect light, and/or transmit light. Reflectance factors are usually measured for opaque and translucent objects. Transmittance is usually measured for hazy and transparent objects. Spectrophotometers measure reflectance and/or transmittance at given wavelengths of the visible spectrum and calculations are done to determine tristimulus values and color coordinates. Colorimeters determine tristimulus values or color coordinates directly during the measurement using broad-band spectral devices. Colorimeters cannot usually detect metamerism. Fluorescent samples not only reflect light but also emit light. Special techniques are required to measure those samples. Measuring materials containing only traditional colorants and metallic flakes require instruments with multiple viewing angles. Measuring materials containing interference pigments require instruments with multiple illumination angles and multiple viewing angles. Measuring lamps and other light emitters require spectral radiometers and radiometric colorimeters.

References

CIE (2004) *Colorimetry*, 3rd edn, Publication No. 15:2004, Central Bureau of the Commission International de l'Éclairage (CIE), Vienna. Available from U.S. National Committee of the CIE (International Commission on Illumination), c/o Thomas M. Lemons, TLA-Lighting Consultants, Inc., 7 Pond St., Salem, MA 01970, www.cie-usnc.org.

ASTM ASTM E 308 (2006) *Practice for computing the colors of objects by using the CIE System, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

CIE (2004) *Method for Assessing the Quality of Daylight Simulators for Colorimetry*, Publication No. 51.2–1981, Central Bureau of the Commission International de l'Éclairage (CIE), Vienna. Available from U.S. National Committee of the CIE (International Commission on Illumination), c/o Thomas M. Lemons, TLA-Lighting Consultants, Inc., 7 Pond St., Salem, MA 01970, www.cie-usnc.org.

ASTM ASTM E 284 (2008) *Standard terminology of appearance, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM D 3134 (1997) *Standard practice for establishing color and gloss tolerances, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

Hunt, R.W.G. (1991) *Measuring Colour*, 2nd edn, Ellis Horwood, Chichester, West Sussex, England.

ASTM ASTM D 4449 (2008) *Standard test method for visual evaluation of gloss differences between surfaces of similar appearance, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM D 1729 (1996) *Standard practice for visual appraisal of colors and color differences of diffusely-illuminated opaque materials, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM D 3928 (2000) *Standard test method for evaluation of gloss or sheen uniformity, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM D 2616 (1996) *Standard test method for evaluation of visual color difference with a gray scale, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM E 313 (2005) *Standard practice for calculating yellowness and whiteness indices from instrumentally measured color coordinates, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM D 156 (2007) *Standard test method for Saybolt color of petroleum products (Saybolt Chromometer Method), Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM D 1209 (2005) *Standard test method for color of clear liquids (Platinum-Cobalt Scale), Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM D 1500 (2007) *Standard test method for ASTM color of petroleum products (ASTM Color Scale), Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM D 6045 (2004) *Standard test method for color of petroleum products by the automatic tristimulus method, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM D 1544 (2004) *Standard test method for color of transparent liquids (Gardner Color Scale) [Historical standard – no longer active], Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM D 6166 (2008) *Standard test method for color of naval stores and related products (Instrumental Determination of Gardner Color), Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM D 1686 (1996) *Standard test method for color of solid aromatic hydrocarbons and related materials in the molten state (Platinum Cobalt Scale), Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM D 2244 (2007) *Standard practice for calculation of color tolerances and color differences from instrumentally measured color coordinates, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

McDonald, R. (1980) Industrial pass/fail colour matching. *J. Soc. Dyers Colour.*, **96**, Part I, 372–376; Part II, 418–433; Part III, 486–495.

AATCC (1992) AATCC Test Method 173–1992, "CMC: calculation of small color differences for acceptability", *AATCC Technical Manual*, AATCC, Research Triangle Park. www.aatcc.org.

CIE (2001) *Improvement to Industrial Colour-difference Evaluation*, Publication No. 142–2001, Central Bureau of the Commission International de l'Éclairage (CIE), Vienna. Available from U.S. National Committee of the CIE (International Commission on Illumination), c/o Thomas M. Lemons, TLA-Lighting Consultants, Inc., 7 Pond St., Salem, MA 01970, www.cie-usnc.org.

DIN DIN 6176. Farbmetrische Bestimmung von Farbabständen bei Körperfarben nach der DIN99-Formel (Colorimetric evaluation of colour differences of surface colours according to DIN99 formula – available in English), DIN Deutsches Institut für Normung e. V., Burggrafenstraße 6, 10787 Berlin, Germany. www.din.de.

CIE (1995) *Geometric Tolerances for Color Measurement*, Publication No. 176–2006, Central Bureau of the Commission International de l'Éclairage (CIE), Vienna. Available from U.S. National Committee of the CIE (International Commission on Illumination), c/o Thomas M. Lemons, TLA-Lighting Consultants, Inc., 7 Pond St., Salem, MA 01970, www.cie-usnc.org.

ASTM ASTM E 179 (1996) *Standard guide for selection of geometric conditions for measurement of reflection and transmission properties of materials, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

Rich, D.C. (1988) The effect of measuring geometry on computer color matching. *Color Res. Appl.*, **13**, 113–118.

Mabon, T.J. (1992) Color measurement of plastics: which geometry is best. Presented at the Regional Technical Conference of the Society of Plastics Engineers, Inc., in *Color Tolerances: Measuring up to Today's*

Standards, Cherry Hill, New Jersey. September 14–16.

Billmeyer, F.W. and Marcus Jr., R.T. (1969) Effect of illuminating and viewing geometry on the color coordinates of samples with various surface textures. *Appl. Opt.*, **8**, 1763–1768.

Budde, W. (1980) The gloss trap in diffuse reflectance measurements. *Color Res. Appl.*, **5**, 73–75.

ASTM ASTM E 259 (2006) *Standard practice for preparation of pressed powder white reflectance factor transfer standards for hemispherical and bi-directional geometries, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

Carter, E.C., Billmeyer Jr., F.W., and Rich, D.C. (2000) Guide to Material Standards and their Use in Color Measurement, *ISCC Technical Report 89-1*. Available from the Inter-Society Color Council, Reston. www.iscc.org.

Clark, F.J.J. (1971) Ceramic colour standards. *Die. Farbe.*, **20**, 299–306.

ASTM ASTM E 1345 (1998) *Standard practice for reducing the effect of variability of color measurement by use of multiple measurements, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM E 991 (2006) *Standard practice for color measurement of fluorescent specimens using the one-monochromator method, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

McKinnon, R.A. (1987) Methods of measuring the colour of opaque fluorescent materials. *Rev. Prog. Coloration*, **17**, 56–60.

Gundlach, D. and Terstiege, H. (1994) Problems in measurement of fluorescent materials. *Color Res. Appl.*, **19**, 427–436.

ASTM ASTM E 2175 (2001) *Practice for specifying the geometry of multiangle spectrophotometers, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM E 2194 (2003) *Practice for multiangle color measurement of metal flake pigmented materials, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM E 2539 (2008) *Standard practice for multiangle color measurement of interference pigments, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM D 1003 (2007) *Standard test method for haze and luminous transmittance of transparent plastics, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM E 1341 (2006) *Practice for obtaining spectroradiometric data from radiant sources for colorimetry, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM E 1336 (1996) *Test method for obtaining colorimetric data from a visual display unit by spectroradiometry, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

ASTM ASTM E 1455 (2003) *Practice for obtaining colorimetric data from a visual display unit using tristimulus colorimeters, Annual Book of ASTM Standards*, ASTM International, West Conshohocken. www.astm.org.

14
Molecular Fluorescence

Bernard Valeur

14.1 Introduction 479
14.2 Characteristics of Fluorescence Emission 481
14.2.1 Radiative and Nonradiative Transitions between Electronic States 481
14.2.1.1 Internal Conversion 482
14.2.1.2 Fluorescence 482
14.2.1.3 Intersystem Crossing and Subsequent Processes 483
14.2.2 Lifetimes and Quantum Yields 484
14.2.2.1 Excited-State Lifetimes 484
14.2.2.2 Quantum Yields 485
14.2.3 Emission and Excitation Spectra 486
14.2.3.1 Emission Spectra 486
14.2.3.2 Excitation Spectra 487
14.2.3.3 Stokes Shift 488
14.2.4 Fluorescence Anisotropy 488
14.3 Effects of Intermolecular Photophysical Processes on Fluorescence 489
14.3.1 Main Photoinduced Intermolecular Processes 489
14.3.2 Fluorescence Quenching 490
14.3.2.1 Dynamic Quenching 490
14.3.2.2 Static Quenching 492
14.3.3 Photoinduced Electron Transfer 494
14.3.4 Formation of Excimers and Exciplexes 495
14.3.4.1 Excimers 495
14.3.4.2 Exciplexes 498
14.3.5 Photoinduced Proton Transfer 498
14.3.5.1 Equations for Excited-State Deprotonation 499
14.3.5.2 Determination of the Excited-State pK^* 500
14.3.5.3 pH Dependence of Absorption and Emission Spectra 501
14.3.6 Excitation Energy Transfer 501
14.4 Applications of Molecular Fluorescence 502
14.4.1 Historical Introduction 502

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

14.4.2 Estimation of Local Physical Parameters 503
14.4.2.1 Viscosity 503
14.4.2.2 Temperature 508
14.4.2.3 Pressure 511
14.4.3 Chemical Sensing 513
14.4.3.1 Fluorescent Molecular Sensors 513
14.4.3.2 Microfabricated Analysis Systems 519
14.4.4 Tracers in Biology 521
14.4.4.1 Fluorescent Proteins 522
14.4.4.2 Semiconductor Nanocrystals 523
14.4.5 Clinical Diagnostics 523
14.4.5.1 Critical Care Analysis 523
14.4.5.2 Angiography 524
14.4.5.3 Bladder Tumor Detection 524
14.4.6 Miscellaneous 526
14.4.6.1 Fluorescence LIDAR 526
14.4.6.2 Food Science 527
14.4.6.3 Forensics 527
14.4.6.4 Counterfeit Detection 528
References 529
Further Reading 531

14.1
Introduction

Molecular fluorescence is extensively used in physical, chemical, material, biological, and medical sciences as a tool of detection/analysis, visualization, investigation of local properties, diagnostics, and so on. In fact, fluorescent compounds can be used not only for mere visualization but also as probes, indicators, sensors, and tracers for providing information on local physical or chemical parameters: pressure, temperature, viscosity, polarity, pH, concentrations of ionic or neutral species, and so on (Valeur, 2002; Lakowicz, 2006; Schulman, 1985–1993; Lakowicz, 1991–2007; Valeur and Brochon, 2001; Kraayenhof and Visser, 2003; Hof, Hutterer and Fidler, 2005; Berberan-Santos, 2007; Jameson, Croney and Moens, 2003; Valeur, Berberan-Santos and Martin, 2007).

Once a molecule is excited by absorption of a photon, it can return to the ground state via many pathways (Figure 14.1). The radiative pathway, that is, emission of photons with retention of spin multiplicity, is called *fluorescence*. Another radiative pathway is *phosphorescence*, but this process first requires an intersystem crossing from the initially singlet excited state to a triplet state. This is the major distinction to be made between fluorescence and phosphorescence, and not only the fact that the decay of fluorescence intensity following light pulse excitation is generally much faster than that of phosphorescence. The nonradiative pathway of de-excitation called *internal conversion* is the direct return to the ground state without emission of photon. In the excited state, the molecule may undergo intramolecular processes: charge transfer, conformational change, and so on.

Interactions in the excited state with other molecules may lead to de-excitation: electron transfer, proton transfer, energy transfer, excimer or exciplex formation, and photochemical reaction. These de-excitation pathways may compete with fluorescence emission if they take place at a time scale comparable with the average time (called *lifetime*) during which the molecules stay in the excited state. This average time represents the *experimental time window* for observation of dynamic processes. The characteristics of fluorescence (spectrum, quantum yield, and lifetime), which are affected by any excited-state process

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

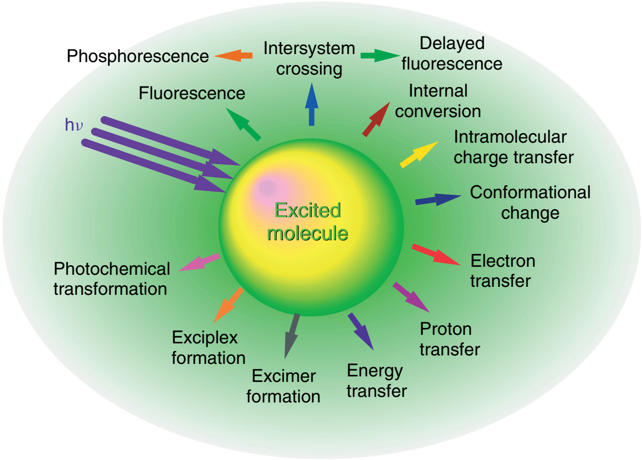
Phosphorescence
Intersystem crossing
Delayed fluorescence
Fluorescence
Internal conversion
hν
Intramolecular charge transfer
Excited molecule
Conformational change
Photochemical transformation
Electron transfer
Exciplex formation
Proton transfer
Excimer formation
Energy transfer

Fig. 14.1 Excited-state processes following excitation of a molecule. (Reproduced with permission from Valeur (2002). Copyright Wiley-VCH Verlag GmbH & Co. KGaA.)

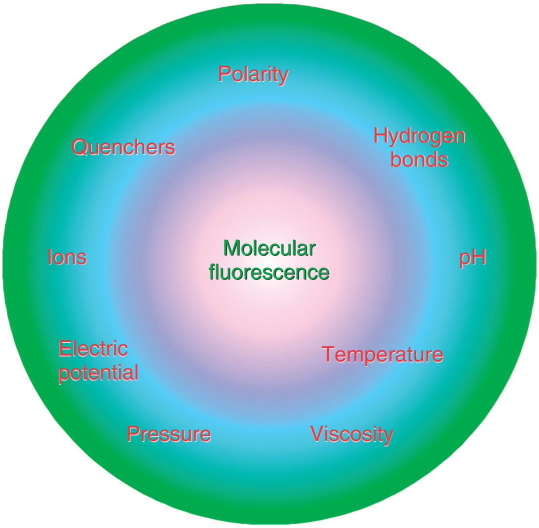
Polarity
Quenchers
Hydrogen bonds
Ions
Molecular fluorescence
pH
Electric potential
Temperature
Pressure
Viscosity

Fig. 14.2 Physical and chemical parameters affecting the emission of fluorescence. (Reproduced with permission from Valeur (2002). Copyright Wiley-VCH Verlag GmbH & Co. KGaA.)

involving interactions of the excited molecule with its close environment, can then provide information on such a microenvironment. Figure 14.2 shows the physical and chemical parameters that characterize a microenvironment and can thus affect the fluorescence characteristics of a molecule.

In this chapter, in the second section the characteristics of fluorescence emission are described. In the third section, the effects of photoinduced intermolecular processes on fluorescence are discussed. The fourth section presents a selection of applications of molecular fluorescence in various fields.

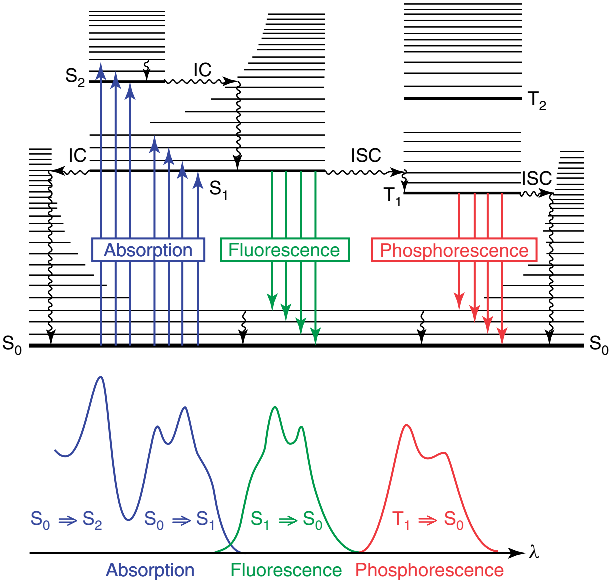

Fig. 14.3 Perrin–Jablonski diagram and relative position of the spectra. IC, internal conversion; ISC, intersystem crossing. (Reproduced with permission from Valeur (2002). Copyright Wiley-VCH Verlag GmbH & Co. KGaA.)

14.2
Characteristics of Fluorescence Emission

In this section, the characteristics of fluorescence emission are described for molecules in dilute solution in the absence of any interaction with other molecules (Valeur, 2002) chap. 3. The photophysical processes resulting from such interactions are discussed in Section 14.3.

14.2.1
Radiative and Nonradiative Transitions between Electronic States

The processes that are involved in the excitation and de-excitation of a fluorescent molecule are conveniently visualized by means of the Perrin–Jablonski diagram (Figure 14.3): photon absorption, internal conversion, fluorescence, intersystem crossing, phosphorescence, and so on. Vibrational levels are associated with all electronic states. The singlet states are denoted S_0 (fundamental electronic state), $S_1, S_2, \ldots$ and the triplet states, $T_1, T_2, \ldots$ It is important to note that absorption is very fast ($\approx 10^{-15}$ seconds) with respect to all other processes (so that there is no concomitant displacement of the nuclei according to the Franck–Condon principle).

In a solution at room temperature, the majority of molecules are in the 0 (lowest) vibrational energy level of S_0. Therefore, the vertical arrows corresponding to absorption start from this level. Absorption of a photon can bring a molecule to one of the vibrational levels of $S_1, S_2, \ldots$. The subsequent possible de-excitation processes are described.

Tab. 14.1 Characteristic times involved in the excitation and de-excitation of a molecule. (From Valeur (2002).)

Process or lifetime	Characteristic time
Absorption	10^{-15} s
Vibrational relaxation	10^{-12}–10^{-10} s
Lifetime of the singlet excited state S_0	10^{-10}–10^{-7} s ⇒ fluorescence
Intersystem crossing	10^{-10}–10^{-8} s
Internal conversion	10^{-11}–10^{-9} s
Lifetime of the triplet excited state T_1	$>10^{-6}$ s ⇒ phosphorescence

The characteristic times of these processes are given in Table 14.1.

14.2.1.1 Internal Conversion

Internal conversion is a nonradiative transition between two electronic states of the same spin multiplicity.

When a molecule is excited to an energy level higher than the lowest vibrational level of the first electronic state, vibrational relaxation – and internal conversion $S_n \to S_1$ if the singlet excited state is higher than S_1 – leads the excited molecule toward the lowest vibrational level of the S_1 singlet state. The excess of vibrational energy is transferred to the solvent during collisions of the excited molecule with the surrounding solvent molecules.

Internal conversion from S_1 to S_0 is less efficient than that from S_2 to S_1 owing to the much larger energy gap between S_1 and S_0. Therefore, internal conversion from S_1 to S_0 can compete with emission of photons (fluorescence) and intersystem crossing to the triplet state.

14.2.1.2 Fluorescence

Emission of photons from an excited molecule without change in multiplicity is called *fluorescence*. Besides a few exceptions, fluorescence emission occurs from S_1 and therefore its characteristics (except polarization) do not depend on the excitation wavelength.

The fluorescence spectrum is located at higher wavelengths (lower energy) than the absorption spectrum because of the energy loss in the excited state by vibrational relaxation (Figure 14.3). The gap (expressed in wavenumbers) between the maximum of the first absorption band and the maximum of fluorescence is called *Stokes shift*.

In general, the differences between the vibrational levels are similar in the ground and excited states, so that the fluorescence spectrum often resembles the first absorption band (“mirror image” rule).

Emission of a photon is a process that is as fast as absorption of a photon ($\approx 10^{-15}$ second), but excited molecules stay in the S_1 state for a certain time (a few tens of picoseconds to a few hundreds of nanoseconds, according to the type of molecule and the medium) before emitting a photon or undergoing other de-excitation processes (internal conversion and intersystem crossing). Thus, after excitation of a population of molecules by a very short pulse of light, the fluorescence intensity decreases exponentially with a characteristic time, called *excited-state lifetime*, reflecting the average time during which the molecules stay in the S_1 excited state (10^{-10}–10^{-7} seconds) before returning to the ground state (see Section 14.2.2.1). The fluorescence intensity decay is formally comparable with a radioactive decay that is also exponential, with a characteristic time called *radioactive period*, characterizing the average lifetime of a radioelement before disintegration.

The emission of fluorescence photons is a spontaneous process. Another emission process, called *stimulated emission*, can be observed under certain conditions, for instance, in dye lasers.

14.2.1.3 Intersystem Crossing and Subsequent Processes

A third possible de-excitation process from S_1 is intersystem crossing toward the T_1 triplet state, followed by other processes (Figure 14.3).

Intersystem Crossing *Intersystem crossing* is a nonradiative transition between two isoenergetic vibrational levels belonging to electronic states of different multiplicities.

Once an excited molecule has relaxed to the 0 vibrational level of the S_1 state, it can move to the isoenergetic vibrational level of the T_n triplet state; then, vibrational relaxation brings it into the lowest vibrational level of T_1. Intersystem crossing may compete with the other pathways of de-excitation from S_1 (fluorescence and internal conversion $S_1 \rightarrow S_0$).

Crossing between states of different multiplicity is in principle forbidden, but spin–orbit coupling can be large enough to make it possible. The presence of heavy atoms (such as I, Br, Pb, etc.) favors intersystem crossing because it increases spin–orbit coupling.

Phosphorescence versus Nonradiative De-excitation The emission of photons from the triplet state T_1 is called *phosphorescence*. Although the transition $T_1 \rightarrow S_0$ is forbidden, phosphorescence can be observed because of spin–orbit coupling. The lifetime of the triplet state T_1 is longer ($>10^{-6}$ second) than that of the singlet state S_1.

In solution at room temperature, the numerous collisions of the molecules in the triplet state state T_1 with solvent molecules favor intersystem crossing and vibrational relaxation to S_0. Nonradiative de-excitation from the triplet state T_1 is thus predominant. On the contrary, at low temperature and/or in a rigid medium, phosphorescence can be observed. Under these conditions, the lifetime of the triplet state may be long enough to observe phosphorescence on a time scale up to seconds, minutes, or even more.

The phosphorescence spectrum is located at wavelengths higher than the fluorescence spectrum (Figure 14.3) because the energy of the lowest vibrational level of the triplet state T_1 is lower than that of the singlet state S_1.

Delayed Fluorescence

- **Thermally activated delayed fluorescence**: When the energy difference between S_1 and T_1 is small and when the lifetime of T_1 is long enough, reverse intersystem crossing $T_1 \rightarrow S_1$ can occur. Thus, the emission spectrum is identical to that of normal fluorescence, but the decay

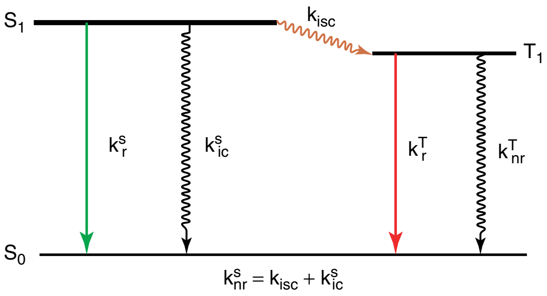

Scheme 14.1 De-excitation processes with relevant rate constants.

time constant is much longer because the molecules stay in the triplet state before emitting from S_1. The efficiency of delayed fluorescence increases with increasing temperature, because the process is thermally activated. It is generally called *delayed fluorescence of E-type* because it has been observed for the first time with eosin. E-type-delayed fluorescence is usually a weak phenomenon, but it is exceptionally intense in the case of fullerene C_{70} because the quantum yield for triplet formation is close to 1, the S_1-T_1 energy gap is quite small, and the intrinsic triplet lifetime is in the order of several milliseconds (Berberan-Santos and Garcia, 1996).

- **Triplet–triplet annihilation**:In concentrated solutions, a collision between two molecules in state T_1 can provide enough energy to allow one of them to return to state S_1. Such a triplet–triplet annihilation thus leads to a delayed fluorescence emission. It is called *delayed fluorescence of P type* because it has been observed for the first time with pyrene. The decay time constant of this type of delayed fluorescence is half the lifetime of the triplet state in dilute solution, and the intensity has a characteristic quadratic dependence with excitation light intensity.

14.2.2
Lifetimes and Quantum Yields

14.2.2.1 Excited-State Lifetimes

The various processes of de-excitation are represented in Scheme 14.1 with the relevant rate constants. The de-excitation processes that result from intermolecular interactions are not considered in this section. They are described in Section 14.3.

The rate constants are denoted as follows:

k_r^S is the rate constant for radiative deactivation $S_1 \rightarrow S_0$ with emission of fluorescence.
k_{ic}^S is the rate constant for internal conversion $S_1 \rightarrow S_0$.
k_{isc} is the rate constant for intersystem crossing.
k_r^T is the rate constant for radiative deactivation $T_1 \rightarrow S_0$ with emission of phosphorescence.
k_{nr}^T is the rate constant for nonradiative deactivation (intersystem crossing) $T_1 \rightarrow S_0$.

It is convenient to introduce the overall rate constant k_{nr}^S for nonradiative deactivation from S_1 : $k_{\mathrm{nr}}^S = k_{\mathrm{ic}}^S + k_{\mathrm{isc}}$ (14.1)

Let us consider a dilute solution of a fluorescent molecule M whose concentration is [M] (in mole per liter). A pulse of light supposed to be infinitely short (i.e., a δ-function (Dirac)) at time 0 brings a certain number of molecules M to the S_1 excited state by absorption of photons. Then, these excited molecules return to S_0 either radiatively or nonradiatively or undergo intersystem crossing. As in classical chemical kinetics, the rate of disappearance of excited molecules is expressed by the following differential equation:

$$-\frac{d[^1\mathrm{M}^*]}{dt} = \left(k_r^S + k_{\mathrm{nr}}^S\right)[^1\mathrm{M}^*] \qquad (14.2)$$

The time evolution of the concentration in excited molecules $[^1\mathrm{M}^*]$ is obtained by integration of this equation:

$$[^1\mathrm{M}^*] = [^1\mathrm{M}^*]_0 \exp\left(-\frac{t}{\tau_S}\right) \qquad (14.3)$$

where $[^1\mathrm{M}^*]_0$ is the concentration in excited molecules at time 0 resulting from pulse light excitation, and τ_S is the lifetime of excited state S_1, given as

$$\tau_S = \frac{1}{k_r^S + k_{\mathrm{nr}}^S} \qquad (14.4)$$

The fluorescence intensity is defined as the amount of photons (in mol, or its equivalent, in einstein; 1 einstein = 1 mol of photons) emitted per unit time (second) and per unit volume of solution (liter) according to

$$\mathrm{M}^* \xrightarrow{k_r^S} \mathrm{M} + \text{photon} \qquad (14.5)$$

The fluorescence intensity i_F at time t after excitation by a very short pulse of light at time 0 is proportional, at any time, to the instantaneous concentration of molecules still excited $[^1\mathrm{M}^*]$; the proportionality factor is the rate constant for radiative de-excitation k_r^S:

$$\begin{aligned} i_F(t) &= k_r^S[^1\mathrm{M}^*] \\ &= k_r^S[^1\mathrm{M}^*]_0 \exp\left(-\frac{t}{\tau_S}\right) \end{aligned} \qquad (14.6)$$

$i_F(t)$, the *δ-pulse response* of the system, decreases according to a single exponential.

In practice, the measured signal I_F representing the fluorescence intensity is proportional to i_F, the proportionality factor depending on instrumental conditions. The numerical value of I_F is thus generally obtained in an arbitrary scale.

If the only way of de-excitation from S_1 to S_0 was fluorescence emission, the lifetime would be $1/k_r^S$: it is called *radiative lifetime* (in preference to *natural lifetime*) and denoted τ_r.

The lifetime of a homogeneous population of fluorophores is in principle independent of the excitation wavelength because internal conversion and vibrational relaxation are always very fast in solution and emission arises from the lowest vibrational level of state S_1.

If a fraction of excited molecules reach the triplet state from which they return to the ground state either radiatively or nonradiatively, the concentration of molecules in the triplet state decays exponentially with a time constant τ_T representing the lifetime of the triplet state:

$$\tau_{\mathrm{T}} = \frac{1}{k_r^T + k_{\mathrm{nr}}^T} \qquad (14.7)$$

14.2.2.2 Quantum Yields

The *fluorescence quantum yield* Φ_F is the fraction of excited molecules that return to the ground state S_0 with emission of fluorescence photons.

$$\Phi_F = \frac{k_r^S}{k_r^S + k_{\text{nr}}^S} = k_r^S \tau_S \tag{14.8}$$

In other words, the fluorescence quantum yield is the ratio of the number of emitted photons (over the whole duration of the decay) and the number of absorbed photons. According to Eq. (14.6), the ratio of the δ-pulse response $i_F(t)$ and the number of absorbed photons is given by

$$\frac{i_F(t)}{[{}^1\text{M}^*]_0} = k_r^S \exp\left(-\frac{t}{\tau_S}\right) \tag{14.9}$$

The fluorescence quantum yield is then obtained by integration of this relation over the whole duration of the decay (mathematically from 0 to infinity):

$$\frac{1}{[{}^1\text{M}^*]_0} \int_0^\infty i_F(t) = k_r^S \tau_S = \Phi_F \tag{14.10}$$

The quantum yields of intersystem crossing (Φ_{isc}) and phosphorescence (Φ_P) are given by

$$\Phi_{\text{isc}} = \frac{k_{\text{isc}}}{k_r^S + k_{\text{nr}}^S} = k_{\text{isc}}^S \tau_S \tag{14.11}$$

$$\Phi_{\text{P}} = \frac{k_r^T}{k_r^T + k_{\text{nr}}^T} \Phi_{\text{isc}} \tag{14.12}$$

Using the radiative lifetime, as previously defined, the fluorescence quantum yield can also be written as

$$\Phi_F = \frac{\tau_S}{\tau_r} \tag{14.13}$$

It is interesting to note that when the fluorescence quantum yield and the excited-state lifetime of a fluorophore are measured under the same conditions, the nonradiative and radiative rate constants can be easily calculated by means of the following relations:

$$k_r^S = \frac{\Phi_F}{\tau_S}, \quad k_r^S = \frac{1}{\tau_S}(1 - \Phi_F) \tag{14.14}$$

14.2.3
Emission and Excitation Spectra

14.2.3.1 Emission Spectra

The variations in fluorescence intensity as a function of the observation wavelength λ_E, for a fixed excitation wavelength λ_F, represents the *emission spectrum*. It is convenient to express the steady-state fluorescence intensity (i.e., under continuous illumination) per absorbed photon as a function of the wavelength of the emitted photons, denoted $F_\lambda(\lambda_F)$ (in m^{-1} or nm^{-1}), normalized so that the integral over the whole spectrum is equal to the fluorescence quantum yield Φ_F:

$$\int_0^\infty F_\lambda(\lambda_F) d\lambda_F = \Phi_F \tag{14.15}$$

$F_\lambda(\lambda_F)$ represents the *fluorescence spectrum* or *emission spectrum*: it reflects the distribution of the probability of the various transitions from the lowest vibrational level of S_1 to the various vibrational levels of S_0.

Assuming that the number of absorbed photons follows the Beer–Lambert law, the intensity of the transmitted light is given by

$$I_T(\lambda_E) = I_0(\lambda_E) \exp\left[-2.3\, \varepsilon(\lambda_E)\, l\, c\right] \tag{14.16}$$

where $I_0(\lambda_E)$ is the intensity of the incident light, $\varepsilon(\lambda_E)$ denotes the molar absorption coefficient of the fluorophore at wavelength λ_E (in liter per mole per centimeter), l is the optical path in the sample (in centimeter), and c is the concentration (in mol per liter).

The quantity $\varepsilon(\lambda_E)lc$ represents the absorbance $A(\lambda_E)$ at wavelength λ_E.

The difference between the intensity of the incident light $I_0(\lambda_E)$ and the intensity of the transmitted light $I_T(\lambda_E)$ represents the absorbed intensity $I_A(\lambda_E)$.

The fluorescence intensity can be written as

$$I_F(\lambda_E, \lambda_F) = k\ F_\lambda(\lambda_F)\ I_A(\lambda_E) \qquad (14.17)$$

where k is a proportionality factor that depends on several parameters and in particular on the optical configuration for observation (i.e., the solid angle through which the instrument collects fluorescence that is, in fact, emitted in all directions) and on the bandwidth of the monochromators (i.e., the widths of the entrance and exit slits).

Finally, the general expression for the fluorescence intensity is the following:

$$\begin{aligned} I_F(\lambda_E, \lambda_F) &= k\ F_\lambda(\lambda_F)\ I_0(\lambda_E) \\ &\quad \times \{1 - \exp[-2.3\varepsilon(\lambda_E)lc]\} \end{aligned} \qquad (14.18)$$

In highly diluted solutions, the exponential term can be approximated to $1 - 2.3\varepsilon lc$, so that the fluorescence intensity becomes proportional to the concentration:

$$\begin{aligned} I_F(\lambda_E, \lambda_F) &\cong k\ F_\lambda(\lambda_F)\ I_0(\lambda_E) \\ &\quad \times [2.3\varepsilon(\lambda_E)lc] \\ &= 2.3\ k\ F_\lambda(\lambda_F)\ I_0(\lambda_E)\ A(\lambda_E) \end{aligned} \qquad (14.19)$$

It should be emphasized that such a proportionality is valid only for low absorbances. Deviation from a linear variation increases with increasing absorbance.

In practice, the variations in I_F measured as a function of wavelength λ_F, for a fixed excitation wavelength λ_E, reflect the variations in $F_\lambda(\lambda_F)$ and thus provide the fluorescence spectrum. Because the proportionality factor k is generally unknown, the numerical value of the measured intensity I_F has no meaning, and generally speaking, I_F is expressed in arbitrary units. It should be noted that k depends on the wavelength because the transmission of the monochromator and the sensitivity of the detector are wavelength dependent. Therefore, correction of spectra is necessary for quantitative measurements.

Spectrofluorometry is a very sensitive technique – up to 1000 times more sensitive than spectrophotometry. This is because the fluorescence intensity is measured above a low background level, whereas in the measurement of low absorbances, two large signals that are slightly different are compared. Thanks to outstanding progress in instrumentation, it is now possible in some cases to even detect a single fluorescent molecule.

14.2.3.2 **Excitation Spectra**

The variations in fluorescence intensity as a function of the excitation wavelength λ_E, for a fixed observation wavelength λ_F, represent the *excitation spectrum*. In Eq. (14.19), when the observation wavelength λ_F is fixed, the variations in fluorescence intensity reflect the evolution of the product $I_0(\lambda_E)\ A(\lambda_E)$. Provided that the wavelength dependence of the incident light $I_0(\lambda_E)$ is compensated, the sole term to be taken into consideration is $A(\lambda_E)$, which represents the absorption spectrum. The corrected excitation spectrum is thus identical in shape with the absorption spectrum provided that there is a single species in the ground state.

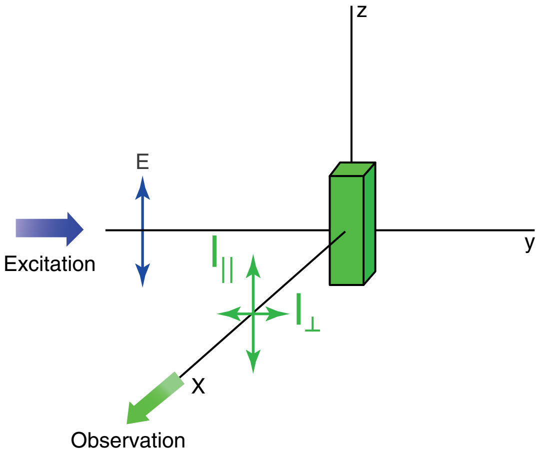

Scheme 14.2 Usual optical configuration for fluorescence anisotropy measurements.

In contrast, when several species are present, or when a sole species exists in different forms in the ground state (aggregates, complexes, tautomeric forms, etc.), the excitation and absorption spectra are no longer superimposable. Comparison between absorption and emission spectra often provides useful information.

14.2.3.3 Stokes Shift

The *Stokes shift* is the difference, expressed in wavenumbers, between the maximum of the first absorption band and the maximum of fluorescence spectrum: $\Delta\overline{v} = \overline{v}_a - \overline{v}_f$.

This parameter depends on the structure of the molecule and on its microenvironment. For instance, when the dipole moment of a fluorescent molecule is higher in the excited state than in the ground state, the Stokes shift increases with solvent polarity. This dependence is a means to estimate the polarity of an unknown environment.

From the practical point of view, the detection of a fluorescent species is easier when the Stokes shift is larger.

14.2.4
Fluorescence Anisotropy

Fluorescence anisotropy, also called *emission anisotropy*, is used to characterize the polarization of fluorescence resulting from photoselection when an assembly of molecules is illuminated by a linearly polarized light (Valeur, 2002 Chapter 5). It is defined as

$$r = \frac{I_{||} - I_{\perp}}{I_{||} + 2I_{\perp}} \qquad (14.20)$$

where $I_{||}$ and $I_{\perp}$ are the intensities measured with a linear polarizer for emission parallel and perpendicular, respectively, to the electric vector of linearly polarized incident light (which is often vertical) (Scheme 14.2). The quantity $I_{||} + 2I_{\perp}$ is proportional to the total fluorescence intensity I.

Fluorescence polarization may also be characterized by the polarization ratio, also called the *degree of polarization*, defined as

$$p = \frac{I_{||} - I_{\perp}}{I_{||} + I_{\perp}} \qquad (14.21)$$

The relation between r and p is

$$r = \frac{2p}{(3-p)} \tag{14.22}$$

For parallel absorbing and emitting transition moments, the (theoretical) values are $(r, p) = (1/2, 2/5)$; when the transition moments are perpendicular, the values are $(r, p) = (-1/3, -1/5)$.

In many cases, it is preferable to use emission anisotropy because it is additive; the overall contribution of n components r_i, each contributing to the total fluorescence intensity with a fraction $f_i = I_i/I$, is given by

$$r = \sum_{i=1}^{n} f_i r_i \quad \text{with} \quad \sum_{i=1}^{n} f_i = 1 \tag{14.23}$$

On continuous illumination, the measured emission anisotropy is called *steady-state emission anisotropy* and is given by

$$\bar{r} = \frac{\int_0^\infty r(t) I(t) dt}{\int_0^\infty I(t) dt} \tag{14.24}$$

where $r(t)$ is the instantaneous anisotropy and $I(t)$ is the fluorescence intensity at time t following a δ-pulse excitation.

The term *fundamental emission anisotropy* describes a situation in which no depolarizing events occur subsequent to the initial formation of the emitting state, such as those caused by rotational diffusion or energy transfer. It also assumes that there is no overlap between differently polarized transitions. The (theoretical) value of the fundamental emission anisotropy depends on the angle α between the absorption and emission transition moments according to the relation $r_0 =< 3\cos^2\alpha > /5$ where $<>$ denotes an average over the orientations of the photoselected molecules. r_0 can take on values ranging from $-1/5$ for $\alpha = 90°$ (perpendicular transition moments) to $2/5$ for $\alpha = 0°$ (parallel transition moments). In spite of the severe assumptions, the expression is frequently used to determine relative transition moment angles.

In time-resolved fluorescence with δ-pulse excitation, the theoretical value at time zero is identified with the fundamental emission anisotropy.

14.3
Effects of Intermolecular Photophysical Processes on Fluorescence

14.3.1
Main Photoinduced Intermolecular Processes

The interaction between an excited molecule M* and another molecule Q may occur of course as long as M is in its excited state (Valeur, 2002, Chapter 4). The key parameter is thus the lifetime of the excited state of M*, which is the experimental time window through which processes of similar duration, and competing with the intrinsic de-excitation, can be observed. It is subsequently denoted τ_0 and is given by Eq. (14.4) rewritten as

$$\tau_0 = \frac{1}{k_r + k_{\mathrm{nr}}} \tag{14.25}$$

Scheme 14.3 illustrates the competition between the intermolecular interaction characterized by the rate constant k_q for the bimolecular process and the intrinsic de-excitation of M* whose rate constant is $k_M = 1/\tau_0$.

The main intermolecular photophysical processes are presented in Table 14.2.

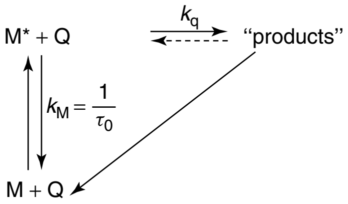

Scheme 14.3 Competition in the excited state between intrinsic de-excitation and interaction with another molecule Q.

Tab. 14.2 Main photophysical processes. (From Valeur (2002).)

Photophysical process	$M^* + Q \rightarrow$ products	**Donor**	**Acceptor**
Collision with a heavy atom (e.g., I^-, Br in CBr_4) or a paramagnetic species (e.g., O_2, NO)	$M^* + Q \rightarrow M + Q + \text{heat}$		
Electron transfer	$D^* + A \rightarrow D^{\cdot+} + A^{\cdot-}$	*D**	A
	$A^* + D \rightarrow A^{\cdot-} + D^{\cdot+}$	*D*	A*
Excimer formation	$M^* + M \rightarrow (MM)^*$		
Exciplex formation	$D^* + A \rightarrow (DA)^*$	D*	A
	$A^* + D \rightarrow (DA)^*$	D	A*
Proton transfer	$AH^* + B \rightarrow A^{-*} + BH^+$	AH*	B
	$B^* + AH \rightarrow BH^{+*} + A^-$	AH	B*
Energy transfer			
Heterotransfer	$D^* + A \rightarrow D + A^*$	D*	A
Homotransfer	$M^* + M \rightarrow M + M^*$	M*	M

It should be noted that most of them involve a fast transfer process from a donor to an acceptor: electron transfer, proton transfer, and energy transfer.

Because of the additional way of de-excitation due to the interaction with Q, the fluorescence quantum yield of M is decreased. The loss of fluorescence intensity is called *fluorescence quenching* irrespective of the nature of the competing intermolecular process and even if this process leads to a fluorescent species (the term quenching applies only to the initially excited molecule). Another consequence is the faster decrease in the fluorescence intensity after excitation by a light pulse.

If Q is in large excess, the excited-state process occurs under nondiffusive conditions. However, if Q is not in large excess and if mutual approach of M* and Q is possible during the excited-state lifetime, the bimolecular excited-state process is then *diffusion controlled*, and the rate constant is time dependent.

14.3.2
Fluorescence Quenching

14.3.2.1 Dynamic Quenching

Stern–Volmer Kinetics In the first approach, let us assume, that the experimental quenching rate constant k_q is time independent. According to the simplified Scheme 14.3, the time evolution of the concentration of M* following a δ-pulse excitation obeys the following differential equation:

$$\frac{d[\mathrm{M}^*]}{dt} = -\left(k_M + k_q[\mathrm{Q}]\right)\,[\mathrm{M}^*] = -\left(\frac{1}{\tau_0} + k_q[\mathrm{Q}]\right)\,[\mathrm{M}^*] \tag{14.26}$$

With the initial condition $[\mathrm{M}^*] = [\mathrm{M}^*]_0$ at $t = 0$, integration of this differential equation leads to

$$[\mathrm{M}^*] = [\mathrm{M}^*]_0\,\exp\left\{-\left(\frac{1}{\tau_0} + k_q[\mathrm{Q}]\right)t\right\} \tag{14.27}$$

The fluorescence intensity is obtained by multiplying the concentration of M* by the radiative rate constant:

$$\begin{aligned} i(t) &= k_r[\mathrm{M}^*] \\ &= k_r[\mathrm{M}^*]_0\,\exp\left\{-\left(\frac{1}{\tau_0} + k_q[\mathrm{Q}]\right)t\right\} \\ &= i(0)\,\exp\left\{-\left(\frac{1}{\tau_0} + k_q[\mathrm{Q}]\right)\,t\right\} \end{aligned} \tag{14.28}$$

where k_r is the radiative rate constant of M*. The fluorescence decay is thus a single exponential whose time constant is given by

$$\tau = \frac{1}{\dfrac{1}{\tau_0} + k_q\,[\mathrm{Q}]} = \frac{\tau_0}{1 + k_q\,\tau_0\,[\mathrm{Q}]} \tag{14.29}$$

Hence,

$$\frac{\tau_0}{\tau} = 1 + k_q\,\tau_0\,[\mathrm{Q}] \tag{14.30}$$

The fluorescence quantum yields in the absence and in the presence of quencher are, respectively,

$$\Phi_0 = k_r\,\tau_0 \tag{14.31}$$

and

$$\Phi = \frac{k}{k_r + k_{\mathrm{nr}} + k_q[\mathrm{Q}]} = \frac{k_r}{1/\tau_0 + k_q[\mathrm{Q}]} \tag{14.32}$$

The ratio leads to the *Stern–Volmer relation*:

$$\frac{\Phi_0}{\Phi} = \frac{I_0}{I} = 1 + k_q\,\tau_0[\mathrm{Q}] = 1 + K_{\mathrm{SV}}[\mathrm{Q}] \tag{14.33}$$

where I_0 and I are the steady-state fluorescence intensities (for a couple of wavelengths λ_E and λ_F) in the absence and in the presence of quencher, respectively. $K_{\mathrm{SV}} = k_q\tau_0$ is the Stern–Volmer constant. Generally, the ratio I_0/I is plotted versus the quencher concentration (Stern–Volmer plot). If the variation is found to be linear, the slope yields the Stern–Volmer constant. Then, k_q can be calculated if the excited-state lifetime in the absence of quencher is known.

Time-resolved experiments in the absence and in the presence of quencher allow us to check whether the fluorescence decay is, in fact, a single exponential and provide directly the value of k_q.

Two cases are to be distinguished:

- If the bimolecular process is not diffusion-limited: $k_q = pk_1$, where p is the probability of reaction for the encounter pair (often called efficiency), and k_1 is the diffusional rate constant.
- If the bimolecular process is diffusion-limited: k_q is identical to the diffusional rate constant k_1, which can be written in the following simplified form (proposed for the first

time by Smoluchowski):

$$k_1 = 4\pi N R_c D \quad (\text{in l mol}^{-1}\ \text{s}^{-1}) \tag{14.34}$$

where R_c is the distance of closest approach (in centimeter), D is the mutual diffusion coefficient (in square centimeter per second), N is equal to $N_A/1000$, N_A being Avogadro's constant. The distance of closest approach is generally taken as the sum of the radii of the two molecules (R_M for the fluorophore and R_Q for the quencher). The mutual diffusion coefficient D is the sum of the translational diffusion coefficients of the two species, D_M and D_Q, which can be expressed by the Stokes–Einstein relation:

$$D = D_M + D_Q = \frac{k\,T}{f\,\pi\,\eta}\left(\frac{1}{R_M} + \frac{1}{R_Q}\right) \tag{14.35}$$

where k is Boltzmann's constant, η is the viscosity of the medium, and f is a coefficient that is equal to 6 for "stick" boundary conditions and 4 for "slip" boundary conditions.

If R_M and R_Q are of the same order, the diffusional rate constant is approximately equal to $8RT/3\eta$.

In liquids, k_1 is about 10^9–10^{10} l mol^{-1} s^{-1}. The diffusion coefficient of molecules in the usual solvents at room temperature is generally of the order of 10^{-5} cm^2 s^{-1}.

Quenching of fluorescence (and phosphorescence) by dioxygen deserves attention because it is always present in solution. The effect on quantum yields and lifetimes strongly depends on the nature of the compound and the medium. Oxygen quenching is a collisional process and is therefore diffusion controlled. Consequently, compounds of long lifetime, such as naphthalene and pyrene, are particularly sensitive to the presence of oxygen. Moreover, oxygen quenching is less efficient in media of high viscosity. Oxygen quenching can be reduced by bubbling with nitrogen or argon; the most efficient method is, however, to use the freeze-pump-thaw technique (especially for phosphorescence studies).

Transient Effects In reality, the diffusional rate constant is time dependent and should be written as $k_1(t)$. Several models have been developed to express the time-dependent rate constant. Using Smoluchowski's theory, the fluorescence decay can be written in the following form:

$$i(t) = i(0)\ \exp\left(-at - 2b\sqrt{t}\right) \tag{14.36}$$

where

$$a = \frac{1}{\tau_0} + 4\pi N R_c D[\mathrm{Q}] \tag{14.37}$$

$$b = 4\sqrt{\pi D} N R_c^2 [\mathrm{Q}] \tag{14.38}$$

It should be noted that, in media of low viscosity, the transient term is significant only at short times (less than 100 picoseconds at viscosities similar to that of water) and can be neglected, whereas in viscous media, this term cannot be ignored (the fluorescence decay is then no longer a single exponential).

14.3.2.2 Static Quenching

Sphere of Effective Quenching In rigid matrices or in media so viscous that M* and Q cannot change their positions in space relative to one another during the excited-state lifetime of M*, Perrin has proposed a model in which quenching of a fluorophore is assumed to be complete if a quencher molecule Q is located inside a sphere (called *sphere of effective quenching, active sphere,* or *quenching sphere*) of volume V_q surrounding the fluorophore M. If a quencher is outside the active sphere, it has no effect at all on M. Therefore, the fluorescence intensity of the solution is decreased by the addition of Q, but the fluorescence decay after pulse excitation is unaffected.

According to this model, the ratio I_0/I is given as

$$\frac{I_0}{I} = \exp(V_q N[\mathrm{Q}]) \tag{14.39}$$

At low quencher concentrations, $\exp(V_q N[\mathrm{Q}]) \cong 1 + V_q N[\mathrm{Q}]$: the plot of I_0/I is thus almost linear (as in the case of the Stern–Volmer plot), but at high concentrations, it shows an upward curvature.

A plot of $\ln(I_0/I)$ versus [Q] yields V_q. The values of $V_q N$ are often found to be in the range of 1–3 l mol^{-1}. This corresponds to a quenching sphere radius of about 10 Å, which is slightly larger than the van der Waals contact distance between M and Q.

Perrin's model has been used in particular for the interpretation of nonradiative energy transfer in rigid media.

Formation of a Ground-State Nonfluorescent Complex In some cases, M can form a nonfluorescent 1 : 1 complex according to the equilibrium

$$\mathrm{M} + \mathrm{Q} \rightleftarrows \mathrm{MQ} \tag{14.40}$$

The stability constant of the complex is

$$K_S = \frac{[\mathrm{MQ}]}{[\mathrm{M}][\mathrm{Q}]} \tag{14.41}$$

The excited-state lifetime of the uncomplexed fluorophores M is unaffected. The fluorescence intensity of the solution decreases with the addition of Q, but the fluorescence decay after pulse excitation is unaffected, in contrast to dynamic quenching. Quinones, hydroquinones, purines, and pyrimidines are well-known examples of molecules responsible for static quenching.

Using the mass conservation law (where $[\mathrm{M}]_0$ is the total concentration of M)

$$[\mathrm{M}]_0 = [\mathrm{M}] + [\mathrm{MQ}] \tag{14.42}$$

the fraction of uncomplexed fluorophores is

$$\frac{[\mathrm{M}]}{[\mathrm{M}]_0} = \frac{1}{1 + K_S\,[\mathrm{Q}]} \tag{14.43}$$

In dilute solutions, the fluorescence intensities being proportional to the concentrations, we obtain

$$\frac{I_0}{I} = 1 + K_S\,[\mathrm{Q}] \tag{14.44}$$

A linear relation is thus obtained, as in the case of Stern–Volmer plot (Eq. (14.33)), but there is no change in excited-state lifetime for static quenching, whereas in the case of dynamic quenching, the ratio I_0/I is proportional to the ratio τ_0/τ of the lifetimes.

Simultaneous Dynamic and Static Quenching When both static and dynamic quenching occur simultaneously, a deviation of the plot of I_0/I versus [Q] from linearity is observed.

In the case of static quenching resulting from the formation of a nonfluorescent complex, the ratio I_0/I obtained for dynamic quenching must be multiplied by the fraction of fluorescent molecules (i.e., uncomplexed):

$$\frac{I}{I_0} = \left[\frac{I}{I_0}\right]_{\text{dyn}} \times \frac{[\text{M}]}{[\text{M}]_0} \qquad (14.45)$$

Using Eqs (14.33) and (14.43), the ratio I_0/I is given by

$$\begin{aligned}\frac{I_0}{I} &= (1 + K_{\text{SV}}\,[\text{Q}])\,(1 + K_S\,[\text{Q}])\\ &= 1 + (K_{\text{SV}} + K_S)\,[\text{Q}]\\ &\quad + K_{\text{SV}}\,K_S\,[\text{Q}]^2 \qquad (14.46)\end{aligned}$$

An upward curvature is thus observed. K_{SV} and K_S can be determined by curve fitting using Eq. (14.46), or alternatively from the plot of $(I_0/I - 1)/[\text{Q}]$ versus [Q], which should be linear.

Alternatively, using the model of sphere of effective quenching, we obtain the following relation:

$$\frac{I}{I_0} = (1 + K_{\text{SV}}[\text{Q}])\ \exp(V_q N_{\text{A}}[\text{Q}]) \qquad (14.47)$$

It should be noted that an upward curvature can also be due to transient effects, which may superimpose the effect of static quenching.

For unambiguous assignment of the dynamic and static quenching constants, it is recommended to carry out time-resolved experiments.

14.3.3
Photoinduced Electron Transfer

Photoinduced electron transfer (PET) is involved in many organic photochemical reactions and is often responsible for fluorescence quenching. It plays a major role in photosynthesis and in artificial systems for the conversion of solar energy based on photoinduced charge separation.

It is well known that the oxidative and reductive properties of molecules can be enhanced in the excited state. Oxydative and reductive electron transfer processes occur according to the following reactions:

$$\text{D}^* + \text{A} \rightarrow \text{D}^{\cdot+} + \text{A}^{\cdot-} \qquad (14.48)$$

$$\text{A}^* + \text{D} \rightarrow \text{A}^{\cdot-} + \text{D}^{\cdot+} \qquad (14.49)$$

Examples of donor and acceptor molecules are given in Table 14.3.

In the gas phase, the variations in standard free enthalpy ΔG^0 for the above reactions can be expressed by using the redox potentials E^0 and the excitation energy ΔE_{00}, that is, the difference in

Tab. 14.3 Examples of electron donors and acceptors. (From Valeur (2002).)

Electron donors	**Electron acceptors**
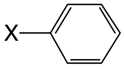	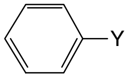
$X = H-, (CH_3)_2N-, CH_3O-, HS-$ Naphthalene Anthracene Phenanthrene Pyrene Perylene	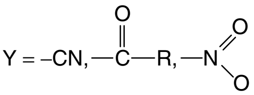 $Y = -CN, -C(=O)-R, -NO_2$ 9,10-dicyanoanthracene

energy between the lowest vibrational levels of the excited state and the ground state:

$$\begin{aligned}\Delta G^0 &= E^0_{\mathrm{D^{\cdot+}/D^*}} - E^0_{\mathrm{A/A^{\cdot-}}}\\ &= E^0_{\mathrm{D^{\cdot+}/D}} - E^0_{\mathrm{A/A^{\cdot-}}} - \Delta E_0(\mathrm{D}) \quad (14.50)\end{aligned}$$

$$\begin{aligned}\Delta G^0 &= E^0_{\mathrm{D^{\cdot+}/D}} - E^0_{\mathrm{A^*/A^{\cdot-}}}\\ &= E^0_{\mathrm{D^{\cdot+}/D}} - E^0_{\mathrm{A/A^{\cdot-}}} - \Delta E_0(\mathrm{A}) \quad (14.51)\end{aligned}$$

These equations are called *Rehm–Weller equations*. If the redox potentials are expressed in volts, then ΔG^0 is given in volts. Conversion into joule per mole requires multiplication by the Faraday constant ($F = 96\,500$ C mol^{-1}).

In solution, two terms should be added for taking into account the solvation effect (enthalpic term ΔH_{solv}) and the Coulombic energy of the formed ion pair:

$$\Delta G^0 = E^0_{\mathrm{D^{\cdot+}/D}} - E^0_{\mathrm{A/A^{\cdot-}}} - \Delta E_0(\mathrm{D}) - \Delta H_{\mathrm{solv}} - \frac{e^2}{4\pi\varepsilon r} \quad (14.52)$$

$$\Delta G^0 = E^0_{\mathrm{D^{\cdot+}/D}} - E^0_{\mathrm{A/A^{\cdot-}}} - \Delta E_0(\mathrm{A}) - \Delta H_{\mathrm{solv}} - \frac{e^2}{4\pi\varepsilon r} \quad (14.53)$$

where e is the electron charge, ε is the dielectric constant of the solvent, and r is the distance between the two ions.

The redox potentials can be determined by electrochemical measurements or by theoretical calculations using the energy levels of the lowest unoccupied molecular orbital (LUMO) and the highest occupied molecular orbital (HOMO).

14.3.4
Formation of Excimers and Exciplexes

Excimers are formed by collision of an excited molecule and an identical unexcited molecule:

$$\mathrm{M^* + M} \rightleftarrows \mathrm{(MM)^*}\ (\text{or } \mathrm{E^*}) \quad (14.54)$$

Excimers are dimers in the excited state (the term excimer results form the contraction of "*exci*ted di*mer*"). The notation (MM)* expresses that the excitation energy is delocalized over the two moieties.

Exciplexes are formed by collision of an excited molecule (electron donor or acceptor) with an unlike unexcited molecule (electron acceptor or donor):

$$\mathrm{D^* + A} \rightleftarrows \mathrm{(DA)^*} \quad (14.55)$$

$$\mathrm{A^* + D} \rightleftarrows \mathrm{(DA)^*} \quad (14.56)$$

Exciplexes are excited-state complexes (the term exciplex comes from "*exci*ted com*plex*").

Excimer and exciplex formations are diffusion-controlled processes. The photophysical effects are thus detected at relatively high concentrations of the species so that a sufficient number of collisions can occur during the excited-state lifetime. Temperature and viscosity are important parameters.

14.3.4.1 **Excimers**

Many aromatic hydrocarbons such as naphthalene or pyrene can form excimers in which the two aromatic rings are facing at a distance of approximately 3–4 Å. The fluorescence band corresponding to excimer is located at wavelengths higher than that of the monomer and does not show vibronic bands because the lowest state is dissociative and can thus be considered as a continuum (Figure 14.4).

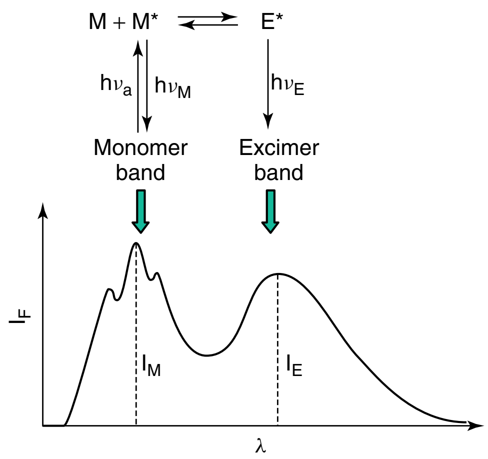

Fig. 14.4 Excimer formation and typical fluorescence spectrum showing the monomer and excimer bands. (Reproduced with permission from Valeur (2002). Copyright Wiley-VCH Verlag GmbH & Co. KGaA.)

Let k_M and k_E be the reciprocals of the excited-state lifetimes of the monomer and the excimer, respectively, and k_1 and k_{-1} are rate constants for the excimer formation and dissociation processes, respectively (Scheme 14.4).

The time evolution of the fluorescence intensity of the monomer M and the excimer E following a δ-pulse excitation can be obtained from the differential equations expressing the evolution of the species according to Scheme 14.4:

$$\frac{d[\mathrm{M}^*]}{dt} = -k_M\,[\mathrm{M}^*] - k_1[\mathrm{M}][\mathrm{M}^*] + k^*_{-1}[\mathrm{E}^*] \qquad (14.57)$$

$$\frac{\mathrm{d}[\mathrm{E}^*]}{\mathrm{d}t} = k_1\,[\mathrm{M}][\mathrm{M}^*] - (k_\mathrm{D} + k_{-1})\,[\mathrm{E}^*] \qquad (14.58)$$

Formation of excimer E* is a diffusion-controlled process. Under the approximation that k_1 is time independent, the δ-pulse responses, under the following initial conditions (at $t = 0$), $[\mathrm{M}^*] = [\mathrm{M}^*]_0$ and $[\mathrm{E}^*]_0 = 0$, are

$$\begin{aligned} i_M(t) &= k_r\,[\mathrm{M}^*] \\ &= \frac{k_r\,[\mathrm{M}^*]_0}{\beta_1 - \beta_2}\big[(X - \beta_2)e^{-\beta_1 t} \\ &\qquad + (\beta_1 - X)e^{-\beta_2 t}\big] \end{aligned} \qquad (14.59)$$

$$\begin{aligned} i_E(t) &= k_r\,k_1\,[\mathrm{E}^*] \\ &= \frac{k_r\,k_1\,[\mathrm{M}][\mathrm{M}^*]_0}{\beta_1 - \beta_2}\big[e^{-\beta_2 t} - e^{-\beta_1 t}\big] \end{aligned} \qquad (14.60)$$

where k_r and $k_r{}'$ are the radiative rate constants of M* and E*, respectively, and

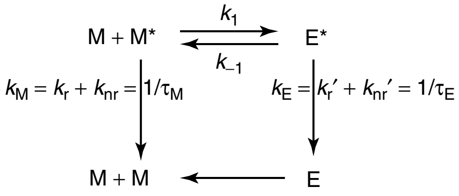

Scheme 14.4 Excimer formation and relevant rate constants.

β_1 and β_2 are given by

$$\beta_{2,1} = \frac{1}{2}\Big\{X + Y \pm \big[(Y - X)^2 + 4\,k_1\,k_{-1}[\mathrm{M}]\big]^{1/2}\Big\} \tag{14.61}$$

where

$$X = k_M + k_1\,[\mathrm{M}] = \frac{1}{\tau_M} + k_1\,[\mathrm{M}] \quad \text{and}$$
$$Y = k_E + k_{-1} = 1/\tau_E + k_{-1} \tag{14.62}$$

The decay of monomer emission is thus a sum of two exponentials, whereas the time evolution of the excimer emission is a difference of two exponentials in which the preexponential factors are of opposite signs. The time constants are the same in the expressions of $i_M(t)$ and $i_E(t)$ (β_1 and β_2 are the eigenvalues of the system). The negative term in $i_E(t)$ represents the rise in intensity corresponding to excimer formation; the fluorescence intensity starts indeed from zero since excimer does not absorb light and can only be formed from the monomer.

If the dissociation of the excimer cannot occur during the lifetime of the excited state ($k_{-1} = 0$), we have

$$i_M(t) = k_r\,[\mathrm{M}^*]_0\,e^{-Xt} \tag{14.63}$$

$$i_E(t) = \frac{k_r\,k_1\,[\mathrm{M}][\mathrm{M}^*]_0}{X - k_E}\left[e^{-k_E t} - e^{-Xt}\right] \tag{14.64}$$

In this case, the fluorescence decay of the monomer is a single exponential with a decay time ($1/X$) equal to the risetime of the excimer fluorescence.

The steady-state fluorescence intensities are obtained by integration of Eqs (14.59) and (14.60). The ratio of the fluorescence intensities of the excimer and monomer bands I_E/I_M (Figure 14.4) is often used for characterizing the efficiency of excimer formation. This ratio is given as

$$\frac{I_E}{I_M} = \frac{k'_r}{k_r}\frac{k_1[\mathrm{M}]}{k_{-1} + 1/\tau_E} \tag{14.65}$$

Under conditions where the dissociation rate of the excimer is slow with respect to de-excitation, this equation is reduced to

$$\frac{I_E}{I_M} = \frac{k'_r}{k_r}\tau_E k_1[\mathrm{M}] \tag{14.66}$$

These equations show that the ratio I_E/I_M is proportional to the rate constant k_1 for excimer formation.

When the two monomers are linked by a short flexible chain, intramolecular excimers can be formed. This process is still diffusion controlled, but is no longer translational; it requires a close approach between the two molecules via internal rotations during the excited-state lifetime. Because intramolecular excimer formation is independent of the total concentration, Eqs (14.59)–(14.66) are still valid provided that $k_1[\mathrm{M}]$ is replaced by k_1.

Assuming that the Stokes–Einstein relation (Eq. (14.35)) is valid, k_1 is proportional to the ratio T/η, η being the viscosity of the medium. A well-known example of application is the estimation of the local fluidity of a medium (see Section 14.4.2.1, Method Based on Intramolecular Excimer Formation).

Another example of application is the determination of micellar aggregation numbers based on self-quenching of pyrene by excimer formation within micelles (Kalyanasundaran, 1987).

14.3.4.2 Exciplexes

The kinetic scheme for exciplex formation is slightly similar to that of excimer formation. Formation of an exciplex with anthracene and *N*,*N*-diethylaniline is a classical example: it results from the transfer of an electron from an amine molecule to an excited anthracene molecule. The quenching of anthracene fluorescence in a nonpolar solvent such as hexane is accompanied by the appearance of a broad structureless emission band of the exciplex at higher wavelengths than anthracene. When the solvent polarity increases, the exciplex band is red shifted. The intensity of this band decreases as a result of the competition between de-excitation and dissociation of the exciplex.

Exciplexes can be considered in some cases as intermediate species in electron transfer from a donor to an acceptor (see Section 14.2.3).

14.3.5
Photoinduced Proton Transfer

Only reactions in aqueous solutions are considered here. Thus, an acid or a base undergoing excited-state deprotonation or protonation, respectively, will always be in close contact with water molecules acting as either a proton acceptor or a proton donor. Therefore, these excited-state reactions will not be diffusion controlled.

Excitation may trigger a photoinduced proton transfer when acids and bases are stronger in the excited state than in the ground state. For instance, the acidic character of a proton-donor group (e.g., OH substituent of an aromatic ring) can be enhanced upon excitation so that the p*K** of this group in the excited state is much lower than the p*K* in the ground state (Table 14.4). In the same manner, the p*K** of a proton-acceptor group (e.g.,

Tab. 14.4 Examples of p*K* and p*K** values. (Adapted from Valeur (2002).)

Reaction		Compound	pK	pK*
Excited-state deprotonation	$ArOH \xrightarrow{-H^+} ArO^-$	Phenol	10.6	3.6
		2-Naphthol	9.3	2.8
		2-Naphthol-6-sulfonate	9.12	1.66
		2-Naphthol-6, 8-disulfonate	9.3	<1
		8-Hydroxypyrene-1,3,6-trisulfonate (pyranine)	7.2	1.3
	$ArNH_2 \xrightarrow{-H^+} ArNH^-$	2-Naphthylamine	7.1	12.2
Excited-state protonation	$ArCO_2^- \xrightarrow{+H^+} Ar\ CO_2H$	Anthracene-9-carboxylate	3.7	6.9
	$ArN \xrightarrow{+H^+} ArNH^+$	acridine	5.5	1.6
		6-methoxyquinoline	5.2	11.8

heterocyclic nitrogen atom) in the excited state can be much higher than in the ground state (pK) (Table 14.4).

14.3.5.1 Equations for Excited-State Deprotonation

The general equations are given only in the most extensively studied case where the excited-state process is proton ejection (p$K^* \ll$ pK); the proton donor is thus an acid, AH*, and the proton acceptor is a water molecule. Assuming that there is no geminate proton recombination, the processes are presented in Scheme 14.5, where τ_0 and τ_0' are the excited-state lifetimes of the acidic (AH*) and basic (A^{-*}) forms, respectively, and k_1 and k_{-1} are the rate constants for deprotonation and reprotonation, respectively. k_1 is a pseudo-first-order rate constant, whereas k_{-1} is a second-order rate constant. The excited-state equilibrium constant is $K^* = k_1/k_{-1}$.

The probability that the back reaction can take place during the excited-state lifetime of A^{-*} is of utmost importance. The rate of this reaction is pH dependent because the process is diffusion controlled ($k_{-1} \approx 5 \times 10^{10}$ l mol^{-1} s^{-1}). If pH ≤ 2, $k_{-1}[\mathrm{H_3O^+}] > \approx 5 \times 10^8$ s^{-1}: the back reaction must then be taken into account. The reciprocal of this value is indeed of the order of the excited-state lifetime of most organic bases.

According to Scheme 14.5, the differential equations expressing the time evolution of the concentration of the species are

$$\frac{d[\mathrm{AH}^*]}{dt} = -\left(k_1 + \frac{1}{\tau_0}\right)[\mathrm{AH}^*] + k_{-1}[\mathrm{A}^{-*}][\mathrm{H_3O^+}] \qquad (14.67)$$

$$\frac{d[\mathrm{A}^{-*}]}{dt} = k_1\,[\mathrm{AH}^*] - \left(k_{-1}\,[\mathrm{H_3O^+}] + \frac{1}{\tau_0'}\right)[\mathrm{A}^{-*}] \qquad (14.68)$$

When AH is selectively excited by a δ-pulse of light, the responses of the fluorescence intensities, under the following initial conditions (at $t = 0$), $[\mathrm{AH}^*] = [\mathrm{AH}^*]_0$ and $[\mathrm{A}^{-*}]_0 = 0$, are

$$i_{\mathrm{AH}^*}(t) = k_r\,[\mathrm{AH}^*] = \frac{k_r\,[\mathrm{AH}^*]_0}{\beta_1 - \beta_2} \times \left[(X - \beta_2)e^{-\beta_1 t} + (\beta_1 - X)e^{-\beta_2 t}\right] \qquad (14.69)$$

$$i_{\mathrm{A}^{-*}} = k_r'\,[\mathrm{A}^{-*}] = \frac{k_r'\,k_1\,[\mathrm{AH}^*]_0}{\beta_1 - \beta_2} \times \left[e^{-\beta_2 t} - e^{-\beta_1 t}\right] \qquad (14.70)$$

where k_r and k_r' are the radiative rate constants of AH* and A^{-*}, respectively, and β_1 and β_2 are given by

$$\beta_{2,1} = \frac{1}{2}\left\{X + Y \pm \left[(Y - X)^2 + 4k_1k_{-1}[\mathrm{H_3O^+}]\right]^{1/2}\right\} \qquad (14.71)$$

where $X = k_1 + 1/\tau_0$ and $Y = k_{-1}[\mathrm{H_3O^+}] + 1/\tau_0'$.

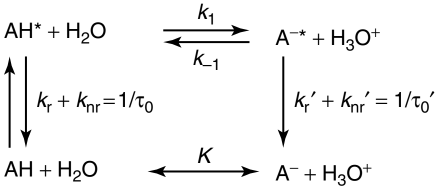

Scheme 14.5 Excited-state deprotonation of an acid AH.

Under continuous illumination, the steady-state intensities can be easily calculated by integration of Eqs (14.69) and (14.70):

$$I_{\mathrm{AH}^*} = C\Phi_0 \frac{1 + k_{-1}\tau_0'[\mathrm{H_3O^+}]}{1 + k_1\tau_0 + k_{-1}\tau_0'[\mathrm{H_3O^+}]} \tag{14.72}$$

$$I_{\mathrm{A}^{-*}} = C\Phi_0{}' \frac{k_1\tau_0}{1 + k_1\tau_0 + k_{-1}\tau_0'[\mathrm{H_3O^+}]} \tag{14.73}$$

where Φ_0 and $\Phi_0{}'$ are the fluorescence quantum yields of AH and A^-, respectively, in the absence of excited-state reaction ($\Phi_0 = k_r\ \tau_0$; $\Phi_0{}' = k_r{}'\ \tau_0{}'$). C is a multiplication factor that is introduced to take into account the experimental conditions (total concentration, choice of excitation and emission wavelengths, bandpasses for absorption and emission intensity of the incident light, and sensitivity of the instrument).

Under conditions where the back reaction cannot take place during the excited-state lifetime ($k_{-1} = 0$), β_1 and β_2 become equal to X and Y, respectively, and Eqs (14.69) and (14.70) are reduced to

$$i_{\mathrm{AH}^*}(t) = [\mathrm{AH}^*]_0\ k_r\ e^{-Xt} \tag{14.74}$$

$$i_{\mathrm{A}^{-*}}(t) = \frac{k_r'\ k_1[\mathrm{AH}^*]_0}{X - Y}\left[e^{-Yt} - e^{-Xt}\right] \tag{14.75}$$

where $X = k_1 + 1/\tau_0$ and $Y = 1/\tau_0{}'$. Therefore, $i_{\mathrm{AH}^*}(t)$ becomes a single exponential and $i_{\mathrm{A}^{-*}}(t)$ is still a difference of two exponentials with a rise time $(1/X)$ (equal to the decay time of AH^*) and a decay time $(1/Y)$ (equal to the lifetime of the basic form).

The steady-state intensities reduce to

$$I_{\mathrm{AH}^*} = C\Phi_0 \frac{1}{1 + k_1\tau_0} \tag{14.76}$$

$$I_{\mathrm{A}^{-*}} = C\Phi_0' \frac{k_1\tau_0}{1 + k_1\tau_0} \tag{14.77}$$

14.3.5.2 Determination of the Excited-State p*K**

Estimation by Means of the Förster Cycle pK^* can be theoretically predicted by means of the Förster cycle associated with spectroscopic measurements. Details on the theory can be found in Valeur (2002) p. 103. Only the final equation is given here at 298 K:

$$\mathrm{p}K^* - \mathrm{p}K = 2.1 \times 10^{-3}(\overline{\nu}_{\mathrm{A}^-} - \overline{\nu}_{\mathrm{AH}}) \tag{14.78}$$

where $\overline{\nu}_{\mathrm{AH}}$ and $\overline{\nu}_{\mathrm{A}^-}$ are the wavenumbers (in cm^{-1}) corresponding to the 0–0 transitions of AH and A^-. The determination of $\overline{\nu}_{\mathrm{AH}}$ and $\overline{\nu}_{\mathrm{A}^-}$ is the major cause of inaccuracy in the estimation of pK^*. They are usually estimated by means of the average between the wavenumbers corresponding to the maxima of absorption and emission :

$$\nu_0 = \frac{\overline{\nu}_{\mathrm{abs}}^{\max} + \overline{\nu}_{\mathrm{em}}^{\max}}{2} \tag{14.79}$$

However, the best approximation is to use the intersection point of the mutually normalized absorption and emission spectra (as it was the practice in Förster's laboratory). In the case of large Stokes shift, it may be difficult to determine an intersection point and it is then preferable to take the average of the wavenumbers corresponding to half-heights of the absorption and emission bands, which is a better approximation than the average of the wavenumbers corresponding to the maxima.

The simplicity of the Förster cycle method explains why it has been extensively used. One of the important features of this cycle is the possibility to use it even in the case where the equilibrium is not established within the excited-state lifetime.

Steady-State Measurements The value of the excited-state pK^* can be determined by fluorometric titration but only when the equilibrium is established in the excited state. Rewriting Eqs (14.72) and (14.73) as $I_{\mathrm{AH}^*} = C\Phi$ and $I_{\mathrm{A}^{-*}} = C\Phi'$, it is easy to obtain the following relation:

$$\frac{\Phi/\Phi_0}{\Phi'/\Phi'_0} = \frac{1}{k_1\tau_0} + \frac{k_{-1}\tau'_0}{k_1\tau_0}[\mathrm{H_3O^+}] \tag{14.80}$$

Consequently, if the excited-state lifetimes τ_0 and τ_0' are known, the plot of $(\Phi/\Phi_0)/(\Phi'/\Phi_0')$ versus $[\mathrm{H_30^+}]$ yields the rate constants k_1 and k_{-1}. The ratio k_1/k_{-1} yields K^*.

Time-Resolved Experiments The most reliable method for the determination of k_1 and k_{-1} is based on time-resolved experiments. They provide the values of the decay times from which the rate constants k_1 and k_{-1} are determined by means of Eqs (14.69) and (14.70).

14.3.5.3 pH Dependence of Absorption and Emission Spectra

The absorption and fluorescence spectra are pH dependent when the acido-basic properties are not the same in the ground and excited states. The discussion is restricted again to the most important case of photoinduced proton ejection ($\mathrm{p}K^* \ll \mathrm{p}K$). Let us recall that for pH values less than approximately 2, the back reaction occurs. Moreover, distinction should be made according to the value of $\mathrm{p}K^*$.

1. If $\mathrm{p}K^*$ is greater than approximately 2, a plateau is observed for the relative fluorescence quantum yield of the acidic form and the basic form between $\mathrm{p}K^*$ and $\mathrm{p}K$ (Figure 14.5) because of the absence of diffusional recombination (see Eqs (14.75) and (14.76)). A typical example is 2-naphthol ($\mathrm{p}K = 9.3$, $\mathrm{p}K^* = 2.8$).
2. If $\mathrm{p}K^*$ is smaller than approximately 2, the acid is very strong in the excited-state and, in general, k_1 is much larger than the reciprocal of the excited-state lifetime so that the fluorescence of the acidic form is not observed at $\mathrm{pH} > \mathrm{p}K^* + 2$ but only for lower pH values (Figure 14.5). This is the case of pyranine ($\mathrm{p}K = 7.7$, $\mathrm{p}K^* = 1.3$).

14.3.6
Excitation Energy Transfer

Distinction between radiative and nonradiative transfer is of major importance (Valeur, 2002 Chapter 4, p. 110, and Chapter. 9); Andrews and Demidov, 1999. In radiative transfer a photon emitted by a molecule D is absorbed by a molecule A (or D). This process is observed when the average distance between D and A (or D) is larger than the wavelength. It does not require any interaction between the partners, but it depends on the spectral overlap and on the concentration. In contrast, nonradiative transfer occurs without emission of photons at distances lower than the wavelength and results from short- or long-range interactions between molecules. For instance, nonradiative transfer by dipole–dipole interaction is possible at distances up to 8–10 nm. Consequently, such a transfer provides a tool for determining distances of a few nanometers between chromophores. It should be

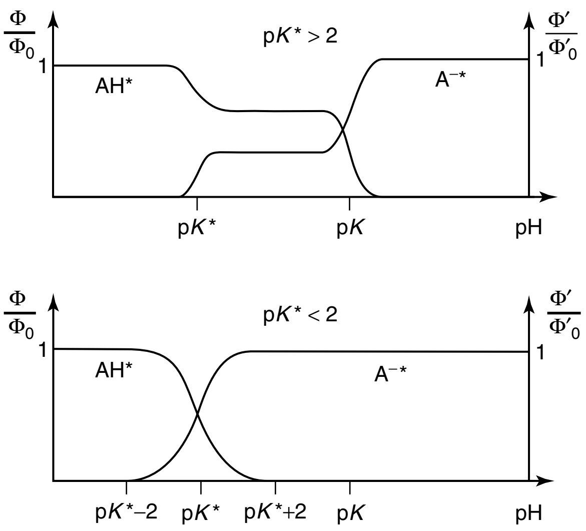

Fig. 14.5 Variations in relative quantum yields of the basic and acidic forms as a function of pH. (Reproduced with permission from Valeur (2002). Copyright Wiley-VCH Verlag GmbH & Co. KGaA.)

noted that the transfer of excitation energy plays a major role in photosynthetic systems.

Nonradiative transfer of excitation energy requires some interaction between a donor molecule and an acceptor molecule, and it can occur if the emission spectrum of the donor overlaps the absorption spectrum of the acceptor so that several vibronic transitions in the donor have practically the same energy as the corresponding transitions in the acceptor. Such transitions are coupled, that is, are in *resonance* and, hence, the term *resonance energy transfer* (*RET*). *Fluorescence resonance energy transfer* (FRET) is often employed, but this expression is not correct because it is not the fluorescence that is transferred but the electronic energy of the donor. Therefore, it is recommended to use either EET (*excitation energy transfer* or *electronic energy transfer*) or RET.

Different interaction mechanisms are involved in RET. The interactions may be Coulombic and/or due to intermolecular orbital overlap. The Coulombic interactions consist of long-range dipole–dipole interactions (Förster's mechanism) and short-range multipolar interactions. The interactions due to intermolecular orbital overlap, which include electron exchange (Dexter's mechanism) and charge resonance interactions, are only short range.

14.4
Applications of Molecular Fluorescence

14.4.1
Historical Introduction

Fluorescence has long been used as an analytical tool for the determination of the concentrations of neutral or ionic species (Valeur, 2007). In his famous book *History of Luminescence*, Harvey (1957) reports that Victor Pierre, who was a professor in Prague and later in Vienna, published a paper in 1862 in which he described the investigation of solutions of single fluorescent compounds and mixtures. He noted that bands of fluorescent spectra were characteristic of a particular substance. He also noted the effect of solvent and acidity or alkalinity. G. G. Stokes – who invented the term fluorescence – had certainly the same idea in mind. In fact, one of the topics of his famous Burnett lectures was " Fluorescence – Its use as a means of discrimination." However, no specific example was given in the text.

The term *fluorescence analysis* was employed for the first time in the paper by Göppelsröder (1868): the drastic enhancement of fluorescence intensity accompanying the complexation of morin (a hydroxyflavone derivative) with aluminum provides a straightforward way to detect this metal.

The use of fluorescent tracers in hydrology is also an old application of fluorescence. Uranin (the disodium salt of fluorescein) was used for the first time in 1877 as a tracer for monitoring the flow of Danube. On all maps, it is shown that the Danube springs in the Black Forest and, after many hundreds of kilometers, flows into the Black Sea. However, because of big sink (swallow holes) in the bed of Danube, it was demonstrated, thanks to the fluorescent tracer, that only a small part of the water from Danube spring arrives at the Black Sea. Most of it flows into the North Sea. Fluorescence tracing is now currently used in hydrology, especially to simulate pollutions.

Numerous applications of fluorescence were developed at the beginning of the twentieth century and reported in several books Radley and Grant (1933); Dake and De Ment (1941); Pringsheim and Vogel (1943). For instance, in the second part of the book of Radley and Grant (1933), the list of applications of fluorescence analysis is impressive: agriculture, bacteriology, botany, ceramics, slags and cements, drugs, foods and food products, fuels, inorganic chemistry, leather and tanning, legal work, medical and biological sciences, minerals and gems, museum work, organic chemistry, paints and varnishes, paper and cellulose and its derivatives, the rubber industry, textiles, and waters and sewage. Pringsheim and Vogel (1943) also gave various examples of application including fluorescent carpets and ceilings in theaters or fluorescent ballets.

The applications of fluorescence that are currently used are now presented.

14.4.2
Estimation of Local Physical Parameters

Physical parameters, such as viscosity, temperature, and pressure, can be locally estimated by means of fluorescent probes, that is, fluorescent compounds whose emissive properties are affected by one of these parameters.

14.4.2.1 Viscosity

Basic Concepts It is important to first recall that viscosity is a macroscopic parameter that loses its physical meaning at a molecular scale (Valeur, 2002 Chapter 8; Valeur, 1993). The term *microviscosity* must be used but with caution. The term *fluidity* is appropriate to characterize, in a very general way, the effects of viscous drag and cohesion of the probed microenvironment (polymers, micelles, gels, lipid bilayers of vesicles or biological membranes, etc.).

In all fluorescence techniques permitting evaluation of the fluidity of a microenvironment by means of a fluorescent probe, the underlying physical quantity is a diffusion coefficient expressing the viscous drag of the surrounding molecules (Figure 14.6).

Let us first recall that the diffusion coefficients for translational and rotational diffusions are, respectively,

$$D_t = \frac{kT}{6\pi\eta r} \tag{14.81}$$

$$D_r = \frac{kT}{8\pi\eta r^3} = \frac{kT}{6\eta V} \tag{14.82}$$

where k is the Boltzman constant, T is the absolute temperature, η is the viscosity, r is the hydrodynamic radius of the sphere, and V is its hydrodynamic volume.

The use of these equations to estimate the viscosity η of an unknown environment from the experimental value of the diffusion constant is questionable. In fact, these equations are valid only for a rigid sphere being large compared to the molecular dimensions, moving in a homogeneous newtonian fluid and obeying the Stokes hydrodynamic law. Many other relations have been proposed, but it should be pointed out that *there is no satisfactory relationship between diffusion and bulk viscosity for probes*. The main reason is that the size of probes is comparable to that of the surrounding molecules forming the microenvironment to be probed. Another difficulty arises in the case of organized assemblies such as micellar systems, biological membranes, and so on, because the microenvironment is not isotropic.

In other words, *viscosity is a macroscopic parameter and any attempt to get absolute values of the viscosity of a medium from measurements using a fluorescent probe is hopeless*.

It is always observed that the diffusion of molecular probes is more rapid than predicted by the theory. Therefore, the "slip" boundary condition is often introduced, and sometimes a mixture of "stick" and "slip" boundary conditions is assumed. Relation (14.82) can then be

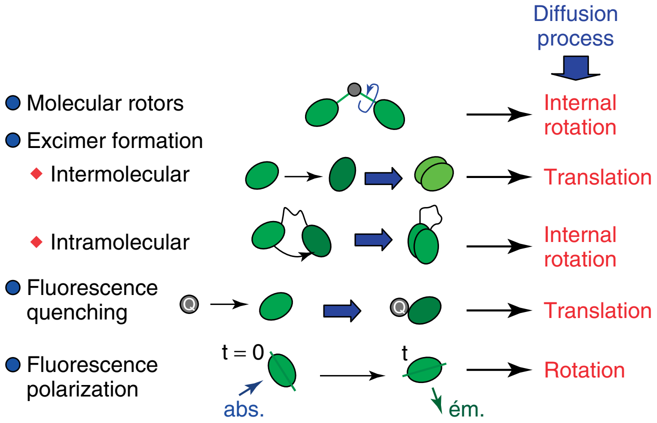
Diffusion process
Molecular rotors
Excimer formation
Intermolecular
Intramolecular
Fluorescence quenching
Fluorescence polarization
Internal rotation
Translation
Internal rotation
Translation
Rotation
Q
Q
t = 0
t
abs.
ém.

Fig. 14.6 Fluorescence techniques for the evaluation of the fluidity of a microenvironment.

rewritten as

$$D_r = \frac{kT}{6\eta Vsg} \tag{14.83}$$

where s is the coupling factor ($s = 1$ for "stick" and $s < 1$ for "slip" boundary conditions) and g is the shape factor. In the case of charged molecules, an additional friction force should be introduced as a result of the induced polarization of the surrounding solvent molecules.

A faster diffusion than predicted can be explained by the fact that a solute molecule moves according to two diffusional processes: a viscous process with displacement of solvent molecules (*Stokes diffusion*) and a process associated with migration into holes of the solvent (*free volume diffusion*). The importance of free volume should be emphasized at a molecular level. An empirical relation between viscosity and free volume was proposed by Doolittle:

$$\eta = \eta_0 \exp\left(\frac{V_0}{V_\mathrm{f}}\right) \tag{14.84}$$

where η_0 is a constant and V_0 and V_f are the Van der Waals volume and the free volume of the solvent, respectively.

The choice of the method for probing fluidity depends on the system to be investigated. The use of intermolecular quenching and intermolecular excimer formation is not recommended for probing fluidity of microheterogeneous media because of possible perturbation of the translational diffusion process. The methods of molecular rotors and intramolecular excimer formation are convenient and rapid, but the time-resolved fluorescence polarization technique provides much more detailed information, including an order parameter if the medium is anisotropic.

Use of Molecular Rotors A *molecular rotor* – as a fluorescent probe – is a molecule that undergoes internal rotation(s) resulting in viscosity-dependent changes of its emissive properties. They belong to the family of diphenylmethane dyes (e.g., auramine O), triphenylmethanes dyes (e.g., crystal violet), and twisted intramolecular charge transfer (TICT) compounds (e.g., *p*-dimethylaminobenzonitrile (DMABN)). Examples of molecular rotors of practical interest are given in Figure 14.7.

In solvents of medium and high viscosity, an empirical relation based on free volume considerations has been proposed to account for the viscosity dependence of the fluorescence quantum yield:

$$\Phi_F \cong b\left(\frac{\eta}{T}\right)^x \tag{14.85}$$

where x is a constant for a particular probe.

In polymers above the glass transition temperature, the fluorescence quantum yields were, in fact, found to be a power function of the bulk viscosity with values of the exponent x lower than 1 (e.g., for *p*-*N*,*N*-dimethylaminobenzylidenemalononitrile, $x = 0.69$ in glycerol and 0.43 in dimethylphtalate).

Molecular rotors are a powerful tool to study changes in free volume of polymers as a function of molecular weight, stereoregularity, crosslinking, polymer chain relaxation, and flexibility. Monitoring of polymerization reactions is possible in this way (Loufty, 1986).

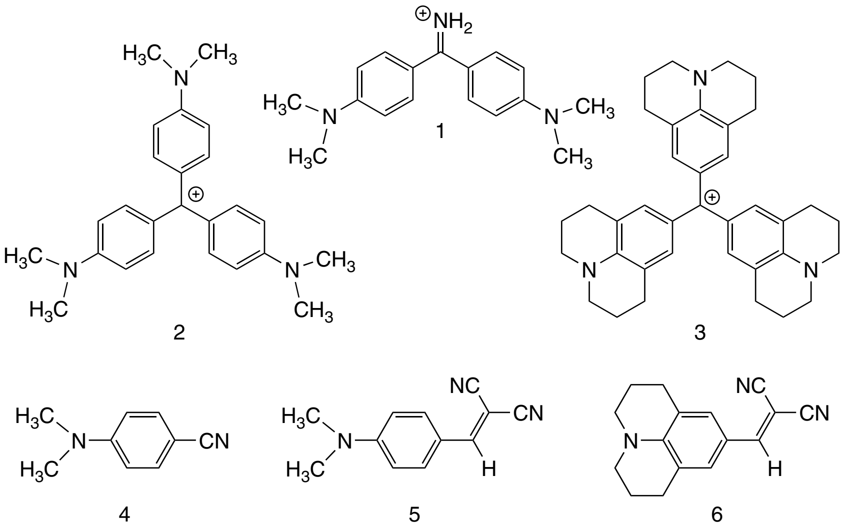

Fig. 14.7 Examples of molecular rotors. 1, auramine O; 2, crystal violet; 4, *p*-*N*,*N*-dimethylaminobenzonitrile (DMABN); 5, *p*-N,N-dimethylaminobenzylidenemalononitrile; 6, julolidinebenzylidenemalononitrile.

Method Based on Intramolecular Excimer Formation Bifluorophoric molecules consisting of two identical fluorophores linked by a short flexible chain may form an excimer. Examples of such bifluorophores currently used in investigations of fluidity are given in Figure 14.8.

The efficiency of intramolecular excimer formation does not depend on the concentration of fluorophores so that Eq. (14.66) should be rewritten by replacing $k_1[\mathrm{M}]$ by k_1:

$$\frac{I_E}{I_M} = \frac{k'_r}{k_r}\tau_E\, k_1 \tag{14.86}$$

It should be noted that when effects of temperature on fluidity are investigated, difficulties may arise from the possible temperature dependence of the excimer lifetime. It is then recommended to perform time-resolved fluorescence experiments.

As in the case of molecular rotors, free volume effects are important, but the free volume fraction measured by intramolecular excimers is smaller than for molecular rotors. The volume swept out during the conformational change required for excimer formation is in fact larger, and consequently these probes do not respond in frozen media or polymers below the glass transition temperature.

The following empirical relation is currently used for the rate constant for excimer formation:

$$k_1 = \alpha\, \eta^{-x} \tag{14.87}$$

It is important to note that in addition to viscosity, the response of a probe depends on several parameters: chemical nature of the solvent, length of the chain, temperature, and oxygen quenching (Viriot *et al.*, 1983).

Intramolecular excimers have been used for probing bulk polymers, micelles, vesicles, and biological membranes (Viriot *et al.*, 1983; Bokobza and Monnerie, 1986; Bokobza, 1990;

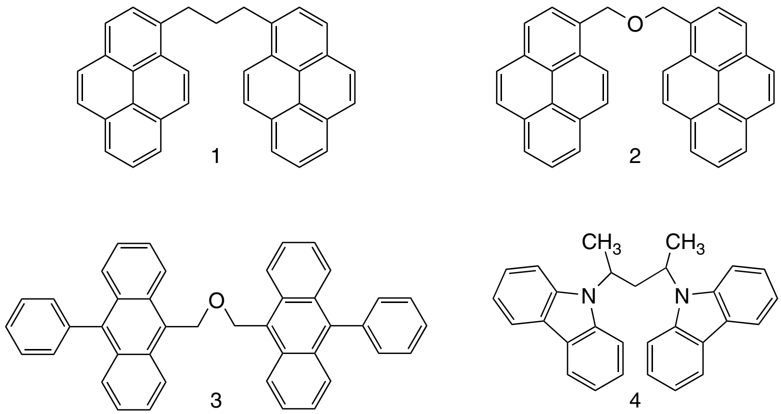

Fig. 14.8 Examples of bifluorophores forming excimers for the study of fluidity. 1, α,ω-di-(1-pyrenyl)propane; 2, α,ω-di-(1-pyrenyl)methylether; 3, 10,10′-diphenyl-*bis*-9-anthrylmethyloxide (DIPHANT); 4, meso-2,4-di(*N*-carbazolyl)pentane.

Georgescauld *et al.*, 1980; Vauhkonen *et al.*, 1990; Zachariasse and Kozankiewicz, and Kühnle, 1983).

Fluorescence Polarization Method The preferred orientation of fluorescent probes resulting from photoselection at time zero is gradually affected as a function of time by the rotational Brownian motions if the medium is fluid enough to permit rotation of the probes during the excited-state lifetime. This causes a depolarization of fluorescence, from which information can be obtained on the molecular motions that depend on the size and the shape of molecules and on the fluidity of their microenvironment.

A molecule undergoes isotropic rotation if it is spherical, but this is also the case of a rodlike molecule whose directions of its absorption and emission transition moments coincide with the long molecular axis (e.g., diphenylhexatriene): the rotations can be considered as isotropic because any rotation about this long axis has no effect on the emission anisotropy.

For isotropic motions in an isotropic medium, the values of the instantaneous and steady-state emission anisotropies are given by

$$r(t) = r_0 \exp(-6D_r t) \tag{14.88}$$

$$\frac{1}{\bar{r}} = \frac{1}{r_0}(1 + 6D_r \tau) \tag{14.89}$$

which allows one to determine the diffusion constant D_r. In principle, a value of the viscosity η could be calculated from the Stokes–Einstein relation (Eq. (14.82) or (14.83)) provided that the hydrodynamic volume V of the probe is known. It should be again emphasized that any "microviscosity" value calculated in this way would be questionable and thus useless. Nevertheless, the changes in D_r upon an external perturbation (e.g., temperature, pressure, additive, etc.) reflect well the changes in fluidity of a medium.

In ordered systems such as lipid bilayers and liquid crystals, the rotational motions of a probe are hindered and the emission anisotropy does not decay to zero but to a steady value r_∞. Equations (14.88) and (14.89) are thus no longer valid. For isotropic rotations (rodlike probe), the emission anisotropy can often be written in the following approximated form:

$$r(t) = (r_0 - r_\infty)\ \exp\left(-\frac{t}{\tau_c}\right) + r_\infty \tag{14.90}$$

where τ_c is the effective relaxation time of $r(t)$, that is, the time with which the initially photoselected distribution of orientations approaches the stationary distribution.

Time-resolved emission anisotropy experiments yield τ_c, r_0, and r_∞. The data can be processed according to the *wobble-in-cone model* (Kinosita, Kawato and Ikegami, 1977; Lipari and Szabo, 1980) in which the rotations of a rodlike probe (with the direction of its absorption and emission transition moments coinciding with the long molecular axis) are restricted within a cone. The rotational motions are described by the rotational diffusion coefficient D_w around an axis perpendicular to the long molecular axis and an order parameter (half angle of the cone θ_c) reflecting the degree of orientational constraint due to the surrounding paraffinic chains. θ_c can be determined from the ratio r_∞/r_0, and D_W from τ_c and θ_c.

The rotations of the probe are thus describe by (i) the wobbling diffusion constant D_w reflecting the fluidity of the medium, and more precisely the chain mobility and (ii) the order parameter θ_c, half angle of the cone, expressing the degree of orientational constraint.

More general theories have been developed (Van der Meer, Kooyman and Levine, 1982; Van der Meer *et al.*, 1984; Zannoni, Arcioni and Cavatorta, 1983; Fisz, 1985; Szabo, 1984; Pottel *et al.*, 1986).

14.4.2.2 **Temperature**

Fluorescent molecular thermometers are based on the temperature dependence of the photophysical properties of fluorescent compounds (fluorescence spectrum, fluorescence quantum yield, and excited-state lifetime) (Chandrasekharan and Kelly, 2004; Uchiyama, de Silva and Iwau, 2006). Temperature is indeed a key parameter of several photophysical processes: intersystem crossing (Section 14.2.1.3), delayed fluorescence (Section 14.2.1.3), and excimer or exciplex formation (Section 14.3.4).

The advantages of fluorescent molecular thermometers over contact temperature sensors are the following: absence of sensitivity to electromagnetic radiation, fast response, and spatial resolution from the nanoscale to the macroscale (temperature-sensitive paints).

The compounds whose fluorescence quantum yield is more or less strongly affected by temperature changes are either organic (e.g., perylene, acridine yellow, perylenedicarboximide, rhodamine B, pyronine B, pyronine Y) or organometallic (e.g., ruthenium or europium complexes : Ru(bpy), $\mathrm{Ru(phen)_3}$, Europium thenoyltrifluoroacetonate (EuTTA)).

For instance, the temperature dependence of the fluorescence quantum yield of a solution of rhodamine B has been exploited for spatial and temporal mapping of temperature within the microchannel networks of microfluidic devices. Changes in fluorescence intensity were used to determine intrachannel temperature variations with a precision of ±3 °C, a spatial resolution of 1 μm, and a temporal resolution of 33 milliseconds (Ross, Gaitan and Locascio, 2001). With the same aim, fluorescence lifetime imaging (FLIM) has been used to quantitatively image temperature distributions in three spatial dimensions within microchannels environments (Benninger *et al.*, 2006). The advantage of lifetime measurements is that, unlike fluorescence intensity, excited-state lifetime is independent of fluorophore concentration, fluctuations of the light source, and photobleaching effects.

Ratiometric measurements, that is the ratio of the fluorescence intensities at two observation wavelengths, offer the same advantage. This requires a change in the shape of the fluorescence spectrum as a function of temperature. This is, for instance, the case of exciplex formation with perylene and *N*-allyl-*N*-methylaniline (NA), leading to the appearance of a broad fluorescence band centered at 550 nm (Chandrashekaran and Kelly, 2001):

$$\underset{\text{(blue fluorescence)}}{\text{Perylene}} + \text{NA} \rightleftarrows \underset{\text{(green fluorescence)}}{\text{Perylene-NA}}$$

A film consisting of perylene and NA encapsulated and anchored, respectively, in a soft elastic polystyrene matrix is a temperature sensor whose working range is 25–85 °C. Figure 14.9 shows the effect of temperature changes on the emission spectra.

An interesting application of fluorescent molecular thermometers concerns real-time polymer processing monitoring. Measurements by thermocouples and infrared imaging are not accurate in this case. Benzoxazolyl stilbene and

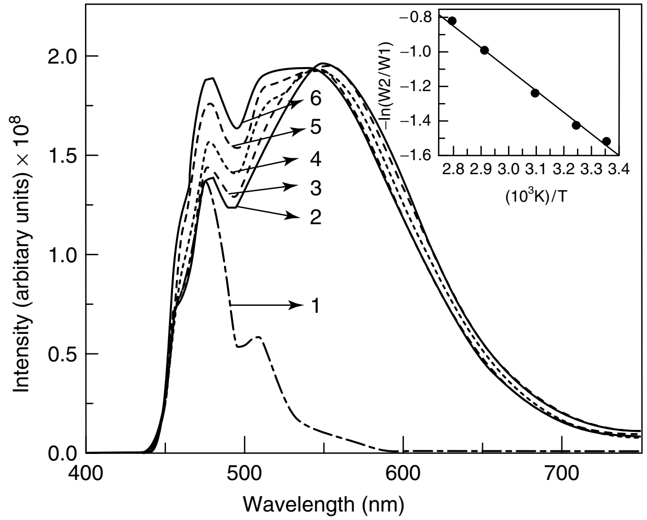

Fig. 14.9 Effect of temperature on the emission spectra of a polystyrene film composed of 0.017 wt % of perylene and 18.96 wt % of *N*-allyl-*N*-methylaniline : (2) 298 K; (3) 304 K; (4) 323 K; (5) 343 K; (6) 358 K; (1) film of perylene alone at 298 K. Excitation wavelength: 386 nm. (Reproduced with permission from Chandrashekaran and Kelly (2001). Copyright 2001 American Chemical Society.)

perylene were chosen because of their compatibility with polymer matrices and their high thermal stability. Morever, the relative intensities of the vibronic bands are temperature dependent, which allow ratiometric measurements. Plots of the intensity ratio versus temperature are linear with a slope of 0.09–0.2% per degree. Temperatures as high as 300 °C can be measured (Bur, Wangel and Roth, 2002).

Optical thermometers based on the thermally activated delayed fluorescence have been proposed. Acridine yellow dissolved in a rigid saccharide glass exhibits a temperature-dependent ratio of fluorescence to phosphorescence intensity, but the effect is weak (Fister, Rank and Harris, 1995). A much stronger effect has been observed with fullerene C_{70}, which shows very intense delayed fluorescence (see Section 14.2.1.3). For C_{70} dispersed in a polymer matrix (poly(*ter*-butyl methacrylate)), the working range extends from −80 to 140 °C (Baleizao *et al.*, 2007). Figure 14.10 shows the increase in intensity of red fluorescence. The working range is even wider when monitoring the lifetime of the delayed fluorescence.

Tab. 14.5 Temperature ranges of temperature-sensitive paints.

Compound	**Temperature range (°C)**
Perylene	0–100
Perylenedicarboximide	50–100
Rhodamine B	0–80
Pyronine B	50–100
Pyronine Y	0–100
Ruthenium complex Ru(bpy)	0–90
Europium complex EuTTA[a]	−20 to 80

[a] Europium theonyltrifluoroacetonate.

Surface temperature mapping can be obtained by using paints containing temperature-sensitive fluorescent compounds (Liu et and Sullivan, 2005). Table 14.5 gives examples of such compounds together with the temperature ranges. In aerodynamics, temperature-sensitive paints are employed in conjunction with pressure-sensitive paints (PSPs) (see the following section) (Liu *et al.*, 1997). Both paints are also used in advanced turbomachinery applications

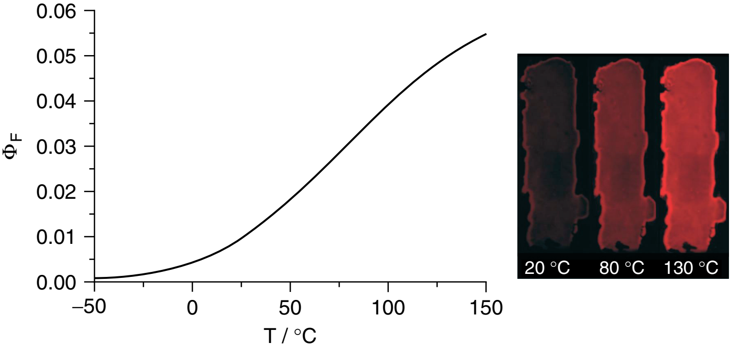

Fig. 14.10 Fluorescence quantum yield versus temperature for C_{70} dispersed in a polymer matrix (poly(*ter*-butyl methacrylate)) and color picture at three temperatures. (Reproduced with permission from Baleizao *et al.* (2007). Copyright Wiley-VCH Verlag GmbH & Co. KGaA.)

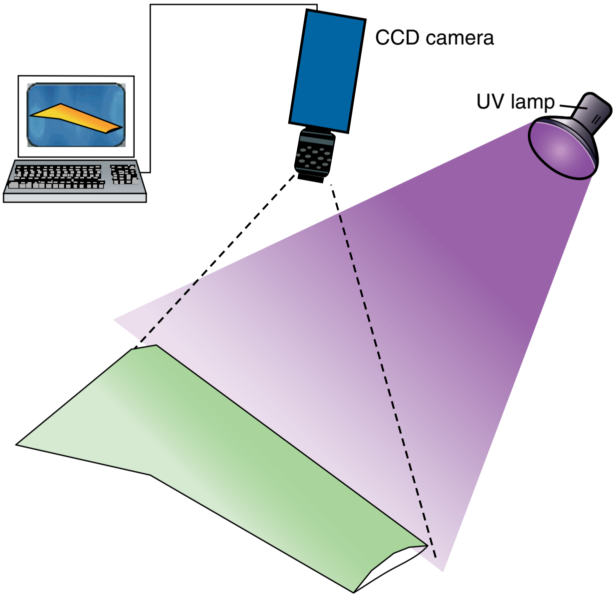

Fig. 14.11 Experimental set-up for temperature and pressure surface mapping using temperature- and pressure-sensitive paints.

(Navarra *et al.*, 2001). Figure 14.11 shows the experimental set-up.

At a microscopic scale, *fluorescence microthermography* is a powerful technique for generating high-resolution thermal maps of integrated circuits and for hot-spot localization in electrical failure analysis (Kolodner and Tyson, 1982; Herzum *et al.*, 1998). The europium complex EuTTA is well suited for these measurements. The lateral resolution is 0.5 μm and the thermal resolution is in the millikelvin range. This technique is thus superior to other routine thermography methods (liquid crystal thermography, infrared thermography, or laser thermoreflectance).

14.4.2.3 **Pressure**

An interesting application of fluorescence quenching is the use of fluorescent paints sensitive to air pressure, with the aim at studying the aerodynamism of aircraft or car models in wind tunnels. Such PSPs sprayed on a model provide a surface pressure mapping, that is, the instantaneous two-dimensional pressure distribution on the surface. They are more convenient and cheaper than the numerous pressure sensors that would be required.

PSPs are oxygen sensitive: they contain fluorophores whose fluorescence is quenched by oxygen (Figure 14.12). Therefore, the larger the air pressure, the larger the partial pressure of oxygen, and the higher the efficiency of quenching of the fluorophores that are embedded in a transparent polymer. Fluorophores should have a long excited-state lifetime: pyrene derivatives and ruthenium complexes are thus suitable for this application.

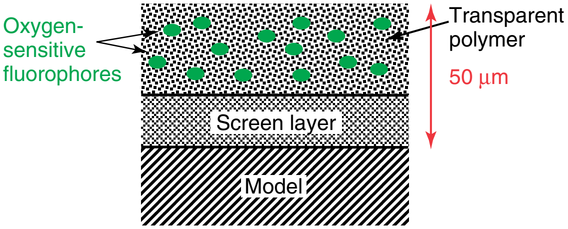

Fig. 14.12 Pressure-sensitive paint on a model.

The Stern–Volmer equation that holds for collisional quenching of fluorescence (Eq. (14.33)) must be rewritten in the following alternative form:

$$\frac{I_0}{I} = A + B\frac{P}{P_0} = \frac{\tau_0}{\tau} \tag{14.91}$$

where I, τ, and P are the fluorescence intensity, the excited-state lifetime, and the pressure, respectively. The subscripts 0 refer to the atmospheric pressure. In some cases, excited-states lifetimes are measured instead of fluorescence intensities.

Since the 1980s the PSP technique has been mainly used in aeronautics for transonic and supersonic flow studies, and it is now currently used in industrial wind tunnels (Liu *et al.*, 1997).

Surface pressure mapping is more difficult to obtain in the case of cars because the variations in pressure are smaller at low speed. Figure 14.13 shows an investigation carried out in ONERA (Engler *et al.*, 2002). The PSP is made of pyrene with gadolinium oxysulfide as reference dye (insensitive to oxygen) that permits correction for the nonhomogeneous distribution of illumination. The pressure map accurately reveals the variations in pressure at the roof/rear window junction.

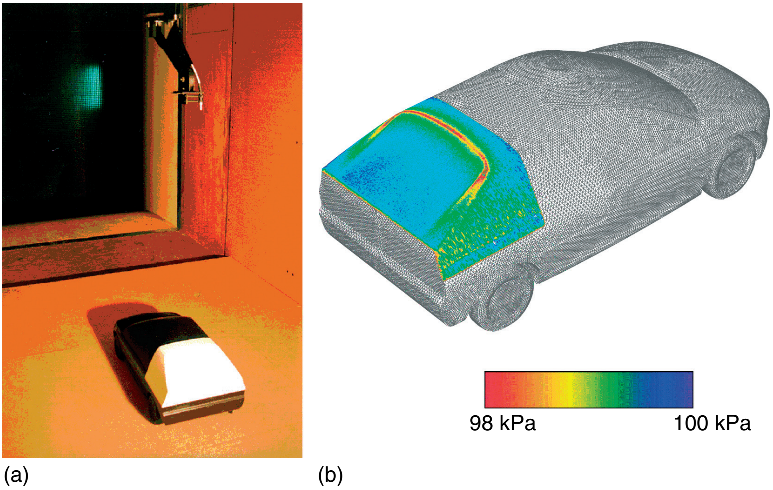

Fig. 14.13 Model of a Peugeot 206 car in a wind tunnel (a) and pressure map (b). (Courtesy of M.C. Merienne, ONERA, France.)

14.4.3
Chemical Sensing

Fluorescence is extensively used for chemical sensing in many fields: analytical chemistry, biology, biotechnology, clinical diagnostics, toxicology, environment, forensics, and chemical oceanography (Lakowicz, 1994; de Silva and Tecilla, 2005; Narayanaswamy and Wolfbeis, 2004). The success of fluorescence can be explained by the distinct advantages that this technique offers in terms of sensitivity, selectivity, response time, local observation under microscope, and remote detection by means of optical fibers.

In principle, a *fluorescent sensor* is the complete optical sensing device containing the light source, the analyte-responsive (supra)molecular moiety properly immobilized (e.g., in plastified polymers, sol–gel matrices, etc.), the optical system (involving an optical fiber or not), and the light detector (photomultiplier or photodiode) connected to appropriate electronics for displaying the signal. When considering only the fluorescent analyte-responsive (supra)molecular moiety, the term *fluorescent molecular sensor* should be used.

14.4.3.1 Fluorescent Molecular Sensors

The design of a fluorescent molecular sensor that is selective of a given analyte is actually a work of molecular engineering that involves many disciplinary fields: photophysics, photochemistry, analytical chemistry, physical chemistry, coordination chemistry, and supramolecular chemistry (Valeur, 2002 Chapter 10; Desvergne and Czarnik, 1997).

The fluorophores that are implied in fluorescent molecular sensors are numerous: naphthalene, anthracene, pyrene, aminonaphthalimide, diaminonaphthylsulfonyl, coumarins, fluorescein, eosin, rhodamines, benzidine, alizarin, seminaphthofluorescein, oligo- and polyphenylenes, porphyrins, ruthenium complexes, and so on.

Much progress has been made in sensitivity and selectivity toward a given analyte. The main targets are the following:

- cations: H_3O^+ (pH), alkali, alkaline-earth, transition, and post-transition;
- anions: fluoride, chloride, carboxylates, phosphates, (adenosine triphosphate) ATP, nitrates, and so on; and
- molecules: hydrocarbons, aminoacids, sugars, urea, ammonia, amines, alcohols, O_2, CO_2, H_2O_2, and so on.

Three classes of fluorescent molecular sensors are to be distinguished (Figure 14.14):

- Class 1: fluorophores that undergo quenching upon collision with an analyte (e.g., O_2, Cl^-).
- Class 2: fluorophores that can reversibly bind an analyte. If the analyte is a proton, the term *fluorescent pH indicator* is often employed. If the analyte is an ion, the term *fluorescent chelating agent* is appropriate. Fluorescence can be either quenched or enhanced upon binding. In the case of enhancement, the compound is said to be *fluorogenic* (e.g., 8-hydroxyquinoline (oxine)).
- Class 3: fluorophores linked, via a spacer or not, to a receptor that ensures appropriate affinity and selectivity toward a given analyte. The changes in fluorescence characteristics of the fluorophore upon interaction with the bound analyte result from the perturbation by the latter of photoinduced processes such as electron transfer, charge transfer, energy transfer, and

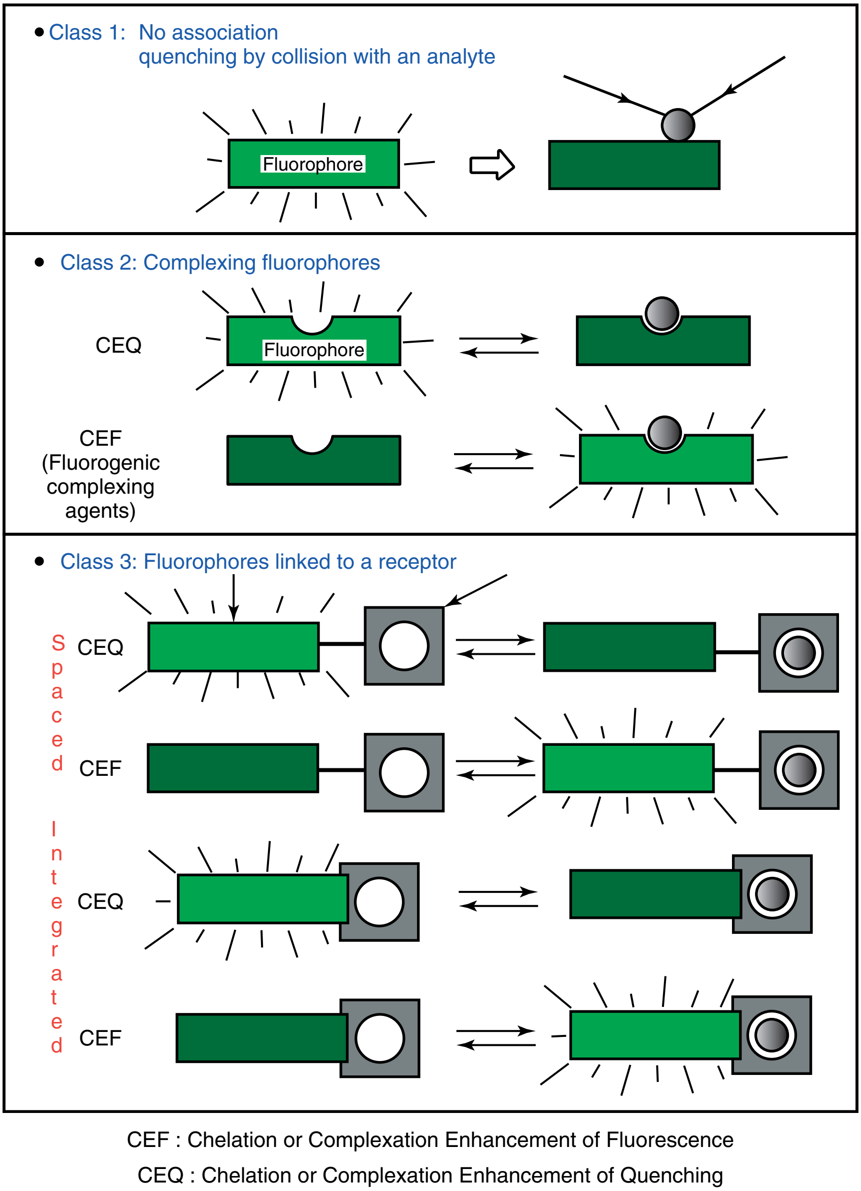
Class 1: No association
quenching by collision with an analyte
Fluorophore
Class 2: Complexing fluorophores
CEQ
Fluorophore
CEF
(Fluorogenic complexing agents)
Class 3: Fluorophores linked to a receptor
Spaced
CEQ
CEF
Integrated
CEQ
CEF
CEF : Chelation or Complexation Enhancement of Fluorescence
CEQ : Chelation or Complexation Enhancement of Quenching

Fig. 14.14 Classes of fluorescent molecular sensors.

excimer or exciplex formation or disappearance. The design of such sensors is relevant to the fields of supramolecular chemistry and photophysics. In the case of ion recognition, the receptor is called *ionophore*, and the whole molecular sensor is called *fluoroionophore*.

Fluorescent pH Indicators Fluorescent pH indicators offer a much better sensitivity than the classical dyes, such as phenolphthalein, thymol blue, and so on, based on color change. They are thus widely used in analytical chemistry, bioanalytical chemistry, cellular biology (for measuring intracellular pH), and medicine (for monitoring pH and pCO_2 in blood; pCO_2 is determined via the bicarbonate couple). In microscopy, fluorescent pH indicators provide spatial information on pH. Moreover, remote sensing of pH is possible by means of fiber-optic chemical sensors.

A fluorescent indicator AH can undergo deprotonation according to the following equilibrium:

$$\mathrm{AH} + \mathrm{H_2O} \rightleftarrows \mathrm{A^-} + \mathrm{H_3O^+} \qquad (14.92)$$

The true value of the acidity constant (that depends only on temperature) must be written with activities:

$$K_\mathrm{A} = \frac{a_{\mathrm{A^-}} a_{\mathrm{H_3O^+}}}{a_\mathrm{AH}} \qquad (14.93)$$

Using the definition of pH

$$\mathrm{pH} = -\log a_{\mathrm{H_3O^+}} \qquad (14.94)$$

Equation (14.93) can be rewritten in the following form:

$$\mathrm{pH} = \mathrm{p}K_\mathrm{A} + \log \frac{a_{\mathrm{A^-}}}{a_\mathrm{AH}} \qquad (14.95)$$

$$\mathrm{pH} = \mathrm{p}K_\mathrm{A} + \log \frac{f_{\mathrm{A^-}}}{f_\mathrm{AH}} + \log \frac{[\mathrm{A^-}]}{[\mathrm{AH}]} \qquad (14.96)$$

where $f_{\mathrm{A^-}}$ and f_AB are the activity coefficients relative to molarities.

In very dilute aqueous solutions, the activity coefficients are close to 1 (reference state: solute at infinite dilution), and Eq. (14.96) reduces to the well-known Henderson–Hasselbach equation:

$$\mathrm{pH} = \mathrm{p}K_\mathrm{A} + \log \frac{[\mathrm{A^-}]}{[\mathrm{AH}]} \qquad (14.97)$$

Otherwise, $\mathrm{p}K_A$ should be replaced by an apparent pK:

$$\mathrm{p}K_\mathrm{app} = \mathrm{p}K_\mathrm{A} + \log \frac{f_{\mathrm{A^-}}}{f_\mathrm{AH}} \qquad (14.98)$$

Hence,

$$\mathrm{pH} = \mathrm{p}K_\mathrm{app} + \log \frac{[\mathrm{A^-}]}{[\mathrm{AH}]} \qquad (14.99)$$

$\mathrm{p}K_\mathrm{app}$ depends on ionic strength, specific interactions between the indicator and the medium, and structural changes of the medium. For instance, $\mathrm{p}K_\mathrm{app}$ differs significantly from $\mathrm{p}K_A$ at the interface of micelles or lipid bilayers.

For fluorometric titrations, the following relation can be used (Valeur, 2002 p. 338):

$$\mathrm{pH} = \mathrm{p}K_\mathrm{app} + \log \frac{I - I_\mathrm{A}}{I_\mathrm{B} - I} \qquad (14.100)$$

where I is the fluorescence intensity at a given wavelength, and I_A and I_B are the fluorescence intensities measured at the same wavelength when the indicator is only in the acidic form or only in the basic form, respectively.

It is recommended, whenever possible, to carry out ratiometric measurements, that is, to determine the ratio of the fluorescence intensities measured at two excitation wavelengths or two emission wavelengths. The ratiometric measurements are indeed preferable because the ratio of the fluorescence intensities at two wavelengths is independent of the total concentration of the dye, photobleaching, fluctuations of the source intensity, and sensitivity of the instrument. Examples of fluorescent

Tab. 14.6 Examples of fluorescent pH indicators allowing ratiometric measurements. (From Valeur (2002).)

Fluorophore	pK_A	Type of measurement
4-MU	7.8	excitation ratio 365/335 nm
Pyranine	7.2	excitation ratio 450/400 nm
fluorescein	2.2, 4.4, 6.4	excitation ratio 490/435 nm
BCECF	7.0	excitation ratio 505/439 nm
SNAFL dyes	7.0–7.8	excitation ratio 490/540 nm or emission ratio 540/630 nm
SNARF dyes	7.0–7.8	emission ratio 580/630 nm
CNF	7.5	excitation ratio 600/510 nm or emission ratio 550/670 nm

4-MU, 4-methylumbelliferone; BCECF, 2′,7′-bis(carboxyethyl)-5(or 6)-carboxyfluorescein; SNAFL, semi-naphthofluoresceins; SNARF, semi-naphthorhodafluors; CNF, carboxynaphthofluorescein.

pH indicators allowing ratiometric measurements are given in Table 14.6.

For ratiometric measurements, the following equation may be used to fit the calibration curve (Valeur, 2002 p. 338):

$$\mathrm{pH} = \mathrm{p}K_{\mathrm{app}} + \log \frac{R - R_{\mathrm{A}}}{R_{\mathrm{B}} - R} + \log \frac{I_{\mathrm{A}}(\lambda_2)}{I_{\mathrm{B}}(\lambda_2)} \tag{14.101}$$

where R is the ratio $I(\lambda_1)/I(\lambda_2)$ of the fluorescence intensities at two excitation wavelengths (or two emission wavelengths) λ_1 and λ_2. R_A and R_B are the values of R when only the acidic form or the basic form is present, respectively. $I_A(\lambda_2)/I_B(\lambda_2)$ is the ratio of the fluorescence intensity of the acidic form alone to the intensity of the basic form alone at the wavelength λ_2 chosen for the denominator of R. It should be emphasized that R, R_A, and R_B are very sensitive to the wavelengths chosen and also to instrumental settings.

According to the preceding relations, the usual working range is about 2 pH units around pK_A. However, the working range is broader when the indicator possesses more that one pK_A (e.g., fluorescein).

Fluorescent Molecular Sensors of Cations

Detecting cations is of great interest in many fields: chemistry, biology, clinical biochemistry, and environment (de Silva *et al.*, 1997; Valeur, 1994; Valeur and Leray, 2000). Sodium, potassium, magnesium, and calcium are involved in biological processes such as transmission of nerve impulse, muscle contraction, regulation of cell activity, and so on. Zinc is an essential component of many enzymes (e.g., in carbonic anhydrase and zinc finger proteins); it plays a major role in enzyme regulation, gene expression, neurotransmission, and so on. In medicine, monitoring of metal ions (e.g., Na^+, K^+, Mg^{2+}, Ca^{2+}, Li^+) in blood and urines is of major importance for diagnosis. In medicine, it is necessary to control the lithium level in the serum for patients under treatment for maniac depression and potassium in the case of high blood pressure. In environment, early detection of toxic metal ions such as mercury, lead, and cadmium is of interest. In chemical oceanography, it has been demonstrated that some nutrients required for the survival of microorganisms in sea water contain zinc, iron, and manganese as enzyme cofactors.

In the design of a fluorescent molecular sensor of cations, called *fluoroionophore*, much attention is to be paid to the characteristics of the ionophore moiety and to the expected changes in fluorescence characteristics of the fluorophore moiety upon binding.

The ionophore can be a chelator, an open-chain structure (podand), a macrocycle (coronand, e.g., crown ether), a macrobicycle (cryptand), a calixarene derivative, and so on. The characteristics of the ionophore, that is, the ligand topology and the number and nature of the complexing heteroatoms or groups, should match the characteristics of the cation, that is, ionic diameter, charge density, coordination number, and intrinsic nature (e.g., hardness of metal cations, nature and structure of organic cations, etc.) according to the general principles of supramolecular chemistry.

The ionophore may be linked to the fluorophore via a spacer, but in many cases some atoms or groups participating in the complexation belong to the fluorophore. Therefore, the selectivity of binding often results from the whole structure involving both signaling and recognition moieties.

The changes in emissive properties of the fluorophore upon cation binding are due to the cation control of photoinduced electron transfer (PET), photoinduced charge transfer (PCT), excimer formation or disappearance, and excitation energy transfer. The discussion is restricted to the most popular PET and PCT sensors.

In a PET sensor (Bissell *et al.*, 1992), excitation of the fluorophore induces the promotion of an electron of the HOMO to the LUMO, which enables PET from the HOMO of the donor (cation-free receptor) to that of the fluorophore, causing fluorescence quenching of the latter (see Section 14.3.3). Upon cation binding, the redox potential of the donor is raised so that the relevant HOMO becomes lower in energy than that of the fluorophore. Therefore, PET is not possible any more and fluorescence quenching is suppressed. Examples of PET sensors are given in Figure 14.15.

In a PCT sensor (Valeur and Leray, 2001), the fluorophore contains an electron-donating group (often an amino group) conjugated to an electron-withdrawing group and thus undergoes intramolecular charge transfer from the donor to the acceptor upon excitation by light. The consequent change in dipole moment results in a Stokes shift that depends on the microenvironment of the fluorophore. Therefore,

- when a cation is in close interaction with the donor moiety (such as an amino group), it reduces the efficiency of intramolecular charge transfer and the fluorescence spectrum is blue-shifted.
- when a cation is in close interaction with the acceptor moiety, it enhances the efficiency of intramolecular charge transfer and the fluorescence spectrum is red shifted.

In addition to these spectral shifts, changes in quantum yields and lifetimes are often observed. All the photophysical effects are obviously dependent on the charge and the size of the cation, and selectivity of these effects is expected.

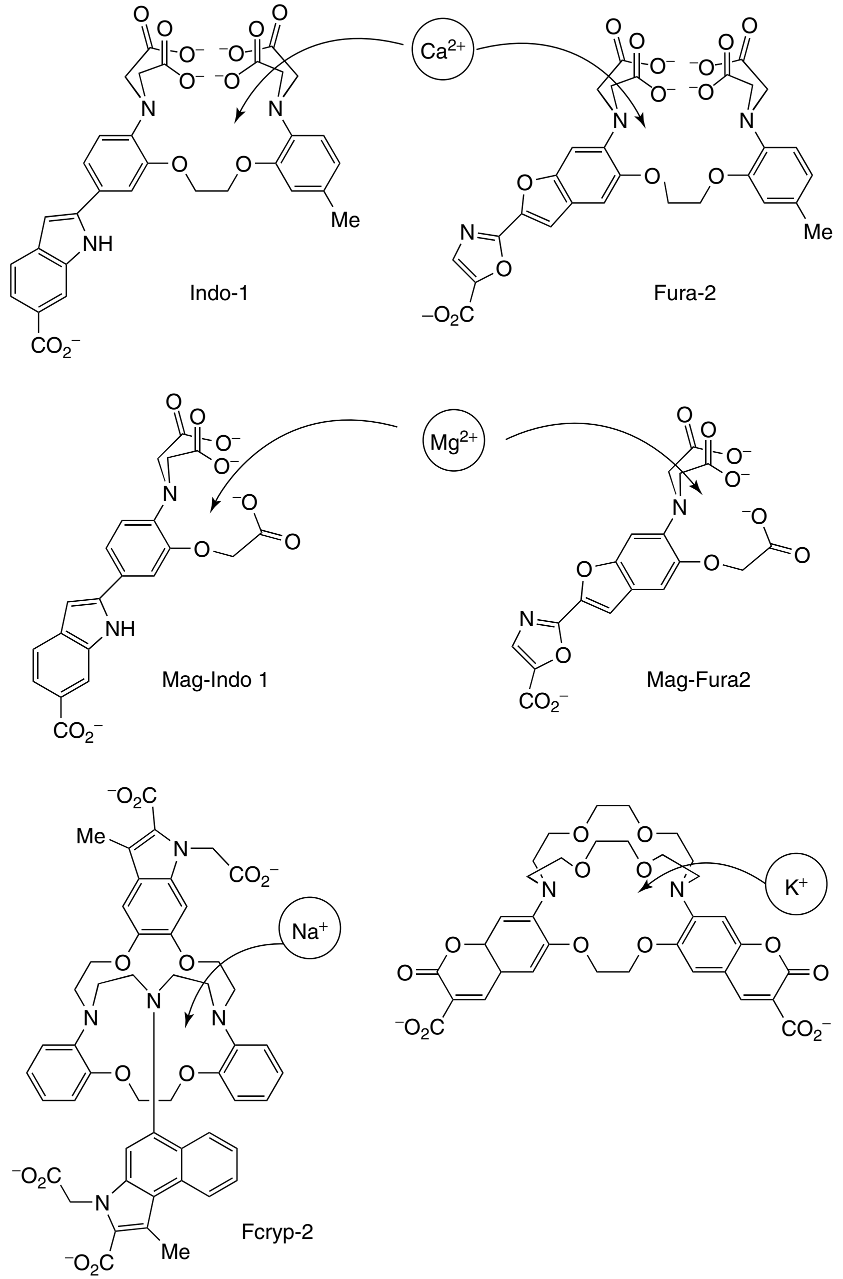
Ca2+
Indo-1
Fura-2
Mg2+
Mag-Indo 1
Mag-Fura2
Na+
K+
Fcryp-2

Fig. 14.15 Examples of PET sensors.

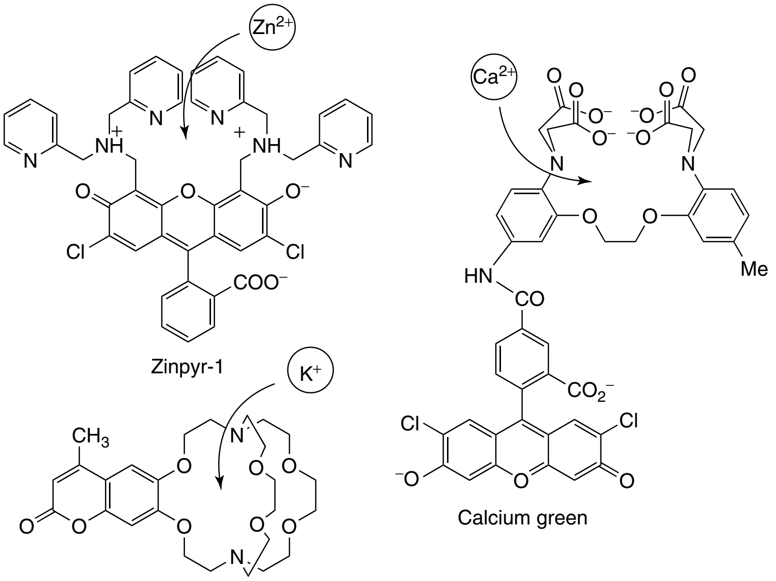
Zinpyr-1

Calcium green

Fig. 14.16 Examples of PCT sensors.

Examples of PCT sensors are given in Figure 14.16.

Fluorescent Molecular Sensors of Anions Anions play key roles in chemical and biological processes. Many anions act as nucleophiles, bases, redox agents, and phase transfer catalysts. Most of enzymes bind anions as either substrates or cofactors. The chloride ion is of special interest because it is crucial in several phases of human biology and in disease regulation. Moreover, it is of great interest to detect anionic pollutants such as nitrates and phosphates in ground water.

Anion molecular sensors are either based on collisional quenching (in general, they exhibit a poor selectivity) or based on recognition by an anion receptor linked to a fluorophore (fluoroionophore), which is described later.

Many fluorescent molecular sensors for halide ions (except F$^-$) are based on collisional quenching of a dye. In particular, the determination of chloride anions in living cells is achieved according to this principle. Examples of halide ion sensors are given in Figure 14.17. The drawback of these molecular sensors is their lack of selectivity. Moreover, the absence of spectral change precludes ratiometric measurements. However, dual-wavelength Cl$^-$ sensors have been designed.

Anion sensors containing an anion receptor have a much better selectivity. Sensors of phosphate and pyrophosphate groups have attracted much attention because of their biological relevance. Figure 14.18 shows some examples.

14.4.3.2 **Microfabricated Analysis Systems**

Miniaturization of sensing devices has been the object of considerable effort. Microfabricated systems indeed offer many advantages that are of particular

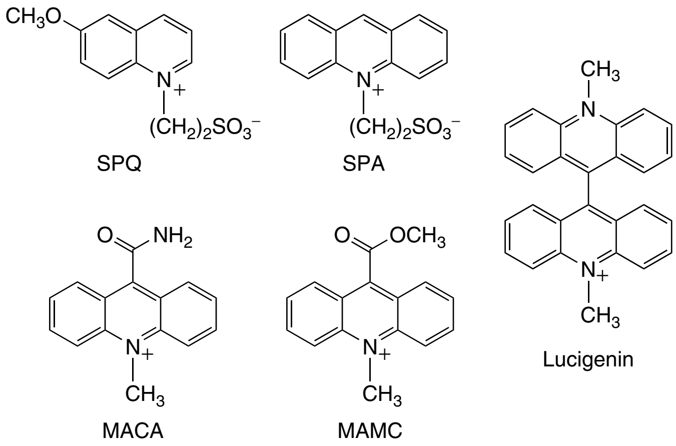
CH3O
N+
(CH2)2SO3−
SPQ
N+
(CH2)2SO3−
SPA
CH3
N
N+
CH3
Lucigenin
O
NH2
N+
CH3
MACA
O
OCH3
N+
CH3
MAMC

Fig. 14.17 Examples of halide sensors.

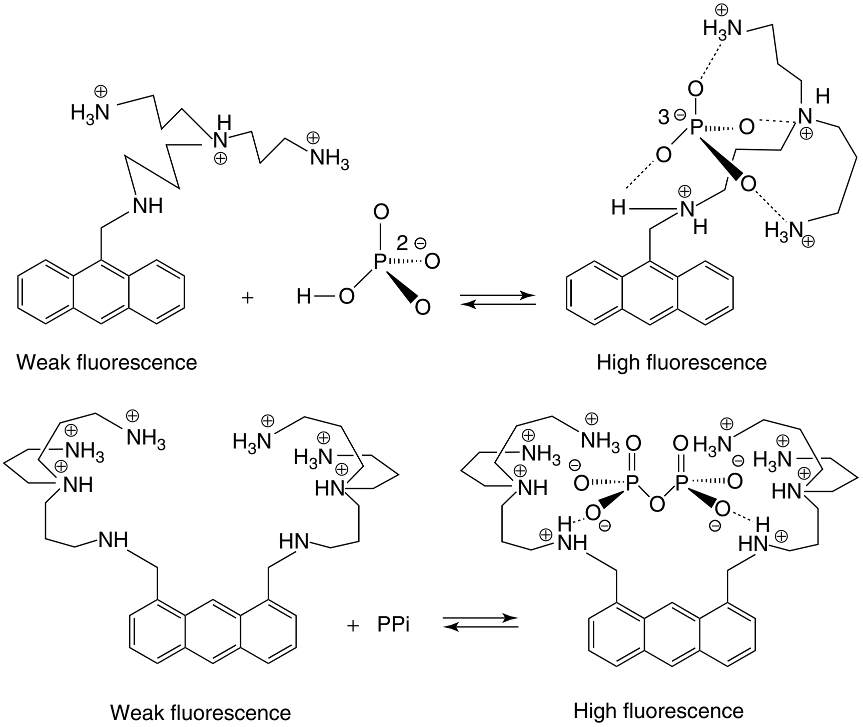
H3N
H
N
NH3
NH
O
2 ⊖
P
O
H—O
O
+
Weak fluorescence
H3N
O
3 ⊖
P
O
H
N
O
O
H—N
H
H3N
High fluorescence
NH3
NH3
NH
NH
H3N
H3N
HN
HN
+ PPi
Weak fluorescence
NH3
NH3
O
O
H3N
H3N
NH
O
P
O
P
O
HN
H
O
O
H
NH
HN
High fluorescence

Fig. 14.18 Examples of sensors of phosphate groups.

interest in the case of biological samples: high speed, online monitoring, and low consumption of samples and reagents. The development of lab-on-chips with, in particular, the DNA gene chips is impressive. Microfabricated analysis systems, called *micro-total analysis systems* (μ-*TASs*), are based on microfluidic technology using microchipchannels, microreactors, valves, and pumps (Reyes *et al.*, 2002; Auroux *et al.*, 2002). Among the detection methods employed in these systems, fluorescence deserves attention. The applications are mainly relevant to life sciences: clinical diagnostics, immunoassays, DNA separation and analysis, sequencing, and so on.

A nice example is the very sophisticated microfluidic device developped by Richard Mathies and coworkers (University of California at Berkeley) for the detection of aminoacids in future missions on Mars (Skelley *et al.*, 2005) (Figure 14.19). It is based on capillary electrophoresis and the use of fluorescamine as a fluorescence marker for aminoacids. The first step of the analysis is sublimation of amines and aminoacids onto an aluminum disk spin-coated with fluorescamine. Successful tests were carried out in Atacama desert in Chile, which can be considered as the best Mars analog site because it is one of the most arid regions in the world. Sensitivity ranges from micromolar to 0.1 nM, corresponding to part-per-trillion sensitivity.

14.4.4
Tracers in Biology

Molecular fluorescence is widely used in biology for cellular imaging using (confocal) fluorescence microscopy. pH imaging and ion imaging (calcium, magnesium, zinc, chloride, etc.) are possible by using appropriate fluorescent pH indicators or fluorescent molecular sensors for ions (see Section 14.4.2). With the aim at tracing cellular components during biological processes, fluorescent species are used to visualize these components.

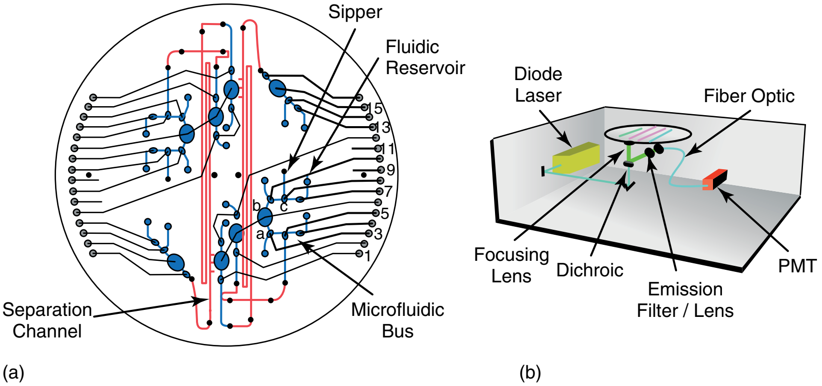

Fig. 14.19 Microdevice for amino acid biomarker detection and analysis on Mars. (a) Top view of the microdevice showing registration of the capillary electrophoresis channel (red), pneumatic manifold (black), and fluidic bus wafers (blue). (b) Schematic of the instrument showing confocal excitation and detection optics. (Reproduced with permission from Skelley *et al.* (2005), Copyright 2005 National Academy of Sciences, USA.)

In contrast to fluorescent probes, they should be insensitive to their microenvironment as much as possible. The four classes of fluorescent tracers in biology are organic fluorophores, fluorescent antibodies, fluorescent proteins, and semiconductor nanocrystals. The first two classes are well known and are not described here, whereas the use of fluorescent proteins and semiconductor nanocrystals is more recent and deserves particular attention.

14.4.4.1 Fluorescent Proteins

Fluorescent proteins become more and more popular as tracers in biology (Zimmer, 2002; Lippincott-Schwartz and Patterson, 2003). The first of them is green fluorescent protein (GFP) from *Aequora Vitoria,* a bioluminescent jelly fish. The green fluorescence is due to a partially cyclic tripeptide (serine, tyrosine, and glycine) in an α-helix. A fusion between a cloned gene and GFP can be created using standard techniques, which leads to a chimera that can be expressed in a cell. Thus, GFP acts as a fluorescent label of a protein of interest whose localization or activity is not affected. It is thus possible to visualize dynamic cellular events and to monitor protein localization.

Spectral variants of GFP emitting in the blue-cyan and the yellowish green have been generated, but the wavelength of the emission maximum is not greater than approximately 530 nm. However, corals (Anthozoa) can provide fluorescent proteins emitting in the red, such as DsRed.

The simultaneous use of several colors is thus possible. A beautiful example of multicolor imaging is shown in Figure 14.20.

At a larger scale, fluorescent proteins also offer an outstanding tool for imaging muscle fibers in vivo, so that one can see the effect of new therapies for muscle diseases.

At an even larger scale, a whole animal can be rendered fluorescent if all cells express the gene of GFP. This has been applied to rabbits, mice, rats, pigs, and silkworms. The interest is

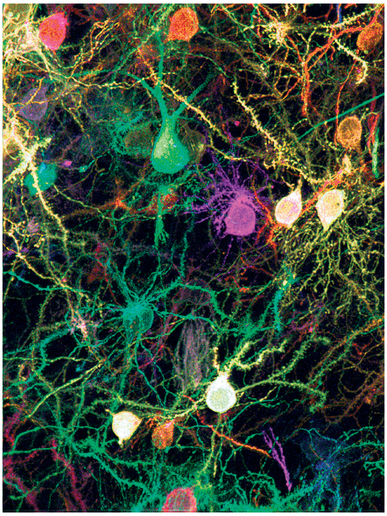

Fig. 14.20 Multicolor imaging of mossy fibers in the cerebellum of a transgenic mouse. This image taken by confocal microscopy shows mossy fibers in one folium of a young adult transgenic mouse. Spectral variants of GFP are expressed in a combinatorial way following genetic recombination mediated by cre-recombinase. The various colors are produced by the particular subset of fluorescent proteins expressed in each axon and their individual concentrations. In this animal, at least 12 distinct spectral fingerprints in different cells can be discerned. (Courtesy of J. Lichtman, Harvard University.)

that GFP makes it possible to follow the fate of cells in embryos, in normal organs, and in tumors, without requiring destruction of the cells. A transgenic fluorescent rabbit generated in the French *Institut National de la Recherche Agronomique* (INRA) appears green under UV illumination, even the eyes show green fluorescence. A bioartist, Eduardo Kac, from the Chicago University wished to breed these rabbits to be used as exotic pets, but this proposition was rejected by the French researchers.

In 2003, a company located in Texas put in the market a genetically modified aquarium fish having the capacity to be red fluorescent under UV illumination. This announcement was followed by protests coming from both activists and the scientific community. Indeed, this event raises several unanswered questions. Are these fish a threat for environment? Are they suffering from their transgene and from UV illumination? Is it ethical to use genetic engineering to generate this particular kind of pets? (Knight, 2003)

14.4.4.2 **Semiconductor Nanocrystals**

The fluorescence of semiconductor nanocrystals is not strictly relevant to molecular fluorescence, but their growing interest deserves attention here. The emissive properties of these nanocrystals, also called *quantum dots*, are due to quantum confinement. The emission wavelength ranges from the UV to the infrared depending on the nature of the semiconductor and the size of the nanocrystal. For quantum dots made of cadmium selenide with a protective coating of zinc sulfide, the emission wavelength ranges from the blue (diameter of 2 nm) to the red (diameter of 7 nm).

Quantum dots offer distinct advantages for tracing owing to their specific characteristics:

- broad absorption spectrum allowing a common excitation wavelength for nanocrystals having different emission wavelengths;
- narrow emission spectrum (gaussian) : 12–15 nm;
- fluorescence quantum yield greater than 50%;
- excellent photostability;
- coatable for water solubility and biocompatibility.

It should be emphasized that the photostability of quantum dots is much better than that of organic dyes, and they are much brighter in tissue. It is thus possible to monitor the trajectories of nanocrystals for 20 minutes, instead of 5 seconds with an antibody coupled to an organic fluorophore such as Cy3 (Dahan *et al.*, 2003).

Quantum dots have also been successfully used for in vivo cancer targeting and imaging (Gao *et al.*, 2004). Applications to encoding are also of great interest. For this purpose, nanocrystals are embedded in various proportions into polymer microspheres with millions of possible combinations. These microspheres can then be attached to DNA sequences, peptides, and so on, and are thus useful for high throughput screening (genomics, proteomics, and drug screening).

14.4.5
Clinical Diagnostics

14.4.5.1 **Critical Care Analysis**

The Osmetech/Roche OSPI critical care analyzer can measure up to eight critical care analytes in whole blood (Tusa and He, 2005): pH, CO_2/Hb, O_2, Na^+, K^+, Ca^{2+}, or Cl^-. Six appropriate fluorescent molecular sensors are embedded in separate disks on a disposable cartridge. It takes not more than 2 minutes to get the result from 120 μl of whole blood. The list of the analytes, the normal range, the pathological range, and the relevant molecular sensors are reported in Table 14.7. Pyranine is used as a fluorescent pH indicator. CO_2 is determined in the following way: a membrane blocks the passage of H_3O^+ and other ions into the sensor but allows CO_2 through, where it forms carbonic acid, increasing the local concentration of H_3O^+ that is determined by pyranine. Cation recognition is achieved by fluoroionophores consisting of the naphthalimide fluorophore linked to an azacrown for Na^+, a cryptand for K^+, and a chelating moiety for Ca^{2+} (Figure 14.21). The dissociation constants are consistent with the pathological ranges.

14.4.5.2 Angiography

Angiography is a good example of application of fluorescence for in vivo diagnostics. Fluorescent tracers are used to visualize the blood vessels of the retina in order to detect anomalies: lesions, aneurisms, and occlusions resulting from specific eye diseases.

A fluorescent dye (fluorescein or indocyanine green) is injected into a vein in the arm of the patient. The dye is conveyed by the circulatory system and reaches the retina vessels.

Figure 14.22 shows an angiogram observed after injection of fluorescein: the existence of bright points leads to the diagnostic of an early diabetic retinopathy showing microaneurisms.

Indocyanine green provides information on deeper tissues, as compared to fluorescein. Because it absorbs near 800 nm and emits in the near infrared, it can reveal deeper vessels called *choroidal vessels*. In the case of an aged-related macular degeneration, leaky vessels, that is, choroidal neovascularization, can be visualized, while the angiogram with fluorescein also shows fluid leakage for the same patient.

14.4.5.3 Bladder Tumor Detection

The measurement of relative fluorescence intensities of tryptophan and NADH by endoscopy permits early detection of bladder tumors in situ. Light pulses from an excimer laser (308 nm) are conveyed by a catheter containing several optical fibers. Fluorescence is

Tab. 14.7 Fluorescent sensors for critical care analytes. (Adapted from Tusa and He (2005).)

	Normal range	Pathological range	Fluorescent sensor
pH	7.35–7.45	6.7–7.7	pyranine
$P(CO_2)$ (torr)	30–45	10–200	pyranine
$P(O_2)$ (torr)	70–100	20–600	Oxygen-sensitive dye
Na^+ (mM)	135–145	100–180	Naphthalimide dye – azacrown[a]
K^+ (mM)	3–5	1–10	Naphthalimide dye – cryptand[a]
Ca^{2+} (mM)	1.0–1.4	0.3–2.5	Naphthalimide dye – chelator[a]

[a] See Figure 14.21.

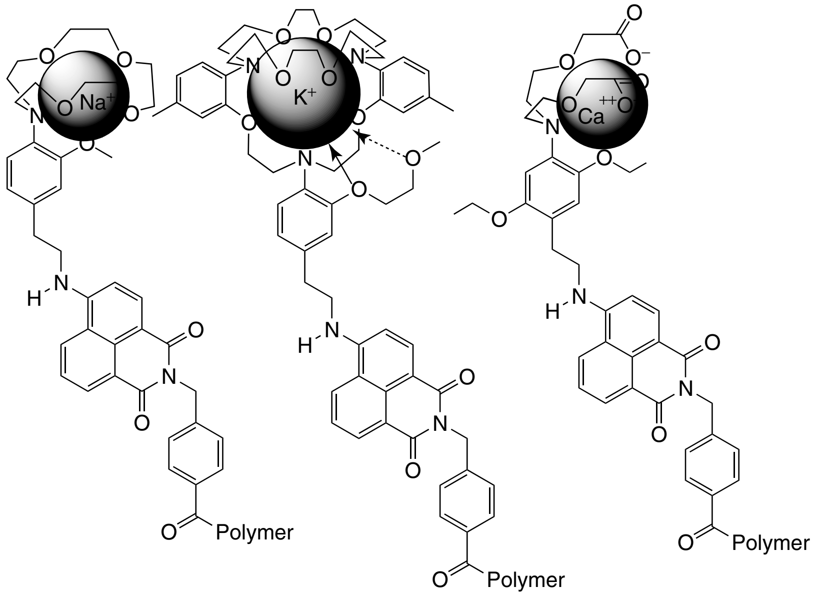

Fig. 14.21 Fluoroionophores for the selective detection of Na^+, K^+, and Ca^{2+}. (Reproduced from Tusa and He (2005) by permission of the Royal Society of Chemistry.)

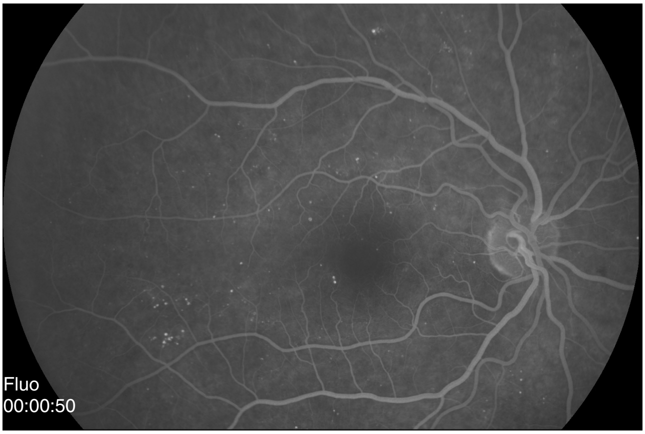

Fig. 14.22 Angiogram showing microaneurisms (bright points). (Courtesy of Dr J.C. Hache, Lille hospital, France.)

collected by other fibers and analyzed by a monochromator and a photomultiplier (Figure 14.23) (Avrillier *et al.*, 1997). The fluorescence spectra are different according to the state of the tissue. For tumors and carcinoma in situ (CIS), the ratio of the fluorescence intensities at 360 and 440 nm is statistically greater than 2. This test is reliable and can be run during routine visit.

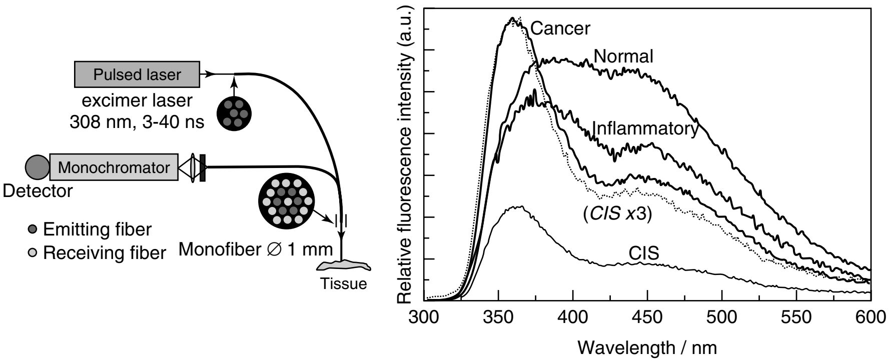

Fig. 14.23 Experimental set-up for the detection of bladder tumors. Fluorescence spectra of normal, inflammatory and cancerous tissues upon excitation at 308 nm (CIS, carcinoma in situ). (Adapted from Avrillier *et al.* (1997).)

14.4.6
Miscellaneous

14.4.6.1 Fluorescence LIDAR

LIDAR is the acronym of light detection and ranging (which is a transposition of RADAR for radiowave detection and ranging). LIDAR is a useful tool for atmospheric monitoring and, in particular, for the detection of SO_2 and NO_2 in atmosphere. This technique that is based on back-scattering and absorption of species can be extended to fluorescence detection by using appropriate emission filters. The applications concern the status of vegetation, the estimation of plankton in sea water, and imaging of the façades of historical monuments.

A Nd:YAG laser is often used for these applications: it provides excitation pulses of about 10 nanoseconds duration at 532 and 355 nm, thanks to frequency doubling and tripling, respectively.

Vegetation It is of great interest to monitor the chemical activities and the status of trees and forests. Early detection and mapping of damaged vegetation as a consequence of pollution can also be made. Remote detection of chlorophyll by a LIDAR is possible. Upon excitation at 532 nm, the chlorophyll concentration can be estimated as a function of time by taking the ratio of the intensity of the images at 740 and 685 nm (Saito *et al.*, 2000). The test is validated by the comparison of the average intensity ratio and the amount of chlorophyll determined by chromatography. The average intensity ratio follows the chlorophyll concentration that increases in summer and decreases in autumn.

Sea Water The detection of fluorescence from chlorophyll in phytoplankton and from phycoerythrin-containing plankton (e.g., cyanobacteria) (both excited at 532 nm) is possible from aircrafts or ships, thanks to a LIDAR. The biomass can thus be estimated. In addition, the fluorescence of dissolved organic substances can be detected upon excitation at 355 nm. NASA has developed an airborne oceanographic LIDAR and a shipboard laser fluorometer to achieve such evaluations.

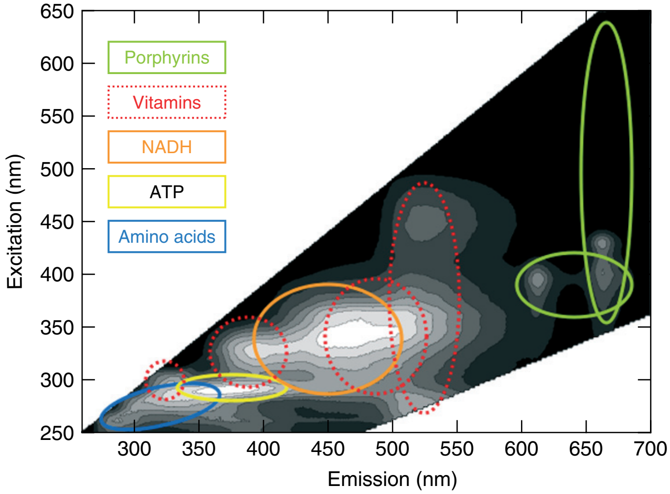

Fig. 14.24 Two-dimensional excitation–emission patterns of typical food-relevant fluorophores. (Reproduced with permission from Christensen *et al.* (2006). Copyright 2006 American Chemical Society.)

Façades of Historical Monuments Svanberg and coworkers have developed a system for imaging the laser-induced fluorescence from the façades of historical monuments (Weibring *et al.*, 2001). The aim is the visualization of areas of biodeterogen (lichens, green algae) and the identification of different stone types with natural surface aging or crust and pollution deposited layers. Façade status assesment is thus obtained in view of restauration planning.

The LIDAR system is placed in a truck at a distance of about 60 m from the façade. A 355-nm pulsed laser beam is swept over the façade row by row. The spectrally resolved fluorescence signals are recorded on each point.

14.4.6.2 Food Science

Food systems (meat, fish, dairy products, oils, cereals, beer, fruits, vegetables, etc.) contain several fluorescent substances such as proteins, vitamins, secondary metabolites, pigments, toxins, and flavoring compounds. The intrinsic fluorescence from intact food can thus provide valuable information on the quality of food products, their authenticity, and the effect of processing and aging. The emitted fluorescence is quite complex because of the presence of several intrinsic fluorophores. Moreover, the fluorescence signals are affected by phenomena such as Rayleigh and Raman scattering, inner filter effects, self-quenching, and so on. The complex signals can be decomposed from a series of two-dimensional excitation–emission patterns into the pure constituent signals using multiway chemometrics (Christensen *et al.*, 2006). Figure 14.24 shows the spectral properties of selected food-relevant fluorophores.

14.4.6.3 Forensics

On a crime scene, the visualization of fingerprints by means of powders is well known. The fluorescent powders are the most efficient.

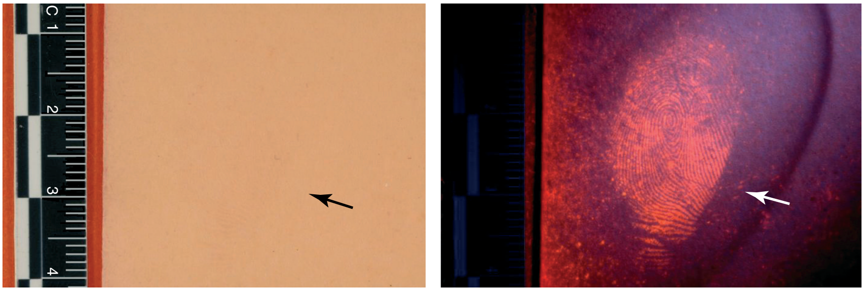

Fig. 14.25 Fingerprint on a piece of paper revealed by fluorescence after treatment with diaza-9-fluorenone. (Courtesy of the Identité judiciaire de la Préfecture de police de Paris.)

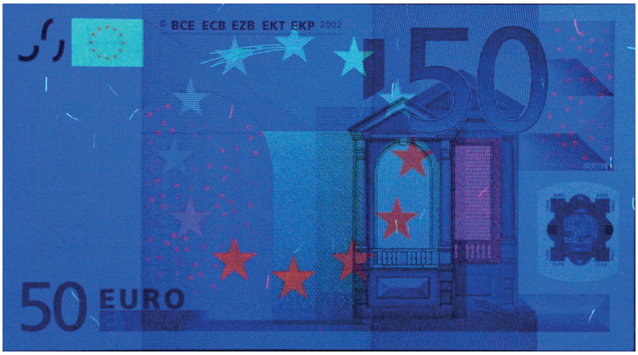

Fig. 14.26 A bill of 50 Euros shows specific color under UV illumination (photograph by the author.)

Fingerprints on papers are more difficult to detect, especially when they are old. In this case, the procedure is to soak the paper in a solution of diaza-9-fluorenone (that reacts with aminoacids present in the sweat deposited by the fingers), and to dry it. Observation under UV illumination can reveal the fluorescent fingerprints provided that the blue fluorescence of paper due to whitening agents is rejected by means of an appropriate filter (Figure 14.25).

14.4.6.4 **Counterfeit Detection**

In many situations, the luminescent pattern serves as a security check for documents, especially for bills. Various luminescent compounds either organic or inorganic are employed. For example, the Euros bills show different colors (e.g., stars, Europe map, etc.) under UV illumination (Figure 14.26). Small fluorescent fibers can also be seen.

The use of semiconductor nanocrystals (Section 14.4.4.4) to counterfeit detection is to be expected in the near future (bank notes, ID cards, badges, etc.).

Fluorescence is also a means to detect counterfeited art works: paintings, ceramics, glassware, sculptures, ivories, old marbles, and so on. In fact, fluorescence depends on ageing. Moreover, traces of restoration, overpainting, falsified signatures, and so on can also be easily detected.

In conclusion, fluorescence is a very powerful tool in both fundamental and applied research. The technological and industrial applications are numerous in variegated fields. The success of fluorescence as a tool of investigation of the structure and dynamics of matter or living systems arises from the high sensitivity of fluorometric techniques, the specificity of fluorescence characteristics due to the microenvironment of the emitting molecule, and the ability of the latter to provide spatial and temporal information. After a long history, no doubt that fluorescence still has a bright future.

References

Andrews, D.L. and Demidov, A.A. (eds) (1999) *Resonance Energy Transfer*, John Wiley & Sons, Ltd, New York.

Auroux, P.A., Iossifidis, D., Reyes, D., and Manz, A. (2002) *Anal. Chem.*, **74**, 2637.

Avrillier, S., Tinet, E., Ettori, D., and Anidjar, M. (1997) *Phys. Scr.*, **72**, 87.

Baleizao, C., Nagl, S., Borisov, S.M., Schäferling, M., Wolfbeis, O., and Berberan-Santos, M.N. (2007) *Chem. Eur. J.*, **13**, 3643.

Benninger, R.K.P., Koç, Y., Hofmann, O., Requejo-Isidro, J., Neil, M.A.A., French, P.M.W., and deMello, A.J. (2006) *Anal. Chem.*, **783**, 2272.

Berberan-Santos, M.N. (ed.) (2007) *Fluorescence of Super Molecules, Polymers and Nanosystems*, Springer Series on Fluorescence, Vol. 4, Springer, Berlin.

Berberan-Santos, M.N. and Garcia, J.M.M. (1996) *J. Am. Chem. Soc.*, **118**, 9391.

Bissell, R.A., de Silva, A.P., Gunaratne, H.Q.N., Lynch, P.L.M., Maguire, G.E.M., and Sandanayake, K.R.A.S. (1992) *Chem. Soc. Rev.*, **21**, 187.

Bokobza, L. (1990) *Prog. Polym. Sci.*, **15**, 337.

Bokobza, L., Monnerie, L. (1986) in *Photophysical and Photochemical Tools in Polymer Science* (ed. M.A. Winnik), D. Reidel Publishing Company, pp. 449–466.

Bur, A.J., Wangel, M.G., and Roth, S. (2002) *Appl. Spectrosc.*, **56**, 174.

Chandrashekaran, N. and Kelly, L.A. (2001) *J. Am. Chem Soc.*, **123**, 9898.

Chandrashekaran, N. and Kelly, L.A. (2004) in *Reviews in Fluorescence* (eds C.D. Geddes and J.R. Lakowicz), Kluwer Academic, New York, p. 21.

Christensen, J., Norgaard, L., Bro, R., and Engelsen, S. (2006) *Chem. Rev.*, **106**, 1979.

Dahan, M., Lévi, S., Luccardini, C., Rostaing, P., Riveau, B., and Triller, A. (2003) *Science*, **302**, 442.

Dake, H.C. and De Ment, J. (1941) *Fluorescent Light and its Applications*, Chemical Publishing Co., Brooklyn.

Desvergne, J.P. and Czarnik, A.W. (eds) (1997) *Chemosensors of Ion and Molecule Recognition*, Kluwer, Dordrecht.

Engler, R.H., Mérienne, M.C., Klein, C., and Le Sant, Y. (2002) *Aerosp. Sci. Techn.*, **6**, 313.

Fister, J.C., Rank, D., and Harris, J.M. (1995) *Anal. Chem.*, **67**, 4269.

Fisz, J.J. (1985) *Chem. Phys.*, **99**, 177; *ibid.* 1989, **132**, 303; *ibid.* 1989, **132**, 315.

Gao, X., Cui, Y., Levenson, R.M., Chung, L.W.K., and Nie, S. (2004) *Nat. Biotechnol.*, **22**, 969.

Georgescauld, D., Desmasez, J.P., Lapouyade, R., Babeau, A., Richard, H., and Winnik, M. (1980) *Photochem. Photobiol.*, **31**, 539–545.

Göppelsröder, F. (1868) *J. Prakt. Chem.*, **104**, 10.

Harvey, E.N. (1957) *A History of Luminescence*, The American Philosophical Society, Philadelphia.

Herzum, C., Boit, C., Kölzer, J., Otto, J., and Weiland, R. (1998) *Microelectron. J.*, **29**, 163.

Hof, M., Hutterer, R., and Fidler, V. (eds) (2005) *Fluorescence Spectroscopy in Biology*, Springer Series on Fluorescence Vol. 3, Springer, Berlin.

Jameson, D.M., Croney, J.C., and Moens, P.D.J. (2003) *Methods Enzymol.*, **360**, 1.

Kalyanasundaran, K. (1987) *Photochemistry in Microheterogeneous Systems*, chap. 2, Academic Press, Orlando.

Kinosita, K., Kawato, S., and Ikegami, A. (1977) *Biophys. J.*, **20**, 289.

Knight, J. (2003) *Nature*, **426**, 372.

Kolodner, P. and Tyson, J. (1982) *Appl. Phys. Lett.*, **409**, 782; 1983, **421**, 117.

Kraayenhof, R. and Visser, A.J.W.G. (eds) (2003) *Fluorescence Spectroscopy, Imaging and Probes*, Springer Series on Fluorescence, Vol. 2, Springer, Berlin.

Lakowicz, J.R. (ed.) (1991–2007) *Topics in Fluorescence Spectroscopy*, Vol. 1–11, Plenum Press, New-York.

Lakowicz, J.R. (ed.) (1994) *Probe Design and Chemical Sensing, Topics in Fluorescence Spectroscopy*, Vol. 4, Plenum, New-York.

Lakowicz, J.R. (2006) *Principles of Fluorescence Spectroscopy*, 3rd edn, Springer, New York.

Lipari, G. and Szabo, A. (1980) *Biophys. J.*, **30**, 489.

Lippincott-Schwartz, J. and Patterson, G.H. (2003) *Science*, **300**, 87.

Liu, T., Campbell, B.T., Burns, S.P., and Sullivan, J.P. (1997) *Appl. Mech. Rev.*, **50**, 227.

Liu et, T. and Sullivan, J.P. (2005) *Pressure and Temperature Sensitive Paints*, Springer, Berlin.

Loufty, R.O. (1986) *Pure Appl. Chem.*, **58**, 1239.

Narayanaswamy, R. and Wolfbeis, O.S. (eds) (2004) *Optical Sensors, Industrial, Environmental and Diagnostic Applications*, Springer Verlag, Berlin.

Navarra, K.R., Rabe, D.C., Fonov, S.D., Goss, L.P., and Hah, C. (2001) *J. Turbomachinery.*, **123**, 823.

Pottel, H., Herreman, W., Van der Meer, B.W., and Ameloot, M. (1986) *Chem. Phys.*, **102**, 37.

Pringsheim, P. and Vogel, M. (1943) *Luminescence of Liquids and Solids and its Practical Applications*, Interscience, New York.

Radley, J.A. and Grant, J. (1933) *Fluorescence Analysis in Ultraviolet Light*, Van Nostrand Co., New York.

Reyes, D., Iossifidis, D., Auroux, P.A., and Manz, A. (2002) *Anal. Chem.*, **74**, 2623.

Ross, D., Gaitan, M., and Locascio, M.E. (2001) *Anal. Chem.*, **73**, 4117.

Saito, Y., Saito, R., KawaHara, T.D., Nomura, A., and Takeda, S. (2000) *For. Ecol. Manag.*, **128**, 129.

Schulman, S.G. (ed.) (1985–1993) *Molecular Luminescence Spectroscopy*, Parts 1–3, Wiley Interscience, New York.

de Silva, A.P., Gunaratne, H.Q.N., Gunnlaugsson, T., Huxley, A.J.M., McCoy, C.P., Rademacher, J.T., and Rice, T.E. (1997) *Chem. Rev.*, **97**, 1515.

de Silva, A.P. and Tecilla, P. Guest editors, (2005) Fluorescent sensors, special issue of *J. Mater. Chem.*, **15**, 2617.

Skelley, A.M., Cherer, J.R., Aubrey, A.D., Grover, W.H., Ivester, R.H.C., Ehrenfreund, P., Grunthaner, F.J., Bada, J.L., and Mathies, R.A. (2005) *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 1041.

Szabo, A. (1984) *J. Chem. Phys.*, **81**, 150.

Tusa, J.K. and He, H. (2005) *J. Mater. Chem.*, **15**, 2640.

Uchiyama, S., de Silva, A.P., and Iwau, K. (2006) *J. Chem. Educ.*, **83**, 720.

Valeur, B. (1993) in *Molecular Luminescence Spectroscopy. Methods and Applications*, Part 3, (ed. S.G. Schulman), Wiley-Interscience, pp. 25–84.

Valeur, B. (1994) in *Probe Design and Chemical Sensing, Topics in Fluorescence Spectroscopy*, Vol. 4, (ed. J.R. Lakowicz), Plenum, New-York, pp. 21–48.

Valeur, B. (2002) *Molecular Fluorescence. Principles and Applications*, Wiley-VCH Verlag GmbH, Weinheim.

Valeur, B. (2007) in *Fluorescence of Super Molecules, Polymers and Nanosystems*, Springer Series on Fluorescence, Vol. 4, (ed. M.N. Berberan-Santos), Springer, pp. 21–43.

Valeur, B., Berberan-Santos, M.N., and Martin, M.M. (2007) in *Analytical Methods in Supramolecular Chemistry* (ed. C.A. Schalley), Wiley-VCH Verlag GmbH, pp. 220–264.

Valeur, B. and Brochon, J.-C. (eds) (2001) *New Trends in Fluorescence Spectroscopy. Applications to Chemical and Life Sciences*, Springer Series on Fluorescence, Vol. 1, Springer, Berlin.

Valeur, B. and Leray, I. (2000) *Coord. Chem. Rev.*, **205**, 3.

Valeur, B. and Leray, I. (2001) in *New trends in Fluorescence Spectroscopy. Application to Chemical and Life Sciences* (eds B. Valeur and J.C. Brochon), Springer, Berlin, pp. 187–207.

Van der Meer, W., Kooyman, R.P.H., and Levine, Y.K. (1982) *Chem. Phys.*, **66**, 39.

Van der Meer, W., Pottel, H., Herreman, W., Ameloot, M., Hendrickx, H., and Schröder, H. (1984) *Biophys. J.*, **46**, 515.

Vauhkonen, M., Sassaroli, M., Somerharju, P., and Eisinger, J. (1990) *Biophys. J.*, **57**, 291.

Viriot, M.L., Bouchy, M., Donner, M., and André, J.C. (1983) *Photobiochem. Photobiophys.*, **5**, 293.

Weibring, P., Johansson, T., Edner, H., Svanberg, S., Sundner, S., Raimondi, V., Cecchi, G., and Pantani, L. (2001) *Appl. Opt.*, **40**, 6111.

Zachariasse, K.A., Kozankiewicz, B., and Kühnle, W. (1983) in *Photochemistry and Photobiology*, Vol. II, (ed. A.H. Zewail), Harwood, London, pp. 941–960.

Zannoni, C., Arcioni, A., and Cavatorta, P. (1983) *Chem. Phys. Lipids*, **32**, 179.

Zimmer, M. (2002) *Chem. Rev.*, **102**, 759.

Further Reading

Baeyens, W.R.G., deKeukeleire, D. and Korkidis, K. (eds) (1991) *Luminescence Techniques in Chemical and Biochemical Analysis*, Marcel Dekker, New York,.

Becker, R.S. (1969) *Theory and Interpretation of Fluorescence and Phosphorescence*, Wiley Interscience, New York.

Berlman, I.B. (1965, 1971) *Handbook of Fluorescence Spectra of Aromatic Molecules*, Academic Press, New York.

Birks, J.B. (ed.) (1975) *Organic Molecular Photophysics*, Vol. 1 and 2, John Wiley & Sons, Londres.

Cundall, R.B. and Dale, R.E. (eds) (1983) *Time-Resolved Fluorescence Spectroscopy in Biochemistry and Biology*, Plenum Press, New York.

Czarnik, A.W. (ed.) (1992) *Fluorescence Chemosensors for Ion and Molecule Recognition*, American Chemical Society, Washington.

Demas, J.N. (1983) *Excited State Lifetime Measurement*, Academic Press, New York.

Guilbault, G. (ed.) (1973, 1990) *Practical Fluorescence*, 1st edn, 2nd edn, Marcel Dekker, New York.

Jameson, D.M. and Reinhart, G.D. (eds) (1989) *Fluorescent Biomolecules*, Plenum Press, New York.

Mielenz, K.D. (ed.) (1982) *Measurement of Photoluminescence*, Academic Press, Washington, DC.

Miller, J.N. (ed.) (1981) *Standards in Fluorescence Spectrometry, Ultraviolet Spectrometry Group*, Chappman and Hall, Londres.

O'Connor, D.V. and Phillips, D. (1984) *Time-Correlated Single Photon Counting*, Academic Press, Londres.

Parker, C.A. (1968) *Photoluminescence of Solutions*, Elsevier, Amsterdam.

Rettig, W., Strehmel, B., Schrader, S., and Seifert, H. (eds) (1999) *Applied Fluorescence in Chemistry, Biology and Medicine*, Springer, Berlin.

Slavik, J. (ed.) (1996) *Fluorescence Microscopy and Fluorescent Probes*, Plenum Press, New York.

Wolfbeis, O.S. (ed.) (1993) *Fluorescence Spectroscopy. New Methods and Applications*, Springer-Verlag, Berlin.

15

Resonance Energy Transfer

David L. Andrews and David S. Bradshaw

15.1 **Introduction 535**
15.2 **History of RET 535**
15.2.1 The First Experiments 535
15.2.2 Early Developments of Theory 536
15.2.3 Förster Theory 536
15.3 **The Photophysics of RET 537**
15.3.1 Primary Excitation Processes 537
15.3.2 Coupling of Electronic Transitions 538
15.3.3 Dissipation and Line-Broadening 538
15.3.4 The Förster Equation 539
15.3.5 Orientation Dependence 540
15.3.6 Förster Radius 541
15.3.7 Polarization Features 542
15.3.8 Diffusion Effects 543
15.3.9 Long-Range Transfer 544
15.3.10 Dexter Transfer 544
15.4 **Applications of RET to Molecular Biology 545**
15.4.1 Spectroscopic Ruler 545
15.4.2 Conformational Change 546
15.4.3 Intensity-Based Imaging 547
15.4.4 Lifetime-Based Imaging 548
15.4.5 Other Applications 550
Acknowledgments 550
References 550
Further Reading 553

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

15.1
Introduction

Resonance energy transfer[1] (RET) is a spectroscopic process whose relevance in all major areas of science is reflected both by the wide prevalence of the effect and through numerous technical applications. It is an optical near-field mechanism that effects the transportation of electronic excitation between physically distinct atomic or molecular components, based on transition dipole–dipole coupling. In this chapter, a comprehensive survey of the process is presented, beginning with an outline of the history and highlighting the early contributions of Perrin and Förster. This is followed by a review of the photophysics and then a discussion of some prominent applications of RET. Particular emphasis is given to techniques used in molecular biology, ranging from the "spectroscopic ruler" measurements of functional group separation to fluorescence lifetime microscopy. Finally, applications to synthetic polymers and chemical sensors are examined.

15.2
History of RET

15.2.1
The First Experiments

RET is a process by means of which the energy of an excited atom or molecule (usually called the *donor*, but known historically as the *sensitizer*) is transferred nonradiatively to an acceptor molecule ("activator"), through intermolecular dipole–dipole coupling. The origins of its discovery can be traced back to 1922, when the phenomenon of RET ("sensitized fluorescence") was first experimentally observed by Franck (1922); Cario (1922); Cario and Franck (1922) in the gas phase. This spectroscopic experiment involved illuminating a mixture of mercury and thallium vapors at a wavelength absorbed solely by the mercury; the resulting fluorescence spectra was proven to include frequencies that could only be emitted from thallium. Such energy transfer in vapors was at first assumed to be uniquely associated with interatomic collisions, but a discovery that transfer could occur at larger separations than the collision radii showed that this was not necessarily the case. Soon, RET was also being observed in solutions (Gaviola

---

[1] RET is also known as Förster – or fluorescence – resonance energy transfer (FRET), or electronic energy transfer (EET).

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

and Pringsheim, 1924), and over the following years in many other physical systems.

15.2.2
Early Developments of Theory

The first theoretical explanation of the phenomenon was proposed by the Nobel laureate Perrin (1927). He recognized that energy could be transferred from an excited molecule to its neighbors amongst closely spaced molecules through dipole interactions; he named this process as "transfert d'activation," and his paper on the subject became the earliest attempt to describe nonradiative (near-field) energy transfer. However, despite its initial success, Perrin's model incorrectly predicted that nonradiative energy transfer should be possible between dye molecules up to an intermolecular distance of 1000 Å, deriving from an inaccurate assumption that the molecules would act as Hertzian oscillators with exactly defined resonance frequencies. Five years later (Perrin, 1932), Perrin's son Francis developed a corresponding quantum mechanical theory of RET, based on the results of Kallman and London (1928). In this work, he recognized a "spreading of absorption and emission frequency" due to the interactions of the dye with the solvent, thus reducing the probability of energy transfer. As a result, efficient transfer was calculated to occur up to 150–250 Å, still approximately a factor of 3 greater than the experimentally observed. A detailed and readable survey of these early contributions of J. and F. Perrin can be found in a review by Berberan-Santos (2001).

15.2.3
Förster Theory

Extending the ideas of J. and F. Perrin, Förster developed the first essentially correct theoretical treatment of RET (Förster, 1946, 1948, 1959). Förster determined that energy transfer, through dipolar coupling between molecules, mostly depends on two important quantities: spectrum overlap and intermolecular distance. Following the observation that "the absorption and fluorescence spectra of similar molecules are far from completely overlapping," he found a means to quantify the spectral overlap integral. The dipole–dipole interaction was known to have an inverse proportionality on the cube of the molecular separation. Because the rate of energy transfer is proportional to the square of this coupling, it thus depends on the sixth power of the separation, that is, the famous R^{-6} distance-dependence law. Moreover, the acceptor distance at which this rate equates to that of spontaneous emission by the donor, now termed the *Förster radius* R_0, has been calculated to be between 10 and 100 Å, in agreement with experimental observations.

Much later, the distance dependence predicted by Förster was fully verified by fluorescence studies of donor–acceptor pairs at known separations (Latt, Cheung and Blout, 1965; Stryer and Haugland, 1967), leading to the suggested employment of RET as a "spectroscopic ruler" by Stryer and Haugland (1967); hence, a technique to measure the proximity relationships and conformational change in macromolecules was realized (see Section 15.3). With the advent of the laser, in the 1960s, the modern understanding of RET led to a raft of modern applications. An excellent in-depth review on the history of RET is given by Clegg (2006).

15.3
The Photophysics of RET

Resonance energy transfer is a mechanism that is now known to operate across a diverse and extensive range of physical systems, encompassing not only gases and dye solutions but also protein complexes, doped crystals, polymers, and so on. Nonetheless, at a fundamental level, it is possible to identify numerous common features in the underlying photophysics.

15.3.1
Primary Excitation Processes

To study the subject in detail, let us commence with the photoexcitation process that creates the conditions for RET to occur. When resonant ultraviolet or visible radiation impinges on any non-homogeneous dielectric material, the primary result of photon absorption is the population of electronic excited states in individual atomic, molecular, or other nanoscale centers – henceforth, to be grouped together under the generic term "chromophore." Typically, such absorption is immediately followed by a rapid but partial degradation of the acquired energy, the associated losses (largely due to vibrational dissipation) ultimately to be manifest in the form of heat. This effect owes its origin to the principle that the release of electronic energy by fluorescence generally occurs from the lowest vibrational level of the excited state. However, if any nearby chromophore has a suitably disposed electronic state, of a similar or slightly lower energy, that neighbor may acquire the major part of the electronic excitation

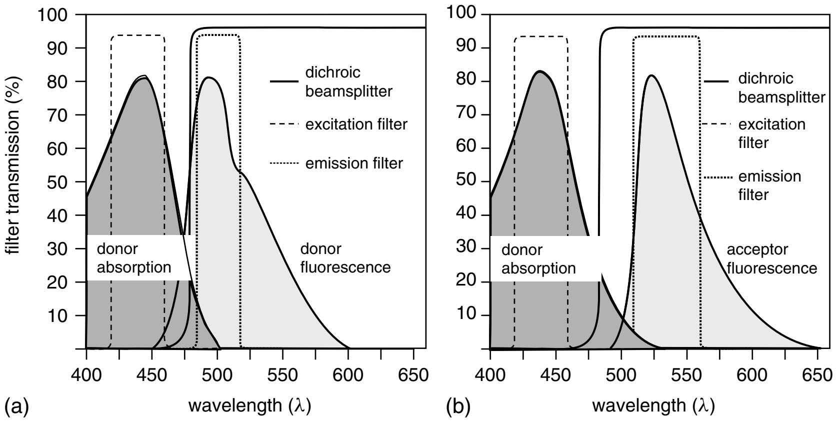

Fig. 15.1 Typical spectral discrimination between the fluorescence from donor and acceptor species (here notionally based on a cyan fluorescent protein donor and a yellow fluorescent protein acceptor): (a) the transmission characteristics of a short-wavelength filter ensure initial excitation of only the donor; a dichroic beam-splitter and another narrow emission filter ensure that only the (Stokes-shifted) fluorescence from the donor reaches a detector; (b) in the same system, a longer-wavelength emission filter ensures capture of only the acceptor fluorescence, following RET.

through RET – a process that takes place well before any further thermal degradation of the excited state energy occurs. The mechanism is most commonly studied through spectrometric differentiation of fluorescence emerging from the initially excited donor and from the acceptor species, as illustrated in Figure 15.1. As is shown in the following, the propensity for energy to be transferred between any two chromophores is severely restricted by distance, and if no suitable acceptor is within reach, the donor will generally shed its energy by fluorescence or local dissipation.

15.3.2
Coupling of Electronic Transitions

In systems where RET occurs, the donors and acceptors are usually fluorophores, that is, chromophores that have the capacity to exhibit fluorescence decay. Moreover, in RET, the transitions of donor decay and acceptor excitation are generally electric dipole-allowed – although other possibilities occasionally arise. Accordingly, the theory of energy transfer, for donor–acceptor displacements beyond the region of significant wavefunction overlap, is traditionally conceived in terms of an electrodynamical coupling between transition dipoles.

Consider the pairwise transfer of excitation between two chromophores D and A. In the context of this elementary mechanism, D is designated as the donor and A as the acceptor. Specifically, let it be assumed that prior excitation of the donor generates an electronically excited species D^*. Forward progress of the energy is then accompanied by donor decay to the ground electronic state. Acquiring the energy, D undergoes a transition from its ground to its excited state. The complete RET process is generally a singlet–singlet coupling mechanism[2], due to the constraints of spin conservation (although this does not apply to Dexter exchange, see below), and its entirety may be expressed by the following chemical equation:

$$ {}^1D^* + {}^1A \xrightarrow{\text{RET}} {}^1D + {}^1A^* \qquad (15.1) $$

The excited acceptor, A^*, subsequently decays either in a further transfer event or by another means such as fluorescence. Because the D^* and A^* excited states are real, with measurable lifetimes, the core process of energy transfer itself is fundamentally separable from the initial electronic excitation of D and the eventual decay of A; the latter processes do not, therefore, enter into the theory of the pair transfer.

15.3.3
Dissipation and Line-Broadening

To delve more deeply into the nature of the process, it needs to be recognized that Eq. (15.1) tells only part of the story, dealing as it does with only electronic excitations. In general, other dissipative processes are also engaged. In a solid, the linewidth of optical transitions manifests the influence of local electronic environments which, in the case of strong coupling, may lead to the production of phonon side-bands. Similar effects in solutions or disordered solids represent inhomogeneous interactions with a solvent or host, while the broad bands exhibited by chromophores in complex molecular systems signify extensively overlapped vibrational levels,

---

2) Triplet-triplet energy transfer is also allowed by the Förster mechanism.

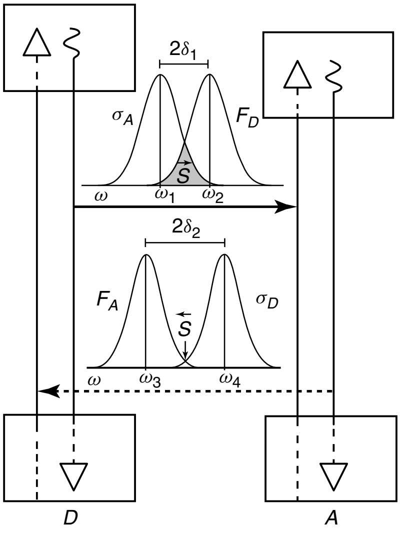

Fig. 15.2 Energetics and spectral overlap features (top) for energy transfer from *D* to *A* (and below, potentially backward transfer from *A* to *D*). For each chromophore, *F* denotes the fluorescence spectrum and σ the absorption. Wavy downward lines denote vibrational dissipation.

including those associated with skeletal modes of the superstructure. In each case, energy level broadening can allow pair transfer to occur at any point within the region of overlap between the donor emission and the acceptor absorption bands, as illustrated in Figure 15.2.

15.3.4
The Förster Equation

The Förster theory delivers an expression for the rate of pairwise energy transfer, w_F, valid for any donor–acceptor separation, R, that is substantially smaller than the wavelengths of visible radiation. For systems where the common host material for the donor and acceptor has refractive index n, at the optical frequency corresponding to the mean transferred energy, the Förster result is as follows (Demidov and Andrews, 2001):

$$w_F = \frac{9\kappa^2 c^4}{8\pi \tau_{D^*} n^4 R^6} \int F_D(\omega)\sigma_A(\omega)\frac{d\omega}{\omega^4} \tag{15.2}$$

In this expression, $F_D(\omega)$ denotes the normalized fluorescence spectrum of the donor, $\sigma_A(\omega)$ represents the linear absorption cross section of the acceptor, and ω is an optical frequency in radians per unit time; the specific form of the integral within which they appear is known as the *spectral overlap* – one of the key determinants of energy transfer efficiency (Andrews and Rodríguez, 2007). Also in Eq. (15.2), c is the speed of light and τ_{D^*} is the associated radiative decay lifetime in the absence of transfer. The latter is related to the measured fluorescence lifetime τ_{fl} through the fluorescence quantum yield $\eta = \tau_{\text{fl}}/\tau_{D^*}$, where τ_{fl}^{-1} is explicitly expressible as

$$\frac{1}{\tau_{\text{fl}}} = \frac{1}{\tau_{D^*}} + \frac{1}{\tau_{D^*}}\left(\frac{R_0}{R}\right)^6 \tag{15.3}$$

The last term on the right-hand side of Eq. (15.3) is expressed with reference to the Förster radius R_0 – the distance at which the rates of donor deactivation by RET and by spontaneous fluorescence become equal. As is evident from Figure 15.2, the propensity for forward transfer is usually significantly greater than that for backward transfer, due to a sizeable difference in the spectral overlaps for the two processes.

15.3.5
Orientation Dependence

The κ factor in Eq. (15.2) depends on the orientations of the donor and acceptor, both with respect to each other and with respect to their mutual displacement unit vector $\hat{\boldsymbol{R}}$, as follows:

$$\kappa = (\hat{\boldsymbol{\mu}}_D \cdot \hat{\boldsymbol{\mu}}_A) - 3(\hat{\boldsymbol{R}} \cdot \hat{\boldsymbol{\mu}}_D)(\hat{\boldsymbol{R}} \cdot \hat{\boldsymbol{\mu}}_A) \tag{15.4}$$

For each chromophore, $\hat{\boldsymbol{\mu}}$ designates a unit vector in the direction of the appropriate transition dipole moment. The possible values of κ^2, as featured in Eq. (15.2), lie in the range (0, 4). It is evident that in the case of fixed chromophore positions and orientations, the result delivered by Eq. (15.4) is a function of three independent angles, as shown and defined in Figure 15.3:

$$\kappa = \cos\theta_T - 3\cos\theta_D \cos\theta_A \tag{15.5}$$

Unfavorable orientations can thus reduce the rate of energy transfer to zero; other configurations, including many of those found in photobiological systems, optimize the transfer rate. The angular disposition of chromophores is therefore a very important facet of energy transfer. It is important to note that transfer is not *necessarily* precluded when the transition moments of the donor and acceptor lie in perpendicular directions – provided that neither is also disposed orthogonally to $\boldsymbol{R}(=R\hat{\boldsymbol{R}})$.

In any, at least partially, fluid or disordered system, the relative orientation of all donor–acceptor pairs may not be identical, and it is then the distributional average of κ^2 that determines the overall measured response. In the isotropic case (completely uncorrelated orientations), the κ^2 factor averages to $^2/_3$; departures from this value provide the quantitative signature of a degree of orientational correlation. In molecules of sufficiently high symmetry, it can also happen that either the donor or the acceptor transition moment is not unambiguously identifiable with a particular direction in the corresponding chromophore reference frame. Specifically, the electronic transition may then relate to a transition involving a degenerate state – as can occur with square planar complexes, for example (Galli *et al.*, 1993). Alternatively, the same

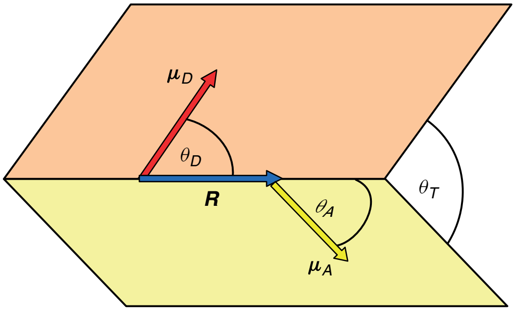

Fig. 15.3 Relative orientations and positions of the donor and acceptor and their transition moments: Here, angles θ_D and θ_A subtended by donor and acceptor transition moments ($\boldsymbol{\mu}_D$ and $\boldsymbol{\mu}_A$, respectively) against the inter-chromophore displacement vector, **R**; the symbol θ_T is the angle between the transition moments.

observational features might indicate rapid but orientationally confined motions. The considerable complication that each of these effects brings into the trigonometric analysis of RET has been extensively researched and reported by van der Meer (1999).

15.3.6
Förster Radius

In applications of RET to spectroscopy, as noted above, it is usually significant that the electronically excited donor can in principle release its energy by spontaneous decay, the ensuing fluorescence also being amenable to detection by any suitably placed photodetector. Because the alternative possibility (that of energy being transferred to another chromophore within the system) has such a sharp decline in efficiency as the distance to the acceptor increases, it is common to invoke the critical distance R_0. The Förster rate equation is itself often cast in an alternative form, exactly equivalent to Eq. (15.2), explicitly exhibiting this critical distance (Lakowicz, 1999):

$$w_F = \frac{3\kappa^2}{2}\frac{1}{\tau_{D^*}}\left(\frac{\overline{R}_0}{R}\right)^6 \tag{15.6}$$

Here $\overline{R}_0$ is defined as the Förster radius for which the orientation factor κ^2 would assume its isotropic average value, $^2/_3$ (Valeur, 2002). For complex systems, the angular dependence is quite commonly disregarded and the following simpler expression employed:

$$w_F = \frac{1}{\tau_{D^*}}\left(\frac{R_0}{R}\right)^6 \tag{15.7}$$

leading to a transfer efficiency Φ_T expressed as

$$\begin{aligned}\Phi_\mathrm{T} &= \frac{1}{1+(R/R_0)^6} = 1 - \frac{\tau_\mathrm{fl}}{\tau_{D^*}}\\ &= 1 - \frac{I_\mathrm{fl}}{I_{D^*}}\end{aligned} \tag{15.8}$$

where, I_fl and I_{D^*} are the intensities of the donor fluorescence with the acceptor present and excluded, respectively, and τ_fl specifically denotes the fluorescence lifetime of the donor

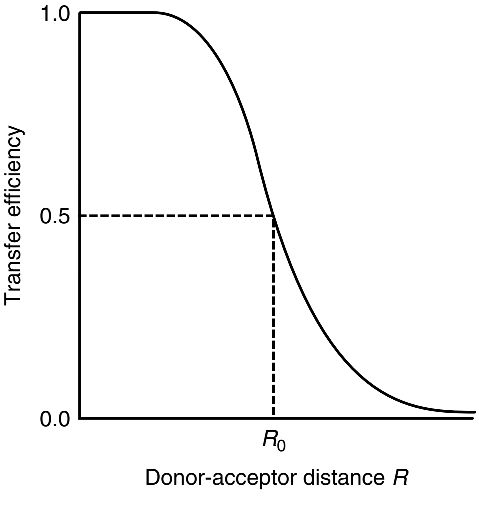

Fig. 15.4 Distance dependence of the transfer efficiency between a pair of chromophores, calculated according to Eq. (15.8).

measured within its RET environment. As graphically depicted in Figure 15.4, a donor–acceptor displacement equal to R_0 corresponds to a transfer efficiency of 50%.

The final equality on the right-hand side of Eq. (15.8), which holds provided decay processes follow single-exponential decay kinetics, provides a formula cast in terms of easily measurable quantities. This is particularly useful because it allows energy transfer efficiencies to be calculated simply on the basis of intensity measurements (for example using a fluorimeter), obviating the separate time-resolved measurements that are otherwise generally necessary for the evaluation of the characteristic decay lifetimes τ_{fl} and τ_{D^*}. When a given electronically excited chromophore is within a distance R_0 of a suitable acceptor, RET will generally be the dominant decay mechanism; conversely, for distances beyond R_0, spontaneous decay (usually fluorescence) will be the primary means of donor deactivation.

15.3.7
Polarization Features

When linearly polarized laser light is used to excite any specific species within a complex disordered solid or liquid system, the probability for excitation of any particular molecule is proportional to $\cos^2\theta$, where θ is the angle between the appropriate excitation transition moment and the electric polarization vector of the input radiation. Consequently, the population of excited molecules has a markedly anisotropic distribution, a phenomenon associated with the term *photoselection*. If radiative decay were to ensue instantaneously, that is, from precisely the initially populated excited level, then the fluorescence would carry the full imprint of that anisotropy and itself exhibit a degree of polarization – the highest value possible. Accounting for the necessary three-dimensional rotational average (Andrews and Thirunamachandran, 1977), it is readily shown that the fluorescence intensity components polarized parallel to and perpendicular to the polarization of the excitation beam, $I_{\parallel}$ and $I_{\perp}$, respectively, would then lie in the ratio 3 : 1. Commonly observed departures from this result thus signify the extent to which the orientation of the emission dipole differs from that of the prior, initial excitation – which may be due to intervening decay, molecular motion, or intermolecular energy transfer.

The two most widely used quantitative expressions of polarization retention are the *fluorescence anisotropy*, r, or the *degree of polarization*, P. Both convey the same information; they are defined and related as follows:

$$r = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + 2I_{\perp}}, \quad P = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}} \quad \Rightarrow$$

$$r = \frac{2P}{3 - P} \tag{15.9}$$

The denominator of the expression for r designates the net fluorescence intensity. In a specific situation where the donor and acceptor have transition dipole moments oriented in parallel, then $r = 0.4$ and $P = 0.5$.

A key molecular factor for determining any loss in polarization is the angle θ between the directions of the absorption and emission transition dipole moments. In terms of this parameter and its influence on the measured fluorescence anisotropy, the case where internal decay intervenes between excitation and fluorescence decay within a single molecule is not different from that of a donor–acceptor pair where the absorption and emission processes are spatially separated – provided the donor and acceptor in the latter case have a fixed mutual orientation (the orientation of the pair being random). The following result, derived by (Levshin, 1925) and Perrin (1929) can be applied in both situations:

$$P = \frac{3\cos^2\theta - 1}{3 + \cos^2\theta} \tag{15.10}$$

In the case of a donor–acceptor pair, θ has to be interpreted as the angle θ_T shown in Figure 15.3. Equation (15.10) thus allows direct calculation of this microscopic parameter, through measurement of the macroscopic quantity P. Moreover, when P proves to exhibit a time-dependent decay, a study of the kinetics provides information on the extent of rotational motion intervening between the absorption and the emission events.

Very different behavior is observed for RET systems in which the donor and acceptor are orientationally uncorrelated, that is, where both are independently, randomly oriented. In such cases, there is a very rapid loss of polarization "memory," and it transpires that the associated degree of anisotropy is precisely 1/25, that is, $r = 0.04$ (Agranovich and Galanin, 1982); two or more energy transfer jumps will therefore usually, to all intents and purposes, totally destroy any polarization in any ensuing fluorescence. However, it should be noted that there is a surprising recovery in the anisotropy at distances approaching the transfer wavelength. The effect is sufficiently strong to warrant attention in dilute solution studies.

15.3.8
Diffusion Effects

Till now, only RET between a donor–acceptor pair has been considered. The discussion is now extended to an ensemble of donors D and acceptors A, all units of which are distributed randomly within an n-dimensional volume. For systems in which translational diffusion is extremely slow compared to the rate of energy transfer, the time-dependence of the donor intensity decay at time t, $I_{D^*}(t)$, as obtained by Förster (1959), is given by the following expression:

$$I_{D^*}(t) = I_{D^*}(0) \times \exp\left[-\frac{t}{\tau_{D^*}} - 2\gamma\left(\frac{t}{\tau_{D^*}}\right)^{n/6}\right] \tag{15.11}$$

The most commonly applied form of this expression is when n equals 3, that is, RET in three dimensions. In Eq. (15.11), the parameter γ is explicitly written as

$$\gamma = \frac{2}{3}\pi^{3/2} C_A R_0^3 \tag{15.12}$$

in which C_A is the concentration of acceptors (number per unit volume) and $(4/3)\pi R_0^3 C_A$ represents the average number of acceptor chromophores in a sphere of radius R_0; the orientational factor is again set as $2/3$.

The case where diffusion is comparable to the transfer rate is very complicated, and calculations by Butler and Pilling (1979) have shown that large errors arise on using Förster theory for systems with diffusion coefficients in excess of 10^{-5} cm^2 s^{-1}. To address such systems, a successful approximation was developed by Gösele *et al.* (1975) This approach involves the insertion of a multiplier G within the second term in the exponential of Eq.(15.11). With $n = 3$, the parameter G is given by the following equation:

$$G = \left(\frac{1 + 5.47x + 4.00x^2}{1 + 3.34x} \right)^{3/4} \tag{15.13}$$

in which $x = D(R_0^6/\tau_{D^*})^{-1/3} t^{2/3}$, where D is the mutual diffusion coefficient. In contrast to the Förster theory, the above method provides an excellent approximation – as was fully verified by Butler and Pilling (1979).

15.3.9
Long-Range Transfer

Förster theory is found to be increasingly inaccurate for RET as donor–acceptor distances extend over and beyond 100 Å. It was originally assumed that a "radiative" process accounted for energy transfer over such donor–acceptor separations, and some recent literature on the subject still perpetuate this initial overstatement. Certain sources wrongly treat Förster "radiationless" energy transfer as exact, distinct, and separable from "radiative" energy transfer – the latter signifying successive but independent processes of fluorescence emission by a donor and capture of the ensuing photon by an acceptor.

Although that certainly is the observed character of RET over very long distances – as for example between donor and acceptor components in a dilute solution – it is now known that both "radiative" and Förster transfer are simply the long- and short-range limits of one powerful, all-pervasive mechanism. The latter, determined from quantum electrodynamical calculations, is the outcome of the unified theory of RET (Andrews, 1989). This theory not only embraces Förster and "radiative" energy transfer, but also addresses the intermediate range in which neither of these mechanisms are fully valid. An expression for the total pairwise energy transfer rate, ranging from molecular dimensions up to interstellar distances, is written as

$$w = w_F + w_I + w_{\text{rad}} \tag{15.14}$$

where w_F represents the Förster rate of Eq. (15.2), w_{rad} is the rate of "radiative" energy transfer – explicitly given by the following equation:

$$w_{\text{rad}} = \frac{9\kappa'^2}{8\pi \tau_D R^2} \int F_D(\omega)\sigma_A(\omega)\, d\omega \tag{15.15}$$

and w_I is the intermediate term that is expressed as

$$w_I = \frac{9c^2}{8\pi \tau_D n^2 R^4} (\kappa^2 - 2\kappa\kappa') \times \int F_D(\omega)\sigma_A(\omega) \frac{d\omega}{\omega^2} \tag{15.16}$$

In both Eqs (15.15) and (15.16), the symbol κ' denotes an orientation factor identical to Eq. (15.4) but with the "3" omitted from the second term. In summary, the unified theory of RET contains not only the R^{-6} term of Förster theory and the R^{-2} term denoting the inverse-square law of "radiative" transfer, but also a previously unidentified R^{-4} intermediate term.

15.3.10
Dexter Transfer

Before concluding this section, it is worth observing that other forms of donor–acceptor coupling are also possible, although considerably less relevant to the systems of interest in the following focused account on applications. For example, the transfer of energy between atomic or molecular components with significantly overlapped wavefunctions is usually described in terms of Dexter theory (Dexter, 1953) – where the coupling involves electron exchange and carries an exponential decay with distance, directly reflecting the radial form of overlapping wavefunctions and electron distributions. Unlike Förster transfer, singlet–triplet energy exchange ($^3D^* + ^1A \rightarrow ^1D + ^3A^*$) may also be allowed by the Dexter mechanism. This is because Dexter transfer does not involve transition dipole moments and, thus, is unaffected by the dipole-forbidden character of the transitions $T_1 \rightarrow S_0$ and $S_0 \rightarrow T_1$ within chromophores D and A, respectively. Compared to materials in which the donor and acceptor orbitals do not spatially overlap, such systems are of less use for either device or analytical applications. This is largely because the coupled chromophores lose their electronic and optical integrity, the Dexter mechanism being operational only at very short distances (<10 Å).

15.4
Applications of RET to Molecular Biology

The field in which measurements of RET have undoubtedly had the greatest impact is molecular biology. The importance of RET to this subject, especially in application to biological macromolecules, was first realized following the construction of spectroscopic equipment for routine fluorescence measurements (Steinberg, 1971; Stryer, 1978; dos Remedios, Miki and Barden, 1987; Wu and Brand, 1994). Toward the turn of the twenty-first century, RET underwent a period of significant redevelopment as a spectroscopic technique (Selvin, 2000). This resurgence arose mainly due to the advent of new experimentation methods, for example single-pair RET (Ha *et al.*, 1996), and further advances in instrumentation. The key advantage of RET techniques over others is that fluorescence measurements are highly sensitive, being made against a zero background; moreover, the UV/visible signals are relatively easy to detect, they are specific and the required instrumentation is noninvasive.

15.4.1
Spectroscopic Ruler

A major use of RET, based on its strong distance dependence, exploits its capacity to supply accurate spatial information about molecular structures. This derives from the quantitative assessment of the interchromophore separations, based on comparisons between the corresponding RET efficiencies (Hohng, Joo and Ha, 2004; Schuler, 2005; Koushik *et al.*, 2006; Zhang and Allen, 2007). Such a technique is popularly known as a “spectroscopic ruler.” The elucidations of molecular structure by such means usually lack information on the relative *orientations* of the groups involved, and as an expedient the calculations usually ignore the kappa parameter (Eq. (15.4)). The apparent crudeness of this approach becomes more defensible on realizing that, even if it were to introduce a factor of two inaccuracy, the deduced group spacing would still be in error by only 12% (because $2^{1/6} = 1.12$). Refinements to the theory to accommodate the effect of fluctuations in position or orientation of the participant groups introduce considerable complexity, although progress is being made in several areas (dos Remedios and Moens, 1995; Tanaka, 1998; Yu, 2007; Isaksson *et al.*, 2007; Jang, 2007).

15.4.2
Conformational Change

Through identification of motions in macromolecules, that is, the variation in proximity of one chromophore with respect to another, a number of valuable RET applications arise; including the detection of conformational changes and folding in proteins (Schuler, 2005; Talaga *et al.*, 2000; Heyduk, 2002; Schuler and Eaton, 2008) and the inspection of intracellular protein-protein (Sekar and Periasamy, 2003; Zal and Gascoigne, 2004; Parsons, Vojnovic and Ameer-Beg, 2004; You *et al.*, 2006) and protein-DNA (Hillisch, Lorenz and Diekmann, 2001; Cremazy *et al.*, 2005) interactions (see, for example, Figures 15.5 and 15.6). These and other such processes can be registered by selectively exciting one chromophore using laser light, and monitoring either the decrease in fluorescence from that chromophore or the rise in the generally longer-wavelength fluorescence from the other chromophore as it adopts the role of acceptor. The judicious use of optical dichroic filters can make this RET technique perfectly straightforward (see Figure 15.1). In cases where the two material components of interest do not display suitably overlapped absorption and fluorescence features in an optically accessible wavelength range, molecular tagging with site-specific "extrinsic" (i.e., artificially attached) chromophores can solve the problem. Located at a molecular site of interest, and being selected on the basis of a significant spectral overlap with the counterpart component, such tags can act either in the capacity of donor or acceptor. Lanthanide ions, with their characteristically prominent and linelike absorption features, prove to be particularly valuable in this connection (Selvin, 2002). The semiconductor nanocrystals known as *quantum dots* are also useful in this regard. These crystalline nanoparticles offer several unique traits, including size- and composition-tunable emission from visible to infrared wavelengths, the possibility of a single light source simultaneously exciting different-sized dots, large absorption coefficients across a wide spectral range, and very high level of

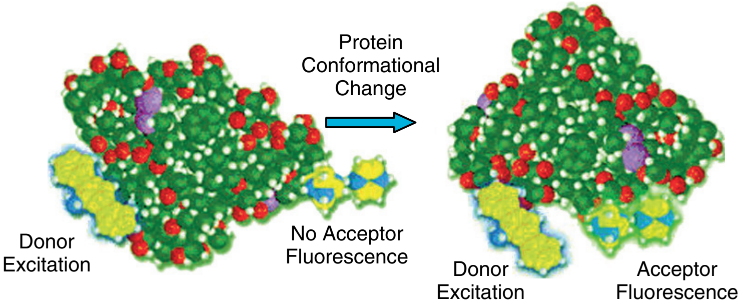

Fig. 15.5 Graphical depiction of RET detection of protein conformational change. (Adapted from Olympus Corporation website.)

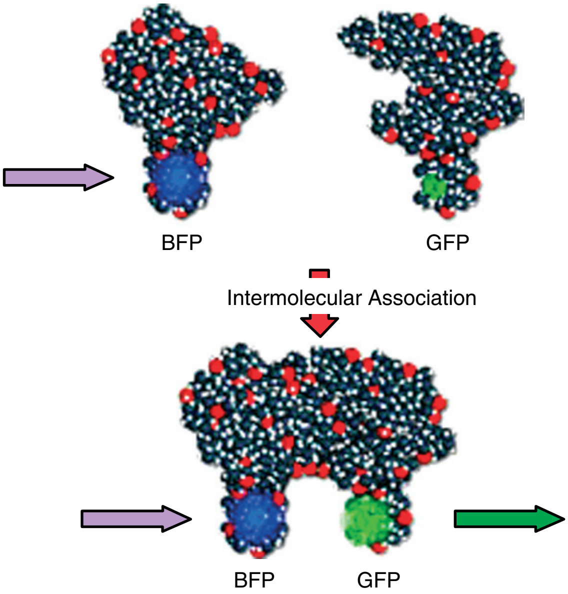

Fig. 15.6 Graphical depiction of RET detection of *in vivo* protein–protein interactions. Purple arrow denotes input laser of wavelength 380 nm and green arrow indicates protein emission at 510 nm. BFP and GFP are acronyms for blue and green fluorescent proteins, respectively. (Adapted from Olympus Corporation website.)

photostability (Chan *et al.*, 2002; Clapp, Medintz and Mattoussi, 2006).

15.4.3
Intensity-Based Imaging

Over the past two decades, there has been burgeoning interest in microscopy based on RET (Clegg, 1996; Day, Periasamy and Schaufele, 2001; Wouters, Verveer and Bastiaens, 2001; Hoppe, Christensen and Swanson, 2002; Jares-Erijman and Jovin, 2003), typical instrumentation for which is illustrated in Figure 15.7. There are three specific types of RET method routinely used in the production of biological images. The principles of sensitized-emission RET have already been described (Figure 15.1). For microscopy purposes, this method is fairly inaccurate; no RET donor–acceptor pair is ideal, that is, there will almost always be some overlap between the donor and the acceptor absorbance bands and also the donor and the acceptor emission spectra. Therefore, filters that completely separate these kinds of spectrum are difficult to design. Various calculational algorithms (Gordon *et al.*, 1998; Xia and Liu, 2001; van Rheenen, Langeslag and Jalink, 2004) have been proposed to compensate for this problem, although the methods are complex and no single procedure has received universal acceptance.

A widely used alternative, experimental approach (Karpova *et al.*, 2003; van

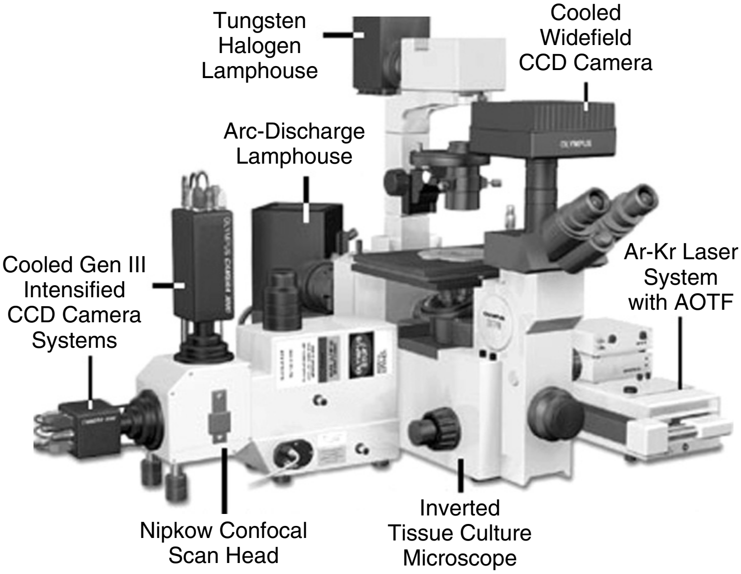
Tungsten Halogen Lamphouse
Cooled Widefield CCD Camera
Arc-Discharge Lamphouse
Ar-Kr Laser System with AOTF
Cooled Gen III Intensified CCD Camera Systems
Nipkow Confocal Scan Head
Inverted Tissue Culture Microscope

Fig. 15.7 Typical commercial setup of a microscope based on RET. (Adapted from Olympus Corporation website.)

Munster *et al.*, 2005) involves deliberately photobleaching the acceptor, the result of which is complete exclusion of RET. In this method, the donor emission is analyzed before and after the acceptor is bleached by the input of an intense laser beam (at a suitable wavelength). The difference between the donor intensities, with and without the laser input, enables the determination of the transfer efficiency by employing Eq. (15.8). Here, account is taken of spectral bleed-through between the two absorbance bands, and equally between the two emission bands. Signal contamination is still not entirely eliminated, due to a small amount of back-transfer through donor excitation by acceptor emission. Often the main disadvantage in prolonged illumination of the acceptor is the possibility of damage to the sample. Therefore, in practice, photobleaching is seldom appropriate for *in vivo* studies.

15.4.4
Lifetime-Based Imaging

Fluorescence need not be characterized from excitation and emission spectra alone; highly significant information can also be secured from lifetime measurements. Thus, when suitable time-resolved instrumentation is available, the determination of decay kinetics (usually on the nanosecond timescale) enables analysis through resonance energy transfer-based fluorescence lifetime imaging microscopy (FRET-FLIM) (Bastiaens and Squire, 1999; Duncan *et al.*, 2004; Wallrabe and Periasamy, 2005; Peter *et al.*, 2005). In this method, spectral bleed-through is no longer an issue because measurements are made only for the determination of donor lifetimes; back-transfer is usually extremely low, within the noise level. The presence of the acceptor within the local

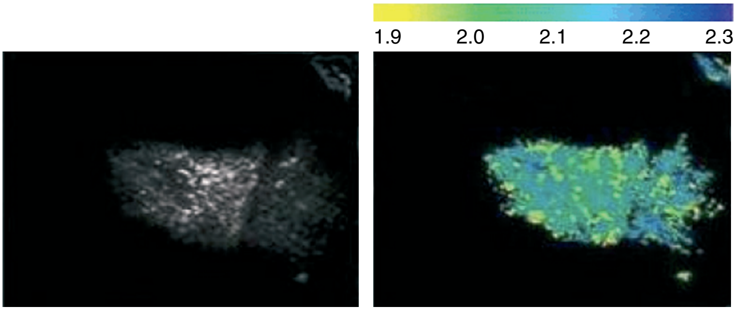

(a) (b)

Fig. 15.8 MDA-MB-231 cancer cell images recorded with argon laser two-photon excitation, RET microscope based on; (a) intensity and (b) fluorescence lifetime (nanosecond). In the latter image, areas of locally reduced lifetime signify clustered intracellular vesicles. (Adapted from Peter *et al.* (2005).)

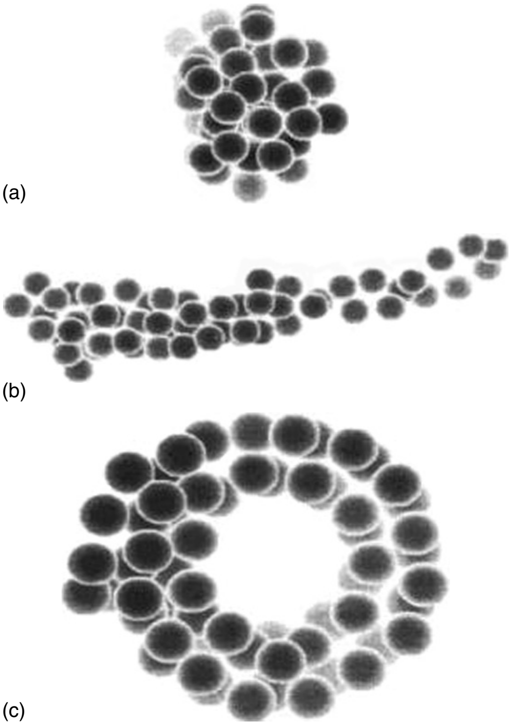

Fig. 15.9 Snapshots of various morphological constructions of a homopolymer chain. The structures (a) – (c) are spherical, rod, and toroidal, respectively. (Adapted from Srinivas and Bagchi (2002).)

environment of the donor influences the fluorescence lifetime of the donor. By measuring the donor lifetime in the presence and the absence of the acceptor, one can accurately calculate the transfer efficiency by use of Eq.(15.8). Drawbacks to FRET-FLIM are the technical challenges the technique presents and the expense of the equipment. Nonetheless, in optical systems that are equipped to provide both intensity and lifetime measurements, a comparison of the two types of image affords a particularly rich source of information, as illustrated by the cancer cell images of Figure 15.8.

15.4.5
Other Applications

Beyond the realm of molecular biology, RET has value in a number of more specifically chemical applications. Two prominent examples are to be found in the fields of synthetic macromolecules and chemical sensors. In polymer science, building on the pioneering principles of Morawetz (1988), RET is now used to determine morphological information on polymer interfaces. Such studies have, for instance, enabled the quantitative characterization of interfacial thickness in polymers of various structures (Farinha and Martinho, 2008). Moreover, RET has been utilized in the study of polymer conformational dynamics. One especially interesting application is the effective differentiation between various collapsed and/or ordered homopolymer chain conformations (spherical, rod and toroidal; as depicted in Figure 15.9) through the associated distribution of transfer effiencies (Srinivas and Bagchi, 2002; Saini, Singh and Bagchi, 2006).

The fabrication of RET-based, analyte-specific sensors has enabled the detection of a variety of species, including dimers of functionalized calixarenes in organic solutions (Castellano *et al.*, 2000), copper(II) in aqueous solution (Cano-Raya, Fernández-Ramos and Capitán-Vallvey, 2006), hydrogen peroxide (Albers, Okreglak and Chang, 2006), phosgene (Zhang and Rudkevich, 2007) and many others. These chemical sensors usually work on the principle of a donor–acceptor system designed such that the presence of the analyte causes the acceptor chromophore to move within closer proximity to a donor, enabling the implementation of an RET process that is not observed in the analyte's absence. Therefore, on irradiation of the system with the relevant chemical present, a strong emission from the acceptor signals the presence of the analyte.

Acknowledgments

We gratefully acknowledge helpful comments from Professor Steve Meech. Research on the theory of RET, at the University of East Anglia, is currently supported by the Leverhulme Trust.

References

Agranovich, V.M. and Galanin, M.D. (1982) *Electronic Excitation Energy Transfer in Condensed Matter*, Elsevier/North-Holland, Amsterdam, The Netherlands.

Albers, A.E., Okreglak, V.S., and Chang, C.J. (2006) A FRET-based approach to ratiometric fluorescence detection of hydrogen peroxide. *J. Am. Chem. Soc.*, **128**, 9640–9641.

Andrews, D.L. (1989) A unified theory of radiative and radiationless molecular-energy transfer. *Chem. Phys.*, **135**, 195–201.

Andrews, D.L. and Rodríguez, J. (2007) Resonance energy transfer: spectral overlap, efficiency and direction. *J. Chem. Phys.*, **127**, 084509.

Andrews, D.L. and Thirunamachandran, T. (1977) On three-dimensional rotational averages. *J. Chem. Phys.*, **67**, 5026–5033.

Bastiaens, P.I.H. and Squire, A. (1999) Fluorescence lifetime imaging microscopy:

spatial resolution of biochemical processes in the cell. *Trends Cell Biol.*, **9**, 48–52.

Berberan-Santos, M.N. (2001) Pioneering contributions of Jean and Francis Perrin to molecular luminescence, in *New Trends in Fluorescence Spectroscopy. Applications to Chemical and Life Sciences* (eds B. Valeur and J.-C. Brochon), Chapter 2, Springer, Berlin.

Butler, P.R. and Pilling, M.J. (1979) The breakdown of Förster kinetics in low viscosity liquids. An approximate analytical form for the time-dependent rate constant. *Chem. Phys.*, **41**, 239–243.

Clegg, R.M. (1996) Fluorescence resonance energy transfer, in *Fluorescence Imaging Spectroscopy and Microscopy* (eds X.F. Wang and B. Herman), Chapter 7, John Wiley & Sons, Ltd, New York.

Clegg, R.M. (2006) The history of FRET: From conception through the labors of birth, in *Reviews in Fluorescence* (eds C.D. Geddes and J.R. Lakowicz), Vol. 3, Chapter 1, Springer, New York.

Cano-Raya, C., Fernández-Ramos, M.D., and Capitán-Vallvey, L.F. (2006) Fluorescence resonance energy transfer disposable sensor for copper(II). *Anal. Chim. Acta*, **555**, 299–307.

Cario, G. (1922) Über Entstehung wahrer Lichtabsorption un scheinbare Koppelung von Quantensprüngen. *Z. Phys.*, **10**, 185–199.

Cario, G. and Franck, J. (1922) Über Zerlegugen von Wasserstoffmolekülen durch angeregte Quecksilberatome. *Z. Phys.*, **11**, 161–166.

Castellano, R.K., Craig, S.L., Nuckolls, C., and Rebek, J. (2000) Detection and mechanistic studies of multicomponent assembly by fluorescence resonance energy transfer. *J. Am. Chem. Soc.*, **122**, 7876–7882.

Chan, W.C.W., Maxwell, D.J., Gao, X., Bailey, R.E., Han, M., and Nie, S. (2002) Luminescent quantum dots for multiplexed biological detection and imaging. *Curr. Opin. Biotechnol.*, **13**, 40–46.

Clapp, A.R., Medintz, I.L., and Mattoussi, H. (2006) Förster resonance energy transfer investigations using quantum-dot fluorophores. *Chem. Phys. Chem.*, **7**, 47–57.

Cremazy, F.G.E., Manders, E.M.M., Bastiaens, P.I.H., Kramer, G., Hager, G.L., van Munster, E.B., Verschure, P.J., Gadella, T.W.J., and van Driel, R. (2005) Imaging in situ protein–DNA interactions in the cell nucleus using FRET–FLIM. *Exp. Cell Res.*, **309**, 390–396.

Day, R.N., Periasamy, A., and Schaufele, F. (2001) Fluorescence resonance energy transfer microscopy of localized protein interactions in the living cell nucleus. *Methods*, **25**, 4–18.

Demidov, A.A. and Andrews, D.L. (2001) in *Encyclopedia of Chemical Physics and Physical Chemistry* (eds J.H. Moore and N.D. Spencer), Vol. 3, Institute of Physics, Bristol, pp. 2701–2715.

Dexter, D.L. (1953) A theory of sensitized luminescence in solids. *J. Chem. Phys.*, **21**, 836–850.

Duncan, R.R., Bergmann, A., Cousin, M.A., Apps, D.K., and Shipston, M.J. (2004) Multi-dimensional time-correlated single photon counting (TCSPC) fluorescence lifetime imaging microscopy (FLIM) to detect FRET in cells. *J. Microsc.*, **215**, 1–12.

Farinha, J.P.S. and Martinho, J.M.G. (2008) Resonance energy transfer in polymer interfaces. *Springer Ser. Fluoresc.*, **4**, 215–255.

Förster, T. (1946) Energiewanderung und fluoreszenz. *Naturwissenschaften*, **33**, 166–175.

Förster, T. (1948) Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Phys.*, **2**, 55–75.

Förster, T. (1959) 10th Spiers Memorial Lecture. Transfer mechanisms of electronic excitation. *Discuss. Faraday Soc.*, **27**, 7–17.

Franck, J. (1922) Einige aus der Theorie von Klein und Rosseland zu ziehende Folgerungen über Fluorscence, photochemische Prozesse und die Electronenemission glühender Körper. *Z. Phys.*, **9**, 259–266.

Galli, C., Wynne, K., Lecours, S.M., Therien, M.J., and Hochstrasser, R.M. (1993) Direct measurement of electronic dephasing using anisotropy. *Chem. Phys. Lett.*, **206**, 493–499.

Gaviola, E. and Pringsheim, P. (1924) Über den einfluß der konzentration auf die polarisation der fluoreszenz von farbstofflösungen. *Z. Phys.*, **24**, 24–36.

Gordon, G.W., Berry, G., Liang, X.H., Levine, B., and Herman, B. (1998) Quantitative fluorescence resonance energy transfer measurements using fluorescence microscopy. *Biophys. J.*, **74**, 2702–2713.

Gösele, U., Hauser, M., Klein, U.K.A., and Frey, R. (1975) Diffusion and long-range energy transfer. *Chem. Phys. Lett.*, **34**, 519–522.

Ha, T., Enderle, T., Ogletree, D.F., Chemla, D.S., Selvin, P.R., and Weiss, S. (1996) Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl. Acad. Sci. USA*, **93**, 6264–6268.

Heyduk, T. (2002) Measuring protein conformational changes by FRET/LRET. *Curr. Opin. Biotechnol.*, **13**, 292–296.

Hillisch, A., Lorenz, M., and Diekmann, S. (2001) Recent advances in FRET: distance determination in protein–DNA complexes. *Curr. Opin. Struct. Biol.*, **11**, 201–207.

Hohng, S., Joo, C., and Ha, T. (2004) Single-molecule three-color FRET. *Biophys. J.*, **87**, 1328–1337.

Hoppe, A., Christensen, K., and Swanson, J.A. (2002) Fluorescence resonance energy transfer-based stoichiometry in living cells. *Biophy. J.*, **83**, 3652–3664.

Isaksson, M., Norlin, N., Westlund, P.-O., and Johansson, L.B.-Å. (2007) On the quantitative molecular analysis of electronic energy transfer within donor-acceptor pairs. *Phys. Chem. Chem. Phys.*, **9**, 1941–1951.

Jang, S. (2007) Generalization of the Förster resonance energy transfer theory for quantum mechanical modulation of the donor-acceptor coupling. *J. Chem. Phys.*, **127**, 174710.

Jares-Erijman, E.A. and Jovin, T.M. (2003) FRET imaging. *Nat. Biotechnol.*, **21**, 1387–1395.

Kallman, H. and London, F. (1928) Über quantenmechanische Energieübertragung zwischen atomaren Systemen. *Z. Phys. Chem. B*, **2**, 207–243.

Karpova, T.S., Baumann, C.T., He, L., Wu, X., Grammer, A., Lipsky, P., Hager, G.L., and McNally, J.G. (2003) Fluorescence resonance energy transfer from cyan to yellow fluorescent protein detected by acceptor photobleaching using confocal microscopy and a single laser. *J. Microsc.*, **209**, 56–70.

Koushik, S.V., Chen, H., Thaler, C., Puhl, H.L. III, and Vogel, S.S. (2006) Cerulean, venus and venus$_{\mathrm{Y67C}}$ FRET reference standards. *Biophys. J.*, **91**, L99–L101.

Lakowicz, J.R. (1999) *Principles of Fluorescence Spectroscopy*, 2nd edn, Chapter 10, Kluwer Academic, New York, USA.

Latt, S.A., Cheung, H.T., and Blout, E.R. (1965) Energy transfer. A system with relatively fixed donor-acceptor separation. *J. Am. Chem. Soc.*, **87**, 995–1003.

Levshin, W.L. (1925) Polarisierte Fluoreszenz und Phosphoreszenz der Farbstofflosungen. IV. *Z. Phys.*, **32**, 307–326.

van der Meer, B.W. (1999) in *Resonance Energy Transfer* (eds D.L., Andrews and A.A., Demidov), John Wiley & Sons, Ltd, New York, pp. 151–172.

Morawetz, H. (1988) Studies of synthetic polymers by nonradiative energy transfer. *Science*, **240**, 172–176.

van Munster, E.B., Kremers, G.J., Adjobo-Hermans, M.J.W., and Gadella, T.W.J. (2005) Fluorescence resonance energy transfer (FRET) measurement by gradual acceptor photobleaching. *J. Microsc.*, **218**, 253–262.

Parsons, M., Vojnovic, B., and Ameer-Beg, S. (2004) Imaging protein–protein interactions in cell motility using fluorescence resonance energy transfer (FRET). *Biochem. Soc. Trans.*, **32**, 431–433.

Perrin, J. (1927) Fluorescence et induction moléculaire par résonance. *C. R. Acad. Sci.*, **184**, 1097–1100.

Perrin, F. (1929) La fluorescence des solutions. *Ann. Phys.*, **12**, 169–275.

Perrin, F. (1932) Théorie quantique des transferts d'activation entre molecules de même espéce. Cas des solutions fluorescents. *Ann. Phys.*, **17**, 283–314.

Peter, M., Ameer-Beg, S.M., Hughes, M.K.Y., Keppler, M.D., Prag, S., Marsh, M., Vojnovic, B., and Ng, T. (2005) Multiphoton-FLIM quantification of the EGFP-mRFP1 FRET pair for localization of membrane receptor-kinase interactions. *Biophys. J.*, **88**, 1224–1237.

dos Remedios, C.G., Miki, M., and Barden, J.A. (1987) Fluorescence resonance energy transfer measurements of distances in actin and myosin. A critical evaluation. *J. Muscle Res. Cell Motil.*, **8**, 97–117.

dos Remedios, C.G. and Moens, P.D.J. (1995) Fluorescence resonance energy transfer spectroscopy is a reliable 'ruler' for measuring structural changes in proteins – dispelling the problem of the unknown orientation factor. *J. Struct. Biol.*, **115**, 175–185.

van Rheenen, J., Langeslag, M., and Jalink, K. (2004) Correcting confocal acquisition to optimize imaging of fluorescence resonance energy transfer by sensitized emission. *Biophys. J.*, **86**, 2517–2529.

Saini, S., Singh, H., and Bagchi, B. (2006) Fluorescence resonance energy transfer (FRET) in chemistry and biology: Non-Förster distance dependence of the FRET rate. *J. Chem. Sci.*, **118**, 23–35.

Schuler, B. (2005) Single-molecule fluorescence spectroscopy of protein folding. *ChemPhysChem*, **6**, 1206–1220.

Schuler, B. and Eaton, W.A. (2008) Protein folding studied by single-molecule FRET. *Curr. Opin. Struct. Biol.*, **18**, 16–26.

Sekar, R.B. and Periasamy, A. (2003) Fluorescence resonance energy transfer (FRET) microscopy imaging of live cell protein localizations. *J. Cell Biol.*, **160**, 629–633.

Selvin, P.R. (2000) The renaissance of fluorescence resonance energy transfer. *Nat. Struct. Bio.*, **7**, 730–734.

Selvin, P.R. (2002) Principles and biophysical applications of lanthanide-based probes. *Annu. Rev. Biophys. Biomol. Struc.*, **31**, 275–302.

Srinivas, G. and Bagchi, B. (2002) Detection of collapsed and ordered polymer structures by fluorescence resonance energy transfer in stiff homopolymers: Bimodality in the reaction efficiency distribution. *J. Chem. Phys*, **116**, 837–844.

Steinberg, I.Z. (1971) Long-range nonradiative transfer of electronic excitation energy in proteins and polypeptides. *Annu. Rev. Biochem.*, **40**, 83–114.

Stryer, L. (1978) Fluorescence energy transfer as a spectroscopic ruler. *Ann. Rev. Biochem.*, **47**, 819–846.

Stryer, L. and Haugland, R.P. (1967) Energy transfer: a spectroscopic ruler. *Proc. Natl. Acad. Sci.*, **58**, 719–726.

Talaga, D.S., Lau, W.L., Roder, H., Tang, J., Jia, Y.W., DeGrado, W.F., and Hochstrasser, R.M. (2000) Dynamics and folding of single two-stranded coiled-coil peptides studied by fluorescent energy transfer confocal microscopy. *Proc. Natl. Acad. Sci. USA*, **97**, 13021–13026.

Tanaka, F. (1998) Theory of time-resolved fluorescence under the interaction of energy transfer in a bichromophoric system: effect of internal rotations of energy donor and acceptor. *J. Chem. Phys.*, **109**, 1084–1092.

Valeur, B. (2002) *Molecular Fluorescence: Principles and Applications*, Chapter 9, Wiley-VCH Verlag GmbH, Weinheim, Germany.

Wallrabe, H. and Periasamy, A. (2005) Imaging protein molecules using FRET and FLIM microscopy. *Curr. Opin. Biotechnol.*, **16**, 19–27.

Wouters, F.S., Verveer, P.J., and Bastiaens, P.I.H. (2001) Imaging biochemistry inside cells. *Trends Cell Biol.*, **11**, 203–211.

Wu, P. and Brand, L. (1994) Resonance energy transfer: methods and applications. *Anal. Biochem.*, **218**, 1–13.

Xia, Z. and Liu, Y. (2001) Reliable and global measurement of fluorescence resonance energy transfer using fluorescence microscopes. *Biophys. J*, **81**, 2395–2402.

You, X., Nguyen, A.W., Jabaiah, A., Sheff, M.A., Thorn, K.S., and Daugherty, P.S. (2006) Intracellular protein interaction mapping with FRET hybrids. *Proc. Natl. Acad. Sci. USA*, **103**, 18458–18463.

Yu, Z.G. (2007) Fluorescent resonant energy transfer: correlated fluctuations of donor and acceptor. *J. Chem. Phys.*, **127**, 221101.

Zal, T. and Gascoigne, N.R.J. (2004) Using live FRET imaging to reveal early protein–protein interactions during T cell activation. *Curr. Opin. Immunol.*, **16**, 418–427.

Zhang, J. and Allen, M.D. (2007) FRET-based biosensors for protein kinases: Illuminating the kinome. *Mol. Biosyst.*, **3**, 759–765.

Zhang, H. and Rudkevich, D.M. (2007) A FRET approach to phosgene detection. *Chem. Commun.*, 1238–1239.

Further Reading

Andrews, D.L. (2008) Mechanistic principles and applications of resonance energy transfer. *Canad. J. Chem.*, **86**, 855–870.

Andrews, D.L. and Demidov, A.A. (eds) (1999) *Resonance Energy Transfer*, John Wiley & Sons, Ltd, New York.

van der Meer, B.W., Coker, G., and Chen, S.-Y. (1994) *Resonance Energy Transfer Theory and Data*, Wiley-VCH Verlag GmbH, New York.

Periasamy, A. (2001) *Methods in Cellular Imaging*, Oxford University Press, New York.

Periasamy, A. and Day, R.N. (2005) *Molecular Imaging: FRET Microscopy and Spectroscopy*, Oxford University Press, New York.

Review of FRET microscopy on Olympus Corporation's web site: http://www.olympusfluoview.com/applications/fretintro.html (accessed July 2008).

Sapsford, K.E., Berti, L., and Medintz, I.L. (2006) Materials for fluorescence resonance energy transfer analysis: beyond traditional donor–acceptor combinations. *Angew. Chem. Int. Ed.*, **45**, 4562–4588.

Scholes, G.D. (2003) Long-range resonance energy transfer in molecular systems. *Annu. Rev. Phys. Chem.*, **54**, 57–87.

16
Optical Spectroscopy of Biological Materials

Valery V. Tuchin

16.1 Introduction 557
16.2 Tissue Optics 557
16.2.1 Tissue Structure 557
16.2.1.1 Tissue Optical Models 557
16.2.1.2 Sizing of Tissue and Cell Compartments 558
16.2.2 Light Interaction with Tissues 559
16.2.2.1 Light Absorption and Scattering 559
16.2.2.2 Theoretical Description for CW 561
16.2.2.3 Index of Refraction of Tissue Compartments 565
16.2.2.4 Short Pulse Interaction with Biomaterials 566
16.3 Measurements of Optical Properties of Tissues and Blood 568
16.3.1 Introduction 568
16.3.2 Integrating Sphere Technique 570
16.3.3 Multiflux Approach 570
16.3.4 The Inverse Adding–Doubling Method 571
16.3.5 Inverse Monte Carlo Method 571
16.3.6 Spatially Resolved Measurements 573
16.3.7 Optical Coherence Tomography 574
16.3.8 Measurement of the Scattering Phase Function 574
16.3.9 Human Tissue Optical Properties Analysis 575
16.3.10 Optical Properties of Blood 613
16.3.11 Refractive Index Measurements 613
16.3.12 Optical Projection Tomography 616
Glossary 617
References 622
Further Reading 625

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

16.1
Introduction

Optics of biological materials determines diagnostic and therapeutic facilities of conventional and laser light. Principle noninvasiveness of low-intensity visible and infrared light, its substantial permeation into biological tissues, and dependence of its propagating properties on tissue morphology and functioning allow for providing structural and functional imaging of many body organs without any risk and limitations. The knowledge of optical properties of tissues and blood is also important for prediction of light–tissue interaction with various follow-up therapeutic effects. Light dosimetry is the key problem in any kind of phototreatment, including cancer treatment using photodynamic or photothermal technologies, as well as laser precise surgery.

In this chapter, optics of biological tissues and blood is briefly discussed. Methods and algorithms for solving the inverse problem of finding tissue and blood optical parameters, such as absorption and scattering coefficients, anisotropy factor, and refractive index are presented. Measuring techniques, such as integrating sphere (IS), spatially resolved, time-resolved, and angular-resolved, optical coherence tomography (OCT), as well as inverse methods, such as inverse adding–doubling (IAD) and inverse Monte Carlo (IMC) are reviewed. Data on optical properties of human tissue and blood measured *in vitro, ex vivo,* and *in vivo* are presented.

16.2
Tissue Optics

16.2.1
Tissue Structure

16.2.1.1 Tissue Optical Models

Biological tissues are optically inhomogeneous and absorbing media. Their average refractive index is higher than that of air, thus partial reflection of the radiation at the tissue/air interface (Fresnel reflection) takes place, while the remaining part of radiation penetrates the tissue. Multiple scattering and absorption are responsible for light beams broadening and eventual decay as they travel through a tissue. The bulk scattering causes a large portion of radiation to be scattered in the backward direction. Light propagation within a tissue depends on the scattering and absorption properties of its compartments:

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

cells, cell organelles, and various fiber structures (Tuchin, 2002; Vo-Dinh, 2003; Tuchin, 2007). The size, shape, and density of these structures, their refractive indices play important roles in the propagation of light in tissues.

The great diversity and structural complexity of tissues require the development of adequate optical models accounting for the scattering and absorption properties of tissues. Two major approaches are currently used: (i) tissue modeled as a medium with a continuous random spatial distribution of optical parameters and (ii) tissue considered as a discrete ensemble of scatterers. The choice of the approach is defined by both the structural specificity of the tissue under study and the kind of light-scattering characteristics that are to be described.

Many tissues can be represented as a random continuum of the inhomogeneities of the refractive index with a varying spatial scale. This approach is applicable for tissues with no pronounced boundaries between elements that feature significant heterogeneity. The process of scattering by these structures may be described under certain conditions using the model of a phase screen (Goodman, 1985; Rytov, Kravtsov and Tatarskii, 1989; Tuchin, 2007).

In accordance with the second approach, biological media are often modeled as ensembles of particles, since many cells and microorganisms are close in shape to spheres, ellipsoids, or rods. A system of noninteracting homogeneous spherical particles is the simplest tissue model. Mie theory rigorously describes the diffraction of light by a spherical particle (Bohren and Huffman, 1983; Born and Wolf, 1999). The further development of this model involves accounting for more complex structure of the spherical particles, namely, the multilayered spheres and the spheres with radial nonhomogeneity, anisotropy, and optical activity (Mishchenko, Hovenier and Travis, 2000; Mishchenko, Travis and Lacis, 2002). Since a connective tissue consists of fiber structures, a system of long cylinders is the most appropriate model for it. Muscular tissue, skin dermis, cerebral membrane (*dura mater*), eye cornea, and sclera belong to this type of tissue formed essentially by collagen fibrils. The solution of the problem of light diffraction in a single homogeneous or multilayered cylinder is also well described (Bohren and Huffman, 1983). Attempts to describe light scattering by a system of interacting particles as a more realistic tissue model (quasi-ordered spherical particles or cylinders) have also been made (see Tuchin, 2007).

16.2.1.2 **Sizing of Tissue and Cell Compartments**

The sizes of cells and tissue structure elements vary in size from a few tenths of nanometers to hundreds of micrometers (Dyson, 1974; Hogan, Alvardo and Weddel, 1971; Silver, 1987; Kessel, 1998; Tuchin, 2002; Vo-Dinh, 2003; Tuchin, Wang and Zimnyakov, 2006; Tuchin, 2007). Blood cells (erythrocytes, leukocytes, and platelets) exhibit the following parameters. A normal erythrocyte in plasma has the shape of a concave–concave disc with a diameter varying from 7.1 to 9.2 μm, a thickness of 0.9–1.2 μm in the center, 1.7–2.4 μm on the periphery, and a volume of $\sim$90 μm^3. Leukocytes are formed like spheres with a diameter of 8–22 μm. Platelets in the blood stream are biconvex disklike particles with diameters ranging from 2 to 4 μm. Normally, blood has about 10 times as many erythrocytes as platelets and about 30 times as many platelets as leukocytes.

Most other mammalian cells have diameters in the range of 5–75 μm. In the epidermal layer, the cells are large (with an average cross-sectional area of about $80\,\mu m^2$) and quite uniform in size. Fat cells, each containing a single lipid droplet that nearly fills the entire cell and therefore results in eccentric placement of the cytoplasm and nucleus, have a wide range of diameters from a few microns to 50–75 μm. Fat cells may reach a diameter of 100–200 μm in pathological cases.

There are a wide variety of structures within cells that determine tissue light scattering. Cell nuclei are on the order of 5–10 μm in diameter; mitochondria, lysosomes, and peroxisomes have dimensions of 1–2 μm; ribosomes are on the order of 20 nm in diameter; and structures within various organelles can have dimensions up to a few hundred nanometers. Usually, the scatterers in cells are not spherical. The models of prolate ellipsoids with a ratio of the ellipsoid axes between 2 and 10 are more typical.

The hollow organs of the body are lined with a thin, highly cellular surface layer of epithelial tissue, which is supported by underlying, relatively acellular connective tissue. In healthy tissues, the epithelium often consists of a single, well-organized layer of cells with en-face diameter of 10–20 μm and height of 25 μm. In dysplastic epithelium, cells proliferate and their nuclei enlarge and appear darker (hyperchromatic) when stained. Enlarged nuclei are primary indicators of cancer, dysplasia, and cell regeneration in most human tissues.

In fibrous tissues or tissues containing fiber layers (cornea, sclera, *dura mater*, muscle, myocardium, tendon, cartilage, vessel wall, retinal nerve fiber layer, etc.) and composed mostly of microfibrils and/or microtubules, typical diameters of the cylindrical structural elements are 10–400 nm. Their length is in a range from 10–25 μm to a few millimeters.

For some tissues, the size distribution of the scattering particles may be essentially monodispersive and for others it may be quite broad. Examples for the two are transparent eye cornea stroma, which has a sharply monodispersive distribution, and turbid eye sclera, which has a rather broad distribution of collagen fiber diameters.

16.2.2
Light Interaction with Tissues

16.2.2.1 Light Absorption and Scattering

Absorbed light is converted to heat or radiated in the form of fluorescence, it is also consumed in photobiochemical reactions. The absorption spectrum depends on the type of predominant absorption centers and water content of tissues. Absolute values of absorption coefficients for typical tissues lie in the range of 10^{-2}–$10^{4}\ \text{cm}^{-1}$ (Anderson and Parrish, 1982; Chance, 1989; Duck, 1990; Frank and Kessler, 1992; Henderson and Dougherty, 1992; Welch and van Gemert, 1992; Niemz, 1996; Tuchin, 2002; Berlien and Mueller, 2003; Vo-Dinh, 2003; Tuchin, 2007). In the ultraviolet (UV) and infrared (IR) ($\lambda \geq 2000$ nm) spectral regions, light is readily absorbed, which accounts for the small contribution of scattering and inability of radiation to penetrate deep into tissues (only through one or two cell layers). Short-wave visible light penetrates typical tissues as deep as 0.5–2.5 mm, whereupon it undergoes an *e*-fold decrease of intensity. In this case, both scattering and absorption occur, with 15–40% of the incident radiation being reflected. In the wavelength range of 600–1600 nm, scattering prevails over absorption, and light penetrates to a depth of 8–10 mm. Simultaneously, the intensity of the reflected radiation increases to 35–70% of the total incident light (due to backscattering).

Absorption spectrum of biological material is expressed usually in terms of the wavelength dependence of absorption coefficient. Such spectra for some tissues and tissue components: water (75%), aorta, skin, epidermis, melanosome, and whole blood are presented in Figure 16.1. The often-used lasers and their wavelengths are presented on the crossings of water-absorption curve, the corresponding laser wavelengths showing what absorption coefficient is expected on absorption in water. Since water is the major component of any soft tissue, presented absorption coefficients will be only slightly corrected for real tissues. As shown in Figure 16.1, for bloodless skin in the visible range, absorption is equal to water (75%), since dermis as the main skin component that is well supplied by water defines absorption coefficient of whole skin in this range.

Light interaction with multilayer and multicomponent tissues, such as skin

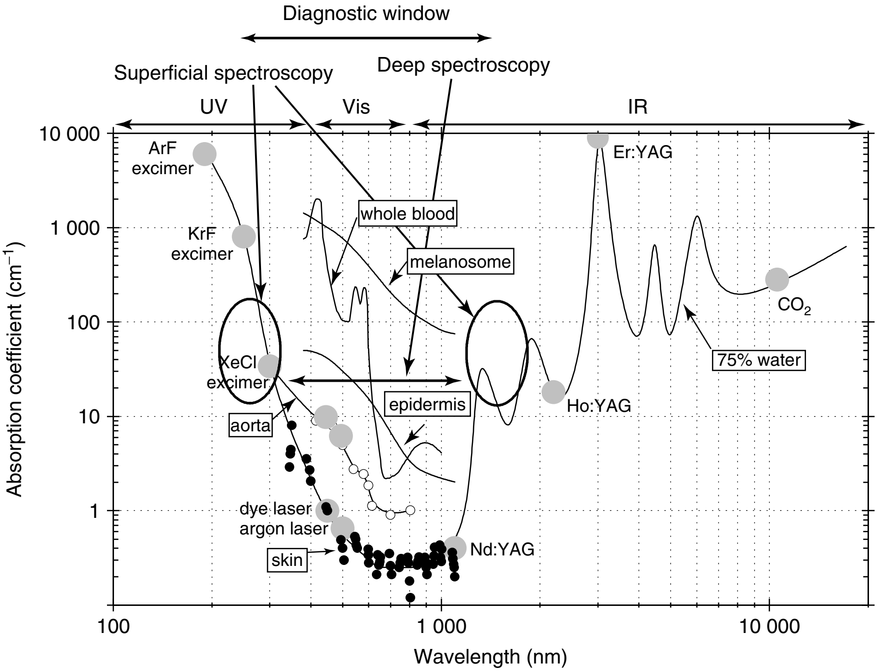

Fig. 16.1 Absorption spectra of some tissues and tissue components: water (75%), aorta, skin, epidermis, melanosome, and whole blood; lasers and their wavelengths as well as diagnostic window and wavelength ranges suitable for superficial and deep spectroscopy are also shown (Tuchin, 2007).

of female breast, is a very complicated process. In skin, the horny layer (stratum corneum (SC)) reflects about 5–7% of the incident light. A collimated light beam is transformed to a diffuse one by microscopic inhomogeneities at the air/horny layer interface. A major part of reflected light results from backscattering in different skin layers (SC, epidermis, dermis, blood, and percutaneous fat). The absorption of diffuse light by skin pigments is a measure of bilirubin content, hemoglobin concentration and its saturation with oxygen, and the concentration of pharmaceutical products in blood and tissues; these characteristics are widely used in the diagnosis of various diseases (Tuchin, 2002; Vo-Dinh, 2003; Tuchin, 2007). Certain diagnostic modalities take advantage of ready transdermal penetration of visible and near-infrared (NIR) light inside the body in the wavelength region corresponding to the diagnostic/therapeutic window (600–1600 nm).

16.2.2.2 **Theoretical Description for CW**

A collimated (laser) beam is attenuated in a thin tissue layer of thickness d in accordance with the exponential law, Bouguer-Beer-Lambert law (Tuchin, 2002; Vo-Dinh, 2003; Tuchin, 2007):

$$I(d) = (1 - R_F) I_0 \exp(-\mu_t d) \qquad (16.1)$$

where $I(d)$ is the intensity of transmitted light measured using a distant photodetector with a small aperture (online or collimated transmittance) in W cm^{-2}; R_F is the coefficient of Fresnel reflection; at the normal beam incidence, $R_F = [(n - l)/(n + l)]^2$; n is the relative mean refractive index of tissue and surrounding media; I_0 is the incident light intensity in W cm^{-2};

$$\mu_t = \mu_a + \mu_s \qquad (16.2)$$

is the extinction coefficient (interaction or total attenuation coefficient) in cm^{-1}, where μ_a is the absorption coefficient in cm^{-1}, and μ_s is the scattering coefficient in cm^{-1}. Strictly speaking, Eq. (16.1) is valid only for a highly absorbing biological media, when $\mu_a \gg \mu_s$.

The extinction coefficient is connected with the extinction cross section σ_{ext} as

$$\mu_t = \rho_s \sigma_{\text{ext}} \qquad (16.3)$$

where ρ_s is the density of particles (tissue and cell compounds). For a system of particles with absorption,

$$\sigma_{\text{ext}} = \sigma_{\text{sca}} + \sigma_{\text{abs}} \qquad (16.4)$$

and

$$\mu_s = \rho_s \sigma_{\text{sca}} \quad \mu_a = \rho_s \sigma_{\text{abs}} \qquad (16.5)$$

The average scattering cross section per particle can be presented in the form suitable for experimental evaluations (Bohren and Huffman, 1983):

$$\sigma_{\text{sca}} = \frac{\lambda^2}{2\pi} \frac{1}{I_0} \int_0^{\pi} I(\theta) \sin\theta \, d\theta \qquad (16.6)$$

where I_0 is the intensity of the incident light, $I(\theta)$ is the angular distribution of the scattered light by a particle, and θ is the scattering angle.

The mean free path (MFP) length between two interactions is denoted by

$$l_{\text{ph}} = \mu_t^{-1} \qquad (16.7)$$

To analyze light propagation under multiple scattering conditions, it is assumed that absorbing and scattering centers are uniformly distributed across the tissue.

Visible and NIR radiation is normally subject to anisotropic scattering characterized by a clearly apparent direction of photons undergoing single scattering, which may be due to the presence of large cellular organelles (mitochondria, lysosomes, and inner membranes (Golgi apparatus)) (Tuchin, 2002, 2007).

When the scattering medium is illuminated by unpolarized light and/or only the intensity of multiply scattered light needs to be computed, a sufficiently strict mathematical description of continuous wave (CW) light propagation in a medium is possible in the framework of the scalar stationary radiation transfer theory (RTT) (Chandrasekhar, 1960; van de Hulst, 1980, 1981; Ishimaru, 1997; Mishchenko, Travis and Lacis, 2000, 2002, 2006; Müller *et al.*, 1993; Niemz, 1996; Sobolev, 1974; Thomas and Stamnes, 1999; Tuchin, 2002, 2007; Tuchin, Wang and Zimnyakov, 2006; Vo-Dinh, 2003; Wang and Wu, 2007; Welch and van Gemert, 1992; Yanovitskij, 1997; Zege, Ivanov and Katsev, 1991). The theory is valid for an ensemble of scatterers located far from one another and has been successfully used to work out many practical problems of tissue optics. The main stationary equation of RTT for monochromatic light has the form

$$\frac{\partial I(\bar{r},\bar{s})}{\partial s} = -\mu_t I(\bar{r},\bar{s}) + \frac{\mu_s}{4\pi} \times \int_{4\pi} I(\bar{r},\bar{s}')p(\bar{s},\bar{s}')d\Omega' \quad (16.8)$$

where $I(\bar{r},\bar{s})$ is the radiance (or specific intensity) – average power flux density at a point $\bar{r}$ in the given direction $\bar{s}$ (W cm^{-2}sr); $p(\bar{s},\bar{s}')$ is the scattering phase function, 1/sr; and $d\Omega'$ is the unit solid angle about the direction $\bar{s}'$, sr. It is assumed that there are no radiation sources inside the medium.

For practical purposes, integrals of the function $I(\bar{r},\bar{s})$ over certain phase space regions $(\bar{r},\bar{s})$ are of greater value than the function itself. Specifically, optical probes of tissues frequently measure the outgoing light distribution function at the medium surface, which is characterized by the radiant flux density or irradiance (W cm^{-2}):

$$F(\bar{r}) = \int_{(\bar{s}\overline{N})>0} I(\bar{r},\bar{s})(\bar{s}\overline{N})d\Omega \quad (16.9)$$

where $\overline{N}$ is the outside normal vector to the domain boundary surface.

In problems of optical dosimetry in tissues, the measured quantity is actually the total radiant energy fluence rate $U(\bar{r})$. It is the sum of the radiance over all angles at a point $\bar{r}$ and is measured by watts per square centimeter:

$$U(\bar{r}) = \int_{4\pi} I(\bar{r},\bar{s})d\Omega \quad (16.10)$$

The phase function $p(\bar{s},\bar{s}')$ describes the scattering properties of the medium and is, in fact, the probability density function for scattering in the direction $\bar{s}'$ of a photon traveling in the direction $\bar{s}$; in other words, it characterizes an elementary scattering act. If scattering is symmetric relative to the direction of the incident wave, then the phase function depends only on the scattering angle θ (angle between directions $\bar{s}$ and $\bar{s}'$), that is,

$$p(\bar{s},\bar{s}') = p(\theta) \quad (16.11)$$

The assumption of random distribution of scatterers in a medium (i.e., the absence of spatial correlation in the tissue structure) leads to normalization

$$\int_0^{\pi} p(\theta)2\pi \ \sin\theta \, d\theta = 1 \tag{16.12}$$

In practice, the phase function is usually well approximated with the aid of the postulated Henyey–Greenstein function:

$$p(\theta) = \frac{1}{4\pi} \cdot \frac{1-g^2}{(1+g^2-2g\ \cos\ \theta)^{3/2}} \tag{16.13}$$

where g is the scattering anisotropy parameter (mean cosine of the scattering angle θ).

$$g \equiv \langle \cos\ \theta \rangle = \int_0^{\pi} p(\theta) \cos\ \theta \cdot 2\pi \sin\ \theta \, d\theta \tag{16.14}$$

The value of g varies in the range from -1 to 1 : $g = 0$ corresponds to isotropic (Rayleigh) scattering, $g = 1$ to total forward scattering (Mie scattering at large particles), and -1 to total backward scattering.

The integro-differential Eq. (16.8) is too complicated; therefore, it is frequently simplified by representing the solution in the form of spherical harmonics. Such simplification leads to a system of $(N+1)^2$ connected differential partial derivative equations known as the P_N approximation. This system is reducible to a single differential equation of order $(N+1)$. For example, four connected differential equations reducible to a single diffusion-type equation are necessary for $N = 1$. It has the following form for an isotropic medium:

$$\left(\nabla^2 - \mu_{\text{eff}}^2\right) U(\bar{r}) = -Q(\bar{r}) \tag{16.15}$$

where

$$\mu_{\text{eff}} = [3\mu_a(\mu_s{}' + \mu_a)]^{1/2} \tag{16.16}$$

is the effective attenuation coefficient or inverse diffusion length, $\mu_{\text{eff}} = 1/l_d$ in cm^{-1};

$$Q(\bar{r}) = (cD)^{-1}q(\bar{r}) \tag{16.17}$$

$q(\bar{r})$ is the source function (i.e., the number of photons injected into the unit volume),

$$D = \frac{1}{3(\mu_s{}' + \mu_a)} \tag{16.18}$$

is the photon diffusion coefficient in cm^2/c;

$$\mu_s{}' = (1-g)\mu_s \tag{16.19}$$

is the reduced (transport) scattering coefficient in cm^{-1}, and c is the velocity of light in the medium. The transport mean free path (TMFP) of a photon (centimeters) is defined as

$$l_t = \frac{1}{\mu_t{}'} = (\mu_a + \mu_s{}')^{-1} \tag{16.20}$$

where $\mu_t{}' = \mu_a + \mu_s{}'$ is the transport coefficient. The TMFP is the distance over which the photon loses its initial direction. It is worthwhile to note that the TMFP in a medium with anisotropic single scattering significantly exceeds the MFP in a medium with isotropic single scattering $l_t \gg l_{\text{ph}}$.

The diffusion approximation is inapplicable for light beam input near the object's surface where single or low-step scattering prevails. When a narrow light beam normally incident upon a semi-infinite turbid medium with anisotropic scattering, it can be considered as converted into an isotropic point source at the depth of one TMFP below the surface.

The refractive index variation in tissues is quantified by the ratio $m \equiv n_s/n_0$, where n_s and n_0 are the index of refraction of scattering particles and background material, respectively. This ratio determines light-scattering efficiency. For example, in a simple monodisperse tissue model, such as dielectric spheres of equal diameter $2a$, the reduced scattering coefficient (Graaff *et al.*, 1992):

$$\mu_s' = 3.28\pi a^2 \rho_s \frac{2\pi a}{\lambda}^{0.37} (m-1)^{2.09} \tag{16.21}$$

where λ is the light wavelength in the scattering medium. This equation is valid for noninteracting Mie scatterers, $g > 0.9$; $5 < 2\pi a/\lambda < 50$; $1 < m < 1.1$. It gives dependencies of reduced scattering coefficient on refractive index mismatch (m), particle size (a), density (ρ_s), and wavelength (λ).

Basing on diffusion approximation attenuation of a light (laser) beam of intensity I_0 at depths $z \geq l_t$ in a thick tissue may be described as

$$I(z) \approx I_0 b_s \exp(-\mu_{\text{eff}} z) \tag{16.22}$$

where b_s accounts for additional irradiation of upper layers of a tissue due to backscattering (photon recycling effect). Respectively, the depth of light penetration into a tissue is

$$l_e = l_d[ln\ b_s + 1] \tag{16.23}$$

Typically, for tissues $b_s = 1\text{–}5$ for beam diameter of 1–20 mm (Star, Wilson and Patterson, 1992). Thus, when wide laser beams are used for irradiation of highly scattering tissues with low absorption, light energy is accumulated in tissue due to high multiplicity of chaotic long-path photon migrations. The light-power density within the superficial tissue layers may substantially (up to fivefold) exceed the incident power density.

Measurement of diffusely reflected light is often used to evaluate bulk tissue optical properties. To solve diffusion equation, the boundary conditions are derived by considering Fresnel's laws of reflection and balancing the fluence rate and photon flow crossing the interface. For the source term modeled as a point scattering source and extrapolated boundary approach satisfying the boundary condition, the spatially resolved steady-state reflectance per incident photon, $R(r_{\text{sd}})$, is expressed as (Farrell and Patterson, 2001):

$$\begin{aligned} R(r_{\text{sd}}) = {} & \frac{F_U}{4\pi}\left\{l_t\left(\mu_{\text{eff}} + \frac{1}{r_1}\right)\right. \\ & \times \frac{\exp(-\mu_{\text{eff}} r_1)}{r_1^2} + (l_t + 2z_b) \\ & \left.\times \left(\mu_{\text{eff}} + \frac{1}{r_2}\right)\frac{\exp(-\mu_{\text{eff}} r_2)}{r_2^2}\right\} \\ & + \frac{F_F}{4\pi D}\left\{\frac{\exp(-\mu_{\text{eff}} r_1)}{r_1}\right. \\ & \left. - \frac{\exp(-\mu_{\text{eff}} r_2)}{r_2}\right\} \end{aligned} \tag{16.24}$$

where r_{sd} is the distance between light source and detector at the tissue surface (source–detector separation), cm; $r_1 = \sqrt{l_t^2 + r_{\text{sd}}^2}$; $r_2 = \sqrt{(l_t + 2z_b)^2 + r_{\text{sd}}^2}$; $z_b = 2AD$ is the distance to the extrapolated boundary, $A = (1 + R_{\text{eff}})/(1 - R_{\text{eff}})$, R_{eff} is the effective reflection coefficient, which can be found by integrating the Fresnel reflection coefficient over all incident angles; and D is the diffusion coefficient. The parameters F_U and F_F represent the fractions of the fluence rate and the flux, which exit the tissue across the interface. These values are obtained by integration of the radiance over the backward hemisphere and depend on refractive index mismatch on the boundary.

16.2.2.3 Index of Refraction of Tissue Compartments

The mean refractive index $\overline{n}$ of a tissue is defined by the refractive indices of its scattering particle material n_s and ground (surrounding) matter n_0. The refractive indices of tissue structure elements and the tissue itself, can be derived using the law of Gladstone and Dale, which states that the resulting value represents an average of the refractive indices of the components related to their volume fractions as (Tuchin, 2002, 2007):

$$\overline{n} = \sum_{i=1}^{N} n_i f_i \quad \sum_i f_i = 1 \tag{16.25}$$

where n_i and f_i are the refractive index and volume fraction of the individual components, respectively, and N is the number of components.

Soft tissue is composed of closely packed groups of cells entrapped in a network of fibers through which water percolates. At a microscopic scale, the tissue components have no pronounced boundaries. They appear to merge into a continuous structure with spatial variations in the refractive index. To model such a complicated structure as a collection of particles, it is necessary to resort to a statistical approach.

It has been shown that the tissue components that contribute most to the local refractive-index variations are the connective tissue fibers (bundles of elastin and collagen), cytoplasmic organelles (mitochondria, lysosomes, and peroxisomes), cell nuclei, and melanin granules. The average background index is defined as the weighted average of refractive indices of the cytoplasm and the interstitial fluid, n_{cp} and n_{is}, as

$$\overline{n}_0 = f_{\mathrm{cp}} n_{\mathrm{cp}} + (1 - f_{\mathrm{cp}}) n_{\mathrm{is}} \tag{16.26}$$

where f_{cp} is the volume fraction of the fluid in the tissue contained inside the cells. Accounting for typical experimental data as $n_{\mathrm{cp}} = 1.367$ and $n_{\mathrm{is}} = 1.355$, and that approximately 60% of the total fluid in soft tissue is contained in the intracellular compartment, it follows from Eq. (16.26) that $\overline{n}_0 = 1.362$. The refractive index of scattering particle material can be determined as the sum of the background index and index variation components (Tuchin, 2007):

$$\overline{n}_s = \overline{n}_0 + f_f(n_f - n_{\mathrm{is}}) + f_{\mathrm{nc}}(n_{\mathrm{nc}} - n_{\mathrm{cp}}) + f_{\mathrm{or}}(n_{\mathrm{or}} - n_{\mathrm{cp}}) \tag{16.27}$$

where the subscripts f, is, nc, cp, and or refer to the fibers, interstitial fluid, nuclei, cytoplasm, and organelles, which are the major contributors to index variations. The terms in parentheses in this expression are the differences between the refractive indices of the three types of tissue component and their respective backgrounds; the multiplying factors are the volume fractions of the elements in the solid portion of the tissue. The refractive index of the connective-tissue fibers is about 1.47, which corresponds to about 55% hydration of collagen, its main component. Taking into account that $n_{\mathrm{nc}} \cong n_{\mathrm{or}} = 1.40$, the mean index variation can be expressed in terms of the fibrous-tissue fraction c_f only:

$$\overline{n}_s = \overline{n}_0 + f_f(n_f - n_{\mathrm{is}}) + (1 - f_f) \times (n_{\mathrm{nc}} - n_{\mathrm{cp}}) \tag{16.28}$$

Collagen and elastin fibers compose approximately 70% of the fat-free dry weight of the dermis, 45% of the heart, and 2–3% of the nonmuscular internal organs (Schmitt and Kumar, 1998). Therefore, depending on tissue type, f_f may be as small as about 0.02 or as large as 0.7. For $n_f - n_{\text{is}} = 1.470 - 1.355 = 0.115$ and $n_{\text{nc}} - n_{\text{cp}} = n_{\text{or}} - n_{\text{cp}} = 1.400 - 1.367 = 0.033$, the mean index variations that correspond to these two extremes are $\overline{n}_s = 1.397$ and 1.452. Indeed, for some hard tissues or melanin granules in skin or eye iris, refractive index of scatterers could be as high as 1.5–1.7.

Measuring refractive indices in tissues and their constituent components is an important focus of interest in tissue optics because index of refraction determines light reflection and refraction at the interfaces between air and tissue, detecting fiber and tissue, and tissue layers; it also strongly influences light propagation and distribution within tissues, defines speed of light in tissue, and governs how the photons migrate (Tuchin, 2007). Although these studies have a rather long history (Duck, 1990), the mean values of refractive indices for many tissues are missing in the literature. According to Duck, most of them have refractive indices for visible light in the 1.335–1.620 range (e.g., 1.55 in the SC, 1.620 in the enamel, and 1.386 at the lens surface). It is worthwhile noting that *in vitro* and *in vivo* measures may differ significantly. For example, the refractive index in rat mesenteric tissue *in vitro* was found to be 1.52 compared with only 1.38 *in vivo*. This difference can be accounted for by the decreased refractivity of ground matter, n_0, due to impaired hydration.

The optical properties of tissues, including refractive indices, are known to depend on water content. Tissue is modeled as a mixture of water and a bioorganic compound of a tissue. For instance, the refractive index of human skin can be approximated by a 70/30 mixture of water and protein (Troy and Thennadil, 2001). Assuming that protein has a constant refractive index value of 1.5 over the entire wavelength range, the following expression for estimation of skin index of refraction was suggested:

$$\begin{aligned} n_{\text{skin}}(\lambda) &= 0.7(1.58 - 8.45 \times 10^{-4}\lambda + 1.10 \\ &\quad \times 10^{-6}\lambda^2 - 7.19 \times 10^{-10}\lambda^3 + 2.32 \\ &\quad \times 10^{-13}\lambda^4 - 2.98 \times 10^{-17}\lambda^5) \\ &\quad + 0.3 \cdot 1.5 \end{aligned} \tag{16.29}$$

where wavelength λ is in nanometers.

16.2.2.4 **Short Pulse Interaction with Biomaterials**

When probing the plane-parallel layer of a scattering medium (biomaterial) with an ultrashort laser pulse, the transmitted pulse consists of a ballistic component, a group of photons having zigzag trajectories, and a highly intensive diffuse component (Müller *et al.*, 1993; Tuchin, 2007). Both unscattered photons and photons undergoing forward-directed single-step scattering contribute to the intensity of the ballistic component. This component is subject to exponential attenuation with increasing sample thickness (see Eq. (16.1)), thus, has a limited utility for practical diagnostic purposes in medicine. The group of snake photons with zigzag trajectories includes photons that have experienced only a few collisions each. They propagate along trajectories that deviate only slightly from the direction of the incident beam and form the first-arriving to detector part of the diffuse component.

These photons carry information about the optical properties of the biomaterial and parameters of any foreign object (i.e., tumor), which they may happen to come across during their travel.

Typically for many scattering biomaterials, the diffuse component is very broad and intense since it contains the bulk of incident photons after they have participated in many scattering acts and therefore migrate in different directions and have different path lengths. Moreover, the diffuse component carries information about the optical properties of the scattering medium, and its deformation may reflect the presence of local inhomogeneities in the medium. The resolution obtained by this method at a high light-gathering power is much lower than in the method measuring straight-passing photons. Two probing schemes are conceivable, one recording transmitted photons and the other taking advantage of their backscattering. If in the diffusion approximation (valid at $\mu_a \ll \mu_s{}'$), the tissue is homogeneous and semi-infinite, the size of both the source and the detector is small compared with the distance r_{sd} between them at the tissue surface, and the pulse may be regarded as single, then the light distribution is described by the time-dependent diffusion equation (Müller *et al.*, 1993; Tuchin, 2002, 2007):

$$\left(\nabla^2 - c\mu_a D^{-1} - D^{-1}\frac{\partial}{\partial t}\right) \cdot U(\bar{r}, t) = -Q(\bar{r}, t) \qquad (16.30)$$

which is, in fact, the generalization of the CW Eq. (16.15). Solving Eq. (16.30) yields the following relation for the number of backscattered photons at the surface for unit time and from unit area $R(r_{\mathrm{sd}}, t)$ (Jacques, 1989; Patterson, Chance and Wilson, 1989)

$$R(r_{\mathrm{sd}}, t) = \frac{z_0}{(4\pi D)^{3/2}} t^{-5/2} \times \exp\left(-\frac{r_{\mathrm{sd}}^2 + z_0^2}{2Dt}\right) \exp(-\mu_a ct) \qquad (16.31)$$

and correspondingly for transmittance

$$\begin{aligned} T(r_{\mathrm{sd}}, d, t) &= (4\pi D)^{-3/2} t^{-5/2} \exp\frac{-r_{\mathrm{sd}}^2}{4Dt} \\ &\times \left\{(d - z_0) \exp\left[-\frac{(d - z_0)^2}{4Dt}\right]\right. \\ &- (d + z_0) \exp\left[-\frac{(d + z_0)^2}{4Dt}\right] \\ &+ (3d - z_0) \exp\left[-\frac{(3d - z_0)^2}{4Dt}\right] \\ &\left. - (3d + z_0) \exp\left[-\frac{(3d + z_0)^2}{4Dt}\right]\right\} \\ &\times \exp(-\mu_a ct) \end{aligned} \qquad (16.32)$$

where $z_0 = (\mu_s{}')^{-1}$ and d is the tissue thickness.

In practice, μ_a and $\mu_s{}'$ are estimated by fitting Eq. (16.31) or Eq. (16.32) with the shape of a pulse measured by the time-resolved photon counting technique. Experimentally measured optical parameters of many tissues and model media obtained by the pulse method can be found in Cheong, Prahl and Welch (1990); Welch and Gemert (1992); Müller *et al.* (1993); Roggan *et al.* (1995); Tuchin (2002); Vo-Dinh (2003); Tuchin (2007). An important advantage of the pulse method is its applicability to *in vivo* studies owing to the possibility of the separate evaluation of μ_a and $\mu_s{}'$ using a single measurement in the backscattering or transillumination regimes.

Many studies are also devoted to image transfer in tissues and the evaluation of the resolving power of optical tomographic schemes making use of the first-transmitted photons of ultrashort pulses (Müller *et al.*, 1993; Tuchin, 2002; Vo-Dinh, 2003; Tuchin, 2007). A contrast image of an object in a scattering medium can be provided by electronical or optical time-gating of the earliest-arriving, minimally scattered light (ballistic and snake photons), which contains geometric information. The first group of schemes uses the electronic time-gating procedure. The time-correlated single-photon counting technique explores a high-repetition rate picosecond laser (e.g., a cavity-dumped mode-locked dye laser). At the detection of the firstly arrived photons, the time delay is measured with a time-to-amplitude converter and a histogram of the arrival times is built up using a large number of low-energy pulses. The time resolution of such a technique is limited to about 50 picoseconds. For more energetic pulses from lasers with a lower repetition rate, the usage of the streak cameras allows for a time resolution down to 1 picosecond. If a synchroscan streak camera is employed, even a high-repetition rate source with low-energy pulses can be used.

The second group of techniques uses the optical nonlinear effects to select photons. For a scheme with an optical Kerr gate, an energetic laser is used. Part of the pulse is transmitted into the tissue and part is used to open the shutter by utilizing the optical Kerr effect (i.e., the cell with CS_2). Since the gate width is determined by the length of the laser pulse only, subpicosecond gate times can be achieved. The Raman amplifier gating technique also uses energetic laser pulses. A Stokes wave generated by stimulated Raman scattering in a gas cell is used to probe the tissue. The low-intensity transmitted light is amplified in a Raman amplifier, which in turn is pumped by an ultrashort laser pulse. This pulse has the proper time delay to strobe on the desired early temporal part of the light under investigation. The third group of schemes uses a time-correlated frequency doubling technique and is frequently used in optical autocorrelators for monitoring laser pulse characteristics. It can be used directly for optical gating of signal photons.

16.3
Measurements of Optical Properties of Tissues and Blood

16.3.1
Introduction

Methods for determining the optical parameters of tissues can be divided into two large groups, direct and indirect methods (Tuchin, 2007). Direct methods include those based on some fundamental concepts and rules such as the Bouguer–Beer–Lambert law (see Eq. (16.1)), the single-scattering phase function (see Eqs. (16.11) and (16.13)) for thin samples, or the effective light penetration depth for slabs. The parameters measured are the collimated light transmission T_c and angular dependence of the scattered light intensity $I\ (\theta)$ in W cm^{-2}sr, for thin samples or the fluence rate inside a slab. The normalized scattering angular dependence is equal to the scattering phase function $I\ (\theta)/I(0) \equiv p(\theta)$, 1/sr. These methods are advantageous in that they use very simple analytic expressions for data processing. Their disadvantages are related to the necessity to strictly fulfill experimental conditions dictated by the selected model (single scattering in thin samples, exclusion of the effects of light polarization and refraction at cuvette edges, and so on; in the case of slabs with multiple scattering, the recording detector must be placed far from both the light source and the medium boundaries.

Indirect methods obtain the solution of the inverse scattering problem using a theoretical model of light propagation in a medium. They are in turn divided into iterative and noniterative models. The noniterative models use equations in which the optical properties are defined through parameters directly related to the quantities being evaluated. They are based on the two-flux or multiflux models. In indirect iterative methods, the optical properties are implicitly defined through measured parameters. Quantities determining the optical properties of a scattering medium are enumerated until the estimated and measured values for reflectance and transmittance coincide with the desired accuracy. These methods are cumbersome, but the optical models currently in use may be even more complicated than those underlying noniterative methods (examples include the diffusion theory (DT), IAD, and IMC methods) (see Tuchin, 2007).

The optical parameters of tissue samples (μ_a, μ_s, and g) are measured by different methods. *In vitro* evaluation is most often achieved by the double-integrating sphere (DIS) method combined with collimated transmittance measurements. This approach implies either sequential or simultaneous determination of three parameters: collimated transmittance $T_c = I(d)/I(0)$ (see Eq. (16.1)), total transmittance $T_t = T_c + T_d$ (T_d being diffuse transmittance), and diffuse reflectance R_d. The optical parameters of the tissue are deduced from these measurements using different theoretical expressions or numerical methods (two-flux and multiflux models, the IMC or IAD methods) relating μ_a, μ_s, and g to the parameters being investigated.

Any three measurements from the following five are sufficient for the evaluation of all three optical parameters (Cheong, Prahl and Welch, 1990): (i) total (or diffuse) transmittance for collimated or diffuse radiation; (ii) total (or diffuse) reflectance for collimated or diffuse radiation; (iii) absorption by a sample placed inside an IS; (iv) collimated transmittance (of unscattered light); (v) angular distribution of radiation scattered by the sample. Iterative methods normally take into account discrepancies between refractive indices at sample boundaries as well as the multilayer nature of the sample. The following factors are responsible for the errors in the estimated values of optical coefficients and need to be borne in mind in a comparative analysis of optical parameters obtained in different experiments: (i) physiological conditions of tissues (the degree of hydration, homogeneity, species-specific variability, frozen/thawed or fixed/unfixed state, *in vitro*/*in vivo* measurements, smooth/rough surface); (ii) geometry of irradiation; (iii) matching/mismatching interface refractive indices; (iv) orientation of detecting optical fibers inside the sample relative to the source fibers; (v) numerical aperture of the recording fibers; (vi) angular resolution of photodetectors; (vii) separation of radiation experiencing forward scattering from unscattered radiation; and

(viii) theory used to solve the inverse problem.

16.3.2
Integrating Sphere Technique

One of the indirect methods to determine optical properties of tissues *in vitro* is the IS technique. Diffuse reflectance R_d, total transmittance T_t, and collimated transmittance T_c are measured. In general, absorption coefficient μ_a, scattering coefficient μ_s, and anisotropy factor g can be obtained from these data using an inverse method based on the radiative transfer theory (see Eq. (16.8)). When the scattering phase function $p(\theta)$ is available from goniophotometry, g can be readily calculated. In this case, for the determination of μ_a and μ_s, it is sufficient to measure R_d and T_t only. Sometimes in experiments with tissue and blood samples, a DIS configuration is preferable, since in this case both reflectance and transmittance can be measured simultaneously and less degradation of the sample is expected during measurements.

The IS technique was used by a number of investigators to determine the absorption coefficient, the scattering coefficient, the anisotropy factor, and/or the reduced scattering coefficient of tissues and blood (see Tuchin, 2007). Barium-sulfate or Spectralon® ISs are used in the experiments. As monochromatic light sources, a laser, an Xe-lamp, and/or a Hg-lamp combined with monochromator are used, while a photomultiplier or an Si-photodiode is employed as a detector. Sometimes, a white light source is used as irradiator and charge-coupled device (CCD) fiber-optic spectrometer as a detector.

16.3.3
Multiflux Approach

To separate the light beam attenuation due to absorption from the loss due to scattering, the one-dimensional, two-flux Kubelka-Munk model (KMM) can be used as the simplest approach to solve the problem. This approach has been widely used to determine the absorption and scattering coefficients of biological tissues, provided the scattering is significantly dominant over the absorption (see Tuchin, 2007). The KMM assumes that light incident on a slab of tissue because of interaction with the scattering media can be modeled by two fluxes, counterpropagating in the tissue slab. The optical flux, which propagates in the same direction as the incident flux, is decreased by absorption and scattering processes and is also increased by backscattering of the counterpropagating flux in the same direction. Changes in counterpropagating flux are determined in an analogous manner. The fraction of each flux lost by absorption per unit path length is denoted as K, while the fraction lost due to scattering is called S. The main assumptions of the KMM: K and S parameters are assumed to be uniform throughout the tissue slab; all light fluxes are diffuse; and the amount of light lost from the edges of the sample during reflectance measurements is negligible. Basic KMM does not account for reflections at boundaries at which index of refraction mismatches exist. Following the KMM and diffusion approximation of the RTT, the KMM parameters were expressed in terms of light transport theory: the absorption and scattering coefficients and scattering anisotropy factor (Cheong, Prahl and

Welch, 1990). Often, such simple methods as the KMM, are used as the first step of the inverse algorithm for estimation of the optical properties of tissues.

16.3.4
The Inverse Adding–Doubling Method

The IAD method provides a tool for the rapid and accurate solution of inverse scattering problem. It is based on the general method for the solution of the transport equation (see Eq. (16.8)) for plane-parallel layers suggested by van de Hulst (1980) and introduced to tissue optics by Prahl, van Gemert and Welch (1993). An important advantage of the IAD method when applied to tissue optics is the possibility of rapidly obtaining iterative solutions with the aid of up-to-date microcomputers; moreover, it is flexible enough to take into account anisotropy of scattering and the internal reflection from the sample boundaries. The method includes the following steps: (i) the choice of optical parameters to be measured; (ii) counting reflections and transmissions; (iii) comparison of calculated and measured reflectance and transmittance; and (iv) repetition of the procedure until the estimated and measured values coincide with the desired accuracy.

The method allows any intended accuracy to be achieved for all the parameters being measured, provided the necessary computer time is available. An error of 3% or less is considered acceptable (Prahl, van Gemert and Welch, 1993). Also, the method may be used to directly correct experimental findings obtained with the aid of ISs. The term *doubling* in the method IAD means that the reflection and transmission estimates for a layer at certain ingoing and outgoing light angles may be used to calculate both the transmittance and reflectance for a layer twice as thick by means of superimposing one upon the other and summing the contributions of each layer to the total reflectance and transmittance. Reflection and transmission in a layer having an arbitrary thickness are calculated in consecutive order, first for the thin layer with the same optical characteristics (single scattering), then by consecutive doubling of the thickness, for any selected layer. The term *adding* indicates that the doubling procedure may be extended to heterogeneous layers for modeling multilayer tissues or taking into account internal reflections related to abrupt change in refractive index.

The IAD is a numerical method for solving the one-dimensional transport equation in slab geometry. It can be used for media with an arbitrary phase function and arbitrary angular distribution of the spatially uniform incident radiation. The IAD method has been successfully applied to determine optical parameters of blood; human and animal dermis; ocular tissues such as retina, choroids, sclera, conjunctiva, and ciliary body; aorta; and other soft tissues in the wide range of the wavelengths. The IAD provides accurate results in cases when the side losses are not significant, but it is less flexible than the IMC technique.

16.3.5
Inverse Monte Carlo Method

Both the real geometry of the experiment and the tissue structure may be complicated. Therefore, the MC method should be used if reliable estimates are to be obtained. A number of algorithms that use the IMC method are available now in the literature (see Tuchin, 2007).

Many researchers use the MC simulation program provided by Jacques.

The MC technique is employed as a method to solve the forward problem in the inverse algorithm for the determination of the optical properties of tissues. The MC method is based on the formalism of the RTT (see Eq. (16.8)), where the absorption coefficient is defined as a probability of a photon to be absorbed per unit length, and the scattering coefficient is defined as the probability of a photon to be scattered per unit length. Using these probabilities, a random sampling of photon trajectories is generated.

The basic algorithm for the generation of photon trajectories can be shortly described as follows (Tuchin, 2007). A photon described by three spatial coordinates and two angles (x, y, z, θ, ϕ) is assigned its weight $W = W_0$ and placed in its initial position, depending on the source characteristics. When an incident photon enters a scattering layer, it is allowed to travel a free path length, l. The l value depends on the particle concentration, ρ, and extinction cross section σ_{ext}. The free path length l is a random quantity that takes any positive values with the probability density $p(l)$:

$$p(l) = \rho\sigma_{\text{ext}} e^{-\rho\sigma_{\text{ext}} l} \tag{16.33}$$

The particular realization of the free path length l is dictated by the value of a random number ξ that is uniformly distributed over the interval [0, 1]:

$$\int_0^l p(l)dl = \xi \tag{16.34}$$

Substituting Eq. (16.33) into Eq. (16.34) yields the value l of the certain realization in the form

$$l = -\frac{1}{\rho\sigma_{\text{ext}}} \ln \xi \tag{16.35}$$

If the distance l is larger than the thickness of the scattering system, then this photon is detected as transmitted without any scattering. If, having passed the distance l, the photon remains within the scattering volume, then the possible events of photon-particle interaction (scattering or absorption) are randomly selected.

The direction of the photon's next movement is determined by the scattering phase function substituted as the probability density distribution. Several approximations for the scattering phase function of tissue and blood have been used in MC simulations. These include the two empirical phase functions widely used to approximate the scattering phase function of tissue and blood, Henyey–Greenstein phase function (HGPF) (see Eq. (16.13)) and the Gegenbauer kernel phase function (GKPF) (Barber and Hill, 1990), and theoretical Mie phase function (Bohren and Huffman, 1983).

When the photon reaches the boundary, part of its weight is transmitted according to the Fresnel equations. The amount transmitted through the boundary is added to the reflectance or transmittance. Since the refraction angle is determined by the Snell's law, the angular distribution of the outgoing light can be calculated. The photon with the remaining part of the weight is specularly reflected and continues its random walk.

When the photon's weight becomes lower than a predetermined minimal value, the photon can be terminated using "Russian roulette" procedure.

This procedure saves time, since it does not make sense to continue the random walk of the photon, which will not essentially contribute to the measured signal. On the other hand, it ensures that the energy balance is maintained throughout the simulation process.

The MC method has several advantages over the other methods because it may take into account mismatched medium-glass and glass–air interfaces, losses of light at the edges of the sample, any phase function of the medium, and the finite size and arbitrary angular distribution of the incident beam. If the collimated transmittance is measured, then the contribution of scattered light into the measured collimated signal can be accounted for. The only disadvantage of this method is the long time needed to ensure good statistical convergence, since it is a statistical approach. The standard deviation of a quantity (diffuse reflectance, transmittance, etc.) approximated by MC technique decreases proportionally to $1/\sqrt{N}$, where N is the total number of launched photons. It is worthy to note that stable operation of the algorithm was maintained by generation from 10^5 to 5×10^5 photons per iteration. Two to five iterations were usually necessary to estimate the optical parameters with approximately 2% accuracy.

16.3.6
Spatially Resolved Measurements

For many tissues, *in vivo* measurements are possible only in the geometry of the backscattering. The spatially resolved reflectance (SRR) $R(r_{\mathrm{sd}})$ is defined as the power of the backscattered light per unit of area detected by a receiver at the surface of the tissue at a distance r_{sd} from the source. $R(r_{\mathrm{sd}})$ depends on the optical properties of the sample, that is, the absorption coefficient μ_a, the scattering coefficient μ_s, and the phase function $p(\theta)$, the refractive index, and the numerical aperture (NA) of the receiving system (Kumar and Schmitt, 1997; Farrell and Patterson, 2001). The corresponding relation for the backscattering intensity as a function of a source and detector positions and optical parameters can be written on the basis of a diffusion approximation (see Eq. (16.24)).

For a semi-infinite medium and source and detector probes (e.g., optical fibers) separated by a distance r_{sd} and normally oriented to the sample surface, the reflecting flux is given by Kumar and Schmitt (1997)

$$R = \frac{z_0 A}{2\pi}\left[\frac{\mu_{\mathrm{eff}}}{r_{\mathrm{sd}}^2 + z_0^2} + \frac{1}{(r_{\mathrm{sd}}^2 + z_0^2)^{3/2}}\right] \times \exp\left[-\mu_{\mathrm{eff}}(r_{\mathrm{sd}}^2 + z_0^2)^{1/2}\right] \quad (16.36)$$

where $z_0 = K/\mu_s'$ is the extrapolation length, K is a dimensionless constant with a magnitude that depends on the anisotropy parameter of the scatterers and the reflection coefficient at the surface, A is the area of detector, μ_{eff} is defined by Eq. (16.16).

The measurement of the intensity of a back-reflected light from a tissue for different source–detector separations, r_{sd}, is the basis of the spatially resolved technique, which allows one to evaluate the absorption and the scattering coefficients using, for example, the analytical expressions (16.24) or (16.36), valid for highly scattering thick tissues.

When optical parameters of skin or mucosa are under investigation, the small source–detector separations should be used, where the diffusion approximation is not valid due to proximity to the tissue boundary. In that case, more sophisticated approximations of the RTT solution should be employed; in particular, a numerical solution of the inverse problem by the MC method is prospective. Such an SRR measurement can be implemented using multifiber probes with a number of fixed source–detector separations or using a CCD with a special optical system, allowing for the depth profiling of tissue optical properties if enough fibers or pixels are employed.

16.3.7
Optical Coherence Tomography

OCT is a newly developed modality that allows one to evaluate the scattering and absorption properties of tissue *in vivo* within the limits of an OCT penetration depth of 1–3 mm (see Tuchin, 2007). The use of OCT to measure the single-scattering coefficient of tissues μ_s has been described by Schmitt, Knüttel and Bonnar (1993). In its simplest form, this method assumes that backscattered light from a tissue decreases in the intensity according to

$$I_b \cong I_0 \exp[-2(\mu_a + \mu_s)z] \qquad (16.37)$$

where $(\mu_a + \mu_s)$ is the total attenuation coefficient and $2z$ is the round-trip distance of light backscattered at a depth z. For most tissues in the NIR, $\mu_a \ll \mu_s$; thus, μ_s can be estimated roughly as

$$\mu_s \cong \frac{1}{2}z \left\{ \ln \left[\frac{I_b(z)}{I_0} \right] \right\} \qquad (16.38)$$

or as the gradient of a graph $\ln[I_b(z)/I_0]$ versus z. Experimental data for tissue OCT images show that the logarithmically scaled average for multiple in-depth scans' backscattered intensity, $\ln[I_b(z)]$, decays exponentially; thus, by performing a linear regression on this curve, the scattering coefficient can be determined. More comprehensive algorithms accounting for multiple scattering effects and properties of a small-angle scattering phase function are available in the literature (Tuchin, 2007; Levitz *et al.*, 2004).

16.3.8
Measurement of the Scattering Phase Function

Direct measurement of the scattering phase function $p(\theta)$ is important for the choice of an adequate model for the tissue being examined. The scattering phase function is usually determined from goniophotometric measurements (GPMs) in relatively thin tissue samples (Jacques, Alte and Prahl, 1987). Measured scattering intensity angular dependence is approximated either by the HGPF (see Eq. (16.13)) or by a set of HGPFs, with each function characterizing the type of scatterers and specific contribution to the angular dependence (Marchesini *et al.*, 1989). In the limiting case of a two-component model of a medium containing large and small (compared with the wavelength) scatterers, the angular dependence is represented in the form of anisotropic and isotropic components. Other approximating functions are equally useful, for example, those obtained from the Rayleigh–Gans approximation, ensuing from the Mie theory, or a two-parameter GKPF (HGPF is a special simpler case of this phase function) (see Tuchin, 2002, 2007). Some of these types of approximations were used to find the dependence of the scattering anisotropy factor g for dermis and epidermis on the wavelengths in the range of 300–1300 nm, which proved to coincide fairly well with the empirical formula (van Gemert *et al.*, 1989),

$$g_e \sim g_d \sim 0.62 + \lambda \times 0.29 \times 10^{-3} \tag{16.39}$$

on the assumption of a 10% contribution of isotropic scattering (at least in the spectral range of 300–630 nm). The wavelength λ is given in nanometers.

16.3.9
Human Tissue Optical Properties Analysis

The above-discussed methods and techniques were successfully applied for estimation of optical properties of a wide number of tissues. Measurements done *in vitro, ex vivo,* and *in vivo* for human tissues are summarized in Table 16.1. It is clearly seen that attention was primarily focused on female breast and head/brain optical properties investigations because of great importance and perspectives of optical mammography and optical monitoring and treatment of mental diseases. Skin and underlying tissues are also well studied.

The most detailed *in vitro* investigations of normal and coagulated brain tissues (gray matter, white matter, cerebellum, pons, and thalamus), as well as of native tumor tissues (astrocytoma WHO grade II and meningioma), using single-IS spectral measurements in the spectral range from 360 to 1100 nm and IMC algorithm for data processing are described by Yaroslavsky *et al.* (2002). All brain tissues under study shared qualitatively similar dependencies of the optical properties on the wavelength. The scattering coefficient decreased and the anisotropy factor increased with the wavelength, which can be explained by the lowering of the contribution of Rayleigh scattering and growing of the contribution of Mie scattering with the wavelength. The wavelength-dependent absorption coefficient behavior of all brain tissues resembled a mixture of oxy and deoxyhemoglobin absorption spectra. This means that in spite of careful preparation of the samples, it was not possible to remove all blood residuals from the tissue sections.

At the same time, the differences in the spectral characteristics of the brain tissues have been observed. For example, the total attenuation coefficients ($\mu_t = \mu_a + \mu_s$) of white matter are substantially higher than those of gray matter. The two brain stem tissues (pons and thalamus) also have different optical properties. The tumors are generally macroscopically less homogeneous than any normal tissues; thus, their scattering coefficients and anisotropy factors are slightly higher than those of normal gray matter. The same tendency of scattering coefficients growing is typical for breast tumors (carcinomas) (Peters *et al.*, 1990).

After coagulation, the values of absorption and scattering coefficients increased for all tissues. The extent of this increase, however, is different for each tissue type, and is characterized by factors from 2 to 5. It was shown (Yaroslavsky *et al.*, 2002) that a significant increase of both interaction coefficients is a result of substantial structure changes, caused mostly by tissue shrinkage and condensation, as well as collagen swelling and homogenization of the tissue. Tissue shrinkage caused by losing water at coagulation makes tissue more dense, which leads to increase of both scattering and absorption coefficients in the spectral range

Tab. 16.1 Optical properties of human tissues measured *in vitro, ex vivo*, and *in vivo* (root mean square (rms) values are given in parentheses).

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	$\mu_s{}'$, cm^{-1}	g	Remarks
			***In vitro* measurements**			
Aorta:						
Normal	308	33	–	77	–	*Postmortem* (6 hr), excised, in 4 °C saline, slab, water bath (85 °C), integrating sphere (IS) technique, IAD Cheong, Prahl and Welch, (1990)
Normal coagulated	308	44	–	270	–	
Fibrous plaque	308	24	–	81	–	
Fibrous plaque coagulated	308	34	–	272	–	
Normal	1064	0.53(0.09)	239(45)	23.9	0.9	*Postmortem*, slab, 70 °C water bath, 10 min, IS, IMC, goniophotometric measurements (GPM) Cheong, Prahl and Welch (1990)
Coagulated	1064	0.46(0.18)	293(73)	29.3	0.9	
Fibro-fatty	355	17.7	–	64.9	–	*Postmortem*, resected, slab (24 hr), photoacoustic (PA) Cheong, Prahl and Welch (1990)
	532	3.6	–	24.8	–	
	1064	0.09	–	7.7	–	
Normal	633*	0.52	316	41	0.87	*Postmortem*, slab, IS, GPM, *diffusion theory (DT), **IMC Cheong, Prahl and Welch (1990)
	1064**	0.5	239	23.9	0.9	
	1064*	0.7	–	22.4	–	
	1320**	2.2	233	23.3	0.9	
	1320*	4.3	–	17.8	–	

Normal	470	5.3(0.9)	–	42.6(6.0)	–	Thin sections (250 µm, intima and media), kept in saline, IS, DT Cilesiz and Welch (1993); corrected data (see Tuchin (1994), p. 379)
	476	5.1(0.9)	–	41.9(5.9)	–	
	488	4.5(0.9)	–	39.9(5.6)	–	
	514.5	3.7(0.9)	–	36.9(5.4)	–	
	580	2.8(0.9)	–	31.1(4.9)	–	
	600	2.6(0.9)	–	29.6(4.7)	–	
	633	2.6(0.9)	–	27.4(4.4)	–	
	1064	2.7(0.5)	–	15.5(2.8)	–	
Adventitia	476	18.1	267	69.4	0.74	Frozen sections, IS, DT Keijzer *et al.* (1989)
	580	11.3	217	49.9	0.77	
	600	6.1	211	46.4	0.78	
	633	5.8	195	37.1	0.81	
	1064	2.0	484	–	0.97	Double IS (DIS), DT Roggan *et al.* (1995)
Intima	476	14.8	237	45.0	0.81	Frozen sections, IS, DT Keijzer *et al.* (1989)
	580	8.9	183	34.8	0.81	
	600	4.0	178	33.8	0.81	
	633	3.6	171	25.7	0.85	
	1064	2.3	165	–	0.97	DIS, DT Roggan *et al.* (1995)
Media	476	7.3	410	45.1	0.89	Frozen sections, IS, DT Keijzer *et al.* (1989)
	580	4.8	331	33.1	0.90	
	600	2.5	323	35.5	0.89	
	633	2.3	310	31.0	0.90	
	1064	1.0	634	–	0.96	DIS, DT Roggan *et al.* (1995)

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	$\mu_s{'}$, cm^{-1}	g	Remarks
Bladder:						
Integral	633	1.40	88.0	3.52	0.96	Excised, kept in saline Cheong, Prahl and Welch (1990)
Integral	633	1.40	29.3	2.64	0.91	DIS, DT
Mucous	1064	0.7	7.5	–	0.85	Müller and Roggan (1995); Roggan *et al.* (1995)
Wall	1064	0.9	54.3	–	0.85	
Integral	1064	0.4	116	–	0.90	
Blood:						
HbO_2 (Hct = 0.41)	665	1.30	1246	6.11	0.995	Whole blood; absorbance, radial reflectance, and/or GPM; Mie theory, transport theory, or IMC
HbO_2 (Hct = 0.41)	685	2.65	1413	14.13	0.990	
HbO_2 (Hct = 0.41)	960	2.84	505	3.84	0.992	
HbO_2 (Hct = 0.4)	810	4.5	–	6.6	–	Cheong, Prahl and Welch, (1990);
HbO_2 (Hct = 0.4)	1064	3.0	–	3.4	–	Roggan *et al.* (1995)
Hb (Hct = 0.41)	960	16.8	668	5.08	0.992	
Hb (Hct = 0.4)	810	4.5	–	3.9	–	
Hb (Hct = 0.4)	1064	0.3	–	6.6	–	
Hct = 0.47 (partially oxygenated)	450	381	2940	8.3	0.9972	Whole blood; IAD and Beer's law; data by Jacques Cheong, Prahl and Welch (1990)
	488	133	3190	4.0	0.9987	
	514	116	3320	4.1	0.9988	
	577	301	3140	7.3	0.9977	
	630	14.3	3660	8.9	0.9976	
	760	15.5	2820	7.9	0.9972	

(Hct = 0.45–0.46, oxygenation >98%)	633	15.5	644.7	–	0.982	DIS, Henyey–Greenstein phase function (HGPF), IMC, whole blood Yaroslavsky *et al.* (1999)
	710	4(0.8)	737(75)	–	0.986(0.006)	
	810	6.5(0.5)	690(80)	–	0.989(0.002)	
	910	8.9(0.4)	649(25)	–	0.992(0.002)	
	1010	8.3(0.4)	645(25)	–	0.992(0.001)	
	1110	4.2(0.3)	630(20)	–	0.993(0.001)	
	1210	5.5(0.5)	654(20)	–	0.995(0.001)	
(Hct = 0.421, oxygenation >99%)	260	375.5(9.0)	631.5(57.6)	136.4(28.0)	0.784(0.030)	IS, fresh erythrocytes from a healthy blood donor diluted in PBS, pH 7.4, hemoglobin concentration 129 g l^{-1}, the temperature was kept constant at 20 °C, turbulence-free cuvette with a laminar flow and a sample thickness of 116 μm, constant wall share rate of 600 s^{-1}; in the wavelength region around 415 nm a cuvette of 40 μm in thickness was used; Reynolds–McCormick phase function ($\alpha = 1.7$), IMC, data were presented by Friebel *et al.* (2006).
	350	368.1	559.5	82.5	0.852	
	415	782.5(62.9)	390.3(61.2)	129.5(17.0)	0.668(0.008)	
	520	120.4(6.9)	766.2(42.4)	24.9(7.6)	0.967(0.009)	
	540	232.3	655.6	35.8	0.945	
	555	178.9	709.3	33.0	0.953	
	575	231.6	658.0	31.7	0.952	
	585	160.2(10.3)	751.7(46.1)	33.5(9.7)	0.955(0.007)	
	630	2.51(0.09)	894.6(28.6)	22.3(3.3)	0.975(0.004)	
	670	1.22	892.3	21.5	0.976	
	780	2.85	821.5	20.5	0.975	
	800	3.27(0.12)	809.9(66.4)	20.2(5.4)	0.975(0.003)	
	830	4.90	798.7	20.1	0.975	
	870	5.10	784.4	20.1	0.974	
	950	6.15(0.35)	712.0(69.8)	20.8(2.7)	0.971(0.002)	
	1050	4.91(0.12)	661.3(12.8)	19.91(0.67)	0.9699(0.0006)	
	1100	3.74	639.5	18.85	0.970	

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	$\mu_s{}'$, cm^{-1}	g	Remarks
Brain:						
Glioma	415	16.6	–	6		Sterenborg *et al.* (1989); data from graphs
(male, 65 yr,	488	12.5	–	3	–	
4-hr *postmortem*)	630	3.0	–	3	–	
	800–1100	≈1.0	–	>1–2	–	
Gray matter	514	19.5	–	85	–	
(male, 71 yr,	585	14.5	–	63	–	
24-hr *postmortem*)	630	4.3	–	52	–	
	800–1100	≈1.0	–	45–20	–	
Melanoma	585	2	–	158	–	
(male 71 yr,	630	20.0	–	75	–	
24-hr *postmortem*)	800	8.0	–	40	–	
	900	4.0	–	30	–	
	1100	2.0	–	25	–	
White matter	415	2.1	–	24	–	
(female, 32 yr, 24-hr *postmortem*)	488	1.0	–	60	–	
	630	0.2	–	32	–	
	800–1100	0.2–0.3	–	40–20	–	
White matter	488	2.7	–	25	–	
(female, 63 yr, 30-hr *postmortem*)	630	0.9	–	22	–	
	800–1100	1.0–1.5	–	20–10	–	

Gray matter	633	2.7(2)	354(37)	20.6(2)	0.94(0.004)	Freshly resected slabs Cheong, Prahl and Welch (1990)
	1064	5.0(5)	134(14)	11.8(9)	0.90(0.007)	
White matter	633	2.2(2)	532(41)	91(5)	0.82(0.01)	
	1064	3.2(4)	469(34)	60.3(2.5)	0.87(0.007)	
Gray matter ($n = 7$)	360	3.33(2.19)	141.3(42.6)	–	0.818(0.093)	DIS, IMC Yaroslavsky *et al.* (1996); Schwarzmaier *et al.* (1997)
	640	0.17(0.26)	90.1(32.5)	–	0.89(0.04)	
	1060	0.56(0.7)	56.8(18.0)	–	0.90(0.05)	
Gray matter coagulated ($n = 7$)	360	9.39(1.70)	426(122)	–	0.868(0.031)	DIS, IMC; 2 hr, 80 °C Yaroslavsky *et al.* (1996); Schwarzmaier *et al.* (1997)
	740	0.45(0.27)	–	–	–	
	1100	1.0(0.45)	179.8(32.6)	–	0.954(0.001)	
White matter ($n = 7$)	360	2.53(0.55)	402.0(91.8)	–	0.702(0.093)	DIS, IMC (Yaroslavsky *et al.*, 1996; Schwarzmaier *et al.*, 1997)
	640	0.8(0.2)	408.2(88.5)	–	0.84(0.05)	
	860	0.97(0.4)	353.1(68.1)	–	0.871(0.028)	
	1060	1.08(0.51)	299.5(70.1)		0.889(0.010)	
White matter coagulated ($n = 7$)	360	8.3(3.65)	604.2(131.5)	–	0.800(0.089)	DIS, IMC; 2 hr, 80 °C Yaroslavsky *et al.* (1996); Schwarzmaier *et al.* (1997)
	860	1.7(1.3)	417.0(272.5)	–	0.922(0.025)	
	1060	2.15(1.34)	363.3(226.8)	–	0.930(0.015)	

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	$\mu_s{}'$, cm^{-1}	g	Remarks
Gray matter ($n = 7$)	450	0.7	117	14.04	0.88	IS, IMC, quasi-Newton inverse algorithm, HGPF; hemoglobin-free cryosections (<48-hr *postmortem*): gray matter – 100–200 μm; white matter – 80–150 μm; coagulation: saline bath 80 °C coagulation: saline bath 80 °C, 2 hr Yaroslavsky *et al.* (2002)
	510	0.4	106	12.72	0.88	
	630	0.2	90	9.9	0.89	
	670	0.2	84	8.4	0.90	
	1064	0.5	57	5.7	0.90	
White matter ($n = 7$)	450	1.4	420	92.4	0.78	
	510	1.0	426	80.94	0.81	
	630	0.8	409	65.44	0.84	
	670	0.7	401	60.15	0.85	
	850	1.0	342	41	0.88	
	1064	1.0	296	32.56	0.89	
White matter coagulated ($n = 7$)	850	0.9	300	36.0	0.88	
	1064	0.1	270	29.7	0.89	
Astrocytoma (grade II WHO, $n = 4$)	400	18.8(11.3)	198.4(55.6)	–	0.93(0.03)	IS, IMC, quasi-Newton inverse algorithm, HGPF; hemoglobin-free cryosections of normal tissues (<48-hr *postmortem*): cerebellum, gray matter, pons, and thalamus – 100 –200 μm; white matter –80 –150 μm; and tumors excised from patients of ≈300 μm in thickness; coagulation: saline bath 80 °C, 2 hr Yaroslavsky *et al.* (2002)
	490	2.5(0.9)	158.5(53.7)	–	0.96(0.02)	
	600	1.2(0.7)	132.4(49.0)	–	0.96(0.02)	
	700	0.5(0.3)	113.2(41.8)	–	0.96(0.02)	
	800	0.7(0.2)	96.7(41.8)	–	0.96(0.01)	
	900	0.3(0.2)	86.4(34.6)	–	0.96(0.01)	
	1000	0.5(0.3)	79.0(34.2)	–	0.96(0.01)	
	1100	0.6(0.2)	73.8(29.6)	–	0.96(0.01)	
Cerebellum ($n = 7$)	400	4.7(0.8)	276.7(19.1)	–	0.80(0.03)	
	500	1.4(0.2)	277.5(32.6)	–	0.85(0.02)	
	600	0.8(0.2)	272.1(12.3)	–	0.87(0.02)	
	700	0.6(0.1)	266.8(12.1)	–	0.89(0.01)	
	800	0.6(0.1)	250.3(17.2)	–	0.90(0.01)	
	900	0.7(0.1)	229.6(15.8)	–	0.90(0.01)	
	1000	0.8(0.1)	215.4(14.7)	–	0.90(0.01)	
	1100	0.7(0.1)	202.1(13.9)	–	0.90(0.01)	

Cerebellum coagulated (*n* = 7)	400	19.3(7.7)	560.0(25.5)	–	0.61(0.01)
	500	5.1(1.7)	512.2(47.8)	–	0.77(0.02)
	600	2.9(1.4)	458.2(65.6)	–	0.78(0.01)
	700	1.7(0.4)	489.9(70.1)	–	0.85(0.01)
	800	1.1(0.2)	458.2(54.0)	–	0.87(0.02)
	900	1.1(0.3)	458.2(65.6)	–	0.89(0.02)
	1000	1.0(0.4)	419.1(49.4)	–	0.90(0.03)
	1100	1.1(0.5)	428.5(40.0)	–	0.91(0.03)
Gray matter (*n* = 7)	400	2.6(0.6)	128.5(18.4)	–	0.87(0.02)
	500	0.5(0.2)	109.9(13.0)	–	0.88(0.01)
	600	0.3(0.1)	94.1(13.5)	–	0.89(0.02)
	700	0.2(0.1)	84.1(12.0)	–	0.90(0.02)
	800	0.2(0.1)	77.0(11.0)	–	0.90(0.02)
	900	0.3(0.2)	67.3(9.6)	–	0.90(0.02)
	1000	0.6(0.3)	61.6(5.7)	–	0.90(0.02)
	1100	0.5(0.3)	55.1(6.5)	–	0.90(0.02)
Gray matter coagulated (*n* = 7)	400	7.5(0.4)	258.6(18.8)	–	0.78(0.04)
	500	1.8(0.2)	326.5(7.7)	–	0.85(0.03)
	600	0.7(0.1)	319.0(15.2)	–	0.87(0.03)
	700	0.7(0.1)	319.0(7.5)	–	0.88(0.03)
	800	0.8(0.1)	252.7(18.3)	–	0.87(0.02)
	900	0.9(0.1)	214.6(10.3)	–	0.87(0.02)
	1000	1.4(0.2)	191.0(18.7)	–	0.88(0.03)
	1100	1.5(0.2)	186.6(13.5)	–	0.88(0.03)

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	μ_s', cm^{-1}	g	Remarks
Meningioma ($n = 6$)	410	4.1(0.5)	197.4(19.8)	–	0.88(0.02)	
	490	1.3(0.2)	188.2(18.8)	–	0.93(0.01)	
	590	0.7(0.2)	171.1(12.7)	–	0.95(0.01)	
	690	0.3(0.1)	155.5(15.6)	–	0.95(0.01)	
	790	0.2(0.1)	141.3(14.2)	–	0.96(0.01)	
	910	0.2(0.1)	116.8(8.6)	–	0.95(0.01)	
	990	0.4(0.2)	163.5(15.3)	–	0.96(0.01)	
	1100	0.6(0.2)	133.7(19.2)	–	0.97(0.01)	
Pons ($n = 7$)	400	3.1(0.7)	163.5(15.3)	–	0.89(0.02)	
	500	0.9(0.3)	133.7(19.2)	–	0.91(0.01)	
	600	0.6(0.2)	109.4(18.5)	–	0.91(0.01)	
	700	0.5(0.2)	93.5(20.9)	–	0.91(0.01)	
	800	0.6(0.3)	83.6(21.0)	–	0.91(0.01)	
	900	0.7(0.3)	74.8(18.7)	–	0.92(0.01)	
	1000	1.0(0.4)	69.9(17.5)	–	0.91(0.01)	
	1100	0.9(0.4)	64.0(17.8)	–	0.92(0.01)	
Pons coagulated ($n = 7$)	410	17.2(1.6)	685.7(63.7)	–	0.85(0.02)	
	510	8.5(0.8)	627.5(73.6)	–	0.89(0.01)	
	610	7.7(0.5)	510.5(70.5)	–	0.89(0.01)	
	710	6.9(0.6)	402.5(67.7)	–	0.89(0.01)	
	810	6.5(0.6)	329.7(55.4)	–	0.89(0.01)	
	910	5.9(1.0)	276.0(46.4)	–	0.88(0.01)	
	1010	5.7(1.0)	241.6(34.4)	–	0.88(0.01)	
	1100	6.5(0.9)	221.1(31.5)	–	0.88(0.01)	

Thalamus ($n = 7$)	410	3.2(1.0)	146.7(49.4)	–	0.86(0.03)
	510	0.9(0.3)	188.7(31.9)	–	0.87(0.03)
	610	0.6(0.2)	176.3(34.5)	–	0.88(0.02)
	710	0.5(0.3)	169.0(28.7)	–	0.89(0.03)
	810	0.7(0.3)	158.5(35.3)	–	0.89(0.02)
	910	0.7(0.3)	155.4(22.3)	–	0.90(0.02)
	1010	0.8(0.3)	139.3(34.9)	–	0.90(0.02)
	1100	0.8(0.3)	146.0(36.6)	–	0.91(0.02)
Thalamus coagulated ($n = 7$)	400	15.0(3.3)	391.1(56.1)	–	0.83(0.04)
	500	4.2(0.9)	399.9(67.7)	–	0.90(0.01)
	600	1.6(0.6)	365.7(43.2)	–	0.92(0.01)
	700	1.4(0.3)	327.0(30.6)	–	0.92(0.01)
	800	1.1(0.3)	286.0(33.8)	–	0.93(0.01)
	900	1.1(0.3)	267.4(31.6)	–	0.93(0.01)
	1000	1.4(0.4)	233.8(39.7)	–	0.93(0.01)
	1100	1.5(0.4)	223.6(32.1)	–	0.94(0.01)
White matter ($n = 7$)	400	3.1(0.2)	413.5(21.4)	–	0.75(0.03)
	500	0.9(0.1)	413.5(43.9)	–	0.80(0.02)
	600	0.8(0.1)	413.5(21.4)	–	0.83(0.02)
	700	0.8(0.1)	393.1(30.9)	–	0.85(0.02)
	800	0.9(0.1)	364.5(28.6)	–	0.87(0.01)
	900	1.0(0.1)	329.5(35.0)	–	0.88(0.01)
	1000	1.2(0.2)	305.4(15.9)	–	0.88(0.01)
	1100	1.0(0.2)	283.2(22.2)	–	0.88(0.01)

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	$\mu_s{}'$, cm^{-1}	g	Remarks
White matter coagulated ($n = 7$)	410	8.7(1.7)	568.7(111.9)	–	0.83(0.03)	
	510	2.9(0.6)	513.2(116.9)	–	0.87(0.02)	
	610	1.7(0.4)	500.2(129.9)	–	0.90(0.02)	
	710	1.4(0.5)	475.2(108.3)	–	0.91(0.01)	
	810	1.5(0.5)	440.0(114.3)	–	0.92(0.01)	
	910	1.7(0.6)	407.4(92.8)	–	0.93(0.01)	
	1010	1.9(0.6)	367.7(95.5)	–	0.93(0.01)	
	1100	2.4(0.5)	358.4(81.6)	–	0.93(0.01)	
Gray matter ($n = 25$)	400	9.778	–	25.878	–	IS, IAD, fixed the anisotropy factor $g = 0.85$ and the refractive index $n = 1.40$ were assumed for every wavelength and for every sample; brain tissue samples were acquired during open craniotomy for tumor resection or temporal lobectomy, hemoglobin-free cryosections with thickness from 0.22 to 1.25 mm were studied, measurements were done at 25 °C, pH 7.4; data were presented by Gebhart, Lin and Mahadevan-Jansen (2006).
	428	16.722	–	26.709	–	
	488	2.272	–	15.957	–	
	550	2.955	–	13.315	–	
	632	0.925	–	10.370	–	
	670	0.809	–	9.480	–	
	750	0.599	–	8.481	–	
	830	0.485	–	7.707	–	
	900	0.503	–	7.055	–	
	1000	0.585	–	6.059	–	
	1064	0.502	–	5.333	–	
	1150	0.815	–	5.070	–	
	1300	0.894	–	4.560	–	

White matter ($n = 19$)	400	9.134	–	88.611	–
	428	15.417	–	80.905	–
	488	1.869	–	70.112	–
	550	2.584	–	62.383	–
	632	0.801	–	53.179	–
	670	0.711	–	50.067	–
	750	0.649	–	45.061	–
	830	0.626	–	40.634	–
	900	0.684	–	37.607	–
	1000	0.883	–	32.603	–
	1064	0.752	–	30.161	–
	1150	1.135	–	27.951	–
	1300	1.274	–	24.250	–
Glioma ($n = 39$)	400	12.393	–	39.009	–
	428	16.124	–	37.076	–
	488	2.592	–	28.933	–
	550	2.768	–	25.300	–
	632	0.846	–	21.068	–
	670	0.741	–	19.608	–
	750	0.679	–	17.343	–
	830	0.662	–	15.481	–
	900	0.707	–	14.138	–
	950	0.768	–	13.646	–
	1000	0.938	–	11.588	–
	1064	0.822	–	10.344	–
	1150	1.231	–	9.654	–
	1300	1.412	–	8.523	–

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	μ_s', cm^{-1}	g	Remarks
Breast (female):						
Fatty normal ($n = 23$)	749	0.18(0.16)	8.48(3.43)	–	–	Excised, kept in saline, 37 °C
	789	0.08(0.10)	7.67(2.57)	–	–	Troy, Page and Sevick-Muraca (1996)
	836	0.11(0.10)	7.27(2.40)	–	–	
Fibrous normal ($n = 35$)	749	0.13(0.19)	9.75(2.27)	–	–	
	789	0.06(0.12)	8.94(2.45)	–	–	
	836	0.05(0.08)	8.10(2.21)	–	–	
Infiltrating carcinoma ($n = 48$)	749	0.15(0.14)	10.91(5.59)	–	–	
	789	0.04(0.08)	10.12(5.05)	–	–	
	836	0.10(0.19)	9.10(4.54)	–	–	
Mucinous carcinoma ($n = 3$)	749	0.26(0.20)	–	6.15(2.44)	–	
	789	0.016(0.072)	–	5.09(2.42)	–	
	836	0.023(0.108)	–	4.78(3.67)	–	
Ductal carcinoma *in situ* ($n = 5$)	749	0.076(0.068)	–	13.10(2.85)	–	
	789	0.023(0.034)	–	12.21(2.45)	–	
	836	0.039(0.068)	–	10.46(2.65)	–	
Glandular tissue ($n = 3$)	540	3.58(1.56)	–	24.4(5.8)	–	Homogenized tissue
	700	0.47(0.11)	–	14.2(3.0)	–	Peters *et al.* (1990)
	900	0.62(0.05)	–	9.9(2.0)	–	

Fatty tissue ($n = 7$)	540	2.27(0.57)	–	10.3(1.9)	–	
	700	0.70(0.08)	–	8.6(1.3)	–	
	900	0.75(0.08)	–	7.9(1.1)	–	
Fibrocystic ($n = 8$)	540	1.64(0.66)	–	21.7(3.3)	–	
	700	0.22(0.09)	–	13.4(1.9)	–	
	900	0.27(0.11)	–	9.5(1.7)	–	
Fibroadenoma ($n = 6$)	540	4.38(3.14)	–	11.1(3.0)	–	
	700	0.52(0.47)	–	7.2(1.7)	–	
	900	0.72(0.53)	–	5.3(1.4)	–	
Carcinoma ($n = 9$)	540	3.07(0.99)	–	19.0(5.1)	–	
	700	0.45(0.12)	–	11.8(3.1)	–	
	900	0.50(0.15)	–	8.9(2.6)	–	
Fatty tissue	625	0.06(0.02)	–	14.3(2.1)	–	Das, Liu and Alfano, (1997)
Benign tumor	625	0.33(0.06)	–	3.8(0.3)	–	
Invasive ductal carcinoma ($n = 10$; 9 in the age group 55–65 yr and 1–35 yr)	450	2.55(0.30)	–	31.5(2.5)	–	Spatially resolved reflectance (SRR); DT; source–detector separation, $r_{sd} > 1.2$ mm; fibers with core diameter 400 μm; tissue slices of thickness ~10 mm; GPM, tissue slices of 20 μm, HGPF; *double HGPF Ghosh *et al.* (2001)
	500	2.22(0.22)	–	29.5(2.2)	–	
	550	2.13(0.23)	–	28.4(2.0)	–	
	600	1.90(0.19)	–	26.8(1.8)	–	
	630	1.64(0.17)	–	26.2(1.5)	–	
	650	1.48(0.15)	–	25.7(1.3)	–	
	633	–	–	–	0.96(0.01)	
	633	–	–	–	0.86(0.02)*	

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	$\mu_s{\prime}$, cm^{-1}	g	Remarks
Adjacent healthy tissue ($n = 10$; 9 in the age group 55–65 yr and 1–35 yr)	450	1.45(0.22)	–	21.7(2.1)	–	SRR; DT; $r_{sd} > 1.2$ mm; fibers with core diameter 400 μm; tissue slices of thickness ~10 mm;
	500	1.24(0.21)	–	20.1(1.8)	–	
	550	1.16(0.26)	–	18.2(1.6)	–	
	600	1.00(0.12)	–	16.4(1.5)	–	
	630	0.82(0.07)	–	15.7(1.3)	–	
	650	0.74(0.08)	–	15.3(1.2)	–	GPM, tissue slices of 20 μm; HGPF; *double HGPF Ghosh *et al.* (2001)
	633	–	–	–	0.88(0.01)	
	633	–	–	–	0.76(0.01)*	
Colon:						
Muscle	1064	3.3	238	–	0.93	Roggan *et al.* (1995)
Submucous	1064	2.3	117	–	0.91	
Mucous	1064	2.7	39	–	0.91	
Integral	1064	0.4	261	–	0.94	
Esophagus	633	0.4	–	12	–	2.5-mm slab Cheong, Prahl and Welch (1990)
Esophagus (mucous)	1064	1.1	83	–	0.86	Roggan *et al.* (1995)
Fat:						
Abdominal	1064	3.0	37	–	0.91	Roggan *et al.* (1995)
Subcutaneous	1064	2.6	29	–	0.91	

Gallstones:						
Porcinement	351	102(16)	–	–	–	Dehydrated, embedded in plastic, and sliced in 1-mm slab, pulsed photothermal radiometry technique Cheong, Prahl and Welch (1990)
	488	179(28)	–	–	–	
	580	125(29)	–	–	–	
	630	85(11)	–	–	–	
	1060	121(12)	–	–	–	
Cholesterol	351	88(7)	–	–	–	
	488	62(15)	–	–	–	
	580	36(7)	–	–	–	
	630	44(10)	–	–	–	
	1060	60(9)	–	–	–	
Head (adult):						
Dura mater ($n = 8$), *postmortem*, <24 h	400	3.08(0.15)	–	22.35(0.89)	–	IS, IAD; excised tissue slabs; in the spectral ranges 480–550 and 600–700 nm: $\mu_s' = 4.54 \times 10^4 \lambda^{-1.23}$, $[\lambda]$ in nanometers Genina *et al.* (2005)
	450	1.51(0.08)	–	22.89(0.92)	–	
	500	1.09(0.05)	–	21.60(0.86)	–	
	550	1.10(0.05)	–	18.48(0.74)	–	
	600	0.80(0.04)	–	17.11(0.68)	–	
	650	0.70(0.04)	–	15.51(0.62)	–	
	700	0.74(0.04)	–	13.99(0.56)	–	

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	$\mu_s{}'$, cm^{-1}	g	Remarks
Scalp ($n = 3$)	805	0.52(0.04)	–	14.09(1.74)	–	Adult scalp *postmortem* (<12 hr), excised, slab, IS, IAD; data averaged for three tissue samples with thicknesses of 6 ± 0.5 mm, 3.5 ± 0.15 mm, and 3.5 ± 0.12 mm Bashkatov *et al.* (2006)
	1000	0.33(0.03)	–	16.83(2.77)	–	
	1100	0.19(0.04)	–	17.10(2.69)	–	
	1200	0.65(0.04)	–	16.70(2.89)	–	
	1300	0.50(0.07)	–	14.70(2.59)	–	
	1400	1.98(0.31)	–	14.28(3.69)	–	
	1430	2.19(0.29)	–	13.15(3.07)	–	
	1600	1.43(0.22)	–	14.16(3.41)	–	
	1700	1.87(0.28)	–	14.71(3.51)	–	
	1800	1.73(0.22)	–	13.36(2.91)	–	
	1900	2.57(0.28)	–	12.15(3.05)	–	
	2000	2.09(0.29)	–	12.00(2.91)	–	
Skull bone ($n = 8$)	801	0.11(0.02)	–	19.48(1.52)	–	Adult head *postmortem* (24 hr), excised, slab from the occipital part, IS, IAD; data averaged for 8 tissue samples $\mu_s' = 1.53 \times 10^3 \lambda^{-0.65}$, [λ] in nanometers (spectral range from 1130 to 1910 nm is excluded) Bashkatov *et al.* (2006)
	980	0.23(0.03)	–	17.38(1.01)	–	
	1100	0.16(0.03)	–	15.92(0.76)	–	
	1200	0.67(0.07)	–	16.77(0.85)	–	
	1300	0.54(0.05)	–	14.78(0.80)	–	
	1400	2.43(0.24)	–	17.22(1.73)	–	
	1465	3.33(0.31)	–	16.84(1.88)	–	
	1600	2.47(0.40)	–	15.84(3.05)	–	
	1700	2.77(0.46)	–	16.12(3.72)	–	
	1800	2.97(0.62)	–	15.42(3.98)	–	
	1930	4.97(1.52)	–	10.92(2.17)	–	
	2000	4.47(1.18)	–	11.48(2.01)	–	

Heart:						
Endocard	1060	0.07	136	–	0.97	Excised, kept in saline, Cheong, Prahl and Welch (1990)
Epicard	1060	0.35	167	–	0.98	Roggan *et al.* (1995)
Myocard	1060	0.3	177.5	–	0.96	
Epicard	1060	0.21	127.1	–	0.93	
Aneurysm	1060	0.4	137	–	0.98	
Trabecula	1064	1.4	424	–	0.97	
Kidney:						
Pars convalescent	1064	2.4	72	–	0.86	Roggan *et al.* (1995)
Medulla renal	1064	2.1	77	–	0.87	
Liver	515	18.9(1.7)	285(20)	–	–	Frozen sections
	635	2.3(1.0)	313(136)	100	0.68	Marchesini *et al.* (1989)
	1064	0.7	356	–	0.95	
Lung	515	25.5(3.0)	356(39)	–	–	
	635	8.1(2.8)	324(46)	81	0.75	
	1064	2.8	39	–	0.91	
Muscle	515	11.2(1.8)	530(44)	–	–	
	1064	2.0	215	–	0.96	
Meniscus	360	13	–	108	–	Frozen, thawed, slab
	400	4.6	–	67	–	Cheong, Prahl and Welch, (1990)
	488	1	–	30	–	
	514	0.73	–	26	–	
	630	0.36	–	11	–	
	800	0.52	–	5.1	–	
	1064	0.34	–	2.6	–	

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	μ_s', cm^{-1}	g	Remarks
Prostate:						
Normal	850	0.6(0.2)	100(20)	–	0.94(0.02)	Shock frozen sections of 60–500 μm, 0.5–3-hr *postmortem*; coagulation in water bath (75 °C, 10 min) Roggan *et al.* (1995)
	980	0.4(0.2)	90(20)	–	0.95(0.02)	
	1064	0.3(0.2)	80(20)	–	0.95(0.02)	
Coagulated	850	7.0(0.2)	230(30)	–	0.94(0.02)	
	980	5.0(0.2)	190(30)	–	0.95(0.02)	
	1064	4.0(0.2)	180(30)	–	0.95(0.02)	
Normal	1064	1.5(0.2)	47(13)	0.64	0.862	Freshly excised, slab, water bath (70 °C, 10 min) Cheong, Prahl and Welch (1990)
Coagulated	1064	0.8(0.2)	80(12)	1.12	0.861	
Sclera ($n = 5$)	404	5.00(0.50)	–	81.40(8.14)	–	IS, IAD; excised tissue slabs, <24-hr *postmortem*, stored in saline at 4 °C; measurements at room temperature; $\mu_s' = 8.95 \times 10^4 \lambda^{-1.16}$, [λ] in nanometers Tuchin (2007)
	499	2.96(0.30)	–	65.17(6.52)	–	
	599	1.95(0.19)	–	53.16(5.32)	–	
	699	1.67(0.17)	–	44.15(4.42)	–	
	749	1.65(0.17)	–	40.10(4.01)	–	
	799	1.58(0.16)	–	37.64(3.76)	–	
Skin:						
	250	1150	2600	260	0.9	Data from graphs; μ_s' is calculated van Gemert *et al.* (1989)
	308	600	2400	240	0.9	
	337	330	2300	230	0.9	
	351	300	2200	220	0.9	
	400	230	2000	200	0.9	

Epidermis	250	1000	2000	616	0.69	Data from graphs; μ_s' and g are calculated using van Gemert *et al.* (1989)
	308	300	1400	407	0.71	
	337	120	1200	338	0.72	
	351	100	1100	306	0.72	
	415	66	800	206	0.74	
	488	50	600	143	0.76	
	514	44	600	139	0.77	
	585	36	470	99	0.79	
	633	35	450	88	0.80	
	800	40	420	62	0.85	
Dermis	250	35	833	257	0.69	Data from graphs; values are transformed in accordance with data for $\lambda = 633$ nm (bloodless tissue, hydration -85%), μ_s' and g are calculated Jacques, Alter and Prahl (1987); van Gemert *et al.* (1989)
	308	12	583	170	0.71	
	337	8.2	500	141	0.72	
	351	7	458	127	0.72	
	415	4.7	320	82	0.74	
	488	3.5	250	60	0.76	
	514	3	250	58	0.77	
	585	3	196	41	0.79	
	633	2.7	187.5	37	0.80	
	800	2.3	175	30	0.85	
Dermis	749	0.24(0.19)	–	23.1(0.75)	–	Frozen sections, DIS Troy, Page and Sevick-Muraca (1996)
	789	0.75(0.06)	–	22.8(1.29)	–	
	836	0.98(0.15)	–	15.9(2.16)	–	

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	μ_s', cm^{-1}	g	Remarks
Caucasian male skin ($n = 3$)	500	5.1	–	50	–	IS, IAD Chan *et al.* (1996)
	810	0.26	–	15.8	–	
Caucasian female skin ($n = 3$)	500	5.2	–	23.9	–	
	810	0.97	–	8.2	–	
Hispanic male ($n = 3$) skin	500	3.8	–	24.2		
	810	0.87	–	7.5	–	
Caucasian skin ($n = 21$)	400	3.76(0.35)	–	71.79(9.42)	–	IS, IAD; tissue slabs, 1–6 mm; postmortem; <24 hr after death; stored at 20 °C in saline; measurements at room temperature; in the spectral range 400–2000 nm: $\mu_s' = 1.1 \times 10^{12}\lambda^{-4} + 73.7\lambda^{-0.22}$, [λ] in nanometers Bashkatov *et al.* (2005a)
	600	0.69(0.13)	–	21.78(2.98)	–	
	800	0.43(0.11)	–	14.02(1.89)	–	
	1000	0.27(0.03)	–	16.83(2.77)	–	
	1200	0.54(0.04)	–	16.71(2.89)	–	
	1400	1.64(0.31)	–	14.28(3.69)	–	
	1600	1.19(0.22)	–	14.16(3.41)	–	
	1800	1.44(0.22)	–	13.36(2.91)	–	
	2000	1.74(0.29)	–	12.01(2.91)	–	
Epidermis	370	1.35(0.16)	–	11.56(1.25)	0.8	IS, IMC, slabs Salomatina *et al.* (2006)
	514	0.63(0.07)	–	6.67(0.66)	0.8	
	633	0.26(0.07)	–	4.76(0.45)	0.8	
	830	0.14(0.06)	–	3.56(0.35)	0.8	
	1064	0.02(0.02)	–	2.97(0.32)	0.8	
	1220	0.07(0.04)	–	2.63(0.31)	0.8	
	1270	0.06(0.04)	–	2.62(0.31)	0.8	
	1470	2.96(0.42)	–	3.08(0.45)	0.8	
	1570	1.01(0.20)	–	2.39(0.34)	0.8	

Dermis	370	0.98(0.14)	–	8.76(1.36)	0.8
	514	0.31(0.04)	–	4.32(0.41)	0.8
	633	0.15(0.02)	–	2.99(0.27)	0.8
	830	0.11(0.02)	–	2.15(0.23)	0.8
	1064	0.05(0.02)	–	1.80(0.21)	0.8
	1220	0.13(0.02)	–	1.65(0.20)	0.8
	1270	0.10(0.02)	–	1.63(0.20)	0.8
	1370	0.48(0.04)	–	1.66(0.19)	0.8
	1470	2.19(0.20)	–	2.13(0.21)	0.8
	1570	0.85(0.07)	–	1.65(0.19)	0.8
Subcutaneous fat	370	1.18(0.21)	–	5.27(0.69)	0.8
	514	0.47(0.07)	–	3.37(0.43)	0.8
	633	0.14(0.03)	–	2.54(0.30)	0.8
	830	0.10(0.02)	–	1.96(0.20)	0.8
	1064	0.07(0.02)	–	1.69(0.15)	0.8
	1220	0.15(0.03)	–	1.61(0.15)	0.8
	1270	0.10(0.03)	–	1.59(0.14)	0.8
	1370	0.27(0.04)	–	1.60(0.15)	0.8
	1470	1.08(0.18)	–	1.81(0.19)	0.8
	1570	0.43(0.07)	–	1.60(0.16)	0.8

(continued overleaf)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	$\mu_s{}'$, cm^{-1}	g	Remarks
Infiltrative basal cell carcinoma	370	0.68(0.08)	–	6.52(0.92)	0.8	
	514	0.26(0.06)	–	4.04(0.30)	0.8	
	633	0.15(0.05)	–	2.81(0.28)	0.8	
	830	0.09(0.04)	–	1.92(0.25)	0.8	
	1064	0.08(0.04)	–	1.26(0.09)	0.8	
	1220	0.17(0.09)	–	1.09(0.10)	0.8	
	1270	0.18(0.12)	–	1.05(0.11)	0.8	
	1370	0.69(0.27)	–	1.09(0.10)	0.8	
	1470	2.75(0.54)	–	1.66(0.32)	0.8	
	1570	1.12(0.31)	–	1.11(0.16)	0.8	
Nodular basal cell carcinoma	370	0.87(0.29)	–	4.62(0.61)	0.8	
	514	0.28(0.11)	–	3.27(0.18)	0.8	
	633	0.12(0.06)	–	2.27(0.12)	0.8	
	830	0.02(0.01)	–	1.49(0.07)	0.8	
	1064	0.00(0.00)	–	1.16(0.06)	0.8	
	1220	0.02(0.01)	–	1.01(0.04)	0.8	
	1270	0.01(0.01)	–	1.00(0.04)	0.8	
	1370	0.32(0.03)	–	1.03(0.06)	0.8	
	1470	1.86(0.16)	–	1.59(0.15)	0.8	
	1570	0.67(0.04)	–	1.06(0.08)	0.8	

Squamous cell carcinoma	370	0.94(0.20)	–	4.36(0.61)	0.8	
	514	0.32(0.04)	–	2.80(0.39)	0.8	
	633	0.13(0.02)	–	1.88(0.25)	0.8	
	830	0.05(0.02)	–	1.22(0.15)	0.8	
	1064	0.04(0.02)	–	0.88(0.12)	0.8	
	1220	0.11(0.03)	–	0.81(0.11)	0.8	
	1270	0.11(0.03)	–	0.78(0.11)	0.8	
	1370	0.43(0.05)	–	0.85(0.11)	0.8	
	1470	2.35(0.21)	–	1.44(0.23)	0.8	
	1570	0.92(0.12)	–	0.92(0.13)	0.8	
Spleen	1064	6.0	137	–	0.90	Roggan *et al.* (1995)
Stomach:						
Muscle	1064	3.3	29.5	–	0.87	Roggan *et al.* (1995)
Mucous	1064	2.8	732	–	0.91	
Integral	1064	0.8	128	–	0.91	
Tooth:						
Dentin	543	4	180	–	–	IS, GPM* Zijp and ten Bosch (1991, 1997); Fried *et al.* (1995)
Enamel	633	4	130	–	–	
	633	6.0*	1200*	672*	0.44*	
	543	<1	45	–	–	
	633	<1	25	–	–	

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	$\mu_s{}'$, cm^{-1}	g	Remarks
Dentin	543	3–4	280(84)	–	0.93(0.02)	GPM, double HGPF, fractions of isotropic scatterers are 0–2% for dentin and 60–35% for enamel; polished plane-parallel sections of 30–2000 μm Fried, *et al.* (1995)
	633	3–4	280(84)	–	0.93(0.02)	
	1053	3–4	260(78)	–	0.93(0.02)	
Enamel	543	<1	105(30)	–	0.96(0.02)	
	633	<1	60(18)	–	0.96(0.02)	
	1053	<1	15(5)	–	0.96(0.02)	
Enamel	200	≈10	≈450	–	–	Compiled data of a few papers, from graphs Fried (2003)
	300	≈5	≈270	–	–	
	400	≈1	≈150	–	–	
	500	<1	≈73	–	–	
	600	<1	≈64	–	–	
	700	<1	≈50	–	–	
	800	<1	≈33	–	–	
	1000	<1	≈16	–	–	
Dentin	2940	2200	–	–	–	Time-resolved radiometry, data from graphs Fried (2003)
	2790	1500	–	–	–	
	9600	6500	–	–	–	
	10600	800	–	–	–	
Enamel	2940	800	–	–	–	
	2790	400	–	–	–	
	9600	8000	–	–	–	
	10600	800	–	–	–	

Dentin	2790	988(111)	–	–	–	Transmission measurements Fried (2003)
	10300	1198(104)	–	–	–	
	10600	813(63)	–	–	–	
Enamel	2940	768(27)	–	–	–	
	2790	451(29)	–	–	–	
	10300	1168(49)	–	–	–	
	10600	819(62)	–	–	–	
Uterus	635	0.35(0.1)	394(91)	122	0.69	Frozen sections Marchesini *et al.* (1989)
Vein (femoral)	1064	3.2	487	–	0.97	Roggan *et al.* (1995)
			***Ex vivo* measurements**			
Aorta:						
Normal ($n = 4$)	1300	–	150–360	–	0.9–1	Optical coherence tomography (OCT); 4 hr of autopsy; $g_{\text{eff}} = \cos\theta_{\text{rms}}$, θ_{rms} – rms scattering angle, $g_{\text{eff}} \geq g$ Levitz *et al.* (2004)
Lipid rich ($n = 4$)	1300	–	0–200	–	0.6–1	
Fibrous ($n = 3$)	1300	–	50–400	–	0.6–1	
Fibrocalcific ($n = 3$)	1300	–	0–200	–	0.8–1	
Fat:						
Abdominal ($n = 2$)	360	3.12(0.78)	–	32.59(3.26)	–	IS, IAD; tissue slabs, <12 hr after surgery, stored at 4 °C in saline; measurements at room temperature; in the spectral range 600–1600 nm: $\mu_s' = 1.23 \times 10^3 \lambda^{-0.59}$, [λ] in nanometers Bashkatov *et al.* (2005b)
	500	2.37(0.59)	–	28.23(2.82)	–	
	600	1.90(0.47)	–	28.58(2.86)	–	
	800	1.87(0.47)	–	25.74(2.57)	–	
	1000	1.77(0.44)	–	20.14(2.01)	–	
	1200	1.79(0.45)	–	17.55(1.76)	–	
	1400	1.75(0.44)	–	17.15(1.71)	–	
	1600	1.47(0.37)	–	16.41(1.64)	–	
	1800	1.92(0.48)	–	17.40(1.74)	–	
	1900	2.48(0.62)	–	20.38(2.04)	–	
	2200	1.65(0.41)		18.95(1.89)	–	

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	μ_s', cm^{-1}	g	Remarks
Subcutaneous ($n = 6$)	400	2.26(0.24)	–	13.39(2.78)	–	IS, IAD; tissue slabs, 1–3 mm; <6 hr after surgery; stored at 20 °C in saline; measurements at room temperature; in the spectral range 600–1500 nm: $\mu_s' = 1.05 \times 10^3 \lambda^{-0.68}$, [$\lambda$] in nanometers Bashkatov *et al.* (2005a)
	600	1.18(0.02)	–	13.39(4.65)	–	
	800	1.07(0.11)	–	11.62(4.63)	–	
	1000	1.06(0.06)	–	9.39(3.32)	–	
	1200	1.06(0.07)	–	7.91(3.17)	–	
	1400	1.08(0.03)	–	7.51(3.31)	–	
	1600	0.89(0.04)	–	7.16(3.21)	–	
	1800	1.21(0.01)	–	7.50(3.48)	–	
	1900	1.62(0.06)	–	8.72(4.15)	–	
	2000	1.43(0.09)	–	8.24(4.03)	–	
Mucous of maxillary sinus at antritis	400	4.89(0.92)	–	36.01(6.41)	–	IS, IAD; tissue slabs, 1–2 mm; <6 hr after surgery; stored at 20 °C in saline; measurements at room temperature; in the spectral range 600–1300 nm: $\mu_s' = 4.4 \times 10^5 \lambda^{-1.62}$, [$\lambda$] in nanometers Bashkatov *et al.* (2005a)
	600	0.45(0.23)	–	13.81(2.43)	–	
	800	0.13(0.16)	–	9.79(1.68)	–	
	1000	0.27(0.21)	–	6.14(0.74)	–	
	1200	0.57(0.31)	–	4.43(0.43)	–	
	1400	4.84(1.79)	–	5.07(0.71)	–	
	1600	2.83(1.01)	–	3.13(0.55)	–	
	1800	3.04(1.15)	–	3.04(0.57)	–	
	1900	9.23(2.69)	–	7.01(3.57)	–	
	2000	9.31(2.28)	–	6.26(3.56)	–	

Skin:						
Caucasian dermis ($n = 12$)	633	0.33(0.09)	–	27.3(5.4)	0.9	A single integrating sphere "comparison" method, IMC; samples from abdominal and breast tissue obtained from plastic surgery, $g = 0.9$ is supposed value in calculations Laufer *et al.* (1998); Simpson *et al.* (1998)
	700	0.19(0.06)	–	23.2(4.1)	0.9	
	900	0.13(0.07)	–	16.3(2.5)	0.9	
Negroid dermis ($n = 5$)	633	2.41(1.53)	–	32.1(20.4)	0.9	
	700	1.49(0.88)	–	26.8(14.1)	0.9	
	900	0.45(0.18)	–	18.1(0.4)	0.9	
Subdermis (primarily globular fat cells) ($n = 12$)	633	0.13(0.05)	–	12.6(3.4)	0.9	
	700	0.09(0.03)	–	12.1(3.2)	0.9	
	900	0.12(0.04)	–	10.8(2.7)	0.9	
Muscle ($n = 1$)	633	1.21	–	8.9	0.9	
	700	0.46	–	8.3	0.9	
	900	0.32	–	5.9	0.9	
Sample/subject −01/01; female (F), age = 51 yr; back of knee, left leg; moderate inflammation in dermis; SC = 40–70 μm; E = 40–150 μm; D = 300 μm	1460	17.88(1.12)	–	10.74(0.49)	–	DIS, IAD; slabs containing stratum corneum (SC), epidermis (E), and dermis (D), taken from 14 subjects; measured within 24 hr of excision; heated to 37 °C; three measurements on each side of the sample; 2.5-cm-diameter sample ports on the setup; total data for 52 wavelengths in the range 1000–2200 nm are available Troy and Thennadil (2001)
	1600	5.35(0.24)	–	8.06(0.29)	–	
	2200	7.46(0.56)	–	7.17(0.26)	–	
07/04; male (M), age = 64 yr; thigh, right leg; mild chronic dermatitis; SC = 20–30 μm; E = 50–90 μm; D = 300 μm	1000	0.69(0.01)	–	10.45(0.61)	–	
	1460	16.64(0.95)	–	10.75(0.81)	–	
	1600	4.96(0.27)	–	7.72(0.40)	–	
	2200	13.04(2.36)	–	9.42(1.57)	–	

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	μ_s', cm^{-1}	g	Remarks
15/10; M, age = 70 yr; scalp; mild chronic dermatitis w/solar elastosis; SC = 4–15 µm; E = 8–10 µm; D = 200 µm	1000	1.04(0.02)	–	12.26(0.44)	–	
	1460	15.95(0.99)	–	10.75(1.20)	–	
	1600	5.09(0.23)	–	8.83(0.92)	–	
	2200	12.65(0.52)	–	8.83(1.94)	–	
20/13; F, age = 53 yr; scalp/facial tissue; mild solar damage; SC = 4 µm; E = 10 µm; D = 200 µm	1000	1.53(0.02)	–	12.89(0.77)	–	
	1460	16.82(1.13)	–	12.01(0.81)	–	
	1600	5.57(0.19)	–	9.47(0.60)	–	
	2200	13.46(0.58)	–	10.41(0.71)	–	
21/14; F, age = 52 yr; abdomen; mild chronic inflammation; SC = 4–5 µm; E = 10 µm; = 200 µm	1000	0.88(0.03)	–	14.96(1.28)	–	
	1460	18.21(2.51)	–	14.20(0.71)	–	
	1600	5.74(0.68)	–	10.58(0.44)	–	
	2200	11.33(0.76)	–	10.40(0.47)	–	
Uterus:						
Postmenopausal	630	0.515(0.054)	–	9.1(1.7)	–	Frequency-domain (FD); intact uteri were obtained by hysterectomy; during measurement period (3–4 hr) wet gauze was applied Madsen *et al.* (1994)
Premenopausal	630	0.193(0.013)	–	7.3(0.9)	–	
	630	0.314(0.030)	–	8.9(1.5)	–	
	630	0.213(0.024)	–	6.0(0.8)	–	
Fibroid	630	0.197(0.030)	–	7.3(1.5)	–	
	630	0.0824(0.0075)	–	7.2(0.9)	–	

			***In vivo* measurements**			
Adenocarcinoma (multiple subcutaneous large-cell, male 62 yrs):						FD, $r_{sd} = 2.2$ cm Fishkin *et al.* (1997); Tromberg *et al.* (1997)
Abdominal, normal tissue	674	0.0589(0.0036)	–	8.94(0.19)	–	
	811	0.0645(0.0032)	–	8.82(0.18)	–	
	849	0.0690(0.0025)	–	8.77(0.14)	–	
	956	0.1110(0.015)	–	7.00(0.62)	–	
Abdominal, tumor	674	0.169(0.02)	–	8.48(0.73)	–	
	811	0.190(0.015)	–	8.30(0.49)	–	
	849	0.276(0.03)	–	9.93(0.87)	–	
	956	–	–	–	–	
Back, normal tissue	674	0.0883(0.006)	–	10.7(0.4)	–	
	811	0.0892(0.005)	–	9.99(0.27)	–	
	849	0.0915(0.0030)	–	9.65(0.15)	–	
	956	0.127(0.03)	–	6.3(0.9)	–	
Back, tumor	674	0.174(0.02)	–	10.4(0.9)	–	
	811	0.177(0.013)	–	9.23(0.5)	–	
	849	0.190(0.01)	–	9.20(0.33)	–	
	956	0.186(0.16)	–	4.7(2.7)	–	
Brain:						
Normal cortex, temporal and frontal lobe	674	>0.2	–	10(1)	0.92	SRR; measurements during brain surgery Bevilacqua *et al.* (1997)
	849	>0.2	–	9.2(1)	0.92	
	956	>0.2	–	8.5(1)	0.92	
Normal optic nerve	674	0.60(0.25)	–	18(1)	0.92	
	849	0.75(0.25)	–	17(1)	0.92	
	956	0.65(0.25)	–	16(1)	0.92	
Astrocytoma of optic nerve	674	1.6(1)	–	14(1)	0.92	
	849	1.1(1)	–	8.5(1)	0.92	
	950	1.8(1)	–	8.5(1)	0.92	

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	$\mu_s{}'$, cm^{-1}	g	Remarks
Normal cortex, frontal lobe	674	<0.2	–	10(0.5)	–	SRR; measurements during brain surgery Bevilacqua *et al.* (1999); Mobley and Vo-Dinh (2003)
	811	<0.1	–	9.1(0.5)	–	
	849	<0.1	–	9.2(0.5)	–	
	956	0.15(0.1)	–	8.9(0.5)	–	
Normal cortex, frontal lobe	674	0.2(0.1)	–	10(0.5)	–	
	811	0.2(0.1)	–	8.2(0.5)	–	
	849	<0.1	–	8.2(0.5)	–	
	956	0.25(0.1)	–	8.2(0.5)	–	
Normal optic nerve	674	0.6(0.3)	–	17.5(2)	–	
	849	0.8(0.3)	–	16(2)	–	
	956	0.7(0.3)	–	15.2(2)	–	
Astrocytoma of optic nerve	674	1.4(0.3)	–	12.5(1)	–	
	811	1.2(0.3)	–	9.5(1)	–	
	849	0.9(0.3)	–	7.6(1)	–	
	956	1.5(0.3)	–	7.3(1)	–	
Normal white matter	674	2.5(0.5)	–	13.5(1)	–	
	849	0.95(0.2)	–	8.5(1)	–	
	956	0.9(0.2)	–	7.8(1)	–	
White matter with scar	674	<0.2	–	6.5(0.5)	–	
	849	<0.2	–	8(0.5)	–	

Medulloblastoma	674	2.6(0.5)	–	14(1)	–	
	849	1(0.2)	–	10.7(1)	–	
	956	0.75(0.2)	–	4(1)	–	
Breast (female):						
Normal (30 Japanese women, averaged for all ages)	753	0.046(0.014)	–	8.9(1.3)	–	Time domain (TD), μ_a (cm^{-1}) $\approx 0.087-8.31 \times 10^{-4}x$, μ_s' (cm^{-1}) $\approx 13-0.08x$, where x = age (20–80 yr) Suzuki *et al.* (1996)
Normal (6 women, 26–43 yr)	800	0.017–0.045	–	7.2–13.5	–	TD, μ_s' (cm^{-1}) $\approx 16.7-7.9 \times 10^{-3}\lambda$, $\lambda =$ 500–1060 nm HeusmannKolzer and Mitic (1996)
Normal (6 women, tissue thickness, 33–49 mm at light compression)	580	0.70(0.12)	–	–	–	Measurements of transmission, $g \approx 0.92-0.95$, $\mu_s' =$ 12–13 cm^{-1} Key *et al.* (1991)
	780	0.23(0.02)	–	–	–	
	850	0.27(0.03)	–	–	–	
Breast cancer (5 patients)	630	0.305(0.16)	–	9.41(7.35)	–	SRR; relapsed cancer, HPD (72 hr) Driver, Lowdell and Ash (1991)

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	$\mu_s{\prime}$, cm^{-1}	g	Remarks
Normal (56 yr)	674	0.04	–	8.5	–	FD, $r_{sd} = 2.2$ cm Fishkin *et al.* (1997); Tromberg *et al.* (1997)
	811	0.035	–	7.4	–	
	849	0.035	–	7.0	–	
	956	0.085	–	6.5	–	
Fibroadenoma with ductal hyperplasia (56 yr)	674	0.055	–	9	–	
	811	0.06	–	8	–	
	849	0.055	–	7.6	–	
	956	0.12	–	7.5	–	
Normal (27 yr)	674	0.035	–	11.1	–	
	811	0.03	–	9.6	–	
	849	0.038	–	9.6	–	
	956	0.09	–	9.7	–	
Fluid-filled cyst (27 yr)	674	0.07	–	7.9	–	
	811	0.07	–	7.0	–	
	849	0.08	–	7.0	–	
	956	0.16	–	7.0	–	
Papillary cancer (55 yr)	690	0.084(0.014)	–	15.0(0.3)	–	FD, DT Fantini *et al.* (1998)
	825	0.085(0.017)	–	12.7(0.3)	–	
Normal (67 yr)	674	0.057	–	9.5	–	FD, diffusion approximation; tumor 1.8 × 0.9 cm; segmented reconstruction Holboke *et al.* (2000)
	782	0.050	–	9.4	–	
	803	0.047	–	9.0	–	
	849	0.054	–	8.9	–	
Ductal carcinoma *in situ* (67 yr)	674	0.17	–	4.1	–	
	782	0.18	–	3.6	–	
	803	0.15	–	4.2	–	
	849	0.21	–	3.3	–	
Calf (11 subjects, 14 measurements)	800	0.17(0.05)	–	9.4(0.7)	–	$\mu_s'(\text{cm}^{-1}) \approx 16 - 8.9 \times 10^{-3}\lambda$, $\lambda = 760\text{–}900$ nm Matcher, Cope and Delpy (1997)

Cervical stromal tissue	849	0.34	61.1	–	0.9	Hayakawa *et al.* (2001)
Cervical tissue:					μ_{bs}	$\mu_{bs} = \mu_s p_b$, p_b – probability of backscattering
Epithelium ($n = 36$)	1300	–	10–140	–	0.1–11	OCT, two-layered model, genetic inverse algorithm Turchin *et al.* (2003)
Stroma ($n = 36$)	1300	–	30–290	–	1.5–12	
Dysplasia II–III	1300	–	40–65	–	1.4–3.6	OCT, single-layered model, genetic inverse algorithm Dolin *et al.* (2004)
Leukoplakia	1300	–	16–32	–	1.3–2.0	
Epithelium	1300	–	80(25)	–	0.28(0.08)	OCT, genetic inverse algorithm two-layered model single-layered model Dolin *et al.* (2004)
Stroma	1300	–	210(30)	–	3.0(2)	
Cancer	1300	–	300(20)	–	1.2(0.6)	
Gastrointestinal tract:					$\gamma = (1 - g_2)/(1 - g_1)$	
Mucosa in the antrum	500	2.5(0.8)	–	16.8(3.4)	1.98(0.20)	Endoscopic SRR, r_{sd} from 0.3 to 1.35 mm; IMC, two-moments HGPF; 35 patients (21 females and 14 males, age from 23 to 87), for each patient, four sites were usually selected – two in the antrum and two in the fundus, average data for normal tissue Thueler *et al.* (1993)
	550	3.6(1.3)	–	13.8(3.4)	1.93(0.15)	
	600	1.0(0.6)	–	12.8(2.1)	1.90(0.12)	
	650	0.5(0.5)	–	11.7(1.8)	1.87(0.12)	
	700	0.4(0.4)	–	10.7(1.6)	1.86(0.12)	
	750	0.45(0.45)	–	9.7(1.6)	1.85(0.12)	
	800	0.5(0.4)	–	9.0(1.6)	1.85(0.12)	
	850	0.7(0.5)	–	8.8(1.6)	1.85(0.12)	
	900	0.8(0.5)	–	8.3(1.5)	1.86(0.12)	

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	μ_s', cm^{-1}	g	Remarks
Mucosa in the fundus	500	3.3(0.8)	–	21.7(3.1)	2.12(0.26)	
	550	4.4(1.5)	–	18.6(3.1)	2.04(0.24)	
	600	1.8(0.5)	–	16.5(2.1)	2.00(0.19)	
	650	0.8(0.4)	–	15.0(2.1)	1.97(0.17)	
	700	0.7(0.4)	–	13.8(1.7)	1.98(0.15)	
	750	0.7(0.5)	–	12.4(1.6)	1.98(0.15)	
	800	0.7(0.3)	–	11.7(1.6)	1.95(0.12)	
	850	0.7(0.3)	–	11.0(1.4)	1.92(0.12)	
	900	0.8(0.4)	–	10.3(1.4)	1.94(0.12)	
Head (7 subjects, 10 measurements)	800	0.16(0.01)	–	9.4(0.7)	–	μ_s'(cm^{-1}) ≈14.5 − 6.5 × $10^{-3}\lambda$, $\lambda = 760$–900 nm Matcher, cope and Delpy (1997)
Forearm (5 subjects, 14 measurements)	800	0.23(0.04)	–	6.8(0.8)	–	TD, μ_s' (cm^{-1}) ≈11 − 5.1 × $10^{-3}\lambda$, $\lambda = 760$–900 nm Troy, Page and Sevick-Muraca (1996)
Forearm	715	0.18	–	3.7	–	FD; r_{sd} – a few centimeters Gratton *et al.* (1997)
	825	0.24	–	3.0	–	
Forehead	715	0.16	–	7.3	–	
	825	0.16	–	6.9	–	
Skin:						
Dermis	660	0.07–0.2	–	9 – 14.5	–	SRR, IMC Graaff *et al.* (1993)
Skin	633	0.62	–	32	–	Doornbos *et al.* (1999)
	700	0.38	–	28.7	–	
Skin (0–1 mm)	633	0.67	–	16.2	–	Kienle, Lilge and Patterson (1994)
Skin (1–2 mm)	633	0.026	–	12.0	–	
Skin (>2 mm)	633	0.96	–	5.3	–	
Epidermis and dermis	750	0.375	–	15	–	SRR; DT Farrell, Patterson and Essenpreis (1998)
Subcutaneous fat	750	0.03	–	10	–	

Arm	633	0.17(0.01)	–	9.08(0.05)	–	SRR, 9 detecting 600-µm fibers; mean separation 1.7 mm; Mie phase function Doornbos *et al.* (1999)
	660	0.128(0.005)	–	8.68(0.05)	–	
	700	0.090(0.002)	–	8.14(0.05)	–	
Foot sole	633	0.072(0.002)	–	11.17(0.09)	–	
	660	0.053(0.003)	–	10.45(0.09)	–	
	700	0.037(0.001)	–	9.52(0.08)	–	
Forehead	633	0.090(0.009)	–	16.72(0.09)	–	
	660	0.052(0.003)	–	16.16(0.08)	–	
	700	0.0240(0.002)	–	15.38(0.06)	–	
Abdominal skin:						
chosen direction	810	0.014	–	20	–	SRR; CCD detector, $r_{\mathrm{sd}} \leq 10$ mm; diffusion approximation Nickell *et al.* (2000)
perpendicular direction (along collagen fibers)	810	0.07	–	10	–	
Forearm (light skin, $n = 7$):						
skin temperature −22 °C	590	2.372(0.282)	–	9.191(0.931)	–	SRR; MC-generated grid Khalil *et al.* (2003)
	750	0.966(0.110)	–	7.340(0.901)	–	
	950	0.981(0.073)	–	6.067(0.847)	–	
skin temperature −38 °C	590	2.869(0.289)	–	9.613(0.894)	–	
	750	1.157(0.106)	–	7.649(0.971)	–	
	950	1.135(0.123)	–	6.234(0.928)	–	
Dermis of a lower arm	1300	–	47	–	–	OCT (Schmitt, Knüttel and Bonnar (1993)
Stratum corneum of finger	1300	–	12	–	–	

(*continued overleaf*)

Tab. 16.1 (*Continued*)

Tissue	λ, nm	μ_a, cm^{-1}	μ_s, cm^{-1}	μ_s', cm^{-1}	g	Remarks
Volar side of lower arm:						OCT Knüttel and Boehlau-Godau (2000)
Epidermis	1300	–	15–20	–	–	
upper dermis	1300	–	80–100	–	–	
Palm of hand:						
stratum corneum	1300	–	10–15	–	–	
epidermis:						
grandular layer	1300	–	60–70	–	–	
basal layer	1300	–	40–50	–	–	
upper dermis	1300	–	50–80	–	–	
Volar side of lower arm (epidermis and dermis)						OCT; depth up to 350 μm; skin treated with a detergent solution (2% of anionic tensides in water) Knüttel, Bonev and Knaak (2004))
Normal	1300	–	140	–	–	
Treated	1300	–	80	–	–	
Skull	674	<0.2	–	9(1)	–	SRR; measurements during brain surgery Bevilacqua *et al.* (1999)
	849	<0.1	–	9(1)	–	
	956	0.15(0.1)	–	8.5(1)	–	

where water absorption is weak (up to 1100–1300 nm). The refractive index microscopic redistribution of a tissue due to cellular and fiber proteins' denaturation and homogenization at thermal action also may have a strong inclusion in alteration of scattering and absorption properties.

16.3.10
Optical Properties of Blood

Fresh human blood placed in calibrated thin cuvette (thickness from 0.01 to 0.5 mm, slab geometry) is usually used for the determination of blood optical parameters. Before the optical measurements, standard clinical tests are necessary to determine the concentration of red and white blood cells, concentration of platelets, hematocrit, mean corpuscular volume and hemoglobin, and the other parameters of interest. If blood sample oxygenation level is of interest, it may be controlled using a conventional blood gas analyzer (Yaroslavsky *et al.*, 1999). In most cases, the experiments are performed with either completely oxygenated or completely deoxygenated blood (Hammer *et al.*, 1998; Roggan *et al.*, 1999; Yaroslavsky *et al.*, 1999; Tuchin, 2002). To obtain complete oxygen saturation, the sample is exposed to air or O_2. To completely deoxygenate blood, sodium dithionite ($Na_2S_2O_4$) is added. To be sure that neither the volume nor the surface area of the blood particles changes during the experiments, the pH of the samples should be maintained in the range of physiological values, at approximately 7.4.

In reality, blood is flowing through the blood vessels and, therefore, it is preferable to study the optical properties of flowing blood. The red blood cells (RBCs) in flow are subject to deformation and orientation. At lower shear rates, reversible aggregation occurs; while under the higher shear rates, erythrocytes are deformed into ellipsoids. The experiments with flowing undiluted and diluted blood are reported (Nilsson *et al.*, 1997; Roggan *et al.*, 1999; Steenbergen, Kolkman and de Mul, 1999).

16.3.11
Refractive Index Measurements

Refractivity measurements in a number of strongly scattering tissues and blood were performed using various techniques (see Tuchin, 2007). Experimental values of mean refractive index for some tissues measured for selected wavelengths are summarized by Tuchin (2007). One of the techniques is a fiber-optic refractometer based on a simple concept that the cone of light issuing from an optical fiber is dependent on the indices of the cladding material (tissue) (Bolin *et al.*, 1989). Using this simple and sensitive technique, it was found that at 633 nm, fatty tissue has the largest refractive index (1.455), followed by kidney (1.418), muscular tissue (1.410), then blood and spleen (1.400). The lowest refractive indices were found in lungs and liver (1.380 and 1.368, respectively). Also, it turned out that tissue homogenization does not significantly affect the refractive indices (the change does not exceed a measurement error equal to 0.006), whereas coagulated tissues have higher refractive indices than native ones (e.g., for egg white, $\overline{n}$ changing from 1.321 to 1.388). Moreover, there is a tendency for refractive indices to decrease with increasing light wavelength from 390 to 700 nm (e.g., for bovine muscle in the limits 1.42–1.39), which is characteristic of the majority of related abiological materials.

The principle of total internal reflection at laser beam irradiation is also used for tissue and blood refraction measurements (Liu and Xie, 1996; Cheng *et al.*, 2002). A thin tissue sample is sandwiched between two right-angled prisms that are made of ZF5 glass with a high refractive index $n_0 = 1.70827$ and angle $\alpha = 29^0 55' 41.4''$. Measurements for fresh animal tissues and human blood at four laser wavelengths 488, 632.8, 1079.5, and 1341.4 nm and room temperature were presented in a form of Cauchy dispersion equation as (Cheng *et al.*, 2002)

$$\overline{n} = A + B\lambda^{-2} + C\lambda^{-4} \qquad (16.40)$$

with λ in nanometers, values of the Cauchy coefficients are given in the following table (Table 16.2):

An expression for human blood plasma received by Cheng *et al.* (2002) was extrapolated to shorter wavelengths from 400 to 1000 nm (Tuchin *et al.*, 2004):

$$\begin{aligned} n_{\text{bp}}(\lambda) = {} & 1.3254 + 8.4052 \times 10^3 \lambda^{-2} \\ & - 3.9572 \times 10^8 \lambda^{-4} \\ & - 2.3617 \times 10^{13} \lambda^{-6} \qquad (16.41) \end{aligned}$$

For modeling of the behavior of refractive index of tissues, blood, and their components, one may use a remarkable property of proteins: that equal concentrations of aqueous solutions of different proteins all have approximately the same refractive index, n_{pw} (Barer, Ross and Tkaczyk, 1953). Moreover, the refractive index varies almost linearly with concentration, C_p

$$n_{\text{pw}}(\lambda) = n_w(\lambda) + \beta_p(\lambda) \cdot C_p \qquad (16.42)$$

where n_w is refractive index of water and β_p is the specific refractive increment; C_p is measured in grams per 100 ml (grams per deciliter). For example, the refractive index of human erythrocyte cytoplasm, defined by the cell-bounded hemoglobin solution, can be found from this equation at $\beta_p = 0.001942$ valid for a wavelength of 589 nm; that is, for normal hemoglobin concentration in cytoplasm of 300–360 g l^{-1}, the RBC refractive index $n_{\text{RBC}} = 1.393$–1.406 (Roggan *et al.*, 1999). Values of specific refractive increment β_p for some other proteins measured by Abbe refractometer at a wavelength 589 nm are given in the following table (Table 16.3) (Barer, Ross and Tkaczyk, 1953):

Other materials of specific biological interest are the carbohydrates, lipoids, and nucleic acid compounds. The first two usually have low values of β; in

Tab. 16.2 Values of the Cauchy coefficients for some tissues.

Tissue sample	A	B x 10^{-3}	C x 10^{-9}
Porcine muscle$_{\|\|}$	1.3694	0.073223	1.8317
Porcine muscle$_{\perp}$	1.3657	1.5123	1.5291
Porcine adipose	1.4753	4.3902	0.92385
Porcine small intestine	1.3563	4.3905	0.92379
Ovine muscle$^{\|\|}$	1.3716	5.8677	0.43999
Ovine muscle$^{\perp}$	1.3682	8.7456	−0.16532
Human whole blood	1.3587	1.4744	1.7103
Human blood plasma	1.3194	14.578	−1.7383

$^{||}$ Porcine (or ovine) muscle samples with the tissue fibers oriented in parallel to the interface.
$^{\perp}$ Same sample with the tissue fibers oriented perpendicular to the interface.

Tab. 16.3 Values of specific refractive increment β_p for some proteins measured by Abbe refractometer at a wavelength 589 nm.

Protein	β_p**, dl/g**
Lipoprotein	0.00170–0.00171
Total serum (human)	0.00179
Total albumin	0.00181
Egg albumin	0.001813
Euglobulin	0.00183

the region of 0.0014–0.0015, and nucleic acids have higher values, 0.0016–0.0020 (Barer, Ross and Tkaczyk, 1953).

Some other techniques of refractive index measurements of biological liquids are also available. One of them allows for refractive index measurements at a few separate wavelengths and uses an equilateral small angle (of 10°) hollow prism made from thin quartz slides (Sardar and Levy, 1998).

Since the refractive index of tissue and blood components defines their scattering properties, measured scattering parameters may have an advantage to evaluate the refractive index of biomaterial components and their mean values (Tuchin, 2007). Determination of the reduced scattering coefficient of a tissue sample using IS or spatially resolved techniques and corresponding algorithms for extraction of the scattering coefficient, such as IAD or IMC, the knowledge of the refractive indices of the scatterers and the ground material at one of the wavelengths, as well as experimental or theoretical estimations for mean radius of the scatterers, allows one to solve the inverse problem and reconstruct the spectral dependence of the refractive index of the scatterers for a given spectral dependence of the refractive index of the ground material. Similar measurements and theoretical estimations done for a tissue sample before and after its prolonged bathing in saline or other biocompatible liquid with known optical characteristics allow one to evaluate the spectral dependencies both of refractive index of the scatterers and the ground material.

A short pulse time delay technique was also successfully applied for refractive index estimation of normal breast tissue and malignant breast tissue (Das, Liu and Alfano, 1997). Using the known thickness of the sample and the measured shift Δt of the transmitted pulse peak relative to the delay time measured through a layer of air of the same thickness, the mean phase refractive index $\overline{n}$ of a tissue sample can be calculated. Very short pulses should be used in such measurements; thus, a group of different wavelengths propagates in a media and the material dispersion $(d\overline{n}/d\lambda)$ should be accounted for by introducing the group refractive index

$$\overline{n}_g = \overline{n} - \lambda \frac{d\overline{n}}{d\lambda} \tag{16.43}$$

The time delay in the pulse arrival for a tissue sample of thickness d is

$$\Delta t = \frac{d}{c_0}(\overline{n}_{g1} - n_{g2}) \tag{16.44}$$

where c_0 is the light velocity in a free space; $\overline{n}_{g1}$ is the effective (mean) group refractive index of a tissue, and n_{g2} is the group refractive index of the homogeneous reference medium (air). The effective group refractive index of a tissue is

$$\overline{n}_{g1} = f_s n_{gs} + (1 - f_s) n_{g0} \tag{16.45}$$

where f_s is the volume fraction of the scatterers composing a tissue, n_{gs} is the group refractive index of the scatterers, and n_{g0} is the group refractive index of the ground material of a tissue. The values of the phase refractive index of the above-mentioned two samples were calculated to be $\overline{n} = 1.403$ for normal and 1.431 for malignant tissue.

OCT provides simple and straightforward measurements of the index of refraction both *in vitro* and *in vivo* (Tearney *et al.*, 1995; Wang *et al.*, 1995; Knüttel and Boehlau-Godau, 2000; Knüttel, Bonev and Knaak, 2004; Ohmi *et al.*, 2000; Tuchin, Xu and Wang, 2002; Tuchin, 2007). The in-depth scale of OCT images is determined by the optical path length Δz_{opt} between two points along the depth direction. Because a broadband light source is used, the optical path length is proportional to the group refractive index n_g and geometrical path length Δz as

$$\Delta z_{\text{opt}} = n_g \Delta z \tag{16.46}$$

Usually, $n_g \cong n$. This simple relation is valid for a homogeneous medium and can be used in *in vitro* studies when geometrical thickness of a tissue sample Δz is known.

Sometimes both refractive index and thickness of a tissue sample should be measured simultaneously. In that case, a two-step procedure can be applied. First, a stationary mirror is placed in the sample arm of an interferometer to get the geometric position of the mirror supposing that the group refractive index of air is 1 (z_1). Then a tissue sample with unknown index n_g, and thickness d should be placed before the mirror in the sample arm. Two peaks from the anterior (z_2) and the posterior (z_3) surfaces of the sample will appear with the distance between them equal to a sample optical thickness, and the position of the mirror (z_4) will be shifted by $(n_g - 1)d$ due to the sample whose group refractive index is greater than that of air. Thus, the calculation of the geometrical thickness and the group refractive index proceeds as follows:

$$\begin{aligned} d &= (z_3 - z_2) - (z_4 - z_1) \\ n_g &= \frac{z_3 - z_2}{d} \end{aligned} \tag{16.47}$$

For *in vivo* measurements of the index of refraction, a focus-tracking method that uses OCT to track the focal-length shift that results from translating the focus of an objective along the optical axis within a tissue was introduced (Tearney *et al.*, 1995) and further developed (Knüttel and Boehlau-Godau, 2000; Knüttel, Bonev and Knaak, 2004; Zvyagin *et al.*, 2003).

16.3.12
Optical Projection Tomography

Optical projection tomography (OPT), based on ballistic and/or quasi-ballistic photon detection, allows for imaging inhomogeneities and quantifying optical properties of biological materials with thicknesses from tens of microns to centimeters (Darrell *et al.*, 2008; Oldham *et al.*, 2008). Major applications are in developmental embryology and gene expression studies, as well as in three-dimensional imaging and quantification within organs of small animals and xenograft tumors. Typically, light from a uniform backlight transverses through the sample to form projection images captured by a CCD camera or fluorescence from fluorophores distributed in biological material and excited by a light orthogonal to the imaging axis is detected at rotation of the sample around its axis. There have been substantial developments in modeling, reconstructing algorithms, reflection, refraction and scattering reduction, and in enhancements of the optical systems used in OPT. As it was noted by Darell *et al.* (2008), several applications of OPT would benefit from quantitative reconstructions, that is, reconstructions in which, for example, the intensities of two areas of gene expression could be reliably compared, thus, they described methods of accounting for the quantitative effects of the isotropic emission and blurring in a fluorescence OPT imaging system.

Glossary

Absorption Coefficient: In a nonscattering medium absorption coefficient μ_a is defined as the reciprocal of the distance d over which light of intensity $I(d = 0) = I_0$ is attenuated (due to absorption) to $I(d) = I_0/e \approx 0.37 I_0$; the units are typically cm^{-1}; behind this definition is a fundamental process of photon absorption that is characterized by a photon absorption cross section.

Anisotropic Scattering: A scattering process characterized by a clearly apparent direction of photons that may be due to the presence of large scatterers.

Aorta: The great artery that leaves the left ventricle; it conducts the whole of the arterial blood supply to all parts of the body other than the lungs; in humans, it carries blood at the rate of 4 dm^3 per minute.

Astrocytoma: A rather well-differentiated glioma (a *tumor* of the brain), which consists of cells that look like astrocytes (a starlike cell of the macroglia of nerve tissue).

Bilirubin: Yellow bile pigment formed in the breakdown of heme (an iron-containing substance; the basic unit of the *hemoglobin* molecule; mammals have four heme units in their hemoglobin).

Cancer: A general term applied to a carcinoma or a sarcoma; the typical symptoms are a *tumor* or swelling, a discharge, pain, an upset in the function of an organ, general weakness, and loss of weight.

Cartilage: A strong, resilient, skeletal tissue; its simplest and most common form consists of a matrix of a polysaccharide-containing protein in which are embedded cartilage cells (chondroblasts); the matrix is without structure and without blood vessels; this type is known as *hyaline cartilage*, is translucent and clear, and occurs in the cartilaginous rings of the trachea and bronchi; elastic cartilage (yellow fibrocartilage) contains yellow fibers in the matrix; it occurs in the external ear and in the epiglottis; white fibrocartilage contains white fibers in the matrix; it occurs in the disks of cartilage between the vertebrae; all types of cartilage contain chondroblasts, which deposit the matrix and become enclosed in the matrix as chondrocytes.

CCD: Charge-coupled device – a solid-state electronic device that serves as an imaging chip and is used in video cameras and fast spectrometers.

Cell Organelle: A part of a cell that is a structural and functional unit, for example, a flagellum is a locomotive organelle, a *mitochondrion* is a respiratory organelle; organelles in a cell correspond to organs in an organism.

Cerebellum: A region of the brain that plays an important role in the integration of sensory perception and motor output; many neural pathways link the cerebellum with the motor cortex, which sends information to the muscles causing them to move and the spinocerebellar tract that provides feedback on the position of the body in space; the cerebellum integrates these pathways, using the constant feedback on body position to fine-tune motor movements.

Cerebral Membrane (*dura mater*): Or pachymeninx, is the tough and inflexible outermost of the three layers of the *meninges* surrounding the brain and spinal cord; the *dura mater* itself has two layers: a superficial layer, which is actually the skull's inner *periosteum*, and a deep layer, the *dura mater proper*.

Collagen: A tough, inelastic, fibrous protein; on boiling it forms gelatin; on adding acetic acid it swells up and dissolves; collagen is formed and maintained in tissues by fibroblasts; it forms white fibers in *connective tissue*; the tropocollagen or "collagen molecule" subunit is a rod about 300-nm long and 1.5 nm in diameter, made up of three polypeptide strands, subunits spontaneously self-assemble, with regularly staggered ends, into even larger arrays in the extracellular spaces of tissues; there is some covalent crosslinking within the triple helices, and a variable amount of covalent crosslinking between tropocollagen helices, to form the different types of collagen found in different mature tissues – similar to the situation found with the α-keratins in hair; a distinctive feature of collagen is the regular arrangement of amino acids in each of the three chains of these collagen subunits; in bone, entire collagen triple helices lie in a parallel, staggered array; 40-nm gaps between the ends of the tropocollagen subunits probably serve as nucleation sites for the deposition of long, hard, fine crystals of the mineral component, which is (approximately) hydroxyapatite with some phosphate; it is in this way that certain kinds of cartilage turn into bone; collagen gives bone its elasticity and contributes to fracture resistance.

Connective Tissue: Various body tissues that bind together and support organs and other tissues, for example, connective tissue surrounds muscles and nerves, connects bones and muscles and underlies the *skin; cartilage* and bone are also connective tissues; typical connective tissue consists of cells scattered in an amorphous mucopolysaccharide matrix in which there are varying amounts of connective tissue fibers (mainly *collagen*, but also *elastin* and *reticulin*)

CW: Continuous wave

Cytoplasm: All the protoplasm of a cell exclusive of the nucleus; it is not just a simple, slightly viscous, fluid; various structures, called *organelles*, each concerned with different functions of the cell, are situated in it; the plasma membrane is part of the cytoplasm.

Cytoplasmic Organelles: See *cell organelle*.

Dermis: The inner layer of the *skin*; it is composed of *connective tissue*, blood and lymph vessels, muscles and nerves; *collagen* fibers are abundant in the dermis and run parallel to the surface of the skin; they give the skin elasticity; sweat glands and hair follicles are scattered throughout the dermis; the dermis is much thicker than the epidermis and is developed from mesoderm.

Diffusion Approximation (Diffusion Theory): The approximated diffusion-type solution of the *RTT*, which is accurate for describing of photon migration in infinite, homogeneous, highly scattering media.

Dysplasia: An abnormality in the appearance of cells indicative of an early step toward transformation into a neoplasia. It is therefore a preneoplastic or precancerous change; this abnormal growth is restricted to the originating system or location, for example, a displasia in the epithelial layer will not invade into the deeper tissue, or a displasia solely in a red blood cell line (refractory anemia) will stay within the bone marrow and cardiovascular systems; the best known form of displasia is the precursor lesions to cervical *cancer*, called cervical intraepithelial neoplasia (CIN); this lesion is usually caused by an infection with the human papilloma virus (HPV).

Elastin: An elastic fibrous protein resistant to boiling and to acetic acid; it forms highly elastic yellow fibers in *connective tissue*; elastin is formed and maintained in tissues by fibroblasts.

Epidermis: The outer layer of the *skin* is a stratified *epithelium* that varies relatively little in thickness over most of the body (between 75 and 150 μm), except on the palms and soles, where its thickness may be 0.4–0.6 mm; the epidermis is conventionally subdivided into (i) stratum basale, a basal cell layer of keratinocytes, which is the germinative layer of the epidermis; (ii) the stratum spinosum, which consists of several layers of polyhedral cells lying above the germinal layer; (iii) the stratum granulosum, which is a layer of flattened cells containing distinctive cytoplasmic inclusions and keratohyalin granules; and (iv) the overlying *stratum corneum*, consisting of lamellae of anucleate thin, flat squames that are terminally differentiated keratinocytes.

Epithelium: A sheet of epithelial tissue; epithelium is derived from ectoderm and endoderm.

Eye Cornea: The transparent covering at the front of the eyeball; it is the modified continuation of the *sclera*; it refracts light and is the most important element in the refractive system of the eye.

Eye Sclera: A dense, white matter, fibrous membrane which, with the *cornea*, forms the external covering of the eyeball; scleral regions: limbal, equatorial, and posterior pole region.

Extinction Coefficient (Attenuation Coefficient): The reciprocal of the distance over which light of intensity I is attenuated to $I/e \approx 0.37I$; attenuation or interaction coefficient, $\mu_t = \mu_a + \mu_s$, where μ_a is the *absorption coefficient* and μ_s is the *scattering coefficient*; the units are typically cm^{-1}.

***Ex vivo*:** Taken from the living organism; pertaining to experiments on animal or human organs that are excised from the living body and kept in conditions very close to the natural ones.

Fat (Adipose Tissue): A modification of areolar *tissue* in which globules of oil are deposited in some of the cells (fat cells); the cells tend to be grouped together and in mammals occur in the tissues under the skin and around the abdominal organs (kidneys, liver, etc.).

Fibrous Tissue: A tissue mainly consisting of conjunctive *collagen* (or *elastin*) fibers, often packed in lamellar bundles.

GKPF: Gegenbauer kernel phase function.

Golgi Apparatus: A netlike mass of material in the *cytoplasm* of animal cells, believed to function in cellular secretion.

Gray Matter: A nervous tissue found in the central nervous system; it contains numerous cell bodies (cytons), dendrites, synapses, terminal processes of axons, blood vessels, and neuroglia; it is internal to *white matter* in the spinal cord and some other parts of the brain; it is external to white matter in the cerebral hemisphere and in the *cerebellum*; coordination in the central nervous system is effected in gray matter; brain nuclei and nerve centers are composed of gray matter.

Hemoglobin: A red iron-containing respiratory pigment found in the blood; it conveys oxygen to the tissues and occurs in reduced form (deoxyhemoglobin) in venous blood and in combination with oxygen (oxyhemoglobin) in arterial blood; it consists of *heme* combined with globin, a blood protein; it is chemically related to chlorophyll, cytochrome, hemocyanin, and myoglobin.

HGPF: Henyey–Greenstein phase function.

Horny Layer: The same as stratum corneum, see *epidermis*.

Interstitial Fluid: A solution that bathes and surrounds the cells of multicellular animals; it is the main component of the extracellular fluid, which also includes plasma and transcellular fluid; on average, a subject has about 11 l of interstitial fluid providing the cells of the body with nutrients and a means of waste removal.

IAD: Inverse adding-doubling.

IMC: Inverse Monte Carlo.

In vitro: In medicine, pertaining to experiments on dead tissue.

In vivo: In medicine, pertaining to experiments on living animals and humans.

Isotropic Scattering: An equality of scattering properties along all axes.

Lysosomes: Membrane-bound particles, smaller than a *mitochondrion*, occurring in large numbers in the *cytoplasm* of cells; they contain hydrolytic enzymes that are released when the cell is damaged; these enzymes assist in the digestion and removal of dead cells, the digestion of food and other substances, and the destruction of redundant *organelles*.

Melanin: A dark-brown or black pigment; melanin in *melanosomes* of normal *skin* is an extremely dense, virtually insoluble polymer of high molecular weight and is always attached to a structural protein; mammalian melanin pigments have one of two chemical compositions: eumelanin, a brown polymer, and pheomelanin, a yellow–reddish alkali-soluble pigment.

Melanosome (Melanin Granular): The cytoplasmic **organelles** on which melanin pigments are synthesized and deposited; normal human skin color is primarily related to the size, type, color, and distribution of melanosomes; melanosomes are the product of specialized exocrine glands: melanocytes.

Meningioma: The most common benign *tumor* of the brain (95% of benign tumors); however, they can also be malignant; they arise from the arachnoidal cap cells of the meninges and represent about 15% of all primary brain tumors; they are more common in females than in males (2 : 1) and has a peak incidence in the sixth and seventh decades.

Mesentry: Sheets of thin **connective** *tissue* by which the stomach and intestines are suspended from the dorsal wall of the abdominal cavity; the mesenteries carry blood, lymph vessels, and nerves to the organs of the alimentary canal.

Mitochondrion (*pl.* Mitochondria): A threadlike, or rodlike, granular *organelle* in the *cytoplasm* of cells, about 0.5 μm in width, and up to 10 μm in length for threadlike mitochondria; mitochondria are bounded by a double membrane; the inner membrane is folded inward at a number of places to form cristae; mitochondria contain phosphates and numerous enzymes that vary in different tissues; their function is cellular respiration and the release of chemical energy in the form of adenosine triphosphate (ATP) for use in most of the cell's biological functions; the cells of all organisms, except bacteria and blue–green algae, contain mitochondria in varying numbers mitochondria are especially numerous in cells involved in significant metabolic activity, such as **liver** cells; mitochondria are self-replicating.

MFP: Mean free path of a photon.

Myocardium: The muscular substance of the heart.

OCT: Optical coherence tomography.

Percutaneous: Pertains to any medical procedure where access to inner organs or other tissue is done through the *skin*, for instance, via needle puncture of the skin, rather than by using an "open" approach where inner organs or tissue are exposed; phototherapy is another example of percutaneous treatment.

Peroxisome: A specialized *organelle* containing the oxidizing enzymes that degrade peroxides.

Photon Diffusion Coefficient: The proportionality coefficient between mean-square displacement of a photon within time interval τ: $\langle \Delta r^2 \rangle \sim D\tau$.

Pons: A structure located on the brain stem; in humans it is above the medulla, below the midbrain, and anterior to the *cerebellum*.

Reduced (Transport) Scattering Coefficient: A lumped property incorporating the *scattering coefficient* μ_s and the *scattering anisotropy parameter* g: $\mu_s{}' = \mu_s(1-g)$ (cm^{-1}); $\mu_s{}'$ describes the diffusion of photons in a random walk of step size of $1/\mu_s{}'$ (cm), where each step involves isotropic scattering; this is equivalent to description of photon movement using many small steps $1/\mu_s$ that each involve only a partial (anisotropic) deflection angle if there are many scattering events before an absorption event, that is, $\mu_a \ll \mu_s'$ (diffusion regime, see *diffusion approximation*); μ_s' is useful in the diffusion regime, which is commonly encountered when treating how visible and near-infrared light propagates through tissues; for many *tissues*, the reduced scattering coefficient obeys a power law, $\mu_s{}' = q\lambda^{-h}$(cm^{-1}, λ in μm).

Ribosomes: In the *cytoplasm* of a cell any several minute, angular, or spherical particles composed of protein and RNA.

RTT: Radiation transfer theory.

Scattering Anisotropy Parameter: A measure of the amount of forward direction retained after a single scattering event; if a photon is scattered by a particle so that its trajectory is deflected by a deflection angle θ, then the component of the new trajectory that is aligned in the forward direction is presented as $\cos\theta$; there is an average deflection angle and the mean value of $\langle\cos\theta\rangle$ is defined as the anisotropy, $g \equiv \langle\cos\theta\rangle$; the value of g varies in the range from -1 to 1: $g = 0$ corresponds to isotropic (Rayleigh) scattering, $g = 1$ to total forward scattering (Mie scattering at large particles), and -1 to total backward scattering.

Scattering Coefficient: A particle with a particular geometrical size redirects incident photons into new directions and so prevents the forward on-axis transmission of photons, this process constitutes scattering; the scattering coefficient μ_s (cm^{-1}) describes a medium containing many scattering particles at a concentration described as a volume density ρ (cm^3); the scattering coefficient is essentially the cross-sectional area σ_{sca} (cm^{-1}) per unit volume of medium: $\mu_s = \rho\sigma_{\text{sca}}$; a power law for dependence of the scattering coefficient on the wavelength is typical for many tissues: $\mu_s \propto \lambda^{-h}$, for different tissue structures parameter h is ranging from 1 to 2.

Scattering Phase Function: The function that describes the scattering properties of the medium and is, in fact, the probability density function for scattering in the direction $\bar{s}'$ of a photon traveling in the direction $\bar{s}$; it characterizes an elementary scattering act: if scattering is symmetric relative to the direction of the incident wave, then the phase function depends only on the scattering angle θ (angle between directions $\bar{s}$ and $\bar{s}'$).

Skin: The external protective covering on a body, joined by *connective tissue* to the muscles; it consists of an inner *dermi*s and an outer *epidermis*.

Stratum Corneum: see *epidermis stroma* The supporting framework, usually of *connective tissue*, of an organ, as distinguished from the parenchyma, for example, the main layer of the *eye sclera* and *cornea*.

Tendon: A cord of white *fibrous tissue*; it usually attaches muscle to bones.

Thalamus: A pair and symmetric part of the brain; it constitutes the main part of the diencephalons; in the caudal (tail) to oral (mouth) sequence of neuromeres, the diencephalons is located between the mesencephalon (cerebral peduncule, belonging to the brain stem) and the telencephalon.

TMFP: Transport mean free path of a photon.

Tumor: An abnormal or diseased swelling in any part of the body, especially a more or less circumscribed overgrowth of new *tissue* that is autonomous, differs more or less in structure from the part in which it grows, and serves no useful purpose; neoplasm.

White Matter: The nervous *tissue* found in the central nervous system; it consists of tracts of medullated nerve fibers in the brain and spinal cord; it also contains blood vessels and neuroglia; it is mainly external to *gray matter*, but is internal to gray matter in the cerebral hemispheres and in the *cerebellum*; the medullated fibers give the tissue its shiny white appearance.

WHO: World Health Organization.

Whole Blood: Blood containing all its natural components: blood cells and plasma.

References

Anderson, R.R. and Parrish, J.A. (1982) in *The Science of Photomedicine*, (eds J.D. Regan and J.A. Parrish), Plenum Press, New York, pp. 147–194.

Barber, P.W. and Hill, S.C. (1990) *Light Scattering by Particles: Computational Methods*, World Scientific, Singapore.

Barer, R., Ross, K.F.A. and Tkaczyk, S. (1953) *Nature* **171**(4356), 720.

Bashkatov, A.N., Genina, E.A., Kochubey, V.I., and Tuchin, V.V. (2005a) *J. Phys. D: Appl. Phys.*, **38**, 2543.

Bashkatov, A.N., Genina, E.A., Kochubey, V.I., and Tuchin, V.V. (2005b) *Opt. Spectrosc.*, **99**, 836.

Bashkatov, A.N., Genina, E.A., Kochubey, V.I., Lakodina, N.A., and Tuchin, V.V. (2006) *Proc. SPIE*, **6163**, 616310.

Berlien, H.-P. and Mueller, G.J. (eds) (2003) *Applied Laser Medicine*, Springer-Verlag, Berlin.

Bevilacqua, F., Piguet, D., Marquet, P. *et al.* (1997) *Proc. SPIE*, **3194**, 262.

Bevilacqua, F., Piguet, D., Marquet, P. *et al.* (1999) *Appl. Opt.*, **38**, 4939.

Bohren, C.F. and Huffman, D.R. (1983) *Absorption and Scattering of Light by Small Particles*, John Wiley & Sons, Ltd, New York.

Bolin, F.P., Preuss, L.E., Taylor, R.C., and Ference, R.J. (1989) *Appl. Opt.*, **28**, 2297.

Born, M. and Wolf, E. (1999) *Principles of Optics*, 7th edn., Cambridge University, Cambridge.

Chan, E.K., Sorg, B., Protsenko, D. *et al.* (1996) *IEEE J. Select. Tops Quant. Electr.*, **2**, 943.

Chance, B. (ed.) (1989) *Photon Migration in Tissue*, Plenum Press, New York.

Chandrasekhar, C. (1960) *Radiative Transfer*, Dover, Toronto.

Cheng, S., Shen, H.Y., Zhang, G. *et al.* (2002) *Proc. SPIE*, **4916**, 172.

Cheong, W.-F., Prahl, S.A., and Welch, A.J. (1990) *IEEE J. Quantum Electr.*, **26**, 2166. updated by W.-F. Cheong; further additions by L. Wang and S. L. Jacques, August 6, 1993).

Cilesiz, I.F. and Welch, A.J. (1993) *Appl. Opt.*, **32**, 477.

Darrell, A., Meyer, H., Marias, K. *et al.* (2008) *Phys. Med. Biol.*, **53**, 3863.

Das, B.B., Liu, F., and Alfano, R.R. (1997) *Rep. Prog. Phys.*, **60**, 227.

Dolin, L.S., Feldchtein, F.I., Gelikonov, G.V. *et al.* (2004) in *Coherent-Domain Optical Methods: Biomedical Diagnostics, Environmental and Material Science*, Vol. 2, (ed. V.V. Tuchin), Kluwer Academic Publishers, Boston, p. 211.

Doornbos, R.M.P., Lang, R., Aalders, M.C. *et al.* (1999) *Phys. Med. Biol.*, **44**, 967.

Driver, I., Lowdell, C.P., and Ash, D.V. (1991) *Phys. Med. Biol.*, **36**, 805.

Duck, F.A. (1990) *Physical Properties of Tissue: a Comprehensive Reference Book*, Academic Press, London.

Dyson, R.D. (1974) *Cell Biology: a Molecular Approach*, Allyn and Bacon, Boston.

Genina, E.A., Bashkatov, A.N., Kochubey, V.I., and Tuchin, V.V. (2005) *Opt. Spectrosc.*, **98**, 470.

Fantini, S., Walker, S.A., Franceschini, M.A. *et al.* (1998) *Appl. Opt.*, **37**, 1982.

Farrell, T.J., Patterson, M.S., and Essenpreis, M. (1998) *Appl. Opt.*, **37**, 1958.

Farrell, T.J. and Patterson, M.S. (2001) *J. Biomed. Opt.*, **6**, 468.

Fishkin, J.B., Coquoz, O., Anderson, E.R. *et al.* (1997) *Appl. Opt.*, **36**, 10.

Frank K. and Kessler, M. (eds) (1992) *Quantitative Spectroscopy in Tissue*, pmi Verlag, Frankfurt am Main.

Fried, D. (2003) in *Biomedical Photonics Handbook*, (ed. T. Vo-Dinh), CRC Press, Boca Raton, pp. 50–51.

Fried, D., Featherstone, J.D.B., Glena, R.E., and Seka, W. (1995) *Appl. Opt.*, **34**, 1278.

Friebel, M., Roggan, A., Müller, G., and Meinke, M. (2006) *J. Biomed. Opt.*, **11**, 034021.

Gebhart, S.C., Lin, W.C., and Mahadevan-Jansen, A. (2006) *Phys. Med. Biol.*, **51**, 2011.

van Gemert, M.J.C., Jacques, S.L., Sterenborg, H.J.C.M., and Star, W.M. (1989) *IEEE Tranc. Biomed. Eng.*, **36**, 1146.

Ghosh, N., Mohanty, S.K., Majumder, S.K., and Gupta, P.K. (2001) *Appl. Opt.*, **40**, 176.

Goodman, J.W. (1985) *Statistical Optics*, Wiley-Interscience Publication, New York.

Graaff, R., Aarnoudse, J.G., Zijp, J.R. *et al.* (1992) *Appl. Opt.*, **31**, 1370.

Graaff, R., Dassel, A.C.M., Koelink, M.H. *et al.* (1993) *Appl. Opt.*, **32**, 435.

Gratton, E., Fantini, S., Franceschini, M.A. *et al.* (1997) *Phil. Trans. R. Soc. Lond. B*, **352**, 727.

Hammer, M., Schweitzer, D., Michel, B. *et al.* (1998) *Appl. Opt.*, **37**, 7410.

Hayakawa, C.K., Spanier, J., Bevilacqua, F. *et al.* (2001) *Opt. Lett.*, **26**, 1335.

Heusmann, H., Kolzer, J., and Mitic, G. (1996) *J.Biomed. Opt.*, **1**, 425.

Henderson, B.W. and Dougherty, T.J. (eds.) (1992) *Photodynamic Therapy: Basic Principles and Clinical Applications*, Marcel-Dekker, New York.

Hogan, M.J., Alvardo, J.A., and Weddel, J. (1971) *Histology of the Human Eye*, W.B. Sanders Co, Philadelphia.

Holboke, M.J., Tromberg, B.J., Li, X. *et al.* (2000) *J. Biomed. Opt.*, **5**, 237.

van de Hulst, H.C. (1980) *Multiple Light Scattering. Tables, Formulas and Applications*, Academic Press, New York.

van de Hulst, H.C. (1981) *Light Scattering by Small Particles*, Dover, New York.

Ishimaru, A. (1997) *Wave Propagation and Scattering in Random Media*, IEEE Press, New York.

Jacques, S.L. (1989) *IEEE Trans. Biomed. Eng.*, **36**, 1155.

Jacques, S.L. (2003) http://omlc.ogi.edu/software/index.html.

Jacques, S.L., Alter, C.A., and Prahl, S.A. (1987) *Lasers Life Sci.*, **1**, 309.

Keijzer, M., Richards-Kortum, R.R., Jacques, S.L., and Feld, M.S. (1989), *Appl. Opt.* **28**(20), 4286.

Kessel, R.G. (1998) *Basic Medical Histology: The Biology of Cells, Tissues, and Organs*, Oxford University Press, New York.

Key, H., Davies, E.R., Jackson, P.C., and Wells, P.N.T. (1991) *Phys. Med. Biol.*, **36**, 579.

Khalil, O., Yeh, S.-J., Lowery, M.G., Wu, X. *et al.* (2003) *J. Biomed. Opt.*, **8**, 191.

Kienle, A., Lilge, L., and Patterson, M.S. (1994) *Proc SPIE*, **2326**, 212.

Knüttel, A. and Boehlau-Godau, M. (2000) *J. Biomed. Opt.*, **5**, 83.

Knüttel, A., Bonev, S., and Knaak, W. (2004) *J. Biomed. Opt.*, **9**, 265.

Kumar, G. and Schmitt, J.M. (1997) *Appl. Opt.*, **36**(10), 2286.

Liu, H. and Xie, S. (1996) *Appl. Opt.*, **35**, 1793.

Laufer, J., Simpson, C.R., Kohl, M. *et al.* (1998) *Phys. Med. Biol.*, **43**, 2479.

Levitz, D., Thrane, L., Frosz, M.H. *et al.* (2004) *Opt. Exp.*, **12**, 249.

Madsen, S.J., Wyst, P., Svaasand, L.O., *et al.* (1994) *Phys. Med. Biol.*, **39**(8), 1191.

Marchesini, R., Bertoni, A., Andreola, S. *et al.* (1989) *Appl. Opt.*, **28**, 2318.

Matcher, S.J., Cope, M., and Delpy, D.T. (1997), *Appl. Opt.* **36**(1), 386.

Mishchenko, M.I., Hovenier, J.W., and Travis, L.D. (eds.) (2000) *Light Scattering by Nonspherical Particles*, Academic Press, San Diego.

Mishchenko, M.I., Travis, L.D., and Lacis, A.A. (2002) *Scattering, Absorption, and Emission of Light by Small Particles*, Cambridge University Press, Cambridge.

Mishchenko, M.I., Travis, L.D. and Lacis, A.A. (2006) *Multiple Sacattering of Light by Particles: Radiative Transfer and Coherent Backscattering*, Cambridge University Press, New York.

Mobley, J. and Vo-Dinh, T. (2003) in *Biomedical Photonics Handbook* (ed. T. Vo-Dinh), CRC Press, Boca Raton, pp. 2–1.

Müller, G., Chance, B., Alfano, R., *et al.* (eds) (1993) *Medical Optical Tomography: Functional Imaging and Monitoring*, **IS11**, SPIE Press, Bellingham, WA.

Müller, G. and Roggan, A. (eds) (1995) *Laser–Induced Interstitial Thermotherapy*, **PM25**, SPIE Press, Bellingham, WA.

Nickell, S., Hermann, M., Essenpreis, M. *et al.* (2000) *Phys. Med. Biol.*, **45**, 2873.

Niemz, H. (1996) *Laser-Tissue Interactions. Fundamentals and Applications*, Springer Verlag, Berlin.

Nilsson, A.M.K., Lucassen, G.W., Verkruysse, W. *et al.* (1997) *Photochem. Photobiol.*, **65**, 366.

Ohmi, M., Ohnishi, Y., Yoden, K., and Haruna, M. (2000) *IEEE Trans. Biomed. Eng.*, **47**, 1266.

Oldham, M., Sakhalkar, H., Oliver, T. *et al.* (2008) *J. Biomed. Opt.*, **13**, 021113.

Patterson, M.S., Chance, B., and Wilson, B.C. (1989) *Appl. Opt.*, **28**, 2331.

Peters, V.G., Wyman, D.R., Patterson, M.S., and Frank, G.L. (1990) *Phys. Med. Biol.*, **35**, 1317.

Prahl, S.A. (2007) http://omlc.ogi.edu/software/iad/index.html.

Prahl, S.A., van Gemert, M.J.C., and Welch, A.J. (1993) *Appl. Opt.*, **32**, 559.

Roggan, A., Dörschel, K., Minet, O., *et al.* (1995) in *Laser–Induced Interstitial Thermotherapy*, **PM25** (eds G. Müller and A. Roggan), SPIE Press, Bellingham, WA, pp. 10–44.

Roggan, A., Friebel, M., Dörschel, K. *et al.* (1999) *J. Biomed. Opt.*, **4**, 36.

Rytov, S.M., Kravtsov, Yu.A., and Tatarskii, V.I. (1989) *Wave Propagation through Random Media*, Principles of Statistical Radiophysics 4, Springer-Verlag, Berlin.

Salomatina, E., Jiang, B., Novak, J., and Yaroslavsky, A.N. (2006) *J. Biomed. Opt.*, **11**, 064026.

Sardar, D.K. and Levy, L.B. (1998) *Lasers Med. Sci.*, **13**, 106.

Schmitt, J.M., Knüttel, A., and Bonnar, R.F. (1993) *Appl. Opt.*, **32**, 6032.

Schmitt, J.M. and Kumar, G. (1998) *Appl. Opt.* **37**(13), 2788.

Schwarzmaier, H.-J., Yaroslavsky, A.N., Yaroslavsky, I.V. *et al.* (1997) *Proc. SPIE*, **2970**, 492.

Silver, F.H. (1987) *Biological Materials: Structure, Mechanical Properties, and Modeling of Soft Tissues*, New York, New York University Press.

Simpson, C.R., Kohl, M., Essenpreis, M., and Cope, M. (1998) *Phys. Med. Biol.*, **43**, 2465.

Sobolev, V.V. (1974) *Light Scattering in Planetary Atmospheres*, Pergamon Press, Oxford.

Star, W.M., Wilson, B.C., and Patterson, M.C. (1992) in *Photodynamic Therapy, Basic Principles and Clinical Applications*, (eds B.W. Henderson and T.J. Dougherty), Marcel-Dekker, New York, p. 335.

Steenbergen, W., Kolkman, R., and de Mul, F. (1999) *J. Opt. Soc. Am. A*, **16**, 2967.

Sterenborg, H.J.C.M., van Gemert, M.J.C., Kamphorst, W. *et al.* (1989) *Lasers Med. Sci.*, **4**, 221.

Suzuki, K., Yamashita, Y., Ohta, K. *et al.* (1996) *J. Biomed. Opt.*, **1**, 330.

Tearney, G.J., Brezinski, M.E., Southern, J.F. *et al.* (1995) *Opt. Lett.*, **20**, 2258.

Thomas, G.E. and Stamnes, K. (1999) *Radiative Transfer in the Atmosphere and Ocean*, Cambridge University Press, New York.

Thueler, F., Charvet, I., Bevilacqua, F. *et al.* (2003) *J. Biomed. Opt.*, **8**, 495.

Tromberg, B.J., Coquoz, O., Fishkin, J.B. *et al.* (1997) *Phil. Trans. R. Soc. Lond. B.*, **352**, 661.

Troy, T.L., Page, D.L., and Sevick-Muraca, E.M. (1996) *J. Biomed. Opt.*, **1**, 342.

Troy, T.L. and Thennadil, S.N. (2001) *J. Biomed. Opt.*, **6**, 167.

Tuchin, V.V. (ed.) (1994) *Selected Papers on Tissue Optics: Applications in Medical Diagnostics and Therapy*, **MS102**, SPIE Press, Bellingham, WA.

Tuchin, V.V. (ed.) (2002) *Handbook of Optical Biomedical Diagnostics*, **PM107**, SPIE Press, Bellingham, WA.

Tuchin, V.V. (2007) *Tissue Optics: Light Scattering Methods and Instruments for Medical Diagnosis*, **PM 166**, 2nd edn, SPIE Press, Bellingham, WA.

Tuchin, V.V., Wang, L.V., and Zimnyakov, D.A. (2006) *Optical Polarization in Biomedical Applications*, Springer-Verlag, New York.

Tuchin, V.V., Xu, X., and Wang, R.K. (2002) *Appl. Opt. – OT*, **41**(1), 258.

Tuchin, V.V., Zhestkov, D.M., Bashkatov, A.N., and Genina, E.A. (2004) *Opt. Exp.*, **12**, 2966.

Turchin, I.V., Sergeeva, E.A., Dolin, L.S., and Kamensky, V.A. (2003) *Laser Phys.*, **13**, 1524.

Vo-Dinh T. (ed.) (2003), *Biomedical Photonics Handbook*, CRC Press, Boca Raton.

Wang, X.J., Milner, T.E., Dhond, R.P. *et al.* (1995) *Opt. Lett.*, **20**, 524.

Wang, L.V. and Wu, H.-I. (2007) *Biomedical Optics: Principles and Imaging*, Wiley-Interscience, Hoboken, NJ.

Welch, A.J. and van.Gemert, M.C.J. (eds.) (1992) *Tissue Optics*, Academic Press, New York.

Yanovitskij, E.J. (1997) *Light Scattering in Inhomogeneous Atmospheres*, Springer-Verlag, Berlin.

Yaroslavsky, A.N., Schulze, P.C., Yaroslavsky, I.V. *et al.* (2002) *Phys. Med. Biol.*, **47**, 2059.

Yaroslavsky, A.N., Yaroslavsky, I.V., Goldbach, T., and Schwarzmaier, H.-J. (1996) *Appl. Opt.*, **35**, 6797.

Yaroslavsky, A.N., Yaroslavsky, I.V., Goldbach, T., and Schwarzmaier, H.-J. (1999) *J. Biomed. Opt.*, **4**, 47.

Zege, E.P., Ivanov, A.P., and Katsev, I.L. (1991) *Image Transfer through a Scattering Medium*, Springer-Verlag, New York.

Zijp, J.R. and ten Bosch, J.J. (1991) *Archs Oral Biol.*, **36**, 283.

Zijp, J.R. and ten Bosch, J.J. (1997) *Appl. Opt.*, **36**, 1671.

Zvyagin, A.V., Silva, K.K.M.B.D., Alexandrov, S.A. *et al.* (2003) *Opt. Exp.*, **11**, 3503.

Further Reading

Ahluwalia, G. (ed.) (2008) *Light Based Systems for Cosmetic Application*, William Andrew, Inc., Norwich, NY.

Baron, E. (ed.) (2009) *Light-Based Therapies for Skin of Color*, Springer, NY.

Drexler, W. and Fujimoto, J.G. (eds.) (2008) *Optical Coherence Tomography: Technology and Applications*, Springer, Berlin.

Drezek, R., Guillaud, M., Collier, T. *et al.* (2003) *J. Biomed. Opt.*, **8**, 7.

Splinter, R. and Hooper, B.A. (2007) *An Introduction to Biomedical Optics*, Taylor & Francis, New York, London.

Tuchin, V.V. (ed.) (2009) *Handbook of Optical Sensing of Glucose in Biological Fluids and Tissues*, CRC Press, Taylor & Francis Group, London.

17
Single-Molecule Fluorescence: Biophysics

Michael Prummer and Christian Hübner

17.1 **Introduction** **629**
17.2 **Photophysical Basics** **630**
17.2.1 Energy Levels and Transitions 630
17.2.2 Energy Transfer 632
17.2.3 The Dye Molecule as a Dipole 634
17.3 **Techniques** **635**
17.3.1 Signal-to-Noise Considerations 635
17.3.2 Wide-Field Fluorescence Microscopy Techniques 636
17.3.2.1 Epifluorescence Microscopy 637
17.3.2.2 TIRF Microscopy 637
17.3.3 Confocal Optical Microscopy 638
17.4 **Applications** **640**
17.4.1 Protein Folding 640
17.4.1.1 The Unfolded State 640
17.4.1.2 Single Molecule Folding/Unfolding Transitions 642
17.4.2 Enzymes 643
17.4.2.1 Static and Dynamic Heterogeneity 643
17.4.2.2 Conformational Changes and Catalytic Activity 646
17.5 **Conclusions** **650**
References **651**

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

17.1
Introduction

The first detection of the fluorescence light of a single molecule dates back to 1976 (Hirschfeld, 1976). At that time, the molecule under investigation contained about 100 fluorescent chromophores. With the advent of extremely sensitive photon detectors and effective filters, the detection of the fluorescence of just a single fluorophore became possible in 1990 (Orrit and Bernard, 1990), however, in a crystalline host at liquid helium temperature. In the same year, single fluorophores were observed in solution at room temperature (Shera *et al.*, 1990), paving the way to the application of single-molecule fluorescence on biological systems in their native surrounding. The next milestones for the perception of single-molecule fluorescence as a useful biophysical technique were the demonstration of fluorescent resonance energy transfer on an individual basis (Ha *et al.*, 1996) and the direct observation of the diffusion of a dye-labeled protein in a lipid bilayer (Schmidt *et al.*, 1995). These two key experiments sparked a tremendous development of the technique over the following years, and already in 2001 the coming of age of single-molecule spectroscopy was stated (Kelley, Michalet and Weiss, 2001). Today, the number of groups performing single-molecule fluorescence detection as well as the number of publications in the field is rapidly growing. The pioneers of single-molecule fluorescence have set out their children, and the grand children are starting their own groups. This chapter therefore cannot be comprehensive, but rather a selection of prototype experiments demonstrating the enormous potential of single-molecule fluorescence for molecular protein biophysics, with an emphasis on protein conformation in the context of protein folding and catalysis. Owing to space limitations, the fields of single-molecule observations of molecular motors and of membrane diffusion are omitted.

Thereby, the aspects of the respective applications will be balanced by the methodological developments. The aim of this approach is to provide the reader with a framework of knowledge that allows for deciding whether a single-molecule approach to a certain problem might be advantageous, and which method is to be preferred.

The basic photophysical principles underlying fluorescence detection, particularly at the high excitation rates necessary

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

to observe the faint light of a single dye molecule, are outlined in the second section. The third section briefly introduce the experimental layouts used, emphasizing their advantages and drawbacks. Finally, selected applications are discussed in the fourth section.

17.2
Photophysical Basics

17.2.1
Energy Levels and Transitions

The fluorescence of single dye molecules is based on the same energy level scheme as ordinary molecular bulk fluorescence, consisting of, in the simplest case, an electronic ground state, an excited state, and a triplet state. In each electronic level, several vibronic states exist. The high repetition rates needed to obtain a measurable signal, however, demand a second look at the energy levels and transitions involved. Besides the levels found in every textbook, higher excited states are shown in the Jablonski diagram in Figure 17.1. Those higher levels exist both in the singlet as well as in the triplet manifold. At high laser intensities, in particular in pulsed excitation, further excitations from the states S_1 and T_1 are possible.

First taking only transitions between the electronic ground state S_0, the first excited singlet state S_1, and the first excited triplet state T_1 into account and neglecting stimulated emission, the following set of rate equations can be derived:

$$\begin{aligned}\dot{n}_1 &= -I\sigma n_1 + k_{\text{rad}} n_2 \\ &\quad + k_{\text{ic}} n_2 + k_{\text{risc}} n_3 \\ \dot{n}_2 &= I\sigma n_1 - k_{\text{rad}} n_2 - k_{\text{ic}} n_2 - k_{\text{isc}} n_2 \\ \dot{n}_3 &= k_{\text{isc}} n_2 - k_{\text{risc}} n_3 \end{aligned} \tag{17.1}$$

where n_1, n_2, and n_3 are the occupation probabilities of S_0, S_1, and T_1, respectively, k_{rad} the radiative rate, k_{ic} the rate of internal conversion, k_{isc} the intersystem crossing rate, k_{risc} the reverse intersystem crossing rate, I the excitation intensity, and σ the absorption cross section of the fluorophore.

The quantum yield of fluorescence Φ_f is the ratio of radiative transitions from

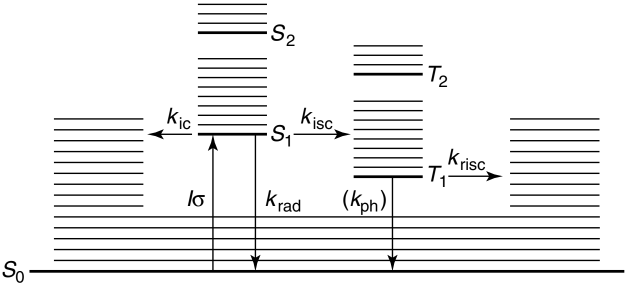

Fig. 17.1 Jablonski diagram showing the electronic ground state S_0, the first electronically and second excited singlet states S_1 and S_2, the first and the second electronically excited triplet states T_1 and T_2, and the most important transitions between the states as detailed in the text.

S_0 to all other transitions from that state:

$$\Phi_f = \frac{k_{\mathrm{rad}}}{k_{\mathrm{ic}} + k_{\mathrm{isc}} + k_{\mathrm{rad}}} = k_{\mathrm{rad}}\tau_f \qquad (17.2)$$

where τ_f is the fluorescence lifetime, which is the experimentally accessible time constant of the decay of fluorescence after pulsed excitation. Therefore, the quantum yield of fluorescence can be determined by measuring the fluorescence lifetime and deriving k_{rad} from the absorption and emission spectra of the dye making use of the relation between the Einstein coefficients for absorption and spontaneous emission (Strickler and Berg, 1962):

$$k_{\mathrm{rad}} = 2.88 \times 10^{-9} n^2 \frac{\int F(\bar{\nu})d\bar{\nu}}{\int F(\bar{\nu})d\bar{\nu}/\bar{\nu}^3} \times \int \frac{\varepsilon(\bar{\nu})}{\bar{\nu}} d\bar{\nu} \qquad (17.3)$$

where $F(\bar{\nu})$ is the fluorescence and $\varepsilon(\bar{\nu})$ the absorption spectrum of the dye, with $\bar{\nu}$ the wave number in cm^{-1}. Fluorophores for single-molecule spectroscopy should have quantum yields of fluorescence close to unity to ensure a strong enough fluorescence signal to be detected. Therefore, the intersystem crossing rate as well as the rate of internal conversion should be small.

Under stationary conditions (all time derivatives are zero), the emission rate R as a function of the excitation rate from the set of rate equations (17.1) is given by

$$R = n_2 k_{\mathrm{rad}} = \frac{k_{\mathrm{rad}}}{1 + \frac{k_{\mathrm{isc}}}{k_{\mathrm{risc}}} + \frac{k_{\mathrm{rad}} + k_{\mathrm{ic}} + k_{\mathrm{ies}}}{I\sigma}} \qquad (17.4)$$

Owing to increasing population of the triplet state T_1 at high excitation rates, the maximum achievable emission rate with infinitely strong excitation is R_∞:

$$R_\infty = \frac{k_{\mathrm{rad}}}{1 + \frac{k_{\mathrm{isc}}}{k_{\mathrm{risc}}}} \qquad (17.5)$$

The emission rate R can be rewritten as

$$R = R_\infty \frac{1}{\left(1 + \frac{I_{\mathrm{sat}}}{I}\right)} \qquad (17.6)$$

with the saturation intensity I_{sat}, which is the excitation intensity where half of the maximum emission rate is reached, given by

$$I_{sat} = \frac{k_{\mathrm{risc}}}{\sigma} \frac{k_{\mathrm{rad}} + k_{\mathrm{ic}} + k_{\mathrm{isc}}}{k_{\mathrm{risc}} + k_{\mathrm{isc}}} \qquad (17.7)$$

The saturation of the average emission rate due to triplet excursions is known as the *triplet bottleneck*.

The solution to the set of rate equations 17.1 with the initial condition of $n_1 = 1$, $n_2 = n_3 = 0$ provides the time evolution of the occupation probability of the first electronically excited singlet state $n_2(t)$, which is related to the fluorescence intensity autocorrelation function or second-order autocorrelation function $g^{(2)}(t)$ defined as

$$g^{(2)}(t) = \frac{\langle n(\tau)n(t+\tau)\rangle_\tau}{\langle n(\tau)\rangle_\tau^2} = \frac{n_2(t)}{\langle n_2\rangle} \qquad (17.8)$$

and reads

$$g^{(2)}(t) = 1 - \frac{k_{\mathrm{isc}}^{\mathrm{eff}} + k_{\mathrm{risc}}}{k_{\mathrm{risc}}} e^{-(k_{\mathrm{rad}}+2I\sigma)t}$$

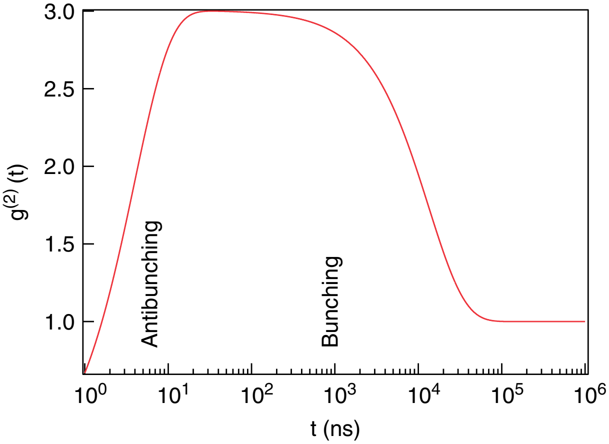

Fig. 17.2 Intensity autocorrelation function of the fluorescence of a single dye molecule under room-temperature excitation.

$$+\frac{k_{\text{isc}}^{\text{eff}}}{k_{\text{risc}}}e^{-(k_{\text{isc}}^{\text{eff}}+k_{\text{risc}})t} \tag{17.9}$$

where the effective intersystem crossing rate $k_{\text{isc}}^{\text{eff}}$ is given as

$$k_{\text{isc}}^{\text{eff}} = \langle n_2\rangle k_{\text{isc}} \tag{17.10}$$

The theoretical autocorrelation function of a single dye molecule according to Eq. 17.10 is shown in Figure 17.2 on a logarithmic timescale. This function drops to zero at the temporal origin, which is intuitively understood taking into consideration that a single quantum system can never emit two photons simultaneously. This behavior is referenced to as photon antibunching. Thereafter, the autocorrelation function takes a value above 1, indicating that the occupation probability of S_1 at this time is higher than on average. This is because the system will always be excited first to S_1, but on average the occupancy $\langle n_2\rangle$ is limited by the triplet bottleneck. Because of triplet excursions, a single molecule emits its photons in bunches (cycling in the singlet manifold under photon emission), separated by dark periods (when the molecule is shelved in the triplet state). Therefore, the second term in Eq. 17.10 is referred to as the *photon bunching*. If dynamical processes are to be studied by single-molecule fluorescence, care has to be taken that photophysical effects do not lead to misinterpretations.

17.2.2
Energy Transfer

Energy transfer between two or more fluorescent entities is a major subject in single-molecule fluorescence. It plays an important role in light harvesting complexes and, on the other hand, is the basis of one of the most popular single-molecule techniques, fluorescence resonance energy transfer (FRET), which is referred to as a *molecular ruler* (Stryer and Haugland, 1967) owing to its capability to measure distances between two fluorophores by spectral means.

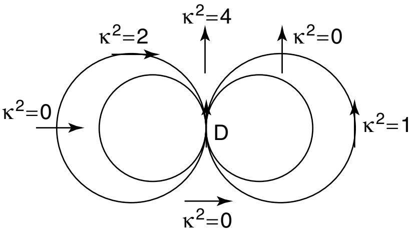

Fig. 17.3 Orientation factors κ^2 for different cases of relative donor–acceptor placement.

In FRET, one fluorophore, called *donor*, is excited, and the excitation energy is radiationless transferred to a nearby fluorpohore, called *acceptor*. The two chromophores may be chemically identical (homotransfer) or distinct (heterotransfer). The transfer mechanism is based on the dipole–dipole interaction between the two fluorophores in the weak coupling regime, that is, the interaction energy is small compared to the transition energies. The transfer efficiency theoretically derived by Förster (Förster, 1946) depends on the relative location and orientation of the transition dipoles of the fluorophores, their distance R, the refractive index of the medium n in the vicinity of both, and the overlap integral J. The latter takes account for the resonance condition and is the integral of the product of the normalized emission spectrum of the donor $F_D(\lambda)$ and the absorption spectrum (molar extinction coefficient in $\mathrm{cm}^{-1}M^{-1}$) $\varepsilon_A(\lambda)$ of the acceptor:

$$J = \int_0^\infty F_D(\lambda)\varepsilon_A(\lambda)\lambda^4 \, d\lambda \tag{17.11}$$

The transfer rate, that is, the rate with which the excited donor transfers its energy to the acceptor, is given as

$$k_t = 8.8 \times 10^{-23}\mathrm{mol}\frac{1}{\tau_\mathrm{D}}\frac{1}{\mathrm{R}^6}\frac{\kappa^2\Phi_\mathrm{D}}{\mathrm{n}^4}\mathrm{J}$$

$$= \frac{1}{\tau_D}\left(\frac{R_0}{R}\right)^6 \tag{17.12}$$

with the distance R taken in Ångstrom and R_0 the Förster radius. Here, the relative position and orientation of the dyes are taken into account via the orientation factor κ^2. Values of κ^2 for some cases are shown in Figure 17.3. The orientation dependence of the energy transfer rate may jeopardize distance measurements. Therefore, a free rotation of the fluorophores on a timescale faster than the fluorescence lifetime is aimed at, leading to an average orientation factor $\langle\kappa^2\rangle = 2/3$. Linkers with several carbon atoms are used to assure unhindered rotation of the dye. However, interactions between the dye and the protein may lead to blocking of the free rotation. Fluorescence anisotropy measurements provide information on the rotational flexibility of the dyes and should therefore always be conducted as a control. Fortunately, even a partial blocking of rotational freedom does not significantly alter the average orientation factor. If, for example, one of the dyes cannot rotate at all while the other one is free to rotate, the average orientation factor takes the value $\langle\kappa^2\rangle = 0.63$ (Berberan-Santos and Prieto, 1988).

In the experiment, the transfer rate is not directly accessible but rather the transfer efficiency, which is the fraction of transferred excitation quanta, and which is related to the distance R as

$$E_t = \frac{1}{1 + \left(\frac{R}{R_0}\right)^6} \tag{17.13}$$

Although in ensemble experiments the transfer efficiency is measured via comparison of donor fluorescence in the presence and absence of the acceptor (e.g., by photobleaching the acceptor), in single-molecule experiments the fluorescence signals of the donor F_A and the acceptor F_D are utilized:

$$E_t = \frac{F_A}{\gamma F_D + F_A} \tag{17.14}$$

The correction factor γ takes into account the quantum efficiencies of fluorescence of the donor and the acceptor and their respective detection efficiencies. If there is direct excitation of the donor (i.e., not FRET mediated), the acceptor signal needs to be corrected by subtraction of that signal. A very robust method to determine the transfer efficiency is based on the fluorescence lifetime of the donor τ_{D0} and τ_{DA} in the absence and presence of the acceptor:

$$E_t = 1 - \frac{\tau_{DA}}{\tau_{D0}} \tag{17.15}$$

The fluorescence lifetime of the donor without the acceptor is determined in a control sample.

In single-molecule FRET experiments, the photophysics of the dyes has to be treated carefully. Although triplet excursions of the donor simply render the whole donor–acceptor pair dark, the situation is more complex if the acceptor is in the triplet state. If only one excited singlet and triplet states are considered, one would expect that no energy transfer occurs and that all excitation quanta lead to an emission of the donor in this case, resulting in an anticorrelated fluctuation of the donor and acceptor fluorescence. The existence of higher excited states for the acceptor, however, in the singlet as well as in the triplet system, gives rise to other pathways for the excitation energy. In particular, FRET-type energy transfer can result in an excitation of the acceptor to a higher triplet state T_n. Depending on the efficiency of this transfer process, the donor fluorescence may increase, stay constant, decrease, or even completely vanish when the acceptor is triplet shelved (Maus *et al.*, 2001). The latter process is known as *singlet–triplet annihilation* (*STA*).

17.2.3
The Dye Molecule as a Dipole

Classically, a dye molecule can be treated as a dipole antenna. Therefore, its absorption and emission has the same anisotropy. Only the component of the electric field parallel to the absorption dipole axis can excite the fluorophore, and the probability of photon emission has the same shape as the emission intensity of a dipole antenna. This holds in an isotropic medium as well as in the vicinity of an interface between media with different refractive indices. If a fluorophore is, for example, located close to an interface to a medium with a higher refractive index (imagine a fluorophore in an aqueous surrounding close to a glass coverslip), most of its fluorescence is emitted toward the interface (Figure 17.4). Additionally, the

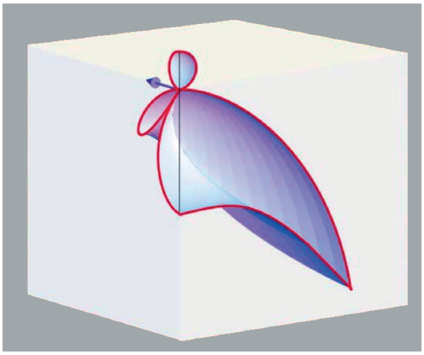

Fig. 17.4 Fluorescence emission characteristics of a dipole close to an interface to a medium with higher refractive index.

light emitted by a fluorophore is linearly polarized in the dipole direction. The anisotropy as well as the polarization of the emission can be exploited for the determination of the orientation of the dipole in space or even for the relative orientation of the dipoles in a donor–acceptor pair (Hübner *et al.*, 2004).

17.3
Techniques

17.3.1
Signal-to-Noise Considerations

The key prerequisite for successful single-molecule fluorescence experiments is a high signal-to-noise (SNR) ratio. The fluorescence signal in single molecule detection (SMD) is typically the photon count within a given time interval. Owing to the stochastic nature of photon emission, this photon number is Poisson distributed. The noise of the signal therefore is the shot-noise of the total number of detected photons. The total number of detected photons is composed of the signal itself and background contributions. The latter can be further divided into background, induced by the excitation laser, and the background signal coming from the detectors in the absence of any source of light, the dark signal.

The laser-induced background originates from elastic Rayleigh and inelastic Raman scattering, as well as from fluorescence in the optical system and the solvent where the molecule under observation is embedded. Its contribution is linearly dependent on the laser intensity. The dark signal, on the other hand, is a constant. The total SNR ratio can be given as

$$SNR = \frac{\eta_{\text{det}} \cdot R \cdot T_{\text{int}}}{\sqrt{\eta_{\text{det}} \cdot R \cdot T_{\text{int}} + C_b \cdot I_L \cdot T_{\text{int}} + N_d \cdot T_{\text{int}}}} = \frac{\eta_{\text{det}} \cdot R}{\sqrt{\eta_{\text{det}} \cdot R + C_b \cdot I_L + N_d}} \cdot \sqrt{T_{\text{int}}} \tag{17.16}$$

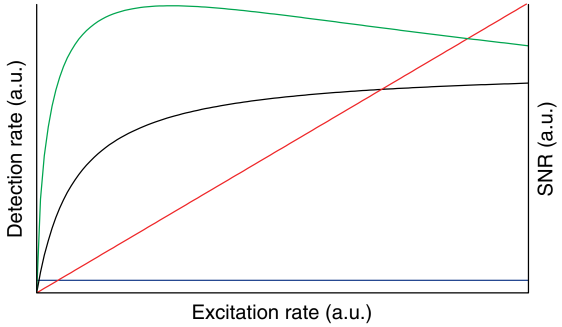

Fig. 17.5 Signal-to-noise ratio as a function of exitation laser intensity according to Eqs 17.16 and 17.6 (green). The fluorescence intensities of the single molecule (black), the dark signal (blue), and the laser-induced background (red) are also plotted.

where R is the emission rate of the fluorophore, T_{int} is the integration time, η_{det} is the detection efficiency, C_b is the constant of proportionality between laser intensity I_L and laser-induced background photon rate, and N_d is the dark count rate of the detector.

Saturation of the fluorescence signal leads to an optimum excitation rate for a given integration time where the SNR reaches its maximum, as shown in Figure 17.5.

17.3.2
Wide-Field Fluorescence Microscopy Techniques

In wide-field fluorescence microscopy, the whole field of view of an optical microscope is illuminated to excite fluorescence, and the whole field of view is also imaged onto a camera. The great advantage of wide-field microscopy is that, on the one hand, many molecules can be observed in parallel and that, on the other hand, the trace of a moving molecule can be followed. However, the time resolution is limited by the cameras in use.

For single-molecule fluorescence imaging, a very sensitive camera is necessary. If slow processes are to be followed, back-thinned cooled charge-coupled device (CCD) cameras are the best choice. Those CCDs convert incoming photons into electrons with an efficiency close to unity. However, the readout process of the charges introduces noise of typically a few electrons. Therefore, single-photon detection is not possible, and longer exposure times are necessary.

If single photons are to be detected with a camera and a short integration time is required, intensified cameras or electron multiplying CCDs (EMCCDs) can be used. Intensified CCDs have an image intensifier coupled to the CCD camera. The image intensifier is a multichannel plate photomultiplier (MCP-PM) with a phosphor behind the anode. Thus, an impinging photon is first converted into a photoelectron, which is multiplied in the PM, and the electron avalanche is back converted into light. The light from the phosphor is then detected by the CCD. Typical quantum efficiencies of the photocathodes of

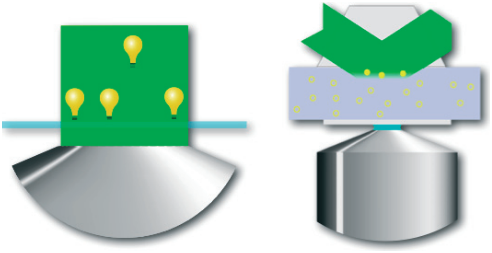
(a) (b)

Fig. 17.6 Sketch of epifluorescence (a) and TIRF (b) illumination modes in wide-field microscopy.

image intensifiers are in the 10–20% range.

A great breakthrough in camera technology was the introduction of an electron multiplying or gain register in the CCD. In this register, the single charges are multiplied in an avalanche process principally similar to a PM. Thus, the single electron generated by a single photon is multiplied to a number that is higher than the readout noise and can therefore be detected. EMCCDs have the same high quantum efficiency of photon detection as CCDs close to unity.

The time resolution of a CCD camera is typically limited by the readout process. However, if only a short sequence of images is to be taken, the charges can be first shifted on the CCD chip and subsequently read out. Depending on the size of the image, exposure times in the submillisecond range are possible this way.

With respect to excitation of fluorescence, two different illumination modes can be applied, namely, conventional epifluorescence and total internal reflection (TIR) excitation as detailed later.

17.3.2.1 Epifluorescence Microscopy

The simplest way to observe the fluorescence of single fluorophores is by using a conventional epifluorescence microscope. Here, the whole field of view is illuminated by parallel laser light (Figure 17.6 (a)). Therefore, the laser is focused into the back focal plane of the microscope objective. This technique, however, requires a sample that is thinner than the focal depth of the microscope; otherwise, out-of-focus molecules would give rise to a high background signal. Such a thin sample is, for example, molecules tethered on the surface of a coverslip or embedded in a thin polymer film. A cell membrane also represents a thin sample, so that epifluorescence microscopy is the method of choice to study, for example, membrane diffusion (Schütz *et al.*, 2000).

17.3.2.2 TIRF Microscopy

If the requirement of a thin sample for epifluorescence microscopy is not fulfilled, illumination of out-of-focus molecules can be circumvented by TIR excitation. The excitation laser is totally reflected typically at a glass–water interface and only fluorophores in the evanescent field at the interface are excited. Depending on the angle of incidence and the wavelength of the laser, the evanescent field penetrates some 10 nm into the aqueous medium. The small penetration depth is advantageous with respect to lowering the background, but, on the other hand, prevents observation of features further away from the interface.

Two different realizations for TIRF excitation are used: objective-type TIRF and prism-type TIRF.

In objective-type TIRF, illumination is carried out similar to epifluorescence by focusing the laser in the back focal plane of an oil immersion microscope objective. In contrast to epifluorescence, this focus is offset from the optical axis so that the emerging parallel light is tilted by an angle larger than the critical angle for total reflection at the glass–water interface of 61°. The highest achievable tilt with an $\mathrm{NA} = 1.4$ oil immersion objective is 67°. Dedicated TIRF objectives have $\mathrm{NA} = 1.45$ corresponding to a maximum tilt angle of 73 °. The advantage of objective-type TIRF is that the fluorescence is collected trough the glass cover slip, in a direction where most of the fluorescence is emitted (cf. Figure 17.4). Thus, a high detection efficiency is assured.

In prism-type TIRF (Figure 17.6(b)), the laser is directed into a prism and the excitation occurs on the prism surface with TIR. In contrast to objective TIRF, the fluorescence is collected through the aqueous medium. Thus, the detection efficiency is considerably lower. The advantage of prism-type TIRF is the higher flexibility in the choice of the angle of incidence of the laser light.

17.3.3
Confocal Optical Microscopy

The main principle behind confocal optical microscopy invented by Marvin Minsky in 1957 (Minsky, 1957) is that a point light source and a point detector are both being imaged in the same position in the sample. The point spread function (PSF) of the microscope, that is, the image that is obtained from a point in the sample, therefore is the product of the two PSFs of illumination and detection. Typically, the microscope objective is used for both focussing the laser and detecting the fluorescence in an epi-illumination scheme, and excitation and fluorescence light are separated in a dichroic beam splitter as can be seen in Figure 17.7. The product of the excitation and detection PSFs defines the confocal volume of about 0.1 fl. The shape of the confocal volume depends on the illumination conditions, but can be approximated quite well by a three-dimensional Gaussian intensity distribution. The advantage of the smaller PSF is not primarily the slightly higher resolution but rather a considerable reduced background. Laser-induced background signal is in a first approximation proportional to the detection volume. Compared to wide-field microscopy, this volume is reduced by a factor of about $2^{3/2}$.

To obtain an image, the confocal volume is raster scanned across the sample. In contrast to wide-field microscopy, where all points of the image are obtained in parallel in a single exposure, in confocal microscopy the image is built in a serial manner, point by point. This results in a drastically prolonged time to record a single image as compared to wide-field microscopy.

To build the image, either the sample or the confocal volume is moved. For sample scanning, piezo-driven scan stages are used, whereas beam or laser scanning is realized by galvanometer mirror scanners. For the latter, a telecentric lens system generates an image of the

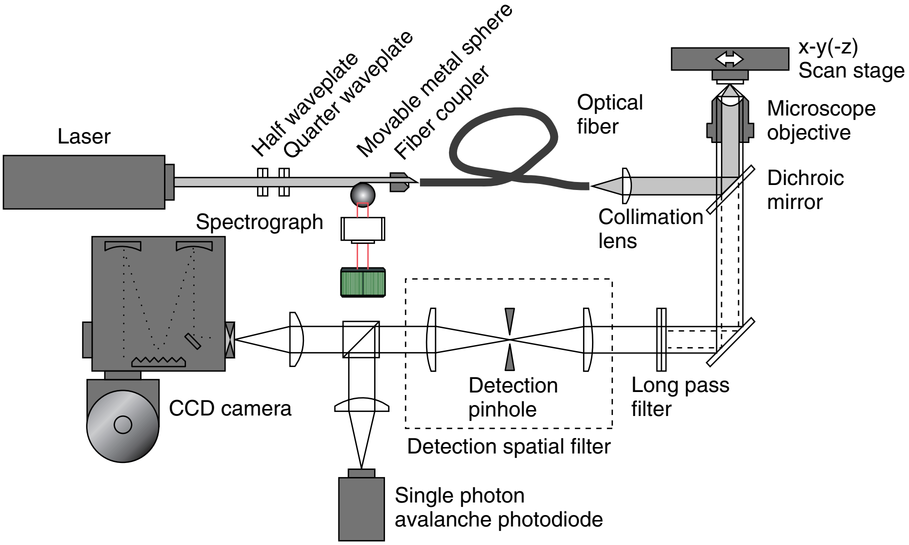
Half waveplate
Quarter waveplate
Movable metal sphere
Fiber coupler
Laser
Optical fiber
x-y(-z) Scan stage
Microscope objective
Dichroic mirror
Spectrograph
Collimation lens
Detection pinhole
Long pass filter
CCD camera
Detection spatial filter
Single photon avalanche photodiode

Fig. 17.7 Schematic drawing of a sample scanning confocal optical microscope.

scanning mirrors in the back focal plane of the microscope objective. Thus, a tilt of the scanning mirror results in a shift to the confocal volume in the sample.

Sample scanning has the clear advantage that the light always propagates on the optical axis through the microscope objective. Therefore, aberrations due to tilt of the beam with respect to the optical axis such as coma or astigmatism are avoided. The allowed size of the image is only limited by the scan range. In contrast, in beam-scanning microscopes the size of the image is limited by the microscope objective. On the other hand, beam scanning also has several advantages: the achievable scan speed is considerably higher, the field of view can be changed by using microscope objectives with different magnifications, and the sample can be mounted in a cryostat.

The pointlike light source required for confocal imaging is realized either by the use of an excitation pinhole or by feeding the laser through a single-mode optical fiber. A fiber typically guarantees a perfect Gaussian mode, whereas a pinhole is advantegeous with respect to polarization control.

The other element defining the confocal volume is the detection pinhole. For single-molecule fluorescence experiments, the size of the confocal pinhole is chosen with respect to minimal loss of fluorescence signal. A diameter corresponding to the diameter of the airy disc turned out to be a good choice.

Key elements of the confocal microscope are the dichroic or multichroic beam splitter and the emission filters. Although in the past holographic notch filters were primarily used to block the back-scattered laser light, the development of even better dielectric filters has rendered them a cost-effective alternative. In particular, sputtered dielectric filters convince with steep edges, high optical density at the laser wavelength, high transmission in the pass band, and their ruggedness. Care should be taken that the filters are subpixel registered, that is, that they do not introduce beam deviations.

The confocal optical microscope is used in two different regimes for single-molecule experiments: in solution experiments, the confocal volume is placed stationary in the sample solution, and freely diffusing molecules lead to fluorescence bursts when they cross it. Thus, distributions of fluorescence properties of the molecules are obtained. If there is dynamics present on a timescale faster than the transit time, this dynamics can be studied. Dynamics on longer timescales, however, is not accessible with this method.

In immobilized experiments, the Brownian motion is prohibited by appro priate immobilization techniques such as embedding in a gel or tethering to a surface. The confocal volume is placed on a particular molecule localized earlier by an image scan, and the (transient) fluorescence is recorded. In this manner, dynamical processes on all timescales can be studied.

17.4
Applications

17.4.1
Protein Folding

The mechanism by which the unstructured polypeptide chain of a protein finds its unique structure remains one of the great mysteries in protein biophysics. It is believed that there is not one single pathway but rather a bunch of ways from the ensemble of unstructured chains to the final three-dimensional structure. This heterogeneity is an appealing target for single-molecule studies. Its ability to resolve subensembles allows for the study of the unfolded state and of folding kinetics and chain dynamics under equilibrium conditions, where either of the folded or unfolded state is weakly populated.

17.4.1.1 The Unfolded State

The best-studied subject in protein folding with single molecules is the unfolded state. The interest for the unfolded state arises from the idea that knowledge about its conformation and dynamics under nativelike conditions provides information about the initial steps on the folding pathway. In the first experimental single-molecule FRET study, the well-investigated protein chymotrypsin inhibitor 2 was specifically labeled with tetramethylrhodamine as the donor and Cy5 as the acceptor (Figure 17.8(a)) (Deniz *et al.*, 2000). The confocal volume was placed in the sample solution, and brownian diffusion gave rise to a burst of fluorescence each time a molecule was passing by. Each burst was analyzed with respect to the transfer efficiency and added to a histogram. In this manner, a distribution of FRET efficiencies is obtained. The histograms shown in Figure 17.8(b) consist of two distinct, rather broad peaks, corresponding to the folded state (high transfer efficiency) and the unfolded state (low transfer efficiency). From the areas under the peaks, the concentrations of the folded and unfolded state can be directly obtained. Strikingly, the center position of the unfolded peak shifts to higher transfer efficiencies with decreasing denaturant concentration, as can be seen in Figure 17.8(c). This shift is indicative of a compaction of the unfolded polypeptide chain under nativelike conditions.

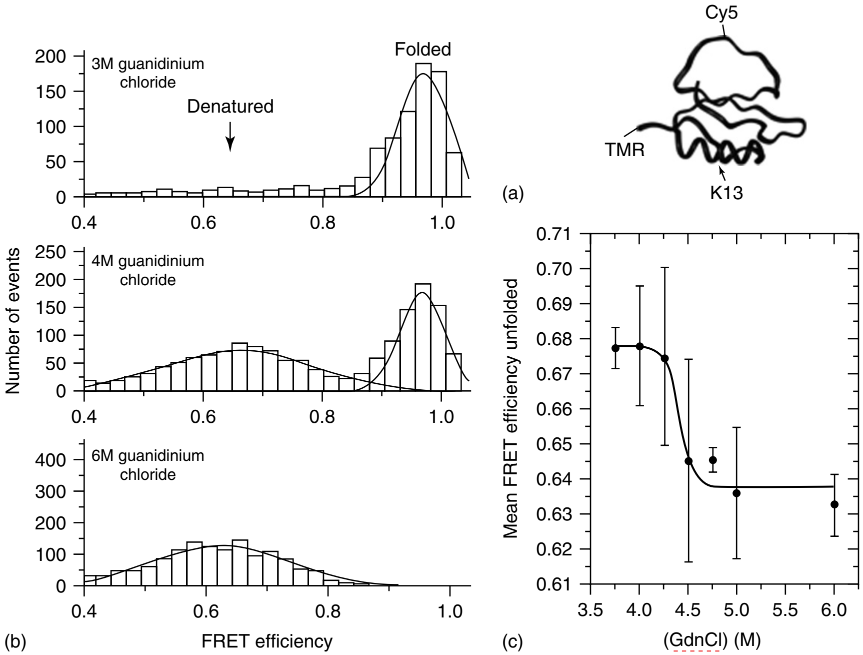

Fig. 17.8 (a) A structural sketch of chymotrypsin inhibitor 2 showing the labeling sites with tetramethylrhodamine as the donor and Cy5 as the acceptor. (b) Energy transfer histograms for different GdnCl concentrations. (c) Mean FRET efficiency as a function of denaturant concentration. (Adapted from Deniz *et al.* (2000).)

Using the same methodology, Schuler *et al.* studied the folding transition of the cold shock protein from *Thermotoga maritima* (Schuler, Lipman and Eaton, 2002). Csp*Tm* was labeled with a donor and an acceptor dye (Figure 17.9(a) and (b)), and single-molecule bursts were analyzed for their FRET efficiency and compiled in a histogram shown in Figure 17.9(c). Again, a compaction of the unfolded chain is observed (Figure 17.9(d)). Furthermore, from the finding that the width of the transfer efficiency distribution is shot-noise limited, it can be concluded that the reconfiguration time is considerably shorter than the transit time of the molecule through the focus. Thus, an upper limit for the reconfiguration time for the polypeptide chain is obtained, which allows for estimating a lower bound for the energy barrier height between the unfolded and folded state of the protein.

Recently, Hofmann *et al.* have studied the influence of different denaturants, namely, guanidinium chloride (GdnCl) and urea, on the collapse of the unfolded chain of the model protein barstar (Hofmann *et al.*, 2008). Using the same FRET pair as Schuler *et al.*, the subpopulation-resolved transfer efficiencies were measured as a function of denaturant concentration. It was found that for low denaturant concentrations, in the salt GdnCl the chain is more compact than in urea. One possible explanation for this finding is that the charges in the polypeptide chain are shielded in GdnCl, thus reducing electrostatic repulsion. This

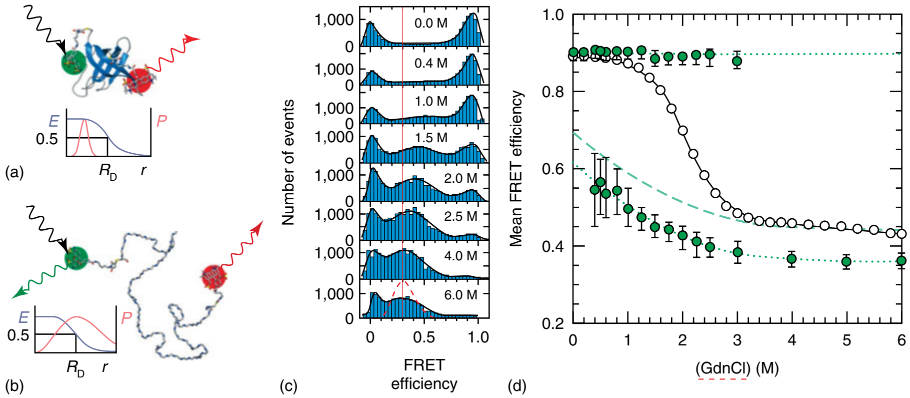

Fig. 17.9 The doubly labeled protein CsP*Tm* in the folded (a) and unfolded (b) state additionally showing the label-distance distribution and the distance-dependent FRET efficiency. The donor dye AlexaFluor 488 is shown in green and the acceptor AlexaFluor 594 in red. (c) The distributions of apparent FRET efficiencies of single CsP*Tm* molecules diffusing through the confocal volume for increasing GdnCl concentrations. The three peaks can be associated to (from left to right) molecules missing: an acceptor dye, unfolded, and folded molecules, respectively. (d) Mean transfer efficiencies derived from the histograms for the folded and unfolded population (the open circles). (Adapted from Schuler, Lipman and Eaton (2002).)

interpretation was further evidenced by unfolding data in urea and the addition of salt, showing a clear compaction of the chain upon addition of salt. This experiment shows that electrostatic interactions are important determinants of the conformation of the unfolded peptide chain.

The dynamics of the unfolded chain of donor–acceptor-labeled Csp*Tm* was directly monitored by Nettels *et al.*, with a sophisticated technique based on a Hanbury-Brown and Twiss detection scheme (Nettels *et al.*, 2007). With this method, the intensity autocorrelation function can be measured down to the nanosecond timescale by circumventing the dead time of the detectors using a pair of them. Chain dynamics associated with fluctuations of the donor–acceptor distance correspondingly results in intensity fluctuations of the donor and acceptor. Thus, the intensity autocorrelation shows a decay on the timescale of the fluctuations. Figure 17.10 shows the donor and acceptor interphoton-time histograms for the unfolded population featuring photon antibunching on the nanosecond timescale typical for a single emitting quantum system. Strikingly, an additional decay of the correlation function is found. The time constant of this exponential decay is 44 nanoseconds, being the reconfiguration time of the chain, which is related to the attempt frequency to cross the energy barrier separating the unfolded and folded state. Thus, single-molecule photon statistics is a valuable tool for the study of chain dynamics associated with the folding reaction.

17.4.1.2 Single Molecule Folding/Unfolding Transitions

A direct observation of folding/unfolding transitions succeeded for the protein adenylate kinase from *Escherichia coli*, immobilized by encapsulation in

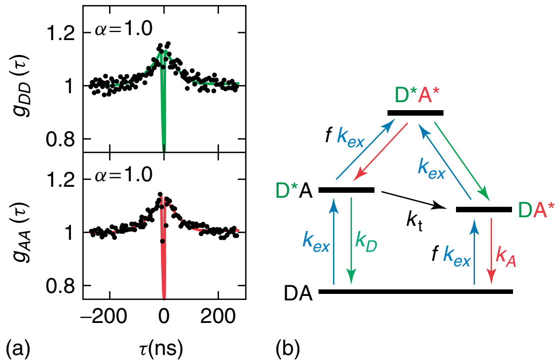

Fig. 17.10 (a) Interphoton-time histogram of donor (top) and acceptor (bottom) fluorescence showing antibunching at short times and a correlation on the 100-nanoseconds timescale. (b) Photophysical model with the ground states A and D and the respective excited states A* and B*. Fluctuations of k_t lead to the decay in the correlation functions shown in (a). (Adapted from Nettels *et al.*, (2007).)

surface-tethered vesicles (Rhoades, Gussakovsky and Haran, 2003). The protein is labeled with a donor–acceptor pair, and their fluorescence is transiently recorded as a function of time. Therefore, the confocal volume is centered on a vesicle with an encapsulated protein. Surprisingly, rather slow transitions between the two states are occasionally observed (Figure 17.11(a)). These slow transitions might indicate the search for the folding pathway in the rugged energy landscape between the folded and unfolded state. Further evidence for the existence of local minima in this landscape is provided by the finding of small steps in the transfer efficiency indicating partial folding/unfolding events (Figure 17.11(b)). Furthermore, there is a broad distribution of step sizes for both, folding and unfolding transitions (Figure 17.11(c)). The stochastic nature of the walk on the energy landscape renders it impossible to observe those in ensemble experiments, clearly pinpointing the advantage of the single-molecule approach.

17.4.2
Enzymes

Single-molecule investigations of enzymatic activity are very appealing for a number of reasons. The most obvious one is common to all single-molecule experiments: the study of heterogeneity. Single-molecule fluorescence, on the other hand, also allows measurements of enzyme kinetics under equilibrium conditions. Furthermore, under nonequilibrium conditions, the system may remain stationary, that is, the concentration of educts and products stays almost constant, due to the low enzyme concentration in single-molecule studies. Eventually, the investigation of conformational changes associated with enzymatic activity may elucidate the mechanisms of catalysis.

17.4.2.1 Static and Dynamic Heterogeneity

The prototype experiment for the study of static and dynamic heterogeneity in enzyme activity was performed by Xie and

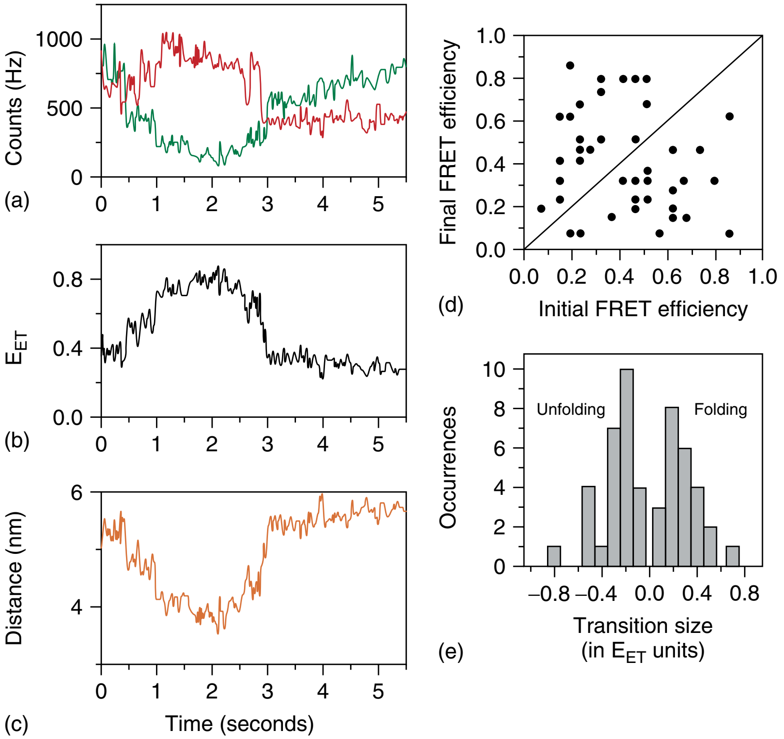

Fig. 17.11 (a) Transient donor and acceptor intensities for a single adenylate kinase molecule inside a vesicle. The anticorrelated behavior is indicative of a change in FRET efficiency shown in (b), which can, in turn, be related to a donor–acceptor distance change shown in (c). (d) Distribution of initial/final FRET efficiencies for folding (below the diagonal line) and unfolding (above the diagonal line) transitions. (e) Distribution of FRET efficiency step sizes of folding/unfolding. (Adapted from Rhoades, Gussakovsky and Haran (2003).)

coworkers with the enzyme cholesterol oxidase (COx) (Lu, Xun and Xie, 1998). FAD, the cofactor of this enzyme, is fluorescent in the oxidized and nonfluorescent in the reduced state. During enzymatic turnover, FAD switches between the oxidized and reduced states.

The COx molecules are embedded in an agarose gel (Figure 17.12(a)), thus freezing their translational mobility while preserving rotational freedom. The latter is very important to avoid fluctuations in the fluorescence signal due to slow reorientation. The confocal volume was centered at one COx molecule, and FAD fluorescence recorded as a function of time.

Binary jumps in the transient fluorescence intensity as shown in Figure 17.12(b) are indicative of cycling between the fluorescent and nonfluorescent state of FAD. Thus, the catalytic activity is directly accessible, where one turnover is just one cycle of high and low fluorescence intensity. The steady-state kinetics can be obtained from the distributions of the on and off states of the fluorescence corresponding to the two modifications of FAD within the reaction cycle. Thereby, both half-cycles of

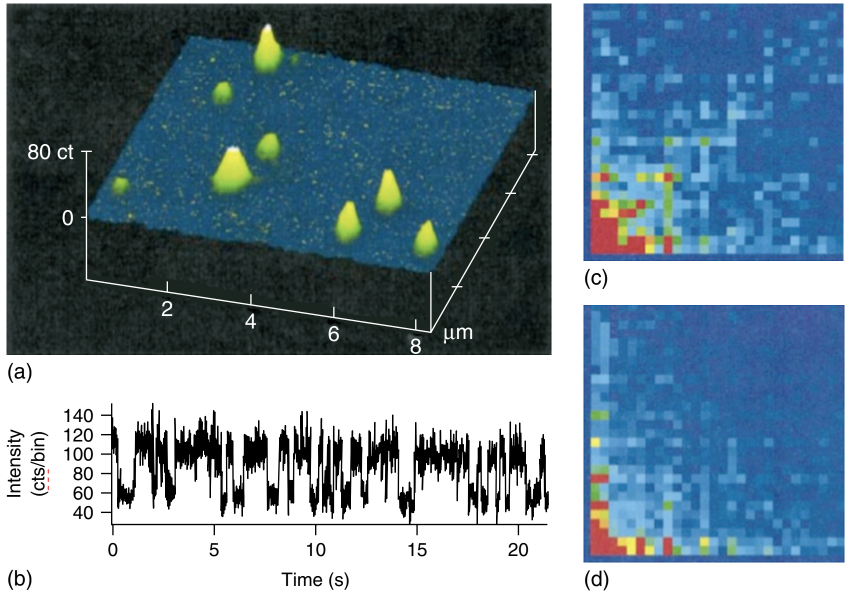

Fig. 17.12 (a) Fluorescence intensity image showing single-Cox molecules. (b) Transient fluorescence intensity of a single-Cox molecule showing binary on/off behavior. (c) and (d) Two-dimensional conditional probability distribution for a pair of two adjacent on-times (a) and for a pair of on-times separated by 10 turnovers. (Adapted from Lu, Xun and Xie (1998).)

the reaction can be analyzed within one single-molecule experiment.

Strikingly, there is a rather broad distribution of the on-times corresponding to the FAD reduction half-cycle for 33 COx molecules, indicative of static disorder, that is, different enzymatic activities of the different enzyme molecules. Now the question arises if there is also dynamic disorder, which is a fluctuation in enzymatic activity over time. To address this question, pairs of on-times separated by a defined number of turnovers are considered and their conditional probabilities calculated. If an on-time of a certain duration is with a high probability followed by an on-time of the same length after this certain number of turnovers, the conditional probability is high. If there is no correlation between the length of the pairs of on-times seperated by the number of turnovers, the conditional probability is low. In a two-dimensional conditional probability distribution histogram, a diagonal feature is evidence for a correlation of the on-time durations of on-states. Such a feature is visible for adjacent on-times (Figure 17.12(c)), but absent for on-times separated by 10 turnover cycles (Figure 17.12(d)). A covariance analysis even allows for the determination of the characteristic timescale for the fluctuations in catalytic activity. Fluctuations of the spectral mean of FAD emission in the absence of cholesterol on the same timescale are interpreted in terms of conformational changes, with the simplest model of interchange between two conformational substates.

Edman and Rigler have used a similar approach for the study of catalytic activity of the enzyme horseradish peroxidase (Edman and Rigler, 2000). Using dihydrorhodamine 6G as a substrate that is nonfluorescent, the catalytic turnover can be followed by the generation of the product rhodamine 6G. The memory landscapes of different proteins are indicative of non-Markovian behavior. Furthermore, oscillations in these memory landscapes can be interpreted by a model where the molecule passes through a number of inactive (conformational) substates and one active substate.

Recently, static heterogeneity was also reported for single-enzyme molecules using a fluorescence-based activity assay with much larger statistics (Rissin, Gorris and Walt, 2008). However, this experiment was based on the determination of product concentration and is therefore not capable of accessing the turnovers of a single-enzyme molecule.

17.4.2.2 **Conformational Changes and Catalytic Activity**

The prime example for the study of conformational dynamics of single-enzyme molecules is the rotation of F_0F_1 ATPase (Kagawa and Racker, 1966). This membrane protein synthesizes ATP from ADP and phosphate driven by a proton gradient across the membrane. The stepwise rotary motion was first directly observed by videomicroscopy of a fluorescent actin filament attached to the F_1 part (Noji *et al.*, 1997). Although ensemble methods are able to show that rotation occurs, only single-molecule experiments allow for resolution of a sequence of steps. After this key experiment, the rotation was investigated by different single-molecule fluorescence methods Rondelez *et al.* (2005); Yasuda *et al.* (2003); Yasuda *et al.* (2001); Hisabori, Kondoh and Yoshida (1999); Iino *et al.* (2005); Itoh *et al.* (2004); Masaike *et al.* (2000); Nishizaka *et al.* (2004); Adachi *et al.* (2007); Kaim *et al.* (2002); Börsch *et al.* (2002); Diez *et al.* (2004); Zimmermann *et al.* (2006); Steigmiller *et al.* (2004). Kaim *et al.* show by polarized detection the stepwise rotation of complete F_0F_1 ATPase from *Propionigenium modestum* both in synthesis and hydrolysis (Kaim *et al.*, 2002). Therefore, the c subunit was labeled by a fluorescent dye with two linkers in order to assure rigid joining (Figure 17.13(a)). From the transient fluorescence detected in two channels corresponding to orthogonal polarizations the polarization P can be calculated according to

$$P = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}} \tag{17.17}$$

The time-dependent polarization in the presence of ATP and Na^+ in Figure 17.13(b) shows stepwise changes between three levels, indicating rotational motion in three steps. A data analysis by means of the autocorrelation function of the transient polarization revealed that a model with three pairs of $90\,^\circ/30\,^\circ$steps as proposed by Yasuda *et al.* (Yasuda *et al.*, 2001) was most consistent with the data.

Single-molecule FRET experiments were also able to show stepwise rotation of F_0F_1 ATPase. Incorporating doubly labeled ATPase (Figure 17.13) in liposomes freely diffusing in solution, Diez *et al.* recorded FRET efficiency transients during passage of the liposomes through the focal volume (Diez *et al.*, 2004). Those transients show stepwise changes both in hydrolysis and in synthesis between three levels (Figure 17.13(d)

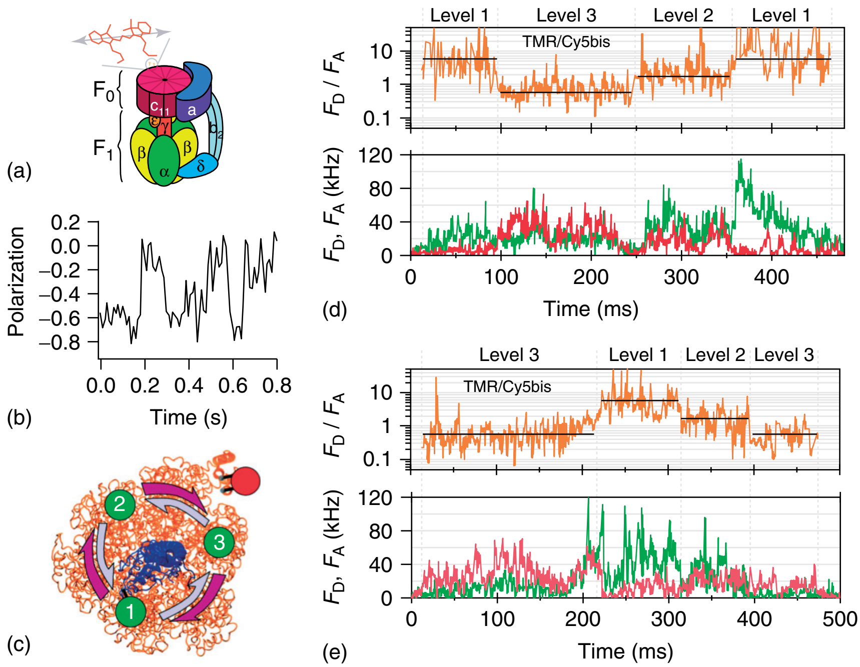

Fig. 17.13 F_0F_1 ATPase labeled at the c subunit (a) shows steps in the polarization transient (b). Donor–acceptor-labeled ATPase (c) exhibits jumps between three levels in the ratio between the fluorescence of the donor and acceptor both in hydrolysis (d) as well as in synthesis (e) with reversed order. (Adapted from Kaim *et al.* (2002) and Diez *et al.* (2004).)

and (e)). Strikingly, the order of the levels is reversed in hydrolysis/synthesis, thus elegantly demonstrating the reverse of rotation. Recently, Yasuda *et al.* were able to identify the transiently populated ATP waiting conformation by single-molecule FRET experiments on immobilized ATPase molecules (Yasuda *et al.*, 2003).

The first single-molecule FRET demonstration of conformational dynamics associated with enzymatic activity succeeded with the enzyme *staphylococcal nuclease* (SNase) in the group of Shimon Weiss (Ha *et al.*, 1999). SNase labeled with a donor–acceptor pair showed fluctuations of the transfer efficiency that can be quantitatively analyzed in terms of the autocorrelation function of the transfer efficiency shown in Figure 17.14. The decay in the E_t autocorrelation function can be fit by an exponential with time constants ranging from some tens to thousands of milliseconds, with an average of 41 milliseconds. In the presence of pTp, a known inhibitor of SNase activity, this mean is shifted to 133 milliseconds. Therefore, it can be concluded that the inhibitor affects the conformational flexibility that is essential for SNase function. By labeling of the SNase with donor and the DNA substrate with the acceptor, SNase activity can be directly monitored via the association time, which is the time period of high FRET efficiency. A clear difference in association times is observed if the DNA is labeled either at the 5′ or 3′ end. This difference is

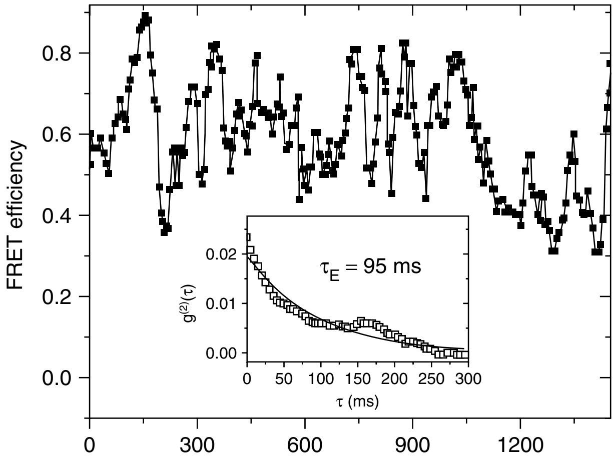

Fig. 17.14 FRET efficiency fluctuations of a single SNase molecule immobilized on a coverslip. The inset shows the autocorrelation function of the the FRET efficiency together with an exponential fit. (Adapted from Ha *et al.* (1999).)

in agreement with SNase processivity in the 3′–5′ direction.

Quenching of the dye AlexaFluor 488 was exploited for the investigation of conformational transitions of the enzyme dihydrofolate reductase (DHFR) (Zhang *et al.*, 2004). DHFR is a two-substrate enzyme catalyzing the reduction in dihydrofolate (DHF) to tetrahydrofolate (THF) by NADPH. To this end, AlexaFluor 488 was specifically attached to the structural loop that closes over the substrates after binding. Single- DHFR molecules were tethered on biotynilated cover slips via a biotin–streptavidin–biotin sandwich. In a TIRF illumination wide-field set-up, the transient intensities of many DHFR molecules could be recorded in parallel by an intensified CCD camera. Upon addition of DHF, conformational dynamics manifests itself in intensity fluctuations of the Alexa dye. In the simplest model, DHFR is switching between two conformational states with an opening and closing rate. Thus, using a threshold criterion, open and closed states can be assigned, delivering the dwell times in the respective states. The histograms of the dwell times provide the rates of opening and closing. From the concentration dependence of the closing rate in conjunction with transient ensemble kinetics measurements a model can be derived where rapid binding of DHF is followed by a conformational change.

Henzler-Wildman *et al.* have investigated the conformational heterogeneity and dynamics of Adk from *aquifex aeolicus* by X-ray crystallography, molecular dynamics simulations, NMR spectroscopy, and smFRET (Henzler-Wildman *et al.*, 2007). Adk is a two-substrate enzyme catalyzing phosphotransfer between AMP and ATP in the presence of Mg^{2+}. From X-ray crystallography in the presence of the substrates it is known that a conformational change occurs when the substrates are bound. Adk was specifically labeled with AlexaFluor 488 as the donor and AlexaFluor

633 as the acceptor at the two substrate-binding lids. First, freely diffusing Adk was investigated, and burstwise histograms of the transfer efficiencies were derived. Surprisingly, even in the absence of any substrate, Adk shows at least two subpopulations, where one corresponds to the conformation where both lids are opened and the other one corresponds to the closed conformation, which can be concluded by comparing the distances from the FRET experiments with crystallographic data. Upon addition of the inhibitor diadenosin pentaphosphate (AP5A), the equilibrium between the two conformations shifts toward the closed conformation.

To study the dynamics of conformational exchange, labeled Adk was immobilized on a functionalized cover slip via a His_6 tag. In spite of the considerable noise level, which results from the short integration time dictated by the required time resolution, transitions between the open and closed states can be observed. The overlap of the transfer efficiency histograms of two conformations renders a single threshold criterion for the discrimination of the two states unusable. Instead, a threshold zone was defined, where a change-of-state required a complete crossing of the zone. After assignment of the open and closed states, respectively, the dwell times in these states are compiled in histograms. From the histograms shown in Figure 17.15(a) and (b) the rate constants for opening and closing are derived. The values found for both cases, $k_{\text{open}} = 6500 \pm 500\,\text{s}^{-1}$ and $k_{\text{close}} = 2000 \pm 200\,\text{s}^{-1}$ for the substrate-free and $k_{\text{open}} = 40 \pm 20\,\text{s}^{-1}$ and $k_{\text{close}} = 350 \pm 200\,\text{s}^{-1}$ for the inhibitor-loaded enzyme, turn out to be very similar to the rates of chemical exchange found by NMR.

The assignment of open/closed states based solely on the apparent FRET efficiency is error-prone due to photophysical processes, which may also lead to fluctuations in the apparent FRET efficiency. It needs to be checked explicitly whether anticorrelated intensity changes of the donor and acceptor fluorescence signals occur. Therefore, a correlation analysis was performed. Clear evidence for anticorrelated behavior of donor–acceptor fluorescence intensity is provided by an anticorrelation in the donor–acceptor cross-correlation function as can be seen in Figure 17.15(c). At shorter times, a correlated fluctuation of donor and acceptor fluorescence is found, which is attributed to triplet dynamics of the donor and/or triplet dynamics of the acceptor and efficient STA. A global fit of the three correlation functions, namely, the donor and acceptor autocorrelation and their respective cross-correlation functions, provides the overall rate of distance dynamics and triplet kinetics. The overall rate of distance dynamics of $k_{\text{tot}} = 7000 \pm 2000\,\text{s}^{-1}$ is in good agreement with the rate from NMR and from the states analysis. During catalysis, opening and closing rates of $k_{\text{open}} = 40 \pm 10\,\text{s}^{-1}$ and $k_{\text{close}} = 1000 \pm 500\,\text{s}^{-1}$ are found matching the NMR data on chemical exchange. Taking into account that NMR can only provide information about the occurrence and timescale of dynamics, but no spatial information, these experiments nicely demonstrate that single-molecule FRET experiments perfectly complement other biophysical methods.

Single-molecule FRET experiments have also been performed on *E. coli* AdK, where opening and closing rates of $k_{\text{open}} = 120 \pm 40\,\text{s}^{-1}$ and $k_{\text{close}} = 220 \pm$

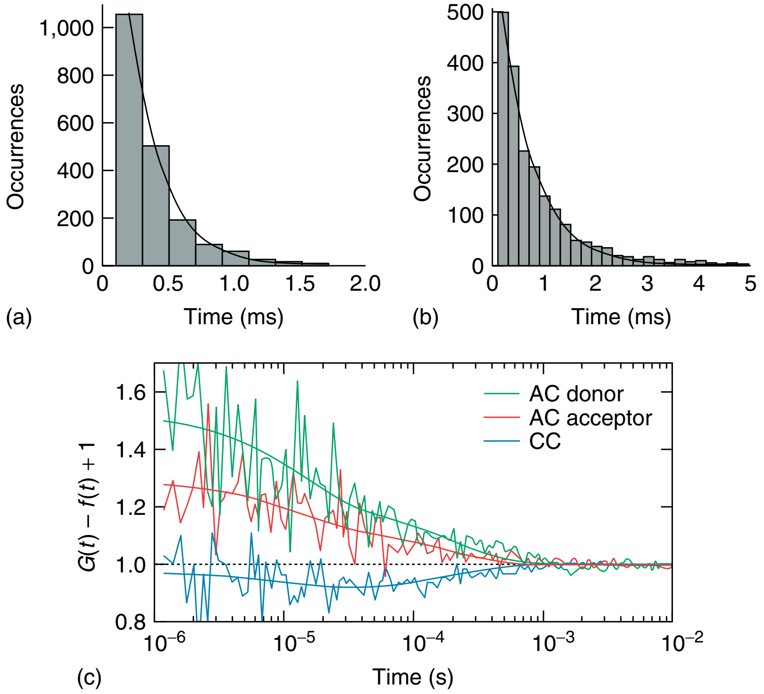

Fig. 17.15 Histograms of (a) low FRET efficiency and (b) high FRET efficiency states in the transient fluorescence of doubly labeled adenylate kinase from aquifex aeolicus attributed to an open and closed conformation. (c) Donor (green) and acceptor (red) autocorrelation and donor–acceptor cross-correlation function (blue) of the same transients. (Adapted from Henzler-Wildman *et al.* (2007).)

70 s^{-1} were found for substrate-free Adk (Hanson *et al.*, 2007).

17.5 Conclusions

In recent years, single-molecule fluorescence has proven to be a valuable tool to obtain information on the conformational dynamics and heterogeneity of proteins that is not accessible by ensemble methods. However, the method is still far from being routinely applicable, mostly for problems with photophysics and photochemistry of the dye molecules. There have been several attempts made to improve the photostability of the dyes, but what works out for a particular dye may be even counterproductive for another. More effort has to be made on both synthesizing new dyes with better stability (Weil *et al.*, 2005) and improving the photostability of existing dyes (Rasnik, McKinney and Ha, 2006).

On the other hand, the noisy signal of a single molecule remains a challenge for data analysis, which might be confronted by methods such as Hidden–Markov modeling (McKinney, Joo and Ha, 2006).

References

Adachi, K., Oiwa, K., Nishizaka, T., Furuike, S., Noji, H., Itoh, H., Yoshida, M. and Kinosita, K. Jr (2007) Coupling of rotation and catalysis in F-1-ATPase revealed by single-molecule imaging and manipulation. *Cell*, **130**, (2), 309–321. DOI: 10.1016/j.cell.2007.05.020.

Berberan-Santos, M. and Prieto, M. (1988) Monte carlo simulation of orientational effects on direct energy transfer. *J. Chem. Phys.*, **88**, (10), 6341–6349.

Börsch, M., Diez, M., Zimmermann, B., Reuter, R. and Gräber, P. (2002) Stepwise rotation of the gamma-subunit of EF0F1-ATP synthase observed by intramolecular single-molecule fluorescence resonance energy transfer. *FEBS Lett.*, **527**, (1–3), 147 –152.

Deniz, A., Laurence, T., Beligere, G., Dahan, M., Martin, A., Chemla, D., Dawson, P., Schultz, P. and Weiss, S. (2000) Single-molecule protein folding: Diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *PNAS*, **97**, (10), 5179–5184.

Diez, M., Zimmermann, B., Börsch, M., König, M., Schweinberger, E., Steigmiller, S., Reuter, R., Felekyan, S., Kudryavtsev, V., Seidel, C.A.M. *etal* (2004) Proton-powered subunit rotation in single membrane-bound F0F1-ATP synthase. *Nat. Struct. Mol. Biol.*, **11**, (2), 135–141.

Edman, L. and Rigler, R. (2000) Memory landscapes of single-enzyme molecules. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, (15), 8266–8271.

Förster, T. (1946) Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Phys. (Berlin)*, **6**, 166–175.

Ha, T., Enderle, T., Ogletree, D., Chemla, D., Selvin, P. and Weiss, S. (1996) Probing the interaction between two single molecules: Fluorescence resonance energy transfer between a sinlge donor and a single acceptor. *Proc. Natl. Acad. Sci.*, **93**, 6264–6268.

Ha, T.J., Ting, A.Y., Liang, J., Caldwell, W.B., Deniz, A.A., Chemla, D.S., Schultz, P.G. and Weiss, S. (1999) Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, (3), 893–898.

Hanson, J.A., Duderstadt, K., Watkins, L.P., Bhattacharyya, S., Brokaw, J., Chu, J.W. and Yang, H. (2007) Illuminating the mechanistic roles of enzyme conformational dynamics. *Proc. Natl. Acad. Sci.*, **104**, (46), 18–055–18–060.

Henzler-Wildman, K.A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M.A., Petsko, G.A., Karplus, M. *etal* (2007) Intrinsic motions along an enzymatic reaction trajectory. *Nature*, **450**, (7171), 838–844. DOI: 10.1038/nature06410.

Hirschfeld, T. (1976) Optical microscopic observation of single small molecules. *Appl. Opt*, **15**, (12), 2965–2966.

Hisabori, T., Kondoh, A. and Yoshida, M. (1999) The gamma subunit in chloroplast F-1-ATPase can rotate in a unidirectional and counter-clockwise manner. *FEBS Lett.*, **463**, (1–2), 35–38.

Hofmann, H., Golbik, R.P., Ott, M., Hübner, C.G. and Ulbrich-Hofmann, R. (2008) Coulomb forces control the density of the collapsed unfolded state of barstar. *J. Mol. Biol.*, **376**, (2), 597–605. DOI: 10.1016/j.jmb.2007.11.083.

Hübner, C.G., Ksenofontov, V., Nolde, F., Müllen, K. and Basché, T. (2004) Three-dimensional orientational colocalization of individual donor-acceptor pairs. *J. Chem. Phys.*, **120**, (23), 10867–10870.

Iino, R., Rondelez, Y., Yoshida, M. and Noji, H. (2005) Chemomechanical coupling in single-molecule F-type ATP synthase. *J. Bioenerg. Biomembr.*, **37**, (6), 451–454. DOI: 10.1007/s10863-005-9489-5.

Itoh, H., Takahashi, A., Adachi, K., Noji, H., Yasuda, R., Yoshida, M. and Kinosita, K. (2004) Mechanically driven ATP synthesis by F-1-ATPase. *Nature*, **427**, (6973), 465–468. DOI: 10.1038/nature02212.

Kagawa, Y. and Racker, E. (1966) Partial resolution of enzymes catalyzing oxidative phosphorylation. 9. reconstruction of oligomycin-sensitive adenosine triphosphatase. *J. Biol. Chem.*, **241**, (10), 2467–2474.

Kaim, G., Prummer, M., Sick, B., Zumofen, G., Renn, A., Wild, U.P. and Dimroth, P. (2002) Coupled rotation within single F0F1 enzyme complexes during ATP synthesis or hydrolysis. *FEBS Lett.*, **525**, (1–3), 156–163.

Kelley, A.M., Michalet, X. and Weiss, S. (2001) Chemical physics - single-molecule spectroscopy comes of age. *Science*, **292**, (5522), 1671–1672.

Lu, H.P., Xun, L.Y. and Xie, X.S. (1998) Single-molecule enzymatic dynamics. *Science*, **282**, (5395), 1877–1882.

Masaike, T., Mitome, N., Noji, H., Muneyuki, E., Yasuda, R., Kinosita, K. and Yoshida, M. (2000) Rotation of F-1-ATPase and the hinge residues of the beta subunit. *J. Exp. Biol.*, **203**, (1), 1–8.

Maus, M., De, R., Lor, M., Weil, T., Mitra, S., Wiesler, U.M., Herrmann, A., Hofkens, J., Vosch, T. and Müllen, K. *etal* (2001) Intramolecular energy hopping and energy trapping in polyphenylene dendrimers with multiple peryleneimide donor chromophores and a terryleneimide acceptor trap chromophore. *J. Am. Chem. Soc.*, **123**, (31), 7668–7676.

McKinney, S.A., Joo, C. and Ha, T. (2006) Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys. J.*, **91**, (5), 1941–1951. DOI: 10.1529/biophysj.106.082487.

Minsky M. (1957) Microscopy apparatus. US patent 3013467.

Nettels, D., Gopich, IV, Hoffmann, A. and Schuler, B. (2007) Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc. Natl. Acad. Sci.* 2007; **104**, (8): 2655–2660.

Nishizaka, T., Oiwa, K., Noji, H., Kimura, S., Muneyuki, E., Yoshida, M. and Kinosita, K. (2004) Chemomechanical coupling in F-1-ATPase revealed by simultaneous observation of nucleotide kinetics and rotation. *Nat. Struct. Mol. Biol.*, **11**, (2), 142–148. DOI: 10.1038/nsmb721.

Noji, H., Yasuda, R., Yoshida, M. and Kinosita, J.K. (1997) Direct observation of the rotation of F/sub 1/-ATPase. *Nature*, **386**, (6622), 299–302.

Orrit, M. and Bernard, J. (1990) Single pentacene molecules detected by fluorescence excitation in a p-terphenyl crystal. *Phys. Rev. Lett.*, **65**, (21), 2716–2719.

Rasnik, I., McKinney, S. and Ha, T. (2006) Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat. Methods*, **3**, (11), 891–893.

Rhoades, E., Gussakovsky, E. and Haran, G. (2003) Watching proteins fold one molecule at a time. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 3197–3202.

Rissin, D.M., Gorris, H.H. and Walt, D.R. (2008) Distinct and long-lived activity states of single enzyme molecules. *J. Am. Chem. Soc.*, **130**, (15), 5349–5353.

Rondelez, Y., Tresset, G., Nakashima, T., Kato-Yamada, Y., Fujita, H., Takeuchi, S. and Noji, H. (2005) Highly coupled ATP synthesis by F-1-ATPase single molecules. *Nature*, **433**, (7027), 773–777. DOI: 10.1038/nature03277.

Schmidt, T., Schütz, G.J., Baumgartner, W., Gruber, H.J. and Schindler, H. (1995) Characterization of photophysics and mobility of single molecules in a fluid lipid membrane. *J. Phys. Chem.*, **99**, (49), 17662–17668.

Schuler, B., Lipman, E.A. and Eaton, W.A. (2002) Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, **419**, 743–747.

Schütz, G.J., Kada, G., Pastushenko, V.P. and Schindler, H. (2000) Properties of lipid microdomains in a muscle cell membrane visualized by single molecule microscopy. *EMBO J.*, **19**, (5), 892–901.

Shera, E.B., Seitzinger, N.K., Davis, L.M., Keller, R.A. and Soper, S.A. (1990) Detection of single fluorescent molecules. *Chem. Phys. Lett.*, **174**, 553–557.

Steigmiller, S., Zimmermann, B., Diez, M., Börsch, M. and Gröber, P. (2004) Binding of single nucleotides to H+-ATP synthases observed by fluorescence resonance energy transfer. *Bioelectrochemistry*, **63**, (1–2), 79–85. DOI: 10.1016/j.bioelechem.2003.08.008.

Strickler, S.J. and Berg, R.A. (1962) Relationship between absorption intensity and fluorescence lifetime of molecules. *J. Chem. Phys.*, **37**, (4), 814–822.

Stryer, L. and Haugland, R. (1967) Energy transfer: a spectroscopic ruler. *PNAS*, **58**, 719–726.

Weil, T., Abdalla, M.A., Jatzke, C., Hengstler, J. and Müllen, K. (2005) Water-soluble rylene dyes as high-performance colorants for the staining of cells. *Biomacromolecules*, **6**, (1), 68–79.

Yasuda, R., Masaike, T., Adachi, K., Noji, H., Itoh, H. and Kinosita, K. (2003) The ATP-waiting conformation of rotating F-1-ATPase revealed by single-pair fluorescence resonance energy transfer. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, (16), 9314–9318.

Yasuda, R., Noji, H., Yoshida, M., Kinosita, K. and Itoh, H. (2001) Resolution of distinct rotational substeps by submillisecond kinetic analysis of F-1-ATPase. *Nature*, **410**, (6831), 898–904.

Zhang, Z.Q., Rajagopalan, P.T.R., Selzer, T., Benkovic, S.J. and Hammes, G.G. (2004) Single-molecule and transient kinetics investigation of the interaction of dihydrofolate reductase with NADPH and dihydrofolate. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, (9), 2764–2769.

Zimmermann, B., Diez, M., Börsch, M. and Gröber, P. (2006) Subunit movements in membrane-integrated EF0F1 during ATP synthesis detected by single-molecule spectroscopy. *Biochim. Biophys. Acta Bioenerg.*, **1757**, (5–6), 311–319. DOI: 10.1016/j.bbabio.2006.03.020.

21
Condensed Phase Ultrafast Dynamics

Stephen R. Meech and Ismael A. Heisler

21.1 Introduction 749
21.2 Ultrafast Lasers 749
21.3 Ultrashort Pulse Width Measurements 750
21.3.1 Autocorrelation Measurements 751
21.3.2 Autocorrelation 752
21.4 Ultrafast Experiments 753
21.4.1 Transient Absorption 753
21.4.2 Ultrafast Fluorescence Up-Conversion 754
21.4.3 Polarization Spectroscopy 756
21.4.4 Nonlinear Optics and Ultrafast Spectroscopy 759
21.4.5 Photon Echoes 763
21.5 The Future 765
References 766

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

21.1
Introduction

Ultrafast laser spectroscopy has evolved over the past 30 years from a somewhat esoteric area on the borders of chemistry and physics to become one of the major tools in molecular, materials, and life sciences. Much of this evolution has been driven by advances in technology, particularly, the development of the Kerr lens mode-locked titanium sapphire laser and subsequent progress in the design of optical parametric amplifiers (OPA). In this chapter, we briefly focus on these developments and outline the common methods of laser pulse characterization. We then describe some of the most widespread methods of modern ultrafast spectroscopy and indicate some of their capabilities.

21.2
Ultrafast Lasers

Short pulses are generated from a laser by modulating the gain at the round trip frequency of the laser. There are many methods of achieving such a modulation, including intracavity acousto-optic modulators and saturable absorbers (Svelto, 1998). The result is a series of pulses at a repetition frequency, which is the inverse of the round trip time and a temporal width, which is ultimately limited by the gain bandwidth of the lasing medium. The earliest studies of molecular dynamics were made with flash lamp pumped mode-locked Nd : YAG and Nd glass lasers (Kaufmann *et al.*, 1975). These sources were, by modern standards, unstable, difficult to use, and dangerous, but nevertheless readily produced picosecond pulses of very high intensities at low repetition rates. However, the first truly ultrafast experiments were made with mode-locked dye lasers. Subpicosecond pulses could be generated from dye lasers because the gain bandwidth of the laser medium is much larger than that for Nd lasers. Although very short pulses were obtained (<50 fs), the technical difficulties associated with these devices were considerable and their stability poor, making ultrafast spectroscopic measurements challenging (Kaufmann *et al.*, 1975; Fork *et al.*, 1981).

Important progress in ultrafast pulse generation was made with the discovery of solid state materials with very broad gain spectra, most notably titanium sapphire, which has a bandwidth capable of supporting pulse widths

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

<10 fs. Titanium sapphire lasers were first mode-locked by traditional intracavity gain modulation methods, but the remarkable discovery that, once initiated, the mode-locking would continue without active modulation was the trigger for the development of a new generation of all solid state ultrafast lasers (Spence *et al.*, 1991). It was quickly discovered that the mechanism for self mode-locking was a nonlinear interaction in the gain medium, which becomes more effective with shorter pulses. With careful correction for group velocity dispersion by intracavity prisms, pulses as short as 15 fs can be extracted directly from the Ti : sapphire laser.

Ti : sapphire also has suitable energy storage properties, which allow it to act as a gain medium for pulse amplification. Amplifying such short pulses to high energies creates enormous peak powers (peak power = energy per pulse per pulse width), which would ultimately destroy the amplifier cavity. To overcome this problem, the method of chirped pulse amplification was developed (Perry and Mourou, 1994). The broad spectrum of the input pulse is spread in time using gratings and prisms such that the pulse width reaches several hundred picoseconds. The broad pulse can then be safely amplified and recompressed to yield pulses of up to millijoule scale energy and tens of femtosecond's width. With careful focusing, terawatt powers are achievable.

Short high peak power amplified pulses from solid state sources also have excellent beam profiles. Such pulses are suitable for pumping OPAs, which, when seeded with broad bandwidth pulses generated through self-phase modulation, can convert approximately 800 nm output of the Ti : sapphire laser into ultrafast pulses at any wavelength between the visible and near IR (Cerullo and De Silvestri, 2003; Wilhelm *et al.*, 1997). Such pulses may thus be used to study the electronic spectroscopy of almost any molecule. The OPA outputs two frequencies, called the *signal* and the *idler*, which may themselves be mixed in another nonlinear crystal to generate the difference frequency in the mid IR region (Towrie *et al.*, 2003). In this way, ultrafast spectroscopy can be performed in the vibrational energy region. The availability of short stable pulses at any frequency has stimulated a range of novel and imaginative range of ultrafast experiments, some of which are mentioned below.

21.3
Ultrashort Pulse Width Measurements

Pulses with tens of optical cycles in the near infrared spectral region are now commonly encountered in research laboratories worldwide. These pulses, which are on the femtosecond time scale, are much faster than any electronic measuring equipment can possibly capture. New methods to time resolve events on the ultrashort time scale have to be developed. In reality, the knowledge related to the generation of ultrashort pulses and the methods for their characterization evolved simultaneously, because it is essential to know the distortions and constraints that limit the production of even shorter pulses. In addition, for experiments utilizing ultrashort pulses, it is necessary at least to know the time duration of such pulses to determine the time resolution of the experiment.

Presently, there are very sophisticated methods that cover a wide range of time scales and produce a complete characterization of ultrashort pulses, that is, not only information about temporal duration but also about phase and amplitude in time and frequency domains. These methods, of which frequency resolved optical gating (FROG) (Trebino *et al.*, 1997) is the best known, are not treated here. Rather we describe the most fundamental technique, which estimates the time duration of ultrashort pulses with good precision. This technique relies on the condition that the ultrashort pulse itself is used to determine its time duration; an autocorrelation process (Jean-Claude and Rudolph, 1996). The setup most widely used is based on the well-known Michelson interferometer. However, linear processes such as a simple interferogram can only produce information about the coherence of the pulse, which is related to the spectral content. To access information about the time duration, it is necessary to work with nonlinear processes, such as second harmonic generation (SHG) in a nonlinear crystal, two-photon absorption, optical Kerr gate, and so on (Boyd, 2003), which enable the measurement of the intensity autocorrelation (Rulliere, 1998).

21.3.1
Autocorrelation Measurements

A typical setup for an intensity autocorrelator is shown in Figure 21.1. A beam splitter divides an incoming pulse into two equal intensity pulses which are then focused to a common spot where a crystal with, for example, a second-order nonlinearity is positioned. The path length difference and consequently the relative delay time (τ) between the two pulses can be adjusted mechanically via a variable delay stage. If the arm length difference is made small, so that the pulses overlap temporally inside the nonlinear crystal, the process of second harmonic (or more generally sum frequency) generation occurs in the phase matched direction separated spatially from the incoming beams and their second harmonics. The intensity of the newly generated beam can be measured

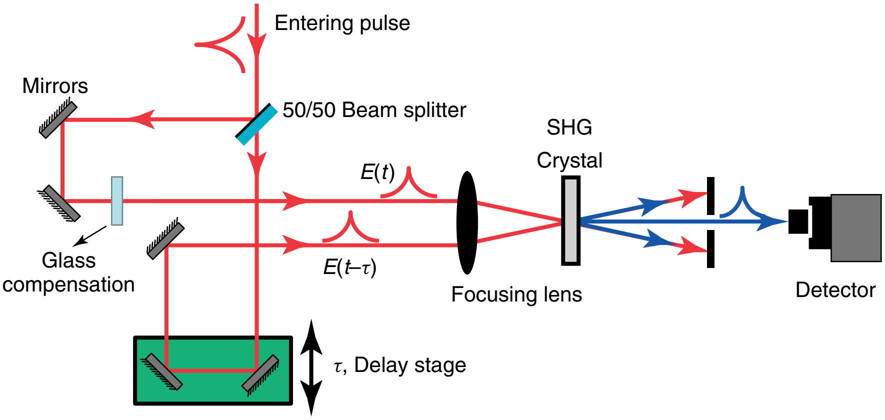

Fig. 21.1 Scheme for an intensity autocorrelator. SHG, second harmonic generation. Solid black lines are the laser fundamental, the dashed line is the second harmonic generated when the pulses overlap.

with a time integrating detector at each delay step introduced by the delay stage. If the relative time delay is increased, so that the overlap of the two pulses in the crystal is reduced, the second harmonic signal becomes weaker and ultimately disappears. The overall result is the intensity of the second harmonic signal as a function of delay time, which carries information about the incoming pulse duration.

21.3.2
Autocorrelation

The mathematical description of an autocorrelation process can be written as follows:

$$G^{(n)}(\tau) = \int_{-\infty}^{\infty} f^{n}(t) f^{n^*}(t-\tau)dt \quad (21.1)$$

where $f(t)$ is some time-dependent function and $f^*(t)$ is its complex conjugate (Jean-Claude and Rudolph, 1996). The function $G^{(n)}(\tau)$ is called the nth-order autocorrelation. Taking $f(t)$ as the resulting electric field from the two pulses inside the crystal, $f(t) = E(t) + E(t-\tau)$, and assuming a second-order process, it is possible to write the resulting signal as

$$G^{(2)}(\tau) = \int_{-\infty}^{\infty} \left|(E(t) + E(t-\tau))^2\right|^2 dt \quad (21.2)$$

where $G^{(2)}(\tau)$ is the second-order autocorrelation as a function of the delay time, τ. This equation can be separated into three terms: one centered at zero frequency, which is known as the *intensity autocorrelation plus a constant background*; the other two terms are centered at the optical frequency, ω, and second harmonic, 2ω, and are known as the *interferometric autocorrelation* (Jean-Claude and Rudolph, 1996). An example of an interferometric second-order autocorrelation measurement is presented in Figure 21.2(b). Most often the intensity autocorrelation, which is easier to measure than an interferometric autocorrelation, produces sufficient information to determine the time resolution of a given experiment. The intensity autocorrelation expression can be written as

$$I^{(2)}(\tau) = \int_{-\infty}^{\infty} I(t)I(t-\tau)dt \quad (21.3)$$

where $I(t) \propto |E(t)|^2$ is the time-dependent intensity of the pulse. An example

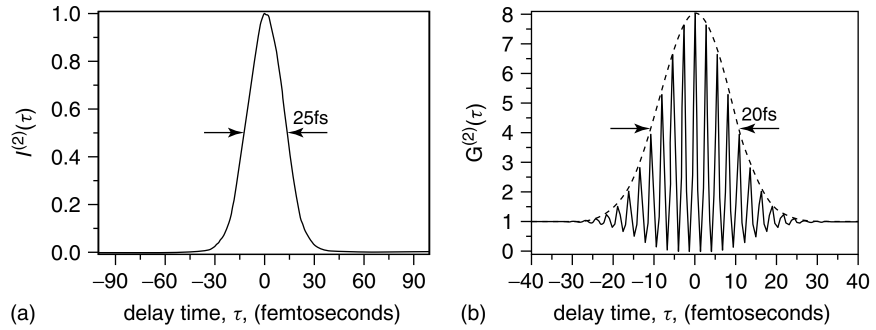

Fig. 21.2 (a) Intensity autocorrelation, without background, showing a FWHM (full width at half maximum) of 25 fs. This corresponds to a pulse with time duration of 16.3 fs, assuming a hyperbolic secant squared pulse profile. (b) Interferometric autocorrelation showing the optical field oscillations. This corresponds to a pulse of 14.8 fs.

of such a measurement can be seen in Figure 21.2(a). To recover the pulse duration, it is necessary to fit the measured data by assuming some functional form for the intensity. Usually ultrashort pulses originating from mode-locked lasers have a temporal shape described by a squared hyperbolic secant function (Jean-Claude and Rudolph, 1996). For a given spectral bandwidth, the attainable theoretical shortest temporal pulse is called *Fourier transform limited* (or simply *transform limited*). For such pulses, the spectral phase difference between the wavelengths that compose the spectrum is constant or equal to zero. This can be tested using the expression: $\Delta\nu\Delta\tau_P \geq c_B$, where $\Delta\nu$ is the spectral bandwidth of the pulse (in Hertz), $\Delta\tau_P$ is the temporal duration, and c_B is a constant that is determined by the model assumed to fit the temporal intensity autocorrelation. For Gaussian and squared hyperbolic secant functions, c_B is equal to 0.44 and 0.31, respectively. The equality in the above expression is achieved for transform limited pulses (Rulliere, 1998).

For the shortest pulses, it is found that transmission of the pulse through any dense medium leads to pulse broadening. This naturally arises from the normal dispersion of the medium, which causes the red portion of the broadband laser spectrum to travel faster than the blue, creating a spread in time (a chirp) (Jean-Claude and Rudolph, 1996). Chirped pulse amplifiers exploit this phenomenon. For high time resolution optical experiments, a correction for this effect must be made. To first order, this is achieved using a simple prism pair. Routing the chirped beam twice through the prism pair, oriented at Brewster's angle, causes the red light to travel further than the blue (negative dispersion). By carefully matching the positive and negative dispersion by controlling the prism separation, the experimenter can deliver the shortest pulses to the sample.

21.4
Ultrafast Experiments

21.4.1
Transient Absorption

This experiment is essentially an extension of the flash photolysis experiment developed by Porter and coworkers (Porter and Wilkinson, 1961) into the ultrafast time domain. It shares with that method the principle of optical excitation of the sample by one pulse and then measurement of the entire spectrum as a function of time with a second broadband pulse. In accessing the subpicosecond time domain, one encounters the same problem as was mentioned in Section 21.3, that no conventional electronic detector is capable of real-time measurements on the subpicosecond timescale. Thus the requirement for the transient absorption experiment is an ultrafast pump and an ultrafast probe pulse, and the latter should have as broad a spectrum as possible. A broad spectrum with an ultrafast temporal width may be generated by the simple expedient of focusing an intense titanium sapphire laser pulse into a transparent medium (a water cell works well). Through a highly nonlinear interaction called self-phase modulation, a fraction of this incident light is converted into a broad band source stretching from the UV to the near IR with a temporal width as short as the input pulse (Shen, 1984).

The pump and probe pulses are then combined in essentially the same geometry as was shown for the pulse width measurement (Figure 21.1), with the sample taking the place of the nonlinear crystal. The pump pulse strikes the sample at a time defined as $t = 0$, where it is absorbed, transferring population from the ground to the excited electronic state. In general, the excited state will have a distinct absorption spectrum. The spectrally broad probe pulse is incident on the same point in the sample at a time τ later, where τ is controlled by the path length difference introduced between the pump and probe paths. On transmission through the sample the probe beam spectrum is measured by a polychromator. For the most accurate measurements, a correction must be made for the dispersion which causes the red and blue components of the probe spectrum to travel at slightly different velocities, thus reducing the time resolution.

Typically, the change in transmission induced by the pump pulse is small – a small percentage at most. To obtain reliable measurements, it is desirable to record an identical reference spectrum taken when the pump pulse is blocked. From these two measurements, a difference spectrum may be created, and the reference spectrum can be used to normalize out intensity fluctuations. The logarithm of the resulting ratio yields the transient optical density difference spectrum, where negative peaks correspond to bleaching of the ground state and positive peaks show transient absorption spectra of newly created states, which may be formed either during the excitation or during subsequent chemical reactions. The transient absorption method has been used by many groups over the years to reveal photo-induced processes in molecules ranging from simple dissociation or ring opening reactions to the dynamics of the photosynthetic system. Changes in optical densities lesser than 10^{-4} can be measured.

21.4.2
Ultrafast Fluorescence Up-Conversion

Fluorescent molecules are widely used as probes of natural and synthetic materials for numerous reasons: fluorescence is very sensitive to its environment; the probe can be introduced to the medium in a site specific fashion; it may be detected with high (single molecule) sensitivity; it affords a noncontact method of probing the material (Lakowicz, 2006). Time resolved fluorescence has long been used to add an additional dimension to fluorescence spectroscopy (Maroncelli and Fleming, 1987). The classical time-correlated single photon counting method has been refined to the extent that excited state decay times of tens of picoseconds can be resolved (Phillips and O'Connor, 1983). However, for true ultrafast time resolution, the fluorescence up-conversion method must be used (Fleming, 1986).

The up-conversion method is essentially an ultrafast sampling of the temporal profile of the sample fluorescence. The experimental geometry is shown in Figure 21.3 (Shah, 1988).

Two ultrafast laser pulses are required, exactly correlated in time. Often the fundamental and second harmonic of the laser are used. The higher energy pulse electronically excites the sample, which relaxes through fluorescence decay at an energy ν_f. The fluorescence is focused into a nonlinear crystal where it is mixed with the second ultrafast laser pulse of frequency ν_p. The nonlinear

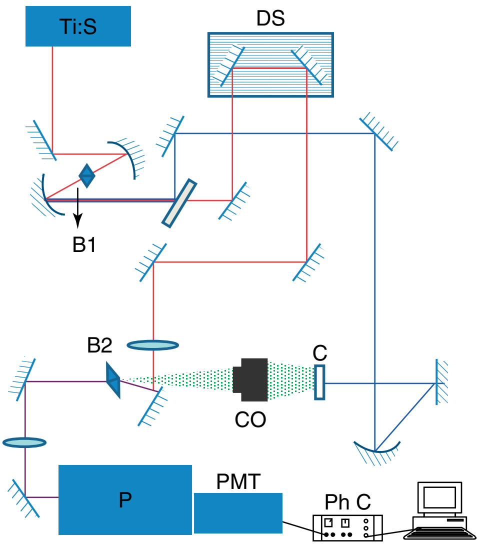

Fig. 21.3 Femtosecond fluorescence up-conversion experimental setup: Ti : S, titanium–sapphire laser; B1, 50-µm BBO crystal; B2, 100-µm BBO crystal; DS, delay stage; CO, collection optics; C, cell containing sample; P, polycromator; PMT, photomultiplier tube; Ph C, photon counter.

interaction in the crystal generates the sum frequency $\nu_{\text{sum}} = \nu_f + \nu_p$ in the phase matching direction. This signal can clearly only be generated when both fluorescence and laser pulse are present in the crystal, and its intensity is a product of their two instantaneous intensities. Thus for constant laser pulse energy, the nonlinear interaction samples the intensity of the fluorescence at the instant of the arrival of the second pulse. By scanning the arrival time of the pulse and measuring the sum frequency intensity as a function of time, the fluorescence decay profile is mapped out. As an example, Figure 21.4 presents a semilog plot of the up-converted fluorescence intensity relaxation of an aqueous solution of the dye molecule auramine O (van der Meer *et al.*, 2000). The signal can be described by a sum of two exponential decays.

As in all ultrafast experiments, the highest time resolution is only achieved if the dispersion is minimized. For up-conversion, this requires independent dispersion compensation in the excitation and up-converting pulses, and as far as possible all reflective optics should be used (Joo and Rhee, 2005).

Fluorescence up-conversion is a particularly powerful tool as it allows detailed studies of a diverse range of phenomena. Excited state reactions (isomerization, electron and proton transfer) (Kang *et al.*, 1990; Todd *et al.*, 1990) are among the simplest known and their dynamics have been characterized in great detail through fluorescence. Similarly, the dynamics of solvation can be measured from the temporal evolution of the fluorescence spectrum (Maroncelli and Fleming, 1987). Such studies have provided very detailed views of condensed phase dynamics, and have been used as benchmarks to test molecular dynamics simulations. Having been well characterized in solution, similar experiments are being conducted in complex media including living cells.

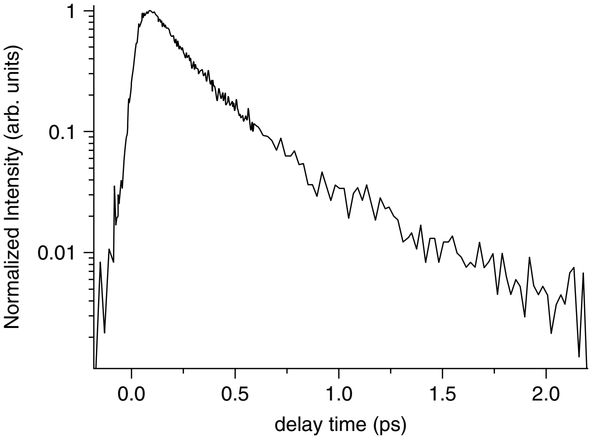

Fig. 21.4 Up-converted signal intensity showing the fluorescence relaxation of dye auramine O in water solution.

21.4.3
Polarization Spectroscopy

A number of techniques that relate spectroscopic information to the polarization properties of electromagnetic radiation are known as *polarization spectroscopy* (Kano and Levenson, 1988). In general, the initial polarization state of the light source is well defined. This source interacts with an anisotropic material resulting in a new polarization state which is analyzed. The anisotropy can be either inherent to the medium (e.g., a crystalline or liquid crystal material) or externally imposed (e.g., the Kerr effect in the presence of strong electric field). Within the realm of ultrafast spectroscopy, a long established and widely applied time resolved technique is the optical Kerr effect (OKE) (Smith and Meech, 2002). In this account, we focus on the OKE technique applied to transparent media, and its analog applied to absorbing media.

One great appeal of the OKE technique is that it measures the low-frequency Raman spectrum (0–500 cm^{-1}) of the medium with extremely high signal-to-noise ratio. This spectrum contains detailed information about the medium dynamics. The OKE signal (Figure 21.5 shows the usual experimental setup) originates from an intense linearly polarized optical pulse, which interacts with an initially isotropic sample (Lotshaw *et al.*, 1995). This pulse, called the *pump*, induces a transient anisotropy through the molecular polarizability, which is probed by a second, weak, time-delayed pulse, incident at the same point in the sample. The probe pulse is linearly polarized and has its plane of polarization at 45° with respect to the pump pulse. The pump-induced anisotropy depolarizes the probe pulse, which emerges from the sample slightly elliptically polarized. A polarizer, crossed with the plane of polarization of the probe pulse, is placed in front of the detector. The pump-induced depolarization causes a fraction of the probe pulse to reach the detector (homodyne signal). The signal is measured as a function of delay time between the pump and probe pulses and so

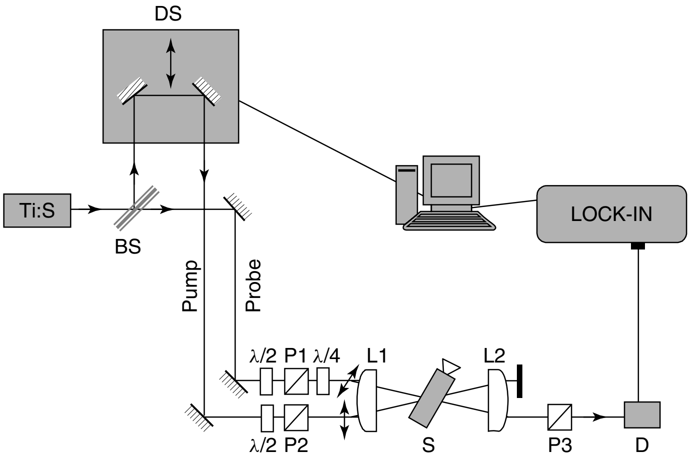

Fig. 21.5 OKE typical experimental setup. Ti : S, Ti : sapphire femtosecond laser; BS, beam splitter; DS, delay stage; P1-3, polarizers; L1-2, lenses; S, sample; and D, detector.

the relaxation of the induced anisotropy in the sample is measured. For a homodyne measurement, the detected signal is quadratically and linearly proportional to the pump and probe pulse intensities, respectively. Additionally, the measured signal is quadratically proportional to the impulse response function of the material, $R_{ijkl}(t)$ (discussed further in Section 21.4.4 below). The response function is related to the frequency-dependent third-order nonlinear susceptibility, $\chi^{(3)}_{ijkl}(\omega)$, through a Fourier transform relationship, and contains both real (dispersive) and imaginary (resonant) parts.

To improve the signal-to-noise ratio and to avoid the complexity of analyzing a quadratic signal, a heterodyned measurement is employed (Kano and Levenson, 1988). This is achieved by the addition of a local oscillator (LO) to the OKE signal. The LO is derived from the probe pulse in either of two ways: (i) by a slight rotation of the analyzer polarizer (P3), which introduces a component *in phase* with the OKE signal; (ii) by the introduction of a quarter wave plate (λ/4) in the probe beam with its fast axis initially aligned with the probe polarization; after a slight rotation of P1 (generally $\leq 1^\circ$) a small component *in quadrature* with the homodyne OKE signal will leak through the analyzer polarizer. The signal measured by the square law detector is:

$$I_D = I_{\mathrm{LO}} + I_S + \frac{nc}{8\pi}\left(E_S^* E_{\mathrm{LO}} + E_S E_{\mathrm{LO}}^*\right) \tag{21.4}$$

where E_i and I_i are the field and the intensity of the LO or homodyne OKE signal field, respectively. The term in brackets is the *heterodyne term*, which scales linearly with $R_{ijkl}(t)$. The LO field amplitude is set to produce a heterodyne signal much larger than the homodyne intensity, but small enough to avoid detector saturation. The LO signal is easily eliminated from measurement using lock-in detection. It is important to note that the real and imaginary contributions to the signal field, E_S, are associated, respectively, with the imaginary (dichroism) and real

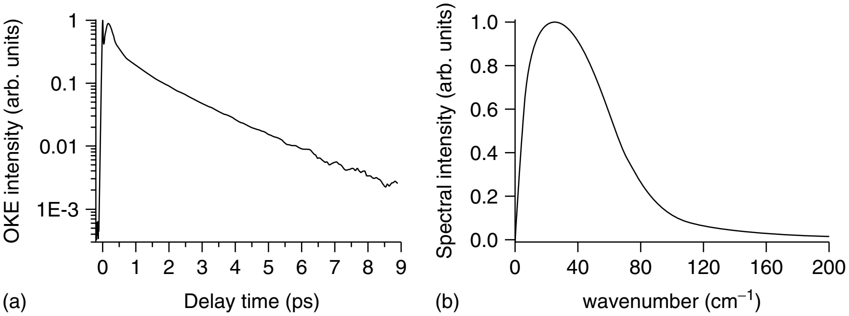

Fig. 21.6 (a) OKE signal for carbon disulfide (CS_2) on a semilog plot. (b) Spectral intensity content of CS_2 OKE signal after subtraction of the diffusive exponential component.

(birefringence) parts of the material response function, $R_{ijkl}(t)$. These may be measured separately depending on the phase introduced in the local oscillator field, E_{LO}.

When the sample is transparent, the real (birefringence) part of the response dominates. The data may be analyzed in either the time or the frequency domain. The latter are accessed through a Fourier transform deconvolution procedure resulting in the Raman spectral density (Lotshaw *et al.*, 1995). Figure 21.6 shows a typical OKE signal for carbon disulfide (CS_2) liquid at room temperature.

CS_2 has one the largest OKE signals, due to its high polarizability anisotropy, and shows many features shared with various other molecular liquids for which OKE measurements have been performed (Castner *et al.*, 2007; Cong *et al.*, 1995; Fourkas, 2002; Friedman and She, 1993; Heisler *et al.*, 2005; Hunt *et al.*, 2007; Hunt and Meech, 2004). At time zero, there is a symmetric instantaneous response which follows the laser pulse autocorrelation. This signal arises from the electronic hyperpolarizability. The nuclear dynamics response of most interest appears with a risetime (around 100 fs) due to the inertial molecular motion that cannot follow the impulsive force exerted by the strong ultrashort laser pump field. The signal then shows a complex subpicosecond relaxation followed by a picosecond time scale, exponential response (Figure 21.6a). This component is well described by a Stokes–Einstein–Debye model of diffusive orientational relaxation (McMorrow *et al.*, 1988). The subpicosecond decay component is usually called *nondiffusive* but scales with the diffusive relaxation (Loughnane *et al.*, 1999). Its origin remains a matter of discussion. On an ultrafast timescale, the single-molecule contribution will take the form of a hindered molecular reorientation or librational motion, which is well fit by a molecule rattling in a cage model. This signal relaxes approximately as a Gaussian function but usually the analysis is made in the frequency domain to avoid model-dependent fitting (Figure 21.6b). Such detailed pictures of liquid state molecular dynamics provide data against which to test both theories and molecular dynamics simulations of the liquid state. In more complex fluids, additional low-frequency intramolecular modes may contribute to the Raman spectral density, and these may be the characteristic of macromolecular structure (Hunt and Meech, 2004, 2003).

When the pump and probe pulses are resonant with a strong electronic transition, then the transient dichroism (imaginary component of $R_{ijkl}(t)$) dominates the response, which may be selectively measured by using an in-phase LO. If the absorption is due to a molecular solute, then the time-dependent signal probes both the solute orientational relaxation and the ground state recovery (because both will ultimately lead to a recovery of the initial isotropic distribution). The experiment can then be used to access information on both molecular orientational motion and excited state kinetics, and has been applied to diverse samples (Cho *et al.*, 1993; Waldeck *et al.*, 1981).

21.4.4
Nonlinear Optics and Ultrafast Spectroscopy

It is apparent that the experiments described above report on multiple interactions of the laser pulses with the sample. Thus it is evident that a natural theoretical picture of such interactions will involve a time domain description invoking nonlinear optics (Mukamel, 1995). It is in general possible to separate optical measurements into linear and nonlinear regimes according to the electric field $\mathbf{E}(t)$ power-law dependence of the process. In such measurements, to properly calculate the signal, one has to start with the Maxwell equations. In these, the only material quantity that appears as a source is given by the optical polarization $\mathbf{P}(t)$ induced by some applied external electric field $\mathbf{E}(t)$. Therefore, a complete knowledge of the optical polarization is essential for the interpretation of any time or frequency domain spectroscopic measurement. Usually a constitutive relation is used to describe the induced polarization in a medium when an electric field is applied. The polarization may be expanded as a power series in the electric field vector (Boyd, 2003):

$$\mathbf{P}(t) = \boldsymbol{\chi}^{(1)}\mathbf{E}(t) + \boldsymbol{\chi}^{(2)}\mathbf{E}(t)\,\mathbf{E}(t) + \boldsymbol{\chi}^{(3)}\mathbf{E}(t)\,\mathbf{E}(t)\,\mathbf{E}(t) + \cdots \tag{21.5}$$

where $\boldsymbol{\chi}^{(n)}$ are material susceptibility coefficients, which in general are tensors (Butcher and Cotter, 1990). As a consequence, the orientation of the induced polarization may be different from that of the applied field. Therefore, it is possible to rewrite the polarization as:

$$\mathbf{P}(t) = \mathbf{P}^{(1)}(t) + \mathbf{P}_{\mathrm{NL}}(t) \tag{21.6}$$

where $\mathbf{P}^{(1)}(t)$ is the linear polarization, accounting for processes such as absorption, reflection, refraction, and so on, and $\mathbf{P}_{\mathrm{NL}}(t)$ is the nonlinear polarization describing higher order processes such as second harmonic generation, optical Kerr effect, two-photon absorption, self-phase modulation, and so on. For time resolved experiments, a better approach to describing optical measurements is through response functions $R(t)$, which have their counterpart in the frequency domain description through the susceptibilities $\chi(\omega)$. The relation between these quantities is given by a Fourier transform:

$$\chi(\omega) = \int_{-\infty}^{\infty} dt\; R(t)e^{i\omega t} \tag{21.7}$$

The response function formalism can be applied to a broad range of linear and nonlinear optical measurements differing in terms of the applied temporal sequence of pulses, frequencies or propagation direction (wavevectors) (Mukamel, 1995). Thus, this formalism enables one to describe and calculate the signals expected for different ultrafast experiments and ultimately compare their information content.

The response function contains information on how a system alters in response to an external applied perturbation. It has been shown that if changes involve small departures from equilibrium, then the equilibrium fluctuations dictate the nonequilibrium response. Thus knowledge of the equilibrium dynamics is useful for predicting nonequilibrium processes.

The polarization at a given time t is related to the applied electric field at all earlier times and can be written, for first and second order, as

$$P_i^{(1)}(t) = \int_{-\infty}^{t} dt_1 R_{ij}^{(1)}(t, t_1) E_j(t_1) \quad (21.8)$$

$$P_i^{(2)}(t) = \int_{-\infty}^{t} dt_2 \int_{-\infty}^{t_2} dt_1 R_{ijk}^{(2)}(t, t_1, t_2) \times E_j(t_1) E_k(t_2) \quad (21.9)$$

and for the nth-order as

$$P_i^{(n)}(t) = \int_{-\infty}^{t} dt_n \ldots \int_{-\infty}^{t_3} dt_2 \int_{-\infty}^{t_2} dt_1 \times R_{ijk\ldots m}^{(n)}(t, t_1, t_2, \ldots, t_n) \times E_j(t_1) E_k(t_2) \ldots E_m(t_n) \quad (21.10)$$

with the convention of summation over repeated indices. The newly defined nth-order response function, $R^{(n)}$, is a tensor of rank n + 1, and a real and causal function of time (Mukamel, 1995). Also, the response is a function of the relative time between perturbation and observation, that is, $R^{(n)}(t, t_0) = R^{(n)}(t - t_0)$. Ordering the fields and labeling the relative time as in Figure 21.7, it is possible to rewrite the expressions for the polarization in a more intuitive fashion as

$$P_i^{(1)}(t) = \int_0^{\infty} d\tau_1 R_{ij}^{(1)}(\tau_1) E_j(t - \tau_1) \quad (21.11)$$

$$P_i^{(2)}(t) = \int_0^{\infty} d\tau_2 \int_0^{\infty} d\tau_1 R_{ijk}^{(2)}(\tau_1, \tau_2) \times E_j(t - \tau_2) E_k(t - \tau_2 - \tau_1) \quad (21.12)$$

$$P_i^{(n)}(t) = \int_0^{\infty} d\tau_n \ldots \int_0^{\infty} d\tau_2 \int_0^{\infty} d\tau_1 \times \begin{pmatrix} R_{ijk\ldots m}^{(n)}(\tau_1, \tau_2, \ldots, \tau_n) \\ E_j(t-\tau_n) E_k(t-\tau_n-\tau_{n-1}) \\ \times \ldots E_m(t-\tau_n-\tau_{n-1} \\ \times \ldots - \tau_2 - \tau_1 \end{pmatrix} \quad (21.13)$$

The optical electric field couples with the material via the microscopic charge distribution. As mentioned before, for spectroscopic measurements, the signal

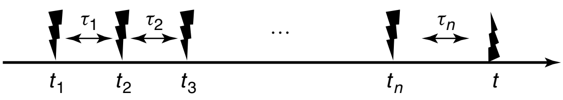

Fig. 21.7 Electric fields time ordering and resulting polarization.

is determined by the macroscopic polarization $\mathbf{P}(t)$, which in turn, is given in terms of the average over an ensemble of particles contained in a volume V, of the expectation value of the dipole operator $\boldsymbol{\mu}$:

$$\mathbf{P}(t) = \frac{1}{V}\overline{\langle\mu(t)\rangle} = \frac{1}{V}\mathrm{Tr}\{\boldsymbol{\mu}\rho(t)\} \quad (21.14)$$

In the above definition, the dipole operator is given by the instantaneous positions of all charged particles of the ensemble contained in volume V (nuclei and electrons):

$$\boldsymbol{\mu} = \sum_i e_i \mathbf{r}_i \quad (21.15)$$

The last term on the right hand side of Eq. (21.14) introduces the density matrix ($\rho(t)$) formalism to calculate the polarization, where Tr is the trace over the matrix representation. The density matrix evolution is determined by the Liouville equation:

$$\frac{d\rho}{dt} = -\frac{\mathrm{i}}{\hbar}[H_I(t), \rho] \quad (21.16)$$

Here we are assuming the interaction picture description where $H_I(t)$ accounts for the light-matter time-dependent interaction, which in the electric dipole approximation can be written as

$$H_I(t) = -\boldsymbol{\mu} \cdot \mathbf{E}(t) \quad (21.17)$$

The general solution for the evolution equation can be written as:

$$\rho(t) = \rho(t_0) - \frac{\mathrm{i}}{\hbar}\int_{t_0}^{t} dt_1 [H_I(t_1), \rho(t_1)] \quad (21.18)$$

It is not possible to solve this equation exactly. The procedure usually adopted is to expand the density matrix operator and then to solve the evolution equation iteratively.

$$\rho(t) = \rho^{(0)}(t) + \rho^{(1)}(t) + \rho^{(2)}(t) + \ldots \quad (21.19)$$

Applying this perturbative approach, solutions for the different orders are

$$\rho^{(0)}(t) = \rho(t_0) = \rho_{\mathrm{eq}} \quad (21.20)$$

$$\rho^{(1)}(t) = -\frac{\mathrm{i}}{\hbar}\int_{-\infty}^{t} dt_1 \left[H_I(t_1), \rho_{\mathrm{eq}}\right] \quad (21.21)$$

$$\rho^{(2)}(t) = \left(-\frac{\mathrm{i}}{\hbar}\right)^2 \int_{-\infty}^{t} dt_2 \int_{-\infty}^{t_2} dt_1 \times \left[H_I(t_2), \left[H_I(t_1), \rho_{\mathrm{eq}}\right]\right] \quad (21.22)$$

and for the nth-order:

$$\rho^{(n)}(t) = \left(-\frac{\mathrm{i}}{\hbar}\right) \int_{-\infty}^{t} dt_n \ldots \int_{-\infty}^{t_3} dt_2 \times \int_{-\infty}^{t_2} dt_1 [H_I(t_n), [H_I(t_{n-1}), \ldots [H_I(t_3), [H_I(t_2), [H_I(t_1), \rho_{\mathrm{eq}}]]] \ldots]] \quad (21.23)$$

The system, before any interaction, is in the equilibrium state; therefore, the assumption $t_0 = -\infty$ may safely be made. Again, to present the expressions in a more physical manner, it is helpful to perform a change of variables so that the time intervals between successive interactions of the electric field with the dipole moment appear explicitly. As an example, we present the expression for the third-order polarization, after developing the commutator terms of Eq. (21.23) and inserting them into Eq. (21.14), resulting in

$$P_i^{(3)}(t) = \frac{1}{V}\left(-\frac{\mathbf{i}}{\hbar}\right)^3 \int_0^\infty d\tau_3 \int_0^\infty d\tau_2 \times \int_0^\infty d\tau_1 \left(\begin{array}{l} \mathrm{Tr}\left\{[[[\mu_i(\tau_1+\tau_2+\tau_3), \right. \\ \quad \mu_l(\tau_1+\tau_2)], \mu_k(\tau_1)], \\ \quad \left. \mu_j(0)]\rho_{\mathrm{eq}}\right\} \\ E_j(t-\tau_3)E_k(t-\tau_3-\tau_2) \\ \quad \times E_l(t-\tau_3-\tau_2-\tau_1) \end{array}\right) \tag{21.24}$$

From this expression, the third-order response function is given by

$$R_{ijkl}^{(3)}(\tau_3, \tau_2, \tau_1) = \theta(\tau_3,\ \tau_2,\ \tau_1)\left(-\frac{\mathbf{i}}{\hbar}\right)^3 \times \mathbf{Tr}\left\{\left[\,[[\mu_i(\tau_1+\tau_2+\tau_3), \mu_l(\tau_1+\tau_2)], \mu_k\ (\tau_1)]\,, \mu_j(0)\right]\rho_{\mathrm{eq}}\right\} \tag{21.25}$$

where $\theta(\tau)$ is the Heaviside function used to impose the causality condition. This expression makes the connection between the microscopic material quantity (the dipole moment which couples to the electric field) and the response function measured through the macroscopic induced polarization. It can be further particularized to a given experiment by choosing the configuration of the applied electric fields and also the particular multipolar moment which is interacting with the field (*i.e.*, fixed dipole moment, polarizability, hyperpolarizability, etc). By working out the commutator expressions, it is possible to rewrite the response functions as

$$R_{ijkl}^{(3)}(\tau_3, \tau_2, \tau_1) = \theta(\tau_3)\theta(\tau_2)\theta(\tau_1) \times \left(-\frac{\mathbf{i}}{\hbar}\right)^3 \sum_{\alpha}^{4} R_\alpha(\tau_3, \tau_2, \tau_1) \times -R_\alpha^*(\tau_3, \tau_2, \tau_1) \tag{21.26}$$

where

$$R_1(\tau_3, \tau_2, \tau_1) = \mathrm{Tr}\left\{\mu_i(\tau_1+\tau_2+\tau_3) \times \mu_l(\tau_1+\tau_2)\mu_k(\tau_1) \times \mu_j(0)\rho_{\mathrm{eq}}\right\} \tag{21.27}$$

$$R_2(\tau_3, \tau_2, \tau_1) = \mathrm{Tr}\left\{\mu_j(0)\mu_l(\tau_1+\tau_2) \times \mu_i(\tau_1+\tau_2+\tau_3) \times \mu_k(\tau_1)\rho_{\mathrm{eq}}\right\} \tag{21.28}$$

$$R_3(\tau_3, \tau_2, \tau_1) = \mathrm{Tr}\left\{\mu_j(0)\mu_k(\tau_1) \times \mu_i(\tau_1+\tau_2+\tau_3)\mu_l \times (\tau_1+\tau_2)\rho_{\mathrm{eq}}\right\} \tag{21.29}$$

$$R_4(\tau_3, \tau_2, \tau_1) = \mathrm{Tr}\left\{\mu_k(\tau_1)\mu_l(\tau_1+\tau_2) \times \mu_i(\tau_1+\tau_2+\tau_3) \times \mu_j(0)\rho_{\mathrm{eq}}\right\} \tag{21.30}$$

The complex terms can easily be worked out in the same fashion.

Third-order nonlinearities describe many of the coherent nonlinear experiments commonly used in research laboratories worldwide: pump-probe, transient grating, photon echoes, coherent antistokes Raman spectroscopy (CARS), degenerate four-wave mixing are some of the most important of these. These experiments are described by some or all of the eight correlation functions that contribute to $R_{ijkl}^{(3)}(\tau_3, \tau_2, \tau_1)$. This shows how complete and general the response formalism can be, and how it can be applied to essentially any ultrafast experiment (Mukamel, 1995). Roughly speaking, once the configuration of electric fields used in the experiment is established, one can refer to these expressions to determine the measured signal outcome. The next step would be to include the relaxation terms involved in a given sample system. This can be done with various

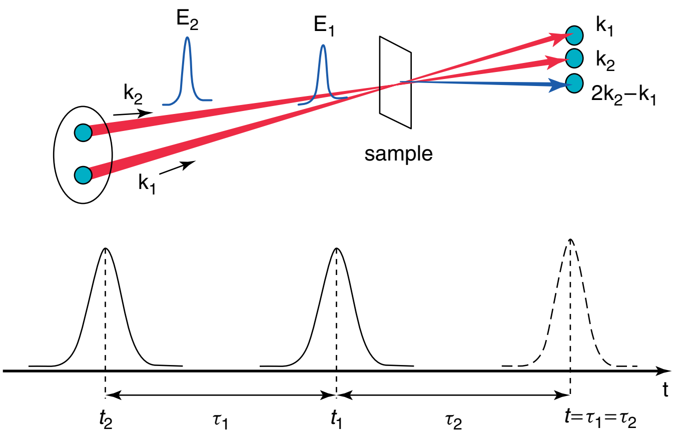

Fig. 21.8 Two-pulse photon echo time ordering and spatial beam geometry. (Adapted from Andrei Tokmakoff, 5.74 Introductory Quantum Mechanics II, Spring 2007. (MIT OpenCourseWare: Massachusetts Institute of Technology). http://ocw.mit.edu/OcwWeb/Chemistry/5-74Spring-2007/CourseHome. Accessed Oct. 1, 2008.)

degrees of complexity, and the outcome depends somewhat on the nature of the sample under study. A number of examples exist in the literature (Mukamel, 1995).

21.4.5
Photon Echoes

The photon echo technique is an interesting example of the kind of multipulse optical experiment, which is currently being developed in ultrafast spectroscopy. The original experiments were an important demonstration of coherent optical phenomena (Kurnit *et al.*, 1964). Measurements on the picosecond timescale revealed the photon echo as being an important tool in molecular spectroscopy, particularly, in allowing the direct measurement of the homogeneous optical lineshape free of inhomogeneous broadening (de Boeij *et al.*, 1998). In the original two-pulse photon echo, two pulses are incident on the sample at times t_1 and t_2 (Figure 21.8).

At a time $\tau_1 = t_1 - t_2$, the sample emits a signal – the photon echo – at the same frequency as the input pulses and in the phase matched direction, $\mathbf{k}_{\mathrm{PE}} = \mathbf{2k}_2 - \mathbf{k}_1$. The analogy with spin echoes in NMR is obvious and quite appropriate. By scanning the pulse separation and measuring the echo intensity, the homogeneous lifetime of the transition can be measured. A further level of sophistication is the three pulse photon echo (Cho *et al.*, 1992). Figure 21.9 shows the time orderings of the pulses and the emitted signal, related to the third-order nonlinear polarization, $\mathrm{P}^{(3)}(t')$, together with the 'box' geometry usually applied. The times τ and T denote the center to center distances between the pulse pairs E_1, E_2 and E_2, E_3, respectively. Also, t_1, t_2, and t_3 denote field–matter interaction points related to the response function

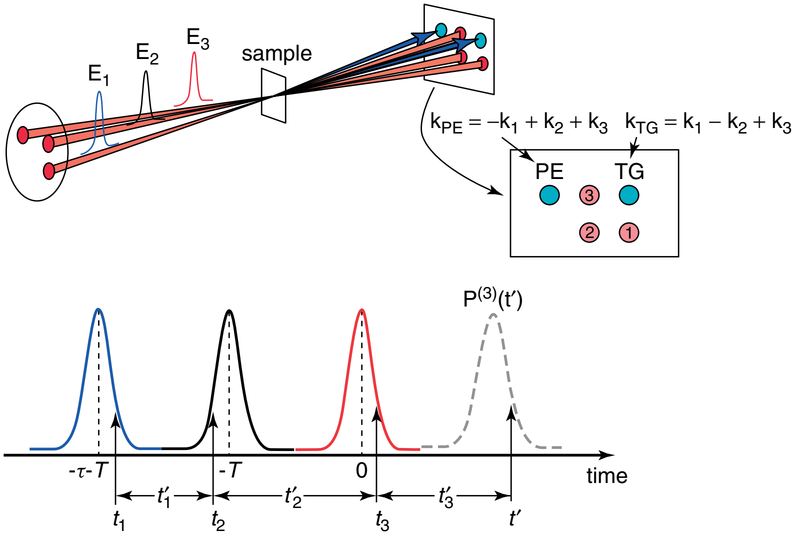

Fig. 21.9 Three pulse photon echo time ordering and spatial beam geometry. Also shown is the box geometry highlighting the phase matched direction for the photon echo signal (PE) and the transient grating signal (TG). (Adapted from Andrei Tokmakoff, 5.74 Introductory Quantum Mechanics II, Spring 2007. (MIT OpenCourseWare: Massachusetts Institute of Technology) http://ocw.mit.edu/OcwWeb/Chemistry/5-74Spring-2007/CourseHome. Accessed Oct. 1, 2008.)

formalism Eq. (21.10), while t'_1, t'_2, and t'_3 denote the relative delays between the interaction points. In this experiment, there are three pulses involved and two delays to be controlled. Changing the delays independently yields quite different dynamic information.

Varying t'_1 yields the homogeneous lifetime, while scanning t'_2 yields the population relaxation time. In analogy with NMR, the measured times are then T_2 and T_1, respectively, although the factors influencing the optical relaxation times are quite different to the nuclear spin case (and the relaxation times in molecules are many orders of magnitude faster). The phase matching direction for the emitted photon echo signal is now given by $\mathbf{k}_{\mathrm{PE}} = -\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3$, as illustrated in Figure 21.9. Also, when $t_1 - t_2 = 0$ and measuring the wave vectors phase matched beam in the direction given by $\mathbf{k}_{\mathrm{TG}} = \mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3$, the experiment is analogous to transient grating scattering. The photon echo and transient grating techniques are also known as *rephasing and non-rephasing methods*, respectively. In terms of the response function contributions (Eqs (21.27)–(21.30)), it is possible to show that the terms that describe the photon echo signal are given by the sum of $R_2 + R_3$ and for the transient grating the terms that contribute are $R_1 + R_4$. Again, this illustrates the generality and completeness of the response functions formalism (Mukamel, 1995).

Unfortunately, as discussed in literature, the photon echo intensity measurements as a function of the delays

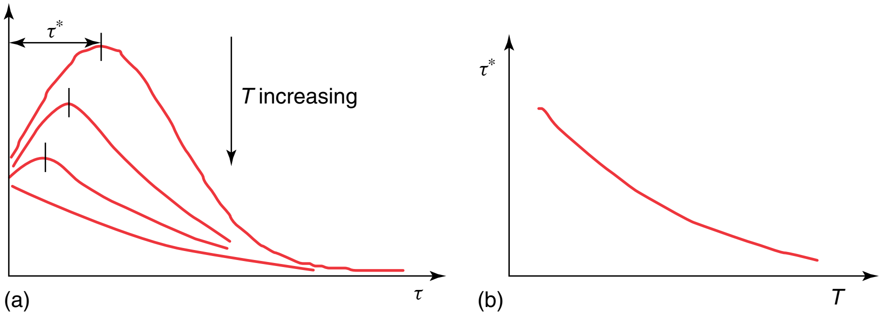

Fig. 21.10 (a) Drawing representing the evolution of the photon echo signal as the population time, T, is increased. (b) Plotting the photon echo peak against T results in the echo peak shift measurement. (Adapted from Andrei Tokmakoff, 5.74 Introductory Quantum Mechanics II, Spring 2007. (MIT OpenCourseWare: Massachusetts Institute of Technology) http://ocw.mit.edu/OcwWeb/Chemistry/5-74Spring-2007/CourseHome. Accessed Oct. 1, 2008.)

t_1' and t_2' provide no direct information on the system-bath interaction dynamics, unless there is a clear and strict separation of timescales as, for example, in the Bloch limit (Bloch, 1946; Nibbering *et al.*, 1994). Usually when working with solutions, the dynamics cannot be classified simply as homogeneous or inhomogeneous. Recently for a probe parameter that could follow the dynamics of the frequency correlation function was sought. An apparently excellent candidate is the measurement of the echo peak shift (Desilvestri *et al.*, 1984; Weiner *et al.*, 1985). With the advent of amplified laser systems and optical parametric oscillator/amplifiers, which in conjunction are able to deliver pulses shorter than 20 fs in a broad wavelength region, the subject was revived and the echo peak shift technique has attracted new interest. In addition, very recently, it has been shown that there is a direct connection between the echo peak shift measurement and the amplitude of the frequency correlation function fluctuations due to the system-bath interaction (Fleming and Cho, 1996). In the three pulse echo peak shift (3 PEPS) experiment, the integrated photon echo intensity is measured for a fixed population delay time, T, while scanning the coherence delay time, τ. The shift of the photon echo maximum with respect to zero delay along the τ axis constitutes the observable, the so-called echo peak shift. It was demonstrated that a plot of the echo peak shift as a function of waiting time T, (Figure 21.10) tracks the frequency correlation function, which carries information about the system-bath interaction (de Boeij *et al.*, 1996). Such experiments represent a powerful tool for unraveling optical dynamics in complex condensed phase systems.

21.5
The Future

The technology and the theory of ultrafast spectroscopy have entered into a mature phase. The future now would appear to lie in establishing the range and power of these new methods. One particularly striking example currently under intense investigation is multidimensional ultrafast spectroscopy. In preceding sections, the analogy of optical and NMR measurements was noted. Extending this idea to the multipulse experiments described above suggests the possibility of a suite of ultrafast experiments to match those associated with multipulse NMR. Numerous researchers are investigating the possibility of exploiting two-dimensional experiments, characterized my multiple pulses and time delays (of which 3PEPS is one example) to determine molecular structural evolution on an ultrafast time scale. In many respects, this is the holy grail of molecular science – a method that demonstrates where the nuclei are during the course of a chemical reaction. Important progress is being made, particularly, in terms of two-dimensional vibrational spectroscopy (Demirdoven *et al.*, 2004; Khalil *et al.*, 2003; Wright, 2002; Zheng *et al.*, 2007).

References

Bloch, F. (1946) *Phys. Rev.*, **70**, 460.

de Boeij, W.P., Pshenichnikov, M.S., and Wiersma, D.A. (1996) *Chem. Phys. Lett.*, **253**, 53.

de Boeij, W.P., Pshenichnikov, M.S., and Wiersma, D.A. (1998) *Annu. Rev. Phys. Chem.*, **49**, 99.

Boyd, R.W. (ed.) (2003) *Nonlinear Optics*, 2nd edn, Academic Press, San Diego.

Butcher, P.N. and Cotter, D. (eds) (1990) *The Elements of Nonlinear Optics*, 1st edn, Cambridge University Press, New York.

Castner, E.W., Wishart, J.F., and Shirota, H. (2007) *Acc. Chem. Res.*, **40**, 1217.

Cerullo, G. and De Silvestri, S. (2003) *Rev. Sci. Instrum.*, **74**, 1.

Cho, M.H., Fleming, G.R., and Mukamel, S. (1993) *J. Chem. Phys.*, **98**, 5314.

Cho, M.H., Scherer, N.F., Fleming, G.R., and Mukamel, S. (1992) *J. Chem. Phys.*, **96**, 5618.

Cong, P., Deuel, H.P., and Simon, J.D. (1995) *Chem. Phys. Lett.*, **240**, 72.

Demirdoven, N., Cheatum, C.M., Chung, H.S., Khalil, M., Knoester, J., and Tokmakoff, A. (2004) *J. Am. Chem. Soc.*, **126**, 7981.

Desilvestri, S., Weiner, A.M., Fujimoto, J.G., and Ippen, E.P. (1984) *Chem. Phys. Lett.*, **112**, 195.

Fleming, G.R. (1986) *Chemical Applications of Ultrafast Spectroscopy*, Oxford University Press, New York.

Fleming, G.R. and Cho, M.H. (1996) *Annu. Rev. Phys. Chem.*, **47**, 109.

Fork, R.L., Greene, B.I., and Shank, C.V. (1981) *Appl. Phys. Lett.*, **38**, 671.

Fourkas, J. (ed.) (2002) *Liquid Dynamics: Experiment, Simulation and Theory*, ACS Symposium, American Chemical Society.

Friedman, J.S. and She, C.Y. (1993) *J. Chem. Phys.*, **99**, 4960.

Heisler, I.A., Correia, R.R.B., Buckup, T., and Cunha, S.L.S. (2005) *J. Chem. Phys.*, **123**, 054509.

Hunt, N.T., Jaye, A.A., and Meech, S.R. (2007) *Phys. Chem. Chem. Phys.*, **9**, 2167.

Hunt, N.T. and Meech, S.R. (2003) *Chem. Phys. Lett.*, **378**, 195.

Hunt, N.T. and Meech, S.R. (2004) *Chem. Phys. Lett.*, **400**, 368.

Jean-Claude, D. and Rudolph, W. (eds) (1996) *Ultrashort Laser Pulse Phenomena*, 4th edn, Academic Press, San Diego.

Joo, T. and Rhee, H. (2005) *Opt. Lett.*, **30**, 96.

Kang, T.J., Jarzeba, W., Barbara, P.F., and Fonseca, T. (1990) *Chem. Phys.*, **149**, 81.

Kano, S. and Levenson, M.D. (1988) *Introduction to Nonlinear Laser Spectroscopy*, Oxford University Press, New York.

Kaufmann, K.J., Dutton, P.L., Netzel, T.L., Leigh, J.S., and Rentzepis, P.M. (1975) *Science*, **188**, 1301.

Khalil, M., Demirdoven, N., and Tokmakoff, A. (2003) *J. Phys. Chem. A*, **107**, 5258.

Kurnit, N.A., Abella, I.D., and Hartmann, S.R. (1964) *Phys. Rev. Lett.*, **13**, 567.

Lakowicz, J.R. (ed.) (2006) *Principles of Fluorescence Spectroscopy*, 3rd edn, Springer-Verlag, New York.

Lotshaw, W.T., McMorrow, D., Thantu, N., Melinger, J.S., and Kitchenham, R. (1995) *J. Raman Spectrosc.*, **26**, 571.

Loughnane, B.J., Scodinu, A., Farrer, R.A., Fourkas, J.T., and Mohanty, U. (1999) *J. Chem. Phys.*, **111**, 2686.

Maroncelli, M. and Fleming, G.R. (1987) *J. Chem. Phys.*, **86**, 6221.

McMorrow, D., Lotshaw, W.T., and Kenney-Wallace, G.A. (1988) *IEEE J. Quantum Electron.*, **24**, 443.

van der Meer, M.J., Zhang, H., and Glasbeek, M. (2000) *J. Chem. Phys.*, **112**, 2878.

Mukamel, S. (ed.) (1995) *Principles of Nonlinear Optical Spectroscopy*, 1st edn, Oxford University Press, Mineola, New York.

Nibbering, E.T.J., Wiersma, D.A., and Duppen, K. (1994) *Chem. Phys.*, **183**, 167.

Perry, M.D. and Mourou, G. (1994) *Science*, **264**, 917.

Phillips, D. and O'Connor, D.V. (1983) *Time-Correlated Single Photon Counting*, Academic Press, London.

Porter, G. and Wilkinson, F. (1961) *Trans. Faraday Soc.*, **57**, 1686.

Rulliere, C. (ed.) (1998) *Femtosecond Laser Pulses*, 1st edn, Springer-Verlag, Berlin.

Shah, J. (1988) *IEEE J. Quantum Electron.*, **24**, 276.

Shen, Y.R. (ed.) (1984) *The Principles of Nonlinear Optics*, 4th edn, John Wiley & Sons, Inc., New York.

Smith, N.A. and Meech, S.R. (2002) *Int. Rev. Phys. Chem.*, **21**, 75.

Spence, D.E., Kean, P.N., and Sibbett, W. (1991) *Opt. Lett.*, **16**, 42.

Svelto, O. (1998) *Principles of Lasers*, Plenum Press, New York.

Todd, D.C., Jean, J.M., Rosenthal, S.J., Ruggiero, A.J., Yang, D., and Fleming, G.R. (1990) *J. Chem. Phys.*, **93**, 8658.

Towrie, M., Grills, D.C., Dyer, J., Weinstein, J.A., Matousek, P., Barton, R., Bailey, P.D., Subramaniam, N., Kwok, W.M., Ma, C.S., Phillips, D., Parker, A.W., and George, M.W. (2003) *Appl. Spectrosc.*, **57**, 367.

Trebino, R., DeLong, K.W., Fittinghoff, D.N., Sweetser, J.N., Krumbugel, M.A., Richman, B.A., and Kane, D.J. (1997) *Rev. Sci. Instrum.*, **68**, 3277.

Waldeck, D., Cross, A.J., McDonald, D.B., and Fleming, G.R. (1981) *J. Chem. Phys.*, **74**, 3381.

Weiner, A.M., Desilvestri, S., and Ippen, E.P. (1985) *J. Opt. Soc. Am. B*, **2**, 654.

Wilhelm, T., Piel, J., and Riedle, E. (1997) *Opt. Lett.*, **22**, 1494.

Wright, J.C. (2002) *Int. Rev. Phys. Chem.*, **21**, 185.

Zheng, J., Kwak, K., and Fayer, M.D. (2007) *Acc. Chem. Res.*, **40**, 75.

23

Surface Second Harmonic and Sum-Frequency Generation

Mitsumasa Iwamoto, Takaaki Manaka and Eunju Lim

23.1 **Introduction 819**
23.2 **Order Parameters and Surface Layer 819**
23.3 **Nonlinear Polarization and Surface Optical Second Harmonic Generation 820**
23.3.1 SHG as a Tool for Characterizing Surface Structure 821
23.3.2 EFISHG as a Tool for Probing Excess Surface Charges 823
23.3.3 Spectroscopic Consideration of the SHG 823
23.3.4 Sum-Frequency Generation 825
23.4 **Measurement Details of SHG and EFISHG 826**
23.4.1 Determination of Surface Structure and Orientational Order Parameters 826
23.4.1.1 Fundamentals 826
23.4.1.2 Experimental Configuration of SHG Measurement 826
23.4.1.3 Experimental Configuration of SFG Measurement 826
23.4.2 Probing of the SHG Intensity Profile as a Convolution Beam Pattern 827
23.4.2.1 Fundamentals 827
23.4.2.2 Experimental Configuration of EFISHG Measurement from OFET Structure 828
23.5 **Results of SHG and EFISHG 829**
23.5.1 Determination of Surface Structure and Orientational Order Parameters 829
23.5.2 Probing of Electric Field $E(0)$ by EFISHG Measurement 830
References 831
Further Reading 832

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.

Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

ISBN: 978-3-527-40773-6

23.1
Introduction

The use of optical second harmonic generation (SHG) measurement for the exploration of dielectric polarization and related surface phenomena is described. A classification of materials using orientational order parameters shows that a surface monolayer is the most generalized system for discussing SHG enhancement from materials, and a theory developed for a monolayer system is broadly extended to other systems such as bulk system materials. On the basis of this background, SHG enhancement from an organic monolayer system is described with a schematic illustration of a typical experimental system. It is shown that SHG is generated from surface monolayers due to their noncentrosymetric structure and is available for the determination of orientational order of surface layers, and so on. It is also shown that SHG is generated from layers with centrosymmetric structure that are subjected to a high static electric field, owing to induced effective dipoles. Results show that enhanced SHG is available for probing excess surface charges. In addition, optical sum-frequency generation (SFG) is briefly described, and we show that this can be available as spectroscopic method for detecting the motion of molecules at the molecular level.

23.2
Order Parameters and Surface Layer

The orientational order parameter is specially important to specify the dielectric property of surface layers that generate SHG and SFG. Material systems composed of rod molecules are classified using orientational order parameters S_n $(n = 1, 2, 3)$ defined by the thermodynamic average of the Legendre polynomials, $P_n(\cos\theta)$ $(n = 1,\ 2,$ and $3)$, of the orientational angle θ of molecules with respect to the surface. For isotropic bulk materials, rod molecules are randomly distributed in space. Thereby $S_1 = \langle P_1(\cos\theta)\rangle = \langle\cos\theta\rangle = 0$, $S_2 = \langle P_2(\cos\theta)\rangle = \langle(3\cos^2\theta - 1)/2\rangle = 0$, and $S_3 = \langle P_3(\cos\theta)\rangle = \langle(5\cos^3\theta - 3\cos\theta)/2\rangle = 0$. On the other hand, orientational direction of molecules is restricted for ordered materials. For example, constituent molecules in nematic phase bulk liquid crystals (LCs) are totally distributed in "up" and "down" directions, without distinction between them. Therefore, $S_1 = 0$, $S_3 = 0$, but $S_2 \neq 0$. Interestingly, for a monolayer composed

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

of rod molecules, all molecules point upward owing to symmetry breaking. Therefore, S_1, S_2, $S_3 \neq 0$. This simple classification suggests that the physical properties of materials are expressed in terms of orientational order parameters. For example, nematic LCs are characterized using $S_2 \neq 0$. Similarly, the physical properties of surface layers such as monolayers are described using nonzero order parameters S_1, S_2, and S_3. Obviously, nonzero values for the parameters, S_1 and S_3 are specific parameters, are connected to symmetry breaking. In other words, the specific physico-chemical properties of surface layers, for example, monolayers are expressed using nonzero S_1 and S_3. Hence, the generation of SH from surface layers can be discussed in terms of orientational order parameters S_1 and S_3.

23.3
Nonlinear Polarization and Surface Optical Second Harmonic Generation

In the absence of a static electric field, that is, a DC field, the SHG signal originates from a nonlinear dielectric polarization produced by the quantum interaction between the electrons in molecules and the external electric field, that is, laser light Bloembergen, 1965; Shen, 1984. Under the electric dipole approximation, SHG is allowed from the noncentrosymmetric system. On account of symmetry breaking, organic monolayers on metal, film, and water surface possess a noncentrosymmetric structure. Hence, SHG is available for characterizing orientational structure of such monolayers.

Interestingly, SHG is also allowed from a centrosymmetric system if the electric quadrupole effect is prominent. SHG due to the quadrupole effect is a bulk phenomenon and observed in thin films composed of centrosymmetric molecules such as phthalocyanine (Pc) Chollet, Kajzar and Le Moigne, 1990, fullerene (C_{60}) Wang *et al.*, 1992, and other organic molecular films. Fortunately, the resonant conditions are different between the electric dipole mechanism and the electric quardrupole mechanism, and one can distinguish the two mechanisms by appropriately choosing a wavelength of laser used in the SHG experiment.

On the other hand, in the presence of a static electric field, the electron distribution of any centrosymmetric molecule is distorted, and an effective dipole can be induced (DC-field-induced dipole). In this situation, a different type of nonlinear polarization is produced in the centrosymmetric molecular system, and SH can be freshly generated. This is called *electic-field-induced SHG* (*EFISHG*). In organic electronic devices, there are many origins leading to the formation of local static electric field Sze, 1981; Sessler, 1987. Obviously external DC biasing used for the organic device operation is one of the possible origins. A displacement of excess charges at the metal–organic material system, injection of carriers form the metal electrode into the organic device, and so on, are other possible origins. In the following sections, our discussion is focused on SHG due to the electric dipole effect and EFISHG, because these SHG processes are very important for probing surface polarization and related phenomena.

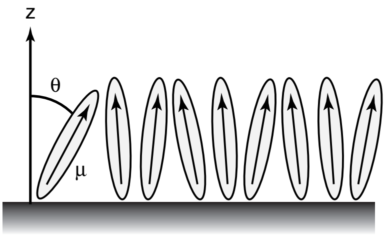

Fig. 23.1 Schematic image of the molecular orientation on a material surface.

23.3.1
SHG as a Tool for Characterizing Surface Structure

The orientational order of organic monolayers comprised of rodlike molecules is specified by using nonzero orientational order parameters, S$_n$ ($n = 1, 2, 3$), defined by the thermodynamic average of the Lagendre polynomials, P$_n$(cos θ) ($n = 1, 2, 3$), of the orientational angle θ of molecules from the normal to the surface (Figure 23.1) Iwamoto and Wu, 2000; Luckhurst and Gray, 1979. Nonlinear dielectric polarization contributing to the SHG is characterized by the second-order nonlinear electronic polarizability $\langle \overset{\leftrightarrow}{\alpha}^{(2)} \rangle$ at an angular frequency ω of the external electric field. This polarization is expressed approximately as

$$\mathbf{P}^{(2)} = \chi^{(2)} : \mathbf{EE} = N_s \langle \overset{\leftrightarrow}{\alpha}^{(2)} \rangle : \mathbf{EE} \quad (23.1)$$

The second-order susceptibility (SOS) $\chi^{(2)} \equiv [\chi_{ijk}]$ of the monolayer is related to the components of the nonlinear molecular electronic polarizability $\alpha_{i',j',k'}$ by $\chi_{ijk} = N_s \langle T_{ijk}^{i'j'k'} \rangle \alpha_{i'j'k'}$, where $T_{ijk}^{i'j'k'} = R_i^{i'} R_j^{j'} R_k^{k'}$ with the usual Euler rotation matrix $R(\phi, \theta, \psi) = [R_i^{i'}]$ describing the coordinate transformation between the molecular system (i', j', k') and the laboratory system (x, y, z) Heinz, 1991. The complete expression of the macroscopic SOS tensor $\chi^{(2)}$ of a monolayer with C_∞-symmetry on a material surface is obtained as a function of the components of the molecular nonlinear electronic polarizability $\alpha_{i',j',k'}$ and the orientational order parameters $S_n \equiv \langle P_n(\cos\theta) \rangle$. With SHG one can deduce $\chi^{(2)}$.

Nonlinear polarization $\mathbf{P}^{(2)} = \mathbf{P}^N$ is a sum of the polarization $\mathbf{P}_{\mathrm{ch}}^N$ associated with the chirality of the monolayer and the polarizations $\mathbf{P}_{\mathrm{ach}}^N$ associated with the nonchirality of the monolayer. The two polarizations in vector form are Iwamoto, Wu and Ou-Yang, 2000

$$\begin{aligned} \mathbf{P}_{\mathrm{ch}}^N = {} & \frac{1}{2} s_{14} [(\mathbf{E} \cdot \mathbf{n})(\mathbf{F} \times \mathbf{n}) + (\mathbf{F} \cdot \mathbf{n}) \\ & \times (\mathbf{E} \times \mathbf{n})] + \frac{1}{2} a_{14} (\mathbf{E} \times \mathbf{F}) \\ & + \frac{1}{2} (a_{36} - a_{14}) \mathbf{n} \cdot (\mathbf{E} \times \mathbf{F}) \mathbf{n} \end{aligned} \quad (23.2)$$

$$\begin{aligned} \mathbf{P}_{\mathrm{ach}}^N = {} & (s_{33} - s_{15} - s_{31})(\mathbf{n} \cdot \mathbf{E})(\mathbf{n} \cdot \mathbf{F}) \mathbf{n} \\ & + s_{31} (\mathbf{E} \cdot \mathbf{F}) \mathbf{n} + \frac{1}{2} s_{15} [(\mathbf{n} \cdot \mathbf{F}) \mathbf{E} \\ & + (\mathbf{n} \cdot \mathbf{E}) \mathbf{F}] + \frac{1}{2} a_{15} (\mathbf{E} \times \mathbf{F}) \times \mathbf{n} \end{aligned} \quad (23.3)$$

with seven independent nonzero elements given as

$$s_{14} = \frac{N_s}{2} S_2 (\sigma_{14} - \sigma_{25}) = \chi_{123} + \chi_{132}$$

$$\begin{aligned}
s_{15} &= \frac{N_s}{5}(S_1 - S_3)(2\sigma_{33} - \sigma_{32} - \sigma_{31}) \\
&\quad + \frac{N_S}{10}(3S_1 + 2S_3)(\sigma_{24} + \sigma_{15}) \\
&= \chi_{131} + \chi_{113} \\
s_{31} &= \frac{N_s}{10}(S_1 - S_3)(2\sigma_{33} - \sigma_{24} - \sigma_{15}) \\
&\quad + \frac{N_S}{10}(4S_1 + S_3)(\sigma_{32} + \sigma_{31}) \\
&= \chi_{311} \\
s_{33} &= \frac{N_s}{5}(S_1 - S_3)(\sigma_{32} + \sigma_{31} + \sigma_{24} \\
&\quad + \sigma_{15}) + \frac{N_s}{5}(3S_1 + 2S_3)\sigma_{33} \\
&= \chi_{333} \\
a_{14} &= \frac{N_s}{3}(\lambda_{14} + \lambda_{25} + \lambda_{36}) \\
&\quad + \frac{N_s}{6}S_2(\lambda_{14} + \lambda_{25} - 2\lambda_{36}) \\
&= \chi_{123} - \chi_{132} \\
a_{15} &= \frac{N_s}{2}S_1(\lambda_{15} - \lambda_{24}) = \chi_{131} - \chi_{113} \\
a_{36} &= \frac{N_s}{3}(\lambda_{14} + \lambda_{25} + \lambda_{36}) \\
&\quad + \frac{N_s}{3}S_2(2\lambda_{36} - \lambda_{14} - \lambda_{25}) \\
&= \chi_{312} - \chi_{321}
\end{aligned} \tag{23.4}$$

where the two 3 × 6 matrices, (σ_{ij}) and (λ_{ij}), are defined by the components of the nonlinear molecular polarizability $\alpha_{i',j',k'}$ in the same way as the conventional contracted notation of (s_{ij}) and (a_{ij}) defined from $\chi^{(2)}$ Shen, 1984; Nye, 1957, that is, $\sigma_{11} = \alpha_{111}$, $\sigma_{14} = \alpha_{123} + \alpha_{132}$, $\lambda_{14} = \alpha_{123} - \alpha_{132}$, and so on, and **E** and **F** are the external electric fields. **n** is unit vector normal to the water surface. In Eqs (23.2) and (23.3), two kinds of external electric fields **E** and **F** are used. In the SHG measurement, the same electric field **E** = **F** from an optical source is used, but in the SFG measurement, and so on, two independent electric field sources **E** ≠ **F** are used. Here, we confine our discussion to the case of SHG, but we can carry similar discussion on SFG, and so on, using Eqs (23.2) and (23.3). In the following analysis, therefore **E** = **F** is assumed.

From a nonlinear polarization point of view, the orientational order parameters S_1 and S_3 are related to the nonlinear polarization associated with the nonchirality of the surface monolayers, whereas S_2 is related to the nonlinear polarization associated with the chirality of the surface layers. In the case of monolayers with $C_{\infty v}$-symmetry, where the constituent molecules are achiral, that is, $s_{14} = a_{14} = a_{36} = 0$, the nonlinear electric polarization in SHG (**E** = **F**) reduces to

$$\begin{aligned}
\mathbf{P}^N &= (s_{33} - s_{31} - s_{15})(\mathbf{n}\cdot\mathbf{E})^2\mathbf{n} \\
&\quad + s_{31}(\mathbf{E}\cdot\mathbf{E})\mathbf{n} + s_{15}(\mathbf{n}\cdot\mathbf{E})\mathbf{E}
\end{aligned} \tag{23.5}$$

The nonlinear dielectric polarization is a function of the order parameters S_1 and S_3 together with the three elements related to the nonchirality, s_{31}, s_{33}, and s_{15}.

Similarly, in case of monolayers with C_∞-symmetry, where the constituent molecules are chiral, that is, $s_{14} \neq a_{14} \neq a_{36} \neq 0$, the nonlinear electric polarization in SHG (**E** = **F**) is given by:

$$\begin{aligned}
\mathbf{P}^N &= s_{14}(\mathbf{E}\cdot\mathbf{n})\cdot(\mathbf{E}\times\mathbf{n}) \\
&\quad + (s_{33} - s_{31} - s_{15})(\mathbf{n}\cdot\mathbf{E})^2\mathbf{n} \\
&\quad + s_{31}(\mathbf{E}\cdot\mathbf{E})\mathbf{n} + s_{15}(\mathbf{n}\cdot\mathbf{E})\mathbf{E}
\end{aligned} \tag{23.6}$$

The first term on the right-hand side of Eq. (23.6) as a function of s_{14} with

S_2 (see Eq. (23.4)) represents the chiral effect. As mentioned earlier, the quantum interaction between the electrons in molecules and the external electric field makes the main contribution to SHG, and thus the nonlinear dielectric polarization given by Eqs (23.2) and (23.3) becomes the main contribution to the SHG. Equations (23.5) and (23.6) suggest that one can determine the orientational order parameters of S_1, S_2, and S_3 by means of SHG caused by the electric dipole effect.

23.3.2
EFISHG as a Tool for Probing Excess Surface Charges

The nonlinear polarization for the SHG in the medium is expressed not only as Eq. (23.1) under the electric dipole approximation but also in well-known tensor form as Shen, 1984

$$P_i = \chi^D_{ijk} E_j(\omega) E_k(\omega) \tag{23.7}$$

where χ^D_{ijk} is the second-order nonlinear susceptibility due to the electric dipole mechanism, and E_j and E_k are the electric fields of fundamental laser light. The effective polarization due to the electric quadrupole also contributes to the second-order nonlinear polarization, and it is expressed as

$$P_i = \chi^Q_{ijkl} \partial_j E_k(\omega) E_l(\omega) \tag{23.8}$$

where χ^Q_{ijkl} is the second-order nonlinear susceptibility due to the electric quadrupole mechanism.

On the one hand, second-order nonlinear polarization is activated by the external static field in case of the DC EFISHG. Since three electric fields $E_j(0)$, $E_k(\omega)$, and $E_l(\omega)$, are concerned with the generation of the nonlinear polarization, the nonlinear optical susceptibility becomes fourth-order rank tensor that does not vanish even in the centrosymmetric system. Thus, the effective polarization is expressed as

$$P_i = \chi^{\mathrm{DC}}_{ijkl} E_j(0) E_k(\omega) E_l(\omega) \tag{23.9}$$

where $E_j(0)$ is the external static electric field induced by the local electric field, and $\chi^{\mathrm{DC}}_{ijkl}$ is the third-order nonlinear susceptibility due to the DC-induced mechanism and $\chi^{\mathrm{DC}}_{ijkl} E_j(0)$ corresponds to the effective dipole induced. Note that although this contribution is the third-order nonlinear optical process, double frequency of the fundamental light is generated because one of the related field is static. As mentioned earlier, there are many possible origins leading to the generation of nonzero local electric field $\mathrm{E}_j(0) \neq 0$. Among them are external electric field, injected carriers, trapped charges, and so on. Hence, we can probe these local electric fields $\mathrm{E}_j(0)$ by means of EFISHG.

23.3.3
Spectroscopic Consideration of the SHG

SHG originates from the quantum interaction between the electrons in molecules and the external electric field. In this section, we briefly summerize the SHGs from the viewpoint of energetics based on the quantum theory. As mentioned in the earlier sections, we should take into account the following three contributions to the SHG:

- electric dipole mechanism (χ_D);
- electric quadrupole mechanism (χ_Q); and
- induced electric dipole mechanism (χ_{DC})

and the nonlinear optical susceptibilities originated from these contributions show the different dependence on the wavelength of each related electric field, owing to the difference in the quantum interaction between the electrons in the molecules and the electric field. Thus, we can distinguish each mechanisms from a view point of spectroscopy. There are many ways to evaluate the frequency-dependent nonlinear susceptibilities. According to the quantum mechanical perturbation theory, the nonlinear susceptibilities for the SHG can be calculated as the second-order contribution to the induced dipole moment. Some of the molecular orbital (MO) calculation programs also enable us to estimate the frequency-dependent nonlinear susceptibilities. First, we discuss the resonant condition for the SHG schematically.

Schematic images for the resonance conditions are shown in Figure 23.2. In this figure, state $|g\rangle$ represents a ground state and $|n\rangle$, $|m\rangle$, and $|p\rangle$ represent the virtual excited states. Selection rules are employed to discuss the SHG, that is, the transitions between states of the same parity are forbidden for the electric dipole transition, whereas transitions between states of the same parity are allowed for the electric quadrupole transition. Considering these diagrams and the selection rules, we can estimate the resonant conditions for the SH frequency, 2ω. Clearly, for the electric dipole mechanism (Figure 23.2(a)), the lack of the centrosymmetry is required to radiate the SH wave, and resonance enhancement occurs only when the SH photon energy is coincident with the allowed excited state. On the other hand, resonance enhancement occurs only when the SH photon energy is coincident with the forbidden excited state for the electric quadrupole mechanism (Figure 23.2(b)). The case of the induced electric dipole mechanism is slightly different. This process is activated by the external field, as mentioned above. As shown in Figure 23.2(c), state $|n\rangle$ overlies $|p\rangle$ state. This indicates that the resonance enhancement occurs when the SH photon energy is coincident with both the allowed and forbidden excited states for this mechanism. Taking into account the resonant conditions discussed above, we can discuss the origin of the SHG from organic film by the spectroscopic observation. In more detail, χ_D, χ_Q and χ_{DC} are described as

$$\chi_{ijk}^{D} \propto \sum_{m,n} \left(\frac{\langle g|\hat{\mu}_i|n\rangle\langle n|\hat{\mu}_j|m\rangle\langle m|\hat{\mu}_k|g\rangle}{\left(\omega_{ng} - 2\omega\right)\left(\omega_{mg} - \omega\right)} \right.$$

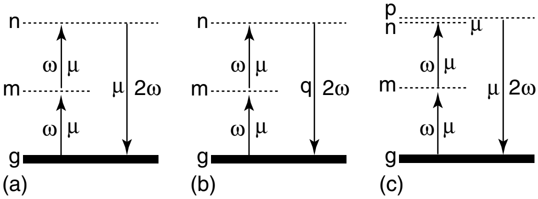

Fig. 23.2 Transition diagrams for the SH process: (a) electric dipole, (b) electric quadrupole, and (c) induced electric dipole.

$$+\frac{\langle g|\hat{\mu}_j|n\rangle\langle n|\hat{\mu}_i|m\rangle\langle m|\hat{\mu}_k|g\rangle}{\left(\omega_{ng}+\omega\right)\left(\omega_{mg}-\omega\right)}$$

$$+\frac{\langle g|\hat{\mu}_k|n\rangle\langle n|\hat{\mu}_j|m\rangle\langle m|\hat{\mu}_i|g\rangle}{\left(\omega_{ng}+\omega\right)\left(\omega_{mg}+2\omega\right)}\Bigg) \qquad (23.10)$$

$$\chi^{Q}_{ijkl} \propto \sum_{m,n}\Bigg(\frac{\langle g|\hat{q}_{ij}|n\rangle\langle n|\hat{\mu}_k|m\rangle\langle m|\hat{\mu}_l|g\rangle}{\left(\omega_{ng}-2\omega\right)\left(\omega_{mg}-\omega\right)}$$

$$+\frac{\langle g|\hat{\mu}_k|n\rangle\langle n|\hat{q}_{ij}|m\rangle\langle m|\hat{\mu}_l|g\rangle}{\left(\omega_{ng}+\omega\right)\left(\omega_{mg}-\omega\right)}$$

$$+\frac{\langle g|\hat{\mu}_l|n\rangle\langle n|\hat{\mu}_k|m\rangle\langle m|\hat{q}_{ij}|g\rangle}{\left(\omega_{ng}+\omega\right)\left(\omega_{mg}+2\omega\right)}\Bigg) \qquad (23.11)$$

and

$$\chi^{\mathrm{DC}}_{ijkl} \propto$$

$$\sum_{m,n,o}\Bigg(\frac{\langle g|\hat{\mu}_i|p\rangle\langle p|\hat{\mu}_j|n\rangle\langle n|\hat{\mu}_k|m\rangle\langle m|\hat{\mu}_l|g\rangle}{\left(\omega_{pg}-2\omega\right)\left(\omega_{ng}-2\omega\right)\left(\omega_{mg}-\omega\right)}+\frac{\langle g|\hat{\mu}_j|p\rangle\langle p|\hat{\mu}_i|n\rangle\langle n|\hat{\mu}_k|m\rangle\langle m|\hat{\mu}_l|g\rangle}{\omega_{pg}\left(\omega_{ng}-2\omega\right)\left(\omega_{mg}-\omega\right)}$$

$$+\frac{\langle g|\hat{\mu}_j|p\rangle\langle p|\hat{\mu}_k|n\rangle\langle n|\hat{\mu}_i|m\rangle\langle m|\hat{\mu}_l|g\rangle}{\omega_{pg}\left(\omega_{ng}+\omega\right)\left(\omega_{mg}-\omega\right)}+\frac{\langle g|\hat{\mu}_j|p\rangle\langle p|\hat{\mu}_k|n\rangle\langle n|\hat{\mu}_l|m\rangle\langle m|\hat{\mu}_i|g\rangle}{\omega_{pg}\left(\omega_{ng}+\omega\right)\left(\omega_{mg}+2\omega\right)}\Bigg) \qquad (23.12)$$

where $\hat{\mu}_i$ and $\hat{q}_{ij}$ are the electric dipole and the electric quadrupole transition moment operator and are expressed as $\hat{\mu}_i = -e\hat{r}_i$ and $\hat{q}_{ij} = -\frac{1}{6}e(3\hat{r}_i\hat{r}_j - \mathbf{r}^2\delta_{ij})$, respectively. $|g\rangle$ represents a wave function of the ground state and $|m\rangle$ and $|n\rangle$ are those of the excited state. The notation ω_{mg} means the frequency difference between $|m\rangle$ state and $|g\rangle$ state, $\omega_m - \omega_g$.

However, we do not need to go into the details of χ_D, χ_Q, and χ_{DC}, when we use the SHGs for probing the orientational order of surface molecules and for probing the local electric field.

Equations (23.10)–(23.12) suggest that it is important to know the origins of SHG enhancement in terms of the resonant frequency.

23.3.4
Sum-Frequency Generation

In this section, SFG is described. SFG also becomes a powerful tool for investigating the surface and interface phenomena of materials, especially on the molecular level. For the SFG process, three optical waves with different frequencies are interacting. There is an energy conservation between two input waves (ω_1 and ω_2) and output wave (ω_3) as $\omega_1 + \omega_2 \to \omega_3$. As mentioned, both SHG and SFG are second-order nonlinear optical processes, and in other words, SFG is a generalization of SHG (ω_1 and ω_2 are degenerate for the SHG process). Accordingly, the SFG output vanishes from the materials with inversion symmetry under the electric dipole approximation as well as the SHG process. We can realize the surface selective vibrational spectroscopy measurement by using SFG, letting the input wave with a frequency ω_1 be visible and the wave with a frequency ω_2 be infrared light.

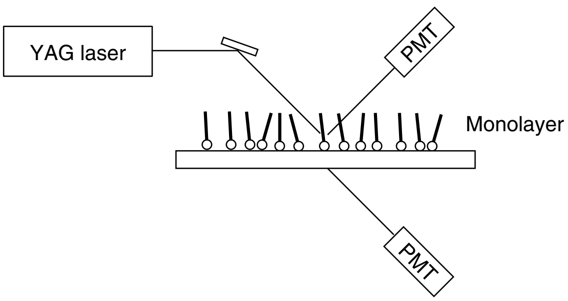

Fig. 23.3 Schematics of experimental setup for SHG measurement system is also attached. The chemical structure of 8CB.

23.4
Measurement Details of SHG and EFISHG

23.4.1
Determination of Surface Structure and Orientational Order Parameters

23.4.1.1 Fundamentals

SH light is generated from the monolayer by laser irradiation. The input light consists of the *s*- and the *p*-polarized waves and is expressed as

$$\mathbf{E}_{\text{in}} = S_\omega \mathbf{s}_{\text{in}} + P_\omega \mathbf{p}_{\text{in}} \tag{23.13}$$

where $\mathbf{s}_{\text{in}}$ and $\mathbf{p}_{\text{in}}$ are unit vectors for the *s*- and the *p*-polarized direction of the waves, and S_ω and P_ω are the amplitude of the *s*- and the *p*-polarized waves, respectively. Substituting $\mathbf{E}_{\text{in}}$ into Eqs (23.5) and (23.6), $\mathbf{P}^N$ generated in surface monolayers is obtained. The output light intensity $I_{2\omega}$ is proportional to $\mathbf{e}_{\text{out}}(2\omega) \cdot \mathbf{P}^N$, where $\mathbf{e}_{\text{out}}$ is unit vector for the output light direction of the wave. Since $\mathbf{e}_{\text{out}}(2\omega) \cdot \mathbf{P}^N$ is a function of S_1, S_2, and S_3, these orientational order parameters of surface monolayers can be determined with the enhanced SHG by the electric dipole effect Zhu, Suhr and Shen, 1987.

23.4.1.2 Experimental Configuration of SHG Measurement

Figure 23.3 shows a schematic diagram of the SHG measuring system for probing nonlinear polarization of surface layers. The SHG experiments can be carried out with the experimental setup illustrated. Briefly, the experimental setup consists of an optical measurement arrangement with a Q-switched Nd : YAG laser (wavelength 0.532 μm, pulse duration <7 nanoseconds, fundamental pulse rate $\leq$ 15 Hz) for the SHG measurement. The laser light is irradiated onto the surface layer with an energy of about 6 mJ at a pulse rate of 2 Hz. The size of the laser spot is about 56.0 mm^2.

23.4.1.3 Experimental Configuration of SFG Measurement

We can observe the SFG output of visible light, and a typical photomultiplier tube (PMT) can be used as a detector. Figure 23.4 represents the typical setup for the SFG measurement. Owing to the photon momentum conservation, the direction of SFG output is not

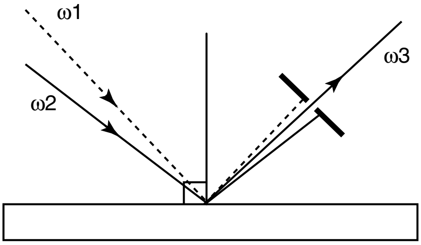

Fig. 23.4 Paths of light rays in SFG measurement.

coincident with those either of reflected fundamental waves (see Figure 23.4).

Experimentally, surface vibrational spectroscopy was first introduced by Zhu, Suhr and Shen 1987. They successful recorded a vibrational spectrum of a monolayer of molecules on a substrate by SFG. They chose second harmonics of YAG laser as a visible input and coumarin dye so that the resonance enhancement of the SFG signal occurred. They also mentioned the possibility to detect the transient surface phenomena because pulsed laser probing in SFG enables us to monitor the transient species appearing on a surface. As mentioned, SHG can also detect the orientation of molecules on a surface. Since the origin of SHG process involves electronic transitions in the molecule, the orientation of molecules directly reflects the orientation of the transition moments. In contrast, SFG can detect the orientation of the functional group of the molecule. In this sense, detailed information of the surface can be extracted from the SFG results.

23.4.2
Probing of the SHG Intensity Profile as a Convolution Beam Pattern

23.4.2.1 **Fundamentals**

For EFISHG, the SHG intensity is proportional to the square of local static electric field as

$$I(2\omega) \propto \left| \chi^{(3)}(-2\omega; 0, \omega, \omega) \times E(0)E(\omega)E(\omega) \right|^2 \quad (23.14)$$

where $\chi^{(3)}(-2\omega; 0, \omega, \omega)$ represents a third-order nonlinear optical (NLO) susceptibility for the EFISHG process, and $E(0)$ and $E(\omega)$ represent the static electric field and electric field of light, respectively. External electric fields, trapped charges, and injected carriers are origins of the static electric field, $E(0)$. Hence, the electric field distribution in materials as well as in organic devices can be obtained at various experimental conditions by the EFISHG measurement. In more detail, SHG intensity profile can be considered as a convolution of the beam pattern of the fundamental laser and the actual electric field distribution formed in organic films as Tojima, Manaka and Iwamoto, 2001

$$I^{2\omega}(x) \propto \left| \int_{-\infty}^{\infty} E(\xi) I^{\omega}(x-\xi) d\xi \right|^2 \quad (23.15)$$

$I^{2\omega}(x)$ represents the SHG intensity at position x, and $E(\xi)$ and $I^{\omega}(\xi)$ represent the electric field distribution and the intensity profile of the fundamental laser, respectively. If the laser beam

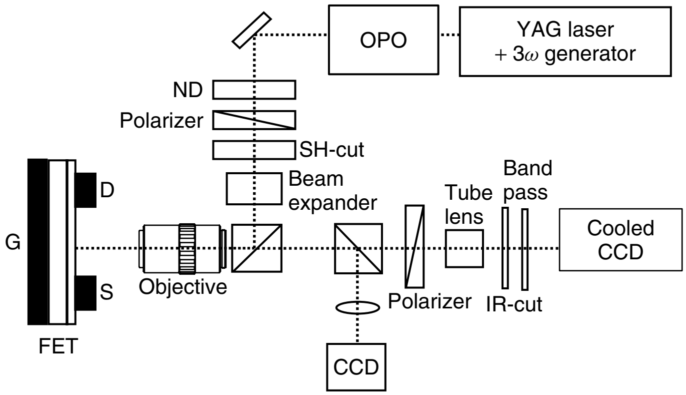

Fig. 23.5 Optical setup for the EFISHG measurement of OFET devices.

profile is approximated by δ-function as $I^{\omega}(x) \equiv \delta(x)$, $\sqrt{I^{2\omega}(x)}$ directly gives the local electric field $E(x)$. However, to obtain the accurate electric field distribution from the SHG, intensity profile requires a deconvolution process taking into consideration the beam profile of the fundamental laser, $I^{\omega}(x)$. Such deconvolution is, in general, a laborious task, because a beam profile and the SHG intensity profile are not expressed by simple mathematical functions, and the evaluated electric field from a scanning measurement sometimes loses accuracy, for example, spatial resolution. We can employ the deconvolution process for the determination of electric field distribution. On the other hand, an SHG imaging technique for visualizing the distribution of the electric field in organic materials can overcome these problems Manaka *et al.*, 2006. The evaluation of the electric field distribution from SHG images does not require the complicated deconvolution process, and the spatial resolution is significantly improved. Moreover, omission of the scanning process results in the reduction of the measurement time.

23.4.2.2 **Experimental Configuration of EFISHG Measurement from OFET Structure**

A detailed configuration about the EFISHG measurement is described in Manaka *et al.*, 2006. Figure 23.5 shows a schematic diagram of the EFISHG measuring system. The light source for the SHG measurement is an optical parametric oscillator (OPO: Continuum Surelite OPO), pumped by a third-harmonic light of Q-switched Nd–YAG laser (Continuum: SureliteII-10), a fundamental wavelength of 1120 or 1320 nm is chosen. Fundamental light from the OPO passes through a prism polarizer, long-pass filters, and a beam expander. Then it is focused on the sample with normal incidence, using a long working distance objective lens (Mitutoyo: M Plan Apo SL20 × ,NA = 0.28, WD = 20.5 mm). SH light generated from the sample is filtered by a fundamental-cut filter and an interference filter to remove fundamental and other unnecessary light. Finally, SH light is detected by a cooled charge coupled device (CCD) camera (Andor technology: DU420-BV). In this configuration, the polarization

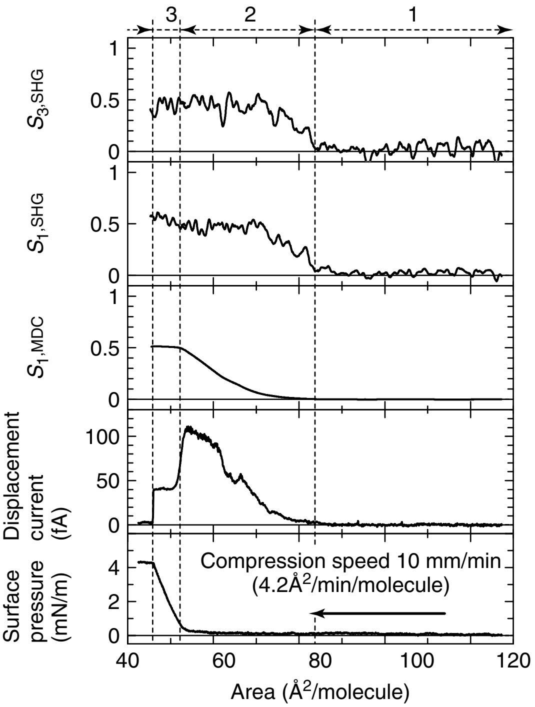

Fig. 23.6 A typical example of SHG and MDCs of 8CB monolayers during the course of monolayer compression. Order parameters, S1 and S3, of 8CB monolayer determined from the MDC-SHG measurement.

direction of the light is chosen in the direction parallel to the sample surface.

23.5
Results of SHG and EFISHG

23.5.1
Determination of Surface Structure and Orientational Order Parameters

In a manner as described in Section 23.3.1, we can carry out SHG experiment for probing orientational order of surface layers. Figure 23.6 shows an example of the determination of the order parameters S_1 and S_3 for 4-*n*-octyl-4′-cyanobiphenyl (8CB) monolayers with $C_{\infty v}$-symmetry on water surface. In region 1, S_1 and S_3 of SHG are very small, although fluctuations coming from the formation of domains are observed. In region 2, S_1 and S_3 increase gradually and show the maximum with decrease in the molecular area. In region 3, they are nearly constant, about 0.5. In contrast, S_1 determined by electrical Maxwell-displacement current (MDC) measurement is also plotted. S_1 is nearly zero in region 1 and increases monotonically in region 2. S_1 is about 0.5 in region 3. These results suggest that 8CB molecules lying on a water surface in large molecular area progressively stand up with a decrease in the molecular area and with an average tilting angle of about 60° in region 2. As mentioned earlier, surface polarization phenomena can be detected and the orientational orders of the surface layers can be well explored by the SHG caused by the electric dipole effect.

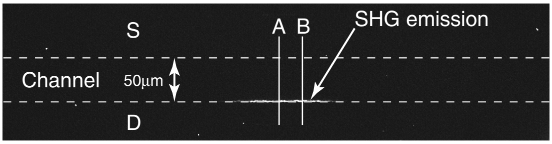

Fig. 23.7 SHG image from the channel of pentacene FET under the application of negative pulse. Channel region lies between two gold electrodes, and edges of the electrode are indicated by dashed lines. SHG emission was observed at the edge of the drain electrode.

23.5.2
Probing of Electric Field $E(0)$ by EFISHG Measurement

As mentioned in Section 23.3.2, the SHG intensity profile obtained on the basis of a scanning measurement does not display an actual spread of SHG emission from the material surface. Direct observations of the SHG image can show a more realistic distribution of the SHG emission within the limits of the system resolution. Figure 23.7 shows the SHG image from the channel of pentacene field effect transistor (FET) under the application of negative pulse. As shown in this figure, strong SHG emission is observed at an edge of drain electrode. High electric fields between the drain and gate electrodes produce a large in-plane component of the electric field, which activates the SHG at the drain edge. It is noteworthy that the width of the SHG emission in the channel clearly decreases compared with a distribution obtained on the basis of a scanning measurement.

Figure 23.8 shows the line scan of the SHG intensity profile across the channel (represented as open squares and filled diamonds) and the in-plane component of the in-plane electric field

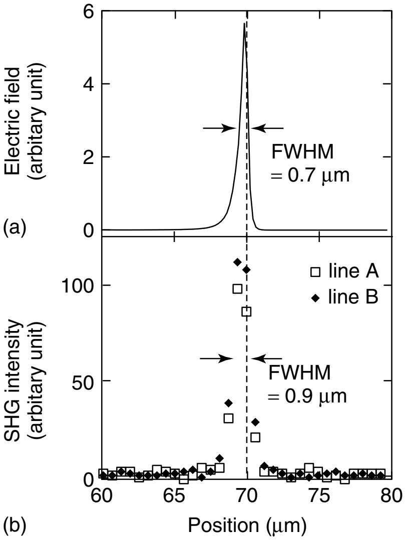

Fig. 23.8 (a) The figure represents the in-plane component of the in-plane electric field distribution in pentacene layer calculated based on a finite element method. (b) The figure shows the line scan of the SHG intensity profile across the channel. Open squares and filled diamonds, respectively, represent SHG intensity profile at line scan A and B as shown in Figure 23.7.

distribution (solid line) in the pentacene layer. Open squares and filled diamonds, respectively, represent the SHG intensity profile at line scan *A* and *B* (see Figure 23.7) as shown in Figure 23.8. The edge of the electrode is located at a position of 70 μm in these figures. It is found that the SHG intensity profile is quite sharp. The sharpness of emission is quantitatively estimated using full width at half maximum (FWHM) values of the profile. FWHM values of SHG profile and the electric field distribution are evaluated as 0.9 and 0.7 μm, respectively. Note that FWHM values of the SHG profile obtained based on the scanning measurement depended on the magnification of the objective lens and have the value of approximately 15 μm for 20 × objective.

For the electric field calculation when there are no excess charges in the device, only the electrode configuration and potential of the electrode are taken into account. In such a case, a Laplacian electric field is formed in the device. A Laplacian field is the electric field caused by the potentials of electrodes in the absence of any charges in the devices between the electrodes. Under negative bias, carrier injection from the electrode into pentacene is prohibited because the injection barrier for electrons at the pentacene/Au interface is quite high, that is, energy difference between the lowest unoccupied molecular orbital (LUMO) of pentacene and the work function of Au electrode is evaluated as 2.7 eV. Thus, the strong electric field around the edge of the electrode is maintained during bias application, and the SHG emission is concentrated around the edge of the drain electrode as shown in Figure 23.7. In other words, the sharp emission of the SHG at the edge of the electrode indicates that the SHG imaging technique successfully visualizes the electric field in the device with a spatial resolution of approximately 1 μm. As mentioned above, the electric field generated by the Laplacian electric field can be detected as the EFISHG, and the distribution of the electric field can be explored using the profile of EFISHG enhanced. Similarly, the electric field distribution caused by injected carriers and trapped charges can be evaluated by the EFISHG Zhu, Suhr and Shen, 1987. It is instructive to note that the migration of the peaks of electric field caused by injected carries is visualized by means of TRM-SHG measurement Manaka *et al.*, 2007a. Further, OFET operation has been successfully probed by SFG, where conformational change of the molecular structure, due to dopings of the molecules by carrier injection from electrodes, is probed Manaka *et al.*, 2007b.

References

Bloembergen, N. (1965) in *Nonlinear Optics* (ed. W.A. Benjamin.), Singapore, World Scientific, pp. 1–61.

Chollet, P.A., Kajzar, F. and Le Moigne, J. (1990), *Proceedings of SPIE*, Vol. 1273, p. 87.

Heinz, T.F. (1991) in *Nonlinear Surface Electromagnetic Phenomena* (eds H.E. Ponath and G.I. Stegemen), Elsevier Science, New York, pp. 397–398.

Iwamoto, M. and Wu, C.X. (2000), *The Physical Properties of Organic Monolayers*, World Scientific, Singapore, pp. 1–200.

Iwamoto, M., Wu, C.X. and Ou-Yang, Z.C. (2000) *Chem. Phys. Lett.*, **325**, 545–551.

Luckhurst, G.R. and Gray, G.W. (1979) *The Molecular Physics of Liquid Crystals*, Academic Press, London, pp. 51–83.

Manaka, T., Lim, E., Tamura, R. and Iwamoto, M. (2006) *Appl. Phys. Lett.*, **89**, 072113–1–072113–3.

Manaka, T., Nakao, M., Yamada, D., Lim, E. and Iwamoto, M. (2007a) *Opt. Express*, **15**, 15964–15971.

Manaka, T., Lim, E., Tamura, R. and Iwamoto, M. (2007b) *Nat. Photonics*, **1**, 581–584.

Nye, J.F. (1957) *Physical Properties of Crystals*, Oxford, pp. 68–169.

Sessler, G.M. (1987), *Electrets*, Springer-Verlag, New York, pp. 1–80.

Shen, Y.R. (1984) *The Principles of Nonlinear Optics*, John Wiley & Sons, Inc., New York, pp. 1–85.

Sze, S.M. (1981) *Physics of Semiconductor Devices*, John Wiley & Sons, Inc., New York, pp. 362–512.

Tojima, A., Manaka, T. and Iwamoto, M. (2001) *J. Chem. Phys.*, **115**, 9010–9017.

Wang, X.K., Zhang, T.G., Lin, W.P., Liu, S.Z., Wong, G.K., Kappes, M.M., Chang, R.H. and Ketterson, J.B. (1992) *Appl. Phys. Lett.*, **60**, 810–812.

Zhu, X.D., Suhr, H. and Shen, Y.R. (1987) *Phys. Rev. B*, **35**, 3047–3050.

Further Reading

Ye, H., Abu-Akeel, A., Huang, J., Katz, H.E. and Gracias, D.H. (2007) *J. Am. Chem. Soc.*, **128**, 6528–6529.

26
Infrared Molecular Vibrational Spectroscopy

Andrew B. Horn

26.1 Principles of Vibrational Spectroscopy 889
26.1.1 Introduction 889
26.1.2 Classical and Quantum-Mechanical Description of Infrared Vibrational Excitation in Diatomic Molecules 889
26.1.3 Group Frequencies and Symmetry in Infrared Vibrational Spectroscopy 893
26.1.4 Gas-Phase Spectra and Rotational Fine Structure 896
26.2 Measurement of Infrared Vibrational Spectra 897
26.2.1 Principles of Infrared Spectrometry 897
26.2.2 Instrumentation and Spectral Processing 897
26.2.2.1 Interferometry in Infrared Spectroscopy 897
26.2.2.2 Processing of Infrared Spectra 901
26.2.2.3 Infrared Detectors 902
26.2.3 Basic Transmission Sampling Methods 902
26.2.3.1 Gases 903
26.2.3.2 Liquids 903
26.2.3.3 Solids 904
26.2.4 Reflection-Based Sampling Methods 905
26.2.4.1 Specular Reflection 905
26.2.4.2 Reflection–Absorption 906
26.2.4.3 Diffuse Reflectance 912
26.2.4.4 Attenuated Total Internal Reflection 914
26.2.5 Infrared Microspectroscopy 921
26.2.6 Extinction Methods from Particles 924
26.3 Interpretation of Infrared Spectra 926
26.3.1 Empirical Interpretation from Group Frequencies 926
Acknowledgments 931
References 931
Further Reading 932

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

26.1
Principles of Vibrational Spectroscopy

26.1.1
Introduction

Infrared spectroscopy is routinely used in analytical laboratories for the identification of materials, molecules, and functional groups through their characteristic absorption patterns in the mid-infrared (4000–400 cm^{-1}) region of the electromagnetic spectrum. In its original conception, analytical infrared spectroscopy was based upon simple dispersive instrumentation measuring the transmitted proportion of light as a function of wavelength through a sample. Over the last 30 years, the development of interferometric methods for the measurement of frequency-domain spectra, sophisticated reflection-based sampling methods applicable to non-transparent materials, and computer-based processing methods has transformed the field into a widely applicable, quantitative method for materials analysis. In this chapter, the basic principles of infrared spectroscopy are outlined, along with a brief account of the measurement and interpretation of infrared spectra. In addition, some topical examples of the use of modern infrared spectroscopy in physics, chemistry, biology, and atmospheric science are also presented.

26.1.2
Classical and Quantum-Mechanical Description of Infrared Vibrational Excitation in Diatomic Molecules

In any form of absorption spectroscopy, the energy absorbed in a *transition* between energy levels is measured. In the mid-infrared region, these transitions correspond to changes in vibrational quantum number. Before treating the system in a quantum-mechanical manner, however, a simpler classical treatment is instructive in order to give a phenomenological perspective. The most basic classical approach relating vibrational transitions to visualizable molecular parameters treats a vibrating diatomic molecule as two masses m_1 and m_2 (in kilograms) connected by a flexible bond with a force constant k (newton per meter) using Hookes' Law, as shown in Figure 26.1.

When the masses are at equilibrium, $r = r_{\text{eqm}}$ and there are no forces acting. However, when the bond is elongated

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

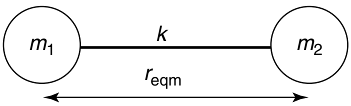

Fig. 26.1 Schematic of the basis of the simple harmonic oscillator model for a heteronuclear diatomic model. The terms are defined in the text.

(or compressed), a restoring force acts to pull (or push) the particles:

$$F(r) = -k(r - r_{\text{eqm}}) \qquad (26.1)$$

Thus, the masses oscillate around the equilibrium distance. It can be shown that the oscillation frequency is given as

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}} \quad (\text{s}^{-1}) \qquad (26.2)$$

where k is the force constant as defined above and μ is the reduced mass (in kilograms):

$$\frac{1}{\mu} = \frac{1}{m_1} + \frac{1}{m_2} \quad \text{or} \quad \mu = \frac{m_1 m_2}{m_1 + m_2} \qquad (26.3)$$

Replacing the masses and force constants with representative atomic values, the oscillation frequencies obtained are typically between 1.2×10^{13} and 1.2×10^{14} Hz.

The nature of the interaction between the electric field of the infrared radiation and the atoms of the molecule can best be envisaged by considering a heteronuclear diatomic molecule, where the atoms at either end of the bond have different electronegativities. This polarization results in a dipole moment, the magnitude of which oscillates with the changing bond length during vibrational motion. The electric field E associated with electromagnetic radiation also oscillates in amplitude with time like a dipole with frequency ν (s^{-1}):

$$E(t) = E_0 \cos(2\pi\nu t) \qquad (26.4)$$

When the frequency of oscillation of the molecule is resonant with the frequency of the infrared radiation, absorption of energy by the molecule can occur. In practice, the requirement for a permanent dipole moment exists only for diatomic molecules and for more complicated polyatomic molecules; several atoms (connected by a number of bonds) may move at once, resulting in a changing (*dynamic*) dipole moment during the course of the vibration even if a static dipole moment does not exist. Figure 26.2 illustrates this for the antisymmetric stretching mode of CO_2. This phenomenon is often expressed as an empirical selection rule for the infrared activity of a given vibration, that is, during the course of a particular vibration the molecular dipole moment must change.

From a quantum-mechanical perspective, the problem of quantifying vibrational absorption activity and of calculating energy levels and intensities relies upon the identification of appropriate operators and wavefunctions to describe

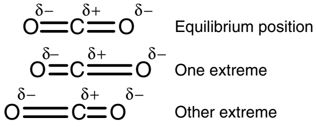

Fig. 26.2 Schematic of the atomic displacements contributing to the antisymmetric stretching mode of CO_2.

the system. Naturally, each energy level of the quantized vibrational motion will be described by a wavefunction, which must be an eigenfunction equation of the Schrödinger equation:

$$\hat{H}\Psi = E\Psi \tag{26.5}$$

The Hamiltonian has both kinetic and potential energy components describing the vibrational motion. A simple treatment that serves to extend the classical treatment into the quantum world treats the potential energy as deriving from Hookes' law, Eq. (26.1). The potential energy operator takes the following form:

$$\hat{V} = \frac{1}{2}kx^2 \tag{26.6}$$

where $x = r - r_{\text{eqm}}$. The kinetic energy operator can be taken directly from the known result:

$$\hat{T} = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} \tag{26.7}$$

Combining these equations together and using the fact that $m = \mu$ for diatomics, we obtain

$$\hat{H}\Psi = -\frac{\hbar^2}{2\mu}\frac{\partial^2\Psi}{\partial x^2} + \frac{1}{2}kx^2\Psi \tag{26.8}$$

This treatment is usually referred to as the *simple harmonic oscillator,* and suitable wavefunctions take the following form:

$$\Psi_{\text{v}} = \sqrt{\left(\frac{1}{2^{\text{v}}v!\sqrt{\pi}}\right)} \times H_{\text{v}}(y) \times \exp\left(\frac{-y^2}{2}\right) \tag{26.9}$$

where v is the *vibrational quantum number* and $H_v(y)$ are called *Hermite polynomials,*

v	$H_v(y)$
0	1
1	$2y$
2	$4y^2 - 2$
3	$8y^3 - 14y$

where y is given as

$$y = \left(\frac{4\pi^2\nu\mu}{h}\right)^{1/2} \times (r - r_{\text{eqm}}) \tag{26.10}$$

Transitions between these energy levels then correspond to vibrational absorption. Solving Eq. (26.8) with these eigenfunctions results in the well-known value for the energy of a given level v:

$$E = h\nu\left(v + \frac{1}{2}\right) \tag{26.11}$$

where ν is the classical frequency given in Eq. (26.2). Also note that these wavefunctions are normalized and orthogonal. Several interesting quantum-mechanical phenomena appear from this treatment. First, the probability of finding a particle at a position x is given by $|\Psi_v(x)|^2$, suggesting that in the ground state ($v = 0$), the oscillator spends most of its time at $x = 0$ (i.e., at the center), directly at odds with classical theory. Second, this also points to the existence of zero-point energy since the lowest state of the oscillator has energy $1/2h\nu$.

The quantum-mechanical treatment gives direct access to the magnitude of the spacing between the energy levels and also the selection rules, since the vibrational transition probability can be derived directly from the wavefunctions. The gross selection rule for the classical treatment given earlier still applies: the electric dipole moment must change during the course of a vibration. This is reinterpreted quantum mechanically through the values of the *transition dipole moment* μ_{fi} on going from an initial state Ψ_i to final state Ψ_f, that is, for an allowed transition, $\mu_{fi} \neq 0$. Clearly,

$$\mu_{fi} = \int_{-\infty}^{+\infty} \Psi_f^* \hat{\mu}_i . d\tau = \langle f|\hat{\mu}|i\rangle \quad (26.12)$$

where the electric dipole moment operator $\hat{\mu}$ has the form

$$\hat{\mu} = \sum_j \mathrm{q}_j \mathbf{r}_j \quad (26.13)$$

where q_j is the charge on the jth particle and $\mathbf{r}_j$ is its position vector with respect to the center of mass of the molecule. Evaluation of these parameters for the wavefunctions of the simple harmonic oscillator leads to the selection rule $\Delta v = \pm 1$. In real molecules, the vibrational potential $V(x)$ is not accurately described by a simple harmonic oscillator: the real potential is anharmonic, resembling a *Morse potential*. The energy levels of the anharmonic oscillator are given by

$$E_\mathrm{v} = h\nu\left(\mathrm{v} + \frac{1}{2}\right) - x_e\left(\mathrm{v} + \frac{1}{2}\right)^2 \quad (26.14)$$

where x_e is called the *anharmonicity constant*. Note that the energy levels become more closely spaced as v increases. This allows transitions across quantum numbers and in reality the selection rule is found to be $\Delta v = \pm 1, \pm 2, \pm 3$, and so on. The first of these transitions is referred to as the *fundamental*, with the higher values being the *harmonics* or *overtones*. Furthermore, the reduction in spacing between the energy levels in the anharmonic oscillator leads to the appearance of the harmonics at a slightly lower frequency than twice that of the fundamental. Nothing in the foregoing treatment forbids excitation from energy levels above the ground state, giving rise to what are know as *hot bands*. The name derives from the fact that the populations of the higher vibrational states $N(v)$ follow the Boltzmann distribution:

$$N(\mathrm{v}) = N_0 \exp\left(-\frac{\mathrm{v}h\nu}{kT}\right) \quad (26.15)$$

where N_0 is the ground state population and T is the temperature (kelvin). Given the typical energy spacing of vibrational levels, the population of the first vibrationally excited state is small at room temperature for all but the lowest frequency vibrations and significant populations are only achieved in hot samples. Again, the reduction in spacing between the energy levels in the anharmonic oscillator leads to the appearance of the hot band at a slightly lower frequency that the fundamental.

The intensity I of absorption upon transition from one energy level to another is proportional to the square of the transition dipole moment:

$$I \propto (\mu_\mathrm{fi})^2 \quad (26.16)$$

In infrared transitions, this value depends upon both the amplitude of the vibration (since large excursions give rise to large changes in the relative positions of the atoms) and the relative polarization of the atom(s) at either end of the chemical bond. For example, as will be seen later, vibrations involving oxygen atoms tend to be intense because of the electronegativity of oxygen, and many vibrations involving hydrogen atoms tend to be intense because small, light atoms often experience large displacements. For further details, see Atkins and de Paula (2009) and Hollas (1992).

26.1.3
Group Frequencies and Symmetry in Infrared Vibrational Spectroscopy

In a simple diatomic molecule, vibration involves only the motion of two atoms. Any movement of one atom affects the atom, and since the center of mass must be conserved only one vibration can occur. This is not the case for polyatomic molecules, for which several combinations of atomic displacement are possible. These combinations are not random, since there are constraints in the retention of a fixed center of mass. For a free atom, there are three degrees of freedom, represented by translation in the x, y, and z directions. For two atoms connected together (taking the z axis as the direction between the two atom centers), a total of six degrees of freedom must therefore exist: three translations of the whole molecule (along x, y, and z axes), two rotations (around x and y axes), and one vibration (along the z axis). By extension of this argument, it can easily be seen that for a polyatomic molecule with N component atoms there must be $3N$ degrees of freedom, from which there must be three translations of the whole molecule (along x, y, and z axes) and three rotations (around the x, y, and z axes). This leaves $3N - 6$ degrees of freedom, which must therefore be accounted for by a series of (different) vibrational motions. It should also be noted, however, that if the molecule is linear, there can only be two rotations; hence, there will be $3N - 5$ vibrations in this case.

Each vibration consists of a number of atomic displacements, since it is impossible for only one bond to stretch without affecting all the other atoms and the center of mass of the molecule must remain stationary (otherwise the displacements lead to translational motion or rotation). This is illustrated for H_2O in Figure 26.3. Since there are three atoms in water, there are three vibrational combinations. The simplest motion involves a displacement of each of the H atoms simultaneously and in phase along their respective O—H bonds. To compensate for the motion of the H atoms, it is necessary for the O atom to move in the opposite direction. To retain the center of mass, the vector distance/displacement products must sum to zero; hence, the O atom will move much less than the H atoms for this to occur. This vibration is referred to as the *symmetric stretch*, for reasons that are discussed later. The second identifiable vibrational combination again involves simultaneous displacement of the H atoms along the O—H bond, but this time out of phase. To compensate for this overall motion of the H atoms to the right, the O atom must move proportionally to the left. This vibration is the *antisymmetric stretch*. The remaining vibration involves a different kind of motion: *bending*. In this case, the O atom must move proportionally upward to counteract the motion of the H atoms.

A unique combination of atomic displacements can be found for each vibrational degree of freedom of any

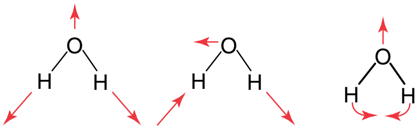

Fig. 26.3 Atomic displacements for the three vibrational modes of H_2O. Note that any allowed combination of displacements must retain the center of mass.

molecule. Each unique combination is referred to as a *normal mode*. As is shown later, some combinations may result in displacements that only differ in the directions in which they occur (e.g., in the x, y plane or the x, z plane) but which occur at the same frequency: such modes are described as *degenerate*. For example, there are three identifiable components of the antisymmetric stretching modes of methane (CH_4), polarized along the x, y, and z axes: all these have the same frequency and are therefore triply degenerate.

In general, for a given normal mode in polyatomic molecules, the individual combination of atomic displacements falls into one of two broad categories: those that are mainly localized on a small subset of atoms and those that involve more complex coupled motion of a significant proportion of the atoms. The former are often readily identifiable with various functional groups of the molecule, the so-called *functional group modes*. A particularly useful example is the stretching of the carbonyl ($>C{=}O$) bond. In many molecules, the vibration of this group is largely decoupled from the rest of the molecule and the mode, therefore, absorbs at roughly the same frequency in a wide range of molecule types. However, functional groups with similar absorption frequencies can couple, a clear example being the functional group modes related to the various components of a secondary amide (containing the $-CO-NH-$functionality). In this case, the carbonyl frequency is close to that of the bending vibration of the NH, and hence strong coupling occurs to give the characteristic *amide I* and *amide II* functional group modes. The former is predominantly, but not exclusively, related to the carbonyl and the latter is related to the NH bending mode. Together, the two bands form a uniquely identifiable motif that strongly features in the infrared spectra of biological materials.

At the other end of the scale, there are normal modes that involve the simultaneous displacement of many atoms within a given molecule. Often, it is very difficult to uniquely assign these modes. However, the combined overlapping collection of these normal modes is a unique characteristic for a given molecule and is responsible for its "fingerprint." Spectra of polyatomic molecules are, therefore, comprised of a complex, unassignable fingerprint spectrum overlaid with more readily identifiable functional group modes.

For simpler polyatomic molecules, the collective displacements of the individual atoms can be described by considering the symmetry properties of the molecule. This leads to a simpler way of determining the vibrational activity of a given normal mode. The first step in this process is the determination of the symmetry properties of the molecule, following the principles of group theory, described in detail elsewhere. This basically involves identifying the collection of symmetry operations (rotations, reflections, improper rotations, centers of inversion, etc.), which can be performed upon the molecule, from which the point group can be determined. For a detailed description of the method, see Vincent (2001).

Once the point group of the molecule in question has been determined, the symmetries of the combinations of atomic displacements for each normal mode can be derived. Three Cartesian displacement vectors are placed on each atom (one for each x, y, and z

component). The effect of the symmetry operations of the point group on these collections of displacement vectors is then determined. The resulting combination is a *reducible representation,* which can then be resolved into a combination of *irreducible representations* by a simple mathematical procedure. Each irreducible representation describes a unique set of transformations of the molecule under each of the operations of the point group, each operation identifying, for example, whether or not a displacement along a given bond within the molecule remains in phase or out of phase when transferred to another (identical) bond in the same molecule. Consequently, the symmetries of each of the normal modes are determined.

In the example below, the symmetries of the normal modes of iodomethane (CH_3I) are determined. The molecule has C_{3v} point group symmetry, the operations of which are E (the identity), two C_3 (threefold) rotation axes, and three σ_v (vertical, containing the principal axis) mirror planes. Defining the principal axis as z, collinear with the C_3 axes, Cartesian displacement vectors are placed on each atom. The effect of each operation of the point group on the full set of Cartesian vectors is then determined, that is, the transformation of each of the x, y, and z vectors is calculated for each atom under the operations of the point group. The sum of these values under each class of operation in the point group then gives the reducible representation, as shown below:

C_{3v}	E	$2C_3$	$3\sigma_v$
χ(15 x, y, z vectors)	15	0	3

In this example, this reducible representation reduces to irreducible representations of $4A_1 + 1A_2 + 5E$ using conventional group theory methods. However, the inclusion of all three degrees of freedom for each atom means that this set of symmetry values accounts for the translations and rotations of the molecule as well as the vibrational normal modes. These nonvibrational degrees of freedom are eliminated by reference to the character table, which for C_{3v} reveals that the three translational degrees of freedom have $A_1 + E$ symmetry, while the three rotational degrees of freedom have $A_2 + E$ symmetry:

Removing these reveals the symmetries of the vibrational degrees of freedom (the normal modes) to be $3A_1 + 3E$. A final check of the numbers shows that this accounts for a total of nine normal modes (E symmetries are doubly degenerate, accounting for two modes simultaneously).

Once the symmetries of the normal modes have been established, the activity of the normal modes in the infrared spectrum can be determined. As described previously, not all modes will give rise to

C_{3v}	E	$2C_3$	$3\sigma_v$	Rotations and translations	Cartesian axes and their products
A_1	1	1	1	T_z	z; $x^2 + y^2$; z^2
A_2	1	1	−1	R_z	
E	2	−1	0	(T_x,T_y); (R_x,R_y)	(x,y); (xz,yz); $(x^2 - y^2)/\sqrt{2}$

an appropriate dynamic dipole change. For energy levels characterized by wavefunctions Ψ_i and Ψ_f, a nonzero transition dipole moment will only arise if the expectation value of the dipole moment operator for the transition is nonzero, that is,

$$\mu_{fi} = \int_{-\infty}^{+\infty} \Psi_f^* \hat{\mu} \Psi_i . \mathrm{d}\tau \neq 0 \qquad (26.17)$$

For this to be the case, the symmetry product of the wavefunctions and the operator should contain the totally symmetric irreducible representation in the point group of the molecule, that is,

$$\Gamma(\Psi_f) \times \Gamma(\mu) \times \Gamma(\Psi_i) \supset A, A_1, A', A_g, \cdots \qquad (26.18)$$

There are a range of procedures for this, which rely upon a knowledge of the symmetries of the various components. For $v = 0{-}1$ transitions (fundamental), the ground state wavefunction Ψ_i is totally symmetric, for example, A_1 in C_{3v}. The final state wavefunction Ψ_f has the symmetry of one of the normal modes, that is, taken from the set of symmetries determined in the example above. The dipole moment operator has the same symmetry as the x, y, and z axes in the point group of the molecule, for example, $A_1 + E$ in C_{3v}, as shown in the character table.

The problem generally simplifies further, since multiplying any symmetry species by the totally symmetric irreducible representation has no effect, for example, in C_{3v}, $A_1 \times E = E$. Consequently, it is only necessary to consider the symmetry of the normal mode and of the dipole operator, that is, to find out if $\Gamma(\Psi_f) \times \Gamma(\mu) \supset A, A_1, A', A_g$, and so on. The process of identifying whether a vibrational transition is infrared active therefore involves determining the molecular point group and identifying the symmetries of the normal modes and of the x, y, and z axes. Where the vibrational normal mode symmetry is the same as one of the axes, the vibration is infrared active. In the iodomethane example, all of the vibrational normal modes are infrared active.[1] Further information on these procedures can be found in Atkins and de Paula (2009) and Vincent (2001).

26.1.4
Gas-Phase Spectra and Rotational Fine Structure

The foregoing discussion has dealt entirely with transitions between vibrational energy levels. However, transitions between rotational levels can also occur simultaneously. As described elsewhere, rotational energy levels typically have spacing of the order of a few cm^{-1}; hence, in cases where natural linewidths are narrow (such as in low-pressure gases of simple molecules), rotational fine structure can be observed. In infrared absorption spectra, this structure is observed as a series of fine lines corresponding to a change of +1 in the vibrational quantum number and ±1 in the rotational quantum number. The intensity and extent of this rotational fine structure depend upon the population of the rotational energy levels in the vibrational ground state (the spacing of the rotational levels is small; hence, the Boltzmann distribution of population in the rotational energy levels typically extends to $J = 15$ or more). A fuller description is given in Hollas (1992).

---

[1] The procedure also applies to the determination of normal mode activity in vibrational Raman spectroscopy, except that the comparison is made between the symmetries of the normal modes and those of the polarizability operator α, which corresponds to the symmetries of the products of any two Cartesian axes. These are also normally shown in character tables. For molecules possessing a center of inversion, infrared and Raman activity are mutually exclusive.

26.2
Measurement of Infrared Vibrational Spectra

26.2.1
Principles of Infrared Spectrometry

The measurement of absorption spectra requires a light source emitting the appropriate range of frequencies, a dispersing element to separate the different frequencies of light, and a detector that responds quickly over the appropriate range of frequencies. At any particular frequency, the incident light intensity I_0 and the transmitted light intensity I are related to the amount of absorption by the sample via the Beer–Lambert law:

$$-\log_{10}\left(\frac{I}{I_0}\right) = \varepsilon c \ell \tag{26.19}$$

where ε is the molar absorption coefficient (meter square per mole), c is the concentration (mole per cubic meter), and ℓ is the path length (meter). The dimensionless logarithm term on the right-hand side (RHS) of Eq. (26.8) is the *absorbance*, A, a dimensionless quantity. Infrared spectra are usually presented as a graph of absorbance versus wavenumber.

As an alternative to absorbance, spectra can be presented as percentage *transmittance*, $\%T$.

$$\%T = \frac{I_0 - I}{I_0} \times 100 \tag{26.20}$$

Clearly, a transmittance of 100% represents no absorption, while a transmission of 10% represents an absorbance of 1.0. This measurement arose historically because transmittance is easier to measure experimentally in dispersive (diffraction grating) instruments. However, absorbance is linear in concentration and path length and is generally preferred for quantitative work.

26.2.2
Instrumentation and Spectral Processing

The basic requirement for benchtop infrared spectroscopy comprises a polychromatic infrared light source, a dispersion method to separate out the individual frequency components, and a suitable detector. In most instruments, infrared light is generated with a blackbody source optimized for output in the mid-infrared. These are generally electrically heated, low-current incandescent devices. Output characteristics of various sources in common use for mid- and near-infrared spectroscopy are reviewed in detail by Buijs (2002).

26.2.2.1 Interferometry in Infrared Spectroscopy

Originally, infrared spectrometers utilized rock salt prisms to disperse the frequency components, eventually moving to diffraction gratings for more flexible operation and improved longevity of the components (Workman, 1998). As for all grating-based monochromators, there are significant trade-offs between the frequency bandpass (a function of the slit widths used at the input and output of the monochromators) and the optical throughput. As a consequence, the measurement of high-resolution infrared spectra (of gases, for instance) was

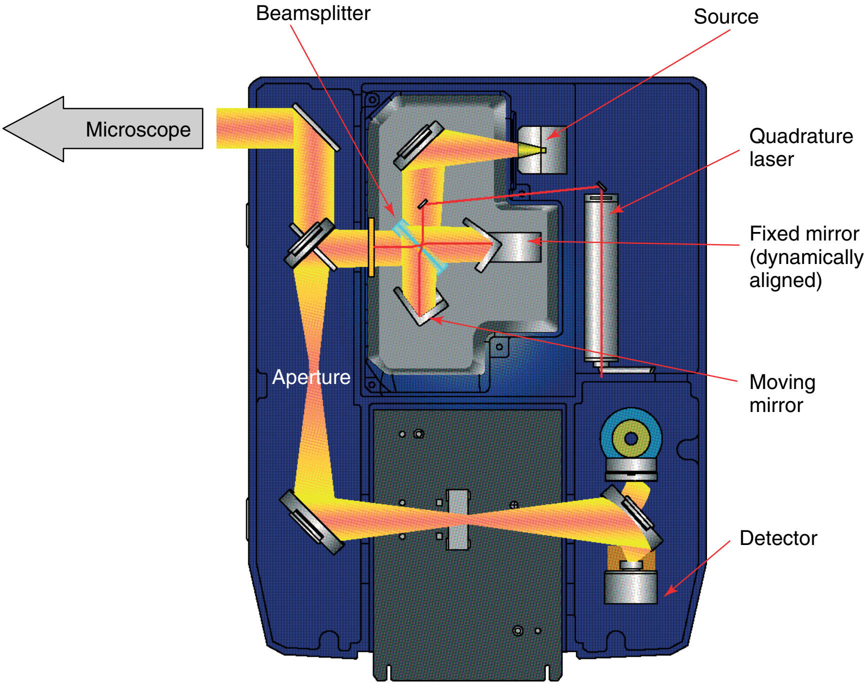

Fig. 26.4 Schematic representation of a Michelson interferometer in a commercial FTIR spectrometer. Reproduced by kind permission of JASCO UK.

a considerable experimental challenge. However, with the advent of improved computational facilities in the 1970s, methods based upon interferometry became the norm. The basic principles of interferometry as applied to infrared spectroscopy involve the separation of the source radiation into two beams using an appropriate beamsplitter. These beams then pass along separate light paths before being recombined (usually at the same beamsplitter). Any path difference experienced by the two beams leads to interference upon recombination, with different frequency components combining with different phases. The superposition of these interferences is recorded at discrete (normally equally spaced) intervals of path difference δ to give a path-difference-dependent signal referred to as an *interferogram*. The most common apparatus used to perform this separation and recombination is the Michelson interferometer, shown schematically in Figure 26.4, although other possibilities are also in widespread use. In many modern spectrometers, the fixed mirror is dynamically aligned with the moving mirror for improved stability and precision. This is normally affected using piezo-electric transducers. The output interferogram, $I(\delta)$, is measured using a detector operating over the appropriate frequency range. After digitization, the interferogram is inverted using a discrete Fourier transform to produce a frequency-domain spectrum, $I(\nu)$, and hence the common acronym *FTIR* (*Fourier-transform infrared*) spectrometry. The spectral resolution (i.e., the spacing between the data points) in the computed spectrum is inversely related to the maximum path difference $\delta_{\max}$ in the interferometer. Typical general purpose benchtop FTIR spectrometers with a maximum resolution of 0.5 cm^{-1} therefore require path differences of 2 cm, which is easily attainable with relatively unsophisticated apparatus. The path difference between the two optical paths is usually measured to an appropriate precision by interference from two orthogonally polarized HeNe laser beams by quadrature. Increased resolution requires very high mechanical stability over path lengths that can be several meters. Further experimental details can be found in sources such as Griffiths and de Haseth (1986) and Jackson (2002).

The acquisition of intensity data in the temporal domain and its transformation into the spectral domain brings a number of interesting features and artifacts that need to be understood and in many cases incorporated into the post-acquisition processing of the resultant data. A typical interferogram is shown in Figure 25.5(a). The mechanical, discrete nature of the temporal domain data acquisition leads to a range of effects. First, the path differences in the two optical paths are usually created by movement (either translation or rotation) of an optical component (usually a mirror) from a path difference of zero to $\delta_{\max}$ at a certain mirror velocity The detector, therefore, observes a signal that contains interference components modulated over a range, with the highest frequency component being modulated at the highest rate. The detector, therefore, needs to be capable of responding to changing signal intensity at this maximum rate. Furthermore, the sampling of this highest frequency component needs to be made at least twice its frequency (according to the Nyquist criterion). This requires suitably optimized amplification and digitization electronics. The collected data also contain a further artifact of scanning δ with a moving optical component: the data are usually "chirped" in the direction of the movement. This effect can be removed by a process called *phase correction*. In practice, interferograms are often sampled double-sided (i.e., from $-\delta_{\max}$ to $+\delta_{\max}$) and in both forward and backward directions, with the information required for effective phase correction being obtained from the asymmetry of the interferogram either side of the zero path difference point (Griffiths and de Haseth, 1986).

The second substantial effect of the acquisition of temporal domain data over a limited range of δ occurs upon Fourier transformation of the data to the temporal domain. Collection of data from $\delta = 0$ to $\delta_{\max}$ is effectively the equivalent of multiplying the full, infinite set of temporal domain data (from $\delta = 0$ to ∞) by a boxcar function (i.e., unity from $\delta = 0$ to $\delta_{\max}$ and 0 beyond $\delta_{\max}$). Consequently, the lineshape observed in the spectral domain is a convolution of the Fourier transform of the temporal data with the Fourier transform of a boxcar function. For example, a pure sine wave signal would transform to a delta function from infinite data, whereas in practice a truncated sine wave (from $\delta = 0$ to $\delta_{\max}$) transforms to a spectral feature with both a finite bandwidth and with additional side-lobes. This effect can be significant for spectral data containing sharp, narrow features (especially in gas-phase spectra). It can be significantly reduced by a process referred to as *apodization*, where the recorded temporal data are

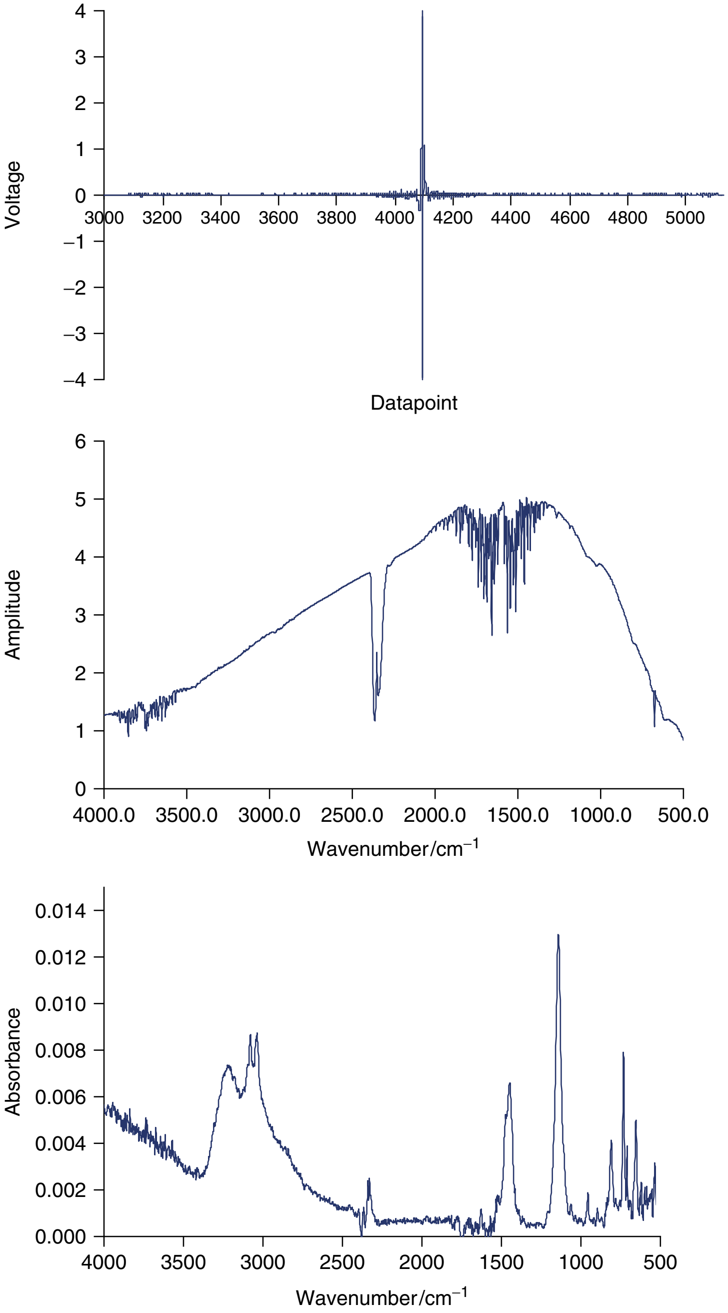

Fig. 26.5 (a) Example of a typical interferogram; (b) single-beam spectrum resulting from the Fourier transform; and (c) absorbance spectrum resulting from the ratio of two single-beam spectra, one containing background information only and one containing additional spectral information.

multiplied by a specific function with a less intrusive spectral domain lineshape. These and other FT-related artifacts are described in more detail in Griffiths and de Haseth (1986) and Kauppinen and Partanen (2001).

A final processing technique, which is frequently encountered, is zero-filling, in which the temporal interferogram data are extended by the addition of further data points (with zero intensity) beyond $\delta_{\max}$. This has the effect of placing additional spectral domain data points between the real values. The additional values are interpolations between the real values, smoothing the data, although they clearly do not add any additional information.

In practice, Fourier-transform-based methods have three main advantages over conventional, dispersive methods. First, the data points are measured temporally with a high degree of accuracy; therefore, spectral accuracy (frequency precision) is high in the computed spectrum. Furthermore, the throughput of light is not limited by the presence of slits as for monochromator-based methods. This leads to much higher working signal levels at the detector, with improved signal-to-noise (S/N) ratios. S/N ratios are also improved by the multiplex advantage, since all frequency components are effectively sampled together. In general, the S/N in FTIR spectra is noise limited and can be improved by the coaddition of interferograms before processing. Typically, the S/N improvement scales with the square root of the number of scans.

26.2.2.2 **Processing of Infrared Spectra**

As described above, FTIR methods therefore involve collection of data as interferograms and subsequent processing to produce a single-beam spectrum, an example of which is shown in Figure 26.5(b). During the course of this processing, a number of mathematical procedures can be performed in order to improve the visual appearance of the output, remove artifacts, and optimize the spectrum. Following the Fourier transformation of the interferogram to produce a single-beam spectrum, the simplest operation is the ratioing of two single-beam spectra to produce an absorbance spectrum. This is necessary because the single-beam operation of an interferometer-based spectrometer produces spectra that are a convolution of the source output profile, the detector response profile, and any limiting optical transmission effects from components in the light path (such as beamsplitters, mirrors, lenses, and sample-handling apparatus). Figure 26.5(b) shows this overall response profile with absorptions from CO_2 and water vapor within the optical path. The absorbance (or transmittance) spectrum of a given sample is obtained from the ratio of single-beam spectra obtained in the presence and absence of a sample. This operation is performed in software with various interventions including the reduction or elimination of water vapor and CO_2 spectral features. Figure 26.5(c) shows a typical absorbance spectrum obtained in such a manner.

A variety of more sophisticated postprocessing methods are commonly available in the software accompanying commercial benchtop FTIR spectrometers for further processing of the absorbance (transmittance) spectra thus obtained. This typically ranges from basic mathematical processing such as spectral subtraction/addition and smoothing to more advanced methods such as Fourier deconvolution and spectral bandshape fitting. Databases of infrared spectra are also widely available, and algorithmic methods for comparison of spectra with these library references are used to suggest a range of possibilities and calculated goodness of fit. At a more sophisticated level, chemometric software is also increasingly available for sophisticated quantitative analysis. The specific details of these methods and packages vary widely between manufacturers. Details of these and other related methods can be found in the ''Quantitative Analysis'' section of (Chalmers and Griffiths, 2002, vol. 3).

26.2.2.3 Infrared Detectors

The detection of infrared radiation in interferometric applications requires apparatus that can detect photons over the relevant mid-infrared range and which is also capable of registering changes in the intensity of such radiation at the maximum modulation frequency of the output of the interferometer (Theocharous and Birch, 2002). In conventional FTIR spectrometers, this is normally achieved using either a bolometer or a photon detector. The commonest example of the former used deuterated triglycine sulfate (DTGS) as the active medium. Thermal changes in the detector upon incidence of infrared radiation are measured electrically and amplified. The response of this material is such that only relatively slow modulation frequencies (a few kilohertz) are possible, but the main advantage is that the linearity of such detectors is high. For this reason, they are the most common choice for quantitative work.

For more sensitive detection of lower levels of infrared radiation, semiconductor-based photon detectors are also in common use. These rely upon the presence of an appropriate bandgap in the detection material, with detection in one form or another of the electrons excited by interaction with the incident infrared photons. To prevent thermal excitations, it is necessary to cool the materials, usually by placing them in thermal contact with a liquid nitrogen reservoir. Although highly sensitive, such detectors have an inherently nonlinear response and care should be taken in quantitative work. The main usage of such detectors is in applications where low light levels occur, especially in sampling accessories with low light throughput. Photon detectors respond much more rapidly to changes in infrared intensity; hence, considerably higher modulation frequencies are accessible. This permits much higher scan rates within the interferometer. The most common material used in modern detectors is mercury cadmium telluride (MCT). There is a trade-off between the range of frequency response of MCT detectors and the sensitivity; with careful bandgap tuning, ''wide band'' MCT detectors can be made to operate across most of the mid-infrared range, although detectivity falls markedly and cost increases. For specific applications in a restricted range within the mid-infrared, other materials such as indium antimonide (InSb) are also available.

26.2.3
Basic Transmission Sampling Methods

In its simplest form, infrared spectroscopy relies upon the preparation of samples through which the infrared beam from a spectrometer can be passed with minimal loss. In some cases, this requires the use of accessories for the presentation of samples to the infrared beam (Belton and Wilson, 1990). For appropriately tractable samples, infrared spectra can be obtained by direct transmission of light. Typical values of the molar absorption coefficient ε require the use of either low concentrations or short path lengths to keep measured absorbance values in the linear region (below ca. 1.2). In practice, an appropriate amount of material is normally supported between two accurately spaced optical windows. The choice of appropriate window material largely depends on application, but some of the most common are shown in Table 26.1, along with their range of practical transmission and also some limiting physical properties.

26.2.3.1 Gases

Apparatus for the acquisition of the infrared spectra of gases normally involves path lengths of 10 cm and upward. High-resolution infrared spectra of gases are often recorded at reduced pressure to reduce the effects of collision-induced (pressure) line broadening, requiring the use of extended path lengths to attain appropriate individual signal levels. A schematic of a typical cell is shown in Figure 26.6. Multiple passes are achieved using White multipass optics, with commercial accessories available with optical path lengths up to 20 m.

26.2.3.2 Liquids

Thin films are often necessary for the acquisition of transmission infrared spectra from liquids. For pure liquids, cells with accurately spaced windows are widely available, with path lengths of the order of a few microns upward. Spectra can also be obtained from solutions, again with an appropriate choice of path length. The contrasting need to avoid solvent absorptions blacking out (reaching absorbance values greater than 1.2) while achieving suitable signals

Tab. 26.1 Properties of infrared transparent materials in common usage.

Material	Transmission range/cm^{-1}	Properties
NaCl	40,000–625	Very soft, easily fogged by water vapour; low cost
KBr	40,000–400	Very soft, easily fogged by water vapour; low cost.
ZnSe	17,000–720	Hard and brittle crystal, pale yellow in colour; sometimes known as Irtran-1.
KRS-5	20,000–250	Very soft, deforms under pressure; soluble in bases and insoluble in acids. Toxic.
CaF_2	77,000–1,110	Strong, acid and alkali resistant; withstands high pressure; insoluble in water. Poor range.
Si	8,300–660	Very hard but quite brittle; relatively inert; insoluble in water.
Ge	5,500–600	Very hard and brittle with a high refractive index (4); insoluble in water.
Diamond	25,000–33	Very hard, resistant to most types of chemicals. Phonon bands around 1900–2600.

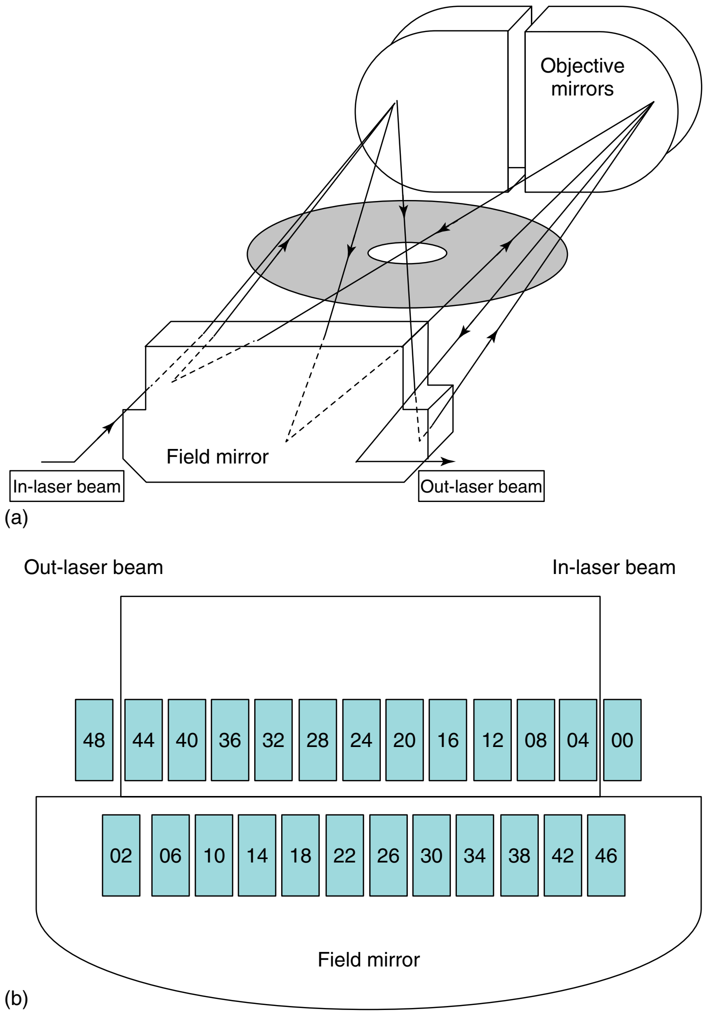
Objective mirrors
Field mirror
In-laser beam
Out-laser beam
(a)
Out-laser beam
In-laser beam
48
44
40
36
32
28
24
20
16
12
08
04
00
02
06
10
14
18
22
26
30
34
38
42
46
Field mirror
(b)

Fig. 26.6 (a) Schematic of the optical configuration of a multipass long path cell for recording infrared spectra of gases. The number of passes can be altered by movement of the objective mirrors. (b) Sequence of images on the field mirror for a 48 pass cell. Every fourth pass is imaged on the top row and the position of every fourth image translates from the entrance (00) to the exit cut-out (48). Taken from (Najera *et al.*, 2009).

from dilute solutes requires a degree of flexibility.

26.2.3.3 Solids

Infrared spectra of solid materials are accessible in a variety of ways, usually obtained after grinding the relevant solid to a fine powder and suspending the particles thus obtained in a suitable medium. Hydrocarbon or halocarbon oils were traditionally used for this purpose, the most common being Nujol.

The resulting oily suspension is then placed between two transparent windows and mounted in a holder in the sample compartment of a spectrometer. Blank windows are normally used as a background in order to nullify the effects of scattering and reflection from imperfect windows. The main intrinsic absorption bands in Nujol arise from aliphatic CH_2 and CH_3 features and thus the material is of little use for detailed examination of organic material containing such functionality, but useful spectral windows are to be found between these resonances. Halocarbon alternatives to Nujol are frequently used to access the CH stretching regions, although these materials do have absorptions in other regions.

A more practical alternative in many organic applications are pressed discs of KBr. In this method, the sample is ground to a fine powder and mixed with dry, powdered KBr and compressed in a die with a hydraulic press. The particles of KBr flow under such pressure and a semitransparent disc is obtained. With careful control of amounts of KBr and sample, this method can be used quantitatively. A significant advantage over mulls is the absence of absorption bands due to the suspending agent, with the principal problem being the hygroscopicity of KBr, which can lead to significant water uptake unless due care with handling and storage of the materials and discs is taken. Another potential problem with this method when applied to inorganic and organic salts is the possibility of ion exchange (e.g., Br^-) between the matrix material and the sample. Care must also be exercised when studying organic solids that exhibit polymorphism, as the pressure required to form KBr discs may also be sufficient to drive a polymorphic phase transition, with a concomitant effect of the infrared spectra of the material. These and other sampling methods in common use are reviewed in detail in Coates (1998), Belton and Wilson (1990), and Hannah (2002).

26.2.4
Reflection-Based Sampling Methods

Reflection methods are finding increasing application in physical sciences as a way of examining materials with a minimum of sample preparation. Using modern sampling accessories, it is possible to acquire infrared spectra from internal, external, and diffusely reflected infrared radiation, although there are many diverse and unusual perturbations to the spectra in reflection configurations that need careful consideration.

26.2.4.1 Specular Reflection

Specular reflection refers to reflection from relatively flat materials such that the angles of incident and reflected beams are the same with respect to the surface normal. The optics of light reflection and transmission from interfaces between both absorbing and nonabsorbing media alike are well understood and accurately described by Maxwell's equations or by a Fresnel equation formalism (Heavens, 1955; McIntyre and Aspnes, 1971). Generally, the reflection behavior is different for light polarized perpendicular and parallel to the plane of incidence (defined as the plane that includes the incident electromagnetic beam and the surface normal), referred to as s- and p-polarization, respectively. In general, the optical properties of each medium are described by their complex refractive index,

$\mathbf{n}$, or the dielectric constant, $\boldsymbol{\varepsilon}$, that is,

$$\mathbf{n} = n - ik \tag{26.21}$$

where n is the refractive index and k is the absorption coefficient. When $k = 0$, the refractive index is real in all cases. Similarly, the dielectric constant has real and imaginary components:

$$\boldsymbol{\varepsilon} = \varepsilon' - i\varepsilon'' \tag{26.22}$$

The two are related.

$$\varepsilon' = (n^2 - k^2)$$
$$\varepsilon'' = 2nk \tag{26.23}$$

For the interface between two phases i and j, the Fresnel coefficients for s- and p-polarization are given as

$$\mathrm{r}_{ij}^{s} = \frac{\mu_j \xi_i - \mu_i \xi_j}{\mu_j \xi_i + \mu_i \xi_j}, \quad \mathrm{r}_{jk}^{p} = \frac{\varepsilon_j \xi_i - \varepsilon_i \xi_j}{\varepsilon_j \xi_i + \varepsilon_i \xi_j} \tag{26.24}$$

where μ_j is the magnetic permeability (ca. 1 in the mid-infrared) for phase j. The function ξ_j for each interface is effectively a complex angle of incidence/refraction term on the interface $(j-1)|j$.

$$\xi_j = \sqrt{\mathbf{n}_j^2 - \mathbf{n}_1^2 \sin^2 \theta_1} \tag{26.25}$$

where θ_1 is the angle of incidence of the light upon the first layer and $\mathbf{n}_j$ is the complex refractive index of the phase j (for multilayer systems, discussed later). The substrate reflectance can be determined as a function of θ using Eqs (26.4) and (26.5). Figure 26.7 shows a comparison of the reflectivity of a typical nonabsorbing dielectric, silicon ($n_{\mathrm{Si}} = 3.44$, $k_{\mathrm{Si}} = 0$), and a metal, platinum ($n_{\mathrm{Pt}} = 3.0$, $k_{\mathrm{Pt}} = 30.0$) (Horn, 1998).

There are large differences between these examples in several ways, which affect the nature of the optical configuration required to acquire spectra from the interfacial region. Most noticeably, the reflectivity of silicon is lower than for a metal (except at 90°) for both polarizations, starting from around 35% at normal incidence. The s-polarized reflectivity rises monotonically from its normal incidence value to 100% at 90° (i.e., for a given angle, less light will be reflected from dielectrics than from metals), whereas the p-polarized reflectivity of silicon falls to zero at the polarizing or Brewster angle θ_B, where

$$\theta_B = \tan^{-1}\left(\frac{n_{\mathrm{Si}}}{n_1}\right) \tag{26.26}$$

When a substrate has frequency-dependent optical constants (i.e., absorption bands), this behavior can be used to collect reflectance spectra. The reflectivity of the substrate is modified across the absorption band by absorption and by dispersion due to the changing refractive index. This gives rise to characteristic lineshapes in reflectance spectra, an example of which is shown in Figure 26.8. The frequency-dependent n and k behavior can be separated out mathematically using a Kramers–Kroenig transform, usually a built-in feature of most modern spectrometer software packages. The k-spectrum thus obtained is directly analogous to the absorption spectrum of the material (Willis and Powell, 1990).

26.2.4.2 Reflection–Absorption

In the case of metals, the reflectivity of the interface can be used to record infrared spectra of adsorbed monolayers and condensed ultrathin films. It was first demonstrated theoretically (Greenler, 1966) and then practically (Pritchard and Sims, 1970) that, under certain

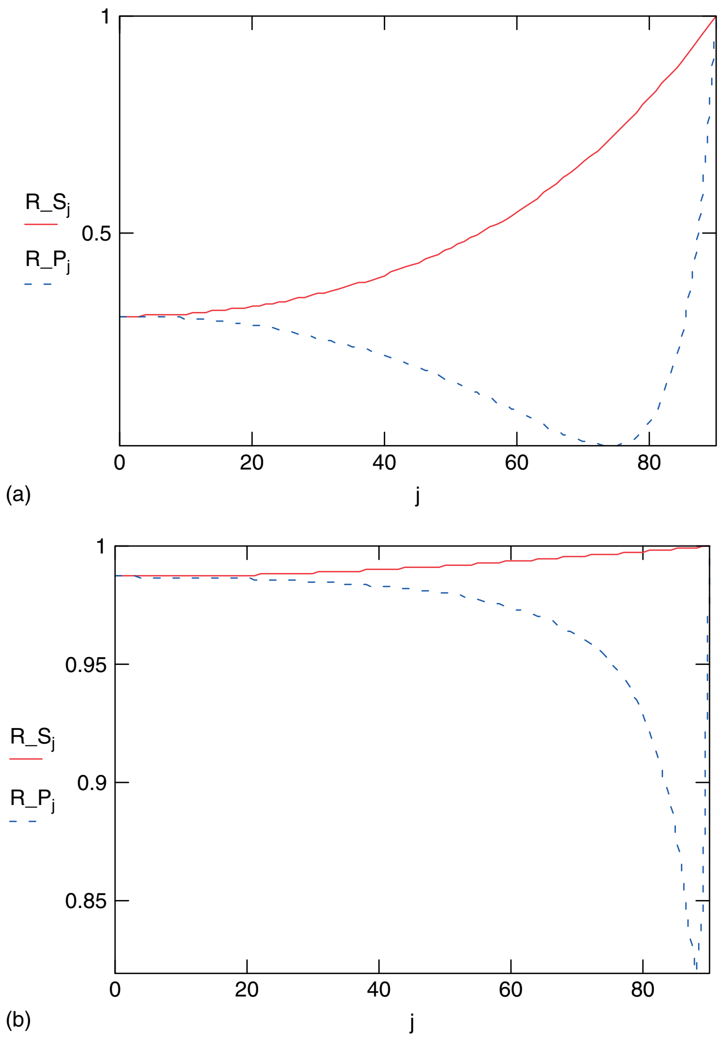

Fig. 26.7 Modeled reflectivity of the vacuum/substrate interface for (a) a typical metal with optical constants $n = 3.0$, $k = 30.0$ and (b) a typical nonabsorbing dielectric with $n = 3.44$, $k = 0$.

circumstances, infrared spectra of monolayers in metals could be obtained in a "reflection–absorption" configuration. The method, referred to variously as *reflection–absorption infrared spectroscopy* (*RAIRS*) or infrared reflection–absorption spectroscopy (IRRAS), has subsequently been used extensively to study adsorption and reaction of a substantial range of adsorbates on a variety of transition metal surfaces that are of interest in catalysis, corrosion, and electrochemistry. Practically, the experiment requires light to be incident upon the metal surface containing the adsorbate at a grazing incidence. This behavior

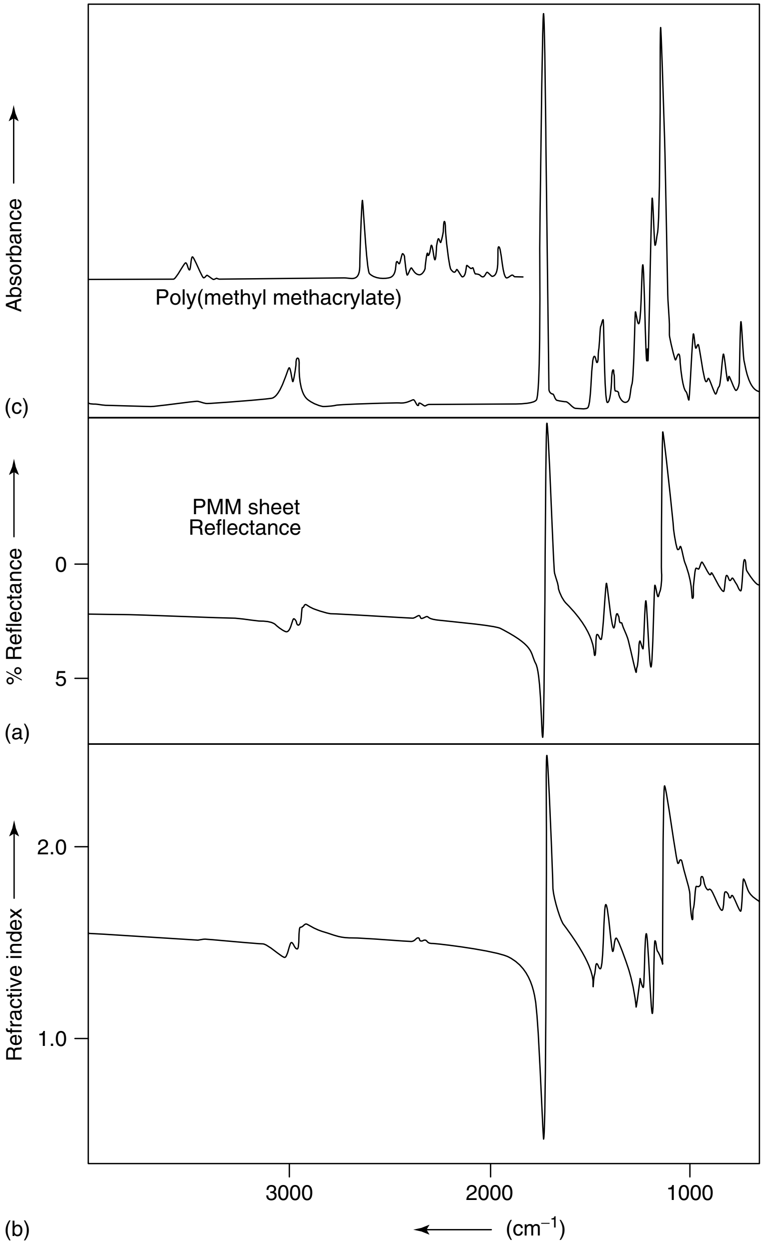

Fig. 26.8 Reflectance spectrum of poly(methyl methacrylate), showing deconvolution of the pure reflectance (a) into the *n*- (b) and *k*- (c) spectral components. Reproduced from (Willis and Powell, 1990.)

again has its origins in the optics of the vacuum/metal interface, and using an extension of the Fresnel model described above it can be shown that only vibrations with a component of their dynamic dipole, which is perpendicular to the surface, are seen. Grazing incidence is therefore necessary both to maximize the surface-normal component of the infrared beam and to interact favorably with the optical properties of the surface. Optically, coupling of the dynamic dipole of vibrational motion parallel to the surface to the infrared radiation is suppressed. A phenomenological description of this behavior was first given by Pearce and Sheppard (1976) who invoked the idea of ''image charge'' in the surface to show that vibrational dynamic dipoles parallel to the surface cancel out, whereas those perpendicular are reinforced, leading to the so-called metal-surface selection rule (MSSR). A detailed description of the optics and practicalities of RAIRS measurements are given elsewhere by Hollins (2000) and Bradshaw and Schweizer (1988).

Typical, RAIRS is used to monitor the adsorption of gaseous species on reactive metal surfaces. This frequently involves the use of a highly polished metal single crystal or polycrystalline metal film as a substrate, normally supported in ultrahigh vacuum to prevent adsorption of contaminants. The infrared beam from a spectrometer is coupled into the vacuum chamber through an infrared-transparent window and is incident upon the sample at a highly grazing angle. The reflected light is recollimated and focused onto a detector externally (in some cases, the source is the external component rather than the detector).

The MSSR is ideal for the identification of adsorbate geometries on metal surfaces, since only those vibrational modes with transition dipole moments perpendicular to the surface can be observed in the resulting RAIR spectrum. If accurate assignment of the normal modes to specific directions with respect to the molecular axes is possible, the presence or absence of individual modes can thus be used to deduce orientation with respect to the surface. For example, the symmetric C—H stretching mode of a methyl group attached to another moiety has its transition dipole oriented the C_3 rotation axis, whereas the antisymmetric CH stretch is polarized in a plane perpendicular to the axis. Simultaneous absence of antisymmetric mode and presence of the symmetric mode in the RAIR spectrum would therefore imply that the C_3 axis for the methyl group is perpendicular to the plane of the surface. The opposite configuration, that is, where the C_3 axis is parallel to the surface, would reverse the activity of the modes in the resulting RAIR spectrum. For whole molecules, the chemical structure of the adsorbates and their orientation with respect to the surface can often be determined by examining the positions and intensities of the absorption bands in comparison to reference spectra from isotropic bulk samples and gases. As an example, Figure 26.9 shows a RAIR spectrum from the adsorption of methyl silane on Cu(111) at 50 K. In this case, the adsorbate remains intact but lies predominantly with the Si–C axis parallel to the surface in the monolayer. Increasing coverage beyond a single monolayer results in a more disordered multilayer with a range of orientations more redolent of an amorphous thin film or liquid (Menard, Tear and Horn, 2008). Information such as this can be vital to the development of the understanding

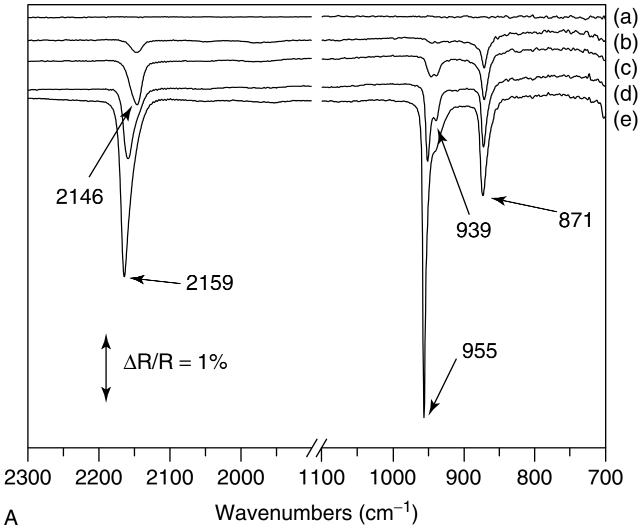

Fig. 26.9 Reflection–absorption infrared spectrum of CH_3SiH_3 (methyl silane) adsorbed on a copper Cu(111) surface at 50 K. The spectra show increasing coverage of the surface from (a) to (e), in which monolayer features are first observed, followed by multilayer features. Under the terms of the metal-surface selection rule, the bands at 870 and 2148 cm^{-1} indicate that the molecule lies with its C_3 axis flat to the surface in the monolayer; further, as the material is deposited, a disordered multilayer forms and many more bands become active. Taken from (Menard, Tear and Horn, 2008).

of catalytic processes at a microscopic level. In less specialized experiments, RAIR spectra can also be obtained from self-assembled monolayers and Langmuir–Blodgett films on metal surfaces under ambient conditions (Horn, 1998). In general, care should be taken in the interpretation of RAIR spectra from thicker films as the optical properties of the ambient phase/thin layer/substrate become less straightforward with increasing thickness. This can lead to a range of band position and shape perturbations, for instance, through the excitation of longitudinal optical resonances (Berreman, 1963). However, with a detailed knowledge of the directionality of the vibrational dynamic dipole moments, orientational information can be extracted with a very high degree of sophistication (Allara and Nuzzo, 1985).

RAIR spectra can also be obtained from nonmetallic surfaces under certain conditions. To a large extent, the MSSR does not apply except for some semimetallic substrates, with vibrational dynamic dipoles in all orientations being observed (although with different intensities). Once again, the behavior is adequately described by a simple Fresnel model. Once again using silicon as a typical example covered with a 1-nm film of isotropic adsorbate (i.e., dynamic dipoles in all directions), the following behavior is observed at the peak of an absorption band. At normal incidence, both the s- and p-polarized responses are equal (as they must be, since both are parallel to the surface at normal incidence) and negative in absorbance units, that is, the presence of an adsorption band results in an *increase* in surface reflectivity compared to the clean surface. With increasing angle, the s-polarized RAIRS response falls monotonically reaching zero at grazing incidence. The p-polarized behavior falls, crosses zero, becomes large and positive at angles of incidence just below θ_B, inverts to become large and negative above θ_B, and again reaches zero at grazing incidence. The p-polarized behavior is largely determined by the reflectivity of the underlying substrate, with the discontinuity around θ_B being a "divide by zero" error. Figure 26.10 shows this behavior schematically.

Reflection–absorption is not confined to gas/substrate interfaces. Recently, there has been much progress in the acquisition of infrared spectra from the surfaces of electrochemical cells or from substrates under high pressures of ambient gas. A number of methods are required to make this possible, largely due to strong absorptions by the electrolyte or gas. In the case of high gas pressures, this can be achieved using the s-polarized (non-interacting) response of a surface as a normalizing signal. In this case, the s-polarized component contains information about the ambient absorption with no spectral information from the surface due to the MSSR rule. The p-polarized response contains the same (isotropic) absorptions from the electrolyte and also the p-polarized surface RAIR information. Polarization modulation and normalization to the s-polarized spectra again allows the gas-phase and

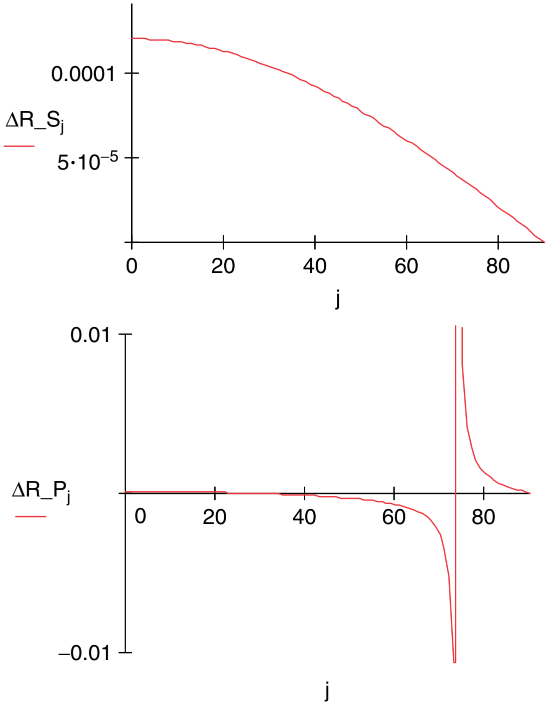

Fig. 26.10 Modeled reflectivity change due to the presence of a thin absorbing layer on a dielectric (silicon) surface as a function of angle of incidence for s- and p-polarized radiation.

surface components to be separated. For electrochemical surfaces, reduction of the ambient (electrolyte) absorption can be achieved using ultrathin electrolyte films above metal surfaces, although transport effects often dominate under such conditions. The use of thicker electrolyte layers requires the use of more advanced methods to remove the effect of electrolyte absorption. Normalized difference spectra are obtained as a function of potential, and the method is frequently referred to as *surface-normalized FT infrared spectroscopy*, (*SNIFTIRS*), and can be very sensitive to very small quantities of adsorbates on electrode surfaces (see, e.g., (Chen and Lipkowski, 1999)). Positive and negative-going bands are seen in the spectra, indicating adsorption and desorption of species from the surface. This differential reflectance behaviour can be modeled to obtain orientational information from the adsorbates under certain circumstances (Pettinger *et al.*, 2001).

In practical terms, accessories are available to fit within the sample compartment of conventional benchtop FTIR spectrometers for recording RAIR spectra from thin films on different substrates. The simplest are fixed angle devices using 80 or 85° for thin films on metallic surfaces. More sophisticated accessories incorporate the ability to change the angle of incidence. A schematic of such a device is shown in Figure 26.11.

26.2.4.3 **Diffuse Reflectance**

Powdered material, whether comprised of regular crystallites or random amorphous particles, scatters light in a range of directions. This light scattering arises from a range of sources, some of which contain spectroscopic information. Where the individual crystallite surfaces are optically flat, some of the scatter arises from specular reflection (from these flat surfaces in the sample), although the range of orientations of the crystallites means that the scattered light exits in all directions. For optically rough surfaces, the scatter is again diffuse. Both of these contributions may contain some spectroscopic information from resonant interaction with material at the surface of the particles, but the effects are usually relatively weak. From a spectroscopic viewpoint, the most useful component arises from light that has passed through the particles before escaping. This component is often internally reflected several times, and again, the scattered light is emitted randomly over a wide solid angle. During its transit through the particles, resonant absorption can occur; hence, comparison of this component to diffusely scattered light from a nonabsorbing reference sample may give considerable spectroscopic information.

The theory of diffuse reflectance (DR) has been described in detail elsewhere, with various structures and nomenclatures used to identify and quantify the contributions to the total collected diffuse scatter (Milosevic and Berets, 2002). The most common treatment is the Kubelka–Munk model, for which a relatively simple solution exists for the case of an infinitely (practically, a few millimeters in a suitable metal cup) thick sample. From an experimental viewpoint, the measurement of spectroscopic information from diffusely reflected light from a sample revolves around the separation of the diffuse specular reflection

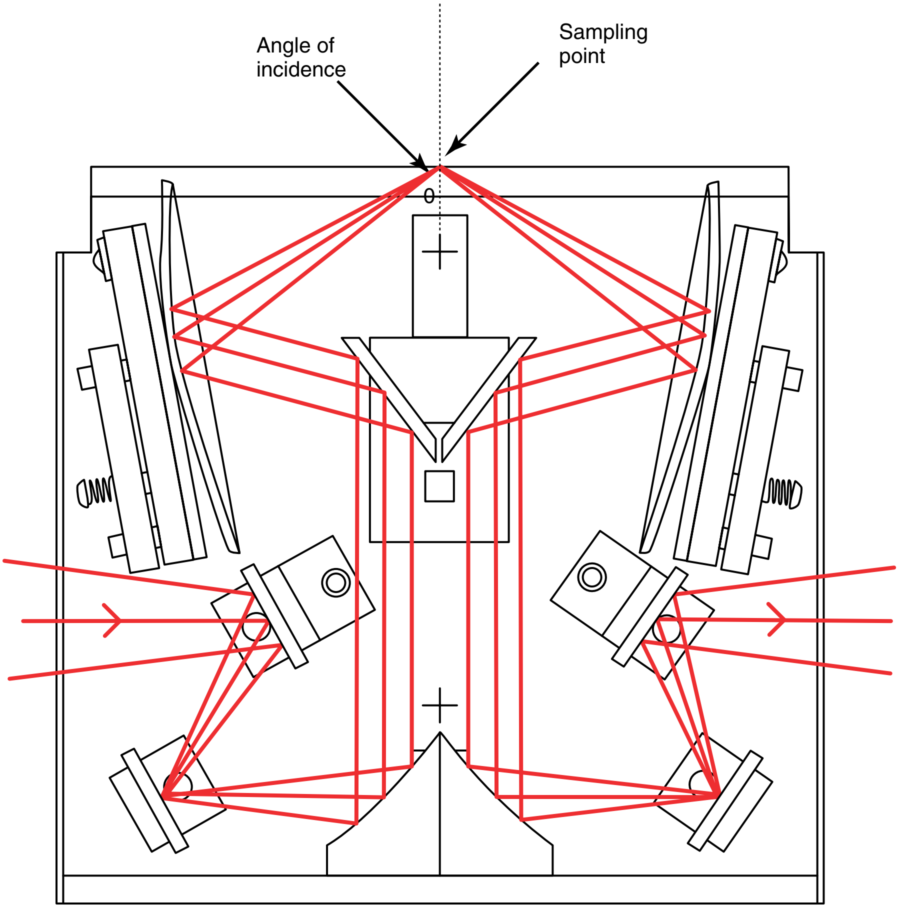

Fig. 26.11 Schematic of the optical path through a variable angle reflection accessory. Reproduced by kind permission of Pike Technologies.

from the internally reflected diffuse component. Practically, sampling utilizes a focused beam incident upon a cup containing the powdered sample. The diffusely reflected light is collected over a large solid angle using a specially shaped mirror, the purpose of which is to recollimate a maximal amount of the internally diffusely scattered radiation and return it to the light path of the spectrometer. The spectral parameter commonly obtained is R_∞, the ratio of the diffuse scatter as a function of wavenumber from a given sample and that from a powdered, nonabsorbing sample. From the Kulbelka–Munk model for an infinitely thick sample with minimal diffuse specular scatter, the absorption k and the scatter s are related to R_∞ such that

$$\frac{k}{s} = \frac{(1 - R_\infty)^2}{2R_\infty} \tag{26.27}$$

A number of sampling methods are in widespread use. The most common application simply requires a powdered sample to be placed in a cup at the focus point of the optics. However, in samples with very strong absorption bands, distortion occurs and it is necessary to dilute the powder. This is normally done by diluting the sample (typically to about 5%) in an inert matrix such as KBr. The sample remains loosely packed (compared to pressed KBr discs for transmission measurements) and thus there are fewer problems with pressure-induced effects on the sample. Another method for the preparation of samples for diffuse reflectance measurements utilizes a SiC disc, which is used to abrade material from otherwise intractable solid samples. The DR spectrum can be recorded directly from this disc, although the optical arrangement is far from the ideal Kubelka–Munk situation and is at best a qualitative measurement.

Quantitation in diffuse reflectance spectroscopy requires great care, as the effects of scattering from the matrix, concentration of the absorber, and dispersion in the optical constants of the sample are complex. For "ideal" powdered samples, the penetration depth of the radiation before diffuse scattering is dependent upon k, which means that diffuse reflectance measurements are particularly sensitive to weak absorptions, for which the spectroscopic path length is effectively longer.

The design of apparatus to collect diffuse scatted from powdered samples is nontrivial, and a number of designs are in use. Broadly speaking, the challenges are threefold: to illuminate the sample with as much incident radiation from the spectrometer as possible, to collect as much of the diffusely scattered light as possible, and to minimize the amount of diffuse specular reflection collected. There is a further trade-off between the solid collection angle and the area of the sample illuminated, which requires optimization in any optical configuration. The high sensitivity of Fourier-transform spectrometers is essential for diffuse reflectance measurements, as the restrictions on collection area and geometry, as well as the inclusion of various methods to reduce the specular scatter, reduce the throughput of diffuse reflectance accessories to as little as 5% of the total power. The optical configuration of a typical diffuse reflection accessory is shown in Figure 26.12. The accessory is located in the sample compartment of the spectrometer, intercepting the spectrometer beam and directing in onto specially shaped mirrors. The diffusely reflected light is collected and returned to the spectrometer light path.

In such a configuration, controlled environment chambers can be added at the sample point, allowing the investigation of reactions on particles of environmental and catalytic interest to be studied. Figure 26.13 shows a modified diffuse reflectance accessory used to observe the uptake of gas-phase NO_2 onto NaCl particles representative of material present in the atmosphere in the marine boundary layer. In this particular case, the adsorption process shows the formation of nitrate during in the reaction, which remains adsorbed onto the particle surface (Vogt and Finlayson-Pitts, 1994).

26.2.4.4 Attenuated Total Internal Reflection

The most significant development in the applicability of infrared spectroscopy to the widest variety of samples without significant preparation over recent years is the advent of attenuated total internal reflection (ATR) based methods. The ATR effect was first described by Newton in *Opticks*, and the seminal work by

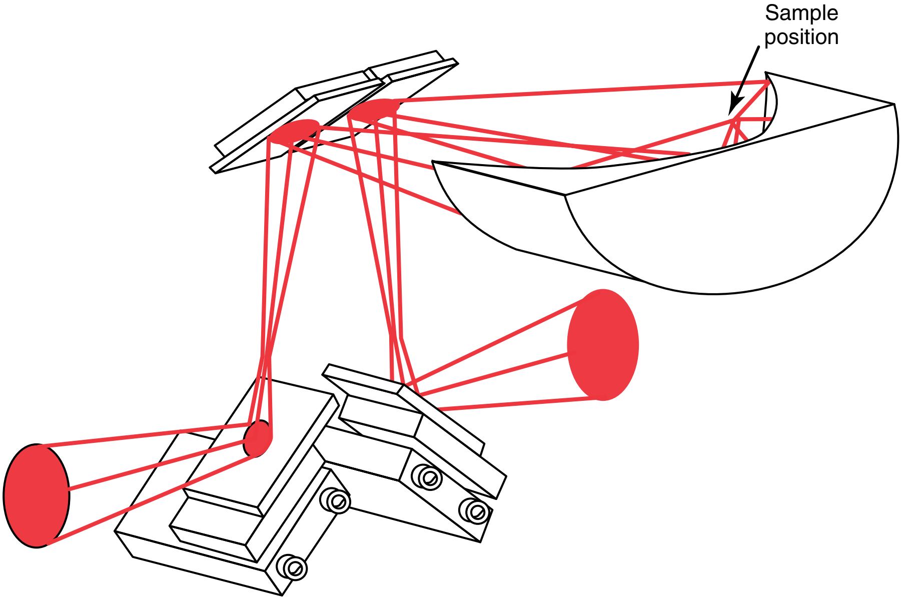

Fig. 26.12 Schematic of the light path through a typical commercial diffuse reflectance accessory. A single complex mirror directs the infrared beam onto the sample and collects the diffusely reflected light for return into the spectrometer light path. Reproduced by kind permission of Pike Technologies Inc.

Harrick (1967) in the 1960s set the groundwork for the wide-scale use of ATR methods in optical spectroscopy.

When light is incident upon the interface between two nonabsorbing media with refractive indices n_1 and n_2 from within that with the higher refractive index, total internal reflection is observed above a certain critical angle, θ_c, given as

$$\theta_c = \tan^{-1}\left(\frac{n_2}{n_1}\right) \tag{26.28}$$

For a germanium substrate in contact with a vacuum, $n_1 = 4$, $n_2 = 1$ and hence $\theta_c = 14°$ from the surface normal. Below this angle, a proportion of the light is internally reflected and the remainder is transmitted through the interface. The variation of the internal component with angle of incidence is again adequately described by the Fresnel equations, as shown in Figure 26.14(a). Subtly different angle-dependent behavior is observed for s- and p-polarized infrared radiation, although the critical angle is the same for both.

The light that is internally reflected at the interface penetrates beyond it in the form of an evanescent wave. The amplitude of the evanescent wave falls exponentially with distance from the interface, such that the field strength at a distance z, denoted as $E(z)$, is given as

$$E(z) = E_0 \exp\left(\frac{-z}{d_p}\right) \tag{26.29}$$

The term d_p is the "depth of penetration", a parameter that depends upon the relative refractive indices of the two

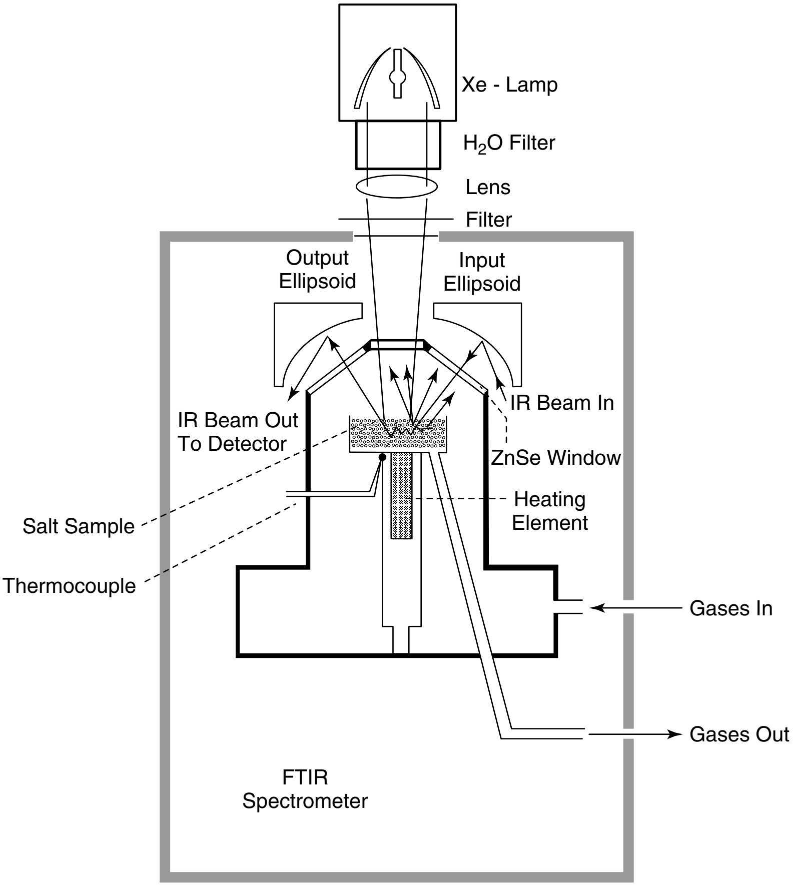
Xe - Lamp
H_2O Filter
Lens
Filter
Output Ellipsoid
Input Ellipsoid
IR Beam Out To Detector
IR Beam In
ZnSe Window
Salt Sample
Heating Element
Thermocouple
Gases In
Gases Out
FTIR Spectrometer

Fig. 26.13 Schematic of apparatus used to collect diffuse reflectance infrared spectra of salt particles during reaction with ambient gases. Taken from Vogt and Finlayson-Pitts (1994).

media (n_1 and n_2), the wavelength (λ), and the angle of incidence (θ), and is effectively the distance at which the evanescent wave falls to $1/e$ of its value at the interface:

$$d_p = \frac{\lambda}{2\pi n_1 \left(\sin^2\theta - \left(\frac{n_2}{n_1}\right)^2\right)^{1/2}} \tag{26.30}$$

The evanescent wave can couple resonantly with vibrational dynamic dipole moments close to the interface. In such cases, energy is lost from the internally reflected wave, the process being referred to as *attenuated total internal reflection*. The magnitude of the attenuation is related to k for the material in contact; hence, a spectrum recorded using an ATR geometry is directly related to the true absorption spectrum (Mirabella, 2002). Figure 26.14(b) shows an illustration of the magnitude of the attenuation of total internal reflection for a semi-infinitely thick film of a

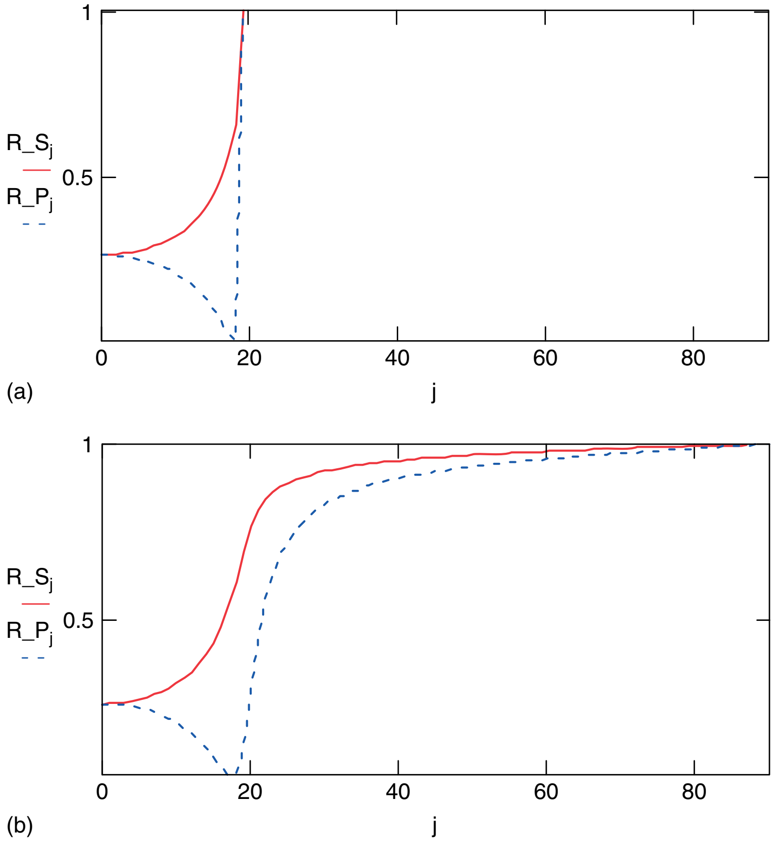

Fig. 26.14 (a) Modeled internal reflectivity of the interface between two transparent media (Ge, $n_1 = 4.0$ and vacuum, $n_2 = 1$) as a function of the angle of incidence for s- and p-polarization. (b) Modeled reflectance of the same interface with an infinitely thick layer of a weakly absorbing material replacing the vacuum ($n_2 = 1.2$, $k_2 = 0.03$).

moderately strong absorber ($n_2 = 1.2$, $k_2 = 0.03$) on a Ge internal reflection element (IRE). The magnitude of the attenuation is greatest close to the critical angle and decreases with increasing angle. There is also a difference in the attenuation of s- and p-polarized light. In practice, angles between 45 and 60° and unpolarized radiation are used, depending upon the nature of the materials to be studied and the application.

From the above equation, it can be clearly seen that there is a linear dependence of d_p upon the wavelength. At $2000\,\mathrm{cm}^{-1}$ for ZnSe (a common choice of material for ATR) at 60°, $d_p \approx 2\,\mu\mathrm{m}$ for example. The concomitant effect of the wavelength dependence of d_p on the evanescent field strength versus distance above the internally reflecting interface across the spectrum is that the intensity of absorption bands across the spectrum is therefore scaled linearly with λ. Most spectrometer software packages contain routines for correction of this phenomenon. However, care is needed while using these corrections, not least because spectrometer optics normally produce a 7–10° cone of incident angles at the IRE surface.

The advent of FTIR spectrometers has made a variety of experimental configurations viable for the collection of ATR infrared spectra, since both the sensitivity is higher and the optical configuration of accessories is not constrained by slits, as was the case in the original ATR-IR apparatus used by Harrick and Fahrenfort (Harrick, 1967) in the earliest practical demonstrations of the technique. The basis of all accessories is the IRE, and arrangements using prisms, trapezoids, parallelograms, cylinders, and hemispheres are or have been available in commercial devices. A primary choice is between multiple reflection and single reflection configurations, each having their own applications and advantages (Fitzpatrick and Reffner, 2002).

Multiple reflection ATR accessories for FTIR applications tend to use horizontal trapezoidal IREs. The effective path length in these IREs is hence a multiple of d_p for each reflection, with the number of reflections being readily calculated from the dimensions of the IRE. Given the limited range of the evanescent wave above the IRE surface, good contact is essential. High surface area horizontal IREs preclude the application of the high pressures necessary to ensure good contact between harder solids and the surface and so are of limited application to powders. However, these devices are very useful for solutions and soft solid materials. Typically, the IRE is mounted in a flat plate or trough. Light from the spectrometer compartment is coupled onto the beveled end of the IRE, undergoes multiple internal reflections, exits the end of the crystal, and is returned to the spectrometer light path. For high-index materials such as Ge, antireflection coatings on the input and output faces of the IRE are essential to increase optical throughput. Figure 26.15

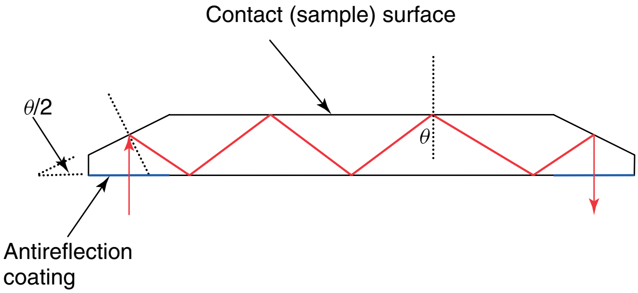

Fig. 26.15 Schematic of multiple internal reflection on a shaped prism (two reflections at the sampling interface). The angle of the beveled ends dictates the internal reflection angle θ. An antireflection coating is often used with high-index materials (e.g., Ge) to reduce reflection losses. In practice, IREs of this type are normally bonded into metal plates for location in optical carriages that translate the infrared beam to and from the sample compartment of a conventional FTIR spectrometer.

shows a horizontal ATR configuration in schematic. Flat plates and troughs are usually widely available, with choice depending upon application. In practice, these are normally removable for cleaning. Multiple internal reflection measurements of adsorption onto surfaces are also feasible, as demonstrated in the experiments of Chabal and coworkers, in which a polished and shaped silicon crystal was used as both a substrate for adsorption and an IRE (Chabal, 1986). Sensitivity well beyond that achievable for reflection–absorption or transmission measurements was achieved as a result of both the surface sensitivity and multiple reflections.

In recent years, single reflection ATR accessories have become more widely available. Figure 26.16 shows an example of the light path through a commercial sample-compartment-mounted single bounce ATR accessory. In this case, the IRE consists of a shaped ZnSe element with a thin diamond as the sample contact patch. Often such IREs also act as focusing elements, optimizing the amount of light directed onto the small diamond element (usually only a few millimeters in diameter). Light is coupled into and out of the IRE by plane mirrors, making a single internal reflection at the diamond/air interface. To improve sample contact between harder materials and the IRE, pressure can be applied with a clamp. ATR infrared spectra of liquids can also be obtained by placing small quantities of the liquid (or solution) analyte onto the diamond ATR, although it is often necessary to take steps to prevent evaporation from volatile media. ATR

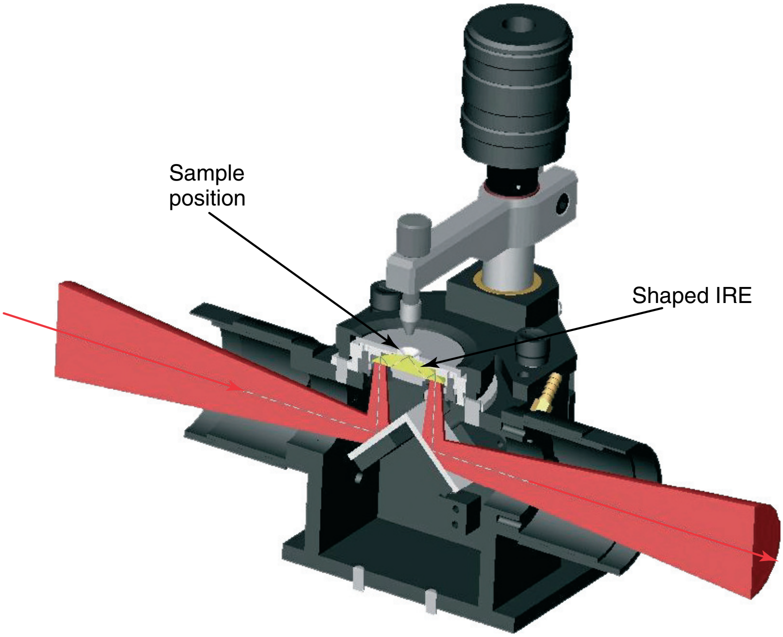

Fig. 26.16 Light path through a commercial single reflection diamond ATR accessory. Reproduced by kind permission of Pike Technologies Inc.

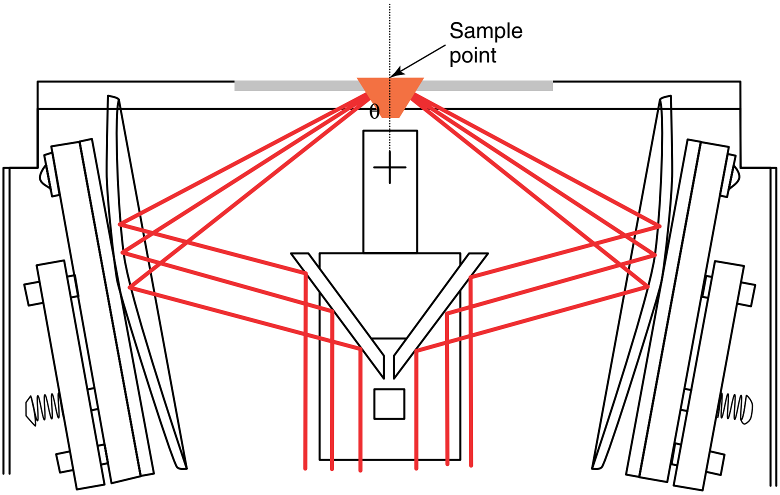

Fig. 26.17 Schematic of the light path through a commercial variable angle ATR accessory (based upon the ATR version of the Pike VeeMax II accessory). Reproduced by kind permission of Pike Technologies Inc.

accessories using composite (diamond-tipped) IREs are also widely used for in situ process monitoring, with the light from the spectrometer being directed to and from the IRE by fiber optics or light pipes. Single reflection accessories are also available with the ability to vary the internal reflection angle. These are useful for a number of purposes, including depth profiling and optical constant measurement. Figure 26.17 shows a schematic of a typical accessory, in which the infrared light from a spectrometer is coupled into and out of a prism IRE in such a way that the angle of incidence can be varied within certain limits.

ATR-based methods are now widely used as the method of choice for routine infrared analysis in a wide range of applications. However, the optics of the configuration also lend themselves to more esoteric applications. The effect of restricted spectroscopic penetration into an ambient medium using the ATR technique can be used effectively to study rates of diffusion in thin films deposited upon the surface of an IRE, as demonstrated by Pereira and Yarwood (1996). The rationale relies upon the diffusion process bringing material from an area beyond the sensitivity of the probe beam toward the IRE surface. As the material approaches the interface, it gives rise to spectral features whose intensity increases with time at a rate determined by the optical properties of the IRE/thin film and the rate of diffusion. The technique can be very sensitive when diffusion occurs upon an appropriate timescale for infrared measurements and has been applied to both polymers and condensed ice films. The rate of this diffusion can be calculated by assuming that the rate of diffusion is Fickian, in which case the absorbance versus time plot convolving spectral intensity with mass transfer will have the following form:

$$\frac{A(t)-A(0)}{A(\infty)-A(0)} = 1 - \frac{8}{\pi d_p\left(1-\exp\left[\frac{-2L}{d_p}\right]\right)} \sum_{n=0}^{\infty}\frac{1}{2n+1} \times \left[\frac{\exp\left[\frac{-2L}{d_p}\right]\left(\frac{2n+1}{2L}\pi\right)+(-1)^n\frac{2}{d_p}}{\left(\frac{2n+1}{2L}\pi\right)^2+\frac{4}{d_p^2}}\right]\exp\left[-\left(\frac{2n+1}{2L}\pi\right)^2 Dt\right] \tag{26.31}$$

where $A(t)$ is the absorbance at time t, $A(0)$ is the absorbance at $t = 0$, $A(\infty)$ is the saturation absorbance, d_p is the penetration depth (as defined in Eq. (26.30)), L is the thickness of the film, and D is the Fickian diffusion coefficient. Figure 26.18 shows a plot of the intensity of one of the characteristic features of ionized HCl diffusing through an ice film along with a fitted curve for a film thickness of 1.5 μm. A diffusion coefficient of $1.5 \times 10^{-15} \mathrm{m^2 s^{-1}}$ is obtained from this data (Horn and Sully, 1997).

26.2.5
Infrared Microspectroscopy

Infrared microspectroscopy (simply infrared microscopy in some references) relies upon the focusing of the infrared beam from an interferometer onto a microscopic sample and subsequent measurement of the infrared spectra of selected sections of the magnified image sampled from the focal plane of the collection optics. The optical considerations for infrared microspectroscopy are practically identical to those of optical microscopes. However, for infrared radiation with a wide range of constituent wavelengths, the use of reflection-based optical components is essential to minimize chromatic aberration. The smallest area from which individual nonoverlapping spectra can be obtained is of the order of the longest wavelength component of the radiation convolved with the numerical aperture of the condenser/objective elements, with about 25×25 μm^2 being a typical practical limit. Trade-offs between wavelength range, magnification, detector element size, and optical throughput must all be considered when selecting instrumentation. For single-point measurements, sample area (size and position) is selected from the magnified image in the focal plane of the objective element with an aperture. Since the amount of light reaching the detector through these various optical stages is relatively small (some 1–2 orders of magnitude less than conventional accessories), the use of high sensitivity, cooled MCT detectors is essential. Further details of the construction and design of conventional infrared microspectrocopy apparatus are given by Sommer (2002). Figure 26.19 shows a plain view of a commercial FTIR microscope, showing the complex light path needed for focusing, imaging, and magnification.

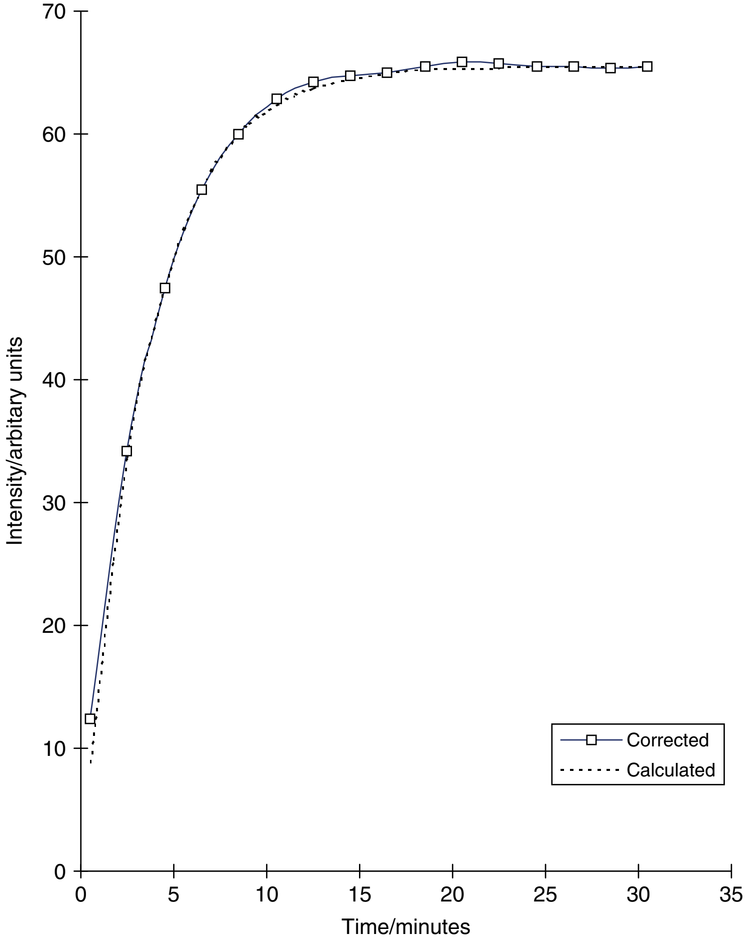

Fig. 26.18 Fickian diffusion of dissociated HCl through an ice film, measured by ATR. Taken from (Horn and Sully, 1997). The solid lines show measured integrated absorbance under the $\nu(H_3O^+)$ band, while the dotted line shows a fit using Eq. (26.31).

A variety of configurations are possible, including transmission, reflection, and ATR. Transmission measurements rely upon the production of films of sample that are sufficiently thin to keep absorbances below 1.0, which can be achieved by either microtoming samples or compression between optically transparent plates. When supported upon optically transparent substrates, corrections for focal length are needed because of refraction effects. Near-normal incidence reflection measurements are very similar in essence to transmission, differing only in the use of a reflective rather than optically transparent substrate and the simultaneous use of the objective as a condenser. For reflection at angles other than near normal, specialized objectives are needed to focus and collect

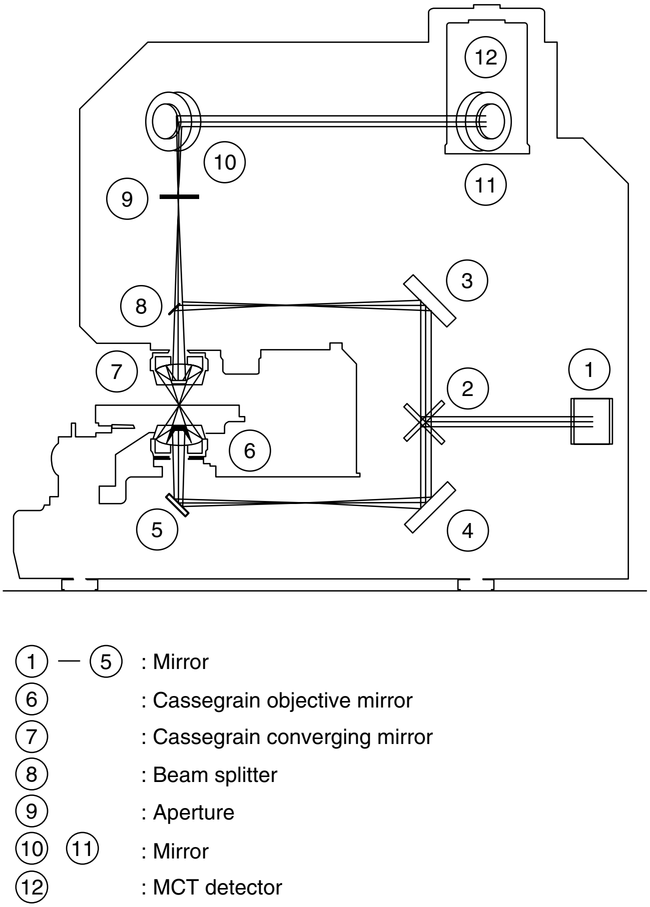
⑫
⑩
⑪
⑨
③
⑧
⑦
②
①
⑥
⑤
④
① — ⑤ : Mirror
⑥ : Cassegrain objective mirror
⑦ : Cassegrain converging mirror
⑧ : Beam splitter
⑨ : Aperture
⑩ ⑪ : Mirror
⑫ : MCT detector

Fig. 26.19 Schematic of the light path through a typical commercial FTIR microscope. Reproduced by kind permission of JASCO UK.

the infrared light. ATR objectives rely upon shaped IREs with the sampling area being determined by the size of the contact patch of the IRE. In practice, such objectives also need to incorporate an optical light path for the alignment and sample point selection. For further examples, see Humecki (1995) and references therein.

Major improvements in the sensitivity and usefulness of infrared microspectroscopy have arisen from two ways. The first is the use of synchrotron radiation sources for high-intensity infrared light. For example, an FTIR microscope recently developed at the Daresbury Laboratory in NW England utilizes an $8 \times 8\ \mu m^2$ spot from the synchrotron at high brightness. This has found application to a wide variety of microscopic samples, including prostate cancer cells with unique spectral features (Gazi *et al.*, 2007). Imaging of samples is also possible, with significant advantages over conventional sources where collection times are long due to the lower throughput.

Imaging the infrared image from the focal plane requires an array detector, and a number of such instruments are now widely available with focal plane array (FPA) detectors from 64 × 64 pixels upward. Step-scan interferometers are required for such applications in order to facilitate the recovery of interferograms from such an array. Technology is continuously improving in this field, with rapid increase in detector pixel number and density. The applications of such technology are widespread, especially in biological sciences, where the ability to extract chemical images from samples by integrating specific absorption bands provides unprecedented detail. In Figure 26.20, images of a eucalyptus leaf section from conventional microscopy, an imaged synchrotron microscope, and an FPA-based detector are shown (Heraud *et al.*, 2006). The authors compare the spatial resolution and the signal-to-noise ratio of the FPA imaging and synchrotron mapping methods. The quality of the images is roughly equal, with the main advantage being that the FPA images were obtained in about 2 minutes, compared to 8 hours using the synchrotron source, although the latter has superior signal-to-noise ratio and potentially better spatial resolution. The major features of the internal structures of the leaf are clearly seen from the integrated amide I band.

26.2.6
Extinction Methods from Particles

When infrared light is incident upon particles with diameters similar to that of the wavelength of the light, scattering occurs. This results in an extinction spectrum that is the sum of a scattering contribution that scales with wavenumber and particle radius and an absorption contribution due to vibrational resonances, which is dependent on concentration, absorption coefficient, and path length of the absorbing species. Typically, Mie theory can be used to relate the observed infrared spectrum to aerosol size and concentration. A number of assumptions must be made for simplification: (i) that scattering is elastic (frequency of scattered light is equal to that of incident light); (ii) only single-scattering occurs (i.e., the particles do not affect each other); (iii) the complex refractive index completely describes the optical properties of the particles; and (iv) the number of particles is large with random separations such that there is no systematic phase relation between waves scattered by individual particles (Bohren and Huffman, 1983). The intensity of light, I, passing through an aerosol is given as

$$\frac{I}{I_o} = e^{(-b_{\text{ext}}\,z)} \tag{26.32}$$

where I_0 is the incident intensity, z is the path length, and b_{ext} is the extinction coefficient. Typical treatments are expressed in terms of τ_{ext}, the light removed from the incident beam by the combined effect of scattering and absorption:

$$\tau_{\text{ext}} = \log_{10}\left(\frac{T_o}{T}\right) = \log_{10}\left(\frac{I_o}{I}\right) = \frac{b_{\text{ext}}\text{z}}{2.303} \tag{26.33}$$

For a distribution of particles with a range of diameters, it can be shown that

$$\tau_{\text{ext}} = \frac{1}{4}\frac{\pi {d_p}^2}{2.303}N(d_p)\text{Q}_{\text{ext}}(d_p)\text{z} \tag{26.34}$$

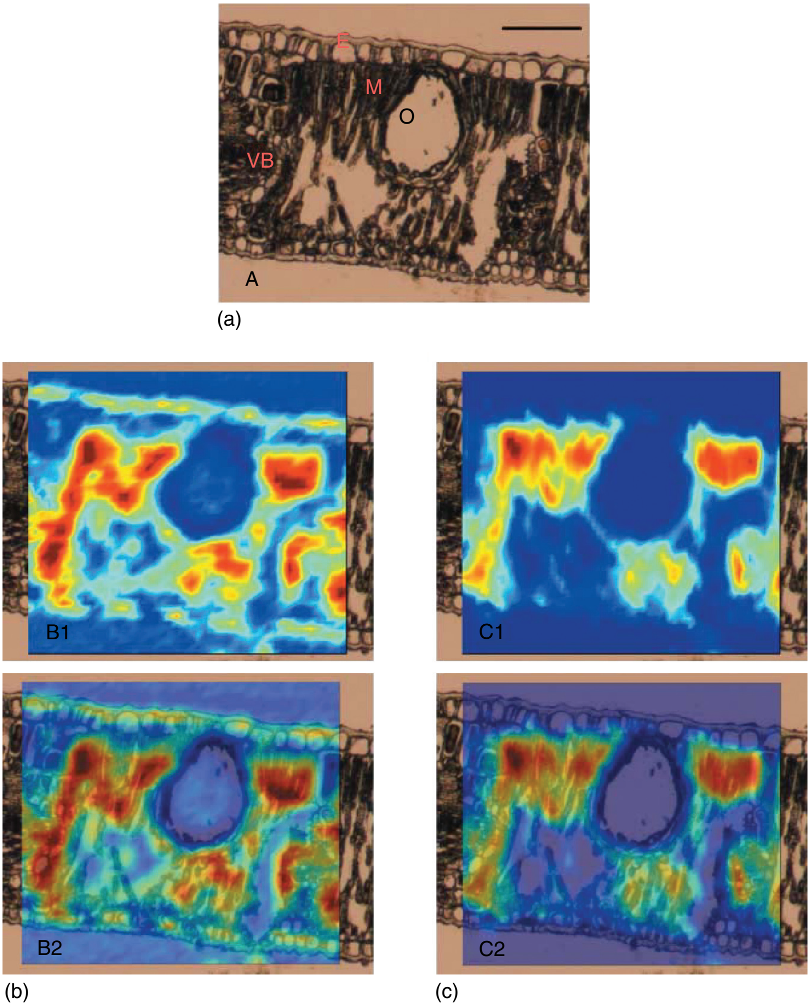

(a)

(b)

(c)

Fig. 26.20 Micrographs and infrared maps of eucalyptus leaves: (a) an optical micrograph and (b) and (c) comparisons between chemical maps of integrated absorbance from 1700 to 1580 cm^{-1} (amide I band) using a point-to point method with synchrotron radiation as a focused source (b) and with an FPA. In each case, the upper panels show the infrared maps at full opacity and the lower panels at 50% opacity. A rainbow scheme has been used to denote absorbance, with the warmest colors (red end of the spectrum). Reproduced from (Heraud *et al.*, 2006).

where d_p is particle diameter, $N(d_p)$ is the size-dependent number concentration, and $Q_{ext}(d_p)$ is the size-dependent extinction efficiency. Light extinction by small particles within the infrared light path has scattering and absorption contributions such that

$$Q_{ext} = Q_{sca} + Q_{abs} \tag{26.35}$$

Mie theory gives a basis for the calculation of the scattering and

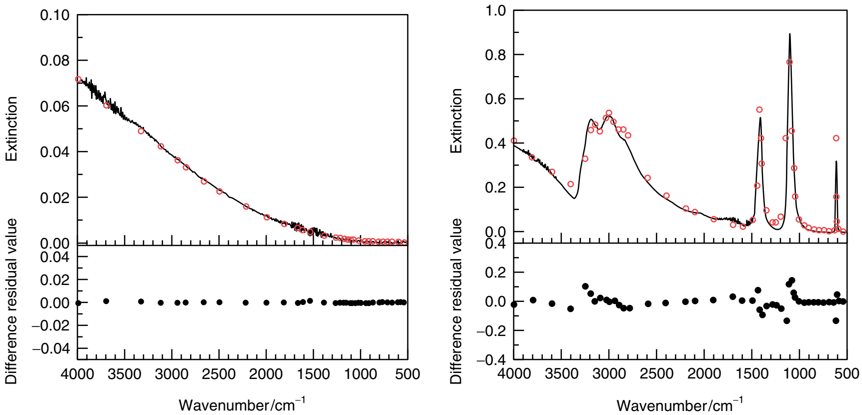

Fig. 26.21 Infrared extinction spectra from aerosols. (a) NaCl aerosols with average diameter 2.1 μm and number density 3.2×10^5 particles cm^{-3}. (b) $(NH_4)_2SO_4$ aerosols with 2.0 μm and $N = 1.1 \times 10^6$ particles cm^{-3}. In each case, the upper panels shown the FTIR extinction spectra (solid lines) and the best fit results based on Mie scattering calculations (circles). The lower panel shows the residual difference between the two spectra. Taken from Najera *et al.* (2009) with permission from the American Institute of Physics.

absorption of light by any spherical particle (with diameter d_p) as a function of the size distribution, number density of the aerosol particles, and the complex refractive index. Q_{ext} and Q_{sca} are usually determined iteratively. In practice, for materials with known wavelength-dependent complex refractive index, a test spectrum can be generated using these extinction coefficients based on an input size distribution and number density, with these parameters iteratively fitted to achieve a match to the real spectrum (Zasetsky *et al.*, 2007). Figure 26.21 shows a fit for two examples using published optical constants. For pure sodium chloride aerosols, there is no absorption component and the spectrum is entirely due to scattering. For ammonium sulphate particles, both scattering and absorption are present. The lower panels show the residual from the calculated versus measured spectra (Najera *et al.*, 2009).

26.3
Interpretation of Infrared Spectra

26.3.1
Empirical Interpretation from Group Frequencies

The interpretation and analysis of infrared spectra draws upon a number of tools, some empirical and some more rigorous. Databases of library spectra have been available for many years (from Sadtler, Aldrich, Hummel among others), and algorithms for the comparison of spectra with those contained in reference libraries for the identification of compounds are constantly evolving. However, to use infrared spectroscopy as a black box method in such a manner would be to miss a substantial part of the richness of molecular information and complexity contained within an infrared spectrum. Therefore, this section concentrates on the methodology of first principles interpretation of infrared spectra using group frequencies and correlation tables. The interpretation of infrared spectra requires both a systematic approach to the assignment of functional group modes and an understanding of the nature of the effect of the intramolecular structure on the precise position of absorption bands. In particular, it is often considered essential to make correlations between the appearances of different types of group modes.

As discussed in the preceding sections, although many individual displacements make up a vibrational mode, the combinations which dominate a given normal mode tend to be of similar frequencies. Adjacent parts of a molecule often move independently of the rest of the molecule when masses are very different or when force constants are very different. A normal mode may then be mostly made up from the displacement of the atoms of a single functional group, with a characteristic absorption frequency almost regardless of the rest of the molecules. Normal modes tend to be dominated by a specific type of motion. For each individual atomic displacement in a polyatomic molecule, two possibilities exist: displacement along the bond direction (stretch) and displacement perpendicular to the bond direction (deformation). A variety of nomenclature exists, but in general functional group modes tend to be labeled ν if they involve mainly stretching, δ if they mainly involve bond bending, ρ if they mainly involve rocking, and τ if they involve twisting. It is common to label symmetric and antisymmetric vibrations with subscripts, for example, $\nu_s(H_2O)$ for the symmetric OH stretching mode of water.

The most useful functional group modes for direct assignment tend to be those which occur in regions of the spectra where there are few other resonances. However, functional groups in spectrally congested regions can frequently be identified when their presence is correlated with other types of vibration of the same group. This might occur in several ways, the simplest being when there are either two functional group modes within a single functionality or there are two types of vibration (e.g., stretching and deformation) of the same functionality.

Aldehydes (R$-$CHO) provide an excellent example of the former. The stretching mode of the aldehydic C$-$H bond, written as ν (C$-$H), typically occurs between 2900 and 2700 cm^{-1} and is usually quite weak, although often clearly distinguishable because the region of the spectrum in which it occurs is normally uncluttered by other resonances. The related aldehydic carbonyl stretching mode, labeled as ν (C=O), typically occurs between 1740 and 1720 cm^{-1} and is normally very strong. These features do not vary much between different aldehydes. Consequently, the presence of both together is a clear marker for the presence of an aldehydic functionality, regardless of the composition of the rest of the molecule. Furthermore, as discussed later, the band position of an aldehydic C=O is often sufficiently different from that of other carbonyl species.

An example of the latter is as follows: the presence of two types of vibration of the same set of atoms occurs widely in organic molecules and in biological materials. The stretching modes of methyl ($-CH_3$) and methylene ($-CH_2-$) groups can be observed in the 3000–2800 cm^{-1} range, a relatively uncluttered region. The corresponding deformation modes occur in the 1500–1350 cm^{-1} range, which is usually highly congested in all but the simplest molecules. Specifically, methylene groups in long alkane chains show symmetric $\nu_s(CH_2)$ and antisymmetric $\nu_{as}(CH_2)$ stretching modes at approximately 2860 and 2930 cm^{-1} and a deformation mode $\delta(CH_2)$ at 1470 cm^{-1}. The former are often readily identifiable: their presence allows confidence in the assignment of a spectral feature around 1470 cm^{-1} to $\delta(CH_2)$. Similarly for a methyl group, the symmetric $\nu_s(CH_3)$ and antisymmetric $\nu_{as}(CH_3)$ stretches occur at approximately 2870 and 2970 cm^{-1}, accompanied by symmetric $\delta_s(CH_3)$ and antisymmetric $\delta_{as}(CH_3)$ deformations at approximately 1375 and 1460 cm^{-1}. At a slightly higher level of complexity, bands of this type also show predictable patterns of relative band intensities and characteristic band positions.

The variation of group frequencies from one molecule to another provides a further challenge to the spectroscopist. Some group modes are relatively immobile, regardless of their molecular environment (e.g., methyl and methylene stretching modes). Others, particularly those derived from functionalities that couple strongly through either inter- or intramolecular interactions, exhibit substantial ranges.

In the latter case, shifts in position (and intensity) arise because of chemical and physical interactions. Double bonds can conjugate, resulting in reduced electron density within a specific region of the molecule, thereby reducing the force constant. For example, in isolated C=C and C=O double bonds, absorptions are observed from $\nu(C{=}C)$ at approximately 1650 cm^{-1} and from $\nu(C{=}O)$ at ca. 1720 cm^{-1}. In conjugation, for example, in C=C−C=O, the $\nu(C{=}C)$ absorption shifts to approximately 1600 cm^{-1} and the $\nu(C{=}O)$ absorption shifts to approximately 1675 cm^{-1}. Intermolecular forces also affect vibrational frequencies, again through perturbation of the electron density within a chemical bond (or set of bonds). For example, H-bonding lengthens and weakens O−H bonds such that $\nu(O{-}H)$ values vary widely (as does the band width). Interactions such as this tend to increase the frequency of deformation modes. Solution effects may also alter frequencies. Being highly polarized, carbonyl compounds interact with solvents, H-bonding with water and alcohols, and so on, with a resultant small shift (usually to lower wavenumber) of the band position.

Carbonyl compounds provide an excellent example of the effects of intramolecular phenomena such as induction (donation or loss of electron density from substituents), angle strain, and resonance upon bond strengths and therefore the positions of absorption bands. Inductive effects due to electron-withdrawing (−I) groups attached to the C atom compete with the $C^+{-}O^-$ resonance form and increase bond strength; hence, $\nu(C{=}O)$ increases, for example, $\nu(>C{=}O)$ in aldehydes $\approx$ 1720 cm^{-1} while $\nu(-COCl) \approx$ 1800 cm^{-1}. Electron-donating (+I) groups attached to the C atom stabilize the $C^+{-}O^-$ resonance form and reduce bond strength; hence, $\nu(C{=}O)$ decreases. Strain effects, where the angle formed between the two substituents is forced away from the preferred 120°, increase $\nu(C{=}O)$. This is particularly prevalent in cyclic systems. Mesomeric effects are those where one or some other forms of the resonance are stabilized by substituents or through interaction with other species. H-bonding may preferentially stabilize one of the mesomeric forms, for example ($C{=}O^- \ldots H{-}X \leftrightarrow C^+{-}O^- \ldots$

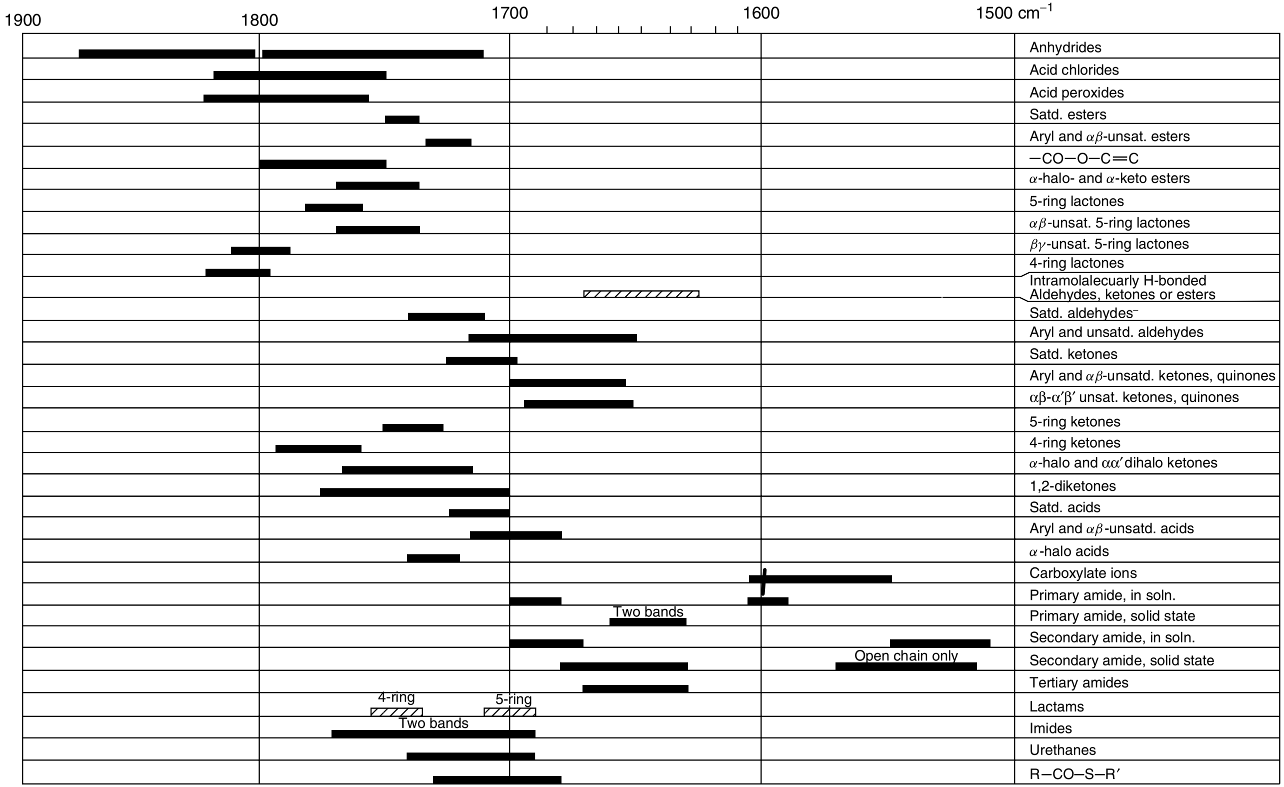

Fig. 26.22 Group frequency correlation table for the carbonyl region. Taken from Williams and Fleming (1995).

H−X), which again lowers the ν(C=O) frequency.

The above-mentioned factors are described in much greater depth in a range of textbooks (e.g., Williams and Fleming, 1995; Socrates, 2004). The information is often summarized in *group frequency correlation tables* such as that shown in Figure 26.22. A systematic approach is required for confidence in assignment. It is often easier to make correlations between the appearance of group modes if the spectrum is considered in a number of well-defined regions. Table 26.2 summarizes the regions in which the main functional group modes occur.

In larger molecules, a substantial proportion of the absorption bands is not easily assigned to vibrational modes of a specific functional group. These absorptions mainly occur below 1500 cm^{-1}, resulting in a congested spectral region. Bands that can sometimes be identified in this region include the C−C skeletal modes of organic molecules. The absorption pattern of molecules in this region is normally strongly characteristic, leading to its use as a fingerprint. However, there are functional group modes that occur in this region, some of which are readily identifiable by virtue of their intensity.

Of particular interest are the absorption bands that arise from the deformation modes of aromatic C−H moieties, especially the out-of-plane ring δ(CH) modes. These show large intensity as a result of both high amplitude of vibration of the H atoms and the large dynamic dipole that arises from the aromatic electron density. Characteristic patterns are observed for the various substitution patterns around aromatic rings. Typically, they lie between 985 and 750 cm^{-1}, with some modes being quite intense. Although they are generally rather insensitive to the nature of the substituent, the pattern of modes observed in the spectrum usually depends upon the number of adjacent hydrogen atoms on the ring (i.e., the number of substituents and where they are with respect to one

Tab. 26.2 Spectral regions in which typical functional group modes can be identified.

Spectral Region (cm^{-1})	**Type**	**Comments/examples**
4000–2300	Hydrogenic stretching modes	H lightest atom, no X=H bonds
2400–1900	Triple bonds and cumulated double bonds	Nitriles, ν(CN) ≈ 2250 cm^{-1} Allenes, ν(−C=C=CH$_2$) ≈ 1940 cm^{-1}
2000–1400 cm^{-1}	Double bonds	Cyano, ν(>C=N-)≈ 1680 cm^{-1} Benzene rings ≈ 1600, 1580, 1500, 1470 cm^{-1} Carbonyls, ν(>C=O) ≈ 1900–1500 cm^{-1}
1600–1000 cm^{-1}	Deformation (bending) modes to hydrogen	Methylene, δ(CH$_2$) ≈ 1450–1480 cm^{-1} Methyl, δ(CH$_3$) ≈ 1450, 1380 cm^{-1} amine, δ(NH$_2$) ≈ 1600 cm^{-1}
1500–400 cm^{-1}	Heavier atom stretches or light atom deformations	Aromatic out of plane ≈ 985–750 cm^{-1} C−O single bonds ≈ 1300–1050 cm^{-1}

Tab. 26.3 Vibrational frequencies for the δCH(op) in phase wagging mode.

Substitution	Wavenumber (cm^{-1})	Adjacent H
Mono	766–736	5 adjacent H[a]
o-di	758–744	4 adjacent H
m-di	791–773	3 adjacent H[a]
p-di	830–804	2 adjacent H

[a] Monosubstituted and *m*-disubstituted benzenes usually show the out-of-plane sextant ring bending mode at approximately 700 cm^{-1} as well.

another). This behavior is summarized in Table 26.3.

In summary, the interpretation of infrared spectra of single component species relies strongly upon the expertise of the analyst. However, once assigned, variations in band position, intensity, linewidth, or band shape from known reference materials in well-defined states convey a large quantity of subtle information.

Acknowledgments

The author gratefully acknowledges the cooperation of JASCO UK and PIKE Technologies for supplying the images used in a range of figures for this work. He also thank Juan Najera of the University of Manchester for prepublication copies of Figures 26.6 and 26.21.

References

Allara, D.L. and Nuzzo, R.G. (1985) *Langmuir*, **1**, 52.

Atkins, P.W. and de Paula, J. (2009) *Physical Chemistry*, 9th edn. Oxford University Press, Oxford.

Belton, P. and Wilson, R.H. (1990) Infrared sampling methods, in *Perspectives in Modern Chemical Spectroscopy*, (ed. D.L. Andrews), Springer-Verlag, Berlin.

Berreman, D.W. (1963) *Phys. Rev.*, **130**, 2193.

Bradshaw, A. and Schweizer, E.K. (1988) IRRAS studies of metal Surfaces, in *Spectroscopy of Surfaces*, (eds R.J.H. Clark and R.E. Hester), John Wiley & Sons, Ltd, Chichester.

Bohren, C.F. and Huffman, D.R. (1983) *Absorption and Scattering of Light by Small Particles*, John Wiley & Sons, Ltd, New York.

Buijs, H. (2002) in *Handbook of Vibrational Spectroscopy*, Vol. 1, (eds J. Chalmers and P.R. Griffiths), John Wiley & Sons, Ltd, Chichester.

Chabal, Y.J. (1986) *Surf. Sci.*, **168**, 594.

Chalmers, J. and Griffiths, P.R. (eds) (2002) *Handbook of Vibrational Spectroscopy*, John Wiley & Sons, Ltd, Chichester.

Chen, A. and Lipkowski, J. (1999) *J. Phys. Chem. B.*, **103**, 682.

Coates, J. (1998) in *Applied Spectroscopy: A Compact Reference For Practitioners*, (eds J.R. Workman Jr and Art.W. Springsteen), Academic Press, San Diego.

Fitzpatrick, J. and Reffner, J.A. (2002) in *Handbook of Vibrational Spectroscopy*, Vol. 2, (eds J. Chalmers and P.R. Griffiths), John Wiley & Sons, Ltd, Chichester.

Gazi, E., Dwyer, J., Lockyer, N.P., Gardner, P., Shanks, J.H., Roulson, J., Hart, C.A., Clarke, N.W., and Brown, M.D. (2007) *Anal. Bioanal. Chem.*, **387**, 1621.

Greenler, R.G. (1966) *J. Phys. Chem.*, **44**, 310.

Griffiths, P.R. and de Haseth, J.A. (1986) *Fourier Transform Infrared Spectrometry*, John Wiley & Sons, Ltd, New York.

Hannah, R.W. (2002) in *Handbook of Vibrational Spectroscopy*, Vol. 2, (eds J. Chalmers and P.R. Griffiths), John Wiley & Sons, Ltd, Chichester.

Harrick, N.J. (1967) *Internal Reflection Spectroscopy*, John Wiley Interscience, New York.

Heavens, O.S. (1955) *Optical Properties of Thin Solid Films*, Butterworths, London.

Heraud, P., Caine, S., Sanson, G., Gleadow, R., Wood, B.R., and McNaughton, D. (2006) *New Phytol.*, **173**, 216.

Hollas, J.M. (1992) *Modern Spectroscopy*, 2nd edn, John Wiley & Sons, Ltd, Chichester.

Hollins, P. (2000) Infrared reflection-absorption spectroscopy, in *Encyclopedia of Analytical Chemistry*, (ed. R.A. Meyers), John Wiley & Sons, Ltd, Chichester.

Horn, A.B. (1998) Infrared spectroscopic techniques for the study of thin interfacial films, in *Spectroscopy for Surface Science* (eds R.J.H., Clark and R.E., Hester), John Wiley & Sons, Ltd, Chichester.

Horn, A.B. and Sully, K.J. (1997) *J. Chem. Soc.-Faraday Trans.*, **93**, 2741.

Humecki, H.J. (ed.) (1995) *Practical Guide to Infrared Microspectroscopy. Practical Spectroscopy Series*, Vol. 19, Marcel Dekker, New York.

Jackson, R.S. (2002) in *Handbook of Vibrational Spectroscopy*, Vol. 1, (eds J. Chalmers and P.R. Griffiths), John Wiley & Sons, Ltd, Chichester.

Kauppinen, J. and Partanen, J. (2001) *Fourier Transforms in Spectroscopy*, Wiley-VCH Verlag GmbH, Berlin.

McIntyre, J.D.E. and Aspnes, D.E. (1971) *Surf. Sci.*, **24**, 417.

Menard, H., Tear, S.P., and Horn, A.B. (2008) *J. Phys. Cond. Matter., 20*, 355002.

Milosevic, M. and Berets, S.L. (2002) Accessories and sample handling for mid-infrared diffuse reflection spectroscopy, in *Handbook of Vibrational Spectroscopy*, Vol. 2, (eds J. Chalmers and P.R. Griffiths), John Wiley & Sons, Ltd, Chichester.

Mirabella, F.M. (2002) Principles, theory and practice of internal reflection spectroscopy, in *Handbook of Vibrational Spectroscopy*, Vol. 2, (eds J. Chalmers and P.R. Griffiths), John Wiley & Sons, Ltd, Chichester.

Najera, J.J., Fochesatto, J.G., Last, D.J., Percival, C.J., and Horn, A.B. (2009) *Rev. Sci. Instrum., 75*, 124102.

Pearce, H.A. and Sheppard, N. (1976) *Surf. Sci.*, **59**, 205.

Pereira, M. and Yarwood, J. (1996) *J. Chem. Soc. Faraday Trans.*, **92**, 2737.

Pritchard, J. and Sims, M.L. (1970) *Trans. Faraday Soc.*, **66**, 427.

Socrates, G. (2004) *Infrared and Raman Characteristic Group Frequencies: Tables and Charts*, 3rd edn, John Wiley & Sons, Ltd, Chichester.

Sommer, A.J. (2002) Mid-infrared transmission microspectroscopy, in *Handbook of Vibrational Spectroscopy*, Vol. 1, (eds J. Chalmers and P.R. Griffiths), John Wiley & Sons, Ltd, Chichester.

Theocharous, E. and Birch, R. (2002) in *Handbook of Vibrational Spectroscopy*, Vol. 1, (eds J. Chalmers and P.R. Griffiths), John Wiley & Sons, Ltd, Chichester.

Vincent, A. (2001) *Molecular Symmetry and Group Theory*, 2nd edn, John Wiley & sons, Ltd, Chichester.

Vogt, R. and Finlayson-Pitts, B.J. (1994) *J. Phys. Chem.*, **98**, 3747.

Williams, D.H. and Fleming, I. (1995) *Spectroscopic Methods in Organic Chemistry*, McGraw Hill Publishing, New York.

Willis, H.A. and Powell, D.B. (1990) Recent advances in vibrational spectroscopy, in *Perspectives in Modern Chemical Spectroscopy* (ed. D.L. Andrews), Springer-Verlag, Berlin.

Workman, J. Jr (1998) in *Applied Spectroscopy: A Compact Reference For Practitioners* (eds J.R. Workman Jr and Art.W. Springsteen), Academic Press, San Diego.

Zasetsky, A.Y., Earle, M.B., Cosic, B., Schiwon, R., Grishin, I.A., McPhail, R., Panescu, R.G., Najera, J.J., Khaliziv, A.F., Cook, A.B., and Sloan, J.J. (2007) *J. Quant. Spectrosc. Rad. Transf.*, **107**, 294.

Further Reading

Chalmers, J. and Dent, G. (1997) *Industrial Analysis with Vibrational Spectroscopy*, Royal Society of Chemistry, Cambridge.

Nakamoto, K. (1997a) *Infrared and Raman Spectra of Inorganic and Coordination Compounds. Part A: Theory and Applications in Inorganic Chemistry*, Wiley Interscience, New York.

Nakamoto, K. (1997b) *Infrared and Raman Spectra of Inorganic and Coordination Compounds. Part B: Applications in Coordination, Organometallic and Bioinorganic Chemistry*, Wiley Interscience, New York.

Sodeau, J.R. (1995) Atmospheric cryochemistry, in *Spectroscopy in Environmental Science* (eds R.J.H. Clark and R.E. Hester), John Wiley & Sons, Ltd, Chichester.

Tranter, G.E. and Holmes, J.L. (eds.) (2007) *Encyclopedia of Spectroscopy and Spectrometry*, Elsevier, Amsterdam.

Workman, J.R. Jr and Springsteen, Art.W. (eds) (1998) *Applied Spectroscopy: A Compact Reference For Practitioners*, Academic Press, San Diego.

28

Biomolecular Structures by Solution Nuclear Magnetic Resonance

Tharin M. A. Blumenschein

28.1 Introduction 965
28.2 Spectra Acquisition 966
28.2.1 Nuclei 966
28.2.2 Sample Preparation 967
28.2.3 Pulse Sequences and Processing Strategies 968
28.3 Chemical Shift Assignment 969
28.3.1 Chemical Shift Assignment in Proteins 970
28.3.2 Chemical Shift Assignment in Nucleic Acids 973
28.3.3 Chemical Shift Assignment in Carbohydrates 973
28.4 Structural Restraints 975
28.4.1 Nuclear Overhauser Effect 975
28.4.2 Torsion Angles 977
28.4.3 Residual Dipolar Couplings (RDCs) 978
28.4.4 Hydrogen Bonds 979
28.4.5 Paramagnetic Centers 980
28.4.6 Structural Information from Chemical Shifts 980
28.5 Molecular Modeling and Structure Calculation 981
References 982
Further Reading 986

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.

Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

ISBN: 978-3-527-40773-6

28.1
Introduction

Molecular structures are essential to understand the function and mechanism of molecules. The use of nuclear magnetic resonance (NMR) spectroscopy to determine the structure of biomolecules in solution has the advantage of reproducing the environment in which these molecules are naturally found.

The aim of this chapter is to provide an introduction to how NMR spectroscopy is applied to structural studies of biomolecules, following the steps described below. It is not the aim to provide an extensive review of structure determination of biological macromolecules using solution NMR, and although NMR spectroscopy can be successfully used to study the dynamics of molecules and molecular interactions (see, for example, Palmer, 2001; Kay, 2005; Latham *et al.*, 2005), these techniques are beyond the scope of this introduction.

Biological macromolecules have a number of characteristics in common, and they present a number of common challenges. The process of structure determination by NMR usually starts with the assignment of chemical shift values to the relevant nuclei in the molecule, followed by the determination of a number of structural restraints. Specific pulse sequences are available for each of those steps. The structural restraints are then used for molecular simulations, using simulated annealing techniques and molecular dynamics. Each type of structural restraint is added to the molecular energy representation as a separate pseudo-energy term, and the simulation explores the conformational space to find the minimal energy structures.

As a solution technique, NMR-determined structures reflect the heterogeneity of conformations presented by biomolecules in solution. Specific challenges in this aspect are intrinsically disordered proteins (see Dyson and Wright, 2005, for a review), which do not have a well-defined tertiary structure in their native state in the absence of their binding partner, and larger oligosaccharides and polysaccharides, that tend to be highly flexible and dynamic (Martin-Pastor and Allen Bush, 1999). In these cases, the structures determined reflect a distribution of conformations, rather than a single structure (Yi *et al.*, 2004; Lindorff-Larsen *et al.*, 2004; Osterhout *et al.*, 1989).

When compared to small organic molecules, the acquisition of NMR spectra of biological macromolecules

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

presents a number of challenges. Experiments are generally performed in water, making good suppression of the solvent signal essential. In addition, the number of signals from the different ^{1}H nuclei in a biological macromolecule inevitably leads to overlapping in the one-dimensional spectrum, requiring high-field spectrometers and multidimensional spectra to resolve the individual resonances.

28.2
Spectra Acquisition

28.2.1
Nuclei

The most common atoms in biological macromolecules are carbon, hydrogen, nitrogen, and oxygen. The magnetic properties of their stable isotopes are summarized in Table 28.1.

The most abundant hydrogen isotope, ^{1}H, has the highest sensitivity, and is the most widely used nucleus for NMR. The most abundant carbon isotope, ^{12}C, has no nuclear spin, whereas the most abundant nitrogen isotope, ^{14}N, has a spin number of 1 and strong quadrupolar moment, leading to fast relaxation. Less abundant isotopes of both, however, have spin number ½. ^{13}C is present as about 1% of the total carbon, and natural abundance spectra are routinely acquired for small molecules. For larger molecules such as proteins, the weak signal would make the acquisition of multidimensional spectra overly long, and ^{13}C labeling of proteins is a standard technique. The same is true of ^{15}N, even less abundant naturally. The only stable magnetically active oxygen isotope is ^{17}O, with a spin number of 5/2, and it is not normally used for solution NMR of biomolecules. ^{31}P is naturally present in all nucleic acids and in some carbohydrates and proteins, and is often used as an additional source of information. Although other nuclei can be used to study biomolecules, structural determination is usually performed using

Tab. 28.1 Magnetic properties of the stable isotopes of biologically abundant atoms. γ is the gyromagnetic ratio.

Isotope	Spin	Natural abundance (%)	γ (10^7 rad T^{-1} s^{-1})	Sensitivity		Frequency at 11.7 T
				Relative to proton	At natural abundance	
^{1}H	½	99.985	26.75	1.00	1.00	500.000
^{2}H	1	0.015	4.11	9.65×10^{-2}	1.45×10^{-6}	76.753
^{12}C	0	98.89				
^{13}C	½	1.108	6.73	1.59×10^{-2}	1.76×10^{-4}	125.721
^{14}N	1	99.63	1.93	1.01×10^{-3}	1.01×10^{-3}	36.118
^{15}N	½	0.366	−2.71	1.04×10^{-3}	3.85×10^{-6}	50.664
^{16}O	0	99.75				
^{17}O	5/2	0.037	−3.63	2.91×10^{-2}	1.08×10^{-5}	67.784
^{18}O	0	0.205				
^{31}P	½	100	10.83	6.63×10^{-2}	6.63×10^{-2}	202.405

multidimensional ^1H, ^{15}N, and ^{13}C spectra. ^{15}N and ^{13}C are indirectly detected, with the magnetization being transferred back to protons for ^1H direct detection. These experiments require triple-resonance probes, capable of pulsing and decoupling ^{13}C and ^{15}N, sometimes simultaneously, while pulsing and detecting ^1H (Cavanagh *et al.*, 1996).

28.2.2
Sample Preparation

Solution NMR of biological macromolecules is usually performed in 90% H_2O/10% D_2O, with the buffer and salt concentrations depending on the requirements for stability of the molecules being studied. Where possible, sample pH should be kept below 7.5 to slow down the exchange rate of the backbone amide protons with the solvent, which leads to loss of signal, and salt concentration should be kept low and buffers of low conductivity should be chosen, as higher buffer conductivity will decrease the signal-to-noise ratio (Kelly *et al.*, 2002).

The large number of atoms in biological macromolecules results in heavy overlapping of resonances in the ^1H spectra, requiring uniform labeling with ^{15}N and/or ^{13}C for resolution of the overlapped peaks in multidimensional spectra. For larger molecules, additional uniform labeling with ^2H is required to decrease transverse relaxation rates (Grzesiek *et al.*, 1993). Deuterated protein samples are usually purified in ^1H_2O-based buffers, resulting in substitution of the deuterium bound to the main-chain nitrogen with protons.

The requirement for labeled protein means that protein samples for NMR spectroscopy are normally obtained by recombinant expression in an organism that can be grown in a medium of controlled composition, to which the only source of nitrogen or carbon can be added with defined isotopic composition. Most commonly, proteins are expressed in *Escherichia coli* grown in minimal media, having glucose as the only source of carbon, and NH_4Cl or $(NH_4)_2SO_4$ as the only source of nitrogen. Uniformly labeled ^{13}C-glucose and $^{15}NH_4Cl$ and $(^{15}NH_4)_2SO_4$ are available commercially. For deuterated proteins, the minimal media can be prepared with D_2O instead of ^1H_2O, and conditioning protocols may be required for good bacterial growth and protein expression (Li, Corson and Sykes, 2002). Rich media containing ^{15}N, ^{13}C, and ^2H are also commercially available, and generally produce much better bacterial growth than minimal media, but are significantly more expensive. When a protein cannot be successfully expressed in bacterial cells, eukaryotic expression systems can be used, such as the yeast *Pichia pastoris* (Morgan, Kragt and Feeney, 2000) or insect cells (Brüggert *et al.*, 2003). More recently, cell-free expression systems have been developed (Vinarov *et al.*, 2004; Torizawa *et al.*, 2004).

For larger proteins or specific experimental requirements, alternative labeling schemes can be required, instead of uniform labeling. Examples include different schemes for selective protonation of methyl groups in amino acid side chains, while the rest of the protein is deuterated (Gardner and Kay, 1998).

For nucleic acids, it is also necessary to obtain pure ^{13}C, ^{15}N-labeled samples in milligram amounts. Most RNA samples are produced *in vitro* by transcription of a DNA template by T7 polymerase (Milligan *et al.*, 1987; reviewed by Fürtig *et al.*, 2003; Flinders and Dieckmann, 2006; Dayie, 2008). Template DNA often includes the sequence for a self-cleaving ribozyme, which will aid in purification and help in overcoming the limitations of T7 polymerase. Uniform labeling is commonly performed by using labeled NTPs as the substrate for the T7 polymerase reaction, but for large RNAs (over 20–30 nucleotides), specific labeling approaches may be required (reviewed in Latham *et al.*, 2005; see Nelissen *et al.*, 2008, for a recent example). Chemical synthesis is an alternative for small RNAs, and has been reviewed by Lagoja and Herdewijn (2002). DNA samples can be obtained by purifying a suitable plasmid containing multiple copies of the desired oligonucleotide from a bacterial culture grown in ^{15}N, ^{13}C-labeled medium, or from a polymerase chain reaction (PCR)-based approach (Louis *et al.*, 1998; René *et al.*, 2006).

Labeled carbohydrate samples for NMR studies can be even more difficult to produce, because the residue sequence in glycans often depends on species-specific synthetic pathways. NMR of small oligosaccharides can be performed on natural abundance ^{13}C, and is often used in combination with other techniques. Glycans from species that can be grown in labeled medium, such as bacteria and fungi, can be normally purified. A review of successful examples of uniform labeling in carbohydrates was done by Duus, Gotfredsen and Bock (2000), but a labeling method generally applicable to carbohydrates is not yet available. Alternative approaches for partial labeling include the enzyme-catalyzed exchange of acetyl groups with ^{13}C-enriched acetic anhydride (Yu and Prestegard, 2006).

28.2.3
Pulse Sequences and Processing Strategies

While specific pulse sequences used to determine chemical shift assignments or structural restraints are mentioned in the following sections, a few general methods can be applied to a range of pulse sequences to make them more effective or faster. Pulse sequences for biological experiments also need to contain solvent suppression pulses, because the experiments are normally performed in 90% H_2O/10% D_2O. Different water suppression techniques can be used in different pulse sequences, and depending on the characteristic of the sample. When observing protons that can easily exchange with the solvent, in nucleic acids for instance, it is important to avoid methods involving presaturation of water (more commonly used in protein studies), and utilize pulse sequences that use jump-and-return techniques (Hore, 1983; Sklenár and Bax, 1987) or the WATERGATE method (Piotto, Saudek and Sklenár, 1992).

Original multidimensional pulse sequences relied on phase cycling to eliminate unwanted signals and minimize artifacts. By changing the phase of specific pulses and the receiver, unwanted signals can be cancelled. This requires repeating the experiment with the different phase combinations, and can result in a number of repetitions much larger than needed for a good signal-to-noise ratio. In probes with pulsed-field gradient capabilities, part of the phase cycling can be replaced by the use of gradient pulses. Field gradient pulses are pulses that create a varying field along the axis of the sample, resulting in defocusing of the magnetization, which can be recovered by applying another gradient pulse, of opposite signal. If those pulses are applied at the right point in the pulse sequence, only the desired signals will be selected. Phase cycling is then used to minimize artifacts, with a smaller number of repetitions.

For larger biomolecules, one of the limiting factors is the fast transverse relaxation, which increases with the rotational correlation time. Fast transverse relaxation broadens the spectral lines, and leads to loss of signal due to relaxation during the pulse sequence. However, the four lines in an undecoupled peak quartet in a heteronuclear correlation spectrum do not relax at the same rate. Transverse relaxation optimized spectroscopy (TROSY) selects the multiplet component with slower relaxation rate, eliminating the other lines (Pervushin *et al.*, 1997; Wider and Wüthrich, 1999). Although only part of the signal is being recovered, the spectral lines show better resolution and signal-to-noise than the comparison standard PEP-HSQC (heteronuclear single quantum coherence) spectrum used for heteronuclear correlation spectroscopy, and this effect is intensified for large molecules. The TROSY effect is field-dependent, being more pronounced at higher fields. It is normally used in conjunction with deuterated samples, which also help to increase the transverse relaxation times. TROSY elements have been incorporated in a number of multidimensional pulse sequences, for use with large molecules (Wider and Wüthrich, 1999).

While TROSY and other techniques aim to increase sensitivity and extend NMR experiments to large and otherwise troublesome proteins, there are various techniques to decrease experimental time when signal is abundant. They include SOFAST-HMQC experiments (Schanda, Kupce and Brutscher, 2005), Hadamard NMR (Kupce, Nishida and Freeman, 2003; Brutscher, 2004), single-scan NMR (Frydman, Scherf and Lupulescu, 2002), projection NMR (Szyperski *et al.*, 1993; Brutscher *et al.*, 1994; Kupce and Freeman, 2004), hyperdimensional NMR (Kim and Szyperski, 2003; Kupce and Freeman, 2008), nonlinear data recording and processing (Mandelshtam, 2000; Hoch and Stern, 2001; Rovnyak *et al.*, 2004), and covariance spectroscopy (Brüschweiler and Zhang, 2004). A number of these techniques were reviewed a few years ago (Freeman and Kupce, 2003).

28.3
Chemical Shift Assignment

There is no single unified method for obtaining chemical shift assignments for polypeptides, polynucleotides, and polysaccharides. Proteins, nucleic acids, and carbohydrates have different characteristics, and the methods for chemical shift assignment differ, even though some of the pulse sequences used are the same. Most of the pulse sequences required for the assignment of biomolecules are provided in the standard installation of modern high-field NMR spectrometers. Chemical shift assignments of biological macromolecules are deposited in the Biological Magnetic Resonance Bank (BMRB) (Ulrich *et al.*, 2008), which is also a useful source for other resonance data, such as average chemical shift values for each amino acid type. Once the chemical shift values seen in the spectra have been assigned to specific nuclei in the biomolecule, structural restraints can be defined for different atoms in the molecule.

28.3.1
Chemical Shift Assignment in Proteins

There are two stages in assigning chemical shift values to individual nuclei in proteins. First, the backbone is assigned using the sequential assignment method, created by Kurt Wütrich (Nobel prize winner in Chemistry in 2002) and coworkers (Wagner and Wüthrich, 1982; Wüthrich, 1986; reviewed in Ferentz and Wagner, 2000; Kanelis, Forman-Kay and Kay, 2002). The side chains of each residue are assigned in a second stage. Modern experiments used for chemical shift assignment are mostly ^{15}N-edited three-dimensional spectra, and are intimately connected to the labeling scheme used for the sample. For most proteins, ^{1}H, ^{13}C, and ^{15}N information is required, from a sample double-labeled with ^{13}C and ^{15}N. For large proteins, it is often not possible to perform three-dimensional NMR experiments in a fully protonated protein, due to the short transversal relaxation time (T_2), which leads to very fast signal decay and line broadening. Main-chain amide protons are usually back-exchanged into deuterated proteins, and specific suites of triple-resonance experiments are the basis for chemical shift assignment (Gardner and Kay, 1998).

Figure 28.1 shows schemes of the magnetization transfer in the most commonly used pulse sequences for backbone sequential assignment in proteins. Sequential assignment starts with a reference spectrum, a 2D {^{1}H, ^{15}N}-HSQC in which a peak corresponds to each main-chain amide, or a 3D HNCO (Kay *et al.*, 1990; Grzesiek and Bax, 1992b), which correlates the proton and nitrogen of a main-chain amide with the carbonyl carbon from the previous amino acid residue. An ideal HSQC or HNCO spectrum will have a peak for every residue in the protein, but N-terminal residues often cannot be seen due to fast exchange with water protons, and loss of magnetization during the water suppression scheme. The faster exchange in this region is caused by the positive charge in the amino terminus of the protein. Conversely, C-terminal residues often give much stronger signals than the rest of the protein. Conformational exchange can also lead to peak broadening beyond detection, and peaks missing from the reference spectrum. Conformational exchange in the protein can also lead to the loss of peaks for some residues.

The basic experiments for backbone assignment of a medium-sized protein are HNCACB (Wittekind and Mueller, 1993) and CBCA(CO)NH (Muhandiram and Kay, 1994; Grzesiek and Bax, 1992a), HNCO and HN(CA)CO (Clubb, Thanabal and Wagner, 1992). HNCACB correlates the amide proton and nitrogen of each residue with the α- and β-carbons of the same amino acid residue more strongly, and with the α- and β-carbons of the previous residue more weakly, while CBCA(CO)NH correlates the amide proton and nitrogen of each residue with the α- and β-carbons of the previous residue only. Sequential assignment is done by matching the weaker peaks in each strip of the HNCACB spectrum with the peaks in a strip of the CBCA(CO)NH spectrum, forming a "chain" of spin systems. Residues with characteristic chemical shift values, such as serine, threonine, glycine, and alanine, help anchor this chain in relation to the amino acid sequence in the protein.

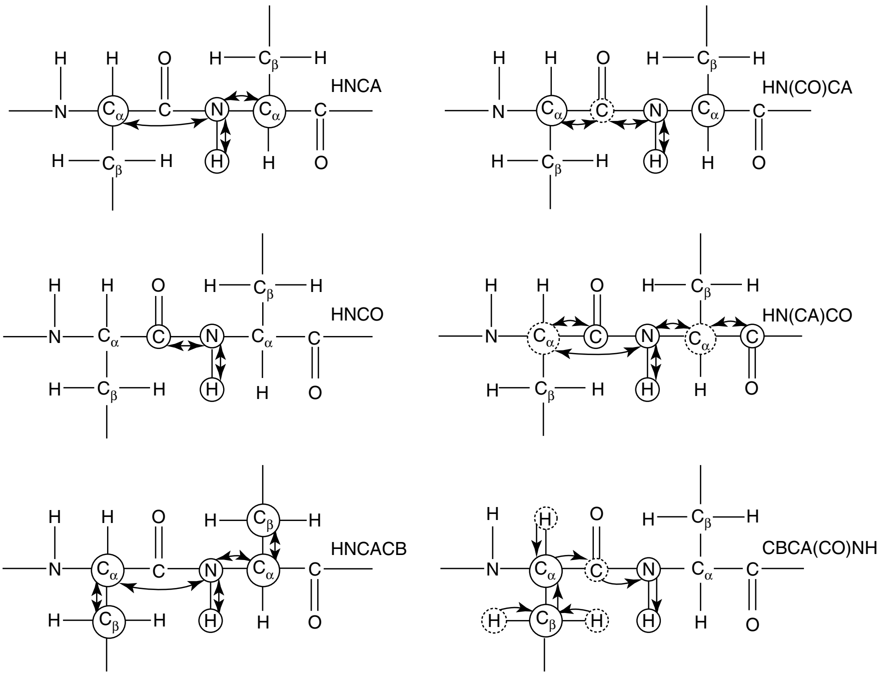

Fig. 28.1 Magnetization transfer steps in pulse sequences used for protein sequential assignment. Nuclei that are not correlated in the final spectra are represented in dashed circles, nuclei correlated in the final spectra are inside solid circles, and magnetization transfer is represented by arrows. With the exception of CBCA(CO)NH, the magnetization in all pulse sequences in the figure starts from the amide proton is transferred to the heteronuclei and back to the amide proton for data acquisition. In CBCA(CO)NH, the magnetization starts in the protons bound to C_β and C_α.

An example of sequential strips in HNCACB and CBCA(CO)NH spectra can be seen in Figure 28.2.

The CBCA(CO)NH experiment is not applicable to deuterated proteins, because it starts by transferring magnetization from the hydrogen atoms bound to the β-carbon (Figure 28.1). When the ^1H is replaced with deuterium, this pulse sequence is not applicable anymore. Instead, the HN(CO)CACB pulse sequence can be used. In large proteins (over 30 kDa), however, the signal that can be obtained with HNCACB and HN(CO)CACB is often too weak, and the combination of HNCA (Grzesiek and Bax, 1992b; Stonehouse *et al.*, 1995) and HN(CO)CA (Bax and Ikura, 1991) experiments can used. Only correlating the α-carbon with the amide, these experiments have better signal-to-noise ratios than HNCACB and HN(CO)CACB under the same conditions. The disadvantage, however, is that they contain less information. Other labeling schemes and pulse sequences have been developed to extend chemical shift assignment procedures to even larger proteins, up to 82 kDa (Tugarinov, Hwang and Kay, 2004).

A number of semiautomated and automated backbone assignment programs

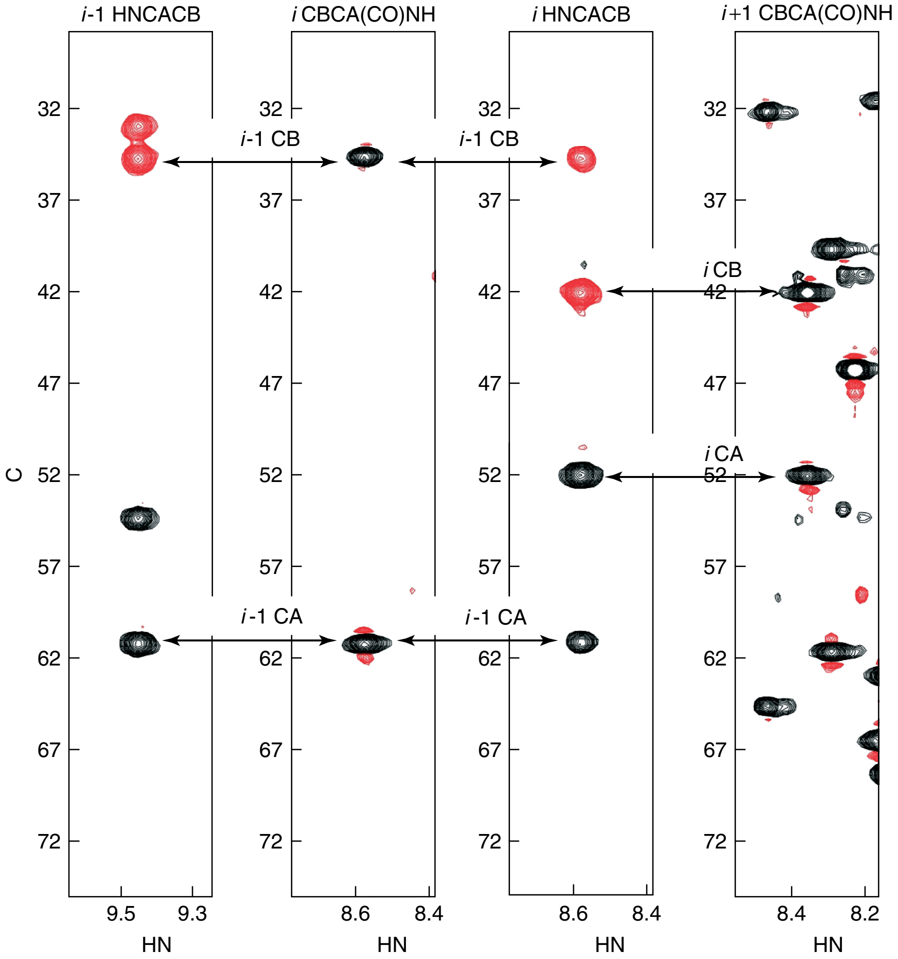

Fig. 28.2 Strips from three-dimensional HNCACB and CBCA(CO)NH spectra for sequential assignment, showing the connectivities between three sequential residues, $i - 1$, i, and $i + 1$. Each strip is a region of a spectrum corresponding to the amide nitrogen plane and proton resonance for the indicated residue. As seen in Figure 28.1, HNCACB spectra correlate the amide of each residue with the α- and β-carbons from the same and previous residues, while CBCA(CO)NH spectra correlate the amide of each residue with the α- and β-carbons from the previous residue. Data was kindly provided by Prof. Brian Sykes, University of Alberta, Canada.

have been developed. Semiautomated assignment programs, such as ANSIG (Helgstrand *et al.*, 2000) and Smartnotebook (Slupsky *et al.*, 2003), suggest possible sequential assignments to be visually verified and confirmed by the user. Automated assignment programs produce the assignment for spin systems in the protein from its amino acid sequence and a peak list. Examples are PASTA (Leutner *et al.*, 1998), TATAPRO (Atreya, Chary and Govil, 2002), MARS (Jung and Zweckstetter, 2004), and MATCH (Volk, Hermann and Wüthrich, 2008).

Such methods still require some level of user input, and in most cases not all amino acid residues in a protein can be assigned automatically.

Once the backbone has been assigned, the side chains in the protein need to be assigned for structural determination. This is achieved through total correlation spectroscopy TOCSY-based experiments that correlate, through bonds, the nuclei already assigned to the rest of the side chain.

28.3.2
Chemical Shift Assignment in Nucleic Acids

Methods for chemical shift assignment of nucleic acids have been reviewed by Flinders and Dieckmann (2006), Tzakos *et al.* (2006), Fürtig *et al.* (2003), Cromsigt *et al.* (2001), Wijmenga and Van Buuren (1998), and Varani, Aboul-Ela and Allain (1996). These reviews also cover structural determination methods for nucleic acids.

Assignment of nucleic acids starts with the nucleobase spin systems, followed by correlation of each nucleobase to the corresponding sugar spin system, using HCN triple-resonance experiments in ^{13}C, ^{15}N-labeled nucleic acids. Assignment of the sugar spin system is then completed using a combination of HCCH-TOCSY and HCCH-COSY-based experiments (Flinders and Dieckmann, 2006). Sequential connections between nucleotides are made using nuclear overhauser effect spectroscopy (NOESY) experiments, or HCP and HCP-TOCSY experiments, which correlate the backbone ^{31}P nuclei to hydrogen and carbon atoms.

Base pairings can be assigned from the imino proton of guanines and uracyls, which have characteristic chemical shifts between 10 and 15 ppm. These protons exchange with water, and therefore can only be observed when involved in hydrogen bonding. A homonuclear 2D NOESY experiment correlates those resonances with neighboring protons, both from the nucleotide involved in the base pairing and from sequential nucleotides in the same strand (Tzakos *et al.*, 2006). HNN-COSY experiments correlate the nucleotides in a base pair using scalar couplings between the nitrogen atom of the hydrogen bond acceptor and the nitrogen and hydrogen of the hydrogen bond donor (Dingley and Grzesiek, 1998; Pervushin *et al.*, 1998). Both methods are illustrated in Figure 28.3. ^{13}C, ^{15}N edited or filtered 2D and 3D NOESY experiments are used to decrease overlap in NOESY spectra, and identify NOEs for assignment and structure determination (Fürtig *et al.*, 2003; Flinders and Dieckmann, 2006).

In single-stranded RNA, the experiments based on base pairing are not possible, and assignment is much more difficult. The lack of base stacking also leads to decreased spectral dispersion. Isotopic labeling of RNA is essential for sequential assignment of single-stranded regions, such as bulges in the RNA structure.

28.3.3
Chemical Shift Assignment in Carbohydrates

Carbohydrates are more diverse than either proteins or nucleic acids in their basic components and connections. A large number of sugars form the possible basic units of carbohydrates and the potential for branched carbohydrate chains indicate that there is not always a primary sequence in the way that polypeptides

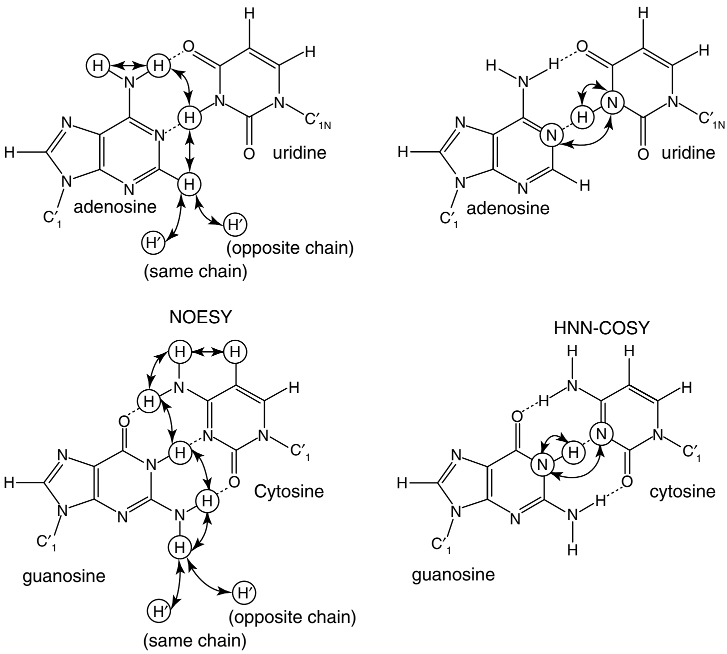

Fig. 28.3 Determination of base pairing in nucleic acids using NOESY or HNN-COSY spectra. Magnetization transfer steps are represented by arrows. In the NOESY experiment, magnetization transfer between protons occurs simultaneously in both directions of each arrow. During the HNN-COSY experiment, magnetization is transferred from the imino proton to the covalently bonded nitrogen, from there to the hydrogen-bonding nitrogen, and back to the imino proton for data acquisition.

and polynucleotides possess. This has led to different approaches to assignment being developed by different groups (Duus, Gotfredsen and Bock, 2000).

Some of the issues in chemical shift assignment of carbohydrates are related to the difficulty in obtaining labeled samples of carbohydrates. A number of experiments can be done using the natural abundance of ^{13}C, but this limitation greatly restricts the size of the molecules that can be studied. The labeling schemes described for carbohydrates in Section 28.2.1 can help to overcome this obstacle.

Unlike proteins and nucleic acids, the sequence of monomers in a carbohydrate is not clearly coded in the organism's genome. This means that the process of assignment often starts with the use of multiple techniques in addition to NMR, such as mass spectroscopy, to determine the sequence of sugar monomers in a carbohydrate.

A summary of the steps used for chemical shift assignment in carbohydrates was reviewed by Duus, Gotfredsen and Bock (2000). In short, the process for oligosaccharides starts by finding a characteristic chemical shift for each sugar monomer in the ^{1}H, and using experiments that correlate nuclei through bonds to assign the other atoms in each monomeric unit. Such experiments can include HSQC/HMQC/HMBC to correlate carbon and protons, TOCSY, DQF-COSY, and ^{13}C-edited HSQC-TOCSY. Intraresidue NOEs or ROEs (rotating-frame NOEs) can also be useful to assign individual resonances inside a monomer, and the anomeric conformation of a sugar can be determined from ^{13}C chemical shifts, and from a number of scalar coupling constants.

More recently, the H2BC (heteronuclear 2-bond correlation) experiment has been shown to provide unambiguous assignments and simplify carbohydrate spectra by correlating protons and carbons over two bonds (Petersen *et al.*, 2006).

A strategy that can be used for carbohydrate assignment is to decrease the dimensionality of the experiment, decreasing acquisition time and focusing on a desired region. With the use of selective pulses, a 1D spectrum can be acquired corresponding to the trace of the 2D spectrum in a specific frequency, reflecting the same connectivities. Similarly, a 2D plane of a 3D spectrum can be acquired, focusing in a specific region of the spectrum (Uhrin and Barlow, 1997; Roumestand *et al.*, 1999).

28.4
Structural Restraints

28.4.1
Nuclear Overhauser Effect

The nuclear overhauser effect (NOE) (Overhauser, 1953) arises from cross-relaxation processes that allow the transfer of magnetization from a nucleus to other nearby nuclei, through dipolar relaxation mechanisms. The build-up of an NOE between two spins i and j is given by the cross-relaxation rate, σ_{ij}, which is the difference between the rate for the flip-flip transition, W_2 (in which both spins are flipped in the same sense), and the rate for the flip-flop transition, W_0 (in which the spins are flipped in opposite senses):

$$\sigma_{ij} = (W_2 - W_0) \qquad (28.1)$$

In the homonuclear case, the resonance frequency of the two spins can be considered to be the same, ω_0, and the cross-relaxation rate will be

$$\sigma_{ij} \propto \frac{1}{r_{ij}^6}\left\{6J(2\omega_0) - J(0)\right\} \qquad (28.2)$$

where r_{ij} is the internuclear distance, $J(2\omega_0)$ is the spectral density function at twice the resonance frequency of the spins (for W_2), and $J(0)$ is the spectral density function at frequency zero (for W_0).

As it can be seen above, NOE build-up rates are proportional to the inverse of the sixth power of the distance, and therefore can be used to calculate distance restraints for structural calculations. NOESY experiments are used to measure the NOE build-up rates, and have a mixing time during which the magnetization evolves according to the NOE. However, the linearity between NOE build-up rates and distance is only complete when the mixing time is zero. In reality, the mixing time is kept small enough to allow this linearity approximation to be used. As the size of the molecule increases, and with it the rotational correlation time, dipolar relaxation mechanisms become more efficient, and the linear approximation is valid for shorter times. Spin diffusion (in which a peak in the NOESY spectrum is created via an indirect relaxation path) also increases with mixing time, as does intensity of the NOE signal before other relaxation processes cause it to diminish. Therefore, mixing times need to be optimized according to the characteristics of the molecule.

Multidimensional NOESY spectra correlate nuclei that are close in space, independently of being connected by bonds. Two-dimensional NOESY spectra correlate protons through space, and are useful for the structural determination of relatively small molecules, including small proteins. As the number of hydrogens in a molecule grows, so does the number of peaks, which start to overlap in a two-dimensional spectrum, and three-dimensional ^{15}N- or ^{13}C-edited NOESY spectra become necessary for resolving the signals. In the three-dimensional spectrum, the peaks from a two-dimensional NOESY spectrum are spread into planes according to the chemical shift of the nitrogen or carbon to which one of the protons is attached (Marion *et al.*, 1989a; Marion *et al.*, 1989b; Zuiderweg and Fesik, 1989).

Each peak in a NOESY spectrum represents a correlation through space between two nuclei, independently of a connection through bonds, and can be used to determine distance restraints for structure calculation. After the peaks in a NOESY spectrum are assigned to the corresponding atoms in the molecule, distance restraints are created for the nuclei correlated by each peak. This step can be time consuming, and often requires multiple rounds of interaction, because commonly many NOEs cannot be unambiguously assigned with chemical shift information only. This problem worsens with increasing protein size, or chemical shift imprecision (Mumenthaler *et al.*, 1997). To circumvent this problem, a preliminary structure is often calculated based on the unambiguously assigned residues, and this structure is then used as additional information to resolve NOE assignment ambiguities, based on atom proximity in the preliminary structure. Alternatively, ambiguous assignments can be used if the different assignment possibilities are taken into account (Nilges, 1995).

This process can be automated in available software for the structural determination of proteins using iterative rounds of automated NOESY assignment, distance restraint determination, and structure calculation. Two examples of this type of software are CANDID (Herrmann, Güntert and Wüthrich, 2002) and ARIA (Linge *et al.*, 2003), which differ amongst other things in their strategies for dealing with multiple assignment possibilities in the calculation process.

Owing to the relationship between distance and build-up rates, the strength of the NOE peaks decreases rapidly as the distance between nuclei increases (Wüthrich, 1986). In practice, that means that peaks in a NOESY spectrum of a biological macromolecule represent a correlation between two nuclei that are no more than 6 Å apart, and that peaks in the same spectrum can be calibrated according to their intensity to reflect the distance between the two nuclei being correlated. A common method for calibrating the distance according to the peak volume involves dividing the peaks into three classes (weak, medium, strong), each corresponding to a range of distances (Clore *et al.*, 1985; Williamson *et al.*, 1985). A different method involves assigning peaks corresponding to nuclei with known distances (such as between the amide proton of one residue, and the α-proton of the previous residue), and use that as a reference to calibrate the remaining distances (Gagné *et al.*, 1997). In either method, it is better to set the upper limit of each distance too large than too small. Distance restraints that are too tight result in distorted molecular structures, while the presence of enough restraints will lead to the correct structure, even if they are too loose compared to the actual distances (Wijmenga and Van Buuren, 1998; Clore and Gronenborn, 1991).

Distance restraints based on NOE correlations are the staple of protein structure determination by solution NMR, and many proteins structures have been determined relying exclusively on them. For nonglobular macromolecules, however, they often do not offer enough information for a well-defined structure. NOE distance restraints do not extend beyond 6 Å and a good structure requires a large number of distance restraints between atoms that are distant from each other in the original residue sequence, but close in space. For elongated structures, cumulative uncertainties in the short-distance NOE restraints can lead to large uncertainties in the structure as a whole, requiring the use of other structural restraints. Examples of this situation are nucleic acids, in which pairwise interactions do not provide the same "long-range" restraints, carbohydrates, and some elongated helical proteins. In those cases, the use of other structural restraints discussed below becomes essential.

When the molecule needs to be deuterated to increase the transverse relaxation time, ^{1}H nuclei will only be available at positions that exchange with water. In this case, the low number of hydrogen nuclei limits the number of NOE restraints. To increase the number of protons available for distance restraints in proteins, while retaining the relaxation advantages of a mostly deuterated protein, different labeling schemes have been developed, that lead to deuteration of most of the protein, with hydrogens present in methyl groups on side chains (Tugarinov, Hwang and Kay, 2004; Tugarinov and Kay, 2004).

28.4.2
Torsion Angles

Using the Karplus relationship (Karplus, 1959), torsion angles can be calculated from three-bond scalar coupling between the nuclei involved. A number of pulse sequences have been developed to measure scalar couplings between relevant nuclei, and the corresponding Karplus relationships have been calibrated.

The scalar couplings $^3J_{\mathrm{HN}\alpha}$, $^3J_{\mathrm{HNCO}}$, $^3J_{\mathrm{HNC}\beta}$, and $^3J_{\mathrm{C}i-1\mathrm{H}\alpha i}$ correlate to backbone ϕ angles, and $^3J_{\mathrm{N}i\mathrm{H}\alpha i-1}$ correlates to ψ angles in proteins (discussed in Smith *et al.*, 1996). $^3J_{\mathrm{HN}\alpha}$ can be measured from the HNHA experiment to generate ϕ structural restraints (Barnwal *et al.*, 2007; Kuboniwa *et al.*, 1994; Pardi *et al.*, 1984). Restraints for side chain χ_1 angles can be derived from $^3J_{\mathrm{NH}\beta}$, measured from the HNHB experiment (Archer *et al.*, 1991). Restraints for ψ angles can be derived from the intensity ratio $d_{\mathrm{N}\alpha}/d_{\alpha\mathrm{N}}$ in a ^{15}N NOESY-HSQC experiment (Gagné *et al.*, 1994).

Alternatively, main-chain torsion angles can be calculated from chemical shift values by computer programs that use a database of protein structures and chemical shift values, such as TALOS (Cornilescu *et al.*, 1999), SHIFTOR (Neal *et al.*, 2006), and the more recent DANGLE (http://dangle.sourceforge.net/).

Numerous experiments can be used to measure scalar coupling constants in carbohydrates, as reviewed by Bush *et al.* (1999) and Duus, Gotfredsen and Bock (2000), and new experiments and Karplus correlations are still being determined (see Coxon, 2007; Mobli and Almond, 2007, for recent examples). Three-bond and one-bond coupling constants can be used to determine the anomeric configuration of the sugar (Duus, Gotfredsen and Bock, 2000).

In nucleic acids, the sugar pucker mode and a number of torsion angles can be determined from scalar couplings. The correlations between a large number of angles and the relevant coupling constants, and the experiments needed to determine such coupling constants, have been reviewed by Wijmenga and Van Buuren (1998). Phosphodiester backbone angles α and ζ in oligonucleotides can be calculated from cross-correlated relaxation rates of ^{1}H, ^{13}C-dipole, dipole and ^{31}P-chemical shift anisotropy (Richter *et al.*, 2000).

28.4.3
Residual Dipolar Couplings (RDCs)

While interaction between two nuclei connected through bonds leads to line splitting because of scalar coupling, the interaction of nuclei through space leads to dipolar coupling, which arises from the interaction between the two magnetic dipoles of the nuclei. The magnitude of an observed dipolar coupling is given by the following equation:

$$D_{ij} = \frac{-\mu_0 \gamma_i \gamma_j \hbar}{4\pi^2 r^3} \left\langle \frac{3\cos^2\theta - 1}{2} \right\rangle \tag{28.3}$$

where γ is the gyromagnetic ratio of the interacting spins i and j, r is the distance between the nuclei, and θ is the angle between the applied magnetic field and the internuclear vector.

In solution NMR, the splitting due to dipolar coupling is not seen, being averaged to zero by rotational diffusion, unlike scalar coupling. (The splitting caused by scalar coupling is normally eliminated by decoupling techniques, but can be seen in undecoupled spectra.) When the molecules in the sample are partially aligned, some splitting due to dipolar coupling can be observed, named as residual dipolar coupling (RDC). By keeping the degree of alignment small, but nonzero, it is possible to have measurable effects from RDCs, and still keep the simplified spectrum obtained from motionally averaged molecules. From RDCs, orientational restraints can be derived, and are widely used in structural determination of proteins, carbohydrates, and nucleic acids. Different aspects of the use of RDCs in structural determination of biological macromolecules have been reviewed by Bax and Grishaev (2005), Blackledge (2005), Lipsitz and Tjandra (2004), and Prestegard *et al.* (2005). The influence of dipolar coupling on peak splitting in biological macromolecules was first shown for the main-chain amides of cyanometmyoglobin, a protein that spontaneously shows detectable partial alignment within the magnetic field from an NMR spectrometer (Tolman *et al.*, 1995).

The effect of alignment on main-chain amide ^{15}N splitting in proteins provides a good example of RDC measurements. In a splitting-enhanced {^{1}H, ^{15}N}-HSQC spectrum (undecoupled in the ^{15}N dimension) (Tolman and Prestegard, 1996), splitting of the ^{15}N resonance in protein main-chain amides can be observed, and the coupling constant can be measured from the distance between the peaks. In solution NMR, when the molecules are tumbling isotropically, this splitting will reflect only the scalar coupling constant $^{1}J_{\mathrm{HN}}$ between the proton and nitrogen. When the molecules are partially aligned by any method, the splitting will reflect the sum of the scalar coupling, $^{1}J_{\mathrm{HN}}$, and the residual dipolar coupling, D_{HN}. Another method to measure the splitting involves the use of in-phase/antiphase (IPAP) spectra, in which the two components of the doublet are seen in separate spectra, therefore not increasing spectral complexity compared to the decoupled spectrum (Ottiger *et al.*, 1998). In *J*-resolved spectroscopy (Aue *et al.*, 1976), the proton resonances evolve with the chemical shift in t_2, and with *J*-coupling in t_1. Unlike the HSQC experiments, that take advantage of heteronuclear signal dispersion in the indirectly detected dimension, *J*-resolved spectroscopy requires good dispersion in the proton spectrum.

From Eq. (28.3) above, the angle between the bond vector and the magnetic field can be determined. It is important to notice that there are always two opposite orientations of the bond vector that satisfy the equation for any RDC value, and therefore RDC information should be obtained from multiple alignment media for structural calculations, or supplemented with other experimental distance information (Blackledge, 2005). RDC information can be used for structural calculation using simulated annealing programs (Tjandra and Bax, 1997). Orientational restraints are very important for the structural determination of carbohydrates, nucleic acids, and large deuterated proteins, for which it is usually not possible to acquire enough NOE distance restraints to resolve the structure. Numerous pulse sequences are available for the determination of RDCs between different nuclei in nucleic acids (reviewed by Latham *et al.*, 2005).

Alignment of molecules with the magnetic field can be generated with the use of paramagnetic metals (Tolman *et al.*, 1995), or by restricting the movement of the molecule with alignment media, which will result in alignment according to the shape of the molecule. Different alignment media that can be used in biological samples include liquid crystal media (Tjandra and Bax, 1997), a suspension of filamentous bacteriophage (Hansen *et al.*, 1998), lipid bicelles (Sanders *et al.*, 1994; Bax and Tjandra, 1997), or stretched polyacrylamide gels (Sass *et al.*, 2000; Tycko *et al.*, 2000).

28.4.4
Hydrogen Bonds

Experimentally detected hydrogen bonds can provide useful structural restraints. In proteins, hydrogen bonds patterns can reveal secondary structure, while in nucleic acids, hydrogen bonds are present in every base pairing. They can be measured either indirectly, from the exchange with the solvent (for example, in Krishna *et al.*, 2004), or directly, via experiments that transfer magnetization across the hydrogen bond through the two-bond scalar couplings (Grzesiek *et al.*, 2004). For structural determination of nucleic acids, the use of NMR experiments that transfer magnetization through H-bond scalar couplings has become widespread.

28.4.5
Paramagnetic Centers

Paramagnetic centers in biological macromolecules can occur naturally, as in metalloproteins, or be added by different methods, such as by replacing a naturally-occurring diamagnetic ion with a paramagnetic ion (Lee and Sykes, 1983), adding a nitroxide spin label (Battiste and Wagner, 2000; Gaponenko *et al.*, 2000), or by engineering a binding site for a paramagnetic ion (for examples, see Ma and Opella, 2000; Donaldson *et al.*, 2001; Wöhnert *et al.*, 2003; Su *et al.*, 2008). Different techniques have been developed to incorporate paramagnetic centers in nucleic acids, for instance using 4-thiouracil or 4-thiothymidine, which is then coupled to a spin label via the sulfur group (Ramos and Varani, 1998), or using site-specifically 2′-amino-modified nucleic acids (Edwards and Sigurdsson, 2007). The use of chemical modifications to create lanthanide-binding sites for paramagnetic structural restraints is becoming increasingly popular (Otting, 2008).

Contact shifts caused by through-bond interaction between the nucleus and the paramagnetic center can be a powerful source of structural information. The presence of an unpaired electron in an NMR sample will also affect the chemical shifts and relaxation of other nuclei close enough to the unpaired electron to present dipolar interactions with it. In practice, this can be as much as 20 Å, which is a longer distance than can be observed by NOE interactions. Having a paramagnetic center in a known position of the molecule, then, can be an effective way to determine structural constraints for nuclei located at close or medium distance from that position. The effect of a paramagnetic center on nearby nuclei can be used to calculate a number of constraints for structural determination, based on pseudocontact shifts, relaxation enhancement, contact shifts, and cross-relaxation (Bertini *et al.*, 2002; Bertini *et al.*, 2005). Pulse sequences may have to be modified when characterizing biomolecules containing paramagnetic centers.

Paramagnetic centers also cause the molecules to align along the external magnetic field, and this alignment can used to extract residual dipolar couplings (Tolman *et al.*, 1995). The combination of RDC and paramagnetic restraints has been used to determine a protein structure without NOE restraints (Hus *et al.*, 2000). As with RDC, the use of paramagnetic restraints can be very beneficial to molecules for which it is not possible to obtain sufficient NOE restraints for accurate structural determination, such as nucleic acids.

28.4.6
Structural Information from Chemical Shifts

Different conformations in biomolecules can have strong effects on the chemical shift values of nuclei in specific positions, and this information can be used to help structural determination. In proteins, regions of secondary structure can be predicted from the chemical shift index (Wishart and Nip, 1998; Wishart and Case, 2001). In brief, chemical shift values of $^{1}H_{\alpha}$, $^{1}H_{\beta}$, $^{1}H_{N}$, $^{13}C_{\alpha}$, $^{13}C_{\beta}$, ^{13}CO, and ^{15}N for each residue can be compared to typical random coil values, and the difference is usually indicative of the presence or absence of helical or sheet secondary structure. For an element of secondary structure to be predicted reliably, it should be predicted by at least three amino acid residues in a row. The use of multiple nuclei also increases the reliability of the method. Methods to extract information from chemical shift values for proteins have been reviewed by Szilágyi (1995) and Case (2000). A number of uses of chemical shift values to aid in assignment and structural determination of nucleic acids have been described by Wijmenga *et al.* (1997), Altona *et al.* (2000), Cromsigt *et al.* (2001) and Fürtig *et al.* (2003).

In partially oriented samples, chemical shift anisotropy (CSA) can also yield structural information (Lipsitz and Tjandra, 2001; Lipsitz and Tjandra, 2003; Ying *et al.*, 2006; Choy *et al.*, 2001; Cornilescu and Bax, 2000; Wu *et al.*, 2001; Yu and Prestegard, 2006). Chemical shifts in a molecule are usually anisotropic by nature, depending on the orientation in the magnetic field. When a molecule tumbles isotropically in solution NMR, the averaging of chemical shift values gives rise to the sharp lines seen in solution NMR. In solid state NMR, the anisotropy can be seen in the "powder pattern" if the molecules are not oriented. In partially aligned samples, the effect of CSA can be seen as an offset of the chemical shift in relation to its value in an isotropic sample.

Recently, methods for determining protein structures based on chemical shift values alone have been developed, CHESHIRE (Cavalli *et al.*, 2007), and CS-ROSETTA (Shen *et al.*, 2008). Both methods use molecular fragment replacement approaches, used previously on the ROSETTA *ab initio* structure prediction method (Bradley *et al.*, 2005) and to analyze RDCs (Rohl and Baker, 2002; Delaglio *et al.*, 2000), amongst other applications. By eliminating the need to acquire separate spectra for structural restraints, the experimental time needed to determine protein structures can be greatly reduced, and structural determination by NMR can be extended to less stable proteins. A current limitation of this approach is protein size; both methods are effective only for proteins up to 15 kDa. The newly developed CamDock procedure extends this approach to protein-protein complexes (Montalvão *et al.*, 2008).

28.5
Molecular Modeling and Structure Calculation

Structural restraints determined as described above are used to calculate the structure by molecular modeling. Simulated annealing and molecular dynamics, either in the Cartesian space of atomic positions or in the torsion angle space, are the most common methods for structure calculation from restraints derived from NMR experimental data, starting from an extended structure or from a preliminary model (Clore and Gronenborn, 1998). The different experimental restraints are incorporated into the calculations as pseudo-energy functions added to the molecular energy representations, forming a hybrid energy term $E_{\text{hybrid}} = E_{\text{phys}} + w_{\text{data}} E_{\text{data}}$, where E_{phys} is the sum of physical energy terms such as bond length and nonbonded interactions, and E_{data} assesses the consistency between the model structure and the different types of experimentally measured restraints (Brünger and Nilges, 1993).

The initial high-temperature stage of the simulated annealing calculations allows for wider exploration of the conformational space by overcoming energy barriers, followed by slow cooling while sampling the conformational space by molecular dynamics. Molecular modeling in Cartesian coordinates is based on the numerical solution of Newton's equations of motion, and the position of each atom is modeled with energy representations for bond angles and lengths (Nilges *et al.*, 1988). Molecular dynamics in the torsion angle space assumes the distance and bond angle between the pair of atoms connected by a bond to be fixed, and allows them to rotate around the bond. The reduced number of degrees of freedom and larger time steps in the simulated annealing lead to faster calculations and convergence (Stein *et al.*, 1997). Common programs that use molecular dynamics for structural calculations from NMR data include Xplor-NIH (Schwieters *et al.*, 2003), CNS (Brünger, 1998; Brunger, 2007) and CYANA (Güntert, 2004), or its predecessor DYANA (Güntert *et al.*, 1997). While computational power was a limiting factor when these methods were first developed, now these calculations are trivial and can be performed in most PCs with the appropriate software.

The iterative steps of NOESY peaks assignment and structural calculation have been automated together in a number of programs, such as ARIA (Nilges *et al.*, 1997; Linge *et al.*, 2003), CANDID (Herrmann *et al.*, 2002), CYANA (Güntert, 2004), and others (Güntert, 2003). These programs require a chemical shift list as input, with the resonance values for the atoms that will be involved in NOE contacts.

More recently, methods based on molecular fragment replacement have been developed for proteins, in which model structures for each fragment are retrieved from a database according to the fragment sequence and experimental data (Bowers, Strauss and Baker, 2000). This is the same approach used for *ab initio* protein structure calculation by the program ROSETTA (Simons *et al.*, 1997; Simons *et al.*, 1999) and has been applied to structure calculation based on chemical shifts only (Cavalli *et al.*, 2007; Shen *et al.*, 2008; Montalvão *et al.*, 2008) and to the analysis of RDC data (Rohl and Baker, 2002). These calculations are more computationally demanding than molecular dynamics-based methods.

Biomolecular structures calculated using any of the methods above need to be validated, and there are a number of programs for structural validation such as VADAR (Willard *et al.*, 2003), PROCHECK-NMR and AQUA (Laskowski *et al.*, 1996), and WHAT IF (Doreleijers *et al.*, 1999; Vriend, 1990).

References

Altona, C., Faber, D.H. and Westra Hoekzema, A.J.A. (2000) *Magn. Reson. Chem.*, **38**(2), 95–107.

Archer, S.J., Ikura, M., Torchia, D.A. and Bax, A. (1991) *J. Magn. Reson.*, **95**(3), 636–641.

Atreya, H.S., Chary, K.V.R. and Govil, G. (2002) *Curr. Sci.*, **83**(11), 1372–1376.

Aue, W.P., Karhan, J. and Ernst, R.R. (1976) *J. Chem. Phys.*, **64**(10), 4226–4227.

Barnwal, R.P., Rout, A.K., Chary, K.V.R. and Atreya, H.S. (2007) *J. Biomol. NMR*, **39**(4), 259–263.

Battiste, J.L. and Wagner, G. (2000) *Biochemistry*, **39**(18), 5355–5365.

Bax, A. and Grishaev, A. (2005) *Curr. Opin. Struct. Biol.*, **15**(5), 563–570.

Bax, A. and Ikura, M. (1991) *J. Biomol. NMR*, **1**(1), 99–104.

Bax, A. and Tjandra, N. (1997) *J. Biomol. NMR*, **10**(3), 289–292.

Bertini, I., Luchinat, C. and Parigi, G. (2002) *Concepts Magn. Reson. A Bridging Educ. Res.*, **14**(4), 259–286.

Bertini, I., Luchinat, C., Parigi, G. and Pierattelli, R. (2005) *Chem. Bio. Chem.*, **6**(9), 1536–1549.

Blackledge, M. (2005) *Prog. Nucl. Magn. Reson. Spectrosc.*, **46**(1), 23–61.

Bowers, P.M., Strauss, C.E.M. and Baker, D. (2000) *J. Biomol. NMR*, **18**(4), 311–318.

Bradley, P., Misura, K.M.S. and Baker, D. (2005) *Science*, **309**(5742), 1868–1871.

Brüggert, M., Rehm, T., Shanker, S., Georgescu, J. and Holak, T.A. (2003) *J. Biomol. NMR*, **25**(4), 335–348.

Brünger, A.T. (1998) *Acta Crystallogr. D Biol. Crystallogr.*, **54**(5), 905–921.

Brünger, A.T. (2007) *Nat. Protoc.*, **2**(11), 2728–2733.

Brünger, A.T. and Nilges, M. (1993) *Q. Rev. Biophys.*, **26**(1), 49–125.

Brüschweiler, R. and Zhang, F. (2004) *J. Chem. Phys.*, **120**(11), 5253–5260.

Brutscher, B. (2004) *J. Biomol. NMR*, **29**(1), 57–64.

Brutscher, B., Simorre, J.P., Caffrey, M.S. and Marion, D. (1994) *J. Magn. Reson. B*, **105**(1), 77–82.

Bush, C.A., Martin-Pastor, M. and Imberty, A. (1999) *Annu. Rev. Biophys. Biomol. Struct.*, **28**, 269–293.

Case, D.A. (2000) *Curr. Opin. Struct. Biol.*, **10**(2), 197–203.

Cavalli, A., Salvatella, X., Dobson, C.M. and Vendruscolo, M. (2007) *Proc. Natl. Acad. Sci. U.S.A.*, **104**(23), 9615–9620.

Cavanagh, J., Fairbrother, W.J., Palmer, A.G. and Skelton, N.J. (1996) *Protein NMR Spectroscopy: Principles and Practice*, Academic Press, San Diego.

Choy, W.Y., Tollinger, M., Mueller, G.A. and Kay, L.E. (2001) *J. Biomol. NMR*, **21**(1), 31–40.

Clore, G.M. and Gronenborn, A.M. (1991) *Annu. Rev. Biophys. Biophys. Chem.*, **20**, 29–63.

Clore, G.M. and Gronenborn, A.M. (1998) *Proc. Natl. Acad. Sci. U.S.A.*, **95**(11), 5891–5898.

Clore, G.M., Gronenborn, A.M., Brünger, A.T. and Karplus, M. (1985) *J. Mol. Biol.*, **186**(2), 435–455.

Clubb, R.T., Thanabal, V. and Wagner, G. (1992) *J. Magn. Reson.*, **97**(1), 213–217.

Cornilescu, G. and Bax, A. (2000) *J. Am. Chem. Soc.*, **122**(41), 10143–10154.

Cornilescu, G., Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13**(3), 289–302.

Coxon, B. (2007) *Carbohydr. Res.*, **342**(8), 1044–1054.

Cromsigt, J., Van Buuren, B., Schleucher, J. and Wijmenga, S. (2001) *Methods Enzymol.*, **338**, 371–399.

Dayie, K.T. (2008) *Int. J. Mol. Sci.*, **9**(7), 1214–1240.

Delaglio, F., Kontaxis, G. and Bax, A. (2000) *J. Am. Chem. Soc.*, **122**(9), 2142–2143.

Dingley, A.J. and Grzesiek, S. (1998) *J. Am. Chem. Soc.*, **120**(33), 8293–8297.

Donaldson, L.W., Skrynnikov, N.R., Choy, W.Y., Muhandiram, D.R., Sarkar, B., Forman-Kay, J.D. and Kay, L.E. (2001) *J. Am. Chem. Soc.*, **123**(40), 9843–9847.

Doreleijers, J.F., Vriend, G., Raves, M.L. and Kaptein, R. (1999) *Proteins Struct. Funct. Genet.*, **37**(3), 404–416.

Duus, J.Ø., Gotfredsen, C.H. and Bock, K. (2000) *Chem. Rev.*, **100**(12), 4589–4614.

Dyson, H.J. and Wright, P.E. (2005) *Nat. Rev. Mol. Cell Biol.*, **6**(3), 197–208.

Edwards, T.E. and Sigurdsson, S.T. (2007) *Nat. Protoc.*, **2**(8), 1954–1962.

Ferentz, A.E. and Wagner, G. (2000) *Q. Rev. Biophys.*, **33**(1), 29–65.

Flinders, J. and Dieckmann, T. (2006) *Prog. Nucl. Magn. Reson. Spectrosc.*, **48**(2–3), 137–159.

Freeman, R. and Kupce, E. (2003) *J. Biomol. NMR*, **27**(2), 101–113.

Frydman, L., Scherf, T. and Lupulescu, A. (2002) *Proc. Natl. Acad. Sci. U.S.A.*, **99**(25), 15858–15862.

Fürtig, B., Richter, C., Wöhnert, J. and Schwalbe, H. (2003) *Chem. Bio. Chem.*, **4**(10), 936–962.

Gagné, S.M., Li, M.X. and Sykes, B.D. (1997) *Biochemistry*, **36**(15), 4386–4392.

Gagné, S.M., Tsuda, S., Li, M.X., Chandra, M., Smillie, L.B. and Sykes, B.D. (1994) *Protein Sci.*, **3**(11), 1961–1974.

Gaponenko, V., Howarth, J.W., Columbus, L., Gasmi-Seabrook, G., Yuan, J., Hubbell, W.L. and Rosevear, P.R. (2000) *Protein Sci.*, **9**(2), 302–309.

Gardner, K.H. and Kay, L.E. (1998) *Annu. Rev. Biophys. Biomol. Struct.*, **27**, 357–406.

Grzesiek, S., Anglister, J., Ren, H. and Bax, A. (1993) *J. Am. Chem. Soc.*, **115**(10), 4369–4370.

Grzesiek, S. and Bax, A. (1992a) *J. Am. Chem. Soc.*, **114**(16), 6291–6293.

Grzesiek, S. and Bax, A. (1992b) *J. Magn. Reson.*, **96**(2), 432–440.

Grzesiek, S., Cordier, F., Jaravine, V. and Barfield, M. (2004) *Prog. Nucl. Magn. Reson. Spectrosc.*, **45**(3–4), 275–300.

Güntert, P. (2003) *Prog. Nucl. Magn. Reson. Spectrosc.*, **43**(3–4), 105–125.

Güntert, P. (2004) *Methods Mol. Biol. (Clifton)*, **278**, 353–378.

Güntert, P., Mumenthaler, C. and Wüthrich, K. (1997) *J. Mol. Biol.*, **273**(1), 283–298.

Hansen, M.R., Mueller, L. and Pardi, A. (1998) *Nat. Struct. Biol.*, **5**(12), 1065–1074.

Helgstrand, M., Kraulis, P., Allard, P. and Härd, T. (2000) *J. Biomol. NMR*, **18**(4), 329–336.

Herrmann, T., Güntert, P. and Wüthrich, K. (2002) *J. Mol. Biol.*, **319**(1), 209–227.

Hoch, J.C. and Stern, A.S. (2001) *Methods Enzymol.*, **338**, 159–178.

Hore, P.J. (1983) *J. Magn. Reson.*, **55**(2), 283–300.

Hus, J.C., Marion, D. and Blackledge, M. (2000) *J. Mol. Biol.*, **298**(5), 927–936.

Jung, Y.S. and Zweckstetter, M. (2004) *J. Biomol. NMR*, **30**(1), 11–23.

Kanelis, V., Forman-Kay, J.D. and Kay, L.E. (2002) *IUBMB Life*, **52**(6), 291–302.

Karplus, M. (1959) *J. Chem. Phys.*, **30**(1), 11–15.

Kay, L.E. (2005) *J. Magn. Reson.*, **173**(2), 193–207.

Kay, L.E., Ikura, M., Tschudin, R. and Bax, A. (1990) *J. Magn. Reson.*, **89**(3), 496–514.

Kelly, A.E., Ou, H.D., Withers, R. and Dötsch, V. (2002) *J. Am. Chem. Soc.*, **124**(40), 12013–12019.

Kim, S. and Szyperski, T. (2003) *J. Am. Chem. Soc.*, **125**(5), 1385–1393.

Krishna, M.M.G., Hoang, L., Lin, Y. and Englander, S.W. (2004) *Methods*, **34**(1), 51–64.

Kuboniwa, H., Grzesiek, S., Delaglio, F. and Bax, A. (1994) *J. Biomol. NMR*, **4**(6), 871–878.

Kupce, E. and Freeman, R. (2004) *J. Am. Chem. Soc.*, **126**(20), 6429–6440.

Kupce, E. and Freeman, R. (2008) *Prog. Nucl. Magn. Reson. Spectrosc.*, **52**(1), 22–30.

Kupce, E., Nishida, T. and Freeman, R. (2003) *Prog. Nucl. Magn. Reson. Spectrosc.*, **42**(3–4), 95–122.

Lagoja, I.M. and Herdewijn, P. (2002) *Synthesis*, **2002**(3), 301–314.

Laskowski, R.A., Rullmann, J.A.C., MacArthur, M.W., Kaptein, R. and Thornton, J.M. (1996) *J. Biomol. NMR*, **8**(4), 477–486.

Latham, M.P., Brown, D.J., McCallum, S.A. and Pardi, A. (2005) *Chem. Bio. Chem.*, **6**(9), 1492–1505.

Lee, L. and Sykes, B.D. (1983) *Biochemistry*, **22**(19), 4366–4373.

Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G. and Kessler, H. (1998) *J. Biomol. NMR*, **11**(1), 31–43.

Li, M.X., Corson, D.C. and Sykes, B.D. (2002) *Methods Mol. Biol. (Clifton)*, **173**, 255–265.

Lindorff-Larsen, K., Kristjansdottir, S., Teilum, K., Fieber, W., Dobson, C.M., Poulsen, F.M. and Vendruscolo, M. (2004) *J. Am. Chem. Soc.*, **126**(10), 3291–3299.

Linge, J.P., Habeck, M., Rieping, W. and Nilges, M. (2003) *Bioinformatics*, **19**(2), 315–316.

Lipsitz, R.S. and Tjandra, N. (2001) *J. Am. Chem. Soc.*, **123**(44), 11065–11066.

Lipsitz, R.S. and Tjandra, N. (2003) *J. Magn. Reson.*, **164**(1), 171–176.

Lipsitz, R.S. and Tjandra, N. (2004) *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 387–413.

Louis, J.M., Martin, R.G., Clore, G.M. and Gronenborn, A.M. (1998) *J. Biol. Chem.*, **273**(4), 2374–2378.

Ma, C. and Opella, S.J. (2000) *J. Magn. Reson.*, **146**(2), 381–384.

Mandelshtam, V.A. (2000) *J. Magn. Reson.*, **144**(2), 343–356.

Marion, D., Driscoll, P.C., Kay, L.E., Wingfield, P.T., Bax, A., Gronenborn, A.M. and Clore, G.M. (1989a) *Biochemistry*, **28**(15), 6150–6156.

Marion, D., Kay, L.E., Sparks, S.W., Torchia, D.A. and Bax, A. (1989b) *J. Am. Chem. Soc.*, **111**(4), 1515–1517.

Martin-Pastor, M. and Allen Bush, C. (1999) *Biochemistry*, **38**(25), 8045–8055.

Milligan, J.F., Groebe, D.R., Witherell, G.W. and Uhlenbeck, O.C. (1987) *Nucleic Acids Res.*, **15**(21), 8783–8798.

Mobli, M. and Almond, A. (2007) *Org. Biomol. Chem.*, **5**(14), 2243–2251.

Montalvão, R.W., Cavalli, A., Salvatella, X., Blundell, T.L. and Vendruscolo, M. (2008) *J. Am. Chem. Soc.*, **130**(47), 15990–15996.

Morgan, W.D., Kragt, A. and Feeney, J. (2000) *J. Biomol. NMR*, **17**(4), 337–347.

Muhandiram, D.R. and Kay, L.E. (1994) *J. Magn. Reson. B*, **103**(3), 203–216.

Mumenthaler, C., Güntert, P., Braun, W. and Wüthrich, K. (1997) *J. Biomol. NMR*, **10**(4), 351–362.

Neal, S., Berjanskii, M., Zhang, H. and Wishart, D.S. (2006) *Magn. Reson. Chem.*, **44**(S1), S158–S167.

Nelissen, F.H.T., van Gammeren, A.J., Tessari, M., Girard, F.C., Heus, H.A. and Wijmenga, S.S. (2008) *Nucleic Acids Res.*, **36**(14), e89.

Nilges, M. (1995) *J. Mol. Biol.*, **245**(5), 645–660.

Nilges, M., Clore, G.M. and Gronenborn, A.M. (1988) *FEBS Lett.*, **239**(1), 129–136.

Nilges, M., MacIas, M.J., O'Donoghue, S.I. and Oschkinat, H. (1997) *J. Mol. Biol.*, **269**(3), 408–422.

Osterhout, J.J. Jr, Baldwin, R.L., York, E.J., Stewart, J.M., Dyson, H.J. and Wright, P.E. (1989) *Biochemistry*, **28**(17), 7059–7064.

Otting, G. (2008) *J. Biomol. NMR*, **42**(1), 1–9.

Ottiger, M., Delaglio, F. and Bax, A. (1998) *J. Magn. Reson.*, **131**(2), 373–378.

Overhauser, A.W. (1953) *Phys. Rev.*, **92**(2), 411–415.

Palmer, A.G. III (2001) *Annu. Rev. Biophys. Biomol. Struct.*, **30**, 129–155.

Pardi, A., Billeter, M. and Wüthrich, K. (1984) *J. Mol. Biol.*, **180**(3), 741–751.

Pervushin, K., Ono, A., Fernández, C., Szyperski, T., Kainosho, M. and Wüthrich, K. (1998) *Proc. Natl. Acad. Sci. U.S.A.*, **95**(24), 14147–14151.

Pervushin, K., Riek, R., Wider, G. and Wüthrich, K. (1997) *Proc. Natl. Acad. Sci. U.S.A.*, **94**(23), 12366–12371.

Petersen, B.O., Vinogradov, E., Kay, W., Wurtz, P., Nyberg, N.T., Duus, J.Ø. and Sorensen, O.W. (2006) *Carbohydr. Res.*, **341**(4), 550–556.

Piotto, M., Saudek, V. and Sklenár, V. (1992) *J. Biomol. NMR*, **2**(6), 661–665.

Prestegard, J.H., Mayer, K.L., Valafar, H. and Benison, G.C. (2005) *Methods Enzymol.*, **394**, 175–209.

Ramos, A. and Varani, G. (1998) *J. Am. Chem. Soc.*, **120**(42), 10992–10993.

René, B., Masliah, G., Zargarian, L., Mauffret, O. and Fermandjian, S. (2006) *J. Biomol. NMR*, **36**(3), 137–146.

Richter, C., Reif, B., Griesinger, C. and Schwalbe, H. (2000) *J. Am. Chem. Soc.*, **122**(51), 12728–12731.

Rohl, C.A. and Baker, D. (2002) *J. Am. Chem. Soc.*, **124**(11), 2723–2729.

Roumestand, C., Delay, C., Gavin, J.A. and Canet, D. (1999) *Magn. Reson. Chem.*, **37**(7), 451–478.

Rovnyak, D., Frueh, D.P., Sastry, M., Sun, Z.Y.J., Stern, A.S., Hoch, J.C. and Wagner, G. (2004) *J. Magn. Reson.*, **170**(1), 15–21.

Sanders, C.R. II, Hare, B.J., Howard, K.P. and Prestegard, J.H. (1994) *Prog. Nucl. Magn. Reson. Spectrosc.*, **26**(PART V), 421–444.

Sass, H.J., Musco, G., Stahl, S.J., Wingfield, P.T. and Grzesiek, S. (2000) *J. Biomol. NMR*, **18**(4), 303–309.

Schanda, P., Kupce, E. and Brutscher, B. (2005) *J. Biomol. NMR*, **33**(4), 199–211.

Schwieters, C.D., Kuszewski, J.J., Tjandra, N. and Clore, G.M. (2003) *J. Magn. Reson.*, **160**(1), 65–73.

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J.M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K.K., Lemak, A., Ignatchenko, A., Arrowsmith, C.H., Szyperski, T., Montelione, G.T., Baker, D. and Bax, A. (2008) *Proc. Natl. Acad. Sci. U.S.A.*, **105**(12), 4685–4690.

Simons, K.T., Bonneau, R., Ruczinski, I. and Baker, D. (1999) *Proteins Struct. Funct. Genet.*, **37**(Suppl. 3), 171–176.

Simons, K.T., Kooperberg, C., Huang, E. and Baker, D. (1997) *J. Mol. Biol.*, **268**(1), 209–225.

Sklenár, V. and Bax, A. (1987) *J. Magn. Reson.*, **74**(3), 469–479.

Slupsky, C.M., Boyko, R.F., Booth, V.K. and Sykes, B.D. (2003) *J. Biomol. NMR*, **27**(4), 313–321.

Smith, L.J., Bolin, K.A., Schwalbe, H., MacArthur, M.W., Thornton, J.M. and Dobson, C.M. (1996) *J. Mol. Biol.*, **255**(3), 494–506.

Stein, E.G., Rice, L.M. and Brünger, A.T. (1997) *J. Magn. Reson.*, **124**(1), 154–164.

Stonehouse, J., Clowes, R.T., Shaw, G.L., Keeler, J. and Laue, E.D. (1995) *J. Biomol. NMR*, **5**(3), 226–232.

Su, X.C., McAndrew, K., Huber, T. and Otting, G. (2008) *J. Am. Chem. Soc.*, **130**(5), 1681–1687.

Szilágyi, L. (1995) *Prog. Nucl. Magn. Reson. Spectrosc.*, **27**(4), 325–443.

Szyperski, T., Wider, G., Bushweller, J.H. and Wüthrich, K. (1993) *J. Am. Chem. Soc.*, **115**, 9307–9308.

Tjandra, N. and Bax, A. (1997) *Science*, **278**(5340), 1111–1114.

Tolman, J.R., Flanagan, J.M., Kennedy, M.A. and Prestegard, J.H. (1995) *Proc. Natl. Acad. Sci. U.S.A.*, **92**(20), 9279–9283.

Tolman, J.R. and Prestegard, J.H. (1996) *J. Magn. Reson. B*, **112**(3), 269–274.

Torizawa, T., Shimizu, M., Taoka, M., Miyano, H. and Kainosho, M. (2004) *J. Biomol. NMR*, **30**(3), 311–325.

Tugarinov, V., Hwang, P.M. and Kay, L.E. (2004) *Annu. Rev. Biochem.*, **73**, 107–146.

Tugarinov, V. and Kay, L.E. (2004) *J. Biomol. NMR*, **28**(2), 165–172.

Tycko, R., Blanco, F.J. and Ishii, Y. (2000) *J. Am. Chem. Soc.*, **122**(38), 9340–9341.

Tzakos, A.G., Grace, C.R.R., Lukavsky, P.J. and Riek, R. (2006) *Annu. Rev. Biophys. Biomol. Struct.*, **35**, 319–342.

Uhrin, D. and Barlow, P.N. (1997) *J. Magn. Reson.*, **126**(2), 248–255.

Ulrich, E.L., Akutsu, H., Doreleijers, J.F., Harano, Y., Ioannidis, Y.E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C.F., Tolmie, D.E., Kent Wenger, R., Yao, H. and Markley, J.L. (2008) *Nucleic Acids Res.*, **36**(Suppl. 1), D402–D408.

Varani, G., Aboul-Ela, F. and Allain, F.H.T. (1996) *Prog. Nucl. Magn. Reson. Spectrosc.*, **29**(1–2), 51–127.

Vinarov, D.A., Lytle, B.L., Peterson, F.C., Tyler, E.M., Volkman, B.F. and Markley, J.L. (2004) *Nat. Methods*, **1**(2), 149–153.

Volk, J., Herrmann, T. and Wüthrich, K. (2008) *J. Biomol. NMR*, **41**(3), 127–138.

Vriend, G. (1990) *J. Mol. Graph.*, **8**(1), 52–56.

Wagner, G. and Wüthrich, K. (1982) *J. Mol. Biol.*, **155**(3), 347–366.

Wider, G. and Wüthrich, K. (1999) *Curr. Opin. Struct. Biol.*, **9**(5), 594–601.

Wijmenga, S.S., Kruithof, M. and Hilbers, C.W. (1997) *J. Biomol. NMR*, **10**(4), 337–350.

Wijmenga, S.S. and Van Buuren, B.N.M. (1998) *Prog. Nucl. Magn. Reson. Spectrosc.*, **32**(4), 287–387.

Willard, L., Ranjan, A., Zhang, H., Monzavi, H., Boyko, R.F., Sykes, B.D. and Wishart, D.S. (2003) *Nucleic Acids Res.*, **31**(13), 3316–3319.

Williamson, M.P., Havel, T.F. and Wüthrich, K. (1985) *J. Mol. Biol.*, **182**(2), 295–315.

Wishart, D.S. and Case, D.A. (2001) *Methods Enzymol.*, **338**, 3–34.

Wishart, D.S. and Nip, A.M. (1998) *Biochem. Cell Biol.*, **76**(2–3), 153–163.

Wittekind, M. and Mueller, L. (1993) *J. Magn. Reson. B*, **101**(2), 201–205.

Wöhnert, J., Franz, K.J., Nitz, M., Imperiali, B. and Schwalbe, H. (2003) *J. Am. Chem. Soc.*, **125**(44), 13338–13339.

Wu, Z., Tjandra, N. and Bax, A. (2001) *J. Am. Chem. Soc.*, **123**(15), 3617–3618.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley & Sons, Inc., New York.

Yi, X., Venot, A., Glushka, J. and Prestegard, J.H. (2004) *J. Am. Chem. Soc.*, **126**(42), 13636–13638.

Ying, J., Grishaev, A., Bryce, D.L. and Bax, A. (2006) *J. Am. Chem. Soc.*, **128**(35), 11443–11454.

Yu, F. and Prestegard, J.H. (2006) *Biophys. J.*, **91**(5), 1952–1959.

Zuiderweg, E.R.P. and Fesik, S.W. (1989) *Biochemistry*, **28**(6), 2387–2391.

Further Reading

Cavanagh, J., Fairbrother, W.J., Palmer, III A.G., Skelton, N.J. and Rance, M. (2006) *Protein NMR Spectroscopy: Principles and Practice*, 2nd edn, Academic Press, San Diego.

Evans, J.N.S. (1995) *Biomolecular NMR Spectroscopy*, Oxford University Press, New York.

Kristina Downing, A. (ed.) (2004) *Protein NMR Techniques*, Methods in Molecular Biology, Vol. 278, Humana Press, Clifton.

James, T.L. (ed.) (2005) *Nuclear Magnetic Resonance of Biological Macromolecules*, Methods in Enzymology, Vol. 394, Academic Press, San Diego.

Markwick, P.R.L., Malliavin, T. and Nilges, M. (2008) *PLoS Comput. Biol.* **4**(9).

Marius Clore, G. and Schwieters, C.D. (2002) *Curr. Opin. Struct. Biol.*, **12**(2), 146–153.

29
Mass Spectrometry

Jürgen H. Gross

29.1 Introduction 991
29.1.1 Concept of Mass Spectrometry 991
29.1.2 Goal of Mass Spectrometry 991
29.1.3 Mass Spectrometry at a Glance 992
29.1.4 Aims and Scope 993
29.2 The Role of Isotopes in Mass Spectrometry 994
29.2.1 Isotopes 994
29.2.2 Atomic Mass Scale 994
29.2.3 Isotopic Patterns 994
29.2.4 Accurate Mass 995
29.2.5 Calculation of Atomic Weights and Molecular Weights 996
29.3 Mass Spectrometric Methods 996
29.4 Ionization Processes 999
29.4.1 Formation of Positive Radical Ions 999
29.4.1.1 Electron Ionization 999
29.4.1.2 Photoionization 999
29.4.1.3 Penning Ionization 1000
29.4.1.4 Charge Transfer 1000
29.4.1.5 Field Ionization 1001
29.4.2 Formation of Positive Ions 1001
29.4.2.1 Proton Transfer 1001
29.4.2.2 Electrophilic Addition 1001
29.4.2.3 Anion Abstraction 1002
29.4.3 Formation of Negative Radical Ions 1002
29.4.3.1 Electron Capture 1002
29.4.4 Formation of Negative Ions 1002
29.4.4.1 Nucleophilic Addition 1002
29.4.4.2 Cation Abstraction 1002
29.5 Isotopic and Elemental Analysis 1003
29.5.1 Thermal Ionization 1003
29.5.2 Spark Source 1005

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

29.5.3 Glow Discharge 1006
29.5.4 Inductively Coupled Plasma 1007
29.5.5 Accelerator Mass Spectrometry 1010
29.6 Ionization of Gaseous Molecules in Vacuum 1011
29.6.1 Electron Ionization 1012
29.6.2 Chemical Ionization 1014
29.6.3 Field Ionization 1017
29.7 Desorption/Ionization 1019
29.7.1 Field Desorption 1019
29.7.2 Fast Atom Bombardment and Liquid Secondary Ion Mass Spectrometry 1021
29.7.3 Laser Desorption/Ionization 1024
29.7.4 Matrix-Assisted Laser Desorption/Ionization 1026
29.8 Ion Formation at Atmospheric Pressure 1028
29.8.1 Atmospheric Pressure Ionization 1030
29.8.2 Electrospray Ionization 1032
29.8.3 Atmospheric Pressure Chemical Ionization and Atmospheric Pressure Photoionization 1036
29.8.4 Atmospheric Pressure MALDI 1039
29.8.5 Desorption Electrospray 1041
29.8.6 Direct Analysis in Real Time 1041
29.9 Activation Methods for Tandem Mass Spectrometry 1042
Acronyms 1046
References 1047
Further Reading 1052

29.1
Introduction

29.1.1
Concept of Mass Spectrometry

Mass spectrometry (MS) includes all techniques for the determination of the masses of single atoms, molecules and their fragments or their associates. The mass spectrometric experiment comprises the sequence of (i) generation of gas phase ions from either inorganic or organic compounds by any suitable method, (ii) separation of these ions by their *mass-to-charge ratio* (m/z), and (iii) detection of the ions qualitatively and quantitatively by their respective m/z and abundance. Each of these steps can be performed in various ways. The first step, *ionization*, can be achieved thermally, by electric fields or by impact of energetic particles such as electrons, small ions or heavy cluster ions, neutral atoms, or photons. The second step, *ion separation by m/z*, is effected by static or dynamic electric or magnetic fields and combinations thereof, forcing the ions to follow a trajectory characteristic of their respective m/z value. Finally, the detection of the ions is accomplished by transforming their impact onto a detector or the image current of their motion into an electric current that is then translated into a signal, a so-called peak, in the output of the *mass spectrometer*, the *mass spectrum* (Figure 29.1). Commonly, the signals are normalized to *relative intensity* by setting the most intensive peak, the *base peak*, to 100 %. A mass spectrum can be represented as bar graph, profile data, or m/z versus intensity listing.

According to these functional requirements, a mass spectrometer can be seen as composed of three basic moieties: (i) the *ion source*, (ii) the *mass analyzer*, and (iii) the *detector*. In contemporary mass spectrometers, these moieties are operated under the common control of a computer system. The principles, types, and operational modes of mass spectrometers are outlined in Chapter 23.

29.1.2
Goal of Mass Spectrometry

The ultimate goal of MS is to identify an analyte from the molecular or atomic mass(es) of its constituents. Knowing just the mass can be sufficient for the identification of elements and for the determination of the molecular formula

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

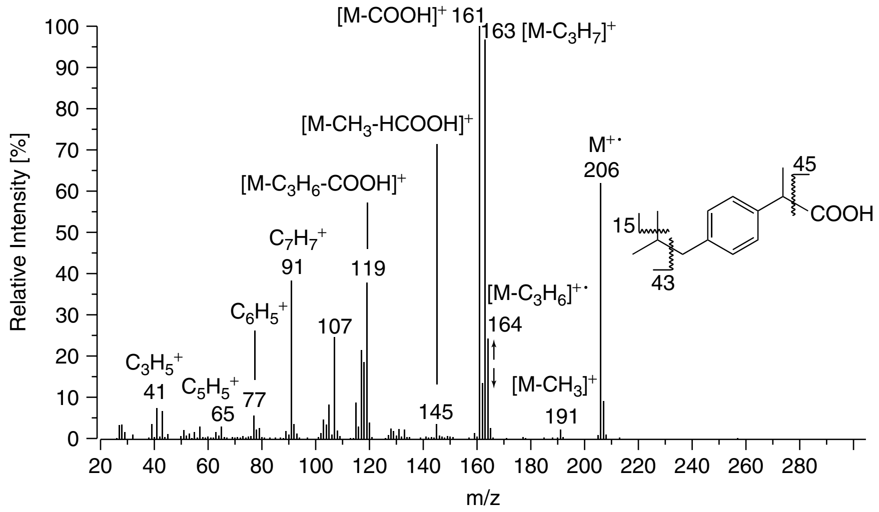

Fig. 29.1 The 70-eV electron ionization mass spectrum of the over-the-counter drug ibuprofen illustrates the principle of MS. The ionized molecule, that is, the molecular ion, fragments along several competing pathways to yield fragment ions that may themselves undergo further decomposition. The *m/z* values observed can – with more or less ease – be correlated to ionic compositions and eventually structures. Putting the puzzle together may reveal the structure of the analyte.

of an analyte. The relative abundance of isotopomers delivers strong supportive information to include or to rule out certain elements in such a formula and to estimate the number of atoms of a contributing element. Fragmentation of molecular species under the conditions of certain mass spectrometric experiments can deliver information on structure such as the connectivity of atoms within smaller molecules, the identification of functional groups, the (average) number and eventually the sequence of constituents of macromolecules, and in some cases even on their three-dimensional structure.

29.1.3
Mass Spectrometry at a Glance

MS has undergone tremendous transformations since J. J. Thomson's work marked its beginning in the early twentieth century (Grayson, 2002). MS developed from a mere mass measurement technique of physicists to one of the major analytical methods in chemistry and life sciences, while still preserving its importance in its fields of origin. Obviously, the fact that MS can be adapted in a very flexible manner to answer analytical questions of vast diversity allows it to become a whole branch of spectroscopy.

- MS serves for elemental identification and isotopic abundance measurement of both short-lived and stable species, thereby gaining relevance not only in physics and radiochemistry but also in geochemistry and more recently in life sciences.
- MS allows for the identification and structural characterization of small

molecules or for very large molecules as provided by either physiological processes or polymer chemistry.
- Mass spectrometric experiments can be stacked. Thus, a mass-selected ion can be studied by *tandem MS* and any product ion of this MS/MS (or MS^2) experiment may eventually be subjected to a third level (MS^3) and so forth (MS^n).
- MS, and especially tandem MS, also provides an elegant means for the study of unimolecular as well as bimolecular reactions of gas phase ions and for the determination of ion energetics.
- MS can be directly coupled to separation methods such as gas chromatography (GC) and all types of liquid chromatography (LC). *Coupling* or *hyphenation* to yield GC-MS or LC-MS instrumentation delivers high selectivity and low detection limits for the analysis of trace compounds in complicated matrices or the deconvolution of complex mixtures, making the hyphenated methods to become a major analytical tool.
- Mass spectra can be obtained from micrometer-sized areas on surfaces. This way, the lateral distribution of compounds on surfaces ranging from microelectronics to slices of tissue can be translated into images that in turn can be correlated to optical images of those samples. This is known as *mass spectral imaging*.
- Mass spectrometers can be very small. Miniaturization has led to portable instruments for environmental on-site analysis, detection systems for explosives and warfare chemicals, and last but not the least for many space missions.

In particular, the life sciences had and still have great impetus on the development of the manifold approaches to expand the mass range to higher molecular weights and increasingly fragile molecules. Environmental and pharmaceutical research present the driving force for the ever-improving limits of detection. There is ongoing research on methods of ion sampling, ion generation, and subsequent ion transfer into mass analyzers with better performance. Every application has unique requirements for the exact way in which the mass spectrometric experiment is performed, that is, on how to acquire a mass spectrum of a given analyte. MS has many facets and does not reveal itself from a single perspective. Like a view onto a globe does not reveal the complete surface of our planet, but roughly just one continent at a time, MS needs to be explored from various vantage points.

29.1.4
Aims and Scope

This chapter is intended to complement the seemingly abstract description of mass spectrometers by putting them into a real-world context of analytical applications. Nonetheless, the vast number of methods in MS and the plethora of applications of MS cannot be comprehensively covered within a single chapter. However, this chapter is intended to provide a convenient entry into the realm of MS by highlighting basics, relevant techniques, and applications. Therefore, the chapter deals with

- isotopes, isotopic abundance, and isotopic mass;
- ionization processes;

- ionization methods and the corresponding ion sources;
- the behavior of the ions formed by these ionization processes; and
- applications of tandem MS.

29.2
The Role of Isotopes in Mass Spectrometry

29.2.1
Isotopes

Elements are classified by (i) their *atomic number* Z (subscripted prefix at the element symbol, e.g., $_6C$), which is equal to the number of protons in the nucleus and also is equal to the number of electrons of a neutral atom and (ii) their *mass number* A (superscripted prefix at the element symbol, e.g., ^{12}C). The mass number is given by the sum

$$A = Z + N \tag{29.1}$$

where N is the number of neutrons in the nucleus.

An element is termed *monoisotopic* if there exists only one nuclide of this element, that is, all of its atoms have equal A. This is the case for 20 of 83 naturally occurring stable elements, for example, F, Na, P, I, or Cs. If radioactive isotopes are also taken into account, there is no element left as monoisotopic. Elements with two naturally occurring stable nuclides are termed *di-isotopic,* for example, Li, B, Cl, Cu, Br, Ag, or Ir, and the others are termed *polyisotopic,* for example, Ru, Sn, or Xe. However, isotopes are distinguished by their mass number, the difference in mass being due to the varying number of neutrons in the nuclei of an element.

Isotopes of different elements sharing the same mass number are termed *isobars,* for example, ^{40}Ar, ^{40}K, and ^{40}Ca. The term also applies to molecules having different compositions yielding equal (nominal) mass, for example, CO, N_2, and C_2H_4.

29.2.2
Atomic Mass Scale

MS has been the key in establishing the atomic mass scale from the beginning (de Laeter, De Bièvre and Peiser, 1992). The atomic mass scale has been fixed relative to carbon. According to the Union of Pure and Applied Chemistry (IUPAC), the *unified atomic mass* [u] is defined as 1/12 of the mass of one atom of the nuclide ^{12}C in its electronic ground state. By defining the molar mass of carbon-12 exactly as 0.012 kg mol^{-1}, the mass of one atom is obtained by dividing this by the Avogadro constant. Therefore, 1 u equaling 1/12 of the atom's mass is obtained as $1\ u = 1/12 \times 0.012\ kg\ mol^{-1}/6.022136 \times 10^{23} = 1.660540 \times 10^{-27}$ kg. Instead of "u" the unit *dalton* (Da) (to honor J. Dalton) is also in use.

29.2.3
Isotopic Patterns

MS separates isotopes by mass, and thus, the isotopic composition of an element is directly reflected by its mass spectrum. In molecules, the isotopes of all atoms are contributing to the isotopic pattern, some more like chlorine and some less like carbon or hydrogen. Consequently, the detection of a molecular ion or fragment translates into a spectrum representing the contribution of all isotopes (Figure 29.2). It is possible

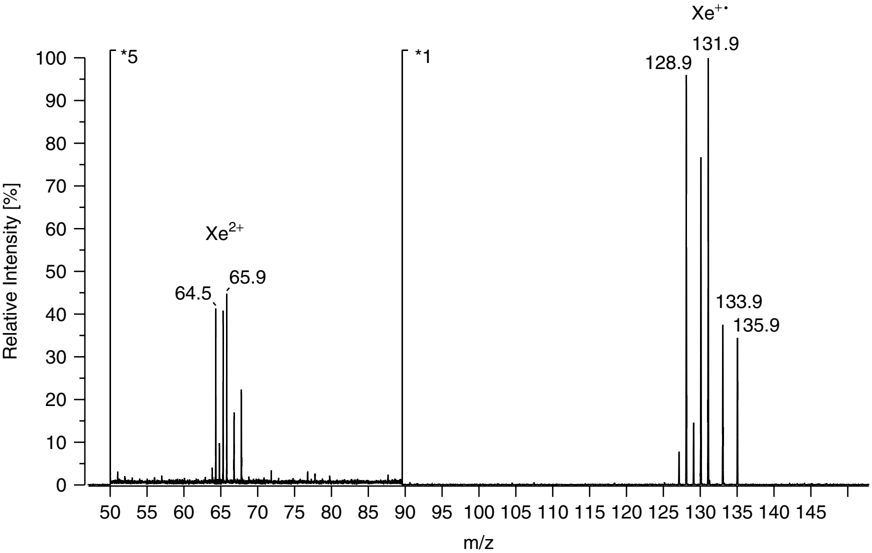

Fig. 29.2 Seventy electronvolts electron ionization mass spectrum of xenon gas. The singly as well as the doubly charged atomic ions reflect the isotopic composition of Xe from their isotopic patterns. The position of Xe^{2+} is at half m/z of $Xe^{+\bullet}$ and the distance between adjacent peaks of Xe^{2+} ion is half that of $Xe^{+\bullet}$ because $z = 2$ in m/z. Note that the intensity scale is increased by a factor of 5 for the doubly charged ion.

to calculate the isotopic pattern of a molecule based on statistical methods, because the isotopic composition of the elements is accurately known in most cases. Vice versa the experimental isotopic patterns can be fitted to theoretical distributions. In fortunate cases, in some cases, the molecular formulas of small molecules, say up to 200–300 u, can be assigned from their isotopic pattern alone.

29.2.4
Accurate Mass

Mass numbers are integers. The *accurate mass* of an isotope, however, slightly deviates from the nominal value of *A* due to the conversion of mass into energy during nucleation. The binding energy per nucleon steeply increases along the mass numbers from ^{2}H to its maximum around ^{56}Fe and then decreases again slightly up to ^{238}U (Figure 29.3).

Translation into isotopic mass reveals that for light elements mass is by some 10^{-3} u above the nominal value (^{1}H or ^{14}N), whereas it is by some 10^{-3} u (^{19}F) to 10^{-1} u (^{127}I) below for heavier elements. This *mass defect* was unveiled by Aston who had already discovered 212 of all 287 stable isotopes (Table 29.1).

To calculate the accurate mass of an ion, the electron mass ($m_e = 9.10939 \times 10^{-31}$ kg or 0.0005486 u) needs to be subtracted from the accurate mass of the corresponding neutral in the case of positive ions or added in the case of negative ions (cf. Figure 29.27).

Measuring the mass of ions with sufficient accuracy (often 0.5–10 ppm) allows to identify isotopes and molecules by

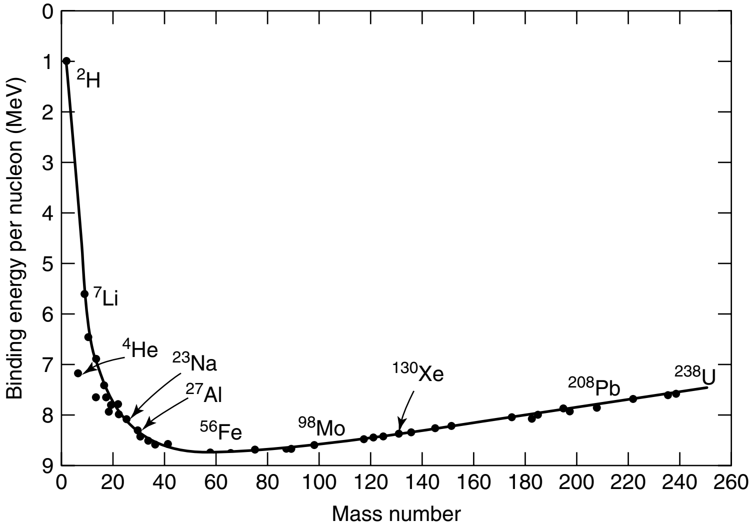

Fig. 29.3 Plot of binding energy per nucleon versus mass number. Reproduced from de Laeter, De Bièvre and Peiser (1992) by permission. © John Wiley & Sons, 1992.

their accurate mass alone (Bristow and Webb, 2003; Bristow, 2006). Assigning molecular formulas unequivocally depends on the mass of the molecule and the diversity of elements that has to be taken into account. For simple organic molecules of the general formula $C_cH_hN_nO_o$, an accuracy of 1 ppm serves this purpose up to about 400 u (Spengler, 2004). In other words, there is no second (reasonable) combination of these elements left that falls within the same mass window of ±0.0004 u. (For an example of the use of high-resolution and accurate mass also refer to Figure 29.23.)

29.2.5
Calculation of Atomic Weights and Molecular Weights

In general, chemistry does not deal with isotopic mass but with *relative atomic mass* A_r or *atomic weight* that is obtained as the average mass of the isotopic mixture of an element:

$$A_r = \frac{\sum_i m_i \times I_i}{\sum_i I_i} \tag{29.2}$$

where m_i is the accurate mass of the isotope i and I_i is its abundance. For molecules, the *relative molecular mass* M_r or *molecular weight* (MW) is obtained analogously.

29.3
Mass Spectrometric Methods

Irrespective of the mass analyzer used, the method depends on the delivery of gaseous ions from an ion source, that is, the formation of ions is a prerequisite for mass spectral analysis of any sample. As MS approaches its 100th anniversary, there is a large variety

Tab. 29.1 Illustrative examples of isotopic mass, isotopic composition, and relative atomic mass (u) of nonradioactive elements. © IUPAC 2001.

Atomic symbol	Name	Atomic number	Mass number	Isotopic mass	Isotopic composition	Relative atomic mass
H	Hydrogen	1	1	1.007825	100	1.00795
			2	2.014101	0.0115	
He	Helium	2	3	3.016029	0.000137	4.002602
			4	4.002603	100	
C	Carbon	6	12	12.000000	100	12.0108
			13	13.003355	1.08	
N	Nitrogen	7	14	14.003074	100	14.00675
			15	15.000109	0.369	
O	Oxygen	8	16	15.994915	100	15.9994
			17	16.999132	0.038	
			18	17.999116	0.205	
F	Fluorine	9	19	18.998403	100	18.998403
Na	Sodium	11	23	22.989769	100	22.989769
P	Phosphorus	15	31	30.973762	100	30.973762
Cl	Chlorine	17	35	34.968853	100	35.4528
			37	36.965903	31.96	
Fe	Iron	26	54	53.939615	6.37	55.845
			56	55.934942	100	
			57	56.935399	2.309	
			58	57.933280	0.307	
Br	Bromine	35	79	78.918338	100	79.904
			81	80.916291	97.28	
I	Iodine	53	127	126.904468	100	126.904468
Xe	Xenon	54	124	123.905896	0.33	131.29
			126	125.904270	0.33	
			128	127.903530	7.14	
			129	128.904779	98.33	
			130	129.903508	15.17	
			131	130.905082	78.77	
			132	131.904154	100	
			134	133.905395	38.82	
			136	135.907221	32.99	
Cs	Cesium	55	133	132.905447	100	132.905447
Au	Gold	79	197	196.966552	100	196.966552
U	Uranium*	92	234	234.040946	0.0055	238.0289
			235	235.043923	0.73	
			238	238.050783	100	

of mass spectrometric methods at hand to accomplish this task for almost any sample and almost any analytical task.

Basically, it is much about selecting a suitable *ionization method*, the choice being the key for the outcome of the experiment. Although some mass spectrometers are built in a way to allow for comparatively rapid changes between several ionization methods, others are tailored to fit exactly one type of ion source.

Ionization methods in MS can be grouped in many ways, the most basic being placing them into coordinates of relative softness and state of aggregation in which ionization is effected (Figure 29.4). Under plasma conditions, for example, any inorganic or organic analyte will immediately be dissociated into its constituent atomic species. It is considerably softer to expose sample vapors to energetic electrons, whereby the molecules are ionized. However, owing the amount of energy imparted during this process, the majority of the molecular ions formed may undergo fragmentation. Such ionization methods are termed *hard ionization*. On the other hand, *soft ionization* methods do not impact high levels of energy onto the analyte. Instead, they supply just enough to achieve ionization and eventually a low level of fragmentation of the ionized molecules. The softness of a method increases when vaporization of the sample is not anymore required *before* ionization. Therefore, techniques employing *desorption/ionization*, that is, the direct transition from the condensed phase into the state of isolated ions in the gas phase, are the softest available. Accordingly, the ions detected in MS span from atomic ions on the hard ionization and low-mass side, over small molecules up to large macromolecules by far exceeding 100 000 u.

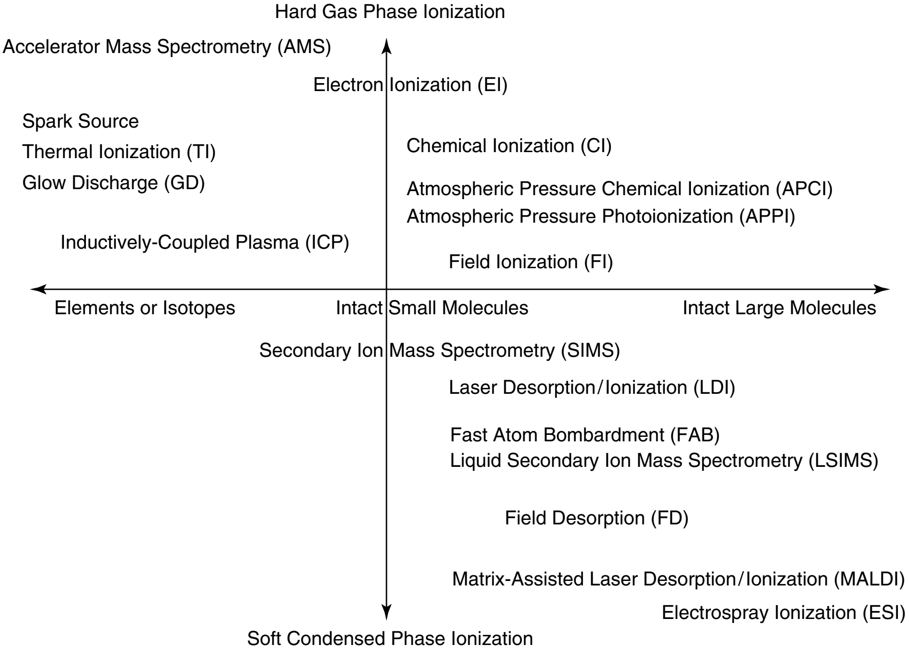
Hard Gas Phase Ionization
Accelerator Mass Spectrometry (AMS)
Electron Ionization (EI)
Spark Source
Thermal Ionization (TI)
Glow Discharge (GD)
Chemical Ionization (CI)
Atmospheric Pressure Chemical Ionization (APCI)
Atmospheric Pressure Photoionization (APPI)
Inductively-Coupled Plasma (ICP)
Field Ionization (FI)
Elements or Isotopes
Intact Small Molecules
Intact Large Molecules
Secondary Ion Mass Spectrometry (SIMS)
Laser Desorption/Ionization (LDI)
Fast Atom Bombardment (FAB)
Liquid Secondary Ion Mass Spectrometry (LSIMS)
Field Desorption (FD)
Matrix-Assisted Laser Desorption/Ionization (MALDI)
Electrospray Ionization (ESI)
Soft Condensed Phase Ionization

Fig. 29.4 Classification of common mass spectrometric methods. Although most of these methods, in particular those of organic and bioorganic mass spectrometry, in fact represent ionization methods that can be run on several types of mass analyzers, some require dedicated instrumentation, for example, SIMS as employed for imaging or AMS, which demands a large facility of its own.

29.4 Ionization Processes

There are many processes leading to the formation of charged species. Although some of these are unique to a certain ionization method, others may occur with several ionization methods. The interaction with electrons, for example, plays a role not only in *electron ionization* (EI) but also in *chemical ionization* (CI), *glow discharge* (GD), and less obvious in other technical implementations of ion generation. This section provides an overview of the most relevant ionization processes in MS.

29.4.1 Formation of Positive Radical Ions

29.4.1.1 Electron Ionization

EI (formerly termed *electron impact*) employs the interaction of energetic electrons with gas phase molecules. For *ionization efficiency* reasons, the *primary electrons* are routinely supplied at 70-eV kinetic energy. The energy transferred to the neutral upon collision with the electron forms a short-lived (<1 femtoseconds) electronically excited state. The excited neutral instantaneously emits one of its own electrons, thereby transforming into a positive radical ion, the *molecular ion* (radical ions are marked with • after the charge sign to indicate the unpaired electron).

$$M + e^- \rightarrow M^{+\bullet} + 2e^- \qquad (29.3)$$

If the energy uptake does not exceed the *ionization energy* (IE), that is, the minimum energy required to ionize a neutral in its ground state (generally 7–15 eV), it may subsequently undergo relaxation by either radiative emission or collisions with thermal neutrals (unlikely under the typical operation conditions of an EI ion source).

Multiple ionization can occur, provided sufficient energy has been imparted during interaction with the primary electron:

$$M + e^- \rightarrow M^{2+} + 3e^- \qquad (29.4)$$

$$M + e^- \rightarrow M^{3+\bullet} + 4e^- \qquad (29.5)$$

EI must not necessary lead to molecular ions. If no minimum on the energy hypersurface exists, EI can cause immediate dissociation, the so-called dissociative ionization:

$$M + e^- \rightarrow [M - B]^+ + B^\bullet + 2e^- \qquad (29.6)$$

Another pathway not leading to molecular ions is *ion-pair formation*:

$$M + e^- \rightarrow [M - C]^- + C^+ + e^- \qquad (29.7)$$

$$M + e^- \rightarrow [M - A]^+ + A^- + e^- \qquad (29.8)$$

Note that ion-pair formation simultaneously yields both cations and anions from the analyte molecule. A typical EI mass spectrum is shown in Figure 29.1. (For practical aspects of EI cf. Section 29.6.1.)

29.4.1.2 Photoionization

The absorption of UV light by a neutral can result in electronically excited states that undergo relaxation either by emission of light or by emission of an electron. Thus, the photon in *photoionization* (PI) serves the same purpose as the energetic electron in EI:

$$M + h\upsilon \rightarrow M^* \rightarrow M^{+\bullet} + e^- \qquad (29.9)$$

As with the electron before, the energy absorbed must at least provide the IE of the neutral, that is, lead into a continuum state.

However, ionization with a single photon is technically difficult, because frequency-quadrupled Nd : Yag lasers deliver photons of 266 nm wavelength (4.6 eV) and even ArF excimer lasers just yield 193 nm (6.3 eV), both being well below the IE of most molecules. Fortunately, the absorption of energy has not to be a one-photon process with a highly energetic UV photon. The stepwise accumulation of energy from less energetic photons, that is, *multiphoton ionization* (MUPI) is also feasible.

Nonresonant ionization has rather low cross sections and consequently requires high power densities of the electromagnetic radiation, or results in poor ionization efficiency. Resonant absorption of photons is more effective by orders of magnitude (Wendt, 2002). The most desirable situation is when resonant absorption of the first photon leads to an intermediate state from where absorption of a second photon, preferably but not necessarily of the same wavelength, can promote the molecule into the continuum (1 + 1 process). This approach is termed *resonance-enhanced multiphoton ionization* (REMPI). Proper selection of the laser wavelengths provides compound-selective analysis at extremely low detection limits (Boesl, 2000; Zenobi, Zhan and Voumard, 1996).

It has been shown that typical nanosecond laser pulses can be inferior to picosecond laser pulses, because the latter do not allow the molecules to undergo unwanted relaxation or fragmentation before molecular ion formation, for example, mass spectra of diphenylmercury exhibit a strong molecular ion peak when subpicosecond laser ionization is employed, whereas $M^{+\bullet}$ is completely absent at nanosecond pulse width (Weickhardt, Grun and Grotemeyer, 1998).

29.4.1.3 **Penning Ionization**

Particles occupying long-lived electronically excited states may transfer energy to ground state neutrals. Ionization of the neutral occurs if its IE is lower than the energy received upon interaction with the excited species. This is known as *Penning ionization*(Penning, 1927):

$$M + X^* \rightarrow X + M^{+\bullet} + e^- \qquad (29.10)$$

Penning ionization requires the presence of at least two different species in the gas phase and the chance for at least one collision between those particles within their dwell time in the ionization volume. The species X, for example, a noble gas, is supplied in excess to achieve a reasonable ionization efficiency for the analyte M. The energy to form the excited particles X^* may be delivered by electrons, photons, or an electric discharge. The associative variant of Penning ionization

$$M + X^* \rightarrow MX^{+\bullet} + e^- \qquad (29.11)$$

is also known as *Hornbeck–Molnar* process.

29.4.1.4 **Charge Transfer**

Soft collisions of radical ions with neutrals present another way of ion creation. The neutrals may be ionized by *charge transfer* (CT) if their IE is lower than the *recombination energy* (RE) of the neutral reagent gas X used to deliver the $X^{+\bullet}$ ions. In CT, an electron is removed from the analyte molecule M to neutralize the charge of the *reactant ion* $X^{+\bullet}$:

$$M + X^{+\bullet} \rightarrow M^{+\bullet} + X \qquad (29.12)$$

The overall process is exothermal, that is, the criterion RE > IE needs to be met. In practice, the reactant ions are generated by EI of a large excess of *reagent gas*, which is jointly admitted to the analyte gas.

29.4.1.5 **Field Ionization**

Strong electric fields in the order of $10–25\ \mathrm{V\,nm^{-1}}$ effectively enable the tunneling of an electron from a gaseous molecule into a solid electrode, because the electric field substantially lowers the potential barrier for the electron's escape. The molecule is thereby transformed into its molecular ion, $M^{+\bullet}$, but in contrast to EI this is achieved without significant deposition of internal energy onto the ion. This process is known as *field ionization* (FI). Thus, positive radical ion formation by FI is given as

$$M_{\text{in electric field}} \rightarrow M^{+\bullet} + e^- \qquad (29.13)$$

To achieve reasonable ionization efficiency, high electric field is required, because electron loss may then proceed within about 1 femtosecond. However, a minor degree of ionization is already achieved at significantly lower field strength.

29.4.2
Formation of Positive Ions

29.4.2.1 **Proton Transfer**

All ionization processes discussed so far delivered positive radical ions, because they essentially caused the loss of one electron from the prospective analyte in the gas phase. If a reactant ion can serve as a Brønsted acid, $[BH]^+$, it can act as proton donor, thus yielding a protonated analyte molecule, $[M{+}H]^+$, by exchange of the proton:

$$M + [BH]^+ \rightarrow [M + H]^+ + B \qquad (29.14)$$

The $[BH]^+$ reactant ions are supplied by multistep reactions of primary reactant ions created by EI of the respective reagent gas with neutrals of the same sort. $[M + H]^+$ ion formation is the widest known ionization process in *CI*. Proton transfer must be exothermal to occur, that is, the *proton affinity* (PA) of M must exceed the PA of X. The $[M + H]^+$ ions formed are termed *quasimolecular ions*, because they occur instead of a molecular ion, but are still intimately related to the molecule M. Different from $M^{+\bullet}$ ions, $[M + H]^+$ quasimolecular ions have intact orbitals. They are *even-electron* (or *closed-shell*) ions as opposed to their *odd-electron* (or *open-shell*) radical ion counterparts. Even-electron ions are less prone to fragmentation than radical ions.

29.4.2.2 **Electrophilic Addition**

There are various possibilities of what can happen when an ion encounters a molecule in the gas phase. An elastic collision accompanied by some exchange of kinetic or vibrational energy does not chemically alter the colliding particles. If the collision is soft enough, it can result in adhesion of both particles, that is, *electrophilic addition* in the case of a positive ion (electrophile) attaching to a neutral. This can also be regarded as solvation of the ion in the gas phase. The resulting *ion–neutral complex*, however, bears a charge and becomes accessible for mass analysis:

$$M + X^+ \rightarrow [M + X]^+ \qquad (29.15)$$

As with $[BH]^+$ ions, such X^+ ions are typically provided by multistep reactions of reactant ions initially created by EI of the respective gas. Thus, electrophilic addition presents another CI process.

29.4.2.3 Anion Abstraction

Anion abstraction may also yield a positive analyte ion, the simplest process of anion abstraction being hydride abstraction. Other anionic substituents (A) such as halogenide ions can also be removed:

$$M + X^+ \rightarrow [M - H]^+ + XH \qquad (29.16)$$

$$M + X^+ \rightarrow [M - A]^+ + XA \qquad (29.17)$$

Again, the abstracting species X^+ is provided by multistep reactions of the ionized reagent gas.

29.4.3
Formation of Negative Radical Ions

29.4.3.1 Electron Capture

Thermal electrons and electrons with energies up to a few electronvolts act in much different manner from energetic electrons. Instead of highly exciting the particle by their passage, they become attached to it, creating a negative radical ion:

$$M + e^- \rightarrow M^{-\bullet} \qquad (29.18)$$

Electron capture (EC) is a resonance process requiring that the *electron affinity* (EA) of the neutral is positive, that is, negative ions may occur with analytes containing highly electronegative elements (Harrison, 1992a).

The internal energy of the anion formed is governed by the heat of reaction of the process. Typically, EC yields molecular anions of low internal energy; in other words, it presents rather soft ionization. Nonetheless, the product of AC may spontaneously dissociate if it does not represent an energetic minimum on the energy hypersurface

$$M + e^- \rightarrow [M - B]^- + B^\bullet \qquad (29.19)$$

which is termed *dissociative EC*.

Finally, the EC can also cause ion-pair formation (cf. EI):

$$M + e^- \rightarrow [M - C]^- + C^+ + e^- \qquad (29.20)$$

$$M + e^- \rightarrow [M - A]^+ + A^- + e^- \qquad (29.21)$$

29.4.4
Formation of Negative Ions

29.4.4.1 Nucleophilic Addition

Nucleophilic addition means the attachment of a negative ion (nucleophile) to an electron-deficient neutral. In analogy to electrophilic addition, this may be regarded as solvation of the ion (nucleophile) by the neutral (analyte) in the gas phase:

$$M + A^- \rightarrow [M + A]^- \qquad (29.22)$$

The *negative-ion–neutral complex* formed can be mass-analyzed (Figure 29.5).

29.4.4.2 Cation Abstraction

Cation abstraction can lead to negative analyte ions, the simplest process of cation abstraction being proton abstraction by an anion A^-:

$$M + A^- \rightarrow [M - H]^- + AH \qquad (29.23)$$

Dissociation of acidic molecules can directly result in the corresponding anions:

$$MH \rightarrow [M - H]^- + H^+ \qquad (29.24)$$

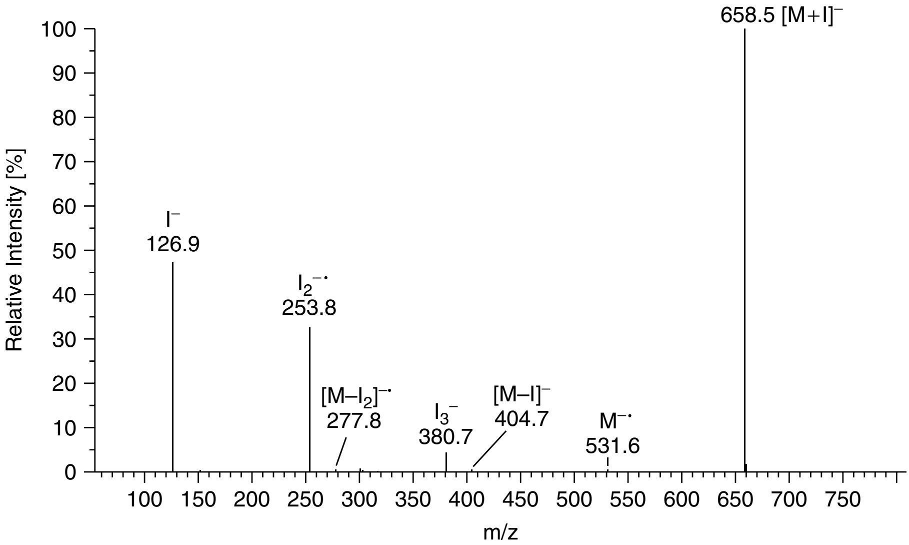

Fig. 29.5 Negative-ion chemical ionization mass spectrum of tetraiodoethene with isobutane reagent gas. The negative molecular ion at *m/z* 531.6 has a relative intensity of only 0.15%, whereas the iodide adduct $[M+I]^-$ at *m/z* 658.5 leads to the base peak of the spectrum. Losses of $I^•$ and I_2 from $M^{-•}$ are also observed. The series of peaks at *m/z* 126.9, 253.8, and 380.7 correspond to traces of iodine present as impurity of tetraiodoethene. The iodine is also ionized by both EC and iodide addition. This spectrum also shows how spectra of two components of a mixture are superimposed. It is not always trivial to tell the corresponding peaks apart. Then, additional experiments such as accurate mass measurements or tandem mass spectrometry are required.

29.5
Isotopic and Elemental Analysis

29.5.1
Thermal Ionization

When a metal salt, metal oxide, or metal placed onto the surface of a rhenium or tungsten filament is heated in vacuum to 400–2300 °C, atomic and eventually molecular ions can be generated by *thermal ionization* (TI), also known as *surface ionization*. Although TI represents another ionization process, and could have been discussed in Section 29.4, it is rather peculiar, because the TI process does not occur with any other ionization method. Basically, its application is limited to *isotope ratio mass spectrometry* (*IR-MS*) of metals, and moreover, TI ion sources are only mounted to dedicated magnetic sector instruments, usually equipped with multicollector systems to measure isotope ratios with highest accuracy (Chapters 3.3.2 and 7.2 in Platzner *et al.*, 1997).

The setup for TI consists of one to three ribbon filaments in close proximity (Figure 29.6). The analyte is supplied on the first filament and evaporated. The neutral atoms or molecules may hit the second filament where they are ionized by electron transfer to (positive ions) or from the bulk metal (anions). With triple filament assemblies, the third filament

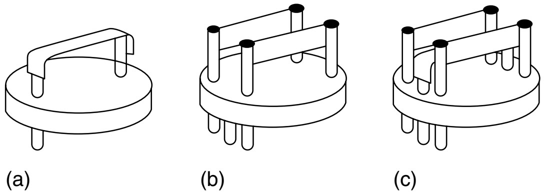

Fig. 29.6 (a) Single filament, (b) double filament, and (c) triple filament assemblies for thermal ionization of metal salts and oxides. Adapted from Platzner *et al.* (1997) p. 27 with permission. © John Wiley & Sons Ltd, 1997.

is normally used to supply a standard in order to sequentially measure unknown and reference under identical ion source conditions. In double and triple filament assemblies, the evaporation and ionization temperatures can be independently controlled. The former is preferably set to low values in order to obtain long-lasting signals and minor isotopic fractionation. The latter is set to high temperature for better ionization efficiency. Consequently, nanograms to micrograms of sample yield signals for hours as evaporation rates are in the order of picograms per second. To avoid frequent breaking of the vacuum, 10–20 filament assemblies are mounted on a carousel-like sample turret.

Blurring of the discrete electronic states of the neutral at high temperature is a key for TI, thereby allowing the Fermi levels of adsorbed neutral and bulk metal to merge. Then, the charge state of the weakly surface-bound particle is purely determined by Fermi statistics, because electrons move freely between the atom and the surface. The Saha–Langmuir equation describes the degree of ionization α^+ expressed as the ratio of ions and neutrals, leaving the surface under the assumption of ideal conditions:

$$\alpha^+ = \frac{N^+}{N^0} = \frac{g^+}{g^0}\exp\left[\frac{e\,(\Phi - IE)}{kT}\right] \tag{29.25}$$

$$= \frac{g^+}{g^0}\exp\left[1.16\times 10^4\frac{(\Phi - IE)}{T}\right] \tag{29.25a}$$

where g^+/g^0 is the ratio of electronic states of the ion and the neutral, Φ is the work function of the filament material, IE is the ionization energy of the atom or molecule to be ionized, and T is the temperature of the filament. As a consequence of this equation, good ion yields are obtained at high temperature with metals of low IE and filament materials with large Φ, which is fulfilled for rhenium ($\Phi_{\mathrm{Re}} = 4.98$ eV, mp = 3180 °C) and tungsten ($\Phi_{\mathrm{W}} = 4.58$ eV, mp = 3410 °C) (Figure 29.7).

Nonmetallic elements with high IE and metal oxides may form negative ions (Kawano and Page, 1983; Kawano, Hidaka and Page, 1983). The degree of ionization α^- is then obtained from the following expression:

$$\alpha^- = \frac{g^+}{g^0}\exp\left[1.16\times 10^4\frac{(EA - \Phi)}{T}\right] \tag{29.26}$$

where EA is the electron affinity of the adsorbed atom.

The low spatial and energetic spread of ions from TI is advantageous for the combination with single-focusing magnetic sector mass analyzers. On the other hand, TI is not suitable for mixture analysis and multielement determinations where ICP-MS is normally employed.

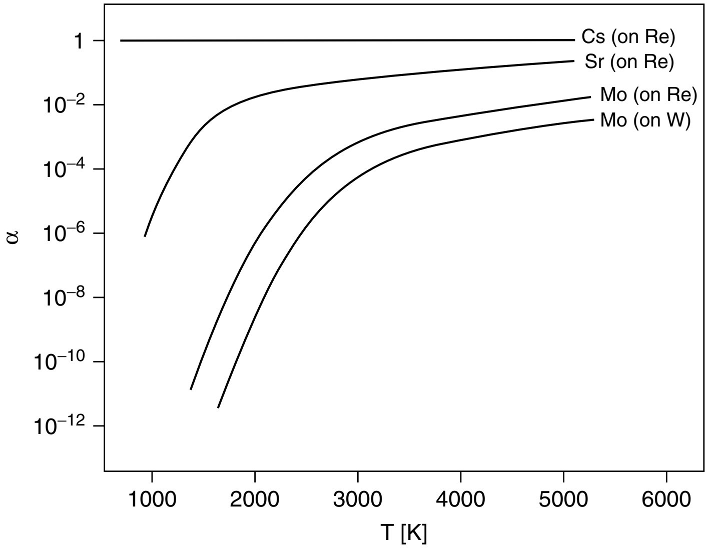

Fig. 29.7 Thermal ionization efficiency α versus filament temperature. The curves for molybdenum reflect the dependence of α on the filament material. The advantage of the Re over the W filament is in accordance with Eq. (29.25). Adapted from Becker (2008) with permission. © John Wiley & Sons Ltd, 2008.

29.5.2
Spark Source

Developed by Dempster in 1935, the spark source provided the first truly multielement and isotopic trace element tool (Koppenaal, 1990). *Spark source mass spectrometry* (SS-MS) is best suited for conducting samples, because it employs a high-voltage electric discharge between two pin-shaped electrodes (about 10 mm long and 1–2 mm in diameter) in vacuum. The electrodes, directly prepared from the solid to be analyzed, need careful cleaning, for example, by etching, before being inserted into the ion source (Chapter 5.2 in de Laeter, 2001, Chapter 2.2 in Becker, 2008) (Figure 29.8).

Nonconducting samples are ground and mixed with high-purity graphite, silver, or gold powder. The mixture is then pressed to yield the electrodes. It is obvious from this procedure that SS-MS is not suited for elemental analysis of small sample amounts; nonetheless, it plays to its strengths for simultaneous analysis of many elements including those present at trace level in alloys, ores, and similar samples.

From the cathode, electrons are emitted by field emission (opposite process to FI), which is possible due to the high local field strengths (10^8–10^9 V m^{-1}) at edges of the microscopically rough surface. The electrons are accelerated toward the anode where they cause evaporation of sample that is then ionized in the gas phase by action of the arriving electrons.

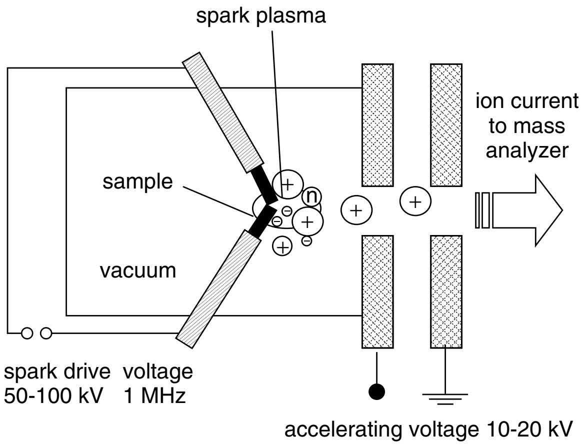

Fig. 29.8 Schematic of a RF spark source. Adapted from Becker (2008) with permission. © John Wiley & Sons Ltd, 2008.

Although operation of a spark source appears simple initially, it is quite demanding. First, it requires a high-voltage supply of several tens of kilovolts, normally at 1 MHz radiofrequency (RF) to achieve uniform ablation of both electrodes. Second, the electrodes need to be carefully adjusted and dynamically readjusted during the measurement (0.1–0.5 mm tip to tip) to maintain a stable ion current. Then, the ions emerge from the discharge with a wide kinetic energy distribution (keV range) that makes a double-focusing mass analyzer mandatory for SS-MS. Still, photoplate detection and subsequent densiometry on a series of exposures, that is, on subsequent spectra acquired for increasing periods of time, are still widespread in SS-MS (Koppenaal, 1990; Verlinden, Gijbels and Adams, 1986). However, multi-ion counting (MIC) systems are promising as they deliver improved sensitivity and precision (Jochum, 1997; Jochum *et al.*, 2001).

29.5.3
Glow Discharge

In 1886, Goldstein was the first to report the formation of positive ions of gases from a *GD* in his so-called channel ray tube (p. 25 in Becker, 2008).

In a GD-MS ion source (Jakubowski *et al.*, 2007; Nelis and Pallosi, 2006; Hoffmann *et al.*, 2005; Stuewer, 1990; King and Harrison, 1990), a discharge current of 1–5 mA is maintained through a dilute argon atmosphere (10^1–10^3 Pa) by applying a DC voltage of 500–1000 V across the ionization volume. The cathode is represented by the solid sample that may be either a disk or a stick and an anode present as the housing and the ion source exit plate (Chapter 5.3 in de Laeter, 2001, Chapter 7.5 in Platzner *et al.*, 1997).

The low-temperature Ar plasma serves as a source of primary ions, excited atoms, and electrons. In the cathode dark space, a region where a large potential drop exists, the positive argon ions are

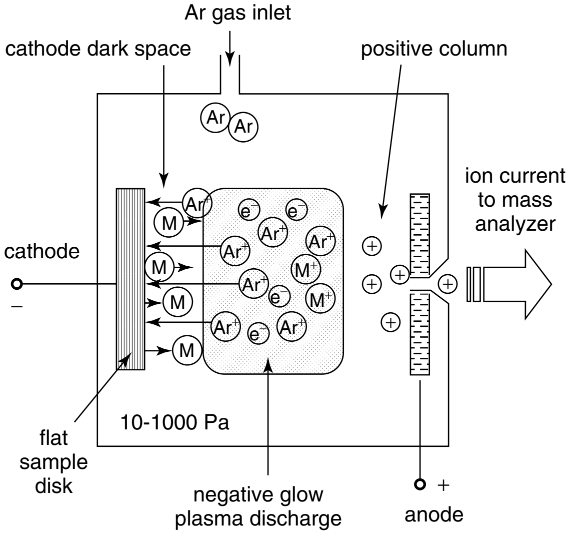

Fig. 29.9 GD ion source with a disk-shaped cathode. Adapted from Becker (2008) with permission. © John Wiley & Sons Ltd, 2008.

accelerated toward the sample cathode (Figure 29.9).

By bombarding the electrode's surface, the impinging ions provide continuous sputtering of neutrals into the gas phase. As the liberated analyte atoms move into the negative-glow region, they may be ionized by either EI or Penning ionization. The negative-glow region emits a bright glow, a phenomenon that led to the construction of the neon lamp. The electron density in the plasma can reach up to 10^{14} cm^{-3} (Penning, 1927) and their energy covers a wide distribution from thermal to energetic. Although their presence in the plasma is less by about a factor of 1000, metastable Ar* atoms are still effective for ionization because they carry 11.55 eV.

Nonconducting samples would rapidly cause the breakdown of the discharge due to electrostatic charging. RF discharges circumvent this problem (Winchester and Payling, 2004). Typically, voltages of 2 kV at RF of 13 MHz are employed.

29.5.4
Inductively Coupled Plasma

Currently, the *inductively coupled plasma* (ICP), developed by Houk *et al.* (1980), probably presents the most frequently employed method of element MS (Chapter 4 in de Laeter, 2001, Chapter 5 in Becker, 2008, Chapter 7.3 in Platzner *et al.*, 1997). Its widespread use is due to a highly versatile sampling mode that, in addition, is comparatively simple and robust because atomization and ionization are achieved in an argon plasma at atmospheric pressure. The ICP provides high ionization efficiency for elements of low IE (0.98 for 7 eV). Although these values drastically drop (0.01 for 13 eV), it is still sufficient to access elements such as P and even Cl.

The plasma is maintained by coupling electric energy (1–2 kW) from a surrounding RF (27 MHz) coil into the gas. The ion motion induced by the magnetic field causes the gas phase to heat up until it forms a continuously flowing plasma of about 8000 K at the center. The analyte is introduced axially into this plasma by a nebulizer that dissipates solutions and suspensions or by admitting gaseous samples.

The plasma torch consists of a coaxially aligned set of three quartz tubes, the outer of which is about 20 mm in diameter and is surrounded by the water-cooled RF coil. An argon stream of 12–15 l min^{-1} is used to cool the walls. The middle tube supplies a stream of Ar (1–2 l min^{-1}) as nebulizer gas to assist the formation of droplets at a size of tens of micrometers from the sample aerosol that is admitted by another gentle Ar stream from the 1–2 mm i.d. center tube (Figure 29.10).

The ions are transferred into the mass analyzer via a differentially pumped interface. A small portion of the plasma enters the first pumping stage through a

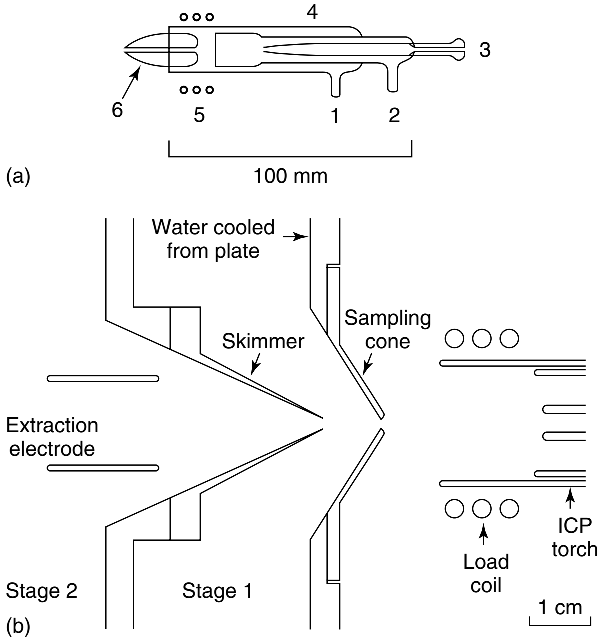

Fig. 29.10 Schematic diagram of (a) an ICP torch where 1 and 2 are inlets for argon sheath and nebulizer gas, 3 inlet for the sample, 4 the outer quartz tube of the torch, 5 the RF coil, and 6 the plasma. (b) An ion extraction interface for ICP-MS. Ions arrive from the right, a small portion is transmitted through the water-cooled sampling cone into the first pumping stage (rotary pump), the center of the free-jet expansion is then transmitted through the second skimmer cone, behind which it is guided by electric potentials through the second pumping stage (turbo pump) into the mass analyzer. Reprinted with permission from Platzner *et al.* (1997) p. 160. © John Wiley & Sons Ltd., 1997.

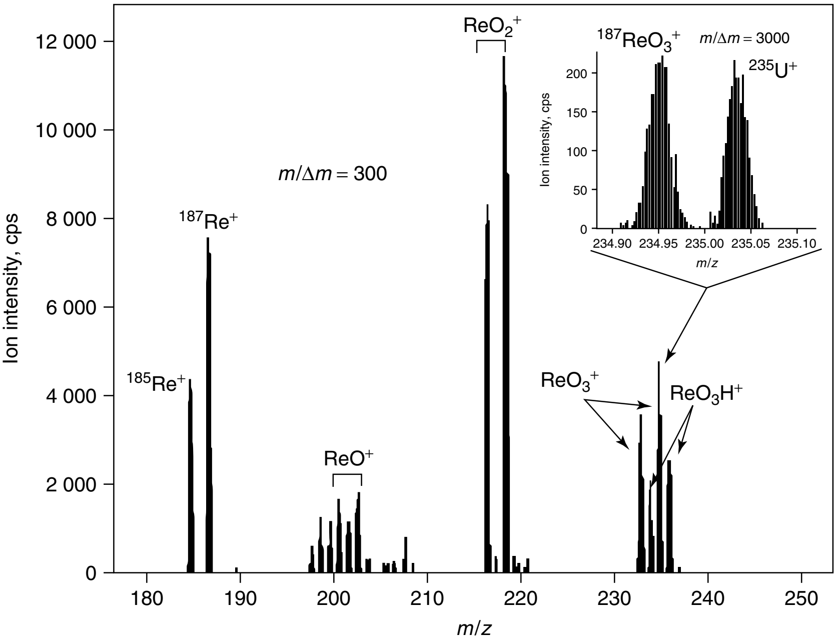

Fig. 29.11 Partial ICP mass spectrum of a radioactive waste solution in the range m/z 180–250 as obtained in low-resolution mode. The insert showing the m/z 235 range is taken from a spectrum as obtained from the same magnetic sector instrument when operated in high-resolution mode at $R = 3000$. Then, the signals due to ^{187}ReO$_3^+$ and ^{235}U$^+$ can easily be distinguished. Adapted from Becker and Dietze (1999) with permission. © The Royal Society of Chemistry, 1999.

hole in the water-cooled sampling cone. Application of suitable electric potentials guides the ions through the entrance of the skimmer, whereas neutrals are pumped off the supersonic expansion in this region. The first implementation of ICP-MS used a quadrupole analyzer due to its acceptance of moderate vacuum conditions (Houk *et al.*, 1980). In between, magnetic sector (Figure 29.11) (Becker and Dietze, 1999), time of flight (TOF), and Fourier-transform ion cyclotron resonance (FT-ICR) mass analyzer are also employed.

To further complete the sampling repertoire, a gas stream may be employed to take up neutrals generated by *laser ablation* (LA). LA-ICP-MS can be used for sampling from solid surfaces of bulk material or alternatively for imaging of element distributions in the upper layers of tissues and similar samples (Figure 29.12). Thus, LA-ICP-MS is being used in biomedical research (Guenther and Hattendorf, 2005; Chery *et al.*, 2006; Becker *et al.*, 2007).

Another advantage of the acceptance of liquid introduction is the coupling of LC and gel electrophoresis to ICP instruments to allow for *element speciation.* This way not only the presence of an element in a sample can be detected but also the assignment to molecular species becomes feasible. This is relevant for the identification of non-C,H,N,O elements in proteins, for example (Ray *et al.*, 2004).

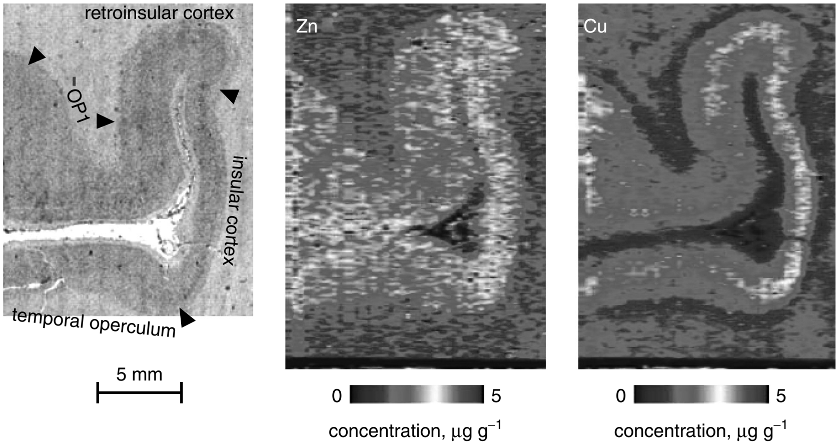

Fig. 29.12 Quantitative images of Zn and Cu in the hippocampus of human brain. A slice of 20 μm thickness was measured by LA-ICP-MS as compared to the light photograph (left). Reproduced from Becker *et al.* (2007) with permission. © The Royal Society of Chemistry, 2007.

29.5.5
Accelerator Mass Spectrometry

Accelerator mass spectrometry (*AMS*) presents an extremely sensitive technique for isotopic analysis (Tuniz *et al.*, 1998). AMS allows to measure isotope ratios for some elements to a level of 1 in 10^{15}, which is by a factor of 10^5 lower than in other mass spectrometers. Moreover, this can be achieved for samples of about 1 mg, that is, only one million atoms of the isotope of interest are present, making it ideal for isotope ratio measurements of low-abundant isotopes, especially those having too long half-lives for counting radioactive decay from small sample amounts such as carbon-14.

The methodology of AMS differs greatly from all other approaches to the analysis of atomic mass. The main characteristic of AMS lies in the energies to which the ions are accelerated. Although other mass analyzers deal with ions having 10 eV–10 keV, in AMS they are at several megaelectronvolts. To achieve this, negative ions are created in an ion source and are mass-analyzed before entering a linear *Van de Graaff accelerator* or a *tandem accelerator*. In a tandem accelerator, negative ions are attracted by positive voltage toward the so-called high-voltage terminal at the center of the vessel. By passing the beam through a thin carbon foil (foil stripper) or a gas collision cell (gas stripper), electrons are stripped off the anion, thereby converting it to a positive ion. The atomic cations are then pushed away from the high-voltage terminal toward the exit of the accelerator. These are then mass-analyzed in a unit that basically represents a large magnetic sector instrument. The typical AMS setup requires some hundred square meters (Figure 29.13).

The higher energies enable ambiguities in the identification of atomic ions to be avoided, because molecular species are completely destroyed at the stage of the stripper. Furthermore, the selection

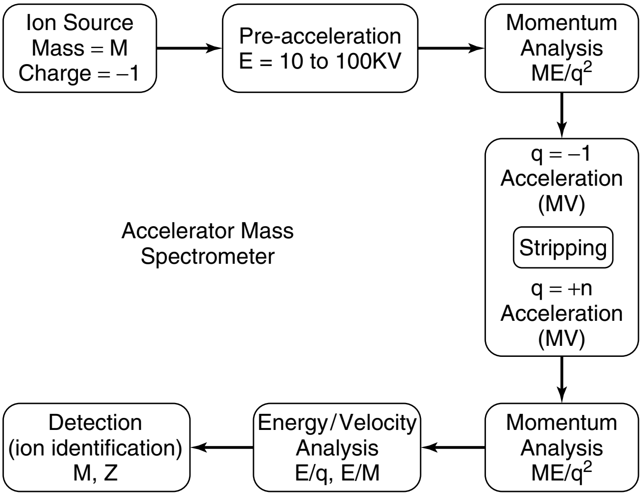
Ion Source
Mass = M
Charge = −1
Pre-acceleration
E = 10 to 100KV
Momentum
Analysis
ME/q²
q = −1
Acceleration
(MV)
Stripping
q = +n
Acceleration
(MV)
Accelerator Mass
Spectrometer
Momentum
Analysis
ME/q²
Energy/Velocity
Analysis
E/q, E/M
Detection
(ion identification)
M, Z

Fig. 29.13 Schematic layout of equipment for AMS. Ions experience marked changes in kinetic energy and charge state during the experiment. In total, the sample is ionized, and ions are preselected in a magnetic sector, accelerated to mega electronvolt energies, charge-inversed in the stripper, separated according to their momentum, charge, and energy, and finally counted individually (Tuniz *et al.*, 1998).

of a suitable high charge state, characteristic of the isotope of interest, permits unequivocal identification, for example, ^{14}C, ^{26}Al, and ^{129}I are already separated from isobaric ^{14}N, ^{26}Mg, and ^{129}Xe due to the instability of the respective negative ions.

Carbon-14 dating was the application that gave birth to AMS (Nelson, Korteling and Stott, 1977; Bennett *et al.*, 1977). Today, AMS is also employed for the determination of isotopes such as ^{10}B, ^{26}Al, ^{36}Cl, ^{41}Ca, ^{129}I, and others including stable isotopes. This versatility has led to applications of AMS in many areas of research including life sciences, geosciences, archeology, and extraterrestrial materials (Brown, Dingley and Turteltaub, 2005; Brown, Tompkins and White, 2006; Ikeda, 2005; Kutschera, 2005).

29.6
Ionization of Gaseous Molecules in Vacuum

All classical ionization methods of organic MS imply ionization in the diluted gas phase. They are applied for small molecule analysis ($M < 800$ u), provided the analyte does not contain too many heteroatoms. Molecules with more than a few oxygen, nitrogen, or sulfur atoms (highly functionalized molecules) are normally very polar – a property preventing their evaporation without prior thermal decomposition. Furthermore, molecular ions of such species often undergo facile fragmentation making them difficult to recognize in the mass spectrum. Thus, the important assignment of molecular mass is rendered impossible and the structure elucidation based

Tab. 29.2 Methods of sample evaporation.

Inlet system	Principle of operation	Cases where applicable	Limitations
Reservoir inlet (also termed *reference inlet, gas inlet*)	Injection into a heated reservoir and admission to the ion source via a needle valve. Inlet is emptied by evacuation	Gases and liquids	Risk of thermal decomposition of the sample. Memory from previous samples
Direct probe (also termed *direct insertion probe or direct inlet system*)	Evaporation from a small vial placed onto the temperature-controlled tip of a probe which is inserted through a vacuum lock	High-boiling liquids, waxes, solids	Risk of thermal decomposition of the sample
Gas chromatograph	Sample solutions are injected into GC and elute from the chromatographic column into the ionization volume	Compounds that can pass through a GC column	Chemical derivatization can be required. Time span of elution is only about 1 s

on fragment ions alone normally fails. Table 29.2 compares methods of sample evaporation.

29.6.1
Electron Ionization

EI, formerly termed *electron impact* (*EI*), employs the interaction of energetic electrons with gas phase molecules. The electrons are generated by thermionic emission from a heated tungsten or rhenium filament. In EI, ions isolated in the gas phase should be formed to avoid ion–molecule reactions. Therefore, the pressure in the ionization volume of an EI ion source is maintained in the range of 10^{-5}–10^{-6} mbar so that their mean free path (about 1 m) is by far longer than the average distance traveled within the ionization zone of ion source (in the order of 1 mm). Their ion source dwell time is about 1 microsecond due to the extracting electric field accelerating them from the zone of creation to generate a continuous ion beam into the mass analyzer (Figure 29.14).

Quantitative energy transfer is required for ionization to occur, when an electron possessing energy just equal to IE collides with a neutral. Such an event is of low probability and, therefore, the *ionization efficiency* is then close to zero. However, a slight increase in electron energy brings about a steady increase in ionization efficiency. Every molecule has an ionization efficiency curve of its own (Figure 29.15). The overall efficiency of EI depends on the intrinsic properties of the ionization process as well as on the ionization cross section of the analyte. Fortunately, the curves of ionization cross section versus electron energy are very similar in shape, exhibiting a maximum at electron energies around 70 eV. In addition, the plateau of the ionization efficiency curve around

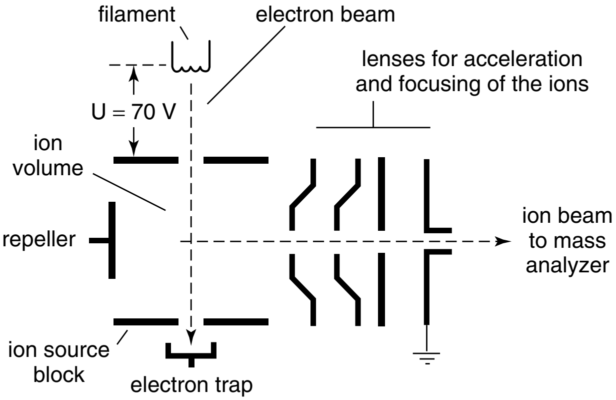
filament
electron beam
U = 70 V
lenses for acceleration and focusing of the ions
ion volume
repeller
ion beam to mass analyzer
ion source block
electron trap

Fig. 29.14 Schematic drawing of an EI ion source. Adapted with permission from Schröder (1991). © Springer-Verlag, 1991.

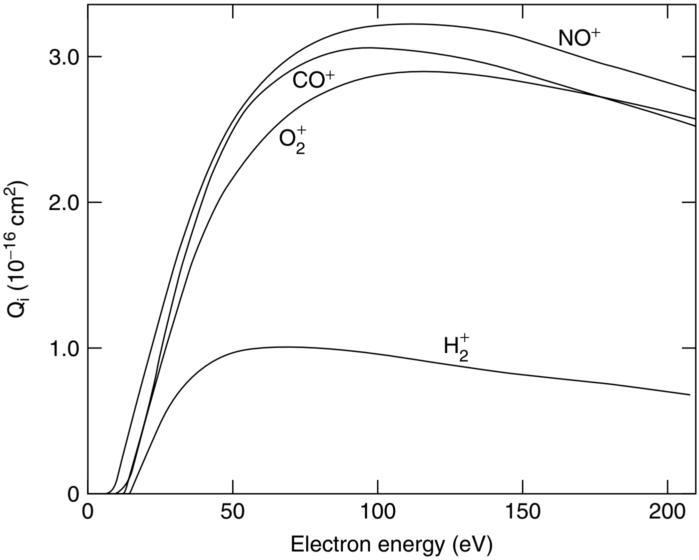

Fig. 29.15 Ionization efficiency curves for NO, CO, O_2, and H_2. Adapted with permission from Kiser (1965). © Prentice-Hall, 1965.

70 eV makes variations in electron energy negligible; in practice, EI works equally well at 60–80 eV. This assures better reproducibility of spectra and, therefore, allows comparison of spectra obtained from different mass spectrometers or from mass spectral databases. Therefore, EI spectra are almost exclusively acquired at 70 eV.

Once formed, the molecular ions "experience a rich life" during this short time span (Figure 29.16).

It takes less than 10^{-15} seconds for an electron of 70-eV kinetic energy to

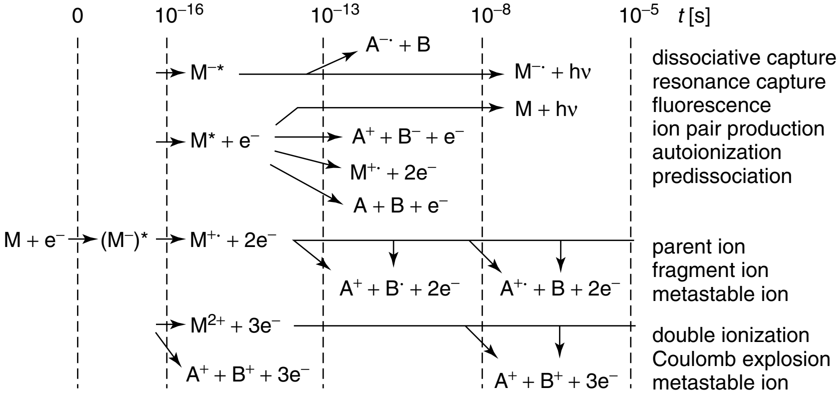
0
10⁻¹⁶
10⁻¹³
10⁻⁸
10⁻⁵ t[s]
A⁻˙ + B
M⁻*
M⁻˙ + hν
M + hν
M* + e⁻
A⁺ + B⁻ + e⁻
M⁺˙ + 2e⁻
A + B + e⁻
M + e⁻
(M⁻)*
M⁺˙ + 2e⁻
A⁺ + B˙ + 2e⁻
A⁺˙ + B + 2e⁻
M²⁺ + 3e⁻
A⁺ + B⁺ + 3e⁻
A⁺ + B⁺ + 3e⁻
dissociative capture
resonance capture
fluorescence
ion pair production
autoionization
predissociation
parent ion
fragment ion
metastable ion
double ionization
Coulomb explosion
metastable ion

Fig. 29.16 Schematic time chart of events possible upon electron ionization. Adapted from Märk (1986) with permission. © John Wiley & Sons Inc, 1986.

pass through a molecule. Vibrations, the fastest internal motions of a molecule, are slower by at least two orders of magnitude, for example, a C–H stretching vibration takes roughly 1.1×10^{-14} seconds per cycle. Thus, the electronic motion and the nuclear motion are in fact separated, that is, the *Born–Oppenheimer approximation* is valid (Born and Oppenheimer, 1927; Seiler, 1969). Now, the *Franck–Condon principle* states that the probability for an electronic transition is highest where the electronic wave functions of both ground state and ionized state have their maxima. Combining these tenets explains the occurrence of the so-called *vertical transitions* (Franck, 1925; Condon, 1926). (Also cf. Section 29.4.1.1 for processes under EI conditions.) Although there is a wealth of chemical reactions, all of them are strictly unimolecular, because the mean free paths are long enough to prevent bimolecular interactions within the lifetime of the ions. Ions of very low internal energy reach the detector and are thus termed *stable ions* (within the below timescale); those that have life times long enough to leave the ion source but decompose during their flight through the analyzer are termed *metastable ions*, and those that fragment inside the ion source are called *unstable ions*.

The kinetics of these reactions are commonly described by the *quasi-equilibrium theory* (QET) (Wahrhaftig, 1986; Märk, 1986; Rosenstock and Krauss, 1963; Wahrhaftig, 1959).

Ions with moderate internal energy undergo decomposition reactions along well-defined pathways, leading to *fragment ions* having structures that are still closely related to the intact molecule. Interpretation of EI spectra following these routes and some well-established rules can therefore lead to the identification of a compound (McLafferty and Turecek, 1993, Chapter 6 in Gross, 2004). A sample EI spectrum of the over-the-counter drug ibuprofen elucidates the basic principle (Figure 29.1).

29.6.2
Chemical Ionization

There are various possibilities of what can happen when an ion encounters a molecule in the gas phase. An elastic collision accompanied by some exchange of kinetic or vibrational energy does not chemically alter the colliding particles. However, if the collision is soft enough it can result in adhesion of both particles or in exchange of charge via mutual transfer of either an electron or some (small) ions such as a proton. All these processes contribute to the manifold approaches to *CI*.

To enable bimolecular processes within the microsecond timeframe, the pressure inside the ion source has to be substantially raised to 10^{-4}–10^{-3} mbar by the introduction of a reagent gas that is then present in more than 1000-fold excess over the analyte. The analyte molecules undergo in the order of 50 collisions (as opposed to none in EI) and, thus, become thermalized and accessible to ion–molecule reactions. To avoid excessive leakage into the remaining mass spectrometer, the ionization volume of a CI ion source has only small holes for the entrance and exit of electrons, reagent gas, sample vapor, and ions produced (Figure 29.17). Sufficient pumping capacity is required at the ion source housing to maintain high vacuum conditions.

Nowadays, there are numerous variations of *CI* introduced in the mid-1960s (Field and Munson, 1965; Harrison, 1992b). *Positive-ion chemical ionization* (PICI or PCI) (Harrison, 1992b; Munson, 2000) employs either proton transfer (Section 29.4.2.1), electrophilic addition (Section 29.4.2.2), and anion abstraction (Section 29.4.2.3) or CT (Section 29.4.1.4) to generate analyte ions. The reagent gases methane and isobutane are preferred for $[M + H]^+$ ion formation, ammonia tends to yield $[M + NH_4]^+$ by electrophilic addition, and gases such as benzene, carbondisulfide, or noble gases serve for the production of $M^{+\bullet}$ ions by CT (Allgood, Ma and Munson, 1991; Hsu and Qian, 1993; Roussis, 1999).

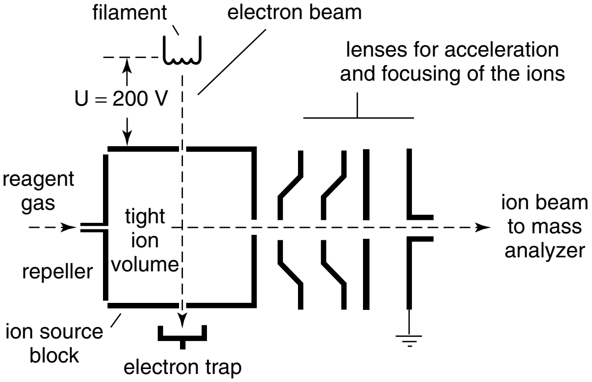

Fig. 29.17 A CI ion source can be readily obtained by replacing the ionization volume of an EI ion source with a tighter one. Many mass spectrometers have EI/CI combination ion sources switching between these methods by pushing a cylinderlike insert into the EI volume in order to close its vents (cf. Figure 29.14). Adapted with permission from Schröder (1991). © Springer-Verlag, 1991.

Negative-ion chemical ionization (NICI or NCI) (von Ardenne, Steinfelder and Tümmler, 1971; Dillard, 1973) makes use of cation abstraction (Section 29.4.4.2), nucleophilic addition (Section 29.4.4.1 and Figure 29.5), or EC (Section 29.4.3.1), the latter being no CI in the strict sense, because the reagent gas (often isobutane) solely moderates the electrons to thermal energy instead of involving any reaction (Knighton and Grimsrud, 1995). Alternatively, the electrons may be delivered from an electron monochromator (Laramée *et al.*, 1996).

The goal of CI is to generate quasimolecular ions or molecular ions of low internal energy in order to reduce the degree of fragmentation. The corresponding ions generally are well recognized in CI spectra of both polarities while the moderate fragmentation still assists structure elucidation of unknowns. It is common to combine the information obtained from both EI and CI spectra (Figure 29.18), or from positive-ion plus negative-ion CI spectra to improve the reliability of the mass spectral analysis.

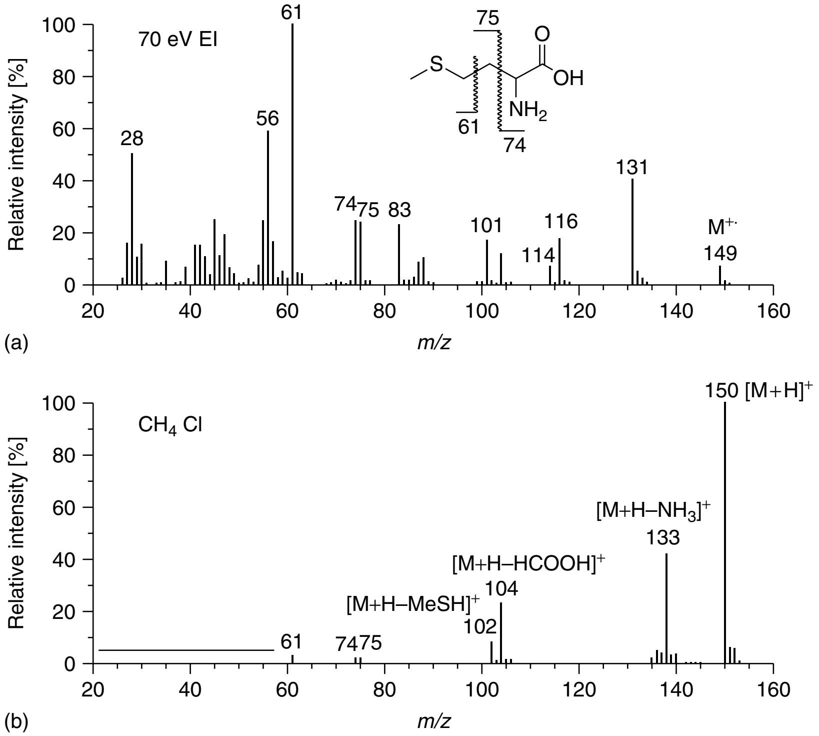

Fig. 29.18 Comparison of (a) 70-eV EI mass spectrum and (b) methane reagent gas CI mass spectrum of the amino acid methionine. The range up to *m*/*z* 60 has been excluded from the CI spectrum, because it is dominated by peaks due to the reagent gas. The reduction in fragmentation as effected by the generation of the closed-shell quasimolecular ions, $[M + H]^+$, in CI is obvious. This also results in easier identification of the sulfur isotopic pattern. Reprinted from Gross (2004) with permission. © Springer-Verlag, 2004.

29.6.3
Field Ionization

In FI, molecular ions, $M^{+\bullet}$, are softly created from gas phase neutrals by action of strong electric fields causing tunneling of an electron from a gaseous molecule into a solid electrode, the field anode (Section 29.4.1.5) (Gomer and Inghram, 1955; Beckey, 1962; Beckey, 1971; Lattimer and Schulten, 1989; Gomer, 1994). Although some micrometer-thin metal wires were employed as electrodes in pioneering work on FI, the use of activated emitters – occasionally specially optimized for FI usage – has become standard in FI-MS. The process of *emitter activation* transforms a tungsten wire of 5–13 μm in diameter into a rather fragile object resembling a hairy caterpillar species (Beckey, Krone and Röllgen, 1968; Beckey, Hilt and Schulten, 1973; Okuyama *et al.*, 1977; Rabrenovic, Ast and Kramer, 1981). These *whiskers* grown on the central wire cause an enormous increase in the electric field strength at their tips as compared to bare wires, because the radii of curvature are drastically reduced. Thus, the activated emitter provides a large number of high electric field sites for FI to occur.

The setup of an FI ion source comprises the field emitter opposed to a counter electrode set to a potential of 10–12 kV negative relative to the emitter (Figure 29.19). The emitter can be resistively heated up to about 1000 °C by passing a current of 50–90 mA through it in order to bake off any condensed sample residues between subsequent measurements. During operation, the resistive heating is normally switched off or set to low levels. The sample vapor can be introduced by any of the means listed in Table 29.2.

As long as the process of sample evaporation is not limiting, FI spectra are mainly characterized by the molecular ion peak of the analyte, eventually accompanied by few generally low-abundant fragment ions (Figure 29.20).

Owing to some degree of surface adhesion in FI, there are two accompanying processes influencing the spectra. On one side, ions formed may encounter a second step of ionization, leading to M^{2+} ions. On the other side, bimolecular processes can cause the formation of $[M+H]^+$ ions analogous to the generation of protonated

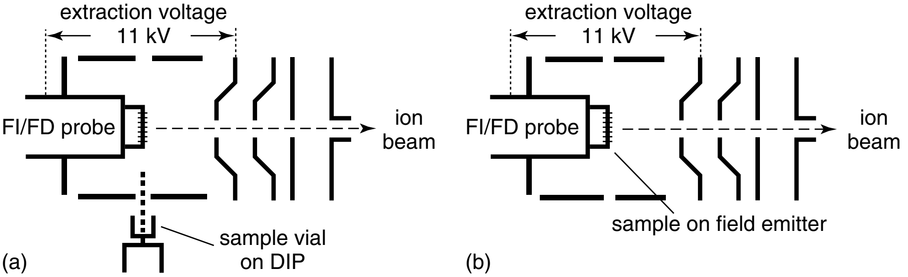

Fig. 29.19 Schematic drawings of (a) FI and (b) FD ion sources. Field ionization has lead to the development of *field desorption* (FD) where the sample is directly placed onto the emitter surface from dilute sample solution by evaporation of the solvent (Section 29.7.1). Adapted from Schröder (1991) with permission. © Springer-Verlag, 1991.

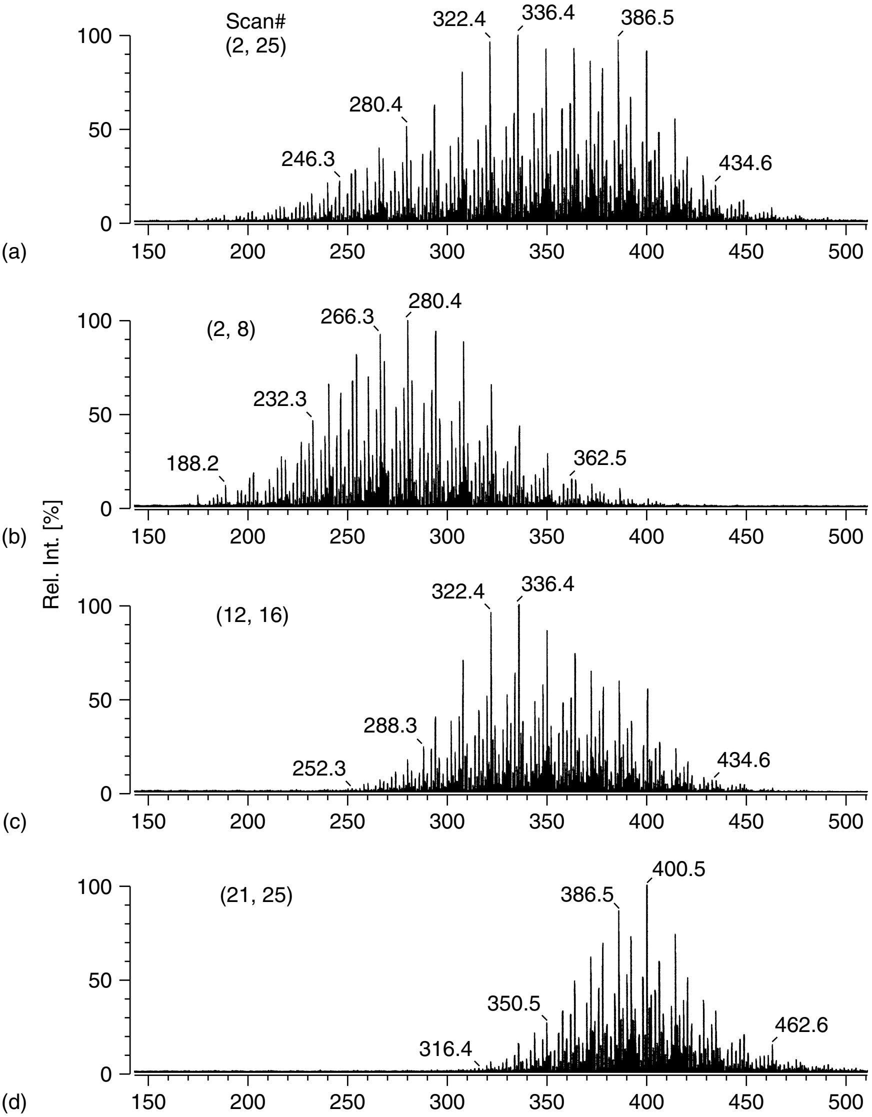

Fig. 29.20 Typical fractionation during the acquisition of the FI mass spectrum of a standard sewing machine oil when evaporated from a direct insertion probe. The majority of peaks corresponds to molecular ions. (a) A spectrum from summing up scans 2–25. (b)–(d) Spectra are sums of a few scans during evaporation of the oil upon heating of the sample vial on the probe tip. Compounds of lower molecular weight evaporate earlier than those of higher molecular weight.

reagent gas molecules in CI (Schulten and Beckey, 1974). Although doubly charged ions are easily recognized in mass spectra by their spacing of isotopic peaks (cf. Section 29.2.3 and Figure 29.2), the partial protonation can lead to misinterpretation of isotopic patterns.

29.7
Desorption/Ionization

When the generation of gas phase ions can be accomplished by combining both evaporation and ionization into a single process, thermal decomposition before ionization can be avoided for the most part. Although CI and FI already belong to the so-called soft ionization methods, these do not possess this key property, and many of their applications have soon been replaced by *desorption/ionization* methods when they became commercially available. In contrast to EI, CI, and FI, desorption/ionization methods handle ionic analytes very well. In fact, ionic analytes are even beneficial, because they only require transfer into the gas phase, while the actual ionization step is not necessary anymore.

The numerous types of ions formed by the various desorption/ionization methods can be conveniently organized by their relevance for analytes of given polarity and charge (Table 29.3). FD (Section 29.7.1) covers the table's left column only, while LSIMS and FAB (Section 29.7.2) as well as LDI (Section 29.7.3) apply to the whole table with a compound-dependent high-mass limit of about 2000–5000 u. Finally, MALDI (Section 29.7.4) can access the whole polarity–analyte mass range, although it is not ideally suited for unpolar analytes and molecules below 500 u.

29.7.1
Field Desorption

FI has been further developed to *field desorption* (FD) where the sample is directly placed onto the emitter surface

Tab. 29.3 Ions formed by desorption/ionization.

Analytes	Positive ions	Negative ions
Nonpolar	$M^{+\bullet}$	$M^{-\bullet}$
Medium polarity	$M^{+\bullet}$ and/or $[M + H]^+$,	$M^{-\bullet}$ and/or $[M - H]^-$,
	$[M + \text{alkali}]^+$,	
	less frequently also,	less frequently also,
	clusters $[2M]^{+\bullet}$ and/or $[2M + H]^+$,	*clusters* $[2M]^{-\bullet}$ and/or $[2M - H]^-$,
	$[2M + \text{alkali}]^+$,	
	adducts $[M + Ma + H]^+$,	*adducts* $[M + Ma]^{-\bullet}$,
	$[M + Ma + \text{alkali}]^+$	$[M + Ma - H]^-$
Polar	$[M + H]^+$, $[M + \text{alkali}]^+$,	$[M - H]^-$,
	exchange $[M - H_n + \text{alkali}_{n+1}]^+$,	*exchange* $[M - H_n + \text{alkali}_{n-1}]^-$,
	clusters $[nM + H]^+$,	*Clusters* $[nM - H]^-$,
	$[nM + \text{alkali}]^+$,	
	adducts $[M + Ma + H]^+$,	*adducts* $[M + Ma - H]^-$,
	$[M + Ma + \text{alkali}]^+$,	
	High-mass analyte $[M + nH]^{n+}$,	*High-mass analyte* $[M - nH]^{n-}$,
	$[M + n\text{alkali}]^{n+}$	$[M + n\text{alkali}]^{n-}$
Ionic	C^+, $[C_n + A_{n-1}]^+$, rarely $[CA]^{+\bullet}$	A^-, $[C_{n-1} + A_n]^-$, rarely $[CA]^{-\bullet}$

from dilute sample solution by a microliter syringe. After evaporation of the solvent, the emitter is inserted into the ion source. Therefore, a FD ion source is basically identical to an FI ion source (Figure 29.19). In FD, the analyte can be field ionized (the mechanism of ion creation) while still adsorbed to the surface or while just being on the leave. This leads to a significant reduction in thermal decomposition of the analyte and thereby allows for the analysis of highly polar molecules up to about m/z 2000 (Beckey, 1977; Wood, 1982; Prókai, 1990a; Prókai, 1990b). To achieve sufficient surface mobility, the emitter is resistively heated by ramping a current at 1–16 mA min^{-1} up to 40–60 mA during the measurement. High *emitter heating currents* (EHC) are employed to remove residual sample. Analytes of low polarity are even successfully analyzed up to higher molecular weights, for example, polystyrenes are accessible by FD up to m/z 10.000 (Lattimer, 2001).

The type of ions produced by FD depends on the analyte's polarity. Compounds of low polarity yield $M^{+\bullet}$ ions, medium polarity molecules are transformed into either $M^{+\bullet}$ ions or quasimolecular ions such as $[M+H]^+$ or $[M+\text{alkali}]^+$, and finally, highly polar compounds only form quasimolecular ions. Ionic analytes (cation C^+ and anion A^-) are desorbed from the surface without prior FI. C^+ is then detected as the most intensive peak, followed by $[C_nA_{n-1}]^+$ cluster ions. Often, nonionic analytes also form cluster ions, for example, $[2M+H]^+$, $[2M+\text{alkali}]^+$, to some degree (Figure 29.21).

Negative-ion FD is a rare exception (Daehling *et al.*, 1982; Dähling *et al.*, 1983), because emission of electrons,

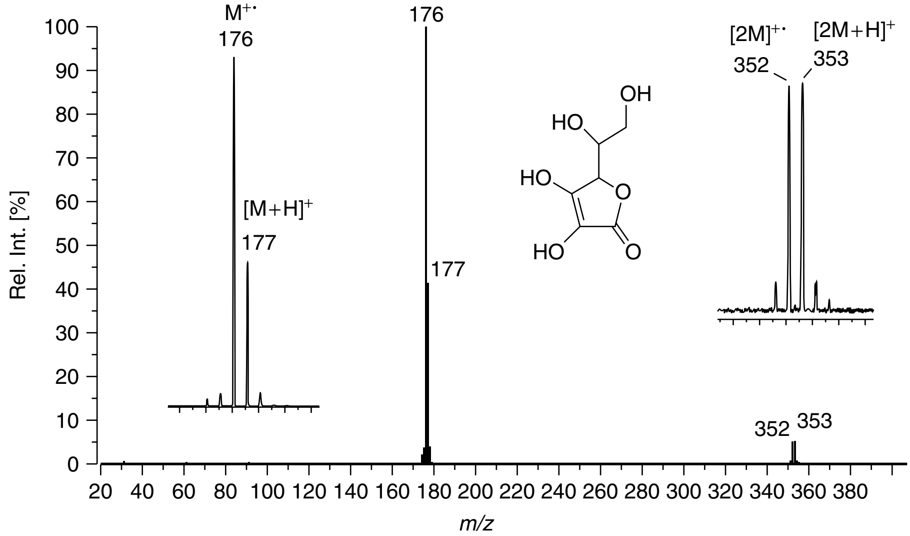

Fig. 29.21 FD mass spectrum of ascorbic acid showing $M^{+\bullet}$ and concomitant $[M+H]^+$ ions. This can be inferred from the overgrown isotopic pattern that cannot be due to $C_6H_8O_6$ alone, because then $[M+1]$ should only exhibit 6.6% relative to $M^{+\bullet}$ instead of the observed 41%. The cluster ion signal shows this phenomenon even more obviously.

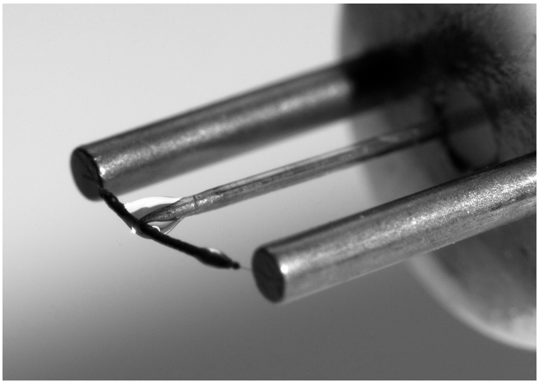

Fig. 29.22 The basic, nonetheless crucial, difference between LIFDI and standard FI or FD is the addition of the capillary. In LIFDI, the field emitter is loaded with analyte solution via this capillary positioned close to the activated area. The liquid spreads over the entire length and the solvent evaporates in vacuum. The thin central tungsten wire connecting this section to the posts is barely visible.

detrimental for the emitter, occurs from activated emitters before analyte ion formation.

The advent of *liquid injection field desorption/ionization* (LIFDI) gave new impetus for the use of FD. In LIFDI, the analyte solution (0.1–0.2 mg ml^{-1}) is transferred onto the emitter inside the ion source by a fused silica capillary (10–75 µm i.d.) (Figure 29.22). In this manner, LIFDI (i) offers inert conditions (Linden, 2004; Gross *et al.*, 2006), (ii) speeds up the measurement process by reducing the number of tuning cycles, and (iii) even admits continuous-flow applications (Schaub *et al.*, 2004). The combination of LIFDI with FT-ICR instruments is gaining importance in industrial analysis, for example, of nonpolar compounds in petroleum products (Figure 29.23) (Schaub *et al.*, 2005).

29.7.2
Fast Atom Bombardment and Liquid Secondary Ion Mass Spectrometry

Energetic particles impacting onto a surface are capable of both sputtering ions as well as neutrals into the selvedge region and ionization of those gas phase neutrals (Figure 29.24). The development of the title methods started with the introduction of ion beams for the production of secondary ions from inorganic surfaces and hence the name *secondary-ion mass spectrometry* (SIMS)(Benninghoven and Kirchner, 1963). Nonconduction materials build up electric charges that prevent further useful ion production by Coulombic repulsion. Thus, primary ion beams were replaced by atom beams (*fast atom bombardment*, FAB) (Devienne, 1967; Barber, Vickerman and Wolstenholme, 1980) that can be obtained by passing rare gas ions through a collision cell where they are neutralized while their kinetic energy and direction are basically retained (Figure 29.25).

The introduction of a *liquid matrix* to dissolve the analyte made FAB a method for the soft ionization of highly polar and eventually large organic or bioorganic molecules (Morris *et al.*, 1981; Barber *et al.*, 1982). In addition, the matrix allowed to employ primary ion beams again, because it circumvented electrostatic charging of the target, thereby giving rise to *liquid secondary-ion mass spectrometry* (LSIMS) (Benninghoven, 1983). FAB and LSIMS work equally well for the

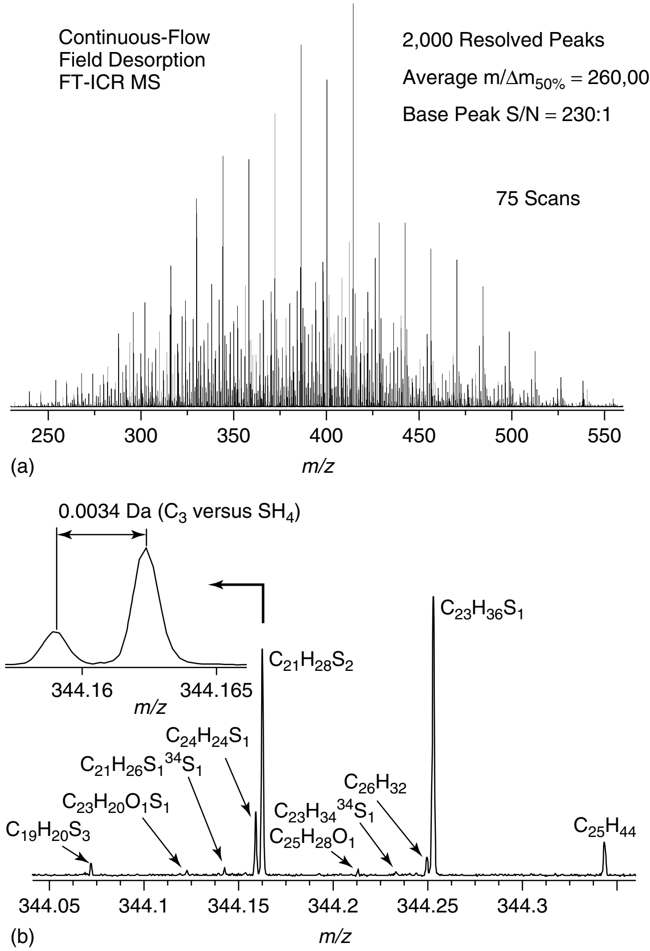

Fig. 29.23 (a) Continuous-flow FD-FT-ICR mass spectrum of a refinery process stream sample; the spectrum is the average of 75 single spectra accumulated during 1 hour. (b) Mass scale expansion at nominal m/z 344 from the broadband mass spectrum. The resolution of the 3.4×10^{-3} u distant C_3 versus SH_4 duplet can be observed at m/z 344.25 and separately in the inset at m/z 344.16. Adapted from Schaub *et al.* (2004) with permission. © John Wiley & Sons, 2004.

whole range of sample polarities and for negative-ion and positive-ion generation (Chapter 9 in Gross (2004) and Table 29.3).

The use of a liquid matrix, for example, glycerol, thioglycerol, 3-nitrobenzylalcohol, 2-nitrophenyloctylether, is key to the abilities of FAB and LSIMS. The matrix absorbs the primary energy, avoids intermolecular interactions of analyte molecules by dissolution, serves as a source of protons for $[M + H]^+$ ion formation by its natural acidity, and mediates the ionization in solution by other

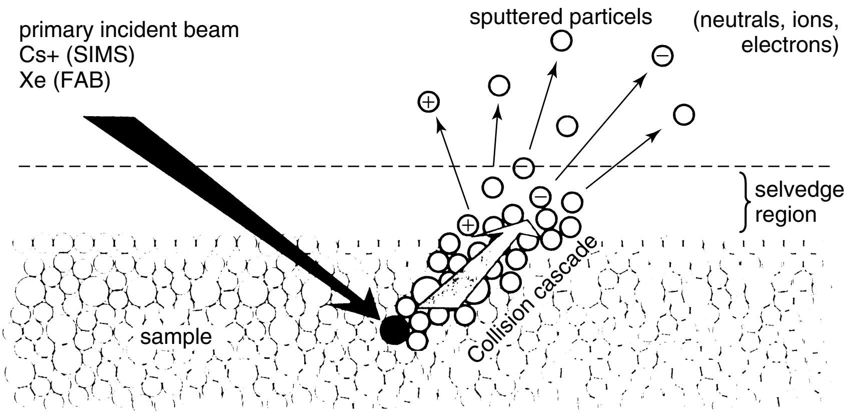
primary incident beam
Cs+ (SIMS)
Xe (FAB)
sputtered particels
(neutrals, ions, electrons)
selvedge region
Collision cascade
sample

Fig. 29.24 Simplified illustration of an instantaneous collision cascade generated as a result of primary particle impact in FAB or SIMS. Adapted from Busch (1995) with permission. © John Wiley & Sons Ltd, 1995.

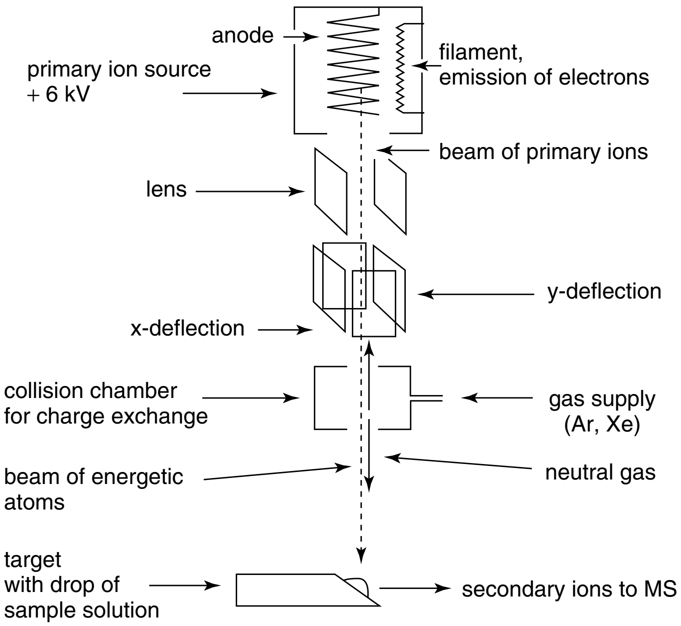
anode
primary ion source
+ 6 kV
filament,
emission of electrons
beam of primary ions
lens
y-deflection
x-deflection
collision chamber
for charge exchange
gas supply
(Ar, Xe)
beam of energetic
atoms
neutral gas
target
with drop of
sample solution
secondary ions to MS

Fig. 29.25 Schematic representation of a FAB gun. The supply of neutral gas also serves as a collision chamber for charge exchange to transform the energetic primary ions generated by EI into a beam of fast neutrals hitting the sample solution on the target. In a FAB ion source, the target takes the same position as the FI/FD probe in Figure 29.19 or the reagent gas inlet in Figure 29.17, respectively; further changes as compared the previously shown ion sources are not required. Therefore, EI/CI/FAB combination ion sources are common. Ion volumes exchangeable through a vacuum lock can reduce ion source contamination.

ionic species, either present as traces such as the ubiquitous alkali ions or by deliberately added ones such as organic acids, salts, and so on (De Pauw, Agnello and Derwa, 1986). Unfortunately, evaporation of the matrix accelerates ion source contamination and limits the time span for sampling of spectra. In addition, abundant ions are produced from the matrix, delivering a strong background of peaks at every m/z with $[n\text{Ma} + \text{H}]^+$ or $[n\text{Ma} - \text{H}]^-$ matrix cluster ions typically dominating those spectra. Therefore to be identified, the analyte ions need to yield strong signals clearly distinguished from that background. Continuous-flow FAB (CF-FAB) (Caprioli, Fan and Cottrell, 1986) and frit-FAB (Ito *et al.*, 1985) not only improve the signal-to-background ratio but also present a means for LC coupling and real-time monitoring of physiological processes by FAB (Caprioli, 1990).

Although still in use, FAB and LSIMS have widely been replaced by more recent desorption/ionization and related methods (see below). One of the advantages of FAB and LSIMS is their ability to actively generate ions from neutrals in unpolar and aprotic matrix solutions, at cryogenic temperatures even from typical organic chemistry solvents (LT-FAB) (Wang *et al.*, 1998; Gross, 1998).

29.7.3
Laser Desorption/Ionization

Different from FAB and LSIMS, *laser desorption/ionization* (LDI) does not provide a continuous beam of ions. Owing to the pulsed nature of a laser, ions are generated in extremely short bursts, the duration of which being slightly longer than the pulse width of the laser. Typically, nitrogen lasers, that is, UV lasers at a wavelength of 337 nm delivering 3 nanosecond pulses are employed. Laser and optics are normally located on top of the ion source housing from where the light is guided onto a sample plate (target) through a quartz window (Figure 29.26). The energy provided by the laser is then absorbed by a (normally solid) layer of the analyte where it causes evaporation

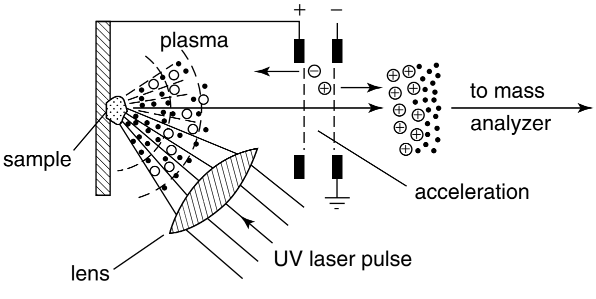

Fig. 29.26 Schematic of an ion source for laser desorption/ionization. The solid sample layer is irradiated with a laser pulse in vacuum. Ions created during this pulse and in the immediate aftermath are extracted from the laser-generated plume and accelerated into a (time-of-flight) mass analyzer. Reproduced from Mamyrin (1994) with permission. © Elsevier Science Publishers, 1994.

and ionization in analogy to the effects of impacting particles as described in Section 29.7.2 (Fenner and Daly, 1966; Vastola and Pirone, 1968; Vastola, Mumma and Pirone, 1970; Posthumus *et al.*, 1978). In addition, nonresonant MUPI in the gas phase can occur (cf. Section 29.4.1.2).

The direct transfer of energy to the analyte, however, creates a significant portion of ions excited well above their dissociation threshold. Therefore, LDI is not as soft as FAB or especially FD. Nonetheless, the fact that LDI is an "energy-sudden" method can compensate for those levels of internal energy, because dissociation kinetics is slower by some orders of magnitude than ion formation (cf. time scale of EI, Figure 29.16).

Stable, UV-absorbing molecules such as fullerenes, polycyclic aromatic hydrocarbons (PAHs), or low-mass organic or inorganic salts are easily analyzed by LDI. In fact, LDI-MS has played an important role in the discovery of the fullerenes (Figure 29.27) (Kroto, 1997).

Although still rather soft and well applicable to analytes up to about *m*/*z* 2000, LDI has its drawbacks in (i) the necessity of an UV chromophore in the analyte molecules and (ii) the conservation of intermolecular forces within the analyte layer that together pose limits upon the application of LDI to highly polar compounds such as oligosaccharides, peptides, and proteins (Posthumus *et al.*, 1978; McCrery, Ledford and Gross, 1982; Cotter, 1987).

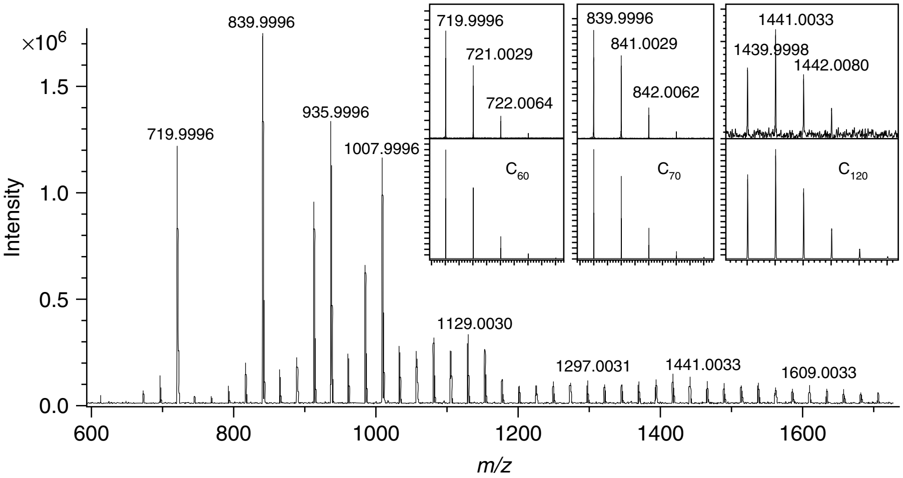

Fig. 29.27 Positive-ion LDI mass spectrum of a fullerene soot synthesized by the Krätschmer–Huffman method (Krätschmer *et al.*, 1990). The spectrum was obtained on a Fourier-transform ion cyclotron resonance (FT-ICR) mass spectrometer; hence, the resolution is 175.000 (1.75×10^5) at *m*/*z* 840, that is, almost by a factor of 1000 higher than on early time-of-flight instruments as used for the discovery of C_{60} and larger fullerenes (Kroto, 1997) The insets show expanded views of the $M^{+\bullet}$ ions for C_{60}, C_{70}, and C_{120} (upper traces) together with the corresponding calculated isotopic patterns. Note that the accurate masses are lower than the nominal values by the mass of an electron, while the difference in mass of 1.0033 u between ^{12}C and ^{13}C can be recognized from the mass increment of the first and second isotopic peaks.

29.7.4
Matrix-Assisted Laser Desorption/Ionization

Tanaka's approach to establish a matrix in LDI was the use of glycerol with an admixture of ultrafine cobalt powder (particle size about 30 nm) serving to transform the UV-transparent liquid into a UV-absorbing medium (Tanaka *et al.*, 1988; Tanaka, 2003) However, the restricted variability of this matrix limited its use to proteins and polar synthetic polymers, and moreover, the poisonous cobalt prevented its adoption as a routine technique.

The other approach, introduced by Hillenkamp and Karas, involved a crystalline organic matrix and started a revolution in terms of what kind of analytes became accessible by MS. Thus, *matrix-assisted laser desorption/ionization* (*MALDI*) with its capability to analyze intact proteins as large as antibodies (150.000 u) meant that MS had really entered the realms of biology and medicine (Figure 29.28)

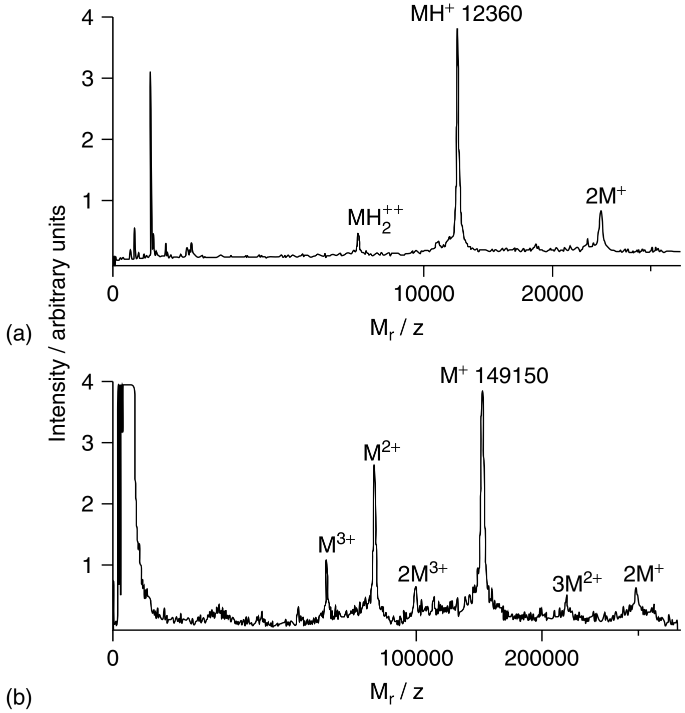

Fig. 29.28 Two UV-MALDI spectra of the early days of MALDI immediately demonstrating the potential of the method. (a) The small protein porcine cytochrome C from 2,5-dihydroxybenzoic acid matrix after accumulation of 20 laser shots and (b) a large monoclonal antibody from nicotinic acid matrix after accumulation of 30 laser shots. Although assigned as $M^{+\bullet}$ or $[M + H]^+$ ions, it may be assumed that the signals are formed by the whole palette of quasimolecular ions possible. Reproduced from Karas, Bahr and Gießmann (1991) with permission. © John Wiley & Sons Inc., 1992.

(Karas *et al.*, 1989a; Karas *et al.*, 1989b; Karas, Bahr and Gießmann, 1991) A majority of portions of the research done in these fields today rely on MALDI as a (high-throughput) method for fingerprinting (Welham *et al.*, 2000), identifing, and sequencing of various classes of biopolymers (Hillenkamp and Peter-Katalinic, 2007).

The tremendous development of TOF instruments was initiated and put forward to a large extent by the demand for increasingly powerful MALDI instrumentation. Indeed, MALDI-TOF was the almost indivisible implementation of MALDI in its first decade.

The instrumental setup for MALDI does not differ from LDI; the matrix is the key to the method's success. The matrix provides effective absorption of the laser energy independent of whether the analyte possesses a chromophore or not. It separates the molecules before evaporation and supports the ionization process, either by acid/base reactions upon mixing with the analyte or during the actual desorption via ion–molecule reactions in the laser-generated plasma (e.g., cf. Sections 29.4.1.4, 29.4.2.1, 29.4.2.2) (Zenobi and Knochenmuss, 1999; Karas and Krüger, 2003).

Nitrogen UV lasers ($\lambda = 337$ nm) are most frequently employed to supply power in excess of the threshold irradiance for ion production ($>10^6$ W cm^{-2}). Nonetheless, excimer lasers ($\lambda = 293$, 248, 308, and 351 nm), frequency-tripled ($\lambda = 355$ nm), and quadrupled Nd : Yag lasers ($\lambda = 266$ nm) are also in use. In IR-MALDI, used for a minority of applications where penetration depth is important, for example, to desorb samples from gels or thin layer chromatographic (TLC) plates, Er : Yag lasers ($\lambda = 2.94$ µm) and CO_2 lasers ($\lambda = 10.6$ µm) are common (Overberg *et al.*, 1990; Overberg, Karas and Hillenkamp, 1991).

Although many standard applications of MALDI do not present very special requirements, MALDI sample preparation can be an art in itself. For certain types of analytes such as oligonucleotides or high-mass (bio)polymers, great care needs to be excised to produce suitable crystal layers where analyte and matrix are preferentially cocrystallized from solution in molar ratios in the 1 : 1000–100.000 (10^5) range. The same applies when ultimate sensitivity is required. Under optimal conditions, the sample consumption is in the low femtomole range, and several attomoles have been demonstrated (Vorm, Roepstorff and Mann, 1994). Special hydrophobic targets with hydrophilic spots effecting the concentration of the preparation on a small diameter spot (200–600 µm) deliver substantially improved sensitivity (Nordhoff *et al.*, 2003).

By selecting a matrix to the analyte's requirements, MALDI can be applied to a large variety of analytes ranging from nonpolar polymers such as (low MW) polyethylene to highly polar polymers such as poly(toluenesulfonic acid) resins. It is suitable for peptides, proteins, oligonucleotides, and DNA, for oligosaccharides (Figure 29.29), and – sometimes overlooked – for small molecule analysis, for example, for dyes, lipids, and porphyrins.

MALDI-MS presents a powerful method for the analysis of proteins and peptides. It may be used for the in-depth analysis of an individual protein to reveal its primary structure, that is, its amino acid sequence, as well as eventual (posttranslational) modifications (Figure 29.30). Proteomics, the

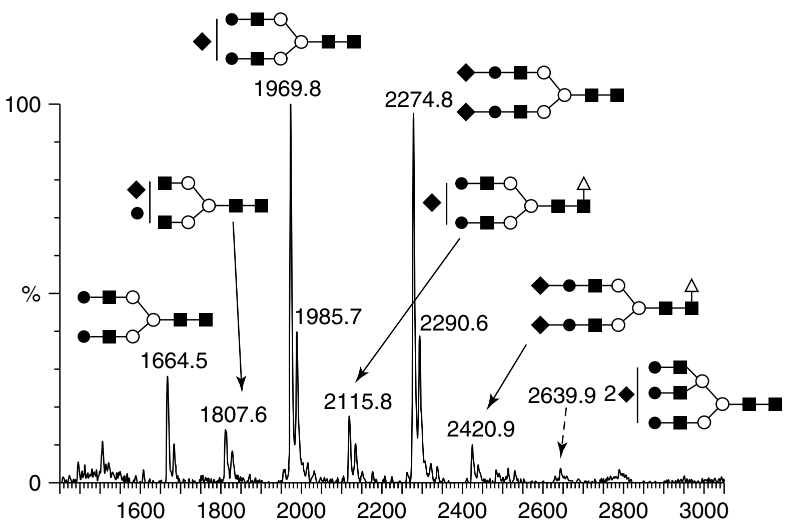

Fig. 29.29 Positive-ion linear mode MALDI-TOF spectrum from 2,4,6-trihydroxyacetophenone (THAP) matrix of native *N*-glycans obtained by in-gel PNGase F digestion from the two-dimensional gel of α 1-antitrypsin spots of a CDG-type-IIx patient. The sialic acid residues were methylated before MALDI analysis. Symbols: ■, GlcNAc; ●, galactose; ○, mannose; ♦, sialic acid; and △, fucose. Adapted with permission from Šagi *et al.* (2005) © Wiley-VCH, 2005.

study of the whole complexity of a protein mixture expressed by given type of cells, is another widespread application of MALDI, which indeed serves as a propellant for the steady improvements in sensitivity, resolution, mass accuracy, and speed of modern MALDI instrumentation. It is a common procedure to obtain survey MALDI spectra of protein tryptic digests and to subsequently measure MALDI tandem mass spectra of the tryptic peptides produced. The sequence of the whole protein may then be derived by putting the subsequences of those peptides together (Lehmann, 1996; Kinter and Sherman, 2000; Snyder, 2000).

In polymer chemistry, MALDI spectra provide identification of the monomer (repeat unit), of the average molecular weight, and of polydispersity, and – provided the low-mass end of the envelope is not beyond m/z 1000 – allow for end group analysis (Figure 29.31) (Nielen, 1999; Murgasova and Hercules, 2003; Montaudo and Lattimer, 2001; Pasch and Schrepp, 2003).

29.8
Ion Formation at Atmospheric Pressure

It is a common feature of the ionization methods discussed so far that the process of ion formation occurs under vacuum conditions, varying from the high vacuum in EI or FI to the low-pressure limit of medium vacuum in the case of CI. Although passing through a steep pressure gradient across the selvedge region, even the deliberation of ions in desorption/ionization methods proceeds from a surface located in vacuum. It is inherent to these methods that they cannot deal with a noteworthy liquid flow, a fact

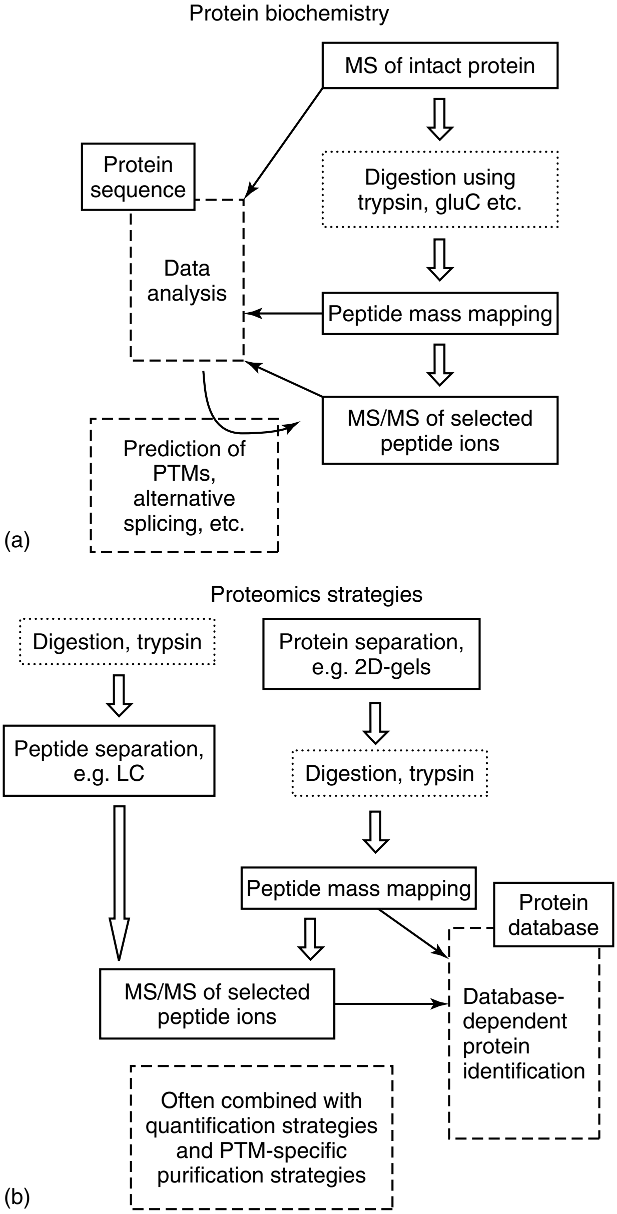
Protein biochemistry
MS of intact protein
Protein sequence
Digestion using trypsin, gluC etc.
Data analysis
Peptide mass mapping
MS/MS of selected peptide ions
Prediction of PTMs, alternative splicing, etc.
(a)
Proteomics strategies
Digestion, trypsin
Protein separation, e.g. 2D-gels
Peptide separation, e.g. LC
Digestion, trypsin
Peptide mass mapping
Protein database
MS/MS of selected peptide ions
Database-dependent protein identification
Often combined with quantification strategies and PTM-specific purification strategies
(b)

Fig. 29.30 Methodologies for protein analysis involving MALDI-MS. (a) Analysis of a purified protein to validate its identity and to detect eventual modifications. (b) Proteomics approach to detect and identify all proteins expressed by a given cell type. Here, two strategies apply, one starting with digestion of the protein before LC-MS/MS and the other employing separation on a two-dimensional gel, followed by digestion and mass analysis (p. 85 in (Hillenkamp and Peter-Katalinic, 2007)).

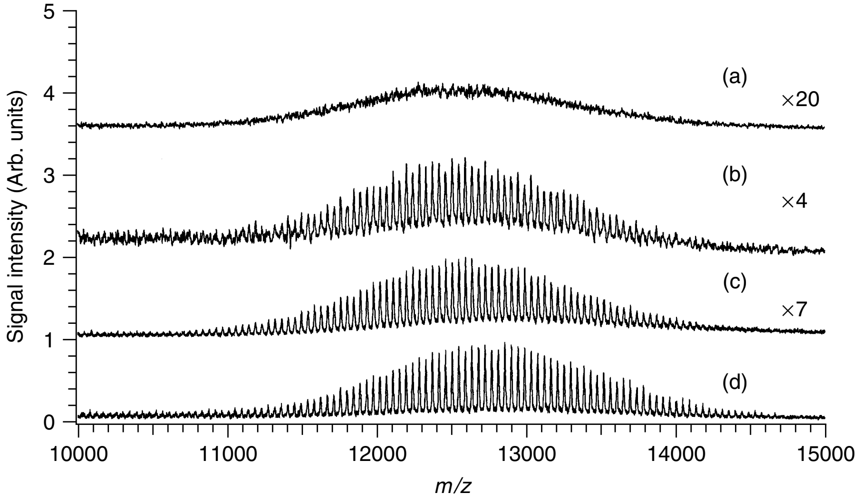

Fig. 29.31 MALDI-TOF spectra of poly(ethyleneglycol) as obtained with different matrices and in different modes of TOF operation. (a) and (d) are measured using 2-(4-hydroxyphenylazo)benzoic acid (HABA) matrix, but (a) without and (b)–(d) with time-lag focusing (TLF); (b) with 2,5-dihydroxybenzoic acid (DHB), and (c) with *trans*-3-indoleacrylic acid (IAA) matrix. These spectra (i) exemplify the enormous improvement in resolving power of TOF instruments upon introduction of TLF (delayed-extraction, DE, and pulsed ion extraction, PIE, have the same effect) and (ii) the importance of the choice of a suitable matrix. Adapted with permission from Whittal, Schriemer and Li (1997). © American Chemical Society, 1997.

preventing – or at least strongly limiting (cf. CF-FAB, Section 29.7.2) – their successful combination with LC. Nonetheless, there was a long-standing desire to develop *liquid-chromatography-mass spectrometry coupling* (LC-MS) in order to address highly polar analyte mixtures.

Atmospheric pressure ionization (API) methods take the very step of ion formation virtually to the outside of the mass spectrometer, or they expand the mass spectrometer into the laboratory atmosphere.

29.8.1
Atmospheric Pressure Ionization

The first implementation of a continuous liquid flow coupled to a mass analyzer was presented by Horning *et al.* as early as 1973 and was termed API (Horning *et al.*, 1973; Horning *et al.*, 1974). In API, the sample solution is evaporated at atmospheric pressure by a heated stream of nitrogen gas. The vapor is then passed along a ^{63}Ni foil, the electrons emitted from which start a series of ion–molecule reactions, leading to $[(H_2O)_n{+}H]^+$ cluster ions that finally cause $[M{+}H]^+$ ions to be formed. Although most of the gas is removed by a rotary pump, a portion is admitted to a quadrupole mass analyzer through a small orifice (Figure 29.32). Although API itself is not anymore in use, the method has coined the name for a whole family of ionization methods.

To improve ionization efficiency, the same group soon introduced *atmospheric*

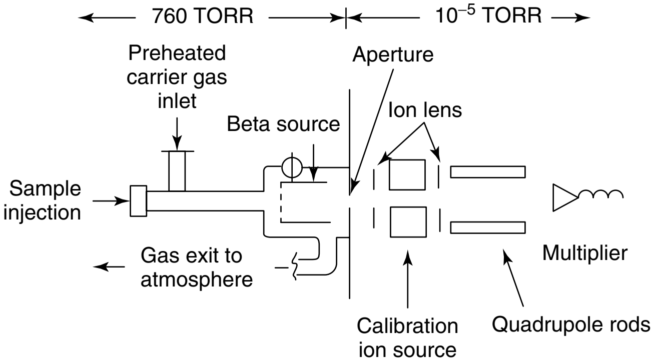

Fig. 29.32 Atmospheric pressure ionization (API) as introduced in 1973. The β-source initiates ion formation via a series of ion–molecule reactions involving water cluster ions. A small portion of the ions is transmitted into a quadrupole mass analyzer through an orifice. The quadrupole instrument shown in this first publication on API even had an EI ion source that could be used for calibration and testing of the instrument. Reproduced from Horning *et al.* (1973) with permission. © American Chemical Society, 1973.

pressure chemical ionization (APCI) where the formation of ions is achieved without the aid of radioactive ^{63}Ni. Here, electrons are delivered by a corona discharge from a needle tip pointing into the sample vapor or by admitting electrons produced from a heated filament as in CI (Carroll *et al.*, 1975; Dzidic *et al.*, 1976). The fact that the ion–molecule reactions take place at atmospheric pressure led to the term. Highly improved setups of APCI still present common LC-MS equipment (cf. Section 29.8.3) (Niessen and Voyksner, 1998; Niessen, 1998).

Another important approach to LC-MS was the *thermospray* (TSP) interface (Blakley, Carmody and Vestal, 1980; Blakley and Vestal, 1983). In TSP, the analyte solution flowing at 1–2 ml min^{-1} is rapidly vaporized by passing it through strong heater. Positive ions are formed by protonation or electrophilic addition of cations, negative ions by deprotonation or nucleophilic addition, for example, by 0.1 M ammonium acetate added to the solution before evaporation. The ions are extracted orthogonally from the vapor phase and enter the mass analyzer through an aperture in the ion exit plate. The ionization efficiency of TSP interfaces can also be increased by the addition of a source for electrons, that is, by changing its operation basically to APCI mode.

In addition to these methods, some others were rather short-lived. The *particle beam interface* (PBI) employed conventional EI or CI ion sources to effect ionization of micrometer-sized solid particles produced upon room-temperature evaporation of low-polarity solvents, followed by removal of the vapor by a jet separator (Willoughby and Browner, 1984; Brauers and von Bünau, 1990). The *moving-belt interface* transported the solid residues of solvent evaporation inside an oven into an EI or CI ion source where they were evaporated by heating the metal belt in analogy to the sample vial in standard EI or CI applications

(McFadden, Schwartz and Evans, 1976; Karger *et al.*, 1979; McFadden, 1979).

29.8.2
Electrospray Ionization

Electrolytic solutions can be dissipated to form a mist of highly charged micrometer-sized droplets by the action of a strong electric field. This process, known as *electrospray*, is based on Zeleny's work on the behavior of surfaces under the influence of electric fields (Zeleny, 1917) and Taylor's observation that spheric droplets then take the shape of a cone (*Taylor cone*) from which a fine liquid jet is emitted (Taylor, 1964) when Coulombic repulsion exceeds the surface tension, that is, when the *Rayleigh limit* is passed (Lord Rayleigh, 1882). Experiments by Dole demonstrated the production of gaseous ions from polymer solutions by electrospray (Dole *et al.*, 1968a; Dole *et al.*, 1968b), and finally Fenn's group developed an interface to a quadrupole mass analyzer. Their *electrospray ionization* (ESI) device was able to deliver multiply charged ions of large biomolecules, a discovery meaning the breakthrough for ESI (Whitehouse *et al.*, 1985; Fenn *et al.*, 1989; Fenn, 2003). Nowadays, both ESI and MALDI are essential tools of biomedical research (Figure 29.33) (Dole, 1997; Pramanik, Ganguly and Gross, 2002).

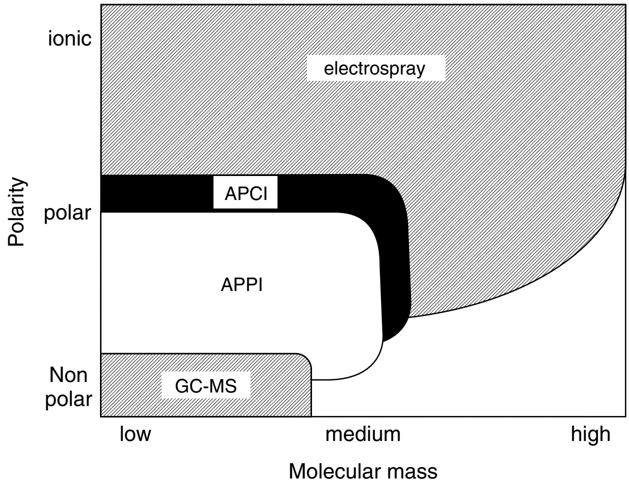

Fig. 29.33 The importance of ESI among chromatography-mass spectrometry applications can be inferred from its wide range of analyte acceptance in terms of both polarity and mass. ESI may deal with medium polar to ionic analytes ranging from as low as 10 u to as high as 10^6u in mass. APCI and APPI are very similar to each other in coverage, both accessing only the lower left segment of the polarity-mass plane in this illustration. Adapted with permission of U. Karst from cover of *Anal. Bioanal. Chem.* **378(4)** (2004).

When an electrolytic solution (typically 10^{-6}–10^{-4} M at 1–5 μl min^{-1}) reaches the end of a capillary (10–100 μm i.d.) that is connected to high voltage

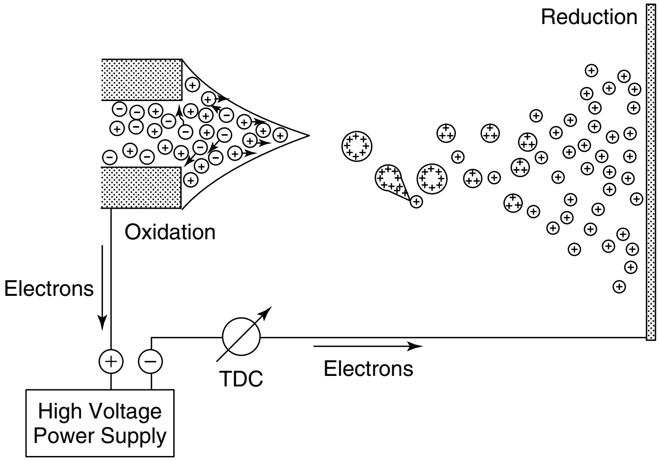

Fig. 29.34 The electrospray process is induced by charge separation at the tip of the spray capillary forcing the emerging droplet to form a cone. The increased electric field strength at cone's apex causes emission of a fine jet that breaks up to form micrometer-sized highly charged droplets that shrink upon evaporation of solvent. The overall process presents an electrolytic cell. To adapt ESI to a mass spectrometer, the counter electrode has a small hole that admits a portion of the aerosol. The sample flow may be supplied by either a liquid chromatograph or a syringe pump. Adapted from Kebarle and Tang (1993) with permission.

(3–5 kV), the electrospray process is induced. First, charge separation at the tip of the spray capillary forces the emerging droplet to form a cone. The increased electric field strength at the apex of the cone then causes emission of a fine jet that soon breaks up to yield micrometer-sized highly charged droplets shrinking upon evaporation of solvent. The overall process represents an electrolytic cell (Figure 29.34) (Kebarle and Tang, 1993). ESI can be adapted to a mass spectrometer when the counter electrode has a small hole to admit a portion of the aerosol via a differentially pumped interface. The principle of a differentially pumped interface has already been outlined (Section 29.5.4, Figure 29.10). However, opposed to the plasma conditions in ICP, the ESI sprayer is operated at room temperature, in fact, in the laboratory atmosphere. To compensate for the heat of evaporation, that is, to prevent droplets from freezing during desolvation while traveling through the interface, some external heating is required. Heating may be realized in either way, using a slow counter current of hot nitrogen gas (1–2 l min^{-1} at 150–250 °C), or transfer of the aerosol through a heated capillary (ca. 0.5 mm i.d., 10 cm length, 120–220 °C), before the first pumping stage.

The shrinking of the droplets causes ejection of much smaller droplets when the critical charge density is reached. Those off-spring droplets carry only a minor portion of the mass, but a significant portion of the charges of their precursor. After repeated *droplet jet*

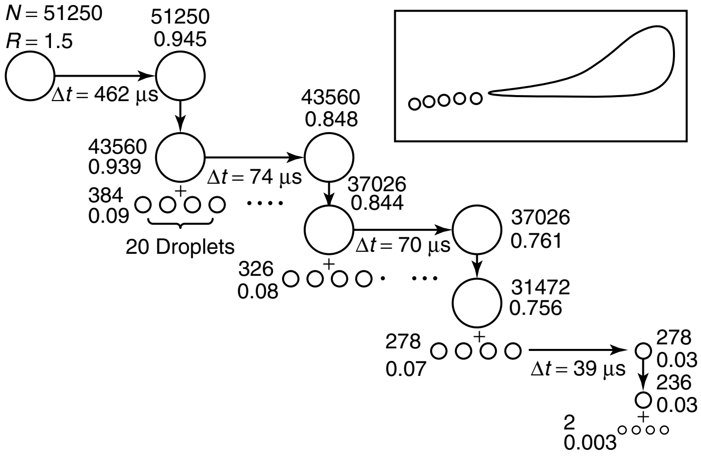

Fig. 29.35 Illustration of droplet jet fission that leads to a stepwise reduction in droplet size. The inset shows the fission process whereby a series of smaller, nonetheless highly charged, off-spring droplets is released from a precursor droplet. Adapted from Kebarle and Tang (1993) with permission.

fission, it comes to the step of production of isolated gas phase ions (Figure 29.35) (Kebarle and Tang, 1993; Cole, 2000; Kebarle, 2000). The ultimate step of ESI, the deliberation of unsolvated gas phase ions, may proceed on one of the following ways: (i) macromolecules can undergo multiple protonation or cationization in solution up to a level where Coulombic repulsion between the charges poses a limit. Those ions may then remain while all solvent molecules evaporate (*charged residue model*). (ii) For small ions, *ion evaporation* may be taken into account, that is, the shrinking droplet may reduce the charge density by emission of ions from its surface.

In recent interface designs, the sprayer is not in line of sight with the analyzer, but roughly at right angle to the entrance (orthogonal sprayer), and eventually the skimmer may be at right angle to the beam direction of the first pumping stage (z-spray). All these adjustments aim at improving the ion transmission, while reducing the load of neutrals (residual gas, solvent vapor) and unwanted ions (buffer salts) onto the system, thereby optimizing the robustness of the interface (Fenn *et al.*, 1990; Ikonomou and Kebarle, 1994; Abian, 1999). To adapt ESI to higher flow rates (10–1000 μl min^{-1}), most modern ESI sprayers offer pneumatic assistance, that is, a nebulizer gas flow emerging from a fine gap around the spray capillary.

On the other hand, to reduce sample flow and thus the extent of use of precious samples, nano-ESI may be employed. In nano-ESI, the analyte solution is supplied in a (single use) glass capillary, drawn to form a fine tip where a small aperture (1–5 μm) allows the generation of a particularly fine spray. High voltage is connected via the metal-coated outside of the capillary. By adjusting the nano-ESI sprayer very close to the entrance of the interface, the ratio of

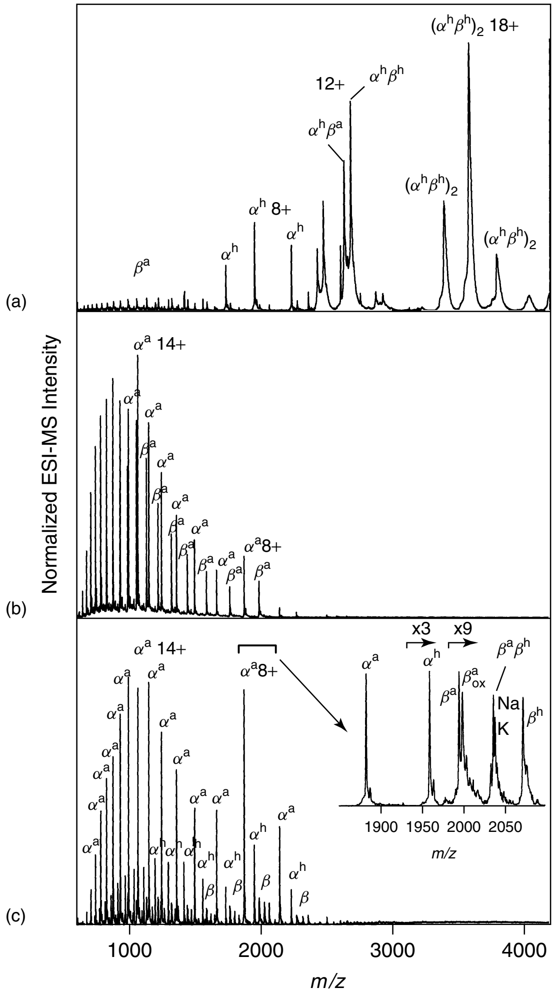

Fig. 29.36 ESI mass spectra of hemoglobin recorded under different solvent conditions. (a) "Native", aqueous solution, pH 8.0; (b) acid unfolded, aqueous solution, pH 2.0; and (c) unfolded protein at pH 10.0 in water/acetonitrile (60:40 v/v). Notation: α^h and α^a represent holo- and apo-α-globin, respectively; β-globin is labeled analogously. Inset in (c) expansion of region around *m/z* 2000, revealing the presence of both α and β subunits in their apo and holo forms. ${\beta^a}_{ox}$ represents an [M + 32] species. The expanded spectrum also shows signals corresponding to $\beta^a\beta^h$ ions carrying Na^+ and K^+ adducts. Reproduced from Boys and Konermann (2007) with permission. © Elsevier Science Publishers, 2007.

admitted to rejected aerosol is significantly improved (Wilm and Mann, 1994; Wilm *et al.*, 1996; Wilm and Mann, 1996). Thus, 1 µl of sample can deliver a 20-min lasting spray at 50 nl min^{-1} while the ion current of the mass spectrometer does not remarkably drop as compared to conventional ESI operation. Beside lower sample consumption, nano-ESI offers operation free of memory from previous samples and access to low volatility solvent systems such as 100 % water or water/dimethylsulfoxide mixtures. On the other side, the metal plated, sometimes precision-cut, capillaries are expensive and fragile and require an experienced operator.

The application of ESI to biopolymers such as proteins, oligosaccharides, or DNA delivers a distribution of charge states, either positive or negative as required by the analyte. Normally, the charge state distribution represented by all peaks belonging to one analyte species forms an even envelope. In the case of proteins for example, it has a maximum at one charge per 1000 u. This way, ESI "folds" high-mass analytes into the *m*/*z* range of most mass spectrometers (Peng and Gygi, 2001; Hanson *et al.*, 2003). The charge state distribution of a protein is not independent of the experimental conditions; on the contrary, it is subject to remarkable changes upon pH and, thus, protein tertiary structure. However, ESI not only provides molecular mass but also allows information on the shape of macromolecules to be derived (Figure 29.36) (Boys and Konermann, 2007). Combined with the extraordinary softness of ion generation, ESI can indeed supply "wings to molecular elephants" (Fenn, 2003).

1. It is possible to calculate charge states for all peaks of the distribution from which, in turn, the mass of the neutral analyte can be deduced (Covey *et al.*, 1988; Mann, Meng and Fenn, 1989).
2. This is due to the fact that the charge state of the ions forming the signal at the adjacent lower *m*/*z* is higher by exactly one.
3. Today, software is available to do automated *charge deconvolution*.

ESI is soft enough to yield isolated gas phase ions of rather labile molecular assemblies such as host–guest complexes, adducts, or solvated species. Therefore, ESI is a well-established analytical tool in supramolecular chemistry (Chapter 5 in Schalley, 2007). In combination with tandem MS, ESI can also deliver information on the stability of such complexes (Figure 29.37).

Advantageous on one side, the inherent softness of ESI yields mass spectra highly reliable for molecular mass and molecular formula determination, but often without fragment ions that could serve for structure elucidation. Therefore, ESI is frequently combined with tandem mass spectrometric methods to compensate for that lack (activation methods cf. Section 29.9). Among these, *collision-induced dissociation* (CID) is the most standardized and widespread technique, for example, CID is the standard tool for peptide sequencing (Figure 29.38) (McLafferty, 2001).

Finally, ESI is well applicable to (ionic) metal complexes. This way, ESI nicely complements the capabilities of FAB and FD, in particular LIFDI. Especially, syringe pump infusion delivers a signal highly constant in time, which provides the basis for advanced experiments, for example, catalyst screening by ion–molecule reactions (Figure 29.39) (Volland *et al.*, 2001; Chen, 2003; Adlhart and Chen, 2003).

29.8.3
Atmospheric Pressure Chemical Ionization and Atmospheric Pressure Photoionization

Modern ion sources for APCI (Keski-Hynnilä *et al.*, 2002) and *atmospheric pressure photoionization* (APPI) (Robb, Covey and Bruins, 2000; Kauppila *et al.*, 2002; Raffaeli and Saba, 2003) are based on ESI interfaces. For APCI, the complete interface from atmospheric pressure to analyzer entrance remains unaltered, whereas the ESI sprayer is exchanged for the heated vaporizer and a corona needle is placed in the space between vaporizer exit and interface entrance. For APPI, the same vaporizer may be used, but the corona needle is replaced by a rare gas discharge UV lamp; for example, a krypton lamp delivers 10-eV photons (Figure 29.40)

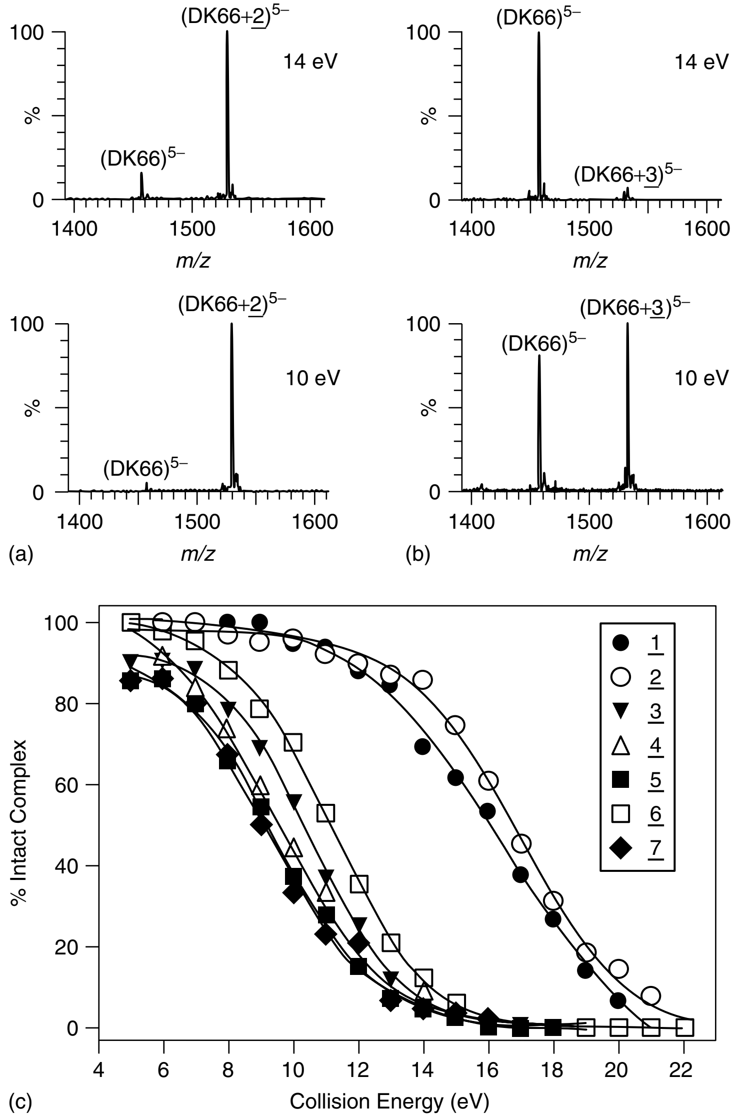

Fig. 29.37 Negative-ion ESI tandem MS experiments on $[1{:}1]^{5-}$ complexes with the d(CGCGAATTCGCG)$_2$ duplex (DK66). Representative tandem mass spectra obtained with different ligands: (a) benzopyridoindole drug (**2**) and (b) benzopyridoquinoxaline drug (**3**), using collision energies of 10 and 14 eV. The increased dissociation upon more energetic collisions is obvious. The argon pressure inside the RF-only hexapole collision cell was 3×10^{-5} mbar. (c) Derived breakdown graphs of different intercalator drugs with DK66 allow to derive the relative binding strengths of those drugs to the duplex; stronger binding causes the curves to shift to the right. Reprinted from Rosu *et al.* (2007) with permission. © Elsevier Science Publishers, 2007.

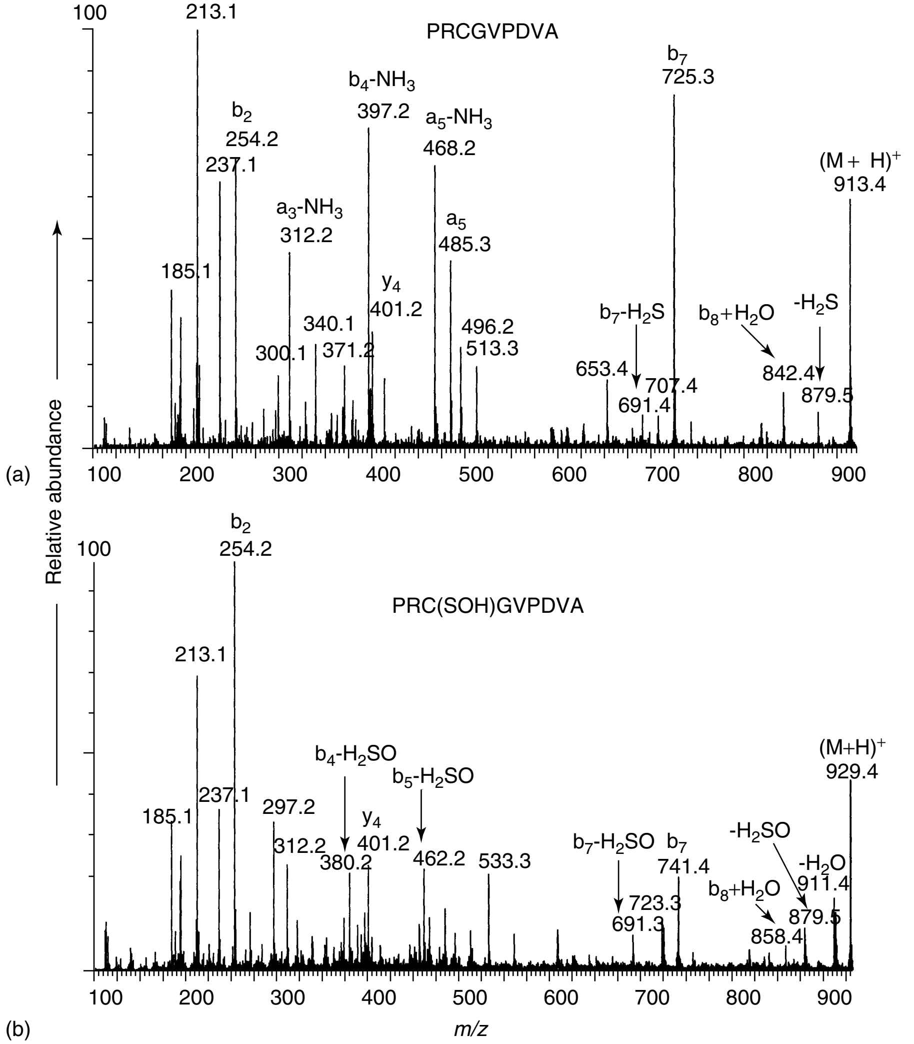

Fig. 29.38 Tandem mass spectra of the ESI-generated $[M + H]^+$ ions of (a) the unmodified cysteine switch peptide PRCGVPDVA, *m/z* 913.4 and (b) the cysteine sulfenic acid product PRC(SOH)GVPDVA, *m/z* 929.4. The b-type fragments containing the sulfenic acid residue are accompanied by peaks due to sulfenic acid loss (−50 u). Reproduced from Shetty, Spellman and Neubert (2007) with permission. © Elsevier Science Publishers, 2007.

(Keski-Hynnilä *et al.*, 2002; Raffaeli and Saba, 2003).

The types of ions are formed depend on the solvent and the analyte. In positive-ion mode, $[M + H]^+$ ions and products of electrophilic addition are preferred with polar combinations, whereas CT will occur to yield $M^{+\bullet}$ ion for unpolar systems (cf. PICI). Although PI of the analyte also plays a role in APPI, it is less

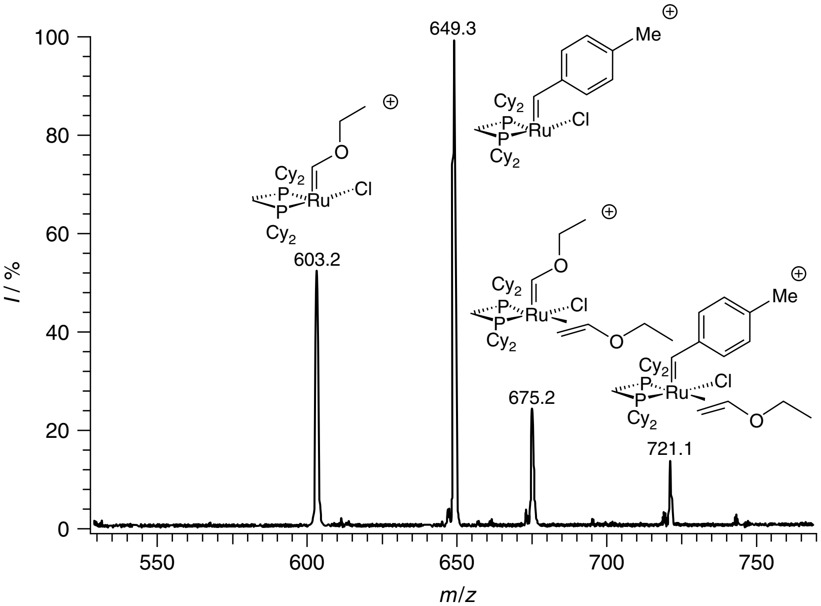

Fig. 29.39 ESI tandem MS experiment for catalyst screening via ion–molecule reactions. The mass-selected precursor ion, [(dcpm-κ^2P)ClRu=CH(p-C_6H_4Me)]$^+$, m/z 649, is reacted with ethyl vinyl ether upon collision in the RF-only octopole collision cell of a triplequadrupole instrument at a collision energy of 1 eV. The product ions incorporating the ether molecule are observed at m/z 603, 675, and 721. Reproduced from Volland *et al.* (2001) with permission. © Wiley-VCH, 2001.

pronounced than one would expect. Unfortunately, there is a tendency in APPI to form radical ions and closed-shell at the same time so that isotopic patterns suffer from superimposition of both species. Analogously to positive-ion mode, operation in negative-ion mode resembles conventional NICI. However, the AP versions of PICI and NICI are slightly softer than their vacuum counterparts, because vaporization from solution reduces "thermal stress" of the analyte and the large number of low-energy collisions at atmospheric pressure serves to keep the ions thermalized.

Both APCI and APPI serve for similar types of analyses. Mostly they are employed as ionization methods in LC-MS applications for low-to-medium polarity analytes, that is, when active ionization of neutrals is required because ion formation in solution is ineffective (Abian, 1999; Tsuchiya, 1995; Reemtsma, 2003). Illustrative examples are the characterization of natural polyamines in spider venom (Figure 29.41) (Chesnov, Bigler and Hesse, 2002), metabolites of pharmaceuticals in human plasma (Yang and Henion, 2002) and PAHs (Hayen and Karst, 2003).

29.8.4
Atmospheric Pressure MALDI

Classical MALDI ion sources are operated under high vacuum conditions, which is a direct consequence of their initial development on TOF instruments.

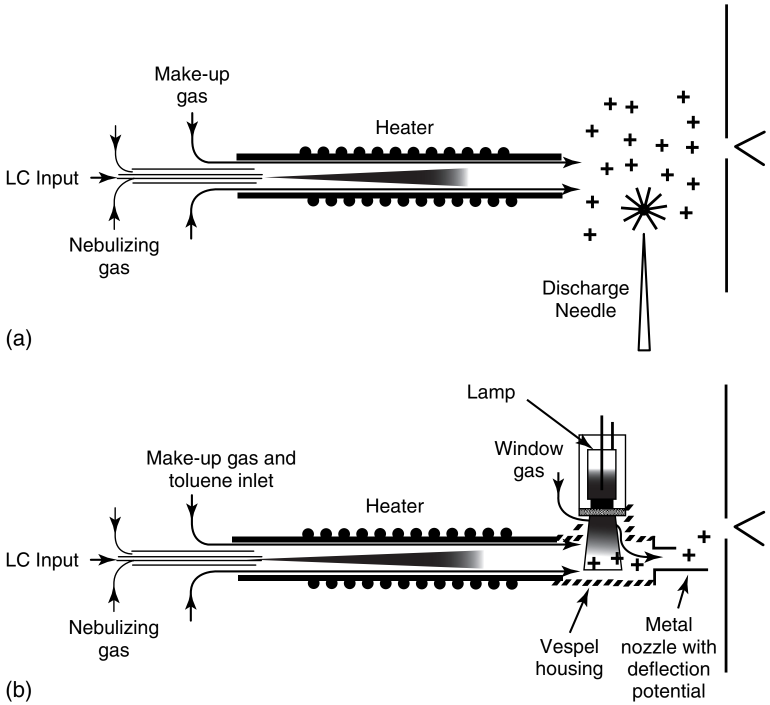
Make-up
gas
Heater
LC Input
Nebulizing
gas
Discharge
Needle
(a)
Lamp
Window
gas
Make-up gas and
toluene inlet
Heater
LC Input
Nebulizing
gas
Vespel
housing
Metal
nozzle with
deflection
potential
(b)

Fig. 29.40 Schematic representations of APCI and APPI interfaces. (a) Conventional heated pneumatic nebulizer interface for APCI. Ionization occurs via the corona discharge needle placed in the vaporized plume of mobile phase and volatilized analytes. (b) Schematic representation of an APPI interface, which is a modification of the probe shown in (a). It incorporates a krypton lamp affixed to the end of the heated nebulizer probe to initiate ionization by photoionization with 10.0-eV photons. Reprinted with permission from Yang and Henion (2002). © Elsevier Science Publishers, 2002.

The desire to have MALDI attached to other analyzers made it necessary to reduce the large energy spread (tens of electron volts) of the laser desorbed ions. This can be accomplished by thermalization with gas, either by having a pulsed valve admitting nitrogen or argon before each laser shot or by moving the event of desorption/ionization into a region of medium vacuum. It turned out that MALDI can even be realized at atmospheric pressure (Laiko, Baldwin and Burlingame, 2000a; Laiko, Moyer and Cotter, 2000b; O'Connor and Costello, 2001; Doroshenko *et al.*, 2002). By positioning a MALDI target in front of the interface, the freshly generated and thermalized ions are guided directly into the instrument (Figure 29.42). Different from APCI and APPI, switching the mode of operation to AP-MALDI is slightly more demanding due to the requirement for accurate alignment of target and laser optics. In addition, sensitivity of AP-MALDI is generally lower than that of vacuum MALDI. The true advantage of atmospheric pressure MALDI (AP-MALDI) is its compatibility with AP

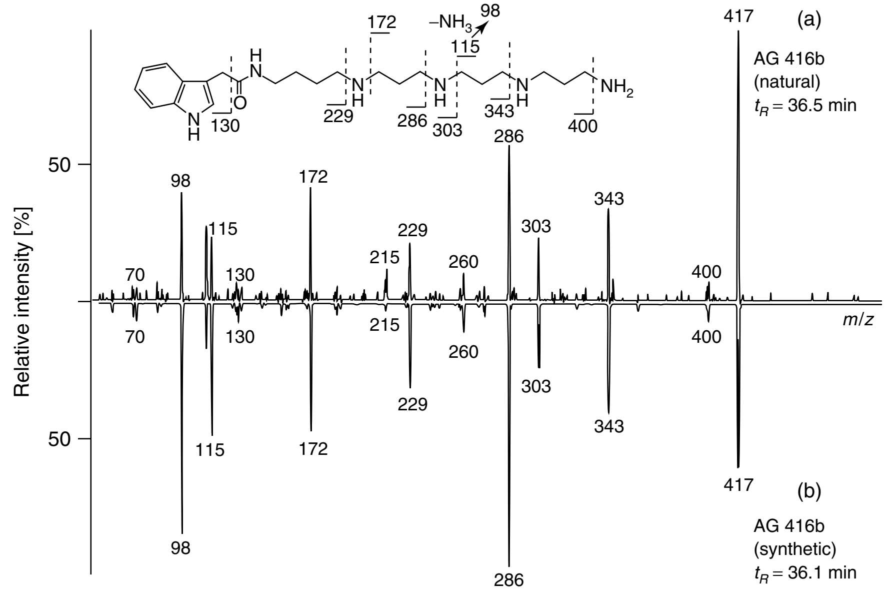

Fig. 29.41 Data from the HPLC-APCI-MS/MS analyses for (a) $[M + H]^+$ ion, m/z 417, of the natural acylpolyamine of a spider and (b) the synthetic compound for reference. Adapted from Chesnov, Bigler and Hesse (2002) with permission. © IM Publications, 2002.

sources, which offers more flexibility for the use of expensive equipment.

29.8.5
Desorption Electrospray

For ESI, the analyte is normally supplied in dilute electrolytic solution. Nonetheless, Takats *et al.* discovered that the processes of electrospraying and sampling can be separated in a way that the electrostatically charged aerosol is blown onto a surface from which analyte molecules are sampled and ionized (Takats *et al.*, 2004; Chen *et al.*, 2005). This method was termed *desorption electrospray/ionization* (DESI). DESI transforms the step of sample preparation into a mere positioning of sample between the sprayer and entrance. DESI obviously plays to its strengths for the quick analysis of surface (adsorbed) compounds such as alkaloids in plants (Figure 29.43), pesticides on fruit, explosives on cloths and similar items, control of bulk compounds, and other applications where the target compound class is defined (Popov *et al.*, 2005; Takats *et al.*, 2005a; Takats, Wiseman and Cooks, 2005; Talaty, Takats and Cooks, 2005).

29.8.6
Direct Analysis in Real Time

Another recent addition to the repertoire for rapid, noncontact analysis at ambient pressure is presented by *direct analysis in real time* (DART) (Cody, Laramee and Durst, 2005) DART differs much from the aforementioned techniques. It is based on the reactions of electronic or vibronic excited-state species with

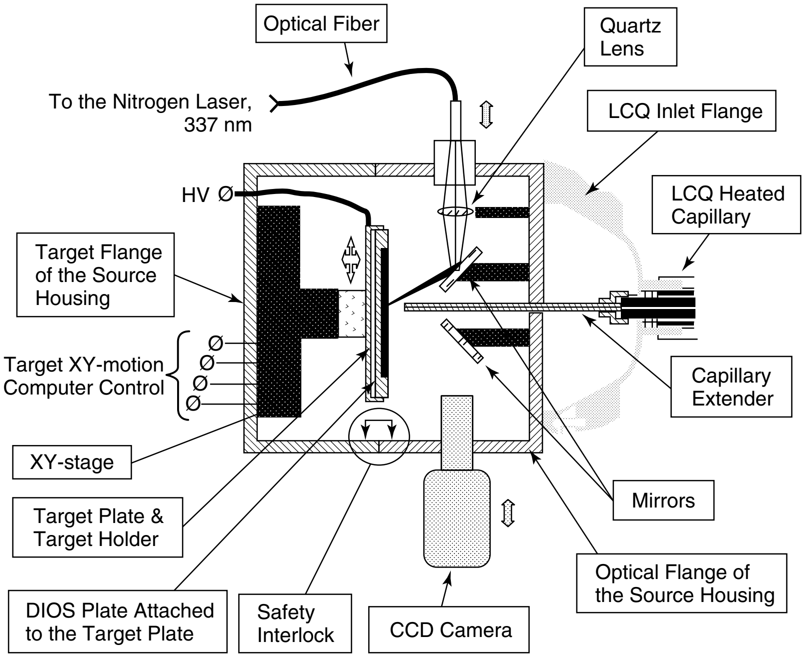
Optical Fiber
Quartz Lens
To the Nitrogen Laser, 337 nm
LCQ Inlet Flange
HV Ø
LCQ Heated Capillary
Target Flange of the Source Housing
Target XY-motion Computer Control
Ø
Ø
Ø
Ø
Capillary Extender
XY-stage
Mirrors
Target Plate & Target Holder
Optical Flange of the Source Housing
DIOS Plate Attached to the Target Plate
Safety Interlock
CCD Camera

Fig. 29.42 AP-MALDI ion source. The heated capillary of an ESI ion source is extended to allow enough space for the laser optics. LCQ is a tradename of Thermo Fisher Scientific for a LC-quadrupole ion trap instrument. This specific drawing incorporates a target for *desorption/ionization on porous silicon* (DIOS) (Wei, Buriak and Siuzdak, 1999) instead of a standard MALDI target. Adapted from Laiko *et al.* (2002) with permission. © John Wiley & Sons, 2002.

reagent molecules, that is, basically on Penning ionization. A stream of nitrogen or helium gas is passed through an high-voltage electrical discharge to generate the excited molecules or atoms that then exit as a directed stream at 1 l min^{-1}. The analyte is exposed to this ionizing gas stream in proximity to the entrance of an API interface.

DART can be applied to the analysis of gases, liquids, and solids on various surfaces. As such, DART provides a very sensitive tool for the detection of explosives, the characterization of inks on paper, and the analysis of residual components from hygiene products (Figure 29.44) (Jones, Cody and McClelland, 2006; Yew, Cody and Kravitz, 2008; Haefliger and Jeckelmann, 2007; Williams *et al.*, 2006).

29.9
Activation Methods for Tandem Mass Spectrometry

The term *tandem mass spectrometry* refers to all techniques where ions mass selected by a first stage are subjected to some type of activation (or less frequently chemical reaction) before the product(s) of this process are mass-analyzed on a second stage of the instrument (McLafferty, 1983; Busch, Glish and McLuckey, 1988). On beam instruments,

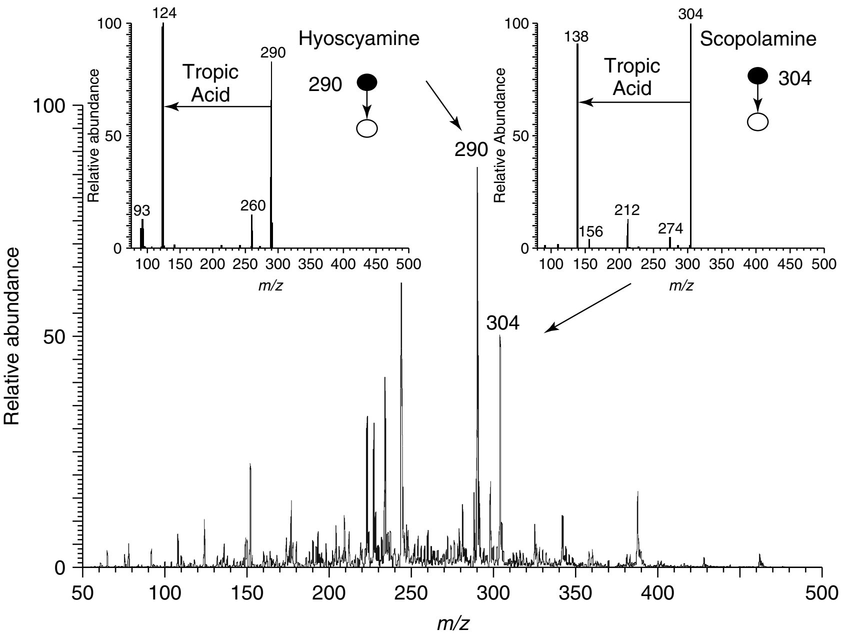

Fig. 29.43 DESI mass spectrum of deadly nightshade (*atropa belladonna*) seed using methanol : water (1 : 1) as spray solvent. Insets show tandem mass spectra of the $[M + H]^+$ ions of hyoscyamine, *m*/*z* 290, and scopolamine, *m*/*z* 304. Both alkaloid ions have the loss of tropic acid in common. Reproduced from Talaty, Takats and Cooks (2005) with permission. © The Royal Chemical Society, 2005.

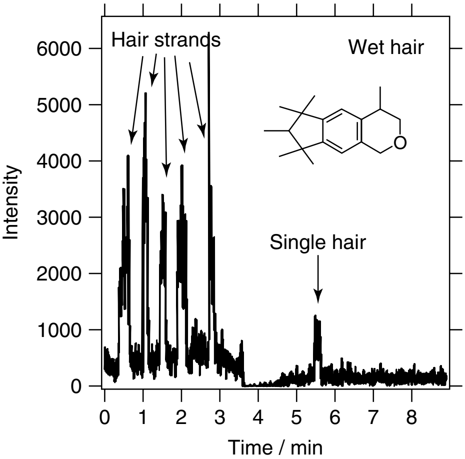

Fig. 29.44 DART-SRM analysis of hair washed with shampoo that contained 0.5% galaxolide® fragrance (musk replacement). The target compound was even identified from a single hair. SRM stands for *single reaction monitoring*, a mode of tandem MS operation where a selected fragment ion, here *m*/*z* 227, is detected only if produced from a given precursor; in this case $[M + H - H_2]^+$, *m*/*z* 257; SRM significantly improves the selectivity of an analysis. Adapted from Haefliger and Jeckelmann (2007) with permission. © John Wiley & Sons Ltd, 2007.

tandem MS requires a second mass analyzer behind the zone of activation; hence, this is sometimes referred to as *tandem in space*. On ion traps of any type, the sequence of selection, activation, and product analysis is usually effected inside the same unit by applying different modes of operation. Owing to its sequential character, this is then termed *tandem in time*. The activation step can be performed in many ways; the most common techniques are listed in Table 29.4. Table 29.5 lists the most important symbols in MS.

Tab. 29.4 Common activation methods for gas phase ions.

Method	**Acronym**	**Basic principle of ion excitation**	**Instrumentation**
Collision-induced dissociation (outdated terms: collisional activation, collisionally activated dissociation)	CID (outdated acronyms: CA, CAD)	Kinetic energy of ions is converted into vibrational excitation upon collisions; high-energy CID refers to keV ions typically once collided with He; low-energy CID refers to ions of 1–100 eV typically multiply collided with Ar (McLafferty, 1983; Busch, Glish and McLuckey, 1988). Suitable neutrals provided, collisions at very low energy can lead to chemical reactions (cf. Figure 39)	Suitable (scan) techniques for all types of mass analyzers available; independent of ion polarity
Infrared multiphoton dissociation	IRMPD	Absorption of IR photons provides vibrational excitation (Woodin, Bomse and Beauchamp, 1978; Watson, Baykut and Eyler, 1987)	On FT-ICR instruments; independent of ion polarity
Electron capture dissociation	ECD	Multiply charged positive ions undergo bond cleavage upon partial neutralization by capturing a thermal electron (Zubarev, Kelleher and McLafferty, 1998; Axelsson *et al.*, 1999)	On FT-ICR instruments; for multiply charged positive ions only
Electron transfer dissociation	ETD	Same effect as ECD, but negative radical ions are supplied (reagent ions) as the source of electrons (Coon *et al.*, 2005; Syka *et al.*, 2004)	Linear and three-dimensional quadrupole ion trap, orbitrap; for multiply charged positive ions only

Tab. 29.5 The most important symbols in mass spectrometry.

Symbol	Term	Explanation
$M^{+\bullet}$ ($M^{-\bullet}$)	Molecular ion	Positive (negative) ion formed from a molecule by removal (addition) of an electron
$[M + H]^+$, $[M + Na]^+$, $[M + NH_4]^+$, and so on	Quasimolecular ion	Ions formed instead of the molecular ion, but very closely related to it. The term also applies to $[M + Na]^+$, $[M + K]^+$ or $[M + Cl]^-$ ions, for example
MS	Mass spectrometry	Refers to the method; do not use it for a spectrum or instrument
MS/MS or MS^2	Tandem mass spectrometry	Mass-selected ions are subjected to fragmentation or reactions, the products are analyzed on a second stage
MS^3, MS^n	Tandem mass spectrometry	Higher order tandem MS experiments
m/z	Mass-to-charge ratio	Dimensionless ratio of the mass number and the number of elementary charges of an ion. The position of peaks on the abscissa of a mass spectrum is stated as "at m/z x"
u	Unified atomic mass	SI unit for atomic mass defined as 1/12 of the mass of the nuclide ^{12}C; $1\ u = 1.66055 \times 10^{-27}$ kg
B	Magnetic sector	Of a sector instrument
E	Electrostatic sector	Also termed ESA, electrostatic analyzer
ICR	Ion cyclotron resonance	To denote the method as well as the ICR cell of an instrument
LIT	Linear quadrupole ion trap	
Q and q	Quadrupole	Q for mass separating, q for RF-only ion guides
QIT	Quadrupole ion trap	A "three-dimensional" trap
TOF	Time of flight	To denote the method as well as theTOF analyzer of an instrument

Acronyms

AMS	Accelerator mass spectrometry
AE	Appearance energy
APCI	Atmospheric pressure chemical ionization
API	Atmospheric pressure ionization
AP-MALDI	Atmospheric pressure matrix-assisted laser desorption/ionization
APPI	Atmospheric pressure photoionization
CE	Capillary electrophoresis
CE	Charge exchange (equivalent to CT)
CF	Continuous flow
CF-FAB	Continuous-flow fast atom bombardment
CI	Chemical ionization
CID	Collision-induced dissociation
CT	Charge transfer
DART	Direct analysis in real time
DESI	Desorption electrospray/ionization
DIOS	Desorption/ionization on porous silicon
EA	Electron affinity
EC	Electron capture
ECD	Electron capture dissociation
EHC	Emitter heating current
EI	Electron ionization
ESI	Electrospray ionization
ETD	Electron transfer dissociation
FAB	Fast atom bombardment
FD	Field desorption
FI	Field ionization
FT-ICR	Fourier-transform ion cyclotron resonance
GC-MS	Gas chromatography-mass spectrometry
GD	Glow discharge
ICP	Inductively coupled plasma
IE	Ionization energy
IR-MS	Isotope ratio mass spectrometry
IR-MALDI	Infrared matrix-assisted laser desorption/ionization
IRMPD	Infrared multiphoton dissociation
LA	Laser ablation
LC-MS	Liquid-chromatography-mass spectrometry
LDI	Laser desorption/ionization
LIFDI	Liquid injection field desorption/ionization
LSIMS	Liquid secondary-ion mass spectrometry
MALDI	Matrix-assisted laser desorption/ionization
M_r:	Relative molecular mass (equal to MW)
MS	Mass spectrometry (not to be used for mass spectrometer or mass spectrum)
MS/MS	Mass spectrometry/mass spectrometry or tandem mass spectrometry
MUPI	Multiphoton ionization
MW	Molecular weight
Nano-ESI	Nanoelectrospray ionization
NCI	Negative-ion chemical ionization
NICI	Negative-ion chemical ionization
PA	Proton affinity
PCI	Positive-ion chemical ionization
PICI	Positive-ion chemical ionization
ppb	Parts per billion
ppm	Parts per million
QET	Quasi-equilibrium theory

RE	Recombination energy
REMPI	Resonance-enhanced multiphoton ionization
RF	Radio frequency
SIM	Selected ion monitoring
SIMS	Secondary-ion mass spectrometry
SRM	Selected reaction monitoring
SS-MS	Spark source mass spectrometry
TI	Thermal ionization
TOF	Time of flight
TSP	Thermospray
UV-MALDI	Ultraviolet matrix-assisted laser desorption/ ionization

References

Abian, J. (1999) *J. Mass Spectrom.*, **34**, 157–168.

Adlhart, C. and Chen, P. (2003) *Helv. Chim. Acta*, **86**, 941–949.

Allgood, C., Ma, Y.C. and Munson, B. (1991) *Anal. Chem.*, **63**, 721–725.

von Ardenne, M., Steinfelder, K. and Tümmler, R. (1971) *Elektronenanlagerungs-Massenspektrographie Organischer Substanzen*, Springer-Verlag, Heidelberg, p. 403.

Axelsson, J., Palmblad, M., Håkansson, K. and Håkansson, P. (1999) *Rapid Commun. Mass Spectrom.*, **13**, 474–477.

Barber, M., Bordoli, R.S., Elliott, G.J., Sedgwick, R.D., Tyler, A.N. and Green, B.N. (1982) *J. Chem. Soc. Chem. Commun.*, 936–938.

Barber, M., Vickerman, J.C. and Wolstenholme, J. (1980) *J. Chem. Soc., Faraday Trans. 1*, **76**, 549–559.

Becker, J.S. (2008) *Inorganic Mass Spectrometry: Principles and Applications*, 1st edn, John Wiley & Sons, Ltd, Chichester, p. 514.

Becker, J.S. and Dietze, H.J. (1999) *J. Anal. At. Spectrom.*, **14**, 1493–1500.

Becker, J.S., Zoriy, M., Becker, J.S., Dobrowolska, J. and Matusch, A. (2007) *J. Anal. At. Spectrom.*, **22**, 736–744.

Beckey, H.D. (1962) *Z. Naturforsch.*, **17A**, 1103–1111.

Beckey, H.D. (1971) *Field-Ionization Mass Spectrometry*, 1st edn, Pergamon, Elmsford, p. 359.

Beckey, H.D. (1977) *Principles of Field Desorption and Field Ionization Mass Spectrometry*, 1st edn, Pergamon Press, Oxford, p. 335.

Beckey, H.D., Hilt, E. and Schulten, H.-R. (1973) *J. Phys. E: Sci. Instrum.*, **6**, 1043–1044.

Beckey, H.D., Krone, H. and Röllgen, F.W. (1968) *J. Sci. Instrum.*, **1**, 118–120.

Bennett, C.L., Beukens, R.P., Clover, M.R., Grove, H.E., Liebert, R.B., Litherland, A.E., Purser, K.H. and Sondheim, W.E. (1977) *Science*, **198**, 508–510.

Benninghoven, A. (1983) *Int. J. Mass Spectrom. Ion Phys.*, **46**, 459–462.

Benninghoven, A. and Kirchner, F. (1963) *Z. Naturforsch.*, **18A**, 1008–1010.

Blakley, C.R., Carmody, J.J. and Vestal, M.L. (1980) *J. Am. Chem. Soc.*, **102**, 5931–5933.

Blakley, C.R. and Vestal, M.L. (1983) *Anal. Chem.*, **55**, 750–754.

Boesl, U. (2000) *J. Mass Spectrom.*, **35**, 289–304.

Born, M. and Oppenheimer, J.R. (1927) *Annalen der Phys.*, **84**, 457–484.

Boys, B.L. and Konermann, L. (2007) *J. Am. Soc. Mass Spectrom.*, **18**, 8–16.

Brauers, F. and von Bünau, G. (1990) *Int. J. Mass Spectrom. Ion Process*, **99**, 249–262.

Bristow, A.W.T. (2006) *Mass Spectrom. Rev.*, **25**, 99–111.

Bristow, A.W.T. and Webb, K.S. (2003) *J. Am. Soc. Mass Spectrom.*, **14**, 1086–1098.

Brown, K., Dingley, K.H. and Turteltaub, K.W. (2005) *Methods Enzymol.*, **402**, 423–443.

Brown, K., Tompkins, E.M. and White, I.N.H. (2006) *Mass Spectrom. Rev.*, **25**, 127–145.

Busch, K.L. (1995) *J. Mass Spectrom.*, **30**, 233–240.

Busch, K.L., Glish, G.L. and McLuckey, S.A. (1988) *Mass Spectrometry/Mass Spectrometry*, 1st edn, Wiley VCH, New York.

Caprioli, R.M. (1990) *Anal. Chem.*, **62**, 477A–485A.

Caprioli, R.M., Fan, T. and Cottrell, J.S. (1986) *Anal. Chem.*, **58**, 2949–2954.

Carroll, D.I., Dzidic, I., Stillwell, R.N., Haegele, K.D. and Horning, E.C. (1975) *Anal. Chem.*, **47**, 2369–2373.

Chen, P. (2003) *Angew. Chem. Int. Ed.*, **42**, 2832–2847.

Chen, H., Talaty, N.N., Takats, Z. and Cooks, R.G. (2005) *Anal. Chem.*, **77**, 6915–6927.

Chery, C.C., Moens, L., Cornelis, R. and Vanhaecke, F. (2006) *Pure Appl. Chem.*, **78**, 91–103.

Chesnov, S., Bigler, L. and Hesse, M. (2002) *Eur. J. Mass Spectrom.*, **8**, 1–16.

Cody, R.B., Laramee, J.A. and Durst, H.D. (2005) *Anal. Chem.*, **77**, 2297–2302.

Cole, R.B. (2000) *J. Mass Spectrom.*, **35**, 763–772.

Condon, E.U. (1926) *Phys. Rev.*, **28**, 1182–1201.

Coon, J.J., Shabanowitz, J., Hunt, D.F. and Syka, J.E.P. (2005) *J. Am. Chem. Soc. Mass Spectrom.*, **16**, 880–882.

Cotter, R.J. (1987) *Anal. Chim. Acta*, **195**, 45–59.

Covey, T.R., Bonner, R.F., Shushan, B.I. and Henion, J.D. (1988) *Rapid Commun. Mass Spectrom.*, **2**, 249–256.

Daehling, P., Röllgen, F.W., Zwinselmann, J.J., Fokkens, R.H. and Nibbering, N.M.M. (1982) *Fresenius Z. Anal. Chem.*, **312**, 335–337.

Dähling, P., Ott, K.H., Röllgen, F.W., Zwinselmann, J.J., Fokkens, R.H. and Nibbering, N.M.M. (1983) *Int. J. Mass Spectrom. Ion Phys.*, **46**, 301–304.

De Pauw, E., Agnello, A. and Derwa, F. (1986) *Mass Spectrom. Rev.*, **5**, 191–212.

Devienne, F.M. (1967) *Entropie*, **18**, 61–67.

Dillard, J.G. (1973) *Chem. Rev.*, **73**, 589–644.

Dole, R.B. (ed.) (1997) *Electrospray Ionization Mass Spectrometry - Fundamentals, Instrumentation and Applications*, 1st edn John Wiley & Sons, Ltd, Chichester.

Dole, M., Hines, R.L., Mack, L.L., Mobley, R.C., Ferguson, L.D. and Alice, M.B. (1968) *Macromolecules*, **1**, 96–97.

Dole, M., Mack, L.L., Hines, R.L., Mobley, R.C., Ferguson, L.D. and Alice, M.B. (1968) *J. Chem. Phys.*, **49**, 2240–2249.

Doroshenko, V.M., Laiko, V.V., Taranenko, N.I., Berkout, V.D. and Lee, H.S. (2002) *Int. J. Mass Spectrom.*, **221**, 39–58.

Dzidic, I., Stillwell, R.N., Carroll, D.I. and Horning, E.C. (1976) *Anal. Chem.*, **48**, 1763–1768.

Fenn, J.B. (2003) *Angew. Chem. Int. Ed.*, **42**, 3871–3894.

Fenn, J.B., Mann, M., Meng, C.K. and Wong, S.F. (1990) *Mass Spectrom. Rev.*, **9**, 37–70.

Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F. and Whithouse, C.M. (1989) *Science*, **246**, 64–71.

Fenner, N.C. and Daly, N.R. (1966) *Rev. Sci. Instrum.*, **37**, 1068–1070.

Field, F.H. and Munson, M.S.B. (1965) *J. Am. Chem. Soc.*, **87**, 3289–3294.

Franck, J. (1925) *Trans. Faraday Soc.*, **21**, 536–542.

Gomer, R. (1994) *Surf. Sci.*, **299/300**, 129–152.

Gomer, R. and Inghram, M.G. (1955) *J. Am. Chem. Soc.*, **77**, 500.

Grayson, M.A. (ed.) (2002) in *Measuring Mass – from Positive Rays to Proteins*, ASMS and CHF, Santa Fe and Philadelphia, p. 149.

Gross, J.H. (1998) *Rapid Commun. Mass Spectrom.*, **12**, 1833–1838.

Gross, J.H. (2004) *Mass Spectrometry – A Textbook*, Springer-Verlag Heidelberg, Heidelberg, p. 536.

Gross, J.H., Nieth, N., Linden, H.B., Blumbach, U., Richter, F.J., Tauchert, M.E., Tompers, R. and Hofmann, P. (2006) *Anal. Bioanal. Chem.*, **386**, 52–58.

Guenther, D. and Hattendorf, B. (2005) *Trends Anal. Chem.*, **24**, 255–265.

Haefliger, O.P. and Jeckelmann, N. (2007) *Rapid Commun. Mass Spectrom.*, **21**, 1361–1366.

Hanson, L., Fucini, P., Ilag, L.L., Nierhaus, K.H. and Robinson, C.V. (2003) *J. Biol. Chem.*, **278**, 1259–1267.

Harrison, A.G. (1992a) *Chemical Ionization Mass Spectrometry*, CRC Press, Boca Raton, p. 26.

Harrison, A.G. (1992b) *Chemical Ionization Mass Spectrometry*, 2nd edn, CRC Press, Boca Raton.

Hayen, H. and Karst, U. (2003) *J. Chromatogr. A*, **1000**, 549–565.

Hillenkamp, F. and Peter-Katalinic, J. (eds) (2007) *MALDI-MS. A Practical Guide to Instrumentation, Methods and Applications*, 1st edn Wiley-VCH Verlag GmbH, Weinheim.

Hoffmann, V., Kasik, M., Robinson, P.K. and Venzago, C. (2005) *Anal. Bioanal. Chem.*, **381**, 173–188.

Horning, E.C., Carroll, D.I., Dzidic, I., Haegele, K.D., Horning, M.G. and Stillwell, R.N. (1974) *J. Chromatogr. Sci.*, **12**, 725–729.

Horning, E.C., Horning, M.G., Carroll, D.I., Dzidic, I. and Stillwell, R.N. (1973) *Anal. Chem.*, **45**, 936–943.

Houk, R.S., Fassel, V.A., Flesch, G.D., Svec, H.J., Gray, A.L. and Taylor, C.E. (1980) *Anal. Chem.*, **52**, 2283–2289.

Hsu, C.S. and Qian, K. (1993) *Anal. Chem.*, **65**, 767–771.

Ikeda, T. (2005) *Radioisotopes*, **54**, 15–21.

Ikonomou, M.G. and Kebarle, P. (1994) *J. Am. Soc. Mass Spectrom.*, **5**, 791–799.

Ito, Y., Takeuchi, T., Ishii, D. and Goto, M. (1985) *J. Chromatogr.*, **346**, 161–166.

Jakubowski, N., Dorka, R., Steers, E. and Tempez, A. (2007) *J. Anal. At. Spectrom.*, **22**, 722–735.

Jochum, K.P. (1997) *Modern Analytical Geochemistry*, (ed R. Gill), Chap. 11 Addison Wesley Longman, Harlow, 188–199.

Jochum, K.P., Stoll, B., Pfänder, J.A., Seufert, M., Flanz, M., Maissenbacher, P., Hofmann, M. and Hofmann, A.W. (2001) *Fresenius J. Anal. Chem.*, **370**, 647–653.

Jones, R.W., Cody, R.B. and McClelland, J.F. (2006) *J. Forensic Sci.*, **51**, 915–918.

Karas, M., Bahr, U. and Gießmann, U. (1991) *Mass Spectrom. Rev.*, **10**, 335–357.

Karas, M., Bahr, U., Ingendoh, A. and Hillenkamp, F. (1989) *Angew. Chem.*, **101**, 805–806.

Karas, M., Ingendoh, A., Bahr, U. and Hillenkamp, F. (1989) *Biomed. Environ. Mass Spectrom.*, **18**, 841–843.

Karas, M. and Krüger, R. (2003) *Chem. Rev.*, **103**, 427–439.

Karger, B.L., Kirby, D.P., Vouros, P., Foltz, R.L. and Hidy, B. (1979) *Anal. Chem.*, **51**, 2324–2328.

Kauppila, T.J., Kuuranne, T., Meurer, E.C., Eberlin, M.N., Kotiaho, T. and Koistiainen, R. (2002) *Anal. Chem.*, **74**, 5470–5479.

Kawano, H., Hidaka, Y. and Page, F.M. (1983) *Int. J. Mass Spectrom. Ion Phys.*, **50**, 35–75.

Kawano, H. and Page, F.M. (1983) *Int. J. Mass Spectrom. Ion Phys.*, **50**, 1–33.

Kebarle, P. (2000) *J. Mass Spectrom.*, **35**, 804–817.

Kebarle, P. and Tang, L. (1993) *Anal. Chem.*, **65**, 972A–986A.

Keski-Hynnilä, H., Kurkela, M., Elovaara, E., Antonio, L., Magdalou, J., Luukkanen, L., Taskinen, J. and Koistiainen, R. (2002) *Anal. Chem.*, **74**, 3449–3457.

King, F.L. and Harrison, W.W. (1990) *Mass Spectrom. Rev.*, **9**, 285–317.

Kinter, M. and Sherman, N.E. (2000) *Protein Sequencing and Identification Using Tandem Mass Spectrometry*, John Wiley & Sons, Ltd, Chichester.

Kiser, R.W. (1965) *Introduction to Mass Spectrometry and Its Applications*, Prentice-Hall, Englewood Cliffs, p. 356.

Knighton, W.B. and Grimsrud, E.P. (1995) *Mass Spectrom. Rev.*, **14**, 327–343.

Koppenaal, D.W. (1990) *Anal. Chem.*, **62**, 303R–324R.

Krätschmer, W., Lamb, L.D., Fostiropoulos, K. and Huffman, D.R. (1990) *Nature*, **347**, 354–358.

Kroto, H. (1997) *Rev. Mod. Phys.*, **69**, 703–722.

Kutschera, W. (2005) *Int. J. Mass Spectrom.*, **242**, 145–160.

de Laeter, J.R. (2001) *Applications of Inorganic Mass Spectrometry*, 1st edn, John Wiley & Sons, Inc., New York, p. 474.

de Laeter, J.R., De Bièvre, P. and Peiser, H.S. (1992) *Mass Spectrom. Rev.*, **11**, 193–245.

Laiko, V.V., Baldwin, M.A. and Burlingame, A.L. (2000) *Anal. Chem.*, **72**, 652–657.

Laiko, V.V., Moyer, S.C. and Cotter, R.J. (2000) *Anal. Chem.*, **72**, 5239–5243.

Laiko, V.V., Taranenko, N.I., Berkout, V.D., Musselman, B.D. and Doroshenko, V.M. (2002) *Rapid Commun. Mass Spectrom.*, **16**, 1737–1742.

Laramée, J.A., Mazurkiewicz, P., Berkout, V. and Deinzer, M.L. (1996) *Mass Spectrom. Rev.*, **15**, 15–42.

Lattimer, R.P. (2001) in *Mass Spectrometry of Polymers* (eds G. Montaudo and R.P. Lattimer), CRC Press, Boca Raton, pp. 237–268.

Lattimer, R.P. and Schulten, H.-R. (1989) *Anal. Chem.*, **61**, 1201A–1215A.

Lehmann, W.D. (1996) *Massenspektrometrie in der Biochemie*, 1st edn, Spektrum Akademischer Verlag, Heidelberg.

Linden, H.B. (2004) *Eur. J. Mass Spectrom.*, **10**, 459–468.

Mamyrin, B.A. (1994) *Int. J. Mass Spectrom. Ion Process*, **131**, 1–19.

Mann, M., Meng, C.K. and Fenn, J.B. (1989) *Anal. Chem.*, **61**, 1702–1708.

Märk, T.D. (1986) in *Gaseous ion Chemistry and Mass Spectrometry* (ed. J.H. Futrell), John Wiley & Sons, Inc., New York, pp. 61–93.

McCrery, D.A., Ledford, E.B. Jr and Gross, M.L. (1982) *Anal. Chem.*, **54**, 1435–1437.

McFadden, W.H. (1979) *J. Chromatogr. Sci.*, **17**, 2–16.

McFadden, W.H., Schwartz, H.L. and Evans, S. (1976) *J. Chromatogr.*, **122**, 389–396.

McLafferty, F.W. (ed.) (1983) *Tandem Mass Spectrometry*, 1st edn, John Wiley & Sons, Inc., New York.

McLafferty, F.W. (2001) *Int. J. Mass Spectrom.*, **212**, 81–87.

McLafferty, F.W. and Turecek, F. (1993) *Interpretation of Mass Spectra*, 4th edn, University Science Books, Mill Valley.

Montaudo, G. and Lattimer, R.P. (eds) (2001) *Mass Spectrometry of Polymers*, 1st edn CRC Press, Boca Raton, p. 600.

Morris, H.R., Panico, M., Barber, M., Bordoli, R.S., Sedgwick, R.D. and Tyler, A.N. (1981) *Biochem. Biophys. Res. Commun.*, **101**, 623–631.

Munson, M.S.B. (2000) *Int. J. Mass Spectrom.*, **200**, 243–251.

Murgasova, R. and Hercules, D.M. (2003) *Int. J. Mass Spectrom.*, **226**, 151–162.

Nelis, T. and Pallosi, J. (2006) *Appl. Spectrosc. Rev.*, **41**, 227–258.

Nelson, D.E., Korteling, R.G. and Stott, W.R. (1977) *Science*, **198**, 507–508.

Nielen, M.F.W. (1999) *Mass Spectrom. Rev.*, **18**, 309–344.

Niessen, W.M.A. (1998) *J. Chromatogr. A*, **794**, 407–435.

Niessen, W.M.A. and Voyksner, R.D. (eds) (1998) *Current Practice of Liquid Chromatography-Mass Spectrometry*, Elsevier, Amsterdam, p. 438.

Nordhoff, E., Schürenberg, M., Thiele, G., Lübbert, C., Kloeppel, K.-D., Theiss, D., Lehrach, H. and Gobom, J. (2003) *Int. J. Mass Spectrom.*, **226**, 163–180.

O'Connor, P.B. and Costello, C.E. (2001) *Rapid Commun. Mass Spectrom.*, **15**, 1862–1868.

Okuyama, F., Hilt, E., Röllgen, F.W. and Beckey, H.D. (1977) *J. Vac. Sci. Technol.*, **14**, 1033–1035.

Overberg, A., Karas, M., Bahr, U., Kaufmann, R. and Hillenkamp, F. (1990) *Rapid Commun. Mass Spectrom.*, **4**, 293–296.

Overberg, A., Karas, M. and Hillenkamp, F. (1991) *Rapid Commun. Mass Spectrom.*, **5**, 128–131.

Pasch, H. and Schrepp, W. (2003) *MALDI-TOF Mass Spectrometry of Synthetic Polymers*, Springer-Verlag, Heidelberg, p. 298.

Peng, J. and Gygi, S.P. (2001) *J. Mass Spectrom.*, **36**, 1083–1096.

Penning, F.M. (1927) *Naturwissenschaften*, **15**, 818.

Platzner, I.T., Habfast, K., Walder, A.J. and Goetz, A. (1997) in *Modern Isotope Ratio Mass Spectrometry*, 1st edn (ed. I.T. Platzner), John Wiley & Sons, Ltd, Chichester, p. 514.

Popov, I.A., Chen, H., Kharybin, O.N., Nikolaev, E.N. and Cooks, R.G. (2005) *Chem. Commun.*, 1953–1955.

Posthumus, M.A., Kistemaker, P.G., Meuzelaar, H.L.C. and Ten Noever de Brauw, M.C. (1978) *Anal. Chem.*, **50**, 985–991.

Pramanik, B.N., Ganguly, A.K. and Gross, M.L. (eds) (2002) *Applied Electrospray Mass Spectrometry*, Marcel Dekker, New York, p. 434.

Prókai, L. (1990) *Pract. Spectrosc.*, **9**, 1–44.

Prókai, L. (1990) *Pract. Spectrosc.*, **9**, 169–225.

Rabrenovic, M., Ast, T. and Kramer, V. (1981) *Int. J. Mass Spectrom. Ion Phys.*, **37**, 297–307.

Raffaeli, A. and Saba, A. (2003) *Mass Spectrom. Rev.*, **22**, 318–331.

Ray, S.J., Andrade, F., Gamez, G., McClenathan, D., Rogers, D., Schilling, G., Wetzel, W. and Hieftje, G.M. (2004) *J. Chromatogr. A*, **1050**, 3–34.

Rayleigh, L. (1882) *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, **14**, 184–186.

Reemtsma, T. (2003) *J. Chromatogr. A*, **1000**, 477–501.

Robb, D.B., Covey, T.R. and Bruins, A.P. (2000) *Anal. Chem.*, **72**, 3653–3659.

Rosenstock, H.M. and Krauss, M. (1963) in *Mass Spectrometry of Organic Ions* (ed. F.W. McLafferty), Academic Press, London, pp. 1–64.

Rosu, F., Nguyen, C.H., De Pauw, E. and Gabelica, V. (2007) *J. Am. Soc. Mass Spectrom.*, **18**, 1052–1062.

Roussis, S. (1999) *Rapid Commun. Mass Spectrom.*, **13**, 1031–1051.

Šagi, D., Kienz, P., Denecke, J., Marquardt, T. and Peter-Katalinic, J. (2005) *Proteomics*, **5**, 2689–2701.

Schalley, C.A. (ed.) (2007) *Analytical Methods in Supramolecular Chemistry*, 1st ed., Wiley-VCH Verlag GmbH.

Schaub, T.M., Linden, H.B., Hendrickson, C.L. and Marshall, A.G. (2004) *Rapid Commun. Mass Spectrom.*, **18**, 1641–1644.

Schaub, T.M., Rodgers, R.P., Marshall, A.G., Qian, K., Green, L.A. and Olmstead, W.N. (2005) *Energy Fuels*, **19**, 1566–1573.

Schröder, E. (1991) *Massenspektrometrie – Begriffe und Definitionen*, 1st edn, Springer-Verlag, Heidelberg.

Schulten, H.-R. and Beckey, H.D. (1974) *Org. Mass Spectrom.*, **9**, 1154–1155.

Seiler, R. (1969) *Int. J. Quantum Chem.*, **3**, 25–32.

Shetty, V., Spellman, D.S. and Neubert, T.A. (2007) *J. Am. Soc. Mass Spectrom.*, **18**, 1544–1551.

Snyder, A.P. (2000) *Interpreting Protein Mass Spectra*, 1st edn, Oxford University Press, New York, p. 544.

Spengler, B. (2004) *J. Am. Soc. Mass Spectrom.*, **15**, 703–714.

Stuewer, D. (1990) *Fresenius J. Anal. Chem.*, **337**, 737–742.

Syka, J.E.P., Coon, J.J., Schroeder, M.J., Shabanowitz, J. and Hunt, D.F. (2004) *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 9528–9533.

Takats, Z., Cotte-Rodriguez, I., Talaty, N., Chen, H. and Cooks, R.G. (2005) *Chem. Commun.*, 1950–1952.

Takats, Z., Wiseman, J.M. and Cooks, R.G. (2005) *J. Mass Spectrom.*, **40**, 1261–1275.

Takats, Z., Wiseman, J.M., Gologan, B. and Cooks, R.G. (2004) *Science*, **306**, 471–473.

Talaty, N., Takats, Z. and Cooks, R.G. (2005) *Analyst*, **130**, 1624–1633.

Tanaka, K. (2003) *Angew. Chem. Int. Ed.*, **42**, 3861–3870.

Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y. and Yhoshida, T. (1988) *Rapid Commun. Mass Spectrom.*, **2**, 151–153.

Taylor, G.I. (1964) *Proc. R. Soc. Lond. A*, **280**, 383–397.

Tsuchiya, M. (1995) *Adv. Mass Spectrom.*, **13**, 333–346.

Tuniz, C., Bird, J.R., Fink, D. and Herzog, G.F. (1998) *Accelerator Mass Spectrometry - Ultrasensitive Analysis for Global Science*, CRC Press, Boca Raton, p. 371.

Vastola, F.J., Mumma, R.O. and Pirone, A.J. (1970) *Org. Mass Spectrom.*, **3**, 101–104.

Vastola, F.J. and Pirone, A.J. (1968) *Adv. Mass Spectrom.*, **4**, 107–111.

Verlinden, J., Gijbels, R. and Adams, F. (1986) *J. Anal. At. Spectrom.*, **1**, 411–419.

Volland, M.A.O., Adlhart, C., Kiener, C.A., Chen, P. and Hofmann, P. (2001) *Chem. Eur. J.*, **7**, 4621–4632.

Vorm, O., Roepstorff, P. and Mann, M. (1994) *Anal. Chem.*, **66**, 3281–3287.

Wahrhaftig, A.L. (1959) in *Advances in Mass Spectrometry* (ed. J.D. Waldron), Pergamon Press, Oxford, pp. 274–286.

Wahrhaftig, A.L. (1986) in *Gaseous ion Chemistry and Mass Spectrometry* (ed. J.H. Futrell), John Wiley & Sons, Inc., New York, pp. 7–24.

Wang, C.H., Huang, M.-W., Lee, C.-Y., Chei, H.-L., Huang, J.P. and Shiea, J. (1998) *J. Am. Soc. Mass Spectrom.*, **9**, 1168–1174.

Watson, C.H., Baykut, G. and Eyler, J.R. (1987) *Anal. Chem.*, **59**, 1133–1138.

Wei, J., Buriak, J.M. and Siuzdak, G. (1999) *Nature*, **399**, 243–246.

Weickhardt, C., Grun, C. and Grotemeyer, J. (1998) *Eur. Mass Spectrom.*, **4**, 239–244.

Welham, K.J., Domin, M.A., Johnson, K., Jones, L. and Ashton, D.S. (2000) *Rapid Commun. Mass Spectrom.*, **14**, 307–310.

Wendt, K.D.A. (2002) *Eur. J. Mass Spectrom.*, **8**, 273–285.

Whitehouse, C.M., Robert, R.N., Yamashita, M. and Fenn, J.B. (1985) *Anal. Chem.*, **57**, 675–679.

Whittal, R.M., Schriemer, D.C. and Li, L. (1997) *Anal. Chem.*, **69**, 2734–2741.

Williams, J.P., Patel, V.J., Holland, R. and Scrivens, J.H. (2006) *Rapid Commun. Mass Spectrom.*, **20**, 1447–1456.

Willoughby, R.C. and Browner, R.F. (1984) *Anal. Chem.*, **56**, 2625–2631.

Wilm, M.S. and Mann, M. (1994) *Int. J. Mass Spectrom. Ion Process*, **136**, 167–180.

Wilm, M. and Mann, M. (1996) *Anal. Chem.*, **68**, 1–8.

Wilm, M., Shevshenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T. and Mann, M. (1996) *Nature*, **379**, 466–469.

Winchester, M.R. and Payling, R. (2004) *Spectrochim. Acta B At. Spectrosc.*, **59B**, 607–666.

Wood, G.W. (1982) *Mass Spectrom. Rev.*, **1**, 63–102.

Woodin, R.L., Bomse, D.S. and Beauchamp, J.L. (1978) *J. Am. Chem. Soc.*, **100**, 3248–3250.

Yang, C. and Henion, J.D. (2002) *J. Chromatogr. A*, **970**, 155–165.

Yew, J.Y., Cody, R.B. and Kravitz, E.A. (2008) *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 7135–7140.

Zeleny, J. (1917) *Phys. Rev.*, **10**, 1–7.

Zenobi, R. and Knochenmuss, R. (1999) *Mass Spectrom. Rev.*, **17**, 337–366.

Zenobi, R., Zhan, Q. and Voumard, P. (1996) *Mikrochim. Acta*, **124**, 273–281.

Zubarev, R.A., Kelleher, N.L. and McLafferty, F.W. (1998) *J. Am. Chem. Soc.*, **120**, 3265–3266.

Further Reading

Ardrey, R.E. (2003) *Liquid Chromatography-Mass Spectrometry – An Introduction*, John Wiley & Sons, Ltd, Chichester.

Assamoto, B. (ed.) (1991) *Analytical Applications of Fourier Transform Ion Cyclotron Resonance Mass Spectrometry*, Wiley-VCH Verlag GmbH, Weinheim.

Becker, J.S. (2008) *Inorganic Mass Spectrometry: Principles and Applications*, 1st edn, John Wiley & Sons, Ltd, Chichester.

Beckey, H.D. (1977) *Principles of Field Desorption and Field Ionization Mass Spectrometry*, 1st edn, Pergamon Press, Oxford.

Budde, W.L. (2001) *Analytical mass spectrometry*, 1st edn, ACS and Oxford University Press, Washington, DC and Oxford.

Busch, K.L., Glish, G.L. and McLuckey, S.A. (1988) *Mass Spectrometry/Mass Spectrometry*, 1st edn, Wiley-VCH Verlag GmbH, New York.

Caprioli, R.M. (ed.) (1990) *Continuous-Flow Fast Atom Bombardment Mass Spectrometry*, John Wiley & Sons, Ltd, Chichester, p. 189.

Castleman, A.W. Jr, Futrell, J.H., Lindinger, W., Märk, T.D., Morrison, J.D., Shirts, R.B., Smith, D.L. and Wahrhaftig, A.L. (1986) in *Gaseous ion Chemistry and Mass Spectrometry* (ed. J.H. Futrell), John Wiley & Sons, Inc., New York.

Chapman, J.R. (ed.) (2000) *Mass Spectrometry of Proteins and Peptides*, Humana Press, Totowa.

Cooks, R.G., Beynon, J.H. and Caprioli, R.M. (1973) *Metastable Ions*, 1st edn, Elsevier, Amsterdam.

Dass, C. (2007) *Fundamentals of Contemporary Mass Spectrometry*, John Wiley & Sons, Inc., Hoboken.

De Hoffmann, E. and Stroobant, V. (2007) *Mass Spectrometry – Principles and Applications*, 2nd edn, John Wiley & Sons, Ltd, Chichester.

Dole, R.B. (ed.) (1997) *Electrospray Ionization Mass Spectrometry – Fundamentals, Instrumentation and Applications*, 1st ed., John Wiley & Sons, Chichester.

Duckworth, H.E., Barber, R.C. and Venkatasubramanian, V.S. (1986) *Mass Spectroscopy*, 2nd edn, Cambridge University Press, Cambridge.

Grayson, M.A. (ed.) (2002) *Measuring Mass – from Positive Rays to Proteins*, ASMS and CHF, Santa Fe and Philadelphia.

Gross, J.H. (2004) *Mass Spectrometry – A Textbook*, Springer-Verlag Heidelberg, Heidelberg.

Harrison, A.G. (1992) *Chemical Ionization Mass Spectrometry*, 2nd edn, CRC Press, Boca Raton.

Henderson, W. and McIndoe, S.J. (2005) *Mass Spectrometry of Inorganic and Organometallic Compounds*, John Wiley & Sons, Ltd, Chichester.

Hillenkamp, F. and Peter-Katalinic, J. (eds) (2007) *MALDI-MS. A Practical Guide to Instrumentation, Methods and Applications*, 1st edn, Wiley-VCH Verlag GmbH, Weinheim.

Kaltashov, I.A. and Eyles, S.J. (2005) *Mass Spectrometry in Biophysics: Conformation and Dynamics of Biomolecules*, John Wiley & Sons, Inc., Hoboken.

Kinter, M. and Sherman, N.E. (2000) *Protein Sequencing and Identification Using Tandem Mass Spectrometry*, John Wiley & Sons, Ltd, Chichester.

de Laeter, J.R. (2001) *Applications of Inorganic Mass Spectrometry*, 1st ed, John Wiley & Sons, Inc., New York.

Lehmann, W.D. (1996) *Massenspektrometrie in der Biochemie*, Spektrum Akademischer Verlag, Heidelberg.

Levsen, K. (1978) *Fundamental Aspects of Organic Mass Spectrometry*, 1st edn, Verlag Chemie, Weinheim.

March, R.E. and Hughes, R.J. (1989) *Quadrupole Storage Mass Spectrometry*, John Wiley & Sons, Ltd, Chichester, p. 496.

March, R.E. and Todd, J.F.J. (eds) (1995) *Practical Aspects of Ion Trap Mass Spectrometry*, Vol. 1, CRC Press, Boca Raton.

March, R.E. and Todd, J.F.J. (eds) (1995) *Practical Aspects of Ion Trap Mass Spectrometry*, Vol. 2, CRC Press, Boca Raton.

March, R.E. and Todd, J.F.J. (eds) (1995) *Practical Aspects of Ion Trap Mass Spectrometry*, Vol. 3, CRC Press, Boca Raton.

March, R.E. and Todd, J.F.J. (2005) *Quadrupole Ion trap Mass Spectrometry*, John Wiley & Sons, Inc., Hoboken.

McLafferty, F.W. (ed.) (1983) *Tandem Mass Spectrometry*, 1st edn, John Wiley & Sons, Inc., New York.

McLafferty, F.W., Turecek, F. (1993) *Interpretation of Mass Spectra*, 4th edn, University Science Books, Mill Valley.

Montaudo, G. and Lattimer, R.P. (eds) (2001) *Mass Spectrometry of Polymers*, 1st edn, CRC Press, Boca Raton.

Niessen, W.M.A. and Voyksner, R.D. (eds) (1998) *Current Practice of Liquid Chromatography-Mass Spectrometry*, Elsevier, Amsterdam.

Pasch, H. and Schrepp, W. (2003) *MALDI-TOF Mass Spectrometry of Synthetic Polymers*, Springer-Verlag, Heidelberg.

Platzner, I.T., Habfast, K., Walder, A.J. and Goetz, A. (1997) *Modern Isotope Ratio Mass Spectrometry*, 1st edn (ed. I.T. Platzner), John Wiley & Sons, Ltd, Chichester, p. 514.

Pramanik, B.N., Ganguly, A.K. and Gross, M.L. (eds) (2002) *Applied Electrospray Mass Spectrometry*, Marcel Dekker, New York.

Prókai, L. (1990) *Field Desorption Mass Spectrometry*, 1st edn, Marcel Dekker, New York.

Roboz, J. (1999) *Mass Spectrometry in Cancer Research*, 1st edn, CRC Press, Boca Raton.

Rossi, D.T. and Sinz, M.W. (eds) (2002) *Mass Spectrometry in Drug Discovery*, Marcel Dekker, New York.

Schalley, C.A. (ed.) (2003) *Modern Mass Spectrometry*, Springer, New York.

Schalley, C.A. (ed.) (2007) *Analytical Methods in Supramolecular Chemistry*, 1st edn, Wiley-VCH Verlag GmbH.

Schlag, E.W. (ed.) (1994) *Time of Flight Mass Spectrometry and its Applications*, 1st edn Elsevier, Amsterdam.

Snyder, A.P. (2000) *Interpreting Protein Mass Spectra*, 1st edn, Oxford University Press, New York.

Sparkman, O.D. (2006) *Mass Spectrometry Desk Reference*, 2nd edn, Global View Publishing, Pittsburgh.

Splitter, J.S., and Turecek, F. (eds) (1994) *Applications of Mass Spectrometry to Organic Stereochemistry*, 1st edn, Verlag Chemie, Weinheim.

Taylor, H.E. (2000) *Inductively Coupled Plasma Mass Spectroscopy*, 1st edn, Academic Press, London.

Tuniz, C., Bird, J.R., Fink, D. and Herzog, G.F. (1998) *Accelerator Mass Spectrometry – Ultrasensitive Analysis for Global Science*, CRC Press, Boca Raton.

Watson, J.T. and Sparkman, O.D. (2007) *Introduction to Mass Spectrometry*, 4th edn, John Wiley & Sons, Ltd, Chichester.

Wilson, R.G., Stevie, F.A. and Magee, C.W. (1989) *Secondary Ion Mass Spectrometry: A Practical Handbook for Depth Profiling and Bulk Impurity Analysis*, John Wiley & Sons, Ltd, Chichester.

Yinon, J. (ed.) (1994) *Forensic Applications of Mass Spectrometry*, CRC Press, Boca Raton.

31
Chemometrics and Multivariate Analysis

Richard G. Brereton

31.1 Introduction 1105
31.1.1 Origins 1105
31.1.2 Development 1106
31.1.3 Software and Methodology 1107
31.1.4 Multivariate Matrices 1108
31.1.5 Notation 1108
31.2 Multivariate Curve Resolution 1109
31.2.1 Motivation 1109
31.2.2 Principal Components Analysis 1110
31.2.3 How Many Components? 1112
31.2.4 Resolution 1113
31.2.4.1 Noniterative Methods 1113
31.2.4.2 Iterative Methods 1115
31.2.4.3 Combining Information 1116
31.3 Calibration 1117
31.3.1 Univariate Calibration 1117
31.3.1.1 Classical Calibration 1118
31.3.1.2 Inverse Calibration 1119
31.3.1.3 Determining the Relationship between Variables 1119
31.3.2 Multiple Linear Regression 1120
31.3.3 Principal Components Regression 1122
31.3.4 Partial Least Squares 1123
31.3.5 Model Optimization and Validation 1125
31.4 Exploratory Methods 1126
31.4.1 Principal Components Analysis 1127
31.4.2 Cluster Analysis 1129
31.5 Supervised Methods 1133
31.5.1 Developing and Assessing Models 1133
31.5.1.1 **Modeling the Training Set** 1133
31.5.1.2 Test Sets, Cross Validation, and the Bootstrap 1134
31.5.1.3 Application 1136

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

31.5.2 Classical Discriminant Analysis 1136
31.5.2.1 Univariate Classification 1136
31.5.2.2 Bivariate and Multivariate Discriminant Models 1137
31.5.3 Discriminant Partial Least Squares 1139
31.5.4 K-Nearest Neighbors 1140
31.6 Multivariate Statistical Process Control 1141
31.7 Conclusions 1143
Glossary 1143
References 1145
Further Reading 1147

31.1
Introduction

31.1.1
Origins

The word *chemometrics* is said to have been coined in the early 1970s by the Swedish organic chemist Svante Wold in submitting a grant application. Over that period there were many groups working in what might in modern parlance be referred to as *chemometrics*, but which was, at the time, called by other names such as *chemical pattern recognition, machine learning*, or *laboratory optimization*. These diverse groups often published the subject in different journals: emerged as a discipline seriously in 1980s, with the organization of a School in Cosenza, Italy, in 1983 (Kowalski, 1984), the founding of two journals in 1986 (Chemometr. Intell. Lab Syst.) and 1987 (J. Chemometrics), and the publication of the first text books (Sharaf, Illman and Kowalski, 1986; Massart *et al.*, 1988; Martens and Næs, 1989; Massart and Kaufmann, 1983; Malinowski and Howery, 1980). These books have been widely used; for example, by January 2008 Martens and Næs (1989) had been cited 3145 times, Malinowski and Howery (1980) including two subsequent editions 2164 times, and Massart *et al.* (1988) 1432 times. As an example of the rapid growth of interest over the past two decades, citations per year for Martens and Næs (1989) are shown in Figure 31.1; now there are two sequels from these authors that are cited more often (Naes *et al.*, 2002; Martens and Martens, 2000). In the next two decades, there were many very significant books written on the subject (Brereton, 2003; Massart *et al.*, 1997; Vandeginste *et al.*, 1998; Otto, 1998; Beebe, Pell and Seasholtz, 1998; Kramer, 1998; Gemperline, 2006; Brereton, 2007; Massart *et al.*, 1990; Brereton *et al.*, 1992; Adams, 2004; Meloun, Militky and Forina, 1992; Meloun *et al.*, 1994; Brereton, 1992; Mark and Workman, 2003; Haswell, 1992; Esbensen, 2002; Eriksson *et al.*, 2001) often from different viewpoints.

The earliest practitioners of the subject were not primarily regarded as chemometrics experts, but worked in areas such as pharmacy, analytical chemistry, process chemistry, and organic chemistry. New subjects are often slow to develop in academia, as they often have to work within the boundaries of quite well established disciplines to fit within historic departmental structures, and a

Encyclopedia of Applied Spectroscopy. Edited by David L. Andrews.
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40773-6

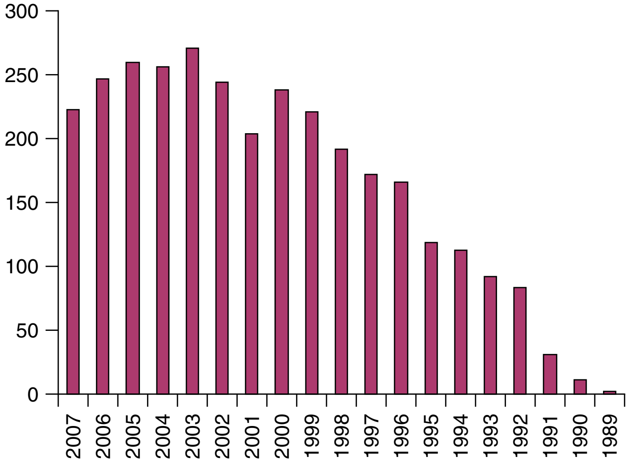

Fig. 31.1 Annual citations for Martens and Næs (1989).

research base needs to be established according to funding patterns that dictate applications.

31.1.2
Development

Historically, however, there was a major development in near infrared (NIR) spectroscopy in the 1980s as cheap, cost-effective probe technology emerged, and there were several important areas of application including process monitoring especially in petrochemical and later in pharmaceutical industries and food chemistry that catalyzed early chemometrics applications. NIR data often involve several highly overlapping peaks that cannot be easily separated or deconvoluted to build a picture of the composition of a mixture. However, NIR is ideal for looking at the overall properties of mixtures and many industrial products, such as food or petrol, that consist of a mixture of compounds. Hence, a method of spectroscopy that is cost effective and can rapidly (and often noninvasively without the need to sample) measure properties that depend on the mixture as a whole has significant advantages. NIR instruments are often stable and have low noise levels. This, however, means that statistically based approaches are required to relate the NIR spectra to the characteristics of a product or process such as acceptability and origins, and this provided a major early driving force for chemometrics. Almost all NIR instruments have a chemometrics capability and this classical area of application is well established with well-described algorithms and a high standard of software. Other spectroscopies have followed suit, especially in cases where technology has improved; for example, cost-effective probes for ultraviolet/visible (UV/vis) spectroscopy have been developed over the past decade, and with most pharmaceuticals having prominent chromophores, the application of chemometrics in reaction monitoring and product quality assessment using UV/vis and chemometrics has grown significantly.

At the other end of the scale, there is a large amount of exploratory work that can be performed using larger instruments such as acoustic spectroscopy, NMR (Nuclear Magnetic Resonance), and MS (Mass Spectrometry), which also provide rich pickings for the chemometrics expert.

31.1.3
Software and Methodology

Often there is a confusion between software capability and chemometrics methodology. In many cases, there are a large number of methods available for the analysis of spectroscopic data but only a small fraction are incorporated into user-friendly software. This is because of the economics of developing software packages, often with a quite specialist market. A package that is tailored for a less common application, such as acoustic spectroscopy, may take as long to develop and cost as much in resources as one for a widespread area such as NIR. Hence, people who are not good programmers are often restricted to prepackaged methods that are primarily targeted toward quite a restricted set of applications. Although basic methods such as principal components analysis (PCA) are widely available, most spectroscopies pose quite specific data analytical problems and require quite specific strategies before pattern recognition. Another problem is that the originators of many of the chemometric-based software companies, most of which were founded in the 1980s, were primarily focused on common methods such as PCA and partial least squares (PLS), and also developed approaches, such as cross validation, that were very efficient in usage of computer power. Moore's law states that computing power doubles approximately every 18 months: on that basis, in a period of 20 years, computing power has increased by a factor of 2^{13} or about 10 000. This means that a calculation that could be performed in a week on a desktop PC in the mid-1980s required only 1 minute 20 years later. Hence, new and much more computationally intense, methods can be conceived and implemented and can provide an assessment of a spectrum within a very short time span compared to those used two decades ago, when chemometrics methodologies started to become commercially viable. The problem here though is that software developers need to keep pace – and the human investment in programming has not changed significantly – programmers are not 10 000 times faster at producing code : with new development tools perhaps code can be developed 2 or 3 times faster than 20 years ago but not more. Therefore, a large amount of the commercial packaged software that is available is based on the philosophy of the mid-1980s, and many spectroscopists who do not have the time or expertise to develop their own code are trapped within this environment, even though if one were to start from the beginning again perhaps quite different methods would become widespread. Areas such as bioinformatics have a more recent vintage and users often do not require real-time desktop solutions. They are happy to perform data mining on a remote mainframe and most users are programmers in their own rights. Hence new methods based on machine learning can easily be incorporated. However, this problem means that there is a gap opening up between the application-based laboratory user of packaged chemometrics software who is often quite limited in the methodology that can be applied and the pioneering research groups that can implement code rapidly but do not have the resources (or investment) to develop user-friendly benchtop software.

31.1.4
Multivariate Matrices

Most of chemometrics is characterized by signals that are multivariate, that is, they are best described in two dimensions and are commonly represented by matrices as in Figure 31.2. Chemometric methods are most valuable when there are several spectra, which are normally represented as rows (often called *samples*) of the multivariate matrix. The columns of the matrix represent the sensors, normally wavelengths (or frequencies). The elements of the matrix are the raw spectroscopic measurements, usually the intensity of a spectrum at a specific wavelength.

We do not discuss experimental design or signal preprocessing (e.g., derivatives and smoothing) in this chapter but these are normally important parts of the data analytical procedure.

31.1.5
Notation

At this point, we digress to give a short introduction to matrices. Matrix notation is generally useful throughout chemometrics.

In this chapter, we employ recommended notation as follows.

- A scalar (or a single number, e.g., a single concentration) is denoted in italics, for example, x.
- A vector (or a row/column of numbers, e.g., a single spectrum) is denoted by lower case bold roman, for example, $\mathbf{x}$.
- A matrix (or a two-dimensional array of numbers, e.g., a set of spectra) is denoted by upper case bold roman, for example, $\mathbf{X}$. Scalars and vectors can also be regarded as matrices, one of whose dimensions is equal to 1.

The dimensions of a matrix are characterized by the number of rows (e.g., the number of spectra) and the number of columns (e.g., the number of wavelengths). Hence, a dataset that consists of 20 spectra recorded at 1000 wavelengths can be arranged as a 20 × 1000 matrix, with 20 rows and 1000 columns.

The inverse of a matrix is denoted by "-1" so that $\mathbf{X}^{-1}$ is the inverse of $\mathbf{X}$. Only square matrices (where the number of rows and columns are equal) can have inverses. A few square matrices do not have inverses and this happens when the columns are correlated. The product of a matrix and its inverse is the identity matrix.

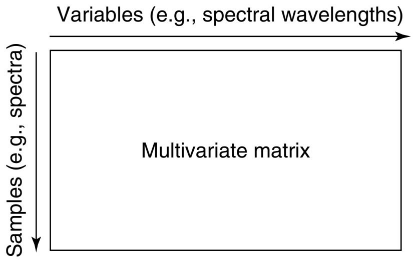

Fig. 31.2 Multivariate matrix as obtained in spectroscopy.

The transpose of a matrix is often denoted by "′" and involves swapping the rows and columns; hence, for example, the transpose of a 5 × 8 matrix is an 8 × 5 matrix.

Most people in chemometrics use the dot ("·") product of matrices (and vectors). For the dot product $\mathbf{X} \cdot \mathbf{Y}$ to be viable, the number of columns in $\mathbf{X}$ must equal the number of rows in $\mathbf{Y}$ and the dimensions of the product equal the number of rows in the first matrix and number of columns in the second one. Note that in general $\mathbf{Y} \cdot \mathbf{X}$ (if allowed) is usually different to $\mathbf{X} \cdot \mathbf{Y}$. For simplicity, sometimes the dot symbol is omitted, although it is always the dot product that is computed.

Finally, we introduce the concept a pseudoinverse of a vector or matrix, which is denoted as $\mathbf{X}^{+}$ so that if

$$\mathbf{Z} = \mathbf{X} \cdot \mathbf{Y} \tag{31.1}$$

$$\mathbf{Y} = \mathbf{X}^{+} \cdot \mathbf{X} \cdot \mathbf{Y} = \mathbf{X}^{+} \cdot \mathbf{Z} \tag{31.2}$$

It is important to note that only square matrices (where the number of columns and rows are equal) have mathematical inverses, but for matrices or vectors whose dimensions are not the same, there are two types of pseudoinverse, both of which can be denoted by $\mathbf{X}^{+}$, the right being

$$\mathbf{X}' \cdot (\mathbf{X} \cdot \mathbf{X}')^{-1}$$

and the left being

$$(\mathbf{X}' \cdot \mathbf{X})^{-1} \cdot \mathbf{X}'$$

The prime (′) implies a transpose, in which rows and columns are interchanged; for example, the transpose of a 5 × 8 matrix is an 8 × 5 matrix. When using vectors, the pseudoinverse mentioned above simplifies still further, but this is not discussed here.

31.2
Multivariate Curve Resolution

31.2.1
Motivation

An early and important application of chemometrics to spectroscopy involves signal resolution; this usually involves the determination of the number, characteristics, and relative amounts of compounds in a series of mixture spectra. Some authors call this *factor analysis*, some *multivariate curve resolution*, and others *deconvolution*. There is a vast array of techniques available to do this; however, only a few are summarized in this chapter. A historic text is by Malinowski and Howery (Malinowski and Howery, 1980) who developed a large array of approaches. This has been updated in a more recent edition (Malinowski, 2002). Multivariate signals often occur in, for example, reactions and involve recording a series of spectra as a function of time as different compounds change in relative concentration and in equilibria studies where each spectrum (e.g., taken at different pHs) is composed of several different spectroscopically active species. The objective is to obtain rate constants or to obtain equilibrium constants, as illustrated in Figure 31.3. The methods described in this section involve the determination of the (relative) concentrations of compounds in mixtures without the addition of external standards.

The initial step is to prepare the data, for example, by smoothing, baseline correction/derivatives, alignment if necessary, and a number of other approaches

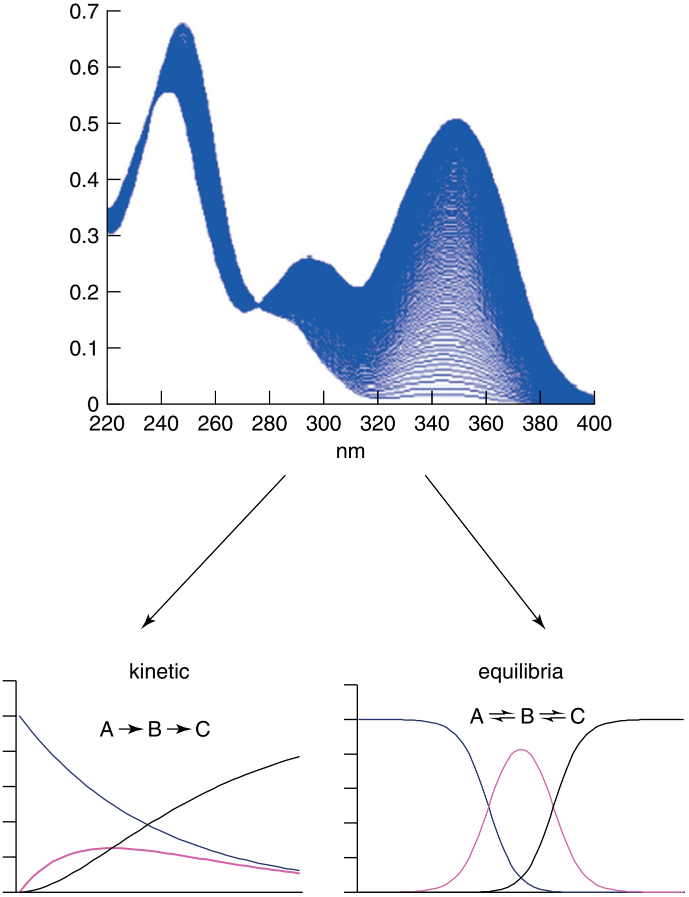

Fig. 31.3 Common motivations for resolving a series of spectra.

to reduce noise. However, these are not discussed in this chapter.

31.2.2
Principal Components Analysis

The first step is normally to perform PCA (Wold, Esbensen and Geladi, 1987; Joliffe, 1987). This common technique aims at converting an experimental data matrix from a series of spectra, which is arranged as a matrix **X** whose rows correspond to spectra (often taken at a specified time in a reaction pathway or pH in an equilibrium study) and whose columns correspond to spectroscopic wavelengths to a number of principal components (PCs) defined as follows:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P} + \mathbf{E} \tag{31.3}$$

where

- **T** are called the *scores* and have as many rows as the original data matrix;
- **P** are the *loadings* and have as many columns as the original data matrix;

- the number of columns in the matrix **T** is equal to the number of rows in the matrix **P**; and
- **E** is an error matrix.

PCs are abstract mathematical entities with certain specific mathematical properties, the most important being the one that scores and loadings are orthogonal, that is, the vector product of any two different scores or loadings vectors is 0. There are several different algorithms for PCA including singular value decomposition (SVD) and nonlinear iterative partial least squares (NIPALS). Using NIPALS (which is being adopted in this chapter), the loadings matrices are also normal, that is, the sum of squares of their elements is equal to 1. Each column of **T** and row of **P** is a characteristic of a PC. The size of each successive PC is often called an *eigenvalue* and reduces for subsequent components. Eigenvalues can be defined mathematically as the sum of squares of each successive scores vector or column of **T**, but this definition depends on the algorithm used to compute PCs (the NIPALS method being very common).

An important function of factor analysis is to relate PCs (sometimes called *abstract factors*) to chemical factors that can be directly interpreted as illustrated in Figure 31.4. Each chemical factor could be regarded as a chemical compound in the spectroscopic mixture, which has various characteristics including a spectrum and a concentration profile as follows:

$$\mathbf{X} = \mathbf{C} \cdot \mathbf{S} + \mathbf{E} \qquad (31.4)$$

- **C** is the matrix consisting of the elution profiles of each compound and
- **S** is the matrix consisting of the spectra of each compound.

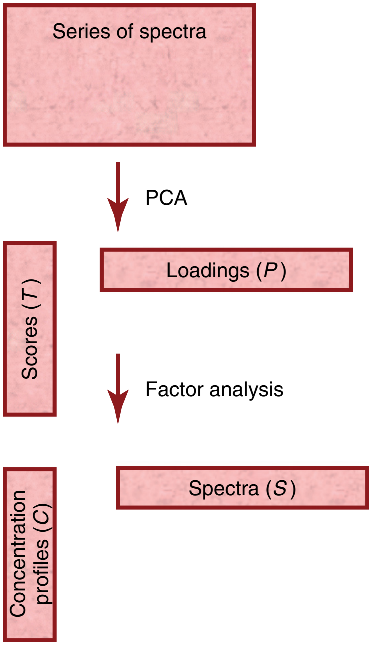

Fig. 31.4 Relationship between PCA and factor analysis.

31.2.3
How Many Components?

Normally, the first stage in factor analysis is the determination of the number of significant components in a series of mixture spectra. The number of significant PCs should be usually equal to the number of compounds in a mixture under most circumstances, although there may be problems. The first is if the concentration of two or more compounds are perfectly correlated, the number of significant PCs or factors will be less than the number of compounds in the mixture. For example, a reaction of the form A + B → C will appear to consist of two factors and not three because there are correlations between the reactants. The second is if there are additional effects such as due to variable baseline, temperature, instrumental drift, or significant noise levels, these could increase the apparent number of factors. There will always be some noise from an instrument; hence, even if there are, for example, four uncorrelated species in a mixture, a four-component model will not exactly fit the data, as there will be a small amount of error (matrix **E** above).

There are numerous methods for the determination of the number of significant components in a series of spectra. There are several groups of methods.

1. The first group of methods uses the size of each successive PC. This is measured by its eigenvalue, which is sometimes defined by the sum of squares of the scores. For each additional PC, the eigenvalue decreases. There are numerous approaches for determining the cut-off threshold, many of which have been developed and described by Malinowski (Malinowski, 2002).
2. The second group of methods is based on cross validation (Wold, 1978). These approaches involve removing one or more samples (or portions of the data) and using the rest of the samples to predict their properties. For example, if one performs PCA on 19 out of 20 spectra, one could determine the loadings of the data matrix using 19 samples. The score of the 20th sample could be predicted and this would then be used to estimate the spectrum of this sample and to see how close it is to the true spectrum. If the PCs are real PCs, they should model factors that are actually in the data, but if too many have been calculated, the later ones will be influenced primarily by noise and as such will provide poor prediction. The procedure is repeated until all the spectra are left out at least once. An optimum number of components will result in a minimum error.
3. The third set of approaches is based on the bootstrap (Efron and Tibshirani, 1993). This approach is not common in traditional spectroscopic data analysis because it is computationally intensive, but as computing power has become more cost effective and is more accessible, it is likely to be adopted more in the future. These methods involve removing several samples at a time, and then iteratively repeating this often a hundred times or more. The PC model is applied to the samples that remain each time and then tested on the samples that are left out. The number of components in the data is usually assessed by majority vote, for example, if the optimum is found to be 4 PCs in 85

out of 100 iterations, then this value is employed.

4. There are a number of techniques, such as orthogonal projection approach and SIMPLISMA (De Braekeleer and Massart, 1997; Windig *et al.*, 1992), and eigenvalue plots (Gampp *et al.*, 1985) that primarily rely on graphical inspection of data.

Then, this first step is used to determine the dimensions of the scores and loadings matrices. Hence, if the original data matrix has dimensions 30 × 1000 and if the number of significant PCs is denoted by A, then

- the dimensions of **T** will be $30 \times A$ and
- the dimensions of **P** will be $A \times 1000$.

The dimensions of **C** and **S** will be the same as **T** and **P**, respectively. Naturally, if the number of significant components is incorrectly estimated, then there can be problems with the reconstruction of the spectra and concentration profiles, but often this is evident visually.

31.2.4
Resolution

Once the number of components has been established, it is necessary to resolve the spectra into individual factors. One important principle behind most methods

$$\mathbf{X} \approx \mathbf{T} \cdot \mathbf{P} = \mathbf{C} \cdot \mathbf{S} \tag{31.5}$$

$$\mathbf{C} = \mathbf{T} \cdot \mathbf{R} \text{ and } \mathbf{S} = \mathbf{R}^{-1} \cdot \mathbf{P} \tag{31.6}$$

so to convert from scores to concentration profiles and from loadings to spectra, it is necessary to find a matrix **R** that is variously called a *transformation*, *rotation*, or *regression matrix*. This method is sometimes referred to as *principal components regression* (*PCR*). Alternatively, knowing **X** and either **C** or **S**, one can find the missing matrix, or if one has a guess of both **C** and **S**, it will be considered a good guess if one can predict **X** well.

There are a plethora of approaches that are summarized below.

31.2.4.1 Noniterative Methods

Many methods for resolution of mixtures involve finding selective variables for each compound: these could be specific wavelengths unique to each particular compound or specific points in time where one species (e.g., in a reaction mixture) is uniquely present. The first set of methods is where every component in a mixture contains some selective or unique information, which is used as a handle for factor analysis, and the second where there are some components without selective information.

Where all components have some selective information, the first step is to find selective or key variables for each component in the mixture. These could be found visually or it is possible to look at small regions (windows) in the mixture spectra and see which regions (if any) are characterized by one significant eigenvalue and thus are likely to correspond to the absorption of a single compound.

Once diagnostic variables have been found, the question is, what is the next stage? Multiple linear regression (MLR) is a common way to complete the picture. If concentration profiles are known (or can be approximated) using key variables such as m/z values, or wavenumbers, we can estimate the spectra by

$$\mathbf{S} \approx \mathbf{C}^{+} \cdot \mathbf{X} \tag{31.7}$$

where superscript + stands for the pseudoinverse.

Some more mathematically oriented chemometricians prefer PCR. There are several sophisticated algorithms in the literature, but a common basis of most of the approaches is to try to find a rotation, transformation, or regression matrix between the scores of a portion of the data and the known information. The first step involves performing PCA over the entire region of the spectra, and keeping a scores matrix **T** that has as many columns as compounds are suspected to be in the mixture. For example, if we know the pure spectral regions for each of the compounds, then we know their concentration profiles **C** from these regions and thus **R** and **S** can be estimated from the scores by

$$\mathbf{R} = \mathbf{T}^{+} \cdot \mathbf{C} \tag{31.8}$$

$$\mathbf{S} \approx \mathbf{R}^{-1} \cdot \mathbf{P} \tag{31.9}$$

Some PCA-based approaches are also called *evolving* (or *evolutionary*) *factor analysis*, but there is a significant terminology problem in the literature, as EFA also denotes the use of eigenvalue plots to determine how many components are in a mixture. There is a lot of history in the names chemometricians use for methods in this area, but the main thing is to keep a very cool head and understand the steps in each approach rather than become confused by the huge variety of names.

When only some components exhibit selectivity, the situation become slightly difficult. A typical example is a UV/vis spectrum of a mixture in which some of the compounds are buried within the spectra and do not have any characteristic wavelengths. Fortunately, PCA can be used here. There are several different ways of exploiting this. One approach uses the idea of a zero-intensity window. The first step is to determine regions for each compound where they are *not* present either in the spectrum or in the profile (e.g., pH profile). This information may be obtained from a number of approaches such as eigenvalue plots. In this region, we expect the intensity of the specific compound to be zero; hence, it is possible to find a vector **r** for each component so that

$$\mathbf{0} = \mathbf{T}_0 \cdot \mathbf{r} \tag{31.10}$$

where **0** is a vector of zeros, if the zero-composition region is for the samples, and $\mathbf{T}_0$ is the portion of the scores in this region. However, there is a problem, that is, one of the elements in the vector **r** must be set to an arbitrary number, usually the first coefficient is set to 1, but this does not have a serious effect on the algorithm. So long as there is always a region of the spectra or samples where at least one component in a mixture is absent, which can be identified, this procedure should always be possible to perform. If one is uncertain about the true number of compounds (or PCs), it is possible simply to change this and to see what happens with different guesses and see whether sensible results are obtained. The matrix **R** can be set up as mentioned above, and the estimated profiles of all the compounds are obtained over the entire region by $\hat{\mathbf{C}} = \mathbf{T} \cdot \mathbf{R}$ and the spectra by $\hat{\mathbf{S}} = \mathbf{R}^{-1} \cdot \mathbf{P}$.

31.2.4.2 **Iterative Methods**

The methods discussed in Section 31.2.4.1 are noniterative, that is, they have a single step that leads to a single solution. However over the years, chemometricians have refined these approaches to develop iterative methods that are based on many of the aforementioned principles but do not stop at the first solution and continue until the solution converges. These methods are particularly useful where only some analytes exhibit selective regions.

First, it is common to produce an initial estimate of the spectrum of each pure species in a series of mixture spectra. In equilibrium studies, there are often regions, for example, of pH, where one compound predominates or is uniquely present. One simple method is called the *needle* search. For each component in a mixture, the datapoints of maximum purity are taken as the initial guesses of the spectra of each component; these are simply spikes at the first stage, that is, one spike for each component: remember that when studying mixtures there may not be any samples that are characteristic of just one compound (e.g., a point in time of reaction mixture), but nevertheless there will be regions of maximum purity for each component in the mixture. Methods such as eigenvalue plots, OPA, and SIMPLISMA can be employed to find these regions, each pure variable being the starting guess of each component. None of these approaches are likely to provide perfect estimates of the spectra and result in initial estimates or guesses.

A feature of most of these methods is that different criteria can then be employed to assess the desirability of the model. Alternating least squares (ALS) (Tauler, Kowalski and Fleming, 1993) is a popular choice. Consider estimating, perhaps imperfectly, the above-mentioned spectra. We can now estimate, in turn, the concentration profiles by

$$\mathbf{C} = \mathbf{X} \cdot \mathbf{S}^{+} \tag{31.11}$$

If a good model is obtained, the concentration profiles should obey certain properties: the most common is nonnegativity, but other constraints can be imposed such as unimodality if desired, especially if there is some knowledge of the equilibrium process. Hence, **C** is altered and a new estimate of **S** can be obtained using these modified versions of concentration profiles and so on, until both **C** and **S** obey the constraints and they are stable (i.e., do not change much each iteration), to get a final solution. The method can be turned on its head in that instead of first guessing pure spectra it is also possible to first guess pure concentration profiles, and then following similar steps, usually iteratively, to obtain a correct solution.

Iterative target transform factor analysis (ITTFA) (Vandeginste, Derks and Kateman, 1985) is an alternative approach used to improve these guesses. PCA is performed first on the dataset, and a rotation or transformation matrix **R** is found to relate the scores to the first guesses of the spectra, just as mentioned above. This, in turn, allows the spectra to be predicted. However, the next stage is the difficult part. The predicted spectral matrix $\hat{\mathbf{S}}$ is examined. In the initial stages, this matrix will usually contain elements that are not physically meaningful, for example, negative regions. These negative regions are changed, normally by the simple procedure of substituting the negative numbers by zero, to provide new estimates of the spectra. This, in turn, results in fresh estimates of the elution profiles that can then be corrected for negativity and so on. This iteration continues until the estimated concentration profiles **C** show very little change. The performance of ITTFA depends on the quality of the initial guesses, but can be an effective algorithm. Nonnegativity is the most common condition, but other constraints can be added.

ALS is based on similar principles, but the criterion for convergence is slightly different and looks at the error between the reconstructed data from the product of the estimated concentration profiles and spectra ($\mathbf{C} \cdot \mathbf{S}$) and the true data. In addition, implemented guesses of either the initial spectra or concentration profiles can be used to start the iterations, whereas for ITTFA it is normal to start with the concentration profiles. The ALS procedure is often considered slightly more general and ITTFA was developed very much with analytical applications in spectroscopy coupled to chromatography or another technique, whereas the idea of ALS was borrowed from general statistics. In practice, neither method is radically different, the main distinction being the stopping criteria.

31.2.4.3 Combining Information

Often different types of information can be combined, for example, we may know the order of a reaction and may want to combine known kinetics with multivariate information, rather than treat each type separately.

There are a wide variety of ways to combine both types of information, but most involve using two alternating criteria. Consider knowing the pure spectra of the reactants and products, as well as the starting material. We can then "guess" a value of k (the rate constant), which gives the profiles of each compound in the mixture, often represented by the matrix **C**. We can then use this to "guess" a value of the original data matrix **X** since $\mathbf{X} \approx \mathbf{C} \cdot \mathbf{S}$ or, in other words, the observed reaction spectra are the product of the concentration profiles and pure compound spectra. If our estimated rate constant is accurate, then the estimate of **X** is quite good. However, if this is not very accurate, we need to change our rate constant, change the profiles, and continue until we get a good fit.

This theoretically simple approach takes into account both the "hard" kinetic models used to obtain **C** and the "soft" spectroscopic models used to see how well the estimated spectra equal the observed spectra. Life, though, is never so simple. We may not know the spectra of every compound perfectly, and certainly not in a quantitative way, we may be uncertain of the starting concentrations, and there may be unknown impurities or some side reactions. Hence, many sophisticated elaborations of these approaches have been proposed involving changing several parameters simultaneously including the initial concentrations.

PCA also can be incorporated into these methods via PCR. If we know the scores **T** of the data matrix and guess **C** using our kinetics models, then we can find a transformation matrix $\mathbf{R} = \mathbf{T}^{+} \cdot \mathbf{C}$. Now we can see how well the scores are predicted, using matrix **R** and adjust k (and any other parameters) accordingly and continue in an iterative way. The most common techniques used are ALS and ITTFA, which been already been introduced, although several tailor-made elaborations have been developed specifically for reaction monitoring.

Kinetics knowledge is not the only information that can be obtained about a series of spectra and these models are by no means restricted to kinetics. There are many other constraints that can be incorporated. Simple ones are those where spectra and concentration profiles are positive, but more elaborate ones can be incorporated. For example, concentration profiles with time are likely to be smooth and unimodal; the profile of the starting materials will decrease with time and for all other compounds it can increase and then sometimes decrease. These can all be incorporated into algorithms, allowing the models to contain as many constraints as required, although the convergence of the algorithms can be quite tricky if a large number of factors are taken into consideration.

31.3
Calibration

Calibration plays an important role in spectroscopy. It involves using a spectral response to predict the concentrations or characteristics of individual components. In chemometrics or analytical chemistry, the calibration problem is often formulated as an *inverse* problem. For example, it is often important to be able to predict a concentration from a spectrum. In classical physical chemistry, often the problem is reversed; for example, can we predict a rate constant from a temperature and thus understand the causal relationship between temperature and observed kinetics? In analytical chemistry, the rate constant could be used as a type of thermometer to predict a temperature, rather than vice versa. In addition, there is a notation difference between traditional analytical chemistry and chemometrics, as illustrated in Figure 31.5, where the x variable, in chemometrics, normally represents spectroscopic data, and the y variable represents a predicted property, often the concentration of a compound in a mixture.

To clarify this, we use c as the property to be predicted and x as the data measured to predict this property, usually spectra; hence, the aim of calibration is to predict c from x. The spectroscopy data is called the *x block* and the concentration (or other properties) that is to be deduced from the spectra is called the *c block*.

Note that although the aim of calibration is to predict, for example, the concentrations of compounds in a spectroscopic mixture, there are usually two stages. The first one is calibration or modeling, which uses a set of samples of known composition to form a relationship between the spectral characteristics and the composition or other properties of the mixture. The second one is called *prediction* where the calibration model is used to predict the composition of unknown samples. A recent review discusses calibration in greater detail (Brereton, 2000) and the text by Martens and Næs (Martens and Næs, 1989) is a classic in the subject.

31.3.1
Univariate Calibration

The most traditional, but still very widely used, method of calibration in chemistry is univariate, involving calibrating a

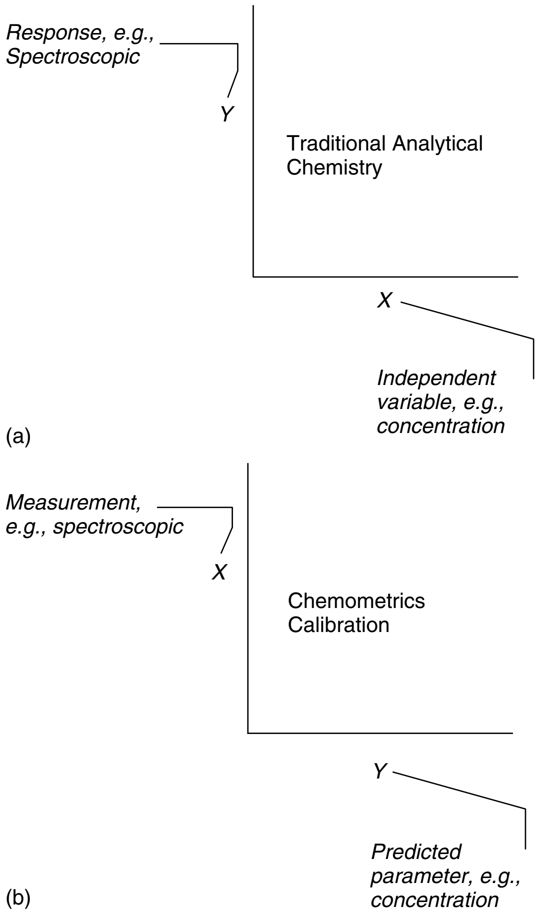
Response, e.g., Spectroscopic
Y
Traditional Analytical Chemistry
X
Independent variable, e.g., concentration
(a)
Measurement, e.g., spectroscopic
X
Chemometrics Calibration
Y
Predicted parameter, e.g., concentration
(b)

Fig. 31.5 Difference in notation in analytical chemistry (a) and chemometrics (b).

single variable (e.g., a spectroscopic intensity) to another variable (e.g., a concentration). The literature goes back a century; however, with the ready availability of matrix-based computer packages ranging from spreadsheets to more sophisticated programming environments such as Matlab, we can approach this area in new ways.

31.3.1.1 **Classical Calibration**

Most chemists are familiar with this approach. An equation of the form

$$x = bc \tag{31.12}$$

is obtained relating, for example, a spectroscopic intensity to a concentration. Sometimes an intercept term is introduced:

$$x = b_0 + b_1 c \tag{31.13}$$

and it is also possible to include squared and higher order terms.

There are several difficulties with classical methods. First, the aim of calibration is to predict c from x and not vice versa: the classical method involves obtaining a model x from c; it may be what a physical chemist wants but not necessarily what an analytical chemist wants to do. The second, and more serious, difficulty is that the greatest source of error is generally in the preparation of sample; hence, the measurement of a concentration (even when determining a calibration model) is likely to be less certain than the measurement of spectral intensity. Twenty years ago, many instruments produced less reproducible results than now, but the significant increase in instrumental performance has not been accompanied by an equally significant improvement in the quality of measurement cylinders, pipettes, and balances.

31.3.1.2 Inverse Calibration

An alternative approach is to assume that the errors are in the c block. This means fitting an equation of the form

$$c = ax \tag{31.14}$$

or

$$c = a_0 + a_1 x \tag{31.15}$$

In terms of both error analysis and scientific motivation, this is more rational than by employing classical calibration, but most chemists still use classical methods even though they are often less appropriate.

If, however, a calibration dataset is well behaved and does not have any significant outliers, straight lines obtained using both forms of calibration should be roughly similar, as illustrated in Figure 31.6. Sometimes there are outlying points, which may come from atypical or erroneous samples, and these can have considerable influence on the calibration lines and there are many technical ways of handling these outliers, which are often said to have high leverage: generally the two types of best fit straight line are influenced in different ways, and this discrepancy can be eliminated by removing any outliers.

31.3.1.3 Determining the Relationship between Variables

Instead of deriving quite complex equations, it is easier to obtain best fit

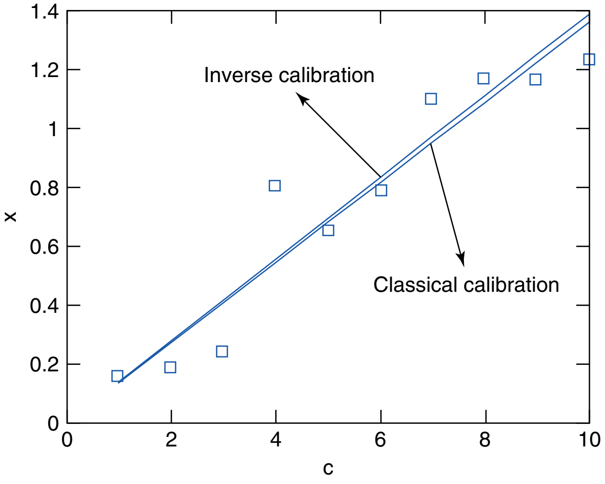

Fig. 31.6 Classical and inverse calibration lines.

calibration parameters using simple vector arithmetic. For inverse calibration, we can use the left pseudoinverse; hence,

$$a \approx (\mathbf{x}' \cdot \mathbf{x})^{-1} \cdot \mathbf{x}' \cdot \mathbf{c} = \mathbf{x}^{+} \cdot \mathbf{c} \qquad (31.16)$$

where **x** is a vector of spectroscopic responses (often peak heights at a single wavelengths) and **c** is the corresponding concentrations. Note that $(\mathbf{x}' \cdot \mathbf{x})^{-1}$ is actually a scalar or a single number in this case, equal to the inverse of the sum of squares of the elements of **x**.

Sometimes it is desirable to include an intercept (or baseline term) of the form

$$c = a_0 + a_1 x \qquad (31.17)$$

Now the power of matrices and vectors becomes evident. Instead of using a single column vector for **x**, we use a matrix whose first column consists of ones and second of the experimentally determined x values, and **a** becomes a row vector of two elements, each element being one coefficient in the equation. The equation can now be expressed as

$$\mathbf{c} = \mathbf{X} \cdot \mathbf{a} \qquad (31.18)$$

so that

$$\mathbf{a} \approx \mathbf{X}^{+} \cdot \mathbf{c} \qquad (31.19)$$

31.3.2
Multiple Linear Regression

MLR is one of the commonest techniques for calibration and regression both in chemistry and statistics.

In the context of the spectroscopy of mixtures, the principle is that a series of mixture spectra can be characterized by

- the spectrum of each individual component;
- the concentrations of these components; and
- noise (or experimental error).

Mathematically, this information can be expressed in a matrix format:

$$\mathbf{X} = \mathbf{C} \cdot \mathbf{S} + \mathbf{E} \qquad (31.20)$$

This equation has already been discussed in Section 2.2. The **X** matrix might, for example, consist of a series of uv spectra of a reaction of the form

$$\mathrm{A} \rightarrow \mathrm{B} + \mathrm{C} \qquad (31.21)$$

consisting of three compounds whose concentrations we wish to monitor as a function of time. Each spectrum will be a row of the matrix **X** and each corresponding row of **C** will consist of concentrations of each component in the reaction mixture at a specific sampling time.

To study the reaction, it is usually necessary to determine **C**. There are many ways of using MLR to obtain this information. The simplest is not really calibration, but involves knowing the spectra of a unit concentration of each pure component, **S**. Using the pseudoinverse, we can write

$$\mathbf{C} \approx \mathbf{X} \cdot \mathbf{S}^{+} \qquad (31.22)$$

However, this supposes that we can actually obtain the spectra of all components in a mixture, that we know all the components and have isolated them, and also that the conditions in the reaction mixture are similar to those of the pure components. An alternative is to take some mixtures, quench them (where necessary), and then measure the concentrations of the components using an off-line method such as chromatography, where each is separable and peak areas can easily be related to the concentration as illustrated in Figure 31.7. We can then determine and estimate $\hat{\mathbf{S}}^{+}$ from this training or calibration set, where the notation "^" means predicted rather than observed, as we know $\mathbf{C}$ from the off-line chromatographic peak areas. This can be used to set up a calibration model and in the future predict the concentrations of compounds from their spectra in the absence of chromatography. If we are satisfied with the calibration model, we can then predict the concentrations in the reaction mixture

$$\hat{\mathbf{C}} = \mathbf{X}_{\text{reaction}} \cdot \hat{\mathbf{S}}^{+} \qquad (31.23)$$

and thus obtain the profiles of each compound with time.

MLR can be extended to include nonlinear such as with squared terms, which can be useful under some circumstances, for example, if there are nonlinearities due to overloaded spectra (which may occur if concentrations in a reaction mixture, for example, are outside the region where the Beer–Lambert law is obeyed). Extra columns are simply added to the $\mathbf{X}$ matrix equal to squares of the intensities. Another common modification is weighted regression, where certain wavelengths have more significance than others.

However, MLR has the disadvantage that all significant compounds in a mixture should be known. If we have information on two out of five compounds in a series of spectra, then the two predicted spectra will also contain features of the spectra of the remaining three compounds, distributed between the two known components (dependent on the nature of the calibration set) and the concentration estimates will contain large and fairly unpredictable errors. It is sometimes possible to overcome this problem if there are selective wavelengths where only two compounds absorb, so that the $\mathbf{X}$ matrix consists of only certain selective regions of the spectrum; however, there are alternative approaches, discussed later, where good models can be obtained using information when only a proportion of the compounds in a mixture can be characterized.

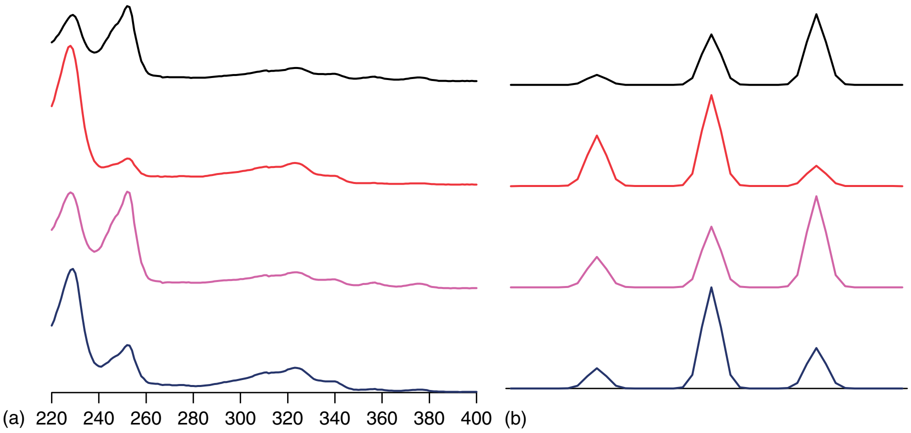

Fig. 31.7 Calibrating spectra (a) to chromatograms (b).

31.3.3
Principal Components Regression

Some consider PCR as a form of factor analysis. Although the two concepts are interchangeable according to some chemometricians, especially spectroscopists, it is important to recognize that there are widespread differences in the usage of the term *factor analysis* by different authors, groups, and software vendors.

PCR has a major role in the spectroscopy of mixtures. It is most useful where only some compounds can be identified in the mixture. For example, we may be interested in the concentration of a carcinogenic chemical in a sample of seawater, but have no direct knowledge or interest in the identities and concentrations all the other chemicals. Can we determine this by spectroscopy of mixtures? MLR works best when all the significant components are identified.

The first step, as always, is to obtain a series of spectra (the calibration or training set), for example, the UV spectra of 30 samples recorded at 200 wavelengths, represented by a 30 × 200 matrix **X**. The next step is to perform PCA as discussed in Section 31.2.2 to obtain scores and loadings matrices, **T** and **P**. A crucial decision is the determination of the number of PCs that should be retained (Section 31.2.3). Ideally, this should equal the number of compounds in a mixture. For example, if there are seven compounds we should retain the first seven PCs, so that the matrix **T** is of dimensions 30 × 7 and **P** of dimensions 7 × 200, in this context. However, in real-world situations, this information is rarely known exactly. In many cases, the number of PCs needed to model the data is very different from the number of real components in a series of mixtures. This is because of correlations between compound concentrations, spectral similarities, and experimental design that often reduce the number of components, and noise and instrumental features such as baseline problems that may increase the number of components.

The next step is variously called *regression, transformation,* or *rotation.* A common approach is to find a relationship between the scores and the true concentrations of a compound of the form

$$\mathbf{c} = \mathbf{T} \cdot \mathbf{r} \tag{31.24}$$

where **c** corresponds to the known concentrations in the mixture spectra and **r** is the column vector whose length equals the number of PCs (seven in this example). This is represented diagrammatically in Figure 31.4. Equation (31.24) can also be expressed as follows:

$$c = t_1 r_1 + t_2 r_2 + \cdots + t_7 r_7 \tag{31.25}$$

if there are seven PCs, where t_1 is the score of the first PC, and r_1 is its corresponding coefficient obtained using regression.

Knowing **c** and **T**, it is not very difficult to estimate **r** using the pseudoinverse by

$$\mathbf{r} = \mathbf{T}^{+} \cdot \mathbf{c} \tag{31.26}$$

This permits the concentration of an individual compound to be estimated by knowing the scores.

If there are several compounds of interest, it is a simple procedure to expand "*c*" from a vector to a matrix, each column corresponding to a compound, so that

$$\mathbf{C} = \mathbf{T} \cdot \mathbf{R} \tag{31.27}$$

As usual, we can determine **R** by

$$\mathbf{R} = \mathbf{T}^{+} \cdot \mathbf{C} \tag{31.28}$$

The extreme case is when the number of PCs is exactly equal to the number of known compounds in the mixture, in which case **R** becomes a square matrix.

The quality of prediction is then demonstrated on a test set (see Section 31.3.5) and finally the method is used on real data of unknowns. It is also useful to be able to preprocess the data under certain circumstances before PCA. This process often involves mean centering or standardizing the spectral matrix before PCA, rather than using the raw spectroscopic data, for example

$$x_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j} \tag{31.29}$$

where x_{ij} is the spectroscopic intensity in spectrum i and wavelength j, $\overline{x}_j$ is the mean intensity over all spectra in the calibration set at wavelength j, and s_j is the corresponding standard deviation: this is valuable if it is desired to weight all wavelengths to be of equal significance, which is often useful in techniques such as NIR and UV/vis where PCR represents a significant improvement over MLR when only partial information is available about the number of components in a mixture, as one can have as many or as few components in the regression model as desired, so long as the number of PCs is determined well.

31.3.4
Partial Least Squares

PLS has a long history in chemometrics. This technique was originally proposed by the Swedish statistician, Herman Wold, whose interests were primarily in economic forecasting. In the 1960s his son Svante Wold, together with a number of Scandinavian scientists, advocated its use in chemistry. Possibly no technique in chemometrics is so mired in controversy, with some groups advocating PLS for almost everything (including classification) with numerous extensions over the last two decades, with other groups agreeing that PLS is useful, but one of very many techniques in the toolbox of the chemometrician. Probably, one confusion is between the availability of good software (there is currently some excellent public domain software for PLS) and the desirability (independently of software) for using PLS. In many areas of calibration, PCR serves almost the same purpose; however, despite this, PLS has been the subject of more extensions and articles than probably any other chemometrics method, and there is a large hard-core of believers who solve most problems (usually quite successfully) by PLS. An important early review of the method was published by Geladi and Kowalski in 1986 (Geladi and Kowalski, 1986).

One principle guiding PLS is that modeling the c block of data is as important as modeling the experimental x information. A drawback of PCR is that the PCs are calculated exclusively on the x block and do not take into account the c block. In PLS, components are obtained using both x and c data simultaneously. Statisticians like to think about modeling covariance as opposed to variance. Most conventional least squares approaches involve finding a variable that maximizes the value of x^2 or the square of the modeled experimental data or variance. PLS finds a variable that maximizes xc or the product of the modeled experimental data with the concentrations, often called the *covariance*. In physical terms, PLS assumes that there are errors in both types of information, which are of equal importance. This makes some sense: the concentrations used in calibration are subject to error (e.g., dilution and weighing) just as much as the spectra or chromatograms. MLR and PCR as commonly applied in chemistry assume that all the errors are in the spectroscopic data and that the concentrations in the calibration set are known exactly.

Often PLS is presented in the form of two equations. There are a number of ways of expressing these, a common one being

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P} + \mathbf{E} \tag{31.30}$$

$$\mathbf{c} = \mathbf{T} \cdot \mathbf{q} + \mathbf{f} \tag{31.31}$$

as illustrated in Figure 31.8. Equation (31.30) appears to be similar to that of PCA, but the scores matrix also models the concentrations, and the vector $\mathbf{q}$ has some analogy to a loadings vector. The matrix $\mathbf{T}$ is common to both the equations. $\mathbf{E}$ is an error matrix for the x block and $\mathbf{f}$ an error vector for the c block. The scores are orthogonal, but the loadings ($\mathbf{P}$) are not. There are various algorithms for PLS, and in many the loadings are not normalized. PLS scores and loadings differ numerically from PCA scores and loadings and are dependent both on experimental measurements (e.g., spectra) and the concentrations or properties of the spectral matrices. In PCA, the scores and loadings depend only on the spectra or x block. In PLS, the product $\mathbf{T} \cdot \mathbf{P}$ does not provide the best fit estimate of the x block. The size of each successive component (as assessed by the sum of squares of each successive vector in $\mathbf{T}$) does not always decline; later vectors may be larger than earlier ones, unlike in PCA. In addition to matrices $\mathbf{T}$ and $\mathbf{P}$ in PLS, there is another matrix called a *weights matrix* $\mathbf{W}$ which is often used, which has analogies to the $\mathbf{P}$ matrix and is a consequence of the algorithm. Some like to use $\mathbf{W}$ instead of or in addition to $\mathbf{P}$. In the PLS literature, it is quite important to read quite closely each paper and determine precisely which algorithm is being employed, there being a much wider range of terminology than in PCA.

The quality of the models can be determined by the size of the errors,

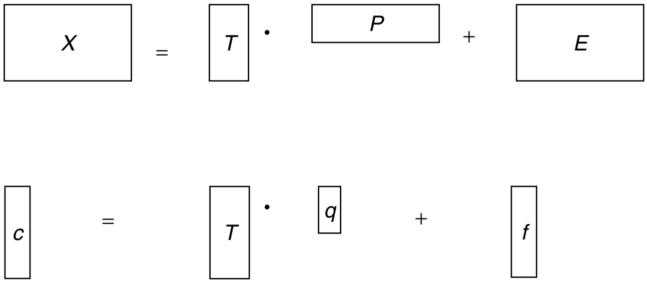

Fig. 31.8 PLS principles.

normally the sum of squares of **E** and **f** is calculated. The number of significant PLS components can be estimated based on these errors, often using cross validation or the bootstrap, although it can be done on the training set (often called *autoprediction*) as discussed in Section 31.3.5. It is important to recognize that there are a large number of criteria for determining the number of significant components that can cause major disagreements between packages and different algorithms.

There are several extensions to PLS.

The standard algorithm described above is sometimes called *PLS1*. One extension, first proposed in the 1980s, is *PLS2* (Hoskuldsson, 1988). In this case, the concentration (or c) block consists of more than one variable, an example being a case in which there are three components in a mixture and hence **C** becomes a matrix with three columns, and Eq. (31.31) changes to

$$\mathbf{C} = \mathbf{T} \cdot \mathbf{Q} + \mathbf{F} \tag{31.32}$$

In analytical spectroscopy, PLS2 is not very popular and often gives worse results than PLS1. However, if there are interactions between variables that correspond to the columns of **C**, this method can have advantages. In most cases in analytical chemistry, however, mixture spectra or chromatograms are linearly additive and independent. An original advantage of PLS2 was also that it was faster in performing several PLS1 calculations, but with modern computing power this is no longer serious.

Another common extension is to include nonlinear (in most case squared) terms in the PLS equation. This is very similar in effect to extending MLR models by including extra terms. Often, however, if a problem is highly nonlinear, completely different approaches such as neural networks may be more appropriate for calibration.

PLS has advantages over MLR in that only some of the components need to be known and over PCR in that the errors can be in both the x and c blocks, and is particularly good when the aim is to determine a statistical or bulk measurement, for example, the amount of wheat in protein. In Section 31.3.5, validation of the models is discussed in more detail. However, if all the components in a mixture are known and are linearly additive, more conventional approaches such as MLR may perform equally well or even better.

31.3.5
Model Optimization and Validation

One of the most important problems in the area of multivariate calibration is determining its effectiveness. The main aim of the model is to use a mathematical function of experimental data to predict a parameter such as concentration. In PCR and PLS, we have to choose the number of components to be retained in the model, which represent one of the dimensions (A) of the scores and loadings matrices. The size of these matrices influences the quality of predictions.

There are two objectives of these methods. The first is to optimize the model and thus determine the optimum number of components (in PLS and PCR). The second is to have an independent assessment about its performance. Often these two aims are mixed up and many people in the area of calibration just cite the calibration and prediction error, the latter being obtained while optimizing the model. We have discussed several approaches in the context of PCA (Section 31.2.3) and briefly introduce this in the context of calibration.

Autoprediction involves looking at the residual, often root mean square, error of calibration. This can be calculated

$$RMSEC = \frac{\sum_{i=1}^{I}(c_i - \hat{c}_i)^2}{D} \qquad (31.33)$$

for the c block, where D is defined as the number of degrees of freedom, which is usually $I-A$, where A is the number of components, I is the number of samples in the calibration or training set, and $\hat{c}_i$ is the predicted value for the tth sample. If the columns are centered, $D = I - A - 1$. This usually decreases the more the variables but the number of components that result in a percentage threshold for the residual error such as 5% could be established for the number of significant PCs. There are variants on this including using the x block, in the literature.

For cross validation (Wold, 1978), similar equations can be established except that the divisor is I, the number of samples rather than the number of degrees of freedom. The root mean square of cross validation (RMSECV) should reach a minimum when the correct number of components is determined. Similar types of equations can be obtained using the bootstrap (Efron and Tibshirani, 1993), likewise searching for a minimum. In both the cases, sometimes there are problems with flat minima or noisy minima that do not have an obvious optimum, and various elaborate rules are available for choosing the number of components to be employed in the model. Both these errors (from cross validation and the bootstrap) should be larger than the autopredictive error.

An external test set is often employed to determine the model performance. This involves samples that have not been used for optimization and have no influence on the number of components in the model, which gives the root mean square error of prediction (RMSEP). Ideally the RMSEP and RMSECV should be calculated on completely different data as one is about using an independent test set to validate the model and the others to optimize the model. Usually $RMSEC < RMSECV < RMSEP$, although there can be exceptions to the rule in unusual situations.

There are several other elaborations to optimizing and validating the models, according to the author and software package, and it is usually important to understand how this is done: most packages are based on different philosophies. However, the above-mentioned methods are some of the most common.

31.4
Exploratory Methods

The approaches in Sections 31.2 and 31.3 aim primarily at determining characteristics of components primarily in a mixture, for example, the concentrations of individual compounds in a series of spectra or the spectra of these components. However, there are other reasons for employing chemometrics in spectroscopy.

Exploratory data analysis involves determining whether spectra form groups or are related. For example, it might be possible to determine the MIR spectra of extracts of the urine of patients with and without a disease and to see if these cluster differently and thus explore whether there are differences and why.

31.4.1
Principal Components Analysis

The most widespread methods are based on PCA. We have already described the theory in Section 31.2.2. In the context of calibration or resolution, the numerical values of the PCs are used to determine physical properties of components in a mixture.

In exploratory data analysis, the main aim is to determine whether spectra fall into groups and why. Most of the output is in the form of graphs. We illustrate this in an example from atomic spectroscopy in which the concentrations of 11 elements obtained using atomic spectroscopy are measured for 58 types of pottery (Bruno *et al.*, 2000) from two origins (or groups). The aim is to determine whether the atomic spectroscopic data can be used to distinguish the pottery and if so which elements are most characteristic of each group.

The first step is normally preprocessing, which involves putting the data on a sensible scale. There are several common procedures.

1. *Using the raw data without any preprocessing*: If all the spectra are roughly on the same scale and the amount of sample in each analysis is similar, this is acceptable. However, usually some additional steps are necessary.
2. *Row scaling*: If the amount of sample to be analyzed is difficult to control, then summing the rows to a constant total by ${}^{\text{nor}}x_{ij} = \frac{x_{ij}}{\sum_{j=1}^{J} x_{ij}}$ is a common procedure.
3. *Mean centering*: Often the columns (variables) are mean-centered. There is no inherent requirement to do this, but the majority of statistical packages perform this automatically as PCA often looks at variation around a mean.
4. *Standardizing*: If all the variables are on quite different scales, then this procedure is quite important. For example, some peaks may be small, but their variation is significant relative to large peaks. This is done by

$$ {}^{stn}x_{ij} = \frac{x_{ij} - \overline{x}_j}{\sqrt{\sum_{i=1}^{I} (x_{ij} - \overline{x}_j)^2 / I}} \qquad (31.34) $$

Note that mean centering and standardizing must be performed after any row scaling. The disadvantage with standardizing is that if there are noisy regions of a spectrum where there is no real information, these will degrade the model. This can be overcome in various ways, for example, by variable selection just keeping useful regions of the spectrum or by an offset on the bottom of the equation.

For the example in this section, we only show the results of standardizing the data.

The scores plot involves plotting the scores of one PC against an other (or in three dimensions three PCs against each other), or in algebraic terms, different columns of **T** against each other. These are represented in Figure 31.9. It can be seen from inspection that there are likely to be groupings. When plotting the scores there are lots of options, for example, any set of PCs can be chosen. Usually, the most significant ones (with the largest eigenvalues) are chosen, but

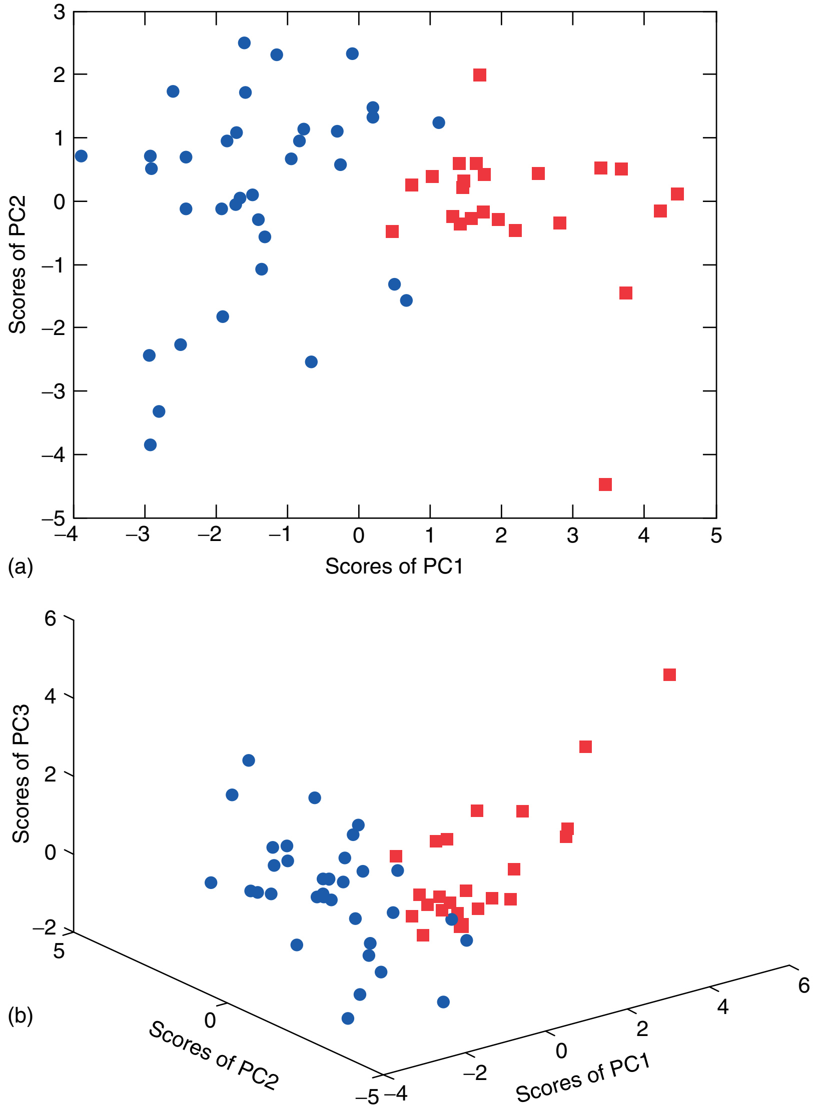

Fig. 31.9 A PC scores plots in which there are two groups in a dataset, in both two (a) and three dimensions (b).

sometimes smaller PCs are important for discriminating between samples, as the largest ones just provide overall information about the data. Often the most suitable PCs can be selected by graphical inspection of various plots.

Parallel to the scores plots are the loadings plots. These involve plotting the rows of **P** against each other, exactly analogous to the scores plots. A loadings plot is illustrated in Figure 31.10 for our example. We can see several features.

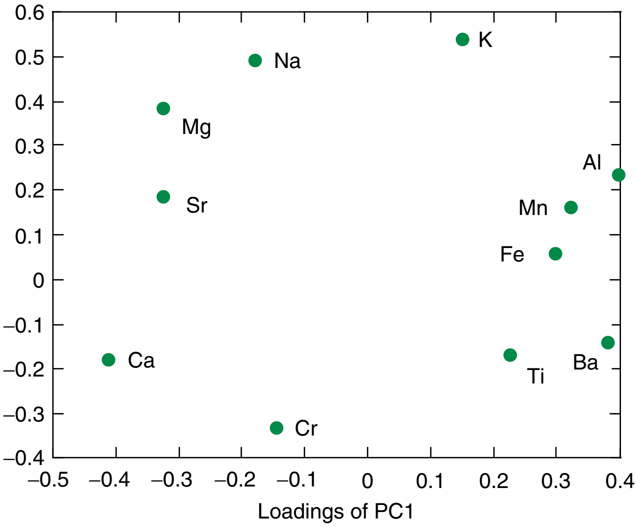

Fig. 31.10 Loadings of the elements corresponding to the example in Section 31.4.1.

The first is that some elements are clustered close together, meaning that they behave in a similar way, for example, Na and Mg. This basically means that if Na is found in high quantities so is Mg, and probably they come from the same source. The second is that one can compare the scores plot, for example, elements on the right-hand side are associated with the group represented by red squares; hence, Al and Mn are likely to be markers for this group of samples, that is, observed in high abundance.

The loadings can be of different forms, such as wavelengths in a spectrum, and would indicate the spectral features associated with specific groups of samples, for example, there may be groups of patients with and without a disease, and if these are separated in a scores plot, the corresponding loadings plot will tell whether there are spectral features diagnostic of specific groups of patients. When comparing scores and loadings plots, it is necessary to ensure that the same PCs are used in each case and that the preprocessing is the same.

PCA is not only used to see groupings but could also be used to see whether there are similarities between samples without any firm groupings, for example, whether there are outliers or whether samples taken as a function of time change with time. There are numerous graphical methods for presenting information about PCs of which the types of graphs, in this section, are the most popular.

31.4.2
Cluster Analysis

Exploratory data analysis, such as PCA, is used primarily to determine general relationships between data (Massart and Kaufmann, 1983; Everitt, Landau and Leese, 2001). Sometimes more complex questions need to be answered, such as, which samples are more similar to each other? Cluster analysis is a well-established approach that was developed primarily by biologists to determine similarities between organisms. Numerical taxonomy emerged from a desire to determine relationships between different species, for example, genera, families, and phyla. However, the spectroscopists also wish to relate samples in a similar manner. Which samples measured by spectroscopy are similar? Cluster analysis involves using a number of methods to group different samples (or objects) using chemical measurements such as spectroscopic intensities or peaks.

The first step is to determine the similarity between spectra. Often the spectra may need to be preprocessed before these distances are measured, for example, standardized, and the results depend on how the data are treated before cluster analysis, just as in the case of PCA.

A number of common numerical measures of similarity are available.

1. *Correlation coefficient between samples*: A correlation coefficient of 1 implies that the spectra have identical characteristics. This value will vary between −1 and +1.
2. *Cosine distance*: This measure involves determining one minus the cosine between the two spectra (treated as vectors).
3. *Euclidean distance*: The distance between spectra k and l is defined as

$$d_{kl} = \sqrt{\sum_{j=1}^{J}(x_{kj} - x_{lj})^2} \qquad (31.35)$$

   where there are j measurements, and x_{ij} is the jth measurement on the sample i, for example, x_{23} is the third wavelength of the second spectrum in a series. The smaller this value, the more similar the spectra; thus, this distance measure works in an opposite manner to the correlation coefficient.
4. *Manhattan distance*: This is defined slightly differently to the Euclidean distance and is given as

$$d_{kl} = \sum_{j=1}^{J} |x_{kj} - x_{lj}| \qquad (31.36)$$

Once a distance measure has been chosen, a similarity (or dissimilarity) matrix can be drawn up. This consists of a symmetric matrix whose dimensions are equal to the number of samples. In a dataset of I samples, the matrix will have dimensions $I \times I$ and is illustrated in Figure 31.11.

The next step is to link the objects. The most common approach is called *agglomerative* clustering, whereby single objects are gradually connected to each other in groups.

- From the raw data, find the two most similar objects.
- Next form a "group" consisting of the two most similar objects. The original I objects are now reduced to $I - 1$ groups.
- It is difficult to decide how to represent this new grouping. There are quite a few approaches, but it is common to change the data matrix from one consisting of six rows to a new one of five rows, four corresponding to original objects and one to the new group. The numerical similarity values between this new group and the remaining objects have

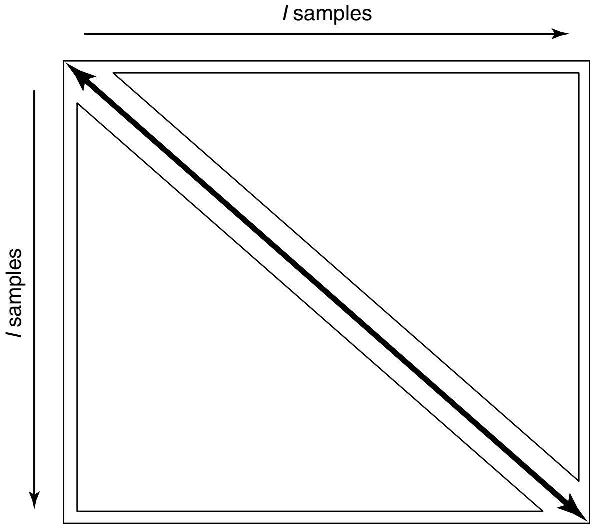

Fig. 31.11 (Dis)similarity matrix produced during cluster analysis. The diagonal values represent maximum similarity (or minimum dissimilarity) and the matrix is symmetrical.

to be recalculated. There are three principal ways of doing this.

1. *Nearest neighbor*: The similarity of the new group from all other groups is given by the *highest* similarity of either of the original objects to each other object.
2. *Farthest neighbor*: This is the opposite to nearest neighbor, and the *lowest* similarity is used.
3. *Average linkage*: The average similarity between two groups is used. There are, in fact, two different ways of doing this, differing according to the size of each group. For the raw data, we compare a group consisting of two members to one consisting of a single object; hence, the calculation is easy. However, in later steps we may want to combine a group consisting of two members with one of four members, and the two methods for calculating similarity indices differ if the number of objects in each group is also different.

The next steps consist of continuing to group the data just as above, until only one group, consisting of all the original objects, remains. It is normal to then determine at what similarity measure each object joined a larger group, and thus which objects resemble each other most. Often the result of hierarchical clustering is presented in a graphical form called a *dendrogram*: note that many biologists call this a *phylogram* and it differs from a cladogram where the size of all the branches are the same. The objects are organized in a row, according to their similarities: the vertical axis represents the similarity measure at which each successive object joins a group. Using nearest neighbor linkage and cosine distance, the dendrogram for the data of Section 31.4.1 is presented

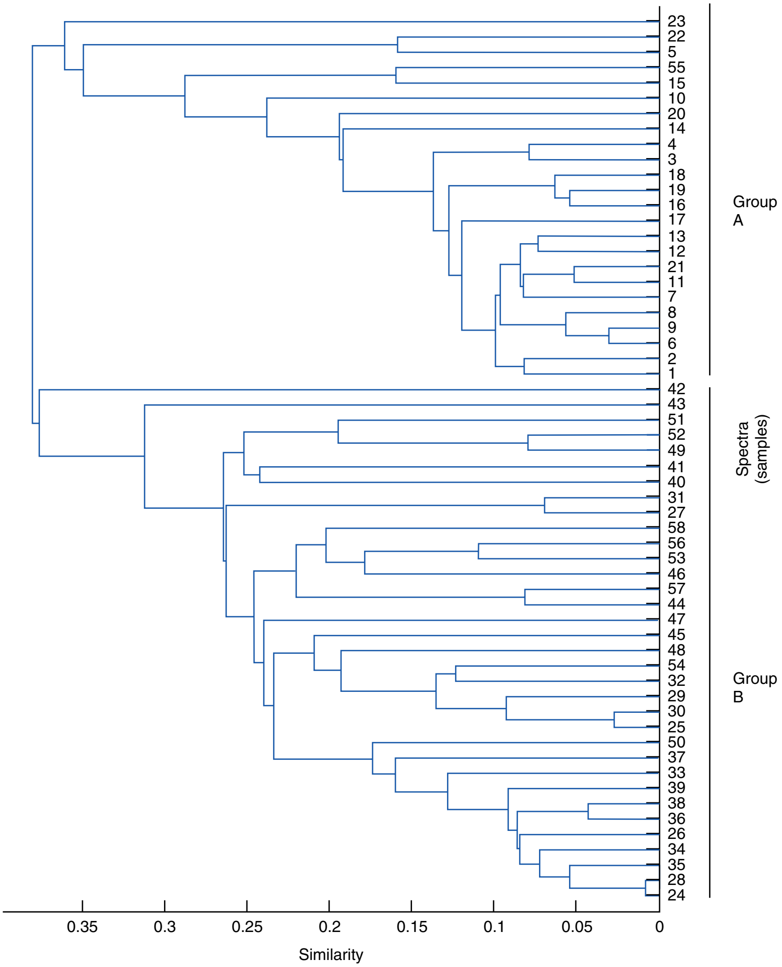

Fig. 31.12 Dendrogram produced by performing cluster analysis on the dataset of Section 31.4.1.

in Figure 31.12. Note that the first group (represented by squares in Figure 31.9) is distinct from the second group. The dendrogram also indicates the samples that are more similar to each other.

Dendrograms do have limitations. There are many clustering methods and each can produce different results. The samples are organized in a linear manner and sometimes the relationships are more complicated and better represented in two or three dimensions. The order of the samples can be variable as it is possible to swap some of the main branches around; hence, parts of the diagram are reflected. However, cluster analysis as described in this section is a well-known and established approach for simplifying and visualizing multivariate data and is often employed in chemometrics.

31.5
Supervised Methods

Classification (often called *supervised pattern recognition*) is at the heart of chemometrics. Can a mid-infrared (MIR) spectrum be used to determine whether a compound is a ketone or an ester? Can the nuclear magnetic resonance spectrum of a tissue sample be used to determine whether a patient is cancerous or not? Can we record the NIR spectrum of an orange juice and decide its origin? Is it possible to monitor a manufacturing process and decide whether the product is acceptable or not? Supervised pattern recognition is used to assign samples to a number of groups (or classes). It differs from exploratory data analysis where, although the relationship between samples is important, there are no predefined groups. There are a very large number of methods for classification and this section summarizes only a few of the most widespread used in chemometrics and spectroscopy; for more comprehensive reviews of methods, see some of the texts referenced in the introduction. However, most common applications in analytical spectroscopy are based on a relatively small number of approaches.

31.5.1
Developing and Assessing Models

Although there are numerous algorithms in the literature, chemometricians have developed a common strategy for developing classification models.

31.5.1.1 Modeling the Training Set

Normally, the first step is to produce a mathematical model between some measurements (e.g., spectra) on a series of objects and their known groups. These objects are called a *training set*. For example, a training set might consist of the NIR spectra of 150 orange juices, 50 known to be from Cyprus, 50 known to be from Spain, and 50 known to be adulterated. Can we produce a mathematical equation that predicts which class an orange juice belongs to from its spectrum? There are several approaches for classification, which are introduced in Sections 31.5.2 – 31.5.4, but the common characteristic is that a model is developed between the spectra and the class each belongs to, which predicts whether a spectrum can be assigned to one class or another.

Once this is done, it is usual to determine how well the model predicts the groups. Table 31.1, often called a

Tab. 31.1 Confusion matrix for classification using a training set.

	Known			Overall
Predicted	**Cyprus**	**Spain**	**Adulterated**	
Cyprus	40	6	6	
Spain	7	44	8	
Adulterated	3	0	36	
Correct	40	44	36	**120**
%CC	80	88	72	**80**

confusion matrix or *contingency table*, illustrates a possible scenario. The higher the relative values down the diagonals, the better the class membership predicted. In this case, of the 150 spectra, 120 are correctly classified. The off-diagonal elements show which class is confused (or mistaken) for another Some classes are modeled better than others, 88% of the Spanish orange juices is correctly classified, but only 72% of the adulterated orange juices. A parameter percentage correctly classified (%CC) can be calculated and is 80% overall. There appears some risk of making a mistake, but the aim of a spectroscopic technique might be to perform screening, and there is a good chance that suspect orange juices (e.g., those adulterated) would be detected, which could then be subject to further detailed analysis. Chemometrics combined with spectroscopy acts like a "sniffer dog" in a customs checkpoint trying to detect drugs. The dog may miss some cases and may even get excited when there are no drugs, but there will be a good chance that the dog is correct. Proof, however, comes only when the suitcase is opened and usually spectroscopy is applied as a cheap and first method, which sifts samples, allowing a smaller number to be examined often by other methods in detail.

Sometimes the number of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) can be computed as an alternative measure of the quality of a classification technique. However, this can be done only for what is called a *one class classifier*, that is, one class against the rest. However, we are interested in knowing whether an orange juice is adulterated or not. Table 31.2 illustrates the way in which such a contingency table can be developed. A "positive" is a spectrum arising from an adulterated orange juice that is known to be adulterated. The percentage of TPs is the percentage of orange juices that are actually positive and the mathematical model predicts to be positive. In the example, the %TP is 72% although the %TN is 97%; hence, the test is quite good at determining whether an orange juice is not adulterated but might miss some adulterated samples.

Tab. 31.2 Stages in developing a 2 × 2 contingency table for determining adulteration.

	Adulterated	**Not adulterated**
Adulterated	36	3
Not adulterated	14	97

	Adulterated	**Not adulterated**
Adulterated	72%	3%
Not adulterated	28%	97%

	Adulterated	**Not adulterated**
Adulterated	TP	FP
Not adulterated	FN	TN

31.5.1.2 Test Sets, Cross Validation, and the Bootstrap

However, it is normal that the training set results in good predictions, but this does not necessarily mean that the method can safely be used in practice. A recommended second step is to test the quality of predictions often using a *test* set. This is a series of samples that has been left out of the original calculations, and is a bit like a "blind test." These samples are assumed to be unknowns at first. Usually the test set predictions are lower than those for the training set, and these are the best indicators as to how well the method can be used in the field. It is particularly important to understand that in spectroscopy it is often possible to obtain several hundreds or thousands of variables per sample. This is a unique property of chemometrics compared to conventional statistics; thus, for the spectroscopist often the sample-to-variable ratio is low, that is, there are many more wavelengths or frequencies than spectra in most typical datasets. In conventional applications of statistics, such as in biology, it is difficult to obtain many measurements per sample. Many classical multivariate methods first reported by biologists were first demonstrated by their application to data obtained by RA Fisher on the iris plant, involving only four variables (Fisher, 1936). However, it can be shown that if the number of variables strongly exceeds the number of samples, it is often possible to get high %CC even starting from completely random datasets (Brereton, 2006); thus, to determine the true effectiveness of the model, it is important to ensure that the model is carefully validated.

Using a test set to determine the quality of predictions is a form of *validation*. The test set could be obtained using samples that are left out of the original dataset. Alternatively, they could have been supplied by an independent laboratory. With modern computers, sometimes it is possible to iteratively generate test sets. For example, if 400 samples are recorded in the original dataset, 100 can be left out randomly at each iteration for testing, and the procedure is repeated many times: computing power increases every year; thus, methods described and implemented 10 years ago can often be improved significantly. What might have taken several hours in 1995 could take a few minutes in 2009 and thus can be repeated many times. With the growth of distributing, computing, and high-performance clusters, we are likely to see significant developments in this direction.

An alternative approach is *cross validation* (Wold, 1978). Only a single training set is required, but usually one (or a group) of the objects is removed at a time, and a model is determined on the remaining samples. Then the prediction on the object (or set of objects) left out is tested. The most common approach is leave one out (LOO) cross validation where one sample is left out at a time. This procedure is repeated until all objects have been left out and is discussed in Section 31.2.3 in the context of PCA. However, the aim in pattern recognition is to determine how well the samples are classified rather than the residual error in modeling the dataset.

Finally, mention should be made of the third alternative called the *bootstrap* (Efron and Tibshirani, 1993). This is a halfway house between cross validation and having a single independent test set and involves iteratively producing several internal test sets: not just removing samples once as in cross validation but also not just having a single test set. A set of samples may be removed, for example, 50 times, each time a different set (although the same samples will usually be part of several of these test sets). The prediction ability each time is calculated, and the overall predictive ability is the average of each iteration.

Naturally, using these it is also possible to calculate the FP or FN rate, and the criterion employed to judge whether a method is suitable or not depends very much on the perspective of the scientist.

If the model is not very satisfactory, there are a number of ways to improve it. The first is to use a different computational algorithm. The second is to modify the existing method – a common approach might involve, for example, instead of using an entire spectrum (many wavelengths are not very meaningful), selecting just the most diagnostic parts of the spectrum. Finally, if everything else fails, change the analytical technique.

It is important to remember that there are two different reasons for using the techniques described in this section. The first is to optimize a computational model. This means that different models can be checked and the one that gives the best prediction rate is retained. In this manner, the samples left out are actually used to improve the model. The second is as an independent test of how well the model performs on unknowns. This is a subtly different reason and sometimes both motivations are mixed up, which can lead to overoptimistic predictions of the quality of a model on unknowns; therefore, care must be taken as discussed by Brereton (2006).

31.5.1.3 **Application**

Once a satisfactory model is available, it can then be applied to unknown samples, using analytical data such as spectra or chromatograms, to make predictions. Usually by this stage, special software is required that is tailor made for a specific application and measurement technique. The software is expected to determine whether a new sample really fits into the training set or not. One major difficulty is the detection of outliers that belong to none of the previously studied groups, for example, if a Brazilian orange juice sample was to be measured when the training set consists just of Spanish and Cypriot orange juices. In areas such as clinical or forensic science, outlier detection can be quite important; otherwise, an incorrect conviction or inaccurate medical diagnosis would be obtained. There are specific protocols according to the application area and type of spectroscopy involved, but outliers and changing instrumental performance are factors that should always be taken into account.

Another problem is to ensure the stability of the method over time, for example, instruments tend to perform slightly differently every day. Sometimes this can have a serious influence on the classification ability of chemometrics algorithms. One way is to perform a small test on the instrument on a regular basis and only accept data if the performance of this test falls within certain limits. If a spectrometer's performance is demonstrably changing with time (as often happens in NIR), and if historic classification (and calibration) models are to be used for future samples, there are special methods called *calibration transfer* that allow adjustments.

31.5.2
Classical Discriminant Analysis

The majority of statistically based software packages contain substantial numbers of procedures, called by various names, such as discriminant analysis, canonical variates analysis, and so on.

31.5.2.1 **Univariate Classification**

The simplest form of classification is *univariate* where one measurement or variable is used to divide objects into groups. An example may be if a peak in an IR spectrum of a known amount of sample is above a certain predefined threshold. If it is, then the sample falls into one group, and if not, it falls into another group. This might be useful, for example, in process analysis, if this peak relates to an unacceptable impurity in the mixture. However, in most problems in the spectroscopy of mixtures, criteria for whether a sample falls into one group or another are more subtle.

31.5.2.2 **Bivariate and Multivariate Discriminant Models**

Often, several measurements are required to determine the group a sample belongs to. Consider performing two measurements, which may involve measuring the intensity of two peaks in a spectrum and producing a graph of the values of these measurements for two groups, as shown in Figure 31.13. The objects denoted by squares are clearly distinct from the objects denoted by circles, but neither of the two measurements, alone, can discriminate between these groups; therefore, both are essential for classification. It is, however, possible to draw a discriminating line between the two groups. Graphically this can be represented by *projecting* the objects onto a line at right angles to this line as demonstrated in Figure 31.14. The projection can now be converted to a position along a single line. Often these numbers are converted to a *class distance*, which is the distance of each object to the center of the classes. If the distance to the center of class A is greater than that to class B, the object is placed in class A and vice versa. It is not always possible to exactly divide the classes into two groups by this method, but the misclassified samples are far from the center of both classes, with the two class distances that are approximately equal.

More sophisticated approaches involve calculating the class centroids (which are the average measurements of all the

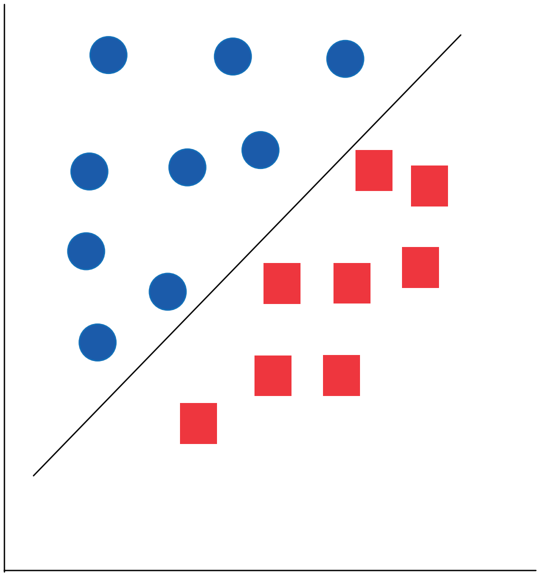

Fig. 31.13 Bivariate example where one measurement is insufficient to distinguish between classes.

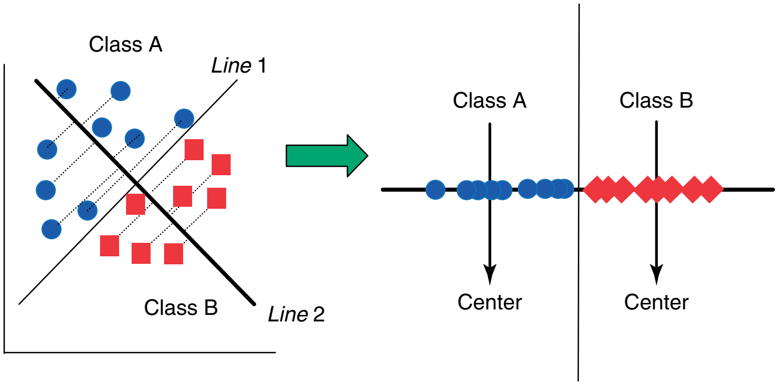

Fig. 31.14 Discriminant function.

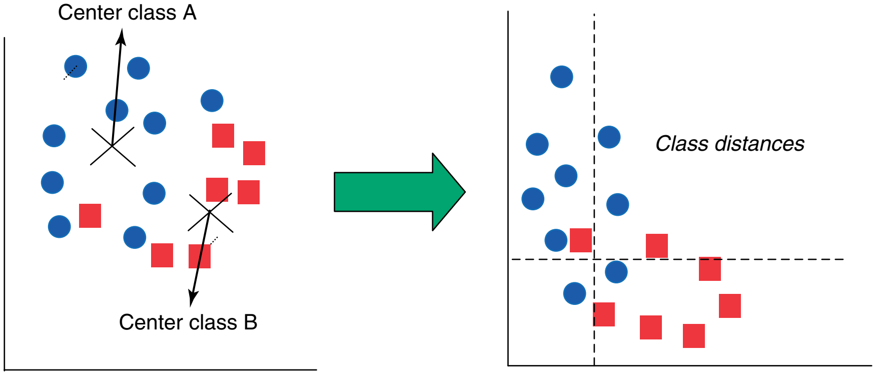

Fig. 31.15 Class distance plot using the Euclidean distance. Top left: almost certainly class A. Bottom left: unambiguous membership. Bottom right: almost certainly class B. Top right: unlikely to be a member of either class, sometimes called an *outlier*.

samples in one class) and the Euclidean distance to the centroids of each class defined as

$$d_A = \sqrt{(\mathbf{x} - \bar{\mathbf{x}}_A)\cdot(\mathbf{x} - \bar{\mathbf{x}}_A)'} \qquad (31.37)$$

where $\mathbf{x}$ is the vector representation of a specific spectrum, preprocessed if necessary, and $\bar{\mathbf{x}}_A$ is the centroid of class A. The distance to two or more class centroids can be calculated. The points no longer fall onto straight lines, but a class distance plot can be obtained as illustrated in Figure 31.15.

It is easy to extend the above-mentioned methods to multivariate situations where instead of two variables many (which can run to several hundreds in spectroscopy) are used to form the raw data. Normally, the class with the smallest distance is taken as the class an object belongs to, but there are several modifications for determining outliers or ambiguous samples or sometimes samples that genuinely belong to more than one class (e.g., a compound might contain both ketone and ester groups and thus exhibit spectroscopic properties of both types of compound and therefore belong to more than one class).

The Euclidean distance is not the only distance measure that can be used; a common enhancement is to employ the Mahalanobis distance defined as

$$d_A = \sqrt{(\mathbf{x} - \bar{\mathbf{x}}_{\mathrm{A}})\cdot\mathbf{S}^{-1}\cdot(\mathbf{x} - \bar{\mathbf{x}}_{\mathrm{A}})'} \quad (31.38)$$

where **S** is a variance–covariance matrix. There are two reasons for this modification to the Euclidean distance.

1. Different classes may be dispersed differently. One class may be quite scattered, whereas another might be very tight. Therefore, a larger Euclidean in from the centroid of the more scattered class has less significance than from the tighter class.
2. Variables may be correlated. This is quite common in spectroscopy, for example, several peaks may arise from a single feature; in NMR of mixtures, each compound will normally result in several features in the spectrum. However, other compounds that could be just as important for the discrimination may have many less spectral features.

A more subtle distinction is between linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). Both use the Mahalanobis distance, but the former takes the variance–covariance matrix of the entire training set (**S**) and in fact forms linear boundaries between classes, whereas the latter involves using a separate variance–covariance matrix for each class in the training set ($\mathbf{S}_{\mathrm{A}}$) and boundaries between classes can be represented as quadratic terms.

As with all classification models, once it is established on a training set, then it can be tested or applied to other samples as discussed in Section 31.5.1.

31.5.3
Discriminant Partial Least Squares

Another technique that can be employed for classification is discriminant PLS. We have already discussed PLS as a calibration method (Section 31.3.3). In that section we showed how PLS can be used to relate on block of data (e.g., "*x*"), such as a spectrum to another (e.g., "*c*"), such as the concentration of one or more components in a spectrum.

PLS can be used for discriminant analysis or classification (Dixon *et al.*, 2007). The simplest approach is when there are only two classes. In such case, one sets up a "*c*" vector, which consists of a number according to which class a spectrum belongs to; there are two common variations, using "1" and "0" or "+1" and "−1" for classes A and B. We recommend the latter convention, although there is no real difference between either of the approaches. PLS can be used to form a model between the *x* block, for example, a set of spectroscopic measurements, and the *c* block, which may correspond to the origins of the spectra. The results of calibration are a number. A simple rule for interpreting this number may be to assign an object to class A if it is positive and to class B if it is negative. There are elaborations, for example, a large positive or negative number much greater than ±1 might indicate an outlier, and there may be a region around 0 where we feel that the classification is ambiguous; hence, it is possible to choose a different decision threshold. If the two classes are, for example, whether a patient is diseased or not, then the decision threshold can be adjusted to reduce the number of FPs or FNs according to the one that is most important for the specific application.

There are many advantages of using PLS. For example, one can choose the number of significant components using cross validation just as in normal calibration to see how well the data are predicted numerically, or one can determine the optimum number of components by looking at the percent correctly classified. There will be scores, loadings, and weights, and these can be used to provide graphical information on the class structure and about the variables (e.g., spectroscopic wavelengths) that are most diagnostic of each class.

When there are more than two classes involved, the method becomes quite tricky. Normally, there is no particular sequential relationship between the classes. A simple approach is to set up a "*C*" matrix of $N - 1$ columns, where N is the number of classes. For example, if there are three classes, set up the first column with "+1" for class A and "−1" for either classes B or C, and the second column with "+1" for class B and "−1" for classes A or C. It is not necessary to have a third column. Obviously there are several (in this case three) ways of doing this. It can be a bit tricky to interpret the results of classification, but one can use an approach similar to that used in discriminant analysis when there are more than two classes. Alternatively, it is possible to use a "*C*" matrix where each column relates to membership of a single class; hence, for three classes we use three columns, with +1 for members and −1 for nonmembers. One should be careful while interpreting the result of calibration. For example, a negative number may be used to imply that an object does not belong to a specific class. What happens when all the three values from the calibration are negative? Do we say that the object is an outlier and belongs to no class? Or do we find the value closest to "+1" and assign an object to that specific class? There are a variety of ways of making choices and interpreting the results.

31.5.4
K-Nearest Neighbors

The methods discussed in Sections 31.5.2 and 31.5.3 involve producing statistical models. Nearest neighbor methods are conceptually much simpler and do not require elaborate statistical computations, and belong to the class of method often called *machine learning*, which had an influence on chemometric pattern recognition in its early days.

The K-nearest neighbor (KNN) method has been used by chemists for over 30 years (Kowalski and Bender, 1972). The algorithm starts with a number of objects assigned to each class.

The method is implemented as follows.

1. Assign a training set to known classes.
2. Calculate the distance of an unknown to all members of the training set. Usually the multivariate Euclidean distance is computed.
3. Rank these in order (1 = smallest distance and so on).
4. Pick the K smallest distances and see what classes the unknown in closest to.
5. Take the "majority vote" and use this for classification; usually K is an odd

number to prevent the problem of tied votes.

6. Sometimes it is useful to perform KNN analysis for a number of values of K, for example, 3, 5 and 7, and see if the classification changes. This can be used to spot anomalies.

This conceptually simple approach works well in many situations, but it is important to understand the limitations.

The first is that the numbers in each class of the training set should be approximately equal; otherwise, the "votes" will be biased toward the class with most representatives. The second is that for the simplest implementations, each variable is of equal significance. In spectroscopy, we may record hundreds of wavelengths, and some will either not be diagnostic or else be correlated. A way of getting around this is either to select the variables or to use another distance measure. The Mahalanobis distance is quite a useful alternative. The third problem is that ambiguous or outlying samples in the training set can cause major problems in the resultant classification. Fourth, the methods take no account of the spread or variance in a class. For example, if we were trying to determine whether a forensic sample is a forgery, it is likely that the class of forgeries has a much higher variance to the class of nonforged samples.

It is possible to follow procedures of validation (Section 31.5.1.2) just as in all other methods for supervised pattern recognition. There are quite a number of diagnostics that can be obtained using these methods.

However, KNN is a very simple approach that can be easily understood and programmed. Many spectroscopists like these approaches, while statisticians often prefer the more elaborate methods involving modeling the data. KNN makes very few assumptions, whereas methods based on modeling often inherently make assumptions such as normality of noise distributions that are not always experimentally justified, especially when statistical tests are employed to provide probabilities of class membership.

31.6
Multivariate Statistical Process Control

A major trend in chemometrics over the past decade has been in multivariate process control (MSPC) (Kresta, MacGregor and Marlin, 1991). This has its origins in industrial processes. Over the past decade, there has been a large development in online process spectroscopy, especially using miniaturized probes for NIR, MIR, UV/vis, and Raman spectroscopy. This allows real-time monitoring of processes and the ability to take spectra every few minutes if necessary. Some probes can be dipped into a reaction vessel, whereas others are pointed at the reaction mixture. In the pharmaceutical industry in particular, it is important to know whether a product is within accepted limits, and there are often legal requirements that impurities, often byproducts of a reaction and of unknown physiological properties, are below limits. If the composition of a product is outside these limits, then the process may have to be stopped or investigated. Chemometrics combined with online spectroscopy has a major role to play. MSPC can be used to decide whether a process is stable or whether there are unexpected changes. A major driving force has been the process analytical technology (PAT) initiative from the US Food and Drugs Administration (FDA) that has encouraged good practice in process monitoring (US Department of Health and Human Services, 2004).

MSPC has been quite slow to get off the ground and most process engineers still use simple univariate measures. A problem with the conventional univariate approach is that there may be an enormous number of parameters measured in time, at regular intervals; thus, in a typical day hundreds of thousands of pieces of information are available especially if spectroscopy is used for process monitoring. Many of the measurements are related; thus, it makes sense to form a multivariate model rather than several univariate models. Methods such as PCA and PLS come to our rescue. Why plot a single parameter when we could plot the PC (suitably scaled) of several parameters? And why not calibrate spectroscopic data to the concentration of a known additive, and use PLS predictions of concentrations as a variable that is monitored with time? In addition to following a number of physical (or process) measurements, the concentrations of specific compounds can be monitored with time, using multivariate calibration.

More sophisticated multivariate methods can be developed and used by chemical engineers who often first define regions of a reaction or batches of a product as normal operating conditions NOC samples; in the case of a reaction or a product, the samples arise from an acceptable process. These NOC samples represent a series of spectra that are considered normal or acceptable, and the aim of MSPC is to look for samples whose characteristics (normally spectroscopic) deviate significantly from the characteristics of these samples, by calculating a statistic for this deviation. There are two common methods.

1. The D-statistic is a measurement of distance a sample from the process is to the center of the NOC samples using a PC model and determines whether a specific sample has a systematic deviation from the steady state region. Usually the Mahalanobis distance from the NOC class of spectra is computed.
2. The Q-statistic calculates the residuals from the new samples when trying to fit them to a PC model, which is based on the NOC samples. The Q-statistic measures nonsystematic variations, which are not explained by the NOC model.

If both statistics are assumed to follow normal distributions, the control limit for these statistics can be computed; usually 95 and 99% limits are calculated. If samples appear to be exceeding these limits, it is probable that there is some difficulty in the process. It is possible to produce control charts in which the values of these statistics are plotted as a function of process time with the control limits indicated. Ninety-five percent limits are often called *warning limits* and allow the operator to investigate why the process appears to be behaving unusually, although one in 20 of the NOC samples will be expected to exceed these limits. The 99% limit is often called an *action limit.*

The use of chemometrics in online process control is highly developed theoretically and slowly being implemented in certain high-tech industries such as the pharmaceutical and food industries for front line products, but there is still a long way to go; the majority of processes are still monitored using fairly straightforward univariate control charts. Nevertheless, this is a major potential growth area, especially in terms of applications of chemometrics to spectroscopy.

31.7
Conclusions

There is a wide applicability of chemometrics to spectroscopy. With the rapid development of fast online computing power, it is possible to do much more compared to a couple of decades ago. What was feasible only off-line using powerful dedicated mainframes in the 1980s is feasible online and in real time now. Most modern chemometrics approaches had their origins several decades ago, but with user-friendly software, the ability to generate large quantities of data and powerful computing are now available to a wide range of users.

This chapter has reviewed some of the most common chemometrics techniques that have relevance to spectroscopy. There is a much wider ranger of methods, but most spectroscopists come across resolution, calibration, and pattern recognition in their basic work. Many approaches not discussed in this chapter have been developed by the machine learning and statistics community and are only slowly being realized by spectroscopists due to the lack of user-friendly software. PCA- and PLS-based approaches, in contrast, have been incorporated into readily available packages for many years and are now commonplace in some spectroscopies. The next few years are likely to see a rapid development in applications of chemometrics methods to a large variety of spectroscopies. However, most methods are based on a relatively small number of techniques such as MLR, PCA, and PLS, and understanding these basic building blocks will allow the spectroscopist to appreciate other approaches reported in the literature.

Glossary

Agglomerative Clustering: Forming relationships between samples by joining together those that are most similar.

ALS: Alternating least squares – a method of modeling usually spectra and concentration profiles that can incorporate constraints such as positivity.

Bootstrap: Alternative to cross-validation involving an iterative algorithm.

Confusion Matrix: See Contingency Table.

Contingency Table: A table whose rows and columns represent each group in a series of spectra, and containing information about how many samples are classified into each group, allowing misclassification rates to be determined.

Cross-Validation: Method for determining the number of PC or PLS components by leaving some samples out and predicting them.

Dendrogram: Graphic representation of cluster analysis.

Discriminant Analysis: Mathematical approaches for determining grouping in samples.

Dissimilarity Matrix: Like a similarity matrix, except that numerical values relate to dissimilarity.

D-statistic: Indicator used usually in MSPC to determine how close a sample is to the centre of an NOC model.

EFA: Evolving factor analysis – used to look at multivariate processes that evolve.

Eigenvalue: Size of PC.

Euclidean Distance: Difference between two spectra obtained by representing each as a point in multidimensional space.

Factor Analysis: Similar to PCA except that the components have an interpretable meaning, for example, as spectra and concentration profiles.

FN: False negative.

FP: False positive.

Inverse: Defined for most square matrices, for which the product is the unit matrix.

ITTFA: Iterative target transform factor analysis – a method similar to ALS.

KNN: K nearest neighbours – a method for classification.

LDA: Linear discriminant analysis – using Mahalanobis distance calculated on a training set consisting of several groups.

Loadings: Numerical values of PCs corresponding to variables.

Mahalanobis Distance: Difference between two spectra obtained by representing each as a point in multidimensional space and scaling each dimension according to variance.

Matrix: Representation of data in a two-dimensional array.

MLR: Multiple linear regression – an alternative to PLS that relates two sets of data, for example, a series of spectra, to concentrations of constituents.

MSPC: Multivariate statistical process control.

Multivariate: Measurements where there is more than one variable, for example, a spectrum at several wavelengths.

NIPALS: Nonlinear lterative partial least squares – a method for performing PCA.

NOC: Normal operating conditions – used in MSPC to indicate the set of spectra that correspond to samples that are of acceptable quality.

PC: Principal component – characterised by scores and loadings.

PCA: Principal components analysis – exploratory visualisation of data matrices, for example, a series of spectra.

PCR: Principal components regression – alternative to PLS and MLR.

PLS: Partial least squares – method for multivariate calibration of data matrices onto another block, calibrating a series of spectra onto concentrations of constituents.

PLS1: PLS algorithm for which there is a single c variable.

PLS2: PLS algorithm for which there is more than one c variable.

Pseudoinverse: Mathematical equivalent to an inverse for a rectangular matrix.

QDA: Quadratic discriminant analysis – using Mahalanobis distance calculated separately on each group in a training set.

Q-statistic: Indicator used usually in MSPC to determine how well a sample fits an NOC model.

Rectangular Matrix: Matrix for which the number of rows and columns are unequal.

RMSEC: Root mean square error of calibration – used for autopredictive models.

RMSECV: Root mean square error of cross validation – often used to determine how many PLS components are suitable for a model.

RMSEP: Root mean square error of prediction – used on an independent test set.

Rotation: Transforming a PC onto a target vector normally consisting of the concentrations of an analyte.

Sample: A row of a matrix usually representing a spectrum.

Scalar: Single number.

Scores: Numerical values of PCs corresponding to samples.

Similarity Matrix: A matrix whose rows and columns represent samples containing values that indicate how similar each sample is to every other sample.

Square Matrix: Matrix for which the number of rows and columns are equal.

Standardizing: Subtracting the mean and dividing by the standard deviation.

SVD: Singular value decomposition – a method for performing PCA.

Test Set: Samples used to determine how well a model performs on samples left out of the model building.

TN: True negative.

TP: True positive.

Training Set: Samples used to establish a model.

Transformation: See rotation.

Transpose: Swapping rows and columns of a matrix.

Univariate: Measurements where there is single variable.

Validation: Determining how effective a model is.

Variable: A column of a matrix usually representing a frequency or wavelength.

Vector: Representation of data in a one-dimensional array, for example, a column or row of a matrix.

References

Adams, M.J. (2004) *Chemometrics in Analytical Spectroscopy*, 2nd edn, Royal Society of Chemistry, Cambridge.

Beebe, K.R., Pell, R.J. and Seasholtz, M.B. (1998) *Chemometrics: A Practical Guide*, Wiley, New York.

Brereton, R.G. (ed.) (1992) *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*, Elsevier, Amsterdam.

Brereton, R.G. (2000) Introduction to multivariate calibration in analytical chemistry. *Analyst*, **125**, 2125–2154.

Brereton, R.G. (2003) *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*, Wiley, Chichester.

Brereton, R.G. (2006) Consequences of sample sizes, variable selection, model validation and optimisation for predicting classification ability from analytical data. *Trends Anal. Chem.*, **25**, 1103–1111.

Brereton, R.G. (2007) *Applied Chemometrics for Scientists*, John Wiley & Sons, Chichester.

Brereton, R.G., Scott, D.R., Massart, D.L., Dessy, R.E., Hopke, P.K., Spiegelman, C.H. and Wegscheider, W. (eds) (1992) *Chemometrics Tutorials II*, Elsevier, Amsterdam.

Bruno, P., Caselli, M., Curri, M.L., Genga, A., Striccoli, R. and Traini, A. (2000) Chemical characterisation of ancient pottery from south of Italy by inductively coupled plasma atomic emission spectroscopy (ICP-AES) statistical multivariate analysis of data. *Anal. Chim. Acta*, **410**, 193–202.

De Braekeleer, K. and Massart, D.L. (1997) Evaluation of the orthogonal projection approach (OPA) and the SIMPLISMA approach on the Windig standard spectral data sets. *Chemometr. Intell. Lab Syst.*, **39**, 127–141.

Dixon, S.J., Xu, Y., Brereton, R.G., Soini, A., Novotny, M.V., Oberzaucher, E., Grammer, K. and Penn, D.J. (2007) Pattern recognition of gas chromatography mass spectrometry of human volatiles in sweat to distinguish the sex of subjects and determine potential discriminatory marker peaks. *Chemometr. Intell. Lab. Systems*, **87**, 161–172.

Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Eriksson, L., Johansson, E., Kettaneh-Wold, N. and Wold, S. (2001) *Multi- and Megavariate Data Analysis: Principles and Applications*, Umetrics, Umeå.

Esbensen, K.H. (2002) *Multivariate Data Analysis in Practice*, CAMO, Oslo.

Everitt, B.S., Landau, S. and Leese, M. (2001) *Cluster Analysis*, Arnold, London.

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179–188.

Gampp, H., Maeder, M., Meyer, C.J. and Zuberbuhler, A.D. (1985) Calculation of equilibrium constants from multiwavelength spectroscopic data–III. Model-free analysis of spectrophotometric and ESR titrations. *Talanta*, **32**, 1133–1139.

Geladi, P. and Kowalski, B.R. (1986) Partial least-squares regression: a tutorial. *Anal. Chim. Acta*, **185**, 1–17.

Gemperline, P.J. (ed.) (2006) *Practical Guide to Chemometrics*, 2nd edn, CRC, Boca Raton.

Haswell, S.J. (ed.) (1992) *Practical Guide to Chemometrics*, CRC, Boca Raton.

Hoskuldsson, A. (1988) PLS regression methods. *J. Chemometrics*, **2**, 211–228.

Joliffe, I.T. (1987) *Principal Components Analysis*, Springer-Verlag, New York.

Kowalski, B.R. (ed.) (1984) *Chemometrics: Mathematics and Statistics in Chemistry*, Reidel, Dordrecht.

Kowalski, B.R. and Bender, C.F. (1972) K-nearest neighbor classification rule (pattern-recognition) applied to nuclear magnetic-resonance spectral interpretation. *Anal. Chem.*, **44**, 1405–1411.

Kramer, R. (1998) *Chemometrics Techniques for Quantitative Analysis*, Marcel Dekker, New York.

Kresta, J.V., MacGregor, J.F. and Marlin, T.E. (1991) Multivariate statistical monitoring of process operating performance. *Can. J. Chem. Eng.*, **69**, 35–47.

Malinowski, E.R. (2002) *Factor Analysis in Chemistry*, 3rd edn, Wiley, New York.

Malinowski, E.R. and Howery, D.R. (1980) *Factor Analysis in Chemistry.*, Wiley, New York.

Mark, H. and Workman, J. (2003) *Statistics in Spectroscopy (Hardcover)*, 2nd edn, Academic Press, New York.

Martens, H. and Martens, M. (2000) *Multivariate Analysis of Quality*, Wiley, Chichester.

Martens, H. and Næs, T. (1989) *Multivariate Calibration*, Wiley, Chichester.

Massart, D.L., Brereton, R.G., Dessy, R.E., Hopke, P.K., Spiegelman, C.H. and Wegscheider, W. (eds) (1990) *Chemometrics Tutorials*, Elsevier, Amsterdam.

Massart, D.L. and Kaufmann, L. (1983) *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*, Wiley, New York.

Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., De Jong, S., Lewi, P.J. and Smeyers-Verbeke, J. (1997) *Handbook of Chemometrics and Qualimetrics Part A*, Elsevier, Amsterdam.

Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Michotte, Y. and Kaufman, L. (1988) *Chemometrics: A Textbook*, Elsevier, Amsterdam.

Meloun, M., Militky, J. and Forina, M. (1992) *Chemometrics for Analytical Chemistry*, Vol. 1, Ellis Horwood, Chichester.

Meloun, M., Militky, J. and Forina, M. (1994) *Chemometrics for Analytical Chemistry*, Vol. 2, Ellis Horwood, Chichester.

Naes, T., Isaksson, T., Fearn, T. and Davies, T. (2002) *A User Friendly Guide to Multivariate Calibration and Classification*, NIR Publications, Chichester.

Otto, M. (1998) *Chemometrics : Statistics and Computer Applications in Analytical Chemistry*, Wiley-VCH, Weinheim.

Sharaf, M.A., Illman, D.L. and Kowalski, B.R. (1986) *Chemometrics*, Wiley, New York.

Tauler, R., Kowalski, B. and Fleming, S. (1993) Multivariate curve resolution applied to spectral data from multiple runs of an industrial-process. *Anal. Chem.*, **65**, 2040–2047.

US Department of Health and Human Services, FDA (2004) *Guidance for Industry PAT – A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance*, September 2004 (http://www.fda.gov/cder/guidance/6419fnl.htm).

Vandeginste, B.G.M., Derks, W. and Kateman, G. (1985) Multicomponent self-modeling curve resolution in high-performance liquid-chromatography by iterative target transformation analysis. *Anal. Chim. Acta*, **173**, 253–264.

Vandeginste, B.G.M., Massart., D.L., Buydens, L.M.C., de Jong, S., Lewi, P.J. and Smeyers-Verbeke, J. (1998) *Handbook of Chemometrics and Qualimetrics Part B*, Elsevier, Amsterdam.

Windig, W., Heckler, C.E., Agblebor, F.A. and Evans, R.J. (1992) Self-modeling mixture analysis of categorized pyrolysis mass-spectral data with the SIMPLISMA approach. *Chemometr. Intell. Lab Syst.*, **14**, 195–207.

Wold, S. (1978) Cross-validatory estimation of number of components in factor and principal components models. *Technometrics*, **20**, 397–405.

Wold, S., Esbensen, K. and Geladi, P. (1987) Principal components analysis. *Chemometr. Intell. Lab Syst.*, **2**, 37.

Further Reading

Adams, M.J. (2004) *Chemometrics in Analytical Spectroscopy*, 2nd edn, Royal Society of Chemistry, Cambridge.

Beebe, K.R., Pell, R.J., and Seasholtz, M.B. (1998) *Chemometrics : a Practical Guide*, John Wiley & Sons, New York.

Brereton, R.G. (1990) *Chemometrics : Applications of Mathematics and Statistics to Laboratory Systems*, Ellis Horwood, Chichester.

Brereton, R.G. (ed) (1992) *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*, Elsevier, Amsterdam.

Brereton, R.G. (2003) *Chemometrics : Data Analysis for the Laboratory and Chemical Plant*, John Wiley & Sons, Chichester.

Brereton, R.G. (2007) *Applied Chemometrics for Scientists*, John Wiley & Sons, Chichester.

Brereton, R.G., Scott, D.R., Massart, D.L., Dessy, R.E., Hopke, P.K., Spiegelman, C.H., and Wegscheider, W. (eds) (1992) *Chemometrics Tutorials II*, Elsevier, Amsterdam.

Coomans, D. and Broeckaert, I. (1986) *Potential Pattern Recognition in Chemical and Medical Decision Making*, Research Studies Press, Letchworth.

De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L. (2000) The Mahalanobis distance. *Chemometr. Intell. Lab Syst.*, **50**, 1–18.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., and Wold, S. (2001) *Multi- and Megavariate Data Analysis: Principles and Applications*, Umetrics, Umeå.

Esbensen, K.H. (2002) *Multivariate Data Analysis in Practice*, CAMO, Oslo.

Gemperline, P.J. (ed) (2006) *Practical Guide to Chemometrics*, 2nd edn, CRC, Boca Raton, FL.

Jurs, P.C. and Isenhour, T.L. (1975) *Chemical Applications of Pattern Recognition*, John Wiley & Sons, New York.

Kowalski, B.R. (ed) (1984) *Chemometrics: Mathematics and Statistics in Chemistry*, Reidel, Dordrecht.

Kramer, R. (1998) *Chemometrics Techniques for Quantitative Analysis*, Marcel Dekker, New York.

Mark, H. and Workman, J. (2007) *Chemometrics in Spectroscopy*, Elsevier, London.

Massart, D.L., Brereton, R.G., Dessy, R.E., Hopke, P.K., Spiegelman, C.H., and Wegscheider, W. (eds) (1990) *Chemometrics Tutorials*, Elsevier, Amsterdam.

Massart, D.L. and Kaufmann, L. (1983) *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*, John Wiley & Sons, New York.

Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., De Jong, S., Lewi, P.J., and Smeyers-Verbeke, J. (1997) *Handbook of Chemometrics and Qualimetrics Part A*, Elsevier, Amsterdam.

Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Michotte, Y., and Kaufman, L. (1988) *Chemometrics : A Textbook*, Elsevier, Amsterdam.

Meloun, M., Militky, J., and Forina, M. (1992, 1994) *Chemometrics for Analytical Chemistry Vols 1 and 2*, Ellis Horwood, Chichester.

Miller, J.N. and Miller, J.C. (2005) *Statistics and Chemometrics for Analytical Chemistry*, 5th edn, Pearson, Harlow.

NIST *Engineering Statistics Handbook*, http://www.itl.nist.gov/div898/handbook/. 20 May 2009.

Otto, M. (2007) *Chemometrics : Statistics and Computer Applications in Analytical Chemistry*, Wiley-VCH Verlag GmbH, Weinheim.

Sharaf, M.A., Illman, D.L., and Kowalski, B.R. (1986) *Chemometrics*, John Wiley & Sons, New York.

Strouf, O. (1986) *Chemical Pattern Recognition*, Research Studies Press, Letchworth.

Vandeginste, B.M.G., Massart, D.L., Buydens, L.M.C., de Jong, S., Lewi, P.J., and Smeyers-Verbeke, J. (1998) *Handbook of Chemometrics and Qualimetrics Part B*, Elsevier, Amsterdam.

Varmuza, K. (1980) *Pattern Recognition in Chemistry*, Springer, Berlin.

Varmuza, K. and Filmoser, P. (2009) *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC, Boca Raton, FL.

5

Weak Interaction

Ken-ichi Hikasa

5.1 Introduction 149
5.2 Elementary Particles 149
5.3 Elementary Processes of the Weak Interaction 150
5.4 Structure of the Weak Interaction 151
5.4.1 Charged Current 152
5.4.2 Neutral Current 152
5.4.3 Effective Four-Fermion Interaction 154
5.5 Basic Constants of the Weak Interaction 154
5.5.1 Charged Current 154
5.5.2 Flavor Mixing 154
5.5.3 Neutral Current 155
5.5.4 W and Z Masses 156
5.6 Processes Induced by Weak Interactions 156
5.6.1 Decay of W and Z 156
5.6.2 Weak Decay of Leptons 156
5.6.2.1 Muon Decay 156
5.6.2.2 τ-Decay 157
5.6.3 Weak Decay of Hadrons 157
5.6.3.1 Neutron and Nuclear β-Decay 157
5.6.3.2 Pion Decay 158
5.6.3.3 Decay of Hadrons Containing Strange Quarks 158
5.6.3.4 Decay of Hadrons Containing Heavy Quarks 159
5.6.4 Top-Quark Decay 159
5.6.5 Flavor Oscillation: Neutral-Kaon and B-Meson Systems 159
5.6.6 CP-Invariance Violation 160
5.6.7 Neutrino Interaction 160
5.7 Gauge Theory of the Weak Interaction: the Standard Model 160
5.8 The Higgs Boson 161
5.9 Origin of Electroweak Symmetry Breaking 162
5.10 Neutrino Masses and Mixing 163
5.11 Beyond the Standard Model 164

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

Glossary 165
References 165
Further Reading 167

5.1 Introduction

The weak interaction is one of the four known fundamental interactions in nature, the other three being gravity, electromagnetism, and the strong interaction. Unlike gravity and electromagnetism, the weak interaction generates an extremely short-range force between matter, which exponentially vanishes at distances larger than 10^{-18} m. The effect of the weak interaction is thus usually unnoticeable unless one directly probes at a short distance comparable to its range. Because of the uncertainty principle, this requires energies larger than 10^2 GeV, which are attainable only by the highest-energy accelerators at present.

There are low-energy processes, however, in which the weak interaction plays a crucial role. These include decay of most elementary particles, decay of hadrons if stable against the strong and electromagnetic interactions, and β-decay and electron capture of unstable nuclei. Processes involving neutrinos are also due to the weak interaction, since neutrinos have no strong or electromagnetic interactions. Particularly important among these is the nuclear-fusion reaction in stars like the Sun. The main branch of the reaction chain, four protons fusing into a helium-4 nucleus by emitting two positrons and two electron neutrinos, includes processes caused by the weak interaction. The Sun would not shine if the weak interaction did not exist! Moreover, the long lifetime of the Sun, which enabled the evolution of life on the Earth, is the consequence of the weakness of the interaction. The weak interaction, however unfamiliar, is vital to our very existence.

5.2 Elementary Particles

The most basic constituents of matter are called *elementary particles*. Twelve such particles are known to date. Six of them are quarks, named u (up), d (down), s (strange), c (charm), b (bottom), and t (top) quark. Three of the quarks (u, c, t) have electric charge $\frac{2}{3}$ (in units of e) and the others (d, s, b) have charge $-\frac{1}{3}$.

The remaining six elementary particles are leptons. Three of them, e^- (electron), μ^- (muon), and τ^- (tauon or tau), have charge -1. The other three are called *neutrinos* and are electrically neutral: ν_e (electron neutrino), ν_μ (muon neutrino), and ν_τ (tau neutrino). All quarks and leptons have spin (intrinsic angular momentum) $\frac{1}{2}$ in units of h and obey Fermi statistics. Each of them

Encyclopedia of Applied High Energy and Particle Physics. Edited by Reinhard Stock
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-40691-3

is associated with an antiparticle. Antiparticles have opposite charge to that of the particles and are denoted as $\overline{u}$, $\overline{d}$, $\overline{\nu}_e$, e^+, and so on.

The main differences between the quarks and leptons are that only the former participate in the strong interaction and that each quark has three internal degrees of freedom called *color*, while leptons have no color. These two facts are deeply related to each other. Because of the confining nature of the strong interaction, quarks cannot be observed singly but form color-neutral composite states called *hadrons*. The proton p, neutron n, and pion π are the most familiar hadrons. Hadrons are classified into mesons and baryons. Mesons are integral-spin hadrons (spin 0, 1, ...) and can be viewed as bound states of a quark and an antiquark. Baryons are half-integral-spin hadrons (spin $\frac{1}{2}$, $\frac{3}{2}$, ...) and are composed of three quarks.

The quarks and leptons are grouped into three generations in the order of increasing masses. Atoms and molecules are composed of the first-generation particles u, d, and e. The electron neutrino ν_e is associated with the electron. The second (c, s, ν_μ, μ) and the third (t, b, ν_τ, τ) generations are identical to the first except for the properties related to the mass. We note that the neutrino states are labeled by the corresponding charged leptons (''flavor eigenstates''), as it is the only means we can detect them (see Section 5.10 for neutrino masses).

In addition to these matter particles, four types of elementary particles having spin 1 and obeying Bose statistics are known to exist: γ (photon), g (gluon), W, and Z. These act as mediators of fundamental interactions. The electromagnetic interaction is due to the exchange of a photon, and the strong interaction is caused by gluons. The W and Z bosons are responsible for the weak interaction. These bosons except W are electrically neutral, and each of them is its own antiparticle. The W boson has charge -1, and its antiparticle has charge $+1$. These weak bosons, $W^\pm$ and Z^0, were discovered in 1983 (Arnison *et al.*, 1983a; Banner *et al.*, 1983; Arnison *et al.*, 1983b; Bagnaia, *et al.*, 1983) and had been by far the heaviest elementary particles until the discovery (Abe *et al.*, 1994a; 1994b; 1995; Abachi *et al.*, 1995) of the top quark.

5.3
Elementary Processes of the Weak Interaction

The weak interaction can be divided into two classes, the charged-current and neutral-current interactions. The former is associated with the W boson and the latter with the Z.

All quarks and leptons participate in the weak interaction. Taking the first-generation particles as an example, the elementary processes of charged-current interaction are

$$e^- \longleftrightarrow \nu_e + W^-, \quad d \longleftrightarrow u + W^- \tag{1}$$

and those obtained from Eq. (1) by ''crossing'' (moving a particle to the other side and changing it to its antiparticle). For the first reaction in Eq. (1), the following five processes result from crossing:

$$\begin{aligned} &e^- + \overline{\nu}_e \longleftrightarrow W^-, \quad e^+ + \nu_e \longleftrightarrow W^+ \\ &\nu_e \longleftrightarrow e^- + W^+, \quad \overline{\nu}_e \longleftrightarrow e^+ + W^-, \\ &e^+ \longleftrightarrow \overline{\nu}_e + W^+ \end{aligned} \tag{2}$$

Elementary neutral-current interactions of the first-generation fermions are

$$u \longleftrightarrow u + Z, \quad d \longleftrightarrow d + Z,$$
$$\nu_e \longleftrightarrow \nu_e + Z, \quad e^- \longleftrightarrow e^- + Z \qquad (3)$$

The second and third generations of quarks and leptons interact with the W and Z boson in the same way as the first:

$$\mu^- \longleftrightarrow \nu_\mu + W^-, \quad s \longleftrightarrow c + W^-$$
$$\tau^- \longleftrightarrow \nu_\tau + W^-, \quad b \longleftrightarrow t + W^- \qquad (4)$$

for the charged current, and

$$c \longleftrightarrow c + Z, \quad s \longleftrightarrow s + Z,$$
$$\nu_\mu \longleftrightarrow \nu_\mu + Z \quad \mu^- \longleftrightarrow \mu^- + Z,$$
$$t \longleftrightarrow t + Z, \quad b \longleftrightarrow b + Z,$$
$$\nu_\tau \longleftrightarrow \nu_\tau + Z, \quad \tau^- \longleftrightarrow \tau^- + Z \qquad (5)$$

for the neutral current.

In addition to these couplings within the same generation, intergeneration processes of *quarks* such as

$$s \longleftrightarrow u + W^-, \quad b \longleftrightarrow c + W^- \qquad (6)$$

occur. These have a weaker strength than the processes involving fermions of the same generation. (For leptons, our definition of the neutrino states precludes intergeneration processes.)

Because a fermion changes its identity by emitting or absorbing a W, the charged-current interaction induces instability of most of the quarks and leptons.

No flavor-changing neutral-current interaction such as $s \leftrightarrow d + Z$ has been observed. Because of this diagonality, the neutral-current interaction does not cause decay of particles, and the existence of the neutral-current interaction was demonstrated (Hasert *et al.*, 1973), only after its existence was anticipated by theoretical consideration (Glashow, 1961; Weinberg, 1967; Salam, 1968). By contrast, a process caused by charged-current was already known at the end of the nineteenth century in the form of nuclear β-decays.

Processes creating a W or Z boson have large threshold energies attainable only at very high energy accelerators. At lower energies, second-order processes with a W or Z boson in the intermediate (virtual) state can occur. These include scattering processes such as

$$\nu_\mu + e^- \longrightarrow \left(\mu^- + W^+ + e^-\right)$$
$$\longrightarrow \mu^- + \nu_e \qquad (7)$$

and decay processes such as

$$\mu^- \longrightarrow \left(\nu_\mu + W^-\right) \longrightarrow \nu_\mu + e^- + \nu_e \qquad (8)$$

Before the discovery of the W and Z bosons in 1983, these kinds of processes involving four fermions were the only experimental manifestation of the weak interaction. The amplitude for these processes is suppressed at low energies by a factor of $1/m_W^2$, yielding a very small cross section or decay rate. This is the reason that the weak interaction was named as such, in spite of the fact that its inherent strength is as large as the electromagnetic interaction. This suppression factor can be translated to the range of the weak force, which is of the order of the Compton wavelength of the W boson, $\hbar/m_W c \sim 10^{-18}$ m. This is 8(3) orders of magnitude smaller than the atomic (nuclear) scale.

5.4
Structure of the Weak Interaction

The electromagnetic interaction can be described as the coupling of the electromagnetic field with the electromagnetic current. The current can be described as the bilinear function of quark or lepton fields at the fundamental level. The strength of the interaction is determined by the elementary charge e.

The form of the interaction of W and Z with quarks and leptons is surprisingly similar to the electromagnetic interaction. The W and Z field couple to weak currents, and the interaction can be described by only a small number of parameters, which are the weak-coupling constant g and the weak "charge" of the quarks and leptons. The Z interaction also depends on the electromagnetic coupling e and the electric charge of the particle. In fact, the weak and electromagnetic interactions are deeply entangled with each other.

5.4.1
Charged Current

It is easiest to write the form of the interaction in terms of Lagrangian density. The charged-current interaction is given by

$$\mathcal{L} = -\left(\frac{g}{2\sqrt{2}}\right) J^{\mu} W_{\mu} + \text{Hermitian conjugate} \tag{9}$$

where J^{μ} is the weak charged current,

$$J^{\mu} = \sum_{i,j} V_{ij}\overline{u}_i\gamma^{\mu}(1-\gamma_5)\, d_j + \sum_i \overline{\nu}_i\gamma^{\mu}(1-\gamma_5)\, l_i \tag{10}$$

Here the particles are represented by the corresponding fields: W_{μ} is the W^{+} field, and

$$u_i = (u, c, t)\,, \quad d_i = (d, s, b), \qquad \nu_i = (\nu_e, \nu_{\mu}, \nu_{\tau}), \quad l_i = (e, \mu, \tau) \tag{11}$$

are the Dirac fields for quarks and leptons. The quantity V_{ij} is the element of a 3 × 3 unitary matrix (Kobayashi and Maskawa, 1973) called the *Cabibbo–Kobayashi–Maskawa matrix* or the *quark-mixing matrix*. This matrix represents the mismatch of charge $\frac{2}{3}$ and $-\frac{1}{3}$ quark mass eigenstates and causes the mixing between generations (Gell-Mann and Lévy, 1960; Cabibbo, 1963; Glashow, Iliopoulos and Maiani, 1970; Kobayashi and Maskawa, 1973).

Up to the unitary matrix **V**, each term in Eq. (10) has the same normalization. The strength of the interaction thus is independent of fermion species. This property is called *universality* of the charged-current interaction.

5.4.2
Neutral Current

The neutral-current interaction has the form

$$\mathcal{L} = -g_Z J_Z^{\mu} Z_{\mu} \tag{12}$$

with

$$J_Z^{\mu} = \sum_f \bar{f}\gamma^{\mu}\left(v_f - a_f\gamma_5\right) f \tag{13}$$

where Z^{μ} is the Z field and f is the fermion field (sum over all quarks and leptons is implied). The overall coupling constant g_Z can be written in terms of e and g. For this purpose, it is convenient to define

$$\sin\theta_W = \frac{e}{g} \tag{14}$$

The angle θ_W is called the *Weinberg angle* or the *weak mixing angle*. Then the Z coupling is given by

$$g_Z = \frac{g}{\cos\theta_W} \tag{15}$$

This relation has something to do with the nature of the four bosons $W^{\pm}$, Z, and γ. Originally, there exist a triplet of bosons (W^{+}, W^{0}, W^{-}) and a singlet B^{0}, whose respective coupling constants are g and g'. The γ and Z are the linear combinations of W^{0} and B^{0} with mixing angle θ_W:

$$\gamma = W^{0} \sin\theta_W + B^{0} \cos\theta_W \tag{16}$$

$$Z = W^{0} \cos\theta_W - B^{0} \sin\theta_W \tag{17}$$

The mixing angle is derived from the couplings by the relation $\tan\,\theta_W = g'/g$. The γ and Z coupling constants are

$$\begin{aligned} e &= \frac{gg'}{\sqrt{g^{2}+g'^{2}}} = g\sin\theta_W \\ g_Z &= \frac{g}{\cos\theta_W} = \frac{e}{\cos\theta_W \sin\theta_W} \end{aligned} \tag{18}$$

Vector and axial couplings for each fermion, v_f, a_f, are given by

$$\begin{aligned} v_f = {} & \frac{1}{2}\left[T_3\left(f_L\right) + T_3\left(f_R\right)\right] \\ & -2Q_f \sin^{2}\theta_W \end{aligned} \tag{19}$$

$$a_f = \frac{1}{2}\left[T_3\left(f_L\right) - T_3\left(f_R\right)\right] \tag{20}$$

where Q_f is the electromagnetic charge of the fermion, and $T_3(f_{L,R})$ is the weak charge (called the *third component of weak isospin*) of the left- and right-handed components of the fermion field as given in Table 5.1. The current (Eq. 13) can be cast in the form

$$J^{\mu}_{Z} = J^{\mu}_{3} - J^{\mu}_{\text{em}} \sin^{2}\theta_W \tag{21}$$

Tab. 5.1 Particle content of the standard model.

Fields	**Spin**	**SU(3)**	**SU(2)**	***T*₃**	***Y***	***B***	***L***
$q_L = \begin{pmatrix} u_L \\ d_L \end{pmatrix}$	$\frac{1}{2}$	3	2	$+\frac{1}{2}$ $-\frac{1}{2}$	$+\frac{1}{6}$	$\frac{1}{3}$	0
u_R	$\frac{1}{2}$	3	1	0	$+\frac{2}{3}$	$\frac{1}{3}$	0
d_R	$\frac{1}{2}$	3	1	0	$-\frac{1}{3}$	$\frac{1}{3}$	0
$l_L = \begin{pmatrix} \nu_{eL} \\ e_L \end{pmatrix}$	$\frac{1}{2}$	1	2	$+\frac{1}{2}$ $-\frac{1}{2}$	$-\frac{1}{2}$	0	1
e^{-}_{R}	$\frac{1}{2}$	1	1	0	-1	0	1
G	1	8	1	0	0	0	0
$\begin{pmatrix} W^{+} \\ W^{0} \\ W^{-} \end{pmatrix}$	1	3	3	$+1$ 0 -1	0	0	0
B^{0}	1	1	1	0	0	0	0
$H = \begin{pmatrix} H^{0} \\ H^{-} \end{pmatrix}$	0	1	2	$+\frac{1}{2}$ $+\frac{1}{2}$	$-\frac{1}{2}$	0	0

[a] Only the first generation of quarks and leptons is shown. Other generations have identical quantum numbers. H denotes the Higgs field. SU(3) and SU(2) entries show the dimension of the representation, T_3 is the third component of the weak isospin SU(2); Y is the weak hypercharge U(1) quantum number. The electric charge is given by $Q = T_3 + Y$. Baryon number B and lepton number L are shown for convenience, but these are not gauge quantum numbers.

where J^{μ}_{em} is the electromagnetic current and J^{μ}_{3} is the third component of the weak-isospin current. With the charged current and its Hermitian conjugate, J^{μ}_{3} comprises all the three components of weak isospin.

The weak currents (10) and (13) contain both vector (γ_μ) and axial ($\gamma_\mu\gamma_5$) terms, which implies that parity invariance is not respected by these couplings. The parity nonconservation of the weak interaction was conjectured (Lee and Yang, 1956) to explain seemingly contradictory properties of kaons, and immediately confirmed by experiments (Wu *et al.*, 1957; Garwin, Lederman and Weinrich, 1957; Friedman and Telegdi, 1957). The violation of parity invariance in the weak interaction is substantial. The effect is maximal for neutrinos in that only helicity -1 neutrino states participate in the weak interactions. In fact, there is no evidence that helicity $+1$ neutrino states exist.

The weak interaction can thus be described by the parameters e, $\sin^2\theta_W$, and V_{ij}, as well as the W and Z masses m_W and m_Z. Furthermore, the masses satisfy the relation

$$m_Z = \frac{m_W}{\cos\theta_W} \tag{22}$$

5.4.3
Effective Four-Fermion Interaction

At energies sufficiently below the W mass, we can use the following effective Lagrangian to describe the weak interaction. For the charged-current interaction,

$$\mathcal{L}_{\text{eff}} = -\left(\frac{G_F}{\sqrt{2}}\right) J^{\dagger}_{\mu} J^{\mu} \tag{23}$$

where

$$\frac{G_F}{\sqrt{2}} = \frac{g^2}{8m_W^2} \tag{24}$$

is called the *Fermi coupling constant* (the factor of $\sqrt{2}$ is of historical origin). The prototype of this Lagrangian (Fermi, 1934) has been expanded and modified several times before the final form was obtained. Notice that only the ratio of the coupling and W mass appears in this effective interaction.

The corresponding Lagrangian for the neutral current is

$$\mathcal{L}_{\text{eff}} = -2\sqrt{2}G_F J_{Z\mu} J^{\mu}_{Z} \tag{25}$$

with the same overall coupling G_F due to the relation (22).

5.5
Basic Constants of the Weak Interaction

5.5.1
Charged Current

The low-energy charged-current interaction of leptons is governed by only one parameter, G_F. This constant can be determined from the lifetime of the muon and is known with high precision:

$$G_F = 1.16637(1) \times 10^{-5}\,\text{GeV}^2 \tag{26}$$

in natural units (the number in the parenthesis is the error of the last digit).

5.5.2
Flavor Mixing

The charged-current interaction of quarks also depends on the Cabibbo–Kobayashi–Maskawa matrix **V**, which is a 3×3 unitary matrix. Although such a matrix generally has nine independent parameters, not all of them are physically observable because some of the phase degrees of freedom can be absorbed into relative phases of the quark fields. Only four physical parameters remain, three of which are real mixing angles and one is a complex phase. If this phase is nonzero, the charged-current interaction violates CP invariance. The complex phase can appear only when at least three generations of quarks exist. In fact, this was the motivation for Kobayashi and Maskawa in anticipating the existence of six flavors of quarks.

The generation structure of the quarks implies that Cabibbo–Kobayashi–Maskawa matrix is approximately diagonal ($\mathbf{V} \approx 1$) and the intergeneration couplings are suppressed. Furthermore, the off-diagonal elements show some hierarchical structure. Our knowledge of $\mathbf{V}$ is summarized in the following "parameterization" (Wolfenstein, 1983):

$$V = \begin{array}{c} \\ u \\ c \\ t \end{array} \begin{array}{c} \begin{array}{ccc} d & s & b \end{array} \\ \begin{pmatrix} 1-\frac{\lambda^2}{2} & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1-\frac{\lambda^2}{2} & A\lambda^2 \\ A\lambda^3(1-\rho-i\eta) & -A\lambda^2 & 1 \end{pmatrix} \end{array} + O(\lambda^4) \tag{27}$$

which is formally an expansion in terms of the parameter $\lambda = V_{\text{us}}$. This and the other three parameters, A, ρ, and η, are determined from the weak interaction processes of hadrons, notably semileptonic decay rates of strange particles and B mesons, as well as CP violating observables:

$$\lambda = 0.2257 \pm 0.0018 \tag{28}$$

$$A = 0.814 \pm 0.022 \tag{29}$$

$$\rho = 0.139^{+0.032}_{-0.016} \tag{30a}$$

$$\eta = 0.358^{+0.015}_{-0.017} \tag{30b}$$

This structure of the matrix has the following phenomenological consequences:

1. suppression of strange-particle decay rates (implying relatively longer lifetimes);
2. stronger suppression of bottom particle decay rates; and
3. that bottom decays preferentially to charmed particles.

The matrix element V_{ud} can be determined from the comparison of muon and nuclear β-decay rates. The result $|V_{\text{ud}}| = 0.97418 \pm 0.00027$, along with the determination of $|V_{\text{us}}|$ and $|V_{\text{ub}}|$ from K and B decays, is compatible with unitarity for three generations,

$$|V_{\text{ud}}|^2 + |V_{\text{us}}|^2 + |V_{\text{ub}}|^2 = 1 \tag{31}$$

5.5.3
Neutral Current

The neutral-current interaction is parameterized by the parameter $\sin^2\theta_W$. This quantity can be extracted from a large number of different neutral-current processes. These measurements give consistent values, which supports the precision of the model. The most precise determination is provided by the studies of Z^0 made at electron–positron colliders LEP and SLC, which produced millions of Z^0 bosons as a resonance between 1989 and 1995, using a number of measurements of angular distribution and polarization asymmetries in the process $e^+e^- \to Z^0 \to \overline{f}f$. An average of these measurements gives

$$\sin^2\theta_W = 0.23119 \pm 0.00014 \tag{32}$$

5.5.4
W and Z Masses

The mass of the Z^0,

$$m_Z = 91.1876 \pm 0.0021\,\text{GeV} \tag{33}$$

is, together with G_F, is the most precisely determined parameter of the weak interaction. The mass of the W boson, $m_W = 80.398 \pm 0.025\,\text{GeV}$, has a larger experimental error. The ratio of the two masses is approximately equal to $(1 - \sin^2\theta_W)^{1/2}$. The slight disagreement arises from the quantum loop effect of the top quark. In fact, the mass of the top quark had been predicted from these measurements before its direct discovery.

5.6
Processes Induced by Weak Interactions

5.6.1
Decay of W and Z

The decay of weak bosons directly reflects their interaction. The charged W^- boson is expected to decay mainly into five modes: $e^-\overline{\nu}_e$, $\mu^-\overline{\nu}_\mu$, $\tau^-\overline{\nu}_\tau$, $d\overline{u}$, and $s\overline{c}$, with the approximate branching ratio of 1 : 1 : 1 : 3 : 3. This results from the universality of charged current, the factor of 3 coming from the three color degrees of freedom of quarks. The partial decay rate for the leptonic mode is calculated to be

$$\Gamma\,(W \longrightarrow e\overline{\nu}) = \frac{G_F m_W^3}{6\sqrt{2}\pi} \tag{34}$$

The measured total width is $\Gamma_W = 2.141 \pm 0.041\,\text{GeV}$.

The neutral weak boson Z^0 decays to any fermion–antifermion pair other than $t\overline{t}$. The decay rate is given by

$$\Gamma\left(Z \longrightarrow \overline{f}f\right) = \frac{NG_F m_W^3}{6\sqrt{2}\pi}\left(v_f^2 + a_f^2\right) \tag{35}$$

where $N = 1$ for leptons and 3 for quarks. The measured branching ratios are approximately 70% for hadrons ($q\overline{q}$ pairs), 20% for invisible final states (neutrino pairs), and 3.4% each for charged-lepton pairs (e^+e^-, $\mu^+\mu^-$, and $\tau^+\tau^-$). The measured total width is $\Gamma_Z = 2.4952 \pm 0.0023\,\text{GeV}$. These widths of W and Z bosons correspond to a lifetime of less than 10^{-24} s. In fact, the weak bosons have the shortest lifetimes we have ever encountered.

5.6.2
Weak Decay of Leptons

The electron, being the lightest charged particle, is absolutely stable. Neutrinos are also stable or at least have an extremely long lifetime. The heavier charged leptons, μ and τ, decay by the weak interaction to their associated neutrino plus a pair of leptons or quarks through a virtual W boson.

5.6.2.1 Muon Decay

The muon decays via the charged-current interaction

$$\mu^- \longrightarrow e^- + \overline{\nu}_e + \nu_\mu \tag{36}$$

The rate of this three-body decay is calculated to be

$$\Gamma\left(\mu \longrightarrow e\overline{\nu}_e\nu_\mu\right) = \frac{G_F^2 m_\mu^5}{192\pi^3} \tag{37}$$

in the approximation of neglecting the electron mass. This m^5 behavior, which can be understood in terms of dimensional analysis (noting that the amplitude is proportional to G_F), is typical of weak three-body decays such as β-decay and decays of charged leptons and heavy quarks (c, b).

The lifetime of muon, $\tau = \hbar/\Gamma \sim 2 \times 10^{-6}$ s, though apparently short, is much longer than the characteristic timescale for the muon, $\hbar/m_\mu c^2 \sim 10^{-23}$ s. This reflects the weakness of the weak interaction at this energy scale. Indeed, the muon lifetime would be of the order of the latter if the decay could be induced by stronginteraction.

5.6.2.2 τ-Decay

The τ-lepton can decay in the same way as the muon:

$$\tau^- \longrightarrow e^- + \overline{\nu}_e + \nu_\tau \tag{38}$$

$$\tau^- \longrightarrow \mu^- + \overline{\nu}_\mu + \nu_\tau \tag{39}$$

In addition, the larger mass of the tau enables semihadronic decays

$$\tau^- \longrightarrow \text{hadrons} + \nu_\tau \tag{40}$$

which at the quark level can be understood as $\tau^- \rightarrow d + \overline{u} + \nu_\tau$ (and a small contribution of $\tau^- \rightarrow \mathrm{s} + \overline{u} + \nu_\tau$). The final hadronic system can be a π^-, $\pi^-\pi^0$, or multimeson states with charge -1.

The partial decay rate for $\tau \rightarrow e\overline{\nu}_e\nu_\tau$ is obtained from Eq. (37) by replacing m_μ by m_τ. The total rate is approximately five times larger (counting two leptonic modes and one semihadronic mode with three colors). The measured leptonic branching fraction $B(\tau \rightarrow e\nu\nu) \sim 18\%$ implies that the hadronic mode is slightly enhanced, an effect of the strong interaction. The lifetime $\tau \sim 3 \times 10^{-13}$ s is much shorter than that of muon, reflecting the difference in mass.

5.6.3
Weak Decay of Hadrons

Hadrons that cannot decay via strong or electromagnetic interactions decay weakly. (The exceptions are the proton and bound neutron, whose decay is forbidden by baryon number conservation.) In general, the states with the lowest mass for a given flavor combination of constituent quarks can decay only by weak interaction. The following is the whole list of weakly decaying mesons, all of which are spin-0 negative-parity particles (their antiparticles are implicit): $\pi^+(u\overline{d})$, $K^+(u\overline{s})$, $K^0(d\overline{s})$, $D^+(c\overline{d})$, $D^0(c\overline{u})$, $D_s^+(c\overline{s})$, $B^+(u\overline{b})$, $B^0(d\overline{b})$, $B_s^0(s\overline{b})$, $B_c^+(c\overline{b})$ (the quark content of each meson is shown in the parentheses). There are also a number of baryons that decays weakly.

5.6.3.1 Neutron and Nuclear β-Decay

The neutron is stable if it is bound into a stable nucleus. However, a free neutron, which is ~1% more massive than a proton, is unstable and decays as

$$n \longrightarrow p + e^- + \overline{\nu}_e \tag{41}$$

with a lifetime of ~10^3 s. This lifetime (much longer than the muon) reflects the small available energy Q in the decay, the rate being proportional to its fifth power.

β-Decays of unstable nuclei are caused by the weak interaction. There are three classes of nuclear β-decays: β^--decay,

Tab. 5.2 Nuclear β-decays.

Category	Reaction	Elementary process
β^--Decay	$(Z, A) \longrightarrow (Z+1, A) + e^- + \overline{\nu}_e$	$n \longrightarrow p + e^- + \overline{\nu}_e$
β^+-Decay	$(Z, A) \longrightarrow (Z-1, A) + e^+ + \nu_e$	$p \longrightarrow n + e^+ + \nu_e$
Electron capture	$(Z, A) + e^- \longrightarrow (Z-1, A) + \nu_e$	$p + e^- \longrightarrow n + \nu_e$

[a] (Z, A) denotes a nucleus with atomic number Z and mass number A, which consists of Z protons and $A - Z$ neutrons.

β^+-decay, and electron capture. All of these are due to the reaction (41) or crossed processes as is summarized in Table 5.2.

In nuclear β-decays, the nucleons can be treated as nonrelativistic. In the nonrelativistic limit, nonvanishing matrix elements arises from the time component of the hadronic vector current (Fermi transition) and the space component of the axial current (Gamow–Teller transition). The former does not change the nuclear spin ($\Delta J = 0$), and the latter has $|\Delta J| = 0, 1$. For example, the decay ${}^{14}_{8}\text{O}(0^+) \rightarrow {}^{14}_{7}\text{N}(0^+)$ (half-life 71 s) is a Fermi transition (spin-parity J^P is shown in parentheses), ${}^{6}_{4}\text{He}(0^+) \rightarrow {}^{7}_{4}\text{Li}(1^+)$ (0.8 s) is a Gamow–Teller transition, and ${}^{3}_{1}\text{H}(\frac{1}{2}^+) \rightarrow {}^{3}_{2}\text{He}(\frac{1}{2}^+)$ (12.3 years) is the mixture of both. Widely varying lifetimes mainly arise from the Q^5 dependence of the rate.

Higher-order and relativistic effects cause “forbidden” transitions with much longer lifetimes. The natural radioactive isotope ${}^{40}_{19}\text{K}$ mainly decays as ${}^{40}_{19}\text{K}(4^-) \rightarrow {}^{40}_{20}\text{Ca}$ with a half-life of 10^9 years. This decay is a third-forbidden transition.

5.6.3.2 **Pion Decay**

Charged pions decay dominantly into a muon and a muon neutrino. The decay rate is

$$\Gamma\left(\pi^+ \longrightarrow \mu^+\nu_\mu\right) = \frac{G_F^2 |V_{\text{ud}}|^2 m_\pi f_\pi^2 m_\mu^2}{8\pi} \times \left(1 - \frac{m_\mu^2}{m_\pi^2}\right)^2 \tag{42}$$

where the pion decay constant f_π is defined by the matrix element for a pion to be created from the vacuum via the charged-current operator (without V_{ud}). From the lifetime of pion, one deduces $f_\pi = 131\,\text{MeV}$. (This quantity is a measure of the size of spontaneous breaking of hadronic chiral symmetry).

Although energetically more preferable, the decay $\pi^+ \rightarrow e^+\nu_e$ is suppressed by a factor $(m_e/m_\mu)^2$, which is the consequence of the current structure of the weak interaction.

5.6.3.3 **Decay of Hadrons Containing Strange Quarks**

The lowest-lying strange mesons decay weakly. The charged kaon K^+ has a lifetime of 1×10^{-8} s. The main decay modes are $\mu^+\nu_\mu$ (64% branching ratio), $\pi^+\pi^0$ (21%), $\pi^+\pi^+\pi^-$ (6%), $\pi^+\pi^0\pi^0$ (2%), $\pi^0\mu^+\nu_\mu$ (3%), and $\pi^0e^+\nu_e$ (5%). The neutral kaon K^0 has two distinct lifetimes as discussed in Section 5.6.5. There are six baryons containing one or more strange quarks (the rest being u or d quarks) that decay weakly. These states are called *hyperons* and all of them have lifetimes in the 10^{-10} s range. For example, the Λ baryon (quark content uds) mainly decays to $p\pi^+$ (64%) and $n\pi^0$ (36%) with a lifetime of 3×10^{-10} s; the Ξ^- baryon (quark content dss) decays to $\Lambda\pi^-$ (lifetime 1.6×10^{-10} s).

5.6.3.4 Decay of Hadrons Containing Heavy Quarks

Weak decay of hadrons containing a charm or bottom quark can be considered in the first approximation as the decay of the heavy quark in the hadron, where the other quarks (spectators) do not play any major role.

Weakly decaying charmed hadrons have lifetimes in the range of 10^{-13}–10^{-12} s. Bottom hadrons have lifetimes of about 1.5×10^{-12} s. The longer lifetime of bottom particles despite the larger mass of b is the consequence of the suppression due to the small Kobayashi–Maskawa matrix element V_{cb}.

5.6.4
Top-Quark Decay

The top quark has an exceptionally large mass of 171.2 ± 2.1 GeV and, unlike other quarks and leptons, can decay to a W boson and to a quark (mostly b). The expected decay width is about 1 GeV. Because of its short lifetime, a top quark produced in a particle collision has no time to form a color-neutral hadron before decaying and no top hadron can practically exist.

5.6.5
Flavor Oscillation: Neutral-Kaon and B-Meson Systems

Neutral flavored mesons such as K^0 and B^0 show an interesting quantum-mechanical phenomenon of particle–antiparticle mixing. The K^0 meson, with its quark content ($d\bar{s}$), is distinct from its antiparticle $\overline{K}^0$ ($s\bar{d}$)). Since the weak interaction can change flavor, transition between these two states is possible at the second order of the effective weak interaction (23). The transition amplitude is large enough that the mixing can occur before the meson decays. Because of this effect, the decay of the produced K^0 does not obey the typical exponential decay law but can be described as the superposition of two exponentially decaying amplitudes corresponding to the following linear combination of K^0 and $\overline{K}^0$ states:

$$|K_S\rangle \sim \frac{1}{\sqrt{2}}\left[(1+\varepsilon)\left|K^0\right\rangle + (1-\varepsilon)\left|\overline{K}^0\right\rangle\right] \tag{43}$$

$$|K_L\rangle \sim \frac{1}{\sqrt{2}}\left[(1+\varepsilon)\left|K^0\right\rangle - (1-\varepsilon)\left|\overline{K}^0\right\rangle\right] \tag{44}$$

Up to the tiny CP-nonconserving effect represented by the parameter ε (see next subsection), the state $K_S(K_L)$ is a CP eigenstate with eigenvalue $+1$ (-1). K_S decays to CP $= +1\, \pi^+\pi^-$ and $\pi^0\pi^0$ states and K_L to CP $= -1 3\pi$ states. These two states have drastically different lifetimes: $\tau(K_S) = 0.9 \times 10^{-10}$ s, $\tau(K_L) = 5 \times 10^{-8}$ s. The oscillation frequency is $1/3.8 \times 10^{-10}\ \mathrm{s}^{-1}$.

Similar mixing is observed in the B^0–$\overline{B}^0$ system, except that the two lifetimes are not very different. On the other hand, only a very small mixing phenomenon is observed for the neutral charmed meson D^0. In fact, the mixing amplitude is expected to be small because of cancellation between s and d intermediate states. Furthermore, the charm decay has no Kobayashi–Maskawa suppression, unlike s or b.

5.6.6
CP-Invariance Violation

Even after the establishment of parity nonconservation in weak interaction, the discovery of the CP-nonconserving decay $K_L \to \pi\pi$ (Christenson *et al.*, 1964) was not anticipated. After this observation, CP-invariance violation was known only in the K^0–$\overline{K}^0$ system for more than three decades until the advent of the B factories. Unlike parity nonconservation, the size of CP violation in kaon decays is very small as it is characterized by the parameter $|\varepsilon| \sim 2 \times 10^{-3}$ in Eqs (43) and (44). In the framework of the Kobayashi–Maskawa scheme, the smallness of the effect can be understood, as it is a higher-order process involving the top quark (the amplitude is proportional to λ^6).

Decays of B mesons are suppressed by a small off-diagonal element of the mixing matrix. This leads to a large CP-nonconserving effect in some decay modes of B mesons, for example, a certain asymmetry between the decays $B \to J/\psi K_S$ and $\overline{B} \to J/\psi K_S$. Study of CP violation in a number of B decay modes at the B factories built at KEK and SLAC confirmed these predictions of the Kobayashi–Maskawa mechanism with a huge amount of B meson data (Abe *et al.*, 2001; Aubert *et al.*, 2001).

5.6.7
Neutrino Interaction

Neutrinos have no strong and electromagnetic interactions and interact only weakly. The cross section of neutrinos with matter is therefore so small that a neutrino can easily penetrate the Earth. The cross section grows as center-of-mass energy squared as long as $E_{\mathrm{c.m.}} < m_W$. For example, the neutrino–nucleon cross section is measured to be $\sigma \sim 0.7 \times 10^{-38} E_\nu\,(\mathrm{GeV})\,\mathrm{cm}^2$ for an incident neutrino energy of E_ν larger than a few giga electronvolts. The neutrino–electron cross section is a few orders of magnitude smaller for the same incident energy.

5.7
Gauge Theory of the Weak Interaction: the Standard Model

The three known fundamental interactions – strong, electromagnetic, and weak – can be very well described by gauge field theory. Gauge theory (Yang and Mills, 1954; Utiyama, 1956; Gell-Mann and Glashow, 1961) is a field theory invariant under gauge transformation. The gauge transformations form a continuous group G (the gauge group) under which the fields transform as representations of the group. Gauge theories are characterized by the invariance under local gauge transformation, the fields at different space-time transforming under different group elements of G: $\phi(x) \to g(x)\phi(x)$, $g \in G$. The theory possesses massless spin-1 bosons called *gauge bosons*, which transform as the adjoint representation of the gauge group. It can also have the spin-$\frac{1}{2}$ and scalar fields in any representation of the gauge group.

The standard model (Glashow, 1961; Weinberg, 1967; Salam, 1968; Glashow, Iliopoulos and Maiani, 1970; Bouchiat, Iliopoulos and Meyer, 1972) is characterized by the gauge group $G = \mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1)$. Besides the gauge fields, the model contains the fermion fields corresponding to three generations of quarks and leptons and a scalar field called the *Higgs field*. The representation contents of these fields are listed in Table 5.1. Note that the model is chiral, that is, the left- and right-handed fermion fields are in different representations. This is the origin of parity noninvariance of the weak interaction.

The interaction of gauge bosons with every particle in the model is strongly constrained by the gauge symmetry. The interaction is unique once the representation of the particle is fixed. The universality of the gauge interactions is the consequence of this fact.

Gauge symmetry prohibits a mass term for gauge bosons. On the other hand, the weak bosons are massive. This apparent discrepancy can be resolved by the notion of spontaneous symmetry breaking and the Higgs mechanism.

If the lowest energy state (the vacuum) is not invariant under the symmetry of the theory (Lagrangian), the physical spectrum loses the expected manifestation of the symmetry. This situation is called *spontaneous symmetry breaking*. In general, spontaneous breakdown of a continuous symmetry is associated with the appearance of a massless scalar boson, called the *Nambu–Goldstone boson* (Nambu and Jona-Lasinio, 1961; Goldstone, 1961; Goldstone, Salam and Weinberg, 1962). However, if the symmetry is a local gauge symmetry, the gauge boson associated with the symmetry becomes massive and the Nambu–Goldstone boson does not appear in the physical spectrum. This phenomenon is called the *Higgs mechanism* (Englert and Brout, 1964; Higgs, 1964a; 1964b; 1966; Guralnik, Hagen and Kibble, 1964; Kibble, 1967).

In the standard model, the potential for the Higgs field can be constructed such that the Higgs field develops a vacuum expectation value. It breaks the $\mathrm{SU}(2) \times \mathrm{U}(1)_Y$ symmetry down to a subgroup $\mathrm{U}(1)_Q$, a linear combination of $\mathrm{U}(1)_Y$, and one of the generators in $\mathrm{SU}(2)$. Three gauge fields that are associated to the three broken generators become massive, which are $W^\pm$ and Z^0 bosons. The remaining state corresponding to the unbroken group is the massless photon.

The masses of the quarks and leptons also have to be generated by spontaneous symmetry breakdown, since gauge symmetry forbids mass terms for these chiral fermions. In the standard model, fermion-mass generation is accomplished by the Yukawa coupling of the fermion to the Higgs field. The generated mass is proportional to the product of the Yukawa coupling and the Higgs vacuum expectation value.

5.8
The Higgs Boson

In the standard model, one physical scalar particle remains from the Higgs field after spontaneous symmetry breaking. This particle has the same quantum numbers as the vacuum, and its coupling strength to a particle is proportional to the mass of the particle. It thus interacts with electrons and nucleons extremely weakly. Its couplings with the W, Z bosons, and top quark are sizable. The preferred decay modes of the Higgs boson are be to a pair of heaviest possible particles which are energetically allowed, such as $b\bar{b}$, W^+W^-, ZZ, depending on its mass.

The mass of the Higgs boson depends on the self-coupling of the Higgs field and cannot be predicted. At present, there is no evidence for the Higgs boson. The current best lower limit of 114 GeV on its mass is derived from the search for the associated production of a Higgs and a Z^0 boson in e^+e^- annihilation. Within the framework of the minimal standard model, available precision data of the Z boson properties and other weak interaction processes imply that the Higgs boson mass should be smaller than 185 GeV. The forthcoming experiments at the LHC (Large Hadron Collider), a proton–proton collider with a center-of-mass energy of 14 TeV, which is scheduled to start running in 2009, will be able to discover the Higgs particle.

In contrast to the three fundamental interactions, all the electromagnetic, strong, and weak interactions are mediated by spin-1 gauge bosons (may be called *vector-type*), the interactions of the Higgs boson (Yukawa interactions and Higgs self-interactions) are not vector-type and are not governed by gauge symmetry. Discovery of the Higgs boson would thus imply the discovery of a new type of fundamental interactions.

5.9
Origin of Electroweak Symmetry Breaking

Although the standard model gives a consistent description of the three interactions and has passed numerous experimental tests, only the gauge-interaction part has been verified. There is no evidence for or against the existence of the extra forces introduced by the model, namely, the Higgs self-interactions and Yukawa interactions, which are vital to explain the masses of weak bosons and fermions. Theoretically, the model contains as many as 18 parameters, which have to be determined by experiments. Despite being a most economical model, it is unlikely to be the final theory. Generally, two possible directions to replace the standard model have been pursued.

One is a weak-coupling scenario in which the model is viable up to very high energies (typically 10^{16} GeV, the scale of grand unified theories, or the Planck scale 10^{19} GeV, where the gravitational interaction becomes strong). Theories of this type tend to have a problem of naturalness (Susskind, 1979; 't Hooft, 1980; Veltman, 1981) to generate and stabilize the weak scale, which is extremely small compared to the typical scale of the theory. The stability problem is solved if the theory possesses supersymmetry (Wess and Zumino, 1974a; 1974b; Ferrara and Zumino, 1974), which is a symmetry between bosons and fermions.

The supersymmetric version of the standard model contains all particles in the standard model and their superpartners, which have the same quantum numbers except that the spin differs by $\frac{1}{2}$.

(To be precise, the model contains two doublets of Higgs fields.) The model implies the existence of a relatively light neutral Higgs boson with mass $\lesssim$130 GeV. From the naturalness of the theory, the superpartners are expected to have masses of the order of TeV.

The other is a strong-coupling scenario, in which the electroweak symmetry is broken by a strong interaction in the TeV energy region (Weinberg, 1979; Susskind, 1979). No light Higgs boson is expected in this scenario. Instead, W bosons strongly interact at TeV energies, at which new phenomena such as WW resonances are expected. A typical example is the so-called technicolor model. Precision measurements at Z^0 put strong constraints on this scenario (Peskin and Takeuchi, 1990; Peskin and Takeuchi, 1992), and a number of models have already been excluded. This type of symmetry breaking may be realized if the dimension of the space-time is more than four, the extra dimensions being compactified with a scale of 10^{-18} cm (corresponding to TeV scale). The existence of extra dimensions is anticipated in superstring theory, a prime candidate for the unification of all interactions including gravity. In fact, the consistency of the theory requires that the space-time should have exactly 10 dimensions.

Determination of the origin of the symmetry breaking is the most important physical target of particle physics at present. It requires an accelerator, which can reach the energy scale of a TeV or more. The LHC is expected to shed lights on this question with the discovery of the Higgs boson or else. An electron–positron linear colliders in the TeV energy range should further clarify the physics of symmetry breaking with its clean environment.

5.10
Neutrino Masses and Mixing

Experiments that directly tried to measure neutrino masses have yielded negative results. The limit for the electron neutrino mass is $m_{\nu_e} < 2\text{eV}$, which is many orders of magnitude smaller than the masses of the charged leptons and quarks. Limits for the other two are not as strong.

Since neutrinos are neutral, two types of mass are possible: the usual Dirac mass and Majorana mass. A Majorana fermion (fermion with a Majorana mass) is identical to its antiparticle. Although neutrinos are exactly massless in the standard model by construction, neutrino masses arise in various extensions of the standard model. In grand unified theories, a very small Majorana mass can be generated naturally (Gell-Mann, Ramond and Slansky, 1979; Yanagida, 1979).

If neutrinos have Majorana masses, lepton number is not conserved and processes such as neutrinoless double β-decay $(Z,\ A) \rightarrow (Z+2,\ A) + 2e^-$ are allowed. There is no firm observation of such processes, implying that the Majorana mass of the electron neutrino should be less than a few electron volts.

Finite neutrino masses are generally accompanied by the mixing of lepton flavor. The mixing causes a phenomenon called *neutrino oscillation* (Maki, Nakagawa and Sakata, 1962; Gribov and Pontecorvo, 1969). For example, if the electron and muon neutrinos mix with a mixing angle θ to form mass eigenstates ν_1 and ν_2, ν_μ produced by a weakly decaying particle can transform itself into ν_e when it is detected by a reaction with matter. The probability of its occurrence is given (for small neutrino masses) by

$$P(\nu_\mu \longrightarrow \nu_e) = \sin^2 2\theta \sin^2\left(\frac{\Delta m^2 l}{4E}\right) \tag{45}$$

where Δm^2 is the mass-squared difference of the two states, l is the distance traveled by the neutrino between its production and detection, and E is the neutrino energy. Oscillation among three species of neutrinos involves two mass-squared differences and a unitary mixing matrix called *Maki–Nakagawa–Sakata matrix*, which is the lepton counterpart of Cabibbo–Kobayashi–Maskawa matrix and contains three mixing angles and CP violating phase(s).

There has long been indications of this phenomenon in the data of neutrinos from the Sun (solar neutrinos) (Davis, 1994; Hirata *et al.*, 1989; Anselmann *et al.*, 1992a; 1992b; Abdurashitov *et al.*, 1994) and neutrinos from cosmic-ray interactions in the atmosphere (atmospheric neutrinos) (Hirata *et al.*, 1992; Fukuda *et al.*, 1994). New generation experiments have clearly demonstrated that oscillations do occur for both solar (Fukuda *et al.*, 1998a; Ahmad *et al.*, 2002) and atmospheric neutrinos (Fukuda *et al.*, 1998b). Experiments using controlled sources of neutrinos confirmed these findings. These utilize neutrinos from reactors (for the solar neutrino mixing) (Eguchi *et al.*, 2003) and accelerators (for the atmospheric neutrino range) (Ahn *et al.*, 2003; Michael *et al.*, 2006) with a long distance between the source and the detector.

The mass-squared differences and mixing angles associated with the two kinds of oscillations have been measured with accuracy. The atmospheric neutrino oscillation is observed between ν_μ and ν_τ with Δm^2 of 2 to 3×10^{-3} eV^2 and almost one-to-one mixing ($\sin^2 2\theta > 0.92$). The solar neutrino oscillation involves a smaller Δm^2 of $(8.0 \pm 0.3) \times 10^{-5}\,\mathrm{eV}^2$ with $\sin^2 2\theta = 0.86^{+0.03}_{-0.04}$ between ν_e and a mixture of ν_μ and ν_τ.

A characteristic pattern of the neutrino mass-mixing emerges from these measurements. The three mass eigenstates have mass differences of at most 0.05 eV and two of them (ν_1, ν_2) are more closely degenerate within 0.01 eV or less. At present, it is not known either the two close-lying states are lighter (called *normal hierarchy*) or heavier (inverted hierarchy) than ν_3. The lightest state can either be massless or have a finite mass smaller than ~1 eV. Both the measured mixing angles are large. In particular, ν_3 is the nearly equal mixture of ν_μ and ν_τ. A third mixing angle is proportional to the ν_e component in ν_3, for which we have no evidence so far. If this angle is nonzero, a CP violating phase in the neutrino mixing matrix can exist, resulting in a difference in neutrino and antineutrino oscillations. If neutrinos are Majorana fermions, additional CP violation is possible in lepton-number violating processes.

Neutrinos pose further questions to the origin of the masses. Neutrino masses are extremely small. The mixing angles are large, which is in sharp contrast to the quarks. These facts may indicate that the origin of neutrino masses may be different from those of quarks and charged leptons.

5.11
Beyond the Standard Model

New processes which are forbidden or extremely suppressed in the Standard Model are the possible windows for new physics beyond the Standard Model. Just as weak decays have been known long before we can directly study the weak bosons, observation of these processes would indicate the existence of a new physics at very high energies currently inaccessible at direct experiments.

Decays such as $\mu^- \to e^-\gamma$ or $\mu^- \to e^-e^-e^+$ violate lepton-flavor conservation and do not occur in the Standard Model. There is no evidence for such processes so far. Neutrino mass mixing can induce these decays but the effect is far too small to be observed. These processes may occur in models with supersymmetry at a rate observable in experiments in the near future.

Similarly important is the nonconservation of baryon number, which can cause the decay of protons and neutrons bound into the nuclei. Grand unified theories (Georgi and Glashow, 1974) generally predict instability of nucleons (Georgi, Quinn and Weinberg, 1974). The expected lifetime is $\sim 10^{33}$ years in some models, which is within the reach of ongoing experiments.

The electric dipole moment of the neutron d_n, a CP-invariance-violating quantity, has been investigated unsuccessfully. The present limit, $d_n < 3 \times 10^{-26} e$ cm, is consistent with the expectation of the Kobayashi–Maskawa scheme, but gives strong constraints on some other models of CP violation. This quantity, as well as the electric dipole moment of other particles including electron, can provide a clue to new physics.

There are evidences from cosmology that the Standard Model is not the final story. Familiar matter comprises only a small percentage of the energy density of the present universe, the rest being dark matter (one-fourth of the total) and dark energy (three fourths). There are no candidates in the Standard Model capable of explaining them. Inflation, a rapid expansion of the early universe which resulted in the uniformity of the current universe, requires a new field to induce it. Asymmetry of matter and antimatter in the universe necessitates the violation of CP symmetry, but the effect of Kobayashi–Maskawa phase is too small to explain the observed asymmetry. New physics at the tera electronvolt scale such as supersymmetry may answer some of these matters. Energy frontier experiments at the LHC and electron-positron linear colliders can probe hitherto unknown physics directly. On the other hand, precision experiments and searches for new processes at lower energies may reveal new physics at very high energy scale, which cannot be directly probed.

Glossary

Charge Conjugation: Operation that exchanges a particle and its antiparticle. All known interactions except the weak interaction are invariant under change conjugation. (Invariance under charge conjugation means that the Lagrangian does not change its form under the operation.)

CP Invariance: Invariance under a combined operation of charge conjugation and parity. The weak interaction violates CP invariance.

Helicity: Component of spin in the direction of the momentum of a particle. A spin-1/2 fermion has two independent states with helicity ±1/2. Helicity changes its sign under parity transformation.

Natural Units: Unit system with $h = c = 1$. In this system there is only one independent dimension, [mass] = [momentum] = [energy] = 1/[length] = 1/[time].

Parity: Parity means space inversion, changing the sign of all three space coordinates. Strong, electromagnetic, and gravitational interactions are invariant under the parity operation, although the weak interaction is not.

Time Reversal: Operation that changes the sign of time coordinate. According to CPT theorem, invariance under time reversal is equivalent to invariance under CP.

References

Abachi, S. *et al.* (D0 Collaboration) (**1995**) *Phys. Rev. Lett.*, **74**, 2632–2637.

Abdurashitov, J.N. *et al.* (SAGE Collaboration) (**1994**) *Phys. Lett. B*, **328**, 234–238.

Abe, F. *et al.* (CDF Collaboration) (**1994a**) *Phys. Rev. D*, **50**, 2966–3026.

Abe, F., *et al.* (CDF Collaboration) (**1994b**) *Phys. Rev. Lett.*, **73**, 225–231.

Abe, F. *et al.* (CDF Collaboration) (**1995**) *Phys. Rev. Lett.*, **74**, 2626–2631.

Abe, K. *et al.* (Belle Collaboration) (**2001**) *Phys. Rev. Lett.*, **87**, 091802.

Ahmad, Q.R. *et al.* (Belle Collaboration) (**2002**) *Phys. Rev. Lett.*, **89**, 011301.

Ahn, M.H. *et al.* (K2K Collaboration) (**2003**) *Phys. Rev. Lett.*, **90**, 041801.

Anselmann, P. *et al.* (Gallex Collaboration) (**1992a**) *Phys. Lett. B*, **285**, 376–389.

Anselmann, P. *et al.* (Gallex Collaboration) (**1992b**) *Phys. Lett. B*, **285**, 390–397.

Arnison, G. *et al.* (UA1 Collaboration) (**1983a**) *Phys. Lett.*, **122B**, 103–116.

Arnison, G. *et al.* (UA1 Collaboration) (**1983b**) *Phys. Lett.*, **126B**, 398–410.

Aubert, B. *et al.* (BABAR Collaboration) (**2001**) *Phys. Rev. Lett.*, **87**, 091801.

Bagnaia, P. *et al.* (UA2 Collaboration) (**1983**) *Phys. Lett.*, **129B**, 130–140.

Banner, M. *et al.* (UA2 Collaboration) (**1983**) *Phys. Lett.*, **122B**, 476–485.

Bouchiat, C., Iliopoulos, J., and Meyer, Ph. (**1972**) *Phys. Lett.*, **38B**, 519–523.

Cabibbo, N. (**1963**) *Phys. Rev. Lett.*, **10**, 531–533.

Christenson, J.H., Cronin, J.W., Fitch, V.L., and Turlay, R. (**1964**) *Phys. Rev. Lett.*, **13**, 138–140.

Davis, R. (**1994**) *Prog. Part. Nucl. Phys.*, **32**, 13–32.

Eguchi, K. *et al.* (KamLAND Collaboration) (**2003**) *Phys. Rev. Lett.*, **90**, 021802.

Englert, F. and Brout, R. (**1964**) *Phys. Rev. Lett.*, **13**, 321–323.

(a) Fermi, E. (**1934**) *Z. Phys.*, **88**, 161–177 (in German); (b) (**1934**) *Nuovo Cimento*, **11**, 1–19 (in Italian).

Ferrara, S. and Zumino, B. (**1974**) *Nucl. Phys.*, **B79**, 413–421.

Friedman, J.I. and Telegdi, T.L. (**1957**) *Phys. Rev.*, **105**, 1681–1682.

Fukuda, Y. *et al.* (KAMIOKANDE Collaboration) (**1994**) *Phys. Lett. B*, **335**, 237–245.

Fukuda, Y. *et al.* (Super-Kamiokande Collaboration) (**1998a**) *Phys. Rev. Lett.*, **81**, 1158–1162.

Fukuda, Y. *et al.* (Super-Kamiokande Collaboration) (**1998b**) *Phys. Rev. Lett.*, **81**, 1562–1567.

Garwin, R.L., Lederman, L.M., and Weinrich, M. (**1957**) *Phys. Rev.*, **105**, 1415–1417.

Gell-Mann, M. and Glashow, S.L. (**1961**) *Ann. Phys. (N.Y.)*, **15**, 437–460.

Gell-Mann, M. and Lévy, M. (**1960**) *Nuovo Cimento*, **16**, 705–726.

Gell-Mann, M., Ramond, P., and Slansky, R. (**1979**) *Supergravity* (eds P. van Nieuwenhuizen and D.Z. Freedman), North-Holland, Amsterdam, pp. 315–318.

Georgi, H. and Glashow, S.L. (**1974**) *Phys. Rev. Lett.*, **32**, 438–441.

Georgi, H., Quinn, H.R., and Weinberg, S. (**1974**) *Phys. Rev. Lett.*, **33**, 451–454.

Glashow, S.L. (**1961**) *Nucl. Phys.*, **22**, 579–588.

Glashow, S.L., Iliopoulos, J., and Maiani, L. (**1970**) *Phys. Rev. D*, **2**, 1285–1292.

Goldstone, J. (**1961**) *Nuovo Cimento*, **19**, 154–164.

Goldstone, J., Salam, A., and Weinberg, S. (**1962**) *Phys. Rev.*, **127**, 965–970.

Gribov, V. and Pontecorvo, B. (**1969**) *Phys. Lett. B*, **28**, 493–496.

Guralnik, G.S., Hagen, C.R., and Kibble, T.W.B. (**1964**) *Phys. Rev. Lett.*, **13**, 585–587.

Hasert, F.J. *et al.* (**1973**) *Phys. Lett. B*, **46**, 138–140.

Hirata, K.S. *et al.* (KAMIOKANDE Collaboration) (**1989**) *Phys. Rev. Lett.*, **63**, 16–19.

Hirata, K.S. *et al.* (KAMIOKANDE Collaboration) (**1992**) *Phys. Lett. B*, **280**, 146–152.

Higgs, P.W. (**1964a**) *Phys. Lett.*, **12**, 132–133.

Higgs, P.W. (**1964b**) *Phys. Rev. Lett.*, **13**, 508–509.

Higgs, P.W. (**1966**) *Phys. Rev.*, **145**, 1156–1163.

't Hooft, G. (**1980**) *Recent Developments in Gauge Theories, Proceedings of the Cargèse Summer Institute, 1979*, NATO Advanced Study Institute Series B: Physics (eds G. 't Hooft, C. Itzykson, A. Jaffe, H. Lehmann, P.K. Mitter, I.M. Singer, and R. Stora), Vol. **59**, Plenum, New York, pp. 135–157.

Kibble, T.W.B. (**1967**) *Phys. Rev.*, **155**, 1554–1561.

Kobayashi, M. and Maskawa, T. (**1973**) *Prog. Theor. Phys.*, **49**, 652–657.

Lee, T.D. and Yang, C.N. (**1956**) *Phys. Rev.*, **104**, 254–258.

Maki, Z., Nakagawa, M., and Sakata, S. (**1962**) *Prog. Theor. Phys.*, **28**, 870–880.

Michael, D.G. *et al.* (MINOS Collaboration) (**2006**) *Phys. Rev. Lett.*, **97**, 191801.

Nambu, Y. and Jona-Lasinio, G. (**1961**) *Phys. Rev.*, **122**, 345–358.

Peskin, M.E. and Takeuchi, T. (**1990**) *Phys. Rev. Lett.*, **65**, 964–967.

Peskin, M.E. and Takeuchi, T. (**1992**) *Phys. Rev. D*, **46**, 381–409.

Salam, A. (**1968**) *Elementary Particle Theory* (N. Svartholm), Almqvist and Wiksell, Stockholm, pp. 367–377.

Susskind, L. (**1979**) *Phys. Rev. D*, **20**, 2619–2625.

Utiyama, R. (**1956**) *Phys. Rev.*, **101**, 1597–1607.

Veltman, M. (**1981**) *Acta Phys. Pol.*, **B12**, 437–457.

Weinberg, S. (**1967**) *Phys. Rev. Lett.*, **19**, 1264–1266.

(a) Weinberg, S. (**1976**) *Phys. Rev. D*, **13**, 974–976; (b) addendum (**1979**) **19**, 1277–80.

Wess, J. and Zumino, B. (**1974a**) *Nucl. Phys.*, **B70**, 39–50.

Wess, J. and Zumino, B. (**1974b**) *Nucl. Phys.*, **B78**, 1–13.

Wolfenstein, L. (**1983**) *Phys. Rev. Lett.*, **51**, 1945–1947.

Wu, C.S. *et al.* (**1957**) *Phys. Rev.*, **105**, 1413–1415.

Yanagida, T. (**1979**) *Proceedings of the Workshop on the Unified Theory and the Baryon Number of the Universe, KEK, Tsukuba, 1979* (eds O. Sawada and A. Sugamoto), National Laboratory for High Energy Physics, Tsukuba, pp. 95–98.

Yang, C.N. and Mills, R.L. (**1954**) *Phys. Rev.*, **96**, 191–195.

Further Reading

Abers, E.S. and Lee, B.W. (**1973**) Gauge theories, *Phys. Rep.*, **9**, 1–142. An early review of gauge theory, but it has not lost its status as a best account despite limitations due to its age.

Amsler, C. *et al.* (Particle Data Group) (**2008**) Review of particle physics, *Phys. Lett. B*, **667**, 1–1340. Gives a most recent comprehensive summary of elementary-particle data as well as nice short reviews on various subjects of high-energy physics. It is widely used and quoted by particle physicists and is updated every other year. Most of the experimental data in this article are quoted from here.

Cahn, R.N. and Goldhaber, G. (eds) (**1989**) *The Experimental Foundations of Particle Physics*, Cambridge University Press, Cambridge. A nice collection of important original papers on experimental high-energy physics, a large part of which is related to the weak interaction. Also contains extensive introductory writeups.

Cheng, T.-P. and Li, L.-F. (**1984**) *Gauge Theory of Elementary Particle Physics*, Clarendon Press, Oxford. A well-organized textbook on the subject at the graduate level. A companion volume (Problems and Solutions, 2000) is also available.

Georgi, H. (**1984**) *Weak Interactions and Modern Particle Theory*, Benjamin/Cummings, Menlo Park. Gives a concise but very nice and clear account of various aspects of the weak interaction based on gauge theory.

Halzen, F. and Martin, A.D. (**1984**) *Quarks and Leptons: An Introductory Course in Modern Particle Physics*, Wiley, New York. A recommendable undergraduate-level textbook on high-energy physics based on gauge theories.

Marshak, R.E., Riazuddin, and Ryan, C.P. (**1969**) *Theory of Weak Interactions in Particle Physics*, Wiley-Interscience, New York. A treatise on the subject in the pre-gauge-theory area.

Nilles, H.P. (**1984**) Supersymmetry, supergravity and particle physics, *Phys. Rep.*, **110**, 1–162. A standard review on application of super-symmetry to particle physics.

Pais, A. (**1986**) *Inward Bound: Of Matter and Forces in the Physical World*, Clarendon Press, Oxford. Describes the history of high-energy physics since the discovery of x rays in great detail from a professional viewpoint and includes an impressive list of literature.

Peskin, M.E. and Schroeder, D.V. (**1995**) *An Introduction to Quantum Field Theory*, Addison-Wesley, Reading. A very modern textbook on field theory. Unlike most textbooks, the authors put particular emphasis on its connection to physical phenomena.

Weinberg, S. (**1995, 1996, 2000**) *The Quantum Theory of Fields, I, II, III*, Cambridge University Press, Cambridge. A very solid modern textbook on quantum field theory. Volume II treats non-Abelian gauge theories and related concepts that are necessary for the modern understanding of the weak interactions. Volume III is devoted to supersymmetry.